首都大学東京都市教養学部人文・社会系
首都大学東京人文科学研究科
「人文学報」第473号
(日本語教育学)
2013年3月抜刷

「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

長谷川　守　寿

「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

長谷川守寿

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。

「規模や学術的信頼性の点で、日本語コーパスの代表格と呼べる」（李ほか 2012）のが『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese：以下 BCCWJ）である。これは厳密な設計を行うことで、代表性と均衡性が保証された、厳密な意味でのコーパスとなっている（国立国語研究所　2011）。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、また Yahoo!ブログにはデータが重複して収録されているという問題がある（田野村　2012）。

研究においては、目的に適合したデータを選択するのは当然である。日本語教育を目的として、日本語の実態を調べるのではなく、教えるに足る（ある程度）正しい日本語を調べる場合には、それに見合ったデータが必要となる。BCCWJ を用いる場合、研究目的によっては、ジャンルなどを限定することによって、正しさの担保されたデータを扱うことができるようになる。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）。初出年などを確認し、目的に合致するデータ

に制限するには、検索後の処理に大きな時間が割かれてしまうことになる。

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、コーパスに求められる要件である、言語資料としての代表性や均衡性は有していない。しかし、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（それでも含まれる誤植には目をつぶらざるを得ない）。また、非常にまれに古い作品を引用の形で含むこともあるが、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。

筆者は、複文の実態や、語彙のコロケーションに興味を持って研究を続けている。そこで対象となるのは、現代日本語であり、しかも正しいとされる文である。このような理由から、言語データとして新聞記事データ集を対象とし、さらに1995年より発売されており、比較的入手が容易な「CD-毎日新聞データ集」を扱っている。

既存のコーパスを使用する場合の注意点として、伊藤（2003）では「コーパスの内容をよく読むこと」を挙げ、「原本が存在しているコーパスの場合は、原本との校訂作業は欠かせない」としている。また金（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。

長谷川（2011）では、1997年の『毎日新聞縮刷版』（毎日新聞社）と「CD-毎日新聞データ集」との照合という形で原本との校訂作業を行った。本稿では、テキストのクリーニング作業を通じて、「コーパスの内容をよく読」み、新聞記事データの特徴を明らかにし、言語研究の資料として使用する際の注意点を述べる（なお、他の新聞社、朝日新聞社、読売新聞社、日本経済新聞社が発行している新聞記事データ集に関しては、別個の検討が必要となるが、それは著者の能力の限度を超えている）。

2．先行研究

BCCWJはどのような方針の下、どのような手順で作成されたか明示されている唯一の均衡コーパスである（国立国語研究所　2011）。それに対し、「CD-毎日新聞データ集」では、どのように作られたか知る術がない。例えば松本（2003）では、以下のように紹介しているが、何をどのように収録しているのか、明らかではない。

> **毎日新聞記事データ**
>
> 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。市販のCD-毎日新聞とは別に、テキストデータの形でCD-ROMに収録されているので、特別なデータ取り出しのためのソフトが不要。毎日新聞社とデータ使用許諾に関する覚書を交わすことで、研究目的に利用できる。

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。

本稿では、複文やある語のコロケーションを調べる際、より正確な結果が得られるように「CD-毎日新聞データ集」を修正していく中で明らかになった毎日新聞記事データの特徴を述べ、これらを日本語研究に使用する際の留意点を提供する。

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、（1）（2）のように、適切な位置に改行記号が付されていないという問題がある。文字データを処理する際は、データの始めから改行記号（エディタなどでは↵と

示される。以下〈ENTER〉で示す）までが読み込まれるが、読み込まれたデータの適切な場所に改行記号が付されていない場合、構文的な調査では問題となってくる（文字や語彙に関する調査ならば影響が少ない）。例えば（2）の場合、「ある自民党閣僚は」と「言った」は別々に処理されることになってしまい、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。

（1） 椿山荘では料亭「錦水」（略）満喫できる『和食レストラン〈ENTER〉花車』を営業。昼（１２～１４時）点心（略）　（1997年12月25日夕刊）

（2） （略）ある自民党閣僚は〈ENTER〉
「チリ沖地震があって、（略）津波がやってくるようなものだよ」〈ENTER〉と言った。自社連立政権は（略）　（1994年８月９日朝刊）

