

音声記号等で表記された言語資料の  
マークアップとコンピュータ処理

**Annotation and Computer Processing of Language Resources  
in Non-Latin Scripts and Phonetic Transcription**

課題番号 15202008  
Grant No. 15202008

平成15年度～平成17年度

科学研究費補助金基盤研究(A)研究成果報告書

2003–2005 Grants-in-Aid for Scientific Research (A) Research Report

研究成果報告書  
Research Report

平成 18 年 3 月  
March 2006

研究代表者 松村 一 登  
(東京大学大学院人文社会系研究科教授)

Project Leader: Kazuto Matsumura  
(Graduate School of Humanities and Sociology, University of Tokyo)



# 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—

千葉 庄寿

麗澤大学外国語学部

[schiba@reitaku-u.ac.jp](mailto:schiba@reitaku-u.ac.jp)

## 1. はじめに

麗澤大学の附属機関として2003年10月に開設された言語研究センター<sup>1</sup>のプロジェクトとして、昨年度からプロジェクト「言語研究のための多言語データベースの構築」(以下「言語情報学プロジェクト」)が始動している。<sup>2</sup> このプロジェクトの最も大きな目標は、言語学的な分析に必要な検索・統計処理機能を備えた多言語の電子化されたテキスト(コーパス)の利用環境を整備すること、情報処理技術を研究に活用するための知識を習得し活用するための支援体制を整備すること、の2つである。

科学研究費補助金によるプロジェクト「音声記号等で表記された言語資料のマークアップとコンピュータ処理」(文部科学省科学研究費補助金 基盤研究 (A)

---

<sup>1</sup> 麗澤大学言語研究センター公式 URL:

<http://www.FL.reitaku-u.ac.jp/LINC/>

<sup>2</sup> プロジェクト URL:

<http://www.FL.reitaku-u.ac.jp/LINC/projects/langTech/>

(2)；課題番号：15202008；研究代表者：松村一登（東京大学）と非常に近い位置づけにある上記プロジェクトは，その活動の一環として，既に公開・販売されているコーパスの収集を進めている。研究開始当初，収集したコーパスは，統一的なインターフェースをもちいた検索システムを構築して順次導入し，共同研究利用者の便宜を図る予定であった。しかし，収集がある程度進んだ段階で，資料収集の段階で，コーパスの中身，つまりデータがどのように記述されているかがコーパスごとに大きく異なり，そのため汎用的に使える検索システムを構築することが非常に難しい，ということが分かったのである。

本稿では，電子化されたテキストを構造化する際にコーパスの形式自体のもつべき要件は何かを検討する。大規模なコーパスを用いた言語研究にあたり非常に有益であると考えられる汎用のコーパス検索システムの構築とその利用のための研究者の技術教育という2点を視野に入れながら，電子化された言語データの構造化のための技術標準である XML (Extensible Markup Language) の規格に沿った言語データを作成することで，どのような検索が可能になるかを示し，言語データの XML 化を念頭に置き，コーパスを利用する言語研究者の新しい知識基盤として，テキスト処理の基礎知識としての「Unicode」と「正規表現」に加え，「XML の基礎知識（検索のための関連規格 XPath の基礎知識を含む）」という新たな柱を設定することを提案する。

## 2. コンピュータによる用例検索の諸形式

まず，コンピュータ上でテキストの用例検索がどのように行われるかを確認しておきたい。

検索ツールを用いる場合、検索は「行」を単位としておこなわれることが多い。キーワード（パターン）を指定することで、そのキーワード（パターン）を含む行が抽出され、リストされるわけである。grep 検索といわれるこの検索形式で得られるデータは、§3. で述べるように、そのコーパスで改行が何をあらわすかによって結果が変わってくる。

以下に、著者が作成した Unicode 対応の grep ツール UniGrep を紹介する。このツールは、Unicode で保存されたテキスト文書を読み込み、キーワードに一致する文字列を含む行を出力する。<sup>3</sup>

