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――

田野村忠温（大阪大学大学院文学研究科）

On Certain Properties of the Web-Based Subcorpora of BCCWJ

Tadaharu Tanomura (Osaka University)

１．はじめに

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、句が複製されて繰り返し現れる。

インターネット文書は「現代日本語書き言葉均衡コーパス(BCCWJ)」にも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。そしてこのたび、それらのサブコーパスにおいてもデータ重複の問題が思いのほか深刻であることが明らかとなった。

以下では、発表者が問題の認識に至った経緯と、２つのサブコーパスにおけるデータ重複の様相に関する調査・分析の結果について述べる。

２　BCCWJ N-gram分析ツール BNAnalyzer

２.１　ツールの概要

発表者は過日、BCCWJの検索結果を用いて語句のコロケーションを調べるための簡易な分析ツールBNAnalyzer（Windows上で作動）を作成し、次のところで公開した。

http://www.tanomura.com/research/BNAnalyzer/

BNAnalyzerは、BCCWJ検索サイト「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表現がよく現れるかを分析する。具体的には、検索語の前後文脈のN-gram（N個の短単位または長単位の連続）の一覧を作成し、エクセルで表示する。

２.２　インストール方法および使用法

上記URLのページの説明に従ってBNAnalyzerをインストールすると、デスクトップ上にアイコンが２つ作られる。それぞれを単純版、circumcollocate版と呼ぶ。