本研究では、上記のような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、年ごとに同じような用例がどのくらいあるのか傾向を見る。

クリーニング作業では、まず（1）（2）のような形で出現することを考慮し、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する。なお、“(” と “)” は、1）のような表記があるので、対象から外す。

プログラムを実行すると（3）（4）のような出力が得られるので（網掛けは筆者。以下同）、その結果を基に改行位置の修正を行っていく。

（3） ◆ピアノを指導していた若林宏子さん（５５） 今も、お姉ちゃんのレッスンで家に伺い、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年６月８日夕刊）

（4） 武蔵丸は優勝決定戦への権利を勇み足という悔やんでも悔やみ切れない結果で逸しただけに▒軍配？　分かんなかった。勇み足？〈ENTER〉

（1994年４月16日朝刊）

本稿では、実際に記事本文を確認しながら修正していく上で発見したことを基に調査を行う探索的方法をとる。なお、「毎日新聞記事データの使用許諾覚書」[1]に従い、CD全体を「CD-毎日新聞データ集」、それに含まれるデータを「毎日新聞記事データ」とし、個別のCDを「CD-毎日新聞データ集1994年版」等とする。

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2003年版」までの10年分に含まれるものを扱う。「CD-毎日新聞データ集」は（株）日外アソシエーツより1991年版以降が発売されているが（http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html、2012年12月14日　取　得）、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るからである。例えば、記事の見出しを示すタグは、1991年から1993年まではＴＩＮ（1994年以降はＴ１）、本文を示すタグはＨＯＮ（1994年以降はＴ２）と、それ以降とは異なるタグのセットを用いているため、処理が煩雑になる。しかし、それ以上に収録の基準が異なっていることが大きい。例えば、（５）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。

（５）［女の気持ち］新しい道　茨城県猿島郡・酒井＊＊　主婦・４３歳
（1991年１月１日朝刊、人名の一部は伏せ字に変更）

他にも特集記事は、記事により収録されていたりいなかったりと統一されていない。これでは、1994年以降と単純に比較ができなくなるため、本研究では、

[1] http://www.nichigai.co.jp/sales/mai_data/contract.pdf より。2012年12月８日取得

1991年から1993年のデータは対象外とする。

なお、本調査の結果が2003年以降にも同じ傾向が見られるのか確認する目的で、一部の項目については2004年から2007年のデータも使用した。

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であるＴ２を持つデータ[2]の総数（記事本文総数）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（括弧数が不一致）、記事本文が読点「、」で終わる数（データ末が「、」）を表１（次頁）に示す。「括弧数が不一致」の件数は延べで数え、（６）のような例は、“「”と“」”の不一致で２件と数える。

（６）　＼Ｔ２＼（略）愚者の船（ＩＰＣ）（略）。　（1995年３月18日朝刊）

括弧数の不一致については、1997年をピークに減少傾向にあるが、これらの修正を通して、新聞データに含まれる内容について明らかにし、言語資料として使用する際の注意点に言及する。

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（７）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。

---

[2] http : //www.nichigai.co.jp/sales/pdf/man_mai.pdf（2012年12月９日取得）ではＴ２のデータ内容を「記事全文」としているが、「全文」という名称は誤解しやすいので、本稿では記事本文とした。また長谷川（2011）に従い、著作権交渉中のためＴ２が表示されていないものを除外した数であり、記事の収録件数とは異なる。

表1　括弧数が不一致の記事本文数と読点で終わる記事本文数

年	記事本文総数	括弧数が不一致	データ末が「、」
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

（7）　＼ＩＤ＼００１４９７９０

＼Ｔ１＼［近聞遠見］高村元外相の「悪の枢軸」批判＝岩見隆夫

（中略）

＼Ｔ２＼　だが、まったくなかったわけではなく、イランと縁の深い高村正彦元外相が、首脳会談翌日の１９日、自民党高村派の総会で、次のように述べている。

＼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、それなりにわかるが、日本には日本の立場がある。〈ENTER〉　　（2002年２月23日朝刊。（８）も同日）

