第 9 章　形態論情報付き統合形式 XML（M-XML）

小木曽 智信　間淵 洋子　前川 喜久雄

9.1　M-XML の概要

形態論情報付き統合形式 XML（Morphology-base XML 以下、M-XML と略記する）は、文字ベースの XML（C-XML）フォーマットをもとにして、固定長・可変長サンプルを統合し、言語構造を一定程度反映させた XML フォーマットである。短単位・長単位の形態論情報を、階層構造を維持したまま埋め込み、言語構造に関わる情報を扱いやすくしている。XML ファイルの文字符号化方式は UTF-8（BOM なし）である。

第 6 章で述べたとおり、M-XML には、数字変換（NumTrans）処理を施した M-XML_NT と、数字変換を行っていない M-XML_OT の 2 種類の本文がある。それぞれのデータの格納場所については第 1 章を参照されたい。

短単位・長単位の形態論情報は、M-XML・TSV の両形式とも同じ内容が付与されており、同一部分の短単位・長単位が異なって解析されていることはない。

9.1.1 固定長と可変長の統合

C-XML では、固定長サンプルと可変長サンプルが別の XML 文書として構造化されている。しかし、2 種類のサンプルは同一の文書から採集されているため、多くの部分が重複している。こうしたデータに形態論情報を付与し整備する場合には、同一内容のテキストを 2 回処理する必要がないように、統合して扱うことができた方が望ましい。しかし、タグが交叉することになるため、別の構造を持つ二つの XML を単純に統合することはできない。そこで、統合形式では以下のような方法によって固定長と可変長を統合することとした。

そもそも、文書構造を意識して採集された可変長サンプルとは違い、均一な長さのサンプルを取得する目的で作られた固定長サンプルでは、文書構造を示すブロック要素タグは大きな意味を持たない。そこで、M-XML では、可変長サンプルの文書構造だけを保持し、固定長の範囲は形態論情報（長単位）タグに付与する属性で示すこととした。可変長部分から固定長部分がはみ出している場合には、はみ出した部分を単純なコンテナ（ ）で囲み、インライン要素だけを保持した。

M-XML は次のような属性を持つ mergedSample 要素をルートとして上記の要素をまとめ上げている。

```
<mergedSample sampleID="サンプル ID" type="BCCWJ-MorphXML" version="1.1">
```

なお、NumTrans 処理が行われた M-XML_NT のサンプルについては、次のように NumTrans 属性を付与して区別している。

```
<mergedSample sampleID="サンプル ID" type="BCCWJ-MorphXML" version="1.1"
NumTrans="true">
```

M-XML_NT のファイルであっても、対象となる数字列が存在せず、NumTrans 処理がなされていないものについてはこの属性は付与されていない。したがって、こうしたサンプルにおいては、M-XML_NT と M-XML_OT のファイルが完全に一致する。

9.1.2 異なる文書型定義の統合

C-XML は、サブコーパス・レジスターによって異なる文書型定義（DTD）が用いられている。Yahoo!知恵袋（OC）、Yahoo!ブログ（OY）、教科書（OT）、韻文（OV）は、おおよそ共通の構造を持ちながらも、一般の可変長サンプルとは異なるそれぞれ独自の文書型定義によっている。そのため、すべてのデータを統一的に処理しようとするとき問題となる場合がある。

そこで、M-XML では、タグセットを一部変更して、すべてのサブコーパス・レジスターについて共通の文書型定義で処理できるようにした。C-XML に比較してやや緩い制約での検証になるが、すべての XML ファイルは単一の XML スキーマで検証済みである。この統合に際してレジスター独自のタグを次のように一部変更している。

OC　:　`<OCQuestion>` → `<article articleID="`サンプル ID-Question`">`
　　`<OCAnswer>` → `<article articleID="`サンプル ID-Answer`">`
OC, OY :　`<br type='physicalLine_original' />` → `<webBr/>`
OT　:　`<root>` → `<squareRoot>`

9.2　要素の階層構造

BCCWJ における短単位・長単位・文節は、その定義から入れ子構造をなす。また、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

文書構造タグや階層化された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。この考え方に従い、M-XML では、次のような階層構造で形態論情報を付与した。

文書構造（ブロック）タグ／sentence（文）／LUW（長単位）／SUW（短単位）／文字

以下はそのサンプルとしてひとつの文（sentence 要素）を抜き出したものである（見やすさのため属性を省略した。形態論情報タグの詳細は第 6 章を参照のこと）。

```
<sentence>
  <LUW><SUW>公共</SUW><SUW>工事</SUW><SUW>請負</SUW><SUW>金額</SUW></LUW>
  <LUW><SUW>の</SUW></LUW>
  <LUW><SUW>動き</SUW></LUW>
（略）
</sentence>
```

以上の形態論情報の階層に C-XML の諸要素を当てはめるならば、図 9-1 のような階層構造が考えられる（網掛けはすべてのテキストに必須の要素）。このとき C-XML における諸要素がこの階層と齟齬を来すことが問題となるが、M-XML では、次節以降に示すように C-XML のタグに修正を加えることで対処している。

