第3回　コーパス日本語学ワークショップ　予稿集

国立国語研究所 言語資源研究系・コーパス開発センター

第3回 コーパス日本語学ワークショップ

予稿集

2013年2月28日、3月1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3回 コーパス日本語学ワークショップ

予稿集

2013年2月28日(木)／3月1日(金)

Program [プログラム]

2月28日（木）

10:00～10:10　**■挨　拶**　前川喜久雄

10:10～12:10　**■口頭発表（1）**

分類器の確信度を用いた合議制による語義曖昧性解消のunsupervisedな領域適応

▷古宮 嘉那子、奥村 学、小谷 善行

格助詞・副助詞類の連続出現パターン

▷佐藤 理史

文節係り受け構造のジャンル依存性

▷高松 亮

多様な音声表現コーパスにおける句末音調のクラスタリング

▷菊池 英明、宮島 崇浩、沈 睿

12:10～13:00　**昼食・休憩**

13:00～14:00　**■ポスター発表（1） Aグループ**

「XをYにして」における形式動詞「して」の脱落について

▷張 麗

日本語複合動詞「V 直す」、「V 返す」、「V 戻す」の特徴

▷木山 直毅

「私的なー名詞」「個人的なー名詞」の使い分け

▷渡邊 ゆかり

対訳対と協調フィルタリングを用いた商品推薦

▷柴田 翔平、古宮 嘉那子、小谷 善行

BCCWJ図書館サブコーパス全テキストへの文体情報付与結果の分析

▷柏野 和佳子、立花 幸子、保田 祥、飯田 龍、丸山 岳彦、奥村 学、佐藤 理史、徳永 健伸、大塚 裕子、佐渡島 紗織、椿本 弥生、沼田 寛

複合機能表現「という」の分類にみるMCNコーパスの方法論検証

▷叢 悠悠、田中 リベカ、中村 絢子、酒向 美帆、佐宗 智子、清水 蘭、劉 月晴、川添 愛、戸次 大介

係り受けアノテーション基準の比較

▷浅原 正幸

「結果、こういうことが言えそうです。」～コーパスにみる名詞の文副詞的用法～

▷東泉 裕子、高橋 圭子

語義曖昧性解消の領域適応のための訓練データの選択法 ～複数ドメインからの選択～

▷堀内 浩史郎、古宮 嘉那子、小谷 善行

談話構成機能からみた外来語の基本語化 ―通時的新聞コーパスを資料に―

▷金 愛蘭

機械学習による中国語助詞の用法解析

▷宋 東旭、浅原 正幸、古宮 嘉那子、小谷 善行

14:00～15:00　**■ポスター発表（1） Bグループ**

TVCMにおける和製英語のパイロット調査 ―文字テクストと音声テクストの対照を軸に―

▷小林 善久

共起語集合の頻度分布と語の属性との相関

▷山崎 誠

BCCWJ係り受け関係アノテーション付与のための文境界再認定

▷小西 光、小山田 由紀、浅原 正幸、柏野 和佳子、前川 喜久雄

書きことばにおける「語りかけ」は何のために用いられるのか

▷保田 祥、柏野 和佳子、立花 幸子、丸山 岳彦

日本語学習者のインタビュー応答時における言いよどみ使用

▷土屋 菜穂子

複数の分野のコーパスを用いた述語項構造解析の比較
―『現代日本語書き言葉均衡コーパス』を用いて―
▷吉本 暁文、小町 守、松本 裕治
「了解」の意味の変遷 ―19世紀末から現代にかけて―
▷中山 健一
CRFを用いたアニメ関連用語の固有表現抽出
▷高瀬 真記、古宮 嘉那子、小谷 善行
外来語使用における言語外的要因の分析 ―書き言葉コーパスの利用可能性―
▷久屋 愛実
国会会議録に見る複合辞の特異な形 ―丁寧形／普通形の不対応―
▷服部 匡

15:00～17:00 **■口頭発表（2）**
筑波ウェブコーパス検索ツールNLTの開発
▷今井 新悟、赤瀬川 史朗、プラシャント・パルデシ
Webを母集団とした超大規模コーパスの設計
▷浅原 正幸、前川 喜久雄
『BTSJによる日本語話し言葉コーパス（トランスクリプト・音声）2011年版』の設計と特性について
▷宇佐美 まゆみ、中俣 尚己
百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング
▷スルダノヴィッチ・イレーナ、スホメル・ヴィット、小木曽 智信、キルガリフ・アダム

17:20～19:20 **■懇親会**

3月1日（金）

10:00～12:00 **■口頭発表（3）**
中古和文における個人文体とジャンル文体
▷小林 雄一郎、小木曽 智信
洒落本コーパスの構造化 ―仕様と事例の検討―
▷市村 太郎、河瀬 彰宏、小木曽 智信
説話のパラレルコーパスの設計 ―平安・鎌倉時代の文体変異の研究に向けて―
▷田中 牧郎
「日本語歴史コーパス 平安時代編」先行公開版について
▷小木曽 智信、須永 哲矢、冨士池 優美、中村 壮範、田中 牧郎、近藤 泰弘

12:00～13:00 **昼食・休憩**

13:00～14:00 **■ポスター発表（2） Aグループ**
医学用語の選択に見られる特徴
▷金子 周司
日本語教育用の形容詞の語彙リストと難易度レベル
▷スルダノヴィッチ・イレーナ、李 在鎬
『枕草子』長単位データを用いた相の類の分析
▷冨士池 優美
接続助詞「けど」の音調と意味用法に関する予備的考察
▷田頭 未希
学習者が犯す誤用の要因・背景からみる日本語作文支援
▷八木 豊、ホドシチェク・ボル、阿辺川 武、仁科 喜久子
近代女性向け雑誌記事における一人称代名詞の分析
―形態論情報付き『近代女性雑誌コーパス』を用いて―
▷近藤 明日子

『虎明本狂言集』コーパスの構造化 ―仕様と事例の検討―

▷小林 正行、市村 太郎

自発発話におけるイントネーション句単位のF0変動の特徴

▷石本 祐一、小磯 花絵

日本語話者の英語発話にみられる日本語の音節構造と母音の無声化との関係
― Japanese AESOPコーパスの分析から

▷近藤 眞理子、鍔木 元

『日本語話し言葉コーパス』における韻律単位の認定基準について

▷小磯 花絵、前川 喜久雄、五十嵐 陽介

音声言語コーパスにおけるspeaking styleの評定と分布 ―転記テキストに着目して―

▷沈 睿、菊池 英明

14:00～15:00 **■ポスター発表（2） Bグループ**

ブラウザベースの動詞語義及び意味役割付与作業システム

▷上野 真幸、竹内 孔一

個人用コーパスの作成とアノテーションを支援する環境の実現

▷山口 昌也

『現代日本語書き言葉均衡コーパス』に対する時間表現・事象表現間の
時間的順序関係アノテーション

▷保田 祥、小西 光、浅原 正幸、今田 水穂、前川 喜久雄

『理研母子会話コーパス（R-JMICC）』構築の試みと研究成果
―対乳児自発音声における日本語特有の韻律的・分節的特徴の解明を目指して―

▷西海枝 洋子、渡辺 和希、小西 隆之、伊藤 直子、金礪 愛、五十嵐 陽介、宮澤 幸希、
西川 賢哉、馬塚 れい子

中学・高校における地学教科書の文体比較 ―学年の進行に伴う文体的特徴の変化―

▷浅石 卓真

web公開予定文法用例検索システム
「日本語文法項目用例文データベース『はごろも』」のレベル付けと学習者コーパスの比較

▷堀 恵子、江田 すみれ

日本語学習者の作文におけるエラータイプの自動分類へ向けて

▷大山 浩美、小町 守、藤野 拓也、松本 裕治

会話分析方式への転記変換におけるデータ間・個人間のゆれに関する分析

▷土屋 智行、伝 康晴、小磯 花絵

日本語における否定の焦点アノテーション

▷松吉 俊、大槻 諒、福本 文代

聞き手行動としての日本語あいづち表現の分析：
転記情報とコーディングによる発話連鎖パターンの認定

▷吉田 悦子

15:00～17:00 **■指定討論・全体討論**

17:00 **■閉　会**

Contents [目次]

口頭発表（1）

2月28日（木）　10:10～12:10

分類器の確信度を用いた合議制による語義曖昧性解消の unsupervised な領域適応

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）
小谷　善行（東京農工大学　工学研究院）

Unsupervised Domain Adaptation in Word Sense Disambiguation Based upon the Comparison of Multiple Classifiers

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

１．はじめに

テストのターゲットとなるドメインとは異なるドメインのデータ（ソースデータ）を利用して学習を行い，ターゲットドメインのデータ（ターゲットデータ）に適応することを領域適応といい，近年さまざまな手法が研究されている．

本稿では，あるドメイン（ジャンル）のターゲットデータに対して，複数のジャンルのコーパスの集合になっているソースデータがある場合，ソースデータの全体集合から，ターゲットデータに適した訓練事例の部分集合を自動的に選択する試みについて述べる．なお，ターゲットデータのラベルは未知とし，語義曖昧性解消（Word Sense Disambiguation, WSD）について領域適応を行った．また本稿では，ターゲットデータの用例ごとに適切な訓練事例は異なると仮定し，用例ごとに訓練事例の選択を行った．具体的には，あるターゲットデータに対して，二つのジャンルからなるコーパスがソースデータとして与えられた際，それぞれのジャンルのコーパスによって訓練する方式と，コーパス全体によって訓練する方式を使って三つの分類器を作成し，用例ごとに学習された分類器の出力する確信度が最大である答えを採用することにより，分類の精度を向上させる手法を示す．

２．関連研究

領域適応は，学習に使用する情報により，fully supervised，semi-supervised，unsupervised の三種に分けられる．（Daumé III, Kumar and Saha, (2010)）によれば，fully supervised の領域適応は，ラベルつきのソースデータに加え少量のラベルつきのターゲットデータを用いて学習を行うもので，訓練事例としてソースデータまたは少量のターゲットデータだけを利用する場合よりも，分類器を改良することを目指す．次の semi-supervised の領域適応は，多量なラベルつきのソースデータに加え，多量なラベルなしのターゲットデータと少量のラベルつきのターゲットデータを利用するものである．また，最後の unsupervised の領域適応は，ラベルつきのソースデータと，ラベルなしのターゲットデータを利用するものである[1]．本研究で扱うのは unsupervised の領域適応である．

領域適応の研究は自然言語処理の分野の内外においてさまざまなされており，supervised のものには（Chan and Ng (2006)），（Daumé III(2007)），（Jiang and Zhai (2007)）などがある．

本稿では，分類器の確信度により領域適応に用いる訓練事例集合を選択する手法について述べる．これに関連した研究として（張本，宮尾，辻井(2010)）や（Asch and Daelemans

†kkomiya@cc.tuat.ac.jp

[1] （Daumé III(2007)）では（Daumé III, Kumar and Saha, (2010)）で unsupervised としているものを semi-supervised としているが，本稿では新しい方を採用した．

(2010)), (McClosky, Charniak, and Johnson (2010)),（古宮, 奥村 (2012)),（Komiya and Okumura (2012)),（古宮，小谷，奥村 (2013)),がある．（張本，宮尾，辻井(2010)）は，構文解析において，分野間距離をはかり，より適切なコーパスを利用して領域適応を行えるようにした．また,（Asch and Daelemans (2010)）は，構文解析において，自動的にタグ付けされたコーパスを用いて，ソースデータとターゲットデータの類似度から性能を予測できることを示した．（古宮，奥村 (2012)）は WSD について supervised な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データとターゲットデータの性質により異なることを示し，最も効果的な領域適応手法を，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ごとに自動的に選択する手法について述べた．また,（Komiya and Okumura (2012)）は,WSD の supervised な領域適応において，本稿でも使用する確信度という尺度を用い, 用例ごとに適切な領域適応手法を自動的に選択した. また,（古宮, 小谷, 奥村 (2013)), unsupervised な領域適応において, あるターゲットデータに対して複数のジャンルのソースデータが混在した場合，確信度と LOO-bound という指標を利用して，領域適応のための訓練事例の部分集合を WSD の対象単語タイプごとに自動的に選択する手法について述べた．

３．用例ごとの訓練事例集合の自動選択

あるドメイン（ジャンル）のターゲットデータを対象に WSD を行うことを考える．このターゲットデータのラベル（語義）は未知であるとする．一方，複数のジャンルのコーパスの集合となっているソースデータが入手可能であるとすると，本稿ではこれらのソースデータの全体集合から，ターゲットデータに適した訓練事例の部分集合を自動的に選択する．この際，以下の手順で訓練事例の部分集合の選択を行う．なお，我々は最も効果的な訓練事例集合は用例ごとに異なると仮定しているため，訓練事例集合の選択はターゲットデータの用例ごとに行う．

(1) 訓練事例集合を変えて複数の分類器を学習する．
(2) 用例ごとに，複数の訓練事例集合による分類器の確信度を比較する．
(3) 分類器の確信度の最も高い訓練事例集合による結果を採用する．

ここでの分類器の確信度（Komiya and Okumura (2012)）は，分類の確からしさの度合いの予測値であり，能動学習においてラベル付けする用例を選択するのによく利用される．本手法では（Komiya and Okumura (2012)）と同様に，この確信度が確率として出力されることに注目し，確信度を比較することで，複数の分類器の合議を行う．

４．実験

４.１WSD のための訓練事例集合

WSD のための訓練事例集合として，本研究では以下に示す三つを用いる．

One : 複数のジャンルのコーパスの集合であるソースデータのうち，ひとつのジャンルのコーパスを訓練事例に用いる．
Another :「**One**」とは別のひとつのジャンルのコーパスを訓練事例に用いる．
Together :「**One**」と「**Another**」で利用したふたつのコーパスを訓練事例に用いる．

分類器としてはマルチクラス対応の SVM（libsvm）(Chang and Lin (2001))を使用した．ま

た，libsvm の確率として出力される分類の確からしさを確信度として用いた．カーネルは予備実験の結果，線形カーネルが最も高い正解率を示したため，これを採用した．また，学習の素性には，以下の 17 種類の素性を用いた．

・WSD の対象単語の前後二語までの形態素の表記（4 種類）
・WSD の対象単語の前後二語までの品詞（4 種類）
・WSD の対象単語の前後二語までの品詞の細分類（4 種類）
・WSD の対象単語の前後二語までの分類コード（4 種類）
・ 係り受け（1 種類）
  - 対象単語が名詞の場合はその名詞が係る動詞
  - 対象単語が動詞の場合はその動詞のヲ格の格要素

分類語彙表の分類コードには（国立国語研究所 (1964)）を使用した．

４.２合議の方法

上記で示した「One」，「Another」の二つ，また「Together」を含めた三つのうちから確信度を用いて，最も適切な分類器の結果を合議により決定した．「One」，「Another」の二種類から選ぶ際には，より確信度が高い方の分類器の結果を採用した．「Together」を含めた三種類の合議の方法は，以下の 4 通りを試した．なお，一番が複数あるときには最も高い確信度の分類器の語義を採用した．

・**Highest:** 最も高い確信度の分類器の結果（語義）を採用する
・**Time:** 語義ごとに，複数分類器から出力された確信度を積算し，最も高い確信度となった語義を採用する
・**Plus:** 語義ごとに，複数分類器から出力された確信度を足しあわせ，最も高い確信度となった語義を採用する
・**Majority:** 分類器ごとに，最も高い確信度となった語義に一票入れ，最も多数の票が入った語義を採用する

４.３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 コーパス）(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 et al. (1998)）の三つのジャンルのデータを利用した．これらのデータには岩波国語辞典（西尾ら (1994)）の語義が付与されている．三つのジャンルのコーパスのうち，ひとつをターゲットデータにし，残りの二つを利用可能なソースデータとして利用することで，全部で3 通りの領域適応を行った．

これらのコーパス中の多義語のうち，三つのコーパス中全てに50 トークン以上存在する単語を実験対象としたところ，全体で22種類となった．

それぞれのジャンルのコーパスにおけるケースごとの最小，最大，平均用例数を表1 に示す．

また，実験には岩波国語辞典の小分類の語義を採用した．語義数ごとの単語の内訳は，2 語義：「場合」，「自分」，3 語義：「事業」，「情報」，「地方」，「社会」，「思う」，「子供」，4 語義：「考える」，5 語義：「含む」，「技術」，6 語義：「関係」，「時間」，「一般」，「現在」，7 語義：「今」，8 語義：「前」，10 語義：「持つ」， 12 語義：「見る」，14 語義：「入る」，16 語義：「言う」，22 語義：「手」である．

表 1　それぞれのジャンルのコーパスにおける単語ごとの最小，最大，平均用例数

コーパスの種類	最小	最多	平均
BCCWJ 白書	58	7610	2240.14
BCCWJ Yahoo！知恵袋	130	13976	2741.95
RWC 新聞	56	374	183.36

５．結果

表 2 に全体の合議の方法別の実験結果を，また，表 3 にターゲットデータと合議の方法別の実験結果を示す．これらの表において，「Self」は，タグつきターゲットデータが手に入ったと仮定して，supervised の学習を 5 分割交差検定を用いて行った結果である．

「ふたつのコーパスの平均」は，ふたつのジャンルのソースデータそれぞれをジャンルごとに分けて訓練事例とした場合の結果の平均である．入手可能なジャンルのコーパスをそれぞれソースデータとして使用した場合の平均的な結果を示している．例えば，Yahoo！知恵袋のデータがターゲットデータの時のソースデータは白書と新聞であるが，このときの「ふたつのコーパスの平均」は，白書の全データで訓練した Yahoo！知恵袋のデータの正解率と，新聞の全データで訓練した Yahoo！知恵袋のデータの正解率の平均となる．

また，「大きい方のコーパス」は，ふたつのジャンルのソースデータのうち，用例数が多いジャンルのソースデータをすべて訓練事例とした場合の結果である．例えば，Yahoo！知恵袋のデータがターゲットデータの時の「大きい方のコーパス」は，白書よりも新聞のほうが全単語タイプで比較したときに用例数が多かったため，新聞の全データで訓練した Yahoo！知恵袋のデータの正解率の平均となる．

最後に，「全てのコーパス」とは，ふたつのジャンルのソースデータ全て（つまり全ソースデータ）を訓練事例とした際の結果である．例えば，Yahoo！知恵袋のデータがターゲットデータの時の「全てのコーパス」は，白書と新聞のコーパス全てを訓練事例として利用した際の結果である．

表 2　全体の合議の方法別の実験結果

	マイクロ平均	マクロ平均
Self	93.29%	85.97%
ふたつのコーパスの平均	76.92%	71.20%
大きい方のコーパス	81.99%	74.25%
全てのコーパス	81.76%	**75.86%**
二種類から選択	82.46%	74.71%
Highest	**82.62%**	74.92%
Time	77.11%	65.85%
Plus	82.48%	74.07%
Majority	80.89%	70.88%

このとき，「Self」は upper bound であり，「ふたつのコーパスの平均」，「大きい方のコーパス」，「全てのコーパス」はベースラインである．表において Self 以外でコーパスごとに

一番高い正解率を太字で示した．またその値をベースラインのうち一番目に高い正解率と比較した際，0.05 水準で有意である場合にはその値に下線を引いた．

表 3　ターゲットデータと合議の方法別の実験結果

	マイクロ平均			マクロ平均		
ターゲットデータ	白書	新聞	Yahoo！知恵袋	白書	新聞	Yahoo！知恵袋
Self	96.07%	79.57%	91.93%	91.53%	78.59%	87.80%
ふたつのコーパスの平均	73.54%	72.94%	79.95%	70.80%	71.23%	71.57%
大きい方のコーパス	80.72%	74.86%	83.50%	75.64%	74.39%	72.73%
全てのコーパス	81.80%	**75.95%**	82.11%	**76.91%**	**74.91%**	**75.76%**
二種類から選択	82.02%	74.81%	83.33%	76.68%	72.71%	74.75%
Highest	**82.28%**	74.94%	83.42%	76.88%	72.80%	75.07%
Time	76.72%	66.39%	78.13%	65.94%	62.28%	69.32%
Plus	81.93%	71.44%	83.67%	75.81%	70.65%	75.75%
Majority	80.10%	67.03%	82.46%	71.45%	67.28%	73.92%

６．考察

まず，表 2 と表 3 においてマイクロ平均を比べると，Yahoo！知恵袋コーパスがターゲットデータの時と全体で比較した際には，「全てのコーパス」の正解率より「大きい方のコーパス」の正解率の方が高い．このことから，訓練事例は必ずしも多ければ良いわけではないことが分かる．

次に，同じ二つの表から，「二種類から選択」のマイクロ平均は新聞がターゲットデータのとき以外は総じて良いことが分かる．しかし「Together」を含めた三種類から選択する「Highest」の方が，マイクロ平均，マクロ平均ともにいつも良い．

その「Highest」は，提案手法で最も高い正解率を示している．特にマイクロ平均においては，ベースライン中で最も高い正解率の「大きい方のコーパス」を有意に上回っている．しかし，マクロ平均についてはどの提案手法も「全てのコーパス」というベースラインを上回ることが出来なかった．マクロ平均をあげることが今後の課題である．

また，二つの表から，「Highest」と「Plus」は「Time」や「Majority」よりも正解率が高いことが分かる．

最後に，表 3 から，マイクロ平均において，新聞がターゲットデータになった際には「全てのコーパス」が全てのうちで最も高い正解率である．これは，訓練事例となった Yahoo！知恵袋と白書のコーパスがふたつとも大きいため，全てのコーパスを利用した場合には片方のコーパスよりずっと大きくなるためであると考えられる． 訓練事例数は必ずしも多ければいいわけではないが，一方で，訓練事例数に大きな差があった場合には，多い方を選ぶと高い正解率となると思われるので，今後は訓練事例数を加味した指標を考える予定である．

７．まとめ

テストのターゲットとなるドメインとは異なるドメインのデータを利用して学習を行い，ターゲットドメインのデータに適応することを領域適応といい，近年さまざまな手法が研究されている．我々は，語義曖昧性解消（WSD: Word Sense Disambiguation）の領域適応を行う際，ターゲットデータの用例によって適切な訓練事例集合は異なると考え，ソースデータとして二つのジャンルによるコーパスが与えられた際，それぞれのジャンルのコ

ーパスによって訓練する方式と，全体のコーパスによって訓練する方式を使って三つの分類器を作成し，用例ごとに学習された分類器の出力する確信度が最大である答えを採用することにより，分類の精度を向上させる手法を示した．用例ごとに自動的に選択された訓練事例集合を用いて領域適応を行うことで，全体のコーパスを使用して学習した時や大きい方のコーパスを利用して学習した時に比べ, WSD の平均正解率がマイクロ平均に関して有意に向上した．マクロ平均を上昇させることが今後の課題である．

謝　辞

本研究は，文部科学省科学研究費補助金[若手 B（No : 24700138）]の助成により行われた．ここに，謹んで御礼申し上げる．

文献

Vincent Van Asch and Walter Daelemans (2010). “Using Domain Similarity for Performance Estimation”. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010*, pp. 31–36.

Yee Seng Chan and Hwee Tou Ng (2006). "Estimating Class Priors in Domain Adaptation for Word Sense Disambiguation." *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp89-96.

Hal Daumé III(2007). “Frustratingly Easy Domain Adaptation.” *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp 256–263.

Hal Daumé III, Abhishek Kumar, Avishek Saha, (2010). Frustratingly Easy Semi-Supervised Domain Adaptation, Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010, pages 53–59.

Koichi Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, and Wakako Kashino (1998). “The Rwc Text Databases”. In *Proceedings of The First International Conference on Language Resource and Evaluation*, pp. 457–461.

Jing Jiang and ChengXiang Zhai (2007). “Instance Weighting for Domain Adaptation in NLP”, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp49-56, pp 264–271.

Kanako Komiya and Manabu Okumura (2012). Automatic Domain Adaptation for Word Sense Disambiguation Based on Comparison of Multiple Classifiers, *Proceedings of 26th Pacific Asia Conference on Language Information and Computation*, pp 77-85.

David McClosky, Eugene Charniak, and Mark Johnson (2010). Automatic domain adaptation for parsing. *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp 28–36.

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

国立国語研究所（1964). “分類語彙表”. 秀英出版.

古宮嘉那子, 奥村学 (2012). “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自動選択”，自然言語処理，Vol.19, No.3, pp.143-166.

古宮嘉那子, 小谷善行，奥村学 (2013). “語義曖昧性解消の領域適応のための訓練事例集合の選択”，第19回言語処理学会年次大会予稿集，In Press.

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.

張本佳子, 宮尾祐介, 辻井潤一 (2010). “構文解析の分野適応における精度低下要因の分析及び分野間距離の測定手法”. 言語処理学会　第16 回年次大会発表論文集, pp. 27–30.

格助詞・副助詞類の連続出現パターン

佐藤 理史　（名古屋大学大学院工学研究科）

Successive Patterns of Kaku- and Fuku-Joshi in Japanese

Satoshi Sato　(Graduate School of Engineering, Nagoya University)

1　はじめに

文を構成する基本ブロックとして「文節」を採用するのであれば、「どれだけの種類の文節が存在するか」を問うことは、自然な成行きである。もちろん、あらゆる可能な文節を列挙することは事実上不可能であるが、適度な抽象化と制約を持ち込めば、日本語の大部分の文節をカバーするリストを作ることは可能であろう。そのような考えのもと、我々は、いわゆる内容語を変数に抽象化した「文節パターン」を列挙することに取り組んでいる。

文節は、大きく、体言を中心とした文節と用言を中心とした文節に分けることができる。このうち、用言を中心とした文末文節は、文末機能表現シソーラスという形で、用言に付属する表現の列挙を進めてきた [1]。これに対して、本稿では、体言を中心とした文節に付属する表現—格助詞・副助詞類—に焦点を当て、それらが、どのような順序で出現しうるかを整理する。

これまで、国語学・日本語学において、助詞の研究は数多く存在する (たとえば、[2, 3, 4, 5] など)。しかしながら、研究の主眼は、助詞の分類と機能の解明にあり、助詞がどのように連続して出現するかについては、あまり注意が払われていない。一方、江副 [6] は、外国人に日本語を教える立場からこの問題に焦点を当て、「日本語の助詞は二列」という仮説を提示している。本研究では、まず、内省に基づき、格助詞・副助詞類の連続出現パターンに対する仮説を立て、その仮説を、現代日本語書き言葉均衡コーパス (BCCWJ) のコアデータを用いて検証する。

2　対象とする助詞の選定

2.1　助詞リストの作成

まず、次に示す 9 種類の文献と辞書から助詞を採取した。

1. 森田良行. 助詞・助動詞の辞典. 東京堂出版, 2007. (以下、[森田] と略記)
   目次より、70 種類の助詞を採取した。複数の箇所で参照されている場合でも、下位区分 (格助詞、副助詞等) が等しい場合は、1 つと見なした。
2. 国語教育プロジェクト編著. 原色シグマ新国語便覧 増補三訂版. 文英堂. 2012. ([シグマ])
   345 頁の表より、35 種類の助詞を採取した。
3. 国立国語研究所. 現代語の助詞・助動詞—用法と実例—. 秀英出版, 1951. ([国語研])
   第一部『助詞』より、109 種類の助詞を採取した。複数の下位分類を持つものは、下位分類のそれぞれを、1 つと見なした。
4. UniDic-2.1.0 ([UniDic])
   語彙素レベルで、助詞と定義されているもの 111 種類を採取した。
5. 益岡隆志, 田窪行則. 基礎日本語文法—改訂版—. くろしお出版, 1992. ([益岡・田窪])
   第 II 部第 9 章『助詞』に示されている 61 種類を採取した。複数の下位分類において示されているものは、それぞれを 1 つと見なした。
6. Juman-7.0 ([Juman])
   品詞が助詞である 96 種類を採取した。

表 1: 助詞の下位区分

[森田]	[シグマ]	[国語研]	[UniDic]	[益岡・田窪]	[Juman]	[JLPT]
格助詞	格助詞	格助詞	格助詞	格助詞	格助詞	格助詞 (類)
副助詞	副助詞	副助詞	副助詞	取り立て助詞	副助詞	副助詞 (類)
係助詞		係助詞	係助詞	提題助詞		
準体助詞		準体助詞	準体助詞			
並立助詞		並立助詞				
接続助詞	接続助詞	接続助詞	接続助詞	接続助詞	接続助詞	接続助詞
終助詞	終助詞	終助詞 間投助詞	終助詞	終助詞	終助詞	終助詞

7. 国際交流基金, 財団法人日本国際教育協会 編著. 日本語能力試験出題規準【改訂版】. 凡人社, 1994. ([JPLT])
   文法 3・4 級の助詞として示されている 40 種類を採取した。
8. Yoko M. McClain. Handbook of Modern Japanese Grammar. The Hokuseido Press, 1981. ([McClain])
   Particles で示されている 49 項目のうち、「か (Colloquial form of *shika*)」と「な (This *na* is not a particle)」を除く 47 種類を採取した。
9. Naoko Chino. All About Particles. Kodansha International, 2001. ([Chino])
   目次に掲載されている 69 種類を採取した。

これらのうち、同一の助詞と見なせるものを集約し、最終的に、209 種類の助詞リストを作成した。なお、これらすべてを助詞とみなすべきかどうかは、別途検討が必要である。

2.2 助詞の下位区分

[McClain] と [Chino] を除く 7 つの文献・辞書において、助詞の下位区分が導入されている[1]。それらを表 1 に示す。7 つの文献・辞書に共通な下位区分は、格助詞、接続助詞、終助詞の 3 種類である。

[益岡・田窪] では、副助詞・係助詞の類を、提題助詞と取り立て助詞に区分する。ただし、[益岡・田窪] に準拠している [Juman] では、これらの下位区分は採用されておらず、一括して、副助詞という下位区分に収められている。

副助詞と係助詞の区分を採用しているのは、[森田]、[国語研]、[UniDic] である。しかしながら、準拠関係にあると思われる [国語研] と [UniDic] の間においても、これらの下位区分が一致しないものが存在する。

以上のことより、副助詞と係助詞 (あるいは、取り立て助詞と提題助詞) の区別は、本研究では重視しない。副助詞類として一括して考える。

準体助詞という下位区分を立てるどうかは、形式名詞との関係をどう考えるかに依存する。本研究では、助詞「の」はすべて特別扱いとし、本稿の検討対象外とする。

並立助詞を立てるかどうかは、主に、接続助詞との関係が問題となり、間投助詞を立てるどうかは、終助詞との関係が問題となる。本稿の対象は、格助詞・副助詞類であるため、ここでは、これらの下位区分については立ち入らない。

2.3 対象とする 23 種類の助詞

表 2 に、本研究で対象とする 23 種類の格助詞・副助詞類を示す。これらは、9 種類の文献・辞書の過半数に収録されていた格助詞・副助詞類のうち、格助詞「の」を除いたすべてである。先に述べ

[1] [Chino] では、終助詞のみが明示的に区別されている。

表 2: 主対象とする 23 種類の助詞

		[森田]	[シグマ]	[国語研]	[UniDic]	[益岡・田窪]	[Juman]	[JLPT]	[McClain]	[Chino]
1	が	格	格	格	格助	格	格	格類 (4)	√	√
2	を	格	格	格	格助	格	格	格類 (4)	√	√
3	に	格	格	格	格助	格	格	格類 (4)	√	√
4	で	格	格	格	格助	格	格	格類 (4)	√	√
5	へ	格	格	格	格助	格	格	格類 (4)	√	√
6	と	格	格	格	格助	格	格	格類 (4)	√	√
7	より	格	格	格	格助	格	格		√	√
8	から	格	格	格	格助	格	格	格類 (4)	√	√
9	まで $_1$	格				格	格	格類 (4)	√	√
10	は	係	副	係	係助	提題/取り立て	副	副類 (4)	√	√
11	も	係	副	係	係助	取り立て	副	副類 (4)	√	√
12	でも	副	副	副		取り立て	副	副類 (4)	√	√
13	しか	係	副	係	副助	取り立て	副	副類 (4)	√	√
14	さえ	係	副	係	副助	取り立て	副		√	√
15	すら	係		係	副助	取り立て	副		√	√
16	まで $_2$	副	副	副	副助	取り立て	副	副 (3)	√	√
17	こそ	係・副	副	係	係助	取り立て	副		√	√
18	など	副	副	副	副助	取り立て	副	副類 (4)	√	√
19	のみ	副		副	副助	取り立て	副			√
20	だけ	副	副	副	副助	取り立て	副	副類 (4)	√	√
21	ばかり	副		副	副助	取り立て	副	副 (3)	√	√
22	くらい	副		副	副助	取り立て	副	副 (3)	√	√
23	ほど	副		副	副助				√	√

たように、助詞「の」は特別扱いするため、ここには含めない。

この表に示すように、「まで」は、格助詞「まで $_1$」と副助詞類「まで $_2$」が存在する。ただし、[国語研] と [UniDic] では、格助詞「まで $_1$」を認めず、副助詞類「まで $_2$」のみを認める。

[益岡・田窪] は、助詞の章において、「ほど」への言及がない。他の章には「ほど」に対する記述があるが、「ほど」の品詞については、明示的な言及はない。[益岡・田窪] に準拠する [Juman] においては、「ほど」は副詞的名詞および接尾辞として定義されている。

以上のように若干の齟齬はあるが、9 種類の文献・辞書における収録状況から判断して、これらを、格助詞および副助詞類の中核的要素と見なすことに問題はないと考える。

3 助詞の分類と出現パターン

3.1 分類の方針

すでに述べたように、本研究の最終目標は、助詞が連続して出現するパターンを列挙し、可能な文節パターンの全貌を明らかにすることである。このために重要なのは、前後の接続関係—どのような語の直後に現れうるか、直後にどのような語が現れうるか—である。助詞の分類では、機能的・意味的側面が重要視されることが多いが、ここでは、それらにできるだけ立ち入らず、直前・直後の接続関係のみで助詞を分類することを試みる。

このような方針の背景には、もう一つの理由がある。それは、現在の形態素解析が、意味を考慮せず、品詞と前後の連接関係だけから形態素認定を行なうという事実である。すなわち、意味に基づく助詞の下位区分を導入しても、それは、形態素解析においては役に立たず、かつ、そのような下位区分は、形態素解析では正しく認定できないということである。[Juman] が、提題助詞と取り立て助詞を区分せず、一括して副助詞としたのは、それらの弁別が形態素解析においては不可能であるという

認識に基づくものと推察される[2]。

3.2 A 群の助詞

まず、次の 2 つのテストを設定する。

Test1 動詞のテ形の直後に、その助詞 X は出現するか？ (「テ→ X」)
Test2 その助詞 X の直後に、「の+体言」が接続するか？(「X →の」)

前述の 23 種類の助詞のうち、これらの 2 つのテストの答がどちらも no である助詞は、「が」、「を」、「に」の 3 つである[3]。これら 3 つの格助詞は、直後に「の+体言」(連体助詞の「の」) の形をとらないという点において、他の格助詞と区別される。これらを、本稿では、**A 群の助詞**と定義する。

もちろん、次のような例を考えることはできる。

(1) a. 「歩いて」が いい。
b. 「歩いて」を 楽しむ。
c. 「歩いて」に 違いない。

しかしながら、これらの例は、引用または省略 (「歩いていくのがいい」) と見なし、通常の接続とは見なさない。

A 群の助詞は、連続して現れることはない。以降の分類は、この A 群の助詞 (特に「に」) との接続を中心に考える。

3.3 B 群の助詞

次のテストを考える。

Test3 A 群の助詞の直後に、その助詞 X は出現するか？ (「A → X」)

A 群の助詞のうち、「が」の直後には、助詞は出現しない。助詞「を」の直後には、「も」しか出現しない。それゆえ、このテストは、事実上、「「に」の直後に出現するか」を問うテストとなる。

このテストの答が no となる助詞は、次のとおりである。

(a) 「で」、「へ」、「と」、「より」、「から」、「まで $_1$(格助詞)」
(b) 「ほど」

なお、「にまで」には、次のような例があるが、この「まで」は、「まで $_2$(副助詞)」と見なす。

(2) その知らせは、学校に まで 届いた。

本稿では、上記の (a) の助詞を **B 群の助詞**と定義する。副助詞「ほど」の判断は、一旦保留とし、後で検討する。

B 群の助詞は、原則として、動詞のテ形に後続しない (Test1: no)。ただし、テ形に接続する「から」に対する判断 (格助詞とみなすか、接続助詞とみなすか) は保留とする。

(3) 書いて から 提出する

なお、「まで」には、次のような例があるが、これは B 群の助詞の出現例とはみなさない。

(4) 走って まで して、追いかける必要はない → 「まで $_2$(副助詞)」と見なす

[2] 益岡・田窪では、助詞「は」は、提題助詞と取り立て助詞の両方に区分される。
[3] テストに対する判断は、特に言及しない限り、筆者の内省に基づく。

B 群の助詞には、「の」が後続する (Test2: yes)。同時に、B 群の助詞は、自然さの度合に差はあるが、直後に A 群の助詞をとる (後述の Test4: yes)。

(5) a. 学校 から を起点にする。
b. 学校 まで をゴールとする。
c. 申請は、本人 より を基本とする。
d. ?通知は、本人 へ を優先する。
e. ?参加は、子供 と を原則とする。
f. ?試験の実施は、教室 で を原則とする。

3.4 C 群および D 群の助詞

残った助詞の Test3 の答は yes である。そこで、さらに、次のテストを考える。

Test4 その助詞 X の直後に、A 群の助詞は出現するか？ (「X → A」)

残りの助詞に対する答は、次のとおりである。

1. no — 「は」、「でも」、「しか」
2. 「が」のみが出現する — 「も」
3. yes — 「さえ」、「すら」、「まで $_2$(副助詞)」、「こそ」、「など」、「のみ」、「だけ」、「ばかり」、「くらい」

以下に示すように、「も」は「が」の直前に現れることがある。

(6) a. 誰 も が驚いた
b. 彼まで も が反対した
c. 彼女さえ も が反対した

最初の例は、「誰も」全体が一語であり、「も」はもはや助詞と見なせない可能性がある。残りの 2 例は、「までもが」と「さえもが」という形式であり、この形式は固定的である。以上のことから、「も」を、yes の仲間ではなく、no の仲間と考える。最終的に、「は」、「でも」、「しか」、「も」を **C 群の助詞**、残りの助詞を **D 群の助詞**と定義する。なお、先ほど保留とした「ほど」も、総合的に判断して、D 群の助詞に含める。(その理由は後に述べる。)

C 群の助詞は、(前述の「もが」の場合を除いて) A 群の助詞の前には現れず (Test4: no)、後ろに現れる (Test3: yes)。これに対して、D 群の助詞は、以下の例に示すように、A 群の助詞の前にも (Test4: yes)、後ろにも (Test3: yes) 現れる。(「ほど」は、例外的に、後ろには現れない。)

(7) a. 彼 だけ に伝える。
b. 彼に だけ 伝える。

D 群の助詞は、B 群の助詞の前に現れうる。

(8) a. その話は、彼 だけ から聞いた。
b. 彼 だけ へ手紙を出した。
c. 会場は、その教室 だけ で十分である。

C 群および D 群の助詞は、「くらい」、「ほど」を除き、動詞のテ形に後続する (Test1: yes)

(9) a. 書いて は いるのだが。

表 3: 助詞の分類

	Test1 テ→ X	Test3 A → X	Test2 X →の	Test4 X → A	
A 群	no	no	no	no	が、を、に
B 群	no	no	yes	yes	で、へ、と、より、から、まで $_1$
C 群	yes	yes	no*	no*	は、でも、しか、も
D 群	yes	yes*	yes	yes	さえ、すら、まで $_2$、こそ、など、 のみ、だけ、ばかり、くらい、ほど

*は例外があることを示す。

(連用修飾の場合) 体言 + D 群 + B 群 + A 群 + D 群 + C 群

(連体修飾の場合) { 体言 + D 群 + B 群 + D 群 / 数量表現 + も } + の + 体言

図 1: 格助詞・複助詞類の出現順序パターン (仮説)

b. 書いて さえ くれれば。

c. ?毎回、出席して だけ いれば、単位は楽勝だ。

C 群の助詞は、原則として、「の+体言」を後続しない (Test2: no)。「も」は、例外的に「もの+体言」の形をとるが、この例外は、数量表現に限られるようである。

(10) 100 万個 も の注文

D 群の助詞は、「の+体言」を後続する (Test2: yes)。たとえば、「すら」には、BCCWJ に次のような実例がある。

(11) 名前 すら の公表を拒み

「さえ」の実例は、BCCWJ では発見できなかったが、上記の例の「すら」と交換できる可能性が高い。

3.5 連続出現パターンに対する仮説

以上の分類結果をまとめると表 3 のようになる。この表から明らかなように、Test1 と Test3 の答、Test2 と Test4 の答は、いずれも一致する。すなわち、A 群から D 群の 4 種類に分類するだけであれば、結果的には、Test1 と Test2 の 2 つのテストで十分だったということになる。

しかし、連続出現パターンを明らかにするという観点からは、Test3 と Test4 が重要である。これらのテストと Test2 の結果より、図 1 に示すような、出現順序が予想されることとなる。

4 BCCWJ を用いた仮説検証

上記の仮説を BCCWJ のコアデータを用いて検証する。具体的には、以下の手順で行なう。

1. BCCWJ のコアデータ (解析済) から、次の条件を満たす助詞列を取り出す。
   (a) 助詞列の直前は、名詞または代名詞。
   (b) 助詞列は 2 個以上の助詞から構成されている。

表 4: 助詞の連続パターン (連体修飾の場合)

名詞	代名	列		D 群	B 群	N 群	代名詞+助詞列
662	11	BN	への		へ.B	の.N	
497	12	BN	との		と.B	の.N	
352	11	BN	での		で.B	の.N	
224	15	BN	からの		から.B	の.N	
105	65	DN	までの	まで.D	(まで.B)	の.N	これまでの
432	2	DN	などの	など.D		の.N	
30	22	DN	だけの	だけ.D		の.N	これ/それ/どれだけの
25	7	DN	ほどの	ほど.D		の.N	
14	9	DN	くらいの	くらい.D		の.N	
11	0	DN	のみの	のみ.D		の.N	
8	0	DN	ばかりの	ばかり.D		の.N	
6	0	DBN	などへの	など.D	へ.B	の.N	
5	0	DBN	などとの	など.D	と.B	の.N	
4	0	DBN	などでの	など.D	で.B	の.N	
19	20	CN	もの		*も.C	の.N	いつもの

注：*は例外を表す。括弧は別解釈の可能性を示す。

(c) 助詞列を構成する助詞は、前述の 23 種類に、格助詞「の」を加えた 24 種類とする。ただし、BCCWJ は UniDic に基づいて短単位解析されているため、格助詞「まで」と副助詞「でも」は存在しない。前者は、副助詞「まで」、後者は、格助詞「で」+係助詞「も」と解析されている。

2. 取り出した助詞列の頻度を数える。頻度は、直前が名詞の場合と、代名詞の場合に分けて集計する。

頻度を、直前が名詞の場合と、代名詞の場合に分けて集計するのは、直前が代名詞の場合は、通常とは少し異なる振舞いを示すことがあるからである。たとえば、「いつも」は、短単位辞書 UniDic では一語とみなされず、「いつ (代名詞)+も (係助詞)」と解釈される。このため、「いつもの」は、「も+の」という助詞の連続を含むことになる。

順序が逆になるが、まず、「の」を介して体言に接続する連体修飾の場合の集計結果を表 4 に示す。ここでは、直前が名詞の場合が、3 回以上観察されたものを示した。なお、「の+体言」の「の」を、便宜上、N 群の助詞とした。この表から、次のことが観察される。

- B 群と N 群の間には、D 群の助詞が入ることができるはずだが、BCCWJ のコアデータにおいては、観察されなかった。このことから、連体修飾文節においては、D 群は B 群の前に現れるのが典型的であり、後に現れるのはまれであると考えられる。
- 「までの」の「まで」は、格助詞または副助詞類のいずれの可能性も考えられる。実際に実例を調査すると、どちらの例も観察された。
- 「もの」は、C 群の助詞が現れるという点において、例外的なパターンである。しかしながら、「いつもの」以外の例は、すべて「数量表現+もの」であった。

次に、連用修飾の場合の集計結果を表 5–6 に示す。ここでも、直前が名詞の場合が、原則として、3 回以上観察されたものを示した。これらの表から、次のことが観察される。

- 前節で示した、助詞の出現順序の仮説に沿う形で、助詞は出現している。ただし、D 群、B 群+A 群、D 群+C 群の 3 つのブロックに分けて整理するのが良さそうである。これらのブロックに、2 個の助詞が入ることは、可能ではあるが、まれである。

表 5: 助詞の連続パターン (連用修飾の場合 その 1)

名詞	代名	抽出列	(並立)	D 群	B 群	A 群	D 群	C 群	(引用)
137	0	DA などが		など.D		が.A			
34	5	DA だけが		だけ.D		が.A			
16	2	BA とが	と.B			が.A			
15	3	DA こそが		こそ.D		が.A			
11	4	DA までが		まで.D	(まで.B)	が.A			
4	0	DA ばかりが		ばかり.D		が.A			
3	0	DA のみが		のみ.D		が.A			
*1	34	CA もが			*も.C	が.A			
*1	0	DCA までもが		まで.D	*も.C	が.A			
313	0	DA などを		など.D		を.A			
30	2	DA だけを		だけ.D		を.A			
14	0	BA とを	と.B			を.A			
13	1	AC をも				を.A		も.C	
8	0	DA のみを		のみ.D		を.A			
7	0	DA ばかりを		ばかり.D		を.A			
5	0	DA までを		まで.D	(まで.B)	を.A			
3	0	DA くらいを		くらい.D		を.A			
2022	114	AC には				に.A		は.C	
673	60	AC にも				に.A		も.C	
200	0	DA などに		など.D		に.A			
136	25	DA までに		まで.D	(まで.B)	に.A			
47	0	AD にまで				に.A	まで.D		
36	9	DA だけに		だけ.D		に.A			
20	2	AC にしか				に.A		しか.C	
8	0	DA のみに		のみ.D		に.A			
8	0	BA とに	と.B			に.A			
7	0	DAC などにも		など.D		に.A		も.C	
7	0	DAC までには		まで.D	(まで.B)	に.A		は.C	
7	0	AD にさえ				に.A	さえ.D		
6	1	DA くらいに		くらい.D		に.A			
6	1	DA ほどに		ほど.D		に.A			
5	0	AB にと				に.A			と.B
4	0	AD にこそ				に.A	こそ.D		
3	0	DAC などには		など.D		に.A		は.C	
3	0	AD にばかり				に.A	ばかり.D		
1675	92	BC では			で.B			は.C	
672	240	BC でも			で.B			も.C	
147	0	DB などで		など.D	で.B				
46	17	DB だけで		だけ.D	で.B				
13	2	DBC だけでは		だけ.D	で.B			は.C	
13	0	DBC などでは		など.D	で.B			は.C	
13	0	DBC などでも		など.D	で.B			も.C	
12	0	BC でしか			で.B			しか.C	
10	3	DBC だけでも		だけ.D	で.B			も.C	
9	0	DB くらいで		くらい.D	で.B				
6	0	DB のみで		のみ.D	で.B				
6	0	BD でさえ			で.B		さえ.D		
5	0	DB ほどで		ほど.D	で.B				
5	0	DB までで		まで.D	で.B				
5	1	BBC とでは	と.B		で.B			は.C	
5	0	BD ですら			で.B		すら.D		
3	0	DBC のみでは		のみ.D	で.B			は.C	
3	0	BB とで	と.B		で.B				
3	0	BD でこそ			で.B		こそ.D		

表 6: 助詞の連続パターン (連用修飾の場合 その 2)

名詞	代名	抽出列	(並立)	D 群	B 群　A 群	D 群　C 群	(引用)
78	0	BB　へと			へ.B		† と.B
8	2	BC　へは			へ.B	は.C	
7	0	BC　へも			へ.B	も.C	
3	0	DB　などへ		など.D	へ.B		
326	9	BC　とは			と.B	は.C	
87	72	BC　とも			と.B	も.C	
21	0	DB　などと		など.D	と.B		(と.B)
10	1	DB　までと		まで.D	と.B		(と.B)
7	0	BC　としか			と.B	しか.C	
4	0	DB　だけと		だけ.D	と.B		(と.B)
4	0	DB　のみと		のみ.D	と.B		(と.B)
92	33	BC　よりも			より.B	も.C	
16	0	BC　よりは			より.B	は.C	
118	22	BC　からは			から.B	は.C	
55	22	BC　からも			から.B	も.C	
33	0	DB　などから		など.D	から.B		
4	0	BB　からと			から.B		と.B
4	0	BBC からでも			から.B	† でも.C	
58	28	DC　までは		まで.D	(まで.B)	は.C	
7	28	DC　までも		まで.D	(まで.B)	も.C	
4	1	DD　くらいまで		くらい.D	† まで.B		
3	0	DC　までしか		まで.D	(まで.B)	しか.C	
69	2	DC　などは		など.D		は.C	
60	1	DC　なども		など.D		も.C	
41	4	DC　だけは		だけ.D		は.C	
13	2	DC　くらいは		くらい.D		は.C	
8	1	CB　はと				は.C	と.B
7	0	DC　くらいしか		くらい.D		しか.C	
6	1	DC　さえも		さえ.D		も.C	
4	0	DC　こそは		こそ.D		は.C	
4	0	DC　ばかりは		ばかり.D		は.C	
3	0	DC　だけしか		だけ.D		しか.C	
3	0	DD　などなど		など.D など.D			
3	1	DC　ほども		ほど.D		も.C	

注：*は例外を表す。括弧は別解釈の可能性を示す。† は注意が必要な箇所を示す。

- 「まで」を格助詞 (B 群) とすべきか、副助詞類 (D 群) とすべきかは、多くの場合、前後の接続関係だけからは決定することができない。
- 助詞「と」には、注意が必要である。格助詞「と」以外にも、並立助詞の「と」や、引用の「と」が考えられる。並立助詞や引用の「と」は、表に示したように、DBADC の並びの外にあると考えるのがよさそうである。ただし、前後の接続関係だけからでは、どの「と」であるかは、決定できない場合が存在する。
- 「へと」の「と」をどう解釈すべきかは、よくわからない。
- BA 群のブロックの後ろに出現する D 群の助詞は、直前の助詞によって限定されるようである。連体修飾文節の場合も勘案すると、D 群の助詞の標準的な位置は BA 群のブロックの前であると考えるのが自然である。つまり、A 群の後ろにしか現れないのが C 群の助詞で、A 群の前に現れるのが D 群の助詞と区別するのがよさそうである。(「ほど」を D 群に含めたのは、このような判断に基づく。)
- 「今 からでも」の「でも」は、「で (B)+も (C)」ではなく、「でも (C)」とみなすのがよさそうであるが、すべての「からでも」をそうみなしてよいかは、もっと多くの実例を観察する必要

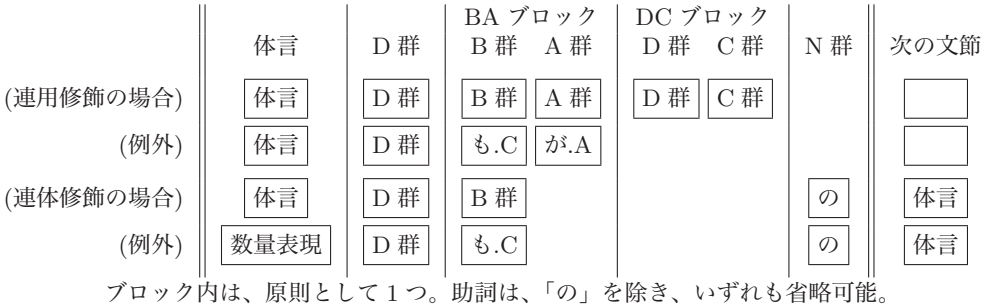


図 2: 格助詞・副助詞類の出現順序パターン (現時点での結論)

がある。

- 「もが」のほとんどは「誰もが」であるが、「大企業の 85%もが」という例が存在した。さらに、「そこに住まう人間 までもが」という例も存在した。助詞「が」は、「も」を含め、直後に副助詞類をとることができない。そのため、例外的に「も」が、「が」の直前に位置に挿入されるのではないかと考えられる。連体修飾の「もの」を含め、「もが」と「もの」は注意が必要である。

5 暫定的な結論

現時点での暫定的な結論を、図 2 に示す。ここで、「暫定的」としたのは、次の理由による。

1. 副助詞類は、動詞のテ系の直後に出現する。基本形・タ形に接続するものもある。これら、用言に接続する場合の連続出現パターンを整理する必要である。(このためには、並行して、接続助詞の整理も不可欠である。)
2. 代名詞がからむ、いくつかの問題を整理する必要がある。たとえば、「いつも」は 1 語と見なすべきか、あるいは、「いつ+も (副助詞類)」とみなすべきか。

今後は、これらの問題を検討し、格助詞・副助詞類の連続出現パターンについて、最終的な結論を得たいと考えている。

謝辞 本研究では、『現代日本語書き言葉均衡コーパス』を利用した。本研究は、JSPS 科学研究費基盤研究 (B) 「平易な日本語表現への工学的アプローチ」(課題番号 24300052) の助成を受けている。

参考文献

[1] 松木久幸, 佐藤理史, 駒谷和範. 文末機能表現シソーラスと述部正規化システム. 第 2 回コーパス日本語学ワークショップ予稿集, pp. 185–194. 国立国語研究所, 2012.
[2] 森田良行. 助詞・助動詞の辞典. 東京堂出版, 2007.
[3] 奥津敬一郎, 沼田善子, 杉本武. いわゆる日本語助詞の研究. 凡人社, 1986.
[4] 田中章夫. 助詞 (3). 岩波講座 日本語 7 文法 II. 岩波書店, 1977.
[5] 国立国語研究所. 現代語の助詞・助動詞—用法と実例—. 秀英出版, 1951.
[6] 江副隆秀. 日本語の助詞は二列. 創拓社出版, 2007.

文節係り受け構造のジャンル依存性

高松　亮（埼玉大学経済学部）†

Genre Dependencies on Phrase to Phrase Modifications of Spoken Japanese

Ryo Takamatsu (Faculty of Economics, Saitama University)

１．はじめに

本報告は，文節間の係り受け関係の構造を木構造（以下，係り受け木と呼ぶ）としてとらえた場合，その形態的特徴が発話のジャンルによってどのように変化するかを定量的に記述・分析することを試みるものである．分析の対象としては，「日本語話し言葉コーパス」（以下 CSJ）の学会講演と模擬講演の発話を用いる．

２．発話のジャンルと係り受けの構造

発話のジャンルによって，そこで用いられる文体，スタイル，レジスタといった属性は異なる．これまでさまざまな視点からこの違いを定量的に捉える試みがなされてきた（樺島(1981)，Biber and Vasquez(2008)，小磯，小木曽，小椋他(2009)）．

本報告では，意味論的な構造を反映している最もプリミティブな要素であると考えられる文節間の係り受けの構造と，発話のジャンルとの関係に注目する．

文節間の係り受けの統計的性質については，係り受け関係を有する文節間の距離が拡張された Zipf の法則に従うことを指摘した例（丸山，荻野(1992)）がある．また，金(1996)は小説に関して，書き手が変わっても係り受けの距離の分布はほとんど変化がないことを示した．しかし，係り受け木の構造とジャンルの関係は調査されていない．

いま，文節をノード，係り元の文節と係り先の文節の関係をエッジと考えたグラフ構造で表わすと，一般的には係り受けの構造を木構造，すなわち係り受け木として表現できる[1]．

個々の係り受け関係は，文節間の修飾-被修飾や原因-結果のような，意味の呼応関係であるから，係り受け木は呼応関係の構造を表現したものである．学会における講演のように，複雑な論理的構造を持つ意味内容を，正確かつ明確に伝達することが必要な場面と，日常会話のような場面とでは，発話者が意味の呼応関係の構造を場面に応じて適応的に変化させている可能性がある．係り受け木の形態を定量的に表す特徴量を観測することができれば，そのようなジャンル毎の傾向の違いが観測値に表れることが期待できる．

３．係り受け木の定義

文節をノード，係り元の文節と係り先の文節の関係をエッジと考え，係り受け関係を木構造で表したものを係り受け木と呼ぶ．以下では，係り受け木の要素をグラフ理論の用語を用いて呼ぶことがある．各文節を「ノード」，係り元がなく，係り先がある文節を「葉」，係り元はあるが係り先のない文節を「根」と呼ぶ．係り受け関係のあるノード間を結ぶ線を「エッジ」，あるノード P から根 R に向かってエッジをたどる最短経路を考えるとき，経路上のノード Q に到達するまでに経たエッジの数を「P と Q の距離」，P から根 R までの距離を「ノード P の高さ」，ある木の葉から根までの高さの最大値を「木の高さ」という．

---

† rtakamat@mail.saitama-u.ac.jp

1 本報告ではグラフが木構造にならないような場合は扱わない．また，ノードの属性に文節の出現順序を加えた順序木も考えられるが，本報告では文節の出現順序の情報を捨象した係り受け木を用いた．

４．分析対象

本報告では，文節係り受けについて手作業によるアノテーションが施されている，CSJ のコア部分について「学会講演」と「模擬講演」の 2 つの発話場面の比較を行なった（表 1）．これは，合計 6 名の話者(以下，共通話者と呼ぶ)が両方の発話場面に収録されており，それらのデータを用いれば，同一発話者の発話場面による差異をも比較可能なためである．

CSJ では係り受け構造の記述を行なう範囲として，文を認定するかわりに節単位という概念を用いている(国立国語研究所 2006)．ほとんどの場合 1 本の係り受け木は 1 個の節単位に対応する．ただし，係り元があって，係り先のない文節が節単位中に複数存在する場合もあり，その場合にはそれぞれの文節を根に持つ複数の木を考えることにする．

表 1 ：分析対象の種類と規模（括弧内は共通話者 6 名についての値）

	話者数	節単位総数	木の総本数
学会講演	70(6)	8516(790)	8723(794)
模擬講演	107(6)	9675(613)	10046(640)

５．特徴量とその傾向

５.１　はじめに

学会講演と模擬講演のデータは，年齢や性別といった話者の属性の分布が同一ではない(国立国語研究所(2006))ため，両者の統計的な性質を単純に比較するべきではないが，両者に共通の話者(共通話者)が 6 名おり，共通話者の場合と全話者の場合それぞれについて比較することで，特徴量の異同の原因がジャンルなのか母集団の違いなのかをある程度判断できる．以下では，係り受け木の形態的特徴を表現する特徴量として，木の高さのような大域的な特徴と，ある文節に対して係る文節の個数やその平均値のような局所的な特徴について分析する．なお，係り受け木のうち，係り元，係り先の両方が存在しない 1 個の文節のみからなる木は，その多くがフィラーなどであるため，分析対象から除外している．

５.２　大域的特徴

５.２.１　木の高さの頻度

係り受け木の高さの相対頻度の分布を図 1 および図 2 に示す．

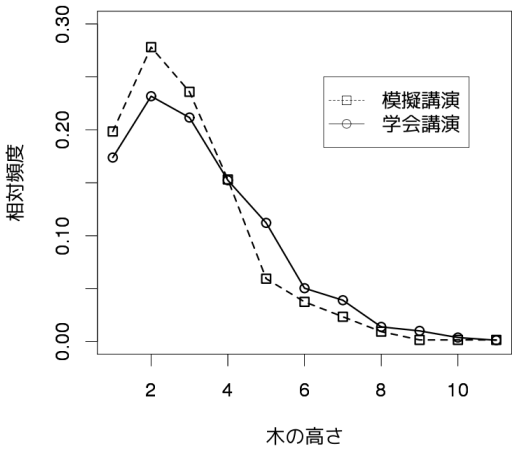


図 1 木の高さの頻度(全話者)

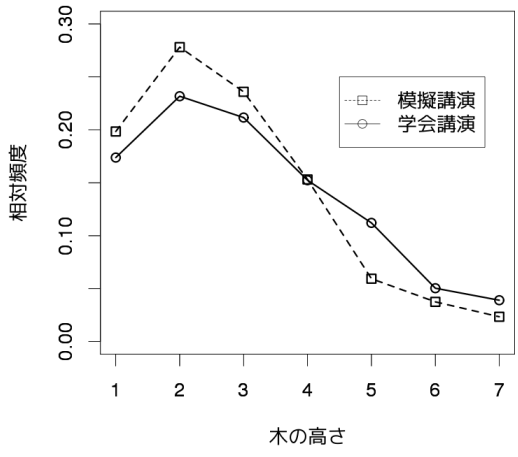


図 2 木の高さの頻度(共通話者)

図 1, 図 2 のいずれも学会講演の方が模擬講演よりも分布の幅が狭く，相対的に高い山の度数が多い．両者に共通する特徴としては，学会講演および模擬講演とも木の高さが 2 で最大値の頻度となり，それよりも木の高さが高くなるにしたがって頻度が単調に減少することがあげられる．学会講演の木の高さの平均値は 3.45(全話者)および 3.27(共通話者)，模擬講演の木の高さの平均値は 2.98(全話者)および 2.88(共通話者)であり，学会講演が模擬講演よりも木の高さの平均値が大きい．

また，全話者と共通話者の双方で同様の傾向を示すことから，学会講演と模擬講演による分布形の差異は，母集団の属性の偏りというよりは，ジャンルに起因する違いであることが推察される．

国立国語研究所(1955)においては「係り受けの次数」という，係り受け木の高さと同等のパラメータを用いて文の構造を分析しており，ニュース音声と日常的な場面における対話音声それぞれに表れる文の次数を比較した結果，ニュース音声(平均値 3.76)が対話音声(平均値 1.77)よりも次数の高い文が頻出すると指摘している(平均値は筆者による再計算)．参考のために本報告における値も含め，木の高さの値の順に並べると，

ニュース音声 > 学会講演 > 模擬講演 > 日常対話

となる．

ニュース音声は独話で，かつ改まり度が高く，本報告における学会講演に近い性質を持っている．また，本報告における模擬講演は比較的くだけた状況における独話であり，日常の対話とニュースや学会講演の中間的な性質を有すると考えられ，このことが木の高さの平均値の大小にも表れているものと考えられる．

５.２.２　文節数の頻度

1 本の係り受け木に含まれる文節の数は，木の規模の大小を表現するパラメターの一つである．図 3 および図 4 に文節数の相対頻度の分布を示す．図より，木の高さの頻度の場合と同様に，共通話者の場合も，話者全体の場合もかなり類似した傾向があることがわかる．すなわち，いずれの場合も文節数 2(図の最も左側のプロット)の頻度が例外的に高く以降単調に減少すること，文節数が 2 においては模擬講演の頻度が高く，3 から 5 程度の範囲ではその差はわずかになり，それよりも文節数が多い領域においては，逆に学会講演の方がわずかに頻度が高いことがわかる．

これらの特徴が図 3 と図 4 に共通して見られることから，文節数の頻度分布の傾向も，学会講演と模擬講演というジャンルの違いから生じていることが推察される．

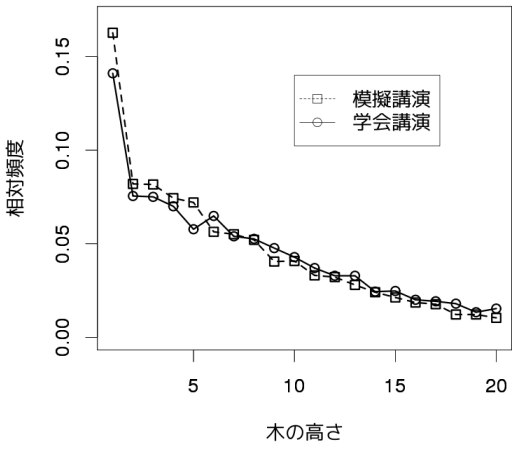


図 3 木に含まれる文節数の頻度(全話者)

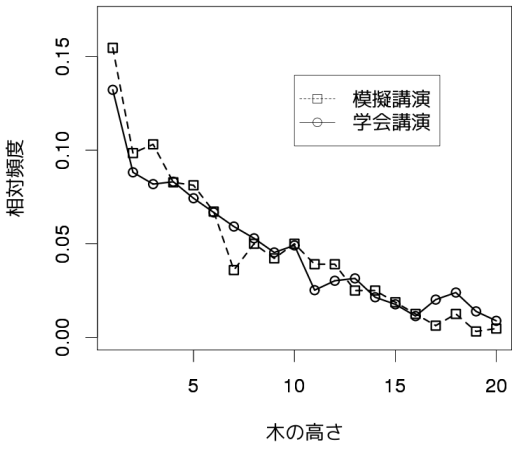


図 4 木に含まれる文節数の頻度(共通話者)

５.３　局所的特徴

５.３.１　係り元の文節数

係り受け木の局所的な特徴のうちもっとも基本的なものとして，ある文節に注目した場合に，その文節に係る文節(係り元)の数が n 個である場合の頻度を考える．図 5 および図 6 に係り元の数の相対頻度の分布を示す．なお，縦軸は相対頻度の常用対数である．

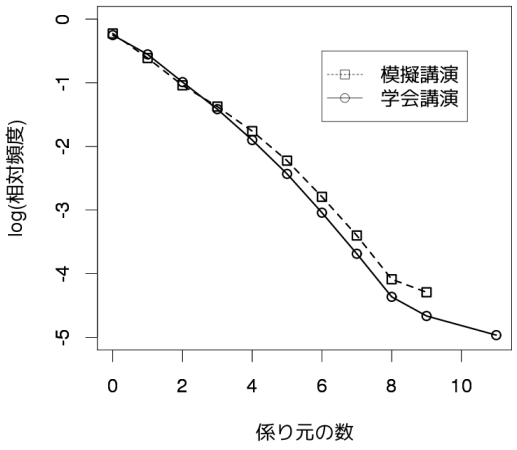


図 5 係り元の文節数の頻度(全話者)

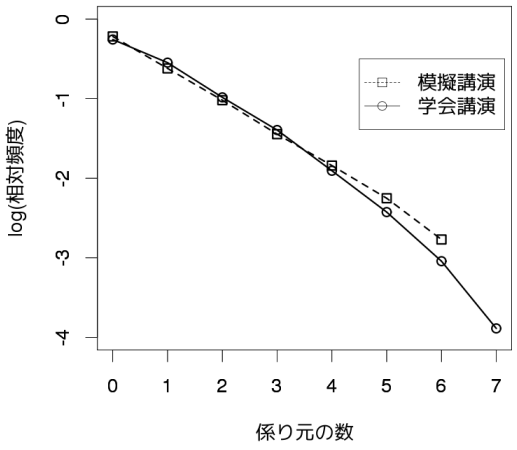


図 6 係り元の文節数の頻度(共通話者)

いずれのグラフもプロットが傾きがほぼ負の直線上にのっていること，係り元の数が 0，すなわち文節が葉である場合の相対頻度が学会講演と模擬講演とで一致すること，係り元の数が 0〜3 ないし 4 個の領域では学会講演が，それ以上の領域では模擬講演が，それぞれわずかずつ頻度が高い．共通話者と全体話者で傾向が一致することから，学会講演と模擬講演の間に見られたわずかな差異が，スタイルの差異から生じたものである可能性がある．

５.３.２　根の文節に係る文節数

根に相当する文節に，n 個の文節が係る場合の相対頻度を図 7 および図 8 に示す．

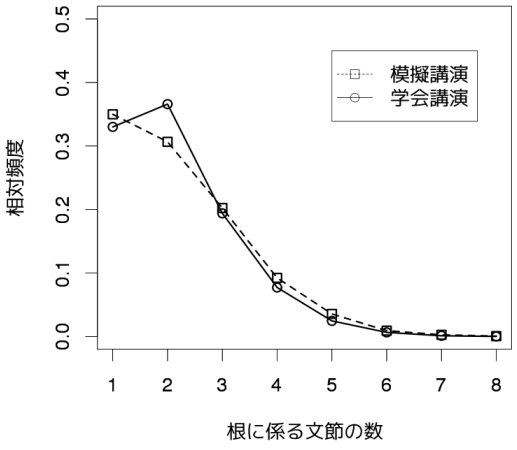


図 7 根の文節に係る文節数の頻度(全話者)

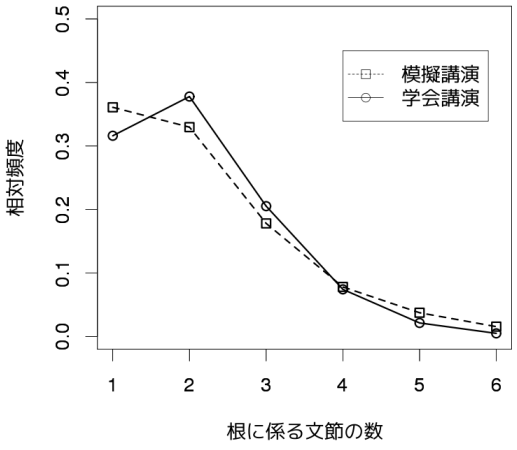


図 8 根の文節に係る文節数の頻度(共通話者)

学会講演は文節数 2 において最大値を，模擬講演は文節数 1 において最大値を取る．また，学会講演の方が分布の幅が相対的に狭い．これらの傾向が両方の図において見られることから，以上の差異が学会講演と模擬講演のスタイルの違いから生じている可能性がある．

５.３.３　葉の高さと葉の累計係り元数

ある葉の高さが n であるとき，葉から根まで辿って行く際に通過する各文節 N_i (i=1,2,..,n) が係り元を d_i 個ずつ持っているなら，d_i の合計数をその葉の累計係り元数と呼ぶことにする．葉の高さと累計係り元数の平均値の関係を図 9 および図 10 に示す．

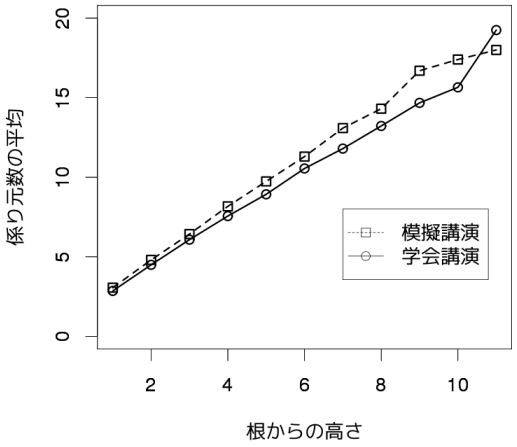


図 9 累計係り元数の平均値(全話者)

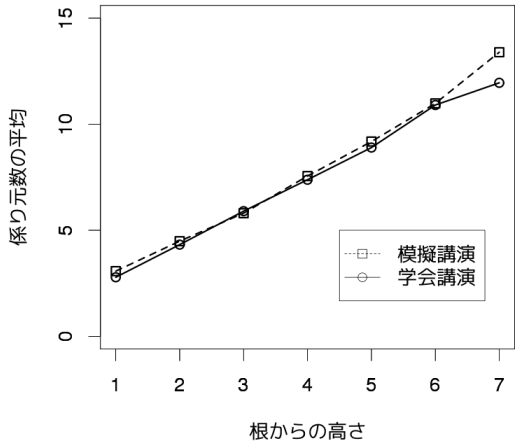


図 10 累計係り元数の平均値(共通話者)

全てに共通する特徴として，葉の高さが 1 から 6 ないし 7 程度までの範囲においては，プロットが傾きが正の直線上に良くのっていることが挙げられる．全話者においてはこの直線の傾きが模擬講演と学会講演とで異なり，学会講演の方が傾きが若干小さく，葉の高さが高くなった場合の係り元数の増加が少ない．一方，共通話者においては，学会講演の方が傾きが小さい点は全話者と同じではあるが，その差はごくわずかである．したがって，傾きの差異が発話ジャンルに起因している可能性はあるが，話者によってはそれほど明確な差が生じないことがあることがわかる．

６．まとめと今後の課題

今回得られた知見のうち，ジャンルによる量的な差異に関するものをまとめると次のようになる(A は学会講演，S は模擬講演を指す)．

- 大域的特徴
  - 木の高さの分布の平均：$A > S$
  - 木の高さの分布の幅： $A < S$
  - 文節数 2 の木の相対頻度： $S > A$
- 局所的特徴
  - 葉の相対頻度：$A=S$
  - 根に係る文節数の分布の最頻値：$A=2, S=1$
  - 根に係る文節数の分布の幅：$A < S$
  - 高さ n の葉から根までの累積係り元数：n に比例して増加(比例定数は $A < S$)

学会講演は木の高さが高く，高さの分布の散らばりも小さいこと，模擬講演は文節数が 2(すなわち高さで言えば 1)の木の相対頻度が相対的に多いことがわかる．また，高さ n の

葉から根までの累積係り元数は n にほぼ比例するが，学会講演の方が比例定数が小さいことから，葉が高い位置にあっても，根からその葉までの経路での枝分れがより少ない．以上より，学会講演は木の高さが高いが，枝分れの少ない構造を持つ傾向があると言える．

係り元を多く持つ文節ほど頻度が急速に減るが，模擬講演の方がよりロングテールな傾向を持つことから，模擬講演には 1 つの文節に多数の文節が係る表現が相対的に多いことがわかる．そのような構造の例を図 11 に示す．

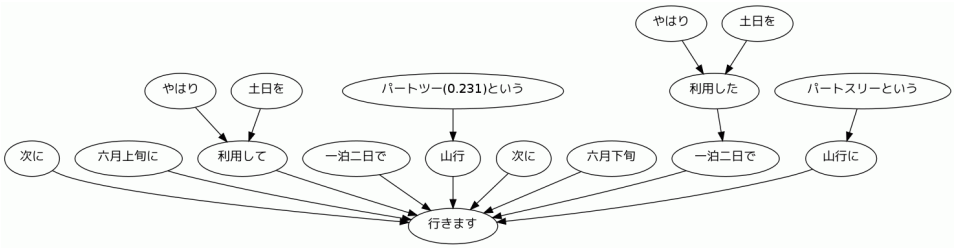


図 11 1 つの文節に多数の文節が係る構造の例

本報告では係り受け木の形態を表す特徴量として，大域的なものと局所的なものとを定義し，それらが学会講演ならびに模擬講演というスタイルの違いによってどのような傾向を持つのかについて調査し，いくつかの知見を得た．

今後の課題としてはまず，より多くの発話ジャンルについての調査を行なう必要がある．また，文節の順序関係についての情報を考慮に入れた場合，どのような傾向が見出されるかについても検討の必要がある．さらに，このような差異の傾向が見られる原因について，発話の生成過程についての知見との接続を行なう必要がある．

謝　辞

本報告でなされた研究は著者が国立国語研究所に外来研究員として滞在している際になされたものです．前川喜久雄先生，小磯花絵先生をはじめ多くの方々の御助力を頂きました．記して感謝を表します．

文　献

樺島忠夫(1981)『日本語はどう変わるか』岩波新書，岩波書店

Biber, Douglas and Cammila Vasquez (2008) "Writing and Speaking", in Handbook of research on writing, ed. C. Bazerman , pp.535--548, Routledge, Oxford, 2007

小磯花絵，小木曽智信，小椋秀樹，他(2009)「コーパスに基づく多様なジャンルの文体比較 -短単位情報に着目して-」言語処理学会 第 15 回年次大会発表論文集，pp.594-597

丸山 宏，荻野 紫穂 (1992)「日本語における文節間係り受け関係の統計的性質」情報処理学会 全国大会講演論文集，45:3, pp173-174. (http://ci.nii.ac.jp/naid/110002889591 よりダウンロード可能)

金 明哲(1993)「文節の係り受け距離の統計分析」社会情報：札幌学院大学社会情報学部紀要，5:2, pp.1-11. (http://hdl.handle.net/10742/754 よりダウンロード可能)

国立国語研究所(1955)『談話語の実態』，国立国語研究所研究報告 8 (http://db3.ninjal.ac.jp/publication_db/item.php?id=100170008 よりダウンロード可能)

国立国語研究所(2006)『日本語話し言葉コーパスの構築法』，国立国語研究所研究報告 124 (http://www.ninjal.ac.jp/csj/k-report-f/CSJ_rep.pdf よりダウンロード可能)

多様な音声表現コーパスにおける句末音調のクラスタリング

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）
沈 睿（早稲田大学 人間科学学術院）

Clustering of Boundary Tones at the Accentual Phrase Edge in the Expressive Speech Corpus

KIKUCHI Hideaki (Faculty of Human Sciences, Waseda University)
MIYAJIMA Takahiro (Faculty of Human Sciences, Waseda University)
Raymond SHEN (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)(CSJ(2011))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))。岩田ら(2012)は、対話調の演技音声資料の文末音節の F0 形状をクラスタリングし、言語学分野における分類と対応させながら、代表的な 6 種類を選定した。

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「表現豊かな音声コーパス」を構築している(菊池ら(2012b))。このコーパスには、同一の発話内容に対して多様な表現で発声された音声が多数収録されているため、話者や内容を統制した条件で句末音調の変動を分析するのに適している。菊池ら(2012a)では、句末モーラにおける F0 変動のパタンを観察して、多様な音声表現に伴う多様なパタンがあらわれていることを確認した。本稿では岩田ら(2012)と同様のクラスタリング手法を用いて表現豊かな音声の F0 変動のパタンを自動分類し、形状の類似性に基づいた分類がどのようになるかを調べた結果を報告する。

２．表現豊かな音声コーパス

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。Miyajima et al.(2011)はこれらのデータについて、「怒り」「喜び」「幸福」などの基本感情語を指示して演技者に表現を委ねる従来的な収集方法によって得られたデータとの比較を行い、物理的・心理的に多

† kikuchi@waseda.jp

様性が高いことを報告している。SEN の収集方法の詳細や多様性の検証については Miyajima et al. (2011)を参照されたい。

なお、この 100 発話には分節単位ラベルと 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．分析方法

発話末のモーラ「カ」における F0 変動のパタンをクラスタリングする。まず、F0 についてはセミトーンで話者正規化したものを 3 次の最小二乗曲線で近似する。これを始端・終端を含めた 10 点でサンプリングし、差分値を 9 次元の特徴ベクトルとしてクラスタリングした。クラスタリング方法としては Ward 法、距離の測度としてユークリッド距離を用いた。なお計算には R を利用した。

図 1 にセミトーンで話者正規化した F0 値(a)と、近似曲線(b)と、サンプリングした 10 点(c)を示す。このように目視で全ての発話について近似の妥当性を確認したところ、大きく外れたものはごく数例だけであった。無声化により F0 値が抽出できないケースや極端に短いために近似ができなかったケースを除き以下では 88 発話を分析の対象とした。図 2 に全発話の近似曲線を示す。

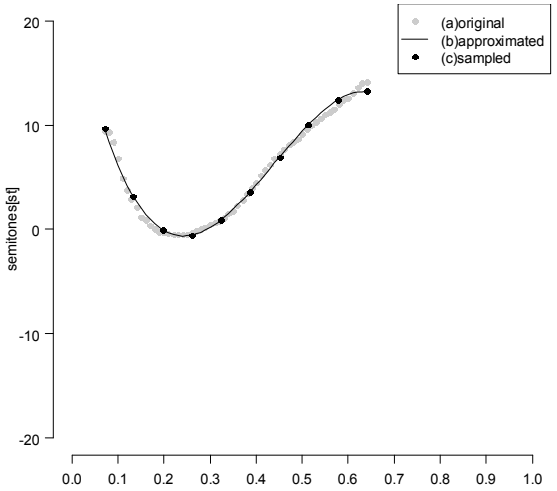


図 1　話者正規化した F0 値と近似曲線とサンプル(例)

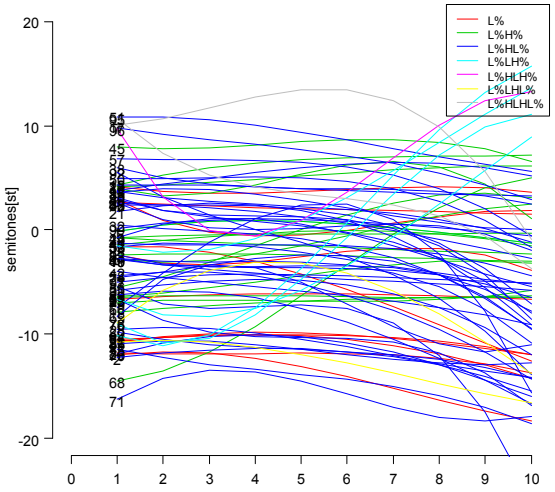


図 2　SEN の句末モーラの F0 近似曲線

４．クラスタリング結果

クラスタリング結果を図３に示す。観察しやすいように便宜上大きく 5 クラスタを認定し、それぞれにクラスタ 1〜5 と番号を与える。以下ではクラスタごとに近似曲線をプロットしてそれぞれの性質を観察する。図４にクラスタごとの近似曲線の分布を示す。

図４より、クラスタリングによって概ね F0 の形状が分離できていることがわかる。ただし、クラスタ 2 と 3 にはそれぞれ明らかに異なる形状が混在しており、下位分類(クラスタ 2A, 2B など)によって分離されている。

次に、このクラスタリング結果に基づいて元の F0 変動パタンをクラスタごとに分けて表示したものを図５に示す。これを、図６のように、人手によって付与された X-JToBI ラベルに基づいて句末境界音調(BPM: Boundary Pitch Movement)毎に観察することにより、クラスタリング結果と BPM の分類との対応関係を考察した。

岩田ら(2012)は上昇音調として疑問型上昇調と強調型上昇調を分けて扱ったが、形状を見る限り、これに相当するのがそれぞれ”LH%”と”H%”であると考えられる。”LH%”はクラスタ 4 とほぼ一対一の関係にあり、クラスタリングによってほぼ分離できているといえる。”H%”については、16 発話中 13 発話がクラスタ 2 の下位分類 2A,2B,2C に分類されている。特に下位分類 2A の 6 発話は全て”H%”の BPM が認定されており、”H%”と対応したクラスタといえる。”HL%”はいわゆる上昇下降調に相当するが、図６を見てもわかるとおり、ここにはゆるやかな下降が長いタイプと短いタイプが存在し、それぞれがクラスタ 3 とクラスタ 5 に分類されている。BPM の認定そのものにも検討が必要であるが、聴取印象の違いを調べたうえで”HL%”の下位分類の検討の必要性を示唆するものと考える。

なおその他の音調(図６の”others”)については数が少なく充分な考察ができない。今後、表現豊かな音声コーパスの資料を利用して出現頻度の少ない音調についても調査する必要がある。

５．まとめ

表現豊かな音声コーパスの一部を用いて、クラスタリング手法によって句末音調の F0 形状に基づく自動分類を行った。X-JToBI ラベルに基づく BPM の分類との対応関係を調べたところ、”LH%”と”H%”などの、クラスタと BPM との対応がよくとれる音調と、”HL%”などの、対応がとれていない音調が存在することがわかった。現在のところ、クラスタリングの特徴量として長さや高さの情報を用いていないなど、クラスタリングの精度を向上させる余地がある。また、今回は一話者の音声のみを対象としたが、ある程度の多様性は確認されているものの表現の種類には話者固有性があると考えられるため、複数話者の音声についても検討する必要がある。今後は表現豊かな音声コーパスの他のデータを用いてさらに大規模な検討を進めていく。

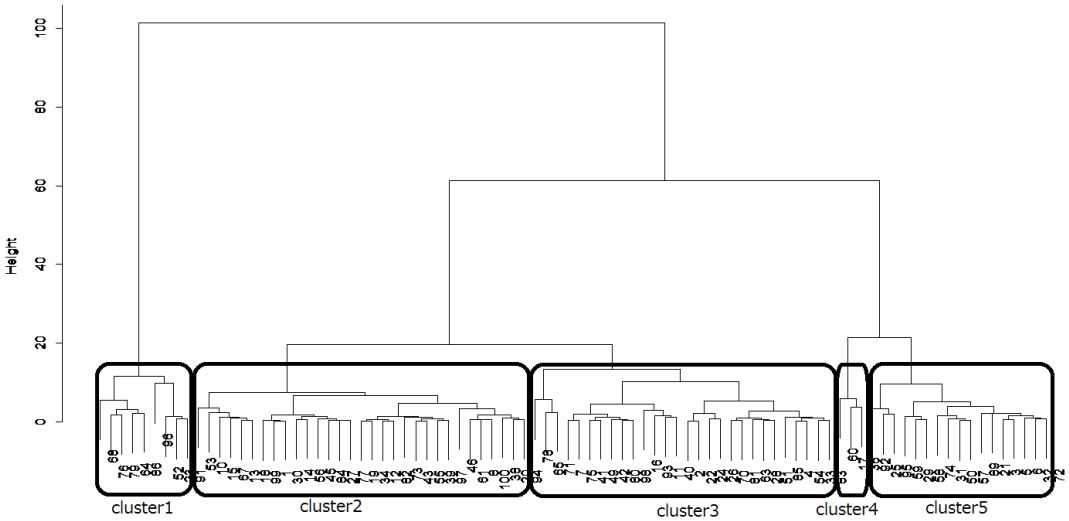


図３　クラスタリング結果
（リーフの番号は発話番号）

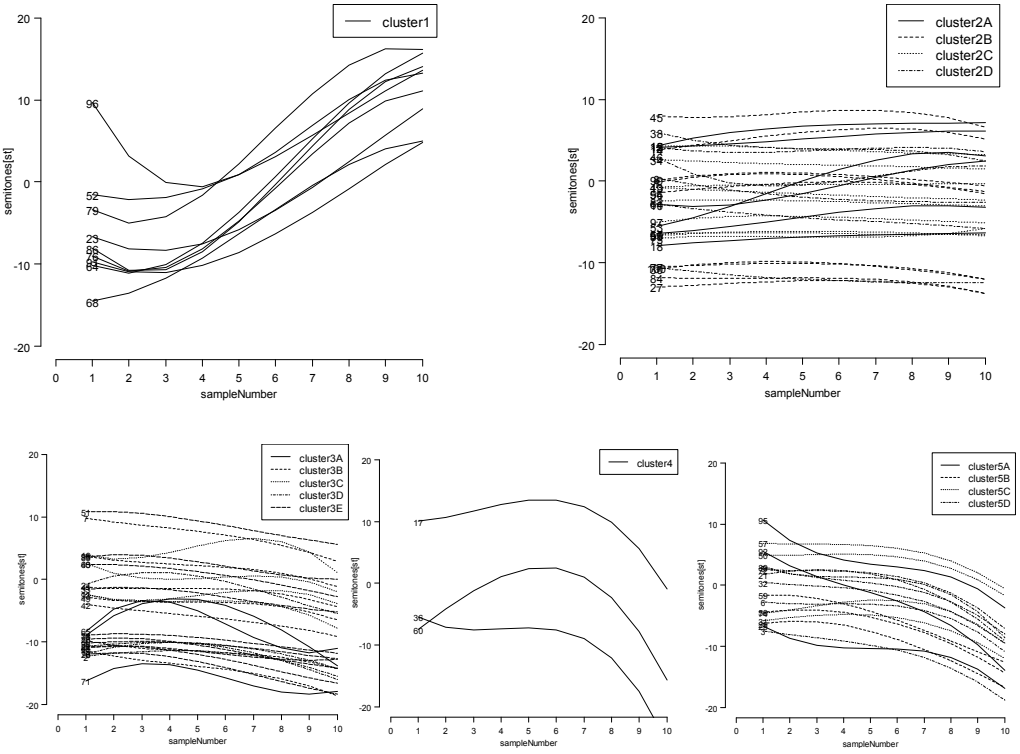


図４　クラスタごとの近似曲線の形状分布
（上段左からクラスタ 1, 2、下段がクラスタ 3,4,5。各曲線左端は発話番号。）

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.

菊池英明, 宮島崇浩 (2012a), 「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」, 第 2 回コーパス日本語学ワークショップ, pp.351-354.

菊池英明, 宮島崇浩, 前川喜久雄 (2012b), 「表現豊かな音声の収集における多様性の追求」, 日本音響学会秋季研究発表会講演論文集, Vol.1-2-16, pp.263-264.

岩田和彦, 小林哲則 (2012), 「終助詞とその音調とによって聞き手に伝わる発話意図の分析」, 電子情報通信学会技術報告, SP2012-77, pp.31-36.

