ＢＣＣＷＪのタグ情報の修正について

長谷川守寿（首都大学東京）

Correction of the Tag Information on BCCWJ

Hasegawa Morihisa(Tokyo Metropolitan University)

１．目的

本発表では、「現代日本語書き言葉均衡コーパス DVD 版」（Contemporary Written Balanced Corpus of Japanese、以後 BCCWJ と略す）の C-XML ファイルに付与されている文タグの修正箇所の検討と、その作業から見えてくる BCCWJ の特徴について述べる。

BCCWJ を使うには、検索サイト「少納言」（www.kotonoha.gr.jp/shonagon/）または「中納言」（https://chunagon.ninjal.ac.jp/）にアクセスするか、BCCWJ-DVD 版を入手し検索ツールを各自が準備して使用するか、どちらかの方法がある。

筆者は日本語学習用のコロケーション情報を抽出するため、文単位のデータが必要となった。そこで BCCWJ-DVD 版を入手し文タグ（〈sentence〉と〈/sentence〉で一つのペア）を用いて、検索用プログラムを作成していたところ、文タグに問題があり、完全には文単位に分けられたデータが得られないことが判明した。最初に発見したのが(1)であり、文タグが示す一文中（〈sentence〉と〈/sentence〉に挟まれた間）に二つ以上の文が入っている。

(1)　〈sentence〉これを「いや、よしてしと読めば、いやがっているという意味である。はたしてそれだけだろうか。句読点をすこし動かして〈quote〉「いやよ、して」〈/quote〉と読めば、（略）。〈/sentence〉　（PB10_00030）

そこで、文タグを修正してから、BCCWJ のデータを研究対象として使用する場合、修正が必要な箇所はどれくらいあるのか、修正箇所の多寡はサブコーパスや媒体により違いはあるのか、そして目的にもよるが、これらのことから文タグを修正して使用するには、どのサブコーパスや媒体が適当かを明らかにする。

２．先行研究

BCCWJ の文書構造タグ（XML）については、『第５章　文書構造タグ』（国立国語研究所 2011）に詳述されており、文タグは階層構造に関するタグ（article、cluster、paragraph など）に含まれ、これ自体は「文に相当する文書要素」を示すのに使われ、実際には〈sentence〉と〈/sentence〉によって区切られている。

文タグの不備については、田野村（印刷中）で「出版と図書館のサブコーパスだけに限っても、少なくとも文の連続の数で約 3,000 か所、文の数で約 11,000 件の文への文タグの付与が漏れている」とし、その原因として「多くの場合、複数の段落にわたる引用、または、注番号の存在のいずれか」を挙げているが、「文タグを参照する処理を自前で行うのではない限り、BCCWJ の利用への影響はほとんどない」としている。本調査のきっかけは、まさに「文タグを参照する処理を自前で行う」ものであるが、田野村（印刷中）以外に文タグの問題点を指摘した調査は見つからないのが現状である。なお国立国語研究所（2011）では「ruby タグ」「correction タグ」等と使われているが、本稿では文タグと呼ぶこととする。

３．方法

本発表の目的は、文タグの追加が必要な箇所はどのくらいあるのか、サブコーパス・媒体によって修正箇所数や出現状況に違いはあるのかを明らかにすることである。これらの考察から、BCCWJ を修正して使う場合、どのサブコーパスが適切かを考え、さらに BCCWJ の特徴にも言及したい。本研究の対象と手続きは以下の通りである。

３．１　対象

本調査の目的は BCCWJ における文タグの調査であるため、出版サブコーパス、図書館サブコーパス、特定目的サブコーパス全てを扱う。またデータ形式は、ファイル数が多く、形態素情報のタグがついていない C-XML（Charactor-base XML）のデータを使用し、文単位を対象とするため可変長（Variable）を用い、ファイルに含まれる全ての文を対象とする。

なお、BCCWJ では文字コードの符号化方式に UTF-8 が採用されているため、以後の正規表現は UTF-8 に対応した表現を用いた。検索には Ruby（1.8.7）を用いて、スクリプトを作成した。総ファイル数は、172,675 ファイルである（BCCWJ ではファイル数をサンプル数と呼んでいるが、本発表では「ファイル数」と呼ぶ）。