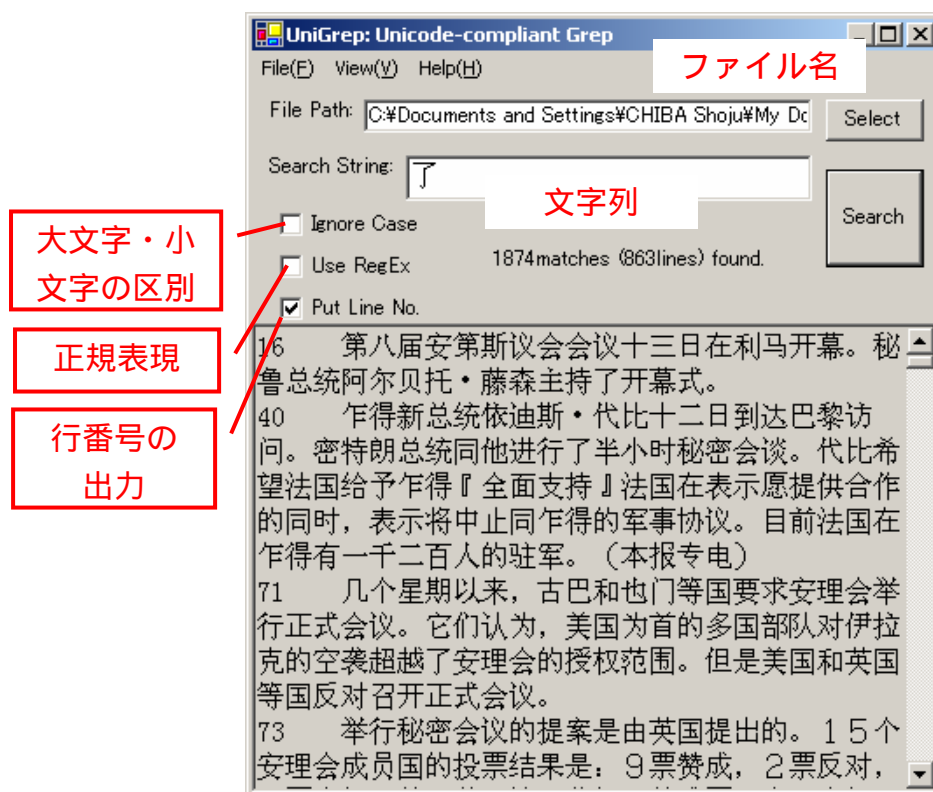


図 1 : UniGrep

<sup>3</sup> 同様の Unicode 対応の grep 検索機能をもつツールとして、Unicode 対応テキストエディタ EmEditor Professional (EmSoft によるシェアウェア) の「ファイルから検索」が挙げられる（メニューバーの[検索] [ファイルから検索]で起動する）。なお、EmEditor の Free 版、および Standard 版には複数ファイルを一括検索する当該機能はない。

また、言語研究で一般的な検索形式に KWIC (KeyWord In Context) 形式がある。検索キーワードを中央 node に置き、その左右に一定の長さの文脈を表示する。KWIC 形式の用例一覧を表示するソフトには、改行をまたいでテキストを検索してくれるものもある。<sup>4</sup>

### 3. 構造化された言語データとは

コーパスにテキストがどのような形で収録されているかにより、検索結果が変わったり、収集できるデータが異なったりする。従って、たとえ原テキストがそのまま電子化されているだけのデータであっても、テキストの電子的な表現方法には一考の余地がある。

例えば、改行記号が何を表すかを考えてみよう。コーパスとして利用される言語データの多くは、コンピュータの種類やソフトに関係なく利用できるよう、単純なテキスト文書として提供される。テキスト文書には、文字情報以外のレイアウト情報は全く入れることができないため、文字として扱われ、コードが決まっている「改行」(および「スペース」「タブ」) はテキストの構造を表現する有効な手段といえる。

Oxford Text Archive (OTA, URL: <http://ota.ahds.ac.uk>) が公開する

---

<sup>4</sup> 日本語を利用できる KWIC コンコーダンサーには、例えば Text Finder (フリーウェア, <http://www.biwa.ne.jp/~aka-san/textfinder.htm> からダウンロード可能。赤瀬川史朗氏作) がある。このソフトはまず用例を grep 検索し、その結果をさらに KWIC 形式で表示するので改行をまたいだデータは表示されない。

テキストデータには、以下のように参照した原著書の改行位置が維持されているものがある（図2、Lewis Carroll *Alice's Adventures in Wonderland* の冒頭）。このようなスタイルは古い文献など体裁を忠実に電子化することを意図しているものに多い。

1. DOWN THE RABBIT-HOLE

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

図2：『アリスの不思議な国』冒頭 (OTA)

また、青空文庫（以前のバージョンで提供されていたテキスト）には、改行がパラグラフの区切りを意味するものがある（図3）。

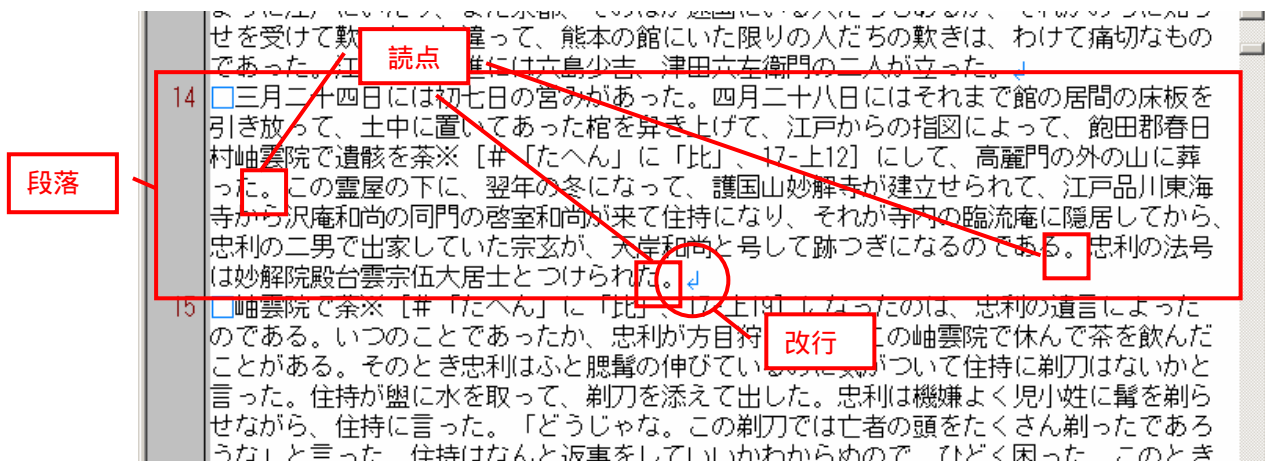


図3：森鷗外『阿部一族』の一部

(画面はテキストエディタ EmEditor (Emurasoft 開発)の一部)

これらのデータを grep 検索する場合、検索結果の内容が異なることは容易に想像できる。改行やタブ、スペースが何を意味しているかがコーパスによって異なる場合、汎用的な検索システムに grep 検索を単純に実装するだけでは充分でない場合がある。言語分析の観点からは、例えば、文末で改行するのが都合がよい場合もある。その場合、上記のデータをさらに編集する必要がある<sup>5</sup>が、その際文より大きな単位である段落をどのように電子的に表現するか、また段落先頭の字下げをどう表現するか、などの詳細を考慮する必要がある。

一方、上述のようなテキスト自体の構造に関わる情報だけでなく、研究目的に応じテキストにさまざまな言語学的な情報（アノテーション）を付加し、検索に利用することが考えられる。例えば、研究用に収集した用例について、研究者が個人で何らかの付加情報を加え、研究に利用することは頻繁に行われる。

