BCCWJに収められた新種の言語資料の特性について

―― データ重複の諸相とコーパス使用上の注意点 ――

田野村　忠　温

キーワード：現代日本語書き言葉均衡コーパス，BCCWJ，サブコーパス，インターネット文書，データの重複

1　はじめに

国立国語研究所において５年間をかけて開発された「現代日本語書き言葉均衡コーパス」が完成し、2011年にはその検索サイト「少納言」「中納言」が公開され、2012年にはコーパスデータを収めた「BCCWJ-DVD版」が利用可能になった。コーパスの名称が長く言及に不便なので、以下ではその英語名“Balanced Corpus of Contemporary Written Japanese”に基づく略称BCCWJを用いる。

すでにほかの所で述べたように（拙論(2009, 2011)）、BCCWJの完成は現代日本語研究史上の画期的な出来事と言ってよい。BCCWJは今後の日本語研究において標準的な言語資料として広範に利用されることであろう。従来よく使われてきた日本語の電子資料と言えば、新聞記事と古めの文学作品だけであった。これに対し、BCCWJは書籍だけでも約２万冊という膨大な数の出所から採取された現代日本語の実例を集成したコーパスであり、そうした一般性の高い資料を利用したいという日本語研究者の長年の念願がBCCWJの完成によって実現した。

しかし、日本語の研究にBCCWJを適切に生かすためには、利用者は少なくとも２つの点を認識し、十分に理解しておくことが必要である。それ

は、BCCWJが複雑な内部構成を有するということ、そして、従来の日本語研究にはあまり使われてこなかった種類のデータを含んでいるということである。“今まで使ってきた「CD-ROM版 新潮文庫の100冊」のテキストの代わりにBCCWJを使いさえすれば日本語の現実をより正確に反映した分析が可能になる”といった素朴な期待は、BCCWJに対する無理解に基づく誤った考えと言わざるを得ない。

以下においては、２節で上述の注意点の意味するところを具体的に説明したうえで、３～４の各節においてBCCWJの使用上注意を要する問題点の一端を調査・分析に基づいて明らかにする。

２　BCCWJの構成

２.１　サブコーパス

BCCWJは出版、図書館、特定目的と名付けられた３つのサブコーパスから成る。それぞれの意味合い、位置付けは、簡単には国立国語研究所のKOTONOHA計画のWebサイト（http://www.ninjal.ac.jp/kotonoha/）、より詳しくは国立国語研究所コーパス開発センター(2011)で説明されている。後者はBCCWJ-DVD版に収められた電子媒体の解説書であり、以後「利用の手引」と呼ぶ。

各サブコーパスのデータは表１に示したように下位区分される。

表１　BCCWJの構成

サブコーパス	下位区分
出版	書籍、雑誌、新聞
図書館	書籍
特定目的	白書、教科書、広報紙、ベストセラー、Yahoo!知恵袋、Yahoo!ブログ、韻文、法律、国会会議録

実は少々困ったことに、これらの下位区分を指す単位名称が一定していない。「利用の手引」の第１章では「レジスター」、第２章では「媒体」、

第３章では「メディア」、第５～７章では ── また中納言でも ── 「サブコーパス」と４通りの用語が用いられている。少納言では書籍関係のデータの扱いに異なるところがあるが、「メディア／ジャンル」という表現が使われている。

言及の都合上いずれかを選ぶ必要があるので、ここでは「サブコーパス」を採る。選択の最大の根拠は、特定目的サブコーパスは特定の母集団 ── すなわち、明確に定められた範囲に収まる日本語の書き言葉の総体 ── を前提としてそれを適切に表すように ── その忠実な見本となるように ── 作られたコーパスではなく、それぞれに独立したコーパス（ないしデータ）の寄せ集めという性格が強いということである。ほかにも、例えばベストセラーをレジスターとは呼びにくいとか、韻文を媒体やメディアとは呼びにくいといったことも考慮した。

サブコーパスという用語を２つのレベルに適用するこの選択により、例えば出版サブコーパスは書籍ほかの３つのサブコーパスで構成されることになる。全体として言えば、BCCWJ は３種13類のサブコーパスから成るコーパスということになる。

２．２　新種の言語資料

出版と図書館のサブコーパスは書籍、雑誌、新聞という一般的な出版物から取られた日本語をその内容としており、従来の日本語研究で広く用いられてきた言語資料とデータの性質上基本的に同等である。

他方、特定目的サブコーパスには一般の出版物以外に、教科書やインターネット上の日本語、国会会議録といったデータも含まれる。話し言葉の記録である国会会議録の「現代日本語書き言葉均衡コーパス」への採録には疑問が残るが、ともあれ特定目的サブコーパスにおいては、教科書あるいは国会の場面で使われる日本語の分析といった、種々の特定の目的の

研究に利用するためのデータが提供されている。

特定目的サブコーパスに含まれる９種類のサブコーパスの日本語は、従来の日本語研究で使われてきた一般的な出版物の日本語を基準とすればさまざまな点で異質性が高い。その意味で、新種の日本語研究資料だと言うことができる。

２．３　サブコーパス併用の危険

研究の目的や方法にもよるが、割り切った言い方をすれば、こうした新種のデータを出版や図書館のサブコーパスと無差別に併用するとすれば、BCCWJのまっとうな使用法にはならない。

BCCWJ全体から「と思っております」「というふうに思っております」という２通りの言い回しを検索すればそれぞれ1,716件、523件の用例が得られる。しかし、そこから現代日本語における後者の使用率が23％だなどと結論付けることはできない。なぜならば、「というふうに思っております」の用例523件のうち517件は国会会議録という特定のサブコーパスに集中して現れるものだからであり、また、出版サブコーパスを構成する３つのサブコーパスどうしを例外として、そもそも異なるサブコーパスにおける用例数を合算したり比較したりすることは意味を持たないからである。

要は、複数のサブコーパスを単純に足し合わせて使うことには一般に問題があり得るわけであるが、データの異質性の高い特定目的サブコーパスの併用にはとりわけの慎重さが必要である。

次節以下においては、特定目的サブコーパスに含まれるサブコーパスを取り上げ ── と言っても、紙幅の制約により２つのサブコーパスに絞らざるを得ないのであるが ── 、データ重複の問題を中心とする注意すべきそれらの特質について述べる。なお、ここでの調査・分析は、主として