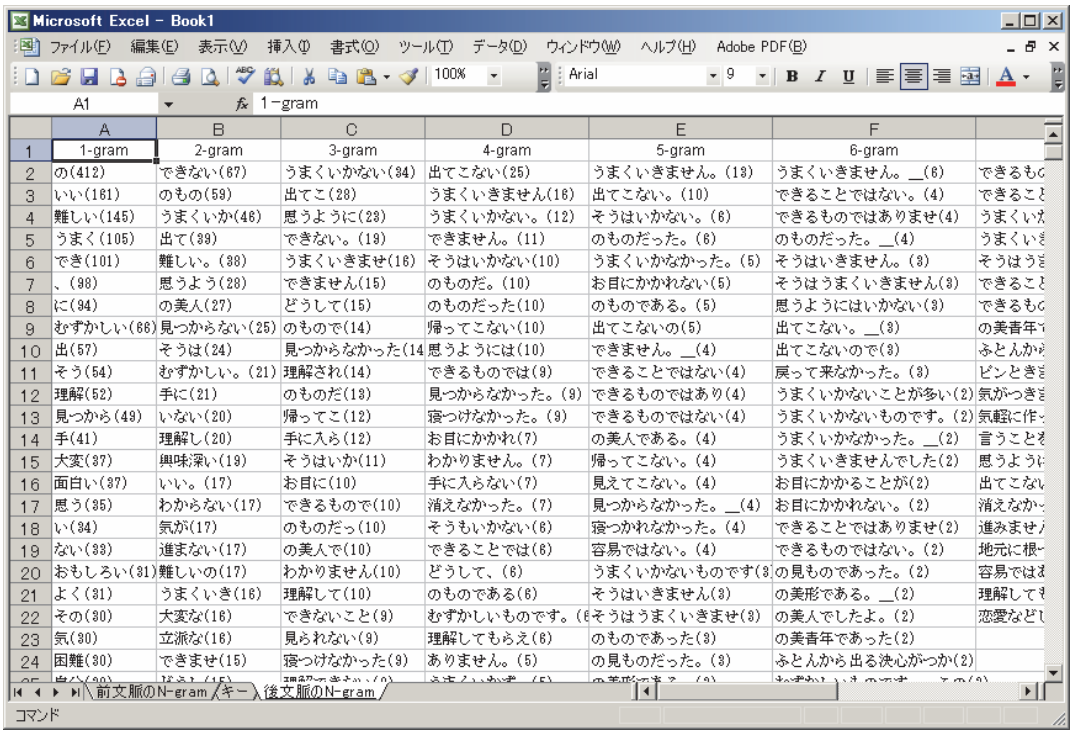
図１　BNAnalyzerのデスクトップアイコン

BNAnalyzerを使ってBCCWJの検索結果からN-gramの一覧を得るには次のようにする。

1) 中納言で［検索結果のダウンロード］によって検索結果（zipファイル）を取得
2) zipファイルをBNAnalyzerのアイコンの上にドラッグ＆ドロップ

これによりエクセルの新しいブックが開かれ、単純版の場合は、検索語（キー）の前文脈の末尾および後文脈の冒頭のN-gram（N＝1～8）が頻度順に表示される。

次に示すのは「なかなか」の検索結果に基づく後文脈のN-gramの一覧である。



	A	B	C	D	E	F
1	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
2	の(412)	できない(67)	うまくいかない(34)	出てこない(25)	うまくいきません。(13)	うまくいきません。_(6)
3	いい(161)	のもの(59)	出てこ(28)	うまくいきません(16)	出てこない。(10)	できることではない。(4)
4	難しい(145)	うまくいか(46)	思うように(23)	うまくいかない。(12)	そうはいかない。(6)	できるものではありませ(4)
5	うまく(105)	出て(39)	できない。(19)	できません。(11)	のものだった。(6)	のものだった。_(4)
6	でき(101)	難しい。(38)	うまくいきませ(16)	そうはいかない(10)	うまくいかなかった。(5)	そうはいきません。(3)
7	、(98)	思うよう(28)	できません(15)	のものだ。(10)	お目にかかれない(5)	そうはうまくいきません(3)
8	に(94)	の美人(27)	どうして(15)	のものだった(10)	のものである。(5)	思うようにはいかない(3)
9	むずかしい(66)	見つからない(25)	のもので(14)	帰ってこない(10)	出てこないの(5)	出てこない。_(3)
10	出(57)	そうは(24)	見つからなかった(14	思うようには(10)	できません。_(4)	出てこないので(3)
11	そう(54)	むずかしい。(21)	理解され(14)	できるものでは(9)	できることではない(4)	戻って来なかった。(3)
12	理解(52)	手に(21)	のものだ(13)	見つからなかった。(9)	できるものではあり(4)	うまくいかないことが多い(2)
13	見つから(49)	いない(20)	帰ってこ(12)	寝つけなかった。(9)	できるものではない(4)	うまくいかないものです。(2)
14	手(41)	理解し(20)	手に入ら(12)	お目にかかれ(7)	の美人である。(4)	うまくいかなかった。_(2)
15	大変(37)	興味深い(19)	そうはいか(11)	わかりません。(7)	帰ってこない。(4)	うまくいきませんでした(2)
16	面白い(37)	いい。(17)	お目に(10)	手に入らない(7)	見えてこない。(4)	お目にかかることが(2)
17	思う(35)	わからない(17)	できるもので(10)	消えなかった。(7)	見つからなかった。_(4)	お目にかかれない。(2)
18	い(34)	気が(17)	のものだっ(10)	そうもいかない(6)	寝つかれなかった。(4)	できることではありませ(2)
19	ない(33)	進まない(17)	の美人で(10)	できることでは(6)	容易ではない。(4)	できるものではない。(2)
20	おもしろい(31)	難しいの(17)	わかりません(10)	どうして、(6)	うまくいかないものです(3	の見ものであった。(2)
21	よく(31)	うまくいき(16)	理解して(10)	のものである(6)	そうはいきません(3)	の美形である。_(2)
22	その(30)	大変な(16)	できないこと(9)	むずかしいものです。(6	そうはうまくいきませ(3)	の美人でしたよ。(2)
23	気(30)	立派な(16)	見られない(9)	理解してもらえ(6)	のものであった(3)	の美青年であった(2)
24	困難(30)	できませ(15)	寝つけなかった(9)	ありません。(5)	の見ものだった。(3)	ふとんから出る決心がつか(2)

図２　「なかなか」の後文脈のN-gram

この一覧は、「なかなか」の後には「の」「いい」「難しい」（1-gram）、「でき-ない」「の-もの」「うまく-いか」（2-gram）、「うまく-いか-ない」「出-て-こ」「思う-よう-に」（3-gram）などの表現がよく現れることを示している。

circumcollocate版を使えば、例えば語彙素「惜しむ」の検索結果から、「寸暇を惜しんで」「努力を惜しまない」「骨身を惜しまず」のように「惜しむ」を前後からはさむように現れる２表現の慣習的な組合せがあることを知ることができる（circumcollocateの用語・概念については拙論(2010)を参照）。

