「CD-毎日新聞データ集」に含まれるデータの特徴について

長谷川守寿（首都大学東京）

1.目的

本稿の目的は、言語研究の対象として「CD-毎日新聞データ集」に収録されている毎日新聞記事データの特徴を調べ、言語研究に使用する際に注意すべき点を述べることにある。

現在、代表性と均衡性を備えた日本語コーパスと呼べるのは『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contempolary Written Japanese： 以下 BCCWJ）（国立国語研究所(2011)）だけである。しかし、BCCWJ に含まれているデータには、書籍や白書などのように、校閲を経ているであろうデータもあれば、Yahoo!知恵袋や Yahoo!ブログなど、文の適格性という面においては、疑問が残るデータも含まれる。また、日本語教育を目的として、教えるに足る（ある程度）正しい日本語を調べるために BCCWJ を用いる場合、ジャンルなどを限定するしかない。しかし、BCCWJ に含まれる書籍データには、出版された年を基準にサンプリングを行っているため、福澤諭吉の『学問のすすめ』など、現代とは言い難いデータも含まれている(丸山 2011)。さらに、コーパスから得られた用例を確認するために原典を調べるのは、非常に困難である。

それに対し、新聞社の発行している記事データ集は、ある年全ての新聞記事を収録したものであるという設計上、代表性や均衡性は有していないが、記事は新聞社各社の校閲の過程を経ており、文の正しさという点では保証されている。また新聞の特性上、記事が執筆された時期と、それが発表された時期が近い。さらに、縮刷版にあたって用例を確認できることも大きなメリットである。

既存のコーパスを使用する場合の注意点として、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「データのクリーニング」(p.11)とし、「コンピュータを用いて、テキストを統計的に分析するには、テキストのクリーニングという作業が必要になる」としている。

本稿では、テキストのクリーニング作業を通じて、新聞記事データの特徴を明らかにする。

2.先行研究

「CD-毎日新聞データ集」について、毎日新聞社と発売元である㈱日外アソシエーツからは、以下のような情報が提供されているが、実際に何が含まれているのか、明示されていない。

(1)　毎日新聞の紙面に掲載された記事データにタグを付け１年ごとにＣＤ（もしくは、ＤＶＤ）に収録。
http://mainichi.jp/contents/edu/03.html

(2)　毎日新聞の東京・大阪本社の朝夕刊最終版を対象とした、毎日新聞 1991 年度以降の全文記事データ集（タグ付テキストデータ）です。
http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

松本(2003)では、以下のように紹介している。

(3)　毎日新聞東京・大阪本社発行の一年分の記事約一〇万件の全文を収録。社会面、解説面、経済面、国際面をはじめ、文化、家庭、総合、芸能、スポーツ面も収録。九一年版から毎年分が発売されている。

長谷川(2011)では、パイロットスタディとして 1997 年の「CD-毎日新聞データ集」と新聞縮刷版を比較する中で、言語データとして扱うには、改行記号が正しい箇所に付与されていないなどの問題があり、このデータを用いた調査を行う場合には、事前の修正が必要となることを指摘した。

3.方法

長谷川(2011)で指摘したように、「CD-毎日新聞データ集」には、(4)(5)のように、適切な位置に改行記号が付されていない問題がある。文が改行記号（以下<ENTER>で示す）によって分断されている場合、構文的な調査では問題となってくる。例えば(5)の場合、「ある自民党閣僚は」と「言った」は別々に処理され、このままでは「閣僚」を「言った」の動作主としては処理できない（なお記事を示す際、“M”は朝刊、“E”は夕刊、“略”は例文の一部省略を示す。また本稿で言及している部分に網掛けを行う）。

(4)　まゆちゃんの大好物だったキャンデーを供えると、「レッスン、<ENTER>　（2002/6/8E 略）

(5)　ある自民党閣僚は<ENTER>
「津波がやってくるようなものだよ」<ENTER>
と言った。<ENTER>　（1994/8/9M 略）

本研究では、上記の問題を発見する自作のプログラムを使用することにより、テキストのクリーニング作業を行う。その中で発見された問題点について、年代ごとに用例数の傾向を見る。

クリーニング作業では、読み込まれたデータ内で、開き括弧（“「”、“『”、“【”、“［”、“〈”、“《”、“＜”）

と、閉じ括弧（“」”、“』”、“】”、“］”、“〉”、“》”、“＞”）の数が一致しないもの、データの最後の文字が「、」で終わるものを検出するプログラムを作成し、問題の箇所を発見し修正する。なお、“（”と“）”は、1）のような表記があるので、対象から外す。

