

# アノテートされた大規模コーパスを用いた 分析ツールの現状と今後の方向性<sup>†</sup>

千葉庄寿  
麗澤大学外国語学部  
schiba@reitaku-u.ac.jp

## 1. アノテーションと大規模コーパス

コンピュータ処理可能なコーパスは、テキストデータをその基軸に据えている。言語の線状性は単純なビットの連続として構成されるテキストデータにとって有利な特徴ではあるが、蓄積・処理すべき情報としてテキスト以外の要素、例えば文書構造やメタ言語情報、発音を含むメディア情報や文法情報などを付加しようとする時、データ構造に手を加える必要が出てくる。バイナリデータとしてテキストデータを他のデータに埋め込むのでなければ、テキストデータの中に「アノテーション annotation」として情報を埋め込む、いわゆる「マークアップ markup」をおこない、付加された情報を元の情報と何らかの形で区別できるようにする (千葉 2006)。

テキストデータ中にアノテーションをおこなう標準的な手法として、応用アプリケーションの整備状況の点でも高い完成度をもつ技術標準が XML (Extensible Markup Language) である (Yergeau *et al.* 2004<sup>3</sup>)。XML を用いることで、一方では大規模なコーパスを構築する場合のように、厳密なデータ構造を検証ソフトウェアによって随時維持しつつ非常に複雑な構造をもつデータを作成することが可能であり、また一方では、データ構造の定義を省略し、XML の最低限の規則に則った簡易なデータを自由に作成することもできるようになる。いずれの場合においても、XML データの構築支援や XML の構造チェック、検索や変換といったデータ処理の各場面で XML の応用技術を活用することができる。

このような状況にあって、XML をデータ構造として採用するコーパスは増えてきている。しかし、XML に準拠したコーパスを利用するための環境は、決して充分整備されてきているとは言えない。本稿では、主として複雑にアノテートされたコーパスの検索と分析という観点から、XML 形式で構築されたコーパスを利用する環境の現状とツール開発上の課題を論じる。

XML 形式のコーパスの利用のうえで問題となりうる点として、まず指摘したいのが、XML データの複雑性である。アノテートすべき情報が増えれば増える

ほど必然的にコーパスの構造は複雑になる。単純なテキストデータ、いわゆる plain text の状態で grep 機能のあるツールを使って簡単に行えるテキスト検索は、品詞情報など次元の異なる情報が付加されると同時に困難になり、ユーザにデータの置換などの高度な情報処理の知識を強いることになる。複雑な構造をもつ高度に洗練された XML コーパスであれば、なおさらである。

もちろん、このような問題は XML コーパスを利用するためのクライアントソフトウェアが充実してくることで将来的に解決できる可能性はある。また、XML 形式で配布されている多くのコーパスは XML データから plain text を抽出する方法について何らかの情報を提供していることも確かであり、XML データを直接扱うことを回避することで、旧来の手法による分析が可能になることも確かである。しかし、ソフトウェアのブラックボックス化という別の問題は常につきまとうので、ユーザは状況によってその複雑な XML データの解析を強いられる可能性から逃れることはできないだろう。

第 2 に、XML 形式のコーパスを構築する際、厳密さと一貫性、データ標準への準拠という異なる基準でそのコーパスのデータ構造を決定することになることに注意したい。例えば、現在公開されている日本語の 2 種類の大規模なコーパスと BNC を比較すると表のようになる

表：3 種類のコーパスのデータ構造の決定要因

	CSJ (日本語話し言葉コーパス)	BCCWJ (日本語書き言葉均衡コーパス)	BNC ( <i>British National Corpus XML Edition</i> )
一貫性	○	△ ジャンル毎に異なるデータ構造	○
厳密さ	○ DTD に従う	○ DTD に従う	○ DTD に従う
標準準拠	× 独自構造を採用	× 独自構造を採用	○ TEI を採用

表 1 から分かるように、厳密さの基準はどのコーパスでも維持され、この点が XML ベースのコーパスの大きな特徴となっている。一貫性の基準とデータ構造の標準規格への準拠は当該コーパスのポリシーにより異なる。また、現在のところ、国立国語研究所が中心になって構築されている 2 種類の日本語コーパスはデータ構造の標準化には関心が薄い。

## 2. XML コーパス分析用ツール

XML でアノテートされたコーパスの検索ツールとして、本稿では以下の 2 つを取り上げる。

- 全文検索システム「ひまわり」...国立国語研究所 (山口昌也氏開発)

URL: <http://www.kokken.go.jp/lrc/index.php?全文検索システム『ひまわり』>

- Xaira (XML Aware Indexing and Retrieval Architecture) ... Research Technologies Service, Oxford University Computing Services, Oxford University

URL: <http://www.oucs.ox.ac.uk/rts/xaira/>

このうち「ひまわり」は国立国語研究所が構築・公開する各種コーパスの検索用に開発されたもので、簡易な KWIC 検索と本文の参照が可能である。バージョン 1.3 (2009 年 2 月時点で β 版) では登録するコーパスの設定によって JAVA の標準機能に従った正規表現の利用も可能になっている。

