第1回 コーパス日本語学ワークショップ予稿集

著者	国立国語研究所 言語資源研究系・コーパス開発センター
雑誌名	コーパス日本語学ワークショップ予稿集
巻	第1回
ページ	1-402
発行年	2012-03-01
URL	http://doi.org/10.15084/00003420

第1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1回 コーパス日本語学ワークショップ

予稿集

2012年3月5日、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１回 コーパス日本語学ワークショップ

予稿集

2012年3月5日（月）／6日（火）

Program [プログラム]

3月5日(月)

10:00～10:10 **■挨　拶**　前川 喜久雄

■口頭発表（1）

10:10～10:40 **程度的な名詞と尺度形容詞類の共起傾向の推移**

▷服部 匡

10:40～11:10 **「少納言」「中納言」検索結果活用ツール**

▷田野村 忠温

11:10～11:40 **統計的機械学習による歴史的資料への濁点の自動付与**

▷岡 照晃

11:40～12:10 **古代日本語の主節の無助詞名詞句 ―活格性との関わりから―**

▷竹内 史郎

12:10～13:00 **昼食・休憩**

13:00～15:00 **■ポスター発表（1）**

『日本語話し言葉コーパス』を用いた「全然」の変化の詳細化

▷佐野 真一郎

「かなしい」と「つらい」の意味について

▷加藤 恵梨

現代日本語におけるコロケーション：検出と分析

▷STRAFELLA Elga Laura、林部 祐太、松本 裕治

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

▷姜 紅

外来語由来の接尾辞「チック」と類義語との比較

▷村中 淑子

語義曖昧性解消のための領域適応手法の決定木学習による選択 ―三手法からの決定―

▷古宮 嘉那子、奥村 学

形態素と文字の情報を用いた中国語形態素解析

▷侯 海霞、古宮 嘉那子、柴原 一友、藤本 浩司、小谷 善行

Web関連度と確率的翻訳モデルを併用した質問応答システム

▷阿部 裕司、森田 一、古宮 嘉那子、小谷 善行

文の長さ分布に見られる対数正規性

▷古橋 翔

言語接触の観点からみた非有生名詞主語の「見る」構文 ―文語体コーパスを利用して―

▷高橋 暦、堀江 薫

文書分類における補集合を併用したNaive Bayes

▷伊藤 裕佑、古宮 嘉那子、小谷 善行

日本語並立助詞「と」・「や」と英語冠詞に関する一考察 ―BCCWJデータに基づいて―

▷川口 裕子

日本語と英語の対訳文対の収集と著作権の考察

▷村上 仁一、藤波 進

テキストの硬さと軟らかさの考察 ―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―

▷柏野 和佳子、立花 幸子、保田 祥、丸山 岳彦、奥村 学、佐藤 理史、徳永 健伸、大塚 裕子、佐渡島 紗織

「語り性」を有する書きことばの典型例の分析

▷保田 祥、柏野 和佳子、立花 幸子、丸山 岳彦

反復語の使用実態から見る話し言葉と書き言葉の連続性 ―コーパスを用いた定量的分析を通して―

▷鯨井 綾希

モラウとイタダクのヲ格名詞・動名詞の違いについて

▷岩井 智哉

コーパスを用いた中国語ネット語の判定システム

▷竇 梓瑜、古宮 嘉那子、小谷 善行

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―

▷小磯 花絵、石本 祐一

コーパス管理ツール「茶器」による中古和文コーパスの利用

▷小木曽 智信

大規模コーパスを用いた用例の典型性評価 ―大規模コーパスを利用した学習辞書作成のために―

▷千葉 庄寿

■口頭発表（2）

15:00～15:30 **テキストの難易度に対する人間の判断と機械の判断**

▷佐藤 理史、柏野 和佳子

15:30～16:00 **大規模コーパスの利用とメタデータの役割**

▷丸山 岳彦

16:00～16:30 **「形容詞＋です」述語の生起要因についての準備的考察**

▷前川 喜久雄

16:30～17:00 **共起語率の分布からみるテキストの語彙的特徴**

▷山崎 誠

3月6日（火）

■口頭発表（3）

10:00～10:30 **多様な様式を網羅した会話コーパスの共有化**

▷伝 康晴、土屋 智行、小磯 花絵

10:30～11:00 **通時コーパスをどう使うか**

▷近藤 泰弘

11:00～11:30 **通時コーパスと言語空間論**

▷山元 啓史、田中 牧郎、近藤 泰弘

11:30～12:00 **近代語史をとらえるための文献選定とコーパス**

▷田中 牧郎

12:00～13:00 **昼食・休憩**

13:00～15:00 **■ポスター発表（2）**

『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

▷渡辺 美知子、清水 信哉

明治初期論説文における一人称代名詞の分析 ―『明六雑誌』コーパスを用いて―

▷近藤 明日子

日英の理工系口頭発表コーパスの構築と検索サイトJECPRESE

▷林 洋子、国吉 ニルソン、野口 ジュディ、東條 加寿子

日本語対話コーパスにおける倒置構文について：聞き手の反応に注目して

▷郭 潔、伝 康晴

現代日本語書き言葉均衡コーパスに基づく外来語音の表記に関する試論

▷単 珊、白勢 彩子

「リアル」を構成要素とする複合名詞の語彙的特徴

▷渡邊 ゆかり

機能動詞結合における動詞の選択制約 ―「影響を与える」と「影響する」―

▷岡嶋 裕子

BCCWJと学習者作文コーパスを利用した日本語作文支援

―表記と共起に関する誤用添削プロトタイプ構築―

▷八木 豊、ホドシチェク・ボル、仁科 喜久子

コーパスに基づく現代語表記のゆれの調査 ―BCCWJ コアデータを資料として―

▷小椋 秀樹

「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを考える際に必要となる視点は何か？ ―「同意要求表現」を中心に―

▷岡田 祥平、江崎 哲也

BCCWJにおける出典情報とトピックおよびレジスターとの関係

▷ホドシチェク・ボル、仁科 喜久子

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―

▷田頭（谷口）未希

用例に基づく複合動詞の構造分析と教育への応用

▷山口 昌也、井上 優、柏野 和佳子、北村 雅則、白井 清昭、千葉 庄寿

日本語話し言葉コーパスにおける句末音調のバリエーション

▷菊池 英明、宮島 崇浩

『日本語話し言葉コーパス』における句末境界音調のピッチレンジ制御

▷五十嵐 陽介、小磯 花絵

『現代日本語書き言葉均衡コーパス』における複合辞の検出と評価

▷鈴木 敬文、阿部 佑亮、宇津呂 武仁、松吉 俊、土屋 雅稔

階層的機能表現辞書の意味的等価クラスおよび対訳用例を用いた機能表現の日英翻訳

▷阿部 佑亮、鈴木 敬文、宇津呂 武仁、山本 幹雄、松吉 俊、河田 容英

日本語音声コーパスにおける促音・非促音の判別

▷天野 成昭、山川 仁子、近藤 眞理子

話し言葉が伝えるものとは、結局何なのか？ ―概念の整理および課題―

▷森 大毅

『日本語話し言葉コーパス』RDBの構築

▷小磯 花絵、伝 康晴、前川 喜久雄

15:00～16:30 **■シンポジウム**

コーパスアノテーションと心理言語学

▷浅原 正幸、小野 創、狩野 芳伸

Contents [目次]

口頭発表（1）

3月5日（月） 10:10〜12:10

程度的な名詞と尺度形容詞類の共起傾向の推移

服部匡（同志社女子大学表象文化学部）

Changes in the Cooccurrence Patterns between Nouns with Gradable Properties and Scalar Adjectivals

Tadasu Hattori (Doshisha Women's College of Liberal Arts)

１．はじめに

筆者は服部(2011a)で、「二字漢語+{性・度・率・量・力}」の形で程度的属性を表わす名詞に対してその値の大きさを述べるのに複数の形容詞類が用いられる場合について、電子的に公開されている 1947 年以降の 60 年間の国会会議録を用いた通時的分析を行った。

例えば、(連体修飾部を伴う/伴わない)「可能性」という名詞では、「大きい・高い・強い・多い・濃い」などの形容詞と主述関係を構成する用例があり、どの場合も (全く同義かはともかく)抽象的な次元では可能性の程度の大きさを述べている。各形容詞は、その最も基本的な用法ではそれぞれ異なる次元の尺度に対応しているが、派生的用法で「可能性」が表わすような抽象的な程度的属性に関して用いられる場合には次元の相違がいわば中和して、どれも(あるいはいくつか)を用いうることがあるのである。しかも各名詞の形容詞選択傾向には通時的変化が明瞭に見られた。

ここでは観察対象を広げ、何らかの意味で程度的と言い得る名詞[1]一般について、各形容詞との共起傾向[2]の推移を分析し、名詞の意味的特徴により、いくつかの推移のパターンが見られることを指摘する。詳細なデータは服部(2011b),服部(2012)に示している。

２．対象とする用例の範囲・用語の定義

国会会議録のデータに形態素解析プログラム MeCab(0.97)と電子辞書 UniDic(1.3.12)による形態素解析を施し、次に当たる表現を調査対象として抽出した。

(1)　調査対象とする表現の範囲

次の語類が「名詞{が・は・も・の}形容詞類」の連接をなし意味的に主述関係にあるもの。

名詞：表記上漢字か片仮名で始まる名詞。ただし、文頭にあるか、直前の字種が「漢字・片仮名」以外のもの。複合的な名詞は除くが、第 2 要素が 1 文字漢語である 2 要素語(「○○性」など)は含む。前に連体修飾等の成分があるものもないものも含む。

形容詞類：　高い、大きい、多い、強い、深い、濃い、重い、大(ダ)、濃厚(ダ)

用例を、発話時期により次の３つに分ける。

(2) 発話時期の区分

Ⅰ期　1947-1966 年
Ⅱ期　1967-1986 年
Ⅲ期　1987-2006 年

---

[1] 程度的ということの判定を厳密にはできず、本研究に関してはその意義も乏しいと思われる。

[2] ある名詞に対して問題となる複数の形容詞が常に意味を変えずに置換可能であると主張するものではない。また名詞そのものに多義を認めるべき場合もあると思われる。

共起用例数とは、ある名詞とある形容詞が(1)に示した形で結合した用例の数である。**総共起用例数**とは、ある名詞に対する、(1)にあげた全形容詞類の共起用例数の合計であり、**共起率**とは、共起用例数を総共起用例数で割ったものである。**共起率差分**とは、Ⅲ期の共起率の値からⅠ期の共起率の値を引いたものである。

少なくとも２つの期に総共起用例数が 100 以上ある名詞を分析に用いることにする。ある名詞に総共起用例数が 100 未満の期がある場合は、その期に*をつけて示す。「会社」「日本」「大臣」「子供」のように、程度的でないことが明らかな名詞は、調査対象から除外した。対象とする名詞は全部で 168 語である。

３．名詞に対する各形容詞の共起傾向の通時的推移

各名詞の形容詞別共起傾向推移の様相を観察する。まず、共起率差分の絶対値が 5 パーセントポイント(以下「ポイント」)以上[3]の名詞の数を形容詞別に示すと次のようである。

表１　共起率変動幅の大きい名詞の数

	高い	大きい	多い	強い	深い	重い	濃い	大	濃厚
上昇	**63**	**43**	3	**19**	1	2	2	0	0
下降	6	**16**	**84**	**27**	8	7	0	4	3

「高い」「大きい」「強い」との共起率の上昇した名詞、「多い」「強い」「大きい」との共起率が下降した名詞が特に多いことが分かる。なお、共起頻度(一定字数当たり)を見ても、「高い」との組合せで大きく上昇した名詞、「多い」との組合せで大きく下降した名詞が多い(服部(2012))。これらの形容詞では、意味用法の拡張/縮小を想定することができる。

次に、いくつかの形容詞について、それとの共起傾向の変化を観察していく。

３．１　「高い」との共起傾向

まず、「高い」との共起傾向が 3 期を通じて強い名詞をあげると、次のようになる。意味的に近いものを便宜的に'/'で分けて示す。価格に関する語[4]や、比率の類の語、「～度」の形の語など多いことが分かる。

(3)　3 期を通じて「高い」との共起傾向の強い名詞
一貫して 99%以上　物価　地価　値段 /　水準　レベル /　格調　精度
一貫して 95%以上　単価　価格　家賃　料金　コスト　運賃　金利 /　確度 /　能率　生産性
一貫して 90%以上　保険料　賃金 / 税率 /　緊急度
一貫して 80%以上　給料　利子　給与 /　価値　質　評価　地位 /　貯蓄率　補助率 /　効率　収益性

次に、「高い」との共起率が 3 期で大幅に上昇した名詞をあげる。具体的には、共起率差分の値が 30 ポイント以上の語である。

(4) 「高い」との共起率が 30 ポイント以上上昇した語　　23 語
必要性　緊急性　必要度　/ リスク 危険性 / 公共性　公益性　信頼性 /
可能性　確率　頻度　率　比率　パーセンテージ　割合　シェア　死亡率 /
ウェート / 所得　人件費　収入 / 能力 / 関心

---

[3] 共起率が３期にわたって単調に増加(Ⅰ期の値＜Ⅱ期の値＜Ⅲ期の値)しているものに限る。以下同じ。

[4] この場合の反義語は一般的には「安い」のように思われるが実際には「低い」の用例もある。

「～性」の形の語、比率を表わす語などが多いことが分かる。「～性」の形の語について、より詳しく、どの形容詞との共起率が増えた(減った)かを見ると、次のようになる。

表２　「～性」の各形容詞との共起率差分(ほとんど出現のない形容詞は略す)

	高い	大きい	多い	強い	濃い
可能性	**65.64**	1.87	**−52.36**	**−12.84**	−1.35
必要性	**63.09**	−6.17	**−21.14**	**−30.29**	0.00
危険性	**59.10**	7.06	**−55.01**	−6.14	0.00
公益性	**55.94**	−0.96	−1.60	**−50.41**	−0.32
公共性	**46.71**	−0.93	−1.11	**−43.01**	−0.05
信頼性	**39.53**	0.00	**−10.00**	**−29.53**	0.00
緊急性	**31.42**	0.00	0.00	**−31.42**	0.00

どの語でも「強い」との共起率が減少しているが、「可能性・必要性・危険性」では「多い」との共起率の減少も多い。その 3 語は、「～スル可能性」のように命題を表わす内容節をとりうるものである。これらについては 3．４で立ち戻る。

次に「比率」の類の名詞について同様のデータを示す。「多い」(と「大きい」)が減少し「高い」が増加する傾向が明瞭である。図１に「比率」の共起傾向推移のグラフを示す。

表３　「比率」等の各形容詞との共起率差分

	高い	大きい	多い	強い
比率	**38.32**	**−13.85**	**−23.16**	−0.33
率	**38.24**	−5.46	**−32.15**	−0.21
確率	**36.20**	−6.56	**−17.88**	−5.88
割合	**35.64**	**−12.11**	**−22.56**	0.00
頻度	**34.57**	−2.21	**−30.87**	−0.75
死亡率	**31.86**	0.00	**−31.86**	0.00
パーセンテージ	**30.21**	9.29	**−38.88**	−0.63

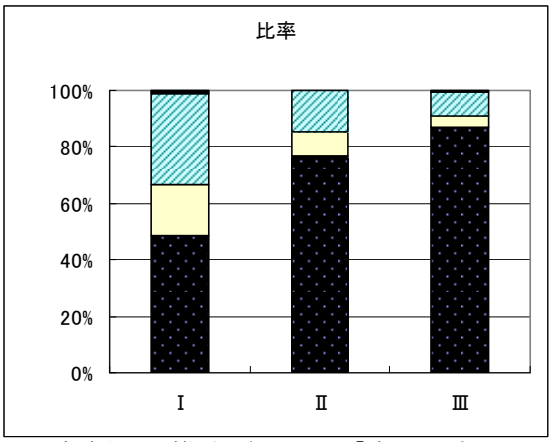


図１　「比率」の共起傾向の推移（下から「高い・大きい・多い・強い」）

３．２　「大きい」との共起傾向

一貫して「大きい」との共起率の高い程度的名詞は少ない。

(5)　3 期を通じて「大きい」との共起傾向の強い名詞
　一貫して 99%以上　(なし)
　一貫して 95%以上　規模
　一貫して 90%以上　(なし)
　一貫して 80%以上　幅　役割

「大きい」との共起率が大きく上昇したのは次の語である。基準との隔たりを表わす語、プラスまたはマイナスの評価を受ける対象の規模を表わす語などが多いようである。詳細は略するが、共起率の減少した形容詞は主として「多い」である[5]。

(6)　「大きい」との共起率が 20 ポイント以上上昇した語　　15 語
　変動 変化 / 開き 差 / 額 / 余地 / メリット 利益 /
　危険 弊害 困難 不安 / 負担 / 意味　意義

３．３　「強い」との共起傾向

一貫して「強い」との共起率の高い程度的名詞は少ない。

(7)　3 期を通じて「強い」との共起傾向の強い名詞
　一貫して 99%以上　(なし)
　一貫して 95%以上　風
　一貫して 90%以上　(なし)
　一貫して 80%以上　　力　意向　性格

「強い」との共起率差分が 20 ポイント以上である名詞は次の 5 語である。およそ、「感じられる」度合いを述べるものである。「感・感じ」の共起傾向変化を表 4 に詳しく示す。

(8)「強い」との共起率が 20 ポイント以上上昇した語　　5 語
　感・感じ・空気・疑い・不満

表４　「感」「感じ」の各形容詞との共起率差分

	高い	多い	強い	濃い	深い	濃厚
感	−0.27	0.73	**52.41**	−0.27	**−51.80**	−0.80
感じ	0.00	2.11	**21.12**	0.76	**−21.96**	−2.03

「感・感じ」では、「深い」との共起率が減少し、おおよそ、「強い」と入れ替わったことが分かる。図２に「感」の共起傾向推移のグラフを示す。

なお、「～{感・感じ}が深い」のような言い回しは、最近はあまり聞かれないように思われるので、いくつか実例をあげておく。

(9)　この食糧管理法の欠陥は、最近それを非常に露呈して参つた感が深いのであります。
(1947　1 参　本会議 10 号　岩木哲夫)

(10)　ともあれ、地方分権いまだしという感が深いわけでありますが、
(2001 151 衆 文科委 7 号 葉山峻)

[5] ただし、「意味」では「強い・深い」、「意義」では「深い」との共起率減少が顕著である。

(11) このような状況を見ますと、これはどうもこの廳はほんとうに日本の工業や商業を発展させて、独立國家としての生産を維持しようという点には、きわめて縁遠いような感じが深い。(1949　5 衆　内閣委 21 号　木村榮)

(12) この数年間振り返ってみますと、同じようなことを伺って同じようなことを答弁を求めて、何たることをしているんだろうかという感じが深いのがこの委員会でありまして、(1986 104 参　大蔵委 3 号　栗林卓司)

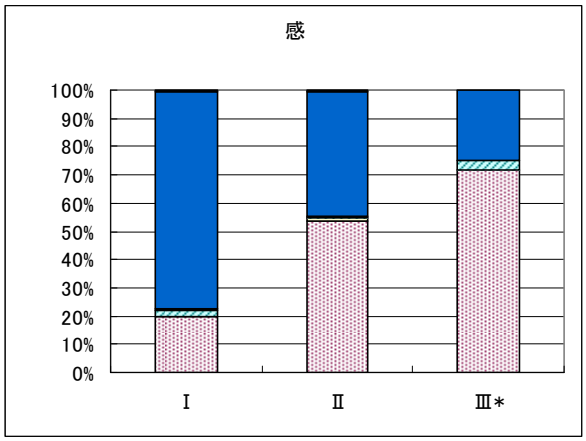


図２　「感」の共起傾向の推移（上から「深い・多い・強い」）

３．４　「多い」との共起傾向

一貫して「多い」との共起率の高い程度的名詞は次のようである。量、あるいは、可算的な数を値とするものなどである。

(13)　3 期を通じて「多い」との共起傾向の強い名詞
一貫して 99%以上　件数　種類　機会
一貫して 95%以上　人数　数　回数 /　交通量　雨量 / 苦労
一貫して 90%以上　人口 規模 /　問題点　欠陥　犯罪　雨
一貫して 80%以上　量 /　支出　輸入 /　課題　分野　議論　トラブル

次に、「多い」との共起率が大きく(30 ポイント以上)減少した名詞 20 語について、その形容詞別の共起率の増減を表 5 に示す。多くの名詞において、「多い」に替って「大きい」(および/または)「高い」の共起率が高まっている[6]ことが分かる。

名詞の種類を見ると、金額を表わすもの、比率を表わすものがある。他に、「危険」の類義語や「弊害・効果・可能性」、「変動・変化」なども含まれる。

6 「障害」は例外である。急激に共起率の伸びた「障害・重い」は、人の心身の障害（障碍）の程度を指す場合が大部分である一方、「障害・多い」はさまざまな種類の障害を表わしている。

表５ 各形容詞との共起率差分

	高い	大きい	多い	強い	濃い	重い	深い	大	濃厚
障害	−0.51	−1.72	**−69.87**	−3.39	0.00	**75.51**	0.00	0.00	0.00
リスク	**52.23**	14.74	**−67.04**	0.07	0.00	0.00	0.00	0.00	0.00
変動	0.00	**65.06**	**−61.24**	−3.83	0.00	0.00	0.00	0.00	0.00
余地	3.91	**54.23**	**−55.83**	−1.54	0.00	0.00	0.00	−0.77	0.00
危険	**26.32**	**30.89**	**−55.30**	1.92	−0.81	0.00	0.00	−2.89	−0.13
危険性	**59.10**	7.06	**−55.01**	−6.14	0.00	0.00	0.00	−1.01	−3.99
変化	1.25	**58.75**	**−52.86**	−7.14	0.00	0.00	0.00	0.00	0.00
可能性	**65.64**	1.87	**−52.36**	−12.84	−1.35	0.00	−0.08	0.54	−1.41
効果	**26.66**	**17.73**	**−40.68**	−1.08	0.00	0.00	0.00	−2.64	0.00
人件費	**37.12**	2.63	**−39.74**	0.00	0.00	0.00	0.00	0.00	0.00
パーセンテージ	**30.21**	9.29	**−38.88**	−0.63	0.00	0.00	0.00	0.00	0.00
利益	3.45	**35.81**	**−36.54**	−0.38	0.00	0.00	−0.76	−1.57	0.00
経費	**27.48**	8.10	**−34.39**	−0.60	0.00	0.00	0.00	−0.60	0.00
率	**38.24**	−5.46	**−32.15**	−0.21	0.00	−0.32	0.00	−0.11	0.00
収入	**30.67**	1.64	**−31.89**	0.00	0.00	0.00	0.00	−0.42	0.00
死亡率	**31.86**	0.00	**−31.86**	0.00	0.00	0.00	0.00	0.00	0.00
所得	**38.88**	−7.20	**−30.96**	0.00	0.00	−0.24	0.00	−0.48	0.00
頻度	**34.57**	−2.21	**−30.87**	−0.75	0.00	0.00	0.00	−0.75	0.00
弊害	0.00	**30.11**	**−30.67**	1.65	0.00	0.00	−0.22	−0.87	0.00
需要	**16.74**	10.64	**−30.55**	3.31	0.21	0.00	0.00	−0.35	0.00

最後のグループについて、少し考察してみたい。「多い」は、1 個・2 個--といった個数について用いられるほか、「かさ」や「目方」のような連続的な値について用いられることもある(久島(2002)に詳しい)。ところで、「危険が多い」は、個々の具体的危険の種類を数えあげるものと解釈しうるが、危険の規模あるいは量を総体的に捉えたものと解釈できる場合もあり、どちらとも決めがたいことがある。下に 2 例あげる。

(14) 難しい工事は河川管理者がかわってやってやる、こういうことをやっているとやはり乱開発も進むし、また災害の危険も多くなるということを言わざるを得ないのですね。(1989 114 衆 建設委 4 号 中島武敏)

(15) つまり所得税等において課税標準が非常につかまえられやすくなるというようなことを納税者が考えて脱税する危険が多い。(1948 2 衆 財政金融委公聴会 徳島米三郎)

「{弊害・効果・可能性}が多い」などについても上と同様の見方を適用することはできる。また、「{変動・変化}が多い」についても、変動の回数とも総体としての変動量ともとれる。

上のようにもともとあいまい性を内在する組合せにおいて「多い」を用いることが少なくなっているようである。また、「比率」のような二次的な量(直接計測できる量より一段抽象的である)においても「多い」の使用が少なくなってきている。

３．５　類義語の共起傾向変化パターン：「危険性」の仲間

次に観点を変えて、意味的(・形態的)に類似性のある複数の語での、形容詞との共起傾向変化パターンの比較を行ってみる。「危険・危険度・危険性・リスク」の４語をとりあげる。

図３～図６を観察すると、どの語でも「高い」との共起率が上昇し、「多い」との共起率が下降している、また、どの語も「大きい」との共起例がある程度は存在する。類義語間で一つの語から他の語へと変化が波及していることが考えられる。

しかし、他の共起形容詞の種類や各形容詞の共起率は語によって異なる。「リスク」以外は「強い」との共起例があり、最も共起形容詞の顔ぶれが多彩なのは「危険性」である。

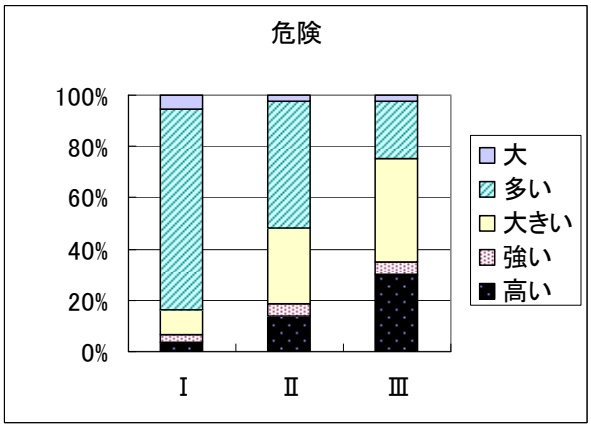


図３

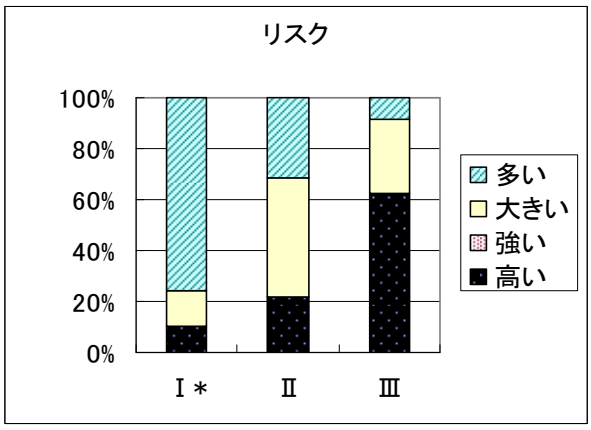


図４

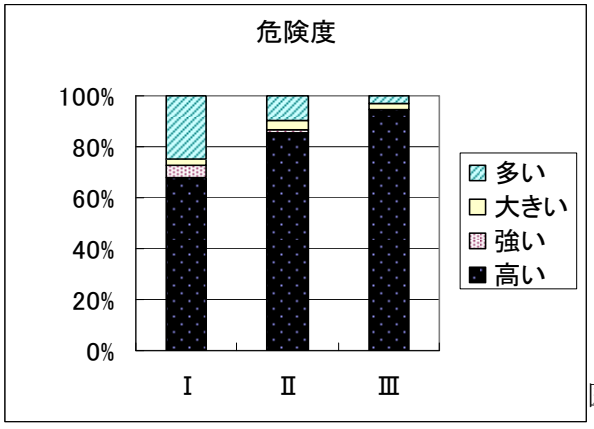


図５

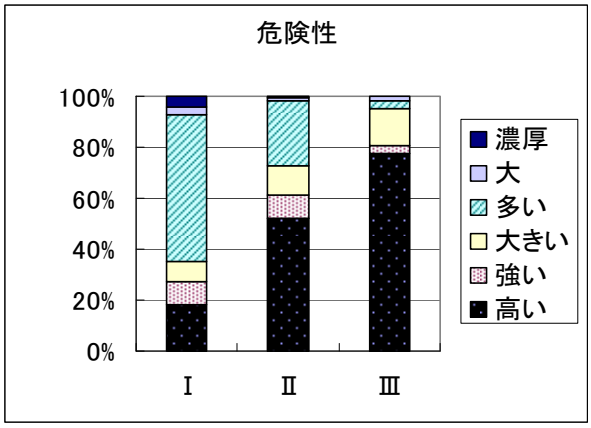


図６

一方では、服部(2011a)で示したように、「～度」の形の語は主に「高い」と共起するといった傾向も存在する(温度計上の点の高低のようなイメージが基本にあるのであろうか)。このように、さまざまな軸における共通性に着目することによって、各名詞の形容詞別共起率変動パターンの相違をある程度は説明しうると思われる。

４　まとめ

戦後 60 年間の国会会議録に記録された発言をデータとし､20 年ごとの 3 期に分けた上、程度性のある名詞と形容詞の共起傾向の推移を分析した。

特に、「高い」、「大きい」、「強い」のいずれかとの共起率が上昇している名詞が多い。各形容詞について、それとの共起率の上昇した名詞・一貫して共起率の高い名詞を眺めると、意味的な共通点のある語群が認められる。これを大局的に見れば、元々は共起形容詞の顔ぶれに関して多種多様であった諸名詞が意味的な類似性を軸として、主に単一の形容詞と共起する方向にまとめられていく変化とみなしうる。

一方、多くの名詞に対して共起率が顕著に減少した形容詞は「多い」である。これは、大局的には、一種の意味変化と考える余地がある。つまり、抽象的な量の大きさを表わす用法を縮小する方向への変化である。

付　記

本研究は、学術研究助成基金助成金(基盤研究(C)「有無・量的大小・増減・出現消滅の述語の総合的研究」、課題番号 23520479)、および、国立国語研究所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。

文　献

久島茂(2002)『《物》と《場所》の意味論 「大きい」とはどういうこと』 くろしお出版.

服部匡(2011a) 「程度的な側面を持つ名詞とそれを量る形容詞類との共起関係―通時的研究―」 『言語研究』140 号, pp.89-116.

服部匡(2011b) 「名詞と尺度的形容詞類の共起傾向の推移―国会会議録のデータから―」『同志社女子大学学術研究年報』62 号, pp.113-141

服部匡(2012) 「名詞と尺度的形容詞類の共起頻度の推移―国会会議録のデータから―」『同志社女子大学大学院文学研究科紀要』12 号, pp.1-11.

「少納言」「中納言」検索結果活用ツール

田野村忠温（大阪大学大学院文学研究科）

Corpus Tools for *Shonagon* and *Chunagon*

Tadaharu Tanomura (Osaka University)

１．はじめに

「現代日本語書き言葉均衡コーパス(BCCWJ)」検索サイト「少納言」「中納言」の検索結果を利用するための小ツールbccwj2excelとsortKWICを作成した。いずれも検索結果をソート（並べ替え）してExcelに収めるものである。日本語版Windows上で作動する。

bccwj2excelは、少納言または中納言で画面上に表示された検索結果を処理の対象とする。sortKWICは、中納言で［検索結果をダウンロード］のボタンを押して取得した検索結果およびその他一般の日本語のKWIC索引を対象とする。

両ツールはそれぞれ次のWebページで公開している。もっともこれらのURLをタイプするより、Googleなどで「bccwj excel」（または「bccwj エクセル」）「sortkwic」などと指定して検索したほうが手っ取り早い。

bccwj2excel: http://www.tanomura.com/research/bccwj2excel/
sortKWIC: http://www.tanomura.com/research/sortKWIC/

２ bccwj2excel

２.１ インストール方法

上記の bccwj2excel のページを開いて「bccwj2excel のインストール」のリンクをクリックし、表示される「ファイルのダウンロード」ダイアログで[実行]ボタンを押す（セキュリティの警告には「実行する」や「はい」で応じる）。続いて表示される「bccwj2excel のインストール」のダイアログで[OK]ボタンを押すと、デスクトップに次のような２つのアイコンが作られる。それぞれを通常版、フルデータ版と呼ぶ。

通常版は検索された用例に著者名と書名だけを添えて出力する。フルデータ版は少納言・中納言の提供するすべてのデータ項目を出力する。

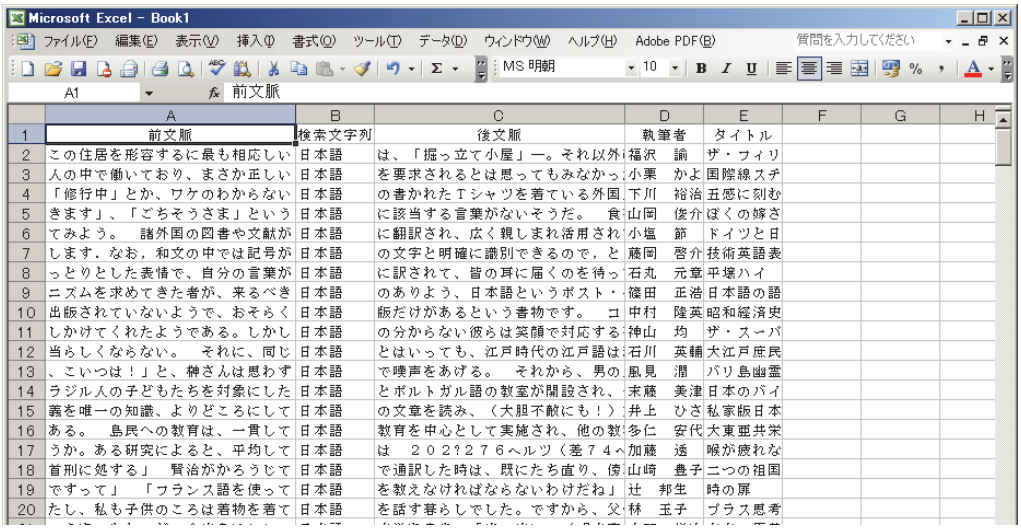
一方だけ使う場合は、使用時に迷わずにすむよう他方を消去するのが便利であろう。アンインストールするにはアイコンをごみ箱に移すだけでよい。

２.２ 用法

少納言または中納言で語句を検索して画面上に表示された検索結果をソートしてエクセルに格納するには次のようにする。

1) Internet Explorer を使って少納言または中納言のサイトで語句を検索
2) 検索結果の画面上で右クリックして「ソースの表示」を選び、表示されたソースを Ctrl+A で全選択して Ctrl+C でコピー
3) bccwj2excel のアイコンをダブルクリック

これによりエクセルの新しいブックが開かれ、各シートにソート済みの検索結果が入力される。必要に応じて列の幅を適宜調整して利用する。前後の文脈は最初各十数文字だけ表示されるが、幅を広げればより広い文脈を見ることができる。



	A	B	C	D	E	F	G	H
1	前文脈	検索文字列	後文脈	執筆者	タイトル			
2	この住居を形容するに最も相応しい	日本語	は、「掘っ立て小屋」一。それ以外	福沢　諭	ザ・フィリ			
3	人の中で働いており、まさか正しい	日本語	を要求されるとは思ってもみなかっ	小栗　かよ	国際線スチ			
4	「修行中」とか、ワケのわからない	日本語	の書かれたＴシャツを着ている外国	下川　裕治	五感に刻む			
5	きます」、「ごちそうさま」という	日本語	に該当する言葉がないそうだ。　食	山岡　俊介	ぼくの嫁さ			
6	てみよう。　諸外国の図書や文献が	日本語	に翻訳され、広く親しまれ活用され	小塩　節	ドイツと日			
7	します．なお，和文の中では記号が	日本語	の文字と明確に識別できるので，と	藤岡　啓介	技術英語表			
8	っとりとした表情で、自分の言葉が	日本語	に訳されて、皆の耳に届くのを待っ	石丸　元章	平壌ハイ			
9	ニズムを求めてきた者が、来るべき	日本語	のありよう、日本語というポスト・	篠田　正浩	日本語の語			
10	出版されていないようで、おそらく	日本語	版だけがあるという書物です。　コ	中村　隆英	昭和経済史			
11	しかけてくれたようである。しかし	日本語	の分からない彼らは笑顔で対応する	神山　均	ザ・スーパ			
12	当らしくならない。　それに、同じ	日本語	とはいっても、江戸時代の江戸語は	石川　英輔	大江戸庶民			
13	、こいつは！」と、神さんは思わず	日本語	で喚声をあげる。　それから、男の	風見　潤	バリ島幽霊			
14	ラジル人の子どもたちを対象にした	日本語	とポルトガル語の教室が開設され、	末藤　美津	日本のバイ			
15	義を唯一の知識、よりどころにして	日本語	の文章を読み、〈大胆不敵にも！〉	井上　ひさ	私家版日本			
16	ある。　島民への教育は、一貫して	日本語	教育を中心として実施され、他の教	多仁　安代	大東亜共栄			
17	うか。ある研究によると、平均して	日本語	は　２０２１２７６ヘルツ（差７４ヘ	加藤　透	喉が疲れな			
18	首刑に処する」　賢治がかろうじて	日本語	で通訳した時は、既にたち直り、傍	山崎　豊子	二つの祖国			
19	ですって」　「フランス語を使って	日本語	を教えなければならないわけだね」	辻　邦生	時の扉			
20	たし、私も子供のころは着物を着て	日本語	を話す暮らしでした。ですから、父	林　玉子	プラス思考			

検索結果は 3 つのモードでソートされ、各シートに収められる。

・モード１：　先行文脈に基づくソート（上図）
・モード２：　検索文字列＋後続文脈に基づくソート
・モード３：　後続文脈に基づくソート

検索文字列が単一の文字列の場合はモード 2 とモード 3 のソートの結果は同一になるので、モード 3 のシートは作成されない。
その他、細かい点を補足すれば以下の通りである。

- 先行文脈は自動的にセルの中で右寄せに配置されるので、列幅の大小にかかわらず途切れなく読むことができる。
- 通常版では書名に副題と巻号を添えて出力する。
- bccwj2excel は処理結果のディスクへの保存は行わない。必要に応じて手動で保存する。
- 実行終了時に表示されるポップアップメッセージは数秒後に自動的に消える。
- コピー後のソースは不要なので、開かれたメモ帳・エディタは閉じる。
- Internet Explorer 以外のブラウザには対応していない（bccwj2excel は動作するがデータを正しく変換できない）。

３　sortKWIC

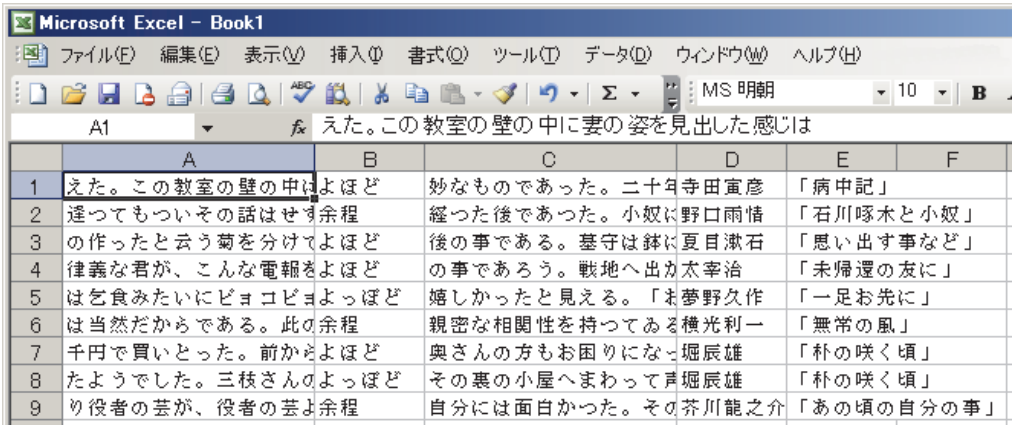
３.１　インストール方法

sortKWIC のインストール方法は 2.1 で説明した bccwj2excel の場合と共通である。sortKWIC のページを開いてインストールする。インストールが終わると、デスクトップに次のような２つのアイコンが作られる。それぞれを通常版、フルデータ版と呼ぶ。

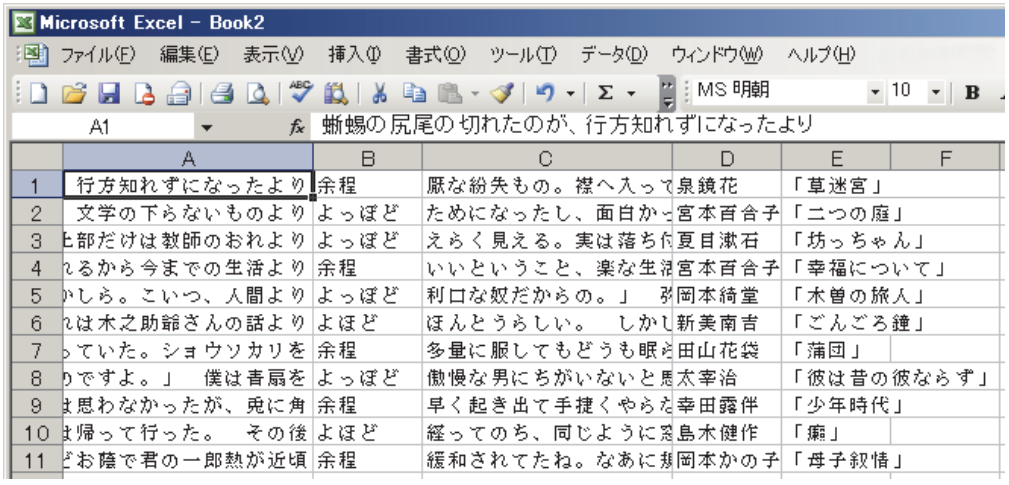
通常版は、中納言の検索結果を処理するとき、用例に著者名と書名だけを添えて出力する。フルデータ版は中納言の提供するすべてのデータ項目を出力する。その他の KWIC 索引を処理するときはどちらを使っても処理結果は同じである。

３.２　用法１──エクセルから

エクセルに次のように KWIC 索引が入っているものとする。



Ctrl+A で全選択して Ctrl+C でコピーし、sortKWIC のアイコンをダブルクリックする。これによりエクセルの新しいブックが開かれ、各シートにソート済みの検索結果が入力される。必要に応じて列の幅を適宜調整して利用する。



bccwj2excel の場合と同じく、検索結果は 3 つ（ないし 2 つ）のモードでソートされ、各シートに収められる。

その他、細かい点を補足すれば以下の通りである。

- 中納言で取得した検索結果を通常版で処理するときは、書名に副題と巻号を添えて出力する。
- sortKWIC は処理結果のディスクへの保存は行わない。必要に応じて手動で保存する。
- 中納言でダウンロードした 3 万件以上の KWIC 索引を処理できることを確認している。ただし、件数の上限はデータや環境に依存する。
- 上の例では先行文脈、検索文字列、後続文脈が A～C 列に入っているが、連続する 3 列ならばどこでもかまわない。

３.３　用法２──タブ区切り形式データ

sortKWIC でソートする KWIC 索引はタブ区切り形式データでありさえすればよく、エクセルに入っている必要はない。

例えば、中納言の［検索結果をダウンロード］を使って検索結果を取得してエクセルに格納するには次のようにする。

1) ［検索結果をダウンロード］ボタンを押し、「ファイルのダウンロード」ダイアログで［開く］を選ぶ
2) メモ帳などで開かれた検索結果全体を Ctrl+A、Ctrl+C によってコピー
3) sortKWIC のアイコンをダブルクリック

コピー後の検索結果は不要なので、開かれたメモ帳類は閉じる。

青空文庫所収の文学作品 3,410 件から語句を用例を検索する「日本語用例検索サイト」（http://www.tokuteicorpus.jp/team/jpling/kwic/）での検索の場合は、「検索結果をダウンロード」にチェックを入れてから［検索］ボタンを押す。あとの手順は上の 1)の後半以下と同じである。

付記

・両ソフトウェアの内容は無保証です。ご自身の責任においてご利用ください。
・両ソフトウェアは日本語版 Windows 上で動作します。動作確認は Windows XP＋Internet Explorer 6＋Excel 2003 で行っています。新しい環境でもおそらく動くと思いますが未確認です。
・両ソフトウェアの作成には Ruby 1.8.7（http://www.ruby-lang.org/）と Exerb 5.3.0（http://exerb.sourceforge.jp/）を使用させていただいています。

統計的機械学習による歴史的資料への濁点の自動付与

岡 照晃 （奈良先端科学技術大学院大学）[†1]

A Machine Learning Approach to Automatic Labeling of Voiced Consonant Mark for Historical Text

Teruaki Oka （Nara Institute of Science and Technology）

1 はじめに

近年，コーパスを利用した日本語研究が増えつつある．

しかし，日本語学や国語学の分野では，古い時代の資料を扱う歴史的研究が現在も大きな位置を占めている．だが，それらの分野で扱われるような歴史的資料は，コーパスとしての整備が現代語のコーパスと比べて進んでいないのが現状である．

歴史的コーパスの整備が進まない原因の一つとして，コーパス整備の際の**校訂**の，作業コストが高いことが挙げられる．校訂作業は専門家にしか行えず，作業人員を大量に集めることが難しい．またその反面，作業対象は膨大であるため，作業を完了するまでに非常に時間がかかる．

そこで本研究では，統計的機械学習手法を用い，歴史的資料の校訂作業を自動化することを最終的な目的とする．これにより，誰でも簡単に低コストかつ大規模に校訂作業を実施することが可能になると考えられる．そしてその第 1 段階として，校訂作業の中から**濁点付与**を取り上げ，自動化に取り組んだ．

2 校訂作業における濁点付与

歴史的資料の記述中にはよく，図 1 のような「濁点を付けて書いてあることが期待されるのに，濁点の付いていない文字」が含まれている．本論文ではこういった文字のことを**濁点無表記文字**と呼ぶ．濁点無表記文字は，コーパスユーザの可読性・検索性を損なわせる原因の一つである．そこで，濁点無表記文字を濁点付きの文字（**濁点文字**）に置き換える校訂作業が行われる．これが濁点付与である．

表 1 に，未校訂の資料中に存在する濁点文字（e.g. が，だ，ば，･･･ ）と濁点を付けることが可能な文字（e.g. か，た，は，･･･ ）の統計データを示した．この表を見ると，総文字数 8,423 文字に対して，その内約 19%の 1,641 文字は濁点を付けることが可能な文字（濁点無表記になっているかもしれない文字）である．人手で作業する場合は，これらすべてに対して濁点無表記か否かを網羅的に確認しなくてはならない．また，濁点が付く可能性のある文字の内，約 22%（368/1,641）が濁点無表記の濁音文字であり，例えば,「ガ」と発音する文字も「か」と表記されている．濁点付与では，見つけた濁点無表記文字の一つ一つに濁点を施していかないといけない．しかし，濁点無表記文字の数が多いだけに，非常に手間のかかる作業である．

[†1] teruaki-o@is.naist.jp

今や廣島は其名大に内外國に顯はれ苟も時事を談するものは同地の形勢如何を知らんと欲せさるはあらす是れ征清の大師一たひ海に航せしより　大元帥陛下大纛を此に駐め大本營となし軍務を親裁し玉ふに因てなり先つ其大勢より叙述して次第に細事に及はんとす

（『太陽』 1925 年 2 号 p.64 より抜粋）

図 1: 濁点無表記の例．このテキストは太陽コーパスの原文から抜きだしたものである．下線を引いた文字には，通常なら濁点が期待されるが，ここでは付けられていない（濁点無表記になっている）．

表 1: 濁点と濁音の統計．近代の雑誌**明六雑誌**第 1 号（2011 年 12 月段階のデータ，総文字数:8,423）中に含まれる濁点文字と濁点をつけることが可能な文字の中から，実際の発音が清音の文字数と濁音の文字数の内訳を調査した．表中の濁点の付き得る文字の内，実際の発音が濁音である文字が濁点無表記文字である．

表記＼発音	濁音	清音	計
濁点文字	78	0	78
濁点をつけることが可能な文字	368（濁点無表記文字）	1,273	1,641
計	446	1,273	

実際，現在コーパス化が進められている近代の雑誌**国民之友**の校訂作業では，1 人の作業者が 23 ページ分 (約 1 万 6,000 文字) に濁点付与するのに，1 日を要したと報告されている[1]．これに対し，例えば，既存の校訂済みコーパスである**太陽コーパス**（国立国語研究所, 2005）は，総文字数約 1,450 万文字の規模である．

よって，濁点無表記のアノテーションを自動化するだけでも，校訂作業の効率向上が大いに期待できる．

3　関連研究: 形態素解析辞書を用いた手法

近代文語論説文（明治普通文）には濁点無表記が多くみられる．そのため，近代文語論説文を対象とした形態素解析辞書である**近代文語 UniDic**[2]（小木曽ら, 2008）には，無濁点の見出し語も少数だが登録されている（e.g.,「す（文語助動詞-ズ，終止形）」）．

形態素解析辞書に無濁点の見出し語を追加することで，濁点無表記を含んだ文の解析を行うことが可能になる．またそれだけでなく，解析結果から濁点付きの表記を得ることもできる（**辞書ベースの手法**）．

しかしながら，辞書ベースの手法を行うためには，校訂対象となる歴史的資料用の形態素解析辞書が必要となる．しかし，現在使用可能なものの中で実用に足るものは，近代文語 UniDic の他に中古和文を対象にした**中古和文 UniDic**（小木曽ら, 2010）のみである．そのため，中世や近世の資料にこの手法を適用することはできない．また，辞書ベースの手法では，濁点付与の性能が形態素解析の精度に依存する．そのため，濁点付与の性能を上げるた

[1] 1 日の作業時間を 5〜6 時間とした場合．
[2] バージョン 1.1

か，き，く，け，こ，　さ，し，す，せ，そ，　た，ち，つ，て，と， は，ひ，ふ，へ，ほ，　ゝ，く（くの字点）

図 2: 校訂対象文字．ここで示した 22 文字が濁点付与の対象となる文字である．

めには，学習用の形態素解析済みコーパスを新しく整備するか[3]，辞書の見出し語を増やす必要がある．しかし，これは一般にコストが高い．

加えて，近代文語 UniDic と中古和文 UniDic はいずれも，基本的に校訂済みの文を解析するために整備されている．そのため，無濁点の見出し語を追加したとしても，未校訂資料の形態素解析に利用した場合，解析結果の精度は決して高くないと考えられる．

そこで本論文では，濁点の自動付与のタスクを文字単位のクラス分類問題として定式化した．具体的には，未校訂の資料中に存在する「濁点の付く可能性のある文字（**校訂対象文字**）」を濁点文字に置き換えるべきか否か分類する問題を扱う．提案手法では，分類は点予測によって行う．点予測とは，周囲の文字に対する分類結果を参照せずに，当該分類を行う手法である．周囲の分類結果を見ないため，分類誤りが伝播することがない．そのため，表記の整っていない未校訂の資料に対しても頑健に動作することができる．また，分類の**素性**[4]にも，分類対象文字の周辺文字列の表層的な情報のみを使用し，周囲の単語境界の情報や，品詞の情報は使用しない．そのため，提案手法では，学習用コーパスとして形態素解析済みコーパスを必要とせず，形態素解析辞書も必要としない．また，形態素解析の精度に濁点付与の性能が左右されることもない．

4　提案手法: 点予測による濁点の自動付与

提案手法では，未校訂資料の中に存在する校訂対象文字に対して，それぞれ独立に「濁点を付けるべきか否か」の分類を実施する．ただし，本論文では，校訂対象文字として平仮名と踊字である**くの字点**（く，ぐ）だけを扱い，片仮名は扱わないこととした[5]．そのため，濁点付与の対象となる文字は図 2 に挙げた 22 種類である．

以下に提案手法の概略を述べる．提案手法の詳しい説明については,（岡ら, 2011）もしくは（Oka et al., 2011）を参照．

4.1　分類に使用する素性

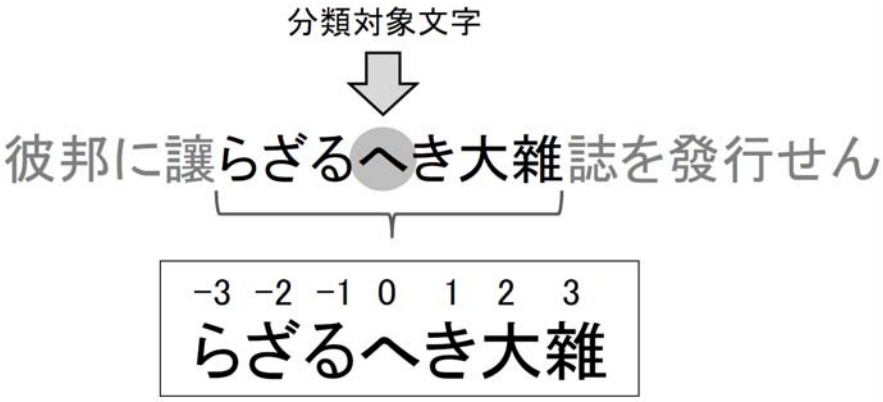
分類時の素性には，当該の分類対象文字とその周囲の文字列の表層的な情報のみを使用する．具体的には図 3 のように，分類対象文字とその左右の 3 文字の範囲内にある文字 n-gram を素性とする．文字 n-gram は，1〜3-gram までを使用する．また，各 n-gram には出現位置（分類対象文字からの相対位置）を添え字として設けている．提案手法で使用する素性は，「その n-gram がその位置に現れたか（1）否か（0）」を表す 2 値素性である．

歴史的資料は，表 1 で示しているように，完全に無濁点になっているとは限らない．所々には濁点が付いた状態の資料もある．濁点の使い方は書き手によって一定でない．そのため，あらゆる濁点の表記状態に対応するためには，こういった濁点を分類時には外しておくべき

[3]各単語のコスト（出現しやすさ）を計算するために使用される．

[4]分類の手がかりとなる情報．

[5]漢字片仮名交じり文を除いて，基本的に片仮名は外来語や固有名詞等の限られた語の表記にしか用いられていない．そのため，本論文では片仮名は対象外とした．



位置-3 の文字 1-gram = ら	位置-3 の文字 2-gram = らざ	位置-3 の文字 3-gram = らざる
位置-2 の文字 1-gram = ざ	(位置-3 の文字 2-gram = らさ)	(位置-3 の文字 3-gram = らさる)
(位置-2 の文字 1-gram = さ)	位置-2 の文字 2-gram = ざる	位置-2 の文字 3-gram = ざるへ
位置-1 の文字 1-gram = る	(位置-2 の文字 2-gram = さる)	(位置-2 の文字 3-gram = さるへ)
位置 0 の文字 1-gram = へ	位置-1 の文字 2-gram = るへ	位置-1 の文字 3-gram = るへき
位置 1 の文字 1-gram = き	位置 0 の文字 2-gram = へき	位置 0 の文字 3-gram = へき大
位置 2 の文字 1-gram = 大	位置 1 の文字 2-gram = き大	位置 1 の文字 3-gram = き大雑
位置 3 の文字 1-gram = 雑	位置 2 の文字 2-gram = 大雑	

図 3: 提案手法で使用する素性．ただし，ここでは値が 1 となる素性のみを示している．

である．しかしながら，予め施されていた濁点は分類の際の証拠として有効な場合もあると考えられる．そこで，分類時には，文字 n-gram 内に含まれる濁点文字の一部〜全てより濁点を外したものも素性として参照することにした．この素性は，図 3 において，括弧付きで示してある．

また，歴史的資料の中では句読点の使い方が一貫していない．句点を読点のように使用する場合もあれば，読点を使って文末を表現する場合もある．そこで提案手法では，資料中の句読点「、。」を全て特殊記号 ⟨PUNC⟩ に置き換えることにした．

4.2 学習用事例の作成手順

学習用の事例は，学習用コーパス中にある学習用文字（図 4 参照）から作成する．学習用コーパスには校訂済みの歴史的資料を使用する．学習用事例作成の手順は以下の通りである．

1. 学習用コーパスから学習用文字を 1 つ取り出す（e.g.,「が」）．
2. 取り出した文字とその左右 3 文字を合わせて 1 つの事例とみなす．この際，取り出した学習用文字が濁点文字であれば，濁点を外しておく（「が」→「か」）．
3. 取り出した学習用文字（「が」）を正解のクラスとする．

4.3 分類器

提案手法では，太陽コーパスのような大規模なコーパスからでも高速かつ高精度に分類器の学習を行うため，分類器として線形パーセプトロンを使用する．実際には，多クラスの

か，き，く，け，こ，	さ，し，す，せ，そ，	た，ち，つ，て，と，
は，ひ，ふ，へ，ほ，	ゝ，く（くの字点）	
が，ぎ，ぐ，げ，ご，	ざ，じ，ず，ぜ，ぞ，	だ，ぢ，づ，で，ど，
ば，び，ぶ，べ，ぼ，	ゞ，ぐ（くの字点）	

図 4: 学習用文字一覧.

PassiveAggressive-I（Crammer et al., 2006）を採用した.

提案手法では，各校訂対象文字ごとに「濁点をつけた文字」か「濁点を付けないまま文字」のいずれかのクラスに分類する規則（**モデル**）を作成する．例えば，校訂対象文字「か」に対して,「か」と「が」のいずれかのクラスに分類するモデルを作成する．そしてそれとは別に,「き」に対して,「き」か「ぎ」か分類するモデルを作成する.

5 濁点付与の性能評価実験

提案手法の有効性を検証するために，濁点付与の性能評価実験を行なった．今回は未校訂の近代文語論説文を対象とし，濁点付与の適合率と再現率を調べた.

5.1 実験に使用したコーパス

本実験では，学習用コーパスとして，以下の校訂済みのコーパスを使用する.

- UniDicMLJ-TRAIN:
  近代文語 UniDic のコスト算出に用いられたコーパス．形態素解析済みコーパスであり，校訂も実施済みである．ただし，量が少なく，原文の情報もコーパス中には保持されていない.

- SUN-TRAIN:
  近代語の大規模コーパスである太陽コーパス[6]から 9 割を学習用コーパスとして利用する．実際には，太陽コーパスの 1895 年 5 号，1901 年 5 号，1909 年 5 号，1917 年 5 号，1925 年 5 号を評価用コーパスとして別に分け (SUN-TEST)，残りを学習に利用することにした (SUN-TRAIN)．太陽コーパスは構造化テキストタグ付きコーパスであるため，校訂は行われているが，形態素解析までは行われていない.

また，評価には以下のコーパスを利用する.

- SUN-TEST:
  太陽コーパスの 1895 年 5 号，1901 年 5 号，1909 年 5 号，1917 年 5 号，1925 年 5 号を評価用コーパスとして利用する.

- NF-TEST:
  現在コーパス化の作業が進められている明治期の雑誌，国民之友も評価用コーパスとして利用する．ここでは，2011 年 3 月の段階で濁点付与が実施されていた 1887 年 10 号，1888 年 20 号，1888 年 30 号，1888 年 36 号を使用した.

[6]2011 年 1 月段階のデータ

表 2: コーパス内の文数と段落数の内訳.

	文数	段落数	文字総数
UniDicMLJ-TRAIN	20,330	-	604,966
SUN-TRAIN	-	70,084	6,380,398
SUN-TEST	-	6,316	619,357
NF-TEST	-	868	172,780
M6-TEST	-	1,450	252,232

表 3: 学習用事例の内訳.

事例のクラス / 学習用コーパス	濁音文字	清音文字	合計
UniDicMLJ-TRAIN	26,123	110,974	137,097
SUN-TRAIN	208,099	962,580	1,170,679

・ M6-TEST:

Oka et al.(2011）では，評価用コーパスとして上記 2 つのコーパスのみを利用している．本論文ではさらに，太陽や国民之友と同じ，明治期の雑誌である**明六雑誌**（全 43 号）[7]も評価用コーパスとして利用する．ただし，明六雑誌はほとんどの記事が漢字片仮名交じり文で記述されている．そのため，ここでは，全ての片仮名文字を平仮名文字に直して用いることにした．

評価に使用する太陽コーパス・国民之友・明六雑誌（3 雑誌コーパス）はいずれも校訂済みのコーパスである．しかし，タグを使って原文が保持されている．そこで，実験を実際のタスクに近づけるため，評価にはタグから再現した原文を使用する．

また，明治期において，句読点の使い方はまだ確定していなかった．そのため，明確な文境界を定めることは難しい．今回使用する 3 雑誌コーパスでも，文境界の明確なアノテーションは行われていない．そこで今回，3 雑誌コーパスは学習・評価の両方において，段落単位で使用することにした．実際のタスクでも文境界が定められていないことが多いため，これは，より実際のタスクに近い設定といえる．ただし，UniDicMLJ-TRAIN は近代文語 UniDic の学習に使用されたコーパスであるため，文境界は明確にされている．そこで，UniDicMLJ-TRAIN のみ，文単位で使用することにした．事例抽出の際には，文（or 段落）の頭と末尾に，それぞれ文頭，文末を表す特殊記号 ⟨BOS⟩ と ⟨EOS⟩ を設ける．

太陽コーパス・国民之友・明六雑誌には口語で書かれた記事も含まれている．この実験では，文語を扱う．そのため，口語の記事や引用は学習用コーパス・評価用コーパスのいずれからも全て除外した．

口語文を除いた各コーパスの文数と段落数，総文字数の内訳を表 2 に示す．また，学習用事例と評価用事例の内訳を表 3 と表 4 に示す.

[7]2011 年 12 月段階のデータ

表 4: 評価用事例の内訳.

評価用コーパス ＼ 事例のクラス	濁音文字	清音文字	合計
SUN-TEST	899	92,803	93,702
NF-TEST	3,842	25,418	29,260
M6-TEST	6,219	39,314	45,533

5.2 比較手法: 辞書ベースの手法

辞書ベースの手法を比較手法として設定した．この手法では，以下の手順で近代文語 UniDic を拡張し，拡張した辞書を用いた形態素解析の結果から濁点付与を行う．

1. 近代文語 UniDic の全ての見出し語から濁点を全て外す．ただし，各語のフィールド中には濁点を外す前の表記を保存しておく．

2. 無濁点にした UniDicMLJ-TRAIN を用いて，1 で作成した辞書の各語のコストを計算する．

3. 各語のフィールド中に残しておいた濁点付きの表記から，濁点の一部〜すべてが補われた見出し語を復元し，辞書に追加する．ただし，この時追加する語のコストは，無濁点の場合のコストと同一とする．

この辞書を使えば，濁点が所々抜け落ちたような文でも形態素解析を行うことができる．また，各語のフィールド中には濁点付きの表記が保存されているため，形態素解析の結果から濁点付きの表記を復元できる．

ただし，辞書ベースの手法では，形態素解析済みの近代語のコーパス（現時点で UniDicMLJ-TRAIN のみ）からでしか学習が行えないという欠点がある．また，評価用コーパスはいずれも文のアノテーションが明確に行われていないため，形態素解析は段落単位で行う．

形態素解析器には MeCab[8] を使用する．

5.3 濁点付与性能評価実験

各手法を用いてそれぞれの評価用コーパスに濁点付与を行い，評価を行なった．ここでは濁点付与の適合率，再現率と，その 2 つの調和平均である F 値で評価した．各値の計算方法は以下の通り．

$$\text{適合率} = \frac{\text{正しく濁点を付けた文字数}}{\text{濁点を自動付与した文字数}} \times 100[\%] \tag{1}$$

$$\text{再現率} = \frac{\text{正しく濁点を付けた文字数}}{\text{評価用コーパス中の濁点無表記文字数}} \times 100[\%] \tag{2}$$

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{3}$$

結果を表 5 に示す．

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

表 5: 濁点付与の性能評価.

評価用コーパス	手法	学習用コーパス					
		UniDicMLJ-TRAIN			UniDicMLJ-TRAIN + SUN-TRAIN		
		適合率.[%]	再現率.[%]	F 値	適合率.[%]	再現率.[%]	F 値
SUN-TEST	辞書ベース	50.9	91.8	65.5	-	-	-
	提案手法	54.7	85.2	66.6	71.2	97.0	82.1
NF-TEST	辞書ベース	93.3	96.5	94.9	-	-	-
	提案手法	95.1	94.5	94.8	96.0	98.3	97.1
M6-TEST	辞書ベース	90.1	95.9	92.9	-	-	-
	提案手法	93.4	92.4	92.9	94.7	98.1	96.4

学習用コーパスを UniDicMLJ-TRAIN でそろえた場合，提案手法は，辞書ベースの手法に比べて低い再現率を示している．しかし，適合率は辞書ベースの手法よりも高いため，F 値においてはほとんど同じ性能が得られた．

また，提案手法は辞書ベースの手法に比べて低コストで学習用コーパスを追加できるという利点がある．学習用コーパスとして SUN-TRAIN を追加したとき，提案手法は適合率，再現率，F 値のすべてにおいて，辞書ベースの手法よりも高い性能を示した．このように，比較的簡単に性能を上げられるという点において，提案手法の優位性が確認できた．

5.4 エラー分析

提案手法のエラー分析を行った結果，以下に挙げた語と語の間で，濁点付与に失敗する傾向がみられた．

- 格助詞「が」・接続助詞「が」と，終助詞「か」・並列助詞「か」
- 打消しの助動詞「ず」と，サ変動詞「す」
- サ変動詞と，ザ変動詞
- 接続助詞「ば」と，係助詞の「は」
- 当時，語形上揺れがあった語（e.g.,「願わくば」と「願わくは」）
- 濁点を付けても付けなくてもどちらでもよさそうな語（e.g.,「結び」と,「結ひ」,「出て」と,「出で」）

また，提案手法と辞書ベースの手法のエラーを比較したとき，濁点付与誤りの傾向に大きな差は見られなかった．ただし，辞書ベースの手法特有のエラーとして，以下の語間での間違いが多くみられた．

- 接続助詞の「て」と接続助詞の「で」
- 接続助詞の「とも」と接続助詞「ども」

提案手法では，単語境界の情報を分類に使用しない．そのため，以下のようなエラーが数例見つかった（濁点付与に失敗している文字を太字にして示している）.

猶ほ兒玉氏の力強**が**りとや思ひけん

表 6: 形態素解析性能の改善度の比較（SUN-TEST）.

	単語分割			品詞認定			語彙素認定		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
原文	99.866	99.965	99.916	99.596	99.498	99.547	99.575	99.478	99.527
辞書ベース	99.886	99.842	99.864	99.300	99.342	99.321	98.815	98.858	98.837
提案手法で前処理	**99.943**	**99.972**	**99.957**	**99.727**	**99.698**	**99.713**	**99.715**	**99.686**	**99.700**

表 7: 形態素解析性能の改善度の比較（NF-TEST）.

	単語分割			品詞認定			語彙素認定		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
原文	98.607	99.406	99.005	95.878	95.112	95.494	95.696	94.932	95.312
辞書ベース	99.860	99.662	99.761	95.859	96.048	95.953	95.225	95.413	95.319
提案手法で前処理	**99.935**	**99.904**	**99.919**	**99.490**	**99.520**	**99.505**	**99.465**	**99.495**	**99.480**

ただし，同じようなエラーは辞書ベースの手法でも生じている．以下の例のように，辞書ベースの手法でも単語分割に失敗し，その結果，濁点付与を失敗することがある．

人 | の | 氣附:名詞（キヅケ） | が:助詞 | さる:連体詞 | 所ろ:名詞 | 、

5.5 形態素解析精度の改善度の比較

近代文語 UniDic は本来，校訂済みの文を解析するために整備されている．そのため，未校訂の資料の形態素解析に利用した場合，結果の精度は高くない．そこで，提案手法を形態素解析の前処理に用いることで，形態素解析の性能がどれほど改善できるか調査した．

ただし実際には，評価用コーパスの校訂済み本文を近代文語 UniDic を用いて形態素解析し，正解データとしている．そして，原文に濁点付与を行うことで，どこまでその結果に近づけるかを評価した．

提案手法は，UnidicMLJ-TRAIN + SUN-TRAIN で学習を行なったモデルを使用する．また，形態素解析器には MeCab を利用した．

評価は，単語分割，品詞認定，語彙素認定の 3 段階で行う．それぞれの段階における適合率・再現率・F 値を調査した[9]．

また，前処理に提案手法を使用せず，辞書ベースの手法で，濁点付与と形態素解析を同時に実施した場合と比較を行なった．ただし，近代文語 UniDic にはもともと少数であるが，無濁点の見出し語が含まれている．2 つの手法を公平に比較するため，提案手法で前処理された資料を解析する近代文語 UniDic からは，無濁点の見出し語を取り除いている．

結果を表 6〜8 に示す．この結果を見ると，提案手法を用いて濁点付与を行うことで，校訂済みテキストを解析するのとほぼ同等の性能を実現することが可能だと分かった．

[9]品詞認定，語彙素認定の性能は MeCab の評価用スクリプト mecab-system-eval を利用して求めた．

表 8: 形態素解析性能の改善度の比較（M6-TEST）.

	単語分割			品詞認定			語彙素認定		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
原文	98.442	99.523	99.980	95.679	94.648	95.161	95.392	94.365	94.876
辞書ベース	99.857	99.788	99.822	95.762	95.828	95.795	94.997	95.063	95.030
提案手法で前処理	**99.898**	**99.930**	**99.914**	**99.398**	**99.367**	**99.382**	**99.353**	**99.322**	**99.338**

6 おわりに

本論文では，点予測を用い，文字単位で濁点の自動付与を行う手法を提案した．太陽コーパスで学習を行い，近代語の資料に対して濁点付与を行なった結果，国民之友，明六雑誌に対しては適合率，再現率共に約 95%以上の性能で濁点付与が行えた．また，提案手法を前処理に用いることで，未校訂の資料でも校訂済みの資料と同程度の性能で形態素解析が行えるようになることが分かった．

謝辞

本研究は，国立国語研究所の共同研究プロジェクト「統計と機械学習による日本語史研究」による研究成果の一部である．

文献

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer (2006) 「Online Passive-Aggressive Algorithms」 *Journal of Machine Learning Research*, 7 pp. 551-585.

[2] Teruaki Oka, Mamoru Komachi, Toshinobu Ogiso, Yuji Matsumoto (2011) 「Automatic Labeling of Voiced Consonants for Morphological Analysis of Modern Japanese Literature」 In *Proceedings of the 5th International Joint Conference of Natural Language Processing* (IJCNLP 2011), pp. 292-300.

[3] 小木曽智信, 小椋秀樹, 近藤明日子 (2008) 「近代文語文を対象とした形態素解析辞書の開発」 言語処理学会第 14 回年次大会発表論文集, pp. 225-228.

[4] 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, 2010-CH-85:4.

[5] 岡照晃, 小町守, 小木曽智信, 松本裕治 (2011) 「機械学習による近代文語文への濁点の自動付与」 情報処理学会研究報告 自然言語処理研究会報告, 2011-NL-201:6, pp. 1-8.

[6] 国立国語研究所編 (2005) 『太陽コーパス』 国立国語研究所資料集 15, 博文館新社.

古代日本語の主節の無助詞名詞句—活格性との関わりから—

竹内史郎（成城大学文芸学部）
siberius@seijo.ac.jp

Bare Noun Phrases of Main Clauses in Old Japanese

Shiro Takeuchi (Seijo University)

1　はじめに

通時コーパスの作成においては、無助詞名詞句の扱いやその統語変化が重要な問題となる (近藤 (2011))。特に名詞句のマークアップにおける統語的な情報の付与については、現代語と異なる部分が大きいことが想定され[1] 、この方面での記述的な研究の進展が望まれるが、古典語における無助詞名詞句やその歴史に関し、答えることのできない基本的な疑問が少なくない。例えば主節の主語や目的語の振る舞いには、歴史上で異なった段階がいくつあり、それぞれの段階の特性や、それぞれの段階をつなぐ変化はどのように記述できるのだろうか。

本稿では、平安時代 (10 世紀) 和文に見える主節を調査の範囲とし、主語や目的語として現れた無助詞名詞句の振る舞いについて考察を加える。この結論として、古代日本語の主節では、〈動作主〉主語と〈対象〉主語の振る舞いが異なり、〈対象〉主語はむしろ他動詞文の目的語と同様に振る舞うこと、古代日本語の無助詞名詞句は意味役割によってその振る舞いが決定されていること等を述べる。さらには、無助詞目的語とヲ格目的語の統語上の振る舞いの違いについても明らかにする。本稿の考察の結果が、上に述べた疑問の解決のための足がかりになればと考える。

2　調査の方針、前提など

今回の調査にあたり、注意すべき諸点を次に示しておく。

- 土左日記（静谿書屋本)、大和物語（大系本、底本は為家本）を調査
- 主節に現れた他動詞文、行為性の自動詞文、非行為性の自動詞文を採取
- ただし主語標示にガ/ノが現れ得る、連体形終止文や係り結び文は主節に含めない
- ハ・モ・ゾ・ナム・ヤ・カ・コソ等が主語ないし目的語に下接した例は、考察の対象としない

また、採取された他動詞文、行為性の自動詞文、非行為性の自動詞文の用例数は、次の通りである。

表 1 採取された主節の他動詞文・自動詞文

他動詞文	行為性の自動詞文	非行為性の自動詞文
89	29	87

[1] 同じく近藤 (2011) の指摘による。

本稿で扱う主節の無助詞名詞句とは、次のように、述語の項が無助詞として現れたものである。ここでは他動詞文の主語と目的語がたまたま共起しており、こうした例はごく少数である。

(1) a. いま、**けふあるひと、ところににたるうた**よめり。(土左日記・36-6)

b. **講師、もの、さけ**おこせたり。(土左日記・6-8)

c. 同じ**帝、狩**いとかしこく好みたまひけり。(大和物語・322-8)

ただし、無助詞名詞句が主節の述語の項であるかどうかを認める際には、注意が必要である。例えば、(2)(3)(4) に示す例では、ゴチックの無助詞名詞句が従属節述語の項か主節述語の項か明らかでない。

(2) a. **あるひと**、あがたのよとせいつとせはてて、れいのことゞもみなしをへて、げゆなどとりて、すむたちよりいでて、ふねにのるべきところへわたる。(土左日記・1-3)

b. **をとこをんな**、からくかみほとけをいのりて、このみとをわたりぬ。(土左日記・26-10)

c. **越前權守兼盛**、兵衞の君といふ人にすみけるを、としごろはなれて又いきけり。(大和物語・256-3)

(3) a. 十一日。**あめ**いさゝかにふりて、やみぬ。(土左日記・37-10)

b. かゝるあひだに、**ふなぎみの病者**、もとよりこち／＼しきひとにて、かうやうのこと、さらにしらざりけり。(土左日記・34-5)

c. 「**われ**、かたきにせめられてわびにて侍り。御はかし暫時かし給はらむ、ねたき物のむくひし侍らむ」(大和物語・315-11)

d. このあるじの、またあるじのよきをみるに、うたておもほゆ。いろ／＼にかへりごとす。**いへのひとのいでいり**、にくげならず、ゐやゝかなり。(土左日記・39-2)

(4) a. 十八日。なほおなじところにあり。うみあらければ、ふねいださず。**このとまり**、とほくみれども、ちかくみれども、いとおもしろし。(土左日記・17-10)

b. かくてふねひきのぼるに、なぎさの院といふところをみつゝゆく。**その院**、むかしをおもひやりてみれば、おもしろかりけるところなり。(土左日記・36-2)

このような例は、今回の考察の対象としない。

また、もちろん次の例のように、ゴチックの無助詞名詞句が従属節の述語の項である場合は考察の対象としない。

(5) つきのあかきにぞわたる。ひと／＼のいはく、「**このかは**、あすかがはにあらねば、ふちせさらにかはらざりけり。」といひて、…(土左日記・ 39-9)

3　無助詞名詞句の振る舞いと解釈

平安時代和文の主節では、主語、目的語が無助詞名詞句として現れるため、他動詞の主語、目的語、自動詞の主語が等しく扱われる中立型として分析することができるかもしれない。しかし古代日本語はSOV ないしSV 言語と考えられ、ある条件の下で他動詞の主語は必ず目的語に先行するから (後述)、語順を考慮すれば両者の扱いが等しいとするわけにはいかない。

しかしながら、語順をたよりにしてそれぞれの名詞句の振る舞いの特徴づけを行うことにも問題がある。例えば英語のようなSVO/SV の語順をもつ言語であれば、どちらのS もV の左側にありO だけが右側ということで、単純に他動詞の主語と自動詞の主語の振る舞いが等しいと見て、主語の振る舞いを主格、目的語を対格と特徴づけられる。これに対し、SOV ないしSV の語順をもつ言語においては、どちらのS もV の左側にあって一見振る舞いが等しく見えるが[2] 、O もV の左側にあるので、自動詞S の振る舞いが他動詞S の振る舞いと等しいのか、あるいはO の振る舞いと等しいのか明らかでない。すなわち、SOV/SV 言語では、語順のあり方からただちに分裂自動詞性[3]を否定できず、分裂自動詞性の検討のためにさらなる考察を加えなければならない。このようなわけで、日本語を含むSOV/SV 言語においては、節中の他の要素に対する振る舞いをたよりにしてそれぞれの名詞句の振る舞いの特徴づけを行うことが望ましい。

以下では、上に述べた観点からそれぞれの名詞句の振る舞いの特徴づけを行う。分裂自動詞性を検証するため、自動詞を行為性の自動詞と非行為性の自動詞の二種に分けて考察を行う。なお、他動詞文の主語と行為性の自動詞文の主語を一括する場合〈動作主〉主語と呼ぶことにする。また、以下で非行為性自動詞文の主語を〈対象〉主語と呼ぶことがある。

3.1　〈動作主〉主語の振る舞い

まずは、他動詞文の主語と行為性の自動詞文の主語の振る舞いから観察する。他動詞文の主語は、(6ab) に示すように述語と隣接することもあれば、(6cd) のように様態副詞や目的語等の要素が介在することもある。

(6) a. これをきゝてよろこびて、**ひと／＼**をがみたてまつる。(土左日記・38-2)

b. 人人よばひ、殿上人などもよばひけれど、あはざりけり。そのあはぬ心は、帝をかぎりなくめでたき物になむ思たてまつりける。**帝**召してけり。(大和物語・322-4)

c. **あるひとのこのわらはなる**、ひそかにいふ。(土左日記・9-1)

d. いま、**けふあるひと**、ところににたるうたよめり。(土左日記・34-5)

行為性の自動詞文の主語でも同様であり、(7ab) は述語と隣接しており、 (7cd) は様態副詞、ニ格名詞句が介在している。

(7) a. かくはいふものか。うつくしければにやあらん、いとおもはずなり。「わらはごとにてはなにかはせん。**おんなおきな**、ておしつべし。あしくもあれ、いかにもあれ、たよりあらばやらん。」とて、おかれぬめり。(土左日記・9-8)

[2]実際に Baker(2001) では、日本語風の語順をもつ言語の主語は英語風の語順をもつ言語と通じるところがあるとされている。

[3]自動詞の主語が何らかの手段で一律に他動詞の主語ないし目的語と等しく扱われるのではなく、自動詞の主語のある部分が他動詞の主語と等しく、もう一方の部分が目的語と等しく扱われることを言う。

b. とよみたりければ、**かの室にとまりたりけるしども**あはれがりけり。（大和物語・243-10）

c. 廿七日。かぜふき、なみあらければ、ふねいださず。**これかれ**、かしこくなげく。（土左日記・24-8）

d. **泉の大将**、故左のおほいどのにまうでたまへりけり。（大和物語・296-1）

より注目したいのは (8)(9) に示す例である。(8) では、他動詞文の主語と述語との間に主節の時を表わす成分が介在している。

(8) a. また、**としこ**、雨のふりける夜千兼を待ちけり。（大和物語・262-8）

b. **故兵部卿の宮**、この女のかゝることまだしかりける時、よばひたまひけり。（大和物語・286-4）

(9) では、行為性の自動詞文の主語と述語のとの間に、やはり主節の時を表わす成分が介在している。

(9) **堤の中納言**の君、十三のみこの母宮すむ所を内裏にたてまつりたまひけるはじめに、「帝はいかゞおぼしめすらむ」など、いとかしこくおもひなげき給けり。（大和物語・252-7）

さらには (10)(11) に示すような例もある。(10a-c) では他動詞の主語と述語との間にド節やテ節が介在している。

(10) a. また、**あるひと**、にしぐになれど、かひうたなどいふ。（土左日記・5-5）

b. **ふぢはらのときざね**、ふなぢなれど、むまのはなむけす。（土左日記・1-7）

c. **おなじ人**、かの父兵衛の佐うせにける年の秋、家にこれかれあつまりて、宵より酒のみなどす。（大和物語・244-10）

行為性の自動詞文の主語の場合でも、述語との間にテ節が介在している例を見出すことができる。

(11) a. こよひ、**ふなぎみ**れいのやまひおこりて、いたくなやむ。（土左日記・35-6）

b. そのうた、よめるもじ、みそもじあまりなゝもじ。ひとみな、えあらでわらふやうなり。**うたぬし**、いとけしきあしくて ゑ（ん）ず。（土左日記・18-10）

c. **平中**、閑院の御にたえてのち、ほど經て あひたりけり。（大和物語・252-12）

以上のように、他動詞文の主語と行為性の自動詞文の主語が共通した振る舞いを示すのは、ともに〈動作主〉であることに起因していると考えられよう。〈動作主〉主語と述語との間には介在する要素がなくてもよいし、ニ格名詞句や様態副詞等のより中核的な要素が介在することができるし、ある種の付加的な成分や節が介在することもできる。以上から、〈動作主〉主語は無助詞名詞句の形で述語から大きく離れることができると見るべきである。

3.2 〈対象〉主語と目的語の振る舞い

次に、非行為性の自動詞文の主語と他動詞文の目的語の振る舞いを観察する。非行為性の自動詞文の主語は、(12) に示すように、そのほとんどが述語と隣接している。

(12) a. 日しきりにとかくしつゝ、のゝしるうちに**よ**ふけぬ。（土左日記・1-6）

b. 廿二日。よんべのとまりより、こととまりをおひてゆく。はるかに**やま**みゆ。（土左日記・22-5）

c. おなじ女、内裏の曹司にすみける時、**忍びてかよひ給人**ありけり。（大和物語・270-1）

同じく他動詞文の目的語でも述語と隣接した例が目立つ。

(13) a. なほかみのたちにてあるじしのゝしりて、郎等までに**もの**かづけたり。（土左日記・2-8）

b. 五條にぞ少将の家あるに行きつきてみれば、いといみじうさわぎのゝしりて**門**さしつ。（大和物語・280-11）

c. これをみて、よくみまほしさに、「この蘆もちたるをのこ呼ばせよ、**かのあし**買はむ」といはせける、…（大和物語・ 318-13）

一方、述語と隣接しない場合では、非行為性の自動詞文の主語と述語との間には、ニ格名詞句、程度副詞、様態副詞等が介在する。

(14) a. むかし、**とさといひけるところにすみけるをんな**、このふねにまじれりけり。（土左日記・26-4）

b. **この女かほ容貌**いときよらなり。（大和物語・320-5）

c. かくいふあひだに、よやうやくあけゆくに、かぢとりら、「**くろきくも**にはかにいできぬ。かぜふきぬべし。みふねかへしてん。」といひて、ふねかへる。（土左日記・17-7）

d. かくうたふをきゝつゝこぎくるに、くろとりといふとり、いはのうへにあつまりをり。そのいはのもとに、**なみ**しろくうちよす。（土左日記・21-7）

また、同じく他動詞文の目的語でも、述語との間には、ニ格名詞句、頻度を表わす副詞、様態副詞、否定に呼応する副詞等が介在する。

(15) a. 「**かれが申さむこと**院に奏せよ。…」（大和物語・ 311-10）

b. かぢとりまたたひもてきたり。**よね、さけ**、しば／＼くる。かぢとり、けしきあしからず。（土左日記・15-8）

c. かくいひつゝくるほどに、「**ふね**とくこげ、ひのよきに。」ともよほせば、（土左日記・30-7）

d. かゝるあひだに、ふなぎみの病者、もとよりこち／＼しきひとにて、**かうやうのこと**、さらにしらざりけり。（土左日記・34-5）

このように、非行為性の自動詞文と他動詞文の目的語から共通した振る舞いが読み取れるのは、それらの意味役割がともに〈対象〉であるからであろう。〈対象〉主語と他動詞文の目的語は、その大半が述語に隣接して現れ、隣接しない場合は、程度副詞、様態副詞、ニ格名詞句、頻度を表す副詞、否定に呼応する副詞等が介在するに過ぎない。〈動作主〉主語の場合とは異なり、介在する要素がより中核的な要素に限られ、ある種の付加的な成分や節が介在することはない。無助詞名詞句の形で実現した〈対象〉主語や他動詞文の目的語は、〈動作主〉主語に比して述語からの距離が制限されていたと考えられる。

最後に、上の観察にとって問題となりそうな例について言及しておく。

(16) **からうた**、こゑあげて いひけり。（土左日記・2-8）

この例では、目的語と述語の間にテ節が介在し、問題例と見えそうであるが、このテ節は付帯状況を表わしており、先に (10c) や (11a-c) で示した、〈動作主〉主語と述語の間に介在するテ節とは異なる。次に (10c)(11a) を再掲する。

(17) a. (＝ 10c)
おなじ人、かの父兵衛の佐うせにける年の秋、家にこれかれあつまりて、宵より酒のみなどす。（大和物語・244-10）

b. (＝ 11a)
こよひ、**ふなぎみ**れいのやまひおこりて、いたくなやむ。（土左日記・35-6）

(17a) は継起、(17b) は原因・理由を表わすテ節であり、目的語と述語との間には、継起や原因・理由を表わすテ節は介在しなかったと考えられる。

また、次に示す例では、目的語と述語の間に副助詞つきの主語が介在している。

(18) からうた、こゑあげていひけり。**やまとうた**、あるじも、まらうども、ことひともいひあへりけり。（土左日記・2-9）

無助詞主題と捉える可能性もあるかもしれないが、処置に悩む例である。

3.3 解釈

Perlmutter(1978) では、自動詞のその唯一の項が、他動詞の目的語が動詞に対してもつような関係[4]をもち得ることが明らかにされた。すなわち非対格仮説の下では、初期層において、自動詞にはその唯一の項が主語である非能格動詞 (unergative verb) と、その唯一の項が目的語である非対格動詞 (unaccusative verb) の二種があるとされた。これをふまえ、Levin and Rappaport Hovav (1995) では、このような自動詞の唯一の項がそのまま目的語として表層に現れる場合を表層非対格 (surface unaccusative)、主語として現れる場合を深層非対格 (deep unaccusative) と呼んでいる。非対格仮説は、複層の理論である関係文法に依拠するものであり、関係文法では初期層と表層の間では文法関係の変更が行われるとされるので、深層非対格の場合は目的語から主語への非対格昇格 (unaccusative advancement) があることになる。したがって、世界の言語には非対格昇格のある表層非対格言語と、非対格昇格のない深層非対格言語が存することになる。この観点からすると、現代の日本語や英語は深層非対格言語ということになる。

[4]Perlmutter(1978) では、他動詞の目的語が動詞に対してもつような関係を 2arc、他動詞の主語が動詞に対してもつような関係を 1arc と称している。

以上のことは、より意味的な側面を重視して捉えなおすことができる。一般に動詞の項は外項 (external argument) と内項 (internal argument) からなり、他動詞は外項と内項をもち、非対格動詞は内項のみをもち外項をもたず、非能格動詞は外項のみをもち内項をもたないとする考え方がある。おおよそ外項には〈動作主〉が、内項には〈対象〉が対応するものと考えられる。これを前提とすると、外項、内項が句構造、語順、格標示の上でどのように実現するかが問題となるが、ここでは、項構造上の外項、内項の区別が何らかの形で保持されて実現する言語と、外項、内項の区別が形式上保持されず実現する言語があることに注意しておきたい。前者は活格性を有する言語、後者は対格性あるいは能格性を有する言語ということになる。

さて、以上をふまえると、先に見た平安時代の主節における無助詞名詞句の振る舞いはどのように解釈できるだろうか。よく知られているように、日本語は、述語が末尾に位置し、かつ、述語は、その述語動詞によって記述される出来事の主要な関与者 (主語、目的語、間接目的語) を個々に表わす表現を必ずしも含まなくてもよい言語である。よって、文法関係や意味役割のあり方に関係なく名詞句が容易に述語と隣接することになるし、また、自動詞の唯一の項が述語と隣接する割合が高くなるのは当然である。だとすれば、今回のデータの解釈において重視されるべきは、述語との隣接というよりは、述語との間にどのような要素を介在させることができるかであるように思える。

先の調査から、無助詞の〈動作主〉主語は大きく述語から離れられるのに対し、無助詞の〈対象〉主語と無助詞の他動詞文の目的語は〈動作主〉主語ほど述語から離れられないことが明らかになった。そうであるならば、古代日本語の主節では、項構造上の外項、内項の区別が、節中の他の要素に対する振る舞いにおいて保持されており、活格性を有しているとの特徴づけが成り立つ。すなわち、古代日本語の主節の無助詞名詞句は意味役割によってその振る舞いが決定されていると考えられる。

4　無助詞目的語とヲ格目的語

これまでに見たように、古代日本語の主節の名詞句は無助詞名詞句として現れるが、実のところ、主節の他動詞文の目的語がヲ格名詞句として実現した例も存する。

(19) a. かこのさきといふところに、かみのはらから、またことひと、これかれさけなにともておひきて、いそにおりゐて、わかれがたきことをいふ。(土左日記・4-4)

b. 「そも／＼まことか」など問はせ給に、鳥飼といふ題を皆人々によませ給けり。(大和物語・310-13)

(19a) は述語に隣接した例であり、(19b) は他の要素を介在させた例であるが、これらの振る舞いは先に見た無助詞目的語と変わらない。

ところが、興味深いことに、次のような振る舞いのヲ格目的語を見出すことができる。

(20) a. をとこもすなる日記といふものを、をむなもしてみんとてするなり。(土左日記・1-1)

b. 三月ばかり、こゝにわたりたるほどにしも、苦しがりそめて、いとわりなう苦しとおもひまどふを、いといみじとみる。(蜻蛉日記・148-16)

c. 忠文がみちのくにの将軍になりてくだりける時、それが息子なりける人**を**、監の命婦しのびてあひかたらひけり。（大和物語・263-6）

d. 扇どものをかしき**を**、その頃は 人々持たり。（紫式部日記・10-4）

e. そのものども**を** 九月つごもりにみないそぎはててけり。（大和物語・232-13）

主節におけるヲ格目的語と無助詞目的語を比較してみると、両者はともに目的語であっても、節中の他の構成要素に対する振る舞いが明らかに異なる。例えば、(20ab) はヲ格目的語と述語との間にトテ節ないしト節が介在した例、(20cd) はヲ格目的語と述語との間に〈動作主〉主語ないし主題句が介在している例、さらに (20e) は、主節時を表わす成分が介在している例であるが、管見では、節中の他の要素とこのような関係にある無助詞目的語を見出すことができなかった。

この観察が正しいとすると、ヲ格目的語は、格助詞ヲを伴うことで無助詞目的語以上の振る舞いを与えられていることになる。言い換えれば、ヲ格目的語は、格助詞ヲによって無助詞かつ〈対象〉としての名詞句の振る舞いから自由になっているということである。この意味で、格助詞ヲは、文法関係を標示するための単なる標識ではなく、目的語に独自の統語的な機能を与えていると見ることができる[5] 。こうした無助詞目的語とヲ格目的語の振る舞いの違いは、古代日本語の無助詞名詞句の振る舞いが意味役割により決定されていることの根拠となる。

5　古代日本語の格標識——ヲ・イ・シ——

古代語の格標識 (有形格助詞) の特性に言及しておく[6]。古代日本語では、いわば二足ないし三足のわらじの一方として格標識が存在している点に特徴があるが、この点について近藤 (2000) にすぐれた考え方が示されている。それまでの議論では、上代語の助詞ヲについて品詞的な性格づけができず、そこから格助詞ヲを取り出すのは困難であるとされ、文体・談話論的な側面からその機能が考察されていた。近藤 (2000) では、助詞ヲを構造上の観点から分析することで品詞分類の根拠が示され、間投助詞、格助詞、終助詞が明確に分類された。(21a) は格助詞、(21b) は間投助詞の例である。なお終助詞の例は省略する。

(21) a. にきびにし家ゆも出でてみどり子の泣く**を**も（哭乎毛）置きて…（萬葉集・ 481）

b. ほととぎすここに近く**を**（知可久乎）来鳴きてよ過ぎなむ後に験あらめやも（萬葉集・4438）

また、格助詞が生起する環境にある助詞ヲの標示のあり方に着目すると、〈対象〉主語ないし目的語に限られ、〈動作主〉主語を標示することがない (竹内 2008a)。こうした分布は、格助詞以外の振る舞いであるとすれば説明がつかないから、このことは助詞ヲに格助詞が認められる根拠となろう。

その後、竹内 (2008a)、竹内 (2008b) においても、助詞イや助詞シに助詞ヲと同様の分類が適用され、それまで均質に間投助詞ないし係助詞とされていた助詞イや助詞シから、格

[5] ヲバによる目的語も含め、無助詞目的語、ヲ格目的語等の関係を表せば次のようになる。
無助詞　≦　ヲ　≦　ヲバ
金水 (2001)、Yanagida(2005) 等で示された法則も考慮しつつ、今後この点を考究していきたい。

[6] 助詞ガ/ノについては、別に扱うこととする。詳しくは野村 (1993) を参照。

助詞を取り出せることが示された。格助詞イを (22a) に、間投助詞イの例を (22b) に、格助詞シの例を (23a) に、間投助詞シの例を (23b) にそれぞれ示しておく。

(22) a. 我が背子が跡踏み求め追ひ行かば紀伊の関守**い**（關守伊）留めてむかも（萬葉集・545）

b. 向つ峰の若楓の木下枝取り花待つ**い**間に（花待伊間尓）嘆きつるかも（萬葉集・1359）

(23) a. 大御舟泊ててさもらふ高島の三尾の勝野の渚**し**（奈伎左思）思ほゆ（所レ念）（萬葉集・1171）

b. 可之布江に鶴鳴き渡る志賀の浦に沖つ白波立ち**し**来らしも（多知之久良思毛）（萬葉集・3654）

助詞ヲの場合と同じく、格助詞が生起する環境にある助詞イ、助詞シの標示のあり方に着目すれば、助詞イの標示は〈動作主〉主語に限られ、助詞シの標示は〈対象〉主語ないし目的語に著しい偏りが認められる。これらの分布が格助詞以外の振る舞いであるとすると説明が難しく、助詞イや助詞シにおいても格助詞を設定する必要がある。格助詞ヲ、イ、シは外項と内項の別に対応した形で分布しており、格助詞イは動作主標識、格助詞ヲ、シは非動作主標識と言うことができる (Vovin1997、竹内 2008a、竹内 2008b)。

以上のように、助詞ヲ、イ、シではいずれも間投助詞と格助詞が認められるが、このことは偶然ではないように思われる。すなわち、いずれの助詞にも、機能的な成分 (間投助詞) から統語的な成分 (格助詞) へと推移していく文法化が生じた (あるいは生じつつあった) と見るべきである 。ヲ、イ、シの文法化は、まず、外項ないし内項と述語の結合を経過して、のちに主語標示ないし目的語標示へと進んでいくと想定され、このような文法化の過程に位置づけてこそ、先述した格助詞ヲ、イ、シの標示の分布はよく理解できる。結果として、ヲの文法化のみが収束し、イとシはその過程で途絶えてしまったと考えられる。

したがって、格助詞ヲ、イ、シには間投助詞に通じる性格が残されていよう。項と述語の結合を明示する統語的な成分であると同時に多分に機能的な成分としての性格もあるであろう。例えば、古代語の主節では、格助詞ヲ、イ、シの使用が任意であり、むしろ無助詞名詞句としてあるのがふつうである。

以上のことをふまえれば、主節において、格標識は発達しているとは言えず、古代語の主節の無助詞名詞句から成る体系は、活格性を含み、かつ、格標識が未発達である体系と言うことができる 。すなわち、語順、格等の手段によってバイアス (対格化ないし能格化) が加えられていないニュートラルな体系と言うことができる。

6　おわりに

以上、本稿では、無助詞目的語とヲ格目的語の比較や格標識ヲ、イ、シにも言及しながら、古代日本語の主節の無助詞名詞句の振る舞いやその体系について論じてきた。

古代語の主節の無助詞名詞句から成る体系にある種の活格性が認められるとすれば、その後の対格性を有する体系への移行が問題となる。また、上代語の主節の無助詞名詞句についても今回の結果との比較考察が必要であろうし、さらには、今回扱うことができなかった無助詞準体の振る舞いや無助詞主題の認定についての考察も今後の課題となるように思われる。

謝辞

本研究は、文部科学省科学研究費補助金若手研究 (B)「日本語の活格性にまつわる記述的研究」（平成 20〜23 年度、研究代表者：竹内史郎）による研究助成を受けています。

用例出典

○萬葉集…『萬葉集 本文篇』『萬葉集 訳文篇』 (ともに塙書房)、○土左日記…『土左日記総索引』(日本大学人文科学研究所)、○大和物語・蜻蛉日記…岩波日本古典文学大系、○紫式部日記…『紫式部日記』 (岩波文庫)

参考文献

金水敏 (1993)「古典語の「ヲ」について」仁田義雄 (編)『日本語の格をめぐって』pp. 191-224, くろしお出版

金水敏 (2001)「助詞から見た日本語文法の歴史」（文法学会研究会 第三回集中講義資料 第一分冊）

近藤泰弘 (2000)『日本語記述文法の理論』ひつじ書房

近藤泰弘 (2011)「通時コーパスの利用法と設計」NINJAL 共同研究「通時コーパスの設計」研究発表会（9 月 16 日、国立国語研究所）

竹内史郎 (2008a)「古代日本語の格助詞ヲの標示域とその変化」『國語と國文學』85 巻 4 号、pp. 50-63.

竹内史郎 (2008b)「助詞シの格助詞性——非動作格性と品詞分類——」『語学と文学』44 号、pp. 9-23、群馬大学語文学会

野村剛史 (1993)「上代語のノ・ガについて (上)」『國語國文』62 巻 2 号、pp. 1-17.

Baker, Mark C. (2001) *The Atoms of Language*, Basic Books.（郡司隆男［訳］(2003)『言語のレシピ 多様性にひそむ普遍性をもとめて』岩波書店）

Levin, B. and Rappaport Hovav, M. (1994) *Unaccusativity: At the Syntax-Lexical Semantics Interface*, MIT Press.

Perlmutter, D. M. (1978) Impersonal passives and the unaccusative hypothesis, *Proceedings of the Fourth Annual Meeting of the Berkeley Linguistics Society*, pp. 157-189, Berkeley Linguistic Society.

Vovin, Alexander (1997) On the syntactic typology of old Japanese, *Journal of East Asian Linguistics 6*, pp. 273-290, Kluwer Academic Press.

Yanagida, Yuko (2005) *The Syntax of Focus and WH-questions in Japanese: A Cross-linguistics Perspective*, Hitsuzi shobo Publishing.

ポスター発表（1）

3月5日（月） 13:00〜15:00

『日本語話し言葉コーパス』を用いた「全然」の変化の詳細化

佐野 真一郎（国際基督教大学）†

Anatomy of the Change of *Zenzen* Using the Corpus of Spontaneous Japanese

Shin-ichiro Sano (International Christian University)

１．はじめに

本研究では、近年話し言葉において顕著に用法の変化が見られる日本語の陳述副詞「全然」を取り上げ、この変化の様相を『日本語話し言葉コーパス』を用いて定量的に明らかにする。「全然」は他の語を伴って現われ（呼応）、その語の意味を強調する機能があるが、昭和前期以降「全然」と呼応する語は「全然大丈夫ではない」、「全然良くない」、「全然おいしくない」などのように打消しの言い方や、「全然違う」、「全然だめ」などの否定的意味を持つものであり、「全然大丈夫」、「全然良い」、「全然おいしい」などの肯定的なものは誤りであるとされてきた（広辞苑第 6 版）[1]。しかしながら、従来誤りとされてきた肯定的用法が近年増えつつある。この用法の変化について、これまでの研究では主に「全然」自体の意味に目が向けられ、また自然発話が定量的研究の対象とされた例はない（若田部 1991、鈴木 1993、新野 1997、野田 2000 他）。

そこで本研究では、「全然」の変化を他の表現との共起関係における文法変化として捉え、その変化の様相を定量的に捉えることを目的とする。具体的には、「全然」を呼応する表現により 1) 否定辞（例、「全然～ない」）、2) 伝統形（例、「全然違う」）、3) 革新形（例、「全然良い」）の 3 種類に分類し、自然発話を対象としてそれぞれの使用実態・経年変化、及び様々な要因が持つ変化への影響を求めた。以下、第 2 節では「全然」の用法の変遷を概観する。第 3 節においてデータ収集、分類・分析方法を述べ、第 4 節で『日本語話し言葉コーパス』を用いた分析を行う。最後に第 5 節で結論を述べる。

２．「全然」の変遷

前節では、「全然」の変化の否定的用法から肯定的用法という一側面を見たが、「全然」の変遷は「否定から肯定へ」のように単純に捉えられるものではなく、複雑な道筋を辿ってきた。「全然」は江戸後期に中国語からの借用語として日本語に入ってきた。これが日本語として定着・確立するのは明治 40 年代以降で、この時代の「全然」は否定的にも、否定辞を伴わずに肯定的にも使うことができた。つまり、現在誤った用法と見なされている肯定的用法がかつては正しい用法だったということである。このことは以下の文学作品においても確認できる。

(1) そこで三人が全然翻訳権を与次郎に委任する事にした　（夏目漱石　三四郎）
(2) この老婆の生死が、全然自分の意志に支配されているということを意識した　（芥川龍之介　羅生門）

(1)、(2)は、明らかに文中に否定辞がなく、否定的な呼応表現と一緒に現れてもいないが、このような用法が明治では正しいものとして使われていたということを示している。以降、元々正しい用法だった肯定的用法が誤りとされ、否定的用法のみが正しいとされるように

† shinichiro@icu.ac.jp

[1] 「全然」の肯定的用法について、その存在を全く否定しているわけではなく「俗な用法で、肯定的にも使う」としている。

なるのであるが、明確な理由は明らかになっていない。しかしながら、大正期の終わり頃から肯定的用法が使われなくなり、結果として否定的用法のみが正しい用法とされるようになったと言われる。近年増えつつある用法が示すように、そこから再度肯定的用法が使われるようになるのであるが、それは昭和 20 年以後であると言われている（鈴木 1993）。これまでの「全然」の肯定的・否定的用法に関する変化の流れを図 1 にまとめる。

図 1　「全然」の用法・呼応表現の変遷

このように見ると、否定的用法に関しては一貫して使われており、この意味では変化がないが、肯定的用法に関しては元々使われていたものが昭和前期頃から使われなくなり、昭和後期頃から再度使われ出したということが見て取れる。つまり、肯定・否定の二項対立に限って言えば、「全然」の変化は「否定から肯定へ」ではなく、「肯定的用法の復活」の方が事実を正しく反映しているとも言える。また、呼応表現の観点から考えれば、「全然」は元々否定辞以外の多様な呼応表現とも共起していたが、昭和前期頃からその呼応表現が否定辞や否定的な意味を持つものに限定されるようになり、昭和後期頃から再度多様な呼応表現を許すようになったと言える。言い換えれば、昭和前期から後期にかけてのみ呼応表現に制限があったということである。なお、本研究ではデータの性質上、明治から昭和前期にかけて起こった肯定的用法の消失に関わる変化については対象とせず、昭和後期から始まり、近年多く観察されるようになった肯定的用法の拡大に焦点を絞ることとする。

３．データ収集、分類・分析方法

本節では、分析の前提となるデータ収集、分類・分析方法について述べる。まずデータ収集では、『日本語話し言葉コーパス』の全 3,302 講演（コア・ノンコア）を対象とした。TRN-SJIS フォルダに格納される「転記テキスト」（拡張子.trn）を用い、片仮名表記の「ゼンゼン」という文字列で検索した。「全然」については、文字列のみによる検索で十分な絞り込みができるため、その他の形態論情報などは参照していない。次に、文字列による検索で収集した用例を文脈を参照し精査する。その基準は以下の通りである。

1. 応答：対話において、前発話に対する「全然」のみによる応答（相づち）
2. 形容動詞：「全然」単独で状態を表すもの（例、「全然。」、「全然です」）
3. 言い指し：発話を中断し、そのまま次の発話に移ってしまうもの
   → 1〜3 については、呼応表現との対応が確認できないため除外
4. 言い直し：他の表現に言い換えられた、又は言い換えるために使われたもの（例、「全然というか」）
   → 除外
5. 繰り返し：近距離で「全然」が 2 度現われており、対応する呼応表現が 1 つであるもの
   → 1 例と見なす
6. その他、間にフィラーが入ったもの（例、「ゼン笑ゼン」）や語断片などの不完全な発話も除外（多くは検索の時点で排除される）

以上の収集過程を経て、今回の分析対象となった「全然」の全用例は 1,534 件である。

次にこれらの用例を分類する。本分析では、梅林（1994）に倣って「全然」をその呼応表現の種類によって「否定辞」、「伝統形」、「革新形」という 3 種類に分類した。図 2 に分

類方法をまとめる。

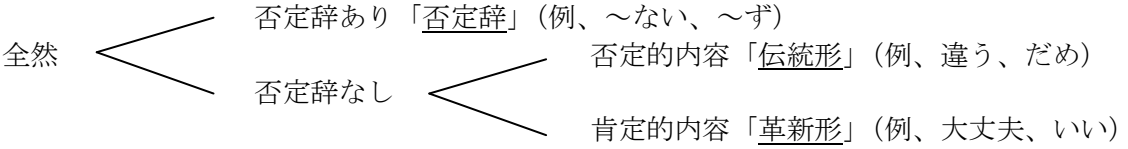


図 2　「全然」の呼応表現による分類

まず、呼応表現が「～ない」、「～ず」などの否定辞である場合、これらを「否定辞」とする。次に、否定辞を伴わないものをさらに 2 種類に分ける。呼応表現が「違う」、「だめ」などのように従来から文法的であると見なされている否定的内容である場合は「伝統形」とし、「大丈夫」、「いい」など近年多く見られるようになった肯定的内容である場合は「革新形」とした（他にも若田部（1991）、新野（1997）などではより詳細な分類があるが、本研究の目的は「全然」の分類ではなく、変化の実態を明らかにすることであるため、今回のデータの分析に過不足がないと思われる 3 種類とした）。

最後に分析方法に関して、「全然」の分布に影響を与え得る要因によって否定辞、伝統形、革新形それぞれの分布がどのように変わるかということを中心的に見る。従って、分析は要因別に進める。本分析で対象としたのは、言語外的要因[2]に相当する生年、講演種、性差、発話スタイルの合計 4 要因である。生年については「全然」の経年変化を求める際の個別の指標として利用するが、その他 3 要因についてはこれらの相互作用も検証する。また、以下では主に各文脈に現れる「全然」の比率に基づいて話を進める。

４．分析

４.１　「全然」、呼応表現の分布

まず、上記の分類方法に従って全用例を分類した。内訳を表 1 に示す。

表 1　「全然」の呼応表現ごとの分布

	頻度	比率
否定辞	1,112	72.49%
伝統形	263	17.14%
革新形	159	10.37%
合計	1,534	100%

表 1 が示すように、「全然」の約 70%が否定辞と共起している一方で、革新形は 10%程度となっており、革新形が近年多く見られるようになってはいるものの、依然として否定辞と共起することがほとんどであることがわかる。つまり、変化はまだ初期段階にあると考えられる（後述）。また、伝統形に関しても否定辞と比べると少ないと言える。

次に、否定辞、伝統形、革新形それぞれの内訳を見る。以下に各呼応表現の中で頻度の高かったものを 8 件ずつまとめる。

[2] 言語内的要因・言語外的要因の詳細については Labov（1994）、（2001）が詳しい。

表 2　各呼応表現の内訳（上位 8 種）

否定辞		伝統形		革新形	
ない	613	違う	198	いい	14
なく	191	別	16	大丈夫	10
なか	160	駄目	14	平気	8
ません	108	変わった（て）	10	普通	4
ず	29	異なる／異質	5	オーケー	3
ねえ	3	少ない	4	うまい	2
なし	2	だけ	3	安全／安心	2
ぬ	2	逆	2	元気	2

まず、否定辞の中でも変異があり、「ない」が 613 件とほぼ半数を占めているものの、その他丁寧語の「ません」（108 件）や「ず」（29 件）なども観察された。伝統形の中では、「違う」が 198 件で過半数を占めている。その他では「別」が 16 件、「駄目」が 14 件で続いている。「違う」の多さが際立っているが、この結果は先行研究における、否定辞を伴わない形式では「違う」、「駄目」が多い（野田 2000 他）という主張とも一致する。革新形では、「いい」、「大丈夫」、「平気」という典型的な例が多く観察された。

４.２　全然の変化

次に、「全然」が具体的にどのような変化を辿っているかということを分析する。ここ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）。以下に「全然」の経年変化を「否定辞」、「伝統形」、「革新形」ごとにまとめた。

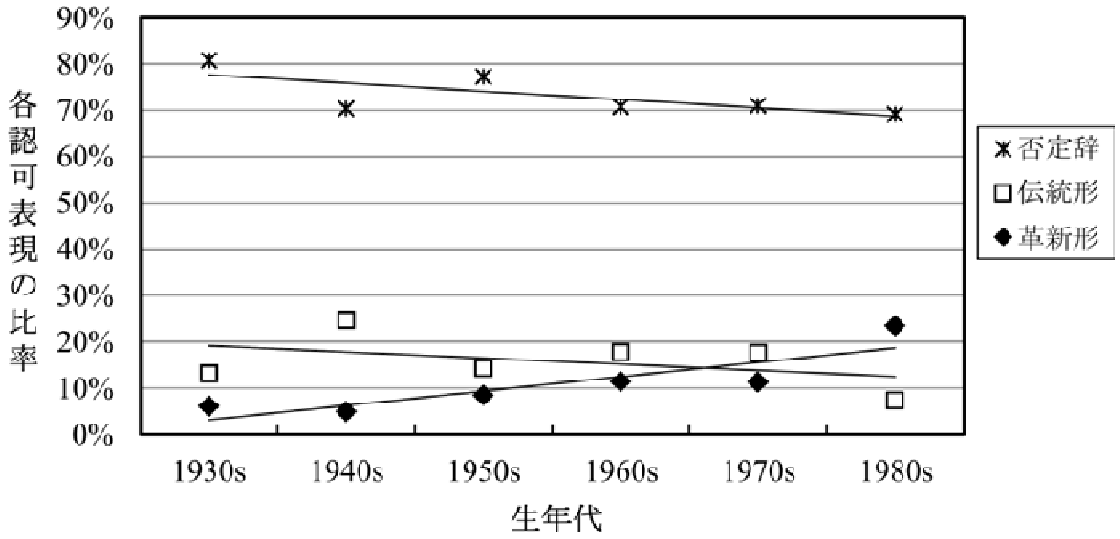


図 3　「全然」の呼応表現別経年変化

図 3 において、否定辞は 1930 年代生まれの話者（以下、「～年代」とする）から 1980 年代まで一貫して最も高い比率を示している。また、1930 年代で約 80%という最も高い比率で、そこから 1980 年代に向かってほぼ一定の速度で徐々に減少している。伝統形は、否定辞と比べると一貫して低い比率を示している。しかしながら、僅かな変動はあるものの、否定辞の変化と同様の傾向を示し、1980 年代に向かって徐々に減少している。また否定辞と伝統形の直線の傾きを比べるとほとんど違いがないことが見て取れる。つまり、経年変化において、両者はほぼ同じ道筋を辿っていると考えることができるのである。一方、革新形の変化は否定辞、伝統形とほぼ正反対の性質を示している。1930 年代、1940 年代で最も低

い比率を示し、そこから 1980 年代に向かってほぼ一定の速度で徐々に増加している。また、比率の最も高い 1980 年代では、否定辞に次いで高い数値を示しており、伝統形の比率を逆転している。

これらの結果から、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ということが実証された。更に、各呼応表現間の関係についても知見が得られた。つまり、単純に肯定的用法が増加してきたということではなく、従来使用されてきた否定辞と伝統形に替わって革新形が使用されるようになるという呼応表現間の役割交替が起こっていると言えるのである。また、伝統形は打消し表現（明示的な否定辞）を伴わないが否定的な意味合いを持つということを考えると、否定的な呼応表現全体が減少傾向にあり、それに代わって肯定的な呼応表現が増加していると考えることができる。図 3 が示すように、肯定的な呼応表現（革新形）が増加し、全ての「全然」の用法のうちある程度の割合を占めるようになったことを考えると、呼応表現を否定辞や否定的な意味を持つものに限定するという制限が弱まり、以前のように否定辞以外の多様な呼応表現とも共起することができるようになりつつあるということも支持される。

このように、経年変化に注目することで、従来注目されなかった各呼応表現間の関係が言語変化の進行過程でどのように変わって行くのかということが明らかになった。つまり、「全然」の変化は肯定的用法だけの問題ではなく、従来の否定的用法にも変化があるということである。なお、今回のデータでは 1910 年代・1920 年代が少なく[3]、肯定的用法が認められていたと思われる時代の分析を行うことができなかった。明治から大正頃までは肯定的用法が見られたことを考えれば、もしこれらのデータを合わせて分析することができたならば、肯定的用法（革新形）の経年変化においてその比率は下降から上昇を描く V 字型になっていたと予測できる。

４.３　講演種

ここでは、講演種の違いによって「全然」の分布がどのように変わるかということを検証する。『日本語話し言葉コーパス』の講演は主に「学会講演」と「模擬講演」に分けられ、前者はあらたまった発話スタイルが特徴的で、後者はくだけた発話スタイルが特徴的である。この学会講演・模擬講演の違いをあらたまった発話・くだけた発話の違いと見なして「全然」の分布を比較する。「全然」の変化があらたまった発話スタイルを選好するのであれば、革新形の比率は模擬講演よりも学会講演で高く、反対にくだけた発話スタイルを選好するのであれば学会講演よりも模擬講演で高くなると予測できる。また、否定辞と伝統形は革新形の分布と反対の性質を示すはずで、変化があらたまった発話スタイルを選好するのであれば模擬講演よりも学会講演で低く、反対にくだけた発話スタイルを選好するのであれば学会講演よりも模擬講演で低くなると予測できる。これは、図 3 が示すように変化の向かう先ではその他の文脈と比べ、新しい形式が増加を示し、古い形式が減少を示すという考えに基づく。以下に結果をまとめる。

表 3　「全然」の講演種別分布[4]

	否定辞	伝統形	革新形	合計
学会講演	68.54% (122/178)	23.60% (42/178)	7.87% (14/178)	178
模擬講演	73.26% (896/1223)	15.62% (191/1223)	11.12% (136/1223)	1223

$\chi^2 = 7.9617, df = 2, p = 0.01867$

[3] 生年がそれ以前の話者のデータは『日本語話し言葉コーパス』には収録されていない。

[4] 以下の分析において、分布の有意差をカイ二乗独立性の検定により検証する（有意確率 5%）。また、期待度数が少ない場合は、フィッシャーの正確確率検定も併せて行う。なお、検定には統計プログラム R (version 2.13.0)を用いた。

革新形に関しては、学会講演では 7.87%であるのに対して、模擬講演では 11.12%という比率を示しており、模擬講演で多く現れるということが分かる。また伝統形に関しては、革新形とは正反対で、学会講演では 23.60%であるのに対して、模擬講演では 15.62%と少ない比率を示している。つまり、革新形の比率が高く、伝統形の比率が低いのは模擬講演であるため、このことから「全然」の変化はくだけた発話スタイルを選好していると考えられる。しかしながら、否定辞に関しては仮説と異なり、学会講演（68.54%）よりも模擬講演（73.26%）の方が全体の比率としてわずかではあるが高くなっている。

結果として、否定辞では確認できなかったものの、革新形、伝統形においては「全然」の変化がくだけたスタイルに特徴的である、より進んでいるということが示された。

４.４　性差

これまでの言語変異・変化研究において、話者の性別が言語変化において大きな役割を果たすことが実証されてきている。中でも多くの場合、女性が言語変化をリードすると言われている（Trudgill 1972、Labov 1990 他）。従って、「全然」の場合も性差の影響を検証することは重要である。「全然」の変化をリードするのが女性であるならば、革新形の比率は男性よりも女性で高くなり、否定辞と伝統形の比率は男性よりも女性で低くなるということが予測できる。以下に分布を示す。

表 4　「全然」の男女別分布

	否定辞	伝統形	革新形	合計
男性	71.11% (544/765)	18.04% (138/765)	10.85% (83/765)	765
女性	73.86% (568/769)	16.25% (125/769)	9.88% (76/769)	769

$\chi^2 = 1.4583, df = 2, p = 0.4823$

表 4 が示すように、「全然」の分布は男女によってほとんど変わらない。つまり、性別に関係なく同じように「全然」を使っているということである。この結果は上記の予測とは異なる。しかしながら、『日本語話し言葉コーパス』では講演種によって男女の構成比が大きく異なる（学会講演では男性の方が多い）。この影響を考慮し、講演種別の分布を男女ごとに求めると以下のようになる。

表 5　「全然」の講演種、及び男女別分布

男性	否定辞	伝統形	革新形	合計
学会講演	67.12% (98/146)	25.34% (37/146)	7.53% (11/146)	146
模擬講演	72.36% (411/568)	14.96% (85/568)	12.68% (72/568)	568
女性	否定辞	伝統形	革新形	合計
学会講演	75.00% (24/32)	15.63% (5/32)	9.38% (3/32)	32
模擬講演	74.05% (485/655)	16.18% (106/655)	9.77% (64/655)	655

男性：$\chi^2 = 10.4087, df = 2, p = 0.00549$
女性：$\chi^2 = 0.0145, df = 2, p = 0.9928$

まず女性に注目すると、否定辞、伝統形、革新形の全てにおいてほとんど差がない。つまり、女性に関しては学会講演でも模擬講演でも「全然」を同じように使っているということである。一方、男性は学会講演・模擬講演で大きな差が見られる。具体的に、革新形は学会講演（7.53%）よりも模擬講演（12.68%）の方が高い比率を示している。否定辞も同様に、学会講演（67.12%）よりも模擬講演（85.00%）の方が比率が高くなっている。伝統形はこれらとは反対に、学会講演（25.34%）の方が模擬講演（14.96%）よりも高い比率を示している。ここで、表 5 の男性の講演種別の分布と表 3 の講演種別の分布を比べるとほと

んど同じであることが分かる。つまり、講演種別の分布は、女性には差がないため、男性の分布の特徴がそのまま反映したと言える。ここまでで、「全然」の変化において男性には学会講演・模擬講演の差があるが、女性には差がないということが明らかとなった。このことから、女性よりも男性の方が場面によるスタイル差が大きいということが読み取れる。しかしながら、従来の性差・スタイルに関する研究によるとむしろ女性の方がスタイル差が大きいと言われており（Labov 1990 他）、矛盾する結果となる。

この原因として、「顕在的威信」（overt prestige）、「潜在的威信」（covert prestige）とこれらに関する男女の振る舞いの違いの影響が考えられる。女性は公に高く評価されるような「正しい言葉遣い」を好む傾向があり、男性は反対にスラングなどの公には高くは評価されないが、男らしさ・たくましさが強調されるような言葉遣いを好む傾向がある（Trudgill 1972）。つまり、女性は顕在的威信を好み、男性は潜在的威信を好むということである。このことを念頭に「全然」の変化を考えると、まず「全然」の肯定的用法（革新形）は公には高く評価されているとは言えない、むしろ誤りであると非難される傾向にある（教育出版、cf. 新野 1997）。従って、「全然」の変化は顕在的威信よりも潜在的威信と結びついていると考られる。そうすると、男性は潜在的威信を好むため「全然」の肯定的用法を積極的に使い、場面によるスタイル差に敏感に反応する。一方、女性はこのような言語表現については積極的ではないため、場面によって特に意識的に使い方をコントロールするということはしないという説明が与えられる。以上のような違いが男女ごとのスタイルの違いとして現われたと考えられる。また、スタイルと性差が言語変化において密接に関係していることが本分析でも実証された。

４.５　発話スタイル

先の講演種に関する分析において、「全然」の変化はスタイルの影響を受ける、その変化は特にくだけた発話において特徴的であるということを確認した。しかしながら、それぞれの講演種の中にもスタイル差があることが予想できる。つまり、学会講演の中でもスタイルの高い講演もあれば低い講演もあるということが考えられる。これは模擬講演でも同様である。従って、講演種とは別にスタイルを検証する必要がある。『日本語話し言葉コーパス』には講演種以外にスタイルを測る方法として「発話スタイル」という指標が付与されているため、これを基にして「全然」の分布を検証する。講演種に関する分析を参考にすれば、革新形の比率はあらたまった発話よりもくだけた発話で高く、否定辞や伝統形の比率はあらたまった発話よりもくだけた発話の方が低いということが予測される。

表 6　「全然」の発話スタイル別分布

	否定辞	伝統形	革新形	合計
1	77.44% (127/164)	6.71% (11/164)	15.85% (26/164)	164
2	74.30% (347/467)	15.63% (73/467)	10.06% (47/467)	467
3	71.74% (457/637)	17.74% (113/637)	10.52% (67/637)	637
4	70.68% (94/133)	23.31% (31/133)	6.02% (8/133)	133
5	70.59% (12/17)	29.41% (5/17)	0.00% (0/17)	17

$\chi^2 = 25.5925, df = 8, p = 0.001233$

表 6 における縦軸の数値は 5 段階の発話スタイルを示している。1 が最もくだけたスタイルで、数字が増えるに従ってあらたまったスタイルとなり、5 が最もあらたまったスタイルである。革新形に関しては、最もくだけた 1（15.85%）で最も高い比率を示していて、発話があらたまるにつれて徐々に下がる。最もあらたまった 5 では 1 例も観察されなかった。伝統形に関しては、革新形とは正反対の分布を示している。最もくだけた 1（6.71%）において最も低い比率で、そこから徐々に高くなり、最もあらたまった 5 では 29.41%を示している。一方否定辞に関しては、 1（77.44%）で高い比率を示していて、わずかではあるがそ

こから 5（70.59%）に向かって徐々に減少する。

この分布は講演種に関する分布と全く同じ傾向を示している。この結果から、否定辞では確認できなかったが、革新形、伝統形においては先述の予測の通り、「全然」の変化がくだけたスタイルに特徴的である、より進んでいるということが再度実証された。

最後に、前節における講演種と性差の関係と同様、発話スタイルと性差の関係を検証する。先述のように、講演種はスタイル差の指標として利用することができるが、それぞれの講演種の中にもスタイル差があることが予想できるためである。これまでの議論において確認された「全然」の変化がくだけた発話スタイルを選好し、且つ男性はスタイルの変化に敏感であるが女性はそうではないという一般化が正しければ、男性の発話では革新形の比率はあらたまった発話よりもくだけた発話で高く、否定辞や伝統形の比率はあらたまった発話よりもくだけた発話の方が低いが、女性の発話ではこのような差が見られないということが予測される。この仮説を検証するために、以下に分布をまとめる。

表 7　「全然」の発話スタイル、及び男女別分布[5]

男性	否定辞	伝統形	革新形	合計
1	74.31% (81/109)	7.34% (8/109)	18.35% (20/109)	109
2	74.46% (172/231)	16.88% (39/231)	8.66% (20/231)	231
3	70.46% (229/325)	17.54% (57/325)	12.00% (39/325)	325
4+5	66.67% (34/51)	29.41% (15/51)	3.92% (2/51)	51
女性	否定辞	伝統形	革新形	合計
1	83.64% (46/55)	5.45% (3/55)	10.91% (6/55)	55
2	74.15% (175/236)	14.41% (34/236)	11.44% (27/236)	236
3	73.08% (228/312)	17.95% (56/312)	8.97% (28/312)	312
4+5	72.73% (72/99)	21.21% (21/99)	6.06% (6/99)	99

男性：$\chi^2 = 20.2337, df = 6, p = 0.00252$
女性：$\chi^2 = 9.5998, df = 6, p = 0.1425$

表 7 において、男性は「1」（くだけた発話）から「4+5」（あらたまった発話）まで大きな差が見られる。具体的に、革新形は「1」（18.35%）で最も高い比率を示し、そこから「4+5」（4.65%）に向かって下降する。伝統形はこれらとは反対に、「1」（7.34%）で最も低い比率を示し、そこから「4+5」に向かって上昇する（34.88%）。否定辞のみこのような目立った傾向は観察されなかった。一方女性に関しては、男性と同様否定辞に関しては目立った傾向は見られない。伝統形、革新形では、表面上男性と僅かに似た傾向を示しているものの統計的に有意な差はない。つまり、女性のスタイル差に関しては一般化できるほどの傾向はないということである。このことから、仮説の通り女性に関してはあらたまった発話でも、くだけた発話でも「全然」を同じように使っているということが再度確認できた。

以上のように、性差について発話スタイルにおいても講演種と同様の傾向が観察された。具体的には、「全然」の変化はあらたまった発話よりもくだけた発話を選好しているが、そのスタイル差は男性のみに見られるものである。この原因として、先述の「全然」自体の性質と潜在的威信の関係に基づく説明が可能性として挙げられる。

５．結論

本研究では、「全然」の歴史を概観し、その呼応表現との関係における変化を『日本語話し言葉コーパス』を使って分析した。分析では「全然」の呼応表現を否定辞、伝統形、革

[5] 発話スタイルの 5 段階のうち、「5」は表 7 のように細分化した場合極めて観測度数が少なく、統計的に信頼できる結果を得ることが難しいため、隣り合う「4」とカテゴリーの併合を行った。

新形の 3 種類に分類し、変化の様相を視覚化すると共に言語外的要因の影響、及びそれらの相互作用を検証した。分析の結果、以下の点が明らかになった。

・変化：「全然」の変化は肯定的用法の変化だけでは捉えられず、従来使用されてきた否定辞と伝統形に替わって革新形が使用されるようになるという呼応表現間の役割交替が起こっている。また、呼応表現に関する共起制限が弱まっている。

・講演種：革新形、伝統形においては「全然」の変化が学会講演（あらたまったスタイル）よりも模擬講演（くだけたスタイル）に特徴的であり、より進んでいる

・性差：性差はそれ自体では影響を与えていないが、講演種・発話スタイルなどスタイル差に関する要因と相互作用を示し、複合的に影響する

・講演種・性差：「全然」の変化において男性には学会講演・模擬講演の差があるが、女性には差がない。これには「全然」自体の性質と潜在的威信の関係に基づく説明が可能性として挙げられる。

・発話スタイル・性差：「全然」の変化はあらたまった発話よりもくだけた発話を選好しているが、そのスタイル差は男性のみに見られる。

その他、言語外的要因に基づく各分布に関して、性別（女性）以外の全ての場合において伝統形が革新形と正反対の分布を示した。つまり、伝統形が多い項目では革新形が少ない、革新形が多い項目では伝統形が少ないという相補分布となっていた。しかしながら、否定辞だけこれら 2 種類とはっきりとした関係が見られず、また振る舞いに一貫性もなかった。このことは、「全然」の変化において、革新形の増加と直接関係があり、革新形に取って代わられるのは否定辞ではなく、伝統形であるという図式を示唆すると共に、伝統形と革新形は統語的に密接に関連していて、否定辞だけが統語上異なる位置付けであるということと一致している。

今後の課題として、要因間の相互作用について本研究では講演種・発話スタイルと性差を検証したが、その他の要因についても研究を拡大する必要がある。また、「全然」の変化は現在初期段階にあるが、今後変化が進行するにつれてこれら要因間の関係がどのように変化するのかということも検証する必要がある。

謝　辞

本稿の執筆、及び調査に際し、加藤泰彦氏、日比谷潤子氏のご協力を賜った。ここに記して感謝を申し上げる。なお、本稿における不備は全て筆者に帰するものである。

文　献

Horn, Laurence (1972) *On the Semantic Properties of Logical Operators in English*, Ph.D. dissertation, UCLA.

Horn, Laurence (1989) *A Natural History of Negation*, Chicago: The University of Chicago Press. [reissued by CSLI, 2001].

Horn, Laurence (ed.) (2010) *The Expression of Negation*, Berlin/New York: De Gruyter Mouton.

Jespersen, Otto (1917) *Negation in English and Other Languages*, Copenhagen: Andr. Fred. Host & Son, Kgl. Hof-Bonghandel.

Kato, Yasuhiko (1985) *Negative Sentences in Japanese*, Ph.D. Dissertation, Sophia University.

加藤泰彦、吉村あき子、今仁生美(編) (2010)『否定と言語理論』、開拓社

Klima, Edward (1964)“Negation in English,”in Jerry Fodor and Jerrold Katz (eds.), *The Structure of Language*, Englewood Cliffs, NJ: Prentice-Hall, pp.246-323.
Labov, William (1990) “The intersection of sex and social class in the course of linguistic change,” *Language variation and change* 2, pp.205-254.
Labov, William (1994) *Principles of linguistic change: Internal factors*, Oxford: Basil Blackwell.
Labov, William (2001) *Principles of Linguistic Change vol.2: Social Factors*, Oxford, UK: Blackwell.
新野直哉(1997)「「"全然"+肯定」について」、『国語論究 第 6 集 近代語の研究』、pp.258-286、明治書院
野田春美(2000)「「ぜんぜん」と肯定形の共起」『計量国語学』22:5、pp.169-182、計量国語学会
太田朗(1980)『否定の意味』、大修館書店
鈴木英夫(1993)「新漢語の受け入れについて―「全然」を例として―」松村明先生喜寿記念会編『国語研究』、pp.428-449、明治書院
Trudgill, Pater (1972)“Sex, covert prestige and linguistic change in the urban British English of Norwich,”*Language in Society* 1, pp.179-195.
梅林博人(1994)「副詞「全然」の呼応について」『国文学解釈と鑑賞』59:7、pp.103-110、ぎょうせい
van der Wouden, Ton (1997) *Negative Contexts: collocation, Polarity, and Multiple Negation*, London: Routledge.
若田部明(1991)「「全然」の語誌的研究―明治から現代まで―」『解釈』37:11、pp.24-29、教育出版センター
吉村あき子(1999)『否定極性現象』、英宝社

関連 URL

教育出版『「全然すばらしい」という言い方は正しいか』
http://www.kyoiku-shuppan.co.jp/view.rbz?nd=1750&ik=1&pnp=101&pnp=113&pnp=566&pnp=1750&cd=19

「かなしい」と「つらい」の意味について

加藤恵梨（名古屋大学留学生センター）

A Semantic Analysis of *kanashii* and *turai*

Eri Kato（Education Center for International Students, Nagoya University）

１．はじめに

意味が類似していると考えられる「かなしい」と「つらい」という感情形容詞について、『現代日本語書き言葉均衡コーパスモニター公開データ（2009 年度版）』などを基に[1]、それぞれの語の意味と、両語の意味の類似点および相違点を明らかにする。

２．「かなしい」の意味分析

はじめに、「かなしい」の意味を分析する[2]。

2.1　先行研究の記述とその検討

森田（1977：166）は「かなしい」の意味を次のように記述している。

> かなしい〔悲しい〕
> 不幸な状況に接し、心が痛む気持ちである場合に使うが、そのような気持ちを人々に起こさせる事物にも言う。
> 「父の死が悲しい」「親友に裏切られるとは悲しいよ」「悲しい最期」

続いて、『学研国語大辞典（第二版）』（1988：363）は「かなしい」の意味を次のように記述している。

---

[1] 例文の後に出典と「*」が付してあるものは『現代日本語書き言葉均衡コーパスモニター公開データ（2009 年度版）』からの引用であることを示す。

[2] 『使い方の分かる類語例解辞典』（2003：246）は、「かなしい」には「悲しい」と「哀しい」という漢字表記があるとし、「悲しい」は「好ましくない事態に接し、心が痛むさま」を表し、「哀しい」は「かわいそうで哀れに思う気持ち」を表すと記述している。しかし、実例を見ると、必ずしも漢字表記の違いによって意味が異なると言うことはできない。

> ・側に行って「昌太」と言って首を撫でたら冷たかった。やはり本当に死んでしまったのだ。とても哀しかった。涙が出そうになったが我慢をした。　（ターキーの気まぐれ日記*）

上の例は、かわいがっていた馬が死んだことに「哀しい」と感じている。これは『使い方の分かる類語例解辞典』の言う「好ましくない事態に接し、心が痛むさま」を表し、「かわいそうで哀れに思う気持ち」を表しているとは考えられない。このように、「かなしい」の漢字表記の違いが必ずしも「かなしい」の意味の違いに対応しているとは考えられないため、本研究では「かなしい」の意味の違いと漢字表記の違いが関係しているとは考えず、「かなしい」の意味を分析する。

かなしい〔悲しい・哀しい〕

① 〔自分の無力や不幸を痛感して、あるいは他人の不幸に同情して〕胸がきつくしめつけられて、泣きたくなるような気持ちだ。

「時々わかったかわかったかと念をおして聞かれるが、大方それがよく分らぬので妙に―・かった」

② 胸がしめつけられて泣きたくなるような気持ちを起こさせる。あわれだ。

「新しきインクのにほひ栓抜けば　飢ゑたる腹に　沁むが―・しも」

上のように、森田は「かなしい」の意味を「不幸な状況に接し、心が痛む気持ちである」と記述し、『学研国語大辞典（第二版)』は「胸がきつくしめつけられて、泣きたくなるような気持ちだ」と記述している。次の例(1)のように、父母、兄弟などの親しい人を失ったことで「かなしい」という感情が生じている場合、森田や『学研国語大辞典（第二版)』の記述で「かなしい」の意味を説明できる。しかし、次の例(2)のように、話し手がやるべきであると考えていることを不動産・建設業界がやろうとしないことに「かなしい」と感じている場合、「かなしい」は「心が痛む気持ち」あるいは「胸がきつくしめつけられて、泣きたくなるような気持ち」というような強いマイナスの感情を表しているとは考えられない。よって、「かなしい」の意味記述について再度検討する必要がある。

(1) 世の中に、なにが悲しいといって、父母、兄弟、姉妹、そのほか親しい人たちを失うことぐらい悲しいものはありません。

（森岡美子『萬葉集物語』冨山房インターナショナル　p. 99）

(2) 物理的・経済的に耐用年数の長い優良な社会資産を築き、暮らしやすく文化度の高い街を形成していくなど、不動産・建設業界には、もっともっと、やるべきことがたくさんあるはずです。

私にいわれなくとも、こんなこと、個人レベルでは業界のだれもが気づいていること。

しかし悲しいかな、そんな当たり前が当たり前でないのが、まだまだ業界の常識、なのです。（住宅購入学入門*）

2.2 「かなしい」の意味分析：〈思いと異なる良くない事態に〉〈気持ちが沈む〉〈さま〉

(3) 世の中に、なにが悲しいといって、父母、兄弟、姉妹、そのほか親しい人たちを失うことぐらい悲しいものはありません。（＝(1)）

(4) 女の子にとって髪はすごく大切。どんどん抜けていく悲しい気持ち、わからないでしょう。（中日新聞　2008 年８月５日）

(5) 私は結婚相手を探そうとまじめに合コンに参加したのですが、いい加減な態度に接

して悲しい思いになりました。　（吉良友佑『お見合い１勝 99 敗』PHP 新書　p. 85）

例(3)は「親しい人たちを失う」こと、例(4)は大切な髪の毛が「どんどん抜けていく」ことに「かなしい」と感じている。よって、例(3)と(4)は、大切なものを失うことによって「かなしい」と感じていると言える。続いて例(5)は、合コンでまじめに結婚相手を探したいという話し手の思いと異なり、他の参加者が「いい加減な態度」であったことに「かなしい」と感じている。例(5)は、話し手の思いと異なる良くない事態に「かなしい」と感じていると言うことができる。例(3)と(4)の、大切なものを失うということも、失いたくないという話し手の思いに反して大切なものを失うことであると考えられるため、例(3)から(5)は、話し手の思いと異なる良くない事態に「かなしい」と感じていると考えられる。

(6) どうしようもなく悲しくて、落ちこんでしまったら、いのちの 110 番に電話で相談しよう。　（小野垣義男『心のけんこう』文芸社　p. 15）

(7) 私の幸せは突然に終わりを告げたんだ。悲しくて悲しくて何をしてもやる気が出てこなかった。彼のことを思うといつも涙が流れた。
（長原千代・小百合ロメイ『Friends』文芸社　p. 243）

例(6)に「悲しくて、落ち込んでしまったら」とあることから、「かなしい」は話し手の思いと異なる良くない事態に気持ちが沈むさまを表していると考えられる。また、例(7)の「悲しくて悲しくて何をしてもやる気が出てこなかった」というのは、話し手の思いと異なる良くない事態に気持ちが沈み、何かをしようという積極的な気持ちが生じない状態を表していると考えられる。よって例(6)と(7)から、「かなしい」は、話し手の思いと異なる良くない事態に気持ちが沈むさまを表すと言うことができる。

例(3)から(7)は話し手の身の上に起きたことに対して「かなしい」と感じているものであるが、次の例のように、他者の言動などに対して「かなしい」と感じているものもある。

(8) 物理的・経済的に耐用年数の長い優良な社会資産を築き、暮らしやすく文化度の高い街を形成していくなど、不動産・建設業界には、もっともっと、やるべきことがたくさんあるはずです。

私にいわれなくとも、こんなこと、個人レベルでは業界のだれもが気づいていること。

しかし悲しいかな、そんな当たり前が当たり前でないのが、まだまだ業界の常識、なのです。　（＝(2)）

(9) ラクをして大学に入る、ということしか眼中にない青春というのは、哀しいものだ。目標があれば、そのために努力する。それが自然な生き方だろう。何か目標があり、そのために大学に入りたいというのであれば、その分野の勉強をするのは当然のことだ。　（パパは塾長さん*）

例(8)は文中に「不動産・建設業界には、もっともっと、やるべきことがたくさんあるはずです」とあるように、話し手がやるべきであると考えていることを不動産・建設業界がやろうとしないことに「かなしい」と感じている。続いて例(9)は、話し手は「目標があれば、そのために努力する」ということは「当然のことだ」と思っているため、話し手の考え方とは異なる「ラクをして大学に入る」という考え方に「かなしい」と感じている。よって、例(8)と(9)の「かなしい」においても、話し手の思いと異なる良くない事態に気持ちが沈むさまを表していると言うことができる。例(8)と(9)は、先程見た例(3)から(7)のように話し手の身の上に生じたことに関して「かなしい」と感じているのではないため、気持ちの沈む程度が例(3)から(7)に比べて低いと考えられる。

(10) 彼らヤンキーの多くは決して性格が曲がっているわけではない。むしろ真っ直ぐ、哀しいほど真っ直ぐなゆえ、一般には「摩擦」と呼ばれる現実に直角に激突してしまうのだ。しかしいかんともしがたい現実に彼らの真っ直ぐは木っ端微塵に吹き飛ばされてしまう。そうして不器用なまでに現実と無謀な格闘を重ねていくにつれ、その美しい正中線は徐々に直線としての形状を保てなくなってしまうのだ。

(中林あきお『泣き虫男、歩いて日本一周してきます』枻出版社　p. 32)

(11) ルートヴィヒ二世が十九歳でバイエルン王に即位したとき、その輝くような美貌は、まさに地上に降り立った神のようだと讃えられた。その澄んだ瞳は哀しいまでに青く、目鼻立ちは女性のように整っていて、彼の乗った馬車が通ると、女たちは思わずうっとりと見とれたという。(中略) ワーグナーも彼を、「あまりに美しくて、夢のように消えてしまわぬかと心配だ。彼こそ私の幸運のすべて。彼がもし死ねば、私も次の瞬間に死ぬ」と書き記している。

(桐生操『世界史　怖くて不思議なお話』PHP 文庫　pp. 213-214)

例(10)は、ある人達の性格について「哀しいほど真っ直ぐ」であると表現している。話し手が好ましく感じている人達の性格があまりに真っ直ぐであるため、現実に臨機応変に対応できず、現実と「直角に」あるいは「不器用なまでに」格闘してしまい、その結果「木っ端微塵に吹き飛ばされてしまう」といった良くない事態となることを「かなしい」と感じていると考えられる。続いて例(11)は、ある人の瞳について「哀しいまでに青」いと表現している。文中に「あまりに美しくて、夢のように消えてしまわぬかと心配だ」とあるように、ある人の美しい瞳がこの世のものとは思えないくらい澄んだ青色であるため、夢のように消えてしまうといった良くないことが生じるのではないかと思い、気持ちが沈むさまを「かなしい」と表していると考えられる。例(10)と(11)のように、好ましく思っている人の身に良くないことが起きる、あるいは起きる可能性があるというのも、話し手の思いと異なる良くない事態であり、そのことに気持ちが沈むさまを「かなしい」と表していると言える。

以上から、「かなしい」の意味は〈思いと異なる良くない事態に〉〈気持ちが沈む〉〈さま〉と記述することができる。

３.「つらい」の意味分析

次に「つらい」の意味を分析する。

3.1　先行研究の記述とその検討

森田（1977：308）は「つらい」の意味を次のように記述している。

> つらい〔辛い〕
> ある状況に置かれて、または、ある事が原因して精神的に耐えられないほど苦痛を感じる状態。

上の意味記述に加え、次のように説明している（森田 1977：308-309)。

「咳が出てつらい」「ずっと立ちっ放しでつらい」のように生理的、身体的な原因もあるが、多くは「つらい仕打ち」「つらく当たる」「部下の首を切るのはつらい」のように精神的なむごさ・悩みに由来する当人の苦悶状態に言う。「立ちっ放しでつらい」も足の苦しさを言うのではなく、そのような状況に置かれた当人の肉体的、精神的ストレスに対する苦しみの感情である。したがって、身体の部分を限定した「足がつらい」「胸がつらい」などの言い方はできない。

続いて、『現代形容詞用法辞典』(1991：368-369）は次のように記述している。

> つらい〔辛い〕
> ① 精神的に苦痛を感じる様子を表す。
> 「かわいがって育てた子を手放すのはつらい」
> ② 冷酷で思いやりのない様子を表す。
> 「彼女は息子の嫁につらく当たった」「彼は世間のつらい仕打ちに耐えて育った」

また②の意味について、②の意味で用いられる時は「つらい」が動詞にかかる修飾語（「つらく当たる」)、または名詞にかかる修飾語（「つらい仕打ち」）として用いられるのが普通で、述語になることは少なく、その場合にはふつう①の意味になると指摘している（『現代形容詞用法辞典』 1991：368)。

森田と『現代形容詞用法辞典』の記述から、「つらい」は苦痛を感じる場合に用いられることが分かる。また、森田は身体の部分を限定した「足がつらい」などの言い方はできないと指摘している。確かに、足に関して「つらい」という場合には、「疲労がたまり、足が張ってつらい」「足がしびれてつらい」のように、足そのものに対してというよりも、足にかかる負担によって感じる精神的ストレスを表していると考えられる。

さらに『現代形容詞用法辞典』は、「つらく当たる」あるいは「つらい仕打ち」のように、「つらい」が動詞あるいは名詞にかかる修飾語として用いられた時、「冷酷で思いやりのない様子を表す」と記述し、「つらい」のもう一つの意味（＝精神的に苦痛を感じる様子を表す）と区別している。一方、森田は「つらく当たる」「つらい仕打ち」のような例も、「部下の首を切るのはつらい」といった例と区別せず、「ある状況に置かれて、または、ある事が原因して精神的に耐えられないほど苦痛を感じる状態」であると記述している。この点に関して次の二つの例を見てみよう。

(12) （前略）女はこの泥亀と関係を持つようになった。そうなると梅吉に辛く当たるようになり、毎日梅吉を役立たず呼ばわりした。
（石川鴻斎『夜窓鬼談』春風社　p. 411）

(13) 学校から帰って夕方暗くなるまで働かねばならないのはつらかったが、学校で喧嘩したとき「小日本」「日本に帰れ……」と罵られるのは余計つらかった。
（井出孫六『終わりなき旅』岩波現代文庫　p. 305）

例(12)の「辛く当たる」は、「役立たず呼ばわり」するというように、梅吉が「つらい」と感じるような態度を女がとることを表している。続いて例(13)は、学校で同級生に「『小日本』『日本に帰れ……』と罵られる」ことを「つらい」と感じている。例(12)の「つらい」は例(13)のように、話し手が「つらい」と感じるような態度を相手がとることを表しているため、本研究では、「つらい」が動詞や名詞にかかる修飾語となった場合と、述語になった場合を区別しないこととする。

以上の先行研究の記述とその検討をふまえ、以下で「つらい」の意味を分析する。

3.2　「つらい」の意味分析

「つらい」に二つの多義的別義を認め、分析の最後に別義間の関連性について示す。

3.2.1　別義１：〈身体に負担がかかり〉〈耐えられないと感じる〉〈さま〉

(14) 自分で育てたカクメロを収穫した青山真一君（18）は「暑い温室内での作業はつらかったが、思ったより真四角な出来栄えでうれしい」と喜んでいた。
（中日新聞　2010 年 7 月 6 日）

(15) やるぞ、と意気込んだのはいいが、苗を植える前かがみの姿勢がつらい。長靴が泥にはまり込み、歩くのも一苦労だ。　（中日新聞　2007 年 12 月 25 日）

(16) 花粉症で目がかゆくてつらいです。　（Yahoo!知恵袋*）

例(14)は室温が高い温室内で作業すること、例(15)は苗を植えるために「前かがみの姿勢」をとること、例(16)は目がかゆいことに「つらい」と感じている。例(14)から(16)は、

身体に負担がかかることによって「つらい」と感じていると言うことができる。

次に、「つらい」とは、身体に負担がかかることでどのように感じる感情であるのかについて、次の例を見てみよう。

(17) あなたが初めてココへ来た時、頭痛の発作で倒れましたがあの様子は異常でした。辛くて耐えられない程の痛みだったんでしょうね。
(尾崎晃『癒しの人』文芸社　p. 55)

(18) 教会の中に入ると、右手の階段を上った所に、礼拝堂がありました。そこは、キリストが磔にされた場所と言われています。中にはたくさんのろうそくが灯され、その中央に磔にされたキリストの像がありました。そこに入った時でした。私は、吐き気と圧迫感に襲われ、泣き出したいぐらい体が辛くなってしまって、どうしてもそこにいることができませんでした。　(うたかたの月*)

例(17)は「辛くて耐えられない程の痛み」、例(18)は「体が辛くなってしまって、どうしてもそこにいることができませんでした」とあることから、「つらい」は、身体に負担がかかり、耐えられないと感じるさまであると考えられる。

以上から、「つらい」の別義１は〈身体に負担がかかり〉〈耐えられないと感じる〉〈さま〉と言うことができる。

3.2.2　別義２：〈望みと異なる良くない事態に〉〈耐えられないと感じる〉〈さま〉

(19) 布団に入ると、いつものようにおっぱいの元に寄って来る。「ごめんね。でも今日はお薬飲んだし、あかんねん」。頭をなでながらおっぱいを隠すと、悲壮な顔つきに。抱っこしても、落ちそうなほど暴れる。あげたい。でもあげられない。こっちもつらくて涙がボロボロ。　(京都新聞　2007年４月12日)

(20) (前略) マザーは貧しい人に尽くすというこの活動を、最初はたった一人で始めたと聞く。彼女は普通の人より愛の量が多すぎて、世の中に貧しさで苦しんでいる人がいることが辛くて耐えられなかったのだろう。　(インド*)

(21) 独特の言葉遣いや決まり事が多くて、お座敷での会話も先輩芸妓のように滑らかにできず、「つらくて何度も辞めたいと思った」　(読売新聞　2011年５月１日)

まず例(19)は、文中に「あげたい。でもあげられない」とあるように、「あげたい」という気持ちに反してあげることができないことに「つらい」と感じている。続いて例(20)は、「世の中に貧しさで苦しんでいる人がいる」というのは、話し手の望みと異なる良くない事態であり、そのような事態に「つらい」と感じている。さらに例(21)は、「先輩芸妓のように滑らかに」会話したいという望みに反し、滑らかに会話できず、「つらい」と感じている。よって、例(19)から(21)は、望みと異なる良くない事態に「つらい」と感じていると

言うことができる。さらに、例(20)は「辛くて耐えられなかった」、例(21)は「つらくて何度も辞めたいと思った」とあることから、「つらい」は、望みと異なる良くない事態に耐えられないと感じるさまを表すと考えられる。

以上から、「つらい」の別義２は〈望みと異なる良くない事態に〉〈耐えられないと感じる〉〈さま〉である。

3.2.3　別義間の関連性について

「つらい」の二つの意味の関連性について考察する。別義１（＝〈身体に負担がかかり〉〈耐えられないと感じる〉〈さま〉）では身体的な負担によって耐えられないと感じているが、別義２（＝〈望みと異なる良くない事態に〉〈耐えられないと感じる〉〈さま〉）では望みと異なる良くない事態という、心理的な負担によって耐えられないと感じている。よって、別義１と２は、〈負担によって〉〈耐えられないと感じる〉〈さま〉という共通の意味を有している。また、別義１から別義２は、身体を通じた経験から、より心理的な経験へと意味が拡張していると考えられるため、別義２は別義１からメタファーによって成り立っていると言うことができる。

４.「つらい」と「かなしい」の類似点と相違点について

分析結果を基に「つらい」と「かなしい」の類似点および相違点について考察する。

4.1　先行研究の記述とその検討

『現代形容詞用法辞典』(1991: 154-155) は、「つらい」と「かなしい」の意味について、「精神的な苦痛を表す意味で『かなしい』は『つらい』に似ているが、『つらい』は意味の範囲が広く、さまざまな感情においてたえがたいという意味を表すのに対して、『かなしい』は悲哀に限定される点が異なる」と述べている。

上の『現代形容詞用法辞典』の記述から、「つらい」と「かなしい」は「精神的な苦痛を表す」という点で意味が類似していることが分かる。また両語の意味の違いについては、「つらい」は「意味の範囲が広く、さまざまな感情においてたえがたいという意味を表す」のに対して、「かなしい」は「悲哀に限定される」点が異なると述べている。確かに、次の例のように、身体に負担がかかることによって苦痛を感じることを「つらい」と表現することはできるが、「つらい」を「かなしい」に置き換えると不自然な表現となる。

(22) 花粉症で目がかゆくてつらい（??かなしい）です。　　（＝(16)）

『現代形容詞用法辞典』の記述により、「かなしい」は「悲哀」を表すのに対し、「つらい」は「かなしい」よりも「意味の範囲が広く、さまざまな感情においてたえがたいという意味を表す」という違いがあることを確認した。以下では、その他の相違点について考察する。

4.2 「つらい」と「かなしい」の類似点と相違点について

(23) 「女の子にとって髪はすごく大切。どんどん抜けていく悲しい（つらい）気持ち、わからないでしょう。」 （＝(4)）

例(23)は、大切な髪が抜けていくことに「かなしい」と感じている。例(23)の「かなしい」を「つらい」に置き換えてもその語を含む文の意味が大きく異ならない。ここでの「つらい」は、別義２（＝〈望みと異なる良くない事態に〉〈耐えられないと感じる〉〈さま〉）を表すことから、「かなしい」と「つらい」の別義２に互換性があると考えられる。また、それらの共通の意味は、〈思いと異なる良くない事態に〉〈精神的苦痛を感じる〉〈さま〉と言うことができる。

次に、「かなしい」と「つらい」の別義２の相違点について考察する。

(24) 政治家と親密な関係を持たない公務員にとって、自分の業績を政治家に見落とされることはとても辛い（??かなしい）。政治家に評価されなければ、せっかくの仕事が昇進のポイントにならないからだ。だから通常、平均的公務員は、自分の職務に関わっている政治家の利益に反するような行為は避けたがる。

（巨大市場インドのすべて*）

(25) 一時とはいえ母親代わりとして育ててきた子供と別れるのは、きっと身を切られるように辛い（??かなしい）ことだったに違いない。

（後藤真理子『マリベル―マヤの国から来た天使』文芸社　p. 7）

例(24)は「自分の業績を政治家に見落とされる」こと、例(25)は「子供と別れる」ことを「つらい」と感じているが、例(24)と(25)の「つらい」を「かなしい」に置き換えると不自然な表現となる。例(24)では、「つらい」と感じる理由は「政治家に評価されなければ、せっかくの仕事が昇進のポイントにならないからだ」とあるように、話し手の望みと異なる良くない事態によって、話し手が大きなダメージを受けるため、耐えられないと感じている。また例(25)では、「子供と別れる」ことは「身を切られるように辛い」とあるように、子供と別れることにより、「身を切られる」ほどの精神的ダメージを受けるため、耐えられないと感じるさまを表している。例(24)と(25)から、「つらい」の別義２は、大きな精神的ダメージを受けることによって、耐えられないと感じるさまを表すと言うことができる。

一方の「かなしい」は、気持ちが沈むさまを表すのであって、大きな精神的ダメージを受けることにより、耐えられないと感じるさまを表すのではない。この点について次の例を見てみよう。

(26) 八月の白山市内での交通死亡事故は二件。いずれも 157 号だ。悲しい（??つらい）

事故が続いている。　（中日新聞　2009 年９月４日）

(27) 私はちょっぴり哀しく（??つらく）なった。室さんと私は、所詮そういう宿命のめぐりあわせでしかないのか。　（演歌の虫*）

例(26)は、他人が交通事故で死亡したことに「かなしい」と感じている。例(26)のように、自身にはあまり関係のない悲劇に対して「かなしい」と表現することはできるが、「かなしい」を「つらい」に置き換えると不自然な表現となる。また、例(27)のように、「つらい」が、程度が低いさまを表す「ちょっぴり」という副詞と共起すると不自然な表現となるのも、「つらい」が表す精神的苦痛が大変強いものであるからであると考えられる。

以上から、「かなしい」と「つらい」の別義２が類似しており、〈思いと異なる良くない事態に〉〈精神的苦痛を感じる〉〈さま〉という共通の意味を有するが、「つらい」が表す精神的苦痛というのは耐えられないというように、大変強いものであるのに対し、「かなしい」が表す精神的苦痛は「つらい」のように強いものではないという違いがあると言える。

５．まとめ

本研究では「かなしい」と「つらい」の意味を次のように記述した。

「かなしい」の意味:　〈思いと異なる良くない事態に〉〈気持ちが沈む〉〈さま〉
「つらい」の意味
　別義１:　〈身体に負担がかかり〉〈耐えられないと感じる〉〈さま〉
　別義２:　〈望みと異なる良くない事態に〉〈耐えられないと感じる〉〈さま〉

また、「かなしい」と「つらい」の意味の類似点と相違点ついて、「かなしい」と「つらい」の別義２が類似しており、〈思いと異なる良くない事態に〉〈精神的苦痛を感じる〉〈さま〉という共通の意味を有するが、「つらい」が表す精神的苦痛というのは耐えられないというように、大変強いものであるのに対し、「かなしい」が表す精神的苦痛は「つらい」のように強いものではないという違いがあると述べた。

文献

遠藤織絵、小林賢次、三井昭子、他（編）(2003)『使い方の分かる類語例解辞典』、小学館
金田一春彦、池田弥三郎（編）(1988)『学研国語大辞典（第二版)』、学習研究社
飛田良文、浅田秀子（1991)『現代形容詞用法辞典』、東京堂出版
森田良行（1977)『基礎日本語Ｉ』、角川書店

現代日本語におけるコロケーション:検出と分析

STRAFELLA Elga Laura（奈良先端科学技術大学院大学）[†1]
林部 祐太（奈良先端科学技術大学院大学）[†2]
松本 裕治（奈良先端科学技術大学院大学）[†3]

Detection and Analysis of Collocations in Contemporary Japanese

Elga Laura Strafella (Nara Institute of Science and Technology)
Yuta Hayashibe (Nara Institute of Science and Technology)
Yuji Matsumoto (Nara Institute of Science and Technology)

1 はじめに

本研究ではコーパスから現代日本語におけるコロケーションを検出し、それらの構文パターンと意味を分析することを目標としている。本稿では、まずコロケーションの定義を行い、「制限コロケーション」・「比喩的イディオム」・「真性イディオム」の違いを簡単に検討する（2 章）。次に、コロケーション検出の研究でよく用いられる代表的な指標を取り上げ、その特徴を述べる（3 章）。そして、それらの違いを定量的に区別するために我々が行ったアンケート調査について説明し（4 章）、アンケート結果を分析する（5 章）。最後に、まとめと今後の方針について述べる（6 章）。

2 コロケーションとは

「コロケーション」とは「連結語句」「連語」「語の配列」なども言われるが、要するに「ある単語と単語のよく使われる組み合わせ」のことである。例えば、“辞書” という単語では、「辞書を引く」「辞書で調べる」「分厚い辞書」とは言えるが,「辞書を読む」「太い辞書」とは通常言わない。このような自然な単語の組み合わせは、そのまま覚える必要があり、日本語だけでなく、全ての自然言語における共起現象の一つである。外国語学習者はできるだけ “正しく、自然な” 言語を話すために, その言語の “正しく、自然な” 組み合わせを使わなければならない。コロケーションは全体的な意味が各単語の意味の組み合わせと比喩的に違ってくることが多く、母国語を参考したり、文字通りに翻訳したりすると “奇妙な表現” になる可能性が高い。

本研究ではコロケーションを単語の結び付きの強さに応じて「制限コロケーション」・「比喩的イディオム」・「真性イディオム」の 3 段階で区別する.