３．２　手順

文タグは「文に相当する文書要素」をマークするタグであり、実際には〈sentence〉と〈/sentence〉でマークされ、(2)(3)(4)のように句点“。”・感嘆符“！”・疑問符“？”の後に“〈/sentence〉”が入力されているのが適切な箇所にタグが付されている例である。

文の終端には句点などいろいろな記号が位置するが、句点で終了する文だけに調査を限定し、感嘆符や疑問符、三点リーダー“…”などで終わる場合については紙幅の都合で省略する（結果は長谷川（印刷中）を参照）。

(2)　〈sentence〉もはやまっしぐら、一直線である。〈/sentence〉　（LBa0_00002）
(3)　〈sentence〉戦前とくらべたら、なんというちがいでしょう！〈/sentence〉　（LBa9_00026）
(4)　〈sentence〉何が大丈夫か？〈/sentence〉　（LBa5_00004）

次に文タグが欠如した例について述べる。(2)(3)(4)に対し、(5)の「…んです。僕…」の部分（場所を明示するために下線を施した。以下同）には、文タグが欠如していると考える。

(5)　〈sentence〉　「カッパというのは二十歳前からのあだ名で、なぜか人々はどんな時でもぼくのことを『カッパ』としか呼ばなくなったんです。僕はそう呼ばれても平気でした。〈/sentence〉　（PB12_00071）

タグが適切に入力されていない箇所を特定する方法に先んじて、修正箇所の数え方について説明しておきたい。本調査では、(5)のような場合、“〈/sentence〉〈sentence〉”を入力すればよいので、修正箇所は１と数える。

文タグが入力されていない箇所は、(5)の「。僕」のように、句点の後に２バイト文字“僕”が続く場合と、(6)のように、句点の後に１バイト文字“〈”が続く場合がある。１バイト文字が続く場合は、後続する文に関する何らかのタグが入力されている。例えば(6)は、句点の後に、後続する文の最初の漢字「真綿」の ruby タグ（以後ルビタグと呼ぶ）が続いてお

り、他に引用情報やサンプリングの開始情報など、いくつかのタグが後続する。

本稿では、〈sentence〉で始まり〈/sentence〉タグで終わっている用例については、表記を省略する。ただし〈sentence type = "quasi"〉のように特殊なものや他のタグは、残すこととする。また、段落の存在を示す行頭の２バイト空白文字は、見やすいように“□”で示す。

(6)　□「鉛色の空から、こまかな雪があとからあとから降ってくる。〈ruby rubyText="ま"〉真〈/ruby〉〈ruby rubyText="わた"〉綿〈/ruby〉をちぎったようなという形容があるけれど、この雪は真綿ではない。（略）　（PB12_00310）

本調査では詳細に調査するために二種類の検索方法を用いた。一つは、(7)のように文字の種類や記号そのものを特定して検索する方法である（以後「文字法」と呼ぶ）。文字の種類は、ひらがな・漢字・カタカナ・アラビア数字（以後、数字）・アルファベット・ギリシャ文字・キリル文字であり、記号については文字コードが連続していないため、個別に特定して“[★☆]”のように記述する（実際には“Å”から“＝”まで）。

(7)　。[ぁ-ん]|。[一-龠々]|。[ァ-ヴー]|。[０-９]|。[Ａ-Ｚ ａ-ｚ]|。[Ａ-Ω α-ο π-ω]|。[А-я ё]|。[Å‰♯♭♪†‡¶∠⊥⌒∂∇≡≒≪≫√∽∝∵∫∬∧∨￣⇒⇔∀∃∈∋⊆⊇⊂⊃∪∩,.・:;ˇ゜ ´ ` ¨ ^ ￣＿ヽヾゝゞ〃仝々〇ー－／＼～∥｜‥‘’“”（〔［｛〈《「『【＋－±×÷＝≠＜＞≦≧∞∴♂♀°′″℃￥＄¢£％＃＆＊＠§☆★○●◎◇◆□■△▲▽▼※〒→←↑↓＝]

