第4章　文書構造情報付き文字ベース XML（C-XML）

山口 昌也

4.1　はじめに

本章では、文書構造情報を付与した文字ベースの（形態論情報を含まない）XML 文書（Character-base XML、以下 C-XML と略記する）の仕様について、(1)文書構造タグ、(2)文字入力、(3)形態論情報付き XML 文書（M-XML：第 9 章参照）との相違点の三つに分けて説明する。なお、本章の内容の詳細については、山口他（2011）、西部他（2011）を参照されたい。

4.2　文書構造タグセットの種類とサブコーパス・レジスターとの関係

BCCWJ は複数のサブコーパス・レジスターから構成される。文書構造タグのセット（タグセット：TS）は、それぞれのサブコーパス・レジスターの特性に合わせて、表 4-1 のように規定される。個々のタグセットは、XML の文書型として定義される。なお、原資料が紙媒体のデータについては、sentence（後述、表 4-2 参照）など一部の要素を除き人手で付与しているが、電子媒体のデータについては、より多くの部分で自動付与を行うなど、個々に方法が異なる。同じタグであってもレジスターの種類によってタグの性質や付与の精度に差が生じることがあるため、注意が必要である。タグ付与方法の詳細については西部他（2011）を参照されたい。

タグセットは、次の 3 種類に大別される。表中で「可変長（一部修正）」とあるのは、可変長タグセットに部分的な変更を加えたタグセットであることを意味する。この後の節では、まず「可変長タグセット」「固定長タグセット」「Yahoo!知恵袋タグセット」について解説し、そのあとレジスターごとに個別の変更部分を説明する。

可変長タグセット（可変長 TS）：　可変長サンプル（ひとつのサンプルがひとつの「記事」に相当するサンプル）を記述するためのタグセット

固定長タグセット（固定長 TS）：　固定長サンプル（ひとつのサンプルに 1,000 文字を包含するサンプル）を記述するためのタグセット

Yahoo!知恵袋タグセット（Yahoo!知恵袋 TS）：　「Yahoo!知恵袋」レジスターのサンプルを記述するためのタグセット

表 4-1: サブコーパス・レジスターとタグセットとの関係

サブコーパス･レジスター	タグセット	原資料の媒体
出版サブコーパス（PB,PM,PN）	可変長 TS、固定長 TS	紙媒体
図書館サブコーパス（LB）	可変長 TS、固定長 TS	紙媒体
白書(OW)	可変長 TS、固定長 TS	紙媒体
教科書(OT)	可変長 TS（一部修正）	紙媒体
広報紙(OP)	可変長 TS	電子媒体
ベストセラー(OB)	可変長 TS	紙媒体
Yahoo!知恵袋(OC)	Yahoo!知恵袋 TS	電子媒体
Yahoo!ブログ(OY)	可変長 TS（一部修正）	電子媒体
韻文(OV)	可変長 TS（一部修正）	紙媒体
法律(OL)	可変長 TS	電子媒体
国会議事録(OM)	可変長 TS	電子媒体

4.3 可変長タグセット

可変長タグセットは、可変長サンプル（ひとつのサンプルがひとつの「記事」に相当するサンプル）を記述するためのタグセットである。タグの種類は、46 種類である。タグの一覧を表 4-2 に示す。また、紙媒体の原資料とタグづけ結果の例を図 4-1 に示す。

本タグセットによって付与される情報は、次の三つに大別される。

- サンプルに関するタグ：サンプルに関するタグには、sample と sampling がある。sample 要素は、ひとつのサンプルの範囲を表す。sampling タグは、サンプル抽出基準点などサンプリングに関する情報を表す。
- 文字・表記に関するタグ：この種のタグの役割は、(1)検索や計算機処理の利便性を高めること、(2)原資料に忠実に電子化テキストを記述することである。前者のタグの例として、correction タグ（誤植を修正した文字を表す）がある。

```
生活基<correction type="erratum" originalText="盟">盤</correction>に
伸びを示し<correction type="omission">て</correction>いる
整備を<correction type="excess" originalText="を" />図るべく
```

