『現代日本語書き言葉均衡コーパス』における 形態論情報付き XML フォーマット

小木曽智信　　間淵洋子　　前川喜久雄

人間文化研究機構 国立国語研究所

1. はじめに

発表者らが構築に関わっている「現代日本語書き言葉均衡コーパス」(BCCWJ)は、2010 年度を以て構築期間を終え 2011 年中に一般公開を開始する予定である。公開の形式としては、Web オンラインサービスのほかに、ディスク媒体でのデータ提供を予定している。これは主に研究者を対象としたもので、ソースとなる XML 形式のデータをすべて納めたものとなる予定である(前川 2008)。

BCCWJ の XML フォーマットとしては、これまでにテキストに文書構造を単純にマークアップした形式を提案し試験公開を行ってきた。この形式は文字列に依拠した利用に対しては十分な対応が可能であったが、BCCWJ の形態論情報を埋め込んで利用するためには不十分な点があった。

本発表では、この文字列ベースの XML フォーマットをもとにして、言語構造を一定程度反映させた新しい XML フォーマットを提案する。これにより、短単位・長単位をはじめとする形態論情報や、文を単位とする情報などの言語構造に関わる情報を付与することを可能にする。

2. 文字ベースの XML とその問題点

2.1. 文字ベースの XML のタグセット

BCCWJ では、ランダムサンプリングによって採集したサンプルから、長さを 1000 字に固定した固定長サンプルと、節や章など文章の意味上のまとまりをとりだした可変長サンプルの 2 種類を作成している。固定長と可変長のサンプルは別個に取得するのではなく、同一のサンプリングポイントから、2 通りの方法によって重複部分を持つ形で作成している。

各々のサンプルは、XML 形式で表 1 に示すタグを用いてマークアップを施される(山口ほか 2008)。マークアップにあたっては単語等の切れ目は意識していない。

なお、文を示す sentence タグは、入れ子構造を許しており、大きな文の中に複数の文が含み込まれることがある。

表 1　文字ベースの XML フォーマットの主なタグ

種類	タグ	説明
サンプル	sample	サンプリングによって 1 サンプルとされた文章の範囲
	sampling	サンプリングポイントに関する情報
階層構造 (文書構造)	article	同一著者による、同一テーマのひとまとまりの文章
	title	ある範囲の文章の内容を代表する記述。章の題、新聞の見出しなど
	cluster	title 要素がまとめる文章の範囲
	list	箇条書きや名詞句の羅列など、列挙された要素
	paragraph	段落に相当する文の集まり
	sentence	文に相当する語の集まり
図 表 (文書構造)	figure	図・表・写真・絵など
	caption	図表等についてのタイトルや説明
引 用 (文書構造)	citation	当該 article 要素とは異なる著作物からの引用
	speech	発話や心内発話の引用・書き起こし
	quote	行内における引用・発話表現
注 記 (文書構造)	noteBody	脚注、後注など、本文と区別して記述される注記
その他 (文書構造)	abstract	article 要素、または cluster 要素の概要に相当する要素
	verse	詩、和歌、俳句、歌謡などの韻文
文字・表記	ruby	ルビ付き文字
	correction	原文の誤植を訂正した文字
	missingCharacter	規定の文字集合に含まれない文字 (JIS 外字)

2.2. BCCWJ の形態論情報

一方、BCCWJ では、すべてのサンプルに対して形態論情報の付与が行われる。形態素解析辞書 UniDic の解析結果に基づく短単位と、これを組み上げた長単位の二つの単位による情報が付与される（伝ほか 2007, 小椋ほか 2010）。

短単位は、単位の認定、品詞や見出しの付与方法などについて詳細な規定を定めた言語単位である。和語の場合は単純語または単純語 2 語の結合を 1 短単位とし、漢語の場合は二字漢語までを 1 単位とするものである。助詞等の付属語や記号も 1 単位となる。次の例文の「/」が短単位境界である。

