『現代日本語書き言葉均衡コーパス』の文境界修正作業の進捗

小西 光 *†　中村 壮範 †　田中 弥生 †　浅原 正幸 †　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 ‡†
（国立国語研究所　† コーパス開発センター　‡ 言語資源研究系）

Revision of Sentence Boundaries in the BCCWJ DVD Edition

Hikari Konishi, Takenori Nakamura, Yayoi Tanaka, Masayuki Asahara, Mizuho Imada
Masaya Yamaguchi, Kikuo Maekawa, Toshinobu Ogiso, Makoto Yamazaki, Takehiko Maruyama
(National Institute for Japanese Language and Linguistics)

1. はじめに

本発表では現在国立国語研究所コーパス開発センターで進められている『現代日本語書き言葉均衡コーパス』(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) 文字列レベルの情報を用いるもの，(2) 形態素列レベルの情報を用いるもの，(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 DVD 第 1 版においては，(1) の文字列レベル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。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（小西ほか (2013)）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り，現在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ている。本稿では実作業の詳細と進捗状況を報告する。

2. 文境界認定手法についての関連研究—手がかり・CSJ における研究動向・BCCWJ の現況

本節ではまず文境界認定基準策定のために必要な手がかりについて述べ，次に BCCWJ の前に作成された『日本語話し言葉コーパス』（以下 CSJ）における文境界認定に関する関連研究を示し，最後に BCCWJ 第 1 版公開時における文境界認定とその後の研究動向について述べる。

2.1 文境界認定基準における手がかり

文境界認定基準において何らかの「手がかり」を用いて規則を人手で記述する必要がある。ある程度文境界認定作業を自動化するために何を「手がかり」に使うかが重要である。以下では「手がかり」として，(1) 文字列レベルの情報を用いるもの，(2) 形態素列レベルの情報を用いるもの，(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) 文字列レベルの情報とは，句点などに基づいて文境界を用いる手法である。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て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を用いた規則を記述し，括弧の対応を取る手法がある。

* hkonishi@ninjal.ac.jp

(2) 形態素列レベルの情報とは，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。一方，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) 係り受けレベルの情報とは，係り受け解析により認定される係り受け関係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 sentence 要素の入れ子を認定したりする。

この三種類以外の手がかりとして，CSJ においては形態素列レベルの情報で認定された節境界の情報や音声のポーズ長などを用いる研究が知られている。次節では CSJ における文境界認定についての様々な取り組みについて紹介する。

2.2　CSJ における文境界認定と関連技術

丸山ほか (2006) は CSJ における統語的単位について議論している。南 (1974) による従属節の分類に基づき，「絶対境界・強境界・弱境界」と呼ばれる三段階のレベルの節境界が設計・定義され，各従属節の境界にラベルが付与された。以降の研究では，「絶対境界」が CSJ における文境界として利用されてきた。

下岡ほか (2004) では CSJ の講演の書き起こしテキストの文境界認定について，話者がとるポーズ長と前後の単語情報に基づいた文境界認定手法を提案した。これに対し，田島ほか (2003) は同じデータでポーズ長が得られないことを想定し，コスト最小法の形態素解析器を用いて，句点を挿入した場合と挿入しない場合との出力コストの比較を行い文境界認定を行う手法を提案した。一種の言語モデル尤度を用いた手法とも言える。福岡・松本 (2005) は田島らの手法を拡張して言語モデル尤度を特徴量とした文境界認定手法を提案している。下岡ほか (2005) は新たに係り受け情報を用いて文境界認定する手法を提案している。この手法においては話し言葉特有の係り受け現象を扱う係り受け解析器を導入し，ポーズ長・節末表現・単語情報・文節間距離・係り受け関係などを複合的に組み合わせて文境界を認定している。西光ほか (2009) は丸山らの三つのレベルを全て認定する手法を提案している。特徴量として局所的な隣接要素間の係り受け関係のみを扱うことにより精度の向上が達成されたことを報告している。またこの論文では音声認識結果からの文境界認定についても議論している。

このように CSJ においてはさまざまなレベルの情報を利用した文境界認定手法が提案されてきた。しかしながら，CSJ 関連の文境界認定の重要な問題として，文末認定（文の最右要素）しか行われておらず，文の最右要素と最左要素の対応が取られていないという点がある。入れ子による観点に基づいた文境界を認定がなされていないために，基本的にはチャンキングなどの系列ラベリングで処理可能なレベルの文境界認定にとどまっている。

2.3　BCCWJ における文境界認定 — 本研究に至るまでの足取り

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 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：浅原・松本 (2013)）構築時に行った文境界認定（小西ほか (2013)）について述べる。

2.3.1　BCCWJ 第 1 版における文境界認定

まず，BCCWJ 第 1 版における文境界について述べる。BCCWJ 第 1 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 形式の二種類の XML 形式のファイルでデータが表現されている。この二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列レベルの情報を用いた自動処理に基づく文境界認定（山口ほか (2011), pp.136–138.）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 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）。文字列レベルの情報として 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 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）。

