第 8 章　文境界情報

浅原 正幸　小西 光　田中 弥生　間淵 洋子

8.1　はじめに

本章では『現代日本語書き言葉均衡コーパス』の文境界情報について説明する。文境界情報を規定する手がかりになるものとして、(1)文字情報を用いるもの、(2)形態論情報を用いるもの、(3)係り受け関係を用いるものなどが考えられる。BCCWJ-DVD 版（Version 1.0）は、文境界情報を含む文書構造タグの整備と形態論情報の整備とを並行して実施していたため、文字情報を手がかりとして用いた文境界認定作業にとどまっていた。また、工数の制約から知恵袋（OC）については論理行 [1] を表す<webLine>タグを付与するにとどめ、実質的な文境界修正作業を行っていなかった。その結果、BCCWJ-DVD 版 （Version 1.0）の文境界認定基準の妥当性については様々な指摘がなされた。

BCCWJ-DVD 版（Version 1.1）へのバージョンアップに際し、M-XML と TSV（第 6 章）に対して、形態論情報を手がかりとして用いた文境界基準を再策定することで、問題の解消を試みた。以下では、BCCWJ-DVD 版（Version 1.0）の問題点を示し、また、それに対しどのような文境界修正作業を行ったのか説明する。

本章の構成は以下のとおりである：8.2 節では BCCWJ-DVD 版 （Version 1.0）の文境界認定基準を示す。8.3 節では BCCWJ-DVD 版 （Version 1.1）の文境界認定作業について述べる。8.4 節では、コアデータに対するその他の文境界情報を紹介する。

8.2　BCCWJ-DVD 版（Version 1.0） の文境界認定基準

本節では BCCWJ-DVD 版 （Version 1.0）の文境界認定基準について述べる。はじめに文境界認定基準における手がかりについて概観する。

8.2.1 文境界認定基準についての手がかり

文境界認定においては、何らかの「手がかり」を用いて規則を人手で記述する必要がある。文境界認定作業をある程度自動化するためには何を「手がかり」に使うかが重要となる。以下では「手がかり」として、(1)文字情報を用いるもの、(2)形態論情報を用いるもの、(3)係り受け関係を用いるものの 3 種類について詳しく述べる。

(1) 文字情報に基づく認定とは、句点などに基づき文境界を認定する手法である。多くの形態素解析の前処理として、句点記号「。」「.」感嘆符「！」疑問符「？」などを手がかりとした文境界認定が行われている。少し高度な情報として、開き括弧や閉じ括弧

[1] 本節では紙面などの物理的制約によって指示される行を「物理行」「表示行」と呼ぶのに対して、改行コードやブロック要素などにより指示される行を「論理行」と呼ぶ。

を用いた規則を記述し、括弧の対応をとるという手法が存在する。

(2) 形態論情報に基づく認定とは、形態素解析により認定される品詞情報などを用いる手法である。句点のリストを第 5 章に示した短単位形態論情報（小椋他 2011）における品詞「記号-句点」などに汎化できるほか、開き括弧や閉じ括弧についても「記号-括弧開」「記号-括弧閉」と汎化して記述することができる。さらに、辞書に登録されている固有名詞や顔文字などに埋め込まれている記号などを文境界候補から除外することができる。その一方で、形態素解析誤りの影響をある程度見込んで処理する必要がある。

(3) 係り受け関係に基づく認定とは、文境界認定に係り受け関係のスパンを用いる手法である。括弧内の要素が文であるかどうかを認定するために括弧内の要素が連結係り受け木をなすかを判定したり、括弧の前後で係り受け関係があるかどうかで文要素の入れ子を認定したりする。

8.2.2 BCCWJ-DVD 版（Version 1.0）における文境界認定基準の概要

まず、BCCWJ-DVD 版 （Version 1.0）における文境界について述べる。BCCWJ-DVD 版 （Version 1.0）においては文字情報のみを含む C-XML（第 4 章）と形態論情報を含む M-XML （第 6、9 章）の 2 種類の XML 形式でデータが表現されている。文境界情報は XML 内の sentence 要素として表現されている。この 2 種類の形式において認定している文境界に差異がある。