もう一つの方法は、特定の文字列の連続ではなく(8)のような正規表現を用いた検索である（以後「否定法」と呼ぶ）。この方法では、“。"”“。〉”“。）”“。〕”“。］”“。〉”“。》”“。」”“。』”“。】”“。□”“。。”は文末としてあり得るので、修正箇所から除外され、ほぼ文字列法と同様の検索結果が得られるが、句点の後にルビ情報が付与された漢字から始まる文などが検出できるのが異なる点である。この方法では(9)のように“。」”を含む部分は文タグが不要なので検出されない。また(10)の“。。”の部分も検出しないが、「。活」の部分は文タグが欠如した部分として検出できるようになっている。

(8)　。[^"〉〕〉」）］』》　。]|。〈r|。〈q|。〈sa|。〈no

(9)　〈sentence type="quasi"〉□「説明をしにおうかがいしたいと思いますが、お名前と住所を教えてください。」〈/sentence〉　（PB11_00006）

(10)　〈sentence type="quasi"〉ココの素材を加工できるソフトを教えてください。。。。活字が打てるといいです〈/sentence〉　（OC02_02989）

４．結果

４．１　検索結果

二つの方法を用いて修正箇所を検出した結果について、説明する。

文字法での検索では、句点にひらがな・漢字・カタカナ等が後続した場所にタグが抜けている例が検出され、(11)では「。また」「。我」「。毎」「。『」のような例が見える。

(11) 〈sentence type="quasi"〉□「―千九百十年は我々の最も得意の時代であつた。『パンの會』は毎週ひらかれた。我々はロダンの銅像の首の唇に寄せた皺の〈ruby rubyText="ねば"〉粘〈/ruby〉こさが何ういふ情を〈ruby rubyText="か"〉藏〈/ruby〉くしてゐるかゞ分るほどになつた。また〈ruby rubyText="（アラビア）"〉亜剌比亜〈/ruby〉物語や近松・三馬などに出てくる青年の心に同情を寄する程の苦勞も覚えた頃である。毎日同じ仲間と交遊して作詩し、作劇して日を暮した。…〈/sentence〉　（PB29_00042）

データは EOS まで読み込まれるが、一読み込み単位で修正箇所数が最大となったのは、48 箇所の修正が必要となる OB3X_00110 である。なお OB3X_00110 にはルビタグが多くつき、多少見にくいので、その次に 45 箇所と多かった PB15_00180 の一部を示す。

(12) □「それもあるかもしれない。それかうまくなるかですね。ゲームの中のハードであるクルマのチューニングも重要だけど、ホントのハードウエアのチューニングは大事です。(略) でも、生産技術は０。　（PB15_00180）

一方、修正すべきでない点を修正箇所として検出した例がある。以下はギリシャ文字を指定し検索した例であるが、(13)はタグが抜けている部分として正しく検出できるが、(14)ではギリシャ文字が顔文字の一部に使用されており、文タグがなくても問題ないため、検出したことは誤りである。文字種により検索精度に違いがあるので、後述する。

(13) Ｘ線のエネルギーは振動数Ｖ＝ｃ／λ）のｈ倍、つまりｈνですから、前方に散乱されるときは、エネルギーは変わりません。θが大きくなると、（略）（PB24_00213）
(14) 〈sentence type="quasi"〉□ピカチュウしか知らなくて（；σ。σ）ゞ　ゴメンネェ．．．〈/sentence〉　（OY14_49261）

４．２　異なる検索方法による違いについて

文字法と否定法を用い修正箇所を検出した結果をファイル単位で見たのが表１である。

表1　修正が必要なファイル数

検索方法	要修正	必要なし	合計
文字法	3,300(1.9%)	169,375(98.1%)	172,675(100%)
否定法	3,738(2.2%)	168,937(97.8%)	172,675(100%)

