『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

国立国語研究所内部報告書（LR-CCG-10-04）

『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

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

平成23年2月

国立国語研究所内部報告書（LR-CCG-10-04）

『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

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

平成２３年２月

序

本報告書は,「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese, 以後, “BCCWJ” と表記) における電子化フォーマットについてまとめたものである。BCCWJ は，2006 年度から構築を開始した日本語のコーパスである。規模は 1 億語，応用対象は言語学，国語教育，日本語教育，辞書編集，自然言語処理などの分野であり，収録対象の資料も書籍，新聞，雑誌などと幅広い。本報告書で記述する電子化フォーマットは，このような背景と応用を考慮しつつ，主として紙媒体の資料を電子テキストに変換する際の方式を定義するものである。

本報告書を作成するまでの過程は，次のとおりである。まず，2005 年度に BCCWJ の設計を行うための小規模な「パイロットコーパス」を構築した。この際，本電子化形式の基本となる仕様を設計し，パイロットコーパスに適用した。その後，パイロットコーパスの評価・検証を経て，電子化形式の見直しを行った。

2006 年度から BCCWJ の構築が開始され，この間，数回の改定を経て，現在に至っている。これまでに，書籍，新聞，雑誌，国会議事録，法律，白書に対して本電子化形式を適用した。本報告書に示す電子化仕様は，現時点での最新版である。

本報告書の執筆・編集は，山口昌也，高田智和，北村雅則，間淵洋子，大島 一，小林正行，西部みちるが共同で行った。また，電子化フォーマットの設計に関しては，上記のメンバーの他，2005 年度に田中牧郎，柏野和佳子が係わっている。

目次

第 1 章

電子化形式の概要

1.1 はじめに

本章では，BCCWJ における電子化フォーマットの概要について述べる。

本電子化フォーマットは，BCCWJ のサンプリング基準によりサンプリングされた原資料を電子テキストに変換する際の形式を定めるものである。BCCWJ に収録される電子化テキストには，原資料に陽に記述されているテキストのほかに，書誌情報，文書構造情報，文字情報といった，さまざまな情報が XML のタグにより付与される。したがって，本電子化フォーマットが規定するのは，テキストの符号化形式，および，付与情報の記述形式ということになる。

本報告書では，２章でテキストの符号化形式について，３章で付与情報を記述するために利用する XML タグの仕様について詳しく説明する。

本電子化フォーマットが記述対象として想定するテキスト，および，電子化されたテキストの利用分野は，次に示すとおりである。これらは，BCCWJ と同一である。

- 記述対象として想定するテキストは，現代日本語の書き言葉とし，1976 年以降の (主として) 出版物を対象とする。実際に想定しているのは，書籍，新聞，雑誌，白書，教科書，議事録，Web データなどである。
- 利用分野としては，言語学，国語教育，日本語教育，辞書編集，自然言語処理など幅広い分野を想定する。

この後の節では，次の順序で，電子化フォーマットの概要を説明していくことにする。まず，次節で電子化フォーマットに対する要求分析を行い，その結果に基づいて，設計方針を決定する。次に，1.3 節で電子化フォーマットの仕様を規定するための XML タグセットを示す。そして，最後に 1.4 節で本章のまとめを述べる。

1.2 電子化フォーマットの設計

1.2.1 電子化フォーマットに対する要求

ここでは，電子化フォーマットの仕様として，何が必要なのかを明確にするために，電子化するテキストの種類，利用方法，コーパスの規模，作成方法という四つの観点から，電子化フォーマットに対する要求分析を行う。

まず，電子化するテキストの種類の観点から要求を考える。BCCWJ の収録対象となる資料としては，書籍，雑誌，新聞，白書，教科書，議事録，Web データ (Yahoo!知恵袋[*1]を予定) などが想定されている (山崎

[*1] 利用者参加型の質問サイト。http://chiebukuro.yahoo.co.jp/

他 2006)。したがって，多様な文書構造を持ったテキストを扱う必要がある。例えば，小説のように，文書の階層構造が単純な資料もあれば，白書のように非常に深い階層構造を持った文書もある。さらに，雑誌の中には，図が多用され，レイアウトが複雑で，文書構造が不明確なものもある。このような文書構造上の多様性に加えて，テキストの特性や利用目的を活かすために，利用目的に特化した情報を付与しなければならないものもある。例えば，非母集団 (特定目的) サブコーパスのテキストは，個別の利用目的に対応できるような情報付与が必要になるだろう。以上のことから，次の要求を挙げる。

要求 1 多様な文書形式に対応できるようにすること

要求 2 利用目的に特化した情報付与に対応できるようにすること

次に，想定される利用方法を見てみよう。BCCWJ の利用分野としては，日本語学，日本語教育，国語教育，辞書編纂，自然言語処理などが挙げられている (山崎他 2006)。まず，すべての利用分野に共通して必要なことは，(1) テキストの文字が適切に符号化されていること，(2) 文字，文法，語彙，文体など言語学的な分析に役立つ文書要素に対して，適切にマークアップがなされ，容易に検索できることである。また，辞書編集のための用例収集のように，実際の用例を検索し，それを人間が詳細に分析するといった用途には，用例を理解しやすい形式で表示するための情報が付与されていることが望ましい。さらに，自然言語処理など，工学的な利用を考慮すると，汎用のツールで処理したり，他の言語資源と連係して利用できることが求められる。以上をまとめると，次のようになる。

要求 3 テキストを正確に符号化できること

要求 4 言語学的な分析に役立つ文書要素が適切にマークアップできること

要求 5 計算機処理に適した形式であること

要求 6 利用者が理解しやすい形式で電子化テキストを閲覧できること

要求 7 他の電子化フォーマットとの連係が取りやすいこと

最後に，コーパスの規模と作成方法の観点から考察する。まず，コーパスの規模は 1 億語で，開発期間は 5 年間と予定されている。また，電子化テキストの作成に際しては，Web データや議事録などの一部を除き，紙媒体からの入力を行う。これは，情報付与がまったくされていない状態から電子化することを意味し，(テキストの著者ではない) コーパスの作成者がテキストを解釈した上で，情報付与を行うことになる。したがって，本電子化フォーマットの利用者，つまり，コーパスの作成者とコーパスの利用者の共通理解を得やすいマークアップが必要であると考える。また，人手によるマークアップを行うことが予想されるため，量的にマークアップすることが可能な付与情報かどうかを考慮することも必要である。そこで，次の二つの要求を掲げる。

要求 8 コーパス作成者，コーパス利用者の共通認識を得やすいマークアップであること

要求 9 人手で構築するのに，実現可能な量の付与情報であること

1.2.2　設計方針

前節で示した電子化フォーマットに対する要求のうち，電子化するテキスト，利用分野，利用者の多様性を鑑み，次の設計方針を立てた。

- 言語学，国語教育，日本語教育，辞書編集，自然言語処理などの幅広い分野への応用を想定した設計にする
- シンプルで，拡張性を考慮した仕様となるように設計する

これらの設計方針の下で，前節に示した要求に対して，次のように対処する。

- 文書中の論理的な役割が明確であり，かつ，紙面上の物理的な構造が明確な文書要素をマークアップの対象とする。

  - 二つの基準により文書構造が認定されるので，コーパス作成者・利用者の両者にとって共通理解を得やすい情報付与が可能になると考えられる ([**要求 8**])。また，論理的な役割が明確な文書要素がマークアップされるので，言語学的な分析に役立つ文書要素が適切にマークアップされることが期待できる ([**要求 4**])。
  - 論理的な構造ごとに閲覧時の表示形式を工夫し，電子テキストを利用者が理解しやすい形式で表示する (例えば，タイトルとしてマークアップされている場合は，フォントサイズを大きくするなど)。([**要求 6**])
- 収録対象の資料に含まれる文字を記述するのに十分な文字規格を採用する。また，ルビ，外字など，文字・表記に関するタグを用意する。 ([**要求 3**])
- 文書記述言語として，XML (eXtensible Markup Language) を用いる。XML は拡張性に優れた文書記述言語であり，多様な文書形式や利用目的に特化した情報付与に対応しやすい ([**要求 1,2**])。また，TEI (Text Encoding Initiative) をはじめとして，多くのコーパスや電子化フォーマットで採用されており，『太陽コーパス』(国立国語研究所 2005) や『日本語話し言葉コーパス』(国立国語研究所 2006) も XML を用いて記述されている。したがって，これらのデータとの整合性も高い。また，XML は，コーパスの記述だけでなく，データ一般の記述に広く用いられており，データ形式の検証，変換，検索などを行う際に，既存のツールを利用できるという利点もある。([**要求 5,7**])
- 量的な観点から，人手でマークアップすることが困難な場合は，自動的，もしくは，半自動的なマークアップを検討する。([**要求 9**])

1.3 電子化フォーマットの仕様

1.3.1 概要

本電子化フォーマットの概要は，次のとおりである。

文書記述言語： XML
文字符号化方式： UTF-16
文字集合： JISX0213:2004

BCCWJ の電子化テキストは XML で記述する。電子化フォーマットは，XML の文書型によって規定する。BCCWJ には，一つのサンプルが一つの「記事」に相当する可変長サンプルと，一つのサンプルに 1000 文字を包含する固定長サンプルがある。したがって，2 種類の文書型を定義する。

文字符号化方式は UTF-16 を，文字集合には JISX0213:2004 を採用した。JISX0213:2004 に含まれる文字数は，約 11000 字である。JISX0213:2004 には，現在最も一般的に利用されている JISX0208 の約 6800 字に，第 3，4 水準漢字・非漢字，約 4000 字が追加されている。

JISX0208 ではなく，JISX0213:2004 を採用したのは，(a) 現時点の国内規格では，最も大きな文字集合を持つこと，(b) 印刷字体を考慮した包接基準を持つこと，(c) 他のコーパスとの連係を考慮したこと，などが挙げられる。(a)(b) は，正確な文字の符号化に寄与すると期待される。(c) の例としては，BCCWJ に収録されているものよりも古い時代の資料[*2] や，今後発展の見込まれる電子データ[*3]がある。詳細については，2 章を参照のこと。

[*2] 例えば，『太陽コーパス』

[*3] PC 用の OS として現在最も普及している Windows の新バージョン (Windows Vista) も JISX0213 を採用しているため，JISX0213 で符号化したデータが流通する可能性がある

1.3.2　タグの仕様

本電子化フォーマットでは，46 種類の XML タグを定義した。タグの一覧を表 1.1 に示す (スペースの関係上，一部のみ)。また，本電子化フォーマットで電子化テキストに変換した例を図 1.1 に示す。本電子化フォーマットで定義する XML タグによって付与される情報は，次の三つに大別される。詳細な仕様については，3 章を参照されたい。

- サンプルに関するタグ
- 文字・表記に関するタグ
- 文書構造に関するタグ

1.3.2.1　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 (表 1.1 参照) がある。 sample 要素[*4]は，一つのサンプルを表す。sampling タグは，サンプル抽出基準点 (丸山岳彦，柏野和佳子，山崎誠他 2007) などサンプリングに関する情報を表す。

sample タグには，サンプルに関する情報が属性として記述されている。sampleID 属性値は，サンプル固有の識別番号である。サンプルの書誌情報は，sampleID をキーとして，書誌情報のデータベースを参照する。書誌情報としては，書名，著者，出版社などが提供される予定である。

sample タグの type 属性は，サンプルの種別 (固定長，可変長) を表す。図 1 では，type 属性が “variableLength” となっているので，可変長のサンプルであることがわかる。一方，固定長の場合は，属性値が “fixedLenghth” となる。

1.3.2.2　文字・表記に関するタグ

文字・表記に関するタグの役割は，二つある。一つは，検索や計算機処理の利便性を高めることである。この役割を持つタグに correction タグがある。このタグは，原文の誤植を訂正した文字であることを表す。次の例は，誤字，脱字，衍字を修正した例である。修正した結果がテキスト本文になり，修正に関する情報はタグによって表現されるので，誤りを意識せずに，検索したり，計算機処理を行うことができる。修正前の文字は，originalText 属性として保持される。

```
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
```

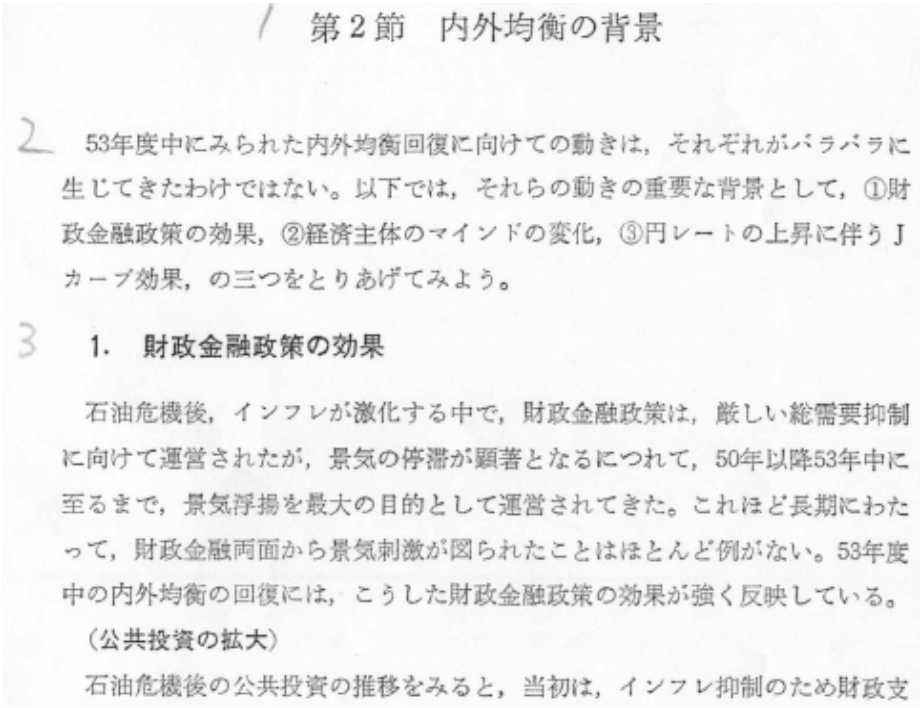
もう一つの役割は，原資料に忠実に電子化テキストを記述することである。この役割を持つタグの例として，ruby，missingCharacter タグの例を次に示す。ruby タグはルビ付き文字を表す。JISX0213:2004 で規定されていない文字は〓で代替され，missingCharacter でマークアップされる。missingCharacter タグは，属性として，文字種を表す attribute 属性，Unicode 番号を保持する unicode 属性，『大漢和辞典』の親字番号を表す daikanwa 属性，字体記述を行うための description 属性などを持つ。

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

[*4] sample タグでマークアップされている文書要素

表 1.1 タグ一覧 (一部)

	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	title 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
	list	箇条書きなど，列挙された文書要素の集まり
	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)	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
注記 (文書構造)	note	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
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
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
	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
	cursive	変体仮名
	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	著作権者の依頼などを受けて削除した本文要素
	br	物理改行
	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



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-16" ?>
<?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 1.1　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)

1.3.2.3 文書構造に関するタグ

文書構造に関するタグは，論理的な役割が明確な文書要素に対して付与される。表 1.1 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。本節では，このうち階層構造に関するタグを中心に説明する。

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 1.1 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 1.1 の article 要素直下の階層には，titleBlock と cluster 要素があることがわかる。

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

■ **article**　article 要素は「記事」を想定した要素で，「同一著者による，同一テーマのひとまとまりの文書要素」を表す。なお，BCCWJ では，一つの article 要素に含まれる文字数の上限が約 1 万字ということになっているため，必ずしも，「同一著者による，同一テーマのひとまとまりの文書要素」すべてを収録できるとは限らない。例えば，図 1.1 の白書のサンプルは，1 章 2 節だけしか収録していない。このような場合，「記事」全体を収録できたか否かを表す isWholeArticle 属性は，“false” となる。

■ **cluster**　cluster 要素は，章，節といったように，タイトル (titleBlock 要素) を持った，ひとまとまりの文書要素を表す。cluster 要素自体には，章，節といった特定の階層を表すための意味づけを行っていないが，入れ子構造により，階層の上下を表す。例えば，上記の例の「(公共投資の拡大)」というタイトルを持つ cluster 要素は，2.1 節に対応する cluster 要素の子要素なることで，2.1 節の下位構造であることを表現する。なお，cluster には必ず titleBlock が含まれる。この制約を課すことにより，紙面上のデザインなどの物理的な特徴に基づいて，cluster が過度に認定されるのを防ぐことができる。

■ **titleBlock**　すでに述べたように，titleBlock 要素は，cluster 要素のタイトルとそれに付随する部分からなる文書要素である。タイトルとその付随部分は，title 要素により，明示的にマークアップされているので，容易にタイトルだけを検索したり，抽出したりすることが可能である。

■ **paragraph, sentence**　それぞれ，段落，文に相当する要素である。これらの要素は，テキスト中に大量に含まれるため，人手でタグを付与することは困難である。そこで，paragraph は行頭の空白，sentence は句点などを手がかりに，自動的にタグを付与している。

1.3.3 他の電子化フォーマットとの関係

テキストを電子的に記述するための形式としては，従来から，TEI や CES (Corpus Encoding Standard) などが提案されている。BCCWJ で新たに電子化フォーマットを策定したのは，次の理由による。まず，TEI は，汎用の電子化フォーマットであるため，仕様が複雑であり，BCCWJ の規模，実施期間を考慮すると，実際に実装するのは困難である。一方，CES は TEI よりもシンプルな仕様であるが，適用範囲として，言語工学やその応用を指向しており，言語学的な分析と工学的な利用の双方を視野に入れた BCCWJ に CES をそのまま適用することは難しい。

それに対して，BCCWJ の電子化フォーマットは，言語学から工学という多様な利用分野を想定しつつ，記述対象のテキストを現代日本語の書き言葉に限定することにより，シンプルで，実際に運用可能なフォーマットを実現するものである。

1.4　おわりに

本章では，BCCWJ における電子化フォーマットの仕様について概要を説明した。我々は本仕様に基づいて，これまで，白書のサンプル (1500 サンプル)，書籍 (約 8000 サンプル)，新聞 (540 サンプル) を電子テキストに変換した。今後，雑誌など，これまで扱ってこなかった種類の資料に対して，本電子化フォーマットを適用するために，随時，仕様を修正・拡張していくことが予想される。本仕様は，Web 上[*5] で一般に公開しているので，最新の情報については，そちらを参照していただきたい。

参考文献

Text Encoding Initiative, The XML Version of the TEI Guidelines, http://www.tei-c.org/P4X/index.html

Corpus Encoding Standard, http://www.cs.vassar.edu/CES/

山崎 誠，丸山岳彦，柏野和佳子 他 (2006) 「現代書き言葉均衡コーパスの現状」，特定領域「日本語コーパス」平成 18 年度全体会議予稿集，pp.9–16

丸山岳彦，柏野和佳子，山崎誠 他 (2007) 「「現代日本語書き言葉均衡コーパス」におけるサンプリングの概要」，「日本語コーパス」平成 18 年度公開ワークショップ予稿集

国立国語研究所 (2005) 『太陽コーパス』(国語研究所資料集 15)，博文館新社

国立国語研究所 (2006) 『日本語話し言葉コーパスの構築』(国語研究所報告書 124)，国立国語研究所

[*5] http://www2.ninjal.ac.jp/densi/public/wiki/

第 2 章

文字入力仕様

本章では，BCCWJ に収録するデータを作成する際の文字入力に関する仕様について述べる。まず始めに，基本仕様を概観した上で，版面に現れる様々な形態の文字を入力する方法について，例示を交えて，具体的に説明する。

2.1 基本仕様

文字入力は，以下の基本方針に基づき行なう。

- 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。

2.2 文字コードと改行コード

文字コードは，Unicode（UTF16LE：Byte Order Mark 付き）を用いる。
また，改行コードは，LF を用いる。

2.3 文字集合

文字集合は，JIS X0213:2004 規格[*1]（日本工業標準調査会 (2004) を参照。以下，改定情報などに言及する必要がない限り，「JIS X0213」または単に「JIS 規格」と呼ぶ）に準拠した独自の文字集合，10,956 文字を用いる。

BCCWJ の文字集合は，JIS X0213 のそれと完全には一致しない。その理由は，以下 2 点による。

(a) コーパスの仕様による制限

(b) データ作成に用いる処理系による制限 [*2]

以下の節で，それぞれについて詳細を示す。

[*1] いわゆる JIS 第 4 水準までの全ての漢字と非漢字を含む 11,233 文字の符号化を規定した JIS の最新規格。

[*2] 今後，文字処理環境の変化などによって，制限が解消される可能性もあるため，それに伴い，本章において示す現在規定の文字集合を，変更する可能性がある。

2.3.1　コーパスの仕様による制限

BCCWJ は，言語研究用のコーパスであるという性質から，言葉をデータ化の対象としている。また，文字の，版面に現われる図形としての側面より，言葉の構成要素としての側面を重視してデータ化する立場を取る。そのため，以下に挙げるものについて，JIS X0213 の文字集合とのずれが生じている。

(1) 入力対象外要素を構成する文字
(2) 装飾・デザインにかかわる文字
(3) 類似の非漢字
(4) 合成文字

2.3.1.1　入力対象外要素を構成する文字

以下の文字は，入力対象外の要素を構成する文字であるため，使用しない。

ソフトハイフン　BCCWJ は，レイアウト上の情報を反映させないため，版面上の改行（行の折り返し）の配慮は必要がない。よって，ハイフネーション（行末の単語内で改行が起こる際に，単語の前半と後半をハイフンで繋いで表示する機能）の結果表示されるハイフンは，電子テキストに反映させる必要はないため，入力対象外となる。

ソフトハイフン（面区点：1-09-09）は，ハイフネーションに用いることが想定される文字であるため，使用しない（→リスト：2.10.1.3「改行関連文字」）。

けい線素片　BCCWJ では，図を入力対象としない。また，文字や文章の囲みなどについても，レイアウト上の情報とみなし，入力対象としない。

JIS X0213 に規定される 47 字のけい線素片は，図形や囲みの記述に用いることが想定される文字であるため，使用しない（→リスト：2.10.1.3「けい線素片」）。

2.3.1.2　装飾・デザインにかかわる文字

同一の機能を持つ文字は，デザインの差や JIS 規格への収録の有無によらず，統一的にデータ化する。装飾・デザインの施された文字は，装飾・デザインを無視して扱う。よって，装飾・デザインにかかわる以下の文字は使用しない。

組み文字　組み文字（複数文字を 1 文字分のスペースに組んだ形で表した文字）は，組まれている文字を全て 1 字ずつ切り離して入力する（→「2.6.4 組み文字」）。例えば，「㍍」は「メートル」と入力される。JIS X0213 には，その他「㈱」「㍾」「㎞」「⁉」等，33 文字の組み文字が定義されているが，一切使用しない（→リスト：2.10.1.1 「組み文字」）。

分数　分数は，分子と分母を「／」で区切って入力する（→「2.6.5 分数」）。例えば「½」は「１／２」と入力される。JIS X0213 には，「½」をはじめ，6 文字の分数が定義されているが，一切使用しない（→リスト：2.10.1.1「組み文字」）。

11 以上のローマ数字　ローマ字は 1〜10 までの文字のみを用いる。それ以上の数は，これらの組み合わせと考えて，切り離して入力する。例えば，「Ⅺ」は「Ｘ Ｉ」によって入力される。この規準に基づき，JIS X0213 に定義されている，ローマ数字 11，12 とそれに対応する小文字は使用しない（→リスト：2.10.1.1「組み文字」）。

囲み文字　囲み文字は，囲まれている文字を入力する（→「2.6.3 囲み文字」）。例えば「①」は，「１」によって入力される。JIS X0213 には，「①」「❶」「ⓐ」「㋐」「㊤」など 136 文字の囲み文字が定義されているが，一切使用しない（→リスト：2.10.1.1「囲み文字」）。

上付き文字 上付き文字は，通常の算用数字などによって入力する（→「2.6.2 上付き・下付き文字」）。例えば，「² 」は，「2」によって入力する。JIS X0213 には，1〜3 の上付き文字が定義されているが，一切使用しない（→リスト：2.10.1.1 「上付き文字」）。

2.3.1.3 類似の非漢字

非漢字においては，細微な形態の差が言葉の意味の差に直接かかわらないと判断される場合，その差異を無視する。JIS X0213 に定義される文字についても，独自の規準によって別の文字と同一視（包摂）して扱う場合がある。その結果，別のある文字に包摂される文字は使用しない（→リスト：2.10.1.2）。

非漢字の独自包摂規準については，「2.4.2 類似記号の独自包摂と意味による使い分け」において述べる。

2.3.1.4 合成文字

JIS X0213 においては単独の文字として定義されているものの，本仕様で符号化に用いる Unicode においては合成によって表現しなければならない文字がある。「ゕ」等一部の半濁点付き仮名，「æ 」等一部の音声記号付きラテン文字，および，「/ 」等一部の声調記号，の合計 25 文字がこれに相当するが，本仕様においては文字合成を行わないため，一切使用しない（→リスト：2.10.1.4「Unicode における合成処理対象文字」）。

これらは，入力可能な代用文字によって入力する。例えば，「ゕ」などの半濁点付き仮名は，半濁点を除いた「か」によって入力される。[*3] 。

2.3.2 データ作成に用いる処理系による制限

データ作成に用いる処理系[*4] によって入力が制限される以下の文字については，代用文字を入力する。

口偏＋「七」（叱） 2004 年の JIS 規格改訂時に追加された 10 字のうち，口偏に「七」の文字（「叱」面区点：1-47-52）は，対応する Unicode（U+20B9F）が，現状の処理系で扱うことができないため，「叱」に独自に包摂する[*5] 。

2.4 包摂規準

2.4.1 漢字の字体包摂

漢字における字体包摂は，JIS X0213 に準拠する。JIS X0213:2000「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。

2.4.2 類似記号の独自包摂と意味による使い分け

非漢字のうち記号類については，独自の包摂規準を設ける。

2.4.2.1 JIS X0213 に定義されていない記号

JIS X0213 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。

[*3] この際，原文の文字情報を以下のように XML タグによって示す。

```
<substitution x0213="1-04-87" unicode="304B,309A">か</substitution>
```

[*4] システム：Microsoft Windows XP, エディタ：Meadow2.0

[*5] ただし，原文の文字情報を以下のように XML タグによって示す。

```
<substitution x0213="1-47-52" unicode="20B9F">叱</substitution>
```

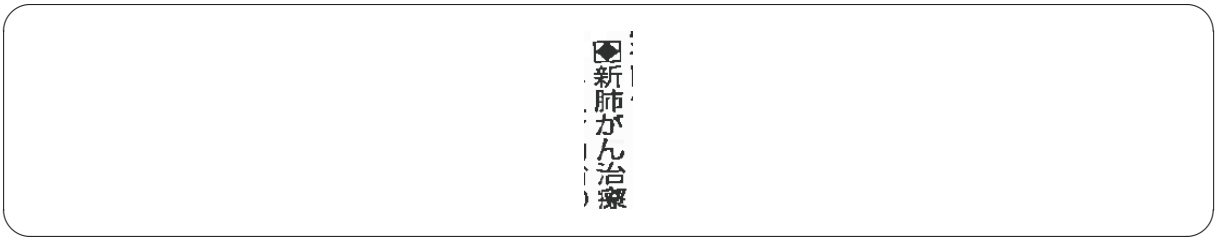


図 2.1　類似の規格内記号に包摂する例（『**毎日新聞**』2002 **年** 6 **月** 13 **日朝刊**）

図 2.1 の入力例

◆新肺がん治療

2.4.2.2　JIS X0213 に定義されている記号

JIS X0213 では，字形の類似した複数の記号類が別字として詳細に分けられている。しかし，これらは，紙面上の見え方でどの文字かが判別できないことも少なくなく，これらを統一的に判別し詳細に入力し分けるのは非常に困難である。また，形態よりも意味によって統一的に入力されている方が，言語研究に用いる際には望ましい。

そこで，類似記号に関しては，形状ではなく用法によって統一的に電子化することを方針として，独自の包摂規準を設けた。例えば，中央位置の横線に類する形状の文字は，JIS X0213 に 9 字が定義されているが，BCCWJ の文字入力においては，意味的な使い分けが必要かつ可能な 4 字に限定して，使用可能な文字と規定した。その上で，原資料における紙面上の形状ではなく，意味によって，4 字のいずれかを判断し入力し分ける方法を取った[*6] 。

表 2.1　線形類似記号一覧

文字	文字名	JIS 面区点	使用条件
ー	長音記号	1-01-28	長音として用いる場合に使用
－	負記号、減算記号	1-01-61	数式等でマイナスの意で用いられる場合に使用
—	ダッシュ（全角）	1-01-29	範囲・経過，引用・挿入句・余韻，項立てなどを示す場合に使用
-	ハイフン（四分）	1-01-30	文節表示・単語連結・英数字連結，住所や電話番号等の区切りの場合に使用
-	ハイフンマイナス	1-02-17	使用不可（独自包摂）
–	二分ダーシ	1-03-92	使用不可（独自包摂）
–	ソフトハイフン	1-09-09	使用不可（入力対象外）
—	横細線素片	1-08-01	使用不可（入力対象外）
―	横太線素片	1-08-12	使用不可（入力対象外）

入力例（右側に×として示したのは，誤った入力例）：

データ	×データ, デ―タ, デ - タ
－２℃	×ー２℃, - ２℃, —２℃
１９７６—２００５	×１９７６ー２００５, １９７６－２００５, １９７６ - ２００５
ＣＤ - ＲＯＭ	×ＣＤーＲＯＭ, ＣＤ－ＲＯＭ, ＣＤ—ＲＯＭ

上記方針に基づき設けた，類似記号の独自包摂と意味による使い分けを次節に一覧にして示す。

[*6] 以下，表内の【文字名】には JIS X0213:2000 附属書 4 表 1〜24 の【日本語通用名称】を用いた

2.4.2.3 類似記号の包摂一覧

ひとつの文字に包摂されるもの 表 2.2 に示した文字は，いずれも近似した別の文字に置き換えても，原文の意味を損なわないとみなし，「代用字」に示した文字に包摂して入力することとする。

表 2.2 類似記号包摂一覧

面区点	文字	UCS	文字名	代用字	面区点	UCS	文字名
1-09-02		00A0	ノーブレークスペース		1-01-01	3000	和字間隔
1-03-92	–	2013	二分ダーシ	—	1-01-29（代用）	2015	ダッシュ (HORIZONTAL BAR)
1-03-91	＝	30A0	二重ハイフン	=	1-01-65（互換）	FF1D	等 号 (FULLWIDTH EQUALS SIGN)
1-09-08	«	00AB	始め二重山括弧引用記号	《	1-01-52	300A	始め二重山括弧
1-09-18	»	00BB	終わり二重山括弧引用記号	》	1-01-53	300B	終わり二重山括弧
1-01-17	￣	203E	オーバーライン	￣	1-09-11（互換）	FFE3	マクロン (FULLWIDTH MACRON)
1-02-18	~	007E	チルド				
1-09-14	·	00B7	中点（ラテン）	・	1-01-06	30FB	中点
1-03-32	•	2022	ビュレット				
1-03-31	◦	25E6	白ビュレット	○	1-01-91	25CB	丸印、白丸
1-02-94	○	25EF	大きな丸				
1-03-26	⦿	29BF	丸中黒	◉	1-03-27	25C9	蛇の目
1-13-64	〝	301D	始めダブルミニュート	“	1-01-40	201C	左ダブル引用符
1-13-65	〟	301F	終わりダブルミニュート	”	1-01-41	201D	右ダブル引用符

意味によって，置き換えるべき文字が複数あるもの 表 2.3 に示す 3 文字は，ISO/IEC646(ASCII) などの 1 バイト文字コードとの互換性を図るために，JIS X0213 に新たに収録されたものである。1 バイトコードで“複数の文字の代替として使用できること”を想定するこれらの文字は，全角文字のみを用い，用法による記号の独自包摂を行う BCCWJ の文字入力仕様の下では，他の文字に吸収されるものであり，必要とならない。従って，それぞれ，用法によって別の文字に包摂する。

表 2.3 1 バイト文字コード互換用文字の置換処理

包摂字	文字名	面区点	使用字	文字名	面区点	使用条件
"	引用符	1-02-16	”	右ダブル引用符	1-01-41	引用表現の終端を示す場合
			“	左ダブル引用符	1-01-40	引用表現の始端を示す場合
			¨	ウムラウト	1-01-15	ウムラウトとして用いる場合
'	アポストロフィ	1-02-15	’	右引用符	1-01-39	引用表現の終端を示す場合縮約形や所有格を示す場合
			‘	左引用符	1-01-38	引用表現の始端を示す場合
			´	アクサンテギュ	1-01-38	アクサンテギュ，プライム記号として用いる場合
-	ハイフンマイナス	1-02-17	-	ハイフン	1-01-30	表 2.1 参照
			−	負記号	1-01-61	表 2.1 参照

2.4.2.4　類似記号の使い分け

使用文字を限定した上で，なお類似文字の組がある場合は，表 2.4 に示す規準によって文字を使い分けることとする。

なお，この使い分け規準を適用した際に，入力すべき文字と紙面に図形として現れた原文の文字が異なる場合があるが（例えば，原文で長音の表記にダッシュ「—」が用いられている場合に，これを「ー」に置き換えて入力した場合），これらは誤植とみなさない。

2.5　外字

上記に示した BCCWJ の文字セットで転記できない文字は，以下のように処理する。

2.5.1　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。

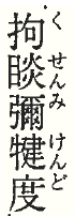


図 2.2　漢字外字の例（**松久保秀胤『安らぎを求めて』**）

図 2.2 の入力例

```
拘〓彌犍度
```

タグ付き入力例

```
拘<missingCharacter attribute="HanIdeograph" unicode="U+7752" daikanwa="M23412"
description="目偏に炎">〓</missingCharacter>彌犍度
```

2.5.2　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。

図 2.3　記号外字の例（『**国民生活白書 平成 2 年版**』）

表 2.4 類似記号の使い分け一覧

使い分け対象	文字	文字名	面区点	使用条件
ー - ――	ー	長音記号	1-01-28	長音を示す場合
	-	ハイフン（四分）	1-01-30	①英数字と共に項目名（例：A - 1） ②文節表示や単語連結 ③住所（郵便番号・番地）や電話番号等の区切り
	−	負記号、減算記号	1-01-61	数式 (マイナスの意) のみ
	―	ダッシュ (全角)	1-01-29	①数, 時間, 月日などの範囲や, 場所の経過を示す ②引用・挿入句・余韻などを示す ③項立て, 質問者の発言などを示す
“” ″	“	左ダブル引用符	1-01-40	始め引用符
	”	右ダブル引用符	1-01-41	終わり引用符
	″	秒	1-01-77	秒を表わす場合のみ
‘’	‘	左引用符	1-01-38	始め引用符
	’	右引用符	1-01-39	①終わり引用符 ②アポストロフィの代用
〈〉 <>	〈〉	始め/終わり山括弧	1-01-50, 1-01-51	括弧・引用符
	<>	不等号 (より小/大)	1-01-67, 1-01-68	①数式 (不等号) ② Web, e-mail 上の引用符 (「>」のみ)
《》 ≪≫	《》	始め/終わり二重山括弧	1-01-52, 1-01-53	括弧・引用符
	≪≫	非常に小さい/大きい	1-02-67, 1-02-68	数式 (不等号) のみ
〓=	〓	げた記号	1-02-14	①外字 ②「〓」文字そのもの
	=	等号	1-01-65	①数式で使用 ②二重ハイフン (欧米人の姓名をカタカナ書きする場合の区切り) の代用
＃ #	＃	シャープ	1-02-84	音楽記号のみ
	#	番号記号，井げた	1-01-84	項立て
× X x	×	乗算記号	1-01-63	①数式 ②「バツ」を表わす
	X x	ラテン大/小文字X	1-03-56, 1-03-88	ローマ字
Φ ϕ ∅	Φ ϕ	ギリシア大/小文字 PHI	1-06-21, 1-06-53	①ギリシア文字 ②直径を表わす
	∅	空集合	1-02-39	数学・論理学での空集合

図 2.3 の入力例

甲種電気用品（２８２品目マーク）

タグ付き入力例

甲種電気用品（２８２品目<image no="1" description="甲種電気用品マーク" />マーク）

2.6　特殊表記

2.6.1　ルビ

行間に小書きされた振り仮名（＝ルビ）が付けられている場合は，これを入力対象とする。ただし，本文行の文字列と区別しておく必要があるため，ルビの付与されている文字列を，ruby タグによって示した上で，タグ内部に rubyText 属性として入力する。

ルビの文字列には，通常小書きする促音・拗音などを，小書きせずに通常の直音文字によって示しているものがあるが，紙面での表記にかかわらず，拗音・促音は小書きの仮名を用いて入力する。

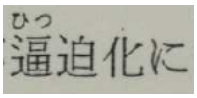


図 2.4　小書き仮名を用いていないルビの例（『**わが外交の近況 昭和** 58 **年版**』）

図 2.4 の入力例

<ruby rubyText="ひっ">逼</ruby>迫化に

なお，傍線，傍点，返り点などは全て入力対象外でありルビにならない。また，通常ルビが示される位置に小書きで添えられた，注番号や注記などは，ルビとして扱わない。これらは，それぞれ「注参照マーカー」（→ 2.6.6）「傍注」（→ 2.6.7）として別途記述する。

2.6.2　上付き・下付き文字

数式や化学式，単位表示等に見られる上付き文字，下付き文字は，全て通常の算用数字やアルファベットによって入力する。その上で，上付きになっている文字は，superScript タグ，下付きになっている文字は，subScript タグによって示す。

