『現代日本語書き言葉均衡コーパス』の文境界修正

小西 光 ♢　中村 壮範 ♢　田中 弥生 ♢♡　間淵 洋子 ♢
浅原 正幸 ♣♢　加藤 祥 ♢　立花 幸子 ♢　今田 水穂 ♢♠
山口 昌也 ♣♢　前川 喜久雄 ♣♢　小木曽 智信 ♣♢　山崎 誠 ♣♢　丸山 岳彦 ♣♢
人間文化研究機構 国立国語研究所 ♢ コーパス開発センター ♣ 言語資源研究系
♡ (現在) 神奈川大学 ♠ (現在) 文部科学省 初等中等教育局
{hkonishi, masayu-a}@ninjal.ac.jp

1 はじめに

本稿では『現代日本語書き言葉均衡コーパス』[5](以下 BCCWJ) に対する文境界修正作業について報告する。文境界の認定には (1) 文字情報を用いるもの，(2) 形態素情報を用いるもの，(3) 係り受け関係を用いるものなどが考えられる。現在公開している BCCWJ 第 1.0 版においては，(1) の文字情報による処理で文境界認定が行われているが，不自然な文境界が残っていることが報告されている [8]。人手による作業にせよ自動処理にせよ，より高レベルのアノテーションに基づくものほど高コストになる一方，より厳密な文境界の認定が可能であり，コアデータに対しては先行研究 [4] において (3) の係り受け関係レベルの文境界再認定が行われた。しかしながら，非コアデータ規模になるとこのレベルの修正は非現実的である。そこで，自動認定された形態論情報に基づく (2) のレベルの文境界修正作業を実施した。本稿では実作業の詳細を報告する。

2 BCCWJ 第 1.0 版の文境界とその周辺

本節では，本研究に至るまでの BCCWJ における文境界認定について述べる。まず BCCWJ 第 1.0 版公開時における文境界認定の基準について述べ，次に係り受けアノテーション（BCCWJ-DepPara：[2]）構築時に行った文境界認定（[4]）について述べる。

2.1 BCCWJ 第 1.0 版の文境界認定

まず，BCCWJ 第 1.0 版における文境界について述べる。BCCWJ 第 1.0 版においては文字情報に基づく C-XML 形式と形態論情報に基づく M-XML 形式の 2 種類の XML 形式のファイルでデータが表現されている。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 形式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（[6], pp.136–138.）が基本となっている。他に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）により制約づけられる。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 1）を示しながら解説する。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，直前文の末尾とみなす処理を行う（例 C-1）。論理行 [1] 頭から一つ以上の sentence 要素の並びが存在する場合で行末に文末記号がない場合は sentence 要素とみなす（例 C-2）。論理行中に一つも 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” と略記し，type=“quasi” がつかない sentence 要素を “正則な sentence 要素” と呼ぶ）。文字情報として 9 種類の括弧（括弧類Ａ）[2] の対応などを用いて，文認定時に sentence 要素の入れ子を許している。

括弧内に一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 C-4）。括弧内に一つ以上の文末記号が含まれる場合，括弧内に sentence

[1] 本稿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。

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

要素を認定する（例 C-5）。括弧内に一つ以上の文末記号が含まれ，且つ，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）。

M-XML における文境界認定：

M-XML 形式（[3]，p. 94.）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。方針提案者は C-XML の問題点として以下の 3 点をあげている：sentence 要素がきわめて長くなる場合がある；形態素解析などの入力となる「文」が定めがたい；データを文番号で管理できない。

C-XML で sentence 要素が入れ子になっている場合に，M-XML ではその最も内側（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は，新たに sentence 要素と見なすとともに type=“fragment” という属性を与えて，文断片（以下 “sentence@fragment 要素” と略記）であることを明示する。この際，括弧記号のみから成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 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 形式と M-XML 形式の 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を新たに 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を 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 要素と sentence@quasi 要素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。

2.2 BCCWJ-DepPara の文境界認定

前小節の状況は，C-XML の方式にしても M-XML の方式にしても係り受けアノテーションにとって好ましくない。係り受けアノテーション従事者は BCCWJ 第 1.0 版 における文境界の問題点として以下の 4 点をあげている：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；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；全要素を xpointer などを用いない一つの 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；実データを見ても，必ずしも報告書通りの処理がなされていない。

そこで，[4] は，係り受けアノテーション向けの文境界認定基準を策定し，BCCWJ 第 1.0 版とは異なる文境界をコアデータに対して人手により付与した。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。“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（[1]）で “EOS” ラベルで区切られる範囲内に出現する文末記号の出現に対し付与される。“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まで調査して，文認定を行っている。