C-XML における文境界認定：

C-XML（第 4 章）においては手がかりとして文字情報を用いた自動処理に基づく文境界認定が基本となっている。話し言葉や既存の書き言葉コーパスと異なり、元媒体のレイアウト情報に基づく文書構造情報（ブロック要素）が利用されている。以下 C-XML における文のスパンを表現する sentence 要素の認定規則について例（図 8-1）を示しながら解説する。自動認定においては句点記号「。」「.」感嘆符「！」疑問符「？」（以下文末記号）やブロック要素開始位置直前を文区切り位置とみなし、直前文の末尾を sentence 要素の始端とみなす処理（sentence タグ<sentence> </sentence> を付与）を行う（例 C-1）。文末記号によって認定される sentence 要素を正則な sentence 要素と呼ぶ。論理行頭からひとつ以上の sentence 要素の並びが存在し、かつ、行末に文末記号がない場合は sentence 要素とみなす（例 C-2）。論理行中にひとつも sentence 要素がなく文末記号もない場合、その論理行全体を sentence 要素とみなす（例 C-3）。これらの文末記号以外によって認定される sentence 要素は、特殊な文として属性 type="quasi"を付与する（例 C-2、C-3：以下 sentence@quasi 要素と略記）。文字情報として 9 種類の括弧の対応（括弧類A[2]）などを用いて、文認定時に sentence 要素の入れ子を許している。

括弧内にひとつも文末記号を含まない場合、括弧内に sentence 要素を認定しない（例

2 括弧類A：「補助記号−括弧開」「補助記号−括弧閉」のうち（ ）〔〕［ ］｛ ｝〈〉《》「」『』【】9 対

C-4)。括弧内にひとつ以上の文末記号が含まれる場合、括弧内に sentence 要素を認定する（例 C-5)。括弧内にひとつ以上の文末記号が含まれ、かつ、閉じ括弧直前に文末記号が出現しない場合、閉じ括弧直前までの部分を特殊な文とみなし、属性 type="quasi"を付与する（例 C-6)。

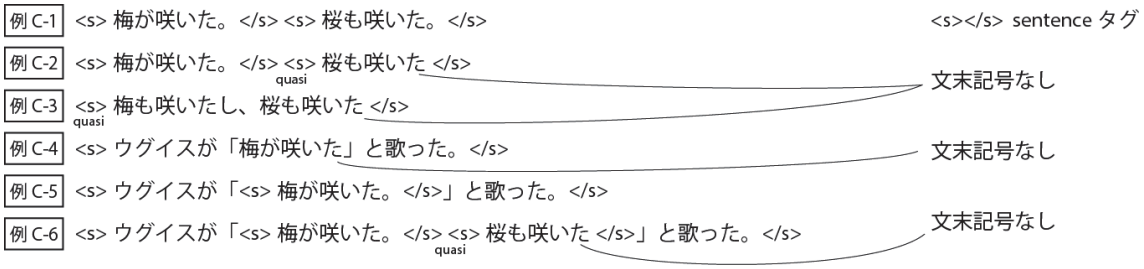

図 8-1:　C-XML における文境界認定

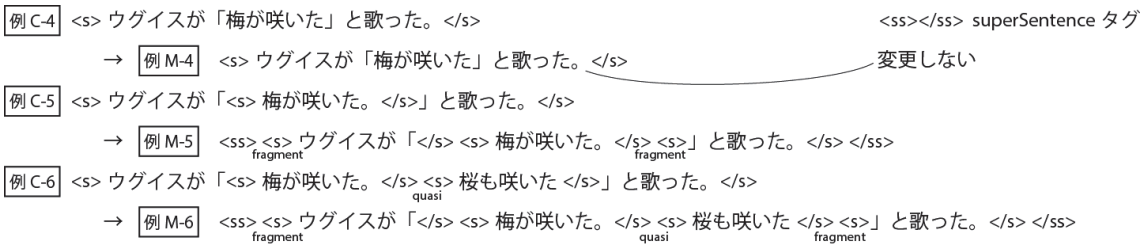

図 8-2:　C-XML から M-XML への変換

M-XML における文境界認定：

M-XML（第 6、9 章）においては、C-XML の文境界認定を基礎としつつ、C-XML とは異なる、より単純化した文境界認定を行う方針を採用した。C-XML の問題点として、sentence 要素がきわめて長くなる場合があること、形態素解析などの入力となる「文」が定めがたいこと、データを文番号で管理できないことの三つがあげられる。

