『現代日本語書き言葉均衡コーパス』の文境界修正

著者	小西 光，中村 壮範，田中 弥生，間淵 洋子，浅原 正幸，立花 幸子，加藤 祥，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
雑誌名	国立国語研究所論集
号	9
ページ	81-100
発行年	2015-07
URL	http://doi.org/10.15084/00000462

国立国語研究所論集 (NINJAL Research Papers) 9: 81–100 (2015)
ISSN: 2186-134X print/2186-1358 online

『現代日本語書き言葉均衡コーパス』の文境界修正

小西　光[a]　中村壮範[b]　田中弥生[c]　間淵洋子[a]　浅原正幸[d]
立花幸子[e]　加藤　祥[f]　今田水穂[g]　山口昌也[d]
前川喜久雄[d]　小木曽智信[d]　山崎　誠[d]　丸山岳彦[d]

[a]国立国語研究所 コーパス開発センター 非常勤研究員
[b]マンパワーグループ株式会社
[c]国立国語研究所 理論・構造研究系 非常勤研究員
[d]国立国語研究所 言語資源研究系／コーパス開発センター
[e]国立国語研究所 コーパス開発センター 技術補佐員［–2015.03］
[f]国立国語研究所 コーパス開発センター プロジェクト研究員
[g]文部科学省／国立国語研究所 コーパス開発センター プロジェクト研究員［–2014.03］

要旨

『現代日本語書き言葉均衡コーパス』第 1.0 版（Maekawa et al. 2014）（以下 BCCWJ）には「文境界」の情報がアノテーションされているが，その認定基準の妥当性について従来から様々な指摘がある（小西ほか 2014, 長谷川 2014, 田野村 2014）。この問題に対処するために，国立国語研究所コーパス開発センターでは 2013 年から 2014 年にかけて，BCCWJ の修正を行った。本稿ではその修正作業について報告する。第 1.0 版における BCCWJ 文境界情報の問題は，コーパス構築の過程において文境界を含む文書構造タグの整備と形態素列レベルの情報の整備とを並行して行ったために，文字情報を用いる文境界処理にとどまったことに由来する。今回，形態論情報に基づいた文境界基準を策定し，問題の解消を試みた。文境界修正の指針を示すとともに，文境界修正に用いた作業環境と，修正件数について報告する *。

キーワード：現代日本語書き言葉均衡コーパス，文境界，アノテーション，修正基準，修正環境

1. はじめに

本稿では『現代日本語書き言葉均衡コーパス』第 1.0 版（Maekawa et al. 2014）（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（i）文字情報を用いるもの，（ii）形態論情報を用いるもの，（iii）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，（i）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（小西ほか 2014，長谷川 2014，田野村 2014）。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る。BCCWJ は人手を介した高精度のアノテーションが付された 100 万語規模のコアデータと自動解析を主とする 1 億語規模の非コアデータからなる。前者のコアデータに対しては先行研究（小西ほか 2013）において（iii）の係り受け関係レベルの文境界再認定が人手によって行われた。しかしながら，後者の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

* 本研究は国立国語研究所コーパス開発センターの予算によって実施したものである。本稿の内容は平成 26 年 12 月 16 日開催第 120 回 NINJAL サロンでの発表をもととしている。

く（ii）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。

本稿の構成は以下の通りである。まず 2 節では文境界認定手法についての関連研究を示す。3 節で今回実施した文境界認定作業の基準について示す。4 節で修正環境と修正件数を示し，5 節でまとめと今後の課題について示す。

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 第 1.0 版の現況

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 1.0 版公開時の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は何らかの「手がかり」を用いて規則を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，（i）文字情報を用いるもの，（ii）形態論情報を用いるもの，（iii）係り受け関係を用いるものの三種類について詳しく述べる。

(i)　文字情報に基づく認定とは，句点などに基づいて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。

(ii)　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 UniDic 品詞体系（小椋ほか 2011）「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。