以下、本章では、Xaira について概説し、現在の開発状況と共にその方向性と現状分析をおこなう。また、次章では、XML 形式でアノテートされたコーパスの本格的な利用ために必要な課題についてまとめ、現状を踏まえた提案をおこなう。

## 2. 1 Xaira とは

Xaira [ˈseərə] は XML 対応版の BNC (British National Corpus XML Edition) の検索ツールとして 2004 年 8 月に開発が着手された。BNC を古いデータ記述形式である SGML から XML に更新することに合わせ、BNC 検索性アプリケーション SARA (<http://www.natcorp.ox.ac.uk/tools/sara/>) を XML 対応に書き直すことになったわけである。

しかし、Xaira は開発当初から汎用コーパス検索ツールとしての性格づけをもって設計されていた。その主な特徴は以下のようなものである。

- Xaira object model (Xairo) によるコーパスへの一貫したアクセス方法の確立と Xaira Indexer による大規模な XML データへの高速なアクセスの実現。
- 多様な形式の XML データを利用可能 (BNC が準拠する TEI のデータ構造でなくともよい)。plain text 形式のデータも取り込み可能であり、Xaira の変換ツールが簡易 XML 文書に変換してくれる。
- Unicode のサポート (英語以外のコーパスの処理も可能)。
- 設計上は OS の縛りなし (ただし、現在公開されているクライアントソフトウェアは Windows 版のみ)。
- オープンソースのソフトウェアとして sourceforge 上で開発し、ソースを公開 (開発用サイト URL: <http://xaira.sourceforge.net/>)。

ソフトウェアとしての Xaira のパッケージは以下のソフトウェアを含む。

- サーバ Xaira Server (Xairo および汎用データインターフェース Xairo API を

実装する)

- インデックス作成ツール Xaira Index Toolkit (コーパス登録)
- クライアントソフトウェア Xaira Client (コーパス検索・分析)

Xaira は Xairo にあるように、コーパス検索に特化したデータアクセス方法を設計し、実装しようという思想で開発が進められた。その結果、コーパス検索用の拡張検索言語 XXQ (Dodd 2006, Burnard 2006) などの仕様の策定もある程度進んでいる (2. 3 節参照)。しかし、大変残念なことに Xaira の開発は現在中断中しており、開発プロジェクトの経過報告 (URL: [http://www.oucs.ox.ac.uk/rts/xaira/projectplan.xml.ID=body.1\\_div.3](http://www.oucs.ox.ac.uk/rts/xaira/projectplan.xml.ID=body.1_div.3)) によれば、Project Work Package 3 までは完遂し、その後ボランティアベースで一部課題を継続中ということである。この状況に関しては、特に ANC (American National Corpus) との共同プロジェクトが不調に終わったことがプロジェクトの進捗に影響したことは否めない。ただし、一部のユーザが中心になりツールの整備とバグ修正は続いているということ、上述のオープンソースソフトウェアの強みが活かされているといえるだろう。また、開発者の一人である Lou Burnard 氏によれば、資金的なバックアップがあればいつでも開発は続行可能な体制にあるということである (本科研による 2009 年 7 月の Oxford University Computing Services 訪問時のインタビュー)。

## 2. 2 Xaira への言語データ登録

本節では Xaira の利用例として、現在構築が進んでいる日本語コーパスである『現代日本語書き言葉均衡コーパス』(BCCWJ) を Xaira で利用する例を紹介する。データは現在構築チーム【注 1】内で利用できる形態素解析済みのコアデータ (60 万短単位強) 【注 2】を XML 形式で構築したものである (2008 年 11 月 14 日配布のバージョン)。

1. 前修正：地のテキスト (以下の例では網掛け) を SUW 要素の内容に変換する【注 3】

置換前：<SUW ([^/]+ )/>([<]+)

置換後：<SUW ¥1>¥2</SUW>

```
<SUW orthToken="日本" IForm="ニッポン" lemma="日本" pos="名詞-固有名詞-地名-国"
Form="ニッポン" pronToken="ニッポン" wType="固" start="121" end="123"
morphID="760" />日本
```

↓

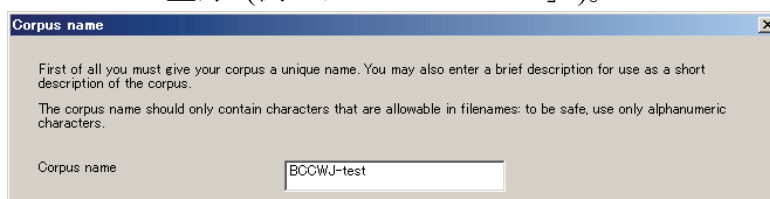
```
<SUW orthToken="日本" IForm="ニッポン" lemma="日本" pos="名詞-固有名詞-地名-国"
Form="ニッポン" pronToken="ニッポン" wType="固" start="121" end="123"
morphID="760">日本</SUW>
```

この作業は EmEditor Professional (Emurasoft によるシェアウエア, URL: <http://jp.emeditor.com/>) など, 正規表現を利用でき, かつファイルの一括置換が可能なツールによっておこなう必要がある。

