首都大学東京都市教養学部人文・社会系
東京都立大学人文学部
「人文学報」第443号
(日本語教育学)
2011年3月抜刷

新聞紙面と新聞記事データ集の相違について

長谷川　守　寿

新聞紙面と新聞記事データ集の相違について

長谷川守寿

1．目的

言語研究のデータとして、近年コーパスが選択されることが多くなっている。市販のコーパスでは、(現在は入手不可であるが)『CD-ROM 版　新潮文庫の100冊』が文法研究の用例抽出によく使われてきた。

また、新聞社は従来から検索ソフトを組みこんだ CD-ANSWER 版[1]を発売していたが、全ての情報にアクセスできる言語研究用の記事データ集の発売を開始したのは、毎日新聞・日経新聞が最初で、読売新聞、朝日新聞なども現在では同様に発行している（松本（2003））。大量のデータを収録し、一般に頻度の低い用例も検出でき、また入手しやすさなどもあり、新聞コーパスから用例を採集する研究も多く行われるようになってきている。

本研究の目的は、このように多用されるようになってきた新聞の記事データ集は、そもそも言語資料として新聞紙面にどれほど忠実なのか、明らかにすることである。当然、新聞紙面と新聞記事データ集では、紙と CD というようにメディアが異なり、さらに写真・広告・文字装飾・段組の有無も異なる。しかし、新聞紙面と新聞記事データ集の間にはどのような違いが見られ、それが文法研究などにどのような影響を与えるか、考察することで新聞記事データ集の適切な使用法が明らかになると思われる。

---

[1] 検索ソフトウエア（CD　Answer）というソフトからのみ検索が可能となるもので、タイトルまたは記事をキーワード（主に名詞）で検索することが可能である。

2．先行研究

新聞紙面と新聞記事データ集を比較した研究には、横山他（1998）がある。横山他（1998）は、朝日新聞CD-ROMと縮刷版を比較し、使用されている文字（特に漢字）の観点から、調査を行ったものである。朝日新聞CD-ROMのテキストデータと新聞の縮刷版の比較から、「見出し部に違いが多く発見された」と述べている（pp. 17－19）。

電子化されたコーパスと元のデータの比較という面では、伊藤（1995）、松田他（2008）が挙げられる。伊藤（1995）では、音声データベースを使い、『平家物語』に曲節をつけて琵琶の伴奏で語るものである『平曲』の録音資料批判を行い、アクセントのゆれや変化を確認した。また、その原因について考察が加えられ、電子化資料批判の必要性を述べている。また、松田他（2008）は、国会議事録（http : //kokkai.ndl.go.jp/）を対象とした研究である。松田他（2008）では、録音資料と国会会議録の対照分析を行い、地方議会の整文規準19項目から、読点を除いた18項目について、調査を行った。その結果、18項目には当てはまらない違いの存在も明らかにし、整文のゆれの指摘と、その要因についての検討を行っていることから、国会会議録データの資料批判といえる。

本調査は、新聞記事データ集の資料批判として、新聞紙面と新聞記事データを比較する。その際、横山他（1998）で指摘された点について、多くの違いが発見されるのは見出しだけで、記事本文にはないのか、あるとすれば、どのような違いなのか、これらの点について明らかにする。

3．方法

3.1．使用データ

本調査では、新聞紙面と新聞記事データ集の検証において、『毎日新聞縮刷版1997年』（毎日新聞社）と『CD-毎日新聞'97データ集』（日外アソシエーツ）を使用する。

『CD-毎日新聞データ集』を選択した理由は、いわゆる三大紙といわれる朝日・読売・毎日の新聞記事データ集の中で最も安価であり、入手が容易なため

```
＼ＩＤ＼００００００３０
＼ＣＯ＼９７０１０１００３
＼ＡＤ＼０１
＼ＡＥ＼Ｎ
＼ＡＦ＼９７０１０１Ｍ０１
＼Ｔ１＼［社告］１月２日の新聞、休みます
＼Ｓ１＼　　’９７．１．１　朝刊　１頁　写図無　（全３０４文字）
＼Ｓ２＼　きょう元日は新聞休刊日で２日の朝刊は休ませていただきます。夕刊は４日（土）から発行します。
＼Ｔ２＼　きょう元日は新聞休刊日で２日の朝刊は休ませていただきます。夕刊は４日（土）から発行します。
＼Ｔ２＼　なお今年の新聞休刊日は、２月９日、３月９日、４月１３日、５月５日、６月１５日、７月１３日、８月１０日、９月１５日、１０月１２日、１１月１６日、１２月１４日の予定です。各休刊日翌日の朝刊は休ませていただきます。ご了承ください。
＼Ｔ２＼×　　×　　×
＼Ｔ２＼　毎日新聞社は休刊日もニュース速報をパソコン通信の「ニフティサーブ」や「ＢＩＧＬＯＢＥ（ＰＣ—ＶＡＮ）」など１６ネットワーク、インターネット上の電子新聞「毎日デイリークリック」、さらにＴＢＳ系テレビや全国約２１カ所のＣＡＴＶ文字放送などでお届けしています。　毎日新聞社
```

図１　「CD–毎日新聞'97データ集」のサンプル

である。『CD–毎日新聞データ集』本社版は、「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～2009年の全文記事データ集（タグ付テキストデータ）」である[2]。タグには、図１に示すように、ＩＤ（データＩＤ）、ＣＯ（索引記事番号）、ＡＤ（掲載面種別コード）、ＡＦ（掲載日付）、Ｔ１（見出し）、Ｔ２（記事本文）等が用意され、タグの後にテキストが追加されている（納品データ仕様書（本社版）[3]より）。なお1997年を選択した理由は、筆者が持つ最新のデータが2007年であり、10年後のデータでも同様の傾向が見られるか検証するためである。以後、Ｔ１のタグがついているものを「見出し」、Ｔ２のタグがついているものを「記事本文」とする。なお、文字コードはSJISである。

