第4回　コーパス日本語学ワークショップ　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

第4回 コーパス日本語学ワークショップ

予稿集

2013年9月5日、9月6日

主催　国立国語研究所 言語資源研究系・コーパス開発センター

会場　国立国語研究所

第4回 コーパス日本語学ワークショップ

予稿集

2013年9月5日(木)／9月6日(金)

Program [プログラム]

9月5日（木）

10:00～10:10　**■挨　拶**　前川 喜久雄

10:10～12:10　**■口頭発表（1）**

コーパスを利用した言葉の意味・用法の変化の研究 ―「敷居が高い」を例に―
▷佐々木 文彦

ニ格とデ格の交替について
▷張 麗

コーパスに基づく日本語擬態語動詞の意味分析
▷菅原 崇、浜野 祥子

条件構文の談話標識化の諸相
▷藤井 聖子

12:10～13:10　**昼食・休憩**

13:10～14:10　**■ポスター発表（1） Aグループ**

接続助詞「から」と「ので」に関する一考察 ―前件のモダリティとの共起を手掛かりにして―
▷李 恵正

BCCWJ教科書データより抽出した頻度情報に基づく日本語ライティング指導教材の作成
▷堀 一成、坂尻 彰宏、石島 悌

接続助詞「けど」の音調と意味用法に関する研究 ―挿入用法についての検討―
▷田頭 未希

『太陽コーパス』における漢語表記の多様性 ―コーパスのXMLタグを利用した研究手法の試み―
▷間淵 洋子

会話コーパスの転記方式の相互変換 ―引き伸ばしに着目して―
▷土屋 智行、伝 康晴、小磯 花絵

弱境界における発話計画に関わる音声的・言語的特徴の分析
▷小磯 花絵、伝 康晴

会話コーパスの転記方式の相互変換
―言語・音響特徴を用いた会話分析方式の音調マーカーの導出―
▷石本 祐一、土屋 智行、小磯 花絵、伝 康晴

14:10～15:10　**■ポスター発表（1） Bグループ**

コーパスを用いた外来語サ変動詞の分析 ―「マークする」を例として―
▷茂木 俊伸

現代日本語における汎用的漢語サ変動詞の抽出とその内部構成の検討
▷李 楓

「一方」という形式に見られる「する」と「やる」の差異について
▷森川 結花、小山 宣子、浜田 秀

文体から見た『今昔物語集』の語彙 ―『日本語歴史コーパス 平安時代編』と比較して―
▷田中 牧郎

『今昔物語集』のテキスト整形
▷冨士池 優美、河瀬 彰宏、野田 高広、岩崎 瑠莉恵

『近代女性雑誌コーパス』の小説会話部分に現れる一・二人称代名詞の計量的分析
▷近藤 明日子

近世口語資料の形態素解析の試み
▷小木曽 智信、市村 太郎、鴻野 知暁

日本語連用形名詞の自立性の段階について
▷沈 晨

日本語話し言葉コーパスを用いた対話音声のイントネーション句の分析
▷石本 祐一、小磯 花絵

15:10～17:10　**■口頭発表（2）**

中古における接続語の使用傾向について

▷岡崎 友子

「ガ／ノ」交替現象についての一考察 ―古代・現代コーパスを対照して―

▷坂野 収

「五国史」宣命のコーパス化

▷池田 幸恵、須永 哲矢

漢語名詞の副詞用法 ～『現代日本語書き言葉均衡コーパス』『太陽コーパス』を用いて～

▷髙橋 圭子、東泉 裕子

9月6日（金）

10:00～12:00　**■口頭発表（3）**

事象の活性化と不活性化を把握する言語資源の構築とその応用
―災害時における問題報告と支援情報のマッチングを例に―

▷佐野 大樹、イシュトバーン・ヴァルガ、鳥澤 健太郎、橋本 力、川田 拓也、呉 鍾勲、大竹 清敬

テキスト関連属性と助詞選択：計量的アプローチに基づく探索的研究
―主語・主題を導く「は」と「が」をめぐって―

▷石川 慎一郎

文節係り受け木の根の構造について

▷高松 亮

〈名詞句＋係助詞〉の格

▷山田 昌裕

12:00～13:00　**昼食・休憩**

13:00～14:00　**■ポスター発表（2） A グループ**

日本語学習者のための名詞と修飾語のコロケーション検索プログラムの開発とその使用例

▷中溝 朋子、坂井 美恵子、金森 由美、大岩 幸太郎、刈谷 丈治

外来語語末長音の表記のゆれについて

▷小椋 秀樹

コーパスコンコーダンサ『ChaKi.NET』の連続値データ型

▷浅原 正幸、森田 敏生

日本語名詞述語文の意味関係アノテーション

▷今田 水穂

コロケーションとシンタクス ―形容詞と名詞のコロケーションを対象に―

▷スルダノヴィッチ・イレーナ

『現代日本語書き言葉均衡コーパス』「図書館書籍」の生年代別分布は何を表しているのか
―「デナイ」「デハナイ」「ジャナイ」の使用割合から見た一考察―

▷森 秀明

現代日本語書き言葉における非外来語のカタカナ表記事情

▷柏野 和佳子、中村 壮範

言語資料としての「判決文」の分析にまつわる問題点

▷矢野 信

現代日本語の従属節に現れるモダリティ形式の分布

▷丸山 岳彦

14:00～15:00　**■ポスター発表（2） B グループ**

クラスタリングを利用した能動学習による語義曖昧性解消の領域適応

▷小野寺 喜行、新納 浩幸

語義曖昧性解消の領域適応における Misleading データの存在と検出

▷吉田 拓夢、新納 浩幸

日伊コロケーション辞書の作成を目指す『現代日本語書き言葉均衡コーパス』からの
コロケーションの検出と分析
▷ STRAFELLA Elga Laura、松本 裕治
NDL Searchによるジャンル名の分析
▷浜田 秀
社会科教科書における教科特徴動詞の用法
―教科書コーパスと図書館コーパスの比較を通して―
▷阿保 きみ枝
「ベテランは足を保護する」が語りかけるとき
▷保田 祥、立花 幸子、柏野 和佳子、丸山 岳彦
多義複合動詞の語義構造の分析
▷山口 昌也
コーパスを活用した法文データの分析に関する問題点
▷矢野 信
同一見出し語の出現間隔の分布と文体差
▷山崎 誠

15:00～17:00　**■指定討論・全体討論**
17:00　**■閉　会**

Contents [目次]

口頭発表（1）

9月5日（木）　10:10～12:10

コーパスを利用した言葉の意味・用法の変化の研究
—「敷居が高い」を例に—

佐々木 文彦（明海大学外国語学部日本語学科）†

A Corpus-based Study on the Change of Meaning and Usage of Words: A Case Study of "*shikii ga takai*"

Fumihiko Sasaki (Faculty of Languages and Cultures, Meikai University)

１．はじめに

平成 20 年度の「国語に関する世論調査」で「敷居が高い」を「本来の使い方とは違う」意味で理解していた人は「本来の使い方」の意味で理解している人を上回ったと報告されている。「誤用」の使用率が「正用」を上回るのであれば、これは既に「誤用」と呼ぶべきではなく、「変化」ととらえてよいと考える。

意味変化と「誤用」の関係については新野(2011)所収の諸論考および国立国語研究所共同研究プロジェクト「近現代日本語における新語・新用法の研究」(プロジェクトリーダー新野直哉)で展開されている諸研究の中で「気づかない変化」として扱われ、興味深い考察がなされているが、本研究も同様に誤用を言語変化の事例ととらえる立場から語の用例を観察、分析しようとするものである。

「敷居が高い」は江戸時代から見られる慣用句であるが、これがどのように用法を変化させて今日に至ったのか、各種コーパス[1]を利用し、江戸時代から現代までの出現数・出現率の変化と実際の用法の変化を観察し、「コーパスを利用した言葉の意味・用法の変化の研究」の可能性と問題点を考察する。

２．誤用の指摘

２.１　世論調査の報告

文化庁月報 No.511(平成 23 年 4 月号)[2]は、平成 20 年度の「国語に関する世論調査」の結果をふまえて、「敷居が高い」の使い方について問答形式で次のように説明している。

> 問 1　「あのレストランは高級すぎて敷居が高いよ。」と言ったら，その使い方はちょっとおかしいと指摘されました。「敷居が高い」の本来の使い方を教えてください。
> 答　「敷居が高い」は，もともと，不義理や面目の立たないことがあって，その人の家に行きにくい，という意味で使われていました。
>
> 問 2　「敷居が高い」について尋ねた「国語に関する世論調査」の結果を教えてください。
> 答　30 代以下の世代では，「高級過ぎたり，上品過ぎたりして，入りにくい」の意味で「敷居が高い」を使う人が多くなっています。
>
> (以上問答部分のみ抜粋)

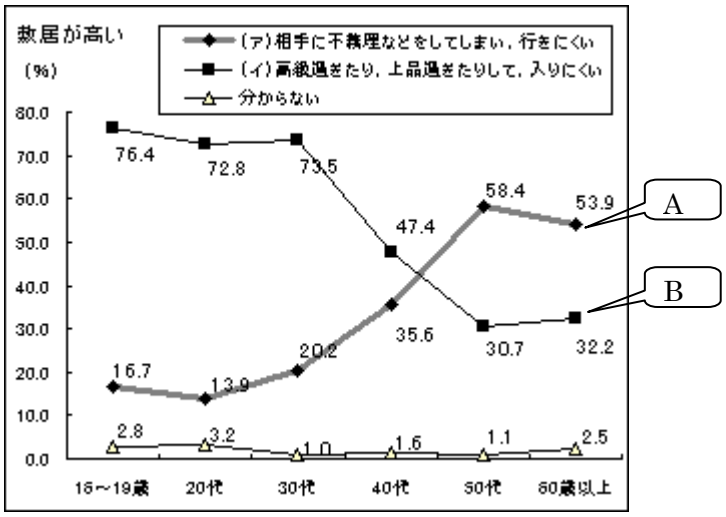
† sasaki-f@meikai.ac.jp

1 『大系本文データベース(国文学研究資料館)』『現代日本語書き言葉均衡コーパス』「各種小説のデータベース」(『青空文庫』『新潮文庫の 100 冊 CD-ROM 版』『カッパノベルス』)「新聞記事データベース」(読売・朝日・日経)「Google Trends」等

2 http://www.bunka.go.jp/publish/bunkachou_geppou/2011_04/series_08/series_08.html

A「不義理や面目の立たないことがあって、その人の家に行きにくい」
B「高級過ぎたり、上品過ぎたりして、入りにくい」

この報告の主旨をまとめると、「敷居が高い」は、本来は A の意味であったが、B の意味で使われるようになってきたというのである。この「問 2」の答の根拠となっているのが次に示すグラフ 1 である[3]。年齢が上の世代から若い世代に向かって(グラフを右から左に向かって)見ていくと、40 代を境に A(ア)から B(イ)に逆転している様子がわかる。次節で見るように、B の用法は辞書や日本語本では「誤用」とされるものであるが、7 割以上の人が B の意味で使っているのだとしたら、これはもはや誤用ではなく「新用法」と言っていいのではないか、という疑問がこの研究の始まりである。



グラフ 1 平成 20 年度「国語に関する世論調査」

２.２　辞書の記述、いわゆる日本語本その他の記述

国語辞典にはもっぱら A の意味のみが示されているのである[4]が、『明鏡国語辞典　第二版』には、「注意」として

程度や難度が高い意で使うのは誤り。「×高級過ぎて僕らには―店」「×初心者には―ゴルフコース」

という参考情報が示されている。これは 2003 年 12 月の初版には記されておらず、2011 年 4 月(平成 23 年)の第二版で追加された情報である。

いわゆる日本語本では「日経おとなの OFF」(2012)に次のような記述がある。

×高級な店は敷居が高い
○先生にご無沙汰して、敷居が高くなった
不義理をしていて、その人の家に入りにくい、合わせる顔がないという意味。程度や難度、格が高くて入りにくいの意ではない。p18

[3]前掲、文化庁月報 No.511web サイトより転載。以下、便宜上(ア)を A、(イ)を B と読み替えることとする。

[4] 『日本国語大辞典(第二版)』『新明解国語辞典第七版』『岩波国語辞典第七版新版』

また、植松(2010)には「『敷居が高い』というと、過去に何があったのかと思われますよ」という項目があり、

> 「値段が高い」とか「高級過ぎる」という場合には、「敷居が高い」という表現は使いません。そういう場合は、「ハードルが高い」などが適当な表現かもしれません。p19

と説明され、いずれも意味 B「高級過ぎたり，上品過ぎたりして，入りにくい」は間違っていると指摘している。

これに対して、山下(2007)は、「着物にはいろいろな決まりごとがあって，敷居が高いように感じる」という広告の表現を元に、

> 私の感覚では，「入りにくい」という意味で，「あの店は高級で敷居が高い」とは言う。しかし，実際の建物でもない「着物」という「モノ」に対して，「とっつきにくい」といった意味で「敷居が高い」を使うのは少しおかしいように思う。

と指摘し、

> 個人的には「○○は敷居が高い」は，ことばからすぐ映像が浮かび，便利な表現だと思う。反対に「入りやすい」ことを「敷居がない（低い）」と言うなど造語力もある。新しい意味もかなり定着しており，辞書どおりの伝統的な意味しか認めないというのは少し納得がいかない。

との見解を示している。つまり、「モノ」に対してまで B の意味を広げるのには反対だが、少なくとも店などの場所については B の意味を認めるべきであるとの立場である。

また、「日経ネット Plus」(2009)には次のような報告がある。

> 新聞・通信・放送各社が加盟する関西地区新聞用語懇談会が行った調査では、「敷居が高い」という表現を記事や放送で単に「行きにくい」の意味で使っていると回答した社が 13 社だったのに対し、使っていないとの答えは 3 社だけでした。新聞や放送でもこの使い方が定着してきたといえそうです。一方で各社からの意見の中には、「敷居が高い」は「行きにくい」の意味を離れて「難しい」「ハードルが高い」といった文脈で使うとの指摘もありました。
>
> 日本経済新聞の記事でも、たとえば「インターネットに不慣れだと外国為替証拠金取引は敷居が高い」などと使っています。近年では「敷居が高い」の反対の意味で「民事再生法は会社更生法よりも敷居が低い」といった表現も登場するようになり、さらに変化した形として「ゴルフの敷居を下げてプレーヤーを増やす」などの言い方も散見されます。敷居は上げたり下げたりするものではないため、ここまで来ると違和感を持つ人が少なくないかもしれません。

「違和感を持つ人が少なくないかも」とは言うものの、2009 年の時点で多くの新聞が「敷居が高い」を B の意味で用いており、しかも山下(2007)では「少しおかしい」と評されていた「モノ」に対して表現する用法も広がりを見せていることが伺える。

２．３　正用と誤用の違い

ここまで見てきたことを整理すると、「敷居が高い」には A・B の意味の他に、

> C「とっつきにくい、難しい、ハードルが高い」(モノに対する用法)

の意味もあり、次の 3 種類の立場があると言える。

1) A のみを認める立場　：国語辞典・日本語本
2) AB を認める立場　：山下(2007)
3) ABC すべてを認める立場　：「日経ネット Plus」(2009)

これら ABC の違いを次の表 1 のように分析してみる。

	ある場所に行く(入る)	自分に原因がある	原因の中身
A	○	○	不義理・不面目
B	○	×	高級・上品
C	×	×	難度・格・繁雑

表 1　ABC の意味成分

A→B→C のように用法が拡大したのだとすると、

(A)「自分が作り出した不義理・不面目のために負い目を感じてどこかに行きにくい」

という意味の「原因の作り手」の部分が変化して、

(B)「場所そのものの持つ高級感や上品さが原因で劣等感を感じて店などに入りにくい」

のような意味になり、さらに「敷居」によって示される空間の境界が単なる心理的障壁のみの意味に変化して

(C)「難度・格などが原因でとっつきにくく敬遠される」

のような順で抽象化されたものと考えられる。
それでは、このような変化はいつごろからどのように生じたのだろうか。次節では各種のコーパスを用いて「敷居が高い」の使用実態を観察することにする。
また、C には「ハードルが高い」との共通性が示されているが、「敷居が高い」と「ハードルが高い」は同じなのか違うのか、これも用例を見ながら考察することにする。

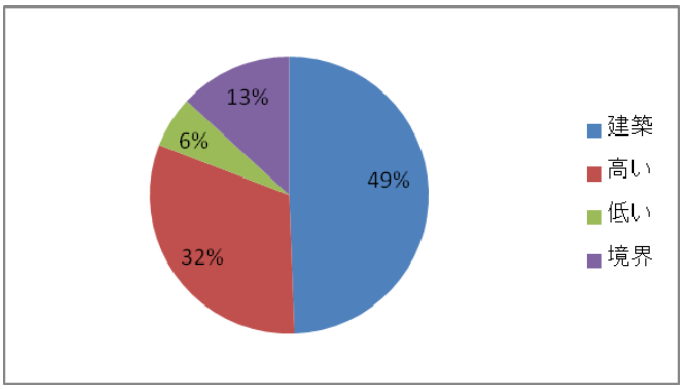
３.「敷居が高い」の用例

３.１　BCCWJ の用例

まず BCCWJ を用いて現代の「敷居」の用例を観察する。BCCWJ において、「敷居」の例は 251 例[5]ある。そのうち(1)のような建物の「敷居」そのものをあらわす「建築」用語としての用例(124 例)、「敷居が高い」の例(79 例)、「敷居が低い」の例(15 例)、そのほかに、(2)のような慣用表現を含め、「敷居」を単なる家の構造部分の名称としてではなく、内と外とを隔てる境界を表現するために用いた例を「境界」(33 例) として分類して使用率を比較したのが次のグラフ 2 である。

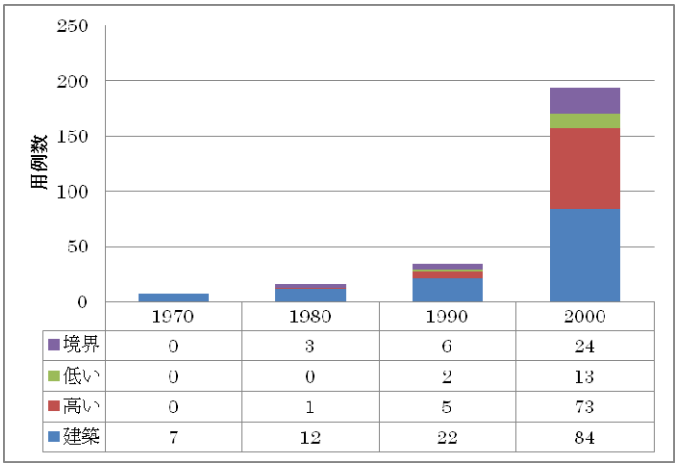
(1) 山田さん宅は、既存の廊下の床天と他の部屋の入口部の敷居との段差が現在一五ミリついています。(佐藤謙一『おじいちゃん・おばあちゃんのための(秘)リフォーム』)
(2) 嘘で塗り固められた手紙をまとめて送り返し、二度とこの家の敷居をまたぐなと宣告します。(歌野晶午『ブードゥー・チャイルド』)

[5] メタ言語的な例は除いてある。以下同様。



グラフ 2　BCCWJ の「敷居」の使用比率

BCCWJ の用例は 1976 年から 2008 年までのものであるが、1970 年代から 2000 年代にかけて、この比率がどのように変化したか、その推移を示すのが次のグラフ 3 である。



	1970	1980	1990	2000
境界	0	3	6	24
低い	0	0	2	13
高い	0	1	5	73
建築	7	12	22	84

グラフ 3　「敷居」の用例数の変化

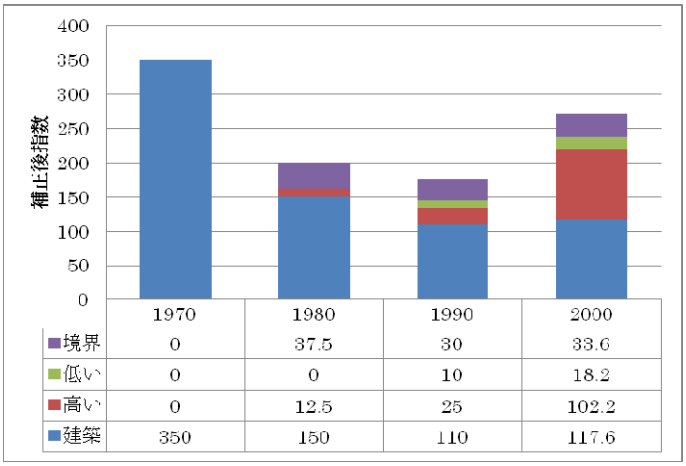
グラフ 3 を見る限り、1970 年代には「建築」(構造物の名称としての用例)しかなかったが、年代を追うごとに増加し、「敷居が高い」という言い方が 80 年代から見え始めて 2000 年代に急増したように見える。けれども BCCWJ はすべての年代にわたって均一に言語データが分布しているわけではないと考えられるので、次の方法で補正を試みた。

用例数の補正の手順[6]

1) 一つの基本語彙を各年代別に検索し、その用例数の全体に対する割合を算出する。
2) いくつかの基本語彙について同様に算出して平均値を求める。
3) 2)で求めた平均値の逆数を求め、「補正係数」とし、これを用例数に乗じて補正する。

6 基本語彙とその用例数、補正係数は次の通り。

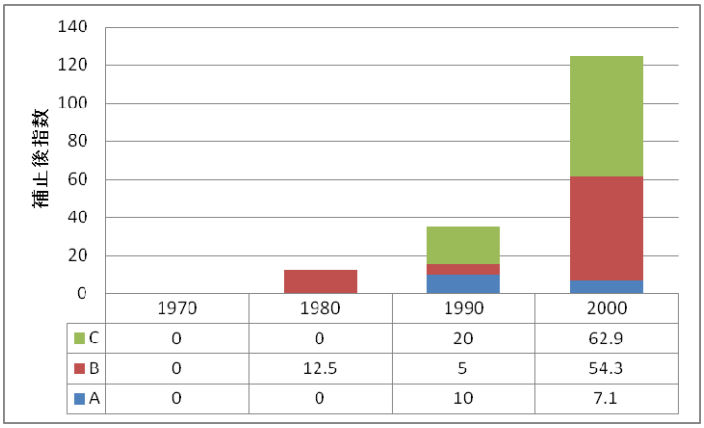
検索語	話		山		人		男		女		年代別
年代	用例数	割合	用例数	割合	用例数	割合	用例数	割合	用例数	割合	補正係数
1970	2291	2%	1622	2%	10873	2%	1286	2%	2335	2%	50.0
1980	9407	7%	8385	8%	48441	7%	5947	8%	10749	7%	12.5
1990	28012	21%	21829	22%	140295	21%	17721	23%	36022	24%	5.0
2000	93821	70%	69663	69%	455610	70%	52655	68%	99406	67%	1.4



グラフ 4　補正後の使用比率変化のイメージ

グラフ 4 を見ると、「建築」の例は年代を追うごとに減少し、「敷居が高い」「敷居が低い」の例は 80 年代から徐々に増え、「境界」の例は 80 年代から 2000 年代までほぼ変化なく出現していることがわかる。

次に、前節で見た ABC の比率の変化を見ると、グラフ 5 のようになっており、「敷居が高い」の本来の用法である A(不義理、不面目)は 90 年代から 2000 年代にかけて減少し、B(高級過ぎたり、上品過ぎたりして、入りにくい)や C(とっつきにくい、難しい)の例は増加していることがわかる。



グラフ 5　ABC の比率の変化(補正後)

３.２　小説等文学作品の用例

BCCWJ によって調査できる用例は 1970 年代以降のものに限られるが、そもそも「敷居が高い」という慣用句は江戸時代から用いられており、近世から現代にかけてどのように変化してきたか観察する必要がある。

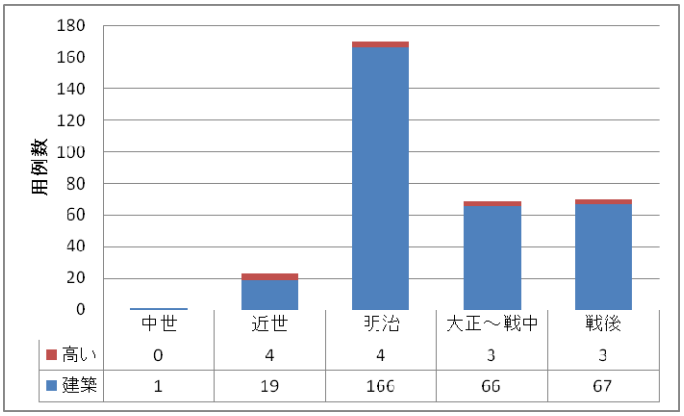
文学作品を対象とする用例を次の各種データベースを用いて検索した。

1) 大系本文データベース(国文学研究資料館)[7]
2) 『CD-ROM 版 新潮文庫の 100 冊』『CD-ROM 版 新潮文庫 明治の文豪』
3) 「青空文庫」の諸作品
4) 「カッパノベルス」の諸作品(138 冊)

上記 1)については上古・中古・中世・近世の時代区分で分け、2)～4)については明治・大

[7] http://base3.nijl.ac.jp/を利用

正～戦中・戦後の 3 つに分け、時代ごとの用例[8]を検索した。



グラフ 6 文学作品における敷居の変遷

その結果がグラフ 6 である。建物の敷居そのものを表す例は明治期をピークとして、現代にも用いられているが、慣用句としての「敷居が高い」は多くなく、データの数から変化の動向を観察するのは難しい。ABC の内訳を表にすると次の表 2 のようになる。

	A	B	C
近世	4	0	0
明治	4	0	0
大正～戦中	3	0	0
戦後	2	1	0

表 1「敷居が高い」の意味

「敷居が高い」の用法はほぼ本来の用法 A に限られ、わずかに B の例として次のものが見られるのみである。

(3) 時折り妻や息子に、本部まで、必要品を届けさせていたが、家族の者も、殺人事件の捜査本部というと、何となく敷居を高く感じるらしく、あまり来たがらない。 (森村誠一『超高層ホテル殺人事件』1971/7)

３.３　新聞の用例

３．３．１　全文検索の用例

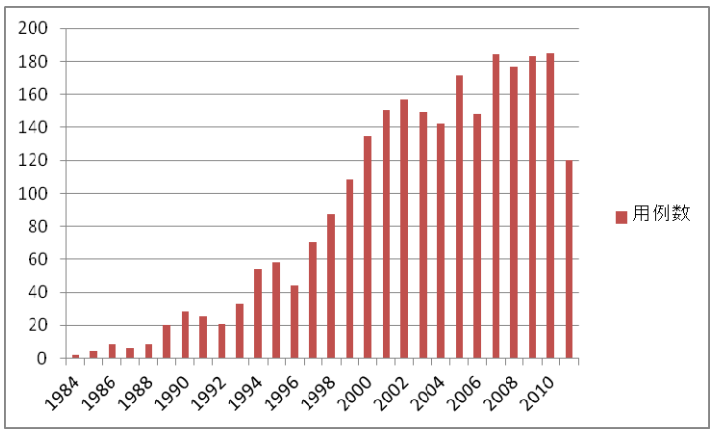
新聞の用例については、朝日・読売・毎日・日経等の各紙のデータベースサイトにおいて、1980 年代以降については全文検索が可能であり、見出しとキーワードによる検索は明治期の記事までさかのぼることができる。

朝日新聞の全文検索で「敷居」を検索すると、1984 年から 2012 年 10 月までの期間の全用例は 2584 例であるが、建物の敷居そのものを表す例はわずかに 20 例であり、その他はすべて「敷居が高い」「敷居が低い」「敷居を下げる」など、慣用句「敷居が高い」を元とする表現である。

[8]各時代区分別の作品数は下記のとおりである。

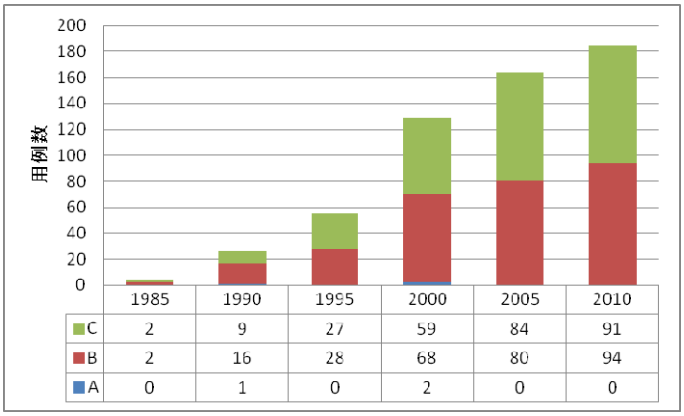
上古	中古	中世	近世	明治	大正～戦中	戦後
29	79	296	176	431	1404	237

朝日新聞における「敷居が高い」の類の用例数の変遷を示すのがグラフ 7 である。



グラフ 7 朝日新聞における「敷居が高い」類の用例数

1980 年代ではほとんど現れないが、90 年ごろからどんどん増加しているのがわかる。1985 年から 2010 年まで、前節と同様に ABC の内訳を 5 年おきに観察して分類したのが次のグラフ 8 である。



	1985	1990	1995	2000	2005	2010
C	2	9	27	59	84	91
B	2	16	28	68	80	94
A	0	1	0	2	0	0

グラフ 8 朝日新聞における「敷居が高い」類の意味

「敷居が高い」類の用例が増加する中で、グラフ 8 の期間に関する限り、本来の用法で用いられているのは次のような例のみであり、むしろ辞書や日本語本で「誤用」とされる例がほとんどであるということがわかる。

(4) ドラマは、家出していた息子（渡辺）と 11 年ぶりに再会した母（李麗仙）との情愛を描く物語で、「男が敷居の高かった家に戻る時の気持ちが、いまの僕の心境とオーバーラップしているが、 (朝日 1990/9/14 夕刊)

３．３．２　キーワード検索の用例

1980 年代以前のキーワード検索を用いて「敷居が高い」がいつごろ A 以外の意味で用いられるようになったか、探してみると、管見の範囲では次の(5)の例が最も古いようである。

(5) 大河内一男東大学長は「都の窓口は敷居が高すぎる。都民がたやすく相談できて、すぐ返事のもらえる相談所をあちこちに設けてほしい」とサービス精神の徹底をうながす。
(朝日 1964/2/8「敷居が高い都の窓口」)

この(5)の例は、都民にとって都の窓口が行きにくい、入りにくいという例であるから B の意味で用いられた例と考えられる。そして、次の(6)の例は、「資格制限」によって入居が制限されるということを「敷居が高い」と表現している例で C にあたる。

(6)「月収十五万円以上ないと、日本住宅公団の新築賃貸住宅にはいる資格はありません」。こんな新しい「資格制限」を、公団が六月から実施していることがわかり、低所得者の入居希望者から「貧乏人締め出しだ」と抗議が出ている。(朝日 1975/10/7「敷居が高くなったー公団の新築賃貸入居資格ー」)

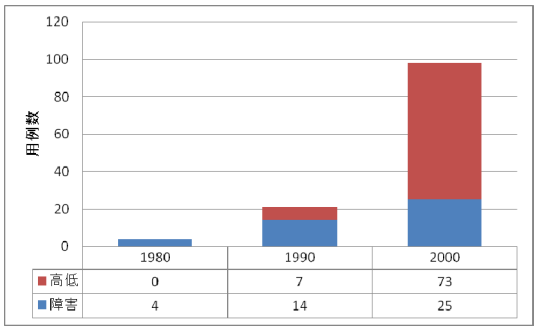
これより少し早く、読売新聞 1975/4/1 朝刊にも「ある閉園ー“補助も敷居が高い”ーミニ保育園に冷たい壁」という見出しで、無認可保育園が補助金の条件が厳しくなったために閉園せざるを得なくなったという記事があり、このころから「敷居が高い」が「条件や制限が厳しい」という意味、つまり C の意味で用いられるようになったことが伺える。ただし、この時期は新聞記事の全文検索が出来ない時期なので、初出例をつきとめたり用法の変遷を観察したりすることは困難である。

４．　ハードルの用例

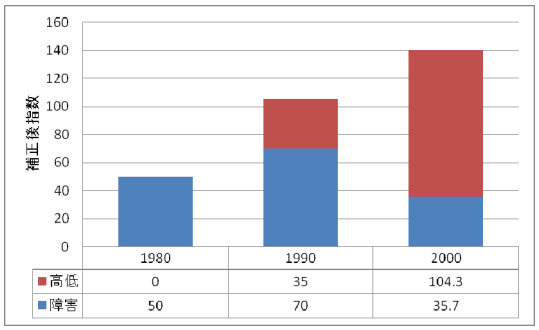
２．３で見たように、「敷居が高い」の C の意味は「ハードルが高い」と同様の意味であるとされることがあるが、「ハードルが高い」についても BCCWJ および新聞で検索してみた。

４．１．１　BCCWJ の用例

BCCWJ で「ハードル」を検索してヒットする用例のうち、陸上競技等の文字通りのハードルを除外すると 143 例ある。このうち「ハードルが高い(低い)」「ハードルを上げる(下げる)」等の慣用句の例が 100 例、「たくさんのハードルが立ちはだかる」のように「ハードル」を「障害」「障壁」を意味する比喩表現として用いる例は 43 例である。この内訳の年代による推移とその補正を行った指数のグラフが次のグラフ 9、10 である。



グラフ 9 ハードル用例グラフ



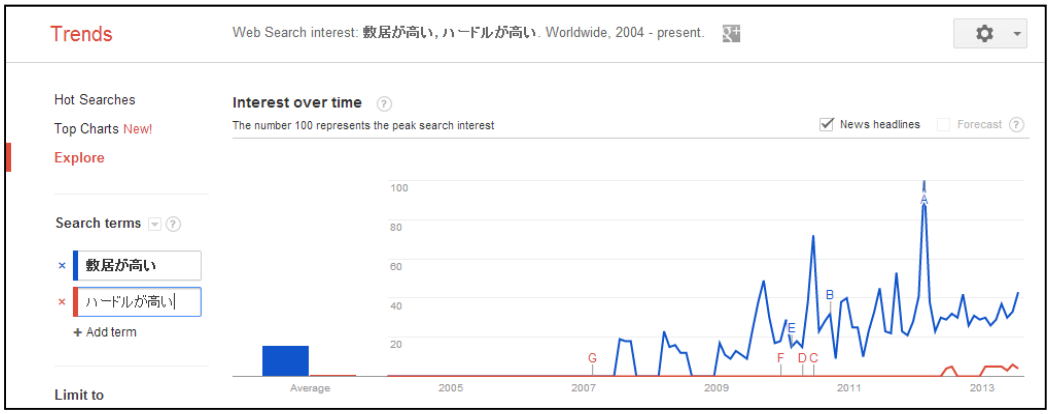
グラフ 10 ハードル補正後指数

これを見ると、時代を追うごとに障害や障壁としての「ハードル」の用例が増加し、「ハードルが高い(低い)」などの慣用句の用例も増えていることがわかる。

４．１．２　Google Trends を利用した用例数変化の比較

Google Trends で「敷居が高い」と「ハードルが高い」を検索して使用頻度を調べた結果、次のグラフ 11 の結果を得た。

「敷居が高い」は 2007 年から例が増え始め、2012 年ごろにピークを示している。これに対して「ハードルが高い」は 2013 に入るあたりで少々の用例を見るが、「敷居が高い」に比べると非常に低い数値を示している。これを、新聞の用例数の推移等とどのように関連づけて考えるべきかは今後の課題である。



グラフ 11 Google Trends による「敷居が高い」「ハードルが高い」

５．まとめ

以上、「敷居が高い」の使用実態を各種コーパスを利用して概観した。小説の例を見ても新聞の例を見ても、「敷居」は「敷居が高い(低い)」等の慣用句としての用法が増加しており、しかもその「本来の意味」である A「不義理・不面目」よりもむしろ B「上品で入りにくい」、C「とっつきにくい」の例の方が優勢となっていることがわかった。

なぜそのような変化が起きているのか、「ハードルが高い」とはどう違うのか、など、用例の前後の文脈を把握しながらさらに詳細な分析をする必要があるが、今回行ったようにさまざまなコーパスを用いて用例数の動向を観察するだけでも、辞書や日本語本で「誤用」と指摘される用法がむしろ多数派の用法として定着しつつあることが見て取れるのである。

付　記

本発表は、2012 年度宮田研究奨励特別研究「日本語語彙体系の史的変遷に関する研究」の研究成果の一部である。

文　献・資　料

植松眞人(2010)『センパイ！　その日本語まちがってます』、保育社
新野直哉(2011)『現代日本語における進行中の変化の研究―「誤用」「気づかない変化」を中心に』、ひつじ書房
日経おとなの OFF(2012)『美しい日本語と正しい敬語が身に付く本』、日経 BP 社
「日経ネット Plus」(2009)「『敷居が高い』変化する意味と形」2009 年 8 月 21 日掲載
http://www.nikkei.com/article/DGXZZO06311970V20C10A4000000/
山下洋子(2007)「敷居が高い」放送研究と調査 2007 年 11 月号、NHK 放送文化研究所、p69
http://www.nhk.or.jp/bunken/summary/kotoba/kotobax3/pdf/031.pdf

関連 URL

青空文庫　http://www.aozora.gr.jp/
大系本文データベース(国文学研究資料館)　http://base3.nijl.ac.jp/
現代日本語書き言葉均衡コーパス(BCCWJ)　http://www.ninjal.ac.jp/corpus_center/bccwj/
聞蔵 II ビジュアル朝日新聞記事データベース　http://database.asahi.com/library2/
ヨミダス歴史館・ヨミダス文書館　https://database.yomiuri.co.jp/rekishikan/
日経テレコン　http://t21.nikkei.co.jp/g3/CMN0F12.do
文化庁月報平成 23 年 4 月号(No.511)
http://www.bunka.go.jp/publish/bunkachou_geppou/2011_04/series_08/series_08.html

ニ格とデ格の交替について

張 麗（大東文化大学）

The Alternation of Auxiliary Word Ni and DE

Zhang Li(Daito Bunka University)

1. はじめに

（1）～（4）が示すように、「とる」「もつ」「かかえる」「だく」のような動詞は格体制の交替（～ヲ～ニ形と～ヲ～デ形）を起こす。

（1a）彼女はグラスを手にとり、一口飲んでみた。
（海老沢泰久『男ともだち』講談社 1998）

（1b）茶碗を右手でとり、左手で扱って、右手で勝手付に仮置きする。
（千宗左『小棚の点前』主婦の友社 1990）

（2a）すると、雑誌を手にもって農家の人が大勢たずねてくるようになった。
（横森正樹『夢の百姓』白日社 2002）

（2b）ときどき六寸ぐらいある基盤を片手でもって、五十匁蝋燭の火を団扇のように煽り消したそうです。
（甲野善紀等『オール讀物』文芸春秋 2004）

（3a）C 男がにわとりを小脇にかかえてしじゅう持ち歩く姿は、幼稚園のなかではいやでも目に付いた。
（友定啓子『子どもの心を支える』勁草書房 1999）

（3b）彼はほとんど沸き返るようなコーヒーの入った陶製のカップを両手でかかえ、その上に身をかがめて、湯気の芳香楽しみながらすすった。
（ウィルバー・スミス著　飯島宏訳『飢えた海』文藝春秋 1995）

（4a）五歳くらいの男の子をうでにだいて、わらっている。
（朝比奈蓉子『へそまがりパパに花たば』ポプラ社 1992）

（4b）ひざを両手でだきながら、『おれ、ここに、いる』と、口をぱくぱくうごかして、
（浜田糸衛『金の環の少年』国土社 1987）

本稿では、「とる」「もつ」「かかえる」「だく」四つのふれあい動詞のニ格とデ格の交替を考察する。

2. 先行研究

2.1 ニ格・デ格についての先行研究

国語学、日本語学からニ格・デ格の分類について、国立国語研究所（1951）、言語学研究会（1983）、益岡・田窪（1987）が挙げられる。第二言語習得から、迫田（1998）などの研究がある。認知言語学から、山梨（1995）、菅井（1997）（2000）、森山（2005）（2008）が挙げられる。

2.2 壁塗り構文のニ格とデ格の交替

格の交替現象が起こるデ格とニ格を中心とする先行研究は川野（2012）などがある。

今までニ格、デ格についての研究は主にニ格、デ格それぞれの分類を中心としている。

更に、場所を表すニ格とデ格の混同やニ格の過剰使用についての研究もなされている。ニ格とデ格の交替についての論文は少し見られたが、主に壁塗り構文の交替に焦点を置き、(1)～（4）のようなふれあい動詞のニ格とデ格の交替についての研究は、管見の限り少ない。「とる」「もつ」「かかえる」「だく」はニ格とデ格の交替が可能だと言われるが、それぞれニ格とデ格の使用率はまだ明らかにされていない。先行研究（言語学研究会 1983：309）ではに格の名詞は主に身体の部分（とくに手）をしめすものであると指摘しているが、「手に～」以外にどんな表現があるかまだはっきりわからない。また、どんな場合、交替ができるかも分からない。

先行研究ではニ格は古い道具を示す指摘もあり、空間の意味を示す研究もある。本稿ではデ格は道具性を表し、ニ格は空間性を表すと考える。

3. 研究の目的

本稿では、「とる」「もつ」「かかえる」「だく」四つのふれあい動詞を取り上げ、以下の三つの問題を明らかにすることを目的とする。

① 四つの動詞それぞれニ格とデ格の使用率はどちらが高いか。
② 四つの動詞の～ヲ～ニ形は「手に～」以外にどんな表現があるか。
③ ～ヲ～ニ形と～ヲ～デ形はどんな場合に交替可能なのか。

4. 調査方法

本稿では、『現代書き言葉均衡コーパス（中納言)』を通して、短単位検索で「とる」「もつ」「かかえる」「だく」について調べる。～ヲ～ニ/デ形と～ニ/デ～ヲ両方使用される可能性があると考え、二つの順序で検索したのである。具体的な調査方法は以下のようである。

4.1～ヲ～ニ／デ形についての調査

「とる」は「書字形出現形」が「と[る、り、って、った]」、前方共起 1 は「書字形出現形」が「[に、で]」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「を」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。

「もつ」は「書字形出現形」が「も[つ、ち、って、った]」、前方共起 1 は「書字形出現形」が「[に、で]」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「を」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。

「かかえる」は「書字形出現形」が「かかえ」、前方共起 1 は「書字形出現形」が「[に、で]」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「を」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。

「だく」は「書字形出現形」が「だ[く、き、って、った]」、前方共起 1 は「書字形出現形」が「[に、で]」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「を」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。

4.2～ニ/デ～ヲ形についての調査

「とる」は「書字形出現形」が「と[る、り、って、った]」、前方共起 1 は「書字形出現形」が「を」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「[に、で]」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。

「もつ」は「書字形出現形」が「も[つ、ち、って、った]」、前方共起 1 は「書字形出現形」が「を」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」

が「[に、で]」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。
「かかえる」は「書字形出現形」が「かかえ」、前方共起 1 は「書字形出現形」が「を」、前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「[に、で]」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。
「だく」は「書字形出現形」が「だ[く、き、って、った]」、前方共起 1 は「書字形出現形」が「を」前方共起 2 は「品詞」の「大分類」が「名詞」、前方共起 3 は「書字形出現形」が「[に、で]」、前方共起 4 は「品詞」の「大分類」が「名詞」で検索する。

5. 調査の結果

5.1～ヲ～ニ/デ形の調査結果

～ヲ～ニ/デ形で検索して得た用例数

	～ヲ～ニ形	～ヲ～デ形
とる （使用率%）	152／500／674 （30.4）	4／500／674 （0.8）
もつ （使用率%）	30／178 （16.9）	2／178 （1.1）
かかえる （使用率%）	12／24 （50）	7／24 （29.2）
だく （使用率%）	7／8 （87.5）	1／8 （12.5）

「とる」が 674 例ヒットし、その中の 500 件が表示された。～ヲ～ニ/デと判断できる例を数えたところ、～ヲ～ニ形は 152 例に対し、～ヲ～デ形は 4 例にとどまった。
「もつ」が 178 例ヒットした。～ヲ～ニ形は 30 例に対し、～ヲ～デ形は 2 例しかなかった。
「かかえる」が 24 例ヒットした。～ヲ～ニ形は 12 例あり、～ヲ～デ形は 7 例あった。
「だく」が 8 例ヒットした。～ヲ～ニ形は 7 例に対し、～ヲ～デ形は 1 例しかなかった。

5.2～ニ/デ～ヲ形の調査結果

～ニ/デ～ヲ形で検索して得た用例数

	～ニ～ヲ形	～デ～ヲ形
とる （使用率%）	2／500／1145 （0.4）	9／500／1145 （1.8）
もつ （使用率%）	9／500／1310 （1.8）	0／500／1310 （0）
かかえる （使用率%）	5／37 （13.5）	6／37 （16.2）
だく （使用率%）	0 （0）	0 （0）

「とる」が 1145 例ヒットし、その中の 500 例が表示された。～ニ～ヲ形は 2 例あり、～デ～ヲ形は 9 例あった。

「もつ」が 1310 例ヒットし、その中の 500 例が表示された。～ニ～ヲ形は 9 例あり、～デ～ヲ形は 0 例であった。

「かかえる」が 37 例ヒットした。～ニ～ヲ形は 5 例あり、～デ～ヲ形 6 例あった。

「だく」が全無であった。

6. 考察

6.1 研究目的①への回答

～ヲ～ニ/デ形で検索したところ、四つの動詞はニ格の使用率が全部デ格の使用率より高いということがわかった。つまり、このような動詞のニ格は「道具性」より、「空間性」で見られる場合が多いと言える。

～ニ/デ～ヲで検索したところ、「とる」と「かかえる」はニ格の使用率と比べ、デ格の使用率が高いことが分かった。「もつ」はデ格の使用率が 0%であった。「だく」はニ格とデ格の使用率が全部 0%であった。

以上から、デ格・ニ格と動詞の距離によって、使用率が異なることが分かった。ニ格・デ格が動詞に近いほう、つまり～ヲ～ニ/デ形のほうは、四つのふれあい動詞の使い方が安定しており、同じ傾向が見られたが、ニ格・デ格が動詞に遠いほう、つまり、～ニ/デ～ヲ形の使い方にばらつきがあると思われる。

6.2 研究目的②への回答

先行研究では、（言語学研究会 1983：309）に格の名詞は主に身体の部分（とくに手）をしめすものであると述べている。今回の調査では、そういう傾向が見られた。更に、手以外の用例もいくつかみつかった。

「とる」

「N にとる」表現は全部で 153 例であり、「手（右手/両手/掌）にとる」用例数は 139 例であった。そのほかを含めて以下の表にまとめる。

N に	手	頭上	右肩	腕	右腰	ひざ	カード	ボール	テープ	綿棒	バケツ	容器
数量	139	4	1	1	1	1	1	1	1	1	1	1

「手にとる」以外の例文は以下のようである。

（5）刀を頭上にとり（6）、重さで降ろした後、膝を囲う姿勢をとり残心…
（小山将生『目で見て学ぶ居合道新陰流』体育とスポーツ出版社 2005）

（6）抜き付けの余勢で刀を右肩にとり撥草になる（4）。
（小山将生『目で見て学ぶ居合道新陰流』体育とスポーツ出版社 2005）

（7）彼は女を腕にとって抱き寄せた。
（シドニィ・シェルダン著　木下望　訳　『氷の淑女』徳間書店　1997）

（8）刀を引き抜きつつ右膝を床に着き（8）、刀を右腰にとった構えを示した後…
（小山将生『目で見て学ぶ居合道新陰流』体育とスポーツ出版社 2005）

（9）竜子が脱いだスリッパを膝にとって視入った。

（矢田津世子『新・ちくま文学の森』筑摩書房　1996）

（10）事項をカードにとったら、それに出てくる人名もカードにとって…

（渡部昇一　『知的生活の方法』講談社 1976）

（11）3 ヨーグルトをボールにとり、いったクミンシード、唐辛子粉、塩を入れてよくかきまわす。

（オッキア・シン　『シンさんの印度料理夜話』NTT 出版　1997）

（12）一度でも実際の会話や口述をテープにとって起こしてみると誰にでもわかる。

（盛山和夫『社会調査法入門』有斐閣 2004）

（13）下まつげは、いったんブラシから液を綿棒にとると塗りやすい。

（分担不明『The マスカラ book』学習研究社 2004）

（14）おばあさんからは、米を研いだ水をバケツにとっておき、後で緑こけむした庭石に、ひしゃくでやさしくかけてやります。

（北阪英一『東洋鬼』日本文学館 2005）

（15）さて、ヒトの血液を容器にとり、これら抗凝固剤を加えて放っておくと…

（平田剛士『そしてウンコは空のかなたへ』金曜日 2004）

「もつ」

「N にもつ」表現は全部で 30 例であり、「手（両手/片手/左手/右手/手元）にもつ」用例は 22 例であった。そのほかの用例を含めて以下の表にまとめる。

N に	手	胸	頭	腕	身	心
数量	22	4	1	1	1	1

「手にもつ」以外の例文は以下のようである。

（16）この星を胸にもつあなたは、生涯食べ物に不自由しません。

（和泉宗章『算命占星学入門』青春出版社 1978）

（17）新しい統治階層を頭にもつ新しい社会、そして新しい国民経済ウクラードである。

（土肥恒之『西洋世界の歴史』1999）

（18）四人は七つの（封印）を腕にもって以来、じつにさまざまな恐怖を味わってきた。

（カイ・マイヤー（著）山崎恒裕（訳）『マンドラゴの恐怖』）

（19）自分の身ぶりを身にもつこと―どのようにして？

（長田弘『読書百遍』岩波書店 1986）

（20）お母さん、自分は自分として、皆きれいなものを心にもっているんだよ。

（実著者不明『心に残るとっておきの話』潮文社）

「かかえる」

「N にかかえる」用例は全部で 12 例であり、「手（両手/片手/左手/右手/手元）にかかえる」用例は 1 例であった。そのほかの用例を含めて以下の表にまとめる。

N に	手	小脇	腕	胸
数量	1	8	2	1

「手にかかえる」以外の例文は以下のようである。

（21）C 男がにわとりを小脇にかかえてしじゅう持ち歩く姿は、幼稚園の中ではいやでも目に付いた。

（友定啓子『子どもの心を支える』1999）

（22）スタンプは別の問題だ―食料品の紙袋を腕にかかえて、自動ドアを通り、

（マーガレット・アトウッド著　大浦暁生訳『食べられる女』新潮社 1996）

（23）そして、風呂敷包みを胸にかかえ、足を北へと向ける。

（清め野塩『雪の積む里』文芸社 2004）

「だく」

「手にだく」の用例は 0 であった。ほかの用例は以下の表にまとめる。

N にだく	膝	胸	腕	おなか
数量	3	2	1	1

（24）これからは、毎晩、抜身をひざにだいて、青い目さんの門先の敷居ぎわに、寝るつもりです。

（リヒャルト・レアンダー著　国松孝二訳　ふしぎなオルガン　岩波書店　1987）

（25）女は赤ん坊を胸にだき、赤ん坊の足の下に女の子が頭を入れて…

（田中小実昌『香具師の旅』河出書房新社 2004）

（26）四、五歳くらいの男の子をうでにだいて、わらっている。

（朝比奈蓉子『へそまがりパパに花々を』ポプラ社 1992）

（27）みえないくらいにちいさなくうきのあわをおなかにだいて。

（いしいじんじ『ぶらんこ乗り』新潮社 2004）

以上から、ニ格とふれあい動詞の結びつきは、ニ格の名詞は主に身体の部分をしめすものであることが先行研究と同じ結論が得られた。しかし、身体表現以外の表現もあることが今回の調査でわかった。「N にとる」はカード、ボール、テープ、綿棒、バケツ、容器のような道具を表す名詞も使われることが分かった。

6.3 研究目的③への回答

～ヲ～デ形の用例は以下のようである。

「とる」

N でとる	右手	ペーパー
数量	3／4	1／4

（28）茶碗を右手でとり、7 左横、右手前でなつめとおき合わせる。

（千宗左『小棚の点前』主婦の友社 1990）

（29）油を熱して 6 を入れ、転がしながら焼きつけて焼き色をつけ、脂をペーパーでとる。

（実著者不明『おせちと気軽なおもてなし』学習研究社 2004）

「もつ」

N でもつ	両手/片手でもつ
数量	2／2

（30）椅子に座っているなら座ったまま、片方の膝の後ろを両手でもち、ゆっくりと足を引き上げる。

（実著者不明『たまった「疲れ」が驚くほどとれる本』永岡書店 2003）

（31）ときどき六寸ぐらいある基盤を片手でもって、五十匁蝋燭の火を団扇のように煽り消したそうです。

（オール讀物編集部『オール讀物』文藝春秋 2004）

「かかえる」

N でかかえる	腕/右腕/両腕でかかえる	両手でかかえる
数量	4／7	3／7

（32）わたしは、今度は、ロッドを右腕でかかえた。

（喜多嶋隆『ルアーに恋した日』光文社）

（33）ルシファードはプレイン・ギヤをかぶった頭を両手でかかえ、わめいた。

（津守時生『三千世界の鴉を殺し』新書館 2001）

「だく」

N でだく	両手でだく
数量	1／8

（34）右太はお婆のひざにぐったり頭を落とすと、ひざを両手でだきながら…

（浜田糸衛『金の環の少年』国土社 1987）

以上、「手にとる」「手にもつ」「手にかかえる」と「手でとる」「手でもつ」「手でかかえる」の用例が全部見つかり、「とる」「もつ」「かかえる」のニ格とデ格の交替可能の用例は身体の部分「手」と結ぶことであると思われる。また、「かかえる」のもう一つ交替可能の用例は身体の部分「腕」と結ぶことであると考えられる。

7. おわりに

本稿では「とる」「もつ」「だく」「かかえる」四つのふれあい動詞のニ格とデ格の交替について考察した。デ格・ニ格と動詞の距離によって、使用率が異なることが分かった。
どんな場合にニ格とデ格の交替が可能なのか、また「手に～」以外の表現も調査で明らかにした。ニ格はほかの助詞との交替を今後の研究課題とする。

文献

言語学研究会（1983）『日本語文法・連語論』むぎ書房

川野靖子（2012）「現代日本語の動詞「詰める」「覆う」の分析―格体制の交替の観点から―」『埼玉大学紀要（教養学部）第 48 巻第 2 号』

国立国語研究所報告 3（1951）『現代語の助詞・助動詞－用法と実例－』秀英出版　pp135－151

迫田久美子（1998）「誤用を生み出す学習者のストラテジー場所を表す格助詞「に」と「で」の使い分け―」『平成 10 年度日本語教育学会秋季大会予稿集』pp128－134

菅井三実（1997）「格助詞「で」の意味特性に関する一考察」『名古屋大学文学部研究論集

文学 43』pp23－40
菅井三実（2000）「格助詞「に」の意味特性に関する覚書」『兵庫教育大学研究紀要　第 2 分冊　言語系教育・社会系教育・芸術系教育 20』pp13－24
益岡隆志　田窪行則（1987）『格助詞　日本語文法　セルフ・マスターシリーズ 3』くろしお出版　pp4－5
森山新（2005）『認知言語学的観点を取り入れた格助詞の意味のネットワーク構造解明とその習得過程（平成 14 年度～平成 16 年度科学研究費補助金研究　基盤研究（C）（2）課題番号 14510615　研究代表者：森山新）成果報告書』
森山新（2008）「認知言語学的観点からの格助詞ヲ、ニ、デの意味構造とその習得－中国語を母語とする日本語学習者を中心として－」（台湾大学大学院との第 2 回所インドゼミ：お茶の水女子大学）『大学院教育改革支援プログラム「日本文化研究の国際的情報伝達スキルの育成」活動報告書　平成 19 年度　海外研修事業編』pp240－244
山梨正明（1995）『認知文法論』ひつじ書房
『現代日本語書き言葉均衡コーパス』

コーパスに基づく日本語擬態語動詞の意味分析

菅原 崇（岐阜工業高等専門学校）、
浜野 祥子（ジョージワシントン大学）

A Semantic Study on Japanese Mimetic Verbs based on Corpus Data

Takashi Sugahara (Gifu National College of Technology),
Shoko Hamano (George Washington University)

１．はじめに

日本語のオノマトペ（特に擬態語、擬情語）は「する」「つく」「めく」などの接辞を伴い動詞化する。

その中で「する」は最も生産的な接辞で、Kakehi, Tamori, and Schourup (1996)においておよそ 300 語のオノマトペが「する」を伴う動詞形を持つとされている。その中でも 2 モーラ反復形のオノマトペに「する」が伴う形（e.g. ぶらぶらする、ざらざらする）は優勢で、Kakehi, Tamori, and Schourup (1996)にはおよそ 170 語が収録されている（本研究ではこれを「する」動詞と呼ぶ）。

2 番目に生産的な動詞化は 2 モーラ非反復形のオノマトペに接辞「つく」を伴う形（e.g. ぶらつく、ざらつく）であるが、「する」動詞ほど数は多くなく Kakehi, Tamori, and Schourup (1996)には 24 語が収録されている（以下「つく」動詞と呼ぶ）。

一見すると、「つく」動詞は「する」動詞と同じ意味を表しているかのように思われる。

(1) a. 私が近所をぶらぶらする。
　　b. 私が近所をぶらつく。
　　c. 床がざらざらする。
　　d. 床がざらつく。

田守・スコウラップ(1999)は「つく」動詞は対応する「する」動詞を持つこと、「つく」動詞は否定的な意味を伝えるものであるとしている（ibid: 56-57）。

しかしながら、「つく」動詞すべてが否定的な意味を持つわけではない（e.g. 雪がちらつく）。さらに、否定的な意味を持つ「する」動詞が必ず対応する「つく」動詞を持つわけでもない（びちゃびちゃする、*びちゃつく）。これらのことは、「つく」動詞が「する」動詞とは異なる意味的特徴を持つ可能性を示唆している。

本研究は、コーパスデータをもとに「する」動詞と「つく」動詞の意味的な違いを明らかにする。

２．データ

Kakehi, Tamori, and Schourup (1996) 収録の 24 の「つく」動詞のうち 23 語が対応する「する」動詞を持つ（例外は「ばらつく」）。本研究はデータとしてこれら 23 の「する」「つく」動詞ペアの現代日本語書き言葉均衡コーパス（少納言）における用例（「つく」動詞 2098 トークン、「する」動詞 2192 トークン）を用いる。

具体的な調査方法は、これら用例を主語が Agent、Experiencer、Theme のどれに属しているのかで大まかに分類したのち、より細かく「する」動詞を好む主語、「つく」動詞を好む主語を特定する形を取る。

３．Theme を主語に取る動詞

Theme として「する」「つく」動詞の主語になっている名詞の中で、「つく」動詞より「す

る」動詞と好んで共起するものがある。それが表 1 にまとめたものである。

表 1 「する」を好む Theme と擬態語の組み合わせ

Theme と擬態語との組み合わせ		「する」と「つく」の頻度
Theme	擬態語	
液体	ねば(ねば)	**19** : 2
食べ物		**10** : 1
その他の集合名詞 (「くつずみ」「膜」など)		**6** : 1
その他の名詞 (「口の中」など)		**13** : 3
目	ごろ(ごろ)	**9** : 0
石や岩		**31** : 2
人や生物 (死体も含む)		**20** : 0
その他散らばっているもの (「話」「企業」など)		**27** : 0
各種不安定なもの (「窓」「ドア」「椅子」など)	がた(がた)	**15** : 7
光沢のあるもの (「ガラス」「電灯」など)	ぎら(ぎら)	**27** : 8
各種垂れ下がっているもの (「足」「ケーブル」など)	ぶら(ぶら)	**16** : 0
人（単数、複数どちらの場合もあり）	ごた(ごた)	**22** : 1
場所		**14** : 1
抽象物		**11** : 0
各種混乱しているもの (「店」「仕事」など)	ばた(ばた)	**15** : 2

表中の Theme と擬態語の組み合わせは恒常的ないしは継続的な状態を表している。例えば、「液体」と擬態語「ねば(ねば)」の組み合わせは、液体が本来持つ粘り気を表すし、食べ物（e.g. 納豆）の場合はその食べ物が本来持つ粘り気を表す。

(2) a. 唾液はねばねばして、いやな味がする。
b. [金のつぶ納豆は] 食べる前にビニールを取るのにネバネバしないからスゲぇ～楽チンっ！

これら液体や食べ物の恒常的な粘り気を表す場合「する」が好まれることを表 1 の頻度は示している。

「目がごろ(ごろ)」はコンタクトレンズなどの異物が目の中に入った場合の不快感を表す際に用いられる。このような状態は異物を取り除くまで継続する。また「ごろ(ごろ)」は石や岩、人などが一つのところに散らばっている（継続的な）状態を表す。どちらの場合においても「する」が好まれている。

(3) a. コンタクトするとごろごろしませんか？
b. 砂浜はなく、岩がごろごろしてますから。

「がた(がた)」「ぎら(ぎら)」「ぶら(ぶら)」「ごた(ごた)」「ばた(ばた)」の場合も同様に、主語となる名詞の継続的な状態を表している。

表 1 の場合とは逆に、「する」動詞より「つく」動詞と好んで共起する主語（Theme）がある。それが Table 2 にまとめたものである。

表 2 「つく」を好む Theme と擬態語の組み合わせ

Theme と擬態語との組み合わせ		「する」と「つく」の頻度
Theme	擬態語	
肌	かさ(かさ)	12 : **21**
雨	ぱら(ぱら)	1 : **37**
雪	ちら(ちら)	1 : **46**
影（「面影」「幻影」も含む）		0 : **21**
姿		2 : **10**
顔		1: **7**
身体部位（「頭の中」「体中」「腹の底」「胸」など）	ざわ(ざわ)	2 : **9**
場所		16 : **38**
贅肉やそれが付いた身体部位	だぶ(だぶ)	3 : **15**
金		0 : **12**
商品としてのもの		0 : **14**
各種余剰なもの（「ニット」「シルエット」など）	もた(もた)	5 : **18**

表 2 にある擬態語のうち「かさ(かさ)」「ぱら(ぱら)」「ちら(ちら)」「ざわ(ざわ)」は、同じく表 2 にある Theme との組み合わせによって一時的な状態や瞬間的な動きを表している。例えば、「肌」と「かさ(かさ)」の組み合わせは体の表面の乾燥を表しているので、そのような乾燥状態はボディクリームなどを塗れば改善する一時的なものである。

(4) a. どこもかしこも乾燥し、ぼくらの身体も、肌がかさかさして白い粉をふき [...]。
b. シャワーで自分の肌に触れると、お？[肌が] かさついてない。

「雨」と「ぱら(ぱら)」の組み合わせは雨の降り始めのような短期間の状態や、すぐに止むような時雨などの一時的な状態を表す。「雪」と「ちら(ちら)」の組み合わせも同様である。

(5) a. 朝は雨がパラパラして・・・とっても寒かった！！
b. 職場を出るときは、雨がぱらついていたが、下町は晴れている。
c. 白いのがちらちらすると見るまに大雪になってしまった。
d. この日は、前日に雪がちらついたものの、朝から快晴となり[...]。

なお、「影」「姿」「顔」を主語に取る場合の「ちら(ちら)」は（比喩的に）瞬間的な動き、すなわちそれらがよぎることを表す。

(6) a. これらの事件の後には、バサエフ野戦司令官の影がちらついている。
b. 昨夜はダイナのナイトガウン姿が頭にちらついてほとんど眠れなかった。
c. 日高玲子の、端整な顔が、内海の目の前にちらついた。

「頭の中」「体中」「腹の底」「胸」などの身体部位と「ざわ(ざわ)」との組み合わせは、身体部位の違和感とそれに伴う不安や昂揚感などの感情を表すが、そのような感覚（またはそれに伴う感情）は長時間継続するようなものではなく、瞬間的ですぐになくなるものである。次に「場所」について、その内訳は「教室」「客席」「会議室」「場内」など通常はそれほど騒がしくないところであり、「ざわ(ざわ)」との組み合わせによって表される喧騒はすぐに収まることが予測される。

(7) a. 心臓が徐々に鼓動を強くしてゆく。体中がざわつき、血が凍りついてゆく。
b. 藤圭子の唄が流れはじめた。唄はいつまでも続き、舞台には誰もあらわれない。客席がざわつきはじめたとき、ピンスポットに照らされてベビードールを着た女が登場した。

「肌」と「かさ(かさ)」に見られた「ものの表面の状態」という特徴は、「だぶ(だぶ)」「もた(もた)」にも見られる。まず「贅肉」と「だぶ(だぶ)」との組み合わせは体の表面に余計なものが付着していることを表す。「金」や「商品」に関してはその比喩的表現といえる。

(8) a. そのひみつは、おなかのだぶだぶした皮ふにあります。
b. 太り過ぎの確かめ方はお腹の肉がだぶついているとか[...]。
c. 郵政民営化で 250 兆円のお金をだぶつかせてどこに使うのかと質問したら[...]。
d. 国内にだぶつく工業製品のはけ口と[...]。

同様に「もた(もた)」についても、衣類などが体の表面に余分にまとわりついている様子を表す。

ここまで見てきた「する」動詞と「つく」動詞の違いは、以下表 3 や表 4 にある Theme と擬態語との組み合わせを対比させることでより鮮明になる。

表 3 「する」を好む Theme と擬態語の組み合わせ（その 2）

Theme と擬態語との組み合わせ		「する」と「つく」の頻度
Theme	擬態語	
胃	むか(むか)	**16** : 8
頭	ぐら(ぐら)	**16** : 1
歯		**16** : 10
頭	ふら(ふら)	**26** : 8
食べ物	ぱさ(ぱさ)	**25** : 15
抽象物 （「関係」など）	べた(べた)	**12** : 0

表 4 「つく」を好む Theme と擬態語の組み合わせ（その 2）

Theme と擬態語との組み合わせ		「する」と「つく」の頻度
Theme	擬態語	
胸	むか(むか)	12 : **29**
「自信」「信念」など	ぐら(ぐら)	2 : **32**
「権威」「立場」など		1 : **14**
体	ふら(ふら)	4 : **18**
足		2 : **19**
「足元」「足どり」		0 : **31**

髪	ぱさ(ぱさ)	5 : **25**
髪	べた(べた)	1 : **7**
化粧品		5 : **25**

「むか(むか)」は「胃」とも「胸」とも共起するが、「胃」との組み合わせの場合、内容物が食道に降りるまで続く比較的長い胃の不快感を表し、それは基本的には胃そのものの疾患による。一方、「胸」との組み合わせの場合は胸やけのように、胃の不快感よりも比較的短い時間で改善する状態を表す。さらに、胸やけの場合は基本的に胃酸の逆流によるものなので、健康な人間であれば食事の直後に横になるなどしなければ起きない一時的な症状である。ここから「胃」の場合「する」を好み、「胸」の場合「つく」を好むことは予想でき、実際、頻度がそれを証明している。

(9) a. 胃が痛くて、ムカムカして、胃ガン検査を受けようかと思いました。
b. レストランを出てしばらくすると、牧師は、急に胸がむかついて気分が悪いと言い出した。

「ぐら(ぐら)」も同様の対比が可能である。「頭」との組み合わせは薬を服用してもある程度の継続するめまいのような体の不調を表す。また、「歯」との組み合わせも同様で、抜くまで継続する歯の不安定さを表す。これらの場合「する」が好まれることを表 3 は示している。一方、「自信」「信念」「権威」「立場」など本来安定しているものの場合、一度不安定になってもすぐに回復することが予測される。表 4 が示すように、これらの場合「つく」が好まれている。

(10) a. 疲労と、久方ぶりに口にしたアルコールの相乗効果で、頭がぐらぐらする。
b. 私が働いていた歯医者にお子さんを連れてきて歯がぐらぐらするって言ってきました。
c. “ＫＧＢきってのエリート女性”という自信がぐらつくのをくのを、マーイヤはどうすることもできなかった。
d. 香取さんのような美人にお目にかかると、ぼくの信念もぐらつきますがね。

「ふら(ふら)」の場合、「頭」を主語に取る場合と、「体」「足」「足元」「足どり」を主語に取る場合に二分できる。前者の場合「ぐら(ぐら)」と同様の継続的なめまいを表し、「する」が好まれている。一方後者の場合、一旦不安定になったとしても、普通は体勢を立て直すことが予測される。これは一時的な状態であり、「つく」が好まれている。

(11) a. 皆が一斉に証言する「変なお香」の効果か、頭がふらふらしているようだ。
b. 疲労のあまり脚はふらつき、背中は汗でぐっしょり。
c. 右手は、右側にある洗濯物のどっかを持ちまして、体がふらつかないようにちょいと腰を入れてね。
d. ハアハア～と速い呼吸で、歩く足取りがふらついたりグッタリとして横になったりします。

「食べ物」と「ぱさ(ぱさ)」の組み合わせは古くなったパンや炊き方を間違えたごはんなどの乾燥状態を表す。そのような食べ物は一度乾燥してしまえば、もう柔らかさやみずみずしさを取り戻すことはない。この場合「する」が好まれることを表 3 が示している。一方、「髪」と「ぱさ(ぱさ)」の組み合わせは髪の毛の乾燥状態を表すが、そのような状態はヘアクリームなどを付ければ改善できる。すなわち、これは表面的で一時的な状態と考えられる。この場合「つく」が好まれていることを表 4 が示している。

(12) a. このターキーサンド、パンもターキーもパサパサしていたの[...]。
b. 妊娠中は、赤ちゃんに栄養を採られるから髪がパサついて纏まりにくくなるんですよね。

「べた(べた)」はものの粘り気を表すが、「関係」などの抽象概念が Theme の場合、すぐにそのような関係が途切れるとは考えられないため、それは継続的な状態といえる。この場合「する」が好まれている。一方、「髪」が Theme の場合、その粘り気は洗えばなくなる表面的で一時的な状態である。この場合「つく」が好まれている。「化粧品」の場合、それを付ける肌（や髪）の表面を前提とし、化粧品の肌への付け心地を表現している。もちろん肌から化粧品はすぐに取り除けるので、これも一時的な状態である。この場合「つく」が好まれている。

(13) a. ベタベタした濃密な人間関係が嫌いだから[...]。
c. 最近は梅雨でジメジメしているので、[髪が]べたつきやすいです。
d. 潤ってベタかない化粧水でオススメのものがあれば教えてください。

ここまでで、主語が Theme の場合、擬態語が継続的な動きや状態を表しているときには「する」が、擬態語が瞬間的な動きや一時的な状態を表しているときには「つく」が好まれることが分かった。また、Theme の表面もしくは Theme が添付されるものの表面の状態を表す場合「つく」が好まれることが分かった。

４．Agent を主語に取る動詞

これより Agent を主語に取る場合の「する」「つく」動詞の振る舞いの違いを見ていく。表 5 は「する」「つく」動詞いずれかで Agent を主語に取ることが優勢な擬態語を、その頻度と共にまとめたものである。

表 5 Agent を主語に取る「する」「つく」動詞の頻度

擬態語	接辞	主語の種類		
		Agent	Experiencer	Theme
ばた(ばた)	する	**139**	0	16
	つく	9	0	4
ぶら(ぶら)	する	**167**	0	16
	つく	99	0	0
ふら(ふら)	する	**134**	0	62
	つく	92	0	100
ごろ(ごろ)	する	**160**	0	98
	つく	2	0	3
いちゃ(いちゃ)	する	**42**	0	0
	つく	41	0	0
もた(もた)	する	**77**	0	5
	つく	39	0	18
うろ(うろ)	する	**401**	0	7
	つく	351	0	0
ぱく(ぱく)	する	**6**	0	0
	つく	31	0	0

表 5 から 8 つの Agent を主語に取る「する」動詞のうち、「ぱくぱくする」以外の 7 つが「つく」動詞より頻度が高いことが分かる。注目すべきは、これら 7 つの「する」動詞は全て身体全体を使った動きであるという点である。

(14) a. レース当日の朝にバタバタしたくないので土曜は山中湖で宿泊。
b. 本当なら久しぶりに心斎橋商店街をブラブラして買い物しようかと思ったけど休日やから人でごった返してるしな～[...]。
c. 今日は買い物がてら、真っ昼間に自転車で出かけまして、近所の公園などをふらふらしておりました。
d. 明日は 1 日中スウェットでゴロゴロしよー[。]
e. 向ヶ丘遊園の喫茶店でウェートレスといちゃいちゃしてやがった[。]
f. 何をもたもたしているんですか。いくら他県の事件にしろ、警部らしくないじゃないですか[。]
g. 昔はよく九州や関西方面を一人でウロウロしていました。

一方、Agent を主語に取る「ぱくぱくする」「ぱくつく」は例外的に後者の方が高頻度である。これらが先の 7 つの「する」「つく」動詞ペアと異なる点は、その動きが身体全体によるものではなく、身体部位、つまり口の動きであるということ、さらに「ぱくつく」については他動詞ということである（なお、「ぱくぱくする」の 6 件のうち、自動詞は 4 件、「口」を目的語にとる他動詞は 2 件）。

(15) a. [魚が]底砂をぱくぱくしている様子もありません。
b. フォルナッティはそこで一呼吸いれ、ロールパンにバターをぬってぱくついた。

５．まとめと考察

本研究では「する」「つく」動詞の違いについて以下の 2 つの発見があった。

I. 主語が Theme の場合、擬態語が継続的な動きや状態を表しているときには「する」が、擬態語が瞬間的な動きや一時的な状態を表しているときには「つく」が好まれる。また、Theme の表面もしくは Theme が添付されるものの表面の状態を表す場合に「つく」が好まれる。
II. 主語が Agent の場合、擬態語が身体全体の動きを表す場合には「する」が、身体部位の動きを表す場合には「つく」が好まれる。

これらは、「つく」動詞は否定的なニュアンスを持つ単なる「する」動詞の代替ではないことを示している。

「する」動詞の「継続性」と「つく」動詞の「一時性」の対比は擬態語部分の形態に由来するとも考えられる。一般に日本語における 2 モーラ反復形の擬音語・擬態語は音や動きの繰り返しを表しているとされている。この対比が本研究の「する」動詞と「つく」動詞の実際の振る舞いの違いに現れたとも言える。

「つく」動詞が持つ「表面性」については自動詞「付く」や他動詞「突く」が語源的に大きく関わっていると考えられる。一方で「する」には「つく」が持つような一般動詞との関連性は見出しにくい（故に、「表面性」のような特徴を持たない）。

Agent を主語に取る場合の上記発見Ⅱは Theme を主語に取る発見Ⅰと関連付けることができる。本研究で最も大きく関わる Agent を主語に取る動詞の用法は Activity verbs 的な用法である。Activity verbs の意味は一般化すれば「意志によってコントロールされる身体全体の継続的な運動」である。これに反復形の擬態語と意味的に軽い接辞「する」の組み合わ

せ容易に符合することができる。一方、非反復形の擬態語と接辞「つく」が持つ意味は Activity verbs の「身体全体」「継続的」という意味と符合しない。そのため、8 つ中 7 つの Agent を取る「する」「つく」動詞ペアで前者の方が後者より頻度が高いという結果を生んだと考えられる。

最後に「ぱくつく」が「ぱくぱくする」より高頻度だったことについて、「ぱく(ぱく)」は口の動きで先の Activity verbs の一般化の「身体全体」の部分に当てはまらない（そのため「ぱくぱくする」の頻度はあまり高くなかった)。一方で、「ぱく(ぱく)」で表される「食べる」という行為は「Agent がその口を食べ物の表面に突き動かす」と見ることができる。この意味は接辞「つく」が持つ「表面性」や一般動詞「付く」「突く」との関連と共存しうる。結果「ぱくつく」が「ぱくぱくする」より高頻度になったと考えられる。

参考文献

Kakehi, Hisao, Ikuhiro Tamori, Lawrence Schourup（1996）*Dictionary of Iconic Expressions in Japanese, 2 vols.* Mouton de Gruyter.

田守育啓、ローレンス・スコウラップ（1999)『オノマトペ：形態と意味』くろしお出版.

関連 URL

少納言 KOTONOHA 現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese). http://www.kotonoha.gr.jp/shonagon/

条件構文の談話標識化の諸相

藤井 聖子（東京大学大学院総合文化研究科）

Conditional Constructions Used as Discourse Markers

Seiko Fujii (Graduate School of Arts and Sciences, The University of Tokyo)

１．はじめに

日本語の条件構文は、「(れ)ば」「たら」「なら」「と」「ても」「ては」等の節接続形態素を含む従属節と主節とが複文を形成する生産的構文である一方、発話冒頭ポジションで、「とすれば」「だったら」「なら」「すると」「とすると」等文頭接続詞的に用いられたり、「例えば」「言ってみれば」「よかったら」「可能なら」「本来なら」「どちらかというと」「要約すると」等ある語彙とともに定型的に副詞的に用いられたりし、談話標識化・語用標識化している用法もある。本発表では、このような条件構文基盤の談話標識化の諸相を探索吟味しつつ、書き言葉コーパス・話し言葉コーパスを用いてこれらの現象を分析する際必要となる指標や類型を考察する。

２．語用標識・談話標識、そして、語用標識化・談話標識化

ここで取り上げる「談話標識」は、Schiffrin (1987) の discourse markers に関する見解・理論に依拠する。Schiffrin（1987）は、談話においてトーク’talk’という単位を見いだした上で、discourse markers を、談話の中でトーク’talk’の単位を区切る連鎖依存の要素と操作的に定義付けている。発話先頭・発話末の標識が前方照応あるいは後方照応しつつトーク‘talk’単位を括り、談話の意味展開の結束性に寄与する標識である。Schiffrin (1987)がその典型的事例として着目したのが、英語では、oh, well, now や so, because, and, but, or, y’know 等である。日本語では、「ああ」「あら」「あれ」「まあ」等いわゆる間投詞といわれてきた類のものや「だから」「だって」「で」「でも」「だけど」「或は」等いわゆる接続語といわれる結束性標識語がその典型的な類例である。

談話における同種の現象に関して、Schiffrin (1987) の discourse marke 以外にも、discourse particle 等他の概念・理論や用語が提示されてきていたが、談話標識 discourse marker や discourse particle 等という範疇も含むより包括的かつ一般化可能な概念・範疇として、Fraser (1996)が語用標識(pragmatic markers)という概念・用語を発案し、語用標識(pragmatic markers)の類型・細分化を提案した。Fraser による語用標識の体系付けの中で、談話標識(discourse markers)を語用標識(pragmatic markers)の一種として位置づけることができる。本研究で扱う現象は、基本的対象としては談話標識であり Schiffrin によるその定義に適合するものであるが、狭義の談話標識の域を越えて語用標識の他の類としての機能を呈するものも含まれる。

このような語用標識・談話標識は、ほとんどの場合、元々そのための語として存在していたというより、他の統語的特徴・意味・機能をもつ語（多くの場合他の品詞・文法範疇）や語の組み合わせが談話の相互行為の中で繰り返し使用される中で、形式と語用的機能との結びつきが顕著になり定着化したものが多い。このような現象を、広く pragmaticalization 「語用(論)化」と捉える理論や通時・共時的事例研究が英語・日本語を含む多くの言語で展開してきていた（Traugott & Heine 1991, Traugott 2004 等, Onodera 1995 等, 他）。藤井(2008 ; 科研課題 2006-2009)では、このような pragmaticalization を射程に、「内容語から文法的機能語へという文法化プロセスのさらに先に生じてくる現象、すなわち文法的機能語が本来の統語的特質を多少薄め、転じてより語用的機能を強化して語用標識に転化していく現象」を、文法的機能語の「語用標識化」と呼んだ。本研究でもこの意味で、「語用標識化」という概念・用語を用い、特にその一種（Fraser 1996 による語用標識の類型と包含関係に基づく）として発話冒頭ポジションで談話結束性や発話の手続き的意味の表象に寄与し談話標識的機能が定着化していくありさまを「談話標識化」と呼ぶ。

前述のpragmaticalization 「語用(論)化」の研究では、多くの言語で、発話左端（発話冒頭）或は発話右端（発話末）で語用化が生じやすいことが示されている。日本語もこの例に漏れず、日

本語の条件構文においても、発話左端（発話冒頭）或は発話右端（発話末）が語用標識化・談話標識化の温床となり易くなっている。発話右端（発話末）ポジションにおいては、助言・提言の発話機能を呈する「〜ば↑」「〜たら↑」で完了する構文や、当為的「義務」機能を呈する「〜ないと」「〜なければ」「〜なきゃ」「〜なくては」「〜なくちゃ」で完了する構文等、本来従属節である条件構文の前件節が単独で独立節構文として用いられ語用標識化しているケース（この場合狭義の「談話標識」とは別の類型）がみられる。本稿が着目するのは、逆に発話左端（発話冒頭）で条件接続形態素を含む条件構文(の一部)が、語用標識化・談話標識化している現象である。

３．日本語の条件構文の語用標識化・談話標識化：構文の形式的特徴に関する類型

日本語の条件構文の語用標識化・談話標識化の探求を目的にして、コーパスに分析用コーディングをする場合、まず構文の形式的特徴に関する作業類型として、「拘束形態素の非拘束化型」「指示詞照応型」「動詞等述語を含む合成型」という類別が少なくとも必要である。

なお、それ以前に必須の基礎的な形式指標は、当然、「(れ)ば」「たら」「なら」「と」等のうちどの接続形態素を用いているか、という明確な形態上のバリエーションであり、その接続形態素と意味・機能との相関・選好性(または制約)の有無が研究問題になるわけであるが、接続形態素という指標に関しては明確であるのでここでは割愛する。

３.１　拘束形態素の非拘束化型

本来拘束形態素(bound morpheme)である形態素が自由形態素的な振る舞いをし、文頭接続語化しているものが筆頭に挙がる。因果関係や逆接・譲歩等の意味範疇では、馴染み深い「だから」「だって」「で」「でも」「だが」等、殆どの一般的国語辞書にも語彙項目としてたてられているほど語彙化している接続語の類である。この類には、明確に語彙化していると考えられるものもあれば、談話の中で使用され新奇性を強く感じさせるものもある。[1]

条件関係の意味範疇では、「ならば」「だったら」「でしたら」「だとしたら」「だと」「ですと」「と」「なら」「では」「じゃ」等が発話冒頭で使われていることがコーパスで認められる。これらの発話冒頭での使用は、(その表現が語彙化していると見なされるかどうかは別にして[2]、少なくともその成り立ち・構成として）本来拘束形態素として用いられる判断詞「だ」や接続形態素（「と」等）が、発話冒頭でむき出しで自由形態素的に使用されている。本来拘束形態素である形態素が発話冒頭(左端位置)に出現することで、発話連鎖依存性を強く感じさせる標識である。

(1) 「そうか！　なら、そっちも手伝ってやるぜ」BCCWJ **書籍:文学:**LBh9_00051b
(2) 「『聖なる石』が欠けたことが、異変の原因なのだろうか…　だとしたら　政堂へはどのように報告をすればよいものやら…」BCCWJ **書籍:文学:**:LBg9_00072b

これらの使用の中には、(1)(2)のように同じ話者による発話冒頭の場合と、(3)(4)のように話者交代後に別の話者が先の話者の発話を受ける場合とがある。また次項(5)のように（形式的には、次項 3.2「指示詞照応型」の用例だが）同じ話者による発話冒頭ではあるものの、先行発話が他者の引用であり、他者引用を照応している用法もある。

(3) 「だったら、早くそれをいってください！」**書籍:文学:**LBm9_00145b

(4)　447.　SJ5:　ato=　piano o hiku koto ka na ?
　　448.　SJ5:　piano .
　　449.　TJ5:　piano nai zyan .
　　450.　SJ5:　kiiboodo ga aru zyan .
　　451.　TJ5:　hiiteru ?

[1] 世代や時代によって受け止め方に差はあるものの、例えば、発話冒頭で用いられる「だもんで」「なもんで」（Fujii 2000）は、後者（新奇性を感じさせる談話標識化途上の事例）にあたるだろう。

[2] 現行 BCCWJ では、「だったら」「では」「じゃ」「だと」は「接続詞」長単位・語彙素としての登録がある一方、その他に関しては（「でしたら」「なら」等も含め）構成形態素それぞれの短単位情報（および活用形）に留まり、その活用形自体が語彙化しているという判断での登録はないようだ。

452. SJ5: hiiteru yo .
453. TJ5: watasi mita koto nai .
454. TJ5: Ikkai mo .
455. SJ5: tama ni hiku mon .
=> 456. TJ5: aa zyaa kondo kikasete morawanakya . ああじゃ今度きかせてもらわなきゃ .
457. SJ5: e .
458. SJ5: e he he he he .
459. TJ5: nori umai no ?
460. SJ5: iya umaku- （Fujii 1995-1997）

話者間・話者内いずれの場合も、「なら」「だとしたら」「だったら」「じゃ」が先行発話を照応し、その先行発話の事態を前提にして後続発話での発話行為を行っていることを示す標識として機能している。その他（「S とすると S」「S となると S」「S とすれば S」等複合辞条件構文に依拠する）「とすると」「だとすると」「となると」等の使用も広汎に認められた。（表 1 参照。）

表 1．発話冒頭で用いられる条件構文基盤の談話標識（一部）： 拘束形態素の非拘束化型「だったら」「なら」「とすると」「だとすると」; 指示詞照応型「そうだとすると」 [BCCWJ]

	文頭「だったら」			文頭「なら」		文頭「とすると」		文頭「だとすると」		文頭「そうだとすると」	
	916			477		269		137		101	
レジスター		ジャンル		R	G	R	G	R	G	R	G
出版・雑誌	37			27		4		3		1	
出版・書籍	299	0 総記	2	121	2	73	3	35	1	21	1
		1 哲学	13		4		3		0		3
		2 歴史	8		6		7		2		0
		3 社会科学	22		20		11		5		10
		4 自然科学	1		6		6		0		0
		5 技術・工学	4		3		5		0		1
		6 産業	3		4		2		0		1
		7 芸術・美術	9		6		8		0		0
		8 言語	2		3		2		2		0
		9 文学	229		62		25		25		5
		分類なし	6		5		1		0		0
出版・新聞	0			4		1		1		0	
図書館・書籍	285	0 総記	1	141	1	139	0	52	0	45	0
		1 哲学	11		6		7		6		3
		2 歴史	3		7		13		0		5
		3 社会科学	21		15		25		5		10
		4 自然科学	2		6		7		2		4
		5 技術・工学	5		5		2		1		1
		6 産業	7		2		3		0		1
		7 芸術・美術	9		4		3		0		1
		8 言語	0		1		5		0		2
		9 文学	210		86		67		36		16
		分類なし	16		8		5		2		2
特定目的・ブログ	83			63		11		4		2	
特定目的・ベストセ	51			15		27		4		6	
特定目的・教科書	0			1		2		0		0	
特定目的・国会会議	11			11		3		8		17	
特定目的・知恵袋	150			94		9		30		9	
特定目的・白書	0			0		0		0		0	
特定目的・法律	0			0		0		0		0	
特定目的・広報誌	0			0		0		0		0	
	916			477		269		137		101	

３.２　指示詞照応型

拘束形態素に加えて「そう」等の指示詞が共起して談話の前方照応をしつつ、ある指示詞と接続形態素との組み合わせが談話標識化している類が頻繁に用いられている（例：5, 6, 7）。3.1.の「拘束形態素の非拘束化型」として出現するものは、「そうならば」「そうだったら」「それだったら」「そうでしたら」「そうだと」「そうなら（そんなら）」「それでは」「それじゃ」「そうだとしたら」「そうだとすれば」「そういうことなら」等、指示詞共起の指示詞照応型の発話冒頭での使用が認められた。

(5) フン、なに…千葉先生もこの梅太郎を叱っていたと…　そんなら、その先生の娘が、お前などは相手にせず、この梅太郎を慕っている事実はどう解釈する？**書籍:文学**:LBa9_00025b

(6) そうだとしたら　こんなにお化粧して香水をふりかけたりするでしょうか。**書籍:文学**:LBk9_00058b

(7)
171 *I57: nizyuusanniti made.
172 *S57: un.
173 *I57: +, zyuuroku kara nizyuusan made de sa.
174 *I57: +, sono ato dooyuu huu ni yaru no ka na?
175 *S57: iya yokuwakannai kedo.
176 *S57: ma zyuuroku kara nizyuusan made wa.
177 *I57: yannai desyo?
178 *S57: zissyuren mitaina kanzi desyo.
179 *I57: sono ato doo suru no ka na?
180 *I57: ## hayaku kimetehosii yo na.
181 *S57: a nanka zyosii no gassyuku toka aru no?
182 *I57: un.å
183 *I57: ikitai n da. [=! laughing]
=> 184 *S57: maa sore dattara ii n zyanai no?　まあ それ だったら いい ん じゃない の ？
185 *I57: e?
186 *I57: ii ka na?
187 *S57: un.
188 *S57: ma hatigatu toka da to tyotto mazui kamo sirenai kedo.
190 *I57: un.　(Fujii 1995-1997)

談話標識化プロセスについては、共時的データの分析に基づく一般化は避けるべきだが、「そうなら」「それなら」の使用基盤から「なら」へ、「そうだとしたら」の使用基盤から「だとしたら」へ、「そうすると」の使用基盤から「すると」へ、というふうに、合成的な指示詞照応型の談話での使用が拘束形態素の発話冒頭での非拘束的新奇使用の基盤となり、拘束形態素の非拘束化の基盤となっているという仮説が妥当であろう。従って、共時コーパスの分析で作業仮説としているのは、「拘束形態素の非拘束化型の使用が観察される場合は、その談話標識に対応しその基盤構文となる指示詞照応型が可能であり使用されている」という仮説である。

この指示詞照応型に参与する指示詞はソ系が最も広汎に使用されている。定着化・慣用性・語彙化度が相対的に強い「そうしたら」「そしたら」等もソ系である。意味機能的にソ系のみで可能なもの（例：「そういえば」）もある一方、コ系やア系も出現可能な意味文脈・談話標識もある。ソ系のみでなくコ系での使用が顕著なのが、「これでいくと」「こうなると」等である（「そうなると」も勿論使用されている）。ソ系かコ系かに関しては、5 節において後述の「発話(間相互行為)における条件付けの仕方」によって動機付けられている変容であると考える。

３.３　動詞等述語を含む合成型

内容述語の条件形が軸となる合成型も広汎に使われている。「具体的にいえば」「言ってみれば」「あえて言えば」「言い換えれば」「詳しく言えば」「なぜかと言えば」「正直にいうと」「まとめると」「例えていうなら」「さらにいうなら」「極言すれば」「極論すれば」「補足すれば」等、ある語彙とともに副詞的に用いられ(ある程度)定型的な条件表現の使用がコーパスで多種認め

られる（表 2 参照）。（「よかったら」「よろしければ」における形容詞等、動詞以外の体言述語が含まれる場合も、この類型の細分とみなす。）自立語・動詞等を含み（3.1 でみた拘束形態素の非拘束化型等と異なり）拘束形態素がむき出しで発話されるわけではなく、内容述語の条件形を基軸に一応節を形成しており、語彙化した談話標識に比べると通常の生産的条件複文により近い。（BCCWJ, CSJ からの用例抜粋を、4 節の(11)(12)(13)に示す。）

とはいえ、ある語彙群とともに使われ(ある程度)定型的であることに加え、その述語の項が（主語も含めゼロになっていること自体は日本語に一般的であるが）特定の指示をもつ特定ゼロ照応(definite null instantiation)ではなく、不定的（indefinite null instantiation）或は文脈指示であることが多い。また、談話標識としての用法では、項を明示的に補っても適切な同機能の発話にはならないことから、文脈において指示認識可能な項がゼロ照応になっていると考えるより項の指示性自体が希薄化していると考えられる。さらに、次節以降でみるように機能的特徴をもつ。

４．発話内の機能的特徴

４.１　基盤条件構文の「事態把握領域」による類型

以上みたような種類の条件前件表現が発話冒頭で使われる際、その発話内での機能に関しても特徴がある。発話内での機能（その機能ラベリングの一つ）の明確化のために、基盤条件構文の「事態把握領域」による類型を援用した。Sweester (1990)は、条件文が異なる「言語概念レベル」・「事態把握領域」において表象・解釈されることを指摘し、内容レベルの事態把握領域、認識レベルの事態把握領域、発話行為レベルの事態把握領域において、内容的条件文(例: 8)のみでなく非内容的条件文が可能であることを示した。後者の非内容的条件文には、認識レベルでは(9)のような認識条件文、発話行為レベルでは(10)のような発話行為条件文等がある（藤井 2012 参照）。

（ｉ）内容(Content)レベル　　(8)名医が早期に手術をすれば、きっと良くなるだろう。
（ii）認識レベル(Epistemic)　(9)良くなったのなら、きっと名医が早期に手術したのだろう。
（iii）発話行為(Speech Act)レベル
　　(10)手術をご検討でしたら、こちらに可能医療機関と医師のリストがあります。

このような事態把握領域による条件文の類型に鑑みて、談話標識・語用標識的な用例を分析すると、内容的条件文（予測的条件文）もみられるが、非内容的条件文（非予測的条件文）が多く、発話行為条件文の類や、認識条件文・メタ言語条件文の特性を呈する類が多いことが分かった。先に3.1節において「なら」「だとしたら」「だったら」「じゃ」が、3.2節において「そんなら」「そうだとしたら」「それだったら」が、談話における先行発話を照応し、その先行発話の事態を前提にして後続発話での発話行為を行っていることを示す標識として機能していることをみた。その発話行為の様相に変容はあれ、概ね発話行為条件文としての特色を共有している。

（11）簡単に言う と 、円レートが上昇すると、外貨建ての輸出価格は上昇し（[括弧内省略]）、輸出数量は、減少する。BCCWJ_OW:OW4X_00505b
（12）白状すれば、わたしも彼女にはひかれていた。（書籍: 文学: LBe9_00203b）
（13）したがって悲しい出来事。ということで。(F え)お話しする場合に。あまり私自身(F えー)残ってないんですが。強いて(F えー)言えば。長いこと(F おー)飼ってた(F あー)犬が。(F あー)老衰で死んだと。いうことを思い出して。（CSJ2004: S02M1698）

４.２　非内容的条件構文の機能細目

日本語の非内容的条件文のこれらの用例を分析するために、Speech act modifiers, Rhetorical connectors（Fujii 1993, 1994, etc.）などの機能細目を用いてきた。「「はっきり言えば」「正直言うと」「何を言いたいかというと」「本音をいえば」「本当のことをいえば」「どっちかといえば」「白状すれば」（例 12）等は、発話に発話者自ら註釈付与をしつつ条件付きの発話・発話行為を行うための談話標識としても使われている（Speech act modifiers 発話行為調節標識）。

「要約すると」「簡単に言うと」（例 11）「言い換えれば」「さらに言えば」「詳しく言えば」「結

論をいえば」「より正確に言うと」等も同様に発話行為の条件付けに寄与するが、同時に談話構造の中での結束性指標や発話間関係の手続き的意味に寄与している。「この立場でいくと、…となる」「…比べると、…」等、表現上内容的条件文の論理構造を帯びている場合も、前件で談話での論旨展開上の条件付けをし、後件でその見地・視点・立場での結論を述べる発話であり、認識条件文と発話行為条件文とメタ言語的条件文との折衷バリエーションであると捉えられる。

表２ 語用標識・談話標識として用いられる条件構文：述語を含む合成型（一部例）[BCCWJ]

		と	（れ）ば	たら	なら
speech act modifier 発話行為調節	B C C W J	**「…言えば」系** 「ほんというと」 「じつをいうと」 「本心をいうと」 「率直にいうと」 「あえていうと」 「正直言うと」 「本音をいうと」 「本当のことをいうと」 「告白すると」 「極言すると」 「極論すると」等	**「…言えば」系** 「実をいえば」 「(もっと」はっきりいえば」 「早くいえば」 「遠慮なくいえば」 「あえて言えば」 「言ってみれば」 「正直言えば」 「本音をいえば」 「本当のことをいえば」 「ほんとをいえば」 「白状すれば」 「極言すれば」 「極論すれば」等		「敢えていうなら」
⇧ {*相互排他的な特質ではない*} ⇩					
meta-linguistic marker メタ言語的		「簡単にいうと」 「一口にいうと」 「どちらかといえば」 「まとめていうと」	「言い換えれば」 「さらに言えば」 「詳しく言えば」 「概していえば」	「簡単にいったら」 「率直にいうたら」	「例えていうなら」 「ついでにいうなら」 「さらにいうなら」
rhetorical connector 修辞連結的		(その)概要を述べると 参考まで述べると 「ついでにいうと」 「何かというと」	「大きくいえば」 「結論をいえば」 「例えていえば」		「なぜなら」
⇧ {*相互排他的な特質ではない*} ⇩	B C C W J	**[　　]かといえば** 「なぜかというと」 「その意味でいうと」 [　　]かというと 「…比べると」 「[スライド,図,表等]をみると」 [TOPIC]というと [TOPIC]について触れると [TOPIC]を参考まで述べると	「わかりやすくいえば」 もっと分かりやすくいえば 「補足すれば」 **[　　]かといえば** 「なぜかと言えば」 「それはなぜかといえば」 「どちらかと言えば」 [TOPIC]といえば	[TOPIC]ときたら [TOPIC]ったら	「常識的に考えるなら」 [TOPIC]なら
perspective-taking marker 視点・観点設定		「これでいくと」 「この立場でいくと」 「この立場でいうと」			
evidential marker 証拠性ソース表示		Xによると [一人称]にしてみると:「私にしてみると」「僕らにしてみると」 [三人称]に言わせると	Xによれば [一人称]にしてみれば:「私にしてみれば」「僕らにしてみれば」 [三人称]に言わせれば		

speech act modifier 発話行為調節	C S J	[　　]かというと [TOPIC]っていうと 「いってしまうと」 「厳しく言うと」 「そういう点でいうと」等	「敢えていえば」 「(簡単に)いってしまえば」 [　　]かといえば [TOPIC]といえば、等		「添えるなら」
				＜**CSJ 例：割愛** (スライドをご覧ください）＞	

５．発話(間)相互行為における条件付けの仕方に関する類型

発話(または発話間相互行為)における条件付けの仕方を考察してみると、異なる条件付けの仕方が観察できる。その大別を、著者の研究で「間主観的条件受け（話者間条件受け)」「主体的条件付け（話者内条件付け)」と呼ぶ。これは通常の複文構造の条件構文においても生じる様相であり、談話標識化した場合も同様の条件付けのバリエーションが可能であることが分かる。

５.１　間主観的条件受け（話者間条件受け）

条件文を用いる際、談話・会話における先行発話・先行文脈で提示された内容を受けて、前件が先行発話・先行文脈の内容に言及し、その内容を前提に後件の発話（疑問や依頼や言明等発話行為を伴うものが多い）が提示されることがある。まず、通常の複文構造の条件構文でのこのような発話を(14)に例示する。

(141) H: 明日サンフランシスコで会議.　ジャパン・タウンで食べようかな.
　　E: サンフランシスコにいくんだったら、紀伊国屋で本買ってきてくれない?

(14)の話者 E の発話は、話者 H がサンフランシスコに行くこと（ほぼ確実な予定）を明示的に知らされた直後に、その計画に条件構文前件で言及しているのであって、かなり蓋然性の強い事態として受け止めた上でのことである。その前件事態に対する認識的態度としては肯定的な態度であるにも関わらず、このような文脈での発話においては条件構文を使用することが多く、特に日本語では名詞化した前件形式「[　S　]のだったら」「[　S　]のなら」を使用するのが自然である。この際の前件事態は、話者以外（会話の聞き手）から会話場で与えられた内容であり、話者は他者から受けた情報を前提にしていることを言語化しつつそれを前提にした発話行為を後件で発話している。このような条件付けの様相を「間主観的条件受け（話者間条件受け)」と呼ぶ。

(15) H: 明日サンフランシスコで会議.　ジャパン・タウンで食べようかな.
　　E: だったら、紀伊国屋で本買ってきてくれない?

さて、この同様の「間主観的条件受け（話者間条件受け)」の発話を、(15)に示すように「だったら」のみで言及することも可能である。先に 3 節で、「だったら」の基盤構文として「[指示詞]だったら」をあげていたが、「[　S　]のだったら」構文も「だったら」の重要な基盤構文である。ここで重要なことは、談話標識化した「だったら」の条件付けの様相が、その基盤構文「[　S　]のだったら」構文のそれを継承していることである。先に 3 節の(3) (7)で例示した「だったら」「そうだったら」も同様に「間主観的条件受け（話者間条件受け)」で用いられている。

５.２　主体的条件付け（話者内条件付け）

一方、発話者内(筆者内)で条件付けを行う発話も多い。特に、語用標識化・談話標識化している条件構文のうち、4 節で Speech act modifier, Rhetorical connector として提示した類型（例 11）では、話者が後件で提示しようとしている発話行為や言明や認識的判断・結論をどのように解釈して欲しいかに関して話者自身が注釈を述べているわけで、その条件付けは主体的であり話者内（話者内発話、話者内思考）において成立しているものである。

５.３　談話標識「だったら」: 発話(間)相互行為における条件付けの仕方

BCCWJ において文頭・発話冒頭で談話標識として用いられている「だったら」907 事例[3]すべてに、(i)間主観的条件受け[話者間条件受け]か (ii)主体的条件付け[話者内条件付け]かに関するコーディングをしたところ、907 トークン中、43%（386 例)が間主観的条件受け（話者間条件受け)、57%（521 例)が主体的条件付け（話者内条件付け）であった。さらに、前文脈の発話機能を分析したところ、後者主体的条件付け 521 例のうち、38%が質問や確認問いかけ発話であった。

[3] 表 1 では文頭・発話冒頭の「だったら」が 916 例抽出されているが、一例ずつ文意・形式を読み取ったところ、9 例は曖昧なケースや誤解析であったためこの分析からは除外した。

６．まとめ

本稿では、条件構文基盤の談話標識化・語用標識化の諸相を、現代日本語書き言葉均衡コーパス（国立国語研究所）や、日本語話し言葉コーパス（同）や会話コーパスを用いて分析するために指標としている（発話冒頭で使用される）条件構文の形式的類型、共起語彙群、意味・機能的類型、条件付けの様相を考察した。長単位・語彙素に関する議論の種にもなれば幸いである。[4] 一方、類型や体系付けや全体像の中での位置付けに加えて、談話標識事例一つ一つが詳細かつ綿密な分析（例えば、孫、山田等）を必要としている対象であることは申し添えるまでもない。

これらに関するコーパスを用いた分析を行うにあたり様々な研究問題・仮説が念頭にあるが、それらの中から、本稿 1-5 節で言及する紙幅のなかった問題を２点以下に挙げる。

A. 書き言葉・話し言葉（異なるレジスター・ジャンル）において談話標識・語用標識として用いられる条件構文の使用傾向： BCCWJ, CSJ, 友人間自由会話という三種のコーパスを分析してみると、談話標識・語用標識として用いられる条件構文の出現傾向が大きく異なることが分かった。BCCWJ 内でも、異なるジャンルを比べると、出現傾向が異なる（表１参照）。

B. 基幹語彙と談話標識・語用標識の機能的特徴との関係： 談話標識化・語用標識化に関する本研究は、談話における相互行為とダイナミックな文法使用の視点から行っているとともに、「語彙と構文」の観点からも興味深い。例えば、動詞等述語を含む合成型では、言動系（特に「言う」系）語彙、及び、思考・認知系語彙が顕著な共起語彙群である。これら共起語彙群の語彙的意味は談話標識としての機能的特徴（特に発話行為条件付け、認識条件付け）に寄与している。

謝　辞

本研究は、日本学術振興会科学研究費補助金基盤研究「構文理論・用法基盤アプローチによる語彙と構文彙の統合的研究」(平成 22～25 年度, 研究代表者:藤井 聖子) による助成を得ています。

参考文献

Fraser, Bruce (1996) Pragmatic Markers. *Pragmatics* 6(2): 167-190. International Pragmatics Association

Fujii, Seiko Y. (1995) Mental-space builders: Observations from English and Japanese conditionals. In Shibatani, Masayoshi & Thompson, Sandra (eds.), *Topics in Semantics and Pragmatics*. Philadelphia/Amsterdam: John Benjamins Publishing Company, 72-90.

Fujii, Seiko Y. (1995-97) Conversations in Japanese: 34 pairs of casual dyadic conversations.

Onodera, Noriko (1995) Diachronic analysis of Japanese discourse markers. In A. Jucker Historical Pragmatics. Pragmatics Developments in the History of English, Philadelphia/Amsterdam: John Benjamins Publishing Company. 393-437.

Schiffrin, Deborah (1987) Discourse Markers. Cambridge: Cambridge University Press

Sweetser, Eve. (1990) *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.

Traugott, Elizabeth C. and B. Heine. (1991) Approaches to Grammaticalization. Philadelphia/Amsterdam: John Benjamins Publishing Company

Traugott, Elizabeth C. (1995) The role of discourse markers in a theory of grammaticalization. Paper presented at the 12th International Conference on Historical Linguistics, Manchester.

Traugott, Elizabeth C. (2004) Historical pragmatics. In L. Horn and G. Ward (eds.) *The Handbook of Pragmatics*. Oxford: Blackwell. 538-561.

国立国語研究所『現代日本語書き言葉均衡コーパス（BCCWJ)』(2008 版、2009 版, 2012 版)

国立国語研究所『日本語話し言葉コーパス（CSJ)』(2004 版、2013 版)

孫羽（in progress)『「那 nà」と「だったら」の対照分析』(仮）東京大学大学院総合文化研究科言語情報科学専攻修士論文（進行中）

藤井聖子（2008)「話しことばの談話データを用いた文法研究：話し言葉で構文機能が強化する？―「～ないと」「～なきゃ」「～なくちゃ」の文法―」長谷川寿一・伊藤たかね・C. ラマール（編)『心とことば―進化と認知科学のアプローチから』、pp. 129-151、東京大学出版会

藤井聖子（2012)「条件構文をめぐって」澤田治美編『構文と意味』pp. 107-131. ひつじ書房

山田彬尭（in progress)『「そういえば」の分析からみた談話管理理論』(仮）東京大学大学院総合文化研究科言語情報科学専攻修士論文（進行中）

4　BCCWJ における単位認定は、コーパス構築段階で様々な要因が考慮され一貫性（かつコアデータからの学習可能度）を優先課題として確立された賜物であるが、本稿で扱った談話標識化に係る現象は BCCWJ のような均衡大規模コーパスが完成して新たな分析・検討が可能になる問題でもあるだろう。

ポスター発表（1） Aグループ

9月5日（木） 13:10～14:10

接続助詞「から」と「ので」に関する一考察
―前件のモダリティとの共起を手掛かりにして―

李　惠正（東北大学大学院文学研究科）†

A Corpus-based Study on *KARA/NODE*:
Focusing on Their Co-occurrence with Modality Expressions

Hyejeong Lee (Graduate School of Arts and Letters, Tohoku University)

１．はじめに

「から」と「ので」は前件と後件の因果関係を表す接続助詞であり、両者の相違点に関しては様々な観点から研究がなされている。中でも「から」と「ので」の使用と「主観/客観」との関係については、永野(1952)による指摘以来多くの研究が行われてきたが、未だ明確な結論には至っていないのが実状である。

本稿では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と呼ぶ）を調査データとして、表現者の心的態度を表すモダリティ表現と「から」「ので」の共起様相から両者の「主観/客観」という観点を再考する。

２．「から」「ので」と「主観/客観」

２．１ 先行研究

「から」「ので」の主観性の議論の始発点となった永野(1952)は以下のように述べている。

> 「から」で結びつけられる前件・後件は元来二つのものであって、それが話し手の主観によって原因結果・理由帰結の関係で結び付られる。これに対して「ので」は事がらのうちにすでに因果関係にたつ前件・後件が含まれていて、ありのままに客観的に描写する場合に使われる。つまり、「から」は話し手の主観が充分の責任を持つという意味の一方で、「ので」は主観の責任がないということになる。
>
> (永野 1952 pp.37-38)

(1)　a. 山に近いので昼間はひどく暑いが、　(永野 1952 p.36)
　　b. ?山に近いから昼間はひどく暑いが、　(筆者による改変)

永野(1952)は、(1a)のような例をあげて、「ので」を「から」に置き換えると不自然な言い方になるか、「ので」よりもしっくりしない感じのものになると説明している。

この永野説による「から」「ので」と「主観/客観」との対応関係は、その根拠の妥当性について山田(1984)、趙(1988)による疑問が呈されたのをはじめ、主張において付け加えはあ

† danahan.j@gmail.com

るものの概ね永野に同意する立場である森田(1989)や、「から」は中立的で「ので」は客観的であるとする今尾(1991)、尾方(1993)などによって活発に議論されてきた。

これらに対して国広(1992)は永野説とは逆の考え方を示している。すなわち、「のだ」の意義素と語のスコープという概念を用いて、「ので」は命題を主観的に、「から」は客観的にとらえるものとしている。

(2)　a. 当地は海岸に近いので健康によい。
　　b. 当地は海岸に近いから健康によい。　(国広 1992 p.32)

国広は(2)の例文について、(2a)は前件の内容を主観的に判断して、私は後件のように思うということを表しているのに対し、(2b)は後件が前件の帰結として生じることを客観的事実として提出し、そのことについてまったく疑いを抱いていないという含みが感じられると述べている。

このように未だに議論されている「から」「ので」の「主観/客観」の対応関係について岩崎(1995)は、両者の結びつきが主観的/客観的というのがどのようなことを意味するのかがはっきりしていないため、多くの研究で「から」「ので」の問題が「主観/客観」の特徴づけや区別問題に還元されてしまうことがあると指摘している。

２．２ 研究目的

本稿では大規模日本語コーパスを用いる実証的な方法で「から」「ので」の使用様相を観察し、「から」「ので」の「主観/客観」の対応に関して再考する。具体的には、BCCWJ の出版サブコーパス「書籍」を対象にモダリティ表現と「から」「ので」の共起関係から考察を行う。

日本語文法におけるモダリティ表現とは、文の構造の 1 つの側面であり、客観的な事柄についての話し手(言語主体)の表現時の態度の概念であると広く同意されている(ナロック 2009 p.35)。さらに、伊藤(2005 p.13)はその話し手の態度を「陳述的側面」と言い、複文の前件[1]にも適用されるとしている。

上記のことから、「から」「ので」を含む複文において前件に現れるモダリティ表現は、後件に結果を導く場合に話し手の態度を表す文の述べ方であり、文全体にも影響力があると考えられる。したがって、「から」「ので」が持つ「主観」「客観」に関する性質に応じて、前件に主に用いられるモダリティ表現形式[2]との共起関係において相違点が存在すると考えられる。共起頻度が高いモダリティ形式から「から」「ので」の「主観/客観」という性質について調査することで従来とは異なる観点から考察が可能になることを期待する。

---

1 一般的に複文の従属節、従属句と呼ぶ。

2本稿におけるモダリティ表現については、モダリティ表現そのものの機能の相違や客観化を許すものとの区別を図るわけではない。各モダリティ表現の詳細な機能は益岡(1991)、森山･安達(1996)、益岡･田窪(1992)に委ねる。

３．データの収集と分類

３．１ 調査対象

本研究では、「から」「ので」とモダリティ表現との共起関係について、国立国語研究所が開発した BCCWJ の出版サブコーパスの「書籍」を検索ツール「中納言」を用いて調査した。

出版サブコーパスの「書籍」[3]は 2001 年から 2005 年の間に国内で発行されたものが対象であり、書き言葉が生み出される出版の実態に着目したものである。また、文字言語ではあるが、地の文と会話文が混ざっており、広範囲で多様な使用場面が存在していることが利点である。

この出版サブコーパスの「書籍」に対し、「から」と「ので」をキーワードとして長単位検索で全数調査を行った。検索条件の詳細は以下のとおりである。

「から」は、語彙素が″から″、品詞が″接続助詞″、「ので」は、語彙素が″のだ″、活用形が″連用形−一般″で検索キーワードを指定して検索を行った。

「中納言」は必要な資料のみを効率よく取り出せるようにキーワードの前・後の条件を指定することが可能なツールである。この「中納言」の機能を使い、前方共起を丁寧体と普通体の二つに分けた。丁寧体とは、「です」「でした」「でしょう」「ます」「ました」「ません」の活用形を、普通体とは「コピュラ(だ/な)」「動詞終止形」「形容詞終止形(~い)」「形容動詞(だ/な)」「助動詞」を指す。これらをそれぞれの前方共起の条件と指定し、検索を行った。中納言を用いた接続助詞としての「から」「ので」のより詳細な抽出方法については李(2013)を参照されたい。

本調査では、出版サブコーパスの「書籍」のコア、非コアデータのすべてを対象として、KWIC データを収集し、目視による確認作業を行い、接続助詞として使われた「から」「ので」がモダリティ表現と共起されている例のみを分析対象とした。「からといって」「からって」などの複合辞、「～は、～からだ。」のような終助詞的な形式、後件が現れてないもしくは、後件に述語がなくて後件の意味判定が曖昧な文は対象外とした。

３．２ 調査結果

本調査は共起数が多いモダリティ表現の相違から「から」「ので」の「主観/客観」について考察を行うことが目的であるため、出版サブコーパスの「書籍」の全データに対して実数調査を行った。そのためモダリティ表現においても実際に現れた形式とその数をカウントしている。しかし、本稿では紙幅の制約で出現数がもっとも多い表現形式順から表示することとし、出現数 15 件未満の形式は表からは省略した。

次の表 1 と 2 に、BCCWJ の出版サブコーパス「書籍」から接続助詞として使われた「から」「ので」複文の総件数と、そこから「から」「ので」とモダリティ表現との共起出現数が多い順をまとめて表示した。また、両者の複文の総出現数が異なるため、両者の複文の総出現数に対する各モダリティ形式の 100 件あたりの調整頻度[4]も合わせて示した。

調査の結果、「から」複文の総出現数は 14,307 件、「ので」複文は 21,415 件であった。その中で前件の末にモダリティ表現が用いられた「から」の文は 1,219 件で、「から」複文全体の 8.52%を占めている。「ので」の文においては 914 件検索され、「ので」複文全体の 4.26%を占める結果であった。

---

[3] 山崎他(2011)に詳しい。

[4] 石川他(2010)『言語研究のための統計入門』の付属 CD-ROM を用いて分析した。

表 1.「書籍」におけるモダリティ表現と「から」

出版サブコーパス「書籍」 / モダリティ表現	から複文（総 14,307 件） 出現数(件)	100 件あたり調整頻度(回)
わけだ	366	2.56
だという	147	1.03
わけではない	59	0.41
たい	58	0.41
はずだ	56	0.39
ようだ	47	0.33
と思っている	39	0.27
そうだ(伝聞)	36	0.25
と思う	35	0.24
しそうだ	34	0.24
しようとする	32	0.22
ということだ	31	0.22
することだ	29	0.20
してもいい	28	0.20
かもしれない	27	0.19
たくない	26	0.18
らしい	18	0.13
しなくてもいい	15	0.10
しなければならない	15	0.10
べきだ	1	0.01
モダリティ総出現数	1219	8.52
全体割合	8.52%	

表 2.「書籍」におけるモダリティ表現と「ので」

出版サブコーパス「書籍」 / モダリティ表現	ので複文（総 21,415 件） 出現数(件)	100 件あたり調整頻度(回)
たい	94	0.44
と思う	87	0.41
わけではない	83	0.39
しなければならない	82	0.38
しそうだ	57	0.27
だという	54	0.25
ようだ	44	0.21
と思っている	36	0.17
しようとする	35	0.16
たくない	30	0.14
と思われる	29	0.14
わけにはいかない	29	0.14
ということだ	29	0.14
すればいい	23	0.11
かもしれない	23	0.11
はずだ	18	0.08
することだ	15	0.07
べきだ	1	0.00
モダリティ出現数	913	4.26
全体割合	4.26%	

次節では、その使用背景と「から」「ので」の「主観/客観」の性質を関連付けて考察する。

4．分析

表 1 と 2 からモダリティ表現形式の詳細を比較すると、両者には次のような相違点が見られた。

第一に、「から」と最も共起頻度が高いモダリティ表現は「-わけだ」、「-だという」のように一般的に受け入れられている事柄や社会的通念を言う時に用いられる表現が多かったのに対し、「ので」は「-たい」、「-と思う」のように述べる事柄に対して話し手の個人的な考えや不確かなことを表す際に用いられる表現が多かった。

第二に、「から」の複文では上位の 3 つのモダリティ表現形式が全体の半数を占めており、使用されるモダリティ表現に偏りがあることが分かる。それに対して、「ので」は「から」に比べて上位の表現形式に偏重される傾向は見られない。

以上の結果から両者のモダリティ表現との共起には異なる様相があることが明らかになった。

益岡(1991)ではモダリティ表現を類型化してそれぞれの特徴と役割によって 9 つに分類し、「モダリティのカテゴリー」と呼んでいる。中でも「真偽判断」、「価値判断」、「説明」、「みとめ方」の 4 つのモダリティ類型[5]が本稿と関連されると考えられる。そこで、益岡(1991)の「モダリティのカテゴリー」の分類とそれぞれの機能を参照し、全体の使用傾向を観察したあと、両者と共起頻度が高いモダリティ表現形式を考察することにする。

表 3.「から」「ので」と共起するモダリティ表現の分類

区分	機能	分類	検索表現	「から」複文 (14,307 件)	「ので」複文 (21,415 件)
第1類	動きに対する評価 (価値判断)	必要	しなければならない　しなければいけない しないとならない　しないといけない せねばならない　しなくちゃならない しなくちゃいけない　しなくてはならない	43 (3.53%)	109 (11.94%)
		許可	してもいい　しなくてもいい　方がいい すればいい　しない方がいい	51 (4.18%)	43 (4.71%)
		義務	するべきだ　するべきではない することだ　せざるをえない するほかない	33 (2.71%)	23 (2.52%)
		禁止 望ましくない	わけにはいかない　してはいけない してはならない　するといけない しちゃいけない	27 (2.21%)	42 (4.60%)
第2類	出来事の確からしさを述べる (真偽判断)	不確かなこと	かもしれない　かもわからない　と思う しかねない　に違いない　って思う と思われる　と思っている	114 (9.35%)	190 (20.81%)
		状況からの判断	はずだ　ようだ　みたいだ　らしい しそうだ　しそうに(も)ない しそうではない	164 (13.45%)	143 (15.66%)
		伝え聞き 確実、確からしさ (説明)	わけだ　わけである　わけではない だという　っていう　ということだ とのことだ　ってことだ　そうだ(伝聞) ~~のである~~	655 (53.69%)	191 (20.92%)
第[6]3類	疑問・確認	尋ねる・疑う	だろう?、のか、もんか、じゃないか、でしょう?		
		確認・同意の求め	終助詞、ね、よね、じゃないか、でしょうなど		
第4類	意志・勧め	話し手の意志的な動作	しようとする　つもりだ(意志) つもりはない　しないつもりだ したつもりだ(主語の考え、思い込み)	45 (3.69%)	43 (4.71%)
		話し手の希望、希求	たい　たくない	84 (6.89%)	124 (13.58%)
第5類	依頼・命令	話し手が聞き手の行動を要求	てほしい　てほしくない	3 (0.25)	5 (0.55)
総出現数(件) 100 件当りの調整頻度(回)				1,219 (8.52)	913 (4.26)
「から」複文の全体に対する割合				8.52%	4.26%

[5] 言及した 4 つの他に「態度伝達」、「ていねいさ」、「表現類型」、「テンス」、「取り立て」のモダリティがある。

[6] 第 3 類に属するものと第 2 類の「-のである」は、複文の前件には接続され難いため、本調査では対象外とした。

４．１「から」「ので」とモダリティ表現の使用傾向

まず最初に、「から」「ので」がどのような機能を持つモダリティ表現と共起し易いかという全体的な傾向を観察するために、本調査で検索されたモダリティ表現形式を森山･安達(1996)の分類[7]に沿って出現数を整理する。

表 3 では、各モダリティ表現を機能によってグループ分けし、同じグループに属する表現の出現件数をまとめて表示した。また、出現件数の下の括弧の中には「から」「ので」のそれぞれの総出現件数である 1,219 件と 913 件において各グループが占める割合を表示した。さらに、3.2 節の表 1 と 2 に省略したモダリティ形式を検索表現のところに全て示した。

「から」は「ので」に比べて前件にモダリティ表現を用いる頻度は高いが、使用する表現形式に偏りが見られる。「ので」よりも「から」との共起頻度が高いモダリティ表現形式は動きに対する評価を表す機能を持つ「義務」のモダリティ表現(第 1 類)と出来事の確からしさを述べる「伝え聞き」を表す表現形式(第２類)である。特に、「伝え聞き」を表す「わけだ」、「という」、「そうだ」などの表現は、モダリティ表現と共起している「から」全体の半数を占める 53.7%の割合で用いられている。これは「わけだから」、「というから」、「そうだから」のように、前件の事柄、もしくは命題が、ある確かな情報源から得られたものであることを説明する場合に「わけだ」、「という」、「そうだ」などのモダリティ表現で限定した後、その事柄について話し手の主観を交えずに後件の帰結に連結する際には「ので」より「から」が選好されることを示している。

また、益岡(1991)で言う「価値判断モダリティ」に属する「義務」(第 1 類)の表現では、「ので」の共起頻度と僅差であるが、「から」のほうが選好されている。「義務」に関するモダリティ表現は、論理的にもしくは一般的に許容される範囲の中で望ましいことや必要だと認められることへの許可と義務の意を表しており、それに対する後件には前件から考えて当然な帰結や必要性を表す事柄がつながる。そのような際に「ので」より「から」が選好されていることから、「から」は前件と後件との妥当性をより強調する役割をすると考えられる。

一方で、「ので」は特定のモダリティ表現に偏らず、多様な表現と共起している。これは「から」より「ので」のほうが多様な場面と話し方に包括的に対応できる機能を持っているためと考えられる。また、「ので」は第 2 類の不確かなことを表すモダリティ表現との共起が多い。これらのモダリティ表現は益岡(1991)の「真偽判断モダリティ」に属するものであり、中でも確信を持って言い切れない場合に何らかの形式を用いて断定を保留する表現である(益岡 1991 p.110)。ある事柄に対して断定を保留する場合、その判断は話し手自身の責任である必要がある。さらに、話し手の判断は何らかの理由で確定的な言い方を避けていることになる。このように話し手の不確かなことに関する判断を「かも知れない」、「はずだ」、「みたいだ」などの共起形式を用いて表しており、その前件の不確かな事柄に対して話し手自身が主観的にとらえるのが「ので」の機能であると言える。さらに、「ので」は「意志」と「希求」の場合(共に第 4 類)と「依頼」の場合(第 5 類)にも「から」よりも選好される傾向が見られる。

以上のように、「から」は特定のモダリティ表現との共起が多く現れ、これは「から」が事柄をそのまま伝える場合に主に使用され、その根拠となる事柄に対してより妥当性を強調する機能を持つためであると考えられる。その反面、「ので」は多様なモダリティ表現と幅

---

7 森山･安達(1996)では「-たい」「-たくない」について言及はなかったが、加藤(2006)では希求を表すモダリティ助動詞として分類してあるため筆者の判断により第 4 類に含めた。

広く共起しており、話し手自身の直感的な判断による感覚で前件と後件を接続する機能を持つと言える。

４．２ モダリティ表現と「から」

「から」と共起数が多い上位のモダリティ表現の中で注目する点は「わけだ」というモダリティ表現の使用である。「わけだから」の形式は調査した表現の中でもっとも多い 366 件(100 件当り 2.56 回)使用された。

益岡(1991 p.145)によると「わけだ」は、説明モダリティの範疇に属するもので、その文に対する帰結を得るためにはその命題が話し手と聞き手が共有する一般的な知識によるものでなければならない。言い換えると、「わけだ」が使用されるためには、話し手のみが知っている知識に根拠するのではなく双方が承知している事柄である必要があるということになる。そのようなモダリティ表現と共起する「から」は一般的な事柄を事実として受けて後件の結果のほうに当然な成り行きで導く役割を持つ。つまり、話し手と聞き手の相互が承知している前件の事柄を「わけだ」というモダリティ表現を使用することでより客観化し、「から」と共起させることで確信を持って説得的に提示している。

(3) a. 次に、効果的な接待の仕方を考えること。接待は相手に喜んでもらうためにするわけだから、相手の好みに合わせることが必要だ。たとえばお酒の飲めないお客をバーで高いお金を使って接待しても、喜ばれるどころか、嫌な思いをさせるだけになる。 (PB13_00699)

b. 接待は相手に喜んでもらうためにするから、相手の好みに合わせることが必要だ。 (筆者による改変)

「相手の好みに合わせることが必要だ」という後件の理由を述べる場合において「接待は相手に喜んでもらうためにするわけだから」という答えと「接待は相手に喜んでもらうためにするから」のという判断の答えにおいて、モダリティ表現を使わない(3b)のほうが「接待というのは相手に喜んでもらうために当然すべきである」というような強い口調に感じられる。このように、「わけだ」というモダリティ表現を用いて一般化させた前件を受ける際には「ので」より「から」を主に用いる。「から」は前件に対する後件の事柄の妥当性を高める役割をする。

次に、共起頻度が二番目に高いのは「−という」の形式である。ある情報が自分の考えではなく、第三者から得た情報である場合、自分の考えを加えずにそのまま伝える際に用いるモダリティ形式である。

(4) a. その討論会では近畿代表に選ばれ、東京まで行ったというから、半端じゃない。 (PB32_00231)

b. 川釣り専門で、年に何回か休みをとっては、世界中の川を釣り歩いていたというから優雅なもんだ。 (PB29_00723)

(4)の例は前件で述べる外からの情報をそのまま受け入れて後件で述べる判断の根拠とする例である。「という」のモダリティ表現を用いて前件の事柄を一般化し、「から」で接続

することで前件の事柄から考えると一般的にそう思われるという意を後件が表している。しかし、(4)の例における「というから」を「というので」に置き換えると、前件が提示する根拠を受けた後件の当然性が、「から」の場合に比べて落ちることになり、話し手のみの感覚を表す感じになる。

以上、「から」と共起するモダリティ表現の中で使用頻度が最も高い「わけだ」、「という」の二つの形式を考察した。両形式は前件で提示したことから後件の事柄が必然的に導かれることを納得する意を表し、「ので」より「から」が選好して用いられる。この際、「から」は前件の根拠をより客観化する役割を持ち、後件で述べる事柄が一般的な当然性を持つようにする。本研究の調査から得られた結果は、従来の研究における「から」は主観的に前件と後件を結びつけるという主張とは異なる結果であると言える。

４．３ モダリティ表現と「ので」

「ので」ともっとも共起数が多いモダリティ表現は「-たい」、「-と思う」、「-わけではない」である。

まず、「たい」は希求を表すモダリティ表現として分類され、話し手の行動や状態について希望を表すものである。森田(1980)によると、「たい」を用いて第三者の希望を表すためには「彼は水が飲みたいのだ/飲みたがっている」のような言い方をしなければならない。したがって、「たい」というモダリティ表現は話し手自身の希望を表しており、正に主観的な表現であると言える。

(5)　a. まさに「子は親の鏡」。私は子どもの笑顔を見たいので、部屋のあちこちに鏡をつるして、そこに自分の顔がうつる度に笑顔を作るように努力をしています。いつも笑顔で子育てをしたいと思います。　(PB13_00234)

b. 私は子どもの笑顔を見たいから、部屋のあちこちに鏡をつるして、そこに自分の顔がうつる度に笑顔を作るように努力をしています。

(筆者による改変)

(5a)の「私は子どもの笑顔を見たい」という前件はモダリティ表現「たい」を用いて話し手の本人にのみ当てはまる事柄として提示されており、「どうして部屋のあちこちに鏡をつるすか」という後件の理由として前件のように考えるのも話し手のみの主観的な感覚である。一方で、(5b)のように「見たいから」に置き換えると同じく「たい」というモダリティ表現を用いているが、「子どもの笑顔を見るために部屋のあちこちに鏡をつるすという行為は一般的に皆が行っていること」という異なる感じになる。言い換えると、(5b)ではそうすることは当然という前件の事柄に対する判断を一般化してしまうのである。つまり、(5a)のように前件の事柄を「ので」を用いて主観的に接続する一方で、(5b)のように「から」を用いると客観的な命題として捉えられるのである。

次に、「-と思う」はその場の証拠や状況からの事柄に対してそれが個人的な考えで確実だとは言い切れない場合に使われる。直前に述べた物事を自然に知覚・認識し、判断内容を直感的に下して自分の意見として示す、または自分の感覚や感情を示してその気持ちを自然に感じる場合に用いるモダリティ表現である。

(6)　a. ･･･ 高回転域になった時に回転が重くなるっていう結果が出たんだよ。みんなも興味があると思うので、もう少し詳しく話をすると、七千五百回転を過ぎたあたりで重くなる。 (PB4n_00003)

b. みんなも興味があると思うから、もう少し詳しく話をすると、七千五百回転を過ぎたあたりで重くなる。 (筆者による改変)

(6a)の前件も(5)と同様に解釈することが可能であると考えられる。「みんなも興味があると思う」というのは話し手自身の考えであり、あくまでも自身の感覚として主観的に捉えている。(6b)のように前件と後件を「から」を用いて接続すると、「みんなが興味を持つのは当然のことである」という言い方にとらえられる。

以上のようなモダリティ表現の機能から分かるように主観的であるというのは、第三者もしくは聞き手を想定に入れないで表現をすること、またはある命題についてそれに関する情報が話し手のみのものであることであると言える。話し手が述べる事柄が一般的に同意されていることではない場合や、ある情報が予め聞き手と共有されていない場合は主観的な事柄になり、「ので」が選好される。その一方で、一般的に容認されているか、その情報が聞き手も承知しているのであればそれは客観的な事柄であり、「から」が選好されると言える。

５．まとめ

本研究では、永野(1952)以来度々論じられてきた「から」の主観的、「ので」の客観的という性質について、話し手の心的態度を表す前件のモダリティ表現との共起関係からその関連性を考察した。BCCWJ の出版サブコーパス「書籍」を対象にして、モダリティ表現を用いる文が前件に現れた際の「から」「ので」との共起様相を調べた。その結果から「から」と「ので」は前件と後件をどのように接続するかを考察し、以下のような結論が得られた。

第一に、「から」は共起されるモダリティ表現に偏りが生じており、主に使われる表現が限定される一方で、「ので」は多様な種類のモダリティ表現と共起されており、「から」より包括的な範囲で使用することが可能である。

第二に、「から」は「わけだから」、「というから」など、話し手が前件の事柄をそのまま伝える際に最も使用される。その場合、話し手は前件の事柄に対する確信が高いことをモダリティ表現を用いて提示し、「から」は後件をより客観化させて一般的な当然の成り行きの結果として接続する機能を持つ。

第三に、「ので」は「たいので」、「と思うので」、「かもしれないので」など、話し手自身の希望や事柄に対する直感的な感覚を表すモダリティ表現と主に共起される。この場合は前件の事柄が話し手の意思に近いことをモダリティ表現で際立たせて、後件を主観的にとらえる機能をする。

以上の結果から、このような本稿の結果は、永野(1952)の主張とは異なる結果となり、国広(1992)が述べる「から」の性質に関する主張を支持する結果となった。すなわち、モダリティ表現と共起する「から」と「ので」は、前件のモダリティ表現の影響を受けて後件をそれぞれ「主観的」、「客観的」にとらえており、モダリティ表現を使用しない「から」「ので」文より後件に強く影響を与えると言える。

本稿では「から」「ので」の「主観/客観」の性質を把握するために前件に用いられるモダリティ表現との共起様相を観察し、全体的な傾向から考察したが、接続に関する制約などの詳細については次回の課題とする。

参考文献

石川慎一郎・前田忠彦・山崎誠(2010)『言語研究のための統計入門』くろしお出版.
伊藤勲(2005)『条件法研究 - いわゆる接続助詞をめぐって』近代文芸社.
今尾ゆき子(1991)「カラ、ノデ、タメ―その選択条件をめぐって―」『日本語学』10:12, pp.78-89. 明治書院.
岩崎卓(1995)「ノデとカラ―原因・理由を表す接続表現―」『日本語類義表現の文法(下)複文・連文編』宮島達夫・仁田義雄編. くろしお出版.
尾方理恵(1993)「「から」と「ので」の使い分け」『国語研究 - 松村明先生喜寿記念編』pp. 844-861. 明治書院.
加藤重広(2008)『日本語文法入門ハンドブック』研究社.
国広哲弥(1992)「「のだ」から「のに」・「ので」へ―「の」の共通性―」『日本語研究と日本語教育』pp. 17-34. カッケンブッシュ寛子他編.名古屋大学出版会.
趙順文(1988)「「から」と「ので」―永野説を改釈する―」『日本語学』7:7, pp.63-77. 明治書院.
永野賢(1952)「「から」と「ので」とはどう違うか」『国語と国文学』29:2, pp.30-41.
ナロック, ハイコ(2009)「モダリティと文の階層構造」『言語』38:1, pp.34-41.
益岡隆志(1991)『モダリティの文法』くろしお出版.
益岡隆志・田窪行則(1992)『基礎日本語文法』くろしお出版.
森田良行(1989)『日本語の類義表現』創拓社.
森山卓郎・安達太郎(1996)『日本語文法セルフ・マスターシリーズ 6 文の述べ方』くろしお出版.
山崎誠(2011)「「現代日本語書き言葉均衡コーパス」の構築と活用」『「現代日本語書き言葉均衡コーパス」完成記念講演会予稿集』pp.11-20.
山田みどり(1984)『研究資料日本文法』鈴木一彦・林巨樹編第 5 巻助詞編.明治書院.
李惠正(2013)「接続助詞「から」と「ので」の接続文体について - 日本語コーパスを用いて - 」『日語日文学』大韓日語日文学会(韓国) 57, pp.49-61.