2. Xaira の Index Toolkit を起動し, 以下の手順で XML コーパスの登録作業をおこなう。

2-1. File → Index Wizard を起動する。

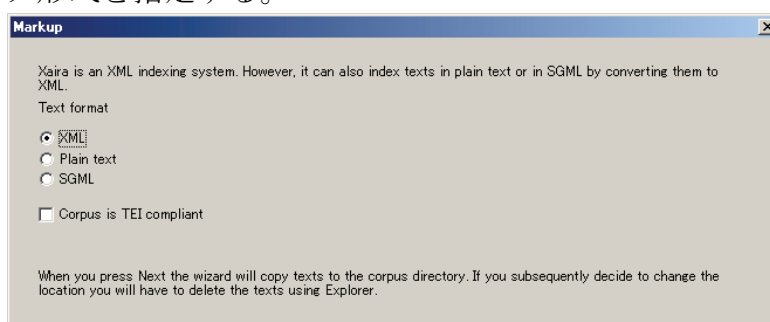
2-2. Corpus Name の登録 (例えば「BCCWJ-test」)。



2-3. コーパスのセットアップ場所 corpus root を指定する。

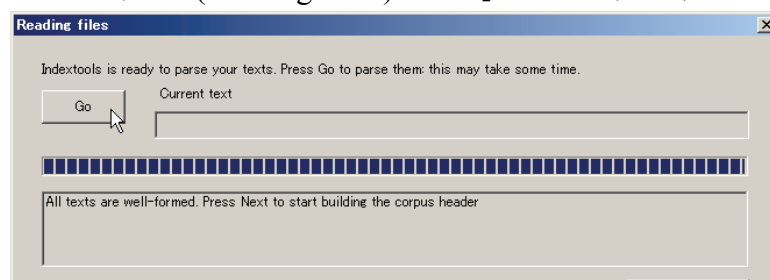
2-4. 登録する XML テキストを指定する。

2-5. データ形式を指定する。



2-6. File Structure (Mode 1), File List を指定する。

2-7. XML のチェック (Reading Files) : 「Go」をクリックすると開始する。



2-8. 言語の指定 (ja = 日本語), ただし, これで全てうまくいくわけではないようだ (cf. Xiao 2006)。

2-9. 【特に重要】 コーパス単位 Text delineation を指定する : mergedSample + Auto-number

2-10. 【重要】 文単位 Unit delineation を指定する : sentence + Auto-number

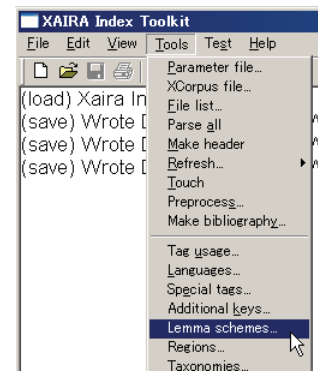
2-11. 【重要】 語トークン Tokenisation を指定する : SUW (短単位)

2-12. 語検索に使うキーが表示される (SUW 要素が持っている属性名の一覧が出る)。

2-13. Bibliography のビルド：今回はメタデータがないので、特に何もしない。

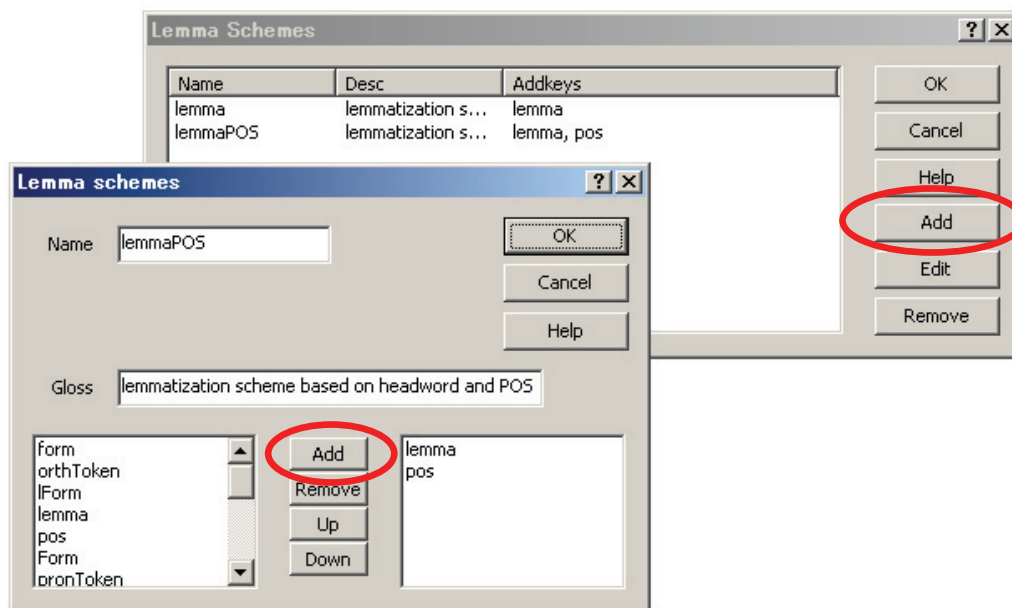
2-14. Indexing：インデックスの作成だが、ここではまだ設定が完了していないので、必ず「キャンセル」をクリックし、Wizard を終了する。

