語の出現分布からみた月刊雑誌と新聞コーパスの特性調査
－用例収集資料としての多様性の検討－

柏野和佳子，　丸山岳彦，　稲益佐知子，　茂木俊伸

国立国語研究所

1.　はじめに

現在，大規模な新聞コーパスがさまざまに研究利用されている。その中で，その用例に新聞独特の偏りのあることが度々指摘されていた。たとえば後藤（1995)は，新聞は「内容の上でかなり広い範囲が扱われてはいるが，政治，経済，社会の分野に大きな比重が置かれており，したがって，それに関連した単語が過大な割合で出現する。」と述べ，例えば「首相」という語が日常生活において使用されるよりも新聞には高頻度で現れることを指摘している。しかし，「首相」以外にどのような単語が新聞に過大に出現しているかの具体的な報告はない。

一方，雑誌は幅広いジャンルにわたっており，豊富な用例収集資料として期待されるものである。しかしながら，大規模な電子化資料は存在せず，また，言語資料としての特性は未検討であった。

我々は，「バランストコーパス」の構築を目指し，用例収集資料としての多様性を探っている。この目的のために，今回，2003年発行の月刊雑誌50誌よりデータを収集し，毎日新聞 2003 年のコーパスと，品詞別に語の出現分布を比較した。その結果，雑誌と新聞とで収集可能な用例の差異を確認することができた。

2.　雑誌と新聞の特性を調べる方法

安本・本多(1981)は，週刊誌は新聞に比べ，以下の特徴があると述べている。

①「文」の長さが短い　②「会話文」の量が多い
③「句点」の数が多い　④「読点」の数が多い
⑤「漢字」の量が少ない　⑥「名詞」の数が少ない
⑦「人格語」の数が多い　⑧「不定止」の数が多い
⑨「名詞の長さ」が短い

陳(2004)は，新聞2紙，週刊誌16誌の文体を25項目の観点によって比較している。そして，現在もなお，週刊誌には安本・本多(1981)が指摘した傾向があり，さらに，「直喩」「声喩」「接続詞」「接続助詞」が多く，「現在止」が多用され，「過去止」の使用は非常に控えられていることなどを指摘している。

今回，茂木ほか(2005)では，語種構成の観点から，月刊雑誌と新聞を比較している。

これらに対し，本研究は，雑誌や新聞を用例収集資料として見るとき，収集可能な用例の多様性を検討することを目的に，品詞別に具体的な語の出現分布を比較する方法を提案し，それを試みるものである。その時の観点として次の3点を考えた。

①各資料から用例が安定してとれ得る語による比較
②各資料により用例のとれ方に差がある語による比較
③各資料ならではの用例がとれそうな語による比較

以上の観点による具体的な調査を4章で報告する。

3.　調査対象

3.1. 月刊雑誌50誌

今回の調査では，2003 年 8～12 月発行の月刊雑誌 50 誌（各 1 カ月分）を対象にした。選定した雑誌名は，以下のとおりである。ジャンルは，『雑誌のもくろく2003年版』（雑誌目録刊行会）による。「趣味」のジャンルは内容が多岐にわたるため，選定された雑誌の数が多くなっている。各雑誌の本文（広告や付録を除く）部分から，文末の句点もしくは記号（「。」「. 」「？」等）を含む文をランダムに 200 文ずつ，50 誌分計 10,000 文を抽出し，Windows 版『茶筌』（WinCha2000R2）[1]で形態素解析を行い，分析用データを作成した。形態素（改行記号を除く）は全部で 224,597 である。（分析用データの詳細は，茂木ほか(2005)を参照。）

ジャンル	雑誌名
「児童・学生」(3 誌)	Animage, My Birthday, 蛍雪時代
「女性」(4 誌)	mini, MORE, 家庭画報, ポップティーン
「家庭」(6 誌)	dancyu, ESSE, QUANTO, 新しい住まいの設計, 壮快, ひよこクラブ
「大衆」(2 誌)	L magazine, Myojo
「総合・文芸」(7 誌)	一個人, 財界人, 小説宝石, 短歌研究, 俳句, 文學界, 文藝春秋
「趣味」(22 誌)	BACKSTAGE PASS, BE-PAL, GOLF DIGEST, HOBBY JAPAN, Lure magazine, Option, SCREEN, Swing Journal, Tennis Classic Break, カメラマン, 月刊碁ワールド, サッカーズ, 月刊ザテレビジョン（首都圏版）, 月刊ジャイアンツ, 趣味の園芸, 鉄道ファン, 月刊バスケットボール, パチスロ必勝ガイド, ヤングマシン, 優駿, ラジオライフ, 旅行読売
「専門」(6 誌)	MONEY japan, Newton, 経済セミナー, 日経PC21, 日経TRENDY, 法学教室