(iii)　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単位の入れ子を認定したりする。

CSJ においては文境界認定のためにこの三種類の手がかりのほかに音声のポーズ長を用いている。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」を CSJ における文境界としたうえで，各種特徴量から文境界を自動認定する手法を検討している。

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）は丸山ほか（2006）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も報告している。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 においては様々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。文要素の入れ子を考慮した文境界の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―本研究に至る経緯

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：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 1.0 版における文境界認定

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 C-XML 形式と 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情報を用いた自動処理に基づく文境界認定（山口ほか 2011: 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[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）。文字情報として九対の括弧（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）。

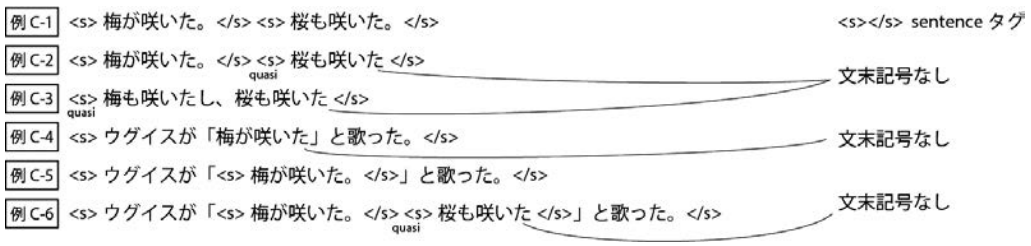

図 1　C-XML における文境界認定

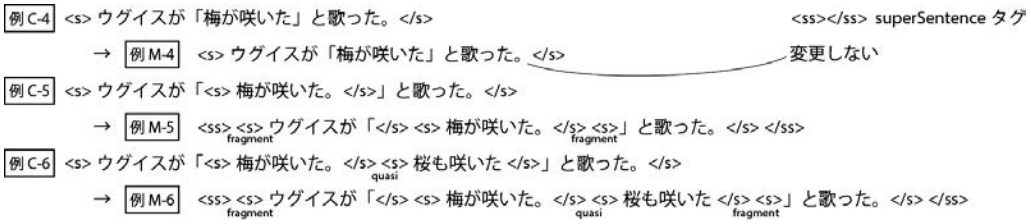

図 2　C-XML から M-XML への変換

【M-XML における文境界認定】

M-XML 形式（国立国語研究所 2011: 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

[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（いわゆる，桁折り）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：UniDic 品詞体系「補助記号 - 括弧開」「補助記号 - 括弧閉」のうち（）〔〕［］｛｝〈〉《》「」『』【】九対。

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をあげている。

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な sentence 要素とし，外側（上位）にある sentence は superSentence とする。そのうえで，superSentence の内側にありながら正則な sentence 要素の外側に位置する部分は，新たに sentence 要素とみなすとともに 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 の二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 要素とみなす（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。正則な 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，正則な 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）。括弧内の二種類の 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 2 では，問題となる例を示した。

2.3.2 BCCWJ-DepPara における文境界認定

前節の状況は，いずれの方式であっ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，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 などを用いない一つの XML ファイルとして表現するために不自然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通りの処理がなされていないことの四つをあげている。

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 第 1.0 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” ラベルの二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・強調の六種類に分類し，要素の意味についてまで調査して，文認定を行っている。

以下具体的な事例を見ながら，BCCWJ に対する係り受けアノテーション BCCWJ-DepPara（浅原・松本 2013）で用いた文境界再認定基準について概観する。

BCCWJ-DepPara では，以下の三点のいずれかを満たすものの結合により文として再認定する。

① 括弧や引用符などの括り記号で括られた発話や引用･補足部分を挟んだり，引用の助詞「と」で受けたりして係り受け関係を結べる要素が前・中・後に接続する
② 箇条書き（改行を伴う）を内包する要素が前・中・後に接続する（主に Web 媒体）
③ 本来一文であるべきものが，書き手による意図的な改行で分割されている（主に Web 媒体）