3.2.『CD-毎日新聞データ集』について

調査の対象とする記事は、新聞記事データ集にも収録されている必要がある。新聞記事データ集は、「最終版を対象とした」ものであるが、新聞紙面の情報

[2] http : //www.nichigai.co.jp/sales/mainichi/mainichi-data.html。以下 url は全て2010年11月23日に取得した。

[3] http : //www.nichigai.co.jp/sales/pdf/man_mai 2008.pdf

が全て収録されているわけではない。

まず、データがテキストファイルのみということで、写真などの画像ファイルが収録されていないのは当然であるが、連載小説、四コマ漫画、風刺漫画なども著作権の問題で収録されていない。また、同様の理由で収録されていない記事が存在する。例えば著名な作家、歌手、研究家などの著名入りの記事であるが、これらの記事本文も収録されていない。見出しを掲載すると、（1）のようなもので、“★”がはじめに付いているものは、見出しだけで記事本文は（2）のようになっていて、個別の記事そのものは収録されていない。

（1） ★［食卓の一品］長芋豆腐そぼろあん＝料理研究家・前田和子
（1月5日朝刊 CD）

（2） 【現在著作権交渉中の為、本文は表示できません】

1997年の場合、総記事数は119,836[4]であるが、（2）のように見出しはあるが記事本文が含まれていないものが8,561記事あった。本調査ではこれらのデータを除外して調査する。なお、語彙調査などでは、これらを除外しないと、8,561回余計にカウントされるため、正確な数値が得られなくなるおそれがある。また、著作権の問題で収録されていない本文記事があることについては、利用許諾契約書[5]、注文書[6]、納品データ仕様書（本社版）にも記述がない。言語研究のためには、重要な情報と思われるので、明記が必要ではないかと考える。

3.3. 手順

まず、新聞記事データ集の中で記事本文が収録されていて、しかも大阪版以外の記事から1,000の記事を無作為抽出し、見出しと記事本文を、新聞紙面と

[4] http : //www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[5] http : //www.nichigai.co.jp/sales/mai_data/order.pdf
[6] http : //www.nichigai.co.jp/sales/pdf/man_mai 2008.pdf

比較する。大阪版を対象外としたのは、対照の際に使用する『毎日新聞縮刷版』（毎日新聞社）が本社版であるため、大阪の記事は掲載されておらず、比較が出来ないためである。比較には『毎日新聞縮刷版1997年１月』から『毎日新聞縮刷版1997年12月』までの当該部分との目視による照合作業を行う。

新聞紙面と新聞記事データ集の異同については、国会会議録のように、準拠すべき基準が見あたらない。そこで、2008年度、首都大学東京大学院人文科学研究科の日本語教育学教室で開講した「日本語教育学研究」で、100記事を対象に予備調査を行った[7]。その結果、横山他（1998）と同様に、見出しには新聞紙面と新聞記事データ集で異なる部分が多く見られたが、記事本文でも異なる種類の違いも見られ、また見出しと記事本文で共通する違いも見られた。そこで異なる部分を、見出し・記事本文に共通のもの、見出しに主に見られるもの、記事本文に主に見られるものという観点から述べることとする。

さらに、比較の結果、異なることが確認され、数値的に比較可能なものについて、1,000記事中でどのくらい起こるか、調査する。なお、1,000データでは出現数が少ないものに関しては、全数調査を行う。

最後に、『CD-毎日新聞2007データ集』から100記事、無作為に抽出し、これらのデータと縮刷版の照合を行い、データの作成に関する変化を調べる。また、『CD-毎日新聞'97データ集』で見られた結果が観察されるのか確認する。

このような手順を通して、新聞紙面と新聞記事データ集ではどのような違いがあり、新聞紙面と新聞記事データを同一の内容にするには、どのような修正が必要となるのか、明らかにする。これは今後新聞記事データ集を言語研究の対象として使用する際の指針となると思われる。

4．結果

新聞紙面と新聞記事データの違いを、前述の通り「見出し、記事本文共通に見られる違い」「見出しに多く見られる違い」「記事本文に多く見られる違い」

---

[7] 受講者であった柴田沙矢香さん、鄭明鎬さん、渡邉千佳子さんに感謝する。

の三点から述べる。新聞の最終版を元に新聞記事データ集が作成されていることから、新聞紙面・該当する新聞記事データの順に示し、また“紙面”“CDデータ”という略称を用いる。

4.1. 見出し・記事本文に見られる違い

見出し・記事本文とともに見られた違いとして、まず算用数字が挙げられる。算用数字は、紙面では（3）のように1バイト文字であるが、CDデータでは（4）のように2バイト文字で表現されている。紙面も数字が1文字の場合、もともと2バイト文字を使用するので、これは2文字以上の場合に発生する。これが1カ所でも見られたサンプルデータは、1,000件中274件あった。具体的に数値や日付が出るのは記事本文であるため、この違いは記事本文で多く見られたが、見出しでも数字の部分はこのような違いが見られた。

なお、新聞紙面は縦書きであるが、横書きで再現したり、紙面を画像ファイルで示す。これは状況により判断した。以下、紙面とCDデータ両方を示す場合には、新聞紙面の例文の後に月・日・朝刊夕刊の別・掲載面を示す。CDデータのみ示す場合には月・日・朝刊夕刊の別を示し、最後にCDと付す。“年”が付されていないのは全て1997年である。また、新聞記事には個人・団体名が掲載されているが、本論文では人名の一部を“＊”に替える処理を適宜行い、個人・団体が特定できないようにする。