<sup>5</sup> 上述の OTA のテキストデータでは文と文を 2 つのスペースで区切っているため、文末での改行の処理は比較的簡単である。また、日本語では読点(。)はもっぱら文末の区切り文字として使われるので、読点の後に改行を挿入することで大雑把な改行処理ができる。

以下の例（1—2）は、筆者が作成しているフィンランド語の用例データの一部である（用例の出典はヘルシンキ大学が提供するフィンランド語の週刊誌 *Suomen Kuvalehti* の 1987 年の記事コーパスである）。網掛けの部分が個人的に追加した情報である。

- (1) sk-36\_001507: Kaksi kertaa sanoo Kiviniemi matkalla ++kehottaneensa\_kehottaa kuljettaja Timo Kuivista\_OBJ\_PAR\_SG, 22, >>ottamaan\_ottaa rauhallisemmin .
- (2) sk-36\_001531: Hän on viime vuoden Suomen juniorimestari ja uraa\_OBJ\_ACC\_SG >>tasoittamaan\_tasoittaa hän ++sai\_saada 50000 markan Mobira-stipendin\_OBJ\_ACC\_SG , jonka Suomen Autourheilijoiden liitto jakaa lupaavalle , uraansa jatkavalle juniorille .

上記は不定詞が現れる用例を集めた用例集であり、主動詞の前に++を、不定詞の前には>>を置き、簡単に検索できるようにしてある。また主動詞と不定詞には、アンダーバーに続けて動詞の辞書形を置き、頻度の調査などに利用している。さらに、不定詞の目的語にも標識 (OBJ) をつけ、格形（例では単数分格形 PAR\_SG）がすぐ分かるようにしてある。また、元のコーパスの出典情報を用例の冒頭につけている（sk-36 がファイル名、001507, 001531 が行数）。

このような情報の付加を共用のコーパスに行う場合、どのような点に気をつけるべきであろうか。Leech (1993: 275) は付加情報の埋め込みによる弊害を避けるため、以下の原理を提案している：

1. 生のコーパスに簡単に戻せるようにすること
2. 付加情報自体を取り出せるようにすること
3. 付加情報の内容や解釈の原則を利用者が閲覧できるようにすること
4. 誰がどのように付加したかが分かるようにすること
5. 付加情報は便宜的なものであり，利用者が自己責任で使うものであること
6. 付加情報はできるだけ偏らず，理論に中立なものにすること
7. 特定の付加情報を絶対的なものとは考えないこと

上記の用例は，オリジナルの用例の単語と区別できるよう，通常使われない記号列等を使って付加情報を明示しているので，必要であれば削除して元に戻すこともできる。従って Leech の原則に従っているとはいえるが，一方で，共用データへのアノテーションを想定する場合，明らかな欠点がある。

- 文書化の必要性：どのような情報がどのような形式で付加されているかが研究者個人にしか分からず，その結果データの共有や永続的な利用が困難になってしまう。
- 情報およびその付加方法の汎用性の欠如：特定の研究用でない大量のデータに対し情報を付加する場合，このような情報の付加はむしろ邪魔になる。

#### 4. SGML から構造化標準としての XML へ

これまでみてきたように，複数の研究者が共同で利用する検索システムを構築する場合，言語データの構造を電子的にどう表現するかという問題と，(言語

学的な) アノテーションをどのように付加するか, という問題など, さまざまなレベルの情報をどのようにコーパス中に記述するかという点が重要になってくる。

原テキストに情報を付加すればするほど, コーパスの構造は複雑になってくる。「構造化」とはまさにこのような複雑なコーパスに含まれる情報をコンピュータで適切に処理するために必要な手順であり, 実際の情報の付加方法を「マークアップ」markup<sup>6</sup> と呼ぶ (Hockey 1998: 107—111; Hockey 2000: 24; Renear 2004: 219—220; 図 4 参照)。

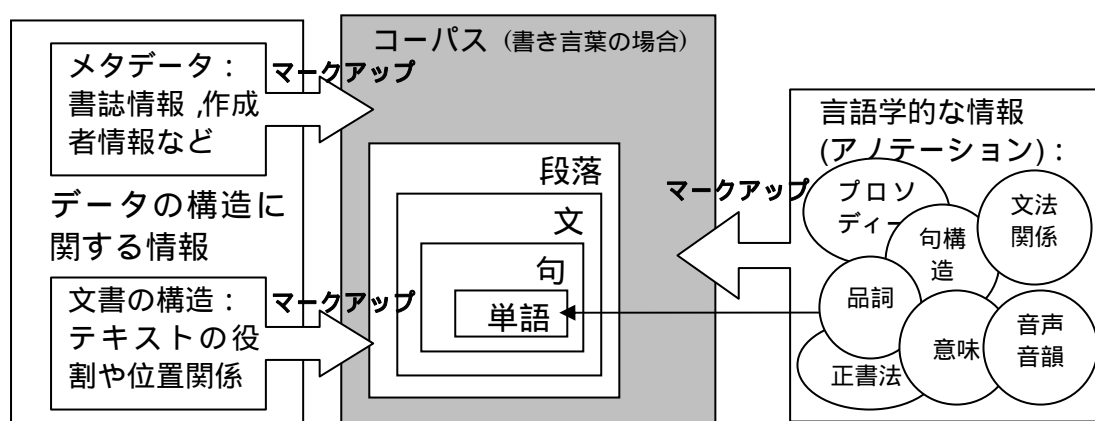


図 4 : さまざまなマークアップ

電子化されたコーパスが作成されはじめた 1970 年代当初から, さまざまなマークアップの方法が提案されてきた (Hockey 1998: 108—115; Hockey 2000:

<sup>6</sup> コーパスの構造化を表す用語には markup のほか, エンコーディング encoding, アノテーション annotation, タグ付け tagging などが使われる。本稿では, 情報の記述方式一般を「マークアップ」と呼び, 「アノテーション」は言語学的な情報の付加を特に指す場合に用いる。Meyer (2002: 81—99) はマークアップとアノテーションを同義に扱い, 「構造的なマークアップ」「品詞のマークアップ」「文法のマークアップ」の 3 種類に分類している。言語学的なアノテーションの種類や標準化の試みについては Leech (1993, 1997) を参照。

24—48; Renear 2004: 219—225)。以下に一例を挙げる：

- COCOA 形式 (Hockey 1998: 108—111)：最も古いデータ構造に関する情報のマークアップ方式。日本語でも、1990年に『源氏物語』コーパスが COCOA 形式で作成され公開されている (近藤 2003: 63—64, 66—67)。
- Brown Corpus のマークアップ方式：行頭にコーパスのファイル番号と行番号を固定長データとして置く。原本の情報を忠実に再現。日本語でもいくつかの古典語コーパスが採用 (近藤 2003: 65)
- ICE (International Corpus of English, Meyer 2002: 82—84) の構造タグ：会話などの間言語的情報の記述方法を規定
- KOKIN ルール (国文学研究資料館, 近藤 2003: 67—68)：岩波古典文学大系・旧版の本文コーパスとして本文をそのまま電子化することを目的に設計

しかし、これらのマークアップ方式は、その殆どが独立した動機で独自に設計されたため、相互変換ができないという大きな問題がある (Hockey 1998: 108; 近藤 2003: 67—68)。さらに、これらのマークアップを活用して高度な検索処理をおこなうためには、そのマークアップに対応したソフトウェアを利用する必要があり、汎用性の点でも問題がある。

このような試行錯誤の中で、文書の交換を目的に 1970 年前後から開発が進められ、1986 年に国際規格 ISO8879 となった SGML (Standard Generalized Markup Language) は電子データの構造化の手法の標準化に画期的な道を開いた。1987 年に Text Encoding Initiative (TEI) という国際組織が設立され、SGML に従いあらゆる言語データを電子化するためのマークアップのガイドライン *Guidelines for Electronic Text Encoding and Interchange* (Sperberg-McQueen & Burnard 1995, 1999) が提案されると、以後 TEI に従

い研究用に設計されたコーパスが数多く作成されるようになった。

TEI はコーパスの構造記述やアノテーションの方法に関する詳細な規定も含んでいる。イギリス英語の優れたコーパスとして知られる British National Corpus (Aston & Burnard 1998) は、1 億語という大規模なデータに話し言葉を含むさまざまなジャンルのテキストをバランスよく収録し、さらにテキストの各単語に品詞情報を付加した画期的なコーパスであるが、一方で各コーパスファイルは TEI に準拠し、テキストの構造、コーパスに関するメタデータ、品詞情報が SGML のマークアップにより一貫した形で記述されている。また、高度な言語分析を可能にする専用の分析ツール SARA, BNCweb はともに SGML を解釈して分析結果を出力する非常に高度なアプリケーションである。以下は、サンプルとして配布されている BNC Sampler のデータの一部である (cf. 図 6)。

```
<head type=MAIN>
<s n=0001 p=Y><w DA>Former <w JJ>Bolivian <w NN1>min
ister <w II>in <w NP1>US <w NN1>court< c YSTP>. </s>
</head>
<head type=BYLINE>
<s n=0002 p=Y><w II>By <w NP1>Mark <w NP1>Tran <w II>
in <w NP1>Washington </s>
</head>
<p>
<s n=0003 p=Y><w AT>THE <w NN2>wheels <w IO>of <w NN1
>justice <w VVD>began <w VVG>turning <w RT>yesterday
<w IF>for <w AT>the <w DA>former <w JJ>Bolivian <w N
N1>Interior <w NN1>Minister< c YCOM>, <w NNB>Mr <w NP
1>Luis <w NP1>Arce <w NP1>Gomez< c YCOM>, <w CS>when
<w PPHS1>he <w VVD>appeared <w II>before <w AT1>a <w
NP1>Miami <w NN1>magistrate <w II>following <w APPGE
>his <w NN1>arrest <w CC>and <w NN1>deportation <w I
I>from <w NP1>Bolivia< c YSTP>. </s>
</p>
```

ヘッダ  
情報

パラグラフ  
(p)

文 (s)

図 5 : BNCA9V.sgm (BNC Sampler Release 1.1 所収)の一部

SGML では、「タグ」と呼ばれる < と > で囲んだ情報でテキストの特定の部分に特定の情報を記述することができる。例えばヘッダ情報（2 種類）は <head>と</head>の間に記述されており，2 種類のヘッダはそれぞれ type が MAIN（タイトル）と BYLINE（著者）であることが，開始タグの内容から分かる，というわけである。原テキストには単語ごとに品詞情報が付加されており，単語の前に<w>というタグで表記される（例えば<w NP1>Bolivia の NP1 は Bolivia が singular proper noun であることを表している）。

テキストのマークアップという形で情報を記述する以外のもう 1 つの重要な SGML の特徴は，文書構造を DTD (Document Type Definition) という別のファイルであらかじめ定義しておくことである。SGML 対応のソフトウェアであれば，DTD を読み，データ構造が正しいかどうかを自動的に判断することができる。例えば，上記<w>タグに終了タグ</w>が必要ない，ということは DTD で定義されている（図 6 参照；定義中の - o のうち，後ろの o が終了タグの設定 (omittable) である）。

```
<!ELEMENT w          - o  (#PCDATA)          >
```

図 6 : BNC サンプラー DTD の一部 (w の定義)

さて，British National Corpus のような幸運な例外はあるものの，SGML には以下のような欠点があり，遅々として普及しなかった（豊島 1992; 村田 1998; 中山他 2001: 23）<sup>7</sup>：