図 3 に結合により文として再認定する例を示す。図中 [ ] 内の 9 桁の英数字は BCCWJ のサンプル ID を表す。また「↲」は改行記号を表す。

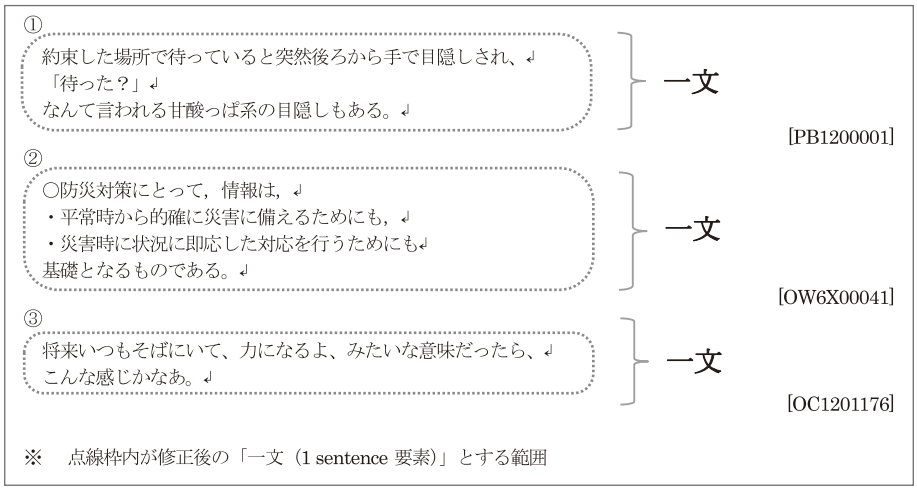

図 3　BCCWJ-DepPara において結合により一文と認定するもの

一方，以下の場合は，現状のまま一つの文にまとめ上げることはしない。

① 倒置部分が改行されている
② 改行を伴って文がねじれている
③ 接続助詞ではなく接続詞「と」「っと」と判断されるものが文頭にくる

④　前後の sentence 要素と括弧や引用符などで括られた要素がそれぞれ独立して係り受け関係にない

図 4 に結合しない例を示す。

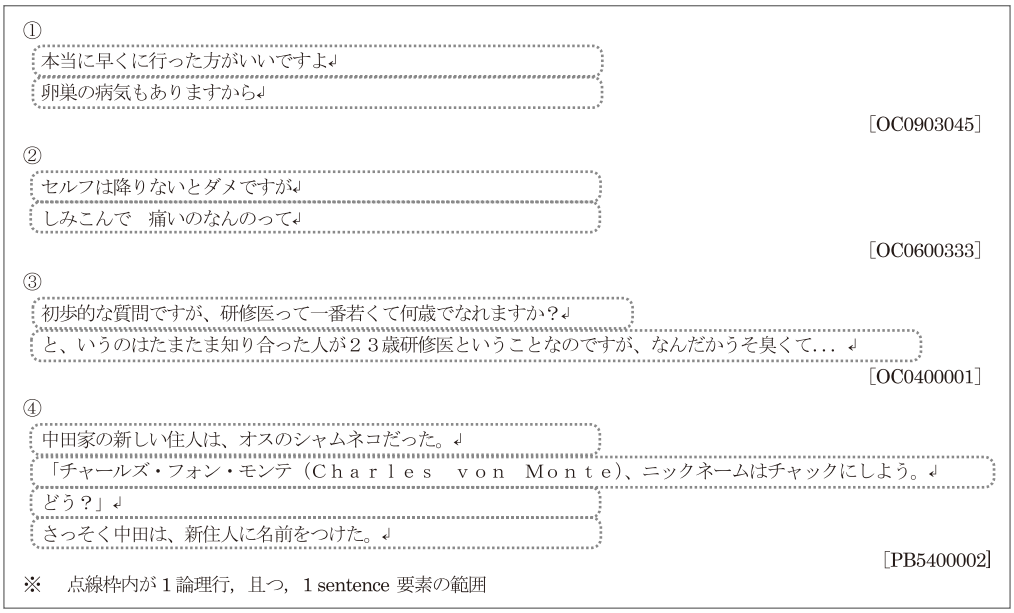

図 4　BCCWJ-DepPara において結合して文認定しないもの

3. BCCWJ 第 1.1 版における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-DepPara では 100 万語規模のコアデータ全体に対し係り受け関係の情報による文境界認定作業を人手によって行った。しかしながら，1 億語規模の BCCWJ 全体に対してこのレベルの文境界認定作業を行うことは非現実的である。一方, BCCWJ 第 1.0 版には自動解析ながらも形態論情報が付与されている。そこで, 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実施する。この文境界認定作業は，基準の一貫性のために非コアデータだけでなく，BCCWJ-DepPara で修正したコアデータについても行う。

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，次に M-XML 形式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 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九対の括弧（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[3] 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 形式レベルで認定できる誤りの検出】

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，形態論情報を用いたパターンを人手で記述した。このパターンの認定で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）に基づいて二種類の処理を行う。

[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。この修正に伴い，形態論情報･文書構造タグの修正が必要な場合がある。この場合，形態論情報･文書構造タグについても修正する。

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として今回踏襲する BCCWJ 第 1.0 版の文境界認定基準三点について示す。

一点目：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。

二点目：付属語から始まる，付属語で終わる[4]，付属語のみの sentence 要素の発生を認める。

三点目：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 を踏襲する）。

以下，3.2.2 節では，上記の処理 M(α)，すなわち括弧内に文末記号が含まれる場合に対してパター

[3] 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。
[4] 付属語は助詞・助動詞からなる。格助詞・接続助詞を含む。

ンを定義して行ったバッチ処理について示す。3.2.3 節では，処理 M(β)，すなわちパターンに基づくバッチ処理で一括処理できない事例を中心に行った，人手作業について示す。3.2.4 節では，今回廃止した BCCWJ 第 1.0 版の属性とタグについて示す。以下 sentence 要素，開始 sentence タグを <s>，終了 sentence タグを </s> とする。全角空白を□で表す。各用例の 9 桁の英数字は BCCWJ のサンプル ID を表す。また，1 行が 1sentence 要素，横線上が修正前・横線下が修正後である。