BCCWJ-DVD版のM-XMLディレクトリに収められたXML文書から拙作ソフトウェアbccwj2textによって抽出したテキストによる。[1] また、字数の統計その他の処理において、空白や記号の類は含め、改行はないものとして扱う。

3　Yahoo!ブログサブコーパス

3.1　言語研究資料としてのインターネット文書

インターネット上に存在する日本語文書は膨大かつ多様で、言語研究資料としてきわめて大きな価値と魅力を有する。そのインターネット文書の短所と言えば、一般に予想されやすいのは特殊な言葉遣いの出現や書き誤りの多さといった点であろうが、実際に最も問題となるのはむしろ同一データの重複出現である（拙論(2010)）。インターネット上にはさまざまな事情で同一の文章、段落、文、句が繰り返し現れる。

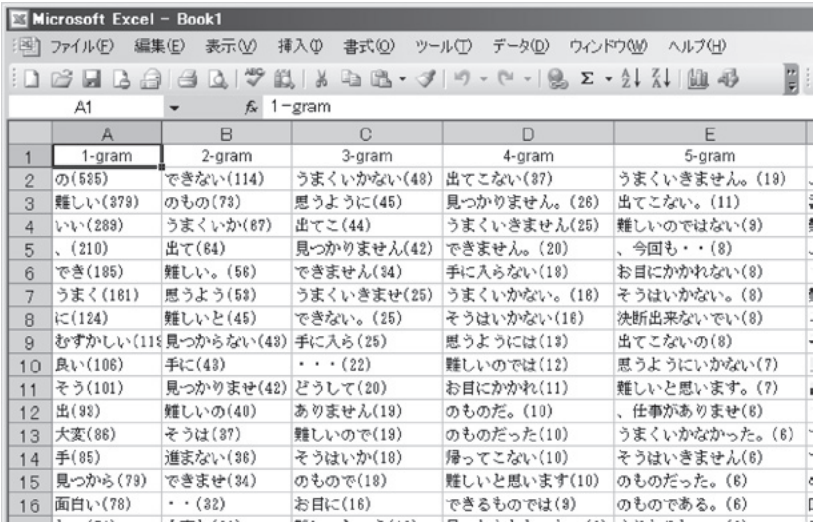
BCCWJにもインターネット文書は特定目的サブコーパスの２つのサブコーパスYahoo!知恵袋、Yahoo!ブログとして収められ、両者を合わせてBCCWJ全体の約２割の分量を占めている。そして、それらのサブコーパスも果たしてデータ重複の問題をまぬがれない。その様相の分析に進む前に、筆者がこの問題の認識に至った経緯を述べ、データ重複がBCCWJによる日本語の分析に与える影響の実例を示す。

3.2　データ重複とその日本語の分析への影響

筆者は最近、中納言によるBCCWJの検索結果に基づいて語句のコロケーションを調べる簡易な分析ソフトウェアBNAnalyzerを作成した。[2] このソフトウェアは、中納言での検索結果をもとに、検索語の前後にどのようなN-gram ── N個の短単位または長単位の連続 ── がよく現れるかを分析し、頻度順に表示する。BNAnalyzerは“BCCWJ N-gram Analyzer”の

略である。

図１に示す分析結果の例は「なかなか」に続くN-gramの一覧である。



Microsoft Excel - Book1

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

A1 1-gram

	A	B	C	D	E
1	1-gram	2-gram	3-gram	4-gram	5-gram
2	の(585)	できない(114)	うまくいかない(48)	出てこない(37)	うまくいきません。(19)
3	難しい(379)	のもの(73)	思うように(45)	見つかりません。(26)	出てこない。(11)
4	いい(289)	うまくいか(67)	出てこ(44)	うまくいきません(25)	難しいのではない(9)
5	、(210)	出て(64)	見つかりません(42)	できません。(20)	、今回も・・(8)
6	でき(185)	難しい。(56)	できません(34)	手に入らない(18)	お目にかかれない(8)
7	うまく(181)	思うよう(53)	うまくいきませ(25)	うまくいかない。(18)	そうはいかない。(8)
8	に(124)	難しいと(45)	できない。(25)	そうはいかない(16)	決断出来ないでい(8)
9	むずかしい(11	見つからない(43)	手に入ら(25)	思うようには(13)	出てこないの(8)
10	良い(106)	手に(43)	・・・(22)	難しいのでは(12)	思うようにいかない(7)
11	そう(101)	見つかりませ(42)	どうして(20)	お目にかかれ(11)	難しいと思います。(7)
12	出(93)	難しいの(40)	ありません(19)	のものだ。(10)	、仕事がありませ(6)
13	大変(86)	そうは(37)	難しいので(19)	のものだった(10)	うまくいかなかった。(6)
14	手(85)	進まない(36)	そうはいか(18)	帰ってこない(10)	そうはいきません(6)
15	見つから(79)	できませ(34)	のもので(18)	難しいと思います(10)	のものだった。(6)
16	面白い(78)	・・(32)	お目に(16)	できるものでは(9)	のものである。(6)

図１　「なかなか」に後続する N-gram

この一覧は、「なかなか」の後には「の」「難しい」「いい」（1-gram）、「でき-ない」「の-もの」「うまく-いか」（2-gram）、「うまく-いか-ない」「思う-よう-に」「出-て-こ」（3-gram）などの表現がよく現れることを示している。こうした情報は語句の用法やコロケーションの考察に役立つ。

ところが、BNAnalyzerを作成したあとBCCWJの検索結果に基づいてあれこれの語句を分析してみると、一見しておかしいと分かる結果が得られるケースがあまりに多いことが分かった。次に２つの例を示す。

	D	E	F	G
1	4-gram	5-gram	6-gram	7-gram
2	わかった。□(80)	見つけたいならＹａｈｏｏ！(25)	見つけたいならＹａｈｏｏ！縁結び(25)	見つけたいならＹａｈｏｏ！縁結び□(25)
3	見つけたいならＹａｈｏｏ(25)	わかった。□「(10)	はわからなかった。□(4)	ダウンロードできます。◆□メルマガ(4)
4	はわからなかった(10)	はわからなかった。(9)	ダウンロードできます。◆□(4)	肌目になるるスキルでは(4)
5	は答えなかった(8)	は答えなかった。(8)	ファーストメールを送りました(4)	売り切れてしまうそうなので(4)
6	わかります。□(8)	忘れてしまいます。(7)	食卓に出す。（4(4)	戻ってきた。□「(4)
7	出てきた。(8)	戻ってきた。(7)	肌目になるるスキルで(4)	いっぱいになってしまいます。(3)
8	分かった。□(8)	気がついた。(6)	売り切れてしまうそうなの(4)	治療を受けたいときなど）(3)