（3） ドキュメント　リマ24時　（2月12日朝刊7面）

（4） ［ドキュメント］リマ２４時（１０日＝現地時間）　ペルー日本大使公邸占拠事件

形態素解析は、解析器と辞書を別々に用意するのが通例である。解析器に茶筌[8]、辞書に「IPA品詞体系辞書」（以後「IPA辞書」と呼ぶ）を使用して形態

[8] http : //chasen-legacy.sourceforge.jp

素解析を行い、語数を数えた場合、算用数字は2バイト文字でも1バイト文字でも、それぞれ1文字1語と数えるため問題はないが、ファイルのバイト数から文字数を推定するには注意が必要である。

4.2. 見出しに多く見られる違い

見出しで多く見られた違いは、「情報の追加」「一部書き換え」「記号の追加」の3点としてまとめられる（なお、複数の項目に該当するものもある)。例えば、（5）と（6）の違いは、網掛けされた部分が情報として追加されており、（7）と（8）では「本紙の記事」が「毎日新聞記事」に書き換えられ、"――"という記号が追加されている（「情報の追加」も行われている)。

（5） 柏田、救援で1失点　　（6月3日朝刊21面）
（6） 米大リーグ　メッツの柏田貴史投手、救援で1失点
（7） 「身代金要求」 本紙の記事　ペルー各紙が報道　　（1月6日朝刊7面）
（8） 「身代金要求」の毎日新聞記事、ペルー各紙が報道――ペルー日本大使公邸占拠事件

「情報の追加」には、「米大リーグ」「メッツ」などという語の追加の他に、（9）のような［社告］・［訃報］・［経済観測］・［みんなの広場］など、記事の種別に関する追加も行われている。

（9） ［社告］ 1月2日の新聞、休みます　　（1月1日朝刊CD)

このように、見出しには情報の追加や語順の変更が多く見られる。これがどのくらいの頻度で起こり、紙面とCDデータがどの程度異なるのか、数量的比較が可能ないくつかの事例については、4.3.6で調査する。

なお、(10) に対する (11) のように語が追加されるだけでなく、語順が変更されているものもあるが、語順の変更については数値化できないため、本稿

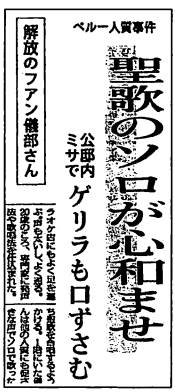
では現象の指摘に留める。また、情報の追加を伴わない語順の変更は、語数に変化がないため違いとしなかった。

(10)　職員国籍条項　松本市も撤廃　　(5月2日朝刊22面)
(11)　職員採用試験から国籍条項を撤廃――長野・松本市

4.3. 記事本文に多く見られる違い

記事本文の違いは、「記号・助詞の追加・変更」「記事本文の追加」「記事本文の削除」「記事本文の統合」という四つの観点から観察できた。「記事本文の統合」は、紙面のいろいろな部分に存在する記事を一つにまとめたもので、スポーツ（特に野球）に多く見られるものである。これは特に語数などでは相違がないため、本稿では除外し、それ以外の違いについて説明する。

なお、記事本文に見られる違いを観察する前に、そもそもCDデータで記事本文としてタグ付けされているものが、紙面でも記事本文なのか、という問題がある。


ペルー人質事件
聖歌のソロが心和ませ
公邸内ミサでゲリラも口ずさむ
解放のフアン儀部さん

(12)　←（左の画像）　　(1月7日夕刊8面)
(13)　(見) 解放のフアン儀部さん「公邸内ミサで聖歌のソロ」――ペルー日本大使公邸占拠事件
　　(本文) ◇聖歌のソロが心和ませ――ゲリラも口ずさむ

前述のデータ仕様書には、紙面のどのような部分を見出し・記事本文としたのか、明確な説明がなく、文字の大きさなどとも関係ないようである。例えば、(12) において、文字のポイントの大きい「聖歌のソロが心和ませ」の部分は、(13) では記事本文のタグが付けられ、逆に文字が小さい「解放のフアン儀部さん」「公邸内ミサで」が見出しになっている（「(見)」は見出し、「(本文)」は記事本文を示す）。同様に、同じ文字の大きさで表示されている文字列が、一

部は見出しのタグが付けられ、一部は記事本文のタグが付けられているというケースもある。

4.3.1. 記号・助詞の追加・変更

CDデータの中には、記事本文の中でも、記号が追加されたり、助詞が追加されているものが存在する。例えば、(14) に対する (15) では、記号“◇”と助詞「は」の追加が行われている。

(14) 三頭体制構築も、支持率頭打ちに (11月18日朝刊7面)
(15) ◇三頭体制構築も、支持率は頭打ちに

これは、4.3で言及したように記事本文とは明らかに異なる、中見出しや小見出しに相当する部分に、記事本文のタグが付けられ、本文記事とされているため、ここで言及しているが、このような違いは4.2.の「見出しに多く見られる違い」と同様に扱われるべき問題であると考える。

4.3.2. 記事本文の追加

紙面とCDデータが異なっているものとして、CDデータには (16) (17) のように、情報が追加されている例があった。これはグラフ・写真・図に限られ、また追加される場所は、記事本文の最後となっている。

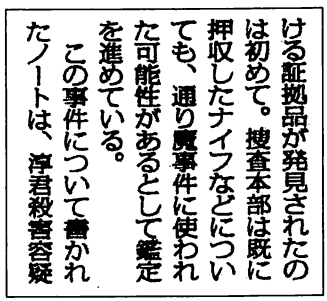
(16) (この記事にはグラフ「脳死を人の死とする臓器移植法案への主要政党別賛否」があります) (4月17日朝刊CD)
(17) ■写真説明 「仮名手本忠臣蔵」七段目の舞台 (12月10日夕刊CD)

4.3.3. 記事本文の削除

次に、記事本文が削除されている例について見てみる。紙面では (18) のように、この後に「この事件について書かれた……」で始まる段落とさらにもう

1段落があるが、CDデータでは（19）のように、この2段落分が削除されている。

(18)