本研究では、修正していく中で発見したことを基に調査を行う探索的方法をとる。

4.対象

毎日新聞記事データは「CD-毎日新聞データ集 1994 年版」から「CD-毎日新聞データ集 2008 年版」までの 15 年分に含まれるものを扱う。「CD-毎日新聞データ集」は 1991 年から毎年継続的に発売されているが、1991 年版から 1993 年版までは扱わない。これは、使われているタグと収録の基準が異なり、単純に比較ができなくなるためである。例えば、(6)には著作権がないことを示すタグがつき、本文は収録されていないが、1994 年以降は［女の気持ち］というコーナーでは本文が収録されている。

(6)　［女の気持ち］新しい道　茨城県猿島郡・酒井
（1991/1/1M 略）

5.結果

読み込んだデータ内で対応する括弧の数が一致していない例、データが読点“、”で終了している例を、記事の特徴から示していく。

新聞記事データは、「＼タグ＼データ」の形をしており、記事本文を示すタグ(T2)を持つデータの総数と、(3)(4)のように記事本文中の括弧数に不一致が見られるデータの件数、記事本文が読点「、」で終わる数を表 1 に示す。

表1. 括弧数が不一致、読点で終わる記事本文数

年	データの総数	括弧数が不一致	文末が「、」
1994	583541	1463	661
1995	672425	1052	521
1996	738154	1899	721
1997	767130	2346	803
1998	801543	1773	623
1999	740676	1450	515
2000	735518	1372	556
2001	712124	463	404
2002	686125	328	562
2003	709076	240	545
2004	650820	164	444
2005	633399	177	289
2006	572330	117	330
2007	536220	95	214
2008	520753	136	270

「括弧数が不一致」数は延べで数え、(7)のような例は、“「”と“』”がそれぞれ不一致で 2 件と数える。

(7)　「愚者の船』（ＩＰＣ）。　（95/03/18M 略）

以下、括弧の数が一致しない記事や、文末が読点で終わる文を修正していく中で、明らかになった特徴を詳述する。

5.1.データの重複

開き括弧と閉じ括弧の数が一致しない（ほぼ）同じデータが続けて 2 度出現することが多く見られた。これは(9)と(8)のようにほぼ内容が同一のデータが続けて現れているものである。

(8)＼ T1 ＼［近聞遠見］高村正彦・元外相の「悪の枢軸」批判＝岩見隆夫　【大阪】
　＼ T2 ＼　「＜悪の枢軸＞は、ブッシュ大統領の立場から、テロへの強いメッセージを述べたもので、
<ENTER>
(9)＼ T1 ＼［近聞遠見］高村元外相の「悪の枢軸」批判＝岩見隆夫
　＼ T2 ＼　「＜悪の枢軸＞は、ブッシュ大統領の立場から、テロへの強いメッセージを述べたもので、
<ENTER>　（2002/2/23M 略。(8)も同日）

(9)は開き括弧（“「”）が一つだけで、対応する閉じ括弧がない例で、次の行以降と統合する必要がある。重複する記事(8)を確認すると、ID が異なり、記事見出しの文言が多少異なり、さらに記事見出し(T1)の末尾に【大阪】（大阪版の記事であることを示す）という文字列を持つかどうかの違いがある。【大阪】を持つ大阪版の記事は、多数の連載記事で見られたが、重複の出現に特徴が見られた「余録」（コラム欄）や「近聞遠見」（岩見隆夫氏が担当しているコラム）「近事片々」（社説）について言及する。各年での出現数の特徴を示したのが表 2 である。“重複”とは、すべて大阪版の記事数であり、1994 年は余録の記事数が 353 件、重複記事数が 0 件、1995 年は余録が 353 件で重複記事数が 1 であることを示す。表 2 から、2002 年と 2003 年の重複している記事数が突出して多いことが分かる。

記事のタイプによっても違いが見られる。「余録」は本文に違いはないが、見出しは(10)の「新年は晴れやかがいい…」と(11)の「今年は心をまるくして」のように、別のものになっている（下線は筆者）。

表2. データが重複している記事数

年	余録	重複	近聞遠見	重複	近事片々	重複
1994	353	0	51	1	295	0
1995	353	1	51	0	298	3
1996	368	14	51	2	313	17
1997	359	6	51	0	300	6
1998	356	3	51	0	295	1
1999	356	3	40	2	298	5
2000	357	3	50	0	297	5
2001	366	13	54	2	303	9
2002	710	354	104	52	587	294
2003	709	354	103	51	588	294
2004	440	84	61	10	364	70
2005	354	0	52	0	293	0
2006	295	0	53	2	295	0
2007	294	0	39	2	356	0
2008	294	0	52	0	356	0