図２　「すぐに」の後続文脈の N-gram

	E	F	G	H
1	5-gram	6-gram	7-gram	8-gram
2	に殉じて自分は(52)	気持ちに殉じて自分は(52)	の気持ちに殉じて自分は(52)	あなたの気持ちに殉じて自分は(52)
3	のことを自分は(35)	すべてのことを自分は(30)	関するすべてのことを自分は(26)	に関するすべてのことを自分は(26)
4	の気持ちに自分は(22)	あなたの気持ちに自分は(22)	あなたに自分はその成果を(11)	てあなたに自分はその成果を(11)
5	、次のような(20)	に自分はその成果を(11)	に殉じてあなたに自分は(8)	あなたの気持ちに殉じて自分が(8)
6	「愛」という(18)	気持ちに殉じて自分が(8)	の気持ちに殉じて自分が(8)	気持ちに殉じてあなたに自分は(8)
7	ている」という(18)	殉じてあなたに自分は(8)	口と言うように「その(8)	くれればあなたの気持ちに自分は(8)
8	」という意味の(15)	あなたのために自分は(8)	ばあなたの気持ちに自分は(8)	白い口と言うように「その(8)

図３　「言葉」の先行文脈の N-gram

いずれの例においても、とうてい一般性を持つとは考えられない「見つけたいならYahoo!縁結び」とか「あなたの気持ちに殉じて自分は」といったN-gramがリストの上位を席巻している。

２例のうち、前者（図２）はYahoo!知恵袋かYahoo!ブログのいずれかのサブコーパスに原因がある可能性が高い。後者（図３）についてはこの分析結果だけからでは事情は分からない。そこで、BCCWJ-DVD版とインターネットを用いて調べてみたところ、真相は次の通りであった。

まず、図２に見る「見つけたいならYahoo!縁結び」という高頻度N-gramはYahoo!ブログサブコーパスに現れるもので、しかし、ブログ記事の書き手が書いたものではなく、ヤフー株式会社の運営する出会い系サイトの宣伝用ブログ「そろそろ恋愛しませんか？」（http://blogs.yahoo.co.jp/yjpartnerblog/）の記事に自動的に張られるリンクのタイトル「［ 結婚相手をすぐに見つけたいならYahoo!縁結び ］」の一部であった。リンクの実例は例えばインターネット上のYahoo!ブログの記事http://blogs.yahoo.co.jp/yjpartnerblog/archive/2008/11/11で見ることができる。この記事はBCCWJにサンプルID OY14_29479として収録されている。

図３にある「殉じて」の奇異な用法を含む多数のN-gramはいずれも「世界で一番、誰よりも愛してる人へ」と題されたブログ（http://blogs.yahoo.co.jp/geoburgher/）に頻出するものであった。BCCWJにはこのブログから「殉じて」を含む記事が36件取られており、そこには「殉じて」が計243回も現れ、BCCWJ全体に含まれる「殉じて」計264例の実

に92％を占めている。図３に見る、「自分」を含むほかのN-gramも同じブログに現れるものであった。

このように、不自然な分析結果の原因は、主にBCCWJのYahoo!ブログサブコーパスにおける同一データの重複出現にあることが判明した。

インターネットに高頻度で現れる一般性の低い表現は、人間がその都度書いたわけではなく、機械生成ないし複製されたものに過ぎない。言語研究上そのようなものを通常の言語データと同列に扱ってはならないことは明らかであるが、BCCWJ評価の観点から引き続き問われるべきは、コーパスにどのような重複がどれくらい含まれているのかという問題である。

３．３　サンプルの完全一致

データの重複と一口に言っても、事例ごとに程度の差がある。まず、サンプル全体の完全な一致について見る。

Yahoo!ブログサブコーパスには52,680件のサンプルが収められているが、調べてみると、うち410件のサンプルについては完全に一致するサンプルが別に存在することが判明した。

410件のサンプルをテキストの同一性に基づいて分類すれば103種類になる。そのうち、もっとも長いものは次の２サンプルで、2,863字の長さである。ここにはその冒頭部分だけを示す。[3)]

サンプルID：OY11_03448＝OY11_03449

もぅ要らない！ＣＩＡ結成の架空自民党。日本統治体制は８１６年間「亀卜政治」
海外では１５０年前～日本の自衛隊員レベルで認識。
狂言的天皇と現政府の相互に「責任を取れない劣等感」を
【弱点】として毎回軍戦略の【標的】に！
米海・陸軍戦略プランに必ず記される。
＜ＣＩＡエージェント岸信介が
＜内政密告１回１０億円報酬金を貰いつづけ＜／
＜５５年自民党結成。
＜ＣＩＡエージェント岸信介党内に３５億円ばらまき＜／
＜５７年首相就任。

＜５８年岸内閣の弟左藤栄作蔵相は「国政選挙資金」を＜／
＜米国大使館を通じて無心。
　（後略）

逆に、最も短いのは次の４サンプルで、記号・空白を含めて21字の長さである。

サンプルID：OY01_00197＝OY01_00544＝OY15_00161＝OY15_01346
オークション　＞　チケット、金券、宿泊予約

もし空白を文字に含めなければ最短は次の２サンプルで、長さは0字である。ここでは空白を「□」で示している。[4)]

サンプルID：OY13_04823＝OY14_33122
□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□

互いに完全一致するサンプル数が多いもののうち、字数が最も多い（429字）のは次の19サンプルである。

サンプルID：OY07_00505＝OY07_00642＝OY07_00680＝OY07_01181＝OY07_01335＝OY07_01353＝OY07_01424＝OY07_01581＝OY07_01627＝OY07_01814＝OY07_01926＝OY07_02204＝OY07_02245＝OY07_02303＝OY07_02622＝OY07_02680＝OY07_02754＝OY07_02842＝OY07_02873
定期的に飲むだけで簡単に痩せれてしまう今話題のサプリメントを入手できるサイトを見つけました！！　私は飲み始めて２週間位でお腹がスッキリしてきて太ももがなり細くなりました。テレビで紹介されてから、品切れ状態で今度いつ入荷されるかわかりませんが、これを見てくれた人だけに教えてます。あまりに安かったので、私は６ヶ月分をまとめ買いしました。
　（中略）
いつも品薄状態なので興味がある方はお早めに。
詳細は　→　ｈｔｔｐｓ：／／ｗｗｗ．ｍｏｓｈｉｍｏ．ｃｏｍ／ｉｔｅｍ／８９５８７／６９８１３２６