2.1 制限コロケーション

制限コロケーションとは、「傘をさす」のように単語の自然な組み合わせのことである。「傘を開く」や「傘を開ける」は「傘をさす」と意味的には、ほとんど変わらないが、母語話者が聞くと、不自然な表現である。ただし、イディオムとは異なり、その組み合わせから特別な意味は生じない。

[†1]elga-s@is.naist.jp
[†2]yuta-h@is.naist.jp
[†3]matsu@is.naist.jp

表 1: 統計指標の例

フレーズ	頻度	相互情報量	T スコア	ダイス係数
腹が立つ	285509	4.71668	529.552	0.170419
歯が立つ	54035	3.13341	222.328	0.0332873
足が出る	19443	-1.44844	-454.072	0.00172974
目が出る	14238	-2.69391	-1645.38	0.000943862
腕が立つ	4845	0.216424	13.5477	0.00239027
首が据わる	4167	4.3758	63.7468	0.00351143
目が据わる	3727	2.57955	56.4299	0.000587353
背が立つ	729	-1.1192	-55.6667	0.000458049
胴が据わる	8	1.55885	2.47487	0.000161713

2.2 比喩的イディオムと真性イディオム

二語以上の単語が固く結びつき、それぞれの単語とは異なる意味を持つものをイディオムとよぶ。ただし、ある程度元々の単語から全体の意味が推測できる場合とそうでない場合がある。

個々の語の意味から構成的に理解できないものを「真性イディオム」と呼ぶ。たとえば、「腹が立つ」や「腹を立てる」は、両方とも “怒る” という意味を指し、それぞれの単語の意味を知っていても、全体の意味を推測できない。このように単語が固く結びついた表現は, 日本語学習者にとって一つずつ覚えるしかない。

一方ある程度元々の単語から全体の意味が推測できる場合、「比喩的イディオム」と呼ぶ。たとえば、「足を伸ばす」とは、2 つの意味を持ち、一つは「足をうーんと伸ばす」という物理的な動作を指し、もう一つは個々の単語の意味から少し離れた「遠出する」という比喩的な意味である。このような結合句は「制限コロケーション」と「真性イディオム」との中間の段階にある。

3 よく用いられる統計指標

最近の言語コーパス研究では、複数の統計指標に基づく共起語検出に対応しているものが多い。しかし、どの統計値を用いれば良いのかは, はっきりとした結論は得られていない。本稿では、一般にコロケーション研究に広く用いられる, 単純頻度・ダイス係数・MI スコア・T スコアという 4 種類の指標を紹介する。アンケート（後述）で使用したフレーズを用いて、表 1 に各指標値の例を示す。

以下, 中心語 A の頻度を f_A, 共起語 B の頻度を f_B, コーパスでの総語数を W と表記する.

3.1 単純頻度

コロケーションを抽出する上で、もっとも単純な指標は、単純頻度である。しかし、より詳細な分析を行おうとすると、単純な頻度だけで判断するのは危険である。もともと頻度の低い語であれば、その語との共起頻度も自然に低くなり, 逆に、もともと多く出現する語であれば、その語との共起頻度が高くなる.

そのため、中心語と共起語の結合形の単純頻度 (共起頻度) だけでなく、 複数の指標を組み合わせて、それぞれが示す共起度を比べる必要がある。なお, 以下共起頻度を f_{AB} と表記

する.

3.2 ダイス係数

ダイス係数は、中心語頻度と共起語頻度の関係だけで 2 語のコロケーション強度を計測する尺度である。共起頻度を中心語頻度と共起語頻度の和で割って 2 倍した値である。式は次のようになる。

$$D = 2 \times \frac{f_{AB}}{f_A + f_B}$$

3.3 相互情報量

相互情報量は、ある語と共起語の統計的な独立性を示す指標である。ただし、頻度が低い語について敏感に反応し過ぎるという欠点をもっている。式は次のようになる。

$$I = \log_2 \frac{f_{AB} \times W}{f_A \times f_B}$$

3.4 T スコア

T スコアは、2 つの語の共起関係の統計的有意性を図り、共起の程度が偶然による確率を超えていると、どのくらいの確かさで言えるかを示す指標である。式は次のようになる。

$$T = \frac{\left(f_{AB} - \frac{f_A \times f_B}{W}\right)}{\sqrt{f_{AB}}}$$

4 アンケート調査

コロケーション研究においてよく使用される何種類の指標のうち、どの指標が有効であるかを調べるためにアンケート調査を行った。

4.1 アンケートの概要

アンケートには「身体名詞-[がをに]-動詞」の組のうち、比較的共起頻度の高いフレーズ 78 個を用いた。頻度は「日本語係り受けコーパス」(JDC) より計算した。JDC とは、約 1 億ウェブページからなる日本語ウェブコーパス 2010(NWC2010) より日本語係り受け解析システム CaboCha を用いて、助詞を介した語と語の係り受けを抽出したコーパスである。

アンケートは、当大学の日本語母語話者 (21 人）と留学生 (日本語能力試験の 1・2 級の 15 人) を対象に実施した。

4.2 母語話者に対する質問

母語話者には、各フレーズの名詞と動詞の結びつきの強さを、自由結合（コロケーションではないフレーズ)・コロケーション・比喩的イディオム・真性イディオムの 4 段階のうち、ど

表 2:「足が出る」に対する母語話者の回答の例

語義	DIC	POW	FREQ	ユーザ番号
0	4	2	2	8
1	2	4	2	8
2	1	4	1	8

表 3: 留学生に対するアンケートから得られたデータの例

フレーズ	意味 1	意味 2	意味 3	名詞の意味	動詞の意味	推測した意味	ユーザ番号
腹が立つ	to get angry			stomach	to stand		2
歯が立たない	be not able to do	too difficult to realize or to do	too hard to chew	teeth	to stand (negative)		60
背が立つ				back	to stand	to get higher	2

の段階に有るのかを回答してもらった。そして、与えられたフレーズの意味を思い付いた順に記入してもらった。

「足が出る」という 2 つの意味を持つフレーズについて母語話者から得られた回答の一部を表 2 に示す。各項目は、それぞれ次の質問に対する回答である。

- **DIC**: 辞書を使わなくても, この意味が出てきましたか?
  1:辞書を引いて初めて意味を知った、2:辞書を引くと意味を思い出した、3:辞書を引かなくても分かった、4:辞書には載っていない意味だった
- **POW**: 名詞と動詞の結びつきの強さはどのくらいですか?どうしても分からない場合は「不明」を選択して下さい.
  2:自由結合、3:コロケーション、4:比喩的イディオム、5:真性イディオム、-2:不明
- **FREQ**: このフレーズが 100 回使われたとして, この意味を表現するために, 使われるのは概ね何%くらいだと思いますか?
  (注 1) このフレーズがよく使われるかどうかという質問ではありません.
  (注 2) このフレーズには意味が 1 つしか無いならば,80%〜を選択してください.
  1:〜29%、2:30%〜、3:50%〜、4:80%〜、

4.3 留学生に対する質問

留学生には、各フレーズの名詞と動詞の意味について答えてもらい、フレーズ全体の意味が分からない場合には、それを推測できるかどうか尋ねた。

表 3 は留学生に対するアンケートから得られたデータの例である。まず、与えられたフレーズに対して最大 3 つまで思いつく意味を列挙してもらい、動詞と名詞の意味も答えてもらった。与えられたフレーズの意味が分からなかった場合、フレーズの意味を推測してもらった。

5 アンケートの結果

本稿では、アンケートに用いた 78 個のフレーズのうち表 1 に示した 9 個のフレーズを使って分析する。

表 4: 複数の意味を持つフレーズに対する母語話者の結び付きの強さの判断の例

フレーズ	意味	真性イディオムと答えた人数	比喩的イディオムと答えた人数
足が出る	予算を超えた支出になる	11	7
	隠しごとが現れる	7	10
目が出る	目玉が飛び出る	3	14
	幸運が巡ってくる	2	13
歯が立たない	固くて噛むことができない	1	7
	自分の力が弱くて対抗や理解ができない	12	10

5.1 一致率と共起頻度

共起頻度の高いフレーズは、母語話者も留学生も回答は似通った。例えば,「腹が立つ」は、母語話者は全員が正確な意味を答え 21 人中 18 人が「真性イディオム」と答えた。一方、共起頻度の低いフレーズは母語話者であってもその意味があまり理解されておらず、コロケーションの強さの判断にばらつきがあった。例えば,「背が立つ」は母語話者は 20 人が辞書を引かないと意味が分からず、8 人が「比喩的イディオム」と答え、9 人が「真性イディオム」と答えた。そのため、回答者の一致率と共起頻度には相関関係があると考えられる。

5.2 複数の意味を持つフレーズ

フレーズには複数の意味を持つものがあった。例えば,「足が出る」には、「布団から足が出る」のように文字通りの意味の他に、「予算を超えた支出になる」や「隠しごとが現れる」という意味がある。その場合は、意味ごとにコロケーションの強さを判断してもらった。その結果の一部を表 4 にまとめた。

表から分かる通り、コロケーションの強さの判断は意味ごとに異なった。しかしながら、3 章で挙げた指標では、意味ごとの分析ができない。

6 まとめ

本稿では、自由結合・コロケーション・イディオムなどの概念を整理して、共起度を測る代表的な指標としてよく使われる単純頻度・ダイス係数・相互情報量・T スコアの 4 つの指標について説明した。

アンケートの結果、共起頻度が極めて高いフレーズは単語間の結び付きの強さの判断の一致率が高かったが、そうでないフレーズは回答者間で判断が割れた。また、複数の意味を持つフレーズは意味ごとに結び付きの強さは異なることがあり、その場合は前述した指標では分析できないことが分かった。

今後は共起に基づく指標だけではなく、フレーズを構成する単語の置き換えに基づく指標を用いることを考えている。コロケーションには、前述したように似たような意味を持つ単語で置き換えると不自然な表現になるという性質がある。そのため、日本語のシソーラス『分類語彙表』を使って、各名詞が共起する動詞の synset を作り、その動詞を同義語と置き換えると、頻度などがどのように変化するか等を探っていきたい。

文献

石川慎一郎 (2006)「言語コーパスからのコロケーション検出の手法：基礎的統計値について」統計数理研究所共同研究レポート 190 巻, pp.225-243

Evert, S. (2004) *The Statistics of Word Co-occurrences: Word Pairs and Collocations*, Ph.D. thesis, University of Stuttgart

金田一秀穂 (2006)『知っておきたい日本語コロケーション辞典：Japanese Collocation Dictionary』学研

庄司香久子 (2010)『日本語言葉のコンビネーション・ハンドブック』（英文版）、講談社

姫野昌子 (2004)『日本語表現活用辞典』、研究社

Seretan, V. (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

国立国語研究所 (2004) 『分類語彙表』、国立国語研究所資料集 14 巻

関連 URL

日本語ウェブコーパス 2010 http://s-yata.jp/corpus/nwc2010/

日本語係り受けコーパス http://hayashibe.jp/jdc/

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

姜　紅（北京外国語大学日本学研究センター・北京第二外国語学院日本語学院）

The Semantic Classification of “Amai” and Its Semantic Distribution Based on Corpus

JIANG Hong（Beijing Foreign Studies University, Beijing International Studies University）

1．はじめに

日本語の形容詞「甘い」は、味覚の基本義以外にも様々な拡張義を持っている。本稿では、2011 年 8 月に国立国語研究所によって公開された検索アプリケーション「中納言」[1]を利用し、「現代日本語書き言葉均衡コーパス」（以下BCCWJと略称）から「甘い」のKWICデータを収集・分類し、従来辞書で見落とされていた意味用法や拡張された用法が存在するかどうかを検証する。さらに、日本語と中国語との違いを意識しながら、外国人日本語学習者の視点から形容詞「甘い」の意味の再分類を試みる。

また、日本語の形容詞「甘い」を対象に、BCCWJ におけるジャンル別の頻度調査を行い、テキストタイプによって、形容詞「甘い」が中心的に担っている機能に違いがあるかどうかを考察する。さらに、叙述用法・連体用法・連用用法という 3 つの用法ごとに、多義語「甘い」の語義分布を調べ、形容詞の文法的機能による語義分布及び意味用法の違いを明らかにする。

2．辞書に見られる「甘い」の意味記述と問題提起

筆者は『日本国語大辞典』『学研国語大辞典』『例解新国語辞典』『デジタル大辞泉』（以下、それぞれ『国語』『学研』『例解』『大辞泉』と略称）の 4 つの辞典を利用し、「甘い」の意味項目を調べてみた。上記の 4 つの国語辞書の意味記述には、次のような特徴がある。

1. 辞書によって、意味立項の配列は多少違っているが、「砂糖や蜜など糖分の味」という語義を第一義として挙げている点では一致している。
2. 各辞書による「甘い」の意味規定から、日本語の味覚形容詞「甘い」が本来味覚を表す語であるが、味覚以外の感覚、物事の状態、人間活動を表すなど、多くの語義を持っていることが分かる。

外国人学習者は日本語の語彙を習得する際に、日本語辞書の助けに頼ることが多い。とりわけ日本語の国語辞書の意味記述は、語彙の意味用法を調べ、理解する上で大きな助けになる。だが、辞書の意味記述は語彙のすべての意味用法を網羅しているわけでもない。例えば、次の「甘い」の使用例を見てみよう。

（1）その変幻自在な歌声と甘いマスクで、女性ファンを獲得しました。
（朝日新聞 2011.4.21）

（2）秋田県の米を与えて子牛を肥育しているのが特徴で、サシ（霜降り）は甘く、とろけるよう。（朝日新聞 2011.5.5）

[1] 利用するには書面による申請が必要である。詳細につき、https://chunagon.ninjal.ac.jp/login を参照されたい。

（3）現実は甘くなかった。人材紹介会社に登録に行くと、まず「年齢がネック」だと言われた。（朝日新聞 2011.4.23）

上記の用例は日本の新聞記事から引いた実際の使用例である。しかし、例（1）と例（2）における「甘い」の意味用法は、上の 4 つの国語辞書の意味項目に該当するものが見つけられない。例（1）「甘いマスク」は、主に男性に使われ、女性から見て甘さを感じるような優しい顔立ちだと言われる。中国語には“甜美的长相”、“甜甜的笑容”などのように、人の容貌について言う表現もあるが、ほとんど女性にしか使わない。この点において、日本語の「甘い」と大きく異なっている。例（2）は、「甘い」が美味の意味で使われ、サシや魚介類などが新鮮で良質な味をするという意味を表す。このような用法は出現頻度が極めて少なく外国人学習者にとっては理解がしにくい。

また、例（3）の「甘い」は打ち消し文に使われ、物事が簡単ではなく軽くみてはいけないという意味を表わす。この意味用法は『学研』に出ている「甘い」の意味項目「大したものではない」という意味に似ているが、『学研』以外の 3 つの国語辞典には似たような語義項目が見当たらない。BCCWJ で調べた結果、「甘い」のこの意味用法は打ち消し文に多く用いられることがわかる。このことは、語義の存在が文法的な形に関わっているということを示唆している。このように、多義語の語彙をより全面的に、より深く理解するためには、辞書の意味記述に頼るだけではなく、言語の使用実態をよく反映する大量の用例を調べることが大切である。

外国語学習には、語彙の語義理解だけでは十分ではなく、語彙の運用能力を高めることも重要な課題である。砂川（2010）が指摘するように、「いくら多くの単語を覚えたとしても、その使い方を知らなければ学習の意味はない」（砂川 2010：106）。特に、第二言語の習得には、母語干渉による誤用が生じやすい。日本語の国語辞書『大辞泉』には、「話しぶりが巧みで、人をたぶらかすさま。うまい。」という「甘い」の意味項目があり、『例解』を除いたほかの 2 つの辞書にも、これと似たような意味が出ている。また、用例として「甘い言葉で誘う」という表現が挙げられている。中国語には“甜言蜜语”という表現があり、「甘い言葉」に近い意味を表す。この点において、日本語の「あまい」は中国語の“甜”とは意味が共通している。また、中国語では、口がうまいことを“嘴甜”というフレーズで表現できる。そこで、中国人日本語学習者は、母語の影響を受け、中国語の“嘴甜”に対応する日本語が「*口が甘い*」のような表現ではないかと勘違いしかねない。このように、母語の干渉を最小限に抑えるためには、母語との相違を意識しながら単語の語義をより細かく記述することが必要である。

われわれ外国人研究者が直面する課題の 1 つは、外国人日本語学習者への配慮という視点から、語義のより一層深い分析に努めることである。日本語の国語辞書による意味記述をもとに、整備された大規模な日本語コーパスを活用することは、多義語の語義に関する理解を深めるための有効な方法だと考えられる。

3. 多義語「甘い」の意味用法の再分類

ここでは、BCCWJ から「甘い」の実例を収集・分析することによって、多義語「甘い」の意味用法を再検討する。ただし、コーパスにも制限があるため、BCCWJ に対する調査で「甘い」の全ての意味用法をカバーできるわけでもない。本研究では、国語辞書や先行研究に見られる「甘い」の意味記述とコーパスによる検索結果を補いあいながら、「甘い」の意味用法を整理・分析する。紙幅の制約上、「甘い」の意味用法の詳細に関する記述は省略

する。再整理した多義語「甘い」の意味用法を以下のように示す。

表 1　多義語「甘い」の意味用法

Ⅰ[身体体験]
1. 味覚
（1）[味：甘味] 糖分があるような味（基本義）……①
（2）[味：旨味] コクがあって良質な味……②
（3）[味：塩味・刺激性] 塩気や辛味が薄い味……③
2. 嗅覚　[匂い：香り] 物の香り・匂いなどが芳醇で快い……④
3. 聴覚　[声・音：歌声・音楽] 音楽や人の歌声・声が心地よい……⑤
4. 視覚　[外見：容貌（男性）]　（男性が）容貌が美しく好感を持たせる……⑥
[外見：服装など] 可愛くて女らしい……⑦

Ⅱ[物事の状態]
1. 物事の機能・品質などに不備があり、不充分な状態である……⑧
2. 物事の状況・程度が満足できない・中途半端な状態……⑨

Ⅲ[人間活動]
1.[精神的行為の生産物] 法律・基準・規定などが厳格ではない……⑩
2.[思考・行為] 人の思考・判断・行為が慎重さや厳密さが欠如する……⑪
3.[態度・接し方] 相手に対する態度や言動が優しい或いは厳しくない……⑫
4.[思考・行為の対象]（打ち消しの形で）物事や相手は簡単で、単純なものではない…⑬
5.[愛情・幸福]愛情や幸福感などがあふれて、うっとりと快い……⑭
6.[誘惑] 人の心を引き付けて迷わせる……⑮

Ⅳ[慣用表現]
1.「甘く見る」: 相手を見くだし、物事を軽く見る……⑯
2.「甘い汁を吸う」: 他人を利用し苦労もせずに利益を得る……⑰
3.「酸いも甘いも噛み分ける」: 世間の事情によく通じている……⑱

以上、「甘い」が慣用表現として使われる場合を除いて、主にⅠ．身体体験、Ⅱ．物事の状態、Ⅲ．人間の活動という 3 つの意味領域にわたり、コーパスや辞書などの用例をもとに「甘い」の意味用法を分類した。「甘い」は基本的には、知覚される外部世界の認知対象の味という属性を表す表現であると同時に、外部世界に対する認知主体の身体的、感性的体験でもある。そのため、「甘い」は食料や植物などのような外部世界の物理的な対象だけでなく、抽象的関係、人間主体や人間活動などに関するものの属性・特性についても使われている。つまり、「甘い」の意味用法は外部世界の状況に関する側面から、認知主体に関する側面へと意味が拡張していることが窺える。

4．形容詞「甘い」の文法的機能

形容詞の文中での機能については、すでに多くの研究がある。鈴木（1972）、西尾（1972）と高橋（1998）は、述語的な用法と比べて、規定語的な用法つまり連体修飾用法の方が、形容詞の主な用法だとみている。また、八亀（2007：61）は、形容詞の文中の機能について、「述語になる」と「規定語になる」という 2 つが中心となるが、テクストタイプによっ

て形容詞の中心的な機能が異なると指摘する。本研究では、BCCWJ におけるジャンル別の頻度調査を行い、テクストタイプによって、形容詞「甘い」が中心的に担っている機能には違いがあるかどうかを考察する。

4.1 調査方法とデータ

本研究では、2011 年 8 月に国立国語研究所によって公開された検索アプリケーション「中納言」を利用し、BCCWJから「甘い」のKWICデータを収集し、調査を行う。「中納言」とは、BCCWJをオンライン検索できる新しいツールであり、短単位・長単位・文字列の 3 つの方法による検索ができるというのが特徴である。2011 年 3 月現在、同コーパスは 11 種類のデータ、合計約 1 億 480 万語からなるという（KOTONOHA「現代日本語書き言葉均衡コーパス」検索デモンストレーションサイト[2]による）。

具体的な調査手順は以下の通りである。

1. 形容詞「甘い」の活用形も考慮に入れ、検索条件を「語彙素が「甘い」」と設定する。
2. 各ジャンル（韻文を除外し、10 種類のジャンル）毎に上記の条件で検索を行い、「甘い」の KWIC データを収集した。データ抽出後に、目視による確認作業を行い、「甘い」の形容詞としての用法のみを絞り出し、データを作成した。
3. 形容詞の文中での機能に基づき、叙述用法・連体用法・連用用法という 3 つの用法別に、2 によって得られたデータを分類した。

4.2 調査結果

以下、まずジャンル毎に検索された各用法の分布を示し、次にジャンルを問わず BCCWJ 全体において「甘い」の各用法の使用頻度を示す。

表 2　各ジャンルと検索結果

ジャンル	叙述用法	連体用法	連用用法	計
書籍	718	1,019	60	1797
雑誌	80	199	12	291
新聞	25	23	3	51
白書	5	2	0	7
教科書	1	2	0	3
広報紙	5	21	0	26
Yahoo!知恵袋	239	239	8	486
Yahoo!ブログ	455	406	12	873
法律	0	0	0	0
国会会議録	51	16	0	67
総計	**1,579（43.85%）**	**1,927（53.51%）**	**95（2.64%）**	**3,601**

BCCWJ の各ジャンルにおいて、叙述・連体・連用という 3 つの用法には以下の図 1 が示すような違いがある。

[2] http://www.kotonoha.gr.jp/shonagon/

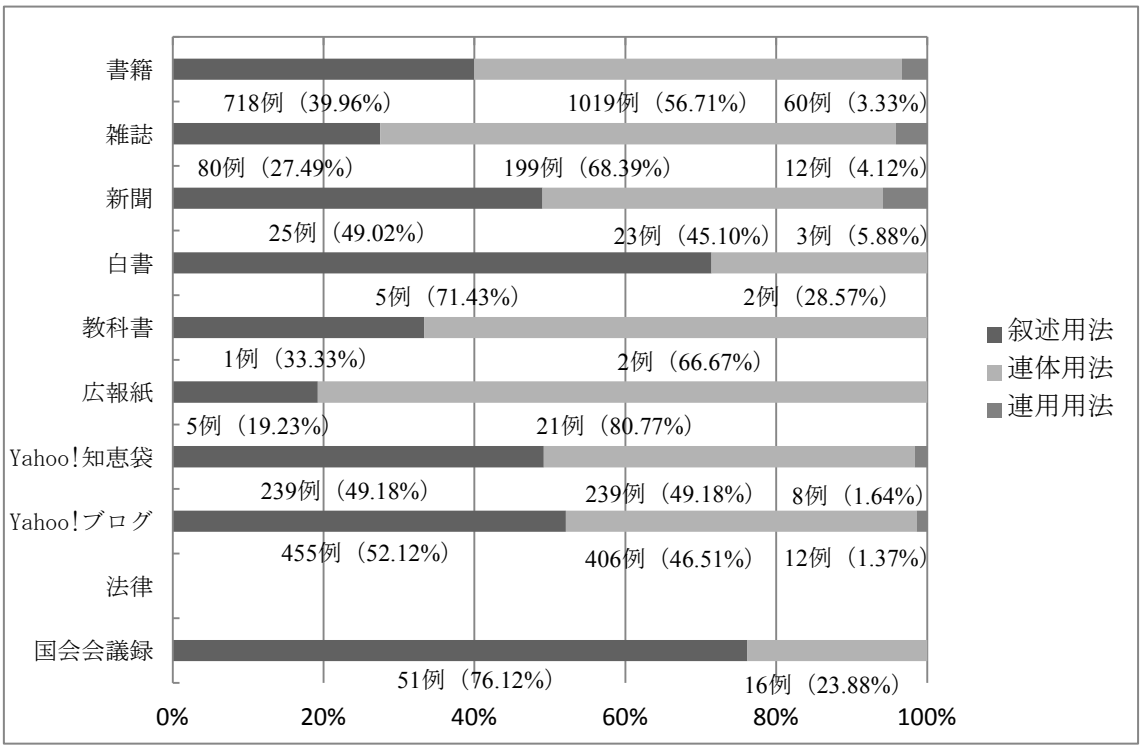


図 1　ジャンル別の用法分布

表 2 と図 1 を総合した結果、以下の分布が観察された。

1. コーパスの相違に関係なく、3 つの用法のうち、連用用法の使用頻度が最も低い。
2. ジャンルを問わず BCCWJ の全体的な分布からみれば、「甘い」においては、叙述用法に比べ連体用法のほうが使用頻度が高いことが読み取れる。
3. ジャンルによって、連体用法と叙述用法のどれが多く用いられるかは異なる。書籍・雑誌・教科書・広報紙においては、連体用法の方が多く使われるのに対して、白書・国会会議録・Yahoo!ブログにおいては、叙述用法の方が多く用いられている。また、新聞・Yahoo!知恵袋の 2 種においては、連体用法と連用用法はほぼ同じ使用頻度で用いられ、大きな差が見られない。
4. テクストタイプによって、形容詞「甘い」が中心的に担っている機能には違いがある。

次に、各ジャンルにおける「甘い」の各用法の使用頻度をもとに、SPSS を用いてピアソン積率相関係数を求め、各ジャンル間の分布の類似を考察した。頻度調査の結果、法律というジャンルにおいて「甘い」の件数がゼロであった。異なるコーパスサイズの比較においてゼロの意味は異なると判断し、法律というジャンルを除いた 9 種のジャンルを最終的な分析対象とした。分析によって、下記の表を得た。

表 3　各ジャンルの相関分析

	書籍	雑誌	新聞	白書	教科書	広報紙	Yahoo! 知恵袋	Yahoo! ブログ	国会 会議録
書籍	1	.933	.923	.581	.978	.875	.952	.916	.500
雑誌	.933	1	.723	.250	.988	.991	.778	.710	.155

新聞	.923	.723	1	.849	.822	.622	.997	1.000*	.794
白書	.581	.250	.849	1	.397	.115	.803	.859	.995
教科書	.978	.988	.822	.397	1	.957	.866	.811	.307
広報紙	.875	.991	.622	.115	.957	1	.684	.607	.018
Yahoo!知恵袋	.952	.778	.997	.803	.866	.684	1	.995	.742
Yahoo!ブログ	.916	.710	1.000*	.859	.811	.607	.995	1	.805
国会会議録	.500	.155	.794	.995	.307	.018	.742	.805	1

（*.相関係数は 5%水準で有意です）

石川・前田・山崎（2010：86）によれば、一般に相関係数の絶対値が.7 より大きければ「強い相関」が、.4 より大きければ「中程度の相関」が、.2 より大きければ「弱い相関」があるとされ、.2 以下の場合は「相関なし」と判断するという。表 6 の結果から、「甘い」の各用法の使用頻度において、BCCWJ の 9 種類のジャンルの間にどの程度の相関が出ているかについて、次の点が窺える。

1. 全体的に、相関の強弱がメリハリのついた形で観察できる。
2. 話し言葉を書き起こした国会会議録は、公的な書き言葉である白書と最も相関の強い（r＝.995）ジャンルの組み合わせである。一方、国会会議録と雑誌、広報紙の間には相関が認められない。
3. 新聞と Yahoo!知恵袋間､新聞と Yahoo!ブログ間でそれぞれ r＝.997 と r＝1.000 という高い相関係数が得られる。また、新聞と Yahoo!ブログ間は、無相関検定でも相関有意が確認されたため、強い正の相関があると結論してよい。
4. 書籍は雑誌、教科書、Yahoo!知恵袋のいずれに対しても r＝.930 以上で強い相関となっている。一方、国会会議録と白書との間ではそれぞれ r＝.500、r＝.581 という中程度の相関が見られる。

4.3 考察

以上、BCCWJ の各ジャンルごとに、「甘い」の各用法の使用頻度を調査し、さらにそれを元に各ジャンル間の相関関係を概観してきた。

まず、BCCWJ の全体的な分布からみれば、「甘い」の 3 つの用法のうち、連体用法が最も多く、叙述用法がそれに次ぎ、連用用法が最も少ないということが明らかである。形容詞の連用用法を副詞とする説もあるくらいで、形容詞の主な機能として認められていない。今回の調査結果で、連用用法の使用頻度が最も低いこともこの点をよく反映している。また、「甘い」の叙述用法と連体用法の使用頻度について、BCCWJ の全体的な分布及び各ジャンル毎の分布をそれぞれ考察した。コーパスの全体的な分布からみれば、連体用法の方が叙述用法より多く用いられることが分かる。しかし、ジャンル毎の分布からみると、ジャンルによって　①連体用法の方が多い　②叙述用法の方が多い　③連体用法と叙述用法がほぼ同じぐらい多い　という 3 つのパターンが観察できる。BCCWJ のジャンルによって、「甘い」の叙述用法と連体用法の使用頻度は違う分布を示すことが明らかになった。すなわち、用例採集に用いられるデータの性格は、調査結果に影響を及ぼすことが考えられる。

次に、「甘い」の各用法の使用頻度をもとに、BCCWJ の 9 種類のジャンルの間にどの程度の相関が出ているかを見てみよう。類似点に注目すれば、新聞と Yahoo!ブログとの 2 つのジャンルにおいて、叙述用法の方が連体用法より出現頻度がやや多い点で似ていて、2 類

のジャンルの間は強い相関を示している。また、Yahoo!知恵袋では「甘い」の連体用法と叙述用法は同じぐらいの使用頻度で使用されている。一方、あるジャンルで連体用法がよく用いられるが、別のジャンルで叙述用法のほうが多く使われるようなずれが見られることが明らかになった。このことから、「甘い」の各用法の使用頻度は BCCWJ の各ジャンルによって違う傾向を見せていると考えることができる。特に、話し言葉を書き起こした国会会議録は、叙述用法のほうが多く用いられ、書籍・雑誌のジャンルとの間に大きな相違が見られる。この調査結果は、八亀（2007：63）が指摘した話し言葉において形容詞の機能が述語中心であることとある程度一致している。

また、八亀（2007：63）では、新聞や評論的な文章においては形容詞が規定語中心であるという指摘が見られる。今回利用した BCCWJ のコーパスでは、新聞ジャンルにおいて「甘い」の連体用法と叙述用法はほぼ同じぐらいの使用頻度で用いられている。この調査結果は、BCCWJ の新聞ジャンルでは「甘い」の出現件数が少なく、連体用法と叙述用法の出現頻度ははっきりとした差が見られなかったことに理由があるかもしれない。また、個々の形容詞によって、叙述用法と連体用法のどちらが多く使用されるのかが異なっているということも考えられる。

5.「甘い」の語義分布

ここでは、叙述用法・連体用法・連用用法という 3 つの用法ごとに、多義語「甘い」の語義分布を調べ、形容詞の文法的機能による語義分布及び意味用法の違いを明らかにする。

5.1 形容詞の語義と用法との関係

形容詞の用法と語義との関係に関与する研究には、次のようなものがある。

宮島（1993）は『現代雑誌九十種の用語用字五十音順語彙表・採集カード』を用いて、合計度数 10 以上の形容詞を対象に、終止・連体・連用という 3 つの用法の量的な調査を行い、「いちおう用法のある語形についても、その用法の量的な面ではかたよりがある」（宮島 1993：94）ということを報告している。丹保（1997）は、IPAL と『学研』の意味区分を参照し、「高い」「広い」「寂しい」の語義を新しく分類した。また、38 冊の国語教科書から上の形容詞の用例を採集し、形容詞の連体・連用・終止用法の出現頻度が語義に大きく依存していることを示した。

形容詞の用法と意味の関係をめぐるこれまでの研究は、そのいずれも形容詞の語義と用法の関連性を理解する上では重要なものであり、示唆に富んだ考察を行っている。その一方で、以上の先行研究が行った理論的一般化に対して、どの程度日本語の使用例に妥当なものであるかを検証するような研究はあまりなされていなかった。また、先行研究が調査を行った際に用いられたデータについて、宮島（1993）と丹保（1997）はそれぞれ雑誌九十種のデータと 38 冊の国語教科書から用例を収集していた。形容詞「甘い」の各用法の出現頻度に対する考察では、頻度調査の結果が用例採集に用いられる資料によって異なることが明らかになった。つまり、1 種類のデータに限らず、様々なジャンルを含めた BCCWJ のようなデータを利用することによって、より精度の高い結果ができると考えられる。

こうした現状に対して、BCCWJ という大規模な言語データを活用しながら、再整理された多義語「甘い」の意味用法に基づき、多義語「甘い」の語義分布を調べ、形容詞の文法的機能による語義分布及び意味用法の違いを明らかにする。

5.2 調査結果と考察

次の表は、多義語「甘い」の語義と文中での用法の関連性を示したものである。

表 4　多義語「甘い」の用法別の語義分布

意味領域	「甘い」の意味用法	叙述用法	連体用法	連用用法	計
Ⅰ身体体験	①味覚[味：甘味]（基本義）	602 (38.32%)	936 (59.58%)	33 (2.10%)	**1571**
	②味覚[味：旨味]	21 (77.78%)	5 (18.52%)	1 (3.70%)	**27**
	③味覚[味：塩味・刺激性]	7 (46.67)	8 (53.33%)	0	**15**
	④嗅覚[匂い：香り]	10 (3.22%)	297 (95.50%)	4 (1.28%)	**311**
	⑤聴覚 [声・音：歌声・音楽]	11 (9.32%)	96 (81.36%)	11 (9.32%)	**118**
	⑥視覚[外見：容貌（男性）]	1 (2.56%)	37 (94.88%)	1 (2.56%)	**39**
	⑦視覚[外見：服装など]	24 (26.37%)	64 (70.33%)	3 (3.30%)	**91**
Ⅱ物事の状態	⑧[物事の機能・品質]	37 (100%)	0	0	**37**
	⑨[物事の状況・程度]	37 (58.73%)	19 (30.16%)	7 (11.11%)	**63**
Ⅲ人間活動	⑩[精神的行為の生産物]	13 (48.15%)	14 (51.85%)	0	**27**
	⑪[思考・行為]	501 (78.65%)	119 (18.68%)	17 (2.67%)	**637**
	⑫[態度・接し方]	149 (74.5%)	46 (23%)	5 (2.5%)	**200**
	⑬[思考・行為の対象]	139 (93.92%)	9 (6.08%)	0	**148**
	⑭[愛情・幸福]	25 (10.97%)	195 (85.53%)	8 (3.50%)	**228**
	⑮[誘惑]	2 (2.25%)	82 (92.13%)	5 (5.62%)	**89**
計		**1579**	**1927**	**95**	**3601**

この表に見られるように、多義語「甘い」の語義分布には以下のような特徴が見られる。

1. 多義語「甘い」の基本義①は 1571 例と比較的高い割合（全体の 4 割以上）で一番多く現れる語義である。「甘い」が表す五感以内の意味領域からみれば、味覚を表す語義（語義①、②、③）が最も多く使用され、次に嗅覚（語義④）、視覚（語義⑥と⑦）、

聴覚（語義⑤）という順で頻出する。また、物事の状態を表す「甘い」の意味用法（語義⑧と⑨）の使用頻度は 100 例であり、極端に少ない。人間活動という意味領域に見られる「甘い」の出現度数は計 1329 例で、全体の 4 割弱を占めている。

2. 叙述用法と連体用法の比重からみると、各語義によって違う分布が見られる。身体体験を表す語義のうち、②以外の語義では、叙述用法より連体用法の比重がはるかに上回っている。特に、嗅覚を表す語義④と視覚を表す語義⑥では、連体用法がそれぞれ 297 例（95.50%）と 37 例（94.8%）でいずれも叙述用法より圧倒的に高い割合を占めている。一方、物事の状態に用いられる場合（語義⑧と⑨）及び人間活動を表す一部の語義（語義⑪、⑫、⑬）においては、連体用法より叙述用法の比率が高い。

3. 多義語「甘い」の全ての語義には叙述用法が見られる。しかし、語義③、⑩、⑬には連用用法がなく、語義⑧には連体用法も連用用法も見られない。このように、語義によって連体用法や連用用法がないものもある。

以上、BCCWJ に対する調査に基づき、多義語「甘い」の語義分布について調べてきた。「甘い」はその語義によって、叙述用法・連体用法・連用用法の比率にかたよりが見られる。表 4 から明らかなように、感覚領域においては、「甘い」の連体用法が中心として用いられている。味覚を基本義とする「甘い」は、物事の状態及び人間活動の意味領域へと意味拡張するとともに、その文中での機能も次第に叙述用法の方へとベクトルが向かうようになる。特に、日本語の「甘い」はネガティブな意味を表す点で、中国語の“甜”と英語の/sweet/と大きく異なっている。表 4 から読み取れるように、「甘い」がマイナス評価を表す場合、その用法は叙述用法に集中している。

6. 終りに

本研究は、日本語の多義語「甘い」をめぐって、コーパス調査に基づき、その意味用法の再分類を試みた。豊富な資料が収集された BCCWJ を活用することによって、日本語の味覚形容詞「甘い」は実際どのような意味として使われるのかがある程度分かるようになった。また、形容詞「甘い」の各用法の使用頻度に対する調査では、BCCWJ の各ジャンルによって「甘い」の叙述・連体・連用用法に違った分布が見られることが分かった。さらに、多義語「甘い」の各語義の量的分布を調べた結果、語義によって叙述用法か連体用法のどちらが頻出するかにはかたよりが見られる。「用法・語形の出現頻度が語義に大きく依存している」という丹保（1997）の指摘があったように、多義語の語義によってその用法が変わってくることは、日本語の形容詞「甘い」の用例によっても検証できた。これは、形容詞の文中での用法は、語彙的意味の存在条件の一つとして考えられるということを示唆する。単語の語彙的意味に影響を及ぼすものは、ほかにもあるはずであるが、それを研究することを今後の課題として、本稿を締めくくりたい。

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

鈴木重幸（1972）『日本語文法・形態論』むぎ書房

砂川有里子（2010）「コーパスを活用した日本語教育研究―日本語学習辞書編集に向けて―」砂川有里子・加納千恵子・一二三朋子・小野正樹編著『日本語教育研究への招待』くろしお出版 pp.99-119

高橋太郎（1998）「動詞からみた形容詞」『言語』27：3，pp.36-43

丹保健一（1997）「形容詞の連体，連用，終止用法の出現頻度と意味との関連性をめぐって：「高い」「広い」「寂しい」を例として」『三重大学教育学部研究紀要』（人文・社会科学）48，pp.9-18
西尾寅弥（1972）『形容詞の意味・用法の記述的研究』秀英出版
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，pp201-214
八亀裕美（2010）「形容詞研究の現在」工藤真由美編『日本語形容詞の文法―標準語研究を超えて』ひつじ書房 pp.53-77
宮島達夫（1993）「形容詞の語形と用法」『計量国語学』19：2，pp94-104

関連 URL

現代日本語書き言葉均衡コーパス（中納言）
https://chunagon.ninjal.ac.jp/login

外来語由来の接尾辞「チック」と類義語との比較

村中淑子（桃山学院大学 国際教養学部）†

The Use Situation of Japanese Suffix *chikku* Derived from English "-tic" and its Synonym

Toshiko MURANAKA (Faculty of International Studies and Liberal Arts, St. Andrew's University)

１．はじめに

外来語が日本語の語彙において重要な地位を占めつつあるという指摘はなされているが、その多くは、単語レベルの現象に着目したものである。しかし、日本語の語彙の動的な実態ということを考えた場合、単語を作り出す接頭辞・接尾辞レベルの外来語に関する研究も必要なのではないかと思われる。

本発表では、接頭辞・接尾辞レベルの外来語に関する研究の一環として、外来語由来の接尾辞「チック」に焦点を当て、類義語との比較も交えつつ、考えてみたい。

野村（1977）は「外来語が（中略）造語能力をもつものとして、漢語につぐ存在となりつつある」と述べており、単語を作り出す造語成分としての外来語に注目している。石野（1992）は英語の接尾辞を含む日本語として「アルバイター」「おとめチック」「がんばリズム」「にやリスト」「キャッシング」「スキンシップ」「ファンタジック」「サイノロジー」を挙げ、日本語として形がはっきりしているのは「チック」「シップ」であるという。米川（1992）は、英語からの接尾辞「チック」は造語力がある、と述べている。しかし接尾辞「チック」の使用実態について詳しく調べたものはほとんど見あたらない。おそらく周辺的な現象にすぎないとみられているためであろう。

２．接尾辞「チック」について

接尾辞チックの語形と文法的性質について、最大公約数的にまとめると次の通りである。

造語成分として働いているものではなく、外来語の一部をなすものとしては、「チック」と「ティック」という表記・音声両面にわたる２種のバリエーションがあるが（ロマンチック／ロマンティック、ドラマチック／ドラマティック、オートマチック／オートマティックなど)、日本語の中で新たなことばを作り出す、すなわち造語力を持つ成分としては、「チック」の形だけである（おとめチック／×おとめティック、漫画チック／×漫画ティック、おばさんチック／×おばさんティック、など)。

通常、前接要素として名詞をとり、ナ形容詞を形作る。

すなわち、「 名詞 チックな 被修飾要素 」という形で使われる。

３．接尾辞「チック」の使用実態

３．１　15 年間の新聞記事における接尾辞「チック」

接尾辞「チック」はどのくらいの頻度で、どのような分野において使われているのか。

まず、15 年間の毎日新聞全記事において接尾辞「チック」が使われた回数を、表 1 にまとめた（ここで用いた「CD-毎日新聞」（1991～2005）は、学術研究向けのタグ付きテキストデータではなく、一般コンシューマー向け商品である)。

† tmuranaka@andrew.ac.jp

15 年間で 39 件、1 年あたり平均 2.6 件とごく低頻度であるが、使われていない年は無く、安定的・継続的な使用が行われていると言ってよいだろう。

前接要素としては、「漫画」「乙女」の類が約半数を占めている。この２つについては、チックのついたナ形容詞の形が固定化し、語単位で定着したと考えられる。それ以外のものについては、造語力を発揮して新しい語をその場その場で作り出しているものであろう。

表１：毎日新聞全記事における「〜チック」の語形と出現度数の推移 [1]
（２回以上出現した場合は、語形の後ろに回数を示した）

	漫画系	乙女系	SF 系	その他	計
1991	漫画チック				1
1992	漫画チック	乙女チック	SF チック 3		5
1993	漫画チック 2				2
1994		乙女チック	SF チック	演歌チック	3
1995	劇画チック				1
1996	少女漫画チック	乙女チック		成金チック	3
1997	漫画チック 2			小説チック	3
1998		少女チック			1
1999	漫画チック 2				2
2000				おもちゃチック 変態チック レトロチック	3
2001	漫画チック	乙女チック		古典チック 劇場チック	4
2002	漫画チック			おばちゃんチック	2
2003		乙女チック		反動チック 絵画チック SM チック	4
2004				映画チック	1
2005	漫画チック 少女漫画チック			絵画チック 2	4
計	14	6	4	15	39

これらの出現した新聞記事をみてみると、本・映画・演劇・テレビ番組・音楽などの紹介記事、エッセイ類、投稿欄、芸能人等へのインタビュー記事、などにほぼ限定されていた。つまり、娯楽提供を目的とする新聞記事において接尾辞「チック」が出現すると言ってもよさそうである。

３．２ 「中納言」における接尾辞「チック」

次に、新聞も含めた現代日本語の書きことば全体で接尾辞「チック」がどれくらい使われているのか調べるために、「中納言」によって BCCWJ を検索し、得られた「〜チック」の語形を全て表２に示す。「〜チック」がゼロであったサブコーパス「新聞」「教科書」「白書」「広報紙」「法律」「韻文」「ベストセラー」は表にしていない。また、サブコーパスごとのデータ発行年の幅が 1 年間〜35 年間とばらばらなので、経年変化は見ないこととする。

表２を見ると、前接要素のバラエティがたいへん豊かであり、「チック」の造語力の強さがうかがえる。異なり件数 68、延べ件数 117 である。ここで検索された「チック」の前接

[1] 表１・２・３は村中（2012）から引用している。

要素には、毎日新聞記事データには見られなかった特徴がみられる。次の３つである。

①「チック」の前接要素は、名詞ではなくナ形容詞である場合もある

②「チック」の前接要素は、固有名詞である場合もある。

③「チック」の前接要素は、語ではなく句である場合もある。

①は、異なり件数 68 のうち、8 件。インスタント、おしゃれ、高級、スキャンダラス、面倒、妖艶、リアル、レトロ、である。「名詞＋チック」でナ形容詞を形成するのではなく、「ナ形容詞＋チック」でナ形容詞となっている。つまりこれら８件は、文法的にはチックが余剰的についているものである。Yahoo!ブログ、Yahoo!知恵袋、雑誌、国会会議録、の４つのサブコーパスに見られる。

②は、異なり件数 68 のうち、３件。「楳図さん」「聖子ちゃん」「ムネオ（鈴木宗男)」である（「セーラー服とほにゃらら編」も「セーラー服と機関銃」を意味しているとすれば固有名詞として数えてもいいかもしれない)。Yahoo!ブログと国会会議録に見られる。

③は、異なり件数 68 のうち２件。「日本の銭湯」「セーラー服とほにゃらら編」であり、それぞれ生産・書籍と Yahoo!ブログに出現した。前者は対談における発言中であった。口語性の強いテキストでは句に接尾辞「チック」がつくことが許容されるのだと考えられる。

また、15 年間の毎日新聞記事データには見られなかったが BCCWJ に多く見られた語形として「メルヘンチック」がある。これもナ形容詞として定着した語形とみてよいだろう。

表２：BCCWJ「中納言」における「～チック」の語形

（語形のアイウエオ順。２回以上出現した場合は語形の直後に出現回数を記した。計は、延べ数。）

Yahoo!ブログ	アニメチック　２、インスタントチック、楳図さんチック、お菓子チック、オカマチック、オカルトチック、乙女チック　４、おまけチック、喫茶店チック、ギャグチック、ギャルチック、求肥チック、クワガタポイントチック、劇画チック、高級チック、サスペンスチック、サバイバルチック、首都圏っチック、聖子ちゃんチック、制服チック、セーラー服とほにゃらら編チック、低音チック、展望台チック、天ぷらチック、箱庭チック、パワプルガールズチック、変態チック、マンガチック、漫画チック、ミリタリチック、ミルフィーユチック、メビウスチック、メルヘンチック　６、やらせチック、夕日チック、リアルチック、レトロチック	計46
流通・書籍	英語チック、オカルトチック、オトメチック　２、乙女チック　４、芸術チック、童画チック、変態チック、ポルノチック、マンガチック、メルヘンチック　６、物語チック、理系チック	計21
生産・雑誌	アジアンチック　３、SFチック、お菓子チック、おしゃれチック、女のコチック、歌謡曲チック、サラシモノチック、少女漫画チック、南国チック、姫チック、マンガチック　２、漫画チック、水商売チック、妖艶チック、レトロチック	計18
生産・書籍	SFチック、オカルトチック、古典チック、自動車チック　２、日本の銭湯チック、廃墟チック、マンガチック　２、メルヘン・チック、メルヘンチック　６	計16
Yahoo!知恵袋	アジアンチック、オカルトチック、お嬢様チック、刈上げチック、カルトチック、金属チック、高級チック、コントチック、哲学チック、東北弁チック、メルヘンチック２、面倒チック、ヤンキーチック	計14
国会会議録	スキャンダラスチック、ムネオチック	計2

表２で「～チック」の出現度数を見ると、Yahoo!ブログが圧倒的に多く、書籍、雑誌、Yahoo!知恵袋がおよそ同程度にみえるが、サブコーパスごとの総語数はかなり異なる。そこで、各サブコーパスの総語数と、総語数を 1000 万語に換算した出現の割合（小数点以下第３位を四捨五入）を示したのが表３である。流通・書籍と生産・書籍をまとめて「書籍」とした。比較のため、「新聞」も示した。これをみると、接尾辞「チック」の出現割合は、「Yahoo!ブログ」が最も多く、「雑誌」もそれに迫る多さである。その３分の一程度の割合で「Yahoo!知恵袋」が続き、「書籍」はさらにその半分以下となる。

表３：接尾辞「チック」の出現度数とサブコーパスごとの出現率

サブコーパス	ブログ	書籍	雑誌	知恵袋	国会	新聞
出現度数	46	37	18	14	2	0
総語数（概数）	1030万	6230万	440万	1030万	510万	140万
1000万語あたりの出現率（概数）	44.66	5.94	40.91	13.59	3.92	0

３．３　接尾辞「チック」の使用実態についてのまとめ

接尾辞「チック」はかなり強い造語力を持ち、次々にあたらしいナ形容詞を形成している。漫画チック・乙女チック・メルヘンチックについては語単位で定着しているといってよさそうだが、そのほかのものは、「チック」の造語力によりその場その場で新しく作られている。低頻度であるが、少なくとも 15 年間、安定的・継続的な使用が行われている。

通常、「普通名詞チックな被修飾要素」という形で使われるが、前接要素が固有名詞であったり、ナ形容詞であったり、語でなく句であったり、というケースも生じている。すなわち、前接要素についての許容範囲が広がりつつある。

新聞においては、娯楽提供を目的とする記事に接尾辞「チック」が出現している。そのほかの媒体では、ブログや雑誌に多く出現している。すなわち、口語的なくだけた親しみやすいテキストで、評価的な内容を記述する際によく使われるということであろう。

４．接尾辞「チック」と類義語との比較

４．１　接尾辞「チック」とその類義語

接尾辞「チック」の類義語として、「的」「～（っ）ぽい」「ライク」などが挙げられる。いずれも、普通名詞　（っ）ぽい／的な／チックな／ライクな　被修飾要素の形で使用することが可能である。次に具体例を挙げる。

表４：「（っ）ぽい」「的」「チック」「ライク」の比較

女の子（っ）ぽい服装	論文（っ）ぽい書き方	ホテル（っ）ぽい内装
女の子的な服装	論文的な書き方	ホテル的な内装
女の子チックな服装	論文チックな書き方	ホテルチックな内装
女の子ライクな服装	論文ライクな書き方	ホテルライクな内装

比べてみると、「（っ）ぽい」は柔らかい感じ、「的」はやや固い感じがある。「チック」はやや子供っぽいような、揶揄するようなニュアンスが感じられる。「ライク」はしゃれた感じで軽さがあるが、「チック」のように子供っぽくはなく、揶揄するようなニュアンスも

無い。

外来語は、一般に、西洋風／近代的／しゃれた／モダン／カッコ良さ／明るさ、などの性質を持つと言われる。しかし、接尾辞「チック」については、「明るさ」以外はあまり当てはまらず、「明るさ」もあまりぴったりした形容ではない。「子供っぽさ」「滑稽味」「揶揄」といったニュアンスを持つ。それはなぜだろうか。／チック／という音声の並びがそのように感じさせるのだろうか。

子供っぽいニュアンスがあるとは言っても、使用者の中には、年配男性の国会議員（BCCWJ の例）や、中年のプロ野球選手、中高年の主婦、芸術家男性など（毎日新聞記事の例）が含まれている。決して若者言葉ではなく、ぶりっ子のことばでもないのである。

４．２　接尾辞「～ライク」について　―「中納言」による検索結果―

「チック」とおなじく外来語由来の接尾辞である「ライク」について、「中納言」で検索した結果を示す。

表５：BCCWJ における「～ライク」の出現数

雑誌	流通・書籍	生産・書籍	ベストセラー	Yahoo! 知恵袋	Yahoo! ブログ	計
43（メンズ 20、レディ６、ビジネス１）	20（ビジネス 17）	15（ビジネス８、レディ１）	2（ビジネス 2）	6（ビジネス 3）	5（ビジネス１、レディ１）	91

「ビジネスライク」の語形が定着しているほか、メンズライク・レディライクがファッション用語としてある程度定着しているようである。カジュアルライク・スポーツライク・デジカメライク・乗用車ライクなど、頻度 1 のものが 22 件あり、趣味の道具などの分野で造語力を発揮しているようである。

４．３　接尾辞「的」について

南雲（1994）および丸山（1997）によると、接尾辞「的」は、次のような性質を持つ。

ア）分野によっては使用される度合いが異なる。(『中央公論』1962 年 11 月号においては、政治、経済、文化一般、教育に多く、社会問題、科学などでは少ない。『中央公論』1992 年 11 月号においては、政治経済に多く、文芸・広告で少ない。)

イ）「的」の前につく語は語種では漢語が圧倒的に多く、文字では漢字が多い。

ウ）専門分野によって、特定の語が繰り返し使用される傾向がある。

「的」の使用頻度は「チック」に比べて桁違いに高いので、「的」の使用の「少ない」分野であっても、「チック」の使用よりは多いのである。「的」が使われない分野を埋める形で「チック」が入り込んで来たのかどうかについては、今後検討する必要がある。

４．４　接尾辞「（っ）ぽい」について

小原（2010）は、BCCWJ2009 を用いて、接尾辞「（っ）ぽい」について調べている。「（っ）ぽい」の新規用法、すなわち「（っ）ぽい」が文や句の後ろに接続する例を中心に述べられている。「チック」と同様、口語的なテキストで多く使われていること、臨時的な用法、す

なわち造語力が発揮されている例が多く見られること、が明らかにされている。

５．おわりに

外来語由来の接尾辞「チック」は、使用範囲が狭く、低頻度ではあるが、造語力を強く発揮する接尾辞として安定的使用がみられる。

外来語の役割として、「外国文化の享受」「新たな概念の導入」などがあるといわれるが、接尾辞については、それらは当てはまらないであろう。外来語由来の接尾辞の場合は、「モノは同じでも新しいニュアンスを加える」という役割しか持たないように思われる。そのニュアンスが、なぜ、どのように生じて来たのか。

今後、類義語との違いを詳しく検討し、明らかにしていきたい。

文献

石野博史（1992）「外来語の造語力」『日本語学』11:5,pp.42-49.
小原真子（2010）「接尾辞「-ぽい」について」『島大言語文化：島根大学法文学部紀要言語文化学科編』29,pp.59-76.
南雲千歌（1994）「現代日本語の「～的」について--雑誌『中央公論』1992 年 11 月号の場合」『ICU 日本語教育研究センター紀要』３,pp.72-98.
野村雅昭（1977）「造語法」『岩波講座日本語９』岩波書店,pp.247-284.
野村雅昭（1984）「語種と造語力」『日本語学』3:9, pp.40-54.
丸山千歌（1997）「英語の接尾辞”—tic”の訳語「～的」について-『中央公論』1962 年 11 月号の場合-」『ICU 日本語教育研究センター紀要』６,pp.15-42.
村中淑子（2012）「接尾辞「チック」について—「CD-毎日新聞」(1991-2005)および BCCWJ を用いて—」『国際文化論集』45 号,pp.115-144.（桃山学院大学 総合研究所）
米川明彦（1992）「新語と造語力」『日本語学』11:5,pp.50-57.

資料

『現代日本語書き言葉均衡コーパス』(略称 BCCWJ) 検索ツール　短単位検索 Web アプリケーション「中納言」　URL：http://chunagon.ninjal.ac.jp/search
『CD-毎日新聞』（1991～2005）日外アソシエーツ
『毎日新聞　縮刷版』毎日新聞社

語義曖昧性解消のための領域適応手法の決定木学習による選択 —三手法からの決定—

古宮　嘉那子（東京農工大学　工学研究院）†
奥村　学（東京工業大学　精密工学研究所）

Determination of a Domain Adaptation Method for Word Sense Disambiguation Using Decision Tree Learning - Determination from Three Methods-

Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institution of Technology)

１．はじめに

語義曖昧性解消（Word Sense Disambiguation，WSD）について領域適応を行った場合，最も効果的な領域適応手法は，ソースドメインのデータ（ソースデータ）とターゲットドメインのデータ（ターゲットデータ）の性質により異なる (Komiya and Okumura 2011)．WSD の対象単語タイプ，ソースデータ，ターゲットデータの三つ組を 1 ケースとして数えるとする．本稿では，このケースごとに，データの性質から，最も効果的な領域適応手法を，決定木学習を用いて自動的に選択する手法について述べるとともに，どのような性質が効果的な領域適応手法の決定に影響を与えたかについて考察する．また本稿では，3 つ以上の n 個の領域適応手法から選択するために, pairwise 方式で nC2 通りの二分決定木をつくり，最終的にそれらを統合することで，ひとつのケースにつきひとつの領域適応手法を決定する方法を，三手法からの選択を例にとって述べる．

２．関連研究

領域適応の研究は様々な分野で研究が行われているが, 本稿に最も近い研究は, (McClosky, Charniak, and Johnson 2010) である．この研究では, 多様なドメインからなる文書を構文解析する際，最も良いモデルは異なるという問題に注目している．彼らは様々な混合モデルによる構文解析の正解率を回帰分析で予測し，それぞれのターゲットデータに対して，最も高い正解率を出すと予測されたモデルを利用して構文解析を行っている．本研究との最も大きな違いは，対象のタスクが構文解析ではなく語彙曖昧性解消である点である．また，彼らは多様なドメインからなる文書があることを想定しているが，我々は想定していない．本研究では決定木学習を用いることで，どのような性質が最適な領域適応の決定に影響を与えるのかについて考察する．

また，二手法からの選択に関しては，(Komiya and Okumura 2011)にて既に述べたため，本稿では三手法以上からの選択にした場合にどのような工夫が必要かという部分を中心に述べる．

３．領域適応手法の自動選択

ケースごとに適切な領域適応手法を自動的に選択し，その手法を適宜用いて領域適応を行えば，どれかひとつの手法を用いるよりも，WSD の性能が向上することが予想される．このため，決定木学習を用いて，領域適応手法の自動選択を行う．決定木学習を用いるこ

† kkomiya@cc.tuat.ac.jp

とで，どのような性質が最適な領域適応手法の決定に影響を与えるのかを明示的に示すことができる．

３．１　WSD のための領域適応手法

WSD のための領域適応手法として，本研究では以下に示す三つを用いる．したがって，pairwise 方式で三つ（Target Only と Random Sampling，Target Only とフィルタリングによる削除，Random Sampling とフィルタリングによる削除）の二分決定木をつくり，最終的にそれらを統合することで，ケースごとに領域適応手法を決定する．

- Target Only (TO)：ソースデータを用いず，ランダムに選んだ少量のターゲットデータにラベル付けしたものだけを訓練事例にする．
- Random Sampling (RS)：ランダムに選んだ少量のターゲットデータの単語トークンにラベル付けしたものとソースデータの両方を訓練事例にする．
- フィルタリングによる削除(FD)：ランダムに選んだ少量のターゲットデータの単語トークンにラベル付けしたものとソースデータの両方を訓練事例にする．このときソースデータは，フィルタリングによりターゲットデータにある一定の閾値以上似ているデータだけを用いる．

なお，追加するターゲットデータのトークン数は常に 10 件とした．また，WSD の分類器としてはマルチクラス対応の SVM（libsvm）の線形カーネルを使用した．また，WSD の学習の素性には，WSD の対象単語の前後二語までの形態素の表記，WSD の対象単語の前後二語までの品詞，WSD の対象単語の前後二語までの品詞の細分類（分類語彙表に(国立国語研究所 1964)よる），WSD の対象単語の前後二語までの分類コード，係り受けを用いた．

３．２　決定木学習のラベル

ケースごとに，最も WSD の正解率がよかった手法によって，領域適応の手法名のラベルか，Same ラベルをつけた．決定木は，ケースごとにソースデータとターゲットデータの性質から，二つのうちどちらの手法を使って領域適応するべきかを判定している．作成する三つの決定木のうちのひとつ，Target Only か Random Sampling を選択する決定木では以下のように付与する．

- Target Only : Random Sampling より Target Only を使用した方が WSD の正解率が良いケース
- Random Sampling : Target Only より Random Sampling を使用した方が WSD の正解率が良いケース
- Same : Target Only と Random Sampling のどちらを使っても WSD の正解率に差がないケース
-

３．３　決定木学習の素性

最適な領域適応手法はソースデータとターゲットデータの分布や距離などの性質によって異なると考えられるため，それぞれの決定木に 24 種類，合計 40 の素性を利用した．これらのうちには，すべての領域適応手法において共通して使用している，ランダムに選んだターゲットデータの 10 トークンを使用して Leave-one-out 法で求めた領域適応手法のシミュレーションの正解率や，その比率，ターゲットデータやソースデータの件数や，ランダムに選び人手でラベル付けしたターゲットデータの 10 トークン中の最も頻度の高い語義

に関する情報，また WSD に使用した素性の JS 距離などが含まれている．

４．実験データ

実験には，現代日本語書き言葉均衡コーパス（BCCWJ）(Maekawa 2008) の白書のデータと Yahoo! 知恵袋のデータ，また RWC コーパスの毎日新聞コーパス (Hashida, Isahara,Tokunaga, Hashimoto, Ogino, and Kashino 1998) の三つのデータを利用し，ソースデータとターゲットデータを変えることで，全部で 6 通りの領域適応を行った．これらのデータには岩波国語辞典(西尾，岩淵，水谷 1994) の語義が付与されている．これらのコーパス中の多義語のうち，ソースデータおよびターゲットデータ中にともに 50 トークン以上存在する単語を実験対象とした．対象単語は「場合」，「自分」，「事業」，「情報」，「地方」，「社会」，「思う」，「子供」，「分かる」，「考える」，「含む」，「使う」，「技術」，「関係」，「時間」，「一般」，「現在」，「作る」，「今」，「前」，「持つ」，「進む」，「見る」，「入る」，「言う」，「出す」，「手」，「出る」 である．

５．決定木学習におけるラベル付きデータの作成方法と学習方法

決定木学習におけるデータのラベル付けの際は，決定木で判定する領域適応手法二手法の WSD の正解率を比較してカイ二乗検定を行い，有意差がないものに Same をつけ，領域適応手法名のラベルを付与した．なお，カイ二乗検定の有意水準は 0.05 を利用した．

また，決定木学習において Same が付与されたケースを訓練事例から削除して決定木で判定する領域適応手法二手法の 2 値分類の決定木学習を行った．なお，テストには全ケースを利用した．

さらに，決定木学習の際は全てのケースに同等の重みがあるとして決定木学習を行った．

領域適応手法決定のための決定木作成アルゴリズムには C4.5 (Quinlan 1993) を利用し，二分決定木を作成した．また，五分割交差検定を行った．決定木作成の枝刈りの閾値は訓練事例の 1/4 を開発用データとした予備実験により最適化した．

５.１　決定木の統合

決定木の統合は，以下のように行った．pairwise の性質上，三つの決定木が三つとも同じ方法がよいと答えることはなく，答えが 2:1 に分かれるか，三つ巴になるはずである．

このうち，2:1 に分かれるときは，かならず 2 つの決定木が出した答えが理論的に一番良くなるため，その答えを選択すればよい．手法 1> 手法 2 のとき手法 1 のほうがよい手法であるとすると，例えば， Target Only>Random Sampling かつ，フィルタリングによる削除>Random Sampling かつ Target Only> フィルタリングによる削除であれば，Target Only> フィルタリングによる削除>Random Sampling なので，Target Only を選択する．

次に，三つ巴のときには，事例が割りつけられた葉についている確率を比較し，一番高い確率のところに割り付けた．確率は，「学習時にその葉に割りつけられた最も多いケース数/学習時に，その葉に割りつけられた全ケース数」として計算した．たとえば，テストデータが，実行時に「学習時に，Target Only が 1 件，Random Sampling が 2 件割り当てられた葉」に割り当てられた場合，そのテストデータは 2/3 の確率で Random Sampling となる．三つ巴の場合には，この確率で比較し，最も高い確率の手法を割り当てた．

三つ巴のときに，ふたつの決定木で割りつけられた葉の確率が同率一位である場合には，Random Sampling>Target Only かつフィルタリングによる削除>Random Sampling なら，フィルタリングによる削除>Random Sampling>Target Only なのでフィルタリングによる削除を選択，というように論理的に選択した．

また，三つ巴でどれも確率が等しい時など，上記のルールを利用してもどうしても領域適応手法が選べない時には，一括的に領域適応を行ったときに正解率が高い順，つまり，フィルタリングによる削除，Target Only，Random Sampling の順で割りつけた．

６．結果

表 1 に，WSD の平均正解率の比較を示す．なお，144 のケースには合計 232116 語義曖昧性解消の対象単語トークンが含まれており，それらのマイクロ平均である．また，人手による選択は，決定木学習を用いる代わりに，ラベルとなっているふたつの領域適応のうち，WSD の正解率の高い領域適応手法をケースごとに人手で選択して，WSD の平均正解率を求めた値であり，upper bound である．

決定木学習を用いて選択した手法を利用した際の WSD の平均正解率は 83.52%であり，個別の手法を用いた際の最高の正解率，フィルタリングによる削除の 82.27%よりも正解率が高いため，決定木を利用して適切な領域適応手法を利用した方が，個々の領域適応手法を使った時よりも正解率が上がることが分かる．またこのとき，カイ二乗検定により十分な有意差が認められた．

表 1　WSD の平均正解率の比較

領域適応手法	WSD の平均正解率
Target Only	81.23%
Random Sampling	80.28%
フィルタリングによる削除	82.27%
決定木により選択された領域適応手法	83.52%
人手により選択された領域適応手法	85.87%

７．考察

五分割交差検定の五回の検定のうち，最も高い正解率だった決定木を付録として示し，生成に特に貢献した素性と素性値について以下に述べる．まず，Target Only と Random Sampling の決定木のルートノードでは，「ふたつの正解率の比=0.70 以上」が no のとき Target Only が割り当てられた．これは「the Other のシミュレーションの正解率/ Target Only のシミュレーションの正解率」の割合が 0.70 以下であれば，Target Only が割り当てられたということである．つまり，10 件のターゲットデータにラベル付けし，Leave One Out 法で評価を行った際の正解率のほうが，ソースデータで分類器を学習し，10 件のターゲットデータにラベル付けしたもので評価した正解率よりも高いときには Target Only が割り当てられたということに等しい．このことから，10 件のラベル付けしたターゲットデータによるシミュレーションの予測が，最適な領域適応の手法を予想する強力な手がかりになることが分かる．

また，Random Sampling とフィルタリングの削除の決定木のルートノードでは，「ソースデータ件数/ターゲットデータに一定以上似ているソースデータ件数=186.85 以上」のときフィルタリングの削除が割り当てられた．フィルタリングの削除は，ターゲットデータに閾値以上似たソースデータだけを訓練事例に利用する手法であるため，ターゲットデータに閾値以上似ていないソースデータ件数が多量にあるときには，ソースデータ全件を利用せず，ターゲットデータに似ているデータだけを利用すればよいことが分かる．このことから，ターゲットデータに十分似ていないデータを足しすぎると，誤った学習が行われてしまうことが推察できる．

また，Target Only とフィルタリングの削除の決定木のルートノードでは，「ターゲットデータ 10 件の MFS の，ターゲットデータに閾値以上似たソースデータ中のパーセンテージ=12.58 以下」である場合に，Target Only が割り当てられた．このことにより，ターゲットデータ 10 件中に最頻出する語義が，フィルタリングの削除の訓練事例として利用される，

「ターゲットデータに一定以上似ているソースデータ」に少ない時には，Target Only を用いた方がよいことが分かる．このことから，二つのデータのラベルが似ていないときは，ソースデータから訓練事例を一切足すことなく，ターゲットデータだけで学習した方がよいと考えられる．

８．まとめ

語義曖昧性解消（WSD; Word Sense Disambiguation）について領域適応を行った場合，ソースデータとターゲットデータのデータの性質により，最も効果的な領域適応手法が異なる．そのため本稿では，決定木学習を用いてソースデータとターゲットデータの性質から，最も効果的な領域適応手法を自動的に選択する手法について述べ，作成した決定木について考察した．

文　献

Hashida, K., Isahara, H., Tokunaga, T., Hashimoto, M., Ogino, S., and Kashino, W. (1998). The RWC text databases. In *Proceedings of The First International Conference on Language Resource and Evaluation*, pp. 457-461.

Komiya, K. Okumura, M. (2011). Automatic Determination of a Domain Adaptation Method for Word Sense Disambiguation Using Decision Tree Learning. In Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 1107-1115.

Maekawa, K. (2008). Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101-102.

McClosky, D., Charniak, E., and Johnson, M. (2010). Automatic Domain Adaptation for Parsing. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 28-36.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

国立国語研究所(1964). 分類語彙表．秀英出版．

西尾実，岩淵悦太郎，水谷静夫(1994). 岩波国語辞典第五版．岩波書店．

生成された決定木

上の枝が yes，下の枝が no に相当する．

```
        ┌──Target Only
    ┌──JS 距離（WSD の対象単語の一つ前の形態素）>=0.61
    │   └──Random Sampling
ふたつの正解率の比>=0.70
    └──Target Only
```

図 1　Target Only と Random Sampling の決定木

```
    ┌──フィルタリングによる削除
ソースデータ件数/ターゲットデータに一定以上似ているソースデータ件数>=186.85　|
    ┌──フィルタリングによる削除
    └──JS 距離（WSD の対象単語の二つ前の形態素）>=0.74
        └──Random Sampling
```

図 2　Random Sampling とフィルタリングの削除の決定木

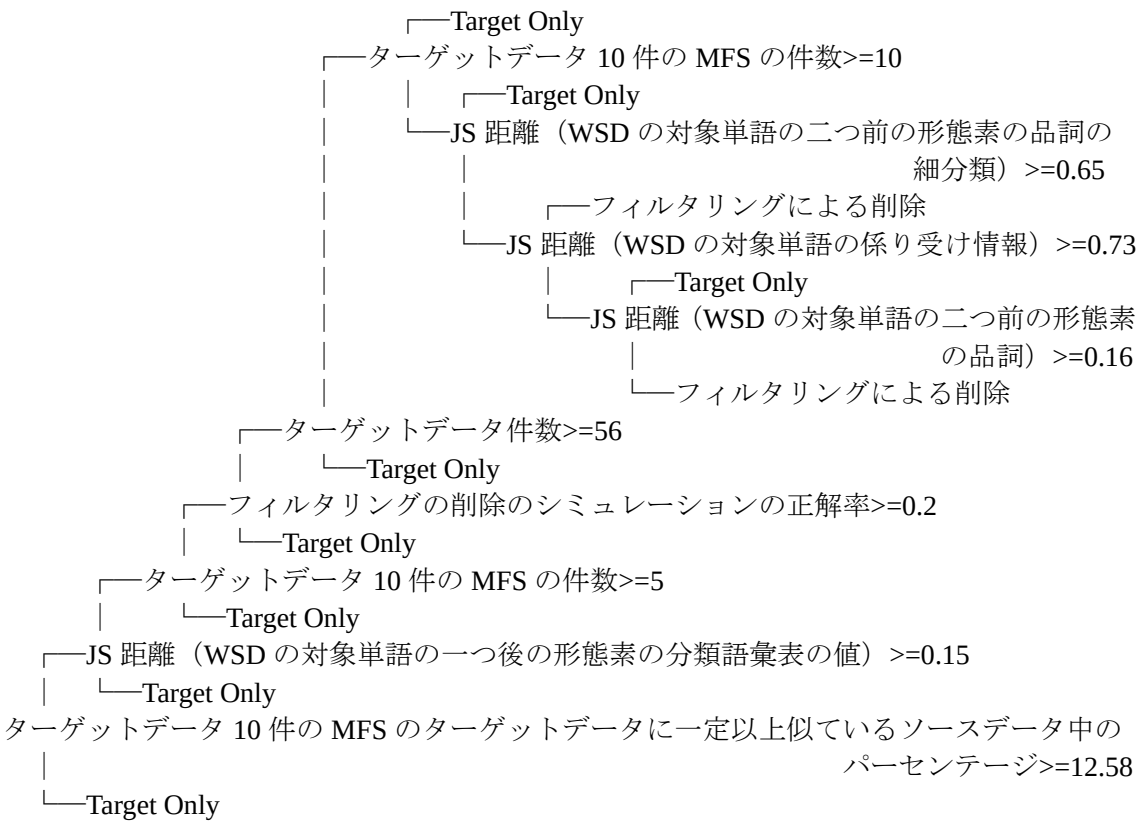


図 3　Target Only とフィルタリングの削除の決定木

形態素と文字の情報を用いた中国語形態素解析

侯 海霞（東京農工大学大学院 情報工学専攻）
古宮 嘉那子（東京農工大学　工学研究院　先端情報科学部門）
柴原 一友（テンソル・コンサルティング株式会社，東京農工大学）
藤本 浩司（テンソル・コンサルティング株式会社，東京農工大学）
小谷 善行（東京農工大学　工学研究院　先端情報科学部門）

Chinese Morphological Analysis Using Morphemes and Characters

Haixia Hou (Graduate School of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Kazutomo Shibahara (Tensor Consulting Co.Ltd., Tokyo University of Agriculture and Technology)
Koji Fujimoto (Tensor Consulting Co.Ltd., Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

形態素解析は自然言語処理の重要な基本技術の一つである。中国語や日本語などには分かち書きがないため、形態素解析は特に重要である。日本語の形態素解析は多く研究されているが、中国語の形態素解析の研究は比較的少ない。形態素解析において重要な課題となっているのは、未知語（辞書に登録されていない、あるいは訓練コーパスに出現されていない、形態素となり得る文字列）をどのように扱うかということである。

ここれまでの研究の多くは、内元(2001)や竹内(1997)などのような形態素の情報だけを使用していた。しかし、形態素の情報だけでは、未知語の扱いが困難であるという問題があった。これに対して、我々は中国語のほとんどは漢字によって書かれるという事実を踏まえ、文字情報を素性として用いることができると考えた。しかし、小田(1999)では文字の情報だけでは、既知語に対する精度が低いと証明した。そのため、我々は、形態素の情報と文字の情報を素性として使用する形態素解析の手法を形態素解析の手法を提案する。

本研究に最も近い研究は、中川哲治(2004)の形態素と文字の情報を併用した中国語の分割であるが、これは文を形態素に分解する部分だけを行い、形態素解析は行っていない。また、我々の手法では、形態素の情報と文字の情報に対して柔軟な素性設計の可能な最大エントロピー(ME)モデルを使用し、品詞遷移や文字位置などの情報を用いてコーパスから未知語の性質を学習した。人民日報のタグ付きコーパスを使用して実験を行った結果、素性として形態素と文字の両方を使用した手法は形態素だけ使用した手法より高い解析精度が得られた。

２　素性の学習

素性の学習には、Berger (1996)の ME モデルを用いる。このとき、ある形態素 x_i がある品詞 y_i である確率 $p(y_i \mid x_i)$ を計算する。ME モデルにおける $p(y_i \mid x_i)$ の計算は

素性に依存する。素性は形態素解析に役に立つ情報によって定義され、素性関数の引数として利用される。素性関数は下記のように定義する。

$$f_{ijk}(x,y)=\begin{cases}1: & x\text{は情報}\ g_{ij}\text{持つ}\quad\text{かつ}\quad y=f_k\\ 0: & \text{その他}\end{cases}$$ 式２－１

x：着目している形態素

y：着目している形態素の品詞

素性(g_{ijk})：x は情報 g_{ij} が持つ　かつ　$y=f_k$

素性集合と訓練データが与えられた時、エントロピーを最大化にすることによって、モデルが生成される。全ての素性 g_{ijk} ごとにパラメータ λ_{ijk} を用い、Berger（1996）のような条件付き確率として表される。

$$p^*(y\mid x)=\frac{1}{Z(x)}e^{\sum_{ijk}\lambda_{ijk}f_{ijk}(x,y)}$$ 式２－２

$$Z(x)=\sum_{y}e^{\sum_{ijk}\lambda_{ijk}f_{ijk}(x,y)}$$ 式２－３

パラメータ λ_{ijk} を推定する際には、下記式２－４のように、訓練コーパスにおける全ての素性 g_{ijk} に対し、ME モデルから計算される（x,y）の確率が訓練コーパスにおいての（x,y）の出現確率と等しくなるようにする。p は訓練コーパスによって計算される確率である。

$$\sum_{x,y}\overline{p}(x,y)f_{ijk}(x,y)=\sum_{x,y}\overline{p}(x)p^*(y\mid x)f_{ijk}(x,y)$$ 式２－４

下記の式２－５で収束するまでパラメータ λ_{ijk} を更新しながら学習する。（括弧の部分が０になるのが一番よい）

$$\lambda_{ijk}^{(n+1)}=\lambda_{ijk}^{(n)}+c\left[\sum_{x,y}\overline{p}(x,y)f_{ijk}(x,y)-\sum_{x,y}\overline{p}(x)p^*(y\mid x)f_{ijk}(x,y)\right]$$ 式２－５

n：推定回数

c：学習率

３．形態素の素性と文字の素性

本論文では、形態素の情報である「形態素の連接品詞」、「形態素の表層」、文字情報である「文字が形態素においての位置」の３種類の素性を定義する。これ以降、形態素の連接品詞を「前品:バイグラム」、形態素の表層を「単品:表層」、文字位置を「文位:文字が形態素においての位置|文字」と表記する。最大エントロピー法では、ラベル（形態素解析器なので本研究では品詞に相当する）ごとに素性を作成する必要があるため、

最も単純な素性でも 46 種類の素性となる。以下に、素性を列挙する。

A.「前品:バイグラム」

本論文が利用しているコーパスにおいては、品詞は 46 種類あるため、合計で 2162（47 種類×46 種類）個の素性を利用する。

A-1　品詞バイグラム（46 種類×46 種類）

下記に例をしめす。

例：「前品：名詞」

「着目している形態素が名詞で、かつ、その直前の形態素の品詞が名詞である」ときに 1 になる素性。

A-2　文頭-品詞バイグラム（46 種類）

例：「前品：文頭」

「着目している形態素が名詞で、かつ、その直前の形態素の品詞が文頭である」ときに 1 になる素性。

B.「単品:表層」

辞書においては、形態素は 57760 個あるため、合計で 2656960（57760 種類×46 種類）個の素性を利用する。下記に例をしめす。

例：「単品:我」

「着目している形態素が名詞で、なおかつ「我」である」ときに 1 になる素性。

形態素の連接品詞と表層を素性としてのパラメータの格納の仕方を表 3-1 に示す。品詞は本論文の訓練コーパスにおいては４６種類があるが、ここでは、品詞を n:名詞、v:動詞、a:形容詞、d:副詞、w:記号の５種類を例にする。

表 3－1　形態素素性に対してのパラメータ

情報番号	情報名	f_1:n	f_2:v	f_3:a	f_4:d	f_5:w
g_{10}	前品:文頭	λ_{101}	λ_{102}	λ_{103}	λ_{104}	λ_{105}
g_{11}	前品:v	λ_{111}	λ_{112}	λ_{113}	λ_{114}	λ_{115}
g_{12}	前品:n	λ_{121}	λ_{122}	λ_{123}	λ_{124}	λ_{125}
…	…	…	…	…	…	…
g_{21}	単品:我	λ_{211}	λ_{212}	λ_{213}	λ_{214}	λ_{215}
g_{22}	単品:是	λ_{221}	λ_{222}	λ_{223}	λ_{224}	λ_{225}
g_{23}	単品:学生	λ_{231}	λ_{232}	λ_{233}	λ_{234}	λ_{235}
…	…	…	…	…	…	…

「前品: 文頭」:現在着目している文字列が文頭である
「前品: v」:直前前の品詞が v である
「前品:n」:直前前の品詞が n である
「単品:我」: 着目している文字列が「我」である
「単品:是」:着目している文字列が「是」である
「単品:学生」:着目している文字列が「学生」である
「品詞」の種類: n:名詞, v:動詞, a:形容詞, d:副詞, w:記号 という 5 種類を例にしている
「品詞」の種類:全部で 46 種類がある
λ_{101}: 情報 g_{10} を持っている文字列が n という品詞になる重みである
λ_{111}: 情報 g_{11} を持っている文字列が n という品詞になる重みである
λ_{ijk}: 情報 g_{ij} を持っている文字列が f_k 品詞になる重みである
λ_{ijk}: 前品に対して 2162 個になる。単品に対して 2656960 個になる

C.「文位:文字が形態素においての位置|文字」

文字の形態素における位置の表記に[7]の表記方法を用いる。この表記を表 3－2 に示す。

表 3－2　文字が形態素においての位置

表記	意味
S	形態素が 1 文字である文字
B	形態素の先頭にある文字
I	形態素の中間にある文字
E	形態素の末尾にある文字

辞書において、文字は 12977 個あるため、合計で 596942 個の素性を利用する。下記に例をしめす。

例 1:「文位: S｜我」

「着目している形態素が名詞で、なおかつ「我」という 1 つの文字である」ときに 1 になる素性。

例 2:「文位: B｜我」

「着目している形態素が名詞で、なおかつ「我」という文字であり、形態素の先頭になる」ときに 1 になる素性。

例 3:「文位: I｜我」

「着目している形態素が名詞で、なおかつ「我」という文字であり、形態素の中間にある（先頭でも末尾でもない)」ときに 1 になる素性。

例 4:「文位: E｜我」

「着目している形態素が名詞で、なおかつ「我」という文字であり、形態素の末尾になる」ときに 1 になる素性。

文字情報である「文位:文字が形態素においての位置|文字」を素性としてのパラメータの格納の仕方を表３-３に示す。

表３－３　文字素性に対してのパラメータ

情報番号	情報名	f_1:n	f_2:v	f_3:a	f_4:d	f_5:w
g_{31}	文位:S｜我	λ_{311}	λ_{312}	λ_{313}	λ_{314}	λ_{315}
g_{32}	文位:S｜是	λ_{321}	λ_{322}	λ_{323}	λ_{324}	λ_{325}
g_{33}	文位:B｜学	λ_{331}	λ_{332}	λ_{333}	λ_{334}	λ_{335}
g_{34}	文位:E｜生	λ_{341}	λ_{342}	λ_{343}	λ_{344}	λ_{345}
g_{35}	文位:I｜一	λ_{351}	λ_{352}	λ_{353}	λ_{354}	λ_{355}
…	…	…	…	…	…	…

「文位:S｜我」：着目している文字列が「我」という１つの文字である
「文位: B｜学」：着目している文字列の先頭が「学」という文字である
「文位: E｜生」：着目している文字列の末尾が「生」という文字である
「文位:I｜一」：着目している文字列の中間に「一」という文字がある
「品詞」の種類: n:名詞, v:動詞, a:形容詞, d:副詞, w:記号 という５種類を例にしている
「品詞」の種類:全部で 46 種類がある
λ_{311}: 情報 g_{31} を持っている文字列が n という品詞になる重みである
λ_{312}: 情報 g_{31} を持っている文字列が v という品詞になる重みである
λ_{ijk}: 情報 g_{ij} を持っている文字列が f_k 品詞になる重みである
λ_{ijk}:全部で 596942 個になる

４．形態素と文字の情報を用いた中国語形態素解析システム

すべての文字列を形態素としてシステムを実行するには時間かかるため、本研究では態素解析候補にする文字列の長さは、あらかじめ 5 文字までとし、辞書に入ってない５文字以上の文字列を考慮しないようにした。これは、中国語において、５文字以上の未知語の形態素が非常に少ないためである。本論文の訓練コーパスにおいて、６文字以上の形態素は僅か 0.62%であった（**表**４－１）。

また、未知語が入っている文章を解析できるようにするため、辞書にある文字列だけではなく、長さが５文字以内の文字列は全て形態素候補とした。この際、形態素候補には全種類の品詞を付けて展開する。ここで、形態素候補と品詞のセットをノードと呼ぶ。

表４-１　訓練コーパスにおいての形態素の長さ

形態素の長さ	出現確率
1 文字	47.34%
2 文字	44.73%
3 文字	4.46%
4 文字	2.12%
5 文字	0.73%
6 文字以上(6 文字含む)	0.62%
形態素の総数：1083411	

また、高速化のため、句読点などのマークは、周り見ずにそのマークを直接一つの形態素として扱った。

作成した形態素候補をリンクで繋いで、ラティスを作成する。この様子を図４－１に示す。ここでは、ノード全てを図に表示するのが困難であるため、一部のみ表示している。

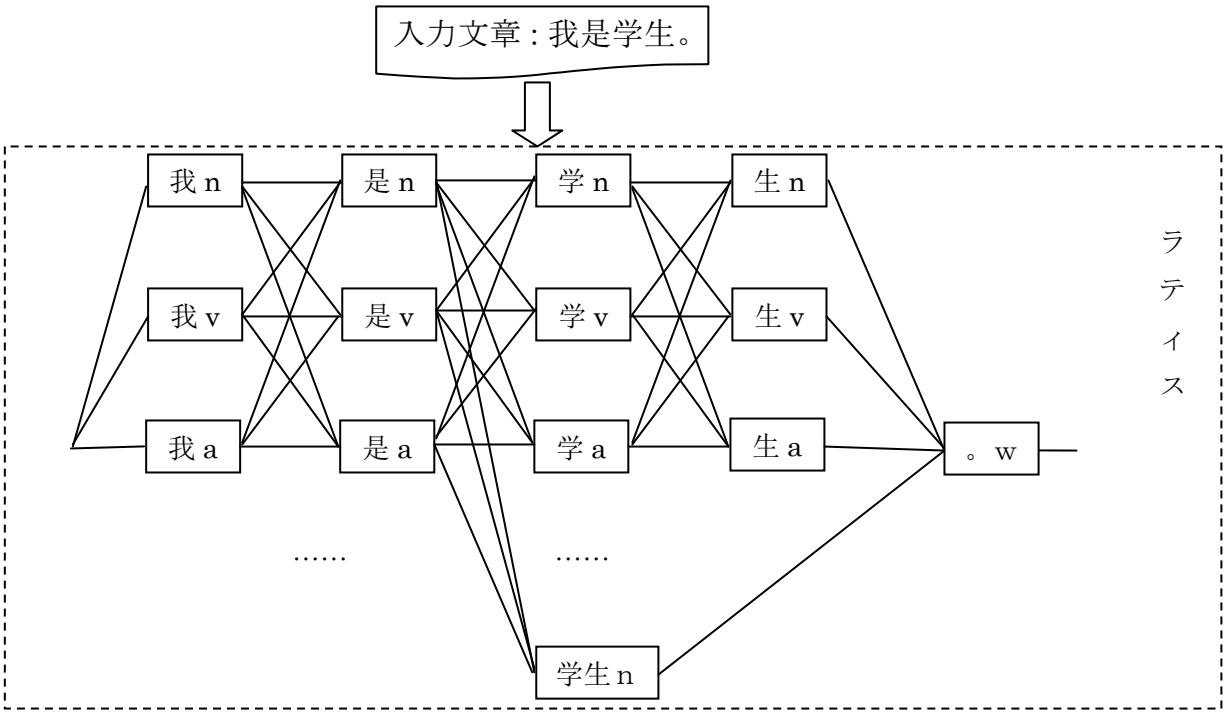


図４－１　ノードセットとラティス

ラティスにあるすべてのルートの確率を ME モデルで計算し、そのうち、確率が最大になっているルートを結果として出力する。

５．実験

形態素と文字の情報を用いた中国語形態素解析を実装し、人民日報、富士通と北京大学によるタグ付きコーパス(人民日報タグ付きコーパス)を使って、形態素と文字の情報を用いた手法と、形態素だけ使用した手法の比較を行った。

人民日報タグ付きコーパスからランダムで取り出した1541文をテストコーパスとして、残りは訓練コーパスとして利用した。人民日報タグ付きコーパスの分類を表５－１にて示す。

表５－１　人民日報タグ付きコーパスの分類

総数	合計	訓練 コーパス	テスト コーパス
文	46, 251	44, 710	1, 541
形態素	1, 083, 411	1, 048, 121	35, 290

解析精度の評価を行うために、Closed テストと Open テストを行った。このテストデータとして closed テストデータと open テストデータをそれぞれ作成する。closed テストデータは訓練コーパスからランダムで取りし、open テストデータは訓練コーパスから分けられたテストコーパスからランダムで取り出す。

上述した形態素と文字の情報を用いた中国語形態素解析の手法で closed テストと open テストをそれぞれ 3 回行ったため、テストデータは毎回 300 文をランダムに 3 回取り出して作成した。表５－２の形態素の数は 3 回のデータの平均値である。

表５－２　Closed test と Open test によるテストデータ

数	Closed テストデータ	Open テストデータ
文	300	300
形態素	7,025	6,870

６．評価

上述した実験の結果を表６-1 にて示す。確率は 3 回実験において得られた確率の平均値である。

表６－１　実験結果

素性	Closed テスト		Open テスト	
	適合率	再現率	適合率	再現率
形態素だけ	96.1%	95.4%	83.71%	89.2%
形態素と文字	**96.1%**	**95.9%**	**90.31%**	**93.2%**
文字情報効果	**0**	**0.5**	**6.6**	**4**

表６－１から、未知語のない closed データでの実験では、文字情報の効果はほとんどないことが分かる。また、同じ表から、未知語のある open データでの実験では、文字情報の効果は明らかである。適合率において 6.6 ポイント、再現率において 4 ポイントを上がっている。以下に具体的な例をとって考察する。

例１、未知語である「细致」

形容詞である「细致」は辞書と訓練コーパスにないが、形容詞である「细心」「细微」「细嫩」「别致」「雅致」はあったため、「文位:B|细」「文位:E|致」との２つの情報を用いて、コーパスから「细致」を学習することができた。

例２、未知語である「圆润」

形容詞である「圆润」は辞書と訓練コーパスにないが、形容詞である「圆满」「圆浑」「滋润」「红润」「湿润」はあったため、「文位:B|圆」「文位:E|润」との２つの情報を用いて、コーパスから「圆润」を学習することができた。

７．おわりに

本論文は、形態素と文字の情報を用いた中国語形態素解析の手法を提案した。実験により、全て漢字により書かれている中国語は形態素解析を行う場合には、形態素の情報も文字の情報も有効であることが分かった。適合率が 6.6 ポイント、再現率が 4 ポイントが上がった。未知語の処理に文字の情報が役に立っていることが分かった。

謝辞

本論文の作成にあたり、人民日報のコーパスを利用させていただきました。人民日報のコーパスを作成した各社に感謝いたします。

文献

Berger,A.L.,Pietra,S.A.D.,and Pietra,V.J.D.,A(1996)「Maximum Entropy Approach to Natural Language Processing」Computational Linguistics,22(1),pp.39-71.

内元清貴、関根聡、井佐原均(2001)「最大エントロピーモデルに基づく形態素解析―未知語の問題の解決策―」,自然言語処理, Vol.8:NO.1,pp.127-142,

竹内孔一、松本裕治(1997)「隠れマルコフモデルによる日本語形態素解析のパラメータ推定」情報処理学会論文誌,Vol.38:NO.3,pp.500-509.

小田裕樹、森信介、北研二(1999)「文字クラスモデルに基づく日本語単語分割」情報処理学会研究報告,99-NL-130,pp.1-8.

中川哲治、松本裕治(2004)「単語レベルと文字レベルの情報を用いた中国語・日本語単語分割」 自然言語処理研究会報告 2004(73),pp.197-204.

Web 関連度と確率的翻訳モデルを併用した質問応答システム

阿部 裕司（東京農工大学大学院 情報工学専攻）
森田 一（東京工業大学 知能システム科学専攻）
古宮 嘉那子（東京農工大学大学院 工学研究院）
小谷 善行（東京農工大学大学院 工学研究院）

Question Answering System Using Web-Relevance and Probabilistic Translation Model

Yuji Abe (Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology)
Hajime Morita (Department of Computational Intelligence and System Science, Tokyo Institute of Technology)
Kanako Komiya (Institute of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institute of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

質問応答タスクは、与えられた自然文の質問に対し回答自体を検索して出力するタスクである。質問応答における回答候補評価モジュールは、文書検索で取得した回答候補に対して詳細な分析・評価を行なうもので、質問応答システムの回答性能に直結する重要な要素である。回答候補は「内容の関連度（質問文に対しどれだけ近い内容が記述されているか）」と「記述の回答らしさ（質問文の記述形式に対応した記述が含まれているか）」によってスコア付けされる。

本稿では、既存の回答候補評価指標二つを統合する候補評価式を用い、より柔軟に回答候補を評価する手法を提案する。

２．既存研究とその理論

石下(2009)では「内容の関連度」を、Web を適用した擬似適合フィードバックによって評価している。これは、質問文の内容語をクエリにして文書検索を行なった際に検索文書内に頻出する語を関連語とみなし、その関連語を多く含む回答候補は「内容の関連度」が高いとみなすものである。

「内容の関連度」と「記述の回答らしさ」を同時に判定する方法として、Soricut(2006)は確率的翻訳モデルを利用した手法を提案している。この手法では、質問文を翻訳前の文、回答文を翻訳後の文とみなし、個々の単語ごとの翻訳確率を QA コーパスから学習する。そして新しい質問が入力された際に、その質問が回答候補文に翻訳される確率を利用して回答候補評価をおこなう。翻訳モデルとしては、単純だが多くのタスクで有効性が確認されている IBM-Model1(Brown(1993))を改変したものを利用している。質問応答タスクにおいて、本来の IBM-Model1 を用いた定式化は式(4.1),(4.2)のようになる。

$$A^* = \arg\max_A p(A \mid Q) = \arg\max_A p(Q \mid A)p(A) \tag{4.1}$$

$$p(Q \mid A) \approx \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(q_j \mid a_i) \tag{4.2}$$

式(4.1),(4.2)において、A^*は最も適切と思われる回答候補文、A $(=a_1, a_2, ..., a_l)$は回答候

補文、$Q(=q_1,q_2,...,q_m)$は質問文、lは回答候補文の単語数、mは質問文の単語数、$p(q|a)$は質問側の単語aが回答側の単語qに翻訳される確率、$p(A)$は回答候補文Aの生成確率、εは回答文から単語数mの質問文が生成される確率である。式(4.2)において、相乗記号の外側の係数はlが小さいほど大きな値となり、したがって回答候補文の単語数が少ないほど文全体の翻訳確率が大きくなってしまうという問題がある。そのため、Soricut(2006)では、式(4.2)の相乗記号の外側の係数を無視した定式化を行なっている。

$$p(Q|A) \approx \prod_{j=1}^{m} \sum_{i=0}^{l} p(q_j|a_i) \tag{4-3}$$

３．既存手法を統合した回答候補評価

「内容の関連度」を評価する際、Soricut(2006)のように回答候補評価に確率的翻訳モデルを用いた場合は同義語や類義語などを柔軟に考慮した評価が可能となるが、複数の語彙に対する共起関係などの情報は利用することができない。一方、石下(2009)のように Web 等を利用した擬似適合フィードバックを用いた場合、複数の語彙に対する共起関係をうまく利用することができるが、類義語などを柔軟に考慮することは困難である。これらを統合的に利用することで、回答候補評価の性能向上が期待できる。

３．１．回答候補評価式

式(4.1)において最大化されるべき部分を$\wp(Q,A)$とおき、石下(2009)が提案する Web 上の内容関連度に基づく回答候補スコアを$Web_relevance(Q,A)$とおく。両スコアを組み合わせた最終的な評価式を次式で定義する。

$$EvalScore(Q,A) = \wp(Q,A)^{1-\gamma} \cdot Web\_relevance(Q,A)^{\gamma} \tag{4-4}$$
$$\wp(Q,A) = p(Q|A)p(A) \tag{4-5}$$

上式において、γは手法の混合比を決める混ぜ合わせパラメータである。$\gamma = 0$ならば翻訳確率単体の評価値、$\gamma = 1$ならば Web 上の内容関連度単体の評価値となる。

３．２．Web 上で評価する内容の関連度の計算

石下(2009)と同様の手法を採用する。まず、入力された質問文から内容語(名詞,動詞,形容詞)を取得し、キーワード集合Kとする。次に、Kに含まれる語の三つ組を全通り作り、それぞれの三つ組から論理積検索のクエリを構成し、Web 検索エンジンにおいて検索をする。$|K|<3$の場合は全ての語の論理積検索を構成する。そして、それぞれのクエリに対してスニペット(Web 検索エンジンが出力した、検索結果ページの要約)を最大上位 100 件取得し、このスニペット内の各内容語w_jを関連語とする。各関連語の関連度$T(w_j)$は次式で定義する。

$$T(w_j) = \max_i \frac{freq(w_j,i)}{n_i} \tag{4-6}$$

iはクエリ(内容語の三つ組)の番号、n_iはi番目のクエリを利用して取得したスニペットの件

数、 $freq(w_j, i)$ は i 番目のクエリで取得したスニペットのうち、w_j を含むものの件数である。Web を利用した内容関連度による回答候補の評価値 $Web\_relevance(Q, A)$ を、回答候補に含まれる語の関連度の総和で定義する。

$$Web\_relevance(Q, A) = \sum_{i=1}^{l} T(w_i) \qquad (4\text{-}7)$$

上式において、Qは質問文、A は回答候補文、l は回答候補文の単語数、w_i は回答候補文に含まれる単語である。

３．３．翻訳確率の学習

学習コーパスとして『Yahoo!知恵袋データ』を用い、翻訳確率の学習ツールとして IBM-Model1 の C++実装である GIZA++[Casacuberta07]を用いた。学習時の EM アルゴリズムのイテレーション回数は 5 回に設定した。学習時に、単語数の多すぎる質問応答事例、質問側と回答側の単語数に差がありすぎる質問応答事例は単語アラインメントの学習に悪影響を与えると考え、質問側または回答側の単語数が 60 語を超える事例、および回答側と質問側の単語数に 5 倍以上の差があるものは学習から除外した。結果として、質問応答事例 1,092,144 件を利用して翻訳確率を学習した。

４．評価実験

提案する回答候補評価式を用いた質問応答システムを実装し、質問応答実験を行なった。

４．１．実験内容

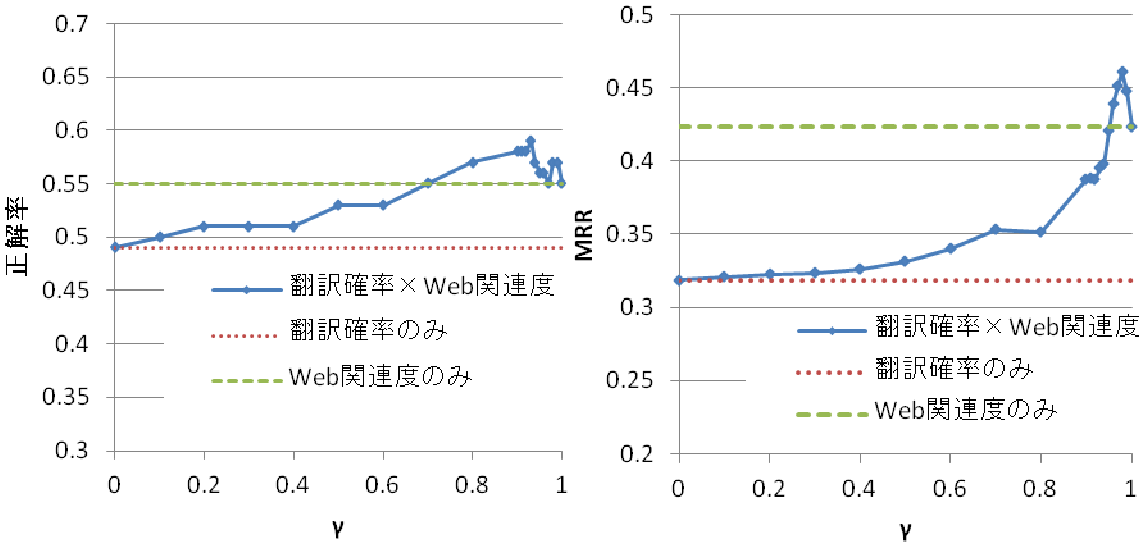
『NTCIR-ACLIA2』(Teruko(2010))の質問応答テストセット 100 問に対し、質問応答実験を行なった。関連文書検索、回答候補抽出は Web 文書を対象として行ない、質問文に含まれる内容語で論理積検索を行ない抽出された文書の出力順位上位 50 件を関連文書とした。抽出された回答候補すべてに対し式(4-4)に基づいた回答候補評価を行ない、スコアの高いものから 5 件を最終的なシステム出力とした。$p(A)$ の計算には通常のバイグラムモデルを単語数で正規化したものを用いた。システムが出力した回答が正解かどうかの判定は人手で行ない、システム性能の評価指標として、Top-5 正解率と MRR を用いた。

４．２．実験結果

γ の値を[0:1]の範囲で変化させていったときの、Top-5 正解率および MRR を図 1 に示す。また、既存手法と提案手法の性能比較を表 1 に示す。γ =0.93 のとき正解率は最大値の 0.59 を示し、γ =0.98 のとき MRR は最大値の 0.461 を示した。ウィルコクソンの符号付順位和検定を行なったところ、MRR については 5%有意水準で既存手法より提案手法が有意に優れていることが示された。

４．３．まとめと展望

本稿では、既存の回答候補評価指標を組み合わせて、「内容の関連度」をより柔軟に評価する手法を提案した。その結果、提案手法は各既存手法よりも優れた質問応答性能を示した。現行のシステムでは「記述の回答らしさ」に関しては十分に評価できていないので、語彙パターンなどを用いて「記述の回答らしさ」を評価する指標を導入することが、今後の課題である。



(a)Top-5 正解率　　(b)MRR

図１：提案手法の質問応答実験結果

表１：提案手法と既存手法の性能比較

	Top-5 正解率	MRR
既存・翻訳確率(γ=0)	0.49	0.318
既存・Web 関連度(γ=1)	0.55	0.423
提案手法(γ=0.93)	**0.59**	0.395
提案手法(γ=0.98)	0.57	**0.461**

謝　辞

本研究を行なうにあたり、ヤフー株式会社が国立情報学研究所に提供した『Yahoo!知恵袋データ』を利用させて頂きました。利用を快諾して下さった各社に感謝いたします。また、評価実験の際に NTCIR の質問応答テストコレクション『NTCIR-8 ACLIA2』を利用させて頂きました。NTCIR の運営にご尽力をいただいている皆様に感謝いたします。

文　献

石下円香、佐藤充、森辰則(2009)「Web 文書を対象とした質問の型に依らない質問応答手法」人工知能学会論文誌,24 巻 4 号, pp.339-350.

Radu Soricut, Eric Brill(2006)”Automatic question answering using the web: Beyond the factoid” Journal of Information Retrieval - Special Issue on Web Information Retrieval, 9,pp.191-206.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer (993) “The Mathematics of Statistical Machine Translation: Parameter Estimation” Computational Linguistics, 19(2), pp.263-311(1993).

Teruko Mitamura, Hideki Shima, Tetsuya Sakai, et al(2010) ”Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access” In Proceedings of 8th NTCIR Workshop Meeting.

文の長さ分布に見られる対数正規性

古橋　翔（東北大学大学院理学研究科）†

Lognormality of the Distribution of Sentence Length

Sho Furuhashi (Graduate School of Science, Tohoku University)

１．はじめに

言語学の一分野である計量文献学では、統計的手法により文献の分析が行われている。その際に着目する量として、単語の長さ、品詞の使用率や句読点の打ち方などがあり、文の長さ（文長）もその一つである。文長は英語など多くの言語で調べられており、日本語でも調べられている。その結果、文長分布は、文字数でみた場合、安本(1958)、佐々木(1976)と新井(2001)らにより対数正規分布、また佐々木によってガンマ分布の場合もあると報告されている。一方で、Ishida, Ishida(2007)により形態素数の場合は Hyper Pascal 分布であると報告されている。

文長分布の特徴に対する考察として、安本は、文長分布の対数正規性は Weber-Fechner の法則に依るのではないかと提案し、佐々木は、対数正規分布のモデルの一例として Kaptyen のアナログマシンの例を挙げている。また、ガンマ分布に対しては、佐々木は、文の構成要素の長さが指数分布に従い、この指数分布のたたみこみ分布であるから文長分布はガンマ分布となるのではないかと考察している。しかしながら、このような考察の一方で計量的な研究は行われていない。

対数正規分布は、文長分布以外でも多くの自然現象や社会現象で見られる。例えば、落下によるガラス破片のサイズ分布や論文の発表数の分布がある。近年、物理学の分野において対数正規分布が注目されるようになり、その生成メカニズムの研究が行われている。本研究では文長分布の対数正規性に着目し、この性質を既存のモデルにより説明できないか試みた。

２．サンプル

本研究では、インターネット上にある著作権の切れた作品を収蔵している青空文庫と京都大学大学院情報学研究科黒橋・河原研究室が提供している京都大学テキストコーパス Version 4.0 を利用した。これらの資料から文を収集するのだが、京都大学テキストコーパスはあらかじめ文ごとに区切られているが、青空文庫は独自に文章を文に分割しなければならない。本研究では、青空文庫の作品から次のような処理により文を収集した。

まず、クローラーを使用してインターネット上の青空文庫から出来る限り多くの作品のテキストファイルを収集する。この時、文字コードを Shift-JIS から UTF-8 へと変換した。次にファイルの最初と最後に作品情報が記載されているのでそれを削除する。この処理を行った後、次のルールに従いテキストファイルを選別する。

- ▲や□などの記号を含んでいない。
- 日本語のみで書かれている。
- 括弧の開閉の数が一致している。
- 詩、俳句などの韻文を含んでいない。
- 箇条書きを含んでいない。章立てになっていない。
- 脚本になっていない。

これらの条件を満たしていない作品は除外した。

残ったファイルを更に加工する。まず、青空文庫に収蔵されている作品は、電子化するに

† furuhashi@cmpt.phys.tohoku.ac.jp

当たり原本の情報を残すため注釈が書かれているので、これを取り除く。次に、踊り字をそれが表している文字に置き換えた。この置換は、形態素解析器などを利用するに当たり、踊り字のままだと誤りが多くなると考えたからである。以上のような処理を行った後、文章を句点で区切っていき文にばらしていった。最後に、得られた文を次のルールに従い選別した。

- カタカナと句読点のみの文ではない。
- 日本語の文字と句読点のみで構成されている。
- 文の長さに制限はない。

得られたサンプルは、青空文庫では、作品ファイルは 2213、著者は 150 名以上で、文の数は 116719、京都大学テキストコーパスでは、文の数は 38397 である。

３．一文当たりの文字数の分布

まず始めに、一文当たりの文字数l_cの分布を調べた。図 1 がその分布であり、平均 42.9 標準偏差 32.9 であった。但し、調べたのは青空文庫のみである。京都大学テキストコーパスは、形態素・構文情報のみ公開しており、元となる毎日新聞データは含まれていないからである。

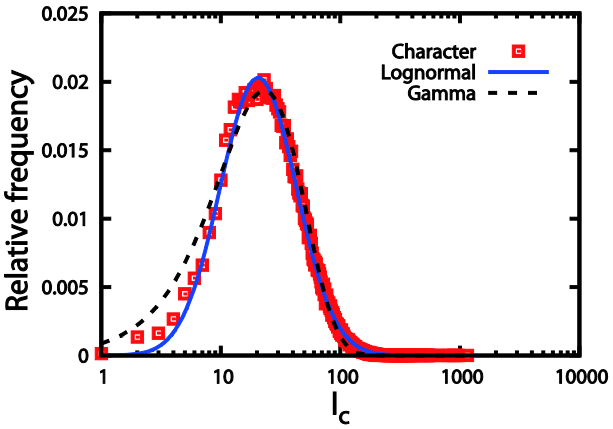


図 1　一文当たりの文字数l_cの分布（青空文庫）

図 1 の分布型が先行研究で報告されていた対数正規分布

$$f_{LN}(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}, \tag{1}$$

とガンマ分布

$$f_G(x)=\frac{x^{k-1}}{\Gamma(k)\theta^k}\exp(-\frac{x}{\theta}), \tag{2}$$

のどちらに近いかを調べた。まず Gnuplot のフィッテイング機能を利用し、パラメータの値を推定した（表 1）。

表 1 パラメータ値（文字数）

	μ	σ^2	k	θ
青空文庫	3.557 ± 0.003	0.536 ± 0.003	2.40 ± 0.02	16.3 ± 0.2

次に、実データとフィッティングで推定したパラメータによる分布関数との差の累積を計算した。

$$R = \sum_{x=x_0}^{x_{\max}} |O(x) - E(x)|, \tag{3}$$

$O(x)$ は実データの長さ x である文の相対頻度、$E(x)$ は当てはめた分布関数である。x が大きい領域ではデータ点はまばらであり揺らぎが大きい。その影響を除くために、x の範囲はデータ点が多い領域に制限した。$x_0 = 1$、$x_{\max} = 242$ として、$E(x) = f_{LN}(x)$ の場合、$R = 0.061$ 、$E(x) = f_G(x)$ では $R = 0.096$ であった。よって、青空文庫から収集した文は、一文当たりの文字数 l_c の分布は対数正規分布に近い。

４．文の構造

日本語の文構造は、文節間の係り受け関係を表した依存構造木（図 2）で表現できる。依存構造木は、文節をノードとして係り元から係り先へ矢印を張り表現される。（矢印の向きを反対に書く場合もある）

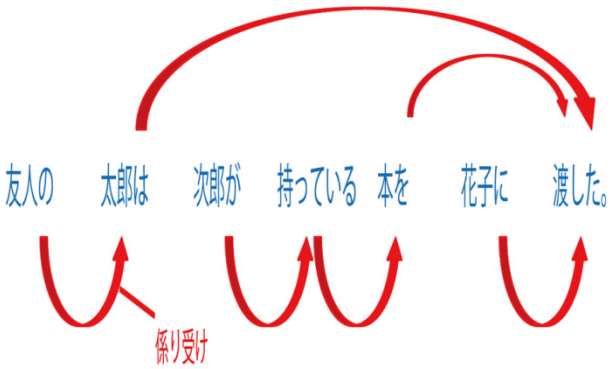


図 2　依存構造木

本研究では、依存構造木に着目して文長分布の対数正規性を生むメカニズムを調べた。京都大学テキストコーパスは既に依存構造木の情報が与えられているが、青空文庫には与えられていない。そのため、青空文庫の文に対して、依存構造木の情報を得るために、形態素解析に MeCab 0.98（辞書は MeCab-Ipadic）を用いた日本語係り受け解析器 CaboCha 0.60 pre4（TinySVM と YamCha なし）を使用した。

５．一文当たりの文節数の分布

依存構造木の構成単位が文節なので、一文当たりの文節数 l_s の分布を調べた（図 3, 4）。

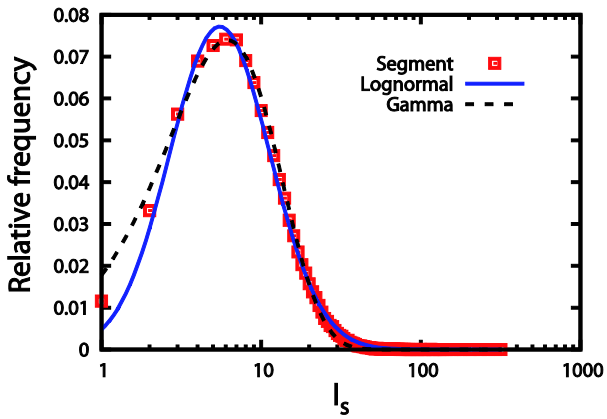


図 3　一文当たりの文節数 l_s の分布（青空文庫）

青空文庫では、平均が 11.1、標準偏差は 8.49、京都大学テキストコーパスの平均が 9.69、標準偏差は 5.27 であった。長さの単位を文字から文節へ切り替えたことで、文長分布型が変化するかどうか確かめるために、文字数の場合と同様に、対数正規分布 f_{LN} とガンマ分布 f_G のどちらがより当てはまるか R の値で比較した。フィッテイングにより得られたパラメータ値を表 2、R の値を表 3 に示す。

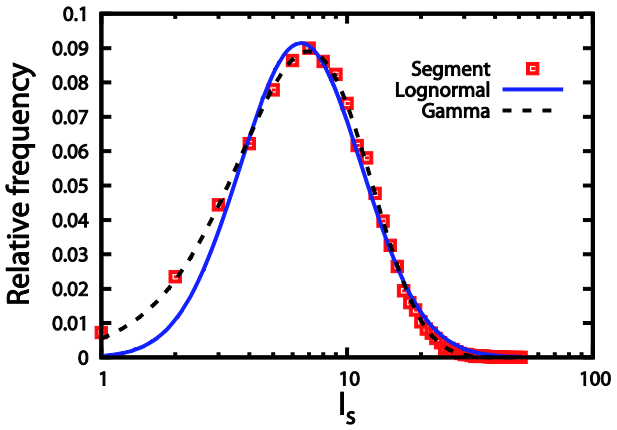


図 4　一文当たりの文節数 l_s の分布（京都大学テキストコーパス）

表 3 より、青空文庫では文字数の場合と異なりガンマ分布と対数正規分布では差が無く、京都大学テキストコーパスではガンマ分布の方が当てはまった。したがって、長さの単位を文節にすると、分布型はガンマ分布に近くなり、対数正規性は小さくなる。

表 2　パラメータ値（文節）

	μ	σ^2	k	θ
青空文庫	2.227±0.005	0.522±0.006	2.45±0.02	4.22±0.05
京都大学テキストコーパス	2.19±0.01	0.32±0.01	3.57±0.03	2.70±0.02

表 3　R の値（文節）

	青空文庫 （$x_0=1$ ，$x_{\max}=88$）	京都大学テキストコーパス （$x_0=1$ ，$x_{\max}=46$）
対数正規分布	0.071	0.11
ガンマ分布	0.070	0.024

６．乗算過程

対数正規分布を生み出すモデルの一つに乗算過程

$$X_n = \alpha_{n-1} X_{n-1} = \prod_{i=0}^{n-1} \alpha_i X_0, \qquad (4)$$

がある。変数 X_n は、ある分布に従う確率変数 α_i を独立に n 回掛け合わせて作られる。X_n を作る試行を多く繰り返すと、n が十分大きければ中心極限定理より X_n は対数正規分布に従う。

７． 依存構造木の枝分かれ過程

本研究では、乗算過程が文中の係り受け過程に表れるのではないかと考えた。係り受け過程を乗算過程と比較するために、依存構造木を図 5 のように書き換えた。図 5 は、葉である文節から係り受け関係に従い文節をまとめていき、最後に根である文になる過程を表している。

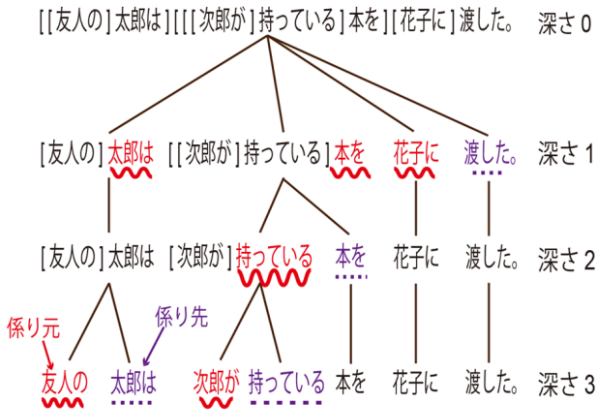


図 5　依存構造木

この依存構造木の作成方法は以下の通りである。

i. 他の文節から係られない文節の集合U を作る。
ii. 文節$b_j \in U$ に対して、その係り先c_j に係る文節が全てU の要素であるか確認する。
iii. ii. が確認された場合、c_j とそれに係る全ての文節をまとめてb_k とする。b_k はc_j の係り受け情報を引き継ぐ。U からc_j とそれに係る全ての文節を削除しb_k を追加する。
iv. U の要素数が一つになるまで ii と iii を繰り返す。

この依存構造木の枝分かれによるノード数増加の過程が乗算過程になっているのではないかと考えた。依存構造木の深さd におけるノード数をS_d として、もし枝分かれ過程が乗積過程であれば、式 4 より$\langle \ln X_n \rangle \propto n$ なので、$\langle \ln S_d \rangle$ はd に比例するはずである。よって、$\langle \ln S_d \rangle$ のd に対する変化を調べた。その際、依存構造木の葉の深さd_l が文ごとに異なる点に注意して、$S_d = l_s (d \geq d_l)$ とした。

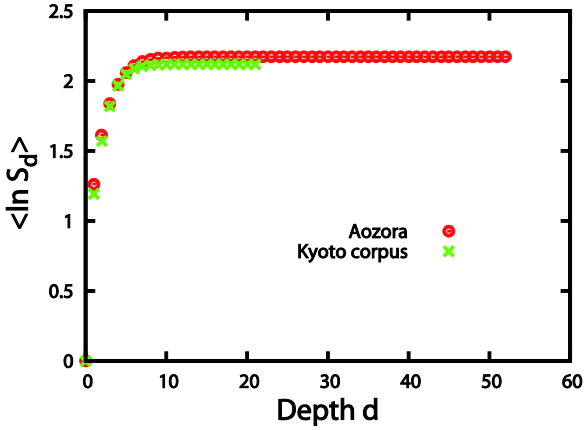


図 6　$\langle \ln S_d \rangle$ の深さd に対する変化

図 6 より青空文庫と京都大学テキストコーパスともに、$\langle \ln S_d \rangle$ はd に比例していなかっ

た。また、文節数l_sとd_lの関係を調べたところ、図６同様に比例関係になってはいなかった（図７）。よって、依存構造木からみた文構造に乗算過程は見られなかった。

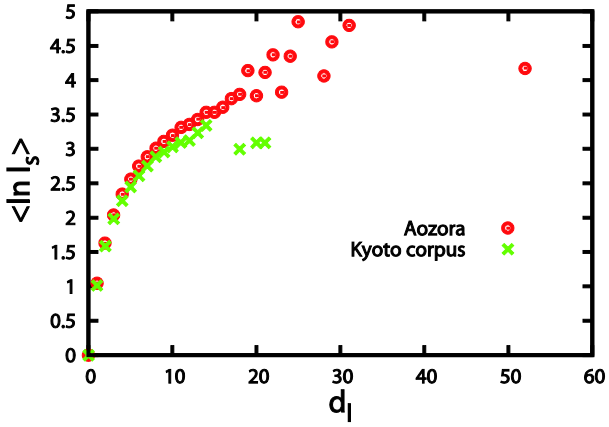


図 7　一文当たりの文節数の自然対数$\ln l_s$の平均と
依存構造木の葉の深さd_lの関係

８．まとめ

文長分布で報告されてきた対数正規性は、文構造に乗算過程が潜在しているのが原因ではなかった。今後は、一文当たりの文節数分布がガンマ分布に近いという結果から、佐々木の考察により文構造を説明できるか確かめるとともに、新たな視点から文長分布の対数正規性を研究していく。

文　献

安本美典(1958)「文の長さの分布型について」計量国語学, 4 号, pp.20-24.

佐々木和枝(1976)「文の長さの分布型」計量国語学, 78 号, pp.13-22.

新井　皓士(2001)「文長分布の対数正規分布性に関する一考察 ： 芥川と太宰を事例として」一橋論叢, 125 号 3 巻, pp.205-223. (http://hermes-ir.lib.hit-u.ac.jp/rs/handle/10086/10418 よりダウンロード可能)

Motohiro Ishida and Kazue Ishida (2007) On distributions of sentence lengths in Japanese writing, *Glottometrics,* 15, pp. 28-44.

関連 URL

Cabocha/南瓜　http://code.google.com/p/cabocha/

MeCab　http://mecab.sourceforge.net/

京都大学テキストコーパス Version 4.0　http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス

言語接触の観点からみた非有生名詞主語の「見る」構文
－ 文語体コーパスを利用して －

高橋 暦（名古屋大学大学院国際言語文化研究科博士課程）†
堀江 薫（名古屋大学）

Inanimate Subject Construction of Miru (‘see’) in Japanese from a Perspective of Language Contact
- Through the Use of a Literary Style Corpus –

Koyomi Takahashi (Graduated School of Languages and Cultures, Nagoya University)
Kaoru Horie (Nagoya University)

１．はじめに

言語接触（language contact）は語彙・文法の両面で言語の構造に大きな影響を与えることが広く知られている（Thomason and Kaufman 1989, Heine and Kuteva 2002）。日本語においては、言語接触は主として語彙の借用現象として観察され、古くは漢語、近代以後は英語から夥しい数の語を借用している（Loveday 1996）。

日本語における「動詞」の借用現象は、ドナー側の言語（例：英語）における動詞を一旦「名詞」として受けいれ、然る後に「する」を付与するといった形で日本語の語彙体系の中に組み入れていくというプロセスを経る。

(1) cut（英語動詞） ➔ カット（名詞） ➔ カット<u>する</u>（日本語動詞）

このような借用プロセスが非常に生産的であるのに対して、ドナー側の言語の動詞の意味・用法をモデルとして、既存の日本語動詞の新たな用法が生み出される(2)のような現象はあまり見られない。

(2) <u>cut</u> a deal（「取引を成立させる」） ➔ *取引を「<u>切る</u>」

本研究では後者の言語接触現象の興味深い事例として、以下のような「見る」の意味・用法（非有性名詞主語は角括弧、「動詞＋目的語」は下線で示す）を取り上げる。

(3) 十一月の［米貿易収支］が大幅に<u>改善をみた</u>ことから、‥
（中日新聞 1988 年 1 月 19 日）
(4) ［受注指数］は同四〇・〇と、ともに二ケタ台の<u>増加をみた</u>。
（中日新聞 1993 年 11 月 18 日）
(5) バブル崩壊後の［不良債権問題］が一応の<u>解決を見て</u>、‥
（東京新聞 2006 年 5 月 16 日）

この「見る」は、(I)基本義（視覚行為）を希薄化させ、「出来事が出現する」といった意

† kym_tkhs@yahoo.co.jp

味を表す（動詞の問題）。また他に、(II)目的語には漢語サ変動詞の語幹（もしくはそれに相当する名詞）が選択されるといった特徴もある（構文の問題）。例えば、上記以外の「一致・合意・完結・進展（を見る）」などの漢語サ変動詞語幹に加え、「（高度・都市）～化（を見る）」などの複合名詞、「ピーク・インフレ（を見る）」といったカタカナ語まで目的語に立つことがある。これらの名詞は全て変化性を含意しており、これを満たすことが目的語として選択される条件となっている（ただし後者 2 タイプは特殊であり、基本的には前者の漢語サ変動詞語幹の名詞が選択される）。

この「見る」の用法については、「目的語＋動詞（「見る」）」という単位でコロケーションの問題として扱う立場もあるが（村木（1991））、本研究では広く構文の問題として扱う。これは、この「見る」に(III)非有生名詞が主語に立つという特徴があるためである。具体的には、常に非有生名詞が主語に立つ訳ではないが、有生名詞よりも非有生名詞の方が当構文との親和性が高く、選択される主語の多くは非有生名詞となるのが特徴である。つまり、この「見る」は「非有生名詞主語の他動詞文」であり、日本語が「ナル型言語」であることを考えた場合、(III)は日本語の特徴に反することになる。

このように、当該の「見る」の用法についてはこれを構文として捉えることが妥当であると考える。そこでこのような分析の立場に基づいて、本研究では当該の「見る」の用法を「「見る」構文」と呼ぶことにする。

問題は、「この「見る」構文が、日本語固有の表現であるかどうか」ということである。本研究はこの問いに対し、外山（1973）、金田一（1981）を始めとする「日本語の他動詞文は非有生名詞を主語に取りにくい」とする主張のもとに、「非有生名詞主語の「見る」構文は、日本語固有の構文ではなく、(6-7)のような英語の視覚動詞 see の用法に基づいて発生した欧文脈である」という仮説を提示する。

(6) ［The year 1861］saw an increase of 49 per cent. in the number of burglaries and 56 per cent. in its cases of housebreaking. (***The Times***, Feb 20, 1863)

(7) ［The first day here］saw a sudden settlement of seven or more causes of considerable importance, … (***The Times***, Aug 09, 1880)

(6-7)における see の意味用法は「見る」構文の持つ(I)～(III)のいずれの性質とも一致する。またこれ以外にも、日英両言語は、(IV)主に、新聞、専門雑誌などの書き言葉に用いられるという言語使用域（register）上の特徴を共有する。以上より、この種の see の意味用法を「見る」構文に対応するものとして「see 構文」と呼ぶことにする。改めて 4 点を整理する。

(I) 基本義（視覚行為）を希薄化させ、「出来事が出現する」といった意味を表す。（動詞の問題）
(II) 目的語には漢語サ変動詞の語幹（もしくはそれに相当する名詞）が選択されるといった特徴もある。（構文の問題）
(III) 非有生名詞が主語に立つ。
(IV) 主に新聞、専門雑誌などの書き言葉に用いられる。（言語使用域の問題）

しかしながら、これら 4 点については、「日英両言語でこれを共有する」と捉えるのではなく、現代語の確立期に当たる明治・大正期において多くの欧文的要素が日本語文章脈へ

と混入、融合した中で、英語 see から日本語「見る」へともたらされた諸特徴であると推認する。

以上を踏まえ、本研究は、上記仮説を検証するとともに、「見る」構文の和文脈化の過程を明らかにすることを目的とする。検証はコーパスデータに基づき行う。具体的に使用するコーパスは、(IV)の言語使用域に合わせた『太陽コーパス』(確立期現代語)、「Times Digital Archive」「BYU-BNC」(現代英語)、「中日新聞・東京新聞記事データベースサービス」(現代日本語)である。

２．先行研究

２．１　現代日本語に関する先行研究

現代日本語に関する先行研究には、村木（1991)、田中（1996)、高橋（2012 採録済）を取り上げる。

村木（1991）は当該の「見る」を機能動詞として捉え、先行名詞との結合により〈他動性の喪失〉が生じると述べる。またこのような働きは、機能動詞の中でも「見る」にだけ認められるものとして、「見る」を他の機能動詞と区別している。

田中（1996）は視覚動詞「見る」の多種多様な意味には、人間の認知能力を基盤とした有機的な連関があるとして、認知意味論の枠組みに基づいて「見る」の多義ネットワークを提案する。当該の「見る」は、この内「状況の経験、出現」の意味に分類されており、特に後者の「出現」の意味については「認知という対自的または対他的行為の側面も失」うと説明されている（pp.133-134)。統語的側面が分析の射程から外れているためか、「自動詞」「他動詞」といった用語の使用は避けられているが、田中（1996）についても村木（1991）に通ずる主張がなされていると考えてよいと思われる。

高橋（2012 採録済）ではこの種の「見る」に分析対象を限定し、構文そのものに非有生性が現れるとして、これを〈自動性の獲得〉であると述べた。またこの中では、目的語の表す〈人間関与〉の度合いが「見る」構文の有生・非有生性に連動すると主張した。

しかしながらこれらの研究は、議論を先述（1 章）の(I)(II)に集約しており、本研究の主眼とする(III)や他の特徴である(IV)については言及していない。

２．２　現代英語に関する先行研究

先行研究を見る前に、「Oxford English Dictionary Online」における「見る」構文に対応すると思われる see の記述を確認する。

(8) 10a: To know by observation (ocular and other), to witness; to meet with in the course of one's experience; to have personal knowledge of, to be a contemporary of and present at the scene of (an event); to be living at (a certain period of time)

（http://www.oed.com/view/Entry/174749?rskey=kSq3zc&result=3&isAdvanced=false#eid）

（(視覚あるいはその他）の観察によって知ること、目撃すること；自分の経験の過程に遭遇する；(出来事に対する）個人的な知識を有する、あるいは出来事の場面に同時代的に居合わせる；(ある特定の時代に）生きている）（筆者訳）

まとめると、当該 see のこの用法は「主体となる主語がその場に臨場し眼前の状況を経験すること」といった意味を表し、これを満たせば主語に有生・非有生の別を問わないものであると言えるかもしれない。またこの点で、非有生名詞が主語に選択される場合、それ

が母語話者に意識されるかされないかは別として、主語にはある種の比喩的作用が働いていると考えられる。この点については、『英語多義ネットワーク辞典』（瀬戸：2007）でも「擬人的用法」と言明されている。

次に、現代英語に関する先行研究として、国広（1967）、Langacker（2008）を挙げる。

国広（1967）は日英両言語の表現構造の相違の一つに非有生名詞主語表現を挙げ、「時」を表す表現の中に英語の視覚動詞 see（及び find, witness）を用いるものがあると述べる。

(9) ［The succeeding days］saw the Talbots restored to peace and ease.
（O. Henry, *The Duplicity of Hargraves*: 国広（1967：153））

(10) ［The early seventeenth century］saw the establishment of the present usage.
（G. L. Brook. (1958). *A History of the English Language*: 国広（1967：153））

英語の主語名詞は日本語では副詞節として処理されること、目的語は名詞句単独であるもの(9)と「動作名詞＋of＋名詞」の構造を有するもの(10)とがあることを述べ、対応する自然な日本語への翻訳規則を提示する。しかしながら、この中で日本語動詞「見る」との関連は指摘されていない。

Langacker（2008）は「Oxford English Dictionary Online」における上記に相当すると思われる see が非有生名詞主語を伴う場合について、これをセッティング主語構文（setting-subject construction）であると述べる。[1] 観察者としての人間は想起されるに止まり、言語表現上に明示的に表されることはないという。また、他のセッティング主語構文と同様、受動化も許容しない。

(11) a. ［The stadium］has seen some thrilling contests.
b. *Some thrilling contests have been seen by the stadium. （Langacker. (2008: 389 (36a-b))

(12) a. ［The last few years］have witnessed some major changes.
b. *Some major changes have been witnessed by the last years.
（Langacker. (2008: 390(37a-b))

さらに(11a)(12a)からも分かるように、主語の場所性は空間から時間にまで及ぶ。

国広（1967）が挙げる「動作名詞＋of＋名詞」（(9)the establishment of the present usage）型の目的語名詞句の用法は「見る」構文により近いものであり、本研究の仮説に対する根拠の一つとなると考えられる。しかしながら、「見る」構文との関連を明示的に指摘する研究は管見の限りはない。

---

[1] セッティング主語構文とは Langacker（2000）の用語で、通常(i)であれば（空間的・時間的な）場所句として表される名詞句（これをセッティング（setting）という。以下の Hilleaman's latest novel がこれに当たる。）が主語位置に立ち取り立てられる(ii)のような構文のことをいう。他動詞構造を備えるが、主語名詞は出来事の参与者としては機能しないため構文は非他動的（non-transitive）となり、それに伴って受動化を許容しないという特徴を持っている。

(i) Hillerman features Jim Chee in his latest novel.
(ii) Hillerman's latest novel features Jim Chee. (Langacker 2000:70(20a-b))

３.「欧文脈」の定義

具体的分析に入る前に、「欧文脈」という用語を確認、規定しておく。欧文脈とは、「近代西欧語による文章表現（これを「欧文章」と呼んでおく）の要素で、明治時代にはいるまでの日本語による文章表現（これを「和文章」と呼んでおく）には見られなかったものを、日本語による文章表現に新しく取り入れた場合、欧文章からきた新しい要素」と定義されるもので（江湖山（1964：133)、名詞（抽象名詞主語、無生物名詞主語、抽象名詞目的語)、動詞（受動態、使役態、進行相、完了相など)、代名詞（人称代名詞、関係代名詞など）といった文法範疇だけでなく、句読点（及びそれに相当する要素（括弧や疑問符など)）の使用といった表現技法にまで及ぶ。[2] いつ以降の表現かといった明確な区分はなく、明治初期から昭和初期の半世紀以上に渡って消長発展する文章脈として歴史性を持っている（木坂（1979)）。[3] 本研究では、以上に該当するものを欧文脈として捉える。

４.『日本国語大辞典（第二版)』における「見る」構文の記述

次に、「見る」構文が欧文脈であるかどうかを検証する判断材料として、『日本国語大辞典（第二版)』により「見る」構文に相応する記述と初出を確認しておく。

(13) （二）物事を経験したり、物事や人に対して身をもって働きかけたりする。
⑥ある行為・作用が実現する。
＊女工哀史（1925)〈細井和喜蔵〉「冬季暖房のおかげで寒さ知らずに働けるに反し、夏季になって温度の上騰を見ることは甚だしい」
＊後裔の街（1948)〈金達寿〉「それは、諸君の方が頑固なものだから今日まで実現を見られないだけだ」
（『日本国語大辞典（第二版)』第 12 巻（p.868)：下線は筆者による）

初出が大正末年（1925 年（大正 14)）であることは、「見る」構文が欧文脈であるとする積極的な根拠とは言い難い。しかしながら、『女工哀史』（同）における他の「見る」構文には興味深い以下のような例(14)もあり、これにより仮説の妥当性が少なからず向上する。

(14) 十二貫にも足らぬ女工の［身体］は消失してなおマイナスを見る訳である。（p.375）

なおこの時期（明治後期から大正期）については、現代日本語の書き言葉が確立した時期であるとして「確立期現代語」と呼ぶことがある。確立期はその社会的要請から、漢語訳（英語を始めとする諸外国語の漢語による翻訳）の急増とそれに伴う新漢語[4]の出現が際立った時期であると言われている。

「見る」構文における(IV)主に新聞、専門雑誌などの書き言葉に用いられるという特徴は、書き言葉として摂取した see 構文が漢語訳を土台に、書き言葉としての「見る」構文へと翻訳、受容された事情がそのまま受け継がれた結果であると推察される。

---

2 欧文脈に見られる特徴の詳細は、森岡（1999）を参照。

3 欧文脈の歴史的発展に関する詳細は、木坂（1979）を参照。

4 「洋学の翻訳より生じたる漢語」（山田 1958）を指す。

５．コーパスデータに基づく仮説の検証

本節では、コーパスデータをもとに、「見る」構文が欧文脈であるとする本研究の仮説の検証を行っていく。以降、『太陽コーパス』により得られたデータを「確立期現代語」、「中日新聞・東京新聞記事データベースサービス」により得られたデータを「現代語」、「Times Digital Archive」「BYU-BNC」により得られたデータを「現代英語」とする。

５．１　確立期現代語の「見る」構文と現代英語 see 構文との類似

本節では、現代日本語の「見る」構文には見られない確立期現代語の「見る」構文と現代英語の see 構文との類似点を述べる。

５．１．１　「見る」構文の出自

本項ではまず、「見る」構文が既に明治期において見られたことをコーパスデータから明らかにする。[5]

(15)　［我外國貿易］が此の如き長足の進歩を見るに至りしは‥　（『太陽』1895 年 2 号）
(16)　［今年の上半季］は昨年の上半季に比して收入に於て實に三萬六千四百十七圓九十六錢乃ち三割二分強の増加を見る。　（『太陽』1895 年 8 号）
(17)　米國を始め葡萄牙、西班牙、露西亞の如きは減退したれども、他は概して増進せり、殊に［我邦］は最も著しき増加を見る、‥　（『太陽』1895 年 10 号）
(18)　今や東京市民の宿題となれる［電車問題］は何等かの解決を見んとして居る。　（『太陽』1909 年 16 号）

これらは意味や主語・目的語の特徴からみて、「見る」構文と言って差支えがないと思われる。従って、少なくとも 1895 年（明治 28）の時点では「見る」構文が成立していたとみることができる。また同時に、『日本国語大辞典（第二版）』の記述も修正される。

５．１．２　主語について：主語の場所性

主語はいずれも非有生性名詞で場所解釈を可能とし、この点で see 構文に共通する。ただし確立期現代語の場合、現代英語ほど場所性が顕現せず、(15)「我外國貿易」、(17)「我邦」、(18)「電車問題」など、目的語の表す変化の主体とも解釈できるような抽象名詞が主語として選ばれる傾向にある。[6]

(19)　［主語で］［目的語を］［見る］　➔　［主語が］［目的語する］
　　場所性　　変化主体

現代語にはこの特徴が顕著であり、現代英語のように空間的・時間的場所として解釈可能な名詞が主語となる例は観察されない（1 節(3-5)参照）。しかし確立期現代語においては、(10)のような時間的場所解釈を可能にする名詞句（あるまとまった期間を表す時間表現）が主語となることは少なくない。つまり、現代日本語にはない現代英語の特徴が、確立期現代語においては観察されるということである。

5　ただし『太陽コーパス』の性格上、1895 年（明治 28）より以前については探ることができない。

6　主語の解釈は目的語の意味に依存的で、主語が変化の引き起こし手そのものとして動作主ないし経験者と考えられうる場合もある。このことについては、高橋（2012 採録済）で詳しく論じている（現代語を考察対象としているが、確立期現代語にも適用できるものである）。

さらに(16)については、「収入に於て（實に三萬六千四百十七圓九十六錢乃ち三割二分強の）増加」といった目的語が、国広（1967）の言う「動作名詞＋of＋名詞」型目的語の構造に近しく、この場合、以下のように主語位置に立つ名詞を時間的場所として（角括弧［ ］）、目的語を修飾する名詞を変化主体として（破線﹏）解釈できるという類似点も見出すことができる。

(10) ［The early seventeenth century］saw the establishment of the present usage. （再掲）
(16) ［今年の上半季］は昨年の上半季に比して収入に於て實に三萬六千四百十七圓九十六錢乃ち三割二分強の増加を見る。 （再掲）

現代語にも類例は見られるが(20)、国広（1967）の翻訳規則の通り、主語位置に立つ名詞副詞節となって現れており（＝）、主語として解釈することは不可能である。

(20) 七十年代以降、電子製版による［カラー印刷技術］が長足の進歩を見た。
（中日新聞 1995 年 8 月 17 日）

以上、本節では、「見る」構文の出自を修正すると共に、現代語には見られない確立期現代語と現代英語との類似点を提示した。これは、確立期において see 構文から借用した「見る」の新規用法が、長い歴史変化の中で日本語独自の用法へと確立していったことの表れであると考えられる。またこれらは当仮説を支持する強い根拠となる。

５．２　確立期現代語の「見る」構文と現代日本語の「見る」構文との差異

本節では、「見る」構文における確立期現代語と現代日本語の差異を明らかにする。また併せて、確立期現代語の「見る」構文においてのみ見られる特徴が、現代英語 see 構文において見られることを指摘する。

５．２．１　「見る」の表記のゆれ：漢字表記とひらがな表記

現代語において「見る」構文の「見る」は漢字だけでなくひらがなによっても表記される（1 節(3)など）。ひらがな表記の目的は、基本義（視覚行為）との差別化である。「見る」構文の場合、差別化を図るというだけでなく「視覚行為を表わさない」ことの強調とも言えるかもしれない。他方、確立期現代語において「見る」は全て漢字表記される。さらに「見る」以外に「視る」の使用まで観察される。

(21) 昨二十七年の臨時帝國議會及通常帝國議會へ［其處分法案］御提出相成候得共成立を視るに至らざりしは‥ （『太陽』1895 年 10 号）
(22) 此の［責任論］を以て第二の責任論とし、大に政府を攻撃せんとするが如し、然りと雖も［是れ］亦未だ輿論の一致を視るに至らず。 （『太陽』1895 年 11 号）

一貫した漢字表記は、see という英語動詞を忠実に日本語に置き換えようという意識的な試みを示唆する。特に「視る」は、「視力・視察・凝視・注視」などの語からも分かるように「自らの視力により注意を払って見る」ことを表す語で、see 構文の持つ臨場性などと少なからず関わりがあるかもしれない。

また、公用文における補助動詞のひらがな表記は、事務次官等会議申合せ（1981 年（昭和 56））により取り決められたものであるが、このような補助動詞の表記事情が「見る」構

文における「見る」の表記のあり方影響を及ぼしている可能性もあり、この点については追加調査が必要である。

５．２．２　目的語について

「見る」に先行する目的語については、確立期現代語と現代語の間で確実な差異が見られる。1 つは漢語サ変動詞語幹名詞の差異である。例えば現代語では「一致」「合意」「完結」「解決」「進展」などの語と共起しやすい傾向にあるのに対して、確立期現代語ではこの内「一致」「解決」のみ使用が確認され、「合意」「完結」「進展」に関しては 1 例も見られなかった（「一致」についても 2 例のみ）。

いずれの語も明治期（またはそれ以前）には既に存在する語であるため語自体の成立時期によって差異が生じるということではないようであるが[7]、現代語の共起語は「変化性を含意する動名詞」であることを基調としているのに対して、確立期現代語ではこのような制限がないこと、また確立期においては一定数見られた「蹉跌」のような共起語が現代語においては一切見られないことなど、確立期現代語との差異を明確に示している。また、「合意」「完結」「進展」などの語は、「見る」構文が構文として安定した地位を確保したのちに共起語となったもので、これらは言わば「見る」構文における新用法と見ることができる。共起語の推移については通時的分析を通じて再検討する必要があるだろう。

最後に、(23-24)のような作用や行為の結果生じた結果物を表す名詞が目的語に選択される例を見ておく。この種の目的語は確立期現代語にしか観察されず、現代語には一切見られない。目的語に選択される要件を、現代語の「見る」構文において「変化性を含意する動名詞」と規定するならば、確立期現代語の「見る」構文においては「変化性を含意する名詞全般」と言うことができるかもしれない。

(23) 若し［英國］が、世の大勢を顧ることなく、頑迷に、勞働者の團結を禁止するといふ法をとつてゐたならば、何時の時代にか、恐るべき革命を見たかも知れないのである。（『太陽』1925 年 13 号）

(24) 而して獨國の意向甚だ妙ならず、後竟に三國同盟を見るに至れり。（『太陽』1895 年 12 号）

(25) the clearer it became that the EC could not, the less eager［the United States］was to see the alliance,…（***Foreign Affairs***, Jul/Aug, 1994）

先述した通りこの種の目的語は現代語の「見る」構文には見られないが、(25)や 2 章(9)(12)のように現代英語の see 構文においてはしばしば観察される。むしろ現代英語 see においては、動作名詞を目的語に取る「見る」構文に近しい用法よりも、このような用法のほうが自然かもしれない。[8] see 構文との並行性を見る限り、確立期現代語における上記のような例については、現代語と乖離があるというのではなく、現代日本語の「見る」構文として定着しなかったと見る方が適当である。また現代英語 see 構文の中で一般的である用法が確

7　近代語における語彙については、森岡（1991）および『日本国語大辞典（第二版）』を参照している。森岡（1991）によれば、確立期から既に「見る」構文の共起語として選ばれていた「一致」「解決」は新漢語である。成立時期が明治以降であることと新漢語であることとは似て非なることであり、この辺りについても検証すべきかもしれない。

8　この点で 5.1 節に含めるべきものであるが、目的語の性質という観点から本節 5.2 の中に組み入れ考察を行った。

立期現代語においてのみ見られ現代語には観察されないことは、本研究の仮説を支持する有力な証拠となる。

５．３　まとめ

本研究では、「非有生名詞主語の「見る」構文は、日本語固有の構文ではなく、英語の視覚動詞 see の用法に基づいて発生した欧文脈である」という仮説を、以下 4 点から検証した。

(26)
（A）「見る」構文の出自の修正
（B）主語の場所性（確立期現代語と現代英語との類似）
（C）「見る」の表記方法（確立期現代語と現代語の差異）
（D）目的語について
　ａ．共起する目的語（確立期現代語と現代語の差異）
　ｂ．作用や行為の結果生じた結果物を表す目的語（確立期現代語と現代英語との類似）

とりわけ（A）主語の場所性、（Ｄｂ）作用や行為の結果生じた結果物を表す目的語については現代英語との連関を肯定的に指摘しうるものであり、これらにより本仮説は概ね支持されたものと考える。

６　今後の課題

本研究の大きな問題点に、主語の扱いが挙げられる。本分析では主格標示の名詞も主題標示の名詞も一括して主語として捉えたが、両者を区別して分析することで主語性についてより詳細な分析結果を提示できるかもしれない。また、以下(27-8)のような受動用法についても考察に加える必要がある。

(27)　基礎研究、医療、先端技術など、あらゆる分野で前世紀とは比べものにならない進展がみられた。　（東京新聞 2000 年 12 月 26 日）
(28)　「‥昨年度は三学期と比較して十一十五時間の授業時間に増加が見られた」とする反面、‥　（中日新聞 2004 年 9 月 15 日）

「見る」構文における受動用法は、少なくとも『太陽コーパス』には実例が見当たらず、昭和以降の極めて新しい用法である可能性が高いことが分かっている。現代語における安定した使用を鑑みるとこのことは大変意外である。また先述の通り、現代日本語の書き言葉の多くは明治後期から大正期にかけて確立したと言われており、この点からも「見る」構文の特殊性を指摘できるかもしれない。

加えて「見る」構文は日英語にだけでなく他言語においても類例が見られている。とりわけ韓国語には「회의가 일치를 보다」（会議が一致を見る）など、日本語と同一の表現がある。またトルコ語にも「３０度を見た」といった「見る」構文に通ずる視覚動詞の用法があるという。[9]「見る」構文が汎言語的に観察される場合、言語類型論的な観点から「見

[9]　この指摘は 2011 年 9 月 17～18 日に行われた認知言語学会第 12 回全国大会（於奈良教育大学）において、池上嘉彦先生に頂いたものです。またその際にトルコ・アンカラ大学のテキメン・アイシェヌール先生にもトルコ語の例文をご教授頂きました。両先生に改めて感謝申し上げます。

る」構文の再分析が可能である。

参考文献

江湖山恒明（1964）「欧文脈」『講座現代語 2 現代語の成立』明治書院, pp.131-153.

金田一春彦（1981）『日本語の特質』日本放送協会.

Heine, Bernd, and Tania Kuteva（2005）*Language Contact and Grammatical Change*. Cambridge University Press.

木坂基（1979）「欧文脈の消長」『言語生活』335, pp.62-69.

国広哲弥（1967）『ELEC 言語叢書　構造的意味論－日英両語対照研究―』三省堂.

国立国語研究所（編）（2005）『太陽コーパス[CD-ROM]―雑誌「太陽」日本語データベース(国立国語研究所資料集 15)』博文館新社.

Langacker, Ronald W.（2000）*Grammar and Conceptualization.* Oxford University Press.

_____.（2008）*Cognitive Grammar: A Basic Introduction.* Oxford University Press.（ロナルド・W・ラネカー（2011）『認知文法論序説』山梨正明訳. 研究社.）

Loveday, Leo J（1996）*Language Contact in Japan.* Oxford University Press.

村木新次郎（1991）『日本語動詞の諸相』ひつじ書房.

森岡健二（1991）『近代語の成立〈語彙編〉』明治書院.

_____.（1999）『欧文訓読の研究: 欧文脈の形成』明治書院.

瀬戸賢一（2007）『英語多義ネットワーク辞典 (英語辞典シリーズ)』小学館.

小学館国語辞典編集部（2001）『日本国語大辞典　第二版　第十二巻』小学館.

高橋暦（2012 採録済）「日本語動詞「見る」における自動性の認知言語学的考察」『認知言語学会論文集』12, 認知言語学会.

田中聡子（1996）「動詞「みる」の多義構造」『言語研究』110, pp.120-142.

Thomason, Sara G., and Terrence Kaufman（1989）*Language Contact, Creolization, and Genetic Linguistics*. University of California Press.

山田孝雄（1958）『国語の中に於ける漢語の研究』宝文館.

関連 URL

Times Digital Archive:　http://infotrac.galegroup.com/itw/infomark/1/1/1/purl=rc6_TTDA

BYU-BNC:　http://corpus.byu.edu/coca/

中日新聞・東京新聞記事データベースサービス:　http://www.cnc.ne.jp/ip/

用例出典

細井和喜蔵（1925）『女工哀史』講談社.

金達寿（1948）『後裔の街』朝鮮文芸社.