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，まず修正箇所自動抽出を行い，人手により例外を確認し，最後にバッチ処理を行うことにより誤りが修正される。

1. 句点類 B[5] のみ，もしくは，句点類 B の前に記号類 C[6] があり，且つ，句点類 B と記号類 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 [7]

(1)　PB2600004

<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〕[8] 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)　PN1b00009

<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>

[5] 句点類 B：UniDic 品詞体系「補助記号 - 句点」。！．？の 4 種。
[6] 記号類 C：UniDic 品詞体系「補助記号 - 一般」（文境界を示す）－ … － ・ ～【】〔〕- ‥ ♩ ♪♫♬《》―ー―の 20 種。
[7] 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。
[8] 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 12 種を用い，それぞれ〔括弧開〕・〔括弧閉〕と呼ぶ：‘’ “” 〈〉《》「」『』【】［］｛｝（）〔〕＜＞

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)　OY1412372

<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 sentence 要素は，前の sentence 要素の末尾に移動

(4)　PN1b00009

<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[10] のみで構成されている場合は，前の sentence 要素の末尾に，それらを移動

(5)　PN2d00008

<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>

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

[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。

<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 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>

<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[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 sentence 要素の末尾に移動

(7)　OC0600325（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)　OY1412372

<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

---

[11] 記号類 F：記号類 C,「補助記号-読点」2 種,「補助記号-括弧閉」12 種。

<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（後述する 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)　PM4100071
<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 sentence 要素の末尾に読点のみを移動