(10)＼T1＼［余録］新年は晴れやかがいい…
　＼T2＼　新年は晴れやかがいい。上から読んでも下から読んでも２００２。　（2002/1/1M 略）
(11)＼T1＼［余録］今年は心をまるくして　【大阪】
　＼T2＼　新年は晴れやかがいい。上から読んでも下から読んでも２００２。

それに対して、「近聞遠見」は、見出しの違いは“【大阪】”の有無などわずかだが、(12)と(13)に示すように記事本文の複数箇所に違いが見られる（“＊＊＊”は、記事の中で同じ部分を表す）。この違いは、単に表記に関係する部分のみならず、テンス・アスペクト、モダリティが異なる部分もある。

(12)　＊＊＊＝岩見隆夫【大阪】＊＊＊胎中楠右衛門たいなかくすうえもん＊＊＊昭和十二（一九三七年）年＊＊＊投げうって＊＊＊過ぎている。＊＊＊類似しているようだ。『井戸べい』＊＊＊雑巾ぞうきん＊＊＊　（1994/5/3M 略。(13)も同日）
(13)　＊＊＊＊＊＊胎中楠右衛門（たいなかくすうえもん）＊＊＊昭和十二年（一九三七年）＊＊＊なげうって＊＊＊過ぎた。＊＊＊類似している。『井戸塀』＊＊＊雑巾（ぞうきん）＊＊＊

なお、「近事片々」については、ルビの有無の違い、空白 2 バイトスペースの違いがあるだけで、本文は全く同じであった。

なぜこの 2 年だけ大阪版の（ほぼ）同じ記事を収録したのかは不明であるが、言語研究の対象として処理を行う場合、これでは同じ文を 2 回扱うことになり、事前の対応や結果の検討が必要となってくる。

5.2.テキスト化について

括弧数の不一致を調べると、以下のように、不一致が連続して出現する例があった。

(14) 農家民宿で「どぶろ｜原料に地域性なく、
　く」提供（財務省）｜コスト回収が困難
　（2003/2/27M 略、(15)も同）

縮刷版で確認したところ、紙面の表に含まれる部分を(15)のようにそのままテキスト化したもので、年代を問わず見られた。これらを修正するには、研究の目的にあわせた修正方法を決める必要があろう。なお、紙面の表内のテキストをクリーニングするには、今回用いた括弧数の不一致や読点とは別の観点（例えば「——」の連続）が必要だろう。

(15) ■焦点の特区構想をめぐる調整状況■
　特区の内容　｜　各省庁の主張
　――――――――――――――
　農家民宿で「どぶろ｜原料に地域性なく、
　く」提供（財務省）｜コスト回収が困難

5.3.読点後に改行記号が入る場合について

文が読点「、」で終わる文には、以下のような記事が含まれ、文が正しく検出されなくなっている。

(16) ついで官邸の施設に触れ、<ENTER>
　「いたれり尽くせりで……」<ENTER>
　と言うと、それも、<ENTER>
　「とんでもない。ネズミも出る」<ENTER>
　と問答がかみ合わない。<ENTER>　（2000/9/5M 略）
(17) こうした流れのなかで、<ENTER>
　――兵<ENTER>
　は、やがて、<ENTER>
　――坂東武者<ENTER>
　となり、そして、<ENTER>
　――武士（御家人）<ENTER>
　となっていったのである。<ENTER>　（2000/1/9M）

(16)には 4 カ所改行記号が入力され、見かけ上は 5 行、(17)は 6 カ所改行記号が入力され見かけ上は 7 行になっているが、いずれも一文として処理されるべきものである。(16)は前出の「近聞遠見」の例であり、(17)は早坂暁著の連載小説『國難』の例である。「CD-毎日新聞データ集」では、著作権の問題から含まれないと考えられていた小説の一部が、実際には収録されていたのである。しかも小説毎に収録の実態には違いがあり、全く収録がないものから、かなりの回数が収録されているものもある（次節で詳述）。

5.4.連載小説

クリーニング作業の過程でデータに連載小説が含

まれていることを発見した。そこで連載小説の本文が CD-毎日新聞データ集に収録されているかを調べた。小説の連載期間は年によって区切られるものではないが、紙幅の都合で約 40 編の小説を年ごとに示したのが表 3 である。