3 BCCWJ 第 1.1 版の文境界認定作業

まず，文境界認定の作業方針について述べる。BCCWJ 第 1.0 版の文字情報による自動処理と，BCCWJ-DepPara の係り受けレベルの情報による人手修正との

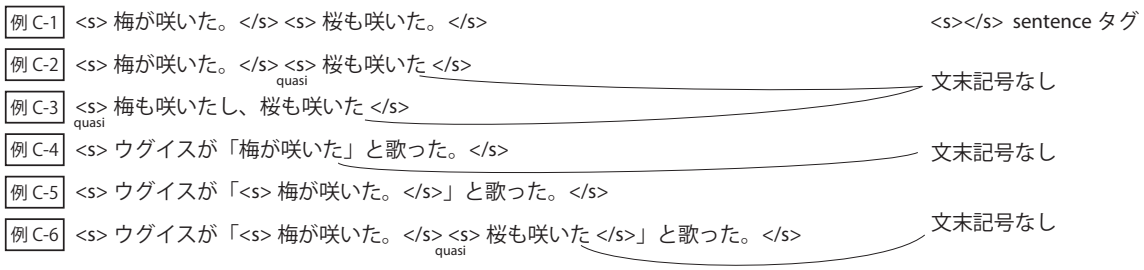

図 1: C-XML における文境界認定

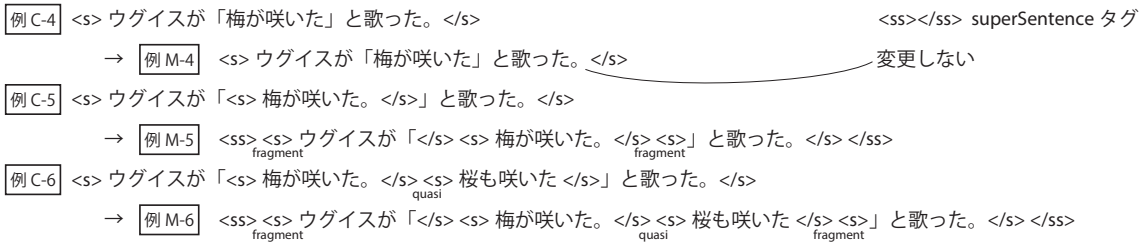

図 2: C-XML から M-XML への変換

中間的な処理として，**形態素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。尚，修正は M-XML のみに処理を行い，C-XML には行わない。中納言のデータは M-XML をベースにしており，本修正が反映される予定である。

修正方法としては，まず文字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，M-XML 形式に変換する際のバグ相当のものを形態素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@quasi 要素，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。まとめると以下のようになる：

処理 C C-XML 形式レベルで認定できる誤りの検出
BCCWJ 第 1.0 版において，文字情報に基づく処理により 9 種類の括弧内（括弧類A）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 [3] 発見された。顔文字に埋め込まれた文末記号があったり，括弧が対応していない事例もあったり，人手で確認した。

処理 M M-XML 形式レベルで認定できる誤り検出
処理 C が完了後，形態素情報を用いた誤り検出を行う。形態素情報を用いた誤り検出においては，国語研コーパス開発センター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，人手で形態素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（自動修正箇所抽出 → 人手例外確認 → バッチ修正処理）。

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（自動修正箇所抽出 → 人手修正処理）。

表 1 に，処理 M の文境界認定基準について示す。まず現存する superSentence 要素を踏襲することを前提に sentence タグを付与する。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，括弧内に文末記号が