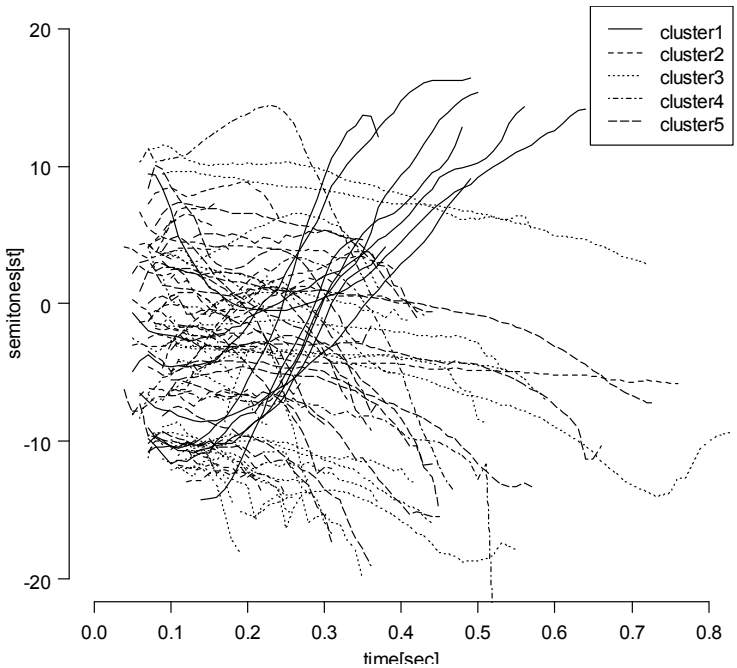


図 5　SEN の句末モーラの F0 変動

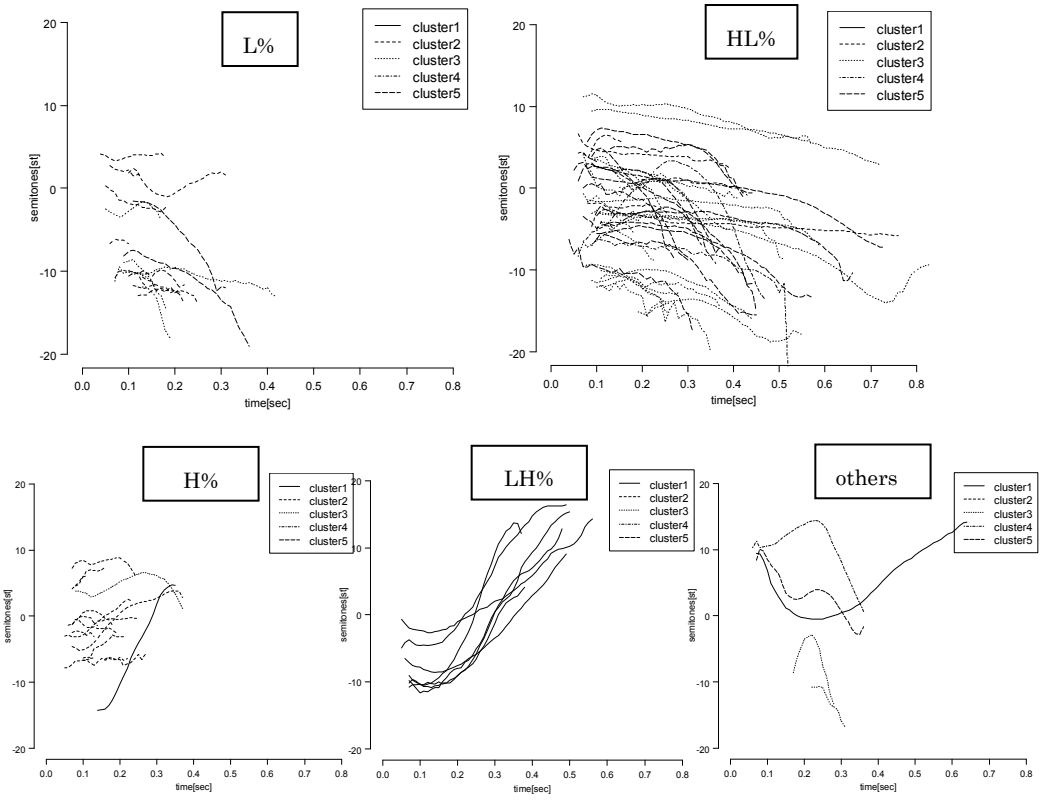


図６　人手で分類した BPM とクラスタリング結果との対応

ポスター発表（1） Aグループ

2月28日（木） 13:00～14:00

「X をYにして」における形式動詞「して」の脱落について

張麗（大東文化大学）

The Omission of the Formal Verb "*shite*" in "X *wo* Y *ni* (*shite*)" Patterns

ZhangLi（Daito Bunka University）

1. はじめに

文の成立には、主要な部分と付加的な部分がある。益岡･田窪(1992)は付帯状況・様態を表す副詞節について、次のように述べている。「付帯状況を表す副詞節は、ある動作に付随する状態や､ある動作と同時並行的に行われている付随的動作を表す。様態の副詞節は、ある動作の特定のやり方を表す。付帯状況を表す表現には、「動詞タ形＋「まま（で)」」、「動詞タ形＋「きり」」「動詞テ形」、「動詞連用形＋「ながら」」、「動詞連用形＋「つつ」」、等がある。付帯状況を表す他の表現として、「ヲ格＋「に」」の形式がある。（地図を手に目的地を探した。）」

村木（1983）では

（1） 超大国のつばぜり合がチャドとスーダンを舞台に激化している。

（2） ソ連のアフガニスタン侵攻をきっかけに米国内で防衛力増強の要望が高まった。

のような言いまわしは、

（1’）……チャドとスーダンを舞台にして……

（2’）……アフガニスタン侵攻をきっかけにして……

のように「して」を補える。（1）（2）のような言いまわしは形式的な動詞「する」の連用の形式「して」が脱落して成立したものであろうと述べている。では、どのような「X をY にして」の「して」が脱落し、「X をY に」の形になれるのか、どのような「X をY にして」の形式動詞「して」が脱落できなくて、「X をY に」に変更できないのか。「X をY に」の意味分布はどうなっているのか。「X をY にして」と「X をY に」はどちらがよく使われるのかという問題は管見のかぎりまだ明らかにされていない。

2.「X をY に」の先行研究

2.1 村木新次郎（1983）（1991）

村木（1983）（1991）は以下のように要約している。

① 「N1 を N2 に」のような言い回しは、発生的には、おそらく、形式的な動詞「する」の連用の形式「して」が省略されて成立したものであろう。「N1 を N2 にして」と「N1

を N2 に」のように「して」がついても、つかなくてもいい表現がある。「して」のつかないもののほうが、「して」のついたものよりも多い。「して」の有無によって、一般に意味の差は生じないようである。

動詞が「して」ではなくて、慣用句を構成する動詞部分が省略されて、同じタイプの表現を作ることがある。

② ＜N1 ヲ＞と＜N2 ニ＞が各々独立して（主）文の成分になることができない。また二つの名詞句の順序を入れかえることもできない。

③ ＜N1 を N2 に＞の＜N2＞は一般に連体修飾をうけることがない。

④ ＜N1 を N2＞（動詞省略）の構造をもつ表現では、副助辞-wa によって名詞句を主題化することができない。

⑤ ＜N1 を N2＞はその知的意味を変えないで、デ、カラ、ニ、トなどの格助辞や（ニ）オイテ、（ニ）ヨッテ、（ニ）対シテなどの後置詞に置き換えられることがある。

⑥ N1 と N2 が結合して複合語をつくることがある。（「子供相手に」）

⑦ ＜N1 を N2＞はひとまとまりとなって構文上の機能をはたす。状況的な成分、付帯的な状況、補語成分に分類している。

村木（1983）（1991）は相当な量の例を踏まえた上で、細かく分類し、今後の研究に非常に参考になるが、どういう条件を満たして、「X を Y に」という構文を作るのかという点は明らかにされていない。

2.2 寺村秀夫（1983）（1993）

寺村（1983）（1993）は「X ヲ Y ニ、S」という構文の成立の条件は少なくとも以下のようなものであろうと考える。

① 名詞 X、Y と S（あるいはそれを構成する主格語と述語）の間に「X が S の Y だ」という意味関係が存在することである。

② Y は、本来的に「何かの Y」であるような性格をもった名詞でなければならない。

③ S が通常その述語の主格に立つ名詞の表すものの意図した成り行き、意図的な行為を表す文であるということ。

以上の三つの条件以外に、更に、表現意図が必要であると述べている。

寺村はどういう条件が満たされれば、この種の構文が成立するかという構文条件を考察した。その構文条件は村木による分類の例にほとんど通用するが、「所持」と分類された Y が身体部分を表す場合は通用できない。更に、その構文条件を満たされても、必ず「X を Y に」という構文が成立できるとは限らない。例：「あんな女を女房にして」という例は、「あんな女を女房に」という構文は成立できないだろう。寺村は「X ヲ Y ニ、S」という構文の成立条件は少なくとも前文で述べた三つの条件であると述べているが、その三つの条件は「X を Y にして、S」の構文条件であり、「X を Y に、S」の構文条件ではないだろうと考える。どのような「X を Y にして」の形式動詞「して」が脱落して、「X を Y に」に変更できるかはまだ明らかにされていない。

2.3 奥田靖雄（1983）

奥田（1983）「を格の名詞と動詞とのくみあわせ」では、「軍人にする」、「亭主にする」、「よめにする」、「人質にする」、「犠牲にする」のような単語のくみあわせも、人をしめすを格の名詞とくみあわさって、社会的な状態変化をあらわす連語を作っていると述べている。また、奥田（1983）「ニ格の名詞と動詞とのくみあわせ」において、結果規定のむすびつきをつくる動詞のうちで一番よくつかわれるのは、「なる」と「する」とであると述べている。（例：すなわち、おおすぎる黒髪をロングカットにしている。）「X を Y にする」表現の例文は少し見えたが、詳しくは述べられていない。

2.4 金子比呂子（1990）

金子（1990）は東京外国語大学付属日本語学校の「中級日本語」に出ている「～を相手に・にして」「～を条件に・にして」「～を理由に・にして」「～を手がかりに・にして」「～を中心に・して」「～を前に・にして」「～を後に・にして」「～を片手に」という 8 つの「N1 を N2 にして」を考察してきた。考察の指標は次のようである。「N1 を N2 に（して）」の「する」は文末にきて、述語としての機能を果たすか、活用するか、意志があるか、「N1 を N2 にして」の「して」は他のどんな動詞に置き換えられるか。「N1 を N2 に」と「N1 を N2 にして」どちらも生起する場合、両者の間にどんな意味の違いがあるか。「N1 を N2 に」の「に」、または「N1 を N2 にして」の「にして」は、「として」に置き換えられるかなどである。金子は教科書に出ている例は細かく分析しているが、「N1 を N2 にして」の「して」の脱落条件には触れていない。

2.5 田中寛（2010）

田中（2010）は＜X を Y に＞形式の中で特に＜きっかけ＞を表す Y の名詞の意味範疇に注目し、下位分類の可能性を考察している。それによれば、Y にはピーク、潮、教訓、皮切り、振り出し、最後などが含まれる。＜X を Y に＞形式の様々な派生形特に連体修飾形についての指摘は「して」脱落現象の考察に一つの道筋を与えていると言えるが、ここでも「して」の脱落許容条件については詳しく言及されていない。

本論文は主に四つの問題を解決しようと思う。

(1) どのような「X を Y にして」の形式動詞「して」が脱落できなくて、「X を Y に」に変更できないのか。
(2) どのような「X を Y にして」の形式動詞「して」が脱落し、「X を Y に」に変更できるのか。
(3)「X を Y に」の意味分布はどうなっているのか。
(4)「X を Y にして」と「X を Y に」はどちらがよく使われるのか。

本論では村木氏の N1、N2 をそれぞれ X、Y としている。Y は形式名詞の場合は、今回

の調査では考察対象として扱っていない。

3. 調査方法

3.1 調査方法 1

本論文は主に「現代日本語書き言葉均衡コーパス」[1]の「少納言」を利用し、調査した。まず「にして」で検索し、ランダムに 500 件の例文が出てきた。その 500 件の「にして」がついた例文の中から、「X を Y にして（いる／おる／いた／くれる…）の例文を抽出した。更に、得た例文から、「～を Y に」で再検索した。たとえば、「にして」で検索して、「あんな女を女房にして」という用例を得た場合、「～を女房に」というキーワードで再検索し、「～を女房に」という用例が出てくるかどうかによって、「～を女房に」という表現を使えるかどうかは判断できる。「X を Y にして」は述語としての表現を考察対象外としている。コーパスでの検索で、「X を Y に」という形がない場合、念のために、Yahoo というエンジンで再検索した。そうすると、用例が大体 2 種類に分けられる。

＜a＞「X を Y にして」という形しかなく、つまり、「X を Y にして」の「して」が脱落できない用例

＜b＞「X を Y にして」と「X を Y に」両方の形が全部そろっている用例、つまり「X を Y にして」の「して」が脱落できる用例

（用例は 2012 年 11 月 2 日に検索　100 例を超えた場合、100 例に限定している。

3.2 調査方法 2

調査 1 で「X を Y にして」と「X を Y に」両方の形がそろっている場合、「X を Y にして」と「X を Y に」のそれぞれの使用数量と使用率を調べる。

4. ＜a＞「X を Y にして」の形があるが、「X を Y に」の形がない用例についての分類

＜a＞の用例を分析すれば、どのような「X を Y にして」の「して」が脱落できないのか、明らかにすることができると考える。

4.1 変える・変わる意味を表す

4.1.1 物の数量、程度、形、色、様子、状態を変える

種類 1　数量、価格、時間などを変える

（1）相手国の輸入規制を免れるため外国業者と共謀し，輸出貨物の価格等を虚偽に低価に

[1] http://www.kotonoha.gr.jp/shonagon/
このサイトでは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と文部科学省科学研究費特定領域研究「日本語コーパス」プロジェクトが共同で開発した『現代日本語書き言葉均衡コーパス』（BCCWJ:Balanced Corpus of Contemporary Written Japanese）のデータを検索できる。BCCWJ には、現代の日本語の書き言葉の全体像を把握できるように集められたサンプルが約 1 億語収録されている。

して輸出手続を行ったり，国内における業界内の取決めを逸脱するために虚偽の輸出手続を行う。

（『警察白書』（1979）昭和 54 年版　警察庁 大蔵省印刷局）

（2）用途利用米についても、平成元年度においてはその生産規模を五十万トン程度というふうにして、いろいろな面で消費拡大に努めているところでございます。

（説明員（高木勇樹君）国会会議録/参議院/常任委員会　第 114 回国会　1989）

（3）立阻止権（59 条）を弱める意味で、衆議院での再議決成立の要件を 3 分の 2 から過半数にして、衆議院の再議決がしやすくなるような改正が念頭におかれているのは想像がつく。

（白崎勇人（2003）『秘書が書く国会議員改革―国会議員を科学する常識・うんちく・論争学』　長崎出版）

種類 2　程度を変える

（4）そこへサバを入れて少し強火のまま煮てから火を中火くらいにして煮て行きます。サバには切れ目を入れてください。

（Yahoo!知恵袋/暮らしと生活ガイド/料理、グルメ、レシピ　2005）

（5）岸や内田には抵抗するようにいったが、改造を小直し程度にして、岸も国務大臣に残すとすれば、もう手がない。

（吉松安弘（1989）『東条英機暗殺の夏』新潮社）

（6）北朝鮮は意図的に資料を隠していて、日本側が弱腰だから出さない。情報を小出しにして、さらに日本から譲歩を引き出そうとしている。

（長谷川慶太郎（2004）『次の世界が見えた』徳間書店）

種類 3　形を変える

ものの形を変えたり、自分の体のある部分の形を変えたりする。

（7）以上をそれぞれ粉末にして、三年酒三升に浸し…

（永山久夫(1998)『たべもの日本史イラスト版』河出書房新社）

（8）伊勢丹新宿店ＬＵＬＵＧＵＩＮＮＥＳＳ大輪の深紅のバラを華麗な花束にして入れたようなソワレバッグ。

（実著者不明『ＢＩＳＥＳ』2004 年 6 月夏号　第 6 巻第 3 号、通巻 30 号）

種類 4　色を変えたり、色が変わったりする

（9）社長は目を伏せ、顔を真っ赤にして、恥ずかしそうに、しかし嬉しそうに笑った。

（大原健士郎(1993)『人はみな心病んで生きる―精神科医の生き方カルテ』講談社）

（10）そのときには、ツェルト（簡易テント）の中で、唇を紫色にして、わななくアキを、すっ裸にして、乾いたタオルでこすってやり、抱き合って、…

（太田蘭三（1994）『被害者の刻印』講談社）

種類 5　物事の様子を変える

（11）だが、筆者の独断では、原型はゴッホでも全面を塗り絵にして一本のゴッホ原線す

ら感ぜられず恐らく適正な補修は不可能な作品と思う故…

（大川栄二（2004）『新・美術館の窓から』財界研究所）

種類６　ものをある状態にさせたり、所持の状態や相手の状態、存在の状態を変えたりする

（ものをある状態にさせる）

（12）［ファイル検索］タブを開く２［ＵＰフォルダ内］ボタンをオン（押された状態）にして、ＵＰフォルダ内のファイルが検索されるようにする。

（実著者不明『Winny トラブル解決最終解答』アスキー　2004）

（13）皆さんはどのようにして寝てますか？扇風機を一晩中つけて寝るのとクーラーをタイマーにして明け方またつけて寝るのとどっちが経済的なんでしょう？

（Yahoo!知恵袋/暮らしと生活ガイド/家事、住宅 Yahoo!　2005）

（14）２の混合ガスを１～１０−１Ｐａ程度まで導入したのち，被処理品を陰極に，容器を陽極にして約６００Ｖの直流電圧をかけ，グロー放電を行わせる。

（渡辺敏(2004)『熱処理技術入門―金属熱処理技能士・受検テキスト』日本熱処理技術協会 日本金属熱処理工業会編著 三澤三郎編　大河出版）

（15）根元からしっかりと上下まつ毛につけます。下まつ毛はブラシを縦にして長さをプラスしていきます。

（実著者不明『JJ』2001 年 9 月号（第 27 巻第 9 号）光文社 2001）

（所持の状態）

（16）タケルはナイロンバッグを肩から斜め掛けにして、足を早めた。引き返すより、さっさと通り抜ける方を選んだ。　（松田美智子(2002)『秘密の地下室』光文社）

（相手の状態を変える）

（17）そして、ぼくは腰かけにすわり、膝の上にコドモを横抱きにして、まんべんなく、ぼくの皮膚にムスメの皮膚がくっつくように揺すりうごかす。

（西成彦(1992)『パパはごきげんななめ』集英社）

4.1.2　ある基準によって、分類、整理する

（18）猶予人員を５歳刻みの年齢層に分けて見たものがⅣ－８表であり，さらにこれらを構成比にして見たものがⅣ－１６図であって，１年齢層別に見た起訴猶予率は両年ともほとんど変わらない。

（『犯罪白書』平成 3 年版　法務省法務総合研究所大蔵省印刷局 1991）

4.1.3 抽象的なことに変える

（19）市民を互いに遠ざけておき、その相互のコミュニケーションを困難なものにし、彼らが危険なしには集まれなくすることである

（ツヴェタン・トドロフ(著)/ 小野潮(訳)（2003）『バンジャマン・コンスタン―民主主義への情熱』法政大学出版局）

4.1.4 親族名詞、職業、活動を表す名詞につけて、社会的な意味を付与する

種類１　親族名詞

（20）目つきがこう、色っぽくて寒気がするほどだ。あんな女を女房にして、高級車を乗り回して、政治家どもを足元に這いつくばらせるのが俺の夢さ。

（鶴田楡（2004）『ダンス・ウィズ・キャット 下』新風舎）

種類２　職業

（21）本邦初の腑分けをおこなったおなじく古方派山脇東洋の流れの山脇某、不遇時代に按摩を生業にして身を起こした賀川流産科の祖賀川玄悦の流れの賀川某など、

（佐藤雅美（2003）『啓順地獄旅』講談社）

種類３　活動

（22）できれば柿渋を塗るメインテナンスを家族の年中行事にして楽しむのも。雑草に悩まされずに雨水を土に返すこともできる。

（三澤文子(著)/ 実著者不明『ミセス』　2001 年 8 月号（通巻第 558 号）文化出版局）

以上から、「X を Y にする」という文型は変える・変わる意味を表す場合、形式動詞「して」の脱落は制限されている傾向が見えた。

4.2　慣用句

慣用句１　「気にする」

（23）声が炸裂するので、慣れていると言えば慣れているが、外に出ると人目がある。人目を気にして、控えめに叱ってくれるのならまだいいのだが、母は違っていた。

（小山田歩美(2005)『ひまわり』日本文学館）

慣用句２　「あとにする」

（24）雪印食品の偽装工作を知ったのは、乞われてやってきた西播磨をあとにして西宮に戻り間もないころだった。

（今西憲之（2003）『内部告発―権力者に弓を引いた三人の男たち』鹿砦社）

慣用句３　「～を異にする」

（25）こうしてモザイクの美は過去の文化財になったように見える。だが、素材と手法を異にして、モザイク画の精神は、コンスタンティノポリスを介し…

（樺山紘一（1992）『世界史への扉』朝日新聞社）

5. 会話、見出し、レシピに見られた形式動詞「して」の脱落

前述したように、「X を Y にする」という文型は変える・変わる意味を表す場合及び慣用句の場合、「して」が脱落しにくい傾向があるが、すべての変える・変わる意味を表す「して」とすべての「X を Y にする」という形の慣用句の「して」が脱落しにくいとは限らない。会話や見出しやレシピなどのような省略が要求される場合、「して」の脱落も見られた。

5.1「X を Y にする」という文型は変える・変わる意味を表す場合見られた「して」の脱落

5.1.1 会話に見られた「して」の脱落

（26a）約四十人が、収穫したレタスをみそ汁やサラダにして、きれいに管理された開放感のある庭園で昼食を楽しんだ。翁長雄志市長も参加した。

（『琉球新報社朝刊』2004/4/13　琉球新報社）

（26b）（会話）コーンも入れれば彩がキレイだったな～（苦笑）最初は「発芽玄米をサラダに ？！」って感じでしたが、試してみたらこれ、かなり美味しかったです。

（Yahoo!ブログ/Yahoo!サービス/Yahoo!ブログ　Yahoo! 2008）

5.1.2　見出しで見られた「して」の脱落

（27a）「今、日本に入っている鰻の大半が、中国産です。中国で養殖した鰻を蒲焼きにして日本に輸出するんですが、鰻を裂く技術が中国にはない。

（深田祐介(1993)『新・新東洋事情』文芸春秋）

（27b）（見出し）浜中の「日帰りさんま」を蒲焼きに　札幌・パイオニアジャパン

http://www.suisan.jp/kakou/003854.html（2012 年 9 月 9 日に閲覧）

5.1.3 レシピに見られた「して」の脱落

（28a）ハムをみじん切りにしてにんにく入れたオリーブオイルで炒め。

（Yahoo!ブログ/生活と文化/グルメ、ドリンク　2008）

（28b）（レシピ）にんにく、ショウガ（市販のチューブ入りでＯＫ）一片ねぎ２～３本をみじん切りに 。

（Yahoo!知恵袋/暮らしと生活ガイド/料理、グルメ、レシピ Yahoo! 2005）

5.2 慣用句に見られた「して」の脱落

慣用句「心を一つにする」

（29a）あるホノルル行きの一機の飛行機のアクシデントにかかわり、皆が心を一つにして無事なフライトができるように努力する物語でした。

（Yahoo!ブログ/エンターテインメント/映画　Yahoo!ブログ　2008）

（29b）（文中）十本の指に満たなくても皆が心を一つに一生懸命修行に専念すれば、それはそのまま道場が盛んであるといっていいのだと…

（酒井大岳（1930）『人生を拓く―正法眼蔵随聞記入門』講談社 1994）

（29c）（見出し）「心をひとつに！山元町ふれあい産業祭」を開催します

http://www.town.yamamoto.miyagi.jp/kankou/fureai-sangyou_fes.html （ 2012 年 9 月 11 日に閲覧）

6. b「X を Y にして」と「X を Y に」両方の形がそろっている用例

今回の調査で、「X を Y にして」と「X を Y に」両方の形がそろっている用例が 96 例ある。

6.1「X をYに」の統語上の使用数量の分布

その 96 例を村木（1991）の分類にしたがい、以下の表にまとめる。

分類	時間	空間	限界	基準	理由	目的	資格	所持	排除	手段	相手	内容
数量	8	14	2	30	1	9	2	15	3	4	4	4

統語論上の役割から見ると、96 例の例文は次のようにまとめる。

統語論上の役割	状況成分	付帯状況	補語成分
数量	64	20	12

以上の表から、状況成分に用いる「X を Y に」が一番多く、全体の 66.7%を占めていることが明らかになった。その中で、特に、基準を表す用法が一番目立っている。つまり、「X を Y に」の各用法の中で、一番よく使われるのが基準を表す用法で（30 例)、全体の三分の一を占めている。次は空間を表す用法で（14 例)、全体の 14.5%を占めている。

6.2 「X をYにして」と「X をYに」のそれぞれの使用数量と使用率

調査方法 1 で、「X を Y にして」と「X を Y に」両方の形がそろっている場合、「X を Y にして」と「X を Y に」のそれぞれの使用数量と使用率を調べる、次のようにまとめている。

① 「X を Y に」の形があるが、「X を Y にして」の形がない例

～を活動拠点に、～を限度に、～を理想に、～を楽しみに、～をねらいに、～を心待ちに、～を小脇に、

以上挙げた表現は「X を Y にして」の「して」が脱落してから、「X を Y に」の形だけ使われるようになり、「X を Y にして」の形がまったく使われなくなったと思われる。

②「X を Y に」の使用率が 50%を超えた例

～をきっかけに（94%)、～を条件に（92.0%）～を契機に（91%)、～を中心に（87.0%)、～を理由に（87%)、～を目的に（82.0%)、～を基に（76.0%）～を根拠に（74.0%)、～を対象に（73%)、～を境に（73.0%)、～を相手に（69%)、～を目標に（65.0%）

～を舞台に（55.0%）

以上から、全体から見れば、「X を Y に」の使用率が高く、「X を Y にして」の使用率が低いことがパーセンテージからはっきりわかった。

7．まとめと今後の課題

本論は「X を Y にして」の形式動詞「して」の脱落について考察した。結論としては、(1)「X を Y にする」という文型は変える・変わる意味を表す場合及び慣用句の場合、「し

て」が脱落しにくい傾向があり、「X を Y にして」の「して」が脱落できなくて、「X を Y に」変更できない。しかし、会話や見出しなどのような省略が必要な場合、一部分の「X を Y にして」の「して」が脱落できることが分かった。

(2)「X を Y にする」という文型は変える・変わるという意味を表さない時、「X を Y にして」の「して」が脱落でき、「X を Y に」に変更できると考える。

(3)「X を Y に」の一番よく使われるのが基準を表す用法で、次は空間を表す用法である。

(4)「X を Y にして」と「X を Y に」両方の形がそろっている用例は、「X を Y に」の使用率が高いことが明らかになった。

「～をきっかけにした N」「～をよそにした N」のような連体修飾表現は今回の調査では触れていない。また、「X を Y にする」と「X を Y とする」との区別は更に検討する余地がある。田中 (2004) は、「7 時にセットするバスのところを 9 時にセットする (ことにする)」のような表現は<X ヲ Y 二>という附帯状況を表すフレーズの一部と考えることができると述べている。それらについての考察は今後の課題としたい。

参考文献：

奥田靖雄（1983）「を格の名詞と動詞とのくみあわせ」「を格の形をとる名詞と動詞とのくみあわせ」「に格の名詞と動詞とのくみあわせ」言語学研究会（編）(1983)『日本語文法・連語論（資料編)』pp281－323　むぎ書房

金子比呂子（1990）「「して」からみた「N1 を N2 にして」の位置付け方」『日本語学校論集』pp17－39　東京外国語大学外国語学部附属日本語学校

田中寛（2004）『日本語複文表現の研究 ： 接続と叙述の構造 』 東京　白帝社

田中寛（2010）『複合辞からみた日本語文法の研究 』 ひつじ書房

寺村秀夫（1983）「「付帯状況」表現の成立の条件―「X ヲ Y ニ……スル」」という文型をめぐって―」『日本語学』2 巻 10 号　明治書院　のちに寺村秀夫（1993）『寺村秀夫論文集 I ―日本語文法編集―』くろしお出版 pp113－126

益岡隆志・田窪行則（1992）『基礎日本語文法』くろしお出版

村木新次郎（1983）「「地図をたよりに、人をたずねる」という言い方」『副用語の研究』渡辺実　編　明治書院 pp267－290

村木新次郎（1985）「慣用句・機能動詞結合・自由な語結合」『日本語学』4 巻 1 月号　明治書院 pp15－27　のちに村木新次郎（1991）『日本語動詞の諸相』に改稿収録

村木新次郎（1991）『日本語動詞の諸相』ひつじ書房

森田良行（1985）「動詞慣用句」『日本語学』4 巻 1 月号　明治書院 pp37－44

日本語複合動詞「V 直す」、「V 返す」、「V 戻す」の特徴

木山直毅[*1] (大阪大学大学院)

Japanese Compounds "V-naosu", "V-kaesu", and "V-kaesu"

Kiyama Naoki (Graduate School of the University of Osaka)

1. はじめに

本稿では、(1) のような日本語複合動詞「V 直す」、「V 戻す」、「V 返す」の 3 つを質的、量的な観点から考察する。

(1) a. 建て直す、言い直す、書き直す、見直す、キャッチし直す
b. 呼び戻す、埋め戻す、差し戻す、奪い戻す、取り戻す
c. 取り返す、引き返す、蒸し返す、捏ね返す、繰り返す

これら 3 つの複合語は、以前に何かを行ったことを、再度、何らかの目的を持って行うことを表す。その点で異なっているようで似ている複合語である。そこで『現代日本語書き言葉均衡コーパス (BCCWJ)』[*2]を用いて (1) の意味や特徴を整理し分析する。

2. 先行研究

2.1. 斎藤 (1992)

「V 返す」の先行研究として、まず斎藤 (1992) を考察する。斎藤の「V 返す」の意味分類は表 1 である。格助詞の現れ方と、事態の構成要素、動作の前提と結果の違いから意味分類をしたものが表 2 である。

<< 斎藤の問題点 >>

表 1 の 4 つの分類を (ii) と (iii) の 2 つに分類することができる。それに加え、表 1 の (iii) の分類における、動作主の相違に関しては無記述である。動作主性とは例えば (2a) と (2b) のような場合である。

(2) a. ... をきっかけに急にまた気力をもり返したようだった。
b. あなたがこちらに姿を見せれば、その手の報道がまた蒸し返されるだけです。

[*1] kiyama.naoki@gmail.com

[*2] https://chunagon.ninjal.ac.jp

表 1　「V 返す」の斎藤 (1992: 183) の意味分類

	説明	例
(i)	物体の表裏の向きを逆にする。	鋤き返す、(土を) 掘り返す
(ii)	ある方向への移動、働きかけに対して、それとは反対方向への移動、働きかけを行う。	
a.	反射、反動作用を表す。	照り返す、(弾を) 跳ね返す
b.	他者からの行為に対して、こちらからもそれに対応する行為を行う。	どなり返す、笑い返す
c.	(こちらに向かってくる事物にある作用を加え) 移動方向を逆にする。	追い返す、送り返す
d.	離れていく事物をこちら側へとひきもどす。	奪い返す、呼び返す
(iii)	もう一度同じ動作、行為を行なう。	(答案を) 見返す、読み返す
(iv)	移動してきた方向へもどる。	引っ返す、(波が) 巻き返す

表 2　表 1(ii) の細かい分類 (ibid: 188)

	文型	構成要素	前提	結果
(ii)a	A ガ C ヲ B カラ (e.g. 太郎が盗まれた品物を犯人から取り返す)	A, C, B	A → B	A ← B
(ii)c	A ガ C ヲ (B ヘ) (e.g. 太郎が荷物を (郷里へ) 送り返す)	A, C, (B)	A ← B	A → B
(ii)d	A ガ C ヲ (e.g. 壁がボールを跳ね返す)	A, C	A ←	A →

(BCCWJ[*3])

(2a) において、気力を盛り返すのは、以前に気力があり、現在は気力を失ってしまった人で、動作主は一貫して同じ人である。一方で (2b) では、蒸し返す人は、かつて噂していた人と現在噂している人では異なっている。では、どのような V_1 の時にこのような動作主の一致、不一致が生じるのだろうか。斎藤はこの点は何も論じていない。

2.2. 王、由本 (2009)

王と由本は語彙意味論の立場から「V 直す」を中国語と比較して分析をしている。彼女らは、「V 直す」の意味を (3a)-(5a) のように 3 つの語彙概念意味 (LCS) で記述している。各 LCS の例は (b) である。

(3)　a.　[x [CONTROL [AGAIN [LCS_1]]]]

[*3] 本紙では例文の下線は全て発表者による

b. 走り直す、橋を渡り直す、飲み直す

(4) a. [x CAUSE [y BECOME [y BECOME [y BE [AT-RIGHT]]] BY [x CONTROLE [AGAIN [LCS_1]]]]]

b. スープを温め直す、論文を書き直す、計画を練り直す

(5) a. [x CAUSE [y BECOME [y BE [AT-z]]] BY [x CONTROL [AGAIN [LCS_1]]]]

b. 古い町家を喫茶店に建て直す、英文を和文に書き直す (王、由本 2009, 王 2011)

<< 王、由本への反論 >>

- 「スープを温め直す」→ 元の「温かい状態」へと戻すことが目的
- 「論文を書き直す」や「計画を練り直す」→ 現状より良いものを生み出すことが目的

表 3 「温め直す」と「練り直す」

	温め直す	練り直す
復元	◯	×
修正	×	◯

(6) 「温め直す」タイプ：ログインし直す、 カーテンを引き直す、ダウンロードし直す、服をかけ直す

(7) 「練り直す」タイプ：まとめ直す、加工し直す、巻き直す、録音し直す、書き直す

それに加え、(3b) の例文は全て (5) に分類されても良いものではないだろうか。由本は (3) が元であることを主張しているが (3) の「反復」を純粋に表すことは可能なのだろうか。

2.3. 姫野 (1999)

姫野は「V 直す」の動作主一致の問題を非常に細かく分類しているため、「V 直す」の動作主一致は姫野を踏襲することとする。表 4 は姫野の分類の一部を簡単にまとめたものである。

3. 研究内容

3.1. 「V 直す」の意味ネットワーク

本研究では「V 直す」を図 1 のように意味を分類する。そして各意味の定義は (8) のようになる。

(8) 修正: 動作を行うことで以前の状態に比べ、状態が改善されることを意図した動作
「前の X が悪かったため」や「より X を良くするために」と共起可

表 4　姫野の動作主の同一性の記述 (ibid: 198-203)

	動作主の異なり: 不可	例
i.	移動など、身体全体の動きに関する語	歩き直す、辿り直す
ii.	着脱など、身体の部分の動きに関する語	(視線を) 向け直す、(顔を) 洗いなおす
iii.	対等な対人行為に関する語	出会いなおす、戦い直す
iv.	感情、思考、知識獲得に関する語	気を取り直す、自身を持ち直す

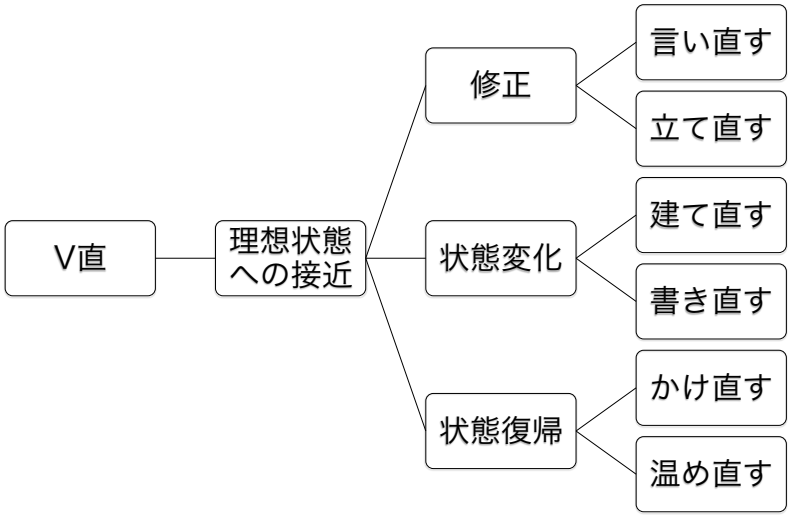


図 1　「V 直す」の意味ネットワーク

状態変化:　動作を行った直後と「V 直す」を行った直後で、動作の対象物が異なる動作
ヲ格と二格の両方が共起することが可能となる。

状態復帰:　以前の状態に戻すことを意図した動作
「V 直す」を「元の X に戻す」と置き換えできる

(9)　a.　材料は、とりあえずはこれで充分です」　キヨシは言い、それで私は急いで椅子に
すわり直した。

b.　データを分析し、組み合わせ、まとめ直し、さもなければ内容を変え、時には送る途中で
新しい情報を創り出してしまう。

c.　た。彼女はもっと一緒に飲んで話していたかった。　初めてきた客でも、ほかの店で
飲み直そうと誘う人がいるが、彼はそうでなかった。

d. 宗教と科学との対話を通して、従来なされてきた仏教思想研究がふたたび解釈し直されなければ、仏教の現代的な意義とそれがもっていた本来の使命は明らかにされない

e. 靴のひもがほどけていたのね』　母親は、歩道にひざまずいて、息子の靴のひもを結び直してやりながら、何気なく眼を上げた。

f. こんな立派な所じゃなくていいから、もっと安いホテルを見つけ直してくれないか、その目はそう言っていた。

g. 「でも、少しずつ、考え直してくれてるプロデューサーが出てきてさぁ」

(10) a. 九尺二間のみそ倉をつぶして車庫にし、母屋をプレハブ造りに建て直した。

b. 先に私や先輩が作成したと思われたくないので、最近では、発送前に私がこっそり作成し直しています。

c. 次いで上方の旧作を仕立て直した新作の、『あかね染』赤根屋半七、『恋の湖』稲野屋半兵衛、『椀久末松山』...

d. することで、Unix 用に書かれたソースコードを Windows 用として全面的に書き直すという膨大な手間を減らし...

e. ルムスキャナでデジタル化して、CD-R などに入れて定期的に別の CD-R などに焼き直せばいい。

(11) a. SP2 を削除した後に再度アプリケーションをインストールし直すと、トラブルを回避することができます。

b. をかけ、居間を抜けて玄関からポーチに出た。キーを使って外から玄関のドアに鍵をかけ直し、ポーチの石段を駆け降りてロッジのまえの小道を横切った。

c. こんどはダルタニャンのほうですぐ一歩後退し、剣を構え直した。

d. 帰る時間がはっきりして、それに合わせて飯を作ったり、冷めた飯を温め直したり、風呂を入れたりすることができて便利ではある。

e. シフトアップ時やアクセルを踏み直したときに、瞬間的に燃料を増量することでレスポンスや瞬発力が得られる。

f. 項目が表示されない場合は、コンピュータの管理者としてログオンし直す必要があります。

g. 水俣病によってずたずたにされた人々の気持ちをつなぎ直す「場」になっているのです。

h. 荒廃した国を建て直し、国民が明日に希望が持てるようにするには、各勢力の和解が不可欠である。 (BCCWJ)

3.2. 「V 返す[*4]」の意味ネットワーク

「V 返す」の表現は、本動詞「返す」の意味を色濃く保持した表現と、「V 返す」に固有の意味の二面性を持つ。次の例を見ていただきたい。

(12) a. アマは情けない顔のまま困ったような笑みを浮かべて、私が弱々しく [相手に/% 再度]微笑み返すと...

b. 私は映子の眼を見つめて頷いた。映子は [相手に/% 再度]頷き返し、私の耳に唇をつけた。

(13) a. より煽情的なさっきの話題を [% 相手に/再度]むし返してきた。

b. 読み返すと楽しかった日々が [% 相手に/再度]思い返されて、何だかおばあちゃんがすぐそこにいるような...。そんな気持ちになれた。 (BCCWJ)

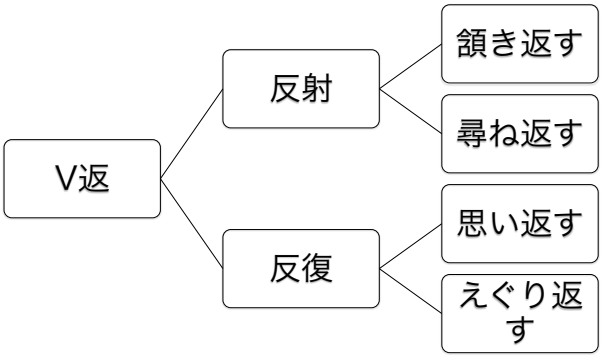


図 2 「V 返す」の意味ネットワーク

(14) 反射: 相手が行なってきた動作を、全く同じ様態によって反復し、相手に返す動作
(「相手に」と共起可 | ヲ格とニ格が共起可)
反復: 以前に行った動作を再度行う動作（「再度」と共起可 | 原則としてヲ格のみが共起可)

次に「V 返す」が「反復」用法の動作主の同一性に関する議論に移る（表 5 参照)。(15) は動作主が同一であり、(16) は動作主が異なっている例である。

(15) a. この３人は過去の実績からしたらこれから絶対 [この３人は]巻き返すでしょう。

b. 彼らから与えられた聖教などは師の説に背いたときには [聖書を与えられた人は]悔い返された。

[*4] 「ひっくり返す」や「ごった返す」に関しては、V_1 の「ひっくり」の語源がはっきりしないため、ここでは今後の課題して扱わないことにする。同時に、「失敗/ミスを取り返す」なども現在は扱えていないため今後の課題とする。

表 5　動作主の同一性

同一でなくてはならない			同一でなくても良い			
巻き返す	悔い返す	煮返す	繰り返す	穿り返す	捉え返す	蒸し返す
吹き返す	思い返す		混ぜ返す	揺れ返す	鋤き返す	掘り返す
盛り返す	読み返す					

c. 何の知識もない我々にいきなり色々説明されても理解できるはずもなく [我々は] 何日もマニュアルを読み返す日々が続きました。

(16) a. ふみの父親が警察庁の警備局長であることから捜査一課には事件を蒸し返すなという空気が漂っている。

b. 夜中に墓地に出かけては、埋葬されたばかりの死体を堀りかえして持ち帰ったのだ。

c. 文子の手話を、施設の先生みたいに上手だと言う斉藤に、先生よりうまいと小池がまぜっ返して、三人は大笑いした。 (BCCWJ)

3.3. 「V 戻す」の意味ネットワーク

「V 戻す」は本動詞「戻す」の意味をほぼそのまま引き継いだ意味となっている。

(17) a. 結果はいわばもとの世界へひき戻されることに終わった

b. 2004 年 3 月に時間を巻き戻してして話を進めていきます

c. 売りをかけて値下がりしてから買い戻すプロでない限り ...

(18) a. 彼を自身の侍医としてソウルに呼び戻す。

b. カセットを何度も巻き戻して。

c. みどりの窓口で払い戻せますか。 (BCCWJ)

以上のことから、「V 戻す」は、これまでに見た複合語に比べ、元の意味を保っている。そこで、「V 戻す」に関しては次のような意味階層を提案する。

3.4. 動詞の特徴付け

本節では、複合語全体がどのような特徴を持つのかを検討する。まず、石井 (2007: 35) に基づく V_1 の「アスペクト・ヴォイスモデル」に基いて意味特徴を見ていく。表 6 の頻度は百分率による割合である。そしてその結果を対応分析にかけた（図 4 参照)。この対応分析の結果から、「V 直す」の次

図 3　「V 戻す」の意味階層

元にはより他動詞に近い V_1 が、他の 2 つにはより自動詞に近い V_1 が集まっていることが分かる。

表 6　石井のモデルに基づく V_1 タイプと V_2 の結合の割合

		戻す	直す	返す	全体の割合
1	主体動作（自）動詞	10.71	5.48	13.51	9.90
2	主体動作（他）動詞	57.14	58.9	56.76	57.60
3	主体動作客体変化動詞	25	26.03	17.57	22.87
4	主体変化動詞	3.57	1.37	6.76	3.90
5	再帰動詞	3.57	8.22	5.41	5.73
	合計	100	100	100	100

3.5.　複合動詞と使用域

Akita (2012) は、特定の形態音韻論が特定の使用域にのみ用いられることを指摘している。そこで本節では対象としている複合語が、使用域にどのような傾向があるのかを考察する（図 5 参照)。ここでは各動詞項目の頻度が 31 以上の語に絞った。図 5 を見ると第 1 次元においては小説や韻文のような、物語などを書いた創作散文と韻文が集まる次元と、情報散文が集まる次元に分けることができる。第 2 次元は、アカデミックとノンアカデミックのジャンルに分けることができる。

図 5 を元に各複合語がどのような特徴を持つのかを概観する。

V 返す: 創作散文やアカデミックのジャンルに偏りが見られ、国会議事録や新聞、白書、広報誌などには出現しない。

V 直す:「V 返す」ほど強い傾向ではないが、「V 返す」に似た傾向を見せる。しかし、「持ち直す」や「見直す」は政治や経済で多く使われる表現であるため、語彙項目によってはノンアカデミックな情報散文にも用いられるが、大部分は創作散文やアカデミック

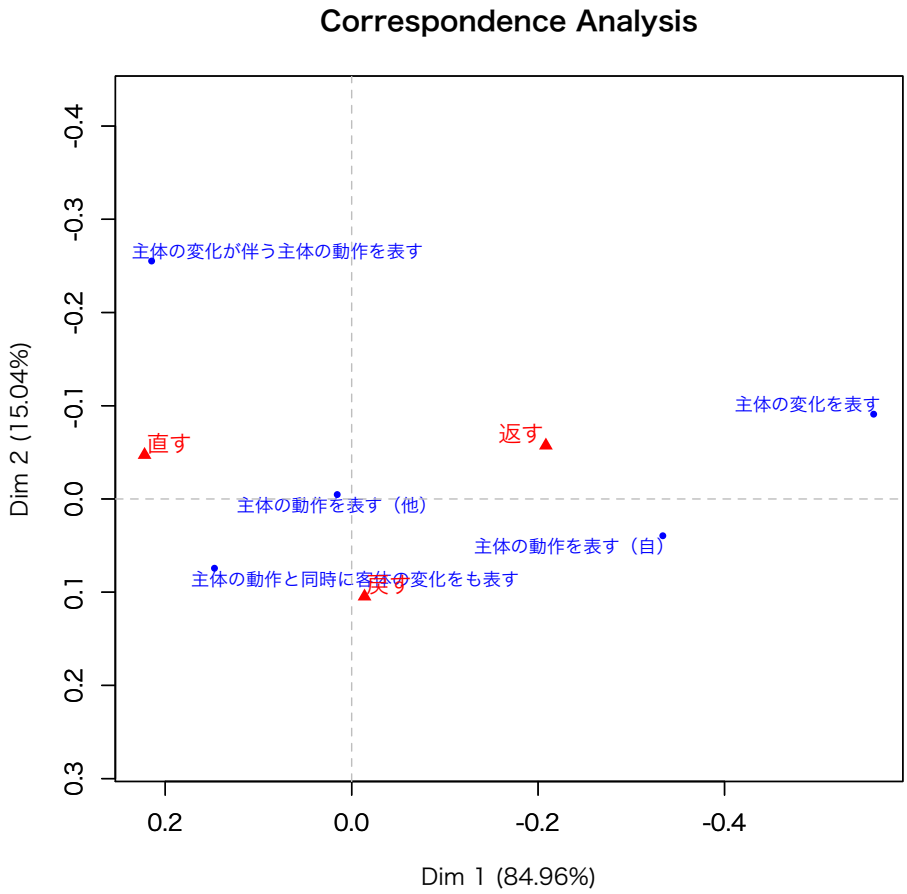


図 4　V_1 のアスペクトと V_1 の関係

なジャンルで多く使われる。

V 戻す: 他の 2 つとは異なり、ジャンルに対しては中立的である。

4.　結論

本論では、日本語の複合動詞「V 直す」、「V 返す」、「V 戻す」の 3 つを比較した。これらは、以前に行った行為をもう一度行う、ということを表している点で共通している。そして V_1 の特徴と V_2 の関係性を石井 (2007) に基いて考察した。その後、使用域を元に複合語全体の特徴を概観した。

参考文献

Akita (2012) “Register-specific morphophonological constructions in Japanese”. *The 38th Annual Meeting of the Berkeley Linguistics Society*. University of California. Berkeley

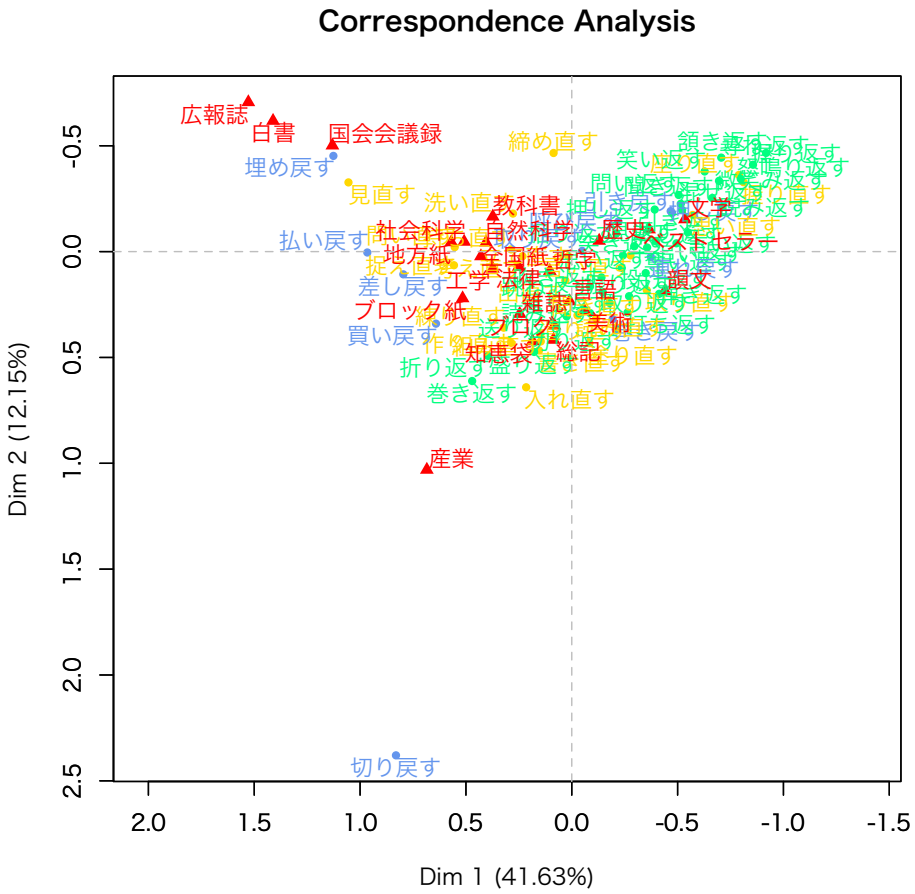


図 5　コレスポンデンス分析の結果

石井雅彦 (2007)「現代日本語の複合語形成論」. ひつじ書房

王蓓淳 (2011)「複合動詞「重 V」の語形成と意味解釈」『言語文化共同研究プロジェクト 2010 自然言語への理論的アプローチ』. pp. 21-30. 大阪大学大学院言語文化研究科

王蓓淳、 由本陽子 (2009)「中国語複合動詞「改 V」の形成と意味–日本語の「V+ かえる」「V+ 直す」との比較を参考に–」. 『日本言語学会第 139 回大会予稿集』. pp. 240-245

姫野昌子 (1999)『複合動詞の構造と意味用法』. ひつじ書房

森田良行 (1977)『基礎日本語–意味と使い方』. 角川書店

由本陽子 (2005)『複合動詞・派生動詞の意味と統語—モジュール形態論から見た日英語の動詞形成』. ひつじ書房

「私的な−名詞」「個人的な−名詞」の使い分け

渡邊　ゆかり（広島女学院大学）

Semantic Differences of '*Shitekina* - Noun' and '*Kojintekina* - Noun'

Yukari Watanabe (Hiroshima Jogakuin University)

１．はじめに

ナ形容詞の連体形として使用される「個人的な」「私的な」は次の（1）のようにほぼ同じ意味で使用することが可能な場合と，（2）（3）のようにいずれか一方が不自然な場合が存在する。

（1）a. 個人的な見解を述べさせていただくと，
　　b. 私的な見解を述べさせていただくと，
（2）a. この問題は個人的な努力で解決するものではない。
　　b. *この問題は私的な努力で解決するものではない。
（3）a. *今日は個人的な会合がある。
　　b. 今日は私的な会合がある。

しかしながら，以下のような国語辞典の意味記述から，これらがどのように使い分けられているかを理解することは難しい。

「個人的」［形動］個人を主体とするさま。個人に関するさま。公的でない立場や，他人と関わりない事柄についていう。プライベート。「―な意見」
「私的」　［形動］個人にかかわっているさま。おおやけでないさま。プライベート。「―な感情」⇔公的。

（Yahoo！JAPAN 辞書の国語辞典『大辞泉』による）

また，管見の限り，これらの使い分けについて考察している先行研究は存在しない。

従って，本研究では，BCCWJ（『現代日本語書き言葉均衡コーパス』）から収集した用例の分析を通し「個人的な＋名詞」「私的な＋名詞」の使い分けのメカニズムを解明することを目的とする。

２．分析方法

本分析においては，まず，BCCWJ から，検索エンジン「中納言」を用い，分析対象とする「個人的な」「私的な」の用例を収集した[1]。次に，これらの用例における「個人的な」「私

[1] 検索日時は，2012 年 9 月 6 日である。「個人的な」の用例収集においては，短単位検索でキーを「個人」と後方共起条件の「的」に指定して用例を収集した後，この中から「個人的な」の形を取るもののみ取り出した。また，「私的な」の用例収集においては，短単位検索でキーを「私的」に指定して用例を収集した後，この中から「私的な」の形を取るもののみを取り出した。

的な」の被修飾名詞のトークン数とタイプ数を調べた。ただし，トークン数については，同一 ID のサンプルから同一の被修飾名詞が複数得られた場合に限り，これを複数のまま数えず，1 例として数えた。従って，本研究で提示する被修飾名詞のトークン数は，同一の被修飾名詞が用いられたサンプルタイプ数に相当する。このような方法を取ったのは，特定の作者の言語感覚がその他の作者の言語感覚を凌いでトークン数に反映されたり，特定のテキストの性格が他のテキストの性格を凌いでトークン数に反映されたりするのを防ぐためである。また，被修飾名詞を抽出するにあたっては，以下のような処理を施した。

処理 1：被修飾範囲が複数の句からなる場合，その主要部に当たる名詞を被修飾名詞選択候補とする。また，被修飾範囲が複数の並立句からなる場合は，1 番目の句の主要部に当たる名詞を被修飾名詞選択候補とする。

処理 2：処理 1 の方法で取り出した被修飾名詞選択候補，ならびに，最初から名詞の形で取り出された被修飾要素のうち，単純語については，これを分析対象とする被修飾名詞として認定する。複合語については，これを，国立国語研究所が UniDic 用に規定した短単位レベルに当たる語彙素レベル[2]に分解した後，語彙素同士の結合関係において主要部的役割を果たしている単位を被修飾名詞として認定する。また，分解した語彙素同士が並立関係にある場合は，1 番目の語彙素を被修飾名詞とする。派生語については，接頭辞は被修飾名詞から除き，接尾辞は被修飾名詞に含めることとする。

処理 3：処理 1，処理 2 の方法で収集した被修飾名詞中に，表記の揺れが見られるものが存在した場合も，個々を異なる被修飾名詞とせず，同一の被修飾名詞として扱う。

次に，「個人的な」「私的な」の被修飾名詞トークン数を『分類語彙表』[3]の意味項目別に割り出し，この結果を基に「個人的な＋名詞」「私的な＋名詞」が保有する意味スキーマを抽出した後，「個人的な＋名詞」「私的な＋名詞」の意味構造と両者の相違を明らかにした。

---

[2] http://r-linc.org/pub/ogiso_20080209.pdfには短単位認定の方法として以下が記されている。

(1)一般
《和語・漢語》2 最小単位の 1 次結合を 1 短単位とする。|母=親| |食べ=歩く| |音=声| |本=箱|
《外来語》原則として 1 最小単位を 1 短単位とする。|コール|センター| |オレンジ|色|

(2)数
「数」以外の最小単位と結合させない。「数」どうしの結合については，一・十・百・千のとなえを取る桁ごとに 1 短単位とする。「万」「億」「兆」などの最小単位は，それだけで 1 短単位とする。小数部分は 1 最小単位を 1 短単位とする。
|十|二|月|二十|三|日| |七百|五十|二|万|語| |五|分|の|二| |二三十|回| |〇|. |四|五|

(3)その他
1 最小単位を 1 短単位とする。
《付属要素》|扱い|兼ねる|
《助詞・助動詞》|豊か|な|暮らし|に|つい|て|
《人名》|星野|仙一| |アンディー|・|シーツ|
《地名》|大阪|府|豊中|市|待兼山町| |六甲|山|
《記号》|図|A| |ＪＲ|

[3] 国立国語研究所編（2004）『分類語彙表増補改訂版』大日本図書

次節では，まず前述の方法で収集した「個人的な」「私的な」の被修飾名詞のトークン数とタイプ数を見ていく。

３．「個人的な」「私的な」の被修飾名詞のトークン数とタイプ数

前節で示した方法を用い分析対象として収集した「個人的な」「私的な」の被修飾名詞のトークン数，タイプ数は，次の表 1 の通りである。

表 1　「個人的な」「私的な」の被修飾名詞のトークン数とタイプ数

	「個人的な」の被修飾名詞	「私的な」の被修飾名詞
トークン数	1,069	296
タイプ数	422	184

表 1 より，トークン数，タイプ数ともに，「個人的な」の被修飾名詞の方が多いことがわかる。「個人的な」の被修飾名詞トークン数は，「私的な」の被修飾名詞トークン数の約 3.6 倍で，「個人的な」の被修飾名詞タイプ数は，「私的な」の被修飾名詞タイプ数の約 2.3 倍である。

以上の結果より，「個人的な＋名詞」「私的な＋名詞」の意味構造の相違について次のような可能性が挙げられる。なお，これらの可能性はいずれも「個人的な＋名詞」「私的な＋名詞」各々が複数の意味スキーマを保有していることを前提としている。

可能性 1　「個人的な＋名詞」は「私的な＋名詞」に比べ，より多くの種類の意味スキーマを保有している。

可能性 2　「個人的な＋名詞」は「私的な＋名詞」に比べ，使用頻度の高い意味スキーマを多く保有している。

可能性 3　「個人的な＋名詞」が保有する意味スキーマの中には，「私的な＋名詞」のそれとほぼ同じものが存在するが，「個人的な＋名詞」のスキーマとしての典型性の方が高いために，対応形式として「個人的な＋名詞」の方が選択されやすい。

次節では，これらの可能性を考慮しながら，『分類語彙表』の意味項目別に分類した「個人的な」「私的な」の被修飾名詞をてがかりに，「個人的な＋名詞」「私的な＋名詞」が保有する意味スキーマを抽出する。

４．『分類語彙表』の各意味項目に属する被修飾名詞トークン数

４．１．　意味項目 X.1-X.5 の各々に属する被修飾名詞トークン数

『分類語彙表』の X.1－X.5 の意味項目別に，「個人的な」の被修飾名詞トークン数，「私

的な」の被修飾名詞トークン数を調べたところ，結果は次の表 2 の通りであった[4, 5]。なお，「個人的な」の被修飾名詞トークン含有率と「私的な」の被修飾名詞トークン含有率とで有意差が認められる項目には，▲，▼のいずれかの記号を付した[6]。▲は被修飾名詞トークンの含有率が他方より有意に高いことを，▼は被修飾名詞トークンの含有率が他方より有意に低いことを示している。

表 2　分類項目 X.1-X.5 に属する「個人的な」「私的な」の被修飾名詞トークン数[7]

	X.1 抽象的関係	X.2 人間活動の主体	X.3 人間活動	X.4 生産物および用具	X.5 自然物および自然現象	該当なし
a.個人的な (含有率%)	290/1069 ▼(27)	21/1069 ▼(2)	750/1069 ▲(70)	61/1069 ▼(6)	8/1069 (1)	20/1069 (2)
b.私的な (含有率%)	105/296 ▲(35)	37/296 ▲(13)	170/296 ▼(57)	31/296 ▲(10)	3/296 (1)	7/296 (2)
a－b	185	－16	580	30	5	13

表 2 より，まず，トークン数の最も高い意味項目に着目すると，「個人的な」「私的な」ともに，X.3 の「人間活動」に属する被修飾名詞トークン数が最も多いことがわかる。ただし，「個人的な」と「人間活動」に属する名詞との結合例は 750 トークンであるのに対し，「私的な」と「人間活動」に属する名詞との結合例は 170 トークンで，両者の間には 580 トークンもの差が存在する，また，「人間活動」に属する被修飾名詞トークン含有率も「個人的な」の方が「私的な」より有意に高い。従って，「個人的な」「私的な」はいずれも，他の意味項目の被修飾名詞より「人間活動」に属する被修飾名詞と結合する傾向にあるものの，その傾向は「個人的な」の方がはるかに強いということができる。

次に，「個人的な」の被修飾名詞トークン数と「私的な」の被修飾名詞トークン数の差に着目すると，X.2 の「人間活動の主体」を除く，X.1 の「抽象的関係」，X.3 の「人間活動」，X.4 の「生産物および用具」，X.5 の「自然物および自然現象」の 4 項目において「個人的な」の被修飾名詞トークン数が「私的な」の被修飾名詞トークン数を上回っている。これら 4 項目のうちトークン数の差が最も顕著なのは，580 トークン差の「人間活動」であり，その後に 185 トークン差の「抽象的関係」，30 トークン差の「生産物および用具」，5 トークン

4　「X」は，『分類語彙表』の「1.体の類」「2.用の類」「3.相の類」のいずれかであることを示す。なお，収集した用例のほとんどは「1.体の類」であったが，一部「2.相の類」のものも存在した。

5　『分類語彙表』に記載されている語の中には，意味の多面性ならびに多義性を考慮し，異なる複数の意味項目に分類されているものが存在する。本調査で収集した被修飾名詞の中にもこのような語が含まれていた。このような語については，意味が多面的であるという理由で異なる複数の意味項目に分類されている場合のみ，各項目のトークン数に計上した。多義的であるという理由で異なる複数の意味項目に分類されている場合については，前後の文脈から用いられている意味を特定し，この意味と合致する意味項目においてのみトークン数に計上した。

6　T 検定を行った結果，$p \leqq 0.05$ で有意差が認められたものについて▲，▼の記号を付した。

7　「該当なし」は，X.1－X.5 のいずれの意味項目にも属さないことを示す。

差の「自然物および自然現象」が続く。

このように，意味項目別に「個人的な」の被修飾名詞トークン数と「私的な」の被修飾名詞トークン数を比較すると，項目により両者のトークン差に偏りが存在することがわかる。そして，この偏りは，「個人的な＋名詞」と「私的な＋名詞」の意味構造の相違に由来する。

従って，次の 4.2 からは，この相違を明らかにするために，「個人的な」の被修飾名詞トークン数が「私的な」のそれを大きく上回る「人間活動」ならびに，「個人的な」の被修飾名詞トークン数が「私的な」のそれを下回る「人間活動の主体」に限定し，「個人的な」「私的な」とこれらの項目に属する被修飾名詞との意味的な結び付きについて考察する。

４．２．　X.3「人間活動」の下位項目に属する被修飾名詞トークン数

「個人的な」「私的な」のいずれにおいても被修飾名詞のトークン含有率が最も高く且つ両者のトークン差が最も大きかった X.3 の「人間活動」は，『分類語彙表』においてさらに「心」「言語」「芸術」「生活」「行為」「交わり」「待遇」「経済」「事業」の 9 種類の下位項目に分けられる。各下位項目に属する「個人的な」「私的な」の被修飾名詞トークン数は次の表 3 の通りである。

表 3　分類項目 X.3.0－X.3.8 の「個人的な」「私的な」の被修飾名詞トークン数

	X.3.0 心	X.3.1 言語	X.3.2 芸術	X.3.3 生活	X.3.4 行為	X.3.5 交わり	X.3.6 待遇	X.3.7 経済	X.3.8 事業
a.個人的な	479	86	12	83	54	41	90	28	10
b.私的な	42	16	2	23	25	33	12	28	7
a－b	437	70	10	60	29	8	78	0	3

まず，表 3 より，「個人的な」「私的な」ともに，X.3.0 の「心」に属する被修飾名詞トークン数が最も多いことがわかる。しかしながら「心」に属する「個人的な」の被修飾名詞トークン数は 479 トークンであるのに対し，「心」に属する「私的な」の被修飾名詞トークン数は 42 トークンと，両者の間には 437 トークンもの差が存在する。

「心」に属する「個人的な」「私的な」の被修飾名詞のうち，トークン数が 2 トークン以上で且つ上位 5 位以内のものには，以下が存在した。なお，（　）内の数値はトークン数を示している。また，名詞の右肩に付された「●」は，その語が「個人的な」の被修飾名詞としても「私的な」の被修飾名詞としても用いられていたことを示す。

「心」に属する被修飾名詞例

個人的な…1 位「意見●」(64)，2 位「問題●」(42)，3 位「経験●」(28)，4 位「感情●」(25)，5 位「感想●」「見解●」(各 23)

私的な　…1 位「意見●」(4)，2 位「思い出●」「感情●」「感想●」「研究●」「調査」(各 2)

「個人的な」の被修飾名詞のうち，トークン数を「私的な」の被修飾名詞トークン数で割った値が最も高いものは「問題」で，「個人的な問題」のトークン数は「私的な問題」のトークン数の 42 倍にあたる。また，この値が 2 番目に高いものは「経験」で，「個人的な経験」のトークン数は「私的な経験」のトークン数の 28 倍に当たる。これらの被修飾名詞と「個人的な」「私的な」との意味的な結び付きは以下の意味スキーマ 1 と対応している。

意味スキーマ 1
A〈他者とは関係なく，自己と密接に関わる〉＋B〈物事〉

従って，以上より，意味スキーマ 1 は，「私的な＋被修飾名詞」より「個人的な＋被修飾名詞」と強く結び付いていると言える。

また，「意見」「感情」「感想」「見解」といった個人の認識を表す被修飾名詞トークン数も「個人的な」が「私的な」を大きく上回る。「個人的な意見」「個人的な感情」「個人的な感想」のトークン数はそれぞれ「私的な意見」「私的な感情」「私的な感想」のトークン数の 16 倍，約 13 倍，約 12 倍に相当する。この他「個人的な見解」のトークン数は 23 であったが，「私的な見解」のトークン数は 0 であった。これらの被修飾名詞と，「個人的な」「私的な」との意味的な結び付きは以下の意味スキーマ 2 と対応している。

意味スキーマ 2
A〈他者より自己の立場，価値観を優先した〉＋B〈心情，言葉〉

従って，意味スキーマ 2 についても，「私的な＋被修飾名詞」よりは「個人的な＋被修飾名詞」と強く結び付いていると言える。ただし，「個人的な＋被修飾名詞」「私的な＋被修飾名詞」のトークン数の比率を見る限り，「個人的な＋被修飾名詞」「私的な＋被修飾名詞」各々とスキーマ 2 との結び付きの程度差は，スキーマ 1 ほど大きくはないと考えられる。

次に，他の意味項目についても，どのような名詞が「個人的な」「私的な」の被修飾名詞として含まれているかを見ていく。

「言語」「芸術」「生活」「行為」「交わり」「待遇」「経済」「事業」に属する「個人的な」「私的な」の被修飾名詞のうち，トークン数が 2 トークン以上で且つ上位 5 位以内のものには，以下が存在した。

「言語」に属する被修飾名詞例

個人的な…1 位「話●」(22)，2 位「相談」(7)，3 位「情報●」(5)，4 位「質問」(4)，
5 位「会話●」「発言●」「魅力」「メール」「メッセージ」「(各 3)

私的な　…1 位「情報●」(2)

「芸術」に属する被修飾名詞例

個人的な…1 位「日記」「悲劇」(各 2)

私的な　…2 トークン以上の語は無い。

「生活」に属する被修飾名詞例

個人的な…1 位「経験●」(28)，2 位「趣味」「生活●」(各 10)，4 位「仕事●」「楽しみ●」

（各 4）

私的な　…1 位「生活●」（7），2 位「労働」（2）

「行為」に属する被修飾名詞例

個人的な…1 位「努力」「能力」（各 6），3 位「仕事●」（各 4），4 位「行為●」「責任」「力」「用事●」（各 3）

私的な　…1 位「行為●」（7），2 位「活動●」「用事●」（各 3），4 位「労働」（2）

「交わり」に属する被修飾名詞例

個人的な…1 位「付き合い●」（14），2 位「相談」（7），3 位「交際●」（4），4 位「対立」（3），5 位「知り合い」（2）

私的な　…1 位「関係」（6），2 位「会合」「付き合い●」（各 3），4 位「加入」「サービス」（各 2）

「待遇」に属する被修飾名詞例

個人的な…1 位「意見●」（64），2 位「アドバイス」「指導」「要請●」「要望」（各 2）

私的な　…1 位「意見●」（各 4）

「経済」に属する被修飾名詞例

個人的な…1 位「利益●」（6），2 位「財産●」「収入」（各 2）

私的な　…1 位「年金」「利益●」（各 5），3 位「利害●」（4），4 位「財産●」（3），5 位「交換●」「サービス」（各 2）

「事業」に属する被修飾名詞例

個人的な…1 位「使用」（2）

私的な　…1 位「施設」（3）

これらの名詞の多くは，「個人的な」「私的な」のいずれとも共起する。しかし，名詞により共起のしやすさに違いが見られる。先に示した意味スキーマ 1，意味スキーマ 2 の B と対応する名詞は，「個人的な」と共起しやすい。それ以外の名詞の中には，以下の意味スキーマ 3 と対応するものも存在した。

意味スキーマ 3：
A〈職務や職場とは関係せず，個人の自由意志と権利に拠る〉＋B〈物事〉

例えば，「個人的な」「私的な」と「仕事」「付き合い」との結び付きは，意味スキーマ 3 と対応する。意味スキーマ 3 の B と対応する名詞のトークン数は，「個人的な」「私的な」とで大差がないことから，意味スキーマ 3 は，「個人的な＋名詞」「私的な＋名詞」のいずれとも同程度に結び付くスキーマであると見ることができる。

また，この他には，以下の意味スキーマ 4 と対応するものも存在した。

意味スキーマ 4：
A〈公的なものとしての資格を持たない〉＋B〈物事〉

例えば，「私的な」と「年金」「施設」との結び付きは，意味スキーマ 4 と対応する。なお，これらの名詞については，「個人的な」とは結び付きにくい。従って，「意味スキーマ 4」は，

「個人的な」よりも「私的な」との結び付きの強いスキーマであると見ることができる。

４．３．　X.2「人間活動の主体」に属する被修飾名詞トークン数

「私的な」の被修飾名詞トークン数が唯一「個人的な」の被修飾名詞トークン数を上回っている X.2 の「人間活動の主体」の項目は,『分類語彙表』においてさらに「人間」「家族」「仲間」「人物」「成員」「公私」「社会」「機関」の 8 種類に分けられる。各々の下位項目に属する「個人的な」「私的な」の被修飾名詞トークン数は次の表 4 の通りである。

表 4　分類項目 X.2.0－X.2.7 の「個人的な」「私的な」の被修飾名詞トークン数

	X.2.0 人間	X.2.1 家族	X.2.2 仲間	X.2.3 人物	X.2.4 成員	X.2.5 公私	X.2.6 社会	X.2.7 機関
a.個人的な	2	0	6	2	9	1	1	2
b.私的な	3	1	4	1	4	6	8	15
a－b	－1	－1	2	1	5	－5	－7	－13

各々の意味項目に属する「個人的な」「私的な」の被修飾名詞のうち，トークン数が 2 トークン以上で且つ上位 5 位以内のものには，以下が存在した。

「人間」に属する被修飾名詞例
　個人的な…1 位「アイデンティ」(2)
　私的な　…2 トークン以上の名詞は存在しない。
「家族」に属する被修飾名詞例
　個人的な…2 トークン以上の名詞は存在しない。
　私的な　…2 トークン以上の名詞は存在しない。
「仲間」に属する被修飾名詞例
　個人的な…1 位「知り合い」(2)
　私的な　…1 位「友人」(2)
「人物」に属する被修飾名詞例
　個人的な…1 位「ボランティア」(2)
　私的な　…2 トークン以上の名詞は存在しない。
「成員」に属する被修飾名詞例
　個人的な…1 位「ボランティア」(2)
　私的な　…2 トークン以上の名詞は存在しない。
「公私」に属する被修飾名詞例
　個人的な…2 トークン以上の名詞は存在しない。
　私的な　…1 位「個人」(6)
「社会」に属する被修飾名詞例

個人的な…2 トークン以上の名詞は存在しない。

私的な　…1 位「施設」(3)，2 位「場●」(2)

「機関」に属する被修飾名詞例

個人的な…2 トークン以上の名詞は存在しない。

私的な　…1 位「機関」「研究会」「施設」(各 3)

上記の名詞のうち，「アイデンティティ」は意味スキーマ 1 の B と，「知り合い」「友人」「ボランティア」は意味スキーマ 3 の B と，「個人」「施設」「場」「機関」「研究会」は意味スキーマ 4 の B と対応する。

以上，本節では，『分類語彙表』の意味項目別に分類した「個人的な」「私的な」の被修飾名詞を手がかりに，「個人的な＋名詞」「私的な＋名詞」が保有する意味スキーマを抽出し，各スキーマと「個人的な＋名詞」「私的な＋名詞」との結び付きの強さを分析した。

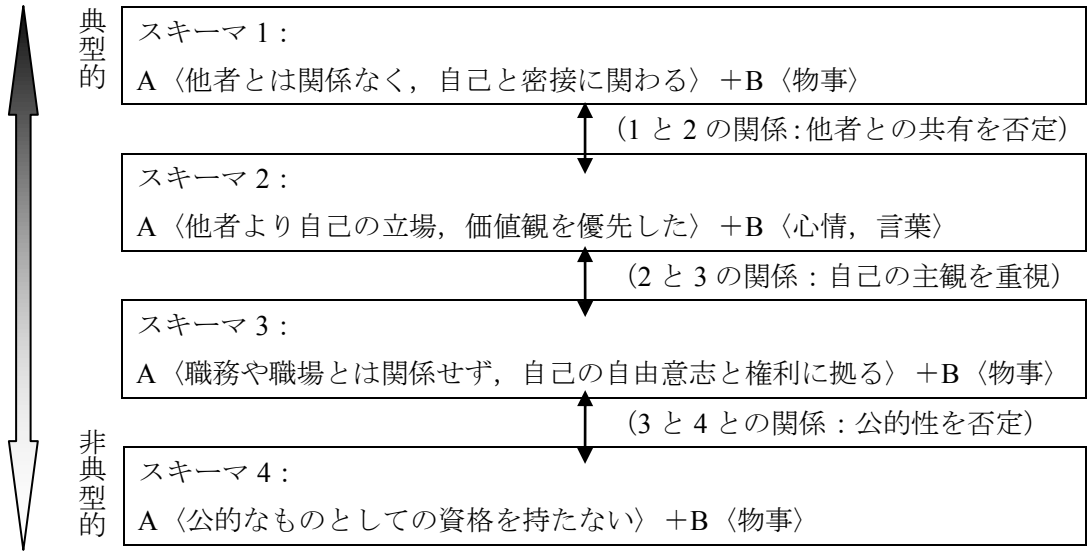
次節では，本分析を踏まえ，「個人的な＋名詞」「私的な＋名詞」の意味構造を記述するとともに，両者の相違について言及する。

５．「個人的な」「私的な」の意味構造の相違

前節での分析より，「個人的な」「私的な」の意味構造は以下のようなものであると考えられる。

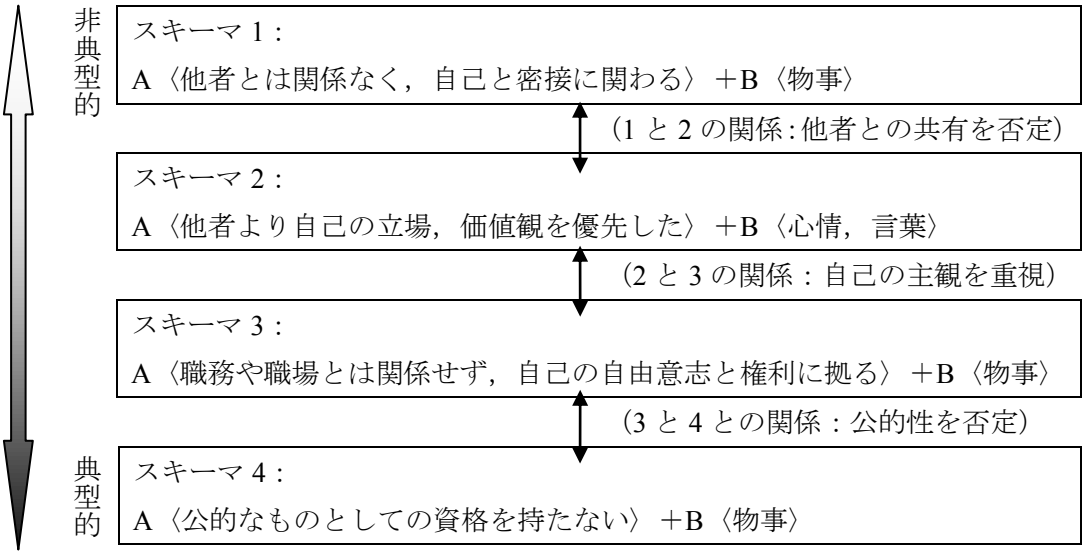
「個人的な」の意味構造

A＝「個人的な」，B＝被修飾名詞



「私的な」の意味構造

A＝「私的な」，B＝被修飾名詞



上記のように，「個人的な」の意味構造と「私的な」の意味構造は極めて類似している。いずれの意味構造においても，スキーマ 1 の A とスキーマ 2 の A には，他者との共有を否定する態度が反映されており，スキーマ 2 の A とスキーマ 3 の A には，自己の主観を重視する態度が反映されており，スキーマ 3 の A とスキーマ 4 の A には，公的性を否定する態度が反映されている。各スキーマは，このような形で結び付き，「個人的な＋名詞」「私的な＋名詞」の意味構造を形成している。

しかしながら，「個人的な＋名詞」「私的な＋名詞」とでは，それぞれの意味スキーマの典型性が異なる。「個人的な＋名詞」では，意味スキーマ 1 の典型性が最も高く，意味スキーマ 4 の典型性が最も低い。一方「私的な＋名詞」では，反対に意味スキーマ 4 の典型性が最も高く，意味スキーマ 1 の典型性が最も低い。従って，このような意味スキーマの典型性の相違より，1 節の（2）において「努力」が「個人的な」と共起し，「私的な」と共起しにくいのは，「努力」がスキーマ 1 の B と対応することによると見ることができる。一方，（3）において「会合」が「私的な」と共起し，「個人的な」と共起しにくいのは，「会合」がスキーマ 4 の B と対応することによると見ることができる。また，「個人的な」の被修飾名詞の方が「私的な」の被修飾名詞よりトークン数，タイプ数がともに多いのは，「個人的な＋名詞」と結び付きの強い意味スキーマ 1，意味スキーマ 2 へのアクセス頻度が他の意味スキーマに比べ高いことに起因すると考えられる。

６．さいごに

本稿では，ナ形容詞の連体形として用いる「個人的な」「私的な」に焦点を当て考察を進めてきたが，類義語に「プライベートな」が存在する。「プライベートな」と「個人的な」「私的な」との比較については今後の研究課題とする。

対訳対と協調フィルタリングを用いた商品推薦

柴田　翔平（東京農工大学　工学部　情報工学科）
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）

Product Recommendation using Translation Pairs and Collaborative Filtering

Shohei Shibata (Department of Computer and Information Sciences
Faculty of Engineering, Tokyo Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo Agriculture and Technology)

１．はじめに

近年，アニメや映画といった日本のメディア作品が海外で人気となり，それに関連する商品を外国人が購入する機会が増えている．しかし，日本語と外国語の間に存在する言語の壁から目的の商品を検索するのは難しい．特に，作品内の登場人物や地名は機械翻訳でも対応できない場合が多く，このような言語の壁が妨げとなって，外国人が目的の商品を購入できない場合がある．

本稿では，商品タイトルの日本語と外国語が対となった対訳対と，日本ユーザと外国ユーザの商品購入情報を基にして協調フィルタリングを用いることで，言語の壁を越え，外国人が目的とする商品を推薦することを目的とする．協調フィルタリングとは，一般的なインターネットのショッピングサイトでも利用される，ユーザの嗜好情報に基づいて推薦を行う方法であり，この方法を用いれば，日本と外国ユーザの嗜好の類似点を見つけ出し，言語をまたいで商品推薦を行うことができると考える．

２．関連研究

商品推薦に関しては，これまで様々な研究が行われている．その中でも，類似しているユーザを選択するための類似度に関する研究が多くある．

ソーシャルネットワークサービス（SNS）上でのつながりをグラフで表現した情報と，商品購入情報の二種類を用いて類似度を定義する研究（Symeonidis and Tiakas and Manolopoulos (2011)）や，SNS 上のつながりの強さまで考慮した類似度を定義する研究（Symeonidis and Tiakas and Manolopoulos (2010)）などがある．

また，類似度には時事性を含ませることが難しいため，時事性を含まない類似度を用いると，ユーザの嗜好の変化に合った推薦を行うことが困難にある．そのような問題に対応する研究も行われている．

意外性のある Web ページをリコメンデーションするため，ユーザの Web ページのブックマーク情報と Wikipedia のコンテンツを照らし合わせる研究（Chang and Quiroga(2010)）や，ユーザ間の関係と嗜好の時間による変化を反映させた協調フィルタリングによる推薦の研究（川前，坂野，山田，上田(1997)）などがある．

しかし，我々の調査した限り，これまでの研究では言語をまたがる商品推薦は考えられていない．そこで本稿では，商品推薦に一つの国のユーザの情報を用いるだけでなく，二つ目の国のユーザの情報を用いることで，商品推薦の結果に幅を持たせ，言語をまたいだリコメンデーションが行えるようなシステムを提案する．

３．対訳対を用いた商品推薦

ある外国ユーザを対象に，日本ユーザの商品購入情報から商品推薦を行うことを考える．外国ユーザと日本ユーザ間の類似度を計算し，協調フィルタリングによる商品推薦を行いたいが，商品購入情報をそのまま用いるだけでは言語の違いが存在するために，外国ユーザと日本ユーザの間で情報の共有ができず，類似度の計算が行えない．そこで，双方の言語の違いを埋めるため，商品タイトルの日本語と外国語との対関係を蓄積した「対訳対」を作成する．商品タイトルの対関係が存在している場合には，日本と外国で異なる商品タイトルでも，同じ商品を購入していると扱うことができる．

提案する商品推薦システムにおいて，外国ユーザと日本ユーザの商品購入情報は，ベクトル化して類似度計算に用いる．このベクトルの素性は商品であり，素性値は商品への評価値である．このベクトルと対訳対を用いて，同じ商品を購入して評価している外国ユーザと日本ユーザ間でコサイン類似度を計算し，協調フィルタリングによる商品推薦を行う．

日本ユーザの商品購入情報は「楽天株式会社」，外国ユーザの商品購入情報は「GroupLens Research」より提供していただいた情報を用いた．対訳対は，Wikipedia より配布されているダンプデータからタイトルの対応関係を抽出することで作成した．

なお，ユーザの商品購入情報や対訳対に存在する商品タイトルには，英語の大文字と小文字やバージョンの違いなど，表記の揺れが存在するため，それを削除した状態の情報も用いることとする．この情報を，商品タイトルを整形した情報と呼ぶ．

４．実験

４.１　商品推薦システムの評価実験

商品推薦システムの出力となる推薦結果の評価は，推薦された商品がユーザの意図するものであったかという判断になる．しかし，その判断は主観的なものであるため，システムへの評価が集まったとしてもその性能について議論することは難しい．

そこで，システムの定量的な評価を行うため，外国ユーザの商品購入情報に存在する商品を対象に商品購入情報をマスキングした上で商品推薦を行った．商品を推薦した数のうち，商品推薦結果に現れるマスキングした商品の割合を「適合率」とし，また，商品推薦結果に現れるマスキングした商品の順位を用いた「平均逆順位（MRR)」を定義し，商品推薦システムの評価実験の指標とした．適合率と MRR は，以下の式で計算される．

$$\text{適合率} P = \frac{\sum_{e \in E} \sum_{j \in J} C_{ej}}{\sum_{e \in E} \sum_{j \in J} N_{ej}} \qquad MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank(i)}$$

適合率 P において，C_{ej} はマスキングされた商品の数，N_{ej} は商品推薦数，e は外国ユーザ，E は外国ユーザの集合，j は日本ユーザ，J は日本ユーザの集合を表す．

MRR において，N はテストデータ数を表し，本稿では類似度が計算された外国ユーザと日本ユーザの組み合わせ数の 2 倍である．また，rank(i)は，i に対する推薦結果中，マスキ

ングされた商品の最高順位を表す．商品推薦の結果に正解が含まれなかった場合には，rank(i) = ∞ とする． MRR が高いほど，推薦結果の上位にマスキングされた商品が出現しているということになる．なお，評価実験は，外国ユーザの購入している商品のうち，マスキング対象の商品を情報を二つに分割して，二分割交差検定によって行った．商品タイトルの整形有無も考慮に入れたため，計四種類の評価実験を行っている．また，評価実験結果だけでなく，実際のリコメンデーション結果についても示す.

４.２　実験結果

外国ユーザ一人あたり推薦される商品数を 1 から 5 まで変化させたときの適合率のグラフを図 1 に, MRR のグラフを図 2 に示す. また, 実際の商品推薦結果の一部を表 1 に示す.

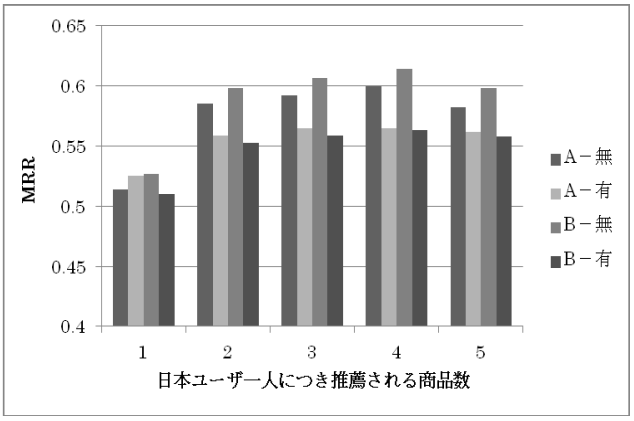


図 1　システムの評価実験における適合率

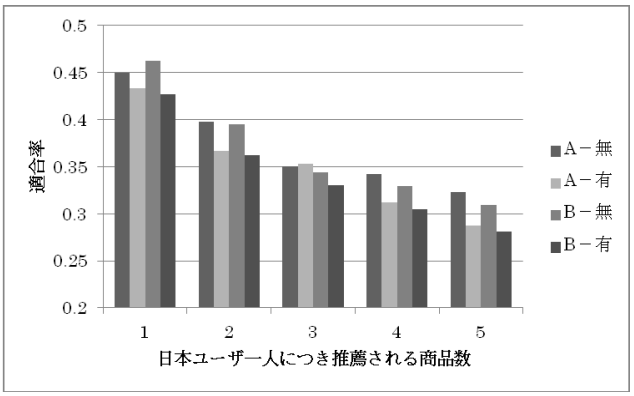


図 2　システムの評価実験における MRR の値

表 1　システムによる実際の商品推薦結果

英語ユーザ ID	日本ユーザ ID	商品番号	商品名
204269	68437	208	pinocchio
204269	68437	260	cinderella
204269	68437	76476	さるかにばなし
204269	68437	76477	三びきのこぶた
252953	16352	11025	Howl's Moving Castle
252953	16352	94061	となりのトトロ新装版

図 1 と図 2 において，凡例の A はマスキング対象を対訳対に存在する商品，B は外国ユーザの購入している商品を示す．また，有無は商品タイトルの整形を施したかどうかを示す．

図 1 と図 2 を見ると，マスキング対象 A,B ともに商品タイトルの整形を行わない方が適合率，MRR ともに高い値を示している．このことは，商品タイトルの整形によって商品タイトルの対応関係が増加したため，商品推薦の数自体は増加したものの，類似度が計算されるユーザも増えることでその推薦内容も多様になり，推薦結果にマスキングした商品が現れにくくなっていると考えられる．

しかし，表 1 に示したシステムによる実際の商品推薦結果を見ると，「pinocchio」（ピノキオ）や「cinderella」（シンデレラ）から「さるかにばなし」が推薦され，童話によるつながりから推薦が行われていると考えられる結果や，「Howl's Moving Castle」（ハウルの動く城）から「となりのトトロ」が推薦され，ジブリ作品のつながりから推薦が行われていると考えられる結果が存在した．このことから，評価実験とは別に，商品推薦システムは本稿の目的に沿った推薦を行っているといえる．

４．３　まとめと今後の展望

商品推薦システムの評価において，適合率は，商品タイトルを整形していない情報を用いた実験で，外国ユーザ一人あたりに推薦される最大商品数を 1 に設定したとき，適合率 46% を得た．また，MRR は，商品タイトルを整形していない情報を用いた実験で，外国ユーザ一人あたりに推薦される最大商品数を 1 に設定したとき，0.61 という値を得た．

また，表 1 に示した実際の商品推薦結果から，商品推薦システムは，本稿の目的に沿った結果を出力することができていると考えられる．

しかし，本稿の対訳対のみでは日本と外国の商品の対応関係が少ないため，今後は日本ユーザと外国ユーザとの間をより広く取り持てるような条件の拡張を行っていく必要がある．

謝　辞

本研究を行うにあたり、楽天株式会社と国立情報学研究所が協力して提供している『楽天データセット』を利用させて頂いた．また，GroupLens Research より提供していただいたデータも利用させていただいた．利用を快諾して下さった各社に謹んで御礼申し上げる．

文　献

Pei-Chia Chang and Luz M. Quiroga (2010). "Using Wikipedia's Content for Cross-Website Page Recommendations that Consider Serendipity". *Proceedings of the Interna-tional Conference on Technologies and Applications of Articial Intelligence,* pp293-298.

Panagiotis Symeonidis and Eleftherios Tiakas and Yannis Manolopoulos (2011). "Product Recommendation and Rating Prediction based on Multi-modal Social Networks" *Proceedings of the ACM Conference Series on Recommender Systems 2011*, pp61-68.

Panagiotis Symeonidis and Eleftherios Tiakas and Yannis Manolopoulos (2010). "Tran-sitive Node Similarity for Link Prediction in Social Networks with Positive and Negative Links" *Proceedings of the ACM Conference Series on Recommender Sys-tems 2010*, pp 183–190.

川前徳章，坂野鋭，山田武士，上田修功 (1997). "ユーザの嗜好の時系列性と先行性に着目した協調フィルタリング". 電子情報通信学会論文誌D Vol.J92-D No.6,pp.767-776.

BCCWJ 図書館サブコーパス全テキストへの文体情報付与結果の分析

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）
飯田 龍（東京工業大学 大学院情報理工学研究科）
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）
椿本 弥生（はこだて未来大学 メタ学習センター）
沼田 寛（はこだて未来大学 メタ学習センター）

Writing Style Annotation for the Library Subcorpus of 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)
Ryu Iida (Dept. Computer Science Graduate School of Information Science and Engineering, Tokyo Instiute of Technology)
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)
Mio Tsubakimoto (Center for Meta-Learning, Future University Hakodate)
Hiroshi Numata (Center for Meta-Learning, Future University Hakodate)

１．はじめに

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」（平成 21～24 年度）の成果報告である。『現代日本語書き言葉均衡コーパス』（BCCWJ）の図書館サブコーパスには，10,551 の書籍サンプルが収録されている。本研究ではそのコーパスをより有効に活用し，テキスト研究を進めるために，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付与，検証とを行ってきた（柏野・奥村 2012，柏野ほか 2012，柏野ほか 2012a，柏野ほか 2012b，保田ほか 2012，保田ほか 2012a，2012b）。

コーパスへ文体情報を付与することの重要性は，EAGLES（1996）等より議論され，例えば Lee（2001）によって，British National Corpus（BNC）への付与が実現されている。また，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がすでにいくつか行われている（小磯ほか 2008，2011，間淵ほか 2010）。しかしながら，サブコーパスに収録される約 1

* waka@ninjal.ac.jp

万という大量の書籍サンプルすべてを精査し，体系的に文体情報を付与するような試みは，本プロジェクトの実践がはじめてのことである。

これまで，柏野ほか（2012），柏野ほか（2012a，2012b）において述べてきたとおり，アノテーション作業は次の二段階で行った。

① 主に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②の細分類の対象とする。
② 内容・表現の細分類をする。「専門度（幼児・小学生～専門家：5段階），客観度（とても客観的～とても主観的：4段階），硬度（とても硬い～とても軟らかい：4段階），くだけ度（とても・どちらかといえば・くだけていない：3段階），語りかけ性度（とてもある・どちらかといえば・特にない：3段階）」の分類指標を付与する。

上記①の段階で図書館サブコーパスの 10,551 の書籍サンプルのうち，8,887（84%）を「構造的に単純なテキストタイプ」と判断し，上記②の内容・表現の細分類の対象とした。柏野ほか（2012），柏野ほか（2012a，2012b）では，その②のアノテーション作業結果に関する報告を重ねてきた。本稿では，これまで取り上げなかった，②の細分類の対象外としたものを取り上げ，それらの類型とアノテーション作業結果について報告する。該当サンプルは，全部で 1,664（16%）である。これまで対象外としたサンプルの分類結果まで分析することにより，図書館サブコーパスに収録される書籍サンプルの全体像と特徴とをより正確に把握することが狙いである。本稿で取り上げる一群のテキストを，以降「特徴的な類型のテキスト」と呼ぶこととする。

２．特徴的な類型のテキストのアノテーション作業

２.１　特徴的な類型のテキストの分類指標

柏野ほか（2009）では，BCCWJ 構築のサンプリングの過程で観察されたサンプルの多様性を報告した。その際に，文章形式に特徴のあるサンプルとして，Q&A 形式（例 1），会話形式（例 2），引用編集形式（例 3）を取り上げた。例 3 は，講義のあまった時間に学生に書かせたものを集めたものであるらしい。編者がそれらを引用して編集しているものとして，「引用編集方式」と呼ぶこととする。さらに，紙面形式に特徴のあるサンプルとして，コマ割りや図，イラストなどの視覚的表現を多用する一群（例 4）を取り上げた。以下，例を示す（サンプルの出典は，BCCWJ のサンプル ID と書名とで記す）。

例 1：Q&A 形式（PB33_00111『環境経営なるほど Q&A　環境先進企業へのヒント』）

Q3－7　マネジメントのための環境会計

マネジメントのための環境会計にはどんなものがありますか？　それぞれの特徴を教えて下さい。

A

■内部環境会計の意義

環境会計は，その目的により，外部報告目的の環境会計と内部管理目的の環境会計とに分類されています。わが国では環境省のガイドラインも推進力となって，多数の企業が環境会計を外部に公表するようになってきた一方，企業の意思決定に役立つ内部管理目的の環境会計の研究も進められています。

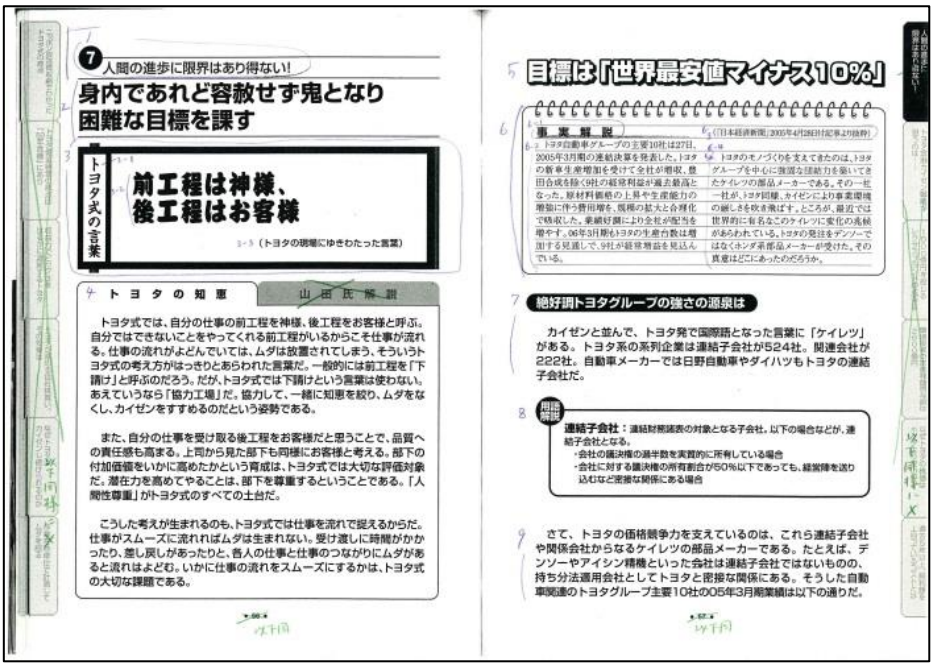
例 2：会話形式（PB53_00480『感性ちゃんと頭脳君の対話』）

感性　そういうことか。分かったわ。つまり，「肌の表面に何を塗っても，その物質がバリアゾーンを通過して有棘細胞層や基底細胞層にまで到達するわけがない」ってことなのね？
頭脳　そうだよ。そんなことは不可能なんだよ。もしもそれが可能だとしたら，肌の防衛網が機能していないことになるから，おそらくそういう人は生きていけないだろうね。
感性　物質のサイズを小さく，細かくしてもダメなの？
頭脳　ダメだよ。無理だね。バリアゾーンが健全な場合には，水の分子一個ですら通さないんだ。

例 3：引用編集形式（PB23_00427『ほろっと本音キラッと青春』）

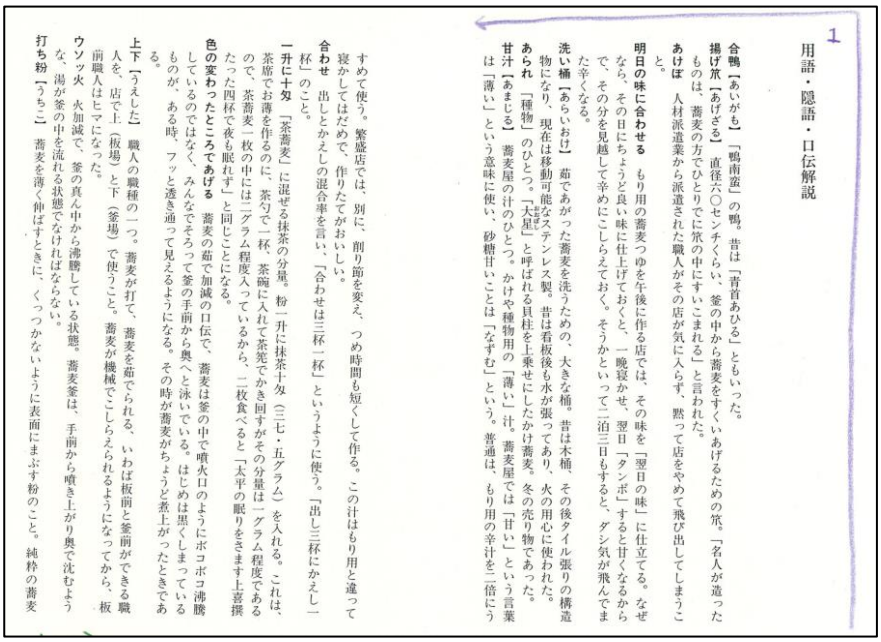
一八歳ってこんなものかなあ。ちょっと予定とはちがう。
なんだか毎日平凡。だけど，毎日平凡に過ごせていることを幸せだと思う。何も特別じゃなくていいと思いながら，毎日を平凡に頑張ってます。
友よ！　おまえらみんなさめすぎや。もっと毎日，感動的に生きろよ。

例 4：コマ割り（PB5n_00141『トヨタだけがなぜ儲かるのか！？』）



以上の観察に加え，辞書形式やカタログ形式をもつテキスト（例 5，6）も文章形式に特徴のあるものと考えられる。

例 5：辞書形式（LBp6_00009『蕎麦屋のしきたり』）



用語・隠語・口伝解説

合鴨【あいがも】「鴨南蛮」の鴨。昔は「青首あひる」ともいった。

揚げ笊【あげざる】直径六〇センチくらい、釜の中から蕎麦をすくいあげるための笊。「名人が造ったものは、蕎麦の方でひとりでに笊の中にすいこまれる」と言われた。

あけぼ　人材派遣業から派遣された職人がその店が気に入らず、黙って店をやめて飛び出してしまうこと。

明日の味に合わせる　もり用の蕎麦つゆを午後に作る店では、その味を「翌日の味」に仕立てる。なぜなら、その日にちょうど良い味に仕上げておくと、一晩寝かせ、翌日「タンポ」すると甘くなるからで、その分を見越して辛めにこしらえておく。そうかといって二泊三日もすると、ダシ気が飛んでまた辛くなる。

洗い桶【あらいおけ】茹であがった蕎麦を洗うための、大きな桶。昔は木桶、その後タイル張りの構造物になり、現在は移動可能なステンレス製。昔は看板後も水が張ってあり、火の用心に使われた。

あられ「種物」のひとつ。「大星」と呼ばれる貝柱を上乗せにしたかけ蕎麦。冬の売り物であった。

甘汁【あまじる】蕎麦屋の汁のひとつ。かけや種物用の「薄い」汁。蕎麦屋では「甘い」という言葉は「薄い」という意味に使い、砂糖甘いことは「なずむ」という。普通は、もり用の辛汁を二倍にうすめて使う。繁盛店では、別に、削り節を変え、つめ時間も短くして作る。この汁はもり用と違って寝かしてはだめで、作りたてがおいしい。

合わせ　出しとかえしの混合率を言い、「合わせは三杯一杯」というように使う。「出し三杯にかえし一杯」のこと。

一升に十匁　「蒸蕎麦」に混ぜる抹茶の分量。粉一升に抹茶十匁（三七・五グラム）を入れる。これは、茶席でお薄を作るのに、茶匁で一杯、茶碗に入れて茶筅でかき回すがその分量は一グラム程度であるので、茶蕎麦一枚の中には二グラム程度入っているから、二枚食べると「太平の眠りをさます上喜撰たった四杯で夜も眠れず」と同じことになる。

色の変わったところであげる　蕎麦の茹で加減の口伝で、蕎麦は釜の中で噴火口のようにボコボコ沸騰しているのではなく、みんなでそろって釜の手前から奥へと泳いでいる。はじめは黒くしまっているものが、ある時、フッと透き通って見えるようになる。その時が蕎麦がちょうど煮上がったときである。

上下【うえした】職人の職種の一つ。蕎麦が打て、蕎麦を茹でられる、いわば板前と釜前ができる職人を、店で上（板場）と下（釜場）で使うこと。蕎麦が機械でこしらえられるようになってから、板前職人はヒマになった。

ウソッ火　火加減で、釜の真ん中から沸騰している状態。蕎麦釜は、手前から噴き上がり奥で沈むような、湯が釜の中を流れる状態でなければならない。

打ち粉【うちこ】蕎麦を薄く伸ばすときに、くっつかないように表面にまぶす粉のこと。純粋の蕎麦

例 6：カタログ形式（LBj6_00025『熱帯魚・水草カタログ』）

その他の仲間の魚たち

以下のページで「その他の仲間」として紹介している魚たちは、本来、別々のグループに属している種類を、編集の都合上、「その他の仲間」として一まとめにしたものだ。したがって、ハゼの仲間やハタの仲間、そしてカワハギの仲間など、色々なグループの魚が登場してくる。多種多様な魚が含まれるということは、個々の魚の飼育方法や、飼育上の注意点などもすべて異なってくることを意味するので注意してほしい。

その他の仲間の魚の中で、特に人気の高い種類は、パープルファイヤー・ゴビー、ディープウォーター・アンティアス、そして、ロイヤルグラマなどである。これらの魚は、最近特に人気が高まりつつあるサンゴなどの無脊椎動物のレイアウト水槽での飼育に適した小型美魚である。この種の水槽では、自然のサンゴ礁に近い環境が再現されるので、こうしたレイアウト水槽に適した小型の美しい魚たちに注目が集まりつつあるのだ。

346

キイロサンゴハゼ

Gobiodon okinawae　~3 分布/西部太平洋

大きさ 3 cm
水温 24~27℃
難易度 普通
水槽 45cm以上
混泳 同大の温和な魚と
入荷 普通
価格 1,000~2,000円
エサ 冷凍エサ各種、フレークフード、顆粒状乾燥餌、アサリのミンチ、クリル、魚肉、魚卵

●全身が鮮やかな黄色一色に染まっているかわいい熱帯産のハゼの仲間である。あまり活発な魚ではなく、普段は岩の窪みや海藻の葉の上などにちょこんと体を乗せて一休みしていることが多い。輸入量はわりと多く、入手は容易だ。とても小さな魚なので、あまり大きな水槽に泳がせるとどこにいるかわからなくなってしまう。むしろ、60cm前後の小型水槽に無脊椎動物のレイアウトを作り、温和な魚とだけ混泳させて飼育するとよい。

347

以上述べたようなものを分類するために，次のような指標を設けた。

(a) 対話系（対話，対談・座談，インタビュー，往復書簡，シナリオ，その他対話形式）
(b) 引用系（Q&A 形式，投稿形式，その他引用編集形式）
(c) 視覚表現多用系（コマ割多用，図解，その他写真やイラストの多用）
(d) データベースやリスト系（用語解説，辞書形式，見本・カタログ形式，その他リスト形式）

さらに，文体を吟味する際，「本文」であるのか「前書き」や「後書き」であるのかは区別すべきと考えた。また，「内容」や「表現」の文体判断が困難になるようなものもそれぞれ別扱いすべきと考えた。その結果設けた指標は次のものである。

(e) 前書きや後書きである
(f) 明治時代より以前の古い言葉が多い
(g) 外国語が多い
(h) 数式やプログラミング言語などが多い
(i) 法律文が多い
(j) 教育現場で使いがたそうである[1]
(k) その他一定量の「本文」が認めがたい

なお, 収録サンプルの中には,「後書き」が「本文」であるテキストが存在する（LBr9_00086『あとがき大全』）。この場合は（e）ではない。引用編集形式であるため，（b）の指標が付与されている。

２.２　アノテーション作業の概要

作業対象と内容は次のとおりである。

●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）の書籍テキスト。
●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（平均3,000語）を範囲とする。
●作業ファイル：サンプルを取得した書籍の紙面コピーを参照する。
●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。
●内容：
下記に該当する場合に指標を付与する。排他的ではなく該当するものすべてを付与する。
(a)対話系，(b)引用系，(c)視覚表現多用系，(d)データベースやリスト系，(e)前書きや後書きである，(f)明治時代より以前の古い言葉が多い，(g)外国語が多い，(h)数式やプログラミング言語などが多い，(i)法律文が多い，(j)教育現場で使いがたそうである，(k)その他一定量の「本文」が認めがたい

３．アノテーション作業結果

３.１　分類指標の付与結果

今回の対象データである 1,664 テキストに対する NDC 別分類指標の付与結果を表 1 に示す。分類指標は排他的ではないため合計は 1,664 を超える。図書館サブコーパス収録サンプルの NDC 別の数と比率は，図 1 に示すとおり「9.文学」と「5.社会科学」が多い。よって，表 1 で「9.文学」「5.社会科学」が全体的に多いのは，もともとのサンプル数の比率の大きさに寄るところがある。しかしながら，図 2 の NDC 別分類指標の付与比率をみてみると，収録サンプル比率とは異なる次のような特徴を確認することができる。

[1] 厳密には，(j)は文体判断が困難な類型ではない。小中学校の教育現場等において用例表示をする際に避けた方が無難だと思われるような，例えば，暴力的な描写や性的な描写を含むものを区別するための指標である。文体情報付与のための指標という目的からは外れるが，コーパス活用のためのテキスト整理の指標として設けたものである。田野村（2009）は，そういったテキストに対し「日本語の学術的研究という観点からそれらを排除すべき理由は本来ない」が，「危うい内容のデータは排除ないし隔離するという処置を講じる必要があるように筆者には思われる」と述べている。この分類はその試みの一つになると考える。

表 1　NDC 別分類指標の付与結果（1,664 テキスト）

NDC	サンプル数	(a)対話系	(b)引用系	(C)視覚表現多用系	(d)データベースやリスト系	(e)前書きや後書きである	(f)明治時代より以前の古い言葉が多い	(g)外国語が多い	(h)数式やプログラミング言語などが多い	(i)法律文が多い	(j)教育現場で使いがたそうである	(k)その他一定量の「本文」が認めがたい
0.総記	46	9	12	5	9	8	0	0	2	1	1	1
1.哲学	75	17	20	3	10	21	1	0	0	0	1	7
2.歴史	143	32	20	20	48	26	5	0	0	0	0	6
3.社会科学	355	112	68	10	66	54	3	0	0	13	17	31
4.自然科学	120	30	18	16	35	15	0	3	4	1	1	2
5.技術	180	18	22	57	71	13	0	1	1	3	0	11
6.産業	54	8	2	13	25	5	0	0	0	1	0	3
7.芸術	177	45	18	59	35	12	0	0	0	0	3	11
8.言語	86	11	14	1	39	7	0	16	1	0	0	5
9.文学	339	77	25	1	16	55	5	0	0	0	115	50
n.なし	89	9	10	30	26	5	1	1	0	0	3	5
計	1664	368	229	215	380	221	15	21	8	19	141	132

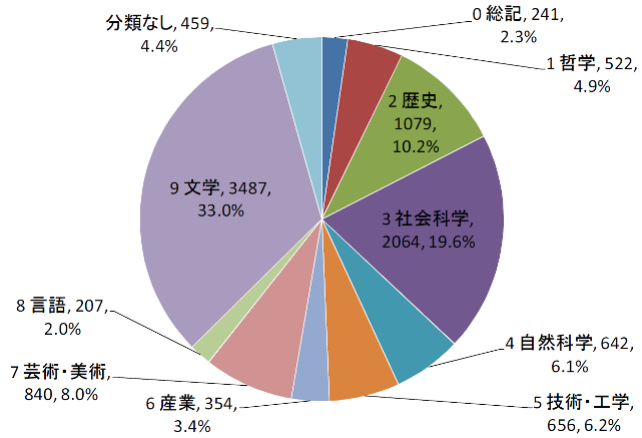


図 1　図書館サブコーパス収録サンプルの NDC 別の数と比率
（DVD 収録『『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版』（2011 年）より）

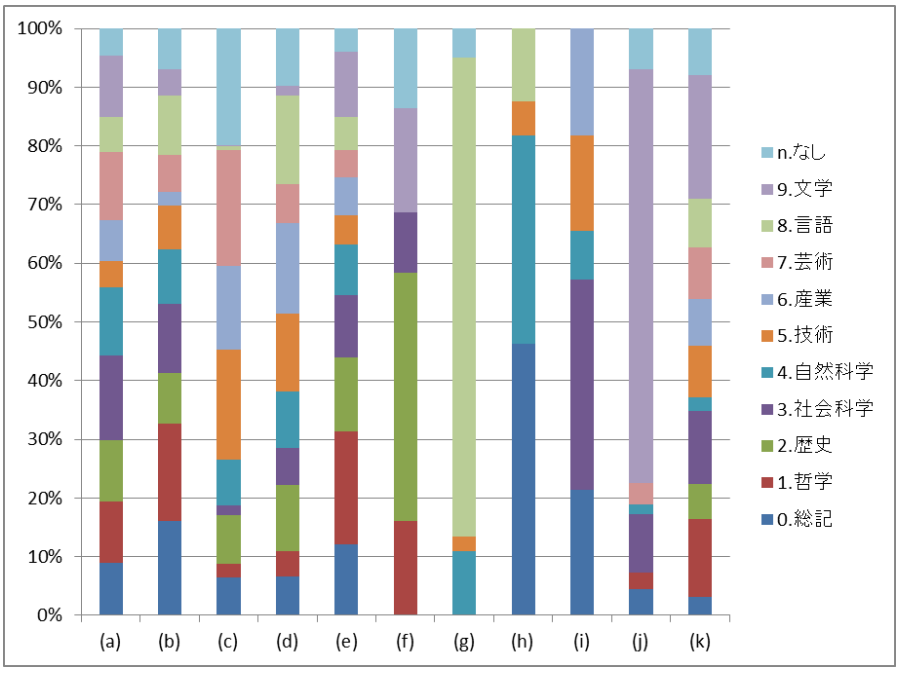


図 2　NDC 別分類指標の付与比率（1,664 テキスト）

・指標の(a)(b)(e)は，NDC の区別なく，広く用いられている形式である。
・指標の(c)は，「5.技術」「7.芸術」，「n.なし」に多い。これは「5.技術」にコンピュータのマニュアル等が多く，そこにキャプチャ画面が多用されていること，「7.芸術」に図画が多く提示されていること，「n.なし」にカタログ状の紙面が多いことに起因すると思われる。
・指標の(d)は，「6.産業」と「8.言語」に多い。「6.産業」には用語解説が，「8.言語」には辞書形式がそれぞれ多用されることによるものと考えられる。
・指標の(f)は，「3.歴史」が多くを占める。歴史を扱うテキストの中で古い言葉が多用されるからであろう。ただし，該当サンプル数はそもそも少ない。
・指標の(g)は，「8.言語」が大半を占める。外国語のテキストで、外国語が本文に入り込んでいるケースが多いためであろう。ただし, 該当サンプル数はそもそも少ない。
・指標の(h)は，「0.総記」「4.自然科学」が大半を占める。前者にはコンピュータのプログラミング言語が，後者には数式が，それぞれ多用されているためであろう。
・指標の(i)は，「3.社会科学」の比率が高い。法学を含むこの NDC で，法律が多く引用されていることがうかがえる。
・指標の(j)は「9.文学」が非常に多くを占める。暴力的な描写や性的な描写を含む小説がこの NDC に入っているためである。

４．おわりに

BCCWJ に収録する図書館サブコーパスの有効活用を可能とするために，「特徴的な類型のテキスト」に分類指標を人手付与した作業結果を報告した。多種多様な形式をもつサンプルがどの NDC にどの程度収録されているかを明らかにした。特に，テキスト形式の選択に関し，(a)対話系，(b)引用系のテキスト形式は NDC の区別なく多用されていること，(c)視覚表現多用系は，「5.技術」「7.芸術」に，(d)データベースやリスト系は，「6.産業」「8.言語」に選択的に多用されていることを確認することができた。

プロジェクト終了に際し, BCCWJ の図書館サブコーパスに収録される 10,551 サンプルの全ての分類結果についてもまとめ中である。その成果報告と分類結果を近いうちに公開する予定でいる。

本成果に基づき，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具体的に考えていきたい。

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「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.