（8）　＼ＩＤ＼００１５１６１０

＼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＼　〈枢軸〉

＼Ｔ２＼　だが、まったくなかったわけではなく、イランと縁の深い、高村正彦元外相が、首脳会談翌日の１９日、自民党高村派の総会で、次のように述べている。

＼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、それなりにわかるが、日本には日本の立場がある。〈ENTER〉

（７）は開き括弧（“「”）が一つだけで、対応する閉じ括弧（“」”）が文頭から改行記号までの間に見られない例であり、次の行以降と統合する必要がある文である。重複する記事（８）についてデータを確認すると、ＩＤ（数字８桁でユニークな情報）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】という文字列を持つかどうかの違いがあることが分かった。【大阪】は、大阪版の記事であることを示している。“【大阪】”を持つ大阪版の記事は、多数の連載記事で見られた。ここでは記事本文において重複の出現に特徴が見られた「余録」や「近聞遠見」「近事片々」について言及する[3]。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、大阪版の記事数であり、1994年は「余録」で大阪版の記事が０件、1995年は「余録」

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	709	353	104	52	587	294
2003	709	354	103	51	588	294

[3]「余録」は毎日新聞のコラム欄のことで、朝日新聞でいうと天声人語、読売新聞で編集手帳に類するものである。「近聞遠見」は岩見隆夫氏が週１回連載しているコラムの名称で、「近事片々」は社説に相当するものである。

で大阪版の記事数が1であることを示す。

表2を見ると明らかなように、2002年と2003年の大阪版の記事数が突出して多いことが分かる。ちなみに2005年は全て0件であり、この2年だけの特徴であることが分かる。

記事のタイプによっても違いが見られる。「余録」の場合、記事本文に違いはないが、記事見出しは（9）の「新年は晴れやかがいい…」と（10）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ことが分かった（下線は筆者）。

（9） ＼ＩＤ＼００００１６３０
＼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。０１年はテ（以下略）
（2002年1月1日朝刊。（10）も同日）

（10） ＼ＩＤ＼００００２７１０
＼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。０１年はテ（以下略）

それに対して、「近聞遠見」は、見出しの違いは“【大阪】”と著者名の有無であるが、（11）と（12）に示すように記事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（～ている／た）、モダリティ（ようだ／φ）が異なる部分もある。ちなみに、本文は（11）が1598文字、（12）が1584文字（記号含む）といくらかサイズも異なっている。

（11） ＊＊＊＝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＊＊＊腹げうって＊＊＊過ぎている。＊

＊＊類似しているようだ。『井戸べい』＊＊＊雑巾ぞうきん＊＊＊馬鹿にして＊＊＊　（1994年5月3日朝刊。(12) も同日）

(12)　＊＊＊＊＊＊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。『井戸端』＊＊＊雑巾(ぞうきん)＊＊＊馬鹿(ばか)にして＊＊＊

なお、「近事片々」については、記事見出しが「近事片々」だけということもあり、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。しかし、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。

他にも、大阪版の記事は「訃報」「ことば」「雑記帳」「憂楽帳」「社告」などがあるが、年によって重複数が大きく跳ね上がるなどの特徴が見られたのは、上記の3つである。他の記事については、重複が見られない「ことば」や、内容によって重複が少ない数であるが見られた「訃報」などがあるが、特徴として抽出することはできなかった。

なお、長谷川（2011）では、新聞紙面とCDデータの比較を行った際、大阪版のデータは紙面の入手が困難で検証できないため対象外としたが、大阪版では（11）のように、ルビをつける際に漢字の後に“(”“)”を囲まないなど問題があるため、長谷川（2011）同様に上記のような大阪版の記事を削除するのも一案ではないかと思う。

5.2. 誤入力

括弧数の不一致を調べた際に、以下のようなデータも出力された。

(13)　告生きる〈ENTER〉　（1999年8月6日朝刊）

データを確認したところ、これは（14）のように、記号を含め記事見出しの3文字以降（この場合「告」以降）が記事本文になっている例であることがわかった。

(14) ＼T 1＼［社告］生きる〈ENTER〉
＼T 2＼告］生きる〈ENTER〉 （1999年8月6日朝刊）

このような例がどのくらい生じているのか確認したところ、表3のような結果となった。おそらく誤入力と思われるが、最大でも39とそれほど大きな数値ではないので、無理なくクリーニングできる。なお翌年の2004年は5件、2005年から2007年は0件であった。

