首都大学東京　人文学報第488号抜刷

BCCWJの文構造タグに関する一考察

長谷川　守寿

首都大学東京人文科学研究科
2014. 3

BCCWJ の文構造タグに関する一考察

長谷川守寿

1. 目的

本稿は、「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書構造タグ、特に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 は、2011 年に国立国語研究所より公開された 1 億語規模のコーパスである。BCCWJ を使うには、検索サイト「少納言」[*1]または「中納言」[*2]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、詳細は後述)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。まず最初に発見したのは、以下の例文である。入力の誤りがあるため(“「いや、よしてし”の部分は“「いや、よして」”が正しいと思われる)理解しづらいが、明らかに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(ママ)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)[*3]

*1 http://www.kotonoha.gr.jp/shonagon/
*2 https://chunagon.ninjal.ac.jp/
*3 以後、BCCWJ に含まれるサンプル ID を示す。

そこで、本稿では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[*4]、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[*5]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。また、どのような理由からタグが欠けているのか原因を考察しながら、修正に関する問題点に言及する。

本調査の結果が、BCCWJ-DVD 版の文タグを用いたいというユーザーに人手での修正が質的・量的に可能なのか、判断材料を与えられれば幸いである。

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)では「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

---

*4 修正しないという対処の仕方もある。例えば、形態素解析器 MeCab の「改行処理」を用いることにより、問題が解消できる部分もあるが、以下のような例文(a)の場合、(b)(c)のように区切られることとなり、問題は全面的には解決しない。

(a)　私は彼に「それは間違っている。こっちが正しい」と何度も言った。(筆者作例)

(b)　私は彼に「それは間違っている。

(c)　こっちが正しい」と何度も言った。

*5 国立国語研究所(2011)では、BCCWJ を構成する三つのデータを“サブコーパス”と呼んでいる。また、“媒体”は国立国語研究所(2011)の 2 章のみに使われ、第 1 章では“レジスター”、第 3 章第 8 章では“メディア”と呼ばれている。本稿は表 2-1 を参照するため、“媒体”という用語を用いる。

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)以外に文タグに関する問題点を指摘した調査・研究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、本稿では田野村（印刷中）にならい、文タグと呼ぶこととする。

3. 方法

本稿の目的は、以下の３点を明らかにすることである。

1．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。
2．文末として認定される記号には“！”や“？”などがあるが、それらの後に文タグを追加する必要がある箇所はどのくらいあるのか。
3．文タグが欠ける原因は何が考えられるか。

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

BCCWJ は、サンプリングの対象という面から、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に分かれる。また、データ形式の面から BCCWJ-DVD 版には、CORE、M-XML、C-XML という三種類のデータが収録されている。CORE データはファイル数は少ないが、形態素解析デー

タに人手による修正を加え、精度を高めたデータである。M-XML(Morphology-base XML)は、形態論情報が付与された形式であり、C-XML(Charactor-base XML)は、文書構造タグが付与されたもので、さらにデータの長さから固定長(Fixed)と可変長（Variable）の二種に分かれる。

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のデータを使用する。さらに、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。表1はサブコーパス別のファイル数である（BCCWJでは、ファイル数をサンプル数と呼んでいるが、本稿では「ファイル数」という用語を使う）。

表1　サブコーパスの詳細

サブコーパス	媒体 括弧内は略称	ファイル (個)	語数 (万語)
出版サブコーパス	書籍(PB)	10,117	2,855
	雑誌(PM)	1,996	444
	新聞(PN)	1,473	137
図書館サブコーパス	書籍(LB)	10,551	3,038
特定目的サブコーパス	白書(OW)	1,500	488
	教科書(OT)	412	93
	広報誌(OP)	354	376
	ベストセラー(OB)	1,390	374
	Yahoo!知恵袋(OC)	91,445	1,026
	Yahoo!ブログ(OY)	52,680	1,019
	韻文(OV)	252	25
	法律(OL)	346	108
	国会議事録(OM)	159	510
合計		172,675	10,493

（国立国語研究所(2011, p. 15, 表2-1)より一部改）

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、まずは、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後述する。