<sup>7</sup> 特に日本において TEI の普及が進んでいない点については，2006 年 5 月 17 日に開催される京都大学 21 世紀 COE プログラム「東アジア世界の人文情報学研究教育拠点」の国際セミナー「TEI Day in Kyoto 2006」で数件の発表が予定

- 仕様の複雑さ
- ツールの不足
- DTD 作成の難しさ
- 論理構造の作成の難しさ
- 既存の電子文書からの変換の難しさ
- 処理パフォーマンスの悪さ (DTD の照合や省略タグの復元をおこなうため)
- 文字コードの処理が不徹底

このような現状を踏まえ，SGML を改良した XML (Extensible Markup Language) 1.0 (Yergeau ほか 1997; 宮下 2005) が 1997 年 12 月にインターネット関連の国際的な標準化団体 World Wide Web Consortium (W3C) により勧告された。<sup>8</sup> XML は SGML の欠点を解消するため，以下のような特徴をもっている：

- DTD に頼らない構造化規則 (well-formedness) : DTD がなくとも表層構造を XML 解析プログラムがチェックできる。
- Unicode への対応 (XML 1.1 で最新の Unicode にも対応)
- 処理効率の向上のための仕様の簡素化 : 例えば終了タグの省略の禁止，最上位要素の義務的配置 (XML データは最初と最後を 1 種類のタグのセットで囲まなければならない) など。

その結果，例えば BNC Sampler のアップデートバージョンである BNC Baby

---

されており，実質的な議論がなされることが期待される。

<sup>8</sup> 最新版は 2004 年 2 月勧告の 1.1 だが，1.0 も引き続き利用できる。

のXMLデータでは,SGMLのDTDに従って許されていた<w>タグに対応する終了タグ</w>が逐一挿入され,開始タグ内部に記述される属性の書式も大幅に整理されるに至っている。以下に図5に対応するXMLデータの箇所を示す:

```
<head type="main">
<s n="0001" p="Y"><w type="DA">Former </w><w type="JJ">B
olivian </w><w type="NN1">minister </w><w type="II">in <
/w><w type="NP1">US </w><w type="NN1">court</w><c type="
YSTP">. </c></s>
</head>
<head type="byline">
<s n="0002" p="Y"><w type="II">By </w><w type="NP1">Mark
</w><w type="NP1">Tran </w><w type="II">in </w><w type=
"NP1">Washington </w></s>
</head>
<p>
<s n="0003" p="Y"><w type="AT">THE </w><w type="NN2">whe
els </w><w type="IO">of </w><w type="NN1">justice </w><w
type="VVD">began </w><w type="VVG">turning </w><w type=
"RT">yesterday </w><w type="IF">for </w><w type="AT">the
</w><w type="DA">former </w><w type="JJ">Bolivian </w><
w type="NN1">Interior </w><w type="NN1">Minister</w><c t
ype="YCOM">, </c><w type="NNB">Mr </w><w type="NP1">Luis
</w><w type="NP1">Arce </w><w type="NP1">Gomez</w><c ty
pe="YCOM">, </c><w type="CS">when </w><w type="PPHS1">he
</w><w type="VVD">appeared </w><w type="II">before </w>
<w type="AT1">a </w><w type="NP1">Miami </w><w type="NN1
">magistrate </w><w type="II">following </w><w type="APP
GE">his </w><w type="NN1">arrest </w><w type="CC">and </
w><w type="NN1">deportation </w><w type="II">from </w><w
type="NP1">Bolivia</w><c type="YSTP">. </c></s>
</p>
```

図7: A9V.xml (BNC Baby v2 所収, 図5のBNCA9V.sgmに対応)の一部<sup>9</sup>

<sup>9</sup> 比較の目的のため,オリジナルのXMLデータにはない改行を適宜挿入している。

## 5. XML ベースの大規模言語資料の公開

XML はインターネットの普及にも後押しされ、技術開発が盛んに行われ、その結果 XML 技術を用いる利用環境やツールが SGML よりもはるかに広く、また急速に普及することになった。例えば、Windows に標準で付属する WWW ブラウザ Internet Explorer は XML で作成された文書を解析するツールを内部に持っており、特別な設定なしで XML 文書を開くことができる。さらに、次節でみるように、W3C を中心になり XML を利用するための技術が着々と策定され、実用化されてきている。

それに後押しされるように、言語コーパスを XML でマークアップし、言語研究者向けに提供する取り組みが、ここ数年さまざまな言語でおこなわれている。英語に関しては XML ベースの BNC Baby のほか、Corpus Encoding Standard for XML (XCES)<sup>10</sup> を用いた新たな英語コーパス American National Corpus<sup>11</sup> の公開が始まっている。また XML をベースとする TEI の新バージョン (TEI P5) の策定も進んでいる。<sup>12</sup>

日本語に関しても、以下のような大規模な XML ベースの言語資料が続々と公開されており、言語研究に XML を利用するための土壌として大いに期待される

---

<sup>10</sup> XCES ホームページ URL: <http://www.xml-ces.org/> XCES は Corpus Encoding Standard (CES, URL: <http://www.cs.vassar.edu/CES/>) を XML に対応させた規格であり、CES 自体は TEI に準拠して作成されている。

<sup>11</sup> ANC ホームページの URL は: <http://americannationalcorpus.org/> ANC は現在 Second Release が公開されており、Linguistic Data Consortium から入手できる。

<sup>12</sup> TEI の次期バージョン P5 は XML に完全準拠する予定であり、それ以前との互換性が保たれなくなる予定である。TEI P5 の更新状況は “TEI: the P5 Release” (URL: <http://www.tei-c.org/P5/>) を参照のこと。