以上の例に含まれる「もぅ要らない」「チケット、金券」「定期的に飲むだけで」といった語句を中納言で検索してみれば ── 文字列検索によるのが手っ取り早い ── 、異なるサンプルIDを持つまったく同じ“用例”が複数表示され、データの重複を容易に確かめることができる。これは以後

の重複の事例についても同様である。

完全一致するサンプルIDの組とその字数の一覧は表２の通りである。スペースの関係で字数が400字以上のサンプルに限って示す。

表２　完全一致のサンプル一覧（部分）

サンプルID	字数
OY11_03448＝OY11_03449	2,863
OY14_40012＝OY14_43821	2,467
OY14_49176＝OY14_49589	2,467
OY01_01375＝OY01_02646	1,365
OY04_01476＝OY04_01665	1,172
OY01_00180＝OY01_00214＝OY01_00260＝OY01_00552＝OY01_00642＝OY11_00260＝OY11_00385＝OY11_01547＝OY11_01553	1,038
OY01_00506＝OY01_00579＝OY02_00125	947
OY01_00055＝OY11_00113	928
OY03_02330＝OY03_03987	855
OY02_00210＝OY02_00212	646
OY11_00268＝OY14_10909	626
OY05_02386＝OY05_02853	593
OY14_29957＝OY14_34602	540
OY05_06006＝OY05_06422	468
OY01_02527＝OY01_02666＝OY01_02900	441
OY01_03129＝OY01_03491＝OY11_08879	441
OY01_02152＝OY01_03450	431
OY07_00505＝OY07_00642＝OY07_00680＝OY07_01181＝OY07_01335＝OY07_01353＝OY07_01424＝OY07_01581＝OY07_01627＝OY07_01814＝OY07_01926＝OY07_02204＝OY07_02245＝OY07_02303＝OY07_02622＝OY07_02680＝OY07_02754＝OY07_02842＝OY07_02873	429
OY07_00235＝OY07_00335＝OY07_00354＝OY07_00392	428
OY14_02964＝OY14_02980＝OY14_03123＝OY14_03268＝OY14_03721＝OY14_05723＝OY14_06940＝OY14_06947＝OY14_07775＝OY14_09850＝OY14_11159	401

３．４　サンプルの部分一致

サンプルの部分的な一致には、完全一致に近いものから小部分の一致に過ぎないものまでさまざまな段階がある。また、一致・不一致のあり方も事例によって異なる。

注意すべきは、部分一致の問題は完全一致のそれに劣らず重要であることである。なぜならば、短いサンプルの完全一致よりも長大なサンプルの程度の高い部分一致のほうが日本語の分析にはるかに大きな影響を与える

からである。

部分一致を含むサンプルの数は完全一致のそれをはるかに上回る。しかし、部分一致は程度問題であるので、それを含むサンプルの範囲を明確にすることは原理的に不可能である。また、大量のテキストから部分一致を検出することには現実の処理上の限界もある。

ここでは、BCCWJを使用するうえで深刻な問題を引き起こす可能性のある、データ一致の程度のはなはだしい事例を２つ見る。

第１の事例は、「犬幼稚園ＢｕｄｄｙＤｏｇ」のブログ（http://blogs.yahoo.co.jp/lovedog111222/）から取られた以下のサンプル38件である（一部に完全一致のサンプルを含む）。

OY05_00030, OY05_00066, OY05_00447, OY05_00486, OY05_00699, OY05_01030, OY05_01096, OY05_01213, OY05_01840, OY05_02017, OY05_02075, OY05_02130, OY05_02232, OY05_02252, OY05_02286, OY05_02335, OY05_02386, OY05_02461, OY05_02834, OY05_02853, OY05_03041, OY05_03152, OY05_03182, OY05_03316, OY05_03593, OY05_03641, OY05_03759, OY05_04364, OY05_04503, OY05_04574, OY05_04712, OY05_04887, OY05_04944, OY05_05154, OY05_05424, OY05_06006, OY05_06217, OY05_06422

これらのサンプルにおいては、１文ないし数文のまとまりが、異なる組み合わせや順序において現れる。例えば、次はOY05_02853＝OY05_02386のテキストである。実際にはもっとこまめに改行されているが、論述の便宜上一定の範囲を段落にまとめた形で示す。各段落末の番号は後の参照のために加えたものである。

サンプルID：OY05_02853＝OY05_02386

犬幼稚園ＢｕｄｄｙＤｏｇに愛犬を預ける飼い主さん。送り迎えの際、愛犬たちがじゃれあう広場でお茶をしつつ、その間に情報交換タイムが始まります。しつけや健康管理の話はもちろん、去勢手術や避妊手術について、詳しい説明や報告をしたり、不安な事を質問したり。フードやおやつの選び方・与え方、犬が喜ぶおもちゃや、留守中に便利なグッズについてなど、話題は多岐にわたります。　……①

犬の幼稚園「ＢｕｄｄｙＤｏｇ」は自由登校システムのため、毎回少しずつ違うメンバーが顔を合わせて、情報は増える一方。みんな子（犬）育て真っ最中で、お互いに相談もしやすいようです。仔犬は本来、目を輝かせ好奇心旺盛・天真爛漫で元気すぎ

る程です！落ち着きがない・無関心、無反応・それは仔犬の本質ではありません。
……②
犬達は犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇで仲良くじゃれあったり、時にはおもちゃを取り合ってみたり・・遊び疲れて寄り添って眠っていたり・・愛くるしい表情をいっぱい見せてくれます。その姿は本当に純粋で愛しい程です。『犬の社会性』を身につけることが、将来に良い子になる秘訣。
……③
「三つ子の魂百までも」は、人間も犬も一緒なんですね。
……④
犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇは、仔犬にとって世界を広める第一歩でもあるわけです。
……⑤
犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇは、きっとあなたと愛犬の間に新しい発見と更なる楽しみをもたらしてくれるはずです。お気軽にご相談ください。
……⑥