[1] http://chasen.aist-nara.ac.jp/hiki/ChaSen/?FrontPage

3.2. 新聞コーパス

今回は毎日新聞2003年，1年分のデータを使用した。なお，雑誌の分析用データのサイズ(1 万文)にあえて揃えなかった。大規模な新聞コーパスに対し，小規模な雑誌データがどれくらいの多様性を示し得るのかを探るためである。このデータに対し，雑誌同様，『茶筌』で形態素解析を行い，分析用データを作成した。全形態素数(改行記号を除く)は35,081,224であり，雑誌の分析用データの約156倍の量である。最も使用度数の大きい形態素を比較すると，雑誌が「、」(読点)の9,504であり，新聞が「　」(空白)の1,484,638である。ちょうど約156倍を示した。すべて単純にはいかないであろうが，おおよそ雑誌の使用度数1は，新聞の約156に相当すると考えることができるだろう。

4.　語の出現分布の比較

雑誌と新聞，それぞれの資料の特性を語の出現分布を比較することによって調べた。

4.1. 各資料から用例が安定してとれ得る語による比較

用例が安定してとれ得る語として，使用度数の大きい上位50語に着目した。品詞別に，雑誌，新聞の各分析用データ中において，使用度数の大きい順に50語を抽出し，比較した。ただし，上位50番目と同じ度数の語はすべてとったため，雑誌の場合，51～53語を抽出したものもある。資料の性格に関わらず，上位50語がほぼ同じ内容になる，助詞，助動詞，非自立語や数詞，代名詞などの品詞は対象外にした。また，資料依存性が高いと考えられる固有名詞も対象外にした。よって，対象にした品詞は，「名詞-一般」「名詞-サ変接続」「名詞-形容動詞語幹」「名詞-副詞可能」「形容詞-自立」「動詞-自立」「副詞-一般・副詞-助詞類接続」「接続詞」の以上である。このうち，本稿では，あまり違いの見られなかった「名詞-副詞可能」「接続詞」を除いた調査結果を表1～表6に示す。各表では上位約50語のうち，雑誌にのみ出現する語，新聞にのみ出現する語，共通に出現する語を，その使用度数の大きさ順に示す[2]。なお,共通は新聞の順位で示す。

はじめに，表1～表3に「名詞」類を示す。特に，「名詞-一般」と「名詞-サ変接続」はもともと大きな語彙量の品詞であるため，資料の内容に即して違いのでやすいことが予想されたが，実際に共通するものは少なく，雑誌，新聞，それぞれにその特徴を示すような語が抽出された。たとえば，雑誌には，商品情報や，個人の嗜好に関する語が多い。一方，新聞には，後藤 (1995)の指摘通り，政治，経済，社会や事件に関する語が多いと言えよう。

表1：名詞-一般

雑誌	白，黒，図，手，作品，形，目，水，部分，効果，データ，ボール，体，言葉，映画，気，大学，基本，相手，モデル，商品，人気，一つ，花，ファン，最後，先生，株，環境，ポイント，状態，心，男，中心，右，家，技術，犯罪，タイプ，一般，左
新聞	首相，政府，大統領，事件，容疑，委員，会社，先，女性，社会，企業，後，政権，国，党，政治，全国，国民，地域，議員，大会，会長，事業，男性，政策，記者，被害，テロ，方針，金融，社長，機関，被告，子供，最終，核，記事，長官，地方
共通	時代，チーム，声，写真，人(ひと)，世界，情報，経済，国際，自分，選手