M-XML では、C-XML において sentence 要素が入れ子になっている場合に、その最も内側（下位）にあるもののみを正則の sentence 要素とし、外側（上位）にある sentence は superSentence とする。その上で、superSentence の内側にありながら正則の sentence 要素の外側に位置する部分については、新たに sentence 要素と見なすとともに type="fragment"という属性（以下 sentence@fragment 要素と略記）を与えて、文断片であることを明示する。この際、括弧記号のみからなる文断片要素を作らないために、内側の sentence 要素に隣接する括弧記号を送り込む。最終的に superSentence と sentence の 2 階層からなる文境界情報が残される（図 8-2)。

例 C-4 においては sentence 要素に入れ子が発生していないため、C-XML と M-XML の sentence 要素は一致する（例 M-4)。

例 C-5 においては、括弧内の最内スパンの sentence 要素 "梅が咲いた。" を M-XML に

おける正則な sentence 要素と見なす（例 M-5）。例 C-5 における最外スパンは新たに superSentence 要素として認定する。正則 sentence 要素に含まれない最外スパンの連続文字列については、sentence@fragment 要素として認定する。ただし、正則 sentence 要素に隣接する括弧記号は sentence 要素に送り込む。

例 C-6 においては括弧内に正則な sentence 要素“梅が咲いた。”と sentence@quasi 要素“桜も咲いた”の二つが認定されている。例 C-6 における最外スパンを新たに superSentence 要素として認定する（例 M-6）。括弧内の 2 種類の sentence 要素（正則な sentence 要素と sentence@quasi 要素）を認定し、これに含まれない前後の連続文字列を sentence@fragment 要素として認定する。ただし、内側の sentence 要素に隣接する括弧記号は内側の sentence 要素に送り込む。

しかし、例 M-5・M-6 における、「内側の sentence 要素に隣接する括弧記号は内側の sentence 要素に送り込む処理」が網羅的ではなかった。今回はこの問題を解決するために網羅的なパターンを記述し、再処理する。図 8-2 では、問題になるパターンを示した。

8.3 BCCWJ-DVD 版（Version 1.1）における文境界認定基準

8.3.1 BCCWJ-DVD 版（Version 1.1）における文境界認定の作業方針

以下に文境界認定の作業方針について述べる。BCCWJ-DVD 版（Version 1.0）の文字情報による自動処理と、BCCWJ-DepPara の係り受け関係の情報による人手修正との中間的な処理として、**形態論情報を用いた自動抽出結果の人手修正**をコアデータ・非コアデータ全体に対して実施する。

修正方法としては、まず C-XML における文字列レベルの情報を用いた文境界認定におけるバグ相当のものを自動抽出して人手修正し、次に M-XML に変換する際のバグ相当のものを、形態論情報を用いて自動抽出して、バッチ処理および人手修正を行う。基本的に最内スパンの正則な sentence 要素を認定するとともに、その作業に伴い発生する sentence@fragment 要素のような文が認定されることを許す。係り受け関係の整合性は検証しないが、括弧内の要素について最低限の確認作業（強調や補足の認定）を行う。詳細を以下に示す：

[処理 C] C-XML レベルで認定できる誤りの検出

BCCWJ-DVD 版 （Version 1.0）において、文字情報に基づく処理により 9 対の括弧（括弧類A）内に文末記号があるが文境界が設定されていない要素が約 6,000 箇所発見された。顔文字に埋め込まれた文末記号や括弧が対応していない事例について、全数人手で確認する。

[処理 M] M-XML レベルで認定できる誤り検出

処理 C が完了後、形態論情報を用いた誤り検出を行う。形態論情報を用いた誤り検出

においては、国立国語研究所に寄せられている様々な誤り報告事例や他のアノテーション作業時に問題となった事例をもとに、人手で形態論情報を用いたパターンを記述した。このパターンの認定においてはそのマッチする事例のうち修正率（真に修正すべき事例数／マッチする事例数）に基づいて 2 種類の処理を行う。