同じブログから取られた他のサンプルOY05_02286＝OY05_02232においては、このテキストの後ろに短いテキストが付け加えられている。また、サンプルOY05_04364では冒頭に短いテキストが付け加えられ、かつ、④以下のテキストが削除されている。

上に段落の形で示したテキストの各部分は、このブログにおいてしばしばテキストの構成要素として ── ときには、わずかに修正された形で ── 繰り返し現れる。それを“要素テキスト”と呼ぶことにすれば、次に示すサンプルでは、上記テキストの４つの要素テキスト③～⑥が使われ、かつ、下線を施した別の要素テキストが⑤と⑥のあいだに挿入された形になっている。また、要素テキスト③の冒頭には「仔」の１字が加えられている。

サンプルID：OY05_02461

仔犬達は犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇで仲良くじゃれあったり、時にはおもちゃを取り合ってみたり・・遊び疲れて寄り添って眠っていたり・・愛くるしい表情をいっぱい見せてくれます。その姿は本当に純粋で愛しい程です。『犬の社会性』を身につけることが、将来に良い子になる秘訣。
……③
「三つ子の魂百までも」は、人間も犬も一緒なんですね。
……④
犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇは、仔犬にとって世界を広める第一歩でもあるわけです。
……⑤
犬幼稚園は犬をしつけるのではなく、犬とじゃれあうことにより社会性を形成する。家族以外の人と接することにより人への信頼・服従を確立する。飼い主は飼い主として必要な知識を学んでいただく所です。
犬幼稚園　ＢｕｄｄｙＤｏｇはきっとあなたと愛犬の間に新しい発見と更なる楽しみをもたらしてくれるはずです。お気軽にご相談ください。
……⑥

程度のはなはだしい部分一致の第２の事例としては、先に見た「世界で一番、誰よりも愛してる人へ」（http://blogs.yahoo.co.jp/geoburgher/）のブログの記事を見る。このブログからはBCCWJに少なくとも次の36件のサンプルが取られている。

OY14_04922, OY14_07061, OY14_09969, OY14_14614, OY14_15863, OY14_17848, OY14_22996, OY14_32704, OY14_33189, OY14_34427, OY14_34837, OY14_34962, OY14_35316, OY14_36252, OY14_38580, OY14_38725, OY14_40012, OY14_42324, OY14_42502, OY14_42913, OY14_43821, OY14_44162, OY14_44932, OY14_45460, OY14_46975, OY14_47226, OY14_47461, OY14_49176, OY14_49589, OY14_50563, OY14_51147, OY14_51273, OY14_53034, OY14_53047, OY14_53827, OY14_54064

このうちの１件のサンプルOY14_53034の冒頭の２文 ── 第２の句点までの内容 ── だけを示せば次の通りである。このブログの記事では文の途中でも改行が頻繁に挿入されているが、文内の改行は省いて示す。

サンプルID：OY14_53034

（３からつづく）今夜から明日にかけて一部の地域で雨に注意する必要があり今日は昨日よりさむいのでさむさにも注意する必要があり花粉症も明日は少ないとなっているけれどまだ終息しないようで黄砂についても気をつけなければならないようでインフルエンザもまだかなりはやっていて個人的な経験で申し訳がないのだけどこの時期も風邪を引きやすく油断がならないからどうかさむさによる悪影響や花粉症による悪影響や雨や急な強い雨による悪影響やインフルエンザや風邪による悪影響が絶対に絶対に絶対に絶対に絶対に絶対に絶対に防がれ絶対にあなたが暖かい健康でゆとりのある毎日を過ごしていてほしいとほんとうにほんとうにほんとうにほんとうにほんとうにほんとうにほんとうに強くおもう。
永久に絶対にどんな場合も現実としても可能性としても当為としてもあらゆるすべてのことについてあなたの気持ちに殉じてあなたのためになるように自分は言葉だけじゃなくて実証されているようにどんな犠牲を払っても自分の命を犠牲にしても自分のみがすべての責任を負って果たして絶対にどんなことでも解決してなんでもできればとおもうのでどうかなにかあればどんなことでもどんな方法でも伝えてくれればとおもう。
（後略）

このブログの記事には全般に際立った特徴があり、きわめて類似度が高く特異な印象を与える長文が繰り返し現れる。かと言って完全な機械生成物というわけでもなく、上の引用に見るようにその時節の話題が添えられていたりもする。また、このブログでは重複の多い長文記事が短い時間間

隔で続々と掲載されていることも注意を引く[5]。ただ、リンクの設置による広告収入を目的としたブログにも見えず、こうした記事の作成・掲載の目的は確認することができなかった。

3.5　語句の極端な反復

上で最後に見た例には、「絶対に絶対に絶対に絶対に～」「ほんとうにほんとうにほんとうにほんとうにほんとうに～」という同一語句の極端な回数の反復も含まれていた。一般の出版物ではこのような反復はめったに見られないが、Yahoo!ブログサブコーパスには同類の事例がほかにも見られる。反復回数の最も多いのは、「眠い」を744回反復した次のブログ記事である。

サンプルID：OY14_38757

眠い
眠い
眠い
（中略）
眠い
眠い
眠い眠い眠い眠い眠い眠い眠い眠い眠い眠い眠い眠い眠い眠い眠い眠いねm（黙れ
勉強すると必ず眠くなる僕
・・・ってかみんな勉強してるとだんだん眠くなると思うよ
だってずっと座ってたら血の巡りが遅くなる（のか？）じゃん
ま、どうでもいいけどね（Σ

この種の反復は人が目で見る分には取るに足りない無害なものであるが ── すべての「眠い」を一々読んだりはしない ── 、語句の頻度に関する機械的な集計や統計分析では研究者の気付かないうちに結果に大きな影響を及ぼす可能性がある。

3.6　Yahoo!ブログサブコーパスとは何か？

Yahoo!ブログサブコーパスのサンプルを観察していると、これはいったい何なのかという疑問が生じ、頭から離れなくなる。無論、このサブ

コーパスがヤフー株式会社の提供によるYahoo!ブログのデータから抽出されてBCCWJに収められたテキスト群であることは言うまでもない。ここで言う疑問とは、Yahoo!ブログに書かれている日本語とはどういう種類の日本語なのか、Yahoo!ブログサブコーパスを使って明らかにすることができるのは日本語のどのような側面なのかという疑問である。