表3 見出しと本文に重複があるデータ数

年	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
出現数	0	0	0	0	16	30	21	9	39	30

5.3. テキスト化について

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

(15) 3月 農水省の第三者機関▒ノリ不作〈ENTER〉
(16) 等調査検討委員会▒が初会合〈ENTER〉（2002年4月16日朝刊）

これは、内容や形式、改行位置から、記事中の年表をそのままテキスト化したものと思われる。これも大阪版の記事だったため、縮刷版で確認できなかったが、記事本文の中には、紙面の表に含まれる部分を（17）のようにそのままテキスト化したものが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、今回の調査で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。

(17)
＼T2＼■焦点の特区構想をめぐる調整状況■〈ENTER〉
＼T2＼ 特区の内容 | 各省庁の主張 | 特区推進室の反論 〈ENTER〉
＼T2＼――――――――――――――――――――――――――――――――〈ENTER〉
＼T2＼ (中略) 〈ENTER〉
＼T2＼外国人医師による | 患者の生命・身体の | 医療の発展に必要な医師は 〈ENTER〉
＼T2＼医療行為（同） | 危険を伴う | 厚労相同意で認めるべきだ 〈ENTER〉
＼T2＼――――――――――――――――――――――――――――――――〈ENTER〉
＼T2＼農家民宿で「どぶろ | 原料に地域性なく、 | 地域独自の酒類製造を認め 〈ENTER〉
＼T2＼く」提供（財務省） | コスト回収が困難 | るべきだ 〈ENTER〉
＼T2＼―――――――――――― (2003年2月27日朝刊、一部改)

5.4. 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。

(18) 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。建物は古いし、（中略）。ネズミも出る」〈ENTER〉
と問答がかみ合わない。 (2000年9月5日朝刊)

(19) 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。　　（2000年1月9日朝刊）

長谷川（2011）では、予測できない場所に改行情報が打ち込まれていることを指摘した。これらは縦書きの小説に見られる従属節や引用文、“——”を含む文言の前後等で改行されており、その意味では改行は予測できる。(18) には4カ所改行記号が入力され、見かけ上は5行、(19) 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、いずれも本来は一文である。

(18) は前出の「近聞遠見」の例であり、(19) は早坂暁著の連載小説『國難』の例である。実はこれまでには言及のないことだが、「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かなり違いがあり、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。これに関して、次節で詳述する。

5.5. 連載小説

前節では、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。縮刷版の索引では「続きもの」とされるため、どの小説も一貫して同じようにタグ付けがされていると予想される。これについて、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ものが、表4である。

例えば、「CD-毎日新聞データ集1994年版」の当初から見ていくと、林真理子著『素晴らしき家族旅行』（1994年1月1日から1994年6月28日）は、全157回連載されているが、そのうちCDに収録されているのは136回分である。これは、新聞紙面には1月1日から掲載されているが、CDの収録は1月14日からで、未収録の連載があるためである（これについては考察で再度言及する）。皆川博子著『花櫓』(1995.1)、辻邦生著『光の大地』(1995.10) などは全く収録されておらず、芦原すなお著『さんじらこ』(1997.4)、宮本輝著『草原の椅

子』(1997.12) 等は、1・2回収録されている。これに対し早坂暁著『國難』(1998.10) や内田康夫著『箸墓幻想』(2000.4) 等は、連載小説全回分が収録されている。このように新聞データには小説も含まれること、しかも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『箸墓幻想』は、(新聞社に) 著作権のない記事にのみ付与される (20) のようなタグがない。このことは新聞社に著作権があることを意味しており、そのためCDにも当然収録されたものと思われる。一方、『花櫓』は全264回連載されているが、CDには全く収録されていない。そこでは全ての回に (20) (21) が付与されているので、形式としては統一されている (なお2001年・2003年のみ“＼Ｔ２＼【著作権交渉中のため本文を表示できません】”であり、表現が多少異なる)。