(10)　PB4500024
<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>

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」の二種類がある。これらは，まず修正箇所自動抽出を行い，その後人手修正処理を行う手順で誤りを修正する。

【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，且つ，その顔文字が文末表示だと考えられる場合，分割する

(11)　OC0602963
<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, 且つ, その（X）が文末表示だと考えられる場合, 分割する

（12） OY1410161
<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も分割する

（13） OY1412372
<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「？」「！」は結合する

（14） PM1100263
<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「？」「！」が含まれる場合，且つ，丸括弧内に含まれる要素が体言で終わる場合，結合する

（15） OY0100185
<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する

（16） PB1n00024
<s> すると、</s>　　←注目点：紙面上に改行があり，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で且つ括弧内に「！」「？」が含まれる場合，結合する

（17） OC0103215
<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>　　←注目点
<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） PN4c00011
<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>

3.2.4 廃止事項

BCCWJ 第 1.0 版に規定されていた以下の属性・要素を，BCCWJ 第 1.1 版 M-XML では廃止する。

- sentence タグの属性 type=“quasi”
- webLine 要素

sentence タグの属性 type=“quasi” は，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」（2.3.1）を表すための属性であり，「quasi（擬似）」の意味が表す通り，文境界認定を留保する意図で設けたものである。

webLine 要素は，web 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（web データ内の改行記号を手がかりとして自動的に認定される行）を連結したうえで認定した，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。web 上の文章では，文末記号の用いられない文や書き手による論理行途中で改行された文が多く存在し，書籍・新聞等の文の様相とは異なるため，文境界認定を留保し，「文を分断していない行」のみを保証する意図で設けたものである。

これらの「文境界認定を留保した」ことを表すタグは，いずれも，今回人手による文境界認定が行われたことで，不要となるため廃止する。

4. 修正環境と修正件数

4.1 修正環境

修正作業には BCCWJ の形態論情報アノテーション支援システムである『大納言』（小木曽・中村 2014）を用いた。文境界情報を含む文章構造タグが関係データベースに格納されており，DVD に収録されている帳票形式のファイルと XML ファイルがシステムから出力されるようになっている。今回新たに『大納言』上に文境界修正モードを作成した。

作業の手順は次の通りである。3.2.3 節で処理 M(β) の手続きを説明した通り，文境界修正箇所はあらかじめ自動抽出されている。作業者は Excel 上の修正対象箇所リストを見ながら，『大納言』画面上で対象となる箇所を検索により表示させ，必要があれば sentence タグ情報を修正する。『大納言』画面上で行う作業は大きく分けて二つある。一つは sentence タグを挿入する作業，もう一つは sentence タグを削除する作業である。図 5 に sentence タグ挿入時の画面を，図 6 に sentence タグ削除時の画面を示す。タグの移動は，この挿入と削除を組み合わせることにより行う。このタグの挿入・削除作業は対象となるデータ（一サンプルに表示される形態素数）・作業者・作業環境（国立国語研究所内・在宅）などにより異なるが，一時間あたりおおよそ 20 件から 100 件程度のペースで行われた。