BCCWJ 教科書データより抽出した頻度情報に基づく日本語ライティング指導教材の作成

堀 一成（大阪大学　全学教育推進機構）†
坂尻 彰宏（大阪大学　全学教育推進機構）†
石島　悌（大阪府立産業技術総合研究所）‡

Creation of Teaching Materials for Japanese Academic Writing, Using Frequency Information Retrieved from the BCCWJ School Text Data

Kazunari Hori (Osaka University, Center for Education in Liberal Arts and Sciences)
Akihiro Sakajiri (Osaka University, Center for Education in Liberal Arts and Sciences)
Dai Ishijima (Technology Research Institute of Osaka Prefecture)

１．概要

大学学部初年次生向け論文・レポートのライティング指導の基礎データとするため、国立国語研究所が開発した「現代日本語書き言葉均衡コーパス（略称:BCCWJ）」の高校教科書データより語彙頻度情報をマイニングし、指導に活用した事例を報告する。

論文・レポートの書き方指導書などでは、文を書く際に使用する用語や言い回しの事例紹介がなされる例が多いが、その用例・文例の根拠が明示されていることはまれである。我々は、BCCWJ を基礎とすることで、特定の著者や学会に偏らないデータが得られ、その成果をライティング指導に活用することで、より広範囲に応用できるレポート作成技能を受講者に身につけさせることができると考えた。

本稿は、このような試みの第二報である。第２回コーパス日本語学ワークショップにおいては、試行との位置づけで、BCCWJ コアデータの白書データに基づく成果を報告した「堀, 坂尻(2012)」。今回は BCCWJ　DVD 版公開データの特定目的サブコーパス教科書（コーパス記号 OT）のうち高校教科書とラベル付けされたものを対象とし、動詞・名詞の頻度情報を得た。一般文でも利用される頻度が高いと判断される語を、頻度上位のリストから除き、論文・レポートで用いることを推奨する用語集として受講者に提供した。また作業を MySQL 上での SQL プログラム実行で行い、一部の自動化を実現した。実際のセミナー授業での活用の様子なども併せて報告する。

２．文章指導における言語資源活用事例と本研究の目的

これまで発行されたライティング関連書籍や教材には、少ないながらも、アカデミックな文章に使われる表現例や文例を提示し、参考にさせる優れたものがある。たとえば「二通他(2009)」は、実際の学術論文から文例をとり、用いるべき表現として紹介している。しかしその表現が採用された根拠（一般文と異なり学術的文章でより用いられやすいとする計量的根拠）は提示されていない。また BCCWJ などのコーパスデータに基づく Web 日本語作文支援システム「なつめ」「仁科（2012)」は、入力した語に対する共起情報を例文根拠情報と共に表示し、用いると良い表現を知ることができる。しかし最初にシステムに入力すべ

† hori@celas.osaka-u.ac.jp,　sakajiri@celas.osaka-u.ac.jp ,　‡ ishijima@tri-osaka.jp

き語（表現）の知識がなければ有効に用いることが難しい。

本研究では、特に大学学部初年次学生のアカデミックな表現に対する知識不足に対応するための教材を開発し、かつその教材が、教員・指導者の経験や内省によるものでなく、コーパスなどの根拠情報から定量的に得られるものとすることを目的としている。それにより、教材の客観性を高めるとともに、受講者からの信頼性をより向上させたいと考えている。

３．頻度リストの作成方法

以下に教材として提示した動詞・名詞の頻度情報を作成した手順を説明する。

（１）BCCWJ 教科書（OT）高校限定データからの情報抽出

まず、BCCWJ 教科書(OT)データを CSV 形式として MySQL に読み込み、各種フィルタリング処理を行った。まず、比較的長い特徴的な単語を抽出するため、長単位情報を基に選択することとし、品詞情報が「動詞－一般」あるいは「名詞－普通名詞－一般」となっているもののみをそれぞれ抽出した。その単語リストの出現頻度を SQL コマンドで計算し、頻度順に並べ替えた。作業の詳細については付録で説明する。

（２）一般文でも利用される頻度が高いと判断される語のフィルタリング

『日本語教育のための基本語彙調査』「国立国語研究所(2001)」に掲載されている語彙のうち、「より基本的な語」とされた約２０００語を、除去参照データとした。２０００語のうち動詞と分類される語、および一般名詞と分類されている語のリストを作成し、（１）で説明した頻度順リストから除く処理をおこなった。

（３）人手による用語選定と整形

上記のように機械的操作によって得られたリストには、大学生のアカデミックライティングにあまり用いることのない単語も含まれているので（理科や音楽の用語など、各教科でのみ使われ一般用語として紹介することが適当でないと判断した）、最後に報告者(堀・坂尻)が実際のライティング指導資料として適当と判断する語に絞り、用語表として学習者に提供した。動詞は上位 330 語、名詞は上位 312 語のリストとなっている。活用法を多様にするため、頻度上位から順に紹介するリストと、その内容を読みの五十音順に並べ替えたリストの両方を提供した。

４．作成データのライティング指導への活用

作成した頻度データを、報告者（坂尻）が担当する 2013 年度ライティング指導セミナー授業で教材として提供した。

４．１　受講者への説明

受講者には、ライティングの実践において口語的な表現を避けるための一つの方法として、あるいは、表現に迷った際の判断基準の一つとして、前述のリストの使用を勧めた。レポート作成の際に、ことば遣いに迷った場合、たとえば、五十音で表を検索してそのことばがあれば、頻度の高い使用可能なことばであることわかる。もし、表になければ、口語的表現、堅すぎる表現、たまたま一覧に無いかのいずれかと判断されるので、国語研の Web ツール（少納言と NLB）を使って文脈での用法や出典（ブログ等か書籍等か）を参照することを提案してみた。

まず、登録等の必要が無い国語研の BCCWJ 検索システム「少納言」を紹介した。配布資料で紹介した用語を利用するに際して、どのような文脈中でその語が使われているかを少納言で検索し、例をよく読んで納得してから使うべきだと指導した。

図 1 報告者（坂尻）が少納言の利用方法を担当授業で説明している場面

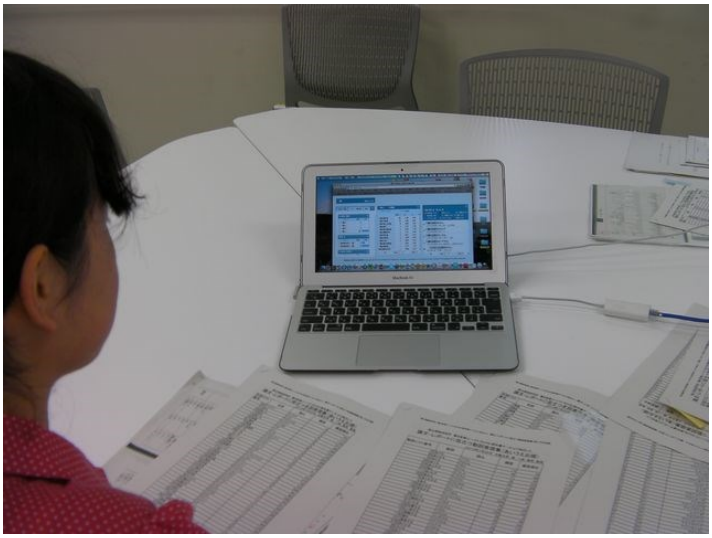


図 2　配布単語頻度データを元に NLB の利用方法を学んでいる受講学生

また、同様の用例検索ツールとして、NINJAL-LWP for BCCWJ（以下、NLB）「Pardeshi,赤瀬川(2012)」も紹介した。NLB は、国立国語研究所のプラシャント・パルデシ氏と Lago 言語研究所の赤瀬川史朗氏が中心になって開発した BCCWJ オンライン検索システムである。NLB はコンコーダンサとは異なるレキシカルプロファイリング手法を用いたコーパス検索ツールで、名詞や動詞などの内容語の共起関係や文法的振る舞いを網羅的に表示できるのが最大の特長とされている。受講者には、用例を調べようとする語について、文法項目で分けられた共起情報が細かく検索できるため、より適切な表現を見つけることができると説明した。

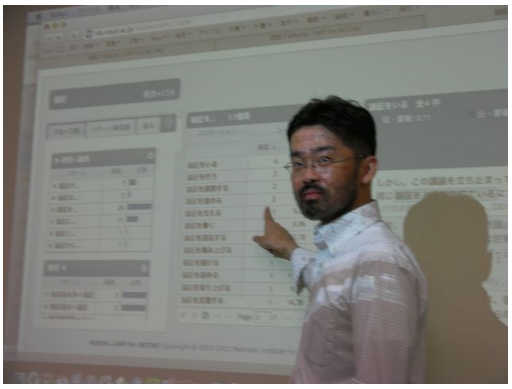


図 3　報告者（坂尻）が NLB の操作法を説明している場面

４．２　受講者の反応

（１）頻度別の一覧を見て

受講者は、論文・レポートに使用することばが、彼らにとって、必ずしも未知で難解なことばではないことに気づき、レポート作成の際に、「ことば遣い」に臆する必要のないことに納得していた。

（２）五十音順の一覧を見て

授業で各自のレポートを見ながら、気になることば遣いを表やサイトで検索し、確認させると学生の理解の度合いも高かった。

総じて一覧表に対する学生の反応は「おもしろい」「参考になる」と良好であったが、表の意味やサイトとの連携などについては細かい指導や解説が必要で、実際に使用する場合には工夫が必要である。

５．今後の展開

本報告は、BCCWJ データを教育に有効活用するための手法開発という位置づけである。今回の手順をきっかけに、さらに大規模・有用な結果がえられる手法開発へと進みたいと考えている。

◎　解析対象コーパスデータの広範化

今回の抽出は、BCCWJ データのうちアカデミックな文章に比較的近い表現（硬い文章）が多く含まれるであろうと判断し高校教科書データを解析対象としたが、BCCWJ の図書館データや書籍データの内、対象とすべきデータはまだ多数あると認識しており、適切な対象を追加選定したいと考えている。さらに BCCWJ だけでなく、学術的文章の参考になりうると判定できるものについては、広く対象としたいと考えている。近年大学が整備を進めているリポジトリに掲載されている論文データや、国立情報学研究所の CiNii 論文情報、Wikipedia の学術的項目説明文などを対象にすべきと考えている。

◎　特徴的な語・表現の抽出方法の改良

今回、特徴語の抽出方法は、得られたリストから基本的な２０００語に含まれるものを除くという、簡易な手法であった。今後適切なデータ集団の差異抽出手法を検討し、より良い抽出結果を得たいと考えている。

◎　作業手順のさらなるプログラム化

今回は BCCWJ の一部データから品詞別頻度情報を得る作業を SQL プログラム化した。それ以外の資料整形等作業は Microsoft Excel を用い、手作業で行った。今後作業対象コーパ

スの拡大を予定しており、R,Mecab なども利用し、広範囲な作業をプログラム化したいと考えている。また開発したプログラムを広く利用してもらえるよう公開する予定である。

◎　資料インストラクション手法の改善

受講生に資料の有効活用法を説明する方法も改善が必要である。前述したとおり、作成した資料のみを渡すだけでは、有効活用は期待できない。頻度リストや関連 Web ツールを使用して、より良い文を選定する具体的な方法を、文章作成指導手順に組み込み提示したいと考えている。そのためのわかりやすい教材も早急に整備したいと考えている。

６．まとめ

BCCWJ より語彙頻度情報をマイニングし、論文・レポートのライティング指導に活用した事例を報告した。BCCWJ 高校教科書データの長単位情報を選び、動詞・名詞の頻度情報を得た。頻度上位のリストから一般文でも利用される頻度が高い語を抜き、用いることを推奨する用語集として受講者に提供し、併せて BCCWJ 活用 Web システムの活用紹介も行った。

謝　辞

本研究は、学術研究助成基金助成金 挑戦的萌芽研究 課題番号: 25540163「XML コーパスからの抽出データに基づく日本語学術ライティング教材作成法の研究」（研究代表者：堀一成）による補助を得ている。

文　献

堀一成、坂尻彰宏(2012)「BCCWJ コアデータの頻度情報に基づく日本語論文・レポートライティング指導の試み」　第２回コーパス日本語学ワークショップ予稿集、pp.1-6

国立国語研究所(2001)「教育基本語彙の基本的研究」国立国語研究所報告 117

「代表性を有する大規模日本語書き言葉コーパスの構築: ２１世紀の日本語研究の基盤整備」総括班(2011)「特定領域研究『日本語コーパス』研究成果報告」

仁科喜久子　監修(2012)「日本語学習支援の構築」凡人社

二通信子、大島弥生、佐藤勢紀子、因京子、山本富美子(2009)「留学生と日本人学生のための　レポート・論文表現ハンドブック」東京大学出版会

Pardeshi, Prashant、赤瀬川史朗(2012)「コーパスを利用した基本動詞ハンドブック作成　－コーパスブラウジングツール NINJAL-LWP の特徴と機能－」言語処理学会第１８回年次大会予稿集 pp.575-578

関連 URL

国立国語研究所　コーパス開発センター　KOTONOHA 計画
http://www.ninjal.ac.jp/corpus_center/kotonoha.html

国立国語研究所 BCCWJ 検索ツール「少納言」　http://www.kotonoha.gr.jp/shonagon/

NINJAL-LWP for BCCWJ (NLB) ホームページ　http://nlb.ninjal.ac.jp/

東京工業大学「なつめ」ホームページ　http://hinoki.ryu.titech.ac.jp/natsume/

付録　BCCWJ データからの頻度情報抽出処理の詳細

以下に、３．頻度リストの作成方法（１）で略述した、BCCWJ データを Windows 上の MySQL に搭載し、頻度情報を抽出するための作業手順を詳述する。実際はステップごとに出力をチェックしつつ作業を進めたが、ここでは根幹となる作業内容のみを記す。

作業した環境は、OS: Windows 8 Enterprise 64Bit、MySQL バージョン:5.6.11 for Win64bit、Excel 2013 Professional for Win64bit である。

まず、BCCWJ　DVD 版公開データ Disk2 の LUW フォルダ下 OT フォルダ内の OT.ZIP を展開する。OT.txt が得られる。これを Excel で読み込み、作業 Index 用に index_id 列を作り、１から順に最後まで連番（最後の行は 924835）を振る。表中の数字データで使われている","を取り除き、文字コードを UTF-8 BOM なしにして、BCCWJ_LUW_OT.csv というファイル名で CSV ファイルとして保存した。

次に MySQL 上でデータを読み込むためのテーブル bccwjluwot を作成する(SQL 省略)。テーブル bccwjluwot にデータを読み込ませる SQL 文は以下のとおりである。

```
LOAD DATA INFILE 'BCCWJ_LUW_OT.csv' INTO TABLE bccwjluwot FIELDS TERMINATED BY ',';
```

以下動詞頻度表の場合のみを紹介する。高校教科書（判定は、articleID の 4 文字目が '3' かどうか）でかつ品詞が「動詞 - 一般」のデータのみを Select し、新しく VIEW とする。

```
CREATE VIEW highverblist AS SELECT index_id, l_orthBase, l_formBase FROM bccwjluwot
WHERE substring(articleID,4,1)='3' and l_pos = '動詞--一般';
```

これで highverblist に高校教科書で使われている一般動詞の長単位情報が抜き出せている。さらに動詞頻度順表をつくる。

```
SELECT l_orthBase,l_formBase, COUNT(*) INTO OUTFILE 'orderedhighverblist.csv' FIELDS
TERMINATED BY ',' FROM highverblist GROUP BY l_orthBase ORDER BY COUNT(*) DESC;
```

これで、動詞頻度順データが orderedhighverblist.csv に保存できた。

Query OK, 5978 rows affected (0.13 sec)

とのメッセージから頻度順に高校教科書使用動詞が 6000 件弱、抽出されたことがわかる。

表 1　教材とした動詞頻度表（出現頻度順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ動詞表現集（出現頻度順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
動詞リスト番号	**動詞**	**読み**	**頻度**	**頻度順位**
1	する	スル	3447	1
2	なる	ナル	3204	2
3	ある	アル	2320	3
4	いう	イウ	1356	4
5	できる	デキル	676	5
6	もつ	モツ	670	6
7	つくる	ツクル	529	7
8	わかる	ワカル	380	8
9	利用する	リヨウスル	355	9
10	みる	ミル	309	10
11	異なる	コトナル	236	13
12	変化する	ヘンカスル	223	14
13	まとめる	マトメル	168	20

表 2　教材とした動詞頻度表（五十音順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ動詞表現集（五十音順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
動詞リスト番号	**動詞**	**読み**	**頻度**	**頻度順位**
26	あげる	アゲル	112	35
90	あたえる	アタエル	48	109
52	あたる	アタル	71	65
249	あてる	アテル	18	375
228	あふれる	アフレル	20	328
324	誤る	アヤマル	11	590
111	あらわす	アラワス	39	138
329	表せる	アラワセル	11	598
75	あらわれる	アラワレル	54	91
3	ある	アル	2320	3
100	あわせる	アワセル	42	124
106	安定する	アンテイスル	40	132
4	いう	イウ	1356	4

表 3　教材とした名詞頻度表（出現頻度順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ名詞表現集（出現頻度順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
名詞リスト番号	名詞	読み	頻度	頻度順位
1	情報	ジョウホウ	384	5
2	物体	ブッタイ	296	8
3	課題	カダイ	176	28
4	しくみ	シクミ	132	40
5	動き	ウゴキ	124	44
6	役割	ヤクワリ	115	49
7	構造	コウゾウ	107	52
8	一定	イッテイ	101	56
9	現象	ゲンショウ	99	58
10	考え方	カンガエカタ	97	60
11	仮説	カセツ	93	63
12	課程	カテイ	85	70
13	手順	テジュン	84	71

表 4　教材とした名詞頻度表（五十音順）の一部

国立国語研究所　書き言葉コーパス（BCCWJ 高校教科書データ）より抽出した				
論文・レポートに役立つ名詞表現集（五十音順）				
２０１３年７月２２日　大阪大学　堀　一成、坂尻　彰宏				
名詞リスト番号	名詞	読み	頻度	頻度順位
294	アイデア	アイデア	10	1781
191	歩み	アユミ	14	1112
200	表し方	アラワシカタ	14	1153
172	安定	アンテイ	16	1012
148	言い方	イイカタ	18	860
185	意義	イギ	15	1089
281	維持	イジ	10	1729
72	意識	イシキ	30	376
179	位置関係	イチカンケイ	15	1047
150	一切	イッサイ	18	869
298	一体	イッタイ	10	1794
8	一定	イッテイ	101	56
80	イメージ	イメージ	29	411

接続助詞「けど」の音調と意味用法に関する研究
−挿入用法についての検討−

田頭 未希（東海大学 教養教育センター）†

Intonation and Discourse Function of *Kedo*
- A Case of Insertion -

Miki Tagashira (Foreign Language Center, Tokai University)

１．はじめに

田頭(2013)では接続助詞「けど」類の用法を 6 つに分類し、音調と用法の関連について、ある音調がある特定の用法と結びつきやすいという関係ではなく、話し言葉の中でゆるやかに対応していることを述べた。また用法に関しては、「けど」類節がなくても文全体の意味には影響を与えない付け足し的な説明や挿入が話し言葉では多用されている点を指摘した。

本稿の目的は、田頭(2013)の 6 つの用法のうち「挿入」用法に注目し、音調との関連性を多角的に考察することである。「挿入」用法は、『日本語話し言葉オーパス』において接続助詞「けど」類の中で使用頻度が最も高く、約半数がこの用法であった。さらに、「挿入」用法は「逆接」や「談話主題の導入」などの他の用法とは異なり、「挿入」用法でありながら、話し言葉の動的な展開の中で逆接的あるいは言いさし的といった他の意味用法を併せ持つことができ、これらの点からも考察を行う。

２．分析データ

２.１　音声資料

『日本語話し言葉コーパス』（以下 CSJ）（Maekawa 2003[1]）のコアデータのうち、韻律情報が付与されている約 18 時間分（模擬講演 107 ファイル）を分析資料とした。

分析資料全体では、接続助詞「けど」類は 1937 例で、田頭（2013）ではランダムに抽出した約半数 1019 例を分析した。本稿では、4.2 節で説明する挿入用法のみに注目しているため、田頭（2013）で扱った挿入用法の 510 例からランダムに抽出した 163 例が最終的な分析データとなる。

２.２　韻律句末の音調

本稿で使用する「韻律句」とは、イントネーションの物理的変化量として基本周波数を考え、時間軸に沿って示される音調の変化のうち、冒頭の上昇から始まり発話末にかけて下がっていく基本周波数で示されるひとつの山のまとまりを指す(Pierrehumbert and Beckman 1988)。韻律句にはIntonation Phrase[2]（以下IP）とAccentual Phrase（以下AP）の2つがある。音調の連鎖という意味では、東京方言では、ひとつのアクセント句は、「相対的に低いピッチ（%L）で始まった後すぐに上昇し（H-）、アクセント核[3]があればそこで下

† t-miki@tokai-u.jp

1 CSJ の概要について説明している論文のひとつである。

2 Pierrehumbert and Beckman（1988）ではアクセント句より階層的に上位の単位として中間句（Intermediate Phrase）と発話（Utterance）を置くが、J_ToBI ではそれらを融合した単位としてイントネーション句（Intonation Phrase）を定めている。

3 語彙的に指定されたアクセントを意味する。なお、この注釈は筆者が加筆したもので、五十嵐他(2008)は本文カギ括弧の表現である。%L、H-、H*+L などは CSJ で採用されている

降し（H*+L)、最後もまた低く終わる（L%)」という基本周波数の一連の変化からなる（五十嵐他 2008)。

CSJ では X-J_ToBI と呼ばれる韻律ラベリングシステムを採用し、韻律句末の音調の型として 5 つの型を定義している。下降調（L%)、上昇調（H%)、上昇下降調（HL%)、低ピッチ区間を伴う上昇調（LH%)、上昇下降上昇調[4]（HLH%）である。

３．接続助詞「けど」類

３.１　音調の型

『明解日本語アクセント辞典』(1997）によると、接続助詞「けど」類について語彙的に与えられている音調は「け」から「ど」にかけてアクセントを持つが、基本的には「けど」類単独で使用されることは少なく、動詞や形容詞、名詞に後続して用いられるため、次のように説明できる[5]。形容詞の場合も以下の説明の動詞の場合と同様に、起伏式形容詞の場合には形容詞の型を変えないで低く下がってつき、平板式形容詞の場合には最後の拍を変え、低く下がってつく[6]。ここでは便宜上、前接要素と比べ、低く下がる音のみを下線付きで表記する。

平板式動詞につく場合：助詞の第一拍から、低く下がってつく
例）なく<u>けど</u>　　（泣くけど）

起伏式動詞につく場合：動詞の型を変えないで、低く下がってつく
例）よ<u>むけど</u>　　（読むけど）

上記の例に示した通り、「けど」類が動詞や形容詞などに後続する場合、語彙情報として持っている音調は、前接要素の品詞やアクセント型に関わらず、前接要素に続いて低く下がってつく下降調である。

３.２　用法

「けど」類に関して、先行研究[7]では分類される用法[8]の数や用法の名称は必ずしも一致していないが、おおよそ以下のような分類があげられる（森田 1980, 渡辺 2000、永田・大浜 2001 他)。（１）談話主題の導入、（２）逆接･対比、（３）並列･累加、（４）前置き、（５）言いきりの回避･言いさし、そして本稿で扱う（６）挿入である[9]。「挿入」は「前置き」と類似しているが、「前置き」は後件の補足、あるいは後件の解釈を阻害する要因を排除するために置かれているのに対し、「挿入」は補足説明を付け加えることを意味し、「けど」節がなくても前後の文意が通ることが条件である。以下に、例をあげる。(a）は先行研究か

韻律ラベリング X-JToBI で使われる記号である。

4 本稿での分析データでは、上昇下降上昇調は全接続助詞 9,518 例のうちわずか 2 例であり、いずれも「て」の例であったため、今回の分析には含まれていない。

5 「新明解日本語アクセント辞典」（秋永 2002）の付録(72)～(74)の表より。まとめは筆者による。韻律句末の音調は、語彙情報として指定された以外の一般的音調以外に特に助詞などの類はイントネーションによって変化しやすいので注意が必要である点が明記されている（秋永 2002)。

6 平板式形容詞の場合は、形容詞の最後の拍を低く変え、低く下がってつく。

7 「けれども」の用法を４つに分類するもの（三枝 2007)、６つに分類するもの（森田 1980, 永田・大浜 2001）などがある。田頭（2013）では、主に森田や永田ら、渡辺（2000）の研究を基に、６つの用法に分類した。

8 定義は、森田（1980）のものを筆者により短くまとめている。

9 それぞれの用法の具体例は田頭（2013）を参照されたい。

らの引用、(b) はCSJから取り出した例（鍵括弧にデータのTalk ID）を示す[10]。

(a) この前貸した本を明日　もし無理だったら明後日でもいいんだ**けど**
返してくれる？（永田・大浜 2001）

(b1) そんな　ことも　ありましたし　娘と　二人で　毎日　あの　猫の　こと　書いて　あのー　夏目漱石じゃない**けど**　あのー　猫の小説でも書けるといいねなんて　[S01F1522]

(b2) あのー　主人が　ゴルフが　好きで　ま　ゴルフ場に　勤めてるんですけ**れども**　ゴルフが　好きっていう　ことも　ありまして　それも　あって　[S00F0014]

(b3) 東京ガスの　んーと　駐車場が　凄く　広く　あるんです**けど**　そこで　それは　住んでから　本当に　気づいたんです**けど**　で　駐車場なので　何も　ないから　凄く　いいなという　風に　思ってた　私　二階なんです**けれども**　そしたら　もう　毎日　雨の　日　以外は　朝八時に　そこで　みんな　ラジオ体操するんですね　[S00F0177]

４．結果

表１に形態に関する内訳を示す。どの分析対象データに関しても、「けれども」「けど」「けども」の順での使用頻度が高い。その他には、「け」「げ」「けお」「けよ」「けれとも」「けろ」などが含まれる。「挿入用法」510 例についてみると、「けれども」「けども」などの副助詞が付いた形態は 338 例で、「けど」「けれど」など副助詞が付かない形態はその約半数の 172 例であった。

表２は句末音調の割合の分布を示している。上昇下降調（HL）が最も頻度が高く、ついで上昇調（H）となっている。形態やアクセントの位置と句末までの距離は、韻律句末の音調変化に影響を与え得る要素と考えられる。形態と音調の関係を表３に、モーラ数[11]と音調の

表１　形態別の度数

	けれども	けーども	けども	けれど	けーど	けど	その他	合計
分析資料全体	708	148	370	44	6	644	15	1935
田頭(2013)の分析データ	414	88	179	13	5	313	7	1019
田頭(2013)の「挿入用法」	188	50	99	7	0	163	3	510
本稿での分析データ	71	8	17	5	0	62	0	163

10 当該要素を太字表記している。また、句読点位置と推定される箇所でのスペースはCSJの転記にて分かち書きされた箇所を示す。

11 CSJの転記が長音表記になっているものは長音も１モーラと数えた。したがって、「けーども」は４モーラ、「けーど」は３モーラに分類する。

表２　句末音調別の度数

	L	HL	H	LHL	合計
分析資料全体	417（21.5%）	881（45.5%）	630（32.5%）	7（0.04%）	1935
田頭(2013)の分析データ	202（21%）	451（46%）	317（33%）	4（0%）	974[12]
田頭(2013)の「挿入用法」	118（23%）	226（45%）	164（32%）	0（0%）	508
本稿での分析データ	32（20%）	69（42%）	62（38%）	0（0%）	163

表３　形態毎の音調の度数

	けれども	けーども	けども	けれど	けど	合計
L	17	0	0	1	14	32
HL	22	4	12	2	29	69
H	32	4	5	2	19	62

表４　モーラ数毎の音調の度数

	２モーラ	３モーラ	４モーラ	合計
L	14	1	17	32
HL	29	14	26	69
H	19	7	36	62

表５　モーラ数毎の音調の度数

	度数
並列･累加	11（1%）
談話主題の導入	71（7%）
挿入	510（52%）
前置き	78（8%）
言い切りの回避	28（3%）
逆接･対比	276（28%）
合計	974

表６　接続関係と音調

	度数
前接に関連あり	31
後続に関連あり	110
前後両方に関連あり	14
合計	155[13]

（田頭 2013 より）

関係を表４に示す。モーラ毎の音調の表出率をカイ２乗検定を用いて検定した結果、有意差が認められた（$p<0.05$）[14]。

表５に示すように、「けど」類を６つの用法に分類すると約半数が「挿入」用法であった

12 1019 例のうち、用法の分類に迷った 45 例が除外されている。

13 判定に迷った 13 例が除外されている。

14 形態毎の音調の表出率についてもカイ２乗検定を用いて検定した結果、一応、有意差が認められた（$p<0.05$）が、分割表におけるセルのうち 20%の期待度数が５未満であった。

（田頭 2013)。「挿入」は補足説明を付け加える用法で、「けど」節がなくても前後の文意が通ると定義されるものである。補足説明を加えるという点で、前後のいずれか、あるいは前後両方の発話と、内容の観点から関連性を持つといえる。そこで、挿入された「けど」節でマークされた発話が前後どの発話とより関連性が深いかを調べた。当該の「けど」節から長単位で前後 10 個までを範囲とし、前接、後続、その両方に関連しているのかを筆者が判定した。表６に結果を示す。後続する発話内容の前置き的に挿入されている例が非常に多くみられた。

５．考察

形態に関しては、「けれども」「けれど」「けども」「けど」の順で丁寧さが薄れていくと言われている。さらに、「けども」「けど」はくだけた話し方とされている。音調の面からは、極端ではない上昇調や上昇下降調を用いると丁寧な感じや優しい感じを与えると考えられている。対象とした発話は、模擬講演データで、決められたテーマに沿って誰かに自分の体験を話すという設定で発話されたもので、友達同士のくだけた会話というよりはもう少し丁寧な発話がなされたと考えられ、そのような場では、発話全体からみると「けど」節がなくても文意は通じる「挿入」のような用法においても、こられの音調が使われる頻度が高いといえる。

「挿入」としての「けど」節と前後の発話との関連性については、後続する発話により関連性が強いと判断されるものが多かった。先にも述べた通り「挿入」用法は他の用法と異なり、「挿入」でありながら、発話の動的な展開の中で補足的、前置き的、あるいは言いさし的、逆接的といった他の意味用法を併せ持つ場合がある。例えば、「挿入」の「けど」節が後続する発話とより関連が深かったということは、後続する内容の前置き的に挿入したり、またはあえて逆接的な内容を挿入したりしながら発話を行っているという解釈ができる。

①　補足的挿入用法

例えば　あの　セブンイレブンとか　の　お赤飯　ありますよね　の　お赤飯の赤い　色は　虫から　取ってるんですよ　で　て　色は　そのー　コチニール貝殻虫って　言うんですけど　その　貝殻虫は　赤い　色を　強く　出す　虫で毛糸屋さんとか　後　藍染めの　人とか　[S00F0082]

②　前置き的挿入用法

パチンコ　パチスロについて　えー　全くの　あの　素人の　立場として　お話します　え　えー　ま　やり始めと　言うか　きっかけというのは　あのー　ま子供の　時から　あのー　兄とかが　いて　よく　パチンコ屋の　話とか　聞かされて　興味を　持ってたんですけど　中学の　時に　あのー　中学　入ってん　すぐ　仲良く　なった　友達の　うちが　うちの　周りが　あのー　ん　パチンコ屋さんで　あの　[S00M0071]

③　逆接的挿入用法

あのー　キラウエア火山　っていうのは　何か　その　ま　私は　あの　ん　三原山ですか　大島とか　あちらの　方は　ちょっと　行った　ことが　ないのであのー　分からないんですけれども　日本の　その　火山の　印象っていうのとは　だいぶ　違いまして　ま　富士山は　昔　登った　こと　ありますし　浅間山も　あの　鬼押出しとかは　行った　こと　あるんですけれども　何か　本当に　そういう　印象　と　違いまして　あーのー　本当に　何か　できたばかりっていう　感じが　とても　強かったんですね　[S00F0014]

また、前接の発話内容との関連が深い「けど」節は言いさし的な印象を与えたり、前後同

じいい方を繰り返しているその間に挟まれた「けど」節は言いかえ的な挿入や並列･累加的な挿入として使われているといえる。

６．まとめ

『日本語話し言葉コーパス』を分析資料とし、接続助詞「けど」類の「挿入」用法に注目し、音調との関連性や、発話内容の観点から話し言葉の動的な展開の中で「挿入」が前後いずれの発話内容と関連が深いのかについて考察した。「挿入」用法は「けど」節がなくても前後の文意が通る補足的な発話だが、模擬講演のような場ではそのような「挿入」でも上昇調や上昇下降調を使い、音声的にも丁寧さを示していると考えられる。また「挿入」用法の下位分類的な意味用法は、前後の発話内容との結びつきによってその位置が決まる可能性が指摘できる。しかしながら、音声と用法においては、田頭（2013）、田頭（谷口）(2012)などでの考察と同様、接続助詞「けど」類や「が」においては、音調と用法が一対一の強い対応をしているわけではなく、ゆるやかな結びつきであることが本データから示された。

謝　辞

本研究は、学校法人東海大学総合研究機構「研究奨励補助計画」の助成を受けて行ったものである。

参考文献

秋永一枝 (2002)「アクセント習得法則」『新明解日本語アクセント辞典』第二版、金田一春彦（監修）秋永一枝（編）、pp.1-99、三省堂

五十嵐陽介・菊池英明・前川喜久雄 (2008)「韻律情報」『報告書 日本語話し言葉コーパス構築法』、(http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/csj_report/よりダウンロード可能)

三枝令子（2007）「話し言葉における「が」「けど」類の用法」『一橋大学留学生センター紀要』10、pp.11-27

田頭未希（2013）「接続助詞「けど」の音調と意味用法に関する予備的考察」第三回コーパス日本語学ワークショップ予稿集、pp.299-306.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_web.pdf　よりダウンロード可能)

田頭(谷口)未希（2012a）「接続助詞「が」の音調と意味用法 −『日本語話し言葉コーパス』の分析を通して−」第一回コーパス日本語学ワークショップ予稿集、pp.343-346.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_web.pdf よりダウンロード可能)

永田良太･大浜るい子（2001）「接続助詞ケドの往訪間の関係について　−発話場面に着目して−」『日本語教育』110、pp.62-71、日本語教育学会

森田良行（1980）『基礎日本語２　−意味と使い方−』、角川書店

渡辺学（2000）「逆接表現の記述と体系　ケド･ワリニ･クセニをめぐって」『現代日本語研究』７、大阪大学大学院

Maekawa, K. (2003) Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition.* 5-8. Tokyo.

Pierrehumbert, B. and M. Beckman (1988) *Japanese Tone Structure*. Cambridge, MA: MIT Press.

『太陽コーパス』における漢語表記の多様性 —コーパスの XML タグを利用した研究手法の試み—

間淵 洋子（国立国語研究所 コーパス開発センター）†

Notational Diversity of Kanji Compounds in the Taiyo Corpus: An Approach Using the XML Tags

MABUCHI, Yoko (Center for Corpus Development, NINJAL)

１．はじめに

現代語ではほぼ統一的に表記されるものの，『太陽コーパス』において複数の表記形で出現するような漢語がある。

(1) 而して大氣が之れに入るときは腹部は膀大し之れを出づるときは【縮小】す、
（1895 年 8 号　石川千代松「蝶の話」P141B20）

(2) 現今支那人が女人の足を【縮少】ならしめんとして鐵骨履を穿たしめて
（1895 年 9 号　福羽美静「教育に就て」P149A26）

上例「縮**小**」「縮**少**」は，いずれも物理的な対象が縮まって小さくなる（対象を縮めて小さくする）意で用いられており，同義の語として機能していると考えられる。この二つの表記で出現する「シュクショウ」という漢語は，現代語においてほぼ統一的に「縮小」と表記されるが[1]，『太陽コーパス』に例を求めると，双方の出現数が殆ど拮抗しており，どちらかをどちらかの誤用・誤字と簡単に言い切ることができない。ここでは，このような関係性にある語を「同義異表記語」[2]と呼ぼう。

このような「同義異表記語」には，「縮**小**」に対する「縮**少**」のように同音・同義(類義)字形による表記の他，以下に見られる「喝**采**」に対する「喝**釆**」のように，異音・異義字形による表記のバリエーションを持つものも少なくない。

(3) 之を小にしては數年前最も多く江戸ツ子の【喝采】を博した藏原惟郭は
（1917 年 6 号　浮田和民「総選挙の回顧的批評（立憲政治と群衆心理）」P011A16）

(4) 其の方に完成せる著書を公衆の前に朗誦して亦異常の【喝采】を博せしが
（1895 年 2 号　森田思軒「紀元前の著名なる航海者（承前）」P047B22）

『太陽コーパス』では，このような表記のバリエーションを，言語コーパスとしての検索性を考慮した上で，現代語における規範的表記に集約・校訂してテキスト化を行なって

† mabuchi@ninjal.ac.jp

[1] コーパス検索アプリケーション中納言を用いて『現代日本語書き言葉均衡コーパス』における語彙素「縮小」を検索したところ，その原文表記形は「縮**小**」2002 例に対して「縮**少**」7 例（うち 1 例は国会会議録の用例，3 例は同一著者による用例）であった。

[2] 田島(1998)は漢字表記語を「音」「義」「表記」の三つの構成要素により①同音・同義・同表記②同音・同義・異表記③同音・異義・同表記④同音・異義・異表記⑤異音・同義・同表記⑥異音・同義・異表記⑦異音・異義・同表記に分類する。これに倣う。

いる。そのため，上の例(2)は「縮**小**」に，例(4)は「喝**采**」にコーパス本文が改められているのだが，その校訂・原文情報は XML タグとして保持されている。

本発表では，この校訂情報を元に，『太陽コーパス』における「同義異表記語」を調査し，その一部について，出現の実態と経年変化を報告する。コーパスを利用した言語研究の方法論として，文字列の情報に加えて，コーパスの XML タグを積極的に用いて情報を抽出する試みの一端を示してみたい。

２．『太陽コーパス』

本発表で調査対象とする『太陽コーパス』は，言文一致を経て口語体による書き言葉が安定し普及する時期（明治時代後期～大正時代）の書き言葉を代表できるコーパスとして作られたものであり，月刊総合雑誌『太陽』（博文館）の明治 28（1895）年，明治 34（1901）年，明治 42（1909）年，大正 6（1917）年，大正 14（1925）年について，著作権処理ができなかった記事を除くほぼ全文を対象にしたものである。

雑誌『太陽』は，分量の多さ，ジャンルの広さ，執筆陣の多彩さ，読者層の厚さなどの点で，当時の文献資料として格別の価値を持っていることが指摘されており（田中 2005），それを裏付けるように，2005 年の『太陽コーパス』公開以来，口語化，濁点や仮名遣い等の表記法の確立・統一化，新漢語の定着など，当時の語用の実態や現代語に連なる変化の過程を明らかにした論考が多く発表された。

『太陽コーパス』の収録記事が発行された明治～大正期は，現在のように表記の規範意識が高くなく，同語に対し多様な表記が存在していたことが知られている(今野 2012)。例えば，明治前期の文献に見られる同義の異表記関係にあった表記には，「華麗」と「麗華」，「遊戯」と「戯遊」のように字順が入れ替わっているペアが多く存在することが報告されており（田島 1998），実際に『太陽コーパス』においても同様の字順の入れ替わった漢語対についての報告がある（吉川 2005）。本研究で取り上げる「同義異表記語」も，そのような明治～大正期の表記の多様性を映す現象の一つであり，『太陽コーパス』はこの現象を観察するのに有用な調査対象資料である。

３．同義異表記語

次に，本研究で扱う「同義異表記語」について確認しておきたい。

「同義異表記語」を捉えるにあたって，極めて近接的な関係にある「同音異表記語」について，意味論上の意味の関係と使用される文字の関係に基づき整理すると，以下のように示すことができる。

① 異義語：**意**義　**異**義
② 類義語：**回**答　**解**答
③ 同義語：
　A) 異体字：**国**語　**國**語
　B) 表意宛字：**本当**　**真実**（ルビに「ほんとう」）
　C) 別字：縮**小**　縮**少**

①，②は，文字の異なりが直接意味の異なりに繋がるもので，語の本来の意味から相互に区別して用いられるべき関係にある。これらの語を混用することは表記の規範という観点からは「誤用」とされるべきものだが，この混用が個人的・一時的の使用ではなく，ある程度社会的に認められ一般的に広く，また意図的に使用されるという状況を伴う場合，

これを「通用」と言って差し支えないだろう。『太陽コーパス』においても，このタイプの表記通用は存在するが，現代語においても同様の事象が多く見られ，現代語とは異なる明治期の特徴的な言語実態を捉えるという今回の目的と異なるため，分析対象とはしない。

③-A)は，「字体」の選択に関わるものであり，問題の所在が異なるため，やはり分析対象としない。

③-B)は，ルビを考慮すれば「同音同義」と言うべきものだが，当該漢語が通常の音で読まれる際に想定される，別語としての性質をも内包する表記であるため分析対象から外す。

③-C)は先に前掲例(1)(2)のようなものだが，このうち現代語における規範的な表記に対して，別字による異表記が特異発生的に出現しているとは考えにくいような例について，今回調査・分析対象とするものである。

なお，③-C)の「同義-別字」の異表記には，「同音」ではない文字によるものがあり，前掲例(3)(4)がこれに相当する。これは，本来異義の別字形として文字レベルで相互に区別して使用されるべきものであり，その混用は「誤字」と判断されるものだが，表記バリエーション発生のメカニズムや，当時の雑誌編集・組版の態度，活字の受容等を勘案する時，「同音-同義-別字」の異表記と連続的な事象として見えてくる。詳しく分析することは別の機会に譲るが，今回の調査手法によって副次的にデータを得ることができるため，その出現実態について同時に報告する。

４．調査概要

４．１　調査対象コーパス

調査には，現在構築中の『太陽コーパス』増補改訂版データを用いた。これは，2005 年の『太陽コーパス』公開時に著作権処理の関係で収録対象とならなかった記事から保護期間終了のものを追加収録したもので，今後形態論情報を付与した上で再データ化される予定のものである。今回の分析では，より多くの用例を収集する目的で，データ量の多い構築中データを用いた。

４．２　調査対象語

本研究で対象とする「同義異表記」の語を抽出するために，『太陽コーパス』の XML タグを用いた。『太陽コーパス』は，XML 形式により様々な情報が付加されているが，その中に，雑誌『太陽』の原文を校訂しテキスト化したことを示す要素「注」がある。この「注」要素は，言語研究用コーパスとしての検索性を考慮し設けられた要素で，現代語の規範と異なるような語用・表記に対して，以下の校訂種別を「分類」属性として示した上で，原文の情報（「原文」属性）と共に情報付けを行なっている（田中 2005）。分類属性に示される校訂の種別は表 1 の通りである。

このうち，「A 誤字通用」については，挙げられた例に，今回分析の対象とすべき「同義異表記語」が含まれる（下線引用者）。

<注 原文="近項" 分類="A 誤字通字">近頃</注>ノースロツプ氏より直接概畧の話を

（1895 年 7 号「樹栽日に就て」牧野伸顕 P150A08）

隨分御<注 原文="氣嫌" 分類="A 誤字通字”>機嫌</注>宜しく

（1895 年 8 号「妄語戒即ち真語律に就て」渡辺龍聖 P163B27）

そこで，今回は XML 文書から，直接このタグの情報を利用して「同義異表記語」の抽出を試みた。具体的な抽出手法を次節に示す。

表１　『太陽コーパス』XML「注」要素「分類」属性

分類	意味
A 誤字通字	誤植，誤用，漢語の漢字表記における通用。誤植・誤用のうち，以下の B～E については別の分類を立てるので，ここには含めない。
B 衍字	なくてもよい文字が紛れ込んだと思われるもの。
C 脱字	あるべき文字が脱落していると考えられるもの。
D 転倒	二つの文字の配列順が，転倒していると考えられるもの。
E 欠損	活字不良などのため，文字が欠けたり，つぶれたりしているもの。
F 濁点脱落	濁音表記が期待される箇所に濁音表記がないもの。
G 仮名遣	仮名遣いが歴史的仮名遣いの規範と異なる使われ方をしているもの。
H 正誤表	次号等の『太陽』誌上で，誤りを告知して訂正している場合は，その告知にしたがって，該当号の該当部分を修正した。

４．３　調査手順

《手順１》

XML 文書から XPath を用いて，「注」要素「分類」属性の値として「A 誤字通用」を持つものを全て抽出した。XPath は，XML 文書内で特定のノードの位置を指定することを目的とした構文である。コンピュータにおけるファイル・システムパスと同様に，スラッシュ区切りで階層を明示した式（ロケーション・パス）を使ってノードを参照することができ，比較的容易に XML 文書内の特定の要素や属性を抽出することができる。今回は，XPath 式「//注[@分類="A 誤字通用"]」を用いて情報抽出を行なった[3]。この際，「注」要素「原文」属性（校訂前の雑誌原文を格納するための属性）の情報を同時に取得した。

これによって抽出されるのは，今回調査対象として扱う「同義異表記語」の他，「。」－「、」，「ぬ」－「ね」といった類形の文字間の誤植，濁点の打ち誤り，「ある」－「あり」のような活用形の誤り等，多種多様な誤用（通用）である。本研究で対象とするのは，一時的な誤用とは言いがたく，規範表記と等価の表記として用いられる漢語表記バリエーションであるため，出現頻度を手掛かりに一時的な誤用を排除することとし，出現度数 10 以上の語を選定した。更に，固有名詞を除く 2 字以上の漢語を今回の調査対象語（規範表記と通用表記の対）と定めた。

なお，分析にあたって，通用表記・規範表記相互の関係性を見る際に，規範表記は『太陽コーパス』の校訂本文によったが，以下の２語に対しては独自に規範表記を当てた。

(ア)「シュドウ」: 通用表記(原文)「主**働**」／コーパス校訂本文「主**導**」→規範表記「主**動**」

『太陽コーパス』原文においては，「主**導**」が 1895 年の号にいずれも「主導者」の形で 4 例現れるが，その後は「シュドウ者」の連接も含めて「主**動**」の表記で 32 例が出現する。

(5) 彼は、實に討幕派の【主導】者にして、

（1895 年 3 月　落合直文「しら雪物語（承前）」P099B14）

(6) 彼等猶世界歴史の【主動】者はアリアン人種にありといふを得る乎。

（1895 年 9 号　田岡嶺雲「十三世紀に於ける蒙古民族の雄図」 P073B03）

(7) 我帝國が最大有力なる【主働】者の地歩を占めざるべからざるや、

（1895 年 12 号　川崎三郎「東邦革新」 P018A01）

[3] 作業環境：Windows8，Cygwin(1.7.17)，perl(5.14.2)＋XPath モジュール

このように「主導」と「主動」が混用されているのであれば，字形の近さと頻度から「主働」の規範表記を「主動」とするのが妥当と判断した。

(イ)「ケンジツ」: 通用表記(原文)「**健**實」／コーパス校訂本文「**賢**實」→規範表記「**堅**實」

「**賢**實」の表記は，『太陽コーパス』原文に 1 例も出現せず，辞書の表記としても見られない。「**健**實」の意に隣接する「**堅**實」を規範表記とするのが妥当と判断した。

(8) わが國の職人にも以上の如き【健實】なる美風があつたのである

（1925 年 2 号　藤原銀次郎「労働問題の解決策として工場の官僚化を排す」P034C12）

(9) 保守的にして【堅實】なる氣風を有する

（1917 年 1 号　安井正太郎「列強海軍とド級艦」P112A16）

以上によって定めた具体的な調査対象語は，表２に示す 88 語である。

表２　調査対象「同義異表記語」リスト

アイソウ(愛想/愛相), イコウ(意向/意嚮), イチズ(一途/一圖), オウタイ(應對/應待), オクビョウ(臆病/憶病),
カイチョク(戒飭/戒飾), カシャク(呵責/苛責), カッサイ(喝采/喝釆), カンケイ(關係/干係), カンケイ(關係/關繫),
カンタン(簡單/簡短), カンニン(堪忍/勘忍), カンネン(觀念/感念), カンパ(看破/觀破), カンマン(緩慢/緩漫),
キオク(記憶/記臆), キガイ(氣概/氣慨), キセツ(季節/期節), キソウ(皮相/皮想), キネン(記念/紀念),
キュウシュウ(吸收/吸集), キョウショウ(狹小/狹少), クフウ(工夫/工風), ゲキレツ(激烈/劇烈), ケネン(懸念/掛念),
ゲンイン(原因/源因), ケンキュウ(研究/研窮), ケンジツ(堅實/健實), コウカ(効[效]果/功果), コウジョウ(向上/昂上),
コウセキ(功績/効(效)績), コウテツ(更迭/交迭), コウテン(公轉/行轉), ゴウモン(拷問/栲問), コジン(個人/箇人),
サイクツ(採掘/採堀), ザンシン(斬新/嶄新), ジコ(自己/自個), シハイ(支配/司配), シマツ(始末/仕末),
シャカイ(社會/社界), シュウカク(收穫/收獲), シュウキ(周期/週期), シュクショウ(縮小/縮少),
シュッパン(出版/出板), シュドウ(主動/主働), ショウバイ(商賣/商買), ショクタク(囑託/囑托), ショチ(處置/所置),
シンコク(深刻/深酷), シンセツ(親切/信切), セイコウ(成功/成効[效]), セイセキ(成績/成蹟[跡迹]),
セイチュウ(掣肘/制肘), センエツ(僭越/潜越), センコウ(銓衡/詮衡), ゼンゼン(全然/全々), センレン(洗練/洗錬),
ソウゴン(莊嚴/壯嚴), ソウホウ(双方/相方), ソガイ(阻害/沮害), ソッセン(率先/卒先), タイテイ(大抵/大低),
タイハイ(頽廢/頽敗), タイヘイ(太平/大平), テイケツ(締結/訂結), ドウサ(動作/働作), トウタ(淘汰/陶汰),
ドウダン(道斷/同斷), ハクセキ(白晳/白皙), ヒエキ(裨益/稗益), ヒジュン(批准/批準), ヒッキョウ(畢竟/必竟),
フウサイ(風采/風釆), フクザツ(複雜/復雜), フクシャ(複寫/復寫), フシギ(不思議/不思儀), ヘキエキ(辟易/避易),
ボウガイ(妨害/防害), ボウシ(防止/妨止), ホウドウ(報道/報導), ホントウ(本當/本統), マスイ(麻醉/魔睡),
ムゾウサ(無造作/無雜作), ヨソウ(予想/預想), ヨロン(輿論/與論), ラチ(拉致/羅致), リュウギ(流儀/流義)

《手順２》

次に，ここで得られた調査対象語リストにおける規範表記を，コーパス本文全体から抽出した。この中には，実際には対になる原文の通用表記（やその他の異表記）が「注」タグの付いた形で含まれるから，先に《手順１》において抽出した注タグ付き出現例を除いて規範表記出現例とした。

５．調査結果と分析

５．１　表記バリエーションの分類

抽出した同義異表記語（表記対）は，規範表記と通用表記の関係性について，以下の二つの観点で分類を行なった。

A) 語義…辞書の記述に基づく語の意味関係（『日本国語大辞典第二版』小学館[4]による）

同義：下記「異義」「類義」に当てはまらないもの。①規範表記により表記される辞書項目において，通用表記が見出し語漢字欄，辞書欄（主要な古辞書，明治期辞書類での記載を示す欄），用例内に現れているもの。②別項目であっても語釈の記述が【「○○」に同じ。】のみのもの。③通用表記が辞書に全く現れないもの。

例：「記念」と「紀念」

き - ねん【記念・紀念】

類義：通用表記と規範表記が異なる辞書項目で，両者に意味の重なりがあるもの（どちらかがより限定された意味のみを持つ場合も含む）。

例：「観念」と「感念」

かん - ねん［クヮン：］【観念】〔名〕 (3)ある（抽象的な）物事に対する考え、意識。

かん - ねん【感念】〔名〕物事についての感じ方、考え方。

異義：通用表記と規範表記が異なる辞書項目で，両者に意味の重なりがないもの。

例：「妨害」と「防害」

ぼう - がい［バウ：］【妨害・妨碍・妨礙】〔名〕さまたげること。邪魔すること。ぼうげ。

ぼう - がい［バウ：］【防害】〔名〕害するものから身を守ること。

B) 字義…表記対における交代字同士の意味関係（白川静『字通』平凡社による）

同義：異体字関係にある文字（今回は対象としない）

類義：部首が交代している文字，字義に重なりや隣接が見られる文字

例：「更迭／交迭」における「更」と「交」

「更」①かえる、あらためる。④こもごも、いれかわる。

「交」②たがいに、こもごも、かわるがわる、入りみだれる。③とりかわす、とりかえる。

異義：上記「同義」「類義」に当てはまらない文字

例：「應対／應待」における「対」「待」

「対」①うつ、あげる、土をうつ。②むかう、あたる、あう、あいて、つい。③こたえる、あわせる、たぐえる、むかえる。

「待」①まつ、まちうける。②ふせぐ、そなえる、もてなす、あてにする。

語義と字義，双方の関係によって同義異表記語を分類した結果を表３に示す（表内下線は辞書に一切掲載のない表記）。

表３から，同義異表記の多くは，漢字列のうちの一文字を類義同音の別字（多くは部首の置換や追加・削除による極めて近接的な別字）に置き換えることにより生じていることが分かる[5]。屋名池(2004)に「「読みさえ定まれば（読めさえすれば）、どのような書き方をしてもよい」（表記と言語は多対一に対応する）という**「一意表記」**の原則がおこなわれていたものらしく、実際、この原則に違反するものはごく少数にとどまる。」と指摘される通りの実態が映しだされている。また，笹原 (2005)は，字体に生じる「同化」「衝突」[6]と

[4] Web 上で利用できる JapanKnowledge 版(http://www.jkn21.com/stdsearch/displaymain)を用いた。字義確認に用いた白川静『字通』も同様。

[5] ここでは規範表記→通用表記という発生順や起源関係を意図しない。

[6] 同化とは字体が近接する他の字体の干渉を受けて同じ構成要素を持つように変化する現象（例：「模糊」…「模」が「糊」の米偏に干渉を受け「糢糊」と表記される）。干渉を受けた結果，既に存在する他の字体と一致するケースを衝突，さらに音義が一致するケースを暗号と呼ぶ（笹原 2005）。

いう現象が『太陽コーパス』に少なからず見られること，またその中で，部首が置換された文字に同化しそれが継続的に使用される「定着」状態にあるものが散見されることを報告している。ここから垣間見えるのは，雑誌『太陽』の著者・編集者・読者にとって，文字の一部の構成要素が置換された字体を，置換前の字体と同義のものとして認識・受容する言語環境があったということである。このような環境下，音の一致によって語としての同一性が確保されれば，なおさら異表記の共存状態は容易に成り立ち得たものと思われる。

表３　同義異表記語の語義・字義による分類

語義＼字義	類義	異義
同義	愛相, 意嚮, 憶病, 苛責, 戒飾, 掛念, 關繋, 勘忍, 簡短, 觀破, 緩漫, 記臆, 氣慨, 紀念, 狹少, 劇烈, 研窮, 源因, 箇人, 交迭, 功果, 効績, 拷問, 採堀, 嶄新, 自個, 收獲, 週期, 縮少, 出板, 主働, 所置, 囑托, 深酷, 制肘, 成効, 成蹟, 洗錬, 潜越, 詮衡, 壯嚴, 沮害, 相方, 卒先, 大低, 大平, 訂結, 陶汰, 働作, 批準, 皮想, 避易, 稗益, 必竟, 不思儀, 復寫, 復雜, 報導, 妨止, 魔睡, 預想, 流義　(62)	【同音】 應待, 工風, 昂上, 行轉, 仕末, 司配, 社界, 商買, 信切, 全々, 頽敗, 無雜作, 與論　(13) 【異音】 喝采,白晳, 風采　(3)
類義	干係, 感念, 期節, 吸集, 健實　(5)	本統, 羅致　(2)
異義		一圖, 同斷, 防害　(3)

一方，異義の文字での置き換えによって生じる異表記も少なからず見られる。これらの内訳をみると，語義からの類推によって異義の文字を当てたと思われるもの（例 10）が多く，また隣接する別語からの干渉による混淆形と思われるもの（例 11），その他，極めて字形の近い異音の別字による置き換え（「采/采」「晳/晳」）によるものがある。

(10) 「**向**上」→語義から上昇のイメージ喚起→「**昂**上」
(11) 「應**対**」→隣接する「接待」等の語による干渉→「應**待**」

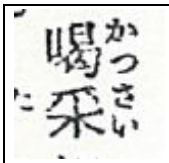


図１

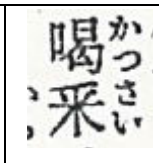


図２

幕のあとで觀客の喝**采**（図１）に應へる時にはベルナアルは立つてはゐるが、左手を輕く他の役者の肩において、それに凭りかかつてゐた。四たび五つたびと幕をあげさせる滿場の喝**采**（図２）は光榮ある彼女の過去の歴史に對する敬意を示すもので、
(1917 年 6 号　厨川白村「老女優サラ・ベルナアル」P167A25-28)

図１，図２は，同一記事の同頁同段の３行の範囲に規範表記「喝采」と通用表記「喝采」が出現している例である。これだけ隣接した箇所の同語の表記に別字「采」「采」が用いられていることから，これらは別字として強く意識されていない可能性が考えられる。

さて，表３を見ると，今回調査対象として取り出した語には，語義において類義の語（同音類義語），異義の語（同音異義語）が含まれていることが分かる。

このうち「語義」を「類義」としたものには，(a)完全に規範表記の語義に内包されるもの（「干係」「吸集」），(b)規範表記の語義と隣接するもしくは規範表記の語義と一部が重なるもの（「感念」「期節」「健実」），(c)規範表記の語義と重なる部分と規範表記の語義には

含まれない異義の両方を持つもの（「本統」「羅致」）があり， (a)は「同義」としても差し支えない。(b)は，通用表記が現代語で殆ど用いられることがないという点で校訂の対象となったもので[7]，本来は類義語の混用と同様の事象であり，厳密には今回取り上げる「同義異表記」とは性質を異にする。 (c)も(b)と同様通用表記が現代語で殆ど用いられることがないという点で校訂の対象となったものだが，弁別される語義において通用表記が用いられている場合は，むしろこれは通用ではなく正用であって，仮に現代語において当該の表記が使用されないとしても，現在では消滅してしまった語の出現と見なすべき事象である。

例えば「ラチ」では，辞書記述を見ると，「羅致」が語義の一部に「拉致」の語義を完全に内包しており，より広い意味で用いられる語であることが分かる。

ら-ち【拉致】〔名〕無理に連れて行くこと。捕えて連れて行くこと。

ら-ち【羅致】〔名〕(1)鳥を網で捕えるように、広く人材を集めること。招きよせること。(2)「らち（拉致）」に同じ。

(12) 海内の學者を闕下に【羅致】したるを以て、濟々たる多士は翰林に充滿せり、
（1895 年 6 号　中西牛郎「清朝全盛の時代」P030A13）

(13) 天心の衣鉢を襲ふの大觀も亦新たに結社するに方つて己れを虚うして逸才を【羅致】した。
（1917 年 13 号　内田魯庵「案頭三尺」P045A10）

一方，2 例のみ出現する規範表記「拉致」についてその用例を確認すると，以下のとおり「羅致」(1)の意で用いられていると思われることから，『太陽コーパス』における「ラチ」は専ら「羅致」(1)の意で用いられ，「拉致」の表記形がむしろ通用している状況と見ることができる。

(14) 一進會を作るや李容九一派の天道教徒を巧みに【拉致】して全國を風靡したる怪手腕を以て、
（190906 号　浅田江村「政治、外交　統監政治の失敗」P028A07）

また，同様に「語義」を「異義」としたもの（「一圖」「同断」「防害」）についても，本来は同音異義語の誤用・混用の現象であり，「同義異表記」とは性格が異なる。これらの現象は，今回の調査において全容を把握することはできないが，調査対象になった一部の語については，異表記通用の実態を報告することを目的として，この先の分析対象から省くことはしない。

５．２　出現率からみた通用表記

次に，それぞれの表記の出現実態について見ていく。表 4 は，同義異表記語の出現度数に占める通用表記形の割合（＝通用表記率）を示したものである。

通用表記率は，極端に高いもの低いものなどさまざまではあるが，今回調査対象とした通用表記が出現度数 10 以上の語の場合，通用表記率は比較的高く，表記の通用という言語運用が広く行われ定着していたことを伺わせる。通用表記の構成要素や辞書への記載の有無による分布の偏りはあまり見られないが，異音異義の文字による通用表記は総じて通用表記率が高い点が特徴的である。これは，先に示した通り字形認識において異なりがさほど意識されていないことに起因するものと考えられる。また，極めて通用表記率の高い「ラチ」は，その用例を精査すると，専ら前掲例(7)(8)に見た「羅致」(1)の意で用いられていることが分かる。この場合，「羅致」は「拉致」が本来持たない語義で用いられているため，「羅致」の表記で現れているものであり，通用表記独自の原義に起因するものである。

[7]『太陽コーパス』では『広辞苑』等中型国語辞典を目安に校訂対象を定めている（田中 2005）。

表４　同義異表記語の通用表記率(使用率の高い順。下線は別義字，網掛けは異音字による異表記)

通用表記率*	辞書記述あり	辞書記述なし
～100%	預想，羅致　(2)	行轉，詮衡，喝采，潜越　(4)
～68%	苛責，魔睡，紀念，緩漫，氣慨，縮少，無雜作，一圖，流義，頽敗，洗錬，同斷，記臆，主働　(14)	復寫，風采，栲問，白晢，避易，戒飾，勘忍，週期，囑托，壯嚴，制肘　(11)
～26%	健實，嶄新，成蹟，交迭，意嚮，本統，狹少，掛念，觀破，收獲，期節，工風，愛相，劇烈，憶病，出板，商買，効績　(18)	卒先，應待，稗益，深酷，採堀，陶汰，沮害，働作，仕末，報導，皮想，信切，所置，批準，訂結，功果　(16)
～7%	成効，必竟，感念，簡短，防害，箇人，吸集，源因，大平，不思儀，研窮，干係，關繋　(13)	相方，妨止，昂上，復雜，輿論，全々，司配，大低，自個，社界　(10)

＊ 分布に開きのあるところで適宜区分して示した。

５．３　経年変化

次に，通用表記の出現頻度が高い上位５語「紀念/記念」「成蹟/成績」「記臆/記憶」「成効/成功」「本統/本當」について，それぞれの表記の出現状況が年を追ってどのように変化するかを確認してみたい。図３に通用表記率の推移を示す。

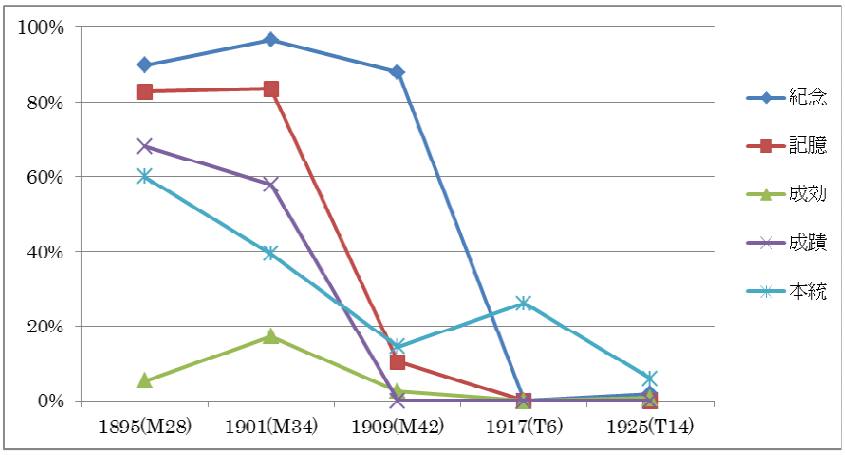


図３　通用表記率の推移（出現度数上位５位まで）

図３より，「紀念」を除く語で 1901(明治 34)年から 1909(明治 42)年で通用表記率が劇的に減少していることが分かる。1917(大正 7 年)には唯一勢力を保っていた「紀念」も例が見られず，1925(大正 14)年には 1917 年に若干勢力を戻した「本統」を含め，全ての語において，通用表記が規範表記に駆逐されている様子が見て取れる。

この背景には，明治期の国語施策が関与しているものと思われる。明治 30 年代から 40 年代にかけては，送り仮名や仮名遣いなど様々な表記についての統一施策が展開された。1916(大正 5)年には臨時国語調査会による「字体整理案」「漢語整理案」等が発表され，漢字表記に関しても統制を指向した言語政策が敷かれた。これらの影響を受け，揺れや通用といった表記のバリエーションを許容してきた言語的環境が崩れ，次第に規範表記への統一化が実現されてきたことの現れと見ることができる。

６．まとめ

本発表では，『太陽コーパス』の XML タグを直接用いて，同義異表記語の抽出を試み，以下の分析・考察を行なった。

- 同義異表記語の語・字構成要素による分類と，異表記発生・受容の背景の考察
- 出現度数による異表記通用状況の実態把握
- 通用表記率の経年変化と国語施策との関連性の指摘

今回の調査・分析は，コーパスの研究用付加情報を生かした研究手法のトライアルケースとして行なったものであり，明治期における漢語表記の多様性についてほんの一端を捉えたものにすぎない。今回の手法の問題点や限界，あるいは時間的制約によって課題として残されたものの幾つかを以下に示す。

- 著者による使用実態・個人差の把握

  表記問題は，個人に帰結する部分も大きく，著者ごとの使用実態把握が欠かせない。今回の調査では，著者別の集計を行なったものの，時間的・能力的制約によって分析に至らなかった。今後の課題としたい。
- 『太陽コーパス』の校訂態度（「A 誤字通用」の対象選定基準）に依拠したことにより抽出できなかった同義異表記語の把握（例：「障碍」「障害」「障礙」[8]）。
- 通用表記の議論として保留した，類義語・異義語の誤用・通用の実態把握（例：「異義」と「異議」[9]）。

  これらについては，別途，辞書や同音異表記のリストなどを用いて，調査すべき語を選定する必要がある。いずれも今後の課題としたい。

参考文献

今野真二(2012)『百年前の日本語――書きことばが揺れた時代 (岩波新書)』岩波書店

武部良明(1981)『日本語表記法の課題』三省堂

国立国語研究所(2005a)『太陽コーパス―雑誌『太陽』日本語データベース―』（CD-ROM）博文館新社

国立国語研究所(2005b)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社

笹原宏之(2005)「漢字文字列における字体の同化と衝突」国立国語研究所(2005b)，pp.293-312

田島優(1998)『近代漢字表記語の研究』和泉書院

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」国立国語研究所(2005b)，pp.1-48

文化庁[編](2006)『国語施策百年史』ぎょうせい

吉川明日香(2005)「字順の相反する２字漢語―「掠奪―奪掠」「現出―出現」について―」国立国語研究所(2005b)、pp.143-156

屋名池誠(2004)「明治語の表記」『日本語学』23-12，pp.65-72

[8] 『太陽コーパス』では校訂対象となっていない。

[9] 同音異議語とされる表記対だが，中納言による BCCWJ の検索結果では，「イギ申し立て」「イギあり/なし」等「異議」が期待される連接において「異義」が見られる（異義 40 例／異議 1574 例）。これらは語彙素がそれぞれ「異義」「異議」となっているが，同一語の異表記として扱うべきものと思われる。

会話コーパスの転記方式の相互変換
―引き伸ばしに着目して―

土屋 智行 (国立国語研究所言語資源研究系)
伝 康晴 (千葉大学文学部/国立国語研究所言語資源研究系)
小磯 花絵 (国立国語研究所理論・構造研究系)

Towards Automatic Transformation between Different Transcript Conventions: Aspect of Stretching

Tomoyuki Tsuchiya (Dept. Corpus Studies, NINJAL)
Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

1. はじめに

近年、大規模な書き言葉コーパスの発展が見られる一方、話し言葉コーパスは音声収録・転記という初期段階にかかる大きな負担が課題となっている。とくに大規模な会話コーパスは未着手であり、コーパス構築の進展が遅れているという状況にある。国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴・2011 年 11 月〜2014 年 10 月）は、既存の会話コーパスを共有化することで、この課題を解決することを目的として立ち上げられた。

既存のコーパスを共有するにあたって、転記方式の不統一や基本アノテーションの欠如といった問題がある。とくに転記テキストは、話し言葉研究の出発点であるにもかかわらず、研究グループごとに異なる転記方式が用いられており、複数のコーパスを集積する場合に問題となる。この問題の解決のためには，個々の会話データの転記テキストと言語・音響情報などから、転記形式を相互に変換することが必要となる。

伝ほか (2012) では、10 数種のコーパスで用いられている転記方式を調査し、全体が『日本語話し言葉コーパス』（CSJ）方式と会話分析（CA）方式に概ね大別できることを示した。土屋ほか (2012) では、イントネーションに注目し、CSJ 方式の韻律ラベルから CA 方式の音調マーカーへの自動変換を試みた。そこでは、CSJ 方式で記述された言語・音響情報から CA 方式の音調マーカーを予測するための多変量モデルを構築し、音調マーカーの予測に貢献する言語・音響特徴を分析した。土屋ほか (2013) では、人手付与や予測結果の音調マーカーが転記者間・データ間でどのように異なるのかを分析し、さまざまなゆれがあることがわかった。とくに、転記者ごと・データごとに音調マーカーの特定に貢献する言語・音響特徴が異なり、また予測精度も音調ごとに異なることがわかった。

本研究の目的は、CA 方式におけるイントネーション以外の音声的なマーキングについて、転記者間・データ間でのゆれと、各言語・音響特徴の貢献度を探り、CSJ 方式と CA 方式の相

互変換の射程を広げることである。具体的には、CSJ 方式と CA 方式の両方で転記されたデータにおいて、CA 方式における音の引き伸ばし記号であるコロン ‘:’ と、CSJ 方式における各単語の時間情報との対応関係を分析する。さらに、CSJ 方式の言語・音響情報から CA 方式のコロンへの変換にあたって貢献する言語・音響特徴を検討する。

2. 方法

2.1 談話資料

本研究で用いる会話コーパスは、土屋ほか (2012, 2013) と同じ千葉大学 3 人会話コーパス (Den and Enomoto 2007) の 2 会話（chiba0232 と chiba0432）、合計約 20 分である。本コーパスには、簡略版 CSJ 方式による転記テキストと、発話単位・形態論情報・韻律情報などの種々のアノテーションが与えられている。

2.2 転記・アノテーション

2.2.1 CSJ 方式

CSJ 方式では、各単語の開始・終了時間の情報が付与されている。各単語の継続時間長は、この時間情報から抽出される。CSJ 方式でも、単語末で音の引き伸ばしがある位置に引き伸ばし記号が使われているが、この付与の形式は CA 方式とは異なり、コロンの個数で継続時間の違いを示すといったことはしていない。また、この記号の付与自体、あまり精密ではない。

2.2.2 CA 方式

CA 方式の転記テキストの例を図 1 に示す。CA 方式では、音の引き伸ばしの程度によってコロンの個数を増減させる（たとえば 04, 05, 06 行目）。

CA 方式による転記は、Gail Jefferson の体系 (Jefferson 2004) に準拠して、会話分析の研究者 3 名（X 氏、Y 氏、Z 氏）によって行なわれた。X 氏は chiba0232 を、Y 氏は chiba0432 を転記し、Z 氏は chiba0232 と chiba0432 の両方を転記した。2013 年現在で、X 氏は約 7 年、Y 氏と Z 氏は約 6 年の会話分析経験を有する。以下に、X 氏、Y 氏、Z 氏それぞれの会話分析経験の概要を示す。

X 氏は、2003 年からカリフォルニア大学ロサンゼルス校で会話分析を学び、2006 年から本格的に会話分析による研究を行なっている。2010 年に日本に帰国した後も、各科研プロジェクトや研究会、データセッションへの参加を継続的に行なっている。また、会話分析以外に音

CA 方式

```
01 A: 俺，まず，行ってないからせ[いじん式.] モ- もう[帰ってへんか]ら.
02 C:                          [あそっか.]       [遠いもんな. ]
03 B: 行きゃいい[のに].
04 A:          [正  ]月に行ったかˆら:,
05 C: あ::.=
06 A: =てゆうか，.hh 成人式，行こうと思ったら:，(>それで<)英語休まないかんﾟからね?ﾟ
```

図 1　CA 方式の転記テキスト

声学（イントネーション）の授業を受けた経験がある。

Y 氏は、2006 年から 2007 年にわたり語用論や談話分析の教科書を通じて会話分析の概念に触れ始め、2007 年からデータセッションへの参加や、独自に収録したデータおよび CSJ の転記を始めている。2008 年からは、カリフォルニア大学サンタバーバラ校で会話分析の授業を受け、2009 年から十数時間程度のデータの収録および転記を行なっている。会話分析以外にも、談話分析の専門的知識を有し、Du Bois 流の記法を学んでいる。

Z 氏は、2004 年から 2007 年までカリフォルニア大学ロサンゼルス校で会話分析を学び、2007 年以降は日本国内のデータセッションや研究会に参加している。大学院博士課程在籍時より、主要な分析手法の 1 つとして会話分析を採用している。また、会話分析以外に、認知言語学と談話機能主義言語学の知識を有している。

2.3　分析単位

コロンが末尾に付与されている単語のうち、アクセント句末以外に付与されているものは、chiba0232 では X 氏が 9 例（コロン全体の 6.1%）、Z 氏が 1 例（0.8%）のみであった。また、chiba0432 では Y 氏 9 例（5.0%）、Z 氏 13 例（6.1%）のみがアクセント句末以外であった。そこで、本研究では、アクセント句末以外のコロンは対象としなかった。[*1]

さらに、アクセント句末の単語の平均モーラ長（= 継続時間長/モーラ数）の分布をコロンの個数（0、1、2 以上）ごとに求めたものを図 2 に示す。アクセント句末の単語の平均モーラ長はコロンの有無（0 とそれ以外）で異なるが、コロンが付与されているアクセント句末の単

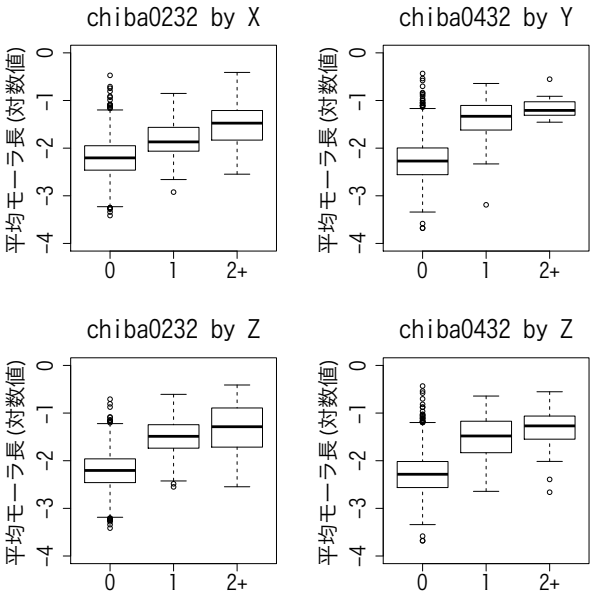


図 2　コロンの個数ごとのアクセント句末単語の平均モーラ長
（chiba0232：N = 987、chiba0432：N = 1241）

[*1] CA 方式では「え::と」のように語末以外にコロンが付与されることもあるが、これらも本研究では対象としない。

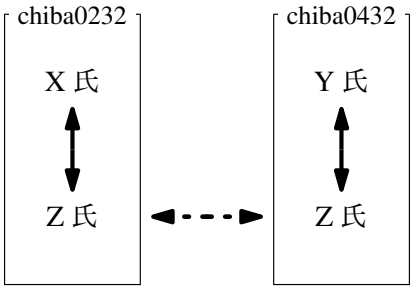


図 3　データ間・個人間比較

語では、コロンの個数が 1 つの場合と 2 つ以上の場合で平均モーラ長に明確な違いは見られなかった。そこで、コロンの個数は対象外とし、コロンの有無のみを対象とした。

2.4　言語・音響特徴

CSJ 方式の言語・音響情報から CA 方式の音の引き伸ばし記号を予測する多変量モデルを構築するために、分析対象アクセント句から以下の言語・音響特徴を抽出し用いた

■言語特徴

Break Index（BI）　アクセント句末の韻律境界の強さ。2, 2+, 3, D, F。中間値としての 2+b, 2+p, 2+bp を 2+ としてまとめた。

句末境界音調（tone）　アクセント句末の句末境界音調。L%, H%, HL%, LH%, D, F。

末尾単語の品詞（lastPOS）　アクセント句末尾の単語の品詞。品詞は以下の 7 種に分類した。体言・用言・助動詞・終助詞・接続助詞・その他の助詞・その他の品詞。

次末単語の品詞（penultPOS）　アクセント句の最後から 2 番目（次末）の単語の品詞

発話中の位置（loc）　当該アクセント句の発話の先頭からの距離（アクセント句数）

発話末からの位置（revLoc）　当該アクセント句の発話の末尾からの距離（アクセント句数）

■音響特徴

アクセント句の最大 F0（f0MaxAP）　アクセント句中の F0 の最大値（標準化得点）

句末単語の最小 F0（f0MinWord）　末尾単語中の F0 の最小値（標準化得点）

句末単語の最大 F0（f0MaxWord）　末尾単語中の F0 の最大値（標準化得点）

句末単語の最大パワー（pwrMaxWord）　末尾単語中のパワーの最大値（標準化得点）

句末単語の平均モーラ長（amdWord）　末尾単語の継続時間長をモーラ数で除したもの（標準化得点）

最右 F0 抽出点の値（lastF0Val）　アクセント句中で最後に抽出できた F0 点の値（標準化得点）

最右 F0 抽出点の位置（lastF0Loc）　最右 F0 抽出点の句末から計った時間（対数値）

末尾 F0 上昇幅（lastF0Rise）　最右 F0 抽出点の値（lastF0Val）から末尾単語の最小 F0 の値（f0MinWord）を引いたもの。句末における F0 の上昇幅にほぼ対応

F0 と平均モーラ長は対数変換後、パワーはそのままで、話者ごとに標準化得点に変換した。

2.5　分析手順

音声の引き伸ばしに関する、データ間・個人間での転記方略のゆれを図 3 のように比較した。まず、データ内・個人間の比較として、chiba0232 では X 氏と Z 氏による転記中でのコロンの有無を、chiba0432 では Y 氏と Z 氏による転記中でのコロンの有無をアクセント句単位で整列し比較した（実線矢印）。

次に、個人内・データ間の比較として、Z 氏によって転記された chiba0232 と chiba0432 のコロンの有無を比較した（破線矢印）。この比較のために、CSJ 方式の言語・音響情報から CA 方式の引き伸ばし記号へ変換する多変量モデルを一方のデータを学習データとして構築し、他方のデータでコロンの有無を予測し、その結果と人手によるコロンの有無とを比較した。多変量モデルとしてランダムフォレスト法 (Breiman 2001) を用い、統計解析ソフト R 言語の `randomForest` パッケージを使ってモデルを構築した。アクセント句単位でデータを分節化し、2.4 で述べた言語・音響特徴を説明変数に用いた。

さらに、各データ・転記者の転記方略を検討するため、4 データのそれぞれに対してランダムフォレスト法による多変量モデルを構築し、OOB 推測に基づく各特徴の貢献度を推定した。

3.　結果

3.1　データ内・個人間のゆれ

同一データ内における異なる転記者同士のコロンの有無の対応を表 1 に示す。一致の程度は比較的高く、chiba0232 では一致率 89.0%（$\kappa = .56$）、chiba0432 では一致率 92.6%（$\kappa = .72$）であった。しかし、コロンが付与されているアクセント句のみに注目すると、chiba0232 で X 氏が 128 箇所、Z 氏が 118 箇所、chiba0432 で Y 氏が 171 箇所、Z 氏が 199 箇所コロンを付与した箇所のうち、両者で一致しているのは、chiba0232 では 6 割程度、chiba0432 では 7〜8 割程度であった。とくに、X 氏と Z 氏の間で不一致が多い。

表 1　データ内・個人間のゆれ

chiba0232
（一致率 = 89.0%、$\kappa = .56$）

	Z 氏		
X 氏	なし	あり	合計
なし	**670**	41	711
あり	51	**77**	128
合計	721	118	839

chiba0432
（一致率 = 92.6%、$\kappa = .72$）

	Z 氏		
Y 氏	なし	あり	合計
なし	**954**	58	1012
あり	30	**141**	171
合計	984	199	1183

3.2　個人内・データ間のゆれ

同一の転記者における異なるデータ間でのコロンの有無の対応を、多変量モデルによる予測結果と人手ラベルとの対応によって調べた。chiba0432/chiba0232 を学習データとして構築した多変量モデルによる chiba0232/chiba0432 の予測結果と人手ラベルとの対応を表 1 に示す。正解率はそれぞれ、92.1%, 90.1% とかなり高かった。しかし、コロンが付与されるアクセント

表 2　個人内・データ間のゆれ（転記者 = Z 氏）

学習 = chiba0432、予測 = chiba0232
（正解率 = 92.1%、κ = .65）

	観測値		
予測値	なし	あり	合計
なし	**699**	44	743
あり	22	**74**	96
合計	721	118	839

学習 = chiba0232、予測 = chiba0432
（正解率 = 90.1%、κ = .57）

	観測値		
予測値	なし	あり	合計
なし	**968**	101	1069
あり	16	**98**	114
合計	984	199	1183

句のみに注目すると、適合率（「あり」と予測して正解したものの割合）は 77.1%, 86.0% と比較的高いものの、再現率（実際に「あり」であるものを言い当てた割合）は 62.7%, 49.2% とかなり低かった。

3.3　言語・音響特徴の貢献度

次に、4 つのデータそれぞれから多変量モデルを構築し、コロンの予測に貢献する言語・音響特徴を調べた。結果を図 4 に示す。予測に貢献している上位 6 位の言語・音響特徴をみると、amdWord, tone, lastF0Val, f0MinWord, f0MaxWord が全て共通しており、そのうち amdWord と tone は全て上位の 2 位を占めていた。とくに、amdWord の貢献度がいずれのデータにおいても高かった。また、chiba0432 では BI も共通して貢献度が高かった。これらのうち、言語特徴は tone と BI のみで、ほかは全て音響特徴であった。

データ内・個人間で比較すると、chiba0232 では、X 氏で revLoc の貢献度が比較的高いのに対して Z 氏では低く、一方、Z 氏で pwrMaxWord の貢献度が高いのに対して X 氏で低かった。chiba0432 では、貢献度の高い言語・音響特徴が Y 氏と Z 氏の間で完全に一致していた。

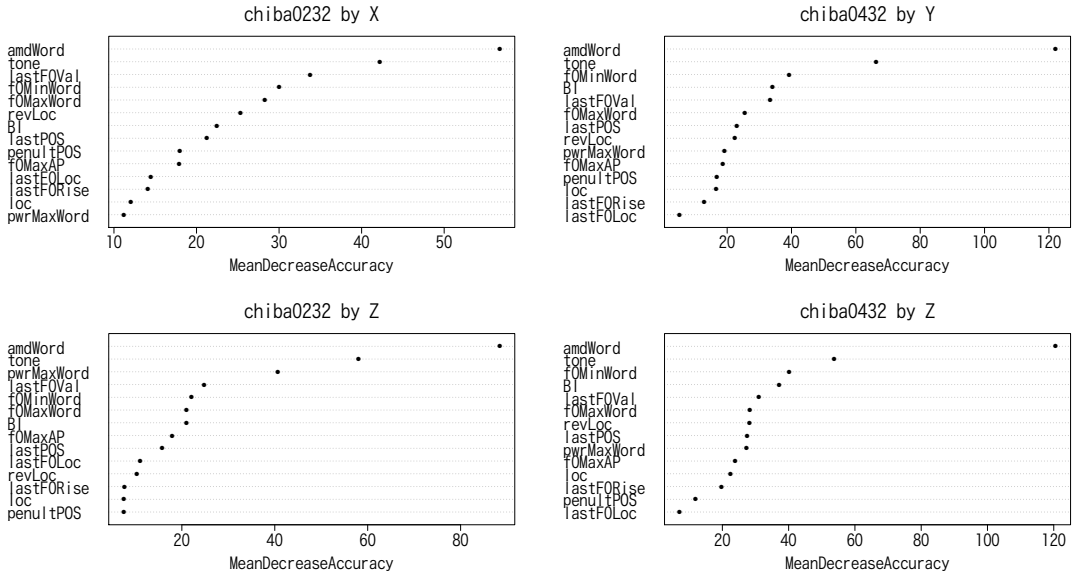


図 4　言語・音響特徴の貢献度

個人内・データ間で比較すると、Z 氏によるコロン付与の予測に貢献している上位 6 位の言語・音響特徴のうち、amdWord, tone, lastF0Val, f0MinWord, f0MaxWord の 5 つはデータ間で共通していた。また、chiba0432 で貢献度の高い BI は、chiba0232 でも比較的貢献度が高かった（7 位）。

4. 議論

本研究では、CSJ 方式の言語・音響情報から CA 方式の音の引き伸ばし記号コロンへの自動変換について検討した。コロンが付与されている箇所の大半は、アクセント句末に位置する単語の末尾であった。CA 方式では引き伸ばしの程度がコロンの個数によって示されているが、コロンの個数によって単語の平均モーラ長に明確な違いはなかった。そこで、アクセント句末のコロンの有無のみを検討の対象とした。

同一データ内における異なる転記者同士のコロンの有無の対応を調べたところ、全体としては比較的高い率で一致していたが、コロンが付与されている箇所のみに注目すると、一致している箇所は 6〜8 割程度であった。とくに、X 氏と Z 氏の間で不一致が多かった。土屋ほか (2013) では、イントネーションの転記に関して、転記者間でどのように異なるか検討したが、そこでは X 氏と Z 氏よりもむしろ Y 氏と Z 氏のほうが不一致が多かった。このことから、音声的なマーキングの種類によって、転記者間の一致の程度は異なることがわかった。

次に、同一の転記者における異なるデータ間でのコロンの有無の対応を、多変量モデルによる予測結果と人手ラベルとの対応によって調べたところ、全体として正解率はかなり高かったが、コロンが付与される箇所のみに注目すると、適合率は 77%〜86% と比較的高いものの再現率は 49%〜63% とかなり低かった。音調マーカーで同様の対応を調べた土屋ほか (2013) でも、音調によっては 24%〜39% とかなり再現率の低いもの（上昇調）があり、現状の言語・音響特徴では自動変換の精度に限界があるといえる。

多変量モデルによるコロンの有無の予測に貢献する言語・音響特徴をデータ間・個人間で比較したところ、貢献の高い特徴はほとんど一致していた。土屋ほか (2013) の分析では、イントネーションの転記方略は転記者間で異なり、品詞などの言語特徴を重視するかアクセント句末の F0 値などの音響特徴を重視するかで大きく分かれたが、音の引き伸ばしに関してはこのような方略の違いはほとんど見られないことがわかる。しかし、データ間での多少の違いも見られ、Z 氏による chiba0432 の転記では chiba0232 よりも Break Index がかなり重視されていた。この点に関しては、土屋ほか (2013) と同様に、同一の転記者であっても、発話者の発話に応じて転記方略を変えていることが考えられる。

貢献度の高い言語・音響特徴を詳細にみると、とりわけ貢献度の高いアクセント句末の単語の平均モーラ長は音の引き伸ばしの有無を直接的に反映する特徴と言えるが、ついで貢献度が高かったのは句末境界音調であった。Z 氏の転記データをみると、句末境界音調が下降調（L%）や上昇調（H%）のときは 90%〜95% はコロンが付いていないのに対して、上昇下降調（LH%）のときは逆に 86% でコロンが付与されていた。また、下降調や上昇調ではコロンの有無によって平均モーラ長の差が大きいのに対して、上昇下降調ではこの差が小さく、コロンがない場合でも平均モーラ長が比較的長かった。このため、平均モーラ長に加えて、句末境界音

調がコロンの有無の予測に貢献しているものと思われる。ただし、平均モーラ長の長い上昇下降調でもコロンが付与されていない箇所もあり、転記者がどのような方略でその選択を行なっているのかは今後より詳しく調べる必要がある。

以上のように、音の引き伸ばしに関する CSJ 方式から CA 方式への転記変換において、関与している言語・音響特徴はデータ間・個人間でかなり共通するものの、引き伸ばしマーカーを付与する箇所にはかなりの不一致がみられ、自動変換の精度も十分ではなかった。今後は、使用する特徴を工夫し、精度を上げる必要がある。

謝辞　会話分析方式の転記を作成していただいた遠藤智子・黒嶋智美・横森大輔の各氏に感謝します。本研究は国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴）による成果である。

参考文献

Breiman, Leo (2001). “Random forests.” *Machine Learning*, 45, pp. 5–32.

Den, Yasuharu, and Mika Enomoto (2007). “A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation.” Toyoaki Nishida (Ed.), *Conversational informatics: An engineering approach*. Hoboken, NJ: John Wiley & Sons. pp. 307–330.

伝康晴・土屋智行・小磯花絵 (2012).「多様な様式を網羅した会話コーパスの共有化」　『第 1 回コーパス日本語学ワークショップ予稿集』　pp. 227–234.

Jefferson, Gail (2004). “Glossary of transcript symbols with an introduction.” Gene Lerner (Ed.), *Conversation analysis: Studies from the first generation*. Amsterdam/Philadelphia: John Benjamins. pp. 13–31.

土屋智行・伝康晴・小磯花絵 (2012).「会話コーパスの転記方式の相互変換に向けて—イントネーションに着目して—」　『第 2 回コーパス日本語学ワークショップ予稿集』　pp. 117–126.

土屋智行・伝康晴・小磯花絵 (2013).「会話分析方式への転記変換におけるデータ間・個人間のゆれに関する分析」　『第 3 回コーパス日本語学ワークショップ予稿集』　pp. 417–424.

関連 **URL**

「会話コーパス」ホームページ：`http://www.jdri.org/kaiwa/`

弱境界における発話計画に関わる音声的・言語的特徴の分析

小磯 花絵（国立国語研究所理論・構造研究系）†
伝 康晴（千葉大学文学部/国立国語研究所言語資源研究系）

An Analysis of Acoustic and Linguistic Features Related to Speech Planning at Weak Clause Boundaries in Japanese

Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)
Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)

1. はじめに

自発性の高い話し言葉では，漸進的に発話内容や言語表現を計画しながら話を進める必要がある。特に，発話の冒頭や，主節からの独立性が高く切れ目の度合の強い節境界の後では，この種の発話計画に関わる認知的負荷が相対的に高いことから，フィラーや語の繰り返し，母音の引き延ばしなどの非流暢性が多く出現することが報告されている (伝 2007, Den 2009, Watanabe 2009)。

たとえば Watanabe (2009) は，切れ目の度合の強い節境界と弱い節境界に着目し，境界の直後に出現するフィラーの比率を比較したところ，弱い節境界よりも強い節境界の後の方がフィラーがより多く出現する傾向にあることを指摘している。強い節境界では，大きなまとまりの発話が終了するため，発話計画に関わる認知的負荷も相対的に高く，直後のフィラーの生起に影響したと考えられる。また Watanabe (2009) は，強い節境界と弱い節境界を対象に，後続する節の長さ（語数）と節境界のフィラー率との関係を調べ，弱い節境界では後続節が長いほどフィラー率が高くなる傾向が見られること，強い節境界ではこのような相関は見られないことを指摘している。更に渡辺・清水 (2012) は，たとえば「**漱石の**小説」と「**漱石の**新聞に連載された小説」では直後の文節に係る前者より三つ先の文節に係る後者の方が言語的に複雑で言語化の負荷が高いと考え，当該文節直後のフィラー率とその文節が係る先の文節までの距離（文節数）との関係を調べた。その結果，係り先の距離が遠いほどフィラー率が高くなる傾向が見られることを明らかにした。

一方，アクセント句末で生じる上昇調や上昇下降調といった句末境界音調（Boundary Pitch Movement, BPM）についても，同じような傾向が観察されることが報告されている。小磯 (2012) は，強い節境界，弱い節境界，節境界以外のアクセント句末（以下，非節境界）を対象に，その境界に生じる BPM の比率を調査したところ，「非節境界 < 弱い節境界 < 強い節境界」の順に，上昇調，上昇下降調ともに出現率が高くなる傾向が見られることを報告している。また，弱い節境界における係り先の距離 *1 や非節境界における係り先の距離 *2 が遠いほど，上

† koiso@ninjal.ac.jp
*1 Watanabe (2009) で検討した弱い節境界における後続節の長さにほぼ相当。
*2 渡辺・清水 (2012) における係り先の距離に関する調査のうち非節境界部分に相当。

昇調，上昇下降調ともに出現率が高くなる傾向が見られることも指摘している。

このように，BPM とフィラーはともに，発話計画による認知的負荷が相対的に高い位置により多く生じるという意味において共通している。しかし，BPM はそもそも発話計画による認知的負荷に関わる母音の引き延ばしを伴うことも多く，BPM の生起が発話計画に直接関係するものではない可能性もある。

そこで Koiso and Den (2013) では，弱い節境界を対象に，発話計画による認知的負荷に関わりうる要因として，境界直後のフィラーの生起の有無，境界直前の BPM の生起の有無，境界直前の最終モーラ長の三つに着目し，これらが弱い節境界後に発話される要素の長さに関わるかを検討した。その結果，BPM を含む全ての特徴が統計的に有意に発話要素長に関わることを明らかにした。しかし，この分析で対象とした弱い節境界には，種類の異なるものが多く含まれており十分な統制がとれていないなどの問題があった。また，弱い節境界後の発話要素長には，節の種類など認知的負荷に関わりうる要素以外の要因も影響する可能性がある。そこで本研究では，節の種類を統制した上で，節の種類や談話の種類なども要因に加えた分析を行い，フィラー，BPM，最終モーラ長の後続発話要素長への効果について改めて詳細に検討する。

2. 方法

2.1 データ

分析には『日本語話し言葉コーパス』（CSJ）を用いた。CSJ は自発性の高いモノローグを中心に構成された話し言葉コーパスであり，学会における口頭発表（以下「学会講演」）と，一般話者による主に個人的な内容に関するスピーチ（以下「模擬講演」）を主対象としている。CSJ 全体は 661 時間の音声から構成されるが，本研究ではこのうち「コア」と呼ばれるデータ範囲の中から学会講演 70（約 29 時間）・模擬講演 107（約 20 時間）を分析対象とした。実際の分析には CSJ 第 3 刷に基づき作成された CSJ-RDB (小磯ほか 2012) を用いた。

2.2 アノテーション

研究には，節単位情報，形態論情報，係り受け構造情報，韻律情報など，CSJ コアに付与されている人手修正を経た精度の高いアノテーションを利用した。

節単位情報は原則「節（clause）」の境界によって得られる文法的・意味的なまとまりを持った単位である (丸山ほか 2006)。節の境界は，構造的な切れ目の大きさの観点から以下の 3 つに分類されている。

絶対境界：いわゆる文末に相当する境界

強境界：　後続の節に対する従属度の低い，切れ目の度合が強い節境界

弱境界：　後続の節に対する従属度の高い，切れ目の度合が弱い節境界

これらの節の境界とその種類は，形態素解析結果に基づき CBAP–csj プログラムにより自動で判別され，人手による修正操作を経た上で，絶対境界か強境界のいずれかで区切られる単位が「節単位」と認定された。本研究で着目するのは，人手修正後に認定される「節単位」の内部に存在する弱境界である。節単位は文節に分割されたのち，文節を単位とした係り受け構造情報が節単位を範囲に付与されている。

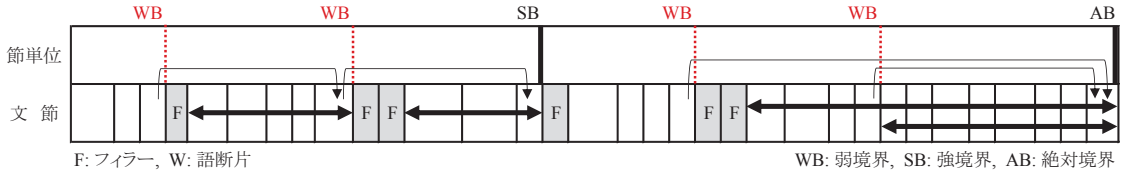


図 2　弱境界後の発話要素長。円弧は弱境界の節とそれが連用修飾として係る先の節との関係を示す。発話要素長は，弱境界直後の文節（フィラーと語断片を除く）の始端から，弱境界の節が連用修飾として係る先の節の終端までの継続時間として計算した。

上記のような形態統語論的情報に加え，CSJ コアには X–JToBI に基づく韻律情報も付与されている (五十嵐ほか 2006)。本研究では，これら韻律情報のうちアクセント句末の境界音調の情報を利用した。CSJ コアが対象とする東京方言では，アクセント句の F0 はゆるやかに下降し（L%），そのあと，単純に F0 が上昇する上昇調（H%）や上昇のあとに下降が生じる上昇下降調（HL%），上昇前に低い F0 区間が持続される上昇調（LH%），上昇下降調のあと更に上昇が生じる上昇下降上昇調（HLH%）などの句末境界音調（Boundary Pitch Movement, BPM）が後続することがある（図 1 参照）。

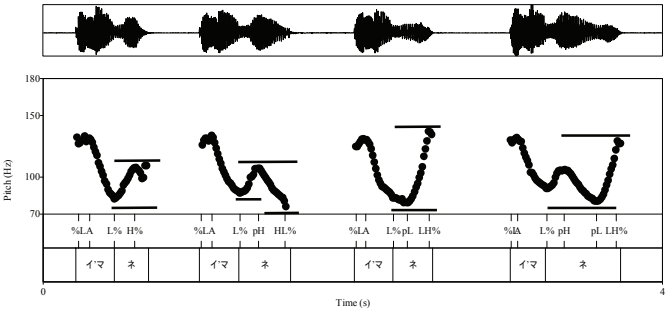


図 1　BPM の例。左から順に，単純に F0 が上昇する上昇調（H%），上昇のあとに下降が生じる上昇下降調（HL%），上昇前に低い F0 区間が持続される上昇調（LH%），上昇下降調のあと更に上昇が生じる上昇下降上昇調（HLH%）。

2.3　分析パラメータ

本研究では，弱境界付近の発話計画による認知的負荷に関わりうる要素などを用いて，弱境界の後に発話される要素の長さを予測するモデルを作成する。そのために，以下の手順で目的変数と説明変数を抽出した。

目的変数は，弱境界の後に発話される要素の長さ（以下「発話要素長」）である。発話要素長として，弱境界直後の文節（フィラーと語断片を除く）の始端から，弱境界の節が係る先の節の終端までの継続時間長を用いた（図 2 参照）。Koiso and Den (2013) とは異なり本分析では，当該の弱境界の節が連用修飾として係る場合に限定し，引用節やトイウ節などは対象外とした。また，「〜に関して」「〜に対して」のように接続助詞「テ」が複合辞の場合や，「〜ては」「〜ても」「〜てから」のように「テ」に助詞が後続する場合，主題の提示に関わるなど連用修飾とは異なる振舞いをすることがあるため，一律分析対象外とした。その他，節境界の種類のラベルに誤りがあるものを除いた。

説明変数としては，発話計画による認知的負荷に関わりうる要素として，次の三つの変数に着目した。

1. 弱境界直後のフィラーの生起の有無
2. 弱境界を終端境界とする節の末尾の BPM の生起の有無
3. 弱境界を終端境界とする節の末尾のモーラの継続時間長

フィラーの有無は弱境界直後の要素，BPM の有無と最終モーラ長は弱境界直前の要素である。また，節や談話の種類の違いも弱境界後の発話要素長の変動に影響する可能性があると考え，次の二つも説明変数に加えた。

4. 節タイプ。節単位情報に含まれる節境界ラベルを用いて，条件節，理由節，並列節，連用節，テ節に分類。
5. 談話タイプ。学会講演，模擬講演の別。

なお，目的変数の発話要素長は対数変換して分析に用いた。説明変数の最終モーラ長については，対数変換した上で話者ごとに正規化して分析に用いた。

2.4　分析対象

上述の通り，本分析では弱境界の節が連用修飾として係る場合に限定した。また，接続助詞「テ」が複合辞の場合や「テ」に助詞が後続する場合，節境界の種類のラベルに誤りがある場合は対象外とした。更に，説明変数としてアクセント句末の BPM の有無を用いるため，弱境界がアクセント句末に一致する場合に限定した。その結果，全 19,184 の弱境界のうち 8,858（条件節：1666，理由節：1390，並列節：1046，連用節：760，テ節：3996）を分析に用いた。

3.　結果

弱境界後の発話要素長と各説明変数との関係を図 3・4 に示す。いずれの節タイプ，談話タイプについても，フィラーや BPM が存在する場合に発話要素長がより長くなり，最終モーラ長が長くなるほど発話要素長がより長くなる傾向が見られる。また，全般的に，節タイプによって発話要素長の差も見られる。一方，談話タイプによる差は見られない。

上記五つの説明変数の発話要素長に対する効果の最適な組合せを検討するため，話者によるクラスターを考慮した線形混合効果モデルを用いた。ランダム効果としてはランダム切片のみ考慮した。分析には R の `lmer` 関数を用いた (Baayen 2008)。まず，フィラーの有無，BPM の有無，最終モーラ長と節タイプの交互作用を検討した。フィラーの有無，BPM の有無，最終モーラ長のそれぞれと節タイプとの交互作用を入れたモデルと，交互作用のないモデルを作成し，尤度比検定を行った結果，いずれの要因についても交互作用は有意ではなかった。次に，この交互作用のないモデルと，談話タイプを加えたモデルを比較し，尤度比検定を行った結果，談話タイプの効果は有意でなかった。

以上の結果から，フィラーの有無，BPM の有無，最終モーラ長，節タイプの主効果のみを含むモデルを採用し，このモデルで発話要素長を予測する階層ベイズモデルを構築した。

$$
\begin{aligned}
y_{ij} &\sim N(\mu_{ij}, \sigma) \\
\mu_{ij} &= \beta_{0j} + \beta_F x_{Fij} + \beta_B x_{Bij} + \beta_D x_{Dij} + \beta_C x_{Cij} \\
\beta_{0j} &\sim N(\mu_0, \sigma_s)
\end{aligned}
$$

j 番目の話者の i 番目のデータの発話要素長 y_{ij} を予測するのに，フィラーの有無 x_{Fij}，BPM の

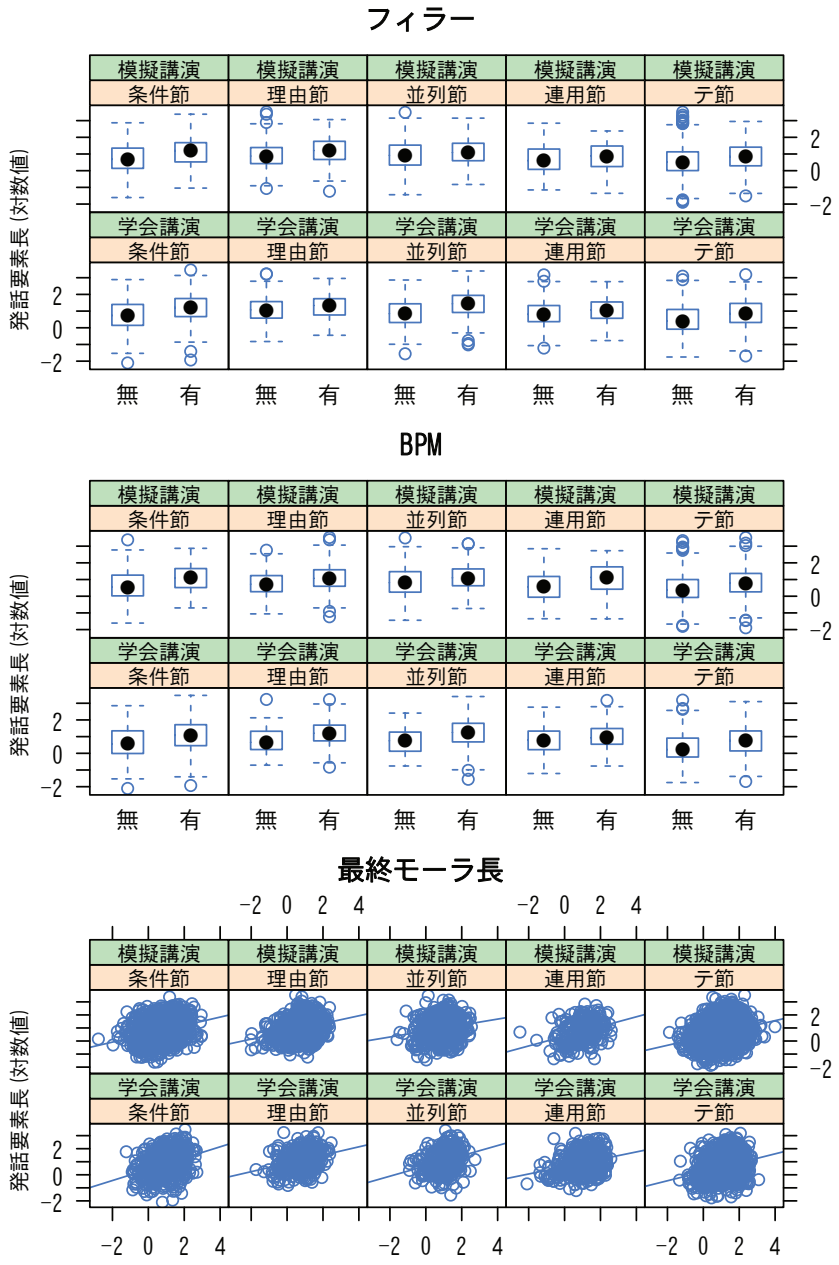


図 3　発話要素長とフィラー・BPM・最終モーラ長との関係

有無 x_{Bij}，最終モーラ長 x_{Dij}，節タイプ x_{Cij} の線形結合を用い，切片 β_{0j} は話者ごとに平均 μ_0 の周りで変動するとした。各説明変数の係数 $\beta_F, \beta_B, \beta_D, \beta_C$ と切片の話者平均 μ_0 の事前分布として正規分布（平均：0，分散：10^{12}）を，残差の分散 σ^2 と切片の話者分散 σ_s^2 の事前分布として逆ガンマ分布（形状：.01、尺度：.01）を用いた。これらのパラメータを MCMC（Markov chain Monte Carlo）法により事後分布からのサンプリングを行い推定した。推定には JAGS と R の `rjags` パッケージを用いた (Kruschke 2011)。

各パラメータの推定結果（事後分布）を図 5 に示す。節タイプについては，各水準間の差の

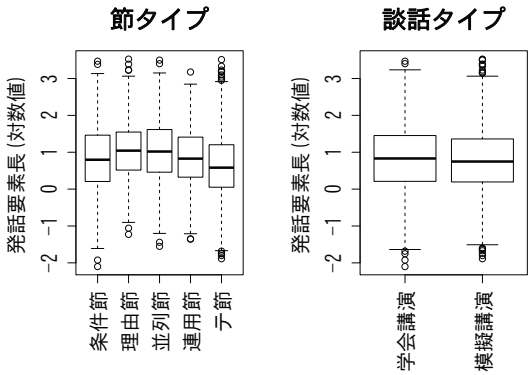


図 4　発話要素長と節タイプ・談話タイプとの関係

事後分布を図 6 に示す。フィラーの有無，BPM の有無，最終モーラ長の係数の事後分布は分布の 95% 範囲（95%HDI）に 0 を含んでおらず，5% 水準で有意であることが分かる。それぞれの平均値から，フィラーと BPM は存在する場合の方が，また最終モーラ長は長い方が，発話要素長が有意に長い。節タイプについては，条件節と連用節，理由節と並列節の差を除き，いずれも分布の 95% 範囲に 0 を含んでおらず，節タイプごとにその後の発話要素長に差があると言える。図 6 の結果を整理すると，「テ節 < 条件節・連体節 < 理由節・並列節」の順に発話要素長が長くなる傾向が見られる。

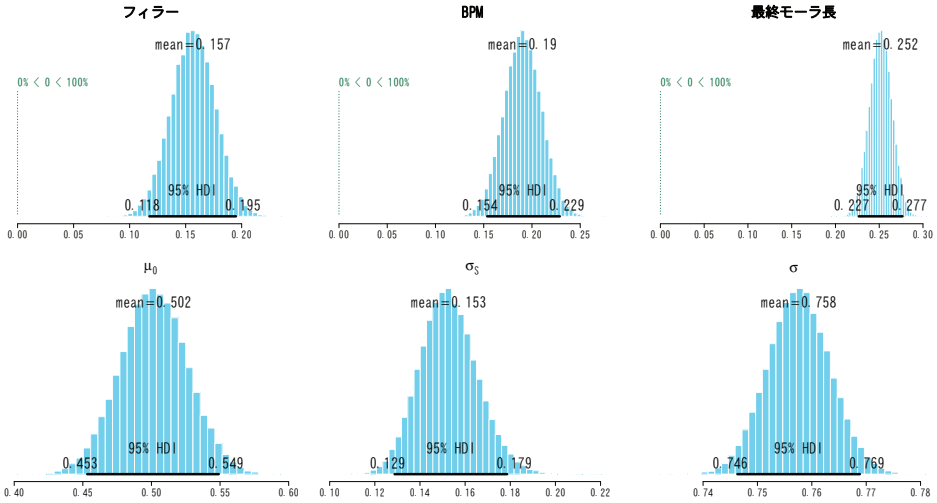


図 5　MCMC に基づく各パラメータの推定値。上段は，説明変数であるフィラーの有無，BPM の有無，最終モーラ長の係数の推定値。下段の μ_0，σ_s，σ は，切片の話者平均，切片の話者標準偏差，残差の標準偏差の推定値。95%HDI は分布の 95% の範囲。この範囲内に 0 を含まなければ 5% 水準で有意であることを意味する。

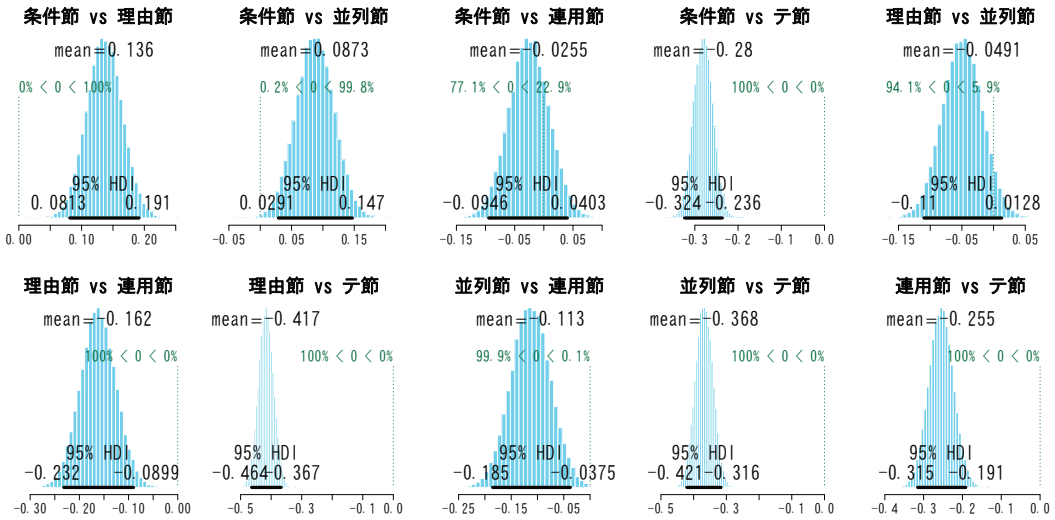


図 6　節タイプの係数に関する推定値。各水準間の差をプロット。たとえば「条件節 vs 理由節」の場合，理由節の係数の推定値から条件節の係数の推定値を引いた値をプロットしている。正の方向に分布し，95% HDI に 0 を含まないため，理由節の方が条件節よりも係数の値が有意に大きいことを意味する。

4.　考察

分析の結果，フィラーの有無，BPM の有無，最終モーラ長，および節タイプのいずれに関しても，弱境界後の発話要素長に有意に影響を及ぼすことが分かった。具体的には，境界直後にフィラーが存在したり，境界直前に BPM が存在したりすると，発話要素長が長くなる傾向が，また境界直前の最終モーラ長が長くなるほど発話要素長が長くなる傾向が見られた [*3]

フィラーや語の繰り返しなどは，発話計画による認知的負荷が高い発話冒頭に頻出することが指摘されている (Clark 2002, Clark and Wasow 1998)。また日本語の話し言葉では，接続助詞や用言連用形などにより複数の節を連結して発話を構成することも多いため，節冒頭にも着目して研究が進められてきた (Watanabe 2009, 渡辺・清水 2012)。発話や節の冒頭では，発話の内容や表現の計画に必要となる時間をかせぐためにこれらの要素が生じやすいと考えることができる。本研究で対象としたフィラーは節冒頭の要素であり，これら先行研究の結果が改めて確認されたことになる。一方，BPM や最終モーラ長は，節冒頭よりもわずかに早い，先行する節の末尾の要素である。節末尾の要素が後続の発話要素長に影響するという今回の結果は，節を完全に終える前に次の発話計画が進められていることの傍証となる。また，日本語話者にとっては，直前の発話や節が終了する前から，発話計画のための時間をかせぐことができる便利なツールを手にしていると言える。

ここで BPM について改めて考えてみたい。1 節で述べたように，BPM はそもそも発話計画による認知的負荷に関わる母音の引き延ばしを伴うことが多く，BPM の存在自体は弱境界後

[*3] これはあくまで統計的関係であり因果関係を示すものではない。後述の通り我々は，次に発話する要素がより長いほど，その内容や表現の計画に必要となる時間をかせぐためにこの種の要素がより生じやすいと考えている。

の発話要素長に直接関係しない可能性も十分に考えられた。しかし分析の結果，最終モーラ長に加えて BPM も発話要素長に影響することが明らかになった。問題はその解釈である。BPM の存在自体が「時間かせぎ」に貢献すると考えにくい。

小磯 (2012) は，「**漱石の**小説」のように直後の文節に係る場合よりも，「**漱石の**新聞に連載された小説」のように二つ以上先の文節に係る方がより多く BPM が生じることを明らかにした。このような統語的構造の違いが BPM の出現に関わるとして，このことを本研究で対象とする弱境界に当てはめるならば，「**このイベントはとても面白そうなので〈理由節ノデ〉**友達と一緒に行ってきます」のように直後の節に係る場合よりも，「**このイベントはとても面白そうなので〈理由節ノデ〉**お金さえあれば〈条件節レバ〉時間を作って〈テ節〉友達と一緒に行ってきます」のように二つ以上先の節に係る方が，BPM が生じる可能性は高くなる。係り先が遠いほど後続要素長は長くなるため，結果として，BPM と発話要素長との間に関係が見られたと考えることができる。この可能性を検証するには，弱境界後の係り先までの節の数と BPM の出現との関係を検討する必要がある。この点については今後の課題とする。

参考文献

Baayen, R. Harald (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Clark, Herbert H. (2002). “Speaking in time.” *Speech Communication*, 36, pp. 5–13.

Clark, Herbert H., and Thomas Wasow (1998). “Repeating words in spontaneous speech.” *Cognitive Psychology*, 37, pp. 201–242.

伝康晴 (2007).「発話冒頭付近での語句の繰り返しの機能」　串田秀也・定延利之・伝康晴 (編)『文と発話 3: 時間の中の文と発話』東京: ひつじ書房　pp. 103–133.

Den, Yasuharu (2009). “Prolongation of clause-initial mono-word phrases in Japanese.” Shu-Chuan Tseng (Ed.), *Linguistic patterns in spontaneous speech*. Taipei: Institute of Linguistics, Academia Sinica. pp. 167–192.

五十嵐陽介・菊池英明・前川喜久雄 (2006).「韻律情報」　『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』　pp. 347–453.

小磯花絵 (2012).「日本語話し言葉コーパスを用いた複合境界音調の発言継続表示機能の検討」　『第 2 回コーパス日本語学ワークショップ予稿集』　pp. 221–230.

Koiso, Hanae, and Yasuharu Den (2013). “Acoustic and linguistic features related to speech planning appearing at weak clause boundaries in Japanese monologs.” *Proceedings of the 6th Workshop on Disfluency in spontaneous speech*. Stockholm.

小磯花絵・伝康晴・前川喜久雄 (2012).「『日本語話し言葉コーパス』RDB の構築」　『第 1 回コーパス日本語学ワークショップ予稿集』　pp. 355–364.

Kruschke, John K. (2011). *Doing bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

丸山岳彦・高梨克也・内元清貴 (2006).「節単位情報」　『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』　pp. 255–322.

Watanabe, Michiko (2009). *Features and roles of filled pauses in speech communication: A corpus-based study of spontaneous speech*. Tokyo: Hituzi Syobo.

渡辺美知子・清水信哉 (2012).「『日本語話し言葉コーパス』における文節境界のフィラーの出現率」　『第 1 回コーパス日本語学ワークショップ予稿集』　pp. 259–264.

会話コーパスの転記方式の相互変換
—言語・音響特徴を用いた会話分析方式の音調マーカーの導出—

石本 祐一（国立国語研究所言語資源研究系）†
土屋 智行（国立国語研究所言語資源研究系）
小磯 花絵（国立国語研究所理論・構造研究系）
伝 康晴（千葉大学文学部/国立国語研究所言語資源研究系）

Towards Automatic Transformation between Different Transcript Conventions: Prediction of Intonation Markers from Linguistic and Acoustic Features

Yuichi Ishimoto (Dept. Corpus Studies, NINJAL)
Tomoyuki Tsuchiya (Dept. Corpus Studies, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)
Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)

1. はじめに

話し言葉コーパスでは音声収録・転記という初期段階にかかる大きな負担が課題となっている。特に大規模な会話コーパスの整備は未着手であり，会話研究の遅れの要因となっている。国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴・2011 年 11 月～2014 年 10 月）は，既存の会話コーパスを共有化することでこの課題を解決することを目的として立ち上げられた。

既存のコーパスを共有する上での問題のひとつとして，転記方式の不統一や基本アノテーションの欠如が挙げられる。土屋ほか (2012, 2013) では，『日本語し言葉コーパス』（CSJ）方式の言語・音響情報から会話分析（CA）方式の音調マーカーを予測するための多変量モデルを構築し，CSJ 方式から CA 方式の音調マーカーへの自動変換を試みた。その結果，転記者ごと・データごとに音調マーカーの特定に貢献する言語・音響特徴が異なり，また予測精度も音調ごとに異なることがわかった。

本研究では，先行研究で構築された多変量モデルにおいて音調マーカーを正しく予測できなかった事例に対して，誤変換の音響的な要因を探る。具体的には，先行研究で特に予測誤りの多かった上昇の音調マーカーであるクエスチョンに着目し，正しく予測された事例の音響特徴と誤って予測された事例の音響特徴の比較を行う。さらに，CSJ 方式から CA 方式の音調マーカーへより精度の高い変換を行うために必要な音響特徴を検討する。

† yishi@ninjal.ac.jp

2. データ

2.1 談話資料

本研究で用いる会話コーパスは，先行研究 (土屋ほか 2012, 2013) と同じ千葉大学 3 人会話コーパス (Den and Enomoto 2007) の 2 会話（chiba0232 と chiba0432），合計約 20 分である。本コーパスには，簡略版 CSJ 方式による転記テキストと，発話単位・形態論情報・韻律情報などの種々のアノテーションが与えられている。

2.2 転記・アノテーション

2.2.1 CSJ 方式

CSJ 方式では，X–JToBI (五十嵐ほか 2006) に基づく韻律情報が提供されており，アクセント句の末尾に句末境界音調が付与される。句末境界音調として，具体的には

(1) 下降調（L%）
(2) 単純な上昇調（L%H%）
(3) 上昇前に一定期間低ピッチが見られる上昇調（L%LH%）
(4) 上昇下降調（L%HL%）
(5) 上昇下降上昇調（L%HLH%）

の合計 5 種類が認定される。ただし，上昇下降上昇調 L%HLH% は本データには出現しなかった。また，上昇下降調 L%HL% も下降調・上昇調に比べて極端に件数は少ない。下降調 L% は，複合境界音調が生じないアクセント句末に付与される音調であり，必ずしも明示的な下降が生じているわけではない。この点において，CA 方式のピリオド '.' とは若干異なる。また上昇調 L%H% および L%LH% は疑問上昇調だけでなく強調上昇調なども含まれており，クエスチョン '?' とは必ずしも一致しない。

2.2.2 CA 方式

CA 方式の転記に使われる種々の転記シンボルのうち，本研究では，

(i) ピリオド '.'（per）
(ii) クエスチョン '?'（ques）
(iii) コンマ ','（com）

の 3 つの音調マーカーに注目した。これらのマーカーはそれぞれ下降・上昇・継続の音調を表す。土屋ほか (2013) では平坦の音調を表すアンダーバー '_'（ub）も取り上げていたが，上記の転記シンボルに比べて事例数が極端に少ないため，本研究の分析対象からは除外した。

本研究で用いるデータには，Gail Jefferson の体系 (Jefferson 2004) に準拠した転記が，X 氏・Y 氏・Z 氏の 3 名の会話分析研究者によって作成されている。本研究ではこのうち，約 6 年の会話分析経験を有する Z 氏によって作成されたデータを用いる。Z 氏は，2004 年から 2007 年までカリフォルニア大学ロサンゼルス校で会話分析を学び，2007 年以降は日本国内のデータセッションや研究会に参加しており，大学院博士課程在籍時より主要な分析手法の 1 つとして会話分析を採用している。また，会話分析以外に，認知言語学と談話機能主義言語学の知識を有している。

2.3 言語・音響特徴

分析の基本単位として，土屋ほか (2013) による先行研究と同様に，強い切れ目で区切られたアクセント句（Break Index が 3 または 2+b, 2+p, 2+bp）を用いた。土屋ほか (2013) では，CSJ の言語・音響情報から CA 方式の音調マーカーを予測する多変量モデルの構築に際し，分析対象アクセント句から以下の言語・音響特徴が抽出され用いられている。[*1]

■言語特徴

句末境界音調（tone）　アクセント句末の句末境界音調。L%, H%, HL%, LH%。

末尾単語の品詞（lastPOS）　アクセント句末尾の単語の品詞。品詞は以下の 7 種に分類した。体言・用言・助動詞・終助詞・接続助詞・その他の助詞・その他の品詞。

次末単語の品詞（penultPOS）　アクセント句の最後から 2 番目（次末）の単語の品詞

発話中の位置（loc）　当該アクセント句の発話の先頭からの距離（アクセント句数）

発話末からの位置（revLoc）　当該アクセント句の発話の末尾からの距離（アクセント句数）

■音響特徴

アクセント句の最小 F0（f0MinAP）　アクセント句中の F0 の最小値（標準化得点）

アクセント句の最大 F0（f0MaxAP）　アクセント句中の F0 の最大値（標準化得点）

句末単語の最大 F0（f0MaxWord）　末尾単語中の F0 の最大値（標準化得点）

アクセント句の最大パワー（pwrMaxAP）　アクセント句中のパワーの最大値（標準化得点）

句末単語の最大パワー（pwrMaxWord）　末尾単語中のパワーの最大値（標準化得点）

アクセント句の平均モーラ長（amdAP）　アクセント句の継続時間をモーラ数で除したもの（標準化得点）

最右 F0 抽出点の値（lastF0Val）　アクセント句中で最後に抽出できた F0 点の値（標準化得点）

最右 F0 抽出点の位置（lastF0Loc）　最右 F0 抽出点の句末から計った時間（対数値）

F0 と平均モーラ長は対数変換後，パワーはそのままで，話者ごとに標準化得点に変換した。

3. 分析

3.1 先行研究の言語・音響特徴を用いたモデル

2.3 節の言語・音響特徴から Z 氏によって転記された chiba0232 と chiba0432 の音調マーカーを予測する多変量モデルを構築した。多変量モデルとしてランダムフォレスト法 (Breiman 2001) を用い，統計解析ソフト R 言語の `randomForest` パッケージを利用した（`mtry = 4` とした)。2 つのデータのうち一方を学習データとし，他方のデータの音調マーカーの予測結果と人手による音調マーカーを比較したところ，chiba0232 を学習データとしたモデルの正解率は 75.9%，chiba0432 を学習データとした正解率は 72.0% となり，比較的高い精度で予測ができていた。

[*1] 音響特徴として，アクセント句の平均 F0・句末単語の平均 F0・句末単語の最小 F0・アクセント句の平均パワー・句末単語の平均パワー・句末単語の平均モーラ長も抽出されたが，これらの特徴との相関が高いため用いられていない。

表 1　先行研究の言語・音響特徴によるモデルの予測結果（上昇調 H% のみ）

学習 = chiba0432，テスト = chiba0232
（正解率 = 54.4%，κ = .35）

	観測値			
予測値	none	per	ques	com
none	**14**	1	3	1
per	6	**38**	32	0
ques	0	1	**10**	0
com	2	1	0	**0**

学習 = chiba0232，テスト = chiba0432
（正解率 = 56.6%，κ = .36）

	観測値			
予測値	none	per	ques	com
none	**16**	0	4	4
per	7	**21**	9	0
ques	2	9	**10**	1
com	0	0	0	**0**

しかし，句末境界音調が上昇調 H% の結果だけを抜き出すと，表 1 に示すように正解率はそれぞれ 55% 前後でかなり低い。特に，H% の典型的な機能と考えられるクエスチョン（ques）をピリオド（per）と誤って予測する例が多い。

3.2　本研究の方法

前節に示したように，先行研究の言語・音響特徴では句末境界音調が H% である事例に対しクエスチョン（ques）の音調マーカーを精度良く予測することができない。以下では ques の予測に焦点をあて，誤った予測へ導く音響特徴の調査を行う。また，ques とそれ以外の音響マーカーを区分できる音響特徴を探る。

3.3　誤った予測へ導く音響特徴

H% の特徴として，句末の F0 の高さが考えられる。人手による音調マーカーが ques の場合における，最右 F0 抽出点の値（lastF0Val）を予測値ごとに分類した結果を図 1 に示す。図 1 からわかるように，lastF0Val が高い場合は正しく ques として予測できている。一方，lastF0Val が低い場合はピリオド（per）と誤って予測されている。すなわち，H% であってもアクセント句の最終 F0 が高くならない場合があり，そのようなアクセント句は正しく ques とは判別されないと考えられる。

しかし，転記者によって句末境界音調が H% と転記されていることから，これらは句末で F0

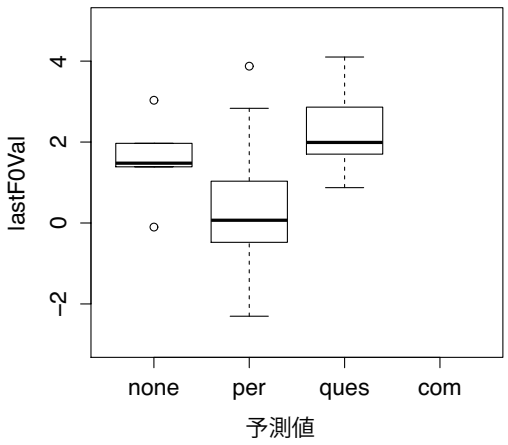


図 1　観測値が ques のときの最右 F0 抽出点の値（lastF0Val）

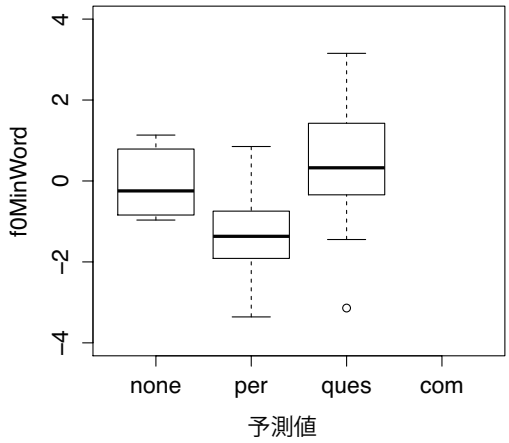


図 2　観測値が ques のときの句末単語の最小 F0（f0MinWord）

の上昇をともなうはずである。そこで，図 2 に示すような，人手による音調マーカーが ques の場合の句末単語の最小 F0（f0MinWord）を観察してみる。すると，lastF0Val と同様に予測値が per となったデータでは ques と比べて f0MinWord が低くなっている。つまり，句末単語（もしくはアクセント句）全体が低い F0 になっており，結果として句末の F0 が上昇しても lastF0Val の絶対的な値は通常の H% よりも低い値となっていると思われる。

3.4　新たな音響特徴の追加

ここで，2.3 節の音響特徴を利用して，句末の F0 上昇を表す音響特徴を導出することを考える。句末の F0 上昇は簡易的に次の音響特徴で表現できる。

末尾 F0 上昇幅（lastF0Rise）　最右 F0 抽出点の値（lastF0Val）から末尾単語の最小 F0 の値（f0MinWord）を引いたもの

人手による音調マーカーと lastF0Rise の関係を図 3 に示す。図 3 から明らかなように，ques でのみ lastF0Rise は大きな値となり，それ以外の音響マーカーでは 0 に近い値となる。すなわち，ques で起こる句末の F0 上昇をうまく表現できているといえる。

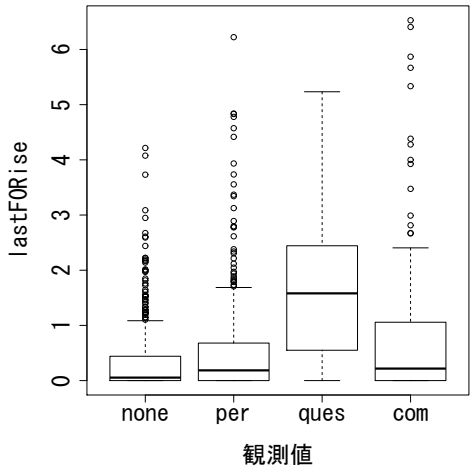


図 3　各観測値における末尾 F0 上昇幅（lastF0Rise）

表 2　句末の F0 上昇を加えたモデルによる予測結果（上昇調 H% のみ）

学習 = chiba0432，テスト = chiba0232
（正解率 = 57.8%，κ = .36）

	観測値			
予測値	none	per	ques	com
none	**14**	1	4	1
per	6	**38**	30	0
ques	0	1	**11**	0
com	2	1	0	**0**

学習 = chiba0232，テスト = chiba0432
（正解率 = 61.4%，κ = .43）

	観測値			
予測値	none	per	ques	com
none	**17**	0	3	4
per	6	**21**	7	0
ques	2	9	**13**	1
com	0	0	0	**0**

3.5　本研究のモデル

2.3 節の言語・音響特徴に `lastF0Rise` を加えて，ランダムフォレスト法による多変量モデルを構築した。chiba0232 を学習データとして chiba0432 の音調マーカーを予測した場合と，chiba0432 を学習データとして chiba0232 を予測した場合の，H% に対する結果を表 2 に示す。句末の F0 上昇を表す音響特徴が用いられているにも関わらず，`ques` の予測はあまり改善されず，chiba0232 の予測の正解率は 3.4 ポイントの上昇，chiba0432 の予測では 4.8 ポイントの上昇にとどまった。

予測に対する各言語・音響特徴の貢献度を図 4 に示す。今回導入した `lastF0Rise` は，chiba0232 では全体として 4 位，音響特徴としては `lastF0Loc` に次ぐ高い貢献度であった。このモデルにおいては `lastF0Rise` が予測に有効な音響特徴として働いているといえる。一方，chiba0432 では `lastF0Rise` は 8 位であり，`lastF0Val` と同程度の貢献しかしていない。

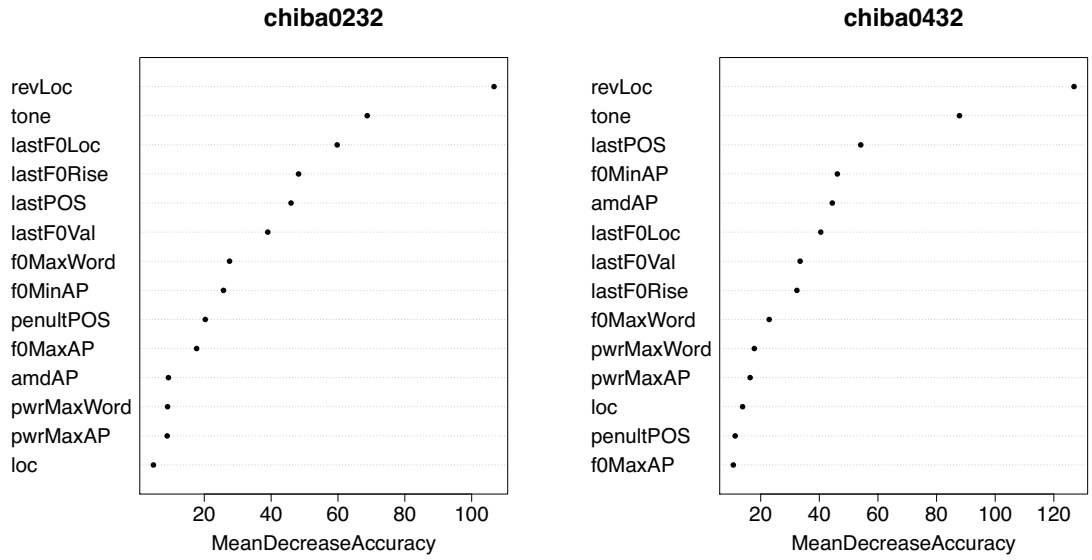


図 4　言語・音響特徴の重要度

4. 考察

句末の F0 上昇を表す音響特徴 lastF0Rise を用いたにもかかわらず ques の予測があまり改善されなかった要因のひとつとして，モデルの学習データにおける ques の件数の少なさが挙げられる。H% に限れば ques の件数は他の音調マーカーとあまり差がないが，L% のアクセント句は H% の約 3〜4 倍の件数がある上，L% では多くのアクセント句が per またはラベルなし（none）となっている。そのため，単純にモデルを構築すると L% の none および per を予測する精度が高くなるように学習され，比較的少数の H% の ques の予測の精度を上げることができないと考えられる。

実際のところ，ランダムフォレスト法の学習時のブートストラップサンプルの作成の際に，ques の件数と同じ件数となるよう他の音調マーカーの事例をサンプリングして学習させると，ques の予測の正解率が高くなるモデルが構築される。しかし，ques の件数が少ないことから全体のデータ数の減少につながり，none や per の正解率は大きく低下してしまう。よって，より高精度な予測を行うモデルを構築するためには，予測に役立つ新たな言語・音響情報を見いだすだけでなく，データ数の増加が必要であると思われる。

5. おわりに

本研究では，CSJ 方式の言語・音響情報から CA 方式の音調マーカーを予測するための多変量モデルを構築した先行研究において特に予測精度が悪かった上昇調 H% におけるクエスチョン（ques）に焦点をあて，誤った予測になる要因の解明を試みた。その結果，H% にも関わらず，句末付近の F0 値が低い事例が存在することがわかった。さらに，ques に関わる句末の F0 上昇を表すことのできる音響特徴を導出した。導出した音響特徴を用いて新たに多変量モデルを構築したところ，わずかに予測精度は向上したがまだ十分な精度には達しなかった。会話コーパス間の転記方式の相互変換に向けて，今後もさらなる言語・音響情報の検討や会話データの増加が必要である。

謝辞　会話分析方式の転記を作成していただいた遠藤智子・黒嶋智美・横森大輔の各氏に感謝します。本研究は国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴）による成果である。

参考文献

Breiman, Leo (2001). “Random forests.” *Machine Learning*, 45, pp. 5–32.

Den, Yasuharu, and Mika Enomoto (2007). “A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation.” Toyoaki Nishida (Ed.), *Conversational informatics: An engineering approach*. Hoboken, NJ: John Wiley & Sons. pp. 307–330.

五十嵐陽介・菊池英明・前川喜久雄 (2006).「韻律情報」　『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』　pp. 347–453.

Jefferson, Gail (2004). “Glossary of transcript symbols with an introduction.” Gene Lerner (Ed.), *Conversation analysis: Studies from the first generation*. Amsterdam/Philadelphia: John Benjamins. pp. 13–31.

土屋智行・伝康晴・小磯花絵 (2012).「会話コーパスの転記方式の相互変換に向けて—イントネーションに着目して—」『第 2 回コーパス日本語学ワークショップ予稿集』 pp. 117–126.

土屋智行・伝康晴・小磯花絵 (2013).「会話分析方式への転記変換におけるデータ間・個人間のゆれに関する分析」 『第 3 回コーパス日本語学ワークショップ予稿集』 pp. 417–424.

関連 URL

「会話コーパス」ホームページ : http://www.jdri.org/kaiwa/

ポスター発表（1） Bグループ

9月5日（木） 14:10～15:10

コーパスを用いた外来語サ変動詞の分析
―「マークする」を例として―

茂木 俊伸（鳴門教育大学 大学院学校教育研究科）†

A Corpus-based Study on Loanword Verbs in Japanese: A Case Study of *maaku-suru* (<*mark*)

MOGI Toshinobu (Graduate School of Education, Naruto University of Education)

１．はじめに

本研究では，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ と呼ぶ）のデータに基づき，外来語サ変動詞「マークする」の語彙・文法的特徴について分析を行う。

以下では，まず，本研究の問題意識を述べたうえで（第 2 節），BCCWJ における「マークする」の用例に基づき，この語の意味的特徴と構文的特徴を記述する（第 3 節）。さらに，類義表現との比較を行う（第 4 節）。

２．「外来語の文法」研究という視点

現代日本語の外来語（カタカナ語）研究の課題として，基本的な語に関する分析が遅れているという点が指摘される（cf. 石野 1996，金 2011 など）。外来語は日本語教育において学習の困難点になることが繰り返し指摘されているものの，そもそも個々の外来語が文中で具体的にどのように使われるかを明らかにする語彙的・文法的研究が手薄なため，応用分野に提供できる基礎的情報が不足した状態にあると言える。本研究は，日本語教育への応用を目指した外来語の記述研究の一つの試みである。

「マーク（する）」は，澤田（1993）で指摘されているように，サ変動詞用法を持つ基本的な外来語であると考えられる。佐々木（2001）は，管見のかぎりこの語の最も詳細な記述がなされている教材であり，(1)のように，すべての語義で他動詞として用いられることが明示されている（例文のルビと英訳，関連語の情報は省略した）。

(1)　マーク＜mark＞ 他動 名
①印をする。印。［―する／―をつける］
・該当する箇所にマークをつけてください。名
②記録や成績を達成する。［―する］
・今回の大会で佐藤選手は自己最高記録をマークした。動
③目をつける。［(…を) ―する・(…に) ―される］
・彼は事件のカギを握る人物として警察にマークされている。動　（同:87）

一方で，(1)からは，「マークする」が他動詞として典型的にどのような名詞をヲ格にとるのかは明確ではなく，例文から推測するしかない。また，語義③には受身形の文型と例が挙げられているが，他の語義にはない。小宮（1997）ではこの意味の「マークする」の理解度が低いことが指摘されており，もし受身形が語義③に特有のものであるならば，日本語学習者にとっては，語義の判別の際の形の手がかりが得られることになる。

以上のことから，外来語の分析においては，意味記述だけでなく文法的な側面を合わせ

† E-mail: tmogi@naruto-u.ac.jp

た「意味と形式の対応」に関する情報を示すことが必要であると言える（cf. 茂木 2011, 2012）。ただし，具体的にどのような現象を記述すべきなのかは，言語直感のみでは十分に検討することが難しい問題である。

したがって，例えば「マークする」のようなサ変動詞であれば，語義や用法の分類ごとに，自動詞か他動詞か，共起する格成分や文末形式に違いはあるか，(1)の「マークをつける」のような動詞相当の意味を表す句（cf. 村木 1982）はあるか，といった情報を，コーパスに基づいて実証的に明らかにしていくことが求められる。

３．事例研究—「マークする」の分析—

以下では，BCCWJ の用例をもとに「マークする」の意味分析を行い（3.1 節），次にこの動詞がどのような格成分や文末形式を伴って構文を形成するのかを見ていく（3.2 節）。

分析対象は，『中納言』（バージョン 1.0.5）を使用して得た「マークする」の用例 374 例である。これは，「マークする」「マークできる」の各活用形の例と，「30 勝をマーク。」のように「マーク」に句読点や記号（！？…）が直接続く例から成る。また，「徹底マークする」のような「漢語＋マーク」型の複合名詞も，意味から判断して対象に含めた。

3.1 「マークする」の意味的特徴

BCCWJ における「マークする」の用例を意味的に分類すると，次の〔表 1〕のようになる。ここでは辞書類も参考にしながら語義[1]〜[4]を立てた（提示順は辞書に準じている）。

それぞれの語義は，おおよそ，“（書いたり塗ったりして）印を作り，ポイントがそこであることを示す”（語義[1]），“人が（意味のある）記録を作る・打ち立てる”（語義[2]），“人が対象に注目し，継続的にその動向に注意を払う・警戒する”（語義[3]），“人が対象の後や周りに貼り付く（ことでその動きを封じる）”（語義[4]）のように定義できる。これらに共通する「マークする」の中核的な意味は，「注目すべき対象を作る」のようなものと考えられる。

〔表 1〕BCCWJ における「マークする」の意味分類

語義		用例数
[1] 印を作る	[場所]を塗りつぶす／強調する [場所]に印をつける	51
[2] 記録する	[数字]（成績）を達成する [記録]を打ち立てる	180
[3] 注意する	[人・モノ]に注目・注視する [人・モノ]に注意を払う・警戒する	64
[4] 人に付く	[人]を尾行・追跡する [選手]に付いて動きを封じる	79
計:		374

それぞれの語義の具体的な用例を，次に挙げる（出典はサンプル ID）。

(2) 語義[1]：印を作る

a. とにかく文書を直す箇所をマークする、どの文書が印刷に必要なのか、その準備を全部私がやってしまっておきたい。（OY14_49757）

b. ２サイズ以上大きめのサイズに、エスニックなベルトでウエストをマーク。（PM11_00996）

(3) 語義[2]：記録する

a. 昨年、長崎県で行われた全国中学大会の覇者。「暑いところの方が得意」という通り予選で自己ベストの十二秒０７をマーク。（PN1e_00003）

b. 白幡は得意の五千メートルで６分三十三秒５４の大会新をマークして１位となり、２位の糸川敏彦（コクド）とともに、五輪代表をほぼ確実にした。（PN1m_00015）
c. 今シリーズも前シリーズに引き続いて高視聴率をマークし、好調の『新キッズ・ウォー』。（PM51_01062）

(4) 語義[3]：注意する
a. 活躍間違いなしの白いバッグをマーク！合わせる色を選ばないうえ、おしゃれ度もピカイチ。春は白バッグ旋風が巻き起こりそうな気配濃厚！（PM51_00199）
b. ［住友重機］は、このところ、かなり高くなっており「そろそろ売りでは」と、マークしていた銘柄だ。（PB33_00700）
c. 今、われわれが、犯人として、マークしているのは、山本です。（LBq9_00258）

(5) 語義[4]：人に付く
a. でも、乃木ありさは、石黒が徹底的にマークしているはずである。マークしていながらここへ来たというのは、石黒をうまくまいたためだろうか。（LBf9_00120）
b. 中村俊輔がマークされるのであれば、彼がおとりになって遠藤にゲームを作らせた方がよかった。（OY15_10654）

語義[3]は，対象の行為や変化によって何らかの利益・不利益が生じるという予測の下で「注意する」のであるが，(4a)のように「注目する」と言い換えられるプラスのニュアンスを伴う例と，(4c)の「疑って警戒する」のようなマイナスのニュアンスを伴う例，さらに(4b)のように文脈を見てもどちらか明確でない例がある[1]。

また，ここでは，語義[3]と[4]を「主体の移動を伴うかどうか」で分けた。文脈上，「マークする」主体が対象にあわせて移動しているかどうかが明確ではない例（例えば(4c)）は，語義[3]に分類している。その意味で，語義[3]と[4]は連続性を持つ（3.2 節で見るように，両者には構文的要素にも共通する点が多い）[2]。

これらの語義[1]～[4]の用例の分布を，BCCWJ のサブコーパスごとに示したものが，次の〔表 2〕である[3]。表右端の「PMW」は，各サブコーパスにおける 100 万語（短単位）あたりの「マークする」（サ変動詞語幹「マーク」）の頻度を示したものである（以下，合計欄の PMW は BCCWJ 全体から見た値）。

〔表 2〕サブコーパスごとの「マークする」の出現

	語義[1]	語義[2]	語義[3]	語義[4]	計	PMW
出版・書籍	23	35	18	27	103	3.61
出版・雑誌	13	36	2	10	61	13.72
出版・新聞	0	40	2	0	42	30.65
図書館・書籍	8	8	20	25	61	2.01
特定目的・ベストセラー	0	0	3	2	5	1.34
特定目的・広報誌	0	1	0	0	1	0.27
特定目的・知恵袋	5	0	10	3	18	1.75
特定目的・ブログ	2	60	9	12	83	8.14
計:	51	180	64	79	374	3.56

1 『例文で読むカタカナ語の辞典（第 3 版）』（小学館辞典編集部 1998）では，この語義の「マークする」が「良い意味にも悪い意味にも用いられる」とされている。
2 『コンサイスカタカナ語辞典（第 4 版）』（三省堂編修所 2010）には，語義[4]から語義[3]が派生したと読める記述があるが，歴史的経緯については不明である。
3 特定目的サブコーパスのうち，韻文，法律，白書，教科書，国会会議録については用例が見られなかった。なお，「出版・書籍」（PB59_00521）と「図書館・書籍」（LBh9_00173）で同一の用例（1 例）が見られたが，ここでは両方にカウントしている。

単純な頻度で見た場合，「マークする」は書籍やブログでよく使われるように見えるものの，PMW は新聞が最も高い。また，語義[2]の用例自体はさまざまな媒体に見られるが，新聞の「マークする」の例はほとんどが語義[2]であることが分かる。

3.2 「マークする」の構文的特徴

次に，3.1 節で示した語義ごとに，共起成分や文末形式といった「マークする」の構文的特徴を見ていく。

3.2.1 連用成分

まず，「マークする」がどのような格成分や副詞的成分と共起するのかを見る。

次の〔表 3〕は，5 例以上見られた共起成分を挙げたものである（パーセンテージは，その語義の用例に占める共起の割合を示す。ただしヲ格とニ格は注 4 の補正を行った）。

〔表 3〕「マークする」の共起成分

語義	用例数	格成分							副詞的成分		
		ヲ	ニ	道具デ	場所デ	期間デ	時ニ	トシテ	時	期間	様態
[1] 印を作る	51	23	14	9						1	
		51.1%	31.1%	17.6%						2.0%	
[2] 記録する	180	168			44	8	27		12		
		100%			24.4%	4.4%	15.0%		6.7%		
[3] 注意する	64	16	1					6		7	6
		32.0%	2.0%					9.4%		10.9%	9.4%
[4] 人に付く	79	36	5					1		10	13
		51.4%	7.1%					1.3%		12.7%	16.5%
計	374	243	20	9	44	8	27	7	12	18	19

まず，対象がヲ格として現れない場合[4]を除くと，374 例中 243 例（65.0%）で「マークする」はヲ格成分と同一文中で共起している。特に語義[2]ではすべての例でヲ格名詞と共起しており，必須成分となっている。この時のヲ格名詞は，「５連勝」のような［具体的数値］の場合と，「世界新記録」のような記録の［種類を表す名詞］の場合がある（先の(3a,b)は「の」を介して両者が共起している例である）。また，語義[3]と[4]のヲ格名詞は，ほとんどが［人］であった[5]。

一方，語義[1]では，次の(6)のようにニ格成分と共起する例が一定数見られる。このニ格名詞は場所的であり，(6b)のように「［場所］に［印となるモノ］をマークする」という形でニ格とヲ格が共起する例も 3 例見られた。語義[3][4]にもニ格との共起例がある（(6c)）。

(6) a. だいたいあの本木が間違えて印をつけたからいけないのだ。間違ったほうにマークしていたのがそもそもの原因だ。（LBt9_00100）
b. 横線を引きたい文の切れ目に‘単独で’タグ〈ｈｒ〉をマークします。（PB35_00263）
c. そしてその前線へのパスの供給源である中田にも、張亨碩が激しくマーク。（LBm7_00043）

[4] 具体的には，対象が，1）ハで主題化されている，2）連体修飾の主名詞になっている，3）受身文の主語になっている，4）分裂文の後項になっている，というケースである。〔表 3〕の各語義のヲ格とニ格の欄の共起の割合は，これらを除いて算出した数値を示してある。

[5] 渡邊（2008:7）は，「［選手が］［選手を］他動詞」という文型をとる動詞として，語義[4]の「マークする」を挙げている。

この他の格成分で特徴的なのは，語義[1]の道具デ格（例：赤ペンで，ベルトで（(2b)）），語義[2]の時のニ格（例：昨年６月に）と場所や種目名を表すデ格（例：プラハ国際で，五千メートルで（(3b)）），期間を表すデ格（例：２年間で），そして語義[3]の複合格助詞トシテ（例：犯人として（(4c)））である[6]。

また，副詞的成分では，「今季／００年」のような時を表す表現が語義[2]に，「最後まで／常に」のような期間を表す表現と，「びったり(と)／激しく／徹底的に」のような様態（特に密着性）を表す表現が語義[3][4]に多く見られた。後者の様態の表現は，「密着マーク／徹底マークする」のような複合名詞（5 例）が語義[4]に見られることとも一致する。

以上の共起の状況をふまえ，10%の用例への出現を目安としてそれぞれの語義の典型的な文型を示したものが，次の〔表 4〕である（[] は名詞，《》は副詞的成分を表す）。

〔表 4〕「マークする」の語義と文型

語義		文型	例
[1] 印を作る	印をつける／強調する	[場所]{ヲ／ニ} ([道具]デ)	回答欄{を／に}（鉛筆で） マークする。
[2] 記録する	記録・成績を打ち立てる	[数値／記録の種類]ヲ (《時》(ニ)) ([場所]デ)	田中選手は （2000年(に) 国際大会で） 世界新記録を マークした。
[3] 注目する	注目・注視する／注意を払う・警戒する	[人]ヲ ([役割]トシテ) (《期間》)(《様態》)	警察は 彼を （容疑者として 先月から 徹底的に） マークしている。
[4] 人に付く	尾行・追跡する／周囲に貼り付く	[人]ヲ (《期間》) (《様態》)	鈴木選手は 相手選手を （最後までぴったり） マークした。

3.2.2 文末形式

次に，助動詞やテ形補助動詞のような，「マークする」に後接する文末形式の特徴を見る。〔表 5〕は，2 例以上見られた文末形式を挙げたものである（表右端には，「する」を伴わず「マーク」に句読点や記号が直接続く例の頻度を参考情報として示した）。

〔表 5〕「マークする」の文末形式

語義	用例数	受身	使役	テイル	テオク	テホシイ	ヨウ	スル略
[1] 印を作る	51	4	0	4	0	1	0	2
		7.8%	0%	7.8%	0%	2.0%	0%	3.9%
[2] 記録する	180	0	0	13	0	0	0	54
		0%	0%	7.2%	0%	0%	0%	30.0%
[3] 注意する	64	23	1	26	0	0	0	1
		35.9%	1.6%	40.6%	0%	0%	0%	1.6%
[4] 人に付く	79	13	2	21	2	1	3	3
		16.5%	2.5%	26.6%	2.5%	1.3%	3.8%	3.8%
計	374	40	3	64	2	2	3	60

まず，受身の助動詞「(ラ) レル」に関しては，語義[3]で後接する率が高い。これは，先の佐々木（2001）の記述(1)を裏付けるものである。語義[4]がそれに続くが，これらの語義の「マークする」は［人］をヲ格としてとるため，「［人］が［人・組織］にマークされる」

6 『日本語語彙大系』（NTT コミュニケーション科学研究所 1997）では，3 種類の「マークする」の構文が挙げられており，ニ格，トシテ格との共起がそれぞれ別の構文パターンとして示されている。

という受身文が作りやすいのだと考えられる。これらの受身文の例では，監視や付きまといに対する不快感やそれによって生じる困難といったニュアンスを伴うことが多い。

「テイル」の後接率に関しても語義[3]と[4]が高く，語義[3]では「受身＋テイル」の割合が最も高い（14.1%）。この語義[3]と[4]のテイル形は，すべて「継続・進行」解釈の例である。このことは，3.2.1 節で見た《期間》を表す副詞との共起とも一致する。

一方，語義[1]のテイル形はすべて「結果状態」解釈の例，語義[2]のテイル形は「経験」解釈の例であった。これを見るかぎり，語義[1]，語義[2]，語義[3][4]という，異なったアスペクト的特徴を持つ 3 種類の「マークする」が存在することになる。

興味深いことに，語義[2]には受身形の例がない。語義[2]の「マークする」には「テイル」以外の文末要素が後接せず，ほぼ基本形として現れると言ってよい。このことは，〔表 5〕に示したように，語義[2]に「する」が省略された例が顕著に多いことと相関していると考えられる。

3.3　分析

以上の BCCWJ の用例の観察に基づけば，外来語サ変動詞「マークする」には，その「意味」（語義）ごとに異なった「形」（文型）をとるという，一定の対応関係があることが指摘できる。この対応関係は，意味的に連続性を持つ語義[3]と[4]には構文的にも類似性が見られる一方で，これらと意味が異なる語義[1]および[2]との間には構文的な差が見られる，という 2 つの形で現れている。

４．関連表現との比較

最後に，「マークする」とそれに類する表現との共通点・相違点について見ていく。

4.1 「マークする」と「マークをつける」

先に(1)で見た佐々木（2001）では，語義①（ここでの語義[1]）の記述に「マークをつける」という表現が挙げられていた。では，語義[1]の「マークする」と「マークをつける」は同義と言ってよいだろうか。まず，BCCWJ に見られた「(「印」を表す）名詞「マーク」（複合名詞を除く）＋格助詞「を」＋動詞」という表現を，次の(7)に示す（以下，カッコ内は用例数。2 例以上のもののみ挙げる）[7]。

(7)　マークを　つける(33)，塗る(5)，描く(4)，書く(3)，貼る(3)，貼付する(2)，隠す(2)

(7)のように，名詞「マーク」は何らかの印を作成するもしくは付着させる動作を表す動詞と共起し，特に「つける」の例が多い。ただし，「マークをつける」には，「マークする」と同義である場合（(8a)）と，そうでない場合（(8b)）がある。

(8)　a. 次にノートの項目同士の関係や何のためにこんなことを書いたのか、つながりの意味が不明な部分は赤でマークをつけておき、翌日、友達や先生に聞きます。(PB33_00141)
　　b. （略）「私が死んだら棺桶にナショナルのマークをつけてくれ」と言う人がいたと聞いたこともあります。　(PB11_00077)

(8a)と(8b)の違いは，「マーク」の性質である。(8b)の「マーク」は，記号や商標のような具体的な形をしたものであり，目的の場所に「つける」前から「マーク」として成立している。これに対し，(8a)やサ変動詞「マークする」の「マーク」は，ペンを使って引いた線

[7] なお，「マークをする」の用例も 6 例見られた（語義[1]が 4 例，語義[3]と[4]が各 1 例）が，ここでは除いた。語義[2]の例がないのは，この語義の「マークする」がヲ格を必須成分としており（3.2.1 節），いわゆる二重ヲ格制約に抵触することからも予測できる。

や鉛筆で塗りつぶした楕円形などの比較的単純な形状をした印であり，動作の結果初めて「マーク」として機能することになると言える。

4.2 「マークする」と「記録する」

語義[2]の「マークする」は，漢語サ変動詞「記録する」に単純に置き換えが可能であるようにも見える。しかし，「記録する」は，「実験方法を記録する」「生活を写真で記録する」のように，ヲ格名詞が必ずしも「何らかの基準以上の顕著な数値や事実」を表さないという点において，「マークする」とは異なっている。

そこで，「マークする」に合わせて「顕著な数値や事実」をヲ格にとる「記録する」のサブコーパスごとの分布を見ると，次の〔表 6〕のようになる。

〔表 6〕「マークする」と「記録する」の分布

	マーク[2]	PMW	記録	PMW
出版・書籍	35	1.23	118	4.13
出版・雑誌	36	8.10	63	14.17
出版・新聞	40	29.19	46	33.57
図書館・書籍	8	0.26	104	3.42
特定目的・ベストセラー	0	0	7	1.87
特定目的・広報誌	1	0.27	8	2.13
特定目的・教科書	0	0	1	1.08
特定目的・白書	0	0	192	39.32
特定目的・国会会議録	0	0	12	2.35
特定目的・知恵袋	0	0	8	0.78
特定目的・ブログ	60	5.89	148	14.52
計:	180	1.72	707	6.74

ここから，「記録する」の方がより多くのサブコーパスに出現すること，また，新聞では両語とも PMW が高いが，白書では「記録する」がより多く使われることが見てとれる。

次に，両語のヲ格名詞を頻度上位 15 種類まで挙げると，(9)(10)のようになる。それぞれ a が記録や成績の［具体的数値］，b が［種類を表す名詞］（3.2.1 節）の例である（a のうち助数詞を伴わない数値表現は〈〉で，b のうち複合名詞・派生名詞は語例を添えて示す）。

(9) **「マークする」（語義[2]）のヲ格名詞**：
- a. ～秒など〈タイム〉(15)，～勝(13)，～km/h など〈速度〉(12)，〈スコア（ゴルフ）〉(7)，～m など〈飛距離〉(6)，～連勝(6)，～本塁打(5)，～割など〈打率〉(4)，～点〈得点〉(4)
- b. タイム（「好-」「トップ-」など）(12)，記録（「世界-」「新-」など）(9)，率（「視聴-」「打-」など）(8)，得点（「高-」など）(5)，新（「自己-」「日本-」など）(4)，数字（「高い-」など）(4)

(10)**「記録する」のヲ格名詞**：
- a. ～%(52)，～円など〈通貨単位〉(26)，～位(23)，～人〈人数〉(22)，～km/h など〈速度〉(21)，震度～(14)，～度〈気温など〉(12)
- b. 最高（「過去-」「戦後-」など）(66)，率（「成長-」「視聴-」「失業-」など）(47)，ヒット（「大-」など）(21)，成長（「経済-」「マイナス-」など）(19)，値（「最高-」「超安-」など）(18)，売り上げ(14)，最低（「過去-」「戦後-」など）(12)，セールス（「好-」など）(11)

(9)(10)のように，「マークする」のヲ格名詞は明らかにスポーツに関わるものが多いのに対し，「記録する」のヲ格名詞は気象や経済に関わるものが複数見られ，より多様である。

また，「記録する」のヲ格名詞には，「過去最低」「マイナス成長」のように程度「低」を表すもの，「失業率」のようなマイナスのニュアンスを伴うものが見られるのに対し，「マ

ークする」のヲ格名詞は「よい成績」を表すプラス評価のものに偏っている。

さらに，(10)のみに見られる名詞を「マークする」と共起させると，文体的に軽い印象を受けることから，「マークする」は有標の文体的特徴を持っていると言える。先に〔表 6〕で見たように，公的な文書である白書において，「マークする」ではなく「記録する」が選択されている理由は，このような文体的特徴にあるものと考えられる。

５．おわりに

茂木（2011）では，多義的な動詞「カットする」の記述の結果，「形から意味が，意味から形が，ある程度予測できる」ことを示し，このことが日本語学習上の手がかりになりうることを指摘した。第 3 節で示したように「マークする」に関しても同様のことが言えるが，さらに第 4 節のように関連表現と比較していくことの必要性も明らかになった。

今回の分析で最も興味深いふるまいが見られたのは，頻度が最も高い一方で，出現媒体や形態・構文が特徴的な語義[2]の「マークする」である。ただし，これらの事実の指摘がただちに教育上の意味を持つことにはならない。今後も外来語に関する基礎的な記述を続けていく過程で，ここで得られたような知見をどのように応用につなげるのかについても検討していく予定である。

謝辞

本研究は，日本学術振興会科研費若手研究(B)「日本語教育用辞書作成に向けた「外来語の文法」の記述的研究」（研究課題番号：25870484）の助成を受けている。

参考文献

石野博史（1996）「辞典における外来語の語義記述―「オープン」の場合―」『言語学林 1995-1996』，pp.273-286，三省堂.

