新聞記事データ集はどれほど新聞紙面に忠実か

長谷川守寿（首都大学東京）

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、『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、毎日新聞・日経新聞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初めて開始したのは、1995 年である（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ることから、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、新聞の記事データ集は、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。

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。

本調査は、新聞紙面と新聞記事データを比較する。その際、横山他（1998）で指摘されたように、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

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』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全文記事データ集（タグ付テキストデータ）」である。本研究では、Ｔ１（見出し）、Ｔ２（記事本文）のタグが付されているテキストデータを使用する（以下、紙面と CD データ両方を示す場合には、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・朝刊夕刊の別のみを示し、最後に CD と付す。また、新聞記事には個人・団体名が掲載されているが、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した)。

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、当然写真などの画像ファイルが収録されていない。さらに連載小説、四コマ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、研究家などの著名入りの記事であるが、これらの記事本文は収録されていない。1)のように見出しに“★”がはじめに付いているものは、記事本文は 2)のようになっていて、個別の記事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＊和子　(1/5 朝 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 119836 であるが、1)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 8561 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。

3.3.手順

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 1000 の記事を無作為抽出し、見出しと記事本文を、新聞紙面と比較する。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 100 記事を対象に予備調査を行った。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、数値的に比較可能なものについて、1000 記事中でどのくらい起こるか、調査する。

4.結果

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、“紙面”“CD データ”という略称を用いる。また、異なる部分は網掛けをした。

なお、新聞紙面は縦書きであるが、状況により横書きで再現したり、紙面を画像ファイルで示す。

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、1000 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、見出しでもいくつか同様の違いが見られた。

3) ドキュメント　リマ 24 時　　(2/12 朝 7)
4) ［ドキュメント］リマ２４時

解析エンジンに茶筌(2.4.2)、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。また、括弧と全角空白文字（以後“□”で表す）については、様々な違いが見られたので、別項として記述する。

4.2.見出しに多く見られる違い

見出しで見られた違いは、「情報の追加」「一部書き換え」「記号の追加」の三つの観点から観察できる（なお、複数の項目に該当するものもある)。例えば、5)と 6)では、5)に網掛けされた部分が追加されており、7)と 8)では 7)に「本紙の記事」が「毎日新聞記事」に書き換えられ、“――”という記号などが追加されている。

5) 柏田、救援で１失点　　(6/3 朝 21)
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　　(1/6 朝 7)
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

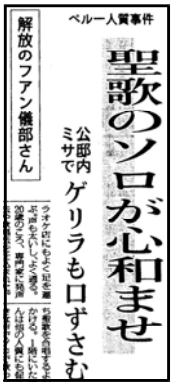
「情報の追加」には、「米大リーグ」「メッツ」などの追加の他に、9)のような［社告］・［訃報］・［経済観測］・［みんなの広場］など、記事の種別に関する追加も行われている。

9) ［社告］１月２日の新聞、休みます　(1/1 朝 CD)

このように、見出しには情報の追加が多く見られる。MeCab(0.98)と IPA 辞書を用いて、紙面と CD データの見出しを形態素解析した結果では、名詞に大きな違いが見られ、CD データでは固有名詞の延べ語数が多く、また“[]”や“――”のような記号の延べ語数が多かった。

4.3.記事本文に多く見られる違い

記事本文は、大きな違いとしては記事本文の「追加」「削除」「統合」が挙げられ、細かい違いとして「記号・助詞」「表記」の違いが挙げられる。また、「括弧」については別のタイプの問題が見られた。記事本文の「追加」「削除」「統合」については、紙幅の都合で省略し、細かい違いについて説明するが、まず、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。



10) ←（左の画像）　　(1/7 夕 8)
11) （見出し）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

「納品データ仕様書（本社版)」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、10)で、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、11)では記事本文のタグがつき、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。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.本文の記号・助詞の違い

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 12)と 13)では、12)に記号“◇“と助詞「は」の追加が行われている。

12) 三頭体制構築も、支持率頭打ちに　(11/18 朝 7)
13) ◇三頭体制構築も、支持率は頭打ちに

これは、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。なお、CD データの場合、助詞の延べ語数が顕著に多いことが明らかになった。

4.3.2記事本文の表記の違い

紙面と CD データでは、いわゆる機種依存文字の扱いに違いが見られた。紙面では丸数字（丸付き数字、①②など）であるが、CD データでは括弧に囲まれた形式（(１)（２）など）となっている。また、紙面では組み文字（キロ・リットル）が使われているが、CD データでは組み文字を使用せずに記述（キロ・リットルなど）している。

解析エンジンに茶筌、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 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、14)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。

14)（略）から蘇（よみがえ）り、（以下略）(11/3 朝 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。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、15)と 16)のように全く異なる結果となり（一部出力結果を修正)、品詞情報とともに表現を検索したり、語数を数える時に問題が生じる。なお、他に、異体字の使用、下駄文字の使用などによる違いが見られた。

15) 蘇り	ヨミガエリ	蘇る	動詞-自立
16) 蘇	ソ	蘇	名詞-固有名詞
（	（	（	記号-括弧開
よみ	ヨミ	よみ	名詞-一般
が	ガ	が	助詞-格助詞
え	エ	える	動詞-自立
）	）	）	記号-括弧閉
り	リ	り	助動詞　文語

4.4.括弧について

括弧には、対応する括弧の形が異なる場合、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。17)は見出しの一部で、開き括弧と閉じ括弧の形が異なり、18)は引用文を閉じるカッコが別の行（19)）に書かれ、20)は書名を表す閉じカッコがない。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、20)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。

17) 盗聴捜査、「賛成」が５６％　凶悪犯罪多発が理由　(9/17 朝 CD)
18) （略）組合員（６６）は「自分が確認した。油くさい。**<ENTER>**
19) 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。　(1/7 夕 CD)
20) 「むしばミュータンスのぼうけん（童心社）　(2/7 朝 CD)

目視で例文を集める場合ならば、一行中になくても、次の行にあれば問題ない。しかし、検索に正規表現を用いる場合、通常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、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。

上の例は句点の後ろで改行されているが、ほかにも句点以外の部分で改行されている例 21)もあった。

21) 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）　(12/25 夕 CD)

また、22)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。

22) □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　(3/10 夕 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、22)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。

なお、このような不一致は見出しにも見られたことを付記しておく。

4.5. 全角空白文字について

全角空白文字“□”が数多く挿入される現象がよく見られる。例えば、23)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、記事本文には全角空白文字がかなりの数使用されていた。

23) 【聞き手・倉田真□□□（略）□□□<ENTER>
西部本社編集局長】□□（略）□□（1/7 朝 CD）

24) （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
(3/7 朝 CD)

24)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、25)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)。

25) 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）
(5/1 朝 CD)

5.まとめ

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、以下の対応が必要になると考えられる。まず、引用文などを研究する場合、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。また、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）であると思われる。

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。

以上のように様々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。

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

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 5 号、明治書院

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

『CD-毎日新聞 1997 年版』『CD-毎日新聞 2007 年版』（日外アソシエーツ）

『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）