以上のようなN-gramの頻度情報は、語句のコロケーションを分析する際の手がかりとなり得る。

３　BNAnalyzerによる不自然な分析結果とその原因

３.１　不自然な分析結果

さて、BNAnalyzerを作ったあと、その動作確認のために随意に選んだ少数の検索語の検索結果を分析してみると、一見しておかしいと分かる結果が得られるケースが非常に多いことが分かった。次に２つの例を示す。

	E	F	G	H
1	5-gram	6-gram	7-gram	8-gram
2	見つけたいならＹａｈｏｏ！(25)	見つけたいならＹａｈｏｏ！縁結び(	見つけたいならＹａｈｏｏ！縁結び__(	見つけたいならＹａｈｏｏ！縁結び__］(25)
3	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(	ダウンロードできます。◆__メルマガやっ(4)
4	はわからなかった。(9)	ダウンロードできます。◆__(4)	駄目になるるスキルでは(4)	駄目になるるスキルではなく(4)
5	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)	売り切れてしまうそうなので、(4)
6	忘れてしまいます。(7)	食卓に出す。〈４(4)	戻ってきた。__「(4)	治療を受けたいときなど）医療(3)
7	戻ってきた。(7)	駄目になるるスキルで(4)	いっぱいになってしまいます。(3)	食卓に出す。〈４人分(3)

図３　「すぐに」の後文脈のN-gram

	D	E	F	G	H
1	4-gram	5-gram	6-gram	7-gram	8-gram
2	」といった(54)	に殉じて自分は(47)	気持ちに殉じて自分は(47)	の気持ちに殉じて自分は(47)	あなたの気持ちに殉じて自分は(47)
3	」などという(53)	のことを自分は(34)	すべてのことを自分は(29)	関するすべてのことを自分は(25)	に関するすべてのことを自分は(25)
4	次のような(53)	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	〉」という(51)	、次のような(20)	に自分はその成果を(11)	の気持ちに殉じて自分が(9)	あなたの気持ちに殉じて自分が(9)
6	ない」という(50)	ている」という(19)	気持ちに殉じて自分が(9)	___【松下幸之助日々の(8)	_____【松下幸之助日々の(8)
7	殉じて自分は(47)	」というような(17)	__【松下幸之助日々の(8)	_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)

図４　「言葉」の前文脈の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」といったN-gramがリストの上位を席巻している。

３.２　原因の調査

上の２例のうち、前者（図３）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４）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」――DVD媒体で頒布されているBCCWJ――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３に見る「見つけたいならYahoo!縁結び」という高頻度N-gramは、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４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４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、機械的に生成または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。

以下においては、BCCWJのYahoo!ブログおよびYahoo!知恵袋の各サブコーパスの特性に

ついて、データ重複の問題を中心に観察する。

４　Yahoo!ブログの特性

４.１　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、サンプル全体が完全に一致するものについて見る。

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分かった。

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。紙幅の関係で冒頭部分だけを示す。

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。飼い主へ甘え切った岸と佐藤。岸らの「借金の肩代わり」に国民の血税が宛てがわれ＜／＜基地の米軍施設や住居の給水高熱費はタダへ！＜恵まれた環境の広範囲にわたる地域を治外法権の基地へ。＜墜落事故が発生しても「現政府は一切対応無し」へ。＜アメリカが何時でも公開すると言う記録を＜現政府は国民へ「密約」と呼ばせ、わざわざ「密約は無かった」と言う。更に年金記録を抹消し平気でいた現政府はなるほど当たり前！。（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。

サンプルID：**OY13_04823＝OY14_33122**

□□□□□□□□□□□□□□□□□□□□□□□□□□

重複の回数の最も多いのは次のサンプルで、23件のサンプルの内容が一致している。

サンプルID：**OY04_00021＝OY04_00397＝OY04_00410＝OY04_00851＝OY04_01012＝OY04_01588＝OY04_02241＝OY04_02813＝OY04_03116＝OY04_03304＝OY04_03669＝OY04_03826＝OY04_04435＝OY04_04634＝OY04_07415＝OY13_00195＝OY15_00223＝OY15_00745＝OY15_08458＝OY15_08930＝OY15_10185＝OY15_11643＝OY15_13183**