おそらく、我々の多くがブログについて抱くのは、個人が毎日あるいは気の向いたときに近況やエッセイを書いてインターネット上に公開するものといったイメージであろう。実際、BCCWJの開発過程においては、ブログ記事の書き手の年齢や性別の情報が利用できれば、そういった観点に着目した日本語の分析が可能になるといった期待もあった。[6)]

ところが、上で見てきたように、ブログの現実はと言えば、むしろその少なからぬ部分を各種の宣伝目的の記事が占めている。だからこそ、代わり映えのしない話が普通の個人的な作文ならばあり得ないほど反復的に書き込まれることになる。組織的な書き込みの場合はそもそも書き手の年齢や性別を問うことが意味を成さない場合もあろう。

宣伝と並んでYahoo!ブログサブコーパスの性格を曖昧にしているのが、外部からの引用によって構成されたブログ記事である。Yahoo!ブログサブコーパスの特に字数の多いサンプルの内容を確かめてみると、その相当数はニュースサイトその他の記事を無断で丸ごと引用したものであることが分かる。そうしたサンプルの大半については、もとの記事を今も出典のサイトで、もしくは、第２、第３のサイトでの引用を通して確認することができる。

次に示すのはYahoo!ブログサブコーパス最長の4,948字のサンプルである。これはニュースの紹介を専らとするブログ「(゜Д゜)新聞」（http://blogs.yahoo.co.jp/pixus_wataru/）の記事で、日本経済新聞のサイトの記事２件 ── ただし、２件目は記事の途中まで ── のコピーである。

サンプルID：OY14_44274

英米のＩＣＴ戦略に学び、日本も戦略産業強化へ政策転換を（ＣＯＬＵＭＮ１）
　未曾有の経済危機に直面するなかで、英国と米国で経済対策・雇用創出策の一環としてのＩＣＴ戦略が動き出した。日本はブロードバンドインフラ整備で世界に先んじたが、その後のＩＣＴ利活用では足踏み状態にあり、このままでは危機をバネにした英米に追い越されかねない。日本も正しい方向でＩＣＴ戦略を強化すべきではないだろうか。
　（中略）
　萌えとエロの関係をどう整理するのか、もし私の予想通り、アイドルが新しいビジネスの主流になるのだとしたら、解決しなければならない大きな課題と

Yahoo!ブログサブコーパスの長大サンプルの上位７位はこのブログから取られた記事が独占している。ちなみに、上のサンプルの長さが4,948字であるのは、Yahoo!ブログにおける記事の字数が最大5,000字に制限されていることによる。当該サンプルには改行が52か所含まれており、それを合わせて5,000字となる。おそらく、コピーした内容のうち字数制限を超える部分がシステム的に削除されたということであろう。

紙幅の関係で各サブコーパスにおけるサンプルサイズの統計的分布について述べる余裕はないが、Yahoo!ブログサブコーパスのサンプルは最小13字から最大4,948字までとサイズの幅が非常に広い。上記のニュース紹介ブログから取られたサンプル以外にも、各種サイトからの丸ごとコピー、ないし、それに申し訳程度のコメントなどを加えたものによって作成された記事がYahoo!ブログサブコーパスの長大サンプルには多い。もちろん、一般には数百字のニュースも多いので、丸ごとコピーが長大サンプルに限られるわけではない。

日本語の分析に対する影響に議論を戻せば、分析に過当の影響を及ぼしやすいのはデータの重複と長大なサンプルである。そして、データ重複は宣伝を目的としたブログの記事に多く、長大サンプルはニュースサイトなどからのコピーによる記事に多い。このようなことでは、我々が期待するようなブログの日本語 ── すなわち、個人が自分の体験や思いをつづっ

て不特定の読者に向けてインターネット公開するときに使う日本語 ―― の分析はできないことになる。

日本語の研究において、言語資料の内容や性質をわきまえることなく用例を取り扱うことの不適切性や危険性はしばしば戒められるところである。さまざまな予想外のデータを含むYahoo!ブログサブコーパスの使用に際しては、データ重複と長大サンプルの弊害に注意するだけでなく、このサブコーパスがどのような実体のものであるかを理解するためにサンプルの内容を自分の目で確認しておくことが必要である。筆者の今回の調査に基づく印象で言えば、全サンプルの１％程度（約500件）を見ればおおよその傾向はつかめるであろう。

引用の関連で付け加えれば、外国語のニュース記事から機械翻訳によって得られたものと見られる日本語文で構成された次のようなブログ記事（2,579字）もある。この記事の後略部分には作成上の不手際によると見られる同一の機械翻訳ニュースの重複も含まれている。スペースの節約と見やすさのため、一部の改行と文中の無意味な空白を省いて示す。

サンプルID：OY04_02243

★ハン・ジェソク韓流スターらしいね！チェンカイコー、チャン・ツイィと慈善バザー雪道

［ニュースには朴世縁記者］

タレントハン・ジェソクが中国四川省地震災害民たちに温情の手助けを伝えた．ハン・ジェソクは１９日中国北京ソピテルワンダベイジングホテルで進行された’たいてい，中アジアスター中国四川省災害民助け合い慈善バザー会’に参加した．中国人口福祉基金回主催で開かれた今度バザー会は去る５月１２日中国四川省で発生した地震で苦痛を経験している子供たちの児童教育基金用意のために企画された公式バザー会でこの日ハン・ジェソクは映画監督チェンカイゴ，ワールドスターザングツイ，シンガポール俳優ウィルリアムなど世界的なスターたちと肩を並べて意味を一緒にした．前作ＭＢＣ’イブのすべて’，ＳＢＳ’硝子くつ’などで中華圏私の寒流スターとして立地を押し堅めているハン・ジェソクは”四川省大震災で疾病とひもじさに苦しむ子供たちを見ていつも気が重かった”と”良い主旨の行事に同参するようになって光栄だ．使い捨て行事ではない持続的な福祉活動を通じて小さな力でも四川省子供たちと災害民たちの希望を探すのに助けになりたい”と念をおした．