金　愛蘭（2011）「20 世紀後半の新聞語彙における外来語の基本語化」『阪大日本語研究』別冊 3，大阪大学大学院文学研究科日本語学講座.

小宮修太郎（1997）「学習者の出身国別に見た外来語の理解度に関する比較考察」『筑波大学留学生センター日本語教育論集』12，pp.43-62，筑波大学留学生センター.

澤田田津子（1993）「日本語教育のための基本外来語について」『奈良教育大学紀要（人文・社会科学）』42:1，pp.225-239，奈良教育大学.

村木新次郎（1982）「外来語と機能動詞―「クレームを つける」「プレッシャーを かける」などの表現をめぐって―」『武蔵大学人文学会雑誌』13:4，pp.226-211，武蔵大学人文学会.

茂木俊伸（2011）「コーパスを用いた外来語サ変動詞の分析―「カットする」を例として―」『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』，pp.103-110，文部科学省科学研究費特定領域研究「日本語コーパス」総括班.
（http://www.ninjal.ac.jp/corpus_center/bccwj/doc/workshop/JC-G-10-02.pdf よりダウンロード可能）

茂木俊伸（2012）「文法的視点からみた外来語―外来語の品詞性とコロケーション―」『外来語研究の新展開』（陣内正敬，田中牧郎，相澤正夫(編)），pp.46-61，おうふう.

渡邊ゆかり（2008）「サッカー中継で用いられる外来語」『広島女学院大学日本文学』18，pp.1-38，広島女学院大学文学部日本語日本文学科.

辞書・教材等

NTT コミュニケーション科学研究所(監修)（1997）『日本語語彙大系 5　構文体系』，岩波書店.

佐々木瑞枝(監修)（2001）『アカデミック・ジャパニーズ 日本語表現ハンドブックシリーズ 5　よく使うカタカナ語』，アルク.

三省堂編修所(編)（2010）『コンサイスカタカナ語辞典（第 4 版）』，三省堂.

小学館辞典編集部(編)（1998）『例文で読むカタカナ語の辞典（第 3 版）』，小学館.

現代日本語における汎用的漢語サ変動詞の抽出とその内部構成の検討

李　楓（神戸大学大学院国際文化学研究科）

Identification of Major Sino-Japanese-Verbs Ending with *-suru* and Analysis of Their Internal Structures

Feng Li（Graduate School of Intercultural Studies, Kobe University）

１．はじめに

世界の中で日本語教育が広く行われるようになるにつれ、日本語の語彙指導についても関心が高まっている。日本語の語彙には、和語、漢語、外来語など、様々な語種が存在するが、数のうえで特に多数を占めるのは漢語であり、その多くが漢語サ変動詞の形を取るとされている。

しかしながら、外国人日本語学習者、特に学習者数の多い中国人日本語学習者の視点で見た場合、漢語サ変動詞の種類や用法などについて、必要とされる情報がすべて明らかにされているとは言えない状況にある。たとえば、国語辞典や日本語語彙教材では、漢語名詞の後に「～する」と添え書きしたり、漢語名詞の用例の一部として「～する」の形を示したりするのが一般的で、漢語サ変動詞そのものを単独の語として認定し、その使い方を体系的に説明しているものはほとんど存在しない。また、無数に存在する漢語サ変動詞の中で、典型的に使用される漢語サ変動詞にどのようなものがあり、それらがどのような内部構成を持っているかといった点についてもはっきりしない。

もちろん、日本語学の中では、漢語サ変動詞の用法に関わる問題は古くより取り上げられてきた。ただし、先行研究の多くはもっぱら理論的な分類や構造の解明を目指しており、現代日本語における実際の使用状況をふまえた研究は必ずしも多くない。しかしながら、学習者のニーズから言うと、まずもって明らかにされるべきことは、日本語を構成する様々な変種において広く使用されている漢語サ変動詞にはどのようなものがあり、それらがどのような内部構成パタンを持つのかがわかりやすく示されることであろう。

以上の点をふまえ、本研究においては、日本語初の大型コーパスである現代日本語書き言葉均衡コーパス（BCCWJ）を用い、各種の日本語変種の中で、汎用的、かつ、高頻度に使用されている漢語サ変動詞を特定したうえで、その内部構成を概観していくこととしたい。BCCWJ には、様々な言語データが収められているが、本研究では、一定数の読者を持ち、外国人学習者にとっても重要であると考えられる書籍、雑誌、新聞、ブログ、白書、知恵袋の 6 種に限り、頻度調査を行っていくこととする。

２．先行研究

すでに述べたように、本研究では、現代日本語の各変種の頻度をふまえ、汎用的な漢語サ変動詞を特定すること、及びその内部構成の特性を明らかにすることを主たる目的とする。前者に関わる先行研究としては、日本語の計量的な語彙調査が古くより活発に行われている。例えば、国立国語研究所（1962；1973；1983）では、雑誌、新聞、教科書などの語彙調査が行われており、漢語を含む日本語語彙の使用状況が一定の範囲で明らかにされている。また、橋本（2009）や松下（2011）は、BCCWJ を用いて、教育的語彙リストの作成を試みている。しかしながら、漢語サ変動詞に対象を絞り、また現代日本語の多様な変種に目配りを行った調査は必ずしも十分ではない。

後者については、特に 2 字漢語から派生した漢語サ変動詞について、理論的立場から、内部構成の分類方法がいくつか提唱されている。ここでは、代表的なものとして、日向（1985）、野村（1999）、小林（2004）を概観する。

日向（1985）は、現代日本語における漢語サ変動詞の構造を、並立関係（添加、開閉など）、修飾関係（直送、軽視など）、客体関係（加熱、失望など）、実質関係（強化、酸化など）の 4 種類に分類している。同研究では、4 分類のうち、特に修飾関係と客体関係に焦点を当てて考察している。

野村（1999）は、漢語サ変動詞の分類を精緻化するため、まず、漢語を構成する個々の文字タイプと、文字間の基本的な結合タイプをそれぞれ 5 種類に下位分類したうえで、これらを組み合わせ、全体で 22 種類からなる内部構造パタンを提唱している。文字タイプの下位分類は、叙述の対象となる物や事を表す「事物類（N)」（鉄、国、水、士、道など)、事物の動作・作用を表す「動態類（V)」（見、増、置、感など)、事物や精神の性質・状態を表す「様相類（A)」（新、軽、大、高など)、動作や状態の程度・内容を限定・修飾する「副用類（M)」（特、再、絶、予など)、語基について形式的な意味を添える「接辞（s)」（不、御、的、性など）の 5 種類である。次に、結合タイプの下位分類は、補足関係（+)、修飾関係（>)、並立関係（・)、対立関係（-)、反復関係（=）の 5 種類である。これらを組み合わせることで、漢語サ変動詞は、「N+V」（気絶、骨折など)、「V+N」（握手、開花など)、「A＋N」（多言、貧血など)、「N>V」（音読、兄事など)、「V>V」（愛用、滑降など)、「A>V」（安眠、軽視など)、「M>V」（一掃、共感など)、「V・V」（引退、救助など)、「V-V」（開閉、屈伸など)、「V=V」（云々、転々など)、「N>N」（金策、原因など)、「V>N」（起因、残業など)、「A>N」（紅葉、粗食など)、「N・N」（影響、葛藤など)、「N-N」（左右、始末など)、「A=A」（清々)、「sV」（殺到、所期など)、「sA」（不精)、「Ns」（液化、酸化など)、「Vs」（欠如、消化など)、「As」（悪化、強化など)、「その他」（運休、軍縮など）の 22 種に整理される。

小林（2004）は、野村（1999）をはじめとする一連の研究を継承しつつ、漢語サ変動詞の範囲を漢語部分が 2 字以外のものに拡張し、幅広い漢語の構成を分析した。このうち、2 字漢語（「動名詞」と称される）については、動詞的要素と名詞的要素で構成される VN-N タイプ（読書、投票など)、動詞的要素と動詞的要素で構成される VN-VN タイプ（使用、殴殺など)、付加詞的要素と動詞的要素で構成される ADJ-VN タイプ（銃殺、病死など)、及び構成要素が抽出できないもの（挨拶、支配、勉強など）の 4 種類に分類している。なお、VN-N タイプについては、内部の名詞的要素と関係づけられた項を取るかどうかという観点から、項を取れないタイプ（飲酒、処刑など)、項を取れるタイプ（投票、登山など)、項を取らなければならないタイプ（開封、除名など）の 3 種類に下位分類している。また、VN-VN タイプについては、語の意味の中心をなし、語全体の品詞を決める主要部の性質や構成要素間の意味的結合関係に基づき、両側主要部タイプ（使用、委託など)、右側主要部タイプ（殴殺、急行など)、左側主要部タイプ（採用、購読など）の 3 種類に下位分類している。

以上の 3 氏の分類を整理すると、おおよそ以下のような対応関係になると考えられる（表 1)。ただし、この対応関係は大まかなもので、必ずしもこのように明示的に述べられているわけではないことに注意されたい。

表 1 先行研究における漢語サ変動詞の内部構成

日向（1985）	並立関係	修飾関係	客体関係	実質関係	
野村（1999）	V・V、V-V、V=V、N・N、N-N、A=A	A+N、N>V、V>V、A>V、M>V、N>N、V>N、A>N	V＋N、N+V	sV、sA、Ns、Vs、As	その他
小林（2004）	VN-VN	ADJ-VN	VN-N	構成要素が抽出できないもの	

以上で、日向（1985)、野村（1999)、小林（2004）の 3 つの枠組みにおける漢語サ変動詞の内部構成パタンの分類方法について概観してきた。これらを概観して気が付くことは、

3 氏はそれぞれ類似した方向性を持ちながらも、細かい点では異なりが多いことである。たとえば、内部構成パタンの数は、日向（1985）では 4 種類、野村（1999）では（基本タイプに限ると）5 種類、小林（2004）では主として 3 種類である。また、構成要素間の結合関係パタンに注目すると、日向（1985）が提唱した客体関係と実質関係は野村（1999）では採用されておらず、一方、野村（1999）は新たに補足関係、対立関係、反復関係を加えている。これに対し、小林（2004）はもっぱら統語的関係性に分類の焦点を当てており、結合関係パタンについては明確に論じられていない。

3 氏の分類はそれぞれに理由のあるものであると考えられるが、教育的観点から漢語サ変動詞の内部構成のパタンを示そうとする場合、これらの先行モデルの統合と精選が必要になるだろう。

３．リサーチデザイン

３.１　本研究の目的

既に述べたように、本研究は、書籍、雑誌、新聞、ブログ、白書、知恵袋の 6 種の変種データを用い、汎用的漢語サ変動詞を計量的に特定したうえで、その内部構成を明らかにすることを目的とする。調査にあたり、以下のリサーチクエスチョンを設定した。

RQ1. 現代日本語の各変種において、高頻度漢語サ変動詞に内容的な違いが見られるか？
RQ2. 各変種の頻度情報を統合することで、どのようなものが汎用的漢語サ変動詞として抽出されるか？
RQ3. 高頻度・汎用的漢語サ変動詞の内部構成パタンとしては、どのようなものが多いか？

３.２　データ

BCCWJ で使用する書籍、雑誌、新聞、ブログ、白書、知恵袋の 6 変種のうち、書籍については、「出版・書籍」と「図書館・書籍」を分析対象とする。前者は、2001 年から 2005 年の間に国内で刊行された書籍で、国立国会図書館の蔵書目録を電子化した「J-BISC」を元に決定された母集団から、日本十進分類法（NDC）及び発行年度ごとに層別に抽出したものである。後者は、東京都立中央図書館の作成した「ISBN 総合目録」に基づき、1986 年から 2005 年までの 20 年間に発行された書籍を母集団としたランダムサンプルである。「収集目的が異なるが、2 種類の書籍データはともに、NDC と発行年をもとにランダムサンプリングされているため、任意に抽出した部分も母集団の NDC の構成比をそのまま表していることが期待される」（大石、2012）ことから、ここでは、2 種類の書籍データをまとめて使用する。

ただし、書籍は BCCWJ の中で、圧倒的なデータ量を占めている。そこで、以下の分析においては、書籍を 1 変種として扱うのではなく、NDC による 10 ジャンル（0.総記、1.哲学、2.歴史、3.社会科学、4.自然科学、5.技術・工学、6.産業、7.芸術、8.言語、9.文学）をそれぞれ独立した言語変種と見なして分析を行う。これにより、他の変種との分量バランスの偏りが改善される。すなわち、本研究で、対象とするのは、書籍の 10 種、その他の 5 種、あわせて 15 種の変種データである。

漢語サ変動詞の頻度情報の取得には、「中納言」を使用する。「中納言」では、短単位検索、長単位検索、文字列検索の 3 種類の検索が可能であるが、本研究の目的に応じて、短単位検索を利用することとした。データの採集は 2013 年 7 月に実施した。

３.３　調査対象の定義

小林（2004）のように、漢語サ変動詞を広義で捉えれば、1 文字のもの、2 文字のもの、3 文字以上のものなど、幅広い語が対象に含まれる。しかし、本研究においては、2 文字からなる漢語サ変動詞に対象を限定する。これは以下の 3 つの理由による。1 点目として、種類や使用頻度の点において、2 字漢語が漢語語彙の圧倒的多数を占めるためである。2 点目

は、3 字以上の漢語に比べて、2 字漢語は「ひとまとまり性」が強く（湯本、1977)、計量的に調査しやすいためである。3 点目は、中日 2 言語の対照性が高く、中国人学習者にとって馴染みの深い語が多いためである。

しかし、対象を 2 字漢語と定めても、具体的に用例を検証していくと、多くの境界例が存在することに気が付く。そこで、統一した処理を行うために、本研究で対象とする漢語サ変動詞の決定ルールを以下のように定めた。

(1)　「X する」(「する」の活用形を含める）の形が存在する。
(2)　「X する」が全体として何らかの動作もしくは変化を表す。
(3)　「X」は原則として 2 文字とする。ただし、「X」が 3 文字以上の場合、後項 2 文字だけで独立した意味を持ち、かつ、「する」と結合しうる場合は、「後項 2 文字+する」の形で対象に含める。

対象になるものと対象にならないものの例を以下の表 2 に示す。

表 2　本研究が対象とする漢語サ変動詞の例

	対象となるもの	対象とならないもの
「X」が 2 文字	感謝する、1 泊する、2 分する	艶々する
「X」が 3 文字	再確認する（→確認する)、一安心する（→安心する）	不自由する、土下座する
「X」が 4～文字	調査検討する（→検討する)、悪戦苦闘する（→苦闘する）	右往左往する、意気投合する

「1 泊する」や「2 分する」は漢数字の異表記と見なして対象に含める。一方、「艶々する」は形容詞的に状態を描写し、上記のルール（2）に抵触するため、対象外となる。また、「不自由する」や「右往左往する」などは、後項 2 文字部だけで「する」と結合することはなく、ルール（3）に抵触するため、同じく対象外となる。

３.４　手法

まず、RQ1 については、15 種の言語変種ごとに、「中納言」を用いて漢語サ変動詞の抽出を行う。その際、検索対象を「スル（語彙素読み)」に指定する。出力された結果をすべてダウンロードし、前述のルールと照合して、対象語の特定を行う。

なお、「X する」の「X」部分が 3 文字以上からなるものについては、前節のルール（3）により、「後項 2 文字＋する」の形が成立すれば対象に含めるわけだが、当該形が日本語として成立するかどうかの判断は時に主観的となりうる。そこで、当該形がコーパス内で 2 例以上出現していることを当該形の成立の判断根拠とする。「不自由する」の例で言うと、「後項 2 文字＋する」形である「自由する」が仮にコーパス内で 2 例以上存在していれば、対象に含めることになる（2 例以上とすることで、特殊な効果を狙った例外的用例などが排除できる)。その後、言語変種別に高頻度に使われている漢語サ変動詞をリストアップし、サンプルとして上位 5 語を概観する。

次に RQ2 については、まず、15 種の言語変種からそれぞれ高頻度上位 50 語、延べ 1250 語を抽出し、重複を除いた 279 語を決める。次に、一定の汎用性を確認するため、15 変種中、1 変種でしか上位 50 語に入っていない 2 語（「出走」、「共起」）を除き、277 語を対象として変種別の頻度調査を行う。その後、主成分分析により、15 変種中の頻度を合成する。最後に、第 1 主成分得点に基づいて 277 語を降順に並べ替え、主成分得点が正（+）となる 93 語を高頻度かつ汎用的な漢語サ変動詞として特定する。

最後に、RQ3 については、RQ2 で特定された 93 語の各々を次節で述べる内部構成パタン別に分類する。その後、タイプベースとトークンベースの 2 種類の観点で計量的に概観し、

最も典型的な構成パタンを明らかにする。タイプ（異なり語数）ベースでは、分析対象語の総語種数（93）を分母として、各パタンに含まれる語種数比率を調べる。また、トークン（延べ語数）ベースでは、93 語の各々が持つ第 1 主成分得点の総計値を分母として、各パタンに含まれる語の第 1 主成分得点の合計値の比率を求める。

３.５　内部構成パタンの分類法

既に述べたように、漢語サ変動詞の内部構成モデルを提唱した日向（1985）、野村（1999）、小林（2004）の 3 氏の枠組みの間には微妙なずれがあり、また一部の分類は過剰に細かなもので、学習者にとってわかりやすいものとは言い難い。

そこで、本研究では、先行研究の分類の融合と精選を行う。

まず、構成要素間の結合関係パタンについては、日向（1985）の枠組みに新たに「補充関係」を加え、全体で 5 分類とする。「補充関係」というのは、「拡大」「縮小」などのように、後項構成要素が前項構成要素の意味を補充するものである。また、品詞的結合パタンについては、野村（1999）で提唱されている 22 分類のうち、「A=A」、「sA」及び「その他」を削除し、「V・V」、「V-V」、「V=V」の 3 種を「V-V」、「A+N」と「A>N」の 2 種を「A>N」にまとめ、さらに「M・s」を加え、合計 18 分類とする。「A=A」、「sA」を除外したのは、当該例における語幹部の名詞性が曖昧であると考えたためである。なお、品詞的結合関係を表す記号については、野村（1999）を参考にしながら、並立関係は「-」、修飾関係は「>」、補充関係は「<」、実質関係は「・」とする。以上により、本研究で使用する分類の枠組みは以下の表 3 に示すとおりとなる。例の大部分は野村（1999）から引用している。

表 3　内部構成の枠組み

構成要素間結合関係パタン	品詞的結合パタン	例	構成要素間結合関係パタン	品詞的結合パタン	例
並立関係	V-V	選択、開閉	客体関係	N+V	骨折、気絶
	N-N	意味、意見		V+N	看病、入院
修飾関係	A>N	紅葉、大病	補充関係	V<A	拡大、縮小
	A>V	短縮、軽減	実質関係	V・s	進化、激化
	M>V	予防、再建		N・s	風化、電化
	N>N	手術、病気		A・s	酸化、強化
	V>V	誤診、傾聴		M・s	特化
	N>V	林立、列挙		s・A	不精
	V>N	起因、残業		s・V	否認、不足

漢語サ変動詞の構成要素となる個々の文字の意味や品詞性については、判断の揺らぎを避けるため、『新潮日本語漢字辞典』（新潮社、2008）、『漢字源　改訂第五版』（学習研究社、2011）、『三省堂常用漢字辞典』（三省堂、2013）の 3 種類の漢字辞書の記述に従い、当該文字の持つ訓や意味を基準として分類する。1 つの漢字が 2 つ以上の訓読みや意味を持つ場合、また、同じ意味を表す訓が複数ある場合などは、機械的に最も先頭に書かれているものを採用する。また、辞書間で記述が揃っていない場合は、3 つの辞書のうち、2 つに出現しているものを採用する。さらに、3 辞書間で統一した意味が見つからない場合は、収録内容が最も多く、記述が最も詳細である『新潮日本語漢字辞典』に準ずる。

以上のような処理手順を取ることで、同じ語が先行研究と異なる分類を与えられる場合もある。例えば、「挨拶」を例にすると、小林（2004）はこれを「構成要素が抽出できないもの」としているが、上記の辞書の解説をふまえると、「挨拶」は「挨（せま）る＋拶（せま）る」と読み下ろすことができるので、本研究ではこれを並立関係の「V-V」と認定する。なお、このとき、「挨拶」という語の意味と「挨る／拶る」（相手に近付く）の意味の間には若干のず

れがあるが、本研究では、構成要素の意味間の関係性を基準として分類する。

４．結果と考察

４.１　RQ1　日本語の各変種における高頻度漢語サ変動詞

15 種の言語変種ごとに高頻度漢語サ変動詞を調査したところ、以下の表 4 のようになった（語幹部のみを示す）。

表 4　言語変種ごとの高頻度漢語サ変動詞（上位 5 語）

	書籍・総記	書籍・哲学	書籍・歴史	書籍・社会科学	書籍・自然科学
1	表示	存在	発見	存在	存在
2	利用	理解	存在	実施	利用
3	実行	説明	主張	利用	発生
4	使用	実現	出土	説明	結合
5	発生	意味	利用	規定	報告
	書籍・技術・工学	書籍・産業	書籍・芸術	書籍・言語	書籍・文学
1	表示	利用	演奏	存在	説明
2	使用	提供	表現	説明	結婚
3	利用	紹介	紹介	意味	発見
4	発生	発生	存在	表現	心配
5	設定	使用	使用	注意	存在
	雑誌	新聞	ブログ	白書	知恵袋
1	発売	発見	発売	実施	表示
2	開催	期待	開催	開催	削除
3	紹介	予想	紹介	設置	質問
4	用意	確認	確認	期待	使用
5	掲載	注目	期待	決定	購入

上記で明らかなことは、変種により、高頻度漢語サ変動詞の内容に大きな違いが存在することである。例えば、15 変種において、共通して上位 5 語に含まれた語は 1 語も存在しない。このことは、現代日本語における漢語サ変動詞を議論するにあたり、各種の言語変種の差に十分に留意する必要があることを示している。

４.２　RQ2　高頻度・汎用的漢語サ変動詞の特定

前述の 277 語について 15 種の変種ごとに頻度を調査し、主成分分析を実行したところ、一般に有効主成分とされる固有値 1.0 以上の主成分が 5 つ抽出された。

このうち、第 1 主成分に注目すると、15 変種すべてに対して負荷量が正となっており、第 1 主成分が複数変数の合成指標となっていることが確認された。第 1 主成分の寄与率は 38.33%となり、元のデータの分散の約 4 割が第 1 主成分に集約されたことになる。以下に示すのは、15 変種の第 1、第 2 主成分負荷量表（表 5）、それらを横軸・縦軸とする二次元象限上にデータを布置した散布図（図 1）、及び、第 1 主成分得点の降順で全体を並べ替え、指標値が正になった 93 語のリスト（表 6）である。

はじめに、表 5 と図 1 に注目する。15 変種の中では、書籍、特に社会科学（0.816）や産業（0.809）の分野に相対的に大きな負荷量がかかっている。しかし、ブログ（0.521）など、書籍以外の変種にも一定の負荷量がかかっており、全体として日本語の多様な言語変種をバランスよく代表した指標値が取り出せたと考えられる。

表 5　主成分負荷量

変種	主成分 1	主成分 2
書籍・総記	0.571	0.313
書籍・哲学	0.704	-0.521
書籍・歴史	0.747	-0.277
書籍・社会科学	0.816	-0.140
書籍・自然科学	0.644	-0.180
書籍・技術・工学	0.742	0.250
書籍・産業	0.809	0.025
書籍・芸術	0.748	-0.101
書籍・言語	0.556	-0.506
書籍・文学	0.575	-0.323
雑誌	0.525	0.535
新聞	0.368	0.361
ブログ	0.521	0.589
白書	0.327	0.332
知恵袋	0.307	0.429

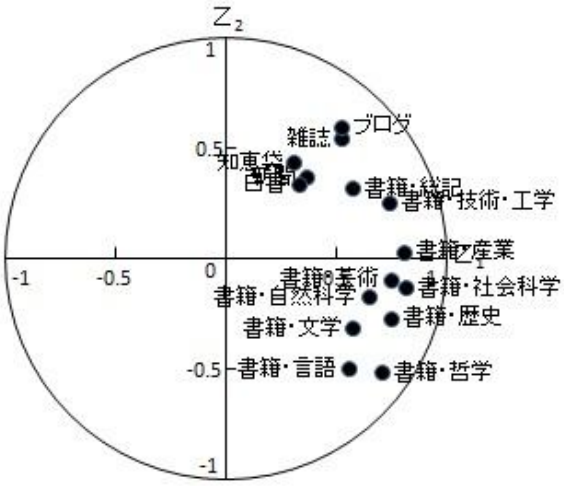


図 1　負荷量散布図

表 6　高頻度・汎用的漢語サ変動詞（主成分得点>0）

順位	語	得点	順位	語	得点	順位	語	得点	順位	語	得点
1	存在	12.457	25	実現	3.353	49	発売	1.841	73	発揮	0.671
2	利用	11.823	26	表現	3.207	50	意識	1.806	74	一致	0.626
3	紹介	8.992	27	展開	3.064	51	移動	1.728	75	無視	0.621
4	説明	8.984	28	参加	3.058	52	指定	1.713	76	維持	0.534
5	使用	8.711	29	検討	3.023	53	強調	1.703	77	適用	0.506
6	確認	8.327	30	意味	2.827	54	発展	1.546	78	確立	0.506
7	理解	7.135	31	対応	2.823	55	増加	1.521	79	記載	0.499
8	表示	6.141	32	主張	2.797	56	開発	1.467	80	勉強	0.469
9	発見	5.946	33	要求	2.787	57	否定	1.402	81	分類	0.451
10	発生	5.726	34	完成	2.649	58	提出	1.368	82	安定	0.368
11	注目	5.148	35	判断	2.642	59	反映	1.332	83	観察	0.360
12	指摘	4.992	36	成功	2.609	60	規定	1.239	84	保存	0.303
13	期待	4.906	37	設置	2.606	61	代表	1.095	85	心配	0.281
14	実施	4.408	38	決定	2.431	62	掲載	1.024	86	重視	0.264
15	用意	4.339	39	作成	2.350	63	支配	1.013	87	考慮	0.253
16	発表	4.018	40	選択	2.208	64	結婚	0.960	88	販売	0.220
17	注意	3.989	41	開始	2.183	65	実行	0.884	89	想像	0.209
18	開催	3.834	42	採用	2.162	66	変更	0.858	90	活用	0.140
19	構成	3.658	43	登場	2.135	67	拡大	0.774	91	限定	0.133
20	評価	3.649	44	認識	2.046	68	解決	0.768	92	減少	0.129
21	設定	3.623	45	予想	1.984	69	解放	0.764	93	区別	0.111
22	変化	3.488	46	成立	1.974	70	比較	0.734			
23	提供	3.436	47	報告	1.901	71	集中	0.727			
24	形成	3.365	48	導入	1.873	72	購入	0.683			

次に、表 6 に注目する。一般に、雑誌や新聞など、情報伝達を主とする言語変種に限定して頻度調査を行えば、当該変種の特性に影響を受け、分野に依存した難語のみが多く選定されがちである。しかし、幅広い言語変種をデータに加えたことで、「存在」「認識」と

いった抽象的な語のみならず、「使用」「説明」などの一般的な語、「結婚」「勉強」といった個人生活に関わる語、さらには「販売」「購入」「発売」といった社会・経済関連語にいたるまで、現代日本語の様々な環境を反映した汎用的な漢語サ変動詞の抽出が行われたと考えられる。

既に述べたように、無数に存在すると思われる漢語サ変動詞のうち、高頻度、かつ、汎用的に使用される語のリストはこれまで十分に整備されてこなかった。表 6 のようなデータがあれば、日本語教育において有効に活用しうるのみならず、日本語研究においても、実際の日本語で典型的に使用される項目に限った漢語サ変動詞のシンプルな用法モデルを構築することが可能になるであろう。以下、本節で得られた 93 語に限定して、漢語サ変動詞の内部構成パタンを概観していく。

４.３　RQ3　内部構成パタンの計量的概観

高頻度・汎用的漢語サ変動詞の 93 語に対して、構成要素間結合関係パタン及び品詞的結合パタンという 2 点から、その内部構成を分類した。結果として、以下の図 2~5 が得られた。

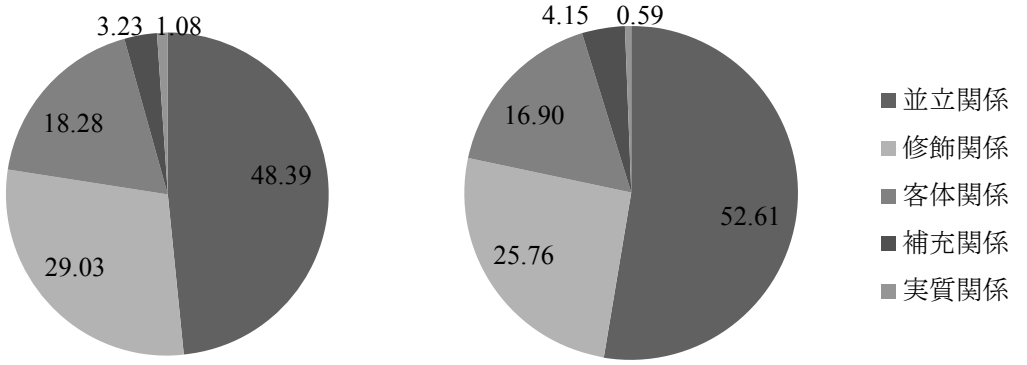


図 2・3　構成要素間結合関係パタン別比率（左図：タイプベース；右図：トークンベース）

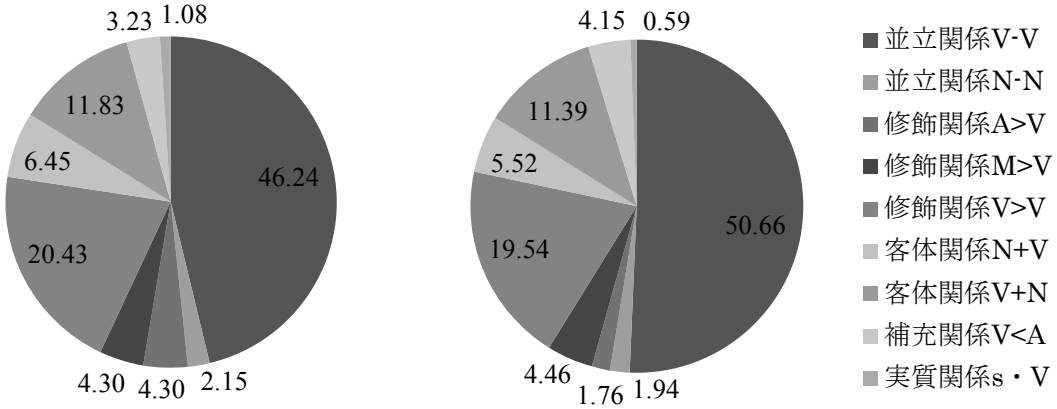


図 4・5　品詞的結合パタン別比率（左図：タイプベース；右図：トークンベース）

先行研究は、漢語サ変動詞の内部構成について様々な分類の枠組みを提案する一方、そもそも何が代表的で典型的であるかについてはほとんど触れていなかったわけであるが、コーパスから得られた高頻度汎用的漢語サ変動詞に限定した計量分析により、当該動詞形

の内部構成の典型性について、極めて重要な知見が得られた。ここでは、4 点に限って言及する。

1 点目は漢語サ変動詞の内部構成の安定性についてである。タイプベースとトークンベースを比較しても、全体の傾向は大きく変化していない。このことは、漢語サ変動詞の内部構成パタンが一定の安定性を持つことを示唆する。

2 点目は漢語サ変動詞の内部構成の多様性の低さについてである。各種のパタンのうち、高頻度に使用されているものはごくわずかしかない。構成要素間結合関係パタン（図 2~3）について言えば、並立関係のみで全体の過半を占め、これに修飾関係、客体関係の 2 種を加えれば全体の 9 割以上を占める。また、品詞的結合パタン（図 4~5）について言えば、並立関係の「V-V」だけで全体の過半を占め、修飾関係の「V>V」と客体関係の「V+N」を含めると全体の 8 割以上を占める。見かけ上の多様性と異なり、実際の言語運用において、漢語サ変動詞の内部構成パタンは極めて限定的であると言える。これは、語形成の観点から見て興味深い結果であると同時に、教育的観点から言えば、典型的な内部構成パタンを明示的に指導することの有用性を示唆するものでもある。

3 点目は、漢語サ変動詞の語形成上の基本的特性についてである。漢語サ変動詞を構成しうる文字には、動詞的なもの、名詞的なもの、形容詞的なものなどが存在するわけであるが、実際には、一般の学習者が想像するように、これらが自由結合して漢語サ変動詞が作られるわけではない。並立関係の「V-V」と修飾関係の「V>V」で、全体の約 8 割が占められることに明らかなように、漢語サ変動詞の基本は動詞的要素を持つ 2 個の漢字同士の結合であると言える。教育学的には、このような漢語サ変動詞の語形成上の顕著な特性もまた学習者に適切に指導されることが望まれる。

最後に、4 点目は、客体関係の品詞的結合パタンについてである。一般に、客体関係は日本語では「(S+) O+V」、中国語では「(S+) V+O」と表現されるわけであるが、「V+N」は「N+V」の 2 倍近く出現しており、客体関係を持つ漢語サ変動詞の内部構成は日本語というより中国語の語順に準拠していると思われる。こうした中日対照という観点からの分析は、今後の研究課題となりうる。

５．まとめと教育的示唆

本研究は、コーパスを用い、多様な日本語変種間に共通して高頻度に使われる汎用的漢語サ変動詞の特定及びその内部構成の特徴の解明を目指して議論を進めてきた。分析により、先行研究ではっきり示されていなかった一般性の高い汎用的漢語サ変動詞にどのようなものがあり、その典型的な内部構成パタンが何であるかが明らかにされた。

本研究は、日本語語彙研究におけるコーパスデータの重要性を改めて示すとともに、コーパスから得られた知見の教育応用の可能性についても一定の示唆を行うものとなった。

今後は、分析対象とするデータの範囲をさらに拡張し、分析の枠組みの一層の精緻化を図るとともに、研究で得られた成果を具体的な教材や教授法の開発に生かす方向を考察していきたい。

文　献

石川慎一郎（2012）『ベーシックコーパス日本語』ひつじ書房
大石亨（2012）「テキストのジャンルとメタファー表現のコレスポンデンス分析―関係のメタファーを例に」『日本認知言語学会論文集』12、pp.52-64.
沖森卓也、三省堂編修所（2013）『三省堂常用漢字辞典』三省堂
国立国語研究所（1962）『現代雑誌九十種の用語用字』秀英出版
国立国語研究所（1973）『電子計算機による新聞の語彙調査』秀英出版
国立国語研究所（1983）『高校教科書の語彙調査』秀英出版
小林英樹（2004）『現代日本語の漢語動名詞の研究』ひつじ書房
新潮社（2008）『新潮日本語漢字辞典』新潮社

藤堂明保、松本昭、竹田晃他（2011）『漢字源改訂第五版』学習研究社
野村雅昭（1999）「サ変動詞の構造」森田良行教授古稀記念論文集刊行会（編）『日本語研究と日本語教育』、pp.1-23、明治書院
橋本直幸（2009）「BCCWJ を利用した日本語教育語彙リスト作成の試み」特定領域研究日本語コーパス平成 20 年度公開ワークショップ（研究成果報告会）予稿集、pp.183-190.
日向敏彦（1985）「漢語サ変動詞の構造」『上智大学国文学論集』18、pp.161-179.
松下達彦（2011）「日本語の学術共通語彙（アカデミック・ワード）の抽出と妥当性の検証」『2011 年度日本語教育学会春季大会予稿集』、pp.244–249.
水本篤（2010）「第 8 章主成分分析：データの情報を圧縮する」石川慎一郎、前田忠彦、山崎誠（2010）（編）『言語研究のための統計入門』、pp.193-217、くろしお書店
湯本昭南（1977）「あわせ名詞の意味記述をめぐって」『東京外国語大学論集』27、pp.31-46.

「―方」という形式にみられる「する」と「やる」の差異について

森川 結花（甲南大学 国際言語文化センター）†
小山 宣子（弘前大学 国際教育センター）
浜田　秀　（天理大学 文学部）

The Differences between “Suru” and “Yaru” When in the “ –kata” Form

Yuka Morikawa (Konan University, Institute for Language and Culture)
Nobuko Oyama (Hirosaki University International Education Center)
Shu Hamada (Faculty of Letters, Tenri University)

１．はじめに

「する」の類義語の「やる」は、日本語学習では初級段階で導入される基本語彙の一つである。現行の主な日本語初級教科書を見ると、「やる」が会話文や例文中に使われている頻度は、『初級日本語「げんき」Ⅰ』『同Ⅱ』で合計 7、『みんなの日本語初級Ⅱ』は 13、『できる日本語初級』は 18[1]で、学習段階の初期から学習者に「やる」を定着させようという意図が感じられる。しかし、「やる」と「する」をどんな場合にどう使い分けたらよいのか、どの教科書にも明確な説明記述はない。

「やる」には「野球をやる」「ゲームをやる」のような「する」と言い換え可能な用法も多数あるが、「英語をやる」「音楽をやる」「テレビでドラマをやっている」「若い頃に結核をやった」のように「やる」独特のヲ格名詞との組み合わせもある。このコロケーションを可能にする「やる」の本質とはどういうもので、それは「する」とはどう違うのか。その答えを見つけようとすることが、本研究の元々の出発点であった。

そして、「現代書き言葉均衡コーパス」BCCWJ中に「する」と「やる」がどのような違いを見せつつ存在しているかという観点からデータを観察していく中で、「名詞をする」「名詞をやる」が「名詞のし方」「名詞のやり方」という文型をとったときに顕著な差異が現れるということを確認した。管見の限りでは、「―方」という文型に現れる「する」「やる」を比較した先行研究はない。そこで、本研究が確認した「名詞のし方」「名詞のやり方」の差異について報告することにする。

２．先行研究

これまでに「する」と「やる」を比較対照した先行研究はいくつかあるが、その中でも特に大塚(2002)と金子(1985)の研究成果を踏まえておきたい。以下にその概略を述べておく。

２．１　大塚(2002)

大塚(2002)[2]では、ヲ格名詞と「する」と「やる」の組み合わせについて、(1)ヲ格名詞の性質（動作性か非動作性か）と、(2)「する」「やる」に機能動詞性がどの程度認められるか、或いは実質動詞的かという観点から、「する」「やる」の用法と性質の連続性が一つの表にまとめられた。それを次ページに表１として引用する。

† morikawa ＠ center.konan-u.ac.jp

[1] ただし、『できる日本語初級』では「やる」を「どうやってチケットを買いますか」のように、「どうやって」に限定している。『みんなの日本語初級Ⅱ』でも「どうやって」が４回使用されており、理由を問う「どうして」と意図的に対照してあるように見受けられる。

[2] 大塚には大塚（1999）を始めとして一連の「やる」「する」を対象とした記述的な研究があるが、本研究では特に大塚（2002）に絞って議論を進めることにする。

表１“「する」と「やる」の動詞機能の連続性とヲ格名詞の性質”（大塚(2002)より）

動詞の機能	ヲ格名詞の種類	ヲ格名詞の性質	する	やる	実質的意味	単独
機能動詞　強	「遊戯・スポーツ」	動作性	○	○	無	不可
	「趣味・習い事」	動作性	×	○	無	不可
	「映画・演劇・放送番組」	動作性	×	○	無	不可
	「様相・様子」	動作性	○	△	無	不可
	〃	非動作性	○	×	無	不可
	「生業」	動作性	○	○	有	不可
	「行事・集団活動・催し物」	動作性	○	○	有	不可
	「役職・役割・役柄」	動作性	○	○	有	不可
	「学問・科目」	非動作性	△	○	有	不可
	「着装物・付帯物」	非動作性	○	×	有	不可
	「映画・演劇・放送番組」	非動作性	×	○	有	不可
弱	「嗜好品」	非動作性	×	○	有	可
実質動詞　弱	「人・動物」	非動作性	×	○	有	可

大塚(2002)は、「する」は機能動詞性が強い動詞であるが、「やる」の方は名詞との組み合わせによって、機能動詞性の強いときもあればそれが弱いときもあり、場合によっては実質動詞として単独（ヲ格名詞なし）で用いられることも可能であると結論づけている。この結論は、本研究の BCCWJ 調査の結果からも支持されるものである。

２．２　金子(1985)

金子(1985)では、話し言葉文字化資料から「する」624 例、「やる」421 例を抽出して分析し、①話し言葉の中で「やる」は 95%が「ある行為を行う」[3]の意味で用いられる、②「やる」はヲ格名詞を省略して単独の形で使われやすい（金子（1985）の調査では 56%の「やる」がヲ格名詞を省略して用いられている）ということが述べられている。

この２点について、森川・小山(2013)は、名大会話コーパスのデータを調査し、金子(1985)を支持することのできる結果が得られたことを報告している。

以上のような大塚(2002)、金子(1985)の研究成果を踏まえつつ、「名詞の一方」という文型をとるときの「する」「やる」という観点から、両語の違いをコーパスデータの中で観察し、分析することにする。

３．「名詞のし方」「名詞のやり方」のデータと分析

３．１　調査の方法

本研究は、データの収集に「現代書き言葉均衡コーパス」BCCWJを利用した。検索システムとして中納言Ver.1.1.0 を使用し、検索対象をBCCWJの全データとして、次のａ） ｂ）の検索条件[4]を指定してデータを抽出した。

ａ）「名詞のし方」の検索条件
　品詞の大分類が「名詞」＋語彙素「の」＋語彙素「仕方」
ｂ）「名詞のやり方」の検索条件
　品詞の大分類が「名詞」＋語彙素「の」＋語彙素「遣る」＋語彙素「方」

[3] 「やる」には「ある行為を行う」という意味の他に、「やりもらい」と「物を移動させる」の意味がある。森田(1989)や森山(2012)では、原義が「物の移動」で、そこから「やりもらい」そして「行為」へと意味が発展したと記述されている。

[4] 単位の区切り方は文字列検索を参照した

その結果、収集できたデータの件数は表２の通りである。

表２　BCCWJ における「名詞のし方」／「名詞のやり方」の出現件数

	総出現件数	出現した種類の総数
a）名詞のし方	2,812	897
b）名詞のやり方[5]	1,050	570

３．２　データの分類

３．１で収集した「名詞をする」「名詞をやる」のデータについて、名詞が「し方」「やり方」の動詞成分「する」「やる」とどのような格関係を結んでいるかという点から、次のⅰ）～ⅲ）のカテゴリーに分類してみた。その結果を下の表３に示す。

ⅰ）名詞が「する」「やる」のヲ格（対象格）の関係にあるもの
　　例：勉強をする　→　勉強のし方　／　勉強をやる　→　勉強のやり方
ⅱ）名詞が「する」「やる」のガ格（主格）の関係にあるもの
　　例：わたしがする　→　わたしのし方　／　わたしがやる　→　わたしのやり方
ⅲ）名詞が「する」「やる」を修飾する関係（修飾格）にあるもの
　　例：昔した　→　昔のし方　／　昔やった　→　昔のやり方

表３　格関係による収集データの分類（値：出現件数）

	名詞のし方	名詞のやり方
ⅰ）ヲ格	2,783（99.0%）	440　（41.9%）
ⅱ）ガ格	3　（0.1%）	381　（36.3%）
ⅲ）修飾格	26　（0.9%）	229　（21.8%）
計	2,812（100%）	1,050　（100%）

また、この分類結果を円グラフに表すと図１、図２のようになる。

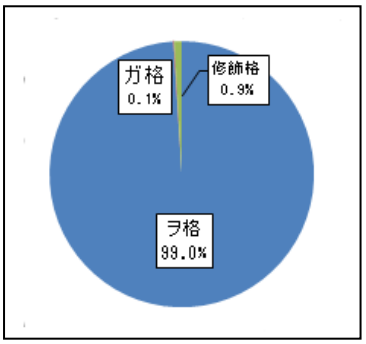


図１「名詞のし方」のデータの格関係による分類

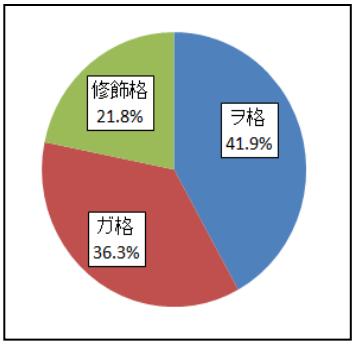


図２「名詞のやり方」のデータの格関係による分類

[5]実際に検索した結果収集できた 1060 件のデータから、「水のやり方」4 件、「肥料のやり方」3 件、「ミルクのやり方」1 件、「おっぱいのやり方」1 件、「目のやり方」1 件の計 10 件を対象外のものとして除き、総出現件数を 1,050 件とした。

３．３　ヲ格名詞と「し方」「やり方」

３．２で「名詞のし方」と「名詞のやり方」のデータを対照した時、まずはヲ格名詞の占める割合の極端な差が目につく。そこで、どのような名詞が「し方」「やり方」とヲ格の関係を持って結びついているのかを詳細に見ておきたい。

「名詞のし方」では全体で 882 種類の名詞が出現し、そのうち頻度５以上のものは 123 種類、「名詞のやり方」では全体で 264 種類、そのうち頻度５以上のものは９種類である。頻度５以上の名詞を下の表４にリストアップする。このリストの中で、「名詞のやり方」のリストに出てくるもので「名詞のし方」のリストにもあがっているものには網掛け強調文字で示す[6]。

表４　「名詞のし方」「名詞のやり方」に現れたヲ格名詞

出現件数	名詞のし方	名詞のやり方
90～99	対応 96	
80～89		
70～79		
60～69	質問 61	
50～59	対処 55、計算 54、処理 54	**仕事 55**
40～49	表現 49、利用 46、 **勉強 42**、説明 41、理解 41	
30～39	生活 35、筆算 34、**仕事 33,**	
20～29	アプローチ 29、設定 28、解釈 25、評価 24、活用 23、運転 20	
16～19	掃除 18、食事 18、手入れ 17、	
15	**運営**、学習、規定、行動、整理、把握、反応、	
14	化粧、検索、調理	
13	管理、梱包、紹介、存在	
12	運用、解決、認識、変化、保存	
11	挨拶、呼吸、作成、登場、表示、料理	
10	活動、**教育**、練習	経営
9	区別、結合、成功、注文、判断、分類、報道	
8	あいさつ、記載、削除、出品、展開、返事	**商売**
7	カット、介護、解凍、作業、指導、処分、操作、表記、剪定	政治
6	ＰＲ、暗算、応援、解除、回答、監督、記述、記入、区分、形成、告白、手当て、接続、選択、注意、陳列、定義、提供、提示、配置、配分、分析	**教育**、研究、
5	アップ、アピール、コピー、プレー、メイク、応対、開発、確認、関係、祈り、決定、交換、交渉、子育て、出現、処置、**商売**、進行、説得、**調査**、調整、投資、答弁、発音、発生、発表、批判、保管	**運営**、**調査**、**勉強**

４．結果の考察

３．２で見た通り、BCCWJ に存在するデータにおいて、「名詞のし方」は 99%の名詞がヲ格名詞由来であるのに対し、「名詞のやり方」はヲ格名詞由来の名詞の割合は半数を割

[6] 表４のリストに網掛け強調で示せなかった「経営」「政治」「研究」の「名詞のし方」の用例は、「経営」が頻度３、「政治」「研究」が頻度１であった。

っていて、「ヲ格名詞：ガ格名詞：修飾格の名詞」の割合が約４：４：２となる。

この言語事実は、大塚（2002）金子（1985）でも述べられていた「する」の機能動詞性の強さと、「やる」の実質動詞的な性質を如実に反映しているものと考えられる。すなわち、「し方」だけでは実質的な内容を表すことができないのでヲ格名詞が必須となるが、「やり方」の方は単独で実質的な内容が表せるのでヲ格名詞は必須ではないということである。

さらに、ヲ格名詞について、表４のリストから以下のことが考えられる。それは、「名詞のし方」のリストに入った名詞群は図３のように分類することができて、それぞれの名詞のグループで意味的な特徴が見いだされるのではないかということである。

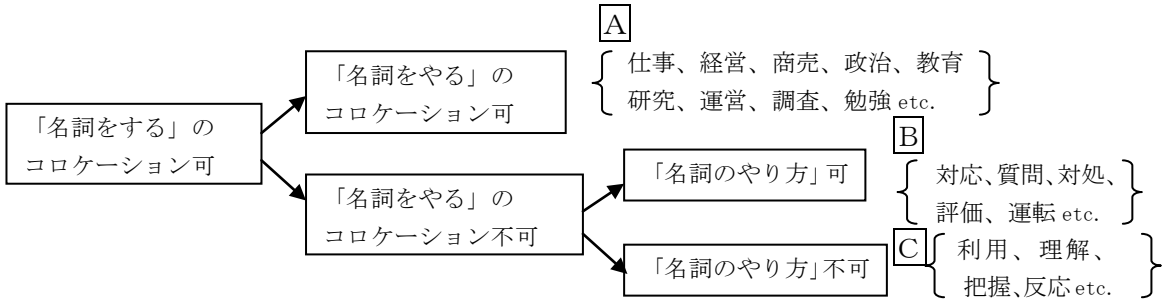


図３　「やる」「やり方」を基準にしたヲ格名詞の分類

この分類は内省に頼って行っており分類基準の実証性に乏しいが、A群は具体的な活動を表す名詞、C群は自発的／自動詞的変化あるいは自己内で完結する認知的な活動を表す名詞で、B群はAとCの中間的な性格を持つものではないかと推測できる。B群は、「する」とは結びつくが「やる」とは結びつくことはない。が、その「やり方」は云々することができるというものであり、興味深い存在である。

この「やる」「やり方」を基準とした分類法について、たとえば、

・勉強のし方が分からないんです。どうやって勉強したらいいか、（　し方　／　やり方　）を教えてください。

のような語彙を分類するためのテストの枠組みを作ったとして、それが有効に働く物であるかどうか、また、「やり方」を問うことが出来るか否かというが動詞のどのような意味特性によるものなのか、こういった点を明らかにすることは、本研究に残された今後の課題である。

５．通時的な観点から見る「し方」と「やり方」

最後に、「し方」と「やり方」を通時的な観点から見ておきたい。「やる」（遣る）は古くは「物や人などを自分から遠い方へ移動させる意。」（小島（2001））であった。堀口（1984）によると、「やる」の用法が広がって「する」の代わりに用いられるようになったのは近世に至ってからであるという。

それでは、「名詞のし方」「名詞のやり方」という文型についてはどうだろうか。

このことを調査するために JapanKnowledge を用いて、『新編日本古典文学全集』を検索した。その結果、「やり方」（遣り方）は用例そのものが存在しなかった。また、「名詞のし方」については、「～の仕方」で検索したところ、近世の作品に 11 例、存在することが分かった。具体的には、以下のような用例である。

・与九「また折を見て、訴訟のしかたもあろう。…」（[古典 81]『東海道中膝栗毛』p.40）

・「死骸おしやり、刀を拭ひ、しづゝゝしまうて立つたりし、武士の仕方のすゝどさよ」（古 典 75］『近松門左衛門集(2)』堀川波鼓 p.515）
・「段々功もゆき、歌学もせんと思ひ、この道に達せんとするときの仕方は、その時にはいかやうとも我が心にも合点もゆけば、学びやうあるべきことなり」（[古典 82］『辺世随想集』排蘆小船p.273）

このような「〜の仕方」がヲ格 3 例、ガ格 3 例、修飾格 5 例と、少数ずつながら三つのタイプの格関係の用例が見いだされた。

つまり、「やり方」という言い方は近代以降に一般的に使われるようになった“新しい”言葉で、それゆえに、現代でも「し方」よりも「やり方」の方がインパクト強く響くのかもしれない。

６．まとめ

以上、「名詞のし方」「名詞のやり方」についての BCCWJ データの観察を通して、この形式に基本動詞「する」と「やる」の本質的な違いが現れるということを確認した。「類義語 A と B の本質的な違いが顕著に表れる特定の形式は？」という観点からの類義語対照研究は従来、それほどされてこなかったのではないだろうか。コーパスデータを観察することから、ある語の本質的なイメージや際だった特徴が現れている「形式」を見いだすことができれば、日本語学習者にとってもそれは日本語の語彙を理解する上で非常にわかりやすい手がかりを提供できることになり、学習者のスムーズな語彙学習のために貢献することができるであろう。

文　献

大塚望(2002)「「する」と「やる」――非動作性名詞がヲ格に立つ場合――」『日本語科学 12』、pp.7-27、国立国語研究所
大塚望(2007)「「する」文の多機能性――文法的機能――」『日本語日本文学』第 17 号、pp.23-39、創価大学日本語日本文学会
（http://libir.soka.ac.jp/dspace/bitstream/10911/2935/1/KJ00004859986.pdf よりダウンロード可能）
大塚望(2012)「「する」文の格構造」『日本語日本文学』第 21 号、pp.33-48、創価大学日本語日本文学会(http://libir.soka.ac.jp/dspace/bitstream/10911/3244/1/nn21-033.pdf よりダウンロード可能)
金子比呂子(1985)「話し言葉における「する」と「やる」」『ＩＣＵ夏期日本語講座』、pp.105-126、国際基督教大学夏期日本語講座
小島聡子(2001)「やる（遣る）」(山口秋穂、秋本守英編『日本語文法大辞典』、pp.814-815、明治書院)
堀口和吉(1984)「動詞「やる」の一考察――「行(や)る」「演(や)る」の誕生――」『山邊道』第 28 号、天理大学国語国文学会、pp.31-49
森川結花、小山宣子(2013)「基本動詞「やる」を日本語教育の中でどう扱うべきか――「する」との比較を通して――」2012(平成 24 年度)　『第 10 回日本語教育学会研究集会予稿集』、pp.25-28
森田良行(1989)『基礎日本語辞典』角川書店
森山新編著(2012)『日本語多義語学習辞典　動詞編』アルク

関連 URL

JapanKnowledge　http://www.japanknowledge.com/top/freedisplay

文体から見た『今昔物語集』の語彙

―『日本語歴史コーパス 平安時代編』と比較して―

田中 牧郎　（国立国語研究所言語資源研究系）†

The Vocabulary of *Tales of Times Now Past* (*Konjaku Monogatarishū*): Comparison with *the Heian Period Series of the Corpus of Historical Japanese*

TANAKA Makiro (National Institute for Japanese Language and Linguistics)

１．はじめに

『日本語歴史コーパス』[1]の設計と構築においては、江戸時代までの口語性の強い資料群を優先してコーパス化のための基礎的研究を行っている。一方、文語性の強い資料も、日本語史研究においては重要であり、そのコーパス化も望まれる。現在、国立国語研究所コーパス開発センターで取り組んでいる日本語史資料のコーパス化は、①平安時代の和文資料（2013 年度完成公開予定）、②室町時代の狂言資料、③平安鎌倉時代の和漢混淆文資料、④江戸時代の洒落本資料の順で作業に着手している。

そのうち、①、②、④が口語性の強い資料群、③が文語性の強い資料群である。③については、『今昔物語集』『宇治拾遺物語』『方丈記』『徒然草』などのコーパス化を試行しているが、その中で量的にも大部で、コーパス化にあたって技術的な問題も多い、『今昔物語集』の研究に力を入れている。本稿は、試作中の『今昔物語集』のコーパスのデータの一部を用いて、その語彙の特徴を、平安時代の和文資料を対象としている『日本語歴史コーパス 平安時代編』（先行公開版）の語彙と比較しながら、文体的な観点で述べていきたい。

２．『今昔物語集』のコーパス化

コーパス化を行う『今昔物語集』（以下、『今昔』と略称）の本文は、小学館の「新編日本古典文学全集」の『今昔物語集 1～4』（馬淵和夫・国東文麿・稲垣泰一校注）により、コーパス構築のために小学館から国立国語研究所に提供された電子テキストを利用している。新編全集の『今昔』は、巻 1～10 の天竺震旦部は収録しておらず、巻 11～31 の本朝部のみを収録しており、コーパス化の対象もこの範囲になる。新編全集の底本は、巻 12･17･27･29 の 4 巻は『今昔』の最古の写本である鈴鹿本（現在は、京都大学図書館蔵、鎌倉時代書写）、巻 11・13･14･15･16･19･20･22･24 は実践女子大学本、巻 23･25･26･28･30･31 は東京大学国語研究室本である。コーパスの試作は、鈴鹿本現存巻である巻 12 から着手しており、本稿でも巻 12 のデータを用いる。

『今昔』に対する形態素解析は、まず「中古和文 UniDic」を用いて自動形態素解析を行

† mtanaka@ninjal.ac.jp

[1] http://www.ninjal.ac.jp/corpus_center/chj/

い、未知語となった語を辞書登録して、『今昔』を解析するための UniDic を整備しているところである。第一段階の作業として『今昔』巻 12 に対して形態素解析を実施し、その結果を目視で確認して、誤解析の修正と揺れの統一を手作業で行った。その作業での単語の認定基準は、小木曽・小椋・須永（2012）が、中古和文を対象として定めた、短単位規程に従うが、『今昔』に適用するにあたって一部変更したところがある。

３.『今昔』巻 12 の性格

『今昔』は、和漢混淆文の作品であるが、全 31 巻（うち 3 巻は欠巻のため、実際は 28 巻）のうち、はじめの方の巻は漢文訓読体としての性格が強く、後の巻に進むにつれて漢文訓読体としての性格は弱まり和文体としての性格が強まっていくという性質を持っている。巻 12 はその半ばよりも少し前に位置付いており、漢文訓読体がまだかなり強いが、説話によっては和文体の要素を含んでいるものも交じっている。巻 12 には、塔の建立、法会の起源、諸仏の霊験、法華経の霊験などの説話 40 話が集められている。『日本霊異記』を依拠資料とするものが 16 話あるほか、『法華験記』、『三宝絵』を依拠資料とするものがそれぞれ 10 話、6 話あり、これらは漢文系の資料で、『今昔』の説話も漢文訓読体である。一方、和文体の『古本説話集』または『宇治拾遺物語』と同文的であり、それらの共通母胎である散逸した『宇治大納言物語』を依拠資料とする説話が 3 話あり、『今昔』の説話もやや和文に近づいた文体になるが、『今昔』の後半の巻に多い和文的な説話に比べるとかなり硬い。このほか、依拠資料が未詳のものが 5 話ある。以上を総合すると、部分的に和文体に少し近い軟らかい文体も交じっているものの、巻 12 全体では漢文訓読体としての性格を持つ硬い文体であると言うことができる。

『今昔』巻 12 の語彙は、記号類と付属語を除くと、短単位で集計して、延べ語数 17,685 語、異なり語数 2,578 語である。比較対象に用いる、『日本語歴史コーパス 平安時代編』は、延べ語数 378,106 語、異なり語数 11,139　語である。なお、同語か異語かの判別は、『中納言』で取得できる UniDic による付加情報のうち「語彙素読み」「語彙素」「語彙素細分類」「品詞」「活用型」の五つのうちいずれか一つが違っていれば、異なる語と認定する基準を立てて行った。

４.『今昔』巻 12 の語彙と平安和文の語彙との比較―品詞と語種―

4.1　品詞

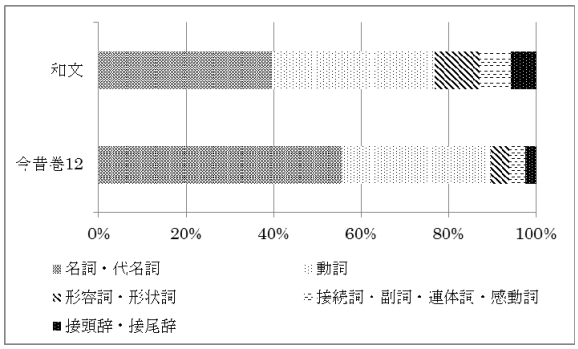


図１　今昔巻 12 と和文の品詞構成（延べ語数）

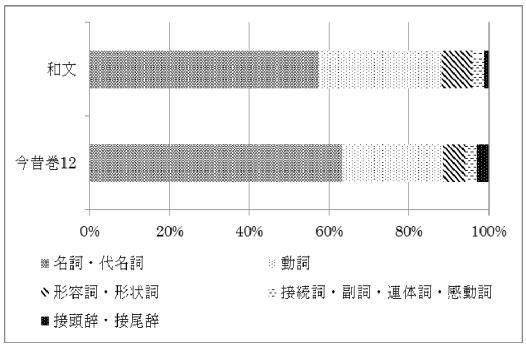


図２　今昔巻 12 と和文の品詞構成（異なり語数）

まず品詞構成を比較しよう。図１は延べ語数、図２は異なり語数の比較である。品詞の分類枠は、UniDic の大分類をもとに、「名詞・代名詞」「動詞」「形容詞・形状詞」「接続詞・副詞・連体詞・感動詞」「接頭辞・接尾辞」の 5 つにまとめた。

延べ語数（図１）では、『今昔』巻 12 は、和文に比較して、名詞・代名詞の比率がかなり高くなっており、形容詞・形状詞の比率は大幅に低くなっている。これは、一般に、漢文訓読文は主語や目的語など名詞に相当する語句を表示しやすく、和文は人物の状態や感情など形容詞・形状詞に相当する語句を表示しやすいという、それぞれの文体で書かれる文章の性質によるものだと考えられる。また、全体に占める分量は多くないが、図１では、接続詞・副詞・連体詞・感動詞、および接頭辞・接尾辞の比率も、和文で高く『今昔』巻 12 で低くなっている。これらの分類には、多様なものを含めてしまっているため、さらに細分類して、このような差が出る理由について、今後、よく分析していく必要がある。

そして、異なり語数（図２）では、『今昔』巻 12 と和文との違いは、延べ語数の場合ほどには大きくない。それでも、『今昔』巻 12 は和文に比べて、名詞・代名詞の比率が高く、動詞、形容詞・形状詞の比率が低い。延べ語数では差がなかった動詞にも、差が見られるようになっている。そして、接頭辞・接尾辞の比率は、延べ語数の場合と反対に、『今昔』巻 12 の方で高くなっている。このような異なり語数のデータに見られる、品詞構成上の特徴が何を意味するのかについても、各品詞の中に含まれる語彙の内訳を見て、よく分析していく必要があるだろう。

4.2　語種

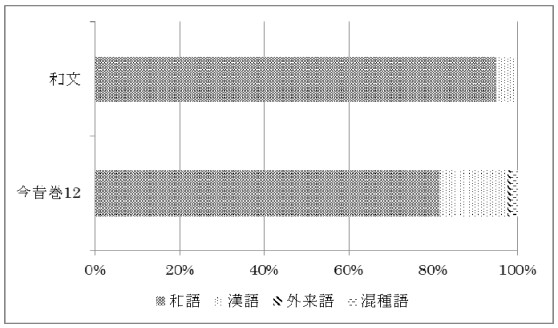


図３　今昔巻 12 と和文の語種構成（延べ語数）

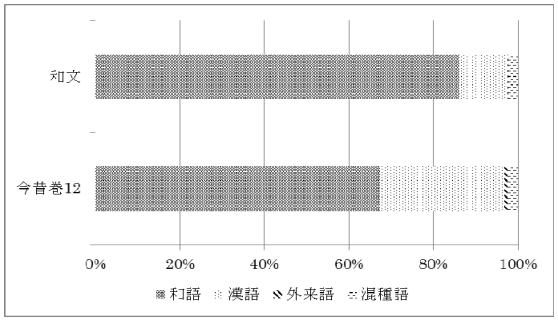


図４　今昔巻 12 と和文の語種構成（異なり語数）

次に語種分布を比較しよう。延べ語数（図３）でも異なり語数（図４）でも、『今昔』巻 12 は和文に比べて、漢語の比率が非常に高くなっており、その分和語の比率が大幅に低下していることが明らかである。『今昔』巻 12 の漢語の比率が、延べ語数よりも異なり語数の方でより高いのは、漢語は、繰り返し用いられる基本的な語には少なく、何度も用いられない周辺的な語には多いことを意味していよう。漢文訓読文としての性格が強い『今昔』巻 12 に漢語が多いことは当然のこととも言えるが、見方を変えれば、『今昔』巻 12 におけるその比率は、延べ語数で 16%程度、異なり語数でも 30%に達しておらず、語彙の大勢は和語が占めていることも確認できる。

前田（1984）は、漢文訓読資料や和漢混淆文資料についての諸調査をもとに、語種比率を表にまとめているが、そこで示された漢語比率と今回の『今昔』巻 12 の調査結果とを一つの表にして示すと表１のようになる。１が漢文訓読資料、２～７は和漢混交文資料である。

表１　平安時代から室町時代の漢文訓読文・和漢混淆文の語種比率

	文献資料	成立期	調査者	漢語比率（延べ語数）	漢語比率（異なり語数）
1	興福寺本大慈恩寺三蔵法師伝	11 世紀末	築島裕	43.4%	85.8%（混種語含む）
2	今昔物語集全巻	12 世紀初	有賀嘉寿子	—	41.9%（混種語含む）
3	今昔物語集巻 12	12 世紀初	田中牧郎	16.2%	29.4%
4	保元物語	13 世紀初	西田直敏	—	27.8%
5	平治物語	13 世紀後	西田直敏	—	34.5%
6	平家物語	13 世紀後	白井清子	23.1%	49.4%
7	史記抄	15 世紀初	柳田征司	—	58.5%

表１にあがる調査は、混種語を漢語に含めるか否かや、一語と認定する単位などが、それぞれの基準に基づいているので、そのままでは比較できない。表１の２番にあがる『今昔』の全巻を調査した有賀（1982）のデータと、本稿のデータの差の要因も、全巻と巻 12 の違いによる面と、調査単位の違いによる面の両方があるだろう。このように表１の諸データは単純に比較できるものではないが、どの文献資料においても、平安和文に比較して漢語の比率は非常に高い。今後は、表１にあるような代表的な漢文訓読資料・和漢混淆文資料に対して、UniDic の斉一な単位認定基準にしたがった語種調査を行っていき、この系統の文体での漢語比率の変化とその背景などを跡づけていくことが望まれよう。

５.『今昔』巻 12 と和文との共通語彙

5.1 頻度による語彙のレベル分け

４節では、品詞や語種の観点で語彙の全体的な比較をしたが、本節と次節では、個別の語について考察していこう。まず、『今昔』巻 12 と和文とで共通する語彙を見ていこう。ここでは、どちらにもよく用いられる共通の高頻度語彙を抽出する。

まず、『今昔』巻 12 の語彙を度数順に並べ、度数 10 以上の高頻度語彙（338 語）、度数５～９の中頻度語彙（330 語）、度数 4 以下の低頻度語彙（1909 語）の三つのレベルに分けた。『今昔』巻 12 において、延べ語数 17,685 の中で、高頻度語彙の累積度数（12,342）が占める比率（カバー率）はちょうど 70%となり、同じく高頻度語彙と中頻度語彙を合わせたもの（14,446）のカバー率は 82%となる。この 70%、82%という数字を基準にして、和文の語彙についても、同じように、高頻度語彙、中頻度語彙、低頻度語彙の三つのレベルに分けた。その概要をまとめたのが、表２である。

表２　頻度による語彙のレベル分け

	今昔巻 12		和文		カバー率
	度数区間	語数	度数区間	語数	
高頻度語彙	10～446	338	131～17,272	461	～70%
中頻度語彙	5～9	330	52～130	562	70～82%
低頻度語彙	1～4	1,909	1～51	10,116	82～100%
全体	1～446	2,578	1～17,272	11,139	—

5.2 共通の高頻度語彙

表２の「高頻度語彙」に配属された、『今昔』巻 12 の 338 語と、平安和文の 461 語を突き合わせると、共通するものは 168 語となる。この 168 語は、和文にも漢文訓読文にもよく使われる、平安時代の基本語彙であると見てよいものと思われる。先に分類した五つの品詞に分けて、次に掲げる。

●名詞・代名詞

尼、哀れ、家、如何、命、今、院〔名〕、院〔接尾〕、上、内、女、彼、方、形、守、川、木、君〔名〕、君〔代名〕、国、車、是、声、心、事、此れ、先、里、十、其、空、其れ、為、使い、常、罪、手、時、所、年、年頃、名、中、何、涙、日、後、母、日、一、人、一人、他、程、仏、前、真（マコト）、身、水、自ら、道、皆、昔、元、者、物、山、故、夢、世、夜、様（ヨウ）、由、童、我

●動詞

会う、合う、明ける、有る、言う、行く、出でる、居る、入る、入れる、失せる、得る、置く、覚える、思う、おわします、返る、掛ける、語る、聞く、聞こえる、来る、籠もる、然り、従う、知る、過ぎる、捨てる、住む、為る、絶える、立つ、奉る、立てる、給う〔四段〕、給う〔下二段〕、遣わす、付く、作る、問う、取る、泣く、嘆く、成す、成る、宣う、上る、乗る、始める、引く、臥す、経る、参る、見える、見る、召す、申す、詣でる、持つ、止む、遣る、許す、読む、寄る、渡る

●形容詞・形状詞

多い、恐ろしい、同じい、限り無い、高い、無い、長い、久しい、深い、やんごとない

●接続詞・副詞・連体詞・感動詞

斯く、必ず、更に、少し、唯〔副詞〕、唯〔形状詞〕、猶、先ず、又〔接続詞〕、又〔副詞〕、良く

●接頭辞・接尾辞

御、日（カ）、大（ダイ）、つ、殿（ドノ）、共（ドモ）、御（ミ）

『今昔』巻 12 の高頻度語の約半数は和文でも高頻度語彙で、基本的な語彙の半分程度は重なることが分かる。一方、『今昔』巻 12 の中頻度語彙 330 語のうち、和文でも中頻度語彙であるものは 65 語に止まる。中頻度語彙は、『今昔』巻 12 と和文とであまり重ならない。今後は、中頻度語彙、低頻度語彙も含めた、語彙の詳細な比較も必要になっていこう。

６．『今昔』巻 12 に特有の語彙

6.1 『今昔』巻 12 の高頻度語彙で和文では使われない語彙

次に、和文と比較した際に『今昔』巻 12 に特有の語彙となるものを抽出し、それがどのような性質のものか考えてみよう。

まず、『今昔』巻 12 では高頻度語彙でありながら、和文には全く使われていない語を抽出すると、次の 26 語となった。

●名詞・代名詞

安置、願主、奇異、ゲンシン（源信：人名）、金堂、聖人、織冠、書生、像、読誦、父母、法会、放生、維摩、霊験、礼拝

●動詞

与える、哀れむ、降りる、住する、在（マシマ）す

●形容詞・形状詞

専（もはら）

●接続詞・副詞・連体詞・感動詞

極めて、暫く、既に

●接頭辞・接尾辞

歳（サイ）

下線を付したのは、漢語または梵語あるいは混種語で、名詞・代名詞と接頭辞・接尾辞には、和語はない。動詞、形容詞・形状詞、接続詞・副詞・連体詞・感動詞には和語が多いが、それらはすべて、山田（1935）、築島（1965）、築島（1969）などによって、漢文訓読語であると指摘されているものである。『今昔』巻 12 における基本的な語彙でありながら和文では全く使われていない語彙は、人名以外では、漢語、梵語、混種語、漢文訓読語のいずれかであるということになる。

6.2 『今昔』巻 12 の高頻度語彙で和文では特に頻度が低い語彙

『今昔』巻 12 の高頻度語彙のうち、和文に使われていても、その使用頻度が極めて低いものは、6.1 で扱った語彙に準じる扱いをしてよい語であると思われる。表 2 において「低頻度語彙」とした和文の語彙は 10,116 語もある。量の少ない『今昔』巻 12 の語彙の分類に合わせて、カバー率 82%以上のところに位置付く語彙がすべてここに分類されているが、量の多い和文の語彙については、これらをさらに段階に分けて、さらに低頻度の語彙だけを取り出すこともできる。そこで、カバー率 97%以上という基準を立てると、度数区間が 1〜4 の特に頻度の低い語彙 6,570 語を抽出することができた。

『今昔』巻 12 で高頻度語彙でありながら、和文では特に頻度の低い語彙となるものを取り出すと、28 語になった。その 28 語の用例を、『中納言』を用いて、『日本語歴史コーパス平安時代編』で検索し、検索結果画面に表示される、地の文か発話部分か、和歌・詞書・手紙・序文などのその他の部分かといった、使用箇所を確認して、表 3 に整理した。表 3 の「語」の列に＊を付けた語は、漢語・梵語・混種語のいずれかのもので、＊の付いていないものは和語である。作品名は頭文字で示し、数字は使用件数で数字のないものは 1 件であることを意味している。

この表から、これらの語は、女性発話の部分に用いられることがないことが分かる。「性不明発話」としたところは、登場人物名が『中納言』の検索結果には表示されないものだが、該当箇所の作品本文を読んで確認をしていくと、女性であるものが 3 件見つかる。次の通りである。

> 大悲者には、他事（ことごと）も申さじ。あが姫君、大弐の北の方ならずは、当国の受領の北の方になしたてまつらむ。三条らも、随分にさかえて返し申しは仕うまつらむ。（源氏物語・玉鬘・三条という女房の発話）
>
> 「一乗の法なり」など人々も笑ふ事の筋なめり。（枕草子・「御方々、君達、上人など、御前に」・女房達の発話）
>
> 昨夜『物言はむ』とて来りしを、（落窪物語・巻一・あこぎの発話）

表３　『今昔』巻 12 の高頻度語で和文で特に頻度の低い語　和文における使用箇所

語	品詞	地の文	男性発話	女性発話	性不明発話	その他
希有*	名詞		1（源：僧都）			
持経*	名詞	2（源 2）				
釈迦*	名詞	3（源、枕 2）			1（源）	
舎利*	名詞					1（古：詞書）
修行*	名詞	2（伊、枕）	1（源：光源氏）			
大門*	名詞	1（枕）				
天皇*	名詞	4（伊、源 3）				
塔*	名詞	2（源、紫）				
童子*	名詞	2（落、枕）				
女人*	名詞		2（源：律師・僧都）			
法（ホウ）*	名詞	2（源 2）			1（枕）	
房（ボウ）*	名詞	1（大）				
諸々	名詞					1（古：序）
薬師*	名詞	4（源 2、枕 2）				
致す	動詞	1（伊）	1（源：律師）			
終わる	動詞	1（伊）	2（源 2：入道・律師）			
来たる	動詞				1（落）	1（源：和歌）
講ずる*	動詞	4（源 4）				
叫ぶ	動詞					1（古：詞書）
然り	動詞	1（土）			1（竹）	2（古：和歌、竹：手紙）
尊ぶ	動詞	1（源）				
説く	動詞	3(枕 2、紫 1)	1（源：光源氏）			
在（マシマ）す	動詞	1（大）	1（源：供人）			1（古：詞書）
速やか	形状詞				1（土）	
但し	副詞				2（竹）	
漸く	副詞	1（土）				2（古：詞書 2）
会（エ）*	接尾辞		1（源:使）			1（古:詞書）
者（シャ）*	接尾辞	2（伊、枕）	1（源：光源氏）		1(源)	

第一例の『源氏物語』玉鬘の三条という女房は田舎者として造型されており、引用箇所の発話部分には「者」のほか、「大悲」「当国」「受領」「随分」など漢語が多く用いられている。「受領」以外は、和文で使われることは極めて珍しいもので、この発話が異常な言葉遣いとして描かれているところである。第二例の『枕草子』での女房達の発話は法華経の引用文である。そして、第三例の『落窪物語』のあこぎの発話は、通常の発話部分であるが、これは「来たる」という動詞ではなく、「来（く）」（動詞）＋「たり」（助動詞）と認定すべきものが誤って動詞「来たる」と認定されたものであろう。こう見てくると、通常の女性の発話で用いられた例は皆無ということになる。男性の発話者も、僧都、阿闍梨、律師といった仏教関係者が目立つ。これらのことから、この 28 語はかなり硬い文体的価値を持った語であると見ることができよう。実際、漢語が多く、和語のうち、「諸々」「致す」「終わる」「来たる」「然り」「尊ぶ」「説く」「在す」「速やか」「但し」「漸く」などは、6.1 で見た諸語と同じく、山田（1935）、築島（1965）、築島（1969）などで漢文訓読語とされているものである。

「叫ぶ」だけは、先行研究で漢文訓読語とされていないものであるが、その和文での使

用例は次のものである。

> 法皇、西川におはしましたりける日、「猿、山の峡に叫ぶ」といふことを題にて歌よませ給うける
> わびしらに猿（ましら）な鳴きそあしひきの山のかひある今日にやはあらぬ
> （古今和歌集・雑体・一〇六七・詞書）

詞書に用いられる「猿、山の峡に叫ぶ」は、新編日本古典文学全集の注によると、漢詩の題詞に「猿叫峡」とあったものであり、「叫ぶ」はその訓読と考えられる。これに相当する内容は、和歌では「鳴く」が用いられている。したがって、「叫ぶ」も、漢文訓読語であると見ることができよう。

以上のように、『今昔』巻 12 の高頻度語彙のうちで、和文の語彙と比較した際の特有語は、漢語・梵語・混種語もしくは漢文訓読語のいずれかであると見ることができる。「叫ぶ」のように従来は見逃されていた漢文訓読語を特定していくこともできる。本稿では、比較対象とした和文で特に頻度の低い語彙を度数４以下の語としたが、度数５以上の語にも、和文での用例の出方が偏っているものもある。比較対象とする和文の語彙の範囲を拡大すれば、「叫ぶ」のような漢文訓読語を、さらに数多く発見していくこともできると考えられる。

７．おわりに

本稿で報告したのは『今昔』のうち巻 12 だけのデータに基づくものであったが、『今昔』の語彙と和文の語彙を比較することで、文体的観点から、語彙を分類していくことが様々に可能になることを示した。『今昔』の他の巻のデータを整備し、『日本語歴史コーパス 平安時代編』と比較していくことで、文体から見た語彙の分析を総合的に進めていくことができるようになるだろう。

付記

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）、及び、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、研究代表者：田中牧郎）による成果の一部です。

文献

有賀嘉寿子（1982）「今昔物語集の語彙」（『講座日本語の語彙 3 古代の語彙』明治書院）
小木曽智信・小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規程集」（科研費報告書）
築島裕（1965）『平安時代の漢文訓読語につきての研究』（東京大学出版会）
築島裕（1969）『平安時代語新論』（東京大学出版会）
前田富祺（1984）「語種構造の漸移相」（『日本語学』3-9、明治書院）
山田孝雄（1935）『漢文の訓読によりて伝へられたる語法』（宝文館）

『今昔物語集』のテキスト整形

冨士池優美　（国立国語研究所　コーパス開発センター）†
河瀬　彰宏　（国立国語研究所　コーパス開発センター）
野田　高広　（国立国語研究所　コーパス開発センター）
岩崎瑠莉恵　（国立国語研究所　コーパス開発センター）

The Text Formatting for *Konjaku-Monogatarishū*

Yumi Fujiike　(Center for Corpus Development, NINJAL)
Akihiro Kawase　(Center for Corpus Development, NINJAL)
Takahiro Noda　(Center for Corpus Development, NINJAL)
Rurie Iwasaki　(Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では「通時コーパスの設計」プロジェクトの中で、古典テキストに対して形態素解析を施す研究を進めている。これまでは平安時代を中心とする和文についてその成果を発表してきたが（小木曽ほか 2010 など)、漢文の要素が交じる和漢混淆文の古典テキストに対して形態素解析を施すためには、和文の場合とは異なる研究が必要になる。

発表者らは、新編日本古典文学全集『今昔物語集』一～四[1]（小学館、以下新編全集『今昔物語集』とする）に基づくコーパス化を行っている。新編全集『今昔物語集』には読解の便宜のために様々な校訂がなされているが、漢文の要素が交じっているため、そのまま形態素解析をすると様々な問題が生じる。この問題をテキスト整形によって解消すべく検討を行った。その結果、テキスト整形が必要な主な要素として、①返読文字、②助詞・助動詞等の省略表記、③捨て仮名、④欠字欠文・破損、⑤字種（片仮名・万葉仮名)、⑥踊り字・くの字点・同の字点があった。本発表では、新編全集『今昔物語集』を例に、各要素の問題点を指摘した上で、問題を解消するためのテキスト整形の事例を示すとともに、テキスト整形後の形態素解析結果を紹介する。

２．問題の所在

和漢混淆文特有の問題点が三つある。

まず、漢文の要素が交じっていることに起因するものである。形態素解析を施すにあたり、語順の転換や形態素の重複・不足があると、文字と形態素との対応を上から順に取れないことが問題になる。『今昔物語集』の場合、①返読文字が語順の転換と形態素の重複、②助詞・助動詞等の省略表記が形態素の不足、③捨て仮名が形態素の重複に該当する。これらについては、語順及び形態素の重複・不足を解消すべく、前もって整備する必要があ

† yfujiike@ninjal.ac.jp

1 新編全集には『今昔物語集』の巻 11 以降（本朝部）が収録されている。なお巻 18、21 は欠巻である。

る。

次に、新編全集『今昔物語集』の本文校訂の問題がある。新編全集では、④欠字欠文・破損に関して、底本に存している空格が記号で表記されるほか、底本に存していないが、校訂者が推定して置いた空格も異なる記号で表記されている。空格が記号で残されると、形態素解析を施す際、その前後に影響を及ぼすため、可能であれば空格に文字列を補うことが望ましい。また、新編全集『今昔物語集』の本文は、「底本を忠実に活字化することを期した」とあり、⑥踊り字・くの字点・同の字点といった繰り返し記号が用いられている。新編全集の中古和文資料の場合、繰り返し記号は用いられず、文字を繰り返して表記されるが、『今昔物語集』では繰り返し記号がそのまま表記された結果、文字とルビが対応しないという問題が生じた。この場合、中古和文資料と同様に、繰り返し記号を文字を繰り返す表記に置き換える必要がある。

さらに、形態素解析辞書に関する問題点がある。『今昔物語集』は漢字片仮名交じり文であり、万葉仮名の歌もあり、複数の⑤字種が含まれている。今回、形態素解析辞書としてUniDic[2]を用いており、片仮名の活用語尾や万葉仮名には対応していないため、字種を変更する必要が生じた。

３.『今昔物語集』のテキスト整形

３．１　概要

2 節で挙げた問題点について、前もってテキスト整形し問題を解消した上で、形態素解析を施すこととした。その際、テキスト整形前の状態を XML タグに記録し、元がどのようなものであったかの情報を取り出せるようにした。以下に、テキスト整形が必要であった要素ごとに、処理の詳細を述べる。

３．２　処理の詳細

①　返読文字

『今昔物語集』には、「不知ズ（シラズ)」「不知リ（シラザリ)」「不知（シラヌ)」のような表記がある。返読文字とはこれらの表記における「不」のような文字を指す。返読を含む文を形態素解析するとき、語順の転換や形態素の重複があり、文字と形態素との対応を上から順に取れないため、「不知ズ（シラズ)」「不知リ（シラザリ)」「不知（シラヌ)」をそれぞれ「知ズ」「知ザリ」「知ヌ」といった形式にする必要がある[3]。この形式の変更にあたり、語順を転換すると同時に、返読文字を形態素解析対象から除外し、返読文字に対応する文字列を挿入した。助詞・助動詞及び「なし」[4]を挿入する際は仮名に改めた。返読であるかどうかの認定は新編全集のルビに従い、山田ほか（1959-1963）に基づき事前に洗い出した返読文字を検索して処理を行った。その際、「所謂」等、形態素解析辞書に 1 単位

---

2 小木曽ほか（2010)、小椋・須永（2012）参照。

3 『今昔物語集』における返読文字の詳細については、冨士池・田中（2012）を参照。

4 形容詞又は形容詞の一部。具体例は 4 節表 2 を参照。

として登録することが妥当と思われる語については、語順変更を要しないため、対象外とした。また、目録・説話タイトル・漢詩文の引用及びルビが付されていない部分も対象外とした。

元の返読文字のほか､返読の通し番号､返読タイプといった情報をXMLタグに記録した。返読タイプとは返読文字に該当する語の仮名表記に着目した分類で、次の4種類がある。

Ａ：返読文字＋□＋語の全体仮名表記
　例）不知ズ（シラズ）　令聞シム（キカシム）　不泣給ヒソ（ナキタマヒソ）
Ｂ：返読文字＋□＋語の一部の仮名表記
　例）不知リ（シラザリ）　令聞ム（キカシム）　難有タキ（アリガタキ）
Ｃ：返読文字＋□
　例）不知（シラヌ）　令聞（キカシム）　于今（イマニ）
Ｄ：上記Ａ～Ｃに当てはまらない変則的なもの
　例）不ジ見（ミセジ）　不被ラ止知ニケリ（シラデヤミニケリ）

返読文字は助動詞・助詞・接尾辞等（以下「助動詞等」とする）と意味が対応する漢文の助字に当たるものであり、□には主に動詞が入る。返読タイプとは助動詞等を助字を用いてどのように表記するかを示したものと言える。タイプＡは助動詞等を漢字で表しつつ仮名で併記したもの、タイプＢは助動詞等を漢字で表したものに送り仮名を付したもの、タイプＣは助動詞等を漢字のみで表したものである。

上に示した例は以下のように処理される。テキスト整形の面から見たとき、タイプＡでは返読文字に対応する文字列の挿入がなく、タイプＢは返読文字で表される語の一部が挿入され、タイプＣは返読文字で表される語全体が挿入されることになる。タイプＤは変則的な表記に合わせて、ルビに合うように対応する語を適宜挿入した。処理対象返読文字数は、タイプＡが最も多く3464箇所、タイプＢは2964箇所、タイプＣは2796箇所、タイプＤは30箇所であった。

Ａ：返読文字のみ除外される
　例）知ズ（シラズ）　聞シム（キカシム）　泣給ヒソ（ナキタマヒソ）
Ｂ：返読文字除外、対応する語の一部挿入
　例）知**ザ**リ（シラザリ）　聞**シ**ム（キカシム）　有**ガ**タキ（アリガタキ）
Ｃ：返読文字除外、対応する語の挿入
　例）知**ヌ**（シラヌ）　聞**シム**（キカシム）　今**ニ**（イマニ）
Ｄ：返読文字除外、ルビに合うように対応する語を挿入
　例）見**ジ**（ミセジ）　**知ラデ止**ニケリ（シラデヤミニケリ）
（**太文字**：挿入箇所）

また、「余日」「余歳」についてはルビに合わせ「日余」「歳余」に語順を変更した。

二十余日（はつかあまり）　→　二十日余　　二十余歳（はたちあまり）　→　二十歳余

②　助詞・助動詞等の省略表記

『今昔物語集』では、「今昔」を「いまはむかし」と読むように、助詞・助動詞等の表記が省略されることが多い。表記されていない文字は形態素解析対象とならないため、このような形態素の不足を補う必要がある。これを補読と呼ぶ。補読処理はルビに基づき行った。助詞・助動詞のほか、「二」に対して「フタリ」のようなルビがある場合に「人」を補う、漢語サ変動詞の「ス」に該当する箇所が省略されている場合に「ス」を補うといった処理もしている。なお、活用語尾の省略については形態素解析辞書 UniDic で対応可能であるため、補読処理の対象外とした。補読処理によって挿入された文字列であることは XML タグに記録した。補読処理の対象は 7334 箇所であった。以下に例を示す。

今昔（いまはむかし）　→　今**ハ**昔　　此二（このふたり）　→　此**ノ**二**人**　　況（いはむや）　→　況**ムヤ**　（**太文字**：補読箇所）

③　捨て仮名

『今昔物語集』には、他の読み方をされる可能性のある漢字の読みの一部を補うことで漢字の読みを明確にする捨て仮名が多用されている。「汝ヂ」の「ヂ」のように最後の一字が主に捨て仮名となるが、「候フウ」のように「さぶらう」の「ぶ」を表記したと思われるものもある。頭注の記述を参考に、捨て仮名を洗い出し[5]、捨て仮名部分は形態素解析対象から除外した。捨て仮名として除外された文字列であること、元の表記を XML タグに記録した。処理対象となった捨て仮名は 1312 箇所であった。以下に捨て仮名の処理例を示す。

汝**ヂ**　→　汝　　此**カク**　→　此　　候**フ**ウ　→　候ウ
努々**メ**　→　努々　　今夜**ヒ**　→　今夜　　（**太文字**：捨て仮名）

表 1 に各巻の捨て仮名出現状況を示した。頻度 1～2 は省略した。「此カク」のような全訓捨て仮名については、「全訓」列に○を付した。捨て仮名は表記にして 240 種類超あるが、およそ半数は頻度 1 となっており、臨時的に付されることが多い様子がうかがえる。

④　欠字欠文・破損

2 節で示したように、新編全集『今昔物語集』では空格が記号で示されている。これを、頭注の記述に基づき[6]、3 種類に分けてテキストの整形を行った。

---

[5] 頭注の記述は一様ではないため、捨て仮名に関する頭注を洗い出し、誤写の可能性があるもの、衍字か捨て仮名かが明確でないもの等については、テキスト処理対象から除外した。

[6] 頭注の記述は一様ではないため、欠字欠文・破損に関する頭注を洗い出し、3 種類の分類、「推定文字列の候補が一つ示されている」とするかどうか、「何の表記が保留されたのか」を判定する基準を定めた上で、欠字欠文・破損に関するテキスト整形の作業を行った。

表１　各巻の捨て仮名出現状況

種類	読み	全訓	11	12	13	14	15	16	17	19	20	22	23	24	25	26	27	28	29	30	31	総計
汝ヂ	ナンジ		24	20	34	44	11	18	51	19	16	1	3	7	3	1	3	0	0	1	5	261
形チ	カタチ		7	1	0	2	1	11	26	12	4	4	4	9	2	4	5	1	5	12	9	119
夜ル	ヨル		3	2	12	4	1	5	3	5	2	0	0	2	3	0	2	2	4	4	2	56
何ニ	ナニ		0	0	1	1	0	3	2	8	1	1	0	6	0	2	6	4	7	4	4	50
智リ	サトリ		13	3	2	0	5	1	3	5	7	0	0	0	0	0	0	0	3	0	3	45
昔シ	ムカシ		8	5	1	3	3	1	1	5	1	0	2	1	0	0	2	0	0	0	0	33
公ケ	オオヤケ		0	0	0	0	0	4	4	9	3	1	2	0	4	0	0	0	1	0	2	30
験シ	シルシ		1	1	1	6	0	4	0	1	8	0	0	1	0	0	0	1	0	1	1	26
者ノ	モノ		0	1	0	1	0	2	1	3	1	0	0	0	4	0	3	4	1	0	1	22
奉ツ	タテマツル		1	2	3	2	1	5	4	0	0	0	0	0	0	0	0	1	0	0	0	19
許リ	バカリ		0	0	1	5	0	0	2	3	1	1	1	2	0	1	0	1	0	0	0	18
後チ	ノチ		0	1	1	0	1	2	3	3	0	0	0	1	0	0	0	3	0	0	1	16
昼ル	ヒル		0	1	1	2	1	0	0	5	0	0	0	0	2	1	1	0	1	0	1	16
上へ	ウエ		0	0	0	0	0	1	1	12	0	0	0	0	0	0	0	0	0	0	1	15
万ヅ	ヨロズ		0	0	0	1	0	0	1	0	1	1	0	0	0	1	1	1	4	2	2	15
為メ	タメ		0	1	2	1	0	0	7	2	0	0	0	0	0	0	0	0	0	0	0	13
現ハ	アラワ		0	0	0	0	3	1	1	2	1	0	0	0	1	1	1	1	0	0	1	13
間ダ	アイダ		0	1	1	1	0	0	1	4	2	0	0	0	0	0	1	1	0	0	0	12
抑モ	ソモソモ		0	1	0	0	0	1	0	0	1	0	0	2	1	4	0	1	1	0	0	12
皆ナ	ミナ		0	0	1	1	1	0	3	0	0	1	0	0	1	0	1	1	1	0	0	11
古へ	イニシエ		0	0	0	0	0	0	2	1	0	0	0	2	0	1	2	2	0	1	0	11
程ド	ホド		0	0	1	0	0	1	1	2	1	0	0	0	0	0	0	2	1	1	1	11
長ケ	タケ		0	0	0	1	0	1	0	1	1	0	0	3	1	0	0	1	0	0	1	10
各ノ	オノオノ		0	0	1	0	0	0	1	1	1	1	0	0	2	0	0	1	0	0	1	9
人ト	ヒト		0	1	1	0	0	1	2	0	1	0	0	0	1	0	0	0	0	0	2	9
適マ	タマタマ		0	1	0	2	0	3	0	0	1	0	0	0	0	0	0	2	0	0	0	9
夢メ	ユメ		0	1	1	3	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	9
志シ	ココロザシ		3	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	8
事ト	コト		0	0	0	3	0	0	2	1	0	0	0	0	0	0	1	0	1	0	0	8
前へ	マエ		0	0	1	0	0	3	2	2	0	0	0	0	0	0	0	0	0	0	0	8
君ミ	キミ・ギミ		0	1	0	3	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	7
罪ミ	ツミ		0	0	1	1	0	0	2	1	1	0	0	0	0	0	0	0	1	0	0	7
其コ	ソコ		0	0	0	0	0	1	0	0	0	0	1	1	0	2	0	1	0	1	0	7
怠タ	オコタル		0	0	2	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	7
男コ	オトコ		0	0	1	0	0	3	0	1	2	0	0	0	0	0	0	0	0	0	0	7
物ノ	モノ		0	0	0	1	0	3	0	0	0	0	0	0	0	0	1	0	0	1	1	7
去キ	ノク		0	0	0	0	0	1	0	2	1	0	0	0	0	0	2	0	0	0	0	6
後ロ	ウシロ		0	0	0	2	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	6
尚ヲ	ナオ		0	0	1	0	0	2	1	0	0	0	0	0	0	0	0	2	0	0	0	6
数タ	アマタ		0	0	0	0	0	1	2	0	1	0	0	0	0	0	0	1	0	0	1	6
注シ	シルシ		2	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	6
度ビ	タビ		2	2	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	6
仏ケ	ホトケ		0	0	0	0	0	2	0	4	0	0	0	0	0	0	0	0	0	0	0	6
豈ニ	アニ		1	0	2	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	6
下ダ	クダル		0	0	0	1	2	0	0	0	0	0	0	0	1	0	1	0	0	0	0	5
共モ	トモ・ドモ		0	1	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	1	5
今夜ヒ	コヨイ		0	0	0	2	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	5
時キ	トキ		0	0	0	0	1	0	2	1	1	0	0	0	0	0	0	0	0	0	0	5
前キ	サキ		0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	1	0	0	5
渡タ	ワタル		0	0	0	1	0	1	0	2	0	1	0	0	0	0	0	0	0	0	0	5
年シ	トシ		0	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	5
偏へ	ヒトエ		0	2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	5
赤タ	マタ		0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	5
為スル	スル	○	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	4
給マ	タマワル		0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	4
経へ	ヘル	○	0	0	0	1	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	4
高カ	タカイ		0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	4
最モ	モットモ		0	0	0	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	4
聖リ	ヒジリ		0	1	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	4
傍ラ	カタワラ		0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4
忝ナ	カタジケナイ		0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	4
為セ	スル	○	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	3
員ズ	カズ		0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
肝モ	キモ		0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	3
強ヨ	ツヨイ		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	3
空ラ	ソラ		0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
此カク	カク	○	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	3
取ト	トル	○	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	3
情ケ	ナサケ		0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	3
心ロ	ココロ		0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	3
祖ヤ	オヤ		0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3
男コ	オノコ		0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	3
殿ノ	トノ・ドノ		0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	3
童ハ	ワラワ		0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	3
如ト	ゴトシ		0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	3
非ラ	アラズ		0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	3
辺リ	ホトリ		0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	3
頻度2　計	37種類																					74
頻度1　計	129種類																					129
総計			68	64	88	122	42	119	192	166	83	16	20	43	36	30	45	53	37	33	55	1312

（１）破損による欠字

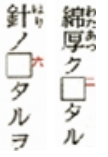
頭注に「破損による欠字」とあるものについては、推定文字列の候補が一つ示されている場合のみ、その読みを検討後、該当文字列を補った。推定文字列が特定できない場合は空格を示す記号を挿入した。

(1)破損

（２）意識的欠字

ａ．漢字表記保留

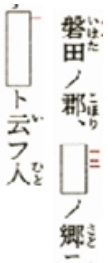
頭注に「漢字表記を期した意識的欠字」のようにあるものについては、推定文字列の候補が一つ示されている場合のみ、その読みを検討後、該当文字列を補った。推定文字列が特定できない場合は空格を示す記号を挿入した。



(2)a. 漢字表記保留

ｂ．具体表記保留

頭注に「(地名・僧名…）の明記を期しての欠字」のように具体表記を保留した欠字であることが示されているものについては、推定文字列が一意に決まる場合でも、欠字の補入は行わず、空格を示す記号を挿入した。



(2)b. 具体表記保留

上記 3 種類とも、空格を示す記号を挿入する際には、1 字は“␣”、2 字以上は“⇦␣⇨”というように文字列長によって空格を示す記号を分けた。3 種類の別は XML タグに記録した。新編全集『今昔物語集』では底本に存する空格と校訂者の判断により補った空格とで記号を分けていたが、この違いについても XML タグに記録した。また、（２）ｂ．具体表記保留については、何の表記が保留されたのか（UniDic の品詞相当の情報[7]）をあわせて XML タグに記録した。

以下に、前ページに示した欠字欠文・破損の処理例を示す。

（１）		鳩**ゾ**現ニ来テ	聴聞⇦␣⇨疑ヒ	
（２）	ａ．	針ノ**サビ**タルヲ	綿厚ク␣タル	
	ｂ．	⇦␣⇨⇦␣⇨ト云フ人	磐田ノ郡、⇦␣⇨ノ郷ニ	（**太文字**：補入箇所）

⑤　字種（片仮名・万葉仮名）

『今昔物語集』は漢字片仮名交じり文である。この片仮名については、全て平仮名に置き換えた。

『今昔物語集』中に出てくる万葉仮名の歌についても、全て平仮名に置き換えた。万葉仮名から平仮名への置き換えについてはルビを参照し、元の文字が万葉仮名であったことと、元の万葉仮名がどの文字であったのかを確認できるよう XML タグに記録した。

---

7 具体表記を保留された内容については、頭注の記述に基づき「人名-一般」「人名-姓」「人名-名」「地名」「数詞」「一般」の 6 種類に分類した。例えば「国司の姓名の明記を期した意識的欠字」とある場合、「人名-姓」「人名-名」に該当する二つのタグを補入した。「一般」は、寺院を含む建物名・官職名・方位・食物名など、UniDic において概ね「名詞-普通名詞-一般」が適用されるもの、及び、具体表記を保留された内容に複数の品詞の可能性がある（例えば「□ノ比（ころ）」の場合、年号＋年代、年号のみ、天皇名、干支等、品詞が特定できない）ものである。

平仮名に置き換えた万葉仮名は 94 箇所であった。例えば、右に示した歌は以下のように置き換えられる。

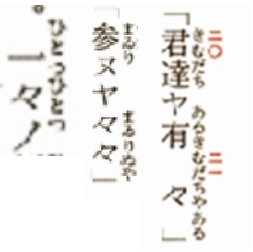
みづはさす　やそぢあまりの　おひのなみ　くらげのほねに　あふぞうれしき

⑥　踊り字・くの字点・同の字点

2 節に示したように、新編全集『今昔物語集』には踊り字・くの字点・同の字点といった繰り返し記号が用いられている。このうち踊り字・くの字点は、原則としてすべて文字を繰り返す表記に改めた。また、同の字点は、複数の文字を繰り返しているもののうち読みが確定しているもの、文節を越えるもの、動詞の終止形が二つ重なる形式、動詞・形容詞の連用形が二つ重なる形式のいずれかに当てはまる場合は文字を繰り返す表記に改めた。元が踊り字・くの字点・同の字点であったこと、元の表記を XML タグに記録した。処理対象となった繰り返し記号は 1495 箇所であった。以下に踊り字・くの字点・同の字点の処理例を示す。

ツヽ　→　ツ**ツ**	給ハゞ　→　給ハ**バ**	ヲイヽヽ　→　ヲイ**ヲイ**
ホロ/＼　→　ホロ**ホロ**	サメ/〃＼　→　サメ**ザメ**	今ヤ/＼　→　今ヤ**今ヤ**
返々ス　→　返**返**ス	穴怖シ々々　→　穴怖シ**穴怖**	

『今昔物語集』の同の字点は繰り返す文字数が一定ではないが、ルビを参照して、原則として文字数が同じになるように、文字を挿入した。同の字点で繰り返す文字のルビが短単位[8]を越えている場合や同の字点と挿入する文字の字数が一致しない場合は、補読処理をあわせて行った。



一々（ひとつひとつ）　→　一ツ一ツ
参ヌヤ々々（まゐりぬや）　→　参ヌヤ参ヌヤ
君達ヤ有々（きむだちやある）　→　君達ヤ有君達ヤ有

なお、読みが確定できないもの（ルビがなく、どこから繰り返しているかが不明確なもの）については、繰り返し記号を処理せずに残した。

３．３　その他

会話中の心中思惟については紙面で「」（「」の太字）と表されていたが、これを＜＞に置き換えた。以下の 1 箇所である。

＜悩スラム所ノ悪鬼ヲ揮へ＞

8 短単位については小椋・須永（2012）参照。

最後に、これは和漢混淆文だけではなく和文と共通の問題点となるが、漢字の字体に関する問題がある。新編全集『今昔物語集』では、原漢字はすべて漢字で転写されているが、漢字のうち、常用漢字字体のあてられるものはおおむねその字体としたほか、「常用漢字に該当するか否かの判定に迷う」場合は「底本の字体に従った」とある。これに対し、『今昔物語集』のコーパス化にあたっては JISX0213 に依拠して文字処理を行っており、JISX0213 外となる文字については別字代用を行うといった前処理を上記テキスト整形前に行っている。漢字の処理方針の詳細については、須永・堤（2012）を参照されたい。

４．形態素解析例

『今昔物語集』巻第十二「遠江国丹生茅上起塔語第二」をテキスト整形後、形態素解析し、人手修正を加えたものの一部を表 2 に示した。キー（本文）とルビは新編全集『今昔物語集』の情報である。これを短単位に分割し、代表形・代表表記に当たる語彙素読み・語彙素、品詞、活用型、活用形等を付与している。表 2 の「テキスト整形」列には 3.2 節①～⑥のうちどの要素に当たるのかを、「紙面」列には新編全集『今昔物語集』における表記を示した。

テキスト整形の結果、漢字片仮名交じり文だった本文は漢字平仮名交じり文となり、返読文字は語順を変換した形、助詞・助動詞等の表記省略箇所は挿入した文字が形態素解析対象となっている。空格は記号に置き換え、空格がどのような欠字欠文・破損であるかについては、XML タグの情報に基づき、人手で情報を付与した。表 2 の空格「⇦␣⇨ノ郷」は具体表記保留によるものである。紙面列の頭注からもわかるように、ここは地名の具体表記を保留した空格であるため、品詞情報として「意識的欠字（地名）」を付与した。「可産（さんすべ）キ」の場合、「可」が返読文字であるため形態素解析対象から除外し、「産す」の「す」がないため「す」を挿入し、返読文字「可」にあたる「べし」の「べ」を挿入した結果、「産すべき」が形態素解析対象の文字列となっている。

５．おわりに

本発表では、和漢混淆文の形態素解析における問題点を、新編全集『今昔物語集』のテキストを例に指摘した。その問題点に対して、テキストをどのように整形することで解決してきたかを具体的に示すとともに、整形後のテキストを形態素解析した結果、どのような情報を付与するのかをあわせて紹介した。各処理の説明で触れたが、テキスト整形の理由と整形前の状態は XML タグに記録し、必要に応じて参照できるようになっている。

語順の転換や形態素の重複・不足は和漢混淆文をはじめとした漢文の要素が交じる資料に形態素解析を施す際の大きな課題である。新編全集『今昔物語集』は漢文の要素が交じる資料にある様々な問題を含んだテキストであった。今回の検討によって、漢文の要素が交じる資料を形態素解析する際の問題点のいくつかに対して、解決がつく見込みが立ったものと考える。

表 2　形態素解析例

キー	ルビ	語彙素読み	語彙素	出現発音形	品詞	解析活用型	活用形	語義	テキスト整形	紙面
遠江国丹生茅上起塔語第二	とをたうみのくにのにふのちがみたふをたつることだいに				題					
					空白					
今	いま	イマ	今	イマ	名詞-普通名詞-副詞可能					
は	は	ハ	は	ワ	助詞-係助詞				助詞・助動詞等の省略表記	
昔	むかし	ムカシ	昔	ムカシ	名詞-普通名詞-副詞可能					
、			、		補助記号-読点					
聖武	しやうむ	ショウム	ショウム	ショーム	名詞-固有名詞-人名-一般					
天皇	てんわう	テンノウ	天皇	テンノー	名詞-普通名詞-一般					
の		ノ	の	ノ	助詞-格助詞					
御代	みよ	ミヨ	御代	ミヨ	名詞-普通名詞-一般					
に		ニ	に	ニ	助詞-格助詞					
、			、		補助記号-読点					
遠江	とをたうみ	トオトウミ	トオトウミ	トートウミ	名詞-固有名詞-地名-一般					
の		ノ	の	ノ	助詞-格助詞					
国	くに	クニ	国	クニ	名詞-普通名詞-一般					
、			、		補助記号-読点					
磐田	いはた	イワタ	イワタ	イワタ	名詞-固有名詞-地名-一般					
の		ノ	の	ノ	助詞-格助詞					
郡	こほり	コオリ	郡	コーリ	名詞-普通名詞-一般					
、			、		補助記号-読点					
⇦⎵⇨					意識的欠字（地名）				欠字欠文・破損：2字以上の意識的欠字（地名）	
の		ノ	の	ノ	助詞-格助詞					
郷	さと	サト	里	サト	名詞-普通名詞-一般					
に		ニ	に	ニ	助詞-格助詞					
、			、		補助記号-読点					
丹生	にふ	ニュウ	ニュウ	ニュー	名詞-固有名詞-人名-姓					
の		ノ	の	ノ	助詞-格助詞					
直	あたひ	アタイ	直	アタイ	名詞-普通名詞-一般					
茅上	ちがみ	チガミ	チガミ	チガミ	名詞-固有名詞-人名-名					
と		ト	と	ト	助詞-格助詞					
云ふ	い	イウ	言う	イウ	動詞-一般	文語四段-ハ行	連体形-一般			
人	ひと	ヒト	人	ヒト	名詞-普通名詞-一般					
有	あり	アル	有る	アリ	動詞-非自立可能	文語ラ行変格	連用形-一般			
けり		ケリ	けり	ケリ	助動詞	文語助動詞-ケリ	終止形-一般			
。			。		補助記号-句点					
（中略）										
父	ちち	チチ	父	チチ	名詞-普通名詞-一般					
有	あり	アル	有る	アリ	動詞-非自立可能	文語ラ行変格	連用形-一般			
て		テ	て	テ	助詞-接続助詞					
母	はは	ハハ	母	ハハ	名詞-普通名詞-一般					
に		ニ	に	ニ	助詞-格助詞					
云く	いは	イウ	言う	イワク	動詞-一般	文語四段-ハ行	ク語法			
、			、		補助記号-読点					
「			「		補助記号-括弧開					
汝	なむ	ナンジ	汝	ナンジ	代名詞				捨て仮名：汝ヂ→汝	
、			、		補助記号-読点					
齢	よは	ヨワイ	齢	ヨワイ	名詞-普通名詞-一般				捨て仮名：齢ヒ→齢	
産す	さんす	サンスル	産する	サンス	動詞-一般	文語サ行変格	終止形-一般		返読文字、助詞・助動詞等の省略表記：可産キ→産す	
べき		ベシ	べし	ベキ	助動詞	文語助動詞-ベシ	連体形-一般		べき	
齢	よはひ	ヨワイ	齢	ヨワイ	名詞-普通名詞-一般					
に		ナリ	なり	ニ	助動詞	文語助動詞-ナリ-断定	連用形-ニ	断定		
非	あら	アル	有る	アラ	動詞-非自立可能	文語ラ行変格	未然形-一般			
ず		ズ	ず	ズ	助動詞	文語助動詞-ズ	連用形-一般			
し		スル	為る	シ	動詞-非自立可能	文語サ行変格	連用形-一般			
て		テ	て	テ	助詞-接続助詞					
産せ	さん	サンスル	産する	サンセ	動詞-一般	文語サ行変格	未然形-一般			
り		リ	り	リ	助動詞	文語助動詞-リ	終止形-一般			
。			。		補助記号-句点					
（中略）										
其	そ	ソ	其	ソ	代名詞					
の		ノ	の	ノ	助詞-格助詞					
塔	たふ	トウ	塔	トー	名詞-普通名詞-一般					
今	いま	イマ	今	イマ	名詞-普通名詞-副詞可能				返読文字：于今→今に	
に		ニ	に	ニ	助詞-格助詞					
有り	あ	アル	有る	アリ	動詞-非自立可能	文語ラ行変格	終止形-一般			
。			。		補助記号-句点					
磐田	いはた	イワタ	イワタ	イワタ	名詞-固有名詞-地名-一般					
寺	でら	テラ	寺	デラ	名詞-普通名詞-一般					
の		ノ	の	ノ	助詞-格助詞					
内	うち	ウチ	内	ウチ	名詞-普通名詞-副詞可能					
の		ノ	の	ノ	助詞-格助詞					
塔	たふ	トウ	塔	トー	名詞-普通名詞-一般					
、			、		補助記号-読点					
此	これ	コレ	此れ	コレ	代名詞					
も	なり	ナリ	なり	ナリ	助動詞	文語助動詞-ナリ-断定	終止形-一般	断定		
と		ト	と	ト	助詞-格助詞					
なむ		ナム	なむ	ナン	助詞-係助詞					
語り	かた	カタル	語る	カタリ	動詞-一般	文語四段-ラ行	連用形-一般			
伝へ	つた	ツタエル	伝える	ツタエ	動詞-一般	文語下二段-ハ行	連用形-一般			
たる		タリ	たり	タル	助動詞	文語助動詞-タリ-完了	連体形-一般	完了		
と		ト	と	ト	助詞-格助詞					
や		ヤ	や	ヤ	助詞-係助詞					
。			。		補助記号-句点					

付　記

本発表は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）、及び、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、研究代表者：田中牧郎）の成果の一部である。

文　献

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」情報処理学会研究報告 人文科学とコンピュータ vol.2010-CH85,No.4

小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規定集」、平成 21（2009）－平成 23（2011）年度科学研究費補助金基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2 （http://dl.dropbox.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf よりダウンロード可能）

須永哲矢・堤智昭（2012）「小学館新全集『今昔物語集』での漢字活字－コーパス化のための調査と処理方針の検討－」、通時コーパスプロジェクト・オックスフォード大 VSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集、pp.15-22

冨士池優美・田中牧郎（2012）「今昔物語集の返読文字について―形態素解析の前処理を通して―」、日本語学会 2012 年度春季大会予稿集、pp.223-228

馬淵和夫・国東文麿・稲垣泰一（1999-2002）『新編日本古典文学全集 今昔物語集』1～4（小学館）

山田孝雄・山田忠雄・山田英雄・山田俊雄（1959-1963）『日本古典文学大系 今昔物語集』1～5（岩波書店）

『近代女性雑誌コーパス』の小説会話部分に現れる一・二人称代名詞の計量的分析

近藤 明日子（国立国語研究所コーパス開発センター）†

First- and Second-Person Pronouns in *Modern Women's Magazines Corpus*: A Quantitative Analysis Focusing on Conversational Sentences in Novels

KONDO Asuko (National Institute for Japanese Language and Linguistics)

１．はじめに

近代日本語の一・二人称代名詞に関する研究はこれまで主に、小説・戯曲の会話部分、落語速記、口語文典などの話し言葉的性質の強い口語文を資料として、当時の話し言葉における実態の解明に焦点をあてて進められてきた。ある程度の年代にわたる複数の資料を対象に複数の語形について考察を行った先行研究として、一人称代名詞を対象とした岡田（1998）・房（2004）・祁（2006a・2006b）や二人称代名詞を対象とした永田（2006・2008a・2008b・2009）などがあげられる。

今後、近代語のコーパスの開発が進むにつれ、コーパス中の話し言葉的性質の強い口語文を利用した一・二人称代名詞の研究も広がることが予想される。コーパスを利用することで、大量のテキストからすべての一・二人称代名詞を網羅的に抽出することがこれまでよりも低コストでできるようなり、一・二人称代名詞の全体像がさらに明らかになることが期待される。また、コーパス利用によって、これまでほとんど不可能であったテキストそのものの言語量の算出やそこから見出される言語的性質の分析も可能となり、対象テキストの性質をふまえた上でのより精緻な一・二人称代名詞の分析が展開されることも期待される。

本稿ではその一つの試みとして、近代語のコーパスの一つである『近代女性雑誌コーパス』（国立国語研究所（編）、2006）に形態論情報を付与したデータを用い、コーパス中の話し言葉的性質の強い口語文の代表として小説（戯曲を含む）に含まれる口語体の会話部分に着目し、その言語量とそこから見出される言語的性質について分析・考察を行う。そしてそこに出現する一・二人称代名詞を網羅的に抽出し、分析・考察を行う。考察では適宜、もう一つの代表的な近代語のコーパスである『太陽コーパス』（国立国語研究所（編）、2005）の分析結果との比較を交え、『近代女性雑誌コーパス』の特徴について考えたい。

２．『近代女性雑誌コーパス』の小説・戯曲の口語会話の抽出

『近代女性雑誌コーパス』は明治後期から大正期にかけて刊行された女性向け雑誌に基づくコーパスである。『女学雑誌』（1894～1895 年刊行分 31 冊）、『女学世界』（1909 年刊行分 6 冊）、『婦人倶楽部』（1925 年刊行分 3 冊）の計 40 冊 1362 記事が収録されており、『太陽コーパス』と対比させながら、当時の女性が読んでいた書き言葉の実態を把握することが可能な資料として設計されている（田中、2006）。

この『近代女性雑誌コーパス』に対して、近代の文語論説文を対象とする形態素解析辞書「近代文語 UniDic」（小木曽、2009）と旧仮名遣いの口語文を対象とする形態素解析辞書（小木曽、2012）を用いて形態素解析を行ったデータが、国立国語研究所の形態論情報データベース（小木曽・中村、2011）に格納されている。本稿ではこのデータベースの 2013

† kondo@ninjal.ac.jp

年 7 月時点のデータに基づき、分析・考察を行う。

コーパス中の小説・戯曲の区別はコーパスの XML の記事タグのジャンル属性に基づき、属性値の NDC 番号の 1 桁目が「9」、2 桁目が「1～9 または X」、3 桁目が「2～3」の記事を調査対象とする。会話部分の区別は引用タグの種別属性に基づき、属性値が「会話」のものを調査対象とする。ただし抽出された会話部分には文語体のテキストも含まれるため、記事タグ・引用タグの文体属性に基づき、口語体のテキストのみを抽出し調査対象とする。

３．小説・戯曲の言語量

まず、コーパス全体および小説・戯曲とその口語会話部分の言語量について見る。表 1 は『近代女性雑誌コーパス』『太陽コーパス』についてコーパス全体、小説・戯曲、小説・戯曲の口語会話それぞれの言語量を示したものである。なお、『近代女性雑誌コーパス』の 1894 年分は 1895 年分にまとめて示す（以下同様）。

表 1　コーパス全体、小説･戯曲、小説･戯曲の口語会話の言語量

		近代女性雑誌コーパス				太陽コーパス					
		1895	1909	1925	通年	1895	1901	1909	1917	1925	通年
コーパス全体	延べ語数	586665	406889	272325	1265879	2031346	1929238	1725992	1619638	1456055	8762269
	記事数	690	407	265	1362	729	635	652	504	889	3409
	号数	31	6	3	40	12	12	12	12	12	60
	1号あたりの平均延べ語数	18925	67815	90775	31647	169279	160770	143833	134970	121338	146038
小説	延べ語数	17889	30148	129620	177657	209529	187423	113857	265431	252996	1029236
	記事数	10	27	47	84	29	30	18	22	58	157
	著者数(異なり)	6	17	32	54	25	14	21	21	23	84
	1記事あたりの平均延べ語数	2982	1773	4051	3290	8381	13387	5422	12640	11000	12253
戯曲	延べ語数	29309	0	0	29309	0	0	96172	42239	0	138411
	記事数	9	0	0	9	0	0	10	4	0	14
	著者数(異なり)	2	0	0	2	0	0	6	4	0	10
	1記事あたりの平均延べ語数	14655	—	—	14655	—	—	16029	10560	—	13841
小説･戯曲	延べ語数	47198	30148	129620	206966	209529	187423	210029	307670	252996	1167647
	記事数	19	27	47	93	29	30	28	26	58	171
	著者数(異なり)	8	17	32	56	25	14	27	25	23	94
	コーパス全体に占める割合	8.0%	7.4%	47.6%	16.3%	10.3%	9.7%	12.2%	19.0%	17.4%	13.3%
小説･戯曲の口語会話	延べ語数	24313	7714	44410	76437	62680	69767	110002	80290	93090	415829
	口語会話を含む記事数	13	26	40	79	22	21	25	26	50	144
	口語会話を含む記事の著者数	5	16	28	48	10	11	20	21	23	85
	コーパス全体に占める割合	51.5%	25.6%	34.3%	36.9%	29.9%	37.2%	52.4%	26.1%	36.8%	35.6%

小説・戯曲の延べ語数は通年で 206966 語、『太陽コーパス』の通年 1167647 語の約 1/6 の言語量となっている。年ごとに小説・戯曲の延べ語数を見ると、もっとも多い 1925 年は 129620 語であるのに対し、もっとも少ない 1909 年は 30148 語と約 4 倍の違いがある。3 カ年の延べ語数の変動係数は 0.77 で、『太陽コーパス』5 カ年の延べ語数の変動係数の 0.20 と比べて高く、『近代女性雑誌コーパス』では年によるばらつきが顕著である。さらに、小説・戯曲中の口語会話の延べ語数においても、3 カ年の変動係数が 0.72 で『太陽コーパス』の 0.23 と比べて高く、年によるばらつきが顕著である。1909 年はもともと小説・戯曲の延べ語数が少ないのに加え、その中に口語会話の占める割合も 25.6%と他の年と比べて低く、その結果、口語会話の延べ語数は 7714 語と極めて少なくなっている。

次に、1 記事あたりの平均延べ語数を見ると、戯曲は通年で 14655 語と『太陽コーパス』の 13841 語と大きな違いはない一方で、小説は通年で 3290 語と『太陽コーパス』の 12253 語の約 1/4 である。その中でも 1909 年は 1773 語と特に少ない。この背景として、1 号あたりの言語量（平均延べ語数）が少ないことや 1909・1925 年は読者投稿による 100 語前後のごく短い作品が多く掲載されていることが考えられる。

このように『近代女性雑誌コーパス』の小説・戯曲はその延べ語数や含まれる口語会話の延べ語数、1 記事当たりの平均延べ語数に『太陽コーパス』と比べて大きな違いがあり、その中でも 1909 年は他の年と比べてその特異性が目立つ。『近代女性雑誌コーパス』の小説・戯曲の口語会話部分の利用においては、こうした性質に留意する必要がある。

4．小説・戯曲の口語会話部分の言語量

次に、小説・戯曲の口語会話について、話者の性別と文体の観点から見ていく。表 2 に『近代女性雑誌コーパス』『太陽コーパス』の小説・戯曲の口語会話について、話者の性別ごとに話者数・延べ語数・会話数を示す。会話数は、コーパスの XML の 1 つの引用タグでマークアップされている範囲を 1 会話としてカウントした。話者の性別は、コーパスの XML の引用タグの話者属性ごとに作品内容から判断した。また、会話の文体は、文末辞「ございます（り）ます」の出現する「ございます体」、「です」「ます」の出現する「ですます体」、「ございます（り）ます」「です」「ます」の出現しない「その他」の 3 種を設定し、会話ごとに文体を決定した。

表 2 小説・戯曲の口語会話の言語量

			近代女性雑誌コーパス				太陽コーパス					
			1895	1909	1925	通年	1895	1901	1909	1917	1925	通年
男	話者数		99	36	164	299	149	173	177	177	249	925
	延べ語数		15732	3051	29208	47991	41972	42766	75552	54482	70697	285469
	会話数	ございます体	38	3	50	91	88	18	34	26	154	320
		ですます体	246	55	378	679	214	425	927	518	910	2994
		その他	850	166	1034	2050	957	1141	2614	1984	2458	9154
		計	1134	224	1462	2820	1259	1584	3503	2528	3522	12396
女	話者数		48	62	106	216	50	47	94	117	98	406
	延べ語数		8557	4544	14601	27702	19493	26084	34138	23893	21130	124738
	会話数	ございます体	53	18	36	107	39	41	152	86	158	476
		ですます体	227	112	320	659	160	255	683	495	340	1933
		その他	305	186	580	1071	360	585	950	689	647	3231
		計	585	316	936	1837	559	881	1857	1270	1145	5712
不明	話者数		6	10	35	51	35	16	21	70	55	197
	延べ語数		24	119	601	744	1215	917	312	1915	1263	5622
	会話数	ございます体	0	0	0	0	6	2	2	0	1	11
		ですます体	2	4	14	20	28	14	0	5	20	67
		その他	7	13	40	60	54	30	30	119	72	305
		計	9	17	54	80	88	46	32	124	93	383
合計	話者数		153	108	305	566	234	236	292	364	402	1528
	延べ語数		24313	7714	44410	76437	62680	69767	110002	80290	93090	415829
	会話数	ございます体	91	21	86	198	133	61	188	112	313	807
		ですます体	475	171	712	1358	402	694	1610	1018	1270	4994
		その他	1162	365	1654	3181	1371	1756	3594	2792	3177	12690
		計	1728	557	2452	4737	1906	2511	5392	3922	4760	18491

表 2 に示した値に基づき、図 1 に会話の延べ語数の性別比率を、図 2 に会話数の性別比率を示す。

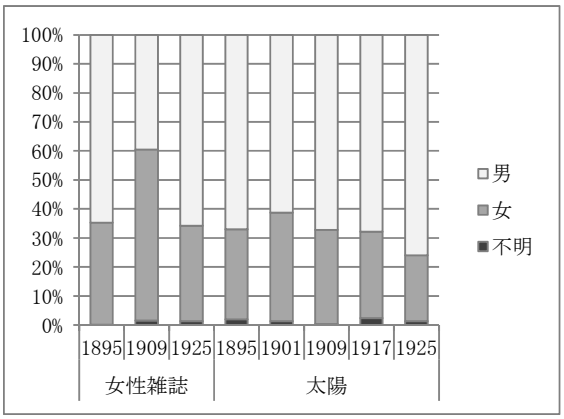


図 1 会話の延べ語数の性別比率

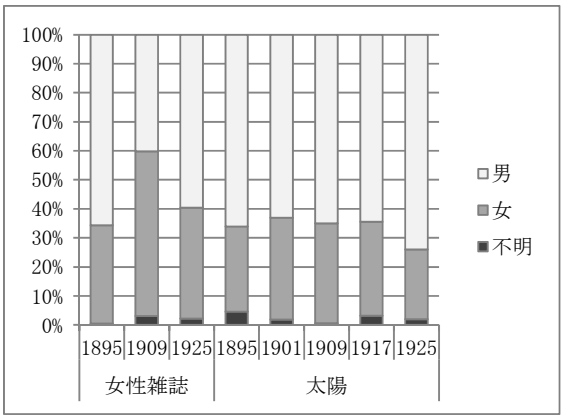


図 2 会話数の性別比率

『近代女性雑誌コーパス』において会話の性別比率は延べ語数・会話数のいずれでも 1909 年が他の年と比べて女性の比率が突出して高い。これは『太陽コーパス』と比較しても高い値である。話者の性別は一・二人称代名詞の分析の観点として重要なものであるが、『近

代女性雑誌コーパス』の利用においては 1909 年の特異性に留意する必要がある。

次に表 2 に示した値に基づき、図 3 に男性の会話における文体比率を、図 4 に女性の会話における文体比率を示す。

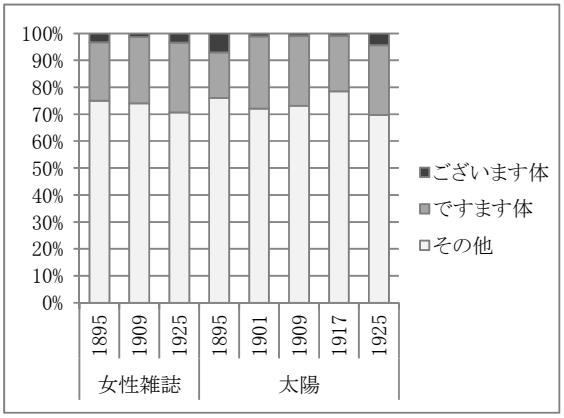


図 3　男性の会話の文体比率

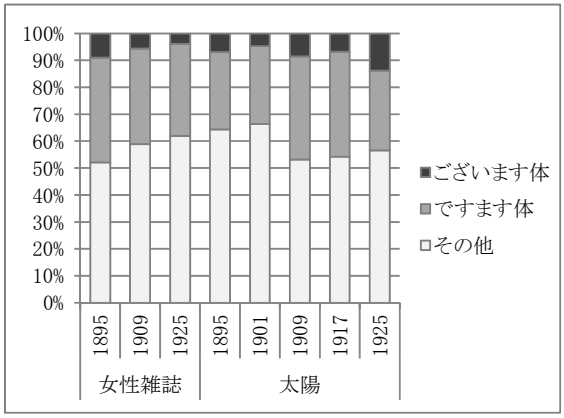


図 4　女性の会話の文体比率

『近代女性雑誌コーパス』において、女性は男性より「ございます体」「ですます体」といった敬体の会話の比率が全体的に高い。これは『太陽コーパス』にも共通する傾向である。会話の文体もまた一・二人称代名詞の分析の観点として重要なものであるが、話者の性別により文体比率に差があることに留意する必要がある。

５．男性の会話に出現する一・二人称代名詞

以上、言語量から見えてきた『近代女性雑誌コーパス』の会話・戯曲とその口語会話の言語的性質を前提として、そこに出現する一・二人称代名詞を抽出・分析する。

分析対象とする一・二人称代名詞は、コーパスに付与された形態論情報の一部の人手修正を経て、形態論情報で品詞が「代名詞」の語彙素（見出し語）から、一人称代名詞・二人称代名詞いずれかの用法のみ持つものを選択した。また、二人称代名詞は接尾辞「さま」「さん」の下接するものは別語形とした。なお、本文テキストで漢字表記されるものについては、ルビ情報のあるものはそこから語形を特定し、ルビ情報のない漢字表記のものは最も一般的と考えられるよみを語形とした。これらのものは本稿中ではカタカナで表記する。ただし、ルビ情報のない漢字表記のもので語形を一つに特定することが困難と考えたものは語形未確定として、本稿中では漢字で表記する。

最初に、男性の会話に出現する一人称代名詞を見ていく。表 3 に男性の会話に出現する一人称代名詞の粗頻度・出現会話数・出現記事数（空欄は 0 を表す）を示す。

『近代女性雑誌コーパス』からは 14 種類の語形が抽出された。『太陽コーパス』で同様の調査をすると 21 種類の語形が抽出され、『近代女性雑誌コーパス』のほうが語形の種類が少ない。年別に見ると、1909 年は 4 種類と特に種類が少ない。

次の図 5 は、表 3 で通年の粗頻度降順上位 4 語「ボク」「オレ」「ワタシ」「ワシ」とその他の語形をあわせて「その他」として、通年での粗頻度の比率を示し、比較のため『太陽コーパス』についても同様に示したものである。

上位 4 語形の比率は合計 82%で『太陽コーパス』の 80%と大きな差はない。ただし、4 語形の中では「ボク」の比率が 52%で『太陽コーパス』の 33%と比べて高い。

表 3　男性の会話に出現する一人称代名詞の粗頻度・出現会話数・出現記事数

	1895			1909			1925			通年		
	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数
ボク	34	31	7	24	20	4	214	139	11	272	190	22
オレ	14	11	3	6	4	2	44	36	11	64	51	16
ワタシ	30	27	8	3	2	2	27	26	8	60	55	18
ワシ	3	2	2				33	22	5	36	24	7
セッシャ							33	28	3	33	28	3
ワタクシ	14	14	3				11	11	2	25	25	5
ワレワレ	3	3	3				10	9	2	13	12	5
オイラ	6	6	3							6	6	3
ヨ							5	5	3	5	5	3
テマエ							4	4	2	4	4	2
私				4	2	2				4	2	2
オラ							2	2	2	2	2	2
オヌシ	1	1	1							1	1	1
ワイ							1	1	1	1	1	1

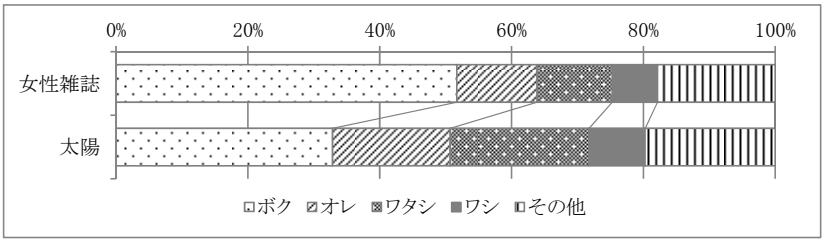


図 5　男性の会話に出現する一人称代名詞の粗頻度（通年）の比率

次に男性の会話に出現する二人称代名詞を見ていく。表 4 に男性の会話に出現する二人称代名詞の粗頻度・出現会話数・出現記事数（0 は空欄）を示す。

表 4　男性の会話に出現する二人称代名詞の粗頻度・出現会話数・出現記事数

	1895			1909			1925			通年		
	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数
キミ	80	70	8	13	13	4	61	48	6	154	131	18
アナタ	41	35	8	5	3	3	106	75	15	152	113	26
オマエ	1	1	1	5	5	2	53	44	11	59	50	14
キコウ							19	16	3	19	16	3
キサマ							17	10	3	17	10	3
ソノホウ							10	10	3	10	10	3
ナンジ	8	5	1				1	1	1	9	6	2
オマイ	7	7	4							7	7	4
ソチ	1	1	1				5	5	1	6	6	2
オマエサン				1	1	1	3	3	2	4	4	3
アナタサマ	3	3	1							3	3	1
オメエ	1	1	1				2	2	2	3	3	3
オマイサン	2	2	1							2	2	1
キデン							2	2	1	2	2	1
オ前				2	2	1				2	2	1
ウヌ	1	1	1				1	1	1	2	2	2
オメエサマ	2	1	1							2	1	1
オメエサン	2	1	1							2	1	1
オ前サン	1	1	1							1	1	1
テメエ							1	1	1	1	1	1

『近代女性雑誌コーパス』からは 20 種類の語形が抽出された。『太陽コーパス』から抽出される 26 種類と比べると少ない。年別に見ると、一人称代名詞同様、1909 年は 5 種類と特に種類が少ない。

次の図 6 は表 4 で通年の粗頻度降順上位 3 語「キミ」「アナタ」「オマエ」とその他の語形をあわせて「その他」として、通年での粗頻度の比率を示し、比較のため『太陽コーパス』についても同様に示したものである。

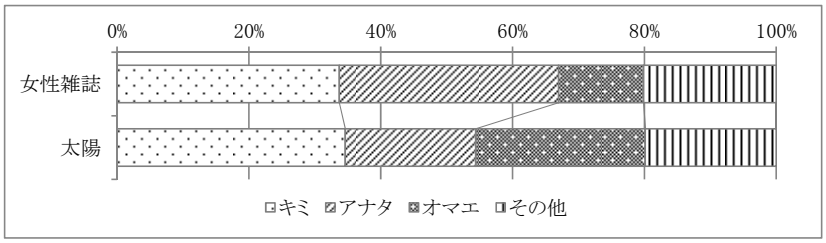


図 6　男性の会話に出現する二人称代名詞の粗頻度（通年）の比率

上位 3 語形の比率は合計 80%で『太陽コーパス』と同値である。ただし、3 語形の中では「アナタ」の比率が 33%で『太陽コーパス』の 20%と比べて高く、一方「オマエ」が 13%で『太陽コーパス』の 28%と比べて低い。

以上のように、男性の会話に出現する一・二人称代名詞は『太陽コーパス』と比べて語形の種類が少なく、また、一人称代名詞は「ボク」、二人称代名詞は「キミ」「アナタ」といった特定の語形が偏って出現する傾向にあることが分かった。

最後に、男性の会話に主に出現する一・二人称代名詞について、文体ごとの出現会話率を概観しておく。出現会話率とは総会話数に対する該当語形の出現会話数の占める割合のことである。図 7 は一人称代名詞「ボク」「オレ」「ワタシ」の、図 8 は二人称代名詞「キミ」「アナタ」「オマエ」の文体ごとの出現会話率を示したものである。

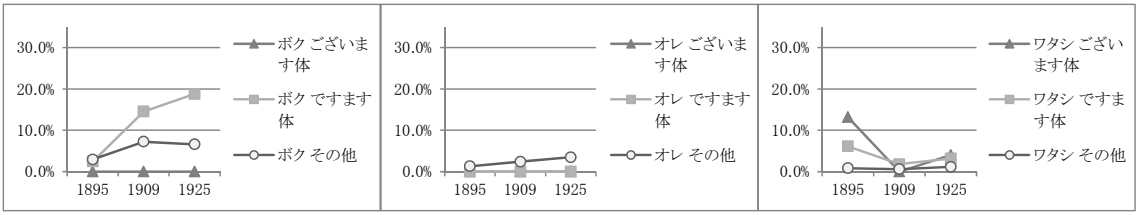


図 7　男性の会話に出現する一人称代名詞の出現会話率

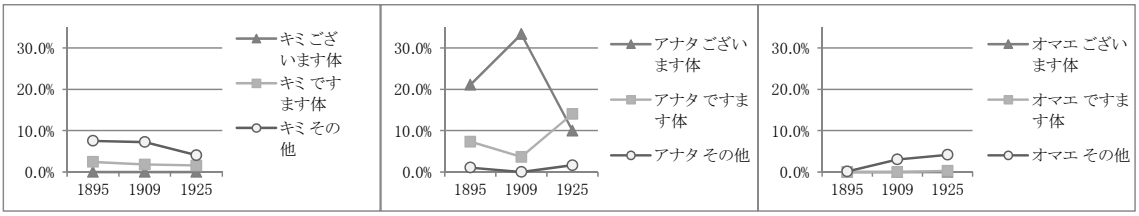


図 8　男性の会話に出現する二人称代名詞の出現会話率

一人称代名詞の「ボク」は「ですます体」と「その他」に、「オレ」は「その他」に、「ワタシ」は「ございます体」「ですます体」に主に出現する。文体別に見ると、「ございます体」に「ワタシ」、「ですます体」に「ボク」「ワタシ」、「その他」に「ボク」「オレ」が主に出現する。

二人称代名詞の「キミ」は「その他」「ですます体」に、「アナタ」は「ございます体」「ですます体」に、「オマエ」は「その他」に主に出現する。文体別に見ると、「ございます体」に「アナタ」、「ですます体」に「アナタ」「キミ」、「その他」に「キミ」「オマエ」が主に出現する。

６．女性の会話に出現する一・二人称代名詞

次に女性の会話に出現する一・二人称代名詞を見ていく。表 5 に女性の会話に出現する一人称代名詞の粗頻度・出現会話数・出現記事数（0 は空欄）を示す。

表 5 女性の会話に出現する一人称代名詞の粗頻度･出現会話数･出現記事数

	1895			1909			1925			通年		
	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数
ワタシ	122	94	9	28	25	4	186	147	16	336	266	29
ワタクシ	1	1	1	4	4	2	31	21	4	36	26	7
ワレワレ							10	6	1	10	6	1
アタシ				3	2	2	4	4	1	7	6	3
私				4	4	3	2	2	1	6	6	4
アタイ				3	3	2	1	1	1	4	4	3
ワシ							4	3	1	4	3	1
ワテ							2	2	1	2	2	1
妾				1	1	1				1	1	1
オレ							1	1	1	1	1	1
テマエ							1	1	1	1	1	1

『近代女性雑誌コーパス』からは 11 種類の語形が抽出された。『太陽コーパス』から抽出される 16 種類と比べると少ない。年別に見ると、1895 年は 2 種類と特に少ない。