Lee, Y. D. (2001) Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle, *Language Learning & Technology*, 5:3, pp.37-72.

柏野和佳子，奥村学(2012)「書籍テキストへの分類指標人手付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『言語処理学会第 18 回年次大会予稿集』pp.1260-1263.

柏野和佳子，立花幸子，保田祥(2012)「書籍テキストをその形式，内容，表現に関わる特徴によって分類する」『ことば工学研究会』41,pp.21-29.
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(2012a)「テキストの硬さと軟らかさの考察 ー『現代日本語書き言葉均衡コーパス』の収録書籍を対象にー」『第 1 回コーパス日本語学ワークショップ』予稿集,pp.131-138.
柏野和佳子，立花幸子，保田祥，飯田龍，丸山岳彦，奥村学，佐藤理史，徳永健伸，大塚裕子，佐渡島紗織，椿本弥生，沼田寛(2012b)「書籍テキストへの文体情報付与の試み」『第 2 回コーパス日本語学ワークショップ』予稿集,pp.155-164.
柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠 (2009). 『『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例』, 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-01), 特定領域研究「日本語コーパス」データ班.
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.
田野村忠温(2009)「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」『人工知能学会誌』24-5,pp.647-655.
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』pp.314-317.
保田祥，柏野和佳子，立花幸子(2012)「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」『ことば工学研究会』41,pp.3-10.
保田祥，柏野和佳子，立花幸子，丸山岳彦(2012a)「「語り性」を有する書きことばの典型例の分析」『第 1 回コーパス日本語学ワークショップ』予稿集， pp.139-146.
保田祥，柏野和佳子，立花幸子，丸山岳彦(2012b)「「語りかけ性」を有すると判断される書きことばの表現」『第 2 回コーパス日本語学ワークショップ』予稿集, pp.43-50.

関連 URL

EAGLES http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html
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/

複合機能表現「という」の分類にみる MCN コーパスの方法論検証

叢悠悠（お茶の水女子大学理学部）
田中リベカ（お茶の水女子大学理学部）
中村絢子（お茶の水女子大学理学部）
酒向美帆（お茶の水女子大学理学部）
佐宗智子（お茶の水女子大学理学部）
清水蘭（お茶の水女子大学理学部）
劉月晴（お茶の水女子大学理学部）
川添愛（国立情報学研究所）
戸次大介（お茶の水女子大学院人間文化創成科学研究科／国立情報学研究所）

Methodology of the MCN Corpus in the Classification of a Functional Compound “toiu”

Yuyu So (Faculty of Science, Ochanomizu University)
Ribeka Tanaka (Faculty of Science, Ochanomizu University)
Ayako Nakamura (Faculty of Science, Ochanomizu University)
Miho Sako (Faculty of Science, Ochanomizu University)
Tomoko Saso (Faculty of Science, Ochanomizu University)
Ran Shimizu (Faculty of Science, Ochanomizu University)
Yuechin Ryu (Faculty of Science, Ochanomizu University)
Ai Kawazoe (National Institute of Informatics)
Daisuke Bekki (Graduate School of Humanities and Sciences, Ochanomizu University / National Institute of Informatics)

１．はじめに

自然言語で記述されるテキストには、書き手にとって真であることが確実な情報と、そうでない情報が混在する。例えば、「太郎が結婚した」という命題について、以下のような文が考えられる。

1. 太郎が結婚した。
2. 花子は太郎が結婚したと言う。
3. 噂によると、太郎が結婚したという。
4. 仮に、太郎が結婚したとする。

このうち、1.は事実だが、2. 3. 4.は事実として捉えてはならない。なぜなら、書き手にとって命題の真偽がはっきりとしておらず、書き手が命題に対して何らかの心的態度を持っているためである。このように、ある命題に対する書き手の認識や態度を表す言語表現をモダリティという。モダリティ表現には、「…という」「…かもしれない」といった様相表現、「…でない」のような否定表現、「～なら…」という条件表現などがある。人間はモダリティ表現から情報の確実性を判断しているのである。

また、上に挙げた例文は、Web 上で「太郎　結婚」というキーワード検索を行ったとき

にヒットする可能性のあるものとしてみることもできる。膨大なテキスト情報の中から確実性の高い情報を選り抜くためには、事実である 1.と、そうでない 2. 3. 4.を区別したいものである。すなわち、モダリティ表現に着目することが必要になる。

本研究では、MCN コーパス（川添ら（2011））のアノテーションガイドラインで使用している言語学的テストの改良を行っている。MCN コーパスは、モダリティ表現に意味アノテーションを付与した言語データである。具体的には、各表現の用法ごとの分類が示されたガイドラインを用いて、テキスト中の表現にラベル付けしたものである。言語学的テストとは、理論言語学の知識に基づいて作成されたテストで、文または文の一部の容認性や適切性を判定するものである。MCN コーパスのアノテーションにおいては、言語学的テストとして「ネガティブテスト」（田中ら（2012a, 2012b））を採用しており、各表現に対するネガティブテストを用意したガイドラインを作成している。本論文では、様相表現「（と）いう」「とする」に対する最新のガイドラインについて、その問題点を考察する。

本論文では以下、第 2 節で MCN コーパスのガイドラインで用いているネガティブテストの概要を述べる。第 3 節では、実際のアノテーション作業で、アノテータ間の意見が分かれやすかった表現について論じる。

２．MCN コーパスのガイドラインにおける言語学的テスト

MCN コーパスのアノテーションでは、アノテータの判断の不一致を避けるために、ネガティブテストを導入している。ネガティブテストは、「文中の表現を別の表現に置き換えたときに文として成立しない、あるいは意味が変化する場合、その用法としてアノテーション不可能（つまりそのカテゴリに分類されない）」という形式をとる。ここで、「置き換え不可能であればアノテーション対象ではない」としているのは、「置き換えが可能」という判断よりも、「置き換えは不可能」という判断の方が、アノテータ間での一致度が高いという傾向が見られるためである（田中ら（2012a, 2012b））。ネガティブテストで置き換え不可と判定された場合、その分類に属さないことが断定できるため、これを用いたアノテーション作業では、消去法で分類先を一つに特定することになる。一つの表現に対する分類先が一意に決定されることは、一貫性のあるコーパスを構築するにあたって重要である。消去法を行った結果として複数のカテゴリが残った場合は、それらのうち、本来の分類先でないカテゴリのテストが不適切であることを意味する。

MCN コーパスのアノテーションで使用しているガイドラインは、「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」（川添ら（2011））をもとにしている。これは、言語情報の確実性に関わる表現にアノテーションを付与し、機械による確実性判断の基盤となるコーパスを構築するために作成されたものである。もともとのガイドラインには、各言語表現について用法別のカテゴリが例文や統語環境などとともに示されている。しかし、これらの基準だけでは、ある表現がどのカテゴリに属するかを判断できない場合がある。例えば、以下の文中の「という」に対し、例文ベースのガイドラインを用いて、他人の認識を表す「（と）いう」としてアノテーション可能かを考える。

1. 太郎が責任をとるべきという人はどうかしている。
2. 太郎が結婚したという話だ。

ガイドラインの記述：

他人の認識 【(と) いう】
分類 : 他人の認識 (他人の報告する事柄や、命題の真偽に関する他人の判断を表す表現)
例 : 今年のインフルエンザの流行は全国的に遅れているという。

1.および 2.の「という」は、ともに名詞句を修飾しており、例文と異なる形をとっているようにみえるが、ガイドライン設計者は、1.は他人の認識としてアノテーション可能であり、2.は不可能であると意図している。両者の違いは、専門的な知識を有しない一般のアノテータが容易に見出せるものではない。

そこで、田中ら (2012a, 2012b) は、「(と) いう」を「との」と「(と) 述べる」のそれぞれに置き換えるテストを作成し、どちらに置き換え可能かで別々のカテゴリに分類するようにした。上の文にこの二つのテストをそれぞれ適用すると、以下のようになる。

1a. 太郎が責任をとるべきとの人はどうかしている。
1b. 太郎が責任をとるべきと述べる人はどうかしている。
2a. 太郎が結婚したとの話だ。
2b. 太郎が結婚したと述べる話だ。

1.の「という」は「と述べる」の置き換えは可能だが、「との」に置き換えると不自然な文になる。一方、2.は「との」に置き替え可能だが、「と述べる」に置き換えることはできない。このように、テストを用いると判断がしやすく、アノテータ間の一致度も高くなる場合が多い。

無論、複数のアノテータによるアノテーション結果が完全に一致するようなテストを作成することは難しい。例えば、「今回の募金は 10 万円を目標とする」という文に対し、「とする」を「にする」に置き換えたとき、容認できるか、不自然に感じるかは人それぞれである。テスト適用時に生じる変化は、個々人の言語感覚に問うものであり、容認の可否が分かれるのは避けられない。また、「太郎は花子を許さないという」という文に、テストとして「わざわざ」を挿入したとき、明らかにニュアンス上の変化が起こるが、それが意味に影響するか否かの判断はアノテータに委ねられる。このような置き換えや挿入による語感の変化がどれだけ大きいと「置き換え不可」となるかを明確に定義するのは、事実上不可能である。

それでもテストを採用しているのは、実際にテストを用いてアノテーションを行った際に、例文ベースのガイドラインを使用するのに比べてアノテータの判断が容易になり、一致度も向上する傾向があるからだ。また、テストを使用しない場合は、例文との類似性のみから判断するほかないが、そうして得られたアノテーション結果に確たる根拠は見出せない。テストを用いたアノテーションは、より信頼性のあるコーパスを得るのに必要な手法であると考えられる。

３．現在のガイドラインの問題

表 1 は、「(と) いう」「とする」のアノテーションに対する最新のガイドラインから一部を抜粋したものである。ガイドラインの完成度は、テストをもとにしてアノテーションを行った際に、分類先が一つに特定できたか、また、アノテータの判断がどれだけ一致しているかによって測られるが、現在のガイドラインは改良の途上にあり、問題点が多くある。本節では以下、アノテーション結果の不一致を招く要因となるもののうち、アノテータの

判断に対する影響が顕著であった 4 つの問題を取り上げる。

表 1：最新のガイドライン（一部抜粋）

sem	表現	別表記	特徴	例文	テスト	統語環境	備考
3	いう		人あるいは物と、その名前を関連づける。「と」の前には原則として固有名詞であり、「インスリン」のような専門用語の一般名詞が現れることもある。 語用論的には、「という」の前の固有名詞あるいは専門用語の指示する対象が、話し手か聞き手の少なくとも一方にとって馴染みのないものであることを表す。	その人は山田という。 そのホルモンはインスリンという。 初めまして。私、山田といいます。 その人は名前を山田という。 そのホルモンは名前をインスリンという。 初めまして。私、名前を山田といいます。 日本には、富士山という山があります。 富士山という山は、どこにあるのですか？ 長崎の名物に、トルコライスというものがあります。	「呼ばれ（てい）る」に置き換えて意味が変化する場合はこのカテゴリではない。 （※主語が 1 人称の場合は違和感がある。） 「～という」の前に「名前を/名を～という」のように「名前を/名を」を補って意味が変化する場合はこのカテゴリではない。	[動作主(NP)]が[名前(NP)]という [動作主(NP)]が名前を/名を[名前(NP)]という	「という5」との区別がつきにくいが、「という」の前が固有名詞でも専門用語でもない場合は「という3」ではないと考えてよい。
2	とする		仮想的な状況を記述する。「想定する」「仮定する」に近い意味をもつ。	太郎が犯人だったとする。その場合、アリバイはどう説明するんだ？ 無人島に、一つだけ物を持っていけるとしよう。君は何を持っていく？ 運転中に視界が悪くなったとします。その場合はどうすればよいでしょうか。 来年三月までの収入の合計を 300 万円とする。その場合、税金はいくらになるか。 直線 AB 上の点を Q とする。	「とする」を「想定する」「仮定する」のいずれにも置き換え不可、あるいは置き換えて意味が変化する場合はこのカテゴリではない。	[S]とする [NP]を[NP]とする	

3.1 「いう 1」「いう 2」の区別について

本ガイドライン中の「いう 1」は、「言葉を発するという意図的な動作を表す」とある。また、動作主を特定の人物以外に「世間一般、人々、みんな」と設定し、「多くの場合は明示される」としている。一方、「いう 2」は、「伝聞の意味を持つ」とある。動作主に関しては、「明示されない」としている。また「～によると」とともに使われる場合が多いとされている。

しかし、動作主が「（と）いう」の直前にない場合、「いう 1」か「いう 2」かの決定が困難なことがあった。以下は、家庭訪問の実態を題材にした新聞記事からの引用である。

1. 家庭訪問は、明治初期に不就学児を登校させるよう親を説得する目的で始まった。「師範学校付属校のような中核校から周辺に広がっていったのでは」と佐藤教授はみている。また、家庭と学校の不干渉が徹底している欧米では、家庭訪問は基本的にないという。

表 2：「いう 1」および「いう 2」

表現	特徴	例文	テスト
いう1	言葉を発するという「意図的な動作」が意味の中心である。 また「NP が」という形の項として動作主を要求する（多くの場合節内に動作主が明示される）。 ※動作主が「世の人」「人々」「みんな」「誰か」である場合、「という2」と意味的に近くなるが、これは「という1」である。	太郎は「昨日渋谷で花子を見た」と言う。 花子はまだ怒っているようで、太郎を絶対に許さないという。 「太郎が責任をとるべき」と言う人は、どうかしている。 「叶わない夢はない」と人はいう。 花子は、太郎を天才だと言う。 その時警官が通りかかったことは、幸運だったというしかない。 花子が「おいしい」という店には行かない方がいいよ。	「話す」「主張する」「述べる」「表現する」「評価する」「判断する」のいずれにも置き換え不可、あるいは置き換えて意味が変化する場合はこのカテゴリではない。 「わざわざ」「口に出して」「あえて」「しつこく」のいずれを挿入しても意味が不自然になる場合はこのカテゴリではない。
いう2	伝聞の意味をもつ。「言葉を発する動作」よりも、むしろ「言説の存在」あるいは「言説が流布している状態」を表しているもの。 動作主が明示されない。 語用論的には、話者が間接的な言語情報として得たことを表す（直接経験して知っていることについては使わない）。 情報源を表す「～によると」と共起することが多い。情報源が明示されない場合、「世間一般」「人々」「専門機関あるいは公的機関の公式発表」である。	ニュースによると、インフルエンザが流行っているという。 警察の調べでは、男は以前から現場付近で目撃されていたという。 駅前の焼肉屋は、このあたりで一番おいしいという。 日本人の9割が何らかのストレスを抱えているという。 世界には自分と同じ顔の人間が7人はいるという。 私たちの普段の生活の中にも、空海が中国からもたらしたというものがあります。それは一体、何でしょう。 私たちの普段の生活の中にも、空海が中国からもたらしたというものがあります。それは一体、何でしょう。	「そう（だ）」と置き換えて違和感がある場合はこのカテゴリではない。 「いわれる」「いわれている」に置き換え不可、あるいは置き換えて意味が変化する（尊敬の意味になる等）場合はこのカテゴリではない。

例文 1.の「という」は、動作主が明示されておらず、「いう 1」と「いう 2」両方のテストが適用可能であるため、どちらか一方に分類することは難しい。ここで注目すべき点は、「欧米では、家庭訪問は基本的にない」という言葉を佐藤教授が実際に発したのか、あるいは話者が他の情報源から伝聞したものなのかというところである。この前後の文を参考にすれば、どちらか特定できる可能性もあるが、実際のアノテーション時におけるアノテータの負担を考慮すると、一つの命題を判断するために広範囲の文章を参照するのはなるべく避けたいものである。

次に、1.の最後の一文を次のように換えてみる。

2. 家庭訪問は基本的に「ない」<u>という</u>。

こちらは、かぎ括弧をつけたことによって、一見佐藤教授が発した言葉のように感じられる。しかし、このかぎ括弧は話者が強調のためにつけたとも考えられ、その場合は佐藤教授から直接聞いた言葉でない可能性がある。かぎ括弧がせりふを表すものであるか、強調のためにつけられたものであるかは、文章全体を読んでもなかなかわかるものではない。新聞等では、既出の人物の発した言葉が、主語を伴わず、かぎ括弧に括られた形で出現することが多々あるが、その場合にこのような問題に直面してしまう。

また、ガイドラインにおいては、「いう 1」の動作主として「世間一般、人々、みんな」等が挙げられている。しかし、これらが動作主となっている場合は、主語を省略する傾向がある。特に、新聞等においてはそれが顕著で、「(と) いう」の前にある命題が「動作主を明示していないが、世間一般で言われていること」であるケースが少なくない。

3. よく「朝食を摂る子供は成績が良い」<u>という</u>。それは本当なのだろうか?

かぎ括弧の中は、世間一般でよく言われているという点においては、不特定多数の人物が意図的に発している言葉である。しかし、この文には動作主が明示されておらず、話者が伝聞したことのようにもとれるため、「いう 2」としてもアノテーションできてしまう。実際、「いう 1」と「いう 2」のテストを適用すると、「いう 1」のテスト「よく～と話す」よりも、「いう 2」のテスト「よく～といわれる」の方が自然である。

これらの問題を根本的に解決する方策として、省略されている動作主を補うことができるか (補ったことによって文が不自然にならないか)、といったテストを追加することが考えられるが、文脈に応じて適切な動作主を補うのは、多くのアノテータにとって容易でないことが推測される。

３.２　「(と) いう」「とする」の命題 / 名詞句の判断

「(と) いう」と「とする」のどちらにも共通して、直前が命題か名詞句かの判断が必要となるカテゴリがある。例えば、「3 は奇数である」「インフルエンザが流行している」は命題であり、「私の赤いドレス」「野球をすること」は名詞句である。

ここで、三平方の定理「直角三角形の斜辺の二乗は他の二辺の二乗の和に等しい」を考える。「三平方の定理」は紛れもなく名詞句である。一方、直角三角形の斜辺を c、他の二辺をそれぞれ a, b とおくと、この定理は「$a^2 + b^2$ は c^2 に等しい」と表せるが、これは命題である。しかし、同じ等式を意味する「$a^2 + b^2 = c^2$」が命題であるか、名詞句であるかの判断は困難である。そのため、「$a^2 + b^2 = c^2$ という式」の「という」に対して、「いう 5」「いう 7」の二つのカテゴリが候補となってしまう。

別の例として、「自分を含めて客が 4 人というライヴに行ったことがある」という文を考える。「という」の前の部分は、一見名詞句のように見えるが、これは「自分を含めて客が 4 人である」から「である」が省略された形となっており、命題であるとされる。

このように、数式の形になっているものや、語尾の「である」が省略されているものは、統語環境を判定できず、分類の決定時に混乱を招く恐れがある。現在のガイドラインにおいては、命題と名詞句に対する定義が不十分であるため、今後より幅広い表現に対応できるよう改善する必要がある。

表 3：「いう 5」および「いう 7」

表現	特徴	例文	テスト
いう5	「という」の前の NP として「今日」「お前」のような直示的表現や、「東京」のような固有名詞、「コーヒー」のような一般名詞が現れることが可能。（前の NP の意味が後ろの NP の意味に含まれていることを表す） 前方の NP の意味が後方の NP の意味に含まれていることが常識的に明らかな場合…強調の効果 後方の NP が「妻」や「生きがい」などのロール概念を表す場合…前方の NP のどの側面に着目するかを限定する効果 前後の NP 間に意味的な包含関係があることが明らかでない場合…それらの間に包含関係があるとする（話者の）主張を強調する効果	今日という日を忘れないようにしよう。 お前という人間がわからなくなった。 長年住んでいるが、東京という町には親しみがわいてこない。 コーヒーという飲み物は実に奥が深いね。 犬や猫の目には、人間という動物がどのように映るのだろう。 ボランティアという生きがいに出会ってから、毎日が楽しくなりました。 相手の女性も、私という妻の存在を知らないわけがありません。 私は、家族という重荷を背負って生きていくのには向いていない。 経済成長という病（*本の名前） アメリカという記憶（*本の名前） 人間というものは、よほどのことがない限り考えを変えようとしない。 それが男というものだと割り切るしかない。 まったく、限度というものを知らないんだから。	（補助テスト*） 「NP という NP」の形で、 (1)前方 NP が固有名詞/専門用語でない場合…という5 (2)固有名詞/専門用語の場合… ①前方 NP の指示対象が話し手・聞き手の少なくとも一方にとって馴染みのないものである→という3 ②双方にとってなじみのあるものである場合→という5
いう7	特に意味的な内容はなく、関係節的な特徴を持つ[命題(S)]と係り先の[NP]との間のつなぎとしての役割のみを持つ。	子供が高校生や大学生という世帯は、全世帯の中でも特に出費が目立つ。 僕もデビューする前は、一年間収入が全くないという時期を過ごしたこともあります。 この人となら結婚してもうまくいくだろう、という人がなかなか現れないのよね。	「（名詞 or 状詞）＋という NP」の場合：「（名詞 or 状詞）＋（の or な）NP」に置き換え不可の場合はこのカテゴリではない。 「（動詞 or 形容詞）連体形＋という NP」の場合：「（動詞 or 形容詞）連体形＋NP」に置き換え不可の場合はこのカテゴリではない。

3.3 「いう 6」の抽象名詞リストについて

本ガイドラインでは、「[命題] という [名詞句]」という形のものを、名詞句が抽象名詞であるかどうかで、それぞれ「いう 6」「いう 7」に分類している。文中の名詞が抽象名詞であるかを判断する際には、表 4 の抽象名詞リストを参照している。しかし、実際のアノテーション作業において、文中の名詞がリストにないが、リスト中のほかの抽象名詞と近い意味を持つ場合、判断が困難であった。例えば、「教室にエアコンを設置してほしいという要望があった」の「要望」はリストに挙げられていないが、それに類似した表現「要求」「声」は含まれている。このように、あらかじめリストの形で提示できる抽象名詞の数は限られてしまい、現実世界の表現すべてに対応するのは原理的に不可能である。

表 4：抽象名詞リスト

事象	事実、真実、事態、事件、こと、出来事、事情、状況、状態、症状、人事、例、事例、判例、現象、問題
ことば	言葉、格言、名言、せりふ、文、文句、文言、遺言、言い方
情報媒体	情報、ニュース、話、記事、報告、知らせ、便り、メール、電話、口コミ、噂、報道、記録、声、音、声明、手紙
言語行為（発話内行為）	命令、忠告、約束、説明、発言、発表、指示、主張、提案、提言、要求、決定、指摘、質問、答え
思考行為	意図、理解、認識、反省、考え、意見、見解、結論、仮定、前提
具体的行為	行動、行為、作業、仕事、習慣
感情	感情、気持ち、意識、感じ、不安、希望、不満、欲望、恐れ、寂しさ、幸せ、喜び、悩み、心配、懸念、疑問、空しさ、自信
概念	概念、思想、主義、知識、理屈、理由、目的、説、学説、理論、法則、印象
モダリティ	可能性、見込み、危険
表出	態度、そぶり、ふり、表情
内容	内容、あらすじ、（話の）流れ、シナリオ
手順	作戦、手順、順序、順番、手続き、計画、企て、予定、プロジェクト
性質	性質、特性、側面、一面、点、利点、長所、短所、特徴、観点

「いう 6」のテスト：
「という」の後の NP が命題（あるいは命題の集合）を意味する抽象名詞（句）でない場合はこのカテゴリではない。
（※ただし、「というもの」の場合は、「もの」と同一指示関係を持つ名詞の種類が抽象名詞（句）であるかを考える）

「いう 7」のテスト：
「（名詞 or 状詞）＋という NP」の場合：
「（名詞 or 状詞）＋（の or な）NP」に置き換え不可の場合はこのカテゴリではない。
「（動詞 or 形容詞）連体形＋という NP」の場合：
「（動詞 or 形容詞）連体形＋NP」に置き換え不可の場合はこのカテゴリではない。
（※「ということ」の場合は、「こと」が抽象名詞句のいずれかに解釈可能な場合が多い）

３.４ 「として 3」の慣用表現について

3.3 項と類似した問題が「として 3」でも生じる。このカテゴリは、「結果として」「時として」などといった慣用表現に特化したもので、例文の項目に代表的な表現がいくつか提示されており、例文に含まれていない表現については、慣用表現かどうかの判断がアノテータに委ねられている。しかし、慣用表現かどうかについての認識には個人差がある。例えば、「感じとして」という表現について、「慣用表現である」という意見と、「単なる言い回しである」という考えに分かれる傾向があった。そもそも、慣用表現と言い回しをどう区別するのかも自明ではない。

表 5 :「として 1」および「として 3」

表現	特徴	例文	テスト
として 1	「～という位置づけで」とほぼ同義。 「～という立場で」、「～という役割で」、「～という名目で」などと言い換えると自然な場合がある。	山田氏を課長として採用する予定。 大人として恥ずかしくないのか。 賞金として二十万円が贈られます。 働きがいのある会社として注目されている企業。	「[NP を][NP として]」という形で出現している場合:「[NP として][NP を]」の順序に入れ替えることができない、入れ替えると意味が変化する場合はこのカテゴリではない。 [NP として]を省略することができない、省略すると意味が変化する場合はこのカテゴリではない。 [NP として]単独で出現した場合:「～という位置づけで」「～という立場で」「～という役割(役職)で」「～という名目で」のいずれにも置き換え不可、あるいは置き換えて意味が変化する場合はこのカテゴリではない。
として 3	慣用的な表現	原則として、部外者の立ち入りを禁ずる。 結果として、その年の合格者はたったの五人だった。 人生には時として、何をしてもうまくいかないことがある。 誰一人として理解してくれない。 一日として忘れたことがない。 遅々として進まない。 彼の行方は、杳として知れない。	

一見、単なる言い回しとの区別がしにくい慣用表現であるが、単独のカテゴリを設けるのは妥当なのだろうか。以下の文を考える。

1. それは原則としてしっかり押さえておく必要がある。
2. 原則として部外者の立ち入りを禁ずる。

1.の「として」は、位置づけの働きを持つので、「として 1」に分類される。一方、2.の「原則として」は慣用表現である。こちらは副詞的な役割であり、文中から除いたときに、程度上の変化はあっても、重要な情報の欠落は生じない。このことから、慣用表現の「として」は、ほかのカテゴリと明白に異なる機能を持つため、特別なカテゴリを設ける必要があると筆者らは考えている。

文中の表現が慣用表現であるか否かの判断は、一般のアノテータには困難である。そのため筆者らは、慣用表現を列挙したリストをガイドラインで提示し、「リストに含まれない表現は慣用表現ではない」というテストを今後追加する予定である。このようなリストの作成にあたって、3.3 項で述べたような問題が危惧されるが、慣用表現は抽象名詞に比べると数が限られている。よって、リストにはないが慣用表現であると判断した場合はテスト設計者にフィードバックし、その過程を通してリストの内容が収束していくことが期待で

きると考えている。

4. おわりに

本稿では、MCN コーパスにおけるアノテーションの問題点を考察した。今後、より一貫性のあるアノテーション結果が得られるよう、テストの改良を行う方針である。

文献

田中リベカ、小池恵里子、戸次大介、川添愛（2012a）「言語学テストに基づく意味アノテーションのガイドライン設計―確実性判断に関わる表現を中心に」言語処理学会第 18 回年次大会発表論文集, pp.401-404.

田中リベカ、川添愛、戸次大介（2012b）「MCN コーパス：言語学的テストに基づくモダリティ・アノテーションの理論と実証」国立国語研究所第 2 回コーパス日本語学ワークショップ予稿集, pp135-144.

川添愛、齊藤学、片岡喜代子、崔栄殊、戸次大介（2011）「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」Technical Report of Department of Information Science, Ochanomizu University, OCHA-IS 10-4.

係り受けアノテーション基準の比較

浅原 正幸 (国立国語研究所コーパス開発センター) *

Comparison of Syntactic Dependency Annotation Schemata

Masayuki Asahara (Center for Corpus Development, NINJAL)

1. はじめに

言語処理の分野でアノテーションデータに基づく統語解析の研究が盛んにおこなわれている。句構造もしくは係り受け構造が付与されたコーパスアノテーションに基づいて、さまざまな統語解析アルゴリズムと構造学習手法が提案されている一方、アノテーションの基準そのものに興味を持つ者は少ない。

英語において係り受け解析器の開発は、句構造がアノテーションされた Penn Treebank (Marcus et al. (1993)) を主辞規則 (Head percolation rules) などにより変換した係り受けアノテーションに基づいて行われている。主辞規則は係り受け解析アルゴリズムの計算量の観点から非交差制約 (projective) に基づいたもの (Magerman (1994), Collins (1999), Yamada and Matsumoto (2003)) が多く、Wh 疑問文・話題化 (topicalization)・分裂文 (cleft)・並列構造などの長距離係り受け関係については単純化されている。係り受け解析器の誤りの多くはこのような係り受け関係であるが、アノテーションの単純化による限界という指摘もあり、Johansson and Nugues (2007) は並列構造や従属節に対する係り受け関係の再定義を行い、分裂文や空所 (gapping) を Penn Treebank に付与されている二次辺 (secondary edge) や痕跡 (trace) の情報を用いて精緻化した。

日本語では文節係り受け構造が京都大学テキストコーパス、KNB コーパス (Kyoto-University and NTT Blog コーパス)、日本語話し言葉コーパス、現代日本語書き言葉均衡コーパスに付与されているが、ほとんどの係り受け解析器が京都大学テキストコーパスのアノテーションに基づいて構成されている。本稿では日本語で係り受け解析器が誤りやすい現象は各コーパスにおいてどのようなアノテーション基準に基づいて表現されているかを明らかにするために、係り受けアノテーション基準の比較を行う。対象は京都大学テキストコーパス基準 (以下KC; 黒橋ほか (2000))、日本語話し言葉コーパス基準 (以下CSJ; 内元ほか (2004))、現代日本語書き言葉均衡コーパス基準 (以下BCCWJ; 浅原 (2013)) の三つとする。KNB コーパスのアノテーション基準は京都大学テキストコーパス基準に準じているものとする。

2. 本稿における係り受け・並列構造の表現

本稿では図 1 にように係り受け・並列構造を表現する。

* masayu-a@ninjal.ac.jp

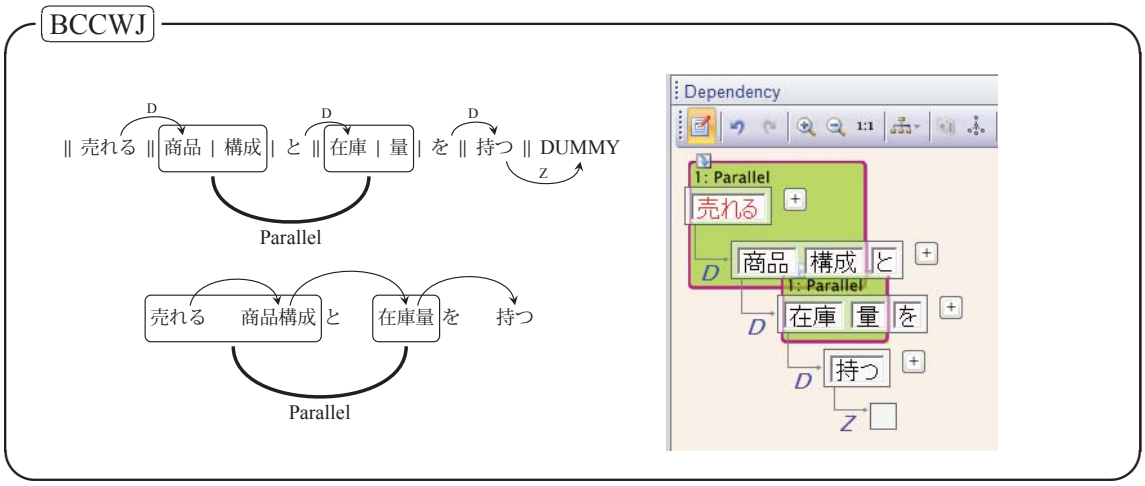


図 1　係り受け・並列構造アノテーションの表現方法

左上図中 ‖ が文節境界、| が短単位形態素境界、例文上のラベル “D” 付矢印が係り受けラベル “D” である係り受け関係を表す。例文下のラベル “Z” 付矢印が文末要素を表現する関係を表す。BCCWJ では並列構造などをセグメントとよばれる短単位形態素境界を最小単位とする範囲で複数切り出し、グループ化する。角丸四角と例文下のラベル “Parallel” 付曲線は並列構造範囲とその対応関係を表現する。他に、点線角丸四角と例文下のラベル “Apposition” 付点線曲線が同格構造範囲とその対応関係、破線角丸四角と例文下のラベル “Generic” 付破線曲線が具体例-総称間同格構造範囲とその対応関係を示す。 “DUMMY” は係り先なしを表現するための要素である。アノテーションツール ChaKi (Matsumoto et al. (2005)) 上では右図のような形で表示される。

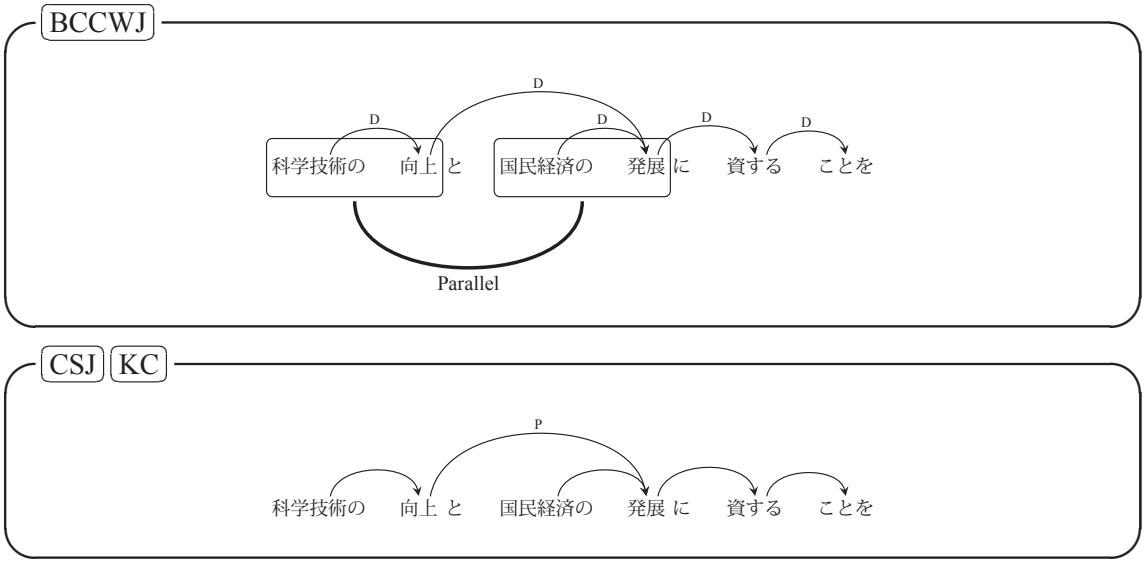
同じ文を、左下図のように略記することもある。文節境界記号と短単位形態素境界記号は範囲指定が不要な場合は省略し、文節境界の間に空白を入れて表現する。文末以外に係り先なしの関係がない場合には “DUMMY” を省略する。「通常の係り受け」は CSJ でラベルなし、KC, BCCWJ ではラベル “D” を用いるが、複数の基準の通常の係り受け関係を表現する際にはラベルなしとする。尚、CSJ においてラベル “D” は言いよどみを意味する。

3.　係り受け関係の比較

以下では三つの係り受けアノテーション基準で差異がある部分を対比的に示す。

3.1　並列構造

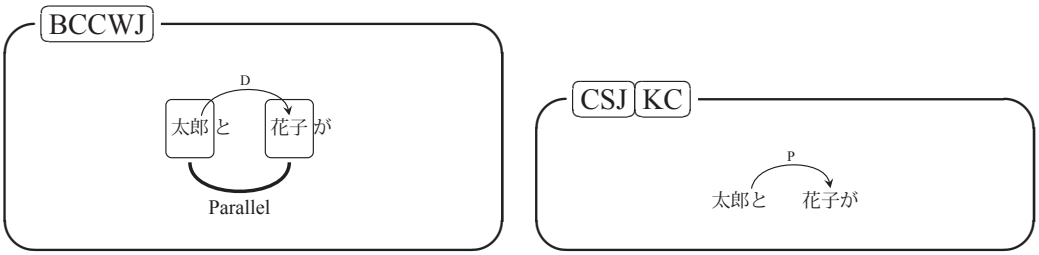
並列構造は日本語係り受け解析において頻出する扱いが難しい構造の一つである。BCCWJ のアノテーション基準の特色として、並列構造の範囲と対応する並列句を、係り受け木とは独立に範囲を付与する点がある。以下の例で、BCCWJ 基準では、係り受け関係ラベルを全て “D” としたうえで、「科学技術の向上」と「国民経済の発展」が対応する並列構造として、セグメント Parallel で切り出され、グループ化される。一方、CSJ、KC では、並列構造の構成句の最右要素動詞をラベル “P” でかける。



以下、様々な並列構造について示す。

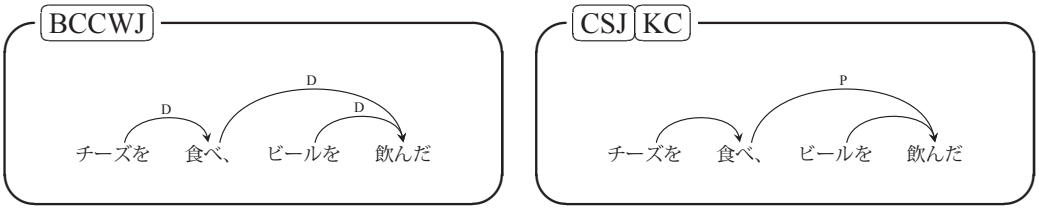
3.1.1 名詞句の並列

名詞句については、対応する名詞句をセグメント Parallel で切り出し、グループ化する。係り受け関係は通常の係り受けと同じラベル “D” を付与する。一方、CSJ KCにおいては、ラベル “P” によりアノテーションを行う。



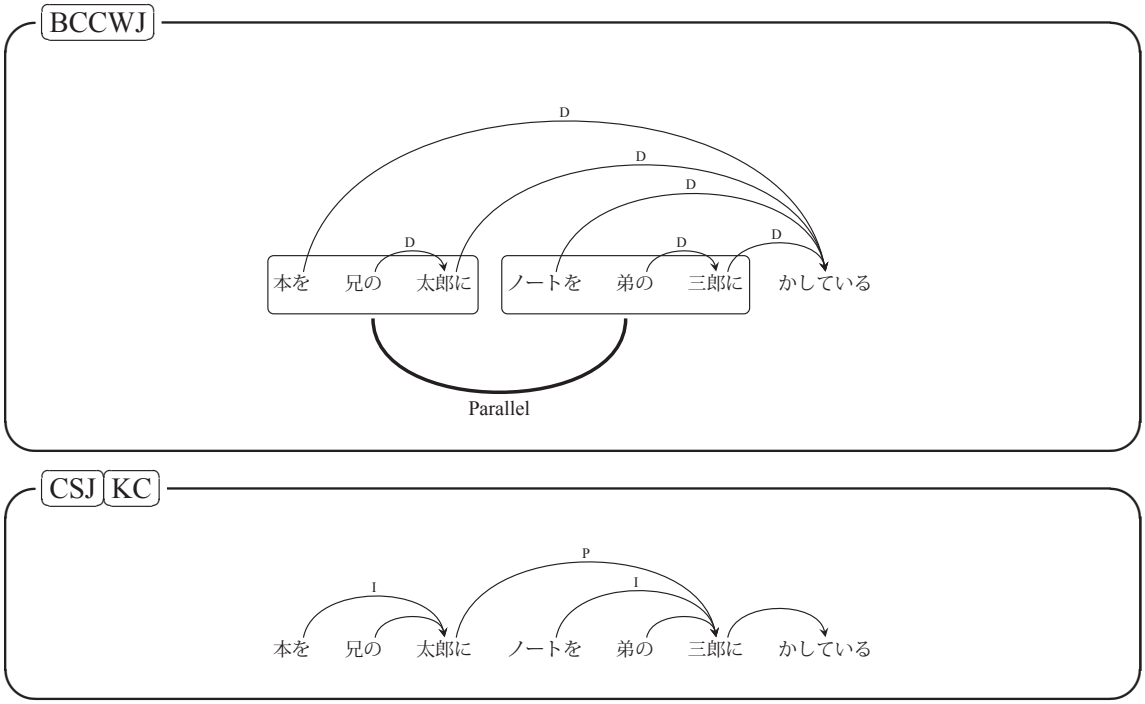
3.1.2 述語並列

CSJ KCでは一部の述語並列について、並列構造を認定しラベル “P” を付与しているが、BCCWJにおいては、全ての述語並列を並列とみなさず、通常の係り受けとして定義する。



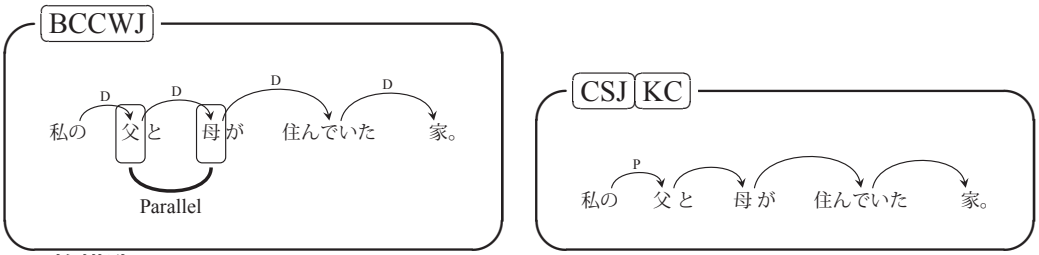
3.1.3 部分並列内の関係

CSJ KCでは以下のような構造について、非交差制約を順守するためにラベル “I” を付与し、真の係り先でないものに係けている。このようにラベルに交差の情報を持たせて、非交差条件を満たす木に変換する手法は pseudo projective と呼ばれる (Nivre and Nilsson (2005))。BCCWJにおいては、範囲を規定したうえで、通常の係り受け関係として真の係り先に係ける。



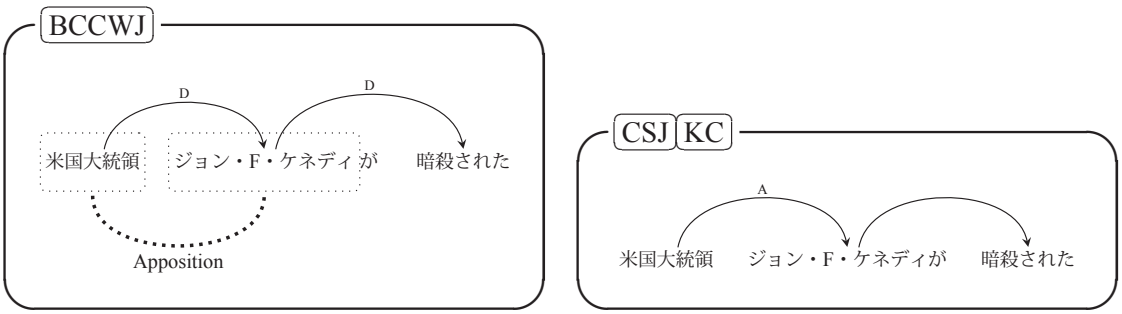
3.1.4　並列構造の複数の要素に左から係る場合

以下のように「オ（リックス）は」は「オーストリア」と「オーストラリア」の両方に係る場合には、BCCWJにおいては当該部分を並列構造範囲から外す。最左要素である「オーストリア」に係けることにより、両方に係っていることを表現する。



3.2　同格構造

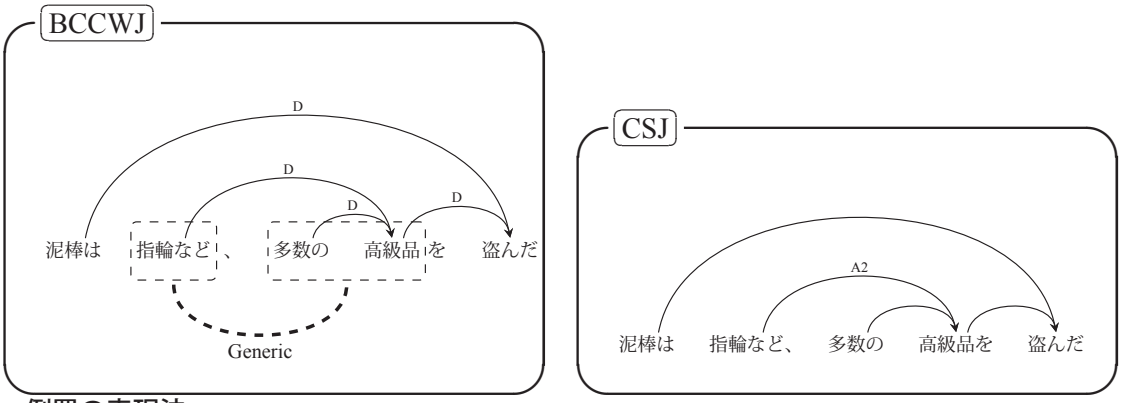
BCCWJにおいて、通常の同格関係は、対応する名詞句をセグメント Apposition で切り出し、グループ化する。係り受け関係は通常の係り受けと同じラベル “D” を付与する。一方、CSJ KCにおいては、ラベル “A” によりアノテーションを行う。



BCCWJとCSJは次に示す広義の同格を認定し、上に示した狭義の同格と区別するのに対し、KCは同格の意味を広めにとる傾向にある。

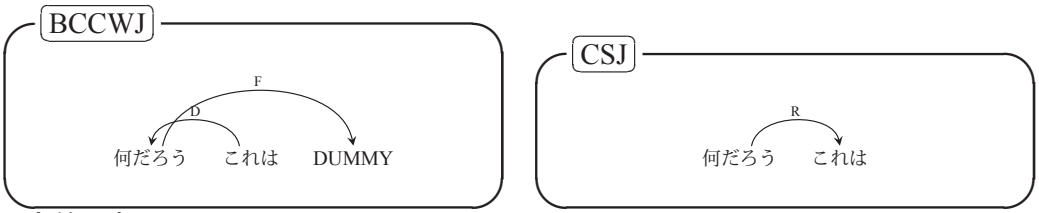
3.3 広義の同格

BCCWJとCSJは広義の同格として具体例と総称の同格関係、具体例と数詞の同格関係を狭義の同格と別のラベルで認定する。BCCWJでは、対応する名詞句をセグメント “Generic” で切り出し、グループ化する。係り受け関係は通常の係り受けと同じラベル “D” を付与する。CSJでは、ラベル “A2” によりアノテーションを行う。KC においてはこの広義の同格を識別する方策は規定されていない。



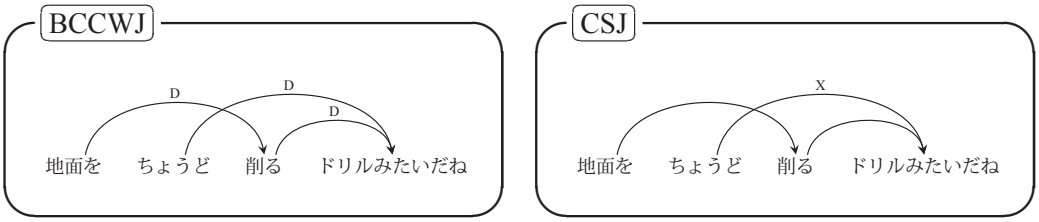
3.4 倒置の表現法

KCの基準においては、Strictly Head Final の原則から常に左から右に係る。BCCWJ CSJの基準においては、右から左に係ることを許す。CSJ では右から左に係ることをラベル “R” を用いて明示するが、BCCWJ においては特に明示しない。BCCWJ において、最初の「何だろう」は係り先なしの根ノードになるが、アノテーションツール上では末尾の DUMMY ノードに係けることにより表現する。



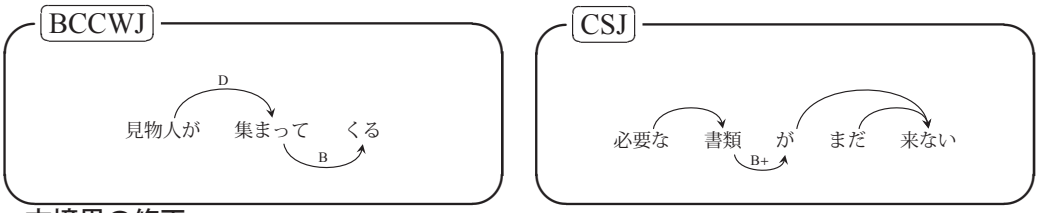
3.5 交差の表現

KCの基準においては、非交差制約の原則から係り受け関係が同格表現以外においては交差することを許さない。BCCWJ CSJの基準においては、係り受け関係が交差することを許す。CSJ では係り受け関係が交差することをラベル “X” を用いて明示するが、BCCWJ においては特に明示しない。ChaKi.NET の Dependency Panel 上では、交差があった場合には係り受け関係の色が自動的にオレンジに変更される。



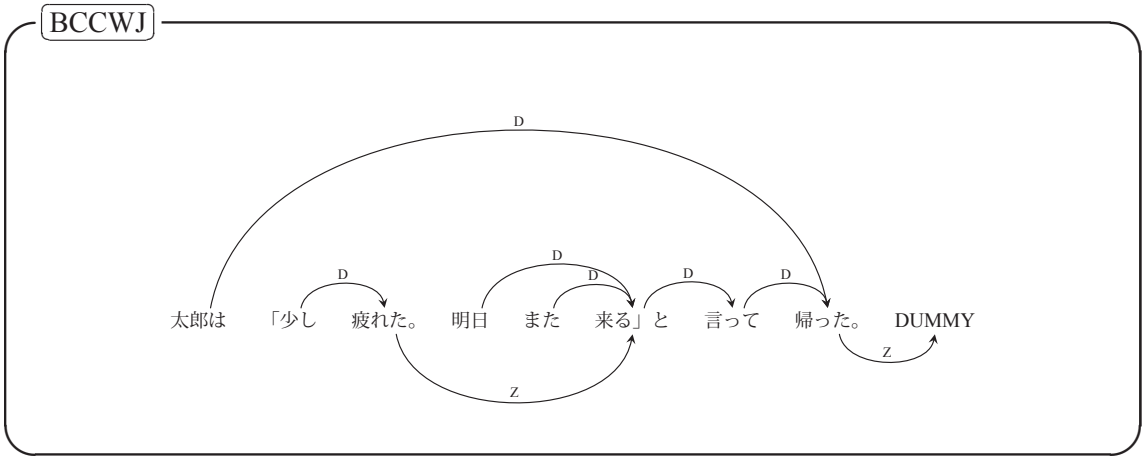
3.6 文節の連結

KCが文節係り受けを付与することを目的として文節単位を規定しているのに対し、BCCWJとCSJは形態論情報のみに基づいて文節単位を規定しており、係り受けを付与するためにそぐわない文節出現する。さらにCSJでは文節および節境界を元の音声ファイルのポーズによっても認定するために、文法的に不自然な単位が認定される場合がある。これに対応するために、文節境界を修正する記述を係り受け関係ラベルBCCWJにおいて “B” ラベル、CSJにおいて “B+” を用いて表現することがある。KCではこのような規定は存在しない。



3.7 文境界の修正

BCCWJ は文単位の定義として文の入れ子を許している。文書構造（レイアウト）に基づいて、一番外側の文について ⟨superSentence⟩ タグが付与されている。本来文の構造としては ⟨superSentence⟩ タグが付与されるべきものであって、文書構造中改行がある場合など ⟨superSentence⟩ タグが付与されていない場合、係り先のない文節が隣接文に出現する場合がありうる。このようなことのないように、BCCWJ 係り受けアノテーションにおいては、係り受けアノテーション向けに前処理で文書構造を考慮せずに、⟨superSentence⟩ 相当情報を追加で付与する。この際、文内に文境界相当の文節端が出現する場合がある。そのような場合には、BCCWJでは係り先なしとし、ラベル “Z” を付与する。一方、CSJは係り受けアノテーションを付与する単位として節を用いておりこのような問題は発生しない。また、KCではこのような規定は存在しない。



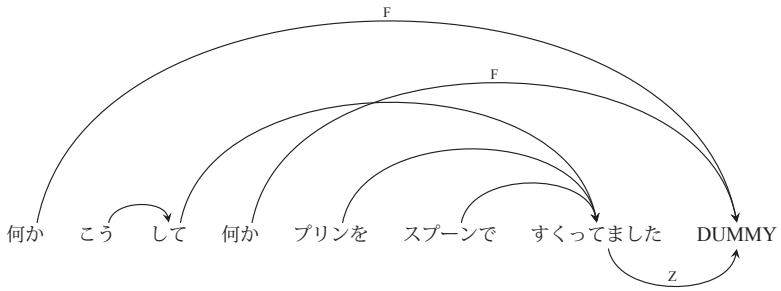
3.8 係り先なしの要素

KCでは係り先なしの文節要素を文末以外に認定していないのに対し、BCCWJとCSJでは係り先なしの文節要素を文末以外にも許している。特にCSJでは係り先なしの文節をラベルで細分化している。以下では、係り先なしの要素について比較する。

3.8.1 フィラー

CSJは、ラベル “F” を用い、フィラーの係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、同様に、ラベル “F” を用い、フィラーの係り先は定義しない。DUMMY に係けることによって係り先なしを示す。

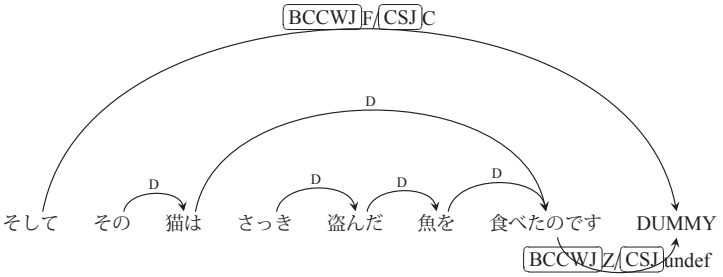
BCCWJ CSJ



3.8.2 接続詞

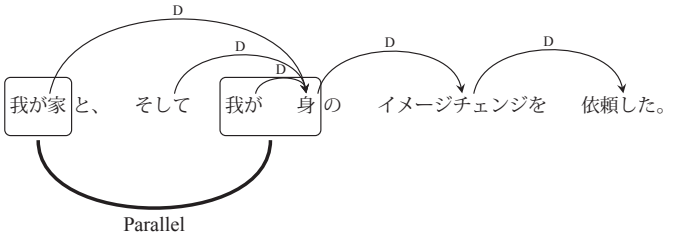
CSJは、ラベル “C” を用い、接続詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、文頭の接続詞で係り先判定が難しい際にラベル “F” を用い、接続詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。

BCCWJ/CSJ



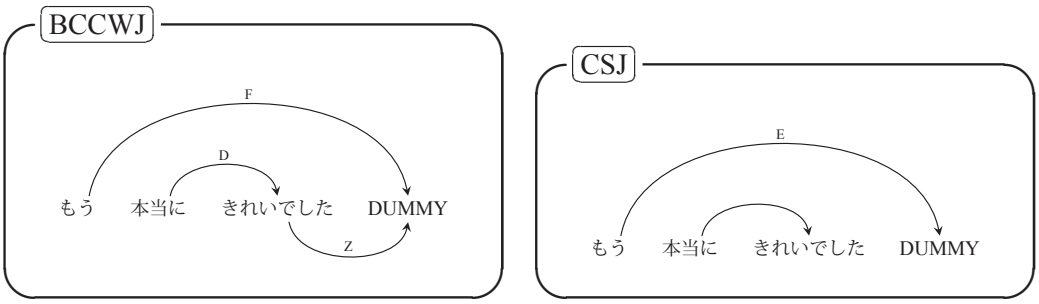
BCCWJにおいて、並列構造などを伴い、並列句の間に接続詞が出現する場合には、右隣接する並列句の最右文節に通常の係り受け関係 (ラベル “D”) として係ける。

BCCWJ



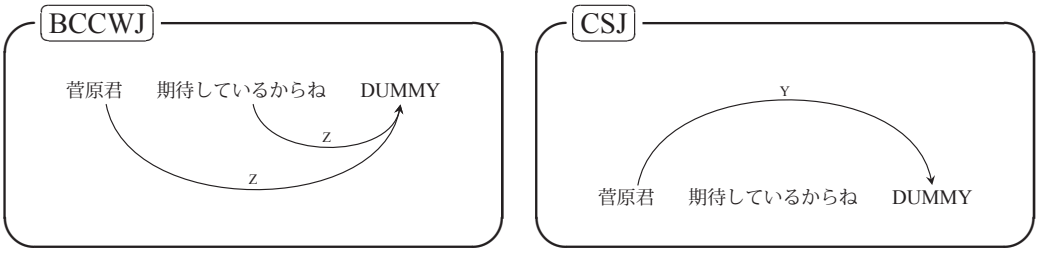
3.8.3 感動詞

CSJは、ラベル “E” を用い、感動詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、ラベル “F” を用い、感動詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。



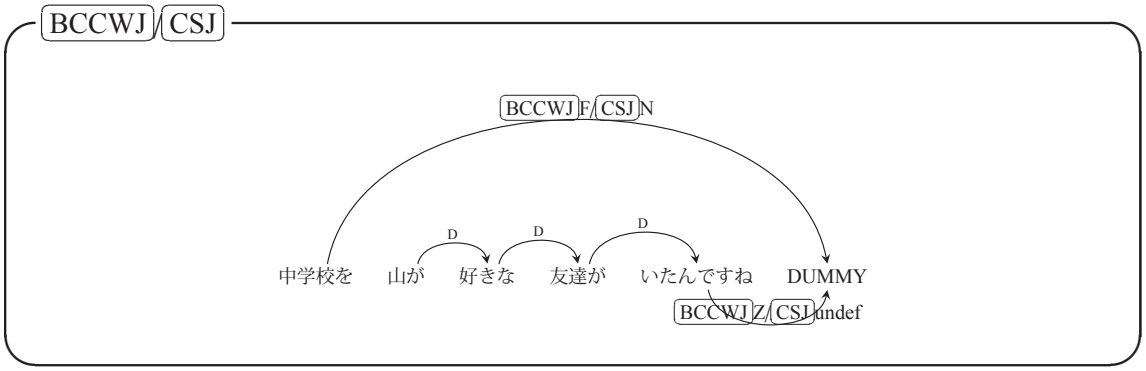
3.8.4 呼びかけ

CSJは、ラベル “Y” を用い、呼びかけの係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、ラベル “Z” を用い、呼びかけのあとに文境界相当の区切りを付与する。DUMMY に係けることによって係り先なしを示す。



3.8.5 係り先が消失している場合に付与するラベル

CSJは、ラベル “N” を用い、DUMMY に係けることによって係り先なしを示す。BCCWJは、ラベル “F” を用い、DUMMY に係けることによって係り先なしを示す。



3.9 格要素が複数の述語に係る場合

係り先を認定するのが難しい事例として、格要素が複数の述語に係る事例がある。並列する複数の述語の場合は等位接続とみなし係りうる遠いものに係ける。一方、複数の述語がそれぞれ従属節・主節に含まれている場合には、主題相当文節（「は」「も」）、主語相当文節（「が」）、それ以外の格要素（「に」「を」）など文節要素ごとに厳密に規定すべきである。BCCWJでは、このあたりの関係を南 (1974) の節分類などに基づき精緻化した。詳細については浅原 (2013)

を参照されたい。

3.10 その他

表 1 に各コーパスの係り受け関係ラベルの違いを示す。

表 1　係り受け関係ラベルの比較

係り受け関係のラベル	BCCWJ	(グループ | セグメント)	CSJ	KC
通常の係り受け	D	-	ラベルなし	D
並列	D	(Parallel)	P	P
部分並列	D	(Parallel)	I	I
同格	D	(Apposition)	A	A
同格（総称、数詞）	D	(Generic)	A2	A
言いよどみ	D	(Disfluency)	D	未定義
倒置	D	-	R	未定義
文節境界に関するラベル	BCCWJ	-	CSJ	KC
後続文節と接続	B	-	B+	未定義
その他	BCCWJ	(セグメント)	CSJ	KC
フィラー	F	-	F	未定義
顔文字	F	-	未定義	未定義
接続詞	F or D		C	D
感動詞	F or D	-	E	D
呼びかけ	Z	-	Y	未定義
非言語音	F	-	ラベルなし	未定義
係り先のない文節	F	-	N	未定義
記号・補助記号	F	-	未定義	未定義
URL・空白	F	-	未定義	未定義
係り受け関係の交差	D	-	X	未定義 (A のみ)
英単語・ローマ字文・漢文	D	(Foreign)	未定義	未定義
古文	D	(Foreign)	K(|S1|E1)	未定義
文境界相当	Z	-	未定義	未定義
コメント	BCCWJ	(セグメント)	CSJ	KC
	未定義	-	S:格表示誤り (「が | を | に」)	未定義
	F	(Disfluency)	S:複数文節の言い直し (|S1|E1)	未定義

以下、言及していない基準間の違いについて簡単に述べる。

- 言い直し・言いよどみ
  CSJでは言いよどみをラベル “D” で付与する。また複数文節の言い直しについては “S:複数文節の言い直し” ラベルに開始タグ (S1) と終了タグ (E1) を付与し範囲指定する。BCCWJでは言いよどみ相当句に Disfluency セグメントを規定し、言い直した表現に通常の係り受け関係で係ける。
- 顔文字・非言語音
  BCCWJでは、格要素などにならない顔文字表現については、副詞的用法であっても、句読法的な用法であっても区別せずに、ラベル “F” とし、DUMMY ノードに係ける。CSJでは、非言語音は通常の係り受けとして扱う。BCCWJでは、顔文字と同様に扱う。
- 記号・補助記号・URL・空白
  BCCWJでは、係り先が判定しにくい、リスト項目マーカー相当の記号・補助記号については、ラベル “F” とし、DUMMY ノードに係ける。URL・空白も同様に扱う。
- 英単語・ローマ字文・漢文・古文
  CSJでは、古文相当を係り受けラベル “K” で扱う。古文が複数文節にわたる場合にはラベル “K” に開始タグ (S1) と終了タグ (E1) を付与し範囲指定する。BCCWJでは、係り受け木とは独立にセグメント “Foreign” として英単語・ローマ字文・漢文・古文の範

囲を指定する。係り受け関係は通常の係り受けとしてみなす。

- 格表示誤り
  CSJでは、発話者の格表示誤りと想定される文節について、ラベル “S” に “格表示誤り(「が | を | に」)” をつけて付与する。

4. おわりに

本稿では、日本語の係り受けアノテーション基準間の差異について概観した。より詳細な比較については浅原 (2013) を参照されたい。

謝辞

本研究は国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

Collins, Michael J. (1999). “Head-driven statistical models for natural language.” Unpublished doctoral dissertation, University of Pennsylvania.

Johansson, Richard, and Pierre Nugues (2007). “Extended constituent-to-dependency conversion for english.” *Proc. of The 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*.

Magerman, David M. (1994). “Natural language parsing as statistical pattern recognition.” Unpublished doctoral dissertation, Stanford University.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). “Building a large annotated corpus of english: the penn treebank.” *Computational Linguistics*, 19:2, pp. 313–330.

Matsumoto, Yuji, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita (2005). “Chaki: An annotated corpora management and search system.” *Proc. of the Corpus Linguistics Conference Series (Corpus Linguistics 2005)*.

Nivre, Joakim, and Jens Nilsson (2005). “Pseudo-projective dependency parsing.” *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pp. 99–106. Ann Arbor, Michigan: Association for Computational Linguistics.

Yamada, Hiroyasu, and Yuji Matsumoto (2003). “Statistical dependency analysis with support vector machines.” *Proc. of 8th International Workshop of Parsing Technologies (IWPT-2003)*.

浅原正幸 (2013).「『現代日本語書き言葉コーパス』係り受け・並列構造アノテーション作業メモ (Version 0.6)」 Technical report, 国立国語研究所コーパス開発センター.

内元清貴・丸山岳彦・高梨克也・井佐原均 (2004).「『日本語話し言葉コーパス』における係り受け構造付与 (Version 1.0)」 Technical report, 『日本語話し言葉コーパス』の解説文書.

黒橋禎夫・居倉由衣子・坂口昌子 (2000).「形態素・構文タグ付きコーパス作成の作業基準 (Version 1.8)」 Technical report, 京都大学.

南不二男 (1974).『現代日本語の構造』 大修館書店.

「結果、こういうことが言えそうです。」

～コーパスにみる名詞の文副詞的用法～

東泉裕子（東京学芸大学）
高橋圭子（東洋大学）

“Result, We Can Say Something like That.”
Usage of Sentential-Adverb-like Nouns in Some Corpora

Yuko Higashiizumi (Tokyo Gakugei University)
Keiko Takahashi (Toyo University)

１．はじめに

現代日本語においては、名詞が副詞のように使われることがある。例えば、「結果」や「挙げ句」には次のような用例が観察される。

(1)　・・・（ゴミの分別の説明）・・・。結果、松戸市では７つに分類されました。
（1998 年 5 月 27 日 NHK ニュース）

(2)　通い続けている鍼治療院の院長から、身体の異変を指摘され、ガンが発見されたのだ。結果、早期治療に結びついた。
（BCCWJ：段勲『私はこうして「がん」を克服した』1997 年）

(3)　向井は中浜にこだわった。挙げ句、迷宮入りした。
（BCCWJ：東野圭吾『週刊プレイボーイ』2002 年）

これらの表現は、従来、「～た結果」「その結果（として）」、「～た挙げ句（の果てに）」「その挙げ句（に）」などとされていたものから、先行の「その」、後続の「として」「に」などが脱落し、単独で用いられるようになったものである[1]。しかし、脱落の前後で意味や機能は変わらない。こういった脱落後の表現は、(1)～(3)の例に見るように、文頭で用いられ、文副詞的に文全体にかかることが多い。そこで、本研究ではこれらの表現を「名詞の文副詞的用法」と呼ぶこととする。そして、歴史語用論の観点から、このような用法の広がりの様相と変化の道筋の持つ意味を考察するために、「結果」と「あげく」（「あげく」「挙（げ）句」「揚（げ）句」を「あげく」で代表させる）を対象として、現代語と近代語のコーパスを用いて調査する。

２．先行研究

２.１　歴史語用論

高田他（2011）によれば、語用論的現象を歴史的に研究する歴史語用論は 1990 年代後半

[1] 見坊 (1988)はこのような「結果」の用法を「副詞的用法」と呼んでいる。また、見坊 (1990)では「究極」「結果」「正直」「事実」の用例を挙げ、「同じ型に属する、定着した語形である」と指摘している。

に登場した新しい学問分野の名称であるが、同様の問題意識をもった研究はかなり以前から行われていたという。これまでの研究成果から、語の意味が、実質的意味を表すものから話し手の主観的な意味を表す方向へと変化していくことが、その逆方向の変化よりは多いという傾向が指摘されている。例えば、英語の名詞 *fact* は、*in fact* という形で「実際において」という意味を表すようになり、やがて「たしかに」「しかしながら」という話し手の真実性に対する判断を表す副詞句となり、次いで「前述したことよりこれから述べることのほうが大切である」ということを知らせる談話標識（discourse marker）として用いられるようになったという。

本研究では、「結果」「あげく」といった名詞の文副詞的用法もこれと同様のプロセスをたどっているのではないかという仮説から出発する。

２.２　コーパスによる量的調査

髙橋（2012）では、『現代日本語書き言葉均衡コーパス（BCCWJ）』と『日本語用例検索』を用い、「実際」「事実」「結果」「正直」「ある意味」といった名詞句の用法の変化を調査した。『BCCWJ』は、2001 年以降のデータが多い。他方、『日本語用例検索』は、『青空文庫』（http://www.aozora.gr.jp）に収録された約 3400 作品の中から用例検索を行えるサイトで、明治から昭和 20 年代頃までのデータが多い。

調査の結果、対象とした名詞句ではいずれも、名詞から文副詞的用法へのプロセスをたどっているが、その時期は語句によって異なることが明らかになった。『BCCWJ』と『日本語用例検索』の間で顕著な相違が現れたのは「正直」「ある意味」である。「ある意味」の各コーパスにおける用法は、表１のように分類できた。

表１　『BCCWJ』と『日本語用例検索』における「ある意味」の用法

	BCCWJ	日本語用例検索
文副詞的	142（ 29%）	0（ 0%）
中間的	340（ 68%）	117（ 84%）
名詞句	15（ 3%）	23（ 16%）
計	497（100%）	140（100%）

表１において、「中間的」とは、対象とする表現（この場合は「ある意味」）自体は名詞句であるが、「ある意味では」「ある意味において」等のように用いられ、文においては副詞句として機能しているもののことである。この中間的な表現の種類も、『BCCWJ』においては「ある意味で（は）」が 329 例と大半を占め、他の形式は「ある意味において（は）」「ある意味から言うと／言えば」「ある意味に」にしぼられていたのに対し、日本語用例検索では、『BCCWJ』で見られる形式に加え、「ある意味から」「ある意味から言えば／言うと／言って」「ある意味からすれば／して」「ある意味から見れば／見て」とバリエーションが豊富であった。つまり、文副詞化の過程をたどる中で、中間的用法も一定の形

式に収斂し、固定化していったと考えられる。

しかし、この調査においては、『日本語用例検索』の限界もまた明らかになった。その最大のものは、出典となる『青空文庫』の底本がまちまちである（雑誌、単行本初版やそれ以降、全集など）ため、厳密な比較には適さないことである。そこで、この限界を補うべく、本研究では『太陽コーパス』を用いることを考えた。『太陽コーパス』は、明治・大正期の雑誌『太陽』の 1985 年から 1925 年までの記事を収録したものである。３節では『BCCWJ』の用例と『太陽コーパス』の用例を比較することによって、文副詞化のプロセスを観察する。

３．調査

３．１　調査方法

本研究では、『BCCWJ』と『太陽コーパス』における「結果」と「あげく」の用例を比較し、文副詞的用法が現れるプロセスを観察する。『BCCWJ』の用例検索には『少納言』を、『太陽コーパス』の用例検索には同梱の全文検索システム『ひまわり』を用いた。

３．２　調査結果

上掲の表１を利用して、『BCCWJ』と『太陽コーパス』における「結果」と「あげく」の用法を分類すると、それぞれ表２と表３のようになる。予想通り、文副詞的用法はどちらの語も、『BCCWJ』のほうが『太陽コーパス』より多かった。

『太陽コーパス』においては、文副詞的用法と考えられる例はそれぞれ１例のみであった。「結果」の例(4)は、上の(2)と同様、文副詞的用法と認められる。

(4) 政本合同をやつて現内閣を倒すなどの荒藝は出來ない。 結果野垂れ死ぬまでズラ〜〜グツタリで行くだらうと云ふのだ。（『太陽コーパス』：鬼谷庵「政界鬼語」1925 年）

一方、「あげく」については、上の(3)のような文頭に現れる例は皆無だった。次の例は、「揚句」が文中に現れているが、文副詞的用法の萌芽だと考えられる。

(5) 一人ならず三人までも行方が知れない。 子供の事だから無論闇雲迷ひ歩つて揚句何處かでのたれ死をするか野獣に咬み殺されるかだ。

（『太陽コーパス』：中村星湖「みじか夜」1917 年）

表２　『BCCWJ』と『太陽コーパス』における「結果」の用法

分類	表現	BCCWJ		太陽コーパス	
文副詞的	結果	10	2.02%	1	0.03%
中間的（副詞寄り）	N の結果	22	4.44%	445	11.53%
	こ/その結果	67	13.54%	308	7.98%
	ル/タ結果	32	6.46%	447	11.58%
中間的（名詞寄り）	N の結果として	0	0.00%	147	3.81%
	その結果として	0	0.00%	102	2.64%
	ル/タ結果として	37	7.47%	74	1.92%
	結果として	0	0.00%	0	0.00%
	その結果において(は)	0	0.00%	5	0.13%
	結果において(は)	0	0.00%	1	0.03%
	結果的には	22	4.44%	1	0.03%
名詞	名詞	305	61.62%	2326	60.24%
動詞	動詞	0	0.00%	4	0.10%
合計		495	100.00%	3861	100.00%

表３　『BCCWJ』と『太陽コーパス』における「あげく」の用法

分類	表現	BCCWJ		太陽コーパス	
文副詞的	あげく	41	4.56%	1	1.85%
中間的（副詞寄り）	あげくのはて	12	1.33%	0	0.00%
	N の/その＋あげく	70	7.78%	5	9.26%
	タあげく	406	45.11%	22	40.74%
中間的（名詞寄り）	あげく＋助詞	71	7.89%	1	1.85%
	あげくのはて＋助詞	127	14.11%	4	7.41%
	N の/その＋あげく＋助詞	37	4.11%	12	22.22%
	タあげく＋助詞/copula	128	14.22%	7	12.96%
その他		8	0.89%	2	3.70%
計		900	100.00%	54	100.00%

さらに、表２と表３からは、表１の「ある意味」ほどは、中間的用法が減っておらず、『BCCWJ』と『太陽コーパス』における「結果」と「あげく」の用例はそれほど変化していないことが分かる。ただし、中間的用法のうち、『太陽コーパス』にみられた「名詞＋の結果として」「その結果として」「その結果においては」は、『BCCWJ』では例がなかった。『太陽コーパス』では多様であった中間的用法の表現が、『BCCWJ』では固定化・定型化していったことが分かる。

「あげく」については、単独での文副詞的用法の増加（0.03%→2%）に加え、「あげくのはて」という定型表現の増加にも注目したい。「あげくのはて＋助詞」が 7%から 14%に倍増したのに加え、「あげくのはて」という名詞句単独の文副詞的用法も『太陽コーパス』では皆無だったが、『BCCWJ』では 1.33%出現している。

４．考察

前節の量的調査の結果から、「結果」「あげく」の用法について、次のような変化のプロセスを仮説として提示することができる。

(6) 名詞 → 中間的用法（名詞寄り） → 中間的用法（副詞寄り） → 文副詞的

先行・後続部分あり　　先行・後続部分脱落

表現の種類は多様　　定型的表現に収斂

これは、髙橋（2012）の「実際」「事実」「正直」「ある意味」の調査結果とも軌を一にする。

そして、個々の用例を観察すると、いずれの名詞句においても、(6)のプロセスを進むに従い、実質的意味の希薄化が起こっているようである。

(7) 結局居住者の承諾を取らず無断で立ち入った案件がありました。結果、居住者は 300 万円相当の腕時計と指輪がなくなったと主張し警察を呼びました。
（BCCWJ：「Yahoo!知恵袋」2005 年）

(8) hanamarin さん的には絶対にもぉない！と、踏んでいて…結果、お年玉クジは終了していて、
（BCCWJ：「Yahoo! ブログ」2008 年）

これらの例における「結果」は、因果関係の結果を表すわけではなく、「そして」のように単に時間の前後関係をつないでいるだけのようである。歴史語用論の観点からみると、実質語がその意味を失い、機能語に近い役割を果たすようになる変化であり、よく観察されるケースである。

また、文副詞的用法を獲得した名詞句の中には、語用論化し、談話標識化したと考えられるものもある。

(9) 真似好き（上手）だから、日本はここまで発展したと思えばしかたないのかな？ある意味、日本らしいのかも〜。
（BCCWJ：「Yahoo!知恵袋」2005 年）

(10) その美容師は「ぱっと見ておかしくなければいいんじゃないですか」と断言。正直、ほとほと閉口しました。（BCCWJ：山田みどり『はじめての接客サービス』2005 年）

(9)の場合、どのような意味で「日本らしい」のか、別の意味なら日本らしさとは別の要因

がみえるのか、といった議論は起こらない。(10)の場合も、正直であるかないかと言った議論は無縁である。どちらの場合も、後続部分の前置きとして、これから述べる表現が適切かどうかわからないが、「ある意味では～のように言える」、「正直な気持ちを述べれば～ということになる」というクッション的・やわらげ的機能を果たしていると考えられる。

５．まとめと課題

小規模ではあるが、今回の調査で、名詞の文副詞的用法が歴史語用論の知見に沿うものであることを提示できた。

今回は量的調査にとどまってしまったが、今後、１つ１つの用法のパターンを、文脈を考慮しつつ、質的に研究する必要がある。また、本研究で例示した名詞句の他にも、「究極（において・のところ）」（見坊 1990）、「基本（的に）」、「（その）瞬間」、「（それに）対して」などの文副詞的用法も増えているようである。こういった変化はどのような語句に生じやすく、それはどのような要因によるのかを綿密に調査・検討していきたい。

謝　辞

本研究の一部は、ひと・ことば勉強会において発表したものです。佐竹秀雄先生（武庫川女子大学）、三宅和子先生（東洋大学）はじめ、ご助言をくださった方々に感謝申し上げます。

文　献

見坊豪紀(1988)「結果（副詞的用法）」『現代日本語用例全集』、pp.41-42、筑摩書房
見坊豪紀(1990)「究極する」『日本語の用例採取法』、pp.86-88、南雲堂
高田博行・椎名美智・小野寺典子編著（2011）『歴史語用論入門』大修館書店
髙橋圭子（2012）「コーパスにみる名詞句の文副詞的用法」第 10 回対照言語行動学研究会
（http://www.ryu.titech.ac.jp/~nohara/taishogengokoudou/files/abst10/abst10_5takahashi.pdf）

コーパス

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

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/
日本語用例検索　http://www.let.osaka-u.ac.jp/~tanomura/kwic/aozora/

語義曖昧性解消の領域適応のための訓練データの選択法 ～複数ドメインからの選択～

堀内浩史郎（東京農工大学　工学部 情報工学科）
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）

Selection of Training Data for Domain Adaptation of Word Sense Disambiguation - Selection from Multiple Domains -

Koshiro Horiuchi (Department of Computer and Information Sciences,
Faculty of Engineering, Tokyo University of Agriculture and 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

ターゲットデータと異なるドメインのデータ（ソースデータ）で分類器を学習し，ターゲットデータに適応することを領域適応という．しかし，語義曖昧性解消ついて領域適応を行う際，分類したいデータごとに適切なソースデータは異なる．近年では複数のドメインの語義タグつきコーパスが入手できるため，分類したいデータに対して適切なソースデータを選択することが望ましい．

本稿では，ソースデータとして利用できるコーパスを複数を持っている場合を想定し，未知のターゲットデータが現れた際に，「用例全体の素性の平均ベクトルの類似度」「用例全体の出現素性の類似度」「用例全体の素性の分布」「分類器の示す確率」を使って訓練データの選択を試みて，訓練データの選択方法として利用できるものがあるか模索する．

２．関連研究

領域適応についての関連研究として，(Vincent Van Asch，Walter Daelemans(2010))の異なるドメインである訓練データとテストデータのコーパス同士の類似度から，品詞タグづけタスクにおける領域適応時の分類器の正解率を予測する研究がある．この研究では，コーパス同士の類似度と正解率の間に線形相関があることが示されている．(Daumé III(2007))は，素性空間を三倍にすることで，さまざまな supervised の領域適応に併用でき，さらに簡単に実装できマルチドメインに拡張も簡単であることを示した．

（古宮，奥村(2012)）は，語義曖昧性解消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データとターゲットデータの性質により異なることを示した．訓練データを選択する研究としては，（Komiya and Okumura(2012)）の訓練データの選択に分類器の確信度を用いる研究がある．全ての訓練データについて分類器をつくり，分類したときの各分類器の示す確信度によって訓練データを選択する手法である．さらに，確信度に加えて LOO-bound という指標を用いる研究も（古宮，小谷，奥村(2013)）によって行われている．本研究でも，確信度と LOO-bound のような指標を用いた実験を行うが，（古宮，小谷，奥村(2013)）では分類器にサポートベクトルマシン（Support Vector Machine，以下 SVM）を用いているのに対して，本研究では最大エントロピー（Maximum Entropy，以下 ME）法（Suarez and Palomar(2002)）の分類器を用いている点が異なっている．

また，（古宮，小谷(2011)）では領域適応が行われる状況によって最も良い手法が異なるとし，与えられたデータの性質を用いて三手法からひとつの手法を選択している．

３．訓練データ選択方法

いくつかのドメイン（ジャンル）のラベルつきデータを全て訓練データとして利用できる際に，ドメインのわからない未知のラベルなしデータを分類したい場合を考える．本稿では，未知のテストデータ（分類対象のターゲットデータ）に合わせた訓練データをいくつかの手法によって選択し，各手法が正解率の向上につながるか調べる．実験は次のようなステップで行う．

I. 各手法による訓練データの選択
II. Iで選択した訓練データでの領域適応（分類実験）
III. 他の手法で選択した訓練データを用いた場合との語義曖昧性解消の正解率の比較

ステップ I において，次に示す手法を試みる．

i. 類似度を用いた訓練データの選択
   - 用例全体の素性平均ベクトルの類似度を利用する手法
   - 用例全体の出現素性ベクトルの類似度を利用する手法
ii. 素性分布の距離を用いた訓練データの選択
iii. 分類器の示す確率を用いた訓練データの選択
   - 分類器の分類確率を利用する手法
   - 分類器の自信度を利用する手法
   - 分類器の分類確率と自信度を利用する手法

なお，語義曖昧性解消の対象単語タイプごとに分類器を作成するため，訓練データの選択は単語のタイプごとに行った．また，iii に関しては，（Komiya and Okumura(2012)）にならい，テストデータの用例ごとに選択を行う実験も行った．

３．１　類似度を用いた訓練データの選択

テストデータと訓練データを表すベクトルを各ひとつずつ，それぞれの素性ベクトル集合を用いて作成し，そのベクトル同士の類似度を訓練データの選択指標として用いる．ベクトルは，各要素を足して用例数で割った「素性平均ベクトル」と，全ての要素の OR を取った「出現素性ベクトル」の二つについて調べる．利用する類似度は，ユークリッド距離（ED），コサイン類似度（CS），ジャッカード係数（JSD），ダイス係数（DSC），シンプソン係数（SSC），ランド類似度（RS）を用いる．なお，全ての類似度についてテストデータとの類似度が最高値であったデータを訓練データにした場合（以後，（最大）と表記）と，最小であったデータを訓練データにした場合（以後，（最小）と表記）の 2 通りを調べる．

３．２　素性分布の距離を用いた訓練データの選択

テストデータと訓練データの素性ベクトル集合において各素性の分布（本稿では 17 個の分布）を作成し，各素性分布同士の距離を測り，その距離の総和が最も小さくなる訓練データを選択する．各素性分布の距離の測定にはジェンセン・シャノン・ダイバージェンスを用いる．

３．３　分類器の示す確率を用いた訓練データの選択

ME 法を用いて分類を行うと，各ラベルに分類される確率が算出される．この確率を「分類確率」と呼び，分類確率の中の最大値を最大分類確率と呼ぶこととする．また，訓練データを 5 分割交差検定した結果を「自信度」と呼ぶこととする．「自信度」は，その分類器

がその訓練データと同じドメインのコーパスをどの程度正確に分類できるかを表す.

訓練データの選択には，最大分類確率の平均値が最大となる訓練データを選択する手法と，自信度を用いて訓練データを選択する手法，そしてこれら二つの値の積を用いて訓練データを選択する手法を試みる.

上記の訓練データ選択実験に加えて，各用例ごとに分類確率を用いて訓練データの選択する手法を試みた．訓練データを変えて分類器を学習し，学習された分類器の中で最高の分類確率を示した分類器の結果を用例ごとに選択する.

４．訓練データ選択実験

４．１　最大エントロピー法

本実験の分類手段として ME 法を用いる．ME モデルの実現には（Le Zhang(2011)）の Maximum Entropy Modeling Toolkit for Python and C++を用いた.

４．２　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（Hashida et al. (1998)）の新聞記事を用いた.

単語の語義は岩波国語辞典（西尾ら (1994)）の小分類の語義を採用した．語義数ごとの単語の内訳は，2 語義 :「場合」,「自分」，3 語義 :「事業」,「情報」,「地方」,「社会」,「思う」,「子供」，4 語義 :「考える」，5 語義 :「含む」,「技術」，6 語義 :「関係」,「時間」,「一般」,「現在」,「作る」，7 語義 :「今」，8 語義 :「前」，10 語義 :「持つ」， 12 語義 :「見る」，14 語義 :「入る」，16 語義 :「言う」，22 語義 :「手」である.

表 1　ドメインごとの単語の最小，最大，平均用例数

コーパスの種類	最小	最多	平均
BCCWJ 白書	58	7610	2240.14
BCCWJ Yahoo！知恵袋	130	13976	2741.95
RWC 新聞	56	374	183.36

実験は，三つのドメインのうちひとつのドメインをテストデータとして利用し，他の二つのドメインのデータから訓練データを選択する．たとえば Yahoo！知恵袋をテストデータとした場合は，訓練データの選択肢は「新聞記事」「白書」「新聞記事＋白書」の三通りである.

各ドメインの各単語ごとに選択実験を行うので，それぞれの手法に対して計 66 回の実験を行う.

５．結果

本実験のベースラインは利用できる訓練データ二つのドメインの両方を利用した場合である．表２に各手法の実験結果を示す．なお，全ての手法の中で最も良い正解率を下線に示す．また，テストデータのドメインごとに最も良かった結果に下線を引いた．マクロ平均・マイクロ平均ともに，新聞記事が素性分布の距離，白書が出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小），Yahoo！知恵袋が出現素性ベクトルのコサイン類似度（最大）で訓練データを選択したときに語義曖昧性解消の正解率が最も高くなった．表３にこれらのドメインごとの結果を示す.

なお，出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小）は各ドメインの各単語について全て同じ訓練データを選択したために，同じ結果となっている.