（7月2日朝刊27面）

(19)（略）……ける証拠品が発見されたのは初めて。捜査本部は既に押収したナイフなどについても、通り魔事件に使われた可能性があるとして鑑定を進めている。

別の都市の合併に関する長文の記事でも、一部の記事本文が削除されており、削除が「殺人事件であるから」という理由でもなく、(19) は短い記事なので記事の長短でもなさそうである。このように記事本文の一部が削除される理由は不明であるが、サンプルデータ1,000例では2例見られた。これは、記事の網羅性や完全性を損なう問題であり、紙面を再現するには修復が必要だと思われる。なお、どのような部分が削除されるのかは規則化できないため、どのくらいあるのか数値化することは不可能である。

4.3.4. 記事本文の表記の違い

文字コードによる表記の問題を取り上げる。いわゆる機種依存文字は、紙面とCDデータでは違いが見られる。

まず、紙面では（20）のように丸数字（丸付き数字、①②③④など）であるが、CDデータでは、(21) のように括弧に囲まれた形式（(1)（2）（3）（4）など）となっている。これに伴って、紙面とCDデータでは（22）(23)

のように文言の違いも見られる。

(20) (略) ①７月のＮＡＴＯ首脳会議で発表する新規加盟国に関する問題
②ＳＴＡＲＴ２批准（中略) (２月８日夕刊１面)
(21) (略)（１）７月のＮＡＴＯ首脳会議で発表する新規加盟国に関する問題 （２）ＳＴＡＲＴ２批准（中略)
(22) 丸数字は学年 (５月30日夕刊９面)
(23) カッコ囲み数字は学年

また、紙面では、(24) のように組み文字（㌔・㍑・kmなど）が使われているが、CD データでは、(25) のように組み文字を使用せずに記述（キロ・リットル・km など）している。

(24) 雪印乳業は、(略) 500㍉㍑入り (１月18日朝刊10面)
(25) 雪印乳業は、(略) ５００ミリリットル入り

これらは、解析器に茶筌、辞書に IPA 辞書を使用して形態素解析を行った場合、丸数字や組み文字は、通常“未知語”と解析されてしまうため、望ましい変更と考えられる（なお、辞書に UniDic[9] を使用して形態素解析した場合、丸数字や組み文字も正しく解析される)。ただし、CD データから新聞記事の正確な文字数などを測定するには、問題が生ずることを付記しておく。

また、紙面では常用漢字表外の漢字にルビがついているが、CD データはテキストデータの形式であるため、ルビはつけることが出来ない。そこで CD データでは、(26) のように、ルビを漢字と送りがなの間に括弧“( )”を入れて表示している。

[9] http : //www.tokuteicorpus.jp/dist/

(26)　(略) から蘇 (よみがえ) り、第二の戦争 (以下略) (11月 3 日朝刊 CD)

なお、ルビが漢字の後につくことによって、形態素解析の精度が下がることが考えられるため、処理には配慮が必要となる。例えば、茶筌を使用して形態素解析を行った場合、「蘇り」と「蘇 (よみがえ) り」は、(27) と (28) のように全く異なる結果となり (表層語・読み・辞書形・品詞の順。一部出力結果を修正した)、(27) では 1 語、(28) では記号も入れて 7 語となってしまう。

(27)	蘇り	ヨミガエリ	蘇る	動詞－自立　五段・ラ行　連用形
(28)	蘇	ソ	蘇	名詞－固有名詞－人名－姓
	(	(	(	記号－括弧開
	よみ	ヨミ	よみ	名詞－一般
	が	ガ	が	助詞－格助詞－一般
	え	エ	える	動詞－自立　一段　連用形
	)	)	)	記号－括弧閉
	り	リ	り	助動詞　文語・リ　基本形

なお、紙面の文字と CD データの文字が異なっていて、これらが異体字の関係となっている場合がある。これについては、データは異なるが横山他 (1998) で詳細な研究が行われているため、ここでは現象の指摘に留める。サンプルデータにおいては、紙面で (29) のように「諫」を用いているが、CD データでは (30) のように「諫」が用いられている例が見られた。

(29)　諫早湾干拓反対派がシンポ　(5 月12日朝刊26面)
(30)　諫早湾干拓反対派がシンポジウム――長崎

4. 3. 5. 下駄文字について

いわゆる下駄文字“〓”は、文字コード (この場合では SJIS) 上に存在し

ない文字の代わりとして使われるもので、新聞社独自の活字をデータ化する際の差異から出る問題である。これも相違として捉えていいかと思うが、これらは人名などで多く使われており、例えば（31）の＝は「鴕」に相当する。また、“＝”は文字コード上に存在しない文字の代わりに使われるものであるが、この代用には恣意性があると思われ、下駄文字にするか否かは、何らかの判断が加わっていると推測される例が見られた。例えば、（32）では「鄧小平」の「鄧」は下駄文字にせず、カタカナに開いている。しかし、「黄長燁」については、下駄文字が使用されている。他にも下駄文字は、スポーツ面や社会面・国際面での人名表記に多く使われている。

（31）（略）連ドラでは、草＝剛の「いいひと。」（略）　（12月19日朝刊 CD）

（32）（略）朝鮮民主主義人民共和国（北朝鮮）の黄長＝書記が亡命申請、と韓国が発表（１２日）…（略）…▼中国の最高実力者、トウ小平氏死去（１９日）　（12月31日朝刊 CD）

下駄文字は、1,000記事内で6記事に見られ、全記事中では1,194記事に見られた（複数使われる場合も1例とした）。この中で、（33）のように下駄文字は何らかの記号の代用として使われているのではないかと思われる例も、いくつか見られた。これらの表記が使われているのは全て大阪発の記事であるため、検証が出来なかったので、今後の課題としたい。