表 3.小説の掲載回数

年	朝刊	夕刊	日曜版	年	朝刊	夕刊	日曜版
1994	276	0	12	2002	0	1	5
1995	0	0	52	2003	0	1	1
1996	0	1	27	2004	0	0	0
1997	2	1	50	2005	0	0	0
1998	2	8	50	2006	0	1	0
1999	2	3	50	2007	0	0	0
2000	1	3	51	2008	0	0	0
2001	1	5	50				

1994 年の朝刊には、林真理子著『素晴らしき家族旅行』が全 157 回連載中 136 回分、村松友視著『同僚の悪口』が全 179 回の連載中 140 回が収録されている。1995 年の朝刊では皆川博子著『花櫓』、辻邦生著『光の大地』が連載されているが、1 回も収録されていない。また、1998 年の日曜版には早坂暁著『國難』が連載小説全回分（50 回）が収録されている。このように新聞データには小説も含まれ、その収録状況にはかなりばらつきがあることが分かる。

さらに詳しくデータを見てみると、全回が収録されている『國難』は、（新聞社に）著作権のない記事にのみ付与される(18)のようなタグがない。つまり新聞社に著作権があり、CD にも収録されたものと思われる。

(18) ＼ ZZ ＼著作権無

(19) ＼ T2 ＼【現在著作権交渉中の為、本文は表示できません】

しかし、日曜版に連載された東野圭吾著『手紙』は、全 68 回中 31 回は(18)のタグがなく本文が収録されており、残り 37 回は(18)と(19)のタグがつき本文は収録されていない。収録上このような契約をしたと考えるのも不自然であり、何らかの原因で統一した処理が行われていないことが分かる。

このように連載小説では、一貫性の面で問題が多く見られた。連載であれば、データとしても同じように扱われているものと思ってしまいがちだが、作品によっては著作権に関する一貫性が保たれていないため、一部が CD に収録されていたりする。これをどう扱うかは、研究者が判断せねばならない。

5.5.その他の特徴

処理を進めていく中で、ルビが特に多く使用された記事(20)を発見した。2001 年にはこうした小学生新聞からの記事が 16 件収録されている。多くのルビが振られていると正しい形態素解析は不可能である。

(20) くるみさんのお店（みせ）に来（く）るのは、カッパや魔術師（まじゅつし）、木枯（こが）らしなど不思議（ふしぎ）なお客（きゃく）さんばかり。（2001/8/4E 略）

本研究では、探索的手法を用いているため、他にも問題はあるだろう。しかし事前に問題の予測を立てることが困難である以上、今後も様々な観点からデータをつぶさに確認することで、埋もれている問題を一つずつ見つけていくしかない。

6.おわりに

本稿では、1994 年から 2008 年までの「CD-毎日新聞データ集」を対象に、テキストのクリーニングを行う過程で明らかになったデータ集の特徴について述べてきた。なお、クリーニングを行うことによる効果については別の機会に行いたい。

いろいろ問題点を指摘してきたが、「CD-毎日新聞データ集」は、テキストファイルだけであり扱いが容易で、各新聞社の記事データの中では、一番安価なのも魅力である。毎年発売されているため、複数年のデータを用いれば、語数だけならば BCCWJ を超えることもでき、もっと見直されてもよいと思う。

伊藤(1995)は「電子化資料批判学の確立」の必要性をあげている。BCCWJ や新聞データ集など、電子化資料は様々な言語研究の対象として手軽に使われるようになった。それを用い言語表現の調査・分析を行う際には、分析の土台となる言語資料そのものの特徴を見極めていく作業も同時に必要であろう。

参考文献

伊藤雅光(1995)「音声データベースによる録音資料批判」『日本語学』7 月臨時増刊号、第 14 巻 8 号、明治書院

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』利用の手引き　第 1.0 版」、BCCWJ-DVD 版収録

長谷川守寿(2011)「新聞紙面と新聞記事データ集の相異について」『人文学報』443、首都大学東京人文科学研究科

松本裕治(2003)「現代語のコーパスの種類とそれぞれの特徴」『日本語学　4 月臨時増刊号コーパス言語学』、第 22 巻第 5 号、明治書院

丸山岳彦(2011)「大規模コーパスの利用とメタデータの役割」『第 1 回コーパス日本語ワークショップ予稿集』、国立国語研究所言語資源研究系・コーパス開発センター