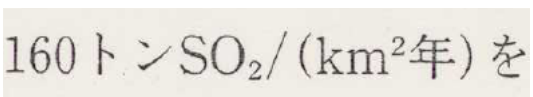


図 2.5　上付き・下付き文字の例（『**環境白書 平成** 7 **年版**』）

図 2.5 の入力例

１６０トンＳＯ２／（ｋｍ２年）を

タグ付き入力例

１６０トンＳＯ<subScript>２</subScript>／（ｋｍ<superScript>２</superScript>年）を

なお，上付き・下付きで示されるもののうち，数式や化学式，単位表示等に相当しない，注番号や注記などは，上付き・下付き文字としては扱わない。これらは，「注参照マーカー」「傍注」として別途記述する。

2.6.3 囲み文字

文字を丸や四角などの図形で囲んで示したものは，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。

図 2.6 略記号として機能する囲み文字の例（『**通商白書 昭和** 55 **年版 各論**』）

図 2.6 の入力例

優秀デザイン商品開発指導事業（優事業）を実施

タグ付き入力例

優秀デザイン商品開発指導事業（<enclosedCharacter>優</enclosedCharacter>事業）を実施

2.6.4 組み文字

複数の文字を 1 文字分のスペースに組んで示した文字は，組まれている文字を全て 1 字ずつ切り離して入力する。

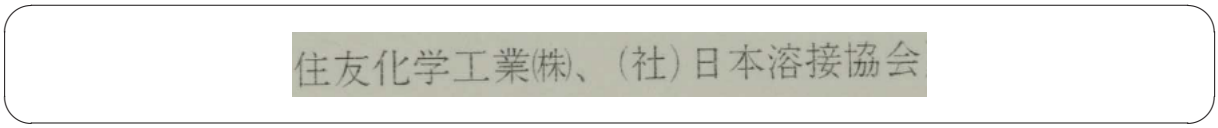


図 2.7 組み文字の例（『**交通安全白書 平成** 5 **年版**』）

図 2.7 の入力例

住友化学工業（株)、(社）日本溶接協会

2.6.5　分数

分数は，分子と分母が，横線を隔てて上下に配されている形式，斜線を隔てて左上・右下に配されている形式のいずれの場合も，「分子／分母」の形式に統一して入力する。

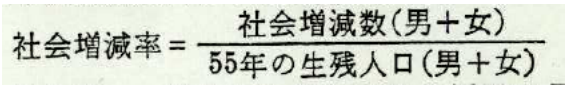


図 2.8　上下に組まれている分数の例（『**建設白書 昭和 57 年版**』）

図 2.8 の入力例

社会増減率＝社会増減数（男＋女）／５５年の生残人口（男＋女）

なお，「１½」のような帯分数の場合は，原文の意味を損なわないよう整数部分と分数部分を区別する必要があるため，分数部分に，fraction タグを付与して示す。

帯分数の入力例

１<fraction>１／２</fraction>

2.6.6　注参照マーカー

本文に対して脚注や巻末注などの参照事項がある場合に，本文の参照位置と注本文とを対応付けるための番号や記号など（以下，注参照マーカー）が，本文行から外れた位置（上付き，下付き，行間など）に示されていることがある。BCCWJ では，これも入力対象として扱う。入力位置は，注参照マーカーが付されている文字の直後とし，本文行の文字列と区別するために，noteMarker タグを付与して示す。

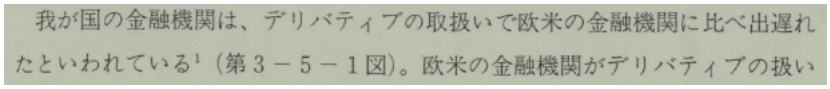

我が国の金融機関は、デリバティブの取扱いで欧米の金融機関に比べ出遅れたといわれている[1]（第３－５－１図)。欧米の金融機関がデリバティブの扱い

図 2.9　上付きで示された注参照マーカーの例（『**経済白書 平成 11 年版**』）

図 2.9 の入力例

我が国の金融機関は、デリバティブの取扱いで欧米の金融機関に比べ出遅れたといわれている１（第３‐５‐１図)。欧米の金融機関がデリバティブの扱い（後略）

タグ付き入力例

我が国の金融機関は、デリバティブの取扱いで欧米の金融機関に比べ出遅れたといわれている<noteMarker>1</noteMarker>（第３－５－１図）。欧米の金融機関がデリバティブの扱い（後略）

2.6.7 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場合（以下，傍注）は，これを入力対象として扱う。入力位置は，注記が付されている語や句の直後とし，本文行の文字列と区別するために，noteBodyInline タグを付与して示す。

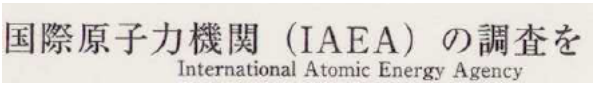


図 2.10 行間に示された注記の例（『**日本の防衛－防衛白書－ 平成 17 年版**）

図 2.10 の入力例

国際原子力機関（ＩＡＥＡ）Ｉｎｔｅｒｎａｔｉｏｎａｌ Ａｔｏｍｉｃ Ｅｎｅｒｇｙ Ａｇｅｎｃｙの調査

タグ付き入力例

国際原子力機関（ＩＡＥＡ）<noteBodyInline>Ｉｎｔｅｒｎａｔｉｏｎａｌ Ａｔｏｍｉｃ Ｅｎｅｒｇｙ Ａｇｅｎｃｙ</noteBodyInline>の調査

2.7 レイアウト

2.7.1 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

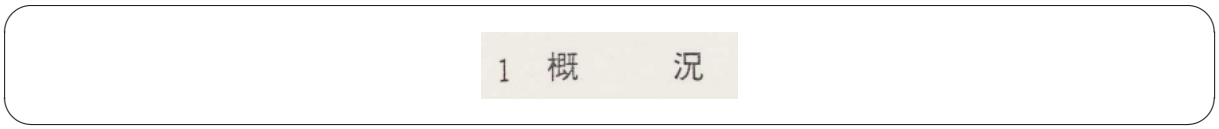


図 2.11　入力対象外となる文字幅調整用スペースの例（『**観光白書 昭和 63 年版**』）

図 2.11 の入力例

```
1　概況
```

2.7.2　改行

改行は，版面の行の折り返しではなく，論理行で行う。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

(1) 段落
(2) 引用
(3) 韻文における行

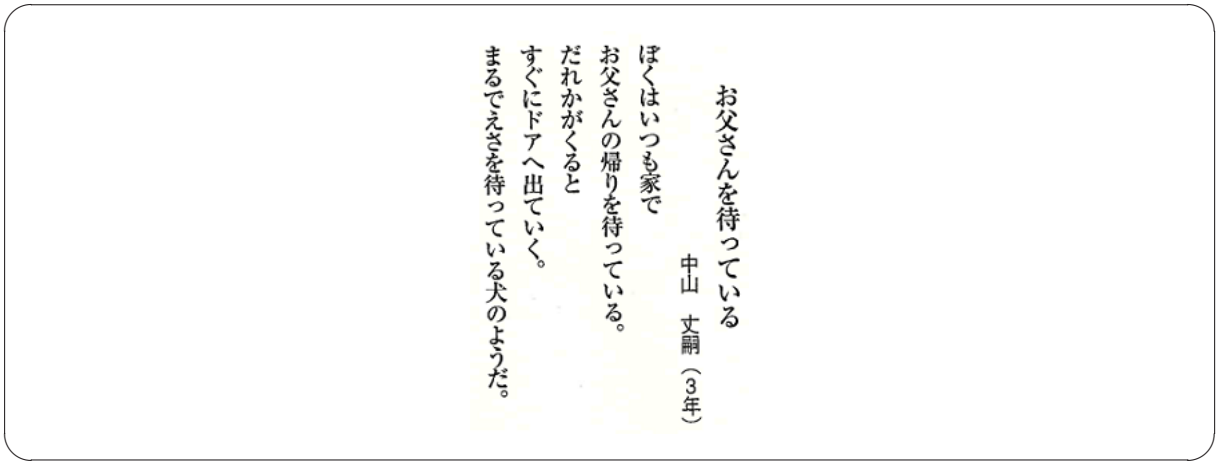


図 2.12　韻文における行の例（『**話を聞いてよ、お父さん！ 比べないでね、お母さん！**』）

図 2.12 の入力例

```
お父さんを待っている
中山丈嗣（３年）
ぼくはいつも家で
お父さんの帰りを待っている。
だれかがくると
すぐにドアへ出ていく。
まるでえさを待っている犬のようだ。
```

版面の行替えと一致しない場合でも改行するもの

(1) タイトル
(2) 表の各セル

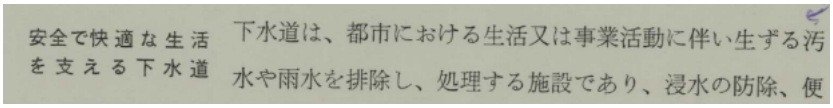


図 2.13　タイトルの例（『**建設白書 昭和 56 年版**』）

図 2.13 の入力例

```
安全で快適な生活を支える下水道
下水道は、都市における生活又は事業活動に伴い生ずる汚水や雨水を排除し、（後略）
```

2.7.3　リーダ・ダッシュ

リーダ・ダッシュは，全て原文に用いられているものを使用する。ただし，リーダ・ダッシュが複数連続するものについては，全て 1 字に置き換える。

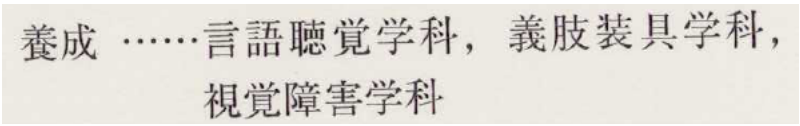


図 2.14　複数リーダの例（『**障害者白書 平成 13 年版**』）

図 2.14 の入力例

```
養成…言語聴覚学科，義肢装具学科，視覚障害学科
```

2.8　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。

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。

図 2.15 の入力例

```
総トン数１００トン未満で長さ３０メートル未満の
```

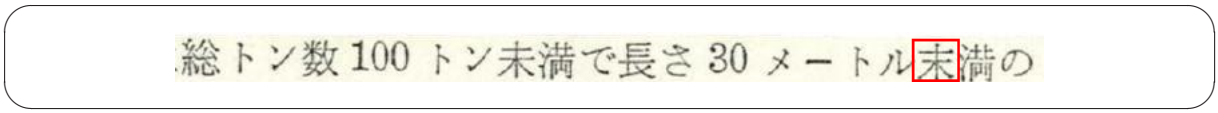


図 2.15　誤字の例（『運輸白書 昭和 51 年度版』）

タグ付き入力例

```
総トン数１００トン未満で長さ３０メートル<correction type="erratum" originalText="末">未
</correction>満の
```

2.9　符号化の実装

2.3 節において定義した BCCWJ の文字集合を，Unicode によって実装する際のコードマッピングは，原則として JIS X0213:2004 の規定によった。

ただし，処理系の制限により，一部の文字で改正前の JIS X0213:2000 の規定により実装した文字があるほか，符号化の実装において，注意を要する文字群があるため，本節で明示する。

2.9.1　JIS X0213:2000 の規定を適用するもの

2004 年の JIS 規格改訂において，対応する Unicode を変更した文字が 363 文字あったが，このうち，現状の処理系で，この変更に対応していない以下の 2 文字については，変更前の JIS X0213:2000 における定義（対応 Unicode）を採用する。

- 全角ダッシュ（「―」面区点：1-01-29）に対応する文字は，「U+2014 (EN DASH)」ではなく，「U+2015 (HORIZONTAL BAR)」を用いる。
- 「予」に「鳥」（「鵆」面区点：2-94-05）に対応する文字は，「U+29FCE」ではなく「U+29FD7」を用いる。

2.9.2　附属書 5 表 2 を適用するもの

算用数字，ラテン文字，その他ラテン文字用図形文字は，JIS X0213:2000「7.3 JIS X0201 のラテン文字と同時に用いる場合の符号」において許容されている，「附属書 5 表 2 数字・ラテン文字・特殊文字の大体名称」に規定される文字を専ら用いることとする。

これは，BCCWJ の本文文字入力には全角文字のみを用いる仕様に対して，JIS X0213:2000 附属書 4 において規定される Unicode への対応付けが，不適切となるためである。

例えば，JIS X0213 では「A（ラテン大文字 A)」を U+0041 に対応付けている。一方，Unicode では，U+0041 を LATIN CAPITAL LETTER A，U+FF21 を FULLWIDTH LATIN CAPITAL LETTER A，と規定しており，前者はいわゆる半角文字の「A」，後者はいわゆる全角文字「A」に対応する（The Unicode Consortium(2007) を参照)。BCCWJ では，本文入力に全角文字のみを使うことになっているため，全角「A」の入力には，FULLWIDTH LATIN CAPITAL LETTER A，つまり U+FF21 を用いなければならない。

このように，附属書 4 の規定において，Unicode のいわゆる半角文字に対応付けられるものについては，附属書 5 表 2 を適用した上で，Unicode における Fullwidth ASCII variants を用いることとする。(→リスト：2.10.2.2「全角文字入力に伴う対応 Unicode の再定義」)。

2.10 【付録】 BCCWJ の符号化文字集合と JIS X0213:2004 規格との差異一覧

2.10.1 使用しない文字の一覧 (計：277 文字)

2.10.1.1 装飾・デザインにかかわる文字 (計：186 字)

組み文字 (47 字)

面区点	文字	UCS	日本語通用名	代用字
1-08-75	!!	203C	感嘆符二つ	！！
1-08-76	??	2047	疑問符二つ	？？
1-08-77	?!	2048	疑問符感嘆符	？！
1-08-78	!?	2049	感嘆符疑問符	！？
1-13-32	㍉	3349	全角ミリ	ミリ
1-13-33	㌔	3314	全角キロ	キロ
1-13-34	㌢	3322	全角センチ	センチ
1-13-35	㍍	334D	全角メートル	メートル
1-13-36	㌘	3318	全角グラム	グラム
1-13-37	㌧	3327	全角トン	トン
1-13-38	㌃	3303	全角アール	アール
1-13-39	㌶	3336	全角ヘクタール	ヘクタール
1-13-40	㍑	3351	全角リットル	リットル
1-13-41	㍗	3357	全角ワット	ワット
1-13-42	㌍	330D	全角カロリー	カロリー
1-13-43	㌦	3326	全角ドル	ドル
1-13-44	㌣	3323	全角セント	セント
1-13-45	㌫	332B	全角パーセント	パーセント
1-13-46	㍊	334A	全角ミリバール	ミリバール
1-13-47	㌻	333B	全角ページ	ページ
1-13-48	㎜	339C	全角 MM	ｍｍ
1-13-49	㎝	339D	全角 CM	ｃｍ
1-13-50	㎞	339E	全角 KM	ｋｍ
1-13-51	㎎	338E	全角 MG	ｍｇ
1-13-52	㎏	338F	全角 KG	ｋｇ
1-13-53	㏄	33C4	全角 CC	ｃｃ
1-13-54	㎡	33A1	全角 M2	ｍ２
1-13-63	㍻	337B	全角元号平成	平成
1-13-66	№	2116	全角 NO	Ｎｏ．
1-13-67	㏍	33CD	全角 KK	Ｋ．Ｋ．
1-13-68	℡	2121	全角 TEL	ＴＥＬ
1-13-74	㈱	3231	全角括弧付き株	（株）
1-13-75	㈲	3232	全角括弧付き有	（有）
1-13-76	㈹	3239	全角括弧付き代	（代）
1-13-77	㍾	337E	全角元号明治	明治

組み文字 (47 **字**)

面区点	文字	UCS	日本語通用名	代用字
1-13-78	㍷	337D	全角元号大正	大正
1-13-79	㍶	337C	全角元号昭和	昭和
1-09-20	½	00BD	2 分の 1	1／2
1-07-88	⅓	2153	3 分の 1	1／3
1-07-89	⅔	2154	3 分の 2	2／3
1-09-19	¼	00BC	4 分の 1	1／4
1-09-21	¾	00BE	4 分の 3	3／4
1-07-90	⅕	2155	5 分の 1	1／5
1-12-31	xi	217A	ローマ数字 11 小文字	x i
1-12-32	xii	217B	ローマ数字 12 小文字	x ii
1-13-31	Ⅺ	216A	ローマ数字 11	Ⅹ Ⅰ
1-13-55	Ⅻ	216B	ローマ数字 12	Ⅹ Ⅱ

囲み文字 (136 **字**)

面区点	文字	UCS	日本語通用名	代用字
1-13-1	①	2460	丸 1	1
1-13-2	②	2461	丸 2	2
1-13-3	③	2462	丸 3	3
1-13-4	④	2463	丸 4	4
1-13-5	⑤	2464	丸 5	5
1-13-6	⑥	2465	丸 6	6
1-13-7	⑦	2466	丸 7	7
1-13-8	⑧	2467	丸 8	8
1-13-9	⑨	2468	丸 9	9
1-13-10	⑩	2469	丸 10	1 0
1-13-11	⑪	246A	丸 11	1 1
1-13-12	⑫	246B	丸 12	1 2
1-13-13	⑬	246C	丸 13	1 3
1-13-14	⑭	246D	丸 14	1 4
1-13-15	⑮	246E	丸 15	1 5
1-13-16	⑯	246F	丸 16	1 6
1-13-17	⑰	2470	丸 17	1 7
1-13-18	⑱	2471	丸 18	1 8
1-13-19	⑲	2472	丸 19	1 9
1-13-20	⑳	2473	丸 20	2 0
1-08-33	㉑	3251	丸 21	2 1
1-08-34	㉒	3252	丸 22	2 2
1-08-35	㉓	3253	丸 23	2 3
1-08-36	㉔	3254	丸 24	2 4

囲み文字 (136 **字**)

面区点	文字	UCS	日本語通用名	代用字
1-08-37	㉕	3255	丸 25	2 5
1-08-38	㉖	3256	丸 26	2 6
1-08-39	㉗	3257	丸 27	2 7
1-08-40	㉘	3258	丸 28	2 8
1-08-41	㉙	3259	丸 29	2 9
1-08-42	㉚	325A	丸 30	3 0
1-08-43	㉛	325B	丸 31	3 1
1-08-44	㉜	325C	丸 32	3 2
1-08-45	㉝	325D	丸 33	3 3
1-08-46	㉞	325E	丸 34	3 4
1-08-47	㉟	325F	丸 35	3 5
1-08-48	㊱	32B1	丸 36	3 6
1-08-49	㊲	32B2	丸 37	3 7
1-08-50	㊳	32B3	丸 38	3 8
1-08-51	㊴	32B4	丸 39	3 9
1-08-52	㊵	32B5	丸 40	4 0
1-08-53	㊶	32B6	丸 41	4 1
1-08-54	㊷	32B7	丸 42	4 2
1-08-55	㊸	32B8	丸 43	4 3
1-08-56	㊹	32B9	丸 44	4 4
1-08-57	㊺	32BA	丸 45	4 5
1-08-58	㊻	32BB	丸 46	4 6
1-08-59	㊼	32BC	丸 47	4 7
1-08-60	㊽	32BD	丸 48	4 8
1-08-61	㊾	32BE	丸 49	4 9
1-08-62	㊿	32BF	丸 50	5 0
1-12-1	❶	2776	黒丸 1	1
1-12-2	❷	2777	黒丸 2	2
1-12-3	❸	2778	黒丸 3	3
1-12-4	❹	2779	黒丸 4	4
1-12-5	❺	277A	黒丸 5	5
1-12-6	❻	277B	黒丸 6	6
1-12-7	❼	277C	黒丸 7	7
1-12-8	❽	277D	黒丸 8	8
1-12-9	❾	277E	黒丸 9	9
1-12-10	❿	277F	黒丸 10	1 0
1-12-11	⓫	24EB	黒丸 11	1 1
1-12-12	⓬	24EC	黒丸 12	1 2
1-12-13	⓭	24ED	黒丸 13	1 3
1-12-14	⓮	24EE	黒丸 14	1 4

囲み文字 (136 **字**)

面区点	文字	UCS	日本語通用名	代用字
1-12-15	⓯	24EF	黒丸 15	1 5
1-12-16	⓰	24F0	黒丸 16	1 6
1-12-17	⓱	24F1	黒丸 17	1 7
1-12-18	⓲	24F2	黒丸 18	1 8
1-12-19	⓳	24F3	黒丸 19	1 9
1-12-20	⓴	24F4	黒丸 20	2 0
1-06-58	①	24F5	二重丸 1	1
1-06-59	②	24F6	二重丸 2	2
1-06-60	③	24F7	二重丸 3	3
1-06-61	④	24F8	二重丸 4	4
1-06-62	⑤	24F9	二重丸 5	5
1-06-63	⑥	24FA	二重丸 6	6
1-06-64	⑦	24FB	二重丸 7	7
1-06-65	⑧	24FC	二重丸 8	8
1-06-66	⑨	24FD	二重丸 9	9
1-06-67	⑩	24FE	二重丸 10	1 0
1-12-33	ⓐ	24D0	丸 A 小文字	a
1-12-34	ⓑ	24D1	丸 B 小文字	b
1-12-35	ⓒ	24D2	丸 C 小文字	c
1-12-36	ⓓ	24D3	丸 D 小文字	d
1-12-37	ⓔ	24D4	丸 E 小文字	e
1-12-38	ⓕ	24D5	丸 F 小文字	f
1-12-39	ⓖ	24D6	丸 G 小文字	g
1-12-40	ⓗ	24D7	丸 H 小文字	h
1-12-41	ⓘ	24D8	丸 I 小文字	i
1-12-42	ⓙ	24D9	丸 J 小文字	j
1-12-43	ⓚ	24DA	丸 K 小文字	k
1-12-44	ⓛ	24DB	丸 L 小文字	l
1-12-45	ⓜ	24DC	丸 M 小文字	m
1-12-46	ⓝ	24DD	丸 N 小文字	n
1-12-47	ⓞ	24DE	丸 O 小文字	o
1-12-48	ⓟ	24DF	丸 P 小文字	p
1-12-49	ⓠ	24E0	丸 Q 小文字	q
1-12-50	ⓡ	24E1	丸 R 小文字	r
1-12-51	ⓢ	24E2	丸 S 小文字	s
1-12-52	ⓣ	24E3	丸 T 小文字	t
1-12-53	ⓤ	24E4	丸 U 小文字	u
1-12-54	ⓥ	24E5	丸 V 小文字	v
1-12-55	ⓦ	24E6	丸 W 小文字	w
1-12-56	ⓧ	24E7	丸 X 小文字	x

囲み文字 (136 字)

面区点	文字	UCS	日本語通用名	代用字
1-12-57	ⓨ	24E8	丸 Y 小文字	y
1-12-58	ⓩ	24E9	丸 Z 小文字	z
1-12-59	㋐	32D0	丸ア	ア
1-12-60	㋑	32D1	丸イ	イ
1-12-61	㋒	32D2	丸ウ	ウ
1-12-62	㋓	32D3	丸エ	エ
1-12-63	㋔	32D4	丸オ	オ
1-12-64	㋕	32D5	丸カ	カ
1-12-65	㋖	32D6	丸キ	キ
1-12-66	㋗	32D7	丸ク	ク
1-12-67	㋘	32D8	丸ケ	ケ
1-12-68	㋙	32D9	丸コ	コ
1-12-69	㋚	32DA	丸サ	サ
1-12-70	㋛	32DB	丸シ	シ
1-12-71	㋜	32DC	丸ス	ス
1-12-72	㋝	32DD	丸セ	セ
1-12-73	㋞	32DE	丸ソ	ソ
1-12-74	㋟	32DF	丸タ	タ
1-12-75	㋠	32E0	丸チ	チ
1-12-76	㋡	32E1	丸ツ	ツ
1-12-77	㋢	32E2	丸テ	テ
1-12-78	㋣	32E3	丸ト	ト
1-12-79	㋤	32FA	丸ロ	ロ
1-12-80	㋥	32E9	丸ハ	ハ
1-12-81	㋦	32E5	丸ニ	ニ
1-12-82	㋧	32ED	丸ホ	ホ
1-12-83	㋨	32EC	丸ヘ	ヘ
1-13-69	㊤	32A4	丸付き上	上
1-13-70	㊥	32A5	丸付き中	中
1-13-71	㊦	32A6	丸付き下	下
1-13-72	㊧	32A7	丸付き左	左
1-13-73	㊨	32A8	丸付き右	右

上付き文字 (3 字)

面区点	文字	UCS	日本語通用名	代用字
1-09-16	¹	00B9	上付き 1	1
1-09-12	²	00B2	上付き 2	2
1-09-13	³	00B3	上付き 3	3

2.10.1.2　類似の非漢字 (計：17 字)

面区点	文字	UCS	日本語通用名	代用字	代-面区点	代-UCS	代-日本語通用名
1-09-02		00A0	ノーブレークスペース		1-01-01	3000	和字間隔
1-03-92	–	2013	二分ダーシ、ダッシュ（二分）	—	1-01-29	2015	ダッシュ（全角）
1-02-17	-	002D	ハイフンマイナス	-	1-01-30	2010	ハイフン（四分）
				−	1-01-61	2212	負符号、減算記号
1-03-91	=	30A0	二重ハイフン、二分二重ダッシュ	＝	1-01-65	FF1D	等号
1-01-17	¯	203E	オーバーライン、論理否定記号	￣	1-09-11	FFE3	マクロン
1-02-18	~	007E	チルド				
1-09-14	·	00B7	中点（ラテン）	・	1-01-06	30FB	中点
1-03-32	•	2022	ビュレット				
1-03-31	◦	25E6	白ビュレット	○	1-01-91	25CB	丸印、白丸
1-02-94	○	25EF	大きな丸				
1-03-26	⦿	29BF	丸中黒	◉	1-03-27	25C9	蛇の目
1-09-08	«	00AB	始め二重山括弧引用記号、始めギュメ	《	1-01-52	300A	始め二重山括弧
1-09-18	»	00BB	終わり二重山括弧引用記号、終わりギュメ	》	1-01-53	300B	終わり二重山括弧
1-13-64	〝	301D	始めダブルミニュート	“	1-01-40	201C	左ダブル引用符
1-13-65	〟	301F	終わりダブルミニュート	”	1-01-41	201D	右ダブル引用符
1-02-16	"	0022	引用符、クォーテーションマーク	“	1-01-40	201C	左ダブル引用符
				”	1-01-41	201D	右ダブル引用符
1-02-15	'	0027	アポストロフィ	‘	1-01-38	2018	左引用符
				’	1-01-39	2019	右引用符

2.10.1.3　入力対象外文字 (計：48 字)

改行関連文字 (1 字)

面区点	文字	UCS	日本語通用名
1-09-09	–	00AD	ソフトハイフン

けい線素片 (47 字)

面区点	文字	UCS	日本語通用名
1-08-1	─	2500	横細線素片
1-08-2	│	2502	縦細線素片
1-08-3	┌	250C	細線素片左上

けい線素片 (47 **字**)

面区点	文字	UCS	日本語通用名
1-08-4	┐	2510	細線素片右上
1-08-5	┘	2518	細線素片右下
1-08-6	└	2514	細線素片左下
1-08-7	├	251C	細線素片左
1-08-8	┬	252C	細線素片上
1-08-9	┤	2524	細線素片右
1-08-10	┴	2534	細線素片下
1-08-11	┼	253C	細線素片中央
1-08-12	━	2501	横太線素片
1-08-13	┃	2503	縦太線素片
1-08-14	┏	250F	太線素片左上
1-08-15	┓	2513	太線素片右上
1-08-16	┛	251B	太線素片右下
1-08-17	┗	2517	太線素片左下
1-08-18	┣	2523	太線素片左
1-08-19	┳	2533	太線素片上
1-08-20	┫	252B	太線素片右
1-08-21	┻	253B	太線素片下
1-08-22	╋	254B	太線素片中央
1-08-23	┠	2520	縦太線横細線素片左
1-08-24	┯	252F	横太線縦細線素片上
1-08-25	┨	2528	縦太線横細線素片右
1-08-26	┷	2537	横太線縦細線素片下
1-08-27	┿	253F	縦細線横太線素片中央
1-08-28	┝	251D	縦細線横太線素片左
1-08-29	┰	2530	横細線縦太線素片上
1-08-30	┥	2525	縦細線横太線素片右
1-08-31	┸	2538	横細線縦太線素片下
1-08-32	╂	2542	横細線縦太線素片中央
1-07-34	⎡	23BE	左上角素片
1-07-35	⎣	23BF	左下角素片
1-07-36	⏀	23C0	丸付き縦線素片
1-07-37	⏁	23C1	丸付き上横縦線素片
1-07-38	⏂	23C2	丸付き下横縦線素片
1-07-39	⏃	23C3	三角付き縦線素片
1-07-40	⏄	23C4	三角付き上横縦線素片
1-07-41	⏅	23C5	三角付き下横縦線素片
1-07-42	⏆	23C6	波付き縦線素片
1-07-43	⏇	23C7	波付き上横縦線素片
1-07-44	⏈	23C8	波付き下横縦線素片

けい線素片 (47 字)

面区点	文字	UCS	日本語通用名
1-07-45	┯	23C9	上横縦線素片
1-07-46	┷	23CA	下横縦線素片
1-07-47	┐	23CB	右上角素片
1-07-48	┘	23CC	右下角素片

2.10.1.4　Unicode における合成処理対象文字 (25 字)

面区点	文字	UCS	日本語通用名	代用字
1-4-87	が	<304B, 309A>	半濁点付き平仮名か	か
1-4-88	ぎ	<304D, 309A>	半濁点付き平仮名き	き
1-4-89	ぐ	<304F, 309A>	半濁点付き平仮名く	く
1-4-90	げ	<3051, 309A>	半濁点付き平仮名け	け
1-4-91	ご	<3053, 309A>	半濁点付き平仮名こ	こ
1-5-87	ガ	<30AB, 309A>	半濁点付き片仮名カ	カ
1-5-88	ギ	<30AD, 309A>	半濁点付き片仮名キ	キ
1-5-89	グ	<30AF, 309A>	半濁点付き片仮名ク	ク
1-5-90	ゲ	<30B1, 309A>	半濁点付き片仮名ケ	ケ
1-5-91	ゴ	<30B3, 309A>	半濁点付き片仮名コ	コ
1-5-92	ゼ	<30BB, 309A>	半濁点付き片仮名セ	セ
1-5-93	ヅ	<30C4, 309A>	半濁点付き片仮名ツ	ツ
1-5-94	ド	<30C8, 309A>	半濁点付き片仮名ト	ト
1-6-88	プ	<31F7, 309A>	小書き半濁点付き片仮名フ	フ
1-11-36	æ̀	<00E6, 0300>	グレーブアクセント付き AE 小文字	æ
1-11-40	ɔ̀	<0254, 0300>	グレーブアクセント付きオープン O 小文字	ɔ
1-11-41	ɔ́	<0254, 0301>	アキュートアクセント付きオープン O 小文字	ɔ
1-11-42	ʌ̀	<028C, 0300>	グレーブアクセント付きターンド V 小文字	ʌ
1-11-43	ʌ́	<028C, 0301>	アキュートアクセント付きターンド V 小文字	ʌ
1-11-44	ə̀	<0259, 0300>	グレーブアクセント付き SCHWA 小文字	ə
1-11-45	ə́	<0259, 0301>	アキュートアクセント付き SCHWA 小文字	ə
1-11-46	ə̒	<025A, 0300>	グレーブアクセントとフック付き SCHWA 小文字	ə̒
1-11-47	ə̓	<025A, 0301>	アキュートアクセントとフック付き SCHWA 小文字	ə̒
1-11-69	ˊ˥	<02E9, 02E5>	声調記号上昇調	˩˥
1-11-70	˧˩	<02E5, 02E9>	声調記号下降調	˥˩

2.10.1.5　漢字 (1 字)

面区点	文字	UCS	代用字	代用面区点	代用 UCS
1-47-52	叱	20B9F	叱	1-28-24	U+53F1

2.10.2 代用コードで運用する文字の一覧 (計：96 字)

2.10.2.1 JIS X0213:2000 の規定コードを用いるもの (2 字)

面区点	文字	JIS 定義 UCS	使用 UCS	日本語通用名・字体説明
1-01-29	―	2014	2015	ダッシュ
2-94-05	鵈	29FD7	29FCE	「予」＋「鳥」

2.10.2.2 全角文字入力に伴う対応 Unicode の再定義 (94 字)

面区点	文字	JIS 定義 UCS	使用 UCS	日本語通用名
1-01-04	，	002C	FF0C	コンマ
1-01-05	．	002E	FF0E	ピリオド
1-01-07	：	003A	FF1A	コロン
1-01-08	；	003B	FF1B	セミコロン
1-01-09	？	003F	FF1F	疑問符
1-01-14	｀	0060	FF40	アクサングラーブ，グレーブアクセント
1-01-16	＾	005E	FF3E	アクサンシルコンフレックス，サーカムフレックスアクセント
1-09-11	￣	00AF	FFE3	マクロン [*7]
1-01-18	＿	005F	FF3F	アンダーライン
1-01-31	／	002F	FF0F	斜線
1-01-32	＼	005C	FF3C	逆斜線
1-01-35	｜	007C	FF5C	縦線
1-01-42	（	0028	FF08	始め小括弧，始め丸括弧
1-01-43	）	0029	FF09	終わり小括弧，終わり丸括弧
1-01-46	［	005B	FF3B	始め大括弧，始め角括弧
1-01-47	］	005D	FF3D	終わり大括弧，終わり角括弧
1-01-48	｛	007B	FF5B	始め中括弧，始め波括弧
1-01-49	｝	007D	FF5D	終わり中括弧，終わり波括弧
1-01-60	＋	002B	FF0B	正符号，加算記号
1-01-65	＝	003D	FF1D	等号
1-01-67	＜	003C	FF1C	不等号（より小）
1-01-68	＞	003E	FF1E	不等号（より大）
1-01-79	￥	00A5	FFE5	円記号
1-01-80	＄	0024	FF04	ドル記号
1-01-83	％	0025	FF05	パーセント
1-01-85	＆	0026	FF06	アンパサンド
1-01-86	＊	002A	FF0A	星印，アステリスク
1-01-87	＠	0040	FF20	単価記号，アットマーク
1-02-54	⦅	2985	FF5F	始め二重パーレン，始め二重括弧
1-02-55	⦆	2986	FF60	終わり二重パーレン，終わり二重括弧

[*7] オーバーラインの代用として用いる。

面区点	文字	JIS 定義 UCS	使用 UCS	日本語通用名
1-03-16	0	0030	FF10	0
1-03-17	1	0031	FF11	1
1-03-18	2	0032	FF12	2
1-03-19	3	0033	FF13	3
1-03-20	4	0034	FF14	4
1-03-21	5	0035	FF15	5
1-03-22	6	0036	FF16	6
1-03-23	7	0037	FF17	7
1-03-24	8	0038	FF18	8
1-03-25	9	0039	FF19	9
1-03-33	A	0041	FF21	ラテン大文字 A
1-03-34	B	0042	FF22	ラテン大文字 B
1-03-35	C	0043	FF23	ラテン大文字 C
1-03-36	D	0044	FF24	ラテン大文字 D
1-03-37	E	0045	FF25	ラテン大文字 E
1-03-38	F	0046	FF26	ラテン大文字 F
1-03-39	G	0047	FF27	ラテン大文字 G
1-03-40	H	0048	FF28	ラテン大文字 H
1-03-41	I	0049	FF29	ラテン大文字 I
1-03-42	J	004A	FF2A	ラテン大文字 J
1-03-43	K	004B	FF2B	ラテン大文字 K
1-03-44	L	004C	FF2C	ラテン大文字 L
1-03-45	M	004D	FF2D	ラテン大文字 M
1-03-46	N	004E	FF2E	ラテン大文字 N
1-03-47	O	004F	FF2F	ラテン大文字 O
1-03-48	P	0050	FF30	ラテン大文字 P
1-03-49	Q	0051	FF31	ラテン大文字 Q
1-03-50	R	0052	FF32	ラテン大文字 R
1-03-51	S	0053	FF33	ラテン大文字 S
1-03-52	T	0054	FF34	ラテン大文字 T
1-03-53	U	0055	FF35	ラテン大文字 U
1-03-54	V	0056	FF36	ラテン大文字 V
1-03-55	W	0057	FF37	ラテン大文字 W
1-03-56	X	0058	FF38	ラテン大文字 X
1-03-57	Y	0059	FF39	ラテン大文字 Y
1-03-58	Z	005A	FF3A	ラテン大文字 Z
1-03-65	a	0061	FF41	ラテン小文字 A
1-03-66	b	0062	FF42	ラテン小文字 B
1-03-67	c	0063	FF43	ラテン小文字 C
1-03-68	d	0064	FF44	ラテン小文字 D
1-03-69	e	0065	FF45	ラテン小文字 E

面区点	文字	JIS 定義 UCS	使用 UCS	日本語通用名
1-03-70	f	0066	FF46	ラテン小文字 F
1-03-71	g	0067	FF47	ラテン小文字 G
1-03-72	h	0068	FF48	ラテン小文字 H
1-03-73	i	0069	FF49	ラテン小文字 I
1-03-74	j	006A	FF4A	ラテン小文字 J
1-03-75	k	006B	FF4B	ラテン小文字 K
1-03-76	l	006C	FF4C	ラテン小文字 L
1-03-77	m	006D	FF4D	ラテン小文字 M
1-03-78	n	006E	FF4E	ラテン小文字 N
1-03-79	o	006F	FF4F	ラテン小文字 O
1-03-80	p	0070	FF50	ラテン小文字 P
1-03-81	q	0071	FF51	ラテン小文字 Q
1-03-82	r	0072	FF52	ラテン小文字 R
1-03-83	s	0073	FF53	ラテン小文字 S
1-03-84	t	0074	FF54	ラテン小文字 T
1-03-85	u	0075	FF55	ラテン小文字 U
1-03-86	v	0076	FF56	ラテン小文字 V
1-03-87	w	0077	FF57	ラテン小文字 W
1-03-88	x	0078	FF58	ラテン小文字 X
1-03-89	y	0079	FF59	ラテン小文字 Y
1-03-90	z	007A	FF5A	ラテン小文字 Z

参考文献

日本工業標準調査会 (2004)『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（追補 1） JIS X 0213:2004』日本規格協会.