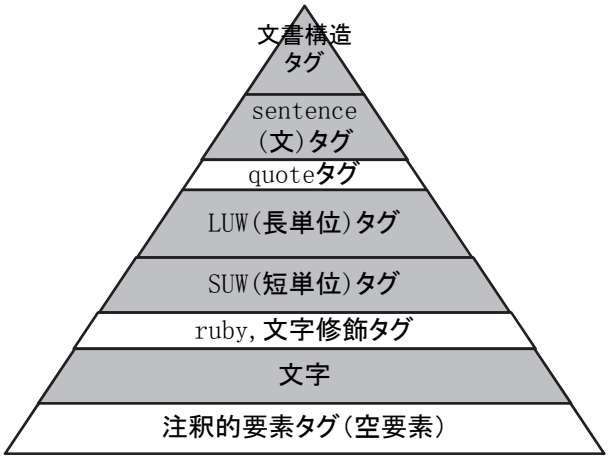

図 9-1: 形態論情報付き統合形式 XML フォーマットの階層構造

9.3 C-XML と M-XML の相違点

これまでに見てきたとおり、C-XML と M-XML の大きな相違は次の 4 点である。

- 固定長と可変長が統合されていること（9.1.1 節）
- レジスターごとに異なる文書型定義が統一されていること（9.1.2 節）
- 短単位（SUW）・長単位（LUW）の形態論情報が付与されていること（第 6 章）
- 文（sentence）タグが修正されていること（第 8 章）

ただし、これ以外にも M-XML ではいくつかの修正がなされている。以下、上記以外の C-XML と M-XML の相違点や追加されたタグについて述べる。

9.3.1 数字変換（NumTrans タグ）

第 6 章で述べたとおり、M-XML_NT においては、数字変換（NumTrans）処理がなされている。この処理が行われた箇所には、次のように NumTrans タグが付けられ、変換前の本文はこのタグの originalText 属性に保存されている。

OT テキスト：１９８６

NT テキスト：<NumTrans originalText="１９８６">千九百八十六</NumTrans>

実際には形態論情報が付与されているから、M-XML の当該箇所は以下のようになる（見やすさのため形態論情報の一部を省略した）。

- M-XML_OT：

```
<LUW l_lemma="一八九六" l_lForm="イチハチキュウロク">
  <SUW lemma="一" lForm="イチ">１</SUW>
  <SUW lemma="八" lForm="ハチ">８</SUW>
  <SUW lemma="九" lForm="キュウ">９</SUW>
  <SUW lemma="六" lForm="ロク">６</SUW>
</LUW>
```

- M-XML_NT：

```
<LUW l_lemma="千八百九十六" l_lForm="センハッピャクキュウジュウロク">
  <NumTrans originalText="１９８６">
    <SUW lemma="千" lForm="セン">千</SUW>
    <SUW lemma="八百" lForm="ハッピャク">八百</SUW>
    <SUW lemma="九十" lForm="キュウジュウ">九十</SUW>
    <SUW lemma="六" lForm="ロク">六</SUW>
  </NumTrans>
</LUW>
```

9.3.2 分数（fraction タグ）

C-XML では帯分数にのみ fraction タグが付与されているが、M-XML では帯分数以外の分数にも fraction タグが追加されている。

```
<fraction>１／１０</fraction>
```

M-XML_NT では、さらに NumTrans 処理によって分子（numerator）・括線（vinculum）・分母（denominator）が次のようにタグ付けされ、分子と分母の順が入れ替えられている。2 桁以上の数字の変換も合わせて行われている。

```
<fraction>
 <denominator><NumTrans originalText="１０">十</NumTrans></denominator>
  <vinculum><NumTrans originalText="／">分</NumTrans></vinculum>
  <numerator>１</numerator>
</fraction>
```

実際には形態論情報が付与されているから、M-XML の当該箇所は以下のようになる（見やすさのため形態論情報の一部を省略した）。

- M-XML_OT :

```
<fraction>
  <SUW lemma="一" lForm="イチ" pos="名詞-数詞">１</SUW>
  <SUW lemma="／" lForm="" pos="補助記号-一般">／</SUW>
  <SUW lemma="一" lForm="イチ" pos="名詞-数詞">１</SUW>
  <SUW lemma="零" lForm="レイ" pos="名詞-数詞">０</SUW>
</fraction>
```

- M-XML_NT :

```
<fraction>
  <denominator>
    <NumTrans originalText="１０"><SUW lemma="十" lForm="ジュウ">十</SUW></NumTrans>
  </denominator>
  <vinculum>
    <NumTrans originalText="／"><SUW lemma="分" lForm="ブン">分</SUW></NumTrans>
  </vinculum>
  <numerator>
    <SUW lemma="一" lForm="イチ">１</SUW>
  </numerator>
</fraction>
```

9.3.3 ルビの処理

形態論情報を付与する際、ルビのタグが形態論情報と齟齬を来す場合があるため、次のように対処した。

BCCWJ では、ふりがなは原則として 1 文字ごとに付与しているが、熟字訓や臨時的な読みでは複数文字を ruby タグで囲んでいる。たとえば次のような例がある。

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)

これらは C-XML では次のようにタグ付けされている。

1a) `<SUW>語<ruby rubyText="い">彙</ruby></SUW>`
2a) `<SUW><ruby rubyText="しぐれ">時雨</ruby></SUW>` もしくは
`<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>`

M-XML では、1a) 2a)のように、短単位タグの内側に ruby タグを置くことができる場合

にはそのままとした。一方、3a) 4a)のように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、元の状態に戻せるようにするとともに、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>

9.3.4 その他の追加されたタグ

改ページ位置を示す参考情報が空要素の info タグに残されている。

以上のように、できる限り互換性を保持するように努めているものの、各種の変更を加えているため、M-XML に付与されたタグと C-XML のタグとの間に完全な互換性はない。

参考文献

小木曽智信・間淵洋子・前川喜久雄（2011）「『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」『言語処理学会第17回年次大会講演論文集』,352-355.

山田篤・小磯花絵 （2008） 『NumTrans マニュアル』, The UniDic Consortium.