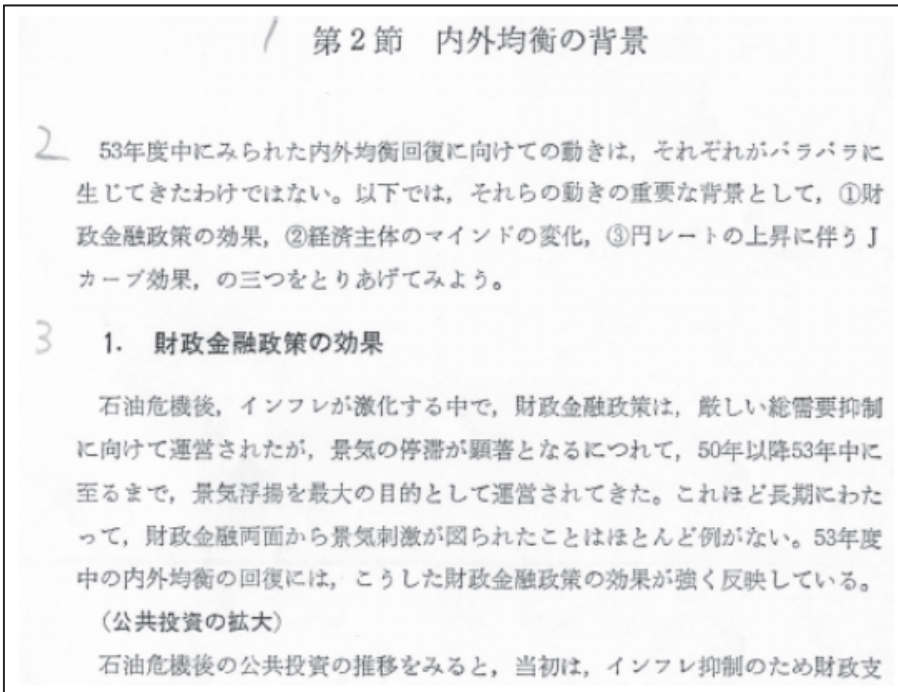
後者の例として、ruby タグ（ルビ付き文字を表す）、missingCharacter タグ（文字セット外字を表す）の例を次に示す。

```
<ruby rubyText="ご">語</ruby><ruby rubyText="い">彙</ruby>
<missingCharacter attribute="HanIdeograph" unicode="U+5AEB"
        daikanwa="M06673" description="女偏に莫">＝</missingCharacter>
```

- 文書構造に関するタグ：文書構造に関するタグは、見出し、概要、キャプション、注記など、文書中における論理的な役割が明確な文書要素に対して付与される。表 4-2 に示したとおり、この種のタグは、(a) 階層構造、(b) 図表、(c) 引用、(d)注記、(e)その他に分けられる。

  このうち、階層構造に関するタグについて、図 4-1 と対応づけて説明する。階層構造に関するタグは、article を最上位の階層として、cluster、paragraph、sentence といった言語的な階層構造を表現する。図 4-1 から、これらの要素に関係する部分を取り出すと次のようになる。なお、字下げは、下位の階層であることを示す。例えば、図 4-1 の article 要素直下の階層には、titleBlock 要素、paragraph 要素、cluster 要素があることがわかる。

```
article
  titleBlock 第２節　内外均衡の背景
  paragraph
  cluster
    titleBlock １．財政金融政策の効果
      cluster
        titleBlock （公共投資の拡大）
```



第２節　内外均衡の背景

53年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。以下では，それらの動きの重要な背景として，①財政金融政策の効果，②経済主体のマインドの変化，③円レートの上昇に伴うＪカーブ効果，の三つをとりあげてみよう。

1.　財政金融政策の効果

石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，50年以降53年中に至るまで，景気浮揚を最大の目的として運営されてきた。これほど長期にわたって，財政金融両面から景気刺激が図られたことはほとんど例がない。53年度中の内外均衡の回復には，こうした財政金融政策の効果が強く反映している。