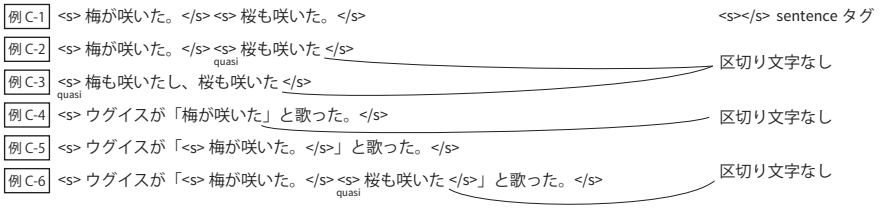

図 1　C-XML における文境界認定

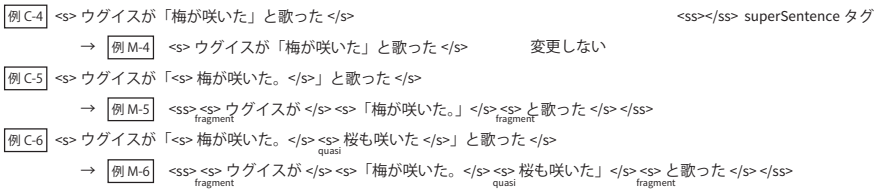

図 2　C-XML から M-XML への変換

[*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（ ）〔〕［ ］｛ ｝〈〉《》「」『』【】9 種。

M-XML における文境界認定：

M-XML 形式（国立国語研究所 (2011)，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三点をあげている：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。

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。

2.3.2　BCCWJ-DepPara における文境界認定

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 DVD データ 第 1 版 における文境界の問題点として以下の四点をあげている：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，小西ほか (2013) 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，コアデータに対して人手による全数確認により，BCCWJ DVD 第 1 版とは異なる文境界を付与した。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（浅原 (2013)）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. 今回の文境界認定作業の概要

3.1 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 第 1 版の文字列レベルの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中間的な処理として，**形態素列レベルの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 形式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，M-XML 形式に変換する際のバグ相当のものを形態素列レベルの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 DVD 第 1 版において，文字列レベルの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，全数人手で確認する。

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，形態素列レベルの情報を用いた誤り検出を行う。形態素列レベルの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，人手で形態素レベルの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

[*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。

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）。

3.2 文境界認定基準

3.2.1 文境界認定基準の前提

文境界認定基準の前提について示す。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）。以下 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修正作業を行う（処理 M(α)）。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。本稿に示すパターンは 2014 年 2 月時点までに設定したものであり，今後作業時に新たなパターンが増えることが考えられる。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，認定を人手で行う（処理 M(β)）。以下 sentence 要素を s 要素と略記する。

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（自動修正箇所抽出 → 人手例外確認 → バッチ修正処理）の手続きで誤り修正される。

1. 句点類Ｂ[*4]のみ，もしくは，句点類Ｂの前に記号類Ｃ[*5]があり，且つ，句点類Ｂと記号類Ｃのみで構成されている s 要素は，前の s 要素の末尾に移動[*6]

例 F-1：PB26_00004 （1 行が 1s 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〕[*7]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

例 F-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>
```

(a) 【例外処理】〔括弧開〕の前がすべて空白[*8]の場合も，それらすべてを次の s 要素の頭に移動

[*4] 句点類Ｂ：「補助記号-句点」。！．？ の 4 種。

[*5] 記号類Ｃ：「補助記号-一般」（文境界を示す） ―…－・～〖〗〔〕‐‥」♪♫《》－－ー の 20 種。

[*6] 条件を規定する演算子は，打消の助動詞を否定とし、「且つ」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。

[*7] 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：〝 〟 〈 〉 「 『 【 □ （ ） < > 〔 〕 。

[*8] 例中□は全角の「空白」を表す。

例 F-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，且つ，〔括弧閉〕と記号類D[*9]のみで構成されたs要素は，前のs要素の末尾に移動

例 F-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>
```

(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s要素が，〔括弧閉〕と記号類D・E[*10]のみで構成されている場合は，前のs要素の末尾に，それらを移動

例 F-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，且つ，〔括弧閉〕に任意の短単位が後続するs要素は，前のs要素の末尾に〔括弧閉〕のみを移動

例 F-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>
----------
<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>
```

(a) 【例外処理】〔括弧閉〕に記号類F[*11]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s要素の末尾に移動

例 F-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>
```

(b) 【例外処理】空白で始まり，〔括弧閉〕と空白のみでs要素を構成する場合は，それ

*9 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。「補助記号-読点」2種は、，からなる。

*10 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。

*11 記号類F：「補助記号-一般」（文境界を示す）20種，「補助記号-読点」2種，「補助記号-括弧閉」12種。

らすべてを前の s 要素の末尾に移動

例 F-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>
```

(c) 【例外処理】上記 2.，4.，4.(a) を適用した結果，「(？)」「(！)」の文字列を s 要素に含む場合には，前後の s 要素をひとま とまりにする
(“文境界認定を打ち消して文を結合する場合” の 1. を参照)

例 F-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>
----------
<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 s 要素の末尾に読点のみを移動

例 F-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>
```