（33）（略）サントリーミュージアム＝天保山＝（略）　（１月９日夕刊 CD）

また、記事本文では、開き括弧と閉じ括弧が一致していない例がある。（34）では開き括弧“「”に対し、閉じ括弧が“]”となり、一致していない。このような例は、機械的に用例が収集できるので、以降で詳しく調査する。

（34）盗聴捜査、「賛成]が５６％（略）　（９月17日朝刊 CD）

4.3.6. 数量的調査

前述したように数量的調査が可能なものである「見出しの相違」、「括弧の不一致」などについて、CDデータに含まれる全ての見出し・記事本文を対象に調査を行う。

4.3.6.1. 見出しの形態的異同について

4.2.で見たように、見出しには紙面とCDデータで異なる部分が頻出した。そこで、紙面とCDデータの見出しについて品詞ごとに語数を比較し、どのような語に違いが見られるか調査する。

なお、見出しのデータについては、紙面上でどこまでを見出しとすべきか判然としないもの、CDデータと極端に異なっているものに限り、CDデータを元に紙面の見出しを疑似的に再現した。疑似的としたのは、CDデータにある見出しを元に、それに相当する紙面の部分をデータとしたからである。前出の(13)では、紙面の見出しを「解放のフアン儀部さん　公邸内ミサで聖歌　ペルー人質事件」とした。

品詞に分ける作業では、まずCDデータの見出しに対して、MeCab[10](ver.0.98)と茶筌(ver.2.3.3)という2種類の解析器で形態素解析を行い、語の区切りの異なる105の見出しを考察した(辞書はともにIPA辞書である)。その結果、どちらも一長一短があるが、MeCabは辞書に単語が登録されていない場合でも品詞を推測するため、カタカナ表記の人名などの解析結果が良好であった。見出しにはこのような未登録の語が多く含まれることが予想されるため、MeCabを解析器に使用した(なお、形態素解析の結果には誤りも多く含まれるが、今回は大まかな傾向を捉えるため、人手による修正は行わなかった)。

その結果、紙面の延べ語数は13,308語、CDデータは16,583語であった。紙面とCDデータで語数に大きな違いが見られたのは、名詞(8,000／9,827、紙

[10] http : //mecab.sourceforge.net

面とCDデータの順。以下同様)、助詞 (1,495／1,969)、接頭詞 (201／239)、記号 (2,911／3,831) である。しかし、他の品詞については特に大きな違いが見られなかった (例えば動詞 (399／407)、形容詞 (149／150))。

品詞の下位区分を見ると (以下IPA辞書の品詞分類)、名詞・助詞について、名詞-固有名詞 (1,437／1,881)、名詞-副詞可能 (143／185)、助詞-連体化 (435／642)、助詞-並立助詞 (39／51)、助詞-格助詞 (729／974) で大きな違いが見られた (なお、品詞の下位区分の精度については問題が見られることを付記しておく)。

まず、CDデータでは「固有名詞」、「名詞-副詞可能」の数が多い結果となった (「名詞-副詞可能」は“1月・2月”のような月を表す語が該当する) が、固有名詞では (36) における“ペルー”、「名詞-副詞可能」では (38) のような“4月”という情報の追加が見られたことによる。

(35) フジモリ大統領、５カ月ぶり　偶然の家族だんらん
(５月12日朝刊２面)
(36) ペルーのフジモリ大統領、偶然の家族だんらん――５カ月ぶり
(37) 20日の東京競馬特別レース　日曜競馬　(４月19日夕刊６面)
(38) ４月２０日の東京競馬特別レース　(日曜競馬)

助詞の「格助詞」「連体化」(“の”のみ) や「並列助詞」(“と”) の例として、(40) (42) のような助詞の追加が見られる。

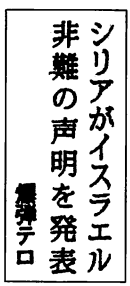
(39) 特殊法人天下り　「 ひどすぎる 」 亀井建設相答弁(２月５日夕刊１面)
(40) 「 特殊法人への天下り、ひどすぎる 」――亀井静香建設相が答弁
(41) (略) 亜大、東洋大が勝ち点　(４月10日朝刊18面)
(42) (略) 亜大と東洋大が勝ち点

接頭詞とは“現・元・旧・新”などが該当するが、CDデータで追加されて

いるもので多かったのは、(44) のような“第”であった。

(43) 全国社会人ラグビー 東芝府中 2年ぶり決勝 (1月26日朝刊24面)
(44) ラグビー 全国社会人大会〈第1 1日〉 東芝府中、2年ぶり決勝へ

記号については、紙面とCDデータで異なる結果となった。まず紙面に多かったのが全角空白文字“ ”(以後全角空白文字が存在することを明示したい場合は、“□”で示すこととする) である。これは、見出しデータの作成の際に入力したものである。例えば、実際の紙面では (45) のように縦書きであるが、改行されている部分に全角空白文字を一律挿入し、(46) のような形にした。このため、新聞紙面では、全角空白文字が多くなる結果となった。



(45) (←左の画像)
(46) シリアがイスラエル□非難の声明を発表□爆弾テロ
(1月4日朝刊7面)

これに対しCDデータで多かったのが“[ ]”と“——”であった。“[ ]”は、(50) の [社説] のように見出しの頭に置かれ、記事の種別を追加するもので、396件見られた。なお、“[”が396回、“]”が397回 (以下に述べる括弧の不一致による) で、新聞紙面では使われていなかった。“——”は、(48) の「ダマスカス」のように情報の追加に用いられ、新聞紙面では2例であったが、CDデータでは497例見られた。

(47) シリアがイスラエル 非難の声明を発表 爆弾テロ ((46) の再掲)
(48) シリア当局、イスラエル非難の声明を発表——ダマスカスの爆弾テロ事件
(49) 社説 景気 晴れ間をもっと広げよう (4月3日朝刊5面)
(50) [社説] 景気 晴れ間をもっと広げよう