（公共投資の拡大）

石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet href="sc_check.xsl" type="text/xsl" ?>
<sample sampleID="OW1X_00000" version="20070208" type="variableLength">
<article articleID="OW1X_00000_V001" isWholeArticle="false">
<titleBlock><title><sentence type="quasi">第２節　内外均衡の背景
</sentence></title></titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。</sentence><sentence>以下では，それらの動きの重要な背景として，...
</paragraph>
<cluster>
<titleBlock><title><sentence type="quasi">１．財政金融政策の効果
</sentence></title></titleBlock>
<paragraph>
<sentence>　石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，５０年以降５３年中に至るまで，景気浮揚を最大の目的として運営されてきた。</sentence> ...
</paragraph>
<cluster>
<titleBlock><title><sentence type="quasi">（公共投資の拡大）</sentence></title></titleBlock>
<paragraph>
<sentence>　石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支出が抑制され，公共事業の伸びは低いものにとどまっていた。</sentence>
```

図 4-1: 原資料とその電子化テキストの例(『経済白書昭和 54 年版』から引用)

表 4-2: 可変長タグセット

	タグ名	内容
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書要素
	sampling*	サンプル抽出基準点などサンプリングに関する情報
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書要素
	blockEnd	意味のまとまりや形式のまとまりを区切るためのマーカー
	cluster	titleBlock 要素が包括する文書要素全体
	titleBlock	title 要素とそれに付随する要素全体
	title	特定範囲の文書要素の内容を代表する記述
	orphanedTitle	不特定範囲の文書要素を代表する記述
	list	箇条書きなど、列挙された文書要素の集まり
	listItem	List 要素を構成する各並立要素
	paragraph	段落を表す文書要素
	sentence*	文に相当する文書要素
図表（文書構造）	figureBlock	図表・写真・絵などの要素と、それに付随する文書要素をまとめた要素
	figure	付随する文書要素のある図・表・写真・絵など
	caption	図表についてのタイトルや説明
	table	表
引用（文書構造）	quotation	当該 article 要素とは異なる著作物からの引用や、発話・心内発話の引用・描写・書き起こし描写・書き起こし図表・写真・絵などの要素と、それに付随する文書要素をまとめた要素
	citation	当該 article 要素の本文において言及される、他文献からの引用要素
	source	引用文献についての情報(文献名、著者名、著者情報など)
	speech	発話の引用・書き起こし、心内発話の描写
	speaker	話者を明示的に表した文字列やマーク
	quote*	当該 article 要素とは異なる著作物からの引用や、発話・心内発話の引用・描写・書き起こし、「 」で表されるさまざまな表現
注記（文書構造）	noteBody	注記とその注記の範囲
	noteBodyInline*	傍注など行外に付随する形式で現れる注記
	noteMarker*	注番号や参考文献番号など、他の文書要素を参照する際の目印として機能する文字列
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文書要素
	authorsData	著作者表示・署名にあたる要素
	contents	目次に相当する文書要素
	profile	著者や登場人物のプロフィールに相当する文書要素
	rejectedBlock	サンプル範囲内において、削除対象となったブロック要素の存在
	verse	詩、和歌、俳句、歌謡などの韻文
	verseLine	韻文における行
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	JIS X 0213:2004 で規定されている文字以外の文字（JIS 外字）
	enclosedCharacter	連続や参照などのラベルとして機能している囲み付きの文字
	cursive	変体仮名
	image	JIS X 0213:2004 が規定する諸記号に含まれていない記号類や絵文字
	superScript	数式や化学式などに用いる上付きの文字
	subScript	数式や化学式などに用いる下付きの文字
	fraction	帯分数の中の真分数部分
	delete	抹消線などによって削除された本文要素
	br	物理改行
	info	補助的な付与情報
	rejectedSpan	サンプル範囲内において、削除対象となったインライン要素の存在
	substitution	別の文字で代用入力されている JIS X 0213:2004 規定文字

※ 表中「*」付きの要素はインライン要素、それ以外の要素はブロック要素。

4.4 固定長タグセット

固定長タグセットは、固定長サンプル（ひとつのサンプルに 1,000 文字を包含するサンプル）を記述するためのタグセットである。可変長のタグセットとほぼ同じ仕様だが、固定長サンプルの収録範囲（文字数を基準に文を単位として限定される）に起因して、次の違いがある。

- 固定長タグセットのブロック要素は、当該要素の定義を満たす要素をすべて含むとは限らない。例えば、可変長タグセットにおける article 要素は「同一著者による、同一テーマのひとまとまりの文書要素」と定義され、記事や章などのまとまった文章範囲に相当するが、固定長の article 要素では、文章のまとまり全体を含まず、titleBlock 要素以外の本文が含まれない場合などもある。
- cluster 要素は認定しない。
- article 要素の isWholeArticle 属性は、IMPLIED （任意）である。

4.5 Yahoo!知恵袋タグセット

「Yahoo!知恵袋」レジスターのサンプルは、質問と回答の組という、一定の論理構造で構成される。しかし、可変長、固定長タグセットでは、この構造を十分記述することができないため、独立した文書型として定義した。タグの種類は、9 種類である。タグの一覧を表 4-3 に示す。また、サンプル例を図 4-2 に示す。

4.6 その他のタグセット

表 4-1 に示したとおり、レジスターの中には可変長タグセットを一部修正して記述しているものも含まれる。ここでは、可変長タグセットとの差異について説明する。

- Yahoo!ブログ
  - rejectedBlock タグの type 属性に ASCIIArt を追加した。これは、サンプル作成時に削除された、いわゆる「アスキーアート」を表す。
- 韻文
  - sample 要素の子要素に複数の article 要素を持つ。これは、「韻文」レジスターのサンプルには、1 サンプルに複数の作品が並列に含まれるためである。なお、可変長タグセットでは、sample 要素の子要素として、article 要素をひとつしか持たない。
- 教科書
  - 可変長タグセットに 5 種類のタグを追加するなど、「教科書」レジスター用に拡張している。詳細は、田中他（2011）「II 教科書コーパスの文字入力・タグ使用」を参照のこと。

表 4-3: 「Yahoo!知恵袋」レジスタータグセット

タグ名	内容
sample	質問本文と回答本文を対にしたもの
OCQuestion	質問本文を表す
OCAnswer	回答本文を表す
br	改行を表す
webLine	Web データに対して、自動で付与される、論理行相当の行を表す
sentence	文に相当するまとまりを表す
rejectedBlock	削除要素を表す
ncr	変換元データの数値文字参照を削除、または「=」に置換したことを表す
info	補助的な付与情報