/国立/国語/研究/所/で/研究/し/て/いる

一方、長単位は、この短単位を組み上げたもので、文節から付属語を取り去ったものが長単位に相当する。付属語は原則として短単位単独で長単位となるが、複合辞として認定した「ている」などは 1 長単位となる。また、「研究し」は漢語サ変動詞として 1 長単位にまとめられる。次の例文の「/」が長単位境界である。

/国立国語研究所/で/研究し/ている/

したがって、短単位・長単位・文節は入れ子の構造を取る。文節はこれが連なって文を構成するし、短単位は文字から構成されるから、BCCWJ の形態論情報は、結局次のような言語単位の階層構造の中に位置づけられることになる。

文章／文／文節／長単位／短単位／文字

XSLT などを用いて形態論情報を活用するためには、この階層構造・包含関係がそのまま XML フォーマットに反映されることが望ましい。

2.3. 文字ベースの XML と形態論情報の齟齬

2.1.で示した文字ベースの XML フォーマットは、2.2.で示した言語単位の階層構造ときれいに対応しない場合がある。ruby タグはその典型的な例である。

ルビ（ふりがな）は、次の 1)～5)のように単語中の一部分の文字に対してつけられる場合から、一文に対して一つのルビが対応するようなものまで様々なものが存在する。BCCWJ の ruby タグは原則として単漢字に対するルビとして付与されているが、熟字訓などでは複数の文字にまたがることになる。

1) 語彙（い） （短単位よりも短いルビ）
2) 時雨（しぐれ） （短単位と一致するルビ）
3) 喜望峰（ケープタウン） （短単位よりも長いルビ）
4) 新しい芸術（アール・ヌーヴォー） （長単位よりも長いルビ）
5) 達者でな（アスタ・ラ・ビスタ） （文全体にかかるルビ）

文字ベースの XML では上記のような例は単純に範囲内の文字列を ruby タグで囲み、ルビ文字を rubyText 属性の値としてきた。これらが短単位の形態論情報タグ（SUW）とともにマークアップされるとき、例 1)のように短単位よりも短い ruby は SUW の子要素とならざるを得ない。一方、例 3)～5)では、逆に ruby は SUW の親要素となるほかない。

```
1a) <SUW>語<ruby rubyText="い">彙</ruby></SUW>
2a) <SUW><ruby rubyText="しぐれ">時雨
    </ruby></SUW> or
    <ruby rubyText="しぐれ"><SUW>時雨
    </SUW></ruby>
3a) <ruby rubyText="ケープタウン"><SUW>喜望
    </SUW><SUW>峰</SUW></ruby>
4a) <ruby rubyText="アール・ヌーヴォー"><SUW>新しい
    </SUW><SUW>芸術</SUW></ruby>
5a) <ruby rubyText="アスタ・ラ・ビスタ"><SUW>達者
    </SUW><SUW>で</SUW><SUW>な</SUW></ruby>
```

ここでは省略するが、長単位タグ（LUW）を考えるときには、関係はさらに複雑なものとなる。したがってこのままでは形態論情報との上下関係が定まらず、利用上不便を来すこととなる。

このほかに引用タグ（quote）も短単位と齟齬を来す場合がある。引用文では、ときに用言の活用語尾の一部分だけが引用され、残りが地の文で補われる場合がある。

```
6) <quote>「解剖後厚く弔」</quote>うべしという指示
```

このとき、短単位「弔う」は quote の終了タグを越えることになる。ただし、これは単にタグだけの問題ではない。引用符（“」”）が短単位内に入り込んでいるため、この文字までが問題となる。

2.4. 文認定の問題

文（sentence）の認定をめぐっては、sentence タグの入れ子が認められているという問題がある。たとえば、次のように文中に引用がある場合には、全体を sentence で囲みつつ、引用部分も sentence でマークアップされている。

```
7) <sentence>驚きながらそう誤魔化した構治の言葉に、
   <quote>「<sentence>落ちた、落ちたって言わないでよ。</sentence><sentence type="quasi">結構辛がってるんだから</sentence>」</quote>言って夕美子
```