表2:名詞-サ変接続

雑誌	登場，発売，撮影，意味，利用，期待，使用，機能，紹介，設定，料理，位置，存在，活躍，変更，仕事，注目，変化，採用，運転，イメージ，デザイン，練習，開催，管理，一緒，演出，展開，編集，記念，契約，行為，表現，注意，移動，試験
新聞	説明，発表，戦争，調査，支援，交渉，選挙，協議，攻撃，改革，会議，優勝，計画，支持，報告，指摘，派遣，会見，影響，判断，著作，関連，対策，経営，協力，検討，死亡，逮捕，実施，決定，連続，施設，担当，出場，捜査
共通	関係，代表，監督，電話，研究，試合，参加，活動，開発，予定，表示，生活，対応，組織，話

表3:名詞-形容動詞語幹

雑誌	きれい，シンプル，不思議，有名，豊富，様々，新鮮，いろいろ，ダメ，正確，残念，大好き，大丈夫，無理，フル，意外，見事，豪華，便利，幸せ，複雑，豊か
新聞	平和，不明，緊急，積極，大幅，主要，疑問，名誉，明確，不正，慎重，不良，正式，公的，困難，確実，深刻，公式，有力，違法
共通	必要，可能，安全，明らか，特別，自由，新た，大量，重要，危険，不安，自然，健康，十分，安定，同様，主，大切，さまざま，完全，好き，独自，有効，普通，元気，非常，大事，大変，確か，簡単

続いて，表4と表5とに「形容詞」「動詞」を示す。形容詞はもともと語彙量が小さく，動詞もよく使われる語彙は限られてくるため，どちらもほとんどが雑誌と新聞に共通するという結果であった。そのためか，先に示した名詞ほど，雑誌と新聞との間に差異が感じにくい。強いて言えば，表4では「すごい」「かわいい」「欲しい」が，新聞より雑誌に確かに現れやすそうだと思われる。また表5では，「感じる」が雑誌らしく，「開く」「認める」が新聞らしいように思われる。

[2]表にまとめる際に同語は一つにまとめた。特に，活用語である形容詞，動詞は同語が多く抽出されたため，異なりが50語より少なくなっている。

表4:形容詞-自立

雑誌	すごい, 軽い, 小さい, 明るい, かわいい, おいしい, 古い, 速い, 欲しい
新聞	激しい, うれしい, 詳しい, 幅広い
共通	ない, 多い, いい, 強い, 高い, 大きい, よい, 厳しい, 少ない, 新しい, 長い, 早い, 悪い, 難しい, 若い, 近い, 深い, 重い, 低い, うまい, 楽しい, 白い, 美しい, 広い

表5:動詞-自立

雑誌	使う, 入れる, やる, つける, 加える, 教える, 食べる, 感じる, 合わせる, つく
新聞	よる, 受ける, 求める, 話す, 示す, 述べる, 決める, 写す, 分かる, 入る, 語る, 向ける, 開く, 認める
共通	する, なる, ある, できる, いう, みる, 言う, 思う, 行う, 出る, いる, 持つ, 思う, 考える, 出す, かける

最後に, 表6に「副詞」を示す。上位50語には共通するものが多かったが, 雑誌と新聞とを比較すると, 新聞の方に文体が硬いものに現れやすいと思われる語が出現している。この傾向については次の4.2節でも取り上げる。

表6: 副詞-一般・副詞-助詞類接続

雑誌	かなり, ずっと, とにかく, まさに, なんと, なかなか, やっぱり, ぜひ, もし, いっぱい, いろいろ, きっと, まるで
新聞	ほぼ, 改めて, 再び, ともに, かつて, 極めて, 突然, 次々, あまり, 少なくとも, はっきり, 一気に, むしろ, 一層, また
共通	どう, 初めて, さらに, そう, まだ, もう, 最も, ほとんど, なぜ, 特に, もっと, 実際, 少し, こう, 既に, すぐ, まず, すでに, 同時に, これから, よく, 当然, 本当に, いつも, わずか, 全く, しっかり, より, もちろん, とても, やはり, 必ず, そのまま, ちょっと, やや