```
<?xml version="1.0" encoding="UTF-8"?>
<sample sampleID="OC01_03216" type="chiebukuro" version="1.0">
<OCQuestion>
<webLine>
<sentence>ｗｉｎｄｏｗｓのＣＭで「税理士Ａの事件ファイル」という漫画をｗｅｂ上で公開しています、という男性が出ていますが、あのＣＭはフィクションですか？</sentence>
<sentence type="quasi">検索かけても出てきませんでした・・・・</sentence>
</webLine>
</OCQuestion>
<OCAnswer>
<webLine>
<sentence>税理士役も俳優さんらしいし、<br type="physicalLine_original" />完全なフィクションでしょう・・・。</sentence>
</webLine>
</OCAnswer>
</sample>
```

図 4-2: 「Yahoo!知恵袋」レジスターのサンプル例

4.7 文字入力仕様

本節では、BCCWJ に収録するデータを紙媒体（表 4-1）から作成する際の文字入力に関する仕様について述べる。なお、原資料が電子媒体のデータについては、データの性質上、この仕様に準拠しない点もある。詳細については、西部他（2011）の第 3 章を参照されたい。

4.7.1 基本方針

文字入力は、以下の基本方針に基づき行なった。

- 装飾、レイアウトなどの図形的情報を除いて文字を入力する（レイアウトの情報は、必要に応じて、タグで表現する）。

- 全ての文字種の入力に、いわゆる全角文字を用いる。
- 文字合成は行わない。
- 上記条件に抵触しない範囲で、原則として、原文を忠実に転記する。

4.7.2 文字符号化方式と文字集合

文字符号化方式は、以下に述べる文字集合を適切に符号化でき、テキストデータに対して施す形態素解析環境に適した方式として、UTF-8（BOM なし）を採用する。

文字集合は、JIS X 0213:2004 を用いる。ただし、次の文字については例外とし、それぞれ独自の方法で処理する。具体的な処理方法は、山口他（2011）を参照のこと。

- 入力対象外要素を構成する文字（例：ソフトハイフン、罫線素片）
- 装飾・デザインにかかわる文字（例：組み文字、分数、11 以上のローマ数字、囲み文字、上付き文字）
- 類似の非漢字
- 合成文字
- 入力が困難な文字（例：口偏に「七」の文字（「叱」面区点：1-47-52））

4.7.3 包摂規準

- 漢字
  - JIS X 0213 に準拠する。JIS X 0213:2000「6.6.3.1 漢字の字体の包摂規準の適用」（日本工業標準調査会 2000 参照）における包摂規準が適用される異体字については、これを区別しない。
- JIS X 0213 に定義されていない記号
  - JIS X 0213 に定義されていない記号であっても、原文の意味を損なわない場合、規格内の類似する記号に包摂してよいこととする。
- JIS X 0213 に定義されている記号
  - 字形の判別が困難な「長音記号」「負記号」「ダッシュ」「ハイフン」については、紙面上の形状ではなく、紙面上の意味によって入力し分けた。
  - その他の類似記号は独自に包摂規準を設けた。

4.7.4 外字

- 漢字、仮名、アルファベット
  - 漢字、仮名、アルファベットの JIS 外字は、当該の文字の代替として「**=**」（ゲタ）を入力すると共に、missingCharacter タグを用いて、タグ内部に属性として文字の情報を表す。

- 一般記号類
  - 入力対象外とする。ただし、語や文の構成要素になっているものについては、記号の代替として、image タグを挿入し、タグ内部に属性として記号の情報を表す。

4.7.5 特殊表記

- ルビ：ruby タグの rubyText 属性値として記述する。
- 上付き・下付き文字：それぞれ、superScript、subScript 要素として記述する。
- 囲み文字：囲みを無視して、囲まれている内部の文字を入力する。なお、連続・参照ラベルとして機能するもの（丸付き数字など）や、ある特定の語の略記号として機能するもの（「秘密」の意を表す丸付きの「秘」など）については、囲みの情報を、enclosedCharacter タグによって表す。
- 組み文字：組まれている文字をすべて 1 字ずつ切り離して入力する。
- 分数：「分子／分母」の形式に統一して入力する。ただし、帯分数の場合は、fraction 要素として記述する。