…の記事を更新しました。本部はこちら↓ＣＩＡ☆こちら映画中央情報局です

以上の例に含まれる語句――例えば、「もぅ要らない」「チケット、金券」「映画中央情報局」――を中納言の文字列検索で検索してみれば、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。

参考までに、完全一致するサンプルのIDと字数を――紙幅の関係で字数が500字以上のサンプルに限定して――示せば表１の通りである。

表１　完全一致のサンプル一覧（部分）

サンプルID	字数
OY14_29957＝OY14_34602	540
OY05_02386＝OY05_02853	593
OY11_00268＝OY14_10909	626
OY02_00210＝OY02_00212	646
OY03_02330＝OY03_03987	855
OY01_00055＝OY11_00113	928
OY01_00506＝OY01_00579＝OY02_00125	947
OY01_00180＝OY01_00214＝OY01_00260＝OY01_00552＝OY01_00642＝OY11_00260＝OY11_00385＝OY11_01547＝OY11_01553	1038
OY04_01476＝OY04_01665	1,172
OY01_01375＝OY01_02646	1,365
OY14_40012＝OY14_43821	2,467
OY14_49176＝OY14_49589	2,467
OY11_03448＝OY11_03449	2,863

なお、ここでの調査は、BCCWJ-DVD版のLUWディレクトリにあるタブ区切り形式データより復元したテキスト（文字コードはShift_JISに変換）によった。他の方法で調査すれば統計（特に字数）に多少の違いが生じる可能性がある。

４.２　サンプルの部分一致

サンプルが部分的に一致するものには、完全一致に近いものから小部分の一致にす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。

部分的な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が、類似は程度問題であるので、部分一致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限界がある。

ここでは、BCCWJを使用するうえで特に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Ｂｕｄｄｙ Ｄｏｇ」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のテキストであるが、

サンプルID：**OY05_02853＝OY05_02386**

犬幼稚園Ｂｕｄｄｙ Ｄｏｇ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。犬の幼稚園「Ｂｕｄｄｙ Ｄｏｇ」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。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。「三つ子の魂百までも」は、人間も犬も一緒なんですね。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。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。

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、「三つ子の魂百までも」以下のテキストが削除されている。

上で色分けして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つの要素テキストが使われ、かつ、波下線を施した別の要素テキストが付け加えられた形になっている。また、最初の要素テキストの冒頭には「仔」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。「三つ子の魂百までも」は、人間も犬も一緒なんですね。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。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。

程度のはなはだしい部分一致の第２の事例は、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/）のブログの記事である。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。

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。こうした記事がどのような方法で作られているかは不明であるが、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。[1]

４.３　その他の問題

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～」という同一語句の極端な反復も含まれていた。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

機械翻訳によって得られた日本語文で構成されたと見られる次のようなブログ記事もある。この記事の省略部分には作成上の不手際による同一の機械翻訳記事の重複も含まれている。

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）

---

[1] 丸山・柏野・田中(2011)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。発表者の憶測によれば、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、各種業者の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。

また、少納言・中納言による検索における“文境界無視”とでも呼ぶべき問題がある。これは特にウェブデータの場合に限ったことではないが、ブログ記事には句点を使わないものが多く、そのような場合、文の境界がないかのように扱われる。

サンプルID：**OY11_06057**

今日も晴れたので散歩に松本城のお堀に二羽の白鳥が仲良く泳いでいたお城もなんか寒そうお城の堀越しに北アルプスの常念岳が白く映えていた緑町まで歩き上土の「花月」ホテルでカレーライスを食べた帰りは縄手通りを通って帰宅　少し欲張って歩いたので左脚の脹脛が痛くなった午後からは曇りになり寒くなった

この記事は実際のブログでは次のように表示される（http://blogs.yahoo.co.jp/sinnshuu28/archive/2008/12/8）。

今日も晴れたので散歩に [見出し]

松本城のお堀に二羽の白鳥が仲良く泳いでいた
　[白鳥の写真]
お城もなんか寒そう
　[松本城の写真]
お城の堀越しに北アルプスの常念岳が白く映えていた
　[常念岳の写真]
緑町まで歩き上土の「花月」ホテルでカレーライスを食べた
　[カレーの写真]
帰りは縄手通りを通って帰宅　少し欲張って歩いたので左脚の脹脛が痛くなった
午後からは曇りになり寒くなった