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.

The Unicode Consortium(2007) *The Unicode Standard, Version 5.0*, Addison-Wesley

第 3 章

タグ仕様

3.1 概要

本章では，BCCWJ のタグ仕様の概要と凡例を示す。現在定義されているタグ (可変長 45 要素，固定長 2 要素) [*1] に関して，それぞれの適用例を具体的に例示しながら説明する。最初にお断りしておくが，版面の関係上，（特に画像など）判読しづらい箇所が散見される。本仕様は Web サイト[*2] においても公開しているので，そちらも併せて参照されたい。

3.1.1 タグの分類と列挙の仕方

タグは，付与する対象という観点から，ブロック要素とインライン要素に分けられる。

- ブロック要素…複数の論理行を包含しうる要素
- インライン要素…常に 1 論理行内に包含される要素

また，機能に着目すると，タグは概して以下のように分類できる。

- サンプルに関する要素
- 言語的な階層構造を記述する要素
- 特定の言語構造を記述する要素
  - 図表に関する要素
  - 引用に関する要素
  - 注記に関する要素
  - その他の要素
- 文字・表記に関する要素

タグには，それ自体独立して完結するものと，有機的に他と関係性を持つものに分かれ，仕様の理解のためには関係するタグも参照することが望まれる。本仕様は，Web サイトでも公開しているが，Web サイトの特性を活かし，各タグ要素について機能ごとに分類し，関連するタグを相互参照できるようにしている。しかし，本章は紙媒体での公開であり，参照の便宜を図るため，各タグ要素を要素名順に列挙した。

3.1.2 可変長と固定長におけるタグ仕様の相違

BCCWJ には，一つの記事 (article 要素) を包含し，1 サンプル中の文字数が，最大 1 万字程度である可変長と，1 サンプル中に含まれる文字数が 1000 文字以上となることを保証された固定長という 2 種類のサンプ

[*1] 本仕様は，ver 2.2 に基づくものであるり，現時点での最新版である。
[*2] http://www2.ninjal.ac.jp/densi/public/wiki/

ルがある。可変長が文章・談話としてのまとまりを重視したサンプルであるのに対し，固定長は文字数を固定し，文字・語・文といったより小さな言語単位に着目して取得されるサンプルである。可変長と固定長は形態や利用目的が異なるため，タグ仕様についても，次にまとめた通り，若干異なる箇所があり注意が必要である。

固定長のタグ仕様における注意点

- 以下の 2 要素について，「可変長」とは定義が異なる。
  - sample 要素
  - sampling 要素
- cluster 要素は認定されない。
- 固定長のサンプル範囲は文字数で限定されるため，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。

3.2　凡例

各タグ要素の説明は，以下のような構成となっている。

- 概要
- 形式
  - 要素
  - 属性
  - DTD
- 説明
- 形式化例

注記

(1) 「要素」に挙げられている ENTITY(実体参照) については，以下のとおり定義した。

```
<!ENTITY % blockElement "article | cluster | paragraph | authorsData | title | titleBlock |
 figureBlock |abstract | quotation | blockEnd | contents | list | profile | rejectedBlock |
 noteBody | orphanedTitle| verse | info">
```

```
<!ENTITY % characterElement "missingCharacter| correction | image | enclosedCharacter |replace|
 jis2004| jis2000 | substitution">
```

```
<!ENTITY % stringElement "rejectedSpan | ruby | fraction|sampling| quote | subScript|
 superScript| noteMarker | noteBodyInline">
```

```
<!ENTITY % inlineElement "%characterElement; | %stringElement;">
```

```
<!ENTITY % inlineText "#PCDATA | %inlineElement;">
```

```
<!ENTITY % character "#PCDATA| %characterElement;">
```

(2) 各タグ要素の最後に示した形式化例について，例示すべきタグが引き立つようにするため，文章の一部や特に示す必要がないタグを省略した箇所がある。省略した箇所については欄外に注記を施した。

タグ一覧 (可変長)

abstract 要素

概要

- article 要素，または cluster 要素の概要に相当する文書要素を表す。

形式

■ 要素

blockEnd, br, cluster, list, noteBody, paragraph, quotation, rejectedBlock, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT abstract
          (blockEnd|br|cluster|list|noteBody|paragraph|quotation|
           rejectedBlock|sentence)*>
```

説明

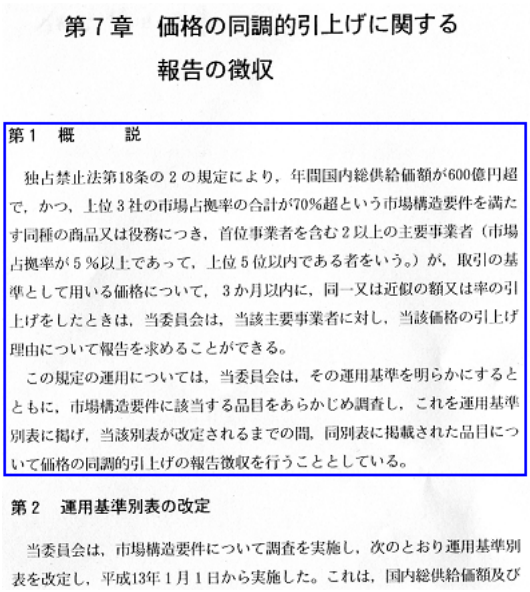
abstract 要素は，article 要素，または cluster 要素の概要に相当する文書要素を表す。abstract 要素に該当する文書要素としては，例えば，次のような文書要素がある。

- 新聞のリード
- 論文の概要やキーワード
- 雑誌の記事要旨・前文・導入文

これらは，ある一定範囲の文書要素の概要として機能すると共に，以下の条件を**いずれも**満たすものとする。

- 囲み，段組の差など，提示形式の差異によって，本文 (文書の主たる構成要素となっている文章の連なり) とは物理的に明確に切り離された文書要素となっている。
- 概要の対象となっている文書要素 (article 要素や，cluster 要素) のタイトルに続く冒頭部に位置している。

この条件によって，本文内の一部の章や一部の段落として現われる概要に相当する文書要素は，abstract 要素の対象からは除かれる (下図『独占禁止白書』における青囲み「第１　概説」を参照。後続する「第２　運用規準別表の改訂」と同様に，「第７章」の本文を構成する章そのものとなっているため，abstract 要素とはならない)。



第７章　価格の同調的引上げに関する
報告の徴収

第１　概　　説

独占禁止法第18条の２の規定により，年間国内総供給価額が600億円超で，かつ，上位３社の市場占拠率の合計が70%超という市場構造要件を満たす同種の商品又は役務につき，首位事業者を含む２以上の主要事業者（市場占拠率が５%以上であって，上位５位以内である者をいう。）が，取引の基準として用いる価格について，３か月以内に，同一又は近似の額又は率の引上げをしたときは，当委員会は，当該主要事業者に対し，当該価格の引上げ理由について報告を求めることができる。

この規定の運用については，当委員会は，その運用基準を明らかにするとともに，市場構造要件に該当する品目をあらかじめ調査し，これを運用基準別表に掲げ，当該別表が改定されるまでの間，同別表に掲載された品目について価格の同調的引上げの報告徴収を行うこととしている。

第２　運用基準別表の改定

当委員会は，市場構造要件について調査を実施し，次のとおり運用基準別表を改定し，平成13年１月１日から実施した。これは，国内総供給価額及び

『独占禁止白書』平成 13 年版

また，上記の条件を満たしていれば，article 要素だけでなく，cluster 要素にも abstract 要素が含まれていてもよいことに注意する (例２)。

なお，「概要」，「Abstract」など，概要となる文書要素のタイトル・代表記述に相当する文書要素がある場合は，abstract 要素の中で title 要素として記述し， title 要素が包括する文書要素の範囲を cluster 要素を用いて記述する (例３)。この場合，abstract 要素が複数の cluster 要素から成り立つ場合もある (例４)。

形式化例

■ 例１：新聞リード (『毎日新聞』2003 年 3 月 2 日朝刊)

原資料

経済がわかる　企業がみえる

湾岸戦争時の日米株価

株価↓下落

長期化なら景気失速

イラク開戦シミュレーション

為替↓

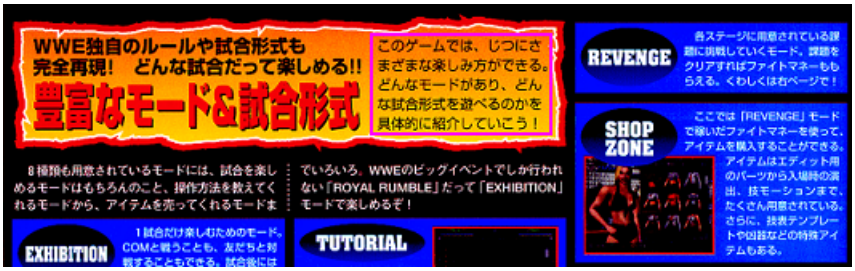
形式化

```
<titleBlock>
 <title>
 <sentence type="quasi">イラク開戦シミュレーション</sentence>
 </title>
</titleBlock>
<abstract>
 <paragraph>
   <sentence>イラク情勢が緊迫化し，世界経済への影響が懸念されている。</sentence>
                                          :
                                          :
 </paragraph>
</abstract>
<cluster>
 <titleBlock>
  <title>
   <sentence type="quasi">株価→下落</sentence>
   <sentence type="quasi">０４年度まで日本マイナス成長も</sentence>
  </title>
 </titleBlock>
                                          :（以下略）
```

※ br 要素を省略。以下同じ。

■ 例２：雑誌導入文 (『電撃ＧＡＭＥＣＵＢＥ』2003 年 12 月号)

原資料

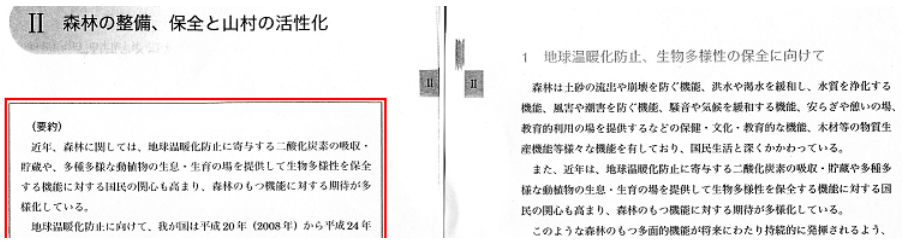


形式化

```
<cluster>
 <titleBlock>
  <sentence>ＷＷＥ独自のルールや試合形式も完全再現！ </sentence><sentence>　どんな
  試合だって楽しめる！！ </sentence>
  <title>
  <sentence type="quasi">豊富なモード＆試合形式</sentence>
  </title>
 </titleBlock>
 <abstract>
  <sentence>このゲームでは、じつにさまざまな楽しみ方ができる。</sentence>
                                          :
 </abstract>
                                          :（以下略）
```

■ 例３：論文概要 (『図説　森林・林業白書』平成 14 年版)

原資料



Ⅱ　森林の整備、保全と山村の活性化

（要約）
近年、森林に関しては、地球温暖化防止に寄与する二酸化炭素の吸収・貯蔵や、多種多様な動植物の生息・生育の場を提供して生物多様性を保全する機能に対する国民の関心も高まり、森林のもつ機能に対する期待が多様化している。
地球温暖化防止に向けて、我が国は平成20年（2008年）から平成24年

Ⅱ　森林の整備、保全と山村の活性化

1　地球温暖化防止、生物多様性の保全に向けて

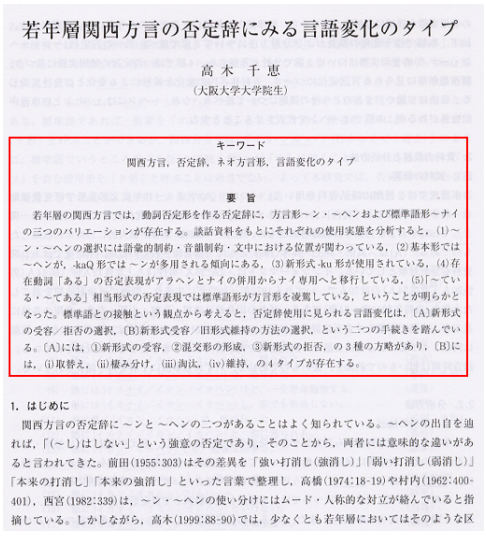
森林は土砂の流出や崩壊を防ぐ機能、洪水や渇水を緩和し、水質を浄化する機能、風害や潮害を防ぐ機能、騒音や気候を緩和する機能、安らぎや憩いの場、教育的利用の場を提供するなどの保健・文化・教育的な機能、木材等の物質生産機能等様々な機能を有しており、国民生活と深くかかわっている。
また、近年は、地球温暖化防止に寄与する二酸化炭素の吸収・貯蔵や多種多様な動植物の生息・生育の場を提供して生物多様性を保全する機能に対する国民の関心も高まり、森林のもつ機能に対する期待が多様化している。
このような森林のもつ多面的機能が将来にわたり持続的に発揮されるよう、

形式化

```
<abstract>
 <cluster>
  <titleBlock>
   <title>
   <sentence type="quasi">（要約）</sentence>
   </title>
  </titleBlock>
<paragraph>
 <sentence>　近年、森林に関しては、地球温暖化防止に寄与する二酸化炭素の吸収・貯蔵や、多種多様な…</sentence>
                                        :
 </cluster>
<abstract>
```

■ 例４：複数の要素 (概要・キーワード) からなる abstract 要素 (『日本語科学』2004 年 10 月，16 号)

原資料



若年層関西方言の否定辞にみる言語変化のタイプ

高木　千恵
（大阪大学大学院生）

キーワード
関西方言，否定辞，ネオ方言形，言語変化のタイプ

要　旨
若年層の関西方言では，動詞否定形を作る否定辞に，方言形～ン・～ヘンおよび標準語形～ナイの三つのバリエーションが存在する。談話資料をもとにそれぞれの使用実態を分析すると，(1)～ン・～ヘンの選択には語彙的制約・音韻制約・文中における位置が関わっている，(2)基本形では～ヘンが，-kaQ 形では～ンが多用される傾向にある，(3)新形式 -ku 形が使用されている，(4)存在動詞「ある」の否定表現がアラヘンとナイの併用からナイ専用へと移行している，(5)「～ている・～てある」相当形式の否定表現では標準語形が方言形を凌駕している，ということが明らかとなった。標準語との接触という観点から考えると，否定辞使用に見られる言語変化は，〔A〕新形式の受容／拒否の選択，〔B〕新形式受容／旧形式維持の方法の選択，という二つの手続きを踏んでいる。〔A〕には，①新形式の受容，②混交形の形成，③新形式の拒否，の３種の方略があり，〔B〕には，(i)取替え，(ii)棲み分け，(iii)淘汰，(iv)維持，の４タイプが存在する。

1．はじめに

関西方言の否定辞に～ンと～ヘンの二つがあることはよく知られている。～ヘンの出自を辿れば，「(～し)はしない」という強意の否定であり，そのことから，両者には意味的な違いがあると言われてきた。前田(1955:303)はその差異を「強い打消し(強消し)」「弱い打消し(弱消し)」「本来の打消し」「本来の強消し」といった言葉で整理し，高橋(1974:18-19)や村内(1962:400-401)，西宮(1982:339)は，～ン・～ヘンの使い分けにはムード・人称的な対立が絡んでいると指摘している。しかしながら，高木(1999:88-90)では，少なくとも若年層においてはそのような区

形式化

```
<abstract>
 <cluster>
  <titleBlock>
   <title>
    <sentence type="quasi">キーワード</sentence>
   </title>
  </titleBlock>
  <sentence type="quasi">関西方言，否定辞，ネオ方言形，言語変化のタイプ</sentence>
 </cluster>
 <cluster>
  <titleBlock>
   <title>
    <sentence type="quasi">要旨</sentence>
   </title>
  </titleBlock>
  <paragraph>
   <sentence>  若年層の関西方言では，動詞否定形を作る否定辞に，方言形〜ン・〜ヘンおよび ...</sentence>
                                                    :
  </paragraph>
 </cluster>
</abstract>
```

article 要素

概要

- 同一著者 (単著では一人，共著では複数) による，同一テーマのひとまとまりの文書要素を表す。

形式

■ 要素

br, sentence, %blockElement;

■ 属性

- *articleID* (必須)
  - 外部データベースから article 要素に関する情報 (著者に関する情報など) を取得するための ID。詳しくは，Article テーブルを参照のこと[*3]。
- *isWholeArticle* (必須) ： 記事全体を格納しているか否かを表す。
  - `true` ... 格納している場合
  - `false` ... 格納していない場合

■ DTD

```
<!ELEMENT article (br|sentence|%blockElement;)+>
<!ATTLIST article articleID CDATA #REQUIRED>
<!ATTLIST article isWholeArticle (true|false) #REQUIRED>
```

説明

同一著者 (単著では一人，共著では複数) による，同一テーマのひとまとまりの文書要素を表す。
article 要素は，*articleID* 属性と *isWholeArticle* 属性を持つ。

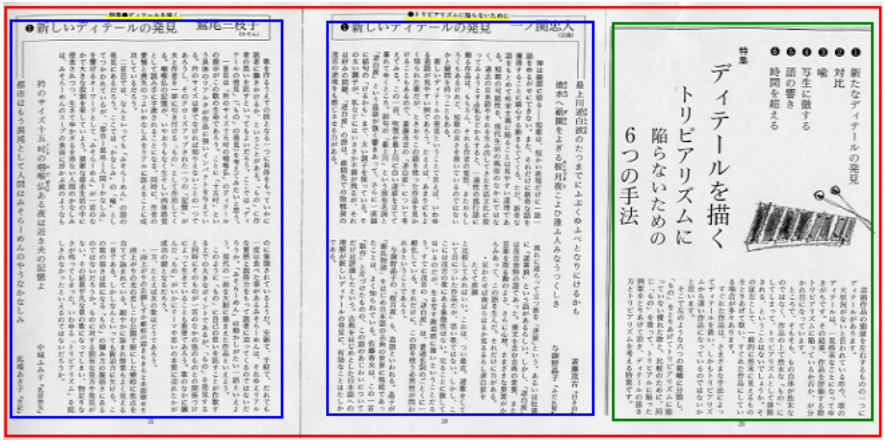
- *articleID* 属性： この属性は，外部データベースから article 要素に関する情報 (著者に関する情報など) を取得するための ID である。参照する外部データベースは，Article テーブル [*3] を想定している。この属性は，必須属性である。
- *isWholeArticle* 属性： サンプルによっては，収録文字数の制限などにより，本来収録すべき文書範囲全体を収録できない場合もある。*isWholeArticle* 属性は，全体を収録できた場合に `true`，できなかった場合に `false` となる。この属性は必須属性である。

[*3] http://www2.ninjal.ac.jp/densi/public/wiki/ から [ver.2.2] → [データベース] を参照のこと。

■ 入れ子の article 要素

入れ子の article 要素の場合，サンプルとして取得される article 要素は，サンプル抽出基準点を包含する最も小さな article 要素となる。ただし，複数の article 要素を包括する枠組み (コーナーなどの親要素としての article 要素) があり，包括する枠組にサンプル抽出基準点があれば，その枠組み自体がサンプル対象の article 要素となる。

例えば，次の記事のように，複数の記事を有する特集について，そのまとめとなる記述が存在する場合である。



『短歌研究』(2003 年 11 月号)

上の例では，特集全体の総括となるような記述 (緑で囲んだ部分；無記名。編集者による) があり，その後ろに，それぞれ別の著者による記事 (青で囲んだ部分；記名) が複数続いている。この場合，個々の記名記事を article 要素として認定した上で，総括部分 (緑) と後続する記事 (青) からなる article 要素 (赤) を認定する。このとき，総括部分 (緑) にサンプル抽出基準点が存在すれば，article 要素 (青) を包含する article 要素 (赤) がサンプルとして取得される。

■ article 要素に含まれる文書サイズの制限

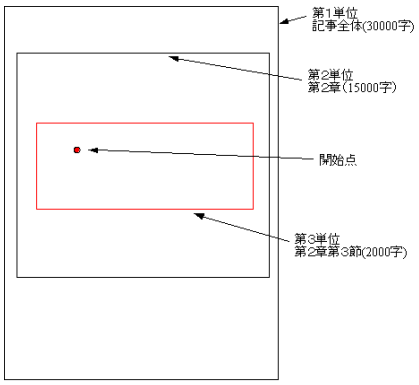
article 要素は，同一著者 (単著では一人，共著では複数) による同一テーマのひとまとまりの文書要素を表す。しかし，小説をはじめとして，サンプリング対象の文書によっては，article 要素に含まれる文書サイズが必要以上に大きくなってしまう場合もある。そこで，article 要素に含まれる文書のサイズを次のように制限する。なお，制限を行った場合は，*isWholeArticle* 属性値は `false` となる。

(1) サンプリングによって指定された「開始点」を含む文章について，同一著者による同一テーマの文書要素を「第 1 単位」と認定する。これが，文書サイズを制限しない場合の article 要素本来の単位である。

(2)「第 1 単位」の文字数を概算し，10000 字を超える場合は，開始点を含む第 1 単位の子要素のうち，開始点を含む cluster 要素，もしくは，blockEnd 要素で区切られる範囲の文書要素を「第 2 単位」とする。この操作を繰り返し，より下位の単位を「第 2 単位」「第 3 単位」…と認定する。そして，10000 字以下になったときに，article 要素とする。

(3) 最小の単位に含まれる文字数が 10000 字以上の場合は，その先頭から 10000 文字を抽出した文書要素を article 要素とする。

この例の場合，第２章第３節がサンプルの単位となる。



■ 著者不明の場合

article 要素は，著者を元に文書要素の範囲を定めるものだが，場合によっては著者が明らかでないこともある。この場合は，目次等の記事情報を手がかりにし，同一著者による文章の範囲を推定する。また，新聞などの，著者の明示も目次も存在しない文書については，内容のまとまりによって，article 要素を認定する。

形式化例

■ 『首都圏白書』　平成１２年版

形式化

```
<article articleID="OW5X_00201_V001">
 <titleBlock>
  <title>
  <sentence type="quasi">９．その他のプロジェクトの推進</sentence>
  </title>
 </titleBlock>
 <cluster>
 <titleBlock>
  <title>
   <sentence type="quasi">（１）　東京湾臨海地域における総合的整備の推進等</sentence>
  </title>
 </titleBlock>
 <paragraph>
 <sentence>　東京湾臨海地域においては、自然環境の保全・回復を図りつつ、都市機能の高質化、都市環境の保全、
 防災性の向上等のニーズに対応した土地利用を進める観点から、低未利用地等を核として周辺地域との一体的かつ
 総合的な整備を進め、各種機能が複合する新たな東京圏を創造する戦略的拠点を形成するとともに、これらの拠点
 間の広域的な連携を促進していくことが重要である。
 </sentence>
 </paragraph>
  :
 </cluster>
 </article>
```

※ br 要素は省略

authorsData 要素

概要

authorsData 要素は，当該文書の著作に関するメタ情報を表すものである。次の二種類がある。

(1) 記事構造上，著作者表示・署名にあたる要素

- 記事を構成するのに参加した人 (著者，対談・インタビュー話者，モデル，カメラマン，イラストレーターなど) の名前や肩書き・役割など

(2) その他，編集情報や記事情報など，その記事そのものに関する情報を表す要素

- その記事の編集に関する取材地，日時，初出情報など

形式

■ 要素

br, info, noteBody, paragraph, rejectedBlock, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT authorsData (br|info|noteBody|paragraph|rejectedBlock|sentence)*>
```

説明

■ 著作者の表示・署名

記事構造上，著作者表示・署名にあたる部分は，authorsData 要素で表す。これは記事の作成にかかわった人の名前や組織，情報の元などを表すものであり，その役割は多岐にわたる。例えば，著者，話者，モデル，カメラマン，イラストレーター，構成，編集などである。

人名と共に，その役割・職業・肩書き，さらに取材地や日時などが記されている場合は，これも含める (→【例 1 】)。ただし，著作者に関する部分でも，著作権にかかわる著作者すべてが authorsData 要素になるわけではないことに注意されたい。authorsData でマークアップする著作者は，(1) 記事の作成にかかわった人を示していて，かつ，(2) 記事の構造として，署名部もしくは著作者表示部 (記事冒頭または末尾に現われる，記事作成関係者をまとめて示す部分) となっている要素である。よって，出てくる名前全てがマークアップされるわけではない。

- 署名部・著作者表示部に，複数の役割と名前が含まれる場合

  - 名前と役割のペアが複数現れる場合，それをまとめて authorsData 要素とする。内部の構造化はされず，改行で区切られる。→ 【例２】
  - 上記表示部において，同一のフォーマットで列挙される項は全て authorsData となる。「協力」「提供」等，著作関係者かどうか判断が微妙なものも入る可能性があるが，それらを別に authorsData の外に出すことはしない。
  - 数人分がまとめて現れている場合は，一人ずつではなくまとめてマークアップされる。
- authorsData を付与しないもの
  - タイトルに含まれる著作者名 (著作者名を冠したタイトル) → title
  - プロフィール中の名前 → profile
  - 記事の下位 cluster 内に示される，インタビュイー，質問回答者，情報提供者などの名前
  - 写真キャプション中の撮影者名 → caption
  - トピック (評論対象) となっている作品の著作者・作者

■ 編集情報や記事情報

著作者名以外に, 編集情報や記事情報など, その記事そのものに関する情報を表す部分も, やはり authorsData 要素となる。これには，その記事の編集に関する取材地，日時，初出情報などがあたる。→【例３】【例４】【例５】

ただしこれも，マークアップされるのは記事冒頭または末尾に現われる要素に限る。

形式化例

■ 例１：著作者（大山正，丸山康則―編「ヒューマンエラーの心理学」）

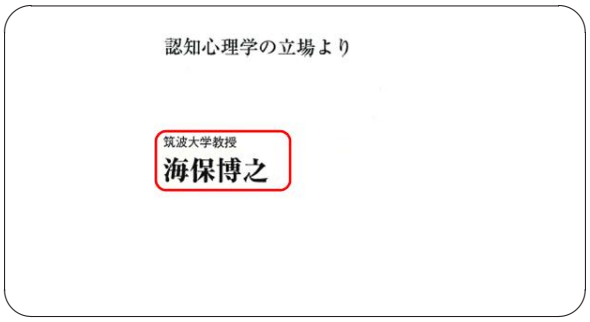
形式化

```
<authorsData>
  <sentence type="quasi">筑波大学教授</sentence><br type="automatic_original"/>
  <sentence type="quasi">海保博之</sentence><br type="automatic_original"/>
</authorsData>
```

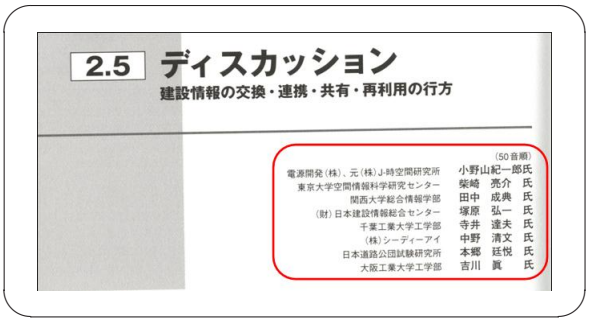
■ 例２：著作者（古田均ほか監修「建設情報の利活用」）

形式化

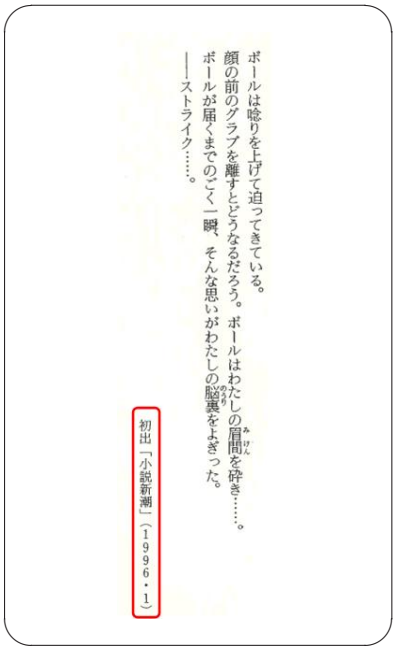
```
<authorsData>
  <sentence type="quasi">（５０音順）</sentence><br type="automatic_original"/>
  <sentence type="quasi">電源開発（株)、元（株）Ｊ‐時空間研究所　小野山紀一郎氏
</sentence><br type="automatic_original"/>
  <sentence type="quasi">東京大学空間情報科学研究センター　柴崎　亮介　氏
</sentence><br type="automatic_original"/>
  <sentence type="quasi">関西大学総合情報学部　田中　成典　氏</sentence><br type="automatic_original"/>
  <sentence type="quasi">（財）日本建設情報総合センター　塚原　弘一　氏
</sentence><br type="automatic_original"/>
  <sentence type="quasi">千葉工業大学工学部　寺井　達夫　氏</sentence><br type="automatic_original"/>
  <sentence type="quasi">（株）シーディーアイ　中野　清文　氏</sentence><br type="automatic_original"/>
  <sentence type="quasi">日本道路公団試験研究所　本郷　延悦　氏</sentence><br type="automatic_original"/>
  <sentence type="quasi">大阪工業大学工学部　吉川　眞　氏</sentence><br type="automatic_original"/>
</authorsData>
```



【例１】原資料

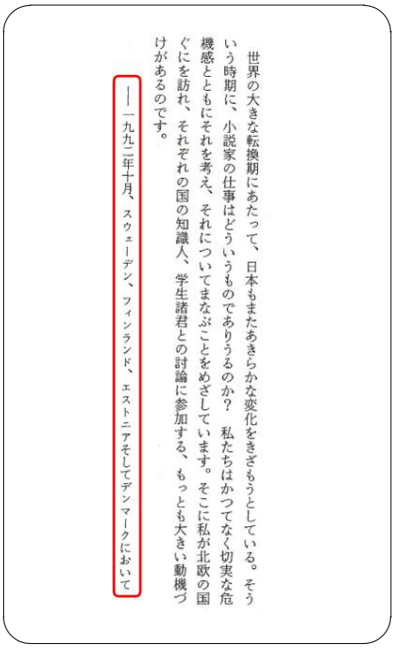


【例２】原資料



【例３】原資料

【例４】原資料



【例５】原資料

■ 例３：記事情報（日本推理作家協会編「自選ショート・ミステリー」）

形式化

```
  <authorsData>
    <sentence type="quasi">初出<quote>「小説新潮」</quote>（１９９６・１）
</sentence><br type="automatic_original"/>
  </authorsData>
```

■ 例４：記事情報（広畑史朗著「警察の視点社会の視点」）

形式化

```
  <authorsData>
    <sentence type="quasi">（平成１２年１２月１日馬頭警察署員を集めての本部長講話）
</sentence><br type="automatic_original"/>
  </authorsData>
```

■ 例５：記事情報（大江健三郎著「あいまいな日本の私」）

形式化

```
 <authorsData>
   <sentence type="quasi">――一九九二年十月、スウェーデン、フィンランド、エストニアそしてデンマークにおいて
</sentence><br type="automatic_original"/>
 </authorsData>
```

blockEnd 要素

概要

- titleBlock 要素が存在しないために，cluster 要素で範囲を記述することができない文書要素を，意味のまとまりや形式のまとまり（＝ブロック）で区切るためのマーカー。
- ブロックの範囲ではなく，まとまりの終端を表す。

形式

■ 要素

- 空要素である。

■ 属性

- なし

■ DTD

```
<!ELEMENT blockEnd EMPTY>
```

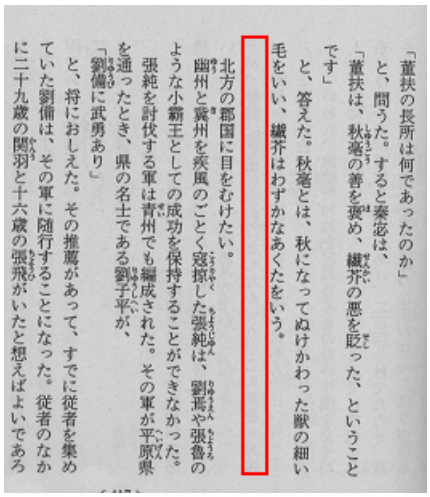
説明

blockEnd 要素は，titleBlock 要素が存在しないために，cluster 要素によって範囲を記述できない，あるまとまりを持った文書要素 (＝ブロック) について，他のまとまりと区別するためのマーカーとして機能し，ブロックの終端に付与される。

blockEnd 要素によるマークアップの対象となるのは，paragraph 要素を除くブロック要素（cluster 要素，quotation 要素，list 要素，figureBlock 要素など）で記述できず，かつ，他の要素との区切りを示す，以下のような形式上のマーカーがあることで，文書の区切れ目が明らかなものである。

- 空行
- 区切り線や区切り記号の挿入
- 後続文書要素先頭の文字強調

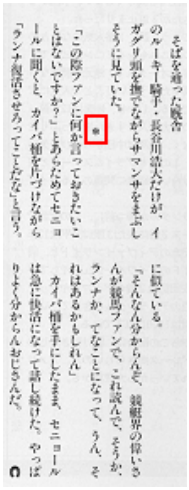
なお，区切りを示すマーカーである空行や，区切り線・記号類は，blockEnd タグによって意味的に置き換えられるため，入力対象とはならない。また，文字強調についても，enclosedCharacter 要素等でマークアップする対象にはならない。



「董扶の長所は何であったのか」
と、問うた。すると秦宓は、
「董扶は、秋毫の善を褒め、纖芥の悪を貶った、ということです」
と、答えた。秋毫とは、秋になってぬけかわった獣の細い毛をいい、纖芥はわずかなあくたをいう。

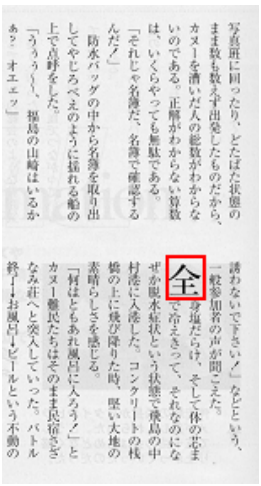
北方の郡国に目をむけたい。
幽州と冀州を疾風のごとく寇掠した張純は、劉焉や張魯のような小覇王としての成功を保持することができなかった。
張純を討伐する軍は青州でも編成された。その軍が平原県を通ったとき、県の名士である劉子平が、
「劉備に武勇あり」
と、将におしえた。その推薦があって、すでに従者を集めていた劉備は、その軍に随行することになった。従者のなかに二十九歳の関羽と十六歳の張飛がいたと想えばよいであろ

【例１】空行（『文藝春秋』2003 年 8 月号）



そばを食った報告
のルーキー騎手・長谷川浩大だけが、イガグリ頭をかきながらサマンサをまぶしそうに見ていた。
「この際ファンに何か言っておきたいことはないですか？」とあらためてセニョールに聞くと、カイバ桶を片づけながら
「ランナ復活させろって言ったな」と言う。
に似ている。
「そんな人分からんぞ、競艇界の偉いさんが観戦ファンで、これ読んで、そうか、ランナか、でなことになって、うん、それはあるかもしれん」
カイバ桶を手にしたまま、セニョールは急に饒舌になって話し続けた。やっぱりよく分からんおじさんだ。

【例２】区切り線や区切り記号の挿入
（『優駿』2003 年 11 月号）



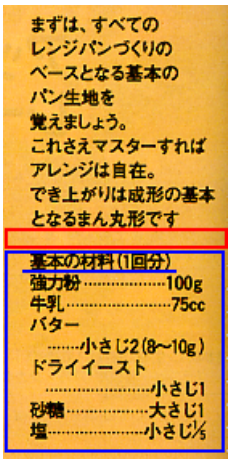
写真班に回ったり、とたばた状態のまま数も数え切れず出発したものだから、カヌーを漕いだ人の総数がわからないのである。正解がわからない算数は、いくらやっても無駄である。
「それじゃ名簿だ、名簿で確認する
んだノ」
防水バッグの中から名簿を取り出してやじろべえのように揺れる船の上で点呼をした。
「うぅぅ～、福島の山崎はいるかあ？　オエェッ」
全身塩だらけ、そして体の芯まで冷えきって、それなのになぜか脱水症状という状態で飛島の中村港に入港した。コンクリートの桟橋の上に疲労が降りた時、堅い大地の素晴らしさを感じる。
「何はともあれ風呂に入ろう！」とカヌー艇民たちはそのまま民宿へとなだれ込んでいった。バトル終了→お風呂→ビールという不動の

【例３】後続文書要素先頭の文字強調
（『BE-PAL』2003 年 11 月号）

■ blockEnd 要素を用いない例

paragraph 要素を除くブロック要素によって，文書の区切れ目を表し得る場合については，blockEnd 要素による区切れ目のマークアップを行わない。よって，ブロック要素の直前直後には，blockEnd 要素は現われない。

例えば，【例 4 】（『ESSE』2003 年 11 月号）では，空行（赤色）によって文書要素の区切れ目が分かるが，空行の次の 1 行（青色下線）が titleBlock 要素となるため，空行より下のブロックは，cluster 要素（青色囲み）となり，必然的に，空行の上のブロックとは切り離される。このような場合，区切れ目と一致する空行を blockEnd 要素によって示すことはしない。