(20) ＼ＺＺ＼著作権無
(21) ＼Ｔ２＼【現在著作権交渉中の為、本文は表示できません】

しかし、例えば「日曜くらぶ」に連載された東野圭吾著『手紙』(2001.7) は、全68回中、31回は (20) のタグがなく本文が収録されており、残り37回は (20) と (21) 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。

なお、小説に関しては、表記の揺れも確認された。例えば、作品名では『木漏れ日の坂』と『木洩れ日の坂』の二通りの表記が見られ、作者名では「村松友視」「村松友＝」(本来の表記は“視”であり、S-JIS外のため“＝”となる) がある。また、統一がなされていないという点では、著者名が「白石一郎」と「白井一郎」(誤り) になっているということもあった。

このように一貫性の面で、問題が多く見られたのが連載小説である。連載であれば、データとしても同じ扱いをさ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、作品がCDに収

表4　データ内の連載小説（朝・夕刊・日曜の別、連載順）

タイトル	作　者	収録回数／掲載回数
〈朝刊〉		
素晴らしき家族旅行	林真理子	136／157
同僚の悪口	村松友視＊	140／179
花櫓	皆川博子	0／264
光の大地	辻邦生	0／160
風の行方	佐藤愛子	0／371
さんじらこ	芦原すなお	2／237
草原の椅子	宮本輝	2／381
天球は翔ける	陳舜臣	2／336
すべて辛抱	半村良	1／371
黒龍の柩	北方謙三	1／470
香乱記	宮城谷昌光	0／525
哀歌	曽野綾子	0／389
〈夕刊〉		
忘れられた帝国	島田雅彦	0／174
かくも短き眠り	船戸与一	0／318
惜別の海	澤田ふじ子	0／509
ボクの町	乃南アサ	1／217
本能寺	池宮彰一郎	10／301
新宿鮫　風化水脈	大沢在昌	3／344
木漏れ日の坂＊	北原亞以子	4／320
大黒屋光太夫	吉村昭	3／320
墓石の伝説	逢坂剛	1／341
〈日曜くらぶ他〉		
橋本治	三日月物語	75／75
津本陽	真田忍侠記	4／94
白石一郎＊	怒涛のごとく＊	100／100
早坂暁	國難	74／74
内田康夫	箸墓幻想	63／63
東野圭吾	手紙	31／68
志水辰夫	ラストドリーム	1／70

（＊は複数の表記があることを示す）

録されていたり、いなかったりする。これをどう扱うかは、研究者が判断せねばならない。しかしながら、仮に（20）のタグを持つデータを研究の対象外とした場合、通信社であるAP通信等からの記事も対象外とせねばならなくなる。そうなると、対象となるデータが新聞社独自の記事のみとなってしまい、新聞紙面から大きく異なるため、慎重な対応が必要となる。

5.6. その他の特徴

以上、データのクリーニング作業を通して、「CD-毎日新聞データ集」に含まれる毎日新聞記事データの特徴と言語研究に使用する際の注意点について述べてきた。ここでは、今までに言及できなかった点について述べる。

まず「CD-毎日新聞データ集」のうち、1994年については、5.5.の「連載小説」で言及したように未収録の記事が存在するという問題がある。1994年の縮刷版では、城山一郎著『もう、きみには頼まない』や夕刊連載小説である笹倉明著『砂漠の岸に咲け』、日曜くらぶの連載小説である小池一夫著『夢源氏剣祭文』があるが、「CD-毎日新聞データ集」には全く情報がなかった。タイトルのみ掲載し、本文には「著作権交渉中」とする形式がこれ以降の連載小説には見られるが、これら上記の小説には見られず、何をどのように収録するか、統一ができていなかったものと推測される。不統一が多く見られる以上、1994年のデータは対象外とするのも一つの方法と考えられる。

本稿では、探索的手法を用いている。そのため、まだ見つかっていない問題もあるかもしれない。例えば、以下の例は、ある接尾辞を持つ語の用例を探している時に偶然発見した記事である。（22）と（23）は、ＩＤは異なるが記事見出し、記事全文が全く同じものである。つまり、言語情報としては完全に同一のものがデータに重複して含まれているのである。今回それを指摘する手がかりとなった大阪版の記事でもない。このように、今までの網にかからないでいる問題がどこかに残されている可能性がある。