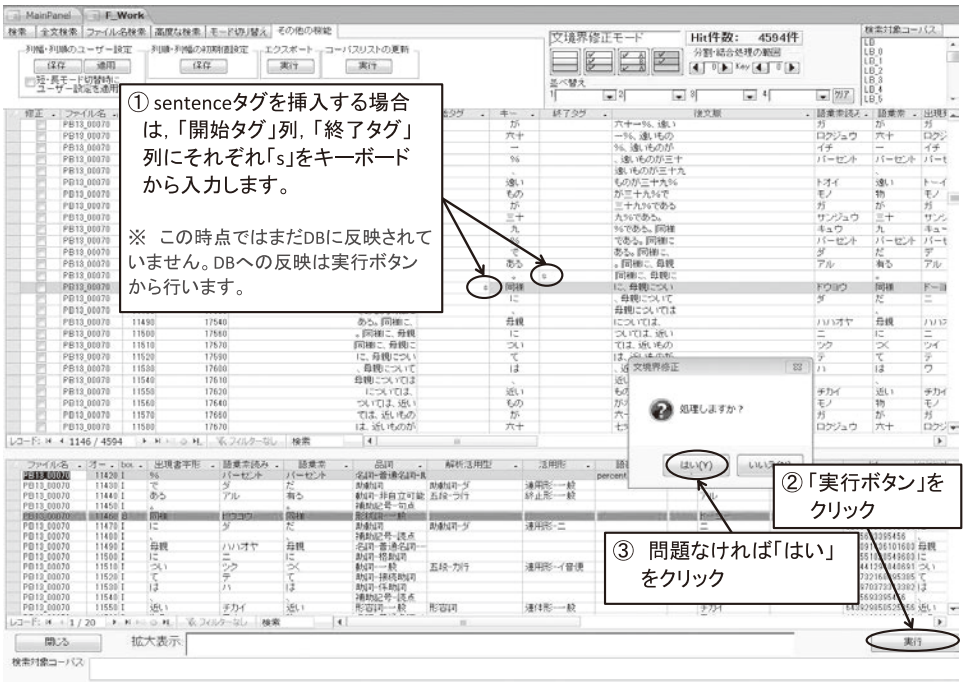

図 5　sentence タグの挿入作業

図 6　sentence タグの削除作業

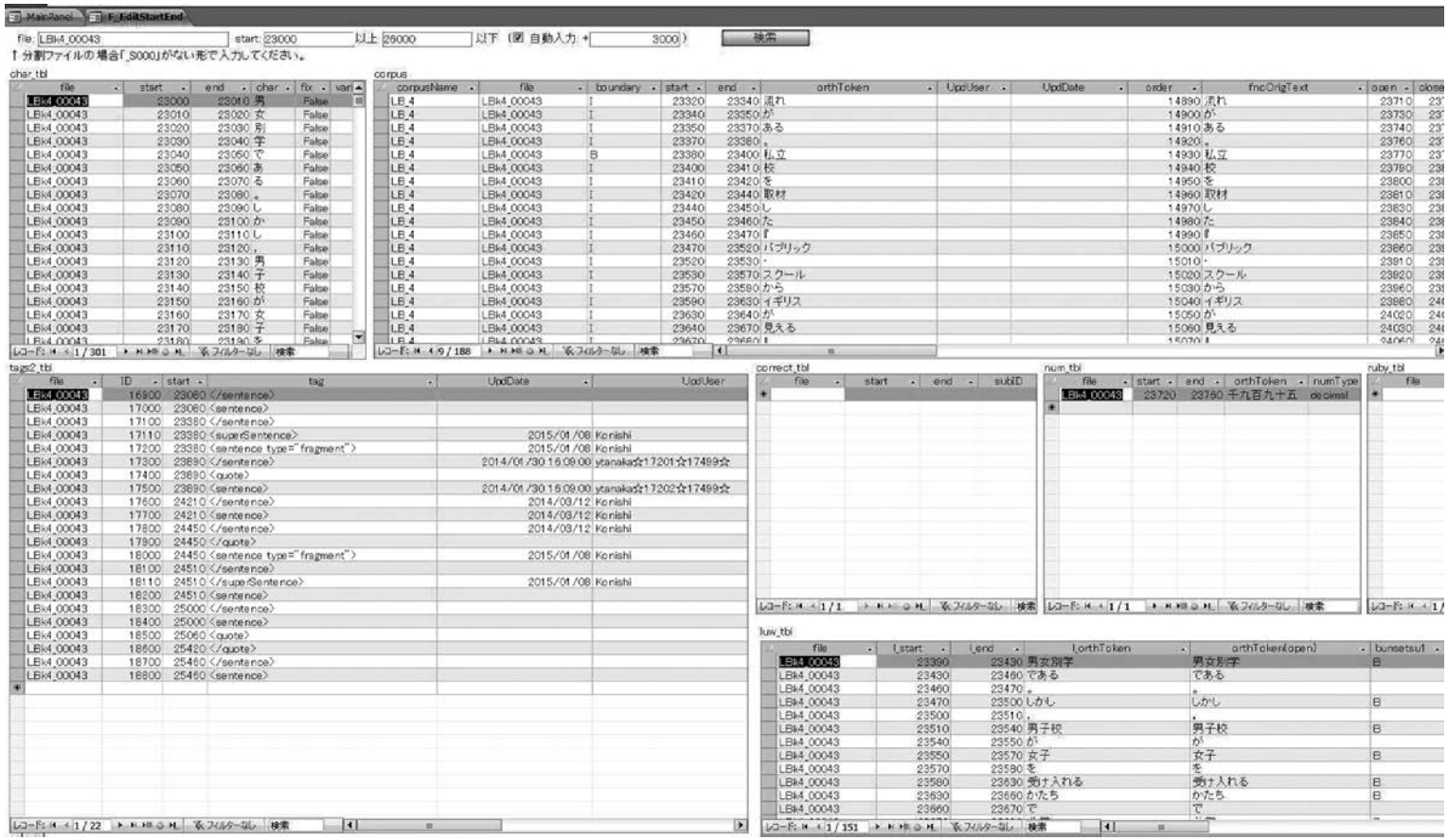

図 7　文書構造タグの直接修正作業

この作業に際し，XML ファイルの正則性を確認するために次の二段階の検証が行われた。

・Well-Formed（整形式）：XML ファイルでタグの交叉等がない状態
・Valid（妥当）：Well-Formed でさらに BCCWJ のタグ仕様を満たしている状態

二段階の検証のうち，Well-Formed（整形式）については修正を元データベースに反映する際にその都度確認が行われ，問題があれば修正は反映されずエラーが表示される。しかしながら，この時点では Valid（妥当）かどうかの検証は行っていない。Valid であるかどうかの検証は，Well-Formed であることが確認されたのちに実施した。Validation エラーが出たサンプルについては，その都度人手で直接修正を行った。この際，文書構造タグを見ながらの作業が必要になる（図 7）。BCCWJ 第 1.0 版には文書構造タグのエラーもあり，文境界側を直すか文書構造タグ側を直すかの判断を行いながら，場合によって原本画像の PDF データを参照しながら作業を行った。この直接修正作業は一時間あたりおおよそ 10 件から 20 件程度のペースで行われた。

4.2 修正件数

表 1 に修正件数を示す。バッチ処理は M(α) に相当する処理で約 16 万件に及ぶ。人手修正は M(β) で約 10 万件にも及ぶ。人手修正作業は，延べ 4 人からなり，人月に換算すると人手修正作業のみで約 30 人月になる。なお，3 節に示した基準に基づく文境界誤りは，一つの誤りが複数のパターンに適合する場合があり，パターンごとの集計は困難である。