2-15. 【重要】Tools → Lemma schemes を選択 (右図)。

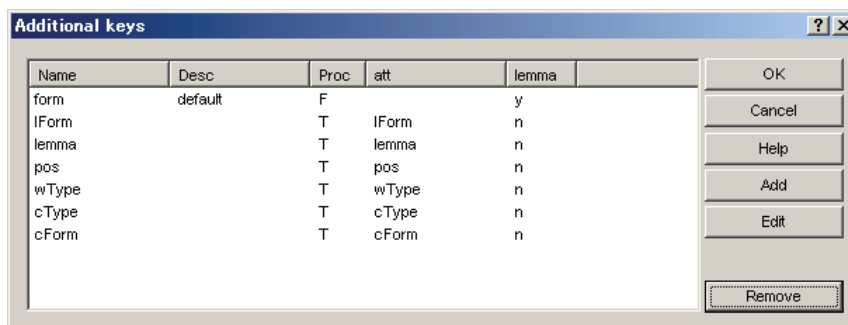


lemma 属性と lemma 属性 + pos 属性等、必要なものをレンマとして登録する。Name に入れる名前は任意。以下は例：

- ◇ Name: lemma → lemma を登録
- ◇ Name: lemmaPOS → lemma と pos をこの順序で登録
- ◇ Name: lemmaCTypeCForm → lemma と cType, cForm をこの順序で登録

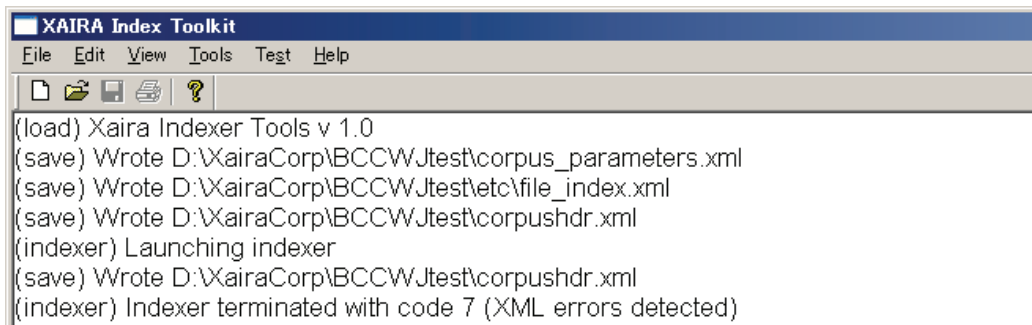


2-16. 語検索に使う属性について、現在登録されているものはかなり数が多く検索結果が見にくくなるので、Tools → Additional Keys で間引く。例：



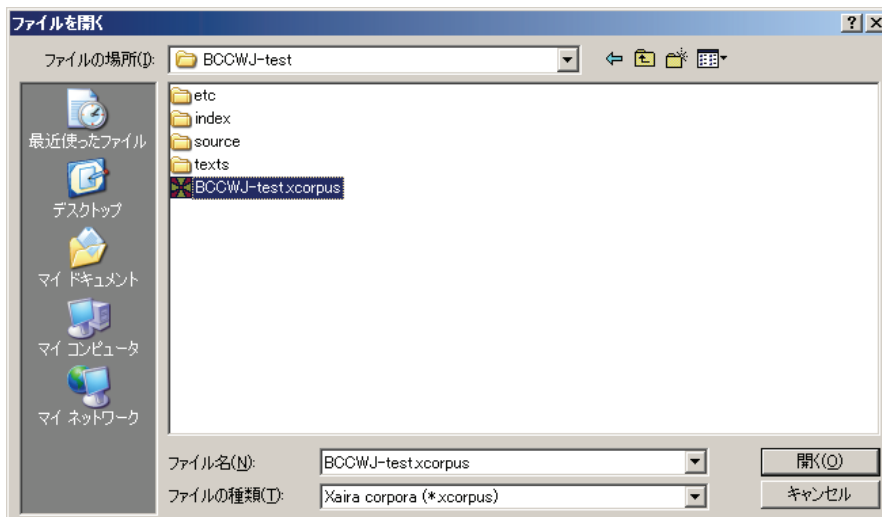
2-17. 【重要】Tools → Indexer → Run でインデックスを作成する。現在の

ところ、以下のようにエラーが出て終了する (ただし、処理自体は終了している)。



```
(load) Xaira Indexer Tools v 1.0
(save) Wrote D:\XairaCorp\BCCWJtest\corpus_parameters.xml
(save) Wrote D:\XairaCorp\BCCWJtest\etc\file_index.xml
(save) Wrote D:\XairaCorp\BCCWJtest\corpushdr.xml
(indexer) Launching indexer
(save) Wrote D:\XairaCorp\BCCWJtest\corpushdr.xml
(indexer) Indexer terminated with code 7 (XML errors detected)
```