は目を伏せ、（中略）うつむいている。</sentence>

複雑な構造をとる文の場合、そのいずれもが文として認められるという点で、このマークアップにも積極的な意味がある。

しかし、(1) 上位の sentence がきわめて長くなる場合がある (2) 形態素解析などの解析ツールの入力となる「文」を定めがたい (3) データを文番号で管理できない、などのデメリットがある。

文についてはまた、sentence タグが付与されない環境が生じているという問題がある。verseLine は詩歌の行を示すタグであるが、これが用いられる場合には、文（sentence）の認定を行わず、原文の改行位置を基準にそのまま verseLine としてきた。そのため、verseLine は sentence を親に持たない特殊な要素となっており、また、原文の状況によっては形態論情報の単位と齟齬を来す可能性がある。

8) <verseLine>霊山の</verseLine>
<verseLine>誓いも深き</verseLine>
<verseLine>君ら西</verseLine>
<verseLine>我ら東と</verseLine>
<verseLine>白馬も雄々しく</verseLine>

2.5. 固定長と可変長の問題

2.1.で触れたとおり、文字ベースの XML では、固定長と可変長を別の XML ファイルとして扱っていた。文字列を対象とした調査を行う場合には別ファイルとなっていることが望ましい場合も多いが、データに対して新たな情報を付与する場合には問題となる。

たとえば、自動で付与された形態論情報に対して人手で修正を施す場合には、重複部分について二度手間が生じるほか、同一箇所に異なる形態論情報が付与される可能性が生じる。

したがって、特に形態論情報を付与する場合には、固定長と可変長を統合した形式をソースとし、そこから固定長・可変長の二つの情報が取得できるようにすることが望ましい。

3. 形態論情報付き XML フォーマット

3.1. 基本方針

以上のような問題点を踏まえ、新しい形態論情報付きの XML フォーマットは、これまでの XML との互換性をできる限り確保しつつ，言語構造と齟齬を来す要素について修正を行うこととした。さらに、新フォーマットから旧フォーマットへは自動変換できるように設計している。

3.2. 階層構造

2.2.で示した形態論情報の階層構造に、表 1 のタグを納めるならば、次のような階層が考えられる（網掛けはすべてのテキストに必須の要素）。

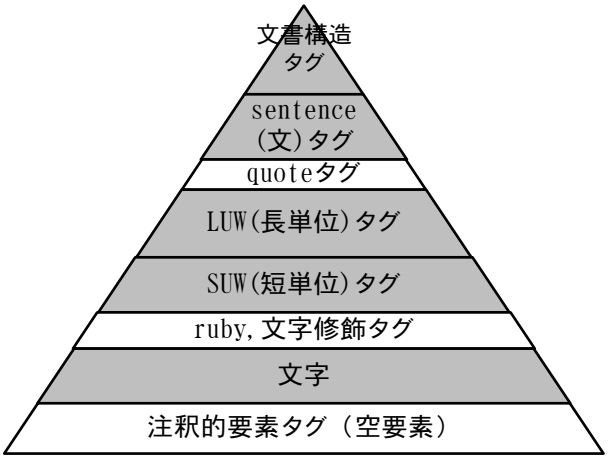

図 1 形態論情報付き XML フォーマットの階層構造

この構造に照らせば、文字ベースの XML フォーマットの問題は、文（sentence）タグの階層が一律に付与されていないことが第一の問題である。そして、ruby が上の階層に飛び出したり、quote が下の階層を侵犯したりすることで、形態論情報と齟齬を来しタグの交叉を招くのが第二の問題だということになる。

3.3. 変更点

これらの問題点を解消するため、新しい XML フォーマットでは固定長・可変長を統合した XML をベースとして、各タグについて次のような対処を行った。

A. 文（sentence）

文タグの階層を整備するために、sentence の入れ子を認めることをやめ、上位の文は superSentence として文書構造タグの一種とした。下位の sentence はそのまま残し、superSentence の一部分を新たに sentence で囲み type="fragment"とした。