(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。問題となる箇所の数え方は、田野村(印刷中)と異なる。田野村（印刷中）では、(5)の場合、連続した2件の文が文タグを欠いていると数えるが、本稿では目的が修正点であるため、(5)は1カ所であると数える。

では、文タグが入力されていない箇所を探す方法について述べる。当該のタグが抜けている部分には、(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“く”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”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、以後<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については、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)。例えば句点の後にひらがなが続く検索表現は“。[ぁ-ん]”のように記述し、句点に漢字が続く検索表現は“。[一-龠々]”のように記述する。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-Ｚａ-ｚ]|。[Α-Ωα-οπ-ω]|。[А-яё]|。[Å‰♯♭♪†‡¶∠⊥⌒∂∇≡≒≪≫√∽

∝∵∫∬∧∨¬⇒⇔∀∃∈∋⊆⊇⊂⊃∪∩，．・：；ˇ ° ´ ` ¨ ^
￣＿ヽゞゝゞ〃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、「句点の後に"〈）〕］〉》」』】”□。以外の文字が続く文字列を探せ」という命令である（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、結果的にはほぼ文字列法と同様の検索結果が得られるが、前述の通り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大きな違いである（4.2.で詳述する）。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。

文字種を指定した文字法での検索では、ひらがな・漢字・カタカナ等が句点に後続した場所にタグが抜けている例が検出された（なお、ローマ数字・

丸数字①等が文の先頭となり、句点に後続している例は検出されなかった）。例えば(11)では、句点の後にひらがなが続いていて、文タグが挿入されていない「。また」や、句点の後に漢字が続く「。我」「。毎」、記号が後続する「。『」のような例が確認できる。

(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（エディタなどでは改行マークで示される部分）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（ルビタグを含む場合も数えているため、田野村（印刷中）とは異なる）。なおこの 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。この話にはもうひとつオチがある。ゲームってのは中にクルマがあろうがなかろうが、所詮はソフトウエア。（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てしまった例がある。例えば、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ているが、(14)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、このような場所を検出したことは誤りである。詳細に見ていくと文字種により検索精度に違いがあるので、この問題については後述する。

(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、散乱X線のエネルギは減ってきます。（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て修正箇所を検出した結果を、ファイル単位で見ると表２のようになった。

表2　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)

なお、本調査では、C-XML 内のファイルを対象としたが、M-XML 内のファイルでも、形態論情報を削除した後で同様の調査を行った結果、同様の結果が得られることを確認した。

次に媒体別に、修正箇所のあるファイル数を調べた結果が表３であり、修正箇所数とその差について調べた結果が表４である。媒体の順序は表１に倣った。

表３　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４　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たとき、PB（出版-書籍）・OC（Yahoo!知恵袋）・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（最大 67、最小 0）に分かれる。否定法でのみ検出されたものを確認してみたところ、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、前述のルビタグを含むもの(6)、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、これらを検出したため、数値が異なっていることが分かった。

(16)　（略）これに対して、自国語に翻訳する手間を省いて意味内容を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、それで解いたと思います。　(LBi9_00146)

これに対し、OC（Yahoo!知恵袋）・OY（Yahoo!ブログ）では、句点に“？”や“！”が続く例を検出してしまっている。これらは、プログラム作成時には、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大量に検出され、文が後続していない部分を誤検出してしまった例が多い。

(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

次に、検出結果の精度について検討する。上述のように、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）について確認した。

表5　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の観点から、修正箇所を再集計し、精度を確認したのが表５である。なお句点に後続する１バイト文字列（この場合は“く”）は除外してある。以後、文字種毎に結果を見ていくが、ギリシャ文字については、後続する例は６カ所だけで、これらを全て確認したところ(14)のように５カ所で顔文字の一部として使われ、文が後続していたのは前出の(13)のみであったため、小節を設けて詳述することはしない。

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。その結果 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例のみであった。筆者の手元にはデータ採取の対象となった書籍がないため確認ができないが、他に同様の部分がないことから、これは誤入力ではないかと思われる。

(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、無作為に抽出した100カ所を検討した。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、仮にこの部分も文とすると、問題なく検出できている。