文字法の検索では、少なくとも一カ所以上の修正点をもつファイルは 3,300 あり、中でも(15)は最も多い 96 カ所の修正点が含まれていた（一部を示す）。

(15) □「多くの人がその過程に吸収されました。一九三〇年ごろ、インドは爆発的な成長をとげました。インドは四七年ごろに弾みをつけ、その後に大きく成長しましたが、今は鈍化しつづけています。一九六二年に私は大学に残るか仕事かを考えあぐね、専門家としての道、研究機関を選びました。今日、人はいっそう奮闘せねばなりませんが、これにはいい面もあるのです。　（PB29_00606）

媒体別に修正箇所のあるファイル数を調べた結果が表２であり、修正箇所数とその差について調べた結果が表３である。媒体の順序は国立国語研究所（2011,p15）に倣った。

表２　媒体別の修正ファイル数

媒体	文字法	否定法
PB	775	825
PM	102	109
PN	5	5
LB	289	306
OW	49	50
OT	22	22
OP	44	45
OB	112	114
OC	821	980
OY	1,059	1,258
OV	15	16
OL	5	5
OM	2	3
合計	3,300	3,738

表３　媒体別の修正箇所数とその差

媒体	文字法(A)	否定法(B)	(B)−(A)
PB	3,938	4,254	316
PM	434	457	23
PN	12	12	0
LB	1,004	1,053	49
OW	103	105	2
OT	86	86	0
OP	98	103	5
OB	670	737	67
OC	1,421	1,589	168
OY	2,111	2,373	262
OV	36	37	1
OL	8	8	0
OM	19	20	1
合計	9,940	10,834	894

文字法と否定法の修正箇所数を比べると、PB（出版-書籍）、OC（Yahoo!知恵袋）と OY（Yahoo!ブログ）のように 100 以上異なるものと、PM（出版-雑誌）、LB（図書館-書籍）のように差が小さいものに分かれる。否定法でのみ検出されたものを確認したところ、PB（出版-書籍）PM（出版-雑誌）PN（出版-新聞）LB（図書館-書籍）OB（ベストセラー）では、ルビタグを含むもの、引用タグ（quote）を含むもの(16)、サンプリングの開始位置を示すもの(17)、傍注を示すもの（noteBodyInline）が出現しており、そのため数値が異なった。

(16)　（略）原語の音とともにひっくるめて借用する場合が〈quote〉「音借用」〈/quote〉である。〈quote〉「借用語」〈/quote〉とか〈quote〉「外来語」〈/quote〉と呼ばれているのは〈quote〉「音借用」〈/quote〉語のことである。　（PB28_00049）

(17)　□「父の少年時代の写真を見ると、変な気持ちになります。〈sampling type="start" /〉私の孫の年齢ですから。五十年の恨のたった一つだけ、（略）　（LBi9_00146）

しかし OC・OY では、句点に“？”や“！”が続く例を検出している。これらは検索条件作成時には文の終端としては想定しておらず、否定法ではこういった用例が検出され、文が後続していない部分を誤って検出したことになる。

(18)　貼り付け方を教えてください。？　（OC02_07366）
(19)　あなたは、リカバリーＣＤを大事に保管していますか。！　（OC02_07489）

４．３　検出結果の精度について

文字種により検出されたものの精度に違いが見られたので、文字種別に検出し、その中から無作為抽出した 100 カ所（用例数が 100 以下の場合は全て）について確認する。

表4　後続文字列別の修正箇所数と検討数、その中で実際に文が後続した数

	ひらがな	漢字	カタカナ	数字	アルファベット	ギリシャ文字	キリル文字	記号	合計
箇所数	3758	3757	647	120	147	6	4	1501	9940
検討数	100	100	100	100	100	6	4	100	610
文後続	99	100	99	100	85	1	0	27	526

句点に後続する２バイト文字の文字種から、修正箇所を再集計し精度を確認したのが表４である。なお句点に後続する１バイト文字（この場合は“〈”）は除外した。以後文字種毎に検討するが、ギリシャ文字は後続する全例（６例）を確認したところ(14)のように５カ所で顔文字の一部として使われ、文が後続したのは(13)のみであったため、小節は設けない。