この例の場合、中納言の短単位検索ではサンプル全体が長い１文として扱われ、「泳い-で-い-た-お-城」や「食べ-た-帰り」といった短単位の連続の“用例”があるものとして処理される。中納言の長単位検索や少納言による検索の場合も同様である。

これはBCCWJにおけるテキスト構造に関わる情報の付与の仕方に関係している。一般論として、行が句点以外の文字で終わる場合、文がそのまま次の行に続く可能性と、文が句点なしでそこで終わる可能性とがある。発表者の断片的な確認によればBCCWJの情報付与では両者の可能性が区別されておらず、その結果として“文境界無視”の現象が生じるものと見られる。

ほかにも細かい問題はあるが、日本語研究上特に問題となり得ると思われるYahoo!ブログサブコーパスの特性としてこれまでに気付いたのは以上である。

５　Yahoo!知恵袋の特性

５.１　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、サンプルの完全一致が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。中納言では質問と回答の境界が考慮されず、問答が上記のような一続きのテキストとして扱われるが、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（丸山・柏野・田中(2011)）。

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２度繰り返されることは常識的に考えがたい。

発表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。BCCWJのマニュアル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。[2] 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。[3]

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、同一の問答を使ったために重複が生じてしまった――といったところかと推測される。

以上、本発表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。

５.２　サンプルの部分一致

サンプルの部分一致についても、Yahoo!知恵袋サブコーパスにはYahoo!ブログサブコーパスにおけるような極端なデータの重複はない。ただし、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：**OC02_07077**
ＤＶＤをプレーヤーに入れたところ、全く動きません。以前正常に作動したプレーヤーもどれも動かなくなりました。原因は何でしょうか。ちなみに６月中旬にＮｏｒｔｏｎを更新しています。それ以降に使えなくなりました。＞正常に作動したプレーヤーもどれもどれもって、何台あるのですか？ウィルスではないですか。スキャンして下さい。

[2] 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。
[3] より詳しくは http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1011740930 などを参照。

また、この問答はYahoo!知恵袋のサイトでは次のように表示されるものであるが（http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q135001663）、

質問：
ＤＶＤをプレーヤーに入れたところ、全く動きません。
以前正常に作動したプレーヤーもどれも動かなくなりました。
原因は何でしょうか。
ちなみに６月中旬にＮｏｒｔｏｎを更新しています。
それ以降に使えなくなりました。

回答：
＞正常に作動したプレーヤーもどれも
どれもって、何台あるのですか？
ウィルスではないですか。スキャンして下さい。

4.3で見た“文境界無視”の事例と共通の理由により、中納言による検索では「どれ-も-どれ」という短単位連続の“用例”があるものとして扱われる。

６　おわりに――BCCWJ使用上の教訓

BCCWJの検索結果に基づくN-gram分析ツールの作成が契機となって明らかになった、BCCWJのウェブデータに含まれるデータ重複の問題ほかについて粗い調査・分析を行ってみた。

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。Yahoo!ブログ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件含まれるというのは一見小さな比率のようでもあるが、上で見た「殉じて」に関わる事例などが示す通り、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。Yahoo!知恵袋サブコーパスも相対的に軽度ながら同様の問題をはらむ。機械生成や単なる複製などによる表現の“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注意が必要である。

今回の調査・分析から得られたBCCWJ使用上の教訓を最後に一般的な形にまとめておく。均衡性の考慮から明確な基準に基づいて収集された出版物（特に書籍）のデータに、それとは採録の手順が異なるだけでなく言語的に異質でデータの重複や信頼性の問題も大きいYahoo!ブログ、Yahoo!知恵袋のデータを単純に加えて使うのは、“均衡”コーパスのまっとうな用法ではない。中納言では検索対象――少納言ではメディア／ジャンル――を目的に応じて適切に指定したうえで検索しなければならない。そして付言すれば、話しことばの書き起こしであり、やはり異質性の高い国会会議録のデータも、“書き言葉”コーパスの他の部分との不用意な併用は避けるべきであろう。

文 献

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』138, pp.1-23.

丸山岳彦・柏野和佳子・田中牧郎(2011)「第３章　サンプリング」『「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版所収のPDF文書）, pp.21-38. 国立国語研究所コーパス開発センター.