表２　訓練データ選択実験結果

手法		マクロ平均（%）	マイクロ平均（%）
ベースライン		75.04	81.1
素性平均ベクトル	ED（最大）	73.59	74.64
	ED（最小）	71.05	79.09
	CS（最大）	71.82	73.97
	CS（最小）	71.16	78.58
	JSC（最大）	74.21	81.02
	JSC（最小）	69.86	72.92
	DSC（最大）	74.21	81.02
	DSC（最小）	69.86	72.92
	SSC（最大）	73.18	74.79
	SSC（最小）	71.68	78.2
	RS（最大）	67.94	72.68
	RS（最小）	74.37	79.75
出現素性ベクトル	ED（最大）	75.34	81.39
	ED（最小）	68.83	73.74
	CS（最大）	74.38	81.55
	CS（最小）	68.99	73.28
	JSC（最大）	73.33	80.89
	JSC（最小）	71.8	74.31
	DSC（最大）	73.37	80.91
	DSC（最小）	71.8	74.31
	SSC（最大）	71.86	74.2
	SSC（最小）	71.64	79.9
	RS（最大）	68.79	73.55
	RS（最小）	75.34	81.39
素性分布の距離		75.02	81.02
分類確率		70.58	77.53
自信度		72.7	80.54
分類確率と自信度		72.7	80.54
分類確率でラベル予測		74.08	80.41

６．考察

表３が示すように，ドメインごとに適当な訓練データの選択手法は異なる．さらに全体のマクロ平均が最も良かった出現素性ベクトルのユークリッド距離（最大）について，語義曖昧性解消の対象単語のタイプごとに結果を詳しく調べると，訓練データよりもテスト

表３　ドメインごとの実験結果

	マクロ平均（%）			マイクロ平均（%）		
手法	新聞記事	白書	Yahoo！知恵袋	新聞記事	白書	Yahoo！知恵袋
出現素性ベクトルの ED（最大）	74.37	76.74	74.9	75.56	80.32	82.66
出現素性ベクトルの RS（最小）	74.37	76.74	74.9	75.56	80.32	82.66
出現素性ベクトルの CS（最大）	71.26	74.28	77.6	73.6	79.66	83.63
素性分布の距離	74.96	74.16	75.95	75.83	79	83.03
ベースライン	74.37	76.57	74.17	75.56	79.86	82.48

データの方が用例数が多い場合に正解率が上がっているものが多いことが分かった．このことから，適当な訓練データの選択手法は語義曖昧性解消の対象単語のタイプごとにも異なることが分かる．

訓練データよりもテストデータが少ない場合について調べると，素性平均ベクトルのユークリッド距離（最大）を用いた場合が最も良い結果となった．ここで，訓練データよりもテストデータが多い場合は出現素性ベクトルのユークリッド距離（最大）を，少ない場合は素性平均ベクトルのユークリッド距離（最大）を用いて訓練データを選択した結果を表４に示す．

表４　二手法を組み合わせたときの正解率

	マクロ平均			マイクロ平均		
手法	新聞記事	白書	Yahoo！知恵袋	新聞記事	白書	Yahoo！知恵袋
二手法組み合わせ	74.37	76.65	77.53	75.56	80.25	83.87
ベースライン	74.37	76.57	74.17	75.56	79.86	82.48

表４より，白書と Yahoo！知恵袋でベースラインよりも語義曖昧性解消の正解率が良くなり，新聞記事でもベースラインと同じ正解率となった．今回の類似度の組み合わせは本研究で利用したデータの特徴から手法を選択しているため，よりデータの性質から適した手法の組み合わせを考える必要があるだろう．

分類確率を用いた実験については，SVM で有効に働いていたが，本実験の ME 法では語義曖昧性解消の正解率を上げることができなかった．

７．まとめ

語義曖昧性解消における領域適応の正解率を向上させるために，「素性平均ベクトルの類似度」「出現素性の類似度」「素性分布の距離」「分類器の分類確率と自信度」を用いて訓練データの選択を行い，どの選択手法が最も優れているかを調べた．全体の平均を見ると，マクロ平均で出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小），マイクロ平均で出現素性ベクトルのコサイン類似度（最大）で選んだ際に，語義曖昧性解消の正解率が最も高くなった．

各ドメインの結果を見ると，マクロ平均・マイクロ平均ともに，新聞記事が素性分布の距離, 白書が出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小）, Yahoo！知恵袋が出現素性ベクトルのコサイン類似度（最大）で訓練データを選択したときに語義曖昧性解消の正解率が最も高くなった．

それぞれの手法は訓練データとテストデータの性質によって異なる様相を見せたため，データサイズによって二手法を組み合わせた実験を行った．二手法を組み合わせた結果，語義曖昧性解消の正解率が全てのドメインでベースライン以上となった．

分類確率を用いた手法については，本研究では ME 法を用いて実験したが，異なる分類器である SVM を用いた関連研究のように語義曖昧性解消の正解率を上げることができなかった．

謝　辞

本研究は，文部科学省科学研究費補助金[若手 B（No : 24700138）]の助成により行われた．ここに，謹んで御礼申し上げる．

文　献

Vincent Van Asch，Walter Daelemans(2010)「Using Domain Similarity for Performance Estimation」，DANLP 2010，pp.31-36.

H. Daumé III(2007)「Frustratingly Easy Domain Adaptation」，ACL 2007，pp.256-263.

H. Daumé III，Abhishek Kumar，Avishek Saha(2010)「Frustratingly Easy Semi-Supervised Domain Adaptation」，ACL 2010，pp.53-59.

Le Zhang(2011)「Maximum Entropy Modeling Toolkit for Python and C++」，http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Kanako Komiya and Manabu Okumura(2011)「Automatic determination of a domain adaptation method for word sense disambiguation using decision tree learning」, IJCNLP 2011, pp.1107-1115.

古宮嘉那子，奥村学(2011)「分類器の確信度を用いた合議制による語義曖昧性解消の領域適応」，言語処理学会第 17 回年次大会発表論文集，pp.552-555.

古宮嘉那子，奥村学(2012)「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―」，言語処理学会第 18 回年次大会発表論文集，pp.1288-1291.

古宮嘉那子，小谷善行(2011)「階層型クラスタリングを利用した文脈によるオノマトペの分類」，NLP 若手の会第 6 回シンポジウム.

Michel Marie Deza, Elena Deza(2012)「Encyclopedia of Distances」，Springer-Verlag.

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.

Koichi Hashida，Hitoshi Isahara，Takenobu Tokunaga，Minako Hashimoto，Shiho Ogino，and Wakako Kashino(1998)「The rwc text databases」，LREC 1998，pp.457-461.

Kikuo Maekawa(2008)「Balanced corpus of contemporary written japanese」, ALR 2008, pp.101-102.

Armándo Suarez，Manuel Palomar(2002)「A Maximum Entropy-based Word Sense Disambiguation system」，COLING 2002，Vol.1，pp.1-7.

談話構成機能からみた外来語の基本語化
—通時的新聞コーパスを資料に—

金　愛蘭（早稲田大学日本語教育研究センター／国立国語研究所）†

Inclusion of Loanwords into the Basic Words in the Japanese Newspaper Vocabulary : From the Viewpoint of Discourse Organizing Function

Eran KIM (Center for Japanese Language, WASEDA University／NINJAL)

１．はじめに

20 世紀後半の新聞文章では、具体名詞のほかに、抽象的な意味を持つ外来語も増加し、基本語化している。言語外的な理由によって説明できる具体名詞の場合と違い、抽象的な外来語の多くは、和語・漢語の同義語・類義語があるにもかかわらず基本語化しており、その増加の理由は言語内的に説明しなければならない。

例えば、基本語化した外来語の中には、その意味が拡大して多義語となり、その結果として、類義語の上位語の位置に立つようになってきたものがある。「トラブル」は、その代表的な例である（金愛蘭 2006a）。また、その用法が拡大して、類義語と文法的な面において分担する傾向を見せるものもある。「ケース」は、その代表的な例である（金愛蘭 2006b）。

しかし、「抽象的な事柄をあらわす外来語の基本語化」問題を追究するためには、このような語彙的な側面についての検討や文法論的な機能の検討とは別に、抽象的な外来語が実際の文章・談話の中でどのような機能を担うようになっているか、という文章論的な検討も必要になる。文章論的な機能には、Halliday and Hasan（1976）の「語彙的結束性 lexical cohesion」、McCarthy（1992）の「談話構成語 discourse-organizing words」、高崎みどり（1988）の「指示語句」などにかかわる機能が考えられる。たとえば、結束性の語彙的な表示である「再叙（語彙的指示の同一性）」には、同一語の繰り返し、同義語や近似同義語、上位語、一般名詞（general noun）、人称指示語などがあるとされるが、「トラブル」は、このうちの上位語および一般名詞としての特徴を持っていると考えられる。実際、後述する「通時的新聞コーパス」からは、次のような例が得られる（氏名のイニシャル化および下線は筆者）。

(1) 大阪地裁で 23 日あった殺人事件の論告求刑公判（K 裁判長）で、殺された娘の遺影を手に傍聴していた母親（53）が、持ち込んだコードで被告の男性（20）の首を絞めたり、遺影の額のガラスを割って破片を法廷に投げつけたりする騒ぎがあった。関係者にけがはなかった。刑事裁判での遺影の持ち込みは、先月から各地で相次いで許可されているが、こうしたトラブルは初めて。大阪地裁は「遺族としての気持ちの高ぶりもある。法的に事件にするかどうかは分からない」と話している。[00 年 10 月 24 日朝刊社会]

ここで、「トラブル」は、「こうした」とともに指示語句を形成しつつ、先行する「騒ぎ」の上位語としてそれを指示するという談話構成機能を発揮している。発表者は、このよう

† kim_eran@aoni.waseda.jp

な「トラブル」の談話構成機能がいつごろから、また、どのように獲得されてきたのかを明らかにすることも、「トラブル」の基本語化の要因を考える上で重要なポイントになるものと考えている。

そこで、本発表では、20 世紀後半の新聞における抽象的な外来語の基本語化現象を、それら基本外来語の新聞文章における文章論的な機能の獲得という側面から考察する。具体的には、自作の通時的新聞コーパス（各年 36 日分増補版）[注1]を用いて、どのような外来語が「指示語＋外来語」からなら指示語句（の名詞＝後要素）を形成し「談話構成語 discourse-organizing words」としての機能を担っているのか、また 20 世紀後半の新聞文章においてその機能をどのようにして獲得してゆくのかを動的にとらえることを試みる。

２．資料―「通時的新聞コーパス」について

調査には、自ら作成した「通時的新聞コーパス」を用いる。同コーパスは、1950 年から 2000 年までの『毎日新聞』から、ほぼ 10 年おきに、毎月 3 日分（5 日・15 日・25 日）、各年 36 日分（全体では 216 日分）の朝刊全紙面の記事を、1950 年・60 年・70 年・80 年は『縮刷版』からテキスト入力し、1991 年と 2000 年については『ＣＤ－毎日新聞データ集』から抽出して作成した。基本的には、広告を除く記事を対象とするが、ラテ欄、都内版、地方版をはじめ、俳句・川柳、証券・株、人事、決算、訃告、競馬などは除外した。抽出比率は、約 10 分の 1 である。作成されたコーパスの規模は、表１（空白は除く）の通り。データの規模は、全体で 1,600 万字を超え、ページ数の極端に少なかった 1950 年、やや少なかった 1960 年を除けば、各年ほぼ 300 万字程度となり、20 世紀後半の通時的な新聞コーパスとしては、他に例を見ない大規模なコーパスを構築することができた。詳しくは、金愛蘭（2011）を参照されたい。

＜表１＞ 各年の文字数

年	文字数
1950	793,692
1960	2,208,396
1970	3,183,297
1980	3,218,737
1991	3,265,786
2000	3,994,933
計	16,664,841

３．先行研究

Halliday and Hasan（1976）は、「結束性は、テクスト内のある要素と、その要素の解釈に欠くことのできない他の要素との間の意味的な関係である」（邦訳 pp.9）とし、テクスト内の「語彙的結束性」は「再叙（語彙的指示の同一性）」と「コロケーション（語彙環境の類似）」によって表示されるとしている。さらに「再叙 reiteration」には、同一語の繰り返し（repetiton）、同義語（synonym）や近似同義語（near-synonym）、上位語（superordinate）、一般名詞（people, stuff, move などのような一般的指示を持つ名詞類、general noun）、人称指示語 it などがある、としている。

一方、McCarthy（1992）は、語のタイプを文法語（grammar words）と語彙語（lexical words）とに区別した上で、テクスト分析の方法として、その中間にあるような機能を持つ

---

1 「通時的新聞コーパス」の作成にあたっては、（財）博報児童教育振興会「第 3 回ことばと教育研究助成」と、文部科学省科学研究費補助金「20 世紀後半の新聞における外来語の基本語化に関する調査研究」（平成 22～23 年度・若手研究Ｂ・課題番号 21720168）および「基本外来語の談話構成機能に関するコーパス言語学的研究」（平成 24～26 年度・若手研究Ｂ・課題番号 23720241）の交付を受けた。本発表では、金（2011）の毎月二日分から、三日分に増補改訂したものを用いる。

語を「談話構成語 discourse-organizing words」とした。例えば、“issue” ”problem” “dilemma”のような語で、「これらの語は、テクストの分節の代わりをしているのである（ちょうど代名詞のように）。分節は、１つの文である場合もあるし、数個の文、パラグラフ全体、あるいは、それよりも広い範囲である場合もある」とする（邦訳 pp.105～pp.107）。つまり、「談話構成語」は、テクストの内容や分野を伝えるための語というより、テクストの構成・構造を示す語で、話題を発展させ広げていくことで、より大きなテクストパターン（例えば、「このジレンマを打開するための解決策としては…」のような場合、「問題－解決」になる）を生成し、談話の全体像を予測させる働きを持つ語である。

以上のような「談話の構成に役立つ語」に関する問題は、文章論研究におけるテーマの一つである。しかし、和語や漢語について触れているものは散見するが、外来語の談話構成機能にふれているものはほとんどない。また、従来の外来語研究においても、このような文章論的なアプローチを採ったものは管見の限りない。今後、語のリスト化と基本語化過程との関連を明らかにしていかなればならない。

４．調査と結果

テスクト内の語彙的結束性は、同語反復のほかにも、同義語・類義語、または上位語（さらに広い語彙概念を表わす総称的な名詞も含む）としてあらわれる。また、上でも述べたように、McCarthy（1992）の「談話構成語」や高崎（1988）の「指示語句」もある。

そこで、本発表では、指示語の中でも「この」を伴った「この＋外来語」の形式に限定し、前で述べた部分をどのように後ろへつなぎ、テクスト性（textuality）を生み出しているかについて、コーパスに基づく調査を行なった。

４．１　調査

データは、上記「通時的新聞コーパス」から、全文検索システム「ひまわり」を使って、「この」にカタカナ文字列が後接するレコードを抽出し、さらに、上記の構造を有するものを目視によって抜き出した。ただし、「この」に後接する後要素が句になっているもの(2)や合成語のもの(3)、および人名・地名など固有名のものは、対象外とした。

(2) ドイツの福祉はかなり進んでいて、盲導犬の訓練施設や訓練師の養成も充実している。このドイツの事情が日本にも役立つのではないか、…［91 年 5 月 5 日家庭］

(3) …株価は業績不振から６月下旬以降急落、今月に入って倒産説も流れ、株価は１０ドルを切った。ゼロックスはこのリストラ策実施で、来年第１四半期からは黒字化すると説明している。［00 年 10 月 25 日経済］

４．２．調査結果

通時的新聞コーパスにおける指示語句「この＋外来語」の用例数の変化は、表 2 の通りである。

＜表２＞ 通時的新聞コーパスにおける用例数の変化

	1950 年	1960 年	1970 年	1980 年	1991 年	2000 年
この＋外来語	19	103	196	175	71	93
	23.9	46.6	61.6	54.4	21.7	23.3

（4.1 で述べた対象外を除いた数、上段は延べ語数の実数、下段は 100 万字当たりの換算値・小数点第二位四捨五入）

５．考察

通時的新聞コーパスにおける指示語句「この＋外来語」を分類した結果、まず、大きく 3 種に分けることができた。以下、具体例をあげながら説明する。

５．１　同一語の繰り返しによる結束性

同じ語を繰り返し使うことにより、文章の結束性が保たれる場合である。これには、同じ語を「そのまま繰り返し」使うタイプと、前に出現した語の類概念となる部分を取り出して用いる「類概念抽出」タイプの、二通りがある。

【そのまま繰り返し】

(4) 一、また大型人工衛星には予備の燃料を持ったエンジンを装備することもできる。このエンジンによって人工衛星が大気の濃い層の中に入る時…［60 年 2 月 5 日外電］

(5) …ような教育を信条とする人たちが大正十二年に「教育の世紀」社というグループをつくって、同名の機関誌をだした。このグループの事業として自由な教育をする私立小学校がたてられ、「児童の村」という名がつけられた。［80 年 3 月 15 日家庭］

(6) …を主力商品として売出し、一五、一三インチも大量生産に踏切るメーカーもある。このメーカーがねらうのは、年収八十万円から百万円の階層。［70 年 7 月 5 日家庭］

【類概念抽出】

(7) アルバイトでためた金で安い中古車を買い、初心者マークをつけて規定通り走って、また驚いた。世の中の名ドライバーはこのマークを軽べつするようである。［80 年 1 月 15 日投書］

(8) ここを拠点に１１月から来年３月まで、ヨーロッパの科学者らが結集して北極圏オゾン層を集中観測する「オゾン・キャンペーン」が実施される。近藤さんは日本からただ一人、世界でたった一つの装置とともに、このキャンペーンに参加する。［91 年 12 月 15 日科学］

(9) 東京・銀座の三越銀座店で、偽造クレジットカードで洋酒八万円相当を買った男が築地署に詐欺の疑いで逮捕された。十四日までの調べで、男は偽造クレジットカードを使った別件の詐欺と改造ピストル所持で起訴されたが逃走、…（中略）…。調べによると、このカードは五月初め、大阪市内の女性が紛失したもの。［80 年 6 月 15 日社会］

５．２　言いかえによる結束性

まとまりのあるテクストを作り出すために、同じ語を繰り返し使う必要はなく、前で述べた言葉の類義語や上位語を用いて言いかえることで、後続の文章へ展開していくことがある。これにも、「同義語または類義語」を使うタイプと、「上位語または一般名詞」を使う、二通りが存在する。

【同義語または類義語】

(10) …が強まっていることが板硝子協会（東京都千代田区）による室内鏡の普及状況調査で分かった。＼ＨＯＮ＼　このアンケートは全国の電話帳から無作為抽出した一万四千五百世帯を対象に［91 年 9 月 25 日家庭］

(11) うっとうしい梅雨期、ひきつづいてやってくる暑い夏―どちらも食欲の減退しがちで心身ともにゲンナリする季節だ。食欲の衰えは体力を低下させ病気を起こしやすくする。しのぎにくいこのシーズンを健康に乗切る方法はないものか。［70 年 6 月 15 日健康］

【上位語または一般名詞】

(12) わたしは遺伝的に喘息があって、若いときはその徴候はなかったが、戦争から帰って、喘息に苦しめられるようになった。＼T２＼　いろんな療法をやったが、このアレルギーに対抗するものはなく、諦めるより仕方がないと観念していた。[00 年 10 月 15 日総合]

(13) 「人間は生きるために食べる。食べた栄養素を体内に蓄え、体に必要なものを作り出しエネルギーとして活用するのに、すい臓から出るインスリンが働くのです」＼HON＼このホルモンが不足すると、血液中のブドウ糖を…[91 年 10 月 5 日家庭]

５．３　とらえ直しによる結束性

「とらえ直し」とは、前で述べた事柄を、指示語を伴いつつ、書き手が何らかの形でとらえ直して、文章を展開していく場合である。その際、書き手は、前で述べた（文に限らず、一つあるいは複数のパラグラフになる場合もある）事柄を要約したり、命名したりするなどの「変容」を行なう。高崎（1988）では、指示語を含む複数の語で構成される一まとまりを「指示語句」と呼び、その変容のありかたとして、「要約」「名づけ」「比喩」「次元変換」「形式化」の 5 種に分けて論じている。以下では、高崎の分類を参考にしながら、調査の結果について考察する。

５．３．１　要約

前の叙述部分を端的にまとめたものを後要素の外来語として、指示語句をつくるタイプである。

(14) ゲームがテレビで放送されたせいで、昨年は男の子の間にローラー・スケートの売れ行きが好調だった。だが、このブームもすでに峠を越したという例もある。[70 年 3 月 15 日家庭]

(15) 一台、五万五千円と値が張るにもかかわらず、予想の二倍近い売れ行き（月八千台）だ。このヒットを生んだ秘密は……。[91 年 12 月 15 日経済]

(16) コンピューター要員の不足は世界的な関心事。西独だけをとっても、一九七五年までに五万人の熟練プログラマーおよびオペレーターが必要だという。このため西独教育科学省はこのギャップを埋めるため、多くの大学に“情報科学学部”を創設する提案を考慮中。[70 年 6 月 25 日経済]

(17) 昨年の輸出入状況をみると輸入約四十億円に対して、輸出は四億円と十分の一程度。だが自分が考案し自分の会社でつくった製品をもって、戦後三回も欧米を歩き回った井上さんからみれば、このアンバランスは必ず解消できるという。[60 年 4 月 25 日商工]

５．３．２　次元変換

前で述べられたことを品詞変換させて再登場させたり、“辞的”なものから“詞的”なものへと次元を変換させたりするタイプである。ただし、外来語の場合、上記の「要約」との区別が必ずしも判然としないものも多い。

(18) 租税収入は七月末までに前年同期比一三．四％増の五兆二千六十億円と好調で、このペースが続けば今年度の税収は当初予算の二十六兆四千百十億円を四千九百億円程度上回る。[80 年 10 月 5 日経済]

(19) …世界選手権の直後、マキシーモフはエジプト代表の監督の座に請われた。年俸３０万ドル（約３１５０万円）という条件だった。ロシア代表監督としての年俸が約１０

万ルーブル（約３７万円）のマキシーモフにとってこのオファーが魅力的に映らないはずがない。［00 年 8 月 5 日スポーツ］

(20) アミノ酸はたんぱく質の倉庫のようなもので、中のたんぱく質は絶えず古いものが分解され新しいものが作られている。「体が小さいほどこのサイクルが早く、アミノ酸の倉庫もいっぱいあるからうまい」と沖谷さん。［00 年 4 月 5 日家庭］

(21) …よう」「薬の副作用、手術の後遺症をしっかり聞こう」「不要と思う検査、手術から逃れよう」――などで、このアドバイスをめぐって活発な討論が行われた。［00 年 5 月 5 日家庭］

５．３．３　命名

広い意味では「要約」に含まれるが、このタイプは書き手が考えや捉えかたを新たなものとして名づけることで、「再語彙化relexicalisation」させるタイプである。ほとんどの場合、臨時一語（石井 1993）のようなものになるが、今回の調査[注2]では、単独用法のみを取り上げたので、このタイプに当てはまるものは出現しなかった。

５．３．４　比喩

前の叙述内容と一見無関係に見えるものを、メタファーやメトニミーなどのような技法によって用いるタイプである。今回の調査では、「ウエート」「プロポーズ」「ギャング」の 3 語が得られた。

(22) でも、現実は難しい。親から「がんばれ」と言われるより「がんばったね」と言ってもらううれしさを知っているのですが……。よい人、よい親、よい嫁（このウエートが大きいかな？）に見られようと、必要以上に演じたり、その中に子供を巻き込んでしまって、自己嫌悪に落ち込んだり、悩んだりの連続です。［00 年 4 月 15 日家庭］

(23) いってみれば文化財保存側が経済開発側に“調和できます”と言い寄ったようなものだ。しかし、開発側がこのプロポーズに応じるだろうか。“片思い”ではないか。今後ほんとうに文化遺産を守る戦いをはじめるのなら…［70 年 9 月 15 日総合］

(24) わけても迫力に富むのは、彼らの敵、大スズメ蜂の襲来を叙した部分だろう。虫の世界で猛威をふるうこのギャングのすさまじさ、恐ろしさ。辛苦を重ねて蜜蜂を養ってきたのに、たちまちのうちに全滅させられて…［80 年 5 月 5 日読書］

５．３．５　形式名詞化

前で述べた事柄をまとめたり名づけたりする等の機能はほとんど認められず、大きな範囲を漠然と受け止めるだけの役割をする。今回は、以下の例しか見られなかった。

(25) とくに北の富士の左差しを完全に封じて先手をとったのは立派。「左差しでないと相撲がとれない」北の富士の速攻が生かされるかどうかはその“左差し”にかかっている。前乃山がこのポイントをおさえて、東土俵から向こう正面へ北の富士を押立てた。［70 年 3 月 15 日スポーツ］

2　今回、「この」の後ろにカタカナ表記がくる場合とともに、“、「、『、【、＜、≪の記号が来る場合も合わせて抽出したが、単独用法では、このタイプのものは見当たらなかった。

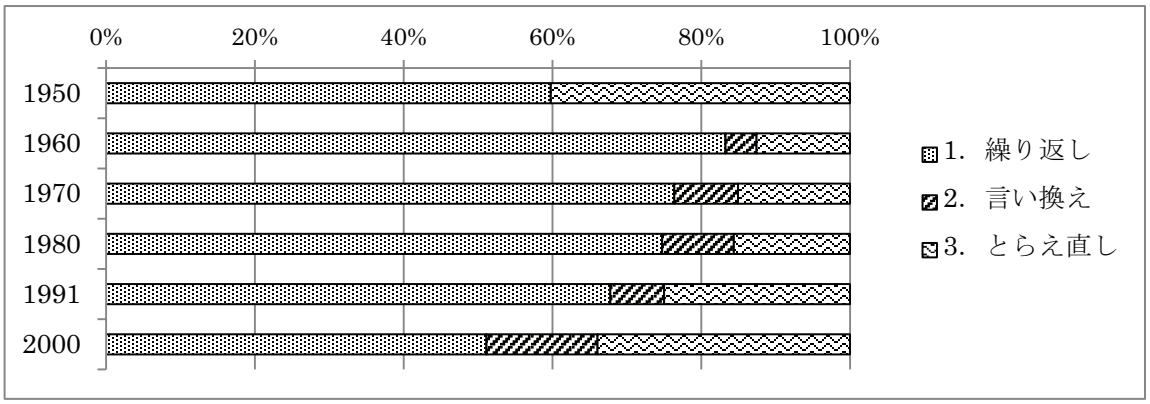
５．４　外来語の文章機能の拡大

以上の分類が「通時的新聞コーパス」においてどのようにあらわれるかを通時的に調査した結果を、表３と図１に示す。

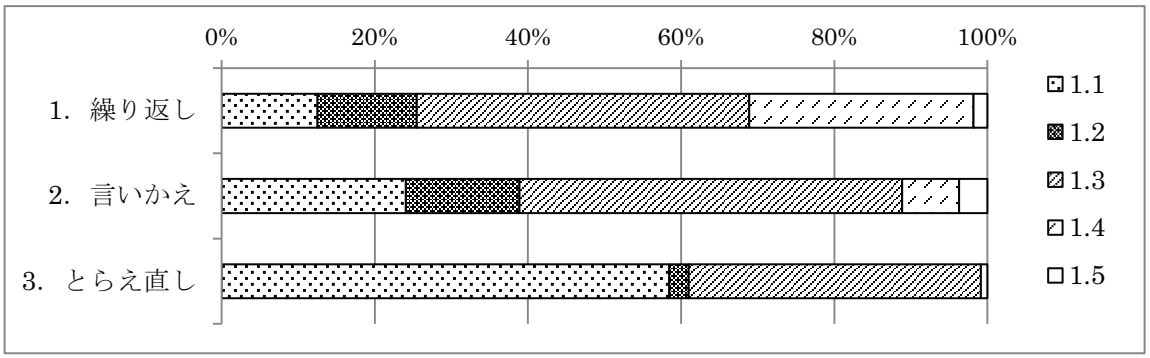
＜表３＞　「この＋外来語」の用法の通時的変化（実数・換算値）

	1950 年	1960 年	1970 年	1980 年	1991 年	2000 年
1. 繰り返し	9	79	142	125	46	46
	11.3	35.8	44.6	38.8	14.1	11.5
2. 言いかえ	0	4	16	16	5	13
	0	1.8	5.0	5.0	1.5	3.3
3. とらえ直し	6	12	28	26	17	29
	7.6	5.4	8.8	8.1	5.2	7.5
4. なし		1	7	5		3
5. 保留	4	7	3	3	3	1

＊上段は延べ語数の実数、下段は 100 万字あたりの換算値



＜図１＞　「この＋外来語」の用法の通時的変化（比率）



＜図２＞ 用法別の意味分野

＜図 1＞からもわかるように、データの規模が小さい 50 年を除くと、「繰り返し」が減り、「言いかえ」と「とらえ直し」が増えてきていることがわかる。つまり、文章論的には、外来語がより複雑な文章構成機能を獲得してきているのではないかと考えられる。

また、『分類語彙表　増補改訂版』（国立国語研究所 2004）の大項目（部門）を用いて、区分ごとの意味分野をみてみた。その結果（図 2）、「繰り返し」は「1.4 生産物･用具」が多

く、「言いかえ」と「とらえ直し」は抽象的な事柄（1.1 抽象的関係と 1.3 人間活動）が多いことがわかった。とくに「とらえ直し」は、「1.4 生産物･用具」の使用例はなく、「1.1 抽象的関係」が他のものより多い。これは、20 世紀後半の「抽象的な事柄を表す外来語の基本語化」現象（金 2011）と合致するところがあり、その背景に「とらえ直し」といった談話構成機能の拡大が関係していることを示唆するものである。

６．今後の課題

今回の調査結果のさらなる検討はもちろん、今回対象外とした合成語用法や「その」「そうした」などのソ系の指示語句を含めた拡大調査を行なう必要がある。さらに、語別に指示語句調査を行ない、金（2011）で行なった語別の基本語化パターンの結果と合わせて検討したいと考えている。

また、これは紙面の大小によるところが大きいためあまり意味を持たない可能性もあるが、紙面構成のない 50 年を除き、よく出現する紙面としては、経済面、スポーツ面、社会面、家庭面、（特集含むともっと多くなる）総合面の順であることがわかった。今後、紙面別（テクストタイプ）にその用例を吟味することも必要である。

謝　辞

本研究は、文部科学省科学研究費補助金「基本外来語の談話構成機能に関するコーパス言語学的研究」（平成 23～25 年度、若手研究 B）による補助を得たものです。

文　献

石井正彦（1993）「臨時一語と文章の凝縮」『国語学』173

金愛蘭（2006a）「外来語『トラブル』の基本語化－20 世紀後半の新聞記事における－」『日本語の研究』2 巻 2 号

金愛蘭（2006b）「新聞の基本外来語『ケース』の意味・用法－類義語『事例』『例』『場合』との比較－」『計量国語学』25 巻 4 号

金愛蘭（2011）『20 世紀後半の新聞語彙における外来語の基本語化』『阪大日本語研究』別冊 3 号

金愛蘭、石井正彦（2012）「同格連体名詞の外来語－文法機能からみた外来語の基本語化」『韓国日本語学会第 25 回学術発表会予稿集』

高崎みどり（1988）「文章展開における“指示語句”の機能」『国文学 言語と文芸』103 号

高崎みどり（2012）「テクストの結束性に与る語彙とその機能について」『第 1 回コーパス日本語学ワークショップ』予稿集、 pp.7-14

Halliday, M.A.K. and Hasan, R.（1976）*Cohesion in English.* London. Longman.［安藤貞雄ほか訳『テクストはどのように構成されるか』ひつじ書房、1997］

Halliday, M.A.K. and Hasan, R.（1985）*Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Deakin University Press.［筧壽雄訳『機能文法のすすめ』大修館書店、1991］

McCarthy, M.（1992）*Discourse Analysis for Language Teachers.* Cambridge Language Teaching Library. CUP.［安藤貞雄・加藤克美訳『語学教師のための談話分析』大修館書店、1995］

機械学習による中国語助詞の用法解析

宋 東旭 (東京農工大学 工学府) *
浅原 正幸 (国立国語研究所 コーパス開発センター)
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)

Comparison of Resampling Strategies for Chinese Auxiliary Word Classification

Dongxu Song (Graduate School of Engineering, Tokyo University of Agriculture and Technology)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

1. はじめに

現代中国語文法の助詞には「語気助詞」、「時態助詞」と「結構助詞」の三種類ある。結構助詞は、日本語の接続助詞にあたり、「的」、「地」、「得」などがこの分類に入る。いずれも/de/と発音され計算機上の入力方法が同じであるため、誤用・混用が多い。

本研究では結構助詞の/de/が出現する位置に「的」、「地」、「得」のいずれの語を選択すべきかを判定する機械学習方法を提案する。対象となる結構助詞の前後二形態素の文脈情報を用い、サポートベクトルマシン (Support Vector Machines; 以下 SVM) (Vapnik (1995)) と多クラスロジスティック回帰（最大エントロピー法）を利用することで入れるべき語を選択する。いずれの手法も分布の偏りに弱く、一方のデータが多い場合にはそちらに引きずられる傾向にある。そこで訓練事例として与えるクラスラベル間の分布を調整することで少ない事例に対応できるかどうかを試みた。SVM は訓練事例が分布する素性空間上の分離超平面を探索する識別学習器であり、ロジスティック回帰は訓練事例がラベルに条件づけられて二項分布で生成されたものとする統計学習器である。それぞれでサンプル数の増減が識別器にどのような影響を与えるのかを検証した結果を報告する。

2. 背景および関連研究

2.1 「的」「地」「得」

「的」はヒト・モノ・コトがどのような態度・状態・程度にある（静態）かを表現するときに利用し、一般に被修飾名詞に前置する。「地」はヒト・モノ・コトがどのような態度・状態・程度で動く（動作）かを表現するときに利用し、一般に被修飾動詞に前置し、修飾形容詞に後置する。「得」はヒト・モノ・コトがどのような態度・状態・程度にする（到達・結果）かを表現

* songdongxu123@gmail.com

するときに利用し、一般に修飾動詞に後置する。

この三種類の結構助詞のうちテキストの出現の観点では「的」が大勢を占めており分布に偏りがある。特に「地」は「的」と誤用され,「得」は「地」,「的」と誤用される (蒋 (2005))。以下にそれぞれの誤用の例を示す。

飞快**的**奔跑	飞快**地**奔跑の誤り	(速く走る)
跑**地**很快	跑**得**很快の誤り	(走るのがとても速い)

現在では中国語学者にも “全て「的」を使用するようにする” という意見もある (王 (2009))。

2.2 機械学習による表現選択

言語処理の分野において機械学習技術を用いた表現選択手法は数多く行われている。古宮ほか (2008) は決定木学習により適切な敬語を選択する規則の獲得手法を提案した。この研究における実験では、判定対象である普通語：尊敬語：謙譲語の比率が 60%:27%:13% もしくは普通語：丁寧語の比率が 49%:51% と比較的均衡が取れたデータであった。竇ほか (2012) はサポートベクトルマシンを用いて中国語のネット語を判定システムを構築した。この研究における評価実験ではネット語 5000 文に対し書き言葉 2000 文で実験しており、極端に分布が偏ったものではない。一方、本研究が対象とする「的」「地」「得」は 95% が「的」である偏った分布であり、これらの先行研究と異なった方策が必要であると考える。次節では分布が偏りがあるデータに対する識別学習の先行研究について示す。

2.3 分布に偏りがあるデータに対する識別学習

分布に偏りのあるデータを Unbalanced Data もしくは Imbalanced Data と呼ぶ。SVM をこのようなデータに適応するにあたり、さまざまな手法が提案されている。その中でデータ集合の均衡を持たせる手法に着目する。Kubat and Matwin (1997) はサンプル数が少ないクラスのデータ量を保持したまま、サンプル数が多いクラスのデータ量を削減することによりデータのバランスを取る方法 (UnderSampling 法) を提案している。UnderSampling 法では、サンプル数が多いクラスのデータを損失してしまうという問題点がある。この問題点に対して少ないクラスのデータ量を重複化させたり、素性空間上で線形補間したりして増やす方法 (OverSampling 法) が提案されている (Japkowicz and Stephen (2002);Chawla et al. (2000))。しかし、SVM の分離平面探索手法の性質から、OverSampling 法ではあまりよい性能が得られず、一般に UnderSampling 法の方がよいことが知られている。そこで Wang and Japkowicz (2009) は UnderSampling 法と OverSampling 法の分布の粒度を変えた SVM を複数作成したものを Boosting する Boosting SVM を提案した。

3. 比較する手法

SVM は二値分類器であり、複数種類のクラスラベルに分類するには one-against-others 法と one-against-one (pairwise) 法などがある (Hsu and Lin (2002)) が、利用する SVM のパッケージ LibSVM は後者を採用している。多クラスロジスティック回帰は正則化手法として L2 正則化と L1 正則化の二つの手法があり、利用する SVM のパッケージ LibLinear はこの両者に対

応している。

本研究では分布が極端に偏った “DE” 相当語の識別について、SVM で提案されているサンプル数を変更する手法の比較を試みる。SVM、L2 正則化多クラスロジスティック回帰、L1 正則化多クラスロジスティック回帰の三つの学習手法について、以下の五つの設定で実験を行い比較する。

- 通常のサンプル分割 (Normal) : 訓練データとテストデータをそのままの分割で訓練する。
- UnderSampling : 訓練データ中最も多いクラスラベル「的」について、20% に削減し、それ以外のクラスラベルについては元のサンプル数のまま訓練する。
- OverSampling : 訓練データ中少ないクラスラベル「得」「地」について、500% に増加し、それ以外のクラスラベルについては元のサンプル数のまま訓練する。
- UnderSampling の多数決 (US Vote) : 訓練データ中最も多いクラスラベル「的」について、20% に削減し、それ以外のクラスラベルについては元のサンプル数のまま訓練した設定で五つ (20% × 5 = 100%) 作成し、その結果の多数決を取る。
- Normal, OverSampling, US Vote の多数決 (Vote) : 上の分割による Normal, OverSampling, US Vote の三つの実験結果の多数決を取る。

4. 評価実験

4.1 実験設定

実験データとして「人民日報タグ付きコーパス (PFR コーパス)」の 1998 年 1 月の記事を用いる。対象となる表現として品詞タグが “u” である「的」「地」「得」を対象とする。表 1 に実験データの分布を示す。

表 1　実験データの分布

ラベル	件数	(割合)
「的」	54487	(95.0%)
「地」	2156	(3.7%)
「得」	661	(1.1%)
合計	57304	

素性として、位置別の前後 2 形態素の単語、品詞情報を固定して用いる。実験において訓練データとテストデータの分割は記事の日付単位の 5 分割交差検定による。

実験設定として五つのサンプル分割を比較する。一つ目は通常のサンプルに基づく識別学習である。二つ目はサンプル数が多い「的」について 500% の (UnderSampling) を行ったものである。「的」のサンプルを五つにデータ分割し、「地」「得」はそのままのものであるものと組み合わせたデータ集合を五つつくり、その結果の平均を取った。三つ目はサンプル数が少ない「地」と「得」について 500% の (OverSampling) を行ったものである。尚、特にサンプル間の線形補間を行ってサンプル数を増やすものではなく、単純な重複化 (duplication) によるものである。四つ目は UnderSampling で評価した五つのデータ集合の結果を多数決を取ったもの

(US Vote) である。五つ目は Normal, OverSampling, US Vote の 3 つのモデルの出力の多数決を取ったもの (Vote) である。

評価は全体の正解率とラベルごとの再現率、精度、F 値による。全体の正解率はラベルが適合したもの/全体の事例数である。再現率はそのラベルで正解した件数/訓練データ中の当該ラベル件数、精度はそのラベルで正解した件数/システムの当該ラベル出力件数、F 値は再現率と精度の調和平均である。

4.2 SVM

SVM のパッケージとして LibSVM を用いる。カーネルは線形カーネルを利用し、その他の設定はデフォールトの設定を用いる。表 2 に結果を示す。

表 2　実験結果 SVM (5 分割交差検定、マイクロ平均)

Normal	正解率 **99.17%**		
	再現率	精度	F 値
「的」	**99.70**	99.48	**99.59**
「地」	92.99	**93.86**	**93.42**
「得」	75.18	**88.59**	**81.34**

UnderSampling	正解率 98.62%		
	再現率	精度	F 値
「的」	98.93	99.69	99.31
「地」	96.09	83.29	89.23
「得」	81.55	72.01	76.49

OverSampling	正解率 99.13%		
	再現率	精度	F 値
「的」	99.67	99.47	99.57
「地」	93.04	93.34	93.19
「得」	74.58	88.03	80.75

US Vote	正解率 98.63%		
	再現率	精度	F 値
「的」	98.90	**99.74**	99.32
「地」	**96.56**	83.18	89.37
「得」	**83.35**	71.09	76.74

Vote	正解率 99.14%		
	再現率	精度	F 値
「的」	99.67	99.49	99.58
「地」	93.13	93.30	93.22
「得」	75.49	87.85	81.20

結果について考察する。まず全体の正解率については通常の SVM が最も性能が高い。OverSampling については、先行研究で言及されているように出力が通常の SVM から少し悪くなった。これは SVM がサンプルからの距離に基づいて分離超平面を設定する学習器であり、同じサンプルを増やすこと自体に本質的には意味がないからだと考える。今後、サンプル数が少ない事例の素性空間上の線形補間のような OverSampling 手法を検討する必要があるだろう。UnderSampling と US Vote は、サンプル数が少ない「地」「得」の再現率を高める一方、精度が悪くなり、結果として各ラベルの F 値および正解率が悪くなった。

4.3 L2 正則化ロジスティック回帰

ロジスティック回帰のパッケージとして LibLinear を用いる。L2 正則化オプション (-s 0) 以外はデフォールトの設定を用いた。表 3 に結果を示す。

結果について考察する。まず、全体の正解率については OverSampling と Vote が最もよかった。L2 正則化ロジスティック回帰は事前分布と事後分布がともに正規分布であることを仮定した事後確率最大化 (MAP) 推定を行う。未知データの対数尤度の期待値を訓練データの対数尤度で近似するが、UnderSampling のような訓練データの削減は元の分布でサンプル数が多い事例を少なくするバイアスをかけるだけでなく、単純に訓練サンプル数を削減するため性能が悪くなると考える。OverSampling は訓練サンプルを増やすために元のバイアスどおりの結果

表 3　実験結果 L2 正則化ロジスティック回帰 (5 分割交差検定、マイクロ平均)

Normal	正解率 99.01%		
	再現率	精度	F 値
「的」	**99.71**	99.32	99.51
「地」	92.25	92.68	92.46
「得」	63.08	91.04	74.53

UnderSampling	正解率 78.83%		
	再現率	精度	F 値
「的」	79.06	**99.71**	88.19
「地」	77.07	81.78	79.35
「得」	64.99	70.94	67.83

OverSampling	正解率 **99.03%**		
	再現率	精度	F 値
「的」	99.43	99.62	**99.53**
「地」	95.50	88.82	92.04
「得」	77.15	84.29	80.56

US Vote	正解率 98.45%		
	再現率	精度	F 値
「的」	98.77	99.69	99.23
「地」	**96.42**	80.11	87.51
「得」	**78.21**	71.31	74.60

Vote	正解率 **99.03%**		
	再現率	精度	F 値
「的」	99.44	99.62	**99.53**
「地」	93.13	**93.30**	**93.22**
「得」	75.49	**87.85**	**81.20**

が得られている。US Vote はこの訓練サンプル数を削減することによる欠点を多数決に補うため、サンプル数の少ない「地」「得」の再現率をあげるという元のバイアスどおりの結果が得られているが、精度が低い傾向にある。特に Vote についてはサンプル数の多い「的」の F 値を保ったまま、サンプル数の少ない「地」「得」の再現率にバイアスをかけたうえで、F 値/精度がよい傾向がみられる。

しかし、残念ながら L2 正則化そのものの結果は全体的に SVM に劣っており、サンプルの変化をしない SVM の方が性能がよいという結果になった。

4.4　L1 正則化ロジスティック回帰

前節の実験と同様にロジスティック回帰のパッケージとして LibLinear を用いる。L2 正則化オプション (-s 6) 以外はデフォールトの設定を用いた。表 4 に結果を示す。

表 4　実験結果 L1 正則化ロジスティック回帰 (5 分割交差検定、マイクロ平均)

Normal	正解率 98.97%		
	再現率	精度	F 値
「的」	99.66	99.34	99.50
「地」	91.88	**91.75**	91.81
「得」	65.05	**89.21**	75.24

UnderSampling	正解率 78.80%		
	再現率	精度	F 値
「的」	**99.71**	79.04	88.18
「地」	81.53	77.00	79.20
「得」	65.03	69.66	67.27

OverSampling	正解率 **99.00%**		
	再現率	精度	F 値
「的」	99.44	99.58	**99.51**
「地」	94.76	89.05	**91.82**
「得」	76.25	83.86	**79.87**

US Vote	正解率 98.29%		
	再現率	精度	F 値
「的」	98.61	**99.69**	99.15
「地」	**96.19**	79.13	86.83
「得」	**78.21**	66.11	71.66

Vote	正解率 98.98%		
	再現率	精度	F 値
「的」	99.46	99.56	**99.51**
「地」	94.48	89.18	91.75
「得」	74.43	84.54	79.16

結果について考察する。L1 正則化ロジスティック回帰は事前分布が Laplace 分布、事後分布が正規分布であることを仮定した事後確率最大化 (MAP) 推定を行う。学習時に不要な説明変数 (素性) に対する重みが 0 に縮退することができる。高次元の疎な素性空間に対して、説明変数を減らす場合には有効であるが、本タスクにおいては全体的に性能が悪くなった。これは、

前後 2 単語の素性を用いており素性空間が疎であるとはいえ、1 事例あたりの発火する素性は 8 つに限定されており、重みが 0 に縮退することで不定となる事例が多々あるからだと考える。L2 正則化と同様に US Vote がサンプル数の少ない「地」「得」の再現率をあげるバイアスがかけられている一方、精度が極端に悪くなる傾向にあり、Vote よりも単純な OverSampling がもっともよい正解率/F 値が得られることがわかった。

5. おわりに

本研究ではラベルの分布が偏った中国語助詞分類タスクにおいて、サンプル数を変えることによる機械学習器ごとと識別性能の変化を評価した。サンプル数が少ないラベルの再現率を上げるもっとも有効な方法として、サンプル数を減らした弱学習器の多数決を取る手法 (US Vote) がどの機械学習器にとっても有効であることがわかった。一方、全体の性能の観点からみると、SVM で通常のサンプル分割で学習するものが最もよかった。これは新聞記事という言語資源であっても、本来「的」を用いるべきではない部分に、「地」「得」を用いる傾向があるからではないかと考える。今後、識別学習の境界値情報をもつ Support Vector となった事例を検証しながら、元データの誤用例について言語学的な分析を進めていきたいと考える。

参考文献

Chawla, N., K. Bowyer, L. Hall, and W. P. Kegelmeyer (2000). “Smote: synthetic minority oversampling technique.” *International Conference on Knowledge Based Computer Systems*.

Hsu, C.-W., and C.-J. Lin (2002). “A comparison of methods for multi-class support vector machines.” *IEEE Transactions on Neural Networks*, pp. 415–425.

Japkowicz, N., and S. Stephen (2002). “The class imbalance problem: A systematic stydy.” *Intelligent Data Analysis*, 6:5.

Kubat, M., and S. Matwin (1997). “Addressing the curse of imbalanced training sets: One-sided selection.” *Proceedings of the 14th International Conference on Machine Learning*.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*.: Springer.

Wang, B. X., and N. Japkowicz (2009). “Boosting support vector machines for imbalanced data sets.” *Knowledge and Information Systems*.

王海峰 (2009).「区别使用“的，地，得”」 編集之友 2009 年 11 期.

古宮嘉那子・但馬康宏・小谷善行 (2008).「決定木を用いた敬語の選択ルールの獲得」 情報処理学会論文誌, 49:7, pp. 2679–2691.

蒋紹愚 (2005).『近代漢語研究概要』.

寶梓瑜・古宮嘉那子・小谷善行 (2012).「コーパスを用いた中国語ネット語の判定システム」第一回コーパス日本語学ワークショップ予稿集, pp. 161–166.

関連 URL

- LibSVM `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
- LibLinear `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

ポスター発表（1） Bグループ

2月28日（木） 14:00～15:00

TVCM における和製英語のパイロット調査

――文字テクストと音声テクストの対照を軸に――

小林善久（一橋大学大学院生）[1]

A Pilot Study of "Wasei-Eigo" in TVCM ---Focusing on the Comparison Between Written Text and Spoken Text---

Yoshihisa Kobayashi

(Hitotsubashi University Graduate School of Language and Society)

1. はじめに

本発表では、和製英語が日本国内における日常生活でどの程度使用されているか、その実態を知るために行ったパイロット調査の結果を報告する。調査に使用したデータは、2012 年 1 月 1 日に放送された TVCM100 本における音声と画面上の文字である。本発表は、そこで使われている和製英語が、記述された文字テクストと音声テクストの面で異なった特徴を出しているかどうかを調査することを主たる目的とする。その特徴分析は、100 本の TVCM をサンプルに、一つは和製英語の数量比較を中心に、もう一つはそれぞれの和製英語の語構成や共起する語句の特徴などを調査することとする。また、文字テクストにおける表記の特徴としてアルファベット表記が多いことなども併せて報告する。

TVCM を対象とした理由は、テレビの広告業界の言葉が、人々の関心を引くことに腐心して創造された言葉であり、様々な場所で繰り返し放映されることで、比較的広範囲の人々が注目する機会が多い言葉だと考えるからである。また TVCM が話し言葉としての音声言語に基づき、なおかつ画面には文字も表示されることで、言葉の本来の機能を十分に発揮している言語として十分に研究の対象になると確信する。

2. 調査概要

今回のパイロット調査では、過去の国立国語研究所の調査方法[2]を参考にして、調査項目を、以下の(1)key/和製英語 (2)CM 商品名 (3)会社名 (4)DATE(日付) (5)時間帯 (6) CH(チャンネル) (7)業種 (8)CM の対象とその属性(性別、年齢、職業等) (9)時間(長さ)(10)英語への言い換え (11)造語パターン (12)語種 (13)語形 (14)外来語の有無 (15)備考 (16)画面上の広告(語/文) (17)音声で流れた広告文 (18)発言者属性(性別、年齢、職業等) (19)混種語サンプル (20)ローマ字(英語文字)サンプル (21)カタカナ英語、の 21 項目とした。

本発表では、上記の調査項目の中でも、(16)画面上の記述広告と (17)音声で流れた広告を中心にそれぞれにおける和製英語の実態調査を報告する。また和製英語と同時に外国語のアルファベット文字が画面上でどのような割合で使用されているのかの数量調査も含めて報告する。

---

[1] lm112008@g.hit-u.ac.jp

[2] 国立国語研究所編（1995)『テレビ放送の語彙調査 I ---方法・標本一覧・分析---』国立国語研究所報告 112

3. 調査内容　―画面に現れる和製英語と音声で現れる和製英語について―

3.1 今発表のための調査に先立つ事前調査

3.1.1 語種調査(TVCM100 本の語種別概観)

はじめにこのパイロット調査 TVCM100 本の言語景観を得るために、語種調査をした。画面の記述文字と音声を文字化したものを合計して数えた。下記の表 1 の語数は短単位[3]に基づく語数調査である。CM100 本の全語数 6322 語のうち外来語が 831 語で 13%の割合は、過去の国立国語研究所の 1990 年代の雑誌 70 種の調査と比べても高い。

表 1：TVCM100 本の語種調査(茶まめ＋手計算[4])

全語数	6322	出現頻度	国研雑誌調査[5]
和語	3305	52.3%	41.6%
漢語	1953	30.9%	45.9%
外来語	831	13.1%	10.6% (外国語を 含む)
混種語	233	3.7%	2%

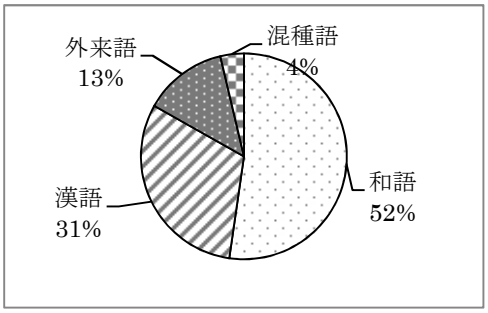


図 1：TVCM100 本の語種調査

3.1.2　和製英語の割合(長単位)

次に、和製英語の語数をまとめて数えた。TVCM 全体の中での和製英語の占める割合を見るためである。この数え方は、表 1 とは異なり、複合語を形態素に分けずに、1 語として数えることにした。固有名詞を含めて数えてある。なお、語種としての外来語の内、和製英語を除いた本来の外来語を、英語以外のものを含めて以降「外国語」として扱う。

表 2：TVCM の中の和製英語数

項目	延べ数	異なり数
CM 本数（13 重複）	100	87
和製英語の語数（33 重複）	229	181
和製英語を含まない TVCM 数	10 / 100	7 /87
外国語を含まない TVCM 数	4 / 100	3 /87

以上が、画面と音声の双方を併せた全体的な分布状況である。和製英語が、9 割の CM に登場し、平均すれば 1 本に 2 語ずつは現れるというのが概観である。また外国語もそれ

[3] 野村雅昭(1973)「複次結合語の構造」国立国語研究所『電子計算機による国語研究Ⅴ』秀英出版

[4] 茶豆の形態素解析では、アルファベット文字を記号という範疇に、また固有名詞という語種分類に相当しない範疇があったりしたので、それを手計算で数え直したものである。

[5] 伊藤雅光(2007)「雑誌に見られる外来語と外国語」の「1990 年代の雑誌 70 種・本文」の語種調査より引用。

以上に多用されている実態がわかった。

3.2 今回の調査

さて、ここから本題の画面と音声に現れる和製英語の比較調査を報告する。

3.2.1《調査１》文字列による総量比較

和製英語には複合語も含まれるので、長単位の解析でなければならないが、今回の調査には間に合わなかったのでそれぞれの文字列数の総和を比較してみた。すると下記の表 3 のように音声として現れる文字列数のほうが画面に現れる文字列数より約 100 字分多いことがわかった。これは全体数を考慮すると、大きな差とは言えない。しかし予想に反して音声テクストの方がわずかだが多かった。画面に記述される文字は、文というより、商品名、会社名、決まり文句などの体言止め表現がかなり多く使われていることがわかった。

表 3：画面テクストと音声テクストの比較

	画面テクスト	音声テクスト
文字列の総和（全角 1 文字＝1）	6320	6418
体言止めの総数	422	235

3.2.2《調査 2》和製英語の出現頻度数の比較

次に、画面と音声に現れる和製英語の出現頻度数であるが、ここでは画面テクストの方に音声で流れる和製英語の数に比べて 2 倍近く出てくる。表 4 に沿って言えば、固有名詞を含めた述べ語数で比較した場合は 1.92 倍、異なり語数では 1.66 倍といずれも大きく画面テクストに傾斜して出現する。固有名詞を除いた数で比較すれば延べ語数の場合は 2 倍を超えている。これはなぜか？

まず数が少なくて扱いやすい異なり語数で考えることにする。次に固有名詞も同様に異なり語数で考察する。

表 4：TVCM の中の和製英語数

	画面 TEXT（A）	音声 TEXT (B)	(A)/(B)
和製英語の数(延べ語数・固有名詞を含む)	175	91	1.92
和製英語の数(異なり語数・固有名詞を含む)	136	82	1.66
和製英語の数(延べ語数・固有名詞を含ない)	126	61	2.07
和製英語の数(異なり語数・固有名詞含まず)	98	56	1.75
和製英語の固有名詞の割合(異なり語数)	27.9%	31.7%	
外国語の数(延べ語数・固有名詞を含む)	175	134	1.31
外国語の数(延べ語数・固有名詞を含ない)	107	94	1.14
外国語の中での固有名詞の割合	38.9%	29.8%	

画面テクストによりはっきりと多く和製英語が現れるその要因は、語構成にあると考え

る。和製英語全体は異なり語数で数えると 117 語になるが、漢語と英語の混種語によるものがその内の 7 割にあたる 82 語を占めている。しかも、その混種語の半分以上の 47 語が画面にだけ出現する。その中でも「臨時一語」の占める割合が非常に大きい。今回の独自の認定[6]では、56 語がそれに相当するが、その中で画面だけに現れるものが 39 語、音声だけに出現するのがたったの 6 語しかない。新聞や雑誌の「見出し語」と同じように、一目見ただけで意味が分かるような機能を持ち合わせている漢語との混種語を TVCM も上手く利用している。言い換えると漢字のもつ表意性が和製英語の混種語という形態で、画面に多く出現するというのが最大の特徴と言える。

3.2.3《調査 3》TVCM の外国語の場合での比較

純粋な外国語の出現頻度は、表 4 に示したとおり、画面表示の方が高い。その出現数差は、和製英語での比較に比べ、あまり大きくはないが、ここで最も特徴的なのが、画面に現れる外国語には、固有名詞の比率がかなり高いことになる。 つまり商品名、会社名などに外国語が多く使われているのだ。[7]　アルファベット文字表記がカタカナに代わる文字表記体として進出してきていることも特徴である。

3.2.4《調査 4》和製英語の定義の分類に基づいての比較

これまで依拠してきた和製英語の定義にそって、語構成分類を試みた。定義は、玉岡賀津雄(2009)[8]の分類を下敷きして独自に作ったものである。

【定義】

和製英語とは
◎日本語の語彙のうち、日本で作られた英語風の外来語および混種語のことを言う。
◎英語の母語話者が意味を理解するのに苦労することが多いものである。
◎外来語を使った造語を指す、という狭い解釈もあるが、ここでは以下のように、少し広めた解釈をするものとする。
①英語として存在するが、英語の意味・用法とは異なる意味・用法で使われるもの。
　(例)「スマート」
②英語として存在するが、原語の発音とは大きく異なるもの。　(例)「ツーダン」
③英語の単語を短縮したり、一部省略したりして使われるもの。(短縮した複合語も含む)　(例)「デパート」
④英語の単語と日本語の和語・漢語とを組み合わせて使われるもの。(例)「スタバる」
⑤英単語には存在しないが、日本語の中で使われるようになったもの。(例)「ナイター」
⑥実際に存在する英単語を組み合わせて造り、新しい意味を付加した合成語(複合語) 。
　(例)「テーブルスピーチ」
⑦実際の英語とは語順や文法配列または単語を変えて、作り出されたもの。
　(例) *EXILE MUSIC VIDEO BEST*

---

[6]林四郎(1982)「臨時一語の構造」『国語学』131 と石井正彦(2007)「第 3 部臨時一語の形成」『現代日本語の複合語形成論』の定義が異なるため、前者をベースに自分の判断を交えながら進めた。

[7] 今回に先立つ「和製英語の語構成──TVCM パイロット調査の分析と考察」で、製品名 58% 会社名 36% となっている。

[8]玉岡賀津雄(2009)「韓国語母語話者による和製英語の理解」

⑧英単語の音と同音の日本語の単語と両方を兼ねて使われるもの。
(例)「 イエー(Yes/家)」
⑨英語や日本語の一部を省略して、アルファベット文字を用いた英語にはない省略語。
(例)「NHK」

以上の定義分類に沿って、今回の TVCM100 本に出現した和製英語の分類を以下の表にまとめた。( )内の数字は、和製英語のそれぞれの語数(異なり語数)を示したものである。

表 5：TVCM 中の和製英語の定義による分類

・ は音声だけで現れたもの。 は画面だけに現れたもの。無印は両方に出現したもの。

1)英語として存在するが、英語の意味・用法とは異なる意味・用法で使われるもの。(6)

メイク(名)/ リフォーム/ スリーショット/ お年玉バーレル/ オフ/ サイン(名)/

2) 英語として存在するが、原語の発音とは大きく異なるもの。(2)

イメージ/ テーマソング

3) 英語の単語を短縮したり、一部省略したりして使われるもの。(短縮複合語も含む) (12)

フラワーアレンジ/ワンセグ TV/ ファンデ/ スマホ/ デコメ/ ハピデコ/ モバプロ/ エネファーム/
アクション RPG/ モンプラ/ギャラ/シンクロ/

4)英語の単語と日本語の和語・漢語とを組み合わせて使われるもの。 (82)

リフォームする/ デザインする/ オフ(動)/ 人気モデル/ 水曜ドラマ/ 天才ジャズピアニスト/
大初夢フェア/ 女子力アップ/低燃費タイヤ/ 髪ドック/専用ソフトウエア/エース対決/ 雪ガール/
売上シェア/ 臨床データ/ スパイ大作戦/ 美女ゴルファー/ ものまねスター/ リーグ戦/ 低燃費エコカー/
初売りフェア/ 証券コード/ 銘柄コード/ バネブラシ/ワイルド現象/ オシャレスモール/プレイヤー同士/
シリーズ史上/ プラスチック製/ イメージ図/ お年玉クーポンパスつき/ パケット通信料/ アクセス可能/
エネルギー不足/ ガソリン車/ パッケージ版/ ダウンロード版/ セール終了後/ メイクアップ効果/
新サービス/ 新燃費測定モード/ 今シーズン/ 超一流パフォーマー/ 一部コンテンツ/ 新エンジン/
ジャスダック上場企業/ ビデオクリップ集/ エコカー減税/ 着うた R 配信中/ お年玉バーレル/酒 DS 計
綾香スペシャルプログラム配信中/ リッター30 キロ/ アルペングループ特選品/ 上下セット/
ジュニア 3 点セット/ 2 時間半 SP/ 1 回 2 カプセル/ 婦人ジャカードストール 24%OFF/
ノンアルコールビールテイスト飲料/ 区間エントリー/ ミッション発令/ 5 枚刃モイスチャージェル BOX/
「カラダまるごと」コントローラーKINECT for XBOX360/ ふくだけコットンさらさらオイルインビオレ
ゲーム画面/ アレルギー体質/ モバイルオンラインプロ野球/ 世界最強エース陣/ JC08 モード走行/
ダブル A 面シングル/ 成人式スーツ・コート/ 2 倍増毛感ボリューム/ SAPPORO 企業 CM ソング/
新型 MR ワゴン/ 7 インチワンセグ TV 内臓/ スター・ドラフト会議/新型アルトエコ誕生// シード決着/
婦人カシミア 100%/ タートルネックセーター37%OFF / CVT 搭載

5)英単語には存在しなかったが、日本語の中で使われるようになったもの。(0)

6)実際に存在する英単語を組み合わせてつくり、新しい意味を付加した合成語(複合語)(6)
フリーダイヤル/ カシスオレンジ/ スマイルバーゲン/ クリアファイル/ エネルギー・フロンティア/ ボリュームマスカラ/
7)実際の英語とは語順や文法配列または単語を変えて、作り出されたもの。(6)
EXILE MUSIC VIDEO BEST/ ジュニアスノーウエア/ アルコールゼロ/ カロリーゼロ/ 糖質ゼロ/ アバターGET!/
8)英単語の音と同音の日本語の単語と両方を兼ねて使われるもの。(3)
イエー(Yes/家)/ ノンアル気分/ キッチン泡ハイター
9)英語や日本語の一部を省略して、アルファベット文字を用いた英語にはない省略語。 (1)
NTT/

表全体から、混種語が圧倒的に多い。英語と和語ないし漢語との混種を機械的に和製英語とした定義に問題があるのかもしれない。玉岡(2009)以外にも、田辺(1989)[9]、野村(1984)[10]、鈴木(2008)[11]や広辞苑他の辞書を通して、作り上げたものだが、更なる精査が必要なのかもしれない。以降、定義の分類順にそって分析してみた。

1) 「メイク」が単独なら英語では動詞扱いなのであるが、「メイキャップ」が短縮して名詞に転成した点、および「オフ」が本来の英語では名詞でも動詞でもないものが、日本語の中で動詞や名詞の機能を果たしている点に新規性がある。「オシャレスモール」の「スモール」も英語の形容詞から日本語の名詞に近づいている現象ではないかと思われる。以上は用法の観点からであったが、意味の上での英語との差異を示したものに「スリーショット」がある。「ツーショット」は日本語から生まれた写真の被写体の数を指したものだが、「スリーショット」は英語ではまだその意味は存在しない。

2) 発音上のユニークさをどこまで厳密にするかで、範囲が変わりやすいもので、扱いにくい項目であるが、これまで「イメージ」と「テーマ」は英語母語話者に通じない場面に何度か遭遇したので避けられない気がした。前者では、英語 2 音節に対し日本語 4 モーラ、後者は 1 音節対 3 モーラの違いが、それぞれ母音と子音の違い以上に大きな差異要素になっている。

3) 短縮による複合構成が予想外に少なかった。窪薗(2002)[12]によれば、2 モーラ＋2 モーラの複合短縮の形が最も広く見られる新語形成パターンのようだが、ここでは「フラワーアレンジ」「ファンデ」のように後半の一部が欠け落ちたり、「スマホ」「デコメ」のように 3 モーラに落ち着くケースもあり、種々雑多な感がある。短縮の動機は「言語の経済性」であろうから、増え続ける外来語のカタカナ表記のウェートを軽くしようとする動きはまだまだ活発化していくと思われる。ここでは画面表示と音声の双方に表われる語がほ

9 田辺洋二(1989)「和製英語の形態分類」『早稲田大日本語研究教育センター紀要２』
10 野村雅昭(1992)「造語法と造語力」『日本語学』5 月号 PP.4-7
11 鈴木俊二(2008)「和製英語の研究 ---その構造と思想」『国際短期大学紀要』第 23 号 PP.1-47
12 窪園晴夫(2002)『＜もっと知りたい！日本語＞新語はこうして作られる』岩波書店

とんで(9 語)、画面だけが 1 語、音声だけが 2 語と偏りはほとんどないと言える。

4)　混種語については、雑多な要素が入っているので、今後更なる階層化した下位分類が必要である。名詞であるがサ変動詞「する」と結合できる動名詞の諸相、接頭辞に漢語・和語が来る場合の造語法の特色、接尾語の場合、省略語の実態等々の分類方法の確立が待たれるところである。特に「臨時一語」の扱いについては、慎重に対処する必要がありそうだ。いずれにしても、混種語の語構成は、音声よりは画面表示にした方が、味わいやすいことがはっきり数字に現れた。

5)　今回の該当例は存在しなかったが、「ナイター」「OL」などがこの例である。ここに該当するものは、現在では英語圏でも使われており、場合によっては辞書にも載っているものもある。1)の項で取り上げた「スリーショット」の類義語でもある「ツーショット」も和製英語として取り上げた研究例があり、日本語が意味を一つ増やしたことになる。

6)　これは英単語どうしの組み合わせによる複合語であり、混種語でもなければ、外来語としての英語でもない日本人による和製英語であるが、比喩というレトリックを使って生み出したものが多いように思える。例えば、「フリーダイヤル」は、受話器の文字盤をさす「ダイヤル」が「電話」という上位語に代わって使用されたシネクドキー（提喩）であり、「エネルギー・フロンティア」は、「エネルギーの開拓者」に喩えるメタファー(隠喩)であり、「スマイルバーゲン」や「ボリュームマスカラ」は、それぞれ「バーゲンで得意顔」になり、「濃厚なマスカラを見てボリューム感」を抱くメトニミー(換喩)であると考えられる。池上(2006)[13]が指摘するように、従来の語の意味の範囲を超えて新規なイメージを作り出そうとする創造性が働いているのだろう。このあたりが TVCM の真骨頂ではなかろうか。このレトリック手法は、当然ながら語だけに限定されるものではないので、この和製英語に共起する前後の語(句)との関連にも目を向ける必要があるが、頁数の制限で割愛する。画面と音声の分布は、同数だった。

7)　ここでは日本語の発想に基づいた語順で英単語を並べるのが主な特徴。Best の位置（「～がベスト」）、SOV の文型パターン（「アバター(を)ゲット」）、数量詞ゼロの使い方等が代表的になっている。今回の調査では見られなかったが、命令文ではない「主語抜き述語文」も時々見かけることがあるが、それもここに入る。

8)　これはよくある同音異義を掛詞にするパターン。和製英語の特徴というよりは、言語一般の言葉遊びと言っていいもの。

9)　今回は「NTT」の 1 語のみだった。CM 全体の中では、頭文字化する前の単語が外国語としての英語であるものが圧倒的に多く、意外にも'KY'のような軽いノリの言葉使いが見られなかった。

13 池上嘉彦(2006)『英語の感覚・日本語の感覚＜ことばの意味＞のしくみ』(日本放送出版協会)

3.3 TVCM 画面における文字表記の特徴について

今回のパイロット調査の中で製品名と会社名におけるアルファベット文字の多さは特筆である。商品名では、以下のグラフが示すように、アルファベットが、他との組み合わせも含めると半数を超えているところに特色がある。人目をひきやすいか、人に訴えかける力が強い、というような新奇なイメージがアルファベットにはあるように思われる。

次にアルファベット文字を使用していないものをピックアップすると以下のものになる。[リーブ 21、ソルマック、ベンザブロック、新コンタックかぜ総合、キッチンハイター、エーザイ/チョコラ BB、リポビタン D、箱根駅伝、初詣]

上記項目で下線を施したものは、医薬品の名称であるが、カタカナが多い。アルファベットで表記されないのは、高齢者を意識して、時には生命に関わるような重大な言い間違い等の回避をねらったものではないかと考えられる。かってはカタカナが翻訳語として斬新な人目を引く魅惑的な要素があったというが[14]、ここではより確実に正確な情報を伝えるという異なる意味合いが込められていて、かっての注目を引くための機能面は、アルファベット文字に移行しているような印象を受ける。

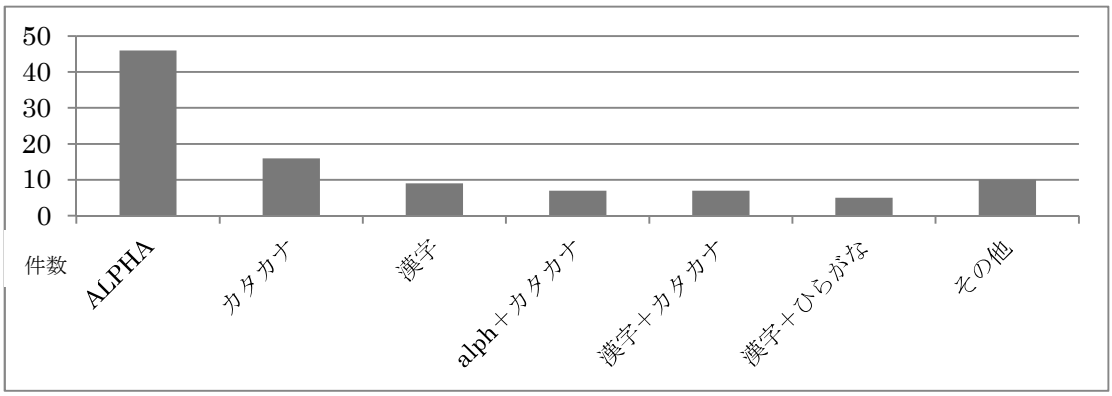


図 2: 商品の文字表記の分類

次に会社名についても、同様な調査をするが、ここで扱う会社名は正式に登録された会社名ではなく、TVCM 上に現れる会社名の表記法を採用した。つまり会社のロゴのようなものである。上記(2)と同様な調査をしてみると、以下のようになった。

表 6：会社名の文字表記

アルファベット	漢字	アルファベット＋カタカナ	漢字＋カタカナ	漢字＋アルファベット
62	15	1	15	3

[14]柳父章(2004)『近代日本語の思想 ― 翻訳文体成立事情』(法政大学出版局)

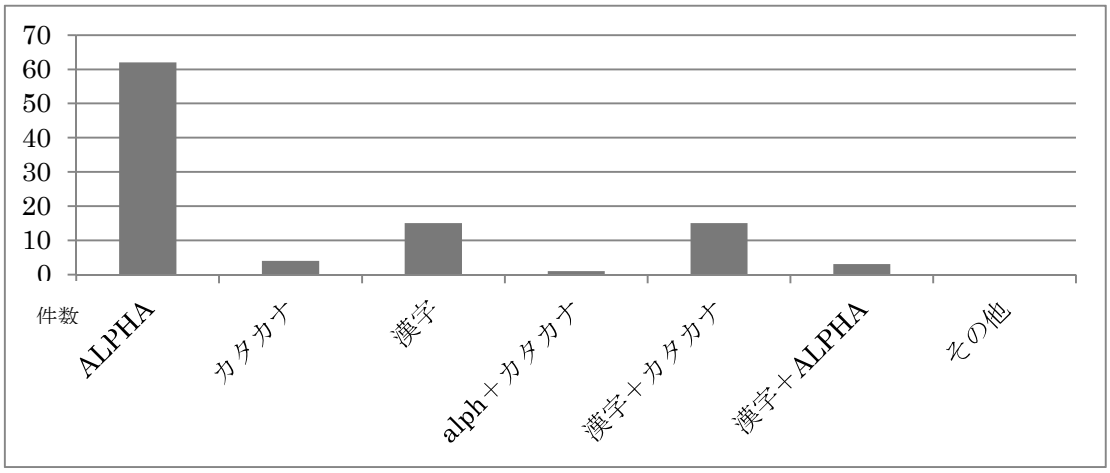


図 3：会社名の文字表記

以上、商品名と会社名の語種と文字表記を比べてみると、商品名は 8 割に外国語が使われているのに対し、会社名では混種語を含めても 6 割弱が外国語であり、漢語の割合が高まっているところが大きな違いとなる。文字表記においては、どちらも外国語の生の形でのアルファベット表記が際立って多いのが特徴であるが、会社名の方でも漢字名が製品名に比べて多いのも共通している。

これまで雑誌や、書籍、様々なコーパス調査を通じて語種調査[15]が行われてきたが、どれも和語と漢語が中心で、それらが大勢を占めるのが趨勢であったが、この 2 項目では、それとは全く異なる現象が出ているのが特徴である。TVCM においては、強調したり、相手の注目を引くための工夫として生の外国語の文字表記を用いる手法が主流になりつつある。端的に言えば、英語が文化として日常生活の中に深く入り込んできている裏返しではなかろうか。是非は別として、英語が政治的経済的に最有力な言語であるが由に、日本だけではなく世界中の多くの国でこのような現象が見られる。[16]

最後に、CM 画面全体に現れたアルファベット文字列の量を調査した。結果は下記の表に示すとおりであるが、算定は、ワードの文字カウントに基づいたものである。アルファベット文字の表示割合の 5.8%が高いかどうかは、明確な比較資料とは言えないかもしれないが、先に示した 1990 年代の国立国語研究所の「現代雑誌 70 種の語彙調査」に見られるラテン文字の割合が 3.9%であることを参照にすれば、その比率は決して低くはない。この数字の意味するところは、伊藤(2002)[17]の予測どおり、アルファベット文字がカタカナ文字に代わって新規性を示す代表的な文字になる方向へ進んでいるのではなかろうかと考える。

---

15 国立国語研究所(1964)『現代雑誌九十種の用語用事　第三分冊　分析』、山崎誠・小沼悦(2004)「現代雑誌における語種構成」(第 10 回言語処理学会ポスター発表要旨, 福田亮・伊藤雅光・塩田雄大 (2007)「日本語の中の外来語と外国語---新聞、雑誌、テレビ」国立国語研究所第 30 回「ことば」フォーラム発表資料

16Armin Mester (2011) 「日本語とドイツ語における英語の影響？」(国立国語研究所研究発表)

17 伊藤雅光(2002)『計量言語学入門』大修館

表 7：CM 画面に出現したアルファベット文字列

アルファベット文字数	CM 全体の文字数	アルファベット文字数の占める割合(%)
366	6320	5.8%

文　献

池上嘉彦(2006)『英語の感覚・日本語の感覚＜ことばの意味＞のしくみ』(日本放送出版協会)

石井正彦(2007)「第３部臨時一語の形成」『現代日本語の複合語形成論』

伊藤雅光(2007)「雑誌に見られる外来語と外国語」の「1990 年代の雑誌 70 種・本文」の語種調査より引用。

窪薗晴夫(2002)『＜もっと知りたい！日本語＞新語はこうして作られる』岩波書店

国立国語研究所編 (1995)『テレビ放送の語彙調査Ｉ ---方法・標本一覧・分析---』国立国語研究所報告 112

鈴木俊二(2008)「和製英語の研究 ---その構造と思想」『国際短期大学紀要』第 23 号　PP.1-47

田辺洋二(1989)「和製英語の形態分類」『早稲田大日本語研究教育センター紀要２』

玉岡賀津雄(2009)「韓国語母語話者による和製英語の理解」

野村雅昭(1973)「複次結合語の構造」国立国語研究所『電子計算機による国語研究Ⅴ』秀英出版

野村雅昭(1992)「造語法と造語力」『日本語学』5 月号　PP.4-7

林四郎(1982)「臨時一語の構造」『国語学』131

柳父章(2004)『近代日本語の思想 ― 翻訳文体成立事情』(法政大学出版局)

共起語集合の頻度分布と語の属性との相関

山崎　誠（国立国語研究所言語資源研究系）†

Correlation between Frequency Distribution of Collocational Set and Key Word's Attribute

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

本稿は、コロケーションを計量語彙論的な観点から記述することを目的とする。コローションの定義はさまざまであるが、本稿では、文脈においてある語と共起する別の語と組み合わせのことと広く捉える。Halliday & Hasan（1976 : 374[1]）では、コロケーションは再叙と並んで語彙的結束性のひとつとされ、「コロケーションによる結束性がテクストに及ぼす効果は，微妙なもので評価しにくい。」（同前 : 379）とされている。本稿では、Hallidayらが取らなかったアプローチ、すなわち、コロケーションという現象を集合としての語彙を量的に観察した場合にどのような特徴が見えてくるかについて考察するものである。

２．共起語集合

本稿で利用する概念「共起語集合」について説明する。計量語彙論では集合としての語彙をもとに、延べ語数や異なり語数、類似度などを利用して分析を進める。本稿ではコロケーションのキーとなる語の前後の一定の距離に現れる語の集まりを考える。例えば、(1)のような語の連続による文脈があった場合、t_i がキーとなる語、t_{i-1} がキーの１語前の語、t_{i+1}がキーの１語後の語などとなる。

(1)　…, t_{i-3}, t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}, t_{i+3},…

対象表現域において、ある単位語 t が、同一の見出し語 m に対応するすべての場合において、t との距離 d の位置にある語の作る集合を $V_{m(d)}$と書くことにする。距離は、相手となる語の相対的位置からキーとなる語の相対的位置を引いた値で表す。したがって、m 自身との距離は 0 であり、前文脈方向がマイナス、後文脈方向がプラスとなる。定義により、$V_{m(0)}$は要素の異なりが見出し語 m のみである集合となる（ただし、m の延べ語数は 1 とは限らない）。このように定義した $V_{m(d)}$ を考えたとき、d の値の変化によって、$V_{m(d)}$の計量的指標がどのように変化するか、また、その変化は見出し語 m の持つ属性とどのように関係するかが本稿の興味の中心となる。

３．データと方法

本稿で利用するデータは『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）である。BCCWJ には 13 のレジスターがあるが、本稿ではそのうち主として図書館書籍（LB）を利用する。分量が多く、結果の安定性が得られるためである。なお、本稿で用いる言語単位は、短単位である。

上で定義した $V_{m(d)}$を求める見出し語の選定は、使用頻度の多い語から品詞を異にするものを適宜選んだ。d の範囲は、文を越えないものとする[2]。したがって、見出し語 m を持つ単位語 t が文末にある場合、後続の文脈がないため、t_{i+1}は存在しない。なお、距離の測定

† yamazaki@ninjal.ac.jp

1 ページは邦訳による。以降の同書からの引用も同じ。

2 文の認定は BCCWJ の DVD に含まれる短単位 TSV ファイルの文頭ラベルを使用した。文頭ラベルが B(=文頭）である語から、次の B が出てくるまでを１文とした。

の対象からは空白と補助記号を除いている。

４．分析１

４．１　概観

図 1～図 8[3]は、適宜選択した見出し語 8 語について、キーとなる語の前後 20 語について延べ語数と異なり語数の推移を示したものである。調査対象は BCCWJ 全体である。図 1 の「思う」の例では、延べ語数はキーのマイナス側は、－1 語まで増え続け、＋1 語以降は下降に転じる。この傾向は図 2 の「見る」、図 3「関係」でも同じである。一方、図 4「人間」、図 5「新しい」、図 6「すごい」、図 7「しかし」、図 8「なお」は、＋1 語まで延べ語数が増え続け、＋2 語以降は下降に転じる。延べ語数の推移は、キーから文頭ないし文末

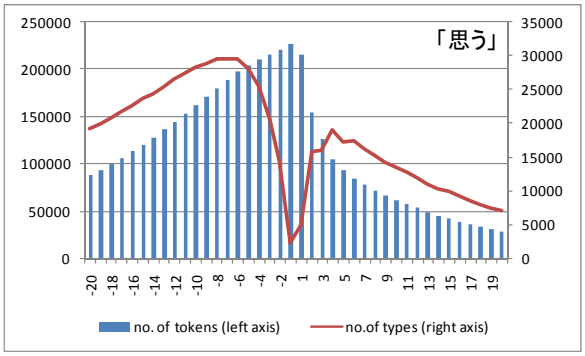


図 1　計量的指標の推移：「思う」

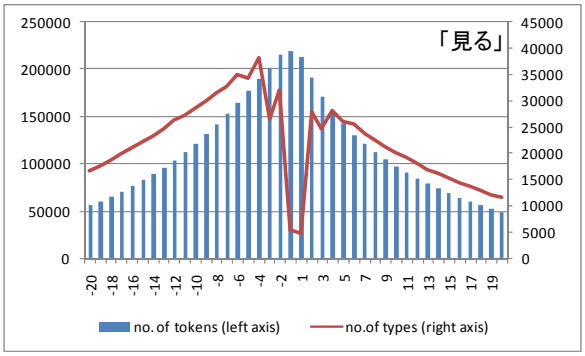


図 2　計量的指標の推移：「見る」

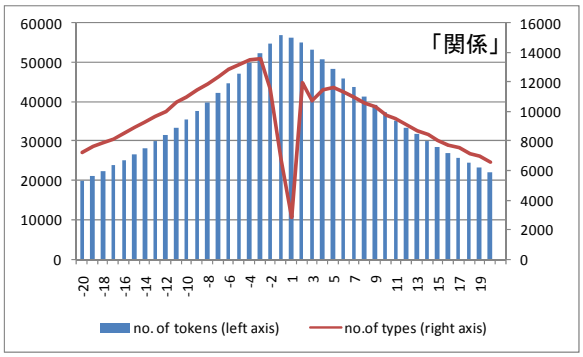


図 3　計量的指標の推移：「関係」

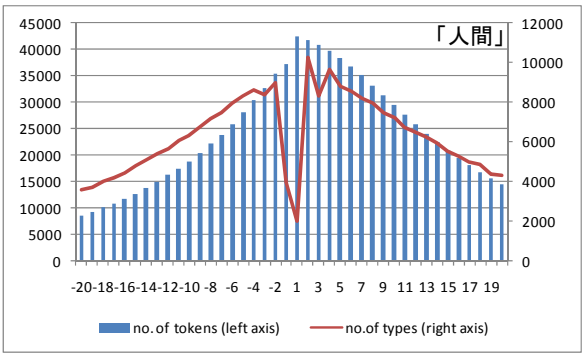


図 4　計量的指標の推移：「人間」

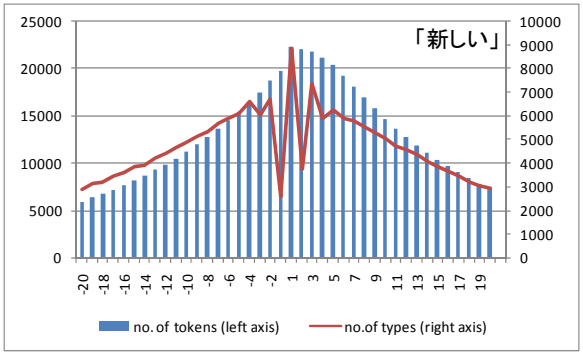


図 5　計量的指標の推移：「新しい」

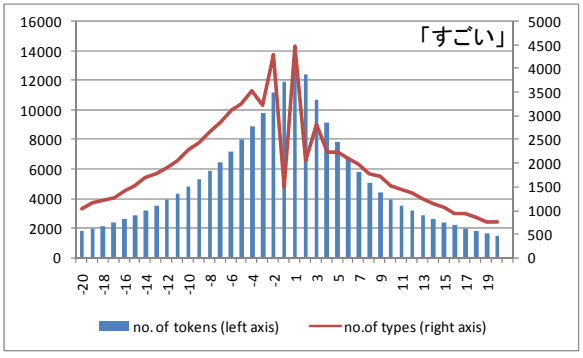


図 6　計量的指標の推移：「すごい」

[3] 図 1～8 のいずれも棒グラフが延べ語数（左軸）、折れ線グラフが異なり語数（右軸）を表す。横軸はキーからの相対的位置である。これは離散的な値をとるため、折れ線グラフにするのは妥当ではないが、見やすさのため、便宜的に使用した。以降のグラフも同様である。

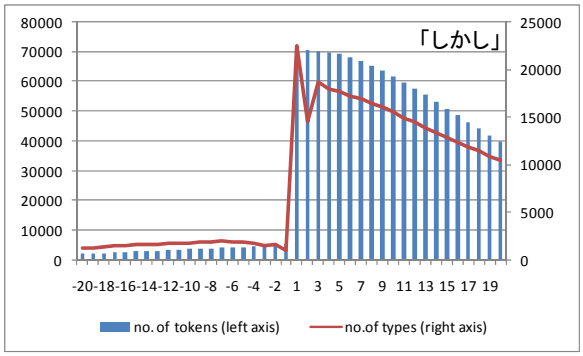


図 7　計量的指標の推移：「しかし」

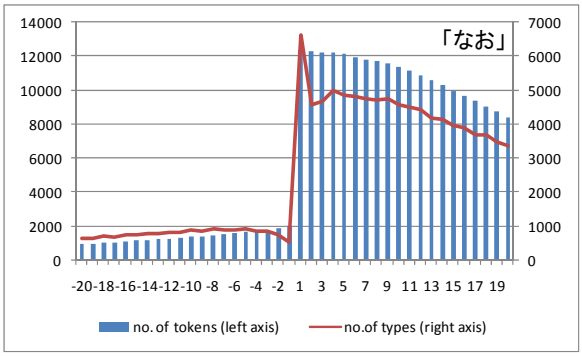


図 8　計量的指標の推移：「なお」

まで何語あるかの分布を意味していることから、その語が平均して文のどの辺に位置しているかを表していることになる。「思う」が「見る」に比べてプラス側の延べ語数の減少が大きいのは文末によく現れることを意味している。また、接続詞である「しかし」「なお」はマイナス側が非常に少ない非対称的な形をしているのもその品詞性の表れである。

図で折れ線で示した異なり語数の推移は、延べ語数の推移とは違って、やや複雑な様相を示している。図 1「思う」では、キーのマイナス側は、－7 語まで上昇し続け、－1 語まで下降し、＋4 語まで再上昇し、＋5 語で若干下降し、＋6 語で上昇、＋7 語以降は下降する。個別に異なる部分はあるものの、「思う」「見る」「関係」「人間」ではキー付近に谷ができる形の分布となっている。図 5「新しい」と図 6「すごい」は－1 語と＋2 語との 2 か所に谷ができる分布であり、図 7「しかし」と図 8「なお」はマイナス側は語数が少なく推移はほぼ一定しているようであるが、プラス側は＋1 語で急に上昇し、いったん小さな谷を作り、下降するという分布になっている。

大局的に見ると、異なり語数の推移は延べ語数の増減に伴う自然増・自然減となっていると見られる部分と、その傾向に反し、延べ語数が増えても減少する、あるいは、延べ語数が減っても増加する部分とに分けられる。前者は語彙の量的な特徴として一般的な現象と考えられるが、後者は当該のキーとなる語の持つ、コロケーションとしての特徴が現れているものと解釈できる。すなわち、キーとなる語の影響によって特定の語の出現が多くなったため、延べ語数の値と異なり語数の値の関係にも影響したものであろう。このようなコロケーションの影響を受けていると思われる部分をマイナス側からの自然増の傾向が破られる箇所（すなわち減少に転じた箇所）、同様にプラス側を値の大きい方から見た場合の自然増の傾向が破られる箇所で特定すると、「思う」が－6 語から＋5 語の範囲、「見る」が－3 語から＋3 語、「関係」が－2 語から＋4 語、「人間」が－3 語から＋3 語、「新しい」「すごい」が－3 語から＋4 語、「しかし」が＋2 語、「なお」が＋3 語となっている[4]。このプラス側の転移箇所、およびマイナス側の転移箇所にはさまれた部分をコロケーションの影響を受けている範囲と考えることができる。

図 9、図 10 は、「思う」について、レジスターごとに延べ語数と異なり語数の推移を見たものである。図 9 の延べ語数では、13 のレジスターのうち 12 個が－1 語目が最大になり、以降下降する傾向を取っている[5]。ちなみに－1 語目における延べ語数がいちばん多いのは、Yahoo!知恵袋(OC)で、以下図書館書籍(LB)、出版書籍(PB)、Yahoo!ブログ(OY)、国会会議録(OM)と続く。国会会議録のマイナス側のカーブは他のレジスターと比べてゆるやかであるが、これは一文が長いということの現れであろう。表 1 は、図 10 の異なり語数の推移について、法律(OL)を除くレジスターごとにキーとなる語に向かってプラス・マイナスそれ

4　接続詞についてはマイナス側は語数が少ないため、評価は行わない。

5　残りの一つのレジスターは法律(OL)で、「思う」が 5 回しか現れないため、傾向を把握することは難しい。

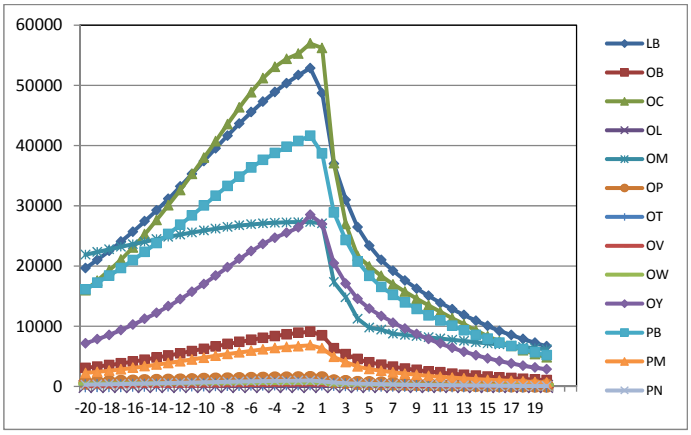


図 9　レジスター別延べ語数の推移「思う」

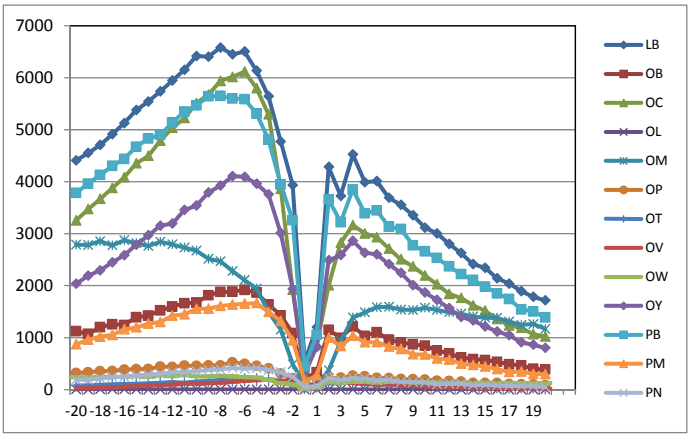


図 10　レジスター別異なり語数の推移「思う」

表 1　レジスターごとの転移箇所

レジスター	マイナス側の転移箇所	プラス側の転移箇所
図書館書籍(LB)	−7	+5
ベストセラー(OB)	−5	+3
Yahoo!知恵袋(OC)	−5	+3
法律(OL)	＊[6]	＊
国会会議録(OM)	−19	+18
広報紙(OP)	−10	+3
教科書(OT)	−11	+17
韻文(OV)	−10	+17
白書(OW)	−18	+16
Yahoo!ブログ(OY)	−6	+3
出版書籍(PB)	−6	+5
出版雑誌(PM)	−4	+17
出版新聞(PN)	−6	+14

6　法律(OL)は用例数が少ないため、転移箇所を判断できない。

ぞれの方向から自然増の傾向が破られる転移箇所を示したものである。BCCWJ 全体では前述のようにこの範囲は－6 語から＋5 語であったが、レジスターで見ると、ベストセラー(OB)、Yahoo!知恵袋(OC)はプラス方向にもマイナス方向にも範囲が狭くなっている。また、Yahoo!ブログ(OY)と広報紙(OP)はプラス方向のみ、出版雑誌(PN)はマイナス方向のみ範囲が狭くなっている。範囲が広くなった主な理由は異なり語数が少なく、値が安定していないためであろう。例えば、白書(OW)のマイナス側の数値の推移は、次のようになっている。「234、242、237、244、244、262、243、275、258、282、256、253、262、255、238、233、186、124、101、12」下線を施した 5 箇所が上昇から下降に転じた点である。

４．２　ＴＴＲによる観察

図 11～14 は $V_{m(d)}$の異なり語数をその延べ語数で割った値、Type/Token Ratio（以下、TTR とする）の推移を示したものである。TTR は語彙の豊かさを表す指標とされ、語彙の計量的な分析や文章の評価によく用いられている。TTR の値が高いほど集合における見出し語の種類が多く、語彙的に豊かであるとされる。本分析でのデータは、キーとなる語から等距離にある語を集めた集合であるため、文脈を有していない見出し語の集合という特徴がある。したがって、そのような集合における TTR の値が意味するものは、データ中に同一文脈がどれだけ複数回使用されているかということの観察になるだろう。

図 11～14 により、TTR の動きはキー付近に谷を形成することから、図 1～8 の異なり語数の推移にやや似ているが違う点もある。図 11 の動詞ではマイナス方向プラス方向ともに相対位置の絶対値が大きくなると TTR の値も高くなる傾向がある。これは図 12 の名詞、図 13 の形容詞でも同じである[7]。図 14 の接続詞ではマイナス方向には TTR が高くなる傾向があるが、プラス方向ではそれがなく、フラットになっているのが特徴的である。図 7、8 からプラス方向で延べ語数の減少が見られることから、延べ語数が一定のためこのようにフラットになったわけではない。キーから離れるにしたがって、TTR の値が大きくなって

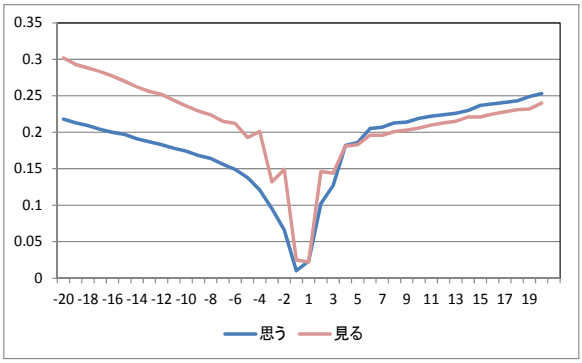


図 11　ＴＴＲの推移：動詞

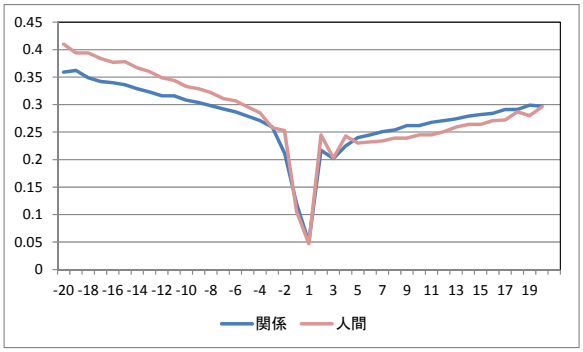


図 12　ＴＴＲの推移：名詞

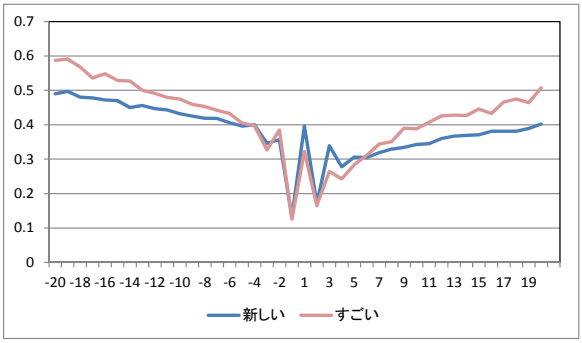


図 13　ＴＴＲの推移：形容詞

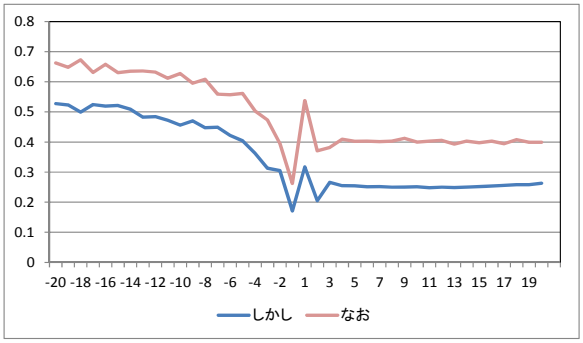


図 14　ＴＴＲの推移：接続詞

[7] 図 13 は谷が二つある型であるが、その外側は絶対値の大きい方向に対して単調増加の傾向が見て取れる。

いくということは、コロケーションの影響がどの辺まで届いているかの判断にも関係する。

TTR の値が一定になるまでコロケーションの影響があるとすると、少なくとも図 11～13 に挙げた語群についてはキーから前後 20 語までコロケーションの範囲ということになる。これは４．１節で述べた異なり語数の推移から見たコロケーションの範囲（ほぼ一ケタ台前半の値）とはずいぶん違っている。どちらがコロケーションの範囲として妥当かは本稿では決めがたいが、その検証方法のひとつとして、延べ語数を一定にしておいて TTR を測ることを次の課題としたい。