【例４】『ESSE』2003 年 11 月号

形式化例

■ 【例１】『文藝春秋』2003 年 8 月号

```
<paragraph>
と、答えた。秋毫とは、秋になってぬけかわった獣の細い毛をいい、繊芥はわずかなあくたをいう。
</paragraph>
<blockEnd />
<paragraph>
北方の郡国に目をむけたい。
</paragraph>
```

※ sentence 要素の形式化は省略

■ 【例２】『優駿』2003 年 11 月号

```
<paragraph>
そばを通った厩舎のルーキー騎手・長谷川浩大だけが、イガグリ頭を撫でながらサマンサをまぶしそうに見ていた。
</paragraph>
<blockEnd />
<paragraph>
「この際ファンに何か言っておきたいことはないですか？」とあらためてセニョールに聞くと、カイバ桶を片づけながら「ランナ
復活させろってことだな」と言う。
：
</paragraph>
```

※ sentence, quote 要素の形式化は省略

■ 【例３】『BE-PAL』2003 年 11 月号

```
<paragraph>
「転覆隊と一緒に行くと、本当に酷い目に遭いますねぇ…」「二度と誘わないで下さい！」などという、一般参加者の声が聞こえた。
</paragraph>
<blockEnd />
<paragraph>
全身塩だらけ、そして体の芯まで冷えきって、それなのになぜか脱水症状という状態で飛島の中村港に入港した。コンクリートの
桟橋の上に飛び降りた時、堅い大地の素晴らしさを感じる。
</paragraph>
```

※ sentence, quote 要素の形式化は省略

br 要素

概要

- 改行を表す。

形式

■ 要素

- 空要素である。

■ 属性

- *type* (必須)：
  - `automatic_original` ... 論理改行（自動付与されたことを表す）

■ DTD

```
<!ELEMENT br EMPTY>
<!ATTLIST br type (automatic_original) #REQUIRED>
```

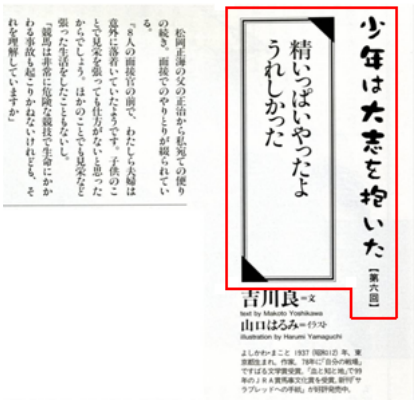
説明

改行を表す。

形式化例

■ 【例１】『優駿』2003 年 11 月号

原資料



形式化

```
<titleBlock>
 <title>
 少年は大志を抱いた【第六回】<br type="automatic_original" />
 精いっぱいやったよ<br type="automatic_original" />
 うれしかった<br type="automatic_original" />
 </title>
</titleBlock>
```

注： sentence 要素の形式化は省略

caption 要素

概要

- figureBlock 要素に含まれる，図表についての タイトルや説明を表す。

形式

■ 要素

br, cluster, info, list, noteBody, paragraph, quotation, rejectedBlock, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT caption
        (br|cluster|info|list|noteBody|paragraph|quotation|rejectedBlock|sentence)*>
```

説明

caption 要素は，figureBlock 要素に含まれ， figure 要素や table 要素として記述された図・写真・絵・表などについての説明を表す。caption 要素に該当するのは，figureBlock 要素の中から，figure 要素・ table 要素を除いた部分である。

【例 1 】における，イラストについての解説文や【例 2 】におけるグラフの説明文 (緑で示した部分) が caption 要素である。

■ caption 要素のタイトル

caption 要素は，

(1) 図表の説明文のみの場合 (【例 1 】)
(2) 図表の説明文と，その説明文に対する見出しがある場合 (【例 2 】)

がある。

このうち，(2) については，タイトルや見出しに相当する部分を titleBlock 要素とし，その titleBlock 要素がまとめる範囲を cluster 要素とする。

【例１】『電撃 GAMECUBE』2003 年 12 月号

	前期日程(人)	後期日程(人)	中期日程(人)	合計(人)
北海道・東北	10,748	3,425	135	14,308
関東・甲信越	19,683	6,986	550	27,219
北陸・東海	9,820	3,575	430	13,825
関西	12,872		611	17,706
中国・四国	10,117	3,394	272	13,783
九州	11,568	3,636	—	15,204
全国合計	74,808	25,239	1,998	102,045
<割合>	74.8%	25.2%	—	—
	73.3%	24.7%	2.0%	—

【例２】『蛍雪時代』2003 年 11 月号

形式化例

■ 【例１】『電撃 GAMECUBE』2003 年 12 月号

```
<figureBlock>
 <figure/>
 <caption>
  <sentence>←赤影といっしょに、鬼を召喚する三角や鳥を呼ぶ鳥寄せも使うと効果的だ。</sentence>
 </caption>
</figureBlock>
```

■ 【例2】『蛍雪時代』2003年11月号

```
<figureBlock>
 <figure/>
 <caption>
   ●１５年一般入試／地区別・日程別募集人員　（表４）
  <cluster>
   <titleBlock>
    <title>
      （注）
    </title>
   </titleBlock>
    ①　１６年「入学者選抜要項」（１５年７月末現在）ベースによる。②　人数は推薦入試、帰国子女、社会人等の特別選抜、専門高校・総合学科卒業生選抜及び産業医大を除く。③　割合の上段は前・後期日程内での割合、下段は総募集人員内での割合。
  </cluster>
 </caption>
</figureBlock>
```

※ caption 要素内の形式化は一部省略

citation 要素

概要

- 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を表す。
- 常に，quotation 要素の子要素として記述される。
- 出典が明らかなもの，書かれたものからの引用であることが明確なものをマークアップの対象とする。
- フィクションにおいては，語り手による「地の文」以外の要素のうち，書かれたものからの引用の体裁を取るものを表す。

形式

■ 要素

authorsData, blockEnd, br, cluster, figureBlock, info, list, noteBody, paragraph, quotation, rejectedBlock, sentence, source, titleBlock, verse

■ 属性

- なし

■ DTD

```
<!ELEMENT citation
          (authorsData|blockEnd|br|cluster|figureBlock|info|list|noteBody|paragraph|
           quotation|rejectedBlock|sentence|source|titleBlock|verse)*>
```

説明

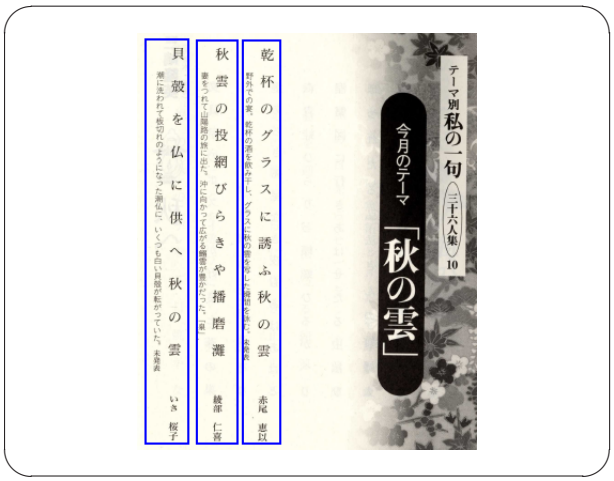
citation 要素は，当該 article 以外の著作物からの引用要素を表す。 article 要素に対応付けられた著者，タイトルによる文章と，それ以外の文章とを区別すること，または，フィクションにおいて地の文とそれ以外の文を区別することをマークアップの目的とする。

citation 要素は，常に quotation 要素の子要素として記述される。改行によって本文と区切られた引用を表す quotation 要素のうち，以下の条件を満たすものを認定する。

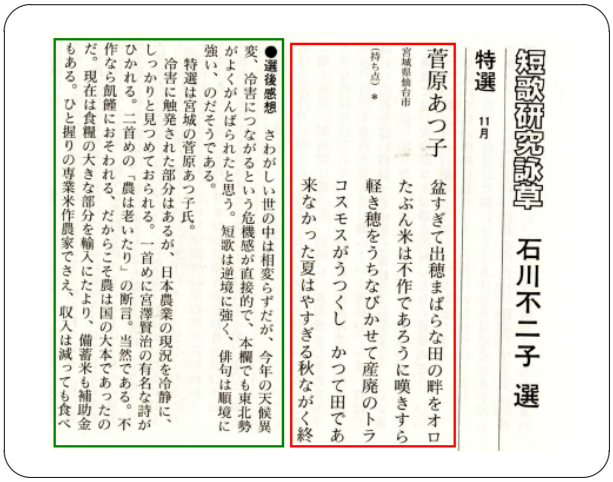
(1) 当該の引用要素に言及する文書要素が同一テキスト内に存在すること
(2) 引用の元となる書記言語による原典（の存在）が明確に示されていること

条件 (1) について，引用要素について言及する本文相当の文書要素が，タイトルを除いて一切存在しない場合は，これを citation 要素とはしない。

例えば，応募された投稿や過去に出版された文献・発表された作品など (以下「原典」と呼ぶ) を元に構成されているような記事の場合，原典を掲載するのみ，あるいは，ある一定の枠組みによって複数の原典を収集，選択，配列するのみであるものは，これを「言及」とみなさず，個々の原典を引用要素としない。この場合，個々の原典は，入れ子の article 要素として記述する。



【例 1】citation 要素とならないもの
(『俳句』2003 年 10 月)



【例 2】citation 要素となるもの
(『短歌研究』2003 年 11 月)

【例 1】においては，タイトル部分を除き，記事を構成しているのは，原典（青囲み）のみであり，原典に対して言及した記事本文は一切ない。この場合，それぞれの原典は引用ではなく，「テーマ別　私の一句 (以下略)」という article 要素を構成する，下位の article 要素となる。

一方，【例 2】においては，タイトル部分の他に，原典（赤囲み）に対して言及した記事本文に相当する要素（緑囲み）が存在している。この場合，原典は citation 要素として記述される。

なお，ここで言う「言及」とは，「○○に……と書いている」のような明示的な表現だけでなく，引用要素そのものを主題として article を構成している場合や，article 要素内部で言及されている事柄について，その関連事項の提示として別の著作物を掲載している場合なども含む。

また，条件 (2) について，原典が明確に示されているとは，以下のような場合を指す。

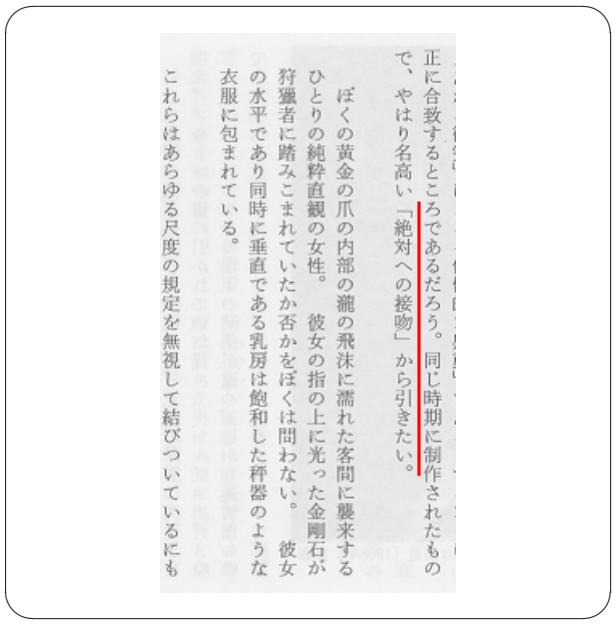
- 記事内に出典情報が示されており，そこからの引用であることが，文脈上明確な場合（【例 3】）
- 一定のスタイルを保った記事であり，表示形態により文献引用であることが容易に分かる場合

なお，citation 要素として本文から切り離されるブロック要素内に，出典情報を示す文書要素が含まれる場合（【例 4】）は，それを source 要素としてマークアップする。source 要素の項を参照のこと。

■ 同一著者の著作物からの引用

citation 要素は，必ずしも著者の別を認定の条件としない。当該の article 要素以外の著作物から引用されたものであれば，citation 要素の著者と article 要素の著者とが同一である場合が生じる。

例えば，【例 5】は「私の代表句」というタイトルから分かるように，著者が自らの作品を挙げて解説を加える記事になっているが，掲載される句（囲み）は，「向こうからとびこんでくる」というタイトルによって包括される article 要素以外の，自らの著作物からの引用であることが，出典情報の提示（傍線）によって明らかである。この場合，掲載句は，citation 要素として記述される。

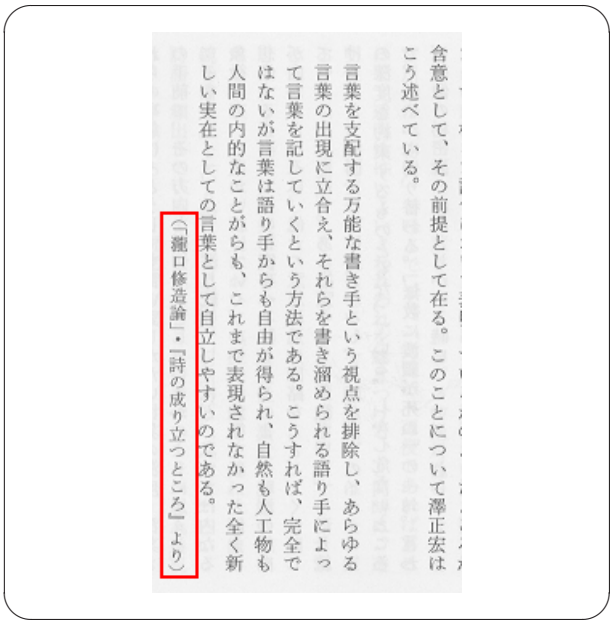


正に合致するところであるだろう。同じ時期に制作されたもので、やはり名高い「絶対への接吻」から引きたい。

ぼくの黄金の爪の内部の瀧の飛沫に濡れた客間に襲来するひとりの純粋直観の女性。彼女の指の上に光った金剛石が狩猟者に踏みこまれていたか否かをぼくは問わない。彼女の水平であり同時に垂直である乳房は飽和した秤器のような衣服に包まれている。

これらはあらゆる尺度の規定を無視して結びついているにも

【例３】出典情報が本文中に示されている例
（『現代詩手帖』2003 年 11 月号）

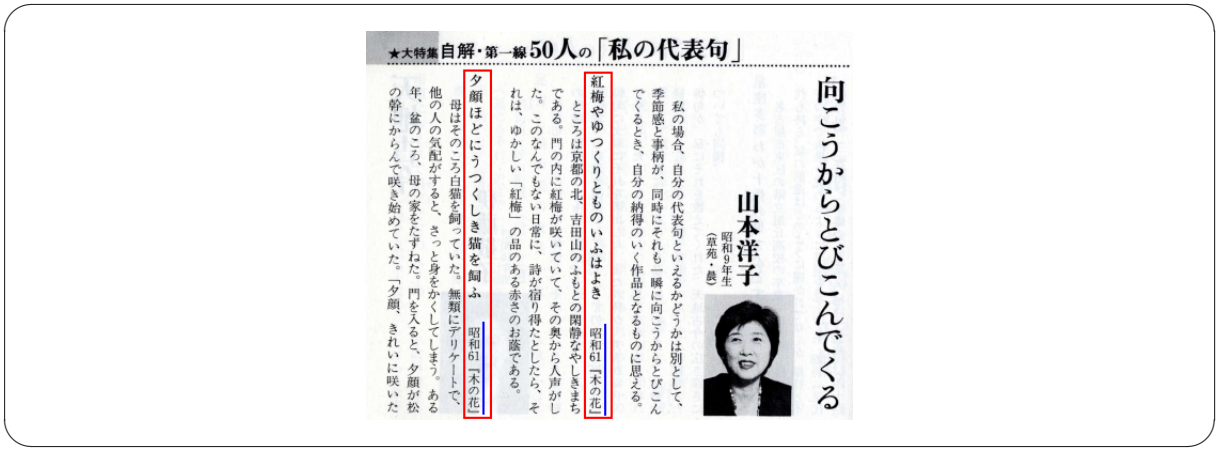


含意として、その前提として在る。このことについて澤正宏はこう述べている。

言葉を支配する万能な書き手という視点を排除し、あらゆる言葉の出現に立合え、それらを書き溜められる語り手によって言葉を記していくという方法である。こうすれば、完全ではないが言葉は語り手からも自由が得られ、自然も人工物も人間の内的なことがらも、これまで表現されなかった全く新しい実在としての言葉として自立しやすいのである。

（「瀧口修造論」・『詩の成り立つところ』より）

【例４】ブロック要素内に出典情報が併記されている例
（『現代詩手帖』2003 年 11 月号）



★大特集 自解・第一線50人の「私の代表句」

向こうからとびこんでくる

山本洋子
昭和9年生
（草苑・蟇）

私の場合、自分の代表句といえるかどうかは別として、季節感と事柄が、同時にそれも一瞬に向こうからとびこんでくるとき、自分の納得のいく作品となるものに思える。

紅梅やゆつくりともの いふはよき　昭和61『木の花』

ところは京都の北、吉田山のふもとの閑静なやしきまちである。門の内に紅梅が咲いていて、その奥から人声がした。このなんでもない日常に、詩が宿り得たとしたら、それは、ゆかしい「紅梅」の品のある赤さのお蔭である。

夕顔ほどにうつくしき猫を飼ふ　昭和61『木の花』

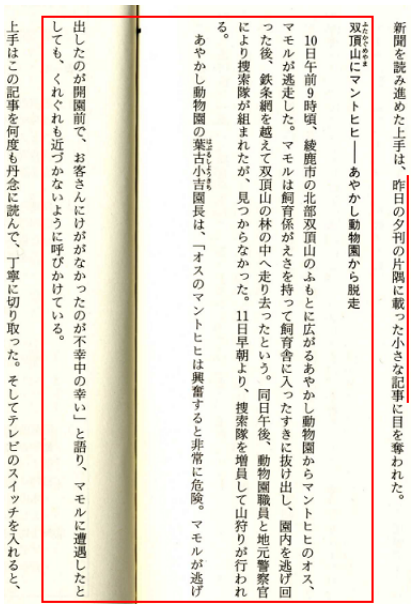
母はそのころ白猫を飼っていた。無類にデリケートで他の人の気配がすると、さっと身をかくしてしまう。ある年、盆のころ、母の家をたずねた。門を入ると、夕顔が松の幹にからんで咲き始めていた。「夕顔、きれいに咲いた

【例５】同一著者の著作物から引用している例
（『俳句』2003 年 10 月号）

■ フィクションにおける引用

フィクションにおいては，以上で規定された，他文献からの引用（例：エピグラフ）の他に，地の文以外の部分を citation 要素とする。具体的には，以下のようなものが，これに相当する。

- 語り手の交代
  - a 作中の手紙
  - b 作中で文献の引用の体裁を取っているもの



新聞を読み進めた上手は、昨日の夕刊の片隅に載った小さな記事に目を奪われた。

双頂山にマントヒヒ――あやかし動物園から脱走

10日午前9時頃、綾鹿市の北部双頂山のふもとに広がるあやかし動物園からマントヒヒのオス、マモルが逃走した。マモルは飼育係がえさを持って飼育舎に入ったすきに抜け出し、園内を逃げ回った後、鉄条網を越えて双頂山の林の中へ走り去ったという。同日午後、動物園職員と地元警察官により捜索隊が組まれたが、見つからなかった。11日早朝より、捜索隊を増員して山狩りが行われる。

あやかし動物園の葉古小吉園長は、「オスのマントヒヒは興奮すると非常に危険。マモルが逃げ出したのが開園前で、お客さんにけががなかったのが不幸中の幸い」と語り、マモルに遭遇したとしても、くれぐれも近づかないように呼びかけている。

上手はこの記事を何度も丹念に読んで、丁寧に切り取った。そしてテレビのスイッチを入れると、

【例６】作中で文献の引用の体裁を取っているものの例
（鳥飼否宇『本格的　死人と狂人たち』2003 年　原書房）

【例６】では，フィクション中で，語り手以外の人物による著作物を引用したという体裁で，新聞記事の内容が示されている。改行および空行によって，地の文と切り離して示された部分（囲み）については，「昨日の夕刊の片隅に載った小さな記事」（傍線部）と出典が明記されおり，citation 要素の認定条件を満たしている。実在する他文献からの引用ではないものの，作中では引用として提示された要素であることが明確であるため，このような要素についても，citation 要素として記述する。

形式化例

■ 【例３】出典情報を含む例（『現代詩手帖』2003 年 11 月号）

```
<paragraph>
（略）…<sentence>このことについて澤正宏はこう述べている。</sentence>
</paragraph>
<quotation>
<citation>
<paragraph>
<sentence>　言葉を支配する万能な書き手という視点を排除し、あらゆる言葉の出現に立合え、それらを書き溜められる語り手によって言葉を記していくという方法である。</sentence><sentence>こうすれば、完全ではないが言葉は語り手からも自由が得られ、自然も人工物も人間の内的なことがらも、これまで表現されなかった全く新しい実在としての言葉として自立しやすいのである。</sentence>
</paragraph>
<source>
<sentence type="quasi">（「瀧口修造論」・『詩の成り立つところ』より）</sentence><br type="automatic_original"/>
</source>
</citation>
</quotation>
```

■ 【例４】 出典情報を含まない例（『現代詩手帖』2003 年 11 月号）

```
<paragraph>
（略）…<sentence>同じ時期に制作されたもので、やはり名高い「絶対への接吻」から引きたい。</sentence>
</paragraph>
<quotation>
<citation>
<paragraph>
<sentence>　ぼくの黄金の爪の内部の瀧の飛沫に濡れた客間に襲来するひとりの純粋直観の女性。</sentence><sentence>彼女の指の上に光った金剛石が狩獵者に踏みこまれていたか否かをぼくは問わない。</sentence><sentence>　彼女の水平であり同時に垂直である乳房は飽和した秤器のような衣服に包まれている。</sentence>
</paragraph>
</citation>
</quotation>
<paragraph>
<sentence>　これらはあらゆる尺度の規定を無視して結びついているにも関わらず、…
```

■ 【例６】 フィクションにおける引用体裁の例（鳥飼否宇『本格的　死人と狂人たち』2003 年　原書房）

```
<paragraph>
（略）…<sentence>　新聞を読み始めた上手は、昨日の夕刊の片隅に載った小さな記事に目を奪われた。</sentence>
</paragraph>
<quotation>
<citation>
<cluster>
<titleBlock>
<title>
<sentence type="quasi"><ruby rubyText="ふたかぐめ">双頂</ruby><ruby rubyText="やま">山</ruby>にマントヒヒ―あやかし動物園から脱走</sentence>
</title>
</titleBlock>
<paragraph>
<sentence>　１０日午前９時頃、綾鹿市の北部双頂山のふもとに広がるあやかし動物園からマントヒヒのオス、マモルが逃走した。<sentence>
                                        （中略）
<sentence>…マモルに遭遇したとしても、くれぐれも近づかないように呼びかけている。</sentence>
</paragraph>
</cluster>
</citation>
</quotation>
<paragraph>
<sentence>　上手はこの記事を何度も丹念に読んで、丁寧に切り取った。</sentence>…
```

※ br 要素は省略

cluster 要素

概要

- 必ず titleBlock 要素を含み，titleBlock 要素が包括する文書要素全体を表す。
- cluster 要素の入れ子構造により，文書の論理的な階層構造を表現する。

形式

■ 要素

abstract, article, authorsData, blockEnd, br, contents, cluster, figureBlock, info, list, noteBody, orphanedTitle, paragraph, profile, quotation, rejectedBlock, sentence, table, titleBlock, verse

■ 属性

- *type* (任意)：cluster 要素の種類

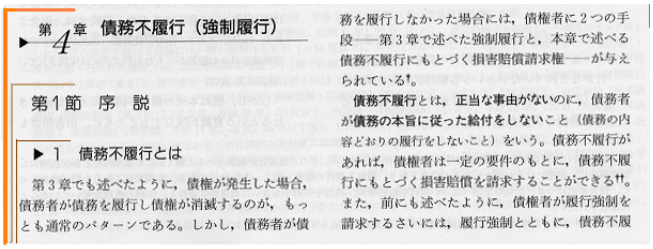
■ DTD

```
<!ELEMENT cluster
        (abstract|article|authorsData|blockEnd|br|contents|cluster|figureBlock|info|
         list|noteBody|orphanedTitle|paragraph|profile|quotation|rejectedBlock|
         sentence|table|titleBlock|verse)*>
```

説明

cluster 要素は，title 要素が代表記述として包括する文書要素の範囲を表したものである。cluster 要素の入れ子構造によって，文書の論理的な階層構造を記述することができる。

cluster 要素によって範囲を記述される文書要素の典型的な例は，論文などにおける章や節である。



第4章　債務不履行（強制履行）

第1節　序　説

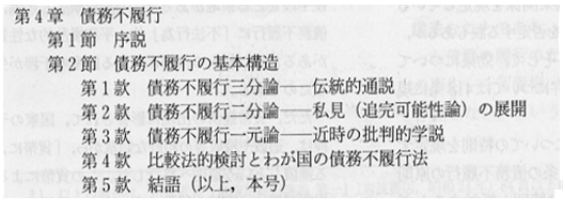
▶1　債務不履行とは

第3章でも述べたように，債権が発生した場合，債務者が債務を履行し債権が消滅するのが，もっとも通常のパターンである。しかし，債務者が債務を履行しなかった場合には，債権者に2つの手段――第3章で述べた強制履行と，本章で述べる債務不履行にもとづく損害賠償請求権――が与えられている†。

債務不履行とは，**正当な事由がないのに，債務者が債務の本旨に従った給付をしないこと**（債務の内容どおりの履行をしないこと）をいう。債務不履行があれば，債権者は一定の要件のもとに，債務不履行にもとづく損害賠償を請求することができる††。また，前にも述べたように，債権者が履行強制を請求するさいには，履行強制とともに，債務不履

【例 1】階層構造を持つ文書（論文）の例（『法学教室』2003 年 11 月号）

例えば，【例 1 】においては，第 4 章，第 1 節，「▶ 1 」がそれぞれ cluster 要素と認定される。それぞれの cluster 要素の包含関係によって，文書構造 (「目次」参照) が示されている。



【参考】例 1 の文書の目次（『法学教室』2003 年 11 月号）

cluster 要素には，必ず titleBlock 要素が一つ含まれることに留意しなければならない。titleBlock 要素の存在しない文書要素のまとまりは，cluster 要素で範囲規定をする対象とはならない。このようなまとまりは，その終端を blockEnd 要素で示す。詳細は，blockEnd 要素のページを参照のこと。

■ cluster 要素の入れ子の認定

cluster 要素の入れ子 (の深さ) には，制限を設けない。titleBlock 要素の存在する文章のまとまりは，どれだけ深い階層になっても，cluster 要素によって範囲を記述する。

ただし，cluster 要素が titleBlock 要素と下位の cluster 要素のみになる場合 (cluster 要素から titleBlock 要素を除いた文書要素が，上位の cluster 要素と下位の cluster 要素で同一になる場合) は，titleBlock 要素と下位の cluster 要素内の titleBlock 要素を合わせて一つの titleBlock 要素とし，入れ子構造としない。典型的なものとしては，連載の記事における連載のタイトルと当該回のタイトルが挙げられる。

【例 2 】では，連載タイトル「愛と涙のドリフト外伝すごいよ！！ オサルさん」と当該回のタイトル「言葉の棒高跳びと石油オバちゃん」が見出せる。表面的な階層構造としては，「愛と涙のドリフト外伝すごいよ！！ オサルさん」をタイトルとするまとまりが第 1 階層目，「言葉の棒高跳びと石油オバちゃん」をまとまりとする箇所が第 2 階層目に見える。

しかし，この例は当該回のタイトルで表される内容で完結しているのであり，連載タイトルは，当該回に対して，内容的に入れ子の関係になっているとは考えにくい。したがって，このような場合は，それぞれの階層の titleBlock 要素を一括して，一つの cluster 要素として統合して記述する。

■ 項立て記号で始まる要素の処理

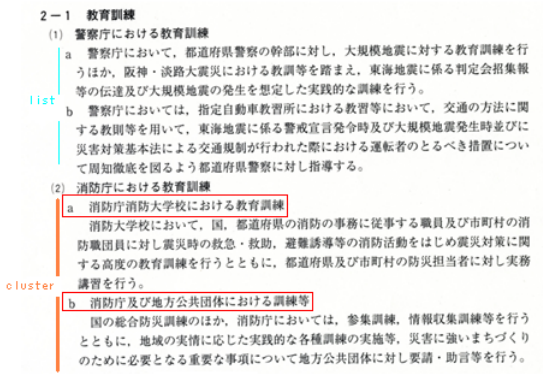
「●」「・」などの記号，または「 1 」「 2 」「 3 」，「イ」「ロ」「ハ」などの順序を表わす文字（以下，これらの記号や文字をラベルと呼ぶ）で始まる要素が列挙されている場合，list 要素とするか cluster 要素とするかで判断が揺れる場合がある。cluster 要素と判断できるのは，以下のような場合である。

- 内部に段落を含む。
- ラベルを含む行の全体が，titleBlock 要素と認定できる（ラベルに加えて，後続する文章に対して見出しとして機能する簡潔な言語表現からなる）。

上記を満たさない場合，ラベルで始まる要素の羅列は，list 要素となる。

【例２】『option』2003 年 11 月号

例えば，【例３】における「２ － １ 教育訓練」の下位の節（１）（２）では，（２）内部のａ，ｂが上記の要件を満たすため，cluster 要素と認定されるのに対して，（１）内部のａ，ｂは上記要件を満たさないため，list 要素となる。



【例３】『防災白書』平成 15 年度版

形式化例

■ 【例１】『法学教室』2003 年 11 月号

```
<cluster>
 <titleBlock><title>
  第４章　債務不履行（強制履行）
 </title></titleBlock>
  <cluster>
   <titleBlock><title>
   第１節　序説
   </title></titleBlock>
    <cluster>
     <titleBlock><title>
     ▲1　債務不履行とは
     </title></titleBlock>
     （略）
    </cluster>
   （略）
  </cluster>
 <cluster>
  <titleBlock><title>
   第２節　債務不履行の基本構造
  </title></titleBlock>
   <cluster>
    <titleBlock><title>
    第１款　債務不履行三分論―伝統的通説
    </title></titleBlock>
    （略）
   </cluster>
  </cluster>
 </cluster>
```

※ paragraph 要素とそれ以下の要素は省略

■ 【例２】『option』2003 年 11 月号

```
<cluster>
 <titleBlock>
 <title>
  愛と涙のドリフト外伝
  すごいよ！！ オサルさん
  「言葉の棒高跳びと石油オバちゃん」の巻
 </title>
 </titleBlock>
                （本文略）
</cluster>
```

※ paragraph 要素とそれ以下の要素は省略

contents 要素

概要

- 目次に相当する文書要素を表す。

形式

■ 要素

cluster, list

■ 属性

- なし

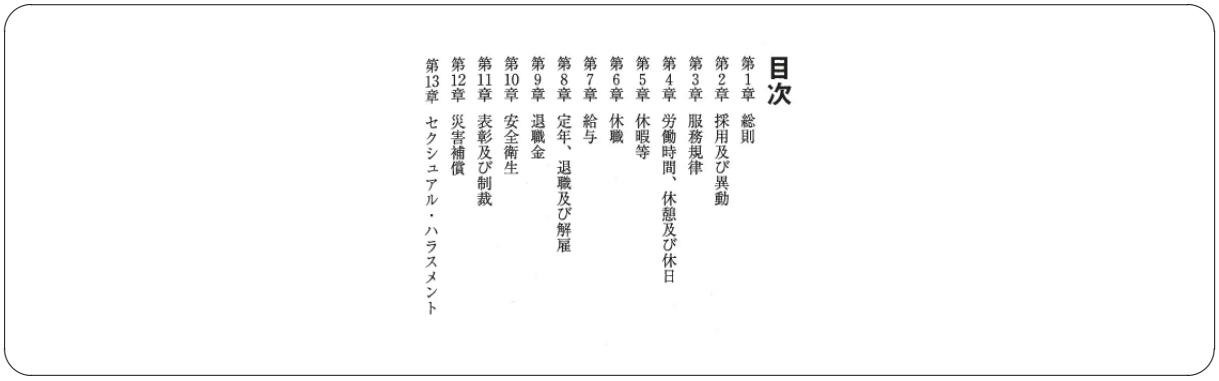
■ DTD

```
<!ELEMENT contents (cluster|list)+>
```

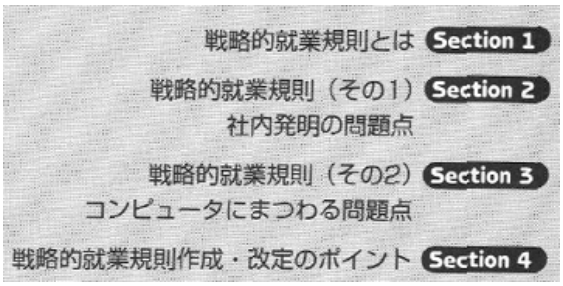
説明

可変長サンプルとしての article 要素より下位の article 要素や cluster 要素の**目次に相当する**文書要素は，contents 要素で表す。次の 2 点のいずれかが，contents 要素を認定する指標となる。

(1)「目次」「もくじ」「ｃｏｎｔｅｎｔｓ」など，目次であることを示すタイトルを持つブロック …【例 1】
(2) その下位に複数ある article や cluster の title 要素をリスト化したもの …【例 2】


目次
第1章　総則
第2章　採用及び異動
第3章　服務規律
第4章　労働時間、休憩及び休日
第5章　休暇等
第6章　休職
第7章　給与
第8章　定年、退職及び解雇
第9章　退職金
第10章　安全衛生
第11章　表彰及び制裁
第12章　災害補償
第13章　セクシュアル・ハラスメント

【例 1】目次を示すタイトルあり（廣岡久生，衛藤寛治著『就業規則の決め方モデル例』ぱる出版 2002 年）



【例２】下位の構造要素のタイトルリスト
（北村庄吾，小林秀星編著『小さな会社のための採用から退職までの実務ができる本』日本実業出版社 2003 年）

形式化例

■ 【例１】目次を示すタイトルあり

```
<contents>
 <cluster>
  <titleBlock>
   <title>
    <sentence type="quasi">目次</sentence>
   </title>
  </titleBlock>
  <list>
   <listItem>
    <sentence type="quasi">第１章　総則</sentence>
   </listItem>
   <listItem>
    <sentence type="quasi">第２章　採用及び異動</sentence>
   </listItem>
   <listItem>
    <sentence type="quasi">第３章　服務規律</sentence>
   </listItem>
    ：(中略)
   <listItem>
    <sentence type="quasi">第１３章　セクシュアル・ハラスメント</sentence>
   </listItem>
  </list>
 </cluster>
</contents>
```

※ br を省略。

■ 【例２】下位の構造要素のタイトルリスト

```
<contents>
 <list>
  <listItem>
   <sentence type="quasi">戦略的就業規則とは</sentence>
   <sentence type="quasi">Ｓｅｃｔｉｏｎ１</sentence>
  </listItem>
  <listItem>
   <sentence type="quasi">戦略的就業規則（その１）</sentence>
   <sentence type="quasi">社内発明の問題点</sentence>
   <sentence type="quasi">Ｓｅｃｔｉｏｎ２</sentence>
  </listItem>
  <listItem>
   <sentence type="quasi">戦略的就業規則（その２）</sentence>
   <sentence type="quasi">コンピュータにまつわる問題点</sentence>
   <sentence type="quasi">Ｓｅｃｔｉｏｎ３</sentence>
  </listItem>
  <listItem>
   <sentence type="quasi">戦略的就業規則作成・改定のポイント</sentence>
   <sentence type="quasi">Ｓｅｃｔｉｏｎ４</sentence>
  </listItem>
 </list>
</contents>
```

※ br を省略。

correction 要素

概要

- 原文の誤植を訂正したことを表す。
- 明らかな誤植のみを対象とし，表記のゆれや知識レベルの誤りは，対象外とする。
- 文字を対象としてタグを付与する。

形式

■ 要素

- *type* 属性の値が `excess` の場合：空要素である。
- *type* 属性の値が `omission` , `erratum` の場合：%character;

■ 属性

- *type* (必須) ： 誤植の種別
  - `omission` ... 脱字
  - `excess` ... 衍字
  - `erratum` ... 誤字
- *originalText* (必須) ： 原文
  - *type* 属性の値が `omission` の場合は，*originalText* 属性不要

■ DTD