４．３．１　ひらがな

ひらがなが後続する例は 3,758 カ所に見られたので、無作為に抽出した 100 カ所を検討した結果、99 カ所でその後に文が続くことが確認された。文が続かないと思われる例は(20)の１例のみであり、誤入力と思われる。

(20)　〈sentence type="quasi"〉□「そう言うと兵をひきいて城外へ突撃した。張巡の兵は勇戦して賊将十四人をとらえ、八百余の首級をあげた。ん〈/sentence〉　（PB49_00170）

４．３．２　漢字

漢字が後続する例は 3,757 カ所で見られたので、100 カ所検討した。括弧“（）”に囲まれた注釈の部分は文に含まれるのか(21)、中国語の文型の部分は文に含めるのか(22)など、文の定義について再考する必要が出てくるが、この部分も文とすると、検出できている。

(21)　“地獄の天使”（オートバイの暴走族。元来はカリフォルニアの暴走族）に補助輪が必要なようにね。　（PB59_00243）
(22)　＊～且―。安〈image description="二重線のダッシュ" /〉。（抑揚）　～でさえ―である。　（OT03_00030）

４．３．３　カタカナ

647 カ所で見られたので、100 カ所を検討した。これらの文の type は「quasi」と「verse」

なので、「カラー：ＢＫ（黒）」「ヤンレ　エエ」を文の一種と考えれば、誤検出はカタカナが顔文字の一部として使われている(25)のみで、それ以外は正しく検出できている。

(23)　〈sentence type="quasi"〉ローイング、プル系トレーニングに。ナイロンパッド付きなので　手首を保護します。カラー: ＢＫ (黒) 〈/sentence〉　(OY07_00073)
(24)　〈sentence type="verse"〉お伝たちまち縄目にかかる。ヤンレ　エエ〈verseLine /〉〈/sentence〉　(LBl2_00041)
(25)　廃人ＯＸです。へ(゜∀゜へ)アヒャ！　(OY03_09472)

４．３．４　数字

120 カ所見られたので、100 カ所を検討した。　“「」”“【】”のような括弧に含まれる注釈(26)(27)や、型番(28)を文と考えれば、抽出できている。

(26)　□発端は、米国最大の商戦期“ブラックフライデー”（感謝祭翌日の金曜日。２００８年は１１月２８日）だった。　(OY14_38930)
(27)　産業振興センター（☎３６０－３１９６で。【有料講座。４時間～、５，１６５円】▸パソコン入門（略）　(OP75_00001)
(28)　〈sentence type="quasi"〉お手入れも楽々。１２３６３２　ＡＴＬＩＵＭ／アトリウムランチョンマット　アイボリー〈/sentence〉　(OY04_01548)

４．３．５　アルファベット

147 カ所で見られたので、100 カ所を検討した結果、85 カ所でその後に文が続いていることが確認された。それ以外ではアルファベットが顔文字やアスキーアートの一部に使用されている。国立国語研究所(2011、p.79)には「サンプル作成時に削除された，いわゆる「アスキーアート」」とあるが、一行単位のアスキーアートは削除されていないようである。

(29)　〈sentence type="quasi"〉（略）外さなきゃならないんですｐ（´⌒`。Ｑ）グスン〈/sentence〉　(OY14_31617)
(30)　〈sentence type="quasi"〉□（Ｔ。Ｔ）〈/sentence〉　(OY15_10298)
(31)　。ｏ○☆＊゜¨゜゜・＊：．．　(OC14_04623)
(32)　初登場第３位キタ――（゜∀゜）⌒Ｙ⌒（。Ａ。）⌒Ｙ⌒（゜∀゜）⌒Ｙ⌒（。Ａ。）⌒Ｙ⌒（゜∀゜）――！！　(OY15_00312)

４．３．６　キリル文字

４カ所で見られ、全て確認したところ、キリル文字の後に文が続く例はなく、全て誤検索である。なお、(34)は一行のデータがこのままで、対応する〈sentence〉がない。