５．分析２

５．１　動詞の場合

この節では、対象を図書館書籍(LB)に、$V_{m(d)}$の距離の範囲を±５にした場合について考察をする。図 15、16 は動詞を対象にして TTR の値を観察したものである。図 15 は UniDic の品詞体系で「動詞－一般」を、図 16 は「動詞－非自立可能」を品詞に持つものである。いずれも似たような動きを示しているが、特徴的なのは、－１語と＋１語の TTR の値が低くなっており、その部分を谷として両側に開いた形を作ることである。また、動詞一般の「考える」「出る」「使う」「聞く」「書く」には－３語目に小さな谷が出来ている。図 16 の非自立可能の方でも「見る」「掛ける」「終わる」の－３語目に小さな谷ある。「始める」「続ける」「切る」は－３語目に谷はないが、－４語目、－５語目まで観察すると値が減少している箇所が認められる。また、－３語目ほど顕著ではないが、＋３語目にも TTR の値が鈍化する部分がある。図 15 では「使う」「聞く」、図 16 では「見る」「掛ける」「切る」である。TTR の値が単調に推移しない理由は、キーから±３語目によく出現する語があることを想定させる。この場合、キーの前後３語目までがコロケーションとして注目すべき範囲であると推測される。

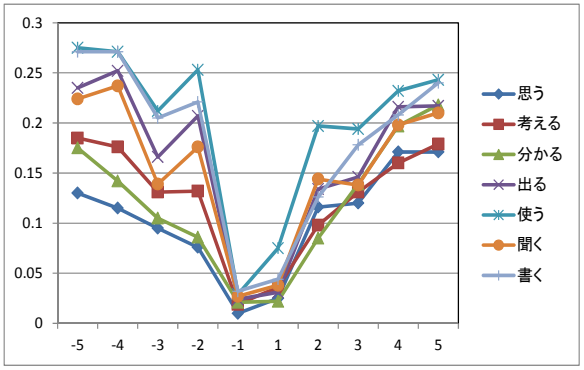


図 15　TTR の推移：動詞・一般

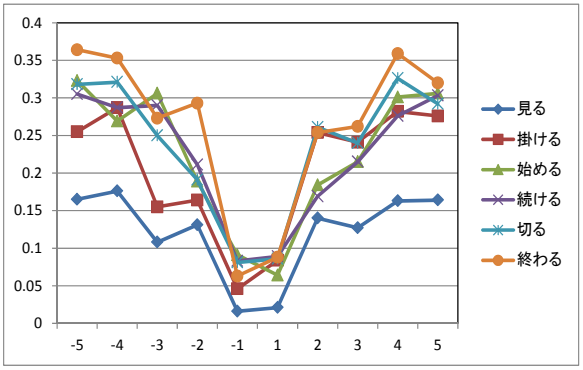


図 16　TTR の推移：動詞・非自立可能

図 15、16 からはキーを挟んで TTR の値が対称的になっているのではなく、全体的にキーの前の方が値が高いように見受けられる。特に図 15 のプラス側は折れ線が混み合っているのに対し、マイナス側はばらけている印象がある。このことを確かめるために、プラス側の TTR の値からマイナス側の TTR の値を引いた値を表 2、3 に示した。この値が 0 より小さければマイナス側の TTR の方が大きいということになる。表 2、3 の網掛けの部分はその差が 0 より小さい部分である。表 2 では 35 箇所中、20 箇所のセルが 0 より小さく、TTR の値に関しては対象でなく、マイナス側のほうが値が高いことが分かる。このことは、キーとなる語の前５語以内に現れる語彙のバラエティーの方が、後の５語以内に現れる語彙のバラエティーよりも多いことを意味する。ただし、表 3 ではその傾向は確認されず、網掛けの箇所は 30 箇所中 16 箇所にとどまるが、キーから１語目の部分を除くと若干傾向が高まる（24 箇所中 15 箇所）。

表 2　キーから等距離はなれた語集合の TTR の差：動詞・一般

キーからの距離	1 語	2 語	3 語	4 語	5 語
思う	0.015	0.04	0.025	0.056	0.041
考える	0.017	-0.034	0	-0.016	-0.006
分かる	0.001	-0.001	0.034	0.055	0.043
出る	0.007	-0.073	-0.02	-0.036	-0.018
使う	0.046	-0.056	-0.018	-0.039	-0.032
聞く	0.011	-0.032	-0.001	-0.039	-0.014
書く	0.012	-0.095	-0.027	-0.063	-0.031

表 3　キーから等距離はなれた語集合の TTR の差：動詞・非自立可能

キーからの距離	1 語	2 語	3 語	4 語	5 語
見る	0.005	0.009	0.019	-0.013	-0.001
掛ける	0.038	0.09	0.086	-0.005	0.021
始める	-0.027	-0.005	-0.091	0.032	-0.017
続ける	0.006	-0.042	-0.075	-0.011	-0.002
切る	0.005	0.07	-0.01	0.005	-0.026
終わる	0.025	-0.039	-0.011	0.006	-0.044

５．２　名詞の場合

名詞における TTR の分布を見てみよう。図 17 は普通名詞、図 18 は固有名詞及び普通名詞だが助数詞としても使うもの（時間、パーセント）を選んだ。名詞は動詞と違い、TTR の谷に相当する部分が 1 例を除いては 1 箇所（＋1 語目）である。この違いは、それぞれ動詞、名詞の前後にくる助詞助動詞の影響ではないかと思われる。前節で述べたように、－3 語目、＋3 語目に TTR　の値が鈍化する箇所があることも同様である。

次に、意味的な違いは TTR の推移にどのように関係しているかを見てみよう。図 17 から「女」と「男」の TTR の値がほぼ重なるくらいによく一致していることが分かる。意味的な類似性のためとも解釈できるが、対比する意味で挙げた「人間」と「子供」もその分布はかなり似ているため、必ずしも意味的な類似が理由とは言い切れないようである。図 18 の「日本」と「アメリカ」は値の大きさは異なるが値の推移の様子は似ている。助数詞にもなる「時間」「パーセント」は谷の位置がずれており、推移が似ているとは言い難い。ちなみに動詞の場合と同じように、キーから等距離にある TTR の値をプラス側からマイナス側を引いた値は、図 17 の 4 語では、20 箇所中 18 箇所が、図 18 では 20 箇所中 13 箇所がマイナスの値をそれぞれ示した。名詞においてもキーの前の語彙の種類の方が、後ろに

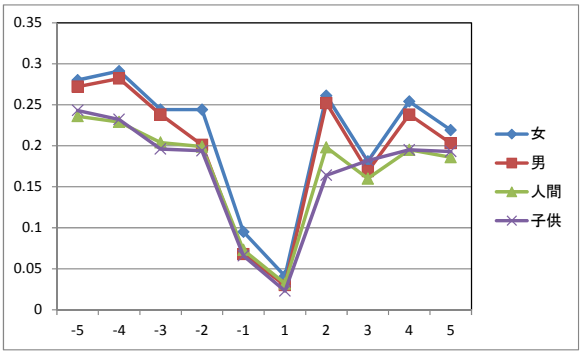


図 17　TTR の推移：普通名詞

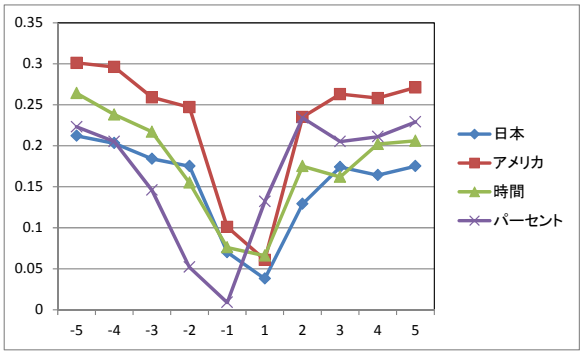


図 18　TTR の推移：固有名詞、助数詞可能

くる語彙の種類よりも多い傾向があることが確認された。

５．３　形容詞の場合

形容詞は図 19 の活用の場合と図 20 のシク活用の場合とに分けた。１例を除いて－１語目に谷を作る分布を示している。動詞、名詞の場合とはやや異なり、TTR の値が鈍化する箇所がマイナス側は－３語目であるが、プラス側が＋２語目と＋４語目の２箇所あるのが特徴的である。図 19 では意味的に関連の深い「良い」「悪い」と「大きい」「小さい」の値の推移がそれぞれ類似していることが分かる。図 20 では「嬉しい」の谷の位置が＋１語目にずれているが、今のところこれを説明する解釈は持ち合わせていない。プラス側とマイナス側の値の差は、図 19 で 20 箇所中 12 箇所が、図 20 で 25 箇所中 16 箇所が 0 より小さく、名詞、動詞と類似の傾向を示すことが確認された。

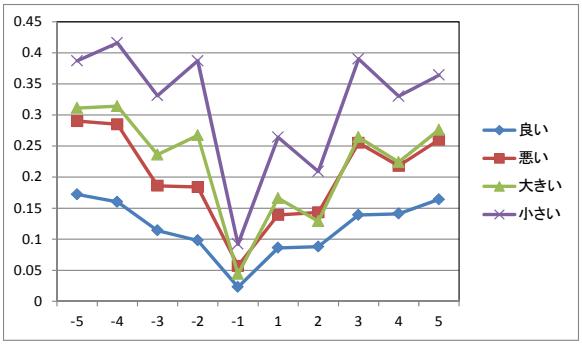


図 19　TTR の推移：形容詞ク活用

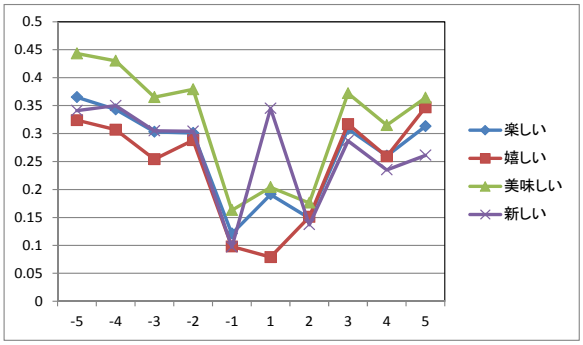


図 20　TTR の推移：形容詞（シク活用）

６．まとめと今後の課題

本稿では、共起語集合（キーとなる語の前あるいは後ろの特定の位置に出現する語の集合）という考えを用いて、BCCWJ においてコロケーションが現れる様子を計量的な指標の観察から記述した。用いた指標は、延べ語数、異なり語数、TTR である。得られた知見をまとめると次の 3 点になる。(1)異なり語数の推移からは異なり語数が自然増ではなくなる範囲をコロケーションとして位置付け、「思う」「見る」などの語について個別の記述を行った。(2)TTR の推移については±20 語目でも値が一定しないことから TTR によりコロケーションの範囲を定めるのは、別の工夫が必要であることが示唆された。(3)図書館書籍（LB）に限ってキーの前後 5 語の TTR の動きを観察した場合、動詞、名詞、形容詞それぞれ特徴的な推移があること、また、キーからマイナス側の方がプラス側よりも値が高い傾向にあることが分かった。

今後の課題としては、調査語の範囲を広げること、共起語集合同士の類似度を用いた分析、特に、類似度を用いた語の分類を試みたい。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

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）

BCCWJ 係り受け関係アノテーション付与のための文境界再認定

小西 光（国立国語研究所コーパス開発センター）†
小山田 由紀（国立国語研究所コーパス開発センター）
浅原 正幸（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）
前川 喜久雄（国立国語研究所言語資源研究系／コーパス開発センター）

Revision of Sentence Boundaries in BCCWJ for Syntactic Dependency Structure Annotation

Hikari Konishi (Center for Corpus Development, NINJAL)
Yuki Oyamada (Center for Corpus Development, NINJAL)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept.Corpus Studies, NINJAL)
Kikuo Maekawa (Dept.Corpus Studies/Center for Corpus Development, NINJAL)

１．はじめに

『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）では，文を機械的に自動認定し，コアデータのみ人手による文境界の修正を行っている．このコアデータの文境界情報を元に係り受け関係アノテーションを付与しようとすると，「係り先のない文」が出現する．これは，文書を電子化した際に認定したレイアウトに基づく階層構造[1]の影響であったり，係り受け関係にあると判断されるものが自動文認定の際に一文とされなかったりしたことに由来する．そこで，BCCWJ のコアデータに関して「係り受け関係アノテーション付与」を目的とした「文」の再認定を行うこととした．

２．自動文認定—sentence 要素—

sentence 要素は，「文に相当するまとまりを表す要素」として機械的に自動認定されている．BCCWJ の電子化フォーマットでは，自動認定は以下のように行われる．

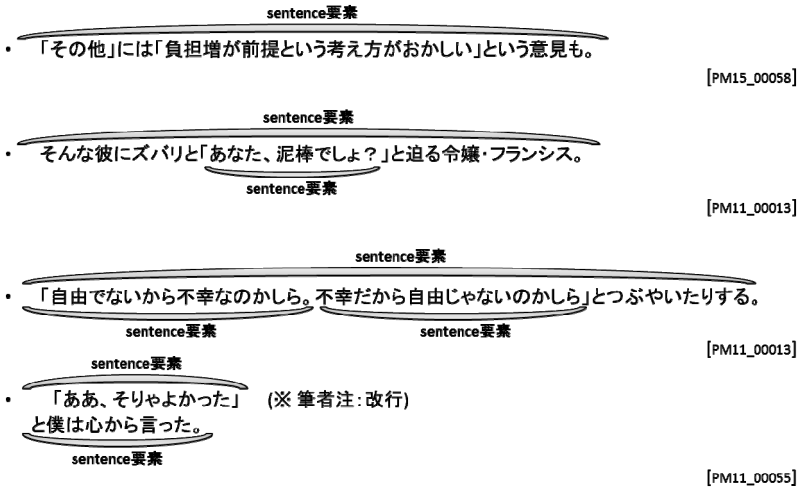


図 1 sentence 要素の自動認定

† hkonishi@ninjal.ac.jp

1 山口ほか(2011)によると上位から article> cluster/titleBlock>paragraph>sentence という階層構造を持つ

自動認定により sentence 要素冒頭と判断された箇所には XML 形式で<sentence>タグを，末尾と判断された箇所には</sentence>タグを挿入する．

現在公開されている BCCWJ（DVD 版）の電子化フォーマットのうち，XML で構成されるものとして文字ベース XML(C-XML)と形態論情報付きの統合形式 XML（M-XML）の二種類がある．C-XML は sentence 要素の入れ子構造を認めるが， M-XML では sentence 要素の入れ子構造を認めず，C-XML で入れ子構造の外側の sentence 要素を<superSentence>としている[2]（図２・上）．また C-XML では，<sentence>タグに属性“quasi”（文区切り文字以外の基準により自動付与された sentence 要素）と“verse”（韻文内の sentence 要素）を付与しており，M-XML では，その二属性に加えて入れ子構造外側の sentence 要素に対して“fragment”という属性を新たに導入している．

M-XML

```
<superSentence>
  <sentence type="fragment">　声明は、同大統領の法案署名へ歓迎と感謝を表明し</sentence>
   <quote>
     <sentence>「米国の支持は、台湾の（ＷＨＯ参加への）努力が既に友邦の理解を得たことを意味する。</sentence>
      <sentence type="quasi">今後、全力を挙げて国際社会の全面的な賛同を得られるよう努力する」</sentence>
   </quote>
  <sentence type="fragment">と述べている。</sentence>
</superSentence>
<br type="automatic_original"/>
```

C-XML

```
<paragraph>
  <sentence>　声明は、同大統領の法案署名へ歓迎と感謝を表明し
    <quote>
       「<sentence>米国の支持は、台湾の（ＷＨＯ参加への）努力が既に友邦の理解を得たことを意味する。</sentence>
        <sentence type="quasi">今後、全力を挙げて国際社会の全面的な賛同を得られるよう努力する</sentence>
       」
    </quote>
       と述べている。
  </sentence>
<br type="automatic_original" />
</paragraph>
```

図２ M-XML(上)と C-XML(下)の比較（PN4g_00001）

２.１　sentence 要素の現状と問題点

係り受け関係アノテーション付与を目的とした場合，BCCWJ の sentence 要素における問題点は，大きく以下の二つに分けられる．

① 文境界と判断されるべき箇所に<sentence>タグが付与されていない．
② 文境界と判断されるべきでない箇所に<sentence>タグが付与されている．

まず①について，前述のとおり<sentence>タグはほぼ自動付与である．そのため，現段階で本来複数の文とされるべき発話や引用・補足などが認定基準が原因でひとつの sentence 要素となっており，分割されていない場合がある．例えば「<sentence>「受験勉強に明け暮れて、東大に入って、官僚になってもちっとも幸福じゃない」最近では、そんなセリフを大人も子供も口にします。</sentence>」(PB33_00032)のように自動認定の基準から外れており，機能的には括弧・引用符と同様に用いられているのだが sentence 要素とされていないような場合である．ただし，これらは人手修正のチェック漏れということも考えられる．

2 小木曽ほか(2011)「上位の文は superSentence として文書構造タグの一種とした。下位の sentence はそのまま残し、superSentence の一部分を新たに sentence で囲み type=”fragment”とした」(p.39)

次に②であるが，これには二つの問題がからんでいる．一つは資料原本のレイアウト情報を元に認定した文書階層構造により発生している問題であり，もう一つは自動及び人手修正による sentence 要素の認定基準と，係り受け関係アノテーションを目的とした「文」の認定基準とが異なるという問題である．

一つ目の問題は，例えば原本が図３のように会話文に入る直前の地の文で改行（<br type=”automatic_original”/>）されるようなレイアウトだと，sentence 要素より上位の文書構造である<paragraph>タグ[3]や<quotation>タグ[4]に阻まれ，文が続いているにもかかわらず一つの文としては認定されていない．

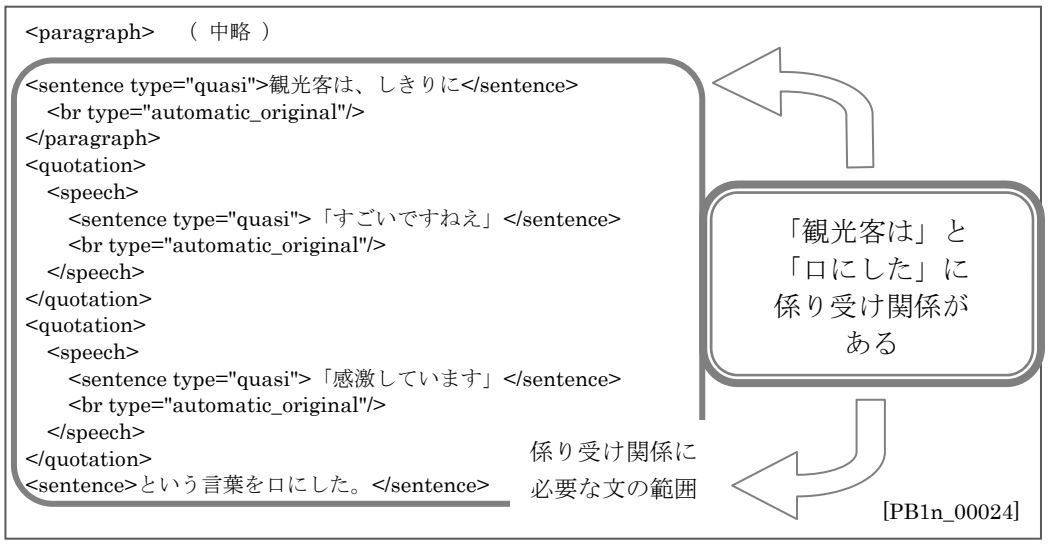


図３ 階層構造により複数の sentence 要素となった例(M-XML)

図３では「観光客は、しきりに」という sentence 要素と「という言葉を口にした。」という sentence 要素が認定されており，「観光客は」の本来の係り先「（口に）した」は「文」を越えて存在することになる．文を越えた係り受け関係は付与できないため，「観光客は」の係り先は不明となり，正しい係り受け関係アノテーションの付与ができないこととなる．

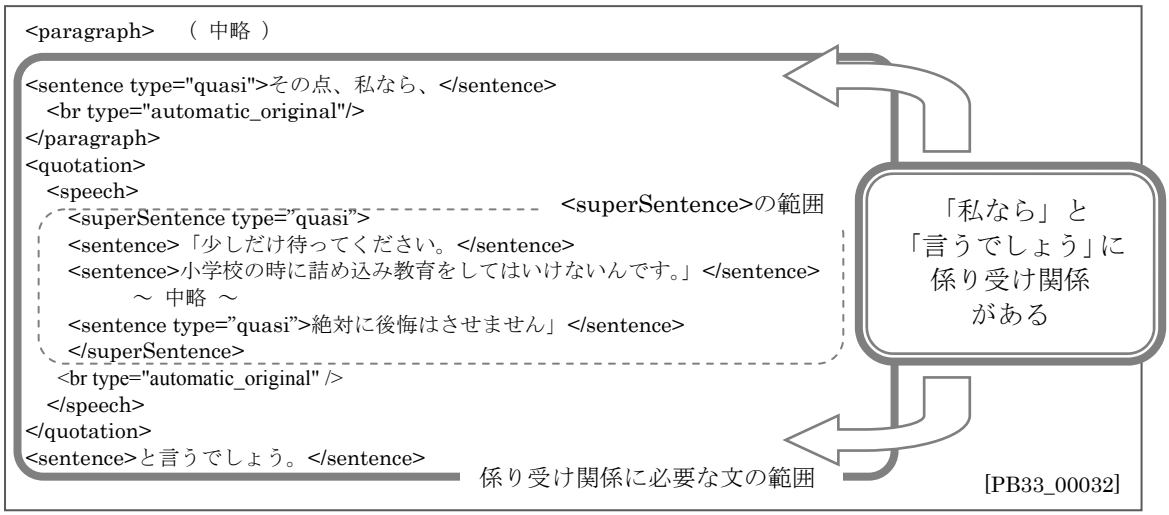


図４ <superSentence>タグの範囲（M-XML）

[3] 段落を表す文書構造要素．原則として，一字下げで始まる．sentence 要素よりも上位階層となり，sentence 要素が上位階層の要素をまたぐことはない．

[4] 山口ほか(2011)「当該記事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・書き起こしを表す」

二つ目の問題は，今回作業対象としたコアデータについては人手による修正が行われているものの，それらは「係り受け関係を付与する」という基準で作業されていないため，係り受け関係アノテーション付与を目的とした文認定が再度必要となる．M-XML に付与された<superSentence>タグを利用した文の認定も可能だが，図４のような過不足のある範囲となっている場合もあるため，自動的に抽出することは難しい．

またこれとは別に，C-XML から M-XML に変換する際に図５のような問題も生じている．

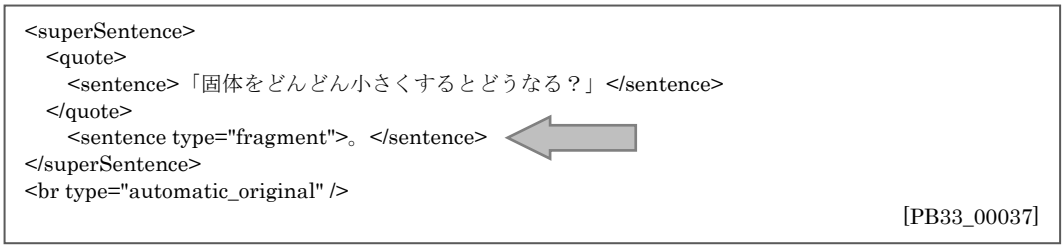


図５　句点「。」のみで 1 文と認定されている例（M-XML）

以上のことからも，係り受け関係アノテーションを付与する場合，現状の「文」境界の認定では問題があるため，今回は係り受け関係アノテーション付与に影響の大きい②の「文境界と判断されるべきでない箇所に<sentence>タグが付与されている」sentence 要素についてのみ文境界の再認定作業を行った．なお浅原(2013)によると，①は係り受け関係アノテーション作業時に「文境界」を表現する係り受け関係ラベル（“Z”ラベル）を導入している．

３．文境界再認定作業

３.１　作業対象

BCCWJ のコアデータ全 60,374 文を対象とする．XML データの<sentence>タグや<superSentence>タグを修正するのではなく，XML データの sentence 要素を参考にした係り受け関係アノテーション用の文を別途認定する．

３.２　認定基準

まず前提として以下の二点を示す．

- 係り受け関係アノテーション付与を目的としたもっとも長い単位としての「文」の認定を行う
- 現在 XML 等に付与されている改行やタグ情報（<sentence>タグ・<paragraph>タグ等）には縛られない

３.２.１　一文と認定するもの

現状の sentence 要素では係り受け関係アノテーション付与に問題があり，以下の三点のいずれかを満たすものを「文」と再認定する．

① 括弧や引用符などの括り記号で括られた発話や引用・補足部分を挟んだり，引用の助詞「と」で受けたりして係り受け関係を結べる要素が前・中・後に接続する
② 箇条書き（改行を伴う）を内包する要素が前・中・後に接続する（主にウェブ媒体）
③ 本来一文であるべきものが，書き手による意図的な改行で分割されている（主にウェブ媒体）

３.２に記したとおり係り受け関係がもっとも長い単位としての文境界越えないことを基準とするが，例えば「「掛け給え」
と部長は言った。」や「手でひたいをおさえて、

「なにをいっているんだ、わたしは？」」のように括られた要素に対して後ろのみ，前のみに接続する場合がある．この場合は「「掛け給え」と部長は言った。」「手でひたいをおさえて、「なにをいっているんだ、わたしは？」」という文を認定した．接続詞のみの場合や助詞「と」だけで括られた要素を受ける場合も同様に処理する．

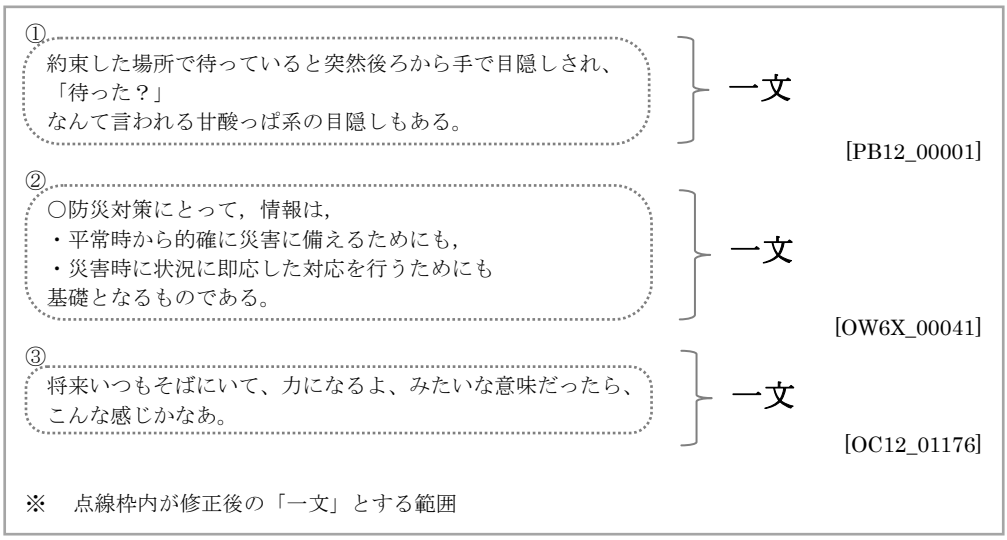


図６　一文と認定するもの

３.２.２　一文と認定しないもの

以下の場合は，現状のままひとつの文にまとめ上げることはしない．

① 倒置部分が改行されている
② 改行を伴って文がねじれている
③ 接続助詞ではなく接続詞「と」「っと」と判断されるものが文頭にくる
④ 前後の sentence 要素と括弧や引用符などで括られた要素がそれぞれ独立して係り受け関係にない

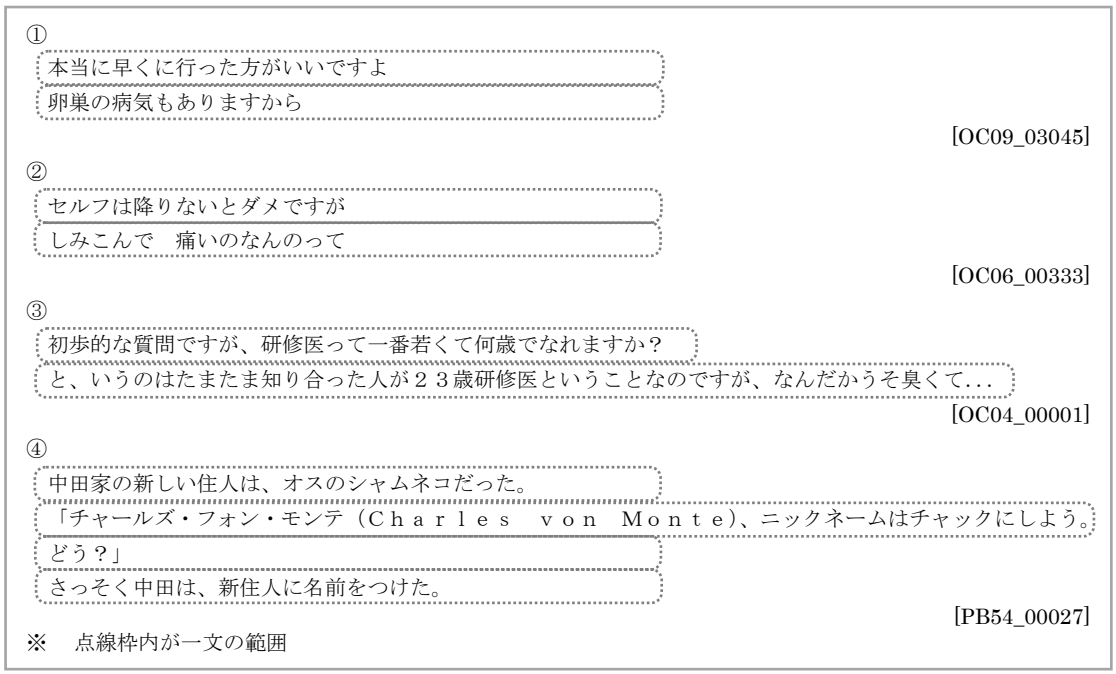


図７　一文と認定しないもの

３.３　作業手順

M-XML を元に sentence 要素, <superSentence>タグの範囲情報, <sentence>/<superSentence>タグの type 情報（fragment，quasi，verse），sentence 要素の冒頭と末尾の品詞情報を抽出し，それらを参考にして作業を行った．

例えば，sentence 要素が助詞や括弧閉，読点・カンマで始まっている場合は，一つ前（もしくはそれ以前から）の sentence 要素を受けていると考えられる．また sentence 要素が読点やカンマ，助詞，括弧開で終わっている場合は，「文」の途中で分割されており，係り先となるはずの文要素がそれ以降に後続していると考えられる．このように手がかりを見つけ次第，その前後の sentence 要素を確認して「文」の範囲を認定した．

４．作業結果

４.１　再認定結果

表１に作業結果をまとめた．

コアデータ全文 60,374 文に対する修正箇所の割合は，4,585 文と約 7.6%である．この 4,585 文を係り受け関係アノテーション用の「文」に修正すると 1,385 文となる．これは修正前の約三文が一つの文にまとまるという割合になる（まとめ上げ「文」数）．

修正箇所全体の約 66%は，「私は「～。」と言った。」のように括弧や引用符等で括られた要素を前と後ろに挟んで係り受け関係を結べるもの（「前後型」とする）である．また約 25%は，「“～。”と言った。」のように括られた要素を後ろのみで受けるもの（「後型」とする）である（表２[5]）．

レジスター別では，Yahoo!ブログ（OY）と新聞（PN）の修正する割合が高い．各レジスターのまとめあげ「文」数を見てみると，それぞれの特徴の一端を示している．

表１　文境界再認定の結果

レジスター	<sentence>タグ数＝「文」数	修正箇所「文」数（チェック率）＝A	修正後「文」数（修正対象のみ）＝B	まとめ上げ「文」数（A/B）
OW（白書）	6,067	385(6.3%)	113	3.41
PB（出版書籍）	10,095	592(5.9%)	166	3.57
PN（新聞）	17,136	**1,530(8.9%)**	415	**3.69**
OC（Y!知恵袋）	6,435	514(8.0%)	161	3.19
OY（Y!ブログ）	7,651	**915(12.0%)**	332	**2.76**
PM（雑誌）	12,990	649(5.0%)	198	3.28
合計	**60,374**	**4,585(7.6%)**	**1,385**	**3.31**

括弧内が複数文

文途中での改行頻度が高い

接続詞・接続助詞で改行

新聞は，約 3.69 文が一つの文にまとめ上げられているのに対し，Yahoo!ブログは約 2.76 文が一つの文にまとめ上げられている．これは，新聞が括り記号内に文区切り文字で区切られる sentence 要素が複数含まれ，再認定作業後の「文」が長文化するのに対し，Yahoo!ブログは，図８のようにブログ執筆者によって接続詞や接続助詞の後ろなど文の途中で改行されている例が修正箇所全体の約三分の一（100 例）と多くを占め，それら

表２　修正箇所の構造（単位：文）

レジスター	前後型	後型	合計
OW	74	10	84
PB	85	47	132
PN	**254**	**75**	329
OC	83	36	119
OY	80	31	111
PM	96	**54**	150
合計	672	253	**925**

[5] 前後型・後型以外にも「前中後型」「前中型」「中後型」「前型」があるが，それらは数が少ないためここでは省いた．

文の断片を結びつけるための単純な再認定である場合が多い.

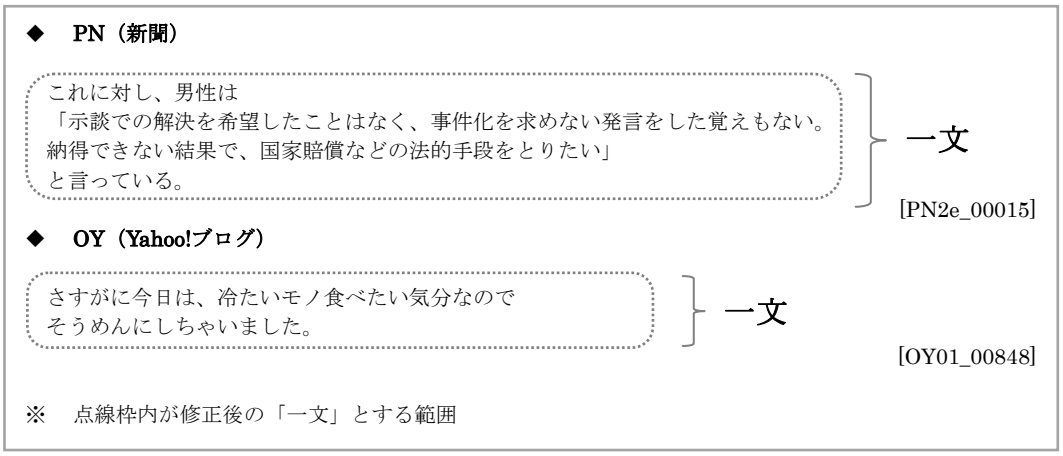


図８　新聞と Yahoo!ブログの比較

４.２　括弧・引用符等の機能別分類

４.１の作業結果をもとに括弧・引用符等の機能別に以下の分類を試みたので，レジスター別の特徴を示す.

- 補足 ： 語や文を補う目的で用いる（主に（ ））
  補足部分がなくても文が成立する
- 発話 ： 「と言う」等で受ける発話
- 心内 ： 「と思う」等で受ける心内語
- 引用 ： 上記以外のもの
- 箇条書き ： 行頭の中点等記号および改行によって複数の項目を列挙したもの
- 強調 ： 主に括弧を用いて他の文字列よりも強調するために用いる
  （書籍名やタイトル等も含む）

表３ 機能別分類（単位：文）

レジスター	補足	発話	心内	引用	箇条書き	強調	合計
OW	40	4	0	23	15	2	84
PB	1	107	4	26	7	1	146
PN	0	307	4	42	0	5	358
OC	12	26	4	62	23	5	132
OY	16	32	24	46	2	8	128
PM	0	108	9	46	3	6	168
合計	69	585	45	240	50	27	1016

表３を見ると，レジスターごとに特徴が表れている.

白書(OW)は，丸括弧による補足と箇条書きが多用されている.

新聞(PN)は，括られた要素の前後で係り受け関係を結べるような発話が多用される．これは，文頭に発話者の情報や状況が来て，続いて引用部分を挟み，引用の「と」等でそれを

受けて係り受け関係を結ぶというある種の「文型」が決まっていると考えられる[6]．また書籍(PB)や雑誌(PM)でも発話が多用されている．

Yahoo!知恵袋(OC)は，Q&A 形式の特徴（答える際に文献からの引用や列挙を用いる）が引用や箇条書きの多用に表れている．

Yahoo!ブログ(OY)は，他のレジスターより心内語が多用され，ブログ執筆者の心情を表わす傾向をとらえている．

５．まとめ

係り受け関係アノテーション付与を目的とした文境界の再認定作業について報告を行った．修正を必要とする 4,585 文（7.6%）のみではあるが，各媒体の特徴の一部が明らかになった．またランダムサンプリングではないデータではあるが，文を単位とした括弧・引用符等の機能別でのアノテーションもレジスター分析に有効な指標を設定するための予備調査に位置づけることができるだろう．

今回の報告により文を自動で認定する困難さが具体的なものとなり，また文分析の可能性の一端を示すことができた．今後はより精度の高い自動文認定解析の確立を待ちつつ，係り受け関係アノテーション付与の研究に着目していきたい．

謝　辞

本研究を行うにあたり，助言いただきました丸山岳彦氏に感謝いたします．また本研究は，国立国語研究所基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国立国語研究所「超大規模コーパス構築プロジェクト」による補助を得ています。

文　献

小木曽智信、間淵洋子、前川喜久雄 (2011)「階層的形態論情報を考慮した『現代日本語書き言葉均衡コーパス』の公開用 XML フォーマット」『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.35-42，JC-G-11-01

山口昌也、高田智和、北村雅則、間淵洋子、大島一、小林正行、西部みちる(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『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2」，JC-D-10-04

浅原正幸 (2013)「係り受けアノテーション基準の比較」本予稿集

6 会話×前後型で 375 例あった．これが全体に占める割合は 36.9%である．

書きことばにおける「語りかけ」は何のために用いられるのか

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）

Why Addressing Expressions are Used in Written Text?

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

書籍テキストの中には，著者が読み手に対して直接語りかけていると解釈できる文体がある（柏野, 2010 など）．たとえば，直感的には「あなた」「みなさん」などのような呼びかけ表現や「ではないでしょうか」「だよね」といった，問いかけもしくは相づちを求める文末表現などを含むテキストがそれにあたる．これらはいわゆるハウツー系の書籍に見られやすい傾向があるが，この場合特定の表現の出現頻度がとりたてて高いとも限らない（保田ほか, 2012b など）．本稿は，これらのテキストを「語りかけ性」があると呼ぶ．

『現代日本語書き言葉均衡コーパス』（BCCWJ）に収録されている図書館サブコーパスの書籍サンプル（全 10,551 サンプル・28,892,944 語）に，文書分類の観点から人手で情報を付与する作業を実施した（柏野・奥村, 2012）．付与した観点の一つにこの「語りかけ性」（とてもある・どちらかといえば・特にない：3 段階）がある．この作業結果から，「語りかけ性」は，話しことば的なテキストから受け取られるというのでもないことが明らかになった（保田ほか, 2012a）．書きことばであっても話しことば的であると判断されるテキストには，リアルタイム性と関わるフィラーや言いよどみ，音声的変化に関わる融合などが現れているが，「語りかけ性」があるとされるテキストにはその種の特徴は現れにくいのである．安藤（2012）は，小説における再現的提示の手法とは，二人称的世界が顕在しないことであるとし，読み手に語りかける言文一致の形がありえたならば，「言」に近い文体が創出されたかもしれないと述べる．すなわち「語りかけ性」は，既存の「言文一致」の範疇にはない表現ということになるのだろう．「語りかけ性」のあるテキストとは，書きことばの形式を保持しながら，疑似的に対話を導入しているテキストであると考えられる．

以下の例は，「語りかけ性」が「とてもある」と判断されたテキストである．特徴的と考えられる表現（保田ほか, 2012b）に下線を施した．

> 1）ここに六〇メートル×六〇メートル×六〇メートルというまったくの箱型の巨船の姿が浮かんでくるではありませんか。「ギルガメシュ叙事詩」は今から四〇〇〇年以上前のものと考えられますから、まさに私たちは単位という糸を伝わって、一気に人類の文化の源までさかのぼる感じです。つまり、人類の文明の発祥とともに、単位は存在していたわけで、文明にとっては単位は切っても切り離せないものだったということがわかります。ということは、単位を考えることで文明そのものを考えていく糸口もつかめるかもしれない、という期待を抱かせます。ま、この点は、あまり気負わずに、ボチボチと本書の中でも試みてみることにしましょう。（高木仁三郎「単位の小事典」）

[†] yasuda_s@ninjal.ac.jp

書きことばでありながら，疑似的な対話形式が用いられているということは，著者が語り手としてテキストから現れているスタイルだとも言えよう．著者が前面に出現しているとすれば，「語りかけ性」のあるテキストが，同時に著者の「主観」が語られるテキストと認識される可能性が期待される．「語りかけ性」のあるテキストが「主観的」であると判断されるのならば，「語りかけ性」は著者の「主観」を語るために用いられているということになる．そこで，本稿は「小説以外の文章の内容」（とても客観的・どちらかといえば客観的・どちらかといえば主観的・とても主観的：4 段階）の観点についての情報付与作業結果を用いることで，「主観的」であることと「語りかけ性」との相関があるのかを調べる．「語りかけ性」が「主観的」と判断されることと関わりがないならば，いったい「語りかけ性」が何のために用いられているのかを考察したい．

２．データ

本稿は，文書分類結果を用い，「語りかけ性」があると判断されたテキストと，「主観的」であると判断されたテキストが，どのように関わっているのかを確かめた．

BCCWJの図書館サブコーパスに含まれる書籍の 10,551 サンプルをランダムに並べ替え，6 人の作業者が文書分類を行った結果を用いた．調査にあたっては，作業結果から約半数をランダムに選び（5,652 サンプル[1]），小説には会話文を含む場合が多いため，小説を全て除いたサンプル（3,750 サンプル・11,630,970 語）を調査対象データとした．

作業者は判断に際し，その根拠等に関するコメントを適宜記述しており，個人によって量は異なるが，それぞれの作業サンプル数の 2%～5%のコメントが得られている．

「語りかけ性」についてのアノテーションは，作業者が「とても（語りかけ性が）ある」「どちらかといえば（語りかけ性が）ある」「とくに（語りかけ性は）ない」の 3 種類の選択肢から該当すると判断した一つを選択する．作業の結果，「とてもある」は 486 サンプル（1,387,665 語・本稿で扱うサンプルの 13.0%），「どちらかといえばある」が 805 サンプル（2,347,671 語・同 21.5%），「とくにない」が 2,459 サンプル（7,895,634 語・同 65.5%）得られた．同様に，「小説以外の文章の内容」についてのアノテーションは，「とても客観的」「どちらかといえば客観的」「どちらかといえば主観的」「とても主観的」の 4 種類の選択肢から一つを選択する．結果，「とても客観的」は 704 サンプル（2,313,220 語・本稿で扱うサンプルの 18.8%），「どちらかといえば客観的」が 1,485 サンプル（4,741,194 語・同 39.6%），「どちらかといえば主観的」が 1,014 サンプル（3,160,066 語・同 27.0%），「とても主観的」が 547 サンプル（1,416,490 語・同 14.6%）得られた．

サンプルの形態素解析には，MeCab 0.993＋UniDic2.1.0 を用いた．分析結果に示す品詞情報や語彙素等の要素は，解析結果に基づく．

３．結果：「語りかけ性」と「主観的」「客観的」の判断

情報付与作業結果を分析したところ，「語りかけ性」のあるテキストが，主観的であるとは受け取られるのでもないことが明らかになった．

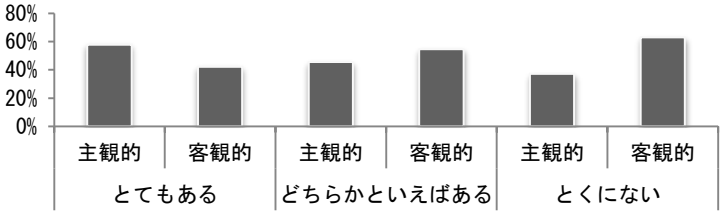


図１　「語りかけ性」の有無と「主観的」「客観的」分類

[1]対談，座談会をはじめ，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，分類対象外（非対象）とされ，本サンプル数には含まない．アノテーション作業者は，分類対象としたサンプルのみ観点付与を行っている．

図１の通り，「語りかけ性」の有無と「小説以外の文章の内容（主観的～客観的)」の判断に関係があるとは言えない.

次に，「語りかけ性」と「小説以外の文章の内容（主観的～客観的)」それぞれの分類群の NDC 分布（図２）・C-code 分布（図３）をあわせて見ておく．いずれかの群で，類似した分布が見られているということはない.

「語りかけ性」は，NDC3・4 番台（社会科学・自然科学）や C-code の「専門」・「実用」の分野で多く用いられている傾向があり，この傾向は同様に「客観的」なテキストに見られている．また，「客観的」～「主観的」の判断においては，NDC の 7 番台（芸術・美術）と 9 番台（文学）が「主観的」と判断されるに従って増加することや，C-code で「主観的」と判断されるのがほぼ「一般」向けであることなどが顕著な特徴と言える．なお， NDC9 番台については，「語りかけ性」が「ない」と判断されるに従って増加する傾向が見られ，「主観的」で「ある」ことと同傾向でもある.

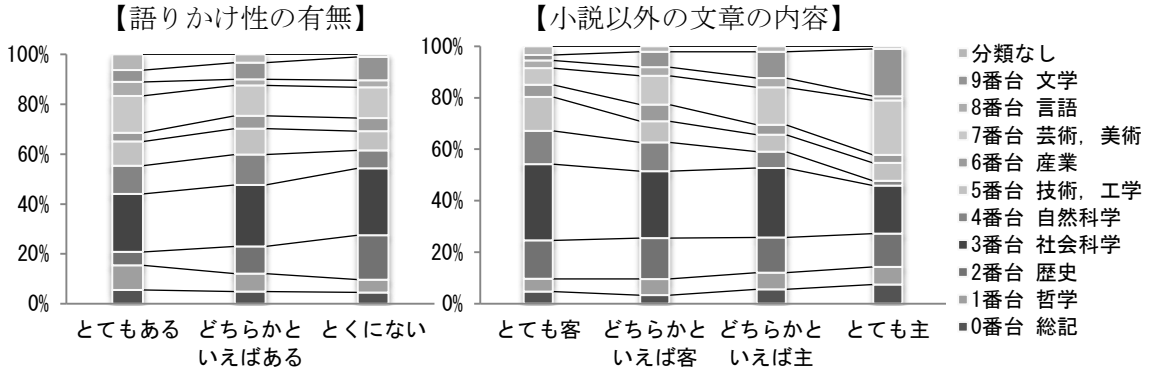


図２　分類群別 NDC 分布

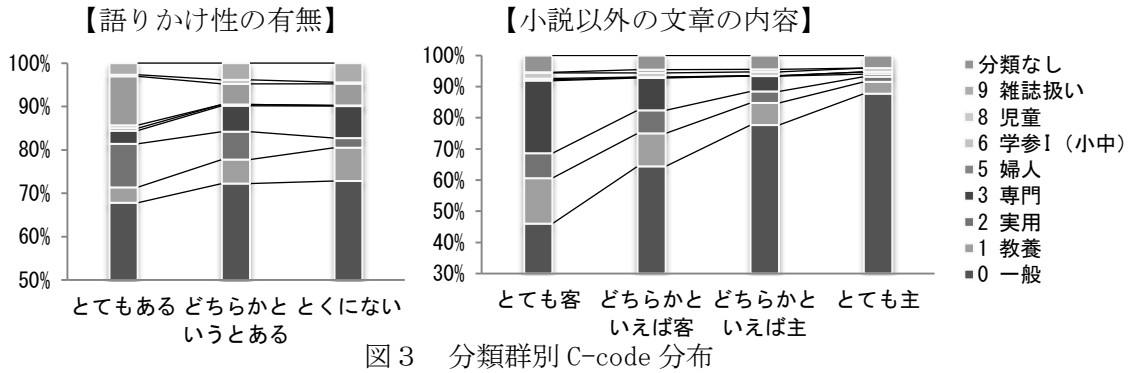


図３　分類群別 C-code 分布

アノテーターコメントの分析により，「内容が哲学・体験談・手記・自伝」であるため「主観的」と判断した記述（内容に重きを置く作業者）や，「内容は主観的であるが表現は客観的」であるとして「客観的」と判断した記述（表現を重視する作業者)，「主観的な意見もあるが，客観的な説明の量が多い」として「客観的」と判断した記述（分量を考える作業者）などのように，個々人で判断基準に差異のあることが推測された．そのため，作業者によって判断に差が生じる場合がある．そして，とくに「語りかけ性」があるとの判断が作業者間で一致した場合，「主観的」「客観的」の判断に差が生じやすい傾向が見られる．同サンプル群における「主観的」「客観的」の判断における 3 人のアノテーターの判断が完全に異なる割合は，11%[2]であったが，「語りかけ性」があると判断された場合のサンプルで

[2] 図書館サブコーパスからランダムに選んだ 485 サンプル（3 人のアノテーターが同サンプル群に観点付与を行った）について，「小説以外の内容」として観点付与が行われた 253 サ

は,「主観的」「客観的」判断の不一致率が 28%に及んだのである.

アノテーターコメントで「主観的」・「客観的」の判断に迷った旨が記載されている際,「解説書・アドバイス・勧誘」であるとの記述が目立っている. これらの書籍タイトルを見ると,「読本」「～の本」「～法」「～バイブル」「～入門」「～知識」「～講座」などが大半であることがわかった.「語りかけ性」は, この種の実用書（いわゆるハウツー本（啓蒙書・指導書）の類）に見られる傾向が確かめられている（保田ほか, 2012, 2013). すなわち,「語りかけ性」があると感じられるテキストは,「主観的」・「客観的」の判断に迷い, 作業者間でも判断の不一致が生じる可能性がある.

それでは, なぜ「語りかけ性」があると感じられるテキストで,「主観的」・「客観的」の判断に迷いが生じるのか.

次節からは,「主観的」「客観的」と分類されたサンプル群に特徴的な表現を調査し,「語りかけ性」との関係を明らかにしたい. 読み手が著者の主観性を感じる表現と「語りかけ性」があると感じる表現の異同から, テキストにおいて何のために語りかけるという表現手法が用いられているのかを考察する.

４. 考察

４.１ 「語りかけ性」の有無群に出現頻度の高い表現と「主観的」「客観的」群

語などの要素の出現頻度を見ると,「語りかけ性」があるとされるテキストと,「主観的」とされるテキストに類似性がある表現がある.「語りかけ性」があるとされるテキストに頻度の高い要素[3]には, 助動詞の「です」「ます」があり,「語りかけ性」がないとされるテキストに頻度の高い要素には助動詞の「た」がある（保田ほか, 2012a). しかし,「主観的」「客観的」群で大きな差異は見られない.

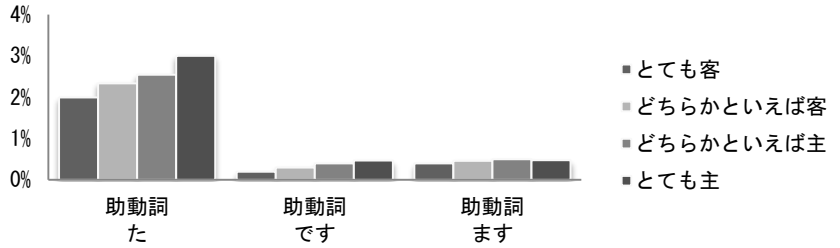


図４ 「主観的」「客観的」群における「語りかけ性」の有無群に出現頻度の高い表現

ンプルを対象に分析を行った.

[3] 「語りかけ性」有無群において, 出現頻度で有意差の見られる要素はほとんど得られなかった. アノテーターが「語りかけ性」があると判断するのに用いたと認識する要素は,「語りかけ性」を形成する表現と言えるが, 個別の出現頻度では影響が捉え難い. そもそも出現頻度を確認することも難しい. まとまった量のテキストにおいて, 種々の表現の総体的な出現量と, 文脈が要されることがわかっている（保田ほか, 2013).

以下に示す例は,「語りかけ性」があるとされるが, 直感的に特徴的と考えられる表現や, 出現頻度の高い表現が見つからないテキストであるといえる. アノテーターのコメントから得られた「語りかけ性」に関わると考えられる表現類に下線を引いた.

例）カップリングコンデンサが大きい場合、オレンジ色の側の配線が同じようにＩＣソケットの足にハンダ付けできればどのように付けても構わない。完成図を見てもらえれば分かると思うが、コンデンサの左の部分は大きくスペースが残してあるので、アキシャルリードのものも基板上に取り付け可能だ。また、大きすぎて基板からはみ出したとしても、特に問題はない。なお、後で説明するが、このコンデンサは無しにも出来る。（酒井智巳「はじめてつくるプリアンプ」）

また,「語りかけ性」のあるテキストでは，相手に対する希望を表す「ほしい」「たい」(「～してほしい」「～されたい」など）や，相手に対する婉曲化の表現として「思う」「感じる」(「～するべきだと思う」など）を用いる傾向がある（保田ほか,2012b）.

この種類の表現は,「主観的」と分類されたテキストにも出現頻度が高いため，類似の表現群が現れる「語りかけ性」があるテキストは,「主観的」と分類されやすくなる可能性が考えられる．図５に「主観的」「客観的」の判断において，出現頻度に特徴的である表現群[4]を示した．但し，頻度が同程度であっても,「主観的」と分類されるテキストで用いられている場合には，文脈的に用法が異なっている可能性がある．「とても主観的」と分類されたテキストの「思う」は，典型的には「それは今につながっているんだけど，やっぱり非常によかったと思う．（坂本龍一「Seldom-illegal」)」のように用いられる．「語りかけ性」がある群のような婉曲化目的というより，個人的な感情や考えを述べていると読める．

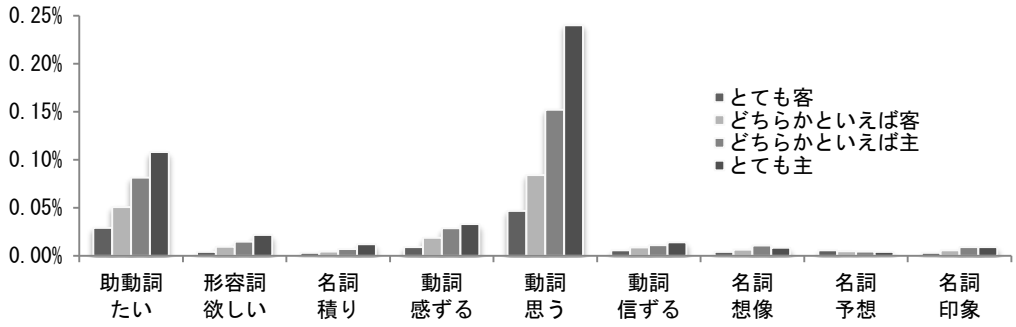


図５　「主観的」「客観的」群に特徴的である表現群

４.２　「主観的」群に特徴的な表現：求められる客観性

平叙文の使用の背後には，常に何らかの問いが存在している（中村,2002）とすれば，読み手に向けて発せられているテキストは，何らかの解答であるという期待を持って受け取られるのだと考えられる．そのため，読み手の同意もしくは共感を得ることが，テキストに求められるはずである.

実際に，アノテーターの「主観的」であるとの判断コメントは,「根拠のない主張である」「事象を理由なしに断定する」「推測が多い」のように，批判的なものが得られている.「客観的」であるとのコメントには「裏付けがある」「納得できる」などの肯定的なものが並び,「主観的」でないことへの批判はないのである．読み手がテキストに対して何らかの解答を期待しているためであろう．但し，書籍タイトルに「体験記」「日記」などが含まれるなど，明らかにエッセイ類と予測されるテキストについては，コメントには「主観的」と判断した際の根拠が記述されるに留まり，否定的な記述は見つかりにくい．エッセイなどは,「主観的」であることが前提とされ，読み手に求められる解答が共感である可能性が考えられる．読み手の要求するものについては，テキストのジャンル性にも関わる.

なお，分類されたテキスト群毎の出現頻度を見ると，断定や推量に関係すると考えられる表現は，図６の割合で出現している.「主観的」と判断されるテキスト群において，意志推量形（「～だろう」など）や「らしい」「そうだ」のような表現はもちろん，断定の助動詞「だ」も多く用いられているのである.「主観的」なテキスト群は，断定や推量が多いというアノテーターのコメントと一致していると言える.

---

[4]図書館サブコーパスからランダムに選び出した約 500 のサンプルのうち「主観的」「客観的」分類について 3 人の作業者の判断が一致したサンプル（51 サンプル・174,961 語）の分析を行った結果から，品詞・活用形・語彙素において，すべての要素の出現頻度について検定を行い，有意差の見られた表現を確認した（調査手順は（保田ほか，2012a）と同様）結果の一部である．その他の表現は注 5 も参照.

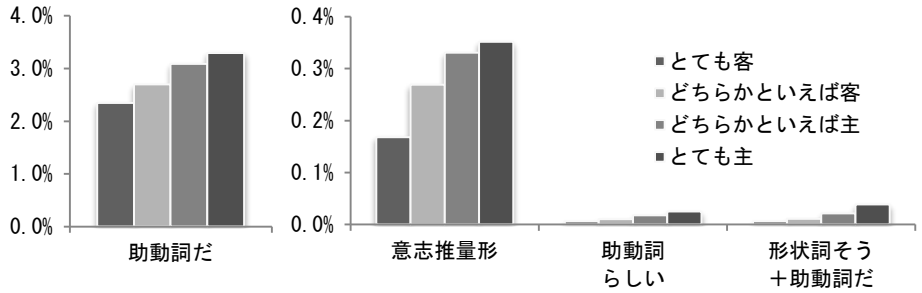


図６「主観的」群に特徴的と考えられる表現

４.３　「客観的」なテキストであるために

それでは,「客観的」であると判断されるテキスト群は，どのような表現が用いられているのか．本稿は，語種と受動文,「という」の伝聞表現に着目し,「主観的」「客観的」群別の出現率などから，どのように客観化が行われているかを確かめる．

４.３.１　客観化（１）: 数値（年号・具体的型番）や具体的名称の割合が多い？

４.２で見たように，読み手はテキストに対して根拠や理由を求めていることが考えられる．そのため，書き手はデータを示すことで客観化を行うはずである．図７に,「主観的」「客観的」分類群別の，普通名詞・数詞・固有語の出現率を示す．特に数詞で「客観的」群での出現率が顕著となっていることがわかる．

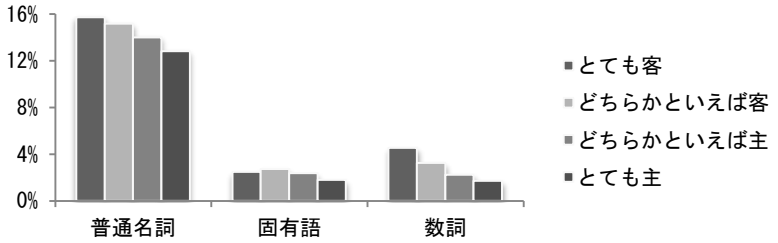


図７　「主観的」「客観的」群別の具体的データ関連要素出現率

４.３.２　客観化（２）: 降格受動文による客観化を行う？

益岡（1991）は，受動文を属性叙述（例：花子の家はビルに囲まれている）と事象叙述に分類し，事象叙述について，受影受動文（例：私は親に叱られた）が主体の経験[5]を表現する主観的な表現で，降格受動文（例：始業のベルが鳴らされた）が客観的表現であると述べている．アノテーターのコメントにも,「報告されている」「評価される」のようなサ変動詞による受動文が，客観的と判断した根拠とされていた旨が散見された．

そこで,「客観的」群と「主観的」群における受動表現の出現率をサ変動詞[6]の受動表現（「される」）について調査した．「される」の出現率は,「客観的」群で 0.34%（23,677 例）「主観的」群で 0.19%（8,599 例）であり,「客観的」群で高くアノテーター判断と一致していると言える．

図 8-1 に,「とても客観的」「とても主観的」二つの分類群からランダムに取得した 900 例の「される」用例を,「属性叙述」と「事象叙述」の「降格受動」「受影受動」に分類した割合を示す．降格受動と受影受動の割合は，客観的と主観的の分類群で明らかに異なっ

[5]動詞の「遣る」「呉れる」「仕舞う」「貰う」でも,「主観的」「客観的」群に差が見られることがわかっている．本文末参考図参照．

[6] 本稿で扱ったサンプル全体における受動表現は，93,873 件あり，うち「される」は 32,275 件と 34%にあたる．

ており，益岡（1991）の指摘に沿う結果となっている．

また，降格受動については，背景化されている動作主を，「私」（例：原因が推定される・場面が想定されるなど），「他の誰か」（例：問題が指摘される・明らかにされるなど），「特定の誰か」（例：商品が値下げされる・遺跡が発掘されるなど），「一般的（不特定の人々）」（例：人命救助が優先される・性能が要求されるなど），「誰かによる何か」（例：金額が記載によって計上される・権利が法に規定されるなど）と分類した．図 8-2 にその割合を示す．動作主が背景化されている中でも，主観的群で動作主が特定されやすく，客観的群では動作主が一般的な例が多いと言える．

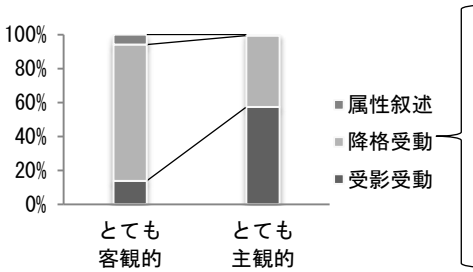


図 8-1　「主観的」「客観的」群別受動文種

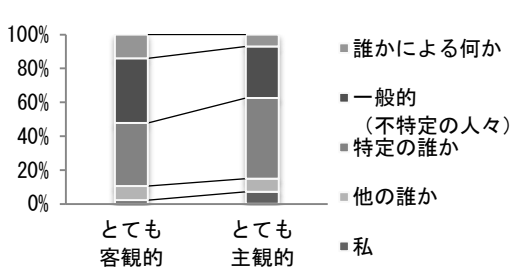


図 8-2　背景化された動作主

４．３．３　客観化（３）：伝聞（「という」）による客観化を行う？

主張の裏付けとして，データの他に引用などを用いることが考えられる．そこで，ここでは伝聞の「という」を用いた表現に着目した．

「という」の頻度[7]のみでは，「とても客観的」群で 9,355 件（0.8%），「とても主観的」群で 8,919 件（1.3%）であり，「とても主観的」に多いということになる．

しかし，「語りかけ性」との関連では，「とても（語りかけ性が）ある」群のうち「とても客観的」「とても主観的」群に出現した「という」2,000 件について調査したところ，このうち文脈上，伝聞（人が～「という」・「という」人がいる・「という」話であるなど[8]）として用いられていたのは，「とても客観的」群で 6.0%，「とても主観的」群で 1.7%であり，伝聞の用例頻度は，「とても客観的」群に多いと考えられる．「語りかけ性」がある群では，「という」を伝聞として用いることで，客観化を行う例が増加する．

４．４　客観性と「語りかけ性」：「語りかけ」は何のために用いられるのか

「語りかけ性」は，「主観的」「客観的」どちらと判断されるテキストからも受け取られる性質であり，「主観的」であることと語りかけることは相関があるとも言い難い．但し，「主観的」と判断されるテキストと類似した要素を含むことから，読み手によっては，語りかける著者が感じられることが，「主観的」と受け取ることもあると考えられる．

読み手に語りかけるという「語りかけ性」は，書きことばにおいては，本来表に現れる

---

[7]「という」を多用するサンプルも見られる．
例）（略）ただそういう不運が生じたということなのだ，ということでしたが，もっと深いなにかがあるのかもしれない，という気もしました．（略）あなたは，なんの理由もなくなにかが起こるという考え，この宇宙はでたらめなのだという考えを，受け入れることができるでしょうか？（H・S・クシュナー（著）／ 齋藤武（訳）「なぜ私だけが苦しむのか」）

[8]1）ギリシャの歴史家ヘロドトスの記述によると，（略）人々が集まったということです．（ユーリィ・ドミトリエフ（著）／佐藤靖彦（訳）「人間と動物の関係」）
2）オルテリウスが，手本にしたという日本地図（清水靖夫「地図で見る世界の形の移りかわり」）
3）『説文通訓定声』をみると，（略）差が出るためだという．（鳥越憲三郎「弥生の王国」）
4）シーザーは（略）失望を隠せない表情をしたという．（谷沢永一「人間通と世間通」）

ことのない著者が現れているのであり，著者の主張が露わであるともいえる．以下に，「語りかけ性」があり，「主観的」でもあるとされる例を示す．「語りかけ性」「主観的」に関わると考えられる表現部分に下線を引いた．

2）私は政治家小泉純一郎には、ほとんど興味がありませんが、人間小泉純一郎には深甚たる興味があります。それは小泉氏が、今日の日本人の一つの典型、つまり徹底して自己愛にとりつかれた人間であり、しかもそれを貫徹することに成功している、今のところそのように見えるということでしょう。自己愛が強いというのは、我が身が可愛いということではありません。小泉総理は、いつでも自分の生命を投げ出す覚悟がある、と私は確信しています。（福田和也「総理の資格」）

しかし，多くのテキストは読み手に何らかの根拠ある解答（少なくとも同意や共感）を求められており，「客観的」であることが期待されている可能性がある．「客観的」であるためには，根拠となるデータを示し，あるいは主体の経験を表す受影受動文の使用を避けることなどの客観化を行うことになる．また，伝聞など，客観化を行う表現手法のうち，特に「語りかけ性」があるテキストに用いられているものも見られるのである．書きことばに対話形式を持ち込むゆえに，「客観的」であることを明らかにするため，客観化に関わる表現が用いられやすい可能性があるのだろう．以下に，「語りかけ性」があり，「客観的」であるとされる例を示す．客観化に関わると考えられる表現に下線を引いた．

3）このようにして、はじめて異なる生物の遺伝子をもち合わせた新種のＤＮＡがつくられたのです。ヒトの三十億のＤＮＡ塩基対の中には、五～十万の遺伝子が存在するといわれています。この中から目的とする遺伝子を取り出す方法をクローニングと呼びます。クローニングはどのように行なうかを見てみましょう。（略）つまり、ファージとヒトのＤＮＡのキメラを大腸菌に感染させると、大腸菌の中で増殖して、百倍以上にもなって大腸菌を溶かして外に出てくるのです。（石浦章一「生命のしくみ」）

また，「語りかけ性」は，「客観的」なテキストに多い NDC3・4 番台（社会科学・自然科学）や C-code の「専門」・「実用」の分野で多く用いられているという特性が見られる（上の例文 3 は，NDC4 番台に分類されるサンプルである）．これは，NDC3・4 番台や C-code の「専門」・「実用」分野に多く含まれるハウツー系書籍に「語りかけ性」が用いられやすい（保田ほか,2012b）ためでもあろう．いわゆるハウツーものとは，「趣味や実用的な事柄の簡便な習得法を説いた書物（スーパー大辞林 3.0)」である．すなわち，読み手が予め目的意識を持ってテキストを読むことが明らかである．よって，「客観的」に解答の要求に応えることを，予め明らかにしているのがハウツー系書籍であるといえる．

以下の例文 4 に，C-code の「実用」に分類されるサンプル，例文 5 に C-code の「専門」に分類され，かつタイトルからハウツー系書籍であることが推測されるサンプルを示す．例文 4・5 ともに，「語りかけ性」があるサンプルという判断がなされている．「語りかけ性」に関わると考えられる表現に下線を引いた．

4）ここでダブルウィッシュボーン式のホイールアライメトについて考えてみましょう。Ｐ．１８９の図を参照しながら読み進めて下さい。まず４本の棒が平行四辺形に結ばれているとします。その平行四辺形の短辺の一方を垂直に固定して他方を上下に動かすと、平行四辺形は上下に変形します。（略）そして残るのは上下アームと考えることができます。（橋口盛典「クルマの基本メカニズム」）

5）したがって，エンジン回転数が上昇した場合，フィールド電流（励磁電流）を減少させて発生電圧を一定に保つためのボルテージ・レギュレータ（ｖｏｌｔａｇｅ　ｒｅ

ｇｕｌａｔｏｒ電圧調整器）を設けねばならない。オルタネータ用のレギュレータは，一般にボルテージ・レギュレータのみで，カット・アウト・リレーはもちろん，カレント・リミッタも特殊な用途の場合を除いて必要ではない。カット・アウト・リレーが不要なのは，オルタネータに取付けたダイオードに，バッテリからの逆流を阻止する働きがあるからである。（竹尾敬三「小型水力発電機製作ガイドブック」）

語りかけるという表現手法は，書き手が読み手の求める解答を提示することを謳い，教示的態度を強調する際に用いられやすいものと考えられる．その場合，解答として「客観的」であることが要求され，著者が現れているという印象があっても，「主観的」であると受け取られないような客観化のための表現が用いられることになる．

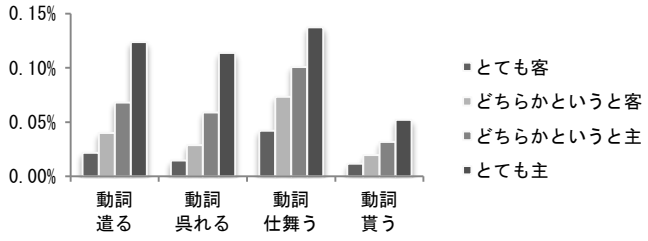
５．まとめ

読み手に「語りかけ」るテキストが，いったい何のために用いられているのかを考察した．「語りかけ性」があるテキストは，著者が前面に現れているということであり，「主観的」であると判断されるのではないかという仮説から，「主観的」あるいは「客観的」と判断されたテキスト群に特徴的な表現やアノテーターコメントの分析を行い，「語りかけ性」との関係性を探った．

結果として，語りかけるという表現手法が，読み手から要求される「客観的」な解答を提示することを示すために用いられやすいということが考えられる．

読み手の求める解答を付与すると明示するために，疑似的な対話を設定し，読み手に相対する書き手が現れることで，教示的態度を強調する．「語りかけ性」があると判断されるテキストでハウツー本が多くを占めるのは，そのためであろう．

また，求められる解答はテキストのジャンルによって異なるが，「客観的」であることが望ましい場合が多く，数値データや，受動文・伝聞などの客観化の効果がある表現の用いられる傾向が見られる．よって，著者が現れていることが，必ずしも「主観的」なテキストであると認識される要因であるとは言えない．



参考図　特徴的動詞の出現率

文　献

安藤宏(2012)『近代小説の表現機構』岩波書店.

柏野和佳子(2010)「「直接的な語り」という表現スタイルをもつ書籍テキストの人手抽出の試み」『ことば工学研究会』35，pp. 63-72.

柏野和佳子，奥村学(2012)「書籍テキストへの分類指標人手付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『言語処理学会第 18 回年次大会』.

中村洋(2002)「「X は Y だ。」と「X が Y だ。」の意味の違いについて」『人工知能基礎論研究会』47，pp. 55-60.

益岡隆志(1991)「受動表現と主観性」仁田義雄(編)『日本語のヴォイスと他動性』pp. 105-121，くろしお出版.

松村真宏，河原大輔，岡本雅史，黒橋禎夫，西田豊明(2007)「メッセージの背後に潜む「問

い」の抽出」『人工知能学会論文誌』22, pp.93-102.
保田祥, 柏野和佳子, 立花幸子, 丸山岳彦(2012a)「「語り性」を有する書きことばの典型例の分析」『第１回コーパス日本語学ワークショップ』予稿集, pp.139-146.
保田祥, 柏野和佳子, 立花幸子, 丸山岳彦(2012b)「「語りかけ性」を有すると判断される書きことばの表現」『第２回コーパス日本語学ワークショップ』予稿集, pp.43-50.
保田祥, 柏野和佳子, 立花幸子(2012)「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」『ことば工学研究会』41, pp.3-10.
保田祥, 柏野和佳子, 立花幸子, 丸山岳彦(2013)「アノテーターコメントを用いた「語りかけ性」分析の試み―頻度情報から捉え難いテキスト性質の解明に向けて―」『言語処理学会第19回年次大会』.

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/

日本語学習者のインタビュー応答時における言いよどみ使用

土屋菜穂子（青山学院大学）

An Analysis of Japanese Language Learners' Hesitations Which Appeared in the Responses to Interview Questions

Naoko Tsuchiya (Aoyamagakuin University)

1　はじめに

日本語を母語としない人々に対して行う日本語教育においては，一般的に「あの」「えーと」「まあ」「なんか」などの言いよどみ[1]に関して，学習者の自然習得に任せている場面が多いと思われる。また，学習者個人によってその習得方法も当然異なるものである。

しかし例えば，日本語を自在に操り自分の専門分野についても論じられる超級レベルの話者らが使う言いよどみと，身の回りの話題に対する質問になんとか「文」で答えることのできる中級レベルの話者らが使う言いよどみとを比べた場合,「それぞれのレベル内で共通する言いよどみの使い方」はあるのであろうか。また，学習者はどのような過程を経て言いよどみを習得していくのか，これをコーパスから調査することは可能であろうか。本研究ではこの二つの点について『日本語学習者会話データベース』[2]（国立国語研究所, 2010 年）を用いて調査・考察を行う。

2　調査の方法

2.1　今回取り上げる言いよどみ

本研究で取り上げる表現は,「あの」「まあ」「なんか」「えーと」，そして「んー」「あー」である。これらのうち「んー」「あー」を除くものは，田窪 (1995) において語彙的形式を持つ言いよどみ系感動詞とされているものである。一方「んー」「あー」は語彙的形式を持たず，特に「あー」に関しては伝統的な国文法の解釈において典型的な感動詞であると捉えることもできるが，近年のフィラー研究においては，石川 (2010)，定延 (2010) などのように，あ系感動詞もフィラーに含めることがある[3]。本研究においても，語彙的形式を持つ言いよどみとの比較の観点からこれらの語を取り上げる。なお，以後，本論内では便宜的に「語彙的言いよどみ」「非語彙的言いよどみ」という呼び方を使うことにする。

2.2　『日本語学習者会話データベース』の特徴と n-gram 処理の使用

データには上記の『日本語学習者会話データベース』のテキストデータを用いる。このデータは，目標言語の口頭能力レベルを測る OPI(Oral Proficiency Interview) の形式[4]で

[1]ここでいう「言いよどみ」とはいわゆるフィラーに相当するものである。土屋（2000）との整合性，また，筆者が早稲田大学日本語教育研究センターで担当しているテーマ科目名「日本語の言いよどみ・あいづち・コメント 5-6」との整合性のためにここでは「言いよどみ」という用語を用いている。フィラーと置き換えて読んでいただいて差し支えない。

[2]https://dbms.ninjal.ac.jp/nknet/ndata/opi/を参照。

[3]『日本語話し言葉コーパス』（国立国語研究所，2004 年）の書き起こしにおいても，あ系感動詞にはフィラーのタグが付けられている。

[4]OPI の概要については牧野，他（2001）を参照。

行われたインタビューのデータである。表 1 に示すように，インタビューデータは OPI のレベルの枠組みに沿って超級・上級・中級・初級の主要 4 レベル，また，超級から初級 - 下までの 10 下位レベルに分類されている。合計 339 名のインタビューデータがある。本研究においても学習者の口頭能力レベルの枠組みはこれに従うものとし，超級から初級 - 中までの主要 4 レベル・9 下位レベル，合計 338 名のインタビューデータを調査の対象とする。下位レベル単位で分析を行うため，データが 1 名のみの初級 - 下は今回は調査の対象から外している。

表 1: 『日本語学習者会話データベース』に収録されている人数と口頭能力レベル

超級	上級	中級	初級
超級　9 名	上級 - 上　24 名	中級 - 上　68 名	初級 - 上　21 名
	上級 - 中　34 名	中級 - 中　84 名	初級 - 中　10 名
	上級 - 下　52 名	中級 - 下　36 名	初級 - 下　 1 名

テキストデータは以下のような漢字仮名交じり，句点なし・読点ありの書き起こしファイルになっており，インタビューアーの発話（T）と学習者の発話（I）が交互に示される形式になっている。発話途中に挟まれた相手のあいづちは〈　〉で示されている。

（超級 0010）
T：あーそうですか〈えー〉，で，えー今日本では〈はい〉何をしてらっしゃいますか
I：今はえーとーパソコン関係の〈ん〉ウェブ関係のお仕事をしてます

なお，本データベースには現段階では形態素情報は付与されてない。以上の特徴から，本データベースのテキストデータに n-gram 処理[5]を行うことが有効であると考える。学習者の発話から任意の長さの文字列をくまなく取り出すことで頻出度の高い表現を抽出することが可能になる。

2.3　n-gram 処理の具体的な方法

『日本語学習者会話データベース』の超級から初級 - 中レベルまでの主要 4 レベル・9 下位レベル，全 338 名のインタビューのテキストデータの学習者の発話部分に対して n-gram 処理を行う。n-gram 処理の前にはテキスト処理言語 awk で，(1) テキストファイルから学習者の発話行を抽出，(2) 抽出した発話行に挿入されているインタビューアーのあいづちを除去，(3) 非言語情報のタグを除去，(4) 時間情報のタグを除去，といった作業を行っている。話者タグ（”I :”）および読点（”, ”）は除去していない。

以上の作業を行った各話者の発話部分のみのテキストファイルに対して n-gram をかける。抽出する文字列は各話者につき頻度 1 以上，2 グラムから 15 グラムまでである。n-gram 処理の後，下位レベルごとに n-gram の結果をマージした対照表を作成し，先に挙げた言いよどみ表現にあたる文字列の出現を手掛かりに分析していく。ただし，「えーと」には，「えーと」「えっと」「えと」「えーっと」といった複数の文字列があったため，これら 4 種類の文字列を調査の対象とし，以下ではこれらをまとめてカタカナの「エート」と表記することにする。なお，n-gram 処理には morogram（師茂樹氏 作）の Windows 実行形式である極悪氏版[6]を使用した。

[5]n-gram の人文科学への応用に関しては長尾・森 (1993)，近藤みゆき (2000)，近藤泰弘 (2000) を参照。
[6]http://www.vector.co.jp/から入手可能のフリーソフトである。同氏のソフトウェアは Perl 環境がなくて

3　各レベルにおける言いよどみの使用方法の調査

3.1　当該文字列の出現回数と調整頻度

まずは単純に，それぞれの下位レベル内における当該文字列の出現と，その文字列が出現した人数を見てみることにする。表 2 では，下位レベルごとに最も出現回数の多かった文字列から順番に表示している[7]。なお，ここで示しているのはあくまでも「出現した文字列」であり，中には言いよどみではない文字列も含まれていることに注意を要するが，その数は全体数のうちのわずかであると判断し，ここでは出現した文字列を言いよどみの語とみなす。

表 2: 当該文字列の出現回数と出現人数

超級 9 名	出現回数 (人数)	上級上 24 名	出現回数 (人数)	上級中 34 名	出現回数 (人数)
あの	774(9 名)	あの	1778(22 名)	あの	1863(32 名)
まあ	462(9 名)	まあ	545(23 名)	なんか	1028(31 名)
なんか	202(8 名)	なんか	539(24 名)	エート	807(17 名)
エート	118(8 名)	んー	389(24 名)	んー	777(34 名)
んー	111(9 名)	あー	323(24 名)	まあ	672(29 名)
あー	68(9 名)	エート	307(20 名)	あー	528(33 名)
上級下 52 名	出現回数 (人数)	中級上 68 名	出現回数 (人数)	中級中 84 名	出現回数 (人数)
あの	1866(45 名)	あの	2805(59 名)	あー	5066(84 名)
んー	1679(51 名)	あー	2514(67 名)	んー	3267(84 名)
なんか	1327(47 名)	んー	1966(68 名)	なんか	1237(54 名)
あー	1211(52 名)	なんか	1514(54 名)	あの	1169(68 名)
まあ	918(35 名)	エート	1383(39 名)	エート	953(37 名)
エート	729(33 名)	まあ	548(38 名)	まあ	202(42 名)
中級下 36 名	出現回数 (人数)	初級上 21 名	出現回数 (人数)	初級中 10 名	出現回数 (人数)
あー	2602(36 名)	あー	2251(21 名)	あー	481(10 名)
んー	1213(36 名)	んー	855(21 名)	んー	367(10 名)
あの	509(31 名)	あの	196(16 名)	あの	38(4 名)
エート	247(19 名)	エート	165 (6 名)	まあ	19(5 名)
まあ	150(15 名)	まあ	19(7 名)	エート	7(3 名)
なんか	93(14 名)	なんか	6(5 名)	なんか	3(1 名)

この結果から，各レベルの言いよどみの使用方法を調査するためのいくつかの手掛かりを見つけることができる。例えば以下の点に注目することができる。

- それぞれの下位レベル内で出現回数が最も多いのは，超級から中級 - 上までは「あの」，中級 - 中から初級 - 中までは「あー」次いで「んー」の順番である。
- 9 つの下位レベルの中間に位置する中級 - 上は，出現回数が最も多いのが「あの」，それに次ぐのが「あー」「んー」である。
- 非語彙的な言いよどみの出現回数に関して，初級 - 中から中級 - 上までは「あー」次いで「んー」の順であるのに対し，上級 - 下から超級までは「んー」の出現回数が「あー」を上回る。
- 語彙的言いよどみの出現に関して,「なんか」の出現回数が上位に上がってくるのは中級 - 中からである。一方,「まあ」の出現回数が上位になるのは上級 - 上からである。

も n-gram 処理を行うことができる。オリジナルの morogram については http://morogram.sourceforge.jp/ を参照。

[7]土屋（2012）では各話者につき頻度 2 以上の文字列を抽出したため，頻度 1 以上の文字列を抽出した今回の表とは数値が若干異なっている。ただし出現回数の順番に大きな変動はない。

次に，それぞれの下位レベルでの当該文字列の出現回数を 100 万語あたりの調整頻度に直して示したのが表 3 である。調整頻度を出す際に必要となる総語数を出すにあたっては，形態素解析に MeCab，辞書は UniDic を使用した。n-gram 処理をする際に作成した学習者の発話部分のテキストを形態素解析にかけ，その結果から話者タグと読点，かぎかっこ等の不要な行を除いたものを総語数として数えた。学習者が誤って発話した部分には誤解析が多少見られたが修正はしていない。このような条件下で出した下位レベル別の総語数は，超級，9 名 33504 語，上級 - 上，24 名 85459 語，上級 - 中，34 名 101273 語，上級 - 下，52 名 143140 語，中級 - 上，68 名 159710 語，中級 - 中，84 名 171496 語，中級 - 下，36 名 63508 語，初級 - 上，21 名 31047 語，初級 - 中，10 名 13452 語である。

表 3: 100 万語あたりの出現回数

	あの	まあ	なんか	エート	んー	あー
超級	23102	13789	6029	3522	3313	2030
上級 - 上	20805	6377	6307	3592	4552	3780
上級 - 中	18360	6622	10131	7953	7657	5203
上級 - 下	13036	6413	9271	5093	11730	8460
中級 - 上	17563	3431	9480	8659	12310	15741
中級 - 中	6816	1178	7213	5557	19050	29540
中級 - 下	8015	2362	1464	3889	19100	40971
初級 - 上	6313	612	193	5315	27539	72503
初級 - 中	2825	1412	223	520	27282	35757

3.2 口頭能力レベルと当該言いよどみの出現回数との関係

これらの数値から，レベルの上昇と出現回数が比例しているのが「あの」「まあ」，反比例しているのが「んー」「あー」であることが分かる。レベルが上がるにつれ，語彙的言いよどみである「あの」「まあ」は出現回数が増えていく。調整頻度を見ると「あの」は中級 - 上からその出現回数が大幅に増え始めるのに対し,「まあ」は上級 - 下から大幅に増え始め，更に超級ではその出現回数が倍に増える。この結果から,「まあ」の習得の難しさがうかがえる。そして，レベルが上がるにつれ出現回数が減少していくのが非語彙的な言いよどみである「んー」「あー」である。ただし，先にも述べたように上級 - 下から上位のレベルでは,「んー」の出現回数が「あー」を上回ることは注目に値する。これについては更に検討を要する。

一方,「なんか」は中級 - 中から大幅に増えるがそのピークは上級 - 中であり，上級 - 上からは再び減少する。同じく語彙的言いよどみであり上位レベルになるほど出現回数が増える「あの」「まあ」とは異なる使用状況が観察される。一つの解釈としては,「なんか」は中級から上級の入り口程度で多用される言いよどみであると考えることができるであろう[8]。

「エート」については「なんか」ほどの特徴を数値からは見出すことはできないが，調整頻度を見る限り，やはり上級 - 上からその出現回数は減少している。

[8]ただし，OPI の性質がかかわっている可能性もある。OPI においては上級 - 上から超級の話者に対して社会的あるいは専門的な話題の中で話者の論理的な意見やその裏付け等を問うため，文脈的に「なんか」が出現しにくい事情がある可能性も否定できない。更に検討を要する。

3.3 当該言いよどみ同士の共起パターン

各レベルに共通した言いよどみの使用方法を更に探るために，次は当該文字列同士が共起する例を調査する。当該文字列同士が共起する例とは以下のようなパターンを指し，隣り合って出現しているものに限定する。

あの，まあ，日本語ではいろいろ（超級 0076）／ あーなんか 外国人なのになんでそ（超級 0015）

これらの例も下位レベルごとに n-gram 処理の結果をマージした対照表から抽出する。仮に「あのー，まあ，なんか」という文字列があった場合には,「あの＋まあ」と「まあ＋なんか」の二つの共起パターンが抽出されることになる。

表 4: 当該言いよどみ同士の共起表　超級～中級 - 上

超級 9 名

	あの	まあ	なんか	エート	んー	あー
あの	3 名 6 回	8 名 33 回	0 名 0 回	5 名 8 回	0 名 0 回	0 名 0 回
まあ	4 名 24 回	3 名 5 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
なんか	3 名 10 回	2 名 4 回	0 名 0 回	0 名 0 回	1 名 1 回	1 名 2 回
エート	2 名 5 回	2 名 3 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
んー	2 名 2 回	3 名 5 回	1 名 2 回	1 名 1 回	0 名 0 回	0 名 0 回
あー	1 名 1 回	0 名 0 回	1 名 1 回	1 名 1 回	0 名 0 回	2 名 3 回

上級 - 上 24 名

	あの	まあ	なんか	エート	んー	あー
あの	9 名 14 回	13 名 49 回	7 名 37 回	3 名 4 回	8 名 11 回	4 名 5 回
まあ	11 名 24 回	6 名 6 回	2 名 2 回	2 名 2 回	2 名 2 回	0 名 0 回
なんか	6 名 13 回	2 名 4 回	2 名 5 回	0 名 0 回	3 名 5 回	4 名 4 回
エート	3 名 3 回	6 名 9 回	1 名 1 回	2 名 4 回	2 名 2 回	0 名 0 回
んー	7 名 7 回	10 名 19 回	5 名 8 回	0 名 0 回	2 名 2 回	2 名 3 回
あー	5 名 6 回	3 名 4 回	0 名 0 回	0 名 0 回	2 名 4 回	4 名 4 回

上級 - 中 34 名

	あの	まあ	なんか	エート	んー	あー
あの	12 名 40 回	7 名 28 回	5 名 19 回	5 名 7 回	7 名 23 回	5 名 19 回
まあ	5 名 10 回	6 名 8 回	7 名 7 回	2 名 16 回	2 名 3 回	5 名 8 回
なんか	7 名 14 回	6 名 6 回	7 名 12 回	3 名 3 回	8 名 12 回	3 名 4 回
エート	2 名 6 回	6 名 45 回	5 名 5 回	7 名 37 回	5 名 15 回	5 名 8 回
んー	7 名 10 回	9 名 28 回	15 名 34 回	4 名 5 回	6 名 10 回	4 名 4 回
あー	5 名 7 回	4 名 6 回	5 名 8 回	3 名 3 回	6 名 6 回	3 名 3 回

上級 - 下 52 名

	あの	まあ	なんか	エート	んー	あー
あの	12 名 18 回	8 名 54 回	6 名 21 回	9 名 37 回	14 名 33 回	9 名 33 回
まあ	7 名 18 回	6 名 15 回	3 名 3 回	3 名 3 回	5 名 9 回	2 名 4 回
なんか	7 名 11 回	2 名 2 回	9 名 12 回	7 名 9 回	13 名 29 回	10 名 13 回
エート	9 名 16 回	8 名 13 回	6 名 8 回	8 名 11 回	9 名 16 回	8 名 14 回
んー	13 名 20 回	15 名 33 回	14 名 29 回	10 名 12 回	17 名 36 回	12 名 21 回
あー	13 名 21 回	7 名 10 回	8 名 12 回	5 名 12 回	10 名 19 回	10 名 25 回

中級 - 上 68 名

	あの	まあ	なんか	エート	んー	あー
あの	18 名 88 回	3 名 6 回	7 名 13 回	3 名 3 回	16 名 50 回	15 名 83 回
まあ	2 名 8 回	8 名 12 回	5 名 8 回	1 名 17 回	4 名 6 回	4 名 6 回
なんか	9 名 15 回	3 名 5 回	7 名 12 回	4 名 41 回	18 名 61 回	10 名 22 回
エート	2 名 7 回	5 名 26 回	7 名 21 回	8 名 23 回	11 名 25 回	8 名 18 回
んー	12 名 25 回	13 名 23 回	17 名 55 回	11 名 15 回	21 名 54 回	15 名 22 回
あー	10 名 14 回	10 名 17 回	12 名 32 回	8 名 16 回	22 名 38 回	25 名 79 回

当該言いよどみ同士の共起パターンを下位レベルごとに示したのが表 4，5 である。各下位レベル内で半数以上の話者に出現した共起パターンには欄内に網掛けを施してある。例えば超級話者の表中にある「8 名 33 回」とは，超級話者全 9 名中，8 名の話者の発話に「あの＋まあ」という共起が合計 33 回出現したという意味を示す。

表 5: 当該言いよどみ同士の共起表　中級 - 中〜初級 - 中

中級 - 中 84 名

	あの	まあ	なんか	エート	んー	あー
あの	11 名 27 回	1 名 1 回	6 名 17 回	1 名 1 回	13 名 38 回	7 名 58 回
まあ	1 名 1 回	9 名 14 回	0 名 0 回	1 名 1 回	3 名 6 回	4 名 4 回
なんか	3 名 4 回	3 名 4 回	6 名 6 回	0 名 0 回	16 名 58 回	15 名 25 回
エート	2 名 4 回	1 名 9 回	3 名 6 回	7 名 13 回	18 名 32 回	12 名 75 回
んー	12 名 19 回	6 名 8 回	20 名 39 回	10 名 40 回	45 名 103 回	42 名 107 回
あー	16 名 30 回	5 名 15 回	10 名 24 回	7 名 34 回	42 名 110 回	43 名 142 回

中級 - 下 36 名

	あの	まあ	なんか	エート	んー	あー
あの	4 名 12 回	1 名 1 回	0 名 0 回	0 名 0 回	3 名 8 回	4 名 28 回
まあ	0 名 0 回	5 名 7 回	1 名 2 回	0 名 0 回	1 名 1 回	1 名 1 回
なんか	0 名 0 回	1 名 5 回	0 名 0 回	0 名 0 回	3 名 4 回	1 名 2 回
エート	2 名 2 回	1 名 1 回	0 名 0 回	1 名 2 回	7 名 12 回	5 名 13 回
んー	3 名 3 回	4 名 6 回	5 名 7 回	2 名 2 回	17 名 35 回	18 名 40 回
あー	5 名 15 回	2 名 3 回	4 名 5 回	2 名 2 回	18 名 52 回	27 名 123 回

初級 - 上 21 名

	あの	まあ	なんか	エート	んー	あー
あの	2 名 4 回	0 名 0 回	0 名 0 回	0 名 0 回	4 名 12 回	5 名 5 回
まあ	0 名 0 回	2 名 2 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
なんか	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
エート	0 名 0 回	0 名 0 回	0 名 0 回	1 名 2 回	2 名 5 回	3 名 34 回
んー	2 名 9 回	0 名 0 回	0 名 0 回	2 名 8 回	13 名 45 回	13 名 51 回
あー	5 名 6 回	3 名 3 回	0 名 0 回	3 名 7 回	16 名 69 回	17 名 149 回

初級 - 中 10 名

	あの	まあ	なんか	エート	んー	あー
あの	1 名 1 回	0 名 0 回	0 名 0 回	1 名 1 回	2 名 2 回	2 名 4 回
まあ	0 名 0 回	1 名 1 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
なんか	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
エート	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回	0 名 0 回
んー	1 名 1 回	1 名 1 回	0 名 0 回	1 名 1 回	5 名 19 回	0 名 0 回
あー	1 名 5 回	0 名 0 回	0 名 0 回	1 名 1 回	4 名 6 回	6 名 17 回

各表中には四つの大きな区切りがある。左上の区切りには，語彙的言いよどみ同士の共起（例．あの＋まあ），左下の区切りには非語彙的＋語彙的言いよどみの共起（例．んー＋まあ），右上の区切りには語彙的＋非語彙的言いよどみの共起（例．あの＋あー），右下の区切りには非語彙的言いよどみ同士の共起（例．んー＋あー）が位置している。

まずは，非語彙的言いよどみが後接するパターンの出現に注目する。超級を見てみると，「あの＋まあ」などの語彙的言いよどみ同士の共起,「んー＋まあ」などの非語彙的＋語彙的言いよどみの共起例はいくつか見られるが，語彙的言いよどみに非語彙的言いよどみが後接するパターン（例.「あの＋あー」）や，非語彙的言いよどみが連続するパターン（例．あー＋んー）はほぼ出てこない。つまり，言いよどみの内側に，非語彙的言いよどみが入り込むパターンがほとんど見られない[9]。

その後，レベルが下がるに従い，表中右上・右下の非語彙的言いよどみが後接するパターンの出現が増えていく。更に中級 - 中から下のレベルでは，表中右下の「あー＋あー」や「あー＋んー」のような非語彙的言いよどみの連続が半数以上の話者に現れる。

非語彙的言いよどみの連続は，多くの場合，話者にとっての言語産出の難しさが反映されているものと解釈できる。例えば以下のような非語彙的な言いよどみが三つ以上連続す

[9]現在調査中の『インタビュー形式による日本語会話データベース』（上村隆一，じんこんもん DATABASE Vol.1，重点領域「人文科学とコンピュータ」総括班，1998）の母語話者 50 名のデータからも同様の結果を得ている。非語彙的言いよどみの連続や，言いよどみの内側に入り込む非語彙的言いよどみの例はきわめて少ない。なお，このデータベースも OPI のインタビュー形式である。