```
<!ELEMENT correction (%character;)*>
<!ATTLIST correction type (omission|excess|erratum) #REQUIRED>
<!ATTLIST correction originalText CDATA #REQUIRED>
```

説明

原文の明らかな誤植（組み版上の誤り。ミスプリント）を訂正したことを表す[*4] 。訂正したテキストを要素内容 (本文) とし，訂正前のテキストは *originalText* 属性として示す。

ここで言う「明らかな」誤植とは，原則として，ある語の語形 (発音した時に実現される形態) または表記 (文字として表現された形態) として辞書などに一切掲載されないものを指す。これは，自動形態素解析に適した語形を本文要素とすることを想定しているためである。

[*4] ただし，誤植のうち，古い出版物に見られる文字の回転に関する組み誤りは，タグ付けの対象外とする。

ただし，一般的な辞書に掲載されない語形・表記についても，慣用的に用いられているもの，認知されていると思われるものは，訂正の対象としない*5。

また，そもそも辞書に登録されないような語については，語形や表記を確定する手がかりがないため，訂正の対象としない。

その他，意図的に用いられていることが明らかな語形や表記，また，知識的な誤りの可能性があるものは，そもそも「誤植」の範疇には入らないため，訂正の対象としない。

誤植として扱わないものの例については，「誤植の対象外となるもの」の項で挙げる。

correction 要素として記述される対象は文字である。語や句ではなく，文字を単位としてタグを付与する。

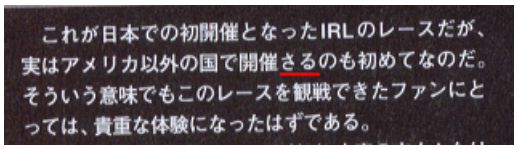
例えば，本来「誤植」とあるべきところを「植誤」とした誤植がある場合は，「植」という文字を「誤」に，「誤」という文字を「植」に訂正するというように，語を単位とするのではなく，文字を単位として訂正を行う。

なお，誤植の種類を以下の 3 種類に分け，*type* 属性によって表す。以下に，それぞれのタイプについて解説する。

- 脱字 ... `omission`
- 衍字 ... `excess`
- 誤字 ... `erratum`

■ 脱字 (omission)

- 明らかに文字（文字列）が脱落しているものを表す。脱落している（挿入すべき）文字を，要素内容とする。

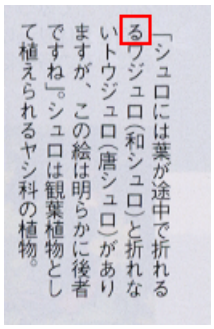


【例 1】脱字の例（『HOBBY JAPAN』2003 年 9 月号）

■ 衍字 (excess)

- 明らかに不要な文字が挿入されているものを表す。不要な文字を *originalText* 属性によって示し，不要な文字があった箇所に空要素タグとして挿入する。
- 誤って挿入された文字によって，同一の文字が複数連続した場合は，先行して現われる文字を本文要素とし，後続する文字を挿入された不要な文字とみなす。

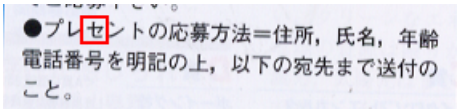
*5 慣用や認知のレベルは様々だが，ある種の慣用が一部で辞書に掲載されるような事象（一部の辞書に「○○の慣用表現」や「○○の誤り」といった形で掲載されるようなもの）については，一様に慣用の可能性があると判断する。例えば，外国語の片仮名表記などでは，外来音を日本語音へ対応付けるルールが不明確であることや，対応付けの際に起こる規則的な発音上の転化があることが知られているが，そこに起因する事象と捉えられる範囲では，訂正対象としない。



【例２】衍字の例（『BE-PAL』11 月号）

■ 誤字 (erratum)

- 明らかに字が誤っているものを表す。訂正後のテキストを要素内容とし，訂正前のテキストを *originalText* 属性によって示す。



【例３】誤字の例（『Newton』2003 年 11 月号）

■ 誤植の対象外となるもの

仮名遣いのゆれや，語の用い方や選択の誤り，知識レベルの誤りなどは，明らかな誤植とは言いがたいため，correction 要素の対象外となる。具体的には，以下のようなものがこれに相当する[*6]。

- 仮名遣いのゆれ，現行の仮名遣いと異なる仮名遣い（『国民生活白書』平成 2 年版）

  今後自動車等の運転にたづさわる高齢者の増加するなかで，高齢者の交通事故の増加が危惧される。

  ※「たづさわる」→「たずさわる」

- 送り仮名のゆれ（『国民生活白書』昭和 53 年版）

  身近かにある既存施設の効率的利用を考えるようになっている。

  ※「身近か」→「身近」

[*6] ただし，著者（または著作権者）からの指摘により，誤植であることが確認された場合に限り， correction 要素を用いることがある。

- 片仮名語表記のゆれ（『環境白書』昭和 58 年版）

> 総合森林レクリェーション・エリアについては，武尊地域について道路等の整備事業を行う。

※「レクリェーション」→「レクリエーション」

- 同音類語の使い分け，同音語の変換ミス（『Ａｎｉｍａｇｅ』2003 年 11 月号）

> 多くのアニメーターやゲームクリエイターを排出しているエンターテインメント総合学園、

※「排出」→「輩出」

- 不適切な助詞の使用（不使用）などの文法的な誤り（『ＤＡＹＴＯＮＡ』2003 年 11 月号）

> 足まわりはランチョの４インチ・リフト・キットを組み込まれている。

※「を」→「が」

- 口語的表現（『環境白書』昭和 61 年版）

> 問題の抜本的解決を図ってくためには，これら施策の充実・強化を図るとともに，

※「図ってく」→「図っていく」

- 語を単位とした置き換え（『環境白書』平成 9 年度版）

> ＣＯ２排出量の内訳を１９９４年（昭和６年）度においてみると、

※「昭和」→「平成」

形式化例

■ 【例１】脱字（『HOBBY JAPAN』2003 年 9 月号）

> これが日本での初開催となったＩＲＬのレースだが、実はアメリカ以外の国で開催さ<correction type="omission" originalText="">れ</correction>るのも初めてなのだ。そういう意味でもこのレースを観戦できたファンにとっては、貴重な体験になったはずである。

■ 【例２】衍字（『BE-PAL』2003 年 11 月号）

> 「シュロには葉が途中で折れ<correction type="excess" originalText="る"/>るワジュロ（和シュロ）と折れないトウジュロ（唐シュロ）がありますが、この絵は明らかに後者ですね」。シュロは観葉植物として植えられるヤシ科の植物。

■ 【例３】誤字（『Newton』2003 年 11 月号）

> ●プレ<correction type="erratum" originalText="セ">ゼ</correction>ントの応募方法＝住所，氏名，年齢，電話番号を明記の上，以下の宛先まで送付のこと。

cursive 要素

概要

- 変体仮名を表す。

形式

■ 要素

%character;

■ 属性

- なし

■ DTD

```
<!ELEMENT cursive (%character;)*>
```

説明

現代通用の平仮名字体とは異なるいわゆる「変体仮名」は，現行の平仮名を入力し，cursive 要素として表現する。

形式化例

■ 【例１】「そば」（立川市内のそば屋の看板）

原資料

形式化

```
\verb|<cursive>そ</cursive><cursive>ば</cursive>|
```

delete 要素

概要

- 著作権者の依頼などを受けて削除した本文要素を表す。

形式

■ 要素

%inlineText;

■ 属性

- *type* (任意) ： 抹消された本文要素の種別
  - `copyright_note_by_author` ... 著作権者の要望による抹消

■ DTD

```
<!ELEMENT delete (%inlineText;)*>
<!ATTLIST delete type (copyright_note_by_author) #IMPLIED>
```

説明

原サンプルの著作権者の要望などにより，伏字で抹消された要素を表す。なお，要素内容は，伏字とした文字数分の「■」とする。

形式化例

■ 【例 1】（作例）

その当時の記録では，重量が<delete>■■■</delete>グラムもあったと伝えられている。

enclosedCharacter 要素

概要

- 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を表す。
- 特定の語の略記号として機能している囲み付きの文字を表す。

形式

■ 要素

%character;, sampling

■ 属性

- *description* (任意)： 連続や参照などのラベルとして機能している場合，囲みの形状などを任意に記述する。

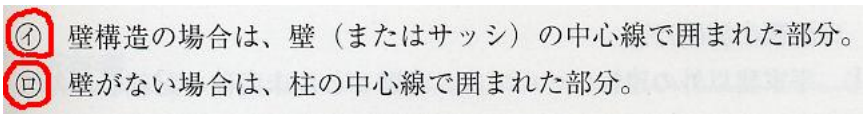
■ DTD

```
<!ELEMENT enclosedCharacter (%character;|sampling)*>
<!ATTLIST enclosedCharacter description CDATA #IMPLIED>
```

説明

■ 連続や参照のラベル

語句や項目などの要素を列挙して示す際に，数字・記号・仮名などの順番を用いて，箇条書きのラベルのように用いることがある。このとき，連続した番号・記号・仮名などが，○や□などで囲まれた形状で実現されているものを，連続のラベルとして機能している囲み文字と呼ぶ。



【例１】連続のラベルとして機能している囲み文字
（日本法令不動産登記研究会編『わかりやすい不動 産登記簿の見方・読み方』日本法令，2003 年）

また，文章中に離れて出現する語句や項目などの要素どうしを参照する際に，同一の番号・記号・仮名などを用いて，参照用のマーカーとすることがある。参照用のマーカーが，番号・記号・仮名などを○や□などで囲んだ形状で実現されている場合，これを参照のラベルとして機能している囲み文字と呼ぶ。

このような連続や参照のラベルとして機能している囲み文字は，enclosedCharacter 要素を用いて表現する。

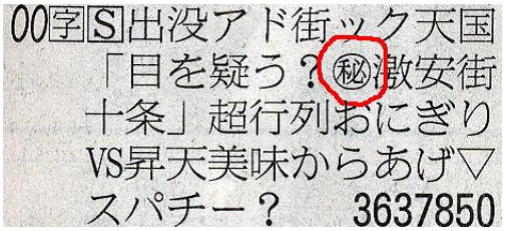
description 属性は，文字囲みの形状などを記述するための属性である。連続・参照ラベルとして機能する囲み文字などのように，囲みの形状によって，指し示されるものを区別する場合に利用する。

例えば，同一の文章の中で，① と❶を区別して用いている場合は，囲みの形状が白い丸なのか黒い丸なのかの情報が必要となる。これらの情報を表わすため，*description* 属性を用いて，囲みの形状を示す。この属性は必要に応じて，任意で用いられるものである。

■ 略記号

特定の語の冒頭 1 文字を取り，これを○や□などで囲み，強調表示をすることで，その語を表す記号として用いるものを略記号と呼び， enclosedCharacter 要素を用いて表現する。略記号の例として，以下のものがあげられる。

- 「秘密」の意味を表す丸囲みの「秘」
- 「編集者」の意味を表す丸囲みの「編」
- 「監督」の意味を表す四角囲みの「監」
- 「日曜日」の意味を表す四角囲みの「日」



【例２】略記号として機能している囲み文字
（『山陽新聞』2006 年 11 月 11 日）

形式化例

■ 【例１】日本法令不動産登記研究会編『わかりやすい不動産登記簿の見方・読み方』日本法令，2003 年

```
<enclosedCharacter description="○">イ</enclosedCharacter>　壁構造の場合は、壁（またはサッシ）の中心線で囲まれた部分。
<enclosedCharacter description="○">ロ</enclosedCharacter>　壁がない場合は、柱の中心線で囲まれた部分。
```

■ 【例２】『山陽新聞』2006 年 11 月 11 日

```
「目を疑う？ <enclosedCharacter>秘</enclosedCharacter>激安街十条」
```

figure 要素

概要

- figureBlock 要素に含まれる図・表・写真・絵などの存在を表す。

形式

■ 要素

- 原則として空要素である。
- ただし，例外として，以下の要素を含む場合がある。
  blockEnd, br, cluster, figureBlock, list, noteBody, paragraph, quotation, rejectedBlock, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT figure (blockEnd|br|cluster|figureBlock|list|noteBody|paragraph|
                  quotation|rejectedBlock|sentence)*>
```

説明

figure 要素は，figureBlock 要素に含まれる図・表・絵・写真など（以下「図表」と呼ぶ）の存在を表すもので，原則的に caption 要素と共に記述される。

図表の存在を示すことを目的とした要素であり，図表の内部に含まれる文字列などは，原則的に表現せず空要素として扱う（図表の内部は入力対象外であり，本文要素とならない）。

ただし，形式的には図表に当てはまらず，通常の cluster 要素，段落，項目などで表現すべきだが，タイトルに「図」「表」「図表」等の表現を含む要素や，タイトルに「図」「表」等の表現を含まないまでも，他の図表付随要素と同様のレイアウトや形式で示されている要素は，空要素ではない figure 要素として表現する。

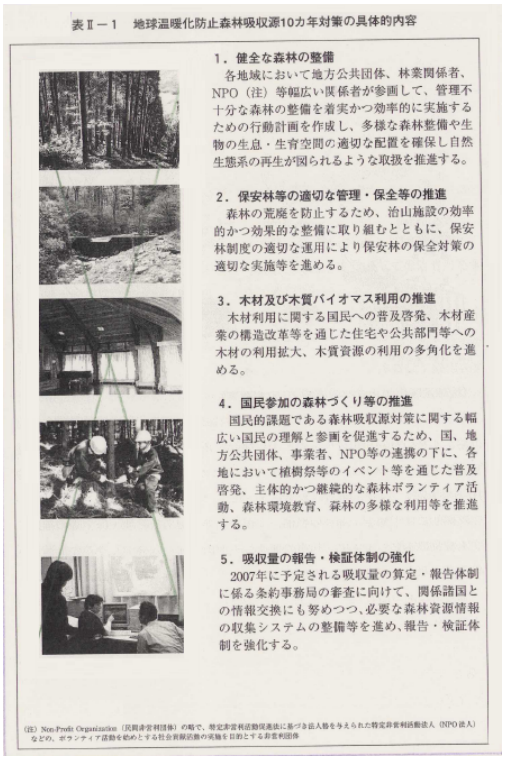
また，figureBlock 要素を構成する付随要素を持つ図表には，一部，内部を入力対象とする表があるが，これは，figure 要素ではなく，table 要素として記述する。

figureBlock 要素，caption 要素，table 要素については，各項を参照のこと。

figure 要素は，図や写真の存在を表わすことを目的としており，サンプル原紙での図表の位置を必ずしも表すものではない。caption 要素に対して，複数の図表が対応する場合であっても，figure 要素は，常に 1 つ，caption 要素に先行して記述される。

【例 1 】において，赤で示したのが figureBlock 要素，緑で示したのが caption 要素，青で示したのが figure 要素である。figureBlock 要素内には，２つの写真が含まれているが，１つの figure 要素として表される。

【例 1】空要素の例：『日経 TRENDY』2003 年 10 月号


表Ⅱ－1　地球温暖化防止森林吸収源10カ年対策の具体的内容

1．健全な森林の整備
各地域において地方公共団体、林業関係者、NPO（注）等幅広い関係者が参画して、管理不十分な森林の整備を着実かつ効率的に実施するための行動計画を作成し、多様な森林整備や生物の生息・生育空間の適切な配置を確保し自然生態系の再生が図られるような取扱を推進する。

2．保安林等の適切な管理・保全等の推進
森林の荒廃を防止するため、治山施設の効率的かつ効果的な整備に取り組むとともに、保安林制度の適切な運用により保安林の保全対策の適切な実施等を進める。

3．木材及び木質バイオマス利用の推進
木材利用に関する国民への普及啓発、木材産業の構造改革等を通じた住宅や公共部門等への木材の利用拡大、木質資源の利用の多角化を進める。

4．国民参加の森林づくり等の推進
国民的課題である森林吸収源対策に関する幅広い国民の理解と参画を促進するため、国、地方公共団体、事業者、NPO等の連携の下に、各地において植樹祭等のイベント等を通じた普及啓発、主体的かつ継続的な森林ボランティア活動、森林環境教育、森林の多様な利用等を推進する。

5．吸収量の報告・検証体制の強化
2007年に予定される吸収量の算定・報告体制に係る条約事務局の審査に向けて、関係諸国との情報交換にも努めつつ、必要な森林資源情報の収集システムの整備等を進め、報告・検証体制を強化する。

（注）Non-Profit Organization（民間非営利団体）の略で、特定非営利活動促進法に基づき法人格を与えられた特定非営利活動法人（NPO法人）などの、ボランティア活動を始めとする社会貢献活動の実施を目的とする非営利団体

【例 2】入力対象の例：『森林・林業白書 平成 14 年度』

【例 2】では，入力対象となる例を示す。「表 II － 1」というタイトルがあるが，内部の記述が図表に該当しない，通常の cluster 要素で表現されるべき形式であるため，figure 要素内部に cluster 構造を示す。

なお，caption 要素を伴わない入力対象外の図表要素は，figureBlock 要素を構成しないため，figure 要素にならない。このような付随する文書要素を一切持たない図表等入力対象外要素は，rejectedBlock 要素（*type* 属性値は，`figure`）によって示すこととする。

rejectedBlock 要素については，当該の項を参照のこと。

形式化例

■ 【例１】『日経 TRENDY』2003 年 10 月号

```
<figureBlock>
 <figure />
 <caption>
  <enclosedCharacter description="○">右</enclosedCharacter>粉末状の光触媒チタンアパタイト
  <enclosedCharacter description="○">上</enclosedCharacter>フィルターに粉末を塗布し、紫外線を
  当てることで細菌などを不活性化する
 </caption>
</figureBlock>
```

※ caption 要素内の形式化は省略。

■ 【例２】『森林・林業白書 平成 14 年度』

```
<figureBlock>
 <figure>
  <cluster>
  １．健全な森林の整備
    各地域において地方公共団体、林業関係者、ＮＰＯ（注）等幅広い関係者が参画して、
  管理不十分な森林の整備を着実かつ効率的に実施するための行動計画を作成し、
  多様な森林整備や生物の生息・生育空間の適切な配置を確保し自然生態系の再生が図られるような取扱を推進する。
   <noteBody>
  （注）Ｎｏｎ‐Ｐｒｏｆｉｔ　Ｏｒｇａｎｉｚａｔｉｏｎ（民間非営利団体）の略で、
    特定非営利活動促進法に基づき法人格を与えられた特定非営利活動法人（ＮＰＯ法人）などの、
    ボランティア活動を始めとする社会貢献活動の実施を目的とする非営利団体
   </noteBody>
  </cluster>
  <cluster>
  ２．保安林等の適切な管理・保全等の推進
    森林の荒廃を防止するため、治山施設の効率的かつ効果的な整備に取り組むとともに、
  保安林制度の適切な運用により保安林の保全対策の適切な実施等を進める。
  </cluster>
       :
       :
  <cluster>
  ４．国民参加の森林づくり等の推進
    国民的課題である森林吸収源対策に関する幅広い国民の理解と参画を促進するため、
  国、地方公共団体、事業者、ＮＰＯ等の連携の下に、各地において植樹祭等のイベント等を通じた
  普及啓発、主体的かつ継続的な森林ボランティア活動、森林環境教育、森林の多様な利用等を推進する。
  </cluster>
  <cluster>
  ５．吸収量の報告・検証体制の強化
    ２００７年に予定される吸収量の算定・報告体制に係る条約事務局の審査に向けて、
  関係諸国との情報交換にも努めつつ、必要な森林資源情報の収集システムの整備等を進め、
  報告・検証体制を強化する。
  </cluster>
 </figure>
 <caption>
  表 II －１　地球温暖化防止森林吸収源１０カ年対策の具体的内容
 </caption>
</figureBlock>
```

figureBlock 要素

概要

- 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を表す。

形式

■ 要素

caption, figure, table

■ 属性

- なし

■ DTD

```
<!ELEMENT figureBlock (caption|figure|table)+>
```

説明

figureBlock 要素は，図表・写真・絵などとそれに付随するタイトルや説明文をまとめた要素を表す。

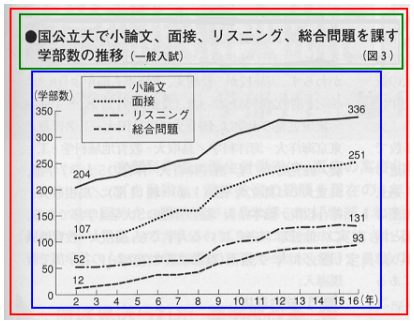
図表・写真・絵など（以下「図表」と呼ぶ）とそのタイトル・説明文（以下「説明文」と呼ぶ）がペアになっており，かつ，図表が主で説明文が従の関係にあるもののみをマークアップの対象とする。

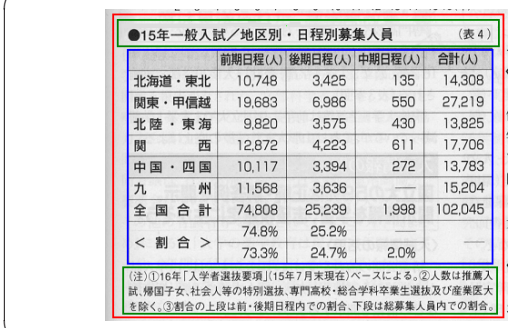
ここで，図表が主で説明文が従の関係にあるものとは，説明文に相当する文書要素の内容を把握したり，article 要素内部における位置づけを捉えたりするためには，図表の存在と，それに付随しているという情報を示すことが欠かせないようなものを指す。

具体的には，次のようなものがこれに該当する。

(1) 写真・図・表のタイトルや注記，解説
(2) 写真に撮られた，または図・絵に描かれた対象の名や，それについての説明

次に，(1)，(2) のそれぞれの例を挙げる。figureBlock 要素を赤，figure 要素を青，caption 要素を緑で示す。





●15年一般入試／地区別・日程別募集人員　（表４）

	前期日程(人)	後期日程(人)	中期日程(人)	合計(人)
北海道・東北	10,748	3,425	135	14,308
関東・甲信越	19,683	6,986	550	27,219
北陸・東海	9,820	3,575	430	13,825
関西	12,872	4,223	611	17,706
中国・四国	10,117	3,394	272	13,783
九州	11,568	3,636	—	15,204
全国合計	74,808	25,239	1,998	102,045
＜割合＞	74.8%	25.2%	—	—
	73.3%	24.7%	2.0%	

（注）①16年「入学者選抜要項」（15年7月末現在）ベースによる。②人数は推薦入試、帰国子女、社会人等の特別選抜、専門高校・総合学科卒業生選抜及び産業医大を除く。③割合の上段は前・後期日程内での割合、下段は総募集人員内での割合。

【例１】『蛍雪時代』2003 年 11 月号

【例２】『日経 TRENDY』2003 年 10 月号

【例１】『蛍雪時代』において緑で囲んだ「●」で始まる要素や「(注)」で始まる要素は，いずれも図や表のタイトルや，表についての説明（注記）である。また，同じく緑で示した【例２】『日経 TRENDY』の丸付き「右」「上」以下の文書要素は，写真の対象物となっている商品についての説明，商品名になっている。

これらの要素は，図表の存在なしに意味を成さない文章であり，図表に付随していることを示さなければ，周囲の文章との繋がりや関係を捉えることができないため，他の cluster 要素や paragraph 要素と区別してマークアップする。

形式化例

■ 【例1】『蛍雪時代』2003年11月号

```
<figureBlock>
 <figure />
 <caption>
    ●国公立大で小論文、面接、リスニング、総合問題を課す学部数の推移（一般入試）（図3）
  </caption>
</figureBlock>
<figureBlock>
 <figure />
 <caption>
    ●15年一般入試／地区別・日程別募集人員（表4）
  （注）
   ①16年「入学者選抜要項」（15年7月末現在）ベースによる。
   ②人数は推薦入試、帰国子女、社会人等の特別選抜、専門高校・総合学科卒業生選抜及び産業医大を除く。
   ③割合の上段は前・後期日程内での割合、下段は総募集人員内での割合。
  </caption>
</figureBlock>
```

※ sentence 要素、caption 要素内 ①〜③の list 要素の形式化は省略

■ 【例2】『日経 TRENDY』2003年10月号

```
<figureBlock>
 <figure />
 <caption>
  <enclosedCharacter description="○">右</enclosedCharacter>粉末状の光触媒チタンアパタイト
  <enclosedCharacter description="○">上</enclosedCharacter>フィルターに粉末を塗布し、紫外線を
  当てることで細菌などを不活性化する
 </caption>
</figureBlock>
```

※ sentence 要素の形式化は省略

fraction 要素

概要

- 帯分数の中の真分数部分を表す。

形式

■ 要素

%inlineText;

■ 属性

- なし

■ DTD

```
<!ELEMENT fraction (%inlineText;)*>
```

説明

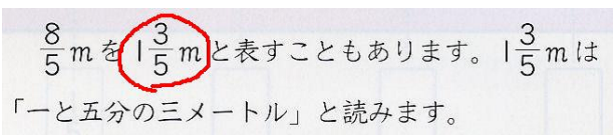
帯分数の中の真分数部分は，分子，「／」，分母の順に入力し， fraction 要素を付与して表現する（分母と分子の区切りが「—」であっても，「／」を用いる）。

なお， fraction 要素は，帯分数を構成する整数部分と真分数部分とを電子化テキストの中で区別して表現するために設けられたタグであり，帯分数以外の分母と分子を「—」や「／」で区切って表現される分数には適応されない。

形式化例

■ 【例１】帯分数 (『新編新しい算数４上』東京書籍，2005 年)

原資料



形式化

８／５mを１<fraction>３／５</fraction>mと表すこともあります。

image 要素

概要

- JISX0213:2004 が規定する諸記号に含まれていない記号類や絵文字などを表す。

形式

■ 要素

- 空要素である。

■ 属性

- *description* (任意)： image 要素で表すものの形状など任意の覚書き
- *no* (必須)： サンプル内での出現番号

■ DTD

```
<!ELEMENT image EMPTY>
<!ATTLIST image description CDATA #IMPLIED>
<!ATTLIST image no CDATA #REQUIRED>
```

説明

JISX0213:2004 が規定する諸記号に含まれていない記号類や絵文字などは，image 要素で表現する。 image 要素で表現するものには，例えば以下のようなものがある。

- 地図記号，天気記号，音楽記号，星座記号などの記号類
- 初心者マーク，ウールマーク，メールマーク，フリーダイヤルマークなどのマーク類
- 携帯メールなどで用いる各種絵文字
- 空欄補充問題などの空欄
- 話者表示としての発話者の似顔絵

image 要素の *no* 属性には，image 要素で表すべき記号類・絵文字のサンプル内での出現番号を記述する。*no* 属性の属性値は，同一のサンプルにおいて，同一の記号類・絵文字が複数回出現しても，それぞれを毎回数える通し番号とする。また，*description* 属性には，記号類・絵文字の形状などを任意に記述する。

形式化例

■ 【例 1】やまみずてんぐ（『補訂版国書総目録』岩波書店，1990 年）

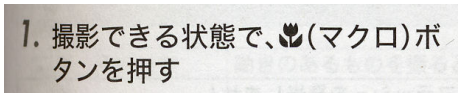
原資料

形式化

```
<image description="やまみずてんぐ" no="1" />
```

■ 【例 2】マクロ（「CaprioR4 使用説明書」株式会社リコー）

原資料



形式化

```
... 状態で<image description="マクロ" no="2" />（マクロ）ボタンを押す
```

info 要素

概要

- 補助的な付与情報を表す。

形式

■ 要素

- 空要素である。

■ 属性

- *arg* (必須)： 付与される情報の種類
- *value* (必須)： 付与される情報の値

■ DTD

```
<!ELEMENT info EMPTY>
<!ATTLIST info arg CDATA #REQUIRED>
<!ATTLIST info value CDATA #REQUIRED>
```

説明

info 要素は，補助的な付与情報を表す。付与される情報の種類は *arg* 属性で指定され，その値は *value* 属性で記述する。

■ article/@isWholeArticle

- 当該の info 要素を包含する直上の article 要素の isWholeArticle 属性に関して，補助的な情報を保持する。
- この種の info 要素は，今後，article 要素の属性として，取り込まれる予定である。
- 例 1 の場合，*arg* 属性により，付与されている情報が「article 要素の種類」であること，また，value 属性によって，「article 要素の種類」が「完結-完全」[*7] であることを示している。

[*7] article 要素が記事全体を完全に含んでいることを表す。

■ **copyright/correction_by_author**

- 著者からの本文訂正依頼の内容を表す。
- 本文自体は修正されない。ただし，correction 要素の適用範囲内であれば，correction 要素として記述され，info 要素は付与されない。

■ **copyright/note_by_author**

- 著者からの依頼のうち，本文訂正依頼以外の注記内容を表す。

形式化例

■ **【例 1】**arg="article/@isWholeArticle"

```
<info arg="article/@isWholeArticle" value="完結-完全"/>
```

■ **【例 2】**arg="copyright/correction_by_author"

- 著者からの依頼内容

```
「切り札」を「決め手」に修正
```

- 結果

```
切り札<info arg="copyright/correction_by_author" value="【切り札】→【決め手】" />
```

■ **【例 3】**arg="copyright/note_by_author"

- 著者からの依頼内容

```
104 ページ 6 行目「アーメン」の直後に「聖書_新共同訳」と表示する。［聖書］と［新共同訳］の間には半角スペース空ける。
```

- 結果

```
アーメン<info arg="copyright/note_by_author" value="出典表示：【聖書 新共同訳】" />
```

list 要素

概要

- 箇条書きなど，列挙された文書要素の集まりを表す。

形式

■ 要素

listItem

■ 属性

- なし

■ DTD

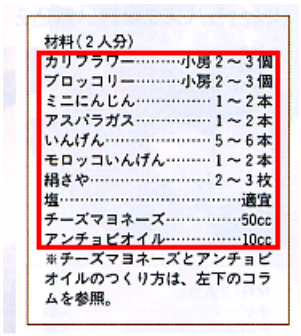
```
<!ELEMENT list (listItem)+>
```

説明

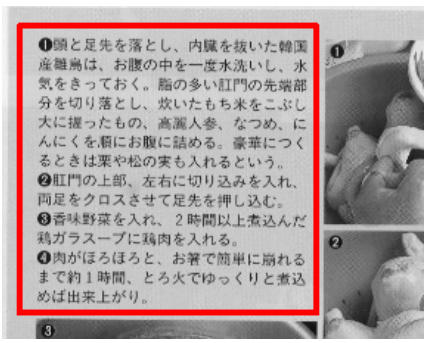
list 要素は，箇条書きなど，列挙された文書要素の集まりを表す。

箇条書きなどの文書要素は，通常の文の連続における文相互の関係と異なり，列挙された個々の文書要素の独立性が高い。さらに，(1), (2), (3) のような順序表現や箇条書きの入れ子など，通常の文の連続とは異なる文書構造を持つ。このような，通常の文との差異を表現するために，個別の文書構造として，形式化した。

list 要素は，改行によって区切られている二つ以上の listItem 要素からなり，list 要素に含まれる個々の文書要素 (列挙される文書要素) は，listItem 要素で記述する（【例１】【例２】参照)。



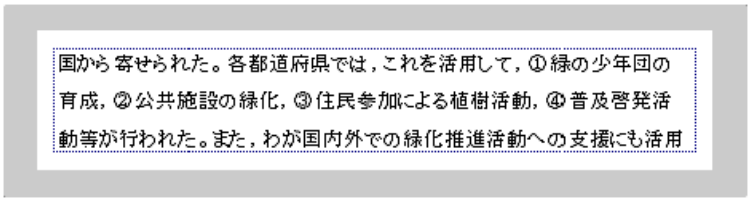
【例１】各項目が改行で区切られている例
(『dancyu』2003 年 8 月号)



【例２】順序表現を先頭に持ち各項目が改行で区切られている例 (『dancyu』 2003 年 8 月号)

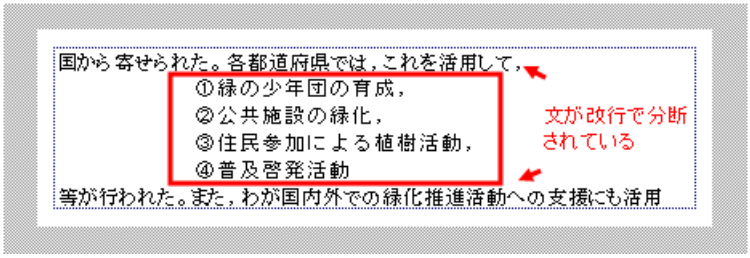
■ 文の中に含まれている項目

文の中に含まれている項目は，原則として list 要素では表さない。



【例 3】 list 要素では表さない例 (『林業白書 平成 10 年度版』)

ただし，項目の前後に改行があって文が断絶されているときに限り， list 要素と認定することができる。



【例 4】 list 要素で表す例

■ list 要素と table 要素（表）

list 要素と似たものに table 要素がある。次のものは，list 要素ではなく table 要素とする。

- 3 列以上のもの
- 列ラベルや行ラベルがあるもの

詳細は，table 要素を参照のこと。

■ list 要素の入れ子

list 要素は幾重にも入れ子になることができる。

形式化例

■ 【例１】『dancyu』2003 年 8 月号

```
<cluster>
<titleBlock>
 <title>
  <sentence type="quasi">材料（２人分）</sentence>
 </title>
</titleBlock>
 <list>
   <listItem><sentence type="quasi">カリフラワー…小房２～３個</sentence></listItem>
   <listItem><sentence type="quasi">ブロッコリー…小房２～３個</sentence></listItem>
   <listItem><sentence type="quasi">ミニにんじん…１～２本</sentence></listItem>
   <listItem><sentence type="quasi">アスパラガス…１～２本</sentence></listItem>
   <listItem><sentence type="quasi">いんげん…５～６本</sentence></listItem>
   <listItem><sentence type="quasi">モロッコいんげん…１～２本</sentence></listItem>
   <listItem><sentence type="quasi">絹さや…２～３枚</sentence></listItem>
   <listItem><sentence type="quasi">塩…適宜</sentence></listItem>
   <listItem><sentence type="quasi">チーズマヨネーズ…５０ｃｃ</sentence></listItem>
   <listItem><sentence type="quasi">アンチョビオイル…１０ｃｃ</sentence></listItem>
 </list>
 <sentence>※チーズマヨネーズとアンチョビオイルのつくり方は、左下のコラムを参照。</sentence>
</cluster>
```

■ 【例２】『dancyu』2003 年 8 月号

```
<list>
  <listItem>
   <sentence>①　頭と足先を落とし、内臓を抜いた韓国産雛鳥は、お腹の中を一度水洗いし、水気をきっておく。
   </sentence><sentence>脂の多い肛門の先端部分を切り落とし、炊いたもち米をこぶし大に握ったもの、高麗人参、なつめ、
   にんにくを順にお腹に詰める。</sentence><sentence>豪華につくるときは栗や松の実も入れるという。</sentence>
  </listItem>
  <listItem>
   <sentence>②　肛門の上部、左右に切り込みを入れ、両足をクロスさせて足先を押し込む。</sentence>
  </listItem>
  <listItem>
   <sentence>③　香味野菜を入れ、２時間以上煮込んだ鶏ガラスープに鶏肉を入れる。</sentence>
  </listItem>
  <listItem>
   <sentence>④　肉がほろほろと、お箸で簡単に崩れるまで約１時間、とろ火でゆっくりと煮込めば出来上がり。</sentence>
  </listItem>
</list>
```

■ 【例４】list 要素で表す例

```
…<sentence type="quasi">各都道府県では，これを活用して，</sentence>
 <list>
  <listItem><sentence type="quasi">①　緑の少年団の育成，</sentence></listItem>
  <listItem><sentence type="quasi">②　公共施設の緑化，</sentence></listItem>
  <listItem><sentence type="quasi">③　住民参加による植樹活動，</sentence></listItem>
  <listItem><sentence type="quasi">④　普及啓発活動</sentence></listItem>
 </list>
<sentence>等が行われた。</sentence>
```

listItem 要素

概要

- list 要素を構成する各並立要素を表す。

形式

■ 要素

blockEnd, br, cluster, figureBlock, list, noteBody, paragraph, quotation, rejectedBlock, sentence, table

■ 属性

- なし

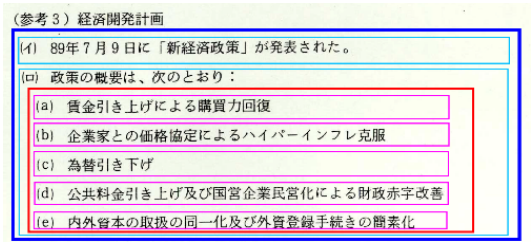
■ DTD

```
<!ELEMENT listItem
          (blockEnd|br|cluster|figureBlock|list|noteBody|paragraph|quotation|
           rejectedBlock|sentence|table)*>
```

説明

箇条書きや名詞句の羅列などが集まったものが list 要素だが，これを構成する各並立要素は，listItem 要素で表現する。listItem 要素は，list 要素の内部の個々の項目にあたり， listItem 要素の直上は list 要素でなくてはならない。

listItem 要素には，list 要素を含むことができる（list の入れ子）。【例 1 】における「（ロ）政策の概要は、次のとおり：」で始まる listItem 要素（水色）には，下位の list 要素（赤）が含まれる。



【例１】『我が国の政府開発援助』 1989 年版（下巻（国別実績）

形式化例

■ 【例１】『我が国の政府開発援助』 1989 年版（下巻（国別実績））

```
<cluster>
 <titleBlock>
  <title>
   （参考３）経済開発計画
  </title>
 </titleBlock>
 <list>
  <listItem>
（イ）８９年７月９日に「新経済政策」が発表された。
  </listItem>
  <listItem>
（ロ）政策の概要は、次のとおり：
    <list>
     <listItem>
        （ａ）賃金引き上げによる購買力回復
     </listItem>
    <listItem>
        （ｂ）企業家との価格協定によるハイパーインフレ克服
     </listItem>
     <listItem>
        （ｃ）為替引き下げ
     </listItem>
     <listItem>
        （ｄ）公共料金引き上げ及び国営企業民営化による財政赤字改善
     </listItem>
     <listItem>
        （ｅ）内外資本の取扱の同一化及び外貨登録手続きの簡素化
     </listItem>
    </list>
   </listItem>
  </list>
 </cluster>
```

※ sentence 要素の形式化は省略

missingCharacter 要素

概要

- 「〓 (ゲタ記号)」で表現されている文字が，JISX0213:2004 で規定されている文字以外の文字 (JIS 外字) であることを示す。
- 属性によって，文字種属性，Unicode 番号などの字体情報を記述する。

形式

■ 要素

%character;