文書分類における補集合を併用した Naive Bayes

伊藤 裕佑(東京農工大学工学部)†
古宮 嘉那子(東京農工大学工学研究院)
小谷 善行(東京農工大学工学研究院)

Naive Bayes using The Complement Set in Text Classification

Yusuke Ito (Department of Computer and Information Sciences Faculty of Engineering),
Kanako Komiya (Institute of Engineering Tokyo University of Agriculture and Technology),
Yoshiyuki Kotani (Institute of Engineering Tokyo University of Agriculture and Technology)

１．はじめに

これまで文書分類に関する研究は数多くなされてきており，これらの研究において Bayes のアプローチがよく用いられている．Naive Bayes を発展させた研究として，Rennie らによる「補集合」を用いた Complement Naive Bayes [1]や、Komiya らの Negation Naive Bayes[2]がある．本研究では，Naive Bayes と Negation Naive Bayes に注目し，補集合を利用した新しい手法を提案する．それには，Naive Bayes と Negation Naive Bayes の統合およびクラスごとの学習量による選択を行う．

２．関連研究

Andrew は Naive Bayes を適用して分類を行う際の事象モデルとして，多項モデルと多変量ベルヌーイモデルの違いを述べ，分類結果から多項モデルの優位性を示している[3]. Komiya らは、Rennie らによる Complement Naive Bayes という手法に注目し，Negation Naive Bayes を提案している．本研究では多項モデルの Naive Bayes と Negation Naive Bayes に注目し，新しい手法を開発する．

３．Bayes の定理を用いた既存の文書分類法

本研究で提案する手法の基礎となる分類器について述べる．これらの分類器は本研究で文書分類の実験を行う際の比較対象となる．

分類器が分類先を推定する際には，Bayes の定理を利用した推定式から事後確率 $P(c|d)$ が最大となるクラス $\hat{c}$ を求める．Naive Bayes は $P(c|d)$ を最大化するように Bayes の定理をそのまま適用した分類器である．これはクラスごとに与えられた文書をそのまま学習に利用する． Negation Naive Bayes は「クラスに属さない文書」，つまり「補集合」を考えることでクラスごとの学習事例数を増やすように工夫している．

３．１　Naive Bayes

確率モデルによる文書分類において，分類対象となる文書を d，ある一つのクラスを c とするとき，事後確率 $P(c|d)$ を最大化するクラス $\hat{c}$ を求めればよい．

Naive Bayes では，$P(c|d)$ に Bayes の定理を適用するが，文書の取り出される確率 $P(d)$ がすべてのクラスについて一定であるので，Naive Bayes はクラスの出現確率 $P(c)$ と各クラスにおける文書の出現確率 $P(d|c)$ の積を最大化するクラスを推定する．ただし，文書 d は $(w_1, w_2, ..., w_n)$ のような単語列からなる．

$$\hat{c} = \arg\max_c P(c) \prod_{i=i}^{n} P(w_i|c) \quad (1)$$

† 50010268502@st.tuat.ac.jp

３．２　Negation Naive Bayes

Naive Bayes による文書分類では，ある一つのクラス c に関する学習に「c に属する訓練事例」を用いていた．それを発展させた分類器として Complement Naive Bayes があり，「c に属さない訓練事例」すなわち「$\bar{c}$ に属する訓練事例(補集合)」を用いて学習する．しかし，Rennie らの研究ではアプローチが特徴的であるものの数学的根拠がない[1]．そのため，Komiya ら[2]は補集合を用いたモデルに基づく方法を開発してきている．

事後確率 $P(c \mid d)$ を最大化するクラス $\hat{c}$ を求める式を変形する．

$$\hat{c} = \arg\max_c P(c \mid d) = \arg\max_c (1 - P(\bar{c} \mid d)) = \arg\min_c P(\bar{c} \mid d) \tag{2}$$

次に，Bayes の定理を用いて Naive Bayes と同様に変形する．

$$\hat{c} = \arg\min_c P(\bar{c}) P(d \mid \bar{c}) \tag{3}$$

したがって，文書 d の属するクラス $\hat{c}$ を次の式で推定する．

$$\hat{c} = \arg\min_c P(\bar{c}) \prod_{i=1}^{n} P(w_i \mid \bar{c}) \tag{4}$$

４．補集合を併用した Naive Bayes の提案

本研究では，Negation Naive Bayes と同様に Bayes の式を変形することで新しい方法を考え，ここではそれを「Universal-set Naive Bayes」と呼ぶ．また，各クラスの $P(c)$ によって Naive Bayes と Negation Naive Bayes を選択して処理する方法を新たに考え, それを「Selective Naive Bayes」と呼ぶ.

４．１　Universal-set Naive Bayes

事後確率 $P(c \mid d)$ の推定に $P(d)$ は不要であるとして式(1), (2)は導出されている．しかし，ここでは式(6)を $P(d)$ について解くことで新たな分類器を考える．式(6)は $P(c \mid d)$ と $P(\bar{c} \mid d)$ を足し合わせた全体の確率

$$P(\text{全体} \mid d) = P(c \mid d) + P(\bar{c} \mid d) = 1 \tag{5}$$

に Bayes の定理を適用している．

$$\frac{P(c)P(d \mid c)}{P(d)} + \frac{P(\bar{c} \mid d)}{P(d)} = 1 \tag{6}$$

式(6)を変形することで $P(d) = P(c)P(d \mid c) + P(\bar{c})P(d \mid \bar{c})$ が得られ, Bayes の式を次のように書きかえられる．

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} = \frac{P(c)P(d \mid c)}{P(c)P(d \mid c) + P(\bar{c})P(d \mid \bar{c})} = \frac{1}{1 + \dfrac{1}{\dfrac{P(c)P(d \mid c)}{P(\bar{c})P(d \mid \bar{c})}}} \tag{7}$$

式(7)の右辺の分数を突き詰めていくと，左辺 $P(c \mid d)$ の最大化が右辺 $\dfrac{P(c)P(d \mid c)}{P(\bar{c})P(d \mid \bar{c})}$ の最大化となるので，Universal-set Naive Bayes は文書 d の属するクラス $\hat{c}$ を次の式で推定する．

$$\hat{c} = \arg\max_c \frac{P(c)}{P(\bar{c})} \prod_{i=1}^{n} \frac{P(w_i \mid c)}{P(w_i \mid \bar{c})} \tag{8}$$

４．２　Selective Naive Bayes

Negation Naive Bayes はばらつきを抑えて Naive Bayes より分類性能を向上させているが，補集合を学習に用いることで逆に学習事例数を減らしてしまう場合がある(図 1)．これを $P(c)$ に基づいてクラスごとに式(1)，(2)で選択し，文書分類を行なう．このとき，0.5 をしきい値とすることで，学習する事例数がより大きくなるように選択する．

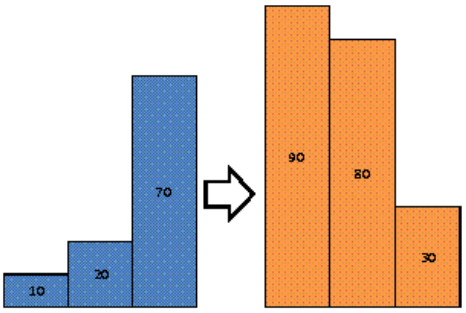


図 1　補集合をとる前後の 3 クラスの学習事例数(例)

分類先のクラスを推定する際には，Naive Bayes や Negation Naive Bayes 単独のときと異なり，$P(d)$ を求める必要がある．この $P(d)$ は以下に示す式(9)のように異なる導出がある．これらは式(6)を変形していくことで得られる．

$$P(d)=\sum_{c}P(c)\prod_{i=1}^{n}P(w_i\mid c)=\frac{1}{|C|-1}\sum_{c}P(\bar{c})\prod_{i=1}^{n}P(w_i\mid \bar{c}) \tag{9}$$

したがって，文書 d の属するクラス $\hat{c}$ を次の式で推定する．

$$\hat{c}=\arg\max_{c}\begin{cases}\dfrac{P(c)\prod_{i=1}^{n}P(w_i\mid c)}{\sum_{c}P(c)\prod_{i=1}^{n}P(w_i\mid c)}, & P(c)\text{が}0.5\text{以上}\\ 1-(|C|-1)\dfrac{P(\bar{c})\prod_{i=1}^{n}P(w_i\mid \bar{c})}{\sum_{c}P(\bar{c})\prod_{i=1}^{n}P(w_i\mid \bar{c})}, & \text{その他}\end{cases} \tag{10}$$

５．実験

ここでは，提案手法の性能を評価するための実験を行い，その結果を示す．

比較する既存手法として，ベースラインの Naive Bayes，および Negation Naive Bayes[2] を取り上げる．なお，ここでは Naive Bayes，Negation Naive Bayes，Universal-set Naive Bayes，Selective Naive Bayes をそれぞれ NB，NNB，UNB，SNB と略記する．

分類性能を評価する実験に用いるコーパスには「現代日本語書き言葉均衡コーパス(BCCWJ)」の一部を用いる(図 2)．五つのジャンルがあり，Yahoo!知恵袋，白書，書籍，雑誌，新聞である．BCCWJ の実験は bag-of-words に加工済みのデータを用いて文書分類を行なっている．5 分割交差検定で実験を行うが，データセットについては 5 クラスの場合だけでなく，3 クラスのデータセットを作った場合も実験を行った．5 クラスよりもシンプルな 3 クラスの分類実験を行なって提案手法と既存手法の違いを確認する．

５．１　実験方法と 3 クラス分類のデータセット

実験は各データセットについて 5 分割交差検定を行い評価する．5 クラスのデータすべてを使った 5 クラス分類と，よりシンプルな 3 クラス分類 10 通りをそれぞれ実験する．5 ジ

ャンルからクラスを 3 つ選ぶ組合せが 10 通りとなる.

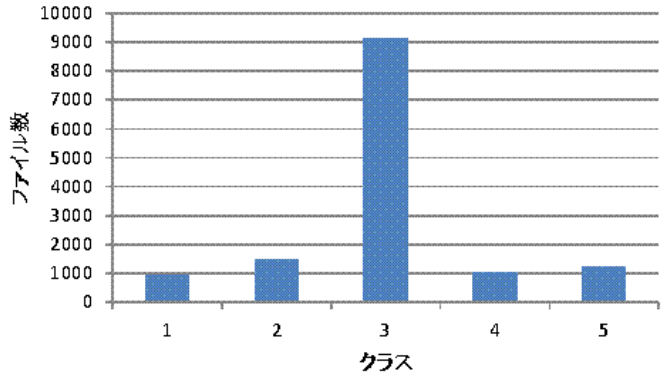


図 2　クラスごとの文書ファイル数(1: Yahoo!知恵袋，2:白書，3:書籍，4:雑誌，5:新聞)

５.２　実験結果

評価実験の結果は表 1 および 2 のとおりであり，シンプルな 3 クラス分類について提案手法が優れており，その違いが有意であることが確かめられた．また，どちらの表からも提案手法の一つである UNB が優れていることがわかり，適用できる問題が広範囲である可能性を持つ.

表 1 5 クラス分類における各手法の評価指標

	NB	NNB	UNB	SNB
Precision	0.7036	0.5574	**0.7442**	0.2505
Recall	**0.8314**	0.3985	0.7573	0.2877
F-measure	**0.7486**	0.3356	0.7381	0.2638
Accuracy	0.7798	0.6683	**0.8048**	0.4500

表 2 3 クラス分類における各手法の評価指標の平均値

	NB	NNB	UNB	SNB
Precision	0.6518	0.6230	**0.6687**	0.5455
Recall	**0.7078**	0.5755	0.7034	0.4748
F-measure	0.5425	0.4744	**0.5641**	0.4550
Accuracy	0.5577	0.5587	**0.5933**	0.5711

６．考察・今後の課題

本研究の開発した新しい手法が文書分類において既存手法に比べて有効であることが示された．Universal-set Naive Bayes が良い手法であることは確かめられたが，Selective Naive Bayes は今後，他の手法との違いをデータセットを変えて詳しく確かめる必要がある.

文　献

J.D.M.Rennie, L.Shih, J.Teevan, and D.R.Karger(2003) Tackling the Poor Assumptions of Naive Bayes Text Classification, ICML2003, pp.616-623

Kanako Komiya, Naoto Sato, Koji Fujimoto and Yoshiyuki Kotani, (2011)Negation Naive Bayes for Categorization of Product Pages on the Web, 2011 (RANLP 2011), pp586-591

Andrew McCallum, Kamal Nigam(1998) A Comparison of Event Models for Naive Bayes Text Classification, AAAI/ICML-98 Workshop on Learning for Text Categorization, pp.41-48

日本語並立助詞「と」・「や」と英語冠詞に関する一考察
—BCCWJ データに基づいて—

川口　裕子（神戸女学院大学大学院研究生）†

On Japanese Parallel Markers and English Articles
—Based on BCCWJ Data—

Yuko Kawaguchi（Kobe College Graduate School, Research Student）

１．はじめに

日本語には冠詞が存在しないが、日本語の格助詞「は」・「が」と定冠詞・不定冠詞との対応関係についてはしばしば論じられてきた。上村(1978)は、「は」は旧情報に言及しており定冠詞(the)に対応し、「が」は新情報に言及しており不定冠詞(a /an)に対応すると述べている。

並立助詞「と」・「や」はそれぞれ「全部列挙」・「一部列挙」を表すとされるが、実際の用例は多岐にわたる。筆者は、並立助詞「と」・「や」の用例の一部と冠詞に関しても対応関係が存在すると考える。本研究では、「と」は新情報的、「や」は旧情報的側面があるという先行研究の主張を踏まえ、BCCWJ データより「と」と「や」の用例を分析し、「と」は不定冠詞、「や」は定冠詞に類似した役割を持つと考える根拠を示す。

２．冠詞

現代英語の冠詞には、定冠詞(definite article)と不定冠詞(indefinite article)があり、それぞれに異なった意味機能を持つ。安藤(2005)によると、定冠詞(definite article)は同定機能(identifying function)を持ち、文脈または場面から指示対象が聞き手にとって唯一的に同定可能(uniquely identifiable)である場合、つまり指示物が聞き手にとって旧情報(old information)に属するという前提が話し手の側に存在する場合に使用される。一方、不定冠詞にはそのような前提は存在しない。

Jespersen(1933)は、英語の定冠詞と不定冠詞について以下のように述べている。

> The chief use of the (definite) article is to indicate the person or thing that at the moment is uppermost in the mind of the speaker and presumably in that of the hearer too. (中略)
>
> The indefinite article is used not only in introductory remarks……, where we expect further information, but also in a great many other cases where the singular of a noun is required, while no identification is possible or important……

関口(1960)はドイツ語に基づいた冠詞論を展開したが、定冠詞と不定冠詞の解釈はJespersen(1933)のものと類似している。

> 定冠詞の機能は、その次に置かれた名詞の表示する概念が、何等かの意味において既知と前提されてよろしいということを暗示するにある。
>
> 不定冠詞の機能は、その次に置かれた名詞の表示する概念が、話者あるいは聴者にとって、

† kawayuko@mx4.canvas.ne.jp

何等かの意味において未知と前提されてよろしいことを暗示するにある。

これら先行研究によると、定冠詞はそのあとに来る名詞が既知情報(旧情報)で唯一的に同定可能であることを前提に用いられ、一方、不定冠詞は指示対象が未知情報（新情報）であり、話し手と聞き手の間に指示対象に対する共通の認識が存在しない場合に用いられるといえる。

３．冠詞と日本語助詞の関係

日本語には冠詞が存在しないが、冠詞と日本語の他の品詞との関連性は先行研究において指摘されている。関口(1962)は、「日本語の『て・に・お・は』も、『意味』というものは持たず、単に『意味形態』だけしか持っていないと云う点で、外国語の『冠詞』という現象にそっくりなところがある。」と述べている。また、鰍沢(2006)によれば、17 世紀の初頭にはすでに西欧語の冠詞と日本語の助詞について言及した文法書 [1]が刊行されていたという。冠詞と助詞の関連については、格助詞「が」と「は」と冠詞に関する先行研究が数多く見受けられる。

３．１　日本語助詞「は」と「が」の機能

大野(1978)は、「私は大野です。」と「私が大野です。」の例を用い、「は」は既知の情報を扱うもの、「が」は未知の情報を扱うものであるとしている。ただし、ここでいう「既知」と「未知」とは、その情報の実際の新旧を問うものではなく、話し手がその情報を「既知扱い」するか「未知扱い」するかのことである。

> 助詞は、客観的世界の事物を直接に指す言葉ではない……その話し手が、物や行為を表す言葉・概念をどう関連づけるか、話し手として主体的にどう扱うかということを表明する言葉である。たとえば、ガという助詞は、たんに未知のものを承けるだけでなく、承けるものを主体的に未知扱いすることを表示する役目を負うのである。

久野（1973）も、「は」と「が」の機能について、「古いインフォーメイション」と「新しいインフォーメイション」という観点から論じている。

> 主文の主語に現われる「ガ」は、名詞句がその文の中で、新しいインフォーメイション(すなわち、文脈から予測することができないインフォーメイション)を表すことをマークする標識である。

上村(1978)によれば、このような情報の新旧に関する考え方は、松下(1924)[2]やChafe(1970)[3]など多くの先行研究に見受けられる。

３．２　冠詞と日本語助詞「は」と「が」の機能

上村(1978)は、「は」と「が」の機能の違いについて論じたうえで、「は」は旧情報に言及しており定冠詞theに対応し、「が」は新情報に言及しており不定冠詞a/anに対応すると述べている。さらに上村

---

1 Ioao Rodriguez (1604-1608) *Arte de Lingoa de Iapan.*

2 松下大三郎(1924)『標準日本文法』

3 Wallace L. Chafe. (1970) *Meaning and the Structure of Language.* The University of Chicago Press.

は、乾(1959)[4]の「日本語の格助辞『は』『が』……の別は、一部は不定冠詞と定冠詞の用法に対応するものがある(There is a book. The book was on the table. 本が……、(その)本は……」という主張以来、「が」と「は」、a(n)とtheとの対応がにわかに論じられるようになったことに触れ、その分析は「単純な事実の単純な指摘」に過ぎず、それぞれの助詞と冠詞の関係について論述したと言えるものではないと述べている。

４．冠詞と日本語並立助詞「と」と「や」の関係

４．１　問題提起

これまで格助詞「は」・「が」の機能、両者と定冠詞・不定冠詞との対応関係に関する先行研究について述べてきたが、筆者は並立助詞「と」・「や」の一部の用法と冠詞の用法が類似していると考える。ここで並立助詞の「と」「や」の意味機能について考え、冠詞の役割との類似点を指摘する。

４．２　並立助詞「と」と「や」に関する先行研究

寺村 (1991)は、「二つ以上の名詞を並立的に結びつけるのに『ヤ』が用いられるときは、それらの名詞が、あるセットの具体例として、そのメンバーの一部として取り上げられていることを示す。先の『ト』が、そのセットのメンバーすべてを挙げるのと違って、『ヤ』は、そのほかにも同類のものがあるという意味を含んでいる。」と述べ、「と」を「全部列挙」、「や」と「一部列挙」と区別している。同様に益岡・田窪 (1992)[5]は、「と」を「総記」、「や」を「例示」とし、柏木 (2006)[6]は「と」を「全体化」、「や」を「類化」と区別している。一方、国広(1973)は「『や』は……いわば‘and’と‘or’の両方にまたがるものである。しかし英語では、‘and’と‘or’は対立概念であって、英語と同じ平面に立って考えれば両方にまたがる概念というものはあり得ない。『や』は全く異なった平面に属しているわけである。」と説明する。

また、朴(2006)は、『黒い雨』の例を用い、全部列挙で「や」が用いられる場合について説明する。

> 「もう池本屋も、広島や長崎が原爆されたことを忘れとる。みんなが忘れとる。あのときの灼熱地獄―あれを忘れて、何かこのごろ、あの原爆記念の大会じゃ。あのお祭り騒ぎが、わしゃあ情けない」

原爆が投下されたのは広島と長崎の二都市であるため、上は全部列挙である。それにもかかわらず、一般的に全部列挙とされる「と」ではなく一部列挙の「や」が用いられている。このような全部列挙に「や」が用いられる理由について、国広は、大した意味もなく「や」を用いる「やわらげ」であると説明する。一方、朴は「『広島』『長崎』は『被爆地』という『カテゴリー』の背景の中に存在しており、『や』は相手に A・B の典型例を挙げることによって『A のような』『B のような』、それにふさわしい『カテゴリーC』を連想させる機能を持つ」と説明している。

さらに、市川(1991)も、「や」には「と」と同様に「全部列挙」機能をもつものが存在するとし、「全部列

---

[4] 乾亮一(1959)「英語と日本語―比較における二三の問題について」『英文法研究』研究社

[5]益岡隆志・田窪行則 (1992) 『基礎日本語文法―改訂版』 東京：くろしお出版.

[6]柏木成章「『全体化』と『類化』： 並立助詞論、特に『と』・『や』を中心として」 (2006) 『別科日本語教育 ： 大東文化大学別科論集』 8：99-107.

挙」の機能を持つ「と」と「や」の使い分けについて以下の傾向があることを指摘している。

1)「や」で結ばれる語は特定のものが多い。
2)「と」は独立的、「や」は述語への依存度が高い。
3)「と」は格助詞的、「や」はムードの助詞的(間投助詞的、とりたて助詞的)。
4)「と」は厳密、「や」は大雑把、全体的。
5)「と」は新情報的、「や」は旧情報的。

5つの傾向のうち、5)の「と」は新情報的、「や」は旧情報的側面を持つという傾向に関し、市川は井伏鱒二の『山椒魚』からの引用を用いて説明している。

> 岩屋の天井には、杉苔と銭苔が密生して、銭苔は緑色のうろこでもって地所とり(小児の遊戯の一種)の形式で繁殖し、杉苔は細く且つ紅色の花柄の尖端に、可憐な花を咲かせた。(中略)サンショウウオは、杉苔や銭苔を眺めることを好まなかった。むしろそれを疎んじさえした。杉苔の花粉はしきりに岩屋の中の水面に散ったので、彼は自分の棲家の水が汚れてしまうと信じたからである。あまつさえ、彼は岩や天井の凹みには、一群ずつのかびさえも生えた。かびはなんと愚かな習性を持っていた事であろう。

市川は、二度目の「杉苔」「銭苔」が出てくるときに「や」が使われていることを、「や」は否定的意味合いをもつ表現と結び付きやすく、「好まなかった」という否定形につながっていることによる可能性を認めつつも、初出でてた「杉苔」「銭苔」を旧情報として「や」で受けたとしている。

４．３　仮説

市川の主張するように全部列挙の「と」を新情報、「や」を旧情報であるとするならば、先述の朴の『黒い雨』からの引用にも異なった観点からの説明が可能である。

(1)1945 年 8 月、広島と長崎に原爆が投下された。
(2) もう池本屋も、広島や長崎が原爆されたことを忘れとる。(朴(2006))
(3)桃太郎は、犬と猿と雉をお供に連れて行った。(寺村(1991))
(*Momotaro* was accompanied by a dog, a monkey, and a pheasant.)
(4)桃太郎は、犬や猿や雉をお供に連れて行ったことを誇りに思った。
(*Momotaro* was proud that he had been accompanied by the dog, the monkey and the pheasant.)
(5)桃太郎は、犬と猿と雉をお供に連れて行ったことを誇りに思った。

(1)と(2)の全部列挙の例文を比較すると、(2)は(1)の内容を前提としており、既知の情報を「や」で受けているということができる。また、(3)(4)(5)も全部列挙であるが、(4)(5)は(3)の情報を前提にしており、(4)(5)で出てくる「犬」「猿」「雉」は唯一的に同定可能である。また、(5)に比べて(4)の方が自然に感じられることは、市川の「『や』は述語への依存が高い」という主張により説明ができる。(4)では、すでに既知の情報である「犬と猿と雉をお供に連れて行った」ことよりも、述部にある「誇りに思った」という未知の情報に焦点を当てていると考えることができる。この全部列挙における「と」を新情報、「や」を旧情報であるとする説明は、不定冠詞を未知情報、定冠詞を既知情報とする冠詞の機能についての説

明に類似している。そのため、筆者は全部列挙の「と」・「や」と不定冠・定冠詞との間にも対応関係が存在するとの仮説を立て、BCCWJ データより全部列挙の「と」と「や」の用例を分析し、不定冠と定冠詞との関係を論じる。

４．４　BCCWJ データに基づく「と」と「や」の分析と冠詞との対応

(6)海援隊のなかで靴をはいた写真を残しているのは、龍馬以外には池内蔵太だけである。内蔵太が靴をはいた写真を残したのは、彼が人一倍龍馬から目をかけられていたためかもしれない。(中略)一方、龍馬や内蔵太が靴をはいた写真を残しているということは、長崎市中を靴で歩いたかどうかは別として、龍馬にとって靴がかなり身近なものになっていたともいえそうだ。(LBa5_00012)

文頭で「海援隊の中で靴をはいた写真を残したのは、龍馬と内蔵太だけである」旨が述べられているため、(6)が全部列挙の文であることは明白である。それにもかかわらず「や」が用いられているのは、「龍馬と内蔵太は靴をはいた写真を残した」という情報を前提とし、その事実を旧情報として扱っているためであるとの解釈が可能である。この「と」と「や」についての新情報と旧情報という説明は、先に見た不定冠詞と定冠詞の機能とも酷似している。

(7)「宿命」というのは、前世から定まっている運命なんだ。例えば剛志や雅子ちゃんが、内海のお父さんとお母さんの間に生まれたということが宿命なんだ。内海の子は嫌だ。○○さんの家の子に生まれたい。と、子供はお父さんやお母さんを選んで生まれることはできない。
（PB54_00148）

(8)「しっ、あまり大きな声を出さないでください。おやじやおふくろが驚きます」
「よし、それじゃ手早く、勝負の日を決めようぞ。明日にしよう。朝早く、ここを出て、伊賀の里まで行くのだ」(中略)
「誰が行くものですか」
「そんなことを言っていいのか。俺は忍びだぞ。足が速いだけではなく、人の命をもらうのも得意だ。おぬしの父親や母親が、ある日首をくくっていても、いいのだな」(PB19_00235)

(7)では、「内海のお父さんとお母さん」という新情報が提示されたあとに、それを前提とした「お父さんやお母さん」という旧情報が提示されている。 (8)の例では、文脈より誰の父親と母親であるかは唯一的に同定可能であると言える。

(9)結婚を一週間後に控えた日曜日。ゆり子は結婚の打ち合わせや準備のため、枝川家を訪れた 。(中略)仰々しい結婚式や披露宴はしたくない、という真行の意思に、ゆり子はすでに同意していた。 このため、次の日曜の朝、二人で婚姻届を出した後、親しい人を家に招いて昼食を取る予定である。(PB29_00634)

(10)たとえば、学校には球技大会をはじめとして多くの行事がある。それらの企画や運営は、各学校の教育方針に沿ったものであり、多種多様である。その実施方法のひとつとして、運営

の一部または全部を生徒に任せるという方法がある。 (LBf3_00004)

(9)、(10)においても、「友里子が結婚を控えている」こと、「学校行事が多く存在する」ことがすでに情報として提示されているため、「結婚式」「披露宴」、「企画」「運営」はそれぞれ旧情報として扱われている。また、そのことにより、誰の結婚式と披露宴であるか、何の企画と運営であるかが唯一的に同定可能であるため、「と」のかわりに「や」が用いられていると考えることが可能である。

ここに挙げた BCCWJ データからの用例を見る限り、全部列挙における「と」は新情報、「や」は旧情報という解釈は妥当であると思われる。また、それぞれの情報の新旧や唯一的同定可能性の有無により、「と」は不定冠詞、「や」は定冠詞に対応すると考えることも可能である。

５．結論

冠詞は日本語に存在しないため、助詞などの他の品詞がその機能を果たしていると言われている。とくに格助詞「は」・「が」と定冠詞・不定冠詞との対応関係が論じられてきた。筆者は、日本語の並立助詞「と」・「や」の機能の一部と、「不定冠詞」「定冠詞」に対応関係があるとの仮説のもと、本研究を行った。定冠詞と不定冠詞の機能、並立助詞「と」と「や」の機能に関する先行研究を検討し、BCCWJ データの「と」と「や」を含む文例を検証した結果、全部列挙表現において、「と」は新情報を提示する機能を持ち、不定冠詞に対応し、「や」は旧情報を提示する機能を持ち、定冠詞の役割を果たすことを示した。今後は、日英対訳のデータ分析、統計的手法などを用い、さらなる調査が必要である。

６．参考文献

安藤貞雄(2005)『現代英文法講義』開拓社

市川保子(1991)「並立助詞『と』と『や』に関する一考察」『文芸言語研究言語篇』20, pp. 61-79. 筑波大学文芸・言語学系

上村和也(1979)「英語の冠詞と日本語の助詞」『英語英文論集』10, pp. 29-49. 鹿児島大学教養部

大野晋(1978)『日本語の文法を考える』岩波新書

鰍沢千鶴(2006)「冠詞と日本語」『上智大学国文学論集』39, pp. 49-67. 上智大学

国広哲弥(1967)「'And'と『と・に・や・も』―日英両語語彙の比較―」『言語研究』50, pp.34-49.

久野暲(1973)『日本語文法研究』大修館書店

関口存男(1960)『冠詞―意味形態的背景より見たるドイツ語冠詞の研究・第一巻定冠詞篇』三修社

関口存男(1961)『冠詞―意味形態的背景より見たるドイツ語冠詞の研究・第二巻不定冠詞篇』三修社

関口存男(1962)『冠詞―意味形態的背景より見たるドイツ語冠詞の研究・第三巻無冠詞篇』三修社

寺村秀夫(1991)『日本語のシンタクスと意味 第 III 巻』くろしお出版

朴 点淑 (2006)「現代日本語における並立助詞『と』『や』」『岡山大学言語学論叢』12, pp. 51-62. 岡山大学

Jespersen, O. (1933) Essentials *of English Grammar*. George Allen and Unwin.

日本語と英語の対訳文対の収集と著作権の考察

村上仁一 (鳥取大学 工学部 知能情報工学科) [*1]
藤波進　（学際統合創研株式会社） [*2]

Japanese-English Parallel Sentences Collection from Digital Media

Jin'ichi Murakami (Faculty of Engineering, Tottori University)
Susumu Fujinami (Cyber Creative Institute CO. Ltd)

概要

日英対訳辞書は，翻訳の研究において必要不可欠のものである．しかし，日本文と英文が文単位に対応していて，量が多く，一般の人が入手可能な対訳文対は，最近まで存在していなかったと言える．本研究では，様々な電子媒体から，日本文と英文が文単位に対応する対訳文対を採取した．電子媒体として，CD-ROM・Internet・電子辞書などを利用した．これらの結果，対訳文対として 1,099,093 万文対を採取した（対訳データベース）．そして，得られた対訳文対から，日本文において単文の対訳文対を 182,113 文対，採取した（単文データベース）．また日本文において重文・複文の対訳文対を 158,633 文対，採取した（重文・複文データベース）．また最後に著作権の問題について述べる

キーワード：機械翻訳，統計翻訳，日英対訳文対，単文，重文・複文，著作権

Summary

We collected large number of Japanese-English parallel sentences from many digital medias. There are many digital media like Japanese English dictionaries, English sample sentences and CD-ROMs. There are 9 types in digital medias. Finally, We collected about 1,099,093 parallel sentences. Also, we extracted 182,113 simple sentences from this parallel sentences. And, we extracted 158,633 complex and compound sentences from this parallel sentences. Also, we described the copyright problem for parallel sentences.

Key words : Machine Translation, Statistical Machine Translation, Parallel Corpus, Simple Sentence, Complex Sentence, Compound Sentence, Copy right

1　はじめに

日英・英日対訳辞書は，翻訳において必要不可欠のものである．そのため，辞書編纂には長い歴史がある．しかし，つい最近まで，電子的に読めて，一般の人が入手可能であり，大規模で，日本文と英文が文単位で対応している対訳文対は，存在していなかったと言える．現在，多くの英日や日英の辞書や例文集などが CD-ROM などの電子媒体で販売されている．しかし，日本文と英文が文単位で対応している対訳文対を大量に採取することは，困難である．電子媒体をある程度加工することで対訳文対を作成することが可能である．しかし，対訳文対を採取することが困難なコーパスは多い．

本研究では，様々な電子媒体のコーパスから得られた日英対訳文対の量や抽出の問題点や得られた対訳文対について述べる．電子媒体には，電子辞書・CD-ROM・Internet などがあるが，本研究では，以下の 9 つに分類した．

1 電子辞書，共通フォーマット　2 電子辞書，独自フォーマット　3 CD-ROM 付の書籍　4 Internet　5 新聞記事
6 対訳文対，無償　7 対訳文対，販売　8 対訳文対，未販売　9 その他

これらの電子媒体から，本研究では，総計として，1,099,093 文対の対訳文対を採取した．そして，得られた対訳文対から，日本語において単文の対訳文対を 182,113 文対，採取した．また日本語において重文・複文の対訳文対を 158,633 文対，採取した．

最後に，これらの対訳文対における著作権の問題について述べる．

2　利用可能な電子媒体

現在，多くの英日や日英の辞書や例文集が，CD-ROM などの電子媒体で販売されている．しかし，日本文と英文が文単位で対応している対訳文対をもつ電子媒体は少ない．しかし，電子媒体をある程度加工することで，対訳文対を採取することが可能である．採取可能な電子媒体は，以下の 9 つに大きく分類することができる．以後は，それぞれの特徴について述べる．

1. 電子辞書，共通フォーマット （分類番号 1）

   現在，コンピュータにおいて検索可能な日英，英日辞書の電子辞書がある．このフォーマットには，一般に公開されているフォーマットを使用している辞書と，各社独自のフォーマットを採用している辞書の 2 種類がある．

   一般に公開されているフォーマットには EPWING 形式と電子ブック形式とロボワード形式の 3 種類が有名である． EPWING は日本独自の電子出版の共通フォーマットで，基本的に JIS コードで記録されている．このフォーマットの辞書は，英日，日英辞書のほかにも広辞苑や漢和辞書などがあり，現在 50 種類を越える辞書が販売されている [41]．フォーマットが公開されているため，テキストを抽出することは容易である．しかし，対訳文対の採取は，かなり困難な場合が多い．通常の辞書では，掲載されている文がそのままテキストになっている．つまり，例文はテキストの中に埋め込まれている．そのため対訳文対は特定の記号や空白などを手がかりにして採取する必要がある．したがって対訳文対の採取は辞書ごとに異なるプログラムを書く必要がある．

   なお，例外として対訳文対が用例ファイルになっている辞書がある．例として，斎藤英和大辞典（3.13 節）があげられる．このような辞書は，簡単に対訳文対を採取することが可能である．しかし種類は少ない．

2. 電子辞書，独自フォーマット （分類番号 2）

   電子辞書では独自のフォーマットをとり，専用のブラウザでなければ見えないものがある．これらの解析は非常に手間のかかる．特に辞書に圧縮がされている場合や外字がある場合，解析は困難である．しかし，ランダムハウス英語辞典（3.17 節）は，歴史のある辞書であるため，フォーマットを解析して EPWING 形式に変換するツールが Internet 上で掲載されている．また，ビジネス技術実用英語大辞典（3.18 節）は解析可能であった．類似の辞書は他にもあるが，解析に時間がかかるため，対訳文対の採取は行わなかった．

3. CD-ROM 付の書籍 （分類番号 3）

   最近は CD-ROM 付の書籍が販売されている． この中から，日英の対訳のある書籍を選んで，簡単なスクリプトをつくることで，対訳文対が採取できる．ただし，1 冊において得られる対訳文対は少ない．例として，“英文ビジネスレター実用フォーマット” （3.35 節）と “英文 E メール文例集”（3.36 節）がある．

[*1] murakami@ike.tottori-u.ac.jp
[*2] http://www.cybersoken.com/

4. Internet （分類番号 4）
   Internet 上に公開されている対訳文対がある．基本的には，中学校や高校における英文法の教育用の定型文である．代表的な例としてアルク社がある． ただし，一定の時期にしか公開されていない．対訳文対を採取した例として，英語教師用データベース（3.24 節）がある．
5. 新聞記事 （分類番号 5）
   大手の新聞社では，日本語の記事と英語の記事が同時に発行されている．朝日新聞と Asahi Evening News，読売新聞と The Daily Yomiuri，毎日新聞と Mainichi Daily News がある．これらは個別に CD-ROM で購入できる．しかし，対訳文対の採取は，原文が記事対応にすらなっていないため，非常に困難である．しかし，日本文と英文の対訳文対を自動的に採取する研究があり，これを利用して，約 20 万文対の対訳文対が作成されている [Utiyama (2003)]．ただし記事対応のテキストから自動的に対訳文対を採取しているため，他のコーパスから採取した対訳文対と比較すると英文の誤りが多い（5.1 節）．また，元の新聞記事の CD-ROM の販売価格は非常に高価である．
6. 対訳文対，無償 （分類番号 6）
   無償で配布している対訳文対がある．多くは個人が収集したコーパスである．研究用には，自由に使用可能と思われる．例としては，田中コーパス (3.32 節) がある．この対訳文対は，元兵庫大学の故田中康仁氏が収集したものである．単文が多くを占めていて，約 20 万文対ある．ただし，学生が英文を作成したため，英文の品質は低い．
7. 対訳文対，販売 （分類番号 7）
   わずかな例ではあるが，英日の対訳文対として販売されているコーパスがある．研究用には，自由に使用可能と思われる．例としては，英文ビジネスレター文例大辞典 (3.8 節) がある．
8. 対訳文対，未販売 （分類番号 8）
   基本的には，個人もしくは会社が翻訳の研究のために収集したコーパスである．翻訳の研究のために作成されたものであるため，対訳文対として最適な文対が得られる．しかし，残念なことに一般の人には入手不可能である．例として IPAL （3.2 節）がある．この分類に属するコーパスの多くは，故池原悟氏が電電公社および NTT に在籍しているときに作成した対訳文対である．
9. その他 （分類番号 9）
   現在，大規模な日英対訳コーパスが公開され始めている．一方で公開されていないコーパスもある．以下に，これらの例を挙げる．

（a）特許
   日本文と英文の特許文から対訳文対を採取し，これを利用して翻訳のコンテストが開催されている [Fujii (2010)]．利用可能な対訳文対は 500 万文対を超える．現在入手可能な日英の文単位の対訳文対として最大であろう．ただし，特許文であるため通常の日本文とかなり異なる．また，同一の日本文と英文の特許から，プログラムで自動的に対訳文対を作成しているため [Utiyama (2003)]，誤った対訳文対がある．

（b）旅行対話タスク
   旧 ATR 現 NICT では，旅行文コーパス（BTEC : Basic Travel Expression Corpus）を収集している．全部で約 70 万文対あると思われる．残念ながら，一般に公開されていない．しかし，IWSLT のコンテストに参加した場合，約 2 万文対が利用可能である．

（c）Wikipedia 日英京都関連文書対訳コーパス
   NICT の MASTAR プロジェクトにおいて，2010 年に『Wikipedia 日英京都関連文書対訳コーパス Version 2.0』が公開された．Wikipedia を翻訳したもので，合計約 50 万文対ある．

（d）青空文庫とプロジェクト杉田玄白
   現在 Internet 上において，著作権が切れた本を掲載する青空文庫が発足している [42]．これらの本に対して翻訳をする杉田玄白プロジェクト [43] がある．このプロジェクトを利用することで対訳文が得られるが，残念ながら文単位の対訳文対になっていない．

（e）みんなの翻訳
   「みんなの翻訳」は，翻訳文を共有することで，翻訳の効率改善と発展を目的として発足した [44]．ただし，プロジェクト杉田玄白と同様，文単位の対訳文対になっていない．

（f）白書
   日本国憲法や政府機関が発行する白書（教育白書など）を英文に翻訳したテキストがある．ただし，訳文は，文単位の対訳文対になっていない．

（g）NHK
   NHK はニュースなどの 2 カ国語放送があるため，大量の対訳文対がある．しかし，放送法による制約のため外部への提供はできないと思われる．

（h）外国語大学
   各外国語大学には，独自で収集した対訳文対を持っている教官がいるが，全容は不明である．

（i）日本語 WordNet
   日本語 WordNet[45] には 48,276 の例文が掲載されている．ただし，句と文が入り交じってる．また文頭が大文字になっていないなどの問題点がある．

（j）EDICT
   EDICT[46] には例文が掲載されている．ただし，日本文と英文の対訳文対に変更するのは，やや困難である．

3 利用したコーパス

以下に本論文で利用した各コーパスの概略を述べる． なお，アンダーラインで囲まれたイタリックの文章は，各コーパスの販売用の案内文や紹介文の抜粋である．なお節の角括弧の番号は参考論文の番号である．また節の丸括弧のアルファベットは表 11 および表 12 の ID 番号である．

3.1 機能試験文集 [1](AA,BC)

分類は 8．一部は分類 4． 故池原悟氏が NTT に在籍のときに作成し，機械翻訳システム評価を目的とした対訳文対である． オリジナルは 5,240 文対であるが，公開されている文対は 3,718 文対である．現在は，“http://www.kecl.ntt.co.jp/mtg/resources/index-j.php” にて公開されている．なお，表 11，表 12 において “機能試験文集（公開）” は公開されている 3,718 文対である．

3.2 IPAL[2] (AB)

分類は 8. 原典は，情報処理振興事業協会 (IPA) が作成した日本語の辞書である．IPAL の辞書の一部の『計算機用日本語基本動詞辞書 IPAL』『計算機用日本語基本形容詞辞書 IPAL』『計算機用日本語基本名詞辞書 IPAL』に含まれる日本文を，NTT の研究所で英文に翻訳した．

なお，現在 IPAL は，GSK[言語資源協会] において無料配布されている．

3.3 学研 アンカー和英辞典，アンカー英和辞典 [3] (AC,AD)

分類は 1．英和約 51,000 語を収録した『ニューアンカー英和辞典』，和英約 25,000 語を収集した『ニューアンカー和英辞典』の CD-ROM である．

3.4 学研 英和辞典 (AE)

分類は 8. 学研が，アンカー和英英和辞書を発売する前に販売していた辞書である．この辞書は現在販売されていない．この辞書の例文を，NTT の研究所が人手で入力した対訳文対が存在する．

3.5 外国人のための基本語用例辞典 [4](AF)

分類は 8. 文化庁から出版されている “外国人のための基本語用例辞典” の日本文を，MTT の研究所で英文に翻訳した．鳥取県岩美町の澤田信一氏が訳したとのデータがある．

3.6 三省堂 英語表現辞典 [5](AG)

分類は 8. 英語表現辞典の中の例文を，MTT の研究所が人手で入力した対訳文対である．

3.7 日本経済新聞 (AH)

分類は 8. 日本経済新聞の日本文を，NTT の研究所で英文に翻訳したコーパスである．新聞記事のため，日本文が非常に長い．そのため英文も長い．

3.8 英文ビジネスレター文例大辞典 [6](AI)

分類は 7. 文単位の英日対訳文対として販売されている，希少なコーパスである．ただし，現在販売が中止されていると思われる．

3.9 外国人のための日本語例文・問題シリーズ [7](AJ)

分類は 8. 外国人のための日本語例文・問題シリーズの中の日本文を，NTT の研究所で翻訳した．助詞，語彙，形式名詞，表記法，談話の構造などの編に分かれている．

3.10 自然発話音声・言語データベース (LDB)[8] (AK)

分類は 7. ATR が作成した．自動音声の研究のためにホテル対話などが収録されている．コーパスには音声とテキスト両者が含まれる．このコーパスには，約 20 万文対の対訳文対がある．

3.11 SENSEVAL 対訳コーパス [9] (AL)

分類は 4. senseval は，語の意味的曖昧性解消のためのコンテストである．この日本語タスクには，辞書タスクと翻訳タスクがあり，翻訳タスクでは，日本語単語に対する適切な英訳を選択することを目指している．このタスクのために対訳コーパスが作成されている．ただし，このコーパスの多くは単語の訳であり，対訳文対は少数である．

3.12 講談社和英辞典 [10] (AM)

分類は 6. 講談社和英辞典から電子技術総合研究所において人手によって入力されたデータである．校正していないと思われる箇所が多く，文字誤りが多い．対訳文対は約 58,000 文対を含んでいる．現在産業技術総合研究所から入手することが可能である．ただし，研究目的に限る．そして，使用のための誓約書を産業技術総合研究所に提出する必要がある．

3.13 斎藤和英大辞典 [11] (AN)

分類は 1. 1938 年に故斎藤秀三郎氏が出版した辞書．見出し語 5 万，用例 12 万，総頁数 4640 頁，当時としては前例のない大和英辞典である．斎藤は「日本人の英語はある意味で日本化されなくてはならない」と，当時としてはユニークな見解を述べており，そのため非常に癖のある辞書となっている．使用されている日本語は，やや古めかしく，かつ差別用語が含まれる．英文も意訳（一部は超訳）が多く，他の辞書に見られない個性豊かな辞書である．賞賛する研究者も多い．

なお，斎藤秀三郎は昭和 4 年に 64 歳で亡くなっているため，この辞書の著作権は切れている可能性がある．ただし，斎藤氏の原稿は 「H 」の項までしか作成していないらしい．なお，対訳文対は用例ファイルとして本文と別になっているため，対訳文対の採取は容易である．

3.14 小倉書店 英語文型・文例辞典 [12] (AO)

分類は 7. 文単位の英日対訳文対が販売されている希少な例である．なお，原文はテキストではなく HTML 形式である．自然科学系の論文，報告書，仕様書あるいは書簡文などの構成に必要な表現例を項目別に編集したコーパスである．

3.15 英辞郎 用例コーパス [13] (AP)

分類は 7.「英辞郎」とは，プロの翻訳者・通訳者で構成されるグループ (EDP) が制作する英和辞書の名称である．英辞郎には，一般的な単語はもちろんのこと，スラング，イディオム，ビジネス用語，経済用語，法律用語，特許用語，コンピュータ用語，科学技術用語，医学用語，固有名詞 (組織名・企業名・人名・国名・映画名) などが含まれている．現在見出し数では，150 万を超えていて，日本最大の辞書になっている．ただし，句の対訳が多く，対訳文対は少ない．

3.16 研究社 新編英和活用大辞典 [14] (AQ)

分類は 1. 例文は比較的分かりやすい規則で収録されている．ただし，得られた対訳文対と公表されている用例数には，大きな差がある．

3.17 ランダムハウス英語辞典 [15] (AR)

分類は 2. このコーパスは独自のフォーマット形式を持っている．しかし，EPWING のフォーマットに変換するツールが Internet で公開されているため，対訳文対を採取することができる．また，英文は高品質で，英語の native でなければ思いつかない文があるが，日本文は，日本語の native にとって奇異な文がある．

3.18 ビジネス技術実用英語大辞典 [16] (AS)

分類は 1. 用例が別ファイルになっている．このコーパスの対訳文対は，他のコーパスと比較すると，理系向きの文章が多く，かつ高品質である．

3.19 コンピュータ用語辞典第 3 版 [17] (AT)

分類は 1. カタログには “例文 12,600 件（延べ）を収録” とあるが，これは英和，和英それぞれの例文の合計を表していると思われる．英和の例文と和英の例文は大部分が重複している．

3.20 佐良木コーパス (AU)

分類は 8. 日本大学の佐良木昌氏が個人的に収集した対訳文対である．

3.21 白井コーパス (AV,BD)

分類は 8. 元 NTT-AT の故白井諭氏が個人的に収集した対訳文対である．

3.22 斎藤健太郎コーパス:比較構文 (AW)

分類は 6. 元鳥取大学工学部知能情報工学科池原研究室の斎藤健太郎氏が，様々な本から比較構文のみを収集した対訳文対である．

3.23 澤田康子コーパス:因果関係構文 (AX)

分類は 6．元鳥取大学工学部知能情報工学科池原研究室の澤田康子氏が，様々な本から因果関係構文のみを収集した対訳文対である．

3.24 アルク 英語教師用データベース [18] (AY)

分類は 4．アルクが公開している英語教師用の対訳文対である．ただし，現在公開が中止されている．

3.25 研究社 総合ビジネス英語文例事典 [19] (AZ)

分類は 1．この辞書では例文が複数行で掲載されている，そのため，対訳文対の採取はかなり複雑になる．

3.26 新実用英語ハンドブック [20] (BA)

分類は 1．この辞書では例文が複数行で掲載されているため，対訳文対の採取はかなり複雑になる．

3.27 研究社 新和英大辞典 [21] (BB)

分類は 1．この辞典の例文は比較的分かりやすい規則で収録されている．公表されている例文数は約 5 万であるが，採取すると約 20 万文対が得られる．

3.28 三省堂 エクシード英和辞典 [22](BE)

分類は 8． 元 NTT-AT の故白井諭氏が個人的に収集した対訳文対である．辞典を人手により入力したコーパスである．

3.29 科学技術日英・英日コーパス辞典 [23](BF)

分類は 2．独自のフォーマット形式である．今回採取した辞書のなかでフォーマットの情報が最も少なかった．用例は多い．

3.30 日本語文型辞典 [24](BG)

分類は 8．日本語文型辞典は，外国人が日本語を勉強するために書かれた日本語の例文集である．この例文を鳥取大学の故池原悟氏が CREST[池原 (2000] の費用で英訳した．

3.31 旺文社マルチ辞書 辞ショック [25](BH)

分類は 1．

3.32 田中コーパス [26](BI)

分類は 6．元兵庫大学の故田中康仁氏が学生と作成した対訳文対である．日本文から英文を作成している．主に学生が翻訳したため，英文の品質にバラツキがある．対訳文対の量は，約 20 万文対ある．この対訳文対は，過去に Internet で一般に公開されていた．そして，誓約書を書くことで全対訳文対が入手可能であった．しかし，現在公開されていないようである．

3.33 読売新聞社説 (BJ)

分類は 8．人手によって読売新聞の社説と The Daily Yomiuri と比較して，文単位に編集された対訳文対である．日本大学の佐良木昌氏が個人的に収集したコーパスである．

3.34 アルク 英語表現辞典 [27](CA,CB,CC,CD,CE,CF,CG)

分類は 4．このコーパスは，7 つに分類される．

3.35 英文ビジネスレター実用フォーマットと例文集 [28] (CH)

分類は 3．高島康司著．数少ない CD-ROM 付の日英の対訳のある書籍である．簡単なスクリプトを作ることで，対訳文対が採取できる．ただし，得られる対訳文対は少ない．

3.36 英文 E メール文例集 [29](CI)

分類は 3．向井京子著．数少ない CD-ROM 付の日英の対訳のある書籍である．簡単なスクリプトを作ることで，対訳文対が採取できる．ただし，得られる対訳文対は少ない．

3.37 読売新聞記事 [30](CK,CL)

分類は 5．読売新聞の英字新聞として The Daily Yomiuri がある．これらは，別々に販売されているが，同じ内容の記事が載っている．しかし，両者の関係は，対訳文対どころか，記事単位の対訳にすらなっていない．しかし，このようなコーパスから文単位の対訳文対を自動的に採取する研究があり，これを利用して，約 20 万文対の対訳文対が作成されている [Utiyama (2003)]．そして，読売新聞の記事を購入することを前提に，入手可能である．ただし，読売新聞の CD-ROM は非常に高価である．1 年分の日本文記事がアカデミック価格で 12 万円，英文記事が 11 万円である．また記事対応になっていないテキストから自動的に対訳文対を採取しているため，他のコーパスと比較すると対訳文対になっていない文対が多い（5.1 節参照）．つまり，日本文から見ると，英文の誤りが多い．

3.38 ATR バイリンガル旅行会話データベース [31](CM,CN)

分類は 7． ATR が音声対話システムのために収集した音声と対訳のコーパスである．対話総数 618，発話総数 16,107 文対で構成されている．

3.39 NHK やさしいビジネス英語実用フレーズ辞典 [32](CO)

分類は 3．この辞典は NHK ラジオ講座「やさしいビジネス英語」（杉田 敏）の Vocaburary Building の内容をまとめた本で，CD-ROM 中に対訳文対を含んでいる．

3.40 自然科学系和英大辞典 [33](CQ)

分類は 1．

3.41 ジーニアス英和・和英辞典 [34](CR)

分類は 1．このコーパスでは例文の中の見出しの単語や連語は “〜” で略されている．そのため，対訳文対を採取しても単語を対応づけることが困難であるため，誤りのない対訳文対を採取することは困難である．本研究では，“〜” を見出しの単語に置き換えて，対訳文対を作成した．

3.42 最新ビジネス英文手紙辞典 [35](CS)

分類は 3．このコーパスは，英文の手紙の例文を収集している．

3.43 機械を説明する英語 [36](CT)

分類は 7．

4　得られた対訳文対の量

4.1　得られた対訳データベースの文対の数

得られた対訳文対の数を表 11 中の「採取文数」に示す．この数字は，文章と推定して採取した文対の数であり，クリーニングが完全でないため，誤った対訳文対が含まれている．様々なコーパスから対訳文対を採取して，総計として 1,099,093 対訳文対を採取した．これらのコーパスでは，約 70% が単文と認定できる文であった．約 20% は重文・複文と認定できる文であった．そして残りの約 10% は．複雑な重文・複文で文長が長い文であった．また大多数は，テキスト文であるが，一部には，対話文があった．対話文の多くは旅行会話である．なお，本論文では，この対訳文対を対訳データベースと呼ぶ．

4.2　得られた単文データベースの文対の数

採取した対訳データベースから，日本文において単文の対訳文対を採取する．一般的には，単文の定義は「述語が一つだけから成る文」であるが，この条件では定義できない文が多い．本研究では，単文の条件を以下のように定義する [西山 (2005)]．

1. 文末以外に動詞が一つもなく，文末が動詞で終わる文
   * 彼は毎日自転車に乗る。
2. 文中に動詞がなく，文末が複合動詞で終わる文
   * ドイツは新しい歴史への一歩を踏みだした。
3. 文中に動詞，複合動詞，形容詞が一つもなく，文末以外に形容詞が一つもなく文末が形容詞で終わる文
   * この林檎はややすっぱい。
4. 文中に動詞，複合動詞，形容詞が一つもなく，文末以外に形容動詞がが一つもなく文末が形容動詞で終わる文
   * 企業の経営戦略は大切だ。
5. 文中に動詞，複合動詞，形容詞，形容動詞が一つもなく，文末が “名詞＋付属語” で終わる文
   * あの人こそ真の英雄だ。
6. 疑問文，命令文，会話文は対象外
   * 疑問文：この本は何について書いてあるか。 * 命令文：そこに私のテントを張れ。 * 会話文：昨日どこかへ行ったかい。

また，日本語で分類を行っているため，英文は複文になっている対訳文対がある．本論文では，この対訳文対を単文データベースと呼ぶ．得られた単文データベースの文対の数を表 11 の「単文」に示す．また，抽出した単文の例を表 1 に示す．

単文データベースは，できるだけ簡潔で問題のない単文のみを選択したため，対訳データベース（4.1 節）の多くの単文が，利用されなかった．最終的には，単文データベースは総計で 182,113 対訳文対となった．ただし，単文データベースは，全ての文対を人手でクリーニングしていない．

表 1　単文データベースの例

日本文 1	猫が縁側をのそのそ歩いている。
英文 1	The cat is walking across the veranda .
日本文 2	新宿のネオンサインが消えた。
英文 2	The neon sign lights of Shinjuku were turned off .
日本文 3	銃声が一発聞こえた。
英文 3	A loud report of a gun was heard .

4.3　得られた重文・複文データベースの文対の数

採取した対訳文対から，人手によって文を解析して，重文および複文を採取した．ただし，分類は日本文で行っている．そのため英文は単文になっている対訳文対がある．採取した重文・複文の文対の数を表 11 の「重文・複文」に示す．総計として 158,633 対訳文対を採取した．なお，本論文では，この対訳文対を重文・複文データベースと呼ぶ．

4.4　重文・複文の文種別

採取した重文・複文データベースを，以下に示す文種別 1 から 5 に従って人手で分類した．分類は日本文で行っているため，英文が単文である対訳文対もある．分類して得られた文対の数を，表 11 中の文種別 1 から 5 に示す．ただし，埋め込みについて修飾要素を持たない用言（連体形）は埋め込み文としない．

1. 文種別 1 （重文）
   文接続を一つ持つ文である．いわゆる重文である．例文を表 2 に示す．

表 2　文種別 1 の例

日本文 1	窓を開けると冷たい風が入ってくる。
英文 1	When you open the window a cold wind blows in.
日本文 2	彼女に会ったおかげで一日が楽しかった。
英文 2	Seeing her made my day.
日本文 3	もし行きたいのなら行きなさい。
英文 3	You can go if you choose to go.

2. 文種別 2 （重文）
   文接続を二つ持つ文である．いわゆる重文である．例文を表 3 に示す．

表 3　文種別 2 の例

日本文 1	昔の友達に久しぶりに会って、夜を更かして語り合った。
英文 1	I met an old friend for the first time in a long time, and we chatted late into the night.
日本文 2	私はカギをなくしてしまったので、妻が帰るまで、待たなければならなかった。
英文 2	I lost my key, so that I had to wait till my wife returned.
日本文 3	その肉はよく煮ないと、かたくて食べられない。
英文 3	If you do not cook this meat well, it will be too tough to eat.

3. 文種別 3 （複文）
   埋め込み文を一つ含む文である．いわゆる複文である．例文を表 4 に示す．

表 4　文種別 3 の例

日本文 1	それを知らぬ者は誰もいない。
英文 1	There is no one but knows that.
日本文 2	誰も完全に幸福な者はいない。
英文 2	None are completely happy.
日本文 3	船が水平線以下に隠れるまで見送った。
英文 3	I followed the ship with my eyes till she disappeared below the horizon.

4. 文種別 4（複文）

埋め込み文を二つ含む文である．いわゆる複文である．例文を表 5 に示す．

表 5　文種別 4 の例

日本文 1	通りの方へ向かっている窓と中庭に向かっている窓がある。
英文 1	Some windows look out on the street, the others look out into the yard.
日本文 2	山を汚す者に山を楽しむ資格はない。
英文 2	People who leave trash in the mountains are not qualified to enjoy them.
日本文 3	彼がこつこつと金を貯めているのは、外国へ旅行するためです。
英文 3	He is diligently saving money in order to travel overseas.

5. 文種別 5 （重複文）

文接続を一つと埋め込み文を一つ含む文である．いわゆる重複文である．例文を表 6 に示す．

表 6　文種別 5 の例

日本文 1	ドアの開く音がかすかに廊下に響いた。
英文 1	The sound of the door opening echoed faintly down the corridor.
日本文 2	昔のことを思い出すと重苦しい悲しみが彼女の心をおおった。
英文 2	A leaden grief swept over her at the thought of her past.
日本文 3	彼は家を建てるために節約してお金を貯めている。
英文 3	In order to build a house, he is economizing and saving money.

5　考察

5.1　採取した対訳文対の誤り調査

採取した対訳文対の精度を調査するために，人手による誤り調査を行った．単文データベースからランダムに 100 文対抽出した．この 100 文対を調査したところ， 4 文対に誤りが検出された．誤りが検出された文を表 7 に示す．

出典を調査したところ，日本文 1 は講談社和英辞典で，日本文 2 はランダムハウス英語辞典で，日本文 3 と日本文 4 は，読売新聞（文対応データ）であった．講談社和英辞典は，入力誤りが多いコーパスである．ランダムハウス英語辞典は，特殊フォーマットであるため，採取に誤りが生じたと考えている．また，読売新聞（文対応データ）は，記事対応の日英対訳文から，プログラムで，自動的に文対応の対訳文対を作成している．そのため誤りが多いと考えている．

なお，重文・複文データベースは，全文を人手により検査して修正を行っている．そのため，重文・複文データベースを単文データベースと同様に調査したが，誤りは見つからなかった．

表 7　単文データベースにおいて誤りが発見された文対

日本文 1	どんな もの か 頓 と 見当 も つか ない 。
英文 1	I have no ideaof what it is like . （“idea of” の誤り.）
日本文 2	彼 から 事業 全体 を 残ら ず 買い取っ た 。
英文 2	We bought the whole business from him,lock,stock , and barrel . （“,lock,stock , and barrel” が不要）
日本文 3	国 の 工業 用 エタノール 買い取り 価格 の 半額 という 安 さ だ
英文 3	The expected price of the produced ethanol will be only half of the 100,000 yen that the government pays to purchase a kiloliter of industrial ethanol , the sources said . （英文が明確に誤っている.）

5.2　対訳文対の品質の問題

採取した対訳文対には，対訳文対として不適切と思われる文対が存在する．意味は解るが英文の品質に問題があると思われる．例文対を表 8 に示す．なお，欠陥英和辞典の研究 [副島 (1989)] において，辞書の例文には，英文として適切ではない例文が存在することが報告されている．

表 8　対訳文対の品質が問題になる例

日本文 1	嫌悪の叫びをあげた。
英文 1	She cried out in revulsion. (日本語に “彼女は” がない.）
日本文 2	顧客サービス部をお願いします。
英文 2	Give me Customer Service. （“Customer Service” は，電話対応における “顧客サービス部” であり，対面の場合は正しいのか？）
日本文 3	皇太子ご夫妻は九三年六月に結婚された。
英文 3	The crown prince and the princess married in June 1993 . （日本語において一九九三年とすべきである．）

本研究では，可能な限り多くの電子媒体から，可能な限り品質の高い対訳文対の採取を試みた．しかし，精度の高い対訳文対を大量に得ることは困難であり，問題のある対訳文対と誤りのある対訳文対の混入は避けられなかった（5.1 節）．著者の意見として，翻訳品質の高い対訳文対を 100 万文対，収集することは，かなり困難であると考えている．

5.3 辞書間の類似文の存在

採取された例文を調査すると，異なる辞書において，似た文章が多く掲載されている，例えば，複数の辞書において，英文で “Wine is made from grapes.” を検索した日本文を表 9 に示す．このような例が多くの辞書において見られる．

表 9 “Wine is made from grapes.” の対訳文

ワインはぶどうから作られる	ワインはぶどうでつくる	ブドウ酒はブドウからつくられる
ワインはブドウから作る	ぶどう酒は葡萄より作られる	ワインはブドウで作る

5.4 過去の辞書の例文における著作権の問題

過去に辞書に集録された例文が問題になった例として，以下の事例が報告されている [副島 (1990)]．1967 年に発行された研究社・新英和中辞典（初版）（岩崎民平・小稲義男）は，開拓社の新英英大辞典 (ISED) とオックスフォード辞書現代英英辞典 (OALD) の例文を，大量に引用した．そのため，ISED と OALD の著者である A.S.Hornby と ISED の発行元である開拓社から抗議をうけて，翌 1968 年，例文を変更して，研究社・新英和中辞典第 2 版が出版された．

6 コーパス等の外部提供等に関する著作権法等との抵触について [藤波 (2012)]

6.1 はじめに ー著作権問題ー

デジタル化・ネットワーク化の急速な進展に伴う著作物の利用形態には，既設の著作権法権利制限規定により可能である行為と実質的に同様の行為も多いのですが，権利制限規定が個別具体の事例に沿って定めていることから，たとえ権利者の利益を不当に害しないものであっても形式的には違法となってしまうとの指摘（「デジタル・ネット時代における知財制度の在り方について」検討経過報告：平成 20 年 5 月 29 日，同．報告：平成 20 年 11 月 27 日，知的財産戦略本部／デジタル・ネット時代における知財制度専門調査会）がなされており，これに対応する著作権法改正が行なわれ平成 22 年 1 月 1 日に施行されました．

この改正著作権法では，研究開発における情報利用の円滑化を図る目的での権利制限規定も新設され，例えば，画像・音声・言語・ウェブ解析技術等の研究開発過程での著作物等の利用について著作権法上の問題が生じるなどの指摘に対して，これを適法化する立法的解決が図られています．

しかし，新設された情報解析のための複製等の権利制限規定（著作権法：第四十七条の七　以後”著 47 条の 7”と略します．）の文言は，著作物その他の記録または創作した二次的著作物の記録を含む翻案への言及に留まり，作成したコーパスの外部提供等やコーパスが依拠した記録の外部提供等についての明文規定は省かれています．このため，著 47 条の 7 等に依拠して作成したコーパスの外部提供等や利用について，主に技術系研究者から様々な意見—-「公開不可で第三者がダウンロードして評価することができない」，「どのような利用方法まで許容されるのか不明であり，確定判決（最高裁判決？）を得るまで他者は利用できない」等々—-が出され，情報解析のための複製等の権利制限規定（著 47 条の 7）の立法趣旨は勿論，著作物利用に係る判例などもが没却されているようにもみえます．

本稿では，著 47 条の 7 の解釈や適用範囲を画する立法事実，著作物性がなく著作権法の保護を受けない表現の利用行為，行為の外形上は著作権法の保護対象となる利用行為に当たるものの著作権法が保護すべきものとして本来想定しているような利用行為とは利用形態が異なる利用行為，違法の疑義がある利用行為についての適正手続き（deu process of law）について，文化審議会資料や関連判例などに依拠し，作成したコーパスや依拠した記録の外部提供等などに係る著作権法等との抵触の存否と適法性を確保する要件について述べます．

なお，著 47 条の 7 の解釈等にかかる考察につき「私見が述べられているにすぎない…」との意見もありますが，法令の公権的解釈権は裁判所が有し，その解釈は具体的争訟性事件性のある法律上の争訟について行なわれる（裁判所法 3 条）日本の司法制度下では，裁判「所」以外が示す法令の解釈や意見は全て「私見」であり，本稿も同様です．

6.2 著作権法：（情報解析のための複製等）第四十七条の七 (著 47 条の 7)

著作権法：（情報解析のための複製等）第四十七条の七を以下に示します．

著作物は，電子計算機による情報解析（多数の著作物その他の大量の情報から，当該情報を構成する言語，音，映像その他の要素に係る情報を抽出し，比較，分類その他の統計的な解析を行うことをいう．）を行うことを目的とする場合には，必要と認められる限度において，記録媒体への記録又は翻案（これにより創作した二次的著作物の記録を含む．）を行うことができる．ただし，情報解析を行う者の用に供するために作成されたデータベースの著作物については，この限りでない．

6.3 情報解析技術の研究開発に着目した権利制限およびその射程範囲

著 47 条の 7 に係る立法者の意思を以下の節に示します（文化審議会著作権分科会報告書 平成 21 年 1 月），本条項の適用・解釈は，「これらの文化的所産の公正な利用に留意しつつ，著作者等の権利の保護を図り，もつて文化の発展に寄与する」（著 1 条）ことを目的として，以下（1,2 節）に沿った適用・解釈が行なわれることになります．

1. 情報解析技術の研究開発に着目した権利制限の根拠
（文化審議会著作権分科会．報告書 88 頁）
高度情報化社会の下で，取り扱われる情報量が爆発的に増大する中，利用者が必要とする情報・知識を抽出し，高度な知的処理を実現する情報解析技術は，デジタル・ネットワーク社会の基盤となるものであり，そのための研究開発も社会的に意義を有する．
情報解析分野の研究開発は，著作物の表現そのものを利用するものではなく，その情報・アイディアの抽出を行うに過ぎないが，その過程で中間的に利用行為に当たる行為を伴うものであり，著作物利用の実質を備えないとの側面もある
2. 権利制限を行う場合の要件
（文化審議会著作権分科会．報告書 88-91 頁）
（a）営利・非営利の別
非営利のものに限定する必要はないと考えられる．その場合に著作権者等の利益が害されるおそれがあるとするならば，次の要件設定ー著作権者等の利益への影響（6.3 節 2b）ーより対応すべきである．
（b）著作権者等の利益への影響
契約によって入手可能なデータベース等の場合には権利制限を認める必要はない．このような意見に照らせば，既存のビジネスの中で研究開発に必要なデータベース等が有償で提供されているような場合，その他，著作物の性質や利用態様等に応じて著作権者等の利益を害すると考えられるような場合には，権利制限の対象外とすることが適当と考えられる．
（c）研究開発の過程で作成された複製物の外部提供等
権利制限が情報抽出のための過程で中間的に行われる複製であることに着目したものであるとの側面からは，基本的に，当該複製物を外部に提供することはその趣旨に反することになるため，当該複製物を研究に参加しない者に提供する行為については権利制限の対象外とすべきと考えられる．なお，研究過程で作成された複製物の外部提供の取扱いと関連して，研究開発を行う者にそのためのデータベース等を提供するような事業があった場合にこれが権利制限の対象となるかどうかについては，このようなデータベース等の作成自体が研究開発目的の者によって行われているかどうかで判断すべきとの指摘がある．

6.4　著作物性のない著作物（非保護著作物）等の利用

表現物の利用については，「既存の著作物に依拠して創作された著作物が，思想，感情若しくはアイディア，事実若しくは事件などの，表現それ自体でない部分又は表現上の創作性がない部分において，既存の著作物と同一性を有するにすぎない場合には，翻案には当たらない」（最一小判平成 13.6.28 江刺追分事件）ことから，思想，感情，アイディアの表現（例：大阪高判平成 6.2.25 脳波数理解析論文事件，東京高判平成 12.3.29 エスニシティ論文事件）はもちろん，著作物性が認め難い，短い表現（例：東京地判平成 13.5.30 チャイルドシート事件，知財高判平成 17.10.6 ヨミウリオンライン事件），ありふれた表現（例：東京地判平成 7.12.18 ラストメッセージ in 最終号事件），選択幅が狭い表現（例：東京地判平成 13.1.23 多摩地図事件，知財高判平成 20.7.17 ライブドア裁判傍聴記事件，東京地判平成 10.11.30 版画写真事件），機能表現（例：東京地判平成 15.1.31 電車線設計プログラム事件，大阪地判平成 4.4.30 丸棒矯正機設計図事件）などは，原則として著作権法では保護されず，利用することができます．また，「著作物の創作的特長を感得できないときは，複製等の著作物の利用には該当しない」とされ，著作物を利用している実態がないときは著作権は動かない（違法ではない）とされています（雪月花事件：東京地裁平成 11 年 10 月 27 日判決，はたらくじどうしゃ事件：東京地裁平成 13 年 7 月 25 日判決）．

更に「複製とは，既存の著作物に依拠し，その内容及び形式を覚知させるに足りるものを再製すること」（最一小判昭和 53.9.7 ワン・レイニー・ナイト・イン・トウキョー事件）とされ，「言語の著作物の翻案とは，既存の著作物に依拠し，かつ，その表現上の本質的な特徴の同一性を維持しつつ，具体的表現に修正，増減，変更等を加えて，新たに思想又は感情を創作的に表現することにより，これに接する者が既存の著作物の表現上の本質的な特徴を直接感得することのできる別の著作物を創作する行為」と定められている（最一小判平成 13.6.28 江刺追分事件）ことから，これらの要件のどの 1 つを欠いても，その利用行為は翻案にも複製権侵害にもなりません．

6.5　論点と対応

情報解析の研究開発における既存著作物等の利用目的は，著作物の思想，表現そのものを感じ取るのではなく，その中から研究開発に必要な部分を探し当てること，アイディアや背景情報等を抽出することなどであって，人間が行ったとするならば視聴行為として著作権が及ばない行為です．しかし，同様の行為をコンピュータ等に実行させるときは，中間的に既存著作物等を蓄積する必要があるために，物理的には，複製行為や翻案行為が行われることになります．著 47 条の 7 はこれらの行為を，権利制限規定を新設することで適法化して立法的解決を図ったものです．

著 47 条の 7 は「著作物は，電子計算機による情報解析（多数の著作物その他の大量の情報から，当該情報を構成する言語，音，影像その他の要素に係る情報を抽出し，比較，分類その他の統計的な解析を行うことをいう．…）を行うことを目的とする場合には，必要と認められる限度において，記録媒体への記録又は翻案（これにより創作した二次的著作物の記録を含む．）を行うことができる．」と規定し，情報解析目的での記録媒体への記録又は翻案を適法としています．但し，「必要と認められる限度において」要件を設けており，「必要と認められる限度」とは何かが論点となります．なお，ただし書きでは「情報解析を行う者の用に供するために作成されたデータベースの著作物については，この限りでない．」と定め，立法時に付された権利制限の要件（6.3 節 2b 参照：契約によって入手可能なデータベース等の場合には権利制限を認める必要はない）を明記しています．

著 47 条の 7 は本条に依拠して作成したコーパス，研究開発の過程で作成された複製物および依拠した記録の外部提供についての明文規定はなく，これらの外部提供（の可否や要件）が論点となります．また，著作権，著作者および実演家の人格権，著作隣接権等に対する侵害は親告罪とされており，その侵害について刑事責任を追及するかどうかは被害者である権利者の判断に委ねられ，犯人を知った日から 6 か月を経過したときは告訴することができませんし（刑訴法 235 条：告訴期間），被害者の被害感情や被害の重み，訴追意思は，公訴提起の判断において重視され，一般に，被害者の意思と全く無関係に訴追が行われることはありません（同．248 条：起訴便宜主義）．従って，当該利用方法が予期しなかった権利侵害の疑義が生じたときは，権利者への対応や権利侵害（民法 709 条：不法行為）を防止する措置の有無や内容により侵害認定が異なることから，対応や防止措置が論点になります．

以下，著 47 条の 7 に係る，次の論点について検討します．

論点 1　必要と認められる限度

「研究開発における情報利用の円滑化」との課題に対応して新設された著 47 条の 7 を，著作権法が本来想定している保護範囲と外形上はその利用行為に当たるものの利用の実質を備えない行為との調整である（文化審議会著作権分科会報告書 86 頁）と解すると，「必要と認められる限度」は，権利の例外を設ける際の一般的要件の充足が要件となります．著作権関係条約における権利例外を設ける際の基準は，一般的には表 10 に示す「スリーステップテスト」が基準になっています．

表 10　スリーステップテストの 3 要素

1	特別の限定された場合であること，	2	通常の利用を妨げないこと，	3	権利者の利益を不当に害しないこと

著 47 条の 7 に依拠する記録媒体への記録又は翻案は，この 3 要件を充足する限度で行為することが求められています．

なお，営利・非営利の要件については，本件権利制限の根拠を情報解析技術に関する研究開発の社会的意義等に求める考え方に照らせば，非営利のものに限定する必要はありません．ただし，著作物の性質や利用態様等に応じて著作権者等の利益を害すると考えられるような場合には，権利制限の対象外とすることが適当と考えられます（同．88-89 頁）．

論点 2　研究開発の過程で作成された複製物の外部提供

本件権利制限が情報抽出のための過程で中間的に行われる複製であることに着目した側面とその趣旨から，当該複製物を研究に参加しない者に提供する行為は権利制限の対象外とすべきと考えられます．なお，研究開発を行う者にそのためのデータベース等を提供するような事業があった場合に，これが権利制限の対象となるかについては，このようなデータベース等の作成自体が研究開発目的の者によって行われているかどうかで判断すべきである（同．89 頁）ことから，研究開発目的の者による研究開発の過程で作成された複製物の外部提供は適法と解され得ます．ただし，この場合であっても，権利例外を設ける際の基準である「スリーステップテスト」の基準を充足した外部提供方法であることを要します．

なお，情報解析のために複製・翻案した複製物や二次的著作物の保存を禁止する著作権法条項はありませんので，複製物や二次的著作物の保存は適法行為と解されます

また，元の複製者が，作成された著作物の複製物や二次的著作物を送信可能な状態にすることを禁止する条項もありませんので，送信や送信可能化する行為も適法と解されます．しかし，複製物・二次的著作物を送信可能な状態におくことは，受信者の目的が法で定められたものであるかの確認ができないことから，当該物を送信可能な状態にした者の不法行為責任（民法 709 条）が生じることがありますので，受信者の目的確認および相応の使用契約に合意した者だけが利用できるなどの措置を採ることが必要になります．

論点 3　コーパスの外部提供

コーパスは，情報抽出過程で中間的に存在する記録から抽出された表現などで構成されています．コーパスはデータベースの著作物であり，通常は，コーパス作成者がデータベースの著作権者であり，コーパスの部分を構成する表現が権利処理された保護著作物または非保護著作物であるときは，当該コーパスを外部提供することができます．

コーパスが著 47 条の 7 に依拠して作成されたときは，その複製物および（作成されたコーパスが新しい著作物か二次的著作物かの論点はあるが）二次的著作物の目的外使用が問題になります．即ち，「第四十七条の七に定める目的以外の目的のために」著作権制限規定の適用を受けて作成された著作物の複製物を利用した場合（著 49 条第 1 項第 5 号）あるいは，同条の規定適用を受けて作成された二次的著作物を利

用した場合（著 49 条第 2 項第 6 号）は違法（複製権侵害）とされます．ただし，著 49 条の文言が「…を利用した者」となっていることに留意した対応が必要になります．

以上により，「第四十七条の七に定める目的」でのコーパスの外部提供は適法行為と解されます（著 49 条第 1 項第 5 号，同. 第 2 項第 6 号の反対解釈）．また，外部提供行為は非営利のものに限定する必要はないと考えられます（文化審議会著作権分科会報告書 88 頁）．

論点 4　コーパスが依拠した原記録の外部提供

コーパスの的確な利用には，コーパス作成に係る原表現記録から当該コーパスの利用領域や利用可能深度を検討する必要があります．故に，言語解析の研究開発過程では，単語や文のつながりなどの用例をウェブ上で検索・表示可能にすることが行なわれて，機械翻訳等に関する研究開発，辞書・文法書の編纂，言語研究等にコーパスが用いられています（文化審議会著作権分科会報告書 85 – 86 頁）．著 47 条の 7 は，著作権法が本来想定している保護範囲と外形上はその利用行為に当たるものの利用の実質を備えない行為との調整である（同．85-86 頁）ことから，著 47 条の 7 に依拠する複製物をウェブ上で検索・表示可能にするなどの行為は適法行為と解されます．但し，本件複製は「必要と認められる限度」で認められることから，検索・表示可能にする記録は「スリーステップテスト」の基準を充足する限度内であることが必要です．

論点 5　外部提供に係る他人の権利や法律上保護される権利の侵害防止

コーパスなどの外部提供ではコーパス作成者も予期しなかった権利侵害の疑義が生じて，正当な権利者や法律上保護される利益を有する者から侵害防止の措置等が求められることがあります．このため，著作権者から保護著作物の削除要求（プロバイバー責任制限法類推）があった場合の削除手続き（オプトアウト：Opt-Out．個人情報保護法 23 条類推）や外部提供に係る権利侵害の申し出手続きと外部提供者が採る権利侵害防止措置などを予め定めて，それらを利害関係人が知り得る状況を作出しておく必要があります．また，著作権法が親告罪であることに鑑み，特にコーパスが依拠した原記録の外部提供に際しては，提供の趣旨，提供情報の非保護著作物化，適法な提供先であることの確認方法など，提供時の適正手続きの制定や実行を可視化して，利害関係人の事前の理解を得ることも必要になります．

6.6　まとめ　ー著作権問題ー

著 47 条の 7 等に係るコーパス等の外部提供等に関する著作権法等との抵触に係わる論点と結論を以下に示します．

論点 1　記録媒体への記録又は翻案における「必要と認められる限度」

「スリーステップテスト」（表 10）基準の 3 要件を充足する限度に限られます．

論点 2　研究開発の過程で作成された複製物の外部提供

研究開発目的の者による研究開発の過程で作成された複製物に限り，研究に参加した者および研究開発を行なう者に限定して，提供することができます．ただし，その提供態様が「スリーステップテスト」の 3 要件を充足しないときは，この限りではありません．

論点 3　コーパスの外部提供

著 47 条の 7 が定める目的に限定して，外部提供できます．また，提供は非営利のものに限定されません．

論点 4　コーパスが依拠した原記録の外部提供

ウェブ上で検索・表示可能にするなどの態様で，外部提供できます．ただし，検索・表示可能にする記録は「スリーステップテスト」の基準を充足する限度内であることが必要であり，それを担保する相応のシステム対応などが必要になります．

論点 5　外部提供に係る他人の権利や法律上保護される権利の侵害防止

権利者から侵害防止の措置等が求められた場合の対応措置（情報削除等の申出・実行・確認等の手続きなど）を予め定めて公開するなど，侵害防止のための適正手続きが必要になります．

6.7　おわりに　ー著作権問題ー

言語解析などの情報解析技術の研究開発で求められていた著作権の権利制限規定等の措置は，平成 22 年 1 月 1 日施行の改正著作権法で，ほぼ終えたものと解されており，新設された権利制限規定等をどのように解し用いて研究開発を推進し成果を社会に供するかは技術者側の対応に委ねられています．もちろん，これらの規定に明文の規定がなく個別具体的な事例に即し判例などで定まる論点もありますが，ある限度で結論が定まる論点がほとんどであり，その限度内で研究開発を進めることが求められています．

なお，コーパス等の外部提供等については，幇助，間接侵害，プロバイダ責任制限法との関係など余の論点も多々ありますが，本稿では割愛します．また，本稿は法律専門語で論じるべき問題を通常の言葉で記述していることから，用いる語彙や意味，論旨展開が法学で用いるモノと異なる箇所があり，内容も法学的厳密性を必ずしも具備していないことに留意してください．

7　まとめ

日英対訳辞書は，翻訳の研究において必要不可欠のものである．しかし，日本文と英文が文単位に対応していて，量が多く，一般の人が入手可能な対訳文対は，最近まで存在していなかったと言える．

本研究では，様々な電子媒体から，日本文と英文が文単位に対応する対訳文対を採取した．電子媒体として，CD-ROM・Internet・電子辞書などを利用した．これらの結果，対訳文対として 1,099,093 万文対を採取した（対訳データベース）．そして，得られた対訳文対から，日本文において単文の対訳文対を 182,113 文対，採取した（単文データベース）．また日本文において重文・複文の対訳文対を 158,633 文対，採取した（重文・複文データベース）．また，重文・複文の対訳文対を採取した重文・複文データベースを，文種別 1 から 5 に従って人手で分類した．

ただし，単文データベースは，コストの問題から全てを人手でクリーニングできていない．そのため 100 文対を調査したところ 4 文対の誤りがあった．一方重文・複文データベースは，全文を人手により検査して修正を行っている．そのため誤りが発見できなかった．しかし，日本文と英文を比較すると，翻訳精度の高い文対になっていない場合がある．これらを考慮すると，翻訳品質の高い対訳文対を 100 万文対，収集することは，かなり困難であると考えている．

なお，統計翻訳の目的のために，様々な電子媒体から対訳文対を採取することは，著作権法：(情報解析のための複製等）第四十七条の七から，問題ないと考えている．そして，採取した対訳文対を特定のグループ内で利用することも，許される行為と考えている．ただし，“必要と認められる限度” を，常識的に判断する必要がある．

謝辞

ここで紹介したコーパスは，長年翻訳に携わってきた多くの方々の努力を，著者が覚え書きとしてまとめたものです．日本大学の佐良木昌氏には，個人的に収集して頂いた対訳文対を利用させて頂きました．電子辞書からの対訳文対の採取には，当時鳥取大学工学部知能情報工学科の片山慶一郎氏（既卒）の助力を得ました．重文・複文の分類は，NTT-AT の，木村淳子氏，小見佳恵氏，阿部さつき氏，竹内（村本）奈央氏，小船園望氏が中心になって行いました．鳥取大学の徳久雅人氏および山梨英和大学（岐阜大学）の池田尚志氏には，この辞書の作成にあたり，様々な協力を得ました．以上の方々に感謝いたします．最後に，元兵庫大学の田中康仁氏と元鳥取大学（NTT) の池原悟氏と元 NTT-AT（NTT,ATR) の白井諭氏は鬼籍に入られました．ご冥福をお祈りいたします．

参考文献

[1] 池原悟, 白井諭, 小倉健太郎, “言語表現体系の違いに着目した日英機械翻訳試験項目の構成”, 人工知能学会論文, Vol.9, No.5, pp.569-579, 1994.

[2] 桑畑 和佳子, 橋本 三奈子, 村田 賢一, “計算機用日本語辞書の開発”, 情報処理学会研究報告, 人文科学とコンピュータ研究会報告 93(42), 27-34, 1993.

[3] “学研 ニューアンカー英和・和英辞典,” B298C49I1, 2000.

[4] “外国人のための基本語用例辞典 （第 3 版）” 文化庁国語課約 4,500 語, 大蔵省印刷局, ISBN 4-17-151302-2, 2000.

[5] “英語表現辞典”, 三省堂編修所, ISBN 4-385-11012-3, 1997.

[6] “英文ビジネスレター文例大辞典 Ver.2[CD-ROM]”, 日本経済新聞社出版局, ISBN 453245509X. 1997.

[7] “外国人のための日本語例文・問題シリーズ”, 荒竹出版.

[8] “自然発話音声・言語データベース（日英対訳)”, http://www.atr-p.com/sdb.html

[9] 黒橋禎夫, 白井清昭, ”SENSEVAL-2 日本語タスク”, 信学技報, NLC 101(351), 1-8, 2001.

[10] “講談社和英辞典”, 講談社, ISBN-13: 978-4061210530, 1982.

[11] 斎藤秀三郎, “斎藤和英大辞典”, 日外アソシエーツ辞書編集部編, EPWING , ISBN4-8169-8078-4, 1999.

[12] “英語文型・文例辞典”, 小倉書店, http://www.ogurashoten.co.jp/kyozai3.html, 1997.

[13] “英辞郎”, アルク, http://shop.alc.co.jp/cnt/eijiro/.

[14] “新編英和活用大辞典” 研究社, ISBN4-7674-3573-0, 1995.

[15] “ランダムハウス英語辞典 第二版 CD-ROM 版”, 小学館, http://ebook.shogakukan.co.jp/scatalog/randam/top/top.htm, 2002.

[16] 海野文男, 海野和子, “ビジネス技術実用英語大辞典”, 日外アソシエーツ, ISBN4-8169-8127-6, T4937695181270, 2000.

[17] “CD-コンピュータ用語辞典 第 3 版英和・和英/用例・文例パッケージ”, コンピュータ用語辞典編集委員会〔編〕, 日外アソシエーツ, ISBN4-8169-8126-8, T4937695181263, 2000.

[18] “英語教師用データベース”, アルク, http://home.alc.co.jp/db/owa/engt_structure?stg=4

[19] “研究社ビジネス英語スーパーパック”, 研究社出版, ISBN4-7674-3590-0, 1998.

[20] “新実用英語ハンドブック”, 大修館書店, ISBN4-469-74233-3, 1995.

[21] “新和英大辞典”, 研究社, ISBN 4-7674-7200-8, 2003.

[22] “エクシード英和辞典”, 三省堂編修所, ISBN 4-385-10650-9, 1998.

[23] “科学技術日英・英日コーパス辞典”, 丸善, ISBN4-621-04991-7, 2002.

[24] “日本語文型辞典”, くろしお出版, ISBN-10: 4874241549, ISBN-13: 978-4874241547, 1998.

[25] “旺文社版マルチ辞書 W 辞ショック”, 株式会社アスク, AWR1-00430, 1997.

[26] 田中康仁 (兵庫大学), “日英・パラレルコーパスの作成”, 言語処理学会 第 8 回 年次大会, B4-2, pp.499-502, 2003.

[27] “英語表現辞典”, アルク, http://www.alc.co.jp/eng/kaiwa/hyogen/index.html

[28] 高島康司, “英文ビジネスレター実用フォーマットと例文集”, ベレ出版, ISBN4-939076-25-3, 2000.

[29] 向井京子, “英文 E メール文例集”, 池田書店, ISBN4-262-16896-4, 2002.

[30] 内山将夫, 井佐原均, “日英新聞記事の対応付けと精度評価”, 第 151 回 自然言語処理研究会第 68 回 情報学基礎研究会, pp.15-22, 2002.

[31] 竹沢寿幸, 白井諭, 大山芳史, “バイリンガル旅行会話コーパスに見られる話し言葉の特徴分析”, 自然言語処理研究会報告, pp.137-144, 2001.

[32] 杉田敏 編, “NHK やさしいビジネス英語実用フレーズ辞典” NHK 出版, ISBN978-4-14-034102-5, 2003.

[33] “自然科学系和英大辞典” 小倉書店, ISBN-13: 978-4902764000, 1997.

[34] “ジーニアス英和・和英辞典 CD-ROM 版” , 大修館書店, ISBN4-469-79057-5, 2001.

[35] フランシス・J・クディラ (著), 朝日出版社, “最新ビジネス英文手紙辞典 CD-ROM 新訂版”, ISBN-13: 978-4255002927, 2004.

[36] “機械を説明する英語”, アスク, ASIN: B00008HV7V, 1999.

[37] “リーダーズ英和辞典”, 研究社, ISBN4-7674-3563-3, 1999.

[38] 木原研三, 小西友七, 他, “新グローバル＆ニューセンチュリー英和・和英辞典”, ISBN4-385-61400-8, JAN コード：T4938641614002, 1994.

[39] “ＣＤ- 科学技術４５万語対訳辞典”, 日外アソシエーツ, ISBN4-8169-8128-4, 2001.

[40] “新英和・和英中辞典”, 研究社, ASIN: B0009EWFA0, 2005.

[41] EPWING, http://www.epwing.or.jp/about/about.html

[42] 青空文庫, http://www.aozora.gr.jp/guide/nyuumon.html

[43] プロジェクト杉田玄白, http://www.genpaku.org/

[44] みんなの翻訳, http://trans-aid.jp/

[45] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/

[46] EDICT, http://www.csse.monash.edu.au/ jwb/edict.html

[池原 (2000] 池原悟, “セマンティック・タイポロジーによる言語の等価変換と生成技術”, (平成 13 年度～18 年度：科学技術振興事業団・戦略的基礎研究 (CREST)), 2000.

[Utiyama (2003)] Masao Utiyama, et al. “Reliable Measures for Aligning Japanese-English News Articles and Sentences”, ACL-2003, pp.72-79, 2003.

[言語資源協会] 言語資源協会 (GSK), E-mail: info@gsk.or.jp, Web: http://www.gsk.or.jp/. http://www.gsk.or.jp/catalog/GSK2007-D/catalog.html

[Fujii (2010)] Atsushi Fujii, et al. , “Overview of the Patent Translation Task at the NTCIR-8 Workshop”, Proceedings of the 8th NTCIR Workshop Meeting, 2010.

[A. S. Hornby (1942)] A. S. Hornby E. V. Gatenby A. H. Wakefield, “新英英大辞典（机上版）(ISED, Idiomatic snd Syntactic English Dictionary)”, ISBN : 978-4-7589-0005-8, 1942.

[副島 (1989)] 副島隆彦, et al, “欠陥英和辞典の研究 別冊宝島”, JICC 出版, 別冊宝島 102, 雑誌 65988-55, 1989.

[副島 (1990)] 副島隆彦, Peter Van Gelder, “英語辞書大論争”, JICC 出版, 別冊宝島 113, 雑誌 65988-77, 1990.

[西山 (2005)] 西山 七絵, 村上 仁一, 徳久 雅人, 池原 悟, “単文文型パターン辞書の構築”, 言語処理学会第 11 回年次大会, pp.372-375 (2005-03).

[藤波 (2012)] 藤波 進, “これだけは知っておきたい技術者のための法律相談 法制度からみた情報流通システムの設計と運用 ”, http://www.cybersoken.com/lawlecture/index.html.

表 11　採取文数

ID	コーパス名	分類	文献	採取文数	単文	重文・複文
AA	機能試験文集	8	[1]	4,853	2,310	1,628
AB	IPAL	8	[2]	15,213	10,813	
AC	学研 アンカー和英辞典	1	[3]	46,108	14,556	14,816
AD	学研 アンカー英和辞典	1	[3]	25,274		7,011
AE	学研 英和辞典	8		4,063	1,626	846
AF	外国人のための基本語用例辞典	8	[4]	26,861	6,113	11,796
AG	三省堂 英語表現辞典	8	[5]	16,316	5,572	6,356
AH	日本経済新聞	8		6,675	952	
AI	英文ビジネスレター文例大辞典	7	[6]	12,544	1,020	5,667
AJ	外国人のための日本語例文・問題シリーズ	8	[7]	14,086	3,601	
AK	LDB	7	[8]	13,107		
AL	SENSEVAL 対訳コーパス	4	[9]	1,205	598	
AM	講談社 和英辞典	6	[10]	45,143	16,554	11,993
AN	斎藤 和英大辞典	1	[11]	94,646		19,313
AO	小倉書店 英語文型・文例辞典	7	[12]	2,133	382	835
AP	英辞郎 用例コーパス	7	[13]	16,613		3,301
AQ	研究社 新編英和活用大辞典	1	[14]	114,531	40,682	31,999
AR	ランダムハウス英語辞典	2	[15]	42,942	14,948	11,317
AS	ビジネス技術実用英語大辞典	1	[16]	13,228	2,727	5,138
AT	コンピュータ用語辞典第 3 版	1	[17]	4,067		1,705
AU	佐良木コーパス	8		1,051	20	160
AV	白井コーパス	8		1,426	43	367
AW	斎藤健太郎コーパス:比較構文	6		199	43	83
AX	澤田康子コーパス:因果関係構文	6		655	29	463
AY	アルク 英語教師用データベース	4	[18]	802	76	429
AZ	研究社 総合ビジネス英語文例事典	1	[19]	2,685	53	451
BA	新実用英語ハンドブック	1	[20]	333	126	86
BB	研究社 新和英大辞典	1	[21]	35,268	9,977	8,597
BC	機能試験文集（公開）	4	[1]	2,146		
BD	白井 2 コーパス	8		46,173		4,278
BE	三省堂 エクシード英和辞典	8	[22]	2,175	670	186
BF	科学技術日英・英日コーパス辞典	2	[23]	14,522		5,860
BG	日本語文型辞典	8	[24]	10,004	4,123	3,952
BH	旺文社マルチ辞書 辞ショック	1	[25]	64,511	27,620	
BI	田中コーパス	6	[26]	198,567		
BJ	読売新聞社説	8		560	69	
CA	アルク なるほど!英語表現データベース	4	[27]	5,997		
CB	アルク 状況別英語表現集	4	[27]	2,742		
CC	アルク 日本を紹介するキーワード	4	[27]	216		
CD	アルク カタカナ表現	4	[27]	454		
CE	アルク 四字熟語	4	[27]	300		
CF	アルク ことわざ・慣用句	4	[27]	459		
CG	アルク 擬音語・擬態語	4	[27]	327		
CH	高島康司 英文ビジネスレター	3	[28]	1,093		
CI	向井京子 英文 E メール文例集	3	[29]	1,091	264	
CJ						
CK	読売新聞 (文対応データ)	5	[30]	150,000	12,806	
CL	読売新聞 (記事対応データ)	5	[30]	5,015		
CM	ATR バイリンガル旅行会話基本構文表現一般	7	[31]	2,144		
CN	ATR バイリンガル旅行会話基本構文表現ダ文	7	[31]	608		
CO	NHK やさしいビジネス英語実用フレーズ辞典	3	[32]	7,276	773	
CP	赤尾好夫 英語基本熟語集	8				
CQ	小倉書店 自然科学系和英大辞典増補改訂新版	1	[33]	10,315		
CR	ジーニアス英和・和英辞典	1	[34]	5,394	2,330	
CS	朝日出版社 最新ビジネス英文手紙辞典 CD-ROM 版	3	[35]	2,338	176	
CT	アスク 機械を説明する英語	7	[36]	2,639	461	
ZZ	自然言語処理専門用語辞書	6				
Total				1,099,093	182,113	158,633

表 12　重文・複文の文種別

ID	コーパス名	分類	文献	重文・複文	文種別 1 (重文 1)	文種別 2 (重文 2)	文種別 3 (複文 1)	文種別 4 (複文 2)	文種別 5 (重複文)
AA	機能試験文集	8	[1]	1,628	772	55	612	54	142
AB	IPAL	8	[2]						
AC	学研 アンカー和英辞典	1	[3]	14,816	7,294	65	497	563	1,335
AD	学研 アンカー英和辞典	1	[3]	7,011	2,990	214	2,954	372	500
AE	学研 英和辞典	8		846	389	16	379	25	38
AF	外国人のための基本語用例辞典	8	[4]	11,796	6,226	1,347	2,461	344	1,420
AG	三省堂 英語表現辞典	8	[5]	6,356	3,344	310	1,964	171	570
AH	日本経済新聞	8							
AI	英文ビジネスレター文例大辞典	7	[6]	5,667	1,553	411	1,903	745	1,055
AJ	外国人のための日本語例文・問題シリーズ	8	[7]						
AK	LDB	7	[8]						
AL	SENSEVAL 対訳コーパス	4	[9]						
AM	講談社 和英辞典	6	[10]	11,993	6,343	445	4,220	292	709
AN	斎藤 和英大辞典	1	[11]	19,313	10,844	1,078	5,396	469	1,578
AO	小倉書店 英語文型・文例辞典	7	[12]	835	308	51	292	59	126
AP	英辞郎 用例コーパス	7	[13]	3,301	1,531	117	1,237	146	307
AQ	研究社 新編英和活用大辞典	1	[14]	31,999	13,153	822	14,240	1,450	2,347
AR	ランダムハウス英語辞典	2	[15]	11,317	5,505	306	4,432	328	772
AS	ビジネス技術実用英語大辞典	1	[16]	5,138	1,497	301	2,071	539	733
AT	コンピュータ用語辞典第 3 版	1	[17]	1,705	541	102	696	160	206
AU	佐良木コーパス	8		160	29	14	51	29	37
AV	白井コーパス	8		367	46	22	77	33	49
AW	斎藤健太郎コーパス:比較構文	6		83	25	14	27	7	10
AX	澤田康子コーパス:因果関係構文	6		463	323	36	46	8	51
AY	アルク 英語教師用データベース	4	[18]	429	185	56	54	30	104
AZ	研究社 総合ビジネス英語文例事典	1	[19]	451	126	35	104	77	109
BA	新実用英語ハンドブック	1	[20]	86	39	2	36	1	8
BB	研究社 新和英大辞典	1	[21]	8,597	4,590	253	2,984	172	598
BC	機能試験文集（公開）	4	[1]						
BD	白井 2 コーパス	8		4,278	2,218	32	1,929	31	83
BE	三省堂 エクシード英和辞典	8	[22]	186	118	1	62	1	9
BF	科学技術日英・英日コーパス辞典	2	[23]	5,860	1,955	562	1,692	578	1,068
BG	日本語文型辞典	8	[24]	3,952	1,950	418	976	134	476
BH	旺文社マルチ辞書 辞ショック	1	[25]						
BI	田中コーパス	6	[26]						
BJ	読売新聞社説	8							
CA	アルク なるほど!英語表現データベース	4	[27]						
CB	アルク 状況別英語表現集	4	[27]						
CC	アルク 日本を紹介するキーワード	4	[27]						
CD	アルク カタカナ表現	4	[27]						
CE	アルク 四字熟語	4	[27]						
CF	アルク ことわざ・慣用句	4	[27]						
CG	アルク 擬音語・擬態語	4	[27]						
CH	高島康司 英文ビジネスレター	3	[28]						
CI	向井京子 英文 E メール文例集	3	[29]						
CJ									
CK	読売新聞 (文対応データ)	5	[30]						
CL	読売新聞 (記事対応データ)	5	[30]						
CM	ATR バイリンガル旅行会話基本構文表現一般	7	[31]						
CN	ATR バイリンガル旅行会話基本構文表現ダ文	7	[31]						
CO	NHK やさしいビジネス英語実用フレーズ辞典	3	[32]						
CP	赤尾好夫 英語基本熟語集	8							
CQ	小倉書店 自然科学系和英大辞典増補改訂新版	1	[33]						
CR	ジーニアス英和・和英辞典	1	[34]						
CS	朝日出版社 最新ビジネス英文手紙辞典 CD-ROM 版	3	[35]						
CT	アスク 機械を説明する英語	7	[36]						
ZZ	自然言語処理専門用語辞書	6							
Total				158,633	73,894	7,085	51,392	6,818	14,440

テキストの硬さと軟らかさの考察
—『現代日本語書き言葉均衡コーパス』の収録書籍を対象に—

柏野 和佳子 *（国立国語研究所 言語資源研究系）
立花 幸子（国立国語研究所 コーパス開発センター）
保田 祥（国立国語研究所 コーパス開発センター）
丸山 岳彦（国立国語研究所 言語資源研究系）
奥村 学（東京工業大学 精密工学研究所）
佐藤 理史（名古屋大学 大学院工学研究科）
徳永 健伸（東京工業大学 大学院情報理工学研究科）
大塚 裕子（はこだて未来大学 メタ学習センター）
佐渡島 紗織（早稲田大学 留学センター）

Analysis of Textual Formality and Informality:
In the Case of the Book Samples
in the Balanced Corpus of Contemporary Written Japanese

Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Sachi Yasuda (Center for Corpus Development, NINJAL)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institute of Technology)
Satoshi Sato (Graduate School of Engineering, Nagoya University)
Takenobu Tokunaga (Dept. Computer Science Graduate School of Information Science and Engineering, Tokyo Instiute of Technology)
Hiroko Otsuka (Center for Meta-Learning, Future University Hakodate)
Saori Sadoshima (Center for International Education, Waseda University)

１．はじめに

我々は大規模なコーパスを様々な学術研究や教育に活用するためには，テキストを所望の目的で分類するための分類指標が必要だと考え，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」において，書籍テキストの多種多様な形式，内容，表現に関わる特徴を捉えるための分類指標の設計と検証とを進めている。

これまでに，テキストの内容や表現に関わる分類指標として「対象読者（難易），主観的・客観的，硬軟，丁寧さ，直接的な語り性の有無」の 5 つの分類指標を設計し，『現代日本語書き言葉均衡コーパス』（BCCWJ）[1]の図書館サブコーパスに収録される書籍テキストを対象に，人手付与を試行した（柏野・奥村 2012）。

それらのうち，「硬軟」と「丁寧さ」の分類指標の付与は，硬い印象を与えるテキスト，軟らかい印象を与えるテキスト，丁寧さを感じるテキスト，くだけた印象を与えるテキストの分類と抽出を目的とするものである。

本稿では，まず，試行したアノテーション作業の概要を述べ，「硬軟」と「丁寧さ」の分類指標の付与結果を取り上げて報告する。そして，付与結果に基づき，テキストの硬軟，丁寧さについて考察する。

* waka@ninjal.ac.jp

[1] http://www.tokuteicorpus.jp/を参照。

２．アノテーション作業

２.１　分類指標の設計

BCCWJ には，出版サブコーパス（10,117 サンプル），図書館サブコーパス（10,551 サンプル），特定目的サブコーパスの一つであるベストセラー(1,390 サンプル), あわせて約 22,000 の書籍サンプルが収録されている。それらには，NDC（日本十進分類法）によるジャンルや，C コード（日本図書コード）による販売対象，発売形態，また，著者情報，形態論情報などが付与されており，それらを利用して，半自動的に種々の観点から分類することは可能である。しかしながら，EAGLES(1996)がコーパスへ付与することが望ましいと挙げる，(A)対象読者に想定される読解レベル（難易度），(B)テキストの作成意図，(C)さまざまな文体情報の 3 種に関する情報は C コード以外には与えられておらず，それらの観点によるテキストの分類や抽出は困難である。そこで，(A)を補う「対象読者（難易）」，(B)を補う「主観的・客観的」，(C)を補う「硬軟」「丁寧さ」「直接的な語り性の有無」という，あわせて 5 つの分類指標を新たに設計した（柏野・奥村 2012）。

「(C)さまざまな文体情報」とは，EAGLES(1996)では定義が困難だと述べられている。複数のパラメータが議論されているが，標準は定まっていないと言う。しかし，たとえば学習者にとって重要な情報になり得るものであり，Joos(1961)の提案（"frozen"，"formal"，"informal"，"colloquial"，"intimate"）や，Halliday et al. (1964)の提案（"colloquial"，"polite"，"casual"，"intimate"，"deferential"）が紹介され，語レベルの文体情報（どのような文体で用いられる語であるか）は各種辞書に工夫されて記載されていると述べられている。

つまり，ここでコーパスに備えることが望ましいと議論されている「文体情報」とは，形式性，親疎性，口語性に関わる文体情報だと言える。よって，その形式性，親疎性を問うものとして「硬軟」と「丁寧さ」の指標を，口語性を問うものとして「直接的な語り性の有無」という指標を設けた。

指標の付与に際しては，「硬軟」は「硬いか軟らかいか」という選択肢にて，「丁寧さ」は「丁寧かくだけているか」という選択肢にて判断することとした。「硬い」とは，かしこまっている感じ，堅苦しい感じであり，「軟らかい」とは，かしこまっていない感じ，親しみやすい感じである。また，「丁寧」とはフォーマルな感じであり，その反対のフォーマルではない感じを「くだけている」という言葉で表すこととした。「くだけている」は「丁寧」の対義語であるという印象を持ちにくいが，「丁寧」の対義語として浮かびやすい「ぞんざい，乱暴，粗雑」といった語はネガティブな印象のみが強いため，「くだけている」を用いることとした。この時，「硬くてくだけている」というテキストは想定できなかったため，次のとおり「硬軟」と「丁寧さ」を組み合わせて同時に問う選択肢を設けた。しかしながら，「硬軟」と「丁寧さ」は関連性は強いが異なる軸として，次の付与作業段階では選択肢を分ける計画でいる。

1 とても硬くて丁寧
2 どちらかといえば硬くて丁寧
3-1 どちらかといえば軟らかくて丁寧
3-2 どちらかといえば軟らかくてくだけている
4-1 とても軟らかくて丁寧
4-2 とても軟らかくてくだけている

２.２　アノテーション作業の概要

アノテーション作業の概要は，次のとおりである。

- 作業目的：人手付与の作業上の問題点の検討，典型例の抽出，分類指標の検証及び基準の検討。
- 対象テキスト：BCCWJに収録されている図書館サブコーパス（10,551サンプル）よりランダムに抽出したサンプルのテキスト。（本稿作成時，合計3,324テキストへ付与済み。）
- 1テキストの範囲と長さ：コーパス収録テキストの分類指標とするため，その一部を字数を揃えて抽出することはせず，1サンプル全体を範囲とする。1テキストの平均はおよそ

3,000語。

●作業ファイル：サンプルを取得した書籍の紙面コピーの電子化ファイルを参照する。

●作業態勢：判断のゆれを検証するために1作業につき，作業者3人を確保した。同一の判定作業を3人がそれぞれ独立して行う。

●作業量：１セット約400～500の書籍テキストに対する指標付与を延べ約10日で行う。

●作業指示：付与すべき指標の種類をごく簡単な説明のみで指示。

また，作業手順は次のとおりである。

①形式による判定を行う。構造的に単純なテキストタイプ（例：章節構造）であれば細分類の対象とする[2]。

②細分類をする。「対象読者」「主観的・客観的」「硬軟」「丁寧さ」「直接的な語り性の有無」の分類指標を付与する。

２.３　アノテーション作業の結果

対象とした 3,324 テキストのうち，さらに細分類の対象となる「構造的に単純なテキストタイプ」と判断されたものは，2,672 テキストであった。このうち，「硬軟」「丁寧さ」の分類指標付与については，2 テキストに付与の欠落があったため，合計 2,670 テキストが付与済みのものとして得られた。付与結果の例を表 1 に示す。

表 1 「硬軟」と「丁寧さ」の付与結果例

タイトル	硬軟		
	A	B	C
犬がころんだ	4-2 とても軟らかくてくだけている		
天才の法則	2 どちらかといえば硬くて丁寧		
夢のハワイ暮らしが実現できる本	3-1 どちらかといえば軟らかくて丁寧		
金田一京助全集	1 とても硬くて丁寧	2 どちらかといえば硬くて丁寧	2 どちらかといえば硬くて丁寧
国民の文明史	1 とても硬くて丁寧	2 どちらかといえば硬くて丁寧	3-2 どちらかといえば軟らかくてくだけている

表 1 にみられるように，3 人の判断が一致するもの，3 人のうち 2 人の判断が一致するもの，3 人ともに一致しないものとがあった。これら判断の一致率，カッパ係数（一致率から偶然の一致率をひいたもの），相関関数については Kashino and Okumura(2010)，柏野・奥村(2012)で報告した。本試行作業においては中～低度の一致であったが，今後のマニュアルの整備等でその一致度が改善する見通しを得ている。

また，判断の一致度をみるために，柏野・奥村(2012)では，各選択肢別に一致した人数とそのテキスト数を示した。「硬軟」と「丁寧さ」に関しては，付与済み 2,670 テキストのうち，全員一致が 387 テキスト，2 人一致が 1,383 テキスト，非一致が 900 テキストであった。各選択肢別の全員一致数，2 人一致数は表 2 のとおりである。

表 2 「硬軟」と「丁寧さ」の選択肢別一致数の内訳

	1.とても硬くて丁寧		2.どちらかといえば硬くて丁寧		3-1.どちらかといえば軟らかくて丁寧	
全員一致	2	0.1%	250	9.4%	60	2.2%
2人一致	50	1.9%	707	26.5%	231	8.7%
小計	52	1.9%	957	35.8%	291	10.9%
	3-2.どちらかといえば軟らかくてくだけている		4-1.とても軟らかくて丁寧		4-2.とても軟らかくてくだけている	
全員一致	53	2.0%	4	0.1%	18	0.7%
2人一致	326	12.2%	34	1.3%	35	1.3%
小計	379	14.2%	38	1.4%	53	2.0%

[2] 対象外とした形式が特徴的なテキスト（例：対談，Q&A 形式，図解，用語解説）については，一定量が分類されてから細分類を検討する予定でいる。

表 2 から「1 とても硬くて丁寧」，「4-1 とても軟らかくて丁寧」，「4-2 とても軟らかくてくだけている」の 3 つの選択肢において全員一致テキストが少なかったことがわかる。

２.４　アノテーション付与済みテキストの NDC 別特徴

「硬軟」と「丁寧さ」の分類指標を付与した 2,670 テキストの NDC 別の特徴を分析した。作業者 3 人の判断一致，不一致に関わらず，各人の選択結果 1 つを 1 点として，各テキストの分類指標の選択肢を点数化した。それを割合になおし，平均との差分を求めた。さらに，各分類指標がどちらに触れているかの尺度を次のとおり求めた。なお，判断のゆれを考慮し，選択肢別に重みづけは行わなかった。

「硬度」（選択肢 1～2 の和と 3～4 の和との差分）

「丁寧度」（選択肢 1,2,3-1,4-1 の和と 3-2,4-2 の和との差分）

この「硬度」を横軸，「丁寧度」を縦軸として NDC ごとの特徴をプロットした結果が，次の図 1 である。

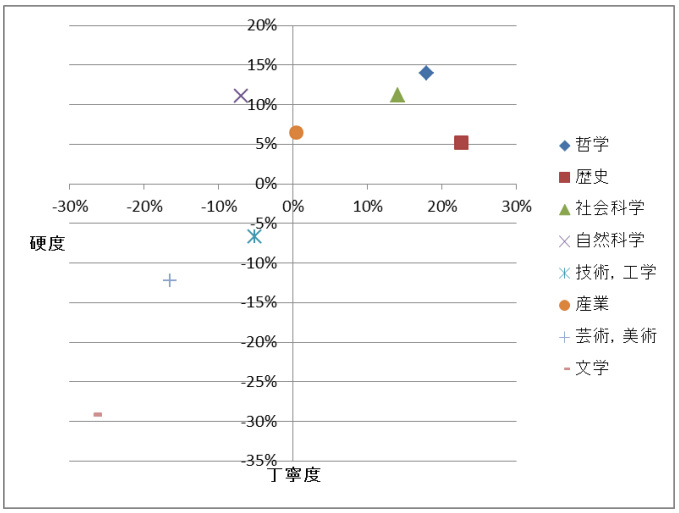


図 1「硬度」と「丁寧度」でみる NDC 別テキストの特徴

図 1 より，各 NDC 別には次の特徴のあることがわかる。

硬くて丁寧：「哲学」，「社会科学」，「歴史」
軟らかくて丁寧：「自然科学」
硬くも軟らかくもなくて丁寧：「産業」
平均的：「技術，工学」
軟らかくてくだけている：「文学」，「芸術，美術」

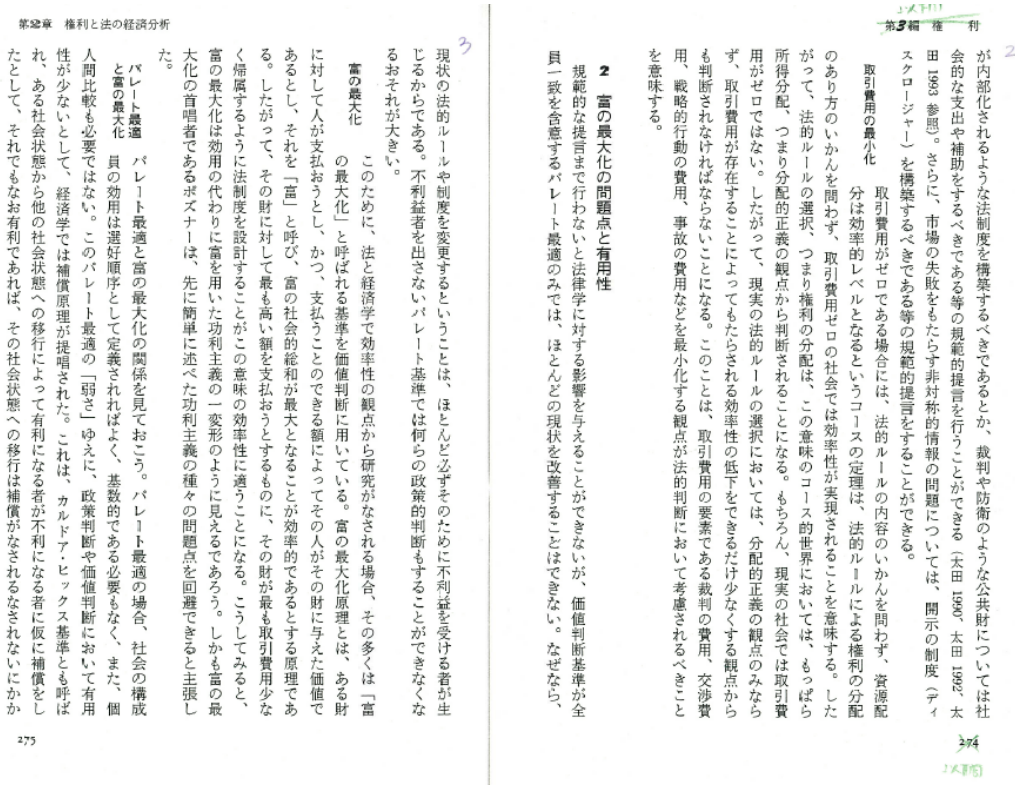
アノテーションの対象を，図書館サブコーパスに収録されたテキストとしたため，この特徴は，BCCWJ の図書館サブコーパスに収録されている書籍の場合の特徴をみていることになる。特に，「自然科学」が「軟らかくて丁寧」なテキストに位置する点が目立つ。また，「文学」だけでなく「芸術，美術」も「軟らかくてくだけている」傾向の強い点も改めて確認することができた点である。アノテーションをすることによって，コーパスに収録されているテキストの特徴分析が可能になる。

３．テキストの硬軟，丁寧さについての考察

３.１　典型例の抽出

テキストの硬軟，丁寧さについて考察するために，アノテーション結果を用いて典型例を抽出した。先の表 2 において全員一致数の少ないところで一致しているものが検討すべき典型例であることがわかった。すなわち，「1 とても硬くて丁寧」からは「硬い印象を与える典型例」として 2 例，「4-1 とても軟らかくて丁寧」からは「軟らかい印象を与える典型例」として 4 例，「4-2 とても軟らかくてくだけている」からは「くだけた印象を与える

典型例」として 18 例を抽出した。なお，「丁寧な印象を与える典型例」としたいものは「軟らかい印象を与える典型例」と重なると考え，典型例は先の 3 種に定めた。次に 1 例ずつ図示する（出典はサンプル ID と書名で記す，色鉛筆は電子化入力の際の指示）。



第2章　権利と法の経済分析

現状の法的ルールや制度を変更するということは、ほとんど必ずそのために不利益を受ける者が生じるからである。不利益者を出さないパレート基準では何らの政策的判断もすることができなくなるおそれが大きい。

富の最大化

このために、法と経済学で効率性の観点から研究がなされる場合、その多くは「富の最大化」と呼ばれる基準を価値判断に用いている。富の最大化原理とは、ある財に対して人が支払おうとし、かつ、支払うことのできる額によってその人がその財に与えた価値であるとし、それを「富」と呼び、富の社会的総和が最大となることが効率的であるとする原理である。したがって、その財に対して最も高い額を支払おうとするものに、その財が最も取引費用少なく帰属するように法制度を設計することがこの意味の効率性に適うことになる。こうしてみると、富の最大化は効用の代わりに富を用いた功利主義の一変形のように見えるであろう。しかも富の最大化の首唱者であるポズナーは、先に簡単に述べた功利主義の種々の問題点を回避できると主張した。

パレート最適と富の最大化

パレート最適と富の最大化の関係を見ておこう。パレート最適の場合、社会の構成員の効用は選好順序として定義されればよく、基数的である必要もなく、また、個人間比較も必要ではない。このパレート最適の「弱さ」ゆえに、政策判断や価値判断において有用性が少ないとして、経済学では補償原理が提唱された。これは、カルドア・ヒックス基準とも呼ばれ、ある社会状態から他の社会状態への移行によって有利になる者が不利になる者に仮に補償をしたとして、それでもなお有利であれば、その社会状態への移行は補償がなされるなされないにかかが内部化されるような法制度を構築するべきであるとか、裁判や防衛のような公共財については社会的な支出や補助をするべきである等の規範的提言を行うことができる（太田 1990、太田 1993、太田 1993 参照）。さらに、市場の失敗をもたらす非対称的情報の問題については、開示の制度（ディスクロージャー）を構築するべきである等の規範的提言をすることができる。

取引費用の最小化

取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。したがって、現実の法的ルールの選択においては、分配的正義の観点のみならず、取引費用が存在することによってもたらされる効率性の低下をできるだけ少なくする観点からも判断されなければならないことになる。このことは、取引費用の要素である裁判の費用、交渉費用、戦略的行動の費用、事故の費用などを最小化する観点が法的判断において考慮されるべきことを意味する。

2　富の最大化の問題点と有用性

規範的な提言まで行わないと法律学に対する影響を与えることができないが、価値判断基準が全員一致を含意するパレート最適のみでは、ほとんどの現状を改善することはできない。なぜなら、

図 2　「硬い印象を与える典型例」（LBi3_00033『現代法社会学入門』）

図86　テコドント

図87　コチロサウルス

恐竜のさいご

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

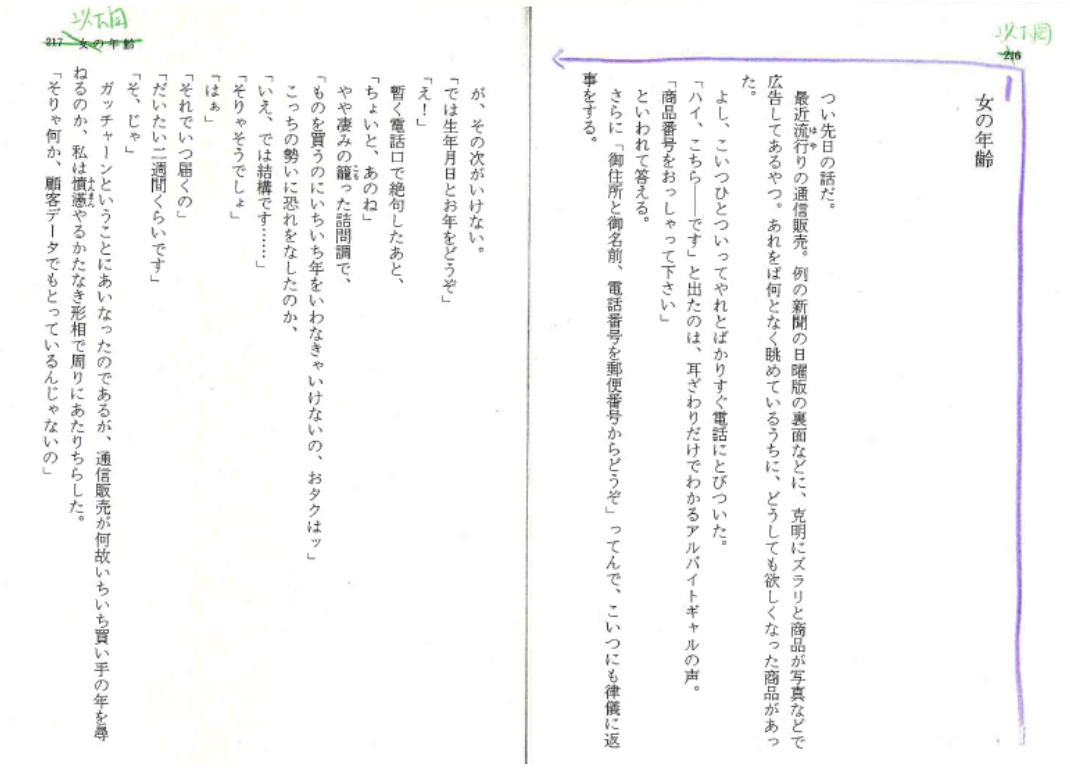
恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三畳紀のはじめにいた、「テコドント」（図86）という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

ではつぎに、テコドントの先祖は、なんだったでしょうか。

古生代のおわりごろ（石炭紀）から中生代のはじめにかけての地層から、「コチロサウルス」（図87）という爬虫類の化石がみつかっています。コチロサウルスのなかまは、四本の足であるき、

図 3　「軟らかい印象を与える典型例」（LBa4_00010『恐竜の世界をたずねて』）



女の年齢

つい先日の話だ。
最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをぼ何となく眺めているうちに、どうしても欲しくなった商品があった。
よし、こいつひとついってやれとばかりすぐ電話にとびついた。
「ハイ、こちら――です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。
「商品番号をおっしゃって下さい」
といわれて答える。
さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

が、その次がいけない。
「では生年月日とお年をどうぞ」
「え！」
暫く電話口で絶句したあと、
「ちょいと、あのね」
やや凄みの籠った詰問調で、
「ものを買うのにいちいち年をいわなきゃいけないの、おタクはッ」
こっちの勢いに恐れをなしたのか、
「いえ、では結構です……」
「そりゃそうでしょ」
「はぁ」
「それでいつ届くの」
「だいたい二週間くらいです」
「そ、じゃ」
ガッチャーンということにあいなったのであるが、通信販売が何故いちいち買い手の年を尋ねるのか、私は憤懣やるかたなき形相で周りにあたりちらした。
「そりゃ何か、顧客データでもとっているんじゃないの」

図 4 「くだけた印象を与える典型例」（LBf9_00067『男はオイ！女はハイ…』）

３.２　典型例の目視による考察

まず，「硬い印象を与える典型例」，「軟らかい印象を与える典型例」「くだけた印象を与える典型例」の各 2 例の本文を読み，目視から得られる特徴を列挙した。次のとおりである。

a.硬い印象を与える特徴
- •文末が「だ・である」の文である
- •断定，定義の文が多い
- •抽象物が主語の受身文が多い
- •接続詞・副詞が硬いものが多い（「すなわち，あるいは，ないし，一方で，にもかかわらず」，など）
- •親密度の低そうな語（学術・専門用語）が多い
- •難解な内容や説明である
- •疑問・回答が対応している

b.軟らかい印象を与える特徴
- •文末が「です・ます」の文である
- •直接的に語りかけてくるような文がある（文末が「ね，よ，でしょう，でしょうか」など）
- •接続詞・副詞に軟らかいものがある（「ですから」など）
- •平易な語がほとんどである
- •平易な内容や説明である

c.くだけた印象を与える特徴
- •体言止めや述語省略がある
- •一人称が主語の文が多い
- •平易な語に加え，俗語がある

•音変化（拗音化，撥音化など）の語がある
•オノマトペが多い
•感覚や感情表現が多い
•卑近な内容や説明である
•回答のない，いいっぱなしの疑問文がある

上記の特徴のうち，目視にて計測した結果を表 3 に示す。特徴として数が多かったところを太字にして表す。

表 3 「硬軟」と「丁寧さ」の特徴の計測

類型		硬い		軟らかい		くだけた	
サンプルID		LBi3_00033	LBr3_00018	LBa4_00010	LBo3_00103	LBg5_00016	LBf9_00067
書籍名		現代法社会学入門	国民の天皇	恐竜の世界をたずねて	ストレスから子どもを守る本	パパはごきげんななめ	男はオイ！女はハイ…
NDC		321	313	457	379	599	914
対象とした文数		96	140	47	29	183	45
文末表現	断定	**34**	**24**	3	2	18	7
	定義	**8**	**1**				
	です・ます			**47**	**29**	**66**	
	語りかけ			**7**	**4**		
	体言止め・述語省略					**3**	**12**
文	受身（される・された）	**7**	**9**	4		2	
	一人称主語		1			**30**	**4**
語	俗語					**39**	**6**
	音変化の文字再現・音表記					**89**	**9**
	感情（うれしい・かなしい）					**5**	**2**

３.３　典型例の計量的考察に向けて

BCCWJ に収録されるテキストの文体を計量的に考察する試みはすでにいくつか行われている（小磯ほか 2008，間淵ほか 2010，小磯ほか 2011）。我々も，今回，抽出した典型例を用いて，計量的考察に着手したところである（保田ほか 2012）。

テキスト数は不揃いであるが，「硬い印象を与える典型例」2 例，「軟らかい印象を与える典型例」4 例，「くだけた印象を与える典型例」として 18 例と，比較用に，アノテーションの作業セット 1 つ分の 463 テキストを形態素解析し，おおよその比較を試みた。その結果，語種については，予測通り，「硬い」ものは漢語率が高く，「軟らかい」「くだけた」ものは和語率が高いことが確認できた。平均との差分を図示したものを図 5 に示す。

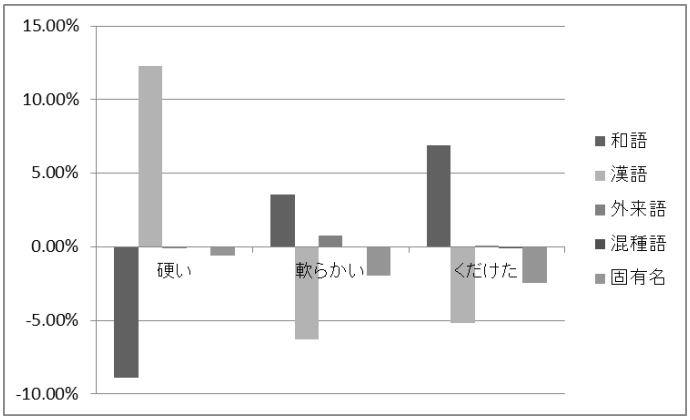


図 5 テキストタイプ別語種の比率の違い

そのほか，形態素解析結果より，品詞に関しては終助詞の比率が「軟らかい」「くだけた」もので高いこと，感動詞は「硬い」ものにはほとんど出現せず，特に「くだけた」もので

比率が高いことがわかった。

BCCWJ には形態論情報だけでなく，間淵ほか(2010)が利用した文書構造情報も付与されており，それらを用いた計量的考察を進めることは今後の課題の一つである。

４．まとめ

BCCWJ に収録する書籍コーパスの有効活用を可能とするための分類指標の人手付与作業の概要を報告した。分類指標のうち，「硬軟」と「丁寧さ」を取り上げ，その付与作業の結果から得られた対象テキストの NDC 別の特徴，典型例の特徴を述べた。

今後，抽出できた典型例の分析を進め，人手及び機械処理で付与する分類指標の正確さの向上を目指す。そして，少なくとも BCCWJ の図書館サブコーパスに収録される 10,551 サンプルの全てに分類指標を付与し，コーパスの研究や教育の利用価値を高めることを目指す。

さらに文体的な特徴を支える言語表現の分析を進め，辞書記述への応用を考えている。

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものです。また，BCCWJ の構築は，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）による補助を得たものです。

文　献

EAGLES. 1996. EAGLES Preliminary recommendation on Text Typology, *EAGLES Document EAG–TCWG–TTYP/P*, Version of Jun 1996. (http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html)

Halliday, M.A.K., A, McIntosh and P, Strevens. 1964 *The linguistic sciences and language teaching*. London: Longman.

Joos, M. 1961. *The five clocks*. New York: Harcourt Brace.

Wakako Kashino and Manabu Okumura(2010),An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese, *Proc. of PACLIC24*, pp.433-438.

柏野和佳子，奥村学(2012 予定)「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』B5-6.

小磯花絵，小木曽智信，小椋秀樹，冨士池優美，宮内佐夜香(2008)「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』pp.192-195.

小磯花絵，田中弥生，小木曽智信，近藤明日子(2011)「評定実験に基づくテキスト分類尺度の体系化の試み」『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.47-52.

間淵洋子，柏野和佳子，山口昌也，高田智和(2010)「コーパスを用いたテキスト分類指標の検討―BCCWJ の文書構造情報分析を中心に―」『言語処理学会第 16 回年次大会予稿集』PA1-11.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012)「「語り性」を有する書きことばの典型例の分析」本予稿集収録.

「語り性」を有する書きことばの典型例の分析

保田　祥[†]（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）
立花　幸子（国立国語研究所 コーパス開発センター）
丸山　岳彦（国立国語研究所 言語資源研究系）

Analysis of Written Japanese Text with Addressing Expressions

Sachi Yasuda (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)

１．はじめに

我々のプロジェクトでは，『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に，人手で文書分類の観点から情報を付与する作業を行っている（柏野・奥村 2012 予定）．

本稿では，人手によるアノテーション作業のうち，「語り性」の観点付与について取り上げる．書籍サンプルからランダムに選び出した約 500[1]のサンプルを 1 セットとし，「語り性あり」「語り性なし」のどちらかに 3 人の作業者の判断が一致したサンプル（本稿では「典型例」と呼ぶ）を分析することにより，どのような観点に基づいて「語り性」の有無が判断されているかを分析する．複数作業者の判断が一致したサンプルをもとに分析を行うことで，「語り性」の観点付与に有効な指標が抽出可能と考えられる．また，あわせて，「語り性」を持つと判断されたテキストを，別のセットの作業で「話しことば的」と判断されたテキストと比較し，両者の異同を確かめる．

２．「語り性」について

書籍テキストの中には，著者が読者に対して直接語りかけていると解釈できる文体がある．

以下の例は，読者に語りかけている文体のサンプルである．

例１）

> ときには一流ホテルのロビーでお茶を飲んで、ゴージャスな盛り花を見てきましょう。物ごとの上達は真似ることから始まるのです。真似るは学ぶの語源だそうです。真似を自分のものにしたいから、夢中になれるのではないでしょうか。
>
> 結婚式などフォーマルなお祝い事がつづくことがあります。そのたびにドレスを新調するのはたまったものではありません。かといって、いつもおなじものでは気がひけます。ある時期から黒い慶弔両用のドレスに、生花のコサージュをつけて出席することに決めました。
>
> コサージュは、お花屋さんでつくってくれます。あなたがすでにお花に詳しいなら、一〜二回の講習会で自分でつくれるようになれます。
>
> （LBo1_00017『ひとりって楽しい』）

[†] yasuda_s@ninjal.ac.jp

[1] セットごとに 456〜485 の幅があり，本稿で扱ったセットは 485 サンプルである．

例２）

> ですから、私たち親としてできることは、子どもの自律神経のバランスを整えるために、食事と睡眠を規則正しくとれるように注意することです。そのためには親子で食事を一緒にとるためにやりくりすることが大切です。親の都合を優先しておいて後から、基本的生活習慣への不安をもち、子どもに“早寝早起き”“三度の食事”を励行させても、それは一朝一夕にはできませんよね。　（LBo3_00103『ストレスから子どもを守る本』）

「あなた」や「みなさん」などの呼びかけ表現や，「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など，「直接的な語り」表現と呼べるような表現が含まれるテキストを，本稿では「語り性を有するテキスト」と呼ぶ（柏野 2010）．

一方，書籍テキストの中には，「話しことば的」として解釈できる文体もある．会話文のみで地の文がない場合，戯曲調で地の文が卜書きの場合，講演会の書き起こし，一人称小説のようなサンプルである．本稿では，これらを「話しことば的」の典型例と呼ぶ．

以下，本稿は，「語り性」を有するテキストの文体的特徴を分析することにより，作業者が「語り性」を有すると判断する際の言語的な指標を抽出することを目的とする．

３．「語り性」に関する先行研究

文章を分類する試みの一つとして，石田（2003）の，量・構文・位置・表現・内容に関する様々な指標の提示がある．小磯ほか（2008）は，テキストに含まれる品詞率，語種率，異なり語率，文の長さなどの 12 の指標を選定してテキストの分類を行っている．また，前坊（2011）のように副詞と文末表現をとりあげて文書分類を行う研究も見られる．

本稿では文章の持つ「語り性」の観点を分析するため，「呼びかけ表現」や「文末表現」を抽出する．そこで，形態素解析を行い，品詞，活用形，語彙素の出現率を調べた．出現率の高い要素を抽出することにより，作業者が「語り性」の観点で分類を行う際に用いる指標を分析することができると考える．

また，小磯ほか（2011）は，調査者から得た評定語を指標として分析を行っている．そのとき，「書きことば的―話しことば的」という尺度に，「読み手に語りかける―語りかけの少ない」という尺度や，改まりの程度などの複数の観点が関与する可能性を示している．

そこで，本稿の「語り性」の観点分析にあたっては，「語り性」があるという分類と「話しことば的」であるという分類に差があるのかという点についても確かめることとした．「語り性」があると作業者が判定を行ったサンプルのセットと，「話しことば的」であると作業者が判定を行ったサンプルのセットを対照し，「語り性」と「話しことば的」の差を調べる．

４．調査データ

本稿で扱うのは，BCCWJ の図書館サブコーパスに含まれる書籍からランダムに選んだ 485 サンプルのセットである．このうち，398 サンプルが作業者 3 人全員に分類対象[2]として選ばれており，分析にあたっては，全作業者が分類を行ったこれらのサンプルを調査対象とする．

また，「語り性あり」「語り性なし」の判断が作業者全員で一致したサンプルは 80%であり，内訳は，「語り性あり」が 6.28%（25 サンプル），「語り性なし」が 73.87%（294 サンプル）である．作業者判断が一致していることから，これらのサンプルはそれぞれ「語り性」

[2] 対談，Q&A 形式，図解，用語解説など形式的に特徴のあるサンプルは，今回は分類対象外（非対象）とされた．作業者は，分類対象としたサンプルのみ観点付与を行っている．

の有無に関する典型例であると考える．「話しことば的」の典型例としては，「話しことば的」か「書きことば的」かの観点でアノテーションを行った 1,890 サンプル中，「話しことば的」の分類で作業者判断が全員一致した 12 サンプルを得ている．

典型例として選び出されたサンプルのセットについて，MeCab 0.98＋UniDic 1.3.12 を用いた形態素解析を行った．以下の分析結果に示す指標は，解析結果に基づく．但し，人手の観点付与作業では，「語り性」の判断に地の文のみが作業対象とされていることから，形態素解析にあたっては，会話文内の話しことばを解析対象としていない[3]．

本稿で扱うデータのセットは表１の通りである．

表１．分析対象データ

	語り性あり	語り性なし	非一致	話しことば的
サンプル数	25	294	79	12
語数	78,013	789,530	254,361	49,225

また，作業者の観点付与結果[4]の NDC 別分類を図１に表す．図１から，「語り性」があると判断されたサンプルは，NDC1 番台（哲学）と NDC4 番台（自然科学）に多い傾向が見られることがわかる．なお，図１の（）内数値は，分析対象としたセット 485 サンプルの NDC 別内訳を表す．

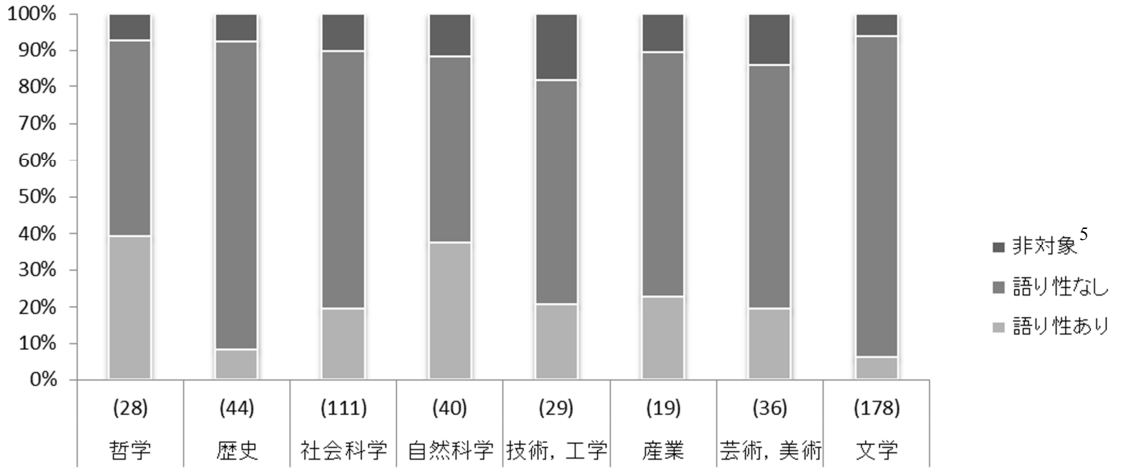


図１．NDC 別の「語り性」観点付与における作業者判断

５．語り性の観点付与結果に現れた指標

「語り性あり」「語り性なし」の典型例のセットを比較対照し，品詞と活用形，語彙素のそれぞれについて，観点付与の判断基準と考えられる指標の抽出を試みる．同時に，「語り性あり」の典型例については，「話しことば的」の典型例との対照も行い，「話しことば的」と「語り性」の差についても明らかにする．

---

[3] 「」内を会話文と見なした．固有表現や引用，強調などの部分に「」が用いられる場合も見られるが，「」は NDC9 番台に集中しており，概ね小説における登場人物の会話に用いられていることが確認された．

[4] 作業者 3 人が行った観点付与作業結果を足し合わせた．

[5] 注 2 を参照．作業者が作業対象外（非対象）と分類した割合を示す．

典型例のセットの比較にあたっては，カイ二乗検定を用いて要素ごとに有意差（有意水準 0.1%以下）の確認を行う．

５.１　語り性の有無と品詞

「語り性あり」「語り性なし」の典型例においては，出現頻度が上位である品詞では，表２のように差があまり見られない．しかし，全品詞を見ると，表３の通り，「語り性あり」の典型例は，「語り性なし」よりも終助詞，代名詞の出現率が高く，「語り性なし」では，「語り性あり」よりも固有名詞の出現率が高いという結果が現れた．

「語り性あり」との観点付与がなされた典型例は，終助詞と代名詞が多い．「語り性あり」の観点付与がなされた典型例は，相手に対する表現として，呼びかけや終助詞の付与が現れるためと考えられる．

また，「語り性なし」との観点付与がなされた典型例では，固有名詞の出現が多いといえる．NDC との関連（図１参照）を見ると，「語り性なし」の観点付与がなされた典型例は，NDC9 番台（文学：小説が多い）が 88%であり，登場人物の人名によって，固有名詞の人名が増えていることが考えられる．NDC2 番台（歴史）も「語り性なし」が 84%であり，固有名詞の地名等が増えることが考えられる．固有名詞の出現率は，ジャンルとの関連性が予測される．

なお，「語り性あり」と「話しことば的」の典型例を対照すると，終助詞，感動詞，固有名詞，代名詞のそれぞれが「話しことば的」に多いという有意差が現れていた．前述した固有名詞のほか，感動詞は，「語り性あり」の典型例よりも「語り性なし」の典型例に出現率が高い．「話しことば的」の判断には感動詞が関連していることが考えられるが，「語り性あり」の判断とは関わりが低い可能性がある．

表２．「語り性」の有無と品詞（頻度上位 10 位）

「語り性なし」典型例			「語り性なし」典型例		
助詞-格助詞	127,879	16.20%	助詞-格助詞	12,577	16.12%
名詞-普通名詞-一般	113,407	14.36%	名詞-普通名詞-一般	10,516	13.48%
助動詞	64,107	8.12%	助動詞	7,225	9.26%
動詞-一般	49,345	6.25%	動詞-非自立可能	5,526	7.08%
動詞-非自立可能	47,863	6.06%	動詞-一般	4,768	6.11%
補助記号-読点	41,238	5.22%	名詞-普通名詞-サ変可能	4,459	5.72%
名詞-普通名詞-サ変可能	35,938	4.55%	補助記号-読点	4,311	5.53%
助詞-係助詞	32,007	4.05%	助詞-接続助詞	3,614	4.63%
助詞-接続助詞	31,725	4.02%	助詞-係助詞	3,436	4.40%
補助記号-句点	30,750	3.89%	補助記号-句点	2,651	3.40%

表３．「語り性」の有無，「話しことば的」の品詞（抜粋）

品詞	「語り性なし」典型例		「語り性あり」典型例		「話しことば的」典型例	
終助詞	1,667	0.21%	384	0.49%	1,145	2.28%
感動詞	506	0.06%	32	0.04%	404	0.80%
固有名詞	21,191	2.68%	436	0.56%	357	0.71%
代名詞	9,586	1.21%	1,133	1.45%	1,014	2.02%

５.２　語り性の有無と活用形

表４は，出現率に有意差の見られた活用形である．「語り性あり」「語り性なし」の典型例で，意志推量形（「～でしょう」「～だろう」など），命令形（「～ください」「～おけ」など）の出現率に差があった．「語り性あり」のテキストでは，読み手に対する表現が用いられているため，意志推量形と命令形の出現率が高いことが考えられる．但し，これらは「語り性あり」と「話しことば的」では差がない．

また，「語り性あり」の典型例には，「語り性なし」と比較すると，融合（例：「～じゃない」「～なきゃ」など）の出現率も高い．しかし，「語り性あり」と「話しことば的」を対照すると，融合は「話しことば的」の典型例で出現率が高いという結果が現れている．「語り性あり」は「話しことば的」とは同一の判断基準ではないことが示唆されよう．

表４．「語り性」の有無，「話しことば的」の活用形（抜粋）

活用形	「語り性なし」典型例		「語り性あり」典型例		「話しことば的」典型例	
意志推量形	1,895	0.24%	360	0.46%	239	0.48%
命令形	442	0.06%	79	0.10%	61	0.12%
撥音便	1,793	0.23%	325	0.42%	199	0.40%
融合	53	0.01%	29	0.04%	107	0.21%

５.３　語り性の有無と語彙素

図２の語種の割合を見ると，「語り性あり」の典型例と「語り性なし」の典型例には有意差がみられることがわかる．特に，「語り性なし」では前述の固有名詞の頻度が高いことから，固有語の割合が高く，また，会話文の補助記号（「」）をはじめとする記号の割合が高い．「語り性あり」の典型例では，「語り性なし」との違いとして，和語の多さがあげられる．

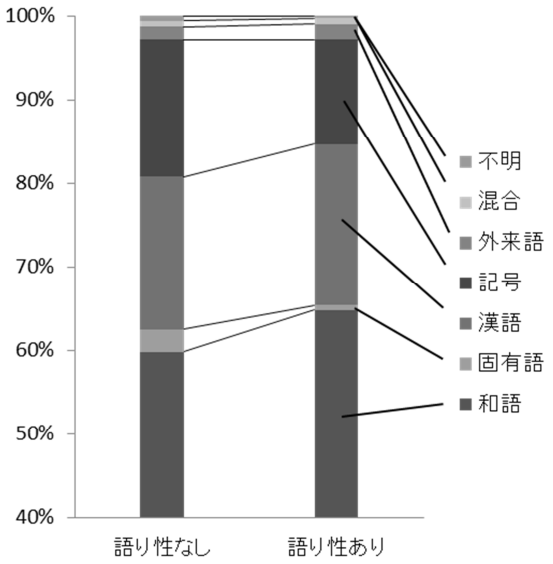


図２．「語り性」の有無と語種

「語り性なし」の典型例と対照すると，「語り性あり」の典型例にあきらかな語彙素は，以下であった．表５に，出現率と「語り性なし」「話しことば的」の典型例との対照を併せ

て示す.

- 「た」・「です」・「ます」（助動詞）が多い
- 「か」・「ね」・「よ」（終助詞）が多い
- 「私」（一人称代名詞）・「貴方」（二人称代名詞）・「自分」（名詞）が多い
- 「の」（準体助詞）・「事」（名詞）が多い
- 「「」・「」」（記号）が少ない

表５.「語り性」の有無,「話しことば的」の語彙素（抜粋）

語彙素	「語り性なし」典型例		「語り性あり」典型例		「話しことば的」典型例	
た（助動詞）	26,939	3.41%	1,285	1.65%	1,018	2.02%
です（助動詞）	388	0.05%	926	1.19%	591	1.17%
ます（助動詞）	698	0.09%	1,051	1.35%	406	0.81%
事（名詞）	4,368	0.55%	820	1.05%	316	0.63%
「（補助記号）	10,285	1.30%	406	0.52%	578	1.15%
の（準体助詞）	5,849	0.74%	825	1.06%	882	1.75%
自分（名詞）	819	0.10%	193	0.25%	125	0.25%
私（代名詞）	946	0.12%	146	0.19%	177	0.35%
貴方（代名詞）	55	0.01%	34	0.04%	14	0.03%
君（代名詞）	7	0.00%	5	0.01%	16	0.03%
か（終助詞）	1,192	0.15%	217	0.28%	245	0.49%
よ（終助詞）	107	0.01%	35	0.04%	280	0.56%
ね（終助詞）	95	0.01%	67	0.09%	286	0.57%

「語り性あり」の典型例には，助動詞の「です・ます」が高い出現率で見られている．しかし,「語り性あり」と「話しことば的」を対照すると,「です」には有意差がないが,「ます」は「語り性あり」の典型例に出現率が高いという違いがある．助動詞の「た」と各種終助詞は,「話しことば的」の典型例に出現率が高い.「語り性あり」の判断に「ます」が関わっており，また,「た」や終助詞は「話しことば的」の判断指標とされている可能性がある．

「語り性あり」における終助詞の多さは，５．１で見た通りであるが，語彙素レベルで見ると，終助詞「か」が「語り性なし」の典型例でも，0.15%の出現率と上位頻度語（55 位／異なり語彙素数 29,790）ながら,「語り性あり」の典型例では，0.28%の出現率となっており,「語り性あり」により多く見られる語彙素であるとわかる．

代名詞についても，５．１で見た通りといえるが,「語り性あり」の典型例と「話しことば的」の典型例では，一人称・二人称代名詞のそれぞれで,「私」のみ「話しことば的」に多く現れることを除き，ほぼ出現率に差がないという結果が現れている.「話しことば的」の判断指標として，一人称代名詞，二人称代名詞が用いられている可能性が考えられる．

また,「～なのだ」のように用いられる準体助詞の「の」や,「～という事」のように用いられる名詞の「事」が,「語り性あり」の典型例に出現率が高いこともわかった．上位頻度語という点では,「語り性なし」にも（「の」20 位：0.74%,「事」22 位：0.55%）多く現れる語彙素であるが,「語り性なし」の出現率を上回っている．とくに「語り性あり」では,「事」の出現率が高い.

なお,「語り性なし」の典型例には，NDC9 番台が多い（小説の会話文が多い）ため，補助記号も多く現れるが，前述の通り，語り性の判断は地の文で行われていることから，直接的な判断指標であるとは言い難い.

６．まとめ

既にアノテーション作業が完了しているサンプルを用い,「語り性」の観点付与に関して,複数作業者の判断が一致した典型例の分析を行った．品詞，活用形，語彙素の出現率を調べ，出現率の高い要素を抽出することで,「語り性」の有無について観点付与を行う作業者が，分類に用いている可能性の高い指標が整理された.

但し,「語り性」観点付与における作業者間の判断基準の差について,考慮しておきたい.ここまで，全作業者の判断が一致した典型例の分析を行って指標を得たが，全作業者の判断が一致しなかったサンプルもある．作業者の判断に個人差があり，作業者によっては判断基準とならない指標がある可能性が考えられる．よって，語り性の有無の観点で 3 人の作業者の判断が一致しなかったサンプル（79 サンプル）との対照を行い，作業者判断に揺れのある例のセットで，抽出した指標の出現率を確かめた.「語り性あり」の典型例に出現率が高くなければ，判断基準としての効果が低い可能性もあるといえよう．表６に対照結果を示す.

表６.「語り性あり」と作業者判断非一致サンプルとの対照

指標	語り性あり		非一致		高頻度
た(助動詞)	1,285	1.65%	6,889	2.71%	非一致
です(助動詞)	926	1.19%	1,416	0.56%	語り性あり
ます(助動詞)	1,051	1.35%	1,909	0.75%	語り性あり
事(名詞)	820	1.05%	2,150	0.85%	語り性あり
の(準体助詞)	825	1.06%	2,690	1.06%	有意差なし
自分(名詞)	193	0.25%	400	0.16%	語り性あり
貴方(代名詞)	34	0.04%	70	0.03%	語り性あり
君(代名詞)	5	0.01%	24	0.01%	有意差なし
か(終助詞)	217	0.28%	664	0.26%	有意差なし
よ(終助詞)	35	0.04%	170	0.07%	非一致
ね(終助詞)	67	0.09%	118	0.05%	語り性あり
意志推量形	360	0.46%	899	0.35%	語り性あり
命令形	79	0.10%	202	0.08%	有意差なし

結果として,「語り性」観点付与のための指標は，以下が得られた.

「語り性あり」
　活用形　　：意志推量形（「～でしょう」「～だろう」など）が多い
　語彙素　　：「です（助動詞)」「ます（助動詞)」「事（名詞)」が多い
　　「あなた（代名詞)」「自分（名詞)」「ね（終助詞)」が多い

「語り性なし」
　品詞・語種　：固有語が多い
　語彙素　　：「た（助動詞)」が多い

また，「語り性」の分類にあたり，作業者が「話しことば的」の観点とは異なる指標で判断を行っている可能性も示唆された．感動詞，融合（「〜じゃない」「〜なきゃ」など），終助詞「よ」のように，「話しことば的」で出現率が高くとも，「語り性あり」の典型例と比較すると，「語り性あり」では出現率が低いという要素があるためである．作業者の「語り性」の分類判断は，「話しことば的」との差異を含め，複雑な条件によって行われているものと考えられる．副詞や接続詞などの品詞毎に詳細な分析を行うことで，出現率が低くとも，判断に用いられる指標が得られる可能性もある．さらに，その他の観点の分析結果とあわせるなどの分析も必要であろう．

今後は，これらの指標とともに典型例を提示したマニュアルを作成し，マニュアルに沿ったアノテーション作業を進めることを予定している．観点付与にあたり，作業者が指標を参照することで，作業の効率化が見込めるとともに，有用なデータの作成が期待される．

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものです．また，BCCWJ の構築は，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）による補助を得たものです．

文　献

石田栄美(2003)「テキストの自動分類に関わる諸要素」『日本図書館情報学会誌』49(2), pp65-78.

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 予定)「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』B5-6.

小磯花絵，小木曽智信，小椋秀樹，冨士池優美，宮内佐夜香(2008)「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』pp.192-195.

小磯花絵，田中弥生，小木曽智信，近藤明日子(2011)「評定実験に基づくテキスト分類尺度の体系化の試み」『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.47-52.