る例は，上級 - 中から超級の話者の発話に見出すことができない。

あー，んー，んー高いー，たかー（上級 - 下 0055）／あー，んー，あー，あちょっとわ（中級 - 上 0030）
あー，んー，んー，あーいちじか（中級 - 中 0330）／あー，んー，んー，先生，は，漢（中級 - 下 0256）

よって，口頭能力レベルが上がるほど，言いよどみに非語彙的言いよどみを後接させて発話することが減少するものと考えられる。

次に，各下位レベル内で出現する共起パターンの種類の多寡に注目する。初級 - 中から中級 - 下までは「0 名 0 回」の欄が目立つが，中級 - 中以降，特に中級 - 上から上級 - 中まではどの言いよどみ同士も共起するという結果である。そして再び超級では「0 名 0 回」の欄が目立つようになる。この結果に対しては以下のように解釈する。初級 - 中から中級 - 下まではそもそも「あの」「まあ」「なんか」「エート」といった語彙的言いよどみを使うのが困難であるために言いよどみの共起パターンの種類が少なく,「0 名 0 回」の欄が目立つ。語彙的言いよどみの習得は中級 - 中程度から始まっていくものと思われるが，超級へ至るまでの過程に語彙的・非語彙的にかかわらず複数の言いよどみを組み合わせて使用する時期があり，それが中級 - 上から上級 - 下，あるいは上級 - 中までの時期に相当する。最終的に超級に至る時期には，非語彙的言いよどみを後接させる使用方法がなくなり言いよどみの使用方法が洗練され，再び「0 名 0 回」の欄が多くなる。

4　結論・研究の応用・次への課題

「それぞれのレベル内で共通する言いよどみの使い方」はあるかという問いに対しては「ある」と答えることができる。言いよどみの習得過程を踏まえながら述べると，次のようになる。

初級から中級 - 中レベルの話者は非語彙的言いよどみを最も多く使用するが，中級 - 上レベルから語彙的言いよどみ「あの」を多く使用し始め，その後も超級に至るまで「あの」を最も多く使用することになる。ただし中級 - 上レベルは非語彙的言いよどみの使用も多く残り，過渡期的な位置を占める。また，中級 - 上から上級 - 中程度までは，語彙的・非語彙的を問わず複数の言いよどみを共起させて発話することがある。超級に至る時期には非語彙的言いよどみの後接がなくなり，言いよどみの共起パターンが減少する。これは使用方法が洗練されることであると解釈できる。

また，言いよどみの語に焦点を当てれば以下のように述べることができる。

語彙的言いよどみ「あの」「まあ」の出現回数はレベルの上昇と比例関係にあり，逆に非語彙的言いよどみ「んー」「あー」の出現回数はレベルの上昇と反比例の関係にある。ただしそれぞれの語には出現状況に特徴があり,「あの」は中級 - 上から,「まあ」は上級 - 下から大幅に増え始める。「んー」と「あー」では初級 - 中から中級 - 上までは「あー」の方が出現回数が多いが，上級 - 下から超級にかけては出現回数の多寡が逆転する。また,「なんか」は中級 - 中程度から上級 - 中の間で多用される。「エート」については不明である。

以上,『日本語学習者会話データベース』のテキストデータと n-gram 処理を用いて調査・考察を行った。冒頭に述べた「学習者はどのような過程を経て言いよどみを習得していくのか，これをコーパスから調査することは可能か」という問いに対しても，ある程度可能であると答えることができる。

本研究の発展は，教育実践の場に応用可能である。学習者の口頭能力レベルに応じた言いよどみの習得を促すことに役立てることができる。ただし今回の調査では取り上げる語を限定しており，かつテキストデータのみを用いたため課題も多い。以下に，次への課題を挙げる。

- 音声データの利用。特に非語彙的言いよどみ「んー」「あー」の音質の調査。初級から超級に至るまで，音質が異なっているものが一律に表記されている可能性がある。
- 「えー」「その」等，他の言いよどみの調査。
- OPI 形式を持つ他のインタビューデータとの結果比較。『KY コーパス』[10]（鎌田修・山内博之，1999 年）や『インタビュー形式による日本語会話データベース』の母語話者のインタビューデータ等。
- OPI 形式を持たない他のインタビューデータとの結果比較。『日本語話し言葉コーパス』の模擬講演インタビュー等。

付記

この研究は，国立国語研究所のプロジェクト『日本語教育データベースの構築-日本語学習者会話データベース』を利用して行われたものである。

参考文献

石川創 (2010)「あいづちとの比較によるフィラーの機能分析」『早稲田日本語研究』19，早稲田大学日本語学会，pp.61-72

近藤みゆき (2000)「n グラム統計処理を用いた文字列分析による日本古典文学の研究」『千葉大学人文研究』第 29 号，千葉大学文学部，pp.187-238

近藤泰弘 (2000)「≪文化資源≫としてのデジタルテキスト―国語学と国文学の共通の課題として―」『国語と国文学』第 77 巻 11 号，東京大学国語国文学会，pp.127-139

近藤泰弘・近藤みゆき (2001)「N-gram の手法による言語テキストの分析方法―現代語対話表現の自動抽出に及ぶ―」『漢字文献情報処理研究』第 2 号，漢字文献情報処理研究会，pp.50-55

定延利之 (2010)「会話においてフィラーを発するということ」『音声研究』第 14 巻第 3 号，日本音声学会，pp.27-39

田窪行則 (1995)「音声言語の言語学的モデルをめざして―音声対話管理標識を中心に―」『情報処理』第 36 巻第 11 号，情報処理学会，pp.1020-1026

土屋菜穂子 (2000)「対話コーパスを用いた言い淀みの統語論的考察」『青山語文』第 30 号，青山学院大学日本文学会，pp.13-26

土屋菜穂子 (2012)「日本語学習者の口頭能力レベル別言いよどみ使用―『日本語学習者会話データベース』の n-gram 分析をもとにして―」『日本語教育学会秋季大会予稿集』日本語教育学会，pp.253-254

長尾眞・森信介 (1993)「大規模日本語テキストの n グラム統計の作り方と語句の自動抽出」『情報処理学会研究報告. 自然言語処理研究会報告』93(61)，情報処理学会，pp.1-8

牧野成一，鎌田修，山内博之，他 (2001)『ACTFL-OPI 入門 日本語学習者の「話す力」を客観的に測る』，アルク

山内博之 (2004)「語彙習得研究の方法―茶筌と N グラム統計―」『第二言語としての日本語の習得研究』7 号，第二言語習得研究会，pp.141-161

山内博之 (2009)『プロフィシェンシーから見た日本語教育文法』，ひつじ書房

[10] http://opi.jp/shiryo/ky_corp.html を参照。

複数の分野のコーパスを用いた述語項構造解析の比較

―『現代日本語書き言葉均衡コーパス』を用いて ―

吉本 暁文 （奈良先端科学技術大学院大学）[†1]
小町　守 （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]

A Comparison of Predicate Argument Structure Analysis on Multi-domain Corpora

— Using the Balanced Corpus of Contemporary Written Japanese —

Akifumi Yoshimoto (Nara Institute of Science and Technology)
Mamoru Komachi (Nara Institute of Science and Technology)
Yuji Matsumoto (Nara Institute of Science and Technology)

1　はじめに

述語項構造解析とは、動作や状態、出来事を表す動詞、形容詞、動作名詞等を述語とし、その述語に関わっている単語や句とその役割を同定するタスクである。例えば「太郎が次郎に本を貸した」という文では「貸す」という単語が述語に相当する。また、「太郎」は本を貸すという動作の主体であり、「次郎」は動作の受け手であり、「本」は動作の対象であり、それぞれガ格、ニ格、ヲ格の項と呼ぶ。述語とこれらの項との関係を述語項構造という。述語項構造解析により、機械翻訳での単語の対応における誤りを低減したり、情報検索における絞り込みに役立てたりすることができるようになる。

しかしながら、従来の述語項構造解析は主に新聞記事を対象にして研究が進められてきた。新聞記事は日本語の書き言葉の中でも文体に揺れが少ないため、新聞記事を訓練・評価両方に用いた場合、高い性能を得られたとしても、それが他の分野においても同様に高い性能を示すとは限らない。

そこで本研究では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）を用いて新聞記事以外の分野を対象に述語項構造解析解析を行った結果について分析する。先行研究における述語項構造解析では、ほとんどの場合新聞記事を対象に評価されているが、本稿ではBCCWJ に含まれる Yahoo!知恵袋における評価結果について報告する。本研究では新たにタグ付けされた Yahoo!知恵袋[†4] を用いて述語項構造解析器を学習し、分野の異なるテキストに対して述語項構造解析を行なった結果について比較する。

[†1] akifumi-y@is.naist.jp
[†2] komachi@is.naist.jp
[†3] matsu@is.naist.jp
[†4] http://chiebukuro.yahoo.co.jp/

本研究の主要な貢献は、以下の 2 点である。

- Yahoo!知恵袋の述語項構造解析アノテーションを用いて述語項構造解析器のトレーニング・テストを行った初めての研究である。
- コーパスのドメインごとに傾向の異なるトレーニングデータが得られることを確認した。

本稿は以下のように構成される。まず 2 節で述語項構造解析で用いられているコーパスについて概観し、3 節で本研究で用いた述語項構造解析器について説明する。そして 4 節で評価実験の内容と定量評価の基準について述べ、5 節で結果の報告と分析を行う。最後に、6 節でまとめと今後の方針について述べる。

2 日本語述語項構造タグ付きコーパス

これまでに様々なジャンルのテキストを対象に、日本語の述語項構造タグ付きコーパスが開発されてきた。

たとえば、京都大学テキストコーパス Version 4.0 [Kaw02] は 1995 年 1 月 1 日から 17 日までの全記事（約 2 万文）と 1 月から 12 月までの社説記事（約 2 万文）の計 4 万文のうち 5,000 文に対して格・省略・照応・共参照の情報がアノテートされている。述語項構造情報は必須格・任意格の両方を表層格でタグ付けされている。また、KNB コーパス (Kyoto-University and NTT Blog Corpus) [橋本 11] は、「京都観光」、「携帯電話」、「スポーツ」、「グルメ」の 4 つのテーマについて書かれた 249 のブログ記事、4,186 文からなる人手による解析済みブログコーパスで、格・省略・照応情報がアノテートされている。アノテーション基準は京都大学テキストコーパスと共通である。京都大学テキストコーパス Version 4.0 と KNB コーパスを比較することで、同じアノテーション基準でアノテートされた新聞記事とブログ記事の述語項構造解析結果を比較することができるが、それぞれ 4,000–5,000 文とテストには十分な規模があるものの、機械学習手法で解析器を学習するには比較的小規模である。

一方、NAIST テキストコーパス 1.4 [飯田 10] は京都大学テキストコーパス 3.0 を元に、4 万文全体に対して照応・共参照・述語項構造の情報がアノテートされている。NAIST テキストコーパスでは述語の基本形に対し、格交替を原型に戻したうえで、必須格となる表現をガ格・ヲ格・ニ格でタグ付けしている。NAIST テキストコーパスは大規模に述語項構造がタグ付けされているものの、新聞記事という 1 つのジャンルでしかタグ付けされていないため、分野をまたいだ比較が難しい。

そこで、本研究では現代日本語書き言葉均衡コーパス (BCCWJ) を対象に、複数の分野における述語項構造解析器を比較した。BCCWJ にはコアデータ [小椋 09] と呼ばれる手作業で形態素情報・構文情報を付与したデータセットがあり、コアデータのうち、書籍 (PB)、新聞 (PN)、白書 (OW)、Yahoo!知恵袋 (OC) には [小町 11] により NAIST テキストコーパスと同じアノテーション基準による述語項構造と照応関係のタグ付けがなされている。

3　機械学習による述語項構造解析

これまでにいくつかの述語項構造解析手法が提案されている。また、述語項構造解析に関連するタスクとして述語の深層格と、述語がとる格を同定する格解析がある。たとえば、KNP [†1]は依存構造解析の過程で、大規模格フレームを用いた確率モデルによって格解析を行っている。他にも、[飯田 04, Ima09, Tai08, Yos11, 笹野 11] などがある。機械学習による手法は表層の語彙素性や大規模な格フレーム情報といった知識を用いることによって性能を向上させることができるが、一般的に機械学習による手法は解析器のトレーニングに用いたデータとテストに用いるデータが異なる分野であるとき、性能が落ちることが知られている。

そこで、本稿では、分野の異なるテキストに対する述語項構造解析器の性能を測るため、機械学習に基づく手法でトレーニングとテストにおいて異なる分野のテキストを用意し、それぞれで性能を比較する。具体的には、トーナメントモデル [飯田 04] によって項の同定を行う。トーナメントモデルでは、トレーニングの際には項（正解）である句と項でない句のペアに対し、その句のペアにおける手がかりから項である句を選ぶように機械学習を行う。テストの際には句のペアを作り、どちらが項候補かを分類してその項候補と残りの句のペアを作り、どちらが項候補かを分類するといった計算を繰り返す。これによって文全体から最終的な項候補を選出する。

本稿では分野の異なるテキストに対する項同定性能を比較するため、述語は既知、かつ文内に項が存在する事例のみをトレーニングとテストに用いた。[†2]

4　異なる分野のテキストを用いた述語項構造解析の比較実験

2 つの異なる分野における述語項構造タグつきコーパスを使用し、機械学習に基づく述語項構造解析器を用いて文内のガ格の項同定性能の比較実験を行なった。まず、それぞれの分野のコーパスでトレーニングした解析器をそれぞれの分野のコーパスでテストすることによって、分野による解析器の性能を評価する。次に、異なる分野のコーパスを追加して再学習することで、分野の異なるコーパスがテストにどのような影響を与えるのか調査する。

4.1　データ

BCCWJ 述語項構造・照応アノテーション[†3] に含まれる BCCWJ のコアデータに対する述語項構造アノテーションのうち、新聞記事 (PN)、Yahoo!知恵袋 (OC) のコーパスを用いて実験を行う。新聞記事データは 2012 年 4 月 4 日版のスナップショット、Yahoo!知恵袋データは BCCWJ 述語項構造・照応アノテーション v0.2（2012 年 10 月 5 日版）を用いた。

新聞記事データは 8,998 文、Yahoo!知恵袋データは 6,321 文である。

[†1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[†2] 述語と句の間の係り受け関係の有無の判定には、人手による係り受け情報ではなく、自動解析結果を用いた。

[†3] http://cl.naist.jp/nldata/bccwj/pas/

表 1　文内ガ格の項同定実験に用いた素性

素性の種類	素性	詳細
文法的	品詞	“名詞-固有名詞”, “名詞-サ変接続” のような項候補の品詞
	格助詞	“は”, “が”, “を” のような項候補に続く助詞
意味的	固有表現	“ORGANIZATION”（組織名）などの固有表現の種類
	生物	項候補が生物である場合に 1
位置的	距離	項候補と述語との間の距離（... , -2, -1, 1, 2, ...）
	文頭	項候補が文頭にある場合に 1
	係り受け	項候補が述語と係り関係にある場合に 1
	連体	項候補が連体節の中にある場合に 1

4.2　ツール

形態素情報は BCCWJ コアデータに含まれる人手修正済みのデータを用いた。また、形態素解析済みのデータを入力として CaboCha 0.65[†4] UniDic モデルを用いて自動的に文節区切り・係り受け解析・主辞解析・固有表現解析を行った。分類器には LIBLINEAR 1.92[†5]を用いた。

4.3　素性

使用した素性は [飯田 04] の素性を基に一部改編したもので、表 1 に示す。これは元の素性から EDR 概念辞書や日本語語彙大系を用いる素性と項と述語の共起、項候補と照応関係にある名詞句の数と Salience Reference List を用いた素性を省いたものである。

4.4　評価尺度

項同定精度の評価には、適合率 (P)・再現率 (R)・F 値を用いた。$P = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}, R = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}, F = \frac{2 \times P \times R}{P+R}$ である。

5　結果と考察

5.1　素性の学習結果

各分野によって学習された素性の一部とその重みの絶対値を表 2 に示す。これらはトーナメントモデルにおいて句のペアから項候補を選ぶ際の選ばれやすさに影響する素性の一部である。素性は特に影響しているものとして句の主辞の品詞と名詞の分類と、直接の係り受け関係を示している。

[†4] https://code.google.com/p/cabocha/
[†5] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 2　項候補の選択に影響する素性と重み

OC				PN			
文頭側		文末側		文頭側		文末側	
代名詞	0.883	代名詞	0.395	代名詞	0.780	代名詞	0.350
名詞	0.154	名詞	0.082	名詞	0.476	名詞	0.166
固有名詞	0.556	固有名詞	0.325	固有名詞	0.493	固有名詞	0.481
普通名詞	0.345	普通名詞	0.209	普通名詞	0.203	普通名詞	0.179
直接係受	0.445	直接係受	0.665	直接係受	0.470	直接係受	0.709

5.2　項同定精度

分野の異なるテキストで文内ガ格の項同定の訓練・評価の違いを見るための比較実験の結果を、述語と項が係り受け関係があるかどうかで分け、表 3、表 4 に示す。

述語と項が係り受け関係にある場合、最も F 値が高くなるのは PN で訓練、PN で評価を行った場合となった。また、OC で訓練、PN で評価の場合が OC で訓練、OC で評価の場合よりも高い。PN は比較的 1 文あたりの項候補の数が多いと考えられるが、これらの結果から今回の実験設定では一番よく解析できる分野であることがわかる。一方で最も F 値が低くなったのは PN で訓練、OC で評価を行った場合となった。

PN で評価した場合について訓練時の各分野を比較しても訓練と評価を合わせた方が良い結果が得られることがわかり、OC で評価した場合についても同様である。

PN で訓練した場合について評価時の各分野を比較すると訓練と評価を合わせた方が良い結果が得られることがわかる。一方で、OC で訓練した場合については係り受け関係にある場合に PN による評価が良い結果を出しており、PN がこの場合よく解析できる分野であることがわかる。

また、OC の訓練データに PN の訓練データが加わると、OC の分野ではほぼ精度変化が見られなかった。一方で PN の訓練データに OC の訓練データが加わると、PN の分野では精度が低下した。PN の分野においては OC の訓練データがノイズになっている可能性が考えられるが、どのような特徴がノイズになっているのかの検討は今後の課題である。

述語と項が係り受け関係にない場合、OC で訓練した場合について評価時の各分野を比べると、係り受け関係ありの場合と異なり PN よりも OC で評価した場合の方が F 値は高くなり、係り受け関係がない場合は PN もあまり解きやすい分野ではなくなることがわかる。また、OC の訓練データに PN の訓練データが加わると、OC の分野では係り関係ありの場合と異なり精度の向上が見られた。今回の実験では項と述語の共起や意味カテゴリ、表層に関する情報を用いていないため、係り受け関係にない場合に項同定を行うための手がかりが不足している可能性が考えられるが、これらの素性を実装することによって、どのように傾向が変わるか調べるのは今後の課題である。

表 3　文内ガ格の項同定精度（述語と項が係り受け関係にある場合）

		評価					
		Yahoo!知恵袋 (OC)			新聞記事 (PN)		
		P	R	F	P	R	F
訓練	OC	0.690	0.901	0.781	0.722	0.877	0.792
	PN	0.665	0.908	0.767	0.777	0.917	0.841
	OC+PN	0.690	0.900	0.781	0.695	0.875	0.774

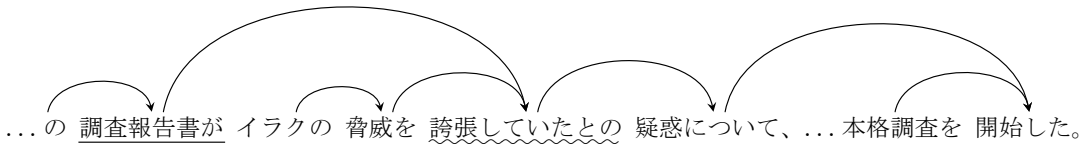
表 4　文内ガ格の項同定精度（述語と項が係り受け関係にない場合）

		評価					
		Yahoo!知恵袋 (OC)			新聞記事 (PN)		
		P	R	F	P	R	F
訓練	OC	0.462	0.533	0.495	0.388	0.596	0.470
	PN	0.426	0.449	0.437	0.524	0.698	0.598
	OC+PN	0.464	0.586	0.518	0.410	0.578	0.480

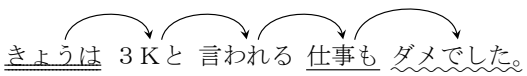
5.3　事例分析

以下に解析結果を示す。下線が正解の項、二重下線が不正解の句、波線が述語を表す。

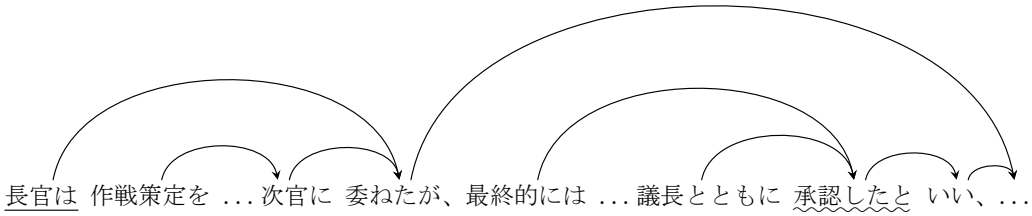
■述語と項が係り受け関係にある場合　これは訓練と評価にともに新聞記事を用いた事例である。PN 分野では物体は組織よりも動作主になりにくいと考えられるが、この事例では述語と項が係り受けの関係にあるため、ガ格の表層格である近くの項を捉えられている。



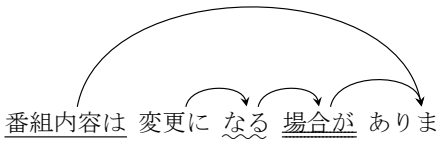
次の事例は、PN で訓練し、OC で評価した事例である。この場合、述語と項の間に係り受け関係があるにも関わらず、正解を出力できなかった。PN と比べ、OC に対する自動係り受け解析結果は頑健ではない可能性があるが、これが述語項構造解析にどのような影響を与えているかは、引き続き調査する必要がある。



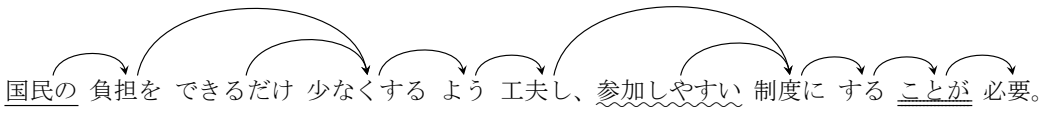
■述語と項が係り受け関係にない場合　以下の事例は訓練と評価にともに新聞記事を用いた事例である。述語と項が離れているが、PN 分野では人名である場合項になりやすいという傾向を学習しているため、文頭の人物を捉えられている。



次の事例も訓練と評価にともに新聞記事を用いた事例であるが、正解を出力できなかった。係り受け関係にある表層格がガ格である近くの句を選んでしまったと考えられる。自動係り受け解析に失敗しているために項同定も失敗している可能性があるので、今後は人手でアノテーションされた係り受け情報[†6]を用いた実験を試したい。



次の事例もまた訓練と評価にともに新聞記事を用いた事例であるが、正解を出力できなかった。「参加する」という述語と「国民」「こと」それぞれの選択選好に関する素性を用いることで、正解を選べるようになる可能性があると考えられる。



6　おわりに

本稿では、複数の分野のテキストで機械学習に基づく述語項構造解析器を訓練し、分野の異なるテキストでの性能を評価した。評価の結果、分野の異なるテキストを訓練に用いた場合、性格の異なるモデルを学習することが分かった。また、訓練と評価に用いるデータの分野が同じ場合が最も文内ガ格の項同定精度が高いことが確認できた。一方、訓練に新聞記事と Yahoo! 知恵袋のデータを両方用いる効果は限定的で、特に評価に新聞記事を用いた場合、Yahoo!知恵袋のデータを新聞記事に加えて訓練すると、係り受け関係にある場合もない場合も大きく精度が下がることが分かった。

今後は共起素性など意味的な素性、そして表層に関する素性を入れた場合にどのように結果

[†6] https://sites.google.com/site/masayua/bccwjdep

が変わるか調査するとともに、白書や書籍といった他の分野のテキストにおいても分野の影響がどのようにあるのか検討してみたい。

参考文献

[Ima09] Imamura, K., K. Saito, and T. Izumi: Discriminative Approach to Predicate-Argument Structure Analaysis with Zero-Anaphora Resolution, in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 85–88, 2009.

[Kaw02] Kawahara, D., S. Kurohashi, and K. Hasida: Construction of a Japanese Relevance-tagged Corpus, in *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 2008–2013, 2002.

[Tai08] Taira, H., S. Fujita, and M. Nagata: A Japanese Predicate Argument Structure Analysis using Decision Lists, in *Proceedings of EMNLP-2008*, pp. 523–532, 2008.

[Yos11] Yoshikawa, K., M. Asahara, and Y. Matsumoto: Jointly Extracting Japanese Predicate-Argument Relation with Markov Logic, in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 1125–1133, 2011.

[橋本 11] 橋本, 黒橋, 河原, 新里, 永田：構文・照応・評判情報つきブログコーパスの構築, 自然言語処理, Vol. 18, No. 2, pp. 175–201, 2011.

[笹野 11] 笹野, 黒橋：大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3328–3337, 2011.

[小町 11] 小町, 飯田：BCCWJ に対する述語項構造と照応関係のアノテーション, 日本語コーパス平成 22 年度公開ワークショップ, pp. 325–330, 2011.

[小椋 09] 小椋, 小木曽, 小磯, 冨士池, 宮内, 渡部, 竹内, 小川, 小西, 原, 中村：「『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予稿集』, pp. 57–64, 2009.

[飯田 04] 飯田, 乾, 松本：文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定, 情報処理学会論文誌, Vol. 45, No. 3, pp. 906–918, 2004.

[飯田 10] 飯田, 小町, 井之上, 乾, 松本：述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から, 自然言語処理, Vol. 17, No. 2, pp. 25–50, 2010.

「了解」の意味の変遷—19 世紀末から現代にかけて—

中山健一(筑波大学／東京外国語大学)

The Change in the Meaning of *Ryookai*

: From the End of the 19th Century to Current Japanese

Kenichi Nakayama (Univ. of Tsukuba / Tokyo Univ. of Foreign Studies)

1. はじめに

現代日本語(東京方言)における「了解」は、次の例のように、他者の行為、あるいは要求・申し出などを理解し、承認・承諾するという意味で使われることが多い。

(1) その日から、私は内藤と朴との王座決定戦を実現するために動きはじめた。私がまず第一にしなくてはならなかったのは、内藤をはじめとする全員に了解をとることだった。これには問題がなかった。私の説明に対して、朴が若く未知のボクサーであることへの不安は表明されたが、決定戦そのものへの反対意見は出されなかった。(沢木耕太郎　一瞬の夏)

しかしながら、明治大正期の書き言葉では、現代語では「理解」を使うような文で「了解」が使われていることが多い。例(2)は、物事の理解を表しており、承認・承諾の意味は含まれない例である。

(2) 幸にして先生の予言は実現されずに済んだ。経験のない当時の私は、この予言の中に含まれている明白な意義さえ了解し得なかった。(任意採集　夏目漱石　こころ)

このように、明治大正期の「了解」の意味と、現代語での「了解」の意味には、違いがみられるようである。本発表は、「了解」の 2 つの意味、便宜的に他の語へ言い換えるのであれば、「理解＋承認」の意味と「理解のみ」の意味の 2 つの意味の、どちらの意味で使われるのかについて、コーパスを用いて、通時的な調査を行なうことを目的とする。

2. 先行研究—辞典類の記述—

管見のかぎり、「了解」の語義、およびその変遷に関する論考は見当たらなかった。本発表では、先行研究における「了解」の語義の捉え方として国語辞典の記述を挙げる。同時に、複数の漢字表記の扱いについてもまとめる。

中型国語辞典として、ここでは「学研国語大辞典」と「大辞林」の語釈を挙げる。

・「学研国語大辞典」

りょうかい【了解】【諒解】【領解】【領会】≪名・他サ≫　物事の筋道・理由・意味などをよくのみこむこと。さとること。また、理解して承認すること。「ーに苦しむ」「伸子がその作を書いた衷心の事情が分れば、あるーが得られるだろうと＜宮本・伸子＞」「議決権なきものとーします＜城山・総会屋錦城＞」　(類)了承

・「大辞林」
りょうかい【了解】【諒解】(名)スル[1]　①事情を思いやって納得すること。理解すること。のみこむこと。了承。領解。領会。「事情をーする」「ーできない」
②無線などの通信で、通信内容を受け取ったことを表す語。「『ただちに行動を開始せよ』『ー』」
③［哲学の専門用語　略］
※加えて、空見出し「「了解①」に同じ」として「りょうかい【領解】」と「りょうかい【領会】」を立てている。

以下、これら 2 つの辞典の記述をもとに、語義、漢字表記、品詞についてまとめる。

まず、語義について、「学研」では、「理解」としての意味を挙げたのち、「また､」として「理解＋承認」の意味を挙げている。「大辞林」では「①」の意味では「理解」としての意味の説明のみだが、別語への言い換えの 1 つとして「了承」を挙げている。

このように、国語辞典の記述では、本発表で問題とする「了解」の 2 つの意味を別義とは捉えていないものの、両方に対して言及がある。

次に、漢字表記について、【了解】【諒解】【領解】【領会】の 4 つの表記が挙がっている。これら 4 つの表記での意味の違いには言及はない。

最後に、品詞について、語釈の前の品詞情報にあるように、名詞、および、「了解する」という形での動詞として使われるのが主である。それ以外に、「大辞林」の「②」の語義のように、感動詞的に使われることがある。

3. 調査の観点・方法

2 節を踏まえて、本発表の調査の観点と方法を述べる。

3.1 「意味」の捉え方

一般に、ある 1 つの語が性質の異なる物事を指し示しうる。その場合「意味」が違う、つまり別義といえる場合も、同じ「意味」として括れるが用法・ニュアンスが違うといえる場合もある。しかしながら、両者の線引きは容易ではなく、線引きの方法について明確な答えを発表者は持っていない。

本発表では、先の例(1)のような使われ方と例(2)のような使われ方のどちらで使われているか、通時的調査を行なうという目的にかんがみ、「了解」の、「理解」としての使われ方と「理解＋承認」としての使われ方を、「意味」の違いとして論を進めることとする。

3.2 表記の違い

2 節でも述べたように、表記としては、【了解】【諒解】【領解】【領会】の 4 つが考えられるが、それぞれを別語ではなく、同一語の異表記として扱う。ただし、表記の違いが本発表で問題としている意味の違いと相関があることも考えられるので、実例分析は表記を区別したうえで行ない、意味との相関の有無を調査する。

以下では、「了解」とカッコ書きにした場合には、「語」を表わすものとし、上記 4 つの表記の代表形として扱う。それに対し【了解】【諒解】など、墨カッコを使う場合には、「表

[1] 「品詞情報」の「(名)スル」は、名詞のうちサ変動詞としても使われるものを表わす。

記」を表わすものとし、それぞれの異なる表記を指すものとする。

3.3 品詞の違い

品詞の違い(名詞か、動詞か)は、意味の違いと相関がある可能性がある。そのため、実例分析は、品詞を区別して行なう。名詞・動詞以外に 2 節で挙げた「大辞林」の「②」の語義のように、感動詞的に使われることがある。これは無線通信の場合に限らず、日常的な話し言葉でもよく使われる。本発表では、この種のものを「大辞林」のようにまったくの別義として捉えることはしないが、他の例とは区別して扱う。

3.4 対象とする年代と媒体

調査対象は、明治後期以降とする。それ以前の言語資料は調査対象とすることができなった。また、媒体は、書かれた言語資料とする。書き言葉に限定する理由として、言語資料(コーパス)の入手のしやすさという調査環境の要因もあるが、「了解」自体が文章語であり、書かれた言語資料に多用されると考えられるからである。使用したコーパスについては、次節で述べる。

4. コーパス

調査対象とした言語資料は以下の 2 つである。(ただし、任意採集の例(2)を除く。)

・国立国語研究所 編(2005)『太陽コーパス ―雑誌『太陽』日本語データベース― 国立国語研究所資料集 15』博文館新社 (以下「太陽コーパス」と呼ぶ)

・新潮社 編(1995)『新潮文庫 100 冊 CD-ROM』新潮社・ボイジャー・NEC インターチャネル (以下、「新潮文庫の 100 冊」と呼ぶ)

明治後期～大正の言語資料として、「太陽コーパス」を使用した。「太陽コーパス」に収められている資料の発行年は、1895(明治 28)年、1901(明治 34)年、1909(明治 42)年、1917(大正 6)年、1925(大正 14)年の 5 つの期間である。市販版のものを使用し、付属の検索ツール「ひまわり」で前述の 4 つの表記を検索し、実例の抽出を行なった。和語動詞「わかる」への当て字など、今回の調査対象外のものは手作業で削除した。

次に、明治大正期との比較・対照のための現代語の言語資料について述べる。本来であれば「太陽」と同様の総合雑誌の記事からとるべきであるが、資料の性質(雑誌記事か書籍か)の違いは、さほど大きく影響しないと判断し、「新潮文庫の 100 冊」から「太陽コーパス」以降である昭和(1926 年～)の資料を、1945 年より前と 1945 年以後に分けて調査した。明治大正期のもの、および、翻訳は除外した。「新潮文庫 100 冊」からの実例の抽出は、市販版を、小木曽智信先生(国立国語研究所准教授)が公開している「新潮文庫 CD-ROM コンバータしおまめ」を使い変換し、「ひまわり」で検索した。具体的な方法は「太陽コーパス」と同様である。

実例数を以下の表 1 にまとめる。それぞれのコーパスおよび年代区分で、実例の数はまちまちである。「新潮文庫の 100 冊」の昭和戦前は合計 4 例と極めて少なく、扱いに注意が必要であろう。

表 1　実例数(全体)

	太陽コーパス						新潮文庫の 100 冊		
年代	1895	1901	1909	1917	1925	合計	1926-44 昭和戦前	1945- 昭和戦後	合計
実例数	20	41	101	103	72	337	4	93	97

5. 調査結果

以下、まず 5.1 節で実例数など調査結果の概要と、本発表での結論の大枠を示す。つづく 5.2 節で、個々の実例を詳しく検討する。

5.1 概要

まず、意味の問題に入る前に、表記、および、品詞ごとに分けて示す。その後、本題である意味の違いごとに実例数を挙げ、表記、および、品詞との相関の有無を調べる。

表記について、表 2 にまとめる。( )の数字は、それぞれの年代ごとでの各表記の占める割合である。表記について、やはり圧倒的に【了解】が多く、次に【諒解】が多かった。【領解】と【領会】は、「太陽コーパス」にみられるが数は少なく、「新潮文庫の 100 冊」にはみられない。

表 2　実例数(表記別)

コーパス	太陽コーパス					新潮文庫の 100 冊	
年代	1895	1901	1909	1917	1925	1926-44 昭和戦前	1945- 昭和戦後
【了解】	19 (95)	34 (83)	72 (71)	79 (77)	37 (51)	3 (75)	73 (79)
【諒解】	0 (0)	1 (2)	14 (14)	19 (18)	35 (49)	1 (25)	20 (21)
【領解】	0 (0)	6 (15)	12 (12)	3 (3)	0 (0)	0 (0)	0 (0)
【領会(會)】	1 (5)	0 (0)	3 (3)	2 (2)	0 (0)	0 (0)	0 (0)
合計	20 (100)	41 (100)	101 (100)	103 (100)	72 (100)	4 (100)	93 (100)

次に、品詞ごとの数を示す。( )の数字は、それぞれの年代ごとでの各品詞の占める割合である。

表 3　実例数(品詞別)

コーパス	太陽コーパス					新潮文庫の 100 冊	
年代	1895	1901	1909	1917	1925	1926-44 昭和戦前	1945- 昭和戦後
名詞	0 (0)	4 (10)	11 (11)	26 (25)	38 (53)	3 (75)	54 (58)
動詞	20 (100)	37 (90)	90 (89)	77 (75)	34 (47)	1 (25)	37 (40)
感動詞的	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)
合計	20 (100)	41 (100)	101 (100)	103 (100)	72 (100)	4 (100)	93 (100)

品詞としては、名詞(「了解が」「了解を」など)と動詞(「了解する」)の場合が主であった。

加えて、とくに現代語の話し言葉において、相手の指示や要求への肯定の返答として感動詞的に使われる場合もある。次に例を 1 つ挙げる。

(3) 「ブンが、フン先生の家にあらわれたのだ。さ、はやく行け！」 「わかってます。で、フン先生というひとの家の所番地は？」「市川市のはずれに下総の国分寺という有名なお寺がある。そのお寺の裏側の畑の中の一軒家だ」「了解。いってきまーす！」(井上ひさし　ブンとフン)

感動詞的なものは、主に話し言葉で使われるものであり、「新潮文庫の 100 冊」では小説の会話文にごく少数見られた。「太陽コーパス」には 1 例も見られなかった。

今、感動詞的なものは措くとして、名詞と動詞を比較した場合、大まかに言って、19 世紀末から 20 世紀はじめころまでは動詞として使われる場合が大多数を占めていたが、現代(昭和戦後)では名詞が過半数を占めていることがわかる。各年代をみても、実例数が極めて少なく確かなことが言えない「新潮文庫の 100 冊」の昭和戦前を除いて、年代が下るにつれて、動詞が少なく、名詞が多くなるという推移を見せている。

以下、本発表で問題となる、「了解」の意味ごとの実例数を挙げる。これ以降、「理解」としての意味を「意味 A」、「理解＋承認」としての意味を「意味 B」とする。

意味 A か意味 B かの判断基準は、最終的にはテクストの読み手である発表者の判断ということになる。しかし、意味 A、意味 B、それぞれの意味が実現される言語的(さらに狭く言えば、構文的)な条件、つまり、「構文的な構造」(中山 2009)を取り出すことが可能である。むろん、すべての実例で明確なわけではないが、可能なかぎりそれを記述し、5.2 節で挙げることとする。ここでは、それぞれの代表的な例をいくつか挙げる。

・「太陽コーパス」意味 A

(4) 市長が自分の俸給三千圓を減じた眞意は、 どう考へて見てもその當時僕は甚だ了解に苦んだ。(東京市長としての奥田男：1917-13)[2]

(5) 併し博士は生物界に於ける共同生存の意義を充分に了解されて居ない様に見える。(『自然界の三大矛盾』に就て：1909-2)

・「太陽コーパス」意味 B

(6) 中央亞米利加に移民を計畫し、 明治廿七年グアテマラを探險して大統領、 内閣員を訪問、その了解を得て廿七年七月の末に日本へ歸つて來た。(実業界の生活を顧みて：1925-10)

・「新潮文庫の 100 冊」意味 A

(7) ホテルからしばらく歩くと、舗道に三、四十人くらい男たちが坐り込んでいるところに出くわした。意外な光景だったが、彼らが坐り込んでいる建物がデイリー・ニューズの社屋だということで了解できた。ニューヨークは新聞ストの真っ最中だった。(開高健　流亡記)

[2] 記事タイトルの後の数字は、それぞれ雑誌の発行年と号数を示す。

・「新潮文庫の 100 冊」意味 B
(8) その日から、私は内藤と朴との王座決定戦を実現するために動きはじめた。私がまず第一にしなくてはならなかったのは、内藤をはじめとする全員に了解をとることだった。これには問題がなかった。私の説明に対して、朴が若く未知のボクサーであることへの不安は表明されたが、決定戦そのものへの反対意見は出されなかった。(沢木耕太郎　一瞬の夏)
※例(1)再掲

(9) 星はとるものもとりあえず、内務省衛生局へかけつけ、依頼した。 「あの原料阿片の積出しはさしつかえないと、早く小樽に電報を打って下さるよう、お願いします。それによって、小樽水上警察署も了解してくれるはずになっております」(星新一　人民は弱し官吏は強し)

意味ごとの実例数を表 4 にまとめる。

表 4　実例数(意味別)

コーパス	太陽コーパス					新潮文庫の 100 冊	
年代	1895	1901	1909	1917	1925	1926-44 昭和戦前	1945- 昭和戦後
意味 A	20 (100)	41 (100)	98 (98)	89 (88)	50 (75)	1 (25)	22 (24)
意味 B	0 (0)	0 (0)	2 (2)	12 (12)	17 (25)	3 (75)	69 (76)
意味 A＋ 意味 B	20 (100)	41 (100)	100 (100)	101 (100)	67 (100)	4 (100)	91 (100)
判断が難しいもの	0	0	1	2	5	0	2
合計	20	41	101	103	72	4	93

やはり、「新潮文庫の 100 冊」では意味 B が大多数を占めるのに対し、「太陽コーパス」では意味 A が大多数を占めるという結果となった。「太陽コーパス」を年代ごとにみると、1895 年、1901 年では意味 A がすべてであったのが、1909 年になって意味 B がごく少数みられ、1917 年、1925 年と年代が下るにしたがって意味 B の占める割合が増えている。

意味 A か意味 B かの「判断が難しいもの」は、これ以降の分析から除外する。「判断が難しいもの」のうち、際立ったものとして、次のような、人(および組織)どうしの関係が問題となる例である。意味 A に近いと言えば近いが、面識・交流をもつというような意味であろうか。現代語の感覚では、なじまないような文脈に現れている。

(10) 交渉團體を爲さぬ無所屬議員にして發言したる者は、 前に長島隆二君、 今回は押川方義、 林毅陸の二君あるも、 この人々は多少とも政黨に關係を有し、了解を有して居つたから其の便宜を得たのであつて、[後略]　(徹頭徹尾党争の府：1917-9)

(11) 大軌社長――大阪奈良間電車の大槻龍治君は元の税關長時代から大阪三長老の一人で鳴した片岡直輝君と諒解があつた。(大阪唯一の社交団たる大阪倶楽部に集る人々：1925-4)

(12) 申合としては綱領調査委員會を設けるに就て――綱領規約調査委員會はその性質上表面の形式は事務機關であるが、 各團體間諒解の絶好の機會であるから之を利用して諒解に力めること――といふことになつた。(無産政党組織準備委員会の主要団体及中心人物―委員会組織の過程及将来―：1925-12)

以下、意味と表記との相関、および、意味と品詞との相関についてみていく。

まず、意味と表記との相関について、表 5 にまとめる。表 5 は、先の表 4 での「判断が難しいもの」を除外し、それぞれの表記・年代ごとに、意味 A と意味 B の数を並べて示したものである。各セルで、上段の数字の左側が意味 A の実例数、右側が意味 B の実例数である。下段の( )のなかの数字は、意味 A と意味 B の実例数の割合である。

表 5　実例数(表記と意味の相関)

コーパス	太陽コーパス					新潮文庫の 100 冊	
年代	1895	1901	1909	1917	1925	1926-44 昭和戦前	1945- 昭和戦後
【了解】	19/0 (100/0)	34/0 0	70/2 (97/3)	74/3 (96/4)	31/6 (84/16)	1/2 (33/67)	19/52 (27/73)
【諒解】	0/0	1/0 0	13/0 (100/0)	10/9 (53/47)	19/11 (53/47)	0/1 (0/100)	3/17 (15/85)
【領解】	0/0	6/0 0	12/0 (100/0)	3/0 (100/0)	0/0	0/0	0/0
【領会(會)】	1/0 (100/0)	0/0 0	3/0 (100/0)	2/0 (100/0)	0/0	0/0	0/0
合計	20/0 (100/0)	41/0 (100/0)	98/2 (98/2)	89/12 (88/12)	50/17 (75/25)	1/3 (25/75)	22/69 (24/76)

【領解】【領会(會)】は意味 A でのみ使われている。しかし、年代との関わりをみると「新潮文庫の 100 冊」には例がなく「太陽コーパス」でも 1925 年にはない。つまり、意味 B が多く使われ始める年代にはそれらの表記の例自体がない。そのため、これら 2 つの表記が意味 A に限られるというより、意味に関わらずその表記自体が使われなくなった可能性が高い。

【了解】と【諒解】について、「新潮文庫の 100 冊」の昭和戦後では、両者とも、意味 A が圧倒的多数となっている。よって、現代語において、2 つの表記と意味の違いの相関は認められない。「太陽コーパス」では、【了解】と【諒解】とを比較すると、特に 1917 年と 1925 年で「諒解」のほうが意味 B で使われる割合が高い。しかし、【諒解】の実例数自体が少ないこともあり、【諒解】のほうが意味 B で使われやすかったとまでは言えない。

結論として、表記の違いと意味の違いの相関については、本調査では明確なことは言えない。どの漢字表記を使うかは、様々な要因があると考えられる。言語の側の要因のみならず、例えば「諒」は常用漢字ではないなど言語政策との関わりもあるだろう。

次に、意味と品詞との相関について、表 6 にまとめる。それぞれのセルの数字の示し方は先の表 5 と同じで、上段の数字の左側が意味 A の実例数、右側が意味 B の実例数、下段の( )のなかの数字は、意味 A と意味 B の実例数の割合である。

表 6　実例数(品詞と意味の相関)

コーパス	太陽コーパス					新潮文庫の 100 冊	
年代	1895	1901	1909	1917	1925	1926-44 昭和戦前	1945- 昭和戦後
名詞	0/0	4/0 (100/0)	9/2 (82/18)	12/12 (50/50)	17/16 (52/48)	0/3 (0/100)	4/49 (8/92)
動詞	20/0 (100/0)	37/0 (100/0)	89/0 (100/0)	77/0 (100/0)	33/1 (97/3)	1/0 (100/0)	18/18 (50/50)
感動詞的	0/0	0/0	0/0	0/0	0/0	0/0	0/2 (0/100)
合計	20/0 (100/0)	41/0 (100/0)	98/2 (98/2)	89/12 (88/12)	50/17 (75/25)	1/3 (25/75)	22/69 (24/76)

前述(表 3)のとおり、意味に関わらず、動詞がほとんどすべてであったのが、名詞が徐々に多くなるという推移をみせている。加えて、表 6 の通り、品詞と意味との相関においても、名詞と動詞との間に際立った差が見られる。「太陽コーパス」では、動詞においては年代を問わず意味 A がほとんどすべてであるのに対し、名詞においては特に 1917 年以降、意味 B も比較的多くみられる。一方、「新潮文庫の 100 冊」の昭和戦後では、名詞の場合ほとんどすべての例が意味 B であるのに対し、動詞の場合は意味 A と意味 B がほぼ半々で、動詞では、意味 B への移行が名詞よりも遅れていることが分かる。

5.2 意味 A・意味 B それぞれの構文的な構造

以下では、紙幅の都合上、ごく簡単にではあるが、意味 A、意味 B それぞれの実現をささえる構文的な構造を、「太陽コーパス」の個々の実例を挙げながら述べる。

5.2.1 「意味 A」をささえる構文的な構造

・動詞の場合

〔名詞(抽象的な事柄)ヲ　了解スル〕

(13) 併し博士は生物界に於ける共同生存の意義を充分に了解されて居ない様に見える。(『自然界の三大矛盾』に就て：1909-2)　※例(5)再掲

(14) 斯くては到底虚心平氣に韓人の眞相を領解すべき餘地あるべからず。治下人民の性情を領解する能はずして一に獨自己の見解によりて萬般の施設を運らす、運らす所巧妙ならざるにあらざるも、殆んど手答へなく、概ね失敗に了るは見易き道理なり。(政治、外交　統監政治の失敗：1909-6)

(15) 其人もし佛語を知るか試に佛語を以て法學上の事を質問すれば、 氏は聲に應じて佛語を了解す、 必要あらば流暢なる佛語を以て答へらる。(フルベツキ博士とヘボン先生：1985-7)

抽象的な事柄を表わす名詞について、具体的には、社会的な事柄(「英国人の生活状態」「時代の眞相」「その国民の思想」)、科学・学術(「精子の作用」「沈殿岩の特性」「日本画の歴史」)、人の内面(「僧の苦心」「政府の意」)、言語(「佛語」「外國語」)など様々なものが来る。「太陽コーパス」ではこの種の例が圧倒的に多く、「太陽コーパス」全体のほぼ半数の約 150 例を占める。

以下、それ以外の構文的な構造についてまとめる。

〔節(疑問詞疑問文＋カ(ヲ) ―抽象的な事柄や一般的事実―)　了解スル〕
(16) 人民は無識にして未だ憲政の何たるかを了解せざるものが多い。(欧州大戦と露国の革命：1917-5)

(17) 日本は過去の二大戦役に於て戦争の物質的精神的代價の如何なるものかを了解したり。戰争は日本の發展を妨げしを悟りたり。(外人の日本観：1909-5)

〔節(一般的事実)コトヲ　了解スル〕
(18) 古來最も有力に統一を妨げたるものは交通の不便なるに在りしことを適切に了解するものは、 同時に世界統一の大業を促進するの大勢力は實に交通機關の發達なることを自ら了解するならん。(平和と世界の統一(強国論)：1917-4)

(19) 若夫、 普通片々たる記者であるならば、 忽ち倨傲尊大の風をなし、 自己廣告を盛にする場合であるのに、 何等如此の態度なかりしは、 決して三文評論家でない事を諒解せしむるに足る。(故春汀鳥谷部銑太郎君：1909-2)

・名詞の場合
〔名詞(事柄)ハ／節(一般的事実)ハ／節(疑問詞疑問文)＋カ、了解ニ苦シム〕
(20) 市長が自分の俸給三千圓を減じた眞意は、 どう考へて見てもその當時僕は甚だ了解に苦んだ。(東京市長としての奥田男：1917-13)　※例(4)再掲

(21) 是は獨逸の爲めには非常に不利益な譯で、 今後獨逸は果して何國をたよりとする積であるか、 吾輩などは如何も了解に苦しむ、[後略]　(米独国交断絶の側面観：1917-3)

〔名詞(事柄)ハ／節(一般的事実)コトハ　了解ガ　デキル〕
形の上から名詞としたが、動詞としての「了解する」の可能形「了解できる」に近い。

(22) 女性が生む力に惠まれてゐる所以は、 此感情の優越性なるを以てしても了解が出來るであらう。それ故女性は、 先天的に男性よりは美の本質に秀れ、 女性でさへあれば如何なる女性でも、 男性が如何なる男性でも美くしいとは云ひ得ざるに反し、 美くしい點を發見し得るものだと考へてゐるのである。(現代の女性美：1925-1)

5.2.2 「意味 B」をささえる構文的な構造

5.1 節で述べたように実例の数は少ないが、意味 B について同様にみていく。

〔名詞(人など)ノ　了解ヲ　得ル〕
(23) 中央亞米利加に移民を計畫し、 明治廿七年グアテマラを探險して大統領、 内閣員を訪問、 その了解を得て廿七年七月の末に日本へ歸つて來た。(実業界の生活を顧みて：1925-10)　※例(6)再掲

(24) 加藤が憲政擁護運動に参加したのは、 平田(内大臣)の諒解を得た後に決心したのだ。平田は寧ろ憲政擁護を煽動したと言つても可からう。(政界煙話…議会解散か内閣瓦解か…：1925-2)

〔名詞(人など)ノ　了解ヲ　求メル・乞フ〕
(25) 『然らば、 何うすればよいのか。』 と、 岩倉公が云つた。 『諸公の腹一つ。』 『勿論、 獨立國の威嚴を保たなくてはならぬが、 その方策は何うするか。』 『その御覺悟なら、 彼等の干渉を斥けなさるがよい。 不肖大隈、 其の任に當つて、 長崎以來の經過をのべ、 彼等の諒解を求めることに致しても差支へない。』(明治初年外交物語(その五)邪教退治の腹芸：1925-2)

(26) 然るに學校の出身者や關係者は、 何故校葬にしないのか と云つて自分を責め、 その辯解に困らされた位であつた。 當日の夕方穗積陳重さんは態態私の家に來て、 是非校葬にして貰ひたい とのことであつたが、 これにも事情――初め自分の專斷で校葬にするつもりでゐた――を話して其諒解を乞ふたやうな始末であつた。(中央大学経営者としての奥田男：1917-13)

6. まとめと課題

本発表の結論をまとめる。「太陽コーパス」では、「了解」がサ変動詞として使われた場合、年代を問わずほとんど全て意味 A で使われている。その場合、動作の対象は、抽象的な事柄や一般的事実が大多数である。一方、名詞の場合は特に 1917 年以降、意味 B が現れている。「新潮文庫の 100 冊」のうち昭和戦後では、名詞の場合は意味 B がほとんどなのに対し、サ変動詞の場合は意味 A と意味 B とが半々と、意味 A もある程度みられる。このことから、意味の変遷を推察すると、「了解」は 1910 年代頃までは名詞であれサ変動詞であれ意味 A で使われていたが、1920 年前後から名詞の場合で意味 B が生じ、その後、特に名詞の場合を中心に意味 B が広がり優勢となったと考えられる。

本発表で明らかとなった別の事実、(意味と関わらず)動詞としての使用が優勢だったのが名詞としての使用が優勢になりつつあることと、上記の意味変化との理論的な関係については、今発表だけでは明確なことは言えない。

最後に、本調査の問題点として、「太陽コーパス」の調査に重きをおいたため、それ以降の年代の調査がやや不十分であった。特に、もっとも「現代」に近い年代区分が「1945 年以降」であり、すでに 60 年以上の期間がある。「太陽コーパス」以後についても、十分な実例を収集し、年代区分を細かく行ない調査を行なう必要がある。

文　献

金田一春彦・池田弥三郎 編(1988[1978])『学研国語大辞典 第二版』学習研究社
松村明・三省堂編修所 編(2006[1988])『大辞林 第三版』三省堂

中山健一(2009)「動詞「くる」と「いく」の多義構造の違いについて」『コーパスに基づく言語学教育研究報告 1』、pp.191-217、東京外国語大学大学院グローバル COE プログラム コーパスに基づく言語学教育研究拠点

CRF を用いたアニメ関連用語の固有表現抽出

高瀬 真記（東京農工大学　工学部　情報工学科）
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）

Named Entity Recognition for Animation-Related Words Using CRF

Masaki Takase(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

近年，日本のコンテンツ産業は「クールジャパン」という名称のもと注目を集めており，その中でも漫画，アニメーションなどのいわゆるサブカルチャーは，商業的な観点から見ても重要なコンテンツとなりえている．また，アニメーションなどの作品には多くの固有表現が含まれている．それはキャラクターの名前であったり，作中に登場するロボットの名前であったり，作品タイトルそのものであったりである．そして，それら固有表現は商品検索や商品同定，推薦などに利用できると考えられる．しかし，従来の研究ではアニメーション関連用語の固有表現抽出システムは基本的に存在しない．そこで，本研究ではアニメ関連用語に特化した固有表現抽出システムを考える．固有表現抽出手法としては CRF を利用する．

２．関連研究

固有表現抽出は，今まで様々な方法で行われている．その中でも大きく分けるとパターン照合による固有表現抽出と，機械学習による固有表現抽出にわけられる．

パターン照合による固有表現抽出とは，あらかじめ人手で固有表現のパターンを作成し，合致する部分をコーパスから発見することによって行われる固有表現抽出のことである（竹本，福島，山田 (2001)）．パターンとは「さん」や「大学」などの固有表現に付属しやすい文字列を指す．しかし，ルール作成のコストが高く，そこに合致しない固有表現は抽出できないので，助詞を含むタイトルなどが多数存在するアニメ関連用語に用いるのは難しい．

人手でパターンを作成するコストや，更新するコストを解決するために，機械学習による固有表現抽出の研究も行われている．機会学習による固有表現抽出は，学習用のコーパスを用意することで，自動で抽出パターンを学習することができる．機械による手法は SVM(Support Vector Machine)（山田，工藤，松本(2002)）を利用した抽出や文節情報を利用した抽出（中野，平井(2004)）などが存在し成果を上げている．その他にも HMM(Hidden Markov Model)や分類機の逐次適応, CRF(Conditional Random Fields)を利用した固有表現抽出なども一般的である．機械学習の問題点としては人手によるコーパス作成（橋本，乾，村上(2008)）のコストが高いことなどがある．

こうした研究を踏まえ，本稿ではアニメなどサブカルチャーの特殊な固有表現に特化した固有表現抽出について行った．

３．アニメ関連用語

アニメ関連用語の固有表現を抽出する前に，固有表現抽出の対象となるアニメ関連用語を定義する必要がある．本研究ではそれらを「内部の固有表現」と「外部の固有表現」に

分けて定義した．具体的には，それぞれアニメ作品内に登場する固有表現とアニメを制作する製作者などを指す固有表現である．定義した固有表現をそれぞれ表 1 と表 2 に示す．

表 1：内部の固有表現

細かな概念	例
1.1.1 作品内のタイトル	となりのトトロ
1.1.2 作品内のタイトルの略称・通称	トトロ
1.1.3 作品内の人間以外の登場キャラ	ポチ
1.1.4 作品内に登場する必殺技	かめはめ波
1.2.1 作品内の登場人物	宿海仁太
1.2.2 作品内の登場人物の略称・通称	じんたん
1.3.1 作品内の登場神	ゼウス
1.4.1 作品内の登場組織名	現代視覚文化研究会
1.4.2 作品内の登場組織名の略称・通称	現視研
1.4.3 作品内の登場一族	波紋の一族
1.5.1 作品内の登場店	MAHO堂
1.5.2 作品内の登場施設	イゼルローン要塞
1.5.3 作品内の登場サイト	地獄通信
1.6.1 作品内の登場製品	波動エンジン
1.6.2 作品内の登場技術・システム	シビュラシステム
1.6.3 作品内の登場言語	メルニクス語
1.6.4 作品内の固有身分	千翔長
1.6.5 作品内の固有職業	ヴァンパイアハンター
1.7.1 作品内で発生する事件	セカンドインパクト
1.7.2 作品内に登場する計画	人類補完計画
1.7.3 作品内に登場するキャラの特殊行動・任務	オペレーショントルネード
1.8.1 作品内に存在する固有物質	飛行石
1.8.2 作品内に存在する固有生物・種族	アーヴ
1.8.3 作品内に存在する固有の部位	空識覚器官
1.9.1 作品内に存在する固有病	黒鉄病
1.9.2 作品内に登場する特殊能力	幻想殺し
1.9.3 作品内に登場する特殊状態	スーパーサイヤ人

表 2：外部の固有表現

細かな概念	例
2.1.1 作品の製作者	石原立也(監督)
2.1.2 作品の原作関係者	谷川流(原作者)
2.1.3 作品製作関係者	杉田智和(声優)
2.2.1 作品の製作会社	京都アニメーション
2.2.2 作品の放送会社	TOKYO MX
2.3.1 作品の関連商品	まんま肉まん
2.3.2 作品の関連商品に関連する会社	コトブキヤ
2.3.3 メディアミックス関連雑誌等	ジャンプスクウェア
2.3.4 作品の関連サイト	アニメの公式サイトなど
2.4.1 作品に関連したイベント	アニメコンテンツエキスポ
2.4.2 作品に関連したプロジェクト	西尾維新アニメプロジェクト

内部の固有表現は作中に出てくる用語，外部の固有表現は作品の製作者や関連商品販売社などを対象としている．

また，アニメ関連用語として，地名は現実世界に存在する実在の地名とかぶることもあり，アニメを対象とした固有表現としにくいため，対象から除外した．

４．アニメ関連用語の固有表現抽出手法

アニメ関連用語の固有表現抽出は，CRF による系列ラベリングで行う．学習用のコーパスをアニメに関連したコーパスにすることでアニメ関連用語に特化した固有表現をおこなう．固有表現のタグには BIOES 形式を使用した．タグの意味は表 3 に示す.

表 3：タグの意味

タグ	意味
B	その形態素が固有表現の始まりであることを示す
I	その形態素では固有表現が継続していることを示す
E	その形態素で固有表現が終わることを示す
S	その形態素一つで一つの固有表現であることを示す
o	その形態素は固有表現ではないことを示す

利用した素性は「表層」「品詞」「品詞細分類」「文字種」「文字数」の五つである．入力された文章を形態素解析し「表層」「品詞」「品詞細分類」を取り出し，表層から「文字種」と「文字数」を作成する.

５．アニメ関連用語の固有表現抽出実験

提案する手法を用いて，アニメ関連用語の固有表現抽出実験を行った．その際に，形態素解析器として MeCab(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)を, 系列ラベリングに CRF++ (http://crfpp.googlecode.com/svn/trunk/doc/index.html)をそれぞれ用いた.

５.１　実験データ

CRF++に学習させる際の学習用アニメコーパスは自身で作成した．対象とした文章は Wikipedia に記事のあるアニメ作品 50 タイトルのあらすじである. その中に含まれる固有表現を先に述べた定義で抜き出し, タグ付けをした. 文字数は 44829 文字で, 表層数は 26948, 固有表現数は 1570 である．その内，S タグで表される固有表現が 742 個で BIE タグで表される固有表現が 828 個である.

５.２　アニメ関連用語の固有表現抽出実験内容

学習用アニメコーパスを五分割交差検定することで評価した．正解のタグとシステムが出力したタグを比較し，S タグの場合は両者が揃った場合，BIE タグの場合は，最初から最後までタグが揃った場合を正解とした.

５.３　アニメ関連用語の固有表現抽出実験結果

表層，品詞，品詞細分類の三つの素性を使った状態の結果をベースラインとし，文字種，文字数，文字種＋文字数の組成を使った状態の結果と合わせて，全部で四種類の素性の組み合わせで出た結果は表 4 のようになった.

表 4：組成ごとの結果

テンプレート	再現率	精度	F値
ベースライン	0.655	0.842	0.737
文字種の素性を追加	0.682	0.844	0.754
文字数の素性を追加	0.667	0.846	0.746
文字種と文字数の素性を追加	**0.687**	**0.848**	**0.759**

全ての値は，文字種＋文字数を使った場合に最大となり，その際，S タグで表される固有表現の精度に関しては，ベースラインから見て 5%有意水準で有意という結果が出た.

６．考察

実験結果から，アニメ関連用語は文字種と文字数の素性を使うとより抽出できていることが分かる．これは，アニメ関連用語には片仮名の単語や，漢字の組み合わせのような単

語が多いことが要因としてあげられる．「スターライトブレイカー」のような形で表される単語は，片仮名であり，さらに表層が区切られにくいため，文字数も多くなりがちである．そういったヒントから，文字種と文字数はアニメ関連用語の固有表現抽出において有用なヒントとなりうると考えられる．しかし，「あの日見た花の名前を僕達はまだ知らない。」や「ジャングルはいつもハレのちグゥ」など，むやみに長いタイトルはうまく抽出できていなかった．この実験では前後二行の素性しか見ていないために，普通の文章と区別がつけられなかったと考えられる．しかし，前後の数を増やすと結果が悪くなる傾向があったので，上手く識別するための素性の改良は今後の課題である．

全体的な結果を見ると，文章が柔らかく，良い結果が出にくい Web 文章を用いたコーパスでの再現率 0.687，精度 0.848，F 値 0.759 という値はよい結果であり，この手法はアニメ関連用語の固有表現抽出に有効である．

７．まとめと今後の展望

本研究では，アニメ関連用語の固有表現抽出を CRF にておこなった．

アニメ関連用語を定義したのちに，自身で学習用アニメコーパスを作成．その学習用アニメコーパスを「表層」「品詞」「品詞細分類」「文字種」「文字数」の素性を持つ形にして，BIOES タグを付け，CRF++に学習させた．アニメ関連用語の固有表現抽出実験を行った結果，BIOES タグによる固有表現の正解率の F 値 0.759 という値を出した．その結果からアニメ関連用語の固有表現抽出にこの手法は有効である．今後，固有表現タグのついた BCCWJ コーパスを利用して，システムの性能をより高めていく予定である．

謝　辞

本研究では，固有表現タグのついた BCCWJ コーパスを参考に素性の設計などを行いました．快くデータをくださった橋本泰一先生に感謝します．

文　献

竹本義美，福島俊一，山田洋志(2001)『辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出』情報処理学会論文誌，Vol42，No.6，pp. 1580-1591

山田寛康，工藤拓，松本裕治(2002)『Support Vector Machine を用いた日本語固有表現抽出』情報処理学会論文誌，Vol.43，No.1，pp.44-53

中野桂吾，平井有三(2004)『日本語固有表現抽出における文節情報の利用』情報処理学会論文誌，Vol.45，No.3，pp. 934-941

橋本泰一，乾孝司，村上浩司(2008)『拡張固有表現タグ付きコーパスの構築』情報処理学会研究報告 2008，pp. 113-120

関連 URL

MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

CRF++: Yet Another CRF toolkit
http://crfpp.googlecode.com/svn/trunk/doc/index.html

外来語使用における言語外的要因の分析
—書き言葉コーパスの利用可能性—

久屋 愛実（オックスフォード大学院 言語学博士課程）[†]

Analysis of Language External Effects on the Use of Loanwords: The Potential of Written Corpus-based Studies

Aimi Kuya (Faculty of Linguistics, Philology and Phonetics, University of Oxford)

０．はじめに

現代日本語において、和語や漢語からなる類義の既存語（以下、単に「既存語」）があるのにもかかわらず頻繁に利用され定着している外来語は数多くある。本稿では、外来語を既存語の社会言語学的語彙変異形とみなし、「ケース」を事例として、『現代日本語書き言葉均衡コーパス』(2011)（以下、BCCWJ）においてその使用に影響すると思われる言語外的要因を調査し、記述することを試みる。以下では、まず本研究の目的を簡単に述べ（第 1 節）、外来語を語彙のバリエーションとして研究するための手法を概観したうえで（第 2 節）、BCCWJ における「ケース」の使用について、書き手の生年代、性別、媒体、スタイルの 4 要因をとりあげ、その影響を検証する（第 3 節）。

１．本研究の目的

外来語の研究においては、語彙調査などの定量的調査がすすむにつれて、日本語における外来語使用が全体としてどう変化してきたかという量的概観が可能になった。こうした研究は、日本語における語種としての外来語の全体像を、いわばマクロにとらえる試みであった。しかし、茂木(2012)が指摘するように、現代日本語で定着を見せている基本的外来語の意味・文法的研究、いわばミクロなレベルでの研究は、まだ不十分なようである。こうした流れを汲み、金(2011)は、20 世紀後半の新聞コーパスで増加している外来語は類義の既存語をもつ抽象名詞[1]に多いことを指摘した。金はさらにその中で「基本語化[2]」したいくつかの外来語の例を挙げ、それらの語が既存語の存在にも関わらずなぜ基本語化するに至ったのかを、意味・用法の側面から通時的に分析している[3]。

金(2011)の研究は、外来語の基本語化を類義語と対照しながら言語内的に説明することが目的であったのだが、本稿では金では言及されなかった外来語使用の言語外的要因について記述する。そのために、ひとつのアプローチとして、外来語を既存語の社会言語学的語

[†] aimi.kuya@ling-phil.ox.ac.uk

[1] この中で、金は日本語で増加している外来語名詞には「具体名詞」と「抽象名詞」の 2 つのタイプがあり、前者が近代化などの理由で外国語から借用され日本語における使用が増えた語群（「テレビ」「ホテル」など）であるのに対し、後者は和語や漢語の類義語があるにもかかわらず生じた語群（「タイプ」「トラブル」「ケース」など）であることを指摘している。これは Myers-Scotton(2006)がいうところの、”Cultural borrowing”と”Core borrowing”という借用語の概念区分とほぼ一致するように思われる。前者は受け入れ言語側の既存語彙では言い表すことができないような事物や概念を表現するために借用され、科学や技術の分野でその多くの例を見ることができる。それに対して、後者は受け入れ言語側の語彙に類似の表現が存在するにもかかわらず、借用されるものである。本稿でも、日本語におけるこうした外来語のタイプの違いを認識したうえで、後者、つまり既存語をもつ外来語に焦点を当てて調査することを前提としている。

[2] 金によれば、「基本語化」とは、当該語彙の周辺部から中心部へと移行して基本語彙（一定の言語使用域において広範囲・高頻度に用いられる語彙）へと仲間入りすることである。

[3] 金は既存語をもつ 2 つの外来語「トラブル」と「ケース」を挙げ、類義語の意味・用法とも比較しながら、両者が新聞における基本語としての地位をどのように確立していったのかを詳細に記述している。

彙変異形（lexical variant）として扱い、外来語の生起率と社会的属性、媒体、スタイルとの関係を「ケース」を事例として見ていく。語彙のバリエーションとしての外来語と、社会属性との関係に注目した研究としては、外来語に対する意識調査研究がある（田中(2007)）。そこでは「キャンセル／解約／取り消し」と「ハッピー／幸福／幸せ」を事例として、この中でどの表現を使いたいかが聞かれており、外来語を使いたいという意識は年齢層により段階的な差があることが例証された。本稿ではコーパスを用いて、実際の言語使用においてこうした属性差を観察することができるかどうか、さらに調査を行う。これにより、先行研究とあわせてより幅広い視野で外来語を捉えることが可能になると思われる。

２．方法論

２．１．語彙のバリエーションとその一変異形としての外来語

本稿では外来語を、和語や漢語からなる既存語の社会言語学的語彙変異形とみなすのだが、語彙を社会言語学的バリエーションとして扱うことの難しさは、意味というものが介在してくるところにある。そもそも、あるものをバリエーションとして扱うことができるのは、変異形の間で交替が起こっても意味の変化が生じないという条件を満たす場合であり、そのもっとも良い例は音韻交替[4]である（Labov(1972)）。一方、語彙交替の場合は、意味の変化が生じないような環境を特定することは容易ではない。なぜなら、それぞれの語が異なる用法や文脈においてもつ意味合い、ニュアンス、語感などが微妙に異なることは往々にしてあるからである。この問題の解決策として、Lavandera(1978)は、「意味的な等価性（semantic equivalence）」を厳密な意味で適用する代わりに、「機能的等価性（functional equivalence）」を条件として変異形の交替環境を定めることを提案している。この概念に基づけば、辞書的な類義性をベースとして、それぞれの類義語が文で同じような機能を果たす場合に、それらを変異形と認めてよいことになる。

語彙のバリエーション研究は、以上のような理由からバリエーション研究の枠組みの中ではあまり手がつけられていないが、Ito and Tagliamonte (2003)による英語における程度副詞（intensifier）のバリエーション研究など、少しずつ研究実績が増えてきている。本稿でも、機能的等価性を手掛かりとして、外来語を語彙変異形として扱うことを試みる。なお「ケース」とその既存語がもつ機能的等価性をどう特定するかは、2．4 で詳しく述べる。

２．２．BCCWJ

本稿ではコーパスとして BCCWJ を用いる。BCCWJ は、異なる観点から設計された 3 つのサブコーパスから構成され、データ量が全体で 1 億語に上る大規模コーパスである（山崎、他(2012)）。出版サブコーパスには書籍、雑誌、新聞が、図書館サブコーパスには書籍が、そして特定目的サブコーパスには白書、国会議事録、教科書、「Yahoo!知恵袋」や「Yahoo!ブログ」などのインターネット上に投稿された書き言葉が含まれ、様々な媒体から抽出された現代日本語書き言葉データにより構成されている。

今回は分析対象としてこの中の出版サブコーパスを利用する。出版サブコーパスは、2001 年から 2005 年までのあいだに国内で出版された書籍・雑誌・新聞を母集団とし、そこからランダムにサンプリングされたデータおよそ 3,600 万語（短単位）からなるコーパスである。このうち、書籍は 2,954 万語、雑誌は 569 万語、新聞は 88 万語のデータからなる（山崎、他(2012)）。のちのち媒体差を分析することを考え、書籍と雑誌に比べて極端に規模の小さい新聞データは分析の対象外とした。また、データには固定長と可変長があるが、今回はなるべくたくさんのデータから調べるために分量の多い可変長データを利用した。書籍と雑誌の可変長データの合計は、およそ 3,500 万語相当となる（山崎、他(2012)）。

本稿で BCCWJ 出版サブコーパスを採用した理由は、1）他の語種より基本的に出現頻度が低い外来語でも、定量的調査に耐えうるサンプル量を抽出できる規模を有すること、2）

[4] 例えば、*fourth floor* は、[r]音が発音されてもされなくても、意味の変化が起こることはない。

社会言語学的調査に必要となる書き手の社会的属性情報が、一定量のサンプルから取り出し可能であること、3）多様な媒体からの書き言葉データを含み、媒体間の比較が可能であることの 3 点に集約できる。

２．３．分析対象語

本稿では、金(2011)の研究と関連付けるために、やはり事例として「ケース」を選択した。これに加え、「ケース」を選択したのは、外来語の中では定量的調査に耐えうるような高頻度語であるという点、また、性別や媒体の影響を見るという目的に照らして、特定の性別や分野に偏った使用が見られないような一般的な語であるという点で、この語が本研究の目的に合致すると考えられるからである。

「ケース」に関しては、金(2011)の事例研究においてその類義語の詳細な選定が行われており、今回はそこで選ばれた「場合」「例」「事例」の 3 語を類義語として採用した。つまり、「ケース」とこれらの既存語は「特定の環境において交替しうる語彙変異形である」とみなす。この「特定の環境」については、次で詳しく述べる。

２．４．分析対象となる環境

2．1 で述べたとおり、「ケース／場合／例／事例」の 4 語を社会言語学的変異形として扱うためにはこれらが機能的等価性を保つような環境を特定することが不可欠である。以下、「ケース」と既存語との交替を可能にする環境はどのような環境であるかを定義する。

金(2011)は、「ケース」がいわゆる「コト」に代表される形式名詞的な用法をもつことを指摘し、その中でも、叙述文が表現する内容を客観的な事柄として名詞化する「客観的同格連体名詞」としての用法を手がかりに、これら 4 語の類義関係を認めている。そこで、本稿ではまず、4 語が機能的等価性をもつのは、文中で形式名詞として「ある内容を客観的事柄として名詞化する」とき、と定義する。これにより、金のいう「場合」の「仮定条件」的用法（1）や「提題」的用法（2）、「例」の単なる「例示」的用法（3）、各語に特有な慣用的用法（4）は排除される（下記の例はいずれも金(2011)から引用）。

（1）賃上げが難しい場合は雇用延長など別のテーマで交渉する“選択”の時代になったと問題提起している。
（2）アサガオの場合、～
（3）例として、～／～を例に挙げ、～
（4）場合によっては／そんなこと言ってる場合かっ／例の悪名高い作家／例によって無言で打つ

形式名詞はふつう修飾語をとるが、金によれば、これら 4 語は「コト」よりもやや具体性のある名詞であるため、とくに修飾語をとらなくてもよいという。これら 4 語がとりうる形式は 4 つあり、修飾部をとらず単独で（5）、合成語の構成要素として（6）、名詞句における被修飾語として（7）、連体修飾節構造における被修飾語として（8）文中に現れる（下記の例は金(2011)から「ケース」の用例を代表として引用）。

（5）しかし、女性の平均賃金は男性より低いため、男女の賠償額にケースによっては 1000 万円近い差が生じている。
（6）ケース１、テストケース、レアケース、重症ケース、脳死・虐待ケース
（7）いじめなどのケース、マドンナさんのケース、京都市のケース、今回のようなケース、4 件のケース、悪質なケース、初めてのケース、似たケース、いろんなケース
（8）a. ネット先進国の米国でも、ネット関連企業は苦戦するケースが少なくない。
b. ～（略）まず母親に『癒（いや）し』が必要なケースも多い」と話す。
c. ～（略）しつけの域を超えて繰り返される暴力・ネグレクトが原因のケースに絞っ

て調べた。

以上の例文を見ると、金の指摘する通り、「ケース／場合／例／事例」は、連体修飾語や連体修飾部を伴わない単独用法や、「ケース 1」のように合成語における接頭辞としても文中に現れている。ただし、金によれば、4 つの形式のうち、20 世紀後半の新聞コーパスにおいて「ケース」が最も多く出現するのは、連体修飾節構造（233/327 例）においてであり、名詞句構造（74/327 例）がそれに続くという。以上のことを踏まえて、本稿では「ケース」およびその既存語が、形式名詞として「ある内容を客観的事柄として名詞化」し、さらに名詞句または連体修飾節構造において出現しているものを分析対象とした。

ただし、この定義でも、何格が接続するか、述語は何をとるかなど、語彙によって表現上のばらつきがある可能性がある。そこで、コロケーションという観点からも 4 語の等価性を高めるために、後続する格と述語の種類をさらに絞った。金は、連体修飾節構造において「ケース」が多少・有無・生起・増減・想定・報告・限定・異同・規定・関与・比較の意味をもつ述語表現と結びつくことを指摘した。金の用例整理を参考にすると、これらのほとんどはガ格と結びつく述語であるため、分析対象に含める述語はガ格をともなう多少・有無・生起・増減・想定・報告の意味をもつもの[5]に限定した。つまり（8c）のようにニ格などを伴う例は、分析対象とならない。ただし、ガ格は（8b）のようにハ格やモ格でも現れうるため、それら 2 つの格が後続する場合も含めた。これで、（9）のように、機能・構造（形式）・コロケーションの 3 つの側面からコントロールされた語彙交替環境が特定できた。

（9）分析対象となる語彙交替環境のモデル
a. 機能：形式名詞として「ある内容を客観的事柄として名詞化する」とき
b. 構造（形式）：名詞句または連体修飾節
c. コロケーション：｛ケース／場合／例／事例｝＋｛ガ／ハ／モ格｝
＋｛多少／有無／生起／増減／想定／報告の意味をもつ述語｝

２．５．サンプルの抽出

ここまでできたところで、分析対象となるサンプルを抽出していく。手順としては、まず BCCWJ 出版サブコーパスの書籍・雑誌から、「ケース」、「場合」、「例」、「事例」の 4 語を含むサンプルをすべて抽出した。そこから、（9）で特定した環境で出現しているもの以外を排除した。その結果、条件にあてはまるサンプルは、「ケース」544 件、既存語 2036 件（「場合」1447 件、「例」461 件、「事例」128 件）の合計 2580 件であった。

表 1：形式の違いと「ケース／既存語」の生起率

		ケース	既存語	合計
名詞句	度数	46	242	288
	%	16.0%	84.0%	100.0%
連体修飾節	度数	498	1794	2292
	%	21.7%	78.3%	100.0%
合計	度数	544	2036	2580
	%	21.1%	78.9%	100.0%

[5] 具体例の一部を以下に挙げる（金(2011)より引用）。
多少（多い・少ない・ほとんどだ・珍しい）、増減（増える・減る）、有無（ある・ない・見られる・認められない）、生起（起きる）、想定（想定される・考えられる・予想される）、報告（挙げられる・紹介される・報じられる）など。

表 1 では、抽出した計 2580 件のサンプルを形式ごとに区分している。「ケース」の出現度数は名詞句（46 件）よりも連体修飾節（498 件）において圧倒的に多く、生起率（%）で見ても名詞句（16.0%）よりも連体修飾節（21.7%）において高くなっている。これにより、既存語全体と比べると「ケース」が名詞句よりも連体修飾節構造で多く使用されていることがわかる。

図 1 は、表 1 における既存語 3 語を区別し、それぞれの生起率を「ケース」の生起率と合わせてグラフ化したものである。ここで注意したいのは、既存語それぞれの出現傾向が形式によって異なるということである。図 1 から、「場合」が名詞句構造よりも連体修飾節構造において生起率が高いのと対照的に、「例」と「事例」は連体修飾節構造よりも名詞句構造において生起率が高いことが読み取れる。このように、形式によるそれぞれの語彙の生起率の違いがみられる以上、両形式を混ぜて分析することはよくないと判断し、今回は全体のサンプル数が圧倒的に多く（2292/2580 件）、かつ「ケース」の出現率がより高い連体修飾節構造に限定してさらに詳細な分析を進めていく。

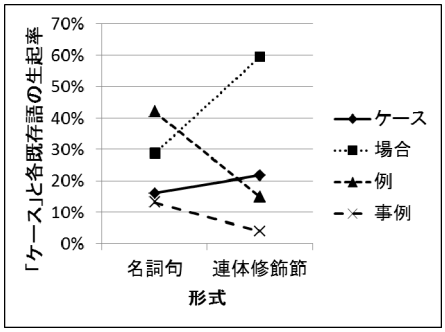


図 1：形式の違いと「ケース／場合／例／事例」の生起率

３．要因ごとの分析結果

本節では、外来語「ケース」の出現に影響を与えうる言語外的要因を選び、要因ごとにクロス表を用いた分析を行う。なお、本稿の目的は外来語が出現する要因を検証することなので、以下、分析をしやすくするために「場合・例・事例」3 語をまとめて「既存語」とし、「ケース」対「既存語」という単純な 2 項対立で見ていく。

３．１．書き手の生年と言語変化

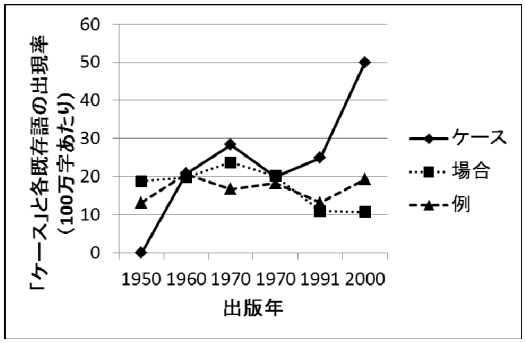


図 2：新聞における「ケース／場合／例」の出現率（100 万字あたり、連体修飾節構造のみ）
（データ元：金(2011)）[6]

[6] グラフは筆者によるもの。金(2011)第 6 章の［表 8］連体修飾節構造の出現率(p.112)のデータをもとにグラフにした 。金は「事例」の用例数が少ないという理由から、年ごとの出現率を出していないため、「事例」はこのグラフでも省かれている。

金(2011)は、20 世紀後半の新聞における「ケース」の出現率（100 万字あたり）が、特に連体修飾節用法において大きく増加していることを通事的に示した（図 2)。この増加が起こった理由のひとつとして、既存語から「ケース」への言葉の使用の変化（言語変化）が起こっている可能性が挙げられる。だとすれば、「ケース」の生起率に、生年による差が存在するのではないか、という予測がたつ。

以上をふまえて、BCCWJ 出版コーパスの書籍と雑誌（2001-2005 年出版）を使い、外来語「ケース」の生起率を書き手の生年という観点から整理し、見かけ上の変化（change in apparent time）が認められるかを検証する。BCCWJ における生年情報は、1930 年代、1940 年代など 10 年刻みで公表されており、それを利用して生年代が最も早いグループ(~1939)、中間のグループ（1940-1959)、そして最も遅いグループ（1960-1979）の３つに区分した。

表 2：生年代の違いと「ケース／既存語」の生起率

		ケース	既存語	合計
～1939	度数	105	458	563
	%	18.7%	81.3%	100.0%
1940-1959	度数	259	946	1205
	%	21.5%	78.5%	100.0%
1960-1979	度数	134	390	524
	%	25.6%	74.4%	100.0%
合計	度数	498	1794	2292
	%	21.7%	78.3%	100.0%
				X^2=7.729, d.f.=2, p<0.05

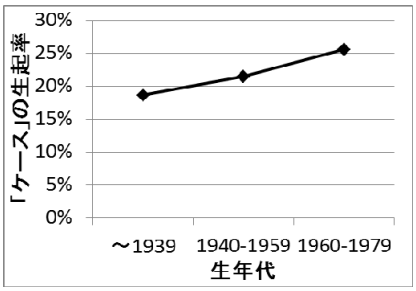


図 3：生年代の違いと「ケース」の生起率

表 2 は「ケース」と既存語の度数と生起率を、3 つの生年代区分ごとに表している。図 3 は「ケース」の生起率のみをグラフ化したものである。図 3 を見ると、書き手の生年代が上がるにつれて「ケース」の生起率が次第に上昇している。つまり、生年が上がる（世代が若くなる）につれて、既存語群に替わって外来語がより用いられていることがわかる。これにより、「ケース」の見かけ上の言語変化が認められ、図 2 で見た、金(2011)における「ケース」の出現率の増加が言語変化とかかわっていることが予測できる。

ちなみに、表 2 でカイ 2 乗検定をかけると有為差が認められた[7]。このことは、若い人ほど「ケース」をより使う、ということを直接示すものではない。しかし、少なくとも「ケース」と既存語の生起率には生年代間により差があることが統計的に認められ、図 3 も参照して総合的に判断すると、特に若い年代が生起率を上げていることが「ケース」の出現率の上昇に影響を与えていると思われる。

[7] 検定には SPSS ver. 20 を使用した。

３．２．書き手の性別

バリエーション研究において、女性のほうが変化をリードする、ということがよく言われる。これを今回の調査語である「ケース」にあてはめると、女性のほうが外来語をより使うことが予測される[8]。そこでまず、書き手の性別という観点から「ケース」の生起率を見ていく。表 3 から、「ケース」の生起率が男性（21.6%）よりも女性グループ（23.0%）において若干高いことが読み取れるものの、これは統計的に有意な差ではなかった。これで、外来語「ケース」の生起率において性差は認められないことがわかった。

表 3：性別の違いと「ケース／既存語」の生起率

		ケース	既存語	合計
男	度数	446	1620	2066
	%	21.6%	78.4%	100.0%
女	度数	52	174	226
	%	23.0%	77.0%	100.0%
合計	度数	498	1794	2292
	%	21.7%	78.3%	100.0%
				X^2=0.242, d.f.=1

３．３．媒体差

次に、媒体間で「ケース」の生起率に差があるかについて検討する。表 4、図 4 から、「ケース」の生起率は書籍（20.6%）よりも雑誌（38.3%）において上昇していることがわかり、これはカイ 2 乗検定により有意であった。

表 4：媒体の違いと「ケース／既存語」の生起率

		ケース	既存語	合計
書籍	度数	444	1707	2151
	%	20.6%	79.4%	100.0%
雑誌	度数	54	87	141
	%	38.3%	61.7%	100.0%
合計	度数	498	1794	2292
	%	21.7%	78.3%	100.0%
				X^2=24.256, d.f.=1, p<0.001

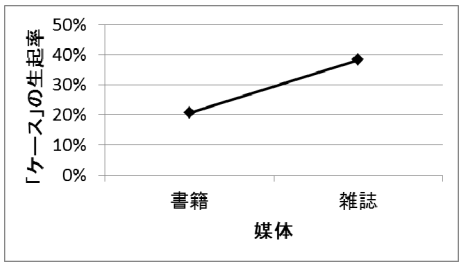


図 4：媒体の違いと「ケース」の生起率

では、なぜ媒体間で差が出たのだろうか。そもそも新聞・雑誌の媒体としての違いはどこにあるのだろうか。考えられる可能性は以下の 2 つである。第 1 に、各媒体の特徴とし

[8] 外来語のなかには、ファッションや美容、スポーツ関連語など、特定の性別のみが使うことの多い特徴的な語彙もある。しかし、本研究では、こうした影響を避けるため、ニュートラルな語を選んだ。

て、スタイル（改まり度）の違いが影響している可能性がある。しかしながら、両媒体とも広いジャンルを網羅する媒体であるため、媒体とスタイルを直結させるのは難しい。よって、たとえば書籍なら専門書か一般書か、雑誌なら専門誌か一般誌かなど、スタイルの異なる種類のものがどのような割合で含まれているのか詳細に調査して、媒体種と改まり度との関連を考察する必要がある。

第 2 に、媒体そのものの特徴が影響している可能性も考えられる。書籍と違い、雑誌は短期間で売り上げを伸ばすために、すぐに読者の目をひくような個性的・魅力的な存在である必要がある。また、常に新しい情報を提供していくという特徴があるため、目新しさという側面も持ち合わせていなければならない。こういった特徴のために、雑誌という媒体は、「スタイリッシュな・おしゃれな・かっこいい・斬新な・目新しい」というようなイメージと結びつきやすいと思われる。一方、梁(2012)によれば、日本語における外来語という語種のもつプラスイメージとして一番多かったものは、「かっこいい」すなわち「洗練されている」という評価だったという。ここで今一度、雑誌において外来語の生起率が高い理由を考えるならば、雑誌という媒体のもつ「スタイリッシュさ」というプラスイメージが、同じく「かっこよさ」というプラスイメージをもつ外来語によって体現しやすいため、と考えることができないだろうか。

最後に、媒体間でのサンプルサイズに大きな違いがあることも考慮する必要がある。書籍コーパスからのサンプルが 2151 件であるのに対し､雑誌コーパスからのサンプル数は 141 件しかない。そのため、雑誌においては少しの度数の差でも全体の割合の差として出やすいという側面もあるのかもしれない。

３．４．スタイルによる差

最後に、BCCWJ の特定目的サブコーパスに含まれる「Yahoo!知恵袋[9]」コーパスを選び、2．5 と同じ手順でデータ（ただし、連体修飾節構造に限る）を抽出し、外来語「ケース」の出現率に、スタイルによる差が出るかどうかを調べてみた。「Yahoo!知恵袋」は、インターネット上に投稿された質問に対して、不特定多数の人が回答を書き込むという形でやりとりされる。インターネットにおける書き言葉は、出版されないという点で、出版を前提としている書籍・雑誌よりも改まり度が低いことが予想され、両者にはスタイルの違いがあると考えられる。よって両者を比較することで、スタイルが「ケース」の生起率に影響を与えているかどうかを検証することができる。ちなみに「Yahoo!知恵袋」コーパスのデータは 2004 年から 2005 年にインターネット上に投稿されたもので、出版コーパスを構成する出版物の出版年(2001-2005)と同時期であり、両者はよい比較対象になると思われる。

表 5：スタイルの違いと「ケース／既存語」の生起率

		ケース	既存語	合計
出版	度数	498	1794	2292
	%	21.7%	78.3%	100.0%
非出版	度数	200	1449	1649
	%	12.1%	87.9%	100.0%
合計	度数	698	3243	3941
	%	17.7%	82.3%	100.0%
X^2=60.633, d.f.=1, p<0.001				

[9] 「Yahoo!知恵袋」コーパスは、ヤフー株式会社から提供された、2004 年 10 月から 2005 年 1 月にかけて投稿された 3,120,839 の質問とそれに対する回答からなるデータがもとになっている。コーパス自体はこのうち抽出された 91,450 サンプルから構成されており、規模は約 1,000 万語にのぼる。なお、1 サンプルは 1 つの質問とそれに対するベストアンサーからなる（山崎、他(2012)）。

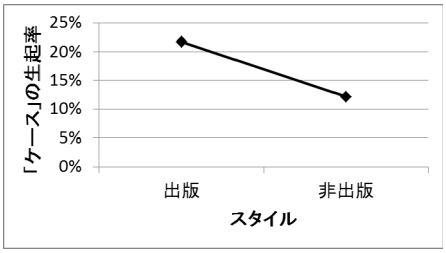


図 5：スタイルの違いと「ケース」の生起率

表 5、図 5 から、「ケース」の生起率は非出版物（Yahoo!知恵袋）（12.1%）よりも出版物（書籍・雑誌）（21.7%）において高いことが見てとれ、この差は、カイ 2 乗検定において有為であった。

ここで、外来語「ケース」の生起率にスタイル差が出たという事実から、外来語「ケース」がもつ、類義語の中での相対的なイメージについて考察してみたい。改まり度が高くなるほど生起率が上昇するということは、「ケース」という語が、既存語に対してよりフォーマルな変異形として機能しているということを意味している。

ここで、こうした傾向は、既存語をもつ外来語群に一般的に当てはまるものなのか、という疑問がわく。筆者の直観としては、外来語が既存語よりもフォーマルであるかどうかは語彙によって異なる可能性がある。なぜなら、たとえば動詞系外来語「チャレンジする」や形容動詞系外来語「クールな」などは既存語に比べて若干インフォーマルな印象があるからである。ただ「チャレンジする」や「クールな」は「ケース」に比べ定着度が低いことから、語のフォーマルさというものが定着度に関係する可能性もあり、定着度に応じてスタイルとの関係を整理する必要もありそうだ。

４．考察・まとめ

本稿では、「ケース」を事例として、外来語と既存語の語彙交替に影響を与えうる言語外的要因のうち、BCCWJ において取得可能な要因（生年代、性別、媒体差、スタイル差）それぞれに関して検証を行った。これまでの外来語研究においては、外来語の生起率について、社会言語学的観点から調査したものはほとんどなく、あっても意識調査にとどまることが多かった。それは第 1 に、語彙を社会言語学的変異形として扱うことが難しいという問題と、第 2 に、外来語という語種の出現頻度が低いために定量的調査に耐えうるだけのデータを収集するのが難しいという問題があったからである。第 1 の問題は、「機能的等価性」を手がかりに分析対象を細かく限定することで解決を試みた。また第 2 の問題についても、BCCWJ が完成したことで外来語のような低頻度語でも一定量のサンプルを得られるようになった。さらに BCCWJ の一部のサンプルには書き手の属性情報（生年代、性別など）もタグ付けされている。これにより、実際の言語資料をもとに、社会的属性を説明変数とした外来語使用についての調査ができるようになった。本研究はその試験的試みである。

本研究での調査の結果、「ケース」の生起率に、生年代の違いと、媒体差（書籍と雑誌）、スタイル差（出版物と非出版物）が影響することがわかった。生年代については、若い世代ほど「ケース」の生起率に上昇が見られたことで、金(2011)が示した 20 世紀後半の新聞における「ケース」の出現率の急増が、言語変化と関連していることがわかった。また、本稿冒頭で紹介した、田中(2007)の意識調査の結果（外来語使用に対する意識は年齢層により段階的な差がある）は、実際の言語使用にも現れていることがわかった。

性差については、女性の「ケース」生起率が男性よりも若干高かったものの、この差は統計的に有為でないことがわかった。ただし「ケース」は日本語においてかなり定着度の高い語彙である。まだ完全に定着していないような外来語を選択して今後調査をすれば、もしかして性差が認められるものもあるかもしれない。

媒体の違いによる生起率の差については、その差の意味を解釈するために、書籍と雑誌

の媒体差の本質が何であるかをより詳しく特徴づけることが今後の課題となった。そのためには、媒体種とスタイル（改まり度）との関わり、そして「ケース」という語のもつイメージを把握することがひとつの糸口となる可能性がある。

スタイルの違いによる生起率の差については、既存語に対する外来語の相対的なイメージ（または地位）を考察するにあたって重要なポイントとなると思われる。今回は「ケース」が既存語と比べてよりフォーマルな語彙であることがうかがえたが、これは、必ずしも一般化できることではないだろうということが、筆者の今の見解である。それぞれの語のイメージや地位はその語の定着度とも関わりがありそうで、語の定着度も参照しながらスタイルとの関係を考えていく必要がありそうだ。

その他、今後の課題としては、外来語使用に影響を与えうる、生年や性別以外の言語外的要因に関しても、可能な限り検証していく必要があるだろう。また、言語外的要因と言語内的要因を合わせて、それぞれの要因の影響度の違いを明らかにすることも重要である。「ケース」とその既存語については、言語内的要因として、それを修飾する節の内容（デキゴト）の「よしあし」や「既然・未然性」（金(2011)）といった意味的分類や、共起する述語の種類（有無・多少など）が一つの指標になると思われる。

いずれにせよ、本研究で得られた傾向が外来語一般に拡張できるのかどうか、「ケース」以外の外来語も調べて事例研究を積み重ねる必要がある。それにより、日本語における外来語の社会言語学的役割の全体像が見えてくるだろう。

謝　辞

本研究で分析したデータは、筆者が 2012 年 11 月から 2013 年 2 月まで外来研究員として国立国語研究所に滞在していた期間中に、BCCWJ から収集したものである。滞在を受け入れて頂いた同研究所所長の影山太郎先生、言語資源研究系系長の前川喜久雄先生に感謝申し上げるとともに、受け入れ教官として滞在中様々な面でご指導ご鞭撻いただいた田中牧郎先生には特に感謝の意を表したい。なお、データ抽出の際には、同研究所コーパス開発センターの中村荘範さん（マンパワージャパン株式会社）に協力して頂いた。また、本稿の準備段階で貴重なコメントをして頂いた同研究所研究員の金愛蘭さんと南部智史さんにもこの場を借りてお礼申し上げる。

参考文献

金愛蘭(2011)「20 世紀後半の新聞語彙における外来語の基本語化」阪大日本語研究 別冊 3.

田中牧郎(2007)「漢語・和語と比較した外来語に対する意識」『公共媒体の外来語―「外来語」言い換え提案を支える調査研究―』国立国語研究所報告 126、pp.302-310.