■ 属性

- *attribute* (必須)： 文字種属性
  - `HanIdeograph` ... 漢字
  - `Hiragana` ... 平仮名
  - `Katakana` ... カタカナ
  - `RomanNumeral` ... ローマ数字
  - `Latin` ... ラテン文字
  - `Greek` ... ギリシア文字
  - `OldHanzi` ... 古代中国文字
- *unicode* (必須)：Unicode4.0 の 16 進コード
  - `U+*****` (4〜5 桁の Unicode の先頭に「U+」を加えた 6〜7 桁の文字列)
  - Unicode 外字の場合は「U+FFFD（REPLACEMENT CHARACTER)」を記述する。
- *daikanwa* (任意)： 『大漢和辞典』の親字番号
  - `M*****` (5 桁の諸橋番号の先頭に「M」を加えた 6 桁の文字列)
  -『大漢和辞典』にない漢字は「M99999」を記述する。
  - *attribute* が `HanIdeograph` となる場合のみ必須 (漢字以外の文字種の場合は記述しない)
- *ref* (任意)： Unicode あるいは『大漢和辞典』にない文字の場合，参照する管理番号
  - `KC****` (4 桁の管理番号の先頭に「KC」を加えた 6 桁の文字列)
- *description* (任意)： 字体記述，属性記述など任意の覚書き

■ DTD

```
<!ELEMENT missingCharacter (%character;)*>
<!ATTLIST missingCharacter attribute
      (Greek|HanIdeograph|Hiragana|Katakana|Latin|OldHanzi|RomanNumeral) #REQUIRED>
<!ATTLIST missingCharacter unicode CDATA #REQUIRED>
<!ATTLIST missingCharacter daikanwa CDATA #IMPLIED>
<!ATTLIST missingCharacter ref CDATA #IMPLIED>
<!ATTLIST missingCharacter description CDATA #IMPLIED>
```

説明

コーパスの入力には，JISX0213:2004 で規定されている文字を使用する。しかし，JISX0213:2004 で規定されている文字だけで，現代日本語文に出現する文字を網羅しているわけではない。そこで，JISX0213:2004 で規定されていない文字は，「〓」を入力し， missingCharacter 要素を付与して表現する（「〓」で表現される JIS 外字は，主として漢字（中国簡体字を含む）が想定される）。

なお，JISX0213:2004 で規定されていない記号や絵文字などは，すべて image 要素を用いて表現する。

形式化例

■ 【例 1】Unicode で表現でき，かつ『大漢和辞典』に掲載されている JIS 外漢字 （秋吉久紀夫編訳『現代中国少数民族詩集』土曜美術社出版販売，2003 年）

原資料

形式化

```
<missingCharacter attribute="HanIdeograph" unicode="U+5AEB"
            daikanwa="M06673" description="女偏に莫">〓</missingCharacter>
```

■ 【例 2】Unicode で表現でき，かつ『大漢和辞典』に掲載されていない JIS 外漢字（遠藤紹徳『早わかり中国簡体字』国書刊行会，1986 年）

原資料

形式化

```
<missingCharacter attribute="HanIdeograph" unicode="U+5E86"
                  daikanwa="M99999" description="慶の簡体字">〓</missingCharacter>
```

■ 【例 3】Unicode になく，かつ『大漢和辞典』に掲載されている JIS 外漢字

形式化

```
<missingCharacter attribute="HanIdeograph" unicode="U+FFFD"
                  daikanwa="M*****" description="">〓</missingCharacter>
```

■ 【例 4】Unicode になく『大漢和辞典』にもない JIS 外漢字 （多賀城市内の案内板）

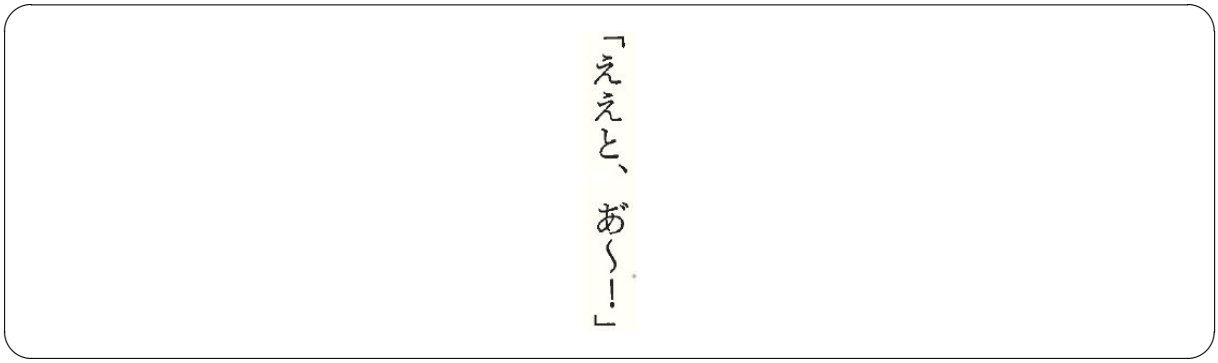
原資料

形式化

```
<missingCharacter attribute="HanIdeograph" unicode="U+FFFD"
                  daikanwa="M99999" ref="KC****"
                  description="脇と巾の合字">〓</missingCharacter>
```

■ 【例５】平仮名「濁点付き平仮名あ」　（日明恩『埋み火』講談社，2005 年）

原資料

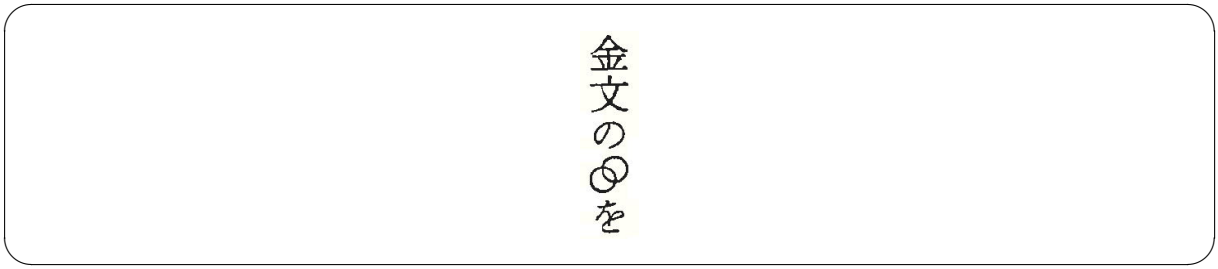


形式化

```
<missingCharacter attribute="Hiragana" unicode="U+FFFD"
                  ref="KC****" description="濁点付き平仮名あ">〓</missingCharacter>
```

■ 【例６】古代中国文字（金文）　（白川静『白川静著作集別巻２金文通釈』平凡社，2004 年）

原資料



形式化

```
<missingCharacter attribute="OldHanzi" unicode="U+FFFD"
                  ref="KC****" description="金文">〓</missingCharacter>
```

noteBody 要素

概要

- 脚注，後注など，本文と区別して記述される注記を表す。
- noteMarker，noteBodyInline 要素も参照のこと。

形式

■ 要素

blockEnd, br, figureBlock, info, list, paragraph, profile, quotation, rejectedBlock, sentence, verse

■ 属性

- なし

■ DTD

```
<!ELEMENT noteBody
        (blockEnd|br|figureBlock|info|list|paragraph|profile|quotation|
         rejectedBlock|sentence|verse)*>
```

説明

noteBody 要素は，脚注，後注など，本文と区別して記述される注記を表す。ただし，傍注[*8]，ママ注[*9]については，noteBodyInline 要素で記述する。この点については，noteBodyInline 要素を参照のこと。

一つの注記は，一つの noteBody 要素に対応させる。例えば，【例 1 】には，19)，20) と二つの注記があるが，それぞれの注記を別々の noteBody 要素として記述する。なお，19)，20) など注記の先頭につけられる注記識別記号も noteBody 要素に含めるものとする。

noteBody 要素の記述位置は，次のようにする。

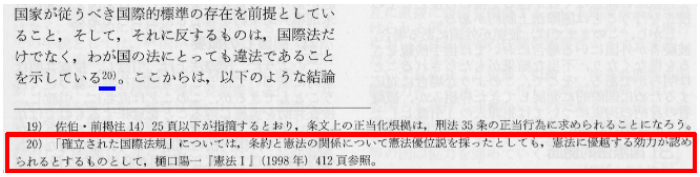
- 基本的には，注記される対象を含んでいる paragraph 要素の直後，もしくは，cluster 要素の末尾に記述する。
  - 注記される対象が citation 要素に含まれ，引用文に対して引用者が新たに付した注である場合は，引用文とは別個の内容と考えられるため，citation 要素の直後に記述する。
- 候補となる paragraph, cluster 要素が複数存在する場合は，階層的に最も近い要素を選択して記述する。

[*8] 「傍注」とは，本文のわきに書き添えた注釈のことである。
[*9] 「ママ注」とは，内容に疑問があるにもかかわらず，原文をそのまま引用したことを示すための注記である。

形式化例

■ 【例１】『法学教室』 2003 年 11 月号

原資料


国家が従うべき国際的標準の存在を前提としていること，そして，それに反するものは，国際法だけでなく，わが国の法にとっても違法であることを示している[20]。ここからは，以下のような結論

19) 佐伯・前掲注 14) 25 頁以下が指摘するとおり，条文上の正当化根拠は，刑法 35 条の正当行為に求められることになろう。
20) 「確立された国際法規」については，条約と憲法の関係について憲法優位説を採ったとしても，憲法に優越する効力が認められるとするものとして，種口陽一『憲法 I』(1998 年) 412 頁参照。

形式化

```
<paragraph>
国家が従うべき国際的標準の存在を前提としていること，そして，
それに反するものは，　国際法だけでなく，わが国の法にとっても
違法であることを示している<noteMarker text="２０）" />。
...
</paragraph>
<noteBody>
２０）　「確立された国際法規」については，条約と憲法の関係に　...
</noteBody>
```

※ sentence 要素と noteBody 要素内の形式化は省略

noteBodyInline 要素

概要

- 傍注やママ注[*10] など，行外に付随する形式で現れる注記を表す。
- noteBody，noteMarker 要素も参照のこと。

形式

■ 要素

- 空要素である。

■ 属性

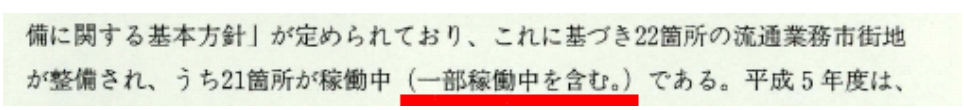
- *text* (必須)：注記
- *info* (任意)：注記自体に付与されるべきタグの情報

■ DTD

```
<!ELEMENT noteBodyInline EMPTY>
<!ATTLIST noteBodyInline text CDATA #REQUIRED>
<!ATTLIST noteBodyInline info CDATA #IMPLIED>
```

説明

傍注やママ注など，行外に付随する形式[*11]で現れる注記を表する。ただし，【例 1 】のように，行外に付随する形式ではなく，行中に含まれる形式の注記は，noteBodyInline 要素として認定しない。また，脚注や後注などブロック要素として表現される注記は，noteBody 要素で記述する。



【例 1 】『建設白書』平成 5 年版

[*10] 「ママ注」とは，内容に疑問があるにもかかわらず，原文をそのまま引用したことを示すための注記である。
[*11] 横書きテキストの本文行の上下にルビ状に付される場合などが想定される。

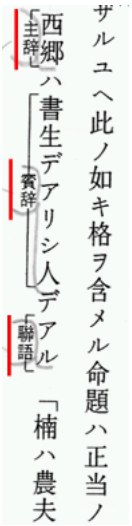
注記自体の文字列は，*text* 属性値とする。当該文字列になんらかのタグ付けをする必要がある場合は，次の規則に基づき，*info* 属性に記述する。*info* 属性の実例は，noteMarker 要素の例 5 を参照されたい。

- タグ名： 元のタグの要素名
- 属性： 属性名と属性値を「=」で繋いで表示。複数の属性を持つ場合は「,」区切りで表示
- 範囲情報： タグ開始文字位置 (x) と包含文字数 (y) を「-」で区切って表示する。文字位置は，「○」を文字として「1 ○ 2 ○ 3 ○ 4」のように数える。包含文字数は，空要素タグの場合 0 とする。タグ範囲が *text* 属性文字列と完全に一致する場合は省略する。

形式化例

■ 【例２】傍注 (『清沢満之全集３』2003 年１月 岩波書店)

原資料

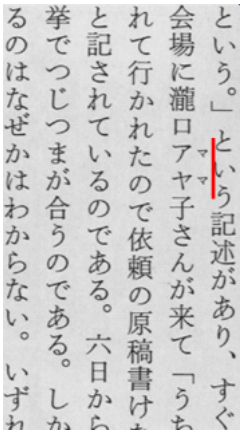


形式化

```
西郷<noteBodyInline text="主辞" />ハ
書生デアリシ人<noteBodyInline text="賓辞" />デ
アル<noteBodyInline text="聯語" />
```

■ 【例 3】ママ注 (『現代詩手帖』 2003 年 11 月号)

原資料



形式化

```
会場に瀧口アヤ<noteBodyInline text="ママ" />子さんが来て
```

noteMarker 要素

概要

- 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（例： † や *1）を表す。
- noteBody 要素も参照のこと。

形式

■ 要素

- 空要素である。

■ 属性

- *text* (必須)： 当該文字列
- *info* (任意)： 当該文字列自体に付与されるべきタグの情報

■ DTD

```
<!ELEMENT noteMarker EMPTY>
<!ATTLIST noteMarker text CDATA #REQUIRED>
<!ATTLIST noteMarker info CDATA #IMPLIED>
```

説明

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（例： † や *1）を表す。主として，行外に付随する参照用の文字列を noteMarker 要素として認定する。

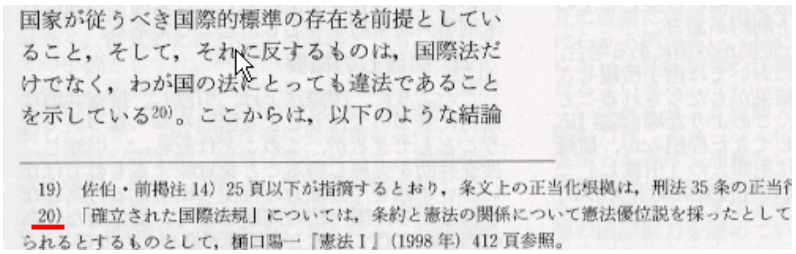
- 注の参照用の番号（→【例３】）
- 参考文献の参照用の番号（→【例４】）

したがって，【例１】のように，行外に付随するのではなく，行中に含まれている文字列や【例２】のように，参照先の参考文献や注の冒頭に付与される文字列は，noteMarker 要素とはしない。

参考文献１）を参照のこと。

新たな方式が開発されている（注１）。

【例１】 noteMarker **要素とはしない例**


国家が従うべき国際的標準の存在を前提としている
ること，そして，それに反するものは，国際法だ
けでなく，わが国の法にとっても違法であること
を示している[20]。ここからは，以下のような結論

19) 佐伯・前掲注 14) 25 頁以下が指摘するとおり，条文上の正当化根拠は，刑法 35 条の正当
20) 「確立された国際法規」については，条約と憲法の関係について憲法優位説を採ったとして
られるとするものとして，樋口陽一『憲法 I』(1998 年) 412 頁参照。

【例 2】『法学教室』 2003 年 11 月号

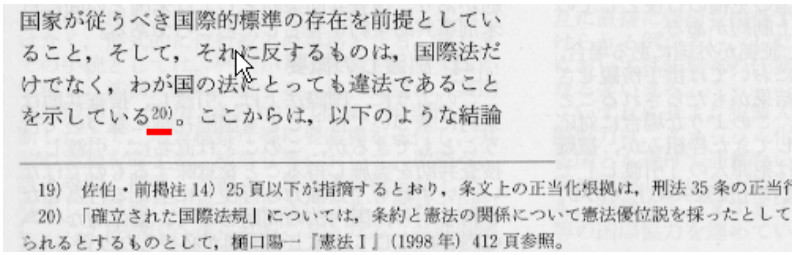
注番号など当該の文字列は，*text* 属性値とする。当該文字列になんらかのタグ付けをする必要がある場合は，次の規則に基づき，*info* 属性に記述する。例 5 に実例を示す。

- タグ名： 元のタグの要素名
- 属性： 属性名と属性値を「=」で繋いで表示。複数の属性を持つ場合は「,」区切りで表示
- 範囲情報： タグ開始文字位置 (x) と包含文字数 (y) を「-」で区切って表示する。文字位置は，「○」を文字として「1 ○ 2 ○ 3 ○ 4」のように数える。包含文字数は，空要素タグの場合 0 とする。タグ範囲が *text* 属性文字列と完全に一致する場合は省略する。

形式化例

■ 【例 3】脚注の参照の目印 (『法学教室』 2003 年 11 月号)

原資料


国家が従うべき国際的標準の存在を前提としている
ること，そして，それに反するものは，国際法だ
けでなく，わが国の法にとっても違法であること
を示している[20]。ここからは，以下のような結論

19) 佐伯・前掲注 14) 25 頁以下が指摘するとおり，条文上の正当化根拠は，刑法 35 条の正当
20) 「確立された国際法規」については，条約と憲法の関係について憲法優位説を採ったとして
られるとするものとして，樋口陽一『憲法 I』(1998 年) 412 頁参照。

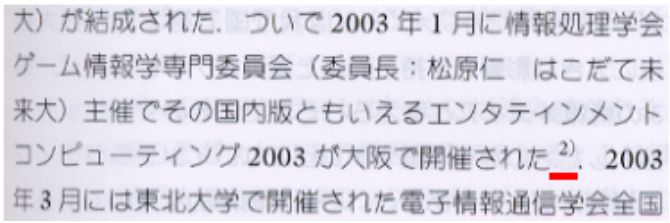
形式化

```
国家が従うべき国際的標準の存在を前提としていること，そして，それに反す
るものは，　国際法だけでなく，わが国の法にとっても違法であることを示して
いる<noteMarker text="２０）" />。
```

※ sentence 要素，br 要素を省略。以下同じ。

■ 【例４】参考文献の参照 (『情報処理』2003 年 44 巻 8 号)

原資料


大）が結成された．ついで 2003 年 1 月に情報処理学会
ゲーム情報学専門委員会（委員長：松原仁　はこだて未
来大）主催でその国内版ともいえるエンタテインメント
コンピューティング 2003 が大阪で開催された[2)]．2003
年 3 月には東北大学で開催された電子情報通信学会全国

形式化

```
その国内版ともいえるエンタテインメントコンピューティング２００３が大阪で開催された<noteMarker text="２）" />.
```

■ 【例５】 info 属性の例（enclosedCharacter の例） (木村宏恒著「インドネシア現代政治の構造」)

形式化

```
出身母体の利害を代弁する浸透機関となった<noteMarker text="１" info="enclosedCharacter:description=○:" />。
```

orphanedTitle 要素

概要

- 不特定範囲の文書要素を代表する記述を表す。

形式

■ 要素

br, noteBody, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT orphanedTitle (br|noteBody|sentence)*>
```

説明

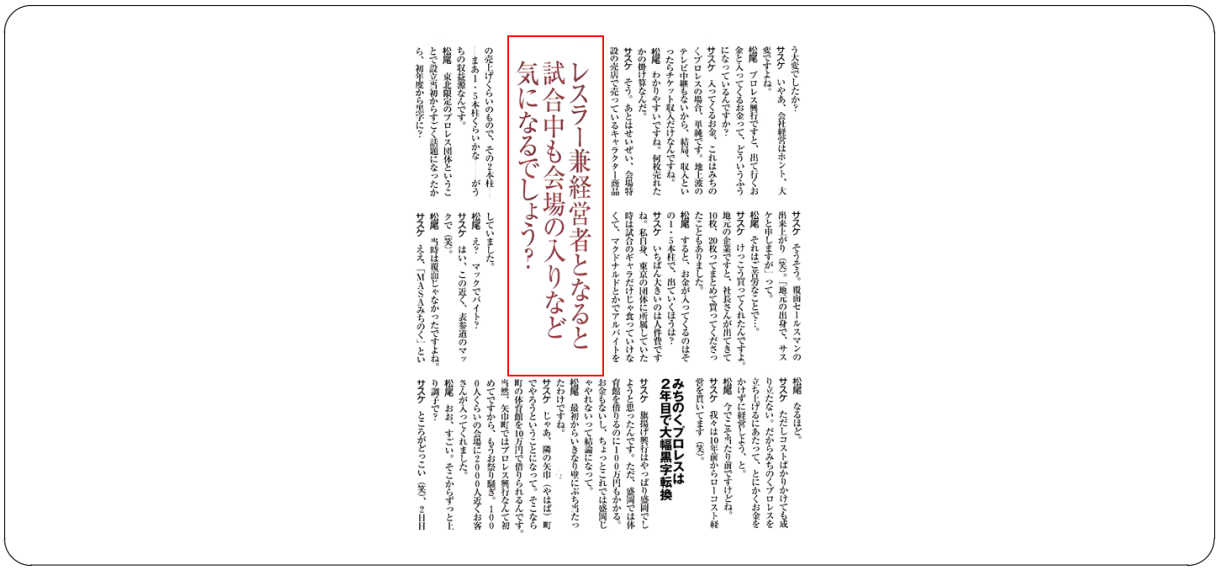
不特定範囲の文書要素を代表する記述 (タイトル) を表す。

orphanedTitle 要素は，cluster 要素中の titleBlock 要素とは異なり，代表する文書要素の範囲が明確でない場合に用いられる。したがって，titleBlock 要素のように，章・節といった，明示的な範囲を持つ文書階層構造に対する代表記述となることはできない。

典型的な例としては，本文から重要な文を抜きだして，本文よりも大きな文字で表示するような例が挙げられる。

例えば，例 1 の場合，「レスラー兼経営者となると試合中も会場の入りなど気になるでしょう？」(囲み部分) が orphanedTitle 要素となる。この orphanedTitle 要素は，本文中では重要な記述である。しかし，cluster 要素中の titleBlock 要素 (「みちのくプロレスは 2 年目で大幅黒字転換」) と異なり，どの文書要素の代表的記述となっているのかが明確でない。

orphanedTitle 要素の記述位置は，orphanedTitle 要素に対して最も内容的に近い cluster 要素 ，もしくは，article 要素の任意の場所とする。



レスラー兼経営者となると試合中も会場の入りなど気になるでしょう？

【例１】『ダイアモンド ザイ』 2003 年 9 月号

形式化例

■ 例１：『ダイアモンド ザイ』 2003 年 9 月号

```
<quotation>
<speech>
<speaker><sentence type="quasi">松尾</sentence></speaker>
<paragraph>
<sentence>　今でこそ当たり前ですけどね。</sentence>
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<speaker><sentence type="quasi">サスケ</sentence></speaker>
<paragraph>
<sentence>　我々は１０年前からローコスト経営を貫いてます（笑）。</sentence>
</paragraph>
</speech>
</quotation>
<orphanedTitle>
<sentence>レスラー兼経営者となると試合中も会場の入りなど気になるでしょう？ </sentence>
</orphanedTitle>
```

※ br 要素を省略

paragraph 要素

概要

- 段落を表す文書要素である。原則として，一字下げで始まる一行を段落として自動認定する。

形式

■ 要素

br, sentence

■ 属性

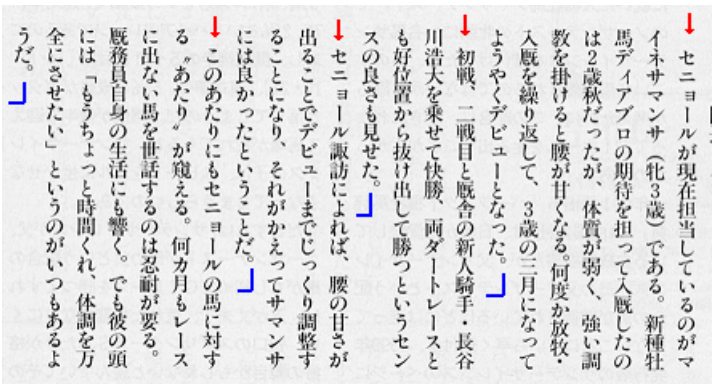
- なし

■ DTD

```
<!ELEMENT paragraph (br|sentence)*>
```

説明

paragraph 要素は，段落を表す文書要素である。原則としてサンプル紙面上，一字下げ (【例 1】　↓) で始まり，改行 (【例 1】　」) によって終わる文章のまとまりを段落とする。なお，現段階では，「一字下げがなされている一論理行」を自動的に paragraph 要素と認定していることに注意されたい。


↓セニョールが現在担当しているのがマイネサマンサ（牝3歳）である。新種牡馬ディアブロの期待を担って入厩したのは2歳秋だったが，体質が弱く，強い調教を掛けると腰が甘くなる。何度か放牧・入厩を繰り返して，3歳の三月になってようやくデビューとなった。」

↓初戦，二戦目と厩舎の新人騎手・長谷川浩大を乗せて快勝，両ダートレースとも好位置から抜け出して勝つというセンスの良さも見せた。」

↓セニョール厩訪によれば，腰の甘さが出ることでデビューまでじっくり調整することになり，それがかえってサマンサには良かったということだ。」

↓このあたりにもセニョールの馬に対する〝あたり〟が窺える。何カ月もレースに出ない馬を世話するのは忍耐が要る。厩務員自身の生活にも響く。でも彼の頭には「もうちょっと時間くれ，体調を万全にさせたい」というのがいつもあるようだ。」

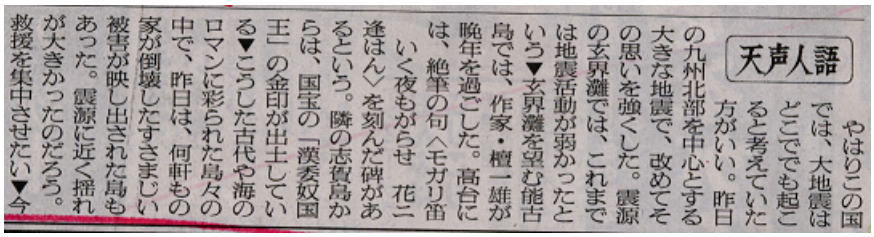
【例 1】paragraph 要素の例 (『優駿』2003 年 11 月号)

■ 補則

段落区切りを別の形式によって示していることが，対象とする資料の範囲内で明らかであれば，その形式によって paragraph 要素の範囲を認定する。

【例２】： 新聞一面のコラム

朝日新聞「天声人語」，毎日新聞「余録」，読売新聞「編集手帳」，産経新聞「産経抄」などは，各社ともに「▼」「▲」「◆」などの記号を段落区切りとしていることが明らかである。よって，新聞一面のコラムについては，これらの記号を段落境界と認定し，paragraph 要素の範囲を定める。



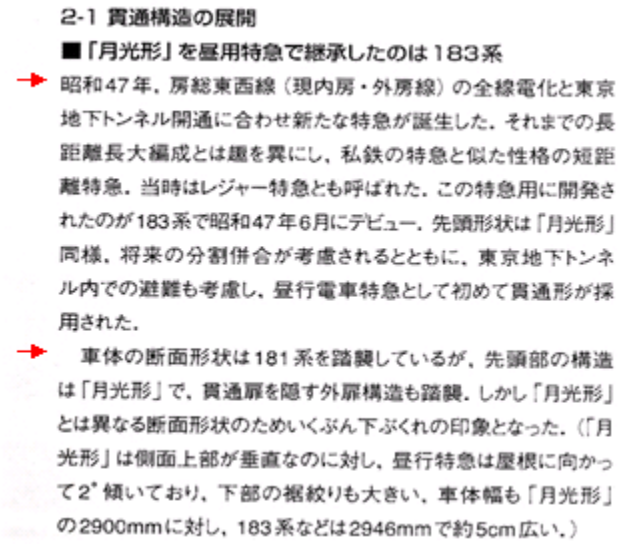
天声人語

やはりこの国では、大地震はどこでも起こると考えていた方がいい。昨日の九州北部を中心とする大きな地震で、改めてその思いを強くした。震源の玄界灘では、これまでは地震活動が弱かったという▼玄界灘を望む能古島では、作家・檀一雄が晩年を過ごした。高台には、絶筆の句〈モガリ笛いく夜もがらせ　花二逢はん〉を刻んだ碑があるという。隣の志賀島からは、国宝の「漢委奴国王」の金印が出土している▼こうした古代や海のロマンに彩られた島々の中で、昨日は、何軒もの家が倒壊したすさまじい被害が映し出された島もあった。震源に近く揺れが大きかったのだろう。救援を集中させたい▼今

【例２】形式による paragraph **要素認定の例** (『**朝日新聞**』「**天声人語**」2005 **年** 3 **月** 21 **日**)

【例３】字下げがない段落の例

【例３】(『鉄道ファン』 2003 年 12 月号) において，一つ目の論理行 (一つ目の→の行) は，行頭の一字下げはされておらず，通常は段落と認定されないが，二つ目の論理行以降は，論理行を一つの段落にしている。したがって，一つ目の論理行が段落であることが明らかなので，段落として認定する。



2-1 貫通構造の展開

■「月光形」を昼用特急で継承したのは183系

昭和47年，房総東西線（現内房・外房線）の全線電化と東京地下トンネル開通に合わせ新たな特急が誕生した．それまでの長距離長大編成とは趣を異にし，私鉄の特急と似た性格の短距離特急．当時はレジャー特急とも呼ばれた．この特急用に開発されたのが183系で昭和47年6月にデビュー．先頭形状は「月光形」同様．将来の分割併合が考慮されるとともに，東京地下トンネル内での避難も考慮し，昼行電車特急として初めて貫通形が採用された．

車体の断面形状は181系を踏襲しているが，先頭部の構造は「月光形」で，貫通扉を隠す外扉構造も踏襲．しかし「月光形」とは異なる断面形状のためいくぶん下ぶくれの印象となった．（「月光形」は側面上部が垂直なのに対し，昼行特急は屋根に向かって2°傾いており，下部の裾絞りも大きい．車体幅も「月光形」の2900mmに対し，183系などは2946mmで約5cm広い．）

【例３】：形式による paragraph **要素認定の例** (『**鉄道ファン**』 2003 **年** 12 **月号**)

形式化例

■ 【例 1】paragraph 要素の例 (『優駿』2003 年 11 月号)

```
 <paragraph>
  <sentence>　セニョールが現在担当しているのがマイネサマンサ（牝３歳）である。</sentence> <sentence>新種牡馬ディアブロの期待を担って入厩したのは２歳秋だったが、体質が弱く、強い調教を掛けると腰が甘くなる。</sentence> <sentence>何度か放牧・入厩を繰り返して、３歳の三月になってようやくデビューとなった。</sentence>
 </paragraph>
 <paragraph>
  <sentence>　初戦、二戦目と厩舎の新人騎手・長谷川浩大を乗せて快勝、両ダートレースとも好位置から抜け出して勝つというセンスの良さも見せた。</sentence>
 </paragraph>
 <paragraph>
  <sentence>　セニョール諏訪によれば、腰の甘さが出ることでデビューまでじっくり調整することになり、それがかえってサマンサには良かったということだ。</sentence>
 </paragraph>
 <paragraph>
  <sentence>　このあたりにもセニョールの馬に対する“あたり”が窺える。</sentence><sentence>何カ月もレースに出ない馬を世話するのは忍耐が要る。</sentence><sentence>厩務員自身の生活にも響く。</sentence><sentence>でも彼の頭には<quote>「もうちょっと時間くれ、体調を万全にさせたい」</quote>というのがいつもあるようだ。</sentence>
 </paragraph>
```

■ 【例 2】形式による paragraph 要素認定の例 (『朝日新聞』「天声人語」)

```
<cluster>
 <title>
 天声人語
 </title>
 <paragraph>
  <sentence>　やはりこの国では、大地震は（…中略…）地震活動が弱かったという</sentence>
 </paragraph>
 <paragraph>
  <sentence>▼玄界灘を望む能古島では、（…中略…）金印が出土している</sentence>
 </paragraph>
 <paragraph>
  <sentence>▼こうした古代や海のロマンに（…中略…）救援を集中させたい</sentence>
 </paragraph>
                          （以下略）
```

※ cluster・paragraph 要素内の形式化は一部省略

profile 要素

概要

- 文書要素著者や登場人物のプロフィールに相当する，article 要素に付随する文書要素を表す。

形式

■ 要素

br, figureBlock, paragraph, quotation, rejectedBlock, sentence, titleBlock

■ 属性

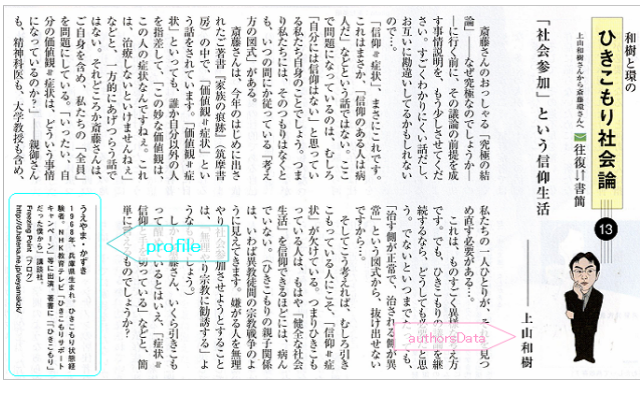
- なし

■ DTD

```
<!ELEMENT profile
        (br|figureBlock|paragraph|quotation|rejectedBlock|sentence|titleBlock)+>
```

説明

記事の著者，インタビューや座談会等の話者，登場人物や話題となっている人物（あるいはグループや組織）についてのプロフィールに相当する文書要素を表す。ただし，【例 1 】で示すように， article 要素に付随する文書要素の場合に限り， article 要素自体がプロフィールで構成されているものは，profile 要素にはならない。

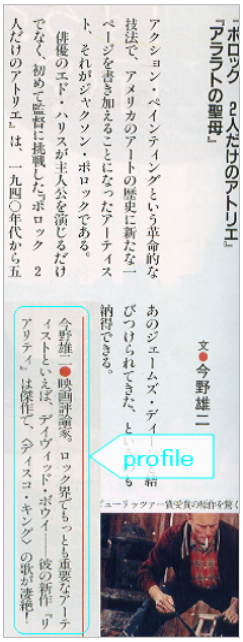


【例 1 】profile (『THE BIG ISSUE』 2006 年 10 月 15 日号)

■ プロフィールとは

プロフィールとは主に，ある人物の経歴・略歴の紹介や寸描，近況報告などをさす。経歴の後に本人のコメントなどがついている場合もあるが，これもプロフィールの一部とみなすこととし，profile 要素に含める。

またその対象は，個人だけではなく，コンビ・グループ・組織などのこともある。



【例 2】本人のコメントがついているもの (『家庭画報』2003 年 11 月号)

■ profile 要素にならないプロフィール

profile 要素になるのは，そのプロフィールが本文に付随する文書要素である場合だけである。よって，article 要素自体がプロフィールの場合は profile 要素とはならない。

- article **要素がすべて詩人の略歴の場合**

詩人略歴
ナー・サインチャクト（納・塞音朝克圖　一九一四～七三）
モンゴル族。内モンゴル錫林郭勒盟の人。一九三七年、日本の東洋大学付属師範留学、帰国後、モンゴル人民共和国留学。内モンゴル日報社。中国作家協会第二回理事。文革期迫害で遭難。詩集に『心の友』『ナー・サインチャクト詩選』などがある。
ニウハン（牛漢　一九二三～　）
モンゴル族。山西省定襄県に生まれる。本名は史成漢。城固の西北大学ロシア文学科入学。学生運動でしばしば逮捕される。人民文学出版社入社。五五年、反右派闘争で逮捕。文革で強制労働に従事。中国作家協会理事。詩集『愛とうた』『沈黙の懸崖』、エッセイ集『蛍火集』など。
：
：

【例 3】秋吉久紀夫編訳『現代中国少数民族詩集』土曜美術社出版販売 2003 年

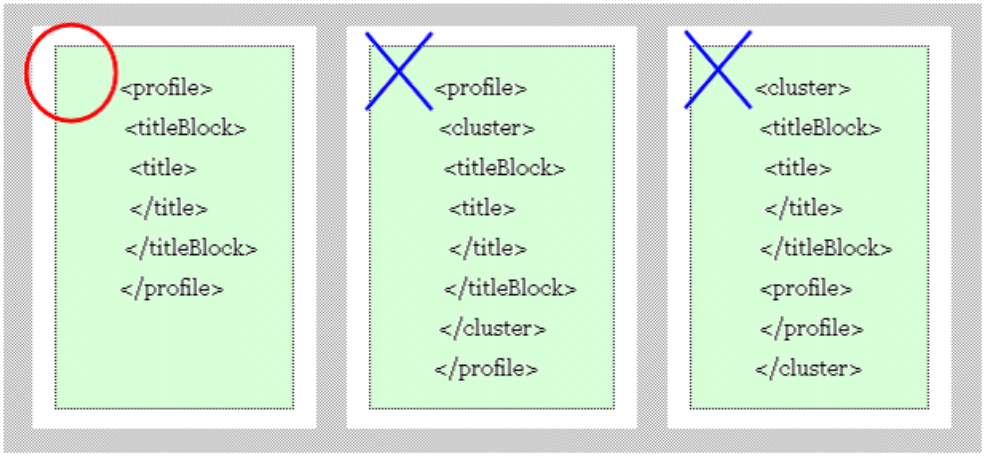
- **「登場人物紹介」などの記事で，物語やドラマの登場人物（複数）の紹介がメインの文書要素**

【例４】プロフィールがメインなので profile 要素とはならない
（『きょうの料理』2003 年 8 月号）

■ プロフィールのタイトル

プロフィールにタイトルがある場合は，該当箇所を titleBlock 要素とする。このタイトルは，「プロフィール」「著者紹介」「○○さん紹介」などのほかにも，そのプロフィールをまとめていると判断できるもの（「○○さんはこんなにスゴイ！」など）も含まれる。

このとき，titleBlock 要素は，profile 要素の内部に，そして cluster 要素を介さず profile 要素の直下に，記述される。



また，たとえタイトル風になっていたとしても，対象となっている人名は titleBlock 要素とはならない。

形式化例

■ 【例 1】profile『THE BIG ISSUE』2006 年 10 月 15 日号