（22）　＼ＩＤ＼００７４４５３０

＼Ｔ１＼［ヒット・ナウ］ポップス　命燃える熱情が聞こえてくる
＼Ｔ２＼　大人の人生を感じさせるようになったらポピュラー音楽も本物である。勢いと情熱だけで押しまくる若さもすがすがしいが、高い山や広い川を越えた後のちょっとした疲労感を漂わせる音楽というのは
（以下略）　（1994年９月26日夕刊。（23）も同日）

(23)　＼ＩＤ＼００７４４６００
＼Ｔ１＼［ヒット・ナウ］ポップス　命燃える熱情が聞こえてくる
＼Ｔ２＼　大人の人生を感じさせるようになったらポピュラー音楽も本物である。勢いと情熱だけで押しまくる若さもすがすがしいが、高い山や広い川を越えた後のちょっとした疲労感を漂わせる音楽というのは
（以下略）

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見つけていくしかない。

また、長谷川（2011）では、形態素解析の際に支障となるので、ルビの削除を提案しているが、今回ルビ削除の処理の中で、ルビが特に多く使用された記事（24）を発見した。これは小学生新聞からの転載のためで、2001年には小学生新聞からの記事が16件収録されている。このように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。こうした文も研究対象とするためには、ルビを削除するということの必要性が今回再度確認された。

(24)　＼Ｔ１＼［読も読もブックランド］新しく出た本　つるばら村の三日月屋さん　【大阪】
＼Ｔ２＼◇パン屋に来る不思議な客たち　（中略）
＼Ｔ２＼　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。やさしい気持（きも）ちになれる夢（ゆめ）いっぱいのお話（はなし）です。

＼Ｔ２＼※「読も読もＢｏｏＫランド」は一部毎日小学生新聞から転載。
（2001年８月４日夕刊）

6．おわりに

本稿では、1994年から2003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と、言語研究として使用する際の注意点について述べてきた。年によって予測できない特徴を含んでいることもあるので、さらに2004年からの言語データを使用するには、当然テキストのクリーニングを行わなければならないであろう。

「CD-毎日新聞データ集」は、文字符号化方式にUTF-8（BOM無）を採用したDVD版BCCWJに比べれば、文字コードがS-JISで特に変換の必要がないこと、テキストファイルだけなのでXMLファイル形式であるDVD版BCCWJよりも扱いが楽なことなど、文系の研究者にとっては利点も多い。各新聞社の記事データの中では、比較的安価なのも魅力である。除外するデータについて言及したが、毎年発売されているため、複数年のデータを使用すれば、語数だけならばBCCWJを超えることもできる。ただ今後も「CD-毎日新聞データ集」の作成の過程が新聞社から明らかにされることはないであろうし、期待するべきでもないであろう。また、BCCWJのオンライン版などとは違い、将来データが更新されることもない。縮刷版との対照などは、むろんデータを扱う側の責任であろう。

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「CD-毎日新聞データ集」など、電子化資料が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が、その一方で言語研究における対象としてのデータそのものの質を問うような「電子化資料批判学」が盛んになっているとは言えない。

言語表現の調査・分析も必要であるが、分析の土台となる言語資料の特徴を見極めていく作業も同時に行っていかなくてはならない。今後もコーパスを言語研究に有用な言語資料として用いていくために何が必要かということを追求

していきたい。

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月号臨時増刊号　パソコンを使う日本語研究』、第14巻第８号、明治書院、127－143.

――――（2003）「コーパスと統計」『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、26－35.

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

田野村忠温（2012）「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――」『第２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、265－274.

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、20－45.

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、54－60.

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、203－210.

李在鎬・石川慎一郎・砂川有里子（2012）『日本語教育のためのコーパス調査入門』、くろしお出版

用例出典

「CD-毎日新聞データ集1994年版」～「CD-毎日新聞データ集2003年版」（日外アソシエーツ）

本稿で用いた「CD-毎日新聞データ集」は、筆者が毎日新聞社と交わした利用許諾契約・覚え書きに基づき、使用した。