次の図 9 は、表 5 で通年の粗頻度降順上位 2 語「ワタシ」「ワタクシ」とその他の語形をあわせて「その他」として、通年での粗頻度の比率を示し、比較のため『太陽コーパス』についても同様に示したものである。

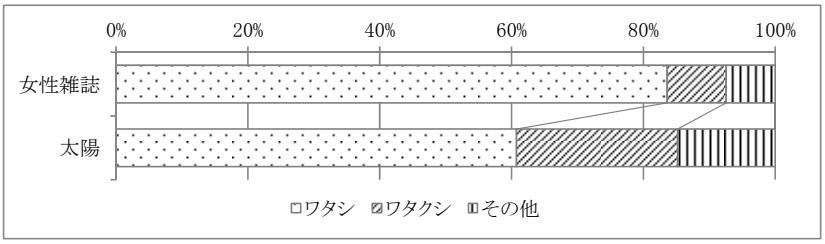


図 9 女性の会話に出現する一人称代名詞の粗頻度（通年）の比率

上位 2 語形の比率は合計 93%で『太陽コーパス』の 85%より高い。また、2 語形の中では「ワタシ」の比率が 82%で『太陽コーパス』の 61%より高く、「ワタクシ」は 9%で『太陽コーパス』の 24%より低い。

次に二人称代名詞を見ていく。表 6 に女性の会話に出現する二人称代名詞の粗頻度・出現会話数・出現記事数（0 は空欄）を示す。

表 6 女性の会話に出現する二人称代名詞の粗頻度･出現会話数･出現記事数

	1895			1909			1925			通年		
	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数	粗頻度	出現会話数	出現記事数
アナタ	108	95	9	9	9	6	98	88	16	215	192	31
オマイ	20	15	4							20	15	4
オマエ				2	2	1	15	14	7	17	16	8
アンタ							9	6	2	9	6	2
オマエサン				1	1	1	5	5	3	6	6	4
オ前				3	3	2				3	3	2
オマイサン	1	1	1							1	1	1

『近代女性雑誌コーパス』からは 7 種類の語形が抽出された。『太陽コーパス』から抽出される 19 種類と比べると少ない。

次の図 10 は、表 6 で通年の粗頻度降順上位 3 語「アナタ」「オマイ」「オマエ」とその他の語形をあわせて「その他」として、通年での粗頻度の比率を示し、比較のため『太陽コーパス』についても同様に示したものである。

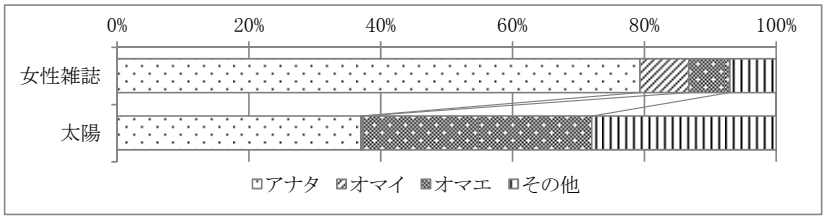


図 10　女性の会話に出現する二人称代名詞の粗頻度（通年）の比率

上位 3 語形の比率は合計 93%で『太陽コーパス』の 72%より高い。3 語形の中では「アナタ」の比率が 79%で『太陽コーパス』の 37%より高く、「オマエ」は 6%で『太陽コーパス』の 35%より低い。なお、「オマイ」は特定の 1 連載記事群のみに出現する語形で、その特殊性を考慮する必要がある。

以上のように、女性の会話に出現する一・二人称代名詞も男性の場合と同様に、語形の種類が少なく、また、一人称代名詞は「ワタシ」、二人称代名詞は「アナタ」といった特定の語形が偏って出現する傾向にあることが分かった。

最後に、女性の会話に主に出現する一・二人称代名詞について、文体ごとの出現会話率を概観しておく。図 11 は一人称代名詞「ワタシ」「ワタクシ」の、図 12 は二人称代名詞「アナタ」「オマエ」の文体ごとの出現会話率を示したものである。

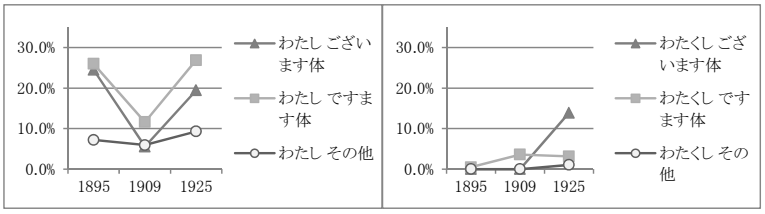


図 11　女性の会話に出現する一人称代名詞の出現会話率

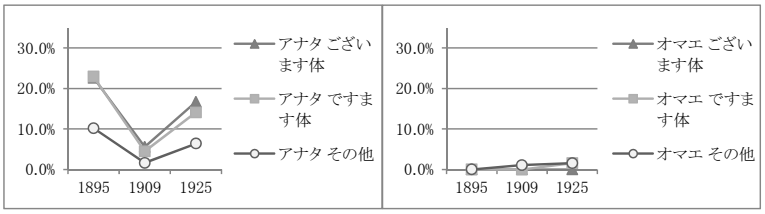


図 12　女性の会話に出現する二人称代名詞の出現会話率

一人称代名詞では「ワタシ」が、二人称代名詞では「アナタ」がどの文体でももっとも多く出現する語形となっている。

7．おわりに

以上、『近代女性雑誌コーパス』の小説・戯曲の口語会話部分に着目し、言語量から見た言語的性質をふまえた上で、そこに出現する一・二人称代名詞を抽出・分析した。

言語的性質については、1909 年が他の年や『太陽コーパス』と比べて言語量や会話の性別比率において異質であることが明らかになった。『近代女性雑誌コーパス』を利用する際には留意すべき点である。特に経年変化を分析する際、年ごとのばらつきは支障となる可

能性がある。

一・二人称代名詞については、特定の語形が偏って出現する傾向にあることが明らかになった。この背景として、会話の話し手・聞き手の社会階層や話し手と聞き手の関係といった一・二人称代名詞の選択に関わる場面要素に偏りがあることが想定される。それは単純に小説・戯曲の口語会話の言語量が多くないために生じたことなのか、それとも女性雑誌ゆえに小説・戯曲の題材が限定されてのことなのか、今後のさらなる調査・分析をまちたいが、いずれにせよ、『近代女性雑誌コーパス』の小説・戯曲の口語会話部分からは当時の話し言葉における一・二人称代名詞の或る一面しか見て取れないことは明らかである。『近代女性雑誌コーパス』の利用にあたっては『太陽コーパス』など他のコーパスと組み合わせるなど工夫が必要であろう。

また、本稿では一・二人称代名詞の分析の観点として話し手の性別と文体をとりあげたが、その他に話し手・聞き手の社会階層や話し手と聞き手の関係など、研究において重要な観点がまだ残されており、それらすべてを組み合わせた分析を行う必要がある。今後の課題としたい。

付　記

本稿は、日本学術振興会科研費（23720242）および国立国語研究所共同研究プロジェクト「通時コーパスの設計」による研究成果である。

文　献

岡田賢二（1998）「明治期の東京語における人称代名詞の研究―明治・大正期の落語の速記本にあらわれた一、二人称代名詞―」『埼玉大学国語教育論叢』2、pp.34-58

小木曽智信（2009）『科学研究費補助金研究成果報告書　近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（http://www2.ninjal.ac.jp/lrc/index.php?UniDic よりダウンロード可能）

小木曽智信（2012）「旧仮名遣いの口語文を対象とした形態素解析辞書」『じんもんこん 2012 論文集』2012:7、pp.25-32

小木曽智信・中村壮範（2011）『特定領域研究「日本語コーパス」平成 22 年度研究成果報告書　『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装　改訂版』（http://www.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-U-10-01.pdf よりダウンロード可能）

祁福鼎（2006a）「明治時代語における自称詞の使用実態と使用規範について」『文学研究論集』24、pp.45-61

祁福鼎（2006b）「明治時代語における自称詞の推移と位相について」『明治大学日本文学』32、pp.95(1)-78(18)

国立国語研究所（編）（2005）『太陽コーパス―雑誌『太陽』日本語データベース』博文館新社

国立国語研究所（編）（2006）『近代女性雑誌コーパス』（http://www.ninjal.ac.jp/corpus_center/cmj/woman-mag/よりダウンロード可能）

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社、pp.1-48

田中牧郎（2006）「『近代女性雑誌コーパス』の概要」『日本学術振興会科学研究費補助金研究成果報告書 基盤研究(B)「20 世紀初期総合雑誌コーパス」の構築による確立期現代語の高精度な記述』pp.55-62（http://www.ninjal.ac.jp/corpus_center/cmj/doc/19w-mag-summary.pdf よりダウンロード可能）

永田高志（2006）「明治前期東京語の対称詞―散切物を通じて」『国語国文』75:6、p.16-33

永田高志（2008a）「国定教科書の対称詞」『国語と国文学』85:3、pp.56-68

永田高志（2008b）「明治後期・大正期東京語の対称詞」『日本文化の鉱脈―茫洋と閃光と』風媒社、pp.95-110
永田高志（2009）「総合雑誌『太陽』に見る対称詞」『国語と国文学』86:9、pp.56-70
房極哲（2004）「近代語における一、二人称代名詞の変遷について」『日本文化學報』21、pp.1-15

近世口語資料の形態素解析の試み

小木曽智信（国立国語研究所言語資源研究系）†
市村太郎（国立国語研究所コーパス開発センター）
鴻野知暁（国立国語研究所コーパス開発センター）

Preliminary Study of Morphological Analysis of Early Modern Japanese

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Taro Ichimura (National Institute for Japanese Language and Linguistics)
Tomoaki Kono (National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所では「日本語歴史コーパス」の一部として、近世の口語を反映した資料群のコーパス化計画を進めている（近藤 2012）。このうち、虎明本狂言集と洒落本の一部については、すでに電子化・形態論情報の付与に着手している。形態論情報の付与にあたっては、高精度で均質なタグ付けのために形態素解析が欠かせないが、従来の形態素解析辞書では近世の口語文を十分な精度で改正することができなかった。そのため、発表者らは UniDic（中古和文 UniDic・近代文語 UniDic）の見出し語データと整備済みの狂言・洒落本等のコーパスを用いて、近世口語文を解析するための辞書の作成に取り組んでいる。本発表では、この近世口語文を対象とした形態素解析の方法、現在達成している解析精度、今後の見通しについて論ずる。

２．狂言・洒落本のコーパス化

狂言は、中世から近世にかけての言語資料として重要な位置を占めている。登場人物が多彩で身分関係が明確であること、対話劇の形で進行し場面・状況が明確であることから、口語資料としての価値は極めて高い。狂言資料の中でも『虎明本』は、寛永 19 年（1642）大蔵流十三世宗家大蔵弥太郎虎明の手による大蔵流の祖本である。本狂言 237 曲を収めており、狂言の類別や詞章の整備された台本として、質・量とも第一級の資料である。その詞章には、中世、室町時代の言葉を伝承していると見られる点、書写当時である近世初期の日常語の影響を受けたと思われる点、舞台言語として整理され固定化・類型化する兆候が見られる点がある。狂言史上の位置を踏まえ、他の台本との比較ということが不可欠であるが、注釈書や総索引が整備され、中世から近世の言語資料として広く利用されてきた（小林・市村 2013）。虎明本では、狂言台本としての性格上、通常であれば漢字表記される語に仮名書きが多い特徴があり辞書未登録の表記が発生しやすく、形態素解析を難しくしている。一方、同一人による写本であるため、全体として均質性も持つ。以下に虎明本「あさう」の一部を掲げる。

例：（あさう）

「〽信濃の国の住人、あさうのなにがしです、そせうの子細あつて、在京仕る処に、安堵の御教書を下され、新地を拝領いたし、あまつさへおいとまを下された、のさ者をよび出し、よろこばせうとぞんずる、藤六あるかやい　〽お前に　〽下六もよべ　〽やい下六めすはやい　〽何とめすといふか　〽あふ　〽とういふてくれひで、お前に〽両人ながらはやかつた、やいなんぢらがよろこぶ事があるは、　〽そはまづめでたひ事で御ざあるが、何事で御ざあるぞ　〽永々在京いたす程にと有て、あんどの御教書を下され、新地を拝領して、おいとままで下されたが、かたじけなひ事ではなひか

† toigso@ninjal.ac.jp

洒落本は、登場人物の会話部分に当時の話し言葉が反映されているとされ、日本語史研究上、近世後期の口語の実態を探る上での重要資料である。大きく分けて江戸版と上方版があり、その口語体の会話部分はそれぞれの地域の言葉を反映する場合も多い。また年代も 18C 後半から 19C 前半までと幅広く、近・現代語への過渡的状況を伺うのに適している。方言や中央語の形成を知る上でも、不可欠な資料である（市村ほか 2013）。洒落本は、作品ごとに内容が異なるだけでなく、江戸・上方で言語そのものが大幅に異なっており、作者・形式も多様で均質性は低い。さらに言葉遊び的な要素をしばしば含むため、形態素解析やコーパス化には課題が多い。以下に洒落本の一つ『聖遊廓』の一部を掲げる。

例：(聖遊廓)

爰に聖人のかよひたまへる郭(くるわ)あり　揚屋(あげや)の亭主(ていす)は李白(りはく)とかや　中にも孔子はくるわにてすいといはれて端手(はで)ならず　ゑちご縮(ちゞみ)のかたびらにもんろの羽織(はおり)すそながく深(ふか)あみがさにあわざうり古金買(ふるかねかい)の目利(めきゝ)にも太夫かいとは見へざりし　李白がかたへ御入りあれば　▲亭主李白 是は仁さまおめづらしい さあ／＼おくへ ともてはやす　▲孔子 なんと李す此中は久しいの 無事で珍重／＼　と座敷へ行　▲李白女房滝(たき)　是はおめづらしいおかほ。 おうはさばつかり申ておりました　▲中居なつ もし仁さま此中横堀でお見うけ申ましたゆへ大かたおよりなさるであろふとぞんじましたに。よふまたせなさつたの　▲孔子 ヲヽよりたかつたけれども行時(ゆくとき)に径(こみち)によらず。

３．学習・評価用コーパス

狂言と洒落本のテキストのうち、現在、表１に示すものが単語情報付きのコーパスとして整備済みである。文書構造のアノテーション、濁点付与・文境界付与等の本文整備を施した後、既存の形態素解析辞書を用いて形態素解析を行って形態論情報データベースに格納し、その後人手によって解析の誤りを修正したものである。

表１　近世口語資料の人手修正済みコーパス

ジャンル	狂言	洒落本	人情本・滑稽本	*合計*
語数（短単位）	9515	25594	14361	*49470*

狂言の内訳は、次の 8 曲である。

「ゑびす大黒」「連歌毗沙門」「入間川」「あさう」「ゑびす毗沙門」「あさいな」「わかな」「腹不立」

洒落本の内訳は次の 5 作品である。本文は、『陽台遺編・妣閣秘言』『風流裸人形』『興斗月』は「洒落本大成」、『遊子方言』『跕婦人伝』は小学館「新編日本古典文学全集」によっている。

『遊子方言』　1770（明和 7）年，江戸
『跕婦人伝』　1749（寛延 2）年，江戸
『陽台遺編・妣閣秘言』　1758（宝暦 8）年，大阪
『風流裸人形』　1779（安永 8）年，京都
『興斗月』　1836（天保 7）年，京都

なお、人情本・滑稽本は、滑稽本が『浮世床』初編の一部、人情本が『春告鳥』初編の一部でいずれも本文は「新編日本古典文学全集」によっている。

表 1 のコーパスのうち、狂言・洒落本のそれぞれ約１割を文単位でランダムサンプリングして精度評価用コーパスを作成した。語数を表 2 に示す。

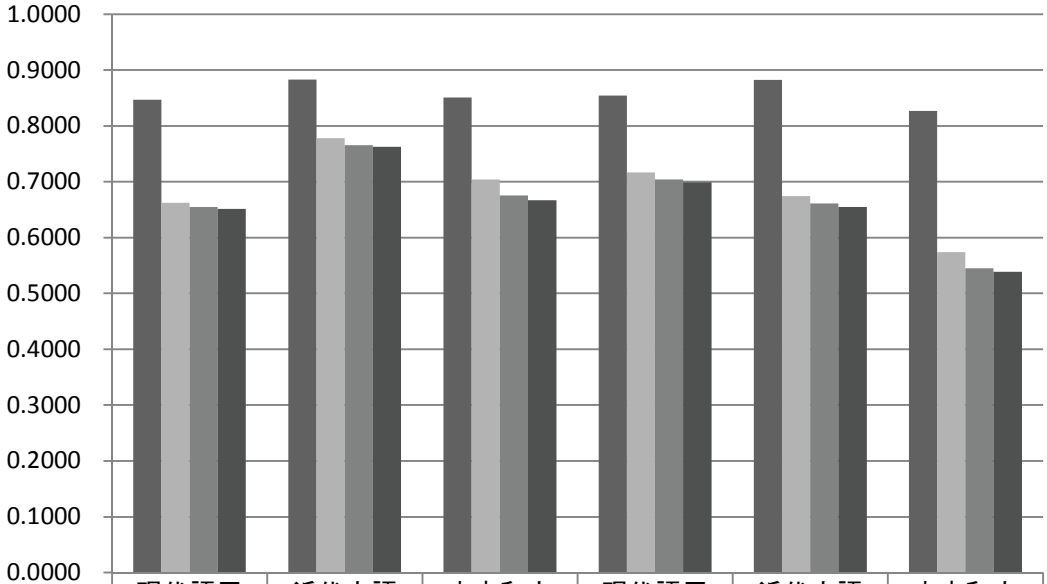
表２　狂言・洒落本の評価用コーパス

ジャンル	狂言	洒落本
語数（短単位）	1030	2448

４．既存の UniDic による解析精度

狂言・洒落本のテキストは、いずれも現代語とは大幅に異なる上に、典型的な古文である平安和文などとも大きく異なる文体で書かれている。発表者らは、これまでに歴史的な資料を対象とした形態素解析辞書として現代語用の UniDic をもとに「中古和文 UniDic」「近代文語 UniDic」を開発・公開してきたが、このいずれも狂言・洒落本の解析には適していない。

形態素解析器に MeCab（Kudo et al. 2004）[1]を用い、現代語用の UniDic と中古和文 UniDic、近代文語 UniDic のそれぞれで狂言・洒落本の評価用コーパスを解析し、精度を評価した。結果を図 1 に示す。数値は F 値（再現率と適合率の調和平均）である。



	現代語用	近代文語	中古和文	現代語用	近代文語	中古和文
	― 狂言 ―			― 洒落本 ―		
lv.1	0.8471	0.8829	0.8506	0.8543	0.8822	0.8269
lv.2	0.6626	0.7778	0.7040	0.7168	0.6744	0.5740
lv.3	0.6551	0.7653	0.6753	0.7044	0.6610	0.5450
lv.4	0.6513	0.7624	0.6667	0.6988	0.6549	0.5385

図 1　既存の UniDic による狂言・洒落本テキストの解析精度

図 1 の lv.1 は、解析結果において単語の境界が正しかったかどうか、lv.2 は境界が正しいことに加えて単語の品詞・活用型・活用形も正しく認定されていたかどうかを見るものである。lv.3 は lv.1 と lv.2 に加えて語彙素の認定も正しかったかどうかを見る。たとえば「金」が「キン」でなく「カネ」と解析されているかどうかという違いに相当する。lv.4 は、発音（読み）の違いが正しく認定されているかどうかを評価するもので、lv.1・lv.2・lv.3 が正しいことに加え、さらに読み方が正しいかどうかを見る。たとえば「言語」が文脈にあわせて「ゲンゴ」ではなく「ゴンゴ」と解析されているかどうかに相当する。

[1] 学習・解析ともに使用した MeCab のバージョンは 0.993 である。

現代語のコーパス構築において必要とされた解析精度が、語彙素認定（lv.3）において 98%であった。歴史コーパスの構築でも、おおむね 96%以上の解析精度が必要とされ、「中古和文 UniDic」「近代文語 UniDic」はほぼその解析精度を実現していた。しかし、図 2 にみるように、既存の解析辞書では近世口語資料の解析は難しく、狂言のテキストが最高で 76.5%、洒落本のテキストが最高で 70%程度の解析精度に留まっている。本格的なコーパス構築に用いるには大幅に精度が不足しており、新たな解析辞書を作成する必要がある。

５．近世口語共通辞書の解析精度

表 1 のコーパスのうち、表 2 の評価用を除く全てを利用して学習を行い、近世口語用の形態素解析辞書を作成した。見出し語は、従来の中古和文 UniDic・近代文語 UniDic で用いていたものに、表 1 のコーパスで出現した語を追加したものを利用した。見出し語は、活用形展開後で総計 134 万に上る。近世口語では利用されない語彙を含むが、①どの語が不要であるかを事前に判断することは必ずしも容易ではないこと、②古文の形態素解析辞書にとって見出し語の肥大化は大きな問題ではないこと、③不要語があることによる解析精度への悪影響は特に認められなかったこと、によりそのまま利用している。

表 3 にこの近世口語共通辞書の解析精度を示す。各レベルの意味は図 1 と同様、数値は F 値である。

表 3　近世口語共通辞書の解析精度

	狂言	洒落本
lv.1	0.9639	0.9681
lv.2	0.8652	0.8635
lv.3	0.8613	0.8545
lv.4	0.8594	0.8491

既存の辞書と比較すると精度は向上しているが、コーパス構築に十分な性能とはいえない。原因として学習用コーパスの絶対的な不足が考えられる。現状の学習用コーパスの量は約 4.5 万語だが、近代文語 UniDic では約 64 万語、中古和文 UniDic では約 82 万語を用いており、現在の学習用コーパスの量では不十分である。

しかし、量の問題とは別に、狂言と洒落本という質的にかなり異なるテキストを近世口語として一括していることにも問題があると考えられる。

６．狂言・洒落本専用辞書の解析精度

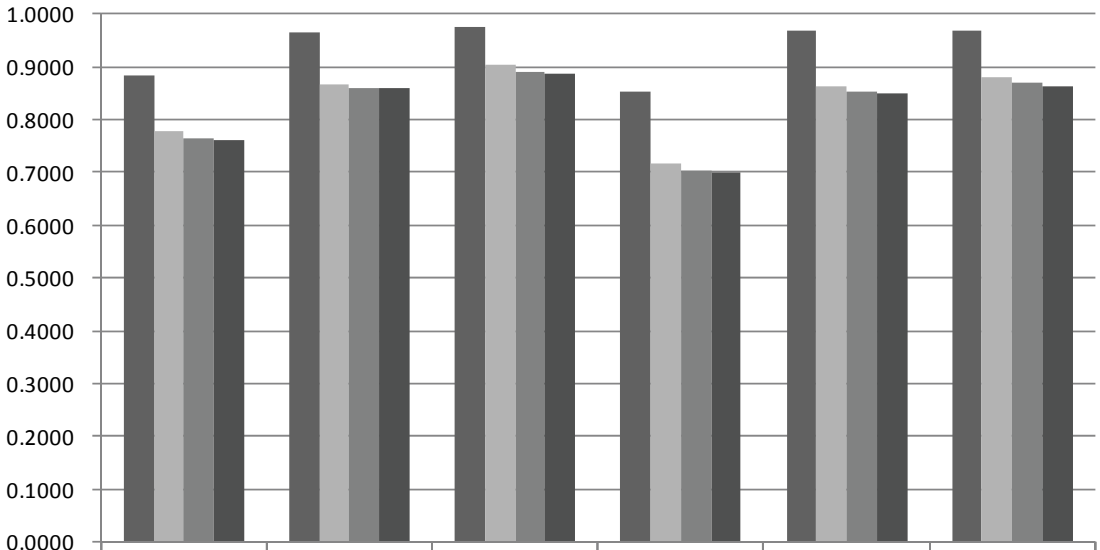
予備的な実験により、歴史的資料の形態素解析を行う際には、異分野のテキストによる学習結果を流用するより、少量であっても専用コーパスによる学習が効果的であることが分かっている。そこで、狂言と洒落本を分割し、それぞれの専用辞書を作成して近世口語共通の辞書と解析精度を比較することにする。分割により、もともと不十分であった学習用コーパスの量はほぼ半減することになるが、専用の学習用コーパスのみを利用することによるメリットがそれを上回る可能性がある。

狂言の学習は狂言のコーパスだけを学習に利用し、洒落本は、滑稽本・人情本に時代的にも内容的にも比較的近いため、これらを利用するもの（A）と純粋に洒落本だけを利用するもの（B）の 2 通りを作成した。それぞれの辞書による解析精度を表 4 に示す。見方は表 3 と同様であるが、表 4 では狂言と洒落本が、それぞれ別の辞書による評価結果となっていることに注意されたい。表 4 の数値は、学習用コーパスの量を大幅に減らしたものであるにもかかわらず、表 3 の共通辞書による解析精度よりも向上している。したがって、狂言と洒落本とでは別の解析辞書を用意すべきであることが分かる。洒落本の（A）（B）についても、滑稽本や人情本を交えない（B）のほうがよりよい精度となっている。

表 4　狂言・洒落本専用辞書の解析精度

	狂言専用	洒落本用（A）	洒落本専用（B）
lv.1	0.9747	0.9695	0.9699
lv.2	0.9026	0.8738	0.8791
lv.3	0.8900	0.8636	0.8688
lv.4	0.8870	0.8583	0.8627

図 2 は、これまでに精度を確認してきた辞書による解析精度をグラフにまとめたものである。既存の辞書による解析結果のうち最良のもの（狂言は近代文語 UniDic、洒落本は現代語用 UniDic）と、近世口語共通辞書、専用辞書の解析結果を比較した。



	既存最良（近代文語）	共通辞書	専用辞書	既存最良（現代語）	共通辞書	専用辞書
	― 狂言 ―			― 洒落本 ―		
lv.1	0.8829	0.9639	0.9747	0.8543	0.9681	0.969919
lv.2	0.7778	0.8652	0.9026	0.7168	0.8635	0.879065
lv.3	0.7653	0.8613	0.8900	0.7044	0.8545	0.868834
lv.4	0.7624	0.8594	0.8870	0.6988	0.8491	0.862696

図 2　各辞書による狂言・洒落本テキストの解析精度比較

このように提案手法により精度はかなり向上したものの、現状の解析精度は狂言・洒落本ともに lv.3（語彙素認定）で 9 割を切っており、コーパス構築のために十分な精度には達していない。これは、学習用コーパスの量の絶対的な不足に起因するものと考えられる。

7．おわりに

新たにコーパスからの学習を行って狂言用と洒落本用の形態素解析辞書を作成し、既存の辞書を上回る精度で解析を行うことが可能になった。また、近世口語を一括した共通辞書を作成する場合と、対象分野を分割した専用辞書を作成する場合とで精度の比較を行うことにより、狂言と洒落本とは別に扱うことが良いことが確認された。

現状では、狂言と洒落本を分割すると学習用のコーパスが大きく不足するため、解析精度は語彙素認定で 90%に達していない。今後、それぞれの学習用コーパスの量を増やし、見出し語の増補を行うことで専用辞書を充実させ、狂言・洒落本のコーパス構築に資するものにしていきたい。

謝　辞

本研究は JSPS 科研費 24520522 の助成を受けたものである。また、本研究の一部は国立国語研究所の共同研究プロジェクト「通時コーパスの設計」および「統計と機械学習による日本語史研究」による研究成果を含む。

文　献

市村 太郎，河瀬 彰宏，小木曽 智信(2012)「近世口語テキストの構造化とその課題」情報処理学会研究報告 人文科学とコンピュータ研究会報告(CH96) pp.1-8

市村 太郎，河瀬 彰宏，小木曽 智信(2013)「洒落本コーパスの構造化 ―仕様と事例の検討―」『第 3 回コーパス日本語学ワークショップ予稿集』pp.249-258

小林 正行，市村 太郎（2013）「『虎明本狂言集』コーパスの構造化 －仕様と事例の検討－」『第 3 回コーパス日本語学ワークショップ予稿集』pp.323-332

近藤泰弘(2012)「日本語通時コーパスの設計について」『国語研プロジェクトレビュー』3 pp.84-92

伝康晴，小木曽智信，小椋秀樹，山田篤，峯松信明，内元清貴，小磯花絵(2007).「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22 号 pp.101-122.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* , Barcelona, Spain, pp.230–237.

関連 URL

MeCab: Yet Another Part-of-Speech and Morphological Analyzer　https://code.google.com/p/mecab/

NINJAL「通時コーパス」プロジェクト　http://historicalcorpus.jp

UniDic　http://sourceforge.jp/projects/unidic/

近代文語 UniDic，中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

日本語連用形名詞の自立性の段階について

沈 晨（北京外国語大学北京日本学研究センター）†

On the Levels of the Independence of Japanese Infinitive-derived Nouns

SHEN　CHEN（Beijing Foreign Studies University）

1. はじめに

日本語には、動詞の名詞化にあたって、動詞の諸活用形中の一形である連用形が、そのままの形で名詞に転化するという、簡単な方式が古くから存続している(西尾 1961)。例えば、

(1)　動き、遊び、扱い、悩み、嗜み、受け入れ、立ち読み

などがある。本稿ではその過程を転成、または名詞化と呼ぶ。それら動詞連用形から形成されたと考えられる名詞を「連用形名詞」[1]と呼ぶことにする。

ただし、上述の転成方法はすべての動詞に適用するのではなく、例えば、「打つ」には「打ち」、「隠れる」には「隠れ」といった名詞はあまり見当たらない。動詞のうち、30%-40%しか名詞化しないとされる（西尾 1961、金 2003）。転成困難な語の中で、一部は前に内項（internal argument）を添加すると名詞化できるようになる[2]。どんな語が名詞化するか、どんな語が名詞化しないのか。また、どんなものが内項を付加することによって名詞化できるのか。名詞として継承されにくい動詞の意味特徴が存在するのか、といった問題が問われる。

また、連用形名詞が使用される際、普通名詞とは異なる独特な様相を呈している。加藤(1987)では、連用形名詞が「ガ格、ヲ格にたちにくい。むしろ「～だ」「～の N」「～に／で…する。」の「～」に現れることが多い」ことを指摘している。また、主格に立つ場合、後続「は」よりも後続「が」のほうが望ましい点が指摘された(西尾 1961)[3]。そして、主語、述語などの位置に立つ場合も、制約が多く、学習者の習得が難しい一面がうかがえる。例えば、

(2)　昨日の(＊働き/ ○仕事)は(＊疲れでした/ ○疲れた)。(玉村 1970)

(3)　それは体の疲れです。

(4)a. (＊走り/ ○ランニング)は体に良い。

b. 走ることは体に良い。

c. 彼は走りが速い。(影山 2011：51)

(4)b と(4)c が適格な文であるのに対して、(4)a の「走り」が単独で主題にたつ場合は非文となる。つまり、「走る」の名詞形「走り」は「彼」という人間の属性の一面を描写するこ

† shenxinchen@gmail.com

[1] この類の名詞について、山田(1936)で 23 例挙げたうえ、副語尾の付属されるものの連用形を以て名詞に転成することも少なくない と述べ、「居体言」と名付けた。ほかに、「転成名詞」と呼ぶ人もいる。

[2] 例えば、「値打ち」「本立て」などである。

[3] 例えば、「腐りが早い」「物覚えがよい」などである。

とが可能であるが、運動の一項目そのものを指示する働きを有していないようである。となると、動作・事態そのものがプロファイルされる場合でも、名詞化を経て、連用形名詞と「動詞+こと」は意味的に同じものではない。それはまた何を意味するものなのか、といった問題が興味深いである。

故に、本稿では、コーパスを利用して、連用形名詞の使用実態を調査し、その自立性により、自立できる語、文脈のサポートを必要とする語、複合語形式での名詞化を必要とする語という三段階に分け、連用形名詞の意味・構文的特徴を検討してみる。そして、動詞の意味とも関連付けて名詞化という過程を考えてみることにする。

2. 調査対象、方法と手順

2.1 調査対象

研究対象「連用形名詞」の規定について、本稿では基本的に西尾(1961)の立場に従う。つまり、「動詞連用形の種々な用法の中で、「（試験を）受けに行く（来る）・受けは（も・さえ等）する（しない）・お受けになる」などの〈受け〉のような用法は、連用修飾語をとることができ、多くの動詞に普遍的にみられる」用法であり、連用形名詞とは考えない。「しかし、「彼は（友人間の）受けがよい」の〈受け〉のような場合には、すでに連体修飾をとり得るから、名詞に転じたものとして、連用形名詞の範囲に含まれる。」また、通時的なものを入れると紛らわしいので、「霧」「境」「相撲」「歌舞伎」などのような語源的なものを考慮せずに、「現代日本の共通語で用いられる名詞のうち、動詞連用形から成り立ち、あるいは動詞連用形を含むもので、しかも普通の言語意識において、特別の知識なしに、多少の反省意識が働けばその語構造を把握しうるもの）」という規定に従う。

なお、本稿では単純和語動詞[4]に絞り、どんな動詞の連用形が名詞として成り立つか、また、成り立つとすれば、どういうふうに使われているのか、いわゆる自立性ということに注目しながら検討する。

2.2 調査方法と手順

本稿では、動詞の連用形が名詞として使用されるときの使用実態を知るために、コーパスを使用して調査することにする。

調査用コーパスは国立国語研究所の KOTONOHA「現代日本語書き言葉均衡コーパス」2008 年データ版を利用する。当該コーパスは書籍・白書・Yahoo!知恵袋及び国会会議録を収録した約 4490 万語規模のコーパスであり、現代日本語の実態を反映できると思われ、本稿には適切だと思われる。

調査手順として、まず形態素解析システム MeCab[5] でコーパスデータに対し形態素解析をし、解析後、品詞情報が標記される。そこで、名詞である語を収集する。次に、それらの名詞語彙を同解析システムの辞書 UniDic [6]に登録されている動詞連用形と対照し、一致

---

[4] ゆえに、「出会い」など（複合動詞「出会う」から転成されたと思われる名詞群）についての分析はまたの機会に譲る。ただし、「立ち読み」のような複合動詞形のない語は、「立ち」 ＋「読み」から複合されたと考え、後項の「読み」が本稿の言う連用形名詞の検討範囲に入る。

[5] MeCab は京都大学情報学研究科–日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンである。詳しくは http://mecab.sourceforge.net/を参照されたい。

[6] UniDic は日本語テキストを単語に分割し，形態論情報を付与するための電子化辞書である。形態素解析器「茶筌(ChaSen)」、「和布蕪(MeCab)」の辞書として利用できる。詳しくは http://www.tokuteicorpus.jp/dist/

したものを抽出する。つまり、
条件①品詞情報：名詞であること
条件②語彙形式：動詞連用形であること
という二つの条件を満たしたものを、調査材料として抽出した。

3. 連用形名詞の自立性の段階

3.1　三段階

連用形名詞のうち、「ガ格」、「ヲ格」などの格に立ち、普通名詞と同じように機能する語、いわゆる最も自立性の高い類から、構文上なんらかの構文パターンの助けを借りて成立する類や、語彙的に項や付加詞のサポートを必要とする、複合語でしか成立できない類など、多様性が見られる。また、そういった自立性の強弱の連続相から、元の動詞の意味とのかかわりが想像されうる。

コーパスより抽出された連用形名詞及び例文をもとに、手作業によって整理してみた。ここで連用形の名詞としての成立可否及び自立度により 3 分類してみた。

まず、基準 A（動詞の連用形が単独で名詞として成立するか否か）により、「遊び」「付き」のような単独名詞形の持つものと、「（海岸）沿い」「（ゴミ）出し」「(びしょ）濡れ」などのような複合名詞として使用するものに大別する。

そして、単独名詞形のあるものの中で、基準 B（特定文脈のサポートが必要か否か）により、タイプ 1「自立タイプ」と、タイプ 2「構文補助タイプ」とに分ける。「遊び」「騒ぎ」などの語はタイプ 1 に属し、「お客様の受けがよい」「ボールの転がりがよい」などの語はタイプ 2 に入る。

上述のことを図で示すと、図 1 のようになる。

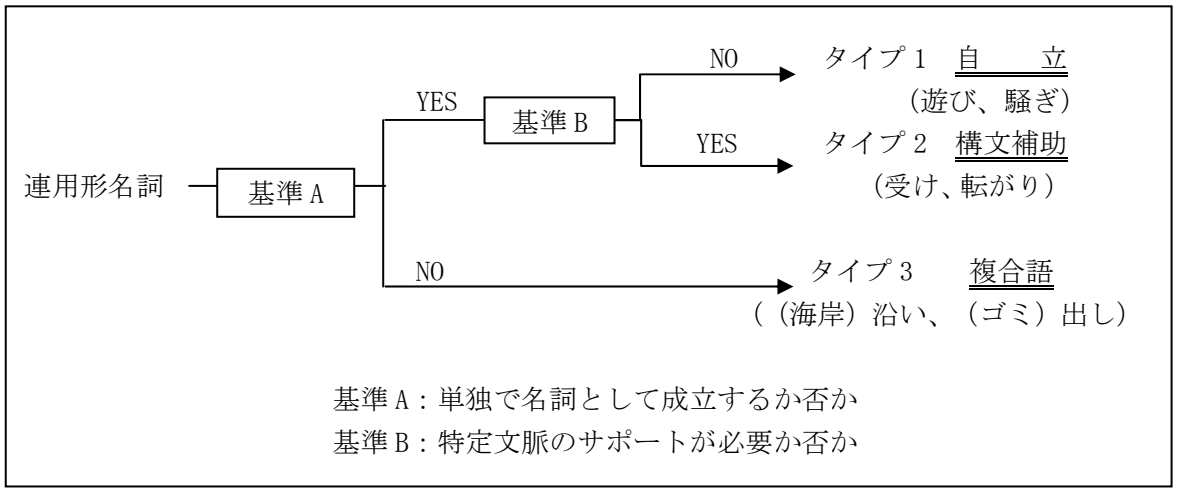


図 1　自立性による動詞連用形名詞の分類

タイプ 1 は単独使用可能な語であり、タイプ 2 は特定文脈の助けを必要とするタイプである。タイプ 3 となると複合形式でしか使用できず、タイプ 1 からタイプ 3 へ、連用形自体が一語としての自立度がどんどん落ちていっているように思われる。動詞を自他動詞に

を参照されたい。

分け、タイプ別に語例を示してみると表 1 になる。

表1　連用形名詞自立性の全体像

	他動詞	自動詞	自立性
タイプ 1 自立	扱い　祈り　戒め　占い　抑え　飾り　考え　調べ　咎め　狙い　始め　守り　報い　など　(104 語)	遊び　踊り　泳ぎ　帰り　重なり　答え　騒ぎ　戦い　疲れ　粘り　残り　笑い　など　(133 語)	高 ↓ 低
タイプ 2 構文補助	受け　下げ　作り　彫り　焼き　磨き　(29 語)	当たり　乾き　転がり　進み　高まり　粘り　伸び　広がり　増え　降り　深まり　減り　(50 語)	
タイプ 3 複合語	(資金) 集め　(毛糸) 編み　(答) 打ち　(メイク) 落とし　(結納) 返し　(証拠) 固め　(値) 決め　(ゴミ) 出し　(ろうそく) 立て　(値) 付け　(ライバル) 潰し　(子供) 連れ　(井戸) 掘り　(家庭) 持ち　(段階) 分け　など　(81 語)	(週・梅雨) 明け　(ホテル) 宛て　(梅雨) 入り　(雲) 隠れ　(八時) 過ぎ　(使用) 済み　(海岸) 沿い　(責任) 逃れ　(人間) 離れ　(ため息) 混じり　(京都) 寄り　など　(50 語)	

3.2 自立タイプ

まず、連用形名詞のうち、「ガ格」「二格」「ヲ格」「デ格」など、様々な格にたつことができ、機能的に普通名詞と似ている類がある。例えば、

(5) a.自動詞

焦り　怒り　痛み　恐れ　驚き　渇き　苦しみ　痺れ　疲れ　慰み　悩み　匂い　励み　誇り　迷い　喜び　遊び　甘え　憩い　飢え　動き　頷き　踊り　泳ぎ　香り　輝き　通い　きらめき　暮らし　叫び　囁き　騒ぎ　戦い　泊まり　流れ　嘆き　鳴り　眠り　登り　働き　響き　瞬き　休み　揺れ　酔い　詫び　空き　集まり　行き　帰り　重なり　固まり　勝ち　曇り　焦げ　染み　ずれ　つながり　並び　濁り　禿げ　外れ　晴れ　負け　乱れ　戻り　破れ　歪み　汚れ　など

b.他動詞

商い　味わい　扱い　誤り　争い　いじめ　偽り　営み　祈り　戒め　彩り　祝い　疑い　写し　訴え　占い　恨み　売り　奢り　教え　脅し　買い　囲い　囲み　賭

け　飾り　語り　悲しみ　借り　狩り　考え　悔い　企て　試み　断り　探り　支え　誘い　定め　悟り　裁き　妨げ　忍び　知らせ　調べ　救い　勧め　責め　攻め　競り　備え　蓄え　貯え　助け　楽しみ　頼み　試し　伝え　慎み　包み　綴り　務め　勤め　繋ぎ　釣り　問い　咎め　届け　流し　眺め　慰め　習い　習わし　盗み　願い　狙い　覗き　望み　呪い　始め　含み　祭り　まとめ　学び　招き　守り　迎え　報い　恵み　儲け　求め　許し　装い　読み　など

(5)の語例を観察すれば分かるように、「恨む」「祈る」「疑う」「悲しむ」などの心理・感情を表すものや、「問う」「語る」「知らせる」「訴える」「調べる」「眺める」など、人間の認識活動・言語活動・表現活動を表すものの連用形が多く見られた。

ただし、動詞からの転成であるゆえ、動詞の意味が受け継がれ、後続する動詞・形容詞などに一定の偏りを示している点では特徴的である。そのうちで、高頻度の「動き」を取り上げてみる。次の表は「動き」のコロケーションパターンである。

表 2　「動きを/が/に」のコロケーションパターン

	コロケーションパターン(用例数)	用例数
～を	見る(271)、する(266)、止める・とめる(207)、示す(125)、見せる・みせる(116)、追う(39)、封じる(33)、知る(29)、観察する(23)、読む(21)…	2709
～が	ある(270)、見られる(169)、出る (120)、止まる(69)、鈍い(59)、強まる(56)、取れる(43)、速い・早い(43)、できる(41)、悪い(39)、激しい(36)、見える(35)、始まる(30)、ない(26)、広がる(26)、良い(26)、分かる(25)、続く(24)、起こる(23)、高まる(21)、遅い(21)…	2199
～に	あわせる(75)、なる(57)、対応する(28)、注目する(20)、注意する(10)、呼応する(10)、反応する(10)…	744

表 2 を通して、以下の事実が確認できる。

（一）「が」格に立つ場合、述語の部分として、「ある」「見られる」「出る」「止まる」「できる」「鈍い」「速い・早い」「ない」などが多く現れる。いわゆる存在詞述語文、存在・必要・充足の動詞述語文、非存在・存在量の形容詞述語文が多い。そして、名詞述語文はほとんどない。

動詞は本来動作・属性を表わすものである。「動き」などの動詞が名詞化されてできた連用形名詞は、動詞の意味素性を受け継いでいる。故に、連用形名詞を主格・主題に据え、認識の中核として認識する際、具体的な「もの」としてではなく、動作・属性の抽象化されたものとしてとらえているのである。それ故、その動作・属性のあり方として、後ろの述語に存在・必要・充足・程度などを表わす語が多くくるだろう。

（二）「ヲ格」では、「する」「止める」「示す」「見せる」など、機能動詞といわれるものが現れる。また、ほかに、「見る」「追う」「知る」「観察する」「読む」ほとんど人間の認識活動を表す動詞がある。

3.3　構文補助タイプ

このタイプはさらにいくつかの種類に分けられる。

A　「主体＋の～」構文が必要なもの

B　「(は/の)～が…」「様態～」構文が必要なもの
C　イディオム

A　「主体＋の～」構文が必要なもの

動詞には、対象物の存否を問わず、主体が必ず関わるわけなので、名詞化された後、「主体＋の＋V 連用形」という形式も一般的に存在するかのように思われる。

ただし、肝心なところは、その「主体＋の」の部分を取れば、文として成立可能かどうかである。例えば、

(6)a.　さらにそれは国内の反戦運動の高まりと、経済の弱体化を招く結果となった。
（『わかりやすいベトナム戦争』）

b. ×さらにそれは国内の高まりと、経済の弱体化を招く結果となった。

(7)a.　この訓示が出ると、現役軍隊の騒ぎはぴったり治まった。
（『「文芸春秋」にみる昭和史』）

b. ○この訓示が出ると、騒ぎはぴったり治まった。

上述の(6)と(7)から分かるように、文を組み立てるとき、「高まり」は必ず「主体＋の」の提示を必要とし、「主体＋の」を削除すると、文としては成立しがたい。それに対して、「騒ぎ」は「主体＋の」提示がなくても、文として成立できる[7]。

そのような語はさらに、「崩れ、高まり、伸び、広がり、増え、深まり、減り、下げ」などがある。

B　「(は/の)～が…」「様態～」構文が必要なもの

連用形名詞のうち、(3)のような「動作・作用のありさま・方法・程度・具合・感じなど」の意味を表すものが多いと指摘されている。

(8)金使い(が荒い)・滑り(がいい)・売れ行き(がすごい)・出来(米の一)・当たり(が柔らかい)　（西尾 1961）

コーパスを調べると、次のようなものが見られる。

(9)a. 自動詞

当たりが強い、お客様の受けがよい、テレビの映りが悪い、写真の写りが悪い、収まりが良い、エンジンの掛かりが遅い、ナイロンのシャツは乾きが速い、この薬は効きが速い、聞こえがいい、包丁がなまって切れが悪い、腐りが早い、ボールの転がりがよい、触りが柔らかい、窓の締まりが悪い、工事の進みが速い、ふすまの滑りを良くする、すわりが悪い、色の染まりが悪い、客の付きが悪い、脳内神経の伝わりが良くなる、今年は米の出来がいい、水の出がいい、声の通りがいい、傷の治りが遅い、粘りがある、化粧ののりが悪い、話の運びがうまい、ボールの弾みが悪い、張りのある声、窓の開きが悪い、頭の回りが早い、血の巡りが悪い、たき火の燃えが悪い、バッテリーの持ちがよい、彼は分かりが速い

b. 他動詞

削りが粗い、刷りが美しい、彫りが深い、焼きが悪い/甘い、巻きが強い、作りがよい、織りが粗い、大きい生地の中央での縫いが簡単、縒りの甘い糸、あの食堂は盛りがいい

[7] 関連する概念として、西山（1990）では「非飽和名詞」、影山(2011：225-231)では「相対名詞」という概念を提起し述べている。

バットの振りが鈍い、体のこなしが柔らかい
押しが強い、魚の引きが強い、魚の食いが悪い
覚えが早い

(9)の例は「ガ構文」のサポートの下で名詞として成立できる。その類は、構文上、自立タイプの使い方と違う。例えば、

(10)〇a. 言葉の遊びをする
×b. 包丁がこんな切れをする。
〇c. そのような騒ぎがあった。
×d. そのような受けがあった。

「遊び」、「騒ぎ」が「をする」・「がある」などの構文にたてるのに対し、「切れ」「受け」などは、「ヲ格」、「ガ格＋動詞」など、自由に文の各位置にたつことは難しく、「…ノ(ハ)～ガ＋形容詞(形容動詞)」という特定の構文パターンで使われることが多い。前の「…ノ(ハ)」の部分は動作の行われる主体であり、後ろに来るものとして、「良い」「悪い」「速い」「遅い」「著しい」などの形容詞が多い。

また、「回」など、デキゴトの助数詞も使えない。

(11)　？　シャツの一回の乾き

つまり、この類の連用形名詞は一つの事件・デキゴト・動作を表しているのではなく、動作の一側面、状態・様態・様子を描いているわけである。

ただし、自立の名詞でもこの用法が可能である。例えば、

(12)a 手は動きを止めていた。
b. 考えると却って動きが鈍くなる。

(12)a は自立の名詞用法であるが、b の方は「ガ構文」を用いて、動詞のあり方・様態を描写している。

C　イディオム

(13)押さえがきかない、潰しが効く、泣きを入れる、逃げも隠れもしない、磨きをかける

3.4　内項・付加詞複合語タイプ

構文上の補足が無理で、必ず複合語形式で名詞化を要求する語がある。例えば、

(14)a.　（値）上がり　（ホテル）宛て　（梅雨）入り　（色）移り　（雲）隠れ　（冬）枯れ　（税）込み　（値）下がり　（八重）咲き　（湯)冷め　（戸)閉まり　（八時）過ぎ　（使用）済み　（計画)倒れ　（高)止まり　（責任）逃れ　（親）離れ　（ため息）混じり　（雨）漏り　（日）焼け　（小)止み　（「オッサン」）呼ばわり　（長）生き　（早）起き　（立ち)消え　（行き）来　（置き)去り　（若）死に　（大)助かり　（逆）立ち　（丸)潰れ　（生）煮え　（びしょ）濡れ　（昼）寝　（中年）太り　（まる）見え　（夏）痩せ　など

b.（旗）上げ　（さつま）揚げ　（資金）集め　（毛糸)編み　（数字）合わせ　（名刺）入れ　（田)植え　（様子）伺い　（口）移し　（穴)埋め　（火）起こし　（メイク）落とし　（荷)降ろし　（雪）下ろし　（水）換え　（席）替え　（真相）隠し　（新聞）掛け　（証拠）固め　（芝）刈り　（値）決め　（ブロック）崩し　（臭い）消し　（頭）越し　（湯)冷まし　（コマ）ずらし　（髪）染め　（ゴミ）出し　（ろうそく）立て　（２階）建て　（根)絶やし　（値）付け　（塩）漬け　（通行）止め　（カビ）取り　（自分）撮り　（裾）直し　（五目）並べ

（味噌）煮　（皮）剥ぎ　（日本）外し　（井戸）掘り　（ごちゃ）混ぜ　（茶碗）蒸し　（ルール）破り　など

例(14)から分かるように、「入れる」「埋める」「換える」「崩す」「取る」など、対象物に具体的な動作を与え、対象物に位置・形態の変化をもたらしたものが多い。それが自立タイプの語と良い対照をなす。

上述のことをまとめて、連用形名詞は自立度に沿って連続相をなしていることが分かる。具体的には次の図のようになる。

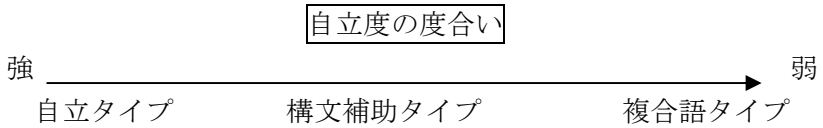


図 2　連用形名詞自立度の連続相

4. 終わりに

本稿では、日本語単純和語動詞の連用形名詞を研究対象とし、従来あまり注目していなかった自立度の問題に焦点をあて、連用形名詞を自立タイプ、構文補助タイプ、複合語タイプに分け、それぞれの類の特徴を観察してみた。

名詞化には「実体化」という働きを持つとされる(池上 1978)。動詞から名詞への転成において、音節といった要因のほかに、元の動詞の意味自体が制約を加えることが予想される。動詞の典型的要素、いわゆる他動性の構成要素、動作性・瞬時性・影響性などが名詞化過程で影響を与えていると思われるが、その影響の優先順位はどうなっているかなどの問題について課題としておく。

文　献

Paul J. Hopper and Sandra A.Thompson(1980)Transitivity in Grammar and Discourse. *Language*Vol. 56, No. 2.

池上嘉彦(1978)『意味の世界―現代言語学から視る―』日本放送出版協会

岡村正章(1995)「「典型的な動詞連用形名詞」に関する一考察」 上智大学国文学論集 28

影山太郎 (1999) 『形態論と意味』くろしお出版

金美淑(2003)「連用形名詞」『日本語論究 7』、pp.299-320、和泉書院

金美淑(2007)「日本語の連用形名詞」名古屋大学大学院文学研究科博士論文

西尾寅弥(1961)「動詞連用形の名詞化に関する一考察」『国語学』43

日本語話し言葉コーパスを用いた対話音声のイントネーション句の分析

石本 祐一（国立国語研究所 言語資源研究系）†
小磯 花絵（国立国語研究所 理論・構造研究系）

F0 Characteristics at the Level of Intonational Phrase in Japanese Dialogs: Analysis of the CSJ

Yuichi Ishimoto (Dept. Corpus Studies, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

1. はじめに

自発音声の発話の韻律的特徴を探るため，小磯・石本（2012）および石本・小磯（2012, 2013）では，『日本語話し言葉コーパス』を対象にイントネーション句（IP）を単位として独話音声の基本周波数（F0）の変動を調べた。その結果，自発性の高い独話において

- IP 単位の F0 最大値・最小値が発話内で徐々に下降する
- 発話の長さ（IP 数）に関わらず，IP はほぼ一定の高さで始まり一定の高さで終わる（発話の長さによって F0 下降の傾きが異なる）
- ただし発話が長い場合，若干高い F0 で発話を開始する
- 発話末に著しい F0 下降がみられる（final lowering）

という傾向がみられた。また，特に発話中に強い統語的境界が存在する場合（二つ以上の節で発話が構成される場合）は

- 発話中の節境界で IP の F0 下降傾向が途切れてリセットされる
- 強い統語境界では F0 最小値は発話末のレベルにまで達せず，final lowering に相当する著しい F0 下降はみられない

という現象がみられることがわかった。発話全体にみられる IP を単位とした F0 下降現象は，発話の長さによって F0 下降の傾きが異なることから，Pierrehumbert & Beckman (1988) などが指摘するような単純な F0 declination（発話に要する時間の関数として単純に F0 が低下する現象）とは考えにくく，自発発話を観察することによって得られた知見と言える。

上記の傾向は，学会講演・模擬講演という自発性の高い独話音声でみられたものであるが，自発性の高い対話音声においても同様の F0 下降現象とその発話中のリセットが観察されるのかはまだ明らかになっていない。本研究では，対話音声におけるイントネーション句単位の F0 推移を調べ，独話と同様の現象が対話でも現れるのか，それとも対話独自の現象が観察されるのかについて検討する。

† yishi@ninjal.ac.jp

2. データ

2.1 発話単位

『日本語話し言葉コーパス（*Corpus of Spontaneous Japanese*:以下 CSJ）』（前川 2004）に収録されているインタビュー形式の対話と課題指向対話を分析対象とした。CSJ 第 3 刷に基づき作成された RDB（小磯ほか 2012）を用い，「コア」と呼ばれるデータ範囲中の 18 対話から，後述の韻律情報が付されたインタビュイーの発話[*1]（約 220 分）を選択して分析した。

発話単位の認定にあたって，CSJ に付与されている節単位情報（丸山ほか 2006）を利用した。節単位は原則「節（clause）」の境界によって得られる文法的・意味的なまとまりを持った単位であり，節境界の構造的な切れ目の大きさの観点から以下の 3 つに分類される。

絶対境界（Absolute boundary）　「〜です」「〜ます」などのいわゆる文末に相当する境界。
強境界（Strong boundary）　「〜けど」「〜が」などの後続の節に対する従属度の低い，切れ目の度合が強い節境界。
弱境界（Weak boundary）　「〜から」「〜で」などの後続の節に対する従属度の高い，切れ目の度合が弱い節境界。

これらの境界は形態素解析結果に基づき自動で判別され，人手による修正・操作を経た上で，絶対境界，強境界のいずれかで区切られる単位が節単位と認定されている。本研究では，絶対境界で区切られる区間を発話に相当する単位として扱う[*2]。

2.2 イントネーション句の F0 特徴量

本研究では IP を単位とした発話の F0 の特徴を探る。

アクセント句 (AP) は，第 1 モーラから第 2 モーラ付近にかけての F0 の上昇と句末への緩やかな下降を有し，かつアクセント核による下降を最大ひとつ持ちうる単位と定義される。イントネーション句 (IP) は AP の上位階層に位置し，AP のピッチレンジを指定する単位と定義される。アクセント核が引き起こす後続 AP のピッチレンジの縮小効果は，IP の範囲で観察される。CSJ にはラベリングスキーム `X-JToBI`（五十嵐ほか 2006）に基づき韻律情報が付与されているが，この中に韻律境界の切れ目の強さに関する情報として Break Index（BI）が存在する。BI=2 は AP 境界，BI=3 は IP 境界，BI=F はフィラー境界，BI=D は言い淀み境界に対応する。そこで，本研究では BI=3 で区切られる範囲を IP と認定し，フィラー，言い淀み部分を除いて分析に用いた。ただし，フィラーを狭んでダウンステップが続く場合はフィラーを内包する形で IP を認定した[*3]。

このように IP を認定した上で，`X-JToBI` に基づく Tone 情報から IP の F0 特徴量として，

F0 最大値　IP 頭の AP の句頭音調（`H-`）あるいはアクセント核（`A`）のうち高い方の F0 値
F0 最小値　IP 末尾の AP の下降音調（`L%`）の F0 値

[*1] 課題指向対話の場合はインタビュー対話でインタビュイーの役割だった話者の発話。
[*2] 明示的な文末表現が置かれるもののほか，「と文末」や「体言止め」なども含む。
[*3] CSJ 第 3 刷に基づき作成された RDB にはアクセント句・イントネーション句に関する情報が含まれており，フィラーや言い淀みの扱いが上記定義と一部異なるが，独話を対象とした小磯・石本（2012）および石本・小磯（2012, 2013）と合わせるため，本研究ではこの定義を採用した。

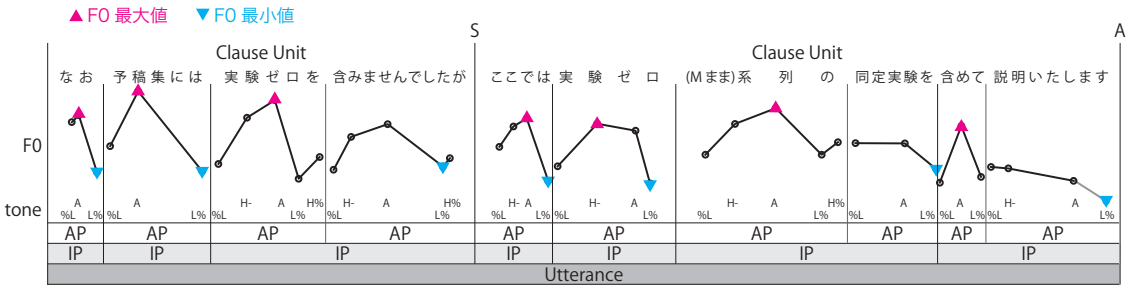


図 1　IP 単位の F0 最大値・最小値

図 2　分析 1 の対象とする節単位（S:強境界，A:絶対境界）

を求めた（図 1）。分析において性差・個人差の影響を小さくするために，F0 値は話者ごとの平均 F0・標準偏差によって Z スコアへ変換し標準化を行っている。

3.　分析 1: 発話内部に強い統語境界を持つ場合の F0 推移

3.1　方法

発話内部に少なくともひとつの強い統語境界を持つ場合における，IP 単位の F0 推移を観察する。具体的には，強境界を末尾に持つ節単位と絶対境界を末尾に持つ節単位の二つの連鎖を分析対象とする（図 2）。データ数は強境界と絶対境界の組合せで 95 件である。

自発性の高い独話を対象とした先行研究（小磯・石本 2012，石本・小磯 2012, 2013）において，節単位内では IP 単位での F0 の下降現象がみられるとともに，強い統語境界ではその下降が途切れてリセットされ，次の節で新たな下降の流れが現れていた。また，発話末尾，すなわち絶対境界直前では final lowering とみられる F0 の急激な下降がみられた。対話においても同様の現象がみられるのかどうかがこの分析の着目点である。

3.2　結果と考察

発話中の IP の F0 最大値の推移を図 3 に，F0 最小値の推移を図 4 に示す。節単位最初の IP，節単位最後の IP，それ以外（中間）の IP に区分している。

それぞれの図から，先行節単位と後続節単位の範囲で F0 の下降がみられるのに対し，強境界で F0 下降が途切れリセットされていることがわかる。これは独話においてみられた現象と類似しており，対話においても強い統語境界で IP 単位の F0 下降がリセットされるといえる。一方，独話においてはっきりとみられていた後続節単位末尾（発話末尾）の F0 の急激な下降が，図 4 においては現れていない。この結果は，対話音声においては final lowering が現れないことを示唆する。

この結果をさらに詳細にみるために，節境界の種類ごとに節境界直後の IP の F0 最大値をま

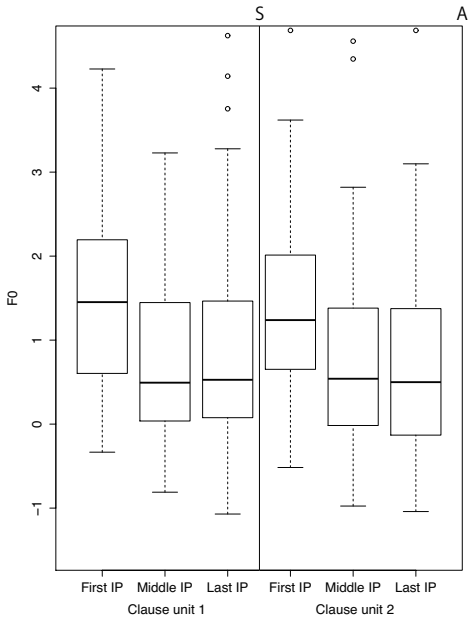


図 3　強境界を内部に持つ発話における IP 単位の F0 最大値の推移

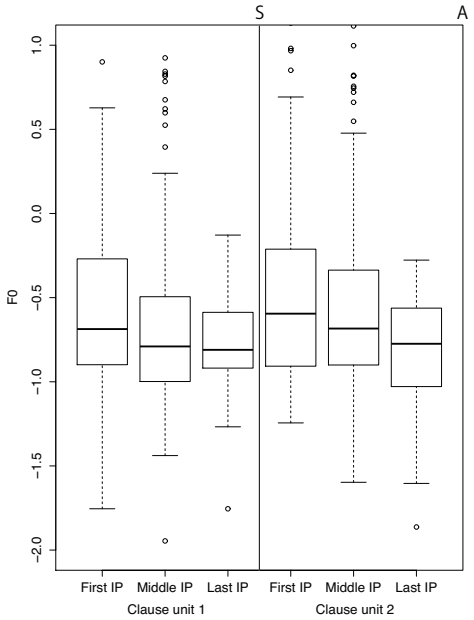


図 4　強境界を内部に持つ発話における IP 単位の F0 最小値の推移

とめたものを図 5 に，節境界直前の IP の F0 最小値を図 6 に示す。比較のため，節境界ではない（非境界 (E) の）IP の F0 最大値・最小値も求めた。なお，対話でよく現れる非常に短い発話，例えば，同意を表す「そう」「そうそう」「そうですね」などを除くために，AP を 1 つしか含まない IP は除いている。

図 5 をみると，非境界直後では F0 最大値が低いことから F0 下降のリセットは生じていないと考えられる。絶対境界直後や強境界直後で F0 最大値が高いのはそれらの境界で F0 下降がリセットされるためであり，図 3 の結果と整合している。一方，弱境界直後では強境界や絶対境界と同程度の F0 最大値となっており，弱境界でも F0 下降のリセットが生じる可能性がある。弱境界が F0 推移に与える影響については今後のさらなる検討が必要である。

図 6 からは，統語的な切れ目が強くなるほど F0 最小値が低くなる傾向がみられる。しかし，独話でみられたような絶対境界直前の F0 最小値が飛び抜けて小さくなるという傾向はみられず，強境界との差はわずかである。これは，final lowering に相当する F0 の急激な低下がみられなかった図 4 の結果を裏付けている。

何故，対話音声では絶対境界における F0 最小値の急激な低下が観察されないのだろうか。一つの可能性として，対話と独話の句末音調の出現率の違いが考えられる。対話では，「〜でしたね」「〜ですよね」のように終助詞「ね」が末尾につくなどして，句末の音調が上昇調や上昇下降調になりやすい（表 1)。その違いが結果に影響した可能性がある。そこで，句末音調ごとに F0 最小値を求めて独話と対話で比較したところ，独話では下降調の絶対境界 (A) で F0 最小値が極端に低くなるのに対し，対話では下降調でも F0 最小値は上昇調や上昇下降調の場合とほとんど変わらず，独話ほど低くならない（表 2)。つまり，対話では上昇調や上昇下降調が多いために絶対境界で F0 最小値の急激な低下がみられないのではなく，下降調の場合に独

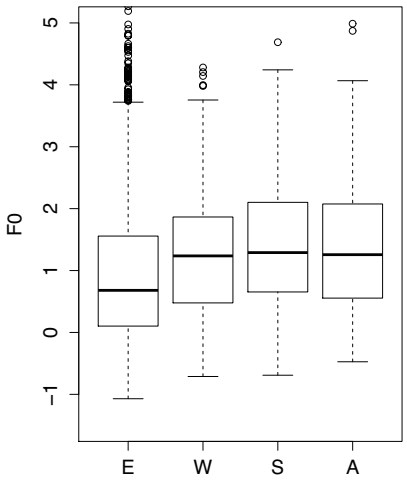


図 5　節境界とその直後の F0 最大値の関係
（E:非境界，W:弱境界，S:強境界，A:絶対境界）

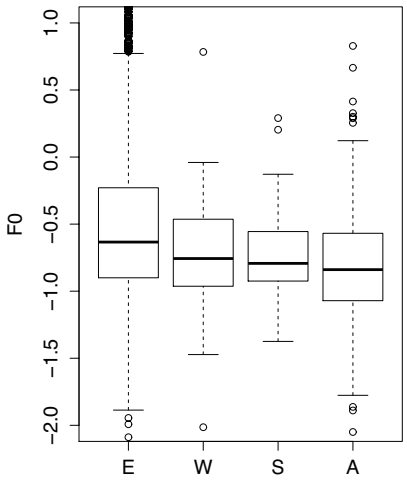


図 6　節境界とその直前の F0 最小値の関係
（E:非境界，W:弱境界，S:強境界，A:絶対境界）

表 1　対話音声の BPM

	H%	HL%	L%
絶対境界 (A)	121	36	135
強境界 (S)	96	54	24

表 2　独話と対話の句末音調と F0 最小値

		絶対境界 (A)	強境界 (S)
独話	下降調	-1.367	-0.906
	上昇調	-1.106	-0.922
	上昇下降調	-0.824	-0.858
対話	下降調	-0.737	-0.717
	上昇調	-0.814	-0.728
	上昇下降調	-0.755	-0.736

話ほど F0 が下がり切らないということであり，少なくとも今回分析対象とした対話では final lowering はあまり生じないと結論付けることができる。

Umeda (1982) は，F0 下降はスタイルに依存しており，final lowering は朗読音声に限られる可能性を示唆している。前川（2013）は，CSJ のコア（対話も一部含むが大半は独話）を対象に諸条件を綿密に統制した分析を行い，final lowering が自発音声にも観察されることを明らかにしたが，スタイルの異なる対話において final lowering が生じない可能性は十分にある。final lowering の役割は発話の終了を表示することである。CSJ に含まれる独話（講演）では，聞き手に対して発話や談話の構造や切れ目が明確に伝わるよう，言語表現や韻律を調整している可能性が高い。final lowering もその一つと考えられる。一方，対話においては，聞き手と発話のターンを交替しながら話を進めるため，発話や談話の構造や切れ目を独話ほど明確に伝える必要が無い可能性がある。また，事前の発話計画が可能な講演とは異なり，対話では相手とのやりとりの中で発話を組み立てる必要があり，そのような調整をするだけの認知的余裕がないとも考えられる。あるいは，発話のターンを保持するためにターン途中の絶対境界では発話の終了性を表示する final lowering を抑制し，ターンの終了時にのみ final lowering が生じている可能性もある。特に今回分析対象とした 18 対話のうち 12 対話はインタビューであり，インタビュイーのターンは長くなる傾向がある。ターン途中の final lowering が抑制された場合，

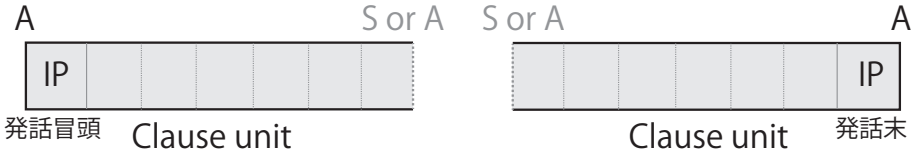


図 7　分析 2 の対象とする節単位（S:強境界，A:絶対境界）

表 3　節単位中の IP 数

IP 数	1	2	3	4	5	6	7	8	9	10〜
発話冒頭	79	68	54	23	19	20	12	7	5	7
発話末	350	126	78	47	38	24	20	14	9	17

今回のような結果が導かれる可能性もある。この点については今後の課題としたい。

4.　分析 2: 発話の長さと発話冒頭の F0 最大値・発話末の F0 最小値

4.1　方法

本節では，発話の長さと発話冒頭の F0 最大値・発話末の F0 最小値との関係について分析を行う。独話においては，発話冒頭の F0 最大値は発話が長くなるほどわずかながら高くなり，発話末の F0 最小値は発話の長さに関わらずほぼ一定となっていた（小磯・石本 2012）。すなわち，F0 の下限は決まっており，長い発話では高い F0 で話し始めることで発話中の F0 下降によって発話末の時点で F0 の下限を下回ることのないようにする配慮を行っていると考えられる。このような傾向が対話でもみられるかどうかがこの分析の着目点である。

ここでは，節単位中の IP 数を発話の長さとみなすことにする。ただし，小磯・石本（2012）では内部に強い統語境界を持たない発話に限定した分析を行っていたが，対話においては同条件では十分なサンプル数が得られないため，絶対境界直後の節単位を発話冒頭，絶対境界直前の節単位を発話末としてそれぞれの節単位について IP 数と F0 最大値・最小値の関係を調べた（図 7）。発話冒頭および発話末の節単位中の IP 数を表 3 に示す。サンプル数が 10 以上のものに限定するために，分析する発話中の IP 数は 7 以下とした。

4.2　結果と考察

節単位中の IP 数と発話冒頭の F0 最大値との関係を図 8 に，発話末の F0 最小値との関係を図 9 に示す。

図 8 にみられるように，節単位中の IP 数が多ければ，すなわち，発話が長ければ F0 最大値は高くなる傾向があるが，独話ほど明瞭に単調増加をしておらずかなりばらついている。これにはサンプル数の少なさが要因として考えられるが，その他の理由として対話の自発性の高さが影響を与えている可能性がある。学会講演や模擬講演では自発発話とはいえ事前の発話の計画が比較的容易であるのに対し，対話では話し始めの時点で十分に自己の発話を計画できず，どの程度の長さの発話をするかが不確かなため発話冒頭の高さを定めにくいと考えられる。

次に図 9 をみると，発話末の F0 最小値は発話の長さによらずほぼ同じ値になっており，独

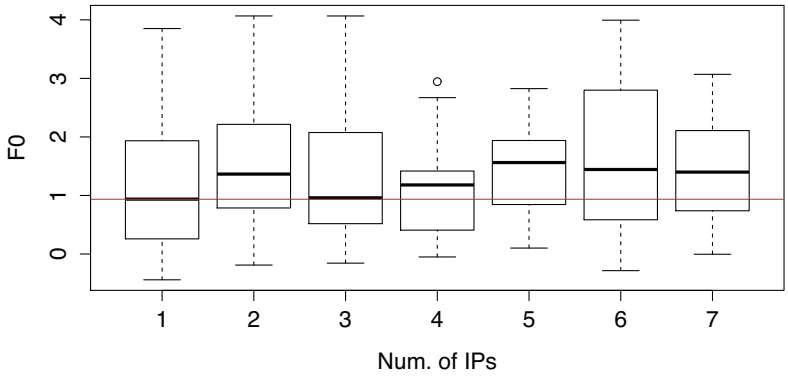


図 8　発話冒頭の F0 最大値と節単位の IP 数

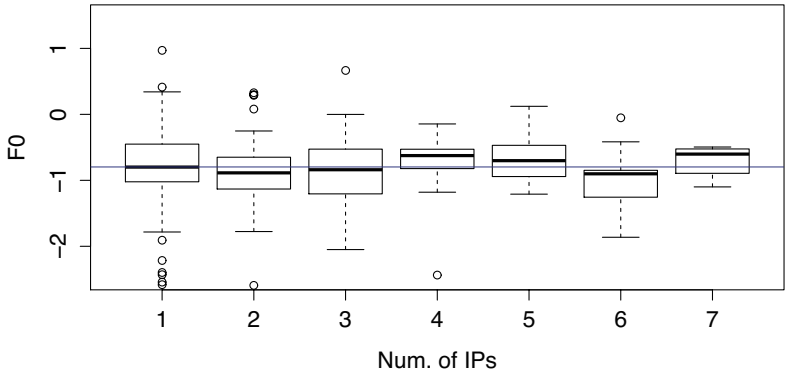


図 9　発話末の F0 最小値と節単位の IP 数

話と同じ傾向を示しているといえる。しかし，独話と比較して F0 値がそれほど低くなっていない。表 2 に示したように，独話では発話末（絶対境界前）の F0 最小値は-1.3 程度になるのに対し対話では-0.7 程度であり，発話の長さによってもそれは変わらない。すなわち，対話の発話末ではある決まった値までは F0 が下降するものの，final lowering といえるほど大幅な低下をみせないということであり，前節の結果と整合的である。

5. おわりに

本稿では，独話でみられた IP 単位の F0 下降現象が対話においてもみられるのかを調べるために分析を行った。その結果，独話と同様に対話でも発話内で IP 単位の F0 が下降する傾向にあること，強い統語境界でその下降がリセットされることがわかった。しかし，final lowring に相当する発話末の急激な F0 下降が対話ではみられず，独話との大きな違いとして現れた。また，発話の長さと発話冒頭・発話末の F0 の関係を調べたところ，発話が長ければ冒頭は高い F0 になる傾向にあるが独話ほど明確に発話の長さと F0 の関係が現れず，また発話末の F0 最小値も発話の長さによらず一定の値になるが独話ほど低い値にならないことがわかった。今後，主に final lowering に関わる部分でみられた独話と対話の違いについて，今回着目した以

外の要因を考慮して調べる必要がある。

参 考 文 献

五十嵐陽介，菊池英明，前川喜久雄 (2006) 「韻律情報」日本語話し言葉コーパスの構築法（国立国語研究所報告 124), pp. 347–453.

石本祐一，小磯花絵 (2012) 「日本語話し言葉コーパスを用いた統語境界におけるイントネーション句変動の分析」第 2 回コーパス日本語学ワークショップ予稿集, pp. 239–246.

石本祐一，小磯花絵 (2013) 「自発発話におけるイントネーション句単位の F0 変動の特徴」第 3 回コーパス日本語学ワークショップ予稿集, pp. 333–342.

小磯花絵，石本祐一 (2012) 「日本語話し言葉コーパスを用いた「発話」の韻律的特徴の分析 –イントネーション句を切り口として–」第 1 回コーパス日本語学ワークショップ予稿集, pp. 167–176.

小磯花絵，伝康晴，前川喜久雄 (2012) 「『日本語話し言葉コーパス』RDB の構築」第 1 回コーパス日本語学ワークショップ予稿集, pp. 393–400.

前川喜久雄 (2004) 「『日本語話し言葉コーパス』の概要」日本語科学, 15, pp. 111–133.

前川喜久雄 (2013) 「日本語自発音声における final lowering の生起領域」『第 27 回音声学会全国大会予稿集』.

丸山岳彦，高梨克也，内元清貴 (2006) 「節単位情報」日本語話し言葉コーパスの構築法（国立国語研究所報告 124), pp. 255–322.

Pierrehumbert, Janet B. and Mary E. Beckman (1988) *Japanese tone structure*, Cambridge: MIT Press.

Umeda, Noriko (1988) “F0 declination is situation dependent” *Journal of Phonetics* 10, pp.279–290.

※ 本研究は萌芽・発掘型共同研究「会話の韻律機能に関する実証的研究」（リーダー：小磯花絵）による成果である。

口頭発表（2）

9月5日（木）　15:10～17:10

中古における接続語の使用傾向について

岡﨑　友子（東洋大学）

The Usage of Conjunction in Early Middle Japanese

Tomoko Okazaki (Toyo University)

１．はじめに

先行研究で指摘されるように（京極・松井（1973）他）日本語には固有の接続詞がなく、その本格的な発達は中世以降となる。しかし、中古においても連語の域を超えて複合語化し、接続詞的に用いられているのではないかと考えられるものもある。

そこで本発表では「日本語歴史コーパス」を利用し、中古における接続表現（これらを「接続語」と呼ぶ）について、１）どのような接続語が用いられているのか、２）それらの接続語について、コーパスの情報である作品別・本文種別等から分析すると、それぞれどのような使用傾向が見られるのかについて報告をおこなっていく。

２．先行研究

京極・松井（1973）で述べるように、古代語において和語系の接続詞を認めることは難しいが、「その変遷を考察するにあたっては、接続詞的はたらきをすると認められる語句を広く対象とするべきであろう」（91 頁）ともするように、日本語の歴史において接続詞を考えるためには、中古における接続語を丁寧に調査する必要がある。

そこでまず、京極・松井（1973）の示す古代語の接続詞をまとめておく。

A 複合接続詞

（a）　指示語を構成要素とするもの

①「か」系　○「かく」類　かくあるほどに等　○「かかり」類　かかりければ等
　　○「かり」類　かれ　かるがゆゑに

②「こ」系　○「ここ」類　ここに等　○「これ」類　これによりて等

③「さ」系　○「さ」類　さいふいふ等　○「さり」類　さらずは等

④「しか」系　○「しか」類　しかありとも等　○「しかり」類　しからば等

⑤「そ」系　○「そ」類　そのゆゑに等　○「それ」類　それ等

（b）　その他の語を構成要素とするもの

①動詞系　あるいは他等　②名詞系　ゆゑは等　③副詞系　ただし等

B 転成接続詞　（あて）　および　かつ　すなわち　はた　また

C 借用接続詞　乃至（ないし）

以上の接続詞について京極・松井（1973）では、（イ）A（a）「か」系（「かり」類除く）・「さ」系は和文に用いられて訓読文に用いられない、（ロ）A（a）「こ」系・「しか」系および「か」系「かり」類は訓読文に用いられて和文に用いられない、（ハ）A（b）および B 転成接続詞・C 借用接続詞はほとんどが訓読文に用いられる、（ニ）A（a）「そ」系は語により異なるが訓読文に用いられる傾向が強いという文体的な傾向があることを指摘する。

また、訓読系接続詞としては築島（1963）「カルガユエニ・カレ・ココニ・ココヲモテ・コノユエニ・コレニヨリテ・コレヲモテ・シカノミナラズ・シカウシテ・シカウシテノチニ・シカルニ・シカニハアラズハ・シカモ・シカルヲ・シカラバ（シカレバ）・シカリトイヘドモ・シカレドモ・ソレ・タダシ・モシソレ・ユエニ・ユエヲモチテ」（328 頁）、また『源氏物語』にも見えるが用法や用例が限られているものとして「ソモソモ」をあげる。

中古和文の接続詞の全体的な調査としては、福島（2008）[1]等もある。本発表ではこれら

[1] 福島（2008）では索引を使用し、各文献の冒頭から 100 文までの間に現れた接続詞数（文のはじめに位置するもの）を調査している。以下例数をあげると、『竹取物語』かかれば１・かくて１、『土佐日記』

の指摘を参考に、「日本語歴史コーパス」から中古の接続語の使用傾向を探っていく。

３．調査範囲・調査方法

これまで述べてきたように、中古和文において接続詞と認められるものは少なく、「日本語歴史コーパス」で「品詞：接続詞」の検索をおこなうと、結果は「また、あるいは、すなわち、ただし、そゑに、さはれ、さて」の 7 語となる。

そこで本発表では、上記の【A】「日本語歴史コーパス」において接続詞とされるもの、【B】和文に見られる訓読系の接続語、そして和文においてもっとも多く用いられていると予測される【C】指示副詞カク系列「かく」・サ系列「さ」に動詞「あり」が複合した「かかり」「さり」による接続語（指示詞系接続語、「かかり」系・「さり」系と呼ぶ）について、どのような使用傾向があるのか分析していく。

【A】「日本語歴史コーパス」において接続詞とされるもの

「また、あるいは、すなわち、ただし、そゑに、さはれ、さて」全 7 語。

【B】和文に見られる訓読系の接続語

「そもそも」および「しかるに・しかれども・しかりとて」（指示副詞「しか」＋動詞「あり」で「しかり」、「しかり」系と呼ぶ。「しか」を用いた語は、築島（1963）でも指摘されているように訓読特有語である）を調査する。

【C】指示詞系接続語「かかり」系・「さり」系列

指示詞系接続語として「かかり」系「さり」系に焦点を当てる。「かかり」系・「さり」系についてはすべての語を取り入れるために、「かかり」「さり」の未然形から命令形まで各活用形を検索、それらすべてを発表者がチェックし、接続語を抽出していく。

○「かかり」系（語彙素読み「カカリ」（語彙素「斯り」）＋各活用形で検索）
未然形：全 29 例、「かからぬ N（N は名詞）」が多く、接続語は「かからで（も）」。
連用形：全 14 例、ほぼ「かかるける N」であり接続語無し。
終止形：全 8 例、接続語無し。
連体形：全 834 例、「かかる N」が非常に多く、接続語は「かかるほどに」「かかるに」等
已然形：全 31 例、接続語は「かかれば」「かかれど（も）」
命令形：全 2 例、接続語無し。
○「さり」 系（語彙素読み「サリ」（語彙素「然り」）＋各活用形で検索）
未然形：全 165 例、接続語は「さらば」が多く、他「さらずとも」「さらずば」「さらで」。
連用形：全 41 例、接続語は「さりければ」「さりぬべくは」「さりける時に」等。
終止形：全 194 例、接続語は「さりとも」が多く（ただし文頭で用いられていないものも多く接続語としていない）、他「さりとて（は・も）」。
連体形：全 853 例、「さる N」が最も多く、接続語は「さるに」「さるは」等。
已然形：全 235 例、接続語「されど」が多く、他「されば」等。
命令形：無

なお、上記の指示詞系接続語については、以下【１】【２】のように語をカウントしている。大まかに言って接続語は文頭にあって、直前の文と接続語を含む文をつなぐものである。そこで、今回は広範囲に傾向を見るため、それぞれの使用の詳細な内容までにはあまり踏み込まず、文頭で用いられているものを接続語として検索結果から抽出している。

【１】「かかり」系・「さり」系の未然形から命令形までの各検索結果の中より、前文が句点「。」で終わり、当該表現が文頭で用いられているもの、または文中にあっても（１）のように、会話文の文頭であるものを抽出する。

かくて２・また２、『蜻蛉日記』かくて６・さて３・また１、『源氏物語』また２、『伊勢物語』さて３、『大和物語』かくてまた１・さて４・さりければ１・また２であり、「和文で書かれた諸文献に見られる接続詞に関しては、まず量的に、文献間の多少のばらつきはあるものの、概して非常に少ない、つまり出現頻度が非常に低いというのがなによりの特徴といえる」（51 頁）とする。

（1）男ども申すやう、「さらば、いかがはせむ」（竹取物語、43 頁）
ただし、この場合あまり例は多くないが、（2）のように接続語と考えられるものが対象外となる。これについては、今後、修正おこなう。
（2）例の、明けてはてぬ。「よし、さらば、この物語は尽きすべうなんあらぬ、また、人聞かぬ心やすきところにて聞こえん。」（源氏物語、橘姫、163 頁）（夜も明けてしまった。「まあよい、それでは、この昔話はとても尽きそうにないことだし、また人に聞かれない安心な所でお話し申すことにしましょう」
【2】抽出したものの中より、以下のように明らかに接続語ではないものを省いていく。
（3）人目もなし。さらぬ人は、とぶらひ参るも重き咎めあり、わづらはしきことまされば（源氏物語、須磨、170 頁）
次に、本発表では「日本語歴史コーパス」にある本文種別を利用し、抽出した接続語が「歌・会話・手紙・地（地の文）・詞章」で傾向が見られるのかについても調査していく。その際、表１の文数を参照していく[2]。

表１　「日本語歴史コーパス」における文数（本文種別）

	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
歌	15	1063	61	235	297	72	147	39	794	18	2741
会話	204	0	30	20	146	1657	218	787	5640	52	8754
手紙	25	0	0	3	3	0	0	0	107	0	138
地	330	189	470	637	972	1371	295	2780	10072	739	17855
詞書	0	787	0	0	0	0	0	0	0	0	787
計	574	2039	561	895	1418	3100	660	3606	16613	809	30275

なお、本発表による表において、二段あるものは上段が用例数、下段は表１の文数よりの%であり、100 文あたりの出現割合となる（小数点第二位以下は切り捨て）。また、その接続語の%が全文数から計算した%より値が大きいものについては強調文字となっている（つまり、「日本語歴史コーパス」全体でその接続語が 100 文あたりどのくらい用いられるかを計の%で示し、各作品での使用された%が大きい場合には強調されている）。

４．中古における接続語

４.１　【A】「日本語歴史コーパス」において接続詞とされるもの

表２に「日本語歴史コーパス」で接続詞とされるものの検索結果を示す。

表２　「日本語歴史コーパス」おける接続詞（計の下段、100 文あたりの出現割合）

	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
また	1	9	13	1	29	32	17	126	353	19	600
あるいは	6	0	0	0	0	0	0	0	0	0	6
すなわち	1	2	0	0	3	3	1	6	1	0	17
ただし	2	0	0	0	0	0	0	0	0	0	2
そゑに	0	1	0	0	0	0	0	0	0	0	1
さはれ	0	0	0	0	0	6	1	5	2	0	14
さて	2	0	5	12	40	29	2	47	93	0	230
計	12 2.09	12 0.58	**18** **3.20**	13 1.45	**72** **5.07**	70 2.25	**21** **3.18**	**184** **5.10**	449 2.70	19 2.34	870 2.87

[2] それぞれの文数については小木曽智信氏にご協力頂いた。ここでの「文」は小学館新編日本古典全集の「。」を基準にカウントしたものであり、また文中に和歌や引用が入る場合にも、その前後で文を切っている。ただし、文が入れ子になる場合などはカウントされていない。また、表では『竹取物語・古今和歌集・土佐日記・伊勢物語・大和物語・落窪物語・和泉式部日記・枕草子・源氏物語・紫式部日記』の下線部で資料名を示す。

接続詞は、「また」600 例、「あるいは」６例、「すなわち」17 例、「ただし」２例、「そゑに」１例、「さはれ」14 例、「さて」230 例である。

そして、作品別では『枕草子（5.10%）・大和物語（5.07%）・土佐日記（3.20%）・和泉式部日記（3.18%）』が多く、『古今和歌集（0.58%）』は接続語の使用が少ない傾向が見られる。

なお「あるいは・ただし」は『竹取物語』のみ使用が認められる。これらの語については、訓読系の語と考えられ（築島（1963）等）、『竹取物語』の本文の属性をよく表していると考えられる（これについては、【B】和文に見られる訓読系の接続語でも述べる）。

（４）日暮るるほど、例の集りぬ。あるいは笛を吹き、あるいは歌をうたひ、あるいは声歌をし、あるいは嘯を吹き、扇を鳴らしなどするに（竹取物語、23 頁）

４.１.１　また

接続詞とされるものの中で、特に使用の多い「また」について分析していく。

表３　接続詞「また」

	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
また	1 0.17	9 0.44	**13** **2.31**	1 0.11	**29** **2.04**	32 1.03	**17** **2.57**	**126** **3.49**	**353** **2.12**	**19** **2.34**	600 1.98

まず、作品別にみると全体的に使用が多いのは『枕草子（3.49%）』であり、次に多いのは、『和泉式部日記（2.57%）・紫式部日記（2.34%）・土佐日記（2.31%）』となる。

次に、本文種別で表にすると表４となる。

表４　接続詞「また」

また	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
歌	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**2** **1.36**	0 0.00	**2**[3] **0.25**	0 0.00	4 0.14
会話	1 0.49	0 0.00	0 0.00	0 0.00	0 0.00	18 1.08	**4** **1.83**	9 1.14	**126** **2.23**	0 0.00	158 1.80
手紙	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**3** **2.80**	0 0.00	3 2.17
地	0 0.00	**9** **4.76**	**13** **2.76**	1 0.15	**29** **2.98**	14 1.02	**11** **3.72**	**117** **4.20**	222 2.20	**19** **2.57**	435 2.43
計	1 0.17	9 0.44	**13** **2.31**	1 0.11	**29** **2.04**	32 1.03	**17** **2.57**	**126** **3.49**	**353** **2.12**	**19** **2.34**	600 1.98

注：詞章例無し。なお、『古今和歌集』9 例中 6 例は「マタハ」であり、3 例の「マタ」はすべて仮名序で用いられている。そして「マタハ」は 6 例すべて次のような形式（左注）で用いられている。「忘らるる身を宇治橋のなか絶えて人もかよはぬ年ぞ経にける　又は、「こなたかなたに人もかよはず」」（巻第 15、恋歌 5、825 番歌）

表４に示すように「また」は傾向として、地の文に多く会話には少なく、さらに手紙・歌には少ない（『古今和歌集（地 4.76%、歌 0%）・枕草子（地 4.20%、会話 1.14%、歌 0%）・和泉式部日記（地 3.72%、会話 1.83%）・大和物語（地 2.98%、会話・歌とも 0%）・土佐日記（地 2.76%、会話・歌とも 0%））。

以上より「また」は、作品としては随筆・日記に多く歌物語には少なく、また本文種別としては（歌・手紙）会話よりも地の文に用いられやすい傾向があるといえる。

[3] 『源氏物語』の歌に見られる「また」について、「見てもまたあふよまれなる夢の中にやがてまぎるるわが身ともがな」（源氏物語、若紫、231 頁）（こうしてお会いすることができても、またお目にかかれる夜はめったにないのですから、いっそこの夢の中にこのまま紛れ消えてしまいたい）のように、２例とも「ふたたび～」といった副詞として機能しているものであると考えられる。他の分析でも同様であるが、今後、検索結果を詳細に見る必要がある。

４.１.２　さて

先の表２で示した「さて」には「さて」＋助詞「も・は」等が含まれるため、「さて」のみ（助詞下接なし）をカウントしたものを表５に示す。

まず、作品別について表５より『大和物語（2.67%）・伊勢物語（1.22%）』に用いられる傾向を指摘できる。これについては岡﨑（2011）で述べたように、歌物語は歌を中心に、各段が短いストーリーで構成されており、その歌がどのように歌われたかという状況を説明する。そこで「状況の説明→「さて」→歌」という形式が多く用いられており、そのためこのような結果が出たものと考えられる。なお、『紫式部日記』には 0 例、『和泉式部日記』には 1 例であるように（『土佐日記』を除き）日記には用いられにくいようである。

（５）伊勢の斎宮の御占にあひたまひにけり。「いふかひなくくちをし」と、思ひたまうけり。さてよみて奉りたまひける。伊勢の海の千尋の浜にひろふとも今はかひなくおもほゆるかな（大和物語、316 頁）

次に本文種別について、物語（『落窪物語・源氏物語』））においては会話文、歌物語・随筆（『伊勢物語・大和物語・枕草子』）では地の文に用いられる傾向がある。また、歌に「さて」が用いられることはない。

表５　接続詞「さて」

さて	竹取	土佐	伊勢	大和	落窪	和泉	枕	源氏	計
会話	0	0	0	1	15(10)	1(1)	8(11)	31(22)	56(44)
手紙	0	0	0	(1)	0	0	0	(1)	(2)
地	2	5	11(1)	38	4	0	28	21(18)	108(19)
計	2 0.34	**5** **0.89**	**11(1)** **1.22**	**39(1)** **2.75**	**19(10)** **0.61**	1(1) 0.15	**36(11)** **0.99**	52(41) 0.31	165(65) 0.54

注：歌、および『古今和歌集・紫式部日記』例無し。（ ）は「さても・さては」等。

この「さて」については、「日本語歴史コーパス」において「さて」（語彙素読み「サテ」）で検索すると接続詞（語彙素「扨」）165 例、副詞（語彙素「然て」）72 例、感動詞（語彙素「さて」）3 例と３つとなる（助詞下接なしのもののみ）。

この結果は岡﨑（2011）で、中古においては未だ副詞として機能する例がまとまってみられるが接続語の方が多いという結果と同様である（接続語 149 例、副詞的 63 例）。[4]そこで、「日本語歴史コーパス」と岡﨑（2011）による「さて」の調査結果を表に示しておく。

表６　「日本語歴史コーパス」と岡﨑（2011）による接続語「さて」

	竹取	土佐	伊勢	大和	枕	源氏
歴史コーパス・接続詞	2	5	11	38	36	52
岡崎(2011)・接続語	2	4	11	38	36	58

４.２　【B】和文に見られる訓読系の接続語

中古和文における訓読系の接続語「そもそも」、「しかり」系の「しかるに・しかれども・しかりとて」について見ていく（築島（1963）等。なお、「日本語歴史コーパス」では接続詞ではない）。「そもそも」は作品別では、『古今和歌集・土佐日記・竹取物語・大和物語・源氏物語』で用いられており、また表７に示すように『古今和歌集』（仮名序）以外は、和

4　岡﨑（2011）調査資料には他に『蜻蛉日記・平中物語』がある。また、接続語はタイプ B・C・D にあたり、表はそれらの合算である。なお「日本語歴史コーパス」にある『古今和歌集・落窪日記・紫式部日記・和泉式部日記』は岡﨑（2011）では調査対象外である。さらに「さて」の対照となる「かくて」については、接続語より副詞として機能するものの方が多いという結果が出ており（接続語 49 例、副詞的 134 例）、「日本語歴史コーパス」で検索したところ「かくて」（語彙素読み「カク」（語彙素「斯く」）＋後方共起：語彙素読み「テ」＋品詞「助詞」）329 例すべて副詞である。

文においてすべて会話において使用されており、また発話者は『土佐日記』を除きすべて男性である[5]。

このように、例は少ないが「そもそも」は、和文において男性の会話に用いられやすいことが指摘される。

（6）翁のいはく、「思ひのごとくものたまふかな。そもそも、いかやうなる心ざしあらむ人にかあはむと思す」（竹取物語）（私の考えと同じことをおっしゃるね。それはそうと、どんな愛情をお持ちの方と結婚しようとお思いか）

表7　「そもそも」

そもそも	竹取	古今	土佐	大和	源氏	計
会話	1	0	1	2	2	6
地	0	1	0	0	0	1
計	**1** **0.17**	**1** **0.04**	**1** **0.17**	**2** **0.14**	2 0.01	7 0.02

注：『伊勢物語・落窪物語・枕草子・紫式部日記・和泉式部日記』および歌・手紙・詞章は例無し。

次に訓読系の接続語として『竹取物語』「しかるに」1例（工匠）・「しかれども」1例（王けいの手紙）、『土佐日記』「しかれども」1例（地の文）、『古今和歌集』「しかりとて」1例（小野篁の歌）が見られ、これについても上記の傾向と同様と考えられる。

（7）「内匠寮の工匠、あやべの内麻呂申さく、玉の木を作り仕うまつりしこと、五穀を断ちて、千余日に力をつくしたること、すくながらず。しかるに、禄いまだ賜はらず。」（竹取物語、34頁）

表8　「しかり」系

	竹取	古今	土佐	計
しかるに	1	0	0	1
しかれども	1	0	1	2
しかりとて	0	1	0	1
計	**2** **0.34**	**1** **0.04**	**1** **0.17**	4 0.01

注：『伊勢物語・大和物語・落窪物語・枕草子・源氏物語・紫式部日記・和泉式部日記』例無し。

4.3　【C】指示詞系接続語「かかり」系・「さり」系

4.3.1　「かかり」系

「かかり」系の調査結果を表9に示す。「かかり」系には連用形、終止形の接続語の例が無い。傾向として「かかり」系の語は、『土佐日記（1.42%)・竹取物語（0.87%)』に多く、それに対し『紫式部日記（0.00%）・枕草子（0.02%）・源氏物語（0.04%)』には少ないことが指摘できる（たとえば、『土佐日記』では以下のように「かかれば」（「かかれど（も)」）が使用されるが、『源氏物語』では「されば」（「されど（も)」）が用いられる)。

（8）この泊の浜には、くさぐさのうるわしき貝、石など多かり。かかれば、ただ、昔の人をのみ恋ひつつ、船なる人のよめる（土佐日記、44頁）（この港の浜には、いろいろきれいな貝や、石などが多かった。だから、ただもう、亡き子ばかりを恋しく思い思いして）

（9）「皇子たちあまたあれど、そこをのみなむかかるほどより明け暮れ見し。されば思ひわたさるるにやあらむ、いとよくこそおぼえたれ」（源氏物語、紅葉賀、329頁）（皇子たちは大勢いるけれど、ただそなただけを、こういう幼い自分から朝晩見ていたものだ。そ

[5] 『土佐日記』の例は複数人の発話であり、性別は特定できない。また、『大和物語』の注に、「そもそも」は「漢文訓読語。男子、僧侶などの漢文に親近することの多い階級の人のことばとしてつかわれる」（415頁）にもあるように、これらは「翁」『竹取物語』、「帝・うへのきぬ来る者（袍)」『大和物語』、「僧都・夕霧」『源氏物語』と『竹取物語』以外はそのように使用されている。

のために、自然そのころが思いだされるせいか、じつにそなたによく似ている）

なお、表９・12 から「未然形＋ば」である「さり」系の「さらば」は多く見いだせるが、「かかり」系の「かからば」は例が見いだせない。さらに「已然＋ば」である「かかれど」「されど」を比較すると、「されど」は多く用いられるが「かかれど」の例は少ない。これについては、指示副詞「かく」と「さ」の用法の違いが影響していることが予想される。これについては、別の機会に論じていく。

表９　「かかり」系

未然形	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
かからで	0	0	0	0	0	0	0	0	1	0	1
かからでも	0	0	0	0	0	0	0	0	1	0	1
連体形											
かかるあひだに	0	0	3	0	0	0	0	0	0	0	3
かかるうちに	0	0	1	0	0	0	0	0	0	0	1
かかるに	0	1	0	0	0	0	0	0	0	0	1
かかるほどに	3	0	0	1	1	6	2	1	4	0	18
かかるままに	0	0	0	0	0	3	0	0	0	0	3
かかるも	0	0	0	0	0	0	1	0	0	0	1
已然形											
かかれば	2	0	1	1	3	5	0	0	0	0	12
かかれど	0	0	1	0	2	0	0	0	1	0	4
かかれども	0	0	2	0	0	0	0	0	0	0	2
計	**5** **0.87**	1 0.04	**8** **1.42**	**2** **0.22**	**6** **0.42**	**14** **0.45**	**3** **0.45**	1 0.02	7 0.04	0 0.00	47 0.15

４.３.２　「かかり」系：「かかるほどに・かかれば」

「かかり」系で例数の多かった語「かかるほどに・かかれば」を、作品・本文種別から見てみる。

これらの語は作品としては『竹取物語・伊勢物語・落窪物語・大和物語』に見られ、『古今和歌集・紫式部日記』（『源氏物語・枕草子』も少ない）には見られない。

次に本文種別から見ると、「かかるほどに」については地の文のみで用いられているのに対し、「かかれば」は会話でも用いられていること、また歌・手紙・詞章で用いられていないことが指摘できる。

表 10　「かかるほどに」

かかるほどに	竹取	伊勢	大和	落窪	和泉	枕	源氏	計
地	**3** **0.90**	**1** **0.15**	**1** **0.10**	**6** **0.43**	**2** **0.67**	1 0.03	4 0.03	18 0.10

※『古今和歌集・土佐日記・紫式部日記』例無し。地の文のみ。

表 11　「かかれば」

かかれば	竹取	土佐	伊勢	大和	落窪	計
会話	**1** **0.49**	**1** **3.33**	**1** **5.00**	**1** **0.68**	**1** **0.06**	5 0.05
地	**1** **0.30**	0 0.00	0 0.00	**2** **0.20**	**4** **0.29**	7 0.03
計	**2** **0.34**	**1** **0.17**	**1** **0.11**	**3** **0.21**	**5** **0.16**	12 0.03

※『古今和歌集・枕草子・源氏物語・紫式部日記・和泉式部日記』例無し。歌・手紙・詞書例無し。

4.3.3 「さり」系

「さり」系の調査結果を表 12 に示す。作品としては『紫式部日記（1.97%）・伊勢物語（1.89%）・枕草子（1.63%）・源氏物語（1.42%)』に多く、『土佐日記（0.53%）・古今和歌集（0.14%)』には少ない傾向が見える。この結果は先の「かかり」系と対照的である（なお『古今和歌集』は【A】0.58%、【C】「かかり」系 0.04%とそもそも接続語が少ない)。

表 12 「さり」系

未然形	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
さらば	3	0	0	1	2	13	1	11	51	1	83
さらずとも	0	0	0	0	0	1	0	0	1	0	2
さらずば	0	0	0	1	0	2	0	2	3	0	8
さらで	0	0	0	0	0	0	0	1	1	0	2
連用形											
さりけるころほひ	0	0	0	0	1	0	0	0	0	0	1
さりける時に	0	0	0	0	2	0	0	0	0	0	2
さりけるものを	0	0	0	0	0	0	0	1	0	0	1
さりけれど	0	0	0	1	0	0	0	0	0	0	1
さりけれども	0	1	0	0	0	0	0	0	0	0	1
さりければ	0	0	0	2	5	0	0	0	0	0	7
さりながらも	0	0	0	0	0	0	0	0	1	0	1
さりぬべくは	0	0	0	0	0	2	1	0	1	0	4
終止形											
さりとて	0	1	0	0	0	1	2	1	9	2	16
さりとては	0	0	0	0	0	0	0	0	1	0	1
さりとても	0	0	0	0	0	0	0	0	1	0	1
さりとも	2	0	0	1	0	4	1	4	33	0	45
連体											
さるに	0	0	0	2	1	0	0	0	0	0	3
さるにては	0	0	0	0	0	0	0	0	1	0	1
さるにても	0	0	0	0	0	0	0	0	1	0	1
さるによりて	0	0	0	0	0	0	0	0	1	0	1
さるほどに	0	0	0	0	0	1	0	0	1	0	2
さるものから	0	0	0	0	0	0	0	0	1	0	1
さるを	0	0	0	1	0	0	0	0	0	0	1
さるは	0	0	1	0	0	1	1	3	51	5	62
已然形											
されど	2	0	0	5	5	3	1	34	71	8	129
されども	0	0	1	0	0	0	0	0	2	0	3
されば	1	1	1	2	6	6	0	1	6	0	24
さればとて	0	0	0	0	0	0	0	1	0	0	1
さればなむ	0	0	0	1	0	0	0	0	0	0	1
計	**8** **1.39**	3 0.14	3 0.53	**17** **1.89**	**22** **1.55**	34 1.03	7 1.06	**59** **1.63**	**237** **1.42**	**16** **1.97**	406 1.34

4.3.4 「さり」系:「さらば・さりとも・さるは・されど」

「さり」系において例数の多かったものを作品・本文種別から見てみる。

まず、作品別において「さらば」が比較的用いられるのが『竹取物語（0.52%）・落窪物語（0.41%）・枕草子（0.30%）・源氏物語（0.30%)』であり、「さりとも」は『竹取物語（0.34%）・源氏物語（0.19%）・和泉式部日記（0.15%)』、そして「さるは」は『源氏物語（0.30%）・紫

式部日記（0.61%）』、最後に「されど」は『伊勢物語（0.55%）・枕草子（0.94%）・源氏物語（0.42%）・紫式部日記（0.98%）』である。そして、これらの語すべて『土佐日記・大和物語』では例が僅かである（「さるは」「されど」については紫式部による『源氏物語・紫式部日記』に多い傾向が見られる）。

表 13　「さらば」（「さり」系・未然形）

さらば	竹取	古今	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
歌	0	0(1)	0	0	0	(1)	0	0	0	0(2)
会話	**3** **1.47**	0 0.00	**1** **5.00**	**2** **1.36**	13(1) 0.78	1 0.45	**11(1)** **1.39**	48(15) 0.85	**1** **1.92**	80(17) 0.91
地	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**3(11)** **0.02**	0 0.00	3(11) 0.01
計	**3** **0.52**	0(1) 0.00	1 0.11	2 0.14	**13(1)** **0.41**	1(1) 0.15	**11(1)** **0.30**	**51(26)** **0.30**	1 0.12	83(30) 0.27

※『土佐日記』および詞章・手紙は例無し。（ ）は、接続語としなかった例数、以下表 14-16 は同じ。

表 14　「さりとも」（「さり」系・終止形）

さりとも	竹取	伊勢	落窪	和泉	枕	源氏	紫式	計
歌	0 0.00	**1** **0.42**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.03
会話	**2** **0.98**	0 0.00	4(1) 0.24	1(3) 0.45	**4(3)** **0.50**	**32(35)** **0.56**	0 0.00	43(42) 0.49
地	0 0.00	0 0.00	0(1) 0.00	0 0.00	0(1) 0.00	1(36) 0.00	0(1) 0.00	1(39) 0.00
計	**2** **0.34**	1 0.11	4(2) 0.12	**1(3)** **0.15**	4(4) 0.11	**33(71)** **0.19**	0(1) 0.00	45(81) 0.14

※『古今和歌集・土佐日記・大和物語』および詞章・手紙は例無し。

表 15　「さるは」（「さり」系・連体形）

さるは	土佐	落窪	和泉	枕	源氏	紫式	計
会話	0 0.00	0 0.00	**1(1)** **0.45**	0 0.00	**17** **0.30**	0 0.00	18(1) 0.20
地	1 0.21	1(1) 0.07	0 0.00	3 0.10	**34(8)** **0.33**	**5** **0.67**	44(9) 0.24
計	1 0.17	1(1) 0.03	1(1) 0.15	3 0.08	**51(8)** **0.30**	**5** **0.61**	62(10) 0.20

※『古今和歌集・伊勢物語・竹取物語・大和物語』および歌・詞章・手紙は例無し。

表 16　「されど」（「さり」系・已然形）

されど	竹取	古今	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
会話	**2** **0.98**	0 0.00	0 0.00	0 0.00	3(1) 0.18	1 0.45	**11** **1.39**	25(2) 0.44	**8** **1.53**	50(3) 0.57
地	0 0.00	0(1) 0.00	**5** **0.78**	**5** **0.51**	0 0.00	0 0.00	**23(1)** **0.82**	**46(5)** **0.45**	0 0.00	79(7) 0.44
計	2 0.34	0(1) 0.00	**5** **0.55**	5 0.35	3(1) 0.09	1 0.15	**34(1)** **0.94**	**71(7)** **0.42**	**8** **0.98**	129(10) 0.42

※『土佐日記』および歌・詞章・手紙は例無し。

次に、本文種別から見ると「さらば」（表 13）は、ほぼ会話で用いられている。「さりとも」（表 14）については、同じく会話で用いられる傾向がある。なお、表 14 に示すように「さりとも」は、本発表では接続語としなかった例の方が『源氏物語』において 71 例と多い（接続語 33 例）。これについては、「さりとも」の用法と（10）に示すように新編古典文学全集における句読点の方針の影響も考えられる（心内語は「」でくくらない等）。

（10）げに言ふかひなのけはひや、さりとも、いとよう教へてむと思す。（源氏物語、若紫、238）（(紫の上が）なるほどたわいのない有様よ。それでも、あの人を是非立派に育てあげてみたいもの）

最後に「さるは」「されど」（表 15・16）については、「さらば」「さりとも」とは違い地の文に多い傾向が見られる。

５．まとめ

これまでの結果を表にまとめる（全文数の%より多かったものを強調している）。表 17 より、先にも述べたが『古今和歌集』は接続語の使用が少なく、それに対し『竹取物語・大和物語・土佐日記』は多い傾向を示している。また、作者が同一である『源氏物語・紫式部日記』は使用傾向がよく似ており、同時代だからであろうか『枕草子』も同様である。

最後に、日記である『土佐日記』と『紫式部日記・和泉式部日記』（「また」の使用が多いということは共通する）は、やはり作者の位相差（『土佐日記』男性、『紫式部日記・和泉式部日記』女性）であろうか、使用傾向は似ていない。また『土佐日記』と『大和物語』（「さり」系除く）、『伊勢物語』と『落窪物語』の使用傾向が似ていることが指摘される[6]。

表 17　まとめ（100 文あたりの出現割合）

	竹取	古今	土佐	伊勢	大和	落窪	和泉	枕	源氏	紫式	計
【A】全体	2.09	0.58	**3.20**	1.45	**5.07**	2.25	**3.18**	**5.10**	2.70	2.34	2.87
【A】また	0.17	0.44	**2.31**	0.11	**2.04**	1.03	**2.57**	**3.49**	**2.12**	**2.34**	1.98
【B】そもそも	**0.17**	**0.04**	**0.17**	0.00	**0.14**	0.00	0.00	0.00	0.01	0.00	0.02
【B】「しかり」系	**0.34**	**0.04**	**0.17**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
【C】「かかり」全	**0.87**	0.04	**1.42**	**0.22**	**0.42**	**0.45**	**0.45**	0.02	0.04	0.00	0.15
【C】かかるほどに	**0.90**	0.00	0.00	**0.15**	**0.10**	**0.43**	**0.67**	0.03	0.03	0.00	0.10
【C】かかれば	**0.34**	0.00	**0.17**	**0.11**	**0.21**	**0.16**	0.00	0.00	0.00	0.00	0.03
【C】「さり」全	**1.39**	0.14	0.53	**1.89**	**1.55**	1.09	1.06	**1.65**	**1.42**	**1.97**	1.34
【C】さらば	**0.52**	0.00	0.00	0.11	0.14	**0.41**	0.15	**0.30**	**0.30**	0.12	0.27
【C】さりとも	**0.34**	0.00	0.00	0.11	0.00	0.12	**0.15**	0.11	**0.19**	0.00	0.14
【C】さるは	0.00	0.00	0.17	0.00	0.00	0.03	0.15	0.08	**0.30**	**0.61**	0.20
【C】されど	0.34	0.00	0.00	**0.55**	0.35	0.09	0.15	**0.94**	**0.42**	**0.98**	0.42

文　献

京極興一・松井栄一（1973）「接続詞の変遷」『品詞別日本語文法講座 6 接続詞・感動詞』鈴木一彦・林巨樹編、pp.90-135、明治書院

塚原鉄雄（1958）「接続詞」『続日本文法講座１　文法各論編』pp.156-174、明治書院.

築島裕（1963）『平安時代の漢文訓讀語につきての研究』東京大学出版会

福島直恭（2008）『書記言語としての「日本語」の誕生―その存在を問い直す―』笠間書院

岡﨑友子（2010）『日本語指示詞の歴史的研究』ひつじ書房

岡﨑友子（2011）「指示詞系接続語の歴史的変化―中古の「カクテ・サテ」を中心に―」『日本語文法の歴史と変化』青木博史編、pp.67-87、くろしお出版

[6] 全体的に女性が書いたとされるものは接続語の使用が少なく、それに対し男性が書いたものとされるものは多い傾向を示している。

「ガ/ノ」交替現象についての一考察
—古代・現代コーパスを対照して—

坂野 収（青山学院大学）

On Ga/No Conversion : A Diachronic Corpus-based Study

Osamu Banno (Aoyama Gakuin University)

１．はじめに

三上（1953）以来、「ガ／ノ」交替について、さまざま議論されてきた[1]。しかし、現代語中心で、内省に基づく議論が主であった。幸いにして、充実されつつある通時コーパスが利用できるようになったので、可能な限りコーパスを利用して検討をしてみた。報告の要旨は次の通りである。

① 中古語の主語の格表示（主格）はゼロ格（φ）である。しかし、連体形節等では「ガ」または「ノ」が用いられることがある。これら主語表示「ガ・ノ」は属格であることが、コーパス分析から確認できる。

② 中古語の連体形節等における「ゼロ格(φ)／ガ・ノ」交替と、現代語における「ガ／ノ」交替は、言語学的にいえば「主格／属格」交替であり、同一現象である。

③ 通時コーパスによって、中古と現代とで「主格／属格」交替現象を比較検討すると、属格主語の出現は、質的にも比率的にも変化はみられない。従って、主語への属格付与メカニズムも変わっていないと考えることができる。

④ 「主格／属格」交替は、名詞性（あるいは、名詞素性）を帯びた節でのみ生ずる現象である。この名詞素性を持つ節を、主要部 C が名詞素性[+N]をもった節タイプの一つと位置付けて、生成統語論の立場で、「主格／属格」交替現象を説明する。

２．中古語の主語表示としての「ガ・ノ」

連体形節等に於いて、主語の格表示として格助詞「ガ」と「ノ」が用いられることがあるのはよく知られている。これらの格助詞の種類について考察する。

２．１　属格「ガ・ノ」の上接語分布

中古語では、属格助詞「ガ」の上接語は限定的であり「ノ」のそれは一般的と言われている[2]が、その分布を中納言(「歴史コーパス」)で検索[3]・調査した。

結果の概要を、表１とそれをグラフ化した図１に示す。共通の上接語も散見されるが、どちらかが主力の場合が多く、真に共通なものは「君」（表１＊ 印の所）くらいである。結論をいえば、互いの上接語の交わりは非常に小さく、ほぼ相補分布しているといっても差支えない[4]

---

[1] Harada(1971), Miyagawa(1993), Watanabe(1994), Hiraiwa(2001b), 大島（2010）など。
[2] 野村（1993)、((2011) pp.75）など。
[3] 検索条件：短単位検索、[名詞 or 代名詞] ＋[格助詞が or の](key)＋ [名詞 or 代名詞]
[4] 分布の内容についての考察は、本稿の目的とは関係しないので立ち入らない。

表１　属格「ガ・ノ」の上接語分布

上接語	の	が	上接語	の	が
こ	2140		大殿	69	
人	899		院	68	
世	864		梅	67	(6)
そ	857	(4)	右	63	
か	716		大臣	62	
心	391		君	*60	*(66)
もの	203		衣	51	
例	199		琴	50	
花	196		上	49	(2)
物	177		今	48	
秋	167		いにしへ	41	
よろづ	151		歌合	41	
山	146		下	39	(2)
身	133		子	39	(2)
宮	132		松	39	(2)
女	132		弁	36	(1)
中	132		まこと	33	
春	129				
昔	126		わ		(629)
夜	122		おの		(63)
前	119		誰		(37)
方	117		これ		(30)
ほど	110		まろ		(14)
こと	105		それ	2	(13)
後	102		帯刀	4	(10)
殿	102		我		(7)
おほかた	100		あこぎ		(5)
東	98		小萩		(5)
事	97	(1)	人麿		(5)
少将	90		惟光	1	(4)
日	86	(1)	貫之		(4)
御簾	85		仲忠		(4)
中将	82		浅茅	1	(3)
納言	77	(3)	た		(3)
空	74		平中		(3)
宰相	72	(1)	典薬	2	(2)
				2960種	142種
				95%	5%
				23344個	1091個
				96%	4%

表２　主語表示「ガ・ノ」の上接語分布

上接語	の	が	上接語	の	が
人	592	(3)	葉	15	
こと	138		北の方	14	
心	115		露	14	
宮	59		御髪	13	
世	59		事	13	
月	52		水	13	
ほど	42		使	12	
もの	42	(1)	色	12	
方	42		心ざし	12	
身	39		中将	12	
人々	38		かぐや姫	11	
花	36		翁	11	
雪	34		子	11	
君	*33	*(12)	車	11	
心地	31		神	11	
さま	29		昔	11	
涙	28		大将	11	
気色	27				
大臣	27		わ		(94)
中	27		誰		(22)
殿	26		おの		(13)
女	24		まろ		(7)
院	23		それ		(6)
何	22		あこぎ		(5)
風	22		なにがし		(5)
声	21		かれ		(4)
夜	21		た		(4)
上	20		右近		(3)
親	19		衛門		(3)
雨	18		汝		(3)
音	18		これ		(2)
男	18		監		(2)
日	18		人麻呂		(2)
ありさま	17		帯刀		(2)
者	16		典薬		(2)
物	16		背子		(2)
				893種	71種
				93%	7%
				3763個	251個
				94%	6%

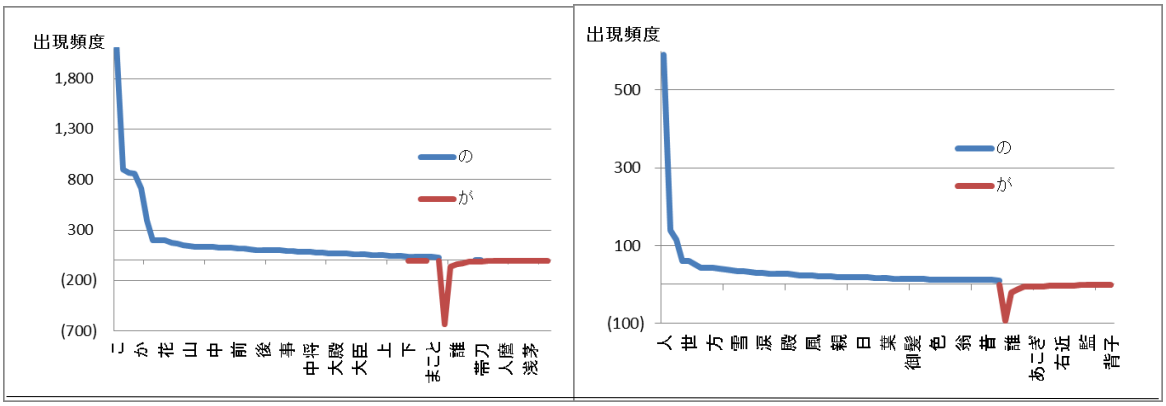


図１　属格「ガ・ノ」の上接語分布

図２　主語表示「ガ・ノ」の上接語分布

２．２　主語表示「ガ・ノ」の分布

連体形終止節、已然形終止節（以降、連体形節、已然形節と称する）、ならびに「未然形＋ば」節では、主語表示に「ガ・ノ」が現れることがある。この主語表示の「格」の種類は何であろうか。本稿では、終止形終止節（平叙文）の主語表示[5]の格を主格とする。

主語表示「ガ・ノ」の上接語の分布を、「中納言」を使って検索[6]・調査した。結果を表２と図２に示す。上接語の分布が同じである上に、特殊環境（連体形節など）にしか現れないことから、これらの格助詞は属格とみても差支えない。Frellesvig ((2010) pp.127-131) では、これらを「属格主語表示」(genitive subject marking)としていることは妥当である。以降、属格表示された主語を**属格主語**と称する。

３．中古語と現代語の「主格／属格」交替

３．１　中古語の格交替

中古語では、連体形節、已然形節、そして「未然形＋ば」節に於いては、（１）～（３）に示すように、いわゆる「主格(ゼロ格φ)／属格(ガ・ノ)」交替が存在する。

（１）連体形節の場合

a.　[いと胸 **φ** いたかるべき]ことなり　(源氏物語)
b.　[こがるる胸 **の** 苦しき] に思ひあまれる炎とぞ見し　(源氏物語)
c.　花すすき[君 **φ**なき] 庭に群れたちて　(古今和歌集)
d.　[鏡にて影見し君 **が** なき]ぞ悲しき　(大和物語)
e.　[梅の香 **φ**をかしき]を見出してものしたまふ。　(源氏物語)
f.　[菊の花 **の**うつろへる]を折りて、男のもとへやる　(伊勢物語)
g.　[日 **φ** たくる]ままに、いかならんと思したるを　(源氏物語)
h.　[日 **の** かさなる]ままにいみじくなむ、　(落窪物語)

（２）已然形節の場合

a.　迎へに[人**φ**あれば]、今またもまゐり来む　(大和物語)
b.　[うちとくまじき人 **の** あれば]、こなたの火は消ちたるに、(枕草子)

（３）「未然形＋ば」の場合

a.　[琴なども習はす人 **φ** あらば]、いとよくしつべけれど、　(落窪物語)
b.　[また見知る人 **の** 侍らば]こそあらめ、　(落窪物語)

３．２　現代語の格交替

現代語では、連体修飾節、補足節[7]などに於いて、（４）に示すように「主格／属格」交替（通称「ガ／ノ」交替）が現れるが、已然形節（の一部）と「未然形＋ば」の後継である仮定形節では、属格主語表示は生じない。

---

[5] 中古の平叙文の主語表示は「ゼロ格(無形)」である(金水・他(2011) pp.94-5,　Frellesvig (2010) pp.129)。
[6] 検索条件：短単位検索、[名詞 or 代名詞]＋[格助詞が or の](key)＋[動詞 or 形容詞]
[7] 益岡・田窪(1992, pp.182)

（４）現代語の格交替

a.　［ 山田 **が/の** 買った］本　　連体修飾節
b.　［ 秋刀魚 **が/の** 焼ける］匂い　　連体修飾節
c.　［ 山田 **が/の** 来たの］を思い出した。　　補足節
d.　［ 山田 **が/の** 買ったの］を食べた。　　補足節
e.　［リンゴ **が/の** 皿の上にあったの］を食べた。　　補足節（主要部内在型関係節）
f.　［ 息子**が/の** いう］ままに、任せていた。　　副詞節

（１）と（４）を比較すれば分かるように、連体修飾節、補足節、そして副詞節など、連体形節に限れば、**属格主語が許される環境は、、中古語も現代語も同じである。**

３．３　中古語と現代語における属格主語の出現割合

中古と現代とで、主格主語と属格主語の生ずる比率をコーパスで調査した。結果を図３に示す。縦軸の節タイプに対応した、主格主語と属格主語の出現割合を横棒で表してある。

重要なことは、連体形節に於いては、**主格／属格交替現象は、中古語と現代語とでは量的(出現比率)にもほとんど変化していないことである。**

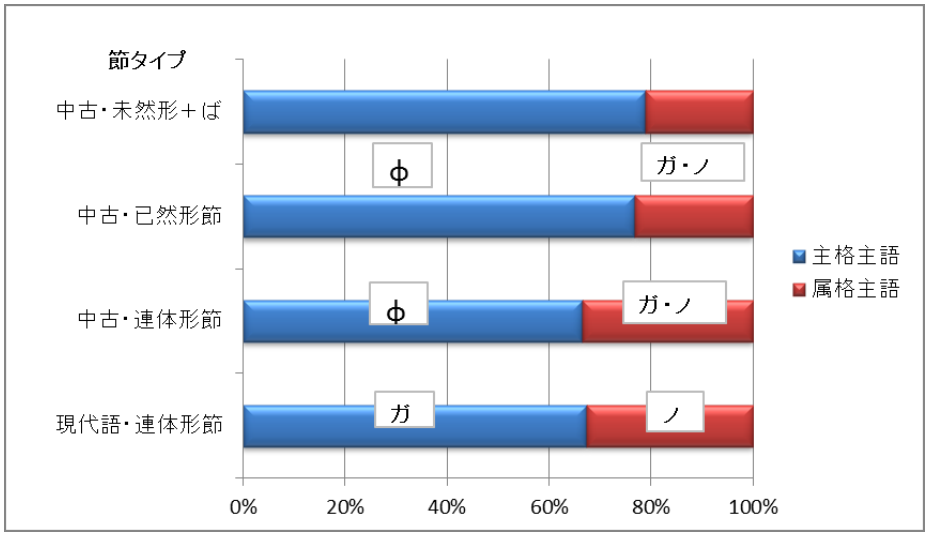


図３　中古・現代語の属格主語と主格主語の割合

検索条件を注[8]に示した。検索結果には、目的に合わない文も多く含まれており、それらの削除(手作業)が完璧でないことや、一定範囲、一定条件(随意に交替が可能と思われる文脈)での検索であることから、図３の数値は、**概略傾向を示す数値である**と認識されたい。

[8] 対象コーパス：中古語＝平安時代データ、現代語＝BCCWJ；コアーデータ
各検索条件は次の通り。ただし連体形節での検索条件のみ記した（いずれも短単位検索）。已然形・未然形はこれに準ずる。
①中古・属格主語表示＝[名詞 or 代名詞](key)＋[ガ or ノ](key)＋[動詞 or 形容詞]＋[(key から３語以内)連体形]　（付録参照）
②中古・主格主語表示＝[名詞 or 代名詞](key)＋[動詞 or 形容詞]＋[(key から３語以内)連体形]
③現代語・主語表示＝[名詞 or 代名詞]＋[ガ or ノ](key)＋[動詞 or 形容詞]＋[(key から３語以内)連体形]

４．属格主語が生ずる環境

現代語において、平叙文、疑問文、そして命令文などには、属格主語は現れない。属格主語が現れるは典型的な例は、主に（４）で述べたような、連体修飾節や補足節である。さらに、（５）（６）に示すように、副詞可能名詞や副助詞などの補部である連体形節でも現れる。これも、中古と現代で同じである。。

（５）副詞可能名詞が使われている例

a.［御都合のよろしい］**おり**に、お訪ねするつもりです。　大西巨人(2003)
b.［時間の許す］**かぎり**、この男の顔を見ていたかった。　藤原万璃子(2002)
c.［隼人の見る］**ところ**、木谷伝八の腕はなかなかのものである。 えとう乱星(2005)

a'.［中将のなき］**をり**に見すれば、心憂しと思へど　(源氏物語)
b'.［水の音の聞こゆる］**かぎり** は心のみ騒ぎたまひて、　(源氏物語)
c'.［さばかりねたげなりし人の見る］**ところ**もありなどこそは思ひはべりつれど、(源氏物語)

（６）副助詞等が使われている例

a.［彼女の帰宅する］**まで** に、決着をつけておきたかった。　折原一（1992）
b.［共産中国の認めた]**より**もさらにいっそう重大化しているのではないかと（ロバート・A・スカラピノ 2003)
c.［私の思う］**に**、これは芸術的衝動というよりもむしろ～　澁澤龍彦(2003)

a'.［月の傾く］**まで** あばらなる板敷に臥せりてよめる　(古今和歌集)
b'.［御使いの申す]**より**も、いますこしあわただしげに申しなせば、(源氏物語)
c'. かならずしも［我が思ふ］**に**かなわねど、　(源氏物語)

従来から、属格主語が現れる説明の一つとして、(非明示も含めて）名詞性の主要部(被修飾部)の存在を仮定してきた。しかし、（６）の例文、ならびに中古の補足節や已然形節を考えたとき、名詞的主要部の存在が必須だとは言えない。**属格主語が生ずるのは、名詞的特性（名詞素性）をもつ節そのものの存在が重要である**。

已然形節、「未然形＋ば」については、属格主語が現れることから、連体形節に近い名詞性を帯びた節と考えられてきた[9]が、図３でもそれが裏付けられると同時に、これらが連体形と同源あるいは、連体形＋αであったとの説明[10]もなっとくできる。現代語では、これらが存在しないこともあり、対照分析のターゲットとしては、連体形節に重点をおいて説明する。なお、「連体形節＋が」という主語表現が存在するが、この「ガ」は補文標識であるという説[11]もあり、今後の検討課題にしたい。

５．属格主語が生ずるメカニズム

５．１　先行研究概観

[9] ホイットマン（2009）
[10] 早田（2010)、金水・他（2011）pp.86-87.
[11] Frellesvig（2010）pp.128-129.

属格主語に関する先行研究は、ほぼ三分類できる。Miyagawa(1993)を中心とした主要部DP仮説、Watanabe(1994)のWH-agree仮説、そしてHiraiwa(2001ｂ)の連体形認可仮説である。詳細は省き、概念のみの説明とする。

主要部DP仮説は、関係節をその典型として、連体形節の外側に付加された主要部DPの存在を仮定し、LFにおいて、主語が主要部DPの指定部に移動して属格が認可されるという主張である。前節で述べたように、中古と現代の実例をみるに、必ずしも外部主要部があるわけではない。

WH-agree仮説は、WH移動が認められる節に於いて、Tns-Agr-Cシステムにより属格付与が可能になるという説である。主要部DPからの属格認可は必要ないが、WH移動があるからには主要部の存在を前提にしていることになろう。さらに、一般の疑問文といったWH移動が関係している文で属格主語が生じないのも説明できない。

最後は、連体形認可仮説であるが、Hiraiwaは、述部連体形のみが属格主語を認可すると述べている。しかし、独立用法の活用形と接続用法の活用形との区別[12]がつけられていないために、例えば「山田が／*の来るはずだ」のような、いわゆる人魚構文（角田 2012）では属格主語が許されないことを説明できない。

このように、いずれの理論も十全とは言い難い。そこで、以下述べるような、節タイプとしての連体形節を提案する。

５．２　連体形節によるメカニズムの説明

前節において、属格主語が生ずるのは、名詞素性の節の存在が重要であると述べた。そこで、この名詞素性をもつ節も、節タイプ(Clausal Type or Clausal Mood)[13]の一つであると規定し、それを改めて「**連体形節**」と称することにする。そして、複雑な理論を立てることなく、この**連体形節の存在のみが、属格主語が生ずる必要十分条件であると主張する**。そして、属格主語の派生を生成文法(特に、ミニマリスト・プログラム)の立場で説明する。

文（節）は、項構造（Argument Structure）を派生するvP（動詞句）相とその上部に位置し表現形式（Expression Structure）を決めるCP相から構成される。CP相の主要部Cが節タイプを決める素性(Illocutionary force＝発話内力）を持っているとされている(Radford 1997 pp.148)。日本語においては、節タイプと述部活用形（独立形式の活用形）は密接な関係にあり、主要部Cは述部の活用形も決定する力を持つと仮定する[14]。

終止形節の一つである平叙文と比較して、連体形節の派生の概要を説明する。図４は平叙文の基本的節構造、図５は連体形節のそれである。

図４に於いて、平叙文CPの主要部C（補文標識）は無標であって、主語に主格を、述部に終止形を認可する素性[Nom, Concl]を持ち、時制句(TP)の主要部 T を経由してそれぞれを認可する。一方図５では、名詞素性[+N]を持つ主要部Cが選択されている。[+N] 素性を持つ節主要部（C[+N]）には、[Nom/Gen、Adnom]素性が与えられているとする。それが TP の主要部 T を経由して、主語に、主格または属格を、随意に認可し、述部には連体形を認可して連体形節を派生する（時制素性[Tns]はどちらにも共通にあるので省略した)。

---

[12] 金水・高山・岡崎・他(2011) pp.79.

[13] 節タイプ(あるいは「節ムード」)とは、平叙文(declaratives)、疑問文(Interrogatives)、命令文(directives)などを言う（cf. Narrog (2009) pp.135-158)。

[14] 活用形と句構造との関係については、研究の緒についたばかりである（cf. 三原 2011、三原・仁田 2012)。拙稿では、節タイプと独立形式の活用形との関係を提案している。

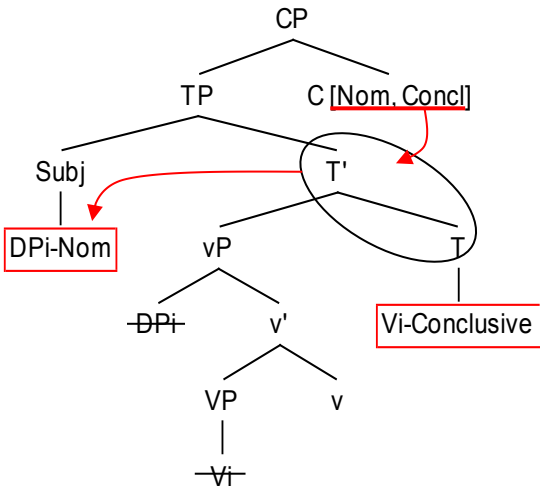


図４　終止形節（平叙文）の基本構造

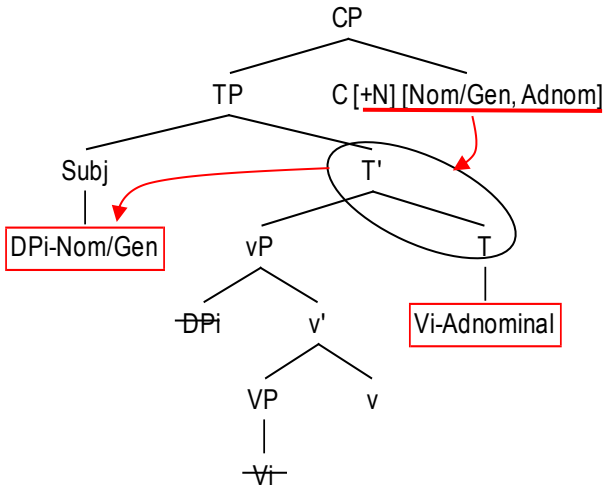


図５　連体形節の基本構造

このように、派生の段階で、節主要部として、名詞素性[+N]を持つ補文標識 C が選択されれば、その節(CP)は連体形節になり、属格主語が可能となる。

現代語の補語節では、、補文標識 C に有声の「の」の存在が一般的である。形容動詞を除いて、終止形の連体形への吸収によって終止形と連体形の区別ができなくなったので、それらの区別と、節境界を明示するために「の」が挿入されたという説があるが、これは[+N]素性の音声化とも考えられる[15]。

６．　付随する問題

６．１　主格目的語

属格付与の特殊な場合として、（７）のよう状態動詞構文における主語あるいは目的語の属格表示がある。

（７）a.　　山田 **が** 英語 **が/？の** できること

　　b. ??山田 **の** 英語 **が/の** できること

内省としては、（７a）は許容の範囲にあるが、（７b）の許容性については意見が分かれるところである。

現代語コーパスの検索結果を（８～１０）に示す。

（８）「が～が」の組合せ

　a.　だれも **が** 予想 **が** つくように、盲流が各地に起こり、　王 文山（1997）

　b.　彼 **が** 気 **が** 済むまでゲームをやっている間　浦賀和宏（2001）

　c.　俺 **が** 関心 **が** あるのは、誤診だけだ。　北方健三（2001）

（９）「が～の」の組合せ

　a.　彼 **が** 人気 **の** あるわけがわかるような気がした。　立川ありあ（1992）

　b.　田中**が** 元気**の**いいころは田中の側近ナンバーワンを自認していた。
　　本澤二郎（1993）

　c.　一介の中将くらい **が** 歯**の** 立つ相手ではない。　水沢蝶児（1991）

[15] 外池滋生氏の示唆による。

（１０）「の～の」の組合せ

a.　自分たち**の**　納得**の**　できることを、自分たちの見通しの中でやれる
　山本冬彦（2002）
b.　「じゃ、私**の**　気**の**　済むようにさせて…ねっ」　　小林光恵（1998）
c.　浪人兵法者**の**　歯**の**　立つような手合いではない、　　司馬遼太郎（2003）

残念ながら、現在のところ、「の～が」の組合せが見つかっていない(筆者の見過ごし)。[[NP の] [NP が] VP] なる組合せは、[[NP の NP]が VP] の読みが最優先される解析上（parsing）の制約により実例が稀少なだけで、統語上の制約とは思われない。

結論として、格表示の四つの組合せに関しては、統語的には随意に選択できると仮定しておきたい。「の～が」の組合せは、認知(解析)上の制約があり、発話が困難と考える。

主語と目的語へ同時に格付与できるのは、「格付与の多重一致理論」(Multiple Agree (Hiraiwa 2001a)) の適用例と考える。

６．２　他動性制約

現代語では、（１１）のような、属格主語のあとにヲ格目的語が介在すると容認度が著しく低下すると言われている。 これは、統語的制約か解析上の制約か、いろいろ議論されてきたが、未だに明確な説明は見当たらない。Watanabe(1994)、Hiraiwa(2001b)では、他動性制約と称して、統語的な制約として議論している。

（１１）　山田 **が/*の**　本を買った店

中古・現代コーパスの検索結果を（１２）に示す。

（１２）「属格主語～ヲ格目的語」の組合せ

a.　秋田市では、博士**を**　蝶**の**　取り巻くこと、大略斯の通りであった。泉 鏡花(2004)
b.　人**の**　花　**φ**見たる形かけるをよめる（φ：ゼロ対格）　　古今和歌集
c.　朱買臣　**が**　妻**を**　教えけむ年には…　　枕草子
d.　隣の家より、風**の**　雪**を**　吹き越しけるを見て、　　古今和歌集

確かに、現代語では、例文を見つけるのは困難で、倒置文一件（１２a）だけであるが[16]、中古語では容易に探し出せる（１２b.c.d.）。従って、統語的制約ではなく、解析上の制約であると考えることが出来る。

属格主語と述語との間に介在物があると、容認度が低下するのは、連体修飾としての「の」が卓立しているため、名詞等が後続すると主語としての解釈が困難となるためと思われる。

７．おわりに

中古語と現代語のコーパスを使った対照分析と、内省だけに頼らない用例検索により、次のことを主張した。

① 連体形節における、「主格／属格」交替現象の本質は、中古と現代とでは何の変化も生じていない。従って、現象が生ずるメカニズムも同じである。

[16] Frellesvig((2011) pp.130)によれば、上代においては、主語とヲ格目的語が共存するときは、必ず、ヲ格目的語が主語より前に位置するとのことであるが、現代語と似ていて興味深い。更なる探究が必要である。

② 格交替は、名詞素性をもつ連体形節そのものの機能に由来する現象である。
③ 主語と述語との間の、目的語などの介在物も、属格主語の派生を妨げるものではない。妨げがあるとすれば、認知的（解析上の）条件である。
④ ただし、属格による主語表示が全く随意であるかは、検討の余地があり、今後の検討課題である（これこそコーパスの出番である）。

このような結果が得られたのも、身近に利用できる通時コーパスのお陰である。益々整備が進み、利用者が多くなれば、すばらしい成果も増えることが期待できる。

文献

B. Frellesvig(2010) *A History of the Japanese Language*, Cambridge U.P.

S. Harada(1971) “Ga-No Conversion and Idiorectal Variations in Japanese”『言語研究』60, pp.25-38. 日本言語学会

早田輝洋(2010)「上代語の動詞活用について」『水門― 言葉と歴史』22, 水門の会

K, Hiraiwa(2001a) “Multiple agree and the defect intervention constraint in Japanese” In *The proceedings of the HUMIT2000*, MITWPL#40, pp.67-80. Cambridge, MA.

K, Hiraiwa (2001b) “On nominative-genitive conversion” In *A view from Building E39,* MITWPL#39, pp.65-123. Cambridge, MA.

ホイットマン、ジョン(2009)「日本祖語の名詞化形と連体形及び已然形の再建」『日本言語学会 第138回大会 予稿集』pp.80-85.日本言語学会

菊田千春(2002)「が・の交替現象の非派生的分析－述語連体形の名詞性」『同志社大学英語英文学研究』74, pp.93-133.

金水 敏、高山善行、岡崎友子、他(2011)『シリーズ日本語史３ 文法史』 岩波書店

近藤康弘 (2000９『日本語記述文法の理論』ひつじ書房

益岡隆志・田窪行則(1992)『基礎日本語文法 - 改訂版 - 』くろしお出版

三上 卓 (1953)『現代語法序説』刀江書院 (復刊 1972 くろしお出版)

三原健一、仁田義雄 編 (2012)『活用の最前線』くろしお出版

三原健一 (2011)「活用形と句構造」『日本語文法』11-1, pp.71-87.

三原健一、平岩建 (2006)『新日本語の統語構造』松柏社

S, Miyagawa(1993) “LF case-checking and minimal link condition”, MITWPL #19 *Papers on case and agreement*, pp.213-254. Cambridge, MA.

H, Narrog (2009) *Modality of Japanese, Amsterdam*, John Benjamin P.C.

野村剛史(1993)「古代から中世の「の」と「が」」『日本語学』12-11, pp.23-33.

野村剛史(2011)『話言葉の日本史』吉川弘文館

大島資生 (2010)『日本語連体修飾構造の研究』ひつじ書房

A, Radford(1997) *Syntactic theory and the structure of English*, Cambridge U.P.