```
<profile>
 <sentence type="quasi">うえやま・かずき</sentence>
  <paragraph>
   <sentence>１９６８年、兵庫県生まれ。ひきこもり状態経験者。</sentence>
   <sentence>ＮＨＫ教育テレビ「ひきこもりサポートキャンペーン」等に出演。</sentence>
   <sentence>著書に『「ひきこもり」だった僕から』講談社。</sentence>
 </paragraph>
 <paragraph>
   <sentence type="quasi">Ｆｒｅｅｚｉｎｇ　Ｐｏｉｎｔ（ブログ）</sentence>
 </paragraph>
 <paragraph>
   <sentence type="quasi">ｈｔｔｐ://ｄ．ｈａｔｅｎａ．ｎｅ．ｊｐ／ｕｅｙａｍａｋｚｋ／</sentence>
 </paragraph>
</profile>
```

※ br 要素は省略

■ 【例 2】本人のコメントがついているもの (『家庭画報』2003 年 11 月号)

```
<profile>
 <paragraph>
  <sentence>今野雄二●映画評論家。</sentence><sentence>ロック界でもっとも重要なアーティストといえば、
デイヴィッド・ボウイ―彼の新作『リアリティ』は傑作で、〈ディスコ・キング〉の歌が凄絶！ </sentence>
 </paragraph>
</profile>
```

※ br 要素は省略

quotation 要素

概要

- 当該記事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・書き起こしを表す。
- フィクションにおいては，地の文とは異なる語り手によるものであることを表す。
- 改行によって，地の文から切り離された要素 (ブロック要素) をマークアップの対象とする。
- インラインの引用は，quote 要素とする。

形式

■ 要素

citation, speech

■ 属性

- なし

■ DTD

```
<!ELEMENT quotation (citation|speech)+>
```

説明

quotation 要素は，当該記事以外の著作物の引用，および発話の引用を表す。 article 要素に対応付けられた著者，タイトルによる文章と，それ以外の文章とを区別すること，または，いわゆる地の文とそれ以外の文を区別することをマークアップの目的とする。

quotation 要素は，著作物からの引用を表す citation 要素と，発話・心内発話の引用・描写・書き起こしなどを表す speech 要素に下位区分される（citation 要素，speech 要素については，それぞれの項を参照のこと)。

ただし，以下に示す条件に適合し，quotation 要素であることが判断できる場合であっても，citation 要素か，speech 要素かの種別を判断できない場合には，いずれの要素にもならないため，citation 要素，speech 要素を含まない quotation 要素として記述されることになる。

quotation 要素は，改行によって他と区切られたブロックの文書要素を対象とし，以下の条件を満たすものを認定する。

(1) 視覚的な差異によって，本文・地の文と明確に区別される文書要素であること
(2) 引用の元となる原典や発話者 (の存在) が明確に示されていること

条件 1 について，本文・地の文との区切り方としては，例えば，以下のようなものが挙げられる。

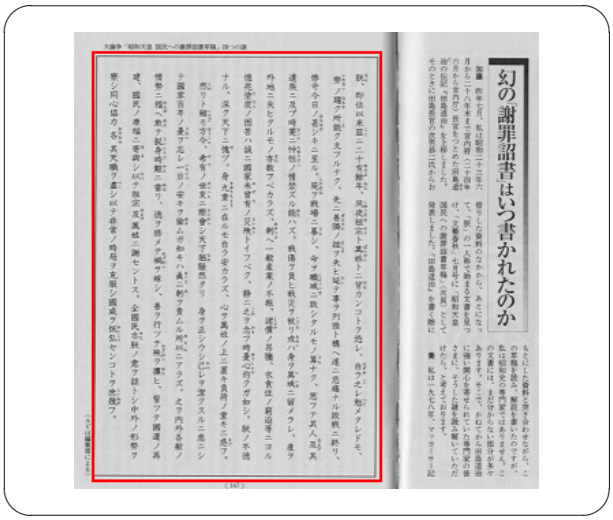
- インデント
- 囲み
- 論理改行
- 引用符やダッシュ

この文書を見てぼくはすぐにやはり三月五日ではないか、と瀧口修造に手紙を書いた。折り返し他の用件もふくめて長い手紙がとどいた。文末の日付は一九六三年二月十二日となっている。この手紙から検挙日に関する部分のみ引用する。

……検挙の月日は御覧になった書類が当のものに間違いないとすれば、私の記憶違いかも知れませんが、案内も三月と記憶しています。私もカン禁されて一カ月ほどして特高部屋に出され桜の花が満開だったのを眺めた記憶があるので、今は私どもに記録としては何も残っていませんが、依然として三月五日頃という風に思い込んでいます。然し福沢氏も四月と書いてをられるので、四月説に対して具体的な反証が何もありません。随分記憶などというものはあやふやなものです。……

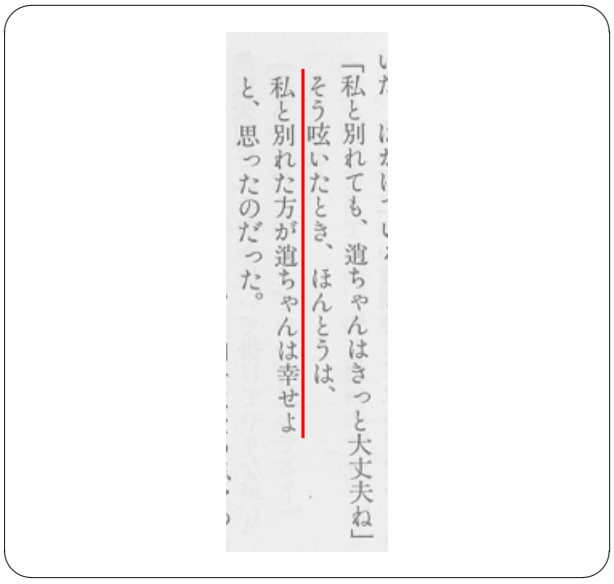
ぼくはその後の訪問の際にこの復刻の書類を持参した。瀧口修造はベージにさっと目を通してすぐさまぼくを見すえて「鵠

【例 1】インデント (『現代詩手帖』2003 年 11 月号)



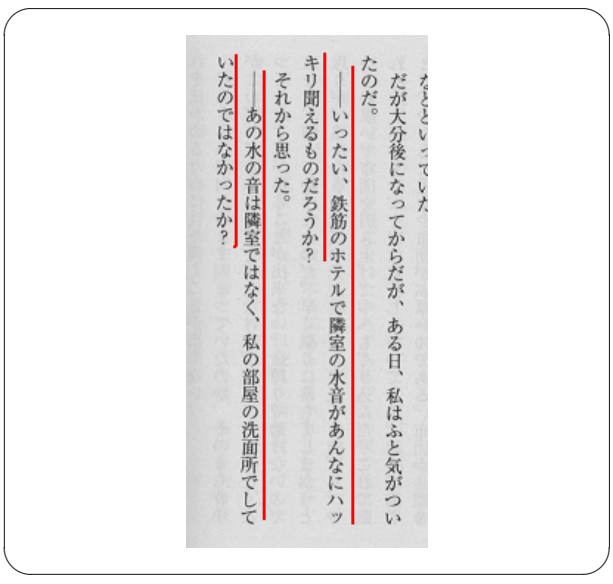
幻の謝罪詔書はいつ書かれたのか

【例 2】囲み (『文藝春秋』2003 年 8 月号)



「私と別れても、道ちゃんはきっと大丈夫ね」
そう呟いたとき、ほんとうは、
私と別れた方が道ちゃんは幸せよ
と、思ったのだった。

【例 3】論理改行 (『文學界』2003 年 11 月号)

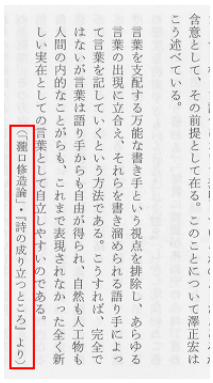


だが大分後になってからだが、ある日、私はふと気がついたのだ。
——いったい、鉄筋のホテルで隣室の水音があんなにハッキリ聞えるものだろうか？
それから思った。
——あの水の音は隣室ではなく、私の部屋の洗面所でしていたのではなかったか？

【例 4】ダッシュ (『小説宝石』2003 年 11 月号)

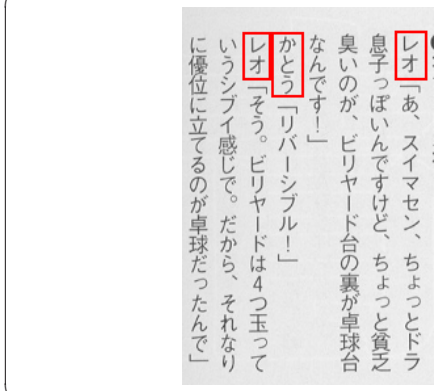
また，条件 2 について，引用元や発話者が明確に示されているとは，以下のような場合を指す。

- 記事内に出典情報や発話者が示されており，当該の原典からの引用であること，または当該の発話者による発話であることが文脈上明確な場合
- 一定のスタイルを保った記事であり，発話であることや文献引用であることが容易に分かる場合


含意として、その前提として在る。このことについて澤正宏はこう述べている。

言葉を支配する万能な書き手という視点を排除し、あらゆる言葉の出現に立合え、それらを書き留められる語り手によって言葉を記していくという方法である。こうすれば、完全ではないが言葉は語り手からも自由が得られ、自然も人工物も人間の内的なことがらも、これまで表現されなかった全く新しい実在としての言葉として自立しやすいのである。

（「瀧口修造論」・「詩の成り立つところ」より）

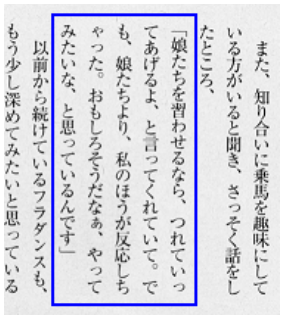

レオ「あ、スイマセン、ちょっとドラ息子っぽいんですけど、ちょっと貧乏臭いのが、ビリヤード台の裏が卓球台なんです！」

かとう「リバーシブル！」

レオ「そう。ビリヤードは4つ玉っていうシブイ感じで。だから、それなりに優位に立てるのが卓球だったんで」

【例５】出典情報提示あり
『現代詩手帖』2003 年 11 月号

【例６】発話者表示あり
『BACKSTAGE PASS』2003 年 8 月号


また、知り合いに乗馬を趣味にしている方がいると聞き、さっそく話をしたところ、

「娘たちを習わせるなら、つれていってあげるよ、と言ってくれていて。でも、娘たちより、私のほうが反応しちゃった。おもしろそうだなぁ、やってみたいな、と思っているんです」

以前から続けているフラダンスも、もう少し深めてみたいと思っている

【例７】かぎ括弧による発話部表示（『ESSE』2003 年 11 月号）

なお，quotation 要素が子要素として，citation 要素，speech 要素を持つ場合，quotation 要素の範囲は，常に子要素である citation 要素，speech 要素の範囲と一致する。よって，例えば 1 話者の 1 発話を単位として認定される speech 要素を包含する quotation 要素の場合は， 1 話者の 1 発話の範囲がマークアップの対象となる。

■ フィクションにおける引用

フィクションにおいては，地の文以外の部分を quotation 要素とする。

(1) 登場人物の発話部
- 戯曲の台詞部
- 小説の会話部

(2) 語り手の交代
- 作中の手紙
- 作中で文献の引用の体裁を取っているもの

(1) 登場人物の発話部は speech 要素，(2) 語り手の交代は citation 要素とする。それぞれ，speech 要素，citation 要素の項を参照のこと。

形式化例

■ 【例 3】『文學界』2003 年 11 月号

```
<quotation>
<speech>
<paragraph>
<sentence type="quasi">「<sentence type="quasi">私と別れても、逍ちゃんはきっと大丈夫ね</sentence>」</sentence>
</paragraph>
</speech>
</quotation>
<sentence type="quasi">そう呟いたとき、ほんとうは、</sentence>
<quotation>
<speech>
<paragraph>
<sentence type="quasi">私と別れた方が逍ちゃんは幸せよ</sentence>
</paragraph>
</speech>
</quotation>
<sentence type="quasi">と、思ったのだった。</sentence>
```

※ br 要素を省略。以下同じ。

■ 【例 5】『現代詩手帖』2003 年 11 月号

```
…<sentence>このことについて澤正宏はこう述べている。</sentence>
<quotation>
<citation>
<paragraph>
<sentence>　言葉を支配する万能な書き手という視点を排除し、あらゆる言葉の出現に立合え、
それらを書き溜められる語り手によって言葉を記していくという方法である。</sentence>
<sentence>こうすれば、完全ではないが言葉は語り手からも自由が得られ、自然も人工物も人間の内的
なことがらも、これまで表現されなかった全く新しい実在としての言葉として自立しやす
いのである。</sentence>
</paragraph>
<source>
<sentence type="quasi">（「瀧口修造論」・『詩の成り立つところ』より）</sentence>
</source>
</citation>
</quotation>
```

■ 【例6】『BACKSTAGE PASS』2003年8月号

```
<quotation>
<speech>
<speaker><sentence type="quasi">レオ</sentence></speaker>
<paragraph>
<sentence type="quasi">「<sentence>あ、スイマセン、ちょっとドラ息子っぽいんですけど、ちょっと貧乏臭いのが、ビリヤード台の裏が卓球台なんです！ </sentence>」</sentence>
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<speaker><sentence type="quasi">かとう</sentence></speaker>
<paragraph>
<sentence type="quasi">「<sentence>リバーシブル！ </sentence>」</sentence>
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<speaker><sentence type="quasi">レオ</sentence></speaker>
<paragraph>
<sentence type="quasi">「<sentence>そう。</sentence><sentence>ビリヤードは４つ玉っていうシブイ感じで。</sentence>
<sentence type="quasi">だから、それなりに優位に立てるのが卓球だったんで</sentence>」</sentence>
</paragraph>
</speech>
</quotation>
```

■ 【例7】 『ESSE』2003年11月号

```
<sentence type="quasi">  また、知り合いに乗馬を趣味にしている方がいると聞き、さっそく話をしたところ、</sentence>
<quotation>
<speech>
<paragraph>
<sentence type="quasi">「<sentence>娘たちを習わせるなら、つれていってあげるよ、と言ってくれていて。</sentence>
<sentence>でも、娘たちより、私のほうが反応しちゃった。</sentence><sentence type="quasi">おもしろそうだなぁ、やってみたいな、と思っているんです</sentence>」</sentence>
</paragraph>
</speech>
</quotation>
<sentence>  以前から続けているフラダンスも、もう少し深めてみたいと思っているし、「習い事にも力を入れたい」今日このごろ。</sentence>
```

quote 要素

概要

- 本文・地の文において，カギ括弧「 」によって，その範囲が示されている文書要素を，自動的にマークアップの対象とする。
- 当該記事とは異なる著作物からの引用や，発話・心内発話の引用，また，「 」で表されるさまざまな表現を表す。
- ブロックの引用は，quotation 要素とする。

形式

■ 要素

%inlineText;, sentence, verseLine

■ 属性

- なし

■ DTD

```
<!ELEMENT quote (%inlineText;|sentence|verseLine)*>
```

説明

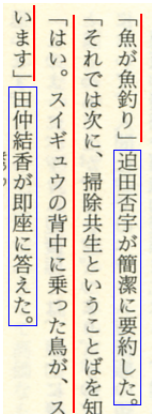
quote 要素は，「 」によって表現されるさまざまな要素であり，当該記事以外の著作物からの引用や発話の引用，「 」による強調表示，など，以下の条件にいずれも当てはまる要素に自動的に付与される。

(1) 本文行内に存在する (インライン)
(2)「 」によって視覚的・形態的に記事本文や地の文と区別して示されている

【例 1】においては，傍線で示した部分が「 」によって取り出される引用表現，囲みで示した部分が引用表現に対する発話者を明示した部分であり，両者が同一の行に位置するインラインの引用 (発話) と認められる。

上記 2 つの条件に当てはまる要素以外は，quote 要素のマークアップの対象とならない。このうち，条件 1 のみを満たさない，インラインでない (ブロックの) 引用・発話は，quotaiton 要素とする。「 」の表示によらないものは，その内容に関わらず，マークアップの対象とならない。

引用以外にも，例えば以下のような「 」で表される表現もマークアップの対象となる。


「魚が魚釣り」迫田否宇が簡潔に要約した。
「それでは次に、掃除共生ということばを知
「はい。スイギュウの背中に乗った鳥が、ス
います」田仲結香が即座に答えた。

【例１】鳥飼否宇『本格的　死人と狂人たち』2003 年 原書房

■ 強調・用語提示

語・表現を強調したり，用語，固有名詞などを提示するために引用符を用いており，その典拠や発話者が明示されていない場合

【例２】『ESSE』2003 **年** 11 **月号**

```
<sentence>新聞などでも報道されて一時期話題になった、いわゆる<quote>「架空請求」</quote>に関する問題も深刻です。
</sentence>
```

■ オノマトペ

人間以外が発した音，鳴き声などを提示するために引用符を用いている場合

【例３】『GOLF DIGEST』2003 **年** 11 **月号**

```
<sentence><quote>「カポ〜ン！」</quote>という打球音はどうかと思うが。 </sentence>
```

形式化例

■ 【例 1】鳥飼否宇『本格的　死人と狂人たち』2003 年 原書房

```
 <paragraph>
%memo wiki
  <sentence>  <quote>「<sentence type="quasi">魚が魚釣り</sentence>」</quote>迫田否宇が簡潔に要約した。
</sentence>
 </paragraph>
 <quotation>
  <speech>
   <paragraph>
    「<sentence type="quasi">それでは次に、掃除共生ということばを知っている人はいますか</sentence>」
   </paragraph>
  </speech>
 </quotation>
 <paragraph>
  <sentence><quote>  「<sentence>はい。</sentence><sentence type="quasi">スイギュウの背中に乗った鳥が、
 スイギュウにつく寄生虫を食べてあげるような関係だと思います</sentence>」</quote>田仲結香が即座に答えた。</sentence>
 </paragraph>
```

rejectedBlock 要素

概要

- サンプル範囲内において，削除対象となった要素の存在を表す。また，原紙の状態により判読できなかった部分の存在を表す。
- 改行によって本文行から切り離された要素 (ブロック要素) をマークアップの対象とする。
- インラインの削除対象要素は，rejectedSpan 要素とする。

形式

■ 要素

- 空要素である。

■ 属性

- *type* (必須)：削除対象要素の種別
  - `copyright` ... 著作権者からの要請など
  - `figure` ... 図表等
  - `formula` ... 数式
  - `foreign` ... 外国語
  - `old` ... 古語
  - `unclear` ... 判読不能箇所
  - `etc` ... その他

■ DTD

```
<!ELEMENT rejectedBlock EMPTY>
<!ATTLIST rejectedBlock type
              (copyright|figure|formula|foreign|old|unclear|etc) #REQUIRED>
```

説明

rejectedBlock 要素は，sample 要素内で，コーパス収録対象外として削除された要素の存在を表す。
コーパス収録対象外の要素には，主に以下のものがある。

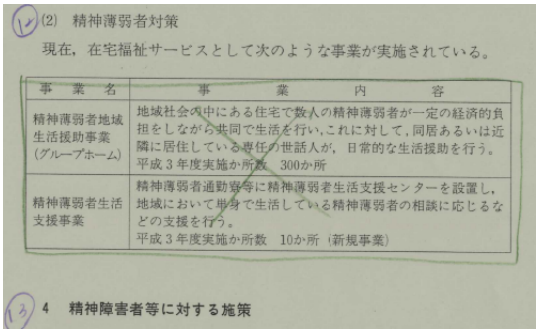
(1) 文字を主たる構成要素としないもの：図，絵，写真，漫画など
(2) 構成要素の図形としての位置関係が意味を持つもの：チャート，表など
(3) 現代日本語の範疇に入らないもの

- 数式
- 外国語
- 古語

(4) 原紙の状態により判読不能な箇所

(5) 著作権処理上，入力が許可されなかった箇所

これらが，サンプル内に現われた場合は入力対象外となるが，文脈を把握する上で，その場所に入力対象外となった要素があるという情報が必要になることがある。



(2) 精神薄弱者対策

現在，在宅福祉サービスとして次のような事業が実施されている。

事業名	事業内容
精神薄弱者地域生活援助事業（グループホーム）	地域社会の中にある住宅で数人の精神薄弱者が一定の経済的負担をしながら共同で生活を行い，これに対して，同居あるいは近隣に居住している専任の世話人が，日常的な生活援助を行う。 平成３年度実施か所数　300か所
精神薄弱者生活支援事業	精神薄弱者通勤寮等に精神薄弱者生活支援センターを設置し，地域において単身で生活している精神薄弱者の相談に応じるなどの支援を行う。 平成３年度実施か所数　10か所（新規事業）

4　精神障害者等に対する施策

【例１】『厚生白書』平成３年度版

【例１】では，囲みと×印の付いた部分が削除対象要素であるが，本文中に「次のような事業が実施されている。」とあり，削除対象要素が「次のような」として参照されるべき要素となっている。この時，削除対象となった要素が存在することを示さずに，続く「４　精神障害者等に対する施策」に連なるのは，文脈上不自然である。

このような場合，入力対象外として削除された要素の存在を示すために，rejectedBlock 要素を用いる。

なお，上記に示した，(1)，(2) の要素については，これらの入力対象外要素に付随する文書要素 (見出しや説明文など) を伴う場合，入力対象外要素と付随する文書要素をまとめて，figureBlock 要素として示し，rejectedBlock 要素は用いない。

rejecteBlock 要素は，削除された要素の内容を *type* 属性によって表す。

- `copyright` ... 著作権者からの要請など
- `figure` ... 図，絵，写真，漫画，チャート，表など
- `formula` ... 数式
- `foreign` ... 外国語
- `old` ... 古語
- `unclear` ... 判読不能箇所
- `etc` ... 上記以外の削除要素（数値主体の項目，人名の羅列など)

形式化例

■ 【例１】『厚生白書』平成３年度版

```
<cluster>
<titleBlock>
<title>
<sentence type="quasi">（2）精神薄弱者対策</sentence>
</title>
</titleBlock>
<paragraph>
<sentence>　現在，在宅福祉サービスとして次のような事業が実施されている。</sentence>
</paragraph>
<rejectedBlock type="figure" />
</cluster>
```

rejectedSpan 要素

概要

- 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，また，原紙の状態により判読できなかった文字の存在を表す。
- 削除対象となる block 要素は，rejectedBlock 要素とする。

形式

■ 要素

- 空要素である。

■ 属性

- *type* (必須)：削除対象要素の種別
  - `formula` ... 数式
  - `foreign` ... 外国語
  - `unclear` ... 判読不能文字
  - `etc` ... その他

■ DTD

```
<!ELEMENT rejectedSpan EMPTY>
<!ATTLIST rejectedSpan type (formula|foreign|unclear|etc) #REQUIRED>
```

説明

rejectedSpan 要素は，sample 要素内で，コーパス収録対象外として削除されたインライン要素の存在，また，原紙の状態により判読できなかった文字の存在を表す。なお，入力対象外要素がブロック要素の場合は，rejectedBlock 要素で記述する。

コーパス収録対象外の要素には，主に以下のものがある。

(1) 文字を主たる構成要素としないもの：図，絵，写真，漫画など
(2) 構成要素の図形としての位置関係が意味を持つもの：チャート，表など
(3) 現代日本語の範疇に入らないもの：数式，外国語，古語など

このうち，インライン要素のものの扱いは以下のように分類される。

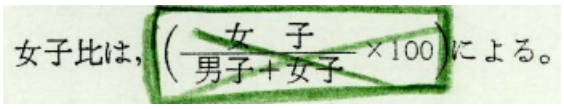
種類	type	備考
数式	formula	要素なし※
外国語	foreign	要素なし※
判読不能文字	unclear	要素なし
図	figure	要素なし
図	–	image 要素
その他	etc	要素なし

※本文行内で入力対象となった場合は，本タグ付与対象外。非現代日本語でも全て本文扱い。

rejectedSpan 要素は，削除された要素の内容を *type* 属性によって表す。

- `formula`：数式
- `foreign`：外国語
- `unclear`：判読不能文字
- `figure`：図
- `etc`：その他

これらが，サンプル内に現われた場合は入力対象外となるが，文脈を把握する上で，その場所に入力対象外となった要素があるという情報が必要になることがある。



【例１】『犯罪白書』昭和 51 年版

上の例では，囲みと×印の付いた部分が削除対象要素であるが，削除対象となった要素が存在することを示さずに，「女子比は，による」と連なるのは，文意・文構造の把握の上で，著しく支障を来たす。

このような場合，入力対象外として削除された要素の存在を示すために，rejectedSpan 要素を用いる。

形式化例

■ 【例１】『犯罪白書』昭和 51 年版

```
女子比は，<rejectedSpan type="formula" />による。
```

ruby 要素

概要

- ルビ（本文に並列して添えられた小書きの文字列）を表す。

形式

■ 要素

%inlineText;

■ 属性

- *rubyText* (必須)： ルビの文字列

■ DTD

```
<!ELEMENT ruby (%inlineText;)*>+
<!ATTLIST ruby rubyText CDATA #REQUIRED>+
```

説明

本文の上下左右など，行間の位置に文字の読みなどを表した文字列（ルビ）は ruby 要素によって表す。ルビを付ける対象を要素内容とし，ルビの文字列は，*rubyText* 属性として記述する。

ルビとして典型的なものは，漢字の読みを示したもの（振り仮名）だが，対象の意味を表す別言語による訳や言い換え語などが，行間位置に小書きされているものも，同様に ruby 要素で表現する。

ルビに対するマークアップは，原則として本文の文字単位で行うが，下の形式化例に示すような熟字訓・当て字や別言語による言い換え，括弧付きのルビなどは，複数の文字列（語全体）に対してマークアップをする。

また，引用部分に用いられる，原文をそのまま転記したことを示す「ママ」は，ルビと同じ位置に現われるが，語の読み・意味等を表したものと区別するために， ruby 要素では表現しない。「ママ」は文章編集上の処置法を示した注として noteBodyInline 要素として記述する。

形式化例

■ 【例１】語彙 [ごい] (振り仮名：熟字訓以外)

```
<ruby rubyText="ご">語</ruby><ruby rubyText="い">彙</ruby>
```

■ 【例２】五月雨 [さみだれ] (振り仮名：熟字訓・当て字)

```
<ruby rubyText="さみだれ">五月雨</ruby>
```

■ 【例３】真名瀬 [（しんなせ）] (振り仮名：括弧付き)

```
<ruby rubyText="（しんなせ）">真名瀬</ruby>
```

■ 【例４】未来 [あした] (類似概念語による言い換え)

```
<ruby rubyText="あした">未来</ruby>
```

■ 【例５】Ｃｈａｒｇｅ　ｕｐ [充電完了](別言語による言い換え)

```
<ruby rubyText="充電完了">Ｃｈａｒｇｅ　ｕｐ</ruby>
```

■ 【例６】特効 [特殊効果] (語義の指示)

```
<ruby rubyText="特殊効果">特効</ruby>
```

sample 要素

概要

- サンプリングによって 1 サンプルとされた文書要素を表す。
- 可変長のサンプルでは，sample 要素は，一つの article 要素からなる。

形式

■ 要素

- article

■ 属性

- *sampleID* (必須) ：サンプルに関する情報を外部データベースから参照するための ID
- *type* (必須) ：サンプルの種別
  - `variableLength` ... 可変長 (この値以外を取ることはない)
- *version* (必須) ：サンプルの版

■ DTD

```
<!ELEMENT sample (article)>
<!ATTLIST sample sampleID CDATA #REQUIRED>
<!ATTLIST sample type (variableLength) #REQUIRED>
<!ATTLIST sample version CDATA #REQUIRED>
```

説明

サンプリングによって 1 サンプルと規定された文書要素を表す。可変長サンプルは，基本的に一つの「記事」を 1 サンプルとする。したがって，sample 要素は，一つの article 要素からなる。ただし，article 要素の中に，複数の article 要素が含まれる場合もあることに注意されたい。

sample 要素は，*sampleID* 属性と *version* 属性，*type* 属性を持つ。

- *sampleID* 属性: サンプルに関する情報 (書誌情報，サンプル長，ジャンル，参照ページなどを格納) を外部のデータベースから参照するための ID である。*sampleID* 属性値の表記規則は，Sample_ID の仕様を参照のこと。
- *version* 属性:サンプルの版を表す。
- *type* 属性: サンプルの種別（可変長，固定長）を表す。可変長の場合は，`variableLength` となる。

形式化例

```
<sample sampleID="OW1X_00001" version="20070208" type="variableLength">
 <article articleID="OW1X_00001_V001" isWholeArticle="false">
    :
    :
 </article>
</sample>
```

sampling 要素

概要

- サンプリングポイントに関する情報を示す。

形式

■ 要素

- 空要素である。

■ 属性

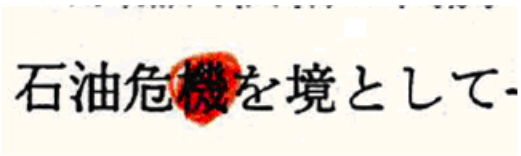
- *type* ： サンプリングポイントの種別
  - `start` ... サンプル抽出基準点

■ DTD

```
<!ELEMENT sampling EMPTY>
<!ATTLIST sampling type (start) #REQUIRED>
```

説明

sampling 要素は，サンプリングに関する情報として，サンプリングポイントを示すための要素である。可変長サンプルにおいては，sample 要素となる article 要素を抽出するための基準となる一文字（サンプル抽出基準点）の位置を示す。



【例 1】『通商白書』昭和 54 年版

上図で塗りつぶされている一文字「機」が，サンプル抽出基準点である。この文字の直前に，空要素のタグを入力することで，基準点の位置を示す。

形式化例

■ 【例 1】『通商白書』昭和 54 年版

```
石油危<sampling type="start" />機を境として
```

※ sentence 要素の形式化は省略

sentence 要素

概要

- 文に相当するまとまりを表す。原則として，句点などの表記上の手がかりに基づいて，自動認定する。

形式

■ 要素

%inlineText;, delete, sentence, verseLine

■ 属性

- *type* (任意)：
  - `quasi` ... 文区切り文字以外の基準により自動付与された sentence 要素
  - `verse` ... 韻文（verse 要素）内の sentence 要素

■ DTD

```
<!ELEMENT sentence (%inlineText; | delete | sentence | verseLine)*>
<!ATTLIST sentence type (quasi|verse) #IMPLIED>
```

説明

sentence 要素は，文に相当するまとまりを表す。原則として，句点などの表記上の手がかりに基づいて，自動認定する。

■ 自動認定の種類

(1) 文区切り文字に基づいて認定される場合

- 次の文区切り文字を文の末尾とする。
  - 「。」「.」「！」「？」
- sentence 要素の範囲は，末尾として認定された文字から，次の条件を満たす範囲のうち，最も狭い範囲とする。
  - 直前の sentence 要素の末尾の次の文字まで
  - 行頭まで (= ブロック要素をまたぐ sentence 要素はないことを意味する)

(2) 文区切り文字以外の基準に基づいて認定される場合

- *type* 属性を `quasi` とする。
- 認定規則は次のとおりである。

(a) 論理行中に一つも sentence 要素が認定されていなければ，行末に文区切り文字がなくても，その論理行全体を sentence 要素とする。

(b) 論理行頭から一つ以上の sentence 要素の並びが存在する場合，行末に文区切り文字がなくても，sentence 要素とする。

```
<sentence>株価が上昇した。</sentence><sentence type="quasi">これは現時点での情報だ</sentence>
```

■ 引用符，かっこ書き内部の sentence 要素の認定

- 次の記号を「引用符」「かっこ」として処理する。
  - ( ) 〔 〕 [ ] { } 〈 〉《 》「 」『 』【 】
- 「引用符」「かっこ」内の文字列は，次のように，一つの論理行として扱い，その文字列に対して，sentence 要素の自動認定処理を実行する。ただし，上記認定規則の 2.a は適用しない。
  - 当該の文字列の先頭を行頭として扱う。
  - 当該の文字列の末尾を行末として扱う。

■ verse 要素内の sentence 要素

verse 要素内の sentence 要素は，必ず *type* 属性の値を `verse` とする。認定は上の認定基準によらず，人手によって認定した文の範囲を 1 論理行としてデータ整形した上で，以下の規則によって自動認定する。

- 1 論理行を sentence 要素とする。
- sentence 要素の入れ子構造は作らない。

形式化例

■ 【例１】さまざまな文区切り文字

```
<sentence>６件中１～３件を表示しています。</sentence>
<sentence>そうなんですか？ </sentence>
<sentence>毎朝新聞を読んで，売上拡大！ </sentence>
```

■ 【例２】文区切り文字以外の基準に基づいて認定された例

```
<sentence>これは３月現在のものです。</sentence><sentence type="quasi">詳細は，付録を参照のこと</sentence>
```

```
<titleBlock>
<title>
<sentence type="quasi">「特集　日本酒」</sentence>
</title>
</titleBlock>
```

■ 【例３】引用内の sentence 要素の例（『小説宝石』2003 年 11 月号）

```
<sentence>白い靴下のまま、僕はおかんの<quote>「<sentence>どこ行くねんなぁー、もう出掛けるでぇー！ </sentence>」</quote>という声を背に受け、坂の上の電話ボックスに走った。</sentence>
```

■ 【例４】かっこ内の sentence 要素の例（『法学教室』2003 年 11 月号）

```
<sentence>～規定する犯罪収益の前提犯罪に，１及び２に掲げる罪（<sentence>現行組織的犯罪処罰法の別表に掲げるものを除く。</sentence>）を加えるものとすること。</sentence>
```

```
<sentence>表３のような結果が出た（<sentence>ただし，途中経過。</sentence><sentence type="quasi">詳細は後日</sentence>）。</sentence>
```

■ 【例５】『短歌』2003 年 6 号

```
<verse>
<sentence type="verse">大いなるロダンは手のみ彫りたれど生き生きとして全体を見す<verseLine /></sentence>
</verse>
```

source 要素

概要

- citation 要素に含まれる，引用文献についての情報 (文献名，著者名，著者情報など) を表す。

形式

■ 要素

br, info, sentence

■ 属性

- なし

■ DTD

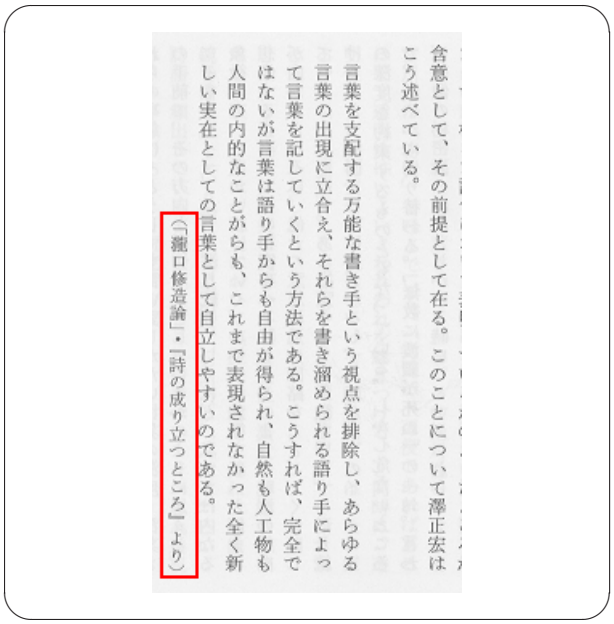
```
<!ELEMENT source (br|info|sentence)*>
```

説明

source 要素は，citation 要素として，本文行とは切り離されたブロックの文書要素に含まれる，引用文献についての情報を表す。引用の出典情報を示すと共に，citation 要素内部における，引用本文と出典情報部を区別することを目的にマークアップを行う。

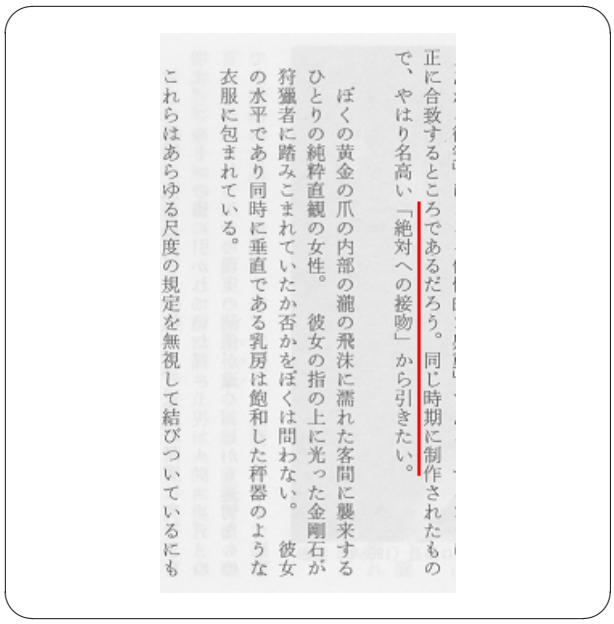
source 要素としてマークアップされる要素の典型的なものは，出典情報である書名・文献名であるが，そのほか，引用文献や投稿の著者名，著者情報，発行年等の文献情報なども含まれる。

source 要素の対象となるのは，citation 要素内部に示された出典情報であり，記事内の全ての引用要素の出典情報を，漏れなく示すものではない。例えば，ブロックで示される citation 要素の前後の本文要素中に，インラインで示される出典情報などは，source 要素の対象とはならない。


含意として、その前提として在る。このことについて澤正宏はこう述べている。

言葉を支配する万能な書き手という視点を排除し、あらゆる言葉の出現に立合え、それらを書き溜められる語り手によって言葉を記していくという方法である。こうすれば、完全ではないが言葉は語り手からも自由が得られ、自然も人工物も人間の内的なことがらも、これまで表現されなかった全く新しい実在としての言葉として自立しやすいのである。

（「瀧口修造論」・『詩の成り立つところ』より）


正に合致するところであるだろう。同じ時期に制作されたもので、やはり名高い「絶対への接吻」から引きたい。

ぼくの黄金の爪の内部の瀧の飛沫に濡れた客間に襲来するひとりの純粋直観の女性。彼女の指の上に光った金剛石が狩猟者に踏みこまれていたか否かをぼくは問わない。彼女の水平であり同時に垂直である乳房は飽和した稀器のような衣服に包まれている。

これらはあらゆる尺度の規定を無視して結びついているにも

【例 1】source **要素の例**
『現代詩手帖』2003 年 11 月号

【例 2】source **要素にならない例**
『現代詩手帖』2003 年 11 月号

上の例において，本文行から改行により切り離された要素 (= citation 要素) の出典が本文行内に示された「絶対への接吻」であることは，文脈上明らかだが，これは，citation 要素として切り出されるブロック要素の外に位置する本文要素であるため，source 要素としない。

形式化例

■ 【例 1】『現代詩手帖』2003 年 11 月号