このように、見出しにおいては、紙面とCDデータで異なる部分が多い。このような理由から、CDデータを元に、新聞の見出しの研究を行うのは困難だと予想される。同様に語彙・文法研究などでも、より紙面に近い結果を求めるのであれば、見出しは対象外とするのがよいのではないかと思われる（横山他(1998)でも同様の扱いをしている)。

4.3.6.2. 括弧の不一致について

括弧の不一致には、対応する括弧の形が間違っている場合、対応する括弧が論理行中にない場合、対応する括弧自体が記事本文にない場合の三つのケースがある。(51)は見出しの一部で、開き括弧と閉じ括弧の形が異なっている。(52)は投稿者の住所・氏名・年齢を閉じるカッコが別の行に書かれているケースで、(53)は書名を表す閉じカッコがないケースである。以下わかりやすさを考え、改行マーク“‹ENTER›”を適宜追加する。なお、(53)は既に新聞紙面の段階で、対応する括弧が抜けていたケースである。

(51) 盗聴捜査、「賛成」が５６％　凶悪犯罪多発が理由　（９月17日朝刊CD)
(52) 即妙に対応していく博士たちのやりとりにこそ注目を。(千葉県習志野市　伊＊＊知　６５ ‹ENTER›
無職)　（６月30日朝刊CD)
(53) 「むしばミュータンスのぼうけん（童心社)　（２月７日朝刊CD)

一行中になくても、次の行にあればいいのではないかと考えられるかも知れない。目視で例文を集める場合ならばそれでも問題ない。しかし、大量のデータを検索する時、通常Perl等のプログラミング言語を用いるが、これらの場合、改行コードが入るまでを一つの単位をして処理するため、改行コードをまたいでも文が続いていることは想定していない。そのため、後掲する引用文などで、開き括弧“「”と閉じ括弧“」”の間に改行が入ると、本来引用文で一文として扱うべきところが上手く処理できないこととなる。

サンプルデータでは、それぞれ1件ずつ見られただけだが、他のケースを調べるために、参考資料として1997年全体での調査を行った。

4.3.6.3.1. 全体の結果

開き括弧と閉じ括弧の有無について調べた結果の一部を表1に示す（出現数が少ない場合、同数の場合などは省略した)。なお、“(” “)” については、a）b）のような形式があるため、数値は参考として掲載した。

表1　CDデータにおける開き括弧と閉じ括弧の有無とその出現数

条件1	条件2	出現数
“「” あり	“」” なし	677
“「” なし	“」” あり	716
“『” あり	“』” なし	6
“『” なし	“』” あり	5
“(” あり	“)” なし	119
“(” なし	“)” あり	166
“【” あり	“】” なし	60
“【” なし	“】” あり	60

表1は、例えば一文中に “「” があり、“」” がないものが出現した数が677例あったことを示す。“【 】” は、対応する括弧が必ず次の行にあったため、「“【” が当該行にあり “】” がない」場合と、「“【” がなく、“】” がある」場合の数値が一致した。

しかし、対応する括弧の不一致や記事本文に対応する括弧がないことがあるため、「あり・なし」の数値と「なし・あり」の数値は必ずしも一致するわけでもない。これらについて、以下詳細に見ていく。

4.3.6.3.2. 対応する括弧の形が違っている場合

開き括弧に対する閉じ括弧の語形が異なるものがある。例えば、開き括弧が “『” の場合は、閉じ括弧は “』” であるべきところが、“」” となっているもの

などである。実際に（54）は見出しで見られ、（55）は記事本文に見られた例である。（54）は新聞紙面では、“診断”の部分は黒字に白の文字になっており、括弧の追加の際の誤植と思われ、（55）は紙面の段階で既に対応する括弧の形が違っている。

（54）［診断」生活設計　長男宅の隣りの土地（略）　（8月10日朝刊CD）
（55）（略）幸さん（１９）は「『支える会」の人が駅前でハンストしているのを見た。（略）と思った」。　（2月27日夕刊CD）

4.3.6.3.3. 対応する括弧が存在しないもの

対応する括弧が記事本文中にない例として、（56）のようなものがある。

（56）「林道から（略）３枚目は『かやが覆い３０センチの大木が横になっているところに目印がついている。（略）ほぼ同じではないですか」
（6月20日朝刊CD）

これはCDデータの問題ではなく、新聞紙面にもともと対応する括弧が存在しない、誤植によるものである。これは閉じ括弧が存在しないケースであるが、対応する開き括弧が存在しないものも見られた。（57）はCDデータにのみ追加されている記事であり、（58）は新聞紙面では“■羽生□谷川”とあるものを文字化したものであり、ともに誤植と思われる（■・□は将棋の駒の形）。

（57）（この記事には第５局の指し手図」があります）　（6月22日朝刊CD）
（58）◇（第２日指し手）［先］羽生［後］谷川」　（6月12日朝刊CD）

CDデータ内の記事本文に関し、対応する括弧が存在しない場合を見てきた。この条件に当てはまるもの全てを元の紙面と照合したわけではないが、元の新聞紙面において既に対応する括弧が抜けている場合がかなりの数存在すること

を指摘しておく。

4.3.6.3.4. 当該行にはないが次の行に存在するもの

次に、対応する括弧が当該行にはなく、改行された次の行に存在するものを見ていく。これは、途中に改行マークが入るということで、記事本文にのみ見られる現象である。(59) (60)、また (61) (62) は本来連続した一つの文である。しかし、(59) は“「 」”で囲まれたいわゆる引用文の途中に改行が入り、対応する括弧“」”が別の行 (60) にある例である。(61) (62) は“『 』”が途中で切られている例である。