(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

カタカナが後続する例は647カ所で見られたので、無作為に抽出した100カ所を検討した。これらの文のtypeは「quasi」と「verse」なので、(23)や(24)のように「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

数字が後続する例は120カ所見られたので、無作為に抽出した100カ所を

検討した。漢字同様に、“「」”“【】”のような括弧に含まれる注釈(26)(27)、さらには型番のようなもの(28)を文と考えれば、問題なく抽出できている。

(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

アルファベットが後続する例は147カ所で見られたので、無作為に抽出した100カ所を検討した。その結果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。なお、ここでは、「アスキーアート」を、「コンピューター上で、等幅フォントの文字（狭義にはアスキーコード）を組み合わせて描くイラストレーション。電子メールの署名・広告や、掲示板の発言などで用いられる。単に絵文字とも」（『スーパー大辞林』より、一部改）と考える。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）後ろの扉のロックを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

キリル文字が後続する例は4カ所で見られ、全て確認した。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>はもともと含まれていない。

(33) <sentence type="quasi">⊂(￣ヨ。Д。)ヨ. . . . . 。
</sentence> (OC14_04563)

(34)) (。ёё)。</sentence> (OY02_00286)

4.3.7 記号

記号は(7)の“A”から“=”までの部分を指すとする。記号が後続する例は1,501カ所で見られたので、無作為に抽出した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、文タグを追加する必要がないのに検出してしまう。このように、記号が後続する場合誤りを含むものが多く、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、全くありませんので、ご了解ください。 (PB13_00114)

(36) 。☆. . . (OY11_06824)

以上のように、検出された部分の検証結果から、文字種により検索精度が異なることが分かった。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、しかもかなりの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以上の媒体について、句点に後続する文字種の内訳を調べたのが表6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してみると、PB（出版-書籍）・LB（図書館-書籍）・OB（ベストセラー）は、ひらがな・漢字・カタカナ・数字の合計数の割合が多く、正しく検出できているといえる。それに対しOC（Yahoo!知恵袋）・OY（Yahoo!ブログ）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（以後“＋”を文字種が連続する意味で使用する）が前接する文字種について調査を行う。

表6 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7である。例えば前出の(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7（次ページ）の記号Aは、前出の(7)の“Ａ”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７　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７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７の「それ以外」というのは(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査を行った結果が表８である。その結果から、記号＋句点＋記号が続く例は、圧倒的に OY（Yahoo!ブログ）が多く、8 割強であり、OC（Yahoo!

知恵袋）が2割弱である。PB（出版-書籍）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[。-゜]となっていたものであろう。

表8　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（Yahoo!ブログ）・OC（Yahoo!知恵袋）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（出版-書籍）・LB（図書館-書籍）・OB（ベストセラー）を対象にするのがよいであろう。

4.4. それ以外の文末記号

通常、句点以外にも“！”や“？”“…”で終わる文が見られる。句点と同様に、“！”や“？”に文が後続していて、文タグの追加・修正が必要な箇所はないだろうか。そこでそのような修正箇所数がどのくらいあるのか、調べてみた。なお、句点はそれ単体だけで文を区切るものとなっているが、「！」「？」「…」は、単体だけでは(43)や(44)のようにタグの一部になっているものもある。そこで、「！□」「？□」という形で「記号＋2バイト空白文字」のみ検索した。

(43)　<sentence type="quasi"><image description="！" no="2" />　ポイント</sentence>　(LBi8_00003)

(44) <image description="横置き辞書と？ボタン" no="53" />をクリックすると、ＩＭＥのヘルプが表示され、使い方を調べられます。(PB15_00074)

その結果、(45)(46)のように“！□”や“？□”の後に文が続く例が見られ、出現数をまとめたのが、表９である。なお、(47)のように、果たして「ポー！」自体が文なのかなど、なかなか難しい問題も出現した。

表9　句点以外の修正箇所数

	！□ (！＋2バイト空白文字)	？□ (？＋2バイト空白文字)	…□ (…＋2バイト空白文字)
出現数	186	117	820

(45) □「宮城県北上町、海と川が出会う、人口四〇〇〇人の河口の町にも、もうひとつのスローフードがある。（略）一年間自家生産している食材にはどんなものがありますか？　いつ頃種をまき、いつ頃収穫しますか？　さらに、それらの食材はどのように調理料理、加工保有をしていますか？　(PB36_00128)