ところである。

- 『日本語話し言葉コーパス』の XML 文書（国立国語研究所，情報通信研究機構 2004, 2005<sup>2</sup>; 前川 2004）<sup>13</sup>
- 『太陽コーパス』（国立国語研究所 2005; 田中 2005: 14—44）

なお，XML 対応の HTML である XHTML で作成されている「青空文庫」のデータは，HTML の構造こそ XML の規則に則って作成しているとはいえ，文書構造を適切にマークアップしているものとは言えない。

## 6. XML データを検索する規格としての XPath

XML 形式のコーパスに対応した汎用の検索ツールを開発する際の現実的な問題として，複雑かつ多様な構造をもつ XML データをどのように扱うか，という問題がある。XML を利用した検索ツールが利用者のニーズを反映していないことについて，豊島（2001:9）は「XML のマークアップ校正を繰り返したテキストに対して，tag を剥ぎ取った形での頒布を求められるのは，しばしばある事である。これは，XML データの検索技術等，（今の処）誰も信用していないからであろう」と述べているが，確かに，XML データを表に出さず，ブラックボックス的に検索をおこなうツールを利用する限り，両者にとっては特定のツールを介してしか利用できない旧来のマークアップ（§3 参照）となんら変わらない。

---

<sup>13</sup> 公開当初のセットに含まれていた XML 文書には一部データの欠落があり，2004 年度に「修正 XML 文書」が購入者に配布されている。

本稿では、むしろ、XMLのマークアップを検索に直接利用できる汎用の検索ツールを構想することで、このような問題の解決の糸口があるのではないかと考える。このように考える理由は3つある。

- XMLを検索する技術がここ数年で普及が急速に進み、多くのソフトやツールに実装されはじめていること。特にXML文書を抽出するための規格XPath (Clark et. al, 1999; Simpson 2002) は既に実用段階にある安定した技術といってよく、他の多くのXML関連技術にも利用されている。
- XML, XPathともに標準規格であり、さまざまな分野で利用され、技術指導ができる専門家を探しやすいこと。
- XML処理用の特別なソフトなしでも稼動するユーザー環境が整ってきたこと。本稿で紹介するプロトタイプのツールは開発環境としてMicrosoft .NET Framework 1.0 を利用しており、実行環境がインストールされていれば安定して動作する。

本稿では、XPathを使った検索システムのプロトタイプとして、searchXMLというツールを開発した(図8)。このツールは、XMLファイルを読み込み、XPathを使ってXMLのなかの特定の構造のみを取り出し、さらにXPathの結果に基づきキーワード検索をおこなうことができる。<sup>14</sup>

---

<sup>14</sup> 本稿で言及するツールUniGrepおよびsearchXMLのアップデート情報などの詳細は言語研究センター(LinC)ホームページ:

<http://www.FL.reitaku-u.ac.jp/LINC/> ないし筆者のホームページ:

<http://www.FL.reitaku-u.ac.jp/~schiba/tools/>

で公開している(searchXMLについては未修正のバグが若干残っている)。なお本文中にも触れたように、実行環境としてMicrosoft .NET Framework 1.0をインストールすることが必要。詳細はツールとともに公開する解説ページおよびMicrosoft社の.NET Frameworkのページを参照していただきたい: URL:

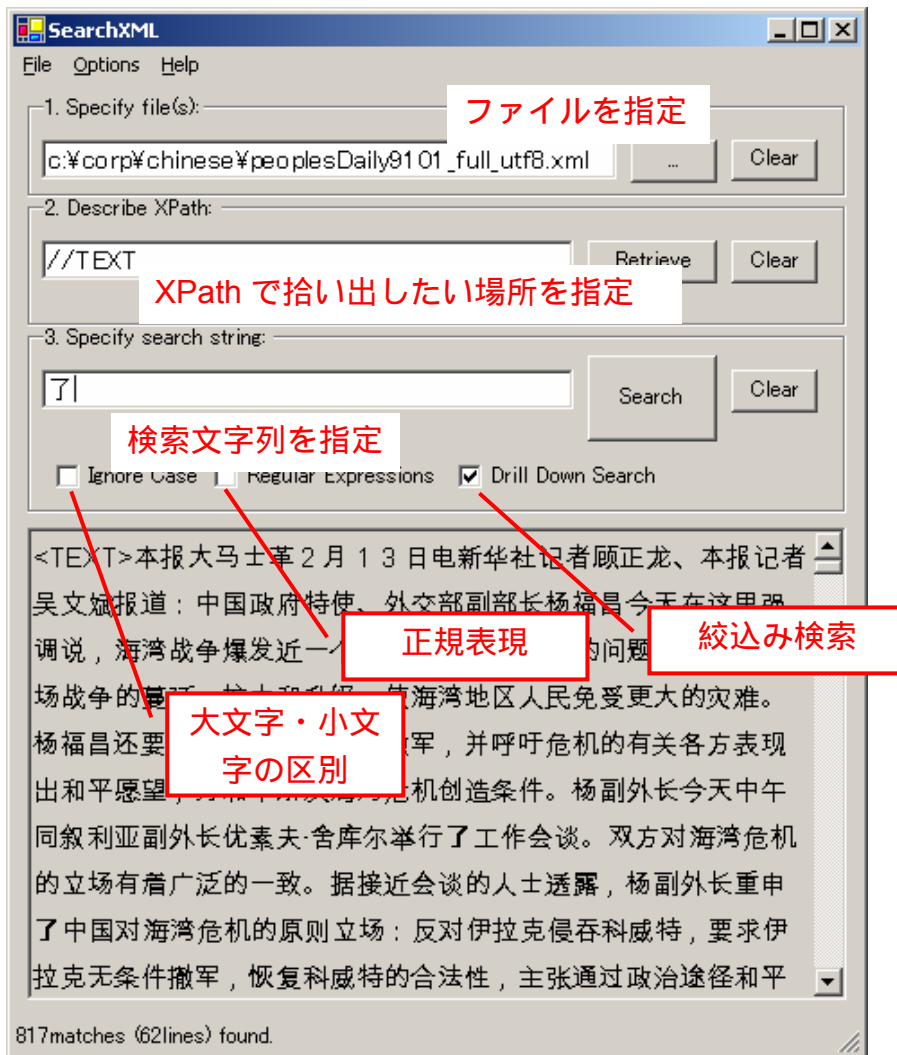


図 8 : searchXML

なお、現段階では未実装だが、Ide (2000a, 2000b) で提案されているように、原文を参照する機能の実現には、XPath の拡張規格ともいえる XPointer (Grosso ほか 2003, XML の一部を参照するための関連規格であり、2003 年 3 月に W3C 勧告となった) を利用することが考えられる。また、XML 文書のより柔軟な検索には、XML の抽出に XPath を援用し、データベース的な操作を可能にする XQuery (2005 年 11 月現在 W3C の勧告候補) などを利用していく

<http://www.microsoft.com/japan/msdn/netframework/>

ことが考えられるだろう。

## 7. XML で構造化されたデータから何が分かるか

XPath を利用することで、アノテーションとして記述した言語学的な情報を含め、厳密な条件を指定して XML の構造を検索することが可能である。さらに、表 1 のように、コーパスデータの特定の位置にあるテキストを条件つきで抜き出す「サンプル・コーパス」的な利用ができる (cf. 後藤 2003: 8) :

検索内容	XPath のロケーションパス (タグは hypothetical)
見出しの文体的、文法的特徴	<code>//h1 //h2 //h3 //h4 //h5 //h6</code>
小説等の最初の数パラグラフと最後の数パラグラフの語彙、文体、文構造、情報構造の比較	(最初の 2 パラグラフ) <code>//body//p[ position( ) &lt; = 2 ]</code> (最後の 2 パラグラフ) <code>//body//p[ position ( ) &gt; = last ( ) - 1 ]</code>
新聞記事などで最初に現れる代名詞を含む文	(文のリスト) <code>//article/p/s/w[ @POS = '代名詞' ][1]</code> (マッチした文の数) <code>count (//article/p/s/w[ @POS = '代名詞' ][1])</code>
名詞の属格形を含む名詞句	<code>//NP[NP/N[ @POS = 'genitive' ]]</code>

表 1 : XPath による XML 文書の検索例

電子データとして記述されていない情報は、コンピュータで検索することは原則として不可能である。しかし、XPath の検索機能を活用することで、このような詳細な構造を指定することができる。このような手法は XML のような明確な構造化手段をもたないデータ形式では困難であり、また今後検索速度が向上すれば大規模コーパスを使った分析でも威力を発揮すると考えられる。

## 8. まとめと展望：言語研究者の新しい知識基盤としての XML

本稿では、汎用コーパス検索システムの設計において言語データの構造化に真剣に取り組む必要があることを論じ、マークアップの標準規格 XML とその検索技術を言語研究に積極的に活用する方向性を示した。

近年の英語を中心とするコーパス言語学の研究成果は目を見張るものがあり、分析手法や理論的な立場を一般言語学的に位置づけようとする試みもなされている (Francis 1993; Hunston & Francis 2000; Tognini-Bonelli 2001)。またコーパス分析ツールについても「言語を問わない」汎用のコロケーション分析ツールが提案されているが (Sinclair ほか 1998)、語や分節の定義がむずかしく、かつコロケーションについての実証研究に乏しい日本語 (cf. 大曾 & 滝沢 2003) など英語以外の言語においては、まず汎用の用例検索ツールを用いた分析手法を開拓することには大いに意義があると思われる。

XML および関連技術の普及はめざましく、特に XML の検索をおこなう XPath はほぼ実用レベルにあると言える。このような技術を積極的に言語研究に活用することで、XML の構造を隠蔽した「ブラックボックス」的な用例検索

を脱却し、個々の研究者が自分の関心に合わせ構造化された言語データを柔軟に検索できるようになる。今後、コーパスデータの XML 化をすすめるとともに、実際に XPath を利用して得られたデータを分析しながら新たな知見を探ること、汎用検索システムの機能の選定とプロトタイピングにつなげていくことができると思われる。

XML とその関連技術 XPath の基礎知識が言語研究の新しいよりどころとすることで、研究者はより高度なデータ収集と分析への足がかりをつかむことができる。これらは国際標準規格として多方面で利用されており、XML を使った研究実績は将来的に言語学以外の分野と連携した新しい研究・応用分野の開拓につながることを期待される。

## 謝辞

本稿は、文部科学省科学研究費補助金 基盤研究 (A) (2)「音声記号等で表記された言語資料のマークアップとコンピュータ処理」(課題番号：15202008；研究代表者：松村一登(東京大学))の研究成果の一部である。

執筆にあたっては、麗澤大学言語研究センター第 9 回研究セミナーにおける研究発表「構造化された言語データからわかること」(2004 年 9 月 30 日、於麗澤大学)、および文部科学省科学研究費補助金基盤研究 (B) (2)「大規模コーパスを利用した英語の構文に関する総合的研究および構文の共起に関する研究」(課題番号：17320073；研究代表者：滝沢直弘(名古屋大学))により開催されたワークショップ「XML をもちいた言語研究の可能性」における講演「構造化され

た言語データが言語研究にもたらすもの」(2006年1月14日, 於名古屋大学大学院国際開発研究科) の内容に加筆訂正をおこなった。

## 参考文献 (邦文文献, 欧文文献の順に挙げる)

大曾美恵子, 滝沢直宏 2003 「コーパスによる日本語教育の研究—コロケーション及びその誤用を中心に—」『日本語学』22/5 (2003年4月臨時増刊号「コーパス言語学」) pp. 234—244.