M. Saito(2012) “Case Checking/Valuation in Japanese: Move, Agree or Merge?” *Nanzan linguistics* 8, pp.109-207.

S. Tonoike(1991) “The comparative syntax of English and Japanese: Relating unrelated languages” In Heizo Nakajima(ed.) *Current English Linguistics in Japan,* pp.455-506. de Gruyter

外池滋生(2011) “A Proposed Excorporation Analysis of Head Movement and the Organization of Grammar” 慶應言語学コロキアム (2011/06/04)

角田太作(2012)『国研プロジェクトレビュー』9，pp3-11.　国立国語研究所
A，Watanabe（1994）“A cross-linguistic perspective on Japanese nominative -genitive conversion and its implecations for Japanese syntax” In *Current topics in English and Japanese*，ed. Masaru Nakamura，pp.341-369.　Hituzi Shobo
山田昌裕（2010）『格助詞「ガ」の通時的研究』ひつじ書房

関連 URL

国立国語研究所(2012)「日本語歴史コーパス」https://maro.ninjal.ac.jp/
国立国語研究（ 2009 ）「現代日本語書き言葉均衡コーパス」
https://chunagon.ninjal.ac.jp/

例文出典（引用順）

大西巨人(2003)『三位一体の神話』光文社
藤原万璃子(2002)『ワイルド・ローズ』心交社
えとう乱星（2005)『ほうけ奉行』ベストセラーズ
折原一(1992)『灰色の仮面』講談社
ロバート・A・スカラピノ(2003)『資料体系アジア・アフリカ国際関係政治社会史』パピルス出版
澁澤　龍彦（2003)『イコノエロティシズム』河出書房新社
王 文山（1997)『七つの中国』文藝春秋社
浦賀和宏（2001)『記憶の果て』講談社
北方謙三（2001)『小説現代』35:14，講談社
立木ありあ（1992)『恋愛の市場心理』講談社
本澤二郎（1993)『裏から見た自民党派閥』エール出版社
水沢蝶児（1991)『獅子と薔薇の銀河』朝日ソノラマ
山本冬彦（2002)『学童保育実践の記』川島書店
小林光恵（1998)『ぼけナース』メディアワークス
司馬遼太郎（2003)『大盗禅師』文藝春秋社
泉 鏡花（2004)『新編泉鏡花集』岩波書店

付録

（中古語・連体形節に於ける属格主語の検索フォーム）

「五国史」宣命のコーパス化

池田　幸恵（長崎大学）
須永　哲矢（昭和女子大学）

Construction of the Corpus of *Gokokushi-Senmyo*

Yukie Ikeda（Nagasaki University）
Tetsuya Sunaga（Showa Women's University）

１．はじめに

天皇の和文詔勅である宣命は、奈良時代から平安時代にかけての日本語を知る貴重な資料であるものの、校本・総索引のある続日本紀宣命を除き、日本語史研究において利用される機会は多くない。その理由としては、宣命が国史や古記録の記事中に分散して存在することや、続日本紀宣命以外の宣命には定本がなく、またそれぞれの宣命の読みが確定していないことなどが挙げられる。

発表者は、宣命の中でも、『続日本紀』から『日本三代実録』までの「五国史」宣命を日本語史の資料として広く利用できるようにすることを目指し、「五国史」に収められた宣命に読みを与え、全体をコーパス化することを計画している。

本発表では、宣命のコーパス化の指針と基本仕様を紹介し、コーパス化により宣命の表記や語彙の研究にどのような利点が生じるのかを示すことにより、研究利用という観点からの現時点における処理方針案を示すこととする。

２．「五国史」宣命のコーパス化の意義

２．１　既存の主な電子資料とその限界

宣命は、一見すると漢文文献のようであるものの、原則として日本語の語順に従い、付属語や用言の活用語尾を万葉仮名で小書しているため、その背後にある当時の日本語をある程度再現できる点で貴重な資料である。宣命の起源は漢文詔勅にあり、語彙・語法に漢文訓読の影響が見られる一方で、儀式の際に口頭で宣布されるという性格上、その語彙には口語性があることも指摘されており、日本語史の資料としての価値は高い（小谷 1986、山口 1993）。

宣命を含む「五国史」（「六国史」）の電子テキストとしては、星野聰・水野柳太郎氏により作成された電子データを XML 化し検索可能にした渡瀬茂氏の「XML による六国史検索の試み（試行版）」や日本文学電子図書館の「六国史〔全文〕」などがある。

しかしながら、これらのデータにおいては、本文テキストの仕様および、データ形式や検索条件の面で限界があり、幅広い利用には至っていない。まず、本文テキストの仕様では、原文の漢字文字列がそのままデータ化されているのみであり、読み下し文は作られていない。そもそも続日本紀宣命以外の宣命には定本が存在せず、読みが確定できていないものも多いため、各所に不確かさを抱えたまま読み下すよりは、もとの文字列をそのまま電子化するという処理は妥当であるといえる。しかし、このような形式のテキストは宣命

にふれる機会の多い専門研究者以外には扱いにくく、同時代の別資料を扱う中で、あるいはより広く通時的研究を行う中で宣命を参照したいという需要には応えられていない。

また、データ形式の面でも、構造化がなされていない単純なプレーンテキスト形式であるために、検索条件も文字列検索のみが可能となっている。漢字と万葉仮名の区別のために、万葉仮名部分を小さく表示したり、〈 〉に入れて表示するデータも存在するが、そのような処理はあくまで書籍同様、読む際の見た目を考えた処理に過ぎず、データ処理で語を取り出すという検索においては無力である。結局、既存の電子テキストでは、例えば漢字の「天」（「天皇・天津神」など）と万葉仮名の「天」（助詞の「て」）などは検索段階では区別することができず、検索結果を個々に確認していくしかない。

２．２　「五国史」宣命コーパス化計画のめざす形式とその意義

本研究では、データ化の中心となる宣命本文の表記を漢字仮名交じり文に改め、それぞれの語に読みを与えたうえで、現在構築が検討されている『日本語歴史コーパス』と共通性をもった形態論情報を付与したコーパスを構築することを計画している。

従来の漢字列のテキストではなく、漢字仮名交じりの「専門外でも読みやすい」テキストにすること、また、形態論情報等の、本文以外の様々な情報をも持たせた構造化テキストにすること、の 2 点が本計画の特徴である。

読みの定まっていない箇所もある本文に対し、全ての漢字に読みを与えたうえで漢字仮名交じり文にするという方式は、その処理の妥当性・確実性において不安を抱えることにもなる。しかし、そのような不安点を差し引いても、読み下し文を本文とした、専門外の研究者にも利用しやすい形式のデータを提供する意義は大きいと判断した。

また、読み下し文にすることの意義は、単に読みやすだけにとどまるものではない。国立国語研究所『現代日本語書き言葉均衡コーパス』の大きな利用価値の一つである、形態論情報の付与も、前提として読みが与えられていればこそ可能になるものである。読み下し文を中心に据えたテキストを構築することにより、専門外の研究者にも利用しやすくなり、かつ、形態論情報など種々の情報を付与することが可能となる。それにより、研究資源としての宣命利用が宣命研究者以外にも広がり、日本語史研究に大きな成果をもたらすことが期待できると考えている。

読み下し文にすることで宣命の原表記を変える点や読みの確実性への不安は生じようが、これについても、構造化テキストとして原表記の情報を付与する予定であり（**３．２**参照）、さらに将来的には原文との対応を図ることにより、解決可能だと考えている。

３．「五国史」宣命コーパスの基本仕様

３．１　本文テキストの基本仕様

前節でも述べたとおり、「五国史」宣命コーパスでの本文は、読み下し文にすることを考えている。

以下に示す**〈原文〉**は、宣命の原文の表記そのままに近い状態である。一見して明らかなように、宣命の表記形式は、「原則として日本語の語順に従い、自立語を大書し、付属語や用言の活用語尾を万葉仮名で小書する」と言っても、小書部分を平仮名に代えると、そのまま現代の漢字仮名交じり文になるというものではない。

宣命は、天皇の和文詔書である内容面の特殊性だけでなく、このような表記面での特殊性により、従来、専門の研究者以外には扱いにくい資料であった。宣命は、たとえ電子化

されたとしても、宣命小書体表記に慣れていない利用者が読み下すことは困難であり、また、語彙研究の一環としてその語彙を利用しようとしても、語を取り出すこと自体に習熟を要する資料であるといえる。

コーパス化の意義の一つは、その資料を利用しやすくすることにより、利用層を拡大することである。宣命の専門研究者のみならず、語彙や文法を研究する中で、さまざまな資料を横断的に検討するその一つとして宣命を利用できるようにするためには、原表記のままではなく、漢字仮名交じり文に読み下した方が有用であると判断した。

「五国史」宣命コーパスでの本文は、その次に掲げる**〈漢字仮名交じり文〉**のような形を基本とする。**〈原文〉**と比べれば、一般ユーザーにとっては、かなり読みやすいものであろう。このような形式で本文を整備することで、より利用しやすいコーパスになると考える。

例）『日本後紀』巻 22、延暦 23 年 10 月 12 日条

〈原文〉

天皇詔旨良万止勅命乎紀伊国司郡司公民陪従司々人等諸聞食止宣此月波閑時尓之弖国風御覧須時止奈毛常母聞所行須今御坐所乎御覧尓磯嶋毛奇麗久海潦毛清晏尓之弖御意母於多比尓御坐坐故是以御坐坐世留名草海部二郡乃百姓尓今年田租免賜比又国司国造二郡司良尓冠位上賜比治賜布目已下及郡司乃正六位上乃人尓波男一人尓位一階賜布又御座所尓近岐高年八十已上人等尓大物賜波久止詔布勅命乎衆聞食止宣

〈漢字仮名交じり文〉

天皇が詔旨らまと勅りたまふ命を紀伊国司郡司公民陪従へる司々の人等諸聞き食へと宣りたまふ。此の月は閑ある時にして国風を御覧す時となも常も聞所し行す。今御坐す所を御覧すに磯嶋も奇しく麗しく海潦も清く晏かにして御意もおだひに御坐坐す。故是以て御坐坐せる名草海部二郡の百姓に今年の田租免し賜ひ又国司国造二郡の司らに冠位上げ賜ひ治め賜ふ。目已下及び郡司の正六位上の人には男一人に位一階賜ふ。又御座す所に近き高年八十已上の人等に大物賜はくと詔りたまふ勅命を衆聞き食へと宣りたまふ。

また、原文を漢字仮名交じり文に読み下す際、また、その過程で漢字に読みを与えていく際にも、一定の指針を設け、テキスト全体としては均質な処理を施したものとなるよう努める。以下に、読み下し、および読みを与える際の基本指針を示す。

〈漢字仮名交じり文に読み下す際の指針〉

（1）一文ごとに句点を付ける。
（2）倒置表記は日本語の語順に改める。
（3）万葉仮名は平仮名にする。
　　例）詔旨良万止→詔旨らまと　　於多比尓→おだひに
（4）読み添える助詞や活用語尾も平仮名で表記する。
　　例）天皇大命→天皇が大命　　勅→勅りたまふ　　聞食→聞き食へ

〈読みを与える際の指針〉

個々の語に読みを与える際には、以下の読みを採用する。
（1）北川（1982）にある読み。

(2)「六国史」の古写本にある読み。
(3)『類聚名義抄』『色葉字類抄』などの古字書にある読み。

・北川（1982）にある読み

天皇（スメラミコト）、詔旨・命（オホミコト）、勅・宣（ノリタマフ）、国司（クニノミコトモチ）、郡司（コホリノミヤツコ）公民・百姓（オホミタカラ）、司々人等（ツカサヅカサノヒトドモ）、諸聞食・衆聞食（モロモロキキタマヘヨ）、此（コノ）、月（ツキ）、時（トキ）、御覧（ミソナハス）、常（ツネ）、聞所行（キコシメス）、今（イマ）、御坐・御坐坐（オホマシマス）、所（トコロ）、奇（アヤシ）、麗（ウルハシ）、清（キヨシ）、御意（ミココロ）、故（カレ）、是以（ココヲモチテ）、二郡（フタコホリ）、今年（コトシ）、田租（タヂカラ）、免賜（ユルシタマフ）、又（マタ）、冠（カガフリ）、位（クラヰ）、上賜（アゲタマフ）、治賜（ヲサメタマフ）、目（サクワン）、已（ヨリ）下（シモツカタ）、及（オヨビ）、正六位上（オホキムツノクラヰノカミツカタ）、男（ヲノコ）、一人（ヒトリ）、一階（ヒトシナ）、近（チカシ）、高年（トシタカキヒト）、八十（ヤソヂ）、上（カミツカタ）、大物（オホミモノ）

・「六国史」の古写本、『類聚名義抄』『色葉字類抄』など古字書にある読み

紀伊　紀伊 岐　（元和三年古活字本倭名類聚抄、巻 5、10ウ）

陪従　陪 ソフ マサシ サカヌ アリ ハムベリ マコト ツカフ マシフ クスク イヨタツ　（観智院本類聚名義抄、法中 23オ5）

従 シタカフ ヨリ ユルス ヲフ ホシイマヽ ソヒク ウツス ヨル キタル ツカフ コノム コトモナシ トモ アレイツ　（観名、仏上 23ウ1）

陪従 ベイジウ　（三巻本色葉字類抄、上 53ウ4）

陪従 ソヘ　（日本書紀、巻 2、鴨脚本・弘安本）

シタカヘ（日本書紀、巻 2、乾元本・丹鶴本）

閑 閑 シヅカナリ ミヤビカナリ ホノカナリ ウヤヒカナリ ヒラク トラフ ウルハシ ヲシフ ナラフ フセリ ノリ ホノメク イタヅラ ヒソカニ イトマ ナヲシ ナホナリ 禾ケン　（観名、法下 39オ6）

国風　跡さきとおなしやとりに行あひてかたるにこそは国ふりもきけ　（言継卿集･435）

磯　磯 イソ カト　（観名、法中 7オ6）

嶋　嶋 シマ　（観名、法上 55オ6）

海　海 ウミ 禾カイ　（観名、法上 2オ5）

潦　潦 ミゾ　（観名、法上 23ウ）

渚 ナキサ 潦 同　（色葉字類抄　中 32ウ1）

晏　晏 オソシ ハル ヤハラカナリ タケヌ ウレシ ヒタク クラシ シツカ ヤスラカ ヨロコフ　（観名、仏中 51ウ6）

名草　名草 奈久佐　（倭名類聚抄、巻 5、24ウ）

海部　海部 阿末　（倭名類聚抄、巻 5、24ウ）

国造　国（クニ） 造（ツコ）　（日本書紀、巻 22、岩崎本）

３．２　構造化テキストとしてのデータ形式

３．２．１　基本方式

「五国史」宣命コーパスは XML 形式の構造化テキストとする。**３．１**に示した本文テキ

スト（読み下し文）の文字列のみのデータではなく、本文テキストに対しマークアップ言語 XML でタグ付けをし情報をもたせることで、コーパスとしての利便性を高める。

例えば読み下して平仮名となった助詞や助動詞などの原表記（万葉仮名・漢文助字）の情報は、タグとして埋め込むことで、表面的には手の加わった漢字仮名交じり文であっても、データとしては手の加わる前の原表記を辿れるようにする。宣命本文を利用しやすい読み下し文に統一することも、宣命資料を構造化テキストタグ付きコーパスとすることも初の試みであるが、原表記の情報も保持するという構造化テキストであることをもって初めて、原文を読み下し文に大幅改編するという処理も許されよう。

３．２．２　宣命コーパスに付与される情報

正式なタグ名や XML タグセットの詳細は、今後の検討を経て確定させなければならないが、現時点では、各宣命を「詔」としてまとめ、「詔」の内部を「文」に区切り、さらに「文」を BCCWJ、『日本語歴史コーパス』等と同様の言語単位である「短単位」（単語）に区切る、という階層を主な構造として考えている。

タグによって付与される情報としては、以下に挙げるものを計画している。なお、以下のタグ例の書式は、理念の説明のために簡略化したものであり、実際の書式とは異なる。

（1）詔

宣命を詔単位でまとめる。詔番号、宣布日、国史名、宣命の型（宣下・奏上・僧綱宣）を付与。

・詔番号…続日本紀宣命は本居宣長『続紀歴朝詔詞解』の詔番号。『日本後紀』から『日本三代実録』までの「四国史」宣命は、馬場（1993）で付された詔番号。
・宣布日…当該宣命が収められた国史の記事の年月日。
・国史名…国史名と当該宣命の収められた巻数。
　例）続日本紀 1…『続日本紀』巻 1
・宣命の型
　宣下…天皇から臣下へ宣布するもの。宣命を「ノリタマフ（詔・勅・宣）」で結ぶ。
　奏上…天皇から神社・山陵に奏上するもの。宣命を「マヲス（奏・申）」で結ぶ。
　僧綱宣…僧官の補任に関するもの。宣命を「マヲス（白）」で結ぶ。

〈タグ例〉

<詔　詔番号="1" 宣布日="文武 1 年 8 月 17 日" 国史名="続日本紀 1" 型="宣下">

（2）文

詔の内部を文ごとに区切った単位。

〈タグ例〉

<文>天皇が詔旨らまと勅りたまふ命を紀伊国司郡司公民陪従へる司々の人等諸聞き食へと宣りたまふ。</文><文>此の月は閑ある時にして国風を御覧す時となも常も聞所し行す。</文>

（3）形態論情報

BCCWJ、『日本語歴史コーパス』と同様に、全ての文を短単位に区切って形態論情報を付与する。付与する形態論情報は品詞・活用型・活用形などである。コーパス内の全ての語

に対し情報を付与することになるが、この作業に関しては形態素解析辞書 Unidic による自動解析結果をもとに、人手で修正する予定である。

〈タグ例〉

<短単位 語彙素="治める" 品詞＝"動詞" 活用型="文語下二段" 活用形＝"連用形">治め</短単位>

<短単位 語彙素="給う" 品詞＝"動詞" 活用型="文語四段" 活用形＝"終止形">賜ふ</短単位>

(4) 読み

漢字に与えた読み。

〈タグ例〉

<読み 読み文字列="アキツミカミ">現御神</読み>

<読み 読み文字列="オホ">大</読み><読み 読み文字列="ヤシマ">八嶋</読み><読み 読み文字列="グニ">国</読み>

(5) 原表記

電子テキスト化に際して万葉仮名・漢文助字を平仮名表記に改めた場合の原表記。

〈タグ例〉

詔旨良万止→詔旨<原表記 原仮名="良万止">らまと</原表記>

天皇之→天皇<原表記 原助字="之">の</原表記>

令有尓→有ら<原表記 原助字="令">しむる</原表記><原表記 原仮名="尓">に</原表記>

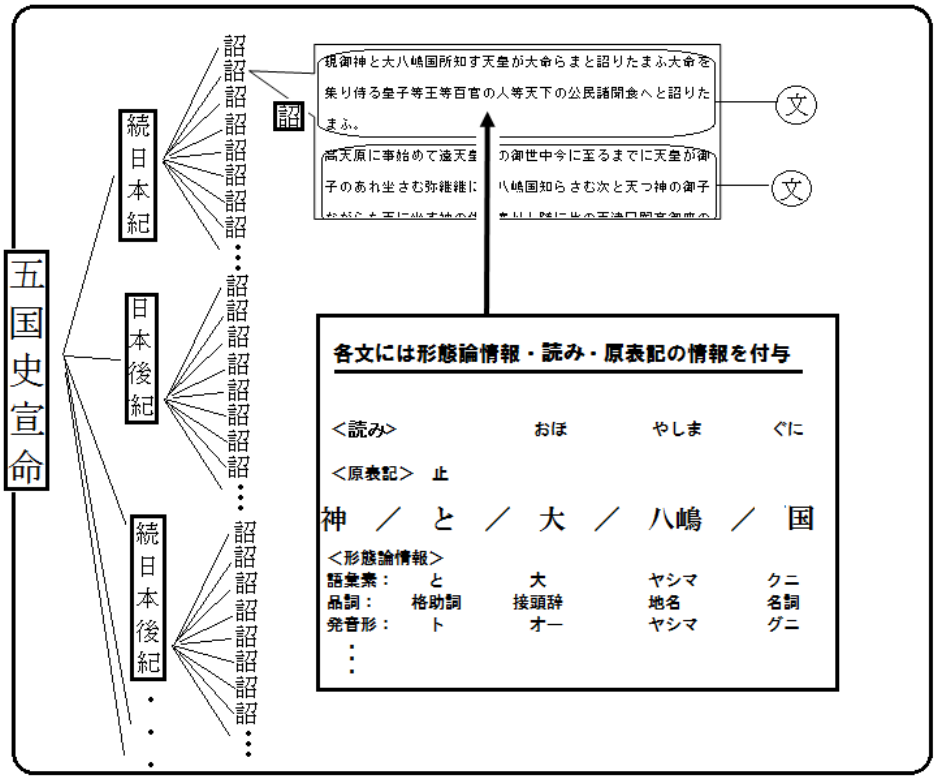


図１　宣命コーパスの階層構造のイメージ

<詔詔番号="66" 宣布日="延暦 23 年 10 月 12 日" 国史名="日本後紀 12" 型="宣下">
<文><読み 読み文字列="スメラ">天皇</読み>が<読み 読み文字列="オホミコト">詔旨</読み><原表記 原仮名="良万止">らまと</原表記><読み 読み文字列="ノ">勅</読み>りたまふ<読み 読み文字列="オホミコト">命</読み><原表記 原仮名="乎">を</原表記><読み 読み文字列="キノ">紀伊</読み><読み 読み文字列="クニノミコトモチ">国司</読み><読み 読み文字列="コホリノミヤツコ">郡司</読み><読み 読み文字列="オホミタカラ">公民</読み><読み 読み文字列="シタガ">陪従</読み>へる<読み 読み文字列="ツカサヅカサ">司々</読み>の<読み 読み文字列="ヒト">人</読み><読み 読み文字列="ドモ">等</読み><読み 読み文字列="モロモロ">諸</読み><読み 読み文字列="キ">聞</読み>き<読み 読み文字列="タマ">食</読み>へ<原表記 原仮名="止">と</原表記><読み 読み文字列="ノ">宣</読み>りたまふ。</文>

図２　XML データの例（試作段階、簡略化のため短単位タグは除いたもの）

４．宣命コーパスの利用例

これまで述べてきたように、宣命のコーパス化の意義の一つは、専門の研究者以外にも宣命資料を利用しやすくすることであるが、宣命研究者にとっても、用例採集や集計の効率化をはじめとして、多大な効果が期待される。以下では、宣命研究の立場から考えうる、コーパスの利用例を挙げる。

４．１　表記研究

(1) 宣命ごと国史ごとの万葉仮名の種類と使用頻度（簡単な字母表作成）
(2) 宣命による漢文助字と万葉仮名の使用頻度の違い
(3) 宣命における読み添え
(4) 仮名書き自立語

４．２　語彙研究

(1) 宣命における高頻度語と低頻度語
(2) 使用される形容詞・形容動詞の語形の変遷
(3) 形容詞・形容動詞をどのくらい重ねるか　例）明支浄支直支誠之心
(4) 動詞をどのくらい重ねるか　例）聞食驚伎悦備尊備念久波

５．公開形式

本コーパスの公開形式としては、以下の 2 種を考えている。

５．１　XML ファイルの利用形態

本文テキストに XML タグによって文章構造・形態論・表記に関する情報を付与した形式が、「宣命コーパス」本体としての基本データとなるが、XML ファイルそのものを扱えるユーザーはまだ限られているのが現状である。そこでまずは利用時の利便性、他資料利用との連動性を鑑み、『日本語歴史コーパス』の一部として、BCCWJ、『日本語歴史コーパス』同様、検索システム「中納言」で検索可能な形での公開を検討している。

５．２　総索引

特に古典語研究においては、このような電子データの利用そのものに不慣れなユーザーもいまだ多いのが現状であり、使い慣れた紙媒体での索引形式の需要も大きい。それにもかかわらず、続日本紀宣命以外には語彙総索引が存在していない。そこで形態素解析結果を用いた総索引作成システム（小木曽・須永 2010）を利用し、宣命コーパスのデータをもとに前後文脈つき総索引を作成し、電子データを使い慣れていないユーザーにも利用しやすい形で公開することも計画している。

付記

本研究は、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究（24320086、研究代表者：田中牧郎）及び、基盤研究（C）「宣命に使用される字音語についての再検討」（25370521、研究代表者：池田幸恵）による成果の一部です。

文献

小木曽智信・須永哲矢（2010）「「近世文語 UniDic」「中古和文 UniDic」を利用した総索引作成システムの開発」（『じんもんこん 2010　論文集』15、pp.119-124）

北川和秀（1982）『続日本紀宣命　校本・総索引』吉川弘文館

小谷博泰（1986）『木簡と宣命の国語学的研究』和泉書院

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」（『国立国語研究所報 122　雑誌『太陽』による確立期現代語の研究　『太陽コーパス』研究論文集　pp.1-48　博文館新社）

馬場治（1993）『五国史所載宣命の国語史的研究』（『telos』11、金沢経済大学人間科学研究所）

山口佳紀（1993）『古代日本文体史論考』有精堂出版

参考 URL

XML による六国史検索の試み（試行版）
http://www013.upp.so-net.ne.jp/wata/rikkokusi/index.html

日本文学電子図書館　六国史〔全文〕
http://www.j-texts.com/sheet/rikkoku.html

日本語歴史コーパス
http://www.ninjal.ac.jp/corpus_center/chj/

漢語名詞の副詞用法
～『現代日本語書き言葉均衡コーパス』『太陽コーパス』を用いて～

髙橋圭子（東洋大学）
東泉裕子（東京学芸大学）

Use of Sino-Japanese Nouns as Adverb:
Evidence from the Balanced Corpus of Contemporary Written Japanese (BCCWJ)
and the *Taiyo* Corpus

Keiko Takahashi (Toyo University)
Yuko Higashiizumi (Tokyo Gakugei University)

１．はじめに

現代日本語では、「結果」などの漢語名詞が、文頭または文中で副詞として使われることがあり、話し言葉から書き言葉にも広がりつつあるようである。例えば、『現代日本語書き言葉均衡コーパス（BCCWJ）』には、次のような用例がある。

(1) 通い続けている鍼治療院の院長から、身体の異変を指摘され、ガンが発見されたのだ。結果、早期治療に結びついた。（LBl4_00040、図書館・書籍、段勲『私はこうして「がん」を克服した：「がん」から生還した 22 人のドキュメント』、日本能率協会マネジメントセンター、1997 年[1]）
(2) それはまた生活の質的な向上にもつながり、結果、加齢とうまく折り合いがつけられるという考え方だ。（LBt4_00034、図書館・書籍、小野繁『ドクター・ショッピング：なぜ次々と医者を変えるのか』、新潮社、2005 年）

このような用法は、「その結果において」「～ル／タ結果として」などの副詞句の一部として使われていたものが、先行の「その」「～ル／タ」や後続の「として」「において」などが脱落し、単独で用いられるようになったものと考えられる。しかし、前後の要素が脱落しても意味や機能は基本的には変わらない。本発表では、「結果」のこのような用法を「文副詞的用法[2]」と呼び、『BCCWJ』および『太陽コーパス』のデータを比較することにより、名詞から副詞への用法拡張のさまを観察する。

２．先行研究

「結果」の副詞用法については、見坊（1988）、坂梨（2011）などにも用例の収集・報告がある。しかし、この用法の発生プロセスの実証的研究は、管見ではわずかである。

[1] 『BCCWJ』の用例には、サンプル ID、レジスター、出版年を付記する。さらに、書籍には執筆者、書名、出版社も記す。

[2] 原則として「文副詞」は文頭に位置するものだが、本発表では(2)のような文中のものも、文頭の「文副詞」につながる例と考え、あわせて、文副詞「的」用法と呼ぶことにする。

髙橋（2012）は、『BCCWJ』および『青空文庫』[3]をデータとして、名詞句「結果」「基本」「原則」「実際」「事実」「正直」「ある意味」の用法を調査した。いずれの名詞句も時の経過とともに文副詞化が進行しているが、その度合いは現代日本語においては(3)に示すように名詞句によって異なるという結果を得た。

(3)

「基本」	「結果」	「正直」	「実際」
「原則」		「ある意味」	「事実」
文副詞化が始まったばかりのもの	←――――→		文副詞化が進行・定着しているもの

東泉・髙橋（2013）は、「結果」「あげく」の用法を、『BCCWJ』および『太陽コーパス』にて調査し、『太陽コーパス』では多様であった中間的用法の表現が『BCCWJ』では固定化・定型化してきていることを確認した[4]。例えば、『太陽コーパス』で多用されていた「その結果として」に代わり『BCCWJ』では「その結果」が多用され、『太陽コーパス』に見られた「（その）結果において」は、『BCCWJ』では用いられなくなっている。

但し、『BCCWJ』を調査した両研究では検索ツールとして『少納言』が用いられており、検索に限界がある、ジャンルごとの分析がなされていない、などの不十分な点がある。また、「結果」の文副詞的用法の中には、因果関係の結果ではなく、単に時間の前後をつないでいるように解釈される用例もあった。「結果」という実質的意味の希薄化が起こっているようであり、文脈を考慮しつつ、質的に研究する必要がある。

３．理論的枠組み

高田他（2011）によれば、歴史語用論のこれまでの研究成果から、語の意味変化は、実質的意味を表すものから話し手の主観的な意味を表す方向への変化が、その逆方向より多いということである。例えば、英語の名詞 *fact* は、*in fact* という形で「実際において」という意味を表すようになり、やがて「たしかに」「しかしながら」という話し手の真実性に対する判断を表す副詞句となり、次いで「前述したことよりこれから述べることのほうが大切である」ということを知らせる談話標識（discourse marker）として用いられるようになったという。

また、Onodera (2004)によれば、「でも」「だけど」の談話標識としての用法は、それぞれ *V-te + mo* や *V + kedo* から拡張したという。このような用法の拡張プロセスは (4)のように表すことができる。

(4)　*V-te + mo* (14th–19th C)　＞　*Demo*　(18th–early 20th C)　＞　*Demo* (PDJ)
　　V + kedo (18th–early 20th C)　＞　*Dakedo*　(early 20th C–PDJ)
（Onodera 2004:111, 113 を改変）

[3] 検索サイト「日本語用例検索　http://www.let.osaka-u.ac.jp/~tanomura/kwic/aozora/」を利用した。

[4] 『太陽コーパス』は総合雑誌の全文コーパス、『BCCWJ』は書き言葉の均衡コーパスであるため、ジャンル差を考慮せずに単純に比較することはできない。しかし、より広いジャンルをカバーしている『BCCWJ』においてバリエーションが少ないということは、現代に向かって用法が固定化・定型化してきていると見なすことができよう。

同様に、「だって」（Mori 1996）や「だから」「なので」（東泉 2012）などの談話標識用法も、従属節末の用法から談話標識的用法へという拡張のプロセスをたどっているようである。「結果」も同じような拡張のプロセスをたどり、文副詞的用法がさらに談話標識的な用法へと、用法を拡張していくと予想される。

４．量的調査

総合雑誌『太陽』の記事を収録した『太陽コーパス』と比較するため、『BCCWJ』所収のデータのうち「総合雑誌」を検索対象とした。検索ツールは、『BCCWJ』はオンライン版『中納言 1.1.0』、『太陽コーパス』は同梱の全文検索システム『ひまわり』を用いた。両コーパスの「結果」の用法を表１にまとめる。また、『BCCWJ』の「総合雑誌」「知恵袋」「ブログ」の「結果」の用法をまとめたものが、表２である。

表１・表２ともに、「名詞」には、格助詞やコピュラが後続する例（省略されている例も含む）、「結果発表」「診断結果」のように複合名詞の一部として用いられている例をカウントした。「文副詞的」には、「結果として」「結果的に」「その結果」などと置き換え可能な例を、「名詞／副詞」にはどちらにも解釈可能な例をカウントした。また、「副詞句」の表現で「N」としたのは名詞句の意である。

表１　『BCCWJ』（総合雑誌）と『太陽コーパス』における「結果」の用法

分類		表現	BCCWJ 総合雑誌 2001-2005		太陽コーパス 1895-1925	
文副詞的		結果	26	3.5%	1	0.0%
名詞/副詞		結果	1	0.1%	0	0.0%
副詞句	副詞句 A（結果＋後続部分）	結果として	21	2.9%	0	0.0%
		結果において	0	0.0%	1	0.0%
		結果的に	59	8.0%	1	0.0%
		小計	80	10.9%	2	0.1%
	副詞句 B（先行部分＋結果）	こ/その結果	94	12.8%	308	8.0%
		N の結果	23	3.1%	445	11.5%
		こ/その N の結果	2	0.3%	0	0.0%
		ル/タ結果	48	6.5%	447	11.6%
		小計	167	22.8%	1200	31.1%
	副詞句 C（先行部分＋結果＋後続部分）	その結果として	7	1.0%	102	2.6%
		N の結果として	0	0.0%	147	3.8%
		そうした N の結果として	1	0.1%	0	0.0%
		ル/タ結果として	5	0.7%	74	1.9%
		その結果において	0	0.0%	5	0.1%
		小計	13	1.8%	328	8.5%
名詞		結果	446	60.8%	2326	60.2%
動詞		結果する	0	0.0%	4	0.1%
合計			733	100.0%	3861	100.0%

表１の「副詞句」からは、東泉・髙橋（2013）で指摘された形式の固定化・定型化に加え、短縮化・抽象化が進行していることがうかがえる。例えば、「先行部分＋結果＋後続部分」という最も冗長な形式の「副詞句 C」はいずれの表現も減少している。また、「先行部分＋結果」という形式の「副詞句 B」においても、「N の結果」「ル／タ結果」は減少し、より短縮化・抽象化された「こ／その結果」が増加している。

表２　『BCCWJ』総合雑誌・知恵袋・ブログにおける「結果」の用法

分類		表現	総合雑誌 2001-2005		知恵袋 2005		ブログ 2008	
文副詞的		結果	26	3.5%	108	7.6%	134	4.7%
名詞/副詞		結果	1	0.1%	23	1.6%	40	1.4%
副詞句	副詞句 A（結果＋後続部分）	結果として	21	2.9%	36	2.5%	74	2.6%
		結果において	0	0.0%	0	0.0%	0	0.0%
		結果的に	59	8.0%	107	7.6%	131	4.6%
		小計	80	10.9%	143	10.1%	205	7.2%
	副詞句 B（先行部分＋結果）	こ/その結果	94	12.8%	65	4.6%	99	3.5%
		N の結果	23	3.1%	50	3.5%	69	2.4%
		こ/その N の結果	2	0.3%	0	0.0%	2	0.1%
		ル/タ結果	48	6.5%	105	7.4%	121	4.2%
		という結果	0	0.0%	0	0.0%	1	0.0%
		小計	167	22.8%	220	15.6%	292	10.3%
	副詞句 C（先行部分＋結果＋後続部分）	その結果として	7	1.0%	4	0.3%	7	0.2%
		N の結果として	0	0.0%	1	0.1%	3	0.1%
		そうした N の結果として	1	0.1%	0	0.0%	0	0.0%
		ル/タ結果として	5	0.7%	1	0.1%	4	0.1%
		結果と	0	0.0%	1	0.1%	0	0.0%
		その結果に	0	0.0%	0	0.0%	1	0.0%
		その結果において	0	0.0%	0	0.0%	0	0.0%
		小計	13	1.8%	7	0.5%	15	0.5%
名詞		結果	446	60.8%	911	64.5%	2162	75.9%
動詞		結果する	0	0.0%	0	0.0%	0	0.0%
合計			733	100.0%	1412	100.0%	2848	100.0%

表２からは、「総合雑誌」より「知恵袋」「ブログ」において、「文副詞的」用法や「名詞／副詞」の例が多いことがわかる。「知恵袋」「ブログ」は、『BCCWJ』の中でも話し言葉に近いレジスターであり、「結果」の文副詞的用法が話し言葉にも書き言葉にも広がっているさまがうかがえる。

５．質的分析

５．１　名詞と副詞の連続性

「結果」の名詞から副詞への用法拡張を考えるヒントとして、ここでは、名詞と文副詞のどちらにも解釈可能な例を検討する。これには、例えば、次のようなものがある。

(5) 投資塩漬けにして上がるのを待つより、どんどん上がっている株に投資していく方が結果儲かります。（OC03_02100、特定目的・知恵袋、2005 年）
(6) 金額は約６万円で結果プラス二十五万になりました。店が遠隔で出してくれたのでしょうか？（OC15_02262、特定目的・知恵袋、2005 年）
(7) トホホホホです。まあ、そんなことで今年のＮＨＫマイルカップは結果残念でしたが、（OY15_01106、特定目的・ブログ、2008 年）
(8) 俺は３点ビハインドで先頭打者が２回も出塁したのに強行作戦を敢行して結果失敗カーッ（゜Д゜≡゜Д゜）、ペッ残塁８つって・・・（OY15_21256、特定目的・ブログ、2008 年）
(9) 近くの個人病院（泌尿器科）で検査。所見：前立腺肥大の疑い有りとのことで血液検査。但し結果異常ない。医師談：疲労からでしょうかね。（LBr9_00153、図書館・書籍、西出真由美『がんばって！っていわないで。：がん患者 180 の本音』主婦と生活社、2003 年）
(10) また、体験談などありましたら結果どうであれ教えて下さい。（OC09_00059、特定目的・知恵袋、2005 年）

これらの例の「結果」はいずれも、後続の助詞「は」ないし「が」が省略された名詞の用例とも、「結果的に」「結果として」の意の文副詞的用例とも解釈できる。

こういった例は、「総合雑誌」「知恵袋」「ブログ」いずれのレジスターにおいても、スポーツ、ギャンブル、投資、病気の診断、といった話題の記事に多く見られ、特に、スポーツ、ギャンブルの話題では、次の(11)・(12)のように、名詞の並置で文が構成されている用例が多い。一方、(5)〜(10)は、名詞並置でない一般的な構造の文であり、その中の「結果」は名詞とも文副詞的とも解釈できる。しかし、「結果」の表す意味・機能自体は、(5)〜(10)も(11)・(12)もほぼ同様ではないだろうか。

(11) スパーキングレディーカップ【結果】　川崎ダ千六百ｍ　　□メイショウバトラー　牝８／鹿毛　　２着逆らわなくてよかった（＾ω＾）（OY15_04928、特定目的・ブログ、2008 年）
(12) 浦和十一Ｒ□７コアレスデジタル３連単７→４、十→３、４、８、十結果　８→５→４　はずれ…（OY15_15534、特定目的・ブログ、2008 年）

また、名詞と文副詞的用法の両方の解釈が可能な例には次のようなものもある。

(13) 歯茎の検査をされまして．．．これが長い、痛い（笑　結果．．．前回よりはあまり良くないとのこと。　（OY15_10541、特定目的・ブログ、2008 年）
(14) あ・・・そぅやｗｗテスト無事今日終わりました〜〜〜ｗｗ　結果・・・聞きたくないなぁ・・・・　（OY14_04018、特定目的・ブログ、2008 年）

(15) あいのりのジュンペイは結果どうなったのですか？最近バイトが多忙で見過ごしてしまいました。（OC01_06891、特定目的・知恵袋、2005 年）
(16) 「とりあえず胃カメラの予約をしときましょう」早くて一週間、遅いと数カ月後に検査となります。そして結果「胃潰瘍がかなり進んでいます。手術をしたほうがいいでしょう」と癌と判明したにもかかわらず、このように説明されます（PB14_00057、出版・書籍、寺下謙三『プライベートドクターを持つということ：新主侍医制度』同友館、2001 年）
(17) 「煮詰まったら外を見よ」とゆう格言が僕の中にありこれがかなりの確率でうまくいく結果一番の近道となる・・・・ことが多いさらに今回タイミング良くお声をかけて頂けていたのだこのたびは（OY15_22114、特定目的・ブログ、2008 年）

(13)・(14)は「結果」のあとに間をおき、結果発表に向け期待と緊張感を高める話し言葉的技法が用いられていると考えられる。(15)・(16)は、名詞として助詞を補うと次のようになるが、不自然さは否定できず、「結果的に」「結果として」の意の文副詞的用法に近づいている例と考えられる。

(15') あいのりのジュンペイは結果はどうなったのですか？
(15'') あいのりのジュンペイの結果はどうなったのですか？
(16') そして結果は「「胃潰瘍がかなり進んでいます。手術をしたほうがいいでしょう」と、癌と判明したにもかかわらず、このように説明されます

(17)は、句読点が付されていないため、名詞・文副詞的用法の両様に解釈できる例である。これも、話し言葉と共通する特徴である。

５．２　実質的意味の希薄化

「結果」の辞書的意味は、ある原因によって達した結末の状態をいう。しかし、文副詞的用法の「結果」には、このような実質的意味が希薄化している例が見受けられる。

(18) この件に近い内容で、結局居住者の承諾を取らず無断で立ち入った案件がありました。結果、居住者は 300 万円相当の腕時計と指輪がなくなったと主張し警察を呼びました。（OC08_01946、特定目的・知恵袋、2005 年）
(19) 『何が皆の成功を邪魔するのか』→□成功体験八割は人の心の在り方を変える→その結果自分の人生を創造するように成る→結果として人生の色んな出来事に振り回されずに済むんだ （OY14_01129、特定目的・ブログ、2008 年）
(20) これだけ借金作っといて私たちはこれっぽっちも悪くないのに痛みだけ分かち合おうと？？結果、増税？？ふざけんのもいー加減にしてくれ。（OC05_01146、特定目的・知恵袋、2005 年）
(21) シミはどこにできても、なっかなか薄くはならないさ…二十一世紀は男も女も黙って美白！結果自然にやけてブロンズカラーになるのは仕方無いんだろうけど…シミはＮＧだ！（OY14_13062、特定目的・ブログ、2008 年）

(18)は、東泉・髙橋（2013）において、因果関係の結果を表すわけではなく、「そして」

のように単に時間の前後関係をつないでいるだけのような例と指摘されたものである。(18)の「結局」「結果」も、(19)の副詞句「その結果」「結果として」も、(19)の「→」と同様、「そして」「そうすると」のように継起する事態を次々に述べるため用いられており、いずれも結果・結末と言える状態には達していない。(20)も、痛み分けの次の段階として増税があるのであり、増税の先にはさらに次の段階がある。決して、増税が最終段階であるわけではない。(21)は、「結果」の前後に因果関係は全く認められないケースである。

６．考察

「結果」の名詞から文副詞的用法への拡張のプロセスを論じるにはまだ調査不足ではあるが、拡張の過程に「その結果」「ル／タ結果として」などの副詞句が関わっており、やがてまず話し言葉において前後の要素が脱落し単独で用いられるようになり、それが書き言葉においても話し言葉に近い特徴を持つ「知恵袋」「ブログ」から、「雑誌」など一般の出版物にも広がりつつある様相の一端は跡付けられたと考える。

文副詞的用法の用例の大半は、脱落の前と意味・機能は変わらず、これが当初の用法であったと考えられるが、やがて、形式の短縮とともに意味・機能にも変化が生じるようになった。１つには、(13)・(14)に見られたような、「これから重大なことを発表する」といった意味合いの談話標識的機能が指摘できる。また、もう１つ、(18)から(21)で挙げたような、実質的意味の希薄化がある。

文副詞的用法の見られる漢語には、「結果」のほかに、「実際」「事実」「基本」「原則」などがある。しかし、和語に比べ漢語は、実質的意味の希薄化は生じにくい。「結果」は漢語ではあるが、使用頻度の高い漢語が和語と同様の現象を示す例は、接頭辞「お／ご」の使い分けなどいくつも挙げられる。文副詞的用法の「結果」の実質的意味の希薄化は、他の漢語より和語の談話標識に近い例と言えるだろう。

しかし、当然のことながら、「結果」には和語の談話標識と異なる点もある。先行研究によれば、「だから」「だって」「でも」などの和語は、従属節末から文頭の談話標識へ用法を拡張してきた、という。「結果」の場合、「ル／タ結果」などは従属節末での用法だが、「その結果」「結果として」といった従属節末以外の用法もかなりの比重を占める。和語や他の漢語との異同を慎重に確認しつつ、調査を進めて行く必要がある。

７．むすび

本発表では、『BCCWJ』『太陽コーパス』を用いて「結果」の文副詞的用法を調査し、名詞から文副詞的用法の拡張に関わる諸側面や、談話標識的用法・実質的意味の希薄化について、他の漢語や和語と比較しつつ観察した。

漢語名詞の文副詞的用法への拡張は、現在進行中の現象である。今後も調査・観察を続けたい。

謝　辞

本研究の一部は、第６回ひと・ことばフォーラムにおいて発表したものです。ご助言をくださった方々に感謝申し上げます。

文　献

見坊豪紀(1988)「結果（副詞的用法）」『現代日本語用例全集』、pp.41-42、筑摩書房

坂梨隆三(2011)「『おられる』の補遺と『ある意味』『ある種』の用例－付、『結果』『正直』」帝京大学日本文化学会『帝京日本文化論集』18、pp.1-33.

高田博行・椎名美智・小野寺典子編著(2011）『歴史語用論入門』大修館書店

髙橋圭子(2012)「コーパスにみる名詞句の文副詞的用法」第 10 回対照言語行動学研究会 (http://www.ryu.titech.ac.jp/~nohara/taishogengokoudou/files/abst10/abst10_5takahashi.pdf)

東泉裕子(2012)「日本語の発話の周辺部：理由を表す接続詞と接続助詞の競合の事例から」青山英語英文学研究会発表（2012 年 11 月 21 日青山学院大学）

東泉裕子・髙橋圭子(2013)「『結果、こういうことが言えそうです』～コーパスにみる名詞の文副詞的用法～」国立国語研究所『第３回コーパス日本語学ワークショップ予稿集』、pp.91-96. (http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_12.pdf)

Mori, Junko (1996) Historical Change of the Japanese Connective *Datte*: Its Form and Functions. In *Japanese/Korean Linguistics. 5*. pp.201-218.

Onodera, Noriko (2004) *Japanese Discourse Markers: Synchronic and Diachronic Discourse Analysis.* Amsterdam: Benjamins.

コーパス

国立国語研究所『現代日本語書き言葉均衡コーパス（BCCWJ）』

国立国語研究所(2005)『太陽コーパス』（国語研究所資料集 15）　博文館新社

口頭発表（3）

9月6日（金）　10:00～12:00

事象の活性化と不活性化を把握する言語資源の構築とその応用 ―災害時における問題報告と支援情報のマッチングを例に―

佐野 大樹　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）
イシュトバーン ヴァルガ　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）
鳥澤 健太郎　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）
橋本 力　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）
川田 拓也　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）
呉 鍾勲　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）
大竹 清敬　（(独）情報通信研究機構 ユニバーサルコミュニケーション研究所）

Construction and Application of Excitation Polarity Dictionary: An Illustration of Usage of the Linguistic Resource for Matching Problem Reports and Aid Messages under a Crisis Situation

Motoki Sano　(National Institute of Information and Communications Technology)
István Varga　(National Institute of Information and Communications Technology)
Kentaro Torisawa　(National Institute of Information and Communications Technology)
Chikara Hashimoto　(National Institute of Information and Communications Technology)
Takuya Kawada　(National Institute of Information and Communications Technology)
Jong-Hoon Oh　(National Institute of Information and Communications Technology)
Kiyonori Ohtake　(National Institute of Information and Communications Technology)

1　はじめに

本稿では、ビッグデータからの矛盾、因果、含意関係などの大規模獲得への活用を目的として構築されている『活性・不活性極性辞書』の概要、構築方法、及び、辞書の利用例について述べる。利用例については、東日本大震災の際に発信されたツイートを対象とした、問題報告ツイートと支援情報ツイートの自動マッチングシステムの構築における、活性・不活性極性辞書の役割について概説する。

本稿の構成は以下の通りである。まず 2 節にて活性・不活性極性について概説し、3 節にて辞書の概要、構築方法について述べる。4 節にて、辞書の利用例について紹介する。

2　活性・不活性極性

意味や機能を述部の分類基準のひとつとして用いる枠組みとしては、Frame Semantics(Fillmore, 1976, 1977)、Conceptual Semantics(Jackendoff, 1990)、Process Type(Halliday, 1985) などが、認知、構造、機能主義的立場から提唱されている。これらの枠組みは、述部がとる項の意味役割の相違点を相対的に捉え、この違いによって述部の意味的性質を細分化し、体系化するものと位置づけることが可能であるが、逆に、述部の意味的性質を集約する枠組みとして、活性・不活性極性という概念が Hashimoto, Torisawa, De Saeger, Oh, and Kazama (2012) によって提唱された。

この枠組みでは、「を生成する」「が消失する」などの助詞と述部の組み合わせ（以下、テンプレート）が、係り元となる名詞の主たる機能、もしくは、効果の発揮に関してどのような作用をもつかを基準として、活性、不活性、中立の 3 タイプに分類される。活性テンプレートには、「を引き起こす」「を使う」「を買う」など当該テンプレートを係り先とする名詞の主たる機能、効果、目的、役割、影

響が、準備あるいは活性化されることを含意するものが該当する。不活性テンプレートには、「を防ぐ」「を捨てる」「低下させる」など当該テンプレートを係り先とする名詞の主たる機能、効果、目的、役割、 影響が抑制あるいは不活性化されることを含意するものが該当する。中立テンプレートは活性でも不活性でもないもので、「を考える」「に比例する」「が言う」などが該当する。つまり、名詞の機能や効果をオンにする、もしくは、オンすることに貢献するような作用をもつものが活性であり、オフにする、もしくは、オフにすることに貢献するような作用をもつものが不活性となる。

名詞の機能に着目するという点で、活性・不活性極性は Pustejovsky (1995) の telic と agentive role に類似した概念と考えられるが、Pustejovsky の枠組みでは不活性について扱われていない。また、Talmy (1988, 2000) の Force Dynamics では、ある項（アゴニスト）が内在する傾向に対して他の項（アンタゴニスト）が及ぼす力関係に着目し、事象間の因果関係を説明しており、一方の力が他に対して作用する状態を活性、一方の力が他に対して作用することに失敗する状態を不活性と考えた場合、一部、活性・不活性極性の概念と共有する部分があると考えられる。しかしながら、活性・不活性極性は、テンプレートに係る名詞の状態を他の項の役割（もしくは、力関係）を問わず把握しようとするもので、2 項間の力関係に限定されずにテンプレートを活性、不活性、中立のいずれかに分類するものであり、この点で、Force Dynamics に比べて適用範囲が広いと考えられる。活性・不活性極性が、先述した通り、因果関係だけでなく、フレーズ間の矛盾 (Hashimoto et al., 2012) や含意関係 (Oh, Torisawa, Hashimoto, Sano, De Saeger, & Ohtake, 2013) の認識に有効なのも 2 項間の力関係に制約されない枠組みとなっているためであろう。また、活性・不活性極性では、名詞が機能したり使用されたりする状態を一種の終極的な状態と考え、そこへ向かうか、遠ざかるかでテンプレートの極性が決まるため、目的論的思考が Force Dynamics に比べて強く反映されていると言える。

さらに、「を改善する」などが活性テンプレート、「を縮退させる」などが不活性テンプレートと分類されることを踏まえると、活性・不活性極性と評価極性とを類似した概念と考える立場も想定できる。しかし、評価極性は評価対象となる事象の good/bad の極性を判定するのに対して、活性・不活性極性は、項の機能や目的に対する極性であるという点で事なる。したがって、例えば、「癌が悪化する」は評価極性は bad となるが、活性・不活性極性では、「癌」の機能が発揮されている状態を「が悪化する」があらわしていると考え、活性となる。

活性・不活性極性は新しい概念であるが、100 万件規模の矛盾関係や因果関係の知識獲得 (Hashimoto et al., 2012)、Why-QA(Oh et al., 2013)、後述する問題報告ツイートと支援情報ツイートのマッチングなど難易度の高いタスクの精度向上に貢献することが実証されている。また、Hashimoto et al. (2012) は、物理学における電子のスピンモデルを利用してテンプレートを自動獲得する方法も提案しており、約 1 万テンプレート規模の活性・不活性極性辞書を少数のシードテンプレートから自動生成している。しかし、Hashimoto et al. (2012) の手法では、出現頻度が高いテンプレートが、一部、獲得できていなかったため、Web6 億ページで高頻度のテンプレートに対して人手で判定を行い、活性・不活性極性辞書を構築・拡張した。活性・不活性極性辞書の拡張により、より多くの因果、矛盾、含意関係などの認識が可能になると考えられる。

3　活性・不活性辞書の構築

3.1　辞書の概要

活性・不活性極性辞書には、約 4 万件の活性・不活性テンプレートが収録されている。現在獲得できている活性・不活性テンプレートの内訳を表 1 に示す[1]。なお、一部のテンプレートは、名詞に

[1] 辞書は現在構築段階にあり、テンプレート数が公開時とは異なる可能性がある。なお、不活性テンプレート数が、活性テンプレート数に比べて少ない原因のひとつとしては、「を接続できない」「は使えない」など活性テンプレートが否定形を伴って、不活性をあらわす場合があるためではないかと考えられる。

よって活性とも不活性とも認識できる場合がある。例えば、「を修理する」は、名詞が「故障」などトラブルをあらわす表現の場合は不活性テンプレートとなるが、「車」など修理されることで機能・使用されやすい状態となる名詞の場合は活性テンプレートとなる。3.2.2 節に詳細を示すが、このようなテンプレートは、「multi」というタグが付与されており他のテンプレートと区別されている。

表 1: 活性・不活性極性辞書のテンプレートの内訳

分類	テンプレート数 (件)
活性テンプレート	32,860
不活性テンプレート	6,836

3.2 構築方法

3.2.1 データ

Hashimoto et al. (2012) の手法によって獲得されたテンプレートに加えて活性・不活性極性のアノテーション対象としたのは、Web6 億ページで出現頻度[2] 上位 4 万位までのテンプレートである[3]。大規模データにおけるテンプレートの出現頻度を使って対象を絞り、獲得できていなかった頻出テンプレートを網羅的にカバーすることが目的である。

3.2.2 アノテーション

テンプレートの活性・不活性極性の判定は、アノテータ 3 名（著者以外）が独立して行い、3 名中 2 名以上が同じ判定をした場合、その判定結果をテンプレートの極性とした。アノテータ 3 者間の κ 値 (Fleiss, 1971) は、0.58（中程度の一致）であった。テンプレートの判定に用いた分類カテゴリと定義を以下にあげる。基本的に、Hashimoto et al. (2012) で提案されたアノテーションスキーマ[4]に準拠する。

活性テンプレート　名詞の指す対象の主たる機能（効能、効果、目的、役割、作用、悪影響も含む影響）が活性化される、あるいは、活性化されている状態にある、または、活性化されるための準備がなされていることを含意、暗示するテンプレート。ここでいう「活性」には、名詞の指す対象の主たる機能（効能、効果、目的、役割、作用、悪影響も含む影響）、あるいは対象それ自身の存在、発生、顕現、出現、生成、維持、使用、利用、準備、整備、入手、所有、隆盛、成長、質や量の高度化、決定、助長、充足、成功、達成、勝利、それへの同調や迎合、それとの対面や出会い、それへの奉仕や貢献、その影響力や効力や機能の発揮、その機能等の発揮の可能性、が含まれる。

不活性テンプレート　名詞の指す対象の主たる機能（効能、効果、目的、役割、作用、悪影響も含む影響）が不活性化される、あるいは、不活性化されている状態にあることを含意、暗示するテンプレート。ここでいう「不活性」には、名詞の指す対象の主たる機能（効能、効果、目的、役割、作用、悪影響も含む影響）、あるいは対象それ自身の不在、消滅、喪失、衰弱、追放、除外、休止、停止、使用不能、抑止、入手不能、不調、低成長、質や量の低下、未定、不足、受難、失敗、不達成、不発、敗北、それへの違反、それとの離反や乖離、それとの矛盾や齟齬、その影響力や効力や機能の喪失、その機能等の発揮の不可能性、が含まれる。

multi テンプレート:活性ディフォルト　名詞によって活性/不活性のいずれかを示すテンプレート。活性テンプレートのほうが典型的な使用方法だと思う場合。

multi テンプレート:不活性ディフォルト　名詞によって活性/不活性のいずれかを示すテンプレート。不活性テンプレートのほうが典型的な使用方法だと思う場合。

中立テンプレート　名詞の指す対象の主たる機能（効能、効果、目的、役割、作用、悪影響も含む影響）が活性化／不活性化されることなく、活性化／不活性化とは無関係で含意も暗示もしないテンプレート。

アノテータには判定対象となるテンプレートのみが提示され、以下の手順で判定が行われた。

[2] 日本語係り受けデータベースの頻度を利用して計算した。https://alaginrc.nict.go.jp/images/documents/DEP_ALAGIN_V1_README.pdf

[3] Hashimoto et al. (2012) の手法で獲得されているテンプレートは除く。今後、追加で上位 5 万件までアノテーションを行う予定である。

[4] http://aclweb.org/supplementals/D/D12/D12-1057.Attachment.pdf

手順 **1** 作業対象となるテンプレートの係り元となる典型的な名詞の例をアノテータ各自が考え、作業ファイルに記載する。

手順 **2** 典型例としてあげた名詞を基準として当該テンプレートの活性・不活性極性を判定する。

手順 **3** multi テンプレートと判定した場合は、活性テンプレートとなる名詞と不活性テンプレートとなる名詞をそれぞれ作業ファイルに記載する。

「multi テンプレート:活性ディフォルト」と判定されたものは、活性・不活性極性辞書において活性テンプレートとして含まれ「multi」というタグが付与されている。同様に、「multi テンプレート:不活性ディフォルト」と判定されたものは、不活性テンプレートとして含まれ、「multi」タグが付与されている。なお、テンプレートの極性を判定する際に利用する名詞を選択する方法としては、名詞とテンプレートの PMI スコアや共起頻度を用いて判定に用いる名詞を自動付与する方法もある。しかし、事前調査で、そのような方法に比べてアノテータ各自が名詞を選出しアノテーションしたほうが κ 値が高かったため、この方法を用いた。

3.2.3　辞書の公開

活性・不活性極性辞書は、高度言語情報融合フォーラム (ALAGIN)[5]より、平成 25 年度中に公開する予定である。同フォーラムにて公開されている『動詞含意関係データベース』[6]や情報通信研究機構より公開されている『日本語 WordNet』[7]と併用することで、さらに多様なテンプレートをカバーできるものと思われる。

4　活性・不活性極性辞書の利用例

2 節にて述べた通り、活性・不活性極性辞書は多様な意味関係の認識に活用することができるが、ここでは、活性・不活性極性辞書を用いて、災害時に発信されるツイートから、問題報告や支援情報を自動認識し、さらに、これらの間の適切なマッチングペアを自動認識する手法[8]について紹介する。支援情報の認識と後述する場所・地名の認識手法の詳細については、スペースの都合上割愛する。

4.1　背景と目的

東日本大震災では通信手段に問題が発生し、「情報の空白地帯」が生じた。そのような中、電話やメールは機能していなかったが、Twitter は機能していた地域もあり、これを介して P1 のような被災者からの問題報告ツイートや A1[9]のような支援団体・ボランティアなどからの支援情報ツイートが発信された (Winn, 2011; Acar & Muraki, 2011; Sano, Varga, Kazama, & Torisawa, 2012)。

P1 友人が粉ミルクがなくて困っています。もし、仙台市で在庫がある店をしっている方いらっしゃいましたら、どうか教えてください。

A1 仙台の者ですが、○○（店名）にはまだ、水、粉ミルクが売っていました。

仮に、A1 を P1 の発信者に伝達できていたとすると、P1 の「友人」は粉ミルクの在庫がある店を知ることができ、問題を解決できた可能性がある。しかし、現実には、多量に発信された情報の中で問題報告ツイートや支援情報ツイートが埋没してしまい、被災者と支援者の間で共有できなかった場合もあった。そこで、問題報告ツイート、支援情報ツイートを自動認識し、さらに、自動で適切な問題報告ツイートと支援情報ツイートのペアをマッチングさせる手法を開発した。問題-支援マッチン

[5] https://alaginrc.nict.go.jp/
[6] https://alaginrc.nict.go.jp/images/documents/ENT_ALAGIN_V1.3.0_README.pdf
[7] http://nlpwww.nict.go.jp/wn-ja/
[8] ここで紹介する研究の詳細については、Varga, Sano, Torisawa, Hashimoto, Ohtake, Kawai, Oh, and Saeger (2013) を参照されたい。機械学習の素性として用いている活性・不活性辞書の規模が、本稿の実験では拡張されている。
[9] 本稿で用いるツイートの例では、店名や病院名などの固有名は、○○などで伏せ、() 内に何が記載されていたかを記述して掲載している。

グペアを被災者や支援者が閲覧することで、問題に対する解決法、もしくは、対処が行われたと報告された問題を確認できるようになり、これによって、被災者が支援者から発信される支援情報ツイートを用いて直面している問題を解決したり、複数の支援者が同じ問題に対応し資源や労力を浪費することを予防することに貢献できると考える。なお、P1 のような問題報告ツイート、A1 のような支援情報ツイート、P1-A1 ペアのような問題-支援マッチングペアを以下の通り定義する。

問題報告ツイート　対処が必要となる問題の発生、もしくは、発生の可能性を報告するツイート。

支援情報ツイート　(1) 問題に対する対処法となる行為や状況について情報提供する、もしくは、(2) 問題が解決したこと、もしくは、解決される予定であることを情報提供するツイート。

問題-支援マッチングペア　問題報告ツイートと支援情報ツイートの関係が、(1) 支援情報ツイートが問題報告ツイートにある問題に対してどのように対応できるのか示す場合、(2) 支援情報ツイートが問題報告ツイートにある問題が解決されたこと、もしくは、されることを示す場合、(3) 支援情報ツイートが問題報告ツイートにある問題の解決に貢献する情報を提供する場合、のいずれかに該当するツイートペア。

なお本研究では、問題報告ツイート、及び、支援情報ツイートには、問題や支援のマーカとして機能する名詞と述部の係り受け関係が含まれると仮定する。例えば、先述した P1 には「粉ミルクがない」、A1 には「粉ミルクが売っていた」という名詞と述部の係り受け関係がそれぞれ問題と支援のマーカとして含まれる[10]。このように問題の核となる係り受け関係（以下、問題核）や支援の核となる係り受け関係（以下、支援核）に着目することで、多量の問題-支援マッチングペアを認識することができる。なお以後、問題核、支援核は、⟨ 名詞, テンプレート ⟩ の形式で表記する。

さらに、問題報告ツイートと支援情報ツイートがマッチングできる場合は、P1⟨ 粉ミルク，が足りない ⟩A1⟨ 粉ミルク，を届ける ⟩ のように問題核・支援核が同じ名詞を共有するものとする。これら 2 つの仮定を設けたことで、問題核と支援核に含まれる述部（「が足りない」と「を届ける」など）の意味関係が、問題核と支援核のペアがマッチングペアとして成り立つか否かを判定する上で重要な鍵となる。この意味関係の把握に、本研究では活性・不活性極性辞書を用いる。

4.2　アプローチ

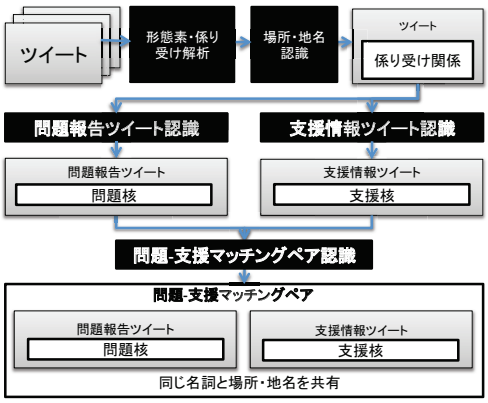


図 1: 問題-支援マッチングシステムの概要

本研究では、機械学習を用いて、問題報告ツイート、支援情報ツイート、問題-支援マッチングペアを認識するシステムを構築した。システムの全体像を図 1 に示す。まず、ツイート本文の形態素解析、係り受け解析を行い、場所・地名の認識を行う。次に、ツイートとこれに含まれる名詞と述部の係り受けのペア（以下、ツイート-核候補ペア）全てに対して、問題報告ツイート認識、及び、支援情報ツイート認識を行う。例えば、先述した P1 の「友人が粉ミルクがなくて困っています。もし、仙台市で在庫がある店をしっている方いらっしゃいましたら、どうか教えてください。」というツイートであれば、「粉ミルクがない」「在庫がある」など当該ツイートに含まれる名詞と述部の係り受け（問題核、もしくは、支援核の候補）とツイートをペアとして、問題報告ツイート認識、及び、支援情報ツイート認識のインプットとする。「粉ミルクがない」と P1、「在庫がある」と P1 は、それぞれ独

[10] 2011 年 3 月 10 日から 2011 年 4 月 4 日に発信されたツイート約 5,500 万件より、500 件をランダムサンプリングし、問題報告ツイートと支援情報ツイートを特定した結果、問題核を含まない問題報告ツイートは 4.5%、支援核を含まない支援情報ツイートは 9.1%であった。核も含まない例としては、名詞の羅列からなる「欲しいものリスト」などが該当する。

表 2: 出力結果の例

P2: いわきの○○病院（病院名）、いわき△△病院（病院名）、××クリニック（病院名）、●●クリニック（病院名）は、１７日から透析を中止します。患者の方は至急連絡してください。
A2: いわき○○病院（病院名）で短時間透析が可能です。受付時間は９時から１６時までです。
P3: ごめんなさい拡散をお願いしてもいいですか。仙台の父親の話ですと携帯の充電がもうない人が続出しているそうです。携帯充電器の支援が必要かと思われます。
A3: 【拡散希望】仙台若林区役所で携帯電話の充電ができるそうです。
P4: 浦安市では、２リットルペットボトルの飲料水、便袋、土のう袋が不足しています。全国の皆さんの救援をお待ちしています。
A4: 浦安市災害対策本部に向けて、土のう袋を発送した（｀´）ゞ。nお役に立ちますように。

自のインプットとして扱われる。最後に、前過程において問題報告ツイート、もしくは、支援情報ツイートと判定されたものをペアにして、問題-支援マッチングペアの認識を行う。出力結果の例を表 2 にあげる。

本研究では、核候補に含まれるテンプレート間の意味関係を活性・不活性極性を用いて把握する。活性・不活性極性を踏まえて問題核を分析してみると、⟨ 一酸化炭素中毒, に苦しむ ⟩、⟨ デマ, が拡散する ⟩ など (A) 名詞がトラブルをあらわす表現で、テンプレートが活性という構造か、⟨ 学校, が崩壊する ⟩、⟨ 充電, が切れる ⟩ など (B) 名詞が非トラブルをあらわす表現で、テンプレートが不活性という構造のいずれかを問題核はもつという傾向が伺えた。(A) に該当する問題核は、あるトラブルが発生、影響、促進されたりする出来事などをあらわすのに対して、(B) に該当する問題核は、物資や対処法などの非トラブルを不全状態にする出来事などをあらわす。この分析に基づき、**(A) か (B) の構造を含むツイートは、問題報告ツイートになる傾向があるという仮説を設けた。**

一方で、支援核を分析すると ⟨ インフルエンザ, が沈静化する ⟩、⟨ がれき, が撤去される ⟩ など (C) 名詞がトラブルをあらわす表現で、テンプレートが不活性という構造か、⟨ 仮設住宅, が建設される ⟩、⟨ インスリン, を届ける ⟩ など (D) 名詞が非トラブルをあらわす表現で、テンプレートが活性という構造を支援核はもつという傾向が伺えた。(C) に該当する支援核は、あるトラブルを不全状態にする出来事などをあらわすのに対して、(D) に該当する支援核は、物資や対処法などの非トラブルが発生、影響、促進されたりする出来事などをあらわす。この分析に基づき、**(C) か (D) の構造を含むツイートは、支援情報ツイートになる傾向があるという仮説を設定した。**表 3 に示す通り (A)〜(D) の構造は、活性/不活性へのテンプレートの分類とトラブル/非トラブルへの名詞の分類のマトリックスとして表現できる。以下、このマトリックスを**核構成マトリックス**と呼ぶ。

表 3: 核構成マトリックス

	トラブル	非トラブル
活性	(A) 問題核	(D) 支援核
不活性	(C) 支援核	(B) 問題核

このように問題核と支援核を特徴づけることのメリットは、問題-支援マッチングペアについて、問題核と支援核が適切なマッチングペアとなる場合、テンプレートの活性・不活性極性は逆になるという傾向を見出せるということにある。例えば P5 と A5 を考慮すると、

P5 いわきに戻ろうと思うんだけど、水道はまだ復旧してないらしい。お風呂がまだダメなのは本当にきびしい。

A5 いわき市の○○寺のお風呂を開放しております。お寺の本堂の脇にある建物です。無料です。

P5 に記載されいている問題のひとつはお風呂に入れないことであるが、A5 の情報はこの問題を解決する、もしくは、解決に貢献する情報であるため、問題-支援マッチングペアと考えられる。P5 と A5 には、「お風呂」という名詞を共有する問題核-支援核ペア（⟨ お風呂, がダメ ⟩、⟨ お風呂, を開放する ⟩）を含んでおり、ここで問題核と支援核のテンプレートの極性は逆になっている。⟨X, がダメ

⟩ は、「お風呂」の機能が不全状態であることを示すため不活性テンプレートであるのに対して、⟨X, を開放する ⟩ は、「お風呂」の機能が活性化されるのに貢献する状態をあらわすので活性テンプレートである。このことは、核の名詞がトラブルをあらわす場合でも成り立つ。例えば名詞が「インフルエンザ」の場合で、問題核の名詞となる場合は、⟨ インフルエンザ, が流行る ⟩ のようにテンプレートは活性になるはずである。一方で、支援核の名詞となる場合は、⟨ インフルエンザ, を撲滅する ⟩ のようにテンプレートは不活性となり、問題核と支援核の活性・不活性極性は逆になる。

以上のように問題核、支援核、問題-支援マッチングペアを活性・不活性極性とトラブル表現を用いて特徴づけることができるが、このような特徴は学習データや評価データを作成する際、アノテータに伝えられておらず、これに反するものもデータには含まれる可能性がある。ここで提起した問題核、支援核、問題-支援マッチングペアの特徴は仮説として設けるもので、本研究は、これらの特徴を機械学習の素性として加え、問題報告ツイート、支援情報ツイート、問題-支援マッチングペアの認識の性能が向上するか否かを検証し、仮説の妥当性を評価する。

4.3　問題報告ツイート、支援情報ツイート、問題-支援マッチングペアの認識

4.3.1　問題報告ツイート、支援情報ツイートの認識

問題報告ツイートと支援情報ツイートの認識には、事前実験で性能が最も高かった線形カーネル Support Vector Machine (SVM) による教師あり学習を用いた。問題報告ツイートと支援情報ツイートの認識には同じ素性を用いており、大別すると、トラブル表現 (TR)、活性・不活性極性 (EX)、活性・不活性極性、及び、トラブル表現との組み合わせ (TR&EX)、語彙評価極性、核候補とその文脈の形態素、及び、係り受け情報、単語意味クラス、要求表現、場所・地名に関する素性がある。ここでは核構成マトリックスに関連する TR、EX、TR&EX について詳細を示す[11]。

A. トラブル表現 に関する素性 **(TR)**

TR は、核候補の名詞がトラブル表現か否かを判定するものである。この判定を行うために、半教師あり学習を用いた De Saeger, Torisawa, and Kazama (2008) の手法を用いてトラブル表現を収集した。収集した表現を人手で確認した結果、「津波」などの災害や「インフルエンザ」などの病気など、多様なトラブル表現 20,249 件が得られた[12]。TR の素性は、作成したトラブル表現辞書に核候補の名詞が含まれるか否かに基づき判定される。

B. 活性・不活性極性に関する素性 **(EX)**

EX は、核候補のテンプレートの活性・不活性極性を判定するものである。テンプレートの活性・不活性極性の判定には 3 節で説明した活性・不活性極性辞書を用いた。なお、「を使わない」「を削除しない」のようにテンプレートが否定表現を伴う場合は、⟨X, を使う ⟩、⟨X, を削除する ⟩ に与えられた活性・不活性極性を反転させることとした。

C. 活性・不活性極性、及び、トラブル表現との組み合わせに関する素性 **(TR&EX)**

TR&EX は、核候補が TR による名詞の分類と EX によるテンプレートの分類の可能な組み合わせのうちどれに該当するか判定する。核構成マトリックスの (A)⟨ トラブル, 活性テンプレート ⟩(B)⟨ 非トラブル, 不活性テンプレート ⟩ に該当すれば、問題報告ツイートとなる可能性が高い。

4.3.2　問題-支援マッチングペアの認識

問題-支援マッチングペアの認識には、問題報告ツイートと支援情報ツイートの場合と同様に、事前実験で最も性能が高かった線形カーネル SVM を用いた教師あり学習を利用した。機械学習の素性には、問題報告ツイートと支援情報ツイートの認識に用いた素性に加えて、活性・不活性極性 (EX)、

[11] その他の素性の詳細については、(Varga et al., 2013) を参照されたい。
[12] ALAGIN フォーラムにて公開。https://alaginrc.nict.go.jp/images/documents/trouble_v1_readme.pdf

トラブル表現と活性・不活性極性の組み合わせ (TR&EX)、類似度、矛盾関係、及び、問題報告ツイートと支援情報ツイート認識で得られた SVM スコアを用いた。ここでも、核構成マトリックスに関連する、EX、TR&EX についてのみ詳細を示す[13]。

A. 活性・不活性極性、及び、トラブル表現との組み合わせに関する素性 **(TR&EX)**

4.2 で述べた通り、問題報告ツイートと支援情報ツイートが、問題-支援マッチングペアとなる場合、問題核と支援核のテンプレートの極性が反対になる傾向があると考えられる。この傾向を捉えるため、テンプレートの極性が同じか反対かを判定する素性を加えた。また、核候補の名詞がトラブル表現か否か、問題核テンプレートの活性・不活性極性、支援核テンプレートの活性・不活性極性の掛け合わせを素性に加えて、問題核と支援核の名詞とテンプレートの関係が 4.2 章で述べたマッチングの条件に合致するか否かを捉えられるようにした。

4.4　実験

4.4.1　災害関連ツイートの収集

問題報告ツイートの認識、支援情報ツイートの認識、及び、問題-支援マッチングペアの認識について評価するため、実験データを用意した。データには、東日本大震災に関連するツイートを利用した。まず 2011 年 3 月 10 日から 2011 年 4 月 4 日までに発信されたツイートを収集し、ここから災害とは関係のないツイートを除くため、「物資」や「断水」などの災害関連用語、被災地の地名、東日本大震災関連のハッシュタグなどを含むキーワード約 300 件によるフィルタリングを行った。結果、5,500 万ツイートが得られた。これら全てのツイートに形態素・係り受け解析[14]を行い実験データとした。以下、このデータを「災害関連ツイート」とよぶ。

4.4.2　問題報告ツイートの認識

A. 学習データと評価データ

問題報告ツイート認識の学習・評価データには、災害関連ツイートから無作為抽出したツイートとこれに含まれる核候補のペア（ツイート-核候補ペア）を用いた。学習データは 13,000 件 (R)、評価データは 1,000 件 (T) のツイート-核候補ペアを含む。各ツイート-核候補ペアは、アノテータ 3 名（筆者以外）により、ツイートが問題報告ツイートで、かつ、核候補が問題核か、それ以外かが独立して判定されている。3 名中 2 名以上が問題報告ツイート・問題核と判定したツイート-核候補ペアをポジティブサンプルとし、3 名中 2 名以上がそう判定しなかったツイート-核候補ペアをネガティブサンプルとした。アノテータ 3 者間の判定の κ 値 (Fleiss, 1971) は、0.74（十分な一致）であった。

B. 実験 **1**

学習・評価データを作成した後、トラブル表現と活性・不活性極性に関する素性の効果を検証するために、提案手法と提案手法に用いた素性から TR、EX、TR&EX を除外して SVM を学習した場合それぞれの適合率、再現率、F 値（SVM スコア=0）を比較した。実験結果を図 2 に示す。X 軸は適合率、Y 軸は再現率を示す。提案手法 (proposed) の再現率は 51.52%、適合率は 75.80%、F 値は 61.34%であった。提案手法から TR、EX、TR&EX を除外して SVM を学習した場合 (proposed-TREX) の再現率は 43.72%、適合率は 72.14%、F 値は 54.44%であり、再現率、適合率、F 値で提案手法が上回った。提案手法のほうが F 値が 6.90%高いことから、TR、EX、TR&EX は問題報告ツイートの認識において有効な素性であると考えられる。また、同様の手法を用いている (Varga et al., 2013) では、活性テンプレート 7,848 件、不活性テンプレート 836 件の活性・不活性極性辞書を用いており、同じテストデータに対して実験を行った結果、再現率は 44.26%、適合率は 79.41%であった。本稿の提案

[13] その他の素性の詳細については、(Varga et al., 2013) を参照されたい。
[14] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN、http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

手法の性能と比較すると、提案手法では適合率が 3.61%低下したものの、再現率が 7.26%増加し、F 値は 4.51%向上した。活性・不活性辞書に頻出テンプレートが含まれるようになったことが、再現率・F 値の向上につながったと考えられる。

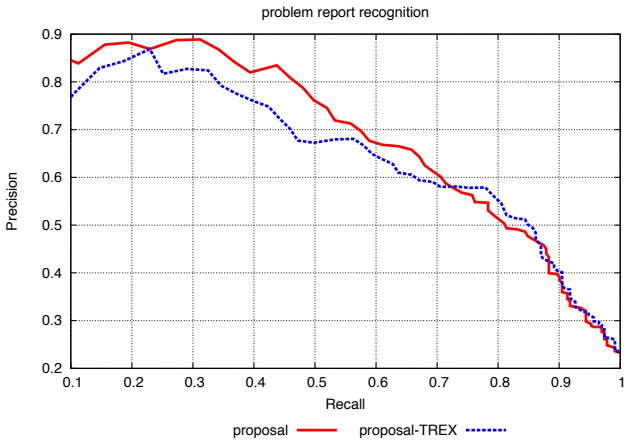


図 2: 問題報告ツイートの認識の再現率と適合率

4.4.3　問題-支援マッチングペアの認識

A. 学習データと評価データ

さらに、問題-支援マッチングペアの認識に関する実験を行った。学習データ、及び、評価データの作成のため、まず、問題報告ツイートと支援情報ツイートの認識を、災害関連ツイート 5,500 万件全てに行った。つぎに、問題報告ツイート、もしくは、支援情報ツイートの認識において SVM スコアが 0 以上であったものに対して、問題核と支援核の核候補の名詞が同じで、かつ、同じ場所・地名を含むのであれば、これらをペアとし、問題-支援マッチングペア候補とした。

学習データは、2 種類の方法でサンプリングした問題-支援マッチングペア候補 ($M1$・$M2$) を含む。$M1$ は、ひとつの問題報告ツイートに対する支援情報ツイートのバリエーションをカバーするために作成したもので、ひとつの問題報告ツイートに対して無作為抽出した最大 30 件の支援情報ツイートをペアとした。$M1$ には、問題-支援マッチングペア候補 3,000 件が含まれる。一方 $M2$ は、問題報告ツイートのバリエーションを考慮して作成したもので、ひとつの核候補の名詞に対して、最大 30 件の問題-支援マッチングペア候補が含まる。$M1$ と違い、ひとつの問題報告ツイートには無作為抽出された 1 件の問題-支援マッチングペア候補のみがペアとなる。$M2$ には、6,000 件の問題-支援マッチングペア候補が含まれる。評価データは $M2$ と同じ方法で作成し、多様な問題報告ツイートに対して適切な支援情報ツイートをマッチングできるか評価できるようにした。評価データには、1,000 件の問題-支援マッチングペア候補が含まれる。問題-支援マッチングペア候補は、アノテータ 3 名（筆者以外）により、当該候補が問題-支援マッチングペアであるか、それ以外かが独立して判定されている。3 名中 2 名以上が問題-支援マッチングペアと判定したものをポジティブサンプルとし、3 名中 2 名以上が問題-支援マッチングペアと判定しなかったものをネガティブサンプルとした。アノテータ 3 者間の κ 値 (Fleiss, 1971) は 0.63（十分な一致）であった。

B. 実験 **2**

トラブル表現と活性・不活性極性に関する素性の効果を検証するため、提案手法と提案手法に用いた素性から TR、EX、TR&EX を除外して SVM で学習した場合それぞれの適合率、再現率、F 値

を比較した。実験の結果、提案手法の再現率は 30.67%、適合率は 68.49%、F 値は 42.37%であった。一方で、提案手法から TR、EX、TR&EX を除外して SVM で学習した場合は、再現率は 28.83%、適合率は 67.14%、F 値は 40.34%であり、問題報告ツイートの認識に比べ差は小さいものの、再現率、適合率、F 値ともに、提案手法が上回った。提案手法のほうが F 値が 2.03%高いことから、TR、EX、TR&EX は、問題-支援マッチングペアの認識においても有効な素性であると考えられる。実験 1 と実験 2 から、トラブル表現とともに活性・不活性極性辞書を用いることで問題報告ツイートの認識と、問題-支援マッチングペアの認識の性能が向上すると考えられる。

5　まとめと今後の展望

本稿では、活性・不活性極性辞書の概要と構築方法、及び、当該辞書の利用方法について述べた。活性・不活性極性辞書は、因果関係、矛盾関係、含意関係などの意味関係の獲得だけでなく、問題報告ツイートや問題-支援マッチングペアの認識など、様々なタスクに活用できる言語資源として位置づけることができるだろう。

今後は、活性・不活性極性辞書の公開準備を進め、さらに、テンプレートのカバレッジが増えるように拡張していければと考えている。また、活性・不活性極性の概念を詳細化してプラン認識などにも活用できるよう、枠組みの細分化と体系化を進めている。

文献

Acar, A. & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *Int. J. Web Based Communities*, **7** (3), 392–402.

De Saeger, S., Torisawa, K., & Kazama, J. (2008). Looking for trouble. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pp. 185–192 Stroudsburg, PA, USA.

Fillmore, C. J. (1976). FRAME SEMANTICS AND THE NATURE OF LANGUAGE. *Annals of the New York Academy of Sciences*, **280** (1), 20–32.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In Zampolli, A. (Ed.), *Linguistic Structures Processing*, pp. 55–81. North-Holland, Amsterdam.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **5**, 378–382.

Halliday, M. (1985). *An Introduction to Functional Grammar*. Arnold, London.

Hashimoto, C., Torisawa, K., De Saeger, S., Oh, J.-H., & Kazama, J. (2012). Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 619–630 Stroudsburg, PA, USA.

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge.

Oh, J.-H., Torisawa, K., Hashimoto, C., Sano, M., De Saeger, S., & Ohtake, K. (2013). Why-Question Answering using Intra- and Inter-Sentential Causal Relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press, Cambridge.

Sano, M., Varga, I., Kazama, J., & Torisawa, K. (2012). Requests in tweets during a crisis: A systemic functional analysis of tweets on the Great East Japan Earthquake and the Fukushima Daiichi nuclear disaster. In *Papers from the 39th International Systemic Functional Congress*, pp. 135–140.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, **12** (1), 49 – 100.

Talmy, L. (2000). *Toward a Cognitive Semantics*. The MIT Press, Cambridge, London.

Varga, I., Sano, M., Torisawa, K., Hashimoto, C., Ohtake, K., Kawai, T., Oh, J.-H., & Saeger, S. D. (2013). Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Winn, P. (2011). Japan Tsunami Disaster: As Japan Scrambles. *Twitter reigns, Global Post*.

テキスト関連属性と助詞選択：計量的アプローチに基づく探索的研究

―主語・主題を導く「は」と「が」をめぐって―

石川　慎一郎（神戸大学国際コミュニケーションセンター／国際文化学研究科）

Japanese Particles *Wa* and *Ga* As Topic/ Subject Markers: A Quantitative Analysis of Text-related Factors on the Particle Choice

Shin'ichiro Ishikawa (Kobe University, SOLAC/ GSICS)

1. はじめに

各種の助詞のうち「は」と「が」の選択は先行研究において様々な角度から問題にされてきた。一般に，「は」が「文で述べようとする事柄を『…について言えば』といった気持ちで話題としてとりたて，それについての説明を導く」のに対し，「が」は「文法的に主語となることを表す」ものとされる（『明鏡』2 版）。これにより，「象は鼻が長い」型の構文や，「僕はうなぎだ」型の構文についても一応の文法的説明が成り立つ。

ただ，実際には，「は」と「が」の選択はそれほど単純ではなく，様々な文法的・意味的・語用論的要因の関与が認められる。野田（1996）や堀川（2010）による先行研究のまとめを整理すると，(A)主格支配範囲（文末までかかれば「は」，節内にのみかかれば「が」），(B)陳述主観性（話し手の主観判断を表明する判断文は「は」，具体的現象をありのまま表現する現象文は「が」，話題を持つ題目的有題文は「は」，話題を持たない平説的無題文は「が」），(C)情報新旧性（既知・既定・不可変の旧情報ないしテーマは「は」，未知・未定・可変の新情報ないしレーマは「が」），(D)主述関係（主格要素の性質を述語が解説する内包・記述・措定の場合は「は」，主格要素と述語名詞が同一であることを表す外延・同定・指定の場合は「は」または「が」），(E)主格要素意味特性（対比的・対照的・単独物的意味を持てば「は」，排他的・総記的・集合物的・強調的意味を持てば「が」）などの観点が主として「は」と「が」の使い分けに関与しているとされる。

野田（1996）は以上の諸原理の妥当性を認めつつも，それら各々が「別々の視点」に基づいているとし，(A')主題を持てるか（持てなければ「が」，持てれば(B')へ），(B')主題を持つか（持たなければ「が」，持てば(C')へ），(C')何を主題にするか（格成分ならば「は」，述語ならば(D')へ），(D')主題を明示するか（明示すれば「は」，暗示すれば「が」）の 4 種と，(E')どう取り立てるか（対比なら「は」，排他なら「が」）からなる統一的な説明モデルを提唱している。

ただ，こうした選択モデルをもってしても，たとえば「これはペンです」と「これがペンです」にまつわる微妙な差のすべてを説明することは困難を極める。実際，日本語教育学などでは，母語話者や学習者を対象として，文中助詞を空欄にして「は」または「が」を埋めさせる実験が広く行われているが，母語話者であっても判断にぶれが見られること

が報告されている。ゲオルギエバ（2008）の実験では，「は」ないし「が」を埋めるべき 46 の項目中，18 項目において母語話者の判断が「は」と「が」に分かれ，母語話者であっても主格を導く助詞が「択一的に選択されているとは限らない」と結論されている。

こうした状況を踏まえると，言語記述においてまずもって明らかにすべきことは，「は」と「が」の優先性ないしデフォルト性であろう。中川（1996）は「これまで説明に用いられてきた概念の対立に準拠し，2 つの助詞を差異化しようとすれば，2 つの選択肢のどちらを選ぶにしてもそれらの概念に応じた動機が必要」になるが，実際には「それが確定できない状況，どちらともつかず選ぶことができない状況」があるとし，そうした無条件の選択に「零度」を認めることで助詞選択の問題を単純化できると指摘している。その上で，「私／先週／京都／行きました」に助詞を添えて文にする場合，ほとんどの日本人が「は」を選ぶことを根拠として，ここでの「は」の選択は何らかの文法規則による判断ではなく，「完全に無条件の選択」であり，「は」こそが選択の「零度」で，これを使ってはいけない場合にのみ「が」が選ばれるのだと結論している。このようなモデルを採用すれば，日本語教育においても「学習者には単純な条件と規則を示」すことができる。

中川 (1996) の言う「は」の零度モデルは現象記述を単純化する上で有益なものだが，「は」の優先性が実際の言語使用においてどの程度一般的に再現されるかについては検証の必要がある。このとき，テキストに関連する言語的・非言語的属性を広く考慮に加えるべきであろう。たとえば，「私／先週／京都／行きました」の場合に「は」が優先的に選ばれるとして，「私」がその他の名詞であってもそうなのか，「私」の位置が文頭でなくてもそうなのか，テキストのタイプや時代に関わらずそうなのか，書き手の性別や年代に関わらずそうなのか，といった点についても確認の手順を踏むことが望ましい。

こうした問題を実証的に調査する場合，前述のように，従来は発話タスクや穴埋めタスクなどの誘引型手法が広く用いられてきたが，日本語における無条件・無意識の選択傾向を信頼できる形で抽出するには大規模な匿名データの解析が不可欠になる。本研究においては，「現代日本語書き言葉均衡コーパス」（以下 BCCWJ）を用い，「は」と「が」の選択におけるデフォルト性の問題を計量的観点から考察してみることとしたい。

2. リサーチデザイン

2. 1　目的と RQ

各種のコーパス研究により，言語の振る舞いは，テキストを取り巻く多様な要因によって複合的に影響されていることが示唆されている。本研究では，テキスト内・テキスト・テキスト外という 3 段階の階層要因モデルを仮定し，テキスト内レベルでは「は」と「が」を含む構文の言語特性として主格名詞の情報新旧性・位置・意味内容に，テキストレベルではテキストタイプ・年代・内容種別に，テキスト外レベルでは書き手の生年・性別にそれぞれ注目しつつ，「は」と「が」の頻度的優先性を多角的に検討してゆく。リサーチクエスチョンは下記の 5 点である。

RQ1 現代日本語の書き言葉全体で「は」と「が」はいずれが頻度上の標準であるか。

RQ2 テキスト内の言語的属性（(1)主格要素の情報新旧性，(2)位置，(3)意味内容）は助詞選択にどのような影響を与えるか。

RQ3 テキスト自身の非言語的属性（(1)テキストタイプ，(2)年代，(3)内容種別）は助詞選択にどのような影響を与えるか。

RQ4 テキスト外の非言語的属性（(1)書き手の生年，(2)性別）は助詞選択にどのような影響を与えるか。

RQ5 非言語的属性で助詞選択をモデル化することは可能か。

2．2 データと処理手順

データはBCCWJで，2013年7月に「中納言」インタフェースを介して調査を実施した。対象は主格要素（名詞・代名詞）の直後位置に出現する「は」と「が」で，検索は語彙素（ないし語彙素読み）単位で行った。このうち，「が」については，「水が飲みたい」や「画面が見にくい」のように他動詞と共起する例があるが，ここでは『明鏡』2版の定義に従い，これらも「主語を表す」用法とみなして他と同列に扱った。また，同辞書が「が」の別用法とする「名詞を修飾する」用法（例：我らが母校）と「同じ名詞をつなぐ」用法（例：親が親だから）については，BCCWJより無作為抽出した1,000例（500例×2回）を質的に検証した結果，当該用例の出現を認めなかったため，該当事例は無視できる程度に少ないものと判断し，検索で得られた値をそのまま使用することとした。このほか，BCCWJの形態素判定では，接続詞の「ところが」が「ところ」＋「が」と分析されている例があるなど（例：|ところ|(が)|、|私|の|職業|は|今|、|漫画|を|描く|こと|だ|。［LBr7_00041］），いくらか問題も認められるが，全体に占める比率は極小であるため，頻度修正は行っていない。なお，以下の議論において，「は」の零度性を検討する際には，「は」の頻度を「は」と「が」の合計頻度で割った「は」選択率を指標として使用する。

まず，RQ1（全般頻度）では，BCCWJ全体で「は」と「が」の総頻度を比較する。

RQ2（テキスト内言語的属性）では，後続動詞は助詞選択にほとんど関与しないとされていることから（ヨフコバ，2007），議論を構文の主格要素に限定した上で，(1)情報新旧性については，BCCWJ全体を対象として主格要素が名詞の場合と代名詞の場合を，名詞については連体詞「その」（語彙素読み）が共起する場合としない場合を比較する。(2)位置については，構文の安定性が高い新聞・雑誌・白書データを用い，主格要素が文頭に来る（直前句点共起）場合と文中に来る（直前句点共起なし）場合を比較する。(3)内容については，まず，BCCWJの全データを用いて主要代名詞別に比較する。ついで，2000年代に刊行された書籍（図書館）に限定して，名詞＋「は」（全188,256件のうちダウンロード上限の10万件），名詞＋「が」（全209,181件中の10万件），代名詞＋「は」（36,056件），代名詞＋「が」（15,977件），合計252,033件の用例データをダウンロードし，両助詞と高頻度に共起する主格要素（名詞・代名詞）を抽出して質的に比較する。

RQ3（テキストの非言語的属性）では，(1)テキストタイプについては，BCCWJを構成する11種を比較する（コア・非コアは区別せず，書籍は生産母集団・図書館母集団・ベスト

セラーを統合)。(2)年代については，長期データを保有する国会会議録と白書を対象に，1970年代，1980年代，1990年代，2000年代を区分して比較する。(3)内容種別については，上述の書籍の25万件データを用い，日本十進分類別に比較する。

RQ4（テキスト外非言語的属性）では，上述の書籍の25万件データの中から，書き手が単独で生年・性別が明示され，かつ，日本十進法分類が明らかな158,780件を抽出した上で，(1)書き手の生年および (2)性別を分けて比較する。なお，(2)に関しては，BCCWJの生年帯（10年単位）ごとのサンプル数に大きなばらつきがあるため，サンプル数が1,000件を超える1890年代生まれ～1970年代生まれ（157,167件）のデータに限定して分析する。

最後に，RQ5（非言語的属性によるモデル化）では，RQ4で使用した約16万件の書籍データに対して Weka v3.77 を用いた決定木分析を実行し，テキストおよびテキスト外レベルの非言語的属性による助詞選択モデルの作成を試みる。これにより，先行研究で注目されることの少なかった非言語的属性の助詞選択に対する影響を計り，あわせて，はっきりした言語的選択動機が「確定できない状況」下での助詞選択傾向を観察する。分析アルゴリズムはJ48（Quinlan の C4.5 に基づき，データを反復的に分割し，情報利得が最大になるものを選ぶ）とする。剪定（pruning）のための confidence factor は0.25，葉（leaf）あたりの最低のインスタンス数は500，相互検証のための fold 数は10とする。なお，言語研究における決定木分析の適用過程の詳細については石川（2013）他を参照されたい。

3. 結果と考察

3．1　RQ1「は」と「が」の全体的標準性

「は」と「が」の出現状況を概観するため，BCCWJ全体で検索したところ，名詞・代名詞に後続する「は」の頻度は1,950,445回，「が」の頻度は2,007,682回で，「は」選択率は49.28％であった。助詞としての「は」と「が」の選択は拮抗しているが，差は有意であり（G^2=844.4, df=1, p<0.1％），「は」の頻度上の優先性は確認されなかった。このことは「は」の零度モデルを日本語の全体に無条件に適用することの問題点を示唆している。

3．2　RQ2 テキスト内言語的属性の影響

3．2．1　情報の新旧性

先行研究の多くは，旧情報ないしテーマは「は」，新情報ないしレーマは「が」とする分類基準を提唱している。では，実際のデータにおいて，情報の新旧性による違いは確認できるのだろうか。

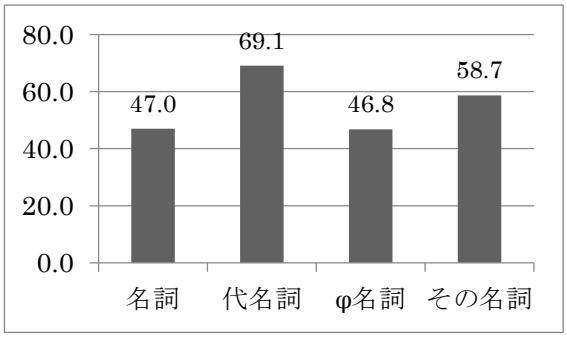


図1 主格要素タイプ別の「は」選択率

BCCWJ全体で名詞・代名詞別の検索を行い，次に，新聞・雑誌・白書に限って名詞前の「その」の共起の有無別に検索を行ったところ，それぞれの条件下での「は」選択率は図1のようになった。

一般に名詞に比べて代名詞が旧情報を指示しやすいこと，また，同じ名詞でも「その」が共起することで旧情報を指しやすいことをふまえて上記のデータを検証すると，主格要素が旧情報となることで「は」の選択傾向が顕著に高まることが確認された。これは，先行研究で広く言われている「は」と既知（扱い）の旧情報，「が」と未知（扱い）の新情報の親和性をデータ面で裏付ける結果と言えよう。

3．2．2　主格名詞位置

野田（1996）のモデルには明示的に含まれていないが，一部の先行研究は「は」と「が」の選択に主格要素の位置が関係する可能性を指摘しており，たとえばヨフコバ（2007）は，小説や学術論文などの文頭文を調査した結果，「圧倒的に『は』が使われている」としている。そこで，主格名詞が文頭に出現する場合（句点＋名詞＋助詞）と，文中に出現する場合（句点なし＋名詞＋助詞）を比較したところ，図 2 の結果が得られた。

右図に明らかなように，主格が文頭に出現する場合，「は」の選択率が高まり，異なるテキストタイプでも傾向は不変である。これは先行研究を広く支持する結果である。談話では，初めに旧情報である話題を提示し，次いでそれに関する新たな情報を付加してゆくのが語用論上自然な展開であり，このことが文頭位置での「は」の選択に影響していると考えられる。

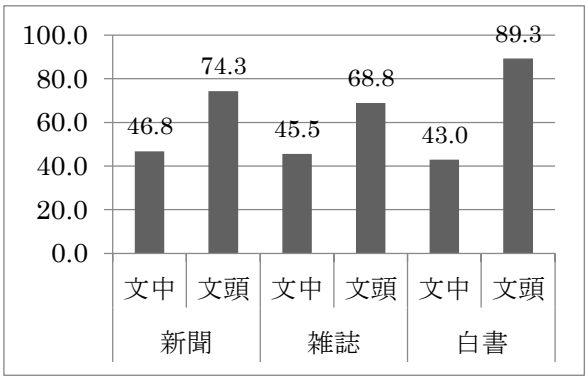


図 2　文頭・非文頭別の「は」選択率

3．2．3　主格要素意味内容

助詞選択に関して，後続する動詞タイプの影響はほとんどないとされているが，先行する名詞タイプの影響はどうであろうか。BCCWJ の全データを対象に，代表的な人称代名詞 3 種（「私」「彼」「彼女」）および非人称代名詞 3 種（「これ」「それ」「あれ」）をサンプルとして「は」選択率を調べたところ図 3 の結果を得た。

人称代名詞 3 種の「は」選択率の平均は 76.1%，非人称代名詞の場合は 73.7%で，主格要素が非人称化（モノ化）することで「は」選択率が低下する傾向が示唆されたようにも見えるが，個々の代名詞ごとに検証すると，傾向は可変的であり（「それ」を除くと，非人称であっても「は」選択率は高い），主格要素の人称・非人称性と助詞選択の間にはっきりした関係は検証されなかった。

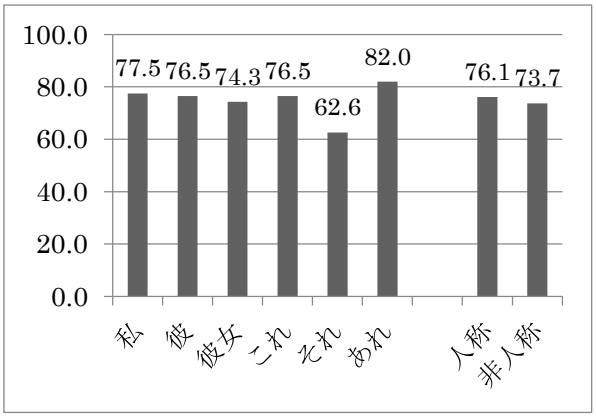


図 3　代名詞タイプ及び主要代名詞別の「は」選択率

続いて，書籍（図書館）の 25 万例の実例データより，「は」と「が」の主格要素のうち，高頻度語上位 30 語を抽出したところ，以下のようになった（「実は」「ほうが」「方が」などの非自立単位も含む）。

（は）私，それ，こと，これ，彼，わたし，彼女，人，僕，ぼく，もの，場合，あなた，とき，俺，実，男，おれ，あれ，ここ，われわれ，自分，今，今度，問題，人間，時，あたし，二人，日本

（が）こと，それ，私，これ，人，もの，彼，気，ほう，ところ，自分，方，何，わたし，あなた，彼女，声，男，必要，音，人間，ぼく，誰，時間，僕，問題，目，俺，言葉，関係

括弧で示したように 30 語中 17 語が重複しており，主格要素の内容には一定の重複性が見られた。ただ，「は」の場合は，「おれ」や「あたし」といった口語的 1 人称代名詞や「あれ」や「ここ」といったくだけた仮名表記代名詞が頻出する。一方，「が」の場合は，「声」「必要」「時間」「言葉」「関係」といった固い文脈で多用される非主観的で抽象的な概念語や，一部の先行研究で指摘されている「何」や「誰」といった疑問詞などが頻出する。以上より，具体的で口語的な談話環境では「は」が，抽象的で書き言葉的な談話環境では「が」が選択されやすいという大まかな方向性が示唆される結果となった。これは，話し手の主観判断を表明する判断文では「は」が，現象をありのままに表現する現象文では「が」が選ばれるという先行研究の見解を異なる観点からサポートするデータと言える。

3．3　RQ3 テキストの非言語的属性の影響

3．3．1　テキストタイプ

助詞選択に関する先行研究の多くは日本語が不可分の実体であるという前提に基づいており，テキストタイプの影響はほとんど考慮されてこなかったわけであるが，実際にはどのような差が見られるのであろうか。テキストタイプ別の分析結果は図 4 のようになった。

コーパス全体で見た「は」選択率は 49.28%であるが（3.1 節参照），特殊性の高い韻文（59.8%）を除くと，すべてのテキストタイプにおいて「は」選択率は 42.5%～51.6%の範囲におさまっており，テキストタイプの影響は比較的限定的であることが明らかになった。

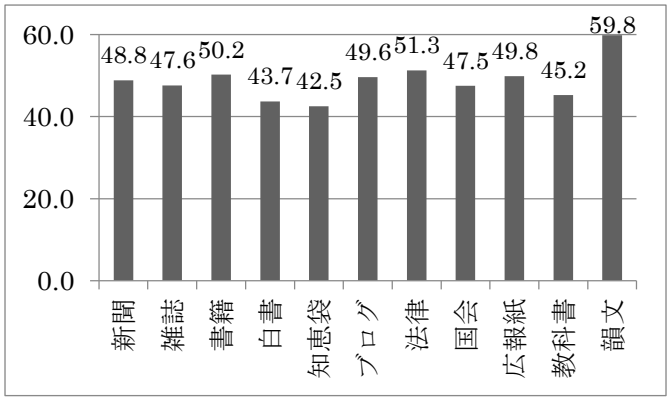


図 4 テキストタイプ別の「は」選択率

ただ，その範囲の中では一貫した差異の傾向も認められる。つまり，書籍・法律・広報紙・新聞など，単なる事実の羅列的伝達に終わらず，何らかの解釈・注

釈・判断が加わるテキストでは「は」の選択率が上昇し，知恵袋・白書・教科書のように，個人的・主観的な解釈が控えられ，具体的で客観的な情報や事実の透明な提示が主となるテキストでは「が」の選択が増えるのである。3.2.3 節で見たとおり，判断文では「は」が，現象文では「が」が選ばれるという傾向が反映された結果と言えるだろう。

3．3．2 テキスト年代

先行研究は年代についてほとんど考慮に入れていないが，「は」の零度性に年代は影響してないのであろうか。長期データが含まれる白書と国会に限って分析を行なったところ，図 5 の結果が得られた。

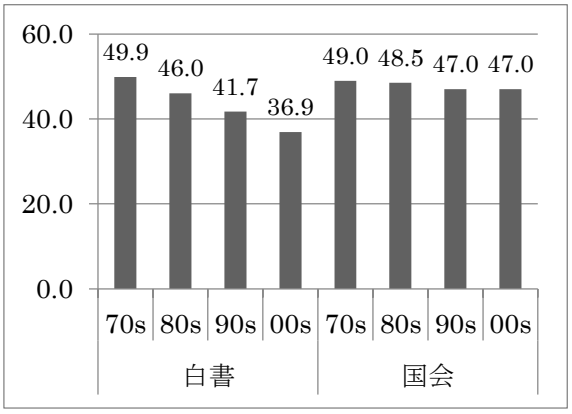


図 5 テキスト年代別の「は」選択率

「は」選択率は国会会議録においては微減しているだけだが，白書においては過去 40 年間に一貫して顕著に低下している。「は」の零度性を考える上でこれはきわめて興味深いデータである。仮に白書に見られる傾向が現代日本語全般の時系列変化を反映したものであるとするなら，40 年間の「は」の減少と「が」の増加は，日本語が，主題中心の伝統的な陳述形態から，主語中心の西欧言語型の陳述形態に接近しつつあることを示唆している可能性がある。

もっとも，BCCWJ は時系列分析用に開発されたコーパスではなく，年代ごとのデータ量も統制されていない。ゆえに，ここで得られた結果を過大に拡張解釈すべきではないが，たとえば複数の新聞社のデータベースなどで同一の経年パタンが確認されるならば，助詞選択の研究や，主語と主題の関係性をめぐる研究において，今後，時代や年代を組み込んだ分析が求められるようになろう。

3．3．3 テキスト内容種別

助詞選択におけるテキストタイプの影響はすでに検討したが（3.3.1 節），それでは，個々のテキストの具体的な内容種別は「は」と「が」の選択にどのように影響しているのであろうか。2000 年代に刊行された書籍（図書館）に絞って，内容種別ごとの「は」選択率を観察したところ，図 6 の結果を得た。

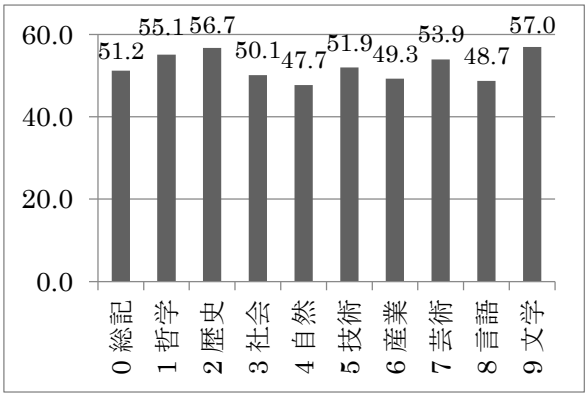


図 6 テキスト内容別の「は」選択率

コーパス全体での「は」選択率は 49.28%，テキストタイプ別では 42.5%～51.6%（韻文の 59.8%を除く）であったが，十進分類法に基づく内容種別では 47.7%～57.0%

となることがわかった。テキストタイプの場合より総じて高めであるものの，上下の幅は10%程度で，テキストタイプの場合とほぼ同等である。だが，ここでも，当該の範囲内においては一定の傾向性が看取できる。具体的には，歴史・文学・哲学など，書き手の解釈と判断に焦点を当てた叙述がなされる場合には「は」が選好され，自然・言語・産業のように。客観的に情報を提示する場合には「が」が選好されている。これは，すでに見たように，判断文と「は」，現象文と「が」の親和性を裏付ける結果と言える。

3．4　RQ4 テキスト外非言語的属性の影響

3．4．1　書き手の生年

過去 40 年間（1970 年代～2000 年代）を対象としたテキスト刊行年代の調査では「は」選択率の低下が示唆されたが（3.3.2 節），過去 90 年間（1890 年代生～1970 年代生）にまたがる書き手の生年の影響はどうであろうか。調査結果は図 7 の通りであった

単回帰による直線のあてはめを行うと，y = －0.7198x + 58.451 という一定の説明力を持った回帰式が得られる（R^2 = .35）。直線の傾きを示す係数は負であり，世代が進むにつれて「は」の選択率が若干ではあるが低下していることがわかる。なお，全体の中で相対的に低い値を示す1890年代と高い値を示す1900年代を除外して，1910年代以降のみで回帰分析をしても，回帰式はy = －0.7297x + 56.893（R^2 = .89）となり，やはり係数は負で，「は」選択率は低下しているように思われる。

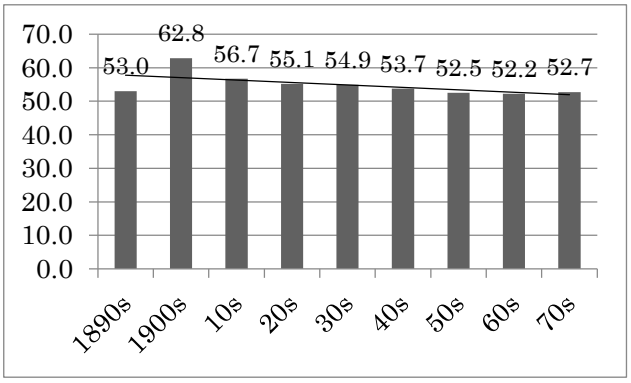


図 7 書き手の生年別の「は」選択率

3．4．2　書き手の性別

性差は言語使用の様々な局面に影響しているとされる。今回のデータで分析を行なった結果，「は」選択率は男性が 46.4%，女性が 53.8%で，男女間の「は」選択率の差は有意であった（G^2 =539.3, *df*=1, *p*<.001）。ただ，結果の再現性を確認するため，データセットを若干拡張し，書籍 25 万件中，単一著者の性別情報を含む 176,082 件全体で調べると，選択率は男性が 53.4%，女性が 53.6%となって選択率の差は有意でなかった（G^2 =0.56, *df*=1, *p*=.454）。このことを踏まえると，性差の影響は仮にあるとしてもごく限定的と思われる。

3．5　RQ5 非言語的属性による決定木モデル構築

以上の分析により，従来の研究でもっぱら分析対象になってきたテキスト内の言語的属性のみならず，テキスト自身およびテキスト外部の書き手に関わる非言語的属性も「は」と「が」の選択に一定の影響を及ぼしている可能性が示唆された。では，非言語的属性だけで「は」と「が」の選択はどの程度説明しうるのであろうか。

決定木分析を行なった結果，図 8 の結果が得られた。モデル中，ジャンル（G）の 0〜9 は書籍内容の十進分類を，書き手生年（B）の 1940 や 1950 はそれぞれ 1940 年代や 1950 年代を，書き手性別（S）の M と F はそれぞれ男性・女性を示す。分岐の結果はたとえば W（15320/6490）のように示されているが，これは当該条件においてモデルが「は」に 15,320 例を分類し，そのうち 6,490 例が誤分類であることを示す。

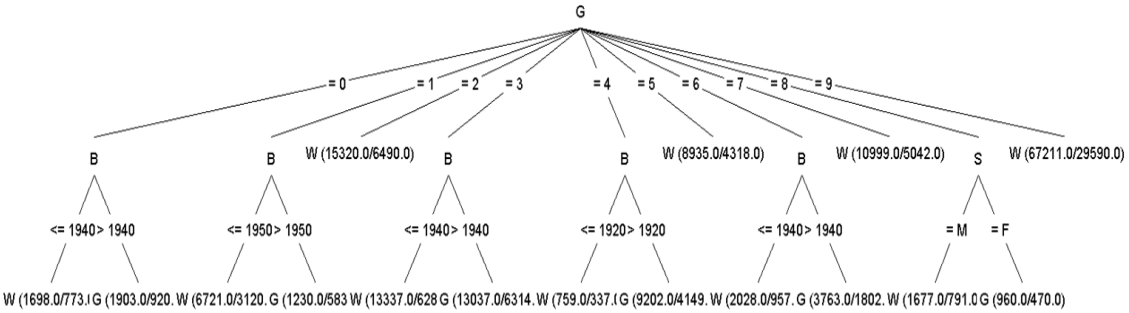


図 8 非言語的属性による「は」と「が」の選択の決定木モデル

モデルはデータの 79.4%を「は」に，20.5%を「が」に分類するもので，正判別率は 54.7%（*Kappa*=.053）である。非言語的属性（内容ジャンル，書き手の生年，書き手の性別）だけで助詞選択の過半が説明されたことは興味深い。また，使用したデータセット中，「は」は 85,223 件，「が」は 73,557 件で，「は」のほうが若干多くなっているが，モデルはより顕著に「は」を優先選択している。

表 1 分類結果（confusion matrix）

	は（分類）	が（分類）
は	69,710	15,513
が	56,404	17,153

非言語的な諸属性のうち，選択に最も影響しているのはジャンルである。書き手の主張に重点が置かれやすい 2（歴史）・5（技術）・7（芸術）・9（文学）類では他の属性の介在なしに「は」が選択される。一方，客観的な事実の紹介が主となる 0（総記）・1（哲学）・3（社会）・4（自然）・6（産業）類では書き手の生年情報が，8（言語）類では書き手の性別が介在的に影響を及ぼして助詞決定がなされる。なお，書き手の生年は 1840 年代から 1980 年代に及んでいるが，分岐の多くは 1940 年代ないし 1950 年代の以前か以後かで起こっており，いわゆる「戦後生まれ」であれば「が」が選択されると言える。この結果はテキスト刊行年代や著者生年の時系列分析（3.3.2 節，3.4.1 節）の結果を精緻化するものである。

4. まとめ

以上，主語・主題標識としての「は」と「が」の選択について，テキストを取り巻く 3 段階の階層要因モデルをふまえた検証を行ってきた。得られた結果についてまとめておく。

まず, RQ1（現代日本語書き言葉全体における標準性）については,「は」選択率は 49.28%で，日本語全般において無条件に「は」が優先されるわけではないことが確認された。

RQ2（テキスト内言語的属性影響）については,「は」の選択は主格要素の (1)情報新旧性に関しては旧情報と，(2)位置に関しては文頭と，(3)意味内容に関しては口語性・具体性・1 人称性と親和性を持つことが示唆された。これらは，話し手の解釈や判断を表明する判断

文で「は」が選ばれやすいという先行研究の主張を支持するものと言える。

RQ3（テキストの非言語的属性影響）については，「は」の選択が，(1)テキストタイプに関しては解釈・注釈・判断を含むテキスト（書籍・法律・広報紙・新聞）と，(2)年代（1970年代〜2000 年代）に関してはより古い年代と，(3)内容種別に関しては解釈・判断が加わるもの（歴史・文学・哲学）と親和性を持つことが示された。判断文と「は」の結合が再確認されたことに加え，テキストの刊行年代が下がるにつれて「は」の優先性が低下する興味深い事実が示された。

RQ4（テキスト外非言語的属性影響）については，「は」の選択は，(1)書き手の生年に関しては古い年代と親和性を持つものの，(2)性別について確定的な結果は得られなかった。

RQ5（非言語的属性によるモデル化）については，6 割弱の正分類を行う決定木モデルが得られ，モデルは「は」を優先するものであった。非言語的属性が助詞決定に関与しうること，いずれかの助詞選択を促す明確な言語的動機を欠く状況では「は」が優先される可能性があること，内容ジャンルが相対的に強く影響しうること，生年が戦前か戦後かで助詞選択が「は」から「が」に変化しうることなどが示唆された。

本研究の結果は，いわゆる「は」の零度性が，頻度の上で，日本語全体に無条件に当てはまるわけではないものの，非言語的要因に基づくモデルにおいては，そうした傾向性が一定程度認められることを示すものであった。ただし，本研究で使用したコーパスは時系列分析や書き手の属性分析を前提として構築されたものではなく，結果の解釈に慎重さが求められるのは言うまでもない。今後，年代や書き手属性ごとのデータ量をより厳密に統制したコーパスの整備がなされ，量的観点に基づく助詞研究が進展することが期待される。

謝辞：草稿に対し，前川喜久雄氏より「が」の用法識別について，森秀明氏より BCCWJ の著者生年の扱いについて貴重な指摘を賜り，一部加筆を施した。記して感謝申し上げる。

文　献

ゲオルギエバ，ベロニカ・トドロバ(2008)『「が」と「は」の使いわけとその理由：母語話者と非母語話者の実態調査を比較して』早稲田大学修士論文.

堀川智也(2010)「日本語の「主題」をめぐる基礎論」『大阪大学世界言語研究センター論集』*4*, 103-117.

石川慎一郎(2013)「語彙難度・語彙多様性・文構成度：母語話者と学習者の区分基準は何か―決定木を用いた学習者コーパス分析―」『統計数理研究所共同研究レポート』*290*, 107-124.

中川正弘(1996)「『は／が』と助詞選択の零度」『広島大学留学生日本語教育』*8* ,11-23.

野田尚史(1996)『新日本語文法選書 1「は」と「が」』東京：くろしお出版.

ヨフコバ四位，エレオノラ (2007)「「は」と「が」に関する一考察 : 外国語としての日本語教育との関連」『横浜国立大学留学生センター教育研究論集』*14*, 159 – 189.

文節係り受け木の根の構造について

高松　亮（埼玉大学経済学部）

Neighboring Structures at Root Node in Dependency Tree of Spoken Japanese

Ryo Takamatsu (Faculty of Economics, Saitama University)

１．はじめに

日本語の文節の係り受け関係を、文節をノードに，係り受け関係をエッジに対応付けたグラフとして表すと木構造になる．木構造の形態的特徴は文の構造を反映したものであり，例えば発話のジャンルが文の構造に与える影響を，木の構造を通じて観察することが可能である．前回の報告において、木の構造は「木の高さ」や「合計文節数」といった大域的特徴や，「各文節に係る文節数の平均」のような局所的特徴を用いて表すことができるため，観察を定量的に行なうことが可能であることを示した（高松(2013)）．その中で、木の大域的特徴の一つである木の高さの頻度について，ジャンルによる分布形の違いがあることが明らかになった．そこで本報告では，特にジャンル間の頻度差が大きかった高さの木について詳細な分析を行なう．その際，木の高さ以外の大域的特徴量として，木に含まれるノード数，すなわち文節数も考慮する．

具体的は，木の高さとノード数をパラメータとした場合の木の頻度分布を調べ，頻度がジャンルによって大きく異なるようなパラメータの値について，木の形態的特徴の傾向を明らかにし，そのような傾向がどのような言語現象に関連したものであるかを考察する．

木の構成要素の中で，根ノード（係り先を持たない文節，以下「根」という）は述語に相当する要素であり，根にかかる文節の数や種類が文の基本的な構造を決める特徴となるため，分析にあたってはこれらの特徴を重要な手がかりとして用いる．

なお，以下では文節の係り受けの情報を木構造として表現したものを係り受け木と呼ぶことがある．また，本報告では，係り元はあるが係り受けがない文節を根にもち，かつ高さ 1 以上(文節数 2 以上)の係り受け木を 1 本の木とする．

２．分析対象

国立国語研究所『日本語話し言葉コーパス(以下 CSJ)』(国立国語研究所(2006))のコア部分に含まれる学会講演と模擬講演を対象に，両者の比較を行なう．

学会講演は実際に行なわれた各種学会での講演を収録したもので，その中の比較的多数を占める理工系の学会では多くの話者が男性の大学院生である．発話のあらたまり度はやや高い．模擬講演は，年齢と性別のバランスをとった一般話者による、日常的話題(各話者に対して，3 種類のテーマから 1 つをあらかじめ指定)についての講演であり，話者の大部分が人材派遣会社からの派遣である．発話スタイルは学会講演よりもくだけたものである．

CSJ では係り受け構造の記述を行なう範囲として，文を認定するかわりに節単位という概念を導入し用いている．ほとんどの場合 1 本の係り受け木は 1 個の節単位に対応する．ただし，係り元があって，係り先のない文節が節単位中に複数存在する場合もあり，その場合にはそれぞれの文節を根に持つ複数の木を考えることにする．

なお，話者数および木の本数は，学会講演が 107 話者／8723 本，模擬講演が 70 話者／10046 本である．

３．木の大域的パラメータ：高さとノード数

木構造の大域的特徴を表わす特徴量の一つに木の高さがある．木の高さは，根から葉に至る経路をたどる際に通過するエッジの数（葉の高さ）のうち最大のものである．言い換えると，係り受け木の高さはある文において，文節が最大で何回の係り受け関係を経て述語に至るかに相当する．これは，国立国語研究所(1955)，小宮(1977)などにおいて「係り受けの次数」と
呼ばれている概念とほぼ同等である．既に高松(2013)において CSJ の学会講演と模擬講演，ならびに国立国語研究所(1955)における木の次数の比較を行なったが，これに小宮(1977)で報告された学校教科書(小中，高校，大学の 3 種類)での値を加えたものを図 1 に示す．なお，"対話"の値は国立国語研究所(1955)をもとに筆者が計算したものである．図より，話し言葉より書き言葉が，対話よりも独話が，あらたまり度の低い発話よりも高い発話が，より木が高い傾向があることが推測される．ただし，木の高さは文認定の基準の影響を受けるものであり，各数値を得た際の文認定の基準が必ずしも一致しているとはいえないため，これらの値を単純に比較することには注意を要する．

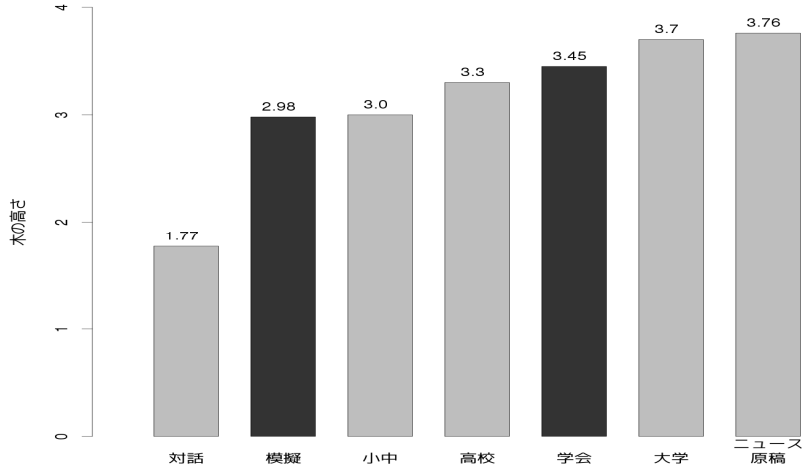


図１　発話ジャンルと木の高さ

大域的特徴として代表的な別の特徴量にノードの合計数(以下，単にノード数)がある．係り受け木の場合，ノード数はその文に含まれる文節数である．したがって，多くの文節を含む，長い文ほどその値は大きくなる．高松(2013)において，学会講演と模擬講演のノード数の相対頻度を比較し，文節数が 2 の木の頻度は模擬講演の頻度が高く，3 から 5 程度の範囲ではその差はわずかになり，それよりも文節数が多い領域においては，逆に学会講演の方がわずかに頻度が高いことから，木の高さ同様，ノード数にもジャンル依存性がみられることが明らかになっている．

以上のように，木の高さとノード数は木の大域的特徴を表わす代表的な量であり，いずれも発話ジャンルに依存して変化する性質を持つが，そもそも両者にはどのような関係があるだろうか．発話のジャンルを固定した場合には，ノード数が増加するにつれて木の高さも高くなることが予想されるが，その詳細は明らかではない．また，発話のジャンルが異なる木の集合同士を比較した場合に，どのような差異が見られるかも不明である．

そこで次節では，ある木の高さおよびノードの合計数を持つ木の相対頻度を求め，そのジャンル依存性を考察する．

４．ある高さとノード数を持つ木の頻度分布

学会講演および模擬講演について，ある高さとノード数を持つ係り受け木の頻度を図 2(学会講演)および図 3(模擬講演)に示す．なお図中では、木の高さに対応する平均ノード数を実線で、平均値に標準偏差を加算および減算した位置を点線でそれぞれ描いている．

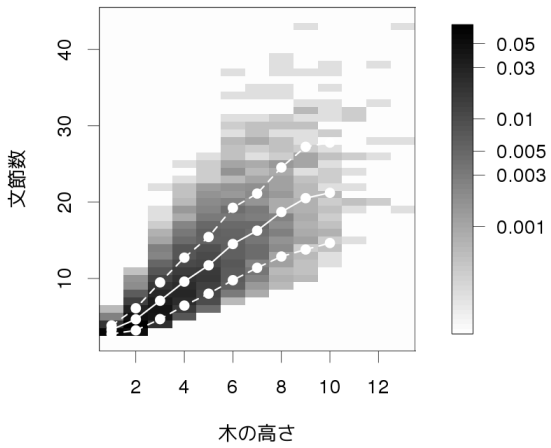


図２　木の相対頻度（学会講演）

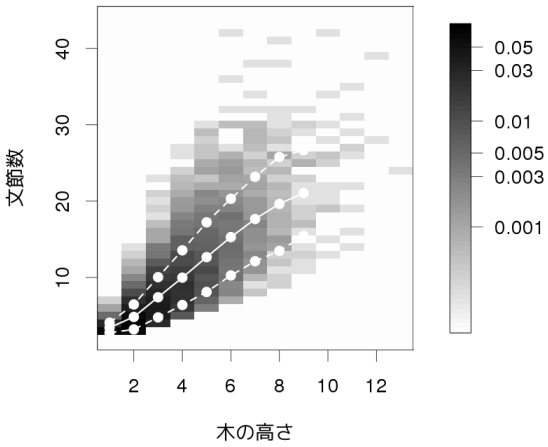


図３　木の相対頻度（模擬講演）

図 2 および図 3 から，木の高さが増加するとノード数も単調に増加すること，学会講演よりも模擬講演の方が増加率が若干大きく，したがって木の高さの割に相対的にノード数が多いこと，木の高さ，ノード数とも学会講演の方が模擬講演よりも広い領域に分布していることなどがわかる．

両者の頻度の差が最も明確に表われている領域を明らかにするために，学会講演の相対頻度から模擬講演の相対頻度を減算したものを図 4 に示す．

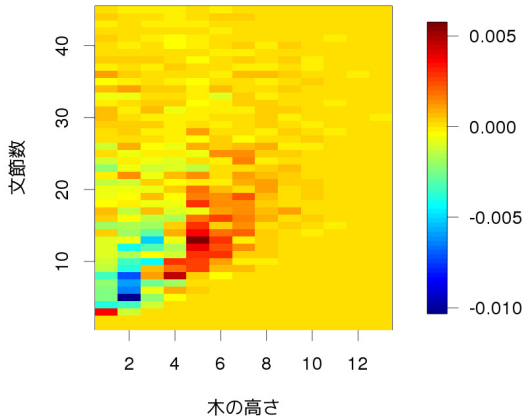


図４　木の相対頻度の差分（［学会講演］－［模擬講演］）

木の高さを h，ノード数を n とすると，図 4 において h=1 および h=2 の領域では，全体的に模擬講演の頻度が高い．特に h=2, n=5 の箇所では大きな頻度の差がある．h=3 および h=4 の領域では，ノード数が少ない部分では学会講演の方が，多い部分では模擬講演の方が，それぞれ頻度が高い．h が 5 以上の領域では全体的に学会講演が模擬講演より頻度が高い．

以上より，学会講演と模擬講演にはつぎのような相対的な差異があることがわかる．すなわち，学会講演は高さが高く，文節数が相対的に若干少ない文が多い．それに対して模擬講演は高さが低く文節数が若干多い文が多い．そこで以下では，図 4 においてそのような差異が最も明確な箇所の１つである h=2, n=5 に注目し，より詳細な調査を行なう．

５．根に係る文節数に基く分析

木の根は述語に相当する要素であり，根に何個のノード，すなわち文節が接続されているかは，文の構造と強い関連がある．そこで，以下では根にかかるノード数の相対頻度を h=2, n=5 において求め分析を試みる．ノード数毎の相対頻度を表 1 に示す。

表 1　根に係る文節数毎の頻度（h=2, n=5）

	学会講演(A)		模擬講演(S)	
根に係る文節数	頻度	相対頻度	頻度	相対頻度
1	17	0.0455	42	0.0766
2	182	0.487	210	0.383
3	175	0.468	296	0.540
合計	374		548	

表 1 より，h=2, n=5 であるような木の頻度は学会講演が 374，模擬講演が 548 である．両講演の全ての係り受け木に対する相対頻度を求めると，それぞれ 0.043(学会)および 0.055(模擬)となり，前節でも既に明らかなように模擬講演の方が頻度が高い．根に係るノード数毎の頻度に注目すると，いずれの講演でもノード数が 2 および 3 が頻度の大半を占めることがわかる．また，学会講演では 2 および 3 の頻度がほぼ同数であり，模擬講演では 2 よりも 3 の頻度がより大きい．さらに，模擬講演では 1 の頻度が学会講演よりも相対的に大きいこともわかる．このように，根に係るノード数の頻度分布が両講演で大きく異なることは，両者の係り受け木の形態の分布が大きく異なることを示唆している．

そこで，以下では根に係る文節数が 1 および 3 の場合について，それぞれどのように発話ジャンルに関係した性質を持っているかを観察する．

5.1　根に係る文節数が 1 の場合

5.1.1　模擬講演

根に係る文節がただ 1 つである場合，その文節の属性はどのようなものかを調査した．全 42 例のうち，言いよどみや述語の言い直しなどによる係り受けの乱れがある 2 例を除いた 40 例について，根に係る文節の最後を構成する短単位（国立国語研究所(2006)）の属性を調べると以下の通りであった．

- 節末（23 個，57.5%）
  - 引用節（格助詞「と(15 個)」，係助詞「は(1 個)」，副助詞「って(1 個)」）
  - タリ節，テ節，トイウ節，条件節タラ，理由節ノデ，並列節ガ（各 1 個ずつ合計 6 個）
- 節末でないもの（17 個，42.5%）
  - 助動詞（5 個），動詞（4 個），形容詞（1 個）の連体形（合計 10 個，25.0%）
  - 助詞（格助詞（3 個），副助詞（2 個））（合計 5 個，12.5%）
  - 助動詞の連用形（2 個，5.0%）

このように、模擬講演においては h=2, n=5 でかつ根にただ 1 つの文節が係るような場合には、その文節は節末、特に引用節であることが多い。引用節が根に係る 17 個の場合について、根の文節を調べるとその大半（15 例）が動詞「思う」によって構成されていた．

5.1.2　学会講演

根に 1 つの文節が係る例(18 例)のうち，言いよどみによる係り受けの乱れがある 1 例を除いた各文節の属性を以下に示す．

- 節末（14 個，82.4%）
  - 引用節（格助詞「と（10 個)」）
  - 「トイウ節」（動詞（連体形）「いう（3 個)」）
  - 「条件節レバ」（接続助詞「ば（1 個)」）
- 節ではないもの（3 個，17.6%）
  - 助詞（格助詞（1 個），副助詞（1 個））（合計 2 個，11.8%）
  - 助動詞の連体形（1 個，5.9%）

節が述語に係るケースが 18 例中 14 例（82.4%）と，模擬講演よりもさらに大きな割合を占めている．その中ではやはり引用節である例が 10 例と多く，これらが係る述語は動詞「思う」(6 例)，「言う」(2 例)，「考え（られ）る」(2 例）で構成される文節であった．

以上から，*h*=2, *n*=5 でかつ根に係る文節数が 1 の文は，述語に節、特に引用節が係る場合が多く，その場合に用いられる述語も限定されることがわかる．また、模擬講演の方が学会講演よりも頻度が高いが，述語に節が係る表現が占める割合は学会講演の方が多い．

5.2　根に係る文節数が 3 の場合

表 1 より，*h*=2，*n*=5 の文のうち模擬講演で最も頻度が高いのは，根に係る文節数が 3 のものであり，模擬講演に 296 本，学会講演に 175 本存在する．なお、*h*=2，*n*=5 の根付き木で根に接続するノード数が 3 であるものは 1 種類のみであり[1]，根に対して 2 つの葉が直接係ることが形態上の特徴である．

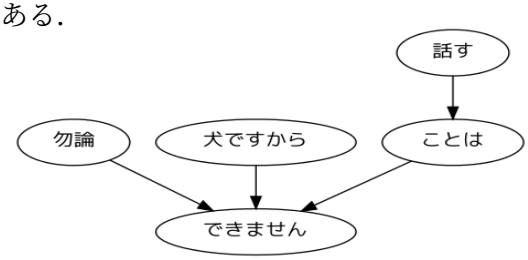


図５　*h*=2, *n*=5　かつ根に係る文節数が 3 の木の例

根に係る文節を構成する短単位のうち，最後のものの品詞の比率を延べ語数で求めたものを表 2，表 3 に示す．表 2 は，根に直接かかる全てのノードに関するもの，表 3 は根にかかる葉のみについてのものである．

表２　根に係る品詞の比率（全ノード）

品詞	高さ2　ノード数5(根に係る全ノード)				全体（根に係る全ノード）			
	模擬講演		学会講演		模擬講演		学会講演	
	個数	割合	個数	割合	個数	割合	個数	割合
助詞	599	0.675	362	0.693	16629	0.701	14602	0.730
格助詞	285	0.321	185	0.354	7607	0.321	8017	0.401
係助詞	132	0.149	90	0.172	2841	0.120	3183	0.159
その他	182	0.205	87	0.167	6181	0.261	3402	0.170
副詞	132	0.149	58	0.111	2059	0.087	1141	0.057
助動詞	58	0.065	29	0.056	2569	0.108	1943	0.097
名詞	52	0.059	39	0.075	992	0.042	865	0.043
形容詞	18	0.020	7	0.013	448	0.019	195	0.010
接尾辞	18	0.020	15	0.029	227	0.010	227	0.011
その他	11	0.012	12	0.023	798	0.034	1018	0.051
合計	888		522		23722		19991	

表３　根に係る品詞の比率（高さ 2，ノード数５の木）

品詞	高さ2　ノード数5(根に係る葉ノード)				全体（根に係る葉ノード）			
	模擬講演		学会講演		模擬講演		学会講演	
	個数	割合	個数	割合	個数	割合	個数	割合
助詞	358	0.605	220	0.632	6188	0.621	5082	0.670
格助詞	174	0.294	108	0.310	3048	0.306	2686	0.354
係助詞	92	0.155	66	0.190	1596	0.160	1611	0.213
その他	92	0.155	46	0.132	1544	0.155	785	0.104
副詞	129	0.218	57	0.164	2016	0.202	1126	0.149
助動詞	41	0.069	16	0.046	519	0.052	422	0.056
名詞	29	0.049	30	0.086	558	0.056	407	0.054
形容詞	16	0.027	6	0.017	292	0.029	99	0.013
接尾辞	12	0.020	11	0.032	133	0.013	116	0.015
その他	7	0.012	8	0.023	255	0.026	329	0.043
合計	592		348		9961		7581	

[1] ここでは同型な根つき順序なし木をもって「同じ種類の木」と考える．

表 3 より，h=2，n=5 における副詞の比率は，任意の（h, n）における副詞の比率にほぼ等しく，一般に模擬講演の方が学会講演よりも副詞の比率が高いことと，この型の係り受け木が副詞の比率に関しては平均的な性質を持っていることがわかる．

一方，表 2 における副詞の比率は学会講演，模擬講演のいずれにおいても h=2，n=5 の方が，任意の（h, n）よりも大きい．すなわち，葉以外のノードも含めた計数を行なうと，この型の木は副詞の割合が高くなる．これは，表 2 と表 3 の副詞の個数を比較すると明らかなように，ほとんどの場合副詞は葉として根に係ること，h=2，n=5 の木のうち本節で扱っている根に係る文節数が 3 のものは，根に係る文節 3 個のうち 2 個までが葉であることが要因となっている可能性がある．

これらの，根に係る葉に含まれる副詞について，どのような語彙が存在するかを調査したものを表 4 に示す．表は頻度の高い順に、第 10 位までの語が示されている。括弧内は頻度である。学会講演は少ない語彙が集中して出現し，模擬講演は様々な語彙が幅広く出現する傾向が見られる．それぞれの発話場面で好んで用いられる語彙の傾向が表れている．

表４　根に係る文節（副詞）

学会講演	まず(8)，ちょっと(5)，例えば(5)，一応(4)，色々(3)， こう(2)，ほぼ(2)，まだ(2)，やはり(2)，全て(2)
模擬講演	やっぱり(8)，あんまり(7)，よく(7)，ちょっと(6)，また(6)， もう(6)，やはり（6)，勿論(6)，色々(5)，大体(4)

６．まとめ

CSJ に含まれる学会講演と模擬講演の 2 つのジャンルを対象に，同じ高さ，文節数をもつ係り受け木の出現頻度がジャンルによって大きく異なる領域を明らかにした．特に大きな差異がある高さと文節数の組合せを対象に，木の根にかかる文節数ごとに，どのような差異があるかを調査した．一般に，2 つの木の高さと合計ノード数が等しいことは，両者が同型であるための必要条件である．今回はその条件に加えて木の根にかかる文節数が 1 または 3 という制約を課した分析を行なった．分析対象とした高さ 2，ノード数 5 の木においてはこの制約は同型であるための十分条件にもなっているため，結果として同型な木について，その出現頻度ならびに種々の言語学的特徴、さらにはジャンル依存性を調査したことになる．今後の課題としては，より広い範囲の高さ文節数を持つ係り受け木を分析することと，また木の同型性に基いた係り受け木の分類・分析手法の検討があげられる．

文　献

高松　亮(2013)「文節係り受け構造のジャンル依存性」第 3 回コーパス日本語学ワークショップ予稿集、pp.17-22
(http://www.ninjal.ac.jp/event/specialists/project-meeting/m-2012/jclws03/ よりダウンロード可能)

国立国語研究所(1955)『談話語の実態』，国立国語研究所研究報告 8
(http://db3.ninjal.ac.jp/publication_db/item.php?id=100170008 よりダウンロード可能)

国立国語研究所(2006)『日本語話し言葉コーパスの構築法』，国立国語研究所研究報告 124
(http://www.ninjal.ac.jp/csj/k-report-f/CSJ_rep.pdf よりダウンロード可能)

小宮千鶴子(1997)「読者層を異にする文章間に見られる文構造の相違(4)-係りの次数による日本史教科書の段階比較-」中央学院大学 人間・自然論叢，第 6 号

金 明哲(1993)「文節の係り受け距離の統計分析」社会情報：札幌学院大学社会情報学部紀要，5:2, pp.1-11 (http://hdl.handle.net/10742/754 よりダウンロード可能)

〈名詞句＋係助詞〉の格

山田昌裕（恵泉女学園大学人文学部）

The Case of "Noun Phrase + *kakari-joshi*"

Masahiro Yamada (Keisen University Faculty of Humanities)

１．はじめに

これまでの係助詞研究は、さまざまな面から行われているが、〈名詞句＋係助詞〉がどのような格成分として使用されているのかという観点からの研究はない。この実態を明らかにすることは、個々の係助詞のみならず、「取り立て」や「焦点」に関しても寄与するところがあると思われる。

そこで本発表では、中納言（国立国語研究所「日本語歴史コーパス」）を用いることにより、〈名詞句＋係助詞〉がどのような格成分となっているのか、その実態を明らかにしたうえで、その実態から見えてくる個々の係助詞の性質、「取り立て」や「焦点」、また副助詞や間投助詞との関連性などについて考察する。

本発表で扱う係助詞は、いわゆる係り結びにかかわる「コソ」「ゾ」「ナム」「ヤ」「カ」を対象とする。〈名詞句＋係助詞〉の用例は、以下の手順によって抽出した。

ａ．短単位検索において、「キー」を「語彙素読み」に指定し、「短単位の条件の追加」で「品詞の小分類が助詞ー係助詞」を選択。
ｂ．「前方共起条件の追加」で「品詞の大分類が名詞」を選択。
ｃ．「コソ」「ゾ」「ナム」「ヤ」「カ」それぞれについて検索。
ｄ．検索結果から『古今和歌集』と散文中の歌を除く。
ｅ．さらに数量詞（例①）や時名詞（例②）は、格成分からは除く。

① 中納言の君といふは、殿の御をぢの右兵衛督忠君と聞えけるが御むすめ、宰相の君は富小路の右の大臣の御孫、それ二人ぞ上にゐて見たまふ（枕草子 p. 412）
② 「三十九なりける年こそさはいましめけれ」（枕草子 p. 290）

得られた〈名詞句＋係助詞〉における名詞句の格が、どのような格成分となっているのかは前後の文脈によって判断する。③④はガ格名詞句の例、⑤⑥はヲ格名詞句の例である。

③ 「のぞきて見れば、顔こそなほいとにくげなりしか」となむ語かたりしとか（大和物語 p. 297）

④ 三郎なりける子なむ、「よき御男ぞいで来む」とあはするに（伊勢物語 p. 165）
⑤ 「三日に当る夜、餅なむまゐる」と人々の聞こゆれば（源氏物語５・p. 274）
⑥ うしろめたげにのみ思しおくめりし亡き御魂にさへ瑕やつけたてまつらん（源氏物語５・p. 194）

２．〈名詞句＋係助詞〉の実態

ここでは、係助詞が下接する名詞句がどのような格成分となっているのか、またヲ格名詞句全体のなかで格助詞「ヲ」による表示と係助詞の承接はどのような様相を呈しているのか、その実態を明らかにする。

２．１　名詞句の格

【表１】は、係助詞「コソ」「ゾ」「ナム」「ヤ」「カ」それぞれが下接する名詞句がどのような格成分なのか数値を示したものである。

【表１】	コソ	ゾ	ナム
ガ格名詞句	349（ 94.5%）	223（ 89.9%）	213（ 86.3%）
ヲ格名詞句	15（ 4.1%）	18（ 7.3%）	23（ 9.3%）
ガ格またはヲ格名詞句	3（ 0.8%）	4（ 1.6%）	4（ 1.6%）
ニ格名詞句	1（ 0.3%）		
ニテ？	1（ 0.3%）	3（ 1.2%）	3（ 1.2%）
不明			4（ 1.6%）
計	369（100.0%）	248（100.0%）	247（100.0%）

	ヤ	カ
ガ格名詞句	200（ 87.0%）	70（ 95.9%）
ヲ格名詞句	28（ 12.2%）	3（ 4.1%）
ガ格またはヲ格名詞句	1（ 0.4%）	
不明	1（ 0.4%）	
計	230（100.0%）	73（100.0%）

「ガ格またはヲ格名詞句」は以下のような例である。

⑦　「ただこの住み処こそ見棄てがたけれ」（源氏物語２・p. 270）
⑧　まづ対の姫君のさうざうしくてものしたまふらむありさまぞ、ふと思しやらるる
（源氏物語２・p. 50）

いずれの係助詞もガ格名詞句に偏って下接していることがわかる。ガ格名詞句は動詞文、形容詞文、名詞文などにおいて使用されるのに対して、ヲ格名詞句は主に他動詞文に限定される。そもそも絶対数に違いがあることが数値に反映していると一旦は捉えておく。

２．２　ヲ格名詞句における係助詞の承接

２．１の数値は係助詞が下接する名詞句の格を見たものであった。ここではヲ格名詞句から見た格助詞や係助詞の承接状況を確認しておく。

【表２】はヲ格名詞句に係助詞が下接する際に、「ヲ」の表示を受けたヲ格名詞句に下接する場合（⑨⑩）とヲ格名詞句に直接下接する場合との割合を示したものである。

⑨　「船に乗りては、楫取の申すことをこそ高き山と頼め」（竹取物語 p. 46）
⑩　何ごとにつけても、故君の御事をぞ尽きせず思ひたまへる（源氏物語５・p. 447）

【表２】	コソ	ゾ	ナム
ヲ格名詞句＋ヲ	74（ 83.1%）	97（ 84.3%）	155（ 87.1%）
ヲ格名詞句	15（ 16.9%）	18（ 15.7%）	23（ 12.9%）
計	89（100.0%）	115（100.0%）	178（100.0%）

	カ	ヤ
ヲ格名詞句＋ヲ	38（ 92.7%）	23（ 45.1%）
ヲ格名詞句	3（ 7.3%）	28（ 54.9%）
計	41（100.0%）	51（100.0%）

ヲ格名詞句に下接する「コソ」「ゾ」「ナム」「カ」はいずれも「ヲ」による格表示を受けている名詞句に下接することが多いことがうかがえるが、「ヤ」だけはその傾向を異にする。

３　実態からの分析

係助詞が下接する名詞句は、そのほとんどがガ格名詞句であった。この点はいずれの係助詞にも共通していた。しかし、ヲ格名詞句に下接する場合には、「ヤ」が他の係助詞と一線を画することが見えてきた。ここでは、２節において確認した実態からどのようなことが読み取れるのか考察する。まずは「コソ」「ゾ」「ナム」「カ」の名詞句に対する下接の仕方から考察し、次に「ヤ」の性質について考える。さらに係助詞と副助詞の関連性についても言及したい。

３．１　「コソ」「ゾ」「ナム」「カ」

もし、これらの係助詞が持つ性質がなんら影響せずに、単純にヲ格名詞句に下接するとすれば、〈ヲ格名詞句＋ヲ＋係助詞〉と〈ヲ格名詞句＋係助詞〉との比率は、ヲ格名詞句における「ヲ」表示と「ヲ」非表示との比率に近い数値になると考えられる。もし、ヲ格名詞句における「ヲ」表示と「ヲ」非表示の割合と、〈ヲ格名詞句＋ヲ＋係助詞〉と〈ヲ格名詞句＋係助詞〉の割合が相違する場合は、そこに何らかの係助詞の性質を読み取ることができるであろう。

そこでまずヲ格名詞句における「ヲ」表示と「ヲ」非表示の割合を確認したい。「ヲ」非表示に関しての検索は、キーが「品詞の大分類が動詞」、前方共起条件の追加が「品詞の大分類が名詞」で行い、名詞がどのような格成分となっているかを前後の文脈により確認した。（例⑪⑫）。

⑪　まさつら、酒、よき物奉れり（土佐日記 p. 21）
⑫　鶏の子抱きて伏したる（枕草子 p. 219）

ヲ格名詞句における「ヲ」表示は 6550 例（59.4%）、「ヲ」非表示は 4475 例（40.6%）となった。ただし、「ヲ」非表示の検索では名詞と動詞が隣り合う例しか拾えず、実際のヲ格名詞句における「ヲ」非表示の割合はもう少し大きな数値となる。ここでは、ヲ格名詞句における「ヲ」表示は 6 割以下、「ヲ」非表示は 4 割以上と大雑把に捉えておきたい。

さて【表２】によれば、「ヲ」表示のヲ格名詞句に下接する係助詞は 83～92%となっていた。ヲ格

名詞句における「ヲ」表示が 6 割、「ヲ」非表示が 4 割であることを考えると、係助詞がヲ格名詞句に下接する場合は、「ヲ」表示名詞句への偏りがあるということになる。裏を返せば、はだかのヲ格名詞句に直接係助詞が下接することは好まれないということになる。

２．１の【表１】において、係助詞が下接する名詞句はガ格名詞句の割合が高く、それはガ格名詞句の絶対数がヲ格名詞句に勝っているからであると捉えておいた。しかし、係助詞がヲ格名詞句との承接において直接の承接が好まれないということを考慮すると、単なるガ格名詞句の絶対数だけではなく、ここには「コソ」「ゾ」「ナム」「カ」に共通する何らかの性格が影響していると思われる。すなわち名詞句に下接する場合、ガ格名詞句や「ヲ」表示のヲ格名詞句とは相性がよく、「ヲ」非表示のヲ格名詞句とは相性が悪いという性質があるのではないか。図に示せば、以下のような状況である。

	格助詞なし	格助詞あり
ガ格名詞句＋係助詞	◎	
ヲ格名詞句＋係助詞	△	◎

ではなぜ係助詞が名詞句に下接する際に、このような一種の相補分布が見られるのであろうか。ここにはいわゆる「焦点」や「強調」という機能が関わっているものと考えたい。

野村（2001）では、係り結びの係りの部分を情報論的に次のように捉えている。本発表でもこの規定にしたがって考察を進めたい（下線は発表者）。

> 連体形結びの係りの意義を考察するためには、提起されてきた用語の中では「焦点」が適当である。「焦点」は、文の選択的指定点であれば結構であるが、それは恐らく規定として強力すぎる。曖昧化してしまうけれど、「選択的指定点ないし文情報の重要点」の如くに弱い規定を与えねば実際的ではなくなってしまう。