```
<quotation>
 <citation>
 <paragraph>
  <sentence>  言葉を支配する万能な書き手という視点を排除し、あらゆる言葉の出現に立合え、それらを書き溜められる語り手によって言葉を記していくという方法である。</sentence><sentence>こうすれば、完全ではないが言葉は語り手からも自由が得られ、自然も人工物も人間の内的なことがらも、これまで表現されなかった全く新しい実在としての言葉として自立しやすいのである。</sentence>
 </paragraph>
 <source>
 <sentence type="quasi">（「瀧口修造論」・『詩の成り立つところ』より）</sentence>
 </source>
 </citation>
</quotation>
```

※ br 要素は省略

speaker 要素

概要

- speech 要素内部で，後続または先行する発話について，その話者を明示的に表した文字列やマークを表す。
- speech 要素の中で，発話部と区別するためにマークアップする。

形式

■ 要素

br, sentence

■ 属性

- なし

■ DTD

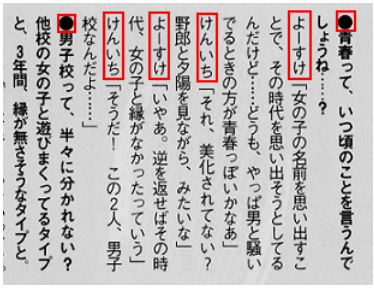
```
<!ELEMENT speaker (br|sentence)*>
```

説明

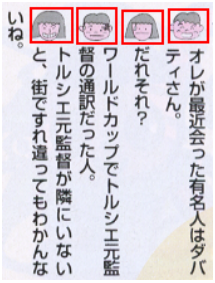
speaker 要素は，speech 要素内部で，後続または先行する発話について，その話者を明示的に表した文字列や記号・マークを表す。speech 要素内部で，実際の発話部以外の要素を区別することをマークアップの目的とする。

speaker 要素は，具体的には，以下のような要素を含む。

- 発話者の名前
- 発話者を表す絵記号
- インタビュアーを示す「—」や「●」などの記号



【例１】発話者の名前／インタビュアーを示す記号の例
『BACKSTAGE PASS』2003 年 8 月号



【例２】 発話者を表す絵の例
（『ひよこクラブ』2003 年 8 月号）

形式化例

■ 【例１】 『BACKSTAGE PASS』2003 年 8 月号

```
<quotation><speech>
<speaker>
<sentence type="quasi">
●
</sentence><br type="automatic_original"/>
</speaker>
<paragraph>
<sentence>　青春って、いつ頃のことを言うんでしょうね…？ </sentence>
</paragraph>
</speech></quotation>
<quotation> <speech>
<speaker>
<sentence type="quasi">
よーすけ
</sentence><br type="automatic_original"/>
</speaker>
<paragraph>
「<sentence type="quasi">女の子の名前を思い出すことで、その時代を思い出そうとしてるんだけど…どうも、やっぱ男と騒いでるときの方が青春っぽいかなあ</sentence>」
</paragraph>
</speech></quotation>
<quotation><speech>
<speaker>
<sentence type="quasi">
けんいち
</sentence><br type="automatic_original"/>
</speaker>
<paragraph>
「<sentence">それ、美化されてない？ </sentence><sentence type="quasi">野郎と夕陽を見ながら、みたいな</sentence>」

</paragraph>
</speech></quotation>
（以下略）
```

■ 【例２】 『ひよこクラブ』2003 年 8 月号

```
<quotation>
<speech>
<speaker>
<sentence type="quasi">
<image description="男性の似顔絵" no="5" />
</sentence><br type="automatic_original"/>
</speaker>
<paragraph>
<sentence>  オレが最近会った有名人はダバティさん。</sentence>
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<speaker>
<sentence type="quasi">
<image description="女性の似顔絵" no="6" />
</sentence><br type="automatic_original"/>
</speaker>
<paragraph>
<sentence>  だれそれ？ </sentence>
</paragraph>
</speech>
</quotation>
（以下略）
```

speech 要素

概要

- 発話の引用・書き起こし，心内発話の描写などを表す。
- 改行によって，地の文から切り離された要素 (ブロック要素) をマークアップの対象とする。
- インラインの発話・心内発話の引用・描写・書き起こしは，quote 要素とする。

形式

■ 要素

blockEnd, br, list, noteBody, paragraph, quotation, rejectedBlock, sentence, speaker, verse

■ 属性

- なし

■ DTD

```
<!ELEMENT speech
          (blockEnd|br|list|noteBody|paragraph|quotation|rejectedBlock|sentence|
          speaker|verse)*>
```

説明

speech 要素は，quotation 要素の子要素であり，発話の引用・書き起こし，心内発話の描写などを表す。文体的差異が予想される，いわゆる地の文と会話文を区別することを，マークアップの目的とする。具体的には，以下のようなものを対象とする。

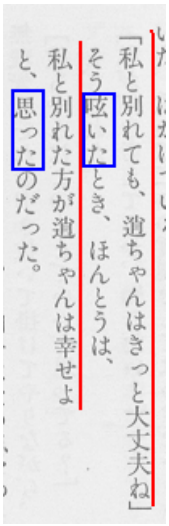
- 発話の引用や心内発話の描写
- 座談会・インタビュー・トーク番組等における発話の書き起こし
- 小説の会話部・戯曲の台詞

speech 要素は，本文と前後を改行によって区切られたブロックの文書要素を対象とし，以下のいずれかの条件を満たすものとする（※インラインの発話・心内発話の引用・描写・書き起こしは，quote 要素とする。quote 要素については，当該の項を参照のこと)。

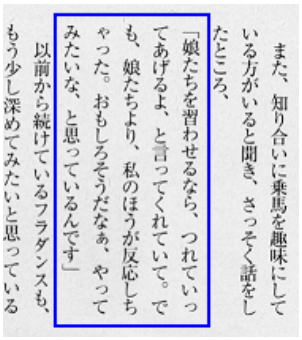
(1) 当該のブロック要素に対応する発話者や思考者が明示されており，「言った」「思った」等の，発話・思考を示す言語表現を伴うこと

(2) article が，戯曲であるか，または，インタビュー・対談・トーク番組等における発話の書き起こしを中心に構成されたもので，一定の表示形態によって発話部が明確なこと

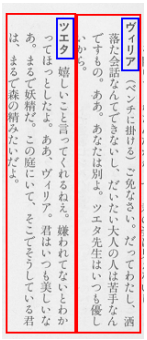
speech 要素は，実際に発声された言葉の描写である発話と，心内で語られた言葉の描写である心内発話の両方を，区別なくマークアップする。



【例１】(1) **発話・思考を示す言語表現を伴う例（小説の会話部）**
（『文學界』2003 年 11 月号）



【例２】(2)-1 **発話部を鉤括弧で示している例**
（『ESSE』2003 年 11 月号）



【例３】(2)-2 **発話者に続けて発話部を示している例**
（『文學界』2003 年 11 月号）

形式化例

■ 【例 1】(1)『文學界』2003 年 11 月号

```
 <quotation>
 <speech>
 <paragraph>
 <sentence type="quasi">  「<sentence type="quasi">私と別れても、逍ちゃんはきっと大丈夫ね</sentence>」
</sentence>
 </paragraph>
 </speech>
 </quotation>
 <sentence type="quasi">そう呟いたとき、ほんとうは、</sentence>
 <quotation>
 <speech>
 <paragraph>
 <sentence type="quasi">私と別れた方が逍ちゃんは幸せよ</sentence>
 </paragraph>
 </speech>
 </quotation>
 <sentence>と、思ったのだった。</sentence>
```

※ speech 要素内の形式化は一部省略

■ 【例 2】(2)-1『ESSE』2003 年 11 月号

```
 <sentence type="quasi">また、知り合いに乗馬を趣味にしている方がいると聞き、さっそく話をしたところ、</sentence>
 <quotation>
 <speech>
 <paragraph>
 <sentence type="quasi">  「<sentence>娘たちを習わせるなら、つれていってあげるよ、と言ってくれていて。
</sentence><sentence>でも、娘たちより、私のほうが反応しちゃった。</sentence><sentence type="quasi">おもしろ
そうだなぁ、やってみたいな、と思っているんです</sentence>」</sentence>
 </paragraph>
 </speech>
 </quotation>
 <sentence>  以前から続けているフラダンスも、もう少し深めてみたいと思っているし、「習い事にも力を入れたい」今日このご
ろ。</sentence>
```

■ 【例３】(2)-2『文學界』2003 年 11 月号

```
<quotation>
<speech>
<speaker>ヴィリア</speaker>
<paragraph>
<sentence>（ベンチに掛ける）ご免なさい。</sentence><sentence>だってわたし、洒落た会話なんてできないし、だいたい大人の人は苦手なんですもの。</sentence><sentence>ああ。</sentence><sentence>あなたは別よ。</sentence><sentence>ツエタ先生はいつも優しいから。</sentence>
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<speaker>ツエタ</speaker>
<paragraph>
<sentence>嬉しいこと言ってくれるねえ。</sentence><sentence>嫌われてないとわかってほっとしたよ。</sentence><sentence>ああ、ヴィリア。</sentence><sentence>君はいつも美しいなあ。</sentence><sentence>まるで妖精だ。</sentence><sentence>この庭にいて、そこでそうしている君は、まるで森の精みたいだよ。</sentence>
</paragraph>
</speech>
</quotation>
```

subScript 要素

概要

- 数式や化学式などに用いる下付きの文字を表す。

形式

■ 要素

%inlineText;

■ 属性

- なし

■ DTD

```
<!ELEMENT subScript (%inlineText;)*>
```

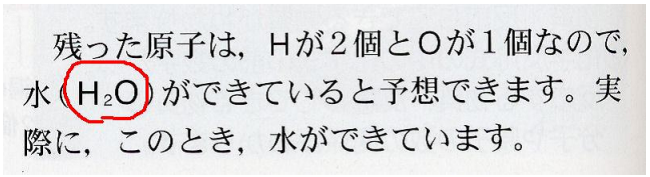
説明

数式や化学式などでは，文字の大きさを小さくして，直前の文字の右下付近に添えるような下付きの文字が使われる。下付きの文字は，文字を入力し，subScript として表現する。

形式化例

■ 【例 1】水の化学式

原資料



形式化

```
H<subScript>2</subScript>O
```

superScript 要素

概要

- 数式や化学式などに用いる上付きの文字を表す。

形式

■ 要素

%inlineText;

■ 属性

- なし

■ DTD

```
<!ELEMENT superScript (%inlineText;)*>
```

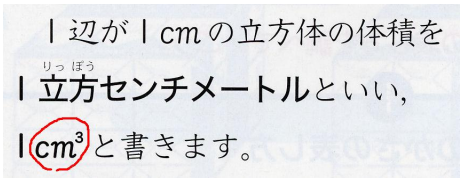
説明

数式や化学式などでは，文字の大きさを小さくして，直前の文字の右上付近に添えるような上付きの文字が使われる。上付きの文字は，文字を入力し，superScript 要素として表現する。なお，横書きの場合，注を参照するための番号が語の右上に小書きで表されることがある。注番号は noteMarker 要素を用いて表現する。

形式化例

■ 【例１】立方センチメートル

原資料



形式化

```
c m<superScript> 3 </superScript>
```

table 要素

概要

- 表を表す。

形式

■ 要素

br, sentence

■ 属性

- なし

■ DTD

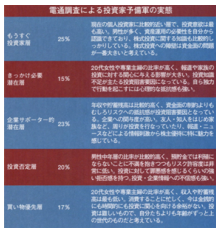
```
<!ELEMENT table (br|sentence)+>
```

説明

■ 表の定義

表とは次のものを指す。罫線の有無は考慮しない。

- ３列以上のもの。…【例１】
- 列ラベル (や行ラベル) があるもの。事項を縦横に配列し，交錯状況によって表現したもの (＝クロス表)。…【例２】



【例１】３列以上のもの
(『MONEY JAPAN』2003 年 12 月号)

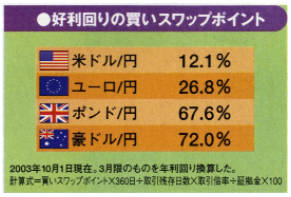
●その他の主な美容機能付き新繊維

製品名（素材メーカー）	特徴
ナノデュウ（カネボウ繊維）	抗酸化作用のあるビタミンE誘導体を配合。高い保湿性と肌への潤い効果が期待できるというナノテクノロジー繊維
ウリアントαホワイト（シキボウ）	化粧品に使用されるα-アルブチンを配合。メラニン色素の合成を抑えるほか、抗菌、抗酸化性の効果も
バイオセイム（小松精練）	肌荒れの抑制やうるおい維持に効果を発揮する新素材。化粧品などに使われるリン脂質系ポリマーを固着させた
VITA-Q10（ダイワボウ）	健康食品などで注目されている補酵素、コエンザイムQ10を固着させた繊維。現在はふとん地が中心だが、今後は肌着などにも

【例２】列ラベルのあるもの
(『日経 TRENDY』2003 年 10 月号)

罫線に囲まれていても，次の両方を満たすものはリストとし，list 要素で表す…【例３】

- ２列以下のもの
- 列ラベル (や行ラベル) がないもの



【例３】表ではなくリスト (『MONEY JAPAN』2003 年 12 月号)

■ 入力対象となる表

サンプル範囲中の表は，原則として入力対象外とすることになっている。しかし，例外として入力対象とされる表は， table 要素として表すものとする。入力対象とならない表は，figure 要素または rejectedBlock 要素で表す。

table 要素は，必ず figureBlock 要素に含まれる。table 要素に付随する文書要素がある場合は，これを caption 要素とし，共に figureBlock 要素の子要素となる。

figureBlock，caption 要素については，各項を参照のこと。

形式化例

■ 【例２】

```
<figureBlock>
<caption>
<sentence type="quasi">●その他の主な美容機能付き新繊維</sentence><br type="automatic_original"/>
</caption>
<table>
<sentence type="quasi">製品名（素材メーカー）</sentence><br type="automatic_original"/>
<sentence type="quasi">特徴</sentence><br type="automatic_original"/>
<sentence type="quasi">ナノデュウ（カネボウ繊維）</sentence><br type="automatic_original"/>
<sentence>抗酸化作用のあるビタミンＥ誘導体を配合。</sentence>
<sentence type="quasi">高い保湿性と肌への潤い効果が期待できるというナノテクノロジー繊維</sentence><br type="automatic_original"/>
<sentence type="quasi">ウリアントαホワイト（シキボウ）</sentence><br type="automatic_original"/>
<sentence>化粧品に使用されるα-アルブチンを配合。</sentence>
<sentence type="quasi">メラニン色素の合成を抑えるほか、抗菌、抗酸化性の効果も</sentence><br type="automatic_original"/>
<sentence type="quasi">バイオセイム（小松精練）</sentence><br type="automatic_original"/>
<sentence>肌荒れの抑制やうるおい維持に効果を発揮する新素材。</sentence>
<sentence type="quasi">化粧品などに使われるリン脂質系ポリマーを固着させた</sentence><br type="automatic_original"/>
<sentence type="quasi">ＶＩＴＡ-Ｑ１０（ダイワボウ）</sentence><br type="automatic_original"/>
<sentence>健康食品などで注目されている補酵素、コエンザイムＱ１０を固着させた繊維。</sentence>
<sentence type="quasi">現在はふとん地が中心だが、今後は肌着などにも</sentence><br type="automatic_original"/>
</table>
</figureBlock>
```

title 要素

概要

- 特定範囲の文書要素の内容を代表する記述を表す。
- 文書の階層構造の標識として機能する。
- titleBlock 要素に包含される。

形式

■ 要素

br, noteBody, rejectedBlock, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT title (br|noteBody|rejectedBlock|sentence)*>
```

説明

title 要素は，titleBlock 要素に包含され， 特定範囲の内容を代表する記述を表す。当該範囲の内容・要約・一部抽出などに相当する表現である。本文とは視覚的にも区別され，文書の階層構造を示すための「標識」として機能する。title 要素の典型的な例は，新聞の見出しや，論文などの文書における章，節などのタイトルである。

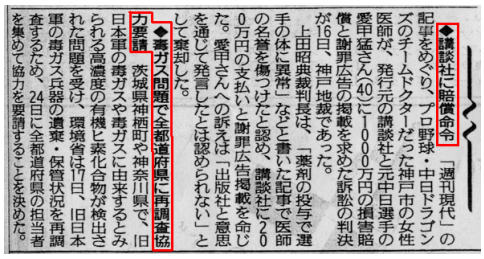
このうち論文等の章，節などのタイトルは，「章」や「節」といった特定範囲を代表する「見出し」であると同時に，文書の階層構造を表すための標識となっている。次に示す例は，章と節の形式化の例である。

```
<cluster>
 <titleBlock>
  <title><sentence type="quasi">第１章　はじめに</sentence></title>
 </titleBlock>
 <cluster>
  <titleBlock>
   <title><sentence type="quasi">１．１　本論文の背景</sentence></title>
  </titleBlock>
  <paragraph>
   <sentence>　本論文の背景としては，　.....
 </cluster>
</cluster>
```

【例１】章や節のタイトル

【例１】において，title 要素は，第 1 章と 1.1 節のタイトルをマークアップしている。章や節の範囲は，cluster 要素により明示する。cluster 要素には，必ず title 要素を含み，章や節の文書構造中の関係は，個々の cluster 要素の包含関係により表現される。

なお，title 要素は，常にブロック要素として記述する。【例２】のような，インラインの title 要素は，改行を挿入してブロックとして取り出してマークアップすることとする。



◆講談社に賠償命令

「週刊現代」の記事をめぐり、プロ野球・中日ドラゴンズのチームドクターだった神戸市の女性医師が、発行元の講談社と元中日選手の愛甲猛さん(40)に1000万円の損害賠償と謝罪広告の掲載を求めた訴訟の判決が16日、神戸地裁であった。

上田昭典裁判長は、「薬剤の投与で選手の体に異常」などと書いた記事で医師の名誉を傷つけたと認め、講談社に200万円の支払いと謝罪広告掲載を命じた。愛甲さんへの訴えは「出版社と意思を通じて発言したとは認められない」として棄却した。

◆毒ガス問題で全都道府県に再調査協力要請

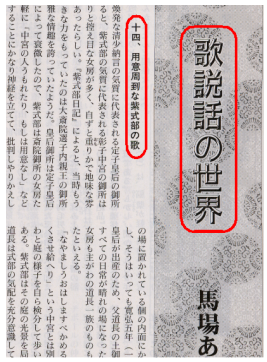
茨城県神栖町や神奈川県で、旧日本軍の毒ガスや毒ガスに由来するとみられる高濃度の有機ヒ素化合物が検出された問題を受け、環境省は17日、旧日本軍の毒ガス兵器の遺棄・保管状況を再調査するため、24日に全都道府県の担当者を集めて協力を要請することを決めた。

【例２】『読売新聞』2003 年 6 月 17 日夕刊

```
<article articleID="PN3c_xxxxx_V0001" isWholeArticle="true">
<titleBlock>
<title>
<sentence type="quasi">◆講談社に賠償命令</sentence><br type="automatic_original"/>
</title>
</titleBlock>
<paragraph>
<sentence>　「週刊現代」の記事をめぐり、プロ野球・中日ドラゴンズのチームドクターだった神戸市の女性医師が、発行元の講談社と元中日選手の愛甲猛さん（４０）に１０００万円の損害賠償と謝罪広告の掲載を求めた訴訟の判決が１６日、神戸地裁であった。</sentence>
                          :　（以下，省略）
```

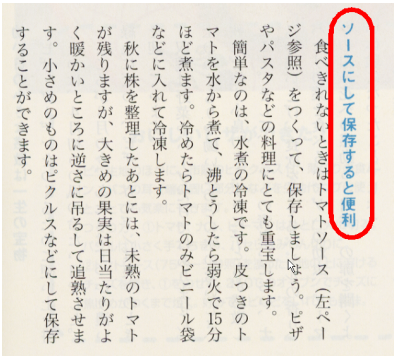
■ title 要素の性質

title 要素が備えている性質としては，【例 3】〜【例 5】に示すようなものが挙げられる。


歌説話の世界

馬場あ

十四、用意周到な紫式部の歌


ソースにして保存すると便利

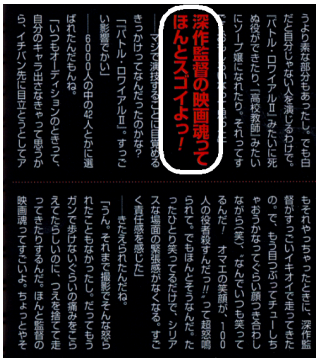
食べきれないときはトマトソース（左ページ参照）をつくって、保存しましょう。ピザやパスタなどの料理にとても重宝します。

簡単なのは、水煮の冷凍です。皮つきのトマトを水から煮て、沸とうしたら弱火で15分ほど煮ます。冷めたらトマトのみビニール袋などに入れて冷凍します。

秋に株を整理したあとには、未熟のトマトが残りますが、大きめの果実は日当たりがよく暖かいところに逆さに吊るして追熟させます。小さめのものはピクルスなどにして保存することができます。

【例 3】特定範囲の文書要素に対するトピック
(『短歌研究』2003 年 11 月号)

【例 4】特定範囲の文書要素に対する要約
(『趣味の園芸』2003 年 8 月号)


深作監督の映画魂ってほんとスゴイよっ！

【例 5】特定範囲の文書要素からの一部抽出 (『ポップティーン』2003 年 9 月号)

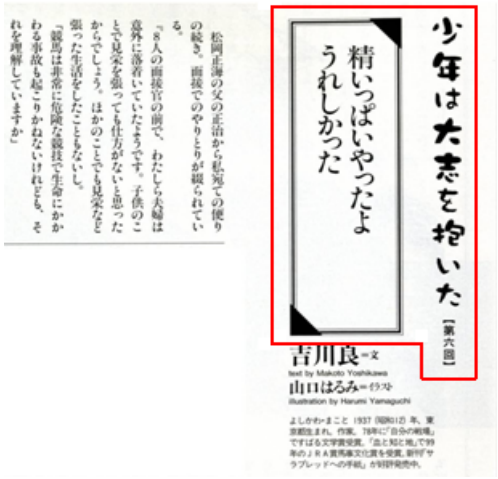
■ title 要素と titleBlock 要素の認定

レイアウト上，本文とは切り離されており，title 要素に付随する要素のように見えるものであっても，文章の内容を判断する上で重要と認められるものについては，title 要素とみなす。例えば，以下のようなものである。

- 連載・枠・コーナーの名称
- 記事種を示す名称
- article 要素・cluster 要素のトピックを表す表現

【例 6】は，連載コラムである。この例における，主たる title 要素は，当該回のタイトル「精いっぱいやったよ　うれしかった」である。連載のタイトル「少年は大志を抱いた【第六回】」は，紙面の表示からは，当該

回のタイトルに付随する要素のようにも見えるが，この文章の内容判断において重要な役割を果たすと考えられるので，これも合わせて title 要素と認定する。



【例６】『優駿』2003 年 11 月号

形式化例

■ 【例３】特定範囲の文書要素に対するトピック (『短歌研究』2003 年 11 月号)

```
<article articleID="PM32\_xxxxx\_0001" isWholeArticle="true">
 <titleBlock>
  <title>
   <sentence type="quasi">歌説話の世界</sentence>
  </title>
 </titleBlock>
 <authorsData>
  <sentence type="quasi">馬場あきこ</sentence>
 </authorsData>
  <cluster>
   <titleBlock>
    <title>
    <sentence type="quasi">十四、用意周到な紫式部の歌</sentence>
    </title>
   </titleBlock>
                          :
  </cluster>
                          :       (以下，省略)
```

titleBlock 要素

概要

- title 要素とそれに付随する要素全体を表す。
- 文書の階層構造の標識として機能する。

形式

■ 要素

br, list, rejectedBlock, sentence, title

■ 属性

- なし

■ DTD

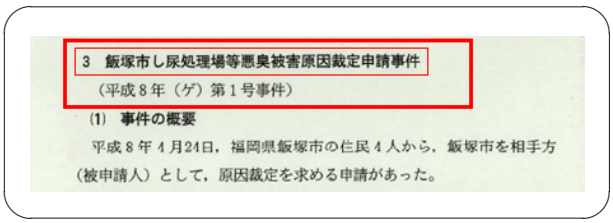
```
<!ELEMENT titleBlock (br|list|rejectedBlock|sentence|title)*>
```

説明

titleBlock 要素は，article 要素や cluster 要素などの冒頭部分に位置し，本文とは明確に切り分けられる部分で，文書の階層構造を示すための「標識」として機能する。title 要素を必ず内包し，これに付随する要素がある場合は，title 要素と付随する要素全体を表す。

title 要素に付随する要素としては，title 要素以外の部分となる。具体的には title 要素の前後に配置される，レイアウト，フォント，文字の大きさなど視覚的に区別されるものを認めることにする。

【例 1】付随要素が title 要素の前にある場合
（『spring』2003 年 1 月号）



【例 2】付随要素が title 要素の後にある場合
（『公害紛争処理白書』平成 9 年版）

【例 1】～【例 3】のごとく，titleBlock 要素は，title 要素に加えて，title 要素に付随するさまざまな文書要素を合わせて形式化する。これにより，タイトル関連部分と本文が明確に区別される。

【例３】付随要素が title 要素の前後にある場合
（『月刊ザテレビジョン』2003 年 10 月号）

形式化例

■ 【例１】付随要素が title 要素の前にある場合（『spring』2003 年 1 月号）

```
<titleBlock>
 <sentence type="quasi">気になる新車はもちろん、おしゃれ有名人が選ぶ中古車から、こだわりマイカー拝見まで</sentence>
 <title>
 <sentence>スタイル重視のおしゃれマイカーを探せ！ </sentence>
 </title>
</titleBlock>
```

※ br 要素を省略。以下同じ

■ 【例２】付随要素が title 要素の後にある場合（『公害紛争処理白書』平成９年版）

```
<titleBlock>
  <title>
  <sentence type="quasi">３　飯塚市し尿処理場等悪臭被害原因裁定申請事件</sentence>
  </title>
<sentence type="quasi">（平成８年（ゲ）第１号事件）</sentence>
</titleBlock>
```

■ 【例３】付随要素が title 要素の前後にある場合（『月刊ザテレビジョン』2003 年 10 月号）

```
<titleBlock>
 <sentence>後藤・石川・藤本が映画初主演！！ </sentence><sentence>　笑って、泣いてモ～たいへん！ </sentence>
 <sentence type="quasi">９／１３<enclosedCharacter>土</enclosedCharacter>公開「青春ばかちん料理塾」「１７才旅立ちのふたり」</sentence>
 <title>
  <sentence type="quasi">ＣＩＮＥＭＡ　ＳＰＥＣＩＡＬ</sentence>
  <sentence type="quasi">１７才それぞれの青春</sentence>
 </title>
 <sentence>料理の勉強に励んだり、友情を深めあったり…３人の少女たちの日常を描く青春映画２本が同時公開！ </sentence>
</titleBlock>
```

verse 要素

概要

- 詩，和歌，俳句，歌謡 (歌詞を含む) などの韻文を表す。

形式

■ 要素

blockEnd, br, info, rejectedBlock, sentence

■ 属性

- なし

■ DTD

```
<!ELEMENT verse (blockEnd|br|info|rejectedBlock|sentence)+>
```

説明

詩，和歌，俳句，歌謡 (歌詞を含む) などの韻文を表すための要素である。

verse 要素は，sentence および blockEnd 要素からなる。verse 要素の中の sentence 要素は，必ず *type* 属性の値を `verse` とする。韻文は，論理行を認定することが困難な場合があること，改行自体が意味を持つ場合があることから，紙面上の改行を verseLine 要素で置き換える。

なお，verse 要素は，ブロック要素であり，行中に現れる韻文は，verse 要素としない。

■ verse 要素内の sentence の認定

verse 要素内の sentence は，以下の基準によって認定する。

(1) 俳句・短歌等： 例外なく 1 句・ 1 首を 1 sentence とする。

(2) 詩

- 連を超える sentence は認めない。
- 文終止記号（句点等），詩行末，および記号類（スラッシュ，空白など）の位置と文末が一致する可能性があると考え，上記注目箇所において，活用形等に鑑み，前後に修飾・係り受け関係がないと判断できる場合は，その位置を文末とみなす。
- 修飾・係り受け関係の判断が困難なもの（一般的な文の認定にそぐわないもの）は，詩行を単位として sentence とする。

形式化例

■ 【例１】『短歌』2003 年 6 号

```
<verse>
<sentence type="verse">大いなるロダンは手のみ彫りたれど生き生きとして全体を見す<verseLine /></sentence>
</verse>
```

■ 【例２】大岡信 他編「群像日本の作家」

```
<verse>
<sentence type="verse">十の蜂舎の成りしとき<verseLine />よき園成さば必らずや<verseLine />鬼ぞうかがふといましめし<verseLine />かしらかむろのひとありき<verseLine /></sentence>
<sentence type="verse">山はかすみてめくるめき<verseLine />桐むらさきに燃ゆるころ<verseLine />その農園の扉を過ぎて<verseLine />苺需めしをとめあり<verseLine /></sentence>
</verse>
```

■ 【例３】大柴晏清著「文学とすし」

```
<verse>
<sentence type="verse">〈灰色に　ひとりぼっちに　僕の夢にかかってゐる／とほい村よ／</sentence>
<sentence type="verse">あの頃　ぎぼうしゅとすげが暮れやすい花を咲き／山羊が啼いて　一日一日過ぎてゐた／</sentence>
<sentence type="verse">やさしい朝でいっぱいであった―／</sentence>
<sentence type="verse">お聞き　</sentence>
<sentence type="verse">春の空の山なみに／お前の知らない雲が焼けてゐる　</sentence>
<sentence type="verse">明るく　そして消えながら／</sentence>
<sentence type="verse">とほい村よ（後略)〉<verseLine/></sentence>
</verse>
```

verseLine 要素

概要

- 韻文における紙面上の行末を表す。

形式

■ 要素

- 空要素である。

■ 属性

- なし

■ DTD

```
<!ELEMENT verseLine EMPTY>
```

説明

韻文における紙面上の行末を表す。

韻文は，改行自体が意味を持つ場合があることから，紙面上における行末の位置を，verseLine 要素で示す。

形式化例

■ 【例１】大岡信 他編「群像日本の作家」

```
 <verse>
 <sentence type="verse">十の蜂舎の成りしとき<verseLine />よき園成さば必らずや<verseLine />鬼ぞうかがふといましめし<verseLine />かしらかむろのひとありき<verseLine /></sentence>
 <sentence type="verse">山はかすみてめくるめき<verseLine />桐むらさきに燃ゆるころ<verseLine />その農園の扉を過ぎて<verseLine />苺需めしをとめあり<verseLine /></sentence>
 </verse>
```

タグ一覧 (固定長)

sample 要素

概要

- サンプリングによって 1 サンプルとされた文書要素を表す。
- 固定長サンプルの sample 要素は，サンプル抽出基準点から，1000 文字以上を含む。

形式

■ 要素

article

■ 属性

- *sampleID* (必須)
  - サンプルに関する情報を外部データベースから参照するための ID[*12]
- *type* (必須) ：サンプルの種別
  - `fixedLength` ... 固定長 (この値以外を取ることはない)
- *version* (必須) ：サンプルの版

■ DTD

```
<!ELEMENT sample (article)*>
<!ATTLIST sample sampleID CDATA #REQUIRED>
<!ATTLIST sample type (fixedLength) #REQUIRED>
<!ATTLIST sample version CDATA #REQUIRED>
```

説明

サンプリングによって 1 サンプルと規定された文書要素を表す。固定長サンプルに含まれる文字数は，サンプル抽出基準点（sample 要素を参照）から 1000 文字以上であることが保証される。なお，固定長の sample 要素は，可変長の sample 要素と異なり，二つ以上の article 要素を含む場合もある。

sample 要素は，*sampleID* 属性と *type* 属性を持つ。

- *sampleID* 属性: サンプルに関する情報 (書誌情報など) を外部のデータベースから参照するための ID である。*sampleID* 属性値の表記規則は，Sample_ID の仕様[*13]を参照のこと 。
- *type* 属性: サンプルの種別（可変長，固定長）を表す。固定長の場合は，`fixedLength` となる。

[*12] http://www2.kokken.go.jp/densi/public/wiki/ から [ver.2.0] → [データベース] を参照のこと。
[*13] Web site 上の仕様 (http://www2.kokken.go.jp/densi/public/wiki/) を参照されたい。

■ sample 要素に収録されるテキストの範囲

sample 要素に収録されるテキストの範囲については，以下の通り規定する。

- サンプル抽出基準点を含む最上位の sentence 要素（sentence 要素がなければ，直上のブロック要素）の**先頭**を，テキストの先頭とする。
- サンプル抽出終了点を含む最上位の sentence 要素（sentence 要素がなければ，直上のブロック要素）の**末尾**を，テキストの末尾とする。

■ 文字数カウントの対象外となる文字と要素

サンプル抽出終了点を決定するために，サンプル抽出基準点より 1000 文字をカウントする際には，以下の文字および要素を数えない。

- 文字: 句読点，空白文字，その他一般記号類[*14]
- 要素: noteMarker, inlineNodeBody, delete

形式化例

```
<sample sampleID="OW1X_00001" version="20070208" type="fixedLength">
 <article articleID="OW1X_00001_F001">
   :
   :
 </article>
</sample>
```

[*14] 本仕様の Web site にカウント対象外文字のリスト (UTF-16LE テキストファイル) を掲載してあるので，参照されたい。

sampling 要素

概要

- サンプリングポイントに関する情報を示す。

形式

■ 要素

- 空要素である。

■ 属性

- *type* ： サンプリングポイントの種別
  - `start` ... サンプル抽出基準点
  - `end` ... サンプル抽出終了点

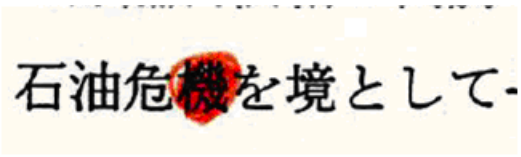
■ DTD

```
<!ELEMENT sampling EMPTY>
<!ATTLIST sampling type (start|end) #REQUIRED>
```

説明

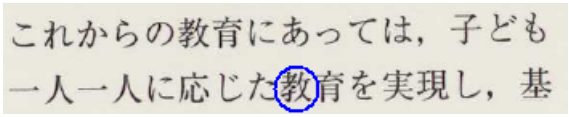
sampling 要素は，サンプリングに関する情報として，サンプリングポイントを示すための要素である。固定長サンプルにおいては，次の 2 種類のサンプリングポイントがある。

- サンプル抽出基準点： sample 要素となる文字列を抽出するための基準となる文字の位置を表す。
- サンプル抽出終了点： サンプル抽出基準点から 1000 文字目の文字を表す。



【例 1】『通商白書』昭和 54 年版

【例 1 】で赤く塗りつぶされている一文字「機」が，サンプル抽出基準点である。この文字の直前に，sampling 要素タグ (空要素) を入力することで，基準点の位置を示す。*type* 属性は，`start` である。



【例 2 】『文部科学白書 平成 13 年度』

一方，【例 2 】の囲みを付した「教」の字が，サンプル抽出終了点である。サンプル抽出終了点は，当該文字の直後に sampling 要素タグ (空要素) を入力する。*type* 属性は，`end` である。

形式化例

■ 【例 1 】『通商白書 昭和 54 年版』

```
石油危<sampling type="start" />機を境として
```

■ 【例 2 】『文部科学白書 平成 13 年度』

```
これからの教育にあっては，子ども一人一人に応じた教<sampling type="end"/>育を実現し，
```

コーパス開発センター (電子化サブグループ)

山口昌也　(言語資源研究系助教，コーパス開発センター (兼))
高田智和　(理論・構造研究系准教授，コーパス開発センター (兼))
北村雅則　(名古屋学院大学商学部講師)
間淵洋子　(コーパス開発センター プロジェクト特別研究員)
大島 一　(コーパス開発センター プロジェクト奨励研究員)
小林正行　(群馬大学教育学部講師)
西部みちる　(コーパス開発センター プロジェクト奨励研究員)

国立国語研究所内部報告書 (LR-CCG-10-04)

『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

平成 23 年 2 月 25 日

執筆者　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
発行者　大学共同利用機関法人 人間文化研究機構 国立国語研究所
連絡先　〒190-8561 東京都立川市緑町 10-2
電話　042-540-4300 (代表)

ISBN978-4-906055-03-6

国立国語研究所