(46) <sentence type="quasi">□３兄妹めちゃいいわぁ～（＾＾ゞ　あのＴシャツ、マジでガチで欲しいですｗ笑　にしても可愛い♥　功一、兄妹の真ん中だし♥　クドカンいるし！豪華すぎるよ！！　Ｔｒｕｔｈ流れた時のにのちゃん、照れちゃって可愛い♥あの問題は簡単だわよ！！！！！（略）　(OY14_21962)

(47) <sentence type="verse">ポー！□ポー！□ポー！□ポー！□ポー！<verseLine /></sentence>　(OV2X_00090)

結果として検出された数値は、句点の修正箇所数（文字法で 9,940 箇所）と比べるとあまり多いとは言えず、また媒体による調査や検証も行っていない。さらにこの方法では(46)の「クドカンいるし！」のように“！”で文が

終わっているが検出されていないものがあることが予想されるため、不十分な点が残るとも言える。そのため、文の修正箇所数として無視していい数値かどうかの判断は保留するが、BCCWJ における文末表記の多様性の一端が垣間見られる結果となった。

なお、“！”“？”“…”には、“</sentence><sentence>”というタグを追加する修正の他に、別の種類の修正が必要となる場合があることが分かった。(48)は“！”が文末を示すものとしてタグ付けされているが、引用文の中の文も同様に<sentence>タグを使用しているため、文が「納豆！」で終わる形になってしまっている。しかし、実際には何らかのタグ（ここではxとする）を使って、(49)のように示せることが望ましい。同様の問題が“？”“…”でも見られた。

(48) <sentence>その頃一人の寡婦のために、毎朝一緒に<quote>「<sentence>納豆！</sentence> (LBa1_00004)

(49) <sentence>その頃一人の寡婦のために、毎朝一緒に<quote>「<x>納豆！</x><x>納豆！</x>」</quote>と言って売り歩いてやったという逸話がある。</sentence> (LBa1_00004 改)

さらに“？”には、文の終端に位置するだけでなく、クエスチョンマーク(50)としての働きもあり、(50)は“？”で文として区切られているので、(51)のように修正されることが望ましい。

(50) <sentence> 甲府での初期微動は、この文章では数秒とよみとれ、また別の報告では、？</sentence> (LBa4_00014)

(51) <sentence> 甲府での初期微動は、この文章では数秒とよみとれ、また別の報告では、？を付して三秒とされているが、これは初期微動ではなさそうである。</sentence> (LBa4_00014)

これは、どのような時に“！”“？”“…”を文末と認めるかという問題だけではなく、文タグ自体の付与の問題でもある。括弧内の文も同じように文タグを使うため、引用文という文の存在を示すタグによって文本来の構造が崩れ、対応するタグが次の行（またはそれ以降）に現れ、対応がわかりにくくなってしまっている。例えば(48)の場合、文頭の<sentence>タグに対応する</sentence>が現れるのは2行後である。この問題については根本的な修正が必要となるため別の機会に述べたい。

4.5. 文構造タグが抜ける原因

田野村（印刷中）では「文タグの欠落には、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が関わっていることが判明した」と述べている。引用を“「”“【”“［”“『”で始まるものと考えれば、本稿では(5)(11)(12)(15)(17)(20)(35)が引用の一部に該当し、妥当な原因だと考えられる。句点ではないが、“？”に文が後続し修正が必要であるということでは(45)も引用が原因だと考えられる。さらに丸括弧“()”も引用とするならば、田野村（印刷中）で説明可能な対象はさらに広がる。しかし(23)のように段落の先頭であることを示す2バイト空白文字“□”がなく、引用とも考えられない場合も見られ、他にも原因があるのではないかと思われる。

そこで、本調査で考えた原因を述べてみたい。それは、対応する括弧や記号が欠けている場合、つまりテキストの誤入力が関わっているということである。もともとのテキストに誤りがあったのか、電子化の際に誤って入力されたのか不明であるが、タグ付けの対象に誤りがあると、当然ではあるが、正確なタグ付けができないのである。