7') <superSentence>
<sentence type="fragment">驚きながらそう誤魔化した構治の言葉に、</sentence>
<quote><sentence>「落ちた、落ちたって言わないでよ。</sentence>
<sentence type="quasi">結構辛がってるんだから」</sentence></quote>
<sentence type="fragment">言って夕美子は目を伏せ、（中略）うつむいている。</sentence>
</superSentence>

また、verseLine については改行位置を空要素の verseLine タグとして残しつつ、文相当の範囲を新たに sentence で囲み、type="verse"とした。

8') <sentence type="verse">霊山の<verseLine/>誓いも深き<verseLine/>君ら西<verseLine/>我ら東と<verseLine/>白馬も雄々しく<verseLine/></sentence>

これにより、すべての短単位はいずれかの sentence に属することとなり、サンプルは sentence の集合としても捉えられることとなった。

B. ルビ（ruby）

短単位を越えるルビについては、先頭の短単位を ruby タグで囲み、そのタグの属性値として本来のルビ範囲のテキストを保持することとした。これにより、元の状態に戻すことを可能にすると同時に、複数単位に渡る特殊なルビを容易に取り出すことを可能にしている。

3a')<SUW><ruby rubyText="ケープタウン" rubyBase="喜望峰">喜望</ruby></SUW><SUW>峰</SUW>

4a') <SUW><ruby rubyText="アール・ヌーヴォー" rubyBase="新しい芸術">新しい</ruby></SUW><SUW>芸術</SUW>

C. 引用（quote）

短単位を分断する引用については、引用符のテキストを移動し、元の場所に空要素タグを残すことで対処した。

6') <quote>「解剖後厚く弔<move type="original" text="」"/>う<move type="modify">」</move></quote>べしという指示

これにより短単位 SUW で引用符（“」”）を含まない「弔う」を囲むことが可能になると同時に、quote と SUW の交叉も解消される。

D. 注釈的要素タグの空要素化

これらのタグ以外に、元のタグセットの仕様では、本来ならば本文テキストとして扱うべきでない文字列がそのまま残されている場合があった。たとえば注釈タグ（noteBody）関連のタグがその一つである。

9) 国際ルールに反しない形でタイド化を行っている<noteMarker>（注１）</noteMarker>

これについては次のような空要素タグに仕様を変更することで問題を解消している。

9') 国際ルールに反しない形でタイド化を行っている<noteMarker text="（注１）"/>

このような空要素化処理をする場合、属性値に入れられるテキスト部分にタグが用いられていることがある。

10) <noteMarker><enclosedCharacter description="○">６６</enclosedCharacter><noteMarker>

これはタグ表記が必要な丸付き数字を含むテキストだが、この場合には次のような記法によって info 属性に元の情報を保持できるようにした。

10') <noteMarker text="６６" info="enclosedCharacter:description=○"/>

3.4. 文字ベースの XML との互換性

3.3.で示した変更点は、原則として元の情報を保持したまま、形態論情報との併存を図ったものである。したがって、この形態論情報付き XML フォーマットから、文字ベースの XML フォーマットに変換することが可能である。可変長・固定長の文字ベースの XML フォーマットは、今後も引き続き提供される予定である。

4. おわりに

以上、BCCWJ の新しい形態論情報付き XML フォーマットについて述べた。一般に書き言葉のテキストでは、言語上の単位と表記法とが一致しない場合があるが、そのような場合の対処事例として参考になれば幸いである。

参考文献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22　pp.101-123

前川 喜久雄（2008）「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」日本語の研究 4-1

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕（2010）『『現代日本語書き言葉均衡コーパス』形態論情報規程集第３版』

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』

付記

本発表は文科省科学研究費特定領域研究「日本語コーパス」（領域代表者：前川喜久雄）による成果の一部を含むものである。