表 1　文境界修正件数（2015/01/14 現在）

	タグ追加	タグ削除	タグ移動
バッチ処理	140	36,985	124,364
人手修正	48,089	53,879	408
合計	48,229	90,864	124,772

5. おわりに

BCCWJ 第 1.0 版の構築は，まずサンプリングを行い，次に文境界タグを含む文書構造タグ付与作業を行い，並行して形態論情報を付与するという工程で進められたため，文書構造タグ付与の段階では形態論情報を参照することができず，文字情報に基づく文境界認定しか行えなかった。さらに，そこで策定された文境界認定基準にも，2.3 節で示したような問題点が含まれていた。今回，形態論情報に基づく文境界認定基準を新たに作成し，人手で修正･確認を行うことにより，その問題点については大幅に改善したものと思われる。

今回提案した修正パターンは先行して行われた処理において発生した問題点に対処するためのものである。今後，「コーパスの設計時にどのような文境界基準を立てるべきか」という点について考察する必要がある。本稿では，文境界認定の一般的な指針として「文字情報」「形態論情報」「係り受け関係」という三つのレベルの基準を提示した。これらの基準は，コーパスに対してど

のレベルまでの情報を付与するのかによって，使い分けられるべきであると考える。さらに，今後の課題として，今回提案した基準や小西ほか（2014）の基準に基づいた文境界認定を自動的に行う解析器の開発があげられる。

参照文献

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.