前坊香菜子(2011)「雑誌コラムに現れる語彙とモダリティ―副詞と文末表現を中心に―」『信学技報』，pp.55-60，電子情報通信学会

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/

特定領域研究「日本語コーパス」　http://www.tokuteicorpus.jp/

反復語の使用実態から見る話し言葉と書き言葉の連続性 −コーパスを用いた定量的分析を通して−

鯨井 綾希（東北大学大学院文学研究科）†

Continuity between Spoken and Written Language as Seen from the Use of Repeated Words; A Quantitative Corpus Analysis

Ayaki Kujirai (Graduate School of Arts and Letters, Tohoku University)

１．本発表の目的

本発表では、『日本語話し言葉コーパス』（以下CSJ）と『現代日本語書き言葉均衡コーパス』（以下BCCWJ）を利用して、文章や談話のまとまりや意味的連続性を形成するのに重要な役割を担う同一名詞の反復（以下これを反復語と記述）の使用実態を定量的な側面から明らかにし、その上で話し言葉や書き言葉といった言語表現上のバリエーションを考察に加え、反復語の量的側面から見たときのそれらの関係性を見出すことを目的とする。

２．調査上の対象・資料・方法

２．１ 調査対象

本発表で対象とするのは、以下の波線部で示したような、文章・談話内における内容を表す名詞の反復使用（以下これを反復語と表記）である。

（1）産業技術が極めて幼稚なうちは、思い込みで適当なことをやっていても一見通用するように見える時代がある。しかし、どのような分野でも産業技術は次第に高度化する。その結果、学問に基づいた本物の技術しか通用しないということがいろいろな産業分野でいま始まっている。
（大見忠弘2004『復活！日本の半導体産業-未来を拓く志-実力を磨いて世の中の役に立とう！-』財界研究所、『BCCWJ』サンプルID:PB45_00024）

（2）R : (F ええ)で猿だったら(F えーとー)聞こえんのか聞こえないのかっていうので(F えーっとですねー)(F えー)何だろう例えばピアノの音とかだと
L : (F はい)
R : (F えー)色んな音の成分がたくさん入ってるんですね
L : (F はい)
R : んでその(D い)一番低い音の成分から(F そのー)高い音の成分まで色々入ってるんだけど （『CSJ』ID:D04M0056を見やすく整形）

† donguri-no-stability@hotmail.co.jp

（1）も（2）も、「産業」「技術」「音」という名詞が反復的に使用されることでそれを中心とした内容展開が構築されている。このような反復語の使用は、文章・談話中において文の意味的連鎖やコミュニケーション上の機能に関わるものとして、従来から話し言葉・書き言葉の双方において定性的な研究が行われてきた（中田 1991、田中 1997、塩澤 2005、馬場 2006 など）。ただ、そもそも反復語が文章・談話内でどの程度現れる存在なのかという定量的側面に注目した研究は管見の限り見られない。よって本発表ではコーパスを利用して話し言葉・書き言葉双方に見られる反復語の定量的分析を行う。

２．２ 調査資料

扱う資料は、CSJ および BCCWJ である。CSJ は話し言葉のデータであり、本発表では対話データを収録したデータと、学会講演と模擬講演という二つの独話データとを取り上げる。対話データは、厳密には講演のインタビュー・課題指向対話・自由対話が含まれるが、それぞれのファイル数の少なさから、一括して扱う。書き言葉は BCCWJ の中の白書・新聞・書籍・雑誌の各サブコーパスを用いる。なお、BCCWJ のデータは、詳細な情報が付与されているコアデータのみ取り上げる[1]。調査資料の大きさは表 1 に示した[2]。

表 1：分析資料の概略

	白書		新聞		書籍		雑誌	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
最小値	216	123	127	97	126	64	134	102
第1四分位値	701.5	238	230	144	254	144.5	273	177.8
中央値	981.5	303	275	169	438	215	521.5	291
平均値	1182.1	339.1	305.1	184.7	548.5	262	598.8	307.3
第3四分位値	1594.2	441.2	342.2	206.2	722	367.5	834.5	381.5
最大値	2925	713	1203	571	1593	719	1954	876
ファイル数	62		340		83		86	

	学会講演		模擬講演		対話	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
最小値	119	82	193	97	176	82
第1四分位値	590	175	278.5	139.5	291.5	119.2
中央値	717	207	337	164	342	146.5
平均値	778	226.1	345	168.1	365.1	151.2
第3四分位値	882	256	403	195	404	171.8
最大値	3509	917	659	280	922	268
ファイル数	987		107		58	

[1] コアデータにはその他にも Web 上のブログや知恵袋という質問サイトのデータも存在するが、データ内に存在する各ファイルが対象にできる大きさに至っていないため、本発表では扱わない。

[2] ここで単位となる「語」の認定基準は後述。また、同じく後述するが、本発表では 50 語ごとの反復語の使用率とそのばらつきを分析に取り入れるため、延べ語数 100 語未満のファイルは分析対象から外した。

CSJ と BCCWJ は、意思伝達の媒体が音声言語か書記言語かという大きな違いがある。これは話し言葉か書き言葉かという区別を行う上で最も大きな要素である。基本的には、その違いによって「話し言葉的」か「書き言葉的」かの性質が決定されると考えられる。また、複数の話し手が話す・聞くという関係を相互に行う点で、CSJ における対話は書き言葉と決定的に異なり、典型的な話し言葉として認められる。一方、白書のような公的機関による書き言葉は、話し言葉的要素をあまり導入しない点で書き言葉の典型と認めうる。

CSJ の学会講演や模擬講演は一人の話し手によってなされ、筋書きが大方決まっている点で書き言葉の音声化とも呼べる。したがって、この二つは対話に比べて書き言葉に近い存在である。また、音声言語と書記言語のそれぞれでも、改まりの程度に違いがあり、白書と学会講演は共に公的な場でのものであるという点で改まりの程度が高い。書籍と模擬講演は、必ずしも改まりの場として捉えなくて良いという点で白書や学会講演よりも改まりの程度が低い。雑誌はそれらに比べるとさらに改まりの程度の低いものであると言える。新聞は改まりの程度としては白書同様の高さであると考えられるが、多くの情報を限られた字数内に収めなければならないという字数制限による表現上の制約が強く、その点で他のサブコーパスに対して特異な性質を持つ。

以上から、本発表では基本的に音声言語・書記言語、独話型・対話型、改まり度の高低、字数制限という四つの基準を設けることができる。これを模式的に表したものが図 1 である[3]。図 1 より、話し言葉と書き言葉は種々の性質の中で連続性を持っていることが分かる。

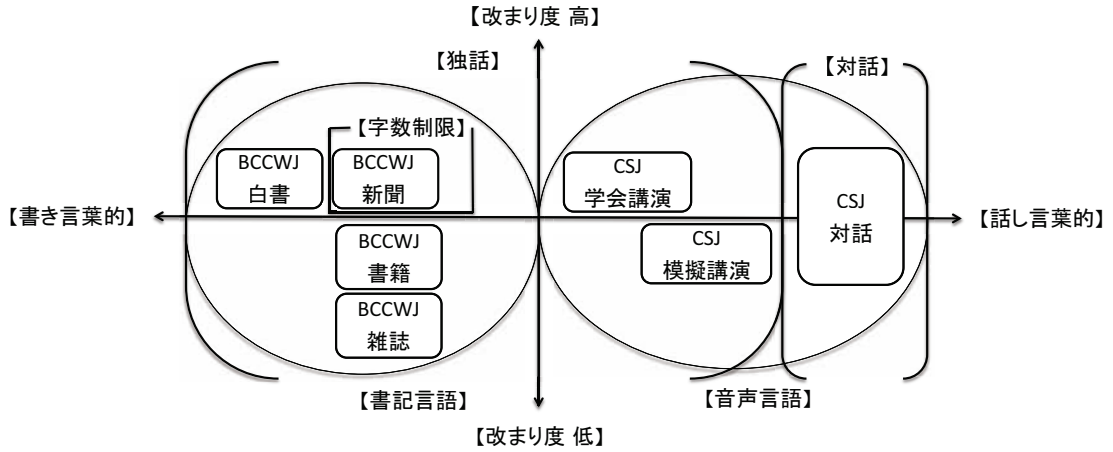


図 1：資料間の関係性と連続性

２．３ 調査方法

定量的分析に際しては、名詞 50 語あたりの反復語の平均使用率・使用率のばらつき、反復語 1 語あたりの平均使用頻度・平均反復間隔という四つの側面に注目した。本発表では各資料における計算結果の記述を行った上で、各資料間の計量結果の横断的に分析するこ

[3] 図 1 において学会講演が模擬講演よりも書き言葉的であるとしているが、これは典型的な書き言葉である白書と、改まりの度合いにおいて類似性が高いのが学会講演であると判断したことによる。

とで、反復語が文章・談話中においてどのように現れ、どのように使用されているのかを多角的に明らかにする。各資料の分析には、「語」相当の集計単位として「短単位」による形態論情報を設定した。CSJ と BCCWJ コアデータには形態論情報が付与されているため、それをそのまま利用した[4]。

３．調査結果

３．１　平均使用率

反復語が文章・談話中でどの程度現れるのかを明らかにするために、始めに名詞 50 語あたりの反復語の使用率の平均値を計算した。反復語は、延べ語数に対する 2 回目以降に使用された名詞の比率を求めれば良い。

本発表で 50 語ごとに計算したのは、文章の長さに影響を受けない形で対象ファイルにおける反復語の使用率を算出するためである[5]。この方法に基づいた対象資料ごとの各ファイルにおける反復語の平均使用率は表 2 に、その箱ひげ図を図 2 に示した[6]。

表 2：平均使用率(%)

	白書	新聞	書籍	雑誌	学会講演	模擬講演	対話
最小値	21.81	6	12.67	5	19.62	15.67	21.67
第1四分位値	26.93	16.4	21.02	16.34	31.83	26.33	31.8
中央値	30.65	19.38	25.6	19.59	35.87	29	34.2
平均値	31.1	19.74	25.56	20.07	35.7	29.44	35.19
第3四分位値	34.27	22.89	30	23.2	39.57	32.73	38.96
最大値	45.47	36	48.8	40.55	53.09	43	55.2

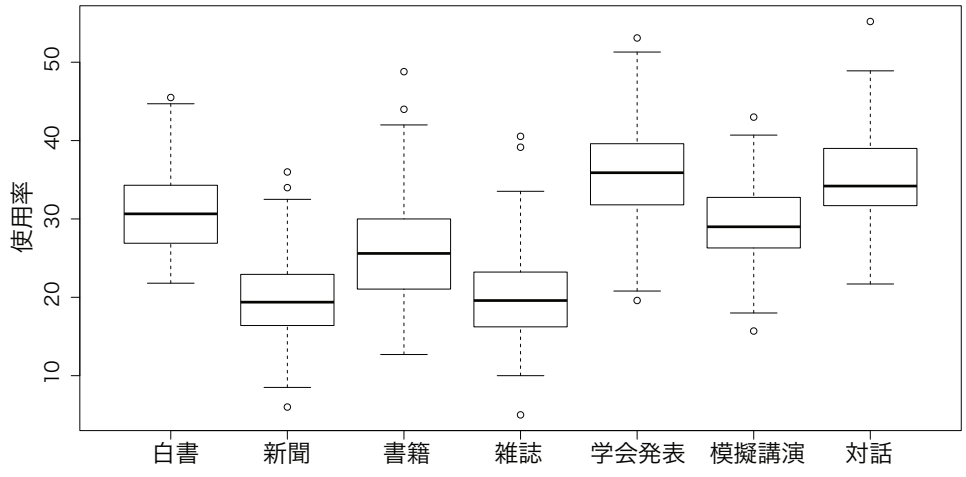


図 2：平均使用率(%)

[4] 短単位については小椋（2006）、小椋（2008）、小椋ほか（2011）を参照。なお、それを用いた集計や計算、作図はスクリプト言語 Perl と統計解析環境 R を利用している。

[5] 50 語ごとに集計した際の余りは切り捨てた。

[6] 図 2 の箱ひげ図は、箱内の線が中央値であり、下辺と上辺がそれぞれ第 1 四分位値と第 3 四分位値を指す。上下の丸は外れ値、上下の点線の先は外れ値を除いた場合の最大値と最小値である。

50 語あたりの反復語の使用率の平均は、書き言葉に比べて話し言葉の方が総じて高い値を示す。また、書き言葉内では、白書の方が書籍よりも使用率が高く、書籍は雑誌よりも使用率が高い。また、話し言葉のうち、学会講演が模擬講演や対話に比べて高い値を取る。この点で、反復語は、話し言葉的であり、かつ改まり度が高い場合において使用率が高まると言える。ただし、新聞の使用率が最も低くなっていることには注意が必要である。これは字数制限により同一語の反復を極力避けることと関係していると考えられる。

３．２　50 語ごとの使用率のばらつき

前節では各テキストで用いられる名詞 50 語ごとの反復語の使用率の平均を計ったが、50 語単位でくり返し計測しているため、計算ごとの結果にはばらつきが生じる。そこで、各データ内に含まれるファイルごとで、50 語ごとの使用率にどの程度のばらつきが見られるのか調査した結果を表 3 と図 3 に示す。なお本発表のばらつきは変動係数によって表した。

表 3：平均使用率のばらつき方

	白書	新聞	書籍	雑誌	学会講演	模擬講演	対話
最小値	0.144	0.055	0	0.065	0.046	0.056	0.032
第1四分位値	0.244	0.273	0.209	0.285	0.197	0.167	0.169
中央値	0.281	0.362	0.295	0.346	0.231	0.223	0.207
平均値	0.293	0.382	0.29	0.36	0.234	0.228	0.204
第3四分位値	0.338	0.467	0.357	0.417	0.269	0.282	0.243
最大値	0.481	1.324	0.609	1.006	0.434	0.487	0.362

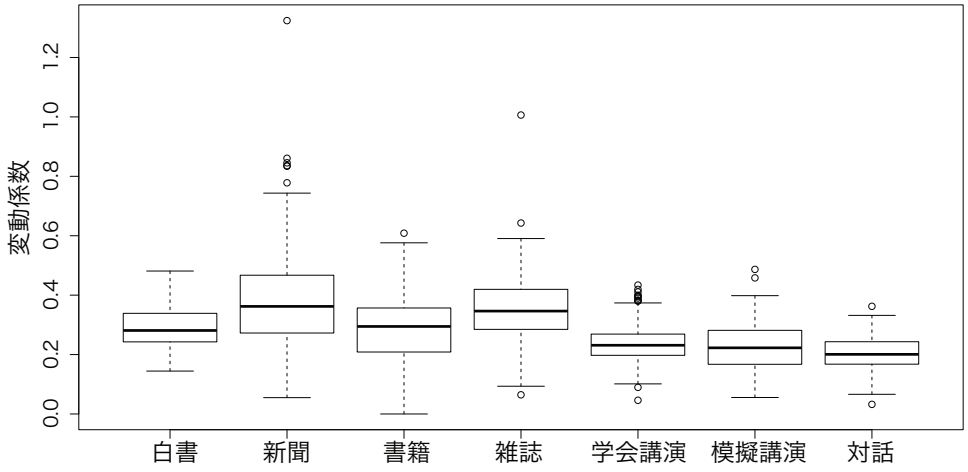


図 3：平均使用率のばらつき方

書き言葉のうち、白書を除く BCCWJ の各サブコーパスは最大値と最小値の差が大きく、ファイルごとの反復語の使用率のばらつきそのものが大きい。つまり、書き言葉には反復語の使用・不使用に関する一貫性が必ずしも見られず、量的に多様な様相を見せると考えられる。

また、変動係数の値の幅は、話し言葉の方が書き言葉よりも総じて小さい。したがって、

話し言葉の方が場面に拘らず反復語を用い、書き言葉の方が反復語を使ったり使わなかったりする傾向にあることが分かる。

３．３　１語あたりの平均反復頻度

反復語は 2 回目以降使用された語が全て含まれるため、反復語に含まれる見出し語自体は多様である。したがって、反復語の使用実態を記述するために、その一つひとつの語がどの程度反復されているのかを把握することも有意義であると考えられる。反復語に含まれる見出し語 1 語あたりの頻度を平均した結果が以下の表 4 と図 4 である。

表 4：反復語 1 語あたりの平均反復頻度

	白書	新聞	書籍	雑誌	学会講演	模擬講演	対話
最小値	3.77	2.35	2.9	2.36	2.83	2.99	3.24
第1四分位値	4.97	2.92	3.47	3.4	5.27	3.68	3.92
中央値	5.73	3.17	4.04	3.93	5.84	3.98	4.25
平均値	6.21	3.27	4.21	4.1	6	4.06	4.42
第3四分位値	7.1	3.51	4.85	4.46	6.58	4.35	4.69
最大値	12.68	9.04	6.47	9.82	11.9	5.37	7.18

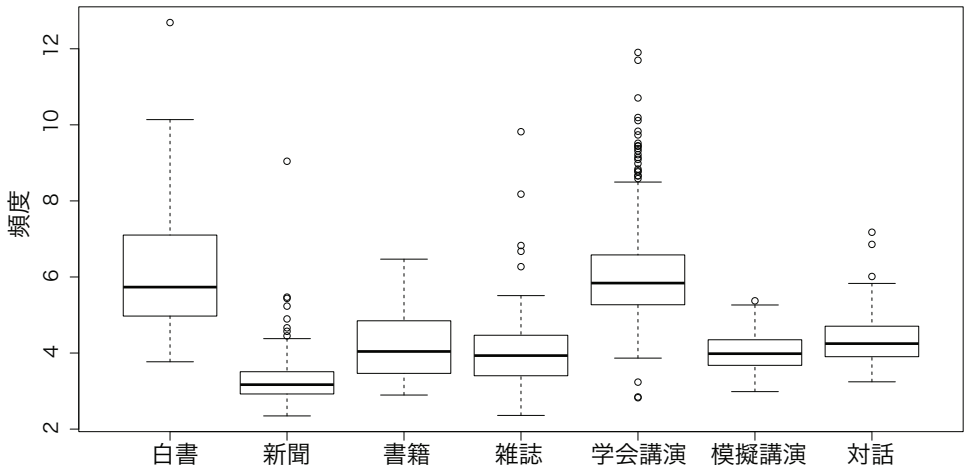


図 4：反復語 1 語あたりの平均反復頻度

学会講演に外れ値が多くあるが、全体としては学会講演と白書の二つにおいて反復語の頻度が高い。つまり、反復語の一つひとつの頻度は、書き言葉・話し言葉の別なく、改まりの程度が高くなると増加してくると考えられる。また、新聞が値の幅、値そのものの両方においてそれ以外のデータよりも小さくなっている。ここでも字数制限による反復語の使用の抑制が伺える。

３．４　１語あたりの平均反復間隔

反復語は文章・談話中で複数回用いられる語の集合である。そのため、反復語となった

語は、それぞれがある間隔を置いて改めて使用されており、前出の語と再使用された語との間には間隔が生まれる。反復語に位置づけられる語のそれぞれにおいて、どの程度の間隔で反復が行われているのかを調査し、その平均間隔を計算した結果が表 5 と図 5 ならびに表 6 と図 6 である。表 5・図 5 では名詞をもとにして距離を計測し、表 6・図 6 では品詞に拘らず全ての語によって距離を計測した。始めに、名詞の語数を利用した結果である表 5 と図 5 を示す。

表 5：反復語 1 語あたりの平均反復間隔(名詞での間隔)

	白書	新聞	書籍	雑誌	学会講演	模擬講演	対話
最小値	13.33	11.77	8.18	12.38	8.13	8.02	5.09
第1四分位値	28.09	22.12	18.77	25.27	17.87	14.5	12.84
中央値	37.88	27.31	29.96	38.03	21.63	18.32	15.73
平均値	38.39	28.44	32.31	40.34	23.43	18.96	15.53
第3四分位値	48.77	32.9	42.81	49.24	27.42	22.67	17.91
最大値	82.9	71.56	76.44	94.52	67.78	34.1	27.74

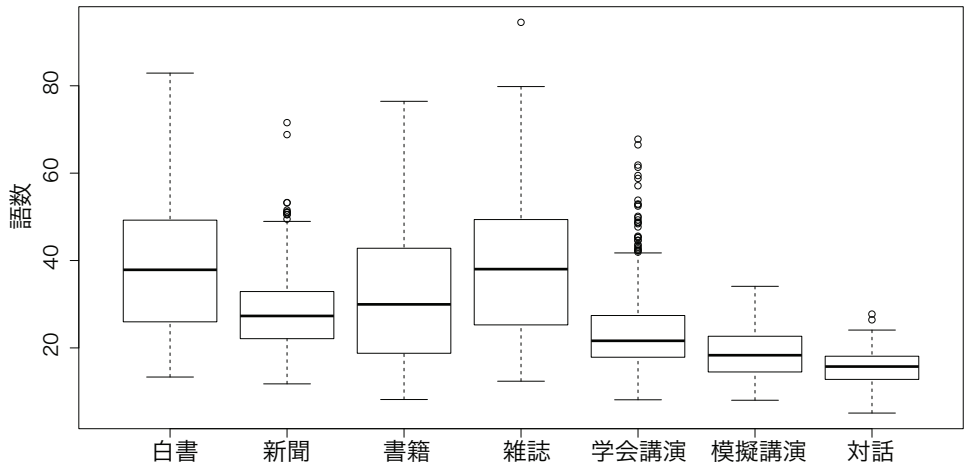


図 5：反復語 1 語あたりの平均反復間隔(名詞での間隔)

名詞間の測定では、BCCWJ と CSJ では、CSJ の方が明らかに間隔が短い。よって、書き言葉よりも話し言葉の方が短い間隔で語が反復されていると言える。また、話し言葉の中でも、話し手と聞き手の交替によって特徴付けられる対話は最も反復の間隔が短くなる。

一方、書き言葉の中で見た場合、特に最大値において新聞が他の書き言葉と明らかに異なった値を示している。他のデータとの比較から、字数制限による影響と考えられるが、その具体的な関係性については、内実をより詳細に観察する必要がある。

次に、全ての語を利用して反復の間隔を測定したものが以下に示した表 6・図 6 である。

表 6：反復語 1 語あたりの平均反復間隔(全語での間隔)

	白書	新聞	書籍	雑誌	学会講演	模擬講演	対話
最小値	37.22	30.52	34.47	48.47	26.21	46.18	29.87
第1四分位値	62.34	59.87	79.82	89.9	63.43	78.86	69.33
中央値	90.45	77.04	113.18	140.19	83.78	96.07	95.09
平均値	96.79	82.04	140.51	152.94	91.04	99.43	95.07
第3四分位値	123.89	100.19	203.26	191.78	110.66	116.51	115.69
最大値	216.15	187.66	368.78	379.94	317.15	169.95	163.32

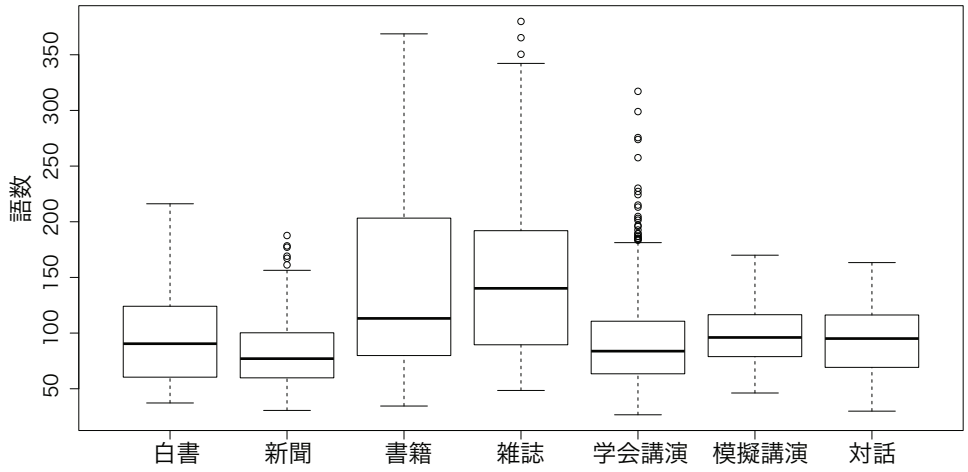


図 6：反復語 1 語あたりの平均反復間隔(全語での間隔)

表 6 と図 6 では、雑誌を除く各データの中央値が概ね 100 語前後の値となっている。つまり、同一の名詞を反復させる間隔そのものは、話し言葉・書き言葉に拘らず、基本的に 100 語前後が平均的であると考えられる。ただし、雑誌の中央値は 140 語と他に比べて間隔が大きい。また、全体としては、書籍や雑誌に見られるように書き言葉は最大値の値が大きくなる傾向にあり、反復語の使用間隔を大きくとることに対しての許容度が高いと考えられる。

新聞については、表 5・図 5 と同様、書き言葉の中では例外的に最大値が小さくなっており、字数制限の影響が存在すると考えられる。また、白書は名詞のみに注目した場合の間隔においては書籍や雑誌と大差がなかったが、全ての語を測定に利用すると、書籍・雑誌に比べて最大間隔が小さくなる。

話し言葉においては、最大値のみを見れば対話から白書にかけて間隔が増加していることが見て取れる。ここでも表 5・図 5 と同様に、典型的な話し言葉である対話においては、反復語の間隔を短くする傾向にあることが分かる。

4．考察とまとめ

本発表では、話し言葉のコーパスとして CSJ、書き言葉のコーパスとして BCCWJ を取り上げ、音声言語を基礎とする話し言葉と、書記言語を基礎とする書き言葉という二つの軸を設定するとともに、それぞれに含まれるサブコーパスを、独話と対話、改まり度の高低

といった諸特性によって分けた。そうした各種の枠組みを設定することを通して、「話し言葉的–書き言葉的」という構図を連続性を持った形で捉えられるようにした。その上で、反復語の使用実態から「話し言葉的–書き言葉的」という典型的な対比構図の様相を観察した。

反復語の使用実態から見た場合、そうした対比構図がよく表されているのは平均使用間隔（表 5 と 6・図 5 と 6）であると考えられる。基本的には、平均使用間隔は書き言葉において大きな値を取ることが可能で、話し言葉的になるほど反復の間隔が短くなる。なお、平均使用間隔については、測定に用いる単位を全ての語に広げると（表 6・図 6)、書き言葉・書き言葉に拘らず概ね 100 語程度が平均的な反復間隔となる。

「話し言葉的–書き言葉的」という観点での連続性は、反復語の使用の一定性・ばらつき（表 3・図 3）についてもある程度までは当てはまると考えられる。この場合、話し言葉的になるほど、反復語を一貫して用いるようになり、書き言葉においては、反復語を使用したりしなかったりといったばらつきが大きくなる。

反復語の使用率（表 2・図 2）についても、全体としては話し言葉の方が書き言葉よりも使用率が上がっており、同様の対比の中で捉えられる。しかし、反復語の使用率では、そうした対比に加えて、改まりの程度差が決定要因として関わっており、音声言語か書記言語かという典型的な話し言葉と書き言葉の対比構図の中だけでは捉えにくくなる。

反復語となる語の平均頻度（表 4・図 4）においては、話し言葉的・書き言葉的という連続性はあまり大きな意味を持たず、むしろ改まりの程度によってその量を変動させている。

また、全体を通して新聞サブコーパスが特異な値を示すことが多く、反復語においては、字数制限による使用制約が、諸要因の中でも強い力を持って作用することが多いと考えられる。

以上の影響関係を概略的に示せば、次の表 7 のようになる。

表 7：反復語の使用実態に関わる諸要因

	話し言葉–書き言葉の影響	改まりの高低の影響	対話–独話の影響	字数制限の影響
反復語の使用率	有り	有り	無し	有り
	話し言葉 ≧ 書き言葉	改まり高 ≧ 改まり低	φ	新聞 ≦ その他
反復語使用のばらつき	有り	無し	無し	無し
	話し言葉 ≦ 書き言葉	φ	φ	φ
反復語1語あたりの頻度	無し	有り	無し	有り
	φ	改まり高 ≧ 改まり低	φ	新聞 ≦ その他
反復語の使用間隔	有り	無し	有り	有り
	話し言葉 ≦ 書き言葉	φ	対話 ≧ 独話	新聞 ≦ その他

また、この結果から、本発表では話し言葉と書き言葉の関係性について、以下の二つの事実を指摘できる。一つは、話し言葉と書き言葉は必ずしもある決定的要因によって性質を分離させているのではなく、典型的話し言葉から典型的書き言葉にかけて大きなグラデーションを描きながら変化させている点であり、これは量的構造から見たときの両者の連

続性を意味している。もう一つは、反復語の 1 語あたりの頻度や、字数制限を受ける新聞というサブコーパスの値において見られたように、状況次第では話し言葉と書き言葉という区分自体が有効に機能しなくなり、それ以外の要因、本発表であれば改まり度や字数制限のような要因によって表現の性質が決定されることがあるという点である。

なお、調査結果の違いはあくまでも量的差異による概略的なものであるため、それが具体的にどのような質的違いに基づいて現れたものなのかという点については、テキストの内実の観察を通した分析により達成されなければならない。この点は今後の課題としたい。

文　献

井上次夫（2011）「書き言葉らしさの判断と測定」『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)』予稿集』, pp.89-96

小磯花絵・田中弥生・小木曽智信・近藤明日子（2011）「テキストの多様性をとらえる分類指標の構築を目指して」『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）予稿集』, pp.431-442

小椋秀樹（2006）「第 3 章 形態論情報」『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.133-186

小椋秀樹（2008）「『日本語話し言葉コーパス』の言語単位」『日本語学』27.5, pp.72-81

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』形態論情報規定集 第 4 版（上）（下)』国立国語研究所内部報告集 LR-CCG-10-05-01, LR-CCG-10-05-02（特定領域研究「日本語コーパス」研究成果報告 DVD 所収）

塩澤和子（2005)「コラムに観察されるくり返しの機能」『文藝言語研究 言語篇』47, pp.15-31

田中妙子（1997）「会話における〈くりかえし〉−テレビ番組を資料として−」『早稲田大学日本語研究教育センター紀要』9, pp.47-67

中田智子（1991）「会話にあらわれるくり返しの発話」『日本語学』10.10, pp52-62

馬場俊臣（2006）『日本語の文連接表現−指示・接続詞・反復−』おうふう

山崎誠（2010）「語の平均使用頻度に現れるテキストの特徴」『特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』, pp.5-14

引用資料

「日本語話し言葉コーパス」（科学技術振興調整費開放的融合研究『話し言葉の言語的・パラ言語的構造の解明に基づく「話し言葉工学」の構築』）

「現代日本語書き言葉均衡コーパス・コアデータ」（特定領域研究「日本語コーパス」データ班，特定領域研究「日本語コーパス」研究成果報告 DVD 所収）

関連 URL

The Perl Programming Language	http://www.perl.org/
The R Project for Statistical Computing	http://www.r-project.org/

モラウとイタダクのヲ格名詞・動名詞の違いについて

岩井　智哉（大阪大学文学研究科）

On the Difference of *wo* Noun Phrases When They are Followed by *morau* and *itadaku*

Tomoya Iwai(Osaka University)

１．はじめに

本発表では、～ヲモラウと～ヲイタダクという形について、ヲ格の名詞・動名詞[1]を比較することで、モラウとイタダクの本動詞的に見える用法に敬意だけではない違いがあることを示す。調査方法としては、昨年の秋に『現代日本語書き言葉均衡コーパス』(以下、BCCWJとする)の検索アプリケーション「中納言」が公開されたことを受けて、BCCWJを利用する。～ヲモラウと～ヲイタダクの直前に来る語・形態素を比較することで、～ヲイタダクは～ヲモラウに比べてスルを下接してサ変動詞として使える動名詞をヲ格に取っていることが多いことを示す。

２．モラウとイタダク

モラウとイタダクには本動詞の用法と補助動詞の用法がある。

(1) 花子が先生に数学の教科書をもらう/いただく。
(2) 花子が先生に数学を教えてもらう/いただく。
(3) 花子が先生に数学をご教授いただく。

例文(1)は本動詞、(2)と(3)は補助動詞の用法である。この補助動詞としての意味・用法については由井 (1993)、菊地(1997)、高見・加藤 (2003)、山田 (2004)などかなり充実した考察がある。オ・ゴ～イタダクの形については菊地(1997)が「お書きになっていただく」「ご指導なさって(になって)いただく」の「になって」「なさって」の部分を端折ったものだと見ると分かりやすいとしている。

しかし、例えば「ご教授をいただく」など、形の上ではヲ格をとっていて本動詞のように見えても、意味は例文(3)とほとんど変わらない用法も見聞きする。このようなものについての研究は管見の限り見当たらない。これについて BCCWJ を用いて調べてみたい。

３．モラウとイタダクのヲ格

３.１　調査方法

「中納言」で BCCWJ のすべてから短単位検索の語彙素読みでそれぞれ「モラウ」「イタダク」と指定し、モラウとイタダクの全用例を検索した。さらにその中からモラウ・イタダクの直前に助詞のヲが来ているものを抜き出し、ヲの直前に来ている名詞・動名詞を集計した。

モラウの全用例は 41954 例、そのうち～ヲモラウは 4597 例あった。

イタダクの全用例は 35438 例あった。そのうち～ヲイタダクは 5063 例であった。

３.２　モラウのヲ格

まず、～ヲモラウについてヲ格の名詞を見る。上位 20 語を挙げる。スルを後接してサ変動詞として使えるもの（動名詞の用法があるもの）に下線を引く。

---

[1] 影山(1993)にならい、名詞のようだがスルと複合して動詞になるものを動名詞と呼ぶ。

～ヲモラウのヲ格名詞・動名詞

金 265　薬 124　給料 117　年金 104　許可 77　賞 69　手紙 64　電話 60　もの 58　元気 57　返事 56　回答 52　プレゼント 45　サイン 39　アドバイス 35　連絡 35　券 35　保険 32　メール 31

～ヲモラウ上位 4 つまでは動作性のない具体的な名詞である。また、ここに見られるサ変動詞の前項になれるものでも、「許可」「賞」「回答」「プレゼント」など動作や行為でなく、その結果できたものを表す「結果名詞」（影山　1993）と解釈できるものが多い。

３.３　イタダクのヲ格

～ヲモラウと同じく、上位 20 語を挙げ、動名詞の用法があるものに下線を引く。

～ヲイタダクのヲ格名詞・動名詞

答え 198　理解 169　答弁 144　回答 129　協力 109　意見 102　説明 80　アドバイス 75　審議 73　言葉 71　指摘 66　連絡 64　メール 58　コメント 52　評価 51　答申 50　時間 50　返事 47　電話 46　手紙 44

動作性のない具体的な名詞の順位は下がる。「理解」「協力」「審議」といった「結果名詞」としての解釈がしにくい語が～ヲモラウに比べて多くなっている。

名詞や動名詞に接頭辞のオやゴが付くものを抽出すると、その傾向がより強くなるように思われる。次に示すのは～ヲイタダクのうち、「を」の二つ前に「お」「ご」「御」が来ているもの、つまり下に示した形態素の直前に「お」「ご」「御」のいずれかがあるものである。

オ・ゴ～ヲイタダクのヲ格名詞・動名詞

答え 186　理解 160　答弁 103　協力 79　意見 77　審議 68　指摘 64　説明 61　話 41　許し 41　議論 35　連絡 31　返事 30　示し 28　言葉 27　祝い 26　茶 26　手紙 26　検討 24　努力 23

オ・ゴ～ヲイタダクでは「理解」「協力」「審議」は以下のように用いられている。下線は引用者。

(4) 先生のおっしゃいましたように、再建監理委員会は当時としての御推計はされましたけれども、最終的には政府においてきちんとした数を定めるということと御理解をいただいて結構だと思います。（参議院　国会会議録　第 104 回国会　1986）
(5)調査にご協力をいただいた聖徳大学の■■■■さんの見解をもとに、ご紹介します。（広報つちうら 2008 年 02 号　茨城県土浦市）
(6) ただ、民営・分割までの間まだ時間がございますし、法案もまだ御審議をいただかなければならない。（参議院　国会会議録　第 104 回国会　1986）

これらの例はイタダクをモラウと置き換えると容認しにくくなるように感じられる。

(7) ?最終的には政府においてきちんとした数を定めるということと御理解をもらって結構だと思います。
(8) ?調査にご協力をもらった聖徳大学の■■■■さんの見解をもとに、ご紹介します。
(9) ?法案もまだ御審議をもらわなければならない。

このような例は本動詞よりも補助動詞用法とされるオ・ゴ～イタダクに近いものと考える方が適切だろう。

３.４　ヲ格名詞の修飾

～ヲイタダクが補助動詞的な用法を持つことは、名詞・動名詞の修飾にも表れている。

ここでは動名詞として解釈しうる語の中で、～ヲイタダクで最も多かった「理解」と、～ヲモラウで最も多かった「許可」を例とする。ヲ格名詞が連体形や連体助詞ノで修飾されているときは名詞的に、連用形や副詞、助詞のトや連用形＋テで修飾されていたり格をとっているときは動名詞的に扱われていると仮定して、どちらであるか実際の用例を見てみる。連用形や副詞などで修飾されていても修飾語句が「充分に」や「ぜひ」などイタダクやそれに続く助動詞などまで含めて修飾すると考えられる場合は、それを根拠に動名詞的に扱われているとは判定しない。

「理解」は～ヲイタダクでは多いが、～ヲモラウでは 2 例しかない。明確に判定できるもののうち、～ヲイタダクにおいて「理解」が名詞的に修飾されているものが 26 例。

(10) 山崎さんや石井君の協力に加え、会長の福川忠昭先生や副会長の伊集院謙信先生にも深いご理解をいただき、ＯＢ・ＯＧ会では勇貴のことを話す時間をとってくださった。（有村　英明『届かなかった贈り物』講談社 2005）

(11) まあ地方公共団体というのはなかなか調整が困難な問題、あるいは環境問題で周辺地域の住民の方々の御理解をいただけなかったというような問題も間々あるわけでありまして、（国会会議録　第 094 回国会　1981）

動名詞的に修飾されているものが 40 例あった。

(12) 私たちの財産である大芝高原を守り育てるこの取り組みにつきましてご理解をいただき、ご参加いただきますようよろしくお願いします。（広報みなみみのわ　2008 年 10 号長野県上伊那郡南箕輪村 2008）

(13) そして最後に、じゃ、社会保険庁を抜本的に見直す、組織を見直すという答えを出そうというこのことで作業を進めてきたと、こういうふうに御理解をいただきたいと思います。（国会会議録　第 162 回国会　2005）

～ヲモラウでは 1 例が動名詞的に扱われていた。

(14) 姑に意見するのは、多少勇気がいると思いますが・・・旦那様にもよくご理解を貰って、頑張って下さい。（Yahoo!知恵袋　　Yahoo!　2005）

理解

	～ヲモラウ	～ヲイタダク
名詞的	0	26
動名詞的	1	40
その他	1	103
総計	2	169

「許可」は～ヲモラウでも～ヲイタダクでも名詞的な扱いが多かった。1

(15) だけど、ボートを使う許可をもらったほうがいいだろう。（C.アドラー/久米穣『ぼ

くたちの宝島』金の星社 1991）

(16)「でも、よくわからないな。そんなの黙って勝手に想像していればいいじゃない。いちいち私の許可をもらわなくたって、君がなにを想像しているかなんて、私にはどうせわかりっこないんだから」（村上春樹『海辺のカフカ』新潮社 2005）

(17) 山村先生はバーネットがあまりお好きでなく、「そんなにうまくいくか」という反応でしたが、なんとかねばって、とうとう「勝手にやれ」という許可をいただきました。（林　昭/バレンチナ･オスタペンコ/宮澤　正顯『からだをなおす』昭和堂 2003）

(18) Ｋさんの許可をいただきましたので、今日、明日、ブログ上写真展を開催します。（Yahoo!ブログ　　Yahoo!　2008）

動名詞的な扱いは少なく、～ヲモラウで 2 例だけだった。

(19) その協会の講師免状をもらっている人（つまり協会から教室を開いてもよいと許可をもらっている人）のうち、実際に教室を開いている方は、なんと全体の三十％しかいないのです。（犬塚義人『はじめよう！カルチャー教室』同文舘出版 2004）

(20) で、自分で修理して良いよと許可を貰えたら、修理に移ります。（Yahoo!知恵袋　Yahoo!　2005）

～ヲイタダクでは「許可」が動名詞的な扱いを受けているものは見られなかった。「許可」は、～ヲモラウ・～ヲイタダクに接続するときは名詞としての解釈が多いと言える。

許可

	～ヲモラウ	～ヲイタダク
名詞的	27	10
動名詞的	2	0
その他	48	10
総計	77	20

「理解」は動名詞的な扱いで～ヲイタダクに前接することが多く、「許可」は名詞的な扱いで～ヲモラウに前接することが多かった。

４．まとめ

本発表では、BCCWJ を用いた調査から、～ヲモラウに比べて～ヲイタダクのヲ格には動名詞になるものが現れやすいこと、それらには連用形や副詞で修飾されるものが少なくないことを示した。

ここから、～ヲモラウと～ヲイタダクに単なる敬意の差があるだけでなく、形の上では本動詞の用法に見える～ヲイタダクに、補助動詞用法に近い抽象的な用法があると考えられる。

文　献

影山太郎(1993)『文法と語形成』ひつじ書房
菊池康人(1997)『敬語』講談社
由井紀久子(1993)「モラウの意味的抽象化・希薄化の過程」『阪大日本語研究』5　pp.83-93
山田敏弘(2004)『日本語のベネファクティブ―「てやる」「てくれる」「てもらう」の文法―』明治書院
高見健一・加藤鉱三(2003)「受益表現の新展開」『言語』32-1～32-6

コーパスを用いた中国語ネット語の判定システム

竇　梓瑜（東京農工大学　工学府　情報工学専攻）
古宮 嘉那子(東京農工大学 工学研究院 先端情報科学部門)
小谷 善行(東京農工大学 工学研究院 先端情報科学部門)

A Detection System of Chinese Netspeak Using Text Corpus

Ziyu Dou(Graduate School of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

１．研究背景

現在、中国では、インターネット利用者が爆発的に増え、それに伴って大量な中国語ネット語（以下、ネット語と書く）が現れた。ネット語の独特な言葉やその使い方は、インターネットだけではなく、徐々に人々の生活にも浸透してきている。しかし、中国の人口はおよそ１３．５億であるのに対して、中国のインターネット利用者は５．１３億と言われており、中国人の半分以上はインターネットを利用していない。そのような人々にとって、ネット語は理解しにくく、意味が分からなかったり、または意味の誤解から、トラブルになったりすることがある。こうした事態を避けるため、我々はコンピューターで自動的にネット語か書き言葉かどうかを区別するシステムを作成した。本システムは任意の中国語の一つ以上の文の入力に対して、ネット語であるかどうかの判断結果を出力する。

２．中国語ネット語特徴の検出システムの構成

中国語ネット語特徴検出システムの構成を、図１に示す。

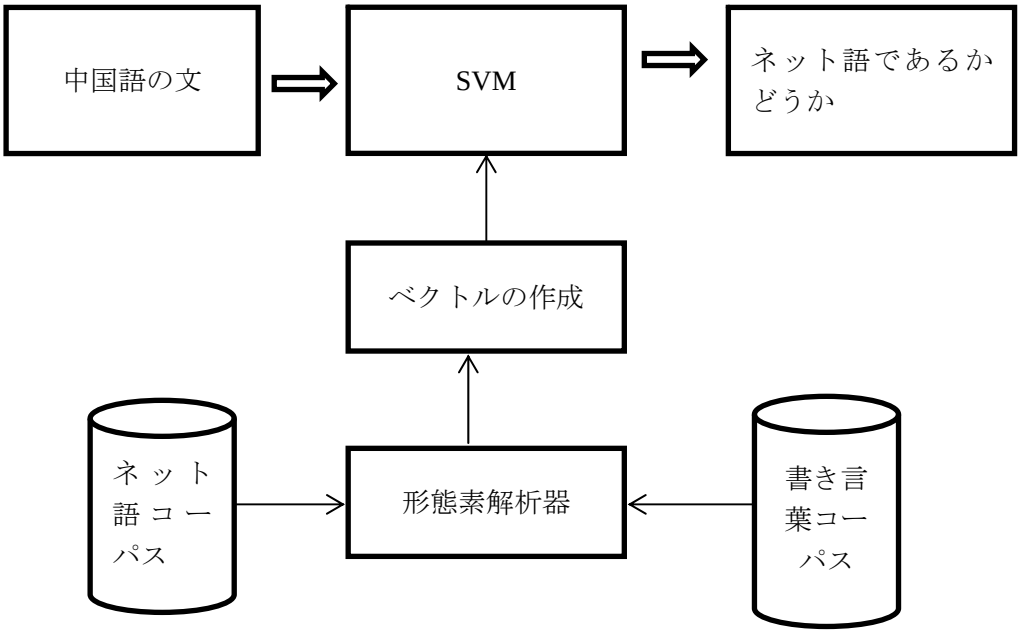


図１　中国語ネット語特徴の検出システムの構成

システムの入力は中国語文であり、出力は、ネット語であるかどうかである。
まず、ネット語コーパスと書き言葉コーパスの形態素解析を行い、それを特徴としてベクトルを作成し、サポートベクターマシン（SVM）（http://otndnld.oracle.co.jp/document/products/oracle11g/111/doc_dvd/datamine.111/E05704-02/algo_svm.htm）を用いて機械学習を行う。次に、入力の中国語文に対して同様に形態素解析を行い、ベクトルを作成し、先ほど学習した学習器を利用して、入力の中国語文がネット語かどうかを判定する。以下に各部分について順次述べる。

２．１　形態素解析

形態素解析の部分はICTCLASという中国科学院計算技術研究所の無料形態素解析プログラム（http://ictclas.org/）を使用した。このプログラムはHMMを基本解析方法として構成されている。プログラムはC言語版とC++言語版が存在するが、本システムで使われているのはそのC++言語版である。なお、この形態素解析プログラムの基本機能は単語分割、品詞つけ、未知語表記で、ユーザ自身でも辞書に単語を入れることが出来る。

このICTCLASプログラムの役割は、書き言葉コーパスの文、ネット語コーパスの文及び入力の文の形態素解析を行い、品詞タグを付けることである。また、ここで使用される品詞タグには、計算所漢語詞性標注集（http://icl.pku.edu.cn）が使われている。

２．２　ベクトルの作成とSVM

SVMにはSVM－lightという無料プログラムを使用した。SVMの入力はベクトルであるため、前処理としてベクトルを作成した。まず、ネット語コーパスと書き言葉コーパスの形態素解析の結果を合わせ、出現した全ての単語を統計し、単語毎に番号を付ける。ただし、ここでは、単語が同じでも、品詞が違う場合には、異なる単語として違う番号を付与した。特にここで統計するとき、アルファベットを除くため、タグ　/x　が付いている単語を全部除いた。ベクトル作成は、コーパスの文ごとに行った。素性は単語であり、素性値はコーパス中の頻度である。

３．実験と結果

３．１　　データ

実験用のデータは、全てインターネットから収集したものである。

このうち、ネット語コーパスは、新浪微博（http://www.weibo.com）から収集した。この新浪微博は、現在２０１２年１月までに、２．５億を超えるユーザを持つ、中国最大のミニブログである。このミニブログはツイッターと同様、一発言として入力できるのは１４０文字までという制限がある。政府などの公式機関のユーザも多数あるが、ほとんどの発言はインターネット利用者の日常的な呟きなので、典型的なネット語があると考えられる。

これらのインターネットユーザ発言を収集するとき、火車採集器（http://www.locoy.com）という無料ウェブデーター収集プログラムを利用し、無作為にユーザを選択してデータを取得した。最初に収集されたデータは既にHTMLタグを全部除いた文である。このような文が一行一文という形で、テキストの中で記録されている。新浪微博から収集したネット語コーパスの文の数は５０００文である。このほかに、百度貼バ（BBS サイト）（http://tieba.baidu.com）から１００文を、ネット語のコーパスとしてテストに利用した。

書き言葉コーパスは前述のように中国国家文字委員会の現代中国語コーパス（http://www.cncorpus.org/）の中の新聞と社論というカテゴリのコーパスから取ったも

のである。このコーパスは、ネット語コーパスと同じく一行一文でテキストの中に記録されている。現代中国語コーパスから利用する文の数は２０００文である。このほかに、SOHU ニュース（http://www.sohu.com/）から１００文を、書き言葉のコーパスとしてテストに利用した。

３．２　　実験設計及び結果

ネット語コーパスと書き言葉コーパスのどちらの文かを判定する制度を見るために、実験１～実験４の四種類の実験を行った。以下にそれぞれについて述べる。

（1）実験１　CLOSED　テスト

まず、ネット語コーパスと書き言葉コーパスのどちらの文かを判定する制度を見るための CLOSED テストを行った。CLOSED テストでは、テストデータは訓練データとして利用したものである。また、CLOSED テストでは、顔文字などの符号を削除していない（略：符号あり）実験を行った。以下に、CLOSED　テストにおける、訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表１に示す。表１から、CLOSED テストでは、全ての文がネット語であるかどうか、正しく判定できていることが分かる。

訓練データ：ネット語コーパス　５０００文、書き言葉コーパス　２０００文
テストデータ：ネット語コーパスからの１０００文と書き言葉コーパスからの４００文
結果：ネット語の率＝SVM がネット語として認識した文/１０００
　　書き言葉の率＝SVM が書き言葉語として認識した文/４００

ここで、ネット語におけるネット語の率はネット語の再現率であり、書き言葉における書き言葉の率は書き言葉の再現率となる。

表１　CLOSED　テスト　結果

	ネット語の率	書き言葉の率
ネット語コーパスからの１０００文（符号あり）	１００％	０％
書き言葉コーパスからの４００文（符号あり）	０％	１００％

（2）実験２　OPEN　テスト（符号あり及び符号なし）

次に、訓練データとテストデータの重複を許さない OPEN テストの二つを行った。OPEN テストでは、符号と英文字によって構成され顔文字や略語の結果への影響を実証するため、符号および英文字がある実験と符号および英文字がない（略：符号なし）実験を行って比較した。以下に、OPEN テストにおける、訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表２に示す。

訓練データ：ネット語コーパス　４０００文、書き言葉　１６００文
テストデータ：ネット語コーパスの上記４０００文を除いて残った１０００文、書き言葉

コーパスの上記１６００文を除いて残った４００文
結果：ネット語の率＝SVM がネット語として認識した文/１０００
　　書き言葉の率＝SVM が書き言葉語として認識した文/４００

表２　OPEN テスト　結果

	ネット語の率	書き言葉の率	正解率
ネット語コーパスからの１０００文（符号あり）	９８．４％	１．６％	９２．６％
書き言葉コーパスからの４００文（符号あり）	２２．０％	７８．０％	
ネット語コーパスからの１０００文（符号なし）	９８．９％	１．１％	８４．９％
書き言葉コーパスからの４００文（符号なし）	５０．２％	４９．８％	

（3）実験３　ネット語１００文と書き言葉１００文　テスト（符号あり及び符号なし）
次に、訓練データとして使用したコーパスとは異なるコーパスとして、ネット語コーパスに BBS、書き言葉コーパスに新聞を利用した際の OPEN テストを行った。以下に、その際の訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表３に示す。この実験の際にも実験２と同様、符号ありと符号なしの実験を行って比較した。
訓練データ：ネット語コーパス　５０００文、書き言葉コーパス　２０００文
テストデータ：　百度貼バ　（BBS サイト）から取った１００文をネット語とし、SOHU ニュース　から取った１００文を書き言葉とし、テストを行う
結果：ネット語の率＝SVM がネット語として認識した文/１００
　　書き言葉の率＝SVM が書き言葉語として認識した文/１００

表３　ネット語１００文と書き言葉１００文　テスト　結果

	ネット語の率	書き言葉の率	正解率
ネット語１００文（符号あり）	８７％	１３％	７１．５％
書き言葉１００文（符号あり）	４４％	５６％	
ネット語１００文（符号なし）	６９％	３１％	８３．５％
書き言葉１００文（符号なし）	２２％	７８％	

（4）実験４　アンケート（符号ありと符号なし）

最後に、比較対象として、人間に符号ありと符号なしの際、どの程度ネット語を判定できるかのアンケートを行った。以下に、アンケート実験における、訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表４に示す。この実験の際にも実験２、実験３と同様、符号ありと符号なしの実験を行って比較した。

テスト方法：数人の中国人インターネット利用者にアンケート

テスト内容：上記の　百度貼バ　（BBS サイト）から取ったネット語５０文を、SOHU ニュース　から取った５０文を符号ありと符号なし二回、被験者に判断してもらった。顔文字で簡単に人間がネット語を判断し、その結果を覚えてしまう可能性を排除するため、先に符号なしのアンケートを行い、その後符号ありのアンケートを行った。また、

結果：ネット語の率＝SVM がネット語として認識した文/５０

　　書き言葉の率＝SVM が書き言葉語として認識した文/５０

として計算した。

表４　ネット語５０文と書き言葉５０文の人工判定テスト結果

	ネット語の率（符号あり）	書き言葉の率（符号あり）	正解率（符号あり）	ネット語の率（符号なし）	書き言葉の率（符号なし）	正解率（符号なし）
２０代中国人男性学生	４４％	１００％	７２％	４４％	１００％	７２％
２０代中国人女性学生	２０％	１００％	６０％	１６％	１００％	５８％
平均	３２％	１００％	６６％	３０％	１００％	６５％

４．考察

まず、表２の OPEN テストの結果が示すように、本研究のネット語コーパスおよび書き言葉コーパスは、確実に区別が存在する。特にネット語コーパスは、符号ありと符号なしの場合、それぞれ、９８．４％と９８．９％の再現率となった。しかし、書き言葉コーパスに対する Open テストは、符号ありのとき、再現率が７８．０％で、符号なしの場合の再現率は４９．８％まで下がった。これは、訓練に使用したネット語コーパスの量が多いためと、ネット語コーパスの中でも、書き言葉のような文が多数存在するためであると思われる。また、符号がある場合とない場合の再現率の差から、機械学習において、符号の影響が大きいことが分かる。

続いて、実際な文に対するテストの結果（表３）を分析する。まず、符号がある場合とない場合と比べると、ネット語１００文に対する認識正解率は１８ポイント上がり、８７％まで達成した。これに対し、書き言葉は符号なしの場合のほうが２２ポイント上回り、７８％まで達成した。これによって、符号がある場合、文がネット語として認識される傾向が強まり、符号がない場合には、書き言葉として認識される傾向が強まることが分かる。

アンケート（表４）の結果を見ると、全部書き言葉の判定は１００％正解したが、ネット語の判定はいずれも４４％と２０％まで止まったことが分かる。また、符号がある場合の、符号がない場合と比べた正解率の上昇はわずかであった。実験４と実験３の正解率を

比べると、機械のほうが、正解率が上回ることがわかった。とこれは、ネット語といっても、BBS では、書き言葉的な表現も多数存在することが、判定の結果に大きく影響したためだと思われる。それに対し、書き言葉は、１００％の正解率で、人間が書き言葉を認識するのは簡単だったことが分かる。ネット語の定義を人間が判断できるものとすれば、再現率にも変化があるだろう。

最後に表３と表４から機械学習と人間の正解率を比べる。表３から、機械学習は最高８３．５％、表４から、人間の判断は最高６６％であるため、作成したシステムの性能が人間に上回ることがわかる。

５．結論

本論文では、文の入力に対してネット語かどうかの判定を行うシステムを作成した。入力の文は形態素解析を行い、ベクトル化したあと、SVM を使ってネット語かどうかを判定した。

実験結果から、本システムは、符号がある場合、ネット語に対する判定の正解率が上がり、符号がない場合、書き言葉に対する判定再現率が上がることが分かった。また、人間に対するアンケートの結果から、人間でもネット語かどうかの判定が難しいことが分かる。特に、BBS やミニブログなどの情報には、ネット語的な特徴がない言葉も多数存在するので、難しかったようである。また、機械学習と人間の正解率を比べると、機械学習は最高８３．５％、人間の判断は最高６６％であるため、作成したシステムの性能が人間に上回ることがわかった。

文　献

佐藤敏紀　「Perl で自然言語処理」　東京工業大学奥村研究室
http://www.slideshare.net/overlast/perl-5460697

谷岡　広樹、丸山　稔(2005)「形態素解析に基づく SVM を用いたアスキーアートの識別」電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 「104:670」, pp.25-30

黒橋禎夫　機械学習に基づく自然言語処理　京都大学情報学研究科
http://nlp.ist.i.kyoto-u.ac.jp/member/kuro/lecture/LIP10/LIP09.pdf

Jin'ichi Murakami, HMM(Hidden Markov Model, 隠れマルコフモデル)
http://unicorn.ike.tottori-u.ac.jp/murakami/doctor/node7.html

語料庫在線　http://www.cncorpus.org/

SVM-light　http://www.cs.cornell.edu/People/tj/svm_light/

ICTCLAS http://ictclas.org/

情報と通信のハイパーテキスト http://www.yobology.info/text/index.htm

Shogo Computing Laboratory http://sora-blue.net/~shogo82148/memo/algorithm/svm/

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―

小磯 花絵（国立国語研究所理論・構造研究系）†
石本 祐一（国立情報学研究所音声メディアグループ）

Prosodic Features of Utterances in the Corpus of Spontaneous Japanese: Intonational Phrase–Based Approach

Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)
Yuichi Ishimoto (Speech Media Group, NII)

1. はじめに

本研究は，『日本語話し言葉コーパス（*Corpus of Spontaneous Japanese*:以下 CSJ）』（前川 2004, 2006）を用いて，自発音声における「発話」の韻律的な傾向を探ることを目的とする。Pierrehumbert & Beckman(1988) では，アクセント句やイントネーション句より上位に「発話（utterance）」という単位を設定している。「発話」は，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）が見られる範囲であり，その末尾で final lowering（平叙文末尾で F0 が局所的に下降し発話の終了を示す現象）が生じるとされる。ニュースのように，ある程度発話内容が準備されたスピーチであれば，「発話」は文末表現で区切られた文に相当する単位ということになろう。しかし自発性の高い話しことばでは，次の例に見られるように，複数の節をつなげて長々と発話を続けることも少なからず見られる。このように，途中に強い統語境界をはさみうる長い「発話」であっても，全体的に declination が生じ末尾に final lowering が観察されるのか，という素朴な疑問が生じる。

> で（エ）絶対音感群に比べますと 多少半音から一音のエラーが見られますが （エ）こちらの条件あるいはこちらの条件の時はさほどオクターブエラーがたくさん見られる訳ではないんですが （アイアル）ＩＲＮ条件の時には（エ）１オクターブエラーがかなり見られてることが分かります
> （談話 ID: A01F0067）

そこで本研究では，F0 の下降現象に着目し，「発話」の長さや統語構造との関係について調べる。手続きとしては，CSJ に付与された節単位情報を用いて暫定的に「発話」を定義した上で，イントネーション句に基づき節単位の韻律的特徴を調べる。Maekawa (2009)・前川 (2011) も同種の分析をアクセント句を単位に行っているが，イントネーション句の中でダウンステップの生じるアクセント句よりも，イントネーション句の方が長い発話の全体的な F0 の推移をとらえる上で適切であると考え，分析単位として採用した。アクセント核の有無など細かな統制のもとで分析を実施した前川の結果と合わせて考察を進める。

† koiso@ninjal.ac.jp

2. 方法

2.1 データ

分析には 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 RDB（小磯ほか 2012）を用いた。

2.2 節単位

「発話」に相当する単位を認定するにあたり，CSJ に付与されている節単位情報を利用した。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り，CSJ において構文・談話レベルの情報を付与するための基本単位として設計されたものである（丸山ほか 2006)。節単位は，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：　いわゆる文末に相当する境界。明示的な文末表現の直後。
強境界：　後続の節に対する従属度の低い、つまり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、つまり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」と認定される。

ここで絶対境界は文末に相当するため，絶対境界と絶対境界で区切られる範囲を「発話」と解釈するという考え方もあろう。しかし大石（1971）が指摘するように，話しことばでは「…ケド」「…ガ」「…ノニ」「…テ」「…タラ」などの接続表現がある種のイントネーションやポーズを伴い文末を表わすのに転用されることがある点に注意しなければならない。例えば「今日はもう帰らないといけないんだけれど」という発話の場合，末尾の音調が句末を強調するような上昇調や上昇下降調で発話された場合には，そのあとにまだ発話が続くことを予感させるのに対し，末尾に final lowering を伴い一定以上のポーズが後続した場合には，必ずではないものの相対的にそこで発話が終了したと感じられる傾向にある。つまり，強境界の末尾は「発話」の末尾であることもあれば「発話」の途中であることもある，ということであり，節単位のラベルからそれを判断することはできないのである。

そこで分析 1 では，直前が絶対境界である（強境界ではない）絶対境界を最後に持つ節単位（図 1 の節単位 4 と 8）を対象に，節単位の長さと F0 の下降現象との関係を見る。この条件で分析対象となるデータは，その内部に強境界など強い統語境界を持たないケースである。

次に分析 2 では，発話内部に強い統語境界が置かれる場合を対象に，そこで F0 の下降が継続するか否かを見る。具体的には，強境界の節単位と絶対境界の節単位の二つの連鎖（図 1 の節単位 2・3 と 6・7）を対象とする。

この条件に相当する事例を幾つか聴取したところ，大石（1971）が指摘するようなある種のイントネーション・ポーズを伴い「発話」が終了したと感じられる事例はあまり無く，「発話」が継続し最後の絶対境界で終了するものが多くを占めた（この点については今後改めて確認する必要がある）。この観察が正しいならば，分析 2 で対象とするデータ（の大半）は一つの発話の中に強い統語的境界（強境界）が（少なくとも）一つ存在するケースということになる。

2.3 イントネーション句

「発話」の韻律的特徴を探るために，本研究ではイントネーション句（Intonational Phrase; 以下 IP）を用いる。IP は，アクセント核が後続するアクセント句（Accentual Phrase; 以下 AP）の F0 ピークを反復的に低下させるダウンステップの生じる領域であり，IP 境界でピッチレンジのリセットが生じる。

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，Break Index (BI）という，韻律境界の切れ目の強さに関する情報が存在する。BI=2 は AP 境界，BI=3 は IP 境界，BI=F はフィラー境界（例：「えっと」「あー」），BI=D は言い淀み境界（例：「す 全ての」）に対応する。ここでは，BI=3, D, F で区切られる範囲を IP と認定した。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。

このように IP を認定した上で，X-JToBI に付与されたトーンの情報に基づき，IP の F0 最大値，F0 最小値，ピッチレンジを次のようにもとめた（図 2）。

F0 最大値： IP 冒頭の 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

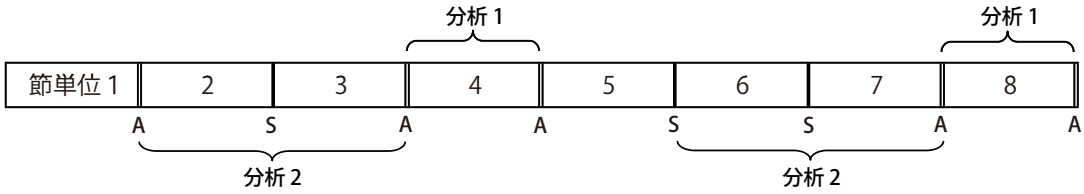


図 1 分析対象とする節単位（A:絶対境界，S:強境界）

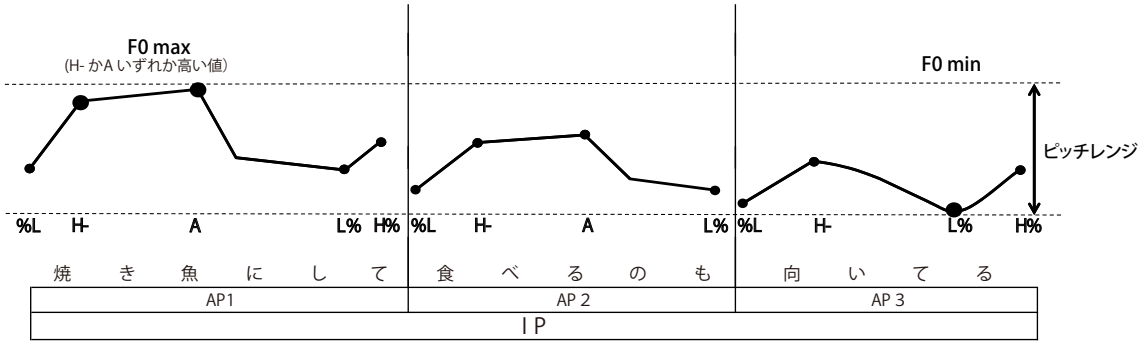


図 2 分析に用いる IP の F0 パラメータ：F0 最大値（F0 max）・F0 最小値（F0 min）・ピッチレンジ

表 1　節単位に含まれる IP 数とその頻度

IP 数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21〜
頻度	570	546	656	569	544	478	364	285	228	192	127	90	68	49	42	30	19	17	14	7	19

ピッチレンジ：　F0 最大値と F0 最小値の差

概ね節単位の方が IP よりも長い単位であり前者が後者を内包するが，まれに交差することもある（例えば接続詞の前にピッチレンジがおかれず前の発話と続けて発話される場合など）。このような交差が生じた箇所は分析対象外とした。

3. 結果と考察

3.1　分析 1

「発話」内部に絶対境界など強い統語境界を持たないケースを対象に，節単位の長さと F0 の推移（特に下降現象）との関係を見る。具体的には，直前が絶対境界である（強境界ではない）絶対境界を最後に持つ節単位を対象とする（図 1）。節単位の長さの指標として，節単位に含まれる IP 数（フィラー・言い淀みを除く）を用いる。

分析 1 で用いるデータは，節単位数 4914，IP 数 26763 であり，1 節単位の平均 IP 数は 5.4 である。表 1 は，節単位に含まれる IP 数ごとの頻度である。以下に節単位中の IP 数が 1 の場合，6 の場合，12 の場合の発話例を記す。

IP 数 1 の場合　長調のメロディーです
IP 数 6 の場合　次に強さの影響を検討する強さパターン制御条件では弱強型強弱型共に高さは一定にしており強さのみを弱強にした音もう一つは強弱にした音です
IP 数 12 の場合　そして（エー）照合システムから（エー）テキストが提示されましたらこの音素HMMを（エー）このテキストに従って接続しパラメーター（ワ）（ン）（エー）スペクトルパラメーターおよびピッチを生成して（エー）合成フィルターにより音声を合成します

図 3 は，節単位中の IP 数ごとに IP の F0 最大値・最小値・ピッチレンジの推移を図示したものである。例えば六つの IP から構成される節単位の場合（図中 len=6），冒頭から数えて 1〜6 番目の IP の各 F0 パラメータの分布がこの順に示されている。また F0 は談話ごとに Z 値に変換している。

図を見ると，節単位に含まれる IP 数に関わらず，節単位冒頭から末尾にかけ，F0 最大値・最小値が徐々に下降する傾向にあることが分かる。特に F0 最大値（IP 冒頭の AP の H-あるいは A の F0 値）と比べて F0 最小値（IP 末尾の AP の L% の F0 値）の下がりが大きいため，ピッチレンジも徐々に下降する。紙面の都合で省いたが，ここに載せていない長さの節単位についても同様の傾向が見られる。またこの傾向は男性・女性に分けた場合にも概ね見られる（図 4・5）[*1]。

[*1] 冒頭二つの IP のピッチレンジが同じあるいは増加していることもあるが，これは冒頭に接続詞が置かれることが多く，その他の IP と異なる性質を有するためと推察される。接続詞の特殊性については Maekawa (2009) でも指摘されている。

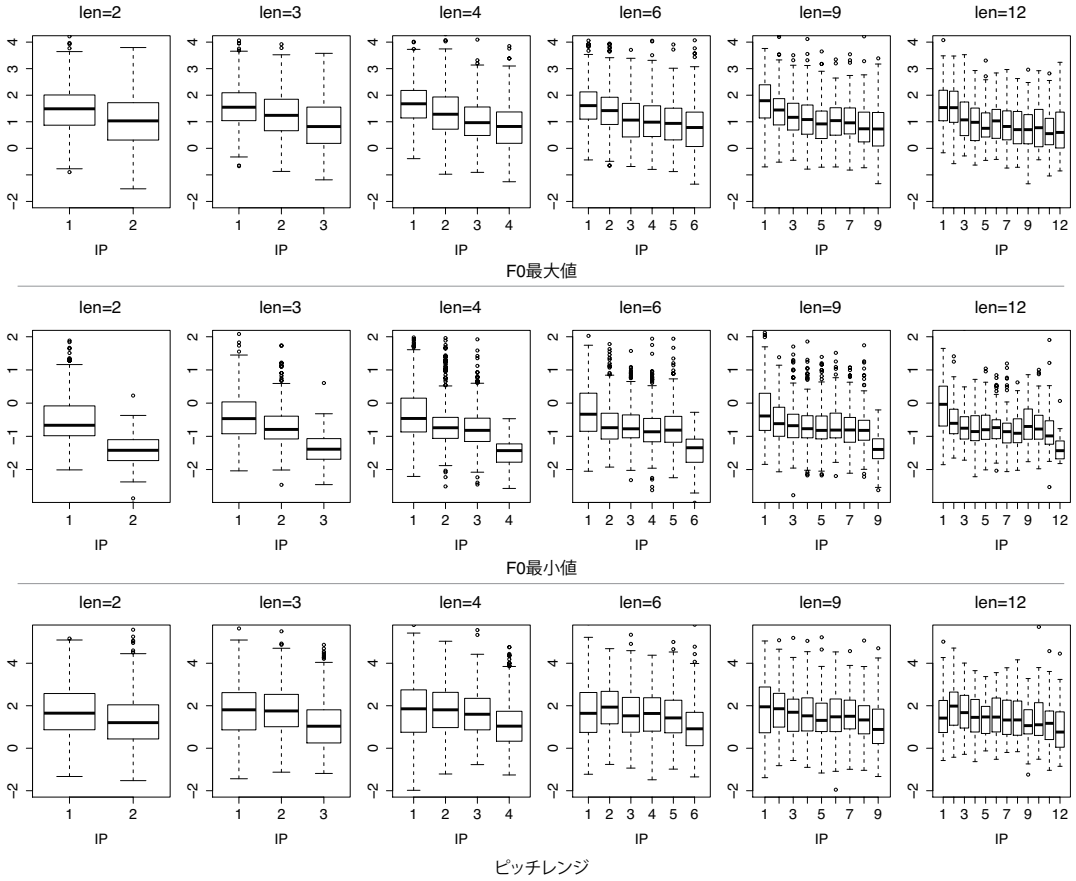


図 3　節単位（絶対境界）内のイントネーション句の F0 パラメータ (Z 値) の推移

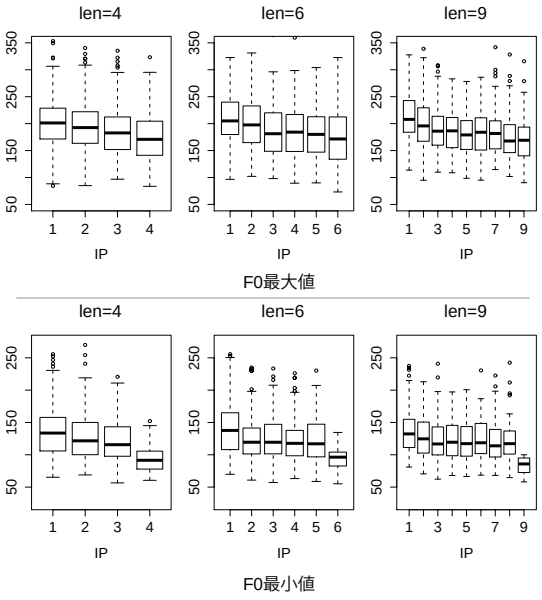


図 4　男性の場合 (Hz)

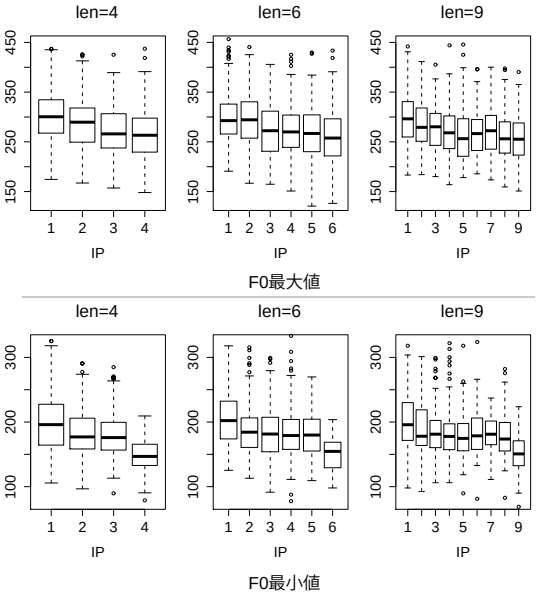


図 5　女性の場合 (Hz)

本分析では IP を分析の単位として見ているため，この下降はダウンステップではない。仮に declination の効果であるならば，declination は節単位の長さに関わらず生じるということになる。また，節単位最後の IP の F0 最小値（図 3 F0 最小値の各図の一番右）がその前の IP と比べてより大きく下降するという傾向が節単位の長さに関わらず観察されるが，これは final lowering の効果と考えられる。以上二つの傾向は，AP を対象に分析した Maekawa(2009)・前川 (2011) でも報告されており，その妥当性が改めて確認されたことになる。

ここで，図 3 を細かく見ると，節単位冒頭（特に接続詞の影響のない 2 番目の IP）の F0 最大値と節単位末尾の F0 最小値が，節単位の長さに関わらずほぼ一定であるという点に気付く。つまり，短い発話をするにせよ長い発話をするにせよ，ほぼ同じ高さから始まりほぼ同じ高さで終わるということである。この点を確認するために，節単位中の IP 数が 2〜13（いずれも頻度 50 以上）の範囲で，節単位頭から 2 番目の IP の F0 最大値と節単位末尾の F0 最小値だけをまとめてプロットした（図 6)。節単位最初の IP ではなく 2 番目の IP を比較するのは接続詞の影響を除くためである。図 6 を見ると，確かに節単位末尾の F0 値はほぼ一定であるのに対し，節単位 2 番目の F0 最大値は節単位中の IP 数が増えるにつれてわずかであるが高くなる傾向が見られる[*2]。

前川 (2011) は，節単位末尾の AP のピッチレンジがほぼ一定範囲に納まる傾向が見られることから，発話長を考慮に入れた F0 制御が行われている可能性があることを指摘している。上記の節単位末尾の F0 最小値がほぼ一定であるという結果は，前川の節単位末尾の AP のピッチレンジの結果とほぼ同じことを指しているが，今回更に，節単位の IP 数の増加に伴い節単位冒頭の F0 最大値が徐々に高くなる傾向にあることが明らかになった点は興味深い。というのも，発話長を考慮し，発話末尾の F0 最小値に向けて発話の途中途中で F0 制御をするというだけでなく，長い発話をする場合には発話の冒頭時点で既に少し高めに発話を開始するといった調整が行われている可能性があるためである。またこの結果は，発話のピッチレンジの下限がかなり固定的であるのに対し，上限は若干幅があるということも示唆する。

ところで図 3 を詳細に見ると，実は節単位冒頭からの IP 数が 5 を越えたあたりで F0 最大値・最小値ともに下降が停滞し，それより長く節単位が続く場合，F0 は一旦上昇した上で下降する傾向にあることに気付く。ただし，上昇に転じる位置（途中で下がり切った位置）の F0 最小値は節単位末の F0 最小値ほど低くはならず，また上昇ののち再び下降に転じる位置（途中で上がり切った位置）も節単位冒頭の F0 最大値ほど高くはならない。つまり発話がかなり長い場合，少し高い位置から発話を開始し下降の傾きを小さくするといった調整をしたとしても，それには限界があり，F0 の下限に達する前に一度 F0 を上昇させ，発話の立て直しをしている可能性があるということであろう。勿論この位置で final lowering は観察されない。

3.2 分析 2

本節では，一つの発話の中に強い統語的境界（強境界）が少なくとも一つ存在するケースを対象に分析を行う。具体的には，強境界の節単位と絶対境界の節単位の二つの連鎖を分析対象とする（図 1)。ここでの着目点は，発話内部に強い統語境界が置かれる位置において，前節

[*2] 数が少ないため明確なことは言えないが，IP 数が 14 以上の場合，このまま上昇し続けることはないようである。

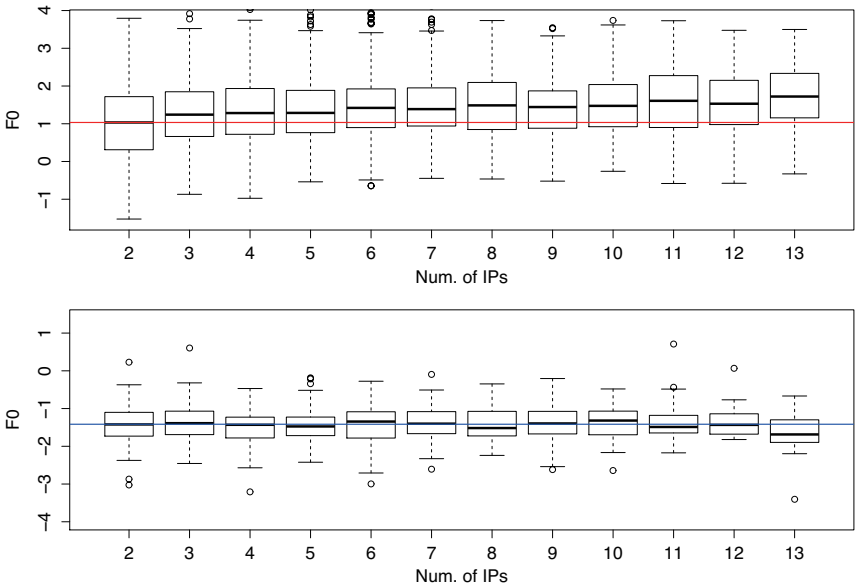


図 6　節単位冒頭・末尾の F0: (上) 節単位 2 番目の IP の F0 最大値，(下) 節単位末尾の IP の F0 最小値

表 2　先行節単位（強境界）と後続節単位（絶対境界）の IP 数とその頻度

		絶対境界の IP 数									
		1	2	3	4	5	6	7	8	9	10〜
強境界の IP 数	1	75	84	65	51	25	20	22	8	2	18
	2	127	136	148	86	58	36	35	19	10	24
	3	116	164	126	98	64	33	33	18	12	31
	4	105	152	107	71	54	46	29	31	17	18
	5	90	121	93	105	49	36	15	21	14	26
	6	76	82	66	67	63	28	22	18	4	27
	7	65	88	45	66	28	28	22	13	5	26
	8	33	45	38	46	43	31	22	10	17	16
	9	40	41	31	20	18	8	8	6	2	10
	10〜	87	124	91	81	65	44	46	26	26	55

で見たような F0 の下降が継続して見られるかどうかということである。表 2 に，先行節単位（強境界）と後続節単位（絶対境界）に含まれる IP 数ごとの事例数を示す。

前節と同じ分析を，先行節単位と後続節単位に含まれる IP 数ごとに行った。組合せが多いため，典型的な事例に限定して IP の F0 最大値と最小値の結果を図 7 と図 8 に示す。強境界 (S) と絶対境界 (A) の位置を縦線で示している。

図 7 と図 8 から，先行節単位と後続節単位の範囲でそれぞれ下降の傾向は見られるのに対し，強境界（S）で下降がリセットされていることが分かる。また後続節単位の末尾，つまり絶対境界（A）の位置の F0 最小値が，いずれのケースでもおよそ-1.5 にまで急激に下がっており（図 7・8（下）），final lowering が確認できる。またこの値は分析 1 の図 6（下）で確認したものと一致する。それに対し先行節単位の末尾である強境界（S）では，-1 程度にまでしか下がっておらず，final lowering も見られない。仮に-1.5 が「発話」のレンジの下限であるとするならば，強境界の末尾ではその下限にまで落ち切らないところで（final lowering を伴わず）リ

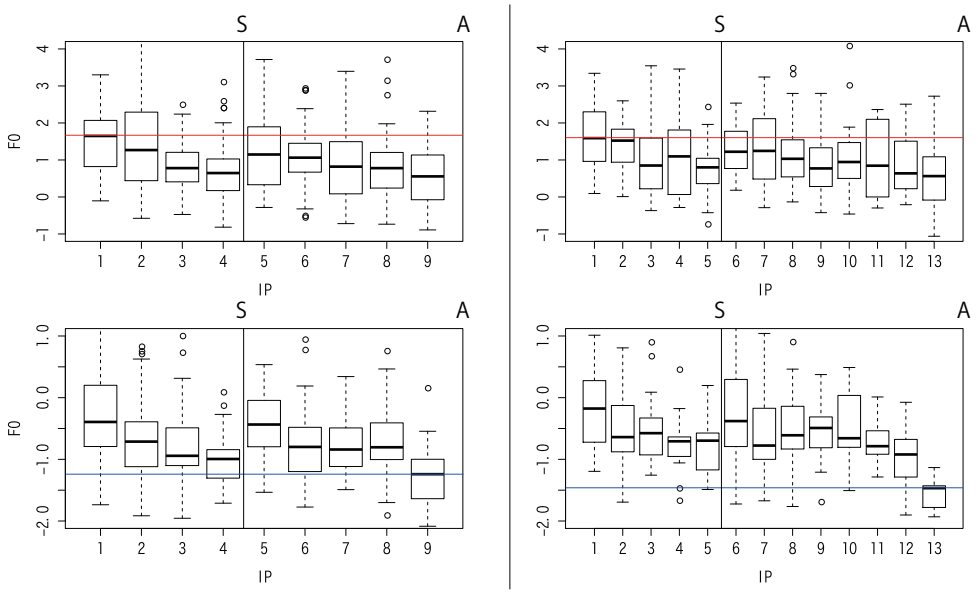


図 7　強境界後の F0 上昇幅が小さい例: (上)F0 最大値，(下)F0 最小値

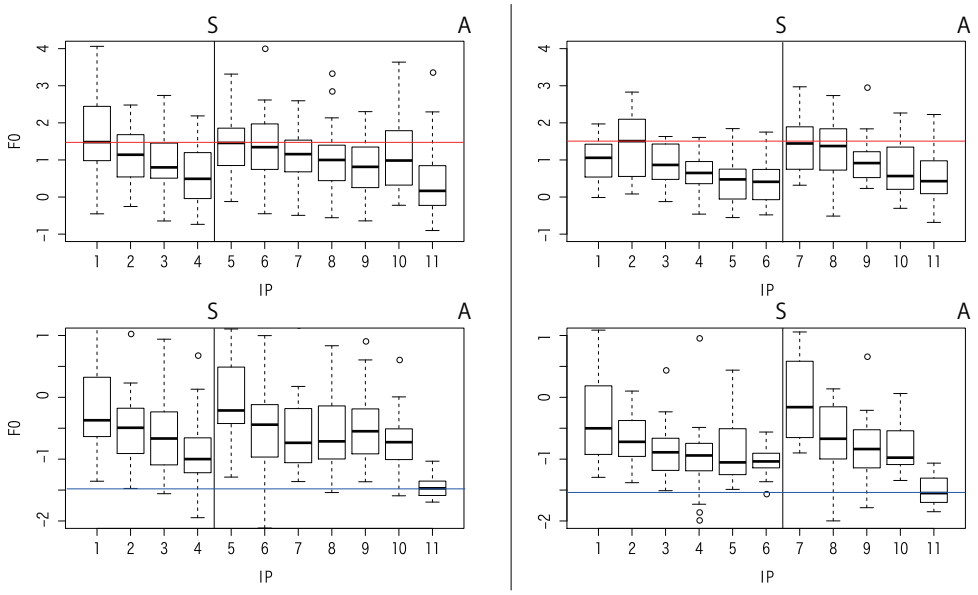


図 8　強境界後の F0 上昇幅が大きい例: (上)F0 最大値，(下)F0 最小値

セットしているということである。

仮に，ここで対象とする強境界が概ね「発話」の途中であり，節単位全般に見られる下降現象が declination（だけ）の効果であり，declination が「発話」全体に見られる現象であるとするならば，強境界の位置で下降は継続することになるが，この位置でリセットされる傾向が見られるということは，この仮定のいずれかに問題があるということになる。勿論，強境界が「発話」末になることがないわけではないが，平均値としてここまで強いリセットが見られることにはならないだろう。また「発話」の末尾に観察される final lowering が後続節単位の絶対境界

にのみ見られるという事実からも，強境界の仮定（のみ）が問題ということはなさそうである。

単純に考えるならば，「発話」の内部であっても強い統語境界があると declination がリセットされるという可能性が指摘できる。強境界の位置では final lowering を伴わないことから，final lowering の生起範囲と declination の範囲が必ずしも一致しないことになる。また declination は本来，発話に要する時間の関数として単純に F0 が低下する現象であるのに対し，本分析で見た下降も Maekawa(2009)・前川 (2011) で観察された下降も，いずれも発話長に応じて下降の傾きが変化するというものであり，declination だけで説明することは難しい（Maekawa 2009）という点にも注意する必要がある。つまりここで見られる下降には declination 以外の要因も関わっており，その別の要因が強い統語境界の範囲に影響する可能性もありうるということである。

なお分析 1 で対象とした，「発話」内部に絶対境界など強い統語境界を持たないケースでも，IP 数が 5 を越えたあたりで F0 の下降の停滞（場合によってはリセット）が観察されることを指摘した。同じ傾向がここで対象とした節単位（特に後続節単位）にも観察される。実は節単位の内部であっても，例えば「ここはちょっとデーターが足りないので（アノ）欠損値になっておりますが」の「ので」の位置のように，いわゆる「弱境界」に相当する統語的な切れ目が存在することもあり，これが丁度 5 番目の IP 前後に置かれる傾向にある，という可能性も存在する。今後，弱境界まで含めた詳細な分析を行う必要がある。

最後に，図 7 と図 8 の結果をもう少し細かく見ておこう。上述の通り，強境界の F0 最小値が-1 程度のところで上昇に転じるという点で共通した傾向を示すのに対し，後続節単位冒頭の F0 最大値は，ここに示していない組合せも含め二つのケースに分類される。一つは図 7 に示すパターンで，先行節単位冒頭の F0 最大値と比較して後続節単位冒頭の F0 最大値があまり高くならない（先行節単位末からの上昇幅が小さい）もの，もう一つは図 8 に示すパターンで，先行節単位冒頭の F0 最大値と比較して後続節単位冒頭の F0 最大値がほぼ同じ高さにまで戻る（先行節単位末からの上昇幅が大きい）ものである。この上昇幅が何によって決まるのかは分からないが，「発話」の途中であってもレンジの上限に戻ることがありうるということであり，「発話」の途中で下限には達しないということと対称的である。

4. おわりに

本稿では，F0 の下降現象に着目し，「発話」の長さや統語構造との関係について分析した。その結果，分析 1 では，「発話」のレンジがある程度固定されておりその範囲で全体的に F0 が下降する傾向にあること，「発話」の長さに応じて下降の傾きが変わる傾向にあること，「発話」が長い場合には若干高い位置で発話を開始する傾向にあることなどが分かった。また分析 2 では，「発話」の内部に強い統語境界があると下限に達する前に F0 の下降がリセットされる傾向にあること，その場合でも final lowering は「発話」末にしか現れないことが分かった。今後，統語的な条件を細かく設定した上で F0 との関係を見ていく必要がある。またポーズなど統語以外の要因についても考慮する必要がある。

参 考 文 献

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 (2006) 「概説」『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. 1–21.

Kikuo Maekawa(2009) “Contributions to corpus phonetics”’(国語研究所 NINJAL サロン).

前川喜久雄 (2011) 『コーパスを利用した自発音声の研究』(博士論文).

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 and Beckman(1988) *Japanese Tone Structure*, Cambridge: MIT Press.

大石初太郎 1971. 『話しことば論』秀英出版.

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), 347–453.

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』

※ 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。

コーパス管理ツール「茶器」による中古和文コーパスの利用

小木曽 智信（国立国語研究所言語資源研究系）†

Application of the Corpus Management Tool ChaKi to the Annotated Corpus of Early Middle Japanese

Toshinobu Ogiso (Dept. Corpus Studies, NINJAL)

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）の完成を受けて、新たに歴史的な日本語のコーパスの構築が進められつつある。国立国語研究所では「通時コーパスの設計」プロジェクト（近藤泰弘リーダー）を中心に構築のための研究が始まっている。こうした歴史的なコーパスのテキストを解析するために、古文のための形態素解析も研究開発が進められており、古典文学作品のテキストについては、実用的な精度で解析を行うことが可能になっている。

しかし、これまでは形態素解析済みの古文のデータの利用が十分に進んでいなかった。その理由の一つに、形態素解析結果を研究者が利用するツールが整っていなかったことがある。古典語は現代語と比較して利用可能なテキスト量が少ないため、研究のためには貴重なテキストに対して高い精度でタグ付けを行う必要がある。そのためには、自動処理に加えて、人手によるタグ付けが柔軟に行えるツールが必要である。また、ツールは多くの文系研究者が利用可能なように、手軽にパソコンにインストールして利用できるものである必要がある。コーパスを用いた日本語の歴史研究の発展のためには、日本語学の研究者が容易に使うことのできるコーパス利用ツールが求められている。

このようなニーズを満たすものとして奈良先端科学技術大学院大学で開発された「茶器」がある。本発表はこの「茶器」に形態素解析を施した中古和文のコーパスを格納し、研究への応用を試みるものである。「茶器」の高度な検索や統計的処理の機能を用いることで、これまでには行えなかった視点からの古典語研究が可能になると思われる。

２．「中古和文 UniDic」

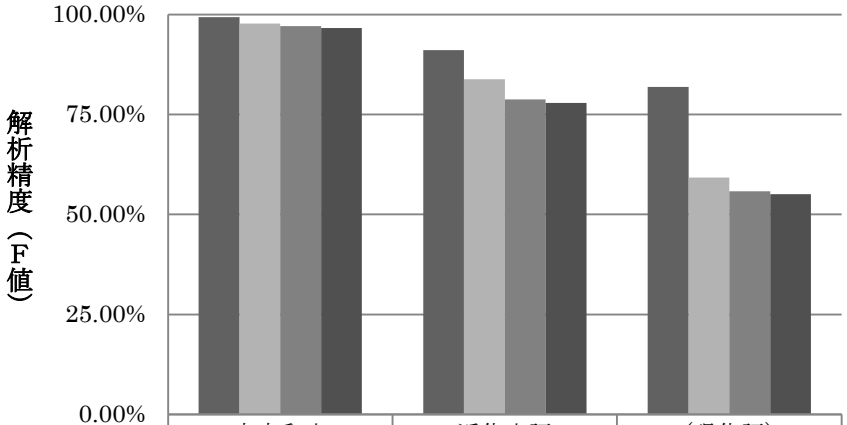
日本語の自動形態素解析は 1990 年代後半から実用化が進んだ。特に ChaSen 以降の機械学習に基づく形態素解析技術は、人手でのルール整備を不要にし、辞書と学習用のコーパスをもとにして高精度の解析を行うことを可能にした。しかし、古典文学作品などの歴史的な資料については，様々な先駆的試みがあったものの、本格的な古語の電子化辞書と機械学習用の古文のデータが不足していたため、実用的な精度で実現することは長らくできなかった。

こうした中、発表者らは歴史的な日本語コーパスの構築に備えるために歴史的な日本語資料を対象とした形態素解析辞書の開発を進めてきた（小木曽ほか 2010）。BCCWJ のタグ付けを行うために開発された形態素解析辞書「UniDic」をベースとし、さまざまな古文の解析に必要な見出し語を追加し、当該古文の機械学習用コーパスを整備することで古文のための形態素解析辞書を実現したものである。現在、近代の文語論説文を対象にした「近代文語 UniDic」と、中古の和文系資料を対象とした「中古和文 UniDic」を作成、公開している。UniDic はもともと言語研究に利用することを念頭に設計されており、短単位という斉一な解析単位、階層化され目的に応じて利用可能な階層化された見出し語を特長としている。

「中古和文UniDic」は、『源氏物語』をはじめとする平安時代の仮名文学作品を主たる対

† togiso@ninjal.ac.jp

象とした辞書である。図 1 は、文単位でランダムサンプリングした平安時代の仮名文学作品のテキストを、現代語用のUniDicと「近代文語UniDic」、「中古和文UniDic」のそれぞれで解析した精度を比較したものである[1]。グラフから分かるように、平安時代の仮名文学作品の形態素解析は、現代語用の辞書による解析では実用にならないものだったが、「中古和文UniDic」を用いることで高い精度で行うことが可能となった。



	中古和文 UniDic0.5	近代文語 UniDic1.1	(現代語) UniDic1.3.12
■単位境界	99.31%	91.09%	81.86%
■品詞認定	97.77%	83.78%	59.25%
■語彙素認定	97.10%	78.73%	55.77%
■発音形認定	96.64%	77.89%	55.05%

図 1　中古和文の解析精度比較

仮名文学作品の文体は、古典文学としてその後の規範となり長い間使われ続けていくため、「中古和文 UniDic」は中世・近世の擬古文をはじめとするさまざまなテキストの解析にも利用できる。明治期の雅文、和歌の解析にも利用可能である。

図 1 に示したとおり、形態素解析の精度はおよそ 95～97%程度である。先述したとおり、限られた資料をもとに研究を進める必要がある古典語研究にとって、この解析精度は目的によっては必ずしも十分でなく、解析結果を修正してさらに精度を高める必要がある。しかし修正を行うためのツールがないため、十分な活用ができないという問題があった。

３．コーパス管理ツール「茶器」

「茶器」は、上述の問題点を解消することが可能な汎用コーパス管理ツールであり、次のような特徴を備えている。

「茶器」は，タグ付きコーパスの検索および管理を支援する目的で作成されたツールである．文字列，単語列，および，係り受け関係による検索機能を備えている．単語列による検索では，単語の表層形以外に，読み，品詞や活用形などの文法情報を指定して検索を行うことができる．係り受け関係による検索では，文節内の単語列の指定と文節間の係り受け関係を指定した文検索が可能である．また，コーパス内の単語の頻度や前後文脈における単語の頻度など，簡単な統計処理を行うことができる．茶器は，タグ付きコーパスを関係データベースシステム（MySQL を使用）に格納し，検索要求を記述し結果を表示するためのインタフェースを提供する．対象言語は，多言語を目指しており，

[1] テストデータは『伊勢物語』『源氏物語』『大和物語』『土佐日記』『紫式部日記』『更級日記』から抽出した約 2.5 万語。ただし未知語なし。

日本語，英語，中国語のデータを取り扱うことが可能である．

（「茶器」使用説明書 version 2.1）

多言語対応を目指していることもあり、茶器は Unicode に対応している（内部文字コードは UTF-16）。そのため、古文には現れるが一般的には使用頻度の低い文字であっても、容易に取り扱うことが可能である。

対応するデータ形式は、MeCab ないし ChaSen による形態素解析結果と、CaboCha による係り受け解析結果である。付属のデータインポート支援ツール「Text Formatter」を用いることで、容易にタグ付きデータをインポートして利用することができる。形態素解析辞書としては IPADIC と UniDic に対応しているため「中古和文 UniDic」で解析された古典語のコーパスも格納することが可能である。

「茶器」は、近年「ChaKi.NET」としてシステムが一新され、簡易なデータベースである SQLite に対応したことによって、いっそう利用のしやすさを増している。SQLite の可搬なデータベースファイルを利用することで、タグ付きの古典語コーパスを広く配布して、研究者のパソコンでローカルに利用することができるようになった。

４．「茶器」による中古和文コーパスの利用

４．１．中古和文コーパスのインポート

「中古和文 UniDic」と「茶器」により、日本語研究者が容易に形態素解析済みの古典語コーパスを利用する環境が整った。図 2 は「茶器」に形態素解析済みの『更級日記』をインポートし、タグ付けを行っている画面イメージである。画面上に配置された各種のパネルによって、コーパスの検索・集計・統計情報の取得から、コーパスの修正、新しいデータのインポートまで行うことができる。

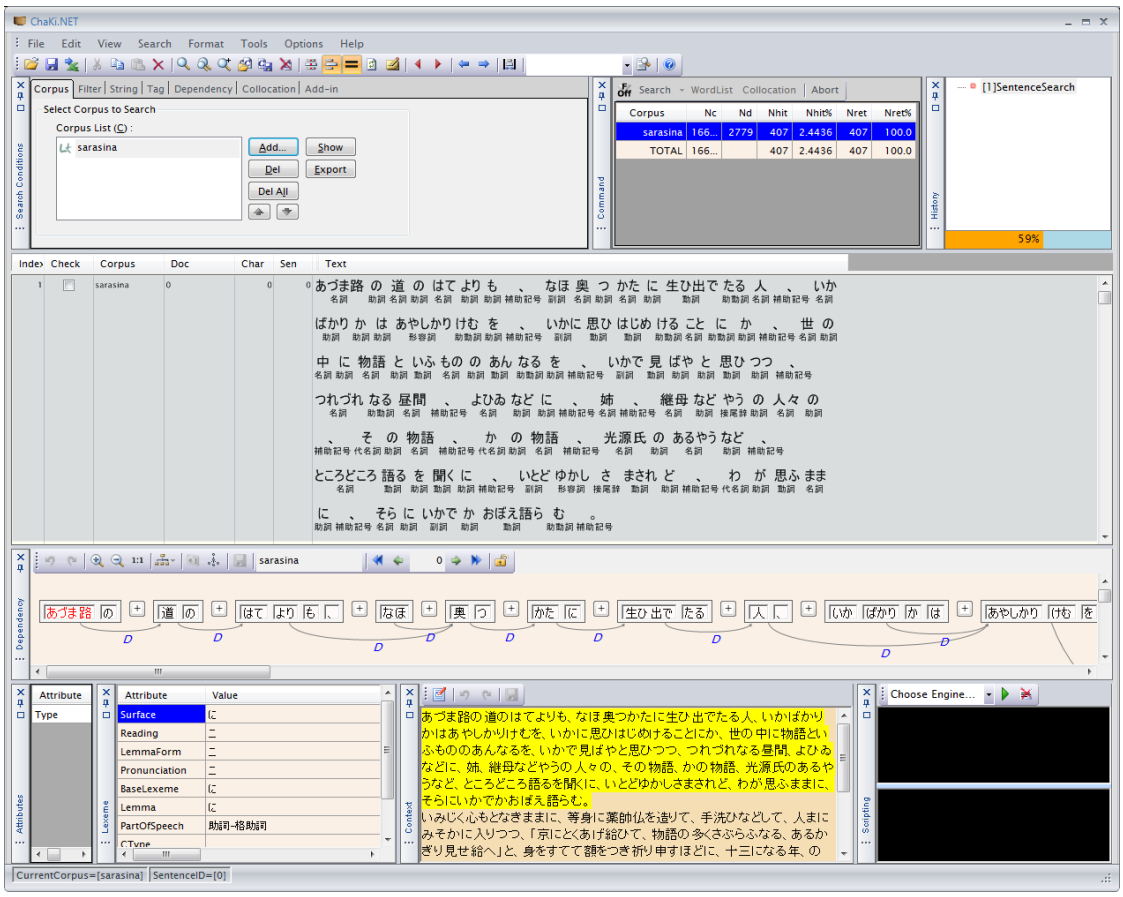


図 2　ChaKi.NET の実行画面（『更級日記』）

今回、表 1 にあるデータを「茶器」にインポートして利用した（以下、このデータを中古和文コーパスと呼ぶ[2]）。合計で約 58.8 万語になる。一部のデータについては、自動解析結果に対して人手による修正を加えている。

表 1　中古和文コーパスの作品一覧

作品名	語数	人手修正
伊勢物語	14624	済み
源氏物語（全）	528734	一部のみ
土佐日記	7948	済み
更級日記	16658	済み
紫式部日記	20353	一部のみ

現時点では、係り受けのタグ付けまで行った古典語のデータは存在しないため、今回用意したデータは、人手修正済みのものを含め、すべて形態素解析までしか行われていないものである。したがって、本来であれば「茶器」には形態素解析結果（*.mecab 形式）の取り込みしか行えない。

しかし、今回、これらのデータに対して実際に文節の係り受けをタグ付けを試行するために、次の簡単なルールに基づいて文節相当と考えられる部分をまとめ上げることにした。

1）助詞・助動詞は，直前の自立語または助詞・助動詞と結合する

2）接尾辞・句読点は直前の短単位と結合する

3）数詞の連続と数詞に続く助数詞は結合する

4）接頭辞は直後の短単位と結合する

このルールにより、簡易検査で 9 割以上の文節が正しく分割されることが確認された。このルールでまとめ上げた*.mecab 形式のファイルを、係り受け情報付きのデータ（*.cabocha 形式）に変換した後、TextFormatter を用いてインポートした。

4．2．形態論情報の格納

「茶器」は、形態論情報として、次の基本 9 属性を取り扱うことができる。括弧内は UniDic での対応する用語である（同一名称の場合は省略した）。

- Surface＝表層形（書字形）
- Reading＝読み（仮名形）
- LemmaForm＝（語彙素読み）
- Pronunciation＝発音（発音形）
- BaseLexeme＝基本形の表層形（書字形基本形）
- Lemma＝（語彙素）
- ParOfSpeech＝品詞
- CType＝活用型
- CForm＝活用形

[2] 「中古和文 UniDic」の作成と、形態素解析済み古典語データ利用の検証のために作成したもので公開予定はない。

UniDic は、語種やアクセント型などの多様な情報を付与することができ、その属性数は計 20 以上に上る。そのため、基本 9 属性に対応しない情報については、カスタムフィールド（custom）に格納して対処した。中古和文コーパスでは 9 属性以外の情報を次のようにカスタムフィールドに格納した。

custom="goshu 和 pronBase ムカシ kanaBase ムカシ formBase ムカシ"

中古和文コーパスのカスタムフィールド内のデータは次の通りである。

goshu＝語種（和語・漢語・外来語等）
pronBase＝発音形基本形
kanaBase＝仮名形基本形
formBase＝語形基本形

５．タグ付けツールとしての利用

５．１．形態素解析結果の修正

先述したとおり、研究に利用する古典語コーパスには、現代語以上に高いタグ付けの精度が求められる。しかし、自動形態素解析だけでその精度を実現するには困難であるため、自動解析結果を人手で修正する必要がある。

「茶器」を用いることで、コーパスのタグ付け・修正を行うことができるため、このような形態素解析の誤り修正のために利用することができる。修正用の辞書見出し語は、インポートしたコーパスから自動生成されているので、既出の語であれば正しい見出し語を選択するだけで解析結果の修正を行うことができる（図 3）。

図 3　解析結果修正のための辞書見出し語選択画面

５．２．文節係り受けのタグ付け

現状では係り受けまでタグ付けされた古典語のコーパスは存在しないが、古典語のコーパスへの係り受けのタグ付けが実現すれば、より高度なコーパス利用が可能になる。「茶器」を用いることで、古典語コーパスに対する係り受けのタグ付けを行うことができる。

図 4 は、「茶器」の文節係り受けのアノテーション画面（Dependency パネル）である。文節間のリンクのドラッグアンドドロップなどマウス操作だけで係り受けのタグ付けができるほか、文節の切り直しなども可能である。

文節係り受けのタグ付けは、文法判断に内省がきかない古典語の場合、極めて難しい作業となる。特に散文では一文が長く、係り先が曖昧な場合が多いため、完全なアノテーションを行うことは困難である。しかし、古典語のデータは量が限られているため、タグ付けする必要がある係り受けの種類を厳選し、十分な時間をかけることで、作品全体にタグ付けを行うことも可能だと思われる。

今回は、検証のためにタグ付けを試みたに過ぎないが、今後、係り受け情報付きの古典語コーパスの実現に向けて、連体修飾や動詞の項構造などに限定して、一部の作品へのタグ付けを進めていきたいと考えている。

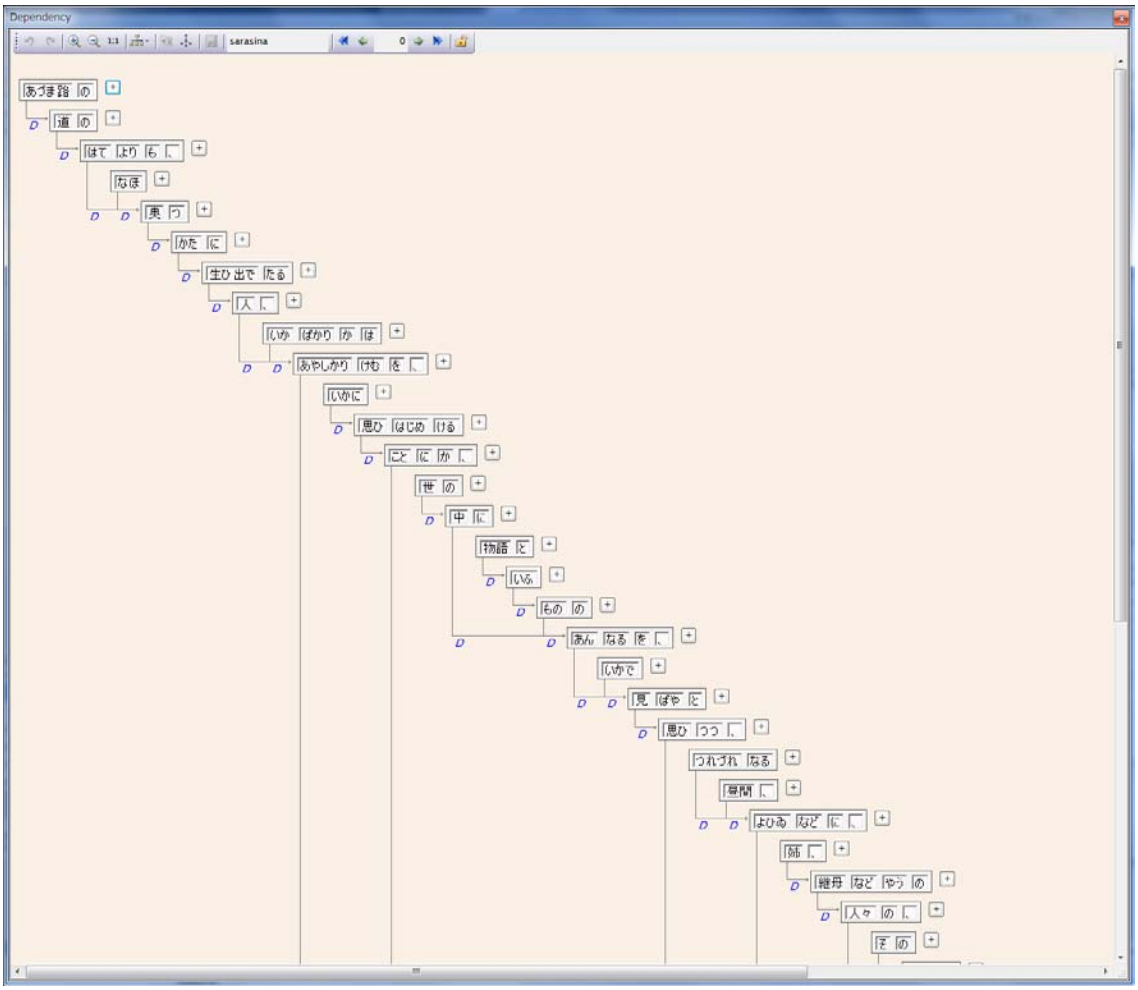


図 4　長い文の文節係り受けアノテーション（『更級日記』冒頭）

６．検索ツールとしての利用

「茶器」の検索機能によって、単純な文字列検索はもちろん、タグ付けされた形態論情報の品詞や活用型を検索キーとしたり、複数の検索語を組み合わせたりした自由度の高い高度な検索を行うことができる。検索結果は KWIC 形式で表示したり、外部にエクスポートしたりすることができる。さらにコロケーションや語彙頻度表などの統計情報を出力することが可能になっている。

電子化され、検索が可能な古典語データはこれまでにもあったが、「茶器」のような高度な検索や集計、統計情報の取得は行えなかった。古典語研究においてこのような処理が可能になったことで新たな発見が期待される。

ここでは、このような新しい利用の例として、正規表現を利用した文字列検索、タグ情報検索、共起検索、ワードリスト、コロケーション統計、そして係り受け検索について紹介する。

こうした処理では、データのサイズと処理速度が問題になるが、古典語のコーパスのサイズは限られているため、快適に利用することが可能である。今回用意した 58.8 万語のコーパスでは、以下の時間のかかる検索であってもすべて 10 秒以内に取得することができた（Windows7 x64, Core i7 2.53GHz, 8GB RAM の環境で確認）。

文字列検索（StringSearch）

「茶器」画面右上の検索条件指定パネル（SearchCondition パネル）で、さまざまな方法での検索を行うことができる。

文字列検索は中でももっとも単純なものだが、「茶器」では正規表現を利用した検索を行うことができる。一般にコーパス検索ツールでは使用できる正規表現に制限があることが多いが、「茶器」では Perl 5 互換の強力な正規表現が利用できる。

タグ検索（TagSearch）

タグ検索は形態素解析によって付与された語のタグ情報を利用して検索を行うものである。先述の 9 属性を自由に組み合わせて、検索に利用することができる。それぞれの項目で正規表現による指定が利用可能である。

UniDic の見出し語は、語彙素・語形・書字形・発音形の四つのレベルに階層化されているため、調査対象に合わせて選択することで、有効な検索ができる。

複数の語を組み合わせ、共起条件を複数設定した検索も可能である。図 5 の例では、動詞の後 2 語以内に助動詞「つ」が来る例を検索している。

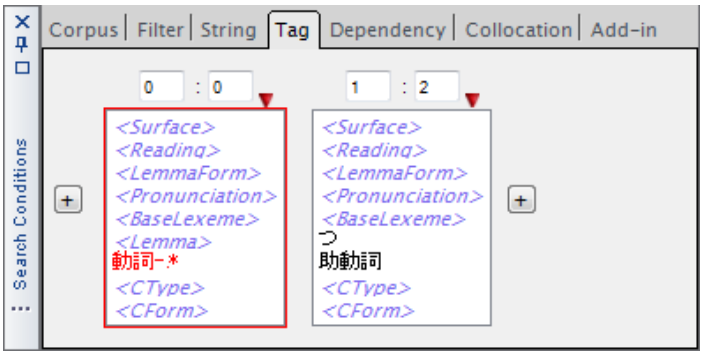


図 5　タグ検索（共起条件の設定）

ワードリスト検索（WordList）

検索条件を指定した後に、コマンドパネルの WordList コマンドを利用することで、条件を満たした語の集計を行うことができる。

図 7 は図 5 の検索結果を元に、助動詞「つ」の前 2 語以内に来る動詞のワードリストを作り、頻度順並び替えたものである。

	Su	Re	Le	Pr	Ba	Lemma_0	Pa	C	Cf	Su	Re	Le	Pr	Ba	Le	Pa	C	Cf	genji	ise	murasa	sarasin	tosa	All	Ratio(%)
▸TOTAL																			1843	37	36	31	18	1965	100
1	*	*	*	*	*	諦める	*	*	*	*	*	*	*	*	*	*	*	*	259	5	2	0	0	266	13.536...
2	*	*	*	*	*	会う	*	*	*	*	*	*	*	*	*	*	*	*	105	0	0	0	0	105	5.3435...
3	*	*	*	*	*	有る	*	*	*	*	*	*	*	*	*	*	*	*	82	0	2	1	1	86	4.3765...
4	*	*	*	*	*	表わす	*	*	*	*	*	*	*	*	*	*	*	*	78	2	0	0	2	82	4.1730...
5	*	*	*	*	*	急ぐ	*	*	*	*	*	*	*	*	*	*	*	*	80	0	0	1	0	81	4.1221...
6	*	*	*	*	*	坐す	*	*	*	*	*	*	*	*	*	*	*	*	73	0	0	1	0	74	3.7659...
7	*	*	*	*	*	入り揉む	*	*	*	*	*	*	*	*	*	*	*	*	63	1	0	3	0	67	3.4096...
8	*	*	*	*	*	承る	*	*	*	*	*	*	*	*	*	*	*	*	62	0	1	0	0	63	3.2061...
9	*	*	*	*	*	置く	*	*	*	*	*	*	*	*	*	*	*	*	39	0	1	0	0	40	2.0356...
10	*	*	*	*	*	遅らす	*	*	*	*	*	*	*	*	*	*	*	*	25	2	0	1	0	28	1.4249...
11	*	*	*	*	*	怠る	*	*	*	*	*	*	*	*	*	*	*	*	24	0	0	0	0	24	1.2213...
12	*	*	*	*	*	押し立てる	*	*	*	*	*	*	*	*	*	*	*	*	20	0	0	1	1	22	1.1195...
13	*	*	*	*	*	大人びる	*	*	*	*	*	*	*	*	*	*	*	*	19	1	0	0	0	20	1.0178...
14	*	*	*	*	*	驚く	*	*	*	*	*	*	*	*	*	*	*	*	20	0	0	0	0	20	1.0178...
15	*	*	*	*	*	驚かす	*	*	*	*	*	*	*	*	*	*	*	*	19	0	0	1	0	20	1.0178...
16	*	*	*	*	*	おわす	*	*	*	*	*	*	*	*	*	*	*	*	13	1	0	5	0	19	0.9669...
17	*	*	*	*	*	おわします	*	*	*	*	*	*	*	*	*	*	*	*	17	0	0	0	0	17	0.8651...
18	*	*	*	*	*	覚える	*	*	*	*	*	*	*	*	*	*	*	*	14	1	0	1	0	16	0.8142...

図 6　ワードリスト検索結果

コロケーション

検索条件を指定して検索を行った後に、Collocation タブで設定することで、表示されている KWIC を対象にした各種の統計を取ることができる。取得できる情報は、粗頻度、MI スコア（相互情報量)、N-gram 頻度、FSM（Frequent Sequence Mining）である。

ただし、ここで計算に使われる頻度は、コーパス全体を対象としたものではなく、検索結果の KWIC に表示されているものが利用されるので注意が必要である。

係り受け検索（DependencySearch）

検索条件パネルの Dependency タブにより、文節の係り受け関係を条件に指定した検索を行うことができる。現在は古典語の係り受け解析は開発途上であるため、現時点では人手修正済みのわずかなデータしか利用できないが、将来的にはこの機能を用いることで、単に隣接しているだけではなく、係り受け関係にある語を検索することが可能になる。

7．おわりに

「茶器」を用いることで、懸案だった形態素解析済みの古典語データの利用環境を整備することができた。今後、単語情報付きの古典語コーパスの豊富な情報と、「茶器」の高度な検索機能や統計情報を用いて、新しい古典語研究が行われることに期待したい。

将来的には、「茶器」で整備した係り受けデータを元に古典語の係り受け解析を実現していきたい。係り受けまで整備されたコーパスが用意できれば、検索結果のいわゆる「ゴミ取り」作業が軽減できるだけでなく、動詞の項構造や連体修飾関係などを使う高度なコーパス利用が可能になる。これによりコーパスを利用した古典語研究がさらに大きく前進するはずである。

文　献

松本 裕治（2009）「コーパスへの自動アノテーションツールとアノテーション支援環境の構築」人工知能学会誌 24(5), pp.632-639

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-CH-85, No.4 pp.1-8

小椋秀樹・須永哲矢・小木曽智信・近藤明日子・田中牧郎（2011）「「中古和文 UniDic」における言語単位の設計」『言語処理学会第 17 回年次大会発表論文集』pp.312-315

小木曽智信・岡照晃・小町守・松本裕治（2011）「コーパス管理ツール「茶器」による単語情報付き古典語コーパスの活用」『人文科学とコンピュータシンポジウム論文集「デジタル・アーカイブ」再考』情報処理学会 pp. 255-260

関連 URL

茶器　http://sourceforge.jp/projects/chaki/

中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

MeCab　http://mecab.sourceforge.net/feature.html

CaboCha　http://code.google.com/p/cabocha/

国立国語研究所基幹型共同研究プロジェクト「通時コーパスの設計」
http://www.ninjal.ac.jp/research/project/a/corpus/

国立国語研究所萌芽・発掘型共同研究プロジェクト「統計と機械学習による日本語史研究」
http://www.ninjal.ac.jp/histlingstat/

大規模コーパスを用いた用例の典型性評価
—大規模コーパスを利用した学習辞書作成のために—

千葉 庄寿（麗澤大学外国語学部）†

How to Evaluate the Typicality of the Corpus Evidence

Shoju CHIBA (Faculty of Foreign Studies, Reitaku University)

１．はじめに

よい辞書にとって，用例は不可欠の要素であることは言を俟たない。用例が提供する文脈情報の重要性はつとに多くの辞書編纂者が述べていることである(Sinclair 1991, Fox 1987)。一方で，辞書には記述スペースの問題があり，全ての用例を辞書に掲載することは難しい。掲載する用例を選択する，ないし適切な用例を編纂者が作成することは，辞書作成作業行程の重要な部分を占めることになるわけである。

用例を吟味する際，その用例をどんな人がどんな目的で参照するかを想定することは重要である。(i) 辞書を使用する人が母語話者かどうか，また (ii) 参照するのが文の理解のためか，それとも文の産出のためかによって，用例に持たせるべき情報の重みは異なるだろう。母語話者にとっては，用例はもっぱら理解，学習者にとっては産出の支援が重要になると思われるが，それに限られるわけではない。

学習者用英語辞典における長い伝統をもつ英語の辞書編纂において，用例にはさまざまな教育的配慮が求められてきた。A. S. Hornby による *Oxford Advanced Learners' Dictionary* 以来，用例は文のテンプレートとしての機能を意識して作られている。一方，Rundell (1998) が指摘するように，このような教育的配慮の結果，しばしば一つの例文に情報を盛り込みすぎ例文が不自然になってしまうこともあった。

学習者むけの日本語辞書において，用例の吟味に関する方針が深く議論されたことはこれまであまりない。優れた国語辞典である『明鏡国語辞典』には，用例の編纂方針として「日常生活で頻繁に用いられる重要語には特に用例を多く載せる」とのみ示されている(北原 2010[2]: ix)。

編集ポリシーとして，用例を実例からとるか，辞書編集者の手による作例を用いるのがよいのかは未だに意見が分かれるところである(Fox 1987, Sinclair 1991 *vs.* Laufer 1992)。しかし，学習辞書の編集にコーパスデータを援用することの意義は，今日の英語学習辞書においてコーパスを使った用例収集が既に標準的な作業となっていることを見るまでもなく明らかであり，事実上「どのようなプロセスで用例に選ばれるのかが異なる」(Rundell 1998) だけである。また，コーパスに基づく辞書であることを謳う学習辞書において，教育的配慮から，コーパスからえられた実例を編集し簡略化することもしばしばあるが，*Collins* 社の *Cobuild* シリーズの辞書のように実例そのままの提示にこだわる辞書も存在する。

† schiba@reitaku-u.ac.jp

２．重要度の認定基準

日本語学習者によって有益な例文とは何であろうか？Fox (1987)は例文を以下のような基準で評価する。

(1) 典型的であること
(2) 自然であること：“the concept of naturalness […] is the well-formedness of sentences not in isolation but in text.” (Fox 1987: 139)
(3) 示唆に富むこと
(4) 典型的な文脈とともに現れること。簡略を期して用例を短縮・修正することは望ましくない (*cf.* Fox 1987:147-149)

また，辞書の他の部品との関係として，

(5) 用例が語義の解説と連動する必要はなく，むしろ用例を語義とは独立して読み，理解することができることが重要である

ことも比較的よく知られている。

では，このような用例の重要度を測定し，大規模なコーパスからできる限り自動で取り出すために有益な情報は何だろうか？上記(2), (4)はコーパスから用例を取得することである程度保証される。一方，(3)は集まった用例を最終的に吟味する段階で確認すべきものと考えられる。本稿では，(5)の語義との連動とともに，(3)については辞書編纂者の手に委ねることを想定し，扱わないことにする。こうして，1 の「典型性」の評価のみが，特に吟味が必要な項目として残ることになる。

本稿では，用例の典型性について，以下の多様なパラメータ(6a-6c, 7)を使い，総合的に評価することを試みる。

(6) 頻度情報：
  a. 典型的な統語パターン：形態統語的情報，連鎖，統語構造
  b. 典型的なコロケーションパターン : 共起頻度とその補正値(ダイス係数, MI スコア)
  c. 典型的な文脈：文のタイプ，ジャンル情報(サブコーパス間の分布)
(7) 比較対象となるコーパスの分布からの逸脱 : BCCWJ のような大規模コーパスから取得した用例を BCCWJ 全体，さらにはサブコーパス間で比較する(対数尤度比など，サイズに影響されにくい指標を用いる, *cf.* 千葉 2011)

本稿では，以下，上述の典型性を測るための頻度情報の一部 [1] と分布の偏りの情報を使い，用例のレポジトリとして大規模なコーパスがどの程度用例の典型性評価に役立つかを示す。また，事例分析を通じて，大規模コーパスを辞書編纂に本格的に活用するための基盤構築の必要性を主張する。

[1] このうち，今回は統語構造および文のタイプに関する考察はおこなわない。

３．事例分析

本節では以下の 6 種類の語彙パターンを考察する。

(8) 動詞「生きる」(動詞-普通，上一段-カ行)
(9) 動詞「足りる」(動詞-普通，上一段-ラ行)
(10) 名詞「信用」(名詞-普通名詞-サ変可能)
(11) 名詞「信頼」(名詞-普通名詞-サ変可能)
(12) 複合助詞「かも」(助詞-副助詞＋助詞-係助詞)
(13) 派生使役動詞「動詞＋せる」(未然形＋助動詞，下一段-サ行)

このうち，(8)〜(11)は特定領域研究「日本語コーパス」言語政策班が作成した BCCWJ の語彙表 [2] において，LB_FL (図書館), PB (書籍), PM (雑誌), PN (新聞), OC (知恵袋), OY (ブログ) の 6 つのサブコーパスのうち 5-6 のサブコーパスにおいてカバー率 (累積頻度) がレベル a (0〜78%) に分類される，いずれも出現頻度が高いものである。(12)は助詞のような文法要素の複合の例，(13)は生産性の高い派生動詞の例である。

今回はテストケースとして，解析ずみの小規模なコーパスを使用して分析をおこなう。データ班が構築した「現代日本語書き言葉均衡コーパス・コアデータ」(特定領域研究「日本語コーパス」研究成果報告 DVD (JC-G-10-03 所収)) を用いる。総形態素数約 130 万語，総文数約 5 万 6 千文という小さなデータであるが，本稿が目指す方向性の出発点としては有効であると考える。

３.１　サブコーパス間の分布

以下に，今回の事例として選んだ 6 つの語彙のサブコーパス間の分布を示す。core 用例数は「現代日本語書き言葉均衡コーパス・コアデータ」での用例数を，WPM は 100 万語あたりの出現頻度を表す。

表１：「生きる」(動詞-普通，上一段-カ行)

	PB (書籍)	PM (雑誌)	PN (新聞)	OC (知恵袋)	OW (白書)	OY (ブログ)	合計
文総数	9,247	11,654	15,672	6,301	5,830	7,272	55,976
形態素総数	234,431	239,877	360,825	110,696	228,272	118,305	1,292,406
形態素数/文	25.35	20.58	23.02	17.57	39.15	16.27	23.09
core 用例数	90	41	72	7	10	22	242
WPM	383.91	170.92	199.54	63.24	43.81	185.96	187.25

表２：「足りる」(動詞-普通，上一段-ラ行)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	10	7	5	5	1	4	32
WPM	42.66	29.18	13.86	45.17	4.38	33.81	24.76

[2] 語彙レベルの詳細については田中(2011)を参照。

表３：「信用」(名詞-普通名詞-サ変可能)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	8	8	17	5	4	5	47
WPM	34.13	33.35	47.11	45.17	17.52	42.26	36.37

*PM のみ語彙レベル b

表４：「信頼」(名詞-普通名詞-サ変可能)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	18	23	35	2	17	5	100
WPM	76.78	95.88	97.00	18.07	74.47	42.26	77.38

*OC のみ語彙レベル b

表５：「かも」(助詞-副助詞＋助詞-係助詞)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	122	101	46	70	0	75	414
WPM	520.41	421.05	127.49	632.36	0.00	633.95	320.33

表６：「動詞＋せる」(未然形＋助動詞，下一段-サ行)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	240	195	313	68	116	58	990
WPM	1,023.76	812.92	867.46	614.30	508.17	490.26	766.01

表から分かるように，平準化した数値をサブコーパス間で比較することで，調査語彙の中で偏りがある場合と，調査語彙間での違いが明らかになることである。例えば，類義語「信用」「信頼」を例にとると，「信用」で OW において出現数が明らかに低い(表３)のに対し，「信頼」では OC, OY など Web 系の使用域において出現数が下がっている(表４)。

さらに興味深いことに，サブコーパス間の出現比率の差は，調査語彙の文法的な抽象度が上がると小さくなる傾向があることである。使役動詞(表６)を参照されたい。

３.２　形態論的情報

以下に，今回の事例として選んだ 6 つの語彙のうち，活用を示す 2 つの動詞の活用形の分布を示す。出現する活用形の分布が大きく異なることが分かる。

表７：「生きる」(*f* = 242) の活用形

意志推量形	仮定形	終止形	未然形		命令形	連体形	連用形	総計
生きよう	生きれ	生きる	生か	生き	生け	生きる	生き	
1	1	13	1	8	3 *	**42**	**173**	242

* 「生きとし**生け**るもの」(1 例); 「**生け**る化石」(2 例)

表８：「足りる」(*f* = 32) の活用形

終止形	未然形	連体形	連用形	総計
足りる	足り	足りる	足り	
1	**29**	1	1	32

３.３　シンタグム (3-gram)

コロケーションは，用例の頻度情報として特に有益である(Sinclair 1991)。以下に，文法的要素としての性格が強い「かも」と「動詞＋せる」の特徴をよく表す連鎖の頻度を示す。

表９：「かも」(f = 414) に後続する連鎖 (1 形態素)

lemma1	lemma2	lemma3	POS3	頻度
か	も	知れる	動詞	319
か	も	。	補助記号	26
か	も	？	補助記号	11
か	も	・	補助記号	7
か	も	ね	助詞	6
か	も	…	補助記号	5
か	も	、	補助記号	5
か	も	」	補助記号	3
か	も	分かる	動詞	3
か	も	！	補助記号	3

* 頻度 3 未満は省略する。

表１０：「(か)も」に後続する連鎖 (2 形態素)

lemma1	lemma2	POS2	lemma3	POS3	頻度
も	知れる	動詞	ない	助動詞	196
も	知れる	動詞	ます	助動詞	116
も	。	補助記号	＃	文境界	24
も	・	補助記号	・	補助記号	7
も	知れる	動詞	ず	助動詞	7
も	？	補助記号	＃	文境界	7
も	ね	助詞	。	補助記号	5
も	。	補助記号	。	補助記号	2
も	」	補助記号	と	助詞	2
も	分かる	動詞	ない	助動詞	2
も	！	補助記号	？	補助記号	2

「かも」については，「知れない」といった否定表現への接続への偏りとともに，「かも」で打ち切りの形で終わる文が多いことが明らかになる。

次に，使役動詞の主動詞となる要素の出現傾向を見る(表１１)。「する」「聞く」といった出現頻度の高い動詞が出ている一方，実に多様な動詞が「せる」をとり出現していることが分かる。低頻度で数多くの種類の動詞が出現することは，この接辞の生産性の高さを示す証拠といえる(図 1, *cf.* Baayen 2001)。このような，生産性の高さに起因する多様な候補ある中で，どのような用例がより典型的と言えるかは，単純な頻度では判断することができない。後述のような，分布の偏り自体を考慮することが望ましい。

表１１：「動詞＋せる」($f = 990$) に現れる動詞

lemma1	POS1	lemma2	頻度	累積
為る	非自立可能	せる	546	546
聞く	一般	せる	19	565
済む	一般	せる	16	581
知る	一般	せる	14	595
思う	一般	せる	13	608
言う	一般	せる	12	620
持つ	一般	せる	12	632
遣る	非自立可能	せる	11	643
行う	一般	せる	9	652
楽しむ	一般	せる	9	661
走る	一般	せる	9	670
行く	非自立可能	せる	8	678
騒ぐ	一般	せる	7	685
膨らむ	一般	せる	7	692
負う	一般	せる	6	698
咲く	一般	せる	6	704
募る	一般	せる	6	710
驚く	一般	せる	5	715
思い出す	一般	せる	5	720
輝く	一般	せる	5	725
滑る	一般	せる	5	730
作る	一般	せる	5	735
飲む	一般	せる	5	740
光る	一般	せる	5	745
待つ	一般	せる	5	750
利く	一般	せる	4	754
困る	一般	せる	4	758
漂う	一般	せる	4	762
出す	非自立可能	せる	4	766
悩む	一般	せる	4	770
覗く	一般	せる	4	774
弾む	一般	せる	4	778
働く	一般	せる	4	782
振り込む	一般	せる	4	786
向かう	一般	せる	4	790
窺う	一般	せる	3	793
歌う	一般	せる	3	796
疑う	一般	せる	3	799
終わる	非自立可能	せる	3	802
通う	一般	せる	3	805
気付く	一般	せる	3	808
食う	一般	せる	3	811
死ぬ	一般	せる	3	814
ちらつく	一般	せる	3	817
尖る	一般	せる	3	820
取る	一般	せる	3	823
滲む	一般	せる	3	826
履く	一般	せる	3	829
挽く	一般	せる	3	832
含む	一般	せる	3	835
休む	一般	せる	3	838
喜ぶ	一般	せる	3	841
笑う	一般	せる	3	844

* 頻度 2 以下の動詞は省略する

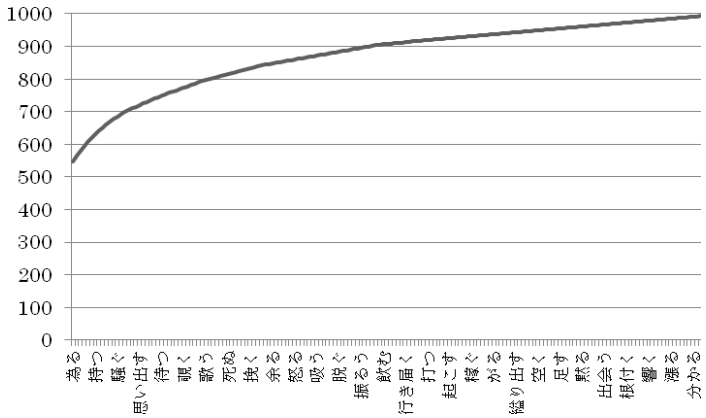


図１：「動詞＋せる」の累積頻度

一方，使役動詞の文末表現のパターンにははっきりと頻度の高い表現群があり，典型的な表現を，通常の動詞(§3.2 参照)と同様あるいはそれ以上に容易に見いだすことができる。テ形接続の一般的な表現パターンとよく似たパターンが現れる一方で，単純な頻度ではなく，後述の分布の偏り自体への考慮(§4 参照)を行うことにより，これらの中から「動詞＋せる」に特徴的な分布を見いだすことができる。

表１２：(動詞＋)「せる」(f = 990) に後続する 2 形態素

lemma1	lemma2	lemma3	POS3	頻度
せる	て	居る	動詞	68
せる	て	頂く	動詞	28
せる	て	呉れる	動詞	20
せる	て	下さる	動詞	15
せる	て	貰う	動詞	13
せる	て	、	補助記号	11
せる	て	仕舞う	動詞	9
せる	て	行く	動詞	9
せる	て	上げる	動詞	6
せる	て	見る	動詞	6
せる	て	欲しい	形容詞	4
せる	て	」	補助記号	3
せる	て	から	助詞	3
せる	て	は	助詞	3
せる	て	ます	助動詞	3
せる	て	来る	動詞	3
せる	て	遣る	動詞	3
せる	て	置く	動詞	2
せる	て	薄皮	名詞	2
せる	て	居る	動詞	2

３.４　用例の特徴を観察・分析するのに有効な情報

これまであきらかになった，典型性を測る調査パラメータと調査語彙との親和性をまとめると表１３のようになる。

表１３：分析パラメータと調査語彙との親和性

	ジャンル	語形	N-gram	語彙頻度	統語構造	係り受け
生きる		◎				
足りる		◎				
信用						
信頼	△					
かも	◎		◎			
動詞＋せる			◎			

「信用」は数十例と得られた用例数が比較的少ないため，どの情報も充分な特徴を観察するものではなかった。「かも」「動詞＋せる」は比較的用例数が多く，容易に特徴を見いだすことができた。

一方，充分な用例数が得られなくとも特徴が判別できる「足りる」のような例もある。これは「足りる」が否定極性に親和性が高いことが，日本語の場合形態素の語形から判断できることが大きいと考えられる。

今後，今回調査の対象から外した統語構造・係り受け構造および文体情報についても引き続き調査を続けていきたい。

４．語彙情報プロファイリングによる分布の分析

千葉(2011)は BCCWJ を比較のサンプルとして他のコーパスの語彙情報を評価する「語彙情報プロファイリング」の手法を開発した。BCCWJ の正式版に基づき，現在オンラインで分析できるプロファイリングシステムを公開準備中である。このシステムを用いて用例の集合の分析を試みる。

語彙情報プロファイリングをおこなうことで，コーパスから特定の語彙パターンを含む用例(コーパスのサブセット)を取得し，その例文集合と BCCWJ (ないしそのサブコーパス)の語彙情報との比較をおこなうことができる。ここではサイズの異なるコーパスの頻度情報を比較するのに有効と考えられている対数尤度比(LLR, 内山ほか 2004)を使用し，bigram (2 形態素の連鎖)の比較をおこなうことにする。

表１４：信用 (f = 47) の bigram の比較

lemma1	POS1	lemma2	POS2	頻度	BCCWJ	LLR *
信用	名詞	為る	動詞	17	494	203.573647
を	助詞	信用	名詞	11	197	142.096939
信用	名詞	出来る	動詞	5	222	55.66316
が	助詞	信用	名詞	4	73	51.508351
の	助詞	信用	名詞	5	357	50.96311
信用	名詞	の	助詞	3	80	36.405926
枠組み	名詞	合意	名詞	2	2	35.928777
た	助動詞	有権	名詞	2	5	33.098369
信用	名詞	を	助詞	3	171	31.912607

* 内山ほか (2004) による補正をおこない用例の特徴のみを正の数値として算出する

用例の集合には当然ながら「信用」と「信頼」という語彙が必ず含まれるため，全体として調査語彙が含まれる連鎖の LLR は高く出る。表１４から，47 例中 17 例(36.2%)が「信用する」という形で，5 例が「信用できる」(10.6%)という連鎖で出現したことがわかる。

一方，表１５で示すように「信頼」の場合，「信頼する」の連鎖は 100 例中 12 例(12%)で，「信用」とは明らかに出現パターンが異なる(LLR 自体は 113.5 と高い)。

表１５：信頼 (f = 100) の bigram の比較

lemma1	POS1	lemma2	POS2	頻度	BCCWJ	LLR *
の	助詞	信頼	名詞	25	793	255.309093
信頼	名詞	を	助詞	21	406	234.783454
信頼	名詞	出来る	動詞	13	241	146.376739
信頼	名詞	が	助詞	10	99	124.676046
信頼	名詞	性	接尾辞	12	441	119.033493
信頼	名詞	為る	動詞	12	557	113.512011
は	助詞	信頼	名詞	7	79	85.516645
信頼	名詞	関係	名詞	8	347	76.735319
信頼	名詞	回復	名詞	5	43	63.666965
科学	名詞	者	接尾辞	6	403	52.369004
で	助詞	信頼	名詞	4	43	49.23556
出来る	動詞	医者	名詞	3	6	45.9852
弾道	名詞	ミサイル	名詞	4	76	44.838018
対する	動詞	信頼	名詞	4	100	42.694878
信頼	名詞	感	名詞	4	107	42.164787

信頼	名詞	の	助詞	4	117	41.4639
関係	名詞	を	助詞	8	3274	41.394149
安全	名詞	で	助詞	3	17	40.535622
た	助動詞	信頼	名詞	3	21	39.359574
者	接尾辞	等	接尾辞	6	1643	35.744162
ポイント	接尾辞	減	名詞	3	40	35.686003
性	接尾辞	の	助詞	8	4818	35.433254

今後, 語彙情報プロファイリングシステムに trigram (3 形態素の連鎖)以上の長さの比較の機能を実装させることで，このような違いはさらに明確に分析抽出できるようになると思われる。

分析対象のコーパスの部分集合である用例リストの語彙情報を分析する意義は，当該語彙の特徴的な組合せを大規模なコーパスの出現比率の観点から判別できることであるが，同時に，用例集合に含まれる分析対象の語彙とは異なる語彙についても観察が可能であることを考えると非常に大きい。明らかに偏った用例集合のもつ LLR の意味，さらにはこのような手法を統計的にどのように位置づけ，典型性の判断に結びつけるかは今後さらに検討していきたい。

５．おわりに

本稿では，頻度情報と分布の偏りの情報を，典型性を測るための情報として用い，コーパスが用例の典型性評価にどの程度役立つかを具体例により検討した。今後，典型性を測るための頻度情報のパラメータを充実させるとともに，より大規模なコーパスデータを用いて検証をすすめ，大規模コーパスを辞書編纂に活用するための基盤の整備をおこなうとともに，辞書の記述とコーパス分析の方法論的な統合のための用例評価のしくみの整備，特に用例の典型性の指標化にむけた研究をすすめていきたい。

一方で，複合的な情報を如何にして組み合わせ，検索された用例の典型性，有用性を測るための指標に練り上げるか，という問題に関して，今後さらに考察する必要がある。特に，指標化はコーパスの中からの典型的用例の抽出にとどまらず，教師や辞書編纂者が作例した文を評価する，といった広い応用可能性をもつ研究成果となる可能性があり，実用化が待たれる。

今後の研究開発の方向性として，まず以下の作業をすすめることが挙げられる：

- 用例評価データベースのプロトタイプの公開
- 語彙レベル情報，統語構造（ごく浅い統語木）や係り受け情報の組み込みと評価
- コーパスの用例を評価するための諸情報の最適な組合せ方法と，指標化およびチューニング方法の検討
- 各種パラメータを総合し，指標化する用例の典型性評価データベースのプロトタイプの開発(*cf.* Kilgarriff *et al*. 2004)

一方，用例の典型性を超えた「よい例文」の判断には，コーパスの情報を使った用例の評価に加え，辞書編集者・教師による判断を併用することが必要になる可能性が高い。「有用」「重要」と辞書編纂者や教師が考える用例を登録し，学習データとして解析・蓄積する機能をもつ用例データベースの開発が望ましい。

また，辞書編纂においては語彙素が包摂する異表記と用例の出現パターンとの関係も自動処理をおこなっておくことが望ましい。例えば，『明鏡国語辞典』(北原 2010[2])では，特殊事例がゴシック体で示されている：

(14) 「重量**挙げ**【上げ・挙げ・揚げ】」「もみじの**ような**手だ【ようだ】」「誇り**高き**騎士【高い】」(北原 2010[2]: xiii)

この種の情報は，コーパスの書字形情報 [3] と語彙素情報を組み合わせることで比較的容易に抽出・分類し，辞書編纂者に事前に提示することができよう。

謝　辞

本研究は、文部科学省科学研究費補助金 基盤研究(A)「汎用的日本語学習辞書開発データベース構築とその基盤形成のための研究」（平成 23～26 年度，課題番号：23242026；研究代表者：砂川有里子）による補助を得ています。

文　献

Baayen, R. Harald (2001) *Word Frequency Distributions*. Dortrecht: Kluwer.

Béjoint, Henri (2010) *The Lexicography of English*. Oxford: Oxford University Press.

千葉庄寿 (2011)「BCCWJ の量的情報の活用：語彙情報のプロファイリングを例に」『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. 89-92.

Fox, Gwyneth (1987) “The case for examples,” in Sinclair, John M. (Ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT, pp. 137-149.

Heid, Ulrich (2008) “Corpus linguistics and lexicography,” in Lüdeling, Anke, and Merja Kytö (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Walter de Gruyter, pp. 131-153.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrž, and David Tugwell (2004) “The Sketch Engine,” in Williams, G., and S. Vessier (eds.) *EURALEX 2004 Proceedings*. Lorient: Université de Bretagne-Sud, pp. 105-116.

北原保雄(編) (2010[2])『明鏡国語辞典』第 2 版. 大修館書店.

Laufer, Batia (1992) “Corpus-based versus lexicographer examples in comprehension and production of new words,” in Tommola, H., K. Varantola, T. Salmi-Tolonen, and J. Schopp (eds.) *EURALEX 1992 Proceedings*. Tampere: University of Tampere, pp. 71-76.

Rundell, Michael (1998) “Recent trends in English pedagogical lexicography,” *International Journal of Lexicography*. 11/4: 315-342.

Sinclair, John M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

田中牧郎 (2011)「語彙レベルに基づく重要語彙リストの作成―国語政策・国語教育での活用のために―」言語政策班報告書 (JC-P-10-01), pp. 77-87. (特定領域研究「日本語コーパス」研究成果報告 DVD 所収.)

内山将夫，中條清美，山本英子，井佐原均 (2004)「英語教育のための分野特徴単語の選定尺度の比較」『自然言語処理』11/3: 165-197

Walter, Elizabeth (2010) “Using corpora to write dictionaries,” in O’Keeffe, Anne, and Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp. 428-443.