茂木俊伸(2012)「コーパスを用いた外来語サ変動詞の分析―「カットする」を例として―」『特定領域「日本語コーパス」』平成 22 年度公開ワークショップ（研究成果報告会）予稿集、 pp.103-110.

山崎誠、小椋秀樹、小沼悦、他(2012)「研究活動・成果の総括：データ班　代表性を有する現代日本語書籍コーパスの構築」『特定領域「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.149-156.

梁敏鎬(2012)「日本語と韓国語の外来語の受容意識―イメージ調査の分析―」、陣内正敬、田中牧郎、相澤正夫編(2012)『外来語研究の新展開』、pp.148-167、おうふう.

Ito, Rika and Sali Tagliamonte (2003) *Well* weird, *right* dodgy, *very* strange, *really* cool: Layering and recycling in English intensifiers. *Language in Society*, 32, pp.257-279.

Labov, William (1972) *Sociolinguistic patterns*. University of Pennsylvania Press.

Lavandera, Beatriz R. (1978) Where does the sociolinguistic variable stop? *Language in Society*, 7, pp.171-183.

Meyers-Scotton, Carol. (2006) *Multiple voices: An introduction to bilingualism*. Wiley Blackwell.

国会会議録に見る複合辞の特異な形
—丁寧形/普通形の不対応—

服部匡（同志社女子大学表象文化学部）

Marked Forms of Compound Particles in the Minutes of the National Diet of Japan

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

1. 概要

いわゆる複合助詞や関連形式の文法的性質に関しては、多くの記述的研究が行われているが、レジスターに強く依存して用いられる形式についてはあまり注目されていない。本研究では、国会の会議に見られるような形式ばったスタイルで特徴的に用いられる形の存在を指摘し、分布からみた使用特徴や通時的推移について述べる。

動詞由来の複合辞形式のうちニを伴い出現頻度の高い「について、において、によって、に関して、に対して」の 5 つと、対応の連体形式、それぞれの丁寧な形式の出現傾向を次の 3 種類のコーパスで調査した。括弧内は対象とする発話の産出年代である。

国会会議録(1947-2006) [1]・日本語話し言葉コーパス(1999-2001?)・BCCWJ(書籍・雑誌)

その結果の概要を示すと次の(1)のようになる。×はコーパスを問わずほぼ出現しない形であり、∆は主に国会会議録に出現するが、従来注目されていない形である[2]。

(1) 複合助詞・対応形式の主な形とその出現状況

において	×におき[3]	×におく	におきまして	∆におきます(る)
		における		∆におけます(る)
について	につき	×につく	につきまして	∆につきます(る)
によって	により	による	によりまして	によります(る)
に関して	に関し	に関する	に関しまして	に関します(る)
に対して	に対し	に対する	に対しまして	に対します(る)

∆を付した丁寧形のうち「におきます(る)」「につきます(る)」は普通形との形式的対応を欠いている。もっとも、「おきます」は意味機能的には普通形「おける」にほぼ対応する[4]。これらの形は、おそらく、「おきまして＞おきます」「つきまして＞つきます」のような、一種の逆成(back-formation)によって生まれたものと思われる。実例をあげておく(前後略)。

(2) まあ終戦の時におきまする問題として、地上に出しておつたものが(1952 参/建設委 5 八

---

1 国会会議録のデータの一部は BCCWJ に収録されているが、ここで用いるのはフルセットのデータで、ある。通時的観点からの日本語研究に用いうるコーパスの種類や各コーパスの話者生年代・産出年代の分布については、服部(近刊)で述べている。

2 国会会議録で「おけます」は「おきます」の約 100 分の 1 の用例数しかなくおよそ 1990 年以降に集中する。「におきます」は BCCWJ にも 2 回出現し、いずれも浅井基文(1940 生, 元外交官, 政治学者)著『平和大国か軍事大国か』の一部であり、講演の記録のように思われる。また、話し言葉コーパスにも 6 回出現し、いずれも学会講演である。どちらのコーパスでも、総字数あたりの出現頻度は国会会議録よりはるかに低い。なお、「におきました(る)」のような過去の連体形式も国会会議録に見られる。

3 「におき」は、国会会議録に少数の用例がある。誤記を疑われるものが多いが、それ以外に、「～におき、また、～におきましても」のような等位接続になっているものなどが僅かな数ある。

4 「おきます」には「～におきますと」などのような言い方も見られ、すべてが「おける」に対応した連体用法というわけではない。他にも、対応関係について検討すべきことがあるが省略する。

嶋三郎)

(3) 私どもは、有事におきます自衛隊の行動につきましては、(1988/85 衆/予算委 2/伊藤圭一)

(4) 災害時におきます被災者に対する心のケアは、(1992/154 参/災害対策特別委 4/高原亮治)

(5) 住民税につきまする基本的な市町村への通達の中で、(1958/28 衆/地方行政委 16/奥野誠亮)

(6) 私、実はこの法案につきまする質疑をいたすに際しまして、(1985/103 衆/公職選挙法特別委 2/上村千一郎)

(7) こうした症状につきます検査、投薬、注射などの診療行為につきましては、(1992/154 参/環境委 2/ 中村秀一)

以下では、調査対象を国会会議録に絞り、テ形式と連体形式とのそれぞれ丁寧形と普通形との分布特徴を観察する。

2. 各形式の分布特徴

国会会議録でのテ/テノ形式および連体形式の総用例数を示すと次のようになる。これは、全期間(60 年間)の合計であり、括弧内は、1 億字あたりの出現頻度である[5]。「におきます」は、低頻度な形ではないことがわかる。「に関する」の頻度が高いが、これは、「～に関する{法律/請願/件}」のような審議案件の標題(の一部)を多く含んでいる。

表 1 各形式の用例数(1 億字あたりの頻度) ：普通形

	～テ	～テノ	連体
につき	1,045,788(30122.4)	18,104(521.5)	**
に関し	54,058 (1557.1)	6,706(193.2)	1,084,222(31229.5)
に対し	517,965(14919.2)	2,7051(779.2)	805,410(23198.7)
により	605,814(17449.6)	2,531 (72.9)	381,935(11001.1)
におき	1,045,078(30102.0)	18,100(521.3)	635,903(18316.3)

表 2 各形式の用例数(1 億字あたりの頻度) ：丁寧形

	～まして	～ましての	～ます
につき	473,969(13652.0)	32,359 (932.1)	1,188 (34.2)
に関し	30,008 (864.3)	1,333 (38.4)	15,524 (447.1)
に対し	140,348 (4042.5)	3,194 (92.0)	28,153 (810.9)
により	253,348 (7297.3)	517 (14.9)	31,118 (896.3)
におき	479,244(13803.9)	7,034 (202.6)	85,197(2454.0)

テ形式と連体形式のそれぞれでの、丁寧形と普通形の比率を図示すると次のようになる。

対応する普通形を欠く「につき」を除くと、連体形式の方がテ形式より、普通形の用例比率が高い。テ形式の場合に比べ連体形式での丁寧形の使用はより丁寧度の高いスタイルを要求するという、三尾(1942)の指摘以来知られる事実の反映と思われる。

5 用例数は、当該複合辞形式の直前・直後の字が漢字か読点の例の数である(「については」等は含まない)が、少数のゴミを含む可能性があり、初期に見られる特殊な表記の一部を見落としている可能性がある。

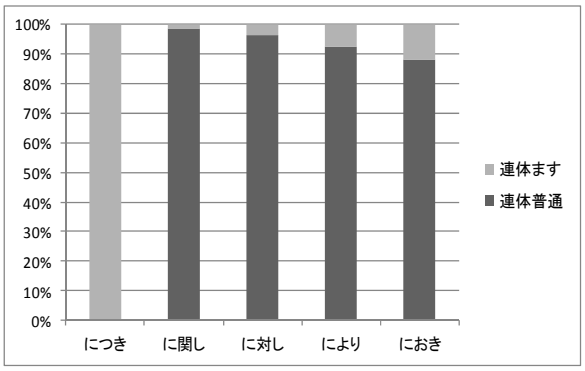


図 1 連体形式での丁寧/普通形の比率(1947-2006)

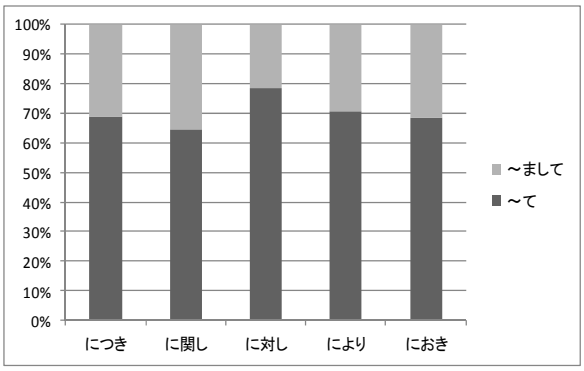


図 2 テ形式での丁寧/普通形の比率(1947-2006)

2.1. 会議種別・発言者別の使用傾向

参議院議員の参議院での発言に限定して、会議の議長・委員長等の発言と一般議員の発言を分け、さらに、一般議員につては会議の種類(本会議/委員会等)によって分けて、各形式の用例頻度の推移を観察する[6]。6 期にわけ、1 億字あたり出現頻度の推移を図示する。

1 期 1947-1956 年　2 期 1957-1966 年　3 期 1967-1976 年
4 期 1977-1986 年　5 期 1987-1996 年　6 期 1997-2006 年

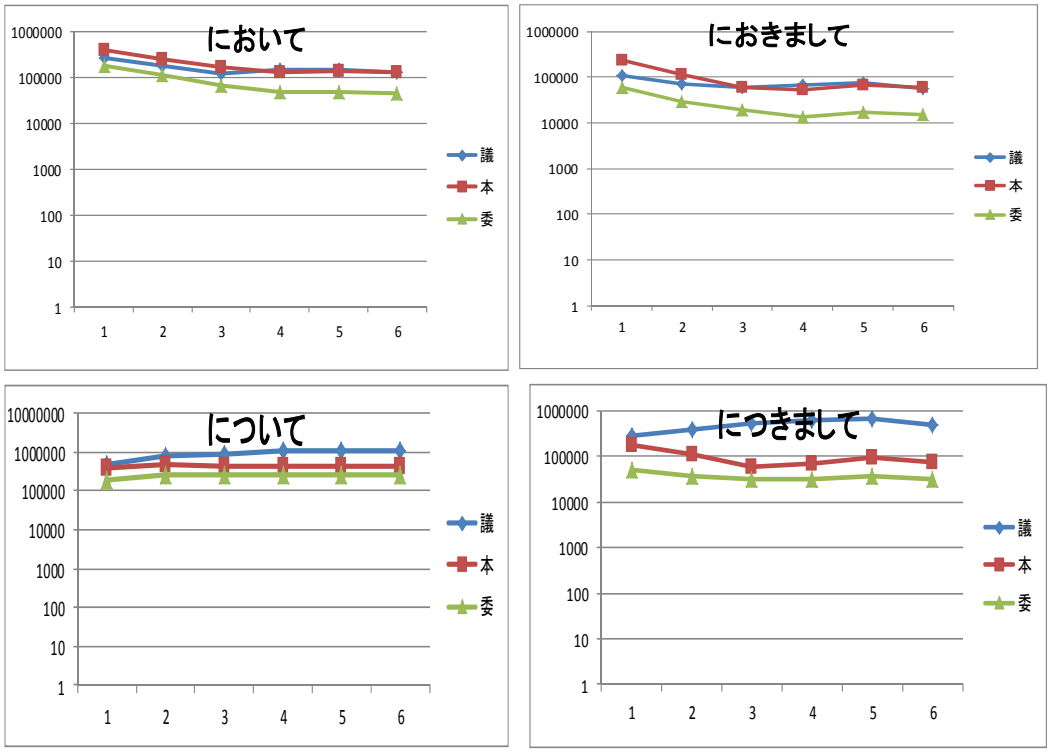
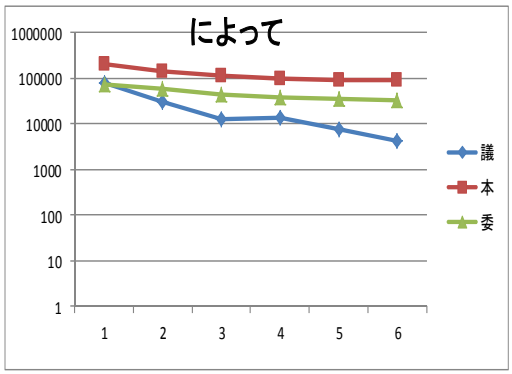
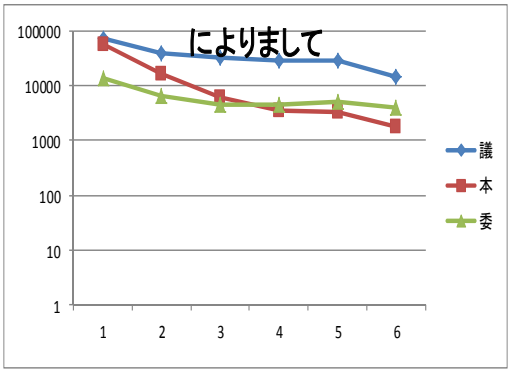
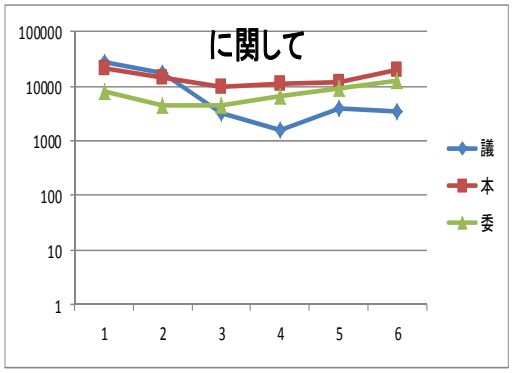


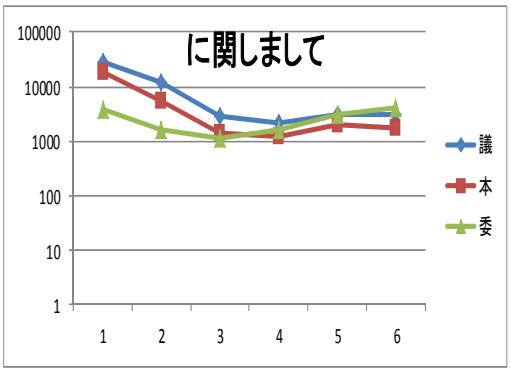
図 3-図 6 頻度の推移 (議長等/本会議/委員会)

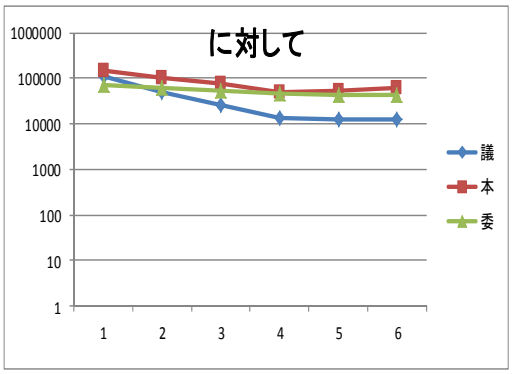
6 用例数は、当該形式の前の文字が漢字の例の数である(後続文字の字種は問わない)。「については」「についてのみ」なども数のうちに含む。

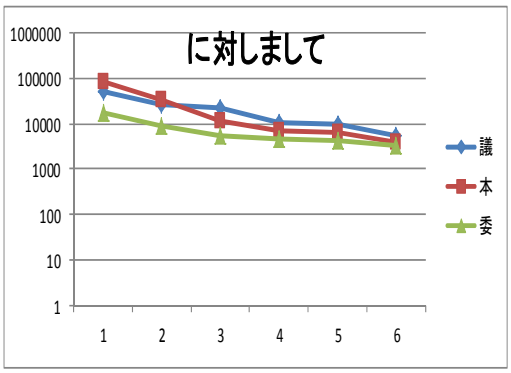


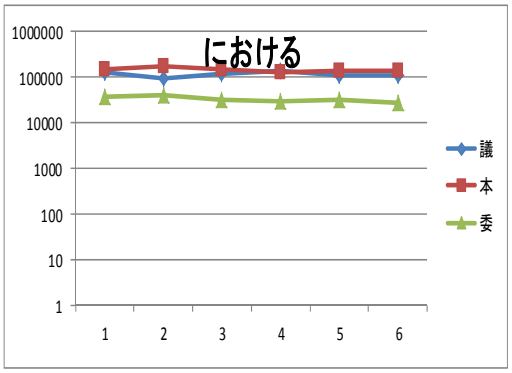












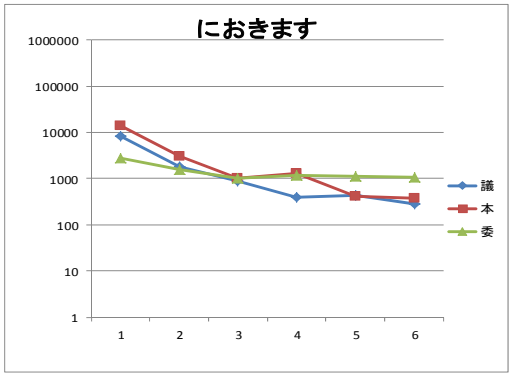


図 7-図 14　頻度の推移 (議長等/本会議/委員会)

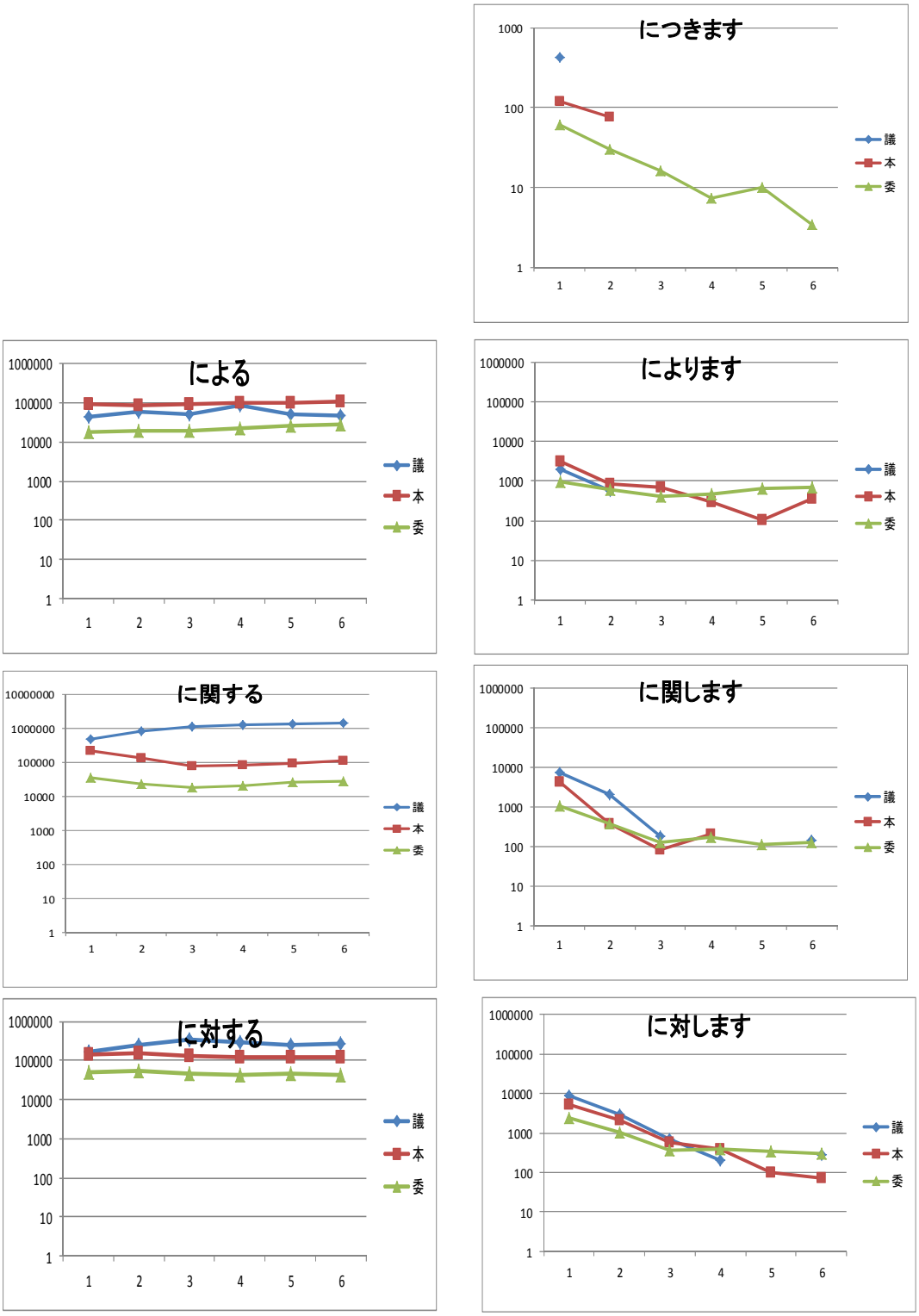


図 15-図 21　頻度の推移 (議長等/本会議/委員会)

「よりまして・関して・対して」などで議長の発言での頻度が高い傾向が見られる。標題や議事進行上の定型的発言の影響かと思われる。また「における・に関する・に対する」などは委員会より本会議での頻度が高いが、「によって」などはその逆の傾向を示している。

2.2. 先行/後続する複合辞的要素

国会会議録では、複合辞的形式が複数連続することがよく見られる。「〜におきます」を例にとると、それに先行するものに次のようなものがある。

(8) 「〜(の)上に」

やってまいります上におきます措置といたしましては

干拓ができました上におきます養殖業の問題でございます

同じような仕事の上におきまする指揮命令の関係において

(9) 「〜(の)際に」

一定のいろいろな措置をいたしまする際におきまする認定は、医学上

審査の内示を与えた際におきます地目は畑が

農薬の登録の際におきます水道水中の農薬の除去技術の検討に関しましては

(10)「〜(の)場合に」

みそを作る場合におきます醗酵過程において

物価が値上りした場合におきまする実態生計費をどうして

会計士補の懲戒の場合におきまする権利を保護しようと

また、後続するものとしては、連体要素の後に挿入されて明確な意味を持たない「ところの」が代表的である。これは、一般的な連体修飾節の後にもよく出現する。

(11) ニジェール国におきますところの探鉱開発を進めるというふうに

一番最近におきますところの暴挙というか、許しがたい行為で

「におきます」の前後両方に複合辞的形式を伴う例も少数ではあるが見られる。

(12) 今後の御審議の上におきまするところの御参考に申上げます。

経済の上におきまするところの規律を確立するにいたしましても

地方自治法の上におきますところの一般的な国、府県、市町村

3. おわりに

国会会議録に特徴に見られる特異な複合辞形式を指摘し、それらを含めた複合辞の丁寧/普通形での分布特徴を観察した。さらに文末形式の丁寧度との関係や、複合辞の前後に現れる特徴語・表現なども分析したい。

文　献

杉本武(2009)　「複合格助詞の連体用法について」　文部科学省科学研究費補助金特定領域研究「日本語コーパス」平成 20 年度研究成果報告書『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発 III』 (研究代表者:田野村忠温)　pp. 166-182.

服部匡(2011a) 「言語資料としての国会会議録の特徴(1) ―本会議と委員会等との比較―」『同志社女子大学日本語日本文学』 23:pp.39-49.

服部匡(近刊)　「現代日本語の通時変化」　『講座日本語コーパス 第 6 巻 コーパスと日本語学』　朝倉書店

三尾砂(1942) 『話言葉の文法(言葉遣編)』 帝国教育会出版部

口頭発表（2）

2月28日（木）　15:00～17:00

筑波ウェブコーパス検索ツール NLT の開発

今井 新悟（筑波大学）
赤瀬川 史朗（Lago 言語研究所）
プラシャント・パルデシ（国立国語研究所）

Development of NLT: the Search Tool for Tsukuba Web Corpus

Shingo Imai (Tsukuba University)
Shiro Akasegawa (Lago Institute of Language)
Prashant Pardeshi (National Institute for Japanese Language and Linguistics)

１．はじめに

本稿では 2013 年に一般公開を予定している NLT（NINJAL-LWP for Tsukuba Web Corpus）の開発とそのシステムの特長について述べる。NLT は 2012 年 6 月に公開した NLB（NINJAL-LWP for BCCWJ）と同一のシステム NINJAL-LWP で動作するレキシカルプロファイリング型のコーパス検索ツールである。検索対象となる筑波ウェブコーパス（TWC）は、ウェブ上から収集した約 11 億語の日本語のテキストデータである。以下では、TWC の構築と検索システムへの実装について述べた上で、公開前のシステムから得られた動詞頻度と動詞（「走る」と「駆ける」）と名詞のコロケーションの結果を NLB と比較し、その有用性と可能性について探りたい。

２．筑波ウェブコーパスの構築の目的と規模

２．１　構築の目的

一般に、コーパス基盤の言語研究においては、研究対象となる言語現象を複数のコーパスで比較して観察することで研究の信頼性や客観性を高めることができる。2011 年に完全公開された『日本語書き言葉均衡コーパス』（以下、BCCWJ）は日本語初の均衡コーパスで、その規模は約 1 億語である。2012 年 6 月にはこの BCCWJ 向けのレキシカルプロファイリング型のコーパス検索ツール NLB（NINJAL-LWP for BCCWJ）が一般公開された[1]。筑波ウェブコーパス（以下、TWC）の開発の最大の目的は NLB と同じ検索システムを利用して BCCWJ と比較できるウェブコーパスを構築することにある。同一のインターフェースを利用することで BCCWJ と TWC の比較が容易になるため、均衡コーパスの「質」とウェブコーパスの「量」の双方のメリットを言語研究や日本語教育に生かすことが期待できる。

２．２　コーパスの規模

均衡コーパスは、厳密な統計的手法に基づいてデータが採取されることから、コーパスの規模に関しては常に時間的・資金的制約がつきまとう。それに対して、ウェブ上のテキストを収集して構築するウェブコーパス[2]には事実上そのような制限はない。つまり、コーパスの規模は限りなく大きくできる。1 億語のウェブコーパスと 10 億語のウェブコーパスを比べれば、10 億語のコーパスのほうがより多くの有用な言語情報を含むと考えてよい。イギリスのコーパス統合ツールサイト Sketch Engine では、10 億語規模から TenTen と呼ばれる 100 億語規模の各国語のウェブコーパスが検索できる。国立国語研究所でも 100 億語を超える超大規模コーパスを開発中である。日本語においても、ウェブコーパスがふつうに活用される時代がすぐそこまで到来している。

TWC については、2012 年夏に 5 億 8 千万語のパイロット版を制作し、2013 年に公開する

[1] URL は文末参考 URL を参照。本稿では、現行の BCCWJ よりもデータサイズがやや小さい BCCWJ の領域公開データ（2009 年版）の 6 千 2 百万語を採録した NLBVer1.10 を用いる。

[2] 英語では、Web As Corpus（WaC）という言い方がよくされる。

一般公開版では 11 億語まで拡張する予定である。10 億語規模にした理由としては、BCCWJ の 10 倍強の規模で比較しやすい大きさであること、英語コーパスを利用した辞書制作のこれまでの経験から見て、10 億語が中型辞書の見出し語の用例を十分に採取できる一つの目安となること[3]、比較的短期間で構築できることなどが挙げられる。

３．筑波ウェブコーパスの構築の過程

３．１　収集方法

ウェブ上からのテキストの収集については、検索エンジンの API を利用して、ウェブページの URL を収集した後、その URL のデータを収集する一般的な手法に従った。具体的な手順については、ウェブコーパス構築ツール BootCaT を参考にしてプログラムを作成した。

【シードおよびタプルの生成】検索エンジンのクエリパラメータに与えるタプルを構成するシードには、NLB の開発過程で作成した BCCWJ（2009 年の領域公開データの一部、約 6 千 2 百万語）の頻度リストを利用した。品詞ごとに分かれた頻度リストのうち、内容語である名詞、動詞、形容詞、副詞のリストをマージして、上位 500 語をシードとして選んだ。ただし、名詞のうち、数詞、固有名詞は排除し、また、動詞、形容詞については活用形も含めた。この 500 語のシードから無作為に 3 語を選び出し、計 50 万組のタプルを作成した。以下にタプルの例を示す。

```
駄目 皆 構造
条件 とても 様々
法律 (答える OR 答え OR 答えよ OR 答えれ OR 答えろ OR 答えりゃ OR 答えん) 人々
```

【検索エンジン API による URL の収集】URL の収集には、Yahoo!ウェブ検索 API を利用した。1 タプル当たりで収集する URL 数は 10 ページとし、2012 年 1 月初旬から下旬にかけて計 500 万 URL を収集した。重複した URL を削除した URL 総数は約 3 割減の約 350 万件になった。

【HTML ページの収集】URL データを 5 万件ごとに分割した上で、3 台の端末を利用して 2 週間をかけて HTML ページを収集した。

３．２　コーパスデータの抽出

【テキストの抽出】次に収集した HTML ファイルからテキストを抽出する作業を行った。具体的には、HTML タグの削除、文字コードの統一（utf8）、日本語以外の言語で書かれたテキストの削除[4]を行った。

【不適正なページの排除】ウェブ上のテキストの収集の目的は日本語の用例を採取することにあるので、単に項目やリンクを列挙しただけのページ、広告と思われる内容の多いページ、センテンス境界の判定が難しいページは、あらかじめコーパスデータの対象から外した。

【センテンスの抽出】レキシカルプロファイリングツール NINJAL-LWP では、センテンス単位にした用例の中にどのようなコロケーションが含まれるかを文法パターン別に抽出する。そのため、コーパスデータはあらかじめセンテンス単位に分割しておく必要がある。一つ前の作業でセンテンス境界の判定が難しいページを排除したのもこの理由による。

【用例データの抽出】センテンス単位のデータのなかには、見出しに相当するものや、メニュー項目に相当するものが含まれる。センテンス中にどの程度名詞が含まれるか、セン

[3] 英語と日本語を比べた場合、同じ語数では英語のほうが情報量が多い。そのため、10 億語の英語と日本語では英語のほうが情報量が多くなる。その意味では 10 億語という数字はあくまでも目安に過ぎない。

[4] Perl モジュール Encode::Guess を利用した。

テンス中に動詞は現れるか、「クリック」や「ログイン」などのウェブページで多用される表現が用いられているかなどの複数の観点から、用例としての適正度を数値化し、用例としてふさわしいデータを抽出した。図 1 は、適正率を示したウェブページのテキストの例である。網がけになったセンテンスは適正率が高く、用例データとしてふさわしいと判断されたものである。さらに、同一ページで同じセンテンスが現れた場合も、最初の 1 件のみを用例として採取し重複を避ける工夫をした。

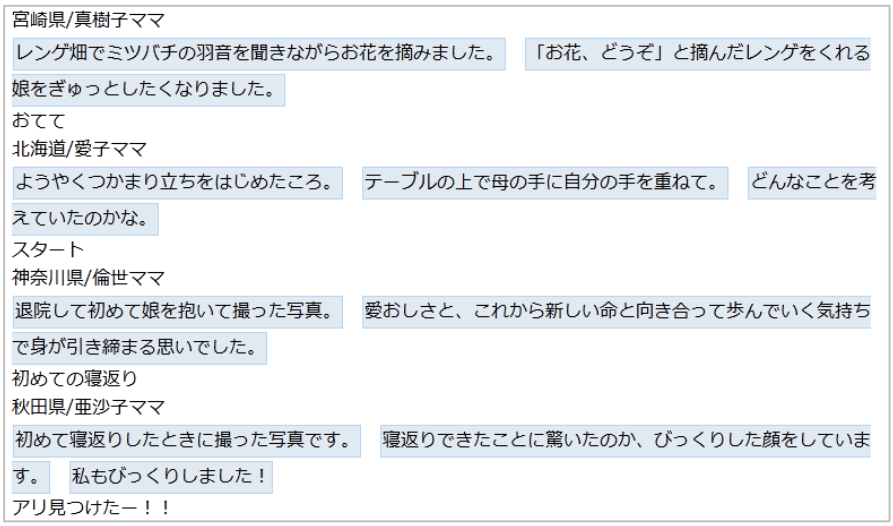


図 1 用例としての適正率

【重複する用例データの削除】一つ前の作業で、同一ページでは同じ用例が複数回採取されないようにしたが､6 億語弱のパイロット版 NLT を開発して実際に運用してみたところ、同一サイトで同一の用例が頻出することが確認された。そのため、URL の情報をもとに同一サイト[5]での同じ用例は一度だけ採取するように改良し、最終的に語数にして 11 億 3781 万語、用例数にして 4672 万 7 千例の筑波ウェブコーパスが完成した。

4．NINJAL-LWP への実装

図 2 NLT の見出し語ウィンドウ

[5] 正確には同一の FQDN（完全修飾ドメイン名）。

NINJAL-LWP は日本語コーパスの汎用的な検索システムである。2012 年の BCCWJ への実装（NLB）に続き、今回の TWC は 2 例目になる。図 2 は、動詞「走る」の見出し語画面である。NLB と同一のインターフェースなので、画面を左右に並べれば BCCWJ との比較が簡単にできる。

５．動詞出現頻度の比較

NLB と TLB で抽出された動詞を頻度順にならべ、それぞれ上位 1 万語を抽出し、どちらか一方に現れない語を削除して、9001 語を得た。両者の頻度を対数変換してピアソンの相関係数を求めると 0.923 であり、NLB での動詞の分布と NLT での動詞の分布は極めて相似している。TWC のデータ収集はウェブのクローリングにより収集されるデータの偏りを克服するため、前述の通り、BCCWJ の語分布を模するという方略（および上記の各種方法）を使った。両者の動詞の相関を見ると、ウェブコーパスの弱点である偏りを克服するという課題は相当程度達成されたと言える。

スピアマンの順位相関は 0.887 である。順位で見てもマクロ的には両者は似ているといえるが、ミクロで見ると違いが現れる。順位の差が大きいものは、表 1 の通りである。

表 1　TWC と BCCWJ の動詞頻度順位の比較

動詞	TWC 順位	BCCWJ 順位	TWC 頻度	BCCWJ 頻度	順位差
答えする	2174	9493	4805	15	-7319
開講する	2546	8720	3726	20	-6174
退会する	3342	9323	2315	16	-5981
許諾する	3740	9696	1910	14	-5956
被曝する	2766	8583	3265	21	-5817
来場する	4184	9932	1538	13	-5748
選考する	4195	9932	1532	13	-5737
研修する	3396	8999	2257	18	-5603
支払いする	2113	7615	5004	30	-5502
祭りする	3747	8861	1904	19	-5114
フォーカスする	4081	9163	1626	17	-5082
リニューアルする	2863	7910	3046	27	-5047
マッチングする	4922	9932	1094	13	-5010
試行錯誤する	4630	9493	1256	15	-4863
拝読する	4141	8999	1576	18	-4858
目指せる	4641	9493	1249	15	-4852
出展する	3402	8215	2248	24	-4813
付帯する	3817	8583	1842	21	-4766
カスタマイズする	3351	8114	2307	25	-4763
正解する	4966	9696	1070	14	-4730
（中略）					
哀願する	9526	4825	190	85	4701
飛び退く	9782	5077	175	77	4705
血走る	9095	4364	220	103	4731

調味する	9542	4734	189	88	4808
舌打ちする	8173	3209	307	171	4964
上気する	9299	4331	205	104	4968
すすり泣く	9247	4259	209	107	4988
言いかける	8043	2940	322	195	5103
後ずさる	9095	3965	220	121	5130
しゃくる	8808	3620	242	143	5188
まさぐる	9359	4011	201	119	5348
にこりする	9552	4166	188	111	5386
微笑する	7997	2601	327	237	5396
くぐもる	9451	3896	195	125	5555
座り直す	9722	4126	179	113	5596
愛撫する	9396	3174	198	174	6222

TWC の方が BCCWJ より相対的に順位が高いもののうち、「答えする」「支払いする」などはそれぞれ「お答えする」「お支払いする」の形で使われているものである。この表では割愛したが、同様に TWC の方が BCCWJ より相対的に順位が高いものの中には、「(お) 届けする」「(お) 預かりする」のように相手を想定した敬体での使用が多い。ウェブ上では顧客相手の情報が多いことの反映であろう。また、「フォーカスする」「リニューアルする」「マッチングする」「カスタマイズする」等のカタカナ語も目立つ。また、「被曝する」のように時事的な話題を反映したと思われるものが入っている。一方で、BCCWJ の方が TWC よりも相対的に順位が高いものには、小説など文学作品においての人物の動作描写に使われそうな語が並んでいる。

６．コロケーション

６．１　「～が走る」

「走る」のガ格に共起する名詞について、BCCWJ（NLB）から頻度 2 以上の共起語を取り出し、120 語を得た。それら 120 語の TWC（NLT）における同様の共起頻度を求め、両者の順位相関は 0.585 となった。このことから、両者の相関はある程度あるものの、収集されているコロケーションにはある程度違いがあることが予想される。なお、NLT の 5 億 8 千万語のパイロット版と今回の NLT の 11 億語版での順位相関は 0.973 であったことから、「～が走る」のコロケーションについては約 5 億語で相当程度安定して収集できることが示唆される。ただし、「～が走る」では、頻度が高いことから比較的安定して収集できたものであり、頻度の低いコロケーションでは、11 億語版であっても安定しないということもありうる。

TWC の「走る」のガ格に共起する名詞で頻度 20 以上のものは 103 語であった。そこから、代名詞、「もの」、「こと」など実質語的意味が希薄な語を除き、意味でカテゴリ化した。例えば、「車」「電車」「自転車」などを「乗り物」というカテゴリにした。表 2 にカテゴリ内の頻度計が 70 頻度以上となったものを示す。なお、右に添えられている数字はそれぞれの出現頻度である。

表２　TWC における「～が走る」の共起語

順位	カテゴリ	共起語例
(1)	乗り物 3284	車 889、電車 513、バス 413、列車 343、自転車 181 など
(2)	人・動物 1896	人 251、馬 197、私 137、自分 121、～たち 112 など
(3)	痛み 1078	痛み 565、激痛 513
(4)	経路 616	道路 178、鉄道 128、道 109、～号線 66、線路 48 など
(5)	動揺・衝撃 473	衝撃 275、激震 81、戦慄 49、動揺 48、電撃 20
(6)	感覚 261	～感 81、悪寒 59、痺れ 38、寒気 36、感覚 25、震え 20
(7)	緊張 248	緊張 248
(8)	線 292	線 236（路線名も含む)、ライン 28、筋 28
(9)	光 212	光 77、閃光 68、稲妻 67
(10)	電気 205	電気 134、電流 71
(11)	溝・亀裂 180	亀裂 102、断層 51、溝 27
(12)	地形 102	～系 45、山脈 32、～帯 25
(13)	虫唾 86	
(14)	線状器官 77	神経 44、血管 33

コロケーションの頻度情報とそのカテゴリ化はコーパス準拠（corpus-based）の辞書編纂に有用である。表 2 の順番に辞書の語義を並べることに特に違和感はなく、ほぼ直観に合っていると言えよう。語義とその配列順序を決めてから例文を探すあるいは作例するという従来の方法とは逆に、コーパスのコロケーションから意味のカテゴリ化を行い、語義を決めるという方法の可能性を示唆している。ただし、「走る」の中心義は「人・動物が足を速く動かして移動する」であり、「乗り物が速く移動する」は意味拡張であろうから、後者の方が圧倒的に頻度が高いものの、辞書編纂においてはコーパス駆動（corpus-driven）ではなく、コーパス準拠（corpus-based）が望ましい。

さて、コーパスのコロケーション頻度の有用性を確認したが、この頻度がある程度高くないと、コロケーションの情報が不安定になり、有用性が損なわれる可能性があるので注意が必要である。TWC では共起語の出現頻度上位 51 語（頻度 48 以上の語）に限ってみても各カテゴリの順位は５番目までは表 2 と変わらない。

(1) 乗り物 2992、(2) 人・動物 1307、(3) 痛み 1078、(4) 経路 529、(5) 動揺・衝撃 453 (6) 緊張 248、(7) 線 236、 (8) 光 212、(9) 電気 205、(10) 感覚 140、(11)溝・亀裂 102、(12) 地形 51

一方、BCCWJ では、頻度上位 50 語（頻度 5 以上の語）で見ると、カテゴリの頻度順は相当変化し、表 2 と等しいのは順位１位の「乗り物」だけになり、コロケーションの情報がやや不安定になっている。コーパス駆動ではなく、コーパス準拠（Corpus-based）だとしても、コロケーション情報は安定して得られる方がよい。

(1) 乗り物 151、(2) 痛み 112、(3) 人・動物 96、(4) 光 73、(5) 感覚 66、(6) 動揺・衝撃 60、(7) 経路 35、(8) 緊張 32、(9) 溝・亀裂 26、(10) 線 25、(11) 電気 15、(12) 地形 10、(13)予感 8

BCCWJ でも頻度 2 以上を採用すると共起語として出現する語数は 120 語となり、以下のような順番・頻度となる。これにより TWC のカテゴリ頻度順に近づく。それでも上位 3 位までは同じになるが、それ以下の順位は異なる。

(1) 乗り物 174、(2) 人・動物 148、(3) 痛み 112、(4) 光 77、(5) 感覚 74、(6) 動揺・衝撃 68、(7) 溝・亀裂 44、(8) 経路 39、(9) 緊張 32、(10) 線 29、(11) 電気 21、(12)地形 16

以上、「〜が走る」のコロケーションの場合は TWC ではコロケーション情報を安定して取り出せるが、BCCWJ の場合はコロケーションの頻度についてはやや不安定になる嫌いがある。BCCWJ においても、出現頻度が 2 までと低いものまで観察の範囲を広げることによって、安定性をある程度向上させられることを見たが、一方、出現頻度が 2 というのは少なすぎて、ノイズ（誤り、個人的な癖など）の影響が高まる懸念も生じる。

６．２　「〜を駆ける」

前節では、「〜が走る」という比較的頻度が高い例を見たが、本節では比較的頻度が低い「〜を駆ける」を見てみる。BCCWJ では共起語のうち、頻度 3 以上のもので、(「なか」、「ウマ」、「間」、「上」のように共起語の分析に適さないものを除くと、道 10、廊下 4、階段 3、戦場 3、山 3、夜道 3、前 3 の 7 語のみである。頻度 2 のものはノイズ（誤り、個人的な癖など）が影響する可能性が高いので、対象外とするが、例えこれらを含めてもあと 13 語増えるのみであり、コロケーションを意味でカテゴリ化して示すことは難しい。一方、TWC では頻度 3 以上の語は 50 ほどあり、以下のようなカテゴリ化が可能である。ただし、頻度が「〜が走る」に比べる大分少ないので、カテゴリの頻度で順序を見るには適さないだろう。

表３　TWC における「〜を駆ける」の共起語

カテゴリ	共起語
空 95	空 61、大空 11、宇宙 8、天空 7、夜空 5、銀河 3
野山 59	草原 16、野 12、野山 11、山 7、森 6、原野 4、荒野 3
経路 47	道 15、廊下 13、路 11、階段 8
戦場 31	戦場 31
世界 27	世界 27
地 22	地 11、大地 8、大陸 3
区域 21	街 8、庭 4、コート 3、町 3、街なか 3
前・後 21	先頭 6、先 5、前 5、後ろ 5
時 16	時代 10、時 6
海 11	海 8、海原 3

７．まとめ

本稿では筑波ウェブコーパス構築に当たり、BCCWJ の「均衡性」に近づけ、ウェブコーパスの弱点であるデータの偏りを回避する方略を提案した。また、NLB（NINJAL-LWP for BCCWJ）と同じレキシカルプロファイリング型のコーパス検索ツール NLT（NINJAL-LWP for Tsukuba Web Corpus）を使ってデータの抽出を行い、双方を比較した。動詞の頻度の比較では非常に高い相関が得られ、個々の動詞には両者の特徴が現れて違いが見られるところがあるものの、概ね両者の動詞の分布が近似していることが実証できた。また、コロケーションについては、データサイズの大きい筑波ウェブコーパスの方が安定的にコロケーション情報を抽出できることを示した。ただし、共起語出現語頻度が本稿で扱ったものより高いコロケーションの場合には､BCCWJ のサイズでも十分な情報が得られるであろうし、一方、共起語出現語頻度が本稿で扱ったものより低いものの場合には、筑波ウェブコーパスのサイズでもなお不十分ということも当然予想される。このような、より稀なコロケーション及びその他の稀なデータについてはさらに大きなサイズのコーパスが要求される。より大きなサイズのコーパスの構築においては、現実的に考えてウェブコーパスとならざるを得ないだろうから、今後の大規模コーパスの構築には本稿での知見が貢献できるとこ

ろも多いと思われる。

謝　辞

筑波ウェブコーパスの構築および NLT（NINJAL-LWP for Tsukuba Web Corpus）の開発には、教育関係共同利用拠点「筑波大学留学生センター　日本語・日本事情遠隔教育拠点」の予算の一部が充てられています。NLT は同上拠点事業としてウエブ上で公開予定です。NLT の基盤となった NLB（NINJAL-LWP for BCCWJ）は、協同研究として、筑波大学留学生センターが国立国語研究所および Lago 言語研究所から使用許可を得て使用しています。

文　献

Baroni, M. and Bernardini, S. (2004) *BootCaT: Bootstrapping corpora and terms from the web.* Proceedings of LREC 2004, Lisbon: ELDA. pp.1313-1316.
(http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf よりダウンロード可能)

Fletcher,W.H.(2007) *Toward cleaner Web corpora: recognizing and repairing problems with hybrid online documents*. Corpus Linguistics 2007, Birmingham pp.27-30.
(http://webascorpus.org/CL07BhamWHFletcher.pdf よりダウンロード可能)

Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (Eds.) (2007) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

今井新悟、赤瀬川史朗 (2012)『日本語ウェブコーパスと BCCWJ コーパスの比較と日本語教育への応用』、2012 年日本語教育国際研究大会パネルセッション「日本語につながるコーパス研究―現状と今後の展望―」、日本語教育国際研究大会名古屋 2012 予稿集第 2 分冊、p.65.

プラシャント・パルデシ、赤瀬川史朗 (2012)『レキシカルプロファイリング手法を用いた BCCWJ 検索ツール NINJAL-LWP とその研究事例』、日本言語学会第 144 回大会ワークショップ「コーパス基盤の日本語研究の新地平」、日本言語学会第 144 回予稿集、pp.364-369.

Sharoff S. 2006. *Creating general-purpose corpora using automated search engine queries.* In Marco Baroni and Silvia Bernardini (Eds), *WaCky! Working papers on the Web as Corpus*, Gedit, Bologna.

関連 URL

NLB（NINJAL-LWP for BCCWJ）　http://nlb.ninjal.ac.jp/
Sketch Engine　https://the.sketchengine.co.uk/
BootCaT　http://bootcat.sslmit.unibo.it/
国研コーパス開発センター 超大規模コーパス　http://www.ninjal.ac.jp/corpus_center/ulc/

Web を母集団とした超大規模コーパスの設計

浅原 正幸 (国立国語研究所コーパス開発センター) *
前川 喜久雄 (国立国語研究所言語資源研究系/コーパス開発センター)

A Design of Web-scale Japanese Corpora

Masayuki Asahara (Center for Corpus Development, NINJAL)
Kikuo Maekawa (Dept. Corpus Studies/Center for Corpus Development, NINJAL)

1. はじめに

国立国語研究所では 2006 年-2010 年の期間に 1 億語規模の書き言葉コーパス『現代日本語書き言葉均衡コーパス』(BCCWJ)(前川 (2007); 前川・山崎 (2008)) を構築し、2011 年より一般公開している。BCCWJ は種々の母集団に沿った無作為抽出を実施することによって、高度な均衡性・代表性を備えた均衡コーパスとなっている。しかし、その規模は、現代のコーパス言語学の趨勢からすれば十分とはいいがたく、生起頻度の低い言語現象の被覆に問題がある。そのためより大規模な日本語コーパスの構築が望まれている。この問題を解消するため、国立国語研究所では 2011 年から 6 か年の期間で、Web を母集団とした 100 億語規模の超大規模コーパスを構築する計画に着手した。本発表では、超大規模コーパスをどのようにして構築するか、どのような情報を付与するか、どのような検索環境を提供するのかなど、設計について概説する。

2. 先行研究

Web スケールの言語資源として、クローラを利用して検索エンジンを運営している企業や掲示板・ウェブサイトをホストしている企業により提供されている語彙表や n-gram 統計情報がある。グーグルは「Web 日本語 N グラム第 1 版」(工藤・賀沢 (2007)) として、元データ 2550 億語/200 億文規模の語彙表・n-gram データを作成し、一般公開した。バイドゥ株式会社 (2010a) は 2000-2010 年にかけてのブログや掲示板のデータ 1000 万文を対象に、月毎のコーパス母集団を元に作成した「Baidu ブログ・掲示板時間軸コーパス」の語彙表・n-gram 統計情報を公開した。また、同時期にバイドゥ株式会社 (2010b) はモバイル検索向けに収集した Web データを元に作成した「Baidu 絵文字入りモバイルウェブコーパス」の語彙表・n-gram 統計情報も公開した。楽天は 2010 年より「楽天データセット」としてレビューデータなどを公開している (楽天技術研究所 (2010))。ヤフー株式会社は「Yahoo! 知恵袋」コーパス（2004 年 4 月-2009 年 4 月）(ヤフー株式会社 (2007); ヤフー株式会社 (2011)) を公開している。

自然言語処理を研究している機関においては、情報通信研究機構 (NICT)、京都大学などが、それぞれクローラを用いて Web アーカイブを構築し、整形したデータを一般公開している。

* masayu-a@ninjal.ac.jp

表 1　主な Web スケールの言語資源

一般企業	
グーグル	「Web 日本語 N グラム第 1 版」 元データ 2550 億語/200 億文規模の語彙表・n-gram データ。2007 年 7 月のスナップショット。
バイドゥ	「Baidu ブログ・掲示板時間軸コーパス」 ブログや掲示板データを対象にした語彙表・n-gram データ 2000-2010 年 7 月にかけてのブログや掲示板のデータ計 1000 万文。月毎に母集団を設定。
バイドゥ	「Baidu 絵文字入りモバイルウェブコーパス」 2010 年 6 月までにモバイル検索向けに収集したデータを元に作成された語彙表・n-gram 統計情報。
楽天技研	「楽天データセット」 (2010 年公開。以下は 2012 年 8 月公開版について) 楽天市場のレビュー (1660 万レビュー)、楽天トラベルのレビュー (465 万評価・レビュー) 、 楽天ゴルフのレビュー (32 万レビュー)、楽天レシピのレシピ情報 (44 万レシピ) ほか。
ヤフー	「Yahoo! 知恵袋」コーパス第二弾 2004 年 4 月-2009 年 4 月の QA 記事。質問数 2600 万、回答数 7300 万。
研究機関・大学・官公庁	
NICT	「日本語係り受けデータベース」Version 1.1 6 億ページ (約 430 億文規模) の係り受け関係 4.8 億対。 収集時期 2007 年 5 月 19 日-11 月 13 日。
京大	「京都大学格フレーム」(Ver 1.0) 2009 年 3 月公開。 約 16 億文規模のテキストから自動構築した約 4 万用言の格フレーム。
NDL	「インターネット資料収集保存事業」 国・自治体・法人・機構・大学などのサイトと電子雑誌の保存事業。
個人	
矢田	「日本語 Web コーパス 2010」 2010 年に ipadic-2.7.0 の見出し語をシードとし Yahoo! Web API から Web ページ。 HTML アーカイブ (1 億ページ, 非圧縮 3.25TB), テキストアーカイブ (非圧縮 395GB), N-gram コーパス (文字, 形態素) を配布。

例えば、情報通信研究機構は検索エンジン基盤 TSUBAKI (Shinzato et al. (2008)) を構築し、約 345GB(非圧縮) 規模の日本語係り受けデータベース (情報通信研究機構 (2011)) を公開した。京都大学は Web データ 16 億文を用いて自動構築した格フレームを公開した (河原・黒橋 (2006); 京都大学大学院情報学研究科黒橋研究室 (2008))。

官公庁においては、国立国会図書館 (NDL) は官公庁自治体のウェブサイトや冊子体から電子版に移行した雑誌の保存を目的として、インターネット資料収集保存事業 (国立国会図書館; 関根 (2010)) を 2006 年より本格事業化している。

個人でも矢田 (2010) が形態素解析用辞書 IPADIC の見出し語の Yahoo! Web API による検索結果を収集することで約 396GB 規模 (非圧縮) のテキストアーカイブを作成し公開している。

表 1 に、一般に公開されている主な Web スケールの言語資源を示す。さまざまな技術の集積により、検索エンジンを運営している企業やコンテンツを保持している企業だけでなく、個人でも Web スケールの言語資源を構築することが可能になっている。

3. 超大規模コーパスの概要設計

本節では既存の技術を用いていかにして超大規模コーパスを構築するか、また、自然言語コーパスとしての可用性をあげるためにどのような工夫を行うかについてくわしく説明する。

表 2　超大規模コーパスの概要設計

- 収集
  - クローラ
    - Heritrix 3.1 系
- 構造化
  - 正規化技術
    - nwc-toolkit
  - 形態素解析
    - MeCab/UniDic (国語研短単位, UniDic 品詞体系)
    - CRF++ (国語研長単位, UniDic 品詞体系)
    - JUMAN (益岡・田窪品詞体系)
    - 教師なし形態素解析 (単語分かち書きのみ)
  - 係り受け解析
    - CaboCha/京都大学テキストコーパス
    - CaboCha/BCCWJ アノテーション基準
  - レジスタ分析
    - BCCWJ メタデータ相当情報推定
    - スパムサイト等判定
    - クラスタリングによる文体論的分析
- 利活用
  - 検索アプリケーション
    - 文字列検索 (+レジスタによるファセット分析)
    - 品詞検索 (中納言相当)
    - 係り受け検索 (ChaKi 相当)
  - 語彙表・n-gram データ
    - 語彙表 (出現形, 形態論情報を含む)
    - n-gram データ (基本形, 形態論情報を含まず)
    - 係り受け部分木 (基本形, 形態論情報を含まず)
  - 言語解析器
    - UniDic 未登録語調査
    - 頻度・共起情報を用いた言語解析器の改善
- 永続保存
  - ファイル形式
    - WARC 形式 (ISO-28500)
  - 情報アクセス
    - Open Source Wayback (ハーベスト)
    - NutchWAX (検索)
  - キュレーション
    - Web Curator Tool

我々が構築する予定の超大規模コーパスの概要設計について、**収集・構造化・利活用・永続保存**の四つの観点から解説する：

収集： Web コーパスを構築するための Web テキストの収集は Web クローラを用いることによる。約 1 億 URL を三か月ごとに収集し、一つの URL に対し、複数の版を取得する。

構造化： Web コーパスを言語研究に利用可能にするためものである。一般的な Web コーパスで用いられている正規化技術・形態素解析だけでなく、係り受け解析・レジスタ推定を行い、言語コーパスとしての信頼性を高める。

利活用： 構造化されたデータから、言語研究に必要な語彙表/n-gram データを整備する。100 億語規模のテキストから特定の形態論・統語論的パタンの事例を効率的に検索するアプリケーションを構築する。

永続保存： 言語の経年変化を観察するための資料として、収集したコーパスは Web アーカイブとして永続保存する。収集時期を時間軸とした組織化を行う。

表 2 に概要設計を示す。以下、各項目について詳説する。

3.1　収集

Web テキストの収集手法はクローラの運用 (Remote harvesting)、コンテンツ会社からの提供 (Database archiving)、検索エンジン/ソーシャルネットワーキングサービス会社が提供する Web API(Transactional archiving) の利用などがある。本研究では継続的に収集を行うために、バルク収集が可能なクローラ Heritrix [(1)] を運用する。Heritrix クローラは、wayback machine と呼ばれる Web アーカイブに実績を持つ米国 Internet Archive が中心となり開発しているク

ローラソフトウェアである。各国国立図書館が Web アーカイブを構築するために利用しており、日本では国立国会図書館がインターネット資料収集保存事業において利用している。アーカイブの保存形式は、後述する Web アーカイブの標準化ファイル形式である WARC 形式が選択できる。

各国国立図書館で運用するクローラは画像ファイル・音声ファイル・動画ファイルも含めたバルク収集ができることが重要である。しかしながら本研究においてはテキストデータの収集が主な目的であるために、.html ファイル・.txt ファイル・.xml ファイルに限定して収集する。

約 1 億 URL をシード URL リストとして、年に 4 回のペースで定点観測的に Web テキストとリンク-被リンク構造の収集を行う。収集対象は基本的に日本語の Web ページとする。日本語であれば スパムサイト (splog) であろうと機械翻訳結果であろうと収集を行い保存する。外国語などのデータは後述するレジスタ推定などにより定期収集対象から除外するが、収集したものを削除することは行わない。

2012 年 7 月に 100 万 URL 規模の第一次収集テスト、2012 年 8-9 月に 5000 万 URL 規模の第二次収集テストを行い、クローラの設定を検討した結果、週次の収集量を 1000 万 URL 程度とし、3 か月ごとに 1 億 URL 規模の収集を行うことにした。2012 年第四四半期から本収集（第一期）開始し、2013 年 1 月現在、本収集（第二期）を行っている。今後一年かけて同様の本収集を行い、URL の更新頻度推定などを行う。二年目以降は更新されない URL を収集範囲から外したうえで、新しい URL を収集範囲として含め、収集範囲の拡充を行う。収集範囲の拡充においては、代表性・均衡性ではなく網羅性を重要視する。コーパスの分布を特徴づける統計量のうち、分散が大きくなるように、また、尖度が小さくなるようにするが、歪度については制御しない。

3.2 構造化

Web テキストは収集しただけではそのままコーパスとして用をなさない。以下では、HTML タグ排除や文字コードの統制などの**正規化**、言語解析としての**形態素解析**、**係り受け解析**、コーパスとしての母集団を規定するための基礎情報となる **レジスタ推定** について説明する。

正規化：収集した Web テキストは、HTML タグを含んでいるだけでなく、文字コードが多様である。さらに言語コーパスとして扱うためには、一般的に分析に利用される単位である文境界の認定が必要になる。この HTML タグの排除・文字コードの統制・文境界の認定を Web テキストの正規化と呼ぶ。Web データの正規化については、2 節に示した先行研究の中で、グーグル「Web 日本語 N グラム第 1 版」[(2)] が採用している手法が事実上の標準となっており、これに準じた正規化が行える日本語ウェブコーパス用ツールキット (nwc-toolkit) [(3)] が公開されている。

Web テキストの正規化の問題のほかに、異なる URL で全く同じ Web ページであるかいなか、同じ URL に対する異なる収集時期の版であるかいなかを検出する技術を重複性・同一性検出と呼ぶ。重複性・同一性検出は Web ページのハッシュ値比較により行うことが一般的であるが、本研究でも同様の重複性・同一性検出を行う。

形態素解析：収集し、正規化を行った Web テキストに対して、形態素解析を行う。UniDic

が採用している国語研短単位は形態論的な情報に基づき、単位に斉一性があり、音韻的な情報が豊富であり、音韻論・形態論的な言語分析を行うには適した単位である。日本語教育などの分野で行われるコロケーション分析では、国語研短単位では粒度が細かく、より長い単位である国語研長単位で言語分析を行う傾向にある。一方、係り受けなどの統語分析を行う研究者は、UniDic が採用している可能性による品詞体系では必要な情報が可能性の名のもとに未定義となり利用できないため、益岡・田窪文法に基づく品詞体系とその品詞に基づいた文節単位を利用する傾向にある。さらに、言語処理の研究者で Web 上に頻出する辞書に登録されない形態を中心に分析するものもいる。

このような多様な利用者を想定して、本研究では形態素解析手法として、MeCab[4] /UniDic[5] による国語研短単位解析、汎用チャンカー CRF++ [6] による国語研長単位解析、JUMAN [7] による益岡・田窪品詞体系に基づく解析、ベイズ階層言語モデルによる教師なし形態素解析持橋ほか (2009) の四つを利用する。

係り受け解析：形態論情報において多様な品詞体系・単位が選択されるように、係り受けアノテーション基準もコーパス間で差異がある (浅原 (2013))。係り受け解析手法として、京都大学テキストコーパス[8] の基準に基づいて学習した CaboCha [9] （益岡・田窪品詞体系に基づく形態論情報）・BCCWJ 基準 (浅原・松本 (2013)) に基づいて学習した CaboCha （UniDic 品詞体系に基づく形態論情報・国語研文節単位）の二つを利用し、双方の基準による係り受け木を作成する。

レジスタ推定：言語学の観点からすると、Web コーパスの信頼性を下げる大きな要因のひとつは、収集されたテキストがどのような目的で書かれているかというレジスタ情報の欠落である。そのため本コーパスでは、収集された Web ページのレジスタ推定を実施する。

収集の時点では、シード URL からリンク構造をたどることによりクロールするため、自然言語コーパスとして均衡性・代表性を持たせた母集団を規定することが困難である。分散を大きく、尖度を小さくするようなクローラ運用ポリシーにより網羅性を重視したうえで、あらかじめ文書分類的な手法を用いて適切な部分サンプル集合をレジスタとして規定することにより、この問題を緩和する。

具体的には、外国語・スパムサイト (splog)・機械翻訳や機械生成されたテキストで非文法的なものを排除するための分類（(半) 教師あり機械学習）、BCCWJ に付与された各種メタデータ・ファイル単位アノテーションを推定するための分類（(半) 教師あり機械学習）、クラスタリングに基づく分類（教師なし機械学習）などを検討している。教師あり機械学習は、多クラスのトランスダクティブ SVM [10] による境界事例分析と、ランダムフォレスト法やブースティング法[11] による有効特徴量分析を行い、クラスタリングによる分類については得られたクラスタに対して言語学（文体論）的な見地からの分析を行う。教師なし機械学習においては、文書集合をどのような特徴量空間に写像するか (持橋ほか (2005)) の検討を行う。

3.3 利活用

構造化されたコーパスとして利活用していくうえで必要な環境整備について、**検索アプリケーション** と **語彙表・n-gram 頻度情報** について説明する。また利活用の事例として想定し

ている**言語解析技術への利用**についても述べる。

検索アプリケーション：構築したコーパスを計算機の扱いが不得手な研究者が利用可能にするために、高速な検索アプリケーションを提供する。レジスタに基づいた絞込（ファセットナビゲーション）を可能にする高速文字列検索機能、コーパス検索アプリケーション「中納言」[12] のような品詞情報に基づいた検索機能、コーパスアノテーション支援環境「ChaKi」[13] の Dependency Search のような係り受けの部分木構造に基づいた検索機能を、100 億語規模で現実的に動作する機能に絞って提供する予定である。

語彙表・n-gram 頻度情報：3 カ月おきにクロールするデータに対して構造化を行ったうえで、語彙表 (1-gram 頻度情報; 形態論情報を含む; 出現形に基づく)・文字列上の n-gram 頻度・形態素列上の n-gram 頻度情報 ($n \geq 1$; 形態論情報を含まない; 基本形に基づく)・文節係り受け木上の部分木頻度情報を収集時期ごとに区切られたサンプル単位で取得する予定である。尚、この語彙表・n-gram 頻度情報を得る母集団は、分散を大きく尖度を小さくするように収集を行うが、歪度については制御しないために代表性は担保されない。n-gram データの構築には FREQT[14] を利用する。また、別処理により、HTML タグの頻度情報・リンク-被リンク関係・同一コンテンツ関係など Web テキスト特有の情報を取得し保持する。可能であれば、レジスタ推定時にこれらの情報を活用する。

言語解析技術への利用：得られたコーパスを用いた言語解析技術の向上手法について検討を行う。形態素解析においては、教師なし形態素解析技術や未知語処理技術により得られた UniDic 未登録語について、人手で形態論情報を付与することにより辞書の拡充を行う。他の言語解析器については、教師なし機械学習に基づく手法の n-gram 頻度情報や部分木頻度情報を用いた各種言語解析技術の性能改善手法について検討を行う。

3.4 永続保存

収集したデータは言語の経年変化を分析するための基礎データとするために永続保存する。IIPC(International Internet Preservation Consortium)[15] における各国国立図書館の活動動向を見ながら、保存のための構造化を行う。

具体的には Heritrix で収集されたデータは、Web アーカイブの保存形式の国際標準 WARC 形式[16] で保存する。WARC ファイルは Internet Archive が公開している Wayback Machine [17] と同じ機能を持つハーベストソフトウェア Open Source Wayback[18] と、情報検索システム NutchWAX (Nutch Web Archive eXtension) [19] により構造化し、Web アーカイブとしての情報アクセスを可能にする。また、選択的な Web クロールを可能にするためのキュレーションツール WCT (Web Curator Tool)[20] の技術調査を行う。日本におけるコーパス言語学は、表層的な情報を用いた統計的手法に基づく分析に偏重しがちだが、用例・事例分析に基づくキュレーション分析に回帰すべく、アノテーションを効率的に行う環境を構築する。

最後に、長期保存可能な記憶媒体を機構内外に確保し、収集し構造化したデータの保存に努める。

4. おわりに

本稿では、現在国立国語研究所コーパス開発センターの「超大規模コーパス構築プロジェクト」で整備を進めている Web スケールのコーパスの概要設計を解説した。表 3 に現状の工程表を示す。

以下、進捗について示す。2011 年度後半に計画立案を行った。2012 年度は主に収集技術・テキストの正規化技術・形態素解析技術・文字列検索技術・保存技術の調査を行った。収集技術に関しては実際にクローラの運用テストを行いながら運用規則の策定を行い、現在クローラの本運用を開始している。今後定期的に運用規則を見直しながら収集作業をすすめていきたい。またテキストの正規化・形態素解析の構造化環境を構築し、係り受け解析のテスト環境を構築中である。2013 年度は、テキストの正規化技術・形態素解析関連技術を運用レベルにあげ、文字列検索技術の調達を開始する。係り受け解析技術の既存技術については調査を行うとともに年度末までに運用レベルにする。技術調査としてはレジスタ分析技術と品詞・係り受け構造に基づく検索技術を対象とする。2014 年度以降、細部については修正の可能性もあるが、大方はこの工程表に準じて構築をすすめる予定である。

本研究に関する意見・要望・疑問点などについては第一著者まで。

謝辞

本研究は国立国語研究所コーパス開発センターの「超大規模コーパス構築プロジェクト」によるものである。本研究を行うにあたり、情報通信研究機構ユニバーサルコミュニケーション研究所の諸氏および統計数理研究所の持橋大地氏よりさまざまな技術指導をいただいた。国立国語研究所コーパス開発センターの諸氏から設計時点での有益なコメントをいただいた。ここに記して謝意を表す。

参考文献

Shinzato, K., T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi (2008). “Tsubaki: An open search engine infrastructure for developing new information access.” *IJCNLP-2008*.

浅原正幸 (2013).「係り受けアノテーション基準の比較」 第 3 回コーパス日本語学ワークショップ.

浅原正幸・松本裕治 (2013).「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」 第 19 回言語処理学会年次大会 (NLP2013).

河原大輔・黒橋禎夫 (2006).「高性能計算環境を用いた Web からの大規模格フレーム構築」情報処理学会自然言語処理研究会 171-12 巻, pp. 67–73.

京都大学大学院情報学研究科黒橋研究室 (2008).『京都大学格フレーム (Ver 1.0)』, http://www.gsk.or.jp/catalog/GSK2008-B/catalog.html.

工藤拓・賀沢秀人 (2007).『Web 日本語 N グラム第 1 版』, 言語資源協会発行 http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html.

国立国会図書館『インターネット資料収集保存事業（ウェブサイト別）』, http://warp.da.ndl.go.jp/search/.

情報通信研究機構 (2011).『日本語係り受けデータベース Version 1.1』, https://alaginrc.nict.go.jp/resources/nictmastar/resource-info/abstract.html#A-8.
関根麻緒 (2010).「国立国会図書館のインターネット情報の制度的収集」 図書館雑誌, 104:5, pp. 288.
バイドゥ株式会社 (2010a).『Baidu ブログ・掲示板時間軸コーパス』, http://www.baidu.jp/corpus/.
バイドゥ株式会社 (2010b).『Baidu 絵文字入りモバイルウェブコーパス』, http://www.baidu.jp/corpus/.
前川喜久雄 (2007).「コーパス日本語学の可能性—大規模均衡コーパスがもたらすもの—」 日本語科学, 22, pp. 13–28.
前川喜久雄・山崎誠 (2008).「『現代日本語書き言葉均衡コーパス』」 国文学解釈と鑑賞, 932(74 巻 1 号), pp. 15–25.
持橋大地・菊井玄一郎・北研二 (2005).「言語表現のベクトル空間モデルにおける最適な計量距離」 電子情報通信学会論文誌, J88-D-II:4, pp. 747–756.
持橋大地・山田武士・上田修功 (2009).「ベイズ階層言語モデルによる教師なし形態素解析」 情報処理学会研究報告:2009-NL-190.
矢田晋 (2010).『日本語ウェブコーパス 2010 (NWC 2010)』, http://s-yata.jp/corpus/.
ヤフー株式会社 (2007).『Yahoo! 知恵袋データ (第 1 版)』.
ヤフー株式会社 (2011).『Yahoo! 知恵袋データ (第 2 版)』, http://www.nii.ac.jp/cscenter/idr/yahoo/chiebkr2/Y_chiebukuro.html.
楽天技術研究所 (2010).『楽天データセット』, http://rit.rakuten.co.jp/rdr/index.html.

関連 URL

(1) クローラ Heritrix-3.1.1：http://webarchive.jira.com/wiki/display/Heritrix/Heritrix

(2) Google Web 日本語 N グラム第 1 版 README：http://www.gsk.or.jp/catalog/GSK2007-C/GSK2007C_README.utf8.txt

(3) 日本語ウェブコーパス用ツールキット：http://code.google.com/p/nwc-toolkit/

(4) 形態素解析器 MeCab-0.995：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

(5) 形態素解析用辞書 UniDic-2.1.1：http://sourceforge.jp/projects/unidic/

(6) 汎用チャンカー CRF++-0.57：http://crfpp.googlecode.com/svn/trunk/doc/index.html

(7) 形態素解析器 JUMAN-7.0： http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/juman/juman-7.0.tar.bz

(8) 京都大学テキストコーパス 4.0：http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0.tar.gz

(9) 日本語係り受け解析器 CaboCha-0.66：http://code.google.com/p/cabocha/

(10) SVMLin：http://vikas.sindhwani.org/svmlin.html（多クラスのトランスダクティブ学習が可能）

(11) BACT：http://chasen.org/~taku/software/bact/（部分木を特徴量とする決定株を弱学習器としたブースティング）

(12) コーパス検索アプリケーション「中納言」1.0.5：http://chunagon.ninjal.ac.jp

(13) コーパスアノテーション支援環境「ChaKi」version 2.3：http://sourceforge.jp/projects/chaki/releases/

(14) 頻出部分木マイニングプログラム FREQT-0.22： http://chasen.org/~taku/software/freqt/

(15) IIPC(International Internet Preservation Consortium)：http://netpreserve.org/

(16) ISO 28500:2009, Information and documentation – WARC file formathttp://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

(17) Wayback Machine – Internet Archive：http://archive.org/web/web.php

(18) Open Source Wayback-1.6.0：http://archive-access.sourceforge.net/projects/wayback/

(19) Nutch Web Archive eXtension-0.13：http://archive-access.sourceforge.net/projects/nutch/

(20) Web Curator Tool-1.6：http://webcurator.sourceforge.net/

表 3　超大規模コーパスプロジェクト：工程表

年	2011	2012				2013				2014				2015				2016
四半期	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q
準備	⇒ 計画立案																	
	⇒ 機材調達 (初回)			⇒ (2 回目)				⇒ (3 回目)				⇒ (4 回目)					
収集	(クローラ関連)																	
		⇒ クローラ運用テスト																
				⇒ クローラ本運用開始														
		⇒ 運用規則策定			⇒ 運用規則見直し (初回)				⇒ 運用規則見直し (2 回目)				⇒ 運用規則見直し (3 回目)					
構造化	(正規化)		⇒ 正規化技術調査															
						⇒ 正規化技術運用開始												
	(形態素解析)				⇒ 形態素解析技術調査（既存技術）													
								⇒ 形態素解析運用開始（既存技術）										
	(係り受け解析)					⇒ 係り受け解析技術調査（既存技術）												
								⇒ 係り受け解析運用開始（既存技術）										
									⇒ 係り受け解析技術調査・開発（新規技術）									
	(レジスタ分析)			⇒ BCCWJ メタデータ関連調査														
						⇒ splog 検出技術調査												
								⇒ クラスタリングによる文体論的分析技術調査										
									⇒ 実装・並列化									
										⇒ レジスタ分析技術運用開始								
利活用	(検索アプリケーション)		⇒ 文字列検索技術調査															
						⇒ 文字列検索技術調達開始												
						⇒ 品詞検索技術調査				⇒ 品詞検索技術調達開始								
						⇒ 係り受け検索技術調査				⇒ 係り受け検索技術調達開始								
															⇒ 内部公開			
																⇒ 外部公開		
	(語彙表・n-gram データ)								⇒ 語彙表作成開始									
									⇒ n-gram データ作成開始									
											⇒ 係り受け部分木データ作成開始							
	(言語解析器)							⇒ 未登録語調査 (初回)				⇒ (2 回目)				⇒ (3 回目)		
													⇒ 言語解析器の改善					
保存			⇒ 技術調査															
						⇒ Open Source Wayback 運用開始												
										⇒ NutchWAX 運用開始								
															⇒ 保存媒体の確保			

『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の設計と特性について

宇佐美 まゆみ（東京外国語大学大学院総合国際学研究院）†
中俣 尚己（京都教育大学教育学部）‡

Design and Characteristics of the "Corpus of Spoken Japanese by BTSJ (Transcription and Audio Recordings) ver.2011"

Mayumi Usami (Graduate School of Tokyo University of Foreign Studies, Institute of Global Studies)
Naoki Nakamata (Kyoto University of Education)

１．はじめに

近年、コーパス日本語学が盛んになりつつあるが、その多くは書き言葉のコーパスに関するものであり、「話し言葉コーパス」に基づくものは多いとは言えない。分析の観点も、書き言葉の特性を考えると当然かもしれないが、形態素解析や語彙や構文の分析、コロケーション研究などが中心で、語用論的分析は未だ手つかずの状態である。一方、「話し言葉のコーパス」も増えつつはあるが、人間の相互作用としての「自然会話（事前の計画がないやりとり）」を編んだコーパスは、ほとんどないといっても過言ではない。日本語学習者の口頭能力試験を集めた学習者コーパスなどはいくつかあるが、これらは口頭能力試験という特殊な状況における相互作用であり、分析の観点も、未だ文法項目の習得などに焦点を当てたものが多い。「会話分析」としては、エスノメソドロジーに端を発する CA（Conversation Analysis）が盛んであるが、基本的に、CA は、対人コミュニケーションの理論化や一般化を目的とはしていないこともあり、その「文字化システム」は、「定性的分析」には適しているかもしれないが、「定量的分析」には適しているとは言えない。昨今公開されている「話し言葉のコーパス」も、語用論的分析に適した「文字化システム」に基づくものはほとんどない。話し言葉コーパスに基づく分析は、講演などのストレート・トークやナラティブ・データに基づいた音声学的な分析などが緒についたところであると言ってもよいだろう。すなわち、人間の相互作用の分析を企図し、会話の定性的分析に加えて、定量的な分析も可能にする形で文字化し蓄積された「話し言葉のコーパス」は、未だほとんどないのが現状である。その理由の一つに、話し言葉をデータとして用いる研究では、会話の収集、文字化といった基礎的作業をはじめ、その後の分析対象のコーディングなどにも膨大な時間と労力を要するということがある。そのため、会話や話し言葉の対人コミュニケーション論的・語用論的分析を、より効率的に進めていくためには、研究者間で自然会話データを共有していくことが不可欠である。また、そのためには、発話の重なりや沈黙などの語用論的分析に必須の情報を記述し、且つ、定量的分析にも適する文字化システムによって蓄積された「話し言葉コーパス」が必須である。このような認識に基づいて、筆者とその研究協力者らは、ここ 15 年来、あくまで人間の相互作用としての「言語の運用」に焦点を当て、対人コミュニケーション論、語用論の観点から「会話の分析」を行い、定量的分析ができる形で文字化したデータを蓄積し、一般公開も行ってきた。それらを改訂し、改めてまとめ直したのが、『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』（以降「BTSJ 話し言葉コーパス」と略記）である。本稿では、その開発・設計の趣旨、及び、その特性と活用方法を簡単にまとめる。

† usamima@tufs.ac.jp ‡ nakamata@kyokyo-u.ac.jp

２.『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の設計の趣旨と特性

本節では、BTSJ 話し言葉コーパスの設計の趣旨と特性を簡単にまとめる。

２．１　BTSJ 話し言葉コーパス設計の趣旨

「１．はじめに」でも述べたように、本コーパス設計の趣旨は、「相互行為としての会話」の対人コミュニケーション論、語用論的分析に適したコーパスを構築することである。そのために重視した点は、以下の３点である。①「言語社会心理学的アプローチ」(宇佐美 1999)、「総合的会話分析」（宇佐美 2008）の方法論に基づき、会話参加者の年齢、性別、話題などを統制したデータ群を収録する。②発話の重なりや沈黙など、語用論的分析に不可欠な情報を記して細やかな定性的分析を可能にするとともに、分析項目のコーディングや集計などの定量的分析も行いやすい「基本的な文字化の原則」である BTSJ（Basic Transcription System for Japanese）によって文字化したトランスクリプトの形で提供する。③「人間の相互作用としての会話分析」は、「会話自体」の分析のみならず、「録音された会話以外の社会的要因」の分析も重視する。そのため、各会話グループのデータ収集条件や話題、話者の年齢・性別・職業、その他の属性をまとめたエクセルファイルも収録する。

２．２　BTSJ 話し言葉コーパスの概要と特徴

『BTSJ による日本語話し言葉コーパス』は 294 の相互作用的会話からなる[1]。会話の総時間は 67 時間 21 分 39 秒、総語数は、789,190 語[2]である。すべての会話は、発話の重なりや沈黙、割り込みなどの語用論的分析に必須の情報を記述するための原則である「基本的な文字化の原則（Basic Transcription System for Japanese：BTSJ）2011 年版」に基づくトランスクリプトの形になっており、約 30%（20 時間分）の会話には、プライバシー保護処理をした「音声資料」がトランスクリプトとともに提供されている。BTSJ トランスクリプトは、多くの人が活用しやすいことを考え、エクセル形式で保存されている。利用する研究者各自が、「発話内容」の右側に「コーディング」の列を追加して分析したい項目をコーディングすれば、エクセルの機能で話者ごとにソートして話者の特徴を概観したり、コーディング項目の頻度の集計などを行うことができるが、2007 年に、エクセルに専用のマクロ機能を搭載して「BTSJ 入力支援・自動集計システムセット」を開発し、対となる記号の自動入力やエラーチェック機能等の「入力支援機能」を付与し、コーディング項目の基本的記述統計を自動集計して表の形で自動表示できるようにした。さらに、2011 年には、同じルールでコーディングした複数の会話ファイルの分析項目の頻度や割合の合計、平均、標準偏差などの自動集計も可能にした。このシステムセットは、現在のところ、「BTSJ 活用方法講習会」[3]（宇佐美 2012）の受講者に無償で配布している。また、テキストファイルに変換して利用することもできる。本コーパスは、事前の計画や準備のない自然会話を中心とするコーパスであるが、一部、電話会話やロールプレイ等も収録されており、日本語母語話者の会話のみならず、接触場面（日本語母語話者と日本語非母語話者）の会話も豊富である。初対面、友人同士、話者の年齢に上下のある会話、同年齢同士の会話、同性同士の会話、異性との会話、教師と学生の面談会話等々、様々な種類の会話が、話者の社会的属性や場面等の諸条件を統制して収集され、収録されている。そのため、話者の社会的属性や話者同士の関係、場面に応じた話し方の特徴や違いを、様々な角度から比較・検討することが可能である。この点が、BTSJ 話し言葉コーパスの最大の長所であり、特徴である。

---

[1] 予稿集では、改訂中の 1 会話を除いた数値を提示したが、本稿では、コーパスに収録されている 294 会話すべてを含めて算出した数値を提示する。これ以降の表 1 などの数値についても同様である。

[2] Mecab+UniDic による。句読点等を除く実質的発話部である。

[3] これまでのところ不定期に、東京、広島、京都、九州、ベルリン、ロンドンで開催している。問い合わせ：言語社会心理学研究会事務局：btsjworkshop@gmail.com

２．３　BTSJ（Basic Transcription System for Japanese）の基本原則と形式

すべてのトランスクリプトは、BTSJ によって記述されており、xlsx 形式のエクセルファイルで提供される。BTSJ によるトランスクリプトの一例を以下の図１に示す。

上部には「会話グループ名」、「会話記号（ファイル名に対応）」、「話者記号の凡例」、「会話番号」、「時間」、「１会話における話者数」の６つの情報が記載されている。その下に発話内容（トランスクリプト）が記される。左には「ライン番号」、「発話文番号」、「発話文終了」、「話者」を記す。

BTSJ では、「発話文」の定義は、「会話という相互作用の中における文」とし、以下のように認定する。基本的に、ひとりの話者による「文」を成していると捉えられる発話を「1 発話文」とする。しかし、自然会話では、いわゆる「1 語文」や、述部が省略されているもの、あるいは、最後まで言い切られない「中途終了型発話」など、構造的に「文」が完結していない発話もある。そのような場合は、話者交替や間などを考慮した上で「1 発話文」であるか否かを判断する。つまり、「発話文」の認定には、「話者交替」、「間」という 2 つの要素が重要になる。そのため、途中で相手の発話が入って話者が一旦交替したため改行され、複数のラインに渡っている発話も、同一話者によって発せられた「１文」を成していると捉えられるものは、複数のラインにまたがる発話をまとめて「１発話文」とする。そして、図１の「発話文番号」の列における「3-1」、「3-2」のように、異なるラインにまたがっていても同じ発話文であることがわかるように同じ番号をつけ、その後に「-」をつけて発話された順を記す。また、完結していないほうの発話には、「発話文終了」の欄に「/」を記す。１会話の「発話文数」は、「発話文番号」が示すとともに、左から３行目の「発話文終了」の列が、発話文が完結していることを表す「*」となっているものを数えてもわかるようになっている。また、「発話内容」の列における「。」も、BTSJ のルールでは発話文の完結を意味するため、質問発話で文末に「?」があっても、文が完結している場合は、「?。」と、必ず、最後に「。」をつける。そのため、「*」と「。」の数が同じになることを利用して、通常のエクセルでも「*」と「。」を数えることによって、発話文数の検算もできる。

会話グループ名：台湾人学習者（上級）と日本人の友人の雑談	会話記号：TF01-JF01	記号凡例： TF：Taiwanese Female JF：Japanese Female
会話番号：142	時間：0分0秒-20分40秒(終)	1会話における話者の数：2

ライン番号	発話文番号	発話文終了	話者	発　話　内　容
1	1	*	TF01	####しゃべってください。
2	2	*	JF01	えっ、何をしゃべるんですか<笑い>。
3	3-1	/	TF01	《沈黙2秒》ね、しゃべらないと,,
4	4	*	JF01	<笑いながら>しゃべれないよ。
5	3-2	*	TF01	また私ずっとしゃべってる感じだ(<笑い>)と思いました。
6	5	*	TF01	あっ[[。
7	6	*	JF01	]]あっ、お菓子食べる<時に…>{(}。
8	7	*	TF01	<そう>(}。
9	8	*	JF01	かな。

図１　BTSJ によるトランスクリプトの例

発話内容には種々の記号を用いて、相互作用に関する情報が付与されている。図１には「沈黙」「笑い」「発話の重なり」「さえぎり」等の情報が記載されている。その他にも「引用部」「イントネーション」「ラッチング」「言い淀み」「文脈情報」などの情報が付与されている（記号の意味など、BTSJ に関する詳細は、宇佐美(2011)を参照）。ただし、各研究者が自身のデータを BTSJ で文字化する場合は、研究目的に応じて、BTSJ で定められた記号を変更しない限りにおいて、独自の記号を追加して特定の現象の記述をより詳細にしたり、逆に、小声のあいづちは文字化しない等の原則を設けて簡略化することも可能である。

２．４　『BTSJ 文字化入力支援・自動集計・複数ファイル自動集計システムセット（2012 年改訂版）』について

BTSJ は、あくまで「文字化のルール」である。そして、『BTSJ 文字化入力支援・自動集計・複数ファイル自動集計システムセット（2012 年改訂版）』は、BTSJ による「文字化」にかかる時間と労力を軽減するための「文字化入力支援機能」と、BTSJ で記されたトランスクリプトにコーディングを行った項目の基本的な記述統計に必要な情報を算出する「自動集計機能」を搭載したシステムセットである。利用者の利便性や汎用性を考えて、Microsoft Excel のマクロ機能を利用して作成されており、「BTSJ 入力支援・自動集計システム(.xlt)」、「BTSJ 複数ファイル自動集計システム(.xls)」の 2 つのファイルから成っている。現在は、日本語版 Windows の Excel 2003、2007、2010 に対応している。（ただし、英語版 Windows でも、Excel 上で日本語を表示できる環境であれば、問題なく使える。また、Mac の場合は、windows をインストールするか、シトリックス http://www.apple.com/jp/business/profiles/citrix/ などの仮想デスクトップを導入する必要がある。）

３．『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の定量的な基本情報

本節では、『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の定量的な基本情報を、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と略記）と比較する形で示す。

３．１　基本情報

本節では、「BTSJ 話し言葉コーパス」に形態素解析を施した結果を示す[4]。まず、表１に、「総語数」、「異なり語数」などの本コーパスの基本情報を示す。

表１　『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の基本情報

会話数	294 会話
総語数	789,190 語
異なり語数	12,079 語
TTR（異なり語数／総語数）	1.530%
Guiraud 値（異なり語数／√総語数）	13.596
発話文数	91,256 文
１文あたりの語数（総語数／発話文数）	8.648 語
総時間	242,499 秒（67 時間 21 分 39 秒）
１文あたりの時間数（総時間／発話文数）	2.657 秒

注）なお、上記の語数には、句読点など、UniDic において「補助記号」に分類されるものは含まない。また、「記号」は、人名などが記号で表されることもあるため（例：F さん）、数値に含めている。

３．２　動詞の高頻度語

「基本語彙」の選定は外国語学習の分野において極めて重要である。本節では、まず、BTSJ 話し言葉コーパスにどのような動詞が多く見られたかを算出し、BCCWJ と比較する。

次頁の表２に、BTSJ 話し言葉コーパスにおける高頻度の動詞、上位 20 と、BCCWJ における高頻度の動詞、上位 20 を、それぞれ 1 万語あたりに換算し頻度とともに示す。太字ゴシックの語は当該コーパスでのみ上位 20 以内に入った語である。表２を見ると、上位 20 語のうち太字ゴシックの語を除く 16 語までが、両コーパスに共通していることがわかる。

[4] エクセルファイルのコーパスを csv 形式に変換後、発話内容の部分だけを取り出し、BTSJ 特有の記号を除去した後、茶まめ(UniDic+mecab)を用いる形で形態素解析を行った。

基本語彙の選定は、これまでも種々の立場から行われているが、中でも最も数が絞られているのは、国立国語研究所の『電子計算機による新聞の語彙調査』をもとに、林(1975)が「文句なしの基本語彙」とした 545 語であろう。その中に動詞は 73 語ある。BTSJ 話し言葉コーパスの高頻度語 20 の中では、「違う」と「書く」を除く 18 語がこの 73 語に含まれている。

また、BTSJ 話し言葉コーパスのみで上位 20 に入った「違う」「聞く」「書く」「取る」の 4 語は、BCCWJ でも、40 位以内に入っている。このことを考えると、これら 4 語は、書き言葉においても基本語に相当すると言ってもよいだろう。つまり、動詞の高頻度語は、コーパスの規模の大小、話し言葉、書き言葉の違いにかかわらず、ほとんど共通しているということと、「基本語彙」（林 1975）との共通性が高いということが明らかになった。

一方、BCCWJ では 12 位に入っている「おる」は、BTSJ 話し言葉コーパスでは、20 位までには入らず、61 例（1 万語あたり 0.77）しかなかったが、実は、BCCWJ の中でも、「国会会議録」に集中的に出現する語であることがわかった。このように、語用論の観点からは、大規模コーパス全体における単なる頻度の比較ではなく、ジャンルごとに分けてみた頻度やコロケーションの分析・考察が重要である。

表２　動詞の高頻度語の比較

順位	BTSJ 話し言葉コーパス	1 万語あたりの頻度	BCCWJ	1 万語あたりの頻度
1	言う	143.10	いる	109.73
2	する	119.69	する	60.51
3	ある	65.66	なる	48.69
4	行く	50.09	ある	47.20
5	思う	48.02	言う	30.43
6	やる	36.14	来る	22.83
7	いる	36.04	思う	20.35
8	なる	32.67	できる	13.29
9	来る	31.37	見る	18.14
10	分かる	23.61	行く	15.36
11	見る	23.12	**しまう**	9.69
12	**違う**	13.56	**おる**	9.60
13	できる	13.29	考える	9.34
14	入る	10.66	**持つ**	8.48
15	出る	10.06	分かる	8.08
16	**聞く**	9.73	出る	8.01
17	**書く**	9.41	やる	7.80
18	知る	8.77	**行う**	5.95
19	考える	7.39	知る	5.79
20	**取る**	7.31	**入る**	5.69

注 1)BCCWJ における「てる」は UniDic では助動詞となっているため、除外した。

注 2)BCCWJ のデータは Ninjal-LWP for BCCWJ Ver.1.10 を使用したため、BCCWJ のうち約 6 千万語分のデータにおける順位である。

３．３　副詞の高頻度語

次に、動詞と同様、BTSJ 話し言葉コーパスにおける副詞の高頻度語を、BCCWJ と比較する形で示す。次頁の表 3 に、BTSJ 話し言葉コーパスにおける高頻度の副詞上位 20 と、BCCWJ における高頻度の動詞上位 20 を、それぞれ 1 万語あたりの頻度とともに示す。

太字ゴシックの語は当該コーパスでのみ上位 20 に入った語である。動詞の結果とは対照的に、BTSJ 話し言葉コーパスの上位 6 語こそ BCCWJ においても上位語になっているが、7 位以下の 14 語のうち 12 語までが、BTSJ 話し言葉コーパスでのみ上位に入った語となっている。同様に、BCCWJ でも上位 20 語のうち 12 語は、BTSJ 話し言葉コーパスでは上位 20 に入っていない。また、先述した林(1975)の基本語彙の中には副詞が 55 語含まれているが、BTSJ 話し言葉コーパスにおける高頻度副詞 20 のうち、この副詞 55 語に含まれているのは、「もう」「やはり」「あまり」「まだ」「もし」「もっと」「いっぱい」「例えば」の 8 語のみであった。林(1975)は、新聞の語彙調査を元にしていることから、副詞の基本語彙は、話し言葉と書き言葉で、かなり異なっていることがわかる。

これらの結果から、副詞の用法は、話し言葉と書き言葉の違いを特徴づける語群の一つであると言えるだろう。

表３　副詞の高頻度語の比較

順位	BTSJ 話し言葉コーパス	1 万語あたりの頻度	BCCWJ	1 万語あたりの頻度
1	そう	167.91	そう	7.62
2	もう	36.21	どう	6.91
3	ちょっと	31.20	もう	5.14
4	こう	27.93	**さらに**	3.77
5	やはり	25.49	やはり	3.24
6	どう	25.22	まだ	3.04
7	**まあ**	22.93	よく	2.87
8	**結構**	16.16	**少し**	2.87
9	**あまり**	12.87	**すぐ**	2.52
10	**多分**	12.16	**まず**	2.52
11	**全然**	11.77	**特に**	2.48
12	まだ	8.52	**まったく**	2.37
13	よく	6.69	ちょっと	2.21
14	**ずっと**	5.12	**すでに**	2.14
15	**色々**	5.08	こう	2.08
16	**なるほど**	4.94	**実際**	2.02
17	**うんうん**	4.80	**ほとんど**	1.85
18	**例えば**	4.75	**最も**	1.74
19	**一番**	4.22	**初めて**	1.73
20	**ちゃんと**	3.78	**もちろん**	1.68

表 3 から、「そう」が双方のコーパスで 1 位であり、「やはり」が BTSJ 話し言葉コーパスで 6 位、BCCWJ で 5 位であることなどから、一見両コーパスで同様の傾向を示しているように見える。しかし順位が同じでも、1 万語あたりの頻度を見ると、話し言葉のほうがかなり多い。また、用例に目を通すと、「そう」は BTSJ 話し言葉コーパスでは、ほとんどが「あ、そうなんだ」「そうそう」「そうですか」のように応答に使われているのに対して、BCCWJ では「母はそう言った」のように具体的な指示内容を持つ用法が多いことがわかった。また、「やはり」の音形に着目すると、BTSJ 話し言葉コーパスでは、「やはり」（4%）、「やっぱり」（61%）、「やっぱ」（35%）であるのに対して、BCCWJ では、「やはり」（68%）、「やっぱり」（28%）、「やっぱ」（3%）となり、話し言葉と書き言葉の違いが顕著に見えてくる。このように、話し言葉と書き言葉の特徴を比較するためには、単なる順位や頻度の比較だけではなく、用例や音形、コロケーションなども考慮に入れた分析が必須である。

４.『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』を用いた語用論的分析

ここまでは、BTSJ 話し言葉コーパスの語彙的特性の概観を定量的観点から示した。しかし、本コーパスは、話者の社会的属性や場面などが統制されて収集されていることが最大の特徴であり、特定の場面やある属性をもつ話者のみを取り出して、その特徴や言語使用の分析を行うことのほうを重視している。この点は、様々のジャンル、媒体、文体等のデータを収録した大規模コーパスに基づいて、データのジャンルや属性の違いをあまり考慮せずに分析している研究が多い従来の「コーパス言語学」においては、あまり重視されていない点である。本節では、語用論的分析の一つとして、話者の属性（母語話者／非母語話者)、話者同士の関係（初対面／友人)、場面（母語場面／接触場面）を統制した形で、それぞれの条件における「異なり語数」と発話文末の「丁寧体率」（丁寧体／総発話文数）を算出することによって、それぞれの状況や話者の属性による語彙数やスピーチレベルの違いを明らかにする。

４．１　話者同士の関係、場面の違う会話における母語話者と非母語話者の「異なり語数」の比較

ここでは、上下関係のない２人の話者の会話において、丁寧体（「です」「ます」）の使用率が、話者の属性（母語話者／非母語話者)、話者同士の関係（初対面／友人)、場面（母語場面／接触場面）によって、どのように異なるかを分析する。そのため、「BTSJ 話し言葉コーパス」の中から、これらの条件に相当する会話を選び出し、「母語場面・初対面」「母語場面・友人」「接触場面・初対面」「接触場面・友人」の４つのグループに分け、それぞれの会話の発話内容のみを取り出し、茶まめ（mecab+UniDic）で形態素解析を行った。母語話者同士の会話である「母語場面・初対面」と「母語場面・友人」は、ファイル内のすべての発話内容を分析対象とし、母語話者と非母語話者の会話である「接触場面・初対面」と「接触場面・友人」では、非母語話者の発話だけを抽出することによって、母語話者と非母語話者という話者の属性による違いを分析した。以下の表４に、各会話グループの条件、属性ごとの話者数を示す。

表４　各会話グループの条件、属性ごとの話者数

グループ名	母語話者・初対面		母語話者・友人		非母語話者[*1]・初対面		非母語話者[*1]・友人	
母語場面／接触場面	母語場面		母語場面		接触場面		接触場面	
話者の関係	初対面		友人		初対面		友人	
性別組み合わせ	女性同士	24	女性同士	18	女性同士	24	女性同士	10
	男性同士	7	男性同士	3	男性同士	4	男性同士	0
	男女	3	男女	3	男女	0	男女	0
話者総数		66		54		28		10
年齢	20 代	54	20 代[*2]	54	20 代	28	20 代	10
	30 代	12	30 代	0	30 代	0	30 代	0
非母語話者の出身					台湾	24	台湾	10
					中国大陸	4	中国大陸	0
非母語話者の日本語レベル					超級	3	超級	0
					上級	22[*3]	上級	10
					中級	3	中級	0

*1　接触場面の会話における非母語話者を対象としている。　*2　10 代後半の数名を含む。
*3　うち４名は中国大陸出身である。

次に、各グループの語数などの基本情報を表５に示す。

表５　各会話グループの語数

グループ名	母語話者・初対面	母語話者・友人	非母語話者*・初対面	非母語話者*・友人
延べ語数	117,497	103,534	39,386	15,801
異なり語数	4,002	4,391	2,113	1,572
Guiraud 値	11.68	13.65	10.65	12.51

*接触場面における非母語話者の発話のみを分析対象としている。

表５において、語彙の豊富さの指標となる Guiraud 値を見ると、全体的には、母語話者のほうがやや高いが、母語話者の初対面会話よりも、非母語話者の友人場面のほうが、Guiraud 値が高くなっている。また、母語話者、非母語話者ともに、初対面会話より友人との会話のほうが、語彙使用の幅が広いということがわかる。これらの結果から、異なり語数については、母語話者、非母語話者の違いよりも、初対面会話か、友人との会話かという場面による違いのほうが大きいということが明らかになった。これは、初対面会話では、互いの自己紹介のように話題が画一的なものになる傾向があり（宇佐美・嶺田 1995)、用いられる語彙も限られたものになりがちであるということを示していると言えるだろう。

４．２　話者同士の関係、場面の違う会話における母語話者と非母語話者の「丁寧体率」の比較

日本語における対人コミュニケーションにおいては、相手や場面に応じて、丁寧体と普通体がどのように使い分けられるかということは、対人関係調整上、重要な意味を持つ。しかし、非母語話者にとっては、その使い分けこそが困難であることが指摘されている（宇佐美 1995、2001)。そこで、ここでは、総発話文数に占める文末の丁寧体（「です」「ます」）の割合を「丁寧体率」と呼び、話者同士の関係、場面の違う会話における母語話者と非母語話者の丁寧体率を比較する。まず、「です」、「ます」それぞれの頻度と総発話文数に占める割合を以下の表６に示す。また、「です」「ます」を合わせた「丁寧体」の頻度とそれが総発話文数に占める割合である「丁寧体率」を次頁の表７に示す。また、「丁寧体率」は、次頁の図２にも示した。

表６　話者同士の関係、場面の違う会話における母語話者と非母語話者の「です」、「ます」の頻度と割合の比較

グループ名	母語話者・初対面		母語話者・友人		非母語話者*1・初対面		非母語話者*1・友人	
丁寧体の頻度（割合）	です	4,497 (31.6%)	です	518 (4.1%)	です	917 (16.6%)	です	61 (3.1%)
	ます	958 (6.7%)	ます	136 (1.1%)	ます	511 (9.2%)	ます	21 (1.1%)
	その他	8,366 (61.6%)	その他	12,121 (94.9%)	その他	4,108 (74.2%)	その他	1,894 (95.9%)
総発話文数	14,221(100%)		12,775(100%)		5,536(100%)		1,976(100%)	

*1　接触場面における非母語話者の発話のみを対象としている。

*2　丁寧体数の欄の括弧内は、発話文数に対する割合である。

表６を見ると、友人との会話における「です」「ます」の総発話文数に占める割合は、母語話者と非母語話者でほとんど差がなく、ともに５％以下と低いことがわかる。（χ^2検定

の結果、5%水準で有意差なし)。一方、初対面会話を見ると、母語話者の「です」が総発話文数に占める割合は 31.6%と高く、非母語話者の約 2 倍にのぼる。(χ^2検定の結果、母語話者と非母語話者の間に 1%水準で有意差が見られた。)逆に、非母語話者は、「ます」の使用率が 9.2%と母語話者よりも高くなっている。(χ^2検定の結果、1%水準で有意差が見られた。)すなわち、非母語話者の方が「ます」を相対的に多く用いていることが明らかになった。このことは、母語話者が、「行くんですか?」というところを、非母語話者は、「行きますか?」と言いがちであるというような報告等を支持しているように思われる。

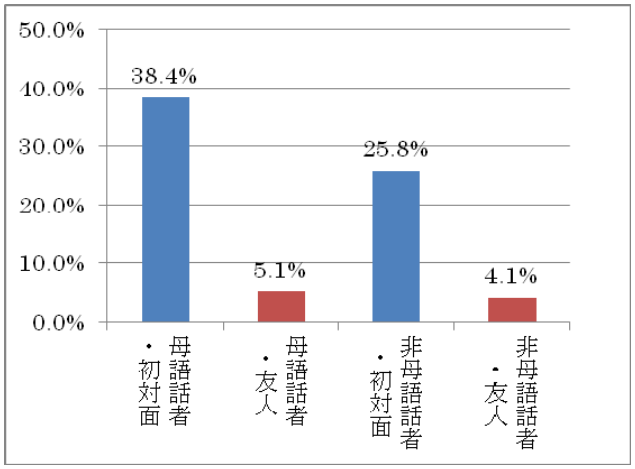
次に、「です」「ます」の頻度を合わせた「丁寧体率」について述べる。表 7 を見るとわかるように、母語話者も非母語話者も、友人との会話より、初対面会話で丁寧体を多く使っていることがわかる。友人同士の会話の丁寧体率は、母語話者、非母語話者ともに、約 5%と低いことでほぼ同様の傾向を見せる。友人同士の会話においては、χ^2検定を行った結果、母語話者と非母語話者の間に 5%水準で有意差は見られなかった。しかし、初対面会話を比較してみると、母語話者が約 40%の丁寧体率であるのに対して、非母語話者の丁寧体率は約 25%と低く、χ^2検定を行った結果、1%水準で有意差が見られた。

つまり、友人同士の会話においては、母語話者と非母語話者の丁寧体の使用に差はないが、初対面の会話においては、丁寧体の使用の違いが顕著であるということである。日本語において、初対面会話における「丁寧体」の適切な使用は、失礼のない円滑なコミュニケーションのために重要な要素の一つである。非母語話者の丁寧体率が、母語話者より有意に低いということは、語用論的に適切でない発話もありうる恐れもある。その点については、別途「定性的分析」と合わせて考察する必要がある。

表 7　話者同士の関係、場面の違う会話における母語話者と非母語話者の「丁寧体率」の比較

グループ名	母語話者・初対面	母語話者・友人	非母語話者*・初対面	非母語話者*・友人
丁寧体数	5,455	654	1,428	82
発話文数	14,221	12,775	5,536	1,976
丁寧体率	**38.4%**	**5.1%**	**25.8%**	**4.1%**

*接触場面における非母語話者の発話のみを対象としている。



5．まとめ

本稿では、『BTSJ による日本語話し言葉コーパス(トランスクリプト・音声)2011 年版』の設計の趣旨と特性を紹介するとともに、本コーパスにおける動詞と副詞の高頻度語を

BCCWJ と比較した。また、話者同士の関係、場面の違う会話における母語話者と非母語話者の発話の「異なり語数」と「丁寧体率」の違いを明らかにした。「BTSJ 話し言葉コーパス」の会話は、諸条件を統制して収集し、相互作用研究に必須である発話の重なりや沈黙などが BTSJ のルールによってきめ細かく記述され、さらに、各会話参加者の社会的属性の情報がコーパス利用者の利便を考慮し、エクセルファイルにまとめられていることが特徴である。「総合的会話分析」（宇佐美 2008）という方法論では、「BTSJ 話し言葉コーパス」のこれらの特徴を活かして、本来は、ここに示した「定量的分析」の中身を「定性的分析」によってより詳細に分析・例示しながら、考察することをもって一研究と捉えることを主旨としている。定量的、定性的双方の分析を行って初めて「総合的会話分析」と言え、その目的である「人間の相互作用のメカニズムの解明」に貢献することができると考えるからである。ただ、今回は、「BTSJ 話し言葉コーパス」の設計と特性について概要を紹介するのが主旨であった。本コーパスを用いた本格的な語用論的、対人コミュニケーション的分析については、今後、稿を改めて発表していく。

謝　辞

本研究は、科学研究費補助金基盤研究（A)「自然会話リソースバンク構築による世界的教材共有ネットワーク実現のための総合的研究」(平成 23 年度～平成 26 年度、研究代表者：宇佐美まゆみ)による補助を得ている。記して感謝したい。

文　献

宇佐美まゆみ(1995)「談話レベルから見た敬語使用: スピーチレベルシフト生起の条件と機能」『学苑』662、pp.27-42. 昭和女子大学近代文化研究所

宇佐美まゆみ・嶺田明美（1995）「対話相手に応じた話題導入の仕方とその展開パターン: 初対面二者間の会話分析より」『名古屋学院大学日本語学・日本語教育論集』2、pp.130-145. 名古屋学院大学留学生別科 (日本研究プログラム).