福岡健太・松本裕治（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』1221–1224.

長谷川守寿（2014）「BCCWJ の文構造タグに関する一考察」『人文学報』488: 23–48.

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（2014）「『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」『第 5 回コーパス日本語学ワークショップ予稿集』127–136.

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.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48: 345–371.

丸山岳彦・高梨克也・内元清貴（2006）「第 5 章　節単位情報」『日本語話し言葉コーパスの構築法』（国立国語研究所報告 124），255–322.

南不二男（1974）『現代日本語の構造』東京：大修館書店.

西光雅弘・秋田祐哉・高梨克也・尾嶋憲治・河原達也（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): 544–552.

小木曽智信・中村壮範（2014）「『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システムの設計・実装・運用」『自然言語処理』21(2): 301–332.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版（上）（下）』，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. 文書管理番号 JC-D-10-05-01, 02.

下岡和也・南條浩輝・河原達也（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): 67–83.

下岡和也・内元清貴・河原達也・井佐原均（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): 3–18.

田島幸恵・難波英嗣・奥村学（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（FIT2003）』155–156.

田野村忠温「BCCWJ の資料特性」（2014）田野村忠温（編）『コーパスと日本語学』（講座日本語コーパス 6），119–151. 東京：朝倉書店.

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる（2011）『『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』，特定領域研究「日本語コーパス」平成 22 年度成果報告書. 文書管理番号 JC-D-10-04.

Correction of Sentence Boundaries in the Balanced Corpus of Contemporary Written Japanese DVD Version 1.0

KONISHI Hikari[a] NAKAMURA Takenori[b] TANAKA Yayoi[c]
MABUCHI Yoko[a] ASAHARA Masayuki[d] TACHIBANA Sachiko[e]
KATO Sachi[f] IMADA Mizuho[g] YAMAGUCHI Masaya[d]
MAEKAWA Kikuo[d] OGISO Toshinobu[d] YAMAZAKI Makoto[d]
MARUYAMA Takehiko[d]

[a]Adjunct Researcher, Center for Corpus Development, NINJAL
[b]Manpower Group Co., Ltd
[c]Adjunct Researcher, Department of Linguistic Theory and Structure, NINJAL
[d]Department of Corpus Studies / Center for Corpus Development, NINJAL
[e]Technical Staff, Center for Corpus Development, NINJAL [–2015.03]
[f]Postdoctoral Research Fellow, Center for Corpus Development, NINJAL
[g]Ministry of Education, Culture, Sports, Science, and Technology / Postdoctoral Research Fellow, Center for Corpus Development, NINJAL [–2014.03]

Abstract
In December 2011, the National Institute for Japanese Language and Linguistics (NINJAL) released a 100-million-word balanced corpus – the Balanced Corpus of Contemporary Written Japanese (BCCWJ) DVD Version 1.0 – which was compiled from 2006 through 2011. Some users have pointed out some issues concerning sentence delimitation in the BCCWJ. To address these issues, we – NINJAL – performed a complete survey and correction, beginning in 2013 and ending in 2014. This article reports the revision work on sentence delimitation in the BCCWJ. The problems with the BCCWJ DVD Version 1.0 derive from the string-based definition. We could not obtain any morpheme information for the sentence delimitation task because of the task parallelism between sentence delimitation annotation and morpheme annotation. The method used this time was morpheme based. We present the morpheme-based annotation guidelines, annotation environment, and basic statistics of the corpus correction.

Key words: BCCWJ, sentence boundary, annotation, error correction standard, error correction environment