[M(α)] 修正率が高いパターン：マッチするほとんどの事例が真に修正すべき事例であるが、例外的に修正しなくてもよい事例が出現するパターン。これらについては、バッチ処理適用前に例外的な事例を排除するように人手で確認する。人手確認後バッチ処理で修正する（修正箇所自動抽出→人手例外確認→バッチ処理)。

[M(β)] 修正率が低いパターン：マッチする事例の一部のみを修正するパターン。全数確認は困難であるが、修正すべき事例が含まれうるパターンを先にバッチ処理で展開し、逐一人手で確認する（修正箇所自動抽出→人手修正処理)。

今回の修正は形態論情報を含む M-XML のみに対して実施し、C-XML については実施しない。この修正にともない、必要があれば形態論情報・文書構造タグも修正する。

8.3.2 BCCWJ-DVD 版（Version 1.1）における文境界認定基準の詳細

8.3.2.1 基準の前提

文境界認定基準の前提として、今回踏襲する BCCWJ-DVD 版 （Version 1.0）の文境界認定基準 3 点について示す。

1 点目は、現存する superSentence 要素を踏襲することを前提に sentence タグを付与することである。

2 点目は、助詞・助動詞から始まる、助詞・助動詞で終わる、助詞・助動詞のみの sentence 要素の発生を認めることである。

3 点目は、括弧内に文末記号が含まれない場合には sentence タグは付与しないことである（例 C-4、例 M-4 を踏襲)。

以下 8.3.2.2 節では括弧内に文末記号が含まれる場合に対してパターンを定義して行った修正作業について示す（処理 M(α))。8.3.2.3 節ではパターンに基づく機械処理で一括処理できない事例を中心に、人手で行う認定作業について示す（処理 M(β))。8.3.2.4 節では、今回廃止した BCCWJ-DVD 版 （Version 1.0）の属性とタグについて示す。以下、例文中、開始 sentence タグを <s>、終了 sentence タグを</s>と略記する。全角空白を□で表す。

8.3.2.2 処理 M(α)：修正率の高いパターン・認定基準

以下修正率の高いパターンについて示す。これらは最初に修正箇所自動抽出を行い、次に人手で例外を確認し、最後にバッチ処理を行うという手続きで誤りが修正される。

1. 句点類Ｂ[3]のみ、もしくは、句点類Ｂの前に記号類Ｃ[4]があり、かつ、句点類Ｂと記号

[3] 句点類Ｂ：「補助記号−句点」。！．？ の 4 種。

類Cのみで構成されている sentence 要素は、前の sentence 要素の末尾に移動[5]

(1) PB26_00004

（9 桁の英数字はサンプル ID、1 行が 1 sentence 要素、横線上が修正前・横線下が修正後）

<s>でも、お客様が並んでしまったら、それより早めに放送してください」 </s>
<s>。</s>

<s>でも、お客様が並んでしまったら、それより早めに放送してください」。</s>

2. 【原則】〔括弧開〕[6]で終わっている sentence 要素は、次の sentence 要素の頭に〔括弧開〕を移動

(2) PN1b_00009

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ（</s> ←注目点
<s>異議なし！</s>
<s>)。</s>

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ</s>
<s>（異議なし！)。</s>

2-a.【例外処理】〔括弧開〕の前がすべて空白の場合も、それらすべてを次の sentence 要素の頭に移動

(3) OY14_12372

<s>□□□□□□『</s> ←注目点
<s>今度は□一緒にファーストで行きたいね□！！</s>
<s>□』</s>

<s>□□□□□□『今度は□一緒にファーストで行きたいね□！！□』</s>

3. 【原則】〔括弧閉〕のみ、もしくは〔括弧閉〕で始まり、かつ、〔括弧閉〕と記号類D

---

[4] 記号類C：「補助記号−一般」(文境界を示す) ― … ― ・ ～ 【 】 〔 〕 - ･･ ♩ ♪ ♫ ♬ 《 》 ― ― ━ の 20 種。

[5] 条件を規定する演算子は、打消の助動詞を否定とし、「かつ」を論理積とし、「もしくは」を論理和とした場合に、この順で優先順位が高い加法標準形で記述する。

[6] 今回は形態論情報により括弧として定義されている「補助記号−括弧開」「補助記号−括弧閉」の全 12 種を用い、それぞれ〔括弧開〕・〔括弧閉〕と呼ぶ：‘’ “” 〈〉《》「」『』【】[ ] { } 等。