(33)　〈sentence type="quasi"〉⊂⌒￣⊃。Д。）⊃．．．．．．．．．．．．．。〈/sentence〉　(OC14_04563)
(34)　）(。ë ë)。〈/sentence〉　(OY02_00286)

４．３．７　記号

記号は(7)の“Å”から“＝”までの部分を指すとする。1,501 カ所で見られたので、100 カ所を検討した。正しく検出したのは 27 カ所で、(35)のように文タグの欠落した場所を検索できることもあるが、(36)のように流れ星のアスキーアートの一部になっている部分も検出してしまう。このような場合、誤りを含むものが多く、また句点が文ではないものの一部で使われていることが多いことが分かる。

(35) □「民事裁判では、訴える人を『原告』、訴えられた人を『被告』と呼ぶことになっています。『被告』イコール犯罪者という意味では、（略）　（PB13_00114）
(36)　。☆．．．　（OY11_06824）

以上から文字種により検索精度が異なり、ひらがな・漢字・カタカナ・数字などが句点に続く場合は文が続くと判断して問題ないが、記号が続く場合は検出例が多く、しかも高い確率で顔文字などのアスキーアートの一部を構成していることがわかった。

そこで、ひらがな・漢字・カタカナ・数字・記号の修正箇所数の合計が 500 以上の媒体を対象に、句点に後続する文字種を調べたのが表５である。ひらがな・漢字・カタカナ・数字を、正しく修正箇所が検出できる指標、記号を修正箇所の誤検出の指標とすると、PB・LB・OB は、ひらがな・漢字・カタカナ・数字の合計の割合が多く正しく検出できているといえる。それに対し OC と OY は記号の割合が多く誤検出の可能性が高いと推測される。

なお、句点に記号が後続する場合、その句点の前が文の終端でない可能性も考えられる。それらを確認するために、「句点＋記号」が前接する文字種について調査を行う。

表５．主要媒体別の句点に後続する文字の文字種とその割合

	ひらがな	漢字	カタカナ	数字	記号	合計
PB	1821	1658	273	12	94	3858
	47.2%	43.0%	7.1%	0.3%	2.4%	100.0%
LB	482	410	63	38	4	997
	48.3%	41.1%	6.3%	3.8%	0.4%	100.0%
OB	350	259	37	1	20	667
	52.5%	38.8%	5.5%	0.1%	3.0%	100.0%
OC	347	432	75	18	514	1386
	25.0%	31.2%	5.4%	1.3%	37.1%	100.0%
OY	469	638	143	32	779	2061
	22.8%	31.0%	6.9%	1.6%	37.8%	100.0%

４．３．８　句点＋記号に前接する文字種

句点＋記号に前接する文（文字列）が何によって終わっているか、文字種別にまとめたのが表６である。例えば(35)では、句点＋記号である“。『”には“す”というひらがなが前接していると考える。表６の記号 A は、(7)の“Å”から“＝”までの記号に、“。”“）”

“…”などを追加したもので、これらは文末を構成することがあるため追加した。

表６　句点＋記号に前接する文字の内訳

	ひらがな	漢字	カタカナ	数字	アルファベット	記号A	それ以外	合計
箇所数	474	42	16	2	6	726	235	1501

表６より、句点＋記号に前接する文字の半数近くが記号 A に含まれていることが分かる。これらは(37)(38)(39)のように顔文字のようなアスキーアートを構成していることが多く、文の終端とはなっていないのである。

(37)　〈sentence type="quasi"〉（＾。＾）～〈/sentence〉　（OC01_00522）
(38)　〈sentence type="quasi"〉（。・m・）クスクス〈/sentence〉　（OC01_00542）
(39)　（。-__-。）。　（OC01_01000）

表６の「それ以外」は(40)(41)のような例で、句点に前接するのは〈sentence〉というタグのみで、文字列がないものである。そもそもこのような文字が文頭に位置し、文のタグが付されること自体がおかしいということも指摘しておきたい。

(40)　〈sentence〉。寝過ぎですよね．．．〈/sentence〉　（OY14_24922）
(41)　〈sentence〉。．〈/sentence〉　（OY03_03891 他多数）