3. Xaira client を起動する : File → Open で .xcorpus 設定ファイルを開く。



以上の手順で Xaira 上で登録したコーパスを使用することができるようになる。ただし、Xaira は開発途中のソフトウェアであり、多言語対応を謳っているが、現時点で日本語コーパスを利用する際に数多くのエラー・不具合・要検討点が発生することが分かっている。以下にその主なものを挙げる。

- SUW 要素が余りに多いと sentence 要素を処理できない。以下は etc フォルダに作成される corpuslog.xml に記録されたエラーログの例である：  
<entry type="error" time="Sun Dec 21 06:35:14 2008"><loc sysid="OW6X\_00000.xml" line="6" col="24"/><msg>Task stack too deep, switched off.</msg></entry>
- メニューが一部文字化けする。
- フレーズ検索では、形態素間に半角スペースがないと検索されない。
- 一部の検索が正しく動作しない (上記エラーが解決されることで、修正される点もあると思われる)。Pattern Query の正規表現は正しく動作しない。

- 一部検索 (XML タグ検索など) が極端に遅い。英語コーパスの場合そのような問題はみられない。

上で指摘した問題のいくつかは単純なソフトウェアのバグであり、今後 Xaira の開発者らと連携し修正することで、より多くの言語で安定して利用できるソフトウェアに改善していくことができるであろう。

### 2. 3 Xaira の拡張検索言語 XXQ

Xaira の画期的な設計思想の一つとして、コーパス処理に特化した機能の実装が挙げられる。そのうちのもっとも興味深いものが Xaira 専用の検索言語の開発であり、その拡張検索言語は XXQ (Xaira Query Language) と呼ばれる。現バージョン (Ver. 1.23) の Xaira には残念ながら未実装であるが、以下にその概要を報告する。

その仕様書ドラフト (Dodd 2006, Burnard 2006) から得られる直感的な印象は「Corpus Query Processor (CQP, Christ 1994) がもつ検索言語の機能を XML で記述したもの」というものである。まず、以下に CQP と XXQ の記述を比較する。

#### 記述例の比較(1): 1~4 語を挟んで現れる「とても」と基本形「ない」の連鎖

CQP (例): "とても"•[{1,4}]•[lemma="ない"]

XXQ (例): <element name="s">  
 <gap/>  
 <lemma>とても</lemma>  
 <rep•min="1"•max="4"><any/></rep>  
 <lemma•scheme="ls">ない</lemma>  
 <gap/>  
 </element>

#### 記述例の比較(2): 任意の連用形と基本形「始める」「はじめる」の連鎖

CQP (例): [pos="\*/\*/連用形"]•([lemma="始める"|lemma="はじめる"])

XXQ (例): <element name="s">  
 <gap/>  
 <addkey•key="pos3">連用形</addkey>  
 <or>  
 <lemma•scheme="ls">始める</lemma>  
 <lemma•scheme="ls">はじめる</lemma>  
 </or>  
 <rep•min="0"><gap/></rep>  
 </element>

XXQ はコーパス分析に効果的な検索が可能になるよう、既存の XML 関連仕様との連携も充実している。以下は、XML の応用技術 (XML のノードを記述・検索) である XPath (Clark *et al.* 1999) でデータ構造を指定した XXQ の例である:

#### XXQ の記述例: 段落頭の接続詞「そして」を検索

```
XXQ (例): <element name="s">
  <predicate>
    ancestor::p/self[position()=1]
  </predicate >
  <lemma>そして</lemma>
  <gap•min="0"/>
</element>
```

同様の検索は、標準的な XML 応用技術 XQuery を用いても可能と思われるが、XXQ に比べて非常に複雑になることが想定される (cf. Ide 2001)。

### 3. XML 形式の大規模コーパスの活用のために

Xaira のような試みは、XML をデータ形式とするコーパスの利用にどのようなインパクトをもたらす (もたらした) であろうか。現状把握に基づき、以下では 2 点を指摘したい。

#### 3. 1 共有形式としての XML データとの関わり

第 1 章で述べたとおり、現在 XML コーパスはその構築ポリシーにより、データ構造の標準仕様への準拠についての対応はまちまちである。現在、XML 形式のコーパスの構築に関連する仕様としては以下のようなものが策定され、一般に公開されている。

- TEI P5 (*Guidelines for Electronic Text Encoding and Interchange*. 2007 年 11 月公開) URL: <http://www.tei-c.org/Guidelines/P5/>
- XCES (Corpus Encoding Standard for XML) URL: <http://www.xces.org/>

前者の TEI は BNC XML Edition が、XCES は ANC (American National Corpus, URL: <http://americannationalcorpus.org/>) が準拠するもので、いずれもコーパス構築にとって必要かつ充分と考えられるデータ構造を標準化したものである (XCES は TEI から派生し、コーパスの構造記述に特化した規格と言える)。