(59) (略) 組合員 (６６) は「自分が確認した。油ください。‹ENTER›
(60) 今の時期はイソで海女さんがノリをとる時期。打撃が心配だ」と双眼鏡を握りしめていた。 (1月7日夕刊 CD)
(61) (略) ぼくは、『大人は勝手や』と思う今の中学生の気持ちがよく分かる。ぼくがこの何年間か感じてきた『意見を言わせて。‹ENTER›
(62) ちゃんと聞いて』という思いは、中学生たちと一緒のはずです」という言葉は、(略) (8月26日朝刊 CD)

(59) や (61) は句点“。”で終わっており、文の形式をしているが、以下の (63) については、文の形式をしていない。紙面を見ると、この部分に全角空白文字“□”があり、この部分に改行マークが挿入されているケースである。句点“。”で区切るのは理解できるが、それ以外での区切りも多いため、検索では注意が必要となる。これは単純に全角空白文字“□”で区切っているのではないかと思われるかもしれないが、(64) (65) のように、同一の文中でも改行されている部分とされていない部分があり、何らかの規則があるとは思えない (なお、(59) ～ (62)、(64)・(65) は大阪発の記事であり、実際に紙面を調べることはできなかった)。

(63) 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) (12月25日夕刊 CD)

(64) ハリウッドに残る世界唯一のサイレント映画館「ＳＩＬＥＮＴ□ ‹ENTER›
ＭＯＶＩＥ」。これを大阪港天保山・海遊館ホールで再現する……

(65) (略) そして「ＳＩＬＥＮＴ□ＭＯＶＩＥ」自体が今、失われる危機にあります。 (９月８日朝刊 CD)

また、(66) のように年表などの箇条書きに近いものの一部でも、区切られるケースが多い。

(66) □３．□４□ＣＤＣが「血友病患者のＨＩＶ感染は血液 ‹ENTER›
□□□□□□製剤が原因とみられる」と警告 (３月10日夕刊 CD)

例えば、“「」”をもとに直接引用文を抽出するときなどには、括弧が対応していない部分があるため、網羅的な検索が出来ないことが予想される。これらについては、括弧の修正を行う前処理を加えることにより紙面に近いデータにすることができ、より正確な結果を得られることが期待できる。

しかし、(66) のように紙面で箇条書きになっているものは、CD データでも改行が加えられていて、その改行位置は上記のように一定の規則がないため、全面的な修正は困難であると思われる。

4.3.6.3.5. 全角空白文字について

次に、括弧“【 】”に関連した問題を見ていく。“【 】”が使われる部分でよく見られる現象であるが、全角空白文字“□”が数多く挿入されている。

(67) 【聞き手・倉田真□□□□□□□□□□□□□□□□□□□□ ‹ENTER›
西部本社編集局長】□□□□□□ (一部略) □□□ (１月７日朝刊 CD)

例えば、(67) の例には、全角空白文字が行末まで入っているが、新聞紙面は“【聞き手・倉田真西部本社編集局長】”のみで、途中に改行も全角空白文字もない。見出しに全角空白文字が複数含まれるデータはないが、(68) や (69) のように、記事本文には全角空白文字がかなりの数使用されている。

(68) □□□□【中西満】□□□□□□□□□□□□□ (1月11日朝刊 CD)

(69) (1) サザエさん□□□□□□□□□□２２．６ (フジ) 日
(2) 名探偵コナン□□□□□□□□□１８．２ (日本) 月
(3) ちびまる子ちゃん□□□□□□□１７．９ (フジ) 日
(3月7日朝刊 CD)

(70)

(日本) 水⑨サイコメトラー映	(フジ)日②名探偵コナン18.2 (日
17.5 (日本) 土⑩踊る大捜査線	本)月⑧ちびまる子ちゃん17.9(フ
.3 (フジ) 火	ジ) 日④ドラゴンボールＧＴ15.9
◇アニメ◇ ①サザエさん22.6	(フジ)水⑤るろうに剣心15.6 (フ

(69) に対応する紙面は (70) のような形であるため、(69) は全角空白文字を挿入して見やすく加工した結果と思われる。このように見やすさを考慮した全角空白文字の使用は、別の問題を引き起こしてしまう可能性がある、(71) は年表内のものを記事としているものであるが、年表を忠実に再現しているため、表現が途中で区切られてしまっている (わかりやすいように一部表記を変えた)。

(71) ７５　国際女性年世界会議、メキシコ市で開催
□□□□「ワタシ作る人、ボク食べる人」ＣＭに性別役割を ‹ENTER›
□□□□固定化するとして、女性グループが (略) (5月1日朝刊 CD)

このように文字数を数える際、全角空白文字を文字として数えると、かなり実態と離れた数値になるおそれがある。語数を数える際、全角空白文字の品詞は「記号」であるため、品詞が「記号」の語を除外して数えるという方針もあ

るが、記号にはアルファベット・句読点・括弧も含まれるため、扱いに困る問題である。少なくともCDデータにおいては、全角空白文字の数は排除したほうがより正確な数字が得られることとなる。また、全角空白文字は箇条書きなどの一部に挿入されることがあり、これらも検索の障害になると思われる。

4.3.6.3.6.『CD-毎日新聞2007データ集』の調査結果

以上97年のCDデータ（以下、「CDデータ97」と呼ぶ）で観察された点について、『CD-毎日新聞2007データ集』（「CDデータ07」と呼ぶ）でも確認されるか、調査する（無作為抽出した100記事を使用するか、全体を使用するかは、調査項目の特徴を考慮して、選択した）。

まず、見出しに記事の種別を括弧に入れて表示する“[社説]”のような形式であるが、CDデータ07では“社説：”のような形に変わっていた。また、見出しにおける「情報の追加」「一部書きかえ」「記号の追加」の問題は、CDデータ97同様に見られる現象であった。また算用数字の使用も、見出し・記事本文で同様に見られた。

記事本文の括弧の問題について、出現数が少ないことが予測できるため、CDデータ07全体で検証したが、(72) や (74) のように“「 」”、“( )”の間で改行記号が入り、それぞれ (73)、(75) と分断されている例が延べ54例見られただけで、他の括弧“『 』”、“【 】”、“[ ]”では見られずデータ全体で誤植が少なくなっているのではないかと思われる。