[7]のみで構成された sentence 要素は、前の sentence 要素の末尾に移動

(4) PN1b_00009

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ（</s>
<s>異議なし！</s>
<s>）。</s>　←注目点

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ</s>
<s>（異議なし！）。</s>

3-a.【例外処理】上記 3.を適用した結果、〔括弧閉〕（と記号類Dのまとまり）を移動した先の sentence 要素が〔括弧閉〕と記号類D・E[8]のみで構成されている場合は、それらを前の sentence 要素の末尾に移動

(5) PN2d_00008

<s>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（</s>
<s>？</s>
<s>）。</s>

<s>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（</s>
<s>？）。</s>　←注目点：ここが記号のみ

<s>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（？）。</s>

4. 【原則】〔括弧閉〕で始まり、かつ、〔括弧閉〕に任意の短単位が後続する sentence 要素は、前の sentence 要素の末尾に〔括弧閉〕のみを移動

(6) PN5f_00020

<s>（咽喉？</s>
<s>）…と其奴がね、異に蔑んだ笑い方をしたものです。</s>

---

[7] 記号類D：句点類B、記号類C、「空白」1 種、「補助記号-読点」2 種。

[8] 記号類E：「記号-一般」2,003 種、「記号-文字」255 種、「空白」1 種、「補助記号-AA-一般」78 種、「補助記号-AA-顔文字」2,405 種、「補助記号-一般」（文境界を示さない）444 種、「補助記号-括弧開」12 種、「補助記号-括弧閉」12 種。

<s>（咽喉？）</s>
<s>…と其奴がね、異に蔑んだ笑い方をしたものです。</s>

4-a【例外処理】〔括弧閉〕に記号類Ｆ[9]が続く場合は、記号類Ｆ以外の短単位が出現するまでの範囲を前の sentence 要素の末尾に移動

(7)　OC06_00325　（この例では〔括弧閉〕と読点を移動）
<s> 峠や市街地でも、追い越し禁止道路で前を走る多少遅い車に接近して（</s>
<s>あおるつもりじゃないが。。</s>
<s>)、車が遠慮して道を譲ってくれた時、だいたい頭を下げて追い抜きます。</s>

<s>峠や市街地でも、追い越し禁止道路で前を走る多少遅い車に接近して</s>
<s>（あおるつもりじゃないが。。)、</s>
<s>車が遠慮して道を譲ってくれた時、だいたい頭を下げて追い抜きます。</s>

4-b.【例外処理】空白で始まり、〔括弧閉〕と空白のみで sentence 要素を構成する場合は、それらすべてを前の sentence 要素の末尾に移動

(8)　OY14_12372
<s>□□□□□□□『</s>
<s>今度は□一緒にファーストで行きたいね□！！ </s>
<s>□』</s>　←注目点

<s>□□□□□□□『今度は□一緒にファーストで行きたいね□！！□』</s>

4-c.【例外処理】上記 4-a.を適用した結果、「(？)」「(！)」の文字列を sentence 要素に含む場合には、前後の sentence 要素をひとまとまりにする（8.3.2.3 の “文境界認定を打ち消して文を結合する場合”の 1. を参照）

(9)　PM41_00071
<s>この業界にしては珍しく（</s>
<s>？</s>
<s>)、可愛らしい女性編集長である。</s>

---

9 記号類Ｆ：記号類Ｃ、「補助記号−括弧閉」12 種。

<s>この業界にしては珍しく（？）、可愛らしい女性編集長である。</s>

5. 読点で始まっている場合は、前の sentence 要素の末尾に読点のみを移動

(10) PB45_00024
<s>「ブオノ・ヴェーロ？」</s>
<s>、美味しいだろうと言ったオジサンはイタリア人で、ここに住む孫のためにナポリの店を引き払いやって来たのだという。</s>

<s>「ブオノ・ヴェーロ？」、</s>
<s>美味しいだろうと言ったオジサンはイタリア人で、ここに住む孫のためにナポリの店を引き払いやって来たのだという。</s>