4.2. 各資料により用例のとれ方に差がある語による比較

雑誌と新聞とにおいて, 用例のとれ方に差が出そうなものとして, 副詞に着目し, 雑誌に出現した全982語と, 新聞において使用度数10以上であった1,395語に対し, 同語異語判定を行い, 度数の大きさによる順位の比較を行った。これによって, 4.1節では, 順位の高いもの同士しか比較しなかったが, 全順位を比較することにより, 一方で高く, 一方で低い, という特徴ある語, すなわち, より雑誌で用例がとれやすい語, より新聞で用例がとれやすい語を抽出できる。前者の結果を表7に, 後者の結果を表8に示す。これにより, 4.1節で見られた, 新聞のみに現れた語には文体の硬い語が多いという傾向がさらに確認できる。なお, オノマトペは, 新聞に出にくく, 雑誌に出やすいと予想したが, 新聞1年分ものデータ量があると, 雑誌よりも多くのオノマトペの異なりが抽出された。しかし, 表7に「キラキラ」「ザーッ」があるように, 全般に雑誌に多く出現する傾向を示している。ただし, 中には, 表8に「バラバラ」があるように, 新聞に出やすいオノマトペもあることがわかった。

表7:より雑誌で用例がとれ得ると考えられる副詞

雑誌度数5以上, 新聞度数90未満	とっても, いちばん, キラキラ, しっとり, 要するに, ただ
雑誌度数3以上, 新聞度数10未満	バッチリ, ザーッ, さっき, ざっくり, ほっそり
そのほか, 新聞度数に比べ雑誌度数の多いもの	きちんと, かなり, なんと, やっぱり, きっと, ちょうど, 全然, いよいよ, いくら

表8:より新聞で用例がとれ得ると考えられる副詞

雑誌度数0, 新聞度数90以上	依然, 一段と, 依然として, 二度と, 再三, 急きょ, ほっと, よほど, ぎりぎり, 一方で, たびたび, ひときわ, 目の当たり, しみじみ, さほど, 万一, 案外, 一向に, 黙々と, 暗に, どうせ, もっぱら
雑誌度数1, 新聞度数90以上	心から, せめて, 精いっぱい, ふと, どれほど, ひたすら, バラバラ, りん, 早くから, いまだ, 一度に, まして, 断固, 到底, 公然, さながら, もとより, 騒然と, 思い通り
そのほか, 雑誌度数に比べ新聞度数の多いもの	改めて, ともに, かつて, 一層, ようやく, いかに, なお, いったん

4.3. 各資料ならではの用例がとれそうな語による比較

今回は, 雑誌ならではの用例がとれそうな, いわゆる若者語, はやり語, 口語的な語, 10語の用例数を調査した。調査語と, 得られた用例数の結果を表9に示す[3]。この調査では, 50誌全体の本文(広告や付録を除く)部分から用例を抽出した。いずれも雑誌の方で用例が多くとれ, 新聞ではほとんど用例がとれないことを予想して調べたが, 結果はほぼその予想通りであった。ただ, 新聞において全く用例のとれなかったものはなく, 1年分ものデータ量があれば, 少なからず用例のとれることが確認できた。10語のうちでは, 「プチ」が新聞にも用例が多かった。ほかの調査語では, 新聞の場合, 投書やコラムから用例がとれている場合が多かったが, 「プチ」では, 「プチ整形」の例が1面や社会面の記事に複数出現していたこともあり, 「プチ」全体で用例が多くとれる結果になった。