[3] なお，各箇所で複数の文境界の修正が発生するために実際に修正する文境界はこの数字より大きい。

表 1: 処理 M の概要

摘要 (“s 要素” = “sentence 要素”)
処理 M(α)：修正率の高いパターン・認定基準
1. 句点類Bのみ，もしくは，句点類Bの前に記号類Cがあり，且つ，記号類Cのみで構成されている s 要素は，前の s 要素の末尾にそれらを移動
2. 〔括弧開〕で終わっている s 要素は，次の s 要素の頭に〔括弧開〕を移動
3. 〔括弧閉〕のみ，もしくは，〔括弧閉〕で始まり，且つ，〔括弧閉〕と記号類Dのみで構成された s 要素は，前の s 要素の末尾に〔括弧開〕（とそれら記号類Dのまとまり）を移動
4. 〔括弧閉〕で始まり，且つ，〔括弧閉〕に何らかの短単位が続いていく s 要素は，前の s 要素の末尾に〔括弧閉〕のみを移動
5. 読点で始まっている場合は，前の s 要素の末尾に読点のみを移動
処理 M(β)： 修正率の高いパターン・認定基準
文境界を認定して分割する場合（特に Web 関連データ）
1. s 要素の中に顔文字を含み，且つ，その顔文字が文末表示だと考えられる場合
2. s 要素の中に（照）（涙）等の（X）を含み，且つ，その（X）が文末表示だと考えられる場合
3. 【特殊事例】空白で文が区切られる場合等
文境界認定を打ち消して文を結合する場合（特に雑誌・Web 関連データ）
1. 「？」「！」等に係り受け関係が結べる要素が後続し，s 要素内に含めるべきと判断される文末記号
2. 補足を表す丸括弧（括弧内に句点類Bを含まないものに限定）
3. 原本レイアウト情報を反映した結果，係り受け関係を結べる要素が二つの s 要素に分割されていて，括弧内に文末記号が含まれていない場合
4. 【特殊事例】〔括弧閉〕に丸括弧で注釈が後続する場合は変更しない

含まれない場合には sentence タグは付与しない（例 C-4, 例 M-4 を踏襲）。処理 M(α) 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修正作業を行う。人手で例外を確認し，必要に応じて新たなパターンを追加する。処理 M(β) 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，認定を人手で行う。句点類Bは「補助記号-句点」。!. ？の 4 種，記号類C :「補助記号-一般」（文境界を示す）―…－・～〖〗〔〕‐･･･♩♪♫♬《》－－ー の 20 種，記号類D :「空白」1 種,「補助記号-一般」（文境界を示す）20 種,「補助記号-句点」4 種,「補助記号-読点」2 種（、，）からなる。

以下，文境界修正に伴う M-XML XML Schema の変更点について示す。BCCWJ 第 1.0 版に規定されていた，文末記号を含まないことを表す sentence タグの quasi 属性は，BCCWJ 第 1.1 版 M-XML では廃止する。sentence タグ内に文末記号があるかどうかにより一意に決まるものであり，人手で付与されたものではないため quasi 属性が不要であると考える。　また，web データに対して自動で付与された論理行相当のスパンを表す webLine タグについても，BCCWJ 第 1.1 版 M-XML では廃止する。今回の作業により，人手により web データの物理行が文に相当するか否か，文に相当する場合には連結するか否かについて判断を行ったために，webLine タグの存在意義がなくなった。

表 2 に処理 M についての修正件数を示す。処理 M では，パターンに適合する用例をコーパス管理システム大納言 [7] 上で帳票形式で枚挙したうえで，修正が必要か不要かを 1 次チェックする。処理 M(α) については修正が不要なものをはじいて，2 次処理でバッチ処理によりデータベース登録作業を行う。登録後，人手で最終確認を行う。処理 M(β) については修正が必要なもののみを残して，2 次処理で人手処理によりデータベース登録作業を行う。

表 2: 修正件数 (2015 年 1 月 14 日現在)

	タグ追加	タグ削除	タグ移動
バッチ処理	140	36,988	124,366
人手修正	48,038	53,839	312
ログに記載なし	42	20	94
合計	48,220	90,847	124,772

4　おわりに

本稿では 2013 年から 2014 年にかけて実施した『現代日本語書き言葉均衡コーパス』に対する文境界修正作業の概要について示した。本稿で示した修正済みのデータ『現代日本語書き言葉均衡コーパス』第 1.1 版を 2014 年度末に公開する予定である。

謝辞

本研究の一部は国語研基幹型プロジェクト「コーパスの管理・運用」国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

[1] 浅原．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, pp. 81–90, 2013.

[2] 浅原, 松本．『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, pp. 66–69, 2013.

[3] 国立国語研究所．『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版, 2011.

[4] 小西, 小山田, 浅原, 柏野, 前川. BCCWJ 係り受け関係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, pp. 135–142, 2013.

[5] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den. Balanced Corpus of Contemporary Written Japanese. Language Resources and Evaluation, Vol. 48, No. 2, pp. 345–371, 2014.

[6] 山口, 高田, 北村, 間淵, 大島, 小林, 西部. 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2. JC-D-10-24, 2011.

[7] 小木曽, 中村．『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用．自然言語処理, Vol. 21, No. 2, pp. 301–332, 2014.

[8] 田野村．コーパスと日本語学，「BCCWJ の資料的特性」．講座日本語コーパス 6. 朝倉書店, 2014.