Xaira は XML 形式のコーパスはもちろん、さまざまなデータ形式のコーパス

に対応している。しかし、Xaira が最も得意とするのは TEI 形式のコーパスであり、TEI 形式で構築されたデータを Xaira で用いることで、メタデータの検索など、コーパスのアノテーションのもつ可能性を生かした多様な分析が可能になっている。このような Xaira の設計思想の背景には、TEI のような汎用のデータ構造を利用したコーパスが多く世に出、広く普及することが汎用のコーパス利用ツールの普及にとって欠かせない、という考えがあると思われる。

この点で、日本の均衡コーパス構築プロジェクトである BCCWJ のデータ構築の姿勢は標準への準拠とは異なる方向性で動いており、標準化に対する姿勢を今後どのような形で示すか興味深い。現在 BCCWJ のコーパス構築グループは、対象となるデータをなるべく忠実にデータ化することに重点を置いており、その結果元データの事例に合わせてコーパスのデータ構造にしばしば ad hoc ともいえる変更が加えられている。最終的に公開された XML による BCCWJ のデータ構造が、コーパスの構造記述標準を射程に置いた「正規化」にどの程度耐えられるか、を考える時、標準化は単なる方針決定以上の意味と効果をもつことに気づかされる。

一方、上記の懸念とは逆に、データ構造の標準化はそれ自体が問題をはらんでいることも考慮すべきであろう。以下の大矢(2006)の指摘を参照されたい：

TEI の利点として、国際化・地域化が意識されていることを挙げたが、実は、これは TEI が抱える課題でもある。TEI で検討されているテキストタイプには、かなり偏りがあり、多くは欧米文化圏(敢えていえば英語文化圏)<sup>21</sup>のものである。また、テキストタイプとして適応可能なものがある場合にも、タグの使用例<sup>22</sup>が欧米文化(英語文化)のものが採られている。そのため、日本文化を形成する資料をマークアップする場合には、1)TEI タグ規定に従い、英語文化圏に倣うか、2) 追加・拡張を行うかを選択する必要がある。1)の選択は、恐らくない<sup>23</sup>。日本で TEI を使用する場合には、2)を選択せざるを得ない。これは、全くのオリジナルタグ集合の作成ではないが、ほとんどオリジナルタグ集合の作成に近い活動が求められることになる<sup>24</sup>。

この場合、3つの活動が必要となる。ひとつは、TEI の中核タグ集合を使用しながら、日本文化の資料に必要な追加タグ集合を検討することである。さらに、それを TEI へ提案することである<sup>25</sup>。人文科学資料の電子化・マークアップデータ化は、単に従来の研究手法をより迅速に、深く進める為の新形式の資料を作成することだけに意義があるのではない。アノテーションという人文科学研究の根本的行為が、電子化の対象となるという、全く新しい研究領域に携わることである。これには、記述対象物としての資料が文化的に多彩である程、手法の一般化に貢献する。文学・歴史・言語圏に閉じた資料を作成することに価値があるのではない。文化を越えた、科学的研究としての共同作業が必要となる<sup>26</sup>。ここに、TEI をベースにした、個別のオリジナルタグ集合を検討・作成する価値がある。

もうひとつは、TEI ガイドラインの国際化を進めることである。現行 TEI ガイドラインには、ソフトウェア開発で検討されてきたようないわゆる「国際化」に対応するためのガイドラインが含まれていない。文化を形成する資料を対象にしたガイドラインであるから、この論議の困難さは容易に想像がつく。これは、TEI の今後の重要課題である。この論議から実質的な成果を得るには、各文化圏からのマークアップ活動報告が重要になってくる。

(大矢 2006:37-38)

大矢 (2006) が述べるように、コーパス構築のデータ構造の標準規格として TEI の現在の仕様がふさわしいのかどうか、また国際化を進めるためにはどのような国際協力が必要なのか、など、日本を含め各国のコーパス構築・利用に携わる研究者はよく吟味する必要があることは確かである。しかし、それ以上に、現在のコーパス構築のノウハウを使って、世界の言語にみられる多様な言語構造が十分に電子化できるのかどうか、という言語学の観点からコーパスのデータ構造の刷新へより具体的にとりくむ必要はないだろうか。この点で、TEI などの構造化標準に依拠せず独自のデータ構造によるコーパス構築を進める日本語コーパスの研究者たちは決して後進的ではない。必要なのは、標準化と体系化の思想であり、コーパス構築の実務の積み重ねから得られるノウハウを、今こそ国際的に共有することが必要なのではないだろうか。

### 3. 2 アノテーションの検索に有効なツールの開発と既存のツールの活用

Xaira のようなソフトウェアは、XML ベースの大規模コーパスの利用に画期的な可能性を与えられると思われる。しかし、Xaira が今後進むべき道は険しく、人文学系研究者にとって XML コーパスの本格的な研究利用の推進の道を拓く「キラアプリケーション」の登場が渴望される場所である。

例えば、XML データを扱うことのできる検索システム「ひまわり」の機能を使い、XML によるアノテーションつきデータを利用することを検討してみよう。開発者の山口氏の私信では、品詞情報の入ったコーパスを「ひまわり」で利用する場合に XML のタグを利用する提案がなされている：