3 メタ的使用や, 曲名, 商品名, 書名, 番組名などの使用は除外した。

表9：若者語・はやり語・口語的な語の用例数

	いけてる	ゲット	はじける	へこむ
雑誌	13	166	17	19
新聞	6	21	5	4

	セレブ	ゴージャス	プチ	ラブラブ
雑誌	74	67	63	41
新聞	6	20	52	3

	っていうか	みたいな
雑誌	96	79
新聞	3	3

調査語について補足説明する。「いけてる」は，格好がいい，の意味の若者語である。「イケてる」や「イケテル」とも表記される。「ゲット」は入手する，の意味のはやり語である。「はじける」は若者語として元気のいいさまの意味で使用されている場合を調査対象にした。「ハジケる」とも書かれる。同様に，「へこむ」は若者語として気持ちが落ち込む意味で使用されている場合を調査対象にした。「へコむ」の表記の方が多かった。「セレブ」は有名人をさすはやり語であるが，特に海外のモデル・デザイナーなどファッション関係者や俳優・監督など映画関係者を指すことがある。時に「セレブな」の形で形容語にもなる。「ゴージャス」は，豪華，を意味するはやりの形容語。「プチ」は，ちょっと，の意味の語である。従来から使われていた「プチトマト」などは対象外にし，新たに「プチ」を用いて形容している「プチ家出」「プチぜいたく」のようなはやりの使用に着目し，その用例のみをとった。「ラブラブ」は，熱々，の意味の若者語である。最後の二つ「っていうか」「みたいな」は口語的な語である。「っていうか」は，文頭，文中，文末いずれで使用される場合も，すべてくだけた表現であるため，調査対象にした。「ってゆーか」とも表記される。一方，「みたいな」は，「ような」の意味で，「違うから，みたいな。」のように，文末で言い切る使用が特にくだけた表現と考え，調査対象にした。つまり，「ドラマみたいなコト」のように文中に現れる場合は対象外にした。

以下，雑誌と新聞から得られた用例の一部を示す。新聞の用例は，月日と面種と，朝刊か夕刊かを示す。

- **イケて**て，熱くて，旬でツウなGALになりたきゃ，じっくりご覧あそばせ!!(ポップティーン)
- 関西の音楽は結構**イケてる**(毎日:1104,社会,朝刊)
- ムリめの恋**ゲット**に大切なコト(My Birthday)
- 60センチ級のメジナを**ゲット**してニッコリするY氏(毎日:0129,総合,夕刊)
- あと，おばさんでも気持ちは若い**はじけ**たおばさんになっていて，「いつまでも若いね」と言われるようになっていたら，いいなーと思う今日このごろです。(ひよこクラブ)
- 「30代最後なので，**はじけ**てみました」(毎日:0826,文化,夕刊)
- これには正直**へコみ**ました。(パチスロ必勝ガイド)
- うまくいかなくて**へこむ**と，口をへの字に曲げる。(毎日:0224,家庭,朝刊)
- **ゴージャス**気分になれる，王冠リング。(mini)
- ぱっと見**ゴージャス**シーフード(毎日:0423,家庭,朝刊)
- **セレブ**なワンピで東洋系ハリウッド女優目指した！(ポップティーン)
- 気分は**セレブ**(毎日:0814,総合,夕刊)
- あっという間の**プチ**おかず。(dancyu)
- これも一種の**プチ**整形か(毎日:0226,1面,朝刊)
- 秋祭りは浴衣で**ラブラブ**。(Animage)
- 娘たちには気持ち悪がられるけれど，これからも**ラブラブ**でいましょうね。(毎日:1216,1面,夕刊)
- **っていうか**，とんこつ以外食べません！(月刊ジャイアンツ)
- 覚悟**っていうか**，プロ意識が伝わって。(毎日:0726,特集,朝刊)
- “いや，もう1回”**みたいな**(笑)。(BACKSTAGE PASS)
- PKOって何？**みたいな**。(毎日:0527,総合,夕刊)

5. まとめ

「バランストコーパス」の構築を目指すとき，資料の種類によって，用例収集の多様性を明らかにすることが欠かせない。そのために，従来の文体差を調査する手法以外の方法を探ることを試みた。今回，使用度数上位語による比較，使用度数の順位に差異のある語の比較，着目語の用例数の比較を行った。これによって，実際にどのような語の用例が雑誌，新聞でとれやすいのか，あるいは，とれにくいのか，ということを示すことができたと考える。しかしながら，この3つの手法，サンプリング，同語異語判定などの妥当性にまだ検討の余地が多くある。今回は順位に注目し，度数そのものについての議論は省略したが，今後は度数にも考慮した，定量的な考察を進めていきたい。

参考文献

後藤斉(1995)“言語研究のデータとしてのコーパスの概念について ―日本語のコーパス言語学のために―”東北大学言語学論集,第4号,71-87.

陳志文(2004)“週刊誌に見られる文体の類型―主成分分析法を通して―”計量国語学,Vol.24,No.6,308-319.

茂木俊伸，山口昌也，丸山岳彦，田中牧郎(2005)“語種辞書『かたりぐさ』の開発と月刊雑誌の語種構成分析”言語処理学会第11回年次大会発表論文集.

安本美典，本多正久(1981) 因子分析法，培風館.