[3] UniDic では書字形情報は基本形ではなく活用していることに注意が必要である。

口頭発表（2）

3月5日（月） 15:00〜17:00

テキストの難易度に対する人間の判断と機械の判断

佐藤理史　（名古屋大学 大学院工学研究科）†
柏野和佳子　（国立国語研究所 言語資源研究系）‡

Which Text is Easier?
—Judgement by Human and Machine—

Satoshi Sato　(Graduate School of Engineering, Nagoya University)
Wakako Kashino　(Dept. Corpus Studies, NINJAL)

1　はじめに

いま、被験者に、1000 字の日本語テキストを印刷した 2 枚の紙を渡し、どちらのテキストがやさしいかを回答してもらうことを考えよう。多くの被験者に、このような課題を与えたとき、彼らの回答は一致するのであろうか。

我々は、母国語である日本語のテキストを読んで、そのテキストの難易を判定する能力を有している。たとえば、上記の課題で、小学校の教科書から抜き出したテキストと大学の教科書から抜き出したテキストを比較するのであれば、その判断は容易であろう。そして、その回答は、日本語を母国語とする成人であれば、まず間違いなく一致するだろう。その一方で、我々は、実生活上の経験から、「テキストの難易の判断には個人差がある」ことを知っている。つまり、任意の 2 つのテキストに対して上記のような課題への回答を被験者に求めた場合、その回答は、被験者間で必ずしも一致しないと予想される。

テキストの難易度に関する研究は、リーダビリティ研究として長い歴史を持ち、特に英語に対して、これまで多くの研究がある [1, 2]。その中核は、テキストの難易度と強い相関を持った特徴量 (使用語彙や文の長さ) の発見と、それらを用いた難易度推定公式の提案である。その一方で、人間にテキストの難易の判定を直接求める被験者実験は、ほとんど行なわれていない。すくなくとも、我々が知る限り、日本語テキストに対して、そのような被験者実験を行なった結果は報告されていない。

本論文では、この未調査領域に着目して実施した、テキストの難易の判定を求める被験者実験とその結果について述べる。今回の実験の主要な調査項目は、次の 2 点である。

Q1 人間の判断の一貫性：人間の判断は、被験者間で一致するのか。
Q2 機械の判断の妥当性：人間の判断と機械の判定結果は、どの程度一致するのか。

2　テキストの理解とテキストの難易

テキストの難易という概念は、かならずしも明確な概念ではない。我々人間がテキストを理解する過程において、理解の促進・阻害には、多くの要因が関係している。これらは、おおまかに、次のように整理される。

1. 読むもの (テキスト) 側の要因
   (a) テキストの表示の見やすさ・見にくさ (legibility)：表示媒体、フォーマット、印刷状態など
   (b) テキスト表現のやさしさ・難しさ (readability)：文章表現、語彙、文体など
   (c) 書かれている内容の複雑さ
2. 読み手側の要因

† ssato@nuee.nagoya-u.ac.jp　‡ waka@ninjal.ac.jp

(a) その言語の運用能力
(b) 書かれている内容に対する背景知識
(c) その時の身体状況

これらの要因のうち、我々は、まず、読み手側の要因を、平均的な読み手を仮定することによって捨象する。次に、読むもの側の要因のうち、1(a) の表示に関わる要因を、同一条件となるように制御して排除する。こうして残る要因は、1(b) と 1(c) となる。

一般に、伝える内容が複雑になれば、伝える文章も難しくなるのが普通である。このため、1(b) と 1(c) を完全に分離することは難しい。たとえば、「ゆで卵の作り方」と「ブフ・ブルギニョン (牛肉の赤ワイン煮) の作り方」では、伝達すべき内容 (料理の作成手順の詳細) は、後者の方が複雑である。それを反映する形で、後者のテキストでは、より多くの語彙が使われ、文や文章の構成が複雑化する。

しかし、その一方で、しかし、ほぼ同じ内容を伝える文章において、平易な文章と難しい文章があるのも事実である。たとえば、プロの料理人向けの「ブフ・ブルギニョンの作り方」にはフランス語由来の専門用語が使われるが、これを日常的な一般語に置き換えれば、すくなくとも語彙という観点では平易になる。また、テキストライティングの本には、伝達内容をあまり落さずに、文や文章を平易にする方法が示されている。

以上のような考えに基づき、我々は、1(b) と 1(c) は概念的には独立な要因であるという立場を採用する。もちろん、1(c) の要因の影響を受けることは避けられない。しかし、我々が「テキストの難易」という用語で指し示すものは、テキストの理解過程における 1(b) の要因とする。これを日本語テキストを対象とする場合において、より平易に言い直せば、「そのテキストがどのような日本語で書かれているか、その日本語は理解しやすいか」ということである。

3 課題設計

テキストの難易の判定を求める被験者実験において、どのような課題を用いるかは自明ではない。我々の知る限り、日本語テキストに対して、このような被験者実験を行なった例はなく、新たに課題を設計する必要がある。

3.1 テキストサイズ

第一に考えるべき問題は、難易の判定を求めるそれぞれのテキストの長さをどの程度にすべきかという問題である。これは、「テキストの難易度というものは、どの程度の長さのテキストに対して定義するのが適切か」という問題でもある。

被験者にテキストを読むことを求めるのであるから、あまり長いテキストを用いるのは非現実的である。他方、文の難易ではなく、テキスト (文章) の難易を計りたいのであるから、それなりの分量があってしかるべきである。

我々は、最終的に、約 1000 字というサイズを定めた。これは、以下に示す理由を勘案した総合的判断に基づく。ただし、難易度を定義するサイズとして、1000 字が最適なサイズであるということを主張するものではない。

1. 我々の経験では、書籍を開いて、その見開き 2 ページを読めば、その書籍がどの程度難しいか、おおよそ見当がつく。書籍の 1 ページ当たりの文字数は千差万別であるが、新書では、約 1000 字程度である。
2. A4 サイズの紙に印刷すると、ほぼ 1 ページの分量となる。紙に印刷して見比べるには都合がよい。

3. テキストを選ぶ母集団に予定している「現代日本語書き言葉均衡コーパス (BCCWJ)[3]」の固定長サンプルのサイズが 1000 字である。
4. テキストサイズが 1000 字あれば、文字や語彙に関する統計量がそれなりに安定する。

3.2 課題形式

次に考えるべき問題は、被験者に課す具体的な課題形式である。我々は、次の 2 つの形式を検討した。

1. 1 対比較課題：2 つのテキストを与えて、どちらがやさしいかを回答させる。
2. 並べ替え課題：n 個のテキストを与えて、それらをやさしい順に並べさせる。

テキストの難易の基本となるのは、テキストの 1 対比較である。1 対比較課題は、これをそのまま課題とするものである。この場合、2 つのテキストに対して回答が 1 つ得られるので、テキスト 1 件当たりの 1 対比較結果は 1/2 個となる。

これに対して、並べ替え課題の場合、読むテキストの数に対して、より多くの 1 対比較結果が得られる。たとえば、$n = 4$ の場合を考えよう。4 つのテキストを A, B, C, D と表した場合、被験者の回答は、たとえば、“$B < A < D < C$” のような形となる。この被験者の回答は、6 個の 1 対比較結果、すなわち、“$B < A$, $B < D$, $B < C$, $A < D$, $A < C$, $D < C$” に同意しているとみなすことができる。このように解釈した場合、4 個のテキストを読めば、6 つの 1 対比較結果が得られることになるので、テキスト 1 件当たりの 1 対比較結果は 3/2 個となる。

一般に、n を大きくすれば、テキスト 1 件当たりに得られる 1 対比較は多くなる。その一方で、課題の遂行が繁雑になり、難しくなる。我々は、まず、$n = 5$ の実行可能性を探ったが、5 枚の紙を見比べるのは、認知負荷的にも作業スペース的にもきつかったため、これを断念し、$n = 4$ を採用した。

4 課題セットの編纂

我々は、今回、上記で述べた並べ替え課題 ($n = 4$) を 20 課題作成した。

4.1 方針

実際の課題の編纂において、次の方針を立てた。

方針 1 課題に使用するテキストとして、「現代日本語書き言葉均衡コーパス (BCCWJ)」の固定長サンプル (1000 字) を用いる。

方針 2 1 つの課題に含まれる 4 つのテキストのジャンルを揃える。具体的には、日本十進分類 (NDC) の 3 桁が一致するテキストを選択する。

方針 3 1 つの課題に含まれる 4 つのテキストに、難易度が異なると思われるものを含める。

方針 1 は、多くの研究者が、使用したテキストにアクセスできるようにするために定めた。被験者実験に使用したテキストが入手可能であれば、実験結果を利用した各種の調査 (たとえば、語彙や文長などの調査) が可能となる。

方針 2 は、「まったく異なるジャンルのテキストを比較することは難しいだろう」という予想に基づき定めた。3 桁の NDC が一致したとしても、かならずしも同じような内容のテキストとは限らない。しかしながら、NDC を揃えないよりは揃えた方が、書かれている内容に対する依存度を軽減することができると考え、このような方針を採用した。

方針 3 は、「似たようなレベルの難易を判定することはかなり難しい」という、これまでの経験に基づき定めた。4 つのテキスト群の中に、相対的にやさしいテキストや相対的に難しいテキストが存

在していれば、並べ替えは比較的容易になると考えられる。テキスト群には、被験者が判断に迷うものも含まれていてよいが、そのようなものばかりだと、被験者に過度の負担を強いることになる。我々は、難易度の値の分布は正規分布に従うと考えており、その仮定が正しいとすると、ランダムサンプリングでは、平均的な難易度のテキストが多数、選ばれることになる。そのため、ランダムサンプリングは採用せず、難易度の値が異なると思われるものを、意図的に選ぶ。

4.2 編纂の実際

課題セットの編纂には、「現代日本語書き言葉均衡コーパス」(2009 年モニター版) のうち、書籍 (BK) の固定長サンプル (9,428 サンプル) を用い、次の手順で 20 課題からなる課題セットを編纂した。

1. 被験者への提示にふさわしいテキストサンプルを選択する。具体的には、そのサンプルにおいて、
   (a) article (同一著者による、同一テーマのひとまとまりの文書要素) は 1 つ、
   (b) title (特定範囲の文書要素の内容を代表する記述) は 2 つ以下、
   (c) caption (図表についてのタイトルや説明)、quotation (当該 article 要素とは異なる著作物からの引用や、発話・心内発話の引用・描写・書き起こし)、rejectedBlock (サンプル範囲内において、削除対象となったブロック要素の存在)、verse (詩、和歌、俳句、歌謡などの韻文) はそれぞれ 1 つ以下、

   という条件を満たすものを選択した。4,032 サンプルが選択された。
2. サンプル集合を、日本十進分類 (NDC コード 3 桁) でグルーピングする。なお、以下では、それぞれのグループを NDC グループと呼ぶ。
3. 含まれるサンプル数および分野のバランスを考慮して、20 個の NDC グループを選ぶ。
4. 選ばれたそれぞれの NDC グループに対して、以下を実行する。
   (a) すべてのサンプルに、obi2/B9[4] で難易度を付与する。この難易度は、1(とてもやさしい) から 9(とても難しい) までの 9 段階の値をとる。
   (b) obi2/B9 難易度で、サンプル集合をソートする。
   (c) ソートした列を 5 グループに等分し、中央のグループを除く 4 グループから、それぞれランダムに 1 つづつサンプルを抜き出す。
   (d) 抜き出した 4 つのサンプルを BCCWJ の ID 順にソートし、A から D の記号を付与する。

作成した 20 個の課題 ($t_1, t_2, \cdots, t_{20}$) に使用したテキストサンプルの一覧を表 1 に示す。この表において、各サンプルは、BCCWJ の ID で表示し、その直後の数字は、そのサンプルの obi2/B9 難易度を表す。

5 被験者実験

編纂した課題セットを用いて、31 名の被験者に対して実験を行なった。被験者は、すべて日本語を母国語とする成人であり、男女比は 6 名対 25 名、年齢層は 20 代 9 名、30 代 8 名、40 代 12 名、50 代 2 名である。

被験者には、次のような指示を与えた。

1. それぞれのセクションにおいて、A から D までの 4 つのテキストが示されています。それぞれのテキストの長さは、約 1000 字です。(テキストは、文章の途中から始まっていることもあります。)

表 1: 課題セットに使用したテキストサンプルの一覧

課題 ID	NDC	A		B		C		D	
t_1	049	LBh0_00001	6	LBi0_00008	3	PB30_00080	4	PB40_00029	5
t_2	159	LBm1_00036	6	PB21_00081	3	PB41_00019	6	PB51_00022	5
t_3	188	LBo1_00028	5	LBo1_00031	4	PB21_00057	6	PB51_00071	6
t_4	210	LBg2_00067	7	LBp2_00050	5	PB12_00061	6	PB42_00048	6
t_5	289	LBc2_00043	6	LBc2_00044	8	LBn2_00008	4	PB22_00135	3
t_6	291	LBm2_00069	5	PB22_00295	7	PB52_00028	5	PB52_00097	3
t_7	302	LBe3_00057	4	LBh3_00090	7	LBn3_00122	7	PB23_00011	5
t_8	312	LBa3_00038	7	LBg3_00076	6	LBm3_00066	8	PB53_00488	6
t_9	335	LBe3_00049	5	LBh3_00101	8	LBo3_00061	6	PB53_00191	8
t_{10}	361	LBi3_00090	6	LBr3_00135	6	PB23_00122	7	PB33_01003	8
t_{11}	367	LBl3_00039	8	LBt3_00058	3	PB23_00258	7	PB43_00904	5
t_{12}	369	LBj3_00094	5	PB33_00465	8	PB43_00470	4	PB53_00341	9
t_{13}	493	LBa4_00018	3	PB14_00183	4	PB14_00232	6	PB24_00023	7
t_{14}	498	LBc4_00017	3	LBo4_00052	5	LBs4_00017	7	PB34_00399	6
t_{15}	673	PB36_00116	3	PB36_00165	4	PB56_00020	5	PB56_00064	8
t_{16}	783	LBm7_00049	7	LBq7_00031	5	PB17_00054	5	PB27_00075	3
t_{17}	913	LBm9_00025	2	LBo9_00127	5	PB39_00681	3	PB59_00309	5
t_{18}	914	LBf9_00067	4	LBk9_00080	5	LBp9_00155	2	PB49_00275	7
t_{19}	916	LBl9_00262	5	LBm9_00267	3	PB29_00053	5	PB49_00126	3
t_{20}	933	LBg9_00165	2	LBj9_00197	3	LBq9_00073	4	PB59_00371	3

2. それぞれのテキストに目を通し、あたなが感じた素朴な印象 に基づいて、やさしい順に並べて下さい。
3. 順位が付けるのが難しい場合は、何度も読み比べてもかまいません。
4. どうしても難易度の順位が付けられない場合は、回答シートの同じ箱に、複数の記号を記述して下さい。(ただし、可能である限りは、順位を付けて下さい。)
5. コメント欄には、難易度の順位付けが容易だったか難しかったかを記述して下さい。
   例： A が最もやさしく、C が最も難しいという判断は容易だったが、B と D は判断に迷った。最終的には $D < B$ としたが、難易度にはほとんど差がないように思う。
6. その他、気付いたこと、感想等があれば、コメント欄に記述して下さい。

6 実験結果の分析

6.1 1 対比較コードへの変換

並べ替え課題 ($n = 4$) の回答は、たとえば、“$B < A < D < C$” のような順序列である。これを 6 個の 1 対比較結果とみなし、6 ビットのコードで表すこととする。それぞれのビットは、上位ビットから、A と B、A と C, A と D、B と C、B と D、C と D の 1 対比較結果を表し、それぞれ、正順ならば 0、逆順ならば 1 と定める。この結果、上記の回答は “10001” と表現されることになる。課題セットは 20 課題から構成されるため、課題セットに対する被験者の回答は、120 ビットのコードで表現されることになる。これを **1 対比較コード**と名付ける。このコードのそれぞれのビットは、ある特定のテキストサンプルの組に対する 1 対比較の結果に対応する。

31 名の被験者 (p_{01}–p_{31}) の回答を、120 ビットの 1 対比較コードで表現したものを、表 2 に示す。この表では、先頭行 (m) に、31 名の被験者の過半数が、正順 (0) と逆順 (1) のどちらを支持しているかを示し (以下、多数派の回答とよぶ)、各被験者の行では、多数派の回答と値が異なるビットのみ、数字で示した。なお、列 d_0 は、各被験者の回答と多数派の回答の相違ビット数 (ハミング距離) を示している。また、下から 2 行目 (av.) に、31 名の被験者の平均値を示した。

表 2 で、縦の列に着目すると、31 名の被験者の結果がすべて一致する (すべて ‘.’) 1 対比較もあれば、判定が割れる 1 対比較もあることがわかる。この表から導ける、調査項目 Q1 に対する答えは、

表 2: 実験結果 (120 ビット 1 対比較コード)

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	d_0	d_c	d_i
m	111000	101001	110000	111011	111111	001111	000011	101001	000110	000100	111011	000100	100001	000001	000000	111111	000100	010100	101001	000001			
ci	−+?+++	−?+−++	++??++	−+−++−	+−−+−+	−+?++	−−+?++	−−+?++	−−+?+	?+?++	++−??+	−+−+?+	?−+−++	+−++??	+−−+−+	?++?+?	++−?++	++−+−	−++−+−	?+++??	120	96	62
p_{01}	.0...	.....	....1.	.....	.....	.....	....1.	.....	.....	1..0..	.....	.....	.....	1....0	.....	0.....	....0.	.....	....1.	....1.	11	2	2
p_{02}	.0...	.1....	.....	.....	.....	....0.	.....0.	.1....	.....	.....	....0.	.....	.....	1	1.....	0	.....	.....	.....	....1.0	11	4	4
p_{03}	.....	.1...0	.1....	.....	.....	...0..	....1..	..0...	.1....	.....	.....	....0.	.....	0	.....	00....	...0..	.....	.....	.....	12	6	6
p_{04}	.0...	.10..0	...1..	.....	0....0	.....	.....	.1....	.....	.....	.00.0.	.....	.....	....1.	.....	0.....	.....	.....	.....	11.1.0	17	8	7
p_{05}	.....	.....	...11.	.....	.....	..00..	...1.0	.1....	.....	1..0..	.....	.....	.....	.....	.....	.....	...0.1	.....	.....	1.....	12	3	3
p_{06}	1	1....0	..11..	.....	.....	.1....	...1..	.1....	.....	.1....	.....	.....	.....	1.1.1.	.....	1.0..0	.....	.....	.1....	....1.	18	7	7
p_{07}	.....	.....	...11.	.....	.....	0.1...	...1..	.....	.....	.....	0..1..	.1..1.	.....	0	.....	.....	...0..	1.....	.....	1.1.1.	15	7	7
p_{08}	.....	..0...	.0...1	.....	0	..0..0	1.1.0.	..0...	....0.	11....	000...	...0..	0.....	...1.0	11....	000...	.1..1.	.....	0.0...	....1.	30	22	18
p_{09}	.00...	.1...0	0.0...	.0..0.	.....	.1....	....0.	.....	.....	.1....	....0.	.1....	0.....	.1.1..	...1..	.....	.....	.....	....1.	.....	18	12	12
p_{10}	.....	..0...	.1..1	.....	.....	.0..0	.11..	.....	....0.	11.0..	...1..	...0..	0.....	.111.	1.....	...00.	.....	.....	0.0...	11.1.0	26	13	12
p_{11}	.0....	....1.	.1111	0..0.	.....	..0...	...1..	.1....	.....	11....	.....	0.1.0	0.....	0	...1..	...00.	.....	.....	..0...	...110	25	11	11
p_{12}	.01..	0....1	.....	.....	.....	.....	.....	.1....	.....	11....	.....	....0.	.....	..1.1.	.....	...000	.....	.....	.....	1.1...	17	7	6
p_{13}	...1..	.1....	.....	.....	.....	.....	...00	.1..0	.1..0.	.....	....0.	.....	0	.....	.....	.....	.0....	.....	.....	1....0	13	7	7
p_{14}	.....	..0...	...1..	.0....	.....	0	111...	.1....	.....	....0.	.....	.1....	0	.....	.....	.....	.....	.....	.....	0	14	8	6
p_{15}	.0...	.1...0	.....	...1..	.....	0	11....	...100	..0..0	.....	...1.0	....1.	0.....	.....	...1..	.0.0..	...0..	.....	..0...	1.1...	23	14	13
p_{16}	.....	.....	00...1	.....	.....	.....	.....	1.....	.....	1..0..	00	...0..	0.....	.....	11....	...0..	.....	.....	.....	.1.1.	20	12	10
p_{17}	.0....	.1....	.0....	.....	.....	.....	.....	....1.	.1.....	.....	....0.	....1.	.....	.....	....1.	.....	.....	.....	....1.	.....	9	5	5
p_{18}	.....	.1...0	.1....	.....	.....	...0..	.....	.1....	.....	.....	....0.	....1.	.....	....1.	....1.	.....	....1.	.0.0..	....1.	1.1.1.	16	8	7
p_{19}	.....	.1....	0.1..	...1..	0.....	.....	...1..	.1...0	.....	1..0..	...1..	...0.1	0.....	.1.1.	1.....	.00000	...0..	.....	1....	0	25	12	12
p_{20}	.....	.....	0..11.	....0.	0.....	.....	111.00	.0...	.11...	11....	0001..	.1..1.	.....	...1.1	1....1	0....0	...0..	.....	.....	...1.0	33	22	17
p_{21}	.....	00.1..	...1..	.....	.....	.....	....0.	....1.	.....	...0..	...1..	....1.	.....	..1.1.	.....	.....	.....	.0....	.....	1....0	14	7	7
p_{22}	.....	..0..0	..11.	....0.	.....	110000	...1..	..0...	....0.	11....	.....	....1.	.....	..0...	..1.1.	...00.	...0..	.....	.....	.....	24	13	10
p_{23}	.....	.1....	.1..1	.....	.....	...0..	...1..	.1....	.....	.1....	.....	.....	.....	.....	.....	.....	.....	.....	....1.	0	11	3	3
p_{24}	.00...	....1.	.11..	.....	0....0	.....	..11..	.1.1.0	....0.	...0..	.0..0.	.1.1..	.....	.....	0	0.....	.....	1.....	.....	..111.	24	12	11
p_{25}	..11.	..0...	.1....	.....	000.00	11...0	...1..	.....	....0.	...0..	...1..	...0.1	0.....	.....	0.1.11	000...	.....	0.....	.....	1.1.1.	31	21	14
p_{26}	.....	..0...	.0....	.....	.....	.0..0	.....	.1....	.....	...0..	.....	.....	0.....	1.....	1....1	0....0	1.....	.....	.....	.....	14	8	8
p_{27}	.00...	.1...0	00...1	.....	0.....	...0...	...11.	.1...0	.1.....	11....	...1..	....1.	.....	.....	0.....	000..0	...0..	.....	.....	...11.	26	12	12
p_{28}	.....	.....	.1....	.....	.....	0..0.0	.....	.0....	.....	...0..	.....	....1.	0.....	.....	.....	...000	...0..	0....	.....	1.....	14	5	5
p_{29}	.00...	.....	.11..	.....	.....	.....	.....	.....	....0.	11....	...1.0	.....	.....	...1.1.	.....	...0..	.....	0.....	.....	11...0	16	5	5
p_{30}	.....	.1....	..11.	.....	0.....	...00.	...1..	..0...	....0.	.....	.0..0.	.1....	0....0	1....0	.....	0....0	...0..	1....	.....	1....0	22	8	8
p_{31}	...1..	.1...0	.0....	....0.	.....	.....	.....	.....	.....	.....	...1.0	...0.1	0.....	111..0	1....1	...000	...011	1.....	.....	.....	23	16	15
av.																					18.8	9.7	8.7
B9	.....	.?....	.0...?	....0?	0.....	.?.....	...?..	....?.	.....?	?..0..	....0.	.1....	0....0	.....	.....	....?..	.....	....?.	?..?.	....?.	14.5	8.5	8.0

「人間の判断が一致するかどうかは、比較するテキスト対に依存する」という予想通りの帰結である。

6.2 1 対比較と有意水準

それぞれの 1 対比較に対して、31 名の被験者の回答を集計すると、多数派と少数派に分かれる。多数派の人数を M とするとき、「多数派 (過半数) である」ことが有意水準 1%で統計的に有意であるのは、$M \geq 22$ の場合に限られる[1]。今回の実験結果では、120 個の 1 対比較中、96 個が統計的に有意であった。(表 2 の行 ci で、'+' または '−' の列が、統計的に有意な 1 対比較を表す。)

ここで、我々は、1 対比較 (2 つのテキストの難易) の「正解」をどのように定義すべきかという問題に直面する。テキストの難易判定は人間しかできないのであるから、多くの人々[2]が同意する回答を、正解とするしか方法がない。ここで、「多くの人々」がどれくらいの割合を意味するものとするか—過半数でよいのか、それとも 2/3 以上必要か—には自由度があるが、その最低ラインは過半数である。

以降、本実験の分析では、多数派の人数 M が $M \geq 22$ を満たす 1 対比較 96 個に対して、多数派の回答を難易の正解順序として定義する。表 2 の列 d_c は、正解順序を定義した 96 ビットにおける、正解と被験者のハミング距離を示した。

列 d_c の値からわかるように、被験者と正解との距離は 2 から 22 まで分布し、その平均は 9.7 である。つまり、平均的な被験者は、96 個の 1 対比較において、10 個程度、正解 (多数派) とは異なる回答をするということである。なお、ここでは、詳細は省略するが、特定の 2 名の結果が非常によく似ているという事実はない。つまり、被験者のサブグループに特定の傾向があるということはなく、被験者の回答は、それぞれの個人によってばらつく。

6.3 独立な 1 対比較

前述のように「正解」を定義すると、96 個の 1 対比較を、独立な 1 対比較と、そうでない 1 対比較に分割することができる。たとえば、課題 t_1 の正解は 11?000 である ('?' は定義されないことを示す)。つまり、定義される正解順は、"$A > B,\ A > C,\ B < C,\ B < D,\ C < D$" となる。このうち、"$B < D\ \ (B < C < D)$" と "$B < A\ \ (B < C < A)$" は、他の 1 対比較の結果と推移律から一意に定まる。これらを除いた 3 つの 1 対比較結果 "$A > C,\ B < C,\ C < D$" は、独立である。それぞれの 1 対比較がどちらの区分に含まれるかを、表 2 の行 ci に、'+'(独立)、'−'(非独立) の記号で示した。独立な 1 対比較は、96 個中 62 個ある。この表の列 d_i は、独立な 1 対比較に限定した場合の、正解と被験者のハミング距離を示している。この表からわかるように、前述の d_c とこの d_i の間には、それほど大きな差はない。すなわち、独立な 1 対比較に限定したとしても、被験者の回答はばらつく。

6.4 機械による難易判定

先に述べたように、並べ替え課題 ($n = 4$) の各テキストには、obi2/B9 の難易度 (9 段階) が付与されている。この難易度を、人間の判断と同様の形に変換する。ただし、比較対象とする 2 つのテキストに、同一の難易度が付与されいる場合は、「判定不能」とする。このようにして求めた obi2/B9 による 20 課題に対する回答を、表 2 の最下行 (B9) に示す。ここで、'?' は、判定不能を意味する。

被験者の場合と同様に、obi2/B9 の結果に対しても、d_0、d_c、d_i を計算した[3]。このうち、d_c と d_i の値は、順に 8.5、8.0 であり、いずれも被験者の平均値を若干下回る。以上のことから、調査項目 Q2 に対し、「機械 (obi2/b9) の判断は、平均的な人間の判断と同程度である」という帰結が得られる。

---

[1] z 検定を用いた。

[2] どのような母集団を想定するかについては、さらに議論が必要であるが、ここでは、十分な日本語能力を持った成人母語話者を想定する。

[3] ただし、判定不能のビットは、距離 0.5 として計算した。

表 3: 被験者の判断と機械の判断

	obi2/B9 難易度の差								
被験者	+5	+4	+3	+2	+1	0	−1	−2	計
31–0	1	3	9	9	3	1	0	0	26
30–1	2	2	4	3	1	0	0	0	12
29–2	0	1	1	4	1	2	0	0	9
28–3	1	0	1	4	2	1	0	0	9
27–4	1	1	1	3	2	0	0	0	8
26–5	0	0	1	3	3	1	2	0	10
25–6	0	0	0	2	1	2	1	0	6
24–7	0	0	0	1	4	0	0	1	6
23–8	0	1	1	3	1	0	1	0	7
22–9	0	0	0	0	3	0	0	0	3
小計	5	8	18	32	21	7	4	1	96
21–10	0	0	0	0	2	2	1	0	5
20–11	0	1	0	3	2	0	0	0	6
19–12	0	0	0	1	2	2	1	1	7
18–13	0	0	0	0	1	0	0	0	1
17–14	0	0	0	1	1	2	0	0	4
16–15	0	0	0	0	1	0	0	0	1
総計	5	9	18	37	30	13	6	2	120

表 3 に、被験者による判定結果と機械の判定結果の、より詳細な比較を示す。この表は、被験者の結果が x 対 y だった 1 対比較に対する、obi2/B9 の判定結果 (難易度の差；差が正の場合が多数派の回答と一致) を示している[4]。なお、正解を定義するのは、22 対 9 以上の場合である。

この表からわかるように、obi2/B9 難易度の差が大きいほど、正解とよく一致する。事実、難易度の差が +3 以上あると判定したものは、正解が定義されれない 1 件を除き、すべて正解と一致していた。判定不能を除けば、obi2/B9 が正解と異なる判定を下すのは 5 件だけであり、そのうち 4 件は、難易度の差は −1 である。これらの結果も、機械の判断が人間の判断とよく一致していることを示している。

7 まとめ

本研究で得られた帰結は、次の 2 点にまとめられる。

1. テキストの難易に対する人間の判断は、比較対象のテキスト対に依存し、多くの人間の判断が一致するものもあれば、一致しないものもある。テキスト間の難易は、多くの人間の判断が一致するテキスト対に対してのみ、定義されるべきであろう。
2. テキストの難易に対する機械 (obi2/B9) の判断は、平均的な人間と同程度の性能である。

謝辞 本研究では、「現代日本語書き言葉均衡コーパス」(2009 年モニター版) の一部を利用した。また、本研究は、国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものである。

参考文献

[1] William H. DuBay. *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, 2007.

[2] William H. DuBay, editor. *Unlocking Language*. Impact Information, 2007.

[3] Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of LREC 2010*, pp. 1483–1486, 2010.

[4] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

[4] 対象とする 1 対比較は、全 120 個である。

大規模コーパスの利用とメタデータの役割

丸山 岳彦　（国立国語研究所 言語資源研究系）†

The Role of Metadata in the Analysis of Large-scale Corpora

Takehiko Maruyama　(Dept. Corpus Studies, NINJAL)

1　はじめに

大規模コーパスを言語分析に利用するためには、メタデータを参照することが欠かせない。また、コーパスを評価する際にも、メタデータの参照が必須となる。本稿では、大規模コーパスの利用・評価にとってメタデータがどのような役割を果たすかについて、『現代日本語書き言葉均衡コーパス(BCCWJ)』の「書誌情報データベース」を例に論じる。具体的には、メディア・ジャンル情報を利用したモダリティ形式の分析（4 節）、初出情報を利用した書籍サンプルの評価（5 節）を行なう。

2　メタデータの種類と役割

コーパスの本体となるテキストデータや音声データに対して、そのデータの出自（書誌情報、収録内容）、メディア、ジャンル、書き手・話し手の属性、社会的な位相などの情報を記録したデータを、「メタデータ」と呼ぶことにする。さまざまな種類のテキストデータ・音声データ（の転記テキスト）を検索して得られたコンコーダンスは、言語表現の断片の集積でしかないため、それぞれの言語表現が本来使われていた使用文脈や発話場面から切り離された状態にある。そこでメタデータを参照することにより、例えば、検索結果を書き手・話し手の性別によって分類したり、使用傾向の違いをジャンルごとに分析したりすることができる。大規模コーパスを用いて言語の使用実態を多様な観点から実証的に明らかにするためには、メタデータの存在が必須である。

ここでは、コーパスを構成する個別のサンプルの中身に対して付与（注釈付け）されたデータを「アノテーション情報」と呼び、メタデータと区別する。アノテーション情報は、文章・談話を構成する言語的階層（音素 < 語 < 文節 < 節 < 文 < 文章・談話）の各階層に対して付与される。これに対してメタデータは、サンプル全体に対して付与されるものとする。テキスト全体から自動的に算出される種々の統計値もメタデータの一種と考える。アノテーション情報とメタデータの関係の例を、図 1 に示す。これらはいずれも、「データのためのデータ（data about data)」として位置づけられる。

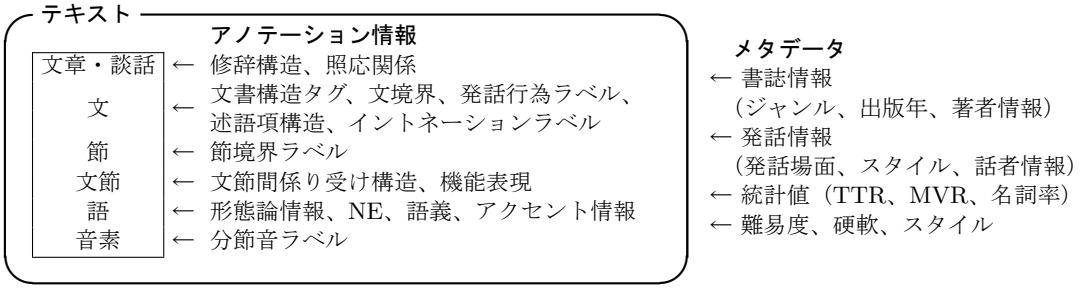


図 1: アノテーション情報とメタデータの例

Burnard (2004) によると、コーパスに付与されるメタデータは、図 2 のように分類できる。以下、本稿では、Burnard (2004) の分類に基づいて BCCWJ に付与されている「書誌情報データ」をメタデータとして整理し、それらを用いることで、どのように BCCWJ を利用または評価することができるかについて述べる。

† maruyama@ninjal.ac.jp

1. Corpus identification： コーパスの識別情報。コーパスの名称、作成者、配布元など。
2. Corpus derivation：サンプルの出自に関する情報。
   (a) Bibliographic description：書誌情報
   (b) Extent：サンプルサイズ
   (c) Languages：使用言語、言語コード
   (d) Classification：層別情報、カテゴリ
3. Corpus encoding：コーパスの編纂に関する情報
   (a) Project Goals：作成の目的
   (b) Sampling and extent：サンプリングの方法と結果
   (c) Editorial practice：編集方法（修正、追加、包摂など）
   (d) Markup scheme：マークアップの方法
   (e) Reference scheme：コーパスを参照するための情報（文番号、単語番号など）
   (f) Classification scheme：カテゴリの分類方法

図 2: コーパスに付与されるメタデータの分類例（Burnard, 2004）

3 BCCWJ の書誌情報データベース

BCCWJ は、「出版サブコーパス」「図書館サブコーパス」「特定目的サブコーパス」という 3 つのサブコーパスから構成される、約 1 億語の均衡コーパスである[1]。各サブコーパスには、「書籍」「雑誌」「新聞」「白書」「教科書」「広報紙」「Yahoo!知恵袋」「Yahoo!ブログ」「韻文」「法律」「国会会議録」から無作為抽出したテキストが、合計 172,675 サンプル収録されている。アノテーション情報として、2 種類の形態論情報（短単位、長単位）、および文書構造タグが付与され、XML 文書として構造化されている。これに加えて、メタデータとして、サンプルを取得した原本の書誌情報やジャンル情報、著者情報、サンプリングの状況などを記録した「書誌情報データベース」が同梱されている。サンプルごとに一意に付与された ID（「サンプル ID」）で関連付けることによって、コーパス本体と書誌情報データベースを結合することができるようになっている。

書誌情報データベースは、次の 4 つのデータから構成される。

書誌情報データ： サンプルを取得した原本やジャンルに関する情報。

サンプル情報データ： サンプルの ID や取得状況に関する情報。

記事情報データ： 記事に含まれる文章の初出および著者に関する情報。

人名録データ： サンプルの著者や著作権者などの人名録。

書誌情報データには、サンプルの原本に関する書誌情報、およびジャンルやカテゴリーに関する情報が記録されている。サンプル情報データには、サンプルごとに一意に付与された ID の他に、出版物（書籍、雑誌、新聞、白書、教科書）のサンプリングを実施した結果（サンプリングの基準となったページ数とページ内の座標）が記録されている。記事情報データは、書籍・雑誌・新聞のサンプルに含まれる「記事[2]」に関する初出情報、および実際に記事を執筆した著者に関する情報が記録されている。人名録データは、書誌情報や記事情報に現れる人名 ID に対応する人名を記録したデータである。4 つのテーブルを関連付けた結果を、図 3 に示す。

[1] 以下では、BCCWJ の全データを記録した「BCCWJ-DVD 版」を前提に話を進める。
[2] 「記事」とは、「同一著者によって、同一のテーマについてまとまりをもって書かれた文章の範囲」のことを指す。

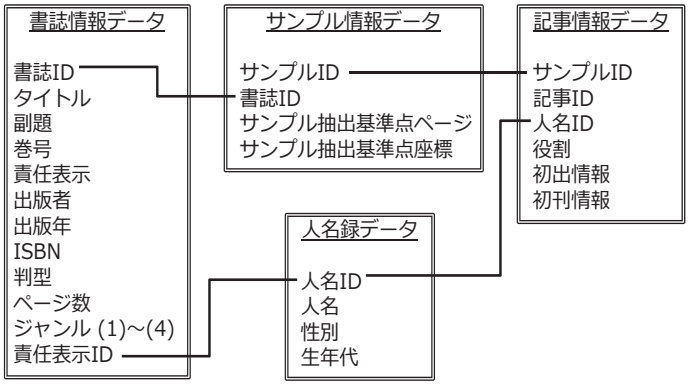


図 3: 書誌情報データベースの構成と関連付け

これらの書誌情報データベースによって提供される情報は、図 2 に示したメタデータのうち、2-(a) “Bibliographic description”、2-(d) “Classification”、および 3-(b) “Sampling and extent”（の一部）に相当する。 1 “Corpus identification” や、2-(b) “Extent”、2-(c) “Language”、および 3 “Corpus Encoding” については、「BCCWJ-DVD 版」に付属するマニュアル (国立国語研究所, 2011) にその中身が詳述してある。将来的には、これらの書誌情報データを、ダブリン・コアに基づくメタデータ記述の基本要素にどのように配置するかという問題を検討することも考えられるが、現時点ではその問題について触れる余裕がないので、ここでは擱く[3]。

BCCWJ を検索して用例を収集したりその出現数を集計したりするとき、分析者がまず考慮することは、どのような性質を持つテキストにどれだけの用例が出現したかという点であろう。多様なテキストの集合体である大規模コーパスを利用する場合、ジャンル、執筆年、執筆者など、収録されているテキストを切り分けるための分類指標ができるだけ多く付与されていることが望ましい。どのような出自や属性を持つデータ集合にどのような言語的特徴が認められるのか、という問いに的確に答えるためには、多角的な観点から詳細に記述されたメタデータの存在が欠かせない。

次節以降では、書誌情報データベースに記録された情報を用いて、BCCWJ をどのように利用・評価することができるかについて論じる。次の 4 節では、書誌情報データベースを用いて検索結果を分類する例として、モダリティ形式の分析例を示す。5 節では、書誌情報データベースに記録された「初出情報」をもとに、BCCWJ に収録された書籍のサンプルを評価する例を示す。

4 書誌情報データベースを用いた BCCWJ の検索と集計 —モダリティ形式の分析—

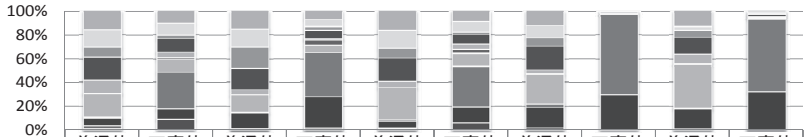
書誌情報データベースを用いて BCCWJ を検索・集計する例として、メディア[4]やジャンルの違いによって文末のモダリティ形式がどのような出現傾向を示すのかについて見てみよう。具体的には、ダロウ・ヨウダ・カモシレナイ・ラシイ・ミタイダという形式を取り上げる。これらは、事態に対する話し手の認識的なとらえ方を表す「認識のモダリティ」として扱われるが (日本語記述文法研究会, 2003) 、実際にそれらが現れやすい（または現れにくい）条件や使用場面についての記述はない。

そこで、「中納言[5]」を用いて、BCCWJ 全体を対象に検索を実施した。上に示した 5 種類のモダリティ形式と、それらを丁寧体にしたデショウ・ヨウデス・カモシレマセン・ラシイデス・ミタイデスという合計 10 通りの形式について、直後に句点が後接する事例を、短単位検索によって検索した。検索条件の例を (1) に挙げる。

[3] 千葉他 (2006) は、「青空文庫」で公開されている書誌情報をメタデータ化した研究である。
[4] サンプルを取得した媒体の種類（書籍、雑誌、白書、法律など）のことを、ここではメディアと呼ぶことにする。
[5] 中納言 RC2 (Released at 2011-11-10)、2012 年 1 月に検索を実施。

(1) (出現書字形 = "だろう" AND 品詞 LIKE "助動詞%") AND 後方共起: 出現書字形 = "。"
ON 1 WORDS FROM キー IN (subcorpusName="生産・書籍" AND core="true")
OR (subcorpusName="生産・書籍" AND core="false") WITH OPTIONS unit="1"
AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"

5 種類・10 通りのモダリティ形式について出現数を集計し、各メディアごとに 100 万語あたりの出現数を求めた。結果を図 4 に示す。



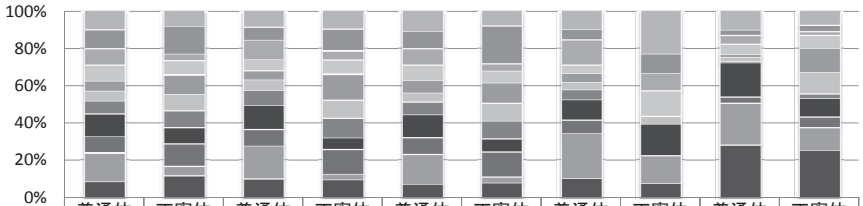
	普通体 (2,268)	丁寧体 (2,192)	普通体 (462)	丁寧体 (722)	普通体 (745)	丁寧体 (445)	普通体 (629)	丁寧体 (97)	普通体 (28)	丁寧体 (131)
	だろう。 (4,460)		ようだ。 (1,184)		かもしれない。 (1,190)		らしい。 (726)		みたいだ。 (159)	
出版書籍 (952)	359	227	70	50	122	41	78	1	4	1
雑誌 (875)	325	220	70	45	110	38	63	2	1	2
新聞 (461)	186	57	80	24	59	7	45	0	1	1
図書館書籍 (1,142)	450	245	86	47	149	36	123	1	4	2
白書 (9)	5	1	1	0	1	0	0	0	0	0
教科書 (455)	247	100	18	9	35	19	23	0	2	1
広報紙 (82)	2	31	1	31	1	15	1	0	0	0
ベストセラー (1,224)	444	244	68	46	202	48	160	0	10	2
Yahoo!知恵袋 (1,285)	16	675	1	268	8	154	18	65	0	80
Yahoo!ブログ (873)	151	187	63	187	43	59	106	29	5	42
韻文 (88)	56	8	4	4	4	0	12	0	0	0
法律 (0)	0	0	0	0	0	0	0	0	0	0
国会会議録 (272)	26	197	0	10	10	27	1	0	0	0

図 4: 100 万語あたりのモダリティ形式の出現数（メディア別）

5 種類のモダリティ形式の出現数は、メディアの違いに関わらず、ダロウが圧倒的に多い。各形式の普通体と丁寧体の違いを見てみると、広報紙、Yahoo!知恵袋、Yahoo!ブログ、国会会議録において、丁寧体が多く用いられていることが分かる。不特定多数の読み手にメッセージを発信したり、特定の聞き手に話しかけたりするスタイルの文体の中では、丁寧体が優先されるためと解釈できる。ラシイとミタイダの丁寧体は、Yahoo!知恵袋と Yahoo!ブログでのみ特に出現数が多いことから、一般人が書く文章に特徴的に現れる形式であると考えられる。法律や白書ではモダリティ形式の出現数が極端に少ないが、これは、法規範の羅列である法律や、国内外の情勢を客観的に記述する白書において、話し手の主観的な認識を表すモダリティ形式が現れにくいためであると解釈できる。

次に書籍（出版、図書館、ベストセラー）について、書誌情報データの「ジャンル (1)」列に記載されている「NDC（日本十進分類法）」によって、各モダリティ形式の分布がどのように異なるかを見てみよう。NDC とは書籍をその主題・内容に基づいて分類したコードであり、BCCWJ では国立国会図書館における NDC の分類に準拠している。NDC の第 1 次区分（10 種）ごとに、100 万語あたりに出現する各モダリティ形式の出現数を求めた。結果を、図 5 に示す。

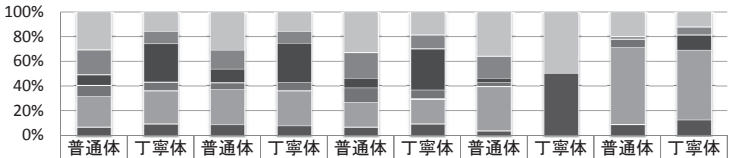
図 5 で普通体と丁寧体を比較すると、特に「文学」において、普通体の割合が丁寧体に比べて顕著に高い。これは、普通体で書かれた小説が「文学」の中に多く含まれるためであると考えられる。一方、「哲学」「自然科学」「工業」などの硬い内容を持つと思われる NDC で丁寧体の割合が高いという（一見すると意外な）結果が出ているが、これらの中には口語的な読み物のサンプルも一定数含まれており（『あなたを変える 3 つのレッスン（哲学）』『ぼくがすすめるがん治療（自然科学）』『電子レンジで朝ごはん（工業）』など）、丁寧体で書かれているこれらのサンプルにモダリティ形式が頻出すること（かつ、モダリティ形式が出現しにくい専門的な内容を持つサンプルは普通体で書かれていること）が原因であると考えられる。



	普通体 (3,271)	丁寧体 (2,696)	普通体 (787)	丁寧体 (594)	普通体 (1,105)	丁寧体 (488)	普通体 (700)	丁寧体 (5)	普通体 (54)	丁寧体 (15)
	だろう。 (5,967)		ようだ。 (1,382)		かもしれない。 (1,593)		らしい。 (705)		みたいだ。 (69)	
総記 (1,100)	396	271	82	68	143	47	82	1	7	1
哲学 (1,299)	392	468	66	83	122	117	49	1	2	1
歴史 (831)	337	115	95	34	114	21	111	1	3	0
社会科学 (857)	318	243	56	53	105	37	39	1	4	1
自然科学 (863)	215	321	42	96	88	63	34	0	1	2
工業 (766)	203	284	54	70	62	55	35	0	1	2
産業 (911)	273	294	76	76	92	55	44	0	0	0
芸術 (1,220)	469	279	121	42	161	41	92	1	12	2
言語 (1,149)	342	378	83	92	117	75	58	0	2	1
文学 (1,375)	589	162	161	21	206	19	200	1	14	2
分類なし (1,103)	328	368	93	67	94	46	84	0	18	4

図 5: 100 万語あたりのモダリティ形式の出現数（NDC 別）

最後に、同じく書籍について、書誌情報データの「ジャンル (3)」列に記載されている「C コード」を利用する。C コードは日本図書コードの一部で、「販売対象コード」、「発行形態コード」、「内容コード」から構成されている。ここでは、「販売対象」による分類（「一般（広く一般が対象）」「教養（知識階層が対象）」「実用（実務家が対象）」「専門（専門家学究者層が対象）」「児童（中学生以下の児童・生徒が対象）」）によって、各モダリティ形式の分布がどのように異なるかを見てみよう。販売対象ごとに、100 万語あたりに出現する各モダリティ形式の出現数を求めた。結果を、図 6 に示す。



	普通体 (1,627)	丁寧体 (1,567)	普通体 (404)	丁寧体 (319)	普通体 (515)	丁寧体 (226)	普通体 (368)	丁寧体 (2)	普通体 (45)	丁寧体 (15)
	だろう。 (3,194)		ようだ。 (723)		かもしれない。 (741)		らしい。 (370)		みたいだ。 (60)	
一般 (1,283)	499	248	126	51	170	43	133	1	9	2
教養 (783)	333	156	63	31	108	25	65	0	1	1
実用 (911)	143	493	45	101	39	75	13	0	0	2
専門 (386)	142	108	22	22	60	17	12	0	3	0
児童 (1,336)	403	412	115	90	104	45	131	0	28	9
その他 (389)	106	149	35	24	34	21	13	1	4	2

図 6: 100 万語あたりのモダリティ形式の出現数（C コード「販売対象」別）

普通体と丁寧体を比較すると、「実用」で丁寧体の割合が極めて高いことが分かる。「実用」の中には、『すぐ役立つ家庭の電気百科』『よくわかる会社更生法改正』のようなハウツー本が多く、読者に語りかける文体のテキストの中で、丁寧体のモダリティ形式が多用されている実態がうかがえる。また、ミタイダの出現数が、全カテゴリ中「児童」で際立って多い。ミタイダには、複数の辞書に「「ようだ」の口語的表現（新明解国語辞典 第六版）」「「ようだ」よりもくだけた言い方（例解新国語辞典 第八版）」のような記述があることから、そのような文体的特徴によって、文体のやわらかい児童書に多く出現しているという結果を解釈することができる。

以上で見たような、ある言語形式の出現傾向について、メディア、ジャンル、文体的な特徴などの

観点から定量的に分析するという方法論は、書誌情報データベースがあって初めて可能になることである。このような分析結果は、例えば辞書の編纂や日本語教育の現場において特に需要が高いと思われるが、従来の記述文法書の中では、このような視点からの記述は皆無であった。多種多様なテキストを収録した BCCWJ と、そのメタデータとしての書誌情報データベースを組み合わせて利用することによって、さまざまな位相における文法現象の分布について、定量的に記述することができる。今後は、どのようなメタデータを用いるとどのような言語形式の出現傾向を記述することができるか、その方法論と実践、そしてそれを可能にするメタデータの検討が求められると思われる。

5　書誌情報データベースを用いた BCCWJ の評価 ―初出情報の利用―

次に、本節では、書誌情報データベースの「記事情報データ」に記録された初出情報を利用して、書籍のサンプルを評価することについて述べる。

ある書籍に含まれる文章は、出版時において初めて世に発表されるものと、そうでないものとに分かれる。このうち前者は、一般的には「書き下ろし」と呼ばれる。一方、後者には、雑誌や新聞に連載されていた小説が単行本として出版される場合や、単行本が文庫として出版される場合などがある。中には、100 年以上前に出版された本が 2005 年に文庫として出版される例もある。

BCCWJ の書籍（出版、図書館、ベストセラー）には合計 22,058 サンプルが含まれているが、中には上記のような理由によって、古い時代に執筆されたテキストも収録されている（例えば、2005 年に出版された夏目漱石『吾輩は猫である』など）。無論、出版の実態を反映した結果であるので、これらの作品が収録されていることはサンプリングの結果としては正しい。しかしながら、『吾輩は猫である』のテキストに対して、「2005 年」という出版年だけでなく、初出に関する情報が付与されていることが、検索結果を利用する上では望ましい。また、あるコーパス（の部分集合）にどのようなテキストが収録されているかを評価する上でも、初出情報は重要である。

書誌情報データベースの「記事情報データ」は、このような問題意識によって作成されたメタデータである。各サンプルに含まれる「記事」を単位として、その文章がそれ以前に出版された経緯の有無を、原本の奥付や目録類などを参照して可能な限り調査した。そして、当該の記事が雑誌や新聞などで初めて発表・出版された年が判明した場合は「初出情報」として、当該の記事が初めて書籍として刊行された年が判明した場合は「初刊情報」として、それぞれ記録した。なお、初刊が確認できなかった場合や、書き下ろしであることが判明した場合は、出版年を初刊情報として記録した。

調査の結果、全 26,915 記事の 25.1%にあたる 6,755 記事から、初出・初刊に関する情報を取得することができた。初出・初刊のうち古い方の年（これを初出年と呼ぶことにする）が、出版年からどのくらい開いているかを集計した結果を、図 7 に示す。

初出年から 3 年程度は、新聞や雑誌での連載が単行本化されたり、単行本が文庫化されたりする場合が圧倒的に多い。初出年から 15 年で 100 記事以下に減少するが、個人全集の出版や名著の文庫化などにより、古い時代に書かれた文章が再度出版されるケースがロングテールで続いている。NDC 別に集計すると、初出年と出版年に開きがある書籍の数は、圧倒的に「文学」が多かった。

なお、BCCWJ に収録された書籍のサンプルのうち、初出年と出版年の開きが最大だったのは、福澤諭吉の『学問のすすめ』であった（1872 年初刊、2002 年出版、130 年の開き）。

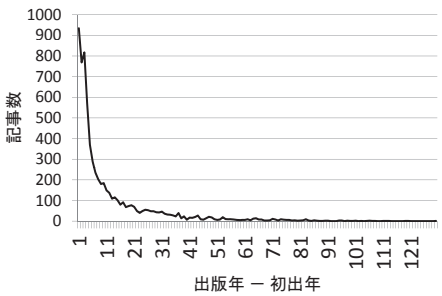


図 7: 書籍における出版年・初出年の開き

以下では、上記の初出情報を利用して、収録されているサンプルを語彙の面から特徴付けることを試みる。図 7 に示した初出年と出版年に開きのある記事のうち、50 年以上の開きがある記事は 281 記事あり、全記事の約 1%を占める。句読点等を含む語数は、合計で 890,314 語である。これに対して、2005 年に出版された書籍で、初出情報のない記事のうち、出版サブコーパスと図書館サブコーパスから 250 記事ずつを無作為に取得した。語数は、前者が 801,976 語、後者が 833,395 語である。

これで、50 年以上の開きがある記事のセットと、2005 年の書き下ろしと見なせる記事のセットを比較することができる。ここでは、2 つのセットのうちどちらかにしか出現しない語を、そのセットの「特徴語」と呼ぶことにする。2 つのセットそれぞれの特徴語（出現書字形）を品詞別に抽出したところ、表 1 のような結果を得た。上段は 50 年以上の開きがあるセット、下段は 2005 年の書き下ろしのセットから得た特徴語の例である。いずれも、出現数の多いものから順に挙げてある。

表 1: 2 つのセットにおける特徴語の例

名詞	まゝ、小吉、兵卒、細君、心持、氣、称呼、君子、攘夷、騾馬、等親、蠅、コサック、黒猫、空地、書生、燈籠、為め、牡丹、山猫、歌壇、爲、聲、義姉、世尊、高直、先き
	保育、ラッキー、メディア、補強、給付、データ、関数、福祉、再生、合金、データ、テレビ、細胞、原価、学習、コミュニケーション、介護、ベジタリアン、フード、暗号
形容詞	いゝ、なかつ、なかる、珍らしい、少い、危く、宜、蒼い、物凄い、新らしい、ゆる、旨く、少、よかつ、柔かい、まるい、よろしゅう、危い、巧い、明かるく、おそろしく
	すごく、美味しい、幼、きつい、やわらかく、温かく、ものすごく、やばい、上手く、厳し、きつい、ややこしく、長かっ、ややこしい、ひくい、嬉しかっ、しかたない
副詞	かう、忽ち、何う、先づ、たとい、悉く、已に、恰も、一ぱい、曾て、いろ、暫らく、少しく、兎に角、至極、すつかり、迚も、終に、頻り、唯、ちやんと、稍、矢張り、嘗て
	じっくり、そー、どー、ササ、わりと、良く、仲良く、ずばり、グスン、やっぱ、適宜、ずーっと、のろのろ、ちょくちょく、ぎくしゃく、ひくひく、とっとと、
動詞	思つ、行つ、言つ、考へ、言ふ、しまつ、云い、來、いつ、云わ、しまひ、起っ、依、をり、起る、向、帰つ、貰、於け、呉れ、御座い、出で、切つ、当る、あらう、言ひ
	超える、変える、分ける、抜き出し、捉える、受け入れる、取り付ける、取り組む、たらす、いかれ、贖い、注ぐ、摂る、図り、振りかぶり、斬り下ろす、欠かせ、囲う、ずれ
接続詞	或ひは、すなはち、只、然し、然も、然して、ち、尤も、もつとも、ぢゃ、乃ち、偖、んて、若くは、併し、して
	じゃぁ、ふんで、んじゃ

（上段：50 年以上の開きがあるセット、下段：2005 年書き下ろしのセット）

記事に対して付与された初出情報を利用することによって、書籍サンプルの中に見られる語彙の時代的な特徴を、表 1 のような形で把握することができる。BCCWJ 全体から見れば少数ではあるが、現代において出版されている書籍に含まれる語彙にどのような時代性が見られるかを、初出情報を参照することによって知ることができるわけである。

初出情報というメタデータの存在は、あるコーパス（の部分集合）に含まれるテキストがどのような性質を備えているかを評価する上で、重要な役割を果たすと思われる。図 7 で見たように、書籍とは部分的に再生産を繰り返す媒体であり、100 年以上前に出版された文章が現在でも出版され続けている。時間的に幅の広い日本語が混在している状態が書籍というメディアの性質であり、そこから無作為抽出した書籍サンプルの中身を把握する上で、初出情報の存在は欠かせないと言える。

6　おわりに

大規模コーパスの利用におけるメタデータの役割について、BCCWJ の書誌情報データベースを例に論じた。具体的には、メディア・ジャンル別に見られるモダリティ形式の分布、および初出情報を利用したテキストの評価と特徴語の抽出、という 2 点について示した。

今後のコーパス日本語学が取り得る方向性として、(1) 既存のコーパスを用いた言語学的分析、(2) 既存のコーパスに対するアノテーション情報・メタデータの付与、(3) 新規コーパスの設計・開発 という 3 つが考えられる。このうち (1) については、従来英語を中心に研究が進んできたコーパス言語学の方法論を現代日本語の諸側面に適用できるという点で、大きな進展が期待される。(2) については、これまでにも自然言語処理の分野においてさまざまなアノテーションが活発に行なわれてきているが、人文系の研究者にとっては使いにくい状況にあった。これに対して、国立国語研究所の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」が 2010 年度から開始されており、共通のデータ集合（BCCWJ のコアデータ）に対して集中的にアノテーションを実施する計画が進行中である。言語学的な分析とアノテーション情報の付与を関連させながら並行して進めることで、より効果的な検索方法や分類の基準、言語の使用実態に関する新たな知見の発見が期待できる。

この流れの中で、既存のコーパスに対する新たなメタデータの設計と付与もまた、今後の課題の一つとなるだろう。個々のサンプルに対する統計値のセットを設計して自動的に値を付与したり、テキストの難易度 (佐藤・柏野, 2012) や硬軟 (柏野他, 2012)、スタイルなどに関する情報を人手で付与したりすることが考えられる。そのようにして構築されたメタデータ群を利用して分析した結果は、メタデータを参照できないデータ（典型的には Web をクロールして得たメタデータのないデータ）を分析した結果よりも、はるかに信頼性と妥当性の高いものになると思われる。

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」、「コーパスアノテーションの基礎研究」によるものである。

参考文献

Burnard, L. (2004). Metadata for corpus work. In Wynne, M. (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pp. 30–46. Oxford: Oxbow Books.

千葉庄寿, 夷石寿賀子, 陳君慧 (2006).「「青空文庫」を言語コーパスとして使おう—メタデータ構築による歴史的・社会言語学的研究への応用の試み—」.『言語処理学会第 12 回年次大会発表論文集』, 915–918.

柏野和佳子, 立花幸子, 保田祥, 丸山岳彦, 奥村学, 佐藤理史, 徳永健伸, 大塚裕子, 佐渡島紗織 (2012).「テキストの硬さと軟らかさの考察—『現代日本語書き言葉均衡コーパス』の収録書籍を対象に—」.本予稿集所収.

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』.（BCCWJ-DVD 版に収録）.

日本語記述文法研究会（編） (2003).『現代日本語文法 4 第 8 部 モダリティ』. くろしお出版.

佐藤理史, 柏野和佳子 (2012).「テキストの難易度に対する人間の判断と機械の判断」.本予稿集所収.

「形容詞＋です」述語の生起要因についての準備的考察

前川喜久雄（国立国語研究所言語資源研究系）†

On the Factors Influencing the Occurrence of Japanese Predicate Consisting of Adjective and '*desu*': A Pilot Investigation Using BCCWJ

Kikuo Maekawa (Dept. Corpus Studies, NINJAL)

１．問題の所在

本稿では現代日本語の書き言葉において形容詞に助動詞の「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件を、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と呼ぶ）を用いて検討する。

日本語のいわゆる述態の文の述語には、名詞述語、形状詞述語、形容詞述語、動詞述語の４種類が存在することは広く認められている。またこれら４種の述語が名詞ないし形状詞述語のグループと形容詞ないし動詞述語のグループとに下位区分されることについても大方の意見は一致していると思われる。

この下位区分の根拠は、名詞ないし形状詞が単独で述語を構成することができず、述語を成立させるためにはいわゆる断定の助動詞（「だ」「です」の類）を必要とするのに対して、形容詞ないし動詞述語はそれら単独で述語を構成しうるという点にあるのだと考えられる。しかし、実際に用いられている日本語の用例中には形容詞に「です」が直接接続して述語を構成している例を容易に見つけだすことができる。以下ではこの種の述語を「A です。」述語と呼ぶ（記号「。」の意味については後述）。

「この花は白いです。」のような「A+です。」述語の文は文法的な容認度が低いと判断する研究者が多い。日本語文法において「A+です。」述語がどのように扱われているかを簡単に調べておこう。

２．日本語文法における「A です。」述語

日本語文法における「A です。」述語の扱いは、文法の執筆目的によって異なっている。日本語の言語学的記述を目的とした文法にあっては、これを日本語述語の一形式に認める立場はほとんど見当たらない。現代日本語文法の研究者に広く読まれていると思われる寺村(1982)や益岡・田窪(1992)などもこの立場である。[1]

逆に、日本語学習者のための教科書では、「A です。」述語を認める立場の方が一般的である。現在内外で広く利用されている教科書であるスリーエーネットワーク（1998）から例を引いておく。以下の対義語の練習問題では正解の形として「A です。」述語があらかじめ与えられている。このような練習問題は他の教科書でも容易に見つけることができる。

† kikuo@ninjal.ac.jp

[1] 昭和 27 年 4 月に国語審議会が建議した「これからの敬語」では「A+です。」を「平明・簡素な形として認めてよい」としているが、ここに紹介した記述文法はこの立場をとっていない

例：　タイは　寒いですか。……いいえ、（暑い）です。

小さい	古い	易しい	忙しい	暑い

1) あした　暇ですか。……いいえ、（　　）です。
2) あなたの　会社は　新しいですか。……いいえ、（　　）です。
3) 日本語は　難しいですか。……いいえ、（　　）です。
4) あなたの　うちは　大きいですか。……いいえ、（　　）です。

(p.70)

日本語教育において「A です。」述語を認める立場がいつから存在するかは明らかでないが、筆者の手許にある文献の調査によって知りえた範囲では、フランス語で執筆された日本語教科書である Mori(1972)に見つかる説明が最も早い時期のものであった。P.28 の説明では「です」が括弧に入って示されているが、P.33 では括弧なしに「A です。」述語が用いられている。

Les mots de qualité variables, à savoir les adjectives japonais, sont susceptibles de modifications morphologiques comme les verbes.

a) Les adjectifs peuvent conclure une phrase comme les verbes.

ex:　*Kore-wa utsukushi-i* (*-desu*). これはうつくしい（です）. Ceci est beau.

(p.28)

Phrase-exemples:

1) *Kore-wa akai hana-desu.* これあかいはなです.　C'est une fleur rouge.
2) *Kono inu-wa shiroi-desu.* このいぬはしろいです.　Ce chien est blanc.
3) *Sono hana-mo akai-desu.* そのはなもあかいです.　Cette fleur, elle aussi, est rouge.

etc.

(p.33)

学習者文法において「A です。」述語が普通に受入れられているのは、現実の日本語の反映である可能性がある。以下ではその可能性を大規模なコーパスを用いて検証し、次いで「A です。」述語の成立要件を検討する。

３．分析

３．１　データ

検証用データとして BCCWJ の形態論情報を利用する。検索には BCCWJ の形態論情報検索用 Web インターフェースである『中納言』を用いる。このデータとインターフェースは公開されているので読者による追試が可能である。『現代日本語書き言葉均衡コーパス』については前川（2008）等参照のこと。

３．２　形容詞述語の生起頻度

最初に形容詞述語全般において「A＋です」述語が占める量的な地位を明らかにしておく。形容詞述語には、形容詞だけで構成されるもの（「A。」）、形容詞に準体助詞「の」（ないし

「ん」）が後続してその後に「です」が後続するもの（「A＋の＋です。」）、「A＋です」に終助詞が後続するもの、「A＋です＋終助詞。」、形容詞に「です」の推量形が後続するもの（「A＋でしょう。」）等があるが、これらの述語については「A＋です。」のように容認度が問題とされることはない。

これらの形容詞述語の生起頻度を『中納言』を用いて検索した。下に「A＋終助詞。」タイプの形容詞述語の検索に利用した検索式を示す。この検索式は「キーに指定された短単位の品詞が形容詞で、その直後に品詞が終助詞の短単位が後続し、さらにその直後に文字『。』が出現している用例を BCCWJ 全体について検索せよ」を意味している。文末の判定は記号「。」で行っており、「？」や「！」は対象としていないことに注意。このようにして検索した形容詞述語各タイプの頻度を表 1 にまとめた。例末尾の記号は BCCWJ のサンプル ID である。

キー: 品詞 LIKE "形容詞%" AND 後方共起: 品詞 LIKE "助詞-終助詞%" ON 1 WORDS FROM キー AND 後方共起: 出現書字形 = "。" ON 2 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"

表 1 で最も生起頻度が高いのは形容詞が単独で述語を構成する「A。」であるが、問題となる「A+です。」述語はそれに次いで高い生起頻度を示している。殊に「A+です。」の頻度が文法的な容認度に問題がないとされる「A＋の＋です。」の頻度よりも高いことは注目に値する。表 1 は、量的にみるかぎり、「A+です。」述語が決して特異な言語現象ではないことを明瞭に示している。

表 1: 様々な形容詞述語の頻度

述語タイプ	頻度	例 (サンプル ID)
A。	108,081	いろいろ問題も多い。(LBk9_00121)
A+です。	11,154	果物もとても多いです。(LBk3_00042)
A+の+です。	6,697	偏食になってしまうことも多いのです。(LBa6_00006)
A+です+終助詞。	10,408	女性が多いですね。(LBd7_00001)
A+でしょう。	5,675	心と体を病んでいる人が多いでしょう。(LBe3_00075)

３．３　「A+です。」述語になりやすい語となりにくい語

次に「A+です。」述語になりやすい形容詞とそうでないものがあるかという問題を検討した。「A+です。」述語になりやすさを以下の式で計算することにした。この値を以下では「%A+です。」と呼ぶことにする。

「A＋です。」述語の頻度÷（「A＋です。」述語の頻度＋「A。」述語の頻度）×100

BCCWJ に含まれる出現頻度が 30 以上の形容詞（短単位）について「%A+です。」を計算

した。最上位 20 語と下位 20 語を表 2, 3 に示す。上位は単純に「%A+です。」が最も高い語を選んでいるが、下位には「%A+です。」がゼロのものが多数並ぶので、表 3 には「A+です。」述語が少なくとも 1 回生じている語について「%A+です。」が低いものを掲載した。

表 2：「%A+です。」最上位 15 語

形容詞	語形	「A。」頻度	「A+です。」頻度	%A+です。
うざい	ウザイ	22	25	53.2
嬉しい	ウレシイ	571	625	52.3
羨ましい	ウラヤマシイ	79	77	49.4
しんどい	シンドイ	30	27	47.4
美味しい	オイシイ	410	362	46.9
辛い	ツライ	228	190	45.5
怖い	コワイ	324	240	42.6
有り難い	アリガタイ	231	157	40.5
寂しい	サビシイ	125	80	39.0
臭い	クサイ	36	22	37.9
痛い	イタイ	271	163	37.6
きつい	キツイ	87	49	36.0
悔しい	クヤシイ	62	34	35.4
待ち遠しい	マチドオシイ	28	15	34.9
悲しい	カナシイ	105	55	34.4

表 3：「%A+です。」 下位 15 語

形容詞	語形	「A。」頻度	「A+です。」頻度	%A+です。
明るい	アカルイ	139	5	3.5
心地良い	ココチヨイ	87	3	3.3
無い	ナイ	59,127	1,926	3.2
疑わしい	ウタガワシイ	67	2	2.9
新しい	アタラシイ	114	3	2.6
荒い	アライ	39	1	2.5
望ましい	ノゾマシイ	489	12	2.4
乏しい	トボシイ	91	2	2.2
珍しい	メズラシイ	221	5	2.2
根強い	ネヅヨイ	53	1	1.9
久しい	ヒサシイ	60	1	1.6
等しい	ヒトシイ	245	4	1.6
凄まじい	スサマジイ	63	1	1.6
鋭い	スルドイ	72	1	1.4
相応しい	フサワシイ	122	1	0.8

３．４　形容詞の意味特性クラス

表 2，3 を比較すると「%A+です」の値に形容詞の意味特性が影響している可能性が窺われる。表 2 にはいわゆる感情形容詞の類が多く、表 3 には状態形容詞の類が多い。日本語の形容詞においては、話者（書き手）の主観的感情を表現するものとそうでないものとで

文法的なふるまいが異なることは多くの研究者が認めている。例えば寺村(1982)は、形容詞による感情表現を「感情状態の直接表現（～がコワイ）」、「感情的判断（～がオソロシイ）」、「属性規定（～が丸イ）」に三分類したうえで、「感情状態の直接表現」タイプの形容詞について、「感情をもつ主体が第三者だと不自然な文になる (p.145) 」と指摘している。

このような形容詞の意味特性と「%A+です。」の間に相関が存在しているかを検討するために、BCCWJ に生じている全形容詞を 3 種に分類した。分類基準は「話し手の感情、主体的感覚を直接表現」していると判断できて、「私が／は～」（「～」の部分が「A＋です。」述語）が自然なものを Type1、「対象の属性についての話し手の感覚ないし主観的価値判断の表現」と判断できて、「私が／は～」が不自然なものを Type2、そして「対象の属性の客観的表現」とみなされるものを Type3 と判断した。

ただし実際の作業にかかると、明らかに Type1 ではないが、Type3 とは判断しにくい形容詞が多数存在していたので、それらは Type2 に分類した。また形容詞に多義性が認められ、それが分類に影響することが一部にあった。その場合は量的に優勢な用法の意味について判断をくだし、個々の用例レベルでの分類は施さなかった（この問題については 3.6 節参照）。

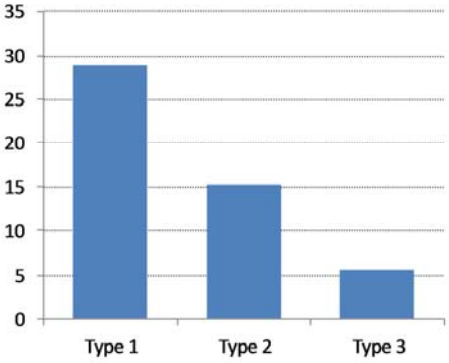


図1: 形容詞の意味特性と「%A+です。」平均値の関係

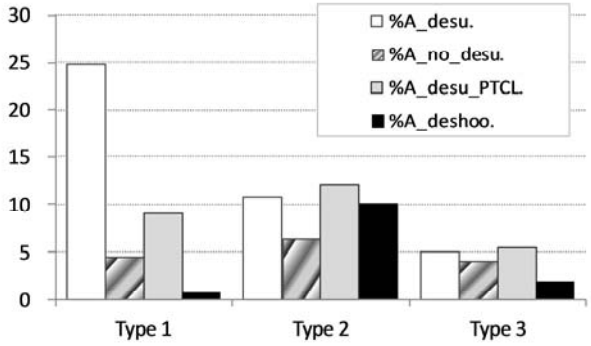


図2：形容詞の意味特性と各種形容詞述語タイプの生起率の関係

この判定作業は筆者が単独で実施した。以下ではこの判定結果を形容詞の「意味特性クラス」と呼ぶことにする。意味特性クラスごとの「%A+です。」の平均値を図 1 に示す。形容詞の主観性が高いほど「A+です。」の生起率が上昇していることがわかる。

このような関係が他の形容詞述語タイプにも認められるかどうかを検討した結果を図 2 に示す。ここでは各述語タイプの生起率を以下の式で計算しているので、図 2 中の「A+です。」述語の生起率（%A_desu）は図 1 における生起値よりも低い値をとっている。また図 2 ではグラフの見やすさのために「A。」述語の生起率を示すバーを省略している。「A。」述語の生起率（%A.）は Type1 で 60.9%、Type2 で 60.4%、Type3 で 83.7%である。

述語タイプ X の生起率 ＝ 述語タイプ X の頻度÷（全形容詞述語の頻度）×100

図 2 においても「A+です。」述語の生起率と形容詞の意味特性の間には図 1 と同様の相関

が存在している。しかし他の形容詞述語タイプについては同様の相関を見てとることができない。形容詞の意味特性は「A+です。」述語に限って機能する制約である。

３．５　レジスターの影響

図 1 では Type3 の形容詞群においても「A+です。」述語が 5%程度は生起している。この事実は、形容詞の意味特性が「A+です。」述語の生起に関わる絶対的な条件とはなっていないことを示すと同時に、形容詞の意味特性以外にも「A+です。」述語の生起に関わる重要な要因が存在する可能性を仄めかしている。

そのような要因の候補として有力と考えられるのが言語のレジスター(register)である。レジスターという用語は音声学と言語学において最も多義的に用いられている専門用語のひとつであるが、ここでは「言語が実際に運用される場の社会的状況―例えば発話の目的、発話者の属性、受容者との関係など―に依存して定まる言語の変種で、発話の全体にわたって分布する言語特徴によって決定されるもの」という意味で用いている。

BCCWJ にはこのような意味でのレジスターについてのアノテーションは施されていないので、サンプリングの際に利用した媒体をもってレジスターの代用とする。

表 4 にレジスターによる「%A＋です。」の変動を示す。表の第 1 列はレジスターの略称であり第 1 文字目の「P」「L」「O」は BCCWJ のサブコーパスである「出版サブコーパス」「図書館サブコーパス」「特定目的サブコーパス」に対応している。詳しくは丸山他(2011)参照。

表 4：BCCWJ のレジスターによる「%A+です。」平均値の変動

略号	レジスター	A。	A＋です。	%A+です。
OL	法律	296	0	0.0
OV	韻文	39	0	0.0
OW	白書	1,772	0	0.0
OT	教科書	1,015	4	0.4
OB	ベストセラー	5,917	35	0.6
LB	図書館書籍	44,058	296	0.7
PN	出版新聞	1,676	11	0.7
PB	出版書籍	35,634	353	1.0
PM	出版雑誌	5,724	131	2.2
OM	国会会議録	924	55	5.6
OY	ブログ	7,879	2,314	22.7
OP	広報誌	197	72	26.8
OC	ネット掲示板	2,950	7,883	72.8

表 4 は、レジスターが「A+です。」述語の生起率に強い影響を及ぼしていることを示している。「法律」「韻文」「白書」における生起率が 0.0%であるのに対して、「ブログ」と「広報誌」の生起率は 20%を超えており、就中「ネット掲示板」（Yahoo!知恵袋）における生起率は 70%を超えており、飛びぬけて高い値となっている。

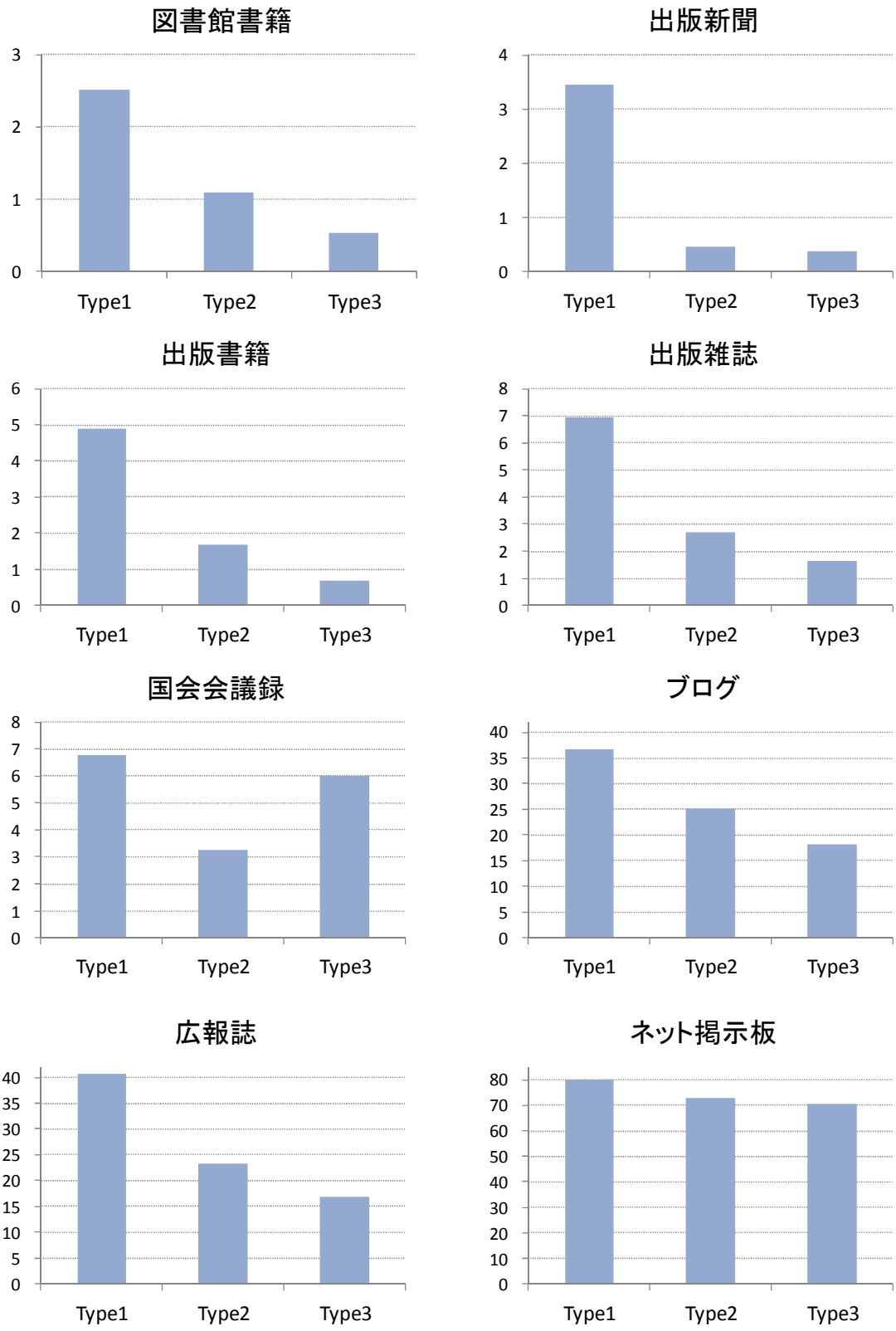


図 3：形容詞の意味特性とレジスターの相互作用

３．６　形容詞の意味特性とレジスターの関係

ここで問題になるのが、形容詞の意味特性とレジスターの関係である。両者は相互に独立して「%A+です。」に寄与するのか、あるいはそこに何らかの相互作用が存在しているのか。この問題を統計的に検討する手始めとして、形容詞の意味特性の影響をいくつかのレジスターにおいて確認しておくことにする。

図 3 は表 4 において「%A+です。」が最も低い 5 つのレジスターを除外した残り 8 種のレジスターについて図 1 と同様のグラフを描いたものである。横軸が形容詞の意味特性、縦軸が「%A+です。」である（縦軸のレンジがレジスター毎に異なっていることに注意）。

ここでは国会会議録を唯一の例外として、図 1 同様、Type1＞Type2＞Type3 の関係が成立している。国会会議録では Type3 に分類される「良い」の頻度が高く(N=81)、「%A+です。」も高い（15.9%）のが例外の原因だが、これは国会で質問者が回答者に対して「もういいです。」「まあいいです。」「それでいいです。」等の発話を多発していることによる。典型的な「良い」の用法とはずれた用法であり、むしろ Type1 に近いと考えられる。

次に形容詞の意味特性とレジスターを要因とする二元配置分散分析を実施した。分析には R 言語（Ver. 2.12.1）を利用した。形容詞の意味特性(df=2, F=4849, p<.0001)とレジスター(df=12, F=1015, p<.0001)はともに有意であり、両者の相互作用項(df=21, F=35, p<.0001)も有意である。

ここで問題になるのが相互作用の性格である。この相互作用は図 3 からわかるように Type3 形容詞の「%A+です。」がネット関係のテキスト、殊にネット掲示板で極端に高いことに起因していると考えられる。つまり国会会議録は除外すれば、他のレジスターにおいて Type3 形容詞の「%A+です。」は Type1 形容詞の 3 分の 1 以下の値をとっているのに対して、ブログでは 2 分の 1 程度、ネット掲示板では 8 割程度に高止まりしていることが原因だと考えられる。

４．議論

４．１　まとめ

前節の分析で「A+です。」述語の生起に影響を及ぼす要因には、言語的要因として形容詞の意味特性、語用論的要因としてレジスターの二つがあることが確認できた。意味特性に関しては主観性の強い形容詞ほど「A+です。」述語になりやすく、レジスターに関してはネット関係（ブログとネット掲示板）と広報誌において「A+です。」述語が生じやすい。両者の効果は基本的には独立しているが、ネット関係のレジスターでは形容詞の意味特性の効果をレジスターが凌駕している。

４．２　待遇表現としての「A+です。」

ネット掲示板などで「A＋です。」の生起頻度が高いのはなぜだろうか。これらのレジスターでは「A+です。」述語が一種の待遇表現として用いられている可能性が高い。ネット掲示板では質問者は「教えてもらえると嬉しいです。」の類の表現を頻繁に用いており、回答

者も特定の質問者を念頭において回答を執筆するので、待遇表現が働きやすいであろう。

これと同じ特徴は広報誌の読者欄にも認められる。また役所が制作する文書であるので、読者として想定する地方自治体の住人に対して待遇意識が働くのは、昨今の世相においては当然である。ブログの内容は様々であるが、読み手からのフィードバックが可能であるという点で、通常の書籍と比較すれば、執筆に際して書き手が読み手を強く意識するレジスターであることは間違いない。

４．３　例外の説明

4.1 節の分析と説明には、検討すべき問題が 3 点あると思われる。第一にこの説明が正しければ、図 1 において Type3 形容詞に生じていた「A+です。」述語はネット掲示板ないしブログに生じていることが期待される（広報誌は形容詞述語の頻度そのものが低いので説明力を期待できない）。データを分析するとこの予想が正しいことがわかる。図 1 の Type3 に生じた「A+です。」述語の 71.4%がネット掲示版に、20.3%がブログに生じており、両者で例外の 9 割が占められている。

４．４　サンプルの文体

レジスターの効果に対しては、そもそも「です体」（敬体）の使用率がレジスター毎に大幅に異なるはずだから、「A+です。」の生起率も、その変動を反映しているに過ぎないという批判が可能である。この問題を検討するため、レジスターごとに「です体」の使用率を計算した。記号「。」を文末とみなして各レジスターにおける文の総数を決定し、文末直前の 3 短単位内に助動詞「です」が生起していれば、それを「です体」の文と認定した。

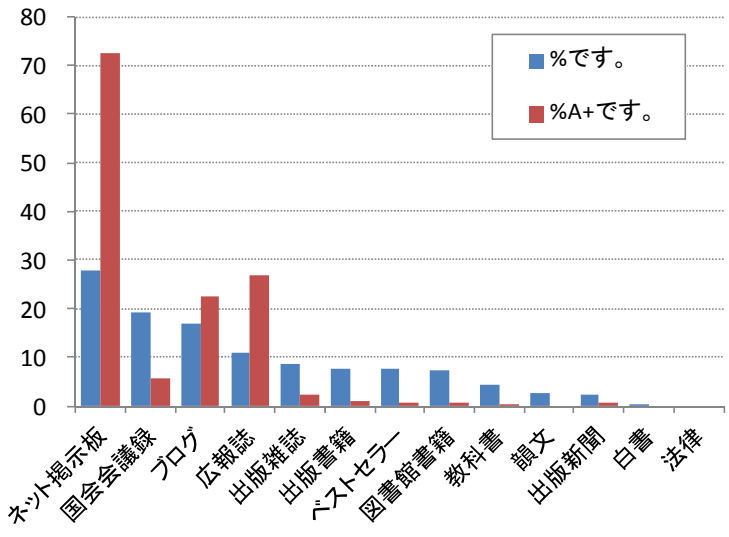


図4: 各レジスターにおける「です体」使用率と「%A+です。」の比較

各レジスターにおける「です体」の生起率（「%です。」）と「%A+です。」とを比較すると図 4 の結果を得る。全体としては高い相関（相関係数 r=0.82）が認められる。しかしネット掲示板と広報誌における「%A+です。」は「%です。」よりも著しく高く、反対に国会会議録における「%A+です。」は「%です。」よりも顕著に低い。

結論として、「です体」の使用率は「A+です。」の生起率に影響を及ぼしているが「A+で

す。」述語の生起頻度の高いレジスターについては、「です体」の使用率から「%A+です。」を予測することは不可能であるか、きわめて困難であると考えられる。

４．５　書き手の年齢

最後に「A+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。図5は、BCCWJのデータのうち比較的時間幅の広いサンプルを収録しており、同時に書き手の生年代の情報が得られていることの多い「図書館書籍」と「ベストセラー」のサンプルを用いて、10年区切りの生年代と「%A+です。」の平均値の関係を調査した結果である。形容詞述語の頻度が100未満の生年代は除外していること、また横軸は書き手の生年代であって、個々のサンプルの執筆年代ではないことに注意。

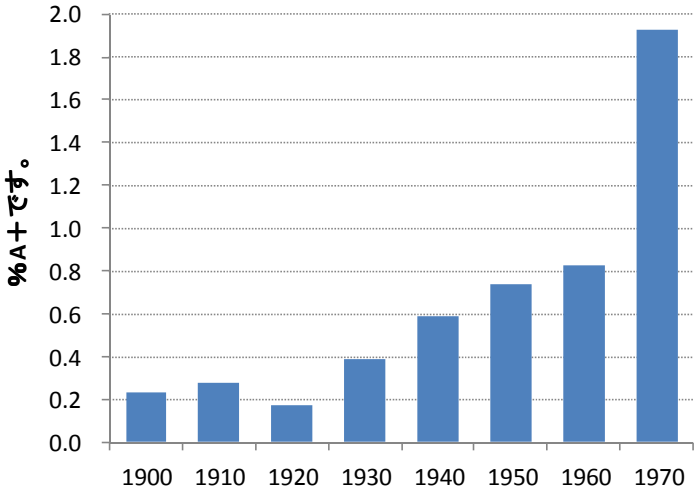


図5:　書き手の生年代と「%A+です。」の関係

図5では、「A+です。」が頻繁に用いられるネット掲示板、広報誌、ブログの3レジスターのデータを利用していないので、「%A+です。」は全体的に低い値にとどまっているが、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傾向を見てとることができる。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。

５．残る課題

今回の調査がカバーしていない形容詞述語の形（例えば過去形）を調査する必要がある。また4.2節で述べた待遇表現としての「A+です。」という仮説をより精緻化する必要がある。

参考文献

スリーエーネットワーク(1988).『みんなの日本語 初級Ⅰ本冊』スリーエーネットワーク.
寺村秀夫(1982).『日本語のシンタックスと意味Ⅰ』くろしお出版.
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.
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.
Mori, Arimasa (1972). *LEÇON DE JAPONAIS*. Taishukan, Tokyo.
丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子(2011).『「現代日本語書き言葉均衡コーパス」におけるサンプリングの原理と運用』国立国語研究所内部報告書 LR-CCG10-01.

謝辞：この研究は国語研基幹型共同研究「コーパス日本語学の創成」によるものです。共同研究発表会を含め、様々な機会にコメントをいただいた方々に感謝します。

共起語率の分布からみるテキストの語彙的特徴

山崎 誠（国立国語研究所言語資源研究系）†

Lexical Characteristics of Text as Seen in the Distribution of Co-occurrence Rate

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

「現代日本語書き言葉均衡コーパス」（Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ と略す）が 2011 年に完成し、それを利用した日本語研究のさまざまな展開が期待されている。BCCWJ の特徴として、多様な日本語を収録していることやアノテーションの充実が挙げられる。それらを生かした研究が今後発多く発表されることと思われる。本発表では BCCWJ のアノテーション情報を利用してテキストの結束性に関する特徴を捉える試みを紹介する。

２．テキストにおける結束性

結束性（cohesion）とは、文章をひとつの統一体としてまとめあげるために必要な性質のひとつとされる。結束性について最初に詳細に研究を行ったのは Halliday & Hasan(1976)である。それによると、結束性について次のように紹介されている。

「結束性が生じるのは，談話のある要素の解釈（INTERPRITATION）が別の要素の解釈に依存する場合である。一方を効果的に解釈するためには他方に頼らなければならないという意味で，一方は他方を前提（PRESUPPOSE）とする。こういうことが生じるとき，結束関係が成立する。その結果，前提語と被前提語という 2 つの要素が、少なくとも潜在的には，統合されて 1 つのテクストになるのである。」（邦訳 p.5）

庵(2007:12)によれば、結束性は推論にもとづくつながりである一貫性(coherence)の下位概念であるとされる。また、結束性には文法的結束性と語彙的結束性とがあり、前者の手段として「指示」「代用」「省略」が、後者には「再叙（reiteration）」と「コロケーション」がある[1] 。再叙には以下の 4 つのタイプがある。

(a)同一語（繰り返し）
(b)同義語（または近似同義語）
(c)上位語
(d)一般語

Károly(2002:162)によれば、英語の作文においては、(a)の同一語の繰り返しよりは(b)～(d)を合わせた「異なる語の繰り返し」の方が多く用いられるということだが、同義語（類義語）や上位語の判断を自動的に行うことが難しいため、本発表では(a)の同一語の繰り返しのみを観察対象とする。同一語の繰り返しは、本発表で用いた図書館書籍のデータでは、10,369 サンプル中同一語の繰り返し[2]が無かったサンプルは 17 個しかなかった。それらはいずれも延べ語数 22 語以下の小さなサンプルで、サンプルの短さがその原因である。ある程度の長さを持つテキストには必ず同一語の繰り返しがあると言ってよいだろう。

† yamazaki@ninjal.ac.jp

1 Halliday & Hasan(1976)では、文法的結束性と語彙的結束性の中間の性質を持つものとして「接続」が挙げられている。

2 ここでは同一語の繰り返しには、助詞・助動詞は含めていない。以下も同様。

３．データ

本発表では、2011 年 12 月にリリースされた『現代日本語書き言葉均衡コーパス』の DVD 版を使用した。Disk1 の M-XML フォルダに含まれる xml ファイルが対象である。この xml ファイルは可変長サンプルと固定長サンプルを統合したもので、短単位、長単位の形態論情報のタグのほか可変長部分には文章構造のタグを含んでいる[3]。

本発表ではこの xml ファイルにおいて<paragraph>というタグが付与された部分を対象にそこに含まれる短単位の形態論情報をもとに分析を行う。結束性を観察するには文も妥当な単位であるが、BCCWJ に付与された文を表すタグ<sentence>は見出しや図表のキャプションにも付与されており、通常の本文との区別をしなければならないため、今回の調査では確実に本文部分を表している<paragraph>タグを対象とした。<paragraph>タグを含むサンプル数は表 1 のとおりである。

表 1　対象サンプル数

媒体	全サンプル数	P サンプル数
出版書籍	10,117	9,742
雑誌	1,996	1,767
新聞	1,473	1,457
図書館書籍	10,551	10,369
白書	1,500	1,496
教科書	412	0
広報紙	354	354
ベストセラー	1,390	1,374
Yahoo!知恵袋	91,445	0
Yahoo!ブログ	52,680	0
韻文	252	0
法律	346	56
国会会議録	159	159
合計	172,675	26,774

教科書、Yahoo!知恵袋、Yahoo!ブログ、韻文は<paragraph>タグを用いていないため、対象サンプル数はゼロである。なお、<paragraph>タグの問題点については西部ほか(2011:232)を参照されたい。

表 2 は、対象となったサンプルの延べ語数、段落数、1 段落あたりの延べ語数、1 段落あたりの異なり語数のそれぞれの平均値である。1 段落当たりの延べ語数を見てみると国会会議録の値が大きい。これは国会会議録における段落の認定（1 発言が 1 段落）が影響しているものである。なお、語数には補助記号、空白、助詞、助動詞は含まれていない。

表 2　各媒体の延べ語数等の平均値

	サンプルの延べ語数	段落数	1 段落の延べ語数	1 段落の異なり語数
出版書籍	1,384.61	43.76	50.51	37.06
雑誌	891.17	29.81	40.05	33.27
新聞	334.33	9.28	38.78	33.33
図書館書籍	1,450.16	54.53	45.76	34.70
白書	1,793.10	29.32	64.74	44.33

[3] タグの詳細については小木曽ほか(2011)を参照。

広報紙	2,903.53	103.14	28.14	23.39
ベストセラー	1,404.46	69.30	29.52	24.28
法律	219.50	6.93	24.04	15.03
国会会議録	17,885.87	144.06	151.30	76.21

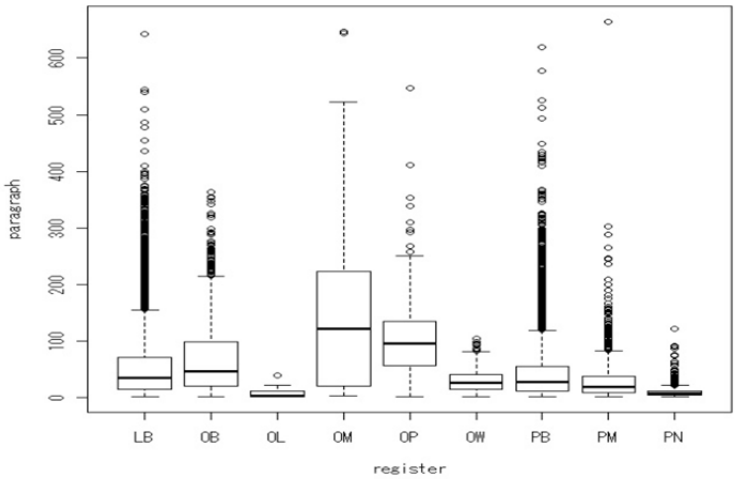


図 1 段落数の分布

図 1 は、サンプルあたりの段落数の分布の様子を媒体ごとに表したものである。全体的に分布が右に（大きい方に）かたよっていることが分かる。また、図書館書籍と出版書籍はほぼ似たような分布を示している。

４．結束性の算出方法

本発表では、ある段落とそれに隣接する段落との間で共通して現れる語の多寡に着目した。語の単純な繰り返しを扱うことのメリットは、他の結束性を表す現象と比べて正確な把握がしやすいこと、また、頻繁に起きる現象であるため、観察がしやすいことである。一方、デメリットとしては観察結果が「語」の単位認定基準に依拠してしまうこと及び同じ語か異なる語かだけの把握にとどまり、意味的な関係が把握できないことである。共通する語だけでなく、類義語等まで含めた計測方法として Hoey(1991)や Károly(2002)があるが、扱っているデータ量はさほど多くない。大量のデータを使って自動的に計測するには語の繰り返しがもっとも適していると思われる。

本発表では、以下の式により結束性の度合いを計り、共起語率と名付けた。

$$\mathrm{C(a,b)} = \frac{F(a,b)}{N_a}$$

a, b：段落番号(1～n)
$\mathrm{C(a,b)}$：段落 a の段落 b に対する共起語率。
$\mathrm{F(a,b)}$：段落 a と段落 b とで共通して現れる語の延べ語数を段落 a 内で数えた数。
N_a：段落 a の延べ語数。

共起語率は、水谷(1980)の非対称類似度を利用した指標である。そのため、連続する 2 つの段落の間の共起語率に 2 つの値が存在する。後続の段落に対する共起語率と前接の段落に対する共起語率である。上述の式では、b=a＋1 のとき、後続段落に対する共起語率とな

り、b=a－1 のとき、前節段落に対する共起語率となる。ただし、文章の冒頭の段落の前接段落及び最後の段落の後続段落は存在しないため、便宜的にその場合の共起語率は 0 とする。

この方法で共起語率を測るにはひとつ制約がある。それは、文章が 2 つ以上の段落から構成されていなければならないことである。そのため、表 1 で対象としたサンプルから 1 段落しかなかったサンプル 340 サンプルを除外した。

なお、計測対象からは言語表現とは見なさない補助記号、空白、及び文章の結束性には影響を及ぼさない助詞、助動詞を除外した。

５．結果

表 3 は、段落あたりの共起語の数と共起語率の平均値である。後続段落との共起語率と前接段落との共起語率とはほぼ等しい値を示している。このことは、どの媒体もそれぞれ同程度の依存関係でつながっていると解釈できる。個々に眺めてみると、法律、白書、国会会議録の共起語率が高く、新聞、ベストセラー、雑誌の共起語率が低いことが分かる。

表 3 共起語の数と共起語率

	後続段落との共起語数	後続段落との共起語率	前接段落との共起語数	前接段落との共起語率
出版書籍	12.98	0.22	12.74	0.22
雑誌	6.89	0.16	6.82	0.16
新聞	5.99	0.15	5.84	0.16
図書館書籍	10.49	0.19	10.36	0.19
白書	20.00	0.31	19.84	0.31
広報紙	5.19	0.18	5.13	0.17
ベストセラー	5.49	0.15	5.47	0.15
法律	12.16	0.48	12.31	0.47
国会会議録	40.45	0.30	39.01	0.30

表 4　NDC 別の共起語の数と共起語率

	後続段落との共起語数	後続段落との共起語率	前節段落との共起語数	前節段落との共起語率
0 総記	12.97	0.22	12.95	0.22
1 哲学	17.55	0.25	17.73	0.24
2 歴史	14.80	0.21	14.60	0.21
3 社会科学	15.02	0.24	14.84	0.24
4 自然科学	14.32	0.24	13.96	0.24
5 技術・工学	10.72	0.22	10.56	0.21
6 産業	11.03	0.21	10.82	0.21
7 芸術・美術	12.02	0.20	11.98	0.20
8 言語	10.40	0.21	10.17	0.20
9 文学	5.07	0.12	4.97	0.12
分類なし	3.46	0.13	3.45	0.13

表 4 は、図書館書籍のデータについて、NDC（日本十進分類法）別の共起語数と共起語率を算出したものである。図書館書籍全体では共起語率は 0.19 であったが、NDC 別に見ると「9 文学」と「分類なし」の値が他と比べて低いことが分かる。「分類なし」についてはデータを見ていないので理由は分からないが、「9 文学」は会話文のような短い段落が多いため、共起語率が低くなったと推測される（表 3 のベストセラーの値の低さもそれに起因しているであろう）。それを確かめるために、1 段落あたりの延べ語数の平均と共起語率の平均との相関を見てみよう。図 2 にその結果を示す。正の相関が認められ、決定係数は 0.799 と高い値を示した。

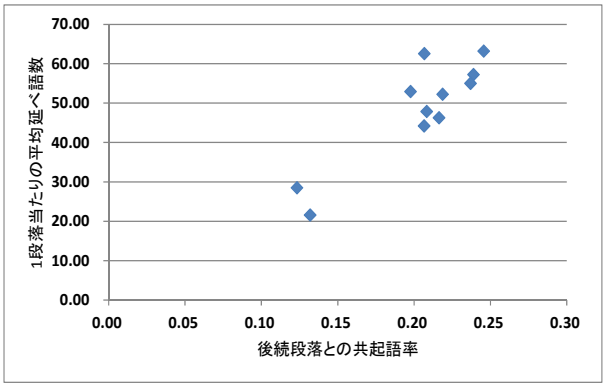


図 2　段落の延べ語数と共起語率との相関

６．文章中の共起語率の推移

共起語率の値はひとつの文章中でどのような変化を示すのだろうか。白書の例を見てみよう。図 3 は OW1X_00000（昭和 54 年版経済白書）というサンプルである。

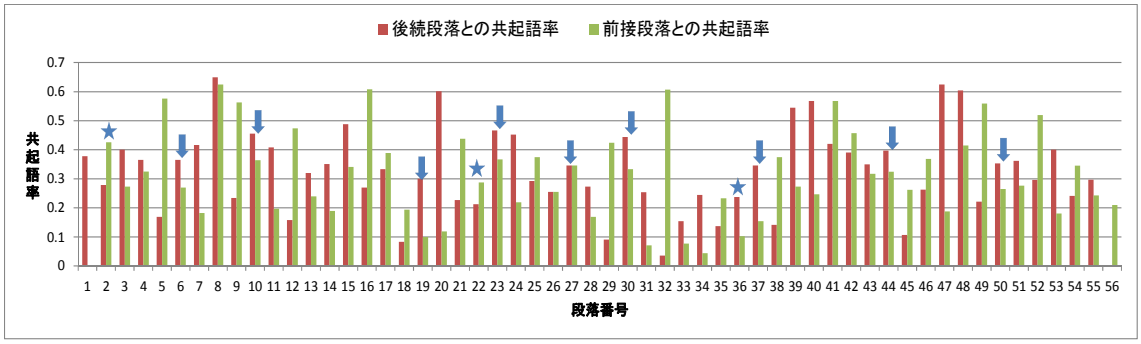


図 3　文章中の共起語率の推移

図 3 で、★を付けた 3 箇所は大きな節が開始する箇所、下向きの矢印を付した 9 箇所はその節の中で小見出しが立っている箇所である。矢印の部分における後続段落との共起語率（左側の棒）と前接段落との共起語率（右側の棒）とを比べてみると、9 箇所のうち 8 箇所が後続段落との共起語率が前接段落との共起語率を上回っている（残りの 1 箇所は同じ値）。このことは、新規の内容になった最初の段落は、新しい話題を展開させるため、その次の段落との結束性が高くなっていると言えるのではないだろうか。

逆に矢印の直前の段落は、あるまとまりの最後の段落を意味する。この部分の後続段落と前接段落の共起語率はどうなっているかというと、9 箇所中 6 箇所で前接段落との共起語率の値のほうが高い。これは一つの例にすぎないが、このような文章中での共起語率の推移を利用して段落のまとまりを自動的に推測することに応用出来る可能性がある。

７．まとめと今後の課題

本発表では非常に単純な指標である共起語率を用いて文章の結束性の度合いを観察した。その結果、法律、白書、国会会議録のように結束性の高い文章と新聞、ベストセラー、雑誌のように結束性の低い文章があることが分かった。NDC 別に観察したデータでは、文学の結束性が低いという結果になった。これは文学に会話文が多く、その会話が 1 段落と認定されているというデータの特徴の現れである。

また、文章中の共起語率の推移をみることにより文章のセグメンテーションへの応用が考えられることを示した。

今後の課題として以下の 3 点を挙げる。これらを通じて文章における結束性について客観的な記述を目指したい。

(1)西部ほか(2011:232)によると、サンプルを構成する文がすべて段落に分割される訳でないと指摘されている。また、<paragraph>の認定は行頭の空白をもとに自動的に認定しているとのことなので段落の実態を確認して分析に問題がないかどうか確認する必要がある。

(2)段落と文の両方を利用した結束性の測定の方法を探る。

(3)指示詞や接続詞など文法的結束性の手段との相関を調べること。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「テキストにおける語彙の分布と文章構造」による研究成果の一部である。データとして利用した BCCCWJ の書籍部分は、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

Halliday, M.A.K. and Hasan, R.(1976) *Cohesion in English*. Longman（邦訳『テクストはどのように構成されるか』、大修館書店、1997 刊）

Hoey,Michael.(1991) *Patterns of Lexis in Text.* Oxford University Press.

Károly,Krisztina.(2002) *Lexical Repetition in Text.* Peter Lang.

庵功雄(2007)『日本語におけるテキストの結束性の研究』、くろしお出版

小木曽智信、間淵洋子、前川喜久雄(2011)「『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き XML フォーマット」、言語処理学会第 17 回年次大会予稿集、pp.352-355.

西部みちる、大島一、間淵洋子、小林正行、田島孝治、高田智和、山口昌也(2011)『『現代日本語書き言葉均衡コーパス』における電子化テキストの構築』、国立国語研究所内部報告書(LR-CCG-10-03)

水谷静夫(1980)「用語類似度による歌謡曲仕分『湯の町エレジー』『上海帰りのリル』及びその周辺」「計量国語学」12(4)、pp.145-161.

口頭発表（3）

3月6日（火）　10:00〜12:00

多様な様式を網羅した会話コーパスの共有化

伝 康晴 (千葉大学文学部/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)

Sharing of Conversation Corpora That Cover Diverse Styles and Settings

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

1. はじめに

近年、書き言葉コーパスの構築は飛躍的な発展を見せている。国立国語研究所では、1 億語を超える規模の『現代日本語書き言葉均衡コーパス』[*1]を開発し、さらに 100 億語を超える規模の Web コーパスの開発を目指している。これに対して、話し言葉コーパスは、音声収録・転記など開発初期段階での負担が大きく、学会講演や模擬講演などの独話を中心とする『日本語話し言葉コーパス』[*2]を除いて、大規模なものは存在しない。とくに日常の言語行動の中心である会話に関しては、個々の研究プロジェクトごとに小規模なデータを独自に収集・利用している状態を脱していない。

これに対する一つの解決策として、既存の会話コーパスの共有化という方式に着目する。小規模データを所有する研究機関は多くあり、それらは音声収録・転記の段階を終え、負担の大きい初期のハードルをクリアしている。これらのコーパスを共有すれば、研究に利用できる会話データの量は従来よりも飛躍的に増加する。しかし、これらのコーパスでは、転記方式は不統一であり、また、韻律情報や発話機能など会話研究に必要な基本情報は必ずしも完備していない。そこで、本研究では、これらの基本情報に関する共通のアノテーションを施し、相互利用可能な形で会話コーパスを共有する方法を考案する。

本稿では、上記の目標を達成するために立案したプロジェクトの概要、および、転記方式に関する予備的な調査結果について述べる。

2. プロジェクトの概要

1 節の目標を達成するため、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）を開始した。本プロジェクトは、2.1 の研究組織のもと、2.2 の研究計画で遂行する。プロジェクトの進捗状況

[†] den@cogsci.l.chiba-u.ac.jp
[*1] http://www.ninjal.ac.jp/kotonoha/
[*2] http://www.ninjal.ac.jp/csj/

表 1　プロジェクト内で共有可能な会話コーパス（予定）。各コーパスの出典については付録 A を参照

名称	参与者数	関係性	様式	内容
千葉大 3 人会話	3 人	友人	対面	雑談
CSJ	2 人	初対面（一方はプロ）	対面	インタビュー
FGI	4 人	初対面 + プロ/アマチュア	対面	インタビュー
言語接触場面 3 人会話	3 人	知人（非母語話者 1 名）	対面	雑談
新聞販売店会話	2 人	店員と顧客	非対面	ビジネスコール
早稲田大自由対話	2 人	友人（ゼミ配属前/後）	対面/非対面	雑談
JPN	2〜3 人	友人/家族	対面	雑談
作業療法会話	主に 2 人	療法士とクライエント	対面	作業療法
宇都宮大音声対話	2 人	友人	非対面	課題指向
三重大地図課題対話	2 人	知人/初対面	対面	課題指向
タングラムパズル対話	2 人	知人/初対面	対面	課題指向
ロゴ積み木対話	2 人	知人/初対面	対面	課題指向
北大 2 人会話	2 人	先輩後輩（非母語話者 1 名）	対面	雑談

および研究成果は、「会話コーパス」プロジェクトホームページ[*3]から随時発信していく。プロジェクト終了時には以下のものを公開する予定である。

- 基本情報の仕様をまとめたマニュアル
- 共有コーパスに付与されたアノテーション[*4]

2.1　研究組織

研究組織は著者たち以外に 9 名の共同研究者・研究協力者からなり、その専門領域は会話分析・談話分析・日本語教育学・日本語学・認知心理学・音声言語情報処理など多岐にわたる（一覧はプロジェクトホームページを参照）。いずれも会話データをみずから収集し、研究に利用している研究者たちであり、本プロジェクト内でコーパスを共有し、共有化に伴う問題点を解決し、共有化の利点を検証するという趣旨に賛同していただいている。

2.2　研究計画

本プロジェクトでは、1 節の目標を達成するため以下のことを行なう。

2.2.1　会話コーパスの調査

会話の諸現象の普遍性と多様性をとらえるために、参与者数・関係性・様式・内容などがさまざまに異なるコーパス群を対象とする。そのためまず、各メンバーが所有するデータ（表 1）を対象として、転記方式や音質・付加情報などを調査し、共有化に伴う問題点を洗い出す。現在、転記方式に関する予備的な調査を進めており、その結果について 3 節で報告する。

2.2.2　共通の基本情報の仕様策定

2.2.1 の調査に基づき、共通に付与できる基本情報（表 2）の仕様を策定し、マニュアルとしてまとめる。

[*3] http://www.jdri.org/kaiwa/

[*4] 公開の可否および形態は各コーパスごとに異なる。

表 2　共通に付与する基本情報（予定）

基本情報	説明	予想される問題点
転記	発話内容の文字化	非流暢性のマークアップや基本単位の不統一
形態論情報	単語分割・品詞	自動解析の可否
韻律情報	発話末音調など	音質によっては主観的付与に限定される
発話機能	談話行為・宛て先など	従来は課題指向対話がおもな対象
局所構造	隣接ペア・発話交換構造	多人数会話への拡張
連鎖構造	先行連鎖・修復連鎖など	これまで明確に策定された基準はない

2.2.3　共通の基本情報の付与

2.2.2 で策定された仕様に基づき、各コーパスに基本情報を付与する。作成したデータはサーバ上で管理し、メンバー間で共有する。

2.2.4　共有コーパスの基礎的分析

2.2.3 の共有コーパスを用いて、各メンバーがこれまで行なってきた研究テーマ（話者交替・あいづち・連鎖構造・成員カテゴリー化など）に関する基礎的分析を行なう。とくに、様式が異なるコーパス間の普遍性と多様性を明らかにする。これによって、多様な様式の会話コーパスを共有することの利点を検証する。

2.2.5　プロジェクト外の会話コーパスの調査

本プロジェクトのメンバーが所有する以外の会話コーパスについて、2.2.1 と同様の調査を行ない、本手法で共有化できるデータがどれくらいあり、どの程度の多様性を網羅できるか調査する。これによって、より大規模な会話コーパスを設計する際の指針とする。

3.　転記方式に関する調査

研究計画の 2.2.1 の最初のステップとして、各メンバーが利用しているコーパスの転記テキストの断片（数分程度）を収集し、さまざまな観点から比較した。表 3 にその概略を示す。また、いくつかのコーパスにおける転記テキストの例を付録 B に示す。

以下、項目ごとに調査結果の概要を述べる。

3.1　転記単位

転記テキストをどのような単位で分割するかについて、CSJ のように明確に定めている（200 ミリ秒以上の無音もしくは明示的な文末表現で分割）場合もあるが、多くのコーパスでは転記単位は不明確であった。転記単位の調査は今後進めたい。

3.2　レイアウト

転記テキストのレイアウトは、多くの場合、1 行に（ないし数行にわたって）1 つの発話単位を記し、発話単位ごとに行を重ねる一般的な書式であったが、中には、参与者ごとに列を区切り、複数の話者が同時に産出した発話単位を同じ行に記すことでタイミングの同時性を表わせるように工夫しているものもあった（付録 B：言語接触場面 3 人会話）。

表 3　転記方式の比較

コーパス	時間情報	文字表記	転記基準
千葉大 3 人会話	単位開始/終了時間	漢字かな混じり	CSJ 方式（簡略版）
CSJ	単位開始/終了時間	基本形・発音形併記	CSJ 方式
FGI	単位開始/終了時間	漢字かな混じり	独自方式
言語接触場面 3 人会話	なし	漢字かな混じり	独自方式
新聞販売店会話	単位内・単位間休止	漢字かな混じり	会話分析方式
早稲田大自由対話	単位開始/終了時間	基本形・発音形併記	CSJ 方式（簡略版）
JPN	単位間休止	ローマ字	Santa Barbara 方式
作業療法会話	なし	漢字かな混じり	独自方式
宇都宮大音声対話	単位開始/終了時間・単位内休止	基本形・発音形併記	談話タグ WG 方式
三重大地図課題対話	単位開始/終了時間・単位内休止	ひらがな	千葉大地図課題方式
タングラムパズル対話	単位開始/終了時間	ひらがな	独自方式
ロゴ積み木対話	単位内・単位間休止	漢字かな混じり	会話分析方式
北大 2 人会話	単位内休止・単位間休止	漢字かな混じり	会話分析方式

コーパス	非言語音	非流暢性	音調	重複位置
千葉大 3 人会話	笑	フィラー・語断片・延伸	（別ファイル）	なし
CSJ	笑・咳・息	フィラー・語断片・延伸	（別ファイル）	なし
FGI	笑	なし	上昇	相づちのみ
言語接触場面 3 人会話	笑・咳	延伸	上昇	あり
新聞販売店会話	笑・咳・息	語中断・延伸	上昇・下降・継続	あり
早稲田大自由対話	笑	なし	なし	なし
JPN	笑	語断片・延伸	上昇・下降・継続	あり
作業療法会話	笑	延伸	上昇	なし
宇都宮大音声対話	笑・息	フィラー・語断片	なし	なし
三重大地図課題対話	笑	なし	上昇	あり
タングラムパズル対話	笑	延伸	なし	あり
ロゴ積み木対話	笑	語中断・延伸	上昇・下降・継続	あり
北大 2 人会話	笑・息	語中断・延伸	上昇	あり

3.3　時間情報

いくつかのコーパスでは、転記単位ごとに開始・終了時間が与えられていた（付録 B : CSJ）。また、開始・終了時間が与えられていない場合でも、発話単位内・単位間に生じた休止の長さを秒や記号（`(.)`：短い休止）で記しているものがあった（付録 B：新聞販売店会話）。

3.4　文字表記

漢字仮名混じりによる表記が半数を占めていたが、実際に発音された音列を仮名で併記しているもの（付録 B : CSJ）や、ローマ字や仮名のみで表記しているものもあった（付録 B : JPN）。

3.5　転記基準

公刊されている転記基準に準拠しているものと、独自の方式で転記しているものがあった。公刊されている転記基準としては、CSJ 方式 (小磯他 2006) に準拠したもの（付録 B : CSJ）と

会話分析方式 (Jefferson 2004) に準拠したもの（付録 B：新聞販売店会話）が多く見られた。これらはともに非言語音の転記や非流暢性に関する豊富なマークアップを規定しているが、その表現の仕方はまったく異なる。

3.6 非言語音の転記の有無

多くのコーパスでは、参与者の笑い声の転記が行なわれていた。咳や息については、省略しているものが多かった。

3.7 非流暢性のマークアップの有無

音の延伸（たとえば「夕刊のほう:::」）は多くのコーパスで特別な記号によりマークアップされていた。フィラー（「あの」「えーと」など）や語断片（語としての形をなしていない音列）は CSJ 方式に基づく一部のコーパスのみでマークアップされていた。会話分析方式では、語が途中で中断された場合のみ（「え-」など）マークアップされていた。

一般に、音の延伸以外はマークアップされていないコーパスが多かった。ただし、マークアップはなくとも、転記としては記されている場合がほとんどであった。

3.8 音調のマークアップの有無

多くのコーパスは上昇調を転記しており、上矢印（↑）（付録 B：言語接触場面 3 人会話）やクエスチョンマーク（?）（付録 B：新聞販売店会話、JPN）などの記号を用いていた。下降・継続の音調をマークアップしているものは、会話分析方式に準拠した一部のコーパスに限られた。ただし、他のコーパスの中にも、韻律情報アノテーションとして別ファイルで音調の情報を与えているものもあった。

どの範囲の音調を転記テキスト中にマークアップするのが利便性がよいか、調査する必要がある。

3.9 重複位置のマークアップの有無

会話分析方式を中心に、複数の話者の発話の重複をマークアップしているものが多く見られた。一方、CSJ 方式に準拠したコーパスでは重複位置はマークアップされていなかった。これらのコーパスでは、単語（千葉大 3 人会話）や音韻（CSJ）ごとに開始・終了時間が与えられており、それらの情報から自動的に重複位置をマークアップすることが原理的には可能である。

重複位置のマークアップ方法もさまざまで、重複箇所や範囲を記号のみで示す方法（付録 B：言語接触場面 3 人会話）と、記号によるマークアップとともに、字下げによって発話開始位置をそろえ視覚的によりわかりやすくする方法（付録 B：新聞販売店会話、JPN）があった。

3.10 まとめ

以上のように、転記方式はコーパスごとにさまざまであるが、CSJ 方式や会話分析方式を中心にいくつかのグループにまとめられそうである。今後、これらのグループ内での細部の差異の調査と、これらのグループを超えた共通の転記基準の策定が可能かなど、調査を進めたい。

4. おわりに

本稿では、既存の会話コーパスの共有化を目標とする研究プロジェクトの概要、および、転記方式に関する予備的な調査結果について述べた。本プロジェクトによって、既存の会話コーパスの共有化の可能性・有効性が示されれば、今後、より広い範囲でコーパス共有を試みたい。同時に、既存のコーパスでは網羅できない様式や内容に関しては新規のデータ収集も必要となろう。将来的には、これらのことを目標に、大規模会話コーパスの構築へとつなげたい。

謝辞　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。調査に協力していただいたメンバー各位に感謝する。

参考文献

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

小磯花絵・西川賢哉・間淵洋子 (2006).「転記テキスト」　『国立国語研究所報告書 124: 日本語話し言葉コーパスの構築法』　pp. 23–132.

関連 URL

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

付録 A. コーパスの出典

※公刊物がない場合は問い合わせ先を記載

千葉大 3 人会話

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.

CSJ

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.

FGI

Morimoto, Ikuyo, Kana Suzuki, Etsuo Mizukami, and Hiroko Otsuka (2008). "Categorization in Japanese group discussion: Its advantages and disadvantages." *Proceedings of the 15th World Congress of Applied Linguistics*. Essen, Germany.

言語接触場面 3 人会話

大場美和子 (2011).「内的場面と接触場面における三者自由会話への参加の調整—談話・情報・言語ホストの役割の分析—」 博士論文（未公刊）, 千葉大学大学院人文社会科学研究科.

新聞販売店会話

Suzuki, Kana (2010). "Other-initiated repair in Japanese: Accomplishing mutual understanding in conversation." Unpublished doctoral dissertation, Graduate School of Intercultural Studies, Kobe University.

早稲田大自由対話

菊池英明（早稲田大学人間科学学術院）

JPN

大野剛（カナダ・アルバータ大学)、鈴木亮子（慶應義塾大学経済学部）

作業療法会話

長岡千賀（京都大学こころの未来研究センター）

宇都宮大音声対話

Mori, Hiroki, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya (2011). "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics." *Speech Communication*, 53, pp. 36–50.

三重大地図課題対話

吉田悦子 (2002).「日本語名称なし地図課題対話コーパスの概要と転記テキストの作成: 報告」 人文論叢（三重大学人文学部文化学科研究紀要）, 19, pp. 241–249.

タングラムパズル対話

吉田悦子（三重大学人文学部）

ロゴ積み木対話

谷村緑、吉田悦子、竹内和広 (2009).「課題遂行対話におけるグラウンディング成立の記述方法の検討—日本人英語学習者の場合」 社会言語科学会第 23 回大会論文集, pp. 162–165.

北大 2 人会話

山本真理（北海道大学大学院国際広報メディア・観光学院）

付録 B. 転記テキストの例

言語接触場面 3 人会話

行	NS1	NNS	NS2
17			あじゃ相当寒いーんです
18	/ん/		よね↑/2 月って一番
19		[[んー]]	寒いんですか [ねやっぱ]
20	2 月とか寒かった		
21		2 月は一寒かった	
22		ちょっと雪が降ってました	

CSJ

```
0147 00252.842-00255.979 R:
ほんの                                & ホンノ<H>
三十分も                              & サンジュップンモ
いなかったと                          & イナカッ(? タ)ト
思うんですけど                        & オモウンデスケド<H>
0148 00255.410-00256.211 L:
(F うーん)                            & (F <VN>)
0149 00256.272-00257.115 R:
乗り換えで                            & ノリカエデ
0150 0150 00257.922-00259.109 L:
やっぱり                              & ヤッパリ<H>
(F そのー)                            & (F ソノー)
```

新聞販売店会話

```
 8 C    え：：：と日経が↑間違って入＜ってます＞。
 9 A    あ：：：＞そうですk-＜<<↑夕刊のほう：：[：です：ね？＝
10 C                                          [はい。      ＝はい。
11 A    え：と、(0.3)  ↑朝日の↓ほう：：：
12      (.)
13 C    はい、そ [うです。
14 A             [↓゜はい゜、入れればえ-（.） [よろしいんですね？
15 C                                          [はい
16      ↓はい。＝
```

JPN

```
216 H: .. jibun dake <X un X>,
217    yasunde,
218    warui na,
219    tte no mo,
220    an ja nai [no]?
221 R:           [un].
222    ... [2 sore de  2],
223 H:     [2 soo ieba 2],
224    goorudenuiiku da <X mon X> na=.
```

通時コーパスをどう使うか

近藤泰弘（青山学院大学/国立国語研究所）†

How to Use the Historical Corpus

Yasuhiro Kondo (Aoyama Gakuin University / NINJAL)

1　はじめに

我々のプロジェクトにおいては、日本語の史的研究に用いることができる本格的な「通時コーパス」を構築する準備段階として，コーパスの設計にかかわる諸問題について研究している。

- コーパスの対象に含める文献資料をどのようにして選定するか
- 選定した資料をどのように電子化しどのような情報を付与するか
- 古典テキストに対応した形態素解析をどのように行うか

など、通時コーパス設計のための重要問題を中心に、基礎的な研究を展開している。こうした研究は、日本語史上のいくつかの時点の主要資料についてコーパスを試作し、これを活用した日本語史研究を実践することを通して行う。また、コーパスの構築作業における他機関との連携の可能性を探り、コーパス公開のために不可欠な著作権処理の問題についての検討も行い、通時コーパスの構築・公開に向けた諸課題に見通しを付けることを目的としている。

言語資源研究系の現代語コーパスにかかわる研究と連携を取り、コーパス開発センターで実施中の現代語コーパスの構築作業、著作権処理業務などとも関連付けて研究を進めている。プロジェクトの Web サイトは以下の URL である。

http://historicalcorpus.jp

2　通時コーパスの構造

基本的には、XML によるマークアップを施した電子化テキストである。従来の同様な試みとしては、国文学研究資料館の KOKIN ルールおよび SGML によるマークアップを施したコーパスがあるが、今回研究をしているコーパスにおいては、記述的マークアップをさらに言語的な要素にまで及ぼすことにしている。

- XML (eXtensive Markup Language) によるマークアップ

† yhkondo@cl.aoyama.ac.jp

- NINJAL BCCWJ (Balanced Corpus of Contemporary Written Japanese) と極力互換性がある
- 全文コーパス
- UTF-8 Encoding
- 形態素単位のマークアップ (自動的な形態論的解析)
- SUW (Short Unit Word) （短単位）

記述的マークアップが必要となる。具体的には次のようなものとなる。少なくとも、形態論的構造まではすべての収録テキストについてマークアップをする予定であるが、統語論的構造や意味論的構造についてはどの程度可能になるかまだ未定である。

1. 論理構造 cf. title, paragraph, citation, kokka-taikan-number
2. 表記構造 cf. ruby, page
3. 形態論的構造 cf. word, part of speech, inflection
4. 統語論的構造 cf. sentence, clause, phrase,
5. 意味論的構造 cf. Agent, Object,

2.1 XML タグセット

XML のタグセットおよびアトリビュートについてもまだ未定の点が多い。現在は、BCCWJ に習っている点も多いが、今後、古典語としての特質を調査しつつ、適切なタグセットを検討していきたい。

1. sample 文書
2. div　　内部構造
3. p 同上
4. pb Page Break
5. note 頭注
6. ruby ルビ
7. sentence
8. SUW　短単位

2.2 XML アトリビュート

1. (sample) ID, no, title, filename, etc.
2. (SUW) orthToken（出現書字形）、lForm（仮名形）、lemma（語彙素）、pos（品詞）、Form（原形）、PronToken (出現発音形)、w Type(語種)、start（開始文字位置）,end（終了文字

位置)、cType（活用型)、cType(活用形)、orderID（単語出現順番号）

3　マークアップの例（竹取物語）

これは、『竹取物語』の冒頭のマークアップ例である。前節で示した仮のタグセットによってマークアップしたものであるが、だいたいのイメージとしてとらえていただきたい。

¡?xml version=”1.0” encoding=”UTF-8”?¿
¡sample sampleID=”1201_竹取物語” no=”1201” title=”竹取物語”
fileName=”1201 竹取物語_100728”¿
¡div id=”00000001”¿¡div type=”古典本文”¿¡p org=”空 1”¿¡sentence¿
¡SUW orthToken=”　” lForm=”” lemma=”　” lemmaID=”23”
kana=”” pos=”空白” Form=”” pronToken=”” wType=”記号”
start=”10” end=”20” orderID=”10” BOS=”True” /¿
¡note org=”1” text=”1”¿¡/note¿
¡SUW orthToken=”いま” lForm=”イマ” lemma=”今” lemmaID=”2460”
kana=”イマ” pos=”名詞-普通名詞-副詞可能” Form=”イマ” pronToken=”イマ” wType=”和”
start=”20” end=”40” orderID=”20” /¿いま
¡SUW orthToken=”は” lForm=”ハ” lemma=”は” lemmaID=”29321”
kana=”ハ” pos=”助 詞-係 助 詞” Form=”ハ” pronToken=”ワ” wType=”和” start=”40” end=”50”
orderID=”30” /¿は
¡SUW orthToken=”むかし” lForm=”ムカシ” lemma=”昔” lemmaID=”37012”
kana=”ム カ シ” pos=”名 詞-普 通 名 詞-副 詞 可 能” Form=”ム カ シ” pronToken=”ム カ シ” wType=”和”
start=”50” end=”80” orderID=”40” /¿むかし

4　利用方法

BCCWJ のコーパスブラウザである「中納言」による利用を考えている。「中納言」では次のような検索が可能である。古典語の場合は、文法的内省が働かないため、特に「形態論情報を利用した検索」がひじょうに有効である。

1. 形態論情報を利用した検索が可能
2. 短単位検索、長単位検索、文字列検索の機能がある
3. 多数の用例が見つかった場合でも、その全体をダウンロードできる
4. 検索語の前後合わせて最大 10 単位まで条件が指定できる

5. 文脈を長めに（最大前後各 500 語まで）表示することができる

5　研究の方法

5.1　平安時代の「て」節の様相

平安時代語の複文は、従属節と主節との区分をすることが厳密には難しく、節が次々と連なっていく、いわゆる「節連鎖」的な特徴を持っていることが言える。その中でも「て」接続助詞による接続は、連鎖的な機能が強く、そのつながり方について詳しく調査することが必要である。

一般に、係助詞は、連用修飾から接続助詞まで広くその後に用いることができるのに対し、副助詞にはその分布がより狭い。また、発表者が論証したように、副助詞は次の二種類に分類できる（1 種・２種の命名は、小柳智一 (1998) による）。

- 第 1 種　ばかり・まで　格助詞に前接・形容詞連用形に後接しない・副助詞に前接
- 第 2 種　のみ・さへ・だに　格助詞に後接・形容詞連用形に後接・副助詞に後接

このうち第 1 種は、連用修飾や格助詞の後に接続しないため、今回の研究からは当面除外して考えることができる。第 2 種については、いわゆる連用修飾や格助詞、そして、接続助詞の一部に至るまで接続することができるのであるが、それが具体的に「て」助詞との関係においてどのような分布になるかは従来わかっていなかった。発表者は、中古語、特に『源氏物語』を資料として、次のようなことを論証した。

接続助詞「て」と言われるものの中には、現代語と同じように、二種類のものがあり、ひとつは、従属節の従属度としては、いわゆる A 類に属するものである。意味的には「付帯状況」を示す。もうひとつはいわゆる B 類に属すると思われるものである。意味的には「継起」「原因／理由」「並列」を示す。そして、A 類に属するものは、第 2 種副助詞を後接することができるのに対し、B 類に属するものは、第 2 種副助詞を後接することが不可能であった。したがって、第 2 種副助詞を後接している文型を調査することで、A 類の「て」の類型を知ることが可能になる。

- （A 類）
  - –（付帯状況）おぼつかなくてのみ年月の過ぐるなむあはれなりける（源氏物語・若菜下）
  - –（付帯状況）おぼししづみてのみおはするを（夜の寝覚）
- （B 類）
  - –（継起）まことに明け方になりてぞ、宮帰り給ふ（源氏物語・梅枝）
  - –（原因理由）道にてやまひしてなむ、死にける。（大和物語）
  - –（並列）知らぬ国に吹き寄せられて、鬼のやうなるもの出で来て殺さんとしき。（竹取物語）

5.2　平安時代の「て」節と疑似分裂文

今回、コーパスによってさらに精密に「て」節を観察することができた。

現代語では、A 類の「て」節は疑似分裂文とすることができるが、B 類の「て」節はそれが不可能なことが知られている（内丸 2006)。

- （A 類）
  - –（付帯状況）太郎はよそ見をして車を運転していた。
  - – 太郎が運転していたのは、よそ見をしてだ。（疑似分裂文）
- （B 類）
  - –（並列）おじいさんは山に芝刈りに行って、おばあさんは川に洗濯に行った。
  - – *おばあさんが川に洗濯に行ったのは、おじいさんが山に芝刈りに行ってだ。（疑似分裂文）
  - –（継起）電車を降りて、改札を抜けた。
  - – ？改札を抜けたのは、電車を降りてだ。（疑似分裂文）
  - –（原因）台風が近づいて、学校が休みになった。
  - – ？学校が休みになったのは、台風が近づいてだ。（疑似分裂文）

平安時代語の場合の「て」節を「なり」が受ける疑似分裂文を調査してみると、先に A 類として分類したものがほとんどをしめる。

- あえかに見えたまひしも、かく長かるまじくてなり。（源氏物語・夕顔）

ただし、原因理由を表す B 類かと思われる「寄りて」だけは疑似分裂文となることができる。

- （よりてなり）
  - 夏虫の身をいたづらになす事もひとつおもひによりてなり（古今集）
  - おぼろげの願によりてにやあらむ（土佐日記）
  - 苦しきによりてにや。（枕草子）

この「よりて」は、表面的には原因理由を示すように読めるが、平安時代語の段階では、「寄る」という動詞の意味がまだかなり強く、付帯状況を示しているという可能性が高いのではないか。

- さらに許されぬによりてなむ、かく思ひ嘆き侍る。（竹取物語）

6　従属節の従属度と副助詞・係助詞

さて、以上調査してきたように、A 類には、副助詞・係助詞が後接し、B 類には係助詞のみが後接する。また、ここには記述しなかったが C 類には副助詞はもちろん、係助詞も後接しない。この

ようにA類からC類にむかって階層的な秩序があるが、個々の接続助詞にはそれぞれ別な制約もあるようである。たとえば、B類のうち、順接の「已然形＋ば」や「未然形＋ば」には係助詞が後接する。それに対し逆接の「も」「とも」には副助詞はもちろん係助詞も接続しない。この制約は統語的なものではなく、意味的なものだろう。

全体としては、おおよそ次のようになる。

類	形態	接続
A類	て（A類）・ながら・ず・で・用言連用形	副助詞・係助詞後接
B1類 （順接）	て（類）・ば・（く）は・つつ・ず・で・用言連用形	係助詞後接
B2類（逆接)	とも・ども・ものの	どちらも後接せず
C類	を・に・が	どちらも後接せず

このように「て」のA類・B類を正確に分離して記述することで、平安時代語の従属節全体についての見通しも明らかになってくるのであり、コーパスを駆使することで記述の精度を格段に上げることが可能になるのである。

文献

[1] 内丸裕佳子 (2006)「動詞のテ形を伴う節の統語構造について一付加構造と等位構造との対立を中心に一」(『日本語の研究』2巻1号)

[2] 小柳智一 (1998)「中古の「ノミ」について一存在単質性の副助詞一」(『國學院雑誌』99巻7号)

[3] 近藤泰弘 (2012)「平安時代語の接続助詞「て」の様相」(『国語と国文学』89巻2号)

[4] 安永尚志 (1989)『日本古典文学作品本文データベース仕様書（第二版)』(国文学研究資料館)

[5] 安永尚志編 (1998)『講座　人文科学研究のための情報処理　[第3巻　テキスト処理編]』(尚学社)

[6] 安永尚志 (1998)『国文学とコンピュータ』(勉誠社)

通時コーパスと言語空間論[*1]

山元啓史（東京工業大学/カリフォルニア大学サンディエゴ校）
田中牧郎（国立国語研究所言語資源研究系）
近藤泰弘（青山学院大学/国立国語研究所言語資源研究系）

Diachronic Corpus and Linguistic Space

Hilofumi Yamamoto (Tokyo Institute of Technology/University of California, San Diego)
Makiro Tanaka (Dept. Corpus Studies, NINJAL)
Yasu-Hiro Kondo (Aoyama Gakuin University/Dept. Corpus Studies, NINJAL)

1 はじめに

国立国語研究所共同研究基幹型プロジェクト「通時コーパスの設計」[*2]は古代から近世までのいくつかの時点における代表的資料により、「通時コーパス」のモデルを試作するものである。そこでは、主に１）資料の選定、２）古典本文の電子化と情報（異文・原文表記・異体字・引用・文体など）付与の問題、３）各時代・各文体に対応した形態素解析などの実務的あるいは技術的問題を扱ってきた。

本稿は、同プロジェクトを通して培われた基本的な概念、共時態と通時態、記述言語学とコーパスの関係、可能性と作業モデル、適応例、通時コーパスを活用する上で、今後、必要性が予想される研究領域について議論する。

2 言語の記述と言語の空間

コーパス開発には、テキスト収集、著作権、電子化、提示情報の決定などの実務的な作業が多く、研究アプローチに関する議論が後回しにされやすい。コーパスの開発は手段であり、それを用いて言語の普遍的形式を探求することが目的である。しかし、コーパス言語学においてはその理論的背景となる「言語の記述」「言語の普遍性」「言語の空間」「共時態と通時態」などの点について、あまり議論されてこなかった。この機会を利用し、これらについて検討し、考え方を整理しておきたい。

2.1 共時態と通時態

Saussure (1983) によれば、通時態とは時間の流れにしたがって、変化していく言語のありさまであり、共時態とは言語の一定時期におけるありさまである。つまり、言語をある時の点と見るか、ある時とある時をつなぐ線と見るかということである[*3]。数理的に整

[*1] 本研究は国立国語研究所共同研究プロジェクト基幹型「通時コーパスの設計」（代表者：近藤泰弘）、および、科学研究費基盤研究C「和歌形態素解析用辞書開発のための用語連接規則に関する基礎研究」（代表者：山元啓史）の助成を得た。

[*2] http://historicalcorpus.jp/

[*3] したがって、現代語のみを取り上げた研究であっても、経年的な変化を問題とした分析であれば、たとえ、時間の幅が短くても、それは通時的分析となる。時間的な変化を無視し、時間的な変化はないものとするならば、共時的分析となる。日本語の研究においては、一時代一言語を分析の対象とする共時的分析を行い、それらの記述をもとに、時代あるいは言語の間を紡ぐ通時的研究が進められる (服部, 1980, p. 249) とい

理すれば、共時態を示す点の集まりが連続して線形に見える時、それは通時態と考えられる。ひとつひとつの点は静的である一方、その連続した線が見せる軌跡（線の内容）は動的である。この動的なありさまを見るには、共時態の層を幾層にも並べ、各層の差分をとり、その差分を層間の変化量として分析するのである。その際のポイントとしては、体系の差を強調するだけでなく、体系の背後にある共通の原理をも抽出することが重要である（図 1）。こうすることにより、従来、現代語だけあるいは古典語だけを分析していたのでは見えなかったものが見えてくるものと考える (近藤, 2000, p. 80)。

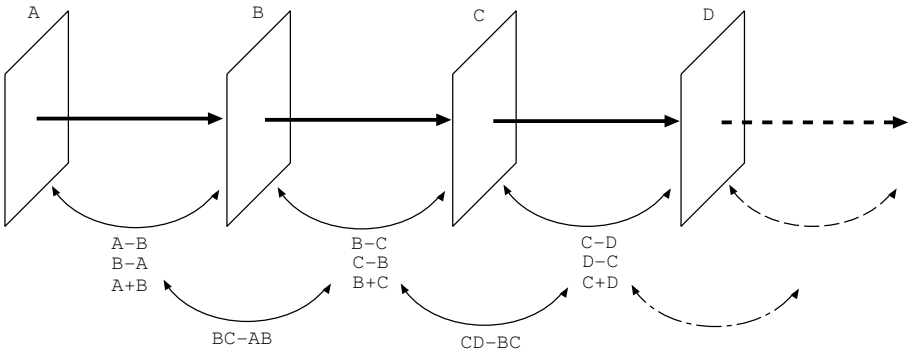


図 1　共時態の各層から差分をとる：たとえば、ＡＢＣＤは時間軸に並べられた任意の資料。差分をとるだけでなく、両者の体系に共通の原理を抽出し、その抽出したものをさらに隣接の抽出したものと比較していく。

2.2　通時コーパスと言語の記述

われわれの脳裏に潜在する langue（言語構造）は、直接観察できず、言語処理の結果として出力された音声や文字列、parole（言語現象）によってしか分析ができない (Saussure, 1983)。そのため、言語現象からさまざまな推論を働かせて、分析する必要がある上、古典語の場合には、その当時の注釈書や古辞書など多角的な資料を利用して分析を進めなければならない。また、現代人には古典語についての内省がなく、その分析には現代語からの推論を利用するため、現代に生きる観察者の知識に依存する要素も多く、観察者の間に差違が認められ、どの基準によれば普遍的な姿としてとらえられるのか、なかなか決定できない[*4]。

一方、コーパスとコンピュータ処理によれば、従来、研究者がテキストを主観的に見て観察していた研究方法から、あらかじめ研究者が設定した客観的なルール（仮説）がテキスト全体にわたって（網羅的に）当てはまるかどうかを徹底的に調べ上げる方法にかわる。これは見ればわかるテキストを、あえて見ないことにより、現代のわれわれには通常認知できないタイプのデータの構造性や規則性を厳しく探り出すことができる。それにより、内省に代わる感知の機構を手に入れることができるのである (近藤, 2001, p. 35)[*5]。

われているが、服部 (1980, p. 230) によると、「上代から現在に至るまでの日本語の変遷史の研究は組織的に行われてきていない」とある。おそらく現時点においてもその事情は同じであろう。

[*4] 服部 (1980, p. 249) は「我々の言語活動は、我々の脳裏に潜在すると推定される langue に支配されているために、或種の特徴がそこに繰り返し現れるものと考えられる。しかしながら、このような langue は、現在の所外部観察することができず、内部観察もほとんど不可能である」と述べている。

[*5] たとえば、近藤 (2000, pp. 301–11) は大量言語処理による観察を通して、内省だけではとらえにくい「のが」「ことが」によって示される名詞節の性質を明らかにし、その記述に成功している。

テキスト自身は静的な、ある時点で表現された言語の現象である。動的な姿をとらえるには、任意のテキストを多重に比較し、その変化量を分析しなければならない。そこで、比較の計画が重要になる。通時コーパスでいえば、図 1 に示すように作品 A〜D を連続したものとしてとらえる仕組みが必要となる。語彙（語種）の場合なら、A–B, B–C, C–D のように互いの 2 者間をとりもつシソーラス（ブリッジシソーラス）を作成し、シソーラス間の差分をとり、変化量とする。文法を明らかにしたいなら、シソーラスの代わりに連接の出現パターンの一覧表（ブリッジテーブル）を作成し、テーブル間の差分を分析することになろう[*6]。いずれの場合も、いったん共時態の間をつなぎ通した上で、時間軸でとらえた動的なマトマリを記述していく[*7]。こうすれば、それぞれの要素は目で見てわかる状態となり、どの段階でどのような要素が変化したのかが考察できよう。

さて、時間軸を紡ぎ、内省を網羅的大量処理で補完することによって準備ができたら、つぎに必要なことは何だろう。おそらく結果をどのように出力するかということであろう。

2.3 言語空間論

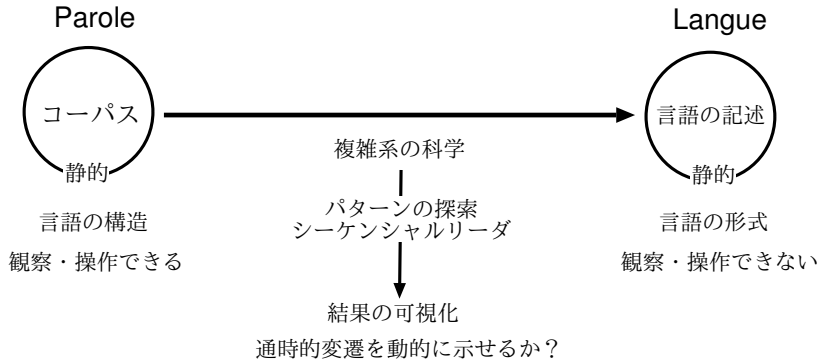


図 2　コーパスと記述、langue と parole: 一般的に記述されたものは静的ではあるが、言語の存在自体は常に変わりつづける動的なものである。その動的な記述はどうすればできるのであろうか。言語の要素はさまざまなものからなり、コーパスにて観察できる姿は複雑多岐にわたる要素が絡み合った現象である。

言語が固定的に見えるのは資料の上だけであって、実際の言語は絶えず揺れ動いている。このような動的なふるまいを明らかにするには、静的な複数の資料を比較し、その差分を分析することによって、動的な変化を見ることは前に述べた。データから得られる動的な変化を分析することにより、過去から現在に至る普遍性だけでなく、さらに将来への予測も期待できる。

言語は時間によって変化するだけでなく、時間の経過につれてヴァリエーションが生じ、それが地域に伝播することによって、言語の地域差が生じる。「地域」は地理的な一要因に過ぎず、文化や産業などさまざまな場面においてもそれぞれのヴァリエーションが生じる。「時間」とともに「空間」が成長し、その中で言語は揺れ動き、変化しつづける。言語の分析として、従来の静的な記述だけでなく、動的な振る舞い（ある部分は枝分かれ

[*6] 後でも述べるが、作品ごとで形態素解析辞書を作成し、その辞書差分をとってもよい。しかし、形態素解析辞書による場合は、隣接パターンに限られるだろう。

[*7] もう少し細かくいえば、資料間の差分がを最も小さくなるようにするために、資料の多重比較を行い、最も近く隣接する資料の間を線で結んでみて、だいたいなめらかな線となるようにモデルを作るのである。

をし、代謝を繰り返し、一部は成長し、またある部分は衰退していく様子）が記述でき、近い将来の言語のふるまいも動的に記述できるはずである。

言語の科学性に関する議論はほとんどの言語学の入門書にあるが、そのいくつかは言語を生物の営みに喩えている。言語の進化や普遍性を考えるために、ここでは生物学での成果を利用することを考える。たとえば、Dong and Searls (1994) は言語学の句構造規則を用いると、遠く離れたところであっても、ある 2 組の DNA のパターンが引き合い、ある細胞を構成する要素になると説明している。

2.4　シーケンシャルリーダ、コーパスロボット

生物学（遺伝子）では、DNA の 4 つのアミノ酸の配列の並びを調べるコンピュータプログラムが数多く公開されている (Dong and Searls, 1994)。これと同じ原理で、コーパスの文字列に見られる言語パターンを調べるコンピュータプログラムを考えてみる。これは研究者の仮説にもとづいて、コーパスの文字列を行き来しながら、何回でも瞬時に仮説を検証することができる機械（コーパスロボット）である[*8]。

開発途上あるいは更新中のコーパスに対応させるには、コーパスの進化に応じた動的なパーシングシステムを考慮しておけばよいだろう。その理論の構築および内部仕様の決定は重要かつ慎重に行われるべきであろうが、同時に実に楽しくなりそうな仕事でもある。コーパスの完成を待たなくても、テキストが質・量ともに充実するにしたがって、コーパスロボットを使って、仮説を立てては何回でも瞬時に検証する仕組みができるならば、言語研究者にとって、それはパワフルで魅力的なものになるにちがいない。次節では、そのための差分の要素（あるいは微分？の要素）はどのように整理すればよいかについて考えてみたい。

3　差分の方法

さまざまな局面で差分を抽出する方法が考えられようが、ここでは任意 2 つのテキストの間の差分をとる方法について簡単に述べる。従来のテキストの比較における問題点のひとつは、テキスト間の内容的な異なり（話の内容）と、言語的な異なり（用語の流行り廃り）を理論的に区別できていないことである。たとえば、ある原作に対する 2 つの翻訳間や原作と現代語訳の間の比較分析を行った論文には、調べたいことが言語の変化にあるのか、翻訳の過程でそうせざるをえなくなったのか、が曖昧になることが多い。これは分析を始める前から研究の枠組みとして区別されていないのである。

田中 (2011) は今昔物語集とその典拠との対応（日本霊異記／宇治拾遺物語）の中で系列比較モデルによる漢語と和語の比較分析を行っている。中国の仏典（法苑珠林：漢文）と今昔物語（和漢混交）を比較すると、漢語がそのまま取り入れられていたり、まったく受け入れられずに捨てられていたり、少々形を変えて、取り入れられていたりなどして、最終的に今昔という形でまとまる。一方、今昔と宇治拾遺物語（ほぼ同じストーリーで和文体）を比較すると、和語がそのまま取り入れられていたり、まったく受け入れられずに捨てられていたり、少々形を変えて、取り入れられていたりなどして、これまた結果的に今昔としてまとめられる。いずれの場合も、2 作品の差から、1）言語の変化により今昔では適宜変更が加えられたもの、2）翻訳者が何らかの基準でことばを取捨選択したもの（翻訳態度）の 2 種が分類できる。従来の比較研究では、このような言語変化による要因

[*8] パターンの探索の点でいえば、DNA の研究がアイデアの発端となるかもしれないが、文字列の数理文法の点でいえば、水谷 (1982, 1983, 2005) などの研究によるところが大きい。

と翻訳者の取捨選択による要因の区別が研究の計画時から一貫しておらず、言語変化のつもりで取り出したデータの中に明らかに翻訳者の操作によるものが紛れ込んでしまうことがあった。上記を研究の枠組みとして、弁別・整理できるように構成したものが系列比較モデルである[*9]。

図 3 は、任意点の時間軸上にある資料を比較する方法を示したモデルである。説明の都合上、任意の 2 点間の違いに限って説明するが、分析の対象は 2 点に限らなくてもよい。A と A' は同じ系列[*10]の言語資料である。A が発生した時を t_1、A' が発生した時を t_2 とする。A と A' の関係は、ある時代の源氏物語の写本とそののちの時代の同作品の写本としてもよいし、1990 年代のプロ野球の実況中継録音と 2000 年代のそれとしてもよい。対応の程度は A と A' の内容をどう捉えるかによる。

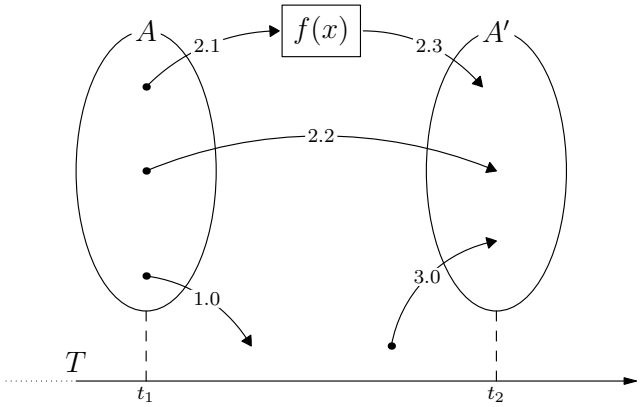


図 3　系列比較のための変遷要素の差分モデル: A は t_1 の時に発生した、あるまとまりを持った内容、A' は t_2 時に発生した、A に対応するまとまりを持った内容。T は時間軸。$f(x)$ は A の任意の要素 x を A' の要素とするための関数。

A に含まれる要素が A' に含まれないことがある。これを 1 類 (1.0) と呼ぶことにする。逆に、A に含まれない要素が A' に含まれることがある。これを 3 類 (3.0) と呼ぶことにする。A にも A' にも含まれる要素がある。これを 2 類と呼ぶことにする。2 類にはまったく同じ要素が A と A' に含まれるものと、A と A' で対応する要素に多少の変動が認められるものがある。前者を 2.2 系とし、後者のうち、A に見られるものを 2.1 系、A' に見られるものを 2.3 系とする。A から A' への時間経過において、取り除かれる要素（1.0)、継続されるが変換の必要な要素（2.1)、そのまま継続される要素（2.2)、継続される際に変換された要素（2.3)、新たに加えられた要素（3.0）の 5 区分で要素の分類を行う。

$$A = \{1.0, 2.1, 2.2\} \tag{1}$$
$$A' = \{2.2, 2.3, 3.0\} \tag{2}$$

A に含まれる要素は、1.0, 2.1, 2.2 の 3 種類、A' に含まれる要素は、2.2, 2.3, 3.0 の 3 種類である。t_1 と t_2 が限りなく近い場合、t_1 と t_2 は共時の資料として捉えてもよい。その場合には、2.x（同じ語、類を同じくする語）が主となり、1.0 や 3.0 の要素が減少する

[*9] 田中 (2011) の例は時間軸というよりも翻訳間という要素で説明した方がよいので、むしろ付録に示す図 4 にしたがうべきである。ただし、これにはまったく時間軸の要素がないわけではない。

[*10] 幅がある概念としておく。狭くは同一内容、原文と翻訳、広くは同じジャンルとする。後に説明する共時モデルにも関わってくる。

が、同一テキストの物理的なコピーや同一内容の録音資料のダビングでもない限り、それらがなくなるわけではない。共時と捉えて分析する資料であったとしても、別々に発生したテキストであれば、時間のずれ（ごく微量な時間経過を伴う要素あるいは単なる言い換え）がある。一方、t_1 と t_2 が限りなく離れている場合には、2.2 が少なくなり、1.0 や 3.0 の要素が増加することが予想されよう。言語の変遷をぜひ見てみたいと思う研究者にとっては、２類 (2.1, 2.2, 2.3) の各要素の分布は注目に値することだろう。

この方法の問題点としては、今、分析している要素が、次の時代に用いられなくなった要素 (1.0) なのか、あるいは何らかの変換が施されて引き続き用いられている要素 (2.1, 2.3) なのか、この両者の判別がつきにくくなることである。しかしながら、表記や読みに関わるマイナーな違いのみを２類の 2.1 あるいは 2.3 とするような厳しい基準であったとしても、この方法で、新たにわかることは多いのではないかと考えている。本稿では、共時的な比較については詳説はしないが、付録にて通時軸を共時軸に置き換えたモデル（図 4)、共時軸を横軸にして通時と共時を一緒に示したモデル（図 5）も紹介しておく。

この系列比較モデルにしたがって、語の弁別に十分利用できるシソーラスを用いたコーパスロボット（任意の t_1 と t_2 における語彙を５区分に自動的に分類するマシン）が作成できれば、近い将来、言語変化の諸相を明らかにしてくれるのではないかと期待している[*11]。

4　今後の研究領域

ここまでの方法を実現するには、必要不可欠な課題がいくつか考えられる。第 1 に処理の単位を柔軟に考えることである。単語の定義は永遠の課題であり、未だ作業的便宜として取り扱われており、確かな理論に基づいて行われているものではない。しかし、これは日本語の問題だけでなくどの言語においても問題とされており、言語学全般に関わる問題である[*12]。語には長いもの短いものさまざまがあるが、短い語は不安定で、長い形で用いられる傾向がある。たとえば、和語については「和語を語形の長さの面からみると、まず『目、葉、荷』のような一拍の語は短すぎて安定しにくく、『はっぱ』『にもつ』のように長い形になって落ち着くばあいもある (西尾, 2002, p. 80)」ということである。また、短い語は多義であるが、長くなればなるほど意味が限定される傾向がある。このことから、おそらく文脈に即してノビチヂミする機構を何らかの方法で開発しなければならないのであろう。しかし、現在のところその有力な手立てはない。

第 2 に形態素解析辞書を一度は資料（作品）別に作って、その資料にとって最も効率のよい辞書を作成し、作品毎の差分・共通を割り出してみることである。それぞれの辞書の連接確率を読むことによって、syntagmatic な側面が、またそれぞれの辞書の語彙差分を読むことによって、paradigmatic な側面が、通時的に把握できるのではないかと考えている。

第 3 に資料（作品）の間を取り持つシソーラスの整備が必要である。系列比較モデルにおいて最も必要なのはシソーラスである。あるトークンが他のトークンと同じであるかどうかを認定するための語彙表が必要なのである[*13]。残念なことに、現状のシソーラスで

[*11] A と A' を系列を同じくする内容を持つものとしたが、共時における系列を異にする A と A' とを比較する際も、上記５区分で系列の異なりを分析することができよう。

[*12] 宮島 (1994, p. 113) は「日本語の語い調査でいちばんこまることは、『単語』という単位が確立していないことである」と述べている。

[*13] 実際には同じトークンであるかどうかを認定することを目標にすると失敗する気がする。生物の世界でも

は、同じ時代の任意の 2 作品（共時的資料）を比べるにも、異なる時代の 2 作品（通時的資料）を比べるにも、単語の長さがまちまちであったり、表記がさまざま（漢字仮名、送り仮名）であったり、上位概念の単語（たとえば花）で、下位概念の単語（たとえば、桜）を示していたりしていて、なかなか 2 作品（資料）の比較が行えない。しかも、単語は時の経過とともに意味を変えることがしばしばあり (Lyons, 1987, p. 212)、作業は困難を極める。意味を決定した符号（たとえば、従来の注釈書で見られるような詳細な意味記述）でデータを処理してしまったら、その符号によってすべての研究の結果は支配されてしまう。したがって、意味的にもきわめて中立的なシソーラスを考案しなければならない。

5　おわりに

本稿では、通時コーパスプロジェクトを進める上での基本的な概念の整理とコーパス言語学の枠組みについて議論し、いくつかのモデルを提案した。しかし、考えれば考えるほど、なさねばならないことは積みあがるばかりである。本稿で取り上げたもの以外にも、音韻、文字、語種など、さまざまな問題があるが、それらに関する通時コーパス流のアプローチは別の機会に考えさせていただきたい。

参考文献

Dong, Shan and David B. Searls (1994) “Gene Structure Prediction by Linguistic Methods”, *Genomics*, Vol. 23, pp. 540–551.

服部四郎 (1980)『言語の本質と機能』，第 1 巻，日本の言語学，第 1 章，大修館書店．

近藤泰弘 (2000)『日本語記述文法の理論』，ひつじ書房，東京．

——— (2001)「コンピュータによる文学語学研究にできること—古典語の「内省」を求めて—」,『文学・語学』，第 171 巻，34–43 頁．特集平成 13 年度夏季大会シンポジウム．

Lyons, John (1987)『Language and Linguistics (言語と言語学)』，岩波書店，第 7 版．

宮島達夫 (1994)『語彙論研究』，麦書房，東京．

水谷静夫 (1982)『数理言語学 (現代数学レクチャーズ D- 3)』，培風館．

——— (1983)『語彙』，第 2 巻，朝倉日本語新講座，朝倉書店，第 1 版．

——— (2005)『言語と数学 POD 版』，森北出版，第 POD 版．

西尾寅弥 (2002)「語種」,『語彙・意味』，第 4 巻，朝倉日本語講座，朝倉書店，第 1 版，79–109 頁．

Saussure, Ferdinand de (1983) *Course in general linguistics...*: McGraw-Hill. tr. of Cours de linguistique generale. from the French by Bally, Charles and Sechehaye, Albert.