[前略]『ひまわり』の設計方針では、(理想的には) 利用者がデータの構造を知る必要がない、というものですので、付与情報は、本文ではなく、XML タグとして記述されることを想定しています。したがって、『ひまわり』に取り込む際に XML へ変換することが必要になると思います。(山口昌也氏私信 Feb. 2008)

山口氏の私信にあるように、「ひまわり」自体は簡単な XML の構造を理解するので、ある程度の工夫をすることで、「ひまわり」でも単純なテキスト KWIC 検索以上の機能の実装を実現することができそうである。

実際に、形態素境界を使った検索は、現在の「ひまわり」の機能として実装できる。以下の図は日本語教育支援システム研究会 (CASTEL/J) による「CASTEL/J 2000 CD-ROM」に収録されているコーパスデータ (講談社ブルーバックスなど約 290 万語、以下 BOOKDATA と略す) について、形態素境界を XML データとして登録し、検索に利用したものである。【注 4】検索文字列の前に ^ (サーカムフィックス)、後に \$ という正規表現を指定することで、「と」という

ひらがな一語のみからなる形態素を検索している。このような検索は、形態素境界の情報がなければ簡単には得られない。また、書籍のメタ情報が得られていることにも注目されたい。これらの効果は検索対象となる XML データに情報が記載されていることによる (下の XML データのサンプルを参照)。



図：「ひまわり」で形態素境界情報を含むパターンを検索

```
<コーパス 名前="書籍コーパス">
<記事 No="1" File="J0105" FileName="J01053.txt" Words="47851" タイトル="タ
テ社会の人間関係" 著者="中根千枝" Publisher="講談社" Year="1967">
<m>—</m><m>—</m><m>序論</m><br />
<m> 1 </m><m> — </m><m> — </m><m> 日本 </m><m> の </m><m> 社会
</m><m>を</m><m>新しく</m><m>解明する</m><br />
<m>理論</m><m>と</m><m>現実</m><m>と</m><m>の</m><m>ずれ
</m><m>の</m><m>あり方</m><m>が</m><m>問題</m><br />
<m> </m><m>日本</m><m>の</m><m>社会</m><m>、</m><m>あるいは
</m><m>文化</m><m>を</m><m>論ずる</m><m>場合</m><m>、
</m><m>従来</m><m>とら</m><m>れ</m><m>て</m><m>き</m><m>た
</m><m>方法</m><m>は</m><m>、</m><m>だいたい</m><m>次
</m><m>の</m><m>二つ</m><m>に</m><m>要約できる</m><m>。
</m><br />
```

図：形態素境界情報付きの「ひまわり」データ

さらに、以下のように文脈情報を KWIC で表示するほかに、前後の形態素を抽出して表示することができる。

出現形(前1)	出現形(後1)	出現形(後2)
アパ	いって	も
集団	いって	も
いる	いって	よい
さす	いって	よい
ある	いって	よい
まったく	いわざる	を
せまい	いわざる	を
大きい	いわざる	を
組織	いわざる	を
大きい	いわざる	を
多い	いわれ	たら
か	いわれ	たり
か	いわれて	いる
など		
など		

図：形態素境界情報を使って「と」の前後の形態素を表示

このような機能追加が「ひまわり」の標準的な XML 処理機能を使って実装できることは、このソフトウェアの優れた設計に負うところが大きい。今後、「ひまわり」のこうした機能性を生かした分析が増えることを望みたい。

CSJについても XML 形式のデータの検索に利用できるツールが開発されている (菊池 2008) が、XML とその関連技術 (XSL Transformations, XPath など) に関する知識が必要となり、現時点で広く普及しているとは言い難い。千葉 (2006) で述べたように、これらの技術に親しむことで、コーパスの利用可能性は大きく広がる。この種の「知識」がコーパスを利用する研究者が標準的に習得しておくべき知識として今後普及するかどうかは、学習用を含む親しみやすく効果的な XML コーパス分析ツールが開発され普及するかどうか、XML コーパスの普及が教育的配慮とともに進むかどうか、といったコーパス構築とは異なる分野の発展にかかっているように思える。

## 謝辞と注

† 本稿は以下の口頭発表の内容に加筆訂正を加えたものである。発表時にコメントを頂いた方々に感謝いたします。

千葉庄寿「アノテートされた大規模コーパスを用いた言語分析のモデル：Xaira を例に」(2008年12月21日，特定領域研究「日本語コーパス」日本語教育班研究連絡会議，於早稲田大学)

また、「ひまわり」開発者であり数多くの情報を提供いただいた山口昌也氏，Xaira 開発についてコメントをくださった Lou Burnard 氏にこの場を借りて感謝申し上げます。