8.3.2.3 処理 M(β) : 修正率の低いパターン・認定基準

以下の例は修正率が低いパターンで、手がかりにより候補を枚挙したうえで人手により修正すべきかどうかを判定する。大きく分けて「文境界を認定して分割する場合」と「文境界認定を打ち消して文を結合する場合」の 2 種類がある。これらは、最初に修正箇所自動抽出を行い、次に人手修正処理をすることで誤りを修正する。

文境界を認定して分割する場合（特に Web データ）

1. sentence 要素の中に顔文字を含み、かつ、その顔文字が文末表示だと考えられる場合は分割

(11) OC06_02963
<s>そーですよ＾＾一番左です＾＾</s>

<s>そーですよ＾＾</s>
<s>一番左です＾＾</s>

2. sentence 要素の中に（涙）等の（X）を含み、かつ、その（X）が文末表示だと考えられる場合は分割

(12) OY14_10161
<s>イプ『</s>
<s>違う！</s>

<s>作りすぎただけだっ（照）ナマモノだから今日中に食え』</s>

<s>イプ</s>
<s>『違う！</s>
<s>作りすぎただけだっ（照）</s>
<s>ナマモノだから今日中に食え』</s>

3. 【特殊事例】空白で文が区切られる場合等は分割

(13) OY14_12372
<s>□□□□□□□□『だね、ローマが一番だったよ□日曜なのでバチカンに行ってミサを聞いた</s>
<s>□□□□□□□□□ミケランジェロも見たよ』□うん、おいらはイタリアはしらない</s>

<s>□□□□□□□□『だね、ローマが一番だったよ□</s>
<s>日曜なのでバチカンに行ってミサを聞いた</s>
<s>□□□□□□□□□ミケランジェロも見たよ』□</s>
<s>うん、おいらはイタリアはしらない</s>