後藤斉 2003 「言語理論と言語資料: コーパスとコーパス以外のデータ」『日本語学』22/5 (2003年4月臨時増刊号「コーパス言語学」) pp. 6—15.

近藤泰弘 2003 「古典語のコーパス」『日本語学』22/5 (2003年4月臨時増刊号「コーパス言語学」) pp. 62—81.

齊藤俊雄, 中村純作, 赤野一郎編 2005<sup>2</sup> 『英語コーパス言語学: 基礎と実践』研究社.

滝沢直宏 2003 「日本語電子化テキストからのコロケーションの抽出」『日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究』(平成13~15年度科学研究費補助金 基盤研究(B)(2) 研究課題番号13480069) 報告論文集』 pp.27—40.

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」国立国語研究所編『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—』博文館新社. pp. 1—48.

田野村忠温 2000 「用例に基づく日本語研究—コーパス言語学—」『日本語学』19/5 (2000年4月臨時増刊号「新・文法用語入門」) pp. 192—201.

豊島正之 1996 「TEI から見た SGML のはなし」『情報処理語学文学研究会会

- 報累積版』情報処理語学文学研究会 (JALLC). pp. 94—110.
- 豊島正之 2001 「XML の骨抜き利用法：アジア・アフリカ言語文化研究所データベースの例」ハンドアウト. 文部科学省科学研究費補助金 特定領域研究「古典学の再構築」情報処理班主宰研究集会「XML pro/con : XML で書く文献学的データ」(於九州大学文学部, 2001年10月27日)
- 中村純作 2004 「コーパス言語学を概観する」『英語青年』2004/2: 650—653.
- 中山幹敏, 奥井康弘編著 (2001) 『改訂版標準 XML 完全解説』全2巻. 技術評論社.
- 前川喜久雄 2004 「『日本語話し言葉コーパス』の概要」『日本語科学』15: 111—133.
- 宮下徹雄 2006<sup>2</sup> 『改訂 XML 入門』SCC 出版局.
- 村田真 1998 『XML 入門』日本経済新聞社.
- Aston, Guy & Lou Burnard 1998 *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- Barnbrook, Geoff 1996 *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.
- Clark, James Clark, and Steve DeRose 1999 *XML Path Language (XPath) Version 1.0*. W3C Recommendation, 16. November, 1999.  
URL: <http://www.w3.org/TR/xpath/>
- Dunlop, Dominic 1995 “Practical considerations in the use of TEI headers in a large corpus,” *Computers and the Humanities* 29: 85—98. (Also in Ide & Véronis (eds.) pp. 85—98.)
- Francis, Gill 1993 “A corpus-driven approach to grammar: principles, methods and examples,” in Baker, Mona, Gill Francis & Elena

- Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins. pp. 137—56.
- Grosso, Paul, Eve Maler, Jonathan Marsh, Norman Walsh 2003 *XPointer Framework*. W3C Recommendation, 25. March, 2003.  
URL: <http://www.w3.org/TR/xptr-framework/>
- Hockey, Susan 1998 “Textual databases,” in Lawler, John & Helen Aristar Dry eds. *Using Computers in Linguistics: A Practical Guide*. London: Routledge. pp. 100—137.
- Hockey, Susan 2000 *Electronic Texts in the Humanities*. Oxford: Oxford University Press.
- Hunston, Susan & Gill Francis 2000 *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.
- Ide, Nancy 2000a “The XML framework and its implications for corpus access and use.” *Proceedings of Data Architectures and Software Support for Large Corpora, Athens, Greece, 30 May, 2000*. Paris: European Language Resources Association, pp. 28-32.
- Ide, Nancy 2000b “Searching annotated language resources in XML: a statement of the problem.” Paper read at the ACM SIGIR 2000 Workshop On XML and Information Retrieval, Athens, Greece, 28 July, 2000.
- Ide, Nancy 2004 “Preparation and analysis of linguistic corpora,” in Schreibman, Siemens & Unsworth (eds.) pp. 289—305.
- Ide, Nancy & Jean Véronis (eds.) 1995 *Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic.

- Leech, Geoffrey 1993 “Corpus Annotation Schemes,” *Literary and Linguistic Computing* 8: 275—281.
- Leech, Geoffrey 1997 “Introducing corpus annotation.” In Garside, Roger, Geoffrey Leech, and Anthony McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, pp. 1—18.
- McEnery, Tony & Andrew Wilson 2001<sup>2</sup> *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Meyer, Charles F. 2002 *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Renear, Allen H. 2004 “Text encoding,” in Schreibman, Siemens & Unsworth (eds.) pp. 218—239.
- Schreibman, Susan, Ray Siemens & John Unsworth (eds) 2004 *A Companion to Digital Humanities*. Oxford: Blackwell.
- Simpson, John E. 2002 *XPath and XPointer*. Sebastopol, CA: O’Reilly.
- Sinclair, John, Oliver Mason, Jackie Ball & Geoff Barnbrook 1999 “Language independent statistical software for corpus exploration,” *Computers and the Humanities* 31: 229—255.
- Sperberg-McQueen, C. M. and Lou Burnard 1999 *TEI P3: Guidelines for Electronic Text Encoding and Interchange*. Bergen: The TEI Consortium. (TEIの最新バージョンは2004年発行のTEI P4であり、XMLにも対応している。URL: <http://www.tei-c.org/P4X/> なお、XMLに完全準拠する予定のTEI P5の更新状況は<http://www.tei-c.org/P5/> を参照されたい。)
- Tognini-Bonelli, Elena 2001 *Corpus Linguistics at Work*. Amsterdam: John

Benamins.

Yergeau, François, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve  
Maler 2004<sup>3</sup> *Extensible Markup Language (XML) 1.0*. W3C  
Recommendation, 4. February 2004. URL: URL:  
<http://www.w3.org/TR/REC-xml/>