文中の成分に焦点を当てたり、あるいは文中の成分を強調したりする場合には、認知的にガ格名詞句へと目が向くのではないだろうか。尾上（2004）では「名詞項と述語との意味関係を大きく変えないで格助詞で言うとすればガが用いられる項」をガ格項と定義し、「多様なガ格項の共通性とは、一言で言えば、事態認識の中核項目ということであろう」と述べる。「事態認識の中核項目」であれば「文情報の重要点」と重なりやすいであろう。「コソ」「ゾ」「ナム」「カ」がガ格名詞句に偏るのは、こうした背景があるのではないだろうか。

またこのような背景があるとすれば、ヲ格名詞句に下接する際に、「ヲ」表示が必要となることも理解される。「コソ」「ゾ」「ナム」「カ」が「ヲ」非表示のヲ格名詞句に下接すると、その名詞句が「文情報の重要点」（焦点）としてガ格名詞句であるという認識が先立ち、情報伝達上不都合となる。先の⑫を例にとれば、「鶏の子抱きて伏したる」ならば「鶏」が行為者、「子」が対象となるが、「鶏の子ぞ抱きて伏したる」となると、「子」に焦点が当たり、「鶏の子」が行為者であるという認識が先立ってしまうのではないか。この情報伝達上の支障を回避するために「鶏の子をぞ抱きて伏したる」のように、「ヲ」表示によってヲ格名詞句であることを明示する必要性があったと考えることができる。

３．２　「ヤ」

「ヤ」はヲ格名詞句への下接の仕方において、「コソ」「ゾ」「ナム」「カ」と異なっていた。ガ格名詞句に下接する割合は同じであるが、「ヲ」非表示のヲ格名詞句（例⑥⑬⑭）に下接する割合が高いということは、「コソ」「ゾ」「ナム」「カ」にはない、「ヤ」の性格がそこに反映されているためであると思われる。他の係助詞とどのような点が異なるのであろうか。

⑬　「あのにくの男や。などかうまどふ。 竈に豆やくべたる」（枕草子 p. 211）
⑭　「大伴の大納言は、龍の頸の玉や取りておはしたる」（竹取物語 p. 49）

「コソ」「ゾ」「ナム」「カ」がガ格名詞句に偏りを見せていたのは、自身が持つ「文情報の重要点」を示すという機能とガ格名詞句が「事態認識の中核項目」であることとの相性のよさゆえであった。そして、ヲ格名詞句に下接する際には、ガ格名詞句と混同されることを回避するため、「ヲ」表示が必要となるのであった。「ヤ」がはだかのヲ格名詞句にも高い割合で下接しているということは、ガ格名詞句との混同が少ない、言い換えればガ格名詞句への志向が弱いということであり、それは「ヤ」が「コソ」「ゾ」「ナム」「カ」に比して、「文情報の重要点」を示すという機能が弱いことを示唆するであろう。

上代から中古にかけて係助詞「ヤ」が係助詞「カ」の領域を侵したことは周知のことであるが、野村（2001）では、「ヤ」の係り結びの成立に関して、「「カ」に近いところのある「ヤ」が「カ」のあるべき場所に侵入した。「ヤ」が間投助詞的であることは、この侵入を容易にしたであろう」と述べ、間投助詞・終助詞の「ヤ」が係り結びの「カ」へと拡がった結果、「ヤ」の係り結びが発生したと見ている。

「ヤ」は、係り結びという点では他の係助詞と同じであるが、「文情報の重要点」を示すという点では、他の係助詞に比して微弱であった。それは係助詞としての「ヤ」の出自に関係することであり、いまだ間投助詞的な性質を多分に持っていると考えることができるであろう。

３．３　副助詞との関連性

まずは「コソ」「ゾ」「ナム」「カ」と副助詞との表現上の連続性について考えてみたい。山田（2012）では、平安期の「ノミ」「サヘ」「ダニ」について同様の調査をしたが、その振る舞いは「コソ」「ゾ」「ナム」「カ」に共通するものがある。数値を示せば以下のとおりである。【表３】は副助詞が下接する名詞句の格成分の数値を示したものである。

【表３】	ノミ	サヘ	ダニ
ガ格名詞句	503(95.4%)	431(91.4%)	436(90.0%)
ヲ格名詞句	24(　4.6%)	34(　7.2%)	37(　7.6%)
ニ格名詞句		3(　0.6%)	11(　2.3%)
ト格名詞句		1(　0.2%)	
不明		3(　0.6%)	3(　0.1%)
計	527(100.0%)	472(100.0%)	487(100.0%)

また、「ノミ」「サヘ」「ダニ」がヲ格名詞句に下接する際には、「ヲ」表示を必要とするという点も同様である。このように副助詞もガ格名詞句に偏るわけであるが、いわゆる取り立てと「事態認識

の中核項目」であるガ格名詞句とは、やはり相性がいいということなのであろう。

「コソ」「ゾ」「ナム」「カ」の機能として、野村（2001）にならい「文情報の重要点」を示すとしたが、名詞句に対する下接の仕方は「ノミ」「サヘ」「ダニ」と同様の振る舞いをしている。ここに表現上の連続性を見ることはできないであろうか。「文情報の重要点」と「取り立て」との上位概念をここでは仮に「焦点」として考えてみたい。同類のものとの比較において、いわば相対的に「焦点」を当てるのが「取り立て」、同類との比較をすることなく、いわば絶対的に「焦点」を当てるのが「文情報の重要点」とすれば、ここに表現上の連続性が認められるのではないだろうか。

次に「ヤ」の位置づけであるが、近藤（2003）では、「卓立の強調は、文の成分のどこにも任意に加えることができ、疑問語やとりたてにも加えることができるのであるが、実は、この特徴は、古典語の係助詞の持っている特徴と等しい」と述べる。係助詞は、確かに「どこにも任意に加えることができ」るが、本発表で見たように、格成分に下接する場合には、「コソ」「ゾ」「ナム」「カ」と「ヤ」には差が見られた。「コソ」「ゾ」「ナム」「カ」がガ格名詞句にだけ偏るのは、「文情報の重要点」を示す機能が強いことによっていた。一方「ヤ」は、そのような機能は弱く、それゆえはだかのヲ格名詞句にも下接しえた。それは「ヤ」がいまだ間投助詞としての性格を帯びていることによっていると考えられた。これを図にまとめると以下のようになるかと思われる。

<table>
<tr><td rowspan="2"></td><td colspan="3">焦　点</td></tr>
<tr><td colspan="2">弱　←　絶対的取り立て　→　強</td><td>相対的取り立て</td></tr>
<tr><td>機能</td><td colspan="2">（ヤ）　　（コソ、ゾ、ナム、カ）</td><td>（ノミ、サヘ、ダニ）</td></tr>
<tr><td>語</td><td>間投助詞</td><td>係助詞（係り結び）</td><td>副助詞</td></tr>
<tr><td>品詞</td><td colspan="2"></td><td></td></tr>
</table>

４．まとめ

本発表では、係助詞が下接する名詞句がどのような格成分となっているのか、その実態を明らかにしたうえで、その背景に何が読み取れるのか考察した。

ａ．「コソ」「ゾ」「ナム」「カ」はガ格名詞句に偏り、「ヤ」にはそれが見られない。

ｂ．「コソ」「ゾ」「ナム」「カ」は「焦点」の当て方が強く、「ヤ」は弱い。

ｃ．「コソ」「ゾ」「ナム」「カ」は表現上、副助詞との連続性が認められる。「ヤ」は間投助詞的な性質が認められる。

５．今後の課題

古典語の無助詞名詞句に関してはまだまだ手つかずの部分が多い。形態素解析が施されている中納言を用いれば、とりあえず名詞句の検索が可能になった。大きな前進であると言えよう。今後は古典語における、すべての無助詞名詞句が文中でどのような役割を担っているのか明らかにする必要があろう。

文　献

尾上圭介（2004）「主語と述語をめぐる文法」『朝倉日本語講座６』、pp. 1-57、朝倉書店

近藤泰弘（2003）「とりたての体系の歴史的変化」『日本語のとりたて―現代語と歴史的変化・地理的変異―』、pp. 243-256、くろしお出版

野村剛史（2001）「ヤによる係り結びの展開」『国語国文』70 巻 1 号、pp. 1-34

山田昌裕（2012）「古典語に見られる〈名詞句＋副助詞〉の格」『青山語文』42 号、pp. 30-40

ポスター発表（2） Aグループ

9月6日（金） 13:00～14:00

日本語学習者のための名詞と修飾語のコロケーション検索プログラムの開発とその使用例

中溝朋子（山口大学留学生センター）
坂井美恵子（大分大学国際教育研究センター）
金森由美（大分大学国際教育研究センター）
大岩幸太郎（大分大学教育福祉科学部）
刈谷丈治（山口大学名誉教授）

Search Program for the Collocation of Nouns and Their Modification for Japanese Learners: Development and Some Application Examples

Tomoko Nakamizo (International Student Center, Yamaguchi University)
Mieko Sakai (Center for International Education and Research, Oita University)
Yumi Kanamori (Center for International Education and Research, Oita University)
Koutarou Ooiwa (Faculty of Education and Welfare Science, Oita University)
Joji Kariya (Professor Emeritus at Yamaguchi University)

１．はじめに

本研究では日本語学習者用に、『現代日本語書き言葉均衡コーパス DVD 版公開データ』(2011)（以下、BCCWJ と略す）を用いて、名詞を中心語とし、修飾語を共起語とするコロケーションの共起頻度とダイス係数を算出するプログラムを開発している。本発表ではその概要と具体的な検索の結果、およびそこから日本語学習者に示せる例について述べる。本プログラムの特徴は、連体修飾表現として共起する「修飾語」について品詞を問わず頻度を一括して計算し、名詞とのダイス係数、および共起頻度順によって提示することができるという点である。これにより、学習者が類義語名詞の違いを修飾語の違いという点から理解し、また作文などの際に、より適切な連体修飾表現を選択するための支援を行いたいと考える。以下、本プログラムについて具体的に述べる。

２．日本語教育におけるコロケーション、および本プログラムの意義

近年、日本語教育におけるコロケーション習得の重要性は多く指摘され、コロケーションに特化された教材やコロケーションが検索できる web サイトも多く開発されている（神田他 2011、山口他 2012 など）。本研究では、コロケーションの中でも修飾語と名詞のコロケーションに注目し、BCCWJ 内で名詞を修飾する語の使用実態を明らかにするためのプログラムを開発している。

例えば日本語学習者にとって「興味」の強さを表現したいときに、「強い興味」なのか「深い興味」なのか、また「イベントが大きい」という意味を表現したいときに「大きいイベント」なのか、「大きなイベント」なのか、「一大イベント」なのかなど、様々な選択肢の中でどれを選ぶかは判断が難しいと考えられる。コロケーションを検索する web サイトとしては、既に NINJAL-LWP for BCCWJ(NLB)や「日本語作文支援システム『なつめ』」など

があるが[1]、本検索プログラムは後述するような特徴を持つ、修飾語と名詞のコロケーションに特化したものである。将来的には、学習者自身が検索できるようインターフェイスを作成し、現在開発中のコロケーション習得教材とリンクさせることで、コロケーションの習得にも結び付けられるよう計画している。以下、３．で本プログラムの概要について述べる。

３．プログラム概要

３.１． データ

データには、BCCWJ(2011)の SUW（短単位）可変長データを使用している。

３.２． 分析方法

３．２．１　分析の手順

本プログラムの最も重要な機能は、上記データの文の構成単語列に対する語彙素 and/or 品詞の正規表現による検索機能である。品詞は前方部分だけ指定すれば良い。分析は、まず文中から対象となる部分を品詞の正規表現で検索し、その後集計等の処理を行う。具体的な検索、および分析の手順は、以下の通りである。

⑴検索したい修飾語に応じて任意に「修飾語パターン」「中心語パターン」「共起パターン」を定義し、それに基づき、全サブコーパス（もしくは任意のサブコーパス）について検索を行う。
⑵修飾語、中心語、共起の頻度を計算し、共起頻度表に修飾語頻度、中心語頻度を取り込み頻度表にする。
⑶サブコーパス毎の頻度表を結合し、全頻度表とする。
⑷サブコーパス語数を使って頻度から１億語あたりの相対頻度とダイス係数を計算する。

以下、３．２．２で本研究において定義したそれぞれのパターンについて説明する。

３．２．２．　本研究における「中心語パターン」、「修飾語パターン」、「共起パターン」

本研究では「中心語パターン」と「修飾語パターン」は、表１のように定義した[2]。なお、表中の品詞分類は、BCCWJ で用いられている品詞分類の名称を用いている。

表１　本プログラムで用いた修飾語と中心語のパターン

修飾語パターン		中心語パターン	共起パターンの例
A	<接頭辞>	**<名詞>＋<接尾辞>*** (<動詞以外 または文末>)	<一大>**<決心>** <新>**<委員><長>**

[1]NINJAL-LWP for BCCWJ(NLB)　http://nlb.ninjal.ac.jp/
「日本語作文支援システム『なつめ』」　http://hinoki.ryu.titech.ac.jp/natsume/
[2]本研究では、従来コロケーションには含まれない「接頭辞」と「名詞」の組み合わせも検索の対象としている。その理由は、接頭辞の中には形容詞や連体詞などの修飾語と同様の意味を表すものがあるためで（例：「一大」と「大きい」など）、日本語学習者が修飾語を選ぶ際には、これらの接頭辞も修飾語の中のひとつの選択肢となり得ると考えたためである。

B	<連体詞>	**<接頭辞>*<名詞>+ <接尾辞>*** (<動詞以外 または文末>)	<大きな>**<変化>**
C	<形容詞><形容詞-非自立可能>?		<青い>**<空>** <少なく><ない>**<被害>**
D1	<接頭辞>*<形状詞>		<最><重要>**<課題>**
D2	<接頭辞>* <形状詞><助動詞>?<形容詞-非自立可能>?		<不><可能><な>**<計画>** <堂々><たる>**<体格>** <格好><良い>**<御><姿>**
E1	<接頭辞>*<名詞>＋<接尾辞>*<助詞>+		**<外国><人><の><新><委員><長>** <御><隣り><さん><の>**<騒音>**
E2	<接頭辞>*<名詞>＋<形容詞>		<思慮><深い>**<行動>**

・中心語はゴシック、修飾語は明朝で記述
＊は、「ゼロもしくは１以上」を表す
？は、「ゼロもしくは１」を表す
＋は、「１以上」を表す

「修飾語パターン」は、中心語である名詞を修飾する表現について、品詞を問わずできるだけ網羅的に検索できるよう、これまで A から E2 のパターンを考えた。「中心語パターン」は、名詞が単独で出現する場合も、名詞に接頭辞が前接する場合にも同じ中心語として扱われ、中心語頻度および修飾語との共起頻度を計算できるよう考慮した[3]。

これらの「修飾語パターン」と「中心語パターン」をそれぞれ共起語と中心語とみなし、A は<接頭辞>が<名詞>と共起する場合を、A 以外は、接頭辞の有無を問わず<名詞>と共起する場合を「共起パターン」として検索した。

検索する正規表現のコマンドでの表現は、表２のとおりである。

表２　各「共起パターン」の正規表現

検索対象	文法	
接頭辞	A	[:"接頭辞"] [#] [(:"名詞")+] [(:"接尾辞")*](""-:"動詞"|$)
接頭辞以外（形容詞、形状詞等）	B	([:"連体詞"] [#]
	C	| [:"形容詞" (:"形容詞-非自立可能")?] [#]
	D1	| [(:"接頭辞")* :"形状詞"] [#]
	D2	| [(:"接頭辞")* :"形状詞" :"助動詞" (:"形容詞-非自立可能")?] [#]
	E1	| [(:"接頭辞")* (:"名詞")+] [(:"接尾辞")* (:"助詞")+]
	E2	| [(:"接頭辞")* (:"名詞")+] [(:"形容詞")*]
		)[(:"接頭辞")* (:"名詞")+] [(:"接尾辞")*] (""-:"動詞"|$)

これらの正規表現を用いた findPattern コマンドでの検索結果は各語の語彙素、品詞、活用形と、該当部分が表示されている文の tsv 形式ファイルである。検索コマンド中で[]で囲まれた部分に一致する語が出力される。[#]は、出力のカラム数を合わせるために空白カラムを挿入する指定である。検索結果２例を表３に示す。

[3]接頭辞の場合と異なり、名詞に接尾辞や名詞が後接する場合は、これらが後接しない場合と別の中心語として検索される。例えば本稿の４．で行う「興味」「関心」の検索では、「御興味」「御関心」の「興味」「関心」は、それぞれ接頭辞のない「興味」「関心」と同じ中心語として計算されるが、「関心事」「興味本位」などは別の中心語として扱われる。

表３　findPattern を用いて検索した結果例　（上段は D1、下段は C の検索例）

検索結果例　（「^」は TAB）
|重要|だ|^ |形状詞-一般|助動詞|^ ||連体形-一般|^ |^ |^ | ^|^役割|^ |名詞-普通名詞-一般|^ ||^ |^ |^ | ^問題は、“冤罪を創作する”ためにマスコミ・メディアが<<<重要な>>>><<<>>>><<<役割>>>><<<>>>を意図してはたす、という新しい状況が派生したことである。
|良い|無い| ^|形容詞-非自立可能|形容詞-非自立可能| ^|連用形-一般|連体形-一般|^ |^ |^ |^ |日|^ |名詞-普通名詞-副詞可能| ^|| ^|^ |^ |^ 調子が<<<良くない>>>><<<>>>><<<日>>>><<<>>>は頭がくらくらし、足がこの通り重くなるんです」と断り、これまでのことを語り始めた。

以下、表４に、本分析に用いた主なコマンドを示す。

表４　本プログラムの主なコマンド一覧

コマンド名	定義
count	指定したカラムで集計、再集計を行う。
select	カラムの文字列としての一致、前方一致、後方一致、部分一致、数値としての大小を判定し、一致、不一致の行を出力する。
columns	カラムの取捨、複製、並べ替えを行う。
sort	文字順、数値の昇順降順で行を並べ替える。
compute	ダイス係数の計算のような簡単な数値計算を行う。
concat	ファイルを連結する。
import	二つの tsv ファイルでカラムが対応したときに指定カラムを取り込む。
findWords	語彙素 and/or 品詞の列で検索する。予備的調査に用いる。
findPattern	語彙素 and/or 品詞の正規表現で検索する。調査に用いる。

以下、4.で本プログラムから得られる結果を用いて日本語学習者に示すことができることの例として、類義語「興味」と「関心」の修飾語を分析した結果について述べる。

４．本プログラムを用いた検索例－「興味」と「関心」－

４．１　辞書による意味の記述

「興味」「関心」については、国語辞典や類語辞典では、表５のように意味が記述されている。

表５　国語辞典・類語辞典による「興味」「関心」の意味の記述（波線筆者）

辞書名	興味	関心
『大辞泉』	①　その物事が感じさせるおもむき。おもしろみ。興。②ある対象に対する特別の関心。「―がわく」③心理学で、ある対象を価値あるものとして、主観的に選択しようとする心的傾向。教育学では、学習の動機付けの一つ。	①ある物事に特に心を引かれ、注意を向けること。「政治に―がある」「幼児教育に―が高まる」「周囲の―の的」
『大辞林』	①物事に心がひかれおもしろいと感じること。おもしろみ。おもむき。「―を覚える」「―をもつ」②〔教・心〕ある対象に対して特別の関心・注意を向ける心的傾向。	①物事に興味をもったり、注意を払ったりすること。気にかけること。「政治に―をもつ」
『類語例解辞典』	(1)「興味」はおもしろいと感じる気持ちや、知りたいと食指を動かされるような気持ちを言う。(2)「興味」が対象のある１点に感情的に向けられることがあるのに対し、「関心」は対象全体に理性的に向けられることが多い。	

このように「興味」と「関心」は、語を定義するのに互いの語が用いられていたり、説明が抽象的であったりなど、辞書の意味の記述だけでは、日本語学習者にとっては必ずしもわかりやすいとは言えない。そこで本プログラムを用いて、より具体的にどのように使用されているかを学習者に示し、両者の違いをさらに明確にさせたいと考える。

４．２　「興味」「関心」と共起する修飾語

本プログラムの結果から、「興味」と「関心」を中心語とする部分を分析する。BCCWJには、「興味」は 7,864 例、「関心」は 6,243 例あった。以下、表６に、検索の対象となる被修飾語の「興味」「関心」が各サブコーパスで実際に出現した頻度（粗頻度）、および各サブコーパスの１億語当たりの頻度（相対頻度）を示す。

表６　「興味」「関心」のサブコーパス別頻度[4]

	頻度	書籍 LB	ベス OB	知恵 OC	法律 OL	国会 OM	広報 OP	教科 OT	白書 OW	ブロ OY	書籍 PB	雑誌 PM	新聞 PN
興味	粗	2597	267	828	0	51	218	59	31	1149	2237	358	69
	相対	7267	6026	6889	0	911	4718	5267	549	8848	6661	6714	4295
関心	粗	2220	179	104	4	301	290	63	413	302	2094	165	108
	相対	6212	4040	865	331	5378	6276	5624	7322	2325	6235	3094	6723

以下、表７に「興味」「関心」とそれぞれの修飾語の共起についてダイス係数上位 20 語を示す。なお便宜上、ダイス係数は 100 万倍して表示している。

表７　「興味」「関心」と修飾語のダイス係数上位 20 語

興味				関心			
順	修飾語	ダイス係数	共起頻度	順	修飾語	ダイス係数	共起頻度
1	強い	3,533	29	1	無	59,736	622
2	デートレードに	3,315	13	2	強い	20,347	169
3	株に	3,306	13	3	重大な	12,047	45
4	子供達の	2,570	13	4	国民の	11,218	60
5	車関係に	2,297	9	5	深い	10,348	61
6	販売に	2,293	9	6	人々の	9,187	44
7	車に	2,186	10	7	大きな	5,768	91
8	物に	2,123	21	8	政治に	3,693	11
9	歴史に	1,895	8	9	高い	3,618	37
10	子供の	1,860	17	10	問題に	3,373	21
11	生徒の	1,817	8	11	問題への	3,194	9
12	話に	1,807	9	12	消費者の	3,032	10
13	科学に	1,761	7	13	性的	2,951	9
14	深い	1,731	12	14	研究者の	2,685	8

[4]表中では省略した BCCWJ のサブコーパスの名称は、以下の通りである。出版サブコーパス（書籍（PB)、雑誌（PM)、新聞（PN))、特定目的サブコーパス（白書（OW)、教科書（OT)、広報紙（OP)、ベストセラー（OB)、Yahoo!知恵袋（OC)、Yahoo!ブログ（OY)、法律（OL)、国会会議録（OM))

15	女性に	1,730	8	15	世間の	2,552	8
16	方に	1,704	10	16	ボランティア活動に	2,506	7
17	食に	1,519	6	17	環境問題への	2,489	7
18	自分に	1,509	10	18	市民の	2472	9
19	純粋な	1,414	6	19	最大の	2,450	10
20	事に	1,385	63	20	主要な	2,397	8

(1)「興味」「関心」を持つ主体を表す修飾語

まず「名詞＋の」で「興味」や「関心」を持つ主体を表す修飾語を比較する。「興味」は、特に「子供（達）の」「生徒の」など年少者に多く使用されている一方、「関心」は「国民の」「人々の」「消費者の」「研究者の」「市民の」など社会的立場や役割を示す集合名詞が多く共起していた。以下、これらの語の一部のサブコーパス別の出現状況を表８に示す。

表８　サブコーパス別「興味」「関心」を持つ主体を表す修飾語（相対共起頻度）

サブコーパス		子供達の(興味)	国民の(関心)	市民の(関心)	人々の(関心)
書籍	LB	2	11	5	89
ベストセラー	OB		22		
yahoo!知恵袋	OC		8		
国会議事録	OM	17	160		
広報紙	OP	21	21	43	
教科書	OT		89		
白書	OW		531		
yahoo!ブログ	OY	7	7		
書籍	PB	23	17	2	32
雑誌	PM	18	18		18
新聞	PN		311	249	
全サブコーパス相対共起語頻度		88	1195	299	139

「興味」と共起する「子供達の」は「書籍(PB)」などで多く見られ、この傾向は「子供の」「生徒の」でも同様であった。また、「関心」と共起する「国民の」は「白書（OW)」で、「市民の」は「新聞（PN)」や「広報紙（OP)」で、「人々の」は「書籍（LB)」で特徴的に出現していた。

(2)「興味」「関心」の強さ・重要性を表す修飾語

次に「興味」「関心」の強さ・重要性を表す修飾語については、「興味」には「強い」「深い」が共起していたのに対し、「関心」には「強い」「重大な」「深い」「大きな」「高い」などが共起しており、語の種類も頻度も「関心」のほうが多いことが観察できた。

以下、強さや重要性を表す修飾語のサブコーパス別の出現状況を表９に示す。「強い」は特に「関心」において頻度が高く、多くのサブコーパスで出現している。一方で「大きな」「重大な」は「国会議事録（OM)」で、「高い」は「白書（OW)」で特徴的に出現しており、「深い」は「ベストセラー（OB)」「書籍（LB)」などで特徴的に出現している。

表９　サブコーパス別「興味」「関心」の強さ・重要性を表す修飾語（相対共起頻度）

サブコーパス		強い(興味)	強い(関心)	深い(興味)	深い(関心)
書籍	LB	36	204	19	75
ベストセラー	OB	90	90	22	135
yahoo!知恵袋	OC		8		
国会議事録	OM		178		53
広報紙	OP	21	21		
教科書	OT				
白書	OW		195		35
yahoo!ブログ	OY	15	15	7	
書籍	PB	26	181	8	59
雑誌	PM		75		37
新聞	PN		124		62
全サブコーパス相対共起語頻度		188	1091	56	456

サブコーパス		大きな(関心)	重大な(関心)	高い(関心)	最大の(関心)
書籍	LB	78	13	22	16
ベストセラー	OB		22		
yahoo!知恵袋	OC				
国会議事録	OM	321	553	53	35
広報紙	OP				
教科書	OT				
白書	OW	141	88	283	
yahoo!ブログ	OY	23		15	
書籍	PB	95	8	11	5
雑誌	PM			56	
新聞	PN	124		62	
全サブコーパス相対共起語頻度		782	684	502	56

(3)「興味」「関心」の対象を表す修飾語[5]

「興味」「関心」の対象を表す語については、「興味」は「科学」「歴史」といった分野を表す語のほかに、「デートレード」「株」といった具体的な事象や「車」「話」「食」など、日常的な語が多く共起しているのに対し、「関心」は「政治」「問題」や、さらに具体的な「環境問題」「ボランティア活動」など、社会的な問題や取り組むべき課題などを表す語と多く共起している。

上記のことから、「興味」と「関心」は、ともに「物事に心ひかれる」「おもしろさを感じる」という意味を表す言葉であることとともに、BCCWJ の検索結果を基に、以下のような特徴を日本語学習者に示すことができると考える。

① 「興味」はより日常的、個人的、「関心」はより社会的、集団的な文脈で用いられることが多いこと
② したがって「関心」では、「関心」を持つ主体を表す言葉には、社会的立場や集団などの意味を表す「国民の」「人々の」などが多く用いられる傾向があること
③ 「興味」「関心」の強さについては、ともに「強い」が最も多く、かつ比較的広い文脈で使用されること
④ さらに「関心」の強さの表現としては、「高い」「大きな」「重大な」「深い」などが共起

[5] 対象を表す修飾語（ニ格）は、「興味」「関心」を修飾する連体修飾語ではないが、本稿では含めて分析を行う。

していたが、これらは特定のサブコーパスで多く使用されており、「強い」に比べると、使用する文脈に注意が必要なこと

⑤ 「興味」「関心」の内容は、「興味」のほうがより具体的、日常的な内容を、「関心」はより社会的な問題や取り組むべき課題に使用される

などが挙げられる。ただし、上記の結果は、書き言葉のコーパスの特徴を反映したものと考えられ、「興味」「関心」の特徴として示すには、話し言葉のコーパスによる調査も必要と考えられる。

５．おわりに

以上、日本語学習者のための名詞と修飾語の検索プログラムの開発の概要と、その検索例と分析例について具体的に記述した。今後は、プログラムの検証、および改善を継続するとともに、日本語学習者が自身で検索できるためのデータの整理とインターフェイスの開発が課題である。

謝　辞

本研究は JSPS 科研費（基盤研究(c) 23520638、および 25370591）の助成を受けています。

文　献

石川慎一郎（2008）『英語コーパスと言語教育』大修館書店
神田靖子、佐尾ちとせ、佐藤由紀子、山田あき子（2011）『連語を使おう』古今書院
坂井美恵子、中溝朋子、金森由美（2011）「類義語『決心』『決意』『決断』の使い分け―コーパスから見たコロケーションの特徴―」『跨文化交際中的日語教育研究１異文化コミュニケーションのための日本語教育』pp.835-837
中溝朋子、坂井美恵子、金森由美（2012）「共起表現から見た『決定』『決心』『決意』『決断』『判断』の異同について」『日本語教育国際研究大会予稿集（第１分冊）』 p.70
山口久代､竹沢美樹､崔美貴（2012）『コロケーションが身につく日本語表現練習帳』研究社
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

資　料

『現代日本語書き言葉均衡コーパス（BCCWJ）DVD 版公開データ』(2011) 国立国語研究所
『大辞泉』　goo 辞書（『デジタル大辞泉』小学館）
『大辞林』　BIGLOBE サーチ（『大辞林　第二版』三省堂）
『使い方のわかる類語例解辞典』（小学館）

関連 URL

現代日本語書き言葉均衡コーパス　http://www.ninjal.ac.jp/corpus_center/bccwj/
日本語学習者のためのコロケーション学習サイト「コロケーション彗星」
http://nagareboshi.susi.oita-u.ac.jp/index.html

外来語語末長音の表記のゆれについて

小椋秀樹（立命館大学文学部）†

Orthographic Variation of Word-Final Long Vowels in Japanese Loanwords

Hideki Ogura (College of Letters, Ritsumeikan University)

１．はじめに

本稿は、小椋(2013)に続き、外来語表記のゆれの実態について、大規模コーパスを活用した実態調査を行うものである。

小椋(2013)では、『現代日本語書き言葉均衡コーパス』(以下、BCCWJ とする。)のコアデータ(1)を資料として、(1)外来語の表記がどの程度ゆれているのかレジスターごとに調査し、レジスターによる差異を明らかにした上で、(2)各レジスターにおける外来語表記のゆれの類型について明らかにした。

その結果、(1)外来語表記のゆれにはレジスターによる差異があること、(2)外来語表記のゆれの類型についても、やはりレジスターによる差異があることを明らかにした。また、外来語表記のゆれについては、具体的にどのようなゆれがあるか調査し、長音に関する表記のゆれ(語末長音を長音符号で書くか省くか、語中長音を長音符号で書くか省くか)が全てのレジスターに見られることを明らかにした。

そこで本稿では、小椋(2013)で全てのレジスターに見られ、また外来語表記のゆれの問題でも取り上げられることの多い、語末長音の表記のゆれを取り上げる。そして BCCWJ を資料として、主として計量的な観点から、表記のゆれの実態を明らかにする。

２．先行研究

外来語表記のゆれについて、大規模な実態調査を行ったものとしては、宮島・高木(1984)が挙げられる。

宮島・高木(1984)は、1956 年発行の雑誌 90 種を対象とした外来語表記のゆれに関する調査報告である。「外来語の表記について」(1952 年、国語審議会部会報告)に示された外来語表記の原則(19 項目)のうち、撥音、イ列・エ列の次の「ア」、外来語音「ティ」「ディ」、語末の-er 等の表記など、7 項目について、どのような表記が見られるのかを語ごとに示している。

このうち、本稿と関わるのは、語末の-er 等の表記である。これは、「外来語の表記について」に、

> 原語(特に英語)のつづりの終りの -er、-or、-ar、などをかたかながきにする場合には、長音符号「ー」を用いる。

† h-ogura@fc.ritsumei.ac.jp

(1) BCCWJ の設計等については、山崎(2007)、前川(2008)を参照。コアデータの設計・構成等については、小椋・小木曽・小磯ほか(2009)を参照。

ライター (lighter)　エレベーター (elevator)

ただし、これを省く慣用のあるものは必ずしもつけなくてもよい。

ハンマ (hammer)　スリッパ(slipper)　ドア(door)

とある規定について、雑誌 90 種での実態を調査したものである。結果は、長音符号で表記した語は異なり語数で 277 語、長音符号を省略した語は同 12 語と、長音符号で表記する語が圧倒的であることを明らかにしている。

宮島・高木(1984)の調査対象年から 57 年経過した現在、語末長音の表記について変化が生じていることは、小椋(2013)の結果から指摘することができる。

また、NHK の放送における外来語表記の基準改定に関連して、放送用語委員会でも外来語の語末長音の表記が議題となっている(山下 2012:77-78)。英語等の語末の-er、-or、-ar、-y は、NHK 及び新聞各社では原則として長音で書き表すことにしている。しかし、語末の長音を省略した表記を目にすることも多いため、長音を表記するという原則を再確認するということで、放送用語委員会で議題として取り上げられている。

これに対して、表音一致の原則を守るという立場から、原則を支持する意見が委員から出されている。なお、「～ティー」「～ディー」については、「ィ」が長音を含むと思っている人が多いのではないか、実際の発音がゆれているのではないか、専門語的な感覚で書きたいということから長音府符号が省略されるのではないかといった意見もある。

以上のような外来語語末長音の表記の現状を踏まえ、本稿では、BCCWJ を資料として、そこに収録された新聞・雑誌・書籍・Web の四つのレジスターを対象に、現代における外来語語末長音の表記のゆれの実態を計量的な手法によって明らかにしていく。具体的には、外来語の語末長音の表記がどの程度ゆれているのかレジスターごとに調査し、レジスターによる差異を明らかにしていくこととする。

３．調査資料・調査対象

３．１　調査資料

本稿では、複数のレジスターのテキストを収録した BCCWJ を資料とした。BCCWJ は、言語単位として長単位と短単位の 2 種類を採用している(2)。今回の調査にはそのうち短単位を用いた。調査したレジスターは、次のとおりである。

出版サブコーパス　：2001 年-2005 年発行の新聞、雑誌、書籍

特定目的サブコーパス　：2004 年 10 月-2005 年 10 月投稿の Yahoo!知恵袋

2008 年 4 月-2009 年 4 月投稿の Yahoo!ブログ

表 1：各レジスターの延べ語数

	延べ語数
書籍	28552283
雑誌	4444492
新聞	1370233
Web	20451020

(2) 長単位、短単位の設計方針、認定規程等については、小椋・小磯・冨士池ほか(2011)を参照。

各レジスターの延べ語数は、表 1 のとおりである(短単位の語数。記号、補助記号、空白は除く。)。本稿では、Yahoo!知恵袋と Yahoo!ブログとをまとめて Web として扱うため、表 1 でもそのように示している。

３．２　調査対象

外来語の語末長音の表記でしばしば問題とされるのが、英語の語末が-er、-or、-ar の語である。外来語の語末長音の表記については、『外来語の表記』(1991 年、内閣告示第 2 号、内閣訓令第 1 号)で次のように規定されている。

> 英語の語末-er、-or、-ar などに当たるものは、原則としてア列の長音とし長音符号「ー」を用いて書き表す。ただし、慣用に応じて「ー」を省くことができる。
>
> 〔例〕　エレベーター　　ギター　　コンピューター　　マフラー
>
> 　　エレベータ　　コンピュータ　　スリッパ

長音符号で表記することを原則として示しつつも、長音符号を省く表記を寛容として認めている。このような、緩やかな性格の基準によって、語末長音の表記のゆれが生じており、山下(2012)に見られるように、しばしば問題となっているのである。また山下(2012)では、英語の語末が-er、-or、-ar の語以外に、語末が-y の語(例：「パーティーーパーティ」「コミュニティーーコミュニティ」)も取り上げられている。

小椋(2013)では、語末長音を長音符号で書くか省くかに関するゆれが、全レジスターに見られることを指摘した。このゆれている語の多くは、英語の語末が-er、-or、-ar、-ty、-dy、-gy、-ry の語であった。

以上のことから、本稿では、英語の語末が-er、-or、-ar、-ty、-dy、-gy、-ry の語を調査対象として取り上げることとした。また、本稿で取り上げる表記のゆれは、「コンピューターーコンピュータ」「セキュリティーーセキュリティ」のような、語末長音を長音符号で書くか省くかというゆれとした。つまり英語の語末が-er 等の外来語について、語末に長音符号があるか否かということのみを調査することとしたのである。そのため、「コンピューターーコンピュータア」のような、語末長音を長音符号で書くか仮名で書くかというゆれは、取り上げなかった。

3.1 節に述べたように、今回の調査では BCCWJ の短単位データを用いる。そのため、本稿でいう語末長音とは、短単位の末尾が長音ということである。また調査の便宜から、短単位データの「語彙素細分類」列に記載された原語の語末が、-er、-or、-ar、-ty、-dy、-gy、-ry の語を対象とした。

４．調査結果

４．１　レジスター別

各レジスターにおいて、どの程度、外来語語末長音の表記のゆれ(長音符号で表記するか省くか)が見られるのか見ていくこととする。

表 2 に、語末長音の表記にゆれの見られる語の異なり語数(「ゆれ」の欄)と、今回の調査対象である英語の語末が-er、-or、-ar、-ty、-dy、-gy、-ry の語の異なり語数(「異なり」の欄)に占める割合を示した。また、表 2 では、語末長音の表記にゆれの見られない語の異なり語数(「ゆれなし」の「計」の欄)を示すとともに、長音符号で書く表記のみが出現する語の異なり語数(「ゆれなし」の「符号」の欄)、長音符号を省く表記のみが出現する

語の異なり語数(「ゆれなし」の「省略」の欄)も示した。

表 2：語末長音の表記にゆれのある語の割合

	異なり	ゆれ	割合	ゆれなし		
				計	符号	省略
Web	1106	254	23.0%	852	787	65
書籍	1233	259	21.0%	974	858	116
雑誌	891	136	15.3%	755	698	57
新聞	394	14	3.6%	380	364	16

表 2 を見ると、語末長音の表記のゆれの割合は、Web が 23.0%で最も高く、書籍が 21.0%でそれに次ぐ。雑誌は 15.3%で、Web・書籍に近い傾向を示している。一方、新聞は、ゆれの割合が 3.6%と最も低い。

語末長音の表記にゆれの見られない語については、各レジスターとも長音符号で書く表記の方が、長音符号を省く表記よりも圧倒的に多いことが分かる。表記にゆれが見られない場合、『外来語の表記』の原則や新聞各社の表記の基準に示されている表記が、そのほとんどを占めているということになる。

ところで、表 2 は度数 1 の語を含んで集計したものである。当然のことではあるが、度数 1 の語には表記のゆれは発生しない。そこで、ゆれが生じる可能性のある度数 2 以上の語に限って集計し直した。結果は表 3 のとおりである。これは、ゆれが生じる可能性のある語が実際にどの程度ゆれているかを示したものである。

表 3：語末長音の表記にゆれのある語の割合(度数 2 以上)

	異なり	ゆれ	割合	ゆれなし		
				計	符号	省略
Web	934	254	27.2%	647	647	33
書籍	1042	259	24.9%	783	695	88
雑誌	720	136	18.9%	584	546	38
新聞	255	14	5.5%	241	238	3

表 2 と比べてレジスター別順位に変動はない。各レジスターとも、ゆれの割合は約 2%～4%程度高くなっている。

以上のように、表 2、表 3 から外来語語末長音の表記のゆれには、レジスターによる差異のあることが分かった。表記にゆれのある語の割合は、Web・書籍では 2 割台、雑誌では 1 割台と高くなっている。先にも述べたように、外来語語末長音の表記のゆれは、外来語表記のゆれの問題の中でも、よく指摘される現象である。今回の調査結果からも、そのことが確認されたといえよう。

新聞は、表記のゆれの割合が一桁台となっていることから、表記がかなりの統一されていることが分かる。また、ゆれのない語において、長音符号で書く表記の占める割合が、度数 2 以上の場合、98.8%と非常に高くなっており、『外来語の表記』や新聞各社の表記の基準に忠実に従っていることが確認された。

なお新聞において、表記にゆれのある語が 14 語見られる。例えば、次のような例である。

「秋津コミュニティ」は、習志野市立秋津小学校区の住民が(PN4d_00022)
農業などの産業基盤、地域コミュニティーを含めた(PN4i_00014)
コメディ・ド・フウゲツという企画を三十年も続けてこられたのは(PN2k_00011)
今度は実際にコメディー映画に出たいですね(PN3j_00007)
半導体メモリー製造販売のエルピーダメモリ（東京）は九日(PN4k_00016)

ここで注意したいのは、長音符号を省く表記は、いずれも固有名のものということである。長音を省いた表記が固有名で採用されている場合、新聞としてもその表記を用いざるを得ない。一方、新聞各社の表記の基準では、長音符号で書くと定めている。その結果、表記にゆれが生じることになるのである。

４．２　英語語末別

本節では、英語の語末別に表記のゆれの割合に差異があるのか見ていくこととする。

表 4 は、度数 2 以上の外来語を対象に英語の語末別にゆれの割合を示したものである。この表では、表記にゆれの見られる語の異なり語数(「ゆれ」の欄)と異なり語数全体(「異なり」の欄)に占めるその割合を示した。

「-er 等」は、『外来語の表記』で「英語の語末-er、-or、-ar などに当たるものは、原則としてア列の長音とし長音符号「ー」を用いて書き表す。」と規定されている語である。「-ty、-dy」は「ティ（ー)」「ディ（ー)」と長音符号の前に小書きの「ィ」が表記されるものである。また「ティーディ」という清濁の関係にあるため、表では一つにまとめた。

なお-ty、-dy、-gy、-ry は、『外来語の表記』に規定のないものである。

表 4：表記にゆれのある語の割合(語末別、度数 2 以上)

	-er等			-ty，-dy			～gy			−ry		
	異なり	ゆれ	割合	異なり	ゆれ	割合	異なり	ゆれ	割合	異なり	ゆれ	割合
Web	759	162	21.3%	87	67	77.0%	17	4	23.5%	71	21	29.6%
書籍	793	170	21.4%	122	64	52.5%	34	5	14.7%	93	20	21.5%
雑誌	553	68	12.3%	85	55	64.7%	14	1	7.1%	68	12	17.6%
新聞	208	3	1.4%	24	8	33.3%	3	1	33.3%	20	2	10.0%

語末が-er 等の外来語の表記のゆれの割合は、Web・書籍が約 21%、雑誌が約 12%、新聞が約 1%で、表 2、表 3 のレジスター別順位と同じである。

一方、語末が-ty、-dy の外来語は、語末長音の表記がかなりゆれているということが分かる。Web・書籍・雑誌では半数以上の語にゆれが見られる。語末が-gy、-ry の外来語のゆれの割合は、語末が-ty、-dy の外来語よりも低いが、それでも、-gy については Web で 2 割を、書籍で 1 割を超えており、-ry については Web と書籍で 2 割を、雑誌で 1 割を超えている。

語末が-ty、-dy、-gy、-ry の語については、『外来語の表記』に規定がないことが、ゆれを生じさせる要因になっている可能性があろう。また語末が-ty、-dy の語については、NHK 放送用語委員会での委員の意見にある、「ィ」が長音を含むと思っている人が多いのではないか、実際の発音がゆれているのではないかといった観点から考察を加える必要もある。

５．終わりに

本稿では、BCCWJ のコアデータを対象に外来語表記のゆれの実態調査を行った。その結果、次のことが明らかとなった。

(1) 外来語語末長音の表記のゆれにはレジスターによる差異が見られる。Web・書籍・雑誌は、ゆれの割合の高いレジスターといえる。一方、新聞は『外来語の表記』、新聞各社の表記の基準に忠実に従い、表記をかなり統一している。

(2) 外来語語末長音の表記のゆれには、英語の語末による差異も見られた。語末が-er 等の語よりも語末が-ty 等の語の方がゆれの割合が高い。特に、語末が-ty、-dy の語では、Web・書籍・雑誌において半数以上の語に表記のゆれが見られる。

本稿は、外来語語末長音の表記のゆれに関する調査報告として位置づけられるものである。レジスターによる差異や英語の語末による差異が見られる要因について、今後、詳細に考察を加える必要がある。

また、外来語語末長音の表記には、どのように発音しているかということも関わってくる。『日本語話し言葉コーパス』を資料として、発音のゆれについても確認しておく必要があろう。今後の課題としたい。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト(基幹型)「コーパス日本語学の創成」(リーダー：前川喜久雄)、同「多角的アプローチによる現代日本語の動態の解明」(リーダー：相澤正夫)、JSPS 科研費「大規模コーパスに基づく現代語表記のゆれの実態解明」(代表者：小椋秀樹)による補助を得た。

参 考 文 献

小椋秀樹・小木曽智信・小磯花絵・冨士池優美・宮内佐夜香・渡部涼子・竹内ゆかり・小川志乃・小西光・原裕・中村壮範(2009)「『現代日本語書き言葉均衡コーパス』における形態論情報付与作業の進捗状況」『特定領域「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集』、pp.57-64.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕(2011)『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版(上・下)』(国立国語研究所内部報告書 LR-CCG-10-05-01、LR-CCG-10-05-02).

小椋秀樹(2013)「現代日本語における外来語表記のゆれ」『現代日本語の動態研究』(印刷中).

前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」『日本語の研究』4-1、pp.82-95.

宮島達夫・高木翠(1984)「雑誌九十種資料の外来語表記」『研究報告集』5(国立国語研究所報告 79)、pp.43-76.

山崎誠(2007)「『現代日本語書き言葉均衡コーパス』の基本設計について」『特定領域「日本語コーパス」平成 18 年度公開ワークショップ(研究成果報告会)予稿集』、pp.127-136.

山下洋子(2012)「外来語の発音・表記について ～[wei]のカタカナ表記と語末の長音～」『放送研究と調査』62-12、pp.74-79.

コーパスコンコーダンサ『ChaKi.NET』の連続値データ型

浅原 正幸 (国立国語研究所コーパス開発センター) *
森田 敏生 (総和技研)

Double Type Data on 'ChaKi.NET'

Masayuki Asahara (Center for Corpus Development, NINJAL)
Toshio Morita (Sowa Research Co., Ltd.)

1. はじめに

ChaKi.NET (Matsumoto et al. (2005)) は、コーパスに付与されたメタデータ・形態論情報・係り受け情報を用いて文単位の検索を行ったり、形態論情報・係り受け情報に基づく頻度統計情報を取得したり、自動解析により付与された形態論情報・係り受け情報を修正したりすることができるコーパス管理システムである。書字テキストからなるコーパスを前提とし、内部におけるデータ型はテキストやアノテーションを格納する文字列型やアノテーションを抽象化した整数型が用いられてきた。このため、音声コーパスを格納するためには書き起こしテキストのみを格納せざるを得なかった。また、書字テキストであってもコーパスを読むときの読み時間など連続値型を格納することはできなかった。

本稿では、コーパス管理システム ChaKi.NET の新しいデータ型について紹介する。ChaKi.NET に新しいデータ型として形態素単位に時刻・時間情報を格納する連続値データ型を設けることで「日本語話し言葉コーパス」(CSJ) などの音声コーパスを時刻情報に対するスタンドオフ形式で形態論情報・係り受け情報を格納することができる。

具体的には既存のデータベースの形態素に対応する単位で、開始時刻 (`start_time`)・終了時刻 (`end_time`)・継続時間 (`duration = start_time - end_time`) の三カラムからなるテーブルを追加する。時間情報を含まない通常のコーパスは各カラムの値を `null` で初期化されている。時間情報が付与されているコーパスの場合、通常のデータベース作成手続きのあと、コマンドラインから実行する `timings.exe` を用いてデータベースに格納することができる。

以下では、発話時刻が収録されている「日本語話し言葉コーパス」(Corpus of Spontaneous Japanese; CSJ) を用いた連続値データ型の活用事例を紹介する。

2. 「日本語話し言葉コーパス」(CSJ) の活用事例

以下では「日本語話し言葉コーパス」(CSJ) を利用した活用事例について示す。利用するのは「日本語話し言葉コーパス」コア RDB 版 (version 1.0) の統語情報サブセットデータベース `csj_syn.db` である。この SQL データベース形式を ChaKi.NET インポート用拡張 CaboCha フォーマットや時刻情報用 TSV ファイルを変換するために必要なプログラム `csj2cab.rb` は

* masayu-a@ninjal.ac.jp

https://github.com/masayu-a/ChaKi-CSJDB2DB から入手することができる。

2.1 初期設定

まず、CSJ のメタデータ・形態論情報・係り受け情報を ChaKi.NET のデータベースとして格納するために拡張 CaboCha フォーマットに変換する。

ChaKi.NET インポート用拡張 CaboCha フォーマットファイル csj.cabocha の生成

```
> ls csj_syn
csj_syn.db
> ruby csj2cab.rb > csj.cabocha
```

以下に生成された拡張 CaboCha フォーマットの例を示す。

ChaKi.NET インポート用拡張 CaboCha フォーマットの例

```
#! DOCID 1 <TalkID>A01F0055</TalkID><Channel>L</Channel>...(省略)
#! DOC 1
...
...(省略)
...
と 助詞,格助詞,,,,,ト,と,ト
いう 動詞,,,,ワア行五段,連体形,イウ,言う,ユー
* 7 10D 0/0 0
こと 名詞,,,,,,コト,事,コト
で 助詞,格助詞,,,,,デ,で,デ
* 8 -1ROOT 0/0 0
えー 感動詞,,,,,,エー,えー,(F エー)
* 9 -1ROOT 0/0 0
す 言いよどみ,,,,,,,,(D ス)
* 10 -1ROOT 0/0 0
発表 名詞,,,,,,ハッピョウ,発表,ハッピョー
し 動詞,,,,サ行変格,連用形,スル,為る,シ
ます 助動詞,,,,,終止形,マス,ます,マス
EOS
```

各行の形式について説明する。

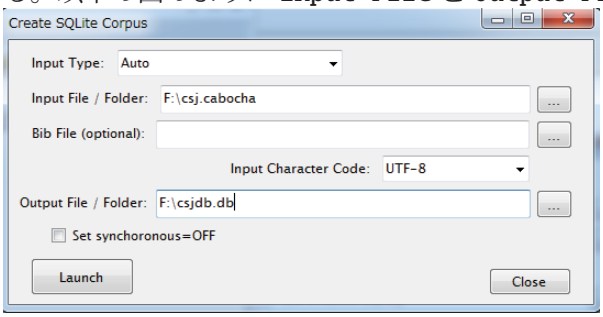
- #! DOCID で始まる行にはメタデータに相当する情報を格納する。CSJ のデータ格納においては、「談話基本情報」「話者基本情報」「対話情報」「再朗読情報」「単独印象評定情報」「集合印象評定情報」を格納する。「集合印象評定情報」については、一つのデータに対し複数の評定者による評定結果があるために、平均値を四捨五入した値を格納する。
- #! DOC で始まる行は、#! DOCID で定義したメタデータに対応するデータが以下の行から開始することを表す。
- * で始まる行は文節 ID と係り先の文節 ID、係り受けラベルが含まれている。それぞれ「文節係り受け」linkDepBunsetsu テーブル の「係り文節 ID」BunsetsuID・「受け文節 ID」ModifeeBunsetsuID・「係り受けラベル」Dep_Label を文ごとの 0-origin の値に変換して格納する。また、格納する制約上、係り受けラベルに対して表 1 のような変更を行った。

表 1　係り受けラベルの修正

元の係り受けラベル	変更したラベル
無印	‘‘D’’
‘‘A2’’	‘‘AA’’
‘‘D’’	‘‘DD’’
‘‘D_X’’	‘‘DX’’
‘‘R_P’’	‘‘RP’’
‘‘S:複数文節言い直し:S1’’	‘‘SS’’
‘‘S:複数文節言い直し:E1’’	‘‘SE’’

- EOS で始まる行は CSJ で規定されている節境界を表現する。
- 上記以外の形態素表層形で始まる行は形態素解析 MeCab の出力形式に変換したものに相当する。

生成された csj.cabocha を [Tools] → [Create SQLite Corpus] から SQLite 形式のデータベースに変換する。以下の図のように Input File と Output File を指定する。



この SQLite 形式のデータに対し、各形態素の開始時刻・終了時刻を格納するために、拡張 CaboCha フォーマットの形態論情報に対応する行（‘‘*’’, ‘‘#’’, ‘‘EOS’’ で始まらない行）に一対一対応する以下のような三列からなる TSV ファイルを作成する(1)。以下が時刻情報保存用の TSV ファイルの例である。

時刻情報保存用 TSV ファイルの例

```
...
...(省略)
...
と          4.00558         4.140573
いう        4.140573        4.196929
こと        4.196929        4.419734
で          4.419734        4.570856
えー        4.696272        4.805055
す          4.805055        4.883036
発表        5.551385        5.978241
し          5.978241        6.078064
ます        6.078064        6.532801
```

一列目は「短単位」subsegSUW の「タグ無し出現形」PlainOrthographicTranscription、二列目は「短単位」segSUW の「開始時間」StartTime、三列目は「短単位」segSUW の「終了

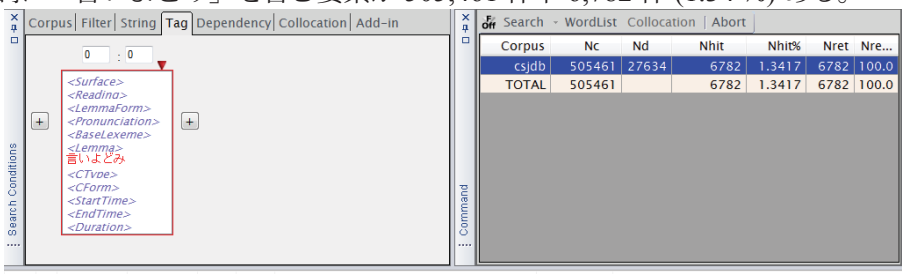
時間」EndTime に対応する。

2.2　検索事例

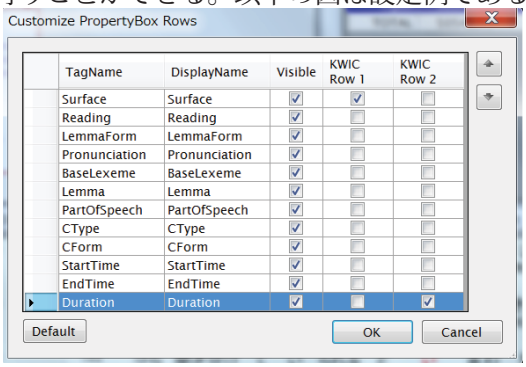
以下では、発話時間 (duration) と形態論・係り受け情報を組み合わせた検索事例について紹介する。

2.2.1　言いよどみと発話時間

以下の図は品詞に「言いよどみ」を含む要素数を [Tag Search] 機能を用いて検索した例である。品詞に「言いよどみ」を含む要素が 505,461 件中 6,782 件 (1.34 %) ある。



[Options] → [Property Box Settings] から KWIC 表示の 2 列目 (KWIC Row 2) に発話時間を表示する設定を行うことができる。以下の図は設定例である。

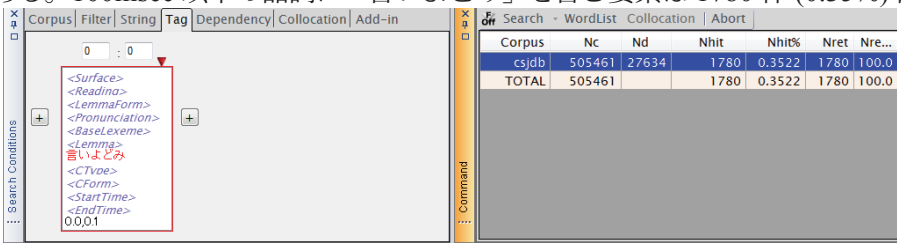


設定を行うと下図のように単語単位の KWIC 表示の下に、発話時間が表示される。発話時間上にある丸の大きさが発話時間の長さを表す。

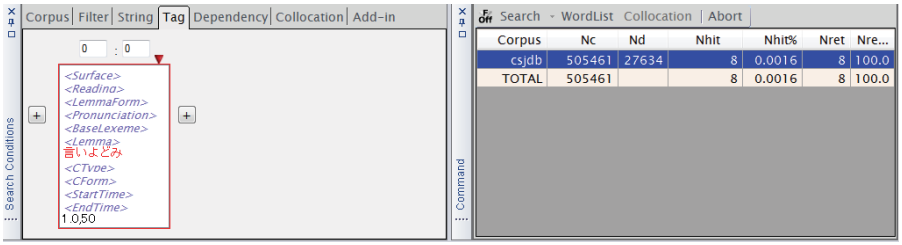
この品詞に「言いよどみ」を含む要素を発話時間を用いて絞込検索をすることができる。具

体的は [Tag Search] の Duration の列に最小値と最大値をコンマ区切りで表現することにより行う。

以下の例は 100msec 以下 (0.0 sec 以上 1.0sec 以下) の品詞に「言いよどみ」を含む要素の検索方法である。100msec 以下の品詞に「言いよどみ」を含む要素は 1780 件 (0.35%) 出現する。

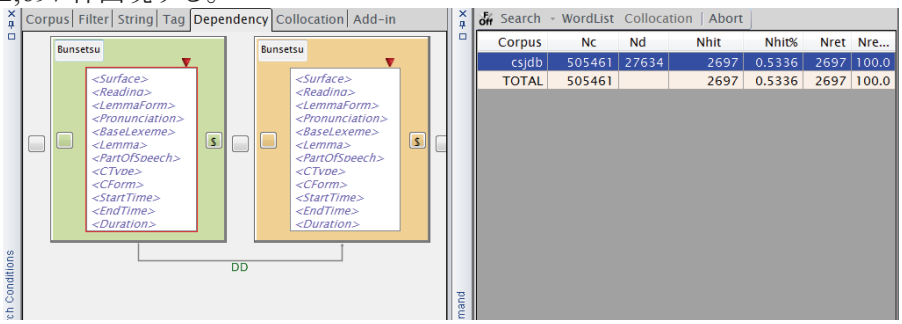


以下の例は 1000msec 以上 (1.0 sec 以上 50.0 sec 以下) の品詞に「言いよどみ」を含む要素の検索方法である。1000msec 以上の品詞に「言いよどみ」を含む要素は 8 件出現する。

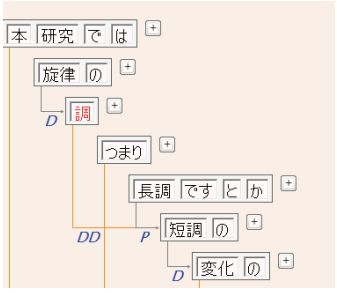


2.2.2　言い直しと発話時間

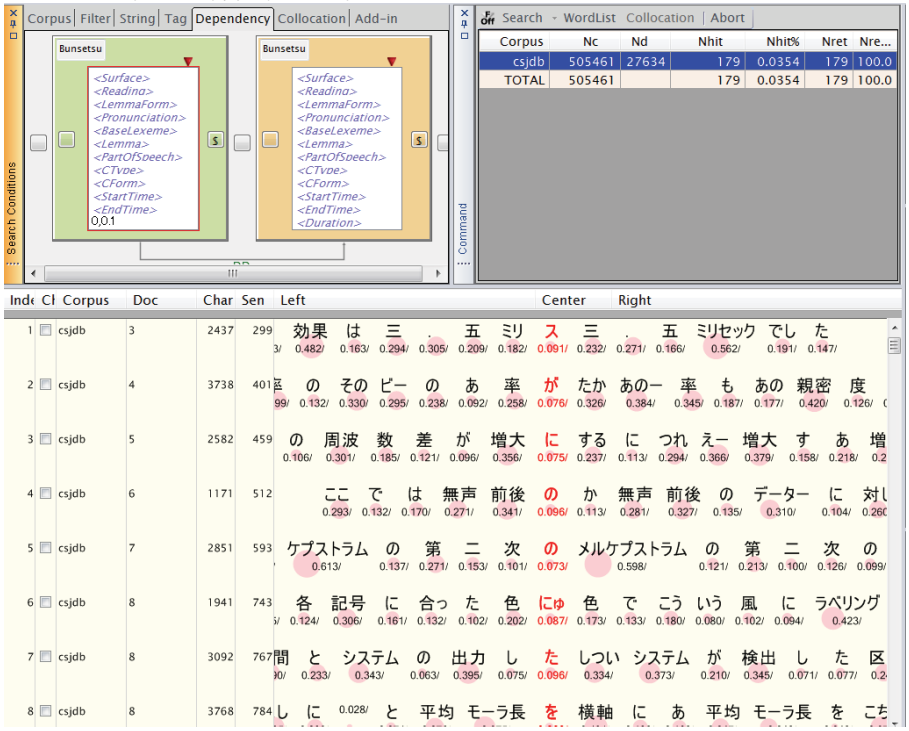
以下の図は「言い直し」を [Dependency Search] 機能を用いて検索した例である。「言い直し」は 2,697 件出現する。



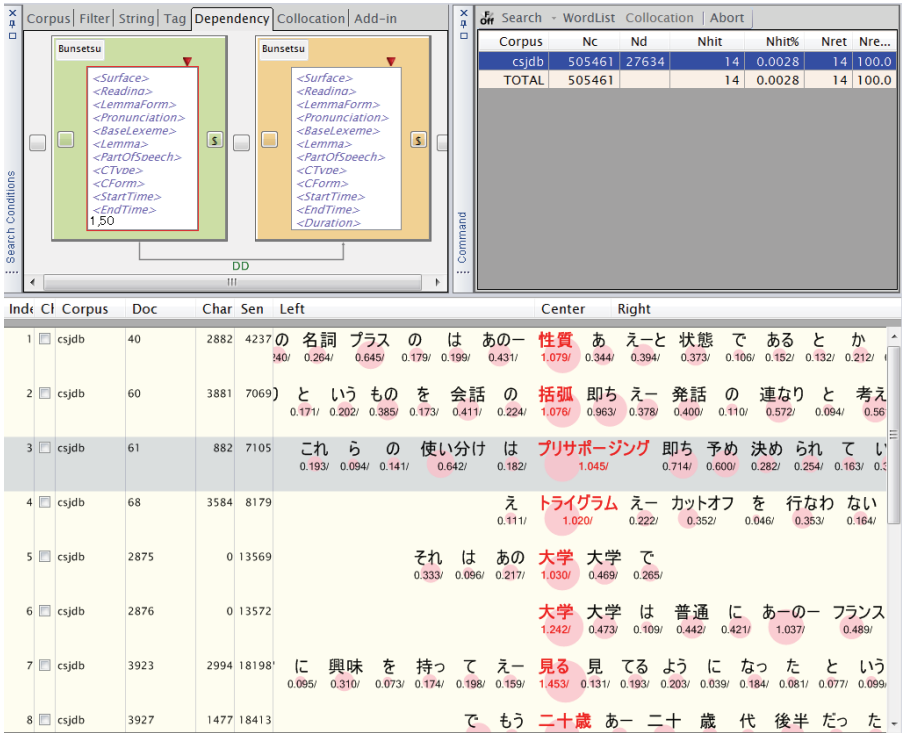
CSJ において「言い直し」は以下のような係り受け関係として表現されている。元のデータベースでは係り受けラベル ‘‘D’’ として表現されているが、他の係り受けアノテーションつきコーパスの多くが通常の係り受け関係をラベル ‘‘D’’ として表現するために、明確に区別するために ‘‘DD’’ として格納している。‘‘DD’’ の係り元が言い直す前の表現、‘‘DD’’ の係り先が言い直したあとの表現の主辞を含む文節を表す。



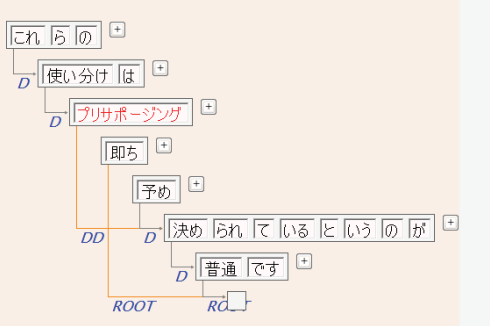
以下の検索事例では、言い直す前の表現が形態素単位で 500msec 以下である「言い直し」を検索したものである。179 件出現する。



以下の検索事例では、言い直す前の表現が形態素単位で 1000msec 以上である「言い直し」を検索したものである。14 件出現する。

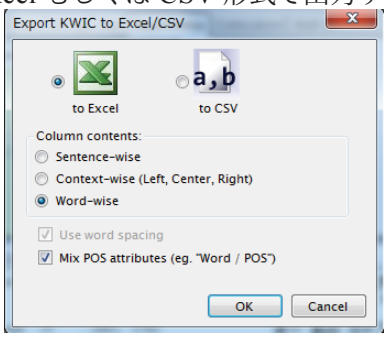


次の図は発話時間が長い形態素の言い直しの例である。長い表現の言い直しの場合、単純な単語の言い直しではなく、以下のような言い換えのようなものが含まれている。



2.3　検索結果の出力

格納されている時間情報は検索結果とともに出力することができる。[File] → [Send To Excel/CSV] より Microsoft Excel もしくは CSV 形式で出力することができる。



出力形式として、文単位 (Sentence-wise)・文脈単位 (Context-wise: 左文脈・KWIC 中央・右文脈)・単語単位 (Word-wise) の三種類を選択することができる。この際に Mix POS attributes を指定すると単語と品詞情報とともに、開始時刻・終了時刻・継続時間の三つの値を ‘‘/’’ 区切りで出力することができる。以下は [Tag Search] で「ちゃう」を検索した際に、単語単位に Excel 出力した例である。

O	P	Q
-1	0	1
崩れ/動詞/424.338043/424.642120/0.304062	ちゃう/助動詞/424.642120/424.845276/0.203146	場合/名詞/424.845276/425.126404/0.281127
違っ/動詞-促音便/446.386780/446.650543/0.263762	ちゃう/助動詞/446.650543/446.870422/0.219855	ん/助詞-準体助詞/446.870422/446.955505/0.085104
てっ/助動詞-促音便/814.408142/814.569031/0.160911	ちゃう/助動詞/814.569031/814.750854/0.181817	と/助詞-格助詞/814.750854/814.845886/0.095027
移っ/動詞-促音便/751.254944/751.534119/0.279165	ちゃう/助動詞/751.534119/751.887451/0.353369	ん/助詞-準体助詞/751.887451/751.987610/0.100138
似/動詞/105.125450/105.345215/0.219768	ちゃう/助動詞/105.345215/105.555038/0.209820	ん/助詞-準体助詞/105.555038/105.606956/0.051925
割っ/動詞-促音便/989.677856/989.826477/0.148596	ちゃう/助動詞/989.826477/990.036377/0.209935	と/助詞-接続助詞/990.036377/990.409424/0.373029
れ/助動詞/429.687408/429.772552/0.085138	ちゃう/助動詞/429.772552/429.964447/0.191878	と/助詞-格助詞/429.964447/430.040039/0.075589
なくなっ/動詞-促音便/430.345825/430.763397/0.417571	ちゃう/助動詞/430.763397/431.096466/0.333084	
れ/助動詞/466.494843/466.601105/0.106247	ちゃう/助動詞/466.601105/466.830414/0.229332	と/助詞-接続助詞/466.830414/467.372040/0.541617
狭まっ/動詞-促音便/476.929352/477.441620/0.512289	ちゃう/助動詞/477.441620/477.666962/0.225334	と/助詞-格助詞/477.666962/477.787842/0.120880
し/動詞/646.292908/646.367004/0.074088	ちゃう/助動詞/646.367004/646.499756/0.132802	言う/動詞/646.499756/646.609375/0.109579
出/動詞/647.586609/647.724670/0.138020	ちゃう/助動詞/647.724670/647.901306/0.176654	ん/助詞-準体助詞/647.901306/647.965637/0.064349
やっ/動詞-促音便/460.530823/460.676758/0.145939	ちゃう/助動詞/460.676758/460.838287/0.161542	って/助詞-副助詞/460.838287/460.984924/0.146635
焦げ/動詞/624.626648/624.800354/0.173747	ちゃう/助動詞/624.800354/625.041748/0.241395	ん/助詞-準体助詞/625.041748/625.158508/0.116716
し/動詞/663.762878/663.828552/0.065692	ちゃう/助動詞/663.828552/663.979797/0.151238	

3.　おわりに

本稿では、コーパス管理システム ChaKi.NET の新しいデータ型について紹介した。形態素単位に連続値型を三つ設けることで、開始時刻・終了時刻・経過時間の値を格納することができる。今後「日本語話し言葉コーパス」でよく利用される他の連続量について格納する方法について検討していきたい。

今回は「日本語話し言葉コーパス」を事例として活用方法について紹介した。他の利用方法

として、書字テキストの読み時間を格納することが考えられる。今後、書字テキストの読み時間を格納した場合の活用事例を紹介したい。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

Matsumoto, Yuji, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita (2005). “Chaki: An annotated corpora management and search system.” *Proc. of the Corpus Linguistics Conference Series (Corpus Linguistics 2005)*.

関連 URL

- 「ChaKi.NET」 Web ページ：http://sourceforge.jp/projects/chaki/releases/
- 「日本語話し言葉コーパス」Web ページ:http://www.ninjal.ac.jp/corpus_center/csj/

日本語名詞述語文の意味関係アノテーション

今田　水穂 (国立国語研究所コーパス開発センター) *

Annotation of Semantic Relations for Japanese Copular Sentences

Mizuho Imada (Center for Corpus Development, NINJAL)

1. はじめに

本研究は日本語名詞述語文の意味論的構造の記述を目的とする。理論的には Jackendoff (2002) の意味/概念構造における記述層 (Descriptive tier) レベルの意味情報の記述を目的とし、将来的に情報構造層など他の言語構造層のアノテーション記述との重ね合わせを想定している。

(1)

Syntax/phonology:	$[_S\ [_{NP}\ \text{Eva}]_4\ [_{VP}\ \text{became}\ [_{NP}\ \text{a doctor}]_5]]_6$
Descriptive tier:	$[_{Event}\ \text{INCH}([_{State}\ \text{BE}([_{Object}\ \text{EVA}]_4, [_{Object}\ \text{DOCTOR}]_5)])]_6$
Referential tier:	4　6
Information structure:	Common Ground_6　Focus_5

Jackendoff (2002:p.396)

Jackendoff (2002) は be に多義性を認めることについて否定的だが (p.396)、本研究は be が述語項構造として取る 2 項の関係を類型化することを目的としている。文の述語項構造を記述する上で構文主義的な考え方と語彙主義的な考え方を区別することができる。構文主義的な考え方とは構文上一定の関係にある言語要素間の意味関係を記述する考え方であり、語彙主義的な考え方とは述語的語彙について構文に関係なくその意味上の項を記述する考え方である。

(2)　震源の深さは約一〇キロ

a. be(震源の深さ, 約一〇キロ)　(構文主義的アノテーション)

b. 深さ (震源, 約一〇キロ)　(語彙主義的アノテーション)

本研究は名詞述語文という構文の網羅的なアノテーションを目的とするため、前者の考え方で意味関係アノテーションを施す。名詞述語文は必ずしも述語的名詞を含むとは限らないので、後者の方法では名詞述語文を網羅的にアノテーションすることが難しい。また、後者の考え方は実質的に述語的名詞の述語項構造アノテーションであり、名詞述語文という構文のアノテーションではない。従って後者の方法については名詞述語文のアノテーションとは独立の研究として、データも名詞述語文に限定せずにコーパス全体から述語的名詞を網羅的に収集してアノテーションすることが望ましい。

* mizimada@ninjal.ac.jp

分類基準の体系性、一貫性の確保とアノテーション作業の効率化のために、本研究ではいくつかの電子化された言語資源を利用する。名詞述語文の抽出は、京都大学テキストコーパス (以下京大コーパス) に付与された形態論情報、構文情報、格関係情報、省略情報、共参照情報などを利用して行う。これらは形態素解析器や構文解析器によって自動解析されたものを人手によって修正したものである。意味情報の付与には、日本語 WordNet と SUMO(Suggested Upper Merged Ontology) という 2 つの意味辞書を利用する。日本語 WordNet は抽出した名詞を語義と対応付けるために、SUMO は取得した語義を少数の上位クラス名に分類するために利用する。以下、使用する言語資源の概要 (第 2 節) を説明し、意味タイプ付与 (第 3 節)、名詞述語文抽出 (第 4 節)、意味関係付与 (第 5 節) の手順の説明と現在の状況の報告をする。

2. 言語資源

2.1 京都大学テキストコーパス

京大コーパスは毎日新聞 1995 年度版の本文約 4 万文に形態論情報、構文情報を付与したものである。そのうち約 5 千文には格関係、省略、共参照情報が付与されている。本研究では後者の約 5 千文を XML 形式に変換したものを使用している。このコーパスは 5127 文、50247 文節 (構文情報の付与単位)、66186 タグ単位 (格関係、省略、共参照情報の付与単位)、132327 語 (形態論情報の付与単位) を含む。

```
<document>
  <sentence S-ID="950101003-001" info="KNP:96/10/27 MOD:2005/03/08 ">
    <chunk id="0" link="26" rel="D">
      <tag id="0" link="1" rel="D">
        <tok id="0" read="むらやま" base="*" pos="名詞-人名" ctype="*" cform="*">村山</tok>
        <tok id="1" read="とみいち" base="*" pos="名詞-人名" ctype="*" cform="*">富市</tok>
      </tag>
      <tag id="1" link="37" rel="D">
        <rel type="=" target="村山富市" sid="950101003-001" tag="0"/>
        <tok id="2" read="しゅしょう" base="*" pos="名詞-普通名詞" ctype="*" cform="*">首相
          </tok>
        <tok id="3" read="は" base="*" pos="助詞-副助詞" ctype="*" cform="*">は</tok>
      </tag>
    </chunk>
```

図 1　京都大学テキストコーパス (XML 形式)

XML 形式は構文解析器 CaboCha の XML 出力形式を参考にして独自に設計したものである。document 要素はファイル全体の最上位ノードである。sentence 要素は文単位に相当し、S-ID 属性は文の ID を表す。chunk 要素と tag 要素はそれぞれ文節とタグ単位に相当し、id 属性は文節やタグ単位の ID を、link 属性は係り先文節やタグ単位の ID を、rel 属性は係り受け関係の種類 (D=通常、P=並列など) を表す。rel 要素はタグ単位が持つ格関係、省略、共参照情報に相当し、type 属性は格関係 (ガ、ヲ、ニ、etc.) や共参照関係 (=、≒、etc.) の種類を、target 属性は関係先の文字列表記を、sid 属性と tag 属性は関係先の文 ID とタグ単位 ID を表す。tok 要素は語単位に相当し、read、base、pos、ctype、cform の各属性は読み、基底形、品詞、活用型、活用形を表す。

2.2 日本語 WordNet

日本語 WordNet は Princeton WordNet の意義 (synset) に日本語を付与したものである。57238 概念 (synset)、93834 語 (word)、158058 語義 (sense) が定義されているとされる。また、概念間には上位語、下位語などの意味関係 (synlink) が定義されている。また、外部資源へのリンク (xlink) をサポートしており、SUMO へのリンクが収録されている。

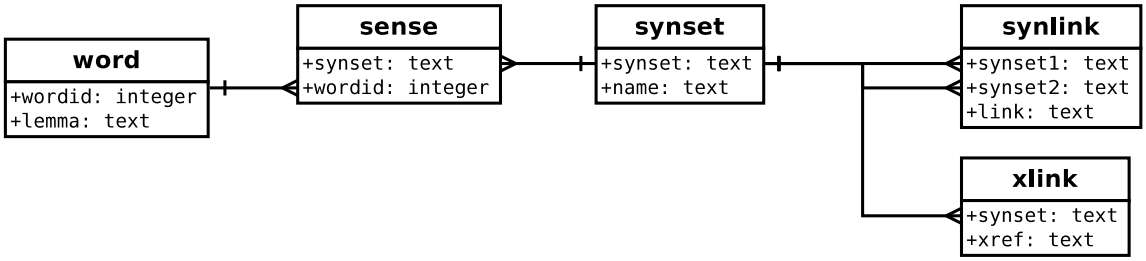


図 2　日本語 WordNet データベース (抜粋)

図 2 はデータベースの主要な部分を抜粋したものであり、実際にはより多くのテーブルやフィールドを含む。word テーブルは語を表し、wordid フィールドは語 ID を、lemma は見出し形を表す。synset テーブルは概念を表し、synset フィールドは概念 ID を、name フィールドは概念名を表す。sense テーブルは語義を表し、語と概念を多対多で関連づけるリレーションシップテーブルである。wordid フィールドは語 ID を、synset フィールドは概念 ID を表す。synlink テーブルは 2 つの概念の間の意味関係を表し、概念と概念を多対多で関連づけるリレーションシップテーブルである。synset1 と synset2 は 2 つの概念の ID を表し、link フィールドは関係の種類 (上位語、下位語、etc.) を表す。xlink テーブルは概念から外部データへのリンクを表し、synset フィールドは概念 ID を、xref は SUMO のエンティティ名を表す。

2.3 SUMO

SUMO は既存のオントロジーを統合するために開発された上位オントロジーである。SUMO は SUO-KIF(Standard Upper Ontology Knowledge Interchange Format) という言語で記述されており、約 25000 語を含むとされる。

```
(subclass Physical Entity)
(partition Physical Object Process)
(documentation Physical EnglishLanguage "An entity that has a location in space-time.
    Note that locations are themselves understood to have a location in space-time.")

(<=>
  (instance ?PHYS Physical)
  (exists (?LOC ?TIME)
    (and
      (located ?PHYS ?LOC)
      (time ?PHYS ?TIME))))
```

図 3　SUMO (SUO-KIF 形式)

括弧は式を表し、括弧内の最初の要素が述語や関数、２番目以降の要素が項を表している。

例えば (subclass Physical Entity) は subclass という 2 項述語が Physical と Entity という 2 つの要素を項に取っていることを表し、これは Physical が Entity の下位クラスであることを意味する。subclass などの述語も SUMO の中で定義が与えられている。

```
(instance subclass BinaryPredicate)
(instance subclass PartialOrderingRelation)
(domain subclass 1 SetOrClass)
(domain subclass 2 SetOrClass)
(documentation subclass EnglishLanguage "(&%subclass ?CLASS1 ?CLASS2) means that ?
    CLASS1 is a subclass of ?CLASS2, i.e. every instance of ?CLASS1 is also an
    instance of ?CLASS2.  A class may have multiple superclasses and subclasses.")
```

図 4　subclass の定義

(instance subclass BinaryPredicate) は subclass が BinaryPredicate クラスのインスタンスであることを表し、(domain subclass 1 SetOrClass) は subclass の第 1 項の定義域が SetOrClass クラスのインスタンスであることを表す (Entity や Physical など全てのクラス概念は SetOrClass クラスのインスタンスである)。documentation は概念に説明を与える 3 項述語であり、概念、説明に使用する言語、説明を表す文字列を項に取る。Entity や subclass などの名前が定項を表すのに対して、?CLASS1 など?で始まる名前は変項を表す。この他、量化を表す forall や exists や、論理演算子を表す and、or、=>、<=>などの要素がある。本研究ではこのオントロジーの一部を独自に RDB 化したものを利用している。

3. 意味タイプ付与

3.1 意味タイプの分類

SUMO のクラスツリーに含まれる図 5 のクラス名を意味タイプのセットとして用いる。

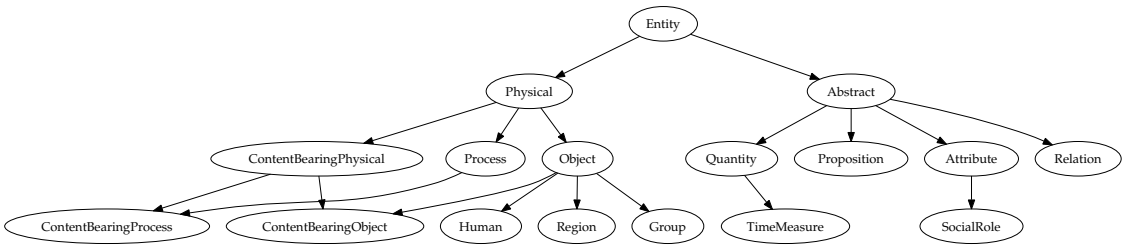


図 5　SUMO 部分集合

意味タイプ情報の自動付与は形態論情報に基づく方法と意味辞書を用いる方法の 2 種類の方法で複合的に行う。まず形態論情報に基づいて人名や地名など一部の品詞と意味タイプの対応付けを行う。次に意味辞書を用いて普通名詞など形態論情報だけでは意味タイプを特定できない語彙と意味タイプの対応付けを行う。

3.2 形態論情報に基づくアノテーション

京大コーパス (関係タグ付き) には延べ語数 132327、異なり語数 13497 の語が含まれる。まずこれらの語に対して形態論情報 (特に品詞) に基づく意味タイプのアノテーションを行った。これにより 132327 語中 50704 語 (38.3%) に意味タイプを付与した。

表 1　品詞と意味タイプの対応付け

品詞		クラス名	品詞		クラス名
名詞-サ変名詞		Process	動詞		Process
名詞-時相名詞		TimeMeasure	形容詞		Attribute
名詞-形式名詞	もの, もん	Entity	副詞		Attribute
	こと, の	Proposition	連体詞		Attribute
名詞-人名		Human	接尾辞-名詞性	助数辞	Quantity
名詞-組織名		Group	接尾辞-動詞性		Process
名詞-地名		Region	接尾辞-形容詞性		Attribute
名詞-数詞	何	Entity			
	その他	Quantity			

3.3 辞書によるアノテーション

次に普通名詞、サ変名詞、動詞、接尾辞-名詞性について、日本語 WordNet と SUMO を用いて意味タイプの推定を行った。まずこれらの語の基本形を日本語 WordNet で検索し、synset 名、synset 番号、SUMO エンティティ名を取得した。次に SUMO を用いて SUMO クラスツリーを遡り[1)]、上位クラス名を取得した。何らかの理由で上位クラス名が取得できなかった場合には synset まで戻り、WordNet における上位語 (hypernym) を取得して以降の処理を繰り返すことにより上位クラス名を取得した。複数の上位クラス名が取得された場合、2 つのクラスの間に上下関係がある場合には下位のクラス名を採用し、上下関係が無い場合には以下の順位に従ってより左側にあるものを採用した。この順位は概ね具象物が上位、下位クラスが上位に来るように配置したものをベースとして、後述する意味関係の自動アノテーションでより望ましい結果が出るように修正を施したものである。

(3) Human > Group > Region > TimeMeasure > Quantity > ContentBearingObject > Object > ContentBearingProcess > Process > ContentBearingPhysical > Physical > Proposition > SocialRole > Attribute > Relation > Abstract > Entity

これにより 21083 語に新たに意味タイプ情報を付与し、8772 語の形態論情報に基づく意味タイプを上書きした。意味タイプの上書きは形態論情報によって付与されたクラス名をその下位クラス名に書き換える場合に限定した (e.g. Process→ContentBearingProcess)。この結果、既

[1)] 例外として UnitOfDuration クラスの下位クラスおよびインスタンスについては、便宜的に TimeMeasure クラスを上位クラスとして与えた。

に付与済みの 50704 語と新規に付与した 21083 語を併せて 132327 語中 71787 語 (54.2%) に意味タイプ情報を付与した。感動詞、助詞、助動詞、接続詞、接頭辞、判定詞、特殊 (記号類) を除くと 79775 語中 71787 語 (90.0%) である。

4. 名詞述語文の抽出

4.1 述語名詞の抽出

名詞述語文の述部は通常「名詞＋コピュラ (だ・です・である)」という構造をしている。京大コーパスではコピュラに判定詞という品詞が与えられている。またコピュラが省略されている場合には CO という情報が付与されている。ただし CO は判定詞ではなくナ形容詞などの活用語尾が省略されている場合にも付与されているので、これは除外する必要がある。

これらの条件を満たすタグ単位を取得するために、まず判定詞または CO を含むタグ単位の抽出を行い、2006 例のタグ単位を抽出した。次にこれらのタグ単位について、コピュラ (またはコピュラが省略されていると思われる位置) に前接する語を調べ、活用型が「ナ形容詞」または「ナノ形容詞」、活用形が「語幹」である場合はナ形容詞などの活用語尾が省略されているものと見なし、抽出対象から除外した。除外したタグ単位の数は 103 例であり、それを除くタグ単位 **1903** 例を述語名詞を含むタグ単位として認定した。

4.2 主格名詞の抽出

主格名詞は京大コーパスの格関係情報を利用して抽出することができる。格関係情報のうち主格に相当するものは「判ガ」「ガ２」「ガ」の 3 種類がある。前節で抽出した述語名詞タグ単位 1903 例について、「判ガ」が与えられている場合にはこれを判定詞に対する主格と見なし、与えられていない場合には「ガ２」および「ガ」を主格と見なして主格名詞の抽出を行った。この結果、「判ガ」125、「ガ２」307、「ガ」2157、併せての 2589 の主格要素が得られた。主格名詞と述語名詞の文構造上の位置関係ごとに集計したものを表 2 に示す。

表 2　主格名詞と述語名詞の構造的位置関係

位置関係	判ガ	ガ２	ガ	計
同一文中	79	17	1330	1426
非同一文中	46	290	827	1163
総計	125	307	2157	2589

ここで非同一文中とは、主格名詞に相当する要素が先行文脈中にあるなど、述語名詞と同一文中に無い場合である。本研究は原則として文構造上主語と述語の位置にある 2 つの名詞の意味関係を記述することを目的としているため、2 つの名詞が同一文中にない 1163 組の事例についてはアノテーション対象から除外した。その他、名詞述語文として分析することが妥当でないと思われる事例を除外し、最終的に 2 つの名詞が同一文中にある 1426 組のうち **1370** 組の事例をアノテーション対象として抽出した。

5. 意味関係付与

5.1 意味関係の分類

主格名詞と述語名詞の意味関係について、まず次の 3 つを区別することにする。

(4) a. 分類的 (Taxonomic)
b. 非分類的 (Non-taxonomic)
c. 分裂文 (Cleft)

分類的とは 2 つの名詞の間に存在論的な上下関係か同一関係が認められる場合である。ここで上下関係とは次の関係を含む。

(5) a. 上位クラスと下位クラスの関係 (e.g. 属性 (Attribute) と色 (ColorAttribute))
b. クラスとインスタンスの関係 (e.g. 色 (ColorAttribute) と赤 (Red))
c. 上位インスタンスと下位インスタンスの関係 (e.g. 赤 (Red) と緋色 (Scarlet))

分類的関係は SUMO における subclass や instance などの 2 項述語に相当する[2)]。

(6) a. `(subclass ColorAttribute Attribute)`
b. `(instance Red ColorAttribute)`
c. `(subAttribute Scarlet Red)`

非分類的とは事物と属性 (e.g. トマトと赤) など、分類的関係以外の関係である。非分類的関係は SUMO における attribute や height などさまざまな 2 項述語に相当する。

(7) a. `(attribute Tomato Red)`
b. `(height SkyTree 634m)`

分裂文とは「トマトを食べたのはハクビシンだ」のように主語が「の」節、述語が「の」節から取り出された要素となっている文のことである。

(8) a. [S ハクビシンがトマトを食べた]
b. [S トマトを食べた] のはハクビシンだ (分裂文)

この構文の述語位置には格成分だけでなく副詞的成分が生起する場合もあるが、ここでは特に区別せず分裂文として扱う。分裂文は主語が命題、述語がその構成要素であり、両者の関係は SUMO における agent や patient などの 2 項述語におおよそ相当する。

2) 実際には SUMO では Red は ColorAttribute の下位クラスである PrimaryColor クラスのインスタンスとして定義されている。また Scarlet は SUMO には含まれておらずここで独自に定義したものである。以下、この節で示す例は説明の便宜上特に断らずに SUMO の体系に従った独自の (必ずしも厳密でない) 定義を含む。

(9) `(exists (?E ?X ?Y)`
`(and (instance ?E Eating) (instance ?X MaskedPalmCivet)`
`(instance ?Y Tomato) (agent ?E ?X) (patient ?E ?Y))`

分裂文の主語と述語の関係は非分類的関係の一種と考えられるが、分類の便宜上、特に区別しておくことにする。参考として本研究における関係の類型と SUMO(一部抜粋) との関係を図 6 に示す。図中の実線は subclass 関係、破線は instance 関係、点線は subrelation 関係を表す。

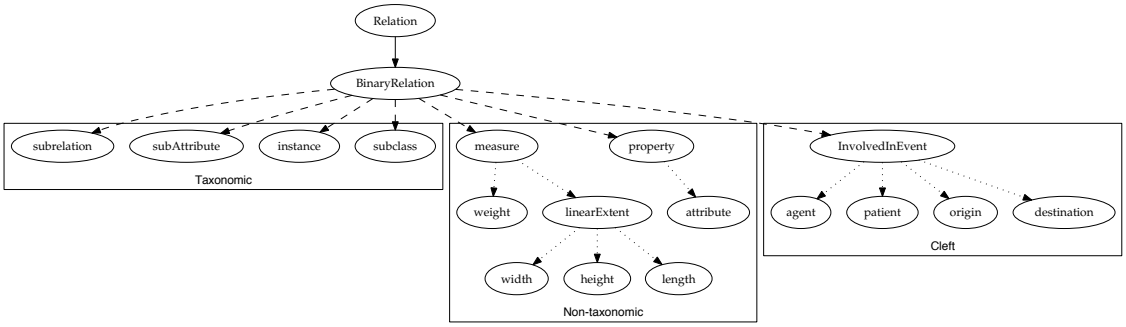


図 6　関係の分類と SUMO 部分集合

5.2　意味タイプ情報によるアノテーション

第 4 節で抽出した主格名詞と述語名詞の対 1370 組について、第 3 節で付与した SUMO 上位クラス名を用いて 2 つの名詞句の意味関係を自動判定した。まず主格名詞が「の」節であるものを分裂文と見なした。次にそれ以外の事例について 2 つの名詞句の意味タイプが同一または上下関係にある場合は分類的と見なした。さらに 2 つの名詞句がいずれも (Human, Group, SocialRole) のいずれか、または (Process, Proposition, ContentBearingPhysical, ContentBearingObject, ContentBearingProcess) のいずれかである場合もおおよそ同系統のタイプと見なして分類的とした。それ以外の事例は非分類的と見なした。また、2 つの名詞句の少なくとも 1 つに意味タイプが付与されていない事例は判定不能とした。結果を表 3 に示す。

表 3　意味関係の自動アノテーション

意味関係	分類的	非分類的	分裂文	判定不能	合計
個数	372	447	198	353	1370

判定の精度を評価するために 1370 組のうち 95/01/01 の記事に相当する部分 277 組について人手で意味関係の判定を行い、正解データを作成した。自動判定の結果と正解データを交差集計したものを表 4 に示す。「除外」は人手判定の段階で新たに見つかった除外対象の事例である。

表 4　意味関係の自動付与と人手付与

自動付与 \ 人手付与	分類的	非分類的	分裂文	除外	合計
分類的	**85**	9	0	1	95
非分類的	38	**40**	0	0	78
分裂文	5	1	**31**	1	38
判定不能	32	32	0	2	66
合計	160	82	31	4	277

分類的は $85/95 = 89.5\%$、分裂文は $31/38 = 81.6\%$ と正解率が高いが、非分類的は $40/78 = 51.3\%$ と正解率が低く、全体では $(85+40+31)/277 = 56.3\%$ の正解率である。また、不正解 121 例のうち 66 例は意味タイプが取得できなかったことによる判定不能の例である。これら 66 例を除外した場合の正解率は $(85+40+31)/211 = 73.9\%$ である。

6. おわりに

日本語名詞述語文に対する意味アノテーション計画の概要と進捗状況について説明した。今後の予定としては、自動アノテーションのカバー率と精度の向上を図り、ある程度の精度が確保できた時点でコーパス全体の自動アノテーションと人手修正を行い、作成したコーパスの公開を目指したい。

参考文献

Bond, F., H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki (2009) "Enhancing the Japanese WordNet," in *The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009*.

Jackendoff, Ray (2002) *Foundation of Language*: Oxford University Press.

Niles, I. and A. Pease (2001) "Towards a Standard Upper Ontology," in *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems(FOIS-2001)*.

Pustejovsky, J., A. Rumhisky, J. L. Moszkowicz, and O. Batiukova (2009) "GLML: annotating argument selection and coercion," in *IWCS-8 '09 Proceedings of the Eighth International Conference on Computational Semantics*.

今田水穂 (2009) 「日本語名詞述語文の意味論的・機能論的分析」，博士論文，筑波大学.

今田水穂 (2013) 「オントロジー体系を用いた名詞述語文の意味記述」,『日本言語学会第 146 回大会予稿集』, pp.156–161.

黒橋禎夫・長尾眞 (1997) 「京都大学テキストコーパス・プロジェクト」,『言語処理学会第 3 回年次大会発表論文集』, pp.115–118.

高橋太郎 (1984) 「名詞述語文における主語と述語の意味的な関係」,『日本語学』, 3:12, pp.18–39.

角田太作 (2011) 「人魚構文: 日本語学から一般言語学への貢献」,『国立国語研究所論集』, 1,

pp.53–75.
西山佑司 (2003)『日本語名詞句の意味論と語用論』，ひつじ書房.
野田尚史 (1996)『「は」と「が」』，くろしお出版.
益岡隆志・田窪行則 (1992)『基礎日本語文法—改訂版—』，くろしお出版.

謝辞

本研究の一部は国立国語研究所コーパス開発センターの「超大規模コーパス構築プロジェクト」によるものである。また、本研究は JSPS 科研費 23720225「Ruby と MSXML による日本語名詞述語文の実例調査とコーパス分析ツールの構築」(研究代表者: 今田水穂) の助成を受けている。

コロケーションとシンタクス
—形容詞と名詞のコロケーションを対象に—

スルダノヴィッチ・イレーナ（国立国語研究所・リュブリャーナ大学）†

Collocation and Syntax: Adjective and Noun Collocations

Irena Srdanović (University of Ljubljana/ National Institute for Japanese Language and Linguistics)

１．はじめに

コロケーション研究は、その始まりから、統計的な面から見た単位と単位の組み合わせに焦点をあてていたが、近年の研究ではシンタクスおよびシンタクスを超えて考えるまでの方法へのシフトが見られる（Grefenstette 1992、Stefanowitsch&Gries 2003、Tanomura 2010、Seretan 2011）。統計的な面から見たコロケーション研究は、語と語の間のスパンを 3〜5 語に設定し、単語と単語の組み合わせの強さを何種類の統計値によって計算する傾向が良く見られる（Hunston 2002）。このアプローチは、現在幅広く使われているツールにおいても、5 単語以内のコロケーションスパン、または、2 単語に絞ったコロケーション抽出などにおいて影響を与えた。それ以外にも、コンコーダンサーにおける用法･意味･パターンの観察方法が多く用いられてきた。一方、このような研究アプローチにシンタクスの面からコロケーションを観察するアプローチを加えることによって、コロケーションデータの取り出し方を更に精密なものにすることができる。

本研究では、「形容詞（連体形）＋名詞」のコロケーションを対象にして、名詞を修飾する形容詞の組み合わせ以外に、そのコロケーションがどのような構成になるか、およびそれぞれの形容詞の用法にどのような傾向があるかを現代日本語のコーパスを利用しながら検討する。従来の研究では、文中の形容詞の機能について多くの指摘があった（鈴木 1972、西尾 1972、高橋 1998、八亀 2008）。これらの研究では主として、形容詞の機能を、叙述用法、連体用法、連用用法の三つに分けて、どの機能が中心的であるかについて議論してきた。さらに、形容詞 136 語を対象にした形容詞辞書 IPAL の研究を紹介した橋本･青山（1992）および宮島（1993）が挙げられる。その研究では、三つの用法のうち、形容詞によってその用法があるかどうかを明確にし、用法のある語形についても、その用法の量的な面では偏りがあるということが述べられている。さらに、大規模データを基にしたジャンル別の分析がある。例えば、形容詞の用法分布･語義分布についての考察（姜 2012）、連体形でしか、または連用形でしか利用されない形容詞（小川他 2008）、形容詞述語のタイプと述語になりやすい語となりにくい語（前川 2012）などである。

本稿では、文節・文章における連体形の形容詞の機能およびそれに関して形容詞ごとの特集の傾向を調べる。以下のような問題点を対象にして、形容詞と名詞のコロケーションの構成を検討する。

統計的に見たコロケーション		**シンタクスを考慮に入れたコロケーション**
高い　建物	……	高い　建物
高い　国	……	教育水準　が　高い　国
高い　国	……	インフレ率　の　高い　国
高い　人	……	コミュニケーション能力　の　高い　人
[形容詞（連体形）＋　名詞]		*[文節・文章における「形容詞（連体形）＋　名詞」]*

† irena.srdanovic@ff.uni-lj.si

統計的な方法だけで形容詞と名詞のコロケーションを抽出すると、形容詞単独で名詞を修飾するコロケーション（「高い建物」）と、不十分だと考えられるコロケーション（「高い国」「高い人」）が同じように扱われている。その区別ができるようにするため、それぞれの形容詞の傾向を把握した上、「形容詞（連体形）＋名詞」のコロケーションを抽出するためのコーパスクエリシンタクス（CQS）のルールを改良する。

２．「形容詞（連体形）＋名詞」の分析およびそれぞれの形容詞の振る舞い

本章では、高･中・低頻度の形容詞（各 3 語）を選び、それぞれの形容詞と名詞のコロケーションデータのコンコーダンスを JpTenTen コーパス（Pomikálek&Suchomel 2012、スルダノヴィッチ他 2013）からランダムな 100 例を取り出し、形容詞の前後文脈まで含めて分析した。分析対象の形容詞は高頻度の「高い」「多い」「寒い」、中頻度の「青い」「甘い」「親しい」、低頻度の「痛々しい」「甘辛い」「野太い」。「手早い」という低頻度の形容詞は、最初に分析対象にランダムに選んだが、連体形+名詞の用法はないので、その代わりに「野太い」にした。

表 1 は、文節・文章における「形容詞（連体形）＋名詞」の構成タイプを形容詞ごとに示す。構成タイプは以下のように分けられる。

- 形容詞単独で名詞を修飾するもの
  （[Ai+N]、例えば「寒い季節」「高い評価」「甘い考え」など）

- 連体修飾節の述語の機能を持つ形容詞で、節が名詞を修飾するもの
  （「地震危険度の高い地域」「雨の多い国」「砂抜きが甘い店」）
  それらは、[N が Ai]N、[N も Ai]N、[N の Ai]N、N は...Ai]N、[N は Ai]N に分類して分析した。ここは、「の」「が」が最もよく使われている助詞であるが、「は」「も」もたまに見られる。時々、かかわる「名詞+助詞」と形容詞の間に副詞および他の単位が現れる。複数の単位の場合には、[N は...Ai]N のように示した。

- 所有の「の」＋形容詞で名詞を修飾するもの
  （[Poss の+Ai+N]、例えば「タレの甘辛い味」「男性の野太い声」「女の子と男の子との間の親しい関係」）

- 名詞＋形容詞という構造をもった複合形容詞で名詞を修飾するもの
  （[N+Ai]+N、例えば「香り高いコーヒー」「誇(ほこ)り高い N」「テンション高い人」）。このタイプは、複合形容詞ではなく、単純に連体修飾の「が」の省略と考えられる場合もある。

合計で見ると、最も高い頻度で現れている構成が形容詞単独で名詞を修飾するタイプであり、続いて述語の機能を持つ形容詞が多い。さらに、各形容詞の構成を量的に見ると、形容詞の振る舞いに偏りがあることが明らかになった。「多い」という高頻度の形容詞は述語の機能の出現は非常に高く、90%を超えている。「高い」という形容詞は、連体形の形容詞が連体修飾節の述語であるケースが全体のおよそ半数を占めていた（「質の高いサービス」）。または、形容詞の前に名詞が付く言語表現が 5 例あった（「香り高いコーヒー」）。「多い」と「高い」は、述語の機能を持つ形容詞として代表であり、量の意味の領域を表す他の形容詞には同じような傾向があることが考えられるが、それを確認するためにさらに数多くの形容詞の用法を検討する必要がある。一方、「青い」「甘辛い」「野太い」「甘い」「痛々しい」「寒い」は、連体形の形容詞が連体修飾節の述語になるケースがないか少ないが、形

容詞によって 7%から 17%まで所有関係の「の」の用法が多く見られる（「日本の青い空」、「男の野太い声」など）。

表 1　各形容詞によって文節における「連体形の形容詞+名詞」の構成の傾向

構成			高頻度の形容詞の例			中頻度の形容詞の例			低頻度の形容詞の例			合計
			多い	高い	寒い	青い	甘い	親しい	痛々しい	甘辛い	野太い	
形容詞（連体形）+名詞	単独[形＋名]	[Ai+N]	7	39	90	82	86	93	83	62	85	627
	述語の機能を持つ形容詞	[Nが…Ai]N		2								2
		[NがAi]N	62	9	1	1	7		6		1	87
		[NもAi]N	9	1								10
		[NのAi]N	18	39								57
		[Nは…Ai]N		3								3
		[NはAi]N	2		1							3
		合計	91	54	2	1	7	0	6	0	1	162
	所有関係	[Possの+Ai+N]		2	8	17	7	7	11	12	11	75
	複合形容詞	[N+Ai]+N	2	5								7
合計			100	100	100	100	100	100	100	74	97	871

さらに、「高い」を例にして、BCCWJ と JpTenTen からランダムに取り出した 100 例ずつのデータには「高い（連体形）+名詞」の構成における違いを調べた。表 2 に示したように、BCCWJ における単独の[Ai+N]のほうが 1 割で多く、JpTenTen における述語の機能を持つ形容詞のほうが多いという傾向が見られる。述語の機能を持つ形容詞の用法のうち、BCCWJ における[N が Ai]N のほうが JpTenTen より多く、JpTenTen における[N の Ai]N のほうが BCCWJ より若干多いと分かった。BCCWJ のレジスターデータを検索した結果、[Ai+N]・[N が Ai]N・[N の Ai]N は、書籍で最も現れていることが分かった。

表 2 文節における「高い（連体形）+名詞」の構成の傾向

構成		BCCWJ	JpTenTen
単独[形＋名]	[Ai+N]	51	39
述語の機能を持つ形容詞	[N が…Ai]N	1	2
	[N が Ai]N	14	9
	[N も Ai]N		1
	[N の Ai]N	28	39
	[N は...Ai]N	1	3
	[N は Ai]N	1	
	合計	45	54
所有関係	[Poss の+Ai+N]	1	2
複合形容詞	[N+Ai]+N	3	5

３．シンタクスを考慮に入れたコロケーション抽出

現在使われている多くのコーパス検索システムは、統計的な方法に頼り、構文的情報を直接扱うことができない。本研究で利用するスケッチエンジンというツール（Kilgarriff 2004）は、シンタクスを考えたパターンの規則化を採用していることで、いわば第 4 世代のコンコーダンスツールであると言える（McEnery&Hardie 2012）[1]。Gahl（1998）によって提案された「corpus query syntax (コーパス検索シンタクス)」を実装し、主に品詞と正規表現を活用したルールによってコロケーション抽出ができる。日本語のコロケーションが抽出できるために、日本語の文法を考えたルールを品詞、正規表現、活用形などで作成した（Srdanović ら 2008、スルダノヴィッチら 2008、2013）。

「形容詞+名詞」のコロケーションを取り出すために、まず、以下のルール 1 を利用した。このルールは、2 単位を取り出すためのルールであり、2 は、連体形の形容詞（「無い」を除外）を取り出し、1 は、名詞を取り出す(数詞を除外)。

ルール 1

```
*DUAL
=modifier_Ai/modifies_N
    2: [tag="Ai.*" & word!="ない|無い" & infl_form="Attr.*"] [tag="Pref"]? 1:[tag="N.*" &
                                                                    tag!="N.num"]
```

ルール 2 は、2 章に紹介した分析を行った上で、文節における「形容詞（連体形）+名詞」の構成を考慮に入れて改善したルールである。このルールにより、述語の役割を持つ連体形の形容詞を除外することが可能になった。述語の役割の場合には、「の」と「が」がよく使われるので、「の」と「が」が形容詞の前にない例文だけを取り出す。[2]　このルールでは、名詞と所有の「の」が形容詞＋名詞の前に来る場合は（例えば「朝の寒い空気」）、その出現はその「形容詞＋名詞」コロケーションには見られない。それでもなお、所有の「の」がないコロケーションのケースのほうが圧倒的に多いため、高頻度のコロケーション結果には差異がなく、中･低頻度のデータにも差異が少ないという結果を得た。

ルール 2

```
*DUAL
=modifier_Ai/modifies_N
        2: [tag="Ai.*" & word!="ない|無い" & infl_form="Attr.*"] [tag="Pref"]?
1:[tag="N.*" & tag!="N.num"] within! [word="が|の" | tag="N.*"] [tag="Ai.*" & word!="ない|
無い" & infl_form="Attr.*"] [tag="Pref"]? [tag="N.*" & tag!="N.num"]
```

以上のルール 1 と同じように、ここに挙げたのは 2 単位を取り出すためのルールである。2 は、連体形の形容詞（「無い」を除外）を取り出し、1 は、名詞を取り出す(数詞を除外)。

---

[1] コロケーションにおけるパターンおよびシンタクスの重要性が明らかになる中で、それによく対応しているツールがスケッチエンジンである、ということは McEnery&Hardie (2012: 257, 129, 43-45)により指摘されている。

[2] 「が/の＋「名詞」＋連体形の形容詞＋名詞」を扱うルールは別のルールにしたため、このようなコロケーションデータも抽出できる。

ただし、この形式の前に「が・の・名詞」が無い場合に限るという点はルール 1 と違う点である。

表 3 は、ルール 1 と 2 によって「高い」という形容詞を例にして、取り出せるコロケーションデータを示す。コロケーションタイプは、modifies_N である。ルール 1 とルール 2 によって抽出したコロケーションリストの違いをピンク色でマークした。「高い」が連体修飾節の述語の機能を持ち、連体修飾節が名詞を修飾する例がルール 1 のリストに見られる（表 3 の左側）。例えば、「背が高い人、コミュニケーション能力が高い人、給料が高い人」、「質が高い作品」などがある。一方、ルール 2 のリスト（表 3 の右側）にはその例がないため、単独で形容詞+名詞の例が多い（「高い精度」、「高いハードル」など）。

表 3 ルール 1 とルール 2 で取り出すコロケーションリストの例（「高い＋名詞」）

jpTenTen, ルール 1
頻度 3842021
modifies_N 1301151

物	52241	値段	5164
所	47269	買い物	4936
評価	34341	品質	4611
事	33885	効果	4484
方	20634	水準	4441
人	18208	信頼	4359
位置	17129	建物	4100
為	15942	サービス	3903
レベル	15793	筈	3836
技術	11889	上	3835
金	10532	状態	3741
声	9533	日本	3644
場所	8679	地域	3403
訳	6671	製品	3301
山	6557	そう	3207
人気	6436	数値	3142
気	6059	ビル	3131
場合	5930	値	3101
作品	5865	価格	3084
商品	5819	天井	3009
音	5566	国	2925
確率	5546	クオリティー	2917
時	5291	デザイン	2896
奴	5228		

jpTenTen, ルール 2
頻度 3842021
modifies_N 585081

所	33988	気	2806
評価	30911	数値	2662
物	23027	値	2629
レベル	13553	価格	2599
位置	13548	天井	2515
金	9636	クオリティー	2509
事	9148	精度	2419
方	8805	筈	2387
技術	8745	支持	2354
声	7709	ハードル	2300
人気	5308	目標	2299
山	5268	ビル	2292
確率	5001	次元	2275
場所	4834	奴	2099
値段	4575	給料	2094
買い物	4467	壁	2063
音	4441	安全	2044
効果	4099	パフォーマンス	2031
為	4067	耐久	1994
信頼	3777	能力	1969
水準	3738	金額	1966
品質	3680	場合	1942
訳	3196	性能	1823
建物	2970		

更に、ルール 2 で取り出したデータの制度を検討した結果、「が」「の」が形容詞から離れたところに現れる場合、「は」「も」が使われている場合だけが残され、単独の形容詞と名詞のコロケーションデータとして取り出してしまう例が 5%である。というのは、95%の制度で結果が取り出せることである。例えば、そのような「高い」の例は、以下のようなものである。

- 最新リマスタリングで揃えられているので 、資料的価値は高いはず。
- そのフォルダというラベルの付いた写真、メディア、または可能性が最も高いデータを見つけることができます。
- その効果はもちろん、成功率がとても高いことでも定評があります。
- 最後の戦いも質は高いものの、あっさり目 。
- 頭の回転が速く、人への関心が強く、社会へのアンテナも高い人です。

なお、NINJAL-LWP for BCCWJ というシステムは、コロケーションや文法的振る舞いの情報を抽出するために、係り受け関係のアノテーションを付与した。本研究で扱った「形容詞+名詞」のコロケーションの結果をみる(表 4)と、係り受け関係でも、連体修飾節の述語の機能を持つ形容詞と単独の「形容詞+名詞」のコロケーションの区別ができない。例えば、「高い」を検索すると、名詞のコロケーションには、「高い国」、「高い男」、「高い[人名]」などの不十分だと考えられるコロケーションが表示される。しかし、「の＋名詞＋形容詞」、「が＋名詞＋形容詞」のコロケーションも別のタイプとして抽出できる。

表 4　NINJAL-LWP for BCCWJ の「高い＋名詞」の結果の一部

名詞	頻	名詞	頻度
の	867	【数字】	109
もの	659	割合	107
ところ	572	場合	98
こと	529	声	94
伸び	238	比率	90
評価	236	地位	88
ん	219	とき	81
人	217	場所	79
水準	206	技術	77
ほう	186	地域	75
【地域】	163	【人名】	73
【一般】	159	値段	64
ため	145	わけ	58
位置	140	数値	58
レベル	134	男	55
山	133	国	54
よう	126	物	53

４．まとめ

本研究では、コーパスから取り出した「形容詞（連体形）＋名詞」のコロケーションを対象にして、単独名詞を修飾する形容詞の組み合わせ以外に、そのコロケーションが文節・文章においてどのような構成になるかを検討した。従来の研究は、形容詞の機能を終止用法、連体用法、連用用法の三つに分けたが、連体形の形容詞の機能を更に分けて、検討する必要があるということが本研究で明らかになった。本研究で、9 語の形容詞の用法を

検討した結果、文節・文章における「形容詞（連体形）＋名詞」の構成に以下のようなものがあった。

- 形容詞単独で名詞を修飾する（「寒い季節」）
- 連体修飾節の述語として名詞を修飾する（雨の多い国」）
- 所有の「の」＋形容詞で名詞を修飾する（「タレの甘辛い味」）
- 名詞＋形容詞という構造をもった複合形容詞で名詞を修飾する（「香り高いコーヒー」）このタイプは、連体修飾の「が」の省略と考えられる場合もある。

さらに、各形容詞を比較した結果、それぞれの形容詞の振る舞いの違いがあることが明らかになった。例えば、「高い」という高頻度の形容詞は、連体形の形容詞が連体修飾節の述語であるケースが全体のおよそ半数を占めていた（「質の高いサービス」）。また、形容詞の前に名詞が付く言語表現が 5 例あった（「香り高いコーヒー」）。一方、「青い」は、連体形の形容詞が連体修飾節の述語になるケースが 1 例しかなく、所有の「の」の用法が多く見られる（「日本の青い空」）。本研究では、高･中･小頻度の 3 語ずつの形容詞を対象にしたが、今後はさまざまな形容詞の意味領域を把握するために、数多くの形容詞を検討することが望ましい。

このようなアプローチにより、確率的な方法で取り出すには不十分だと考えられるコロケーション（例えば、「高い＋コーヒー」、「高い＋サービス」）は、更に正確に取り出せるようになる（例えば、「香り高いコーヒー」、「質の高いサービス」）。今後のコロケーション研究においても、コロケーション分析に統語的アプローチを取り入れつつ、実証的および確率的に、語およびその組み合わせの振る舞いを検討し、記述することが望ましい。確率論的アプローチに統語的アプローチを加えることによって、コロケーションデータの取り出し方を更に精密なものにすることができ、実証的なコーパス分析の有意義さが明確に示される。

謝　辞

本研究は、博報財団第 7 回「日本語海外研究者招聘事業」による研究「日本語教育における語の共起関係」（平成 24～25 年度、受入機関：国立国語研究所、招聘研究員：スルダノヴィッチ・イレーナ）の補助を得ています。

文　献

鈴木重幸（1972）『日本語文法・形態論』むぎ書房
スルダノヴィッチ・イレーナ, 仁科喜久子（2008）「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 国書刊行会, pp.59–80.
スルダノヴィッチイレーナ・スホメルヴィット・小木曽智信・キルガリフアダム (2013)「百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング」『「第 3 回コーパス日本語学ワークショップ」予稿集』 国立国語研究所, pp.229-238.
高橋太郎（1998）「動詞からみた形容詞」『言語』27:3, pp.36-43.
小川典子・李在鎬・横森大輔・土屋智行(2008)コーパス調査による形容詞の連体形と連用形の頻度, ICJLE
西尾寅弥、国立国語研究所報告 44(1972)『形容詞の意味・用法の記述的研究』秀英出版
前川喜久雄(2012)「「形容詞＋です」述語の生起要因についての準備的考察」,第 1 回コーパス日本語学ワークショップ予稿集, pp.211-220.
宮島達夫 (1993)「形容詞の語法と用法」『計量国語学』19:2, pp.94-104.
橋本三奈子・青山文啓（1992）「形容詞の三つの用法：終止，連体，連用」『計量国語学』

18:5, pp.201-214.
橋本和佳(2007)「名詞とそれを修飾する形容詞の関係」『日本語学』pp.26-10.
八亀 裕美（2008）『日本語形容詞の記述的研究―類型論的視点から』明治書院
姜 紅(JIANG Hong)(2012)コーパスに基づく多義語「甘い」の意味再分類及び語義分布調査 第 1 回コーパス日本語学ワークショップ予稿集, pp.59-68.
Cantos-Gomez, Pascual and Aquilino, Sánchez (2001) Lexical Constellations: What Collocates Fail to Tell. *International Journal of Corpus Linguistics* 6:2, pp.199–228.
Gahl, Susanne (1998) Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus, ms., ICSI-Berkeley
Grefenstette, John (1992) Use of syntactic context to produce term association lists for text retrieval, *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* , New York: ACM, pp.89-97.
Hunston, Susan (2002) *Corpora in Applied Linguistics*. Cambridge University Press
Kilgarriff, Adam, Rychly, Pavel, Smrž, Pavel & Tugwell, David (2004) The Sketch Engine. *Proceedings of EURALEX*. France: Université de Bretagne. pp.105-116.
McEnery, Tony and Hardie, Andrew (2012) *Corpus Linguistics: Method, Theory and Practice.* Cambridge Textbooks in Linguistics. Cambridge University Press
Pomikálek, Jan, Suchomel, Vít (2012) Efficient Web Crawling for Large Text Corpora. ACL SIGWAC Web as Corpus (at conference WWW)
Seretan, Violeta (2011) *Syntax-Based Collocation Extraction*. Berlin: Springer
Srdanović, Irena, Erjavec Tomaž & Kilgarriff, Adam (2008) A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing*) 15:2, pp.137-159.
Stefanowitsch, Anatol and Gries, Stefan. Th. (2003) Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8:2, 209-243.
Tanomura, Tadaharu (2010) Retrieving collocational information from Japanese corpora: Its methods and the notion of “circumcollocate”. Peter Grzybek et al.(eds.) *Text and Language: Structures, Functions, Interrelations*, pp.213-222.

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/
スケッチエンジンツール Sketch Engine　http://www.sketchengine.co.uk/
中納言コーパス検索アプリケーション https://chunagon.ninjal.ac.jp/search
NINJAL-LWP for BCCWJ http://nlb.ninjal.ac.jp

『現代日本語書き言葉均衡コーパス』「図書館書籍」の生年代別分布は何を表しているのか
ー「デナイ」「デハナイ」「ジャナイ」の使用割合から見た一考察ー

森　秀明（東北大学大学院文学研究科）

What Does the Birth Year Distribution in The Library Sub-Corpus in the BCCWJ Represent; A Study Based on the Percentage of *"denai","dewanai"* and *"janai"*

Hideaki Mori (Graduate School of Arts and Letters, Tohoku University)

１．はじめに

『日本語話し言葉コーパス』（以下 CSJ と呼ぶ）や『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と呼ぶ）には、生年代情報が付与されているサンプルがある。分析者がこれを利用して、調査対象の生年代別分布を調べることも多いだろう。しかし、そもそもこれらのコーパスで得られた生年代別分布は何を意味しているのだろうか。また、その分布を無批判に分析に利用しても良いものだろうか。ここでは前半で二つの先行研究を取り上げ、生年代別分布の捉え方や分析の問題点を考察する。後半では具体的な言語形式を取り上げ、BCCWJ の「図書館書籍サブコーパス」（以下「図書館書籍」と呼ぶ）を利用して生年代別分析を行い、実年代データと比較しながらその分布が何を意味するのかを考える。

２．先行研究

２．１　佐野(2012)における生年代別分布の捉え方と分析の問題点

佐野(2012)では陳述副詞「全然」の用法において「全然大丈夫、全然いい」などの「革新的」な用法が近年増加しているかどうかを確かめるため、CSJ を使用した生年代別分析が行われている。この分析に当たって佐野(2012)では「話者の生年における差を時間の流れに見立てて考える（見かけ時間）」(p. 36)として「見かけ時間」という捉え方が提示されている。

見かけ時間とは、時間軸上のある一点で言語データの収集を行い、その分析結果に見られる年齢差や世代差から時系列的な変化を推定する考え方である。この考え方の基盤には、人間が言語形成期（10 歳前後）に獲得した言語は終生保持されるとする臨界期記憶仮説がある。実時間で何度も調査を行うには膨大な時間や費用がかかるほか調査対象者の確保など様々な問題も多いため、社会言語学では多用されている分析手法である。

CSJ の調査年は非公開であるが、佐野(2012)はこれを時間軸上のある一点で言語データの収集を行ったものとみなして、見かけ時間で捉える立場を取る。その上で「全然」の使用法の「経年変化」をまとめたものとして、図 1 を示している。

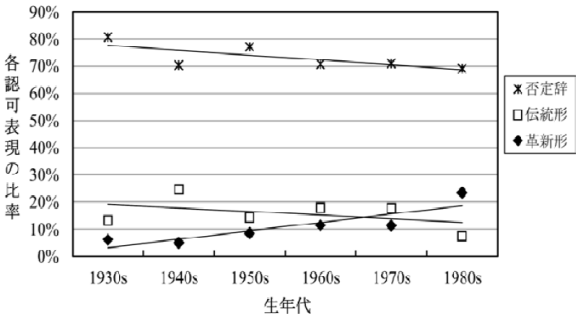


図 1　「全然」の呼応表現別経年変化
（佐野, 2012, p. 36 より転記）

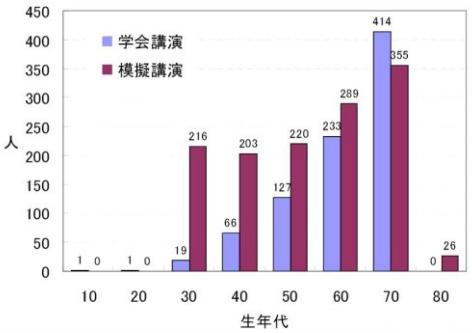


図 2 学会講演と摸擬講演話者の生年による分布
（前川, 2006, p. 6 より転記）

図 1 の「否定辞」とは「全然～ない」などの用例、「伝統形」とは「全然～違う」などの用例、「革新形」とは「全然～いい」などの用例を指す。

この分析を検討するために、簡単に CSJ の性格を見ておく。CSJ とは、国立国語研究所等が作成した、日本語の自発発話の音声を格納した大規模コーパスである。録音音声ファイルのほか、文字化して形態論情報が付与された XML 文書も提供されているため、文法や意味の分析にも利用できる。語数はおよそ 750 万語で、その内訳は学会講演が 328 万語、摸擬講演が 360 万語、その他が 62 万語となっている。学会講演とは理工学を中心とした研究発表のライブ録音で、講演者は男性の大学院生が多く、あらたまり度の高い発話が多い。摸擬講演は人材派遣会社から派遣された一般話者が、「人生で一番うれしかったこと」等について行った 10～15 分程度のスピーチで、学会講演よりもくだけた発話になっている。

このような CSJ の性格からすると、佐野(2012)の言う「革新形」は学会講演より、摸擬講演で多く用いられていることがあらかじめ予想される。それを考慮に入れた上で、前川(2006)に記載されている CSJ の講演種類別の講演者の人数を見てみよう。図 2 を見ると、学会講演と摸擬講演で人数に著しい差があることが分かる。しかも 1980s の人数はわずか 26 人しかいない。さらにその全てが摸擬講演を行っている。これを図 1 で確認すると、やはり 1980s のみ「革新形」の割合が高い。この年代を図 1 から除いた場合、「革新形」の推移はほぼ並行になっているようにも見える。これで果たして「近年多く見られるようになった革新形が確かに徐々に増加してきていることが実証された。」(p. 37)と言えるであろうか。

このような分析結果となったのは、もともとのコーパスが持っているデータの偏りを検討せず、無批判に分析に使用したためと考えられる。もし、各年代ごとに比較的人数が安定している模擬講演だけで分析を行ったら、もっとはっきりした傾向が見えた可能性がある。ただし、分析対象を摸擬講演だけに絞った場合、今度は話し言葉の一部の位相でしか時系列的な変化を見ていないことになる。

見かけ時間とは、年齢差や世代差から時系列的な変化を推定することであった。しかし、何をどれぐらいの精度で推定するかは、それぞれの分析目的によって異なるだろう。一口に話し言葉と言っても、そこには様々な姿がある。一人の人間が話す言葉でも、聞き手の年齢や地位、あるいはプライベートな場かフォーマルな場かなどによって使用される言葉は変化する。これら様々な社会的要因によって特徴づけられた言語使用の現象を位相と呼ぶなら、できるだけ多くの位相を対象にして現代日本語の平均的な姿を推定しようとするか、言語変化が現れやすい一部の位相を対象にして変化の傾向を推定しようとするかでは、結果が当然異なる。佐野(2012)では CSJ の全データを分析対象としているから、フォーマルな話し言葉も含めた現代日本語でどのような変化が起きているかを調査する目的があったと思われる。もしそうであれば、学会講演と摸擬講演のデータ数のばらつきを補正した上で分析すれば、より「推定したい現代日本語」に近似した結果が得られた可能性がある。

一方、分析の精度についてはどうだろう。CSJ では調査年代が非公開になっているが、これを仮に 2000 年に行われたものだと仮定すると、生年代 1970s の世代は調査当時 30 代、1930s は 70 代になる。実時間で 2000 年の「現代日本語」の姿とは、おおよそこれら 30 代～70 代の平均的な言葉遣いだといえよう。この平均が仮に 40 代のものと同じだとしたとき、70 代はおよそ 30 年前、30 代はおよそ 10 年後の平均的な言葉遣いを表していると推定される。もし 2010 年に CSJ と同様の調査を行い、その平均が 30 代の言葉遣いと一致すれば、この見かけ時間による推定は精度が高かったことになる。

しかし、このような推定の精度が必ずしも高くないことは日比谷潤子編著(2012)などの入門書でも繰り返し指摘されている。見かけ時間を実時間と見なす上での問題点を横山(2011)で示された枠組みを利用して確認してよう。図 3 は横山(2011)で図示されている言語運用能力の生涯習得モデルである。言語運用能力はいつ生まれたかに最も影響される。これは臨界期記憶仮説をもとにした要因で「世代効果」と言える。さらに年齢を重ねた分の「加齢効果」と、調査した年代の言葉遣いの影響である「時代効果」が加わる。この二つの効果は調査した年によって決定されるため「調査年効果」とまとめることができる。先の仮定

に沿って生年代1930sの例で考えれば世代効果とは1930s＋10歳=1940s頃に言語形成期を迎えたことによる言葉遣いの影響、加齢効果とは平均の40代より30年分年齢を重ねたことによる知識の増加や70代にふさわしい言葉遣いを選択することなどによる影響、時代効果とは2000年当時の平均的な言い回しや、流行語などによる影響ということになる。

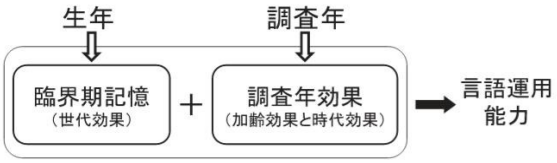


図3　生涯習得モデルの枠組み（横山, 2011, p. 29より転記）

佐野(2012)では、調査年効果による影響が考慮されていない。CSJの生年代別比較では時代効果は同一と仮定して無視できるとしても、加齢効果が使用割合に影響している可能性がある。やはり「実証」というのは難しく、おおまかな推定を示した分析と言えよう。

以上、佐野(2012)から生年代別分布を見かけ時間で捉える考え方があること、その分析に当たっては調査年効果の影響やコーパスに潜在しているデータの偏りを考慮する必要があることを述べた。

２．２　前川(2012) における生年代別分布の捉え方と分析の問題点

前川(2012)では、「大きいです。」「小さいです。」のように、形容詞に助動詞の「です」が直接後続して文末を形成しているタイプの述語の成立要因を、BCCWJを用いて検討している。この4.5節に「A（形容詞）＋です。」述語の社会言語学的動向を知るために、「使用者の年齢を検討する」として、生年代別分布の分析が行われている。これは生年代別分布を見かけ時間（いわば横の変化）で捉える立場を取らず、年齢による差（いわば縦の変化）で捉える考え方だと言えよう

前川(2012)では、「図書館書籍」と「ベストセラー」を用いて分析が行われているため、はじめにこれらの性格を確認しておく。「図書館書籍」はBCCWJの中核をなすサブコーパスの一つで、語数はおよそ3,000万語である。東京都内の公立図書館に所蔵されている書籍のうち、1986年から2005年の20年間に出版されたものを対象とし、ランダムサンプリングによってデータが集積された。書き言葉（書籍）が社会に流通している実態を公立図書館の所蔵状況によって近似的に捉えたもので、「流通実態サブコーパス」とも呼ばれる。出版年とジャンルに偏りが出ないように構成比を割り当てて10,551個のサンプルを選び、そこからさらにランダムに範囲を指定して各サンプルごとに約1,000文字を抽出している。このようにサンプルごとの文字数まで均等に抽出したデータを「固定長サンプル（約670万語）」と呼ぶ。この他に章や節などのまとまりのある範囲で最大1万字のデータを抽出したものを「可変長サンプル（約2,889万語）」という。この可変長サンプルはサンプルごとに語数のばらつきがあるため統計分析に使用する場合は注意が必要となる。これら固定長サンプルと可変長サンプルはその一部が重なっているものが多いため、この二つを足して重なりを除いた語数が約3,000万語となる（このサンプルを可変長、固定長に対して「両方」と呼ぶことにする）。

一方「ベストセラー」は、無作為抽出を施していない「特定目的サブコーパス」の一部を構成する、いわばサブコーパスのサブコーパスとでも言うべきものである。1976年から2005年までの30年間に出版された書籍のうち、『出版年鑑』などで年間20位に入った書籍951冊のうち、「作業上の理由」から695冊を選び、可変長のみを抽出した（以上、詳しくは『現代日本語書き言葉均衡コーパス』利用の手引第1.0版参照のこと）。

まず、生年代を年齢と捉える考え方から検討する。「図書館書籍」のサンプル数は、10,551

個である。ただし、サンプルには生年代が不明のものや共著・共訳による書籍がある。共著・共訳ではすべての執筆者の生年代が一致する場合（以後これを「同年代共著」と呼ぶ）のみが生年代分布の分析に使用できる。ただし BCCWJ の Web 検索サイトの『中納言』で公開されている資料（BCCWJ_WC_LUW_v10.xlsx）には、翻訳者・共著者の区別ができる情報がないため、ここでは単著で生年代が判明しているサンプルのみを扱う（表 1）。これにより生年代が判明した 7533 個のサンプルを生年代別（年齢別）に示したのが図 4 である。

表 1　「図書館書籍」生年代の判明・不明サンプル数

判明	不明	翻訳・共著	合計
7533	1164	1854	10551
71.40	11.03	17.57	100.00

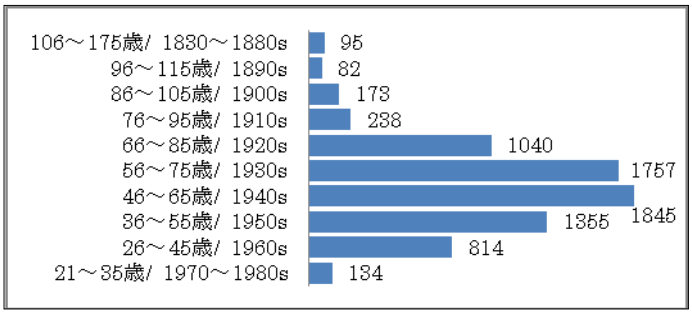


図 4　「図書館書籍」の生年代別サンプル数（翻訳・共著を除く）

縦軸にしている「年齢/ 生年代」を検討してみる。「図書館書籍」は 20 年間に渡って出版された書籍からサンプリングしている。このため一つの生年代における年齢の幅は、20 歳あることにある。理論的には 1970~1980s は 6 歳～35 歳となるが、1970s は 1993 年以後、1980s は 2001 年以後のものしかない。これはいわば「観察打ち切り」のデータでサンプリング対象が 2005 年で区切られているため、20 歳分のデータが集積できなかったことによる。

ここから縦軸を順に見ていくと、1910s 以上は実際に執筆した年齢とは見なしにくい。特に最上段の 1830~1880s では人間が生存して執筆できる年齢とは考えられない。これと横軸のサンプル数を比べると、1910s から極端に少なくなっている。個別に確認しないと確実ではないが、1910s 以前の生年代は、生前に執筆した書籍が再出版されている可能性が高い。そのように見なしたときこれらの生年代は全生涯のかなりの部分に渡る年齢となる。

以上のように検討すると生年代を年齢で捉えた場合、およそ 20 年幅の年齢と全生涯のかなりの部分に渡る年齢とが混在することになる。年齢という一つの指標に 2 つの基準が混在するため、生年代を年齢で捉える考え方には問題があると言えるだろう。

次に前川(2012)の分析内容を検討する。前川(2012)の分析結果を転記したものが図 5 である。図 5 では横軸が生年代となっているが、前川(2012)ではこれをあくまで年齢で見ている。

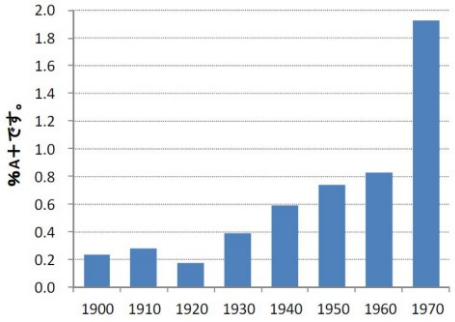


図 5　書き手の生年代と「%A+です。」の関係（前川, 2010, p. 220 より転記。縦軸は%）

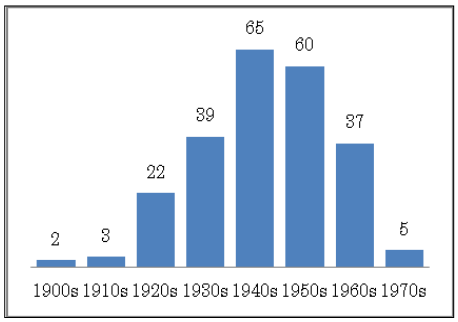


図 6　「図書館書籍」＋「ベストセラー」における「A です。」の頻度（縦軸は用例頻度）

「%A+です。」とは、「A。」の頻度と、「A＋です。」の頻度合計に対する「A＋です。」の割合を指す。この分析から前川(2012)では「書き手の年齢が低下するにつれて「%A+です。」

述語の使用率が上昇する」ことを指摘した上で、「特に 1960 年代から 1970 年代にかけて著しく上昇していることは注目に値する。」と述べている。

図 6 は図 5 のもとになった「A＋です。」の頻度を再調査した結果である。調査には『中納言』を使用し、検索対象を「図書館書籍」＋「ベストセラー」に限定した上で、キー: 形容詞、後方共起 1: 書字出現形「です」、後方共起 2: 書字出現形「。」、サンプルは「両方」で検索した。なお検索データの分析には田野村忠温氏が開発し、無償でネット公開されている sortKWIC を使用した（3 節の分析も同じ）。

図 6 を見ると、両端の生年代で著しく頻度が小さい。特に前川(2012)で「注目に値する。」と述べられた 1970s の頻度はわずか 5 例しかなく、これで 1%の有意差を評価するのは統計的に困難だと思われる。そもそも両端の頻度が極端に少なくなるのは、分析対象としたコーパスが持っている分布の偏り（図 4）のためである。「図書館書籍」では、出版年やジャンルには厳格な構成比の割り当てがなされているが、生年代においてこの点は考慮されていない。生年代別分布の分析に当たっては、特にコーパスに潜在している分布の偏りに留意する必要がある。さらに 1970s 以後は、「観察打ち切り」によって十分なデータが集積されていない。この年代に特異な振る舞いが観察された場合は、慎重な判断が必要であろう。

以上、前川(2012)の検討から「図書館書籍」の生年代を年齢で捉えるのは問題があること、生年代別分布の分析に当たっては、コーパスに潜在しているデータ分布の偏りや観察打ち切りデータの扱いに留意すべきことを述べた。

２．３　「図書館書籍」を見かけ時間で捉える場合の問題点

佐野(2012)と前川(2012)から生年代分布の捉え方には見かけ時間と年齢という二つの捉え方があることが分かった。しかし「図書館書籍」の場合、これを年齢で捉えると一つの指標に二つの基準が存在することになり問題があった。それでは「図書館書籍」を見かけ時間で捉えることは妥当なのだろうか。CSJ はある一点と見なしうる調査年代に実際に発話されたデータを集積したコーパスだが、「図書館書籍」は 20 年に渡って出版された書籍を集積している。しかも 1910s 以前は生前に出版された書籍が再出版されている可能性がある[1]。

この問題を図 3 の生涯習得モデルの枠組みから考えると、世代効果は一義的に決まるが、加齢効果は 1920s ～1960s 頃までは 10 年ごとの重なりを持つ 20 歳幅の効果、1910s 以前は全生涯のかなりの部分における各加齢効果の合計となる。図 3 の調査年を出版年とした場合、そもそも 1910s 以前は、出版年によって加齢効果が決まるわけではない。これは時代効果も同じで 1920s 以後は出版年である 1986 年～2005 年の時代効果の影響を受けていると考えられるが、1910s 以前は執筆者の生年－没年で時代効果が異なる。

このように考えると生年代別分布を見かけ時間で捉えた場合、理論的には解決が難しい問題が存在する。理論的にはこれ以上踏み込めないため、次節ではケーススタディを行って生年代別分布と実時間データの分布を比較し、生年代別分布を見かけ時間で捉えた場合、実際にどのような問題が生じるのかを確かめてみる。

３．ケーススタディ

３．１　用例の抽出

検索対象は BCCWJ の「図書館書籍」、調査対象の言語形式には名詞述語とナ形容詞述語の否定形に使用される「デナイ」「デハナイ」「ジャナイ」（以後「コピュラ否定形」と呼ぶ）を使用し、その生年代別分布を観察する。コピュラ否定形を使用するのは、生年代によって使い分けがあるものの、その合計数はあらゆる生年代で一定であると考えられるからである。一般的に考えて、特定の年代で動詞述語やイ形容詞述語だけが使われやすかったり、肯定形に比べて否定形が少なく使用されるなどの変動は考えにくいだろう。

[1] 1910s 以前に限らず、書籍にはそもそも執筆年と出版年が必ずしも一致しないという問題がある。丸山(2012)によれば、書籍の大部分は初出から 3 年以内に出版されているため、ここではその問題には触れない。

ただし、名詞述語とナ形容詞述語の否定形には丁寧体において「デハアリマセン」等も使用される。「図書館書籍」では、これらがコピュラ否定形の 1 割程度使用されているが、これは今回の分析の対象とはしない。ただし、同じ丁寧形でも「デハナイデス」等の形式は含まれる。

用例の抽出には『中納言』を使用して検索対象を「図書館書籍」に限定し「デナイ」の場合、キー: 未指定、後方共起 1: 書字出現形「で」、後方共起 2: 語彙素「無い」＋品詞「形容詞」で検索した。サンプルは「両方」、単位は「短単位」を選択した。このように抽出したデータから単著で生年代情報が得られているものに加え、単独の訳者による翻訳、同年代共著を含めて形式・年代別に示したのが表 2 である。以後の分析はこれをもとに行う。

表 2 「図書館書籍」におけるコピュラ否定形の形式・生年代別頻度

形式	~1880s	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s~	小計	不明	合計
デナイ	150	129	307	359	1270	1763	1854	1211	753	76	(7872)	1460	9332
%	30.80	22.09	23.85	21.23	19.28	14.76	15.36	12.11	12.19	8.47	(15.22)	17.34	15.51
デハナイ	287	333	787	1052	4370	7552	7601	6015	3362	458	(31817)	4654	36471
%	58.93	57.02	61.15	62.21	66.34	63.21	62.95	60.16	54.42	51.06	(61.50)	55.28	60.63
ジャナイ	50	122	193	280	947	2633	2619	2773	2063	363	(12042)	2305	14348
%	10.27	20.89	15.00	16.56	14.38	22.04	21.69	27.73	33.39	40.47	(23.28)	27.38	23.85
合計	487	584	1287	1691	6587	11948	12074	9999	6178	897	(51732)	8491	60151
%	0.81	0.97	2.14	2.81	10.95	19.86	20.07	16.62	10.27	1.49		14.00	100.00

＊「~1880s」にはそれ以前の年代、「1970s~」にはそれ以後の年代を含めた。

＊生年代情報がないもの、共著・共訳で執筆者の年代が異なるものは不明とした。

３．２　実時間データとの比較

実時間データと比較する前に、「図書館書籍」の場合、何をもって実時間データとみなすのかを定義しておく必要がある。一般に実時間と言った場合、書籍が執筆された年と考えやすい。しかし「図書館書籍」は、そもそも流通実態を反映するために設計されたコーパスである。古い時代に執筆出版された書籍が再出版された場合、そこで使用されている言語も流通実態としては出版年に流通している言語の一部を構成している。そこでここでは「図書館書籍」における実時間の使用言語を「ある年代に出版され図書館に所蔵された書籍の総体に使用されている言語、あるいはそれらの平均」と定義する。なお、古い年代においては『太陽コーパス』を実時間データとみなして使用する。『太陽コーパス』は、2005 年に国立国語研究所が刊行した近代語研究のためのコーパスである。明治末期から大正末期にかけて、当時最もよく読まれ、多彩なジャンルに渡る記事を掲載していた雑誌『太陽』を、1895,1901,1909,1917,1925 年の各年ごとにコーパス化したもので、約 1400 万字が収録されている。『太陽』は一つの雑誌に過ぎないが、これまでの先行研究において当時の書き言葉の使用実態を反映していると評価されているため、これをここでは実時間データとみなすことにする。

実時間データの『太陽コーパス』と「図書館書籍」の出版年別にコピュラ否定形の頻度を調査した結果が表 3、表 4 である。表 4 の「図書館書籍」出版年は 1986～1989 年の 4 年分を 1988 年、1990～1999 年の 10 年分を 1994 年、2000～2005 年の 6 年分を 2002 年と表示した。これらを使って、表 2 の「図書館書籍」の生年代分布との比較を行う。

表 2 のデータと表 3、表 4 のデータを重ね合わせてグラフを描いたものが図 7 である。ただし、図 7 において実時間と見かけ時間の時間幅は同じではない。また、グラフを重ねるに当たっては最も割合が近似すると思われる年代に重ねたが、この重なりが直ちに同年代を意味するものではない。あくまで生年代別分布を何の補正もせず、単純に見かけ時間と

見なした場合、実時間とどのような関係になるかを比較するためのグラフである。

表 3 『太陽コーパス』の出版年別分布

形式	1895	1901	1909	1917	1925	合計
デナイ	162	394	986	1047	843	3432
%	38.39	46.08	45.04	40.52	32.98	39.88
デハナイ	197	316	947	1152	1427	4039
%	46.68	36.96	43.26	44.58	55.83	46.93
ジャナイ	63	145	256	385	286	1135
%	14.93	16.96	11.69	14.90	11.19	13.19
合計	422	855	2189	2584	2556	8606

表 4 「図書館書籍」の出版年別分布

形式	1988	1994	2002	合計
デナイ	1222	5046	3064	9332
%	17.53	15.64	14.65	15.51
デハナイ	4095	19574	12802	36471
%	58.75	60.65	61.23	60.63
ジャナイ	1653	7653	5042	14348
%	23.72	23.71	24.12	23.85
合計	6970	32273	20908	60151

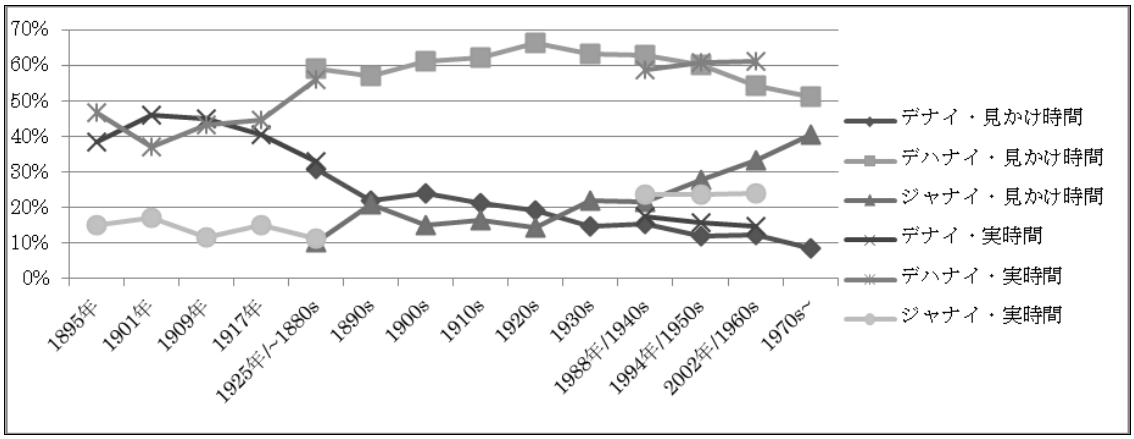


図 7 実時間『太陽コーパス』、見かけ時間「図書館書籍」生年代、実時間「図書館書籍」出版年の比較

初めに実時間から検討する。実時間では 1895~1917 年頃まで、デナイとデハナイが同程度の割合で使用されている（1895 年と 1901 年でグラフが変動するが、この頻度を表 3 で確認するとこの区間のみ他の区間の 1/6~1/3 程度しかない。このためこの変動はデータ不足による可能性が高いと思われる）。それが 1925 年になると大きく使用傾向が別れ、デナイよりデハナイを使用する割合が増えている。この時期は言文一致運動の成長期で、特に 1922 年には全国誌の新聞でも全紙面で言文一致を採用するに至っている。この変化はそのような時代の流れを反映したものと思われるが、これについては稿を改めて述べる。

一方、1988~2002 年（正確には 1986～2005 年）の区間では、ジャナイは 24%程度でほぼ一定で、デハナイが 6 割弱から 6 割強へ微増、デナイが 17%半ばから 14%半ばへ微減している。わずかな変化は見られるが、ほぼこの区間のコピュラ否定形の使用割合は一定だと言えるだろう。これらは実時間のデータをもとにした分析であり、流通実態としての書き言葉の平均的な姿を、ある程度正確に反映していると考えられる。

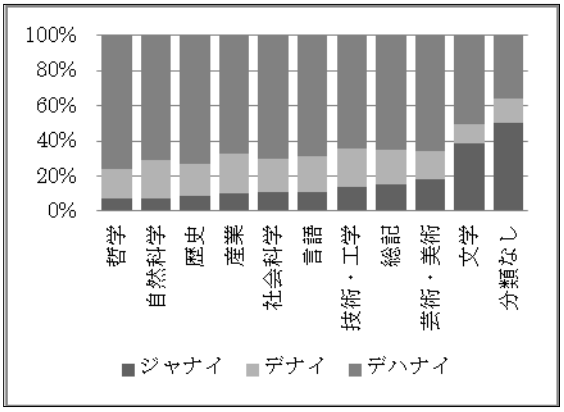
次に見かけ時間を検討する。~1880s～1940s までは、生年代によって多少のでこぼこがあるもののほぼ一定の傾向を示している。デハナイは 6 割程度であまり変化がない。デナイは 3 割程度から徐々に半減していく。ジャナイは 1890s で特異に増加した後はあまり増加せず、1920s で急に増加した後、1930s から 1940s の区間ではまた水平になる。全体的な傾向ではデハナイはほぼ横ばい、デナイは 30%から 15%へ半減、ジャナイは 10%から 20%強への増加となっている。~1880s から 1910s までの区間はデータ数が少ないものの、1890s の特異な変動を除けば、実年代の 1917 年から 1925 年までの傾向をそのまま自然に推移させた姿のように見える。また３．２節の考察によれば、1910s と 1920s には加齢効果と時代効果に違いがあるはずである。しかし図 7 ではこの区間で大きな変化があるようには見えない。

一方、1950s 以降はデハナイとジャナイに急激な変化が見られる。デナイはそれまでの傾向通り徐々に割合を下げていくが、デハナイは 1940s の 6 割強から 1970s の 5 割強へおよそ 10%減少、ジャナイは同様に 21％強から 40％強へ増加している。これを実時間と比較すると、大きく異なる値となっているのが分かる。表 2 で頻度を確かめても、1950s と 1960s には十分な頻度が確認されることから、この区間における実時間と見かけ時間の乖離は、データ不足が原因ではないと考えられる。

以上、実時間と見かけ時間を比較することにより、見かけ時間の 1940s 以前は、1890s の特異な変動を除けば、おおよそ実時間の傾向のまま推移しているように見えるが、1950s 以降は実時間と大きく乖離していく傾向があることを観察した。

３．３　ジャンル効果

前節では見かけの時間の 1950s～1970s~の区間が、実時間と乖離していく傾向を観察した。しかし、なぜこのような傾向が観察されるのだろうか。言語変化は若い世代ほど起きやすいと言われている。その加齢効果が強く影響しているのであろうか。それにしてはあまりにも急激な変動に思われる。そこで加齢効果以外の要因を検討するため、まずジャンルごとにコピュラ否定形の使用割合に変化があるかどうかを調査した。



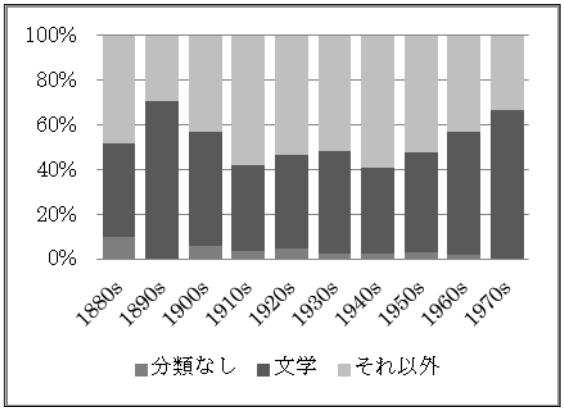


図 8 「図書館書籍」ジャンル別コピュラ否定形割合　図 9 「図書館書籍」生年代別ジャンル構成比率

図 8 を見ると文学と分類なしで他のジャンルの数倍ほどジャナイが使用されていることが分かる。分類なしには明らかに小説と考えられる書籍が多く含まれているから、基本的に会話の描写が多い書籍でジャナイが使用されやすいと考えて良いだろう。

次に図 9 で生年代別のジャンル構成を見ると、1950s から次第に「文学＋分類なし」が増加していることが分かる。図 7 の見かけ時間において 1950s～1970s~の年代でジャナイの比率が増加していたのはこの影響による要因が大きいと考えられる。ただし、図 9 ではこの年代以外にも偏りが見られる。1880s～1900s では総じて「文学＋分類なし」の割合が高く、特に 1890s の偏りが著しい。これを図 7 で確認すると、この年代だけ特異にジャナイの比率が高くなっている。これを補正すればこの区間はもっとなめらかに推移すると考えられる。

「図書館書籍」で古い年代の文学比率が高くなるのは、時代を超えて読み継がれる書籍には文学が多いからだろう。一方、新しい年代で高くなるのは若い世代が書籍を出版できるとしたら、専門書などより文学の可能性が高いからであろう。これら「図書館書籍」に潜在する要因で、分布が偏っていたのである。

以上から、コーパスの生年代別分布には、ジャンルによる対象言語の使われやすさ（ジャンル別使用傾向）と、生年代別のジャンルの偏り（生年代別ジャンル偏向）が影響していることが分かった。さらに「図書館書籍」でジャンルごとの可変長語数を比較すると、最も少ない「言語」に対して最も多い「文学」では約 1．7 倍になっている。このため検索

オプションを固定長以外にして検索すると、この影響も受ける。これを（ジャンル別可変長偏向）と呼ぶことにする。これらは連動して働き、分布に影響を及ぼす。この３つを一括してジャンル効果と呼ぶと、コーパスの生年代分布はジャンル効果によって影響されていると考えられる。

３．４　調査年効果

２．３節で考察したように 1910s 以前の加齢効果と時代効果は「図書館書籍」の出版年とは無関係であると考えられる。これらの調査年効果を調べるには個別のサンプルを一つ一つ確認していくしかない。BCCWJ の DVD 版には書誌情報が収録されているが、丸山(2012)によれば初出・初刊の情報はおよそ 25%ほどしか取得できないため、この区間の調査年効果を正確に観察することは難しい。また、1970s~は 1993 年以後のサンプルしかない。よってここでは 1920s～1960s のみを対象に調査を行う。

調査は表 2 のデータを 5 年間の出版年ごとに区切り、コピュラ述語否定形の割合をグラフに描いて観察する。本来 1 年ごとに比較を行うべきだが、表 2 のデータを 20 分割すると頻度が少なくなる上に、グラフの線が重なりすぎて分かりにくくなる。このため 5 年分ずつの比較とした。図 10～12 の 1986-等の表示は 1986-1990 年等の 5 年間であることを表す。

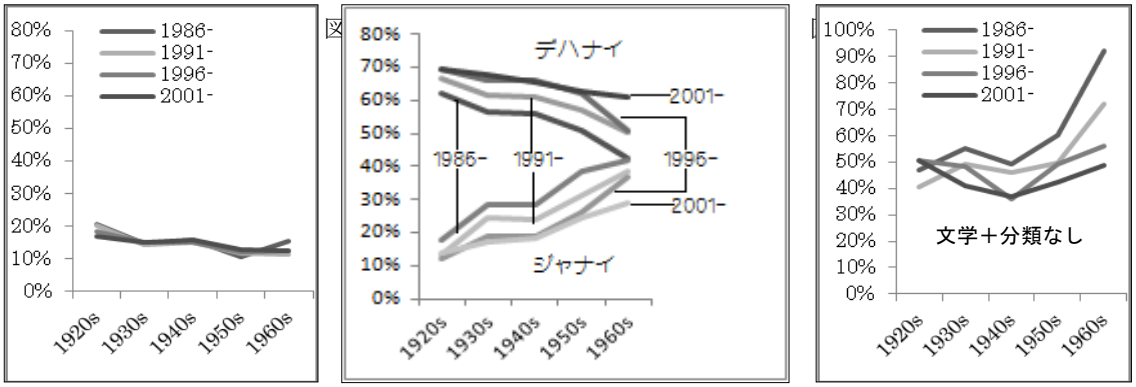


図 10 は、デナイだけを取り出したグラフである。図 10 ではグラフがほぼ一直線にまとまることから、デナイには調査年効果がほぼ観察されない。デナイの使用割合が緩やかに減少していく要因は、生年代による世代効果によるものと考えられる。図 11 はデハナイとジャナイの割合のみを示したグラフである。これを見ると互いががほぼ相補的に推移していることが観察される。1960s を除くと、1986-1990 年、1991-1995 年、1996-2000 年の 2 つの間隔は全ての生年代でほぼ等しくなっている。また 1996-2000 年と 2001-2005 年は全ての年代で重なっている。これは 1920s～1950s の生年代にほぼ同一の効果が及んでいることを示唆している。調査年効果には、時代効果と加齢効果があるが、加齢効果は各生年代で異なっていると考えられるため、このように均一に働く効果は時代効果である可能性が高い。1986-1990 年、1991-1995 年の各 5 年間は、ジャナイに変わってデハナイが使われやすくなる傾向が見られるが、1996-2000 年の 5 年間でこの傾向が見えなくなる。

このような傾向にジャンル効果が関わっているかどうかを確認するために、５年区切りの出版年別に文学＋分類なしのジャンル割合を調査したのが図 12 である。これと図 11 を比較すると、図 11 にジャンル効果が一定程度及んでいるように思われるが、図 11 の均等な傾向は図 12 のジャンル割合では説明できない。このため図 11 の傾向はジャンル効果以外に時代効果が働いていることを示唆していると考えられる。

なお、1960s は 1996-2000 年の 5 年間でジャナイの割合が特異に高くなる。これはこの生年代のこの区間だけに観察される現象である。このためこの区間の用例を個別に確認した。すると、文学や分類なし以外のジャンルでも会話が多用されている傾向が見られた。図 12

ではこの区間の文学＋分類なしのジャンル割合が 56%しかなく、1986-1995 年に比べて低いように思われる。しかし 1 年ごとに調査すると、例えば 1996 年では文学＋分類なしのジャンル割合が 48%しかないのに、ジャナイの割合は 49%になっているなどの現象が観察された。1996 年の場合、社会科学に分類されている『中国てなもんや商社』、歴史に分類されている『超貧乏旅』、美術・芸術に分類されている『リチャードレスター』などで会話が多用され、ジャナイが多く出現する。1960s で 1996-2000 年の区間が特異なふるまいを見せる原因には、ジャナイが多用されるサンプルが集積されたことが考えられる。

以上、調査年効果の調査から 1920s～1960s の生年代では、時代効果が観察され、1986-1995 年の 10 年間ではデハナイが一定の割合で増加する傾向が見られるが、1996 年からはその傾向が見えなくなることを述べた。

４．まとめ

先行研究では、生年代別分布を見かけ時間で捉える立場と年齢で捉える立場があった。「図書館書籍」では、およそ 1910s を境に年齢の基準が変わるため、生年代分布を年齢で捉えることは難しい。これを見かけ時間で捉えた場合にも 1910s を境に調査年効果が変化すると考えられる。本稿ではケーススタディを行って「図書館書籍」の生年代分布を見かけ時間で捉えた場合の問題点を考察した。

その結果、生年代分布に作用する因子として、ジャンル効果と時代効果が観察された。ジャンル効果とは、コーパスが潜在的に持っている分布の偏りのため、特定の生年代にあるジャンルが偏り、そのジャンルに現れやすい言語の頻度が高くなる効果のことである。言語の変化は書き言葉より先に会話に現れやすいため、会話の描写が多い文学の比率が高まると、その年代であたかも言語変化が進行しているかのように見える。

コピュラ否定形における「図書館書籍」の生年代別分布は、ジャンル効果などの影響により、無意味な内容を表している。コーパスに潜在する分布の偏りを補正することで、生年代別分布ははじめて意味を持つと言えるだろう。

参考文献

佐野真一郎(2012)「『日本語話し言葉コーパス』を用いた「全然」の変化の詳細化」『第 1 回コーパス日本語学ワークショップ予稿集』pp. 33-42.