- 注記参照マーカー：「専門用語[2]」の上付きの「2」のような本文行から外れた位置にある注記参照用のマーカーは、noteMarker タグを付与する。
- 傍注：本文行の語や句の脇（行間など）に、注記が示されている「傍注」は、注記対象の語句の直後に、noteBodyInline タグを付与して示す。

4.7.6 レイアウト

- 空白
  - 入力対象となるもの：版面に現れる空白は、以下の場合に入力対象とする。その際、空白文字は常に 1 字分のみを入力する。
    - 段落冒頭の 1 字下げ
    - 語や文の区切り目を表すための空白
    - 「？」「！」などの後ろに挿入される空白
  - 入力対象とならないもの：上記以外の空白は、全てレイアウトによるものとみなし、無視する。例えば、以下のようなものをレイアウトとして入力対象としない。
    - 引用文、例文、項目等を本文行と区別するためのインデント
    - 中央揃え・右揃え・下揃え等の配置に伴うインデント
    - 文字幅を調整するためのスペース
- 改行

  改行は、版面の行の折り返しではなく、論理行（論理的に意味のある行。段落など意味のある切れ目で改行が施された行）で行う。具体的には、以下の要素の前後に改行を入れる。
  - 版面の行替えと一致する場合に改行するもの

- ✧ 段落
- ✧ 引用
- ✧ 韻文における行

➢ 版面の行替えと一致しない場合でも改行するもの

- ✧ タイトル
- ✧ 表の各セル

● リーダー・ダッシュ

リーダー・ダッシュが複数連続するものについては、すべて 1 字に置き換える。

4.7.7 誤植

原文に明らかな誤植がある場合は、これを訂正して入力する。ただし、原文の誤植を訂正した文字は、correction タグを用いて示し、原文の情報をタグ内部に originalText 属性として表す。以下に例を示す。

原文：

総トン数１００トン未満で長さ３０メートル<u>末</u>満の

タグづけ、および、修正：

総トン数１　０　０　トン未満で長さ３　０　メートル<correction type="erratum" originalText="末">未</correction>満の

なお、明らかな誤植とは、近似の字形の文字を誤って写植したもの（誤字）、前後の文字を逆に写植したもの（転倒）、脱字、衍字を指す。誤用や表記のゆれ、旧仮名遣い、仮名遣いの誤りなどは、これに含めない。詳細は、山口他（2011）を参照のこと。

4.8 M-XML との相違点

C-XML は、BCCWJ-DVD 版（Version 1.0）から変更されていない。特に、BCCWJ-DVD 版（Version 1.1）で加えられた文関連の修正が適用されていないため、次の点において、M-XML （第 9 章参照）と内容的に相違が生じている。利用する際は、注意されたい。

- C-XML では、文認定基準は Version 1.0 と同一であり、新規に追加された文認定基準が適用されていない。
- C-XML では、Version 1.0 と同様、文区切りを自動的に行っており、人手修正を行っていない。
- C-XML では、文認定の人手修正に伴い発見された文書構造タグの誤りが修正されていない（Version1.1 の M-XML では修正済み。8.3.1 節参照）。

参考文献

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる （2011） 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書「『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.2.2」

西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也 （2011） 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書「『現代日本語書き言葉均衡コーパス』における電子化テキストの構築」

田中牧郎・相澤正夫・斎藤達哉・棚橋尚子・近藤明日子・河内昭浩・鈴木一史・平山允子（2011）特定領域研究「日本語コーパス」平成 22 年度研究成果報告書「言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用」

日本工業標準調査会（2000）『7 ビット及び 8 ビットの 2 バイト情報交換用符号化拡張漢字集合 JIS X 0213:2000』日本規格協会.