田中牧郎 (2011)「平安時代末期における語彙の文体的変異: 同文説話の語彙比較を通して」．第 99 回国語語彙史研究会, 於: 大阪大学．

人間の世界でもそうであるようにまったく同じものは 2 つとして存在しないのであり、厳格に分析するとどこかで違うと判定されてしまうだろう。むしろ、任意 2 つのトークンが同じ類に属するかどうか判別する仕組みを作るべきである。

付録

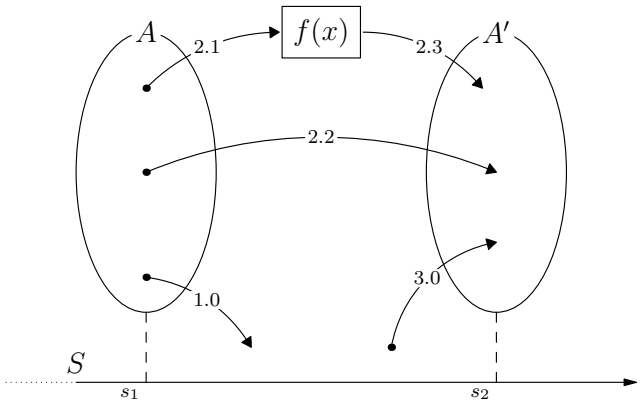


図 4　系列比較モデル（共時）: 通時のモデルの時間軸 T を共時軸 S にしただけである。ただし、T は時間しか表さないが、共時軸 S は、同じ時に発生した同じテキストの異なる言い方や文化、翻訳、方言など、さまざまな場合が考えられる。

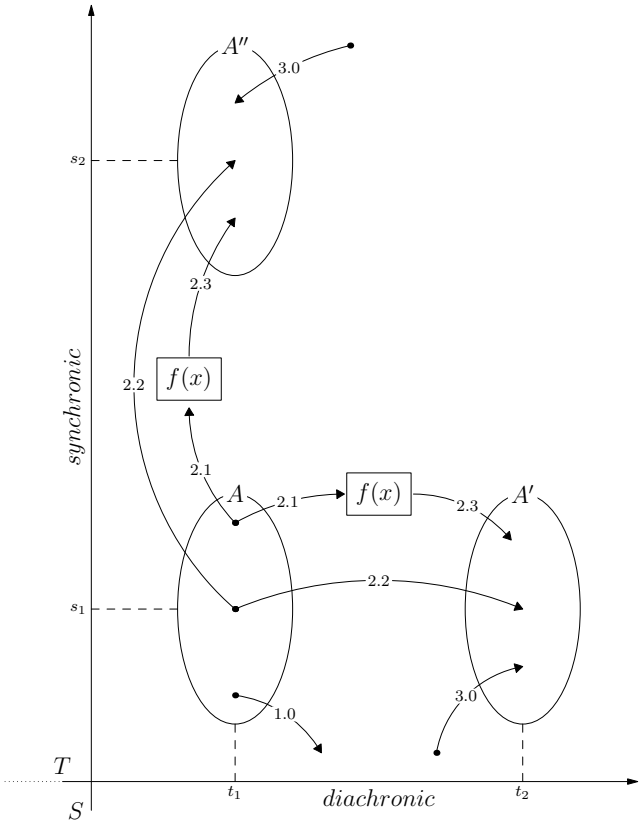


図 5　系列比較モデル（共時／通時）: 縦軸が共時 (synchronic)、横軸が通時 (diachronic)。共時と考えられる関係であっても時間の幅を持つ要素が含まれることもある。

近代語史をとらえるための文献選定とコーパス

田中　牧郎（国立国語研究所言語資源研究系）†

Material Selection and Corpus Compilation for Historical Study of the Modern Japanese

Makiro Tanaka (Dept. Corpus Studies, NINJAL)

１．はじめに

本稿は、平成 21 年度から行っている国立国語研究所の共同研究プロジェクト「近代語コーパス設計のための文献言語研究」で研究している内容のうち、近代語のコーパスの対象にする文献資料をどのように選ぶかという課題について考えるものである。

２．近代語のコーパスへ

２．１　近代語コーパスの位置

『現代日本語書き言葉均衡コーパス』が完成したが、日本語研究のためのコーパスはこれで十分というわけではもちろんなく、国立国語研究所においてはすでに、「超大規模コーパス」「通時コーパス」など新たなコーパスの設計が始まっている。

『現代日本語書き言葉均衡コーパス』（1 億語以上）は、明治時代から現代に至る近現代日本語の全体を把握するためのコーパス群（KOTONOHA）の最重要構成要素と位置付けられて開発が始まった（前川 2008）。KOTONOHA を構成する国立国語研究所のコーパスとしては、『日本語話し言葉コーパス』（約 750 万語、2004 年公開）、『太陽コーパス』（約 750 万語、2005 年公開）が先行している。現代共通語話者の独話を対象とした『日本語話し言葉コーパス』は、東京工業大学・情報通信研究機構と協力して構築されたもので、工学的応用の側面も色濃く反映した設計になっており、この性質は『現代日本語書き言葉均衡コーパス』にも継承され、さらに「超大規模コーパス」へと受け継がれていくことが見込まれる。明治後期から大正期の総合雑誌『太陽』を対象とした『太陽コーパス』は、国立国語研究所に従来あった近代語研究や史的国語辞典編集といった文献言語研究の系譜の中から生まれたものであり、その側面はやはり『現代日本語書き言葉均衡コーパス』に受け継がれており、さらに「通時コーパス」へと連なっていくものと考えられる。

「通時コーパス」の設計については、平成 21 年度から新規に始まった「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘客員教授）において扱われているが、そこでは奈良時代から江戸時代までが対象とされており、明治時代以後『現代日本語書き言葉均衡コーパス』までをつなぐコーパスとしても、近代語コーパスの重要性は高い。

２．２　『太陽コーパス』

近代語のコーパスとして公開済みの『太陽コーパス』（国立国語研究所 2005a）について概観しておこう[1]。『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、月刊の総合雑誌『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、その全文（著作権処理ができなかった記事を除く）を対象にしたものである。年次が 6 年または 8 年刻み

† mtanaka@ninjal.ac.jp

[1] 同種のコーパスに、国立国語研究所『近代女性雑誌コーパス』があり、CD-ROM で公開している（http://www2.ninjal.ac.jp/lrc/）。これは、『太陽コーパス』とほぼ同時期の女性を読者とした 3 誌（『女学雑誌』『女学世界』『婦人倶楽部』）を対象とした約 120 万語の小規模なコーパスである。

となっている点はサンプリングコーパスと言えるが、対象になった年次の全体を含んでいる点では全文コーパスとも言える。

コーパスの重要な要件のひとつである代表性の担保については、対象とした総合雑誌『太陽』が、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は格別の価値を持っていることに、根拠を置いている（田中 2005）。例えば、図１は、『太陽コーパス』のジャンル（NDC）別の記事数とその比率を『現代日本語書き言葉均衡コーパス』（出版サブコーパスの書籍、図２）のサンプル数（丸山ほか 2011）と比較できる形で示したものであるが、社会科学が最も多く、文学がこれに次ぐところなど、『現代日本語書き言葉均衡コーパス』（出版サブコーパス書籍）と『太陽コーパス』は似ている面があることが分かるだろう。しかし、明治後期から大正期の書き言葉の広がりを母集団として把握した上で『太陽』の代表性を検証して設計したものではなく、一資料のみを対象としたコーパスにとどまる限界は否めない。

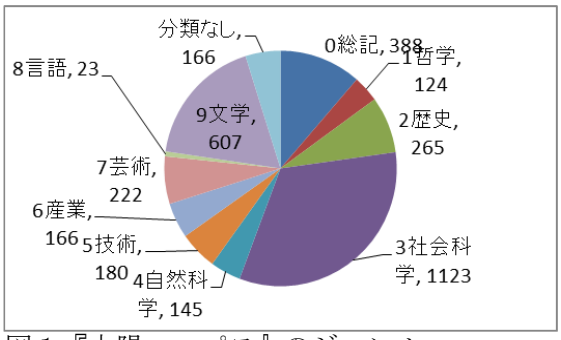


図１『太陽コーパス』のジャンル

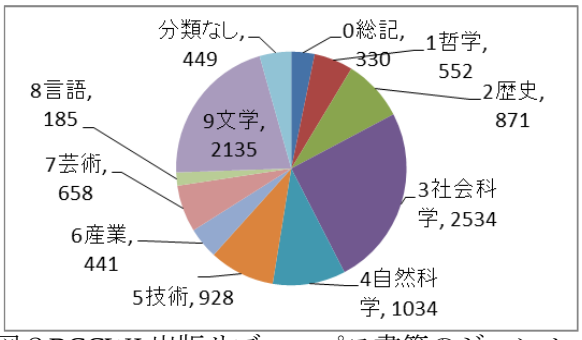


図２BCCWJ 出版サブコーパス書籍のジャンル

コーパスの要件としてやはり重要な、言語研究に役立つ付加情報については、総合雑誌という特性を生かせるように、記事や引用などの範囲をマークアップし、そこにジャンル・文体・著者（話者）などの情報を属性として埋め込んだ構造化テキストを実現し、校訂注記や異体字情報などもタグを付与して豊富に表現した（田中 2005）。あわせて、それらの情報を自在に利用できるように、『ひまわり』（山口 2005）、『プリズム』『たんぽぽ』（小木曽 2005）などの検索ツール群を開発した。このように、『太陽コーパス』は、文献資料を対象としたコーパスとしては画期的なものであったが、『日本語話し言葉コーパス』や『現代日本語書き言葉均衡コーパス』に付与されている形態論情報が全く付与されていないなど、不十分なところも残されていた。

『太陽コーパス』は公開後約７年が経過したが、ほぼ同時期に公開した『日本語話し言葉コーパス』に比較すると、これを利用した研究成果は、必ずしも多くない。これは、上述した代表性や付加情報の不十分さに起因するほか、明治後期から大正期の約 30 年間だけを切り取っていて近代語史の中で浮いた存在になっていることにもよっている。前後の時代をも対象に加えて『太陽コーパス』を相対化し、近代語史の全体がとらえられる近代語のコーパスを設計し構築を始めることが求められている。

３．近代語の文献リストの作成

３．１　「国語辞典編集準備資料」

先に『太陽コーパス』は国立国語研究所の史的国語辞典編集の系譜から生まれたと述べたが、その史的国語辞典編集を行う準備研究のために設置された国語辞典編集準備室によって、用例採集の対象とすべき近代語資料をまとめた目録が、三つ作成されている。

(1)『用例採集のための主要文学作品目録』（国語辞典編集準備資料２、1980 年）
主要文学全集に収録された、明治元（1868）年～昭和 41（1966）年の 1506 作品をリスト化したもので、有識者 10 名が投票を行い得点化し、高得点の主要作品 139 点が「用語索引を作る作品」として選定されている。

(2)『用例採集のための主要雑誌目録』(国語辞典編集準備資料３、1983 年)
国立国会図書館の和雑誌目録の中から、昭和 25(1950)年以前に創刊され 20 年間以上発行されている雑誌 2778 件をリスト化したもので、有識者 10 名が投票を行い得点化し、高得点の主要誌 120 点が選定されている。
(3)『用例採集のためのベストセラー目録』(国語辞典編集準備資料４、1984 年)
ベストセラーに関する参考書に掲載された、明治元(1868)年～昭和 53(1978)年の書籍、1882 件をリスト化したもの。このリストについては得点化や主要作品の選定は、行われていない

実際の史的国語辞典編集のための用例採集事業は紙媒体で開始されたが、すべての用語・用例を採集できるようにする「総索引方式」と、任意の用語・用例を選抜して採集する「スカウト式」の二段構えで着手された。総索引方式では国定国語教科書を対象とした『国定読本用語総覧』(国立国語研究所 1985-1997 として完成公開)が作成され[2]、スカウト式では雑誌『太陽』の用例採集が進められた。ところが、この事業に本格的にコンピュータが導入されたことがきっかけとなって、『太陽』は途中からスカウト式を止めコーパス化の対象にされ、『太陽コーパス』が作成されたのである(この間の経緯は、木村・加藤・田中 1999 参照)。『太陽コーパス』の完成に先立って史的国語辞典編集のための用例採集作業は中断された形になっているが、実質的にはコーパス構築事業にその考え方は継承されており、平成 21 年度から通時コーパスと近代語コーパスの設計に関わるプロジェクトが同時に始まったことで、その側面はより色濃くなってきたと言える。近代語コーパスに含めるべき文献を検討する際に、上記の目録類は第一に参考にすべきものである。

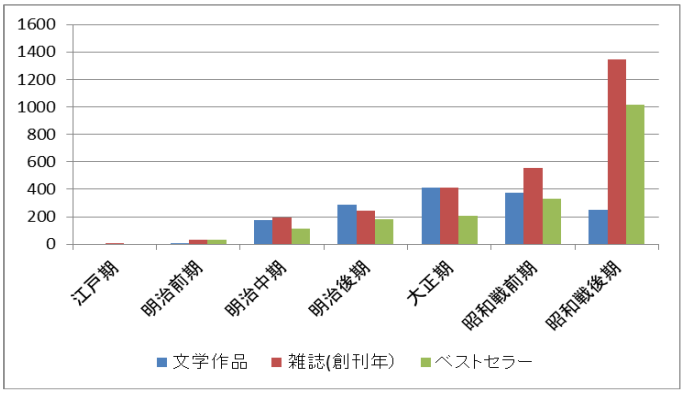


図３　国語辞典編集準備資料に掲載された文献数(時代別)

図３は、上記の三つの目録に掲載された文献の数を時代別にまとめたものである。時代区分は、明治から大正期をほぼ 15 年ごとに４つに区切り、昭和期を戦前と戦後に分けた

明治前期：明治 1～15(1868-1882)年　　明治中期：明治 16～30(1883-1897)年
明治後期：明治 31～44(1898-1911)年　　大正期：大正 1～14(1912-1925)年
昭和戦前期：昭和 1～20(1926-1945)年　昭和戦後期：昭和 21(1946)年～

明治・大正期と昭和期とで時間幅が異なっていて比較しにくい面はあるが、雑誌とベストセラーは時代を追って増加傾向にあり、文学作品は大正期まで増加し、昭和期に入って減少していると見ることができよう。こうした傾向はそれぞれの媒体が各時代にどの程度の量発行されたかという実態を反映している面もあるかもしれないが、直接的には目録作成の材料に何が使われたかということを反映しているのではないかと思われれる。また、明治前期・中期が全般的に少ないのは、この目録作成が 20 世紀を主たる対象にしていたということも関係しよう。

[2]教科書については資料目録は作成されていない。国定読本の他には国定算数教科書の用語索引が作られたが、公開されてはいない(木村・加藤・田中 1999)。

雑誌とベストセラーは、『現代日本語書き言葉均衡コーパス』でも対象としており、文学作品は『現代日本語書き言葉均衡コーパス』では書籍の下位に NDC 分類に即して配置されている。『現代日本語書き言葉均衡コーパス』にはこのほか、新聞、教科書、白書、広報誌、Yahoo!知恵袋、Yahoo!ブログ、法律、国会会議録などが含まれている。このうち、新聞、教科書、国会会議録などは、史的国語辞典編集のための文献目録作成は行われていないが、用例採集作業の対象として研究は行われており、対象文献の候補にはなっていた。一方、白書、広報誌という媒体は、昭和戦前期までは存在しておらず、Yahoo!知恵袋、yahoo!ブログのようなインターネット上の文章もまた同様である。しかし、政府や役所から国民や住民に告知する文書は戦前にもあり、知恵袋やブログを私的性格の強い文章と考えれば、手紙や日記など近代から存在していた媒体は多い。近代語コーパスの対象に含めるべき文献の候補は、さらに幅を広げて検討していくことが望まれよう。

３．２　叢書類

国語辞典編集準備資料の目録３冊は、近代語コーパスに含めるべき文献を考えるのにきわめて有益な資料であるが、不十分なところも多いため、他の材料を用いて増補していくことが必要である。特に、明治前期の文献の手薄さが目立つため、この時期の文献を豊富におさめる叢書類をもとに文献リストを増補していくことにした。用いた叢書は次の４つである。

(1) 明治文化全集　全 24 巻（1927～1932 年、日本評論社）
(2) 明治文化資料叢書　全 12 巻（1959～1963 年、風間書房）
(3) 日本近代思想大系　全 24 巻（1988～1992 年、岩波書店）
(4) 新日本古典文学大系 明治編　全 30 巻　（2001 年～刊行中、既刊 29 巻、岩波書店）

これらの叢書は、言語研究を目的として編纂されたものではないが、文化・思想・文学の分野の重要文献が選ばれていると考えられ、そこには、言語資料としても価値の高いものが含まれていると思われる。

表１　叢書類に収録される文献の数（時代別）

	江戸期	明治前期	明治中期	明治後期	計
明治文化全集	16	265	196	16	493
明治文化資料叢書	2	20	50	39	111
日本近代思想大系	70	959	504	7	1540
新古典大系明治編	1	26	99	14	140
計	89	1270	849	76	2284

表１は、４つの叢書に収録された文献の数を発行された時代別にまとめたものである（発行年代が大正期以後のものや不明のものは集計から除いてある）。明治前期・明治中期に集中しており、国語辞典編集準備資料の目録で不十分だった部分を補うことができよう。

この４つの叢書以外にも、文献リスト増補の材料として有用な叢書や図書目録は色々と考えられるが、まずは、上記の３つの目録と４つの叢書とから作成した文献リストの中身を分析することで、近代語史をとらえるための文献選定をどのように行っていくのがよいかを考えていきたい。ここでは、明治前期・明治中期を例に取り上げたい。

４．文献リストの分類と文献選定の考え方―明治前期・中期を例に―

４．１　文体の観点

4.1.1　文体の流れ

上記の文献リストのうち明治前期・明治中期の部分には、2000 点余りがおさめられてい

る。これについて、文体・ジャンル・媒体の３つの観点から分析を加えていこう。はじめに文体の観点から見る。

言文一致による口語体書き言葉の成立は、近代語史における最重要の出来事のひとつだが、その文体の流れを、森岡（1991）が示す図式をもとにまとめる表２の通りである。明治初期には、文語体も口語体も多様な文体があったが、次第に統合されていき、明治 40 年代には言文一致体という口語体ひとつに統合されていく流れがあった。統合以前に多様に分かれていた文体は、研究者によって様々な分類や名付けがなされており、森岡説はそのひとつである。各文体は連続し交錯し、相互の識別が難しい場合も多い。要点は、近代の文体史は多様性から均質性へという明確な方向性をもっており、まずは文語体・口語体それぞれの内部で統合され、やがて口語体が全体に及んでいき、明治時代のうちにそれが完結するというところにある。文語体の内部、口語体の内部での文体の識別は、その指標が立てにくいが、文語体か口語体かの別については、文末辞を指標として明確に識別することが可能である[3]。

表２　近代語の文体統合の流れ（森岡 1991 に基づき作成）

<table>
<tr><th colspan="2"></th><th>明治初期</th><th>明治 10 年代</th><th>明治 20 年代</th><th>明治 30 年代</th><th>明治 40 年代</th></tr>
<tr><td rowspan="6">実用文系統</td><td rowspan="3">文語体</td><td>漢文訓読体</td><td rowspan="3">和漢折衷体</td><td colspan="2" rowspan="3">明治普通文</td><td rowspan="8">言文一致体</td></tr>
<tr><td>和漢折衷体</td></tr>
<tr><td>候文</td></tr>
<tr><td rowspan="3">口語体</td><td>問答体</td><td rowspan="3">演説体</td><td rowspan="3">演説体</td><td rowspan="3">初期言文一致体</td></tr>
<tr><td>講述体</td></tr>
<tr><td>談話体</td></tr>
<tr><td rowspan="2">文学系統</td><td>口語体</td><td>俗文体</td><td>講釈体</td><td>初期口語体</td><td>初期言文一致体</td></tr>
<tr><td>文語体</td><td>和漢折衷体</td><td colspan="2">雅俗折衷体</td><td>（雅俗折衷体）</td></tr>
</table>

4.1.2　文語体と口語体

表３　明治前期・明治中期の文体

	明治前期	明治中期
文語体	1187（93.1%）	773（91.1%）
口語体	31（2.4%）	47（5.5%）
文語体・口語体	3（0.2%）	0（0%）
その他	55（4.3%）	29（3.4%）
計	1276（100%）	849（100%）

表３は、明治前期・中期の 2000 点余りの文献について、文語体か口語体かを認定しその数と比率をまとめたものである[4]。文語体と口語体が混用されているものは、基調をなす文体がどちらであるかによって区別した。「文語体・口語体」と記したのは、両者が同等であ

[3]文末辞が「なり」「たり」「き」「けり」などで終わる文体は文語体、「だ」「である」「た」「です」「ます」などで終わる文体は口語体と識別できる。『太陽コーパス』の文体情報もこの基準で付与してある。

[4]明治前期には国語辞典編集準備資料と叢書類の両方を集計し、明治中期には叢書類のみを集計した。これは、国語辞典編集準備資料が示す文献のすべてを実際に見ることができなかったため、文体が未確認のものが残ったことによる。

るもの、「その他」は漢文や英文あるいは文章でないもの（名簿など）である。明治前期では文語体がほとんどで、明治中期には口語体が数パーセント増加するものの、まだ大部分が文語体である。この時期、文語体が圧倒的に優勢であったことが確かめられる。

4.1.3　文語体

明治前期の文語体を、森岡（1991）は、漢文訓読体、和漢折衷体、候文の３種に分類するが、それぞれ、次のような文体のことを指す。上記の文献リストに含まれるものから１例ずつをあげてみよう。

○漢文訓読体

吾輩日常二三朋友ノ盍簪ニ於テ偶當時治亂盛衰ノ故政治得失ノ跡ナド凡テ世故ニ就テ談論爰ニ及ブ時ハ動モスレバカノ歐洲諸國ト比較スルコトノ多カル中ニ終ニハ彼ノ文明ヲ羨ミ我ガ不開化ヲ歎ジ果テ果テハ人民ノ愚如何トモスルナシト云フコトニ歸シテ亦欷歔長大息ニ堪ザル者アリ

（西周「洋字を以て国語を書するの論」、『明六雑誌』1、1874 年、明六雑誌原本による）

○和漢折衷体

輕重長短善惡是非等ノ字ハ相對シタル考ヨリ生ジタルモノナリ輕アラザレバ重アル可ラズ善アラザレバ惡アル可ラズ故ニ輕トハ重ヨリモ輕シ、善トハ惡ヨリモ善シト云フコトニテ此ト彼ト相對セザレバ輕重善惡ヲ論ズ可ラズ斯ノ如ク相對シテ重ト定リ善ト定リタルモノヲ議論ノ本位ト名ク諺ニ云ク腹ハ脊ニ替ヘ難シ又云ク小ノ虫ヲ殺シテ大ノ虫ヲ助クト

（福沢諭吉『文明論之概略』、1875 年、文明論之概略原本による）

○候文

浜田御預り所村々百姓共、衆訴落印と二つに相分り候に付、今度鶴田御役所より御役人様御上下拾六人、書添村へ御出張に相成、

（津山藩岡熊治郎による監察記録、1868 年、日本近代思想大系による）

候文は文末などに「候う」を伴うもので、文体類型として確立し、この類型に属する文章を特定していくことができるが、漢文訓読体と和漢折衷体との識別は難しい。漢文訓読体に和文や俗文の要素が交じった福沢諭吉の文章などが和漢折衷体の典型とされるが、個々の文章を漢文訓読体と和漢折衷体とに判別する明確な指標は立てることはできない。

4.1.4　口語体

森岡（1991）は、明治前期の口語体には、実用文系統に３種、文学系統に１種あったと見ているが、それぞれ、次のようなものを指すと思われる。やはり、上記の文献リストに含まれるものから例をあげよう。

○問答体の例

開化文明　サア／＼英吉君。是こそ僕が舊宅だ。

西海英吉　ホヽウ成程、茅葺の門長屋、廣庭の植ごみ、こなし部屋から牛部屋の景況、なんとなく古色を帶て、歴然たる舊家の豪農殿が兵衛が宅に來たやうだね。ソシテアノ異な歌を大勢が唱つて居るあれは何ンだね。

（横河秋濤『開化の入り口』、1873 - 1874 年、明治文化全集による）

○講述体の例

世の諺にも｜不治是天福［しらぬがほとけ］と申す通りで、成程世の事國の事も自身に識らざる時は、更に心に掛［かゝ］らずして一向心配することはありますまい。だが、右の如く人間が箇［か］様［やう］に世間の物事を識らずして濟むものでありませう歟［か］。

（植木枝盛『民権自由論』、1879 年、明治文化全集による）

○談話体の例

なぐさみながら、よみあげまする。お経の文句はなにがなんだと、たずねてみたれば、

作州五郡の庄屋がねんらい、あんまりおうきな盗みをしおった。そのしりだん／＼百姓がほりかけ、あちらもこちらも村々さわだち、中々ちよっこりちよっとにゃおさまりませんが、そのわけあらまし申してみふなら、ぬすんだそのかずおふひが中にも、とりわけ大きな事からあげます。

（本多応之助「鶴田騒動の阿呆陀羅経」、1868 年、日本近代思想大系による）

○俗文体の例

モシあなたヱ牛［ぎう］は至［し］極［ごく］高［かう］味［み］でござすネ此［この］肉［にく］がひらけちやアぼたんや紅葉［もみぢ］はくへやせんこんな清［せい］潔［けつ］なものをなぜいままで喰［く］はなかつたのでござうせう

（仮名垣魯文『安愚楽鍋』、1871 年、明治文学全集による）

明治前期の口語体文献は 31 点あるが、それらが上の 4 種の文体のいずれであるかに分類するのは難しい場合も多く、これらの種別は明確な類型としてではなく、口語体の多様な広がりの範囲を考える目安として考えるのが適切であろう。

4.1.5　文献選定における文体の扱い

以上見てきたように、明治前期に多様であった文体について、明確な類型を立てて指標にしたがって個々の文章を分類していくことは困難である。一方、文語体と口語体の識別は文末辞を指標として明確に判別していくことが可能である。したがって、文献選定においては、文語体か口語体かの別については、これを選定の際の判断材料に用いることができるが、それぞれの中の細分類は、材料として採用しにくいと考えられる。

また、明治前期・中期は、口語体の比率はきわめて低いが、それを理由として、当期のコーパスにおける口語体文献の構成比率をぐっと低くするのは適切でないと考えられる。なぜなら、後代にすべての文体を統合していく新しい文体がどのように広がり定着していくかを歴史的に把握するためには、まだ少数派だった初期段階のそれを積極的に採り、その発展過程を研究できるようにしていくべきであるからである。

４．２　ジャンルの観点

ジャンルの枠組みは、『現代日本語書き言葉均衡コーパス』の書籍や、『太陽コーパス』では、NDC（日本十進分類法）が用いられている[5]。上述の文献リストに収録される文献についても図書館に収録されている書籍の場合は、NDC番号が取得できる場合がある。そこで、国立国会図書館の「近代デジタルライブラリー」を検索し、そこに収録されているものにNDC番号を引き当て、明治前期・中期のジャンル分布を図４に表した。

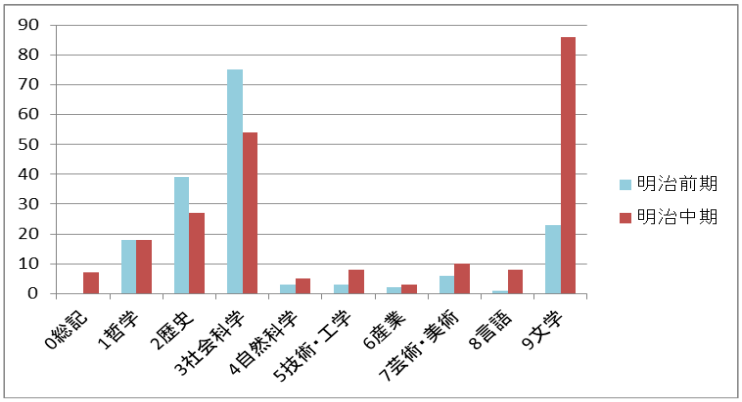


図４　明治前期・中期の文献のジャンル

[5]『現代日本語書き言葉均衡コーパス』では国会図書館の書誌データに付されている NDC 番号を利用したが、『太陽コーパス』ではコーパス作成者が記事を読んで番号を付与した。

明治前期は、社会科学が最も多く歴史がこれに次ぎ、さらに文学、哲学の順に多い。ところが、明治中期では文学が最も多くなっており、社会科学がこれに次ぎ、そして歴史、哲学という順となり、時代的な変容が大きい。これも、時代によるジャンルの多寡の違いが反映している面と、資料とした目録や叢書の性質を反映している面とがあろう。このような大きな変容があるところでは、単純に実際の構成比率にしたがってサンプルの比率を決めることは適切でないように思われる。むしろまずは、文献リストの中身を見ながら、当期の当該ジャンルの文献として重要性の高いものであれば採ることを検討し、そうでなければ別に典拠とすべき叢書や目録がないか検討していくような研究段階が必要であろう。例えば、当期の自然科学や技術・工学の文献はきわめて少ないが、表４のような文献が含まれている。これらの文献を実際に見て、コーパス化の適否を考えていくことが望まれよう。

表４　明治前期の「4 自然科学」「5 技術・工学」の文献（部分）

文献	著者	NDC	文体	西暦	叢書	叢書巻
訓蒙 窮理図解	福沢諭吉	420	文語	1868	日本近代思想大系	科学と技術
物理了案	宇多健齋	420	文語	1880	明治文化全集	科学編
舎密局開講之説	三崎嘯輔	430	文語	1870	明治文化全集	科学編
天変地異	小幡篤次郎	440	文語	1868	明治文化全集	科学編
西洋時計便覧	柳河春三	535	文語	1870	明治文化全集	風俗編
男女普通家政小学	小林義則	590	文語	1880	日本近代思想大系	風俗 性
女房の心得	望月誠	590	文語	1878	日本近代思想大系	風俗 性
服製年中請負仕様書	鈴木篤右衛門	593	文語	1868	明治文化全集	風俗編
西洋料理通	仮名垣魯文	596	文語	1872	明治文化全集	風俗編
通俗男女自衛論	三宅虎太	598	文語	1878	日本近代思想大系	風俗 性

４．３　媒体の観点

文献リストを見ていくと、先に「ジャンル」として設定した NDC とは別の枠組みで分類した方がよいのではないかと思われるものが目につく。例えば、表５に示したものは、明治８（1875）年に発行された新聞・雑誌の一群の一部である。

表５　明治８（1875）年の新聞・雑誌（部分）

文献	著者	NDC	文体	西暦	叢書	叢書巻	出典
評論新聞	海老原穆	—	口語・文語	1875	明治文化全集	雑誌編	
仮名読新聞	—	—	口語	1875	日本近代思想大系	言論とメディア	
萬国叢話	—	—	文語	1875	明治文化全集	雑誌編	
国民気風論	西周	150	文語	1875	日本近代思想大系	天皇と華族	明六雑誌
華士族論	島地黙雷		文語	1875	日本近代思想大系	天皇と華族	共存雑誌
善良なる母を造る説	中村正直	370	文語	1875	日本近代思想大系	教育の体系	明六雑誌
真影の禁を論ず	高木登		文語	1875	日本近代思想大系	天皇と華族	朝野新聞

明治前期に次々に創刊される新聞や雑誌それ自体が叢書におさめられている場合（上の三つ）と、叢書に採られた文献の出典が新聞・雑誌である場合（下四つ）とがある。飛田

（1973）は、新聞・雑誌は、近代に存在する多様な言語資料の性格をすべて合わせもっている「総合資料」という扱いをしており、雑誌『太陽』がそれ単体で代表性を持つと考えて『太陽コーパス』を設計したのも、そのような考え方に立ってのことであった。コーパス作成にあたっては、新聞・雑誌は、その総合性が生きるように、多様な文献をまとめて採集できる資料として扱うのが適切だろう。具体的には、総合性の高い新聞や雑誌をいくつか定め、その新聞や雑誌につついては、等間隔の期間を置く方法などによってサンプリングを行うこが考えられる。『太陽コーパス』と同様の方法である。

新聞・雑誌以外で目を引くのは、教科書、法律、文書の類である。教科書や法律は、『現代日本語書き言葉均衡コーパス』の「特定目的サブコーパス」に採られた枠組みである。文書は、公文書については、同じく白書や広報誌と通じるところがあろう。また、日記や手紙などの一群もあるが、これらのうち私的な性質を持っているものは、同じく Yahoo!知恵袋や Yahoo!ブログと共通する性格があろう。これらは、近代の重要文献として一群をなしているだけではなく、『現代日本語書き言葉均衡コーパス』への接続という点でも重要性の高いものである。こうした NDC によるジャンルとは別に立てることが必要だと思われる分類枠は、広い意味で「媒体」と呼ぶことができるだろう。

なお、上記の文献リストでは目立たなかったが、近代語研究の重要資料には他に、演説や落語などの速記、日本語について記述した文典・辞書などが存在する。速記は、『現代日本語書き言葉均衡コーパス』における国会会議録や『日本語話し言葉コーパス』に対応づけられるものとしても重要であり、明治後期以後には演説や落語の録音資料も存在しており、近代語コーパスに話し言葉資料をどのように取り込むかという課題につながっていこう。また、文典・辞書などは、コーパスの直接の対象にはしにくい面もあるが、コーパスから記述できる近代語の文法や語彙の実態と対照すべき資料として重要性は高いので、コーパス設計時において、その関連づけの方法を検討しておくことも有意義なことだろう。

４．４　その他の観点

上に記した、文体、ジャンル、媒体のほか、ある文献をコーパスに入れるかどうかを検討する際に考慮すべき点が、ほかにも想定される。まず、原本の参照可能性の高さという点である。文献資料に基づく日本語史研究においては、コーパスができれば原本を見なくてもよいということにはおそらくならず、コーパスのもとになった本文がどのような姿であったかを参照したいという要求が研究者には強く存在すると考えられる。そうした要求に応えられるように、コーパス作成と同時に原本の影印や画像などを提供することも考えられるが、現実にはそこに開発コストをかけることは難しい面がある。そこで、複製本が出版されていたり、国立国会図書館などの電子図書館で画像が公開されているものをコーパス化することが考えられる。同じような理由で、本文についての研究成果が反映した校訂本や注釈書類などが整備されている文献も、コーパス化する価値が高いであろう。

次に指摘するのは、コーパスとして用いられる場合でなくとも、文献資料による言語研究一般において、価値が高いとされる文献は、コーパスの対象としても価値が高いという点である。例えば、振り仮名がついているものは語形が確定できる優位性があり、著者の自筆本に基づいているものは別人による改変の心配がないという優位性がある。

以上のような、コーパス化する文献そのものの優位性にかかわる情報も、文献リストに書き入れておき、選定の際の判断材料に使えるようにしておけるとよいと思われる。

５．文献選定の実施に向けて

最後に、以上述べてきたことを踏まえて、近代語コーパスを設計する際に、今後どのようにして文献を選定していけばよいかについて、現段階での見通しを記しておきたい。

(1) 発行年代、媒体、ジャンル、文献の四層を立て、この枠組みで分類しながら文献のリストを増補していく。利用する叢書や目録は、現在手薄となっている媒体やジャンルを中心に、範囲を広げていく。

(2) 第Ⅰ層には時代を立てる。時代区分は5年を一単位とし、明治・大正期は三つの単位をまとめた15年ごとの明治前期・明治中期・明治後期・大正期というまとまりを設定する。昭和戦前期は20年でひとまとまりとし、昭和戦後期も当面分割しない。
(3) 第Ⅱ層に媒体を立て、書籍（初出が雑誌・新聞等のものも含む）、新聞・雑誌、教科書、法律、文書、日記・手紙、速記（会議録を含む）、文典・辞書などを立て、近代語資料として重要なもの、『現代日本語書き言葉均衡コーパス』の媒体と対応付けられるものを生かした枠組みとする。なお、文学作品とベストセラーの目録から収集した文献はまとめて「書籍」に入れる。
(4) 第Ⅲ層にジャンルを立て、書籍はNDCの第1階層を枠組みとし、NDCでは細かすぎる場合は、部分的に統合する。書籍以外は各媒体の性質に応じて枠組みを検討するが、第Ⅲ層が不要な（直下の層が文献である）媒体もある。
(5) 第Ⅳ層は個々の文献とするが、文献リストには、各文献について、発行年、媒体、ジャンル、文献名のほか、著者名、文体、出典、複製本、注釈書、所蔵図書館、表記法、底本の状態等、選定作業において有用と思われる情報を加え、選定作業の判断材料とする。目録における優先候補、叢書における扱いなどもできるだけ書き入れる。
(6) 四つの層による分類を見わたしながら、バランスを考慮して文献選定の考え方を議論し、各層各枠の中で文献に優先順位を付けていく。
(7) 近代語コーパスの開発期間、開発予算、開発手順などが具体化してきたら、文献リストを活用して文献選定案を作成する。

ここに記したことの中には、プロジェクトで十分に議論を行っていない案も交じっているが、このような作業仮説を立てて候補になる文献を実際に見ながら分類し、採否の基準やバランスの取り方を工夫していくことが重要だろう。近代語研究の最大の障壁は資料が多すぎることだと言われることもあるが（湯浅2000）、資料論を重ねながらコーパスを設計することで、その障壁を乗り越えていく道筋も見えてくるのではないだろうか。

文　献

小木曽智信(2005)「構造化テキストを直接利用するアプリケーション―『プリズム』と『たんぽぽ』―」（国立国語研究所2005b所収、pp.83-113）
木村睦子・加藤安彦・田中牧郎(1998)「国語辞典編集のための用例データベース」（『日本語科学』5、国書刊行会、pp.109-127）
国立国語研究所(1985-1997)『国定読本用語総覧』（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）
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）
飛田良文(1973)「近代語研究の資料」（『文学・語学』66、三省堂、pp.45-60）
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.82-94）
丸山岳彦・柏野和佳子・田中牧郎（2011）「第3章 サンプリング」（『現代日本語書き言葉均衡コーパス 利用の手引　第1.0版』、国立国語研究所コーパス開発センター、pp.21-38）
森岡健二(1991)『近代語の成立 文体編』（明治書院）
湯浅茂雄(2000)「近代語研究の要点と課題」（『日本語学』19-11、明治書院、pp.138-148）
山口昌也(2005)「構造化テキストに対応した全文検索システム『ひまわり』」（国立国語研究所2005b所収、pp.49-82）

ポスター発表（2）

3月6日（火） 13:00〜15:00

『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

渡辺美知子（東京大学・特任研究員）†
清水信哉（東京大学・大学院生)‡

The Probability of Fillers at Bunsetsu-Phrase Boundaries in Presentation Speeches of "The Corpus of Spontaneous Japanese"

Michiko Watanabe (The University of Tokyo)
Shinya Shimizu (The University of Tokyo)

１．はじめに

「アノー」や「エート」などのフィラーは，日常会話や講義・講演など即興性のある発話には頻繁に観察されるが，朗読のような，内容も言語表現も予め定められたテキストの音声的表出には稀である。そのため，このような現象は時間的制約のあるオンライン発話生成時のトラブルに関連していると考えられている。すなわち，発話が不自然に長く途切れたり，対話者からの問いにすぐには答えられないようなとき，続く発話を準備中であることを相手に知らせるために話し手はフィラーを発すると考えられている（定延＆田窪，1995)。では，淀みない発話生成を妨げる要因としてどのようなことが考えられるだろうか？これまでの研究では，心理的要因（Christenfeld & Creger, 1996)，発話内容や表現上の選択肢の多少（Christenfeld, 1994)，統語的複雑さ（Clark & Wasow, 1998)，発話内容の質への配慮（小出，1983)，情報へのアクセスのしやすさの度合い（Arnold et.al, 2004）などが取り上げられている。本研究では，このうち，統語的複雑さに着目する。

伝統的な発話生成モデルでは，①メッセージの生成（conceptualizing)，②言語化(formulating)，③発音(articulating)という３つの段階が想定されている（Levelt, 1989)。換言すると，心に浮かんだメッセージを，何から先にどのように話したらよいかを考えながら言語化し，声に出すというプロセスである。このモデルは，メッセージのある部分の言語化が完結して初めて次の部分の処理が始まることを想定しているわけではなく，メッセージの複数の構成部分の処理が，異なるレベルで並行して進行することを想定している。言語化の単位や大きさについては議論があるが，Levelt は，全てのレベルに共通する単位を特定するのは不可能で，処理単位は処理のレベルに依存するとしている。では，各処理レベルにおいて，どのような生成単位が考えられているのだろうか？本稿では，多くの研究が行われている統語上の処理単位に着目する。

英語では，動詞を中心としたまとまりである節（clause）が発話生成の最小単位と主張されている。たとえば Ferreira (2000)は，英語の平叙文の場合は，最初の名詞句と主動詞の言語化が終われば，話者は発話を開始できるとしている。また，Ford（1982）は，200ms 以上のポーズの出現率とその長さに，深層節（動詞１つとその項を含む節）境界と，複数の深層節を含む定形動詞節境界とで差がなかったことから,深層節が発話の生成単位であること，複数の深層節を含む定形動詞節でも，話者は初めに発話全体を計画して話し始めるわけではなく，深層節単位で，逐次，発話を生成していることを主張している。これに対し，Ferreira は，通常は深層節一つより先まで発話を計画して話し始める話者でも，実験環境で早く発話

† watanabe@k.u-tokyo.ac.jp, ‡ s_shimizu@gavo.t.u-tokyo.ac.jp

するようせかされると，発話可能な最小単位が生成できた時点で発話を開始している可能性のあることを指摘している。しかし，Ford の実験の被験者は発話をせかされているわけではなく，この批判は Ford については当てはまらない。一方，自発発話における冠詞，代名詞のくり返しの出現率を調べた Clark & Wasow (1998)は，後続構成素が複雑なほどくり返しの出現率は上昇することを見出し，構成素の複雑さ，言い換えると，言語表現によって担われる情報量の違いは発話生成の負荷に影響するとし，プラニングの負荷に対する後続発話の複雑さの影響を認めない Ford を批判している。

筆者等も，後続構成素の複雑さに反映される情報量が，発話プラニングの負荷，ひいては言い淀みの出現率に影響していると考えた。そして，節境界のフィラーの出現率と後続節の複雑さとの関係を調べた（Watanabe, 2009）。複雑さの指標としては，複雑さと高い相関があるとされている語数を採用した（Wasow, 1997）。その結果，南（1974）の分類（ただし，田窪（1987）による修正を採用）による B 類の従属節の後では，フィラーの出現率と後続節の語数との間に正の相関が見られたが，C 類の後や文境界ではそのような相関は観察されなかった。B 類の従属節は，C 類に比べ，形態的に主節からの独立度が低いが，このことは，内容の独立度も低いことを示唆している。すなわち, B 類の従属節は, 発話している時点で, 続く主節の内容はすでにほぼ決まっている可能性が高い。言い換えると，B 類の従属節の後では，後続節の内容のプラニングの負荷は比較的小さく，主としてメッセージの言語化が行われていると考えられる。一方，C 類の後や文境界では，後続の節や文の自由度は，形式，内容ともに高い。したがって，そのような深い境界では，次のセクションの内容のプラニングが行われている可能性が高い。フィラーの出現率と後続節の複雑さとの対応が，内容プラニングの負荷が小さいと考えられる B 類の境界でのみ観察され，C 類の後や文境界では観察されなかったことから，筆者等は，フィラーは，発話内容の生成というよりは，比較的ローカルな言語化のプロセスに深く関連した現象ではないかと考えている。そこで, 本研究では, 節内における内容のプラニングや言語化の負荷とフィラーの出現率との対応を調べることにした。

「父が描いた絵」と「父が去年フランスで描いた絵」という句を比べると，どちらも，表現しようとする絵のイメージは発話冒頭で既にできていると考えられる。しかし，後者の句の方が言語的に複雑なため, 言語化の負荷は後者の方が高いことが想定される。したがって, フィラーが言語化の負荷に関連した現象であるとすると，句頭や「父が」の後のフィラーの出現率は後者の句の方が高いことが予測される。より一般的には，フィラーの出現率は，まとまった概念を表す構成素の複雑さに対応することが予測される。本研究では，修飾文節から被修飾文節までをまとまりある概念を表す単位と考え, それが複雑なほど, 言い換えると, 修飾文節から被修飾文節までの距離が長いほど，言語化の負荷が大きいためフィラーの出現率は上昇すると予測した。

２．方法

『日本語話し言葉コーパス（CSJ）』コア中の模擬講演 107 講演のデータを分析対象として用いた（国立国語研究所，2006）。CSJ のコアデータでは，節単位ごとに各文節の係先文節が示されている。係り先を持つ文節について，係先文節までの文節数を計測し，これを係り先までの距離とした。その際，フィラーや語断片は計測対象から除外した。そして，係り先を持つ文節を係り先までの距離毎にグループ化し，各グループの文節直後のフィラーの出現率を計測した。

一方，節単位を，1 つまたは複数の節からなる，内容的にまとまりのある単位とすると（丸山他，2006），発話内容のプラニングの負荷は，節単位頭で大きく，終わりに近づくにつれて小さくなることが予測される。そこで，内容プラニングの負荷のフィラーへの影響を探るために，節単位内での文節の位置（節単位頭から数えて何番目の文節か）とフィラーの出現率との関係を調べた。また，節単位内で，伝えるべき情報がどのぐらい残っているかもフィラーの出現率に影響する可能性があると考えた。そこで，ある文節から節単位末までの文節数を，その時点で残されている伝えるべき情報量を示す指標として採用し，その値とフィラーの出現率との関係を調べた。

３．結果

まず，係先までの文節数とフィラーの出現率との対応を図 1 に示す。係先までの文節数が 21 以上のグループは，母数が 50 を切るため信頼性に問題があると考えて省略してある。フィラーの出現率は，隣接文節に係るか 2 つ先の文節に係るかで 10%違う（それぞれ，5%と 15%）が，距離が 2～11 文節の間は，1 文節増えるごとに約 2%ずつ，ほぼ線形に増加している。距離が 12 以上になると，出現率は 25%と 40%の間で変動している。

次に，節単位中の位置とフィラーの出現率との対応を図 2 に示す。文節数が 33 以上の節単位は，母数が 50 を切るため省略してある。フィラーの出現率は，第一文節直後では約 17%と，第 2 文節目以降に比べて 4～5%高いが，第 2 文節～第 28 文節境界間では 11～14%の間を変動しており，大きな違いは観察されない。すなわち，節単位中の位置効果は第一文節においてのみ観察される。

最後に，節単位中の残りの文節数とフィラーの出現率との対応を図 3 に示す。残りの文節数が 32 以上の節単位は母数が 50 を切るため省略してある。フィラーの出現率は，残りの文節数が 1 つのときは約 6%，2 つのときは約 11%と低いが，3 つ以上の場合，ほぼ 12～16%の間を変動しており，残文節数による明らかな違いは観察されない。

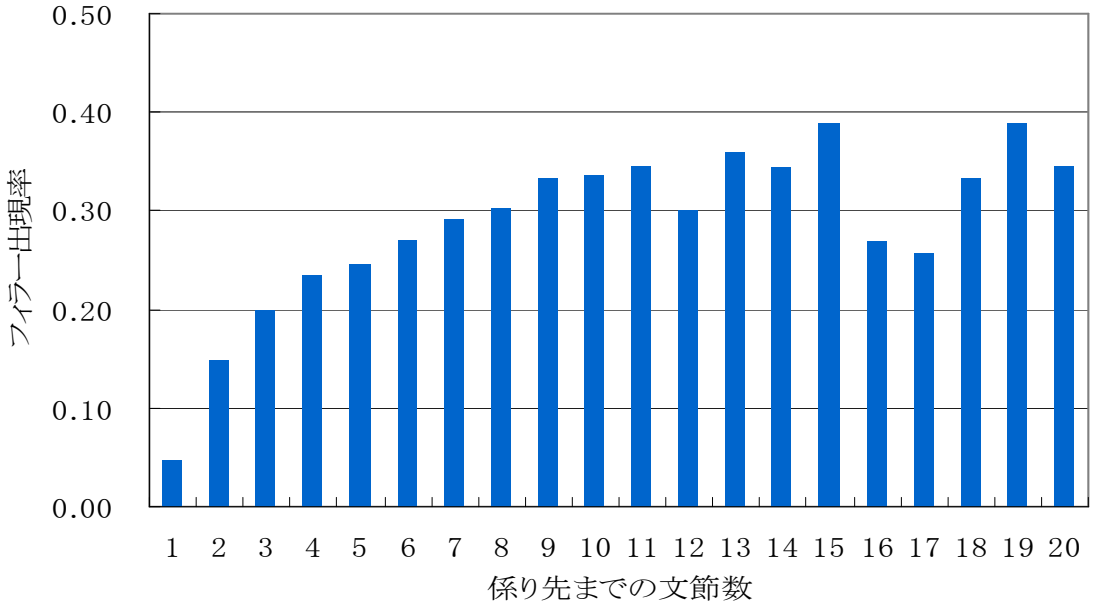


図 1　係先までの文節数とフィラーの出現率

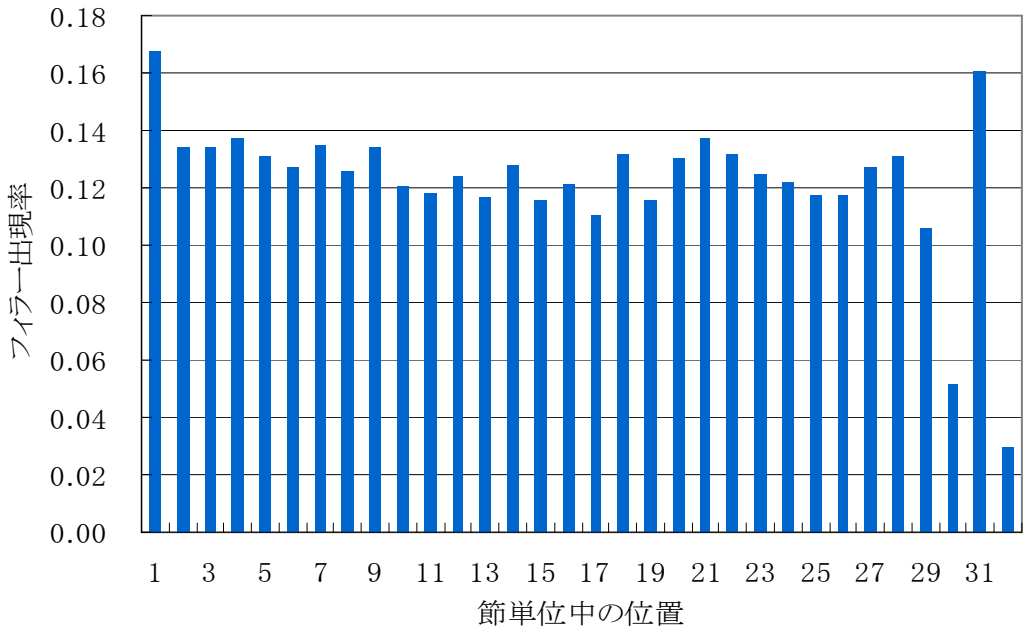


図 2　節単位中の位置とフィラーの出現率

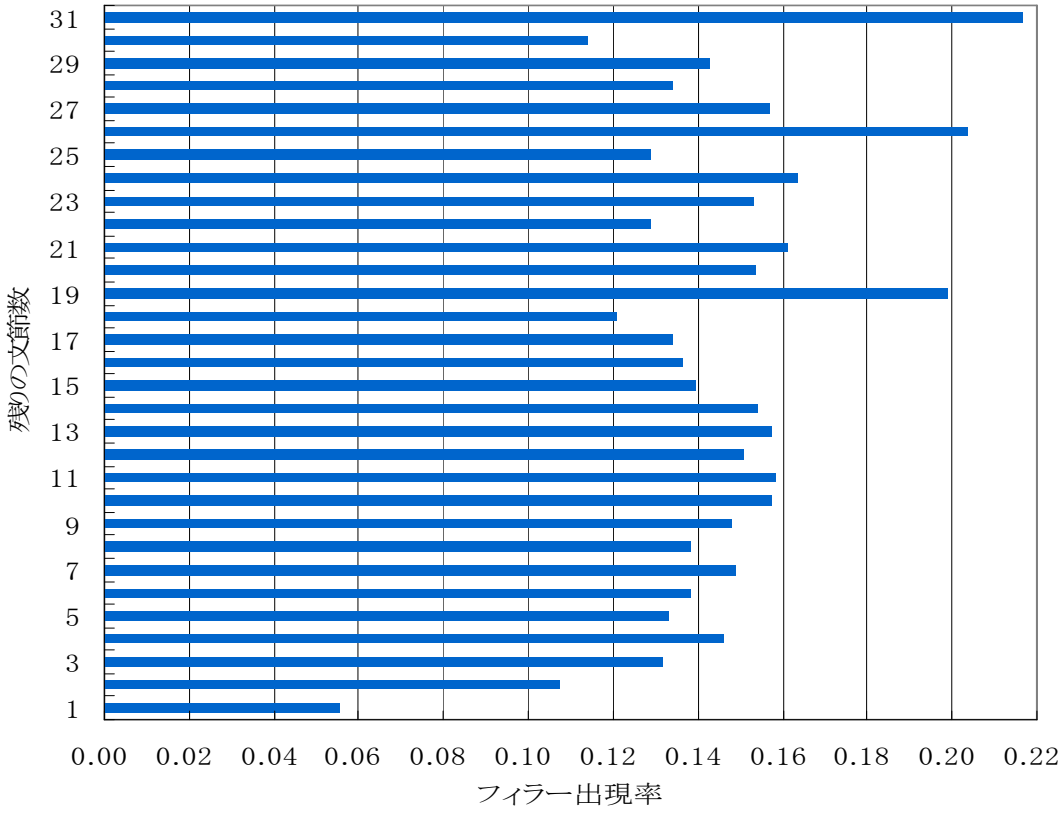


図 3　節単位中の残りの文節数とフィラーの出現率

４．考察

まず，係り先までの距離とフィラーの出現率の間には，予測通り，明らかな対応が見られた。フィラーの出現率は，係り先までの距離が長くなるほど上昇する傾向があった。この結果は，フィラーが，ローカルな言語化プロセスの負荷に関連した現象であるという仮説を支持するものである。フィラーの出現率が後続要素の統語的複雑さに対応しているという結果は，英語における冠詞や代名詞のくり返しの出現率が後続構成素の複雑さに対応するという Clark & Wasow (1998)の結果と方向を一にするものである。英語のくり返しと日本語のフィラーの働きに共通性のあることが推測される。ただし，係り先までの距離と共にフィラーの出現率が増加していくのは 11 文節までで，それ以降は明確な対応は見られなかった。この結果は，日本語では，一度にできる言語化の範囲が 11 文節ぐらいまでであることを示唆している。係り先がそれよりも遠いことが予測される場合，話者は最初から途中で区切るつもりで言語化している可能性が考えられる。

次に，節単位中の文節の位置でフィラーの出現率を比べると，第一文節直後のみで高く，それ以外の箇所で大きな違いは見られなかった。この結果には，２つの解釈が可能である。１つは，フィラーは発話内容のプラニングにも関与するが，内容のプラニングはもっぱら節単位冒頭で行われるため，フィラーの出現率も第一文節直後のみで高く，第 2 文節以降では差がないという解釈である。もう一つは，内容のプラニングは冒頭だけでなく節単位内でも行われており，その負荷は発話が進むにつれて軽減しているが，フィラーは内容のプラニングに強く関連した現象ではないため，フィラーの出現率には変化が見られないという捉え方である。確かに，第一文節直後のフィラーの一部は内容のプラニングに関連していると考えられる。しかし，内容のプラニングが第二文節以降でほとんど行われないとは考えにくく，筆者等は 2 番目の解釈が妥当なのではないかと考えている。

最後に，節単位内の残りの文節数とフィラーの出現率との関係を調べると，残りが，二文節，一文節となるにつれ，出現率は，13%から，11%，6%と，大きく低下した。一方，残りの文節が 3 文節以上の場合，一貫した傾向は見られなかった。前段落で提示した解釈同様，この結果も，フィラーが，発話内容のプラニングというよりは，ローカルな言語化に対応した現象であることを示唆しているように思われる。

フィラーの出現率は，節単位中の位置よりも後続要素の統語的複雑さに顕著に対応していた。この結果は，フィラーが，マクロなレベルの概念生成というよりは，節と同じかそれよりも小さい言語単位生成の負荷に強く関連した現象であることを示唆しているものと考える。

謝　辞

本研究は、文部科学省科学研究費補助金（基盤研究 (C)）「大規模コーパスを用いた日英語言い淀みの対照研究」（平成 21～23 年度，課題番号 21520467，研究代表者：渡辺美知子）による助成を受けて行われた。

文　献

小出慶一（1983）「言いよどみ」水谷修（編）『講座日本語の表現３ 話しことばの表現』pp. 81-87. 筑摩書房

定延利之，田窪行則（1995)「談話における心的操作モニター機構」『言語研究』，pp. 74-93.

田窪行則（1987)「 統語構造と文脈情報」『日本語学』6（5)，pp. 37-48. 明治書院

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『日本語話し言葉コーパスの構築法』，pp. 255-322.

南不二男（1974）『現代日本語の構造』大修館書店

Arnold, J. E., Altmann, R., Fagnano, M. & Tanenhaus, M. K. (2004) The old and thee, uh, new. *Psychological Science*, 578-582.

Christenfeld, N. (1994) Options and Ums. *Journal of Language and Social Psychology* 113/ 2, 192-199.

Christenfeld, N. & Creger, B. (1996) Anxiety, alcohol, aphasia and ums. *Journal of Personality and Social Psychology*, 70, 451-460.

Clark, H. H. & Wasow, T. (1998) Repeating words in spontaneous speech. *Cognitive Psychology 37*, 201-242.

Ferreira, F. (2000) Syntax in language production: An approach using tree-adjoining grammars. Wheeldon, L. R. (Ed.) *Aspects of language production*, 291-330. Psychology Press: Hove, UK.

Ford, M. (1982) Sentence planning units: Implications for the speaker's representation of meaningful relations underlying sentences. J. Bresnan (Ed.) *The mental representation of grammatical relations.* Cambridge MA: MIT Press.

Levelt, W., J., M. (1989) *Speaking*. The MIT Press: Cambridge, Massachusetts.

Wasow, T. (1997) Remarks on grammatical weight. *Language Variation and Change,* 9, 81-105.

Watanabe, M. (2009) *Features and Roles of Filled Pauses in Speech Communication -A corpus-based study of spontaneous speech* (Hituzi Linguistics in English No.14), Hituzi Syobo Publishing.

明治初期論説文における一人称代名詞の分析
—『明六雑誌』コーパスを用いて—

近藤 明日子（国立国語研究所コーパス開発センター）†

First Person Pronouns in the Articles Written in the Early Meiji Era: An Analysis of the *Meiroku Zasshi* Corpus

KONDO, Asuko (Center for Corpus Development, NINJAL)

１．はじめに

日本の近代の言語資料のコーパス化とそれを用いた近代語研究は今後一層の発展の期待される分野である。コーパス言語学的手法による近代語研究には、形態論情報の付与されたコーパスの開発が必須であるが、近代の文語論説文を対象とした形態素解析辞書「近代文語 UniDic」の開発により、その環境整備は飛躍的に進んだ。現在は、実際にその技術を用いた近代語の形態論情報付きコーパスの開発が始まっており、国立国語研究所においても、明治初期に刊行された『明六雑誌』の形態論情報付きコーパスの開発が進行中である。

本発表は、この『明六雑誌』コーパスを用いて、そこに出現する一人称代名詞の分析を行うものである。用例の抽出や分析では、形態論情報をはじめとするコーパスに付与された情報を用い、コーパスの特長を活かした研究となることを目指す。そして、『明六雑誌』というほとんどが論説文よりなる資料を用いることで、当時の書き言葉的要素の強い資料における一人称代名詞の使用実態の一端を明らかにしたい。

２．『明六雑誌』コーパスの概要

国立国語研究所で開発中の『明六雑誌』コーパスは、明治 7（1874）年から明治 8（1875）年にかけて刊行された、明六社の機関誌である『明六雑誌』の全文コーパスである。明六社は当時の洋学者によって結成された学術団体であり、そこで行われた演説や討論を広く一般に発表する媒体として『明六雑誌』は刊行された。よって、そこに掲載された記事はほとんどすべてが、ある物事について論じ解説する論説文となっている。

この『明六雑誌』に基づく本コーパスは、本文テキストに書誌・文書構造・形態論・文字等に関する情報を付与する設計となっている。付与される情報のなかで特に注目されるのは形態論情報であろう。なぜなら、これまで形態論情報の付与された近代語のコーパスはほとんど例がなく、近藤・小木曽・加藤（2010）の『高等小学読本』コーパスといったものがわずかに存在するだけだからである。本コーパスの形態論情報は、『高等小学読本』コーパス同様、近代の文語論説文（明治普通文）を対象とする形態素解析辞書「近代文語 UniDic」を用いて本文を形態素解析した後、人手修正を加えたものが付与される。それにより、語の単位として揺れのない斉一な単位である「短単位」（小椋・小磯・冨士池・他、2011）を採用し、表記の揺れや語形の変異にかかわらない見出し語を付与した、日本語研究に適した構造を持つ情報となっている。

本コーパスに付与された形態論情報をはじめとする情報に基づき、コーパスの規模を概観すると、全 43 号に掲載された記事の総数は 155 記事、著者（翻訳者含む）は異なりで 16 名、延べ語数は約 18 万 3 千語（記号類を除く）となる[1]。

† kondo@ninjal.ac.jp

[1] 本稿に示すコーパスに基づく数値は 2011 年 12 月時点のデータに基づくものであり、今後コーパスデータの変更に伴い、数値も変更となる可能性がある。

表 1 は、著者別に記事数を示したものである。これを見ると、記事数の多い上位 3 名（津田真道・西周・阪谷素）によって著された記事が計 74 記事と、全記事数の約半分を占めていることがわかる。本コーパスの分析から導き出される実態が、当時の論説文の一般的なありようではなく、特定の著者による個別的なありようである可能性があることになり、本コーパスを言語資料として扱う際には、そのことを十分に念頭に置いておく必要があるであろう。

さらに、記事の地の文の文体について見ると、全 155 記事のうち、文語文体の記事が 150 記事、口語文体の記事が 4 記事、文語口語混合文体の記事が 1 記事となっており、ほとんどが文語文体の記事で占められ、口語文体の記事はごくわずかしかない。文体の面でもデータに偏りがあることにも留意する必要がある。

表 1　著者別記事数

著者	記事数
津田真道	29
西周	25
阪谷素	20
杉亨二	13
森有礼	12
西村茂樹	11
中村正直	11
神田孝平	9
加藤弘之	8
箕作麟祥	5
柏原孝章	4
福沢諭吉	3
清水卯三郎	2
箕作秋坪	1
津田仙	1
柴田昌吉	1
合計	155

３．分析対象とする語の抽出とその度数の概観

以下、この『明六雑誌』コーパスに出現する一人称代名詞の分析を行う。近代語の人称代名詞の研究は、これまで話し言葉的性質の強い資料（小説の会話部分、落語速記、口語文典など）を中心に行われてきた。よって、論説文といった書き言葉的性質の強い資料における実態は未だ明らかになっていない部分も多い。本稿の分析によりその実態の一端を明らかにしたい。

分析のためには、まず一人称代名詞の抽出が必要となるが、抽出作業は次にあげる手順でおこなった。

① 本コーパスの形態論情報を用い、品詞が代名詞となっている見出し語を抽出する。
② 国語辞典等を参照し、①から一人称代名詞の可能性のある見出し語を選別する。
③ 一人称代名詞と関わりの深い見出し語として本コーパスでは連体詞となっている「わが」「おのが」を②に追加する。
④ ③までの作業で得られた見出し語に属する用例について、文脈を確認し、実際に一人称代名詞として用いられているものを選別し分析対象とする。さらに、関連の深い用法として、人称にかかわらず対象それ自身を指す、いわゆる反射指示代名詞として用いられている用例も分析対象とした。

この手順により、異なりで 15 語、延べで 1202 語の一人称代名詞および反射指示代名詞が抽出された。語ごとに記事の文体別の度数と表記の種類を表したものが表 2 である[2]。

これを見ると、度数の多い上位 5 語「わが」「よ」「われ」「おのれ」「ごじん」の度数を合計すると 1110 語と全体の 90%以上を占め、これら 5 語が『明六雑誌』で主たる語であったことがわかる。

[2] 表 2 にあげられた表記の中には、読みの特定が困難なものもある。例えば、「我」「吾」一字の表記は、「わが」と読むのか「われ」と読むのか（それともそれ以外で読むのか）、はっきりしない場合がある。また、「吾輩」二字の表記は「ごはい」と読むのか「わがはい」と読むのか、断言することは難しい。そこで、「我」「吾」表記は、文脈から判断して「わが」「われ」いずれかに割り振り、それ以外の漢字表記はそれぞれ一種類の読みに倒して度数を数えた。よって、例えば「吾輩」表記はすべて「ごはい」と見なし、「わがはい」として数えることはしなかった。また、「己」表記は、助詞「が」が後続する場合「おの」と読むことも多分に考えられるが、「己レガ」という「おのれ」＋「が」とほぼ確定できる表記があったため、すべて「おのれ」と見なし、「おの」として数えることはしなかった。

また、『明六雑誌』では濁音を表記する仮名に濁点が用いられていない場合があり、「わが」の「が」も「カ」と表記されることがあるが、それらはすべて「ガ」に校訂した上で表記ごとの度数を数えた。

表 2　表記の種類と記事文体別度数

語	表記の種類	度数			
		文語記事	口語記事	混在記事	合計
わが	我(474)、我ガ(67)、吾(19)、吾ガ(13)	538	27	8	573
よ	余(189)、予(6)	195	0	0	195
われ	我(127)、吾(20)、我レ(13)、予レ(1)	159	1	1	161
おのれ	己(101)、己レ(31)	129	3	0	132
ごじん	吾人(49)	49	0	0	49
ぼく	僕(17)	17	0	0	17
ごはい	吾輩(15)	15	0	0	15
わがはい	我輩(14)	14	0	0	14
せっしゃ	拙者(12)	0	12	0	12
ごせい	吾儕(10)	10	0	0	10
よはい	余輩(10)	10	0	0	10
それがし	某(4)、某シ(3)	4	0	3	7
わたくし	私(4)	0	4	0	4
よせい	余儕(2)	2	0	0	2
ぼくはい	僕輩(1)	1	0	0	1
合計		1143	47	12	1202

また、記事の文体別の度数を見ると、「せっしゃ」「わたくし」の 2 語はすべての用例が口語記事中に出現しており、これらの語の話し言葉的性質の強さがうかがえる。しかしながら、２．で述べたように『明六雑誌』の口語記事はごくわずかであり、その少量のデータに基づいて、語と文体との対応関係を分析し、当時の口語文体の論説文における一人称代名詞および反射指示代名詞の実態について論じるには限界がある。よって、以後は文語記事中に出現する語に限って分析を進めることとする。

４．語と後続助詞との対応関係

次にあげる表 3 は、文語記事中に出現する語ごとに代名詞としての用法別度数を示したものである。

表 3　代名詞用法別度数

語	一人称	反射指示	合計
わが	456	82	538
よ	195	0	195
われ	107	52	159
おのれ	0	129	129
ごじん	49	0	49
ぼく	17	0	17
ごはい	15	0	15
わがはい	14	0	14
ごせい	10	0	10
よはい	10	0	10
それがし	4	0	4
よせい	2	0	2
ぼくはい	1	0	1
合計	880	263	1143

ここから、一人称用法を持つ語は 12 語、反射指示用法を持つ語は 3 語あることがわかる。同じ用法を持つ語が複数存在する場合、内部でさらに何らかの使い分けがなされていると考えられるが、そうした語の間の違いについて探るため、後続する助詞・助動詞ごとに度

数を示した表 4 を用い、コレスポンデンス分析を行った。コレスポンデンス分析は、データ表の行や列に含まれる情報を少数の成分（次元）に圧縮し、それらの関係を散布図上に布置することで、行カテゴリー間の関係、列カテゴリー間の関係、および行カテゴリーと列カテゴリー間の関係を視覚的に捉えることができる分析手法で、コーパス言語学においても活用範囲が広いとされるものである（石川・前田・山崎（編）、2010、pp.245-249）。

表 4　後続助詞別度数

語	が_体	ナシ	の_体	が_用	に	を	は	の_用	と	も	より	てふ	なり	など	合計
わが_一	444	0	0	12	0	0	0	0	0	0	0	0	0	0	456
わが_反	75	0	0	7	0	0	0	0	0	0	0	0	0	0	82
よ	24	100	5	26	1	3	22	3	3	7	0	0	1	0	195
われ_一	0	58	8	0	24	5	5	3	2	0	1	0	1	0	107
われ_反	0	22	9	0	7	6	0	1	4	0	1	2	0	0	52
おのれ	50	14	31	4	11	16	0	2	0	0	1	0	0	0	129
ごじん	0	28	13	0	0	2	0	6	0	0	0	0	0	0	49
ぼく	1	13	0	2	0	0	0	1	0	0	0	0	0	0	17
ごはい	0	10	2	0	0	0	1	1	0	0	0	0	0	1	15
わがはい	0	10	1	0	0	0	0	3	0	0	0	0	0	0	14
ごせい	0	6	1	0	0	0	0	3	0	0	0	0	0	0	10
よはい	0	5	2	0	0	0	1	1	0	1	0	0	0	0	10
それがし	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4
よせい	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
ぼくはい	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
合計	594	273	72	51	43	32	29	24	9	8	3	2	2	1	1143

表 4 では、一人称と反射指示の両方の用法を持つ「わが」「われ」については、用法ごとにカテゴリー化し、一人称用法のものを「わが_一」「われ_一」、反射指示用法のものを「わが_反」「われ_反」として示した。また、後続する助詞・助動詞のうち「が」「の」については、後ろの体言にかかる連体用法をとるものと後ろの述語にかかる連用用法をとるものとを分けてカテゴリー化し、前者を「が_体」「の_体」、後者を「が_用」「の_用」として示した。助詞・助動詞の後続しないものについては「ナシ」としてカテゴリー化した。さらに、「わが」については、「わが」の「が」を後続する助詞と見なして度数をカウントした。

コレスポンデンス分析に用いたのは表 4 全体ではなく、網掛けを施した部分である。「わが」については、そもそも後続の助詞・助動詞という観点からの分析にはそぐわないため、分析対象から外し、また、外れ値の影響を考慮して、合計の度数が 10 未満のカテゴリーについては分析対象から外したものである。分析には、統計分析パッケージ R の MASS ライブラリーの corresp 関数を用いた。

分析結果から、もっとも寄与率の高い第 1 次元（47.62%）と第 2 次元（29.94%）の得点を 2 次元空間上に布置したものが図 1・図 2 で、図 1 は後続助詞の得点の散布図、図 2 は語の得点の散布図である。

まず、図 1 の第 1 次元を見ると、正の方向に「が_体」「を」「の_体」「に」が布置され、負の方向に「は」「の_用」「ナシ」「が_用」が布置されている。負の方向に布置される助詞群が受ける語は、多くの場合、述語に対し動作主や経験者といった意味的役割を担う[3]。一方、正の方向に布置される助詞群が受ける語は、述語に対し動作主や経験者といった意味的役割を担うことは「が_体」「の_体」の場合はもちろんなく、「を」「に」の場合も多くはない。つまり、第 1 次元は動作主や経験者といった意味役割を担うか否かに基づくものであることになる。これを図 2 と対応させてみると、他の語から大きく離れて正の方向に布置されている「おのれ」は、動作主や経験者といった意味的役割を担うことが少ないといった点で特徴付けられることになる。

[3] 「は」の場合は相当する格助詞に置き換えた場合の意味的役割について言う。

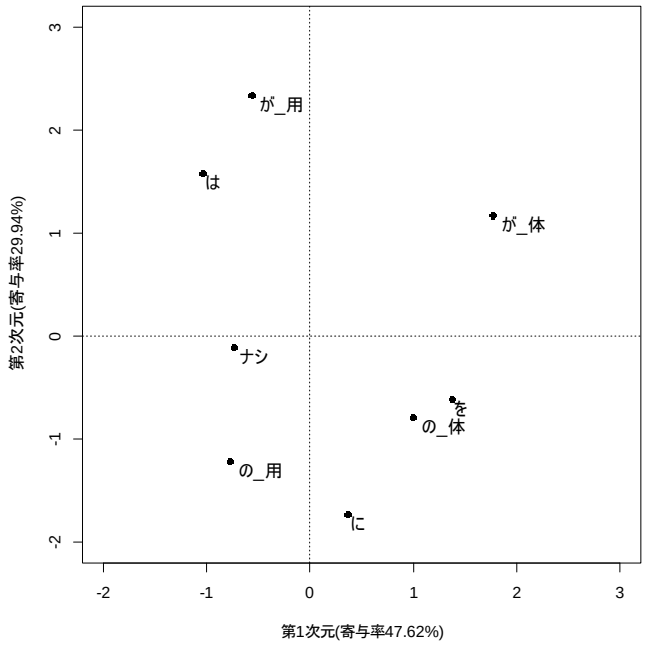


図１　後続助詞の散布図

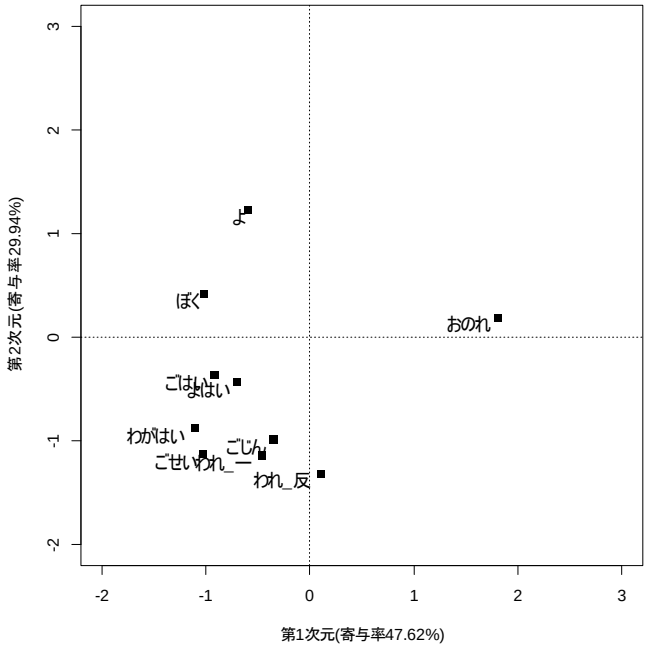


図２　語の散布図

次に、図 1 の第 2 次元を見ると、正の方向に「が_用」「は」「が_体」が布置され、負の方向に「に」「の_用」「の_体」「を」が布置されている。この軸の解釈は難しいところがあるが、正の方向に助詞「が」が、負の方向に助詞「の」が集まっている点には留意される。「が」「の」は人を表す体言をうける場合、待遇表現上の区別が認められ、「が」の用いられる場合はその人物に対する親愛・軽蔑・憎悪・卑下等の感情を伴い、「の」が用いられる場合には敬意あるいは心理的距離があると言われている。とすると、第 2 次元は待遇の程度に基づくものであることが考えられる。図 2 と対応させて見ると、正の方向に布置され

る「よ」「ぼく」は、負の方向に布置される「われ_反」「われ_一」「ごせい」「ごじん」「わがはい」「よはい」「ごはい」と比較して、相対的に待遇の程度が低いことになる。

以上のように、後続助詞と語との間には明らかな対応関係があり、それにより語は大きく次の 3 つのグループに分けることができると考えられる。

A　おのれ
B　よ・ぼく
C　われ_一・われ_反・ごじん・ごはい・わがはい・ごせい・よはい

このグループ分けと語の代名詞としての用法との関係を考えると、まず A グループの「おのれ」は反射指示の用法を専らとする点で他の語と区別される。B グループの「よ」「ぼく」は一人称で、かつ書き手自身のみを指す単数用法を専らとする点で他の語と区別される。C グループの内、「われ」を除いた「ごじん」「ごはい」「わがはい」「ごせい」「よはい」は一人称で、かつ書き手だけでなく書き手を含めた複数の人を指す複数用法を取り得る点で他の語と区別される。本コーパスでの用例を見ると、「われ」を除く C グループの中で最も度数の多い「ごじん」は専ら複数用法をとり、「ごはい」「わがはい」「ごせい」「よはい」も複数用法が認められる。

このように、後続助詞との対応関係に基づく語のグループは、代名詞としての用法に基づく語の分類とほぼ一致することがわかる。

なお「われ」は、代名詞としての用法から見ると、一人称・反射指示両方の用法を持つ点で他の語とは区別されるが、後続助詞という観点からは「ごじん」等と同じ C グループに属する結果となった。「われ」については別の観点によるさらなる分析が必要であると言える。

５．連体用法における語と被修飾体言との対応関係

次に４．で分析の対象外とした「わが」について見てゆく。表 4 に示したように、「わが」は一人称・反射指示のどちらの用法でも連体用法をとることが多い。そこで、連体用法をとる「わが」および「の_体」「が_体」を伴う他の語について、被修飾体言との対応関係について検討し、語の間の違いについて見ていく。

表 5 は、各語が連体用法をとる場合の被修飾体言を示したものである。（　）内は各体言の度数を示す。体言の種類の多い場合は、代表的な体言のみを示し以下は省略した（「…」で表記）。また「如し」にかかるものもここに含めて示してある。

表 5　連体用法における被修飾体言

語	が_体	の_体
わが_一	国(195)、帝国(40)、人民(17)、大日本帝国(13)、国内・地球・政府(7)、邦人(6)、国民・民・国産・社(5)、日本帝国・日本・心(4)、アジア・法律・今上天皇陛下・性・東州・東方(3)…	
わが_反	国(5)、身・父(4)、為・同生同人・同人・日本・自由・父母・用・物品・子・本体(2)…	
よ	ロジック・考・言(3)、胸臆・頭脳・所見(2)…	喜び・憶説・論・意・如し(1)
われ_一		有・文章・障子ガラス・義務・民…(1)
われ_反		如し・三法・下・精神・国・父…(1)
おのれ	力(5)、為・身体(3)、用・自由・三宝・意・身・一身・利・鋭利・労(2)…	意(3)、欲・如し(2)、迷信・子・力・胸中・責・権利・国…(1)
ごじん		為・性・心裏(2)、進歩・生活・感覚・天性…(1)
ぼく	論(1)	
ごはい		雲仍(2)
わがはい		目(1)
ごせい		如し(1)
よはい		首唱・鄙見(1)

ここから、語と被修飾体言との関係を見てゆく。

まず「わが」については、特に一人称用法の「わが」は、被修飾体言が「わが」の「わ」にとっての「所属先」という関係になる場合が多いということが言える。「わが」以外の語では、被修飾体言は各語にとっての「所有物・所属物」という関係をとることが多いのとは対照的である。典型的なのは最も度数の多い「わが国」で、「「わ」の所属する国」の意となる。「わが帝国」「わが地球」「わが社」「わがアジア」「わが東州」等も同様である。さらに、体言が「所属先の所有物・所属物」、特に「所属する国の所有物・所属物」という関係になることもある。例えば「わが人民」とは「「わ」の所属する国に所属する人民」の意で用いられている（「「わ」の統治する人民」「「わ」の所有する人民」の意ではない）。「わが政府」「わが邦人」「わが国民」「わが民」「わが法律」等も同様の関係にある。このように、一人称用法の「わが」は、被修飾体言の関係が「所属先」「所属先の所有物・所属物」となる点で特徴付けられる。なお、反射指示用法の「わが」については、その被修飾体言が「所属先」「所属先の所有物・所属物」の関係となる割合は一人称用法のものほど高くなく、「身」「父」「自由」「物品」等の「所有物・所属物」の関係となる場合も比較的多くなっている。

「わが」以外の語は、先に述べたように、被修飾体言が「所有物・所属物」の関係になることが多い。その中で、「よ」は被修飾体言が「ロジック」「考」「言」といった「所有する考え・意見」を意味する語で多く占められる点で特徴付けられる。「ぼく」「よはい」も被修飾体言に「論」や「首唱」「鄙見」をとり、「よ」と同様の傾向があるものと見られる。

以上のように、連体用法において語と被修飾体言との間にはいくつかの対応関係が見いだされることがわかった。

６．主な語の特徴

以上の分析結果に基づき、主要な語についてそれぞれの特徴をまとめる。取り上げる語は文語記事での度数の多い上位 5 語「わが」「よ」「われ」「おのれ」「ごじん」である。

まず「わが」であるが、連体用法をとる主たる語であり、(1)(2)のように被修飾体言が「所属先」「所属先の所有物・所属物」の関係となる点で特徴的である。

(1) 夫レ我ガ國ノ文字先王始メ之ヲ漢土ニ取テ之ヲ用ウ（1 号「洋字ヲ以テ国語ヲ書スルノ論」西周）[4]

(2) 目今諸省ニ於テ許多ノ洋人ヲ雇テ其學術ヲ傳取スル如ク彼尤善尤新ノ法教師ヲ雇テ公然我人民ヲ教導セシメバ奈何（3 号「開化ヲ進ル方法ヲ論ズ」津田真道）

次に、「よ」であるが、一人称単数の用法をとる主たる語である。述語に対し動作主や経験者といった意味的役割を担い、(3)のような著者の個人的な体験を語る文脈でも用いられるが、論説文という文章の性質上、(4)(5)のように著者の意見や主張を述べる文脈で用いられることが多い。連体用法をとる場合も同様で、(6)のように著者の意見や主張を意味する語が被修飾体言となる。著者個人を指し示す語ゆえに、卑下の感情を伴う助詞「が」のほうが「の」よりも後続しやすい。

(3) 余會テ歐洲ニ遊テ煉火石造ノ家屋ヲ見ル（4 号「煉火石造ノ説」西周）

(4) 故ニ余敢テ謂フ我邦人倫ノ大本未ダ立ズト（8 号「妻妾論ノ一」森有礼）

(5) 余ハ思フニ政府ハ猶精神ノ如ク人民ハ猶躰骸ノ如クナリ（2 号「学者職分論ノ評」津田真道）

(6) 余ガ考ニハ狗ヲ連ルヨリモ兎ヲ輸入シテ錢ヲ取ラルヽ方遙ニ恐ル可シト思フ位ノコトナリ（26 号「内地旅行西先生ノ説ヲ駁ス」福沢諭吉）

次に「おのれ」であるが、反射指示用法をとる主たる語である。(7)(8)のように連体用法をとることが多く、述語に対して動作主・経験者といった意味的役割を担うことは少ない。

(7) 是皆個々人々日夜孜々汲々己ガ勞ヲ厭ハズ己ガ力ヲ盡シテ之ヲ求ムベキ者ニシテ

[4] 本文の引用に際しては、末尾の（　）内に号数・記事題名・著者名を示す。

（38 号「人世三寶説（一）」西周）

(8) 今日ニ至リテハ諸邦ノ君主タトヒ聰明衆ニ超タリトモ己ノ意ヲ以テ命令ヲ下スコトナシ（12 号「西学一斑（前号ノ続）」中村正直）

次に「ごじん」であるが、一人称複数の用法をとる主たる語である。著者の個人的な意見について述べる文脈で用いられやすい「よ」とは異なり、(9)(10)のように、より一般性のある説や論を述べる文脈に用いられることが多い。また、著者自身のみならず他の人も含めて指し示す語ゆえに卑下の意味を伴う助詞「が」が後続することはない。

(9) 想像ハ瞑目思想ノ間吾人覩見スル所ノ形象事歴ニシテ頗ル蜃氣樓ト相類似ス（13 号「想像論」津田真道）

(10) 若夫レ吾人ノ性中情欲ヲ缺ク時ハ人類何ニ由テ生々蕃植スルコトヲ得ンヤ（34 号「情欲論」津田真道）

最後に「われ」であるが、一人称・反射指示の両用法をとる語であり、一人称用法の「われ」は「よ」「ごじん」との違いを、反射指示用法の「われ」は「おのれ」との違いを明らかにしたいところである。

一人称用法の「われ」は、後続助詞との対応関係から「ごじん」と同じグループに属し、さらに(11)のように複数用法と思われる用例が見いだされ点でも「ごじん」と共通する。

(11) 米利ノ戰艦一旦江戸海ニ侵入シ請求スルニ通信ノ約ヲ以ス是ニ於テ我之ヲ託シ始テ彼ニ日本來往ノ便ヲ得シム（7 号「独立国権義」森有礼）

反射指示用法の「われ」は、(12)のように述語に対し動作主や経験者といった意味的役割を担うことが少なくなく、その点が「おのれ」とは異なる。

(12) 自由ヲ伸シ羈絆ヲ脱シ租税ハ吾之ヲ増減スベシ官吏ハ吾之ヲ進退スベシ是人民ノ利ナリ（39 号「政府与人民異利害論（六月一日演説）」西村茂樹）

本稿の分析からは、このような「われ」と他の語との類似点・相違点が指摘できるが、ではさらに進んで「ごじん」との違いはどのような点にあるのか、「おのれ」との違いは何によってもたらされるのかといったことについては、明らかにはできなかった。今後の課題としたい。

7．おわりに

以上、『明六雑誌』の形態論情報付きコーパスを用いて分析を行い、当時の文語論説文における一人称代名詞（および反射指示代名詞）について実態の解明を試みた。語と後続助詞、語と被修飾体言との間には明らかな対応関係が見いだされ、それにより一部ではあるが各語の特徴が明らかになった。今後は、別の観点からの分析を加え、一人称代名詞間の違いについてより詳細に考察したい。また、他のコーパスを用いて分析を行い、近代の一人称代名詞の通時的変遷についても考察する予定である。

文献

石川慎一郎、前田忠彦、山崎誠（編）（2010）『言語研究のための統計入門』、くろしお出版

小椋秀樹、小磯花絵、冨士池優美、宮内佐夜香、小西光、原裕（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』形態論情報規程集　第 4 版（上）（下）』

近藤明日子、小木曽智信、加藤文明子（2010）「『高等小学読本』の形態論情報付きコーパス」、情報処理学会シンポジウムシリーズ Vol.2010, No.15 人文科学とコンピュータシンポジウム論文集 人文工学の可能性～異分野融合による「実質化」の方法～、pp.189-194

関連 URL

近代文語 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

日英の理工系口頭発表コーパスの構築と検索サイト JECPRESE

林　洋子（大阪大学国際教育交流センター）
国吉 ニルソン（早稲田大学理工学術院）
野口 ジュディ（武庫川女子大学）
東條　加寿子 （大阪女学院大学）

Building a Japanese-English Corpus of Presentations in Science and Engineering and JECPRESE

Hiroko Hayashi (Center for International Education and Exchange, Osaka University)
Nílson Kunioshi (School of Creative Science and Engineering, Waseda University)
Judy Noguchi (School of Pharmacy and Pharmaceutical Sciences, Mukogawa Women's University)
Kazuko Tojo (Department of International & English Interdisciplinary Studies, Osaka Jogakuin College)

１．はじめに

現代日本語のコーパスは、書き言葉のみならず話し言葉についても整備されつつある。しかし、留学生を含む大学院の学生が最も必要とする修士論文口頭発表のデータは未だ明らかにされていないのが現状である。これは、知的財産権の問題があること、および学位取得のための口頭発表が専門分野に属するものとされ言語教育の対象とは思われてこなかったことなどによると考えられる。

そこで、我々は許可を得て修士論文口頭発表のデータを収集し、英語の発表と比較検討する研究に取り組んでいる。本稿ではデータ収集の歩み、およびデータを載せた検索サイト JECPRESE の開発、およびデータの解析結果の一部について概観する。

２．データ収集の歩み

我々は日本語、英語、化学の研究者の混成チームであり、本稿に述べるように 400 近くの発表データを含む理工系のプレゼンテーション・コーパスを構築することができた。どのようにそれが可能であったかと、理系の研究者との共同研究を願っている日本語の研究者から問われる機会が多い。そこで、本稿では、データ収集の歩みについて詳細に述べることにする。

1991 年当時、大阪大学には留学生のための日本語教室はなかった。そこで、工学研究科に日本語クラスが設置されることになり、偶然にも林が担当することになった。その時の留学生相談室の担当教員は「ぜひ理系という特色を生かした日本語教育を」と熱意に燃えていたが、日本語が話せない留学生と漢字能力が高い中韓の留学生が混在するクラスで、実験等で休みがちの留学生に、研究できるまでの日本語力がつくよう指導することは難しかった。試行錯誤を続けたが、特にアジア圏の留学生から「第二次世界大戦で焼け野原になったのに、欧米に飲み込まれずにトップクラスの科学技術レベルを持つにいたったのはなぜか、を知りたいため日本に留学した、それを可能にした日本の文化についても学びたい」という声が聞かれた。そのため、林が担当した日本語クラスでは、「豊かさとは何か、コミュニケーション、フェミニズム」など社会科学的トピックについてのディスカッションを通じて語彙・表現を増やすという手法をとった。（使用したテキスト「日本語で考える」は工学研究科によって発行されている。）ディスカッションには、理工系で必要な語彙・表現を使用するとしたが、「理系の日本語語彙」についての参考文献はなかったため、林は工学研究科の教員を対象として語彙・表現についてのアンケート調査を行った。得ら

れた調査結果はアンケートで恣意的である可能性もあるため、網羅性のあるデータ収集方法を模索し、各専攻公聴会の聴講を重ねた。当時の工学研究科の教官の反応は「日本語の先生には専門の語彙・表現は難しいでしょう」というものであった。個人的に論文や口頭発表の入手も試みたが、大規模なコーパス構築は難しかった。その時、偶然に、ある大学の農学部卒業論文発表の録音を入手・文字化し、序論部を解析した論文を出すことができた。それを旧知の教官に見せ「頂いたデータはこのような形で公表するため専攻に迷惑をかけることはない」とデータ収集をお願いした。その教官は専攻の許可をとってくださったが「データは序論部に限る」という限定付きであった。林は「論文すべてのデータが必要」と主張し、許可を得るため専攻の教授会に出席し必要性を訴えることになった。幸いにも許可が得られ、データの解析結果を論文で発表した。その論文を他専攻の教官にお見せしその専攻でのデータ収集の許可をお願いした。

このようなやり取りの中で、林は個人ではなくチームでコーパス構築し・解析する研究の必要性を感じ、日本語の関係者に共同研究を申し入れたが賛同はえられなかった。当時、工学教育ではますます英語教育が重視されるようになっており、「理工系では英語で指導するため専門日本語教育は不必要」と考えている日本語教員も多かった。

林は工学英語（野口）を聴講させていただき、ESP (English for Specific Purposes)の考え方に触れ、コーパスの重要性を理解していた野口と、英語による化学教育を始めようとしていた国吉とチームを組むことになった。その後、国吉が早稲田大学に転出することになり、さらに東條（英語教育）がチームに加わった。

英語のデータ収集はチームの初めからの目標で、野口を中心にかなり働きかけをした。しかし、欧米では許可を取ることは極めて難しいことがわかった。その後アメリカの大学生の発表のデータが得られたが、日本の修士論文口頭発表とはかなり異なっていた。国際学会のデータは国吉の働きかけで得られた。日英語の比較については後に発表する予定である。

データを収集するにあたっては各専攻・発表者と厳しい契約を取り交わした。その内容はおおむね以下のようである。

1. 序論のみならず内容までに関する公聴会の聴講は許可する。
2. 公聴会での記録はテープレコーダーのみ許可する。
3. 学生発表原稿の内容は発表前日のものとする。学生によっては、手書きの学生、ワープロで作成する学生、原稿を必要しない学生、等々であるが、なるべく学生にはワープロ作成するように指導する。しかし、学生全員の原稿の収集は不可能である。
4. 公聴会において専攻内で配布する「発表内容概要（集）」は提供する。
5. 上記の研究内容に関する記録（2)、原稿（3)、概要集（4）等を第 3 者に閲覧させること及びそれらのコピーを第 3 者に配布することを厳禁とする。
6. この調査で得られたデータは加工して語彙・表現の集計として公表する。
7. 各研究内容について公表することなく，著作権・特許権等に触れることはしない
8. 調査結果を公表するに際しては事前に専攻の許可を得る。

得られたデータは録音・原稿・発表のパワーポイントなど様々である。契約でチームの 4 人以外はデータにアクセスできないため、未だ文字化できていない録音も多いが、この契約により、データは学生個人というより専攻の承認を得たものとなり、信頼性が高まったと考えている。日本においても欧米と同様、研究発表の現場データの入手は、今後さらに困難になると予想される。その意味でも本コーパスは極めて貴重なデータになろう。表 1 に本コーパスの収録データ数を掲げる。

表 1 収録されたコーパス

	専攻等	英語	日本語
2003 年	知能機能創成工学	0	30
2005 年	応用生物工学	0	39
2006 年	物質化学	0	34
	分子化学	0	31
	知能機能創成工学	0	13
	電気工学	0	13
	化学系 COE	0	5
2007 年	船舶海洋工学	7	0
	機械工学	0	69
	化学系国際学会	37	0
	環境・エネルギー工学	1	0
	バイオテクノロジー英語特別コース	4	0
2008 年	環境・エネルギー工学	0	77
	アメリカ学部学生	16	0
		65	311

３．オンライン検索サイト JECPRESE のインタフェース設計

国吉は、理工系研究者が直感的に利用できるように，JECPRESE のインタフェースを設計した．検索対象の単語・表現，言語，move(表現意図)，分野，講演者の対象言語における経験，表示結果の詳細など入力・変更する方法を容易にした．すなわち，move および研究分野の検索をクリック一つでできるようにした．また，検索方法や move 記号の詳細を「Help」ページにて，書き起こしファイルの名称の意味を「Filenames」にて，説明した．さらに，単語・表現の検索結果を，前後の文脈を基準にして簡単にソートできるようにし，検索した単語・表現がどのような文脈でよく現れるかが素早くわかるようにした．現在はさらに使いやすいインターフェースにするための変更を目指しており、今年度中には完成する予定である。

４．本コーパスの解析結果

４.１　理工系専門語彙の特定

林(2004)では「工学系の基本語彙はさらに細分化していると思われる。従って、代表的な学術雑誌に掲載された論文を抽出し分析するこれまでのような語彙調査においても、より詳細に研究分野（ロボット、マテリアルなど）を検討した上で、その分野における頻度数の高い語彙を抽出する必要があると考えられる。また、各研究分野の高頻度語彙を階層的に積み上げていくことによって機械・材料・電気・情報などの語彙のグループ、さらに工学系・医学系などの語彙のグループと、語彙の階層性が明らかになり、同時にいずれの分野にも共通の専門記述語が判明すると思われる。」とした。なお、解析には奈良先端科学技術大学院大学の自然言語処理学講座によって開発された形態素解析システム「茶筌」を用いた。林ら(2010)ではさらにコーパスを広げて解析し「頻度数上位 10 語だけで各品詞の延べ数のほぼ半数以上がカバーできる．また，それらの語彙は化学系と機械系でほぼ共通しており，さらに，全語彙中の各品詞は化学系，機械系，知能機能系，いずれの分野におい

てもほぼ同じ割合を示している．基本的な文を基本的な語彙を用いて定型化し，そこに研究分野によって多彩な名詞・漢語動詞を入れ込むことによって研究内容を表していることがわかった．異なる分野においてこのような結果がみられることは，理工系口頭発表における文の形態，発表の構成が類似・標準化されていることを示唆しており，これらの語彙は工学教育基準の基本語彙の可能性があることを示している.」とした。表 2、3 には分野・品詞別の頻度数・割合を示す。また、表４～17　には品詞ごとの頻度数上位 10 語を掲げる。

表 2 分野・品詞別の頻度数

	化学	機械
提供された発表原稿数	40	69
協力 labo 数	13	22
形容詞	689	1287
格助詞-連語	1574	2179
接続詞	1069	1449
接頭詞	905	1232
動詞	8892	13517
名詞-サ変接続	9784	13683
副詞	855	979
名詞	13624	21503
名詞-形容動詞語幹	1045	1538
名詞-接尾	5735	6458
名詞-非自立	1927	1019
名詞-副詞可能	1746	2023
連帯詞	1155	1946
名詞-代名詞	867	1051

表 3 分野・品詞別の割合（%）

	化学	機械	知能機能創成工学
形容詞	1.4	1.8	2.1
格助詞–連語	3.2	3.1	3.3
接続詞	2.1	2.1	1.8
接頭詞	1.8	1.8	1.5
動詞	17.8	19.3	17.4
名詞–サ変接続	19.6	19.6	19.5
副詞	1.7	1.4	1.2
名詞	27.3	30.8	29.6
名詞–形容動詞語幹	2.1	2.2	2.5
名詞–接尾	11.5	9.2	9.4
名詞–非自立	3.9	1.5	3.9
名詞–副詞可能	3.5	2.9	4
連帯詞	2.3	2.8	2.6
名詞–代名詞	1.7	1.5	1.2
計	99.9	100	100

表 4 名詞の頻度数上位10語

化学	13624	機械	21503	知能機能創成工学	8639
錯体	358	粒子	341	表面	168
分子	340	条件	329	ロボット	167
触媒	294	図	302	温度	145
活性	228	熱	236	材料	138
光	205	システム	222	モデル	126
構造	198	方向	218	ナノ	122
炭素	181	軸	215	熱	118
基質	129	速度	197	形状	97
蛍光	128	工具	196	原子	91
金属	124	モデル	182	特性	85
10 語の割合	16%	10 語の割合	11%	10 語の割合	15%

表 5 形容詞の頻度数上位10語

化学	689	機械	1,287
高い	112	大きい	255
よい	87	小さい	134
大きい	64	高い	84
低い	47	長い	65
ない	32	ない	64
強い	31	多い	64
新しい	23	硬い	46
多い	23	厚い	40
小さい	21	よい	39
長い	19	少ない	37
10 語の割合	67%	10 語の割合	64%

表 6 名詞-形容動詞語幹の頻度数上位10語

化学	1,045	機械	1,538
可能	114	可能	127
同様	96	安定	86
明らか	84	必要	82
様々	65	同様	70
非常	58	困難	52
必要	54	十分	48
安定	49	主	42
重要	34	自由	41
新た	31	不安定	35
主	29	多様	32
10 語が品詞全体に占める割合	59%	10 語が品詞全体に占める割合	40%

表 7 接続詞の頻度数上位10語

化学	1069	機械	1449
また	276	また	338
および	178	および	150
そこで	108	そして	147
次に	94	そこで	143
一方	61	次に	142
しかし	52	しかし	91
つまり	36	それでは	64
そして	35	つまり	48
たとえば	23	一方	38
あるいは	18	なお	31
10 語が品詞全体に占める割合	82%	10 語が品詞全体に占める割合	82%

表 8 副詞の頻度数上位10語

化学	855	機械	979
さらに	137	まず	295
まず	116	さらに	106
ほとんど	57	次に	85
もっとも/最も	39	実際	62
ほぼ	34	ほぼ	52
より	33	もっとも/最も	38
全く	31	特に	33
特に	26	ほとんど	25
既に	20	同時に	23
よく	19	常に	22
10 語が品詞全体に占める割合	60%	10 語が品詞全体に占める割合	76%

表 9 接頭詞の頻度数上位10語

化学	905	機械	1232
本	142	本	333
脱	100	各	113
超	49	高	84
当	43	非	71
環	42	超	48
再	42	低	40
約	42	被	37
重	38	約	36
単	35	再	34
不	33	第	31
10 語が品詞全体に占める割合	63%	10 語が品詞全体に占める割合	67%

表 10 名詞-接尾の頻度数上位10語

化学	5735	機械	6458
化	544	的	376
的	439	率	279
性	394	化	259
物	343	部	251
位	300	法	221
体	229	物	196
基	215	性	189
剤	204	値	182
率	166	流	182
量	163	面	174
10 語が品詞全体に占める割合	52%	10 語が品詞全体に占める割合	36%

表 11 格助詞 - 連語の頻度数上位10語

化学	1574	機械	2179
により/よって/よる	491	により/よって/よる	504
として/しまして	288	について	486
について	235	において/おける	446
において/おける	228	として/しまして	335
に対し/対して	142	に対し/対して	197
という/といった	91	という/といった	89
に関して/関する	36	に関して/関する	89
とともに	33	とともに	17
に従い/従って	12	を通して	4
につれ	8	に従い/従って	3
10 語が品詞全体に占める割合	99%	10 語が品詞全体に占める割合	99%

表 12 名詞-副詞可能の頻度数上位10語

化学	1746	機械	2023
結果	351	結果	378
ため	234	時間	188
場合	177	場合	149
ところ	99	以上	136
以上	89	それぞれ	88
それぞれ	77	以下	82
時間	60	ため	80
とき/時	56	とき/時	67
今回	43	今回	53
中	37	従来	46
10 語が品詞全体に占める割合	70%	10 語が品詞全体に占める割合	63%

表 13 名詞-非自立の頻度数上位 7 語

化学	1927	機械	1019
こと	1164	よう/様	752
の	102	もの	105
もの	101	こと	72
よう/様	424	ほう/方	44
ン	103	ン	17
ほう/方	21	点	17
点	12	の	12
8語が品詞全体に占める割合	100%	8語が品詞全体に占める割合	100%

表 14 連体詞の頻度数上位10語

化学	1155	機械	1946
この	790	この	1261
その	263	その	370
同じ	35	同じ	98
大きな	30	大きな	69
どの	15	小さな	62
さらなる	7	どの	50
なんらかの	3	ある	16
ある	2	なんらかの	5
いわゆる	2	いかなる	4
どういう	2	こうした	4
10 語が品詞全体に占める割合	99%	10 語が品詞全体に占める割合	99%

表 15 名詞-代名詞の頻度数上位10語

化学	881	機械	1051
これ	215	これ	266
こちら	183	こちら	260
これら	154	これら	161
それ	105	ここ	147
ここ	55	それ	88
それら	36	それら	32
いずれ	34	どちら	19
私	31	我々	14
そこ	12	そこ	12
どちら	9	その他	12
10 語が品詞全体に占める割合	95%	10 語が品詞全体に占める割合	96%

表 16 動詞の頻度数上位12 語

化学	8892	機械	13517
する	3526	する	4360
用いる	447	なる	671
行う	388	示す	513
なる	383	用いる	480
示す	376	行う	447
考える	313	できる	451
得る	305	わかる	383
できる	254	考える	288
よる	192	ある	281
わかる	175	みる	195
ある	143	得る	182
有する/有す	136	求まる	167
10 語が品詞全体に占める割合	75%	10 語が品詞全体に占める割合	62%

表 17 名詞-サ変接続の頻度数上位10語

化学	9784	機械	13683	知能機能創成工学	5701
反応	1015	研究	451	学習	164
生成	367	計算	353	研究	156
酸化	299	加工	331	結晶	151
進行	223	実験	321	計算	136
合成	220	制御	220	組織	132
検討	213	説明	211	溶融	118
研究	209	変化	199	変化	103
結合	208	発生	184	実験	95
選択	202	影響	173	解析	83
配	178	解析	168	凝固	80
10 語が品詞全体に占める割合	32%	10 語が品詞全体に占める割合	19%	10 語が品詞全体に占める割合	21%

これらは「工学教育」に採択された論文に載せたデータであるが、論文の採択によりこの語彙リストは工学教育のエキスパートの承認を得たと考えている。

４.２　ムーヴ（表現意図）の特定

専門的な職業・学問に携わる人々の集団（ディスコース・コミュニティ）はその共通目的を達成するためにコミュニケーションを行うが、それが繰り返されることによりパターン化してジャンルが形成される。ジャンルは文全体の構造・文法要素・単語・フォーマットなどの統合体であり、書き手や読み手など当該ディスコース・コミュニティ内のメンバーに共通して利用されている。Swales は Genre Analysis (1990)においてジャンルの重要性について指摘し、理工系論文のジャンルにおけるムーヴ（move）解析を行った。ムーヴ解析は「その表現形式は，ディスコース・コミュニティのどのような表現意図を示すものか」を探る。しかし、Swales のムーヴ解析は Introduction、Method の一部に限定されていた。我々は本コーパスを解析し、英語による発表より日本語による発表の方がより形式化されておりムーヴ解析が容易であることを見出した。ムーヴのリストについては林ら(2009)において明らかにしたが、現在 JECPRESE の改良にあたり、リストも見直している。

５．まとめ

日英語による口頭発表の信頼できるデータを集めたコーパスを構築し、コーパスの一部を解析し頻度数の高い語彙・表現を抽出し、ムーヴ解析も行った。また、研究・口頭発表に用いられる表現を容易に検索できるサイト JECPRESE を開発した。本コーパスは今後も拡大していく可能性が高い。

６．今後の課題

我々のコーパスは理工系の発表に特化した信頼できるデータと考えられるため、今後さらに詳細な解析を行いたいと考えている。すでに林(2004)、林ら(2010)で指摘したように、形容詞などの各語彙・条件表現などに特徴的な傾向がみられ、また、Theme/Rheme，Given/New 概念を含む Information structure（情報構造）の日英語における違いが明らかになった。今後は対象を広げ日英語の比較を考慮しながら精密な解析を行いたいと考えている。

謝　辞

貴重なデータをご提供くださった皆様に感謝申し上げます。また、本研究は平成 21 年度より科学研究費基盤研究（C）補助金を受けています。

文　献

林洋子(1999)「大阪大学工学部教官の認識に関する調査」『専門日本語教育教材作成に向けてー教官へのアンケート調査からー』，大阪大学工学部国際交流委員会

林洋子(2001)『日本語で考える：理工系専門日本語　基礎コースⅠ』

林洋子(2002)「考える」プロセスを重視して-多文化クラスの試み-，専門日本語教育研究，第 4 号, pp. 37

米田由喜代・林洋子(2003)「口頭発表の序論部の談話構造と語彙・表現-農学部卒業論文発表の分析から-」，専門日本語教育研究，第 5 号, pp. 37-43

林洋子(2004)「工学系修士論文口頭発表に用いられた語彙・表現」専門日本語教育研究,第 6 号、pp. 25～32

野口ジュディ・国吉ニルソン (2005)‘ESP education based on JSP research’JACET Kansai Chapter 2005 Spring Conference, June 4, 2005, Wakayama University

野口ジュディー，林　洋子，国吉ニルソン，東條加寿子(2007)理工系日本語・英語口頭発表における move・表現が検索可能なオンラインコーパスの開発，言語処理学会第 14 回年次大会発表論文集,pp.516−519

林　洋子，国吉ニルソン，野口ジュディ，東條加寿子(2008)若い研究者の言語獲得，電子情報通信学会、技術研究報告, IECE Technical Report, TL2008-3,2008-05, pp.11−16, 2008

林　洋子, 国吉ニルソン, 野口ジュディ(2009)工学系修士論文口頭発表のムーヴ解析, 工学教育, 57−6,pp.137−143, 2009
林　洋子, 国吉ニルソン, 野口ジュディ(2010)化学系と機械系の基本語彙, 工学教育, 58−6,pp.130−136, 2010
国吉ニルソン, 野口ジュディ、東條加寿子、林　洋子(2011) “Building a bilingual corpus of presentations in science and engineering: Purpose, issues and procedures”　The 16th World Congress of Applied Linguistics, August 27, Beijin, China.
東條加寿子 (2011) “Analysis of rhetorical strategies to identify moves in English research presentations in science and engineering fields”　The JACET 50th Commemorative International Convention, September 2, Fukuoka, Japan.

関連 URL

JECPRESE, The Japanese—English Corpus of Presentations in Science and Engineering, http://www.jecprese.sci.waseda.ac.jp

日本語対話コーパスにおける倒置構文について：聞き手の反応に注目して

郭潔（千葉大学大学院融合科学研究科）
伝康晴（千葉大学文学部）

Inversion Expressions in Japanese Conversation: From the Viewpoint of Hearer's Reactions

Jie Guo (Graduate School of Advanced Integration Science, Chiba University)
Yasuharu Den (Faculty of Letters, Chiba University)

1. はじめに

大坊(2009)によると、コミュニケーションは、「伝える」ことと「伝えられる」ことから成り立つ。他者に伝える方法は言語的、非言語的チャネルがあり、言語的コミュニケーションはシンボルとして「言葉」を用いる。言葉を表出したものが文や発話であり、さまざまな表現形式を取る。一般に、自然言語の表現には決まった語順があり、日本語では、主語が前、述語が後ろ、修飾語が前、被修飾語が後ろという順序で配列される。しかし、実際の言語運用において、特に話し言葉では、必ずしもすべての文がこの語順を守るのではなく、さまざまな理由によって語順が変っていく。そのうちの一つに、倒置構文というものがある。

藤井(1991、1992)は、日本語の語順は類型論的分類ではＳＯＶ型と認められているが、実際の発話においては目的語が主語の前に来るケースを始め、述語の後ろに文の要素が来るケースまで多様性を示しており、特に述語の後ろに文の要素が来る倒置現象は、話し言葉において特徴的なものであると述べている。そこでは、述語の後ろに倒置された文の要素を主に主語、修飾語などに分類しているが、詳細な形態論情報に基づく分析はなされていない。倒置構文の使用条件、特に話し手が一般にいつ倒置構文を使うかという要因について、郭、伝(2011)は、倒置構文の形態論的特徴と倒置構文をよく使う話し手の年齢,性別などの社会言語学属性の影響を明らかにした。そのほか、多くの言語研究者が伝達機能、情報提示方略と文法論など様々な観点から分析してきた。しかし、倒置構文の形態論的特徴と言語外的要因、特に聞き手の存在による影響に関する議論はまだ不十分だと言える。

コミュニケーションは、話し手の行為と聞き手の行為の間の相互参照的関係を必須のものとして含むプロセスで、聞き手による話し手の心的状態の解釈は、聞き手の反応的な行動を通じて、元の話し手によっても解釈可能なものとなる(高梨・榎本、2009)。話し言葉でよく出現する倒置構文が、このような聞き手の反応的な行動を話し手が参照することを通じた調整的な行為であるという視点から、本稿では倒置構文を用いる要因を検討する。

2. 方法

2.1 対話データ

分析対象は『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese、以下 CSJ と略称)(前川、2004)の 2 人対話(インタビューデータ)である。

CSJ には全体で約 660 時間(語数にして約 700 万語)の自発音声が格納されている。その一部、約 45 時間(50 万語)分は、同コーパスの「コア」と呼ばれ、書き起こしテキストに加えて、(人手修正済み)形態論情報(小椋、2006)や節単位情報(丸山ほか、2006)、さらに分節音ラベルやイントネーションラベルが提供されている。本稿では、2 人対話(インタビ

ュー形式)の 12 対会話(3.6 時間、約 13 万語)を分析対象とした。インタビューデータは、学会講演ないし模擬講演関してインタビュアーが様々な質問を発し、話者がこれに答える形式の対話である。

2.2 倒置構文の認定

CSJ において、「コア」と呼ばれる約 50 万語分のデータでは、 書き言葉の文に相当する「節単位」が付与されている(丸山ほか、2006)。ただし、対話データでは、節単位は一方の話者に対してしか与えられていない。そこで、最近になって開発された CSJ-RDB(小磯ほか、2012)中の対話データを分析に用いた。このデータでは、両方の話者の節単位が認定されている。

節単位は以下のような考えに基づいている。「文」は、語とともに、言語表現を構成する最も基本的な単位として捉えられてきた。書き言葉では、書き手自身によって文末位置に句点が付与されるため、「文」の範囲を取り出すことは比較的容易である。それに対して、話し言葉では，話し手自身が文末位置を明示的に示すということはない。とすると、話し言葉の中から「文」の範囲を取り出すためには、何らかの手がかりをもとに「文」の範囲を認定しなければならない。節単位では、主に節末の表現形式に従って、「文」の切れ目を認定し、切れ目の強さに応じて「絶対境界」、「強境界」、「弱境界」の三種類を分類している(丸山ほか、2006)。

倒置構文は一つの節単位とみなされる。引用節中での倒置構文や言い換え、呼びかけなどは今回の分析対象と認めなかった。これらの倒置構文について以下のように本体と倒置要素を規定した。

帰るんですね、実家に
本体　　倒置要素

「帰るんですね、実家に」という文は倒置構文であり、本体の「帰るんですね」と倒置要素の「実家に」の二つの部分に分けた。倒置要素が付加される要因を検討するという観点から、次の二つの語の形態論情報に注目した。

です：本体の次末(penultimate)単語
ね：本体の末尾単語

これらを倒置構文を含まない発話(以下、非倒置文と呼ぶ)の次末単語、末尾単語の形態論情報と比較した。ただし、あいづちやフィラー文などは非倒置文に含めなかった。

本稿では、倒置構文 49 個と非倒置文 1561 個を分析対象とした。

2.3 聞き手反応の認定

本稿では、倒置構文の使用の要因として、聞き手の反応に注目する。そのため、倒置構文本体末尾付近(タイプ①)と倒置要素末尾付近(タイプ②)および非倒置文末尾付近(タイプ③)での聞き手の反応を抽出した。

聞き手の反応を抽出する範囲として、末尾位置の 0.5 秒前から 0.5 秒後までを「末尾付近」と見なした（図 1 参照)。

対話における発話の産出・理解には相互信念が必要であり、一方、その相互信念は対話の進行に伴う発話の産出・理解を通じて蓄積される(Clark,1996)。Clark(1996)によると、理解の証拠は主張(assertion)、前提(presupposition)、表示(display)と例示(exemplification)に分類され、その中で、主張の表現は主にうなずいたり、「うん」「はい」の類のあいづちを打つことである。一方、関連する次ターンを産出することは理解の「前提」とされる。本稿では、聞き手の反応を「あり」と「なし」に分けて、「あり」の場合は「あいづち」と「次ターン」に分けた。

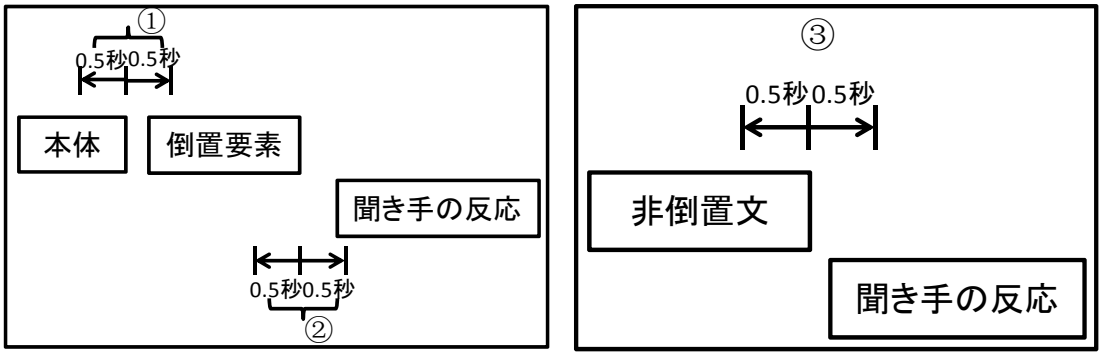


図 1　聞き手の反応

3.　結果と考察

3.1　形態論的要因

3.1.1　本体の末尾単語の品詞の分布

倒置構文の本体の末尾単語および非倒置文の末尾単語の品詞の度数を図 2 に示す。倒置構文の本体の末尾単語において、「よ」「ね」などの終助詞が 38 個あり、倒置構文全体の 77.6%を占めた。そのほかに、接続助詞が 12.2%を占め、助動詞が 8.2%を占めた。これに対して、非倒置文で末尾単語が終助詞である発話は全体の 47%に過ぎず、倒置構文の終助詞率はその 1.7 倍高い。そのほか、この図から、倒置構文の末尾単語の種類は非倒置文より少ないことが分かった。たとえば、格助詞、係助詞、動詞、形容詞、感動詞は非倒置文の末尾単語にしか出現しない。

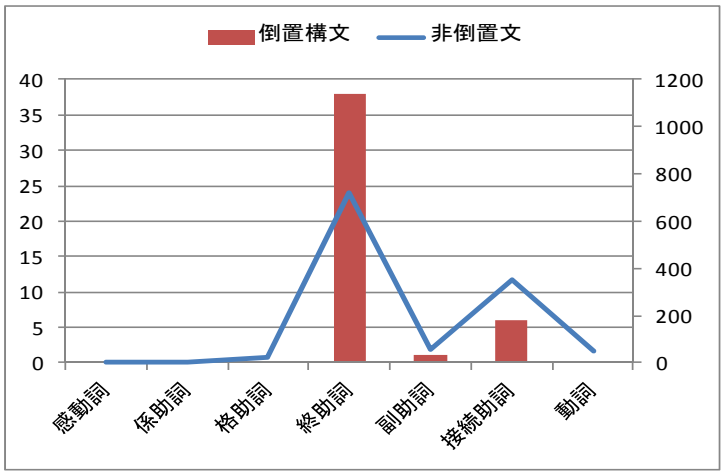


図 2　本体の末尾単語の品詞

3.1.2　本体の次末単語の品詞の分布

倒置構文の本体の次末(最後から 2 番目)単語および非倒置文の次末単語の品詞の度数を図 3 に示す。倒置構文の本体の次末単語において、「です」「ます」「だ」といった助動詞が 27 個で、本体の次末単語の 55.1%を占めた。しかし、非倒置文でも助動詞の比率は 52.9%であり、倒置構文との差は大きくない。それに対して、頻度が二番目の終助詞(13 個)は倒置構文全体の 26.5%であり、非倒置文の終助詞率 8.7%と比べて 3 倍あった。一方、動詞は 3 個で、倒置構文本体の次末単語の 6.1%しかなく、非倒置文の動詞率(14.5%)の半分以下で

ある。他には、格助詞、係助詞、形容詞や形状詞、感動詞などは非倒置文だけで出現し、倒置構文の次末単語として出現しなかった。以上のことから、倒置構文の本体の次末単語においても終助詞が多いことが明らかになった。

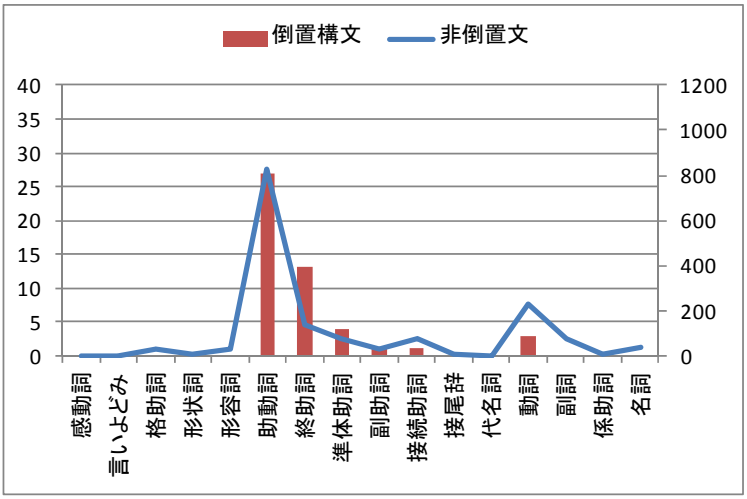


図 3 本体の次末単語の品詞の分布

本節の分析から、倒置構文の本体においては、末尾単語でも次末単語でも終助詞の比率が非倒置文と比べてかなり高いことが分かった。終助詞は一般に聞き手に働き掛ける形式であるため、次に聞き手の反応に注目する。

3.2 聞き手要因

倒置構文の本体末尾付近と非倒置文の節末付近の聞き手反応の度数を節の種類ごと、聞き手反応のタイプごとに表 1 に示す。

表 1 聞き手の反応

	絶対境界			強境界			弱境界		
	なし	あいづち	次ターン	なし	あいづち	次ターン	なし	あいづち	次ターン
倒置構文本体末尾付近	18 46%	9 23%	12 31%	2 50%	2 50%	0 0%	1 17%	4 67%	1 17%
非倒置文節末付近	366 39%	271 29%	294 32%	126 35%	194 54%	37 10%	144 53%	78 29%	51 19%

倒置構文本体末尾付近において、強境界と弱境界では倒置構文の絶対数が少なすぎるため、主に絶対境界の聞き手の反応を分析した。

表 1 から、絶対境界の場合、倒置構文本体末尾付近では、聞き手の反応がない場合が 46%あり、非倒置文の節末付近の 39%より高いということが分かった。

そこで次に、これらの本体末尾付近で聞き手の反応がない 18 事例について、倒置要素末尾付近での聞き手の反応を分析した。結果を表 2 に示す。

表 2 本体末尾付近で反応「なし」の時の
倒置要素末尾付近での反応

	なし	あいづち	次ターン
倒置要素末尾付近	11	1	6
	61%	6%	33%

表 2 から、本体末尾付近で聞き手反応がなかった事例のうち 40%近くで、倒置要素末尾付近で聞き手反応が得られていることが分かる。このようなパターンに当てはまる事例にはたとえば以下のようなものが見られた。

L:今何人ぐらいで(0.588)(F あ)(0.29)(D おひてぃ)基本的には一人で(0.121)やってらっしゃる(0.104)ですもんね　<- 本体
(0.12)
この(0.585)研究　<- 倒置要素
R:(F えっとー)(0.434)(F あ)(D む)(0.137)僕の所属してる研究室で(0.44)は話し言葉の認識をやってるのは(0.143)(D む)そうですね僕一人ですね

(D04M0010:597.2014-612.0856)

このことは、話し手が本体を話した直後に聞き手の反応がないことに気づいて、倒置要素を付加した結果、聞き手の反応が得られ、相手との相互信念を形成することに成功したことを示唆する。これは倒置構文を使用する一つの要因と言えるだろう。

4. おわりに

本研究では、まず、倒置構文の本体の次末単語、末尾単語の形態論的特徴を分析して、非倒置文の対応する位置の単語の品詞と比較して、以下の結果が得られた。

① 倒置構文の本体の末尾単語において、「よ」「ね」などの終助詞が倒置構文全体の77.6%を占め、非倒置文の末尾単語の終助詞率よりかなり高かった。

② 倒置構文の本体の次末単語においても、終助詞は非倒置文と比べてずっと多かった。

これらのことから、倒置構文の使用は、聞き手への働き掛けと関係していると言える。

そこで、倒置構文本体末尾付近での聞き手の反応を抽出した結果、絶対境界では非倒置文節末付近と比べて、聞き手の反応がない比率が高いということが分かった。さらに、これらの倒置構文では倒置要素を付加することにより聞き手反応を引き出せているケースが多く見られた。これらのことから、話し手が発話中に聞き手の反応がないことに気づいて、倒置要素を付加することで、相手との相互信念の形成に成功していることが示唆された。

本稿の聞き手の反応の分析結果は、話し手が聞き手に反応を求め、聞き手の反応が得られない場合に、聞き手の理解を助けるため、倒置要素を産出しているという可能性を示唆する。これは、倒置構文の使用に聞き手との相互信念の形成の役割を認める、まったく新しい視点である。今後,雑談などの会話データを使ってほかの会話タイプでもこのような分析を進めたい。

文　献

Clark, H. H. (1996) *Using language*,Cambridge University Press

大坊郁夫、磯友輝子(2009)「対人コミュニケーション研究への科学的アプローチ」大坊郁夫、永瀬次郎（編）『講座社会言語科学 3：関係とコミュニケーション』, pp. 2-35, ひつじ書房

江口巧（2000）「日本語の倒置文：情報提示の方略」『言語文化論究』12 , pp. 81-93 、九州大学大学院言語文化研究院

藤井洋子（1991）「日本語文における語順の逆転―談話語用論的視点からの分析」『言語研究』99, pp. 58-81
藤井洋子（1992）「日本語の会話文における主節前置の談話語用論的分析」『放送大学研究年報』10, pp. 103-122
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』
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
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書 124: 日本語話し言葉コーパスの構築法』, pp. 255-322
小椋秀樹（2006）「形態論情報」『 国立国語研究所報告書 124: 日本話し言葉コーパスの構 築法』, pp. 133-186
高梨克也、榎本美香（2009）「『特集-聞き手行動から見たコミュニケーション』編集にあたって」『認知科学』16(1), pp. 5-11

現代日本語書き言葉均衡コーパスに基づく外来語音の表記に関する試論

単　珊（東京学芸大学大学院）
白勢　彩子（東京学芸大学）

Notation of the Loanword Based on the Balanced Corpus of Contemporary Written Japanese

San Zen (Tokyo Gakugei University)
Ayako Shirose (Tokyo Gakugei University)

1. はじめに

外来語とは、「自国語に組み入れられた外国語のことで、厳密にはr a d i o(レイディオウ)[reidiou]がラヂヲ、ラジオなどと語形をとるように、受け入れた国の音韻組織によって再構成された語形をもつ語」（杉本, 2008）を指す。現在では、外来語と言えば主として欧米系の諸言語に由来するものを指すことが多く、片仮名で書き表すのが一般的である（国語審議会, 1991）。外来語の片仮名表記については、基本的には「ア」、「カ」、「キャ」など国語の音に対応する仮名が用いられているが、これらに加え、外来語の原音に応じた、特別の仮名表記が工夫され、現在、表記のよりどころとして、平成三年（1991 年）に公布された「外来語の表記」がある。この「外来語の表記」には、外来音に対応する仮名として「シェ」、「ジェ」、「トゥ」、「ドゥ」などが挙げられている。これらは、例えば「いつもそう書かなければならないことをその意味するものではない」とされ、慣用により「セ」「ゼ」などとしてよいとされている。すなわち、同一の外来音に複数の片仮名が用いられることが許容されることとなっているのが現状である。

外来語の表記には、複数の表記、いわば「ゆれ」があることになり、この問題についてはいくつかの研究がある。福盛（2010）は、「ブ」と「ヴ」など数種に対して、調査を行ない、例えば、「ブ」と「ヴ」では、比較的に近年に取り入れられた外来語について「ヴ」の使用が増えていることが得られ、また、固有名詞で登録されている場合に、より多くの使用が認められることがわかった。この結果は地名・人名は原音に近い形で表記されることが多いことによると考えられた。松崎（1992）では、辞書類における外来語音の表記の「ゆれ」の実態を調査して、ゆれの類型が明らかとされ、続く松崎（1993）では、17 種の辞典類の外来語表記を調査し、「ゆれ」が視覚化され、ある拍が外来語音寄りか日本語音寄りかは語により異なることが数量的に再確認されている。しかしながら、これらの研究では調査対象の範囲が限られている、辞書を調査したもので必ずしも表記の実態を捉えたものではないなどの問題がある。さらにいえば、そもそも外来語の原語の音に対し、どのような片仮名が対応するのかといった基本原則も明瞭に示されてきていないようである。

そこで、筆者らは外来音と仮名表記の対応について、基本則を明らかとし、使用の実態を明らかにすることを目的とした研究を進めている。本稿では、原音と表記の関係についての概略を示し、これに基づき「現代日本語書き言葉均衡コーパス（BCCWJ）」を調べた結果について報告する。

なお、本稿では、「シェ」「ジェ」「トゥ」といった外来音に対応する仮名を「外来音仮名」と呼び、「ア」「カ」「キャ」など国語の音に対応する仮名を「国語音仮名」とする。

2. 原音と表記の対応関係

2.1 外来音仮名と国語音仮名の関係

現行の「外来語の表記」では、「ミルクセーキ」と「ミルクシェーキ」の二様の表記が可能となる。松崎（1992, 1993）は、外来音仮名と国語音仮名の両用がある時、二者にどのような関係があるかを調査した。これによると、1 拍にあたる外来音仮名と国語音仮名は、A : 1 拍にあたる外来音と国語音 2 拍分がゆれているもの、B : 母音を共通要素として、外来音と国語音がゆれているもの、C : 子音を共通要素として、外来音と国語音がゆれているものの 3 類に関係性が整理されている。A の例としては、「シェ」と「ジェ」、B の例としては、「シェ」と「セ」、C の例としては、「シェ」と「シュ」がある。このように、外来語の表記において、同一の音に対し、仮名が複数、存すると考えられるが、原語での発音に対して片仮名表記がどのようになされているかについては、明瞭でない。そこで、原音（原語の発音）と片仮名表記がどう対応しているか調べるため、以下の調査を行った。

2.2 原音に基づく表記調査

調査は、『コンサイスカタカナ語辞書　第 4 版』を資料とし、前述の松崎（1992）で指摘された外来音仮名とそれを国語音で表記する場合の仮名のいずれかが含まれ、「ア」「イ」で始まる見出し語を調査対象とした。これらの語について、語の原綴りと原発音記号を調べた。調査の対象となった語は合計 2,036 語であった。うち、国語音仮名でのみ表記される発音があり、これらは分析対象外とした。データの結果を表 1、表 2 に示す。表 1、2 の「分類」にある「c」、「v」は、それぞれ、子音、母音を指す。「cv」 は、子音＋母音の音節、「cvv」は子音＋二重母音の音節、表 2 の「c.v/c」は、子音音節の後に母音、あるいは子音が続く音節、「cc.v/c」は二重子音の後に母音、あるいは子音が続く音節、「c#」とは、子音で終わる場合を示している。[　]内に原音の音声記号を示し、ハイフンで対応する仮名を示した。

表 1 に示すのは、原音に対し仮名表記が 1 種のみ、単独に対応する関係である。「cv」では[da]に対する「ディ」、[dø]に対する「ドゥ」、「cvv」では[fəu]に対する「フォ」などがある。仮名を軸にして捉えると「ドゥ」は[dø]、[du]の 2 種の原音に対し、用いられていることがわかる。

表 2 に示すのは原音に対し仮名表記が 2 種以上、複数に対応する関係である。最大で 7 種の仮名で表記される音があることが示されている（[tə]に対する「チ」、「ツ」、「テ」、「ト」、「チュ」、「ティ」、「トゥ」）。対応する仮名の数は 2~7 種の範囲にあるものの、2 種の仮名が用いられる音が最も多いことがわかった。2 種の仮名が用いられるのは例えば[tu]に対する「ツ」、「トゥ」などである。

表 1 と表 2 を比較すると総数に大きな差はない（「出現数」の総数：表 1 が 22、表 2 が 19）。このことは必ずしも外来音仮名で表記される音が複数の表記を持つわけではないことを意味する。表 1 のように音と仮名が一対一に対応するものは原音からみて「ゆれ」がなく、一方、表 2 のように複数あるものは原音からみて「ゆれ」があるといえる。

表 1　単独に対応する仮名

分類	原音と仮名の対応	出現数
cv	[da]―ディ　[dø]―ドゥ　[du]―ドゥ　[fa]―ファ　[fæ]―ファ [fe]―フェ　[fɛ]―フェ　[fə]―フィ　[sə]―シュ　[ʃu]―シュ [tʃɛ]―チェ　[tʃu]―チュ　[dʒe]―ジェ　[dʒu]―ジュ　[dj]―ディ [dju]―デュ　[wɛ]―ウェ	17
cvv	[fəu]―フォ　[tʃuə]―チュ	2
c.v/c	[dju]―デュ	1
cc.v/c	[fy]―フュ	1
c#	[ʒ]―ジュ	1

表 2　複数に対応する仮名

対応する仮名の数		原音と仮名の対応	出現数
2	cv	[tu]―ツ・トゥ　[fə]―ファ・フォ [fɔ]―ファ・フォ　[va]―バ・ヴァ [vɛ]―ベ・ヴェ　[tʃə]―チュ・フュ [dʒi]―ジ・ジェ　[tju]―チュ・トゥ [we]―ウェ・クエ	11
	cvv	[ʃei]―シエ・シュ	
	c.v/c		
	cc.v/c		
	c#	[ʃ]―シェ・シュ	
3	cv	[fi]―ヒ・フィ・フェ [se]―セ・ゼ・チェ [dʒə]―ジ・ゼ・ジェ	5
	cvv		
	c.v/c	[d]―ジ・ディ・ド [s]―ズ・セ・シュ	
	cc.v/c		
	c#		
4 以上	cv	[di]―ジ・デ・ティ・ディ [ti]―ズ・チ・テ・ト・ド・ティ [tə]―チ・ツ・テ・ト・チュ・ティ・トゥ	3
	cvv		
	c.v/c		
	cc.v/c		
	c#		

表 2 にあるように、原音に対し、仮名表記が複数ある場合、音と仮名がどのような関係にあるかについて、音の側面から規則性があるかどうか検討した。原語の発音を仮名の発音と比較し、母音が同じかどうか、子音が同じかどうかの観点により整理した。例えば、原語の発音記号が[va]であるものと対応する仮名表記「バ」の音声は、[ba]である。母音が同じ[a]であるが、子音はそれぞれ[v] 、[b]と異なっているといえる。

[va]とバ[ba]の関係は、子音が非共通で母音が共通する「子音非共通型」に分類できる。現在、この「子音非共通型」について、音源（有声・無声）・調音点・調音様式の面から、相違点を整理し、詳細に分析している。結果としては、例えば、[dʒi]とジ[dzi]のように硬口蓋性の有無で関係づけられるものが多く見られるなど法則性があることが示唆されている。子音共通で母音が非共通であるものは「母音非共通型」に分類でき、[tu]とトゥ[tɯ]、[di]とデ[de]などの例がある。これについても母音の発音の特徴から整理することを試みている。

3. コーパス調査

3.1 調査概要

上記のように、原音と片仮名表記の対応関係を把握したが、この結果に基づいて、BCCWJ における外来音表記の用いられ方を調査し、表記と発音間のゆれの現状を把握したい。上記では、原音と片仮名表記の対応関係を「子音非共通型」、「母音非共通型」として捉えた。これらのタイプのうち、今回は「子音非共通型」の「ヴァ」、「母音非共通型」の「トゥ」を抽出し、BCCWJ にて検索し、結果を整理する。前述の仮名を含む語のうち、出現語数がもっとも多かった語の上位 5 位から、それぞれ「子音非共通型」2 件、「母音非共通型」1 件を抽出した。各自の語彙素で BCCWJ により検索し、得た結果をジャンル別の出現頻度から、相違点を検討する。

3.2 調査結果と分析

外来音仮名「ヴァ」の検索結果より、語数がもっとも多かった上位五位のうち、「ヴァ/バイオリン」と「ヴァ/バイオレット」の二語を抽出し、それぞれのジャンル別の出現頻度を表 3 と表 4 にまとめた。

表 3 に示したように、「ヴァ/バイオリン」という語においては、「ヴァ」の表記を用いる語が 415 語あった。それに対して、「バ」の表記を用いる語は 293 語であった。「ジャンル」により各語の出現傾向を見ると、「ヴァイオリン」については、第五位までの語で出現頻度が全体の 73.7%を占め、その他のジャンルは 26%程度である。つまり、「ヴァイオリン」の出現頻度が、第五位までのジャンルに偏っているといえる。特に「芸術・美術」のジャンルに多く出現している（43.3%）。

一方、「バイオリン」では、ジャンル中、第五位までの語の合計が全体の 52.2%を占め、その他のジャンルは 47.8%となっている。「ヴァイオリン」に見られたような第五位までに集中して出現する傾向はないといえる。第五位までの各ジャンルについては、「文学」19.5%、「芸術・美術」11.6%、「社会科学」8.2%、「エンターテインメントと趣味」6.8%、「関東地方」6.1%であり、「文学」での出現頻度が高いものの、いずれかのジャンルに偏って出現する結果とはなっていない。つまり、「バイオリン」は平均的に分布しているといえる。

以上、まとめると、「ヴァイオリン」については、「芸術・美術」や「文学」という特定のジャンルに傾いて出現する傾向がみられる。これに対し、「バイオリン」は各ジャンルに平均的に分布されていることがわかった。

表 4 に示した「ヴァ/バイオレット」という語について見ると、「ヴァ」の表記を用いる語が 115 語、「バ」の表記を用いる語が 38 語あった。「ヴァイオレット」においては、ジャンルの第五位までで出現頻度が 97.5%となり、その他のジャンルは 2.5%に過ぎない。中でも、「文学」のジャンルにおける出現率がもっとも多く、91.3%に達している。つまり、「ヴァイオレット」は非常に偏って特定のジャンルに出現しているといえる。一方、「バイオレット」では、第五位までの出現頻度が 76.3%で、その他のジャンルが 23.7%である。上位第五位までのジャンルの中で、それぞれ、「技術・工学」26.3%、「総合」23.7%、「文学」10.5%、「分類なし」7.9%、「科学」7.9%となっている。「バイオレット」については、特定のジャンルにやや偏っているものの、「ヴァイオレット」に比しては平均的に分布しているといえる。

表 3「ヴァ/バイオリン」の検索結果

語例	総語数	ジャンル	出現頻度
ヴァイオリン	415	芸術・美術	43.3(%) 180
		文学	14.5 60
		エンターテインメント	8 33
		芸術	4.8 20
		Yahoo!サービス	3.1 13
		その他	26.3 109
バイオリン	293	文学	19.5 57
		芸術・美術	11.6 34
		社会科学	8.2 24
		エンターテインメントと趣味	6.8 20
		関東地方	6.1 18
		その他	47.8 140

表 4「ヴァ/バイオレット」の検索結果

語例	総語数	ジャンル	出現頻度
ヴァイオレット	115	文学	91.3(%) 105
		分類なし	3.5 4
		哲学	0.9 1
		社会科学	0.9 1
		自然科学	0.9 1
		その他	2.5 3
バイオレット	38	技術・工学	26.3 10
		総合	23.7 9
		文学	10.5 4
		分類なし	7.9 3
		科学	7.9 3
		その他	23.7 9

1「総語数」はそれぞれ、語彙素である「バイオリン」と「バイオレット」で検索した語数のことである。

2「ジャンル」の欄では、検索結果の上位五位までを載せる。五位以下は「その他」に含めた。

3「出現頻度」の上段には、総語数に対する相対頻度をパーセンテージで示し、下段には実数を示した。

以上、まとめると、「ヴァイオレット」については、特定のジャンルに非常に偏向していており、これに対して、「バイオレット」ではいくつかのジャンルに偏るものの、「ヴァイオレット」に見られるような特定のジャンルに集中して分布する様相はないことがわかった。

続いて、外来音仮名「トゥ」の検索結果の中、語数がもっとも多い上位五位のうち、「トゥール」を抽出し、調査結果を表 5 にまとめた。「トゥール」に対応する「ツール」についても同様に表 5 に示す。

表 5「トゥ/ツール」の検索結果

語例	総語数	ジャンル	出現頻度
トゥール	107	文学	63.6(%) 68
		歴史	12.1 13
		芸術・美術	6.5 7
		総合	6.5 7
		社会科学	4.7 5
		その他	6.6 7
ツール	2972	インターネット、PC と家電	20.4 607
		総記	16.2 482
		技術・工学	12.6 374
		社会科学	5.6 166
		自然科学	4.0 117
		その他	41.2 1226

1「総語数」はそれぞれ、語彙素である「トゥール」と「ツール」で検索した語数のことである。

2「ジャンル」の欄では、検索結果の上位五位までを載せる。五位以下は「その他」に含めた。

3「出現頻度」の上段に、総語数に対する相対頻度をパーセンテージで示し、下段には、実数を示した。

表 5 に示したように、「トゥ/ツール」という語では、「トゥ」の表記を用いる語が 107 語あり、「ツ」の表記を用いる語が 2972 語あった。「ヴァ/バイオリン」、「ヴァ/バイオレット」では外来音表記の方が多く出現する傾向が見られており、「トゥ/ツール」には異なる特徴が見受けられる。これまでと同様、「ジャンル」の出現傾向を見ると、「トゥール」では、第五位までの語の出現頻度が全体の 93.4%を占め、その他のジャンルの出現頻度は 6.6%である。加えて、上位五位のジャンル中、「文学」のジャンルに集中して出現しており、2 位の「歴史」と合わせると 75%を超える。このことから、「トゥール」もまた、特定のジャンルに偏って出現しているといえる。

「ツール」では、第五位までの語の出現頻度で 58.8%となり、各ジャンルの出現頻度がそれぞれ、「インターネット、PC と家電」20.4%、「総記」16.2%、「技術・工学」12.6%、「社会科学」5.6%、「自然科学」4%であり、「インターネット、PC と家電」での出現頻度が高いものの、いずれかのジャンルにのみ偏って出現する結果とはなっていない。つまり、「ツール」は比較的平均して複数のジャンルに分布しているといえる。

以上、まとめると、「トゥール」が特定のジャンルに傾く傾向があり、これに対し、「ツール」では、特定のジャンルへの偏りはなく分布しているといえる。

4. まとめ

本研究は、外来音の仮名表記について、表記の現状を把握するため、まず、原音と仮名表記の関係を示し、次いで、「シェ」、「チェ」といった外来音に対応する仮名の表記を調査の対象とし、外来語表記の用いられ方を、BCCWJ により検索し、ジャンルによる出現の様相の相違について、検討した。

検討の結果、次の点が明らかとなった。

1. 原音と仮名の関係を整理したところ、外来音仮名に対応する音が必ずしも複数の表記を持つわけではないことがわかった。
2. 「ヴァイオリン」については、「芸術・美術」という特定のジャンルに傾く傾向がみられた。これに対し、「バイオリン」は各ジャンルに平均的に分布されていることがわかった。
3. 「ヴァイオレット」については、特定のジャンルに非常に偏向して出現しており、これに対して、「バイオレット」ではいくつかのジャンルに偏るものの、「ヴァイオレット」に見られるような特定のジャンルへの偏向はないことがわかった。
4. 「トゥール」については、「文学」という特定のジャンルに傾く傾向があり、これに対し、「ツール」では特定のジャンルへの偏りはなく、比較的平均して分布していることがわかった。

以上をまとめると、「ヴァ」、「トゥ」の外来音仮名が用いられた場合では、いくつかの特定のジャンルに傾いて出現する傾向がみられるといえる。これに対し、国語音仮名が用いられた場合では、特定のジャンルへの偏りはなく、複数のジャンルに分布する様相があることが示された。今回、こうした分析を行うことにより、外来語の表記において、これまで把握されにくかった原音と仮名の関係を明瞭化する指針を得ることができ、また、大規模コーパスを用いることによって外来語表記における仮名の使用状況の実際を整理して、仮名による相違点を検討することができ、今後の研究の基礎的な論点を示すことができたと考えている。

上記のような点が明らかとなったが、以下の問題点も存在している。辞書に基づく表記

調査においては、収集したデータの結果が一部分の外来音仮名に偏る現象が見られ、十分なデータを得たとは言いがたい。また、コーパス調査においては、語彙素を用い、検索を行うことなど調査方法の妥当性について、検討の余地がないとはいえない。コーパス調査のデータに関しては、未整理のデータが含まれていることも今後の課題である。例えば、外来音「トゥ」の検索結果では、もっとも数が多かった外来音の「トゥ」と対応する原語が「to」と「two」の両方が含まれており、今回は、「トゥ」についての分析を省略した。今後は、まず、十分なデータを得たうえ、コーパス調査の手法について十分検討した上でデータを議論する必要がある。さらには、「ジャンル」以外の項目についても、詳細な分析を進め、外来語を仮名表記することに関連・影響する要因を明らかにしていきたいと考えている。

文　献

福盛貴弘(2010)『基礎からの日本語音声学』東京堂出版.

国語審議会(1991)「外来語の表記」前文『公用文の書き表し方の基準(資料集)増補二版』, pp.197-201.

松崎寛(1992)「外来語音におけるゆれの類型：辞書類の表記を中心として」言語学論叢, 10:11, pp.43-56.

松崎寛(1993)「外来語音の表記のゆれに関する定量的研究」東北大学文学部日本語学科論集, 3, pp.83-94.

三省堂(2010)『コンサイスカタカナ語辞典　第4版』三省堂編修所.

杉本つとむ(2008)「外来語」『日本語学研究事典』pp.408-411, 明治書院.