日比谷潤子編著(2012)『はじめて学ぶ社会言語学－言葉のバリエーションを考える 14 章－』ミネルヴァ書房

前川喜久雄(2006)「第 1 章　概説」『日本語話し言葉コーパスの構築法』　pp. 1-18.

横山詔一(2011)「言語変化は経年調査データから予測可能か？」『国研プロジェクトプレビュー』№6　pp. 27-37.

前川喜久雄(2012)「「形容詞＋です」述語の生起要因についての準備的考察」『第 1 回コーパス日本語学ワークショップ予稿集』pp. 211-220.

丸山岳彦(2012)「大規模コーパスの利用とメタデータの役割」『第 1 回コーパス日本語学ワークショップ予稿集』pp. 203-210.

関連 URL

『中納言』BCCWJ 短単位語数　https://maro.ninjal.ac.jp/wiki/index.php?BCCWJ 短単位語数

『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版
http://www.ninjal.ac.jp/ corpus　_center/bccwj/doc.html

『第 1 回コーパス日本語学ワークショップ予稿集』
http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_web.pdf

現代日本語書き言葉における非外来語のカタカナ表記事情

柏野 和佳子*（国立国語研究所 言語資源研究系）
中村 壮範（国立国語研究所 コーパス開発センター）

Frequency of Katakana Representation for Japanese Non-loan Words as Observed in the BCCWJ Corpus

Wakako Kashino （Dept. Corpus Studies, NINJAL）
Takenori Nakamura （Center for Corpus Development, NINJAL）

１．はじめに

「カタカナ語」といえば，主として外来語をさす。しかしながら，非外来語である，和語や漢語にカタカナ表記が用いられることは，少ないことではない。

これまで，新聞や若者雑誌を中心に，非外来語である，和語や漢語のカタカナ表記が増加する傾向にあることが指摘されてきた。佐藤(1991)は若者を対象にカタカナ使用についてのアンケート調査を行った。中山(1998)と成田・榊原(2004)は新聞記事の表記を調査した。堀尾・則松(2005)と則松・堀尾(2006)は，若者雑誌の表記を調査した。堀江(2001)は，少し広範囲に，小説，マンガ，雑誌，新聞記事の表記を調査した。

また，柏野・奥村(2012)は，『現代日本語書き言葉均衡コーパス(Balanced Corpus of Contemporary Written Japanese; 略称 BCCWJ)』(前川 2013)に収録される書籍テキストデータを用いて，書籍における非外来語のカタカナ表記の実態を調査した。その結果，雑誌や新聞同様にカタカナ表記率の高い語がある一方，ひらがな表記率や漢字表記率が高い語もあり，表記の実態は，新聞，雑誌，書籍といった媒体で差があることを明らかにした。

非外来語のカタカナ表記については，教育的な関心も高く，高原(2012)，茂木(2012)，黎(2012)らが取り上げ，調査や分析を行っている。

このように，非外来語のカタカナ表記に関する先行研究はいくつもある。しかしながら，そのほとんどは調査対象語や着目語の選定が主観的に行われている。そこで，本研究は，調査対象語をコーパス分析によって客観的に選ぶことで，現代日本語書き言葉全般における，非外来語のカタカナ表記の実態を明らかにすることを目的にする。その方法は，現代語の均衡コーパスである，BCCWJ の全収録語の自動解析から得られる，同一見出し語の表記別の頻度情報を用いるというものである。本研究の一つ目の成果として，本発表では，カタカナ表記頻度の多い非外来語の語彙表を報告する。さらに，カタカナ頻度の多い語を，カタカナ表記率，ひらがな表記率，漢字表記率，の三つの観点により捉え直して得られた語彙表を報告する。

２．非外来語のカタカナ表記の実態調査

２.１　カタカナ表記頻度の多い非外来語の語彙表の作成

『現代日本語書き言葉均衡コーパス(BCCWJ)』の DVD 版(国立国語研究所 2012 年)の短単位データを用いた。短単位の全収録語数は，約 1 億 5 百万語である。サブコーパス別の収録語数は，次の表 1 のとおりである。

BCCWJ の短単位データの作成は自動形態素解析で行われている（「コア」と呼ばれる一部には人手修正あり）。形態素解析器に「MeCab[1]」，解析用辞書に「UniDic[2]」が使用されている。この BCCWJ の形態論情報付与に使用されている UniDic では，表記が異なっても同じ語であれば一つの見出し語にまとめられている。よって，解析結果より，見出し語単

* waka@ninjal.ac.jp

1 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

2 http://sourceforge.jp/projects/unidic/

表 1 サブコーパス別収録語数（短単位）

レジスター名	略表記	概要	語数	割合
出版・書籍	PB	出版実態サブコーパス・書籍	28552283	27.22%
出版・雑誌	PM	出版実態サブコーパス・雑誌	4444492	4.24%
出版・新聞	PN	出版実態サブコーパス・新聞	1370233	1.31%
図書館・書籍	LB	図書館実態サブコーパス・書籍	30377866	28.96%
特定目的・ベストセラー	OB	特定目的サブコーパス・ベストセラー	3742261	3.57%
特定目的・白書	OW	特定目的サブコーパス・白書	4882812	4.65%
特定目的・広報紙	OP	特定目的サブコーパス・広報紙	3755161	3.58%
特定目的・法律	OL	特定目的サブコーパス・法律	1079146	1.03%
特定目的・国会会議録	OM	特定目的サブコーパス・国会会議録	5102469	4.86%
特定目的・教科書	OT	特定目的サブコーパス・教科書	928448	0.88%
特定目的・韻文	OV	特定目的サブコーパス・韻文	225273	0.21%
特定目的・知恵袋	OC	特定目的サブコーパス・Yahoo!知恵袋	10256877	9.78%
特定目的・ブログ	OY	特定目的サブコーパス・Yahoo!ブログ	10194143	9.72%
合計			104911464	100.00%

表 2 カタカナ表記頻度上位 100 語の語彙表

番号	見出し	読み	カタカナ		頻度総数
			頻度	率	
1	物	モノ	3798	1.3%	285804
2	俺	オレ	3455	11.4%	30265
3	馬鹿	バカ	3179	40.3%	7890
4	本当	ホントウ	2383	6.8%	35043
5	僕	ボク	2282	4.8%	47657
6	良い	ヨイ	2198	1.1%	198994
7	癌	ガン	2188	32.9%	6652
8	ねた	ネタ	2037	97.5%	2089
9	嫌	イヤ	1991	15.6%	12741
10	塵	ゴミ	1918	40.9%	4690
11	車	クルマ	1856	8.6%	21623
12	人	ヒト	1828	1.2%	156890
13	子	コ	1719	5.8%	29535
14	こつ	コツ	1643	91.1%	1803
15	勧める	ススメル	1582	13.5%	11683
16	此れ	コレ	1571	0.8%	199500
17	薔薇	バラ	1534	72.5%	2116
18	蛋白	タンパク	1528	48.3%	3164
19	奴	ヤツ	1356	11.1%	12186
20	格好	カッコウ	1345	20.9%	6432
21	漫画	マンガ	1329	37.9%	3510
22	猫	ネコ	1301	18.5%	7037
23	丸	マル	1287	18.6%	6919
24	奇麗	キレイ	1262	10.8%	11715
25	林檎	リンゴ	1186	49.3%	2408
26	苛々	イライラ	1125	68.6%	1639
27	私	ワタシ	1108	2.5%	45106
28	びっくり	ビックリ	1106	23.1%	4790
29	葉書	ハガキ	1097	36.3%	3025
30	猿	サル	1071	46.8%	2288
31	事	コト	996	0.1%	740923
32	嘘	ウソ	993	17.9%	5537
33	金	カネ	961	4.9%	19742
34	どきどき	ドキドキ	942	79.7%	1182
35	雌	メス	936	47.2%	1985
36	まじ	マジ	925	65.5%	1412
37	阿呆	アホウ	901	65.0%	1386
38	鼠	ネズミ	888	55.9%	1588
39	台詞	セリフ	866	46.5%	1861
40	巣	ス	856	34.8%	2461
41	携帯	ケイタイ	852	10.1%	8408
42	家	ウチ	851	59.6%	1429
43	木	キ	826	5.8%	14356
44	葡萄	ブドウ	824	44.4%	1855
45	餌	エサ	816	32.2%	2531
46	はい	ハイ	809	10.5%	7697
47	烏	カラス	802	59.5%	1348
48	烏賊	イカ	801	63.5%	1261
49	雄	オス	792	41.4%	1913
50	徽	カビ	792	87.0%	910

番号	見出し	読み	カタカナ		頻度総数
			頻度	率	
51	蝦	エビ	784	47.9%	1636
52	兎	ウサギ	779	51.4%	1517
53	喧嘩	ケンカ	772	23.9%	3229
54	ぱちんこ	パチンコ	771	97.3%	792
55	散らし	チラシ	769	91.0%	845
56	親父	オヤジ	746	23.6%	3155
57	熊	クマ	746	40.9%	1824
58	はっきり	ハッキリ	744	7.1%	10452
59	餓鬼	ガキ	742	71.8%	1033
60	ばらばら	バラバラ	738	57.2%	1291
61	怪我	ケガ	731	25.5%	2864
62	眼鏡	メガネ	730	30.9%	2363
63	茸	キノコ	692	55.8%	1241
64	ぴん	ピン	685	72.2%	949
65	ぱっ	パッ	682	43.8%	1557
66	壺	ツボ	675	43.5%	1550
67	ちび	チビ	670	68.0%	986
68	鸚鵡	オウム	668	86.9%	769
69	貴方	アナタ	667	1.4%	48127
70	伯父	オジ	666	10.8%	6191
71	歯	ハ	657	11.5%	5706
72	ぎりぎり	ギリギリ	653	45.9%	1422
73	訳	ワケ	650	0.9%	73097
74	草	ソウ	646	41.1%	1572
75	茄子	ナス	640	48.6%	1317
76	蛸	タコ	630	58.4%	1079
77	蟹	カニ	629	56.3%	1117
78	葱	ネギ	626	45.1%	1389
79	にこにこ	ニコニコ	622	64.0%	972
80	虎	トラ	622	38.4%	1621
81	蝉	セミ	621	65.3%	951
82	鍵	カギ	620	16.1%	3846
83	癖	クセ	619	15.8%	3913
84	杉	スギ	614	44.5%	1380
85	内	ウチ	611	1.4%	45175
86	犬	イヌ	594	7.0%	8495
87	彼れ	アレ	594	4.8%	12404
88	桜	サクラ	589	13.1%	4506
89	すっきり	スッキリ	585	30.4%	1922
90	惚け	ボケ	585	74.0%	791
91	蛙	カエル	584	79.2%	737
92	大蒜	ニンニク	578	39.1%	1479
93	有り	アリ	572	11.0%	5177
94	乗り	ノリ	571	43.6%	1311
95	手	テ	558	1.2%	47213
96	蝶	チョウ	555	41.0%	1353
97	此処	ココ	553	1.0%	56797
98	行く	イク	549	0.2%	220638
99	段	ダン	548	11.4%	4804
100	勘	カン	546	44.9%	1216

位の表記別の頻度情報を自動的に得ることができる。その頻度情報をもとに，カタカナ表記頻度の多い語を抽出した。その際，数詞，助詞，接頭辞，接尾辞は対象外とした。もっとも頻度の多かったのは「モノ（物）」で，その頻度は 3,798 である。298 位は 4 語あり，その頻度は 246 である。頻度 246～3,798 の 301 語が調査範囲として扱いやすい適当なサイズであると判断し，それら 301 語を今回の調査対象語として選定することとした。そのうち，上位 100 語で作成した語彙表を表 2 に示す。

表 2 を見ると，頻度総数に対する，カタカナ表記率が語によって大きく異なることが目につく。たとえば，1 位の「モノ（物）」は，ひらがな表記，漢字表記とをあわせた頻度総数が 285,804 と多く，そのカタカナ表記率はわずか 1.3%に過ぎない。

２.２　表記の比率別にみる語彙表の作成

調査対象の 301 語を，カタカナ表記率，ひらがな表記率，漢字表記率の高い順に並び替え，それぞれの上位 50 語で作成した各語彙表を順に，表 3～5 に示す。

表 3 カタカナ表記率上位 50 語の語彙表

番号	見出し	読み	カタカナ		ひらがな		漢字		頻度総数
			頻度	率	頻度	率	頻度	率	
1	もてる	モテル	285	99.7%	1	0.3%	0	0.0%	286
2	女	メ	419	99.1%	0	0.0%	4	0.9%	423
3	ねたばれ	ネタバレ	263	98.5%	4	1.5%	0	0.0%	267
4	蛇	ジャ	320	98.2%	0	0.0%	6	1.8%	326
5	壁蝨	ダニ	301	97.7%	6	1.9%	1	0.3%	308
6	ねた	ネタ	2037	97.5%	52	2.5%	0	0.0%	2089
7	海豚	イルカ	430	97.5%	5	1.1%	6	1.4%	441
8	ぱちんこ	パチンコ	771	97.3%	21	2.7%	0	0.0%	792
9	ぶす	ブス	349	91.6%	32	8.4%	0	0.0%	381
10	こつ	コツ	1643	91.1%	160	8.9%	0	0.0%	1803
11	散らし	チラシ	769	91.0%	70	8.3%	6	0.7%	845
12	軟派	ナンパ	250	90.9%	0	0.0%	25	9.1%	275
13	びら	ビラ	285	90.5%	30	9.5%	0	0.0%	315
14	ばりばり	バリバリ	380	88.6%	49	11.4%	0	0.0%	429
15	こく	コク	401	87.4%	58	12.6%	0	0.0%	459
16	劫	コウ	406	87.3%	0	0.0%	59	12.7%	465
17	かりかり	カリカリ	252	87.2%	37	12.8%	0	0.0%	289
18	黴	カビ	792	87.0%	55	6.0%	63	6.9%	910
19	鸚鵡	オウム	668	86.9%	46	6.0%	55	7.2%	769
20	鸚哥	インコ	265	86.6%	19	6.2%	22	7.2%	306
21	焜炉	コンロ	290	85.0%	46	13.5%	5	1.5%	341
22	でぶ	デブ	394	82.8%	82	17.2%	0	0.0%	476
23	ぽち	ポチ	498	81.8%	111	18.2%	0	0.0%	609
24	蛾	ガ	536	81.0%	10	1.5%	116	17.5%	662
25	がんがん	ガンガン	365	80.8%	87	19.2%	0	0.0%	452
26	がらがら	ガラガラ	282	80.6%	68	19.4%	0	0.0%	350
27	鍍金	メッキ	252	80.5%	61	19.5%	0	0.0%	313
28	齣	コマ	426	80.2%	65	12.2%	40	7.5%	531
29	どきどき	ドキドキ	942	79.7%	240	20.3%	0	0.0%	1182
30	どん	ドン	517	79.5%	133	20.5%	0	0.0%	650
31	蛙	カエル	584	79.2%	99	13.4%	54	7.3%	737
32	ぶな	ブナ	290	78.6%	71	19.2%	8	2.2%	369
33	片仮名	カタカナ	445	78.2%	9	1.6%	115	20.2%	569
34	家鴨	アヒル	268	77.5%	52	15.0%	26	7.5%	346
35	鮑	アワビ	318	77.2%	34	8.3%	60	14.6%	412
36	ちん	チン	276	77.1%	82	22.9%	0	0.0%	358
37	蜻蛉	トンボ	397	76.8%	104	20.1%	16	3.1%	517
38	麒麟	キリン	400	75.5%	26	4.9%	104	19.6%	530
39	とん	トン	246	75.5%	80	24.5%	0	0.0%	326
40	とんとん	トントン	246	74.5%	84	25.5%	0	0.0%	330
41	わん	ワン	311	74.2%	108	25.8%	0	0.0%	419
42	栗鼠	リス	416	74.0%	106	18.9%	40	7.1%	562
43	惚け	ボケ	585	74.0%	158	20.0%	48	6.1%	791
44	上	カミ	452	73.0%	51	8.2%	116	18.7%	619
45	けち	ケチ	397	72.7%	138	25.3%	11	2.0%	546
46	薔薇	バラ	1534	72.5%	88	4.2%	494	23.3%	2116
47	ぴん	ピン	685	72.2%	264	27.8%	0	0.0%	949
48	蜥蜴	トカゲ	250	72.0%	29	8.4%	68	19.6%	347
49	餓鬼	ガキ	742	71.8%	82	7.9%	209	20.2%	1033
50	括弧	カッコ	372	71.0%	15	2.9%	137	26.1%	524

表 3 に示すカタカナ表記率上位 50 語は，その比率が 71.0%以上といずれも高く，主にカタカナ表記される語であることが確認できよう。ただし，たとえば，16 位の「コウ（劫）」のように，あまり見慣れない語も入っている。実際の用例を確認してみると，囲碁用語や，仏教で長い時間を表す語としての「劫」のカタカナ表記の使用例のほか，音を表すだけの「コウ」が「劫」と誤解析された結果もまじっているため，頻度や比率が高くなってしまっていた。44 位の「カミ（上）」もわかりにくい例であるが，これは，「カミサン（上様）」（＝主婦を呼ぶのに用いる）のカタカナ表記の使用例であった。ただし，その中に，「神」のカタカナ表記を「上」と誤解析したものも含まれていた。このように，自動解析にはある程度の誤解析は含まれはするが，表記実態の傾向は自動解析結果より十分にみてとれる。

表 4 は，ひらがな表記率上位 50 語の表である。上位を占める半分以上の語が，ひらがな表記が主な語である。それらは，カタカナ表記率は低いが，使用総数が多いためにカタカナ表記の頻度が多くなっている語であることがわかる。しかし，カタカナ表記率が 20%を超える語は，ひらがな使用が主であるといっても，カタカナ使用の目立つ語の一群であるとの印象を受けるものになっている。

表 4 ひらがな表記率上位 50 語の語彙表

番号	見出し	読み	カタカナ		ひらがな		漢字		頻度総数
			頻度	率	頻度	率	頻度	率	
1	為る	スル	395	0.0%	2563160	100.0%	305	0.0%	2563860
2	此れ	コレ	1571	0.8%	195971	98.2%	1958	1.0%	199500
3	一寸	チョット	344	1.2%	29147	98.1%	206	0.7%	29697
4	此処	ココ	553	1.0%	55682	98.0%	562	1.0%	56797
5	此方	コチラ	322	1.8%	17508	97.5%	126	0.7%	17956
6	私	ワタシ	1108	2.5%	43951	97.4%	47	0.1%	45106
7	あっ	アッ	406	4.2%	9256	95.8%	1	0.0%	9663
8	彼れ	アレ	594	4.8%	11779	95.0%	31	0.2%	12404
9	事	コト	996	0.1%	702885	94.9%	37042	5.0%	740923
10	うん	ウン	273	5.2%	4995	94.8%	0	0.0%	5268
11	物	モノ	3798	1.3%	268202	93.8%	13804	4.8%	285804
12	訳	ワケ	650	0.9%	68478	93.7%	3969	5.4%	73097
13	筈	ハズ	271	1.1%	23796	93.4%	1408	5.5%	25475
14	貴方	アナタ	667	1.4%	44900	93.3%	2560	5.3%	48127
15	はっきり	ハッキリ	744	7.1%	9708	92.9%	0	0.0%	10452
16	所	トコロ	427	0.4%	100940	90.2%	10506	9.4%	111873
17	えっ	エッ	448	9.8%	4116	90.2%	0	0.0%	4564
18	内	ウチ	611	1.4%	40576	89.8%	3988	8.8%	45175
19	はい	ハイ	809	10.5%	6888	89.5%	0	0.0%	7697
20	わし	ワシ	290	6.0%	4196	87.4%	317	6.6%	4803
21	へえ	ヘエ	301	15.3%	1664	84.7%	0	0.0%	1965
22	只	タダ	485	11.0%	3694	83.7%	235	5.3%	4414
23	きつい	キツイ	373	16.7%	1854	83.3%	0	0.0%	2227
24	有り	アリ	572	11.0%	4307	83.2%	298	5.8%	5177
25	御免	ゴメン	436	10.0%	3555	81.9%	349	8.0%	4340
26	でかい	デカイ	247	18.7%	1068	80.9%	5	0.4%	1320
27	良い	ヨイ	2198	1.1%	160909	80.9%	35887	18.0%	198994
28	凄い	スゴイ	454	3.1%	11569	78.4%	2724	18.5%	14747
29	さっ	サッ	405	21.6%	1445	77.1%	24	1.3%	1874
30	わあ	ワア	401	24.0%	1272	76.0%	0	0.0%	1673
31	ぴったり	ピッタリ	541	24.0%	1716	76.0%	0	0.0%	2257
32	びっくり	ビックリ	1106	23.1%	3600	75.2%	84	1.8%	4790
33	ぐっ	グッ	306	27.7%	796	72.2%	1	0.1%	1103
34	すっきり	スッキリ	585	30.4%	1337	69.6%	0	0.0%	1922
35	ちら	チラ	340	33.5%	675	66.5%	0	0.0%	1015
36	御洒落	オシャレ	322	17.3%	1216	65.2%	327	17.5%	1865
37	菠薐	ホウレン	255	35.0%	465	63.9%	8	1.1%	728
38	やばい	ヤバイ	437	36.1%	772	63.9%	0	0.0%	1209
39	胡麻	ゴマ	425	19.9%	1358	63.6%	353	16.5%	2136
40	餡	アン	311	25.3%	779	63.3%	140	11.4%	1230
41	行く	イク	549	0.2%	134413	60.9%	85676	38.8%	220638
42	増し	マシ	496	32.2%	910	59.0%	136	8.8%	1542
43	ずれ	ズレ	288	41.1%	412	58.9%	0	0.0%	700
44	奇麗	キレイ	1262	10.8%	6895	58.9%	3558	30.4%	11715
45	伯母	オバ	542	11.8%	2682	58.5%	1358	29.6%	4582
46	大蒜	ニンニク	578	39.1%	854	57.7%	47	3.2%	1479
47	己等	オイラ	477	42.6%	641	57.2%	2	0.2%	1120
48	ばれる	バレル	516	42.9%	682	56.7%	5	0.4%	1203
49	ぱっ	パッ	682	43.8%	875	56.2%	0	0.0%	1557
50	にやり	ニヤリ	283	44.3%	356	55.7%	0	0.0%	639

表 5 は，漢字表記比率上位 50 語の表である。上位 50 語のうち，カタカナ表記率が 20%を超える語は少ない。しかし，20%以下でも，カタカナ表記頻度の多い語を中心に，カタカナ表記が多いと感じる語がこの表にも多くあがっている。たとえば，22 位の「ケイタイ（携帯)」は，そのカタカナ表記の内訳が，「ケータイ(797)/ケイタイ(54)/ケータィ(1)」であるのだが，「ケータイ」という使用例がもはや標準であるかのような印象を与える語である。しかし，それはあくまでも，「携帯電話」をさす「ケータイ」であり，「携帯電話」以外にも，「携帯する」「携帯手荷物」といった用法をもつ「ケイタイ（携帯)」という語は，漢字表記が圧倒的に多いということである。また，この表の中では，従来の先行研究ではあがってこなかった 34 位の「ソン（損)」に注目したい。漢字表記率が高い一方，カタカナ表記率も 20.5%あり，カタカナ表記頻度が 382 ある。よって，この語も，カタカナ使用の目立つ実態をもつ語の一つであると言えるだろう。

表 5 漢字表記率上位 50 語の語彙表

番号	見出し	読み	カタカナ		ひらがな		漢字		頻度総数
			頻度	率	頻度	率	頻度	率	
1	手	テ	558	1.2%	32	0.1%	46623	98.8%	47213
2	日	ヒ	388	0.7%	589	1.0%	58221	98.3%	59198
3	機	キ	421	1.9%	190	0.9%	21124	97.2%	21735
4	山	ヤマ	277	1.5%	254	1.4%	18094	97.1%	18625
5	人	ヒト	1828	1.2%	3932	2.5%	151130	96.3%	156890
6	村	ムラ	360	2.8%	152	1.2%	12398	96.0%	12910
7	奥	オク	252	2.2%	219	1.9%	11159	96.0%	11630
8	彼	カレ	290	0.3%	4254	4.6%	87495	95.1%	92039
9	金	カネ	961	4.9%	15	0.1%	18766	95.1%	19742
10	茶	チャ	262	3.9%	94	1.4%	6279	94.6%	6635
11	薬	クスリ	247	4.1%	98	1.6%	5671	94.3%	6016
12	首	クビ	457	4.6%	119	1.2%	9412	94.2%	9988
13	子	コ	1719	5.8%	233	0.8%	27583	93.4%	29535
14	個	コ	529	4.4%	265	2.2%	11203	93.4%	11997
15	黒	クロ	470	5.0%	174	1.8%	8802	93.2%	9446
16	無理	ムリ	400	3.9%	365	3.6%	9511	92.6%	10276
17	裏	ウラ	417	7.0%	33	0.6%	5529	92.5%	5979
18	犬	イヌ	594	7.0%	86	1.0%	7815	92.0%	8495
19	鳥	トリ	377	8.5%	2	0.0%	4047	91.4%	4426
20	車	クルマ	1856	8.6%	71	0.3%	19696	91.1%	21623
21	切れる	キレル	276	8.9%	6	0.2%	2822	90.9%	3104
22	携帯	ケイタイ	852	10.1%	2	0.0%	7554	89.8%	8408
23	傷	キズ	258	7.8%	93	2.8%	2945	89.4%	3296
24	木	キ	826	5.8%	753	5.2%	12777	89.0%	14356
25	楽	ラク	419	8.5%	207	4.2%	4277	87.2%	4903
26	段	ダン	548	11.4%	72	1.5%	4184	87.1%	4804
27	歯	ハ	657	11.5%	86	1.5%	4963	87.0%	5706
28	体	カラダ	381	1.6%	2690	11.5%	20379	86.9%	23450
29	松	マツ	252	11.3%	84	3.8%	1897	85.0%	2233
30	虫	ムシ	429	12.8%	94	2.8%	2829	84.4%	3352
31	大根	ダイコン	271	12.6%	104	4.8%	1776	82.6%	2151
32	変	ヘン	504	8.7%	547	9.4%	4741	81.9%	5792
33	皿	サラ	502	15.3%	107	3.3%	2669	81.4%	3278
34	損	ソン	382	20.5%	0	0.0%	1480	79.5%	1862
35	鍵	カギ	620	16.1%	203	5.3%	3023	78.6%	3846
36	椅子	イス	539	8.8%	866	14.1%	4749	77.2%	6154
37	誰	ダレ	375	1.0%	8112	22.2%	28105	76.8%	36592
38	猫	ネコ	1301	18.5%	380	5.4%	5356	76.1%	7037
39	鶏	ニワトリ	381	17.8%	146	6.8%	1616	75.4%	2143
40	暇	ヒマ	373	12.5%	362	12.2%	2242	75.3%	2977
41	無駄	ムダ	480	12.9%	472	12.7%	2762	74.4%	3714
42	下手	ヘタ	288	12.1%	330	13.9%	1758	74.0%	2376
43	桜	サクラ	589	13.1%	608	13.5%	3309	73.4%	4506
44	豚	ブタ	426	22.8%	75	4.0%	1365	73.2%	1866
45	米	コメ	534	25.7%	62	3.0%	1483	71.3%	2079
46	本当	ホントウ	2383	6.8%	7724	22.0%	24936	71.2%	35043
47	麺	メン	310	15.8%	302	15.4%	1348	68.8%	1960
48	嘘	ウソ	993	17.9%	799	14.4%	3745	67.6%	5537
49	蛇	ヘビ	392	27.2%	82	5.7%	969	67.2%	1443
50	稲	イネ	257	31.3%	21	2.6%	544	66.2%	822

３．今後の課題

BCCWJ 全体の自動解析結果を用いた非外来語のカタカナ表記の調査分析は着手したば

かりである。先に，柏野・奥村(2012)においては，その時の調査対象語に対し，カタカナ表記になりやすい語のタイプ（「表外漢字」「表外音訓」「表外熟字訓」などを含む語，動植物名の語，擬音語・擬態語，第一義でない意味の語，特別な意味を加味する用法の語）やケース（強調用法のある場合，同字語を避けたい場合，音を明示したい場合）別に使用実態を分析した。BCCWJ で得られる今回の調査対象語に対しても，そうしたタイプやケース別の分析を進める予定でいる。また，表 1 に示したとおり，BCCWJ は複数のレジスタから構成されており，レジスタ別の分析が可能である。試行的に行っている予備調査では，たとえば，教科書や白書ではカタカナ表記が避けられ，逆に，新聞，雑誌，知恵袋，ブログでは，カタカナ表記が好まれる，という傾向がみてとれている。今後，レジスタ別の使用傾向の差異についても明らかにしていきたい。

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト基幹型「コーパス日本語学の創成」による補助，並びに，文部科学省科学研究費補助金基盤研究(C)「コーパス分析に基づく辞書の位相情報の精緻化」（課題番号: 23520572）の助成を受けたものです。

文　献

柏野和佳子・奥村学(2012)「和語や漢語のカタカナ表記:『現代日本語書き言葉均衡コーパス』の書籍における使用実態」『計量国語学』28(4), pp.153-161.

佐藤栄作(1991)「若者のカタカナ使用と外来語表記―語種意識から―」『日本語学』10-7, pp.76-88, 明治書院.

高原真理(2012)「日本語学習辞書における表記情報の課題と今後の取り組み」(2012 年度マレーシア国際研究集会「日本語学習辞書開発の支援を考える」口頭発表資料 http://jishokaken.sakura.ne.jp/doc/Malaysia/M3.pdf）.

中山恵利子(1998)「非外来語の片仮名表記」『日本語教育』96,pp.61-72, 日本語教育学会.

中山恵利子・桐生りか・山口昌也(2007) 第 3 部第 3 章「新聞に見る基幹外来語」『国立国語研究所報告 126　公共媒体の外来語　―「外来語」言い換え提案を支える調査研究―』.

成田徹男・榊原浩之(2004)「現代日本語の表記体系と表記戦略―カタカナの使い方の変化―」『人間文化研究』2,pp.41-55, 名古屋市立大学.

則松智子・堀尾香代子(2006)「若者雑誌における常用漢字のカタカナ表記化―意味分析の観点から」『北九州市立大学文学部紀要』72,pp.19-32.

堀江紫野(2001)「カタカナ表記の研究―非外来語系を中心に」『国文目白』40,pp.16-24, 日本女子大学.

堀尾香代子・則松智子(2005)「若者雑誌におけるカタカナ表記とその慣用化をめぐって」『北九州市立大学文学部紀要』69,pp.35-44.

前川喜久雄（編）(2013)『コーパス入門』（講座 日本語コーパス）朝倉書店.

茂木俊伸「第 5 課「チョー恥ずかしかったヨ！」なカタカナの不思議」, 定延利之（編著）『私たちの日本語』朝倉書店, pp.47-57.

黎婉珊(2012)「非外来語のカタカナ表記―「中納言」検索および日本語教科書についての調査から―」（日本語教育国際研究大会名古屋 2012 年ポスター発表）A494.

関連 URL

国立国語研究所コーパス開発センター　http://www.ninjal.ac.jp/corpus_center/

言語資料としての「判決文」の分析にまつわる問題点

矢野 信（株式会社法学館法教育研究所）†

Problems in Corpus-based Studies of Judgment Documents

Makoto Yano (Japan Research Institute of Law Related Education, HOUGAKUKAN CO.,LTD.)

１．はじめに[1]

「判決文」は，今でも分かりにくい日本語の代表格のように認識されていると考えられるが[2]，その“分かりにくさ”には，複数の次元における種々の要因が関係していると考えられる。

しかしながら，「判決文」の日本語の分かりにくさについて言語学者や裁判実務家によって従来述べられてきたのは，その多くが経験的なものであり，定量的な分析を試みたものは少ない。

本発表では，「判決文」の言葉について定量的な分析を行なっていく前提として，法律実務家・法律教育者側の視点を踏まえつつ，①主としてコーパスの手法によって「判決文」の言語を分析することの位置付けと，②「判決文」についてコーパスを具体的に考えていく場面での留意点について整理する。加えて，③「判決文」のデータの分析から得られる情報の例を挙げることとする。

２．「判決文」の言語に関係する先行研究[3]

２．１　自然言語処理の観点から

阪野(2005)，阪野(2006)を初めとする研究では，自然言語処理（テキスト自動要約）の観点から，最高裁判決の原文データから構築した判例コーパスにより，いわゆる判例要旨に相当する部分を抽出する手法の提案がなされていた。

ここでは，判例（判決）の文体・構造についての（法学の）既存の知見を用いて，判例要旨の抽出作業の自動化が図られている。このため，判決の言語自体の分析を目的としてコーパスを用いるという本稿とは目的を異にするといえる。

２．２　法言語学の観点から

一方，橋内(2012a)では，法言語学の観点から，裁判員制度施行前後の第一審刑事判決の「罪となるべき事実」について，文体情報・特徴語の抽出を行い，“市民に分かりやすく”という裁判員制度導入の目的との関係での検証に相当する考察が行われていた。

これは，本稿の方向性とアプローチは共通しているといえるが，本稿がひとまず直接の目的とするのは，そのような考察の前提となる判決の言語（ことば）それ自体についての分析の段階である。

---

† klagegrund@gmail.com

[1] 根拠として法令の条文を挙げる場合，法令名＋条文番号のみを掲げ，書物あるいはインターネット上の法令集を文献として掲げることはしない（なお，インターネット上でアクセス可能な法令集を末尾の関連 URL に記載しておく。）。

[2] 古くは岩淵(1979)が有名であるが，比較的最近のものとして，大橋(2010)などがある。

[3] 電子データを扱うものに限る。この他にも，法律実務家・理論言語学者などによる「判決」の言葉についての指摘・言及は数多くある。

３.「判決文」の言語を分析することの位置付け

３.１　用語の定義

３.１.１　「判決文」の意義

「判決文」という語は，日本の現行法令中に存在しない[4] [5]。

本稿で分析・検討対象として取り上げる「判決文」は，直接的には，日本の通常裁判所[6]における判断の形式としての「判決」[7]の内容部分を言語の形で表現したものをいうとしておく。また，裁判所における公権的な法的判断[8]の形式には，判決のほかに「決定」「命令」等があるが[9]，それらを言語の形で表現したもの（「決定文」などという。）も，「判決文」の集合に含めることがある。

このうち，「内容部分を言語の形で表現した」というのは，法学の解説書などであまり触れられない。この部分をあえて書いたのは，「判決文」が作成されない「判決」があるからである（後述「調書判決」）。

３.１.２　類義語・関連語の整理

- 「判決書」というのは，「判決」の内容を記載した文書・書面のことをいう[10]（したがって，物体として存在するものである。）。なお，「決定」という判断形式に対応するのは「決定書」である。
- 「判例」とは，判決・決定などに示された法的な判断に着目した用語である。広義には「過去に下された裁判」一般を指すが，狭義には「それらに含まれる原則のうち，現在拘束力を持つもの」をいう（金子(2008)）。実務（特に裁判所）では，後者の観点から専ら最高裁判所の判決・決定を指して使われる。

  これに対して，「裁判例」「判決例」は，法律面や事実面における判断の事例（ひとつのケース）という意味合いで個々の判決を指す場合に用いられる。[11]

  これらはいずれも，「判決」で示される判断の内容・中味に着目した表現であることから，判決文それ自体の言語的性質の分析を目的とする本稿では，基本的に用いない。
- 「判例集」とは，「判例」を検索できるように年代別・内容別に整理して編集した書物（あるいはデータベース）のことである。後述のように，「判決文」に一般にアクセスできるのは「判例集」を通じてあることから，その性質を理解しておくことは本稿の分析目的との関係で重要な意味を持ってくる。

３.２　「判決文」の言語を分析することの意義

言語学的な関心の下，言語資料としての「判決文」を分析する意義を最大限広く考える前提として，「判決」の性質として，かなり抽象化した次の点で捉えておく。

事実を確定し，それルール（規範）に当てはめて，一定の判断を下す行為

このうち，「事実」とは，法律家が使う用語としての意味である。[12] また，ルール（規範）

---

[4]「法令データ提供システム」（総務省）の法令用語検索による。

[5] また，末川(1991)，金子(2008)などの法学の辞典にも，項目としては存在しない。

[6] 最高裁判所及びその下に属する各種の裁判所のことを指す（憲法 76 条 1 項）。なお，国会の下に「裁判官弾劾裁判所」があるが（憲法 64 条），これを含まない。

[7] 民事訴訟法 243 条，刑事訴訟法 329 条以下

[8] これを訴訟法の概念で「裁判」という（金子(2008)）。

[9] 民事訴訟規則 50 条，刑事訴訟 43 条

[10] 民事訴訟法 253 条，刑事訴訟規則 218 条

[11] なお，この 2 語はいずれも，末川(1991)，金子(2008)には項目としては存在しない。

[12] 少なくとも，日常用語とは異なり，真である（truth）という意味合いを全く含んでいない（この点は，

には「法律」をはじめとする「法令」が含まれるが，それらに限られない。[13]

このように広く捉えれば，「判決」と類似の作用は，およそ団体・組織のあるところにはすべて存在するものであるといえそうである。[14] 例えば，ある非行を理由として社員を懲戒にする場合の会社の判断，あらかじめ決めてある一定の資格を満たした者に対して入会を認めるという団体の判断，ルールブックに則ってストライク／ボールを決する野球の審判の判断などなど，すべて，上記の構造をとっている。[15]

そうすると，そのような意味での「判決」という作用を理解できることは，市民社会の構成員としてのリテラシーの一つであるといってよいかも知れない。[16] このことから，「判決文」の分析には，法律の専門家や専門家になるための教育にとって必要な知見を得ることにとどまらない面があると考えている。[17]

４．「判決文」コーパスを考える際の留意点

４.１　言語資料としての「判決文」データについて

前川(2013)では，現代的なコーパスに求められる要件として代表性，均衡性，真正性などの点や公開にまつわる問題点を挙げている。「判決文」のコーパスを考える際におけるそれらの問題点のいくつかについてまとめてみる。

４.１.１　著作権

著作権法 13 条 3 号は，「裁判所の判決，決定，命令及び審判並びに行政庁の裁決及び決定で裁判に準ずる手続により行われるもの」は著作権の目的とならない（著作権が発生しない）ことを定めている。このことは，言語資料として「判決文」を分析する場合における大きな利点であろう。

なお，判例集・判例雑誌には，しばしば，当事者（検察官・弁護士など）が書いた「上告理由」等が判決本文の後に掲載される。特に，出版される「最高裁判所判例集」には，上告理由等が必ず掲載されることになっている。この部分は，上記の著作権法の定義には当てはまらないことから，この部分を含めたコーパスを構築した場合には，著作権処理の問題が出てくることになる。

４.１.２　「判決」の母集団（コーパスの代表性の問題１）

まず，すべての「判決」において「判決文」が作成されるわけではないことに留意しておく必要がある。

民事訴訟においては，訴えの提起を受けた被告が事実を争わないなどの一定の場合に，

日弁連(2007)で指摘がなされていた。)。厳密に定義しようとすれば，例えば“法的判断の前提問題としてその存否が確定の対象となる歴史的な出来事”などという表現が考えられるが，このような定義は，末川(1991)，金子(2008)などの法学辞典や法学書等には記載されていない。

[13] この点については，矢野(2013)も参照。

[14] なお，最終的に裁判所で争うことが可能であるような各種の紛争解決手段における判断（国税不服審判所の審判，特許庁による審決，海難審判など）は，当然のこととして省略する。

[15] なお，あらかじめルールが共通に定められていない世界・集団においても，条理や過去の判断などを参考にしてルールを定立しながらその判断を行うことが可能である（英米法における判例法は，そのような判断の積み重ねとして生成されてきたものということができる。)。したがって，あらかじめ定められているルールの存在と，上記のような意味での「判決」作用の存在は，一方の存在が他方の前提となる関係にはないと考えられる。

[16] 法教育研究会(2004)では，法教育において取り扱うべき主たる内容の一つとして「ルールに基づいてどのように紛争を解決していくのか」の点が掲げられていた。

[17] 現時点でも，法言語教育の実践として，高校の国語科において判決文を題材とした授業が行う例がある（橋内(2012b)）。

判決書の作成を省略できる（民事訴訟法 254 条)。この場合は，裁判所書記官が，口頭弁論調書に，判決の主文・理由の要旨などを記載する。また，刑事訴訟においては，控訴・上告の申立てがなく，かつ，判決書の謄本の請求がない場合に，判決書の作成を省略できる（刑事訴訟規則 219 条)。この場合も，口頭弁論調書に判決主文などが記載される。このことから，これらの制度はいずれも「調書判決」と呼ばれる。

「司法統計」では，このような「調書判決」の件数は明らかにされていないが，判決件数のうちの一定の割合が「調書判決」によると考えられる。

民事・刑事いずれの場合も，裁判所の行為としての「判決」は存在するが，その内容を言語で表現した「判決文」は存在しないことになる。したがって，コーパスを構築して「判決文」の言語を分析するという本稿においては，取得の対象から外れることになる。

４.１.３　「判決書」の母集団（コーパスの代表性の問題２）

判決文の入手は公開された判決書によることになる。

民事・刑事とも，判決書は，原則として公開というのが建て前であるが[18]，それは，裁判所（民事の場合)・検察庁（刑事の場合）の庁内での閲覧が可能であるというにとどまる。また，刑事事件判決は，実際には，プライバシー保護などを理由に事件の関係者などでなければ閲覧の許可もされない場合がある（福島(1999)など)。

したがって，実際には，判決文の入手は，裁判所ホームページで公開されたデータや，出版された判例集・判例雑誌，商用データベースなどによることになる。

ここで，2011（平成 23）年に最高裁判所で処理がなされた民事事件・刑事事件の件数[19]，最高裁判所ホームページ[20]・商用データベース（Westlaw Japan[21]）に登載されている同じ期間の最高裁判決の本数を表 1 に掲げる。また，同年の地裁第一審事件の件数に関する同様のデータを表 2 に掲げる。

表 1　2011（平成 23）年の最高裁判所の事件件数等

	既済事件数	「最高裁判所判例集」の公開データ数	Westlaw Japan で入手可能なデータ数
民 事	6,654	76 (1.14%)	109 (1.64%)
刑 事	3,854	37 (0.96%)	37 (0.96%)

表 2　2011（平成 23）年の地裁第一審事件の件数等

	既済事件数	「下級裁判所判例集」の公開データ数	Westlaw Japan で入手可能なデータ数
民 事	774,183	67 (0.01%)	3512 (0.45%)
刑 事	80,886	35 (0.04%)	43 (0.05%)

これらの表からわかるように，現実的に一般にアクセス可能な判決文データは，全体の数%にも満たないのであるが，そのようにして公開・出版されるものは，ほぼ間違いなく，

[18] 民事訴訟法 91 条 1 項，刑事確定訴訟記録法 4 条
[19] 本稿執筆時点で公開されている最新の「司法統計」（年報）からのデータである。
[20] 最高裁判所ホームページ内の「最高裁判所判例集」
[21] 出典ウェストロー・ジャパン株式会社

法律実務家の観点から実務上の意義を有するもの（法的に新しい判断をしたものなど）に限られる（判例集・判例雑誌の目的から明らかであろう。最高裁判所の場合について後述）。

したがって，我々が入手可能なデータは，「判決」「判決文」全体の集合との関係で考える限りは，代表性・均衡性の点でかなり問題があるデータであるということになる。

しかし，そのようなデータであっても，ある程度の分量をそなえれば，ある種の文体情報など判決文の言語に関する分析にとって，意味のある情報を導くことができるのではないかと考えている。[22]

４.１.４　「判決文」の構成内容

判決文は，「主文」と「理由」[23]からなる。判決文と類似の性質を有するその他のものにおいても，この限度では同じである。

「理由」でどのような項目・内容をどの程度記載するかは，制度によっては定まっている（民事訴訟法 253 条，刑事訴訟法 335 条など）。判決文の分析における主たる関心はこの「理由」部分の言語になってくる。

なお，最高裁判所の判決・決定では，各裁判官が個別に意見を述べることができる（裁判所法 11 条）。[24] この部分については，「判決文」の本体とは区別をしておく必要があろう（基本的には分析対象から外すことになるが，分析内容・目的によってはそれらを対象とすることもあり得る。）。

４.１.５　仮名処理の点

我々が入手可能な状態の判決文の多くは，固有名詞について仮名処理がなされている。

この点について，例えば，次に述べる最高裁判所ホームページで公開されているデータについては，「文中の固有名詞などには，プライバシーなどへの配慮から，「A」「B」「C」等の記号に置き換えているものがあります。」[25]と注記されていることから分かるように，固有名詞について一律の処理をしているわけではない。[26]

いずれにしても，データの分析を行う段階で留意しておく必要がある点である。

４.２　「判決文」データの公開態様と入手[27]

４.２.１　裁判所ホームページ

裁判所ホームページに，次の 6 種類の検索フォームがあり，PDF 形式によって判決文のデータが公開されている。

最高裁判所判例集，高等裁判所判例集，下級裁判所判例集，
行政事件裁判例集，労働事件裁判例集，知的財産裁判例集

---

[22] この点の厳密な検討は今後の課題とする。

[23] これは，「主文」を導くに至った理由全体を表す（刑事訴訟法 44 条の「理由」はこの意味である。）。後掲民事訴訟法 253 条 1 項 3 号の「理由」より広い意味である。

[24] 裁判官が個別に付す「意見」には，「補足意見」「反対意見」「意見」の 3 種類がある。これらをまとめて「少数意見」という。「少数意見」ではない判決理由本体に示された内容を，これと対比する意味で，「多数意見」「法廷意見」などということもある（金子(2008)）。

[25] 「各判例について」（http://www.courts.go.jp/picture/hanrei_help.html）

[26] 行政に関係する事件では，行政主体（東京都，大阪市など）が仮名処理されていないものが多い。刑事事件の犯罪場所の記載は，都道府県名・市区町村名は仮名処理されていないものが多い。その他，行政が関係しない事件でも，仮名処理が行われていない箇所が見られた。

また，最高裁判所ホームページ内の「知的財産裁判例集」に収録されている判決や商用データベースに収録されている判決の中には，仮名処理を一切おこなっていないものも見られた。

[27] ここでは，コンピュータを用いたコーパスによる分析を行うという観点から，紙ベースで入手可能なものについては省略する。

このうち，「最高裁判所判例集」には，最高裁判所の発足後，1947（昭和 22）年から[28]毎年発行されている「最高裁判所民事判例集」「最高裁判所刑事判例集」（判例集），「最高裁判所裁判集民事」「最高裁判所裁判集刑事」（裁判集）に収録されている判決・決定の全文が収録されている。[29] これらの「判例集」「裁判集」は，最高裁判所の判決・決定から，先例としての価値があることその他の点を考慮して，最高裁判所に設置された「判例委員会」が選定して公表するものである（田原(1965)，川口(2010)）。

また，「下級裁判所判例集」は，2002（平成 14）年以降に各地の下級裁判所[30]が独自にホームページに公表した判決のデータが収録されている。ここには，どの判決を公表するかについて全国統一の基準は存在しないとされる。[31] したがって，公表される判決の種類・公表する裁判所などについて，かなりの偏りが生じている。

４.２.２　各種商用データベース

商用の判例データベースが複数ある。ここには，公式・民間の各種判例集[32]や，「判例時報」（判例時報社　旬刊），「判例タイムズ」（判例タイムズ社　月刊[33]）をはじめとする判例雑誌など，紙媒体で得られる判決文等の多くが（遡って）データ化されて収録されている（基本的には HTML データになっているが，一部，画像を PDF 形式にした状態のものもある。）。また，運営会社が独自に入手した判決等も収録されている。[34]

多くのデータベースで「全文検索」をすることが可能であり，検索件数については比較的容易に得られるといえる。ただ，それを超える複雑な分析を行うことはできない。[35]

また，データベースに収録されているもの多くは判例集・判例雑誌の掲載に由来することから，データの代表性・均衡性の点で検討を要するのは先に述べたとおりである。

５.「判決文」のデータの分析から得られる情報の例

今回は，「判決文」についての分析の第一段階として，最高裁判所ホームページの「最高裁判所判例集」で入手可能な判決・決定の PDF データをテキスト化したものについて，全体の傾向を見るため，判決 1 件あたりの文字数と 1 文あたりの文字数を算出した。[36]

ここでは。全区間を 5 年ごとに区切り（例えば 1955 年区間は 1951～1955 年を表す），判決 1 件あたりの文字数と 1 文あたりの文字数（句点（。）の数を文の数とみなし[37]，総文字数を文の数で割った数）を概算した（図 1）。なお，今回は，先に述べた仮名処理のバラつきの点について別途処理を行わなかった（PDF データに含まれる状態のまま）。

・　判決 1 件あたりの文字数は，事件の複雑さ・事件全体の情報量や，理由の判示の詳細さなどが影響すると考えられる。事件の複雑さ・情報量が同じであっても，理由部分を分かりやすく表現しようとすれば，全体の文字数は増えることになると考えられる。

---

[28] 初年（1947 年）については，年の途中からの収録となっている（中野(2009)）。
[29] これらは，2010 年 4 月に，遡って 1947 年以来の全データの公開が開始された（最高裁ホームページの当時のデータによる。http://web.archive.org/web/20100421055110/http://www.courts.go.jp/）。
[30] 高等裁判所，地方裁判所，家庭裁判所，簡易裁判所
[31] 担当者によって公開する判決の選定・件数にかなりのバラつきがあるといわれる（裁判官出身者の方の私信による。）。
[32] 紙媒体の判例集にどのようなものがあるかについては，中野(2009)が詳しい。
[33] 2012 年 12 月までは月 2 回刊であった。
[34] データベースを契約している弁護士が判決を持ち込む場合などがある。
[35] 例えば，正規表現による検索に対応したものは，現段階では見当たらない。
[36] いずれも，少数意見（脚注 24）の部分は除外してある。
[37] パーレン（ ）やカギ括弧「 」内に現れる句点（。）は除外してある。

これについては，1955 年区間以来ほぼ一貫して値が増加している。これが事件の複雑さ・情報量の増加によるものか，理由の判示の詳細かによるものかについては，今後の検討課題である。

・　また，1 文あたりの文字数については，一般的には，1 文が短いほど読みやすい文章であるといえる。渡辺（2010），白取（2010）などでは，判決の平易化の一つの表れとして，短文化を挙げていた。

しかし，最高裁判決に関する限り，1955 年区間が最小であった。また，2005 年区間で一旦極小となった後は増加の傾向にある。これらの要因が何であるかについては，今後の検討課題である。

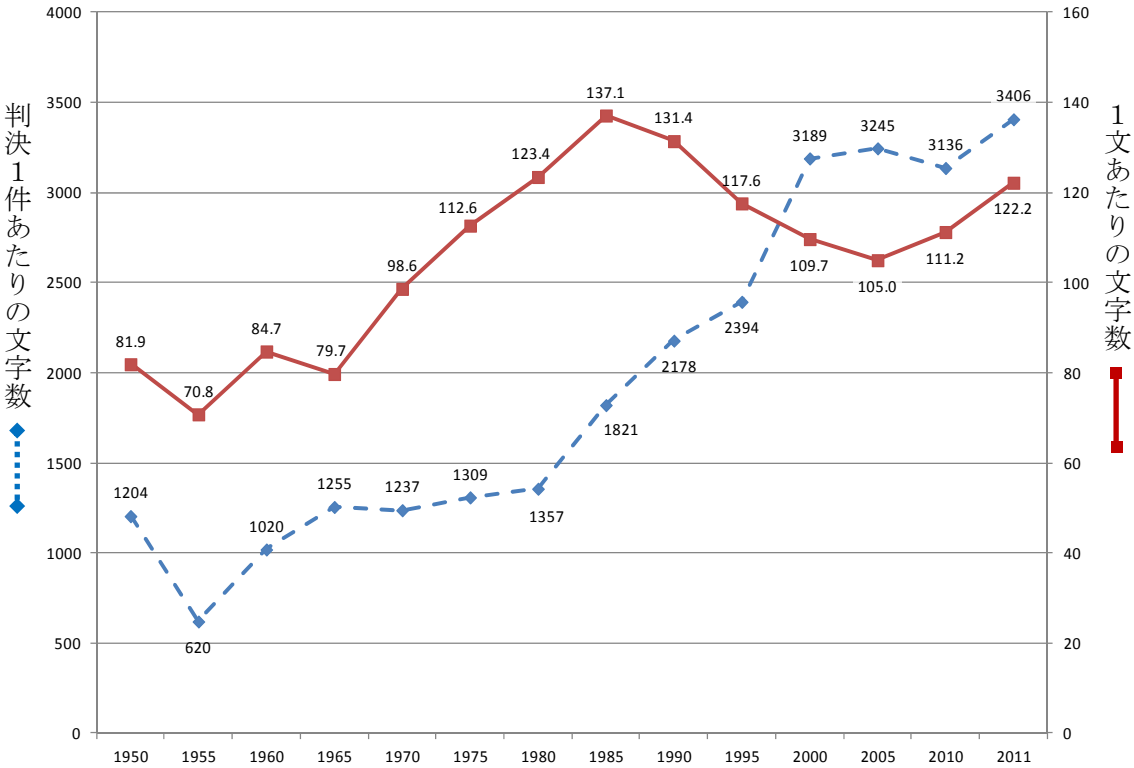


図 1　判決 1 件あたりの文字数／1 文あたりの文字数

6．おわりに

本稿では，主に，「判決文」の言語の分析のためにコーパスを用いる際の前提となる事項を整理した。今後，ここで整理した内容を踏まえて，「判決文」の言語の特徴についての分析を進めていきたい。

文　献

岩淵悦太郎，他(1979)『悪文』日本評論社
大橋將(2010)「法の日本語」専門日本語教育 12 号，pp.15-18
金子宏，新堂幸司，平井宜雄(2008)『法律学小辞典　第 4 版補訂版』有斐閣
川口冨男(2010)「判例委員会で学んだこと（裁判エッセイ 35)」中央総合法律事務所季刊ニュース 60 号，pp.14

（http://www.clo.jp/office_news.html よりダウンロード可能）
白取祐司(2010)『刑事訴訟法』日本評論社
末川博，他(1991)『新法学辞典』日本評論社
田原義衛(1965)『最高裁判決の内側』一粒社
中野次雄，他(2009)『判例とその読み方』有斐閣
日弁連(2007)「法廷用語の日常語化に関するＰＴ最終報告書（裁判員制度実施本部法廷用語の日常語化に関するプロジェクトチーム)」
http://www.nichibenren.or.jp/ja/citizen_judge/program/nichijyougoka.html
橋内武，他(2012a)「判決文はどう変わったか　裁判員制度以前と以後（その 1)」桃山学院大学総合研究所紀要，37:3，pp.223-231
（http://stars.andrew.ac.jp/modules/xoonips/detail.php?id=AA11337282-20120330-1223 よりダウンロード可能）
橋内武，他(2012b)『法と言語　法言語学へのいざない』くろしお出版
半田正夫，松田政行(2009)『著作権法コンメンタール』勁草書房
阪野慎司，他(2005)「判例コーパスを用いた判決文の要約手法」ディジタル図書館 27 号，pp.3-8
（http://www.dl.slis.tsukuba.ac.jp/DLjournal/No_28/1-banno/1-banno.html よりダウンロード可能）
阪野慎司，他(2006)「機械学習に基づく判決文の重要箇所特定」言語処理学会第 12 回年次大会発表論文集，pp.1075-1078
（http://www.anlp.jp/proceedings/annual_meeting/2006/pdf_dir/C5-2.pdf よりダウンロード可能）
福島至(1999)『コンメンタール刑事確定訴訟記録法』現代人文社
法教育研究会(2004)「法教育研究会報告書　我が国における法教育の普及・発展を目指して－新たな時代の自由かつ公正な社会の担い手をはぐくむために－」
http://www.moj.go.jp/shingi1/kanbou_houkyo_houkoku.html
前川喜久雄(2013)「コーパスの存在意義」『講座日本語コーパス 1 コーパス入門』，pp.1-31，朝倉書店
矢野信(2013)「コーパスを活用した法文データの分析に関する問題点」，本予稿集収録
渡辺咲子(2010)『刑事訴訟法講義』不磨書房

関連 URL

法令データ提供システム（総務省）　http://law.e-gov.go.jp/cgi-bin/idxsearch.cgi
最高裁判所規則集（最高裁判所）　http://www.courts.go.jp/kisokusyu/
司法統計（最高裁判所）　http://www.courts.go.jp/search/jtsp0010?
裁判例情報（最高裁判所）　http://www.courts.go.jp/
Westlaw Japan（新日本法規）　https://go.westlawjapan.com/wljp/app/signon/display

現代日本語の従属節に現れるモダリティ形式の分布

丸山 岳彦　（国立国語研究所 言語資源研究系）†

Distribution of Modal Expressions in Japanese Subordinate Clauses

Takehiko Maruyama　(Dept. Corpus Studies, NINJAL)

1　はじめに

2011 年に『現代日本語書き言葉均衡コーパス』（BCCWJ）が一般公開されたことにより、さまざまなレジスター（言語使用域、言語変種）における言語使用の実態を定量的に分析するための基盤が整えられ、どのような場面でどのような言語表現が使われるか（あるいは使われないか）を知ることができるようになった。例えば、言語の使用場面が変わることによって、複数のバリエーションを持つ文法形式の使用傾向に変化が見られるか否かを定量的に観察できるようになった。また、多様なレジスターを含む BCCWJ を参照コーパスとして利用し、これまでに行なわれてきた定性的・定量的な文法記述の結果を検証したり、追試をしたりすることも可能になった。このような状況の出現は、母語話者の内省に基づく従来の記述的文法研究に対して、言語使用の実態に即した「現実的な文法研究」を大きく前進させる可能性をもたらしたと言うことができる。

そこで本稿では、現代日本語における複文を構成する従属節のうち、特に連用節として機能する従属節がさまざまなレジスターの中でどのように出現しているかを定量的に分析することを目的とする。同時に、連用節を構成する述語句内にどのようなモダリティ形式が現れるかについても検討する。現代日本語文法の研究において、連用節とモダリティ形式の共起関係という問題はさまざまな角度から論じられてきたが、本稿では BCCWJ に基づいて、どのようなレジスターでどのような文法形式が観察されるかという点について、先行研究の検証・追試を交えながら論じていくことにしたい。

2　先行研究

日本語の述語句内部にどのようなモダリティ形式が現れ得るか、という問題については、国語学における助動詞の相互承接から見た構文論的研究 (渡辺, 1953; 北原, 1970) や、三上章による活用形の研究 (三上, 1953) などを背景として、現代における記述的文法研究の中に受け継がれている (仁田, 1991; 益岡, 1991, 2007)。また、連用節の述語句に現れ得るモダリティ形式の範囲を手がかりとして、連用節を 3 つのクラスに分類した南不二男の研究 (南, 1974, 1993) により、文の階層的な成立段階がモデル化され、この枠組み（通称「南モデル」）によって多様な文法現象を説明できることが明らかにされた (田窪, 1987; 野田, 1995)。

一方、定量的な観点からは、南 (1991) がさまざまな連用節の内部に取り得る要素（主題、補足語、修飾語、助動詞類）の出現数を調査し、数量的な分析を行なっている。ただし、調査資料の規模が少なかったためか、モダリティ形式を含む従属節（ガ、ケレドモ、カラ、ノデ、ナラ、タラなど）の用例数が十分に拾えているとは言い難く、調査では専ら、連用節内にある主題や補足語、修飾語が主要な分析対象となっている。

時代が下り、電子テキストが一般に普及した 2000 年代以降では、ナロック (2006) が、各種電子ブックや「CASTEL/J」などの電子テキストを大量に準備してブレンドし、そこから 17 万例以上におよぶ接続助詞の用例を検索・抽出して、連用節内部に現れる各モダリティ形式の分布を詳細に分析している。

† maruyama@ninjal.ac.jp

ナロック (2006) は、対比節（ガ、ケレドモ）・理由節（カラ・ノデ）・逆接節（ノニ・ニモカカワラズ）という 3 グループ、合計 6 種類の連用節を対象として、そこに現れるモダリティ形式を調査した。調査したモダリティ形式は、「認識的モダリティ」「証拠的モダリティ」「根源的モダリティ」の 3 グループ、各 4 形式の、12 種類である。ナロックによる調査の結果をまとめ直したものを、表 1 に示す。なお、論文中で、出現数は 10000 語あたりの数に正規化されている。

表 1: ナロック (2006) による調査結果のまとめ (1)

		対比節		理由節		逆接節	
		ガ	ケレド	カラ	ノデ	ノニ	ニモ…
認識的モダリティ	ニチガイナイ	36	67	111	20	39	69
	ハズ	37	60	83	8	375	46
	ダロウ	29	67	42	0	49	0
	カモシレナイ	96	130	51	18	15	0
証拠的モダリティ	ソウ 1（様相）	14	11	37	22	12	0
	ソウ 2（伝聞）	67	72	139	2	0	0
	ラシイ	38	79	45	26	6	9
	ヨウ・ミタイ	66	78	45	18	4	0
根源的モダリティ	ナケレバナラナイ	18	9	90	39	66	0
	タイ	145	37	57	18	24	2
	テモイイ	31	164	205	4	125	0
	ベキ	6	5	4	0	43	3

連用節の種類とモダリティ形式の種類を一つにまとめた上で、10000 語あたりの出現数を算出したナロックの調査結果を図式化すると、図 1 のようなグラフとなる。

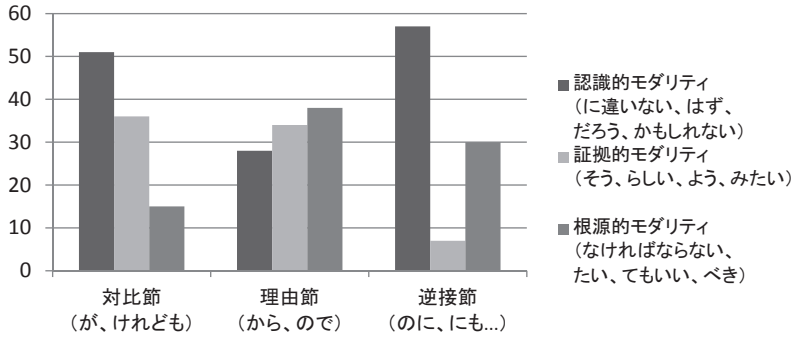


図 1: ナロック (2006) による調査結果のまとめ (2)

ここからナロックは、理由節では「根源 → 証拠 → 認識」の順で使用頻度が下がり、対比節では「認識 → 証拠 → 根源」の順で使用頻度が下がる、すなわち、理由節と対比節におけるモダリティ形式の使用傾向には逆の関係があることを指摘している。ナロック自身が述べている通り、使用したデータの集合が日本語のどの側面を代表しているのかが明確でないという問題点はあるものの、17 万例以上という大量の用例を集計し、各連用節が取るモダリティ形式の分布が明らかに異なることを実証したという点において、コーパス日本語学としての優れた実践例と言うことができる。

3　本稿の目的

コーパスを利用して言語の使用実態・使用傾向を調査することの利点は、母語話者の内省では知り得ない言語事実を実証的に明らかにすることができる点にある。文法研究に即して言えば、どのような場面でどのような文法形式がどれだけ使われるか（または使われないか）を記述することは、現実の発話状況や使用場面をも取り込んだ、より「現実的な」文法記述への可能性を開くことになる。

これは、翻れば、「頻度情報つきの日本語記述文法書」が存在しないという現状の問題に結びつく。英語の記述文法書（レファレンスグラマー）としては、Quirk et al. (1972, 1985) や Biber et al. (1999)、Huddleston and Pullum (2002)、Carter and McCarthy (2006) など、コーパスに基づく大規模な記述文法書がすでに数多く公刊されている。一方、体系立てられた日本語記述文法書としては、益岡・田窪 (1992) や「日本語記述文法研究会」による 7 巻組の記述文法書（日本語記述文法研究会 (2010) など）があるものの、日本語コーパスで観察されたレジスターごとの頻度情報を正面から取り上げて掲載した日本語記述文法書は、まだない。

そこで本稿では、言語の使用場面によって文法形式の分布がどのように異なるかを記述することを目的として、連用節とモダリティ形式の分布について分析を行なう。BCCWJ を分析対象データとして、異なる性質の書き言葉の中でどのような連用節・モダリティ形式が現れるかという問題を考えたい。分析の立脚点として、先に見たナロック (2006) を踏襲する。すなわち、ナロック (2006) における調査方法を BCCWJ における各レジスターに適用し、先行研究の検証・追試としながら、議論を進めることにする。

4　分析対象データ

4.1　『現代日本語書き言葉均衡コーパス』（BCCWJ）

ここでは、『現代日本語書き言葉均衡コーパス』（BCCWJ）の DVD 版に収録された長単位データを用いる。分析対象は、『BCCWJ-DVD 版』に収録された形態論情報付きテキストデータのうち、「韻文」を除く全体とした。韻文（俳句・短歌・詩）は、通常の書き言葉（いわゆる散文）とは明らかに異なる文体を持つため、連用節やモダリティを定量的に分析する上で、対象から除外することが適当であると判断した。タブ区切りの表形式になっている形態論情報付きテキストデータを Microsoft SQL Server に格納し、RDB に対してクエリを実行して用例の抽出と分類を行なった。

『BCCWJ-DVD 版』に収録されたデータ全体の語数を、レジスターごとに表 2 に示す。なお、本稿での「語数」とは、長単位として解析された単位数を指すものとする。

表 2: BCCWJ に収録された語数（長単位、概数）

出版サブコーパス	
書籍 (PB)	2762.3 万語
雑誌 (PM)	428.4 万語
新聞 (PN)	122.1 万語

図書館サブコーパス	
書籍 (LB)	3027.4 万語

特定目的サブコーパス	
白書 (OW)	382.4 万語
教科書 (OT)	92.5 万語
広報紙 (OP)	309.0 万語
ベストセラー (OB)	386.3 万語
Yahoo!知恵袋 (OC)	1030.0 万語
Yahoo!ブログ (OY)	1093.3 万語
法律 (OL)	83.3 万語
国会会議録 (OM)	449.8 万語

4.2　連用節の抽出手順

本稿で分析対象とする連用節は、ナロック (2006) での分析対象を一部入れ替えて、以下の 13 種類の形式とした。便宜的に、「並列節」「理由節」「条件節」という 3 つのグループに分類しておく。

並列節：ガ、ケレドモ、ケドモ、ケレド、ケド

理由節：カラ、ノデ

条件節：タラ、タラバ、ト、ナラ、ナラバ、レバ

ナロック (2006) で論じられていた「逆接節（ノニ、ニモカカワラズ）」は、「条件節（タラ、タラバ、ト、ナラ、ナラバ、レバ）」に入れ替えた。これは、条件節が日常の書き言葉の中に多く観察される代表的な連用節の一部であると考えられること、南モデルの B 類・C 類に分類される連用節であり、モダリティ形式との共起関係がしばしば問題とされること、大規模コーパスから逆接節としての「ノニ」を自動的に収集することが困難であること、などの理由による。

各形式の連用節について、RDB に格納された BCCWJ のデータ全体から用例を取り出した。その際、当該の語を中心として、前接する 10 語、後接する 10 語を結合し、前後 10 語ずつの KWIC 形式のデータを作成した。各語には、長単位として付与されている形態論情報のうち、書字形出現形（OT）、語彙素（LM）、品詞（POS）、文節境界（B）の情報を持たせた。KWIC データから一部を抜粋した例を、図 2 に示す。

	Sample_ID	order	CBL	B2	OT02	POS02	B1	OT01	POS01	OT00	POS00	OT_1	POS_1	B_2	OT_2	POS_2
1	PB13_00019	24220	keredomo	B	あっ	動詞-一般		た	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞	B	書か	動詞-一般
2	PB13_00045	14940	keredomo		ませ	助動詞		ん	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
3	PB13_00045	11590	keredomo	B	ござい	動詞-一般		ます	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
4	PB12_00356	34480	keredomo		な	助動詞		んだ	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
5	PB12_00356	19560	keredomo	B	あり	動詞-一般		ます	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
6	PB12_00356	13230	keredomo		な	助動詞		んです	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞	B	（	補助記号-括弧開
7	PB13_00045	18490	keredomo		は	助詞-係助詞	B	わかる	動詞-一般	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
8	PB13_00045	19580	keredomo	B	思う	動詞-一般		のです	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
9	PB13_00045	22570	keredomo	B	思い	動詞-一般		ます	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点
10	PB13_00045	23710	keredomo		ませ	助動詞		ん	助動詞	けれど	助詞-接続助詞	も	助詞-係助詞		、	補助記号-読点

図 2: BCCWJ から作成した KWIC データの例（一部抜粋）

ただし、この状態では、連用節以外の用例が KWIC データに含まれている。例えば、文節の冒頭に出現する「だけど」や「ですので」、「なら」、「そうすると」といった形式は、実質的には接続詞に相当するが、形態素解析の結果としては接続詞ではなく、「だ_助動詞 / けど_接続助詞」「です_助動詞 / ので_助動詞」「なら_助動詞」「そう_副詞 / する_動詞 / と_接続助詞」として解析されている（かつ、これは設計上、正しい解析結果である）。これらは、形式的には連用節の検索条件に適合するため、連用節の用例として抽出されてしまうことになる。そこで、以下に示すような表現が文節の始端に位置している場合は、接続詞に相当する用例と見なし、分析データから除外した。

> そうすると、そうすれば、そしたら、なら、そうしたら、だけど、ですので、だとしたら、だが、だとすれば、だったら、だとすると、ですが、なので、だから、でしたら、すると、であれば、ですから、であるなら、けど

4.3　モダリティ形式の抽出手順

次に、連用節の述語句内に現れるモダリティ形式を抽出した手順を示す。ここでは、以下に示す 13 種類のモダリティ形式を検討する。これは、ナロック (2006) が「認識的モダリティ」「証拠的モダリティ」「根源的モダリティ」という 3 つのグループに分類して検討したモダリティ形式に対応する。

> **認識的モダリティ**：だろう、かもしれない、はずだ、にちがいない
>
> **証拠的モダリティ**：（し）そうだ、ようだ、みたいだ、らしい、（する）そうだ
>
> **根源的モダリティ**：なければならない、たい、てもいい、べきだ

図 2 に示した KWIC データからこれらのモダリティ形式を抽出するために、検索用クエリを作成した。上記に挙げた各形式を基本形として、一部が漢字表記になる場合（「かもしれない」→「かも

知れない」など)、丁寧形になる場合（「だろう」→「でしょう」など)、過去形になる場合（「はずだ」→「はずだった」など)、否定形になる場合（「べきだ」→「べきではない」など)、交替形を取る場合（「なければならない」→「ないといけない」「なくちゃならん」など）をカバーするように検索式を作成した。

なお、形態素解析の結果によっては、同じモダリティ形式が異なる結果で解析されている場合がある。たとえば「かもしれません」というモダリティ形式は、「かもしれません_助動詞」だけでなく、「か_副助詞 / も_係助詞 / しれ_ 動詞 / ませ_助動詞 / ん_ 助動詞」と解析されている場合があり得る。これらの異なりを可能な限りカバーするように検索式を調整した。

さらに、形態素解析に伴う必然的な問題として、一定の割合で誤解析が含まれる、という点に注意する必要がある。例えば、「もっと長くいたかったのだが」というガ節の例には、願望を表すモダリティ形式「たい」が現れているが、実際の解析結果は「もっと_副詞 / 長く_形容詞 / いたかっ_形容詞 / た_助動詞/ のだ_助動詞 / が_接続助詞」となっており、「いたい」の部分が形容詞「痛い」として誤解析されてしまっている。このような場合は「たい」の用例としては抽出できないことになる。BCCWJ 全体の解析精度は 98%という高い精度を達成しているが (国立国語研究所, 2011)、それでも一定の割合で誤解析が含まれること、それに伴うモダリティ形式の抽出漏れがあることは不可避的な前提として、以降の分析を進める。

5　分析：連用節とモダリティ形式の分布

5.1　連用節とモダリティ形式の出現数

以下では、BCCWJ から抽出された連用節の各形式、およびモダリティ形式の分布について、分析を行なう。まずはじめに、分析対象データから抽出された連用節全体の出現数、および連用節中に現れたモダリティ形式の出現数を、表 3、4 に示す。

表 3: 連用節の出現数

分類	形式	出現数
並列節	ガ	383,523
	ケレドモ	21,014
	ケドモ	695
	ケレド	10,741
	ケド	58,607
理由節	カラ	159,187
	ノデ	137,430
条件節	タラ	99,886
	タラバ	136
	ト	237,910
	ナラ	22,997
	ナラバ	6,253
	レバ	169,252
総計		1,307,631

表 4: 連用節内部のモダリティ形式の出現数

分類	形式	出現数
認識的	だろう	9,731
	かもしれない	8,641
	はずだ	2,581
	にちがいない	393
証拠的	（し）そうだ	2,754
	ようだ	7,162
	みたいだ	1,525
	らしい	2,442
	（する）そうだ	1,936
根源的	なければならない	2,932
	たい	12,891
	てもいい	1,659
	べきだ	630
総計		55,277

表 3 のうち、並列節について見ると、ガの出現数がケレドモ類（ケレドモ、ケドモ、ケレド、ケド）の合計数を大きく上回っている。両者とも、基本的には逆接的または対比的な関係を表す並列節であるが、書き言葉の中ではガが圧倒的に多く選ばれているということになる。これは、ナロック (2006) で報告された結果と同様である。また、ケレドモ類に含まれる 4 形式の出現数を比較すると、

ケドが最も多く、ケレドモ、ケレドの順に続き、ケドモの出現数は極めて少ない。次に理由節について見ると、カラの方がノデよりも若干多いものの、出現数の上では、両者に大きな開きはない。これもナロック (2006) での報告と符号する。さらに条件節について見ると、ト、レバ、タラ、ナラの順に出現数が多くなっている。条件節を導くこれら 4 形式が持つ文法的性質の異同については、従来の文法研究の中で盛んに取り上げられてきたテーマであるが (益岡, 1993) 、書き言葉における出現数という点から見ると、4 つの形式に対してこのような順列を与えることができる。なお、ナラに対するナラバの出現数に対して、タラに対するタラバの出現数は極めて少なくなっている。

一方、表 4 に示したモダリティの各形式の出現数については、ナロック (2006) での報告と大きく異なる分布となっている。ナロック (2006) では、各連用節におけるモダリティ形式の出現数として、「カモシレナイ → ソウ 2 →ヨウ・ミタイ → ハズ → テモイイ → ニチガイナイ → ラシイ→ ダロウ →ナケレバナラナイ → タイ → ソウ 1 → ベキ」という順列が挙げられているが、今回 BCCWJ から抽出した結果とはかなりの差異がある。このような出現数の分布の違いが、使用したコーパスの性質によるものかどうかは、現段階では不明である。

5.2　各レジスターにおける連用節の分布

次に、BCCWJ を構成する各レジスターに連用節がどのように分布しているかを見る。ここでは、各レジスターにおける連用節の出現数を 100 万語あたりの出現数に正規化した上で、比較を行なった。各レジスターにおいて 100 万語あたりに出現する並列節・理由節・条件節の出現数を、図 3 に示す。

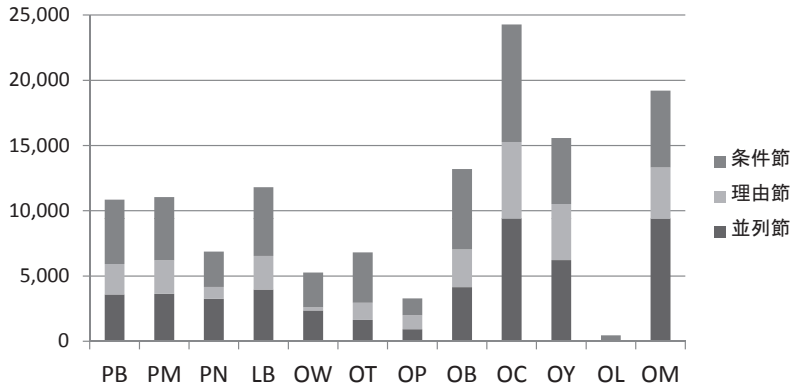


図 3: 各レジスターにおける連用節の分布（100 万語あたり）

図 3 を見ると、Yahoo!知恵袋 (OC)、国会会議録 (OM)、Yahoo!ブログ (OY) の順に、連用節全体の出現数が多くなっていることが分かる。これらは、一般人の書いた比較的くだけた書き言葉 (OC、OY)、または話し言葉を転記した書き言葉 (OM) である。一方、連用節の出現数が少ないのは法律 (OL)、広報紙 (OP)、白書 (OW) であり、これらは公共性の高い書き言葉である。ここから、書き言葉のスタイルの違いが連用節の出現に影響を与えていることが推測される。すなわち、比較的くだけたスタイルの書き言葉では連用節が多用されるのに対し、硬いスタイルの書き言葉では連用節はあまり使用されない、という傾向を見て取ることができる。この結果は、テキストの特徴的な傾向として、前者では多重的な節連鎖構造が発生することによってダラダラと続く長文や接続助詞で終わる「言いさし文」が現れやすいことを、後者では比較的短い文の連続が現れやすいことを、それぞれ予測させる。

5.3　連用節内部のモダリティ形式の分布

以下では、連用節の各形式の述語句部分に現れたモダリティ形式の分布について分析を行なう。はじめに、各連用節の全数のうち、各モダリティ形式を含む用例を集計した結果を、図 4 に示す。

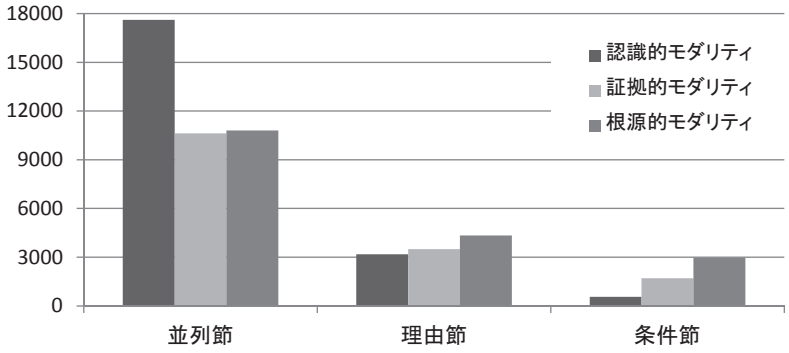


図 4: モダリティ形式を含む連用節の分布 (1)

図 4 を、図 1 で示したナロック (2006) による調査結果をと比べると、大まかには似た分布が観察される。ただし、(1) 並列節において根源的モダリティが証拠的モダリティを上回っている点、(2) 理由節の出現率が全体的に低く押さえられている点、という 2 つの差を見出すことができる。これは、5.1 節の最後に述べた、モダリティ形式の分布そのものがナロック (2006) と本稿との間で大きく異なるという点に起因している可能性がある。

次に、連用節の各形式について、その内部にどのような種類のモダリティを取るかを集計した。各モダリティ形式を含む割合の高い順に連用節を並べ替えた結果を、図 5 に示す。

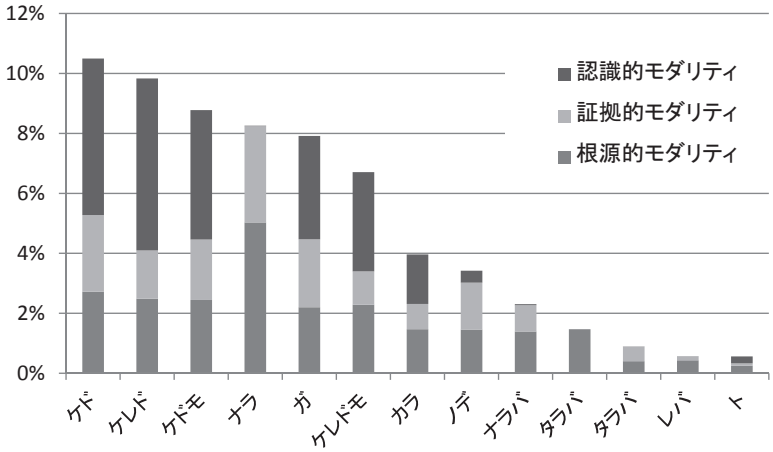


図 5: モダリティ形式を含む連用節の分布 (2)

図 5 を見ると、モダリティ形式を含む比率の高いから低いものに移るにしたがって、並列節、理由節、条件節へと移行していくことが分かる。これは、南モデルにおける C 類に分類される並列節が、より広い範囲のモダリティ形式を取り得るという統語的な特徴を反映しているものと考えられる。

唯一、条件節のナラが並列節の中に混じって高い比率を示しているが、これはグラフを見れば分かるように、ナラが根源的モダリティを突出して多く取っていることに起因している。元データを確認したところ、ナラ節が根源的モダリティを取る 1157 例のうち、約 93%の 1075 例が「たい」で占められていた。さらに、そのうち約 45%は Yahoo!知恵袋 (OC) からの例であった。すなわち、投稿された質問に対して回答が寄せられる知識検索サービスの文脈において、「〜したいなら」という言い

回しが定型的に多用されていることに原因があると考えられる。

なお、ナロック (2006) には、モダリティ形式を内部に含む割合の高い連用節として、「カラ ＞ ケレド ＞ ノニ ＞ ガ ＞ ノデ ＞ ニモカカワラズ」という順列が与えられているが、今回 BCCWJ を分析した結果からは、この順列を再現することはできなかった。

5.4　各レジスターにおけるモダリティ形式の分布

最後に、図 4 と同様、各連用節がどのようなモダリティ形式を含むかについて、レジスターごとに集計を行なった。結果を図 6 に示す。なお、ここでは縦軸を出現数とする。

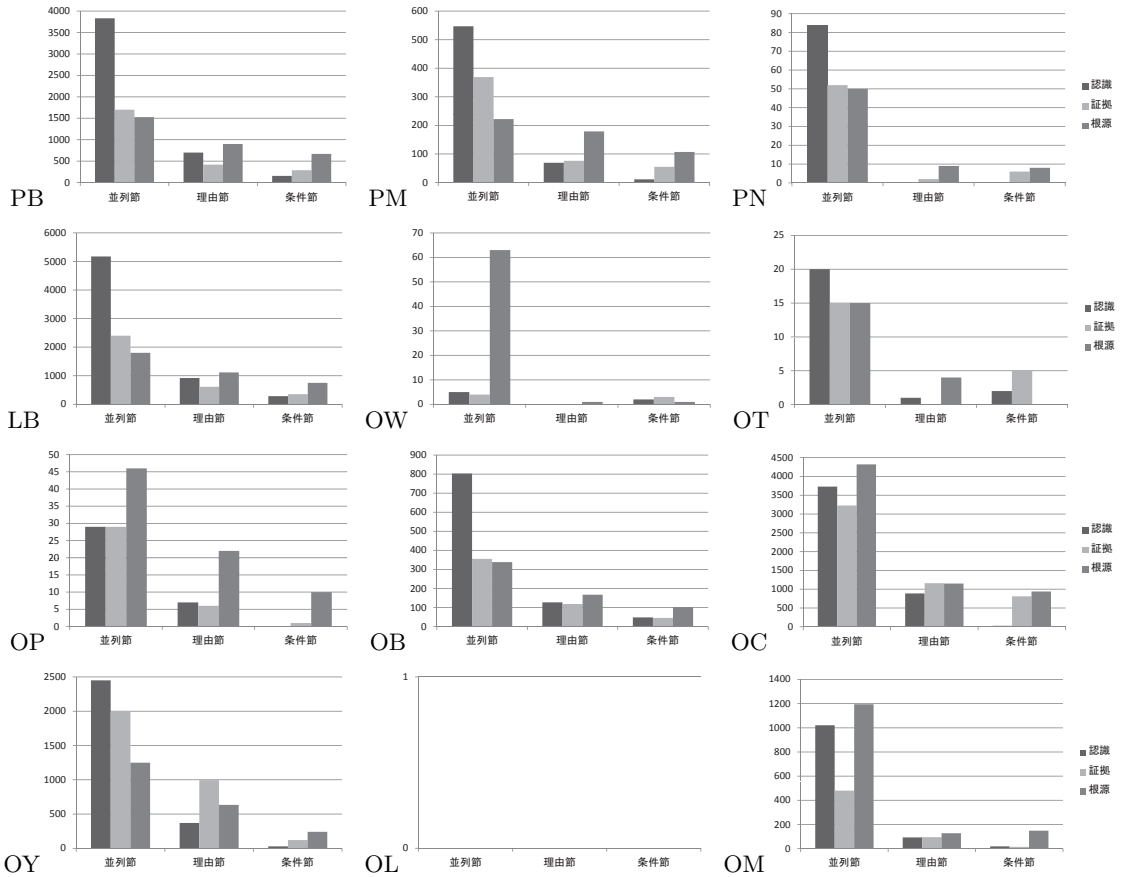


図 6: 各モダリティ形式を含む連用節の割合（レジスターごと）

図 6 からは、レジスターによって連用節が含み得るモダリティ形式の分布に違いがあることが分かる。出版サブコーパスの書籍 (PB)、図書館サブコーパスの書籍 (LB)、ベストセラー (OB) はほぼ同じ分布を示していることから、母集団の性質は異なるものであっても、同じ書籍として共通した文法的特徴を備えていると見てよい。雑誌 (PM) は図 1 で示したナロック (2006) の調査結果と最も近似した分布になっている。雑誌には、文芸小説やファッション誌、さまざまな情報誌、対談記事など、多彩なテキストが含まれていることから、多くの電子ブックやさまざまな電子テキストをブレンドして使用したナロック (2006) のデータと、テキスト集合としての性質が近いのかもしれない。白書 (OW) は他のレジスターと大きく分布が異なり、「義務・必要・許可」などを表す根源的モダリティが並列節に突出して多く見られる一方、その他のモダリティ形式はほとんど見られない。特に認識的モダリティのように、書き手・話し手の不確かな捉え方を表すための文法形式は、客観的な事実を記述するスタイルを取る白書のテキストとしてふさわしくないのだろう。また、法律 (OL) では、連用

節中にモダリティ形式が一切現れない（連用節の数そのものも非常に少ない）。そもそも法律とは、ある行為を命令・禁止したり、ある権利を保障したりすることを明示的かつ曖昧性のないように述べるためのテキストであるため、話し手の判断や態度を表すモダリティ形式は、法律文というテキストが持つ機能にそぐわないものと考えられる。

6　考察

ここまでの分析で明らかになったのは、BCCWJ のレジスターごとに、使われる連用節の形式と数、およびモダリティ形式の種類と数が大きく異なるという点である。言語の使用場面によって使用される文法形式は異なる、という直感的には理解される言語の使用実態に伴う特徴が、BCCWJ のレジスターの違いを利用して実証的に明らかにできたことになる。母語話者の（少なくとも筆者の）内省に基づくだけでは、異なるレジスター間で連用節・モダリティ形式の使用傾向にどのような差があるかを正確に把握することは難しい。

本稿での分析の立脚点としたナロック (2006) では、図 1 のような結果をもとに、対比節と理由節におけるモダリティ形式の使用傾向に逆の関係があることを結論として述べていた。これは重要な知見の一つであり、BCCWJ の一部のレジスターでもほぼ同じ分布と傾向を確認することができた。しかしながら、それがどのような場面でも成り立つ使用傾向であるかと言えば、本稿での分析でも明らかなように、必ずしもそうではない。コーパスからある文法的特徴を抽出できたとき、それが現代日本語に広く観察される特徴であるのか、ある特定の使用場面に依存した特徴であるのかという区別に、分析者は常に意識的である必要がある。

性質の異なるテキストの間で文法形式の分布に差が観察されるとき、それを裏付けるための説明の仕方については、さまざまな観点があると考えられる。例えば、本稿でも一部で触れた書き言葉のスタイル差、すなわち、テキストの硬さ・軟らかさ、客観的か主観的か、改まった書き言葉かくだけた書き言葉か、などといった差は、文法形式の分布の違いを裏付ける指標として考えられるだろう。さらに、書き手の性別・年齢、読者対象、執筆年代といったメタ情報の違いや、あるテキストが執筆された目的やテキスト自体が担う機能の違い、さらに名詞率、MVR、TTR、品詞構成比率などといったテキスト自体から得られる統計的指標なども、文法形式の分布の違いを説明するための手がかりになり得ると考えられる (小磯, 田中, 小木曽,・近藤, 2012)。レジスターごとに観察される言語的特徴と、それを裏付けるためのさまざまな指標・観点を組み合わせて考えることにより、社会における言語活動のどのような場面にどのような実態が観察されるのか、その現実的なありさまを記述していくことが可能になると思われる。

7　おわりに

本稿では、現代日本語における複文を構成する連用節、およびモダリティの諸形式が、実際の書き言葉の中でどのように分布しているかについて分析を行なった。ナロック (2006) が実施した調査・記述を立脚点として、BCCWJ を用いてさまざまなレジスターにおける連用節・モダリティ形式の分布を見た。その上で、ナロック (2006) による観察結果と同様の結果を、一部のレジスターで確認することができた。一方、別のレジスターではそれとは異なる分布を示していることを示した。

文献

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education.

Carter, R. & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge University Press.

Huddleston, R. D. & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

日本語記述文法研究会（編） (2010).『現代日本語文法 1』. くろしお書房.

北原保雄 (1970).「助動詞の相互承接についての構文論的考察」.『国語学』, **83**, 32–59.

小磯花絵, 田中弥生, 小木曽智信,• 近藤明日子 (2012).「テキストの多様性をとらえる分類指標の体系化の試み (2)」.『言語処理学会 第 18 回年次大会 発表論文集』, pp. 739–742.

益岡隆志 (1991).『モダリティの文法』. くろしお出版.

益岡隆志・田窪行則 (1992).『基礎日本語文法 —改訂版—』. くろしお出版.

益岡隆志（編） (1993).『日本語の条件表現』. くろしお出版.

益岡隆志 (2007).『日本語モダリティ探究』. くろしお出版.

三上章 (1953).『現代語法序説』. 刀江出版.

南不二男 (1974).『現代日本語の構造』. 大修館書店.

南不二男 (1991).「現代日本語の従属句についての小調査」.『日本語学』, **10** (12).

南不二男 (1993).『現代日本文法の輪郭』. 大修館書店.

ナロックハイコ (2006).「従属節におけるモダリティ形式の使用」.. **6** (1), 21–37.

仁田義雄（編） (1991).『日本語のモダリティと人称』. ひつじ書房.

野田尚史 (1995).「文の階層構造からみた主題ととりたて」. 益岡隆志, 野田尚史,• 沼田善子（編）,『日本語の主題と取り立て』, pp. 1–35. くろしお出版.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A Grammar of Contemporary English*. Longman.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

田窪行則 (1987).「統語構造と文脈情報」.『日本語学』, **6** (7), 37–48.

渡辺実 (1953).「叙述と陳述 —述語文節の構造—」.『国語学』, **13/14**, 20–34.

国立国語研究所コーパス開発センター (2011).『『現代日本語書き言葉均衡コーパス』利用の手引 第 1.0 版』. 国立国語研究所.

ポスター発表（2） Bグループ

9月6日（金） 14:00～15:00

クラスタリングを利用した能動学習による語義曖昧性解消の領域適応

小野寺 喜行 (茨城大学 工学部 情報工学科) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Domain Adaptation for Word Sense Disambiguation by Active Learning Using a Clustering Method

Yoshiyuki Onodera (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応の問題に対して，初期段階でデータ選択にクラスタリングを利用した能動学習手法を提案する．

自然言語処理のタスクにおいて帰納学習手法を用いる際，訓練データとテストデータは同じ領域のコーパスから得ていることが通常である．ただし実際には異なる領域である場合も存在する．そこである領域（ソース領域）の訓練データから学習された分類器を，別の領域（ターゲット領域）のテストデータに合うようにチューニングすることを領域適応という[3]．

領域適応の問題をターゲット領域のラベル付きデータの不足からくる問題として見なせば，能動学習 (Settles (2010)) や半教師あり学習 (Chapelle et al. (2006)) を利用することは有効である．ここでは WSD の領域適応の問題に対して，能動学習を利用する．

一般に能動学習はラベルなしデータの集合から学習効果の高いデータを選択し，そのデータにラベルを付けて訓練データに追加することで，徐々に分類器の精度を高めてゆく．能動学習のポイントはどのようにして学習効果の高いデータを選択するかである．通常はその時点で保持しているラベル付きデータを利用してその選択が行われるが，領域適応では能動学習の初期の段階ではソース領域のラベル付きデータで占められるため，上記の選択が有効に行える保証がない．ここでは能動学習の初期の段階ではターゲット領域のデータをクラスタリングし，そこから代表点を選ぶことを提案する．

実験では BCCWJ コーパス (Maekawa (2007)) の 2 つ領域 PB (書籍) と OC (Yahoo! 知恵袋) から共に頻度が 50 以上の多義語 17 単語を対象にして，能動学習を利用した WSD の領域適応の実験を行い，提案手法の有効性を示す．

2　クラスタリングを利用した能動学習

2.1　能動学習

精度の高い分類器を構築するためにはラベル付きデータを増やせばよい．ただしラベルを付けるコストは高いため，少量のラベル付けで精度の高い分類器を構築する方法が望まれている．能動学習もそのような背景から考案された学習手法である．

能動学習では学習効果の高いデータをシステムが選択し，ユーザがそのデータにラベルを付ける．これを繰り返すことで分類器の精度を徐々に向上させてゆく．ランダムに取り出されたデータにラベルを付けるよりも，少量のラベル付けで精度の高い分類器が構築できる．

能動学習には様々な手法が存在するが，ここでは簡易でありながら効果が高い Schohn の手法 (Schohn and Cohn (2000)) を利用する．Schohn の手法を図 1 に示す．

[2] 10t4019s@hcs.ibaraki.ac.jp
[2] shinnou@mx.ibaraki.ac.jp
[3] 領域適応は機械学習の分野では転移学習 (神嶌敏弘 (2010)) の一種と見なされている．

(1) ラベル付きデータから分類器を作成する．

(2) 作成した分類器によりラベルなしデータを識別する．このとき識別の信頼度も求める．

(3) 識別の信頼度が最も低いデータにラベルを付け，ラベル付きデータに追加する．

(4) (1) に戻る

図 1: 能動学習の手順

ここでは分類器として SVM を用いる．また SVM ツールには libsvm[4] を用いた．そこでは `-b` オプションにより識別の信頼度が求められる．

2.2 初期データ選択

領域適応の問題解決のために，ターゲット領域のラベル付きデータを用意する戦略をとれば，能動学習が利用できることは明かである．ただしこの場合，能動学習の初期の段階ではソース領域のラベル付きデータの占める割合が高く，そこから新たにラベルを付けるデータを選択する手法 (図 1) が有効に働かない可能性がある．

そこで本論文では能動学習の初期の段階のみ，クラスタリングを利用してデータを選択を行う．具体的には，ターゲット領域のデータを K 個のクラスタにクラスタリングする．次に各クラスタ C_i の中から代表点 c_i を選ぶ．得られた $\{c_1, c_2, \cdots, c_K\}$ を初期データとする．代表点の選び方は，そのクラスタ内の他のデータとの類似度の和が最大のデータをとる方法である．

$$c_i = \arg\max_{c \in C_i} \sum_{x \in C_i} sim(c, x)$$

このようにして選択されたデータはターゲット領域のみに依存しているので，ソース領域の影響を受けない．

また本論文では $K = 5$ に設定している．クラスタリングのツールには CLUTO[5] を用いた．

3 実験

実験には BCCWJ コーパスの OC(Yahoo!知恵袋) と PB(書籍) を使った．その中から表 1 に示す 17 単語について OC から PB，PB から OC の 2 通り領域適応の実験を行う．

[4] `http://www.csie.ntu.edu.tw/ cjlin/libsvm/`
[5] `http://glaros.dtc.umn.edu/gkhome/views/cluto`

表 1: 実験対象単語

単語	辞書上の語義数	PB での頻度	PB での語義数	OC での頻度	OC での語義数
言う	3	1114	2	666	2
入れる	3	56	3	73	2
書く	2	62	2	99	2
聞く	3	123	2	124	2
来る	2	104	2	189	2
子供	2	93	2	77	2
時間	4	74	2	53	2
自分	2	308	2	128	2
出る	3	152	3	131	3
取る	8	81	7	61	7
場合	2	137	2	126	2
入る	3	118	4	68	4
前	3	160	2	105	3
見る	6	273	6	262	5
待つ	4	153	3	62	4
やる	5	156	4	117	3
ゆく	2	133	2	219	2
平均	3.35	193.9	2.94	150.6	2.88

実験では図 1 の能動学習手法 (従来手法) と提案手法による語義識別の平均正解率で比較する．提案手法では，クラスタリングから初期の 5 点を選び訓練データに追加する．その後，通常の能動学習で 5 点選び訓練データに追加し，それを用いて SVM の学習を行う．また従来手法では，クラスタリングを用いずに通常の能動学習で 10 点を訓練データに追加する．

従来手法と提案手法の学習の平均正解率を OC から PB の領域適応について表 2 と図 2 に，PB から OC の領域適応について表 3 と図 3 に示す．結果，PB から OC の領域適応では提案手法の効果はあったと言える．しかし OC から PB の領域適応では，8 点目までは提案手法の方が平均正解率が高かったが，最終的な 10 点目での平均正解率は従来手法の方が高かった．

表 2: 平均正解率 (%)(OC → PB)

追加データ数	従来手法	提案手法
1	73.20	
2	73.14	
3	73.96	
4	75.21	
5	76.11	76.32
6	77.05	77.31
7	76.72	77.80
8	78.00	78.58
9	78.88	78.82
10	79.47	79.04

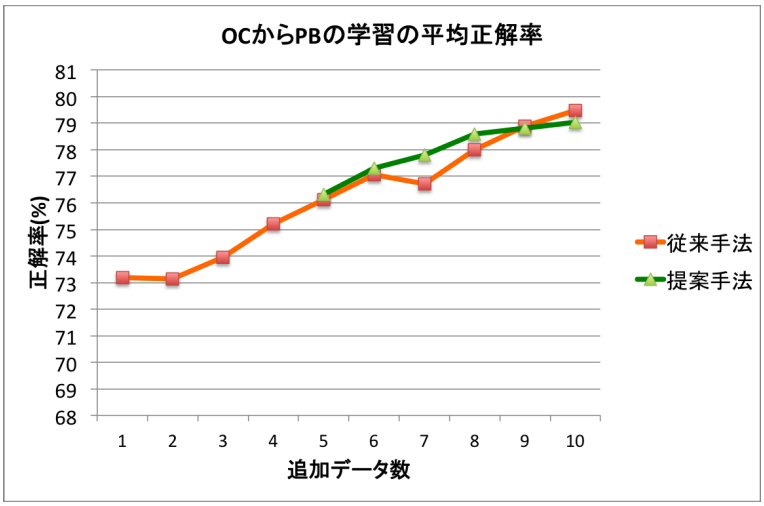


図 2: 平均正解率の変化 (OC → PB)

表 3: 平均正解率 (%)(PB → OC)

追加データ数	従来手法	提案手法
1	72.12	
2	73.26	
3	74.22	
4	76.02	
5	76.46	75.78
6	76.92	77.27
7	77.18	77.49
8	77.61	79.03
9	78.02	79.76
10	78.15	79.54

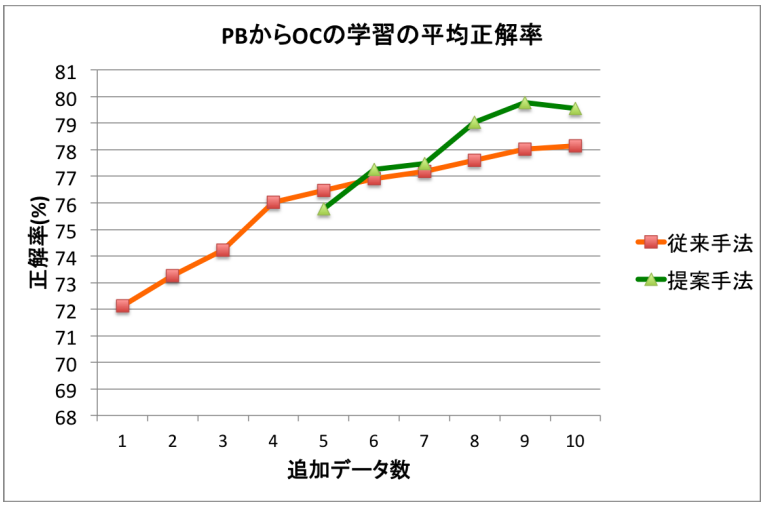


図 3: 平均正解率の変化 (PB → OC)

4　考察

4.1　領域間距離

領域適応ではソース領域とターゲット領域との距離が本質的な役割を担う．距離が近ければ，ソース領域の知識の多くはターゲット領域においても有効である．逆に距離が離れていれば，ソース領域の知識はターゲット領域においてあまり役立たない．本論文における提案手法のアイデアは，ソース領域とターゲット領域との距離がある程度離れていることを想定している．距離が近い場合，従来手法がそのまま使えるはずである．そのため提案手法の効果はソース領域とターゲット領域との距離に相関があると考えられる．ここでは，この点を確認する．

まずソース領域 S とターゲット領域 T との距離の測り方が問題である．これは様々な手法が提案されているが，ここでは以下の形で領域間の類似度 $sim(S,T)$ を測定した．注意としてここでの類似度は方向性があり必ずしも $sim(S,T) = sim(T,S)$ とはなっていないことに注意する．

領域 X 内のデータ x は素性リスト $x = \{f_x^1, f_x^2, \cdots, f_x^{m(x)}\}$ で表せる．集合 F_x はこの素性リストの素性を集めたものである．

$$F_x = \bigcup_{x \in S} \{f_x^1, f_x^2, \cdots, f_x^{m(x)}\}$$

素性 $y \in F_x$ の頻度を $g(y)$ 表す．ソース領域 S とターゲット領域 T と類似度 $sim(S,T)$ を以下で定義する．

$$sim(S,T) = \frac{\sum_{y \in F_s \cap F_T} g(y)}{\sum_{y \in F_s} g(y)}$$

上記の類似度と前述した実験結果との対応関係を表 4 と表 5 に示す．表 4 は OC から PB の領域適応，表 5 は PB から OC の領域適応である．

それぞれ類似度と手法の効果（正解率の差）について相関係数を求めると，OC から PB の学習では 0.17，PB から OC の学習では-0.22 であった．この値からは領域間距離と提案手法の効果との間に関連性は認められない．ただしここで行った領域間の距離の測定は簡易なものであり，適切に測定できていない可能性が高い．提案手法が効果が出るのは，領域間距離が離れている場合であると考えられるので，今後は適切な領域間距離の測定法を考察したい．

表 4: 領域間類似度と手法 (%) の効果 (OC → PB)

単語	sim(OC,PB)	正解率の差	従来手法	提案手法
言う	0.53	2.42	79.26	81.68
入れる	0.13	-1.67	80.36	78.69
書く	0.20	-3.76	90.32	86.57
聞く	0.32	-0.77	79.67	78.91
来る	0.26	-1.79	99.04	97.25
子供	0.34	-4.95	56.99	52.04
時間	0.27	0.60	90.54	91.14
自分	0.36	-0.28	97.73	97.44
出る	0.26	-0.65	60.53	59.87
取る	0.25	6.35	30.86	37.21
場合	0.23	0.49	86.13	86.62
入る	0.32	-1.67	61.02	59.35
前	0.23	0.97	88.13	89.09
見る	0.33	0.28	84.62	84.89
持つ	0.34	-0.62	79.74	79.11
やる	0.31	-0.42	93.59	93.17
ゆく	0.29	-1.90	92.48	90.58

表 5: 領域間類似度と手法 (%) の効果 (PB → OC)

単語	sim(PB,OC)	正解率の差	従来手法	提案手法
言う	0.45	-0.76	82.58	81.82
入れる	0.16	2.69	78.08	80.77
書く	0.30	1.26	73.74	75.00
聞く	0.29	-1.17	70.16	68.99
来る	0.32	1.02	80.42	81.44
子供	0.27	11.86	45.45	57.32
時間	0.18	1.30	84.91	86.21
自分	0.28	0.44	88.28	88.72
出る	0.25	3.36	68.70	72.06
取る	0.17	5.61	45.90	51.52
場合	0.23	-1.44	97.62	96.18
入る	0.17	1.91	72.06	73.97
前	0.21	-1.47	92.38	90.91
見る	0.34	0.63	86.26	86.89
持つ	0.24	-1.01	93.55	92.54
やる	0.28	-0.61	94.87	94.26
ゆく	0.34	0.14	73.52	73.66

4.2　スペクトラルクラスタリングの利用

提案手法ではクラスタリングを利用して，初期のデータ選択を行う．このためクラスタリングの精度が正解率に影響を与えている可能性がある．ここでは精度の高いクラスタリング手法として知られるスペクトラルクラスタリング (Von Luxburg (2007)) を利用して提案手法を試す.

従来手法，提案手法およびスペクトラルクラスタリングを用いた手法の 3 手法による平均正解率を，OC から PB の学習について表 6 と図 4 に，PB から OC の学習について表 7 と図 5 に示す.

実験の結果，OC から PB の領域適応では提案手法より平均正解率が向上し，従来手法の結果も上回っている．また，PB から OC の領域適応において平均正解率は提案手法より下がったものの，従来手法の結果より高い正解率となった．つまり提案手法はクラスタリングの精度とも関連しており，

精度の高いクラスタリング手法を利用することで，効果が現れると言える．

表 6: スペクトラルクラスタリングの利用 (OC → PB)

追加データ数	従来手法	提案手法	スペクトラルクラスタリングの利用
1	73.20		
2	73.14		
3	73.96		
4	75.20		
5	76.11	76.32	75.74
6	77.05	77.31	76.34
7	76.72	77.80	77.80
8	78.00	78.58	78.20
9	78.88	78.82	78.87
10	79.47	79.04	79.49

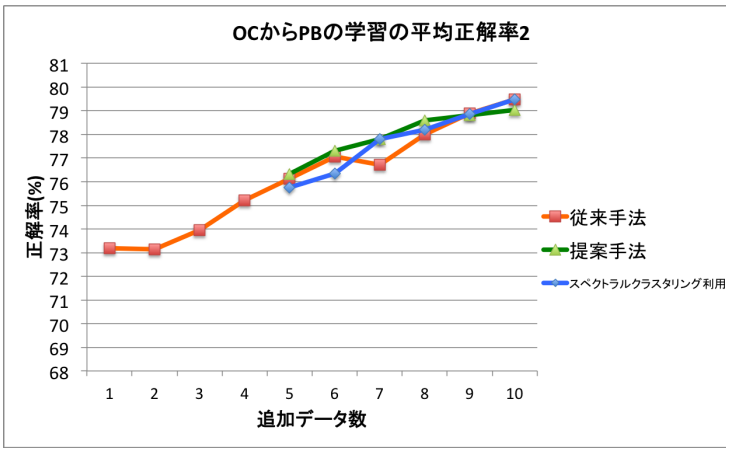


図 4: スペクトラルクラスタリングの利用による平均正解率の変化 (OC → PB)

表 7: スペクトラルクラスタリングの利用 (PB → OC)

追加データ数	従来手法	提案手法	スペクトラルクラスタリングの利用
1	72.12		
2	73.26		
3	74.22		
4	76.02		
5	76.46	75.78	74.48
6	76.92	77.27	76.09
7	77.18	77.49	76.84
8	77.61	79.03	77.25
9	78.02	79.76	78.38
10	78.15	79.54	79.14

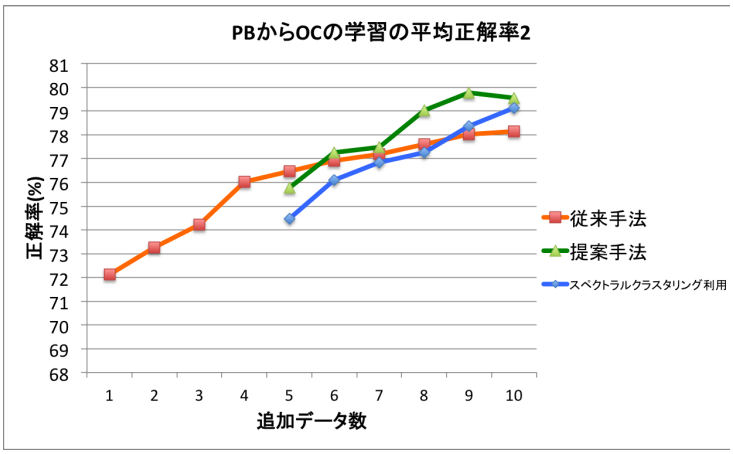


図 5: スペクトラルクラスタリングの利用による平均正解率の変化 (PB → OC)

5　おわりに

本論文では WSD の領域適応の問題に対して能動学習を試みた．領域適応に能動学習を利用する場合，初期の段階ではソース領域のデータが占める割合が高いため，データ選択が適切に行えない可能性がある．そこでここでは能動学習の初期の段階ではターゲット領域のデータをクラスタリングすることでデータ選択を行うこと提案した．

BCCWJ コーパスの OC(Yahoo!知恵袋) と PB(書籍) の 2 つの領域を用いた実験では，PB から OC の領域適応では提案手法の効果はあったが，OC から PB の領域適応では 8 点目までは提案手法の効果はあったが，最終的な 10 点目では従来手法の方が平均正解率は高かった．ただし精度の高いクラスタリング手法を使うことで，提案手法の効果を確認できた．

提案手法を効果的に利用するには，領域間距離を適切に測ることが重要だと考えている．この測定法を考案することが今後の課題とする．

文献

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien et al. (2006) *Semi-supervised learning*, Vol. 2: MIT press Cambridge.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Greg Schohn and David Cohn (2000) “Less is more: Active learning with support vector machines,” in *ICML*, pp. 839–846.

Burr Settles (2010) “Active learning literature survey,” *University of Wisconsin, Madison*.

Ulrike Von Luxburg (2007) “A tutorial on spectral clustering,” *Statistics and computing*, Vol. 17, No. 4, pp. 395–416.

神嶌敏弘 (2010) 「転移学習」，人工知能学会誌，第 25 巻，第 4 号，pp.572–580.

語義曖昧性解消の領域適応における Misleading データの存在と検出

吉田 拓夢 (茨城大学 工学部 情報工学科) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Existence and Detection of Misleading Data in Domain Adaptation for Word Sense Disambiguation

Hiromu Yoshida (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応の問題に対して，負の転移を起こす Misleading データの存在と検出について議論する．

自然言語処理のタスクにおいて帰納学習手法を用いる際，訓練データとテストデータは同じ領域のコーパスから得ていることが通常である．ただし実際には異なる領域である場合も存在する．そこである領域（ソース領域）の訓練データから学習された分類器を，別の領域（ターゲット領域）のテストデータに合うようにチューニングすることを領域適応という[3]．

領域適応の問題を扱う場合，ソース領域のラベル付きデータは全て使う方がよいと考えられるが，必ずしもそうではない．ソース領域のラベル付きデータの一部は利用しない方が逆に精度が向上することがある．これは負の転移現象 (Rosenstein et al. (2005)) と呼ばれる現象である．Jiang はこのような悪影響を及ぼすデータを Misleading データと呼び，そのデータを検出・削除してから学習を行うことを提案している (Jiang and Zhai (2007))．一方，新納は WSD の領域適応の問題を，語義分布の推定の問題とスパース性の問題に分けて考え，語義分布の推定に影響を及ぼさなければ，ソース領域のラベル付きデータは全て使う方が良いことを主張している (新納浩幸・佐々木稔 (2013))．

WSD の領域適応の問題は共変量シフトの問題 (Shimodaira (2000) 杉山将 (2006)) と見なせるため，密度比を用いてデータの重要性を測ることができる．ここでは密度比の低いデータを Misleading データとして設定できるかどうかを調べる．

実験では BCCWJ コーパス (Maekawa (2007)) の 2 つ領域 PB (書籍) と OC (Yahoo! 知恵袋) から共に頻度が 50 以上の多義語 17 単語を対象にして，WSD の領域適応の実験を行なった．密度比を用いてデータの重要度を測り，その値が低いデータを Misleading データと考え，それらデータを外した場合の平均正解率を調べた．結果，密度比から Misleading データを検出する効果はなかった．総当たり的に各データが Misleading データかどうかを確認したところ，Misleading データの存在は確認できた．また密度比からは Misleading データを検出することが困難であることも示した．有効な Misleading データの検出法を今後の課題とする．

2　語義曖昧性の領域適応

WSD の対象単語 w の語義の集合を $C = \{c_1, c_2, \cdots, c_k\}$，$w$ を含む文（入力データ）を x とする．WSD の問題は最大事後確率推定を利用すると，以下の式の値を求める問題として表現できる．

$$\arg\max_{c \in C} P(c)P(x|c)$$

つまり訓練データを利用して語義の分布 $P(c)$ と各語義上での入力データの分布 $P(x|c)$ を推定することで WSD の問題は解決できる．今，ソース領域を S，ターゲット領域を T とした場合，WSD の領域適応の問題は $P_S(c) \neq P_T(c)$ と $P_S(x|c) \neq P_T(x|c)$ から生じている．

[2] 10t4067l@hcs.ibaraki.ac.jp
[2] shinnou@mx.ibaraki.ac.jp
[3] 領域適応は機械学習の分野では転移学習 (神嶌敏弘 (2010)) の一種と見なされている．

新納は $P_S(x|c) = P_T(x|c)$ は成立していると考え，$P_T(x|c)$ の推定を困難にしているのはコーパスのスパース性の問題として捉えた．つまり $P_T(c)$ の推定に影響を及ぼさなければ，ソース領域のラベル付きデータは全て利用しても問題ないことを示した (新納浩幸・佐々木稔 (2013))．

一般に識別モデルである SVM は語義分布の影響をあまり受けずに識別境界を定める．このため，新納の結果に基づければ，SVM で学習するのであれば，ソース領域のラベル付きデータは全て利用しても精度の低下は生じない，あるいは非常に小さいと考えられる．

本論文では WSD に SVM を用いる．その際に，Misleading データと見なせるものを省くことで精度が向上するかどうかを確認する．

3　密度比による Misleading データの検出

語義が曖昧な単語 w を含む文 s があったとき，この s がどのような領域のコーパスに出現したとしても w の語義が変化するとは考えられない．この点から Kikuchi は WSD の領域適応の問題は共変量シフトの問題と見なせることを示し，共変量シフトの問題の解法手法を利用して WSD の領域適応の解決を図った (Kikuchi and Shinnou (2013))．そこではデータ x に対して密度比 $r = P_T(x)/P_S(x)$ を測り，その r を x の重みとして学習する．密度比はデータ x の重要度を表しているので，ここでは密度比の小さなものを Misleading データと見なすことにする．

密度比の測り方だが，ここでは論文 (Kikuchi and Shinnou (2013)) で示された方法を使う．

対象単語の w の用例 $\boldsymbol{x}$ の素性リストを $\{f_1, f_2, \cdots, f_n\}$ とする．求めるのは領域 $R \in \{S, T\}$ 上の $P_R(\boldsymbol{x})$ である．まず以下を仮定する．

$$P_R(\boldsymbol{x}) = \prod_{i=1}^{n} P_R(f_i)$$

領域 R のコーパス内の w の全ての用例について素性リストを作成し，素性 f の頻度を $n(R, f)$ とおく．また素性の総頻度を $N(R)$ とおく．つまり $N(R) = \sum_{f \in R} n(R, f)$ である．次に領域 S と領域 T における w に関する素性の種類数を M とする．

$P_R(f)$ を以下で定義する．

$$P_R(f) = \frac{n(R, f) + 1}{N(R) + M}$$

4　実験

実験では BCCWJ コーパスの OC(Yahoo!知恵袋) と PB(書籍) を用いる．両領域で頻度 50 以上の表 1 に示す 17 単語を対象単語として，OC から PB，PB から OC の 2 通りの WSD の領域適応を行う．

表 1: 対象単語

単語	辞書上の語義数	PB での頻度	PB での語義数	OC での頻度	OC での語義数
言う	3	1114	2	666	2
入れる	3	56	3	73	2
書く	2	62	2	99	2
聞く	3	123	2	124	2
来る	2	104	2	189	2
子供	2	93	2	77	2
時間	4	74	2	53	2
自分	2	308	2	128	2
出る	3	152	3	131	3
取る	8	81	7	61	7
場合	2	137	2	126	2
入る	3	118	4	68	4
前	3	160	2	105	3
見る	6	273	6	262	5
持つ	4	153	3	62	4
やる	5	156	4	117	3
ゆく	2	133	2	219	2
平均	3.35	193.9	2.94	150.6	2.88

ソース領域の各データについて密度比を測り，その値が閾値 θ 以下のものを Misleading データと見なし，訓練データから Misleading データを除いた．この訓練データから SVM で学習し，語義識別の平均正解率 (%) を調べた．結果を表 2 と表 3 に示す．表 2 は OC から PB の領域適応であり，表 3 は PB から OC の領域適応である．表の「そのまま」は Misleading データの検出を行わずに，訓練データをすべて利用した場合の識別に対応する．

実験は $\theta = 0.1$ と $\theta = 0.05$ で試したが，どちらのケースにおいても Misleading データを除いて学習する効果はなかった．

表 2: 実験結果 (OC → PB)

単語	そのまま	Misleading を削除	
		$\theta = 0.1$	$\theta = 0.05$
言う	78.37	80.61	79.80
入れる	71.43	67.86	71.43
書く	67.74	80.65	85.48
聞く	64.23	65.04	65.04
来る	97.12	97.12	97.12
子供	31.18	30.11	26.88
時間	87.84	85.14	85.14
自分	92.21	84.74	86.69
出る	59.21	59.87	62.50
取る	27.16	27.16	27.16
場合	85.40	82.48	85.40
入る	46.61	45.76	36.44
前	78.13	74.38	73.75
見る	82.78	83.52	82.78
持つ	78.43	69.93	69.93
やる	92.95	92.95	92.95
ゆく	88.72	87.22	87.22
平均	72.32	71.44	71.51

表 3: 実験結果 (PB → OC)

単語	そのまま	Misleading を削除	
		$\theta = 0.1$	$\theta = 0.05$
言う	79.13	79.43	80.63
入れる	72.60	71.23	49.32
書く	73.74	73.74	73.74
聞く	67.74	58.06	56.45
来る	79.89	79.89	79.89
子供	23.38	46.75	25.97
時間	83.02	83.02	83.02
自分	87.50	87.50	87.50
出る	67.18	51.15	66.41
取る	37.70	32.79	34.43
場合	86.51	84.13	85.71
入る	57.35	45.59	45.59
前	86.67	86.67	86.67
見る	55.73	56.49	56.49
持つ	83.87	46.77	59.68
やる	94.02	94.02	93.16
ゆく	68.49	68.49	68.49
平均	70.85	67.40	66.66

5　考察

5.1　Misleading の存在

実験では Misleading データを除いて学習を行う効果は確認できなかった．原因は Misleading データが存在しない，あるいは密度比では Misleading データを検出できない，のいずれかである．

本節では，上記の点を明らかにするために，総当たり的に各データが Misleading データかどうかを調べることで Misleading データの存在を調べる．具体的には先の実験で使用した各々の訓練データ $D = \{d_1, d_2, \ldots, d_N\}$ について i 番目のデータ d_i を除いた訓練データ $D_i = \{d_1, d_2, \ldots, d_{i-1}, d_{i+1}, \ldots, d_N\}$ を用意し，D_i を用いて学習した際の正解率を調べ，正解率が上昇するデータを Misleading データとした．このようにして設定した Misleading データを除いて学習を行った結果を表 4，表 5 に示す．OC から PB また PB から OC のどちらの領域適応においても，平均正解率の向上が確認でき，Misleading データは存在すると結論づけられる．

表 4: Misleading データの存在確認実験 (OC → PB)

単語	そのまま	Misleading を削除	Misleading の個数 (データ数)
言う	78.37	82.50	159 (666)
入れる	71.43	75.00	6 (73)
書く	67.74	82.26	21 (99)
聞く	64.23	73.98	26 (124)
来る	97.12	97.12	0 (189)
子供	31.18	39.78	5 (77)
時間	87.84	90.54	1 (53)
自分	92.21	97.08	13 (128)
出る	59.21	64.47	14 (131)
取る	27.16	32.10	6 (61)
場合	85.40	85.40	0 (126)
入る	46.61	45.76	36 (68)
前	78.13	84.38	8 (105)
見る	82.78	84.98	10 (262)
持つ	78.43	72.55	8 (62)
やる	92.95	92.95	0 (117)
ゆく	88.72	88.72	17 (219)
平均	72.32	75.86	19.41 (150.59)

表 5: Misleading データの存在確認実験 (PB → OC)

単語	そのまま	Misleading を削除	Misleading の個数 (データ数)
言う	79.13	82.43	127 (1114)
入れる	72.60	68.49	19 (56)
書く	73.74	73.74	0 (62)
聞く	67.74	68.55	26 (123)
来る	79.89	80.42	1 (104)
子供	23.38	46.75	12 (93)
時間	83.02	83.02	0 (74)
自分	87.50	87.50	0 (308)
出る	67.18	64.89	39 (152)
取る	37.70	42.62	10 (81)
場合	86.51	91.27	7 (137)
入る	57.35	60.29	38 (118)
前	86.67	89.52	10 (160)
見る	55.73	56.87	3 (273)
持つ	83.87	85.48	2 (153)
やる	94.02	94.02	0 (156)
ゆく	68.49	70.32	15 (133)
平均	70.85	73.31	18.18 (193.94)

5.2 密度比による Misleading の検出

前述の実験により Misleading データの存在は確認できた．次に密度比によりどの程度の Misleading データが検出できるのかを調べた．まず前述の実験により得られた Misleading データを検出の正解データと見なした場合，密度比による Misleading データの検出での再現率を表 7 に正解率を表 8 に示す．

表 7 より，PB から OC への領域適応の場合，密度比による Misleading データの検出の再現率は

高いが，OC から PB への領域適応の場合はあまり高くないことが確認できる．次に表 8 からは，どちらの方向の領域適応においても検出の正解率は低い．

検出の F 値は表 6 となる．

表 6: 密度比による Misleading データ検出の F 値

	OC → PB	PB → OC
$\theta = 0.1$	0.2087	0.2281
$\theta = 0.05$	0.1984	0.2185

本論文の実験結果（表 4 と表 5）から考えて，上記程度の検出能力では WSD の領域適応には効果が出ない．結論としては，密度比による Misleading データの判別は有用ではないと考える．

表 7: 密度比による Misleading データ検出の再現率 (%)

単語	OC → PB		PB → OC	
	$\theta = 0.1$	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.05$
言う	70.44	64.15	83.46	71.65
入れる	50.00	50.00	89.47	89.47
書く	85.71	80.95	-	-
聞く	57.69	53.85	88.46	65.38
来る	-	-	100.00	100.00
子供	80.00	60.00	83.33	75.00
時間	0.00	0.00	-	-
自分	69.23	61.54	-	-
出る	78.57	78.57	89.74	87.18
取る	66.67	66.67	100.00	100.00
場合	-	-	71.43	71.43
入る	75.00	69.44	78.95	71.05
前	62.50	50.00	90.00	80.00
見る	80.00	50.00	100.00	100.00
持つ	50.00	25.00	100.00	100.00
やる	-	-	-	-
ゆく	100.00	100.00	66.67	53.33
平均	66.13	57.87	87.81	81.88

表 8: 密度比による Misleading データ検出の正解率 (%)

単語	OC → PB		PB → OC	
	$\theta = 0.1$	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.05$
言う	25.63	26.02	11.76	11.45
入れる	4.76	5.17	32.69	34.00
書く	20.22	20.99	-	-
聞く	16.13	15.73	23.23	19.10
来る	-	-	1.03	1.08
子供	5.56	4.29	12.05	11.25
時間	0.00	0.00	-	-
自分	10.23	9.64	-	-
出る	10.38	11.22	25.36	25.37
取る	7.55	7.55	13.70	13.89
場合	-	-	4.24	4.50
入る	46.55	45.45	28.85	27.00
前	6.17	5.26	5.96	5.67
見る	3.77	2.58	1.25	1.38
持つ	7.69	4.08	1.44	1.50
やる	-	-	-	-
ゆく	8.81	9.60	8.93	7.69
平均	12.39	11.97	13.11	12.61

5.3　語義分布推定への影響

ここでは「Misleading データは存在する」と結論づけたが，これは新納の結果 (新納浩幸・佐々木稔 (2013)) と矛盾するものではない．論文 (新納浩幸・佐々木稔 (2013)) の主張からは「Misleading データは存在しない」ことになるが，これは「語義分布の推定に影響を与えなければ」という前提が存在する．本論文で発見できた Misleading データを除くことで訓練データ中の語義分布が変化している可能性がある．この点は早急に確認したい．

6　おわりに

本論文では WSD の領域適応における Misleading データの存在と検出について論じた．密度比の低いものを Misleading データと見なす方法を試したが，効果はなかった．総当たり的に各データが Misleading データと見なせるかどうかを調べることで，Misleading データの存在を確認できた．また密度比を利用した Misleading データの検出については，その検出能力が低いことも示した．今後はまず Misleading データの削除により語義分布に変化が生じているかどうかの確認する必要がある．また精度の高い Misleading データの検出法も考えたい．

文献

Jing Jiang and Chengxiang Zhai (2007) “Instance weighting for domain adaptation in NLP,” in *ACL-2007*, pp. 264–271.

Hironori Kikuchi and Hiroyuki Shinnou (2013) “Domain Adaptation for Word Sense Disambiguation under the Problem of Covariate Shift,” 情報処理学会自然言語処理研究会報告，pp.NL–212–4.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich (2005) “To transfer or not to transfer,” in *NIPS 2005 Workshop on Transfer Learning*, Vol. 898.

Hidetoshi Shimodaira (2000) “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, Vol. 90, No. 2, pp. 227–244.

新納浩幸、佐々木稔 (2013)「k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応」，情報処理学会自然言語処理研究会報告，pp.NL–211–13.

神嶌敏弘 (2010)「転移学習」，人工知能学会誌，第 25 巻，第 4 号，pp.572–580.

杉山将 (2006)「共変量シフト下での教師付き学習」，日本神経回路学会誌，第 13 巻，第 3 号，pp.111–118.

日伊コロケーション辞書の作成を目指す『現代日本語書き言葉均衡コーパス』からのコロケーションの検出と分析

STRAFELLA Elga Laura（奈良先端科学技術大学院大学）
松本裕治（奈良先端科学技術大学院大学）

Towards a Japanese-Italian Collocation Dictionary: Detection and Analysis of Collocations from the *Balanced Corpus of Contemporary Written Japanese*

Elga Laura Strafella (Nara Institute of Science and Technology)
Yuji Matsumoto (Nara Institute of Science and Technology)

１．はじめに

本研究では、日伊コロケーション辞典を編集するために、コロケーション検出に広く用いられる単純頻度、相互情報量 (PMI)、対数尤度比 (Log-Likelihood Ratio)、ダイス係数 (Weighted Dice) という4つの指標を使用し、大規模な『現代日本語書き言葉均衡コーパス』(BCCWJ) からコロケーションを検出しながら、文法と意味解析を行う。データセットとしては、日本語能力試験の過去の出題基準となっていた１級の語彙リストを用い、コロケーションを検討する。統計的なアプローチで BCCWJ から連語 （「名詞＋動詞」「名詞＋イ形容詞」）を検出した上、統語論・意味論の特徴によって各語句の分類を行う。

次に、本研究の背景（２章）について述べてから、分析の流れと検出方法に関する特徴について簡潔に説明する。また、検出したコロケーションのいくつかの例を挙げて、それらの意味と例文も挙げる（３章）。最後に、まとめと今後の課題について述べる（４−５章）。

２．研究の背景

本研究は日本語における「コロケーション」の使い方を身に付けるために行なっている。日本語学習者は、何かを言いたいと思って辞書を引き、言いたいことに当たる名詞を見つけても、その名詞と共にどの動詞を使えばよいか分からなければ困るのであろう。現在に存在している日本語辞典には名詞に連続する動詞の用例が載っていないし、載っていても例が短いため、実際の用法が分かりにくいことが多い。例えば、'Have you **made** any **plans** for the summer vacation? ' と聞きたいと思っても、「計画」の後にどんな動詞を使えばよいのか分からないと、「夏休みの**計画を**何か**しました**か」「夏休みの**計画を**何か**作りました**か」と言ったりしてしまいがちである。しかし、ここでは「計画を立てる」というコロケーションを使わないといけないのである。また、「頭」には「頭が**いい**」という形容詞の他に「頭が**悪い**」「頭が**固い**」「頭が**古い**」などのコロケーションがあり、これらを使うことによって日常生活で的確な表現をすることができる。こうした知識を得るために、まずコロケーションを整理する必要がある。

第３章「分析の流れ」は、検索コーパスやデータセットの設計などについて詳しく述べる。

３．分析の流れ

データセットとしては、日本語学習者の初級者にも役に立つ資料を編集するため、日本語能力試験の過去の出題基準となっていた１級の語彙リスト（約8,000語）を用い、『現代日本語書き言葉均衡コーパス』を検索した。一般の会話だけではなく、文学書や新聞などでも使われている現代日本語を検討するため、均衡コーパスを使うことにしたのである。

３.１　抽出

はじめに、 BCCWJ からの係り受けを抽出した。具体的にいうと、 BCCWJ Dependency Extraction Toolkit [1] というプログラムで CaboCha [2] で解析した結果から、「名詞が動詞またはイ形容詞に係っている事例」を抽出した。「名詞＋格助詞＋動詞」という係り受けは 385,470 あり、 「名詞＋格助詞＋イ形容詞」という係り受けは 13,353 あった。それから、それぞれの係り受けの頻度と PMI、Log-Likelihood Ratio、Weighted Dice を計算した。以前の研究でも明らかになった通り、コロケーションの検出ではよく使われているいくつかの統計的な指標は語と語の結びつきの強さについてそれぞれの特徴を現すので、一つだけを使うと分析結果を信頼しにくくなる。[3] 頻度だけでも判断してしまうと役に立てないデータが多くなる。例えば、「方が良い」(f = 7,012)、「方が多い」(f = 845)、「方が高い」(f = 254) などはとても高い頻度があっても、日本語としてはそのままには使えないと考えられる。つまり、その前に来る表現によって伝えられる意味が変わってしまう。例えば、「方が良い」の場合、「寝た**方が良い**」と言えるし、「出た方が良い」とも言え、伝えている意味は違う。

しかし、ここで連続している名詞「方が」とそれぞれの形容詞「良い」「多い」「高い」の単純頻度は比較的に高いため、一緒に現れる頻度も多い。こういう偏ったデータを避けるために各コロケーション候補に対して、今回利用した３つの共起尺度 (PMI, Log-Likelihood Ratio, Weighted Dice) による共起度の順番をもとめ、その逆数の平均値を示す Mean Reciprocal Rank (MRR) [4] を計算し、分析を進めた。 MRR によって各連語をソートすると高い数値を示すものは３つの共起尺度によって総合的にコロケーションである可能性が高いと支持された表現であると言える。

３.２　分析

抽出したデータの分析は一つずつ手動で行われているため、時間が非常に掛かる。それで、以前の研究 [5] では「名詞＋格助詞＋動詞」しか扱われてなかったため、今回は「名詞＋格助詞＋イ形容詞」だけを分析した結果について述べる。

まず、関連研究に基づいていわゆる frequency threshold [6] を適用することにした。広く用いられるのは $f \geq 3$、$f \geq 5$、$f \geq 10$ であり、データを削減するために一番高い数値を適用し

---

[1] 解析プログラムは奈良先端科学技術大学院大学　自然言語処理学研究室　博士後期課程の林部祐太に作られた.