- 【1】文部科学省科学研究費補助金 特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21世紀の日本語研究の基盤整備」に属する研究者が内部的に利用できるデータである。特定領域外の研究者はモニター公開データ(約2,800万語)を特定領域研究に利用申請のうえ入手できる(申請方法はホームページを参照：[http://www.kokken.go.jp/kotonoha/ex\\_8.html](http://www.kokken.go.jp/kotonoha/ex_8.html))が，現在のモニター公開データには形態素解析済みのデータは含まれていない。
- 【2】官公庁白書，書籍，新聞が各20万短単位以上，全体で485サンプルからなる。
- 【3】Xaira では地の文の登録に「語」レベルのタグによるマークアップが必要である。当該の前作業はデータの構造上の理由から，SUW要素が空要素になっているために必要となる。この変更により特殊な処理が必要となるコーパスサンプルは今回のXMLデータにはなかった。
- 【4】本データは，2008年2月9日に麗澤大学で開催された言語研究センター第7回ワークショップ「言語情報学」のための実習用データとして，講師である山口昌也氏に整備いただいたものである。ここに記して感謝いたします。

## 参考文献

大矢一志 (2006) 「マークアップの課題を syntax から見た分類と解決のステップ」 *Proceedings of the TEI Day in Kyoto 2006, 17. May, 2006, Kyoto*. Kyoto: Kyoto University. Pp. 29-39. URL: <http://coe21.zinbun.kyoto-u.ac.jp/tei-day/>

小木曾智信 (2008) 「白書・書籍・新聞 XML 版コアデータ」 BCCWJ 領域内公開データマニュアル. 2008年11月14日公開.

菊池英明 (2008) 「『日本語話し言葉コーパス』のXML文書」『日本語学』2008年4月臨時増刊号 (27/5), 114-128.

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

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:

43-66.

山口昌也, 高田智和, 北村雅則, 間淵洋子, 小林正行, 西部みちる (2008) 『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.0』文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築: 21世紀の日本語研究の基盤整備」平成19年度研究成果報告書 (JC-D-07-03)

山口昌也 (2007) 「全文検索システム『ひまわり』/設定ファイル作成の手引き」 URL: <http://www.kokken.go.jp/lrc/index.php?plugin=related&page=全文検索システム『ひまわり』/設定ファイル作成の手引き>

Aston, Guy & Lou Burnard (1998) *The BNC Handbook*. Edinburgh: Edinburgh University Press. (北村裕監訳 (2004) 『The BNC Handbook コーパス言語学への誘い』松柏社.)

Burnard, Lou (2004) “Metadata for corpus work,” in Wynne, Martin (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Pp. 30-46. Also available from the URL: <http://ahds.ac.uk/creating/guides/linguistic-corpora/>

Burnard, Lou (2006) “XXQ: a query language for XML corpora,” *Proceedings of the TEI Day in Kyoto 2006, 17. May, 2006, Kyoto*. Kyoto: Kyoto University. P. 124.

Burnard, Lou & Syd Bauman (2007) *Guidelines for Electronic Text Encoding and Interchange* (P5). URL: <http://www.tei-c.org/Guidelines/P5/>

Clark, James Clark & Steve DeRose 1999 *XML Path Language (XPath) Version 1.0*. W3C Recommendation, 16. November, 1999. URL: <http://www.w3.org/TR/xpath/>

Cummings, James (2006) “Exploring TEI XML Documents with XQuery,” *Proceedings of the TEI Day in Kyoto 2006, 17. May, 2006, Kyoto*. Kyoto: Kyoto University. Pp. 99-115.

Dodd, Tony (2006) *XXQ: An Informal Introduction*. Manuscript. Oxford: Oxford University Computing Services. URL: <http://www.oucs.ox.ac.uk/rts/xaira/Doc/XXQdoc.xml>

Garside, Roger, Geoffrey Leech & Anthony McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Ide, Nancy (2000) “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, Patrice Bonhomme & Laurent Romary (2000) “XCES: An XML-based encoding standard for linguistic corpora,” *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association. Pp. 825-830.

Xiao, Richard (2006) “Review: Xaira – an XML aware indexing and retrieval architecture,” *Corpora* 1/1: 99-103.

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: <http://www.w3.org/TR/REC-xml/>

## **Xaira に関する参照 URL**

*Indexing a Corpus with XAIRA: a Tutorial*. URL:

<http://www.oucs.ox.ac.uk/rts/xaira/Doc/xairaIndexing.xml>  
*Introducing XAIRA Briefly.* (Aston, Guy) URL:  
<http://www.oucs.ox.ac.uk/rts/xaira/Doc/GuysbriefXairaIntro.htm>  
*Using the ANC with Xaira.* URL: <http://americannationalcorpus.org/xaira.html>  
*Using the BNC XML Edition with Xaira.* URL:  
[http://www.natcorp.ox.ac.uk/tools/bncXml\\_search.xml](http://www.natcorp.ox.ac.uk/tools/bncXml_search.xml)  
*Using Xaira Under Windows: Reference Guide to the Windows Client.* URL:  
<http://www.oucs.ox.ac.uk/rts/xaira/Doc/refman.xml>  
*Xaira Page.* URL: <http://www.oucs.ox.ac.uk/rts/xaira/>  
*Xaira Sourceforge Home.* URL: <http://xaira.sourceforge.net>