（後略）

理想を言えば、このような記事はBCCWJへの収録からは外されるべきものであった。インターネット文書に限って見られる固有の不適切な言語データである。

4　Yahoo!知恵袋サブコーパス

4.1　サンプルの完全一致

Yahoo!知恵袋サブコーパスにはYahoo!ブログサブコーパスに見られるほどのデータの重複はないが、完全一致のサンプルが1例だけあった。

サンプルID：OC08_00706＝OC08_05640

国勢調査って何のためにするの？

国勢調査の人口は、議員定数や地方交付税算定の基準など、法定人口として利用されます。
また、男女・年齢別人口・産業別帯・高齢者のいる世帯などの統計は、国や市町村の社会福祉雇用政策、環境設備政策、号際対策などの行政資料として利用されます。

Yahoo!知恵袋は、利用者どうしの知識の交換を目的とする問答形式の掲示板である。上のサンプルの場合、冒頭の「国勢調査って何のためにするの？」が質問で、残りの部分が回答である。また、Yahoo!知恵袋のサイトで1つの質問に対して複数の回答があった場合は、BCCWJにはそのうち「ベストアンサー」とされたものだけが回答として収録されている（「利用の手引」第3章）。

それにしても、問答の完全一致がなぜ生じたのか。上記のようにそれなりの長さを持つ問答が一字一句違わず繰り返されることは常識的に考えがたい。

筆者の推測をあえて通俗風に表現すれば、これはYahoo!知恵袋を運営するヤフー株式会社が“自作自演”の問答を半ば不手際で2度掲載してしまったものである。

そのことを理解するには、BCCWJのYahoo!知恵袋サブコーパスのデー

タが何物であるかを知る必要がある。「利用の手引」には正確な記載がないが、BCCWJに収録されているのは、実は“Yahoo!知恵袋”のデータではなく、その準備段階の“Yahoo!知恵袋ベータ版”のデータである。

Yahoo!知恵袋は2005年11月7日に正式運用を開始したが、その約１年半前の2004年4月7日にはその試験段階としてYahoo!知恵袋ベータ版の運用が始められた。[7] Yahoo!知恵袋ベータ版では必ずしも実際の利用者が問答のやり取りをしたわけではなく、一部の質問や回答はヤフー株式会社によって用意されたのであった。[8]

現行のYahoo!知恵袋のサイトにはYahoo!知恵袋ベータ版の時期の問答も併せて掲載されており、当該の２サンプルは今も次の異なる問答として参照することができる。

> http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q116243661
> （質問日時：2005/9/23 10:27:12、解決日時：2005/9/23 10:35:34、回答数：1）
>
> http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q136197063
> （質問日時：2005/9/26 13:42:59、解決日時：2005/9/27 09:38:14、回答数：7）

Yahoo!知恵袋の正式運用開始を間近に控えた2005年9月にあって、１度目の掲載は問答の模範例を示すために行われ、その３日後の２度目の掲載は複数回答からのベストアンサーの選び出しの実験あるいは例示の目的で行われた ── ただし、同一の問答が使われ、そして、BCCWJではベストアンサー以外の回答が削除されたために問答の完全一致が生じた ── といったところかと推測される。

以上、小論の目的の中心を外れるが、“Yahoo!知恵袋”サブコーパスに収められたデータが正確には“Yahoo!知恵袋ベータ版”のデータであることを明らかにする目的も兼ねて、サンプルの完全一致の生じた背景に関する推測を述べた。

4.2　サンプルの部分一致

Yahoo!知恵袋サブコーパスにおけるサンプルの部分一致については、相手の発言をコピーして引用するインターネット掲示板の慣習がデータの重複を多数生じている。

サンプルID：OC03_01870

出来高ってどうやって計算するのですか？あと、出来高をみると何がわかるのでしょうか？宜しくお願いします。

＞出来高ってどうやって計算するのですか？
約定株数を集計します。
＞出来高をみると何がわかるのでしょうか？
人気度、注目度、…など
＞宜しくお願いします。
いえいえ、どういたしまして。

もっとも、質問とそれに対する１件の回答という組合せの関係上、引用による同一表現の重複は基本的に高々２回である。したがって、多数回の重複の多く見られるYahoo!ブログサブコーパスに比べれば問題は軽い。

4.3　その他の問題

Yahoo!知恵袋サブコーパスについては、上で見たような多少のデータ重複以外に特に気の付いた大きな問題はない。

強いて言えば、質問か回答が欠けて問答の体を成さないサンプルが一部にあり、これはコーパスへの収録から外すほうがよかったと思われる。

サンプルが質問とベストアンサーの対で構成されているのになぜ一方が欠けることがあるかと言うと、BCCWJへの収録に際してアスキーアート（文字を使って描いた絵）、外国語、数式などは削除されているため、結果的に質問や回答がテキストを含まないという事態が生じ得るのである。

ただし、アスキーアートに吹き出しなどの形で書かれているせりふが質問に対する回答になっている場合も多く、そうしたせりふは回答として残

すという判断もあり得たと思われる。また、処理の誤りによって回答が欠けているサンプルもある。例えば、サンプルID OC02_00414は、Yahoo!知恵袋ベータ版の元のデータ[9]によれば本来次のようになるべきであるが、

サンプルID：OC02_00414

ＦＡＱって何の略？

Ｆｒｅｑｕｅｎｔｌｙ　Ａｓｋｅｄ　Ｑｕｅｓｔｉｏｎです（良くある質問の略です）.

BCCWJではこの回答の部分が欠けている。これはおそらく作業者か処理プログラムが回答の出だしの部分だけに基づいて外国語と誤認したものと推測される。また、サンプルID OC15_00429 ── 問答の内容の性質上、引用は控える ── も回答が欠けているが、これについても元データにある回答「CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC」は、その内容から考えて、日本語として意味を持つテキストとして残すべきものであったと思われる。

もっとも、中納言などの既存のBCCWJ関連ソフトウェアではYahoo!知恵袋サブコーパスのサンプルにおける質問と回答の関係を扱うことができないので、BCCWJの大多数の利用者にはほとんど関わりのないことではある。また、不完全問答のサンプルがYahoo!知恵袋サブコーパス全体に占める割合は大きくなく、いずれにしても重大な問題ではない。

5　おわりに

中納言の検索結果に基づくコロケーション分析ツールの作成を契機として気付いたデータ重複の問題を中心として、BCCWJに収められた新種の言語資料の特性の一端を調査・分析してみた。Yahoo!ブログ、Yahoo!知恵袋以外のサブコーパスにおけるデータ重複の様相や、データ重複以外の種々の問題に関して述べるべきことはなお多いが、すでに紙数が尽きた。