「リアル」を構成要素とする複合名詞の語彙的特徴

渡邊　ゆかり（広島女学院大学文学部）

Lexical Features of Japanese Nouns Compounded from '*RIARU*' and Other Morphemes

Yukari Watanabe (Hiroshima Jogakuin University)

１．はじめに

英語の形容詞'real'から作られた外来語の「リアル」は、近代に登場して以降、日本語の中に広く浸透し、その用法も多様化していった。近年では、「リアルな」「リアルに」といった連体修飾用法や副詞的用法のみならず、「リアルタイム」「リアルクローズ」「リアル友達」など様々な複合名詞の構成要素としても多用されている。このような「リアル」を構成要素とする複合名詞には、英語（あるいはヨーロッパの言語）の「'real'＋名詞」という形をとる名詞句からの借用にあたるものから日本で独自に作られたものまで数多く存在する。また、その語彙バリエーションは、時代や使用ジャンルといった位相の影響を大きく受けている。本研究では、Web 上で利用可能なコーパスや検索システムを用いて収集した用例の調査に基づき、「リアル」を構成要素とする複合名詞の語彙バリエーションが、こうした位相とどのような関わりを持つのかを明らかにする。

２．調査方法

調査に際しては、次の表 1 に示す、Web 上で利用可能なコーパスと検索システムを用い、「リアル」を構成要素とする複合名詞を収集した。なお、③については、「リアル」を検索語として文字列検索したところ、検索結果が表示上限数の 500 件を超えていたので、表示された 500 例の中から、「リアル」と語基形態素が結び付いてできている複合名詞のみを取り出し、分析を行った。

表 1　調査に利用したコーパス

①KOTONOHA 現代日本語書き言葉均衡コーパス小納言の書籍ジャンル（1971 年 - 2005 年、21,943 件、約 6,230 万語）（以下「書籍」と略称する）
②KOTONOHA 現代日本語書き言葉均衡コーパス小納言の雑誌ジャンル（2001 年 - 2005 年、1,989 件、約 440 万語、以下「雑誌」と略称する）
③KOTONOHA 現代日本語書き言葉均衡コーパス小納言の Yahoo!ブログジャンル（2008 年、52,680 件、約 1,030 万語、以下「ブログ」と略称する）
④Web 上で公開されている 1947 年 - 2010 年の国会会議録

その後、①から収集された語彙の分析結果を基盤に、専門領域の相違という観点からまず

①と④の［リアル＋名詞］の語彙バリエーションを比較し、次に、情報の伝達目的の相違という観点から①と②③各々の［リアル＋名詞］の語彙バリエーションとを比較した。以下、これらの分析結果を見ていく。

３.書籍における［リアル＋名詞］

次の表 2 は、書籍に存在した［リアル＋名詞］もしくは ［名詞＋リアル］を構成要素として含む普通名詞の異なり語数と延べ語数である。

表 2　書籍中の［リアル＋名詞］［名詞＋リアル］を構成要素として含む普通名詞の異なり語数と延べ語数

	前項 or 後項が外来語	前項 or 後項が外来語以外	合計
異なり語数	47	13	60
異なり語数の割合（%）	78.33	21.67	100.00
延べ語数	257	21	278
延べ語数の割合（%）	92.45	7.55	100.00

表 2 から「リアル」と結び付く名詞の多くは外来語であることがわかる。なお、［名詞＋リアル］を構成要素とするものは、「［バーチャル＋リアル］ 体験」「［ばーちゃん＋リアル］体験」の 2 例のみであった。このことから、「リアル」は基本的に［リアル＋名詞］という形で後項名詞の意義を限定するのに使用されることがわかる。次に、［リアル＋名詞］を構成要素とする普通名詞に限定し「リアル」の意義を調べたところ、表 3 に挙げた A1－H の意義が存在した。

表 3　［リアル＋名詞］における「リアル」の意義（書籍の普通名詞）

A1	「偽物の」に対する「本物の」「正真正銘の」の意
A2	「簡易的な」「一般的な」に対する「本格的な」「ハイクラスの」の意
B	「真実らしさに欠けている」に対する「真に迫っている」の意
C1	「空想の」「虚構の」に対する「現実の」「実際の」「実社会の」の意
C2	「サイバー上の」に対する「現実の」「実際の」「実社会の」の意
D	「現在と一致しない」に対する「現在と一致する」の意
E	「リアルタイムの」の意
F	「80286 以降の CPU を、これ以前の 8086 の CPU として扱う」意から転じた「実際よりも能力が劣っている」の意
G	「理念を重視する」に対する「現実の力関係や利益を重視する」の意
H	ドイツの ‘realschule（実科学校）’の形態素の一部が表す「実科の」の意

なお、A1 と A2 の「リアル」は、後項名詞がその名に相応しいものであることを表してい

る点において共通しているので、A の下位カテゴリーとして位置づけた。また、C1 と C2 も、「リアル」が〈現実の〉〈実際の〉〈実社会の〉という意義を含んでいる点で共通しているので、C の下位カテゴリーとして位置づけた。

各意義に対応する［リアル＋名詞］には、次の表 4 に挙げる要素が存在した。

表 4　意義ごとに見た［リアル＋名詞］の異なり例（書籍の普通名詞）

A1	［リアル＋モカシン］1995、［リアル＋ファイト］2001、［リアル＋夫婦本番］2002、［リアル＋ドキュメント］2003、［リアル＋ファー］2005
A2	［リアル＋インターネット］2001
B	［リアル＋描写］1989、［リアル＋イメージ］1990、［リアル＋映像］2002、［リアル＋体験］2002
C1	［リアル＋ランド］1994、［リアル＋画像］2005
C2	［リアル＋世界］1997、［リアル＋イベント］2001、［リアル＋店舗］2001、［リアル＋空間］2002、［リアル＋市場］2002、［リアル＋社会］2002、［リアル＋ワールド］2002、［リアル＋商店］2004、［リアル＋美少女］2005、［リアル＋マネー］2005
D	［リアル＋タイム］1987
E	［リアル＋ネットワークシステム］2001、［リアル＋オッズ］2003
F	［リアル＋モード］1996
G	［リアル＋ポリティック］1998
H	［リアル＋シューレ］2001

表 4 より、第一に、1997 年から登場しているC2 タイプの［リアル＋名詞］の異なり数が 10 例と他と比べて多めであることがわかる。その背景には、1997 年から 2010 年にかけてのインターネット利用人口の急激な増加が大きく関与していると考えられる。総務省の「通信利用調査」[1]によれば、1997 年末のインターネット利用人口率は、9.1%であるが、2010 年末には、78.2%に達している。またインターネット利用人口の増加とともに、一人あたりのインターネット利用時間やインターネットコミュニティ、Eコマース（電子商取引き）などの利用者も増加している。このような我々の生活スタイルの変化が、 C2 の「リアル」を用いた複合名詞のバリエーションの増加に繋がったものと見られる。

第二に、D、F－H の各タイプの「リアル」は、その他のタイプに比べて、意義が特殊化しており、後項名詞との結び付きが強い。それゆえ、造語力（生産性）に乏しく後接可能な名詞は限定されている。

最後に、2001 年に登場している E の「リアル」は、D の「リアルタイム」の縮約形にあたり、リアルタイムで動くものやリアルタイムで伝達されるものを表す名詞が後接する。

[1] 総務省が Web 上で公開している統計調査データベースにおいて調査結果が公開されている。（www.soumu.go.jp/johotsusintokei/statistics/statistics05a.html）

以上、本節では、書籍における［リアル＋名詞］の語彙バリエーションについて分析した。次に、専門領域の相違という観点から、国会会議録における［リアル＋名詞］の語彙バリエーションを書籍の語彙バリエーションと比較する。

４．国会会議録における［リアル＋名詞］

次の表 5 は、国会会議録に存在した［リアル＋名詞］を構成要素とする普通名詞における「リアル」の意義である。なお、意義カテゴリーの種類を表すアルファベットは、表 4 に準じ、表 4 に存在しないものは新たに追加した。以後も新たな「リアル」の意義を挙げる際は、この方法を用いることとする。また、表 4 の B と意義的に近いものが存在したので、これを B2 とした。従って、以後、表 4 の B は B1 として扱う。また、K1 - K4 は、「実質」という意義を含んでいるという点で共通しているので、K の下位カテゴリーとして位置づけた。

表 5　［リアル＋名詞］における「リアル」の意義（国会会議録の普通名詞）

A1	「偽りの」に対する「本物の」「正真正銘の」の意
B2	「実体・実態からかけ離れている」に対する「実体・実態通りの」の意
C1	「空想上の」「虚構の」に対する「現実の」「実際の」「実社会の」の意
C2	「サイバー上の」に対する「現実の」「実際の」「実社会の」の意
D	「現在と一致しない」に対する「現在と一致する」の意
E	「リアルタイムの」の意
G	「理念を重視する」に対する「現実の力関係や利益を重視する」の意
I	英語の‘real estate（不動産）’という連語を構成する単語の一部、「実質の」という意の解釈とスペイン語の‘real’に由来する「王の」という意の解釈がある
J	「三次元時空の」に対する「三次元時空に時間を加えた四次元時空の」の意
K1	「物価変動の影響がある」に対する「物価変動の影響を除外した（実質的な）」の意
K2	「関税が付加された」に対する「関税が付加される前の（本来の、実質的な）」の意
K3	「国の経済対策のうち経済成長率（GDP）を直接（実質的に）押し上げる効果のある」の意
K4	「今後かかることが予想される」に対する「実際にかかった（実質的な）」の意
L	「無形の」に対する「実体のある・実物の」の意
M	「資産市場の」に対する「消費財や投資財の生産と分配に関わる財市場、労働市場の」の意
N	「後項要素のイベントの実現」の意

各意義に対応する［リアル＋名詞］には、次の表 6 に挙げる要素が存在した。

表 6　意義ごとに見た［リアル＋名詞］の異なり例（国会会議録の普通名詞）

A1	［リアル＋デイフエンスフオース］1953、［リアル＋ウイル］1956、 ［リアル＋ピース］1991
B2	［リアル＋画像］1998、［リアル＋イメージ］1999
C1	［リアル＋ビジネス］1994、［リアル＋リスク］2003、［リアル＋ワールド］2010
C2	［リアル＋空間］1999、［リアル＋ワールド］2000、［リアル＋取引］2001、 ［リアル＋スペース］2002
D	［リアル＋タイム］1986
E	［リアル＋通信］2000
G	［リアル＋ポリティック］1999
I	［リアル＋エステート］1999
J	［リアル＋ワールド］1995
K1	［リアル＋ウェイジ］1953、［リアル＋ターム］1956、［リアル＋ベース］1982、 ［リアル＋エコノミー］1998
K2	［リアル＋プライス］1964
K3	［リアル＋マネー］1987、［リアル＋ウオーター］1993
K4	［リアル＋コスト］2003
L	［リアル＋リソーセス］1976
M	［リアル＋経済］2002
N	［リアル＋バンククリーンアップ］2002

表 5 に挙げた「リアル」の意義のうち書籍に存在しなかったものは、B2 と I - N である。B2 を除く I - N の「リアル」は、政治、経済、科学技術と関わる特殊な意義を表しており、後項名詞との結び付きが強く、後接可能な名詞は限られている。また、いずれも、専門用語的性格が強く、その中には、K1 の［リアル＋ターム］と［リアル＋ベース］、K3 の［リアル＋マネー］と［リアル＋ウオーター］のように、同一の知的意味を表しながら、用語が固定していないものも存在する。いずれを用いるかは、発話者の相違による要因もなくはないが、主に時代の相違と対応しており、延べ 6 例存在した［リアル＋ターム］（発言者は異なり数 2 名）は 1956 年と 1964 年に使用されており、延べ 5 例存在した［リアル＋ベース］（発言者は異なり数 1 名）は 1982 年と 1984 年に使用されている。また、延べ 12 例存在した［リアル＋マネー］（発言者は異なり数 7 名）は 1987 年に使用されており、延べ 3 例存在した［リアル＋ウオーター］（発言者は異なり数 3 名）は 1993 年に使用されている。

次に、「リアル」の意義ごとに見た［リアル＋名詞］の異なり数の割合と延べ数の割合を、書籍と国会会議録とで比較した。次頁の表 7 は、書籍、国会会議録の各々における、「リアル」の意義ごとに見た［リアル＋名詞］の異なり数と延べ数ならびにそれぞれの割合を示している。なお、延べ数の割合は、いずれも D が他のタイプより格段に高いので、D 以外については、 D の延べ数を除外した上で比較することとした。表 7 中の（　）の数値は、

D の延べ数を除外して算出した数値に相当する。以後も、各意義の延べ数とその割合を挙げる際は同様の方法を取ることとする。

表 7　［リアル＋名詞］の異なり数、延べ数とその割合（書籍と国会会議録の普通名詞）

	異なり数とその割合				延べ数とその割合			
	書籍		国会会議録		書籍		国会会議録	
	数	割合（%）	数	割合（%）	数	割合（%）	数	割合（%）
A1	5	17.86	3	10.71	18	(33.96)	5	(6.85)
A2	1	3.57	0	―	1	(1.89)	0	―
B1	4	14.29	0	―	4	(7.55)	0	―
B2	0	―	2	7.14	0	―	5	(6.85)
C1	2	7.14	3	10.71	2	(3.77)	4	(5.48)
C2	10	35.71	4	14.29	19	(35.85)	9	(12.33)
D	1	3.57	1	3.57	223	80.80	1,271	94.57
E	2	7.14	1	3.57	3	(5.66)	1	(1.37)
F	1	3.57	0	―	1	(1.89)	0	―
G	1	3.57	1	3.57	4	(7.55)	4	(5.48)
H	1	3.57	0	―	1	(1.89)	0	―
I	0	―	1	3.57	0	―	5	(6.85)
J	0	―	1	3.57	0	―	5	(6.85)
K1	0	―	4	14.29	0	―	14	(19.18)
K2	0	―	1	3.57	0	―	1	(1.37)
K3	0	―	2	7.14	0	―	15	(20.55)
K4	0	―	1	3.57	0	―	1	(1.37)
L	0	―	1	3.57	0	―	2	(2.74)
M	0	―	1	3.57	0	―	1	(1.37)
N	0	―	1	3.57	0	―	1	(1.37)
計	28	100.00	28	100	276 (53)	100.00 (100.00)	1,344 (73)	100.00 (100.00)

表 7 のうち、［リアル＋名詞］の異なり数の割合が 15%以上のものに着目すると、書籍では、C2 が 35.71%と最も高く、「リアル」の借用元に当たる英語の‘real’のプロトタイプ的意義でもある A1 が 17.86%と次いで高かった。一方、国会会議録には、15%以上のタイプは存在しなかった。また、D の数値を排除した延べ数の割合が 15%のものに着目すると、書籍では、先と同じく C2 が 35.85%と最も高く、A1 が 33.96%と次いで高かった。一方、国会会議録では、K3 が 20.55%と最も高く、K1 が 19.18%と次いで高かった。いずれも、経済と関わる特殊な

意義である。

以上、本節では専門領域の相違という観点から、国会会議録における［リアル＋名詞］の語彙バリエーションについて書籍と比較してきた。次に、情報の伝達目的の相違という観点から、雑誌における［リアル＋名詞］の語彙バリエーションを書籍と比較する。

５．雑誌における［リアル＋名詞］

次の表 8 は、雑誌に存在した［リアル＋名詞］を構成要素とする普通名詞における「リアル」の意義である。

表 8　［リアル＋名詞］における「リアル」の意義（雑誌の普通名詞）

A1	「偽物の」に対する「本物の」「正真正銘の」の意
A2	「簡易的な」「一般的な」に対する「本格的な」「ハイクラスの」の意
B1	「真実らしさに欠けている」に対する「真に迫っている」の意
B2	「実体・実態からかけ離れている」に対する「実体・実態通りの」の意
C1	「空想の」「虚構の」に対する「現実の」「実際の」「実社会の」の意
C2	「サイバー上の」に対する「現実の」「実際の」「実社会の」の意
D	「現在と一致しない」に対する「現在と一致する」の意
O	「実生活に適さない」に対する「実生活に適した」の意

各意義に対応する［リアル＋名詞］には、表 9 に挙げる要素が存在した。

表 9　意義ごとに見た［リアル＋名詞］の異なり例（雑誌の普通名詞）

A1	［リアル＋ファー］2001、［リアル＋レザー］2001、［リアル＋サーファー］2004、 ［リアル＋ユースカルチャー］2004、［リアル＋カーボン］2005、［リアル＋クロコ］2005、 ［リアル＋ファイト］2005
A2	［リアル＋ヒップホップスタイル］2002、［リアル＋サーフ］2003、 ［リアル＋ブラック］2003、［リアル＋スポーツ］2004、 ［リアル＋アメリカンラグジュアリー］2005、［リアル＋志向］2005
B1	［リアル＋話］2002、［リアル＋イラスト］2003、［リアル＋エンターテイメント］2004 ［リアル＋タイプ］2003、［リアル＋フィギュア］2005
B2	［リアル＋ビュー］2003、［リアル＋サウンド］2004、
C1	［リアル＋体験談］2002、［リアル＋頭身］2003、［リアル＋ライフ］2005
C2	［リアル＋ワールド］2001
D	［リアル＋タイム］2001
O	［リアル＋クロージング］2004、［リアル＋クローズ］2004、［リアル＋スタイル］2005

表 8 に挙げた「リアル」の意義のうち、書籍に存在しなかったものは B2 と O の二つであ

る。このうち、2004 年から現れた O の「リアル」はファッション業界の専門用語に用いられる特殊な意義で、後接可能な名詞は限定されている。

次に、「リアル」の意義ごとに見た［リアル＋名詞］の異なり数の割合と延べ数の割合を調べたところ、次の表 10 のようであった。

表 10　［リアル＋名詞］の異なり数、延べ数とその割合（雑誌の普通名詞）

	異なり数とその割合		延べ数とその割合	
	数	割合(%)	数	割合(%)
A1	7	25.00	11	(26.83)
A2	6	21.43	7	(17.07)
B1	5	17.86	9	(21.95)
B2	2	7.14	2	(4.88)
C1	3	10.71	3	(7.32)
C2	1	3.57	1	(2.44)
D	1	3.57	44	51.76
O	3	10.71	8	(19.51)
計	28	100	85 (41)	100 (100)

表 10 のうち［リアル＋名詞］の異なり数の割合が 15%以上のものに着目すると、A1 が 25.00%と最も高く、次いで A2 が 21.43%と高く、B1 が 17.86%と三番目に高かった。これらの「リアル」は、人々の趣味，趣向，関心事と関わる名詞と結び付き、その名詞の表す事物の品質の良さをアピールしたり、その事物が現実志向であることを表したりするのに用いられている。

また､D の数値を排除した延べ数の割合が 15.00%以上のものに着目すると､A1 が 26.83%と最も高く、次いで B1 が 21.95%と高く、続いて O が 19.51%と 3 番目に高く、最後に A2 が 17.07%と四番目に高かった。従って、D の数値を排除した延べ数の割合が 1 位の「リアル」は、異なり数の割合が 1 位の「リアル」と順位が一致している。

以上、本節では情報の伝達目的の相違という観点から、雑誌における［リアル＋名詞］の語彙バリエーションについて書籍と比較した。次に、同様の観点から、ブログにおける［リアル＋名詞］の語彙バリエーションを書籍と比較する。

６．ブログにおける［リアル＋名詞］

次頁の表 11 は、ブログに存在した［リアル＋名詞］を構成要素とする普通名詞における「リアル」の意義である。

表 11　［リアル＋名詞］における「リアル」の意義（ブログの普通名詞）

A1	「偽物の」に対する「本物の」「正真正銘の」の意
A2	「簡易的な」「一般的な」に対する「本格的な」「ハイクラスの」の意
B1	「真実らしさに欠けている」に対する「真に迫っている」の意
B2	「実体・実態からかけ離れている」に対する「実体・実態通りの」の意
C1	「空想の」「虚構の」に対する「現実の」「実際の」「実社会の」の意
C2	「サイバー上の」に対する「現実の」「実際の」「実社会の」の意
D	「現在と一致しない」に対する「現在と一致する」の意
E	「リアルタイムの」の意

各意義に対応する［リアル＋名詞］には、次の表 12 に挙げる要素が存在した。

表 12　意義ごとに見た［リアル＋名詞］の異なり例（ブログの普通名詞）

A1	［リアル＋ダーリン］2008、［リアル＋年齢］2008、［リアル＋ファー］2008、 ［リアル＋ファイト］2008
A2	［リアル＋ダウンヒルマシン］2008、［リアル＋ブラック］2008
B1	a1．人口的事物が自然の事物（＝真）のように見える ［リアル＋カモ柄］2008、［リアル＋ステッカー］2008、［リアル＋ドール］2008 a2．虚構的事物が実際の事物（＝真）のように見える ［リアル＋アクション］2008、［リアル＋バラエティー］2008、 ［リアル＋フィギュア］2008、［リアル＋路線］2008、［リアル＋ロボット］2008 b1．自然の事物が人口的事物（＝真）のように見える ［リアル＋招き猫］2008 b2．実際の事物が虚構的事物（＝真）のように見える ［リアル＋季封村］2008)、［リアル＋コナン］2008
B2	［リアル＋カラー］2008
C1	［リアル＋出所］2008、［リアル＋すれ違い］2008、［リアル＋ライフ］2008
C2	［リアル＋店舗］2008、［リアル＋友］2008、［リアル＋友達］2008、［リアル＋腐友］2008
D	［リアル＋タイム］2008
E	［リアル＋閲覧者数］2008、［リアル＋視聴］2008、［リアル＋情報］2008、 ［リアル＋ショット］2008

表 11 の「リアル」の意義のうち、書籍に存在しないものは B2 のみであるが、表 12 に B1 の下位類として表示した b1、b2 も同じく書籍に存在しない。b1、b2 は、現在のところ、どちらかというと臨時的であり、新奇性はあるが規範性は低い。

次に、「リアル」の意義ごとに見た［リアル＋名詞］の異なり数の割合と延べ数の割合を、調べたところ、次頁の表 13 のようであった。

表 13　［リアル＋名詞］の異なり数、延べ数とその割合（ブログの普通名詞）

	異なり数とその割合		延べ数とその割合	
	数	割合（%）	数	割合（%）
A1	4	13.33	5	(11.90)
A2	2	6.67	2	(4.76)
B1	11	36.67	13	(30.95)
B2	1	3.33	2	(4.76)
C1	3	10.00	4	(9.52)
C2	4	13.33	6	(14.29)
D	1	3.33	128	75.29
E	4	13.33	10	(23.81)
計	30	100.00	170 (42)	100 (100)

表 13 のうち［リアル＋名詞］の異なり数の割合が 15.00%以上のものに着目すると、B1 が 36.67%と最も高く、これ以外は 15.00%に達していなかった。この B1 の「リアル」は、人々の趣味，趣向，関心事と関わる名詞と結び付き、その名詞の表す事物が現実志向であることを表すのに用いられている。

また、Dの数値を排除した延べ数の割合が 15.00%以上のものに着目すると、同じくB1 が 30.95%と最も高く、次いでEの「リアル」が 23.81%と高かった。Eの「リアル」は、書籍において 2001 年に初めて登場しており、国会会議録においても 2000 年に初めて登場しているが、「リアルタイム」の略語に当たるという点において規範性がやや劣る。このように、ブログでは他のコーパスに比べ規範からずれた使い方が好まれる（あるいは受容される）傾向が見られる[2]。

7．さいごに

以上の考察より、すべてのコーパスの共通点として、「リアル」は［リアル＋タイム］の構成要素として用いられることが最も多いことが明らかとなった。また、それ以外の傾向として、書籍ではサイバー世界と現実世界の対比の中で用いられやすく、国会会議録では政治、経済、科学技術と関わる特殊な意義で用いられやすく、雑誌では人々の趣味、趣向、関心事と関わる事物の品質の良さをアピールするのに用いられやすく、ブログでは人々の趣味、趣向、関心事と関わる事物が現実志向であることを表すのに用いられやすいことが明らかとなった。さらにブログでは、他のコーパスに比べ規範からずれた使い方が好まれる（あるいは受容される）傾向にあることが確認できた。

[2] ブログでは、他のコーパスよりも後項名詞に外来語以外のものが現れる割合が高い。この点からも、ブログが他のコーパスより規範性からずれた使い方を好む（あるいは受容する）傾向にあることがうかがえる。

機能動詞結合における動詞の選択制約
—「影響を与える」と「影響する」—

岡嶋裕子（東京大学大学院博士課程総合文化研究科）

Selectional Restrictions on Verbs in Light-verb Combining
—“Eikyo wo ataeru” and “Eikyo-suru”—

Yuko Okajima (Graduate School of Arts and Sciences, University of Tokyo)

1. はじめに

動作・状態・現象を示す名詞を事態性名詞といい、その事態性名詞と結び付き文法的な機能を果たす動詞が機能動詞である。また、その事態性名詞と機能動詞が結び付いたものを機能動詞結合といい、「散歩をする」「迷惑をかける」「煙が立つ」などの例が挙げられる。この機能動詞結合は慣用性があるため、日本語学習者は、どの事態性名詞とどの機能動詞が結び付くのかを個々に習得しなければならない。

岡嶋（2011）は、中国語を母語とする日本語学習者の作文を分析し、学習者の機能動詞結合使用において誤用を生じる要因の一つに、機能動詞結合そのものが複雑であることを挙げている。岡嶋の調査で、結合が難しいために誤用が生じた機能動詞結合で用いられていた事態性名詞は、「影響」と「注意」の二つだった。「影響」について、辞書では「他に作用が及んで、反応・変化があらわれること（広辞苑）」と一つの意味しか記載していない。しかし、「影響」に接続する機能動詞が「する」「与える」「ある」等と異なった場合、機能動詞結合全体としてその意味・用法も異なってくる。機能動詞結合は語と語の慣用的な結び付きであることから、コロケーションの一種と考えられる。近年、単語単位の語彙学習の限界からコロケーション学習が注目されているが、結び付く語の組み合わせによってどのような使用制約があるのかについての研究はまだなされていない。

本研究では、日本語学習者にとって習得が困難である機能動詞結合の中から、「影響を与える」と「影響する」を取り上げ、両者にどのような意味・使用の違いがあるのかを明らかにする。この具体的な個別事例の解明を通して、習得困難な機能動詞結合が一般的に抱える問題点を明らかにする端緒としたい。

2. 先行研究

藤井・上垣（2008）は、本稿の調査と同じ国立国語研究所の『現代日本語書き言葉均衡コーパス（BCCWJ）』を用いて機能動詞結合の事例を分析したが、「与える」が参与する事態性名詞のほとんどが一桁の事例しかない中で、「影響」のみが 316 例と群を抜いていた。

谷部（2002）は、事態性名詞に漢語を取る機能動詞結合を取り上げ、新聞における使用実態を調査した。谷部は、機能動詞結合のうちヴォイス的意味を担う機能動詞である「－スル／サレル」形（ex. 影響する／される）と「－ヲアタエル／－ヲウケル」形（ex. 影響を与える／受ける）の出現件数を分析し、「影響」の特殊性について報告している。報告では、調査対象とした事態性名詞の中で「影響」1 語のみが、「－ヲアタエル」形の方が「－ヲウケル」形を上回っていた。また、「影響」は「－ヲアタエル／－ヲウケル」形とも、「－ス

ル／サレル」形を上回る唯一の語であり、さらに能動表現は「－スル」形と「－ヲアタエル」形とで大差がないのに、受動表現は「サレル」形と「－ヲウケル」形では約 1 対 6 の開きがあった。谷部は、「影響」は「影響をおよぼす」「影響をこうむる」といった表現もあり、名詞表現[1]が優勢であると述べている。日本語学習者の作文に出てくる機能動詞結合を調査した岡嶋では、「影響」を用いた誤用が 7 例あったが、そのすべてで、機能動詞に「する」が用いられていた。出てきた誤用例は以下である。

a ＊一人が一日にいくつかたばこを吸ってもいい。でも別の人にえいきょうしないはずだ。（影響を与えない）

b ＊ 吸う人が吸わない人に、影響させるので、（影響を与えるので）

（＊は誤用、（ ）内は下線部を正しく言い換えたもの：以下同様）

村木（1991）によると、「する」は実質的な意味が希薄な典型的機能動詞であり、生産性が高く、多くの事態性名詞と結び付く。したがって、日本語学習者にとって最も用いやすい機能動詞だと考えられる。では、なぜ事態性名詞「影響」では、「与える」「受ける」などの他の機能動詞との結合ではなく、「影響する」と「する」を用いた場合だけに誤用が見られたのだろうか。

3. 調査概要

A 調査目的

機能動詞結合「影響を与える」と「影響する」を調査対象とし、それぞれがどのように使い分けられているのかを明らかにする。

B 調査資料

多義語の意味・使用の差別化、分析を行った Fillmore and Atkins (1992)、Atkins. S. et al. (2003) にならい、コーパスを用いる。用いたコーパスは次のものである。

国立国語研究所　コーパス　KOTONOHA 『少納言』[2]

『少納言』では大量のデータの中から、無作為データを抽出することができるが、その中から「影響を与える」「影響する」それぞれの用例 100 を抽出した。

C 分析方法

(1) 「影響を与える」と「影響する」とが入れ替え可能か否かを 3 段階評価。

可　／　どちらとも判断できず　／　不可

3 人の日本語母語話者が判断し、2 人以上が一致した結果を採用。

(2) 各用例で「影響」を引き起こしている“causer”と、その“対象”を、国立国語研究所「分類語彙表」の区分に基づいて次の 6 つに分類

①抽象的関係　ex. 要因、事情、特徴、進展、増加、変化、数量、時間

②人間活動の主体　ex.人間、子孫、日本人、メンバー、世界、国、山村、会社、銀行

③人間活動―精神及び行為

ex. 意思、態度、教育、文化、権利、法律、経済、利益、運動、行為、産業

④人間活動の生産物－結果及び用具　ex. 食べ物、蒸気機関、ソフト

⑤自然―自然物及び自然現象　ex. 味、気候、波、生物、からだ、分娩、成育、疾病

⑥不明: “causer”や“対象”が文の中に表れないもの、またはコーパスに表示された 80

---

[1] 名詞が核になっている表現構造。ex.注意を与える　（cf. 動詞表現：ex.注意する）

[2] 『少納言』のうち、書籍、雑誌、新聞、教科書、Yahoo!知恵袋、Yahoo!ブログを使用し、白書、韻文、広報、法律、国会会議録は除いた。

字前後の文脈の中では特定できないもの。

(3)「影響」の結果を 3 段階に分類

良い　／　どちらとも判断できず　／　悪い

3 人の日本語母語話者が判断し、2 人以上が一致した結果を採用。

判断基準

①言い換え

「影響」に修飾語が付いている場合、言い換えにあたっては、連用修飾は連体修飾に、連体修飾は連用修飾に置き換えた。

ex. 大きな影響を与えた。　⇔　大きく影響した。

言い換えが可能かどうかは、言い換えた後の文が同じ内容を伝えており、かつ自然かどうかで判断した。

②「影響」の結果の良し悪しの判定

文中に良い結果を示す表現があるもの、なくてもそう読み取れるものは良い結果とした。

ex. タイムに影響を与えられるようになってきたことが、うれしくて仕方ないという。

文中に悪い結果を示す表現があるもの、なくてもそう読み取れるものは悪い結果とした。

ex. 自分が病気になると生まれる子に影響するだろう。

4. 分析結果

「影響を与える」の分析結果を表 1 に、「「影響する」の分析結果を表 2 にまとめた。”causer”と対象が「主体」であるかどうかで大きな違いを見せたため、“causer”と、“対象”の下位分類のうち、分析方法で述べた①抽象的関係、③人間活動、④人間活動の生産物、⑤自然の 4 つは「その他」にまとめた。したがって、表 1、表 2 では、“causer”と、“対象”は「人間活動の主体」（以下、「主体」）、「その他」、「不明」の 3 つに分類されている。

表 1　「影響を与える」分析結果

		「影響を与える」	「影響する」との言い換え		
			可	判断つかず	不可
計		100	51	15	34
“causer”	「主体」	10	0	0	10
	その他	70	42	12	16
	不明	20	9	3	8
“対　象”	「主体」	24	2	5	17
	その他	69	44	10	15
	不明	7	5	0	2
「影響」の結果	良い	6	1	1	4
	中立	81	41	12	28
	悪い	13	9	2	2

表 2　「影響する」分析結果

		「影響する」	「～与える」との言い換え		
			可	判断つかず	不可
計		100	93	5	2
“causer”	「主体」	0	0	0	0
	その他	93	87	4	2
	?	7	6	1	0
“対　象”	「主体」	3	3	0	0
	その他	91	87	4	0
	?	6	3	1	2
「影響」の結果	良い	1	1	0	0
	中立	73	69	2	2
	悪い	26	23	3	0

5. 考察

5.1　causer と対象

事態性名詞「影響」に機能動詞「与える」がついた場合と「する」がついた場合で“causer”と“対象”に違いがあるのかを見てみると、“causer”と“対象”が「主体」であるかどうかで大きな違いがあった。「主体」には、人間だけでなく、日本人、国、山村、会社、銀行、世界等が含まれる。

「影響を与える」（以下『与える』）の分析結果を示した表 1 で顕著なのは、“causer”が「主体」のものは 10 あったが、すべて「影響する」（以下『する』）に言い換えができなかったことである。c-1 は“causer”が「主体」である『与える』のデータだが、c-2 のように『する』に言い換えることはできないと判定された。

c-1 ユスティニアヌス帝の妃テオドラは、夫の政策決定に大きな影響を与えた

c-2 ＊ユスティニアヌス帝の妃テオドラは、夫の政策決定に大きく影響した

『与える』の“対象”はどうかというと、“対象”が「主体」であるデータ 24 の内 17（71%）が「する」に言い換えることができないと判定された。d-1 は“対象”が「主体」である『与える』のデータだが、それを『する』に言い換えた d-2 は容認されなかった。

d-1 ギンギンに効かせたサーフ・ギターのスタイルは、全世界のギタリストに直接的な影響を与えているはずだ

d-2 ＊ギンギンに効かせたサーフ・ギターのスタイルは、全世界のギタリストに直接的に影響しているはずだ

つまり『与える』の場合、“causer”と“対象”に「主体」を取るもののほとんどは、「する」と言い換えできないと判定されたことになる。

「影響する」の場合を表 2 で見ると、『与える』の場合とは違って、“causer”に「主体」をとるものは 1 つもなく、“対象”で「主体」をとるものは 3 データだけであった。この 3

データに該当する「主体」は、「胎児」2、「将来の社会」1 であった。「胎児」は「活動の主体」とは言えず、また‘現在の社会’とは異なり‘将来の社会’では「活動の主体」としての意思形成はまだなされていない。e は“causer”が「胎児」だった『する』のデータである。

e 自分が病気になると生まれる子に影響するだろう。

したがって、『する』では、“causer”または“対象”に「主体」を取るものはまったくなかったと言える。

以上から、“causer”または“対象”に「主体」が来る場合、「影響を与える」は用いることができるが、「影響する」を用いることはできないと結論される。

岡嶋の誤用例は次のようであった。

（再掲）a ＊一人が一日にいくつかたばこを吸ってもいい。でも別の人にえいきょうしないはずだ。（影響を与えない）

b ＊ 吸う人が吸わない人に、影響させるので、（影響を与えるので）

これらは、いずれも“対象”に活動の主体である“人”が来ているために許容されないと考えられる。

5.2 良い結果か悪い結果か

f 円高が企業経営に影響を与えた。

g 円高が企業経営に影響した。

g の場合、円高によって経営が悪化したと思われるが、f の場合には、円高が経営に良い影響を及ぼしたとも、悪い影響を及ぼしたとも、どちらとも言えない。このように、「する」を用いた場合、「影響」の結果がよくないことを含意すると直感的に思われるが、実際にそうであるのか、コーパスデータで調査してみた。

表 3 は、「影響を与える」と「影響する」それぞれのデータ文で、影響が良い結果を生じていたか、悪い結果を生じていたかをまとめたものである。

表 3 「影響」の結果の良し悪しクロス表

	良	中立	悪	計
「影響を与える」	6	81	13	100
「影響する」	1	73	26	100
計	7	154	39	200

『与える』の場合、良い結果は 6、悪い結果は 13、『する』の場合、良い結果は 1[3]、悪い結果は 26 だった。χ^2 検定で『与える』と『する』の違いを見てみると、χ^2 (2, N=200) = .016、P< .05 で、有意差が見られた。したがって、『与える』の場合には良い結果、悪い結果、どちらとも言えない中立のもの、どれでもが含まれうるが、『する』の場合、『与える』よりも、良い結果は含まれにくく、悪い結果が含まれやすいといえる。

[3] 「影響する」で 1 つだけよい結果を表すものであると判定された事例

ex.「組織化された集団の 3 が，目標設定と目標達成に向かって，努力するよう影響するプロセス（または行動）である」と定義づけられている。

例 h は日本語学習者の作文から取ったものだが、日本語教師 3 人が『する』は誤用であり、『与える』に修正すべきであるとした。

h ＊インターネットは、新しい生活方式として、われわれの生活に非常に影響した。

インターネットは生活に良い影響を与えるものと一般的にみなされるから、e の文脈では『する』とそぐわないのである。i のようにマイナスの結果を引き起こすと解釈されるものと置き換えると、容認される文となる。

i 昨年度の消費税引き上げは、われわれの生活に非常に影響した。

5.3 相互の言い換え

どちらとも判断がつかなかったものを除き、「与える」と「する」相互に入れ替えができたものとできなかったものだけを表 4 にまとめた。「与える」から「する」に言い換えが可能なものは 51、不可のものは 34 であるのに対し、「する」から「与える」に言い換えが可能なのは 93、不可のものは 2 だけだった。したがって、「影響する」はほとんど「影響を与える」と言い換えが可能だが、「影響を与える」から「影響する」に言い換えられるものには制約があるということになる。

どのようなものが「影響を与える」から「影響する」に言い換えられないかを見ると（表 5)、「与える」から「する」に言い変えられないもの 34 の内、影響が良い結果であるものが 4、影響の“causer”または“対象”に「主体」が含まれるものが 22 あった。

表 4　言い換え可・不可数

	可	不可
与える→する	51	34
する→与える	93	2

表 5　〈与える→する〉言い換え不可要因

良い結果	4
「主体」	22
その他	11
計	37

注)・“causer”と“対象”両方が主体である場合は 1 例と数えた。
・よい結果と「主体」がダブっているものが 3 例

5.4 「影響を与える」と「影響する」の使用制約

以上見てきた結果を表 6 にまとめた。「与える」は“causer”にも“対象”にも「主体」を取ることができるが、「する」は取ることができない。また、影響によって良い結果が引き起こされる場合には、「する」は用いられづらいが、「与える」はどんな結果の場合でも用いることがほぼ可能である。したがって、『する』は『与える』と比べ、多くの使用制約があり、「与える」から「する」に言い換えられない場合が多くあるが、「する」から「与える」への言い換えは、ほとんどの場合問題がない。図 1 に、表 6 の『与える』と『する』の関係を図示した。

谷部は、「影響」は「－ヲアタエル／－ヲウケル」形とも、「－スル／サレル」形を上回る唯一の語であると報告しているが、それはこのように『与える』よりも『する』の方が使用制約が多く、選別的であるためと考えられる。また、岡嶋の「影響」を用いた誤用で全て「する」が用いられていたのも、このように「する」のほうが使用制約が多いので、誤用が生じやすいからである。

表 6　使用制約

	「影響を与える」	「影響する」
①causer・対象に「主体」	可	不可
②「影響」の結果が良いもの	可	不可
使用制約	無	有

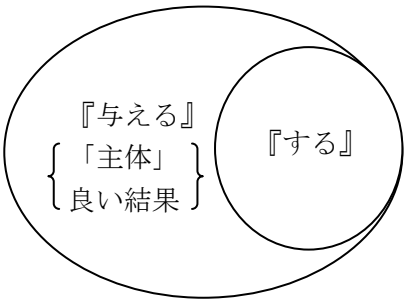


図 1　「与える」と「する」の文中での使用範囲

5.5　語彙概念構造試案

はじめにで述べたように、事態性名詞「影響」の辞書で記載されている意味は一つだけであるが、機能動詞が付加されることによって、機能動詞結合として、新たな意味・使用制約が生じる。日本語学習者が、機能動詞結合を実際に文中で用いようとする際に、適切に使い分けし、また、誤用を犯した場合に、何が問題であるかを理解するためには、結び付く動詞によって、機能動詞結合の意味・用法がどのように異なってくるのかを、体系的に示す必要がある。

自然言語処理の分野では、影山（1996）の分析に基づいて岡山大学が作成した日本語語彙概念辞書がWeb上で公開[4]されている。そこで提示されている「影響する」の語彙概念構造は次のようである。

([動作主]の働きかけで)　<対象>が何らかの影響を受けた状態になる。

[[　]y BE AT [　]z]

y、z は項を表し、BE AT は項 y が項 z の状態である、あるいは場所に存在することを表している。この岡山大学の辞書は、基本的な意味の骨組みによって語彙をパターンごとにまとめることが目的であるので、基本構造だけを取り出している。しかし、本研究の目的は逆に、類似した意味を持つ機能動詞結合の差異化を行うことである。そこで、この枠組みを元にして、日本語学習のために、「影響を与える」と「影響する」の違いを明示した機能動詞結合の（複合的）語彙概念構造（試案）を次のように提起する。

［$_{N}$影響］［$_{P}$を］［$_{V}$与える］
causer x が対象 y に働きかけた結果、対象 y が変化する。
[[　]x ACT ON [　] y] CAUSE [BECOME [　]y [BE AT CHANGED]］

（N: 名詞 noun 、P: 助詞 particle 、V: 動詞 verb　／　点線下線は任意の要素を表す）

これに、x＝円高、y ＝貿易　である例 j を当てはめると次のようになる。

j 円高が貿易に影響を与えた。

[[円高] ACT ON [貿易]] CAUSE [BECOME [貿易] [BE AT CHANGED]]

[4] 岡山大学語彙概念構造辞書　http://cl.it.okayama-u.ac.jp/rsc/lcs

［$_V$影響する］
causer x（ヒトではない）が対象 y（ヒトではない）に働きかけた結果、対象 y が（中立的・悪い状態に）変化する。
[[ ]x impersonal ACT ON [ ]y impersonal] CAUSE [BECOME [ ]y impersonal [BE AT CHANGED not good]]

両者を比較すると、基本構造は同じだが、「影響する」の概念構造では「影響を与える」に 2 つの制約が付加されている(網掛部分)。「影響する」の 2 つの制約とは、ひとつは“causer”と“対象”にヒトを用いることはできないこと、そしてもうひとつは、影響によって生じる結果には、悪い結果、中立的結果のみが来て、良い結果は来ないということである。

なお、概念構造の最初に「影響を与える」「影響する」の品詞分類を示したが、それは、「影響を与える」の「影響」は名詞であるのに対し、「影響する」は全体で 1 つの動詞で品詞が異なるという情報提供が、日本語学習者にとって重要だからである。辞書（大辞林）に「影響」は名詞であると記載されてあることもあり、学習者は k のように、「影響する」の「影響」を名詞ととらえて連体修飾してしまう誤用が多く見られるからである。

k ＊他の人に悪い影響することがある。

今回は「影響を与える」と「影響する」だけを対象としたが、事態性名詞「影響」は他に「影響を受ける」「影響される」「影響がある」「影響を及ぼす」と多様な機能動詞結合があるので、さらにそれらの調査分析も行いたい。また、「影響」に限らず、他の複雑な結び付きを行う機能動詞結合の総合的な分析を行うことを今後の課題とする。

文 献

岡嶋裕子（2011）「漢字圏日本語学習者の機能動詞結合習得」第十回世界日本語教育大会予稿集、pp.464-465.

影山太郎（1996）『動詞意味論―言語と認知の接点―』くろしお出版

国立国語研究所（1964）『分類語彙表』秀英出版

藤井聖子・上垣渉（2008）「支援動詞構文における事態性名詞と動詞との項共有と連結性:『日本語コーパス』を用いた分析」日本言語学会第 136 大会予稿集、pp.432-437.

村木新次郎（1991）『日本語動詞の諸相』ひつじ書房

谷部弘子(2002)「日本語中級段階の漢語運用に関する一考察―漢語動名詞の機能動詞結合を中心に―」『東京学芸大学紀要 2 部門』第 53 号, pp.147-155.

Sue Atkins, Michael Rundell, and Hiroaki Sato (2003) The Contribution of FrameNet to Practical Lexicography, *International Journal of Lexicography*, 16:3, pp. 333-357.

Charles J. Fillmore and Sue Atkins (1992) Towards a frame-based organization of the lexicon: The semantics of RISK and its neighbors, Lehrer, A and Kittay, E. (eds.) *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization.* Hillsdale: Lawrence Erlbaum Associates. pp. 75-102.

関連 URL

岡山大学語彙概念構造辞書　http://cl.it.okayama-u.ac.jp/rsc/lcs

国立国語研究所『現代日本語書き言葉均衡コーパス』KOTONOHA「少納言」http://www.kotonoha.gr.jp/shonagon/

BCCWJ と学習者作文コーパスを利用した日本語作文支援
－表記と共起に関する誤用添削プロトタイプ構築－

八木　豊　　（株式会社ピコラボ）†
ホドシチェク・ボル（東京工業大学）
仁科　喜久子　　（東京工業大学）

Japanese Writing Support System Using the BCCWJ and Learners' Corpus
－ Error Correction Prototype for Misspelling and Misuse of Co-occurrence Representation －

Yutaka YAGI (Picolab Co., Ltd.)
Bor Hodošček (Tokyo Institute of Technology)
Kikuko NISHINA (Tokyo Institute of Technology)

１．はじめに

近年、BCCWJ に代表される大規模な日本語コーパスが利用可能になったことや、形態素解析器や係り受け解析器といった自然言語処理基盤のソフトウェアがより身近なものになったことで、それらを利用した日本語作文支援システムも数多く見られる。jcorrect は、形態素解析および係り受け解析の結果を利用して、技術文章に含まれる誤りの可能性を指摘し、日本語文章の校正を補助する機能を提供している。Chantokun では、Google 日本語 n-gram や日本語ウェブコーパスといった大量のデータから収集した統計情報を利用して、格助詞誤りをチェックする機能を提供している。

仁科らの日本語作文支援システム「なつめ」では、BCCWJ を含む日本語のコーパスから大量の共起情報を収集し、日本語学習者がそれらを効果的に閲覧できる環境を提供することで作文支援における一定の成果をあげている（仁科他(2011) ）。しかしながら、使用しているコーパスは基本的に正しい日本語として収集したもので、誤用に関する情報は含んでいない。我々は、学習者作文コーパスを利用することで、現行の「なつめ」とは異なる観点からの日本語作文支援機能を実現することを目的として、日本語学習者が書いた作文を対象に誤用タグの付与および誤用の分析を進めてきた（曹他(2010)、八木他(2011b) ）。これらの内容に基づいて、学習者の作文に含まれる誤用を検出・特定し訂正例を提示する誤用添削システムを開発中であり、将来的には「なつめ」の共起表現検索および例文表示機能と組み合わせて一つの日本語作文支援システムとすることを検討している（図　１）。

本稿では、正用データとして BCCWJ および「なつめ」プロジェクトで独自に収集したコーパスを利用し、我々がこれまでに構築した学習者作文コーパスを誤用データとして利用した日本語作文支援機能の中から、学習者の作文に含まれる表記の誤り訂正および共起表現の誤り訂正に焦点を当てて報告する。

† yagi@picolab.jp

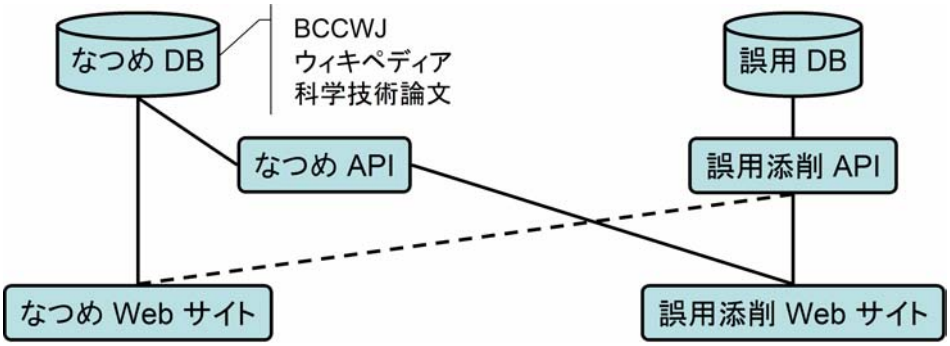


図 1 「なつめ」プロジェクト

２．利用したデータ

表記の誤り訂正および共起表現の誤り訂正において利用したデータについて、正用データと誤用データに分けて説明する。

２．１．正用データ

正用データとしては、BCCWJ に収録されている各種サブコーパスに加えて、ウィキペディアおよび、国立国語研究所の支援により「なつめ」プロジェクトで独自に収集している科学技術論文のデータを利用した。我々は、大学あるいは大学院に在籍する日本語学習者が書くレポートや論文を本システムによる当面の作文支援対象と位置付けており、科学技術論文のデータは、作文支援対象に近い文体のデータを拡充するものである。

「なつめ」では、名詞が格助詞を介して動詞と係り受け関係にあるもの（以下、「名詞＋格助詞＋動詞」と表記する）などをこれらのコーパスから共起表現として抽出しており、本システムでも「なつめ」に搭載されている共起表現データを利用する。表 1 にコーパスごとの共起表現数を示す。

表 1 「なつめ」搭載のコーパス（単位：千）

コーパス名	述べ共起表現数	異なり共起表現数	文字数
BCCWJ：			
書籍	2,955	1,608	53,801
Yahoo!知恵袋	427	260	9,763
国会会議録	422	202	8,712
検定教科書	99	69	1,819
白書	410	172	8,444
Yahoo!ブログ	195	144	5,246
雑誌	21	18	456
新聞	62	51	1,188
ウィキペディア	18,550	7,022	372,901
科学技術論文	340	174	6,108
計	23,482	8,711	468,439

２．２．誤用データ

誤用データとしては、我々がこれまでに構築した学習者作文コーパスに含まれる誤用タグ付き作文データを利用した。学習者作文コーパスは、複数の日本語教師の協力のもとに、大学あるいは大学院に在籍する日本語学習者が日本語の授業で書いた作文を収集したもので、作文の主な内容は、日本語の授業の中で設定した特定のテーマについてのレポートである。これらの作文に対して日本語教師による添削を行い、誤用箇所ごとに「表記の誤り」や「語の共起（コロケーション）の誤り」など独自に定義した誤用種別および誤用の訂正例をタグ付けしたものが「誤用タグ付き作文データ」である。また、学習者作文コーパスは、誤用タグ付き作文データの他に、作文を行った日本語学習者の情報として、国籍、母語、性別、日本語レベル、学習時間なども保持している。

現在も汎用アノテーションツール Slate（徳永他(2010) ）を用いて誤用タグの付与を実施しているが分析の途中であるため、本稿では、試験的な誤用タグの付与を行った以前のデータを使用した。164 人の日本語学習者から 261 作文（総文数 5,600 文）を収集し、そのうちのおよそ 3,500 文に対して誤用タグの付与を行ったものである。

３．誤用添削

誤用添削処理は、まず、学習者作文を CaboCha+UniDic で形態素解析、係り受け解析し、その解析結果および前述の正用データと誤用データに基づいて大まかに以下の流れに沿って行う。

(1)　誤用判定対象箇所の特定
(2)　誤用か否かの判定
(3)　訂正候補の提示

本稿では、「表記の誤り訂正」、「共起表現の誤り訂正」を独立して行った。共起表現の誤りは表記の誤りを含んでいる場合もあるため、本来は双方を関連付けて行うべきところではあるが、現時点でそこまでの連携はとれておらず今後の課題である。以降では、「表記の誤り訂正」、「共起表現の誤り訂正」それぞれについて、上記、誤用添削処理の流れにしたがって説明する。

３．１．表記の誤り訂正

(1) 誤用判定対象箇所の特定

形態素解析の結果、未知語とされた文字列を誤用判定の対象箇所とする。

(2) 誤用か否かの判定

八木他(2011a)では、学習者が表記誤りをした場合の多くは編集距離が 2 以内のところに訂正例の文字列が含まれていることを示した。そのことに基づいて、(1)で特定した対象箇所の文字列から編集距離が 2 以内の文字列リストに展開し、その中から正用データに単語として出現している文字列のみを取り出して訂正候補リストとした。訂正候補リストに 1 件以上の単語が含まれている場合に対象箇所を誤用であると判定する。

(3) 訂正候補の提示

正用データから抽出した単語の頻度情報および誤用データから抽出した編集操作の頻度情報に基づいて訂正候補リストに含まれる単語を順位付けし、上位のものを訂正候補として提示する。

３．２．共起表現の誤り訂正

(1) 誤用判定対象箇所の特定

係り受け解析の結果から特定の係り受けパターンを抽出して対象箇所とする。本稿では、「名詞＋格助詞＋動詞」、「名詞＋格助詞＋形容詞」、「形容詞＋名詞」を対象の係り受けパターンとして対象箇所を抽出した。

(2) 誤用か否かの判定

下記(a)、(b)の二通りの判定を試みる。

(a)　誤用パターンとの一致

(b)　レジスターの妥当性確認

まず(a)では、対象箇所が誤用データに含まれる実際の誤用例あるいは誤用パターンに一致するか否かを確認し、一致した場合に誤用であると判定する。ここでいう誤用パターンとは、実際の誤用例から日本語 WordNet（Bond 他(2009) ）および正用データの頻度情報に基づいて拡張したものである（八木他(2011a) ）。本稿では、誤用データから抽出した「名詞＋格助詞＋動詞」、「名詞＋格助詞＋形容詞」、「形容詞＋名詞」の誤用例 117 個とそこから拡張した誤用パターンを使用した。

次に(b)では、使用される表現や記述内容が、作文支援対象として位置付けているレポートや論文に最も近いと思われる科学技術論文および BCCWJ の白書を準正用データ、口語が含まれておりレポートや論文とは最も遠いと思われる BCCWJ の Yahoo!知恵袋、Yahoo!ブログ、国会会議録を準誤用データとして、対象箇所の共起表現の出現分布を元にカイ二乗検定を行い、準誤用データにおける出現が有意に多い場合にその共起表現はレポートや論文のレジスターとしてふさわしくないものとして誤用であると判定する。ホドシチェク(2011)では、人手でレジスターの誤用であると判定したもののおよそ 8 割がこの手法で自動的に判定可能であることを評価実験により明らかにしている。

(3) 訂正候補の提示

上記(a)の誤用パターンの一致に該当した場合は、誤用データに記載されている訂正例を訂正候補として提示する。上記(b)のレジスターの妥当性確認でレジスターとしてふさわしくないとなった場合は、正用データに含まれるコーパスから準正用データ、準誤用データに分類して利用しており、訂正例を含んでいないため訂正候補の提示をなしとした。

４．適用実験および実験結果の考察

学習者作文コーパスに含まれる作文データのうち、誤用タグを付与していない 36 作文（476 文）に対して誤用添削処理を適用する小規模な実験を行った。

４．１．表記の誤り訂正結果

表記誤りの訂正では、異なりで 17 語（延べ 26 語）を誤用判定対象箇所として特定し、うち 15 語を誤用であると判定した。誤用であると判定したものに対して適切な訂正候補を提示できたのは、「フットサール（→フットサル)」、「アシア（→アジア)」、「ビビムバ（→ビビンバ)」など 6 割ほどであった。しかし、誤用判定対象箇所の特定で取り漏らしている表記誤りも多く存在する。まずはこの点について改善が必要である。特に今回は、形態素解析で未知語とされた文字列を誤用判定の対象箇所としたのみであったが、例えば「うとんを食べる」のような文では、「うとん」の部分が「う（感動詞)」＋「とん（飛ぶの連用形撥音便)」として形態素解析され未知語にはならないため、誤用判定対象箇所から漏れて

表 2 レジスターの妥当性確認の結果

判定結果	名詞＋格助詞＋動詞	名詞＋格助詞＋形容詞	形容詞＋名詞
共起データなし	598	16	11
誤用	41	6	17
判定不可（有意差なし）	468	14	26
正用	3	1	2
計	1,110	37	56

しまう。このような場合に対応するために、正用データに現れにくい品詞の並びになっている箇所や、レポートや論文では使用することの少ない感動詞の周りなど、レジスターに特化した形での特定方法を検討している。

４．２．共起表現の誤り訂正結果

誤用パターンとの一致では、「私＋が＋思う」をレポートらしく「私＋が＋考える」に訂正するなど、マッチしたものは数件しかなかった。

レジスターの妥当性確認では、そのほとんどが「名詞＋格助詞＋動詞」の係り受けパターンではあるが、1,203 件の共起表現が妥当性確認の対象となった。係り受けパターンごとの判定結果を表 2に示す。判定結果の大半は共起データが全くないかあるいは、準正用データと準誤用データとの間で出現分布に有意な差がみられず判定不可となったものであったが、全体の 5%ほどをレジスターとしてふさわしくない誤用であると判定することができた。

正しく誤用であると判定できた共起表現としては以下のようなものが挙げられる。こうした表現は日本語として誤っているわけではないがレポートや論文の場合には別の表現に書き換えたほうがよく、まさにレジスターに関する誤用であるといえる。

- ことがある
- 問題が起きる
- 結論を出す
- 一緒にする
- いい経験

反対に誤って誤用であると判定してしまった共起表現としては以下のようなものが挙げられる。判定の際、科学技術論文および BCCWJ の白書を準正用データとして使用したが、その中にこういった話題に関する文章が少なかったことが、誤用であると判定された要因として考えられる。こうした表現はレポートや論文のテーマによって通常使用するものであるので、準正用データとして適切なコーパスを選択し、それが十分に大きいものであれば出現する可能性があると思われる。

- 子供がいる
- 仕事をする
- 大学に行く

５．まとめ

本稿では、BCCWJ を含む正用データと誤用データを利用して表記の誤り訂正および共起

表現の誤り訂正を行う誤用添削処理を提案し、提案手法を用いた実験結果を報告した。

レジスターの妥当性確認を除く誤用添削処理では正用データ全体をそのまま利用したが、作文支援対象と位置付けているレポートや論文に対してより適切な作文支援をするためには、レジスターの妥当性確認と同様に正用データに含まれるコーパスを内容に応じて使い分けることが望ましい。一方で、誤用データを利用した誤用パターンとの一致ではマッチしたものが数件と少なく、誤用データの質量ともに強化するために引き続き誤用タグの付与を実施していく予定である。

また、こうした誤用添削の結果を学習者に対して効果的に提示するためのユーザインタフェイスを構築する必要がある。

文献

曹紅荃、黒田史彦、八木豊、鈴木泰山、仁科喜久子(2010)「学習者作文支援システムのための誤用データベース作成－動詞の誤用分析を中心に－」世界日語教育大会論文集, pp.1571-1-1571-9.

徳永健伸、Dain Kaplan、飯田龍(2010)「Slate - A multi-purpose annotation tool」情報処理学会自然言語処理研究会報告, 情報処理学会, NL-199, 19.

仁科喜久子、村岡貴子、因京子、Joyce Terence Andrew、鎌田美千子、阿辺川武(2011)「バランス・コーパス利用による日本語作文支援システム『なつめ』の構築と評価」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.215-224.

Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. (2009)「Extending the Japanese WordNet」言語処理学会第 15 回年次大会発表論文集, 言語処理学会.

ホドシチェク・ボル、仁科喜久子(2011)「作文支援システムにおけるレジスターの扱い」世界日本語教育研究大会　異文化コミュニケーションのための日本語教育 2, pp.522-523.

八木豊、鈴木泰山、仁科喜久子(2011a)「BCCWJ と誤用コーパスを利用した日本語作文支援に関する一考察」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.119-124.

八木豊、鈴木泰山、仁科喜久子(2011b)「学習者作文コーパスの構築および BCCWJ と併用した日本語作文支援」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.119-124

関連 URL

日本語作文支援システム「なつめ」：http://hinoki.ryu.titech.ac.jp/

jcorrect を利用した技術文章校正のヒント：http://www.ispl.jp/~oosaki/research/tips-jcorrect/

Chantokun -統計的日本語校正-：http://cl.naist.jp/chantokun/

コーパスに基づく現代語表記のゆれの調査
― BCCWJ コアデータを資料として ―

小椋秀樹（国立国語研究所言語資源研究系）

Corpus-Based Survey of the Orthographic Variation in Contemporary Japanese: Analysis of the BCCWJ-Core

Hideki Ogura（Dept. Corpus Studies, NINJAL）

１．はじめに

音節，語など，種々の言語単位において，形式が一つに定まらず，複数の形式が許容されることがある。この複数の形式が共時的に存在する現象を「ゆれ」と呼ぶ。

語のレベルにおけるゆれには，語形やアクセントのゆれのほか，日本語においては表記のゆれが多く見られる。例えば，「俺－おれ」「さくら－サクラ」のような異なる文字体系間の対立によるゆれのほか，「付属－附属」のような異なる漢字の対立によるゆれ，「行う－行なう」のような送り仮名の違いによるゆれ等がある。また，「上げる－挙げる－揚げる」のような異字同訓も，それぞれを別語とせず同一の語と見なした場合，動詞《アゲル》に「上げる」「挙げる」「揚げる」という複数の表記が共時的に存在すると捉えられ，《アゲル》の表記のゆれとして扱うことができる。

日本語の語表記のゆれについては，これまでに次の三つの調査が行われている。

宮島達夫（1997）　：　1956 年発行雑誌 90 種の調査
国立国語研究所（1983）　：　1966 年発行朝日・毎日・読売 3 紙の調査
国立国語研究所（2006）　：　1994 年発行雑誌 70 誌の調査

しかし，これらの調査については，二つの問題点がある。1 点目は，いずれも調査対象が単一の媒体という点である。現代語表記のゆれの実態解明という面からは，複数の媒体を対象に調査を行い，語表記のゆれに媒体差があるのか明らかにする必要がある。

2 点目は，国立国語研究所（2006）の調査対象年（1994 年）から既に 18 年が経過しているという点である。1990 年代から，情報機器の急速な普及に伴って書記環境が大きく変化するとともに，それに伴う漢字使用の増加が指摘されている。その結果，語表記のゆれの実態にも変化が生じていることが予想される。そこで，より現在に近い時期における語表記のゆれを調査する必要がある。

本研究は，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）のコアデータ(1)を資料として，そこに収録された白書・新聞・雑誌・書籍・Web という五つの媒体を対象に，より現在に近い時期における語表記のゆれの実態を明らかにしようとするものである。

２．調査対象

本研究の目的は，複数の媒体を対象として，より現在に近い時期における語表記のゆれの実態を明らかにすることにある。そこで，複数の媒体のデータを収録した BCCWJ を調

(1)　コアデータの設計・構成等については，小椋秀樹・小木曽智信・小磯花絵ほか（2009）を参照。

査対象とした。ただし BCCWJ 全体を対象とするのではなく，今回はコアデータのみを対象とした。

コアデータは，以下の媒体から成り，延べ語数は，短単位で約 110 万語，長単位で約 84 万語（短単位，長単位とも記号，空白，補助記号を除く。）となっている。

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：2001 年～ 2005 年発行の白書
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

コアデータは，自動形態素解析をした後に，全体に対して人手による確認，修正を行ったデータで，解析精度は長単位・短単位とも約 99%以上である。

BCCWJ を対象とした語表記のゆれの調査では，BCCWJ 全体を対象とすることも考えられるが，コアデータ以外のデータの精度が約 98%と少し低いことから，調査に当たって誤解析がどの程度影響するかが気になるところである。そこで，今回は BCCWJ 全体を対象とした語表記のゆれの調査を行う前の予備調査として，より精度の高いコアデータのみを対象とすることとした。

本研究では，固定長・可変長サンプルの両方を対象とした。また，長短 2 種類のデータのうち短単位を用い，固有名詞・数詞・感動詞・助詞・助動詞・記号・補助記号を除く，いわゆる一般語を対象とした。

３．語表記のゆれの認定方法

本研究では,媒体別・語種別に計量的な観点から語表記のゆれの実態を明らかにするが,その際，どのように表記のゆれを認定し，集計するかが問題となる。以下，本研究における語表記のゆれの認定方法について述べる。

BCCWJ の形態素解析には，形態素解析辞書 UniDic(2)を用いた。UniDic では，表記や語形の違いにかかわらず，同じ語であれば，同一の見出しを与えるという方針を取り，語を階層化した形で登録している。この階層の最上位を語彙素（国語辞典の見出しに相当）と呼んでおり，この語彙素の下に語形（語形の違いを区別する層），更に語形の下に書字形（表記の違いを区別する層）という階層を設けている。

語彙素	語形	書字形
ヤハリ【矢張り】	ヤハリ（副詞）	やはり
		ヤハリ
		矢張り
	ヤッパリ（副詞）	やっぱり
		ヤッパリ
		矢っ張り

図 1：UniDic の階層構造

この階層構造は，BCCWJ の形態論情報にも反映しており，コーパス中の全ての短単位に対して語彙素・語形・書字形という階層的な見出しや品詞情報が付与されている。本研究では，この階層的な見出しを利用して表記のゆれを認定することとした。具体的には，

(2) UniDic の概要については伝康晴・小木曽智信・小椋秀樹ほか（2007）を参照。

「任意の二つの書字形が，同じ語彙素・語形・品詞を持つ場合，同じ語の表記のゆれと認める」とした。つまり，図 1 で言えば，「やはりーヤハリー矢張り」は，同じ語彙素「ヤハリ【矢張り】」，語形「ヤハリ」，品詞「副詞」を持つので，これらは語表記のゆれと認められる。同様に，「やっぱりーヤッパリー矢っ張り」も語表記のゆれと認められる。

語表記のゆれの認定に語形まで含めたのは，同一の語彙素を持つ場合に語表記のゆれと認めるとすると，語形が異なることによる表記の差異も表記のゆれとして扱うことになるからである。例えば，図 1 の書字形欄に掲げた六つの表記は全て《ヤハリ》という語の表記のゆれとなる。しかし「やはりーやっぱり」「ヤハリーヤッパリ」「矢張りー矢っ張り」の対立は，語形の違いによるものであり，本研究では除外する必要がある。

ただし，このようにして認定した語表記のゆれは，語形「ヤハリ」「ヤッパリ」における表記のゆれであり，語の表記のゆれを調査するという目的からは，問題があるという指摘もあろう。しかしながら，本研究では BCCWJ から自動で取得可能な情報を活用することとし，上記のような方法を取った。

このように，語形レベルで集計を行うため，以下に示す語数は，全て語形の数である。表 1 に媒体別の異なり語数・延べ語数を示した。本研究では，Yahoo!知恵袋と Yahoo!ブログとを併せて Web として集計することとした。

表 1：媒体別語数（異なり・延べ）

媒体	異なり	延べ
Web	12,652	93,594
書籍	12,408	105,248
雑誌	14,674	105,610
新聞	15,809	166,630
白書	6,474	120,636

語表記のゆれとして扱う範囲については，冒頭でも述べたように，異なる文字体系間の対立によるゆれ，異なる漢字の対立によるゆれ，送り仮名の違いによるゆれ等のほか，異字同訓も語表記のゆれに含める。また，公用文の表記の基準では，《サラニ》について，副詞は「更に」，接続詞は「さらに」と書き分けることとしている。本研究では，このように基準によって書き分けられているものも，語表記のゆれとして扱う。

４．調査結果

４．１　語表記にゆれの見られる語の割合

まず，媒体別にどの程度の語に表記のゆれが見られるのかを見ていく。語表記にゆれのある語の異なり数，割合を媒体別及び語種別にまとめ，表 2 として示した。

表 2 から，Web・書籍では約 1 割の語に表記のゆれが見られることが分かる。雑誌も 9.0%と Web・書籍に近い割合である。一方，新聞（5.8%）は Web・書籍のおよそ半分程度，白書はそれよりも更に低く，3.3%となっている。このように，語表記のゆれには媒体差が見られ，今回の調査では，語表記のゆれの割合が 1 割程度の Web・書籍・雑誌と 6%以下の新聞・雑誌とに大きく分けられる。

表 2：表記にゆれのある語の割合（媒体別・語種別，異なり）

媒体	異なり	ゆれ	%	語種	異なり	ゆれ	%
Web	12,652	1,299	10.3%	和	4,640	899	19.4%
				漢	5,367	170	3.2%
				外	2,233	198	8.9%
				混	412	32	7.8%
書籍	12,408	1,343	10.8%	和	5,086	1,117	22.0%
				漢	5,852	176	3.0%
				外	1,097	17	1.5%
				混	373	33	8.8%
雑誌	14,674	1,323	9.0%	和	5,189	889	17.1%
				漢	6,622	155	2.3%
				外	2,388	248	10.4%
				混	475	31	6.5%
新聞	15,809	916	5.8%	和	4,944	670	13.6%
				漢	8,330	152	1.8%
				外	2,064	77	3.7%
				混	471	17	3.6%
白書	6,474	211	3.3%	和	1,311	173	13.2%
				漢	4,334	25	0.6%
				外	688	11	1.6%
				混	141	2	1.4%

今回調査した五つの媒体が，このように大きく二分される要因としては，その媒体に共通の表記の基準があるかないかということが考えられる。白書は，全省庁とも常用漢字表，送り仮名の付け方等の国が定めた表記の基準に基づいて書かれている。新聞は，漢字使用に関しては，各社共通の基準として，日本新聞協会が常用漢字表を基に定めた新聞漢字表がある。送り仮名等についても，各社とも国の定めた表記の基準によっている。このように，新聞・白書には共通の表記の基準があるため，語表記のゆれも低く抑えられているものと思われる。

一方，雑誌・書籍については，出版社や雑誌，著者ごとに独自の用字の方針を持っているとしても，新聞漢字表のような各社共通の基準は見られない。特に書籍は，著者個人の自由度が高いと予想される。Web（Yahoo!知恵袋・ブログ）は，どのような表記を用いるかは全く個人の自由である。このように共通の表記の基準を持たないことから，出版社や著者個人による表記の差が多くあり，その結果，語表記のゆれの割合が高くなっているものと思われる。

表 2 には，語種別の語表記のゆれの割合も示した。これを見ると，四つの語種の中で和語が最も語表記のゆれの割合が高い。Web・書籍では 2 割程度にゆれが見られ，媒体別に見た場合にゆれの割合の少なかった新聞・白書でも和語は約 13%にゆれが見られる。一方，漢語は，Web の 3.2%が最も割合が高く，最も割合の低い白書では 0.6%となっている。

外来語については，Web で 8.9%，雑誌で 10.4%と他の媒体に比べて割合が高くなっている点が注意される。これは，外来語が英字で表記されたものも加えていることによるものである。英字表記の外来語を除外して，語表記のゆれの割合を算出したところ，Web は 1.9%（異なり 2,085，ゆれ 40），雑誌は 0.9%（異なり 2,141，ゆれ 20）で，かなり低い割合となる。

４．２　表記の種類数

本節では，媒体別・語種別に，何種類の表記が，それぞれ何語見られるのか見ていくこととする。その結果を，媒体別・語種別に表 3 として示した。

媒体別・語種別のいずれにおいても，表記が 1 種類の語，つまり表記にゆれのない語が最も多く，表記の種類が多くなるほど，語数が少なくなっている。

次に媒体別に見ると，語表記のゆれの割合の高い Web・書籍・雑誌では 3 種類以上の語も比較的多く見られるのに対し，ゆれ割合の低い新聞・白書は，3 種類以上の語は少ない。特に白書は，表記が 3 種類以上の語が 22 語と，全ての媒体の中で最も少ない。

表 3：表記の種類数（媒体別・語種別，異なり）

媒体	語種	表記種類数						
		1	2	3	4	5	6	7
Web	和	3,741	717	139	36	4	1	2
	漢	5,197	141	27	2	0	0	0
	外	2,035	153	40	3	1	1	0
	混	380	26	6	0	0	0	0
書籍	和	3,969	909	172	31	3	1	1
	漢	5,676	158	16	2	0	0	0
	外	1,080	17	0	0	0	0	0
	混	340	31	2	0	0	0	0
雑誌	和	4,300	742	121	23	3	0	0
	漢	6,467	140	13	1	1	0	0
	外	2,140	208	36	4	0	0	0
	混	444	28	3	0	0	0	0
新聞	和	4,274	580	76	11	2	0	1
	漢	8,178	140	12	0	0	0	0
	外	1,987	70	7	0	0	0	0
	混	454	14	3	0	0	0	0
白書	和	1,138	155	18	0	0	0	0
	漢	4,309	21	1	3	0	0	0
	外	677	11	0	0	0	0	0
	混	139	2	0	0	0	0	0

語種別に見ると，和語には表記の種類の多い語が見られる。7 種類の語が Web に 2 語，書籍に 1 語あり，さらにゆれの少ない新聞にも 1 語ある。6 種類の語も Web，書籍にそれぞれ 1 語ずつ見られる。

6 種類の表記を持つ語と 7 種類の表記を持つ語とを，その表記とともに，表 4 として示した。

表 4：表記種類数 6，7 の和語

媒体	見出し	度数	表記（度数）
Web	トル	141	とる（58），取る（50），執る（1），採る（2），撮る（26），獲る（1），録る（1）
	ドウ	366	だう（1），ど～（3），どう（352），どぉー（1），どー（5），ドォー（2），如何（2）
	ワカル	243	わかる（123），ワカル（3），分かる（94），分る（2），判る（15），解る（6）
書籍	カワル	64	かわる（6），代る（1），代わる（3），変る（5），変わる（47），替る（1），替わる（1）
	トル	182	とる（140），取る（32），採る（2），撮る（5），獲る（2），盗る（1）
新聞	トル	150	とる（94），取る（48），執る（1），捕る（1），採る（2），摂る（3），撮る（1）

7 種類の表記が見られる語は，Web では動詞《トル》と副詞《ドウ》，書籍では動詞《カワル》，新聞では動詞《トル》である。6 種類の表記が見られる語は，Web では動詞《ワカル》，書籍では動詞《トル》である。

動詞《トル》《カワル》《ワカル》は，異字同訓の語である。常用漢字表には，《トル》を訓に持つ漢字として「採」「執」「取」「捕」の 4 字が，《カワル》を訓に持つ漢字として「換」「代」「替」「変」の 4 字が掲げられている。《トル》《カワル》ともに意味によっ

てこれら 4 字での書き分けが求められるものであり，元々表記の種類が多くなる可能性のある語と言える。それに加えて，《トル》には「撮る」「獲る」「録る」という表外訓による表記や平仮名表記があり，《カワル》には送り仮名の違いによるゆれと平仮名表記がある。その結果，表記の種類が 7 種類と最も多くなっている。

《ワカル》については，常用漢字表では《ワカル》を訓に持つ漢字として「分」のみを掲げているが，「解る」「判る」という表外訓による表記のほかに，送り仮名のゆれ，平仮名表記，片仮名表記があり，表記の種類が多くなっている。

なお，表記が 5 種類ある和語を見ると，Web では動詞《アラワレル》《カカル》《ススメル》《ツクル》，書籍では動詞《アラワス》《オサエル》《ヒク（他動詞）》であり，これらも異字同訓の語である。

動詞のうち異字同訓の語は，表内字・表外字を含めて書き分けがなされることで元々表記の種類が多く，さらにそこに送り仮名の異なる表記や平仮名表記，片仮名表記が用いられることで，更にゆれが大きくなる傾向があると考えられる。

４．３　動詞の語表記のゆれ

前節で見たように，異字同訓の動詞は，漢字の書き分けに加え，送り仮名の違い等のゆれもあり，表記の種類が多くなっていた。そこで，本節では，動詞に注目し，どのような語表記のゆれの類型があるのかを見ていく。語表記のゆれの類型については，国立国語研究所（1983）で用いられた以下の類型を用いた。なお，この調査では媒体別に集計せず，コアデータ全体でまとめて集計した。

a. 異なる漢字の対立 　例：付属－附属，会う－合う b. 送り仮名の対立 　例：行う－行なう c. 仮名遣いの対立 　例：行う－行ふ d. 外来語表記法の対立 　例：バイオリン－ヴァイオリン	e. 漢字と平仮名の対立 　例：俺－おれ，微妙－びみょう f. 漢字と片仮名の対立 　例：俺－オレ，微妙－ビミョウ g. 平仮名と片仮名の対立 　例：さくら－サクラ h. 文字と記号の対立 　例：国国－国々

コアデータ全体に動詞は異なりで 3,326 語あり，そのうち語表記にゆれのある語は，1,169 語（35.1%）である。動詞の大半は，和語に分類されるが，表 2 に示した和語のゆれの割合を上回っている。動詞は，ゆれの割合の高い語群ということができる。

語表記にゆれのある動詞について，上に示した類型に分類した結果を表 5 として示した。類型 d，h に分類されるものはなかったので，表 5 には，この二つの類型を示していない。なお，動詞の中には，複数の類型に分類されるものがある。例えば，動詞《アタル》の表記として「あたる」「当たる」「当る」の 3 種類がある場合，「当たる－当る」は，「b.送り仮名の対立」に分類され，「あたる－当たる・当る」は「e.漢字と平仮名の対立」に分類される。したがって，動詞《アタル》は類型 b と e との二つに分類されることになる。このような語があるため，各類型の異なり語数の合計が上に示したゆれのある語の異なり語数 1,169 を超えている。

表 5：動詞における語表記のゆれの類型

類型	異なり	%	延べ	%	語　例
a	238	16.5%	33,704	17.1%	暖める－温める，打ち込む－撃ち込む
b	66	4.6%	5,273	2.7%	荒す－荒らす，打ち上げる－打上げる
c	9	0.6%	20,670	10.5%	うなずける－うなづける，考える－考へる
e	1,050	72.7%	88,775	45.1%	あう－合う，あえる－和える
f	33	2.3%	10,140	5.2%	オソレイル－恐れ入る，ハネ上がる－跳ね上がる
g	48	3.3%	38,064	19.4%	するースル，いじめるーイジメる

表 5 を見ると，類型 e に属する語が最も多く，異なりで 72.7%，延べで 45.1%を占める。類型 a がそれに次ぎ，異なりで 16.5%，延べで 17.1%となっている。

次に，単純動詞・複合動詞に分けて集計した結果を表 6，表 7 として示した。

表 6：語表記のゆれの類型（単純動詞）

類型	異なり	%	延べ	%
a	185	18.6%	33,151	17.3%
b	38	3.8%	4,657	2.4%
c	9	0.9%	20,670	10.8%
e	688	69.3%	84,797	44.4%
f	29	2.9%	9,988	5.2%
g	44	4.4%	37,911	19.8%

表 7：語表記のゆれの類型（複合動詞）

類型	異なり	%	延べ	%
a	53	11.8%	553	10.1%
b	28	6.2%	616	11.3%
c	0	0.0%	0	0.0%
e	362	80.3%	3,978	73.0%
f	4	0.9%	152	2.8%
g	4	0.9%	153	2.8%

単純動詞，複合動詞共に，1 位は類型 e，2 位は類型 a となっている。複合動詞では，類型 e の割合が異なり・延べとも動詞全体（表 5），単純動詞（表 6）より高くなっている。

類型 e に属する語を見てみると，単純動詞では度数順で上位から《イル》《アル》《イウ》《ナル》《クル》が挙げられる。これらは，基本動詞であり，また複合辞の構成要素にもなっている語である(3)。複合辞の構成要素については，実質的な意味が薄れていることから，平仮名表記される傾向が見られる。《イル》等が類型 e に属することには，複合辞での使用例も関わっていると思われる。

このほか，動詞《トル》《キク》《ツクル》といった異字同訓の語や，《ツナガル》《マトメル》といった漢字表記した場合に表外漢字となる語がある。

複合動詞では，全体を平仮名表記にしたものや，前項又は後項を平仮名表記したものが見られる。例えば，「とりくむー取り組む・取組む」「くりかえすーくり返すー繰り返す・繰返す」などである。

類型 a に属する語は，基本的に異字同訓の語である。「聞くー聴くーきく」「換わるー代わるー替わるー変わるーかわる」といった常用漢字表内で書き分けるもののほか，「見るー診るー観るーみる」「取るー執るー採るー撮るー獲るー録るーとる」「言うー云うーいう」のように，表外訓・表外字（下線部）が見られるものもある。また，異字同訓については，前節で見たように，多くの場合，平仮名表記も共に用いられている。類型 a に分類されている 238 語のうち，171 語が類型 e にも分類されている。

異字同訓の語で平仮名表記も用いられる要因の一つとして，書き分けの難しさから，平仮名表記が選択されるということが考えられる。例えば，次に挙げる動詞《トル》の仮名

(3) これらの動詞を構成要素に持つ複合辞として，例えば「ている」「てある」「という」「ことになる」「てくる」が挙げられる。

表記例は，いずれも「取」で表記して問題ない例ではあるが，動詞《トル》の中心的な意味用法ではないため，どの漢字で表記するか判断に迷う面があり，平仮名表記が選択された可能性がある。

コミュニケーションをとる手段を身につけさせる（教育再生！）
首相がこうした言動をとることで（読売新聞）
防災に関しとるべき措置と地域防災計画の作成（消防白書）

なお，これらについては，中心的な意味用法ではないため，漢字表記すること自体に違和感があり，平仮名が選択されたという可能性もあろう。異字同訓の語については，どのような意味用法の時に，漢字表記が選択されるのか，また平仮名表記が選択されるのかといったことを調べていく必要がある。

５．終わりに

本研究では，BCCWJ のコアデータを対象に語表記のゆれに関する調査を行った。その結果，以下のことが明らかとなった。

(1) 語表記のゆれには媒体による差異がある。コアデータに収録した五つの媒体については，語表記のゆれの割合の高い Web・書籍・雑誌とゆれの割合の低い白書・新聞とに大きく分けられる。
(2) 語種別に見た場合，和語が最も語表記のゆれの割合が高い。
(3) 動詞は，異なりで約 35%の語に表記のゆれが見られる。類型別に見た場合，平仮名と漢字の対立によるゆれ，異なる漢字の対立によるゆれが多い。
(4) 異字同訓の動詞については，漢字の書き分けにより表記の種類が多く，さらに送り仮名の違いや平仮名表記により，更にゆれが大きくなる傾向が見られる。

今回の調査を基に，他の品詞等についても調査を進めていく必要がある。また，2 節に述べたように，今回の調査は予備調査である。BCCWJ 全体を対象として，語表記のゆれの調査を行い，現代における語表記のゆれの実態をより一層明らかにしていくことが必要である。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）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.

国立国語研究所（1983）国立国語研究所報告 75『現代表記のゆれ』.

国立国語研究所（2006）国立国語研究所報告 125『現代雑誌の表記― 1994 年発行 70 誌―』.

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, pp.101-123, 国書刊行会.

宮島達夫（1997）「雑誌九十種表記表の統計」『日本語科学』1, pp.92-104, 国書刊行会.

「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを考える際に必要となる視点は何か？ —「同意要求表現」を中心に—

岡田　祥平（九州共立大学共通教育センター）†
江崎　哲也（山梨大学留学生センター）‡

What Should be Considered When Developing Corpus of Spoken Japanese Annotated of Tones and the Speaker's Intention: Focus on "Speaker's Expectation of the Hearer's Agreement" Expressions

Shohei Okada (Kyushu Kyoritsu University)
Tetsuya Esaki (University of Yamanashi)

１．はじめに

近年，音声言語の言語学的研究の分野において，イントネーション研究が精力的になされるようになったのは，前川（2006）も指摘するとおりである。しかし，その一方で，イントネーション・音調[1]の研究の手法には，まだまだ発展の余地があると思われる。そこで，本発表では，イントネーション・音調研究のさらなる発展を模索するものとして，話し言葉の音声コーパスにおいて，音調，中でも「文末音調と発話意図とを統合したアノテーション」を施すことの可能性，必要性と，その際に留意すべきであろう視点について述べる。

本稿では，まず第 2 節で，文末音調と発話意図との関わりとを模索する研究の問題点について述べた上で，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備の必要性について述べる。続く第 3 節では，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備の端緒として，東京方言，および愛媛県宇和島方言の「同意要求表現」に着目すればいいのではないかという視点を指摘する。ただ，その際には，「そもそも『同意要求表現』とは何か」という点を明確にする必要性を，第 4 節で述べる。第 5 節では，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを整備した後の研究の展開の可能性として，「研究者の内省・観察に基づ」く（田中 2011）と「同意要求表現」に使用されるという，東京方言における「とびはね音調」（田中 1993・2010・2011 などを参照。本稿 3.1 節でも説明する）が，実際にそのように言えるのか，定量的に調べられるのではないか，という点を述べることにする。最後に第 6 節では，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを整備した後の研究の展開の可能性として，第 5 節で論じた点以外のことを，簡単にまとめることにする。

２．「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備の必要性

御園生（2000）は，「文末諸形式の表す話し手の発話表現意図に関連付けてしばしば取り

† okadash@kyuko-u.ac.jp
‡ esakit@yamanashi.ac.jp

[1] 斎藤（2011）に簡潔にまとめられているように，「音調」という用語は「さまざまな意味で用いられてきた」（斎藤 2011）。そこで，本稿における定義，立場を明らかにしなければならないだろう。本稿では，「音調」という用語を，「単語レベルのアクセントと文レベルにおけるイントネーション」（上野 2011）を包括した，「話し言葉における相対的な声の高さの様相」（斎藤 2011）のことを指すこととする。

上げられる」「文末イントネーションの研究法」を，以下の二つに分類している。

①モデル分析　「取り上げる形式の用法とイントネーションの音調が持つ機能を関数として扱おうとするもの」。この場合，「用法と文末のピッチパタンの関係は『確認要求の場合は上昇，情報提供の場合は下降』というように少数のパタンに還元され対応づけられて説明される場合が多い」。

②実例分析　「自然談話から文末形式の使用例を収集して，表現意図と文末イントネーションとの関係を分析していこうとするもの」。

御園生（2000）がいう「モデル分析」の研究法としては，具体的には，話者（や研究者）の内省を利用するものが主流であろうが，他に「シミュレーション法」（郡 2006）とも称される手法が考えられよう。郡（2006）の言を借りれば，「シミュレーション法」とは，以下のような手法のことである。

> たとえば，「目上の人に言うつもりで」のような指示を与える，あるいはそのような場面を設定した対話を演じてもらうなどの方法で発音を求め，その音声的特徴を比較するという手法である。

しかし，現実世界で行われている対話の音声の文末音調を観察していると，内省や「シミュレーション法」では，到底捉え切れない数々の変種が存在していることに気付く。また，郡（2006）も指摘するように，「シミュレーション法でわかるのは，ステレオタイプとしての表現法」であり，「現実の表現法」ではない[2]。

このように考えると，文末音調と発話意図との関係を観察するには，御園生（2000）がいう「実例分析」の手法が必須になる。実は，郡（2006）は，「まだまだ未開拓と言える対人音声」の研究の第一の手法としては，「自発音声の分析」を挙げているのだが，「実例分析」（御園生 2000），「自発音声の分析」（郡 2006）には，大きな問題点がある。それは，「実例のイントネーションは多様な現れ方を見せ，用法と単純には対応づけられない結果となっている」（御園生 2000），「多くの話者にあてはまるような一般的なことをこの手法だけで言おうとすると，膨大な録音資料が必要になる」（郡 2006）といった点である。

しかし，そのような問題点は，自発発話の音声コーパスを整備することによって，解決の糸口が見出せるように思える。実際，『日本語話し言葉コーパス』（前川 2004 など）の登場によって，少なくとも，独話の自発音声のイントネーション分析は，飛躍的な発展を遂げた（あるいは，遂げている途上である）。ただ，独話中心の『日本語話し言葉コーパス』では，我々の日常生活における音声言語の主流である対話の「現実の表現法」（郡 2006）を捉えるには，不十分な側面もある[3]。

以上のような議論を踏まえるならば，文末音調と発話意図との関係を観察するためには，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備が必要になってくると考えるのである。

３．「同意要求表現」とその音調

前節では，文末音調と発話意図との関係を観察するために，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備の必要性を論じた。しかし，ただ漠然と自然談話を収集し，そこに現れる文末音調と発話意図との関係性を見出そうとしても，膨大かつ多様なデータを前にしては，雲を掴むような話である。そこで，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを整備し，それから文末音調と発話意図との関係を観察する端緒として注目したいのが，「同意要求表現」である。

---

[2] 川上（2000）の，「アクセントやイントネーション研究の実験台にされた人が真に自然なイントネーションで発音することは，老練の役者などは別にして，まずまず有り得ないことである。大方は要らざる所で声を高めてしまう。」という「警鐘」にも，耳を傾けたい。

[3] 前川（2006）も，「自然な対話音声を大量に観察すれば，比較的容易に新しいカテゴリや変種を発見できるのではないかと思う」と述べている。

というのも，「同意要求表現」には，その表現特有の文末音調が生起するという報告が見られるため，文末音調と発話意図との関係を観察する端緒としては好適だと思われるからである。

３．１　東京方言の場合

東京方言には，「とびはね音調」と呼ばれる音調が存在する。この音調は田中ゆかり氏が1993 年春の日本方言研究会で報告（田中 1993）したことによって広く知られるようになったと思われるが，具体的には，以下のような音調である（田中 2011）。

(1)「～ナイ？」形式を取る問いかけ音調の昇調の一種である

(2)「～ナイ？」部分にだけに焦点を絞れば，「浮き上がり調（川上蓁,1963）」の一種に聞こえる。「～ナ］イ」（引用者註: ］はアクセントの下がり目を表す）の下がり目が無効化される。

(3)「～ナイ？」の前接形式に起伏型アクセントの単語が入る場合，その単語のアクセントの下がり目が無効化される。前接形式には，形容詞（ク形），動詞（＋タク），形容動詞（＋ジャ），名詞（＋ジャ）などが入る。

(4) 「～ナイ？」の前接形式に平板形アクセントの単語が入った場合，その単語のアクセント型は保持される。

(5)(1)～(3)を踏まえると，「カワイク（形容詞Ⅱ類）＋ナイ？」の場合，「カワイ］ク」の下がり目ならびに「ナ］イ？」の下がり目が無効化され，「カワイクナイ↑↑」と上昇を続ける音調として実現される。田中ゆかり（1993）では，「最後の拍の上昇を備えるように語アクセントを破壊してまで，早い段階からピッチが上がったままになる」と表現している。

通常は「語アクセントにイントネーションが加わってもアクセント型に変化は生じない」（都染 2007）とされる（森山 1989 なども参照）。しかし，田中（2011）の(3)にあるように，「とびはね音調」は「クナイ」「ジャナイ」に前接する語が有核語の場合，その単語のアクセント核が「無効化」される。その点で，「とびはね音調」は，東京方言において，特徴的な振る舞いを見せる音調といえる。

そして，この「とびはね音調」は，「『同意求め』として出現することが多い」（田中 2010）とされている（蔡 1996，湧田 2003 も参照）。

なお，「とびはね音調」は，典型的には形容詞否定形「クナイ」が同意要求表現として使用される場合に生起するのだが，近年では，井上史雄氏によって，「名詞＋ジャネ」という形式の場合において，「ジャネ」に前接する名詞のアクセントが破壊，無効化されるという現象が報告されている（井上 2008a・b）。上に引用した田中（2011）の(3)を見てもわかるように，田中ゆかり氏は，井上（2008a・b）が言及する現象も「とびはね音調」に分類している（田中 2010 も参照）。

３．２　宇和島方言の場合

工藤・八亀（2008）では，宇和島方言は，質問文と平叙文が語形によって明確に区別されるため，質問文末の「イントネーションは下降調が普通になる」という記述がある。ただし，工藤・八亀（2008）によると，上述のような宇和島方言の疑問文末イントネーションの特徴は「基本的な場合であり，絶対に下降調イントネーションしか使わない，と言っているのではない」とし，次のようにまとめている（傍点部は筆者）。

①中立的な質問: 下降調が普通

②話し手の想定や断定に対する質問（同意を求める場合）：上昇後下降するイントネーションが義務的

ここまで見てきたように，中立的・基本的な質問文（つまり，聞き手に何か情報を求めるような質問文）では，下降調のイントネーションが義務的である。しかし，次のような場合は，一度上がった後で下がるようなイントネーションが義務的である。最初の二つは否定形が，後の二つは断定形が使われていて，話し手の想定や断定について，聞き手に同意を求める特別の質問のタイプである。

(28)外寒ーあらへん↬（外は寒いんじゃない？）
(29)確か昨日は暇やあらへなんだ↬（確か昨日は暇だったんじゃない？）
(30)外寒いわいなー↬（外は寒いよね？）
(31)確か昨日暇なかったいなー↬（確か昨日は暇だったよね？）

４．発話意図の分類・ラベルの問題

ここで問題となるのが，発話意図のラベリングである。前節では，東京方言や宇和島方言の「同意要求表現」に特徴的な音調が観察されるという先行研究を紹介したが，そもそも「同意要求表現」とは何か，という問題がある。そこで参照しなければならないのが現代日本語文法の研究の成果であろうが，実は「同意要求」を明確に定義した研究はそれほど多くないようである。管見の限りでは，「同意要求」を明確に定義した研究は，1990 年代になって，ようやく現れる（鄭 1992，ザトラウスキー1993，三宅 1996，神部 1997）[4]。「同意要求」と類似した用法である「確認要求」が，1960 年代から定義されていた（国立国語研究所 1960）のとは，対照的である[5]。

しかも，ここで留意したいのは，「同意要求」と「確認要求」とは，必ずしも截然とは区別できるとは限らないようだ，という点である。先行研究を概観すると，「同意要求」と「確認要求」の差異を明確には論じないもの（小針 1998，石井 2000 など）や，「同意要求」を「確認要求」に含めるもの（三宅 1996），さらには「同意要求」と「確認要求」とが区別できない場合もあることを指摘するもの（谷川 2003）も存在する[6・7]。したがって，ある発話に対し，「同意要求」，あるいは「確認要求」という発話意図のラベリングを付与することは，意外と困難な作業になるかも知れない，ということが予想される。実際，井上・山口

---

[4] 厳密には，江端（1977）に，既に「同意要求」という表現が見られるのであるが，江端（1977）は「同意要求」に明確な定義を与えていない。しかも，以下のように記述され，「同意要求」と「確認要求」が同一カテゴリとして扱われている。

「ダラー」の表現機能は,単純な想像・推測・推量よりも,確実な推察・推定・確認・同意要求の方が多い。

[5] 現代日本語文法の研究において，「同意要求」が近年になりようやく注目されるようになったという事実は，以下のデータベースの検索結果数からも推察できる。

●雑誌「国語学」全文データベース（http://db2.ninjal.ac.jp/SJL/）●CiNii（http://ci.nii.ac.jp/）●国立国会図書館雑誌記事検索（http://opac.ndl.go.jp/Process）●日本語研究・日本語教育文献データベース（https://dbms.ninjal.ac.jp/bunken/data/）●国文学論文目録データベース（http://base1.nijl.ac.jp/~ronbun/）

付表　「同意要求」という語を含む論文数と「確認要求」という語を含む論文数

	同意要求	確認要求
雑誌「国語学」全文データベース	7	22
CiNii	8	50
国立国会図書館雑誌記事検索	4	38
日本語研究・日本語教育文献データベース	6	52
国文学論文目録データベース	2	20

[6] 平塚・原田（2012）では，鹿児島県北薩方言の「セン」が，「全年齢層に共通してみられる基本的な用法は同意要求であるが，若年層では確認要求としても用いられるようになるという用法面の変化」が起きていることを指摘しているが，その理由としては，「同意要求の用法が拡張したもの」と説明している。

[7] 現代日本語の音声を中心に研究を進めてきた筆者にとっては，現代日本語文法の論文の中で，「確認要求」の例として提示された用例・文例が「同意要求」のように感じたり，逆に「同意要求」の例として提示された用例・文例が「確認要求」のように感じたりしたことがあったことを，正直に告白しておく。

（2002）では，「確認要求・同意要求の表現」と，両者を併記して立項されている[8]。

そのような事情を踏まえると，「同意要求」と「確認要求」との定義・差異を明確に（あるいは対照的に）示した現代日本語文法（方言文法も含む）の研究を見ておく必要があると思われる。そのような研究としては，鄭（1992），三宅（1996），高木（2011），平塚・原田（2012）がある。それぞれの定義を以下に示す。

表　先行研究による「同意要求」と「確認要求」の定義

	「同意要求」	「確認要求」
鄭（1992）	情報的に同等の関係にあると思われる聞き手に同意を求めるニュアンス	話し手より情報的に優位にあると思われる聞き手に確かめるニュアンス
三宅（1996）[9]	同意や同感を求めるといったことが表されるもの	話し手にとって何か不確実なことを，聞き手によって確実にしてもらう
高木（2011）	同一の判断・認識を聞き手が持っていることの表明を要求	話し手が判断したこと（内容）があっているかどうかの判断を要求
平塚・原田（2012）	話し手と同じ判断を形成していることを聞き手に問いかける	発話命題が真であることの確認や，命題が表す情報を聞き手が持っていることの確認を，聞き手に要求する

以上の表を踏まえるならば，現代日本語日本語文法研究の世界では，「同意要求」と「確認要求」とは，「同意要求→聞き手に判断を必ずしも求めない」，「確認要求→聞き手に何らかの判断を求める」という点で違いがある，とまとめられるのではないだろうか。もし，「とびはね音調」を手掛かりとして「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを整備するとするならば，この点を留意しなければならないだろう[10]。

５．「とびはね音調」は「同意要求」の際に現れる音調と言えるのか

５．１　蔡（1996），湧田（2003）の再検討

３．１節で述べたように，「とびはね音調」は「同意要求」の際に現れる音調だとされる。そのことを「実証的に」論じた研究としては，蔡（1996）や湧田（2003）がある。しかし前節で指摘したように，「同意要求」という発話意図のラベルを付与する際には，「確認要求」との差異にも留意する必要が出てくると思われる。

そこで，「とびはね音調」は「同意要求」表現だとした蔡（1996）湧田（2003）は，どのような手法でそのような結論を導き出したのかを確認しよう[11]。

---

[8] 井上・山口（2002）では，終助詞ネの用法のうち，「情報の妥当性を話し手側で確認中であることを聞き手に述べる」「内部確認用法」で，「その場で誰でも考えそうなことを述べる場合には，同意要求，共感表明の意味が加わる」と説明されている。これと全く同様の説明が，井上（2002）にも見られる。

[9] 三宅（1996）は，「同意要求」を「確認要求」の下位分類とみなす立場に立っていることは，本文中で述べたとおりである。

[10] さらにいうなれば，「新情報認知要求」（簡 2011）の用法（「聞き手が知らないあるいは分からないはずの情報を提示して当該の情報の認知を要求するもの」簡 2011）との関連も考慮しなければならないと考えている。ただし，「新情報認知要求」は，先行研究において，「用法がごく少数で，〈推量〉や〈確認要求〉と並ぶ信用法として扱われていない」（簡 2011）が現状であり，「新情報認知要求」について論じるのは筆者の能力を超える作業である上，紙幅の都合もあるため，ここではこれ以上の言及は避けることにする。

[11] 「とびはね音調」の存在を知らしめた田中ゆかり氏は，実は少なくとも発話調査では，「とびはね音調」が「同意要求」であることを裏付けるような調査（「同意要求」以外の場面も設定し，「同意要求」場面での音調とそれ以外の場面の音調とを対比する調査）はなさっていない（田中氏からの私信によると，「『と

蔡（1996），湧田（2003）は，いずれも郡（2006）でいう「シミュレーション法」を用いている。

蔡（1996）では，以下のような文脈を提示し，下線部の音調を分析している。

B: へー，面白いものって，例えば？

A: （カバンからあるものを出して）ほら，見て，例えばこれ，<u>面白くない？</u>

B: 本当だね。面白いというか，可愛いね。その店で買ったの？

しかし，下線部は「同意要求」とも，「確認要求」とも解釈可能ではないか。「同意要求」という解釈の場合であれば【A は B に「面白い」と同意してもらいたかった】ということになるし，「確認要求」という解釈の場合であれば，【A は B に「面白い」かどうかを判断してもらいたかった】ということになる。蔡（1996）で提示された文脈での，B の応答は，「本当だね」と同意しつつ，「可愛い」という新たな判断を示している。はたして，この用例は，「同意要求」なのであろうか，それとも「確認要求」なのであろうか。調査対象者がどちらの解釈に立って発音したのか，にわかに判断がつきにくい。

次に，湧田（2003）の場合である。湧田（2003）では，以下の図 1 に示したようなイラストを提示して，調査対象者に発話することを求めている。しかし，図 6.1 の A の発言は「同意要求」（＝同意求め）と断言できるのか，筆者にはにわかに自信がない。というのも，図 6.1 の A は B に「おもしろかった」と同意してほしい（＝「同意要求」）のか，それとも「おもしろかった」と判断してもらいたい（＝「確認要求」）のかが，図だけでは判断できないからである（図 6.1 は B から話を振っているので，B は当該映画に関する知識を持っている＝「確認要求」の対象としての条件を満たしている）。

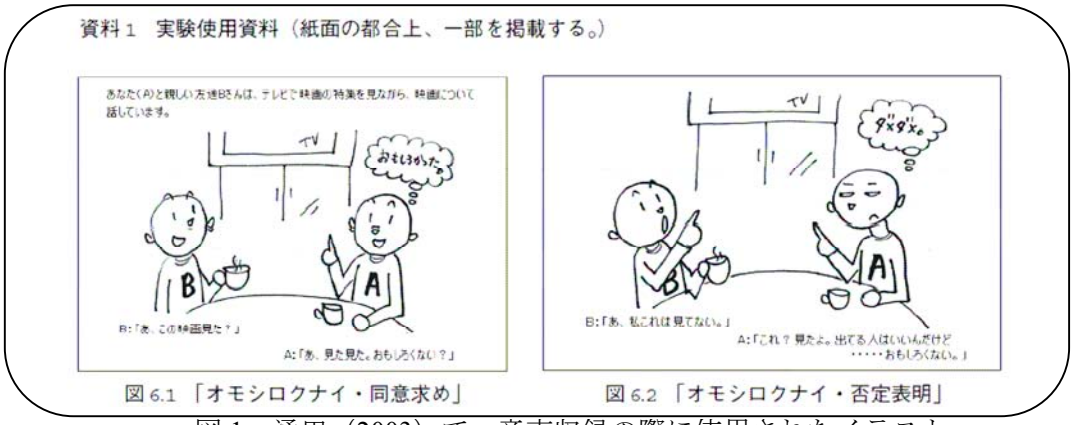


図 1　湧田（2003）で，音声収録の際に使用されたイラスト

以上のように考えると，「とびはね音調」は「同意要求」の音調であると結論づけた先行研究には，再考の余地があるように思われる。そもそも，「同意要求」（と思われる）場面しか提示していないにもかかわらず，「とびはね音調」が「同意要求」の音調であると結論づけるのは，拙速なのではないだろうか（他の用法でも「とびはね音調」が使用されている可能性もあるが，「同意要求」場面しか提示してないのであれば，「同意要求」以外の場面でも「とびはね音調」が使用されるということは実証できないであろう）。そのような議論を踏まえると，「シミュレーション法」で，文末音調と発話意図との関係を模索するためには，複数の解釈が生まれないような場面設定を提示する必要があるが，先行研究には，その点に再考の余地はあろう。そもそも，調査対象者が「うまく演じてくれたかどうか」

びはね』が最大出現するとしたら，という調査意図なので，この調査では、『友達に同意を求めるように』という指示以外では発音を求めていません」とのことである）。このことは，田中氏が「とびはね音調」を「直感的」に「同意要求表現」と結び付けられてお考えだったと思われる（5.2 節で言及した田中 2011 の引用も参照）。

という点にも，議論の余地がある。そのような，再考，議論の余地を埋めるには，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備が有用であるとも考えられるのである。

５．２　田中（2011）の指摘

実は,「とびはね音調」が「同意要求」以外の場合にも使用されるということは, 田中 (2011) で既に指摘されている。

田中（2011）は,「研究者の内省・観察に基づいた，昇調は単純な問いかけの『質問（意見求め)』,『とびはね音調』は問いかけ音調形式を用いた『同意要求』，という機能の振り分け観はなんとなく共有されている」とした上で,「『～ナイ？』形式にかかわる首都圏に共存する複数音調の機能について，小調査データから検討」した結果，以下のような結果と，将来の展望を述べている。

> 首都圏大学生を対象とした小規模な聞き取りアンケート調査の結果から,「～ナイ？」形式をとる問いかけ音調の機能について検討してきた。その結果，たしかに「とびはね音調」は「同意要求」としての機能が優勢ではあるものの,「質問」という機能を果たすという認識も少なくないことがわかった。また,「とびはね音調」に限らず, 現代の首都圏に共存する「～ナイ？」形式にあらわれる複数の問いかけ音調は，いずれにおいても「質問」「同意要求」といった機能と一対一の対応はしていなかった。これらの結果は，音調と機能の対応について，量的観点を導入してみていくと，きれいな一対一対応をしていない，という側面が前景化してくることを示していると考えられる。
>
> （中略）
>
> 音調と機能の対応をどのように記述しているかという問題は，研究を進めていく上でかなり重要な問題となるものといえそうである。一対一対応をかならずしもとらないことが想像される音調と機能の対応をいかに記述していくか，また，調査の場においてどのような指示で発話してもらうか，というような調査方法の問題など，多くの課題がみえてくる。

この，田中（2011）の指摘は示唆的である。その一方で，田中（2011）の分析結果は，あくまで「聞き取りアンケート調査の結果」であり，実際の発話データをもとに分析した結果ではない。聞き取りアンケート著差結果の傾向と，発話の傾向とが異なる可能性もある。「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを整備できれば，文末音調と発話意図との関係を観察する際に,「実例分析」に基づく「モデル分析」（御園生 2000）が可能となるであろうし，「音調と機能の対応」（田中 2011）を定量的に示すことも可能となるではないか。さらには，現代日本語文法研究の成果に対し，音声研究が何らかの貢献をすることが可能なのではないかと考えられる[12]。

６．おわりに

本稿での主張をまとめるならば，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備すると,「コミュニケーション研究として，あるいは学習・教育のために我々が知りたい」対象である「現実の表現法」（郡 2006）に肉薄する手段の一歩となるに違いない，という点にある。

また，このような試みは,「文法研究者と音調研究者のコラボレーション」（田中 2011）にもつながってくる[13]。

---

[12] たとえば，音調を手掛かりにすることによって,「同意要求」と「確認要求」との区別を明確化できるという知見を，現代日本語文法研究の世界に提供できる，といった可能性が考えられよう。

[13] 田中（2011）では,「『とびはね音調』に優勢にあらわれる「同意要求」機能についてどのように考えていくべきか」という問いに対し,「文法研究者と音調研究者のコラボレーション」という視点を提示されて

従来の現代日本語を対象とした文末音調の研究は，音調と発話意図との関連を模索するものが多かった[14]。以下の図 2[15]でいう①の関係である。一方，従来の文法研究は，言語形式と発話意図との関連を模索するものが多かった[16]。以下の図 2 の②の関係である。しかし，「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備するということは，音調，発話意図，言語形式の 3 つの要素を統合した研究（イメージとしては図 3）ということになり，音声研究者と文法研究者とのコラボレーションの研究の可能性が開けるであろう[17]。

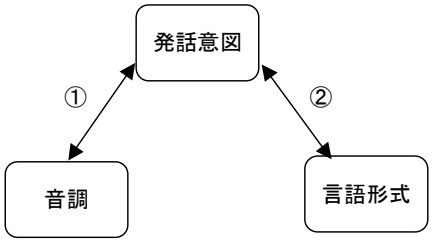


図 2

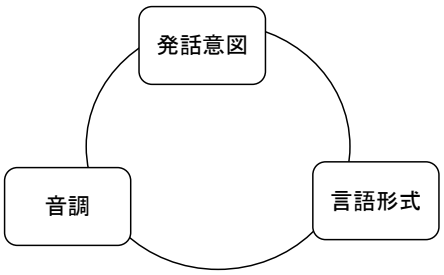


図 3

「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスの整備することで，さらに，以下のような研究の展開の可能性が考えられる。

・「とびはね音調」自体の音声的特徴の記述も可能

いる。

[14] このような研究の例のうち，東京方言を対象としたものとしては，吉沢（1960），上村（1989），郡（2003）などが挙げられよう。また，全国諸方言の質問文の音調を類型化しようとする木部（2010）も，音調と「質問」という発話意図を模索している点で，図 2 の①に分類できる研究かも知れない。

[15] 図 2，図 3 をめぐる議論は，（方言）文法研究者の立場から本共同研究に参加している高木千恵氏とのやりとりから生まれた結果であることを付記しておく。

[16] たとえば，4.節で言及した，現代日本語文法の研究成果などを参照。

[17] 実は，音調，発話意図，言語形式の 3 つの要素を統合した研究も，少なくはない。中でも，文末音調，発話意図，文末詞の 3 つの要素の関係について模索した研究は，精力的になされてきている。東京方言の場合には，轟木靖子氏の一連の研究（轟木 2008 やその参考文献欄を参照）や森山（1989・2001），小山（1997），杉藤（2001）などがある。しかし，言語形式に対する音声研究者の理解不足，逆に音調に対する文法研究者の理解不足の側面があったのではあるまいか。たとえば，森山（1989）の終助詞「ね」の音調の観察に対して，服部（1999），川上（2000）が疑義を呈している（なお，森山氏は森山 2001 で，服部 1999，川上 2000 の指摘を受け，自身の音調観察に「修正が必要である」と述べていることを付記しておく）。また，文末詞よりも「大きな」言語形式において，音調，発話意図，言語形式の 3 つの要素を統合した研究は，これから，音声研究者と文法研究者の協力によって，進めることができるのではないかと考えている。いずれにせよ，服部（1999）や川上（2000）の以下の一節は，それらが発表されてから十数年経過した現代においても傾聴に値し，「文法研究者と音調研究者とのコラボレーション」（田中 2011）は，これからの音調研究の方向性として，模索して生き続けなければならない問題と考える。

・服部（1999）

「音調に関する事実の詳細に十分な注意を払わない文法研究者が少なくないように見受けられる現状」

「例えば↑や↓といった記号の定義の不明確さから来る混乱，不正確または断片的な音調の観察，方言差への不注意，などの理由により有効性を疑われる議論を，特に主として文法を専門とする研究者の論文の中に散見する。」

・川上（2000）

「氏（引用者註: 森山氏）はイントネーション研究をもっぱらにしてこられたお人ではなく，音調現象は多岐微妙なので，多少の聞き漏らしがあったのであろう。さらに言えば，イントネーションに関わる何らかの『理論』を樹立したくなると，それになじむ音調現象のみが増幅されて聞こえてくることは，氏に限らず人間としてほとんど避けがたいことなのかも知れない。」

・「現在のところ人文系の理論的研究ではもっとも有力なイントネーションの捉え方になっている」自律分節音韻理論に基づくイントネーション表記方式である（郡 2011）ToBI(Tone and Break Indices)の日本語版 X-JToBI（前川ほか 2004・五十嵐ほか 2006・五十嵐 2008 を参照）では記述できない（想定していない）の音調の記述法の開発にも寄与[18]

いずれにしても，我々の試みは胎動しはじめたばかりである。我々の目論見が成功するか否かは，現段階では未知数である。しかし，この種の視点による研究にご興味を覚えられた方と協力できる体制を作ることができることを祈りつつ，本稿を閉じることにする。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト（領域指定型）「文末音調と発話意図とを統合した話し言葉のアノテーションの可能性－日本語諸方言の同意要求表現を中心に考える－」（平成 22～25 年度、代表者：岡田祥平）による補助を得ています。

文　献

石井和仁(2000)「London-Lund Corpus に見る英語の同意・確認要求表現について」『福岡大学人文論叢』第 32 巻第 1 号，pp.191-206

五十嵐陽介(2008)「『日本語話し言葉コーパス』の韻律情報」『日本語学』第 27 巻第 5 号

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』国立国語研究所 http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/csj_report/07.pdf(2012 年 1 月 11 日閲覧)

井上史雄(2008a)「ことばの散歩道 123　新方言じゃね」『日本語学』第 27 巻第 9 号，p.21

井上史雄(2008b)「地域語の経済と社会―方言みやげ・グッズとその周辺―第 27 回『古株じゃん　新米じゃね』」　http://dictionary.sanseido-publ.co.jp/wp/2008/12/13/%E5%9C%B0%E5%9F%9F%E8%AA%9E%E3%81%AE%E7%B5%8C%E6%B8%88%E3%81%A8%E7%A4%BE%E4%BC%9A-%E7%AC%AC27%E5%9B%9E/(2012 年 1 月 10 日閲覧)

上村幸雄(1989)「日本語のイントネーション」『ことばの科学』3，pp.193-220，むぎ書房

上野善道(2011)「アクセント」城生佰太郎・福盛貴弘・斎藤純男編『音声学基本事典』，pp.305-311，勉誠出版

川上蓁(1963)「文末などの上昇調について」『国語研究』pp.25-46

川上蓁(2000)「服部氏のネの音調の説に同調」『国語学』第 51 巻 3 号，pp.33-34

簡月真(著)・真田信治(監修)『海外の日本語シリーズ 1　台湾に渡った日本語の現在―リンガフランカとしての姿―』明治書院

神部宏泰(1997)「播磨方言における同意・確認要求の表現法―婉曲表現を中心に―」『ノートルダム清心女子大学紀要　国語国文学編』第 21 巻第 1 号，pp.1-11

木部暢子(2010)「イントネーションの地域差」小林隆・篠崎晃一編『方言の発見　知られざる地域差を知る』，pp.1-20，ひつじ書房

工藤真由美・八亀裕美(2008)『複数の日本語　方言からはじめる言語学』講談社

郡史郎(2003)「イントネーション」上野善道編『朝倉日本語講座 3　音声・音韻』，pp.109-131　朝倉書店

郡史郎(2006)「対人関係・対人態度を反映する韻律的特徴」土岐哲先生還暦記念論文集編集

[18] 『日本語話し言葉コーパス』の対話を検索したが，「とびはね音調」が生起するとされた「形容詞否定形＋ナイ？」といった形式は，1 件もヒットしなかった。これは，『日本語話し言葉コーパス』の対話がカジュアルなものではないこと（常体が使用できる関係の対話ではないこと）が関係していると思われる。『日本語話し言葉コーパス』で「とびはね音調」が観察されない以上，『日本語話し言葉コーパス』の韻律ラベリングに採用されている X-JToBI では，「とびはね音調」は記述できないと思われる。註 3 で引用した前川（2006）の言説も参照。

委員会編『日本語の教育から研究へ』，pp.167-176，くろしお出版
郡史郎(2011)「イントネーション」城生佰太郎・福盛貴弘・斎藤純男編『音声学基本事典』，pp.338-348，勉誠出版
国立国語研究所(1960)『国立国語研究所報告 18　話し言葉の文型(1)―対話資料による研究―』
小針浩樹(1996)「同意要求文の位置と形式」『国語学研究』第 35 集
小山哲春(1997)「文末詞と文末イントネーション」音声文法研究会編『文法と音声』, pp.97-119, くろしお出版
蔡雅芸(1996)「同意要求的疑問文のアクセント核破壊型音調―『これ、面白くない?』について―」『東北大学文学部日本語学科論集』第 6 号，pp.35-46
斎藤純男(2011)「音調」城生佰太郎・福盛貴弘・斎藤純男編『音声学基本事典』，pp.314-315，勉誠出版
杉藤美代子(2001)「終助詞『ね』」の意味・機能とイントネーション」音声文法研究会編『文法と音声Ⅲ』，pp.3-16，くろしお出版
ザトラウスキー・ポリー(1993)『日本語研究叢書 5　日本語の談話の構造分析　勧誘のストラテジーの考察』くろしお出版
高木千恵(2011)「日本語諸方言における同意要求表現～否定疑問形式由来の文末形式を中心に～」日本音声学会第 323 回例会シンポジウム「日本語諸方言における同意要求表現とその音調の諸相」発表時パワーポイント提示資料
田中ゆかり(1993)「『とびはねイントネーション』の使用とイメージ」『日本方言研究会第 56 回研究発表大会発表原稿集』，pp.59-68
田中ゆかり(2010)『首都圏における言語動態の研究』笠間書院
田中ゆかり(2011)「『とびはね音調』は同意要求表現か？」『論集Ⅶ』（アクセント史資料研究会），pp.25-39
谷川依津江(2003)「『ね』のイントネーションと話題にかかわる機能―上昇下降イントネーションと平坦イントネーションを伴う『ね』に注目して―」『日本語日本文化研究』第 13 号，pp.149-159
鄭　相哲(1992)「いわゆる確認要求の『ネ』と『ダロウ』―情報伝達論的な観点から―」『日本学報』11，pp.105-121
都染直也(2007)「音声と音韻」荻野綱男編著『現代日本語学入門』pp.24-40，明治書院
轟木靖子(2008)「東京語の終助詞の音調と機能の対応について―内省による考察―」『音声言語Ⅵ』，pp.5-28，近畿音声言語研究会
服部匡(1999)「終助詞ネの音調に関する森山説への疑問」『国語学』199 集，pp.90-92
平塚雄亮・原田走一郎(2012)「鹿児島県北薩方言の文末詞セン―用法の変化に着目して―」『日本語の研究』第 8 巻 1 号，pp.1-13
前川喜久雄(2006)「イントネーション研究発展の要因」『音声研究』第 10 巻第 3 号，pp.-7-17
御園生保子(2000)「文末に現れるジャナイの用法と韻律の分析をめぐる問題について」山田進・菊地康人・籾山洋介編『日本語　意味と文法の風景―国広哲弥教授古希記念論文集―』，pp.343-355，ひつじ書房
三宅知宏(1994)「否定疑問文による確認要求的表現について」『現代日本語研究』1，pp.15-26
三宅知宏(1996)「日本語の確認要求的表現の諸相」『日本語教育』89，pp.111-122
森山卓郎（1989）「文の意味とイントネーション」宮地　裕編『講座日本語と日本語教育 1 日本語学要説』，pp.172-196，明治書院
森山卓郎（2001）「終助詞『ね』のイントネーション―修正イントネーション制約の試み―」音声文法研究会編『文法と音声Ⅲ』，pp.31-54，くろしお出版
吉沢典男(1960)「イントネーション」『話しことばの文型(1)』，pp.249-288，国立国語研究所
湧田美穂(2003)「『い形容詞+ナイ』の韻律的特徴―アクセント・イントネーション・持続時間の側面から―」『早稲田大学日本語教育研究 III』，pp125-139

BCCWJ における出典情報とトピックおよびレジスターとの関係

ホドシチェク・ボル　（東京工業大学大学院社会理工学研究科）†
仁科　喜久子　（東京工業大学留学生センター）

Comparison of Metadata with Topic and Register in the BCCWJ

Hodošček Bor　(Tokyo Institute of Technology)
Nishina Kikuko　(Tokyo Institute of Technology)

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）には様々なメタ情報が付与されており、その中にはメディア名（サブコーパス名）、出典情報（作家名、出版年、出版社名）、またはジャンルを記述する NDC（日本十進分類表）情報などがある。ある現象のジャンル別傾向を調査するときにメディアに頼ることはしばしばあるが、メディアの間ではそれぞれジャンルとして非常に似ている文書もあれば、著しく異なるものもある。そこで、本稿では BCCWJ の様々なメディアにおけるメタ情報とトピックおよびレジスターとの関係を分析する。

２．トピックとレジスター

まずトピックとレジスターについて簡単に述べる。言語の構造を単純化していえば、文は内容的な要素（内容語）とそれらの内容を構成する機能的な要素（機能表現）からなるといえる。例えば、普通名詞「野球」、動詞「投げる」は内容語、接続詞「なぜならば」、副詞「極めて」は機能表現である（Biber and Conrad, 2009；松吉ら, 2007）。

内容語は本稿で扱うトピック、つまり話題に関連する。言語学では「一文中の主題」として「topic」という用語を用いることがあるが、本論のトピックは文書単位における「話題」に近いものである。

一方、機能表現はレジスターに深く関わるものといえる。本稿では、レジスターを「言語の共時的、機能的な変異であり、状況によって語彙・文法の使用が変化するものおよびコミュニケーションの目的とコンテクストにおいて明瞭に定量的なパターンによって特徴づけられる」と定義する（Biber and Conrad, 2009）。日本語には機能表現が多種多様に存在し、豊富であることから、言語の機能的な変異であるレジスターを分析する上で機能表現を用いることが考えられる。また、文書の文体に深く関わる品詞比率もレジスターを分析する上で有効であると考えられる。

２．１．トピックモデル

トピックモデルは、確率的生成モデルであり、代表的なものとしては Latent Dirichlet Allocation（LDA）がある。LDA はトピック分布の多項分布でモデル化し、トピックの分布に対してディリクレ分布を仮定する（Blei et al., 2003）。本稿では、Yahoo! LDA を用いて 1000 トピックのモデルを BCCWJ で学習した（Smola and Narayanamurthy, 2010）。トッピクモデルの素性は、形態素解析辞書 UniDic の品詞名によって名詞（数詞を除く）、動詞（非自立可能なものを除く）、形容詞（非自立可能なものを除く）、形状詞（助動詞語幹のものを除く）、副詞の品詞から語（短単位）を選択した。表 1 は、LDA モデルにおけるそれぞれのメディアごとのトピックの関連語をそれぞれ示したものである。

例えば、書籍（PB、LB、OB）のグループでは、「顔、目、手、声」「言う、思う」などが共通して出現しており、小説などの創作における具体的な人間の行為、想念などに関連していることが推測できる。韻文（OV）では、「月、花、秋、夜」「赤、白、風、空」など詩歌の題材に見られるトッピクが抽出されている。国会会議録（OM）では「大臣、政府、委

† hodoscek.b.aa@m.titech.ac.jp

員」のような役職のグループ、「言う、訳、こと、風」など言語行為に関する語群が抽出されている。

このように LDA によるトピックの抽出は、それぞれのメディアの特色を提示していることが分かる。

表 1: 各メディアにおけるトピック（高頻度順）

	1 位	2 位	3 位
PB	顔，目，声，そう，手	言う，事，思う，そう，時	事，於く，つく，因る，物
LB	言う，事，思う，そう，時	顔，目，声，そう，手	言う，声，顔，事，そう
OB	言う，事，思う，そう，時	顔，目，声，そう，手	言う，出る，男，電話，入る
PM	人気，スタイル，感，デザイン，使う	バッグ，スカート，ニット，パンツ，スタイル	シャツ，ブランド，ジャケット，プリント，カラー
PN	優勝，回，大会，初，決勝	大統領，米国，関係，政府，外交	首相，自民，コイズミ，政治，総理
OC	方，教える，どう，分かる，出る	言う，今，人，時，知る	言う，事，思う，そう，時
OY	所，もう，後，前，気	言う，事，思う，そう，時	今日，明日，笑い，頑張る，まあ
OW	於く，為，つく，行う，図る	年，パーセント，増加，図，別	事業，整備，年度，施設，実施
OV	夜，日，花，秋，月	風，白い，中，空，赤い	姿，巨大，物，光，今
OT	実験，調べる，事，考える，分かる	計算，数字，数，答え，桁	運動，力，速度，時，物体
OP	月，日，市，申し込み，センター	課，■■，月，平成，市	時，日，午後，分，午前
OL	条，項，規定，当該，於く	条，項，規定，因る，業務	事業，事，指定，定める，大臣
OM	委員，つく，事，大臣，政府	言う，訳，そう，事，風	案，国会，提出，つく，法案

略称：LB、PB、OB：書籍；PM：雑誌；PN：新聞；OC：Yahoo! 知恵袋；OY：Yahoo! ブログ；OW：白書；OV：韻文；OT：教科書；OP：広報紙；OL：法律；OM：国会会議録

２．２．レジスター

本稿では、レジスターとして下記の３種類の異なるデータを均等に重み付けをし、計量する。

a) 松吉ら（2007）の「つつじ：日本語機能表現辞書」に含まれる機能表現

b) Srdanović ら（2008）で用いる推量副詞

c) 品詞比率からなる Modifier Verb Ratio (MVR) という指標

以下、それぞれについて述べる。

２．２．１．機能表現

「つつじ」では、機能表現が階層構造によって構成されている。本稿では、つつじのレベル L2 の区分から異なり表現 435 種類を用いた。表２は各メディアごとの上位５位までの機能表現を示している。

表 2: 各メディアにおける機能表現（高頻度順）

	1 位	2 位	3 位	4 位	5 位
PB	から	こと	よう	という	です
LB	から	こと	という	よう	です
OB	から	こと	です	よう	という
PM	から	です	という	こと	よう
PN	から	など	こと	では	という
OC	です	から	ので	って	こと
OY	です	から	ので	こと	よう
OW	こと	について	から	において	として
OV	から	よう	なり	とき	こと
OT	よう	から	など	こと	には
OP	など	から	です	こと	とき
OL	とき	において	により	による	なら
OM	という	こと	です	から	よう

２．２．２．推量副詞

表３では、Srdanović ら（2008）で使用したものと同じ推量副詞（合計 18 種類）を用い、各メディアごとの上位５位までの推量副詞を示す。

表 3: 各メディアにおける推量副詞（高頻度順）

	1 位	2 位	3 位	4 位	5 位
PB	あるいは	必ず	絶対（に）	恐らく	きっと
LB	あるいは	絶対（に）	必ず	恐らく	きっと
OB	あるいは	必ず	絶対（に）	きっと	恐らく
PM	絶対（に）	必ず	あるいは	きっと	多分
PN	絶対（に）	必ず	あるいは	きっと	必ずしも
OC	絶対（に）	必ず	多分	きっと	もしかして／たら／すると
OY	絶対（に）	多分	きっと	どうも	必ず
OW	あるいは	必ずしも	絶対（に）	必ず	大抵
OV	あるいは	きっと	恐らく	多分	どうやら
OT	あるいは	絶対（に）	必ず	必ずしも	恐らく
OP	必ず	絶対（に）	あるいは	きっと	必ずしも
OL	必ず	あるいは	よほど	ひょっとして／たら／すると	もしかして／たら／すると
OM	あるいは	どうも	恐らく	必ずしも	絶対（に）

この分布を見ると例えば Yahoo! 知恵袋と Yahoo! ブログ、新聞、雑誌などでは、「絶対に」という感情的な表現が高頻度に出現し、白書（OW）、国会（OM）では「あるいは、必ず（しも）、どうも、多分、おそらく」などの婉曲的な表現が出現している。このような特色もレジスターを区別する有用なデータとなると考えられる。

２．２．３．MVR

MVR は文章中における用の類（動詞）とそれらを修飾する相の類（副詞、連体詞、形容詞、形状詞）の比率（100× 相の類の比率／用の類の比率）であり、文章の文体を計る指標とされる（樺島忠夫、寿岳章子, 1965; 冨士池ら, 2011; Hodošček, 2011）。名詞比率が低い場合、MVR が高いほど「ありさま描写的」、低いほど「動き描写的」であることから、ジャンルによる特色を読み取ることができると考えられる。また、名詞比率が高いと「要約的」という。本稿では、サ変名詞などの影響を少なくするために品詞比率を計算する際、BCCWJ の長単位データを用いた。

表 4: 各メディアにおける名詞比率と MVR

		PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL	OM
N	平均値	31.06	28.66	25.81	35.00	40.12	25.01	28.35	44.07	34.67	33.23	49.56	42.75	30.77
	SD	7.01	6.13	4.35	7.16	5.86	6.79	10.30	4.76	6.68	6.15	4.30	3.38	4.90
MVR	平均値	74.98	75.75	76.17	86.23	51.45	94.92	116.08	59.74	54.34	61.55	57.98	30.99	69.71
	SD	24.63	21.54	20.64	32.79	19.43	86.04	151.25	25.49	42.83	23.49	9.15	12.76	15.19

表４では、Yahoo! 知恵袋と Yahoo! ブログにおける MVR の標準偏差が大きいのに対し、広報誌、新聞、国会会議録、法律書における MVR の標準偏差が小さいことが明らかになった。つまり、Yahoo! 知恵袋とブログは、多様なテキストが混在している一方で、広報誌や新聞などでは、事のありさまを描写する文書から成りたっていることが推測できる。名詞比率からは、広報誌、白書、法律および新聞が要約的なメディアであることが分かった。

３．考察

BCCWJ 中の様々なメディア間の差異を計量するために、前述のトピックモデルとレジスターの観点から BCCWJ に含まれる全サンプルをメディアごとにまとめて観察した。表５はスピアマンの順位相関係数でメディア間のトピックとレジスターのそれぞれの相関を示したものである。

表 5: メディアにおけるトピックおよびレジスターの相関

トピック												
	PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL
LB	0.85											
OB	0.50	0.66										
PM	0.58	0.56	0.43									
PN	0.56	0.55	0.29	0.59								
OC	0.25	0.18	0.28	0.54	0.33							
OY	0.18	0.22	0.34	0.60	0.37	0.77						
OW	0.33	0.17	−0.09	0.10	0.49	0.06^{ns}	−0.11					
OV	0.28	0.42	0.42	0.27	0.16	0.16	0.33	−0.20				
OT	0.52	0.54	0.35	0.34	0.38	0.19	0.20	0.24	0.29			
OP	0.14	0.09	-0.02^{ns}	0.17	0.37	0.10	0.13	0.29	−0.01	0.13		
OL	0.24	0.04^{ns}	−0.15	−0.08	0.29	0.04^{ns}	−0.17	0.61	−0.22	0.12	0.23	
OM	0.18	0.10	-0.04^{ns}	-0.01^{ns}	0.36	-0.03^{ns}	−0.15	0.53	−0.18	0.08	0.21	0.56

レジスター												
	PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL
LB	0.93											
OB	0.89	0.90										
PM	0.88	0.88	0.87									
PN	0.78	0.77	0.78	0.81								
OC	0.81	0.80	0.79	0.83	0.74							
OY	0.86	0.85	0.83	0.85	0.76	0.87						
OW	0.66	0.65	0.65	0.67	0.76	0.60	0.62					
OV	0.64	0.64	0.65	0.66	0.69	0.59	0.62	0.63				
OT	0.74	0.73	0.74	0.76	0.82	0.70	0.71	0.77	0.71			
OP	0.72	0.71	0.71	0.74	0.76	0.72	0.72	0.71	0.61	0.78		
OL	0.41	0.40	0.40	0.42	0.49	0.38	0.38	0.60	0.44	0.53	0.51	
OM	0.72	0.72	0.72	0.72	0.72	0.70	0.70	0.68	0.57	0.69	0.71	0.44

* 注意：$n.s.$ 以外の値はすべて $p < .05$

4．まとめ

以上の分析から、あるメディアがほかのメディアと大凡どの程度トピックおよびレジスターが異なるかが分かった。今後の課題としては、メディアごとのサンプルに分析を拡大することが必要である。

文　献

樺島忠夫、寿岳章子（1965）『文体の科学』綜芸舎

冨士池優美、小西光、小椋秀樹、小木曽智信、小磯花絵（2011）「長単位に基づく媒体・カテゴリ間の品詞比率に関する分析」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. 273–280.

松吉俊、佐藤理史、宇津呂武仁（2007）「日本語機能表現辞書の編纂」自然言語処理，Vol. 14，No. 5，pp. 123–146.

Biber, Douglas, and Susan Conrad (2009) Register, Genre, and Style. Cambridge: Cambridge Textbooks in Linguistics.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.

Hodošček, Bor (2011) “Word Class Ratios and Genres in Written Japanese, Revisiting the Modifier-Verb Ratio”, Acta Linguistica Asiatica, Vol. 1, No. 2, pp. 53–62.

Smola, A., and S. Narayanamurthy (2010) “An Architecture for Parallel Topic Models”, In The Proceedings of the VLDB Endowment (PVLDB), Vol. 3, No. 1, pp. 703–710.

Srdanović, I., B. Hodošček, A. Bekeš, and K. Nishina (2009)「ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用」自然言語処理，Vol. 16，No. 4，pp. 29–46.

関連 URL

つつじ：日本語機能表現辞書	http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/
Yahoo! LDA	https://github.com/shravanmn/Yahoo_LDA

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−

田頭(谷口) 未希（東海大学 外国語教育センター）†

Boundary Pitch Movement and Usage of Conjunctive Particle “ga”
- The Analysis of the CSJ -

Miki TAGASHIRA-Taniguchi (Foreign Language Center, Tokai University)

１．はじめに

発話の音調は様々な要因が関わっている。日本語では、文末や文節末の音調と、語と語のアクセントの強弱関係が発話の音調を決定する重要な役割を担っているとされる（郡 2003）。文末や文節末の音調を考えた場合、終助詞など文末の音調とその意味機能については多くの先行研究がある一方で、統語的にみた文節末での音調に関する研究は非常に少ないといえる。そこで『日本語話し言葉コーパス』（以下、CSJ と略す）を用い、文末・文節末の音調について、意味・用法との関連を定量的に分析し、対応関係を体系的に記述することを研究全体の目的とする。田頭(谷口)(2011)では 1 モーラからなる接続助詞の全体的傾向を分析するとともに接続助詞「が」について主に音調パタンに着目し分析を行った。本稿では、田頭(谷口)(2011)で用いた「が」の意味・用法の分類を見直し、下降調とその意味・用法との関係について報告する。

２．分析データ

分析資料として CSJ 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）を分析した。このうち、本稿では特に音調パタンが下降調という条件に限定して分析を行う。接続助詞「が」の総数は 643 例であるが、そのうち下降調は 79 例で､意味･用法の判定ができなかった 1 例を除いた 78 例が本稿での分析対象データである。本稿で下降調のみに注目した理由は、CSJ のコアデータ全体の傾向として、韻律句末の約 75%が下降調であるのに対して、CSJ にあらわれる接続助詞「が」に関しては下降調がわずか 10~20%しかないこと、この傾向は他の 1 モーラからなる接続助詞の句末音調の出現頻度ともかなり異なること(田頭(谷口) 2011)があげられる。接続助詞「が」において極端に下降調が低い理由を意味・用法との関係から探る。

３．音調パタンと意味用法

３.１　韻律句と句末音調

本稿で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、「韻律句」は時間軸に沿って示される、冒頭の上昇から発話末にかけて下降していく基本周波数のひとつの山のまとまりと定義する。CSJ では Intonation Phrase（以下、IP と略す）と Accentual Phrase（以下、AP と略す）の２つの韻律句がある。

CSJ では韻律句末の音調の型として、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%）、そして上昇下降上昇調（HLH%）の５つの音調が定義されている。

３.２　「が」の音調

「新明解日本語アクセント辞典」（秋永 2002[1]）によると、接続助詞「が」の語彙的情報として指定された音調は以下の２つである。

† t-miki@tokai-u.jp

[1] 付録(74)～(79)の表より。

(1)平板式動詞[2]に付く場合には、助詞の第一拍から低く下がってつく（例）わらうが[3]
(2)起伏式動詞に付く場合には、動詞の形を変えないで、低く下がってつく（例）よむが
したがって、語彙的には接続助詞「が」はそれ自身では音調変化を持たず、前接要素に続けて自然に下降していく音調をとる。ただし、これらはあくまでも語彙的情報として指定された音調であり、イントネーションによる影響をうけることがある点は秋永（2002）にも明記されている。

３.３　「が」の意味用法

接続助詞「が」[4]の意味機能・用法は、主に「逆接」や「対比」関係を表すほかに、「談話主題の提示」「前置き」「注釈」などに用いられるとされる（森田 1980）。先行研究であげられている意味・用法を基に、本稿では以下の６つの意味・用法を設定し、分類を行った[5]。

(1)「逆接・対比[6]」
- 授業は厳しいが、楽しい（M）
- 夏は日が長いけれども、冬は短い（M）
- よく猫は撫でてあげようと思って近づくと逃げてしまうので　私も追いかけるのが大変なんですが　その点犬は撫でられると尻尾振って喜んじゃいますから全くかわいいもんです（S00F0031）
- このアドバイザーグループとは　サークルとは違って　あ　似たようなものなんですが　え　違うとことがありまして　その違うところというのは（S00F0088）

(2)「並列・累加」
- 英語ももちろんできるが、フランス語も話せる（M）

(3)「談話主題の提示」
- 昨日の話ですけれど、どうなりました（S）
- 比較的　あの　空室の部屋がないようなさまで　え　非常に　あのー　心地良く　過ごしております　今では　あのー　私がスポーツクラブへ　毎日通っているんですが　んー　スポーツクラブの（S03F1477）

(4)「補足説明・前置き」
- 1912 年というのは明治の終わった年ですが、この年に私の姉は生まれました（S）
- おたくの大学に入りたいという学生がいるんですが、手続きはどのようにしたらいいんですか（S）
- どうも　あの　遠因　ま　遠い原因と書きますが　あのー（S02M0068）
- それで　あのー　どうしたら　こう　痩せられるだとうと思い　常に　こういろいろ研究をしてたんですが　ま　どんぐらい太っていたかと言うと（S00M0065）

(5)「注釈[7]」

---

2 動詞を例に挙げたが、他の品詞でも同様の音調パタンをとる。

3 本稿では便宜上、前節要素と比べ低く下がる音を下線付きで示すことにする。

4 接続助詞「けど（「けれども」）と意味機能や用法が似ていることから「が・けど」類とも呼ばれる。用例には「が」または「けど」「けれども」を示す。

5 用例は M は森田（1980）、S は齋藤（2011）による。CSJ からの用例には Talk ID を明記している。当該要素の「が」は太字で示し、CSJ のデータについては句読点位置と推定される箇所でのスペースは筆者による。

6 「逆接」か「対比」かは、前後の節関係に因果関係がみとめられるか、意味的コントラストをなしているかによって分類する研究もある（渡辺 2000）が、その違いには依然曖昧さが残されているため、本稿ではひとつの分類として扱う。

7田頭(谷口)(2011)で、発話は切り出されているが主要なテーマを補いたい場合や主題に付随する情報を付

- 夜分遅く恐れ入りますが、太郎君はいらっしゃいますか（S）

(6)「言いさし・言い切りの回避」

- やるだけはやってみます**が**‥‥（M）
- まー　昼の生活夜の生活っていろいろありますけれども　う　多分　分かるようなねたなんで　えー　これ以上触れません**が**　ま　ということで（S00M0112）
- ま　ちょっと　教育じみた話で　私も　あの　長年　人事社員教育とやってたもんで　どうしてもそのような　あ　ことに　落ち着くような気がしておりま**す**　えー　あー　その辺が　これのお話のことをご披露申し上げました（S05M0412）

４．結果と考察

接続助詞「が」はその品詞の特徴として文節に置かれ、統語的切れ目となる。音声的には、「が」の直後にポーズの挿入や次にくる韻律句でピッチの立て直しが起こる可能性が十分に考えられる。実際、ここで分析対象とした「が」の 78 例の下降調は全て韻律境界が IP で、AP は 1 例もなかった。これは、下降調となる「が」では、次の韻律句頭でピッチの立て直しが起こるか、長いポーズを伴っている場合などであることを意味する。また、「あー」「えー」などの複数のフィラーの連続が生じている例もみられた。

IP 境界だからといって直後に必ずポーズが挿入されるとは限らないが、長いポーズは IP 境界を決めるひとつである。そこで、「が」の直後の 200msec 以上のポーズの有無を調べた。ポーズを伴わないものは 78 例中 13 例で、短いものではポーズの長さが 50msec 程度であった。ポーズの挿入がある 65 例は 230〜4500msec と様々であった。文節末の音調について、上昇調は直後にポーズを伴うのが普通である（郡 2003）という報告もある。「が」に関しては下降調でも直後にポーズを伴うことが多く、しかもかなり長いポーズを伴う場合もあ

表 1 意味・用法の頻度とポーズの平均時間長

意味・用法	生起数	ポーズの平均時間長 msec
補足説明・前置き	46 (60%)	761.67
逆接・対比	24 (31%)	691.13
言いさし	6 (8%)	1747.49
談話主題の提示	1 (1%)	523.04

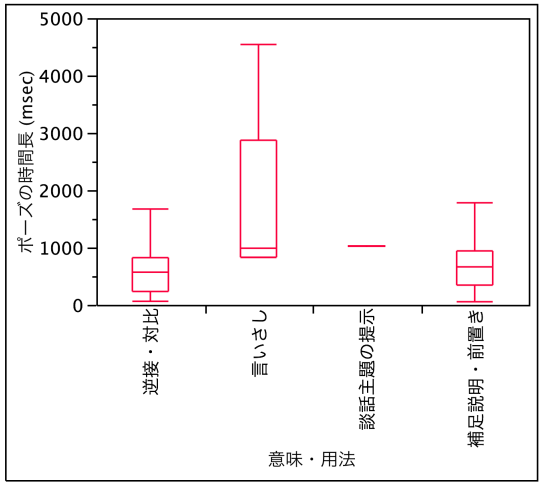


図 1 意味・用法によるポーズの時間長の平均と分散

け加えたい場合と定義した「注釈」は、本稿では「補足説明・前置き」と「注釈」の 2 つに分けて分類することにした。田頭(谷口)(2011)では決まり文句のような単なる前置き的表現（本稿での「注釈」）と発話の内容を補足する前置き的表現（本稿でのを「補足説明・前置き」）同じカテゴリに分類したが、本稿では両者を区別してより詳細な分類の観点から再度分析を行うことを目的としたためである。

るようである。

次に意味・用法と音調パタンとの関係をみてみる。3.3 節で 6 つの意味・用法の分類をあげたが、下降調では「並列・累加」と「注釈」用法に分類される例はみられなかった。表 1 に下降調の意味・用法別の頻度と、「が」直後のポーズの平均時間長を示す。下降調をとる場合、補足説明や前置きの用法であることが最も多く、半数以上を占めた。2 番目に多かった逆接・対比の用法を合わせると「が」の下降調の 9 割がこの 2 つの用法のいずれかであることになる。先にも述べた通り、「が」の下降調は IP 境界のみであったことを考えると、補足説明・前置きや逆接・対比の用法では韻律的により大きな境界を作っていることになる。談話構造の観点から、補足説明・前置きは付け足しではあるけれど、その部分だけで意味的にひとかたまりになっていると考えられ、そこが音声的にも大きな切れ目を作って独立したかたまりとなっていると解釈できる。さらに補足説明や前置きである内容が特別強調すべきことでなければ、音声的際立ちは必要でないため、下降調が用いられる割合が高くなる。一方、最も少なかったのは談話主題の提示の用法であった。これは逆に、境界という側面からは、新しいトピックを導入するという意味で次に続く発話とは強く関連しているため、IP となる可能性が低くなると予測される。一方で新しいトピックの導入は、場合によってはそこを際立たせる必要が生じることも考えられ、この点では下降調以外の音調を使う方が音声的にトピックの導入としてのひとかたまりを示すことができるといえよう。しがたって、談話主題の提示には下降調を用いないというストラテジーは十分に考えられる。

直後にポーズを伴う場合のポーズの平均時間長は、言いさし用法の場合が最も長く、最も平均が短かった談話主題の提示の用法に比べ、3 倍以上の長さであった。言いさし用法は「が」に続く事柄と直接的に関係していない事柄が多いと考えられ、ポーズが長くなるのは全体の生起数が少ないことを考慮しても、ある意味妥当な結果であると考えられる。

５．まとめ

CSJ をデータとし、接続助詞「が」が下降調をとる場合の意味・用法との関係について分析した。「が」が下降調である場合、境界は IP のみが観察され、「が」の直後にポーズを伴うことが多かった。意味・用法の観点からは補足説明や前置きの場で下降調が用いられることが最も多かった。これは補足説明や前置きはあくまで付け足しであるため、そこには音声的際立ちを置く必要性が低く、そのため下降調が用いられやすいといえる。

謝　辞

本研究は、文部科学省科学研究費補助金 若手研究(B)「日本語の自発音声における韻律句末音調と意味機能の分類に関する研究」による補助を得ています。

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

郡史郎(2003)「イントネーション」『朝倉日本語講座 音声 3 音声・音韻』上野善道（編）、pp.109-131、朝倉書店

齋藤美穂(2001)「接続助辞ガ・ケレドモの意味・機能と文法的制約」『阪大日本語研究』23、pp.33-55（http://ir.library.osaka-u.ac.jp/metadb/up/LIBNIHONGOK/23-02.pdf からダウンロード可能）

田頭(谷口)未希(2011)「話し言葉にみられる接続助詞の音調−１モーラ接続助詞の場合−」『人工知能学会研究会資料』ISSN 0918-5682 SIG-SLUD-B003-04、pp.19-22

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

用例に基づく複合動詞の構造分析と教育への応用

山口昌也　(国立国語研究所言語資源研究系) †
井上 優　(麗澤大学外国語学部)
柏野和佳子　(国立国語研究所言語資源研究系)
北村雅則　(名古屋学院大学商学部)
白井清昭　(北陸先端科学技術大学院大学情報科学研究科)
千葉庄寿　(麗澤大学外国語学部)

Analysis of Japanese Compound Verb Based on Examples and its Application to Education

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)
Masaru INOUE (Faculty of Foreign Studies, Reitaku University)
Wakako KASHINO (Dept. Corpus Studies, NINJAL)
Masanori KITAMURA (Faculty of Commerce, Nagoya Gakuin University)
Kiyoaki SHIRAI (School of Information Science, JAIST)
Shouju CHIBA (Faculty of Foreign Studies, Reitaku University)

1　はじめに

本稿では，大量の用例に基づいて，複合動詞とその構成動詞について，格要素間の関係を分析する。また，分析用のデータ，および，分析結果を日本語教育へ応用する計画についても述べる。

なお，本研究は，国立国語研究所の共同研究プロジェクト「文脈情報に基づく複合的言語要素の合成的意味記述に関する研究」の一環として行っている。本研究は複合的言語要素として，複合動詞を扱ったものである。本稿では，日本語教育への応用について触れるが，プロジェクトとしては，国語辞典編集，語義の自動分類，テンス・アスペクト研究など，国語学，言語学，自然言語処理などと連携させつつ，研究を進めている。

2　複合動詞の構造分析

2.1　概要

本稿では，日本語の複合動詞の格要素に関して，構成する動詞の格要素との関係を分析する。具体的には，(1) 複合動詞と構成動詞の格要素集合が重複する度合いを実際の用例に基づいて計算し，(2) 格要素集合の重複度と複合動詞・構成動詞の格支配構造の関係について考察する。なお，本稿で扱う複合動詞は,「語彙的複合動詞」（影山 1993）である。

日本語の複合動詞と構成動詞との関係分析として，山本 (1984) の格支配構造による分析がある。この研究によれば，複合動詞 (Vc) は，前項動詞 (V1)，後項動詞 (V2) との格支配構造の関係に基づいて，次の四つに分類できるとしている。

- I 類：V1，V2 どちらも Vc の格要素と格支配関係を有するもの (例：「投げ捨てる」「叩き切る」)
- II 類：V1 だけが Vc の格要素と格支配関係を有するもの (例：「見上げる」「書き込む」)
- III 類：V2 だけ Vc の格要素と格支配関係を有するもの (例：「打ち重なる」「引き起こす」)
- IV 類：V1,V2 どちらも Vc の格要素と格支配関係を有しないもの (例：「繰り返す」「取り組む」)

このうち，格支配関係を有する場合，格要素の名詞は複合動詞，構成動詞のいずれの文中でも適格である。例えば，I 類の用例 E1 は，E1a，E1b のように，V1, V2 の用例を作ることができる。

†masaya@ninjal.ac.jp

- (E1) 太郎が煙草を投げ捨てる
- (E1a) 太郎が煙草を投げる
- (E1b) 太郎が煙草を捨てる

このように複合動詞と構成動詞が格支配構造上の関係を有する場合，格要素の名詞も対応関係を有する。したがって，理論上は，複合動詞の格要素の集合は，構成動詞の格要素の集合の部分集合となるはずである。

そこで，本稿では，複合動詞と構成動詞の格要素集合の重複度を調査し，複合動詞・構成動詞の格支配構造における関係の有無が，格要素集合の重複度とどのように関係しているのかを分析する。

2.2 格要素の重複度

本稿では，文における重要性や，格ごとの出現頻度を考慮し，他動詞の場合はヲ格，自動詞の場合はガ格の格要素の重複度を計算する。

格要素の重複度 OV_i は，複合動詞の格 i が取り得る格要素集合 E_{ci} を基準とし，それらが構成動詞の格 i の格要素集合 E_{si} と重複する割合を表す。定義は，次のとおりである。なお，w_a，w_b は格要素の名詞，$n(w)$ は w の出現ページ数を表す。

$$OV_i = \sum_{w_a \in E_{ci} \cap E_{si}} n(w_a) / \sum_{w_b \in E_{ci}} n(w_b)$$

2.3 分析データの構築

2.2 節の分析を行うには，特定の分野に偏らない，大量の用例とそれを格解析した結果が必要となる。そこで，本研究では Web から用例を収集することにした。収集手順は，次のとおりである。

(1) 『複合動詞資料集』(野村・石井 1987) から，複合動詞の構成要素として多用される動詞上位 10 語を選択し,「種」とする。そして，それぞれ 10000 ページ（前項の動詞用に連用形で 5000 ページ，後項の動詞用に終止形で 5000 ページ) を Baroni(2006) の方法で収集する。
(2) 収集した Web ページを形態素解析した後,「種」動詞を含む動詞の連続を抽出し，複合動詞候補とする。そのうち，50 ページ以上に出現した候補を目視確認し，分析対象の複合動詞とする。
(3) 分析対象の複合動詞に対して，Baroni(2006) の方法で 2000 ページの Web ページを収集する。
(4) 収集した Web ページを形態素解析した後，収集対象の複合動詞を含む文を抽出し，構文解析，および，格解析を行う。なお，格解析には，KNP (ver.3.01, http://nlp.ist.i.kyoto-u.ac.jp/) を利用した。
(5) 収集した複合動詞の構成動詞を種として，再帰的に 1〜4 を繰り返す。

以上の手順で複合動詞 783 語，構成動詞 194 語の用例を収集した（原稿執筆時)。平均用例数は，それぞれ 1312.5，14078.4 例である。

2.4 実験

構築した複合動詞のうち，次の条件を満たす複合動詞をランダムに 100 個抽出し，分析対象とした。

- 複合動詞の用例が 1000 個以上収集されていること。また，構成動詞の用例が前項・後項双方とも 2000 個以上収集されていること
- $\sum_{w \in E_{ci}} n(w) \geq 50$，$\sum_{w \in E_{si}} n(w) \geq 50$ であること。ただし，収集した Web ページに出現する割合が，複合動詞の場合，0.25% 未満，構成動詞の場合，0.05% 未満の名詞は除外する。

表 1: 複合動詞の内訳

分類	複合動詞数	構成動詞数	一致率 (%)
I 類	39	48	81.0
II 類	24	30	80.0
III 類	21	26	70.0
IV 類	16	21	64.0
全体	100	80	76.5

以上の条件を満たす複合動詞 100 個に対して，複合動詞と構成動詞との格支配構造上の関係の有無を人手で与えた。山本 (1984) の 4 分類で集計した結果を表 1 に示す。なお，複合動詞が多義の場合，いずれかの語義で格支配の関係が認められれば，関係ありとしている。

表 1 の複合動詞を対象に，人手による格支配関係の判別結果と重複率との関係を見てみる。図 1 に関係のある場合の重複度 ($\mu = 61.1, \sigma = 23.7$)，図 2 に関係のない場合の重複率 ($\mu = 31.8, \sigma = 23.8$) をヒストグラムとして示す。横軸は重複度，縦軸は頻度である。

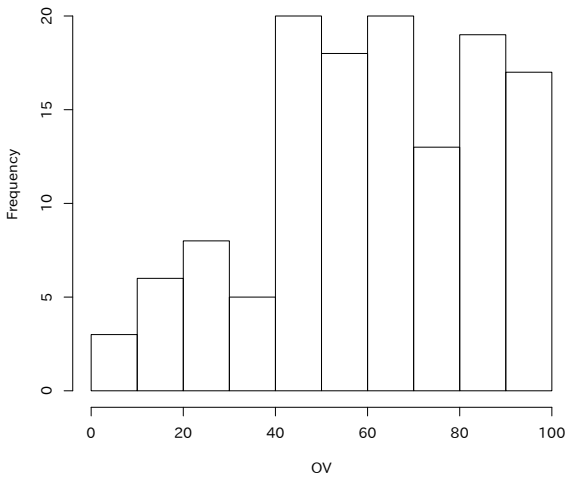


図 1: 重複度 (格支配関係あり)

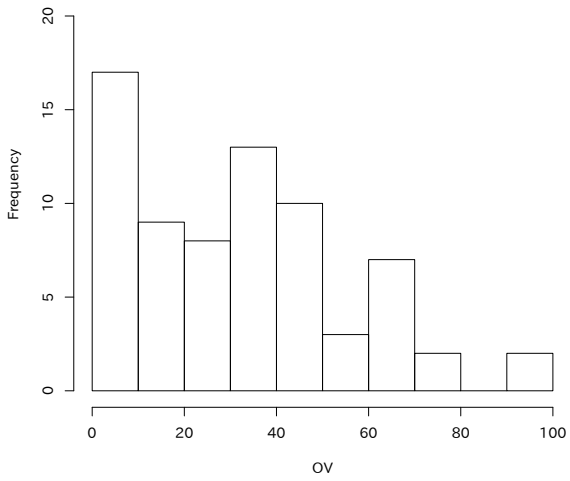


図 2: 重複度 (格支配関係なし)

2.5 考察

まず，人手で付与した格支配関係の有無と重複度による判別結果との一致率を求める。ここでは，閾値 t を設け，$OV_i \geq t$ のとき，格支配関係があると判定するものとする。閾値 t は，0.5 刻みで変化させ，人手で付与した格支配関係の有無との一致率が最大になる値を求めたところ，$t = 38.5(\%)$ となった。表 1 の「一致率」欄に結果を示す。

以上のことより，一致率の面からは，重複度 38.5%に格支配関係の有無の境界がある。この結果や図 1 では，理論とは異なり，格支配関係ありの場合も，重複度の低い場合が存在する。そこで，図 1 の閾値以下の複合動詞を見てみると，最も大きな原因は複合動詞の多義性にあった。例えば,「過ぎ去る」と「去る」のガ格の格要素集合の重複を求めると,「嵐」「台風」「ブーム」などは「去る」の用例と重複するに対して,「時間」「 1 年」などは重複しない。この二つの名詞のグループは，大辞林 (松村 2006) では，二つの語義に対応している。

また，逆に格支配関係なしの場合も，重複度が高い複合動詞が存在する。図 2 の閾値以上の複合動詞を調べてみると，構成動詞の格要素が過剰に重複することが主な原因だった。例えば,「取り扱う」の前項動詞「取る」は，接頭辞的に用いられているため，格支配関係はない。しかし,「取り扱う」のヲ格の格要素集合に「商品」「製品」など,「取る」と共通する名詞が多数含まれ，重複度が高くなる。

3　日本語教育への応用

日本語学習者にとって複合動詞の習得が困難なことは，従来より指摘 (松田 2002 など) されている。そこで，上記の研究成果を日本語教育に応用することを計画している。

その試みの一つとして，収集した用例を検索するシステムを，試験的に全文検索システム『ひまわり』(http://www2.ninjal.ac.jp/lrc) で実現した。実行例を図 3 に示す。この図のとおり，用例を検索し，格要素の一覧などを閲覧することができる。今後，格要素の重複度を応用して，複合動詞と構成動詞との関連を示す仕組みを導入する予定である。

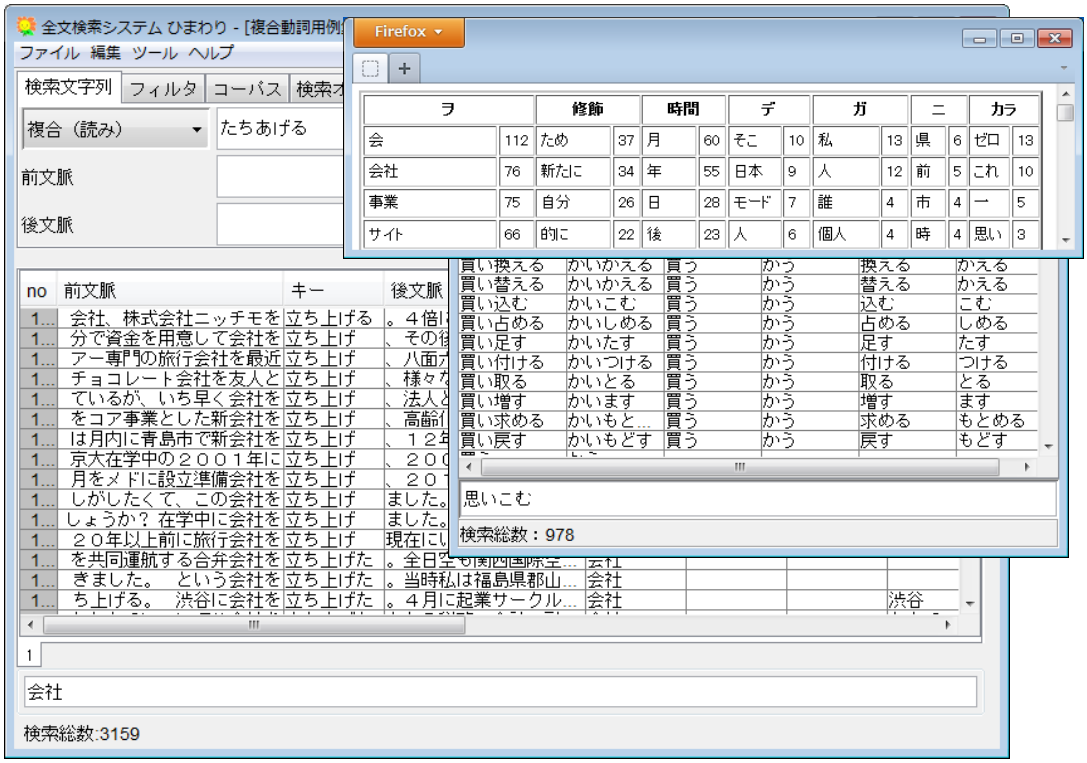


図 3: 全文検索システム『ひまわり』での実現例

4　おわりに

本稿では，日本語の複合動詞の格要素に関して，構成する動詞の格要素との関係を分析するとともに，その応用例として，日本語教育向けの複合動詞用例検索システムを示した。

参考文献

影山太郎 (1993) 文法と語形成，ひつじ書房

山本清隆 (1984) 複合動詞の格支配，都大論究，Vol.21，pp.32-49

M. Baroni and S. Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004

野村雅昭，石井正彦 (1987) 複合動詞資料集，科研費特定研究 (1) 言語データの収集と処理の研究

松田文子 (2002) 複合動詞研究の概観とその展望 —日本語教育の視点からの考察—，言語文化と日本語教育 増刊特集号，pp.170–184

松村 明 (編) (2006) 大辞林第３版，三省堂

日本語話し言葉コーパスにおける句末音調のバリエーション

菊池英明（早稲田大学人間科学学術院）
宮島崇浩（早稲田大学人間科学学術院）

Variation of Tones at the Accentual Phrase Edge in the Corpus of Spontaneous Japanese

KIKUCHI Hideaki (Faculty of Human Sciences, Waseda University)
MIYAJIMA Takahiro (Faculty of Human Sciences, Waseda University)

1. はじめに

表現豊かな音声が伝える様々な情報について、科学的解明や工学的応用の関心が高まっている(Erickson(2005), Schuller(2009))。発話の速さや大きさ、イントネーション、声質など、音声表現を豊かにする音響特徴は多数あるが、その中でもアクセント句末の音調が様々な非言語的情報を伝達することがわかっている。Venditti et al.(1998)は、アクセント句末に生じるピッチの変動を”BPM: Boundary Pitch Movement”と表現して、日本語東京方言における句末音調(ピッチの変動のない音調は含まない)について、生成・知覚の双方の観点で 5 種類の音調が独立して存在することを明らかにした。　日本語話し言葉コーパス(CSJ: Corpus of Spontaneous Japanese)には X-JToBI のスキーム(前川ら(2001))に基づいてラベリングがなされており、付与されたラベル系列のパタンからは、日本語(の主に東京方言)の話し言葉においては主に 7 種類の句末音調(ピッチの変動のない音調を含む)が存在するといえる(前川(2011))。筆者らは、表現豊かな音声の特性を調べることを目的に、声優や俳優などに多様な状況設定を与えて演技音声を収集することにより多様な音声表現コーパスを構築している(Miyajima et al. (2011))。これまでに収集した 3000 発話以上もの音声においても、上述の X-JToBI によって 8 種類のパタンで句末音調を表現できる見通しを得ている。

しかしながら、CSJ におけるイントネーションラベリングにおいて、BPM の種類の判断やラベル付与位置に困難をおぼえたケースは数多くあり、実際の基本周波数(F0)変動のパタンが CSJ においてどのように分布しているかを詳細に観察する必要がある。本研究では、CSJ における句末音調のバリエーションについて、まず X-JToBI ラベルに基づく BPM の出現頻度分布分析結果を示し、句末モーラ区間での F0 変動のパタンを上述の多様な音声表現コーパスとの比較とともに観察する。

2. データ

2.1 CSJ

2011 年にリリースされた CSJ 第三版(CSJ(2011))のコア(分節単位ラベルとイントネーションラベルが付与された 201 講演)と、それを用いて作成されたアクセント句単位 XML を用いる。アクセント句単位 XML は、原則として X-JToBI の BI ラベルの値が 2 以上の位置を境界とする単位でアクセント句を構成し、論理的に包含するトーンラベルの情報を下位要素として表現する形で記述される。これによって、アクセント句ごとに BPM の種類と句末モーラの始端・終端時間位置、句末モーラの無声化の有無などの情報を容易に取得できる。

なお、CSJ にはモノローグを中心に、学会講演(APS)、模擬講演(SPS)、再朗読(R)、対話(D)の 4 種類のスタイルの談話が存在する。以降はこれらの種別を略称で表記する。

2.2 多様な音声表現コーパス

筆者らは、声優や俳優に指示を与えて多様な音声表現を収集してコーパス(通称「千の声コーパス」、以降"SEN"の略称を用いる)を構築する試みを 2008 年より続けている。指示の具体的な例を表 1 に示す。

以下では、こうした指示を受けて 1 名の 40 代女性声優が発声した発話内容「あーそうですか」の 100 発話のデータを用いる。収集方法の詳細やこのデータにおける物理的・心理的多様性の検証については Miyajima(2011)を参照されたい。

CSJ と同様に、分節単位ラベルと X-JToBI ラベルを付与しており、以降の分析ではこれらのラベルを用いる。

表 1 表現豊かな音声表現を得るための指示の例

共通	発話時の場所・状況	大家族を取り扱った特集において(テレビ番組)
	発話者と聞き手の関係	親子
聞き手	年齢／性別	１０歳未満／男
	職業・役柄	小学生
	人物像	典型的なやんちゃな小学生。元気があり待っている状態
発話者	年齢／性別	３０代／女
	職業・役柄	主婦
	人物像	元ヤンのヤンママと言った感じ。言葉遣いはキレイではない。
	発声時の背景	子供のだらしなさに対し、思わず声を張って叱る様子

3 句末音調タイプの出現割合

本章では、句末音調の種類の出現割合を、コーパスごとおよび談話のスタイルごとに観察する(図 1 参照)。なお、"L%>"と"H%>"は、X-JToBI においてそれぞれ"L%"(BPM を伴わない音調)と"H%"にトーンの引き延ばしをあらわす">"(エクステンダー)が付与されたものを示し、いずれも"L%", "H%"の内数である。図 1 より、CSJ においては再朗読(R)を除いて概ね出現割合が似通っていることがわかる。再朗読に出現しない HL%や L%>、H%>は話し言葉を特徴づける音調であるといえる。CSJ と SEN の比較においては、多様な音声表現の収集を意図したコーパス SEN に比べて CSJ は"HLH%"と"LH%"、"L%>"、"H%>"などの音調の出現割合が著しく低いことがわかる。

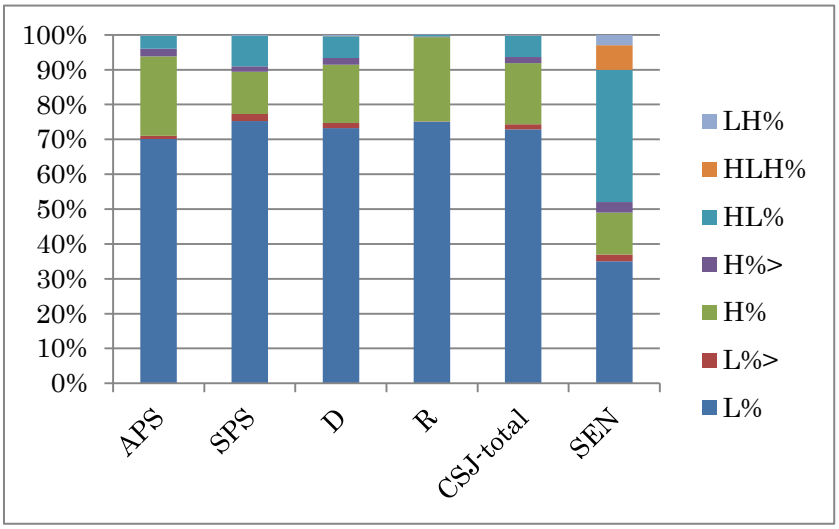


図 1 CSJ と SEN における談話のスタイルごとの句末音調タイプ出現割合

4　句末モーラにおける F0 変動

本章では、アクセント句末のモーラにおける F0 変動のパタンを CSJ と SEN で比較観察する。いずれも F0 についてはセミトーンで話者正規化した。まず図 2 に、CSJ の講演種別ごとにランダムに抽出した 50 か所の F0 変動の分布を示す。図 3 に SEN のランダムに抽出した 50 音声の F0 変動の分布を示す。なお、図 3 には、「怒り」「喜び」などの典型的な感情表現のみを指示として与えた際の同一話者による音声表現(Miyajima(2011))における分布を比較のために示す。　視認による判断ではあるが、APS と R が比較的分布が狭いのに対して、SPS と D は高低および時間方向のいずれも広がりを見せており、図 3 の SEN と同様に多様な F0 変動パタンが出現していることがわかる。

今後はこれらの分布の違いを定量的に計測するとともに、BPM の分類と F0 変動の形状との関係を調べる予定である。

謝辞

本研究は国立国語研究所（言語資源研究系）基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）および萌芽・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。

日本語話し言葉コーパスのアクセント句単位 XML は、小磯花絵、西川賢哉、野口広彰の 3 氏との共同作業によって構築した。ここに記して感謝の意を表する。

参考文献

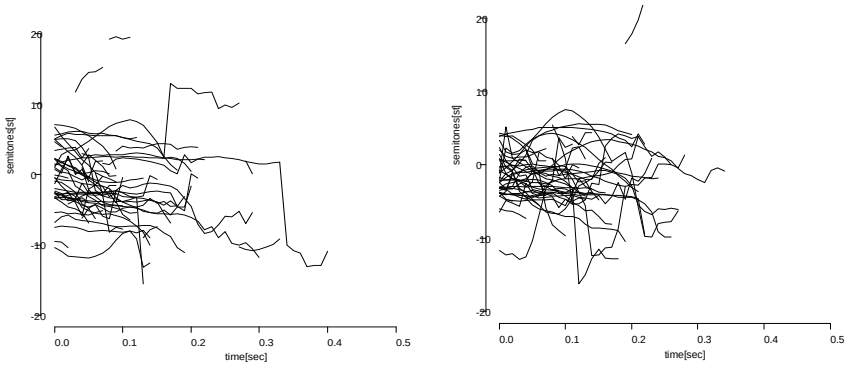
D. Erickson (2005). “Expressive speech: Production, Perception and Application to Speech Synthesis”, Acoust. Sci. & Tech., vol.4, no.26, pp.317-325.

B. Schuller, S. Steidl, A. Batliner (2009). “The INTERSPEECH 2009 Emotion Challenge”, Proc. of INTERSPEECH 2009, pp.312-315.

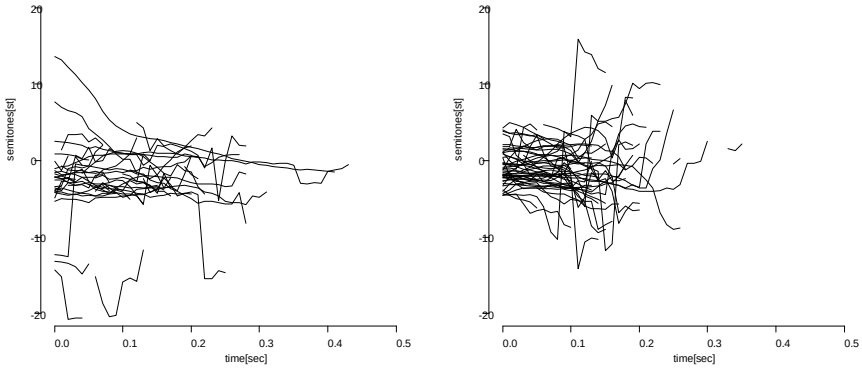
J. Venditti, K. Maeda, and J. P. H. van Santen (1998). “Modeling Japanese boundary pitch movements for speech synthesis.” Proc. of the 3rd ESCA Workshop on Speech Synthesis.

前川喜久雄, 菊池英明, 五十嵐陽介 (2001). 「X-JToBI: 自発音声の韻律ラベリングスキーム」, 電子情報通信学会技術報告(NLC2001-71, SP2001-106), pp.25-30.

前川喜久雄 (2011). 「コーパスを利用した自発音声の研究」, 東京工業大学大学院博士論文.
CSJ(2011). 「日本語話し言葉コーパス」, 国立国語研究所, http://www.ninjal.ac.jp/csj/
T. Miyajima, H. Kikuchi, K. Shirai (2011). "Collection and analysis of emotional speech focused on the psychological and acoustical diversity", Proc. of ICPhS2011, pp.1394-1397.