6. 【原則】発話者「～」や発話者：「～」で表記される場合は，一つの s 要素とする

例 F-11：OY13_01632

```
<S>弟子：「</S>
<S>どうすれば先生のようになれるのでしょう？」</S>
----------
<S>弟子：「どうすれば先生のようになれるのでしょう？」</S>
```

(a) 【例外処理】括弧内が二つの s 要素に分かれる場合は，括弧内一つめの s 要素のみを前の s 要素に移動

例 F-12：OY14_08173

```
<S>ゆり「</S>                    <S>ゆり「美羅！</S>
<S>美羅！</S>                    <S>真ちゃん何があったの？」</S>
<S>真ちゃん何があったの？」</S>
```

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で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 種類ある。これらは（自動修正箇所抽出 → 人手修正処理）の手続きで誤り修正される。

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，その顔文字が文末だと考えられる場合

例 F-13：OC06_02963（縦線左が修正前・縦線右が修正後）

```
<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. s 要素の中に（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

例 F-14 : 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『違う！</S>
<S>作りすぎただけだっ（照）</S> ←注目点
<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

例 F-15 : 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.「？」「！」等に係り受け関係を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

例 F-16 : 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（括弧内に句点類Ｂを含まないものに限定）

例 F-17 : 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セタ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セタ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 s 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

例 F-18 : PB1n_00024

<S>すると、</S>　←注目点：紙面上にて改行があり，s 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>

(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

例 F-19 : 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変更しない

例 F-20 : 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>

4. おわりに

本発表では『現代日本語書き言葉均衡コーパス』の文境界修正作業について報告した。

以下作業進捗について示す。M-XML に基づく文境界は約 587 万箇所にのぼる。2013 年度

内に処理 C の 文字列レベルの修正作業（約 6,000 箇所）が完了する予定である。表 1 に処理 M の想定修正箇所数の現時点での集計結果を示す。処理 M(β) の非コアについては今後集計するが，母集団の規模比率からコアの 100 倍を見込んでいる。2013 年度末までに形態素列レベルの修正作業のうち処理 M(α) に着手し，2014 年度は処理 M(α) と M(β) を進める。その過程で今後新たな例外事例が認定された場合には新たなパターンを記述することになるだろう。本稿におけるパターンは先行して行われた処理において発生した不備に対処するためのものである。本作業終了後に「最初にどういう文境界基準を立てるべきであったか」という課題に取り組む必要がある。これについては機械学習などを用いて文境界認定規則を自動獲得したうえで人間にとって可読性の高い規則に汎化するということを行いたい。

表 1　処理 M の想定修正箇所数

摘要 (“s 要素” = “sentence 要素”)	コア	非コア
処理 M(α)：(3.2.2 節) 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	104	2,898
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	222	36,086
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	59	4,547
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	220	35,672
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	5	979
6. “発話者「～」” や “発話者：「～」” で表記される場合は，一つの s 要素とする	66	2,062
合計	676	82,244
処理 M(β)：(3.2.3 節) 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	35	未集計
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	194	未集計
3.【特殊事例】空白で文が区切られる場合等	1,397	未集計
合計	1,626	未集計
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	1,547	未集計
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）	77	未集計
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	902	未集計
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	400	未集計
合計	2,926	未集計

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

浅原 (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, pp. 81–90.
浅原・松本 (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, pp. 66–69.
福岡・松本 (2005).「Support Vector Machines を用いた日本語書き言葉の文境界修正」　言語処理学会第 11 回年次大会発表論文集, pp. 1221–1224.
国立国語研究所 (2011).『『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版』.
小西・小山田・浅原・柏野・前川 (2013).「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.
丸山・高梨・内元 (2006).「第 5 章 節単位情報」　国立国語研究所報告書 124 日本語話し言葉コーパスの構築法, pp. 255–322.
南 (1974).『現代日本語の構造』　大修館書店.
西光・秋田・高梨・尾嶋・河原 (2009).「局所的な係り受けの情報を用いた話し言葉の節・文境界の推定」　情報処理学会論文誌, 50:2, pp. 544–552.
下岡・南條・河原 (2004).「講演の書き起こしに対する統計的手法を用いた文体の整形」　自然言語処理, 11:2, pp. 67–83.
下岡・内元・河原・井佐原 (2005).「日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化」　自然言語処理, 12:3, pp. 3–18.
田島・難波・奥村 (2003).「形態素解析器を利用した講演書き起こしの文境界検出について」　情報科学技術フォーラム (FIT 2003), pp. 155–156.
山口・高田・北村・間淵・大島・小林・西部 (2011).『特定領域研究「日本語コーパス」平成 22 年度研究成果報告書『現代日本語書き言葉均衡コーパス』における電子化フォーマット Ver.2.2』, JC-D-10-24.