(72) （略）「日本画は勢いで描くことができない。（略）一筆に覚悟を持たなければならなければならな ‹ENTER›
(73) い、そうした意味で日本画は武士道」と語った。
（2007年5月25日夕刊CD）
(74) キヤノン賞＝デジタル一眼レフカメラ（ＥＯＳＫｉｓｓデジタルＸレンズキッ ‹ENTER›
(75) ト）
（2007年5月19日朝刊CD）

なお、“【 】”の使用法はCDデータ97では、統一されていなかったが、CDデータ07では、“【姓名】”という形で統一された形で使用されている。

また、これに関連して指摘した全角空白文字“□”の多用であるが、これはCDデータ07でも同様に見られた。例えば（76）は“□”が一つ入っている例で、（77）は“□”が文字列の先頭を揃えるために挿入されている例である。

（76） キヤノン賞＝コンパクトデジタルカメ□ラ（ＰｏｗｅｒＳｈｏｔＳ３ＩＳ） （2007年6月29日朝刊CD）

（77） 【得点者】２９分・服部（広）‹ENTER›
□□□□□７１分・ウェズレイ（広） （2007年8月19日朝刊CD）

他の下駄文字の使用、記号の追加などは、CDデータ07でも同様に見られた。このようなことから、いくつかの問題点も残っているが、修正されている部分もあり、できるならば、新しいデータを選択することが望ましいのではないかと思われる。

最後に、CDデータ07に見られた問題を挙げておく。通常新聞記事では（78）のように、発言の引用部分で改行されることはない。しかし、（79）のように、引用部分で改行される記事本文が見られた。このような形式は、小説などでは多く見られ、検索の際の困難点となる問題であるが、新聞記事でもこのような形式が見られた。これについては、いつからどの程度出現しているのかも含め、今後の課題としたい。

（78） 一方、松岡＊＊農相は「事務所費として報告すれば、説明責任を果たしている」と言い、（略） （2007年1月27日朝刊3面）

（79） （略）細かに諭し、励ました。安倍側近の一人は、‹ENTER›
「指南役をお願いした、ということではないか」‹ENTER›
と言う。 （2007年1月27日朝刊3面）

5．まとめ

本調査では、新聞の紙面とCDデータの比較作業を通じ、新聞記事データ集を扱う際の問題点を明らかにしてきた。

新聞記事データ集を言語研究に使用する際、より正確な結果を求めるならば、以下の条件が必要と考えられる。まず、引用文を研究する際、事前に括弧を修正しておくか、改行が挿入されていることを承知して検索を行うか、どちらかの対応が必要になる。次に、新聞記事データ集内の見出しであるが、これは研究対象からはずした方がよいのではないかと思われる。あえて研究対象とするならば、新聞紙面を元にした全面的な修正と、「見出し」か「記事本文」かという現在の二分法についての再考が必要となると考えられる。

この二者については、修正・排除という対応が取れるが、“□”でレイアウトされた記事本文については、特に規則性が見当たらないため、修正が不可能（または、非常に困難）である。

このような新聞記事データ集に対し、正確に文字列検索し、情報抽出を行うには困難な点がいくつかある。網羅的に検索を行っているつもりでも、そのままでは網から外れてしまう用例がたくさんあることも注意する必要がある。

以上のようにさまざまな異なる点を挙げたが、新聞記事データ集を対象として行う調査においては、その結果について慎重に扱う必要があるということをまとめとして指摘できるのではないかと思う。今回の結果は、『CD-毎日新聞データ集97年』を対象とした場合に限定されるもので、朝日新聞や毎日新聞など、他の新聞記事データ集を対象とした場合でも同様の検証が必要となろう。執筆現在、均衡のとれたコーパスは存在しておらず、国立国語研究所が作成中である『現代日本語書き言葉均衡コーパス』が2011年にその唯一のものとなる予定である（山崎他（2006））。このようなコーパスに関しても、同様の検証が必要なのではないかと考える。

また、新聞記事データ集はWEBデータなどに比べ、規範性の高い文が収集できるコーパスである。そのため、コーパス言語学などの分野では使用されることが多いものである。今回見たように、新聞紙面の情報にさまざまな要素を

追加するというCDデータ内の見出しの変更は、いつ読んでも理解できるようにというような配慮による修正であろう（詳細については、作成者に問い合わせる必要があるが、時間の関係で行えなかった）。しかしこれが逆に紙面の姿から遠ざかる結果となってしまっていることも事実である。言語研究に要求される新聞記事データ集とはどんな姿か、一顧に値すると思われる。

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巻第8号、明治書院

松田謙次郎・薄井良子・南部智史・岡田裕子（2008）「第2章　国会議事録はどれほど発言に忠実か？　—整文の実態を探る—」『国会議事録を使った日本語研究』松田謙次郎編、ひつじ書房

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、第22巻第5号、明治書院

山崎　誠・前川喜久雄・田中牧郎・小椋秀樹・柏野和佳子・小磯花絵・間淵洋子・丸山岳彦・山口昌也・秋元祐哉・稲益佐知子・吉田谷幸宏（2006）「代表性を有する現代日本語書き言葉コーパスの設計」、言語処理学会第12回大会発表論文（NLP2006）

横山詔一・笠原宏之・野崎浩成・エリクロング（1998）『新聞電子メディアの漢字—朝日新聞CD-ROMによる漢字頻度表—』国立国語研究所プロジェクト選書1、三省堂

用例出典

『毎日新聞縮刷版1997年1月』～『毎日新聞縮刷版1997年12月』（毎日新聞社）

『CD-毎日新聞'97データ集』『CD-毎日新聞2007データ集』（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・覚書にもとづき、使用したものである。