ひらがなのように、句点＋記号が後続する例も３割強あるが、半数近くを占めるのが記号 A の例である。そこで記号 A（726 例）が、どの媒体に多いのか調べた結果（表７）、記号＋句点＋記号が続く例は、圧倒的に OY に多く、8 割強であり、OC が 2 割弱である。PB の例は、(42)のようにかなり変わった例で、実際の紙面では１バイト文字［。-゜］であろう。

表7　媒体別の記号＋句点＋記号の出現数

媒体	PB	OT	OC	OY	合計
箇所数	4	19	133	570	726
割合	0.6%	2.6%	18.3%	78.5%	100.0%

(42)　１字以上の半角カタカナをワイルドカードで検索する場合は半角文字で“［ヲ-゜］｛１，｝”（半角カタカタの記号も含める場合は“［。-゜］｛１，｝”のようにします.（略）　（PB35_00023）

このように記号を多く含む OY や OC は、顔文字やアスキーアートであることが多いため修正対象にはなりにくいであろう。修正するならば、句点に文が後続する可能性の高い PB・LB・OB を対象にするのがよいであろう。

５．結論

以上のように、修正箇所の検出とサブコーパス・媒体による修正箇所数と出現状況を見

てきた。不必要な文タグを削除したり、タグのない部分につけたりする対象としては、出版サブコーパス（PB・PM・PN）や図書館サブコーパス（LB）が適していると思われる。PM・PN は修正箇所数が少なく、PB・LB は、修正箇所数は多いが真に修正が必要な箇所である可能性が高く、文タグの追加などの修正後、データとして用いることが可能である。逆に、特定目的サブコーパスは、OM・OL・OV のように修正箇所数が少ないものも多いが、OY や OC は修正箇所数が多い割には、その場所が文の終端ではない可能性も高く、さらに修正して文として一様に扱うためには、タグの追加だけではなく、タグの削除やデータ自体の修正が必要になり、困難が予想される。また、OY では、文の終端が様々で、文末を探すのが難しいという問題があり、例えば(43)(44)では“♡”や“♥”、“♫”“♫”などが文の終端に位置し、他に“w笑”なども見られた。

(43)　〈sentence type="quasi"〉傑作ポチを戴けるととても嬉しく思います♡　御協力に感謝申し上げます（o＊。__。）oペコッ〈/sentence〉　（OY14_28649）

(44)　〈sentence type="quasi"〉σ（・・；　早く元気になってねぇ〜♥　（略）　「カシャカシャ」って、ケーキなんかも作りたい♫　デジカメ持って散歩もしたい♫　私、欲張りか？？？（略）〈/sentence〉　（OY14_35446）

さらに、OY（Yahoo!ブログ）には文自体が切れているサンプルがあり、(45)は元々のブログが検索したサイトの一部を貼り付けたような形式で、途中で文が切れていて、文単位で取り出すこと自体無理なデータも含まれている。

(45)　（略）ホースです。大事に長く乗りたい方には必需品です。■■仕様変更によりグレードアッ...　（OY14_01602）

BCCWJ は出版サブコーパス、図書館サブコーパスそれぞれが生産実態、流通実態を反映するために作成されている。特定目的サブコーパスは、上記二つのサブコーパスでは「十分な分量が集まりにくい資料を中心に収録」（国立国語研究所(2011,p.16)）されているため、DVD 版を使用する場合、用例数を増やすなど安易な目的で三つのサブコーパスを同様に扱ってはならず、研究目的に合わせ、対象とするサブコーパスを慎重に選び、さらに修正を加えていくことが重要となってくる。

参考文献

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」、BCCWJ-DVD 版収録

田野村忠温（印刷中）「BCCWJ の資料的特性――コーパス理解の重要性――」『講座日本語コーパス６　コーパスと日本語学』、朝倉書店、
（http://www.tanomura.com/temporary/bccwj_tanomura_2.pdf、2013 年 2 月 24 日取得）

長谷川守寿（印刷中）「BCCWJ の文構造タグに関する一考察」、『人文学報』第 488 号、首都大学東京