Yahoo!ブログサブコーパスにおけるデータの重複は、用例の頻度に着目する研究にしばしば破壊的な影響を与える。同サブコーパスのサンプル52,680件に完全一致の相手を持つサンプルが410件含まれる（3.3）というのは一見小さな比率のようでもあるが、3.2で見た事例からも分かる通り、コーパス全体の用例数が少ないときにデータ重複による“用例”が多数あれば、分析は致命的にゆがんだものとなってしまう。そして、サンプルの部分一致は完全一致よりはるかに多いので、問題は410件のサンプルにとどまらない。機械生成や複製による表現の“用例”をそれと知らずに通常の用例と同列に扱ってしまうことのないよう十分な警戒が必要である。

すでに2.3で述べた通り、複数のサブコーパスの単純な併用 ── BCCWJ丸ごとの使用を含む ── には問題があり得る。中でも特定目的サブコーパスに収められた新種の言語資料は特に危険性が高い。

「現代日本語書き言葉均衡コーパス」の名称を素直に受け止めれば、コーパス全体を使うことで日本語の平均像を調べることができるかのようである。しかし、筆者の理解では、実際のところは“均衡”の形容は出版と図書館のサブコーパスにのみ、それも、それぞれ独立に適用し得るものである。拙論(2009)で述べた通り、BCCWJの使用者には、“その各部分がそれぞれどのような意味で均衡であり、あるいは、ないのか、そしてまた、どのような日本語を代表し、あるいは、しないのかを十分に把握・認識したうえでコーパスを研究に用いる”ことが求められる。これは、日本語コーパスの原始時代から急速な進展を遂げたBCCWJ時代の日本語研究者に課せられた、先端のコーパスを正しく用いるために怠ることのできない課題である。

注

1）拙作ソフトウェアbccwj2textは筆者のWebサイトで公開している。本稿ではバージョン1.40を使用した。
http://www.tanomura.com/research/bccwj2text/

2）拙作ソフトウェアBNAnalyzerは筆者のWebサイトで公開している。
http://www.tanomura.com/research/BNAnalyzer/

3）このサンプルに複数含まれる「＜」「＜／」は、もとのYahoo!ブログのページに含まれていたHTMLタグ「<u>」「</u>」の一部が消されることなく残ったものである。

4）こうした無内容のサンプルが生じた明確な理由は確かめようがないが、BCCWJ-DVD版に収められたXML文書によれば何らかの特殊な文字・記号が消去された結果のようである。

5）「利用の手引」第３章によれば、Yahoo!ブログサブコーパスの作成においては、Yahoo!ブログの元データからサンプルを取得する際、“抽出時点で1,000記事以上あるブログからの記事”であることが条件の１つとされた。筆者の憶測によれば、その方針がサブコーパスにおけるデータの重複を増やす原因となった可能性がある。なぜならば、毎日欠かさず記事を１件書いても1,000件の記事を書くには３年かかり、内容のある記事を個人が人手で書くとすれば相当の勤勉を要する。しかし、機械的な記事生成や定型文の使用に頼る場合には ── そしてまた、営利組織などによる宣伝を目的としたブログのように複数の人間が記事を掲載するブログにおいては ── 1,000件達成のハードルは格段に低くなる。「世界で一番、誰よりも愛してる人へ」のブログには、本稿を執筆している2012年7月15日の午前（0時～12時）だけでも40件の長文記事が掲載されている。

6）しかし、自己申告による年齢や性別が事実かどうか分からないといった事情もあり ── ほかにも理由はあったのかも知れないが ── 、BCCWJにブログ記事の書き手の属性は収められなかった。

7）それぞれの日付はhttp://chiebukuro.yahoo.co.jp/docs/whats2004.html、http://chiebukuro.yahoo.co.jp/docs/whats2005.htmlによる。

8）より詳しくはhttp://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1011740930などを参照。

9）ベータ版を含むYahoo!知恵袋のデータは国立情報学研究所から配布されている。
http://www.nii.ac.jp/cscenter/idr/yahoo/chiebkr2/Y_chiebukuro.html

文献

国立国語研究所コーパス開発センター(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(国立国語研究所コーパス開発センター)

田野村忠温(2009)「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」『人工知能学会誌』第24巻第5号

田野村忠温(2010)「日本語コーパスとコロケーション ── 辞書記述への応用の可能性 ── 」『言語研究』第138号

田野村忠温(2011)「コーパス言語学の新たな展開」『日本語学』第30巻第14号(2011年11月臨時増刊号『言語研究の新たな展開』)

付記　本稿は国立国語研究所の第2回コーパス日本語学ワークショップ(2012年9月6日～7日)で「BCCWJに含まれるウェブデータの特性について ── データ重複の諸相とBCCWJ使用上の注意点 ── 」と題して行ったポスター発表の内容に大幅な加筆を施したものである。

(文学研究科教授)

SUMMARY

An Analysis of Data Duplication Observed in the Web-Based Subcorpora of BCCWJ

Tadaharu TANOMURA

The Balanced Corpus of Contemporary Written Japanese (BCCWJ), which is a first-ever large-size balanced corpus of the Japanese language, was completed in 2011 after five years' construction. The corpus is now accessible either through two Web applications named *Shonagon* and *Chunagon* or directly in the DVD format. BCCWJ will henceforth be employed extensively as the standard corpus of Japanese, replacing electronic texts of novels and newspapers, which have been used commonly as a substitute for the non-existent reliable corpus of the language.

Although the significance of this long-awaited corpus cannot be overestimated, it is nevertheless necessary for its user to pay sufficient attention to two facts concerning the corpus. First, BCCWJ has a complex internal structure, i.e., it consists of three subcorpora, each of which in turn consists of smaller component subcorpora. Second, some of the subcorpora comprise of distinct types of text taken from various sources, such as the Internet or the minutes of the National Diet of Japan, which may be quite different in many ways from the kinds of texts generally used in the study of Japanese.

Here we will take up the two Web-based subcorpora of BCCWJ, *Yahoo! Blog* and *Yahoo! Chiebukuro*, the latter of which is the Japanese counterpart of *Yahoo! Answers*, and analyze aspects of duplication of data contained therein, which may affects in a highly detrimental way statistical analyses of phenomena of Japanese based on BCCWJ. The high extent of data duplication observed makes us conclude that the Web-based subcorpora need to be handled with particular caution, and that a simply combined use of those subcorpora and other subcorpora should be avoided.