例えば、以下の文は何らかの問題で対応する括弧（この場合ともに“「”）が欠けている。(52)は「選択と集中」、(53)は「浮いている」が正しいのではないかと思われるが、これらが正確に入力されていたならば問題なく文タグが付加されたであろう。前出の(1)も誤入力により“」”が欠け、(13)も対

応する“（”が欠けているためであろう。

(52) <sentence>□得意な分野に経営資源を集中させ、効率のよい経営をめざすことを畑(ママ)選択と集中」などと呼ぶ。しかし、アウトソーシングを導入したものの、（略）、本末転倒というものである。（略）(PB13_00111)

(53) <sentence>それらは彼女には、いかにも浮いている感じ」がする。「先生は、タテマエばっかりできらい」。この（略）表現は、学級崩壊を経験した六年生の女の子からも聞いた。</sentence> (PB13_00292)

このように対応する記号が入力されているというのは重要な要因のようである。それは、同じタグを持つ(54)と(55)を比較したとき、(54)の問題点は、対応する「“」に対応する「”」がないこと以外考えられないからである。

(54) <sentence>コミットメントとは、“仕事を達成するための決意(ママ)“である。経営者マインドや事業家精神を持った人材に対し、経営者陣がコミットメントを引き出そうとすることは、エグゼクティブコーチングには欠かせない行為である。そして、コミットメントを引き出すためには、上記４つの条件を考慮しながら行う必要がある。</sentence> (PB23_00220)

(55) <sentence>それから、向こうの政府に協力して、輸出拠点にするということになると、篠原さんの言われる“ブーメラン現象”になって、今度は本社の企業と競合的な関係に立つ。</sentence> (LBa2_00020)

以上のように対応する括弧や記号が欠けている場合、文タグに誤りが見られる。長谷川(2013)では、『CD-毎日新聞』の問題点の指摘の際に、誤入力を探すヒューリスティックな方法として、対応する括弧の数が一致しない部分を見つけることを挙げたが、ここでも有効に機能するのではないかと思われる。

さらに、<sentence type="quasi">というタグを持つ場合、複数の文が入って

いても、それらを分けるようにはなっていない。では、どのように<sentence>と<sentence type="quasi">を区別してタグ付けを行ったのかについては、具体的な方法が分からないため不明であるが、文字列の終端のみに着目しているのではないかと思われる(タグ付けを行った方に確認したわけではないので、あくまで推測である)。

なお、<sentence type="quasi">というタグを持つ場合に多く見られたので、このタグも関係しているようであるが、原因の特定までには至らなかった。

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。

金(2009)では「テキストの中の必要ではない記号・文字列(ゴミ)を取り除いたり、間違った文字列を訂正したりすること」を「データクリーニング」としている。不必要なタグを削除したりタグのない部分に必要なタグをつけることもクリーニングと考えれば、クリーニングの対象としては、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(出版-雑誌)・PN(出版-新聞)は修正箇所数が少なく、PB(出版-書籍)・LB(図書館-書籍)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(国会議事録)・OL(法律)・OV(韻文)のように修正箇所数が少ないものも多いが、OY(Yahoo!ブログ)・OC(Yahoo!知恵袋)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(Yahoo!ブログ)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る。例えば、(56)(57)では“♡”や“♥”、“♫”“♪”などが文の終端に位置し、前出の(46)ではさらに“w笑”などの形も見られた。

(56) <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)

(57) <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った。(58)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。

(58)　(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、対象とな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。

また、上記のような原因で文タグが抜けているデータが見られたが、文タグを研究者個々人で追加した場合、研究データの共有というコーパスの持つ重要な特性が失われてしまうため、『日本語話し言葉コーパス』のようにタグの追加・修正が施された第2版が入手できることを希望したい。

参考文献

金明哲(2009)『テキストデータの統計科学入門』、岩波書店

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(2013)「「CD-毎日新聞データ集」に含まれるデータの特徴と使用上の注意点について」、『人文学報』第 473 号、首都大学東京、pp.31-49