(a)SPS (b)APS



(c) D (d) R

図 2 CSJ における句末モーラの F0 変動

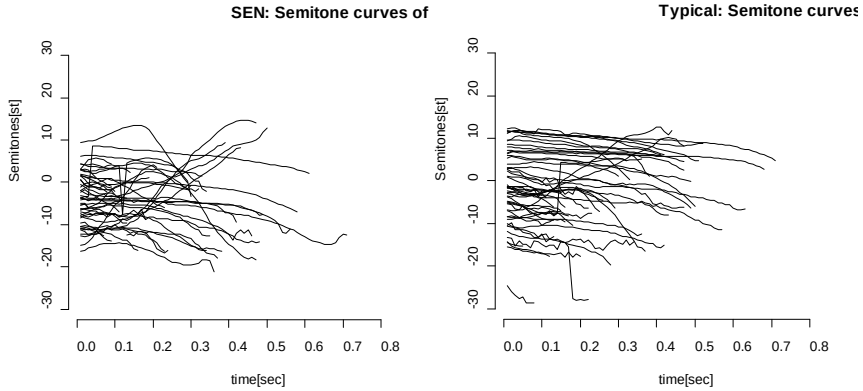


図 3 SEN における句末モーラの F0 変動

(左は SEN の 50 音声、右は典型的な感情指示語に対する音声表現 50 個)

『日本語話し言葉コーパス』における句末境界音調のピッチレンジ制御

五十嵐 陽介（広島大学）†
小磯 花絵（国立国語研究所理論・構造研究系）‡

Pitch Range Control of Boundary Pitch Movements in the Corpus of Spontaneous Japanese

Yosuke Igarashi (Hiroshima University)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

１．はじめに

韻律句（prosodic phrase）末尾に生じる音調で発話の語用論的解釈（質問、継続、強調など）に貢献する音調を句末境界音調（Boundary Pitch Movement, BPM）という。どのようなBPMがいくつあるのかに関する研究や、BPMの機能に関する研究は古くからなされているが（金田一 1951, 大石 1959, 川上 1963, 郡 1997, Venditti, Maeda, and van Santen 1998）、BPMのピッチレンジ（pitch range）[1]に関する研究は管見の及ぶ限りほとんどなされていない。

日本語の韻律句のピッチレンジを決定する主要な要因のひとつとして、アクセント句のピッチレンジを縮小させるダウンステップ（downstep）と呼ばれる現象が知られている（Pierrehumbert and Beckman 1988）。BPMにダウンステップは観察されるのであろうか。それともダウンステップの効果はアクセント句の主要部（BPMを除いた部分）に限定されるのであろうか。もしBPMにダウンステップが観察されないのであれば、BPMのピッチレンジ制御と発話の他の部分のピッチレンジ制御は、ある程度独立していることとなる。その場合、発話のピッチレンジ制御を扱う従来のモデルでは、BPMのピッチレンジを扱えないことになる。

本研究は日本語のBPMのピッチレンジ制御を検討するものであるが、その目的は『日本語話し言葉コーパス』（前川他 1998, 以降 CSJ）の分析を通じて、BPMにダウンステップが観察されるかを明らかにすることにある。第2節ではBPMの記述のために必要な諸概念を導入するとともに本研究が取り組む課題を明確にする。第 3 節では用いたデータを記述する。第4節では分析結果を報告し、第5節で結果の考察を行う。第6節で結論を述べる。

２．日本語の韻律構造の記述

２.１　ダウンステップ

日本語におけるダウンステップとは、アクセント核（lexical pitch accent）が、それを含むアクセント句に後続するアクセント句の基本周波数（F0）ピーク（ピッチレンジの上限）を、反復的（iterative）に低下させる現象である（Pierrehumbert and Beckman 1988）。図1は「旨い飴がありました」（左）と「旨い豆がありました」（右）の音声波形と基本周波数（F0）曲線を示したものである。双方の発話とも、最初のアクセント句（ウマ┐イ）はアクセント核を持つ有核句である。そのため2番目のアクセント句（アメガ／マメ┐ガ）にダウンステップが生じ、F0 ピークが低下する。一方、2 番目のアクセント句は、左の発話ではアクセント核を持たない無核句（アメガ）であるのに対して、右の発話では有核句（マメ┐ガ）である。したがって、3 番目のアクセント句（アリマ┐シタ）にダウンステップが観察されるのは、右側の発話のみとなる。

† igarashi@hiroshima-u.ac.jp, ‡ koiso@ninjal.ac.jp

[1] 本研究ではピッチレンジを特定の時間区間（例えばBPMの開始時刻から終了時刻）における基本周波数の最大値と最小値の差分と定義する。

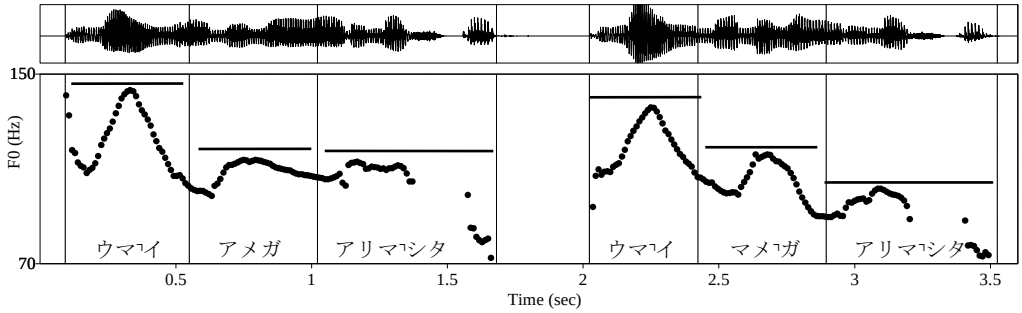


図 1　ダウンステップ.

左の発話は「旨い飴がありました」、右の発話は「旨い豆がありました」。縦の点線はアクセント句境界を表す。各アクセント句のピッチレンジの上限を水平方向の実線で示している。発話者は第 1 筆者。

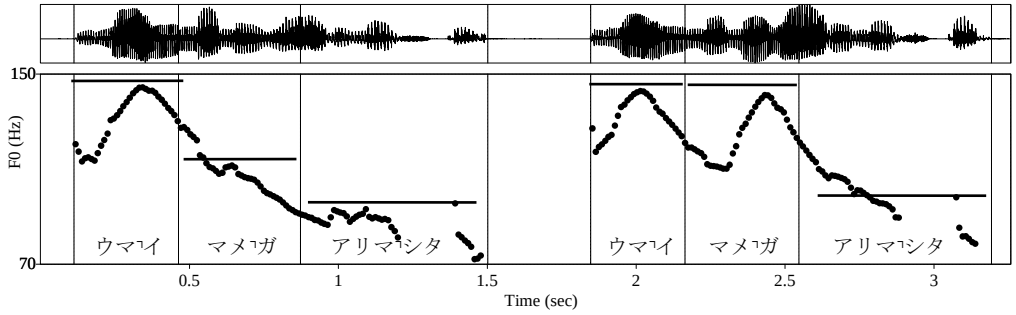


図 2　フォーカスによるダウンステップの阻止.

発話は左右とも「旨い飴がありました」。左の発話は「旨い」に、右の発話は「豆」にフォーカスが置かれている。縦の点線はアクセント句境界。水平方向の実線はピッチレンジの上限。発話者は第 1 筆者。

ダウンステップは、アクセント句の F0 ピークを低下させるが、F0 ボトム（ピッチレンジの下限）はあまり低下させない（前川 1998）。このため、アクセント句のピッチレンジ（F0 ピークと F0 ボトムの差分）は、有核句が先行することにより反復的に縮小することになる。以降、F0 ピークを低下させるだけでなく、ピッチレンジを縮小させるダウンステップを「狭義のダウンステップ」と呼ぶことにする。

２.２　イントネーション句

ダウンステップの効果は、一定の統語構造やフォーカスによって阻止されることが知られている（Pierrehumbert and Beckman 1988; Venditti et al. 2008）。図 2 は「旨い豆がありました」の F0 曲線を示したものであるが、左の発話は「旨い」にフォーカスが置かれており、発話は「豆」にフォーカスが置かれている。左側の発話では 2 番目以降のアクセント句にダウンステップが観察される。フォーカスは後続要素のピッチレンジをさらに縮小させる効果があるため（post-focal prosodic subordination, cf. Venditti et al. 2008）、図 1（右）の発話と比較して、ピッチレンジ縮小の程度がより顕著になっている。

一方図 2（左）では、フォーカスを受けた語「豆」を含むアクセント句（マメ┐ガ）の F0 ピークは、それに先行するアクセント句（ウマ┐イ）の F0 ピークとほぼ同水準となっており、ダウンステップが観察されない。この現象を記述するために、Pierrehumbert and Beckman (1988)の韻律理論では、アクセント句（マメ┐ガ）の始端に、アクセント句より階層的に上位の韻律句の境界を仮定する。この韻律句は、CSJ が採用している日本語の韻律ラベリング体系である X-JToBI（Maekawa et al. 2002）およびその前身である J_ToBI（Venditti 1995, 2005）では、イントネーション句（intonation phrase）と呼ばれおり、ダウンステップの生じる領域、あるいはピッチレンジが指定される領域として定義される。X-JToBI では、図 2 の発話に図 3 に示す韻律階層が仮定される。

図 3　図 1 の発話の韻律階層.

左側は「旨い」にフォーカスが置かれた発話、右側は「豆」にフォーカスが置かれた発話。AP はアクセント句を IP はイントネーション句を表す。AP より下位の韻律単位は省略してある。

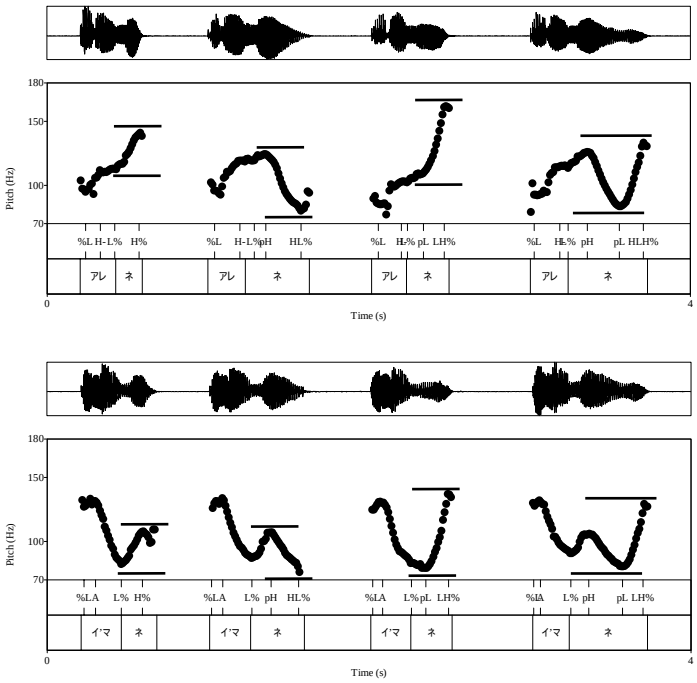


図 4　BPM.

上は無核句（アレネ）の句末に BPM を伴う発話（4 発話）、下は有核句（イ┐マネ）の句末に BPM を伴う発話（4 発話）である。BPM のタイプは上下ともに左から H%, HL%, LH%, HLH%である。水平方向の実線によってピッチレンジの上限と下限を示している。発話者は第 1 筆者。

２.３　句末境界音調（BPM）

日本語にどのような BPM がいくつあるかに関する研究は数多くあるが、見解の一致は得られていない（金田一 1951, 大石 1959, 川上 1963, 郡 1997, Venditti, Maeda, and van Santen 1998）。日本語の BPM の種類と数を確立することは重要な研究課題であるが、本研究ではこれに取り組まない。本研究では X-JToBI の枠組みに基づいて BPM を記述する。

X-JToBI では主要な BPM として図 4 に示す 4 種類が認められている。H%（上昇調 1）は F0 が単純に上昇するタイプの BPM であり、HL%（上昇下降調）は上昇の後に下降が生じるタイプの BPM である。LH%（上昇調 2）は、F0 が上昇する点は H%と同様であるが、上昇の前に低い F0 が一定時間継続する点が異なる。HLH%（上昇下降上昇調）は、上昇の後下降が生じ、その後さらに上昇が生じるタイプの BPM である（五十嵐他 2006）。

BPM は複数の F0 屈曲点によって特徴づけられるため BPM の物理的実現の分節音列上の生起時刻と F0 値を記述するためには、複数のトーンラベルが必要となる。この目的のために X-JToBI では、BPM のトーンラベルを分解し、複数の屈曲点の生起位置を記述している。この際、トーンラベルを単に分解しただけでは検索に支障をきたすので、分解されたラベ

ルの一部には補助記号を付与することにしている。4 種類の BPM それぞれの記述に用いられるトーンラベルと、そのラベルが記述する物理的なイベントは表 1 に要約されている。トーンラベルを用いた BPM のラベリング例は図 4 に示されている。

表 1 BPM ラベル

BPM タイプ	用いられるトーンラベルとそれが記述する物理的イベント
H%	L%（上昇開始点），H%（上昇終了点）
HL%	L%（上昇開始点），pH（下降開始点）， HL%（下降終了点）
LH%	L%（低 F0 区間開始点），pL（上昇開始点），LH%（上昇終了点）
HLH%	L%（上昇開始点），pH（下降開始点），pL（上昇開始点），LH%（上昇終了点）

２.４　BPM のピッチレンジ

BPM にダウンステップは観察されるのであろうか。Pierrehumbert and Beckman (1988)には BPM のダウンステップに関する記述がある。彼らは、発話末に BPM（H%）を伴う「もうちょっと右側が上げられる？」（モーチョㇰット ミギガワガ アゲラレル）と「もうちょっと右側が下げられる？」（モーチョㇰット ミギガワガ サゲラレㇽ）を比較しており、最終アクセント句にアクセント核を伴う後者（サゲラレㇽ）は、核を伴わない前者（アゲラレル）より BPM の F0 ピークが低くなることを報告している。したがって、彼らの分析結果に従えば、BPM にもダウンステップが生じることになるが、分析対象となった発話が限定されているため、この結果が自発音声を含めた日本語一般に敷衍できるかは不明である。また、BPM の F0 ボトムについては詳述されておらず、アクセント核が先行することによってピッチレンジが縮小されるかどうか、すなわち狭義のダウンステップが観察されるかも不明である。

図 4 は第 1 筆者によって発話された 4 種類の BPM であるが、この特定の発話を検討する限り、BPM にアクセント核が先行しない場合（上図）と比較して、BPM にアクセント核が先行する場合（上図）は、BPM の F0 ピークがはるかに低くなっている。しかしながら、BPM にアクセント核が先行する場合は、HL%を除いて、F0 ピークが低下するとともに F0 ボトムも顕著に低下しており、結果として BPM のピッチレンジの縮小はほとんど観察されない。すなわち、HL%の場合を除いて、BPM には狭義のダウンステップは観察されないようである。このことは、BPM はダウンステップに関してアクセント句の主要部とは異なるふるまいをすることを示唆している。一方 HL%のピッチレンジは、図 4 を見る限り、アクセント核が先行すると顕著に縮小するようであり、狭義のダウンステップが観察されるようである。その理由は、アクセント核が先行する場合は F0 ピークが低下する一方で、アクセント核が先行しない場合の F0 ボトムとアクセント核が先行しない場合の F0 ボトムとがほぼ同水準にとどまる（等しく低い）ことにありそうである。以上の観察から、BPM に狭義のダウンステップが観察されるか否かは BPM のタイプによって異なることが示唆される。

３．データと計測

データとして使用したのは CSJ Core に含まれる学会講演 70 ファイル（19 時間）と模擬講演 107 ファイル（20 時間）である。以下、用いたデータと計測方法ついて詳述する。

３.１　分析対象としたアクセント句と BPM

対象としたアクセント句は、BPM を持つイントネーション句に含まれるアクセント句である。アクセント句主要部の諸特徴は、先行アクセント核数と当該アクセント句の有核無核ごとに集計した。先行アクセント核数が 5 以上のアクセント句は分析対象から除外した。その結果、表 2 に内訳を示した合計 32360 のアクセント句が分析対象となった。

BPM は先行アクセント句の数と BPM のタイプごとに集計した。先行アクセント核数が 6

以上のものは除外した。アクセント句の次末モーラ以前から上昇を開始する BPM の変種 も対象から除外した。H%には、F0 上昇終了後に同水準の F0 値が持続される変種が観察されるが、今回の分析ではこの変種と通常の変種との区別はせず、双方とも H%として集計した。HLH%は数が少ないので（7 件）除外した。先行アクセント核数が 4 以上である LH%の件数は 0 であった。その結果、表 3 に内訳を示した合計 32353 の BPM が分析対象となった。

表 2 分析対象としたアクセント句のクロス統計表

当該アクセント句の有核無核	先行アクセント核数					
	0	1	2	3	4	計
無核	4750	3157	956	201	36	9100
有核	13017	7537	2218	415	73	23260
計	17767	10694	3174	616	109	32360

表 3 分析対象とした BPM のクロス統計表

BPM タイプ	先行アクセント核数						
	0	1	2	3	4	5	計
H%	3759	12178	6232	1779	321	46	24315
HL%	959	3816	2179	611	130	27	7722
LH%	32	176	80	28	0	0	316
計	4750	16170	8491	2418	451	73	32353

３.２　先行アクセント核数の計算法

F0 ピーク、F0 ボトム、ピッチレンジは、先行するアクセント核の数別に集計した。アクセント句主要部の場合、当該アクセント句より前に位置するアクセント句で、かつ、当該アクセント句が所属するイントネーション句と同一のイントネーション句に所属するアクセント句に存在するアクセント核の数の合計を、先行アクセント核数とした。

BPM の場合の先行アクセント核数は、アクセント句主要部と同様の方法で計算したのち数に、BPM が所属するアクセント句の有するアクセント核を加えたものとした。つまり BPM を伴うアクセント句にアクセント核が存在する場合は、その句に先行するアクセント句が持つアクセント核の数にさらに 1 をプラスした数が、先行アクセント核数となる。

３.３　F0 値の抽出

F0 値は CSJ の XML ファイルに記録されている F0 値（Hz）を利用した。原則として X-JToBI のトーンラベルひとつにひとつの F0 値が与えられているが、母音の無声化等の理由により F0 値が与えられていない場合もある。このような欠損値は分析対象から除外した。性差や個人差の影響を最小限にするために、F0 値は談話ごとに Z スコアに変換した。

アクセント句主要部（BPM を除いた部分）のピッチレンジに関する F0 値は以下の方法で抽出した。F0 ピークの値としては、当該アクセント句に属するラベル[H-]あるいは[A]の持つ値を採用した。当該アクセント句に[H-]と[A]の双方が存在する場合は、ふたつのラベルの持つ値を比較して、より高い値を採用した。当該アクセント句内に[H-]と[A]の双方ともが存在しない場合は、[%L]および[L%]の持つ値を比較して、より高い値を採用した。一方、F0 ボトムの値として採用したのは、[%L]の値と[L%]の値を比較してより低いと判断された値である。

BPM のピッチレンジに関する F0 値の抽出方法は、BPM タイプごとに異なる。まず F0 ピークであるが、H%の場合は[H%]の持つ値を、HL%の場合は[pH]の持つ値を、LH%の場合は[LH%]の持つ値を採用した。一方 F0 ボトムの値として採用したのは、H%と LH%の場合は BPM の直前の[L%]の値を採用し、HL%の場合は[HL%]の値を採用した。

４．分析

４.１　アクセント句主要部のダウンステップ

はじめに、アクセント句の主要部にダウンステップが観察されることを確認するための分析をおこなった。図 5〜図 7 は、先行要素（当該アクセント句に先行するアクセント句）のアクセント核数ごとにみた、当該アクセント句の F0 ピーク、F0 ボトム、ピッチレンジの値を表す箱髭図[2]である。左が当該アクセント句が無核句の場合、右が有核句の場合である。

F0 ピークの中央値（図 5）は先行要素のアクセント核の数が増加するにしたがって低下している。一方 F0 ボトム（図 6）には、先行要素のアクセント核の数が増加するにしたがって低下する傾向があるが、F0 ピークの場合と比較するとその程度が小さい。ピッチレンジ（図 7）は、先行するアクセント核の数が増加するにしたがって、一貫して縮小している。

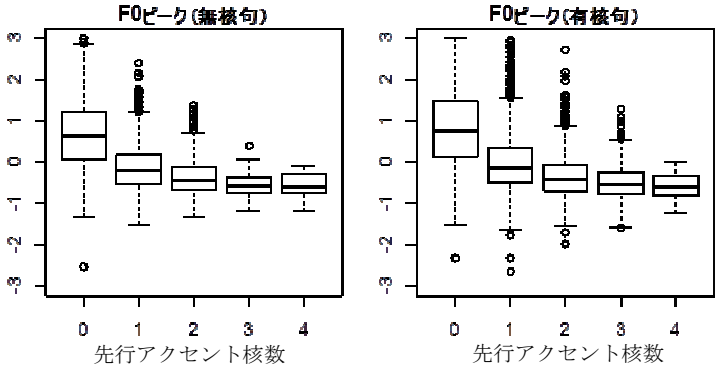


図 5　アクセント句主要部の F0 ピーク.

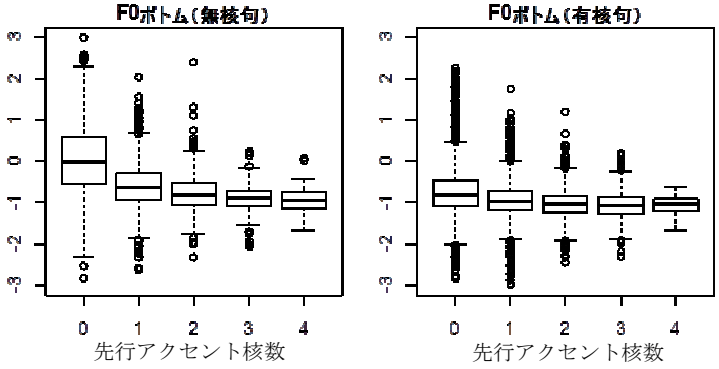


図 6　アクセント句主要部の F0 ボトム.

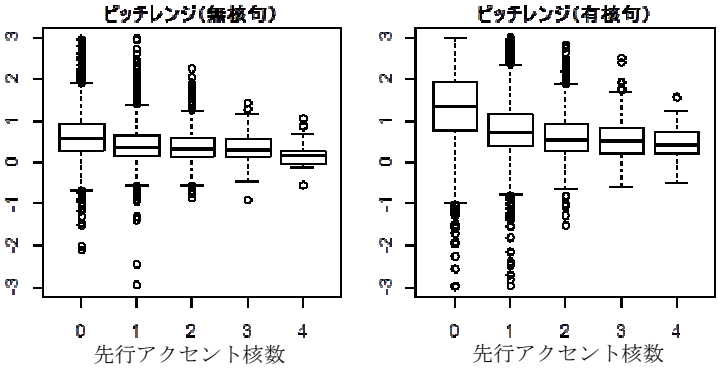


図 7　アクセント句主要部のピッチレンジ.

[2] それぞれの平均値は付表 A1-A3 に示す。

以上の結果から、先行するアクセント核はアクセント句の F0 ピークを低下させるが、F0 ボトムはあまり低下させず、その結果ピッチレンジが縮小する狭義のダウンステップが、アクセント句主要部に生じることが確認された。これは従来の報告の通りである。

4.2 BPM のダウンステップ

次に、BPM にダウンステップが観察されるかどうかを BPM タイプごとに検討した。

図 8〜図 10 は、先行アクセント核数ごとにみた、当該アクセント句の F0 ピーク、F0 ボトム、ピッチレンジの値を表す箱髭図[3]である。

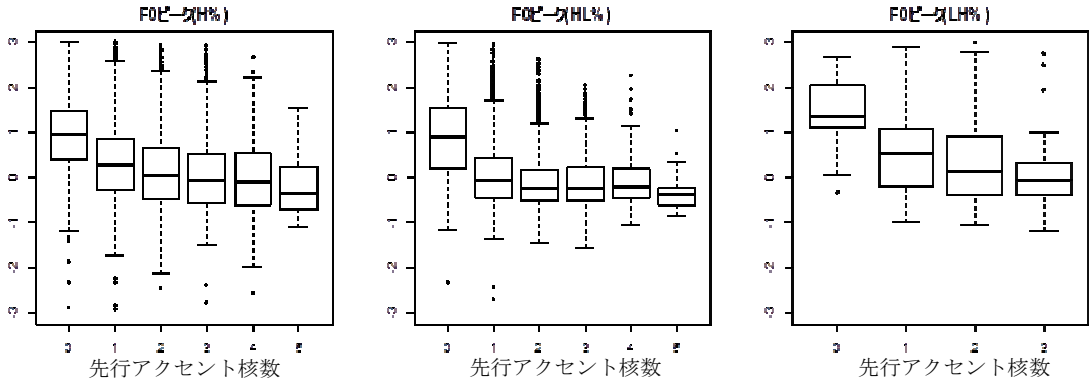


図 8　BPM の F0 ピーク.

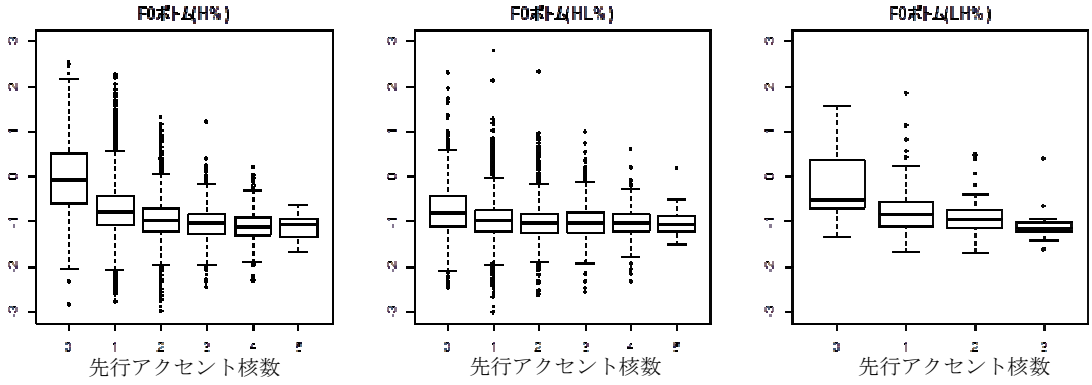


図 9　BPM の F0 ボトム.

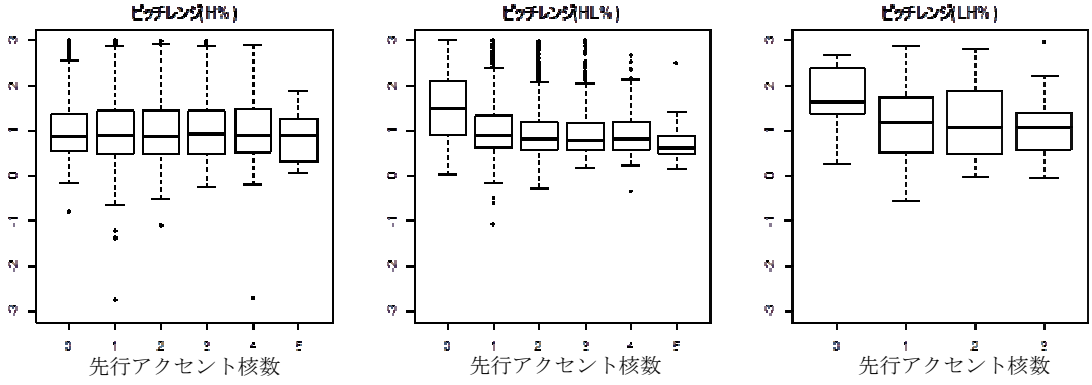


図 10　BPM の F0 ボトム.

[3] それぞれの平均値は付表 A4-A5 に示す。

まず F0 ピーク（図 8）であるが、H%に関する限り、先行アクセント数が増えるにしたがって F0 ピークがほぼ一貫して低下していることがわかる。一方 HL%は H%とは明らかに異なる傾向を示している。HL%の F0 ピークは、アクセント核が 0 の場合と 1 以上の場合との間に差が見られることを除いては、ほぼ同水準にとどまっている。LH%は H%と似たパタンを示している。

次に F0 ボトム（図 9）であるが、H%の F0 ボトムは、ピークの場合と同じように、先行アクセント数が増えるにつれてほぼ一貫して低下している。低下の程度もピークの場合と同程度である。一方 HL%の F0 ボトムは、アクセント核が 0 の場合と 1 以上の場合との間にわずかな差が見られることを除いては、ほぼ一定の値にとどまっている。ここでもまた、HL%と H%との間に明確に異なる傾向が観察できる。LH%は H%と似たパタンを示している。

最後にピッチレンジ（図 10）を検討しよう。H%のピッチレンジは先行アクセント核数に関わらずほぼ一定であることがわかる。この BPM に関する限り狭義のダウンステップは観察されない。一方 HL%のピッチレンジは、アクセント核が 0 の場合と 1 以上の場合との間に差が見られることを除いては、ほぼ一定にとどまっている。言い換えると、この BPM のピッチレンジはひとつ以上のアクセント核が先行するときに縮小するが、アクセント核の数が増えるにしたがって反復的に縮小する現象は観察されない。HL%に観察されるピッチレンジの縮小は、アクセント句主要部に観察されるダウンステップとは性質が異なるものと言える。LH%のピッチレンジもまた HL%と類似したパタンを示している。

５．考察

今回の分析結果によって、H%のピッチレンジは、先行アクセント核の数に関わらず、一定に保たれることが示唆された。したがって、この BPM には狭義のダウンステップは生じないとみなすことができそうである。H%の F0 ピークは先行アクセント核の数にしたがって反復的に低下するが、同様の反復的低下は F0 ボトムにも観察される。H%の F0 ボトムは、この BPM が所属するアクセント句の F0 ボトムと基本的に一致することを考慮すると、BPM のボトムに観察される反復的低下は、先行アクセント核が BPM に与えた効果ではなく、先行アクセント核が BPM の所属するアクセント句主要部の F0 ボトムに与えた効果とみなすことができる。H%のピッチレンジ（上昇幅）が、今回の結果が示唆するように一定であると仮定すると、H%の F0 ピークに観察される反復的低下は、アクセント句主要部の F0 ボトムが反復的に低下した結果に過ぎないとみなすことができる。以上から、H%には広義のダウンステップさえも観察されないと結論することができるであろう。

一方、先行アクセント核が HL%のピッチレンジに与える効果は、H%に与えるものとは異なることが示唆された。H%とは異なり、HL%のピッチレンジは、先行アクセント核の効果を受けて縮小する。しかしながらその縮小のパタンは、アクセント核の数が増えるにしたがって反復的に縮小するようなパタン（アクセント句主要部に観察されるパタン）ではない。今回の分析結果によれば、この BPM の F0 ピークは、先行アクセント核数がひとつ以上の場合低下するが、数がひとつ以上であれば、その数に関わらず一定となる。一方 F0 ボトムは、先行アクセント核数が 0 の場合わずかに高めであることを除けば、先行アクセント核数に関わらずほぼ一定となる。したがって HL%のピッチレンジは、アクセント核が先行しない場合（大きい）とアクセント核がひとつ以上先行する場合（小さい）との間で差が観察されることになる。HL%のピッチレンジの縮小は、先行アクセント核によってもたらされる点でアクセント句主要部のダウンステップに類似するが、アクセント核の数が増えるにつれて反復的に縮小するものではない点で異なる。したがって、アクセント句主要部に観察されるものと同種のダウンステップは、HL%には観察されないと結論付けることができるだろう。HL%のピッチレンジ、とりわけ F0 ピークに対する先行アクセント核の効果がどのようなものかを明らかにするのは今後の課題である。

LH%のピッチレンジ制御は、F0 ピークとボトムのふるまいは H%に類似するが、ピッチレンジのふるまいは HL%に類似するというものであった。しかしながら、分析対象となっ

たこの BPM の数が他の BPM と比較してはるかに少ないので、この結果の解釈には注意が必要である。

６．結論

本研究は『日本語話し言葉コーパス』の分析に基づいて、句末境界音調（BPM）にダウンステップが観察されるかを検討した。その結果、BPM には、少なくともアクセント句主要部に観察されるものと同種のダウンステップは、観察されないことが明らかになった。また、BPM のピッチレンジ制御は BPM のタイプごとに異なることが示唆された。

日本語のピッチレンジを取り扱う従来のモデルは、アクセント句の主要部の分析に基づいて提案されたものであり、BPM のピッチレンジ制御はこれまでほとんど検討されてこなかった。本研究の分析結果が示唆するように、BPM はアクセント句主要部とは異なるピッチレンジ制御が行われている。現行のモデルでは BPM のピッチレンジ制御を扱うことができない。今後は、BPM をも組み込んだピッチレンジの理論を構築することが必要である。

謝　辞

本研究は、国立国語研究所（言語資源研究系）基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）および萌芽・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）による補助、および文部科学省科学研究費補助金・若手研究（B）「イントネーションの音韻論と音声学を峻別する実験手法の確立」（研究代表者：五十嵐陽介）による補助を得ています。

文　献

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集 124：日本語話し言葉コーパスの構築法』，347-453，国立国語研究所

大石初太郎（1959）「プロミネンスについて―東京語の観察に基づく覚え書―」『国立国語研究所論集 I：言葉の研究』，87-102，国立国語研究所

川上蓁（1963）「文末などの上昇調について」『国語研究』16, 25-46.

金田一春彦（1951）「コトバの旋律」『国語学』5, 37-59.

郡史郎（1997）「日本語のイントネーション－型と機能」杉藤美代子(監), 国広哲弥他(編)『日本語音声 2：アクセント・イントネーション・リズムとポーズ』，169-202，三省堂

前川喜久雄（1998）「音声学」『岩波講座言語の科学 2：音声』, 1-52, 岩波書店

前川喜久雄、籠宮隆之、小磯花絵、小椋秀樹、菊池英明（2000）「日本語話し言葉コーパスの設計」『音声研究』4(2), 51-61.

Maekawa, K., H. Kikuchi, Y. Igarashi and J. Venditti (2002) X-JToBI: An extended J_ToBI for spontaneous speech, *Proceedings of the 7th International Conference on Spoken Language Processing*, 1545-1548, Denver, Colorado.

Pierrehumbert J. and M. Beckman (1988) *Japanese Tone Structure*, Cambridge: MIT press.

Venditti, J. (1995) Japanese ToBI labelling guidelines, ms Ohio State University. (Also printed in 1997: In: K. Ainsworth-Darnell and M. D'Imperio (eds.) *Papers from the Linguistics Laboratory. Ohio State University Working Papers in Linguistics* 50, 127-162).

Venditti, Jennifer J., Kazuaki Maeda, and Jan P. H. van Santen (1998) Modeling Japanese boundary pitch movements for speech synthesis. In M. Edgington (ed.) *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 317-322, Jenolan Caves, Australia.

Venditti, J. (2005) The J_ToBI model of Japanese intonation, In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*.172-200, New York: Oxford Univ. Press.

Venditti, J., K. Maekawa, and M.E. Beckman (2008). 'Prominence marking in the Japanese intonation system', in S. Miyagawa, and M. Saito (eds.), *Handbook of Japanese Linguistics*, 456-512, New York: Oxford University Press.

付 表

A1 アクセント句主要部の F0 ピークの平均値

当該アクセント句の有核無核	先行アクセント核数				
	0	1	2	3	4
無核	0.664	-0.135	-0.358	-0.519	-0.553
有核	0.884	-0.039	-0.346	-0.476	-0.574

A2 アクセント句主要部の F0 ボトムの平均値

当該アクセント句の有核無核	先行アクセント核数				
	0	1	2	3	4
無核	0.023	-0.599	-0.753	-0.886	-0.910
有核	-0.750	-0.949	-1.046	-1.065	-1.087

A3 アクセント句主要部のピッチレンジの平均値

当該アクセント句の有核無核	先行アクセント核数				
	0	1	2	3	4
無核	0.685	0.460	0.377	0.363	0.196
有核	1.542	0.836	0.644	0.540	0.487

A4 BPM の F0 ピークの平均値

BPM タイプ	先行アクセント核数					
	0	1	2	3	4	5
H%	0.993	0.346	0.169	0.063	0.023	-0.202
HL%	0.909	0.082	-0.077	-0.071	-0.059	-0.340
LH%	2.080	0.597	0.416	0.192		

A5 BPM の F0 ボトムの平均値

BPM タイプ	先行アクセント核数					
	0	1	2	3	4	5
H%	-0.034	-0.738	-0.946	-1.063	-1.080	-1.104
HL%	-0.723	-0.975	-1.022	-1.005	-1.028	-1.094
LH%	-0.325	-0.773	-0.991	-1.050		

A6 BPM のピッチレンジの平均値

BPM タイプ	先行アクセント核数					
	0	1	2	3	4	5
H%	1.050	1.094	1.108	1.147	1.105	0.844
HL%	1.640	1.071	0.959	0.954	0.993	0.764
LH%	2.350	1.336	1.554	1.220		

『現代日本語書き言葉均衡コーパス』における複合辞の検出と評価

鈴木敬文　(筑波大学大学院システム情報工学研究科)
阿部佑亮　(筑波大学大学院システム情報工学研究科)
宇津呂武仁　(筑波大学システム情報系)*
松吉俊　(山梨大学大学院医学工学総合研究部)
土屋雅稔　(豊橋技術科学大学情報メディア基盤センター)

Detection of Compound Functional Expressions in "Balanced Corpus of Contemporary Written Japanese" and its Evaluation

Takafumi Suzuki　(University of Tsukuba)
Yusuke Abe　(University of Tsukuba)
Takehito Utsuro　(University of Tsukuba)
Suguru Matsuyoshi　(University of Yamanashi)
Masatoshi Tsuchiya　(Toyohashi University of Technology)

1. はじめに

機能表現[1]とは,「にあたって」や「をめぐって」のように，2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現である．一方，この機能表現に対して，それと同一表記をとり，内容的な意味をもつ表現が存在することがある．例えば，文 (1) と文 (2) には「にあたって」という表記の表現が共通して現れている．

(1) 出発する にあたって，荷物をチェックした．

(2) ボールは壁 にあたって，跳ね返った．

文 (1) では，下線部はひとかたまりとなって,「機会が来たのに当面して」という機能的な意味で用いられている．それに対して，文 (2) では，下線部に含まれている動詞「あたる」は，動詞「あたる」本来の内容的な意味で用いられている．このような表現においては，機能的な意味で用いられている場合と，内容的な意味で用いられている場合とを識別する必要がある．

我々は，このような日本語機能表現の解析の課題に対して，これまでに，国立国語研「現代語複合辞用例集」 [国研 01] に収録されている 125 機能表現の異表記を展開した 300 表現について，新聞記事中の用例に対して機能的用法・内容的用法を判別した用例データベース [土屋 06] を作成・公開した．また，機能的用法・内容的用法の自動判別ツールを作成し，係り受け解析ツールとの統合により，複合辞としての機能的用法を考慮した係り受け解析を実現した [注連 07]．また，日本語機能表現の全表記を網羅した辞書として，日本語機能表現の全表記約 17,000 を網羅的に収録した「つつじ」 [松吉 07, 松吉 08][2] が公開されたのを受けて，日本語機能表現の全表記約 17,000 を網羅的に収録した辞書「つつじ」の階層的構造および言語学的特性を活用して，全 17,000 表記を対象とした網

*utsuro @ iit.tsukuba.ac.jp

[1]機能表現は，複数形態素からなる複合辞と一つの形態素からなる機能語から構成される．本論文において対象とする機能表現は，いずれも複数形態素から構成される複合辞に相当するため，本論文においては，複合辞と同等の意味で機能表現という用語を用いる．また，本論文では，特に，機能表現を構成する表記が，複合辞として用いられる機能的用法となる場合と，複合辞を構成する形態素が本来の意味で用いられている内容的用法となる場合の曖昧性を持つ場合に注目する．

[2]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 「(1) 複合辞として一長単位を構成する」または「(2) 複数の長単位から構成される」という曖昧性を持つ短単位列の例

	表記	短単位列	長単位	意味	例文
(1)	に当たって	に (助詞) +当たっ(動詞) +て (助詞)	に当たって (助詞)	状況	証券化 **に当たって**、より有利な 商品設計が可能である点など でメリットがあると考えられる。
(2)			に (助詞) +当たっ(動詞) +て (助詞)	-	… 紫外線 **に　当たっ　て** も 分解されず、においの 成分が長もちするんです」
(1)	ことがあっ	こと (名詞) +が (助詞) +あっ(動詞)	ことがあっ (助動詞)	経験	誰かに似ているな、と 彼はこれまでにも時折 思った **ことがあっ** たが、…
(2)			こと (名詞) +が (助詞) +あっ(動詞)	-	どんな人にも … 、得意なことと 苦手な **こと　が　あっ** て、…
(1)	ところが	ところ (名詞) +が (助詞)	ところが (接続詞)	逆接	**ところが**、それは意外に素早く 簡単に済んだのだった。
(2)			ところ (名詞) +が (助詞)	-	… 部屋の北側に一段高い **ところ　が** あって、…

羅的な日本語機能表現表記の用法判定 [鈴木 12]，および，日本語機能表現の集約的翻訳の枠組み [島内 10, Nagasaka10, 阿部 12] を提案した．

ここで，これらの研究のうち，日本語機能表現表記の機能的用法・内容的用法の分析および自動判定手法の研究 [土屋 06, 注連 07, 鈴木 12] は，いずれも，新聞記事という限定されたジャンルのコーパスを対象としたものであった．そこで，本研究では，大規模な均衡コーパスである『現代日本語書き言葉均衡コーパス』において，上述した機能的用法・内容的用法の曖昧性を持つ機能表現表記を対象として，機械学習により用法判定を行う手法を適用し，その性能を評価した結果を報告する．具体的には,『現代日本語書き言葉均衡コーパス』コアデータ [BCCWJ 総括班 09] を対象として，複合辞となり得る表記 (機能表現表記) を構成する短単位列が,「全体として 1 つの機能的長単位」となるのか，それとも,「複数の長単位から構成される列」となるのかという曖昧性を解消することを目的とする．適用する手法としては，条件付き確率場 (Conditional Random Fields, CRF) [Lafferty01] を利用したチャンキングを用い，ツールとしては CRF++[3]を用いた．評価実験の結果，機能的な長単位の検出において 97%近い F 値を達成した．

2. 『現代日本語書き言葉均衡コーパス』における複合辞

2.1 複合辞の短単位列・長単位の分析

本研究ではまず,『現代日本語書き言葉均衡コーパス』(以下，BCCWJ) コアデータ [BCCWJ 総括班 09] を用いて，複合辞となり得る表記 (機能表現表記) を構成する短単位列について，以下の曖昧性の有無を調査した．

1. 短単位列が全体として一つの機能的長単位 (=複合辞) を構成する．

2. 一つの短単位が一つの長単位を構成する．

[3] http://crfpp.sourceforge.net/

表 1 にこれらの曖昧性の例を示す．例えば，表中の「ところが」が，機能的な長単位となる場合は，非構成的に「逆接」の意味で用いられており，その品詞は接続詞である．一方，複数の長単位から構成される長単位列となる場合は,「ところ」は場所を表す名詞,「が」は格助詞として用いられている．

2.2 検出対象の複合辞の選定

前節を踏まえて，本節では，検出対象となる機能表現表記 (複合辞となり得る短単位列) の選定手順およびその結果について述べる．

BCCWJ においては，[小椋 11] において，助詞相当 75 語，助動詞相当 55 語の複合辞が収録されているが，これらは，

1. [国研 01, グループ・ジャマシイ 98, 森田 89] における収録状況に基づき，重要度を判断．
2. BCCWJ において一定以上の頻度 (50 程度) がある．
3. BCCWJ において，機能的用法の割合が 80%程度である．
4. 複合辞前後の短単位から，機能的用法であると判定できる．

という条件のもと選定されたものである．

一方，本論文では，前節で述べたように,「複合辞として一長単位を構成する」，もしくは,「複数の長単位から構成される」の曖昧性を持つ短単位列に対して，その曖昧性の解消を目的とする．そこで，まず，以下の手順により，検出対象とする短単位列 (機能表現表記) を選定した．

1. BCCWJ コアデータを対象として，品詞が助詞，助動詞，接続詞[4]となる長単位を列挙することにより，1,010 種類の長単位が得られた．
2. 文字長が一文字である長単位，78 種類を除外し，932 種類となった．
3. 口語調の崩れた日本語や誤字，古語，方言など，合計 213 種類を人手で除外し，719 種類となった．
4. 719 種類の長単位に対して，短単位列の種類数は 727 種類となった．これらの短単位列のうち,「複合辞として一長単位を構成する」，もしくは,「複数の長単位から構成される」の曖昧性を持つ短単位列は，201 種類 (助詞 63 種類，助動詞 112 種類，接続詞 26 種類) となった．なお，これらの短単位列のうち，[小椋 11] において選定された複合辞，および，その他連語と重複する短単位列の数は，助詞 60 種類，助動詞 70 種類，接続詞 18 種類の合計 148 種類 ([小椋 11] での単位に従い，表記の揺れ，活用形の違いをまとめた場合は，助詞 43 種類，助動詞 25 種類，接続詞 16 種類の合計 84 種類) である．

上記の手順により除外した長単位の一例を表 2 に示す．

3. 条件付き確率場を用いたチャンキングによる複合辞の検出

3.1 条件付き確率場

本論文では,「複合辞として一長単位を構成する」，もしくは,「複数の長単位から構成される」という曖昧性を持つ短単位列を対象として，複合辞としての長単位を検出する手法として，条件付き確率場 (Conditional Random Fields, CRF) [Lafferty01] を適用する．CRF は正しいラベル系列を他の

[4]「つつじ」においては接続詞型の機能表現が収録されていることをふまえて，本論文でも，接続詞の長単位を検出の対象とする．なお，これらは，[小椋 11] においては、その他連語 86 語として選定されている．

表 2: 除外した長単位 (短単位列) の例

分類	例
長単位としての表記長 1 字	ぬ (助動詞)，た (助動詞)
古語	てゐる (助動詞)，といふ (助詞)
誤字	て ”くる (助動詞)，な　かっ(助動詞)
口語調の崩れた日本語	てくださぁ〜ィ(助動詞)，でぇえええええええす (助動詞)
方言	でっしゃろ (助動詞)，のお (助詞)
長単位としての 曖昧性を持たない	にもかかわらず (助詞)，かも知れない (助動詞)， あるいは (接続詞)

全ラベル系列の候補と弁別する学習を行う．本論文では，CRF による学習・解析用のツールとして CRF++[5]を利用する．正規化項としては，L1 正則化，L2 正則化，MIRA の 3 通りを評価し，最も性能のよかった L1 正則化を採用した．

3.2 チャンキングタグの表現法

本論文では，短単位を最小単位として，検出対象とする機能表現表記を構成する短単位列に対して，共通のチャンクタグを付与するという手順で，機能的な長単位の検出を行う．チャンクタグは，そのチャンクタグが付与された短単位が，検出対象とする機能的な長単位のいずれかに含まれるか否かを表し，チャンクの範囲を示す要素によって表現される．チャンクタグの範囲を示す要素の表現法としては，以下で示す IOB2 フォーマット [Tjong Kim Sang00] を使用する．

I 機能的な長単位 (=複合辞) を表すチャンクに含まれる短単位 (先頭以外)

O 機能的な長単位 (=複合辞) を表すチャンクに含まれない短単位

B 機能的な長単位 (=複合辞) を表すチャンクの先頭の短単位

3.3 素性

学習・解析に用いる素性は，[土屋 07,注連 07] で用いられているものに従う．また，本論文では，[土屋 07,注連 07] で述べている形態素の情報を，全て短単位における該当情報に置き換えるものとする．

文頭から i 番目の短単位 m_i に対して与えられる素性 F_i は，形態素素性 $MF(m_i)$，チャンク素性 $CF(i)$，チャンク文脈素性 $OF(i)$ の 3 つ組として，次式によって定義される．

$$F_i \quad = \quad \langle MF(m_i), CF(i), OF(i) \rangle$$

[土屋 07, 注連 07] では，形態素素性 $MF(m_i)$ は，形態素解析によって形態素 m_i に付与される情報である．IPA 品詞体系の形態素解析用辞書[6]に基づいて動作する形態素解析器 ChaSen[7]による形態素解析結果を入力としているため，以下の 10 種類の情報 (表層系，品詞，品詞細分類 1〜3，活用型，活用形，原形，読み，発音) を形態素素性として用いていた．

[5] http://crfpp.sourceforge.net/
[6] http://sourceforge.jp/projects/ipadic/
[7] http://chasen-legacy.sourceforge.jp/

一方，本論文では，これに対応するものとして，UniDic[8]の品詞体系に従い付与された，BCCWJのコアデータ中の短単位における以下の 10 種類 (書字体，品詞，品詞細分類 1~3，活用型，活用形，語彙素，仮名形，発音形) を利用する．

チャンク素性 $CF(i)$ とチャンク文脈素性 $OF(i)$ は，i 番目の位置に出現している機能表現表記に基づいて定まる素性である．下図の短単位列 $m_j \cdots m_i \cdots m_k$ からなる機能表現候補 E が存在したとする．

$$m_{j-2} \quad m_{j-1} \quad \underbrace{\boxed{m_j \ldots m_i \ldots m_k}}_{\text{機能表現表記 } E} \quad m_{k+1} \quad m_{k+2}$$

チャンク素性 $CF(i)$ は，i 番目の位置に出現している機能表現表記 E を構成している短単位の数 (機能表現表記の長さ) と，機能表現表記中における短単位 m_i の相対的位置の情報の 2 つ組である．チャンク文脈素性 $OF(i)$ は，i 番目の位置に出現している機能表現表記の直前の 2 つの短単位および直後の 2 つの短単位の形態素素性とチャンク素性の組である．すなわち，i 番目の位置に対する $CF(i)$ および $OF(i)$ は次式で表される．

$$\begin{aligned} CF(i) &= \langle k-j+1,\ i-j+1 \rangle \\ OF(i) &= \langle\ MF(m_{j-2}), CF(m_{j-2}), MF(m_{j-1}), CF(m_{j-1}), \\ &\qquad MF(m_{k+1}), CF(m_{k+1}), MF(m_{k+2}), CF(m_{k+2})\ \rangle \end{aligned}$$

素性の詳細な定義については，[土屋 07, 注連 07] を参照されたい．

4. 評価

4.1 評価手順

本節では，評価手法，及び，データセットに関して述べる．本論文では，評価にあたり，BCCWJコアデータ中の 50,693 文のうち，201 種類の機能表現表記 (短単位列) を含む 37,231 文を利用して 10 分割交差検定を行った．表 3 に示すように，評価対象となる個所は，2.2 節で選定した 201 種類の短単位列が出現する個所で，それらの個所に対応する長単位の総数は，48,178 個である．本論文では，これらの長単位の個所を基本単位として，以下で定義する適合率，再現率，F 値を測定し，評価尺度として用いる．

$$\begin{aligned} \text{適合率} &= \frac{\text{検出に成功した長単位数}}{\text{システムによって検出された長単位数}} \\ \text{再現率} &= \frac{\text{検出に成功した長単位数}}{\text{評価データに存在する評価箇所の長単位数}} \\ \text{F 値} &= \frac{2 * \text{再現率} * \text{適合率}}{\text{再現率} + \text{適合率}} \end{aligned}$$

4.2 評価結果

評価結果を表 4 (1) に示す．短単位列全体が一つの長単位となる個所に対しては，97%近い F 値を達成した．また，ベースラインとして，全ての評価対象個所に対して，「短単位列が全体として一つの長単位となる」として判定した場合の評価結果を表 4 (2) に示す．結果として，全ての評価対象個所に対しては，75%程度の F 値にとどまり，提案手法はベースラインに対して，F 値を 20%以上改善した．

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

表 3: 評価箇所に関する統計情報

単位	内訳		総計
	短単位列を構成する 複数の短単位が それぞれ一つの長単位となる	短単位列を構成する 複数の短単位が 1 つの長単位となる	
短単位の数を集計	19,274 (23.3%)	63,349 (76.7%)	82,596 (100%)
長単位の数を集計	19,274 (40.0%)	28,904 (60.0%)	48,178 (100%)

表 4: 評価結果

(1)　CRF によって出力された長単位

類型	適合率 (%)	再現率 (%)	F 値 (%)
タスク 1: 短単位列が全体として 一つの長単位となる個所を検出	96.1	97.5	96.8
タスク 2: 短単位列を構成する短単位それぞれが 一つの長単位となる個所を検出	97.9	90.5	94.0
合計 (タスク 1 + タスク 2)	96.8	94.7	95.7

(2)　ベースライン：評価対象の短単位列を全て「一つの長単位」となると判定

類型	適合率 (%)	再現率 (%)	F 値 (%)
タスク 1: 短単位列が全体として 一つの長単位となる個所を検出	60.0	100	75.0
タスク 2: 短単位列を構成する短単位それぞれが 一つの長単位となる個所を検出	0	0	0
合計 (タスク 1 + タスク 2)	60.0	60.0	60.0

5. 関連研究

[首藤 88, 首藤 98, Shudo04] は，機能表現や慣用表現を含む複数の形態素からなる定型的表現をできるだけ網羅的に収集し，機能表現間に類似度を定義して，機能表現の言い換えや機械翻訳に利用することを提案している．特に，文献 [Shudo04] では，機能表現を検出することを目的として，機能的用法と内容的用法を識別するための規則を人手で作成している．しかし，人手で規則を作成するにはコストがかかるため，網羅できる機能表現の規模には限界がある点が課題であると言える．

一方，我々は，[鈴木 12] において,「つつじ」 [松吉 07] の全 16,801 表現を対象とした方式を提案している．この方式においては,「つつじ」の階層性を利用し，階層において下位に位置する派生的表現の用法判定に際して，用法が類似するより上位の代表的表現の用例を参照することで用法判定を行っている．

また，[松吉 08] においては,「つつじ」中の機能表現を対象として，意味を保存する言い換えが可能な機能表現の分類を規定している．その他，機能表現の検出・係り受け解析等の解析を対象とした研究 [土屋 07, 注連 07, 小早川 09]，内容語と口語的な機能表現を対象として，代表的表現への言い換えを介した機械翻訳の方式 [山本 02] 等が知られている．同様に,「つつじ」 [松吉 07] の機能表現を

対象として，代表的表現への言い換えを介した機械翻訳を行う手法の研究事例として，日本語文型辞典 [グループ・ジャマシイ 98] 中の例文を対象とした集約的英訳 [坂本 09]，特許文を対象とした集約的英訳 [島内 10, Nagasaka10, 阿部 12]，及び集約的中国語訳 [劉 10] についての手法が提案されている．

6. おわりに

本論文では，大規模な均衡コーパスである『現代日本語書き言葉均衡コーパス』において，機能的用法・内容的用法の曖昧性を持つ機能表現表記を対象として，機械学習により用法判定を行う手法を適用し，その性能を評価した結果を報告した．具体的には，『現代日本語書き言葉均衡コーパス』コアデータ [BCCWJ 総括班 09] を対象として，複合辞となり得る表記 (機能表現表記) を構成する短単位列が，「全体として 1 つの機能的長単位」となるのか，それとも，「複数の長単位から構成される列」となるのかという曖昧性を解消することを目的とした．適用する手法としては，条件付き確率場 (Conditional Random Fields, CRF) [Lafferty01] を利用したチャンキングを用い，評価実験の結果，機能的な長単位の検出において 97%近い F 値を達成した．

参考文献

[阿部 12] 阿部佑亮, 鈴木敬文, 宇津呂武仁, 山本幹雄, 松吉俊, 河田容英：対訳用例および意味的等価クラスを用いた機能表現の日英翻訳, 言語処理学会第 18 回年次大会論文集 (2012).

[グループ・ジャマシイ 98] グループ・ジャマシイ（編）：教師と学習者のための日本語文型辞典, くろしお出版 (1998).

[小早川 09] 小早川健, 関場治朗, 木下明徳, 熊野正, 加藤直人, 田中英輝：単語格子とマルコフモデルによる日本語機能表現の解析 — 日本語機能表現辞書「つつじ」を用いて —, 電子情報通信学会技術研究報告, NLC2009-1, pp. 15–20 (2009).

[BCCWJ 総括班 09] 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: BCCWJ 領域内公開データ（2009 年度版） (2009).

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[Lafferty01] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. 18th ICML*, pp. 282–289 (2001).

[劉 10] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊：意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析, 言語処理学会第 16 回年次大会論文集, pp. 194–197 (2010).

[松吉 07] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[森田 89] 森田良行, 松木正恵：日本語表現文型, NAFL 選書, 第 5 巻, アルク (1989).

[Nagasaka10] Nagasaka, T., Shimanouchi, R., Sakamoto, A., Suzuki, T., Morishita, Y., Utsuro, T. and Matsuyoshi, S.: Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Translation Rule Acquisition from Parallel Patent Sentences, *Proc. 7th LREC*, pp. 1778–1785 (2010).

[小椋 11] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕：『『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 (下)』, 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 (2011).

[坂本 09] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集, pp. 654–657 (2009).

[島内 10] 島内蘭, 長坂泰治, 坂本明子, 宇津呂武仁, 松吉俊：日英特許翻訳における日本語機能表現の集約的英訳可能性の調査, 言語処理学会第 16 回年次大会論文集, pp. 611–614 (2010).

[注連 07] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史：日本語機能表現の自動検出と統計的係り受け解析への応用, 自然言語処理, Vol. 14, No. 5, pp. 167–197 (2007).

[首藤 88] 首藤公昭, 吉村賢治, 武内美津乃, 津田健蔵：日本語の慣用的表現について—語の非標準的用法からのアプローチ—, 情報処理学会研究報告, 第 1988-NL-66 巻, pp. 1–7 (1988).

[首藤 98] 首藤公昭, 小山泰男, 高橋雅仁, 吉村賢治：依存構造に基づく言語表現の意味的類似度, 電子情報通信学会研究報告, 第 NLC98-30 巻, pp. 33–40 (1998).

[Shudo04] Shudo, K., et al.: MWEs as Non-propositional Content Indicators, *Proc. 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 32–39 (2004).

[鈴木 12] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係を利用した日本語機能表現の解析方式の評価, 言語処理学会第 18 回年次大会論文集 (2012).

[Tjong Kim Sang00] Tjong Kim Sang, E.: Noun Phrase Recognition by System Combination, *Proc. 1st NAACL*, pp. 50–55 (2000).

[土屋 06] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一：日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).

[土屋 07] 土屋雅稔, 注連隆夫, 高木俊宏, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一：機械学習を用いた日本語機能表現のチャンキング, 自然言語処理, Vol. 14, No. 1, pp. 111–138 (2007).

[山本 02] 山本和英：換言と言語変換の協調による機械翻訳モデル, 言語処理学会第 8 回年次大会発表論文集, pp. 307–310 (2002).

階層的機能表現辞書の意味的等価クラス および対訳用例を用いた機能表現の日英翻訳

阿部佑亮　(筑波大学大学院システム情報工学研究科)
鈴木敬文　(筑波大学大学院システム情報工学研究科)
宇津呂武仁　(筑波大学システム情報系)*
山本幹雄　(筑波大学システム情報系)
松吉俊　(山梨大学大学院医学工学総合研究部)
河田容英　(株式会社ナビックス)

Japanese to English Machine Translation of Functional Expressions based on Semantic Equivalence Classes of Hierarchical Lexicon and Translation Examples

Yusuke Abe　(University of Tsukuba)
Takafumi Suzuki　(University of Tsukuba)
Takehito Utsuro　(University of Tsukuba)
Mikio Yamamoto　(University of Tsukuba)
Suguru Matsuyoshi　(University of Yamanashi)
Yasuhide Kawada　(Navix Co., Ltd.)

1. はじめに

日本語には 16,000 種類以上の機能表現 (助詞・助動詞・接続詞相当語句) の異形が存在する．日本語機能表現には非常に多様な異形が存在するが，それらの異形を網羅的に正しく翻訳することは難しい．この問題に対応する手法として，先行研究では，日本語機能表現を網羅的に列挙した大規模日本語機能表現階層辞書における機能表現の意味的等価クラスを利用して，日英対訳特許文中に出現する日本語機能表現の日英翻訳を対象として，日本語機能表現の集約的な日英機械翻訳を行う手法を提案している．この手法を 53 の意味的等価クラスに適用した結果，20 クラスについては，意味的等価クラスに属する日本語機能表現の翻訳規則を 1 規則ないし 2 規則に集約出来ることが分かった．しかし，一方で，他の 33 クラスについては，意味的等価クラスに属する日本語機能表現の翻訳規則を集約することが出来なかった．これは，日本語機能表現を英訳する際の曖昧性のためであった．より正確な翻訳を行うためには，これら機能表現表記のもつ曖昧性を考慮した翻訳の仕組みが必要不可欠である．

以上を踏まえて，本論文では，NTCIR-7 の特許翻訳タスクで配布された 1,798,571 件の日英対訳特許文対から得た対訳用例を用いて，日本語機能表現を英訳する方式を提案する．この方式においては，機能表現の意味的等価クラスごとに，様々な対訳用例からデータベースを構築し，英訳対象となる機能表現表記の用例と最も類似した対訳用例の訳語を適用することで，上記の曖昧性に対応する．評価実験として，句に基づく統計的機械翻訳モデル Moses [Koehn07] を，日英対訳特許文を用いて訓練したものとの翻訳精度比較を行った．両手法の作成時に参照するテキストと同ジャンルである特

*utsuro @ iit.tsukuba.ac.jp

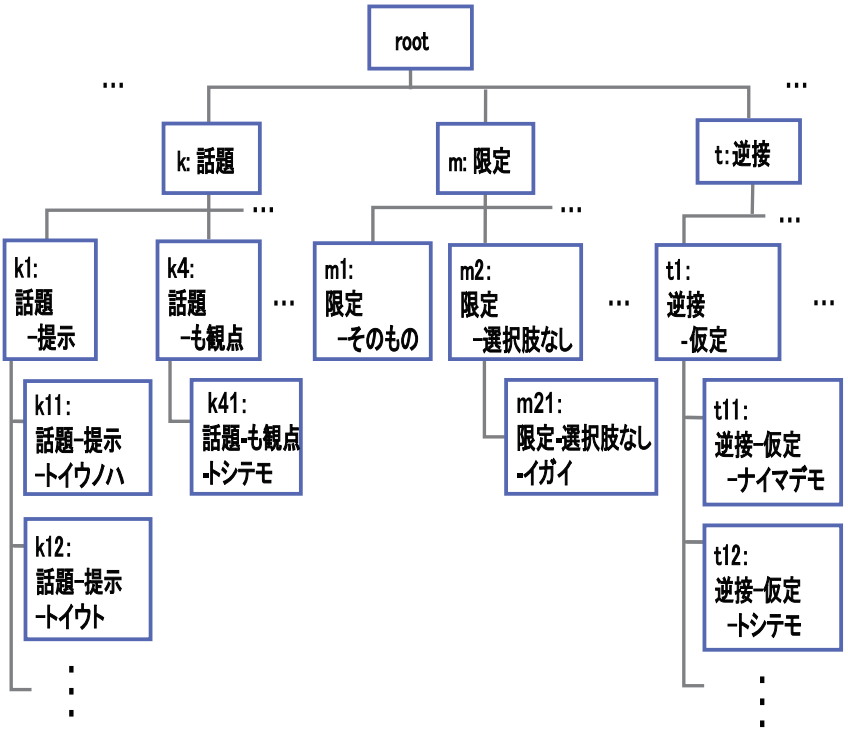


図 1: 意味的等価クラスに基づく階層構造

許文における翻訳精度は，多くの意味的等価クラスにおいて Moses の方が優れていたが,「日本語書き言葉均衡コーパス」および「日本語学習者用用例集」における翻訳精度は，多くの意味的等価クラスにおいて提案手法の方が優れていた．このことから，対訳用例を選定したテキストとは異なるジャンルのテキストにおける英訳においても，提案手法は比較的安定した翻訳性能を示すことを実証できた．

2. 階層的日本語機能表現辞書

[松吉 07, 松吉 08] は，日本語の機能表現の異型を，機能表現の構成要素の組み合わせとして階層的に収録した辞書を編纂した (日本語機能表現一覧「つつじ」[1])．日本語機能表現一覧「つつじ」には，16,801 の機能表現が収録されており，この辞書によって，日本語機能表現の網羅的取り扱いが可能になった．

また，日本語機能表現一覧「つつじ」では，図 1 に示すように，辞書に収録されている見出し語について，意味的等価クラスという形での階層的分類も行っている．この最下層に位置する全 199 個の意味的等価クラスについて，同一クラス内の機能表現は，日本語文中で言い換え可能であるとされている [松吉 08]．この意味的等価クラスを用いることにより，日本語機能表現の言い換え候補を網羅的に取り扱うことが可能となった．

3. 機能表現表記の曖昧性

日本語機能表現表記の適切な英訳を行うためには，日本語機能表現表記の持つ曖昧性に対応する必要がある．日本語機能表現表記を英訳するにあたって，対応すべき曖昧性は，大きく分けて 3 種類

[1] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 機能表現表記の曖昧性の例

(a) 機能的用法/内容的用法の曖昧性

	表記	例文	用法
(1)	ものの	乾燥に供した加熱空気は蒸発した水蒸気を含み，多くの熱エネルギーを持っている ものの ，回収して循環利用するには限界があり，多くの場合廃棄されている．	機能的用法 t24(逆接-確定-モノノ) (～ ものの= *although* ～)
(2)	ものの	ここで，ブロックが存在しない場合は，探索対象段の位置を，保持されたアベイラブルエリアで最後の ものの 左上隅点とし（ステップ１１０６），その後，後述する図１２に示される処理を実行する。	内容的用法 (～ ものの = *of*)

(b) 複数の機能的用法間の曖昧性

	表記	例文 (英訳文)	用法
(3)	としても	このため、誤って装置に物等を落下した としても ，その衝撃は反射ミラー８ｆに伝わり難くなっている．	機能的用法 t12(逆接-仮定-トシテモ) (としても= *even when*)
(4)	としても	さらに，ブレード４５は接触ローラ３７の外周面３７ａの汚れを除去するクリーニング手段 としても 作用する．	機能的用法 k41(話題-も観点-トシテモ) (としても= *as*)

(c) 対訳英語の曖昧性

	表記	例文 (英訳文)	用法
(5)	による	原稿台１１側からの光のミラー１４ による 反射光路上には結像レンズ１６とプラテン２０がこの順に配置されている。	機能的用法 c11(仲介-原因-ニヨッテ) (による = *by*)
(6)	による	本発明 による 可変差動制限装置２の制御は、以下の (1), (2), (3) の３種の制御の組合せから構成される。	機能的用法 c11(仲介-原因-ニヨッテ) (による= *according to*)
(7)	による	つまり、放電開始 による 電圧の低下が、極間異常状態と判定されてしまうことがある。	機能的用法 c11(仲介-原因-ニヨッテ) (による= *due to*)

ある．1つは，文中の表現が機能表現の意味として用いられているもの (機能的用法) と，その表現を構成する語本来の意味で用いられているもの (内容的用法) との間の曖昧性である (表 1 (a))．もう 1 つは，機能表現の意味が文脈によって異なるという機能的用法の曖昧性である (表 1 (b))．そして最後の 1 つは，対訳英語の曖昧性である (表 1 (c))．

4. 対訳用例データベースの構築

本論文では，NTCIR-8 の特許翻訳タスク [Fujii10] で配布された日英対訳特許文の文対応データのうち，1,798,571 件をフレーズテーブルの訓練用データとして使用した．この文対応データに対して，句に基づく統計的機械翻訳モデルのツールキットである Moses [Koehn07] を適用し，日英の句の組および日英の句の組が対応する確率を示したフレーズテーブルを作成する．

このフレーズテーブルを用いて，先の約 180 万件の日英対訳特許文対から，対訳用例データベースを構築した．その構築手順を，図 2 に示す．具体的には，対訳用例データベースの構築対象としている意味的等価クラスに属する各日本語機能表現表記について，以下の条件を満たす「日本語機能表現表記- 英訳語」組をフレーズテーブルから抽出する．

- 日英対訳特許文対における日本語機能表現表記の出現頻度が 20 以上
- Moses によって，日英対訳特許文対における「日本語機能表現表記 - 英訳語」組が句対応していると自動判定された箇所の頻度が 10 以上
- フレーズテーブルにおける日英翻訳確率が 0.05 以上

そして，抽出した各「日本語機能表現表記 - 英訳語」組について，この表記および英訳語が対応関係であると人手で判断された対訳文対を，約 180 万件の日英対訳特許文対から収集し，対訳用例デー

表 2: 対訳用例データベースを構築した意味的等価クラス

意味的等価クラス			表現数	表現の例
日本語機能表現表記の用法の曖昧性	大きい	M11(不必要 - 不必要 - ナクテヨイ)	299	なくてもよい，までもない，ずともよく
		P11(例示 - 程度 - クライ)	6	くらい，ばかり，ほど
		c11(仲介 - 原因 - ニヨッテ)	15	により，をもって，によります
		m12(限定 - そのもの - ノミ)	5	きり，だけ，のみ
		n12(添加 - 非限定 - ダケデナク)	12	のみならず，だけじゃなく，上に
		s11(理由 - 因状況 - イジョウハ)	9	からには，うえは，以上
		t12(逆接 - 仮定 - トシテモ)	21	にしても，としましても，たところで
	小さい	D11(判断 - 当為 - ナケレバナラナイ)	213	ないといけない，ねばならない，べき
		b11(対象 - 関連 - ニツイテ)	26	に関する，について，につきまして
		u12(対比 - 般 - カワリニ)	14	代わりに，代りに，かと思うと

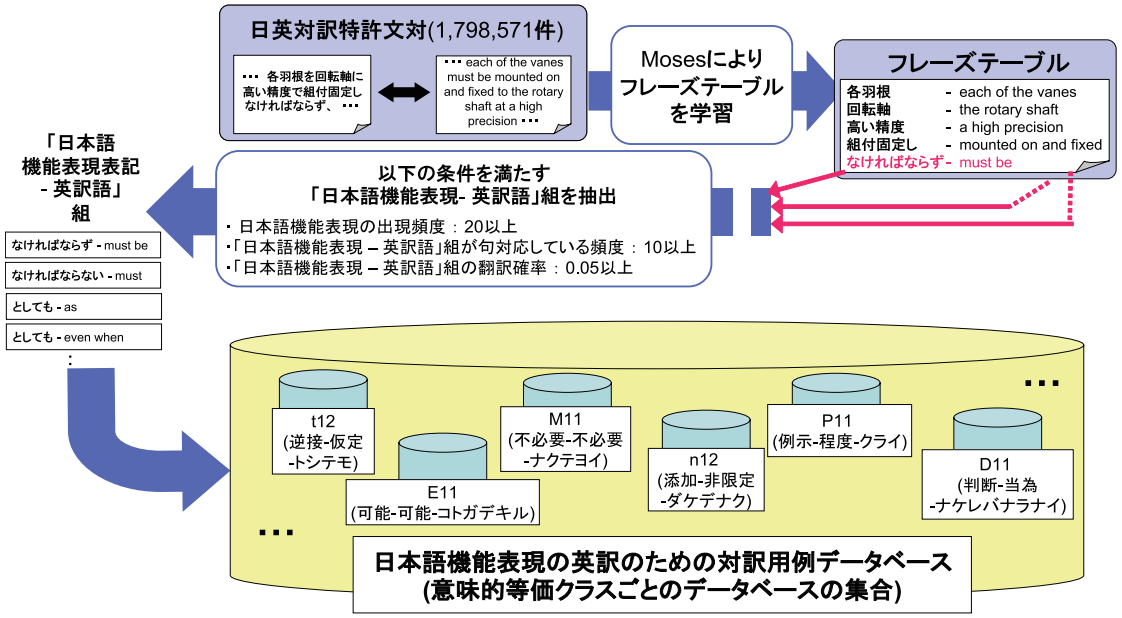


図 2: 意味的等価クラスごとの対訳用例データベースの構築手順

タベースへ登録する．ただし，本研究では，意味的等価クラスごとのデータベースを用意し，当該日本語機能表現が属する意味的等価クラスのデータベースにのみ，対訳用例を追加する．この構築手順に従って，表 2 に示した 10 の意味的等価クラスの対訳用例データベース構築を行った．その結果，表 3 に示すように，10 クラス合計で 5,253 用例の対訳用例データベースを構築することができた．

5. 対訳用例を用いた機能表現の日英翻訳

対訳用例を用いた機能表現の日英翻訳においては，翻訳対象である日本語用例が与えられると，まず，用例中の機能表現が属する意味的等価クラスの対訳用例データベースが参照される．次に，用例間の類似度に基づいて，与えられた用例中の日本語機能表現表記と用法の最も類似した対訳用例を選択する．そして，その対訳用例における英訳語を，翻訳対象の日本語機能表現表記の英訳語として採用する．

以下，まず，入力された日本語用例を $e_j = \langle m_{pre}, M_c, m_{suf} \rangle$，その日本語用例 e_j 中の日本語機能表現表記を $f_j(e_j)$ とする．ただし，m_{pre}，m_{suf} はそれぞれ，日本語機能表現表記に前接する，あるいは，後接する形態素を表し，M_c は，日本語機能表現表記を構成する形態素列を表す．また，データベース中のある用例を $e_{je}^{db} = \langle e_j^{db}, t_e^{db} \rangle$ とする．ただし，e_j^{db} は，データベース中の用例の日本語部分を，t_e^{db} は e_j^{db} 中の機能表現部分の英訳を，それぞれ表す．ここで，日本語用例 e_j 中の日本語機能

表 3: 対訳用例データベース中の日本語機能表現表記数・用例数

意味的等価クラス			日本語表記数	「日本語機能表現表記 - 英訳語」組数	用例数
日本語機能表現表記の用法の曖昧性	大きい	M11(不必要 - 不必要 - ナクテヨイ)	5	16	96
		P11(例示 - 程度 - クライ)	3	5	113
		c11(仲介 - 原因 - ニヨッテ)	5	15	1489
		m12(限定 - そのもの - ノミ)	2	4	773
		n12(添加 - 非限定 - ダケデナク)	6	12	797
		s11(理由 - 因状況 - イジョウハ)	2	5	494
		t12(逆接 - 仮定 - トシテモ)	6	16	778
		上記 7 クラスの合計	29	73	4540
	小さい	D11(判断 - 当為 - ナケレバナラナイ)	6	9	114
		b11(対象 - 関連 - ニツイテ)	6	19	455
		u12(対比 - 般 - カワリニ)	4	7	144
		上記 3 クラスの合計	16	35	713
合計			45	108	5253

表現表記 $f_j(e_j)$ の属する意味的等価クラスの集合を $SS(f_j(e_j))$ とし[2]，$SS(f_j(e_j))$ 中の一つの意味的等価クラスを S，S の対訳用例データベースを EDB_S とすると，参照すべき対訳用例全体の集合は，以下の $EDB(e_j)$ で表される．

$$EDB(e_j) \quad = \quad \bigcup_{S \in SS(f_j(e_j))} EDB_S$$

また，日本語用例 e_j と対訳用例データベース中の用例 e_{je}^{db} との類似度 $Sim\big(e_j, e_{je}^{db} = \langle e_j^{db}, t_e^{db} \rangle\big)$ は，以下で定義される．

$$\begin{aligned} Sim\Big(e_j, e_{je}^{db} = \Big\langle e_j^{db}, t_e^{db} \Big\rangle\Big) \quad = \quad & Sim_{pre}\Big(m_{pre}(e_j), m_{pre}(e_j^{db})\Big) + \; Sim_c\Big(M_c(e_j), M_c(e_j^{db})\Big) \\ & + \; Sim_{suf}\Big(m_{suf}(e_j), m_{suf}(e_j^{db})\Big) \end{aligned}$$

ここで，Sim_{pre} および Sim_{suf} はそれぞれ，e_j，e_j^{db} の前接形態素，および，後接形態素の類似度であり，Sim_c は，e_j の構成形態素列と e_j^{db} の構成形態素列の類似度である．厳密には，Sim_{pre} および Sim_{suf} はそれぞれ，前接形態素 $m_{pre}(e_j)$，$m_{pre}(e_j^{db})$，および後接形態素 $m_{suf}(e_j)$，$m_{suf}(e_j^{db})$ の品詞・活用形の一致数を正規化した値を表し，Sim_c は，構成形態素列 $M_c(e_j)$，$M_c(e_j^{db})$ の品詞・活用形の一致数 [3]を正規化した値を表している．図 3 は，用例間の類似度 $Sim\Big(e_j, e_{je}^{db} = \Big\langle e_j^{db}, t_e^{db} \Big\rangle\Big)$ の定義を，具体例で図示したものである．この図中の 2 つの用例の場合，前接・後接形態素および構成形態素列の品詞・活用形が，いずれも完全に一致しているため，Sim_{pre}，Sim_{suf} および Sim_c はいずれも 1 となり，この 2 つの用例間の類似度 $Sim\Big(e_j, e_{je}^{db}\Big)$ は，最大値である 3 となる．

以上に基づき，参照すべき対訳用例全体の集合中で最も類似する対訳用例に基づいて機能表現の翻訳を行う関数 $tran_{fe}\big(f_j(e_j)\big)$ は，以下で定義される．

$$tran_{fe}\Big(f_j(e_j)\Big) \quad = \quad t_e^{db} \quad s.t. \; \Big\langle e_j^{db}, t_e^{db} \Big\rangle = \underset{e_{je}^{db} \in EDB(e_j)}{\operatorname{argmax}} \; Sim\Big(e_j, e_{je}^{db} = \Big\langle e_j^{db}, t_e^{db} \Big\rangle\Big)$$

提案手法による訳語選択の例を，図 4 に示す．

[2]ここでは，一つの機能表現表記が複数の意味的等価クラスに属する場合があることを想定した定式化となっている．

[3]2 つの構成形態素列 $M_c(e_j)$，$M_c(e_j^{db})$ の形態素数が異なる場合，形態素列の前側から順に形態素を対応付け，品詞・活用形の一致数を計上していき，形態素列が長い方について，対応付ける形態素が無い場合，その形態素との品詞・活用形の一致数は 0 とする．

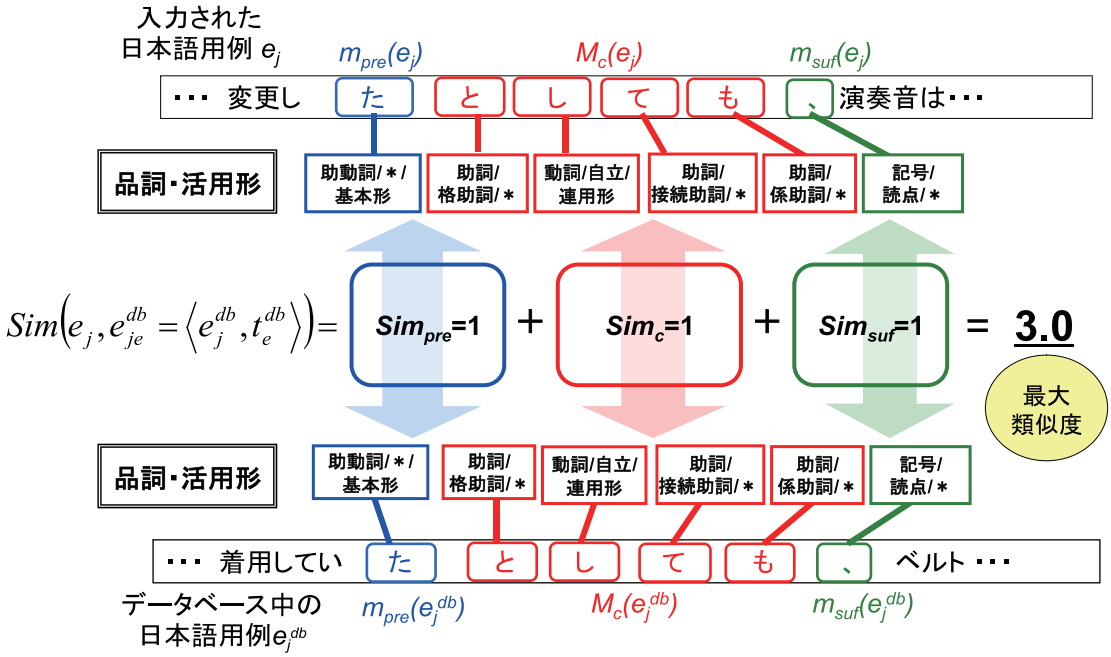


図 3: 機能表現表記の用例間の類似度

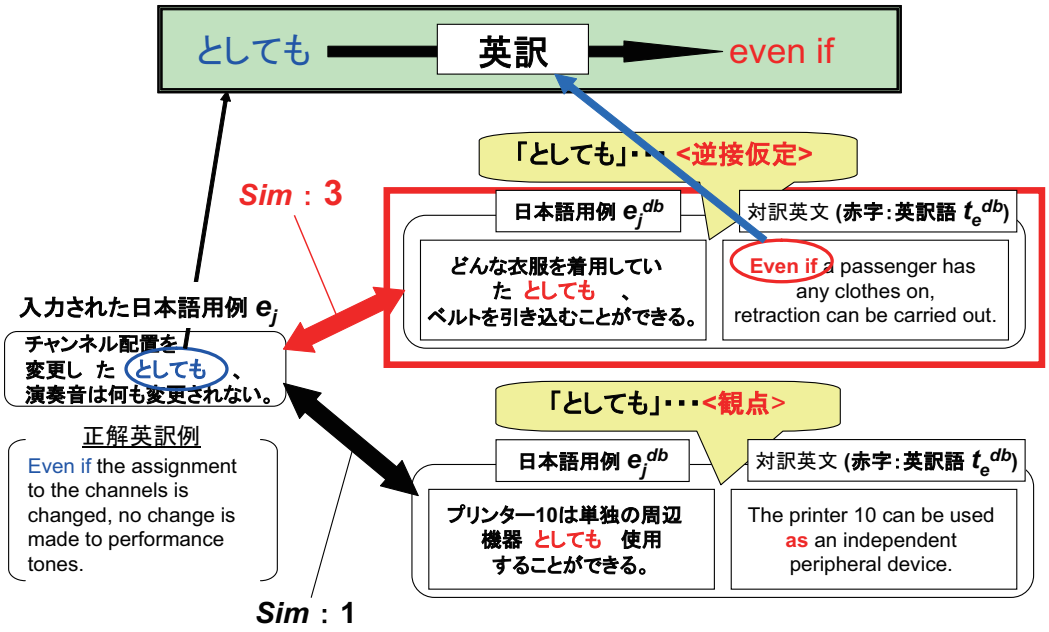


図 4: 日本語機能表現表記の用例間の類似度を用いた訳語選択

6. 評価

提案手法の評価として，提案手法および句に基づく統計的機械翻訳モデルである Moses [Koehn07] を日英対訳特許文を用いて訓練したものについて，翻訳精度の比較を行った．評価は，4 節で述べた手順に従って対訳用例データベースを構築した 10 の意味的等価クラスを対象として行った．

評価文は，NTCIR-8 の特許翻訳タスク [Fujii10] で配布された日英対訳特許文のうちの約 140 万件 (以下,「特許文」),「現代日本語書き言葉均衡コーパス」 [BCCWJ 総括班 11](以下,「書き言葉コー

表 4: 評価文の収集対象となるテキストの内訳

テキストの種類	テキストの特徴	文数
日英対訳特許文 (特許文)	2001-2007 年発行の日本公開特許および，それと対応する米国公開特許の文書について，互いに対訳関係にある度合いの高い部分として「背景」および「実施例」の部分を抽出し，日英間で文対応を付けられている.	1,387,713
現代日本語書き言葉均衡コーパス (書き言葉コーパス)	新聞や小説，雑誌，ウェブページ，国会議事録等，さまざまなジャンルの文書から収集された文を収録している.	4,212,638
日本語学習者用用例集	日本語学習者向けに，日常会話文を中心とした文を収録している.	9,125

表 5: 評価結果

評価対象		評価文数	翻訳精度 (%)			
			ベースライン	Moses	提案手法	提案手法 (上限値)
テキストジャンル別	特許文	140	53	**66**	65	70
	書き言葉コーパス	292	53	34	**77**	78
	日本語学習者用用例集	168	51	42	**56**	62
評価文集合別	評価文集合 (1) (提案手法：英訳が容易, Moses：英訳が容易)	210	54	**66**	**66**	73
	評価文集合 (2) (提案手法：英訳が容易でない, Moses：英訳が容易)	180	48	47	**54**	58
	評価文集合 (3) (提案手法：英訳が容易, Moses：英訳が容易でない)	90	58	31	**74**	82
	評価文集合 (4) (提案手法：英訳が容易でない, Moses：英訳が容易でない)	120	55	20	**63**	65
合計		600	53	46	**63**	69

パス」)，および,「日本語学習者用用例集」 [グループ・ジャマシイ 98]，の 3 種類のテキストから収集した．各テキストの特徴および文数を，表 4 に示す．また，評価文は,「対訳用例データベース中の各用例との類似度の最大値 $Sim_{max}(e_j)$ [4]の範囲」および「評価文 e_j 中の日本語機能表現表記 $f_j(e_j)$ が Moses の学習データである約 180 万件の対訳特許文中に出現するか」の 2 つの尺度で分類した．前者に関しては，$2.33 \leq Sim_{max}(e_j) \leq 3$ を満たす評価文 e_j を「提案手法での英訳が容易」，$0 \leq Sim_{max}(e_j) < 2.33$ を満たす評価文 e_j を「提案手法での英訳が容易でない」と，それぞれ判断した．同様に後者に関しては，日本語機能表現表記 $f_j(e_j)$ が Moses の学習データである約 180 万件の対訳特許文中に出現する評価文 e_j を「Moses での英訳が容易」，出現しない評価文 e_j を「提案手法での英訳が容易でない」と，それぞれ判断した．この 2 つの尺度に従って，各意味的等価クラスについて 4 種類の評価文集合を作成した．また，評価文は，各評価文集合から選定された評価文の数が均等になるように選定した.

評価結果を表 5 に示す.「提案手法」,「Moses」は，それぞれの手法での翻訳精度を示しており,「提案手法 (上限値)」は，提案手法において複数の英訳語が出力され，その中に評価文中の機能表現表記

[4] $Sim_{max}(e_j) = \max_{e_{je}^{db}} Sim\left(e_j, e_{je}^{db} = \left\langle e_j^{db}, t_e^{db} \right\rangle\right)$ と定義する.

に対して適切な英訳語と不適切な英訳語が混在する場合，その中から適切な英訳のみを選択し，出力した場合の翻訳精度を示している.「ベースライン」は，各意味的等価クラスの対訳用例データベースにおける対訳用例の英訳語 t_e^{db} のうち，頻度最大のものを英訳語として出力した場合の精度である．評価の結果，評価文書集合全体では，提案手法の方が Moses よりも優れた翻訳精度となった．また，ベースラインと比較して，どの評価文集合においても，提案手法の方が翻訳精度が高いことから，提案手法によって機能表現表記の曖昧性に対応した翻訳が行われていることが分かる．

テキストのジャンル別の翻訳精度を見てみると，両手法の作成時に参照するテキストと同ジャンルである「特許文」では，Moses の方が精度が高いが，それ以外のジャンルのテキストでは，提案手法の方が精度が高いことが分かる．このことから，提案手法は，特定のジャンルのテキストから対訳用例を集めても，それとは異なるジャンルのテキストへ適応できる可能性が高いことが分かる．通常，一般的なジャンルにおいては，二言語対訳コーパスの作成は高コストであり，使用できるものが限られるため，この特性は重要であると考えられる．

評価文中の日本語機能表現表記が訓練データ中に出現しない評価文集合 (3)，(4) における翻訳精度の比較結果をふまえると，Moses は訓練データから作成したフレーズテーブルを用いて翻訳を行うため，訓練データ中に出現しない表記の翻訳を行うことが困難であることが分かる．一方，提案手法は用例間の類似度として品詞・活用形の情報のみを利用しているため，訓練データ中に出現しない表記の翻訳にも対応することができている．

7. 関連研究

本論文の関連研究としては，大きく分けて「日本語機能表現の検出および解析に関する研究」と「機械翻訳に関する研究」の 2 種類がある．以下，それぞれについて述べる．

7.1 日本語機能表現の検出および解析に関する研究

機能表現の検出および解析を目的とした研究としては [土屋 07,注連 07,小早川 09,鈴木 12b] が，機能表現を考慮した係り受け解析に関する研究としては [注連 07] が，それぞれ挙げられる．このうち，機能表現の検出を目的とした研究は，機能表現表記の持つ曖昧性のうち，機能的用法/内容的用法の曖昧性の解消を行う研究と言える．特に， [鈴木 12a] においては，階層的日本語機能表現辞書 [松吉 07] の機能表現の間の階層的関係を利用して，上層に位置する代表的な機能表現の用例を利用して，その派生表現にあたる機能表現の検出を行う手法を提案している．

7.2 機械翻訳に関する研究

本論文で採用した**用例ベース翻訳** [Nagao84,佐藤 92,Sommers03] の枠組みは，大きく分けて，**対訳用例データベース**と**用例を参照するための用例間の類似度**の 2 つの要素から構成される．用例ベース翻訳の代表的研究の一つとして，[佐藤 91] では，翻訳対象を動詞とその必須格の名詞からなる動詞フレームに限定して，用例ベースの英日翻訳を行っている．具体的には，対訳用例を動詞フレームという枠で扱い，注目する動詞の必須格である名詞の出現頻度の傾向が類似する動詞フレームを参照している． [村田 01] では，日英翻訳における動詞の時制や進行形，完了形や助動詞の選択を，用例に基づいて行っている．参照する用例を決定するための類似度として，文末からの一致文字列の長さという，簡便なものを用いているが，分類語彙表 [国立国語研究所 64] の分類番号や語の変化形といった情報も利用している．

その他には，代表的表現への言い換えを介した機械翻訳の研究として，内容語と口語的な機能表現を扱った [山本 01,山本 02] が知られている．同様に，階層的日本語機能表現辞書 [松吉 07] の機能表

現を対象として，代表的表現への言い換えを介して機械翻訳を行う手法の研究として,「日本語学習者用用例集」 [グループ・ジャマシイ 98] 中の例文を対象とした集約的英訳についての研究事例 [坂本 09]，特許文を対象とした集約的英訳についての研究事例 [島内 11]，および，集約的中国語訳についての研究事例 [劉 10] がある．

8. おわりに

本論文では，対訳用例に基づく日本語機能表現表記の英訳手法を提案した．対訳用例データベースは，約 180 万件の日英対訳特許文より収集した．意味的等価クラス別の対訳用例データベースとすることで，参照する用例の探索範囲を，用法の類似する用例に限定した．さらに，機能表現表記の用法の区別をせずに対訳用例を収集し，品詞・活用形の情報に基づく用例間の類似度を利用して参照する用例を選択することで，機能表現表記の持つ曖昧性への対応も可能とした．評価実験として，句に基づく統計的機械翻訳モデル Moses [Koehn07] を日英対訳特許文を用いて訓練したものとの翻訳精度比較を行った．両手法の作成時に参照するテキストと同ジャンルである特許文における翻訳精度は，多くの意味的等価クラスにおいて Moses の方が優れていたが,「日本語書き言葉均衡コーパス」および「日本語学習者用用例集」における翻訳精度は，多くの意味的等価クラスにおいて提案手法の方が優れていた．このことから，対訳用例を選定したテキストとは異なるジャンルのテキストにおける英訳においても，提案手法は比較的安定した翻訳性能を示すことを実証できた．

今後の課題としては，対訳用例データベース構築の手順において，対訳用例収集の対象とする「日本語機能表現表記 - 英訳語」組の条件を修正するか，あるいは,「日本語機能表現表記 - 英訳語」組の人手による選定を行い，より多くの英訳語を含むようデータベースを拡張することが挙げられる．また，他の意味的等価クラスについても同様に対訳用例データベースを構築し，その翻訳性能を検証することも挙げられる．そして，Moses のような統計的機械翻訳の手法との併用によって，翻訳精度向上が可能であると考えられるので，機械学習の導入により，それを実現することを目指す．

参考文献

[Fujii10] Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Echizen-ya, H., Ehara, T. and Shimohata, S.: Overview of the Patent Translation Task at the NTCIR-8 Workshop, *Proc. 8th NTCIR Workshop Meeting*, pp. 371–376 (2010).

[グループ・ジャマシイ 98] グループ・ジャマシイ（編）：教師と学習者のための日本語文型辞典, くろしお出版 (1998).

[小早川 09] 小早川健, 関場治朗, 木下明徳, 熊野正, 加藤直人, 田中英輝：単語格子とマルコフモデルによる日本語機能表現の解析 — 日本語機能表現辞書「つつじ」を用いて —, 電子情報通信学会技術研究報告, NLC2009-1, pp. 15–20 (2009).

[Koehn07] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th ACL, Companion Volume*, pp. 177–180 (2007).

[BCCWJ 総括班 11] 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: 特定領域研究「日本語コーパス」研究成果報告 (2011).

[劉 10] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊：意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析, 言語処理学会第 16 回年次大会論文集, pp. 194–197 (2010).

[松吉 07] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[村田 01] 村田真樹, 馬青, 内元清貴, 井佐原均：用例ベースによるテンス・アスペクト・モダリティの日英翻訳, 人工知能学会論文誌, Vol. 16, No. 1, pp. 20–28 (2001).

[Nagao84] Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, Elithorn, A. and Banerji, R. (eds.), Artificial and Human Intelligence, Elsevier Science Publishers, B.V (1984).

[国立国語研究所 64] 国立国語研究所：分類語彙表, 秀英出版 (1964).

[坂本 09] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集, pp. 654–657 (2009).

[佐藤 91] 佐藤理史：MBT1: 実例に基づく訳語選択, 人工知能学会誌, Vol. 6, No. 4, pp. 592–600 (1991).

[佐藤 92] 佐藤理史：実例に基づく翻訳, 情報処理, Vol. 33, No. 6, pp. 673–681 (1992).

[島内 11] 島内蘭, 阿部佑亮, 鈴木敬文, 宇津呂武仁, 松吉俊：特許文における日本語機能表現の集約的英訳規則の作成と評価, 言語処理学会第 17 回年次大会論文集, pp. 396–399 (2011).

[注連 07] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史：日本語機能表現の自動検出と統計的係り受け解析への応用, 自然言語処理, Vol. 14, No. 5, pp. 167–197 (2007).

[Sommers03] Sommers, H.: An Overview of EBMT, Carl, M. and Way, A. (eds.), Recent Advances in Example-Based Machine Translation, pp. 3–57, Kluwer Academic (2003).

[鈴木 12a] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔：『現代日本語書き言葉均衡コーパス』における複合辞の検出と評価, 『コーパス日本語学ワークショップ』予稿集 (2012).

[鈴木 12b] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係を利用した日本語機能表現の解析方式の評価, 言語処理学会第 18 回年次大会論文集 (2012).

[土屋 07] 土屋雅稔, 注連隆夫, 高木俊宏, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一：機械学習を用いた日本語機能表現のチャンキング, 自然言語処理, Vol. 14, No. 1, pp. 111–138 (2007).

[山本 01] 山本和英, 白井諭, 坂本仁, 張玉潔：SANDGLASS: 両言語換言機構を基軸とする音声翻訳, 言語処理学会第 7 回年次大会発表論文集, pp. 221–224 (2001).

[山本 02] 山本和英：換言と言語変換の協調による機械翻訳モデル, 言語処理学会第 8 回年次大会発表論文集, pp. 307–310 (2002).

日本語音声コーパスにおける促音・非促音の判別

天野成昭†（愛知淑徳大学）
山川仁子（愛知淑徳大学）
近藤眞理子（早稲田大学）

Discrimination between Single and Geminate Stops in Japanese Speech Corpus

Shigeaki Amano (Aichi Shukutoku University)
Kimiko Yamakawa (Aichi Shukutoku University)
Mariko Kondo (Waseda University)

１．はじめに

日本語において「っ」で表記される促音の主な音響的特徴は，音声波形上の無音区間の時間長にある。例えば促音を含む「いった」では無音区間の時間が長く，非促音の「いた」では無音区間の時間が短い。言い換えるならば，促音と非促音の生成範疇境界は無音区間の時間長に関係していると言える。

しかし，この無音区間の時間長は，発声速度によって変動することが分かっている。すなわち，発声速度が速ければ無音区間の時間長は短くなり，逆に発声速度が遅ければ無音区間の時間長は長くなる。したがって，無音区間の時間長だけでは，様々な発声速度における促音と非促音の生成範疇境界を正しく表すことはできない。

この発声速度による変動を吸収し，発声速度に依存しない促音・非促音の生成範疇境界の表現を求めることを目的として，Amano & Hirata (2010)は，Hirata & Whiton (2005)が録音した日本語２音節単語における促音・非促音の生成データを分析した。彼らはその分析結果に基づき，「促音・非促音の生成範疇境界は，閉鎖区間の時間長とそれを含む単語の時間長の２変数による１次式で表される」と主張した。これに対し Hirata & Whiton (2005)は，「促音・非促音の生成範疇境界は，閉鎖区間の時間長とそれを含む単語の時間長の比で表される」と主張している。Amano & Hirata (2010)および Hirata & Whiton (2005)が分析に用いた生成データは，話者数が 4 名と少ないため，それぞれ主張の正否を判断することは難しく，またどちらの主張も一般化するにはやや難があった。

そこで，多数の話者による生成データによって Amano & Hirata (2010)および Hirata & Whiton (2005)の主張を検証することを目的として，『日本語話し言葉コーパス』RDB 版(小磯・伝・前川, 2012)から複数話者が発声した促音・非促音を抽出して判別分析を行い，促音・非促音の生成範疇境界の表現を特定することにした。

なお，Amano & Hirata (2010)および Hirata & Whiton (2005)が用いた単語は全て２音節単語であるので，彼らの主張はそれぞれ「促音・非促音の生成範疇境界は，閉鎖区間の時間長とそれを含む２音節の時間長の２変数による１次式で表される」および「促音・非促音の生成範疇境界は，閉鎖区間の時間長とそれを含む２音節の時間長との比で表される」と読み替えることができる。本研究では，この読み替えた２つの主張をそれぞれ検証する。

２．解析

２.１　解析対象

『日本語話し言葉コーパス』RDB 版(小磯・伝・前川, 2012)から，以下の条件を満たす促音を含む音韻列 $C_1V_1QC_2V_2X$，および非促音を含む音韻列 $C_1V_1C_2V_2X$ を抽出し，解析対象とした。ただし，C_1: 先行子音，V_1: 先行母音，Q: 促音，C_2: 後続子音，V_2: 後続母音，

† psy@asu.aasa.ac.jp

X: 任意の音韻である。

・V_1, V_2 は/a/, /i/, /u/, /e/, /o/のいずれかである
・V_1, V_2 は無声化・長音化していない
・C_1 は任意の子音または空である
・C_2 は/p/, /t/, /ts/, /ch/, /k/のいずれかである
・Q および C_2 は長単位で区切った場合の単語の語頭に位置しない
・C_1, C_2 は拗音を含まない
・X は撥音・促音ではない

この条件は Amano & Hirata (2010)が解析対象の促音・非促音を含む２音節単語の選択に用いた条件よりも緩和されている。緩和された点は，以下のとおりである。

・C_1 は任意の子音である
・C_2 に/ts/, /ch/が含まれる
・V_2 に/i/, /u/が含まれる
・単語長が任意である
・促音・非促音の単語内位置が語頭以外の任意の位置である
・促音と非促音を含む音韻列がミニマルペアをなさなくてもよい

抽出の結果，得られた促音を含む音韻列の数は 6056，非促音を含む音韻列の数は 34012，合計数は 40068 であった。話者の異なり数は 201 名であった。

２.２　解析方法

抽出した音韻列の各音韻の時間情報から，促音および非促音の閉鎖区間の時間長 y (ms)と，閉鎖区間を含む２音節 $C_1V_1(Q)C_2V_2$ の時間長 x (ms) を計算した。時間長の計算は Amano & Hirata (2010)の計算方法と同一にするために，次の 2 条件を設けて実施した。

・C_1, C_2 に無音区間を含めない
・V_1, V_2 の終了後の声帯振動の持続時間は母音の時間長に含める

計算で得た y と x を独立変数とし，促音・非促音のカテゴリーを従属変数として１次式による判別分析を行った。さらに y と x の比(y/x)を独立変数とし，促音・非促音のカテゴリーを従属変数とする判別分析を行った。

２.３　解析結果

閉鎖区間の時間長 y (ms)とそれを含む２音節の時間長 x (ms)を変数とする１次式を用いて判別分析を行った場合，得られた判別関数は，

$$y = 0.0588x + 38.4 \quad \text{(式１)}$$

であった。図１に，この判別関数と判別対象の促音・非促音のデータを示す。この判別関数による促音と非促音の誤判別率は 7.30%であった。

一方，閉鎖区間の時間長 y (ms)とそれを含む２音節の時間長 x (ms) の比 y/x を変数として判別分析を行った場合，得られた判別関数は

$$y/x = 0.255 \quad \text{(式２)}$$

であった。この判別関数による促音と非促音の誤判別率は 17.3%であった。

３．考察

閉鎖区間の時間長 y (ms)とそれを含む２音節の時間長 x (ms) を変数とする１次式によって判別分析を行った場合の誤判別率 7.30%は Amano & Hirata (2010)が得た誤判別率 8.72%とほぼ同じである。この結果は促音と非促音の判別が，閉鎖区間の時間長 y (ms)とそれを含む２音節の時間長 x (ms) の２変数による１次式を用いてほぼ可能であることを示している。すなわち促音と非促音の生成範疇境界はこの２変数の１次式によって表せると言える。この結果は Amano & Hirata (2010)の主張を支持する結果である。ただし, Amano & Hirata (2010)が得た判別関数は

$$y = 0.446x - 17.1 \qquad (式３)$$

であり，本研究で得られた判別関数(式１)とは異なっている。この違いは，Amano & Hirata (2010)では「速い・普通・遅い」の３種の発声速度による音声を用いているのに対し，本研究では『日本語話し言葉コーパス』RDB 版(小磯・伝・前川, 2012)から抽出した普通の発声速度が大半を占める音声を用いていることに一因があると考えられる。すなわち，促音においても非促音においても，速い発声速度では y と x は小さくなり，遅い発声速度では y と x は大きくなるので，速い発声速度のデータや遅い発声速度のデータが普通の発声速度のデータと同数程度存在した場合，判別関数の傾きは大きな正の値になりやすく，かつ切片は負の値になりやすくなる。この傾向によって Amano & Hirata (2010)の判別関数(式３)の傾きは大きくかつ切片が負となり，本研究の判別関数(式１)の傾きは小さくかつ切片が正になったと考えられる。

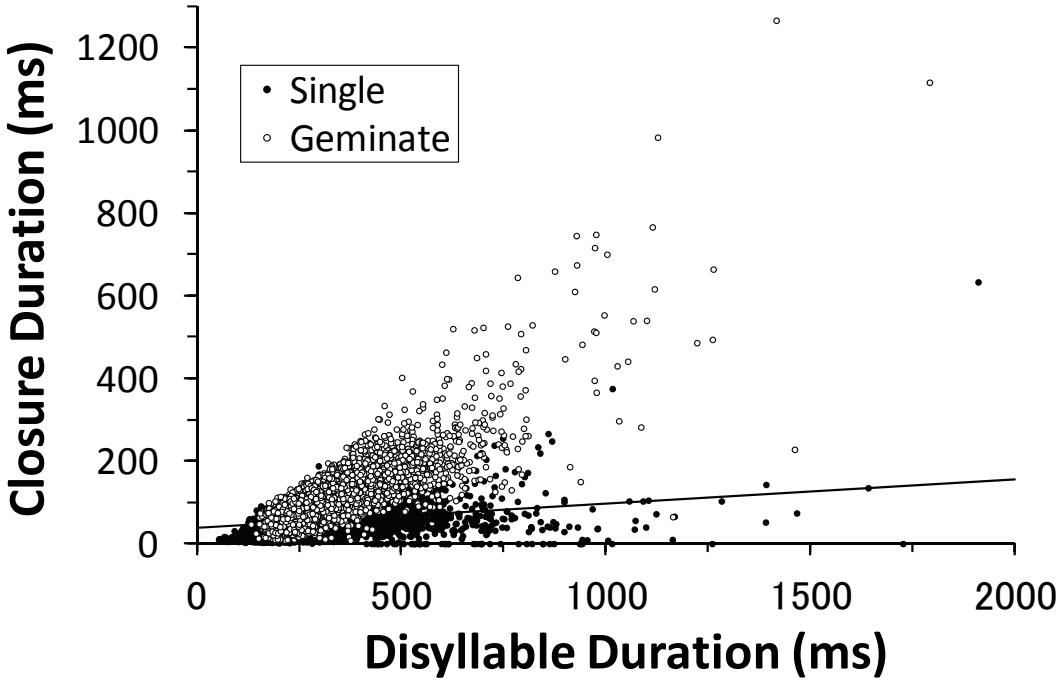


図１　閉鎖区間の時間長 y と閉鎖区間を含む２音節の時間長 x で表される平面上にプロットした促音(Geminate)・非促音(Single)のデータおよび判別関数(y = 0.0588x + 38.4)

一方，閉鎖区間の時間長 y (ms)とそれを含む２音節の時間長 x (ms) の比 y/x を変数として判別分析を行った場合の誤判別率は 17.3%と大きい。これは，比 y/x を用いた場合，促音と非促音の判別が十分にできないことを意味している。したがって，促音・非促音の生成範疇境界は比 y/x では正確に表せないと言える。これは「促音・非促音の生成範疇境界は，閉鎖区間の時間長とそれを含む２音節の時間長との比で表される」という Hirata & Whiton (2005)の主張に反する結果である。本研究で用いた音声は Hirata & Whiton (2005)よりも，発声数，話者数，および母音や子音の種類が多いので，本研究の結果のほうが正しい可能性が高いと考えられる。

多数話者の生成データを用いた本研究の結果, 閉鎖区間の時間長 y (ms)とそれを含む２音節の時間長 x (ms) の２変数による１次式を用いれば，促音と非促音の範疇境界が表せるという結論が得られた。しかし，この結論はいくつかの制限条件の下で得られた解析結果に基づいているので，これがどのような条件においても成り立つという保証はない。特に次に示す場合に関しては，この結論が正しいか否か不明である。

- V_1, V_2 が無声化した場合
- V_1, V_2 が長音化した場合
- V_2 に撥音が後続する場合
- V_2 に促音が後続する場合
- C_1, C_2 に拗音が含まれる場合
- C_2 が摩擦音である場合

より一般性のある結論を導き出すために，これらの場合について今後さらに解析を進める予定である。

謝　辞

本研究の一部は国立国語研究所（言語資源研究系）基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）による成果である。また本研究は、愛知淑徳大学研究助成費(共同研究)「音声コーパスを用いた促音・非促音の範疇境界を表す判別関数の検証」（平成 23～24 年度）による助成を受けた。

文　献

Amano, S., and Hirata, Y. (2010). “Perception and production boundaries between single and geminate stops in Japanese,” The Journal of the Acoustical Society of America 128, 2049-2058.

Hirata, Y., and Whiton, J. (2005). “Effects of speaking rate on the single/geminate stop distinction in Japanese,” The Journal of the Acoustical Society of America 118, 1647–1660.

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.

話し言葉が伝えるものとは、結局何なのか？
— 概念の整理および課題 —

森　大毅（宇都宮大学大学院工学研究科）

So What Should We Call the One That Spoken Language Convey?

Hiroki Mori (Utsunomiya University)

1　はじめに

話し言葉の本質は、それによって伝達される情報が言語だけではないところにある。逆の見方をすれば、話し言葉を文字言語に書き起こした時に失われる情報、それが話し言葉が持つ書き言葉にはない特質であるとも言える。

珍しく、夜遅くまで机に向かう息子。心配した家族が「なにやってんの？」と問うと、息子は「まだ宿題が終わらないんだ」と答える。これが、実は彼がまだ彼の年齢には早すぎる動画を眺めていたのを見つかった場面だったらどうだろう。家族の「なにやってんの！」の声を聞いて息子は驚き、叱られたと感じるだろう。しかし、この場面では、先の例とは違い、何をやっているかについての答えを求められているとは考えない。このように、単語列としては同一の発話内容であっても、その韻律的特徴 (ピッチ、パワー、テンポ) の違いによって伝わる情報は異なる。

ここでの問題は少なくとも 2 つある。第 1 の問題は「その情報は何なのか」であり、第 2 の問題は「その情報がどのように伝わっているか」である。第 2 の問題に関連する研究は、これまでにも盛んに行われてきた。音声科学に限ってみても、話者の年齢・性別・健康状態と声質との関係に関する研究 (粕谷 2011) や、音声から話者の感情を推定する研究、様々な発話スタイルによる音声合成の研究など数多くある。ところで、第 2 の問題は明らかに第 1 の問題をクリアしない限り成立しないはずだが、実際には「その情報は何なのか」を体系的に明らかにし、また音声研究者に広く受け入れられたものは存在しない。そこで、先の「なにやってんの！」の例ならば、伝わる情報は話者の「意図」であり、それぞれ「質問」と「叱責」の違いだ、といった具合にひとまず仮定して話が進められることになっていた。そして、研究者・研究コミュニティ間で共通の体系がない状態が続けられてきたために、後述するように、この種の研究の基本的な部分を論じるための用語および概念に深刻な混乱が見られるようになってきたのである。

国立国語研究所共同研究プロジェクト「パラ言語・非言語情報の研究における基本概念の体系化」(研究リーダー: 森　大毅) は、第 1 の問題に対する答を見出すことを目的として平成 22 年度から開始された。本発表では、本プロジェクトの研究成果の中間報告として、これまでに実施した関連研究の分析および問題点の洗い出しを行うとともに、今後の研究の方向性を示す。

2　音声が伝達する情報: 文献

2.1　藤崎の三分法

話し言葉が伝達する情報の分類は藤崎 (1996) を嚆矢とする。藤崎は、音声に含まれる情報を言語的情報・パラ言語的情報・非言語的情報の 3 つに大別した。

(藤崎 2005) より

> ··· 一方、音声は上記のような離散的な情報ばかりでなく、それ以外の情報も表現することができる。たとえば、文字による表記では同じ平叙文でも、断定／疑問／勧誘／反論など、さまざまな**意図**を込めて発音し、その意図をかなり明瞭に相手に伝えることができる。また、丁寧／ぞんざい、改まった／くだけた、などの話者の**態度**の区別を表すことができる。さらに、ゆっくり／早口、大声／小声、などの話し方（**スタイル**）を変えることにより、発話がどのような聞き手やその置かれた状況を対象としたものかを表すこともできる。··· 筆者はこの種の情報を言語的情報と区別して、パラ言語的情報と定義する。ただし、言語的情報とパラ言語的情報に共通なのは、いずれも話者が音声によって表現するべく、意識的に選択するという点である。···
>
> 音声により表現される第３の種類の情報は、たとえば話者の**個人**的な特徴や、年齢・性別・健康などの**身体的な状態**に関するもの、あるいは気質・**感情**などの心理的な状態に関するもので、特定の発話の言語的な内容とは関係なく存在し、また、一般には、話者が意識的に制御していないものである。もちろんこれには例外もあり、個人的な特徴・年齢や感情も、話者が意識的に模擬することは可能であって、いわゆる声帯模写や、演劇における感情の表現はそのよい例である。

この第 3 の種類の情報は非言語的情報と呼ばれている。藤崎の分類は我が国では強い影響力を持っており、「意図・態度はパラ言語情報」「個人性・感情は非言語情報」というのが一種のステレオタイプになっている。

本共同研究プロジェクトの課題名に「パラ言語・非言語情報」という文言が入っているのは、まさに《話し言葉が伝えるもの》を藤崎流に表現し、(自明である) 言語情報を取り去ったものに他ならない。この分類は広く受け入れられており、我々が目指す基本概念の体系化においても出発点とすべき重要なものである。しかし、研究対象をひとたび従来の音声学が対象としてきたような「きれい」で規範的なものから「きたない」自然な発話へと、さらに音声から手話を含む身体的コミュニケーションへと拡張しようとするときには、いくつかの問題が生じる。

「パラ言語情報」の範囲

Paralanguage という用語の使用は Trager (1958) まで遡ることができる。Trager が論じたのは主として音声言語を構成する非言語的 (nonverbal) 手がかり (例えば話者性、声質、発声の種類) であり、それが伝えるものについての分析的考察はあまりない。Crystal (1969) がいう paralinguistic feature は Trager よりもさらに狭く、コミュニケーションの構成要素としての音声現象 (韻律的特徴を除く) に限定されている。一方近年では、paralanguage あるいは paralinguistic feature は、それがコミュニケーションにおいて果たす機能という文脈で論じられることが普通である (Ladd 1980)。藤崎の言う paralinguistic information = パラ言語的情報は、この文脈に沿った用語法であると言える。

ただし、藤崎の分類では、感情はパラ言語情報には含まれない。藤崎の定義では、パラ言語情報は話者が意識的に選択したものでなければならないが、感情は意識的に制御したものではないからである。感情の模擬は例外だというのだが、実際の対人コミュニケーションでは、感情を出さないことも含め、感情表出はほとんどいつもある程度の制御下にあるのだから、むしろ例外の方がありふれているとも言える。感情に関連した問題は 2.2 で述べる。

近年は、paralinguistic という語が非常に広い意味で使われるようになり、言語以外のコミュニケーションチャネル全般を指す語になりつつある。例えば音声以外のノンバーバル要素 (例えばボディランゲージ) を指す場面もある (Wikipedia)。極端な例として、Interspeech 2010 Paralinguistic Challenge というコンテストにおける同定対象は、話者の年齢・性別・感情であった。

「伝えようとしたもの」vs「伝わったもの」

藤崎の分類におけるパラ言語情報は、話者が意識的に選択したものであるという点で言語情報と共通している。確かに、先に挙げた「なにやってんの」の例では、パラ言語情報は話者の発話意図の違いに応じて選択したもののように思える。

しかし、実際の対人コミュニケーションでは、発話意図を客観的に同定することは困難である。音声インタラクションへの談話行為タグの付与作業は発話意図の同定と本質的に類似しているが、実際の会話では、会話構造は聞き手の解釈によって事後的に形成されることも多い高梨 (2001)。すなわち、インタラクションは話し手だけが作るものではなく、聞き手が欠かせないということである。このことは、発話の言語的・パラ言語的特徴の違いを発話の原因に求めることを無意味にする。そこで、発話の原因よりも、その効果を記述することが妥当性を帯びる。「伝えようとしたもの」ではなく「伝わったもの」を記述しようとする態度である。

発話意図のほかにも、自然なインタラクションに見られる感情の起伏については、話者がそれを意識的に選択したものか否かを判断することは一般に難しい。多くの場合、話者自身も例外ではない。その場合、それらを藤崎の分類でパラ言語情報とみなすべきか否かを判断する材料が得られないことになる。

藤崎も含め、パラ言語情報のようなノンバーバル情報は、常に言語コミュニケーションとのアナロジーで解釈されることが多かった。パラ「言語」というネーミングが何よりの証拠である。しかし、表情・視線・体動などを考えればわかるように、話し手において産出され聞き手に伝達されるノンバーバル情報の全てが意志的 (volitional) な制御下にあるわけではない。ただ、それらは言語だけでなく、身体的コミュニケーションを形成する様々な構成要素と大なり小なりの相関を持っているというだけのことである。話し言葉が伝えるものは、意図して伝えようとしたものから、意図せず伝わったものまで、幅広い程度を有すると考えるのが適当であろう。

「非言語情報」という用語

1950 年代、Birdwhistell や Hall などによって始められたノンバーバル・コミュニケーション (nonverbal communication) 論は、《話し言葉が伝えるもの》と密接な関係がある。Vargas (1986) によれば、ノンバーバル・コミュニケーションの構成要素の 1 つが paralanguage、すなわち声によってもたらされる語以外の刺激である。

ところで、ノンバーバル・コミュニケーションはしばしば「非言語コミュニケーション」と訳される。また、藤崎の三分法における nonlinguistic information は、和訳すれば非言語情報である。前者はパラ言語情報の上位概念であり、後者はパラ言語情報とは区別されるものであるから、似たような「非言語」という用語で呼ばれているがそれらの意味は大きく異なる。混乱を避けるために、今後は nonlinguistic information に代わる用語が必要になるかもしれない。

2.2 感情

用語と定義

《話し言葉が伝えるもの》において、話し手の感情は重要な地位を占めている。感情は太古からの関心事であり、その全容の解明が簡単でないとは言え、心理学研究に分野を限ってみても、用語やそれが指す概念の混乱は深刻である。

「感情」と関連する用語に「情動」がある。心理学辞典 (1991) によれば、情動 (emotion) は急激に生起し短時間で終わる比較的強い感情であり、主観的な内的経験であるとともに行動的・運動的反応として表出され、生理的活動を伴うものとされる。一方、感情は経験の情感的あるいは情緒的な面をあらわす、より広い概念を指す語である。ところが、「感情」を自覚した感情体験 (feeling) に限る用語法もある。この場合は、心理学辞典にもあるように、情動が感情の上位概念となる。

一方、英語の場合には emotion のほか、関連する用語として affect あるいは affection が使われる。これらは類義語としても用いられるが、affect は刺激に対して良い／悪いの評価をもたらす、より直接的な反応を限定的に指すことがある。したがって、affect は positive あるいは negative という形容詞を伴うことが多い。

Izard (2010) は、35 名の著名な感情研究者に対しアンケート調査を行い、emotion の定義・機能・喚起要因・調整・認知および行動との関係・今後の研究テーマについての質問を集計した。その結果、emotion を定義づける特徴のセットに関して研究者間で共通したとみなされるものはなかった。また、「“emotion” は曖昧で科学においては位置付けが定まっていない」「研究者は “emotion” を文脈化し、何を意味するかを明確にすべきである」に対する同意の度合 (10 点満点) の平均値がそれぞれ 6.2, 8.2 と高い値を示した。

英語と日本語の用語の対応については、心理学の専門書においては emotion を情動、feeling や affection を感情とする場合が多い。狭義の affect はアフェクトと表記する場合もある。しかし、日本語の論文では感情という用語が好まれ、その訳語として emotion をあてる場合が多い。

研究の対象

心理学では感情を表情との関係で検討している場合が多く、その大多数は基本感情すなわち喜び・驚き・恐れ・悲しみ・怒り・嫌悪を対象としている。音声研究においても基本感情を前提としたものは多く、その場合、喜び・驚き … 以外は平静とか中立としてまとめられる。

これら基本感情の枠組に立つ研究では、ある一定の感情喚起刺激を呈示するか、さもなくば演技をしてもらって反応を収録することになる。これは特殊な状況であり、日常的なコミュニケーションにおける感情とは様相が異なる危険がある。日常的なコミュニケーションにおいても、社会的相互行為としての感情の表出は何らかの表示規則に従ってなされるのが普通であり、これも一種の演技による感情表出とみなすことができる。他方、中立文や無意味音節に怒りや喜びなどの感情を込めて発声した音声資料の類は、同じ演技による感情表出と言っても同列に扱うことは難しいと思われる。

自然なコミュニケーションにおける感情を研究の対象とする場合には、いわゆる「正解」の感情を付与するために、どのようなカテゴリが必要で、どのように推定すればよいのかという問題を解決しなければならない。また、感情とそれに関連した意図・態度などとの峻別も問題になる。

3　話し言葉が伝えるものの再分類

過去の研究を分析して得られた問題点を踏まえ、これまでパラ言語情報・非言語情報と呼ばれてきたものの再分類を試みた。

再分類にあたって考慮すべき特性には、メディアの性質、記号性、普遍性 (文化依存性)、原因／効果の別などがあり得る。今回は、藤崎の分類でもポイントになっていた「話者が意志をもって選択したか否か」を重視した。藤崎の分類では、感情が非言語情報となっていたが、意志を持って (volitional) 選択された感情表出と、話者の制御下にない感情表出を区別した点が異なる。また、話者のメッセージや状態が複数の側面に同時に影響を与えていることを明確にした。分類木を図 1 に示す。

新しい分類木に従って、これまで問題となっていた境界事例の分類を試みた。

- 感情
  - 話者の真正な (authentic) 感情が知覚されるときは、意志的でないので心理状態。
  - 多くの「感情音声」研究の素材は演技によるもので、意志的だからパラ言語情報。
  - 日常会話に現れる話者の感情状態は、意志的かどうか判別しがたいので分類は難しい。
  - 言語情報にも現れる。
- 性差の模倣、物真似、話者の詐称、職業口調 (駅員、物売りなど)
  - 意志的なのでパラ言語情報。
  - 言語情報にも現れる。
- 咳払い、ため息、舌打ち
  - 意志的であればパラ言語情報。そうでなければ心理状態。
- 笑い
  - 笑ってみせている場合はパラ言語情報。
  - おかしさにたまらず笑ってしまっている場合は心理状態。
- 質問
  - 終助詞などの言語形式は言語情報。
  - 上昇調音調は言語形式を取らないのでパラ言語情報。
- あいづち
  - 意志的なので言語情報またはパラ言語情報。
- フィラー
  - 意志的でないので心理状態。

4　おわりに

本発表は、共同研究プロジェクト「パラ言語・非言語情報の研究における基本概念の体系化」の中間報告である。当該研究に関わる基本概念には用語法・定義に多くの問題点があることを指摘した。

問題点に対する対応は一提案であり、研究の立場が違えばそれが新たな問題を産むこともある。今後は、立場の違いによる基本概念の捕らえ方の違いを明確にするとともに、統一が

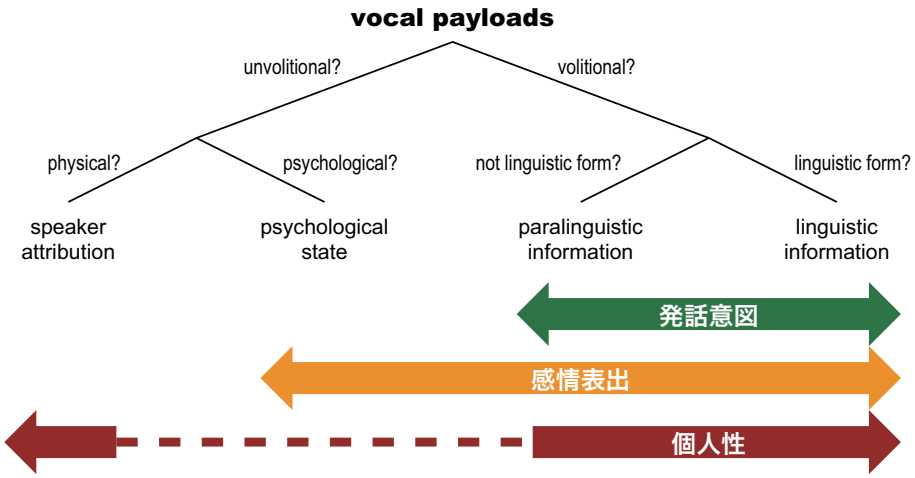


図 1: 音声が運ぶもの — 分類木

可能である部分については統一し、そうでない部分は文脈化を徹底することにより、コーパスアノテーション共通化を視野に入れた基本概念の体系化を進めて行く。

謝辞

いつも熱心にご討論いただく共同研究プロジェクトメンバーの皆様に感謝します。

参考文献

粕谷英樹、木戸博 (2011) 「声質の伝える情報とその関連量」日本音響学会秋季研究発表会講演論文集, pp. 249–252.

Fujisaki, H. (1996) “Prosody, models, and spontaneous speech” in *Computing Prosody* (Sagisaka, Y., Campbell, N. and Higuchi, N. eds.), Springer-Verlag, pp. 27–42.

藤崎博也 (2005) 「音声の音調的特徴のモデル化とその応用」文部省科学研究費特定領域研究「韻律に着目した音声言語情報処理の高度化」研究成果報告書.

Trager, G. L. (1958) “Paralanguage: A first approximation” *Studies in Linguistics*, 13, pp. 1–12.

Crystal, D. (1969) *Prosodic Systems and Intonation in English*, Cambridge University Press.

Ladd, D. R. (1980) *Structure of Intonational Meaning: Evidence from English*, Indiana University Press.

高梨克也、森本郁代 (2001) 「「孤独な」発話を救済する (1)「質問」や「言明」はどこにあるのか?」人工知能学会言語・音声理解と対話処理研究会資料, 31, pp. 49–54.

Vargas, M. F. (1986) *Louder than Words: An Introduction to Nonverbal Communication*, Iowa State University Press.

Izard, C. E. (2010) “The many meanings/aspects of emotion: Definitions, functions, activation, and regulation” *Emotion Review*, 2:4, pp. 363–370.

『日本語話し言葉コーパス』RDB の構築

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)
前川 喜久雄 (国立国語研究所言語資源研究系)

Construction of Relational Database for the *Corpus of Spontaneous Japanese*

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

1. はじめに

『日本語話し言葉コーパス』（*Corpus of Spontaneous Japanese*, 以下 CSJ）は，1999 年から 5 年間かけ，国立国語研究所・情報通信研究機構（旧通信総合研究所）・東京工業大学が共同で開発した，約 640 時間の日本語自発音声からなるデータベースである。2004 年に公開を開始して以降，音声言語情報処理，自然言語処理，日本語学，言語学，音声学，心理学，社会学，日本語教育，辞書編纂など，幅広い領域で利用されており，第 2 刷（2008 年），第 3 刷（2011 年）と順調に版を重ねている。

このように CSJ は様々な分野の研究者の関心を集める一方で，多くの（主に人文系）ユーザから，具体的にどのように利用してよいか分からないといった相談が数多く寄せられているのも事実である。CSJ には多種多様な研究用付加情報が付されており, 各種情報を統合して表現した XML 文書も提供されているが，XML を操作する技術を持たないユーザには手も足も出ないのである。

そこで筆者らは，CSJ 第 3 刷に基づき，XML 文書で表現された情報をもとに各種情報を相互に関連付けて表現した RDB を構築した。本稿ではその設計について紹介する。

2. 『日本語話し言葉コーパス』の概要

CSJ には，転記情報，文節情報，形態論情報（長単位・短単位），節単位情報，分節音情報，韻律情報，係り受け構造情報，談話境界情報，要約・重要文情報，印象評定データなど，多様な研究用付加情報（アノテーション）が付されている。ただし，これらの情報は CSJ の全体に対して斉一的に付されているわけではなく,「コア」と呼ばれるデータ範囲（約 50 万語，約 45 時間）を対象に集中的に付与されている。図 1 はコアとそれ以外における情報付与の異同の概念図である。

このように CSJ には多様な研究用付加情報が付されているが，例えば，節単位末尾の短単

† koiso@ninjal.ac.jp

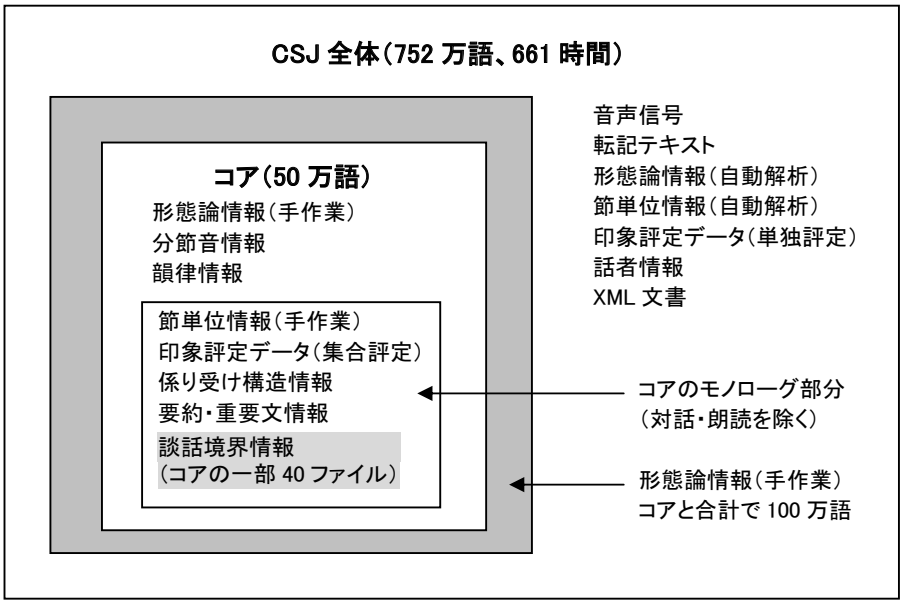


図 1　CSJ の研究用付加情報の階層構造（(国立国語研究所 2006) より）

位冒頭のモーラの時間長を知りたいといったように，複数の言語単位に関わる分析を効率的に行うためには，各種情報を相互に関連付けて表現することが求められる。また，コーパス構築の観点からは，情報間の整合性を検証・修正した上で精度の高いデータを作り上げる必要がある。そこで CSJ では，多様な情報の記述に適し，かつ，構造の妥当性を検証する仕組みを持つ XML を用いて，各種情報を統合した文書（以下，CSJ-XML）を作成し，ユーザに提供している。CSJ-XML 文書の例を図 2 に示す。

これを見ると分かるように，各種情報が多種多様なタグで表現された複雑な文書である。ここから必要な情報を効率的に抽出するには，XSLT という XML 文書の書式変換用言語を用いる必要があるが，プログラミングの経験のないユーザには敷居が高く，CSJ の豊富な研究用付加情報を高度に利用した研究の推進を阻んでいたといえる。

3.　『日本語話し言葉コーパス』RDB

この状況を踏まえ，筆者らは，図 1 のコアと呼ばれるデータ範囲を対象に，CSJ-XML で表現された情報をもとに各種情報を相互に関連付けて表現した RDB を試作した（以下，CSJ-RDB）。RDB（リレーショナルデータベース）とは，相互に関連付けられた複数のテーブルから構成されるデータベースであり，複数の言語単位に関する研究用付加情報を有する CSJ の表現に適したデータ構造である。個々の情報はテーブル，つまり行と列で構成される表の形式で記述され，直感的に把握しやすい。また，各テーブルは相互に関連付けられ，ばらばらにデータが提供される場合とは異なり，複数の言語単位に関わる検索も比較的容易に行うことができる。

3.1 では，CSJ-XML のデータ表現方式の概要を述べ，その問題点を指摘する。3.2 で CSJ-RDB における各種単位の基本構成について言及した上で，3.3 で具体的なデータベース構成について述べる。

```
<IPU IPUID="0065" IPUStartTime="00142.148" IPUEndTime="00142.524" Channel="L">
 <LUW LUWID="1" LineID="001" IsNewLine="1" LUWDictionaryForm="マタ" LUWLemma="又" LUWPOS="接続詞">
  <SUW SUWID="1" ColumnID="001" OrthographicTranscription="また"
       PlainOrthographicTranscription="また" PhoneticTranscription="マタ"
       SUWDictionaryForm="マタ" SUWLemma="又" SUWPOS="接続詞"
       ClauseUnitID="12" ClauseBoundaryLabel="&lt;接続詞&gt;" Dep_BunsetsuUnitID="0"
       SE_Subject1_50p="1" SE_Subject2_50p="1" SE_Subject3_50p="1">
   <TransSUW TransSUWID="1">
    <Mora MoraID="1" MoraEntity="マ">
     <Phoneme PhonemeID="1" PhonemeEntity="m">
      <Phone PhoneID="1" PhoneStartTime="142.184389" PhoneEndTime="142.227423"
             PhoneEntity="m" PhoneClass="consonant"/>
     </Phoneme>
     <Phoneme PhonemeID="2" PhonemeEntity="a">
      <Phone PhoneID="1" PhoneStartTime="142.227423" PhoneEndTime="142.305318"
             PhoneEntity="a" PhoneClass="vowel">
       <XJToBILabelTone Time="142.230742" ToneClass="ibt" F0="234.8590">%L</XJToBILabelTone>
      </Phone>
     </Phoneme>
    </Mora>
    <Mora MoraID="2" MoraEntity="タ">
     <Phoneme PhonemeID="1" PhonemeEntity="t">
      <Phone PhoneID="1" PhoneStartTime="142.305318" PhoneEndTime="142.343411"
             PhoneEntity="SclS" PhoneClass="others"/>
      <Phone PhoneID="2" PhoneStartTime="142.343411" PhoneEndTime="142.357172"
             PhoneEntity="t" PhoneClass="consonant"/>
     </Phoneme>
     <Phoneme PhonemeID="2" PhonemeEntity="a">
      <Phone PhoneID="1" PhoneStartTime="142.357172" PhoneEndTime="142.486984"
             PhoneEntity="a" PhoneClass="vowel">
       <XJToBILabelTone Time="142.486954" ToneClass="fbt" F0="243.6610">L%</XJToBILabelTone>
       <XJToBILabelWord Time="142.486984" PerceivedAccPos="0">mata</XJToBILabelWord>
       <XJToBILabelBreak Time="142.486984" FillerStart="1">3</XJToBILabelBreak>
      </Phone>
     </Phoneme>
    </Mora>
   </TransSUW>
  </SUW>
 </LUW>
</IPU>
```

図 2　CSJ-XML 文書の例

3.1　CSJ-XML 文書のデータ表現方式

CSJ-XML では，原則として階層構造が認められる 6 つの単位「転記基本単位（一定以上のポーズで区切られた単位)」「長単位」「短単位」「モーラ」「音素」「分節音」を設定して各種情報を表現している (国立国語研究所 2006)。しかし，ポーズを基準に認定される転記基本単位と言語的に認定される長単位はつねに階層関係を維持するわけではなく，「国立 ⟨ポーズ⟩ 国語研究所」のように，長単位内にポーズが生じて複数の転記基本単位に長単位がまたがることも少なからず存在する。このような場合，XML 表現の階層性を保つために長単位をいったんポーズで分割して表現した上で，分割されたことを示す情報を記すなど，複雑な記述となっている。

長単位よりも上位の（より長い）単位である節単位や文節では，同様の問題がより頻繁に生じる。これらは言語単位として直接的には表現されず，その境界の情報が短単位の情報に埋め込まれている。アクセント句と長単位の間にも同様の問題が生じる。このように，節単位，文節，アクセント句では言語単位が CSJ-XML 文書中で直接的に表現されないため，扱いがより複雑化する。

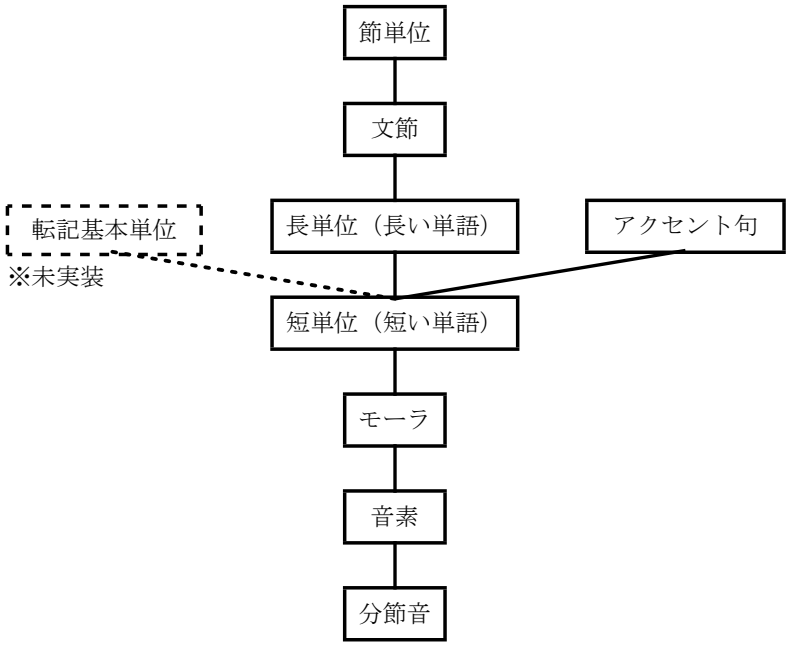


図 3　CSJ-RDB のデータ表現方式

なお CSJ では，節単位，文節，長単位，短単位といった統語形態論的な階層関係の認められる言語単位に特化した別の XML 文書も提供されているが，これとベース XML 文書にあるアクセント句の情報とを組み合わせて使うといったようなことは容易ではない。

これらの問題は，利用面での困難さを招くだけでなく，データの整合性の検証作業を過度に複雑化し，検証が不十分なままデータを提供することにつながりかねない。

3.2　CSJ-RDB のデータ表現方式

前節で指摘した問題点を踏まえ，CSJ-RDB では，CSJ-XML で採用されていた入れ子構造による階層関係の表現をやめ，多層的アノテーションの表現方法として主流となっているスタンドオフ形式による表現を採用した。この方式では，各要素の談話中での生起位置を開始時間と終了時間の対で表し，時間区間の包含関係によって階層関係を表現する。ただし，以下に述べるように，実際には利用の簡便のため，単位間の親子関係を別テーブルで陽に表現している。この結果，図 3 に示すように，節単位～短単位の系列からなる統語形態論的階層関係と，アクセント句～分節音の系列からなる韻律音韻論的階層関係を同時に表現できるようになった。

3.3　CSJ-RDB の構成

CSJ-RDB は，伝・小磯 (2012) で提案したコーパス管理用 RDB を研究用 RDB に作り変えることで構成されている。コーパス管理用 RDB では，談話中の要素を記述するセグメントと，セグメント間の関係を記述するリンクによって，アノテーションを極めて一般的に表現している。セグメントは，図 3 のように層化されており，どの層とどの層の間に親子（先祖・子孫）関係があるか指定されている。研究用 RDB では，利用の簡便を考え，層（単位）ごとに別々のテーブルでセグメントを表現している。層（単位）間の親子（先祖・子孫）関係もそれぞれ

表 1　セグメント・テーブルの個別記載情報

テーブル名	列名	説明	取りうる値 or 例
segClause （節単位）	OrthographicTranscription ClauseBoundaryLabel CU_ObligateComment	基本形 節境界ラベル 節単位義務的コメント	そこに行きましたが /並列節ガ/ 引用節構造
segBunsetsu （文節）	OrthographicTranscription	基本形	国立国語研究所では
segLUW （長単位）	OrthographicTranscription	基本形	国立国語研究所
segSUW （短単位）	OrthographicTranscription word nMorae	基本形 音素記号列 モーラ数	国語 kokugo 3
segAP （アクセント句）	OrthographicTranscription break fbt prm misc	基本形 Break Index 句末境界音調 プロミネンス 注釈情報	これが 2+bp HL% PNLP AYOR
segMora （モーラ）	MoraEntity PerceivedAcc	モーラ記号 アクセント核の有無	ユ 0/1
segPhoneme （音素）	PhonemeEntity	音素記号	kj
segPhone （分節音）	PhoneEntity PhoneClass Devoiced StartTimeUncertain EndTimeUncertain	分節音記号 分節音記号のクラス 無声化の有無 開始位置不明 終了位置不明	kj consonant 0/1 0/1 0/1
pointTone （トーン）	tone F0Uncertain CategoryUncertain PositionUncertain	トーン記号 F0 の不明確さ カテゴリーの不明確さ 位置の不明確さ	HL% 0/1 0/1 0/1

の層（単位）ごとに別々のテーブルで表現している。

RDB 全体は，セグメント・テーブル，サブセグメント・テーブル，親子関係テーブル，リンク・テーブル，メタ情報テーブルの 5 種類のテーブルから構成される。以下，各テーブルの概要について説明する。

3.3.1　セグメント・テーブル

セグメント・テーブルは，図 3 の各単位ごとに談話中の要素を記述したテーブルである。すべてのセグメント・テーブルに共通する属性として，

談話 ID，セグメントの ID，開始時間，終了時間，話者ラベル

の 5 つがある。これらによって，各セグメントの生起位置が一意に特定される。特別な場合として，トーン（アクセントや句末境界音調）のようにある瞬間に生起する（開始時間と終了時間が等しい）要素もある。

表 2　サブセグメント・テーブルの個別記載情報

テーブル名	列名	説明	値の例
subsegLUW	LUWDictionaryForm	代表形	イク
（長単位）	LUWLemma	代表表記	行く
	LUWPOS	品詞	動詞
	LUWConjugateType	活用の種類	カ行五段
	LUWConjugateForm	活用形	連用形
	LUWMiscPOSInfo1	その他情報 1	格助詞
	LUWMiscPOSInfo2	その他情報 2	促音便
	LUWMiscPOSInfo3	その他情報 3	連語
subsegSUW	PlainOrthographicTranscription	タグ無し出現形	行き
（短単位）	PhoneticTranscription	発音形	イキ
	SUWDictionaryForm	代表形	イク
	SUWLemma	代表表記	行く
	SUWPOS	品詞	動詞
	SUWConjugateType2	活用の種類 2	カ行五段 2
	SUWConjugateForm2	活用形 2	連用形 2
	SUWMiscPOSInfo1	その他情報 1	副助詞
	SUWMiscPOSInfo2	その他情報 2	語幹
	SUWMiscPOSInfo3	その他情報 3	言いよどみ
	ClauseBoundaryLabel	節境界ラベル	<テ節>
	CU_preBracket	節単位前ブラケット	<<
	CU_postBracket	節単位後ブラケット	>>
	CU_OperationSign	節単位操作記号	-
	CU_ObligateComment	節単位義務的コメント	体言止め

これらの共通情報に加えて，各単位に固有の情報が記されている。表 1 に各テーブルの個別情報を挙げる。

3.3.2　サブセグメント・テーブル

自発音声では，複数の語が融合して，分割できない一つの要素を形成することがしばしばおこる。例えば，「僕は」が融合して「ボカー」のように発音される場合である。CSJ では，これを「(W ボカー; ボクワ)」のように転記したうえで，形態論情報としては「僕」と「は」の 2 つの要素に分けて記述している。しかし，「僕」と「は」の境界は実際の音声中には存在しないため，開始・終了時間に依拠したセグメントとしては表せない。そこで CSJ-RDB では，単語の階層（長単位と短単位）は一般に，時間的に分節化できる部分をセグメントで表し，時間的に分節化できない部分はその下位にあるサブセグメントとして表している。

サブセグメント・テーブルは，

談話 ID，サブセグメントが帰属するセグメントの ID，

セグメント中のサブセグメントの位置，セグメント中のサブセグメントの総数

の 4 つの属性を共通に持つ。これらの共通情報に加えて，各単位に固有の情報が記されている。表 2 に各テーブルの個別情報を挙げる。

表 3　リンク・テーブルの記載情報

テーブル名	列名	説明	値の例
linkDepBunsetsu （文節係り受け関係）	SourceID DestinationID Dep_Label	係り文節 ID 受け文節 ID 係り受けラベル	3 10 D
linkTone2AP （トーンの帰属先）	SourceID DestinationID	トーン ID 帰属先アクセント句 ID	3 1

3.3.3　親子関係テーブル

親子関係テーブルとは，図 3 に表された階層関係に従って，セグメント間の親子（先祖・子孫）関係を ID の対で表現したものである。例えば節単位を親（先祖）とする親子関係テーブルは，文節，長単位，短単位，モーラ，音素，分節音のいずれかを子（子孫）とするものが合計 6 種類作成される。一方，節単位とアクセント句とは階層関係をなさないため，これらの間には親子関係テーブルは作成されない。

親子関係テーブルには，

談話 ID，親（先祖）セグメントの ID，子（子孫）セグメントの ID，

親セグメント中の子セグメントの位置，親セグメント中の子セグメントの総数

が共通して記されている。

セグメントに基づくアノテーション表現では，親子（先祖・子孫）関係は時間的包含関係から導出できる。しかし，CSJ-RDB では利用の簡便から，親子（先祖・子孫）関係をあらかじめ導出し，テーブルとして提供している。この親子関係テーブルを利用することで，複数の単位に関わる分析（例えば，節単位末尾の短単位冒頭のモーラの時間長）が容易に行える。

3.3.4　リンク・テーブル

セグメント間の関係としては，親子関係以外にも様々なものが考えられる。例えば，文節係り受けは文節同士の間の関係である。CSJ-RDB では，このようなセグメント間の関係をリンク・テーブルで表現している。

現在のところ，リンク・テーブルとしては，「文節係り受け関係」と「トーンの帰属先」の 2 つがある。後者は，韻律ラベルで与えられているアクセントや句末境界音調などのトーンがどのアクセント句に帰属するかを表わしたものである（トーンの生起位置が帰属先アクセント句の範囲をはみ出すことがしばしばあり，時間情報からだけでは帰属先を決められない）。

リンク・テーブルには，

談話 ID，リンク元（source）セグメントの ID，リンク先（destination）セグメントの ID

が共通して記されている。これらの共通情報に加えて，各リンクに付随する情報が記される場合もある。表 3 に各テーブルのリンク元，リンク先の内容と付随情報を挙げる。

3.3.5　メタ情報テーブル

以上のテーブル群に加え，談話や話者など言語単位以外の情報を納めたテーブルが含まれる。表 4 に示すように，談話の基本情報を納めた「談話基本情報」，話者に関する情報を記した「話者基本情報」，個々の談話の各種印象を調査した「印象評定情報」（46 の印象項目を 7 段

表 4　メタ情報テーブルの記載情報

テーブル名	列名	説明	取りうる値 or 例
infoTalk （談話基本情報）	TalkID	談話 ID	S01F0001
	TalkType	談話タイプ	独話/対話/朗読
	Genre	ジャンル	学会/模擬
	SpeakerID	話者 ID	116
	SpeakerAge	話者年齢（5 年刻み）	20to24
infoSpeaker （話者基本情報）	SpeakerID	話者 ID	116
	SpeakerSex	話者性別	男/女
	SpeakerBirthGeneration	話者生年代（5 年刻み）	70to74
	SpeakerBirthPlace	話者出生地	東京都
infoImpression （印象評定情報）	（省略）		

階で評定した結果を収めたもの）がある。

4. おわりに

筆者らは，人文系のユーザが CSJ に付された豊富な情報を使いこなせることを目標に，XML 文書と比べて直感的に構造が把握しやすく，かつ，操作技術の取得がより容易と考えられる RDB によって CSJ のデータを表現することを試みた。

昨年 11 月には，筆者らがそれぞれ代表者として関わる国立国語研究所の共同研究プロジェクトのメンバーを中心に試作版 CSJ-RDB を限定公開し，CSJ-RDB の構成，RDB から情報を抽出するための技法（SQL の書き方）などについて解説する講習会を 2 回開催した。受講者の多くは，いわゆるプログラミングの経験をほとんど持たない人文系の研究者であったが，多くの受講者から「これなら使えそう」という感想を頂いた。

CSJ-RDB を構築する過程で，CSJ 第 3 刷に含まれる様々なバグが発見された。現在，そのバグを修正しつつ，CSJ-RDB の完成に向けて作業を進めているところである。各種検証作業を経た上で，来年度を目途に CSJ 購入者に対する一般公開を予定している。

謝辞　CSJ-RDB の構築に際し，西川賢哉氏（理研），山田篤氏（ASTEM）の協力を得た。ここに記して感謝する。本研究は国立国語研究所萌芽・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），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），基幹型共同研究「コーパスアノテーションの基礎研究」（リーダー：前川喜久雄），基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）による成果である。

参考文献

国立国語研究所 (2006).『国立国語研究所報告書 124: 日本語話し言葉コーパスの構築法』.

伝康晴・小磯花絵 (2012).「RDB と既存のアノテーションツールによる統合的コーパス開発環境」　言語処理学会第 18 回年次大会発表論文集.

関連 URL

『日本語話し言葉コーパス』ホームページ：http://www.ninjal.ac.jp/csj/

シンポジウム
コーパスアノテーションと心理言語学

3月6日（火） 15:00〜16:30

コーパスアノテーションと心理言語学

浅原 正幸（国立国語研究所コーパス開発センター）†
小野 創（近畿大学理工学部）
狩野 芳伸（科学技術振興機構さきがけ）

Corpus Annotation and Psycholinguistics

Masayuki Asahara (Center for Corpus Development, NINJAL)
Hajime Ono (School of Science and Engineering, Kinki University)
Yoshinobu Kano (PRESTO, Japan Science and Technology Agency)

シンポジウム「コーパスアノテーションと心理言語学」開催趣意

言語のカタチをとらえるために数多くのコーパスが構築され共有されてきた。日本語においては『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）が完成し、昨年末より公開が始まった。今後コーパスを用いて言語運用を定量的に評価する日本語研究が盛んになることが期待される。また、言語処理研究者を中心に、コーパス上に、データの出自などのメタデータ以外に、形態論情報や統語意味論情報などのアノテーションが付与されてきた。これまでは、新聞記事など共有化できるコーパス上にアノテーションを重ね合わせてきたが、今後、他の多様なジャンルを含む BCCWJ 上に異なるレベルのアノテーションを重ね合わせ、共有されていくことだろう。

コーパスのアノテーションは、コーパスコンコーダンサなどを介して検索や統計量取得の際に利用され、言語学研究者による調査のための手がかりとして用いられる。また、言語処理研究者による言語解析器の開発のための教師ありデータとして用いられ、さらに解析手法の性能評価のためのベンチマークデータとしても用いられる。いずれの分野の研究者も、コーパスのアノテーションに対し、誤りやゆれが少ないことを求める。アノテーションに携わる者は基準を検討したり、ツールを整備したりすることにより、アノテーション誤りを少なくし、複数人のアノテータ間のゆれだけでなく 1 人のアノテータによる時間経過によるゆれを少なくする努力を続けてきた。しかしながら、この問題について決定的な解決方法は未だ提案されていない。

ここで、誤りやゆれの原因について、アノテーション過程の観点から考えてみたい。アノテーションのないコーパス（生コーパス）はテキストの原著者による言語の生成過程の産物であるが、アノテーションそのものは生コーパスのテキストを読むアノテータの受容過程の産物である。テキストの生産者（書き手）と受容者（読み手）との間に認識の齟齬があり、このことがアノテーションの誤りやゆれをもたらす。生コーパスそのものを調査することは、言語の生成過程をとらえることにほかならない。これに対して、アノテーションを調査することは言語の受容過程をとらえることであり、書き手と読み手の間にある個人の経験の差という要素に左右され誤りやゆれが必然的に入り込む。このような観点を考慮すべきである。

心理言語学の分野では、人間の心理的な観点から、人間が言語を獲得する過程、言語の生成過程と受容過程をそれぞれ研究対象としてきた。研究の方法論として、文生成課題、自己ペース読解課題、視線追跡読解課題をはじめとして、近年では脳波計を用いた事象関連電位の調査など被験者実験を中心としており、社会的に共有される言語規範だけでなく、共有されない個人の経験に基づく言語規範についても明らかにしてきた。これに対して、従前のコーパス言語学では、新聞など統制されたテキストを調査対象とし、社会的に共有されている言語規範を明らかにしてきた。今後、多様な書き手を含む BCCWJ を用いて、

† masayu-a@ninjal.ac.jp

どのような書き手が文法規範から外れた表現を生成するかを調査することにより、言語の生成過程における個人の経験の差という要素を明らかにすることができるようになろう。

ここで、コーパスアノテーションを言語の受容過程と考え、心理言語学的な観点から見つめなおすことを考えたい。コーパスアノテーションにおいて、基準やツールを用いて、社会的に共有されている言語規範については統制して誤りを減らす努力はすべきである。しかしながら、そうではない言語現象に対しては、誤りやゆれを認めたうえで、これを受容過程における個人の認識の差異として分析対象とし、定量的評価を行いたいと考える。

では、具体的にどういったことができるだろうか。ここで 2 つの方法論について考えたい。1 つは、コーパスアノテーション過程そのものを心理言語学における被験者実験ととらえ、アノテーションの誤りやゆれそのものを評価することである。基準やツールを用いても統制できない誤りやゆれについて、アノテータの作業過程で何が起こっているのかを深く検討することができるであろう。もう 1 つは、コーパスに対し心理言語実験で用いられている手法を用いて可読性などを定量評価した情報を付与し、その情報をコーパスアノテーション作業に生かすことである。適切にサンプリングされたテキストデータに対して、複数人による心理言語実験を行い、リーダビリティの低いテキストを定量的に評価して、その情報をもとにアノテーション誤りを検出することができるであろう。

言語解析器の開発の立場からは、定量的に評価された誤りやゆれの情報をどう扱えばよいだろうか。例えば、識別モデルに基づく機械学習手法を用いて、社会的に共有されている言語規範として集積されたアノテーションを正例とし、アノテータによって生成された明らかなアノテーション誤りを負例として推定することが考えられる。また、生成モデルに基づく機械学習手法を用いて、アノテータ間の受容過程のゆれを考慮して適切な分布を推定することが考えられる。

本シンポジウムでは、浅原、小野、狩野の 3 人の若手研究者より上に述べた観点から問題提起を行う。浅原が上に述べた問題意識について説明し、小野が実験言語学的な研究の方法論について概説し、狩野がコーパス言語学と心理言語学を結びつける方法について話題提供する。

登壇者プロフィール

浅原正幸（国立国語研究所コーパス開発センター特任准教授）

奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。博士（工学）。学術振興会特別研究員（DC1）、奈良先端科学技術大学院大学情報科学研究科助手、助教を経て、2012 年 1 月より現職。形態素解析器、構文解析器の開発と言語処理向けのコーパス整備に従事。現在はコーパスに対する統語意味論構造のアノテーションと言語研究向けの超大規模コーパスの設計・開発に従事。

小野創（近畿大学理工学部専任講師）

米国メリーランド大学人文学研究科博士課程修了。Ph.D（Linguistics）。広島大学特別研究員、関西外国語大学特任講師を経て、2009 年 4 月より現職。統語論を中心にした理論言語学の研究､特に日本語と英語の感嘆文の wh 依存関係に関する統語･意味論研究に従事。近年は、日本語かき混ぜ文の文処理（行動実験や脳波計測実験）、疑問詞などの wh 依存関係の処理、受身文の産出、存在文の産出と視線の関係などの研究に従事。

狩野芳伸（科学技術振興機構さきがけ研究者・国立情報学研究所外来研究員）

東京大学情報理工学系研究科博士課程単位取得退学。博士（情報理工学）。東京大学情報理工学系研究科特任研究員等を経て、2011 年 10 月より現職。言語資源（ツール・コーパス）の互換性と相互運用性の研究、相互比較や視覚化を行う統合環境の設計・開発に従事。現在はさらに大規模処理や機械学習を統合した自動化システムと、心理学的妥当性を考慮した言語処理モデルとその応用の研究に従事。

書　名　第1回 コーパス日本語学ワークショップ予稿集
発行日　平成24年3月1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）