[2] code.google.com/p/cabocha/

[3] Strafella, 2013.

[4] en.wikipedia.org/wiki/Mean_reciprocal_rank

[5] Strafella, Hayashibe, Matsumoto, 2012.

[6] Evert, 2007: 5.

た結果、997 の事例を得た。これを一つずつ分析し、統語論・意味論の特徴によって分類を行った。表１には7つの名詞に関する例を挙げている。MRR の一番高い数値は太字になっている。

表１データサンプル

連語	頻度	PMI	Log-Likelihood Ratio	Weighted Dice	MRR
数が多い	455	1.764420904	1439.231531	0.316377828	0.5
数が少ない	348	2.984089786	2216.775566	0.87863168	**1**
数が大きい	18	0.186751859	0.86599955	0.025617287	0.305555556
数が高い	15	-0.390397562	3.845040344	0.014466393	0.277777778
事が多い	339	1.499015408	808.4376951	0.224693869	**0.833333333**
事が良い	31	-1.059835123	78.46334711	0.010324712	0.388888889
事が嬉しい	14	2.089499352	49.58983067	0.061022288	0.611111111
運が良い	304	1.876938586	1106.195885	0.171873048	0.666666667
運が悪い	119	2.469259606	566.2383853	0.241212205	**0.833333333**
運が強い	15	0.852461806	12.2699912	0.025782384	0.333333333
感じが良い	117	0.882616398	112.1124388	0.055053262	**0.833333333**
感じが強い	20	1.100675457	25.47720714	0.037820416	0.666666667
感じが悪い	18	0.541039449	6.540102027	0.021985545	0.333333333
幅が広い	95	4.930777293	1134.423075	1.197958316	**1**
幅が大きい	46	2.196344774	180.0766338	0.100587431	0.305555556
幅が狭い	40	4.196325143	380.3666159	0.440739387	0.5
幅が小さい	16	2.311900078	66.07301857	0.068965517	0.277777778
頭が良い	253	0.915291199	259.4077687	0.135060678	0.25
頭が痛い	156	3.715031998	1291.841864	0.938112048	**1**
頭が悪い	82	1.318846185	143.5533177	0.138391098	0.261111111
頭が可笑しい	65	3.27056167	445.6198181	0.382279207	0.5
頭が重い	26	2.461868949	118.0045722	0.123633215	0.194444444
頭が固い	24	2.947460102	141.207311	0.128926889	0.244444444
頭が大きい	12	-0.314097396	1.929147147	0.014380595	0.130952381
頭が高い	10	-0.891246817	16.0490946	0.008077637	0.136904762
気が重い	94	3.407324739	688.5074666	0.488218187	**1**
気が強い	60	1.121004539	79.25590641	0.121707911	0.197619048
気が弱い	50	2.397068731	219.6111854	0.200706123	0.388888889
気が小さい	40	1.721740933	105.8618708	0.127700734	0.261111111
気が早い	30	1.330232622	52.29419161	0.084334986	0.163888889
気が短い	24	1.897810298	73.7165589	0.083967264	0.186507937
気が良い	20	-1.96210847	252.43963	0.005676478	0.25
気が荒い	12	3.161715443	78.3686147	0.042998051	0.26984127

表１に示してあるように、MRR の一番高い数値を示す連語はコロケーションであり、単純頻度と違っている場合もある（「運が悪い」「頭が痛い」）。

関連研究においてはデータをソートするためによく使われる方法がもう一つあり、１つの指標だけを選んでソートすることである。しかし、本研究は３つの指標を使っているため、それらを総合的に考慮するため MRR を用いることにした。

また、それぞれの指標を検討してみると数値がばらばらになっているため、統計的な解析ができないことが分かる。つまり、コロケーションの研究には統計的なアプローチでデータを得られるが、その後研究者の判断が必要になるため、手動で分析をする必要がある。次に、我々が分析した結果について述べる。

３.３　結果

上記に述べた 997 の「名詞＋格助詞＋イ形容詞」の連語を分析した結果、約 300 のコロケーションが得られた。これの中には「拡張語彙意味単位」と言えるものもあり、特別な意味を持っていないが、よく使われる表現もある。前者は、「頭が高い：傲慢な、誇りを持って」「腕が良い：有能な、上手な」のように全体的な意味は語と語の意味から少し離れていて、比喩的な意味を表す表現である。後者は、「水が欲しい」「車が欲しい」「山が多い」「火山が多い」のように日常生活でもよく使われる表現である。

はじめに述べた通り、本研究の最終的な目的は日本語の初級者にも役に立つ資料を作ることであるため、それぞれのコロケーションの実際の用法が分かるために用例を挙げないといけない。それは、『国立国語研究所』が構築した NINJAL-LWP for BCCWJ [7] という BCCWJ を検索するためにオンライン検索システムを用いて挙げたいと思っている。具体的な例を挙げると、上記に挙げた「**腕が良い**」の用法を表示する例としては、

a.「板前の**腕が**めっぽう**いい。**」

b.「その年で自分のお店を始められるなんて、よっぽど**腕がいい**んだ。」が挙げられる。

また、「**頭が高い**」の場合、

a.「**頭が高い**からお客さんがつかない。」

b.「**頭が**やや**高く**、少しイビツなフォーム。」という例が挙げられる。

この4つの用例はコロケーションの一つの特徴を示している。つまり、コロケーションはイディオムと違い、文型が固定してないため、中心語と関連する語のあいだに別の語が入ることが可能である。この例では、入っている語は副詞である。

こうした統語論・意味論の分析を行うと、それぞれのコロケーションの実際用法はもっとも分かりやすくなり、学習者に役に立つ情報も得ることができる。

４．まとめ

本稿で行った調査は博士課程の主要議題であり、2010年に始まった。はじめに、関連研究で Multiword Expressions の検出にあたって共起度を測るためによく使われるいくつかの指標を深く調べ、その中で四つ（単純頻度、PMI、Log-Likelihood Ratio、Weighted Dice）を用いることにした。しかし、連語といってもコロケーションだけでなく、自由結合・イディオムなども存在しているため、その中でコロケーションだけを検出するのはコンピュー

[7] nlb.ninjal.ac.jp/

タを使っても複雑なタスクであることが明らかになった。従って、統計的なアプローチでコーパスから強く結びついている連語を検出し、手動で一つずつ分析を行っている。こうして辞典の基になるコロケーションリストを作っている。

５．今後の課題

本研究は、日本語を学習している学習者達に役に立つ資料を作ろうとしているため、「名詞＋イ形容詞」だけでなく、「名詞＋ナ形容詞」「名詞＋動詞」「副詞＋動詞」なども検討する必要がある。今後はそれぞれの文型を検討し、コロケーションリストを完成することを考えている。しかし、データの量が多くなければ多くなるほど分析しにくくなり、それを自動的に分析ができる方法を考える必要がある。また、網羅性の高いコロケーションリストを作成したら、各コロケーションの用例も挙げて、イタリア語に翻訳をしたり、当のコロケーションとの同じ意味を表すイタリア語の表現を挙げたりすることを考えている。

謝辞

本研究は、日本語学術振興会「外国人特別研究員（欧米短期）」（平成 24～25 年度）による補助を得ている。

本研究の推進にあたっては、共同研究者各位、とくに自然言語処理学研究室博士後期課程林部祐太に多くのご指導・ご助言をいただいた。感謝を申し上げます。

文献

Evert, Stefan (2005) *The Statistics of Word Co-occurrences: Word Pairs and Collocations,* Ph.D. thesis submitted to the Institut für machinelle Sprachverarbeitung, Universität Stuttgart.

Evert, Stefan (2007) *Corpora and collocations*, Institute of Cognitive Science, University of Osnabrück, Germany. (Extended Manuscript, 13 October 2007).

姫野昌子 (2012) 『研究社日本語コロケーション辞典』研究社.

Pereira, Lis W.K., Manguilimotan, Erlyn, and Matsumoto Yuji (2013) *Collocation Suggestion for Japanese Second Language Learners*, 情報処理学会研究報告　第 210 回自然言語処理研究、Vol.2013-NL-210, No.3, pp. 1-5, January 2013.

Pereira, Lis W.K., Maguilimotan, Erlyn, and Matsumoto Yuji (2013) Data Coverage vs. Data Size: A comparison of two large-scale corpora in Collocation Suggestion for Japanese Second Language Learners. In *Proceedings of the Nineteenth Annual Meeting of the Association or Natural Language Processing (NLP-2013)*, Nagoya, Japan, March 2013.

Seretan, Violeta (2010) *Syntax-Based Collocation Extraction*. In N. Ide and Véronis J. (eds.) *Text, Speech and Language Technology*, vol. 44, Springer Dordrecht Heidelberg London New York.

Strafella, Elga Laura (2013) *Detection and Analysis of Collocations in Contemporary Japanese – A corpus-based language study*, Ph.D. thesis submitted to the Department of Asian, African, and Mediterranean Studies, University of Naples “L’Orientale”.

Strafella, Elga Laura, 林部祐太、松本裕治 (2012) 『現代日本語におけるコロケーション：検出と分析』、第 1 コーパス日本語学ワークショップ、 pp. 53-58、国立国語研究所、5-6 March 2012.

関連URL

『現代日本語書き言葉均衡コーパス』(BCCWJ)　http:// www.ninjal.ac.jp/corpus_center/bccwj/

BCCWJ Dependency Extraction Toolkit　https:// github.com/shirayu/misc/tree/master/bccwj

Mean Reciprocal Rank (MRR)　http://en.wikipedia.org/wiki/Mean_reciprocal_rank

MeCab　http:// mecab.googlecode.com/svn/trunk/mecab/doc/index.html

NINJAL-LWP for BCCWJ (NLB) 国立国語研究所　http:// nlb.ninjal.ac.jp/

NDL Searchによるジャンル名の分析

浜田 秀（天理大学文学部）†

Analysis of Genre Names Based on the National Diet Library Search

Shu Hamada (Faculty of Letters, Tenri University)

１．はじめに

ジャンル名の用法はおおむね同時代の存在として実践的に使用されるもの（実践的用法）と、歴史的存在として客体的に使用するもの（歴史的用法）に分けることができる。たとえば「新体詩」という用語を目次や作法書に使用するのは実践的用法、文学史の記述に使用するのは歴史的用法である。実践的用法の存在は、そのジャンル名が積極的にジャンル構築に関わっていることを示す。

NDL Search（国立国会図書館サーチ）は国立国会図書館のHPの検索サービスであるが、コーパスとして使用することで、ジャンル名の用法の変遷がたどることができる。NDL Searchの検索結果を検討すると、「口語詩」という用語が1940年代に歴史的用法へと推移したことが分かる。

２．ジャンル名の推移について

明治から大正・昭和初期にかけてはジャンルの激動期である。ジャンル名の意味するところにもまた変動があった。

現在「詩」と言えば、漢詩や短歌、俳句から区別される、口語自由詩としての詩形を通常思い浮かべる。つまり、「口語詩」「自由詩」と取り立てて言わずとも、単に「詩」と言えば通じるのである。

現在の「詩」の語義は、このプロトタイプ[1]としての口語自由詩を指す用法と、漢詩や短歌、俳句を含めた詩歌の総称としての用法の二義を中心とする。

だが「詩」という用語は近世までは中国の古典詩形、すなわち漢詩を意味した。「漢詩」という語が使用されるようになったのは明治以降である。「詩」という語彙は、いつプロトタイプ的用法としての「口語自由詩」という意義を獲得したのであろうか。また「口語詩」という用語は、いつその実践性を失ったのであろうか。

文学史の記述は、「口語詩」という用語は1907年（明治40年）5月に、人見東明によって使用されたのが初出であること、作品としての口語詩は実質上同年9月の川路柳虹の「塵溜」に始まることを教える。当時、五七調文語文で書かれた「新体詩」から口語自由詩を区別して表現する用語として「口語詩」「自由詩」「言文一致詩」などが乱立した。しかしながら、「口語詩」が単に「詩」と呼ばれるようになった事は、語義の変化であっても事象の変化としては捉えられないため、文学史家の注意をさほど引かないようである。

ジャンル名の変遷は、言語の変化であると同時にそれを使用する一般の人々の常識の変化を示すものでもある。本調査では近代文体成立期のめまぐるしいジャンル名の変遷に定量的にアプローチする道を探りたい。

３．ジャンル名の用法

本調査では、ジャンル名「口語詩」を（a）歴史的用法、（b）理論的用法、（c）所属的

† s-hamada@tenri-u.ac.jp

1 認知意味論におけるプロトタイプ概念については、レイコフ(1993)、テイラー(2008)を参照。

用法の三種に分けた。

（a）歴史的用法とは、過去の歴史的事象としての「口語詩」に言及するものである。
　　例：口語詩は、1907 年に川路柳虹によって始められた。
（b）理論的用法とは、「口語詩」の一般的性格に言及するものである。
　　例：口語詩は、心のままに書かれるべきである。
（c）所属的用法とは、当該のテクストがどのジャンルに属するのかを決定するものである。たとえば、目次や、題名の横に「口語詩」と書かれる例が相当する。
　　例：口語詩　哀れにもまた勇ましき古い合戦の物語　星野水裏

（b）の理論的用法は、「口語詩」という語彙が何を意味するのかについて主張するものであり、我々の語義に対する常識を構築する行為であると言える。また（c）の所属的用法は当該の作品がどのジャンルに所属するのか、その社会的コンテクストを受手に明示し、そのようなものとして受手に提示する、遂行的表現である。いずれも「口語詩」という語彙が、他のジャンルとの区別を表現するために生きて働いていることを示す。両者を「実践的用法」と一括することができる。

（a）歴史的用法
（b）理論的用法　｝実践的用法
（c）所属的用法

ただし（a）と（b）の区別はしばしば曖昧である。

（1）　口語詩は俗語で書かれた詩である。その発生は 1908 年にさかのぼる。
（2）　口語詩は俗語で書かれた詩である。卑俗にならぬよう注意する必要がある。

同一の文であっても、前後のコンテクストによって(1)は歴史的用法、(2)は理論的用法として区別される必要がある。

（c）の所属的用法は、いくつかの特徴を持つ。第一は、これはテクストそのものの一部ではなく、テクストに付随するという特殊な形式をもつ表現であるということである。ジュネット(2001)は、このようなタイトル・作者名・献辞・序文・挿絵・奥付など一連のテクストに伴う存在のことを「パラテクスト」と呼んだ。ジャンル名の所属的用法はパラテクストの機能の一部として考えられる。

第二は、これが文ではなく、語として提示され、あるテクストのジャンル所属を決定するという言語行為を遂行しているということである。ジュネット(2001)はタイトルにジャンル指示的機能を持つものがあることを指摘しているが、これは言語行為の一種として考えることができる。単語がそのままで言語行為を遂行するということは書記言語には豊富に見られる現象である。

第三は、これがテクストの社会的コンテクストを構成するものとして提示されるということである。テクストにアプローチする読み手は、これによってジャンルを知らされ、ジャンルの制度に従って読解を行うことになる。

第四は、これが読み手には、テクストと切り離せないものとして提示されるということである。題名とは異なり、ジャンル名は編者によって作者の意図に反したものが付けられる可能性もあるが、読者には、書籍に埋め込まれたテクストと不可分のものとして提示され、テクストをそのジャンルに所属するものとして受け取らざるを得ないということになる。

４．NDL Search とは

NDL Search（国立国会図書館サーチ）とは、7300 万件の文献情報、200 以上のデータベースと連携した統合検索サービスである。本格的な稼働は 2012 年 1 月であり、現在も発展し続けている。

NDL Search は、他の検索サービスと連携しており、ここには NDL-OPAC、ゆにかねっと、CiNii Articles、CiNii Books、国立国会図書館デジタル化資料などが含まれる2。

NDL Search の利点の一つは、デジタル化された資料にリンクが貼られていることにより、現物を確認可能だということである。これは歴史資料を再入力したコーパスでは実現できないことである。現物を確認することで、誤記入の確認や、用法の判定が容易になる。

資料のデジタル化は急速に進んでいる。廣瀬（2013）によれば、2012 年 3 月末の時点で、和図書 436 万冊のうち 90 万冊、和雑誌 431 万冊のうち 112 万冊、全所蔵資料 965 万冊のうち 225 万冊のデジタル化が進んでいる。和図書については、明治期・大正期に刊行された図書のデジタル化をほぼ終え、1968 年に受け入れたものまではデジタル化を完了したということである。

現在国会図書館のデジタル化資料は「国立国会図書館デジタル化資料」や「近代デジタルライブラリー」のように外部から閲覧可能なものと、館内でのみ閲覧できるものに分かている。著作権法の改正により、2014 年 1 月から他の図書館に送信サービスが始まる。これは国会図書館の館内閲覧資料のうち、入手困難な資料、雑誌、博士論文などについて、他の図書館から閲覧可能になるサービスである。ただし、現在も市場で流通しているものは申し出によって閲覧ができなくなる場合もある（廣瀬 2013）。

５．NDL Search の規模

NDL Search は言語分析のために設計・収集されたものではなく、狭義のコーパスには当てはまらない。これをコーパスとして使用するためには、当然検証が必要となる。ここではまずコーパスの規模について考える。なお調査結果はいずれも 2013 年 7 月のものである。

田野村(2009)によれば、「新潮文庫の 100 冊」が 2000 万字、BCCWJ2008 の書籍データが 6000 万字である。文献情報 7350 万件との単純な比較は難しいが、コーパスとして使用可能な規模の言語量はあると思われる。

ところで、本調査が対象とする「口語詩」という用語についてはどうであろうか。「口語詩」で検索すると 1450 万字あるとされる「太陽コーパス」では 0 件、「少納言」でも 1 件しかヒットせず、これらのコーパスでは分析に耐える用例数を集めることができない。

一方、NDLsearch による簡易検索では 322 件ヒットする。yahoo のサーチエンジンで引用符に入れて検索すると 9740 件のヒットであるから、Web 全体に対しても約 30 分の 1 程度の検索結果を得られたということになる（ただし、NDL Search によるヒットするは検索語を含む文献数であって検索語の度数ではない）。田野村(2009)によれば、Web 上の日本語文書は 26 兆字ということであるから、単純に考えると 1 兆字弱の web 情報から検索するのと同様の検索結果を入手できたわけである。これはかなり効率がいいといえる。

なぜこのような現象が生じたのであろうか。それは NDL Search が書誌情報から構成されているという特殊性に起因する。

NDL Search で検索できるのはあくまで書誌情報であり、「文」ですらない。助詞や助動詞といった機能語の検索に向かないのは言うまでもない。しかしながら、書誌情報であるからこそ含まれるものもある。それはジャンル名、とくにその所属的用法である。

NDL Search のデジタル情報化資料には、目次情報が含まれる。これにより、本調査で言うところの所属的用法に相当するものが検索され、ヒットするわけである。NDL Search がジャンル名のコーパスとして使用可能な所以である。

---

2 詳細は国立国会図書館ホームページの「検索対象データベース一覧 http://iss.ndl.go.jp/information/target/」を参照されたい。）

なお、国会図書館の HP による「口語詩」の検索結果を示す。

国立国会図書館デジタル化資料　9 件

NDL Search　簡易検索　322 件
　検索結果の内「デジタル資料」　239 件
　検索結果の内「国立国会図書館デジタル化資料」　237 件

NDL Search　簡易検索＋「すべての連携先を検索」　445 件
　検索結果の内「デジタル資料」　295 件
　検索結果の内「国立国会図書館デジタル化資料」　237 件

国会図書館の HP「国立国会図書館デジタル化資料」は NDL Search 内の「国立国会図書館デジタル化資料」と異なり、外部から閲覧できるもののみが検索される。ここには「近代デジタルライブラリー」が含まれる。検索結果は 9 件であり、送信サービス開始後はこの状況は改善されると思われるが、現時点では数が少なすぎる。

また NDL Search で「すべての連携先」にチェックを入れると、322→445 といささか件数は増える（ただし「国立国会図書館デジタル化資料」については同数である）。しかし、その内容は、インターネット資料収集保存事業（WARP）によるネット上のものや、チェックを外した検索結果と重複の多い「国文学研究資料館国文学論文目録」などであって、これをあえて含める意味はあまりない。

コーパスの内実が十分に明らかになっていない現時点では「口語詩」を調査するためには、国会図書館に赴き、デジタル化された資料を確認するのがのぞましい。総数 322 件に対して「デジタル資料」239 件はさほど遜色のない数字である。

ところでこの「デジタル資料」と「国立国会図書館デジタル化資料」とは概念の相違がある。「デジタル資料」はデータベースの種別を意味しており、国会図書館内で全て確認できるとは限らない。たとえば CiNii の情報も出てくるが、CiNii 自体がデジタルデータと考えられているためかリンク先に pdf が存在していないものも「デジタル資料」カウントされているようである。一方「国立国会図書館デジタル化資料」とは国会図書館でデジタル化した資料である。これは全て館内で閲覧可能であるが、CiNii で pdf のリンクを持つものも排除されてしまう。

「デジタル資料」のうち、「国立国会図書館デジタル化資料」は 237 件であり、それに含まれなかった 2 例は JAIRO、CiNii Articles のリンクから現物の確認が可能であった。結果として 239 点全ての確認が可能であった。

つまり、「口語詩」という用語については、総件数に比してデジタル資料が少ない、また「デジタル資料」が常に閲覧できるとは限らないという二つの問題をそれほど考慮に入れる必要はないと思われる。

６．NDL Search の通時的構成

歴史的検証に NDL Search を使用するためには、年度により資料の偏りがないかどうか確認する必要がある。NDL Search では検索結果に対して「出版年」ごとの内訳を見ることが出来る。これを利用してその通時的構成を調べた。

詳細検索画面で分類記号 9（文学）を検索したところ、1672930 件がヒットした。分類記号 9 の検索結果でもっとも古い「出版年」を示すものは 0002（紀元 2 年）の仙石廬元坊『花供養』と邵寶『刻杜少陵先生詩分類集註』であるが、「詳細情報」の「出版年月日等」欄を見るとそれぞれ「寛保 2（1742）序」「明暦 2 年（1656）」であり、出版年の「0002」は明らかに誤記入である。検索結果についてはある程度批判的に取り扱う必要があるようだ。

図 1 に明治初年（1868 年）以来の 9 門全体と、そのうち「デジタル資料」の該当数の年

度ごとのヒット件数を示す（2013 年は年度途中のために省略した）。なお検索時に「連携先」のチェックは外しているが、このチェックの有無で検索総数に変化は無かった。
経年で見ると、以下のことがわかる。

・総件数は漸増の傾向にある。特に戦後は激増している。
・デジタル資料も増加の傾向にあるが、総件数ほどではない。1968 年以降はデジタル化が進んでいないこともあり、件数が激減する。
・総件数・デジタル資料共に 1944 年、1945 年は激減する。

分類記号 9 の検索結果は、1868～1912 年で 1397002 件、うち「デジタル資料」152765 でおよそ 10.9%となる。これは 1968 年以降も総件数が増加するのに比して、デジタル資料がほとんどヒットしないせいである。1868～1968 で見ると、総数 489124 に対してデジタル資料 140545 件は 28.7%とおよそ 3 割に近くなる。1968 年までの資料に関しては、戦時の 2 年を除き、「デジタル資料」の検索でもそれなりの検索結果が期待できると思われる。

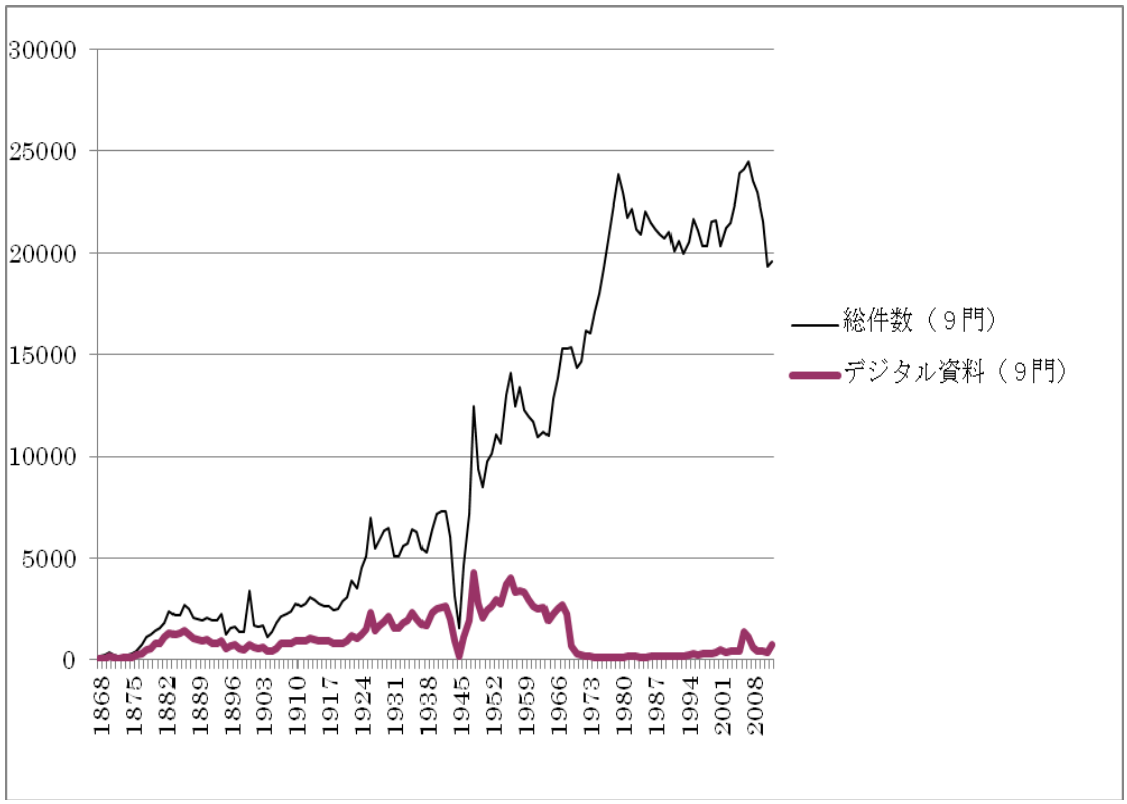


図1　NDL Search の通時的構成　（9 門　1868～2012）

７．NDL Search による「口語詩」調査

７．１　調査方法

NDL Search で「口語詩」に「デジタル資料」をチェックして簡易検索し、カウントされた 239 文献のうち、以下のものを除いた 229 件を対象とした。

・1969 年以降の 8 件
・同一年度に出版された明らかに同一の書籍 2 点

カウントの仕方・用法の判定は以下の基準によった。

・「口語詩」および「口語詩」を含む語彙（「口語詩問題」など）をすべて検索事例に含めた。「口語詩人」「口語詩集」は「口語詩＋人」「口語詩＋集」ではなく、「口語＋詩人」「口語＋詩集」であるが、内容に鑑み検索対象に含めた。
・書籍・雑誌の 1 点を 1 カウントとした。一つの文献内に、複数の「口語詩」を含む文献が含まれるケースも 1 カウントとした。複数の用法が併用している例は見られなかった。
・明らかに文語詩に付属しているものも件数に入れた。
・選集等に再録されているもの、年度を変えて再版されているものもその時点での用例として意味を持つため、件数とした。
・著者の死後の出版は、理論的用法のものであっても歴史的用法に入れた。
・ある章の題名の一部として「口語詩」が使用されている時には、著書そのものではなく、章のレベルをコンテクストとして用法を判定した。

７.２　結果

結果は表 1 のようになった。

実践的用法としての理論的用法と所属的用法の出現はほとんど重なっている。いずれも 1939～40 年を境に全く見ることができなくなる。

歴史的用法は実践的用法より 10 年ほど遅れて出現し、現在にいたる。現在でも使用されているという意味では「口語詩」という用語は決して古語ではないが、ジャンルを実践的に弁別・構築する機能は既に果たし終えていると言える。

歴史的用法の初出は 1919 年であるが、これはその年出された『抱月全集　第 1 巻』に「口語詩問題」（初出は 1908 年）というエッセイが収められたためである。死後の所収であり基準に従い「歴史的用法」に入れたが、前年の抱月の急逝を受けた出版であり、口語詩の発生からおよそ 10 年しか立っておらず、歴史的用法が発生したと考えるにはやや特殊な事例である。はっきりとした歴史的用法といえるのは 1924 年の 2 点以降である（白鳥省吾『現代詩の研究』「前期の口語詩論」「相馬御風の口語詩論」と福井久蔵『日本新詩史』「口語詩」）。口語詩発生から 17 年を経ており、それなりに現象が回顧可能となったと考えられる。

我々は現在口語自由詩を「詩」とも「口語詩」とも呼ぶことができるはずである。しかし、「口語詩」という用語は口語自由詩を一般的に論じる時には使用されず、歴史的な経緯を背景知識とした時にのみ使用されるようになってきているのである。

NDL Search による理論的用法の初出は 1908 年であり、文学史上の初出から 1 年しか遅れていない。これは当時口語詩をめぐって議論が激しく交わされ、文献が多く出されたためである。

驚くべきことに、用語および口語詩テクストの出現からほとんど間をおかず所属用法が出現している。NDL Search でヒットしたのは『中学文壇』『新文林』『少女の友』といった青少年向けの雑誌である。「口語詩とは何か」をめぐってまだ議論がされている最中に、「口語詩」という語彙はジャンル構築の力をこれらの層に対して持っていたのである。

理論的用方は 1908 年の 3 例を除き、『少女詩の作り方』「文章詩と口語詩」（水谷まさる 1922）、『新しい詩とその作り方』「口語詩と文語詩との区別」（室生犀星 1925）といった啓蒙的なものにみられるのが主流である。概念を巡る戦いが終わり、常識の啓発のためにこの用語が使用された。しかし、それも 1940 年ごろを境に全く見られなくなる。「詩」が「口語で自由に書かれたもの」という概念が常識となり、「口語」という用語で他ジャンルと弁別する必要が無くなっていったのだろう。

表 1　「口語詩」用法の通時的変化

年代	実践的用法		歴史的用法
	理論	所属	
1968 昭和43			2
1967 昭和42			
1966 昭和41			2
1965 昭和40			1
1964 昭和39			1
1963 昭和38			5
1962 昭和37			5
1961 昭和36			11
1960 昭和35			2
1959 昭和34			1
1958 昭和33			5
1957 昭和32			1
1956 昭和31			
1955 昭和30			4
1954 昭和29			6
1953 昭和28			6
1952 昭和27			
1951 昭和26			1
1950 昭和25			
1949 昭和24			1
1948 昭和23			
1947 昭和22			
1946 昭和21			
1945 昭和20			
1944 昭和19			1
1943 昭和18			
1942 昭和17			
1941 昭和16			1
1940 昭和15			
1939 昭和14		2	
1938 昭和13	1		

年代	実践的用法		歴史的用法
	理論	所属	
1937 昭和12	1		
1936 昭和11	1		
1935 昭和10			
1934 昭和9			1
1933 昭和8		1	2
1932 昭和7			
1931 昭和6			1
1930 昭和5	1		
1929 昭和4	1		
1928 昭和3		1	2
1927 昭和2			1
1926 昭和1			1
1925 大正14	1	3	
1924 大正13		9	2
1923 大正12		4	
1922 大正11	1	6	
1921 大正10		7	
1920 大正9		7	
1919 大正8		15	1
1918 大正7	1	16	
1917 大正6		16	
1916 大正5		7	
1915 大正4		11	
1914 大正3		11	
1913 大正2		4	
1912 大正1			
1911 明治44		5	
1910 明治43			
1909 明治42		20	
1908 明治41	3	6	
合計	11	151	67

８．まとめ

NDL Search を使用し、1908 年から 1968 年にわたる「口語詩」の用法を２類３種に分けた上で、その推移を分析した。

理論的用法・所属的用法といった実践的用法は、そのジャンル名が生きて働いており、他の諸ジャンルと区別するために積極的な力を持っていることを示す。

「口語詩」の使用においては、実践的用法が先行し、実践的用法・歴史的用法の共存の時期を経た上で、歴史的用法のみの時期が続く。現在の我々にとって、口語詩という用語は歴史的文脈を前提としたものへと変化してしまっている。

ジャンル名の用法は、このようなジャンル構造の変遷を示すものである。特にジャンル名用法の推移を調べることは、ジャンルに所属するテクストの属性の変化のみならず、ジャンルをカテゴライズする我々の知識の変化をも調べることになる。我々がいかにジャンル名を使いこなし、ジャンルを認知・構築・管理しているのか、ジャンル名の調査が教え

てくれるところは大きいと言える。

謝　辞

本調査は、平成 25 年度天理大学 学術・研究・教育活動助成費による研究結果の一部である。

文　献

国立国会図書館サーチパンフレット（日本語）
（http://iss.ndl.go.jp/information/wp-content/uploads/2013/05/pamphlet_japanese_201303.pdf よりダウンロード可能）
国立国語研究所編(2005)『太陽コーパス―雑誌『太陽』日本語データベース』博文館新社
ジュネット(2001)『スイユ　テクストから書物へ』水声社.
田野村忠温(2009)「コーパスからのコロケーション情報抽出―分析手法の検討とコロケーション時点項目の試作―」『阪大日本語研究』21、pp.21-41
テイラー(2008)『認知言語学のための 14 章　第三版』紀伊國屋書店.
服部嘉香(1963)『口語詩小史―日本自由詩前史―』昭森社.
浜田秀(2010)「ジャンル」『認知物語論キーワード』、pp.87-94、和泉書院.
人見円吉(1975)『口語詩の史的研究』桜楓社
廣瀬信己（2013）「国立国会図書館のデジタル化資料の図書館等への送信に向けて」『図書館雑誌』107(2)、pp.86-88
レイコフ(1993)『認知意味論：言語から見た人間の心』紀伊國屋書店.

関連 URL

NDL Search　簡易検索　http://dl.ndl.go.jp/
国立国会図書館デジタル化資料　http://dl.ndl.go.jp/

社会科教科書における教科特徴動詞の用法

―教科書コーパスと図書館コーパスの比較を通して―

阿保きみ枝（一橋大学大学院言語社会研究科）

Usages of Characteristic Verbs in Social Studies Textbooks

-Through a Comparison between Textbook Corpus and Library Corpus-

Kimie Abo(Graduate School of Language and Society, Hitotsubashi University)

１．本調査の目的

本調査の目的は、社会科教科書において特徴的に現れる動詞の用法の特性を探ることである。日本語を母語としない児童・生徒が学校教育での教科内容を理解する際には、教科書内の言語表現が壁となる場合がある。この壁を取り除くためには適切な指導や教材が必要であり、そのためには、まず、教科書の言語使用実態を的確に把握する必要があるだろう。

教科に特徴的な語というと、まずは科目ごとの専門語が挙げられる。一般的に専門語として意識されやすいのは名詞だろう。例えば、文部科学省によるＪＳＬカリキュラムでは、教科ごとの用語対訳一覧を公開しているが、そのほとんどが名詞で、専門的な語義を持つものである。

今回注目したいのは、このような専門語ではないが、ある特定の教科で頻出する語、つまり特徴語である。そのような語の振る舞いを調べることによって、気づかれにくい、教科ごとの言語使用実態の特徴が見えてくるのではないかと考えている。

最終的には、日本語を母語としない児童生徒が教科内容を理解するために必要な語を教科別にまとめ、頻出する用法・用例で提示する教材を作成することを目指している。

２．調査概要

調査に使用したのは、特定領域研究「日本語コーパス」言語政策班作成の「教科書コーパス」内の社会科ジャンル（以下「社会科教科書」）と『現代日本語書き言葉均衡コーパス』の図書館・書籍ジャンル（以下「図書館・書籍」）である。

「教科書コーパス」では特に校種を限定せず、小学校から高校までを対象とした。比較対象として図書館・書籍ジャンルを選んだのは、一般的に受容されている書籍を対象としているため言語表現が特定の分野・傾向に偏りにくく、一般的な言語使用を見るのに適していると考えたためである。

「教科書コーパス」の中で社会科のみの特徴語[1]となっている動詞を抽出した結果、以下の 18 語が取り出された。

[1] 教科書コーパスにおける特徴語の選定方法については、田中・近藤（2011）を参照。

表 1　社会科特徴動詞

	語彙素	社会科出現度数		語彙素	社会科出現度数
1	通ずる	253	10	強まる	108
2	巡る	239	11	改める	87
3	設ける	196	12	遂げる	82
4	強める	176	13	敗れる	63
5	唱える	168	14	衰える	61
6	広まる	162	15	重んずる	56
7	経る	123	16	押し進める	55
8	栄える	110	17	打ち立てる	18
9	滅ぼす	109	18	絡み合う	16

表 1 を概観すると、「栄える」「滅ぼす」などのように社会科教科書での用例が容易に想像できるものから、「絡み合う」のような社会科との関係がわかりにくいものまであることがわかる。これらの動詞は使用度数のみで抽出されたもので、用法に特徴があるかどうかは分からないため、各動詞の具体的な用例を見る必要がある。

用例が 50 例を超える語についてはランダムに 50 例を取りだし、用例を観察することとし、本稿では、これら 18 語の中から特徴的ないくつかの動詞について紹介する。

３．調査結果

３．１．事例 1「巡る」

『新明解国語辞典』で「巡る」の語義は以下の３つに分類されている。

> めぐ・る【巡る】（自五）
> 一　ぐるりとひと回りして元へもどる。「因果は ―― ／ ―― 年月」
> 二　〈どこヲ ―― 〉
> 目指す所を次つぎと訪ねてまわる。「名所旧跡を ―― 旅／知人の家を巡り歩く」
> 三　〈なに・だれヲ ―― 〉
> そのものを中心として物事が展開する。「彼を ―― 五人の女性／賛否をめぐって会議が紛糾する／消費税を ―― 論議」

この語義分類に従って、社会科教科書及び図書館・書籍での「巡る」の用例を語義別に筆者が分類した結果、表 2 のようになった。

表 2　「巡る」の用例分布

	社会科教科書	図書館・書籍
一	1	10
二	2	5
三	47	35
計	50	50

これらの語義の中で、三のみが複合辞的に用いられるものであり、一や二と異なる性質をもっている。よって、語義一と二を統合した上で比較すると、社会科教科書において複合辞的な用法が多いことがわかる（χ^2=9.8、p＜0.001）。

具体的に用例を見てみると、図書館・書籍では「そんな思いがめぐってきた。」（語義一）「今日は街を巡って、位置関係を把握しよう。」（語義二）などの例もあるのに対し、社会科教科書は「国境を巡る争いが戦争に発展することも少なくない」（語義三）などに偏っている。

３．２．事例２「通じる」

「巡る」と似たふるまいを見せるのが「通じる」である。「通じる」の語義は『新明解国語辞典』には以下のように記述されている。

つうじる 【通じる】 [2]

[一]（自上一）

一 〈（どこ・なにカラ）どこ・なにニ ―― 〉

何かを伝って一方から他方へ到達出来る状態になる。「山頂へ ―― 道／電話が ―― 〔=a 電話で連絡が取れる。b 電話が架設される〕／電流が ―― 〔=回路を通って流れる〕／大便が ―― 〔=体外へ出る〕／窮すれば ―― 〔=どうにもしようがない土壇場まで追い込まれると、かえってうまい知恵も出てくるものだ〕」

二 〈どこ・なに・だれニ ―― ／なに・だれト ―― 〉

一方から他方に何かが伝わり、つながりがつく状態になる。「先方に話が通じていない／冗談が通じない人／ここでは英語が通じない／隣国と ―― 〔=交流がある〕／敵方に ―― 〔=内通する〕／人妻と ―― ／気心が通じ合う」

三 〈なにニ ―― 〉

関係者以外には分からないはずの情報を得ている。「その辺の事情に ―― 〔=詳しい〕」

四 〈なにニ ―― ／なにト ―― 〉

限定された範囲よりも はるかに広い方面にまで関連がある。「現代にも ―― 問題／一脈 相アイ ―― ものがある〔=⇒ 一脈〕」

[二]（他上一）

一 〈だれニなにヲ ―― ／だれトなにヲ ―― 〉

つながりをつける。「よしみを ―― 〔=親しく交わる〕／相手に意志を ―― ／刺シを ―― 〔=名刺を差し△上げる（上げて面会を申し込む）〕／気脈を ―― 〔=相手方とひそかに連絡をとる〕／情を ―― 〔=密通する〕」

二 〔「…を通じて」の形で〕ある状態がその範囲のすべてにわたることを表わす。「一年を通じてあたたかい」

三 〔「…を通じて」の形で〕何かを間に立てて仲介とすることを表わす。「あらゆる機会を通じて〔=利用して〕／ラジオやテレビを通じて〔=…により〕知らせる」

[2] 『新明解国語辞典』では「通ずる」は「通じる」を見出しとして統合されている。

表 3　「通ずる」の用例分布

	社会科教科書	図書館・書籍
[一]一	0	5
[一]二	1	7
[一]三	0	2
[一]四	0	7
[二]一	0	0
[二]二	4	8
[二]三	45	21
計	50	50

「巡る」と同様に、[二]二および三が複合辞的用法と考えられるため、これらに当たる用例をまとめ、それ以外と比較すると、やはり社会科教科書において複合辞的用法が非常に多いことが分かる（χ^2=9.8、p＜0.001）。

上述の２つの動詞は、社会科教科書において特定の用法で使われやすいことが分かるが、この用法の偏りは社会科という教科に特定のものではなく、教科書全体に共通するものである可能性もあるため、今後の調査が必要である。また、この他にも同様の振る舞いを見せる動詞があるかどうかも課題として残っている。

３．３．事例 3「唱える」

「唱える」の語義は『新明解国語辞典』では以下の３つに分類されている。

となえる【唱える】
（他下一）
一　〔決まった文句を〕△調子をとって（口の中で繰り返し）言う。「△お題目（念仏）を ―― 」
二　〔短い文句を〕大きな声で叫ぶ。「万歳を ―― 」
三　〔自分の意見・主張を〕大衆に広める目的で発表する。「△異（絶対反対）を ―― ／平和を口に ―― 」

表 4　「唱える」の用例分布

	社会科教科書	図書館・書籍
一	3	24
二	0	0
三	46	25
その他[3]	1	1
計	50	50

[3] 「その他」に分類したのは、以下の 2 例である。
・「民本主義といふ文字は、日本語としては極めて新しい用例である。従来は民主々義といふ語を以て普通に唱へられてられて居ったやうだ。」（社会科教科書）
・「何時から、誰が禅宗などと唱えたとも伝えられてはいないのだ。」『正法眼蔵』

３．４．事例 4「強まる」

「強まる」は多義語ではないため語義による違いはないが、主格名詞にどのような分布があるか調べた。下の表 5 および 6 に共起頻度が 2 以上の名詞をまとめた。

表 5　「社会科教科書」で「強まる」と共起する主格名詞（頻度 2 以上）

	社会科教科書
結び付き	11
動き	7
統制	3
傾向	2
度合い	2
批判	2
発言力	2

表 6　「図書館・書籍」で「強まる」と共起する主格名詞（頻度 2 以上）

	図書館・書籍
傾向	2
勢力	2
風	2

上記のように、社会科教科書では「結び付き」および「動き」が「強まる」とよく共起していることがわかる。図書館・書籍ではこのように特定の語に偏ることはないため、社会科教科書の特徴である可能性がある。このような共起語の偏りが何に起因するものか、社会科教科書のみの傾向かどうかなどは今後の課題だが、このような偏りがあることは、児童・生徒に対する日本語教育上、留意すべき特徴であるだろう。

４．まとめ

今回は、社会科教科書の動詞という非常に限定された条件で調査を行っているが、上述のように①複合助詞的用法への偏り②意味の偏り③共起語の偏り、という 3 点の特徴が観察できた。

もちろん、他の品詞や他の教科、小学校・中学校・高校の区別についても調査が必要であるが、日本語を母語としない児童・生徒に何を優先して教えるべきかを考察する際には、このような可能性を考慮に入れるべきだと考える。

例えば、学習者用の教材を作成する際には、それぞれの語の特徴を満たした例文で提示する、特徴的な用法や意味から教えるなど、児童・生徒が教科内容を理解しやすくなるように、日本語教育側からの配慮・工夫の余地はまだ残っているだろう。

文　献

田中牧郎・近藤明日子（2011）「教科書コーパス語彙表」『特定領域研究「日本語コーパス」言語政策班報告書　言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用』、pp.55-64. (http://www.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-P-10-01.pdf よりダウンロード可能)

関連 URL

文部科学省「帰国・外国人児童生徒教育等に関する施策概要」
http://www.mext.go.jp/a_menu/shotou/clarinet/003/001.htm

「ベテランは足を保護する」が語りかけるとき

保田　祥[†]（国立国語研究所 コーパス開発センター）
立花　幸子（国立国語研究所 コーパス開発センター）
柏野　和佳子（国立国語研究所 言語資源研究系）
丸山　岳彦（国立国語研究所 言語資源研究系）

When a Sentence such as “Experienced people protect their feet” is Addressed to Readers

Sachi Yasuda (Center for Corpus Development, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept. Corpus Studies, NINJAL)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)

１．はじめに

本稿は，語レベルでは特徴的表現の出現頻度が高くないが，全体としてある種の文体が感じられると判断されるテキストについて考える．

『現代日本語書き言葉均衡コーパス』（BCCWJ）に収録されている図書館サブコーパスの書籍サンプル（全 10,551 サンプル・35,732,431 語[1]）に，文書分類の観点から人手で情報を付与する作業を実施した（柏野・奥村,2012）．付与された観点の一つに「語りかけ性」（とてもある・どちらかといえばある・特にない：3 段階）がある．「語りかけ性」とは，直感的には「あなた」「みなさん」などのような呼びかけ表現や「ではないでしょうか」「だよね」といった，問いかけもしくは相づちを求める文末表現などを含むテキストに見られ，著者が読み手に対して直接語りかけていると解釈できる文体（柏野,2010 など）である．本稿は，このような文体が含まれると判断されたテキストを「語りかけ性」があるテキストと呼び，この観点付与結果を分析対象とする．

このような「語りかけ性」があると判断されたサンプル群は，判断根拠となる特徴的な表現が，語レベルの表現について出現頻度を分析することでは得にくい傾向がある（保田ほか,2012a など）．そこで，「語りかけ性」があると判断されたテキストにあっても，「語りかけ性」に特徴的と考えられる表現が含まれないような文について考察し，それらが語りかけるテキストとどのように関連しているのかを考えたい．

そのため，本稿は，特徴的表現の出現頻度ではなく，テキストのまとまりを単位としたサンプルあたりの出現量を確かめる（４．１）．次に，語レベルではなく文などの大きな単位レベルでの特徴的な表現の頻度情報を調べる（４．２）．さらに，文脈の提示が書籍タイトルから予めなされている場合の多いことを確かめる（４．３）．これらの結果により，一見「語りかけ性」に特徴的と考えられる表現が見当たらなくとも，「語りかけ性」があると判断されるテキストがどのようなものか考察する．

２．先行研究と本研究

書籍は読まれることが前提であり，広義には読み手に向けて語りかけるべく書かれているものと考えられるが，話しことばと等しいのではない．三宅（2005）は，「話しことば」の典型としてイメージされるものが「おしゃべり」であるとする．保田ほか（2012a）の調査では，書きことばであっても話しことば的であると判断されるテキストには，リアルタイム性と関わるフィラーや言いよどみ，音声的変化に関わる融合などが現れているが，「語

[†] yasuda_s@ninjal.ac.jp

[1] 空白・記号を除く場合は，28,892,944 語（中納言「語数」を参照．https://maro.ninjal.ac.jp/wiki/）．

りかけ性」があるとされるテキストにはその種の特徴が現れにくいという差異が見られた．「話しことば的」と判断されるテキストは，戯曲調で地の文が ト書きの場合や講演の書き起こし，一人称小説などに限定され，書籍サンプルの 0.6%[2]と僅少で，「語りかけ性」とは異なるものと判断されるのである．

以下に，「語りかけ性」があると判断されたテキスト例を示す．「語りかけ性」があると分類されたテキスト群において，語レベルの出現頻度で特徴的であった表現（保田ほか，前掲）に下線を施した．例 1 は，疑似的な対話形式が用いられており，読み手に対する人称（「あなた」）や接頭辞（「お」）のほか，文末に終助詞（「ね」「よ」）や勧誘（意志推量形「ましょう」）のような特徴的表現が見つかる．

1）お年寄りや赤ちゃんを連れた人，重い荷物を持った人には，席をゆずりましょう．
　席をゆずるのは，あなたが立っていても平気なときだけよ．病気やケガをしているときは無理をしなくてもいいの．
　また，ゆずられた方は，「ありがとう」とお礼を言い，会釈をします．また自分が降りるとき，あるいは替わってくれた人が降りるときも，もう一度軽く「ありがとう」と言うのをおわすれなくね．　（バーバラ寺岡「魅女ってみませんか」）

岸本（2005）は，「ネット日記」の「読み手意識表現」として，「読み手めあて」で用いられる丁寧体（調査結果における頻度順位 1 位）や伝達態度を示す終助詞（「よ」「ね」：同順位 2 位）を分析している．上に示した例 1 は，そのような「読み手意識表現」としての特徴が見られているものと考えられる．

また，野田（2012）は，「ブログや軽い文体のエッセイの文章」について，典型的な書きことばとは異なり「書き手と読み手とのコミュニケーションが意識されている」とし，エッセイ末の「読み手を意識した表現」を調査している．結果として，エッセイ末には読み手の存在を特に意識した表現（丁寧体・終助詞・疑似独話・余韻を感じさせる表現など）が多く現れていることを示すが，同時に，表現の種類や頻度には個人差も大きいことを示している．実際に，保田ほか（2012b など）の調査結果でも，例 1 で見たような「語りかけ性」があると判断されたサンプル群に高頻度な語レベルの特徴的表現（「読み手」を意識した表現とされる丁寧体や終助詞といった特定の表現はこれらに含まれる）の出現頻度が，テキストによってはとりたてて高いとも限らず，ばらつきのあることが確認された．

以下の例 2 に，語レベルで「語りかけ性」があると判断されたサンプル群に特徴的であった表現（出現頻度の高い表現）を含まない例を示す．但し，「語りかけ性」があると判断したアノテーターのコメントから，その判断基準とされた表現を下線で示す．語レベルの出現頻度からは特徴的であると判断されにくいが，それでもなお「語りかけ性」を感じるとされた表現である．読み手を意識したと考えられる婉曲化（「思う」など）や読み手に対する動作（「説明する」など）のほか，評価（「構わない」「問題はない」）や可能（「可能」「出来る」など）も，読み手に対してのアドバイスと感じたとのコメントが得られており，これらも「語りかけ性」に特徴的な表現といえる．

2）カップリングコンデンサが大きい場合，オレンジ色の側の配線が同じようにＩＣソケットの足にハンダ付けできればどのように付けても構わない．完成図を見てもらえれば分かると思うが，コンデンサの左の部分は大きくスペースが残してあるので，アキシャルリードのものも基板上に取り付け可能だ．また，大きすぎて基板からはみ出したとしても，特に問題はない．なお，後で説明するが，このコンデンサは無しにも出来る．
（酒井智巳「はじめてつくるプリアンプ」）

[2] 図書館サブコーパスからランダムに選出した 1,890 サンプル中，3 人のアノテーターが「話しことば的」と判断したサンプルは 12 サンプルにすぎなかった．

例 2 に現れたような種類の表現群が見られることで，「語りかけ性」が感じられるという可能性が考えられるが，例 1 に見られたような特徴的表現と異なり，この種類の表現は「語りかけ性」の有無によって分類したサンプル群ごとに出現頻度を比べても，そもそもの出現頻度が僅少であったり，文脈によって語感が異なったりするため，大差が見られない．しかし，アノテーターが分類のための根拠としたとすれば，サンプルとしてのテキスト内で印象に残ったということでもあり，テキストのまとまりを単位とすれば，出現数に差の見られる可能性はあろう．

さらに，語レベルの特徴的な表現が見つかりにくい文であっても，文脈によって「語りかけ性」が感じられる場合はあり得る．以下の例の下線部を施した部分は，「語りかけ性」があると判じられたサンプル群に出現頻度の高い表現も，アノテーターが「語りかけ性」があるとの判断根拠にしたという表現も見つからない．しかし，「語りかけ性」があるとされたサンプルの多くを占める，特徴的な表現のないテキストである．

3) すべると危険です．釣り上げられて跳ねた魚や自分の釣りバリでケガをするかもしれません．ベテランは，釣り物によっては真夏でも釣り用ブーツで足を保護しています．小物釣りでも滑らないスニーカーなどを履きたいものです．（井田玲子「超明解！船釣り入門 ABC」）

4) パンツは化学繊維などの混紡で，通気性のある，肌にまとわりつかない物がよいでしょう．伸縮性のあるジーンズも市販されています．よく綿のジーンズを見受けますが，厚手のそれは濡れると硬くゴワついて動きにくくなるし，乾きも悪いのです．（同上）

例 3・4 は，前後に「履きたいものです」や「よいでしょう」が現れていることで，文脈上「語りかけ性」が読み取れるとも考えられるが，下線部内に例 1・2 に示されたような表現があるのではないため，テキスト全体として，特徴的と考えられる表現の出現頻度は減少することになる．特徴的な表現の出現頻度が低くとも，印象的な表現がテキスト内に出現することで「語りかけ性」があると判断されるという可能性が考えられる．但し，アノテーターのコメントによれば，特徴的な表現が見つかってもそれらが多くはないと感じた場合は，全体として「語りかけ性」があるとは判断しなかった旨が記述されていた（保田ほか，2012）．読み手は，特徴的な表現のテキスト全体における出現量を鑑み（４．１で後述），「語りかけ性」を感じている可能性があろう．

しかし，なぜ特徴的表現が少ないサンプルでも「語りかけ性」があると判断されるのだろうか，という疑問がある．そこで，上記の例 1 から 4 は全て，これまでに読み手を意識した表現が出現しやすいと考えられてきたブログやエッセイではないテキストであるということに着目したい．

「語りかけ性」は，上記の例に見られたいわゆるハウツー系の書籍にも出現しやすい傾向がある（保田ほか，2013a など）．ハウツー系の書籍は，読み手がテキストに要求する解答に応えることを明示している可能性があり，客観的であることが望ましい社会科学・自然科学分野に多く，「語りかけ性」と主観的なテキストであることに相関はない．教示的態度を強調するために「語りかけ性」が用いられているのだといえる（保田ほか，2013b）．ハウツー系の書籍の文脈に見られやすい語よりも大きな単位の文レベルで照応があるような表現（４．２で後述）や，予めハウツー系の書籍であることが明示されているなどで書き手と読み手の関係が設定されること（４．３で後述）によって，「語りかけ性」が感じられる可能性が考えられるのではないだろうか．

３．データ

本稿は，BCCWJ の図書館サブコーパスに含まれる書籍の 10,551 サンプルをランダムに並べ替え，6 人の作業者が文書分類を行ったアノテーション結果（柏野・奥村，前掲）を用いて分析を行った．

全 10,551 サンプル（可変長サンプル，35,732,431 語）のうち，「語りかけ性」がある[3]と判断されたサンプルは 2,211 サンプル（7,128,885 語），ないと判断されたサンプルは 6,607 サンプル（23,350,311 語）である．

特徴的な表現の検索に「中納言 1.1.0（https://chunagon.ninjal.ac.jp/），短単位データ 1.0・長単位データ 1.0」を用いたほか，サンプルの形態素解析に，MeCab 0.993＋UniDic2.1.0 を用いた．分析結果に示す品詞情報や語彙素等の要素は，解析結果に基づく．

４．調査と結果

４.１　特徴的表現のサンプルあたりの出現量

語を基準としたサンプル群における出現頻度が 0.1%以下[4]の表現では，サンプル群ごとの出現頻度の差異は見えにくい．しかし，語レベルの表現であっても，アノテーターが「語りかけ性」を有したテキストであるという判断の根拠にした表現（例 2 参照）や，直感的にハウツー系の書籍に出現しやすい印象を与える表現がある．それらは，出現頻度の高い表現ではないが，「語りかけ性」がある印象を形成する表現と考えられる．但し，アノテーターは，目立った表現があったとしても「多くない」と感じた際には，テキスト全体に「語りかけ性」があると判断しないことがわかっている（保田ほか，2012）．

そこで，特徴的と考えられる表現例（保田ほか，2013a[5]）が，語数やサンプル群全体ではなく，テキストのまとまりとしてのサンプルあたり（当該表現を含むサンプルあたり），どの程度の量出現していたのか，「語りかけ性」有無のサンプル群（小説サンプルを除く）毎に確かめた．

特徴的表現の含まれるサンプルにおけるその表現の量を 10,000 語あたり[6]の平均数として表１に示す．「語りかけ性」の有無によって，テキストのまとまりを単位とした特徴的表現の出現量に差異が見られるといえる．特徴的な表現は，語数を基準として算出した頻度としては目立たなくとも，「語りかけ性」があると判断されるテキスト（サンプル）を単位とすれば，数として多く出現し，読み手の印象に残る可能性があるのだと確かめられた．

表１　当該表現を含むサンプルあたりの特徴的表現量(1 サンプルを 10,000 語に調整した際の平均数)

	という＋準体助詞 という＋形状詞助動詞語幹	という＋＜もの＞ という＋＜こと＞	＜出来る＞	＜構わない＞ ＜構いません＞	＜私達＞ ＜我々＞ ＜我等＞	＜便利＞ ＜大切＞
語りかけ性あり	35.0	40.5	20.2	2.2	11.9	8.2
語りかけ性なし	25.8	33.2	16.3	1.7	10.5	5.4

４.２　文脈を作り出す文レベルの特徴的表現とサンプル群別出現量

次に，「語りかけ性」を形成すると考えられる表現のうち，語レベルではない表現について見てみたい．「語りかけ性」があると判断されるテキストにハウツー系書籍が多い傾向が

[3] 「語りかけ性」のアノテーション結果は「とてもある・どちらかといえばある」「ない」の三種類であるが，本分析においては，「ある」「ない」として扱った．

[4] たとえば＜構わない＞＜構いません＞は，10,551 サンプル中に計 916 件出現する表現であるが，「語りかけ性」の有無で分類した群内の出現率（語数あたり）は，両群ともに 0.001%と低いため，単純な出現頻度では有意差は見えにくい．同様に，＜便利＞＜大切＞は計 4,582 件（各々サンプル群内の出現率は 0.005%程度），＜私達＞＜我々＞＜我等＞も計 16,658 件（同 0.01～0.02%程度）である．

[5] 保田ほか（2013a）では，客観化のために主張の裏付けとして伝聞の引用表現「という」を用いることを示し，出現頻度が「語りかけ性」がある群内における「とても客観的」と分類されたサンプル群に多いことを明らかにした．このほか，アノテーターコメントから得られた表現例を用いて調査を行った．

[6] サンプルによって語数が異なるため，サンプルの総語数を 10,000 語あたりの調整頻度に変換した．また，表現の形態素数によって表現量を算出した．

あるのならば，教示的態度（読み手の解答要求への応え）によって現れやすい文脈が予測される．たとえば，読み手の求める条件（例示）があって，その対応方法（結果や評価など）が示されるという文脈である．

例 1 では，「病気やケガをしているとき」「無理をしなくてもいい」という照応が見られる．例 2 でも，「カップリングコンデンサが大きい場合」「構わない」をはじめ，「完成図を見てもらえれば」「分かる」，「はみ出したとして」「問題はない」などが多数見られる．

そこで，以下のパターンの出現量を調査した．調査にあたっては，後文脈は 10 語以内に検索語の出現する例を取得した．取得した用例のうち，文が途切れるなどによって後文脈につながりがないと考えられるものは対象外とした．また，「語りかけ性」有無群については，会話文を除外するため小説を対象外とした（「語りかけ性」あり：1,824 サンプル，なし：4,074 サンプル）．

① 前文脈が条件で，後文脈に特徴的表現例（名詞）を含む
- 前文脈：動詞＋「と」（接続助詞）or 仮定形 or＜時＞＜場合＞
- 後文脈：＜必要＞＜可能＞＜大切＞＜便利＞を含む

> 例）デザイン事務所といえば，場所のイメージもやはり大切です．
> （高橋慈子，伊藤華子「家庭をオフィスにする SOHO 読本」）

742 件：「語りかけ性あり」群 244 件・「語りかけ性なし」群 360 件

② 前文脈が条件で，後文脈に動詞・形容詞を含む
前文脈・後文脈の種類は，頻度上位 5 種をそれぞれ脚注に示す．

A.
- 前文脈[7]：動詞＋「と」（接続助詞）
- 後文脈[8]：形容詞（終止形）

> 例）コンニャクの球根もかじると舌が割れるほど辛い．（奥山久「山菜」）

6,607 件：「語りかけ性あり」群 743 件・「語りかけ性なし」群 813 件

B.
- 前文脈[9]：動詞＋「時」
- 後文脈[10]：動詞

> 例）壁に御影石などを貼るときは，必ず石の裏側に防水を施しましょう．
> （浜口和博「プロも見落とす家づくりの急所」）

1,599 件：「語りかけ性あり」群 380 件・「語りかけ性なし」群 493 件

C.
- 前文脈[11]：動詞＋「場合」
- 後文脈[12]：動詞

---

[7] すると：14.2%・みると：12.8%・いうと：8.3%・なると：7.2%・おくと：3.9%

[8] よい：34.2%・ない：27.5%・多い：2.5%・おもしろい：1.9%・わるい：1.5%

[9] いるとき：17.3%・するとき：15.9%・いうとき：9.8%・あるとき：4.8%・みるとき：1.6%

[10] する：13.4%・いる：6.9%・ある：5.5%・いう：5.3%・なる：4.3%

[11] する場合：23.1%・いる場合：16.4%・ある場合：9.9%・いう場合：8.1%・なる場合：3.2%

[12] ある：21.9%・する：18.2%・なる：3.6%・いる：3.1%・いう：3.1%

例）４回以上の支給がある場合は，月々の報酬に含まれます．
（高橋徹編「社会保険・労働保険のすべてがわかる事典」）

4,999 件：「語りかけ性あり」群 1,381 件・「語りかけ性なし」群 2,114 件

D.　・前文脈[13]：仮定形
・後文脈[14]：動詞

例）発芽してきたら，覆っていた新聞紙をとってやる．
（稲山光男「まるごと楽しむキュウリ百科」）

54,783 件：「語りかけ性あり」群 9,913 件・「語りかけ性なし」群 16,792 件

それぞれの表現について，「語りかけ性」の有無群別に 1,000 サンプルあたりの調整頻度を以下に示した．①の結果を表２に，②を表３に示す．それぞれ，「語りかけ性」があると判断されたサンプル群における出現量が多いことがわかる．文レベルでも，特徴的な表現が出現しているといえよう．

表２「語りかけ性」の有無とサンプル群別出現量（調整頻度 1,000 サンプル，条件＋後文脈内要素）

	＜必要＞	＜可能＞	＜大切＞	＜便利＞	4 種計
語りかけ性あり群	11	48	18	43	121
語りかけ性なし群	7	39	6	12	64

表３「語りかけ性」の有無とサンプル群別出現量（調整頻度 1,000 サンプル，条件別）

	動詞＋と_形容詞(終止形)	動詞＋時_動詞	動詞＋場合_動詞	仮定形_動詞
語りかけ性あり群	353	208	757	5,435
語りかけ性なし群	200	121	519	4,122

４.３　書籍タイトルに見る書き手と読み手の関係性の設定

本稿で扱っているサンプルが書籍サンプルに限られるため，書籍タイトルから「語りかけ性」が含まれるテキストであるかという推測が可能かを調べた．ハウツー系書籍であれば，読み手が予め目的意識を持ってテキストを読むことが推測され，そのためテキスト内容の代表たるタイトルに，解答の要求に応える旨の明示が期待されるからである．

４.３.１　方法と手順

調査は，図書館サブコーパスの書籍タイトル（10,551 件）を用い，2 名の作業者（第一発表者・第二発表者）が，（１）タイトルから語りかけていると予測可能か（ハウツー本であるか）否か，（２）タイトルに判断根拠たる指標が含まれているか否か・指標等はないがそれらしいと感じるか否か，を付与した．

作業者がタイトルに付与した「語りかけ性」の予想結果と，実際にテキストを読んで「語りかけ性」があると判断された結果の対照を行った．

４.３.２　結果（１）

10,551 サンプル中 8,818 サンプルがアノテーション対象であり，そのうち作業者 A が

[13] た（たら）：25.6%・ない（なけれ）：22.3%・だ（なら）：16.4%・ず（ね）：5.8%・する（すれ）：4.3%

[14] なる：25.5%。・する：9.1%・いう：4.8%・ある：3.2%・できる：1.8%

2,150 サンプル，作業者 B が 2,011 サンプルについて，本文に「語りかけ性」が予測されると判定した．実際のアノテーション結果で「語りかけ性」があると判断されたサンプル数は 2,211 サンプルである．

作業者 A と B の判定の一致率（F 値）は 77.3%であった．しかし，作業者 2 名が共通してタイトルから推測したサンプルと実際のアノテーション結果の一致率（F 値）は 42.5%に留まる．タイトルから推測される「語りかけ性」は，作業者間で 8 割程度一致するが，実際の結果とは半数程度しか一致しないという結果になった．

それでは，タイトルから判定できない「語りかけ性」は何であったのか．図 1 は，タイトルからの推測結果とアノテーション結果について，それぞれNDC分類とC-code分類による内訳を示したものである．網点部分が 2 種類の差異が明確な部分である．NDC分類では 9 番台（文学），C-code分類では 8 番台（児童）でアノテーション結果に多いという違いの生じていることがわかる．タイトルからの推定と本文のアノテーション結果が不一致となったサンプルを見ると，NDC9 番台・C-code8 番台ともに，物語・小説と推測されるタイトルが並んだ[15]．以下に例を挙げる．

・NDC9 番台（文学）の不一致例：「野に出た小人たち」「短篇ベストコレクション」「窯変源氏物語」「世界探偵小説全集」「メグレたてつく」など
・ C-code8 番台（児童）の不一致例：「目をさませトラゴロウ」「イワンのばか」「動物園ほのぼの日記」「いたずらまじょ子のヒーローはだあれ」など

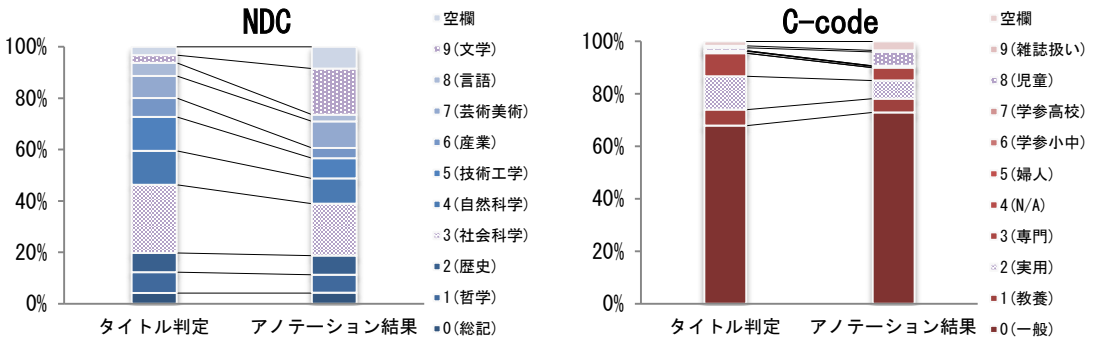


図 1　タイトル判定とアノテーション結果の NDC・C-code による内訳

なお，タイトルから判定された結果は，NDC3 番台（社会科学）と C-code2 番台（実用）でアノテーション結果より多い．NDC3 番台と C-code2 番台は，それぞれハウツー系書籍の多い分類であり，ハウツー系書籍と判断されたタイトルでも，本文においては「語りかけ性」がないと判定されたサンプルが見られるということでもある．但し，サンプル部分が「語りかけ性」なしという判定であっても冊全体がハウツー本でないというわけでもない．「語りかけ性」があるとタイトルから推定されたが，アノテーション結果では判定されなかった例の画像データを見ると，図鑑的部分がサンプルとなっており，レシピ部分は「語りかけ性」があると判定されそうであるが範囲外という書籍（例：「とっておきの山菜利用術」）や，インタビュー部分が「」内であってアノテーションの判断に用いられた地の文が僅かであった書籍（例：「夢を夢で終わらせないためのペット・ケアの仕事と資格」）なども含まれている．

[15] 小説を除外し，タイトル判定とアノテーションで結果が不一致だったサンプルの NDC 分布を調べると，NDC3 番台（23.0%）と 7 番台（17.2%）の割合が大きい（他番台は平均的）．NDC7 番台（芸術・美術）には，芸能人やアスリートのエッセイ等（例：「ほんじょの眼鏡日和。」「ジャイアンツ愛」など）が含まれる．

４.３.３　結果（２）

タイトルから「語りかけ性」があると推測するにあたり，タイトルに判断根拠たる指標が含まれている・指標等はないがそれらしいと感じるという判定の付与を行った．作業者 A が「語りかけ性」があると推測したタイトルの 81%に指標が含まれるとし，作業者 B が同 76%に含まれるとした．作業者 2 名の判定一致率（F 値）は以下の通りである（「講座～」のような，指標はあるが「語りかけ性」はない気がするとコメントがついたタイトルを含む）．

- それらしき指標がある一致率（F 値）　76.6%
- 指標はないがそれらしい（F 値）　40.9%

作業者 2 名の判定結果から，「語りかけ性」があると推定されたタイトルのうち，判断根拠となる特徴的な表現を含むタイトルが 8 割程度あり，また，作業者間の感覚も 8 割程度で一致することがわかった．

タイトルに含まれた特徴的指標は多様であるが，「語りかけ性」があるテキストと判断される根拠となる表現（呼びかけ（「あなた」など）・可能・勧誘・命令・評価・希望など）をはじめ，対象読者を示す表現（「～のための」・初心者向けであると示す「入門」「はじめて」「やさしい」など），方法解説であると示す表現（「テクニック」「レッスン」「手引き」「教える」）などに分類され，ハウツー系書籍であることを明示する傾向がある．「語りかけ性」を有する書籍（特にハウツー系書籍）においては，タイトルに「語りかけ性」を推定させる特徴的指標を含む場合が多く，特徴的指標を含まない場合は，読み手の感覚に個人差が見られるといえる．

４.３.４　書籍タイトルと「語りかけ性」

書籍タイトルから，予めテキストに「語りかけ性」があると推測する場合，2 名で 8 割程度の推測が一致することがわかった．実際に本文を読んだアノテーション結果との一致は半数程度に留まったが，その理由として，小説や一部随筆など推測のしにくいジャンルや，「語りかけ性」があるサンプルの多いジャンルにおいても語りかけていないと受け取られる場合があるためと考えられる．また，書籍タイトルから「語りかけ性」があると推測される場合，その判断根拠として特徴的表現が含まれるという印象は 2 名で 8 割一致する．

以上により，書籍タイトルのみでもハウツー系書籍であるなどと判断されることで，語り手と読み手の教授関係が設定され，語りかけるという教示的態度が予め示されているという可能性はあろう．

５．まとめ：「ベテランは足を保護する」が語りかけるとき

５.１　「語りかけ性」を作り出す表現

「語りかけ性」は，特徴的表現によってその印象が作り出される．

まず，直感的に語りかけていると感じられる「あなた」「みなさん」のような呼びかけや，終助詞「ね」や丁寧体「です」「ます」のような読み手を意識したと考えられる表現は，実際に「語りかけ性」があると判断されるテキスト群における出現頻度も高く，「語りかけ性」があるという印象を形成するのに特徴的といえる．

次に，一つ一つの表現がそれだけで特徴的とは呼び難く，語レベルの出現頻度を調べても目立たないが，文脈上，読み手が「語りかけ性」を受け取る種類の表現がある．直感的に語りかける表現がなくとも，テキストのまとまり全体の中に，この種類の表現が多いと感じた場合には，読み手は「語りかけ性」があると判断するのである．そのため，テキストのまとまりごとの出現量を調べると，「語りかけ性」があると判断されたテキストのまとまりには，その表現の出現量が多いという傾向が確かめられる．これらの表現もまた，「語りかけ性」に特徴的な表現と呼ぶことができるだろう．

しかし，読み手が語りかけている感じを受けると考えられる特徴的な表現が僅かであっ

ても，「語りかけ性」があると判断されるテキストもある．たとえば，「そのようにしましょう」「やめましょう」というような教示などは，記述せずとも読み手に明らかであるのならば省略され得る．すなわち，「語りかけ性」を有する文脈によって，特徴的な表現が含まれずとも語りかける感じが生じるためであろうと推測される．

たとえば，ハウツー系書籍では，前文脈に読み手の期待する条件があり，後文脈にその解答（結果・評価など）が示されるというような，文脈的に出現量の多い表現のあることが確かめられた．文を形成するそれぞれの語は特徴的な表現ではなく，文として特徴があるとも言い難くとも，語りかける印象を形成している可能性は考えられる．また，予めハウツー系書籍であるということが書籍タイトルから明らかであって，語りかけるという教示的な書き手の態度が推測されている場合もあろう．このように，「語りかけ性」が文脈や書き手と読み手の関係性の設定によって明らかな場合，特徴的な表現が省略されていても，読み手は語りかける感じを受ける可能性がある．よって，ハウツー系書籍でも読み手の要求に応じることが明示的であれば，「語りかけ性」を有する必要はないともいえる．タイトルから「語りかけ性」があることを予測した結果が，実際のアノテーション結果よりもNDC3 番台（社会科学）と C-code2 番台（実用）で多かったのは，このためもあろう．

５.２　「ベテラン」であること

ハウツー系書籍においては，使用される語彙が特徴的であることも考えられる．

ここでは，「ベテラン」に着目したい．例 3 では，初心者に向けて，「危険」や「ケガ」の恐れなどが述べられ，結論として「滑らないスニーカーなどを履」くことを勧めている．その途中に「ベテラン」が「ブーツで足を保護」していることが挿入される．ここでの「ベテラン」は，「初心者」との対照として用いられていると読める（例 5 ではそのことが明示されている：網掛け部分を参照）．

> 3) すべると危険です．釣り上げられて跳ねた魚や自分の釣りバリでケガをするかもしれません．ベテランは，釣り物によっては真夏でも釣り用ブーツで足を保護しています．小物釣りでも滑らないスニーカーなどを履きたいものです．　（再掲）

この文においては「ベテラン」が足を保護するという条件に対し，「ケガ」の恐れがないという結果の記述すら省略されている．このように，読み手の求める条件に対する解答を記述するという文脈も不要であるのは，「ベテラン」が，ハウツー系書籍の読み手にとって理想的なモデルを想起する語であるためと考えられる．また同時に，「ベテラン」でも「足を保護」するのであるから，初心者もそのように「しましょう」と勧める例示とも読めようが，その明示的な記述はない．

> 5) 一日も早く花が見たいということから，ベテランは「早蒔き」をします．その時期は７月下旬～８月上旬．発芽適温が十五度のパンジーや，なかでもヴィオラにとってはまだ暑く，そのままでは発芽させにくい上級者向けです．「標準蒔き」は８月下旬から９月の上旬ですが，この時期でも外気温のまま発芽させるにはまだ暑すぎます．初心者や確実を期したい場合には，９月中旬の「遅蒔き」がベスト．　（著者不明「人気のパンジー」）

なお，例５は，「ベテラン」は「上級者（網掛け部分）」と言い換えられ，「初心者」と対照される存在であるが，「ベテラン」であるからできるのであって，「初心者」には無理であるから真似をするなという教示に用いられていると読める例である．この場合，「ベテラン」は「初心者」から遠い性質が焦点化されているが，やはり，初心者は「やめましょう」との記述は直接的になされてはいない．「ベテラン」のようなハウツー系書籍に特徴的な語彙によって，「語りかけ性」を有する文脈が明確化することで，「語りかけ性」を感じる特徴的表現類が省略されている可能性が考えられよう．

５.３　まとめ

「語りかけ性」という文体がどのように形成されているのか，「ベテランは足を保護する」という文を含むテキストを取りあげて考察した．

語りかける感じを与えると考えられる特徴的な表現を含まない文であっても，特徴的な表現が文脈上不要なテキストであれば，語りかけると捉えられる可能性がある．そのため，「語りかけ性」がどのような特徴的な表現から成っているのか，特徴的と考えられる各々の出現頻度からは捉え難いという現象が生じている．その場合，文脈上のパターンとして蓄積されたり（４．２），書籍タイトルのようなもので前提的に書き手と読み手の関係性が設定されたり（４．３）することによって，直接的な表現がテキストから取得しにくくなっている可能性があろう．

ハウツー系書籍は，「語りかけ性」の文体を戦略的に使用しており，「語りかけ性」が積極的に用いられる傾向がある．語りかけるという本来書きことばにない疑似的対話表現を用いることは，読み手の要求する解答を与える教示的な態度をその異質性によって明示するマーカーとして機能していると考えられる．特徴的な表現が出現頻度として高くなくとも，テキストのまとまりあたりの出現数として多い傾向（４．１）は，その表現が目立つように用いられれば良いためでもあろう．「語りかけ性」に特徴的な表現は，使用する必要がなければ省略可能であり，語りかける印象があっても，まったく出現しない場合すらあり得るのだと考えられる．

このように，「語りかけ性」のような文体的な性質は，直接的な表現をはじめ，文脈や，語り手と読み手の関係性などによっても形成され，多様に読み手へと語りかける印象を与えるのである．

文　献

柏野和佳子(2010)「「直接的な語り」という表現スタイルをもつ書籍テキストの人手抽出の試み」ことば工学研究会，35，pp. 63-72.

柏野和佳子，奥村学(2012)「書籍テキストへの分類指標人手付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」言語処理学会第 18 回年次大会予稿集，pp. 1260-1263.

Levinson, S. C. (1990) *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

岸本千秋(2005)「ネット日記における読み手を意識した表現」メディアとことば 2「特集」組み込まれるオーディエンス，ひつじ書房，pp. 204-231.

三宅和子(2005)「携帯メールの話しことばと書きことば」メディアとことば 2「特集」組み込まれるオーディエンス，ひつじ書房，pp. 234-261.

野田春美(2012)「エッセイ末における読み手を意識した表現」人文学部紀要，32，pp. 39-54.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012a)「「語り性」を有する書きことばの典型例の分析」第１回コーパス日本語学ワークショップ予稿集，pp. 139-146.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012b)「「語りかけ性」を有すると判断される書きことばの表現」第２回コーパス日本語学ワークショップ予稿集，pp. 43-50.

保田祥，柏野和佳子，立花幸子(2012)「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」ことば工学研究会，41，pp. 3-10.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2013a)「アノテーターコメントを用いた「語りかけ性」分析の試み―頻度情報から捉え難いテキスト性質の解明に向けて―」言語処理学会第 19 回年次大会予稿集，pp. 358-361.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2013b)「書きことばにおける「語りかけ」は何のために用いられるのか」第３回コーパス日本語学ワークショップ予稿集，pp. 143-152.

多義複合動詞の語義構造の分析

山口昌也 (国立国語研究所言語資源研究系) †

Analysis of Polysemy of Compound Verbs

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

1　はじめに

本発表では，複合動詞の多義性が発生するメカニズムに構成動詞がどのように関わっているかを分析する。対象とする複合動詞は,「動詞 (連用形) ＋動詞」タイプの複合動詞である。また，ここで言う多義性とは，複数の格フレームを持つこととする。

まず，多義の複合動詞の例として,「塗り替える」と「噴き出す」の語義と用例を示す。語義と用例は大辞林第３版 (松村 2006) から引用した。

- 「塗り替える」

  (1) 前に塗ってあったものを改めて，新しく塗り直す。「壁を塗り替える」
  (2) すっかり変える。また，記録などを更新する。「世界記録を塗り替える」

- 「噴き出す」

  (1) 中から外に激しい勢いで出る。「温泉が噴き出す」
  (2) 内から外へ吹いて出す。「タバコの煙を噴き出す」

このうち,「塗り替える」の語義 (2) は，既存の記録などを新しいもので更新することを「塗り替える」という語義 (1) のメタファーで表現したものとなっている。この種の多義性（以後,「比喩による多義性」）を説明するためのモデルとして，メタファー，シネクドキー，メトニミーといった，比喩による意味拡張に基づいた認知意味論的なモデルがある (松本 2003，国広 2006 など)。一方,「噴き出す」の語義 (1) は非対格自動詞，語義 (2) は他動詞である。このタイプの多義性（以後,「多重項構造による多義性」）は，一つの語が複数の項構造を持つことによるものと考えられる。

以上のように，二つのタイプの多義性を説明するための枠組みはすでに整っており，実際の辞書に応用されている[1]。しかし，これらの枠組みでは，複合動詞の構成動詞が多義性にどのように関わっているかを明らかにしていない。そこで，本稿では，(a) 比喩による複合動詞の多義性に構成動詞が関与する方法の類型化とその特徴，(b) 多重項構造による多義性を持つ複合動詞の量と特徴，を複合動詞と構成動詞の用例データベースを用いて分析する。

2　分析の方法

2.1　使用するデータ

分析方法を説明する前に，分析に用いる『複合動詞用例データベース』[2](以後,「用例 DB」) について説明する。

用例 DB は，Web データをもとに半自動的に構築したデータベースで，複合動詞の用例，語構成情報に加えて，構成動詞の用例を収録している。収録対象の複合動詞は,「動詞 (連用形) ＋動詞」タイプの「語彙的複合動詞」(影山 1993) である。収録語数は，複合動詞が 3912 語，単一動詞が 1148 語である。収録語の用例数（中央値）は，複合動詞が 977 例，単一動詞が 5858 例である。本稿で分析対象とする複合動詞は，1000 例以上の用例を持つ 1939 語である。

†http://www2.ninjal.ac.jp/masaya

[1]例えば，瀬戸 (2007) では認知意味論的な語義わけの基準を採用している。また，国語辞典において，動詞の自他は語義わけの基準として一般的に用いられている。

[2]http://csd.ninjal.ac.jp/comp

用例 DB 中の用例は形態素解析[3]・格解析[4]され，当該動詞が持つ格とその格要素がわかるようになっている。ただし，自動解析のため，語の認定や係り先の誤りなどが含まれることに注意されたい。

次の例は，複合動詞「聞き出す」の格と格要素の情報（一部）である。カッコ内の数字は，出現ページ数を表す。今回の分析では，誤解析の影響を軽減するため，使用する格要素は，(a) 出現ページ数が 3 ページ以上であること，(b) 格標識が格助詞の格要素であることを条件としている。

ヲ　情報 (166)/話 (68)/番号 (60)/名前 (39)/本音 (33)/住所 (31)/場所 (31)/秘密 (24)
カラ 人 (15)/本人 (11)/相手 (9)/者 (9)/男 (8)/彼女 (6)/彼 (6)/こちら (5)/口 (5)/子供 (4)
デ　電話 (7)/中 (7)/会 (5)
ニ　人 (6)/中 (5)/時 (4)/前 (4)

2.2　比喩による多義性の分析方法

比喩による複合動詞の多義性が構成動詞とどのように関係しているのかを分析するためには，複合動詞と構成動詞の語義を意味的に比較する必要がある。山口 (2013) では，複合動詞と構成動詞の意味的な関係を分析するために，格要素の「重複度」を用いた。重複度 $Overlap_v(w_c, w_s)$ は，複合動詞の格 i の取りうる格要素が構成動詞の格 i で取りうる割合であり，次式で定義される。なお，w_c，w_s はそれぞれ複合動詞，構成動詞，W_c, W_s はそれぞれ w_c，w_s と共起する語の集合，$F_c(w)$ は語 w が w_c と共起する頻度である。

$$Overlap_v(w_c, w_s) = \frac{\sum_{w_a \in W_c \cap W_s} F_c(w_a)}{\sum_{w_b \in W_c} F_c(w_b)}$$

重複度が高くなると共通して取りうる格要素の割合が高くなるので，意味的な類似性も高くなると考えられる。ただし，重複度では多義性は考慮されておらず，用例 DB の用例にも語義情報がない。そこで，本稿では，人手で複合動詞の語義ごとに格要素を分類し，複合動詞と構成動詞が共通して取りうる格要素を定性的に分析することにする。そして，比喩による複合動詞の多義性に構成動詞が関与する方法を類型化する。分析の手順は，次のとおりである。なお，複合動詞と構成動詞対象とする動詞の種類は他動詞，対象とする格はヲ格である。また，比喩の種類はメタファーとする。

(1) 多義の複合動詞を用意する。今回は，構成動詞 (前項動詞，後項動詞) とのヲ格の重複度が共に 0.3 以上 0.7 以下の複合動詞をランダムに 50 語選択し，その中から大辞林で複数の語義を持つ 26 語を抽出した。重複度を制限したのは，構成動詞の格要素と部分的な共通性を持つ複合動詞は非共通部分が別義になりうるので，多義になりやすいと考えたためである。

(2) 用意した複合動詞のヲ格の格要素を語義ごとに分類する。語義分類は，大辞林を参考にした。分類の結果，各語義ごとに 5 語以上の格要素が集まらない語義は分析対象から外した。

(3) 複合動詞と構成動詞と共通して取りうる格要素を用例 DB で調べ (前項動詞，後項動詞別々に行う)，複合動詞の多義性に構成動詞が関与する方法を類型化する。基本的には，(1) 格要素の共通性が低ければ比喩的な多義性でないかを検証し，(2) 共通性が高ければ構成動詞の多義性との関係を検証する。

2.3　多重項構造による多義性の分析方法

分析の手順としては，(1) 複数の項構造を持つ複合動詞を探し，(2) 構成動詞の項構造と比較する。探索の対象となる複合動詞は，2.1 節で述べたように，用例 DB 中に 1000 以上の用例を保持してい

[3]JUMAN ver.6.0 (京都大学黒橋・河原研究室，http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN）を利用した。
[4]KNP ver.3.01 (京都大学黒橋・河原研究室，http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP) を利用した。

る 1939 語である。これらをすべて人手でチェックするには大きな手間がかかる。また，格ごとに見ていくと，分析するのに十分な用例がない場合もある。そこで,「用例 DB では，二つの項構造を持つ動詞はガ格・ヲ格両方の出現率が高い」という，一種のヒューリスティクスを使って，候補の複合動詞を絞り込む。

まず，実際のデータを用いて，このヒューリスティクスを説明する。図 1 は，横軸・縦軸をガ格・ヲ格の出現率として，1939 語の複合動詞をプロットしたものである。

一般的に自動詞はヲ格を取らないので[5]，解析誤りを考慮しても，ヲ格出現率は 0 に近くなるはずである。したがって，一定値以上のガ格出現率を持ち，図 1 の横軸近辺に並行して分布している複合動詞は，自動詞の可能性が高い。一方，他動詞は一定値以上のヲ格出現率を取ることが予想されるが，該当する点を見ると，ガ格出現率が全体的に低いことがわかる。実際，ヲ格出現率が 0.2 以上の動詞 (1046 動詞) のガ格出現率は，0.02 だった。この原因としては，主語が省略されている用例の存在や，用例 DB では格助詞以外は格マーカと認めていない[6]ことが考えられる。以上のことから，ガ格・ヲ格両方の出現率が高い動詞は，二つの項構造を持っている可能性が高い，というヒューリスティクスを設定した。

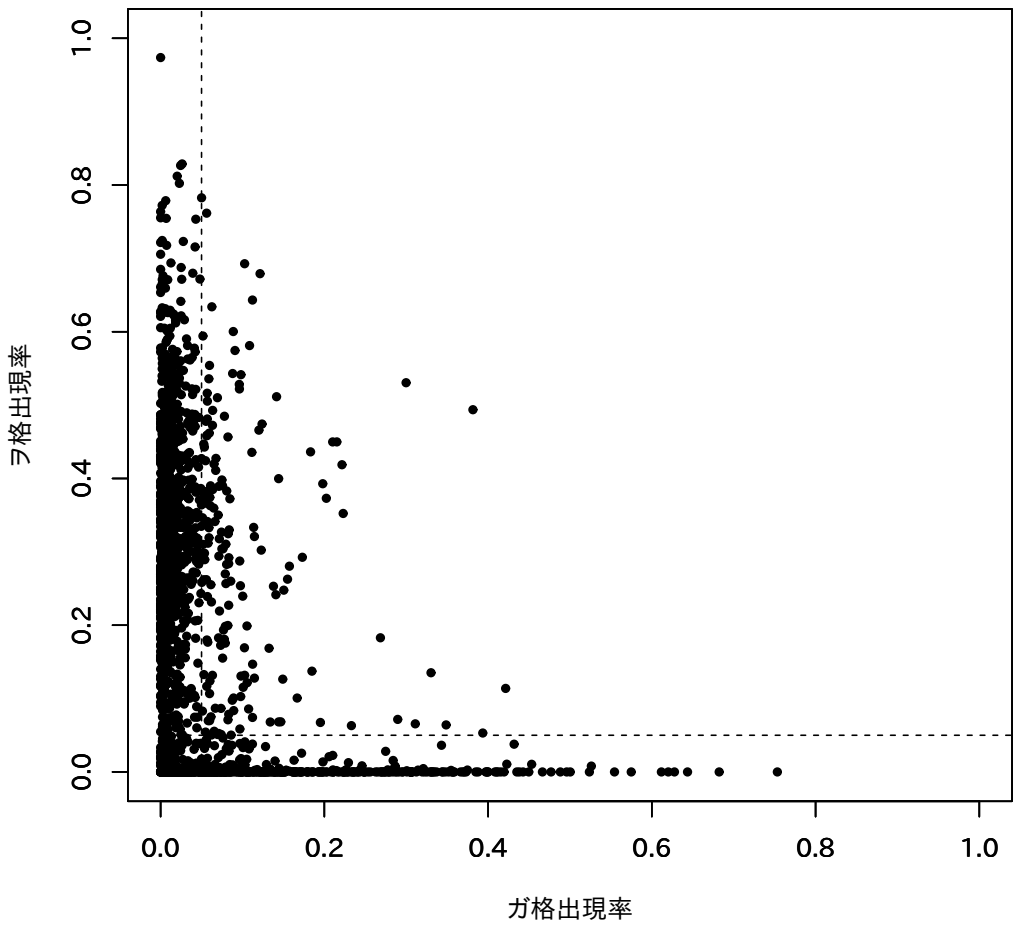


図 1: ガ・ヲ格の格要素出現率

上記のヒューリスティクスを用いて，ガ格・ヲ格の出現率が共に 0.05 以上の複合動詞を抽出した。

[5]自動詞でも「野山を走る」のように，動作の場所を表すヲ格を取る場合がある。
[6]これは，係助詞による格の曖昧性を排除するためである。

この条件だと，どちらの格も少なくとも 50 例以上の用例を参照することができる。該当する複合動詞は，168 語 (図 1 では，点線内部の点) であった。用例 DB 中の用例を参照しつつ，これらの複合動詞の中から，複数の項構造を持つ複合動詞を探したところ，次の 17 語を得ることができた。3 章ではこれらの複合動詞の項構造と構成動詞の項構造を比較・分析する。

噛み込む，持ち直す，吹き込む，吹き出す，吹き付ける，注ぎ込む，張り詰める，
張り込む，張り出す，突っ張る，噴き出す，ぶり返す，消し込む[7]，照り返す，
乗り出す，盛り返す，立て込む

3　分析結果

3.1　比喩による多義性

2.2 節に示した方法で，複合動詞の格要素と構成動詞の格要素の共通部分を調べたところ，構成動詞の語義と複合動詞の語義間の関係には，大きく分けて，二つのパターンが存在した。表 3.1 に「刻みつける」「投げつける」「建て直す」の例を示す。なお，語義 2 は語義 1 のメータファーとしての語義である。また，格要素は出現ページ数順に上位 5 語 (降順) を示した。括弧内の数字の 1，2 は，それぞれ前項動詞，後項動詞の用例にも存在した格要素であることを示す。

表 1: 複合動詞と構成動詞の語義間の関係

動詞	語義番号	語義説明 (大辞林から引用)，格要素 (ヲ格)
刻みつける	語義 1	彫って，あとをつける。
	1	名 (1)，記憶 (1)，文字 (1)，印 (1,2)，溝 (1,2)
	語義 2	しっかりと心にとどめておく。
	2	存在，言葉 (1)，印象 (1)，名 (1,2)，記憶 (1)
建て直す	語義 1	今までの建築物をこわして新しく建てる。改築する。
		家 (1,2)，建物 (1)，自宅 (1)，住宅 (1)，病院 (1)
	語義 2	つぶれそうになった会社などを，もとどおりにする。再建する。
		日本，国 (1)，経済，会社，生活 (2)
投げつける	語義 1	物を目標に向けて投げる。また，手荒く投げる。
		物 (1,2)，石 (1,2)，ボール (1)，瓶 (1)，爆弾 (1)
	語義 2	強い口調で相手にものを言う。
		言葉 (1,2)，視線 (1)，疑問，不満，皮肉

まず,「刻みつける」は，比喩による構成動詞の多義性が複合動詞に反映されている例である。この例の場合，語義 2 の「存在」を除き,「刻む」は「刻みつける」と共通する格要素を持ち，しかも,「刻む」にとっても，語義 2 は語義 1 のメタファーになっている。したがって，前項動詞「刻む」の多義性が複合動詞に反映されていると考えられる。同様の構造を持った複合動詞としては,「投げつける」「産み出す」「搾り取る」「吸い取る」があった。

次に，もう一つのパターンとして,「建て直す」を挙げる。「建て直す」は，構成動詞の比喩からではなく，複合動詞自体の比喩により多義性が生じている例である。これは，メタファー側の語義 2 では，いずれの構成動詞とも格要素を共有しないことからわかる。同様の構造を持った複合動詞としては,「落し入れる」「見合わせる」があった。

以上のように，用例データから全体的な傾向を把握することはできるが，問題もある。一つは，前項，後項動詞の関係を用例データから把握できていない点である。例えば,「刻みつける」の後項動

[7]釣りに関連する用語である。例：「魚がエサをくわえた時に，ウキが消し込みます」「気持ちよくウキを消し込む豪快なアタリは最高ですね！」(用例 DB から引用)

詞「つける」は語義 1 の場合，比喩的な用法でないが,「刻みつける」と共通する格要素は限定的である。したがって，本稿で仮定している格要素の共通性と意味的な類似度の面から言えば，比喩的な意味と解釈されてしまう。共通する格要素が限定的なのは,「刻む」が「つける」を含意するという関係を持っており,「つける」単独での選択制限とは異なるためだと考えられる。

もう一つは，頻度の低い格要素の扱いである。三つ目の例「投げつける」は,「刻みつける」と同様に前項動詞の多義性が複合動詞に反映されていると考えられる。しかし，表 3.1 に示したように，用例 DB 中には「(疑問— 不満—皮肉) を投げる」という用例はほとんど現れない。これらの三つの格要素は，出現ページ数が 3 ページであり，用例数が十分でなかった可能性がある。

これら二つの問題の解決は，今後の課題である。

3.2　多重項構造による多義性

2.3 節で抽出した複合動詞の項構造と構成動詞の項構造を調査した結果を表 2 に示す。「項構造 A」「前項$_A$」「後項$_A$」列は，それぞれ，自動詞と解釈された場合の複合動詞，前項動詞，後項動詞の項構造である。項構造は，非能格自動詞 (表中では「能」)，非対格自動詞 (対)，他動詞 (他) で表す。「項構造 B」「前項$_B$」「後項$_B$」列は，他動詞と解釈された場合の項構造である。なお,「込む」のように自立しない構成動詞の場合など，項構造が明確でない場合は,「—」としている。

表 2: 複合動詞と構成動詞の項構造

	項構造 A	**前項**$_A$	**後項**$_A$	**項構造** B	**前項**$_B$	**後項**$_B$
噛み込む	対	対	—	他	他	—
持ち直す	対	対	他	他	他	他
吹き込む	対	対	—	他	他	—
吹き出す	能・対	能・対	他	他	他	他
吹き付ける	対	対	他	他	他	他
注ぎ込む	対	対	—	他	他	—
張り詰める	対	対	他	他	他	他
張り込む	能	他	—	他	他	—
張り出す	対	対	他	他	他	他
突っ張る	対	—	対	他	他	他
噴き出す	対	対	他	他	他	他
ぶり返す	対	—	他	他	—	他
消し込む	対	他	—	他	他	—
照り返す	対	対	他	他	対	他
乗り出す	能	能	他	他	能	他
盛り返す	対	他	他	他	他	他
立て込む	対	他	—	他	他	—

複合動詞と構成動詞の項構造の関係で特徴的なことは，次の 2 点である。

(1) 11 の複合動詞の構成動詞自体が複数の項構造を持つ
(2) 「他動性調和の原則」(影山 1993) を満たさないなど，11 の複合動詞が項構造上の問題を持つ

(1) に該当する複合動詞は，表 2 の上から 11 個の動詞である。複数の項構造を持つ構成動詞は,「噛む」「持つ」「吹く」「噴く」「注ぐ」「張る」の 6 動詞である。次の例は,「噛む」と「噛み込む」,「吹く」と「吹き込む」の用例を，項構造ごとに対応付けて示したものである。

- 噛む/噛み込む
  - (対) くさびの歯が正しく噛むように調整する / くさびの歯が木材に噛み込む
  - (他) バネが異物を噛んだ / バネが異物を噛み込む
- 吹く/吹き出す
  - (能) (その話に) 太郎は吹いた / その話を聞いて，太郎は吹き出した
  - (対) 風が吹く / 風が煙突から吹き出す
  - (他) ろうそくに息を吹く / 穴から外へゴミを吹き出した

このように，構成動詞の項構造の多重性が複合動詞の項構造の多重性に影響を及ぼしており，その際，構成動詞と複合動詞の意味的な整合性は保たれている。筆者の内省の範囲では，上記の２動詞以外の４動詞でも，意味的な整合性が保たれる例を確認できた。

その一方で，(2) のとおり，前項動詞と後項動詞の項構造の関係に，不整合が発生している場合が多い。最も顕著なのが,「他動性調和の原則」(影山 1993) を満たさない場合である。他動性調和の原則は，同一タイプの項構造を持った動詞の組み合わせで複合動詞が構成されるとする規則であり，他動詞＋他動詞，非能格自動詞＋非能格自動詞，他動詞・非能格自動詞の組み合わせが許される。上記の 17 動詞に照らし合わせると，8 動詞で他動性調和の原則が満たされていない。例えば,「吹き出す」の「対＋他」の場合である。このとき,「吹く」の内項はガ格に相当するが,「出す」ではヲ格に対応するので,「風が吹き出す」を同一の表層構造のまま「出す」に適用すると,「＊風が出す」のように不適格な文になる。

他動性調和の原則以外でも,「立て込む」「消し込む」「盛り返す」の自動詞としての解釈において，項構造上の問題が確認できる。具体的には，前項動詞がすべて他動詞なので，前項動詞の内項が空の構造になってしまう。

4　おわりに

本稿では，複合動詞の多義性が発生するメカニズムに構成動詞がどのように関わっているかを分析し，(a)「比喩による多義性」には２種類の発生メカニズムがあること，(b)「多重項構造による多義性」を持つ 17 動詞を示し，11 動詞で構成動詞の多重項構造が複合動詞の多義性につながっていることなどを明らかにした。

謝辞

本研究の一部は，科学研究費補助金基盤研究 (B)「文脈依存の意味情報を判別する機能表現抽出 WEB システムの開発と運用実験」(代表者：松田真希子) の支援を受けた。

参考文献

山口昌也 (2013)「複合動詞・構成動詞間の意味的な共通性を計る指標の評価」，言語処理学会第 19 回年次大会予稿集，pp.757-760

影山太郎 (1993)「文法と語形成」，ひつじ書房

松本曜 編 (2003)「認知意味論」，大修館書店

国広哲弥 (2006)「日本語の多義動詞　理想の国語辞典 II」，大修館書店

瀬戸賢一 (2007)「英語多義ネットワーク辞典」，小学館

松村明 編 (2006)「大辞林第３版」，三省堂

コーパスを活用した法文データの分析に関する問題点

矢野 信（株式会社法学館法教育研究所）†

Problems in Corpus-based Studies of Legal Codes

Makoto Yano (Japan Research Institute of Law Related Education, HOUGAKUKAN CO., LTD.)

１．はじめに[1]

『現代日本語書き言葉均衡コーパス』（BCCWJ）の「特定目的サブコーパス」の一つに「法律」がある。これは，「公的な性格の強い書き言葉であり，これらの分析により言語政策に関わる基礎資料を提供することが期待できる」（丸山(2011)）として，独立したサブコーパス（以下「SC」という。）とされている。

しかし，例えば第 1～3 回の「コーパス日本語学ワークショップ」において法律 SC に言及する発表は 11 件であるなど，その分析・研究がいまだ活発であるとはいえない。

本発表では，コーパスの手法によって法文を分析する前提として，①法律（法文）の言語を分析することの位置付け，②法律 SC やその他のデータベースのコーパスとしての性質について，法律実務家・法律教育者側の視点を交えながら検討する。加えて，末尾で法律 SC などを用いて得られるデータの例を挙げる。

２．法律（法文）の言語に関係する先行研究

松田(2011)などでは，法令の文言を純粋に言語資料として用い，法令の言葉の中に見られる「ゆれ」の観察から，変異の内的要因についての考察を行なっている。そこでは，法言語学との関係についても言及がなされている。これは，分析・考察の方向性や関心において本稿とは異なるが，手段の第一として法令の言語そのものを分析する（そこにコーパスの手法を用いる）という点で共通性を有する。

このほか，法律の言葉について，法哲学者・言語哲学者による研究は数多くあるが，言語学のアプローチをとるものではないことから，ここでは省略する。法的推論を計算機で実現するという研究（自然言語処理・人工知能の観点）についても同様である。

３．法律（法文）の言語を分析することの位置付け

まず，本稿で用いる「法文」という概念について定義をした上で，（コーパスの手法による，よらないに関わらず）「法文」の言葉を分析することの位置付けを確認しておく。

３.１　用語の定義

３.１.１　「法律」

これは，大きく分けて，①国会の議決を経て成立する法形式としての「法律」の意味，②法律（法）に関係する分野・領域の２つの意味で用いられる（金子(2008)）。[2]

† klagegrund@gmail.com

[1] 法律の条文は，本来，縦書き・漢数字で表記されているが，本稿では，横書き・算用数字で表す。また，条文中の読点は本来「、」であるが，本稿では「,」に置き換えた。

なお，根拠として法令の条文を挙げる場合，法令名＋条文番号のみを掲げ，書物あるいはインターネット上の法令集を文献として掲げることはしない（ただ，インターネット上でアクセス可能な法令集を末尾の関連 URL に記載しておく。）。

[2] 一般に“法律の言葉は難しい”といわれるときは，①②どちらの意味で使われているか分からない。

本稿では，「法律」の語を基本的には①の意味でしか用いていない。②の意味では，「法律分野」などとする。

３.１.２　「法令」

公権力によって制定され，成文の形式をとる規範のことをいう（条約は含まない。）。これには，上記の意味の「法律」のほか，内閣が定める「政令」，各省庁の「○○省令」，地方自治体の「条例」，衆議院・参議院の議院規則，裁判所が定める裁判所規則などがある。（一般的な定義，法学上の定義と同じ意味で用いる。）

３.１.３　「条文」

法令を構成する個々の文のことを指す。法令を構成している規範（ルール）の最小単位であるといえる。“項目”という意味合いを込めて「条項」ということもある。次の「法文」と区別して使う。

３.１.４　「法文」

法律・法令の内容を構成している文のことをいうが，この言葉は，法学書・法律学辞典などで定義がなされていない。[3] 本稿で今後，分析の対象として取り上げるのは，この意味の「法文」である。[4]

従来，この語は，主として法哲学・言語哲学の観点から，規範（ルール）の内容を表している言葉の意味で用いられることが多かった（規範が言葉によって表されるということについての関心である。）。

ただ，本稿では，言語資料（書き言葉）としての法律・法令を分析するという観点から，成文化された法令の内容を構成する言葉を含むことにする。その中心的部分は，従来の意味の“法文”であるが，それだけではなく，例えば制定文[5]，前文[6]などの部分も含む。

３.２　「法文」と類似の性質を有するものの例

上記のような意味の「法文」の言葉は，コミュニケーションのために用いられるものではなく，個々の法文の言葉を発するという発話行為も観念できない。[7] このような意味の「法文」には，「法令」を構成する言葉のほか，表 1 に掲げるものがみな含まれると考えられる。

表 1「法文」と類似の性質を有するものの例（◎○の区別は後に触れる）

◎　解釈通達といわれるものの一部（所得税基本通達など。法令の解釈に関して，条文の形で作られている。）
◎　「行政指導指針・指導要綱」といわれるものの一部
○　各種組織の内部規則（例：○○大学学生委員会規則）
○　「公認野球規則」，「旅客営業規則」，就業規則，約款，定款，町内会規則，校則

[3] 末川(1991)，金子(2008)などには項目が存在しない。また，町田(1980)でも特に定義をせずに用いている。

[4] なお，Tiersma（アメリカの法言語学者）は，文書化された法の言葉を，①効力のある文書（法律・判決文・契約書など），②解説文書（依頼者へ事件を説明する文書，法律の教材など），③説得を目的とした文書（裁判所に提出される準備書面など）の 3 つに分類しており（Gibbons(2013)），そこでは法律と判決文とが同分類とされている。しかし，筆者の当面の目的として，規範それ自体を表した言葉とそれ以外との分離を行いたいことから，さしあたり，その分類にはよらないこととした。

[5] 法令の冒頭に置かれ，その法令の制定根拠を表現するための文章のこと。例：「弁理士法（大正 10 年法律第 100 号）の全部を改正する。」（弁理士法（平成 12 年法律第 49 号）の制定文）

[6] 法令の冒頭に置かれ，その法令の制定趣旨・目的などを述べた文章のこと。例：日本国憲法前文

[7] なお，いわゆる「公布文」（法令の公布者の意思を表明する文章。「○○法を公布する。」）は，行為遂行的な言葉（発話行為）である。

また，対等な当事者間で作成される「契約書」「和解調書」などの文書を構成する言葉も，上記のような意味の「法文」の言語的性質を備えていると考えられるが，その点自体検討を要する問題であることから，ひとまずは分析の対象から外しておく。

３.３　「法文」の分析の意義

本稿では，「法文」の言葉を分析する意義について，法令を“使う”側の視点（法律実務家・法律教育者側の視点）から，次のような点を意識しておきたい。

いわゆる「法律」の読み方については，法学部などの授業で最低限の事項は触れられるものの，法文の作り方について体系だって整理されているものは，いわゆる「法制執務」の実務の現場にしかほぼ存在しない。[8] しかし，前項のとおり，「法文」と類似の言語的性質を有するものには多様なものが含まれ，日常生活でごく一般的に触れるものもある。

「ルール」が広く集団社会一般に必須のものであるとすれば，法律家等の専門家教育の場面にとどまらず，法文の読み方・作り方についてのリテラシーを向上させる必要がある。「法文」の言語の分析をすることで，そのようなリテラシー向上に資する知見を提供することが考えられる。

以下では，そのような「法文」全体の分析のための第一歩として，主に「法律」を分析することを考えていく。

４．法律 SC やその他のデータベースのコーパスとしての性質

４.１　「法文」を言語的に分析する際の留意点

４.１.１　「法令」の著作権

著作権法 13 条には，権利（著作権）の目的とならない著作物として，「憲法その他の法令」(1 号)，「国若しくは地方公共団体の機関（中略）が発する告示，訓令，通達その他これらに類するもの」(2 号）などが定められている。

表 1 で◎としたものはみな 2 号に含まれる。一方，〇としたものは，制定主体が独立行政法人などの場合には 2 号に含まれる。

これら著作権の目的とならないものは，コーパスを公開する際に著作権の処理を必要としない点がメリットとして指摘されている（松田(2011)）。

４.１.２　法令が通用する時点・時間の問題

法令の言葉が他の種類の書き言葉と異なる最大の性質として，それが通用している時点・時間の問題がある。

法令以外の書き言葉はみな，それが作成された時点における言語という，いわば歴史的意味しか持たないと考えられるのに対して，法令の言葉は，それが効力を有するものとして通用している限り，“現在”の書き言葉としての側面を有し続ける。つまり，過去に制定・公布された法令がみな，現在の我々の日常生活を直接規律する。[9]

法令を“使う”側の視点においては，現在（現時点）において有効なすべての法令が関心の対象となり，その点を踏まえてコーパスのデータを分析・考察する必要があろう。

４.１.３　いわゆる改正法令の問題

前項に述べた法令の通用時間の点に関連して，いわゆる改正法令をどのように扱うかと

[8] 公共政策大学院では，法制執務に関しての講義が開講されているところもある。

[9] ちなみに，「法令データ提供システム」に収録されている最古の法令は，（法律ではないが）明治 5 年太政官布告第 337 号（改暦の布告）である（1872 年）。これは，大日本帝国憲法の公布（1889（明治 22）年）より 17 年も前のものであり，これが，現在も有効な法令として存在する。

いう問題ある。改正法令とは，ある法令を改正するために制定される法令のことである（以下，「法律」を例に説明するが，原理的には「法令」のすべて，また，表 1 に掲げたものについても当てはまることである。）。

既に施行されている法律Ｘを改正する際，法律Ｘを改正するための法律Ｙが公布・施行されることによって，法律Ｘの内容がＸ′に修正・変更される。このとき，法律Ｙは，法律Ｘに「溶け込む」ことによって“消えてしまう”。

「法令データ提供システム」においても，市販されている各種の法令集（六法）においても，（当然ながら）溶け込み処理がなされた後のＸ′の状態（に加えて，法律Ｙの附則部分。制定後附則といわれる）が収録されている（図 1）。

国会が制定した歴史的文書の全体という意味では，Ｘ／Ｙの側を見るべきであるともいえるが[10]，法令を“使う”側の視点においては，主にＸ′の状態に関心がある。研究目的に応じて何をコーパスの母集団と見るかにおける留意点の一つであると考えられる。

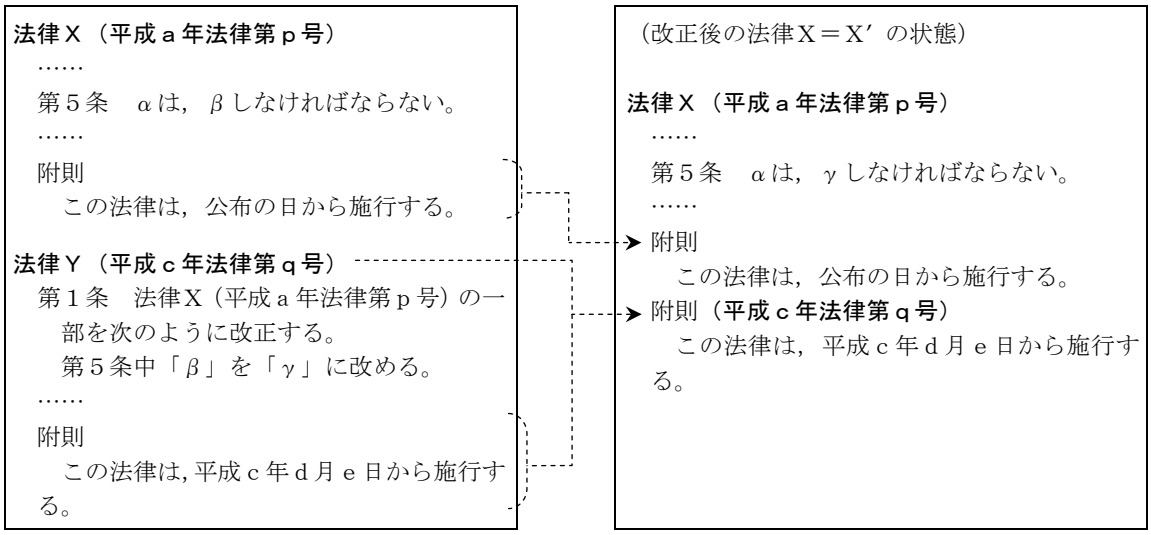


図 1　改正法令と溶け込みの関係

４.２　BCCWJ / 法律 SC のデータについて

ここでは，BCCWJ の法律 SC に収録されているデータの特徴について，丸山（2011）などから整理をした。

４.２.１　「法律」の原文との差異

データの真正性の問題として，法律の原文との違いの可能性を確認しておく必要がある。

法律 SC に収録されているデータは，「法令データ提供システム」に収録されている電子データをそのまま用いている。このことから，紙ベースの資料を電子データ化する場合に生じるような誤りの可能性は存在しない。[11]

ただ，「法令データ提供システム」に収録されている法令の電子データは，“正しい”ものであることが保障されてはいない。あくまで，法令として“正しい”ものとして効力を有するのは「官報」に掲載された状態であるというのが建て前である。[12] [13]

[10] 「日本法令索引」では，このような改正法律を含む，全ての制定法令の全文データが検索可能である。
[11] HTML タグ除去の段階の問題は残る。
[12] この意味で，「法令データ提供システム」はあくまで，日本国の公式法令集ではない。

また，ブラウザでアクセス・検索可能な状態（HTML ファイル）にする必要上，外字・ルビ・数式などの表示は原文とは異なる。また，システムの管理上，原文とは異なる処理をしている部分もある。

これらの点については，同システムの「最初にお読みください」[14] 等に掲載されている注意点がそのまま当てはまろう。

４.２.２　収録範囲の点

データの代表性に関連して留意すべき点が，収録範囲の点である。

BCCWJ では，法律 SC も含めて，テキストの収録年代範囲が限られている。法律 SC については，「1976 年から 2005 年までの間に公布され，2009 年 9 月の時点でも施行されている法律」が「対象データ」であるとされている。

この対象データには，前記の意味の改正法律は含まれていないが，さらに細かく，次のようなことが見られた。

- 制定時附則の中に含まれる「改め文」（他の法律を改正する条文）は，対象データに含まれている。[15]
- いわゆる「整備法」（大規模な法律（例えば商法，民事訴訟法など）を全面改正する際に，多数に上る改正規定・経過規定[16]などをまとめて作られる単独の法律）については，対象データに含まれている。[17]

  ただ，「整備法」の本則部分の「改め文」は含まれていないと考えられる。[18]

これは結局，BCCWJ へ収録するために「法令データ提供システム」へアクセスした時点（2009 年 9 月時点）において，「法律」として収録されていたデータのうち，廃止法令等[19]を除いたものと考えられる。上記の「対象データ」の意味は，このようなものとして理解をする必要があろう。[20]

４.２.３　書誌情報データ「出版年」の意味

法律 SC のデータには，書誌情報「出版年」として各法律の公布年が付与されている。

ところが，前記の述べたような改正法令／溶け込みの仕組みによって，実際には，「出版年」よりも後の年に生成された言葉が混入していることがある。

例として，民事執行法（昭和 54 年法律第 4 号）は，「出版年／1979 年」として表示されるが，このうち，「財産開示手続」に関する同法第 196 条以下の条文は，2004（平成 16）年に施行された改正[21]によって追加されたものである。

検索結果として得られた各データが何年に生成された言葉であるかの点を厳密に確認す

---

[13] 本稿執筆中に発見したものとして，「仮登記担保契約に関する法律」（昭和 53 年法律第 78 号）附則 4 条中の文言がある。「法令データ提供システム」「少納言」では，「ものである。」と出てくるが，正確には「ものである」（句点が不要）である（官報の画像データ（脚注 22 参照）で確認）。
なお，「日本法令索引」のデータは，正しく，「ものである」となっていた。

[14] http://law.e-gov.go.jp/readme.html

[15] 例えば，国際捜査共助等に関する法律（昭和 55 年法律第 69 号）の制定時附則の文言など多数。

[16] 旧法時代に起こった事態の効力などを定めるための条項（鈴木(2003)など参照）

[17] 実際，郵政民営化法等の施行に伴う関係法律の整備等に関する法律（平成 17 年法律第 102 号）の制定時附則の条文が「少納言」の検索結果に現れた。

[18] 法令データ提供システムに収録されないことから。

[19] 廃止法令，失効法令，実効性喪失法令のこと

[20] 例えば，動詞「改める」の頻度を得ようと法律 SC を検索するのは意味がないと考えられる。

[21] 「担保物権及び民事執行制度の改善のための民法等の一部を改正する法律」（平成 15 年法律第 134 号）による。

るには，その法律についての改正法律をすべてたどっていくしかない。[22]

４.２.３　その他の点

法律 SC においては，「法令データ提供システム」に設定されている 50 の事項分類が書誌情報「ジャンル」として設定されているが，BCCWJ に収録する際のサンプリングの結果として，全部で 43 項目になっている。ジャンルごと／年代ごとの語数の一覧表[23]には，法律 SC の収録年代範囲を 5 年ごとに区切ったセグメントで，ジャンルごとの語数が掲載されているが，そこでは，全 258 個の小セグメント中 143 個（55.4%）が語数 0 となっている。これらのことから，法律 SC を「ジャンル」「年代」を区切って検索した結果は，頻度を算出するのに意味を持たないことが多いと考えられる。

また，収録年代範囲が限られていることから，法律用語の検索には向かないことが多いと考えられる。前述のように法律は明治期以来のものが現在そのままの形で有効に存在し，かつ，改正を経ても原則として法律番号が変わらないことから[24]，我々の日常生活に密接に関連する基本的な法律[25]は，法律 SC にはあまり含まれていないことが予想される。

５．法律 SC などを用いて得られるデータの例

５.１　対照で用いたデータ

「法令データ提供システム」では，システムにアクセスした時期に有効な[26]現行法令のデータを全文入手すること可能であり，このデータ（以下「MIC データ」という。[27]）をもとにコーパスを構築することができる。「法令データ提供システム」のサイト上では，任意語による単純な全文検索は可能であるが，正規表現など複雑な検索には対応していないことから，複雑な分析を行うには，必要な範囲のデータをダウンロードして用いることになる。[28]

一方，BCCWJ の検索アプリケーション『中納言』では，形態論情報を用いた検索，共起条件を指定した検索などが可能であること，他のレジスターとの比較が可能であることなど，MIC データのみでは得られない情報を得られる可能性がある。

５.２　検索結果の例[29]

今回，検索結果のサンプル例として，法律分野の言葉に特徴的であるといわれてきた 2

---

[22] 主要な法律については過去の法令集（六法全書）を調べればよい。また，原理的には，「日本法令索引」の制定法令を検索することで可能である。政令・規則（○○省令など）については，過去の官報を検索するしかない（官報には，全ての法令の原文・全文が掲載され，それによって法令が「公布」される。）。過去の官報は，「官報情報検索サービス」で全文検索が可能である（有料）。

[23] 「中納言オンラインマニュアル」に掲載されている。

[24] 例えば，「民法」は明治 29 年法律第 89 号，「刑法」は明治 40 年法律第 45 号といった具合である。

[25] 多くの人がイメージする法律としては，明治期に近代国家としての法整備をするために制定された法律群（民法，商法，刑法など）や，戦後間もなくの時期に新憲法下での新しい国家体制を整備するために制定された法律群（地方自治法，国家公務員法，労働組合法など）のものが多いと考えられる。これらは，実務でよく使われる法律でもある。

[26] システムの更新頻度の都合から，アクセス日現在の法令に厳密には一致しない。更新は約 1 か月ごとに行われているようである。

[27] MIC は総務省の英語略記である。

[28] 今回，MIC データをダウンロードしてテキストファイル化するところまでを自動化する Perl スクリプトを作成して，2013 年 5 月 20 日現在でアクセスして得られた MIC データを分析に使った。

[29] ここで示す MIC データは，法律 SC の収録対象データと同じく，法律番号が昭和 51（1976）年から平成 17（2005）年の間の法律のみから検索した結果である。ただ，元データの取得時点が異なることから（脚注 28），法律 SC のデータと厳密には同じではない（法律 SC のデータ取得時点以降に改正・廃止された法律の部分が異なることになる。）。

つの表現について，頻度を算出した。

５.２.１　「この限りで（は）ない。」

法文のただし書きなどで，一定の条件の下で本体のルールが及ばないことを表す意味に用いられる。検索結果を表 2 に示す。

- 「は」が入っているかどうかで法文としての意味は変わらないと思われるが，実際に検索で出現するのは，圧倒的多数が「は」を含まない形であった。
- 表には記載していないが，BCCWJ の「雑誌」「新聞」「教科書」における検索結果は，いずれの形も 0 であった。これは，この言い回しを学ぶ機会が稀である可能性が高いことをうかがわせる。この点は，法教育の観点から検討の必要があろう。

表 2 「この限りで（は）ない。」

	「少納言」（法律 SC）		MIC データ（法律）	
この限りでない。	281	100.0%	1,506	99.7%
この限りではない。	0	0.0%	4	0.3%

５.２.２　「ものである（。）」

法律分野でよく見られる言い回しの例として，しばしば挙げられる表現である（判決におけるこの表現について言及するものとして，田中(2012)）。検索結果を表 3 に示す。

表 3 「ものである（。）」

	「少納言」（法律 SC）		MIC データ（法律）		MIC データ（政省令等）	
ものであること。	168	46.4%	692	43.3%	2,486	56.5%
ものであること…	24	6.6%	145	9.1%	409	9.3%
ものであるとき	69	19.1%	367	23.0%	404	9.2%
ものである場合	55	15.2%	299	18.7%	580	13.2%
ものである。	0 *	0.0%	5 *	0.3%	0	0.0%
その他	46	12.7%	127	7.9%	525	11.9%
（合計）	362	—	1,598	—	4,404	—

(*)脚注 13 で述べた 1 件を除外した値である。

法律 SC と MIC データ（法律）とでは，出現形の比率は大きくは異ならないことが分かる。一方，MIC データ（法律）と MIC データ（政省令等）とでは，出現形の比率の分布が少し異なる。法律と政省令とでは，規定される対象・内容が異なるが[30]，その点と関係がある可能性がある。この点は今後の検討課題の一つである。

また，MIC データ（法律）には「ものである。」が 5 件あるが，これはいずれも，法律の前文の文字列であった。[31]

[30] 法律は，国民の権利・義務に直接関わる事項を定めるのに対して，政省令は，法令から委任を受けた事項や法令の施行のために必要な技術的な細目を定めるというように，対象分野・内容がやや異なる。

[31] 感染症の予防及び感染症の患者に対する医療に関する法律，ハンセン病療養所入所者等に対する補償金の支給等に関する法律など。

６．おわりに

本稿では，「法文」の言語の分析のためにコーパスを用いる際の前提事項を整理しつつ，「法文」の代表格としての「法律」の分析の例を挙げた。ここで整理した内容を踏まえて，今後，「法文」の具体的な特徴の分析を進めていきたい。

文　献

John Gibbons (1976) *Forensic Linguistics : an Introduction to Language in the Justice System*
（邦訳『法言語学入門　司法制度におけることば』，東京外国語大学出版会，2013）
金子宏，新堂幸司，平井宜雄(2008)『法律学小辞典　第 4 版補訂版』有斐閣
末川博，他(1991)『新法学辞典』日本評論社
鈴木達也(2003)「経過規定と旧法令の効力」立法と調査 237 号
（参議院法制局「法制執務コラム集」として次の URL に転載されている。
http://houseikyoku.sangiin.go.jp/column/column051.htm）
田中伊式(2012)「『この事件は，～したものです』などの表現をめぐって」放送研究と調査 2012 年 5 月号，pp.72
（https://www.nhk.or.jp/bunken/summary/research/kotoba/037.html よりダウンロード可能）
法制執務研究会(2007)『ワークブック法制執務』ぎょうせい
前川喜久雄(2013)「コーパスの存在意義」『講座日本語コーパス 1 コーパス入門』，pp.1-31，朝倉書店
町田顕(1980)「法文の長さ」判例タイムズ，400 号，pp.38
松田謙次郎(2011)「法令の言語変異を探る」トークス，14 号，pp.23-43
（http://ci.nii.ac.jp/naid/110008095658 よりダウンロード可能）
松田謙次郎(2012)「法令に見られるサ変動詞の五段化・上一段化について　2001 年から 2011 年のデータ分析」トークス，15 号，pp.37-48
（http://ci.nii.ac.jp/naid/110008799628 よりダウンロード可能）
丸山ほか(2011)「『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装」国立国語研究所内部報告書(LR-CCG-10-02)
矢野信(2013)「言語資料としての「判決文」の分析にまつわる問題点」，本予稿集収録

関連 URL

法令データ提供システム（総務省）　http://law.e-gov.go.jp/cgi-bin/idxsearch.cgi
日本法令索引（国立国会図書館）　http://hourei.ndl.go.jp/SearchSys/
衆議院規則（衆議院）　http://www.shugiin.go.jp/index.nsf/html/index_houki3.htm
参議院規則（参議院）　http://www.sangiin.go.jp/japanese/aramashi/houki/kisoku.html
最高裁判所規則集（最高裁判所）　http://www.courts.go.jp/kisokusyu/
自治体 Web 例規集へのリンク集（洋々亭）　http://www.hi-ho.ne.jp/tomita/reikidb/reikilink.htm
官報情報検索サービス　（独立行政法人国立印刷局）　https://search.npb.go.jp/
少納言（国立国語研究所）　http://chunagon.ninjal.ac.jp/
中納言オンラインマニュアル　https://maro.ninjal.ac.jp/wiki/

同一見出し語の出現間隔の分布と文体差

山崎　誠（国立国語研究所言語資源研究系）†

Distribution of Gaps between Successive Occurrences of the Same Word and Stylistic Differences

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

計量語彙論の基本的なテーマのひとつに同一の見出し語の出現間隔の問題がある。水谷(1983)によれば、古くは、Epstein(1953)がプーシキンの『大尉の娘』をデータとしてロシア語の前置詞 κ の出現間隔の分布を、 $F(x) = 1 - e^{-\lambda x}$ で近似した例がある（ただし、x は出現間隔、λ は当該の語の使用率。この場合 λ は、8.3×10^{-3}）。Herdan(1966:127-130)は、これをさらに進めて機能語の出現はポワソン分布に従うのではないかということを示唆した。

本稿では、同一見出し語の分布がテキストの持つ特性、とくに、文体ないしはレジスターと呼ばれるものと何らかの関係を持つのではないかという想定のもとに 2 つの調査でそれを確かめることを目的とする。ひとつは、高頻度語の出現間隔の分布、もうひとつは、ひとつのテキストに現れるすべての同一見出し語の出現間隔の分布である。

２．出現間隔の測り方

テキスト中に現れる語は、使用頻度 1 の語以外は、2 回以上繰り返して出現することになる。その際、同一の語が 2 回出現する場合、それらの間の距離を、間に含まれる言語単位の数で測った値を出現間隔とする。本稿では、同一語の 2 回出現の間に他の語が x 語存在している場合、その 2 語の間の出現間隔を x+1 とする。したがって、同一語が隣り合って出現する場合はその出現間隔は 1 となる。また、出現間隔は、着目している出現と直前あるいは直後の出現との距離を測り、間にひとつ以上同じ語をはさんだ距離は出現間隔の測定の対象外とする。以上の手続きにより、例えば、あるテキストで使用頻度 x(x≧2)の語の持つ出現間隔の総数は、x－1 回となる。

３．データ

本稿で使用するデータは、『現代日本語書き言葉均衡コーパス』(BCCWJ)のコアデータである。意味的なまとまりを重視するため、可変長データを用い、言語単位は短単位を利用した。また、語数のカウントにあたって品詞が空白ないしは品詞欄に「記号」の文字列を持つ語[1]を除外した。以下、本稿では語数という場合はこの方法による。

４．高頻度語（機能語）の出現間隔

ひとつめの調査は、コーパス開発センターで公開されている、BCCWJ の短単位語彙表データにおいて BCCWJ 全体の順位の上位 10 語を高頻度語とし、調査対象とした。具体的には、以下の 10 語である。「の（格助詞）、に（格助詞）、て（接続助詞）、は（係助詞）、だ（助動詞）、を（格助詞）、た（助動詞）、する（動詞）、が（格助詞）、と（格助詞）」

表 1 は、高頻度における出現間隔の分布のようすである。いずれも平均よりも中央値が小さく、分布が低いほうに偏っていることが分かる。例として格助詞「の」の出現間隔のヒストグラムを図 1 に挙げた。表 1 に挙げた語はいずれもこのような分布を示している。

† yamazaki@ninjal.ac.jp

1 品詞が補助記号であるものと記号であるものとが該当する。

ただし、表 2 に示したように、出現間隔の値を対数（自然対数）で表すと、正規分布に近くなるように見える[2]。

表 1　高頻度語の出現間隔の分布

語	出現間隔数	平均値	最小値	中央値	最大値
の（格助詞）	46,485	18.848	1	13	315
に（格助詞）	31,007	27.401	1	19	461
て（接続助詞）	27,465	30.314	2	20	1289
は（係助詞）	26,043	32.115	2	23	784
だ（助動詞）	24,395	43.610	1	20	2212
を（格助詞）	27,288	30.245	1	20	481
た（助動詞）	22,982	35.068	2	20	1196
する（動詞）	22,845	35.663	1	23	745
が（格助詞）	20,820	39.350	2	26	737
と（格助詞）	18,439	43.438	1	29	776

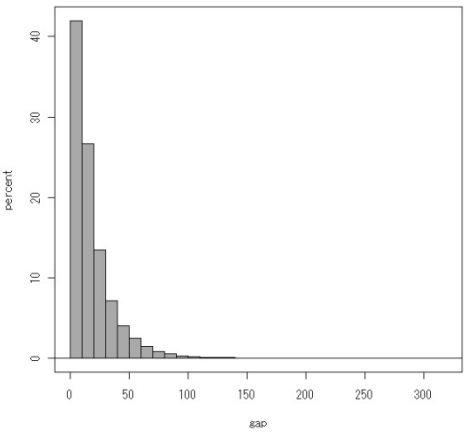


図 1 格助詞「の」の出現間隔の分布

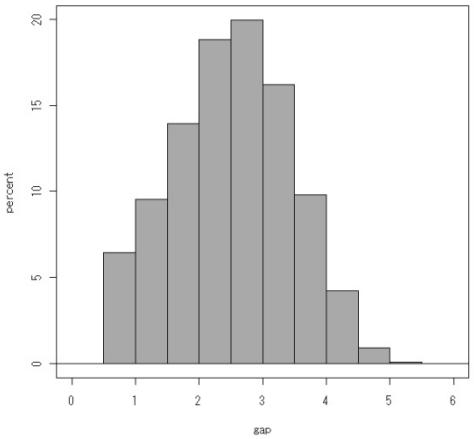


図 2 格助詞「の」の出現間隔の分布（対数）

出現間隔を対数に変換したものを使い、表 1 の 10 語について、R の多重比較（チューキーの HSD 検定）を行った。表 2 は、各語ごとに、コアデータを構成する各レジスターの組み合わせのどれに有意差があったかを示したものである。表 2 から、以下のようなことが見てとれる。(1)多くの組み合わせに有意差が認められること。(2)「の」「を」の出現間隔の分布は、他と比べて有意差の組み合わせが少ない。(3)PB（出版書籍）と OY（Yahoo!ブログ）の組み合わせはここで挙げた高頻度語の出現間隔の分布では大きな差がない。

表 2 レジスターの組み合わせと有意差

組合せ	の	に	て	は	だ	を	た	する	が	と
OW-OC	***		***	***	***		***	**	***	***
OY-OC		***	***	***		*	**	***	***	***
PB-OC		***	***	***	***		***	***	***	***
PM-OC		***	***	***	***		***	***	***	***

[2] ただし、バートレット検定を行うと各群（OC、OW、OY、PB、PM、PN）の分散は等しくない。(df=5、p=2.2e-16)

PN-OC		***	***	***	***		***	***	***	***
OY-OW	***	***	***	***	***		***	***	***	***
PB-OW	***	***	***	***	***		***	***	***	***
PM-OW	***	***	***	***	***		***	***	***	***
PN-OW	***	***		***	***	***	***	***	***	***
PB-OY					**	*	*			
PM-OY			***		***		***	*	*	***
PN-OY			***		***	***	***			***
PM-PB		***	***	***	***		***			***
PN-PB		**	***	***	***	***	***	***		***
PN-PM	***		***		***	***		***		

（注）***：0.1%水準で有意差、**：1%水準で有意差、*：5%水準で有意差

５．全出現間隔の分布

テキストに出現する語は、頻度 1 以外は出現間隔を持つが、それらすべての出現間隔の分布はどのようになっているだろうか。LB（図書館書籍）、OW（白書）、OY（Yahoo!ブログ）の 3 つのレジスターについて観察した。それぞれのレジスターにおいて可変長部分で短単位の語数が 2000 語～2099 語のサンプルを選んだ。OW と OY は該当するサンプルのすべて（26 サンプルずつ）、LB は、ランダムに抜き出した 22 サンプルである。その一覧を表 3 に挙げる。

表 3 全出現間隔の分布

サンプル ID	Token[3]	Type[4]	総個数[5]	平均[6]	標準偏差	ジャンル（NDCなど）
LBd0_00011	2078	623	1473	136.3252	251.0421	0 総記
LBe0_00003	2049	607	1458	130.7236	245.2134	0 総記
LBl1_00007	2069	418	1663	105.8443	215.3948	1 哲学
LBq1_00035	2086	417	1676	120.3365	232.6101	1 哲学
LBe2_00056	2052	536	1528	128.0118	224.7713	2 歴史
LBr2_00028	2096	619	1490	113.943	224.3516	2 歴史
LBg3_00067	2073	525	1770	166.4633	307.9518	3 社会科学
LBo3_00020	2021	390	1643	112.5526	229.1028	3 社会科学
LBh4_00036	2059	414	1658	111.7835	214.3982	4 自然科学
LBo4_00014	2022	522	1531	137.1633	245.3251	4 自然科学
LBm5_00009	2061	470	1601	132.6552	245.3482	5 技術･工学
LBm5_00011	2047	562	1502	129.8755	252.5623	5 技術･工学
LBb6_00012	2021	535	1516	133.0389	234.9972	6 産業
LBj6_00011	2095	605	1507	132.8653	249.6481	6 産業
LBo7_00002	2025	589	1449	154.8075	271.912	7 芸術･美術
LBs7_00075	2039	676	1384	130.2536	250.1484	7 芸術･美術
LBb8_00005	2030	325	1749	90.52144	196.4048	8 言語
LBs8_00014	2079	656	1454	125.3521	221.0104	8 言語
LBr9_00016	2047	657	1404	152.6489	282.0645	9 文学

3 延べ語数。

4 異なり語数。語彙素、語彙素読み、語彙素細分類、品詞の 4 つの属性が一致したものを同一の語として集計した。

5 当該サンプルの可変長部分における同一語の出現間隔の全個数。

6 当該サンプルの可変長部分における同一語の出現間隔の総和を全個数で割ったもの。

LBr9_00093	2078	557	1976	143.3011	295.4928	9 文学
LBmn_00016	2050	485	1681	130.8917	247.1816	N 分類なし
LBsn_00024	2019	565	1624	142.3565	296.3724	N 分類なし
OW1X_00089	2056	301	1755	116.159	186.6517	科学技術
OW1X_00457	2053	421	1632	119.4424	203.7048	安全
OW1X_00540	2086	350	1736	119.4764	217.0649	国土交通
OW2X_00030	2033	357	2179	140.8123	256.6013	農林水産
OW2X_00949	2008	489	1637	151.843	269.7998	福祉
OW2X_00960	2016	272	1744	96.06479	159.5637	安全
OW3X_00044	2062	450	1825	142.5326	290.3834	国土交通
OW3X_00046	2083	582	1706	179.2093	315.1336	外交
OW3X_00198	2000	508	1809	152.382	278.9863	農林水産
OW3X_00205	2076	451	1625	138.4732	246.2405	安全
OW3X_00359	2051	452	1599	138.9962	236.5418	環境
OW3X_00435	2059	419	1640	113.1311	196.4604	安全
OW3X_00562	2069	479	1590	110.4755	204.0817	安全
OW4X_00145	2060	416	1940	143.2495	264.8059	環境
OW4X_00197	2061	325	1736	108.8278	183.9437	福祉
OW4X_00238	2031	343	1942	122.7549	239.4015	環境
OW4X_00536	2098	416	1682	125.9417	228.489	安全
OW4X_00648	2084	318	1766	106.9468	188.3023	国土交通
OW5X_00025	2081	266	1815	105.4997	160.1334	福祉
OW5X_00170	2093	624	1469	150.8754	263.0187	環境
OW5X_00663	2033	585	1448	149.9841	269.2003	外交
OW5X_00853	2032	481	1551	113.0129	206.1083	福祉
OW5X_01851	2012	508	1730	145.9046	284.9876	安全
OW6X_00035	2007	264	2103	140.8783	242.0626	経済
OW6X_00090	2004	447	1557	129.9332	249.6389	環境
OW6X_00190	2082	467	1615	124.2533	240.2149	環境
OY01_02573	2063	400	1663	120.9784	223.2255	ビジネスと経済
OY01_02830	2099	525	1574	155.1652	244.3484	ビジネスと経済
OY04_02358	2096	623	1473	123.7502	236.6436	エンターテインメント
OY04_02691	2046	428	1618	116.2979	210.2759	エンターテインメント
OY04_03782	2077	532	1545	116.2979	210.2759	エンターテインメント
OY04_04825	2067	529	1538	128.4018	247.8479	エンターテインメント
OY04_06866	2013	513	1500	145.018	259.8423	エンターテインメント
OY06_01586	2094	625	1469	140.998	275.7961	政治
OY13_00240	2006	611	1395	131.0616	244.4963	芸術と人文
OY13_03195	2039	508	1531	127.1346	238.8284	芸術と人文
OY13_03612	2090	671	1419	142.561	260.089	芸術と人文
OY14_03639	2090	478	1612	127.6917	222.3718	Yahoo!サービス
OY14_06486	2009	617	1392	119.403	227.533	Yahoo!サービス
OY14_07107	2000	503	1497	134.5137	248.3512	Yahoo!サービス
OY14_12047	2063	675	1388	149.438	258.3771	Yahoo!サービス
OY14_12048	2059	675	1384	149.8215	258.5773	Yahoo!サービス
OY14_36191	2026	651	1375	125.8211	250.4715	Yahoo!サービス
OY14_37664	2045	616	1429	155.3611	288.0154	Yahoo!サービス
OY14_50772	2008	438	1570	135.235	234.4837	Yahoo!サービス
OY15_00485	2069	757	1312	148.3941	259.358	趣味とスポーツ

OY15_01982	2074	803	1271	148.8891	279.3363	趣味とスポーツ
OY15_02404	2015	697	1318	148.9484	278.0393	趣味とスポーツ
OY15_07093	2074	691	1383	121.5582	220.2368	趣味とスポーツ
OY15_17336	2084	749	1335	153.4022	285.8187	趣味とスポーツ
OY15_17721	2008	628	1380	151.259	259.3086	趣味とスポーツ
OY15_18790	2037	515	1522	119.4087	238.4768	趣味とスポーツ

前節で高頻度語においては、レジスターによって出現間隔の分布が異なっているものが多いことが確認されたが、テキスト全体の出現間隔はレジスターにより異なるだろうか。出現間隔の値を対数に変換して[7]、R の多重比較を用いたところ、表 4 のような結果を得た。どの組み合わせにおいても 5%水準で有意差は認められなかった。しかし、表 5 に示したように、出現間隔の総個数[8]で比較したところ、いずれも組み合わせも 1%の水準で有意差があった。また、表 6 は Type（異なり語数）による比較であるが、OW（白書）と LB（図書館書籍）、OY（Yahoo!ブログ）と白書（OW）に 1%水準で有意差が認められた。出現間隔全体の分布はレジスターによる特徴がないことから、むしろテキストの普遍的な属性として考えられる可能性がある。この点についてはさらに多くのレジスターのテキストについて考察が必要である。

表 4 出現間隔の平均値による比較

組み合わせ	P 値
OW-LB	0.08873
OY-LB	0.13782
OY-OW	0.13463

表 5 出現間隔の総個数による比較

組み合わせ	P 値
OW-LB	0.0023
OY-LB	0.0062
OY-OW	$<1e^{-04}$

表 6 Type による比較

組み合わせ	P 値
OW-LB	0.000402
OY-LB	0.185692
OY-OW	$<1e^{-04}$

ところで、出現間隔の総個数が多くなると、理屈の上では、平均出現間隔が小さくなると期待されるが、実際にデータを見てみると図 3 に示したように、出現間隔の総個数と平均出現間隔との間には相関は見られない（相関係数は－0.1867）。一方、出現間隔の総個数

図 3 出現間隔総個数と平均出現間隔

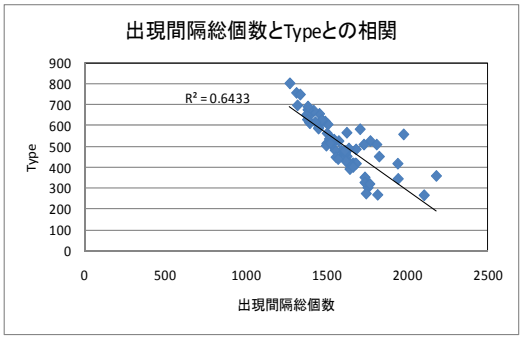


図 4 出現間隔総個数と Type

[7] バートレット検定を行ったところ各群（LB、OW、OY）の分散には有意差がなかった（df=2、p=0.1411）。

[8] 対数に変換した値を用いた。表 6 の Type も同じ処置である。

は、図 4 に示すように Type（異なり語数）と強い負の相関がある。出現間隔が増えることは個々の見出し語の使用頻度が多くなることを意味する。今回扱ったサンプルは延べ語数がほぼ一定なので、使用頻度の多い見出し語が増えれば、相対的に Type が少なくなるものと想定される。また、当然のことであるが、出現間隔の総個数と Token（延べ語数）との間には相関はない。

６．まとめと今後の課題

本稿では、テキストにおける見出し語の出現間隔の分布と文体（レジスター）との関係を考察した。その結果、高頻度語（主に機能語）の出現間隔は、BCCWJ のレジスターによって違いがあることが分かった。また、テキストにおける出現間隔の平均は、レジスターによる違いは見られないが、出現間隔の総個数がレジスターにより違いがあった。

今回は、出現間隔と文との関係は考慮しなかった。出現が 1 文の中で起きたものか、それとも文を超えるものであるか、その違いも考察の対象となるだろう。また、テキストを文の連続と考え、何番目の文に同じ見出し語が出現しているかというとらえ方も可能であろう。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

Epstein, B. (1953) "Some Remarks on the Length of Gap Between Successive Occurrences of High Frequency Words" in Josselson, H.H.(1953) "Russian Word Count", Wayne University Press.

Herdan, G. (1966) The Advanced Theory of Language as Choice and Chance. Springer-Verlag

水谷静夫(1983)　『朝倉日本語講座 2 語彙』朝倉書店

書　名　第4回 コーパス日本語学ワークショップ予稿集
発行日　平成25年9月2日
発行者　国立国語研究所 言語資源研究系・コーパス開発センター
http://www.ninjal.ac.jp/organization/chart/03/
http://www.ninjal.ac.jp/organization/chart/06/
連絡先　〒190-8561　東京都立川市緑町10-2
大学共同利用機関法人 人間文化研究機構 国立国語研究所コーパス開発センター内
電　話　042-540-4300（代表）