文境界認定を打ち消して文を結合する場合（特に雑誌・Web データ）

1. 係り受け関係を結べる要素が後続し、sentence 要素内に含めるべきと判断される「？」「！」は結合

(14) PM11_00263
<s>今が買い！</s>
<s>の中古ＭＦ一眼レフ</s>

<s>今が買い！の中古ＭＦ一眼レフ</s>

2. 補足を表す丸括弧（括弧内に句点を含まないものに限定）内に「？」「！」が含まれる、かつ、丸括弧内に含まれる要素が体言で終わる場合、結合

(15) OY01_00185
<s>この大会のチラシを、今夜（</s>
<s>昨夜？</s>

<s>）のハードルの練習中にわざわざ七夕ホールまで持ってきてくださったのです！</s>

<s>この大会のチラシを、今夜（昨夜？）のハードルの練習中にわざわざ七夕ホールまで持ってきてくださったのです！</s>

3. 【原則】係り受け関係を結べる要素が、原本レイアウト情報を反映した結果二つの sentence 要素に分割されていて、括弧内に文末記号が含まれない場合は結合

(16) PB1n_00024

<s>すると、</s> ←注目点：紙面上にて改行により sentence 要素が分割されている

<s>「溶岩流が危険だから、逃げるんです」という答えが返ってきたのである。</s>

<s>すると、「溶岩流が危険だから、逃げるんです」という答えが返ってきたのである。</s>

3-a.【例外処理】括弧が強調やタイトル等の目的で用いられている場合

(17) OC01_03215

<s>ゆうべＰＭ９時から日本テレビ「</s>

<s>ものまねバトルオール新ネタ！</s>

<s>夏祭りＳＰ</s>

<s>」に出てましたよ。</s>

<s>ゆうべＰＭ９時から日本テレビ「ものまねバトルオール新ネタ！夏祭りＳＰ」に出てましたよ。</s>

4. 【特殊事例】〔括弧閉〕に丸括弧で注釈が後続する場合は結合しない

(18) PN4c_00011

<s>□だが、農業団体の韓国農業経営人中央連合会は、</s>

<s>「通貨危機で金利負担が膨らみ、農家は今も借金に苦しんでいる。</s>

<s>対策は成功していない」</s>

<s>（政策調整室）と批判的だ。</s>

8.3.3 BCCWJ-DVD 版（Version 1.1）における廃止事項

- BCCWJ-DVD 版（Version 1.0）に規定されていた以下の要素・属性を、BCCWJ-DVD 版（Version 1.1） の M-XML では廃止する。
  sentence タグの属性 type="quasi" : sentence タグの自動付与にあたり、文末記号以外によって認定される特殊な文であることを表すための属性である。今回、文末記号によらない新たな基準に基づき人手で文境界を認定したことで、文の属性（「quasi」は「擬似」の意）として不適切となるため廃止する。
- webLine 要素：「Yahoo!知恵袋」データに対する sentence タグの自動付与にあたり、「文を分断しない範囲で」データ上の物理行（改行記号により自動的に認定される行）を連結した上で認定した、論理行（意味的なまとまりを伴う行）相当のスパンを表す要素である。今回の文境界認定基準と、BCCWJ-DVD 版（Version 1.0）作成時に任意に「文を分断しない」と判断した行との間には矛盾が生じる場合もあるため、不要と判断し廃止する。

8.4 BCCWJ-DepPara における文境界認定

BCCWJ-DVD 版 （Version 1.0） （C-XML、 M-XML） や BCCWJ-DVD 版 （Version 1.1） （M-XML）とは異なる文境界認定基準として、係り受けアノテーションである BCCWJ-DepPara における文境界認定基準（小西 2013）がある。コアデータのみを利用する際には、BCCWJ-DepPara の文境界基準を利用することも考えられる。

係り受けアノテーション従事者は BCCWJ-DVD 版 （Version 1.0） における文境界の問題点として、基準の手がかりが文字列に基づく手法であるために係り受けを分断するような文境界が大量に発生すること、sentence@quasi 要素や sentence@fragment 要素においては要素内に係り先が存在せず、離れた別の sentence 要素に係り先を認定するような現象が起きること、全要素を xpointer などを用いないひとつの XML ファイルとして表現するために ad hoc な後処理がなされ文単位認定に無理が生じていること、実データを見ても必ずしも報告書どおりの処理がなされていないことの四つをあげている。

BCCWJ-DepPara は BCCWJ に対する係り受け・並列構造アノテーションである（浅原・松本 2013)。2012 年 10 月に BCCWJ-DVD 版（Version 1.0）を対象とした最初のバージョンが公開[10]されている。

基本方針として、元の文書構造タグを用いず、文の内容に即して “EOS” ラベルと“Z” ラベルの 2 種類の文境界を認定している（浅原 2013)。 “EOS” ラベルは、係り受け関係がつながる範囲で文を連結したもので、C-XML の最外スパンや M-XML の superSentence 要素に近い基準となっている。“Z” ラベルは、係り受け関係ラベルの一種で “EOS” ラベルで区切られる範囲内に出現する文末記号に対し付与される。“Z” ラベルは文末要素にしか付与されないが、“Z” ラベルを根とする係り受け木の最大スパンを確認することで、局所的な

[10] https://github.com/masayu-a/BCCWJ-DepPara

文の文頭要素が認定できるために、実質的に文の入れ子構造を認定している。括弧内の要素の扱いにおいては、コアデータに出現する括弧で括られた要素の機能を補足・発話・心内・引用・箇条書き・強調の 6 種類に分類し、要素の意味についても調査して、文認定を行っている。

参考文献

浅原正幸（2013）「係り受け関係アノテーション基準の比較」『第 4 回コーパス日本語学ワークショップ予稿集』,81-90.

浅原正幸・松本裕治（2013）「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」『言語処理学会第 19 回年次大会発表論文集』,66-69.

小西光・小山田由紀・浅原正幸・柏野和佳子・前川喜久雄（2013）「BCCWJ 係り受けアノテーション付与のための文境界再認定」『第 4 回コーパス日本語学ワークショップ予稿集』,135-142.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011a）「『現代日本語書き言葉均衡コーパス』形態論情報規程集 第４版 （上）」国立国語研究所内部報告書 LR-CCG-10-05-01

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011b）「『現代日本語書き言葉均衡コーパス』形態論情報規程集 第４版 （下）」国立国語研究所内部報告書 LR-CCG-10-05-02.