宇佐美まゆみ(1999)「談話の定量的分析-言語社会心理学的アプローチ-」『日本語学』18:11、pp.40-56、明治書院.

宇佐美まゆみ(2001)「『ディスコース・ポライトネス』という観点から見た敬語使用の機能 -敬語使用の新しい捉え方がポライトネスの談話理論に示唆すること-」『語学研究所論集』6、pp.1-29、東京外国語大学語学研究所

宇佐美まゆみ(2008)「相互作用と学習―ディスコース・ポライトネス理論の観点から」『講座社会言語科学 第 4 巻 教育・学習』、pp.150-181、ひつじ書房.

宇佐美まゆみ(2011)「基本的な文字化の原則(Basic Transcription System for Japanese: BTSJ)2011 年版」http://www.tufs.ac.jp/ts/personal/usamiken/btsj2011.pdf

林四郎(1975)「第二章　基本語彙はきめられるか」『新・日本語講座 1　現代日本語の単語と文字』、pp.37-54、汐文社.

関連 URL

宇佐美まゆみ研究室　http://www.tufs.ac.jp/ts/personal/usamiken/

宇佐美まゆみ監修(2011)『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』について http://www.tufs.ac.jp/ts/personal/usamiken/btsj_corpus_explanation.htm

宇佐美まゆみ（2012）「BTSJ 活用方法講習会の趣旨」
http://www.tufs.ac.jp/ts/personal/usamiken/btsj_koushuu_0_shushi.pdf

付表　『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』に収録されている会話グループとその概要

会話グループ番号と会話グループ名		会話の通し番号	データの特徴	データ数	総分数	音声付き
1	親しい同性友人同士(男女)の雑談	1-19	同性の友人同士の会話	19 会話	444 分 24 秒	
2	初対面と友人同士の女性の雑談	20-42	女性の、親しい友人同士と初対面の会話	23 会話	482 分 5 秒	
3	論文指導	43-52	教師と学生の面談の会話	10 会話	311 分	
4	女性同士の断りの電話会話	53-91	ある学生(女性)をベースに、電話で、先輩・同輩・後輩に依頼の電話をかけた会話	39 会話	78 分 35 秒	○
5	同性同士男女の依頼を含む電話会話	92-111	同性の友人同士の会話	20 会話	53 分 02 秒	
6	友人同士の女性の雑談	112-116	女性の友人同士の会話	5 会話	91 分 55 秒	
7	OPI インタビュー	117-120	OPI インタビュー形式に基づく、フランス語母語話者の縦断データ	4 会話	40 分	
8	韓国人学習者(中級)と日本人の初対面雑談	121-129	韓国人日本語学習者の接触場面データ	9 会話	249 分	
9	台湾人学習者(上級)と日本人の初対面雑談	130-141	台湾人日本語学習者の接触場面データ	12 会話	234 分 20 秒	
10	台湾人学習者(上級)と日本人の友人の雑談	142-151	台湾人日本語学習者の接触場面データ	10 会話	173 分 51 秒	○
11	初対面女性ベース雑談(接触、母語)その 1	152-160	20 代前半の日本人女性(学生)が、対同世代の日本人女性、対日本語中級話者、対日本語超級話者と 3 通りの会話を行っている	9 会話	159 分 32 秒	○
12	初対面女性ベース雑談(接触、母語)その 2	161-172	20 代前半の日本人女性(学生)が、対同世代の日本人女性、対日本語初級話者、対日本語上級話者と 3 通りの会話を行っている	12 会話	120 分 11 秒	
13	初対面男性ベース雑談(性差、年齢差)	173-190	35 歳男性が、年上(45 歳)・同等(35 歳)・年下(25 歳)の話者(男/女)と 6 通りの会話を行っている	18 会話	295 分 39 秒	○
14	初対面同性同士雑談(男、女)	191-206	20 代前半大学生・大学院生、初対面の雑談	16 会話	272 分 18 秒	○
15	友人同士女性雑談	207-209	20 代女性学生、親しい友人同士の雑談	3 会話	63 分 37 秒	

16	友人同士男女(雑談、討論)	210-233	10 代後半～20 代大学生友人同士の会話、ベース話者(男女同数)が、同性/異性の友人との雑談/討論という 4 通りの会話を行っている。	24 会話	401 分 16 秒	
17	友人同士男女間討論	234-238	20 代-30 代学生、友人同士の討論	5 会話	88 分 16 秒	
18	初対面女性討論	239-242	20 代女性、大学生・大学院生、初対面の討論	4 会話	44 分 33 秒	
19	友人同士女性誘い	243-250	20 代大学生友人同士。話者の一方が協力者である。協力者が「気軽に行うこと」を誘うように依頼した。	8 会話	172 分 53 秒	
20	初対面女性雑談(母語・接触)	251-262	日本語母語話者同士の会話と、日本語母語話者と日本語学習者の会話	12 会話	186 分 20 秒	○
21	謝罪の会話	263-294	2 人の話者が、負担度の軽い場合と重い場合の 2 つの謝罪場面についてロールプレイを行っている。	32 会話	78 分 52 秒	○
計				294 会話	4041 分 39 秒 (約 67 時間)	

データ提供者は、下記の通り。(50 音順)。
李恩美、伊集院郁子、宇佐美まゆみ、カチマレク・ミロスワバ、北見奈津子、木林理恵、金銀美、木山幸子、黄瓊芸、施信余、鄭賢児、関崎博紀、蘇玉萍、高森絵美、張鈞竹、鄭榮美、藤田朋世、松本剛次、松本紫帆、宮武かおり、林君玲

百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング

スルダノヴィッチ・イレーナ（国立国語研究所・リュブリャーナ大学）†
スホメル・ヴィット（マサリック大学言語処理センター）
小木曽智信（国立国語研究所）
キルガリフ・アダム（レクシカルコンピューティング・リーズ大学）

Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen

Irena Srdanović (National Institute for Japanese Language and Linguistics/University of Ljubljana),
Vit Suchomel (Natural Language Processing Centre, Masaryk University)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Adam Kilgarriff (Lexical Computing Ltd./Leeds University)

１．はじめに

近年、一億語を超えた大規模な現代日本語書き言葉均衡コーパスが完成し、その大きなプロジェクトの成果として新しいアノテーションツール、電子化辞書、コーパス検索ツールなどの日本学以外の様々な分野に応用できるリソースが作成されてきた。次の段階として、コーパス量を増やす必要性が明らかになり、今までのデータでは十分把握できず、抽出できなかった言語的情報を得るために超大規模なウェブコーパス構築が始まった。こうした中、様々な言語でウェブコーパス作成の重要性が認識されてきて、多言語のための TenTen と呼ばれるウェブコーパス群の構築が行われている。本論文において、まず新たに作成された JpTenTen という日本語の 100 億語の超大規模なウェブコーパスを紹介する。このコーパスは、SpiderLing（Pomikalek and Suchomel 2012）などのツールでデータをクロールし、クリーニングを行った上で、MeCab と UniDic2 (小木曽ら 2011)で形態素解析し、短単位と長単位アノテーションを付与した。コーパスは Sketch Engine というレクシカルプロファイリングツール（Kilgarriff ら 2004）に搭載した。このツールは既に 4 億語の日本語コーパス JpWaC を基にした語彙･文法プロファイリングを可能にしているが（Srdanović ら 2008）、本研究によって新たに可能になった成果は以下の通りである。

- 超大規模なコーパスを構築し、スケッチエンジンツールに載せた。その結果、今までできなかった言葉の組み合わせなどの言語情報を取り出せるようになった。
- 長単位と短単位のアノテーションを利用したことで、以前より統一された短単位のデータと、以前には存在しなかった長単位のデータが利用可能になった。
- 品詞タグだけでなく、UniDic の活用形および活用型等の英訳アノテーションを利用し、以前にはなかった活用形に関する詳細な情報を取り出せるようになった。
- 「文法関係ファイル」のデータを更に整備し、今まで取り出せなかった語と語の組み合わせおよびその振る舞いの情報が抽出できるようになった。

以上の外に、２語以上の共起抽出などの新しく開発した機能により、以前にはできなかった情報習得および表示ができるようになってきた。

本論文では、第２章においてコーパスの構築を紹介した上で、第３章においてコーパスのアノテーションおよび短単位と長単位の語彙プロファイリングのメリットについて述べる。第４章は、新しい「文法関係ファイル」によって抽出できるようになった語彙・文法情報を紹介し、第５章では、具体的な例を取り出し、百億語の日本語のコーパスからどのような言語的情報が得られるかについて述べる。

† irena.srdanovic@ff.uni-lj.si

２．TenTen コーパス群と JpTenTen コーパス構築

近年、ウェブデータを用いたコーパス構築のメリットが認識され、それに関する研究が増加してきた。最初の日本語大規模ウェブコーパス JpWaC は、Baroni and Kilgarriff (2006)、Sharoff（2006）が提案した方法を利用し、WaC 群の一つとして開発されたものである（Srdanović ら 2008)。近年「Corpus Factory」（コーパスファクトリ）（Kilgarriff 2010）というプロジェクトの枠組みで、10^{10}（百億語）の TenTen という新しいウェブコーパス群の開発が始まり、さまざまな言語のコーパスが構築された。TenTen 群の一つとして日本語超大規模コーパス JpTenTen が 2011 年に作成された（Pomikalek and Suchomel 2012)。正式な名前は「JpTenTen11」[1]である。

JpTenTen は以下の手順で構築された。

（１） 日本語の言語モデル作成。日本語のウィキペディアからデータを利用し、モデル学習を行った。約 1000 ページの日本語ウェブページをさまざまなエンコーディングで取得した。（Kilgarriff 2010）

（２） 言語コーパス作成用の SpiderLing クローラー（Pomikalek and Suchomel 2012）によって、前述したモデルを利用し、日本語のウェブページをクロールした。

（３） JusText を利用し（Pomikalek 2011）、「文にあるテクストだけ」（text in sentences only）を収集し、それ以外のテクストではないデータおよび「ボイラープレート」（boilerplate）を削除した。

（４） 「オニオン」というツールで、段落レベルの情報で重複したデータを削除した（de-duplicate）（Pomikalek 2011)。

（５） 形態素解析ツール MeCab 0.98 および電子化辞書 UniDic 2.1.0 を利用し、全体のコーパスを処理し、アノテーションを付加した(小木曽ら 2011)。その際、UniDic の品詞・活用形・活用型のマッピングを行い、英訳のタグセットを作成した。

（６） Comainu 0.60 を利用し、UniDic の長単位の処理およびアノテーションを行った。このステップは時間を要するため現時点では作業中であり、サンプルコーパスが完成しているところである。

（７） 以前作成した日本語の「文法関係ファイル」を基にして（Srdanović ら 2008、スルダノヴィッチ・仁科 2008)、UniDic の英訳タグセットと正規表現を利用し、新しい日本語の「文法関係ファイル」を作成した。

（８） データの記号化（encoding）とワードスケッチのコンパイルは、Sketch Engine（Kilgarriff 2004）が利用している Manatee というシステムで行った。

UniDic の短単位でタグされた JpTenTen は、10,321,875,665 語のデータである。15,553,207 のウェブページ、734,758 のドメインからのものである。高頻度のドメインは 28,474 のウェブページからなっており、一つのウェブページからなるドメイン数は 224,293 である。表１は、コーパスにあるトップ頻度の５つのドメインを示す。

表１　コーパスにあるトップ頻度の５ドメインおよび
ドメインごとのウェブページ割合

ドメイン	Com	jp	net	info	Other
ページ割合	50%	32%	9%	5%	4%

[1] 「Jp」は、日本語を指す 2 文字のコードである（ISO-6390-1pcode)。数年後更新するモニターコーパスとして計画されているため、「11」は、2011 年にウェブから得られたデータのことを示す。

３．UniDic 短単位と長単位アノテーションを付加した JpTenTen

日本語は単語の分かち書きがなされず多様な表記法を持つため、日本語のコーパスにとって単語情報（形態論情報）のアノテーションは重要である。特に、単語の区切り方をどうするのか、多様な表記をどのようにまとめ上げるのか、という点は大きな問題となる。

JpWaC コーパスでは、従来 ChaSen 標準の辞書である IPADIC を利用してきたが、この辞書では、単語の区切り方の揺れや、表記のまとめ上げなどは言語の研究にとって十分であるとは言えない点があった。たとえば、区切り方の面では、「株式会社」が 1 語である一方で「有限／会社」「合資／会社」は 2 語に分割されるような揺れがあった。また表記の面では、「ネギ」「ねぎ」「葱」を見出し語として一つにまとめ上げることができなかった（読みとしてはまとめられるが、そうすると「禰宜」と区別されない）。

今回、JpTenTen では、UniDic を利用することによってこうした問題に対処した。UniDic は、BCCWJ の開発にあたって整備された形態素解析辞書で、このような問題を解決することができる。UniDic は、短単位と呼ばれる厳密な規定によって単語の区切り方が定められており、揺れが少ない斉一な単位による解析が可能になっている（小椋ら 2011）。また、語彙素・語形・書字形・発音形という見出し語の階層構造を持っており、利用者が必要に応じて、見出し語のレベルを選択して利用することができる（伝ら 2007）。たとえば、表記そのものに関心があるのであれば「書字形」を、語形の差異に関心があるのであれば「語形」を、辞書見出し（lemma）のレベルでまとめ上げたいのであれば「語彙素」を利用すればよい。

UniDic では、前述の「株式会社」は規程に従って他と同様に「株式／会社」と 2 語に分割され、「ネギ」「ねぎ」「葱」には共通して語彙素「葱」・語彙素読み「ネギ」の情報が付与される。さらに、新しい JpTenTen では、BCCWJ と同様に長単位による解析も行い、短単位と長単位の両方で利用することを可能にした。長単位とは、文節を基準とした語の単位で、まず文節を区切りとし、さらに文節のうちの付属語を切り出したサイズになる。また、短単位で分割される漢語サ変動詞や一部の複合辞は 1 長単位となる。次の例は、同じ文を短単位と長単位で分割した例である。

短単位：私/は/国立/国語/研究/所/で/日本/語/を/研究/し/て/いる/
長単位：私/は/国立国語研究所/で/日本語/を/研究し/ている/

短単位が辞書の見出しとしてあらかじめリストアップされたかなり短い単位であるのに対し、長単位は実際にコーパスに出現する形に基づいて作られる比較的長い単位である。ただし、多くの事例では短単位と長単位は一致する。長単位は、長単位解析器 Comainu により、UniDic を使って行われた形態素解析結果である短単位を組み上げる形で作成される（小澤ら 2011）。

JpTenTen に利用した UniDic の品詞、活用形、活用型は英訳した上でコーパスに載せた。品詞マッピングの例を表２に示す。

表２　UniDic の品詞マッピング

品詞	品詞(英訳)	記述
代名詞	Pron	pronoun
副詞	Adv	adverb
助動詞	Aux	auxiliary_verb
助詞-係助詞	P.bind	particle(binding)
助詞-副助詞	P.adv	particle(adverbial)

４．スケッチエンジンに載せた JpTenTen

４.１　コンコーダンス

JpTenTen コーパスをスケッチエンジンに搭載することにより、ツールのウェブページからアクセスができ、標準的なコンコーダンスとしての機能が利用できる。コンコーダンスは、語彙素、語句、単語、文字および CQL 機能（Corpus Query Language、コーパス検索言語）で正規表現とデフォルト属性を基にした共起、文法的パターンなどの項目の検索方法が指定できる。ここは、UniDic の短単位と長単位で分析されている語彙素で検索ができる。図１は、コンコーダンスにあるデフォルト属性の選択肢を示している。以前は単語（word）、語彙素（lemma）、タグ（tag）での検索だけが可能だったが、現在は活用形（infl_form）、活用型（infl_type）また語彙素読み（lemma_kana）で言語的情報の検索ができるようになった。

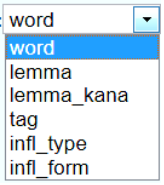


図１　コンコーダンスにあるデフォルト属性の選択肢

図２は、コンコーダンスで可能な表示の例を示している。キーワードだけのアノテーションを表示するか周りの単位のアノテーションも表示するか、またどのアノテーションタイプを表示するかを選択できる。図の例は上から順に、(1)キーワードの語彙素、(2)キーワードの語彙素と品詞、(3)キーワードの単語・語彙素・読み方・品詞・活用型・活用形、(4)キーワードとコンテクストの語彙素と品詞を表示したものである。

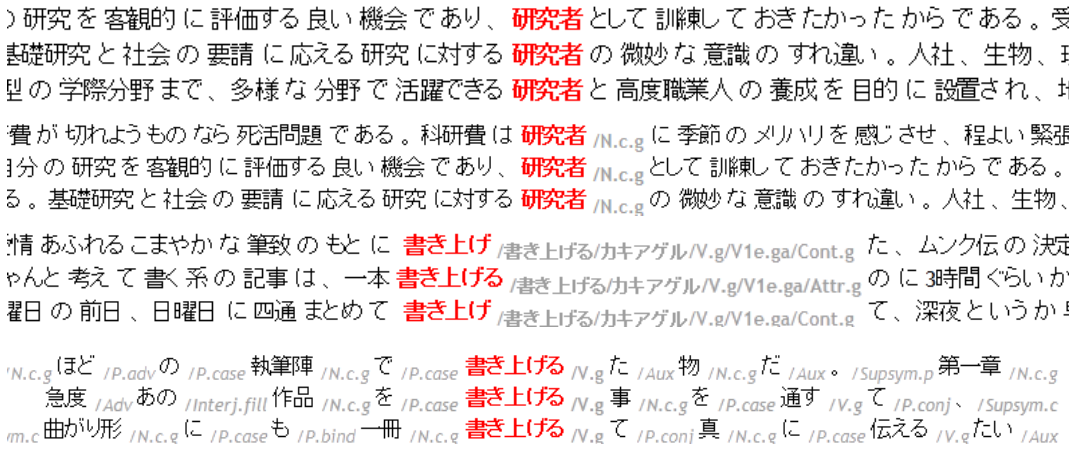


図２　コンコーダンスにある可能な表示

４.２文法関係ファイルとワードスケッチ

日本語の「文法関係ファイル」において語彙・文法的関係を決定した結果、コンコーダンスだけでなく、キーワードの語彙・文法的プロファイリング、キーワードのシソーラス、類似した語の差異と共通点などをウェブ上で１ページにまとめた言語的情報が見られる。

日本語のための「文法関係ファイル」は 2007 年に初めて作成された（Srdanović ら 2008）。ファイルの作成においては、Gahl（1998）によって提案された「corpus query syntax (コーパス検索シンタクス)」を実装し、主に品詞と正規表現を利用した。日本語の語彙・文法的規則を作成するにあたって、日本語の動詞、名詞、形容詞、副詞、接尾辞、接頭辞、助動詞

などの単位をカバーし、それぞれの品詞がどのような語と語で組み合わされ、どのようなパターンで現れるかをさまざまな言語的データから簡単に抽出・観察できるようになった。

本研究では、既存の「文法関係ファイル」を様々な面で整備・更新した。内容を以下にまとめる。[2]

(1) 第 3 章に説明した MeCab-UniDic の短単位と長単位のアノテーションを採用するため、「文法関係ファイル」に以前利用した ChaSen‐IPADIC のタグから MeCab-UniDic へのタグマッピングを行った。
(2) 品詞だけでなく、新たに活用型・活用形に基づいて正規表現で語彙・文法パターンを作成した。
(3) 以前はカバーされなかった文法関係を新しく作成した。

それぞれの改善点は、以下の「文法関係ファイル」のパターンの例、または図１と図２のワードスケッチの例に見られる。

*DUAL[3]
=modifier_Ai_cont/modifies_N+する
2:[tag="Ai.*" & word!="なく|無く" & infl_form="Cont.*"] [tag="Pref"]? 1:[tag="N.c.vs"]

語彙・文法関係は、主に 2 項（dual）関係タイプとして設定する。たとえば、以上の例は名詞‐普通名詞‐サ変可能（tag="N.c.vs"）を検索すると、それを修飾する連用形の活用形（infl_form="Cont.*"）にある形容詞が現れる（tag="Ai.*"）。また形容詞をキーワードにして検索すると、それに呼応する名詞‐普通名詞‐サ変可能の例が現れる。このパターン（いわゆる「文法関係」）に「modifier_Ai_cont/modifies_N+する」という名前を付けた。

以上に利用した省略の説明は以下のとおりである。

tag="N.c.vs"- noun.common.verb_suru の品詞省略
infl_form="Cont.*"Continuous_ren'yo の活用形省略
tag="Ai.*"　adjective –i の品詞省略

このパターンの内容には、前述した(1)のタグマッピングの結果の例、(2)の活用形の利用、(3)新しく作成した文法関係の例を含んでいる。

図３は以上のパターンを利用したワードスケッチの例を示す。たとえば、「結婚」というサ変名詞がどのような連用形の形容詞と結びつくかを検索した結果である（例えば、めでたく結婚する、早く結婚する、仕方無く結婚するなど）（図３の１欄目）。また、「素晴らしい」という形容詞が連用形の活用形の場合、どのサ変名詞と結びつくかを示した結果である（例えば、素晴らしく洗練する、素晴らしく感動する、素晴らしく調和する、素晴らしく充実するなど）（図３の２欄目）。

検索した結果は短単位であり、「する」は別の単位と扱われているため、結果にはサ変名詞だけが表示される。一方、図３の３欄目は、UniDic の長単位で検索し、「結婚する」を一つの単位として扱った例である。JpTenTen サンプルコーパスから取り出した結果なので、高頻度の組み合わせの「めでたく結婚する、早く結婚する」だけが表示されているが、長単位のキーワードで検索できるメリットがある。全体のコーパスが利用可能になると、抽出できる結果が増加する。

図３のそれぞれのの欄に表示されている数字は、1 列目がコーパスの中の共起頻度を示し、2 列目がその共起の統計的な重要度（salience）を示している。[4]

---

[2] 以前のデータの評価および問題点について Srdanović ら（2011）、Kilgarriff ら（2010）を参考されたい。

[3] 語彙・文法関係の 2 項（dual）などの設定について詳細は Srdanović（2008）らを参考されたい。

[4] 1 列目の数字をクリックすると、コーパス中にあるキーワードとそれぞれの共起語が含まれる例文がコン

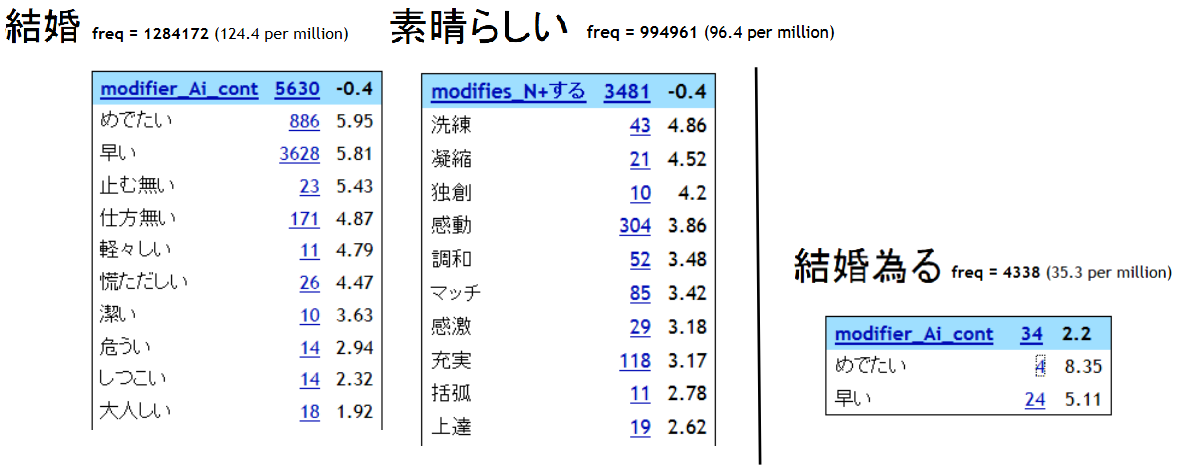


図３「結婚＋形容詞連用形」および「素晴らしい＋名詞普通名詞サ変可能」のワードスケッチの例（JpTenTen,UniDic2 短単位）。また、「結婚する＋形容詞連用形」のワードスケッチの例（JpTenTen のサンプル,UniDic2 長単位）

以上の例は、前述した長単位による情報抽出のメリットを示すもので、以前には抽出できなかった「サ変名詞＋する」の組み合わせパターン、および活用形のタグを用いた抽出の例である。

５．JpTenTen を用いた語彙・文法情報のプロファイリング

本章では、百億語の JpTenTen コーパスから取り出せる語彙・文法情報プロファイリングのいくつかの例を紹介する。

５．１　まとめた形のキーワードのプロファイリング

図４は、ワードスケッチの「女性」というキーワードの様々なパターンの例で、新しく抽出できるようになった文法関係のバラエティーを示す。パターンは、「女性＋助詞」「女性＋名詞」「女性＋の＋名詞」「名詞＋の＋女性」「女性＋に＋動詞」「女性＋が＋動詞」などである。キーワードがどのようなシンタクスの中でよく利用されているか、どの助詞と結びつくか、どの形容詞・形状詞に修飾されるかなどの細かい語彙の振る舞いが観察できる。

スペースの制限のため、それぞれのパターンの結果を省略し、一番重要度が高い 3・4 語を示した。

コーダンスの中で表示される。文法関係用語のリンク（modifies_N など）をクリックすると、その文法関係が正規表現と品詞を利用して、どのように決定されているかを確かめることができる。

スルダノヴィッチ・仁科（2008）に示したように、このような情報により、キーワードの意味を把握できるため、キーワードの意味記述などのために辞書学によく応用される。それ以外にも、言語学、言語教育などの分野に幅広く利用できる。

図４JpTenTen から取り出せる「女性」という名詞のさまざまなパターン
（パターン結果は省略した）

５．２　短単位および長単位で見る語彙プロファイリングのメリット

3 章に既に紹介したように、UniDic にはさまざまなメリットがある。本章では短単位および長単位で取り出せる言語的情報の例を上げるが、特に強調するポイントは、以下のとおりである。

- 短単位により、言語単位がどのような部分から構成されているのか、調べられること。特に派生語と関連して、接尾辞、接頭辞、非自立可能な品詞のそれぞれの特徴、振る舞い傾向を細かく調べることができる。例えば、どの形容詞が「～らしい、～こい、～臭い」などの接尾辞とよく結びつく傾向があるか、また「研究」という名詞の後ろにどの接尾辞がよく付くかといった情報が大規模なデータにより把握できる。

- 長単位で、複数の単位からできている言語単位の振る舞いを検討することができる。以前は抽出できなかったサ変動詞、複合名詞、複合動詞、のような複合語が語彙素になり、これらの語彙を単位とする組み合わせとして抽出できるようになった。これにより、長単位をキーワードとして調べることができるだけでなく、他の語をキーワードとして調べたときに長単位にもとづく情報が得られるという二つ面でメリットがある。

図５は、長単位でタグづけされたサンプルコーパスから取り出した「研究者」という名詞および「興味深い」という形容詞の例を示す。

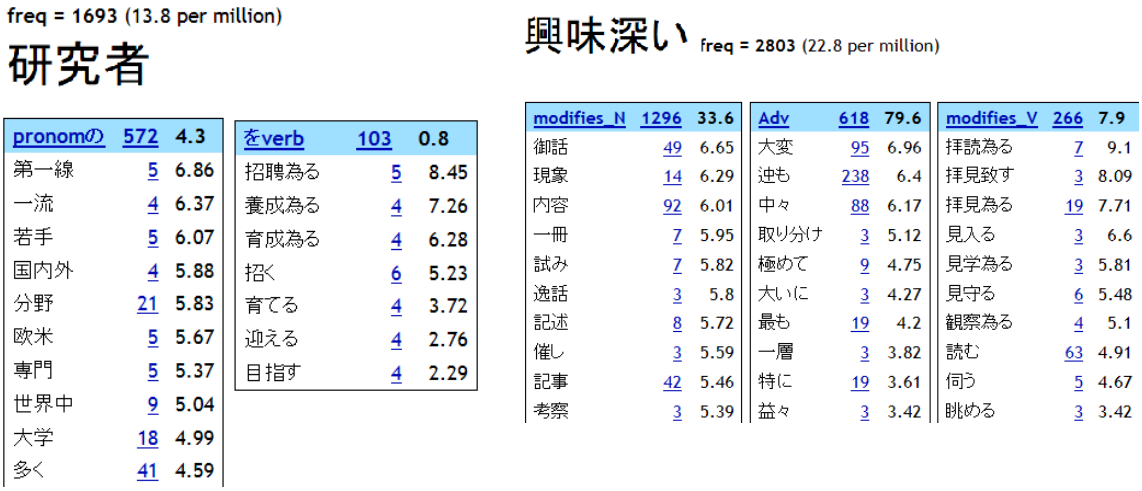


図５JpTenTen（長単位、サンプル）から取り出した「研究者」および「興味深い」のプロファイリング

これらのキーワードは短単位では取り出せなかった語であり、このような複合語のプロファイリングができるのは非常に重要である。取り出した結果にも複合語のデータが多い。例えば、「第一線の研究者、国内外研究者、世界中の研究者、研究者を招聘する」また「興味深いお話、興味深く拝見致す、興味深く拝読する」などである。

５．３　複数単位の抽出

新しく追加された機能で、それぞれのパターンにある単位からマルチワードスケッチページ（Multiword sketches）に飛ぶことができるようになった。図６はこのようなページ結果の例を示す。例えば、「最近の研究」から「最近の研究成果」、「新たな研究」から「新たな研究領域」、「とても興味深い」から「とても興味深く読む」などの複合語が並んだ例が見られる。

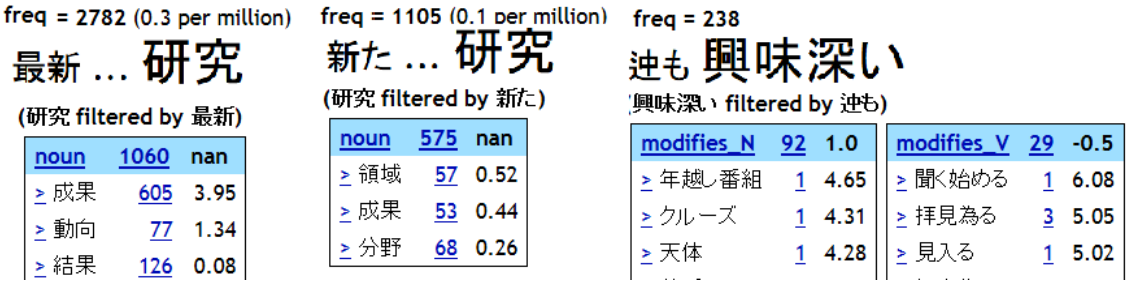


図６マルチワードスケッチの例（最近の研究～、新たな研究～、とても興味深い～）

５．４　語彙・品詞・活用形・活用型・パターンの頻度リスト

スケッチエンジンツールでは、さまざまな語彙・品詞・活用形・活用型の頻度リストが取り出せる。図７にその例を示す。1 欄目は、UniDic 短単位で解析された 100 億語のコーパスに現れる品詞の高頻度順リストである（名詞-普通名詞-一般、助詞 - 格助詞、助動詞、名詞-普通名詞-サ変可能など）。2 欄目は、もっとも高頻度の活用形（連用形-一般、終止形-一般、連体形-一般、連用形-促音便など）、3 欄目は、もっとも高頻度の活用型のリストである（助動詞-ダ、五段-ラ行、助動詞-タ、サ行変格など）。

図７の 4 欄目は、コーパスに現れる「助詞 - 接続助詞の「て」＋動詞 - 非自立可能」というパターンの頻度リストである[5]。　頻度の高いほうから、「ている、て来る、てしまう、て行く、て見る、てくれる」の順番で日本語の（テ形に接続する）補助動詞が現れる。Martin（2004、512 ページ）は、1964 年の国立国語研究所の「現代雑誌九十種の用語用字」のデータを基にして、日本語の主な補助動詞を相対頻度で並べて表に示している。現れている高頻度の補助動詞は、ほとんど図７の 4 欄目と統一している。並んだ順番もほとんど類似しているが、微妙な違いが見られる。例えば、「てしまう」は Martin（2004）の表では「行く、くれる、くださる」よりやや低い頻度で、JpTenTen のデータでは「しまう」のほうがやや頻度が高くなっている。

特定の単語、単語のグループ、一つの品詞の語彙、一つパターンなどを対象にして、超大規模コーパスから活用形、頻度などのデータを取り出すことで、今後の教育シラバス作成などに応用できる豊富な情報が得られるといえる。

tag	Freq
N.c.g	1588826682
P.case	1359247326
Aux	863345234
N.c.vs	554425638
V.bnd	532961623
V.g	506815227
Supsym.p	447290326
N.num	414314084
Supsym.c	376814114
P.conj	372440519
P.bind	355183028
Supsym.g	280555426
N.c.adv	215750266
Suff.n.g	206247617
Sym.ch	204585993
Adv	173481056
Supsym.bo	171503010
P.adv	158837049
Supsym.bc	153531627
N.c.count	128986663
P.fin	119879442

infl_form	Freq
Cont.g	607632121
Concl.g	485365140
Attr.g	469486993
Cont.t	158394931
Irr.g	116420893
Cont.ni	48221528
Cond.g	36496602
Cont.i	31451799
Vol_tent	29269791
Irr.sa	22309244
Cont.n	14146926
Concl.n	13031503
stem.g	12867278
Imp	12830635
Cont.int	6790980
Attr.n	1550658
Irr.se	1347788
Cond.int	1187705
Real.g	683018
Attr.abbr	642454
Cont.u	640659

infl_type	Freq
Aux.da	260552830
V5.ra	233862083
Aux.ta	214041705
sa_irr	196667604
Adj	167134814
V5.wa_a	131979751
V1i.a	110162200
Aux.masu	96666486
Aux.desu	83078022
V5.ka	73080756
Aux.nai	54693529
Aux.reru	52457777
V5.sa	44028481
V1e.ta	37135919
V1e.ra	35918383
V1e.ka	29983141
V5.ma	26376522
Aux.nu	23359675
V1i.ka	22916985
V1e.ma	22535426
V1i.ma	21574027

lemma	Freq
て 居る	5259778
て 来る	1107200
て 仕舞う	716055
て 行く	551444
て 見る	509031
て 呉れる	506095
て 下さる	331408
て 貰う	251620
て 頂く	193982
て 置く	183006
て 有る	115184
て 遣る	98028
て 上げる	63500
て 参る	16593
て 見せる	11584
て 為る	10239
て いらっしゃる	9846
て 回る	8578
て 出来る	7959
て 成る	7727

図７JpTenTen における品詞・活用形・活用型・パターンの頻度リスト（UniDic 短単位）

６．まとめ

本論文では、新規の超大規模な日本語のウェブコーパス JpTenTen の構築とそのアノテーションを紹介した上で、百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリングの実例を紹介した。日本語の様々な語彙・文法情報を UniDic の短単位と長単位の品詞・活用形・活用型に適用することで、以前できなかった語と語の振る舞いの情報を抽出できるようになってきた。この成果は日本語学、対照言語学、日本語辞書学、日本人学習者用英語辞書学、日本語教育、日本語言語処理、心理学などの研究分野に活用できると期待される。

[5] 利用した検索パターンは[tag="P.conj"& word="て"] [tag="V.bnd"]

謝　辞

本研究は、博報財団第 7 回「日本語海外研究者招聘事業」による研究「日本語教育における語の共起関係」（平成 24～25 年度、受入機関：国立国語研究所、招聘研究員：スルダノヴィッチ・イレーナ）およびチェコ教育科学所によるプロジェクト「LINDAT-Clarin LM2010013」（研究員：スコメル・ヴィット）の補助を得ています。

文　献

スルダノヴィッチ・イレーナ, 仁科喜久子（2008）「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 国書刊行会, pp. 59–80.

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.

小椋秀樹・小磯花絵・冨士池優美・宮内左夜香・小西光・原裕（2011）国立国語研究所内部報告書『『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）』

小澤俊介, 内元清貴, 伝康晴（2011）「BCCWJ に基づく中・長単位解析ツール」, 特定領域「日本語コーパス」平成 22 年度公開ワークショップ予稿集, pp. 331-338.

小木曽智信・伝康晴（2011）「UniDic2.0:言語資源としての電子化辞書」特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集, pp. 411-418.

Baroni, Marko and Kilgarriff, Adam (2006) Large linguistically-processed Web corpora for multiple languages, In Proceedings EACL Trento, Italy

Gahl, Susanne (1998) Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus, ms., ICSI-Berkeley

Kilgarriff, Adam, Rychly, Pavel, Smrž, Pavel and Tugwell, David (2004) The Sketch Engine. Proceedings of EURALEX. France: Université de Bretagne. pp. 105-116.

Kilgarriff, Adam, Kovář, Vojtěch., Krek, Simon, Srdanović, Irena, Tiberius, Carole (2010) A Quantitative Evaluation of Word Sketches. Proceedings of the XIV Euralex International Congress. Leeuwarden:Fryske Academy. pp. 7

Kilgarriff, Adam, Reddy, Siva, Pomikálek, Jan and Pvs, Avinesh (2010) A corpus factory for many languages. In proceedings of LREC, Malta

Martin, Samuel E. (2004) A reference grammar of Japanese. University of Hawai'i Press, Honolulu

Pomikálek, Jan (2011) Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis, Masaryk University, Brno

Pomikálek, Jan, Suchomel, Vít (2012) Efficient Web Crawling for Large Text Corpora. ACL SIGWAC Web as Corpus (at conference WWW)

Sharoff, Serge (2006) Open-source corpora: using the net to fish for linguistic data, International Journal of Corpus Linguistics, 11 (4), pp. 435–462.

Srdanović, Irena, Erjavec Tomaž and Kilgarriff, Adam (2008) A web corpus and word-sketches for Japanese. Shizen gengo shori (Journal of Natural Language Processing) 15/2. pp. 137-159.

Srdanović, Irena, Ida, Naomi, Shigemori Bučar, Chikako, Kilgarriff, Adam, Kovář, Vojtěch (2011) Japanese Word Sketches: Advantages and Problems. Acta Linguistica Asiatica, 1 (2), pp.63-82.

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/

スケッチエンジンツール Sketch Engine　http://www.sketchengine.co.uk/

クローラ SpiderLing　http://nlp.fi.muni.cz/trac/spiderling

Comainu に関する参考文献 https://maro.ninjal.ac.jp/Comainu/related_paper/

形態素解析辞書 UniDic http://download.unidic.org/

MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.googlecode.com

口頭発表（3）

3月1日（金）　10:00～12:00

中古和文における個人文体とジャンル文体

小林 雄一郎（日本学術振興会）†
小木曽 智信（国立国語研究所言語資源研究系）

Styles and Genres in Early Middle Japanese

Yuichiro Kobayashi (Japan Society for the Promotion of Science)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)

1．はじめに

現在，国立国語研究所では，日本語歴史コーパスが構築中である (近藤 2012)。そして，2012 年 12 月には，平安時代の仮名文学作品 10 作品の短単位解析済みデータが先行公開された。また，言語資源の整備にともない，中古和文の語彙や文体に関する研究も進められている（須永 2011，小木曽 2012）。

本研究の目的は，中古和文コーパスを分析対象とし，個人文体とジャンル文体の関係を明らかにすることである。安本 (1982) は，現代日本の作家の文章における 15 の文体項目を対象に，因子分析を用いて，個人文体とジャンル文体の類型化を行っている。しかしながら，中古和文を対象とする場合，個人文体とジャンル文体の区別はそれほど簡単にはできない。なぜならば，歴史的な資料は現代語の資料と比べて圧倒的に数が限られているからである。その結果，あるテクストと別のテクストの間に見られる言語的差異が，書き手の差によるものなのか，ジャンルの差によるものなのか，はたまた年代の差によるものなのか，を見分けることが難しくなる。そこで本研究では，紫式部の『源氏物語』と『紫式部日記』，そして『更級日記』における助詞・助動詞の使用傾向を調査し，多変量解析などの統計的手法を用いて，書き手による文体差とジャンルによる文体差の関係について検討する。

2．中古文学の計量文体研究

中古文学の文体研究の多くは，この時代を代表する文学作品である『源氏物語』を対象にしている。また，計量文献学の分野においても，『源氏物語』の作者の推定に関わる研究がなされてきた。たとえば，安本 (1958) は，『源氏物語』を宇治十帖 10 巻とそれ以外の 44 巻に分けて統計的検定を行った結果，両者の作者が同一人物であるとは言い難いと結論付けている。これに対して，新井 (1997) は，五十音図の頭子音行列と母音列別の頻度データに基づいて，宇治十帖の作者は他の諸巻の作者と別人であるとは考えられないとする。同様に，土山・村上 (2012) も，名詞，動詞，形容詞，形容動詞，副詞，助詞のそれぞれを変数とする主成分分析とランダムフォレストを行い，宇治十帖他作者説を退けている。『源氏物語』の成立論に関して，村上・今西 (1999) は，高頻度の助動詞を変数とする数量化 III 類を行い，(1) 第 1 部の紫の上系物語，(2) 第 2 部全てと第 3 部の宇治三帖，(3) 第 1 部の玉鬘系物語，(4) 第 3 部の宇治十帖の順で執筆されたという仮説を提唱している。さらに，他の文学作品との比較に関して，土山・村上 (2011) は，名詞，動詞，形容詞，形容動詞，副詞，助詞，助動詞のそれぞれを変数とする主成分分析とクラスター分析を行い，『源氏物語』の使用語彙と『宇津保物語』の使用語彙の間には顕著な差が見られると報告している。

また，物語文学と日記文学を計量的に比較した研究として，坂東 (1990) が挙げられる。この論文では，『枕草子』と『紫式部日記』における名詞率，MVR（動詞数に対する形容詞，形容動詞，副詞，

† kobayashi0721@gmail.com

連体詞の総和数の割合），形容詞，色彩語についての比較が行われている。ただ，その目的は「それぞれの作品の個別的文体」を明らかにすることであり，個人文体とジャンル文体の関係に光を当てるものではない。

3. 調査方法

3.1 資料

本研究で調査対象とする資料は，新編日本古典文学全集の『源氏物語』と『紫式部日記』である。『源氏物語』に関しては，第1部の「桐壺」と「若紫」（いずれも紫の上系物語）と第3部の「橋姫」と「夢浮橋」（宇治十帖の最初の巻と最後の巻）を対象とする。

これらの紫式部による作品に加えて，『更級日記』も調査対象に含める（このデータは，西端ほか 1996 に基づいている）。菅原孝標女による『更級日記』を含めたのは，個人文体とジャンル文体の関係，言い換えれば，書き手による言語的特徴の違いとジャンルによる言語的特徴の違いの関係を明らかにするために，紫式部以外の手によるテクストが必要であるからである。なお，菅原孝標女は『源氏物語』を愛読していたとされ，『更級日記』の文体も『源氏物語』の強い影響を受けていると言われている（上野 1991，上野 1994）。

また，これらの資料に対して自動形態素解析を行い，解析誤りを手作業で修正した。データに付与されている単語情報は，形態素解析辞書中古和文 UniDic（小木曽ほか 2010）で採用されている短単位に基づくものである。

3.2 変数

本研究では，各テクストにおける助詞と助動詞の語彙素の頻度を変数とする。これらの変数を選んだ理由は，日本語が膠着語であり，助詞や助動詞が表現の論理や情緒を表すにあたって重要な働きを持っているからである（此島 1971）。なお，個々のテクストの総語数が異なるため，分析にあたっては，必要に応じて，観測頻度を出現率に変換した相対頻度を用いる。

3.3 手法

各テクストにおける助詞と助動詞の頻度を分析するにあたって，最初にカテゴリー別の頻度をバープロットで視覚化する。次に，個々の助詞や助動詞の観測頻度に基づく相関係数を算出し，テクストの類似度を求める。さらに，テクストと変数の関係を把握するために，多重因子分析とクラスター分析による次元縮約を行う。そして最後に，対数尤度比を用いて，個々のテクストに特徴的な助詞と助動詞を抽出する。

4. 結果と考察

4.1 変数の分布

図1は，中古和文 UniDic による形態素解析の結果に基づき，格助詞，係助詞，終助詞，副助詞，接続助詞，助動詞の頻度の相対頻度を視覚化したものである。この図を見ると，他のテクストと比べて，『更級日記』における格助詞の頻度が高い。そして，物語文学と比べて，日記文学における助動詞の頻度が低い。これらについては，後段で詳しく見る。

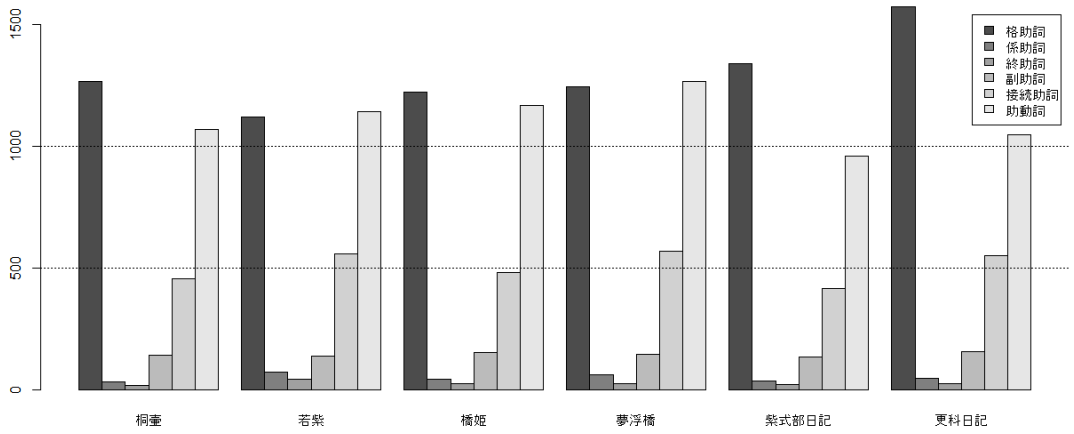


図1　各テクストにおける助詞・助動詞の相対頻度

4.2　テクストの相関関係

図2は，個々の助詞と助動詞の観測頻度に基づき，各テクスト間の相関係数（Pearsonの積率相関係数）を求めた結果である。これを見ると，最も高い値は「桐壺」と「橋姫」の0.988，最も低い値でも「夢浮橋」と『紫式部日記』の0.893であり，全体的に非常に高い値となっている。本研究の調査対象がいずれも11世紀のテクストであることから，大まかな助詞と助動詞の使い方は極めて類似したものになっていると思われる。

4.3　テクストと変数のクラスタリング

前節までの分析手法では，助詞と助動詞の使用傾向に関して，各テクスト間における顕著な差異は認められなかった。しかしながら，これまでの多くの研究において，『源氏物語』と『紫式部日記』，あるいは『源氏物語』における宇治十帖と他の巻との文体差が指摘されてきた（上野 1990, 野村 1970a, 野村 1970b, 安本 1958）。本節では，多重因子分析とクラスター分析を用いて，高次元のデータを低次元に圧縮し，その結果を直感的に解釈しやすい形式での視覚化を試みる。

以下は，個々のテクストをケースとし，助詞と助動詞の頻度を変数とした多重因子分析の結果である。多重因子分析は（因子分析ではなく）主成分分析の一種であり，変数をグループ（群）に分けて指定することができる (Wong *et al.* 2002, Pagès 2004)。この分析では，格助詞，係助詞，終助詞，副助詞，接続助詞，助動詞という6つのカテゴリーを変数のグループとして指定した。まず，図3は大局的ケース図であり，partial pointsとの結線を表示したものである。これを見ると，第2象限に『紫式部日記』と『更級日記』という日記文学がプロットされている。次に，図4は大局的負荷図であり，格助詞の多くが左上（第2象限）を向いていることは注目に値する。また，図5は群表示であり，終助詞以外の5つのカテゴリーが第1主成分に同等に寄与しており，第2主成分には助動詞が最も寄与している。そして，図6が群の大局への寄与であり，接続助詞と係助詞の第1主成分が大局分析の第1主成分に高く相関し，それ以外の4つのカテゴリーの第1主成分が大局分析の第2主成分に高く相関していることが分かる。

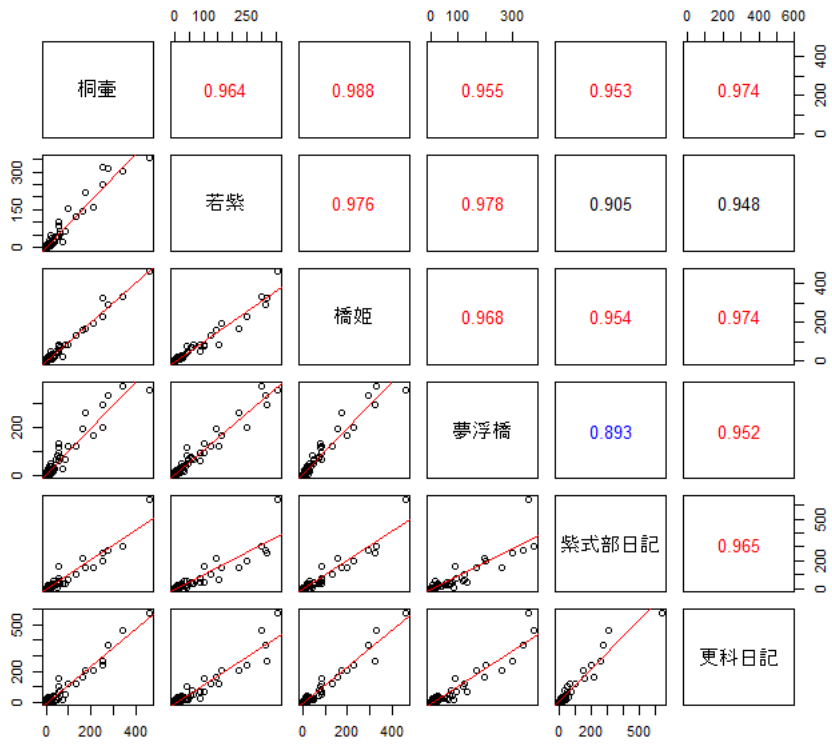


図2　各テクストの相関関係

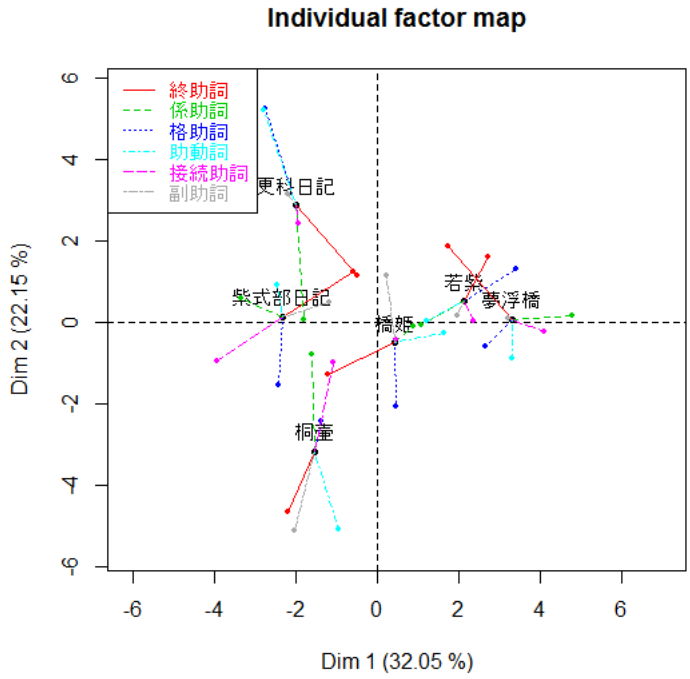


図3　大局的ケース図

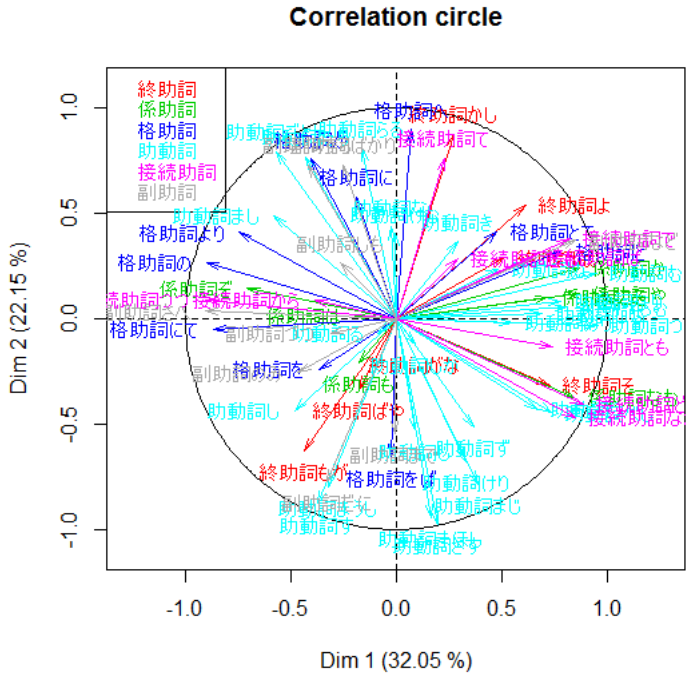


図4　大局的負荷図

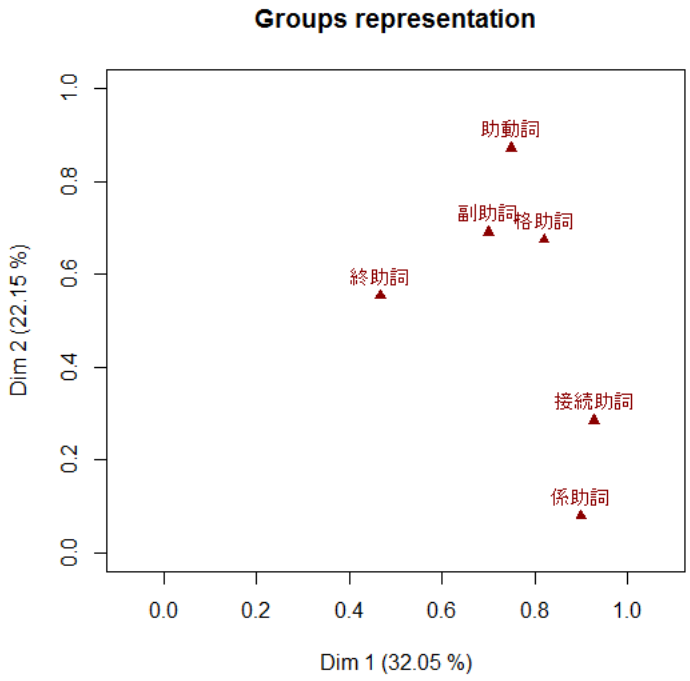


図5　群表示

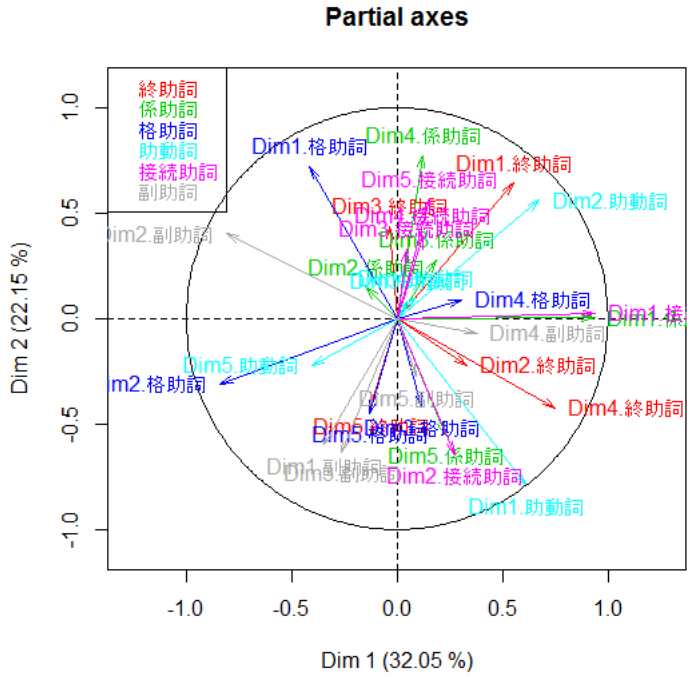


図6　群の大局への寄与

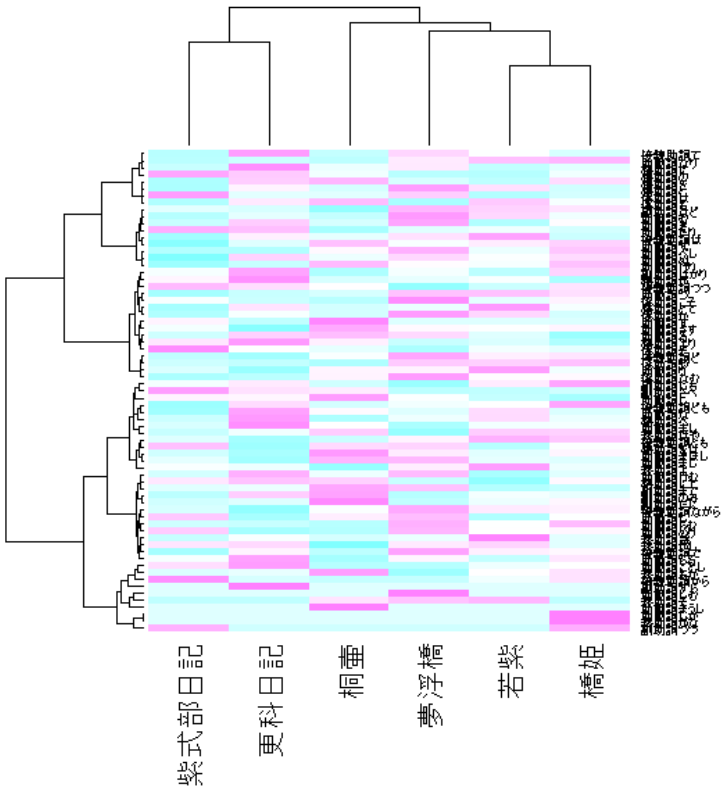


図7　ケースと変数のクラスター分析とヒートマップ

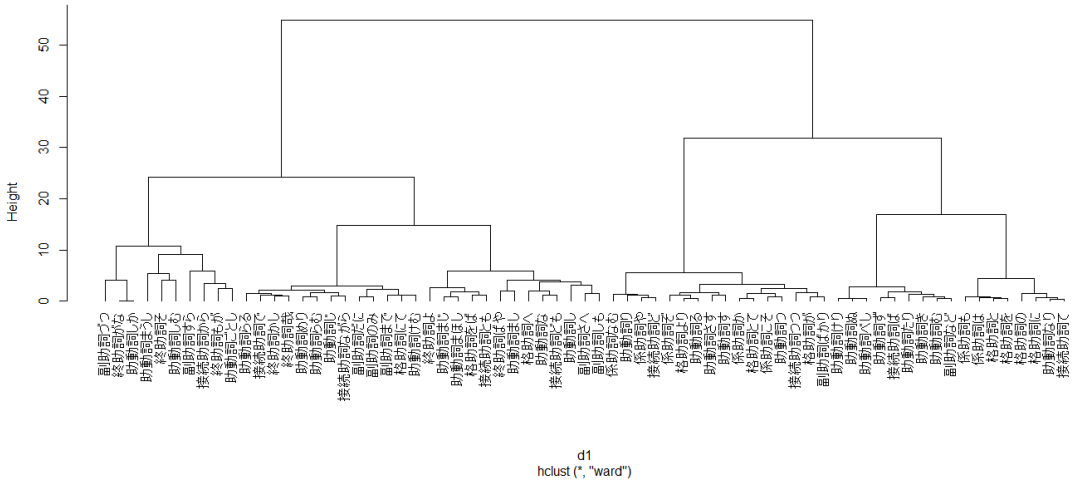


図8　変数のクラスター分析（拡大）

図7は，ケース（テクスト）と変数（助詞・助動詞）の双方に対して階層型クラスター分析を行い，ヒートマップとともに表示した結果である。分析にあたって，ケース間の距離の計算には，値が小さく差が少ないデータ同士に対しても非常に感度が高いとされているキャンベラ距離 (Gorden 1999) を用いた。また，クラスター間の距離の計算には，クラスターの各値からその質量中心までの距離を最小化するため，他の距離関数に比べて分類感度が高いとされているウォード法 (Anderberg 1984) を用いた。[1] そして，ヒートマップとは，クラスター分析に使われている元データ（ここでは，助詞・助動詞の頻度行列）に含まれる値の大小を色で表す視覚化手法である (Chaussabel 2004)。

図7におけるケースクラスタリングの結果を見ると，左側に日記文学（『紫式部日記』，『更級日記』），右側に物語文学（「桐壺」，「夢浮橋」，「若紫」，「橋姫」）がクラスターを形成している。また，『源氏物語』における紫の上系物語（「桐壺」，「若紫」）と宇治十帖（「橋姫」，「夢浮橋」）の差異は認められない。この結果をまとめると，少なくとも本研究で調査対象としたテクストにおける助詞と助動詞の頻度を変数とした分析では，書き手による文体差（個人文体）よりもジャンルによる文体差（ジャンル文体）の方が大きいことを示している。上野 (1990) は，「『源氏物語』と『紫式部日記』の文章が類似し，同一の傾向を潜在的に共有しているにせよ」，「日記と物語とは，やはり別種の作品」であるとする。本研究の結果は，助詞と助動詞の使用傾向から，この説を計量的に裏付けるものと言えるだろう。

なお，図8は，図7における変数クラスタリング結果を拡大し，右に90度回転させて表示したものである。図8を見ると，大きく2つのクラスターが形成されており，左側のクラスターには副助詞と終助詞が，右側のクラスターには格助詞と係助詞が多く含まれている。

4.4　各テクストに特徴的な変数

本節では，対数尤度比 (log-likelihood ratio, LLR) (Dunning 1993) を用いて，各テクストに特徴的な変数を抽出する。この抽出指標は，『古典対照語い表』を用いた古典作品別の特徴語抽出 (宮島・近藤 2011)

[1] ウォード法にはユークリッド距離を用いるのが基本であるが (Romesburg 1973)，計量文献学の分野ではキャンベラ距離とウォード法の組み合わせを用いることもある (石田 2008, 金 2009)。

にも利用されているものである。なお，あるテクストに特徴的な変数を抽出するにあたっては，それ以外の 5 つのテクストを比較対象とする。

表 1 は，対数尤度比を用いて，各テクストに特徴的な助詞，助動詞を抽出した結果である。この表を見ると，「桐壺」を最も特徴付ける助動詞として，敬語の「す」と「さす」が抽出されていることである。これらの助動詞は，他のテクストと比べて，「桐壺」に高い頻度で生起している（図 9）。

表 1　各テクストに特徴的な助詞・助動詞

桐壷		若紫		橋姫		夢浮橋		紫式部日記		更科日記	
語	LLR	語	LLR	語	LLR	語	LLR	語	LLR	語	LLR
す	30.183	り	42.378	なむ	25.314	なむ	27.367	の	134.304	に	42.650
さす	9.548	ば	37.110	む	12.629	む	14.841	たり	44.588	き	16.289
けり	9.104	む	16.905	き	9.274	き	10.283	は	36.804	まし	11.119
なむ	6.522	なり	9.333	ど	8.368	と	8.466	ぞ	34.139	けむ	10.481
まで	6.101	と	9.208	と	6.415	ど	8.232			て	9.714
のみ	6.059	なむ	9.039	こそ	4.974	こそ	5.404			ばかり	6.732
を	5.496	哉	7.732	べし	4.107	しむ	5.279			が	6.632
だに	5.377	とて	7.493	らむ	3.381	か	4.218			らる	6.613
まうし	4.846	つ	7.012	か	3.224	べし	3.447			たり	6.036
		や	6.254							へ	5.712
		ど	5.368							より	5.435

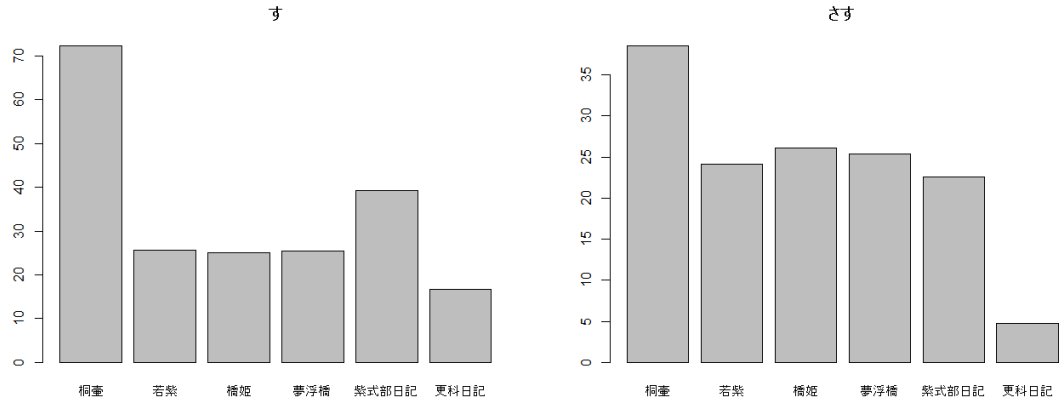


図 9　助動詞「す」，「さす」の分布

また，宇治十帖の「橋姫」と「夢浮橋」に特徴的な助詞と助動詞が極めて類似している。そして，『更級日記』から「に」，「が」，「へ」，「より」という格助詞が抽出されている。『更級日記』に格助詞が多く生起することは前述のとおりである（図 1）。図 10 は，格助詞「に」，「が」，「へ」，「より」の頻度分布を視覚化したものである。

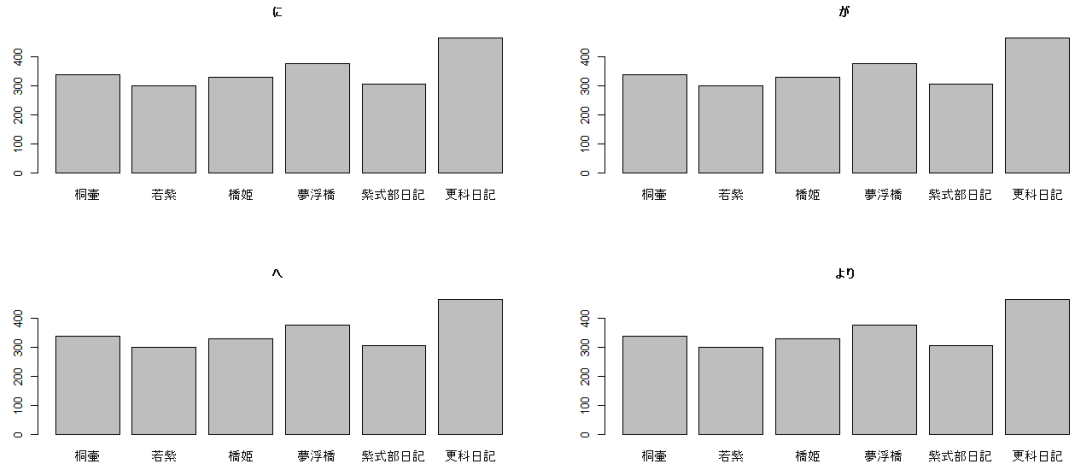


図 10　格助詞「に」,「が」,「へ」,「より」の分布

5. おわりに

本研究では，紫式部の『源氏物語』と『紫式部日記』，そして『更級日記』における助詞・助動詞の使用傾向を調査し，多変量解析などの統計的手法を用いて，書き手による文体差（個人文体）とジャンルによる文体差（ジャンル文体）の関係について検討してきた。その結果，少なくとも今回分析したデータにおいては，個人文体よりもジャンル文体の方が大きいことが明らかにされた。

今後の課題としては，まず，同時代の他のテクストや他の言語項目の分析を積み重ねていかなければならない。また，助詞と助動詞を変数とする場合でも，「なら-じ」(2 語連結）や「なら-ぬ-を」(3 語連結）のような「助詞・助動詞相互の連結関係」(宇都宮 1966) を扱うことも考えられる。さらに，テクスト全体を 1 つのケースとするだけでなく，「会話」,「歌」,「手紙」,「地の文」といった本文種別情報 (小木曽 2012) を活用し，より詳細な文体分析を行う必要がある。

文　献

新井皓士 (1997).「源氏物語・宇治十帖の作者問題：一つの計量言語学的アプローチ」『一橋論叢』117(3), pp. 397-413.

石田基広 (2008).『R によるテキストマイニング入門』森北出版.

上野英二 (1990).「紫式部における日記と物語」『成城國文學論集』20, pp. 11-50.

上野英二 (1991).「更級日記と文学史」『成城國文學論集』21, pp. 1-36.

上野英二 (1994).「菅原孝標女と源氏物語」『成城國文學論集』22, pp. 1-27.

宇都宮睦男 (1966).「紫式部日記の文体―助動詞・助詞の連結から見た」『国語教育研究』11, pp.65-71.

小木曽智信 (2012).「中古和文における語彙の文体差」『NINJAL「通時コーパス」プロジェクト・Oxford VSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集』国立国語研究所, pp.41-50.

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), pp. 1-8.

金明哲 (2009).『テキストデータの統計科学入門』岩波書店.

此島正年 (1971).「源氏物語の助詞」山岸徳平・岡一男（編）『源氏物語講座 第 7 巻 表現・文体・

語法』 有精堂出版, pp. 266-293.
近藤泰弘 (2012). 「日本語通時コーパスの設計について」 『国語研プロジェクトレビュー』3(2), pp. 84-92.
須永哲矢 (2011).「コロケーション強度を用いた中古語の語認定」『国立国語研究所論集』2, pp. 91-106.
土山玄・村上征勝 (2011). 「源氏物語と宇津保物語における語の使用傾向について」 『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. 125-132.
土山玄・村上征勝 (2012). 「語の使用頻度の計量分析による宇治十帖他作者説の検討」 『情報処理学会研究報告』2012-CH-94(5), pp. 1-8.
西端幸雄・木村雅則・志甫由紀恵 (1996). 『平安日記文学総合語彙索引：土佐日記・蜻蛉日記・和泉式部日記・紫式部日記・更級日記』 勉誠社.
野村精一 (1970a). 「宇治十帖の言語と文体」 『源氏物語文体論序説』 有精堂出版, pp. 228-262.
野村精一 (1970b). 「紫式部の文体—紫式部日記について」 『源氏物語文体論序説』 有精堂出版, pp. 263-305.
坂東久美 (1990). 「『枕草子』と『紫式部日記』における文体の比較研究」 『徳島大学国語国文學』 3, pp. 64-69.
宮島達夫・近藤明日子 (2011). 「古典作品の特徴語」 『計量国語学』28(3), pp. 94-105.
村上征勝・今西祐一郎 (1999). 「源氏物語の助動詞の計量分析」 『情報処理学会論文誌』40(3), pp. 774-782.
安本美典 (1958). 「宇治十帖の作者—文章心理学による作者推定」 『心理学評論』2, pp. 147-156.
安本美典 (1982). 「文章様式論」 宮地裕・樺島忠夫・安本美典 (編) 『講座日本語学 8 文体史 II』 明治書院, pp. 1-22.
Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.
Chaussabel, D. (2004). Biomedical literature mining: Challenges and solutions in the 'omics' era. *American Journal of Pharmacogenomics,* 4(6), pp. 383-393.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics,* 19(1), pp. 61-74.
Gordon, A. D. (1999). *Classification.* 2nd ed. Boca Raton: Chapman and Hall.
Pagès, J. (2004). Multiple factor analysis: Main features and application to sensory data. *Revista Colombiana de Estadística*, 27(1), pp. 1-26.
Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont: Lifetime Learning Publications.
Wong, S., Gauvrit, H., Cheaib, N., Carré, F., & Carrault, G. (2002). Multiple factor analysis as a tool for studying the effect of physical training on the autonomic nervous system. *Computers in Cardiology,* 29, pp. 437-440.

関連 URL

統計と機械学習による日本語史研究

`http://www.ninjal.ac.jp/research/project/c/statisticsja/`

日本語歴史コーパス

`http://www.ninjal.ac.jp/corpus_center/chj/`

中古和文 UniDic

`http://www2.ninjal.ac.jp/lrc/index.php?UniDic`

洒落本コーパスの構造化　―仕様と事例の検討―

市村 太郎　（国立国語研究所 コーパス開発センター）†
河瀬 彰宏　（国立国語研究所 コーパス開発センター）
小木曽 智信（国立国語研究所 言語資源研究系）

Structuring the Corpus of *Share-bon*

Taro Ichimura　(National Institute for Japanese Language and Linguistics)
Akihiro Kawase　(National Institute for Japanese Language and Linguistics)
Toshinobu Ogiso　(National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所「通時コーパス」プロジェクトの一環として検討されている、『洒落本大成』の XML 形式での電子化について、資料の電子化に際し、いかなる要素を認定し、どのように構造化するのが適切かについて検討し、モデルを示す。

市村・河瀬・小木曽(2012)では、洒落本に狂言を加え「近世口語テキスト」全体での基礎的な構造化仕様を検討した。本発表では、さらに実際の作業を経たうえでの再検討を加え、洒落本の文書構造や語・文字についてどのように処理し、その結果いかなるデータができるかを、いくつかの作品を例にとり、提示する。

２．洒落本のコーパス化の意義

洒落本は、登場人物の会話部分に当時の話し言葉が反映されているとされ、日本語史研究上、近世後期の口語の実態を探る上での重要資料である。

大きく分けて江戸版と上方版があり、その口語体の会話部分はそれぞれの地域の言葉を反映する場合も多い。また年代も 18C 後半から 19C 前半までと幅広く、近・現代語への過渡的状況を伺うのに適している。方言や中央語の形成を知る上でも、不可欠な資料である。

洒落本の電子化資料としては、先駆的なものとして国文学研究資料館の「大系本文データベース」がある。上方の洒落本については「忍頂寺文庫洒落本データベース」に大阪大学忍頂寺文庫所蔵の洒落本類のデータがあり、これも貴重な資料である。いずれも、主に紙面にもとづく外形的な面でマークアップがなされており、また「忍頂寺文庫洒落本データベース」では、漢字を仮名に開いた「解釈データ」もあり、有用である。

しかしながら、現在のところ、品詞分解された索引や、形態論情報付きの大規模なコーパスはなく、また江戸・上方、宝暦期から文化・文政期まで広く見渡せるものはない。『洒落本大成』には幅広い作品が収められているが、利用に際しては、他の電子化データと重複する一部の作品を除き、個々の作品をその都度目視して用例を拾い集める他ない。もし一定の数量を持ち、アノテーションされた形態論情報付きコーパスが完成すれば、近世・近代語史研究に画期的な成果をもたらすことが期待できる。

３．本コーパスの設計方針

底本には『洒落本大成』を用いる。洒落本を対象とした大規模な叢書であり、多くの作品が収録されている。

コーパスの主な利用者としては、言語研究者を想定する。現在、上記のように、「大系本文データベース」のような、紙面にもとづいて外形的なマークアップがなされたテキストデータが存在するが、言語研究の観点からは、さらに言語構造面に重きを置いた構造化が求められる。たとえば、単純なテキストデータで検索する場合、近世のような仮名遣いや

† tichimura@ninjal.ac.jp

漢字表記が多岐にわたる資料では、文字列検索でいちいちすべての表記を想定したうえで検索しなければならず、大きな足枷となる。その点「忍頂寺文庫洒落本データベース」は読みのデータを付記することで特に漢字表記の面での解決がはかられており、また、簡易なマークアップがされているため、ある程度の要素の抽出が可能である。

これらをふまえさらに、たとえば「ここからここまでがタイトルである」「ここからここまでが〇〇の発話である」などといった、文書構造情報を認定し、一定の構造で発話等がマークアップされることによって、得られた用例について、どのような要素の用例なのかを判断することができるのならば、さらに言語データとして有用なものになるだろう。用例が文書構造中のどの箇所で得られたものなのか（たとえば発話なのか地の文なのか）というのは、近世語研究にかぎらず、極めて重要である。さらにそこに形態論情報が付記され、どのような語のどの活用形か、などがあらかじめ特定されていれば、極めて質の高いデータが、電子データ検索によって瞬時に得られる。

本研究では、このような、いかなる要素の、いかなる性質を持つ語の表記体であるという情報が付された用例を一覧表として短時間で取り出すことが可能なコーパスを目指す。

そのため、記述には XML を用い、国語研が作成した『太陽コーパス』の仕様や BCCWJ の仕様、『明六雑誌コーパス』の仕様を継承しながら、TEI P5 を参考に必要なタグを選択・追加し、構造化する。構造化されたデータには、さらに形態素レベルでのタグを付し、品詞情報や活用形など、形態論情報を付与する。

４．洒落本テキストの構造とタグセット

洒落本テキストは，会話部分を主とし，その他序文・前置き的な地の文・後書きで構成されることが多い（図 1)。

①序文・内題・署名等（＋目録・人物解説等） 構成要素：タイトル・本文・日付・署名（時折和歌・漢文等）
②状況描写など前置き的な地の文 構成要素：本文・記号としての話者表示のない発話・引用
③会話部分（中心部） 構成要素：四角囲みなどの話者・発話・地の文・割書（時折小見出しを伴う複数セクション）
④後書き・尾題・出版情報等 構成要素：タイトル・本文・日付・署名・和歌等

図 1　洒落本テキストの構造概略

作品によってばらつきはあるが、全体としては分量上、また構成上、会話部分が中心となり、またその会話部分も話者表示と発話が中心で、その間に地の文や割書が配置される。

洒落本のコーパス化にあたっては、このような文書構造の各要素について、できるだけ均質に、過不足なくマークアップする枠組みを設定することが求められる。

４．１文書の階層構造に関する要素

表 1　文書の構造に関するタグ（太線は階層上の大きな切れ目）

タグ（要素）	説明	属性
<text>	作品（演目）全体、作品のシリーズ・タイトル等を開始タグ内に記述	@textID（必須） @series シリーズ名（必須） @title　作品タイトル（必須） @yomi 作品名の読み（任意） @year　　西暦成立年（必須） @year_w　和暦成立年（必須）
<front>	前付け	
<body>	主本文	
<back>	後付	
<article>	記事	@type（任意）
<titleBlock>	<article>レベルでのタイトル等の記述	
<p>	タイトルや注釈等を除く本文の塊	
<block>	<p>で記述された本文とは区別されるタイトル・注釈等のブロック要素	@type（必須）
<s>	文	
<SUW>	短単位（語）	（多岐にわたるため省略）

洒落本の図 1①～④のような、テキスト構造を表す大きな構成単位は、１つのテキスト全体を表す<text>と、それを構成する<front>①・<body>②③・<back>④の 3 要素から成る。作品に関する情報は、属性値として<text>内に記述する。

さらに、これらの内部は基本的には<article>に分割され、さらにその内部は<p>か<block>に、そしてその内部は<s>に分割される。さらにその文が、形態論情報を記述した短単位<SUW>に分割される。本コーパスでは、<p>は非常に大きな本文の塊に付されるだけであるので、テキストを分割する単位としては<s>と<SUW>が軸となる。

article 要素　前付・後付を除いた中心的本文は、小見出し等を伴う複数の要素から成ることがある。また、前付や後付内には、自序とともに他人が記した文章や出版情報などが併存することがある。このような階層の要素を表すものとして、<article>を用いる。@type 属性で、序・跋・刊記の別等を記述する。

p 要素　<article>内の本文の塊全体で１つ付与する。視覚上、また内容上いわゆる段落を認定するのは困難である。本研究では「主たる本文かそれ以外か」に重点をおいている。

block 要素　視覚上または構成上、明らかに主本文の塊と区別される要素を表す。@type 属性で、タイトル・内題・尾題・小見出し・著者・日付・表・注釈等の別を記述する。

titleBlock 要素　テキストのタイトル（外題）のほか、序文等の後に再度作品のタイトル（内題）等や尾題等が示される場合がある。これらを厳密な階層構造の中に組み込むことは難しい。そのため、<article>と同階層でマークアップし、並列的に扱う。

s 要素　すべてのテキストは文に分割される。ただしいわゆる「文」とは完全に同一ではなく、発話や割書の区切りでも切る。なお、<s>が<s>を含むような階層性は認めない。

SUW 要素　短単位（おおよそ語に相当）を表す。すべての文は短単位に分割される。本研究での基本的な単位である。語彙素・語形・書字形・活用型・活用形・発音形等語に関する多くの情報が、属性で記述される。開発中の「近世口語 UniDic」による解析結果を人手で修正して付与する。

キー	語彙素	発音形出現形	品詞	活用型	活用形
おゆき	オユキ	オユキ	名詞-固有名詞-人名--一般		
さん	さん	サン	接尾辞-名詞的-一般		
はやふ	早い	ハヨー	形容詞-一般	形容詞	連用形-ウ音便
お	御	オ	接頭辞		
いで	出でる	イデ	動詞-一般	下一段-ダ行	連用形-一般
わたし	私	ワタシ	代名詞		
も	も	モ	助詞-係助詞		
これ	此れ	コレ	代名詞		
から	から	カラ	助詞-格助詞		
かみゆひ	髪結い	カミユイ	名詞-普通名詞-一般		
さん	さん	サン	接尾辞-名詞的-一般		
に	に	ニ	助詞-格助詞		
かみ	髪	カミ	名詞-普通名詞-一般		
を	を	オ	助詞-格助詞		
ゆふ	結う	ユー	動詞-一般	五段-ワア行	連用形-ウ音便
て	て	テ	助詞-接続助詞		
もろ	貰う	モロ	動詞-非自立可能	五段-ワア行	連用形-ウ音便
て	て	テ	助詞-接続助詞		
こんや	今夜	コンヤ	名詞-普通名詞-副詞可能		
から	から	カラ	助詞-格助詞		
おしろい	白粉	オシロイ	名詞-普通名詞-一般		
も	も	モ	助詞-係助詞		
し	為る	シ	動詞-非自立可能	サ行変格	連用形-一般
て	て	テ	助詞-接続助詞		
べに	紅	ベニ	名詞-普通名詞-一般		
も	も	モ	助詞-係助詞		
つけ	付ける	ツケ	動詞-非自立可能	下一段-カ行	連用形-一般
て	て	テ	助詞-接続助詞		
おき	置く	オキ	動詞-非自立可能	五段-カ行	連用形-一般
ましよ	ます	マショ	助動詞	助動詞-マス	意志推量形
。	。		補助記号-句点		
ありや	彼れ	アリャ	代名詞		
いろ	色	イロ	名詞-普通名詞-一般		
め	奴	メ	接尾辞-名詞的-一般		
が	が	ガ	助詞-格助詞		
いに	往ぬ	イニ	動詞-一般	五段-ナ行	連用形-一般
おつ	居る	オッ	動詞-非自立可能	五段-ラ行	連用形-促音便
た	た	タ	助動詞	助動詞-タ	終止形-一般
これ	此れ	コレ	代名詞		
のふ	ノウ	ノー	感動詞-一般		
ををい	おい	オーイ	感動詞-一般		
/ \	/ \		補助記号-一般		

図 2　短単位解析済みデータの例（一部項目省略・8 巻『風流裸人形』p.277 上段 2 行～）

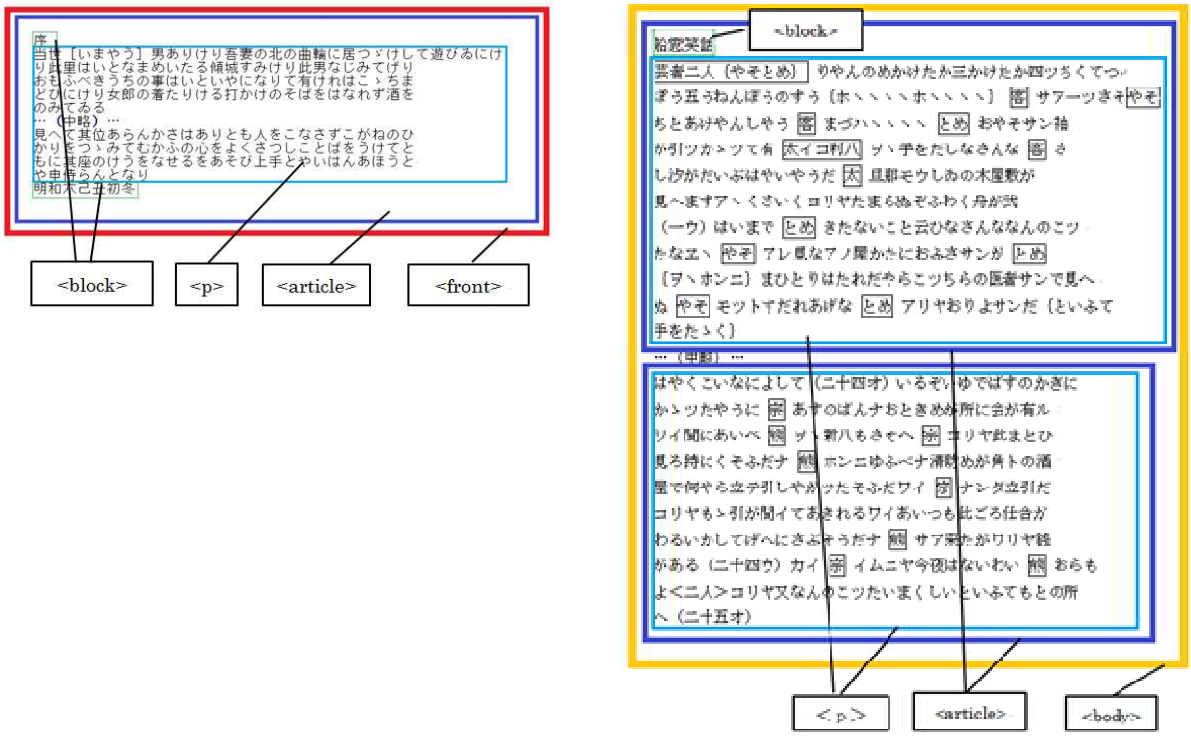


図 3　作品冒頭～本文のマークアップ（4 巻『郭中奇譚』pp.297-308）

４．２文・語の機能に関する要素

表 2　文・語の機能に関するタグ

タグ（要素）	説明	属性	
<speech>	会話	@source（任意） @type（任意）	
<quotation>	単純な発話以外の引用要素	@type（任意） @source（任意）	
<warigaki>	割書き		↑文以上
<speaker>	話者		↓文未満
<delivery>	発話等のスタイルの表示		
<verse>	韻文		

speech 要素　1 回的な会話文の連続を表す。<speaker>を内部に認定し、一体として扱う。会話文内に話者が示されていない場合には@source 属性で話者を可能な限り記述する。なお、割書内にもごく簡単な会話文が出現することもあるが、割書中では認定しない。

quotation 要素　和歌・手紙等、単純な会話文以外の引用要素を表す。@type でどのような種の引用かを、@source で出典を記述する。

warigaki 要素　多くは細字二行で、会話部分における地の文または注釈として発話間に現れる。ただし笑い声や間投詞の類が小書きで 2 行に渡るものは割書とは認めない。

speaker 要素　会話文に付属する、話者の表示である。主に囲みや小書きで表される。

delivery 要素　会話文の冒頭には、話者だけでなく、「歌」などと、スタイルを小書き等で記してある場合がある。その場合に本要素を付与する。

verse 要素　和歌・俳句・歌等明らかな韻文ついて、文未満の単位で（主に文毎に）付す。

４．３主に語・文字単位で外形・機能等を表す要素

表 3　語・文字単位で外形等を表す要素

タグ(要素)	説明	属性	
<hi>	文字列（語）に対する装飾	@rend（必須）	↑短単位以上
<lRuby>	左ルビ	@rubyText（必須） @rubyBase（任意）	↓短単位未満
<ruby >	ルビ	@rubyText（必須） @rubyBase（任意）	
<odoriji>	踊り字を開いた文字	@originalText（必須）	
<gap/>	抹消・破損等で判読できない文字の存在（空要素）		
<corr> <corr/>	本文修正	@type（必須） @originalText（任意）	
<unclear>	推読された文字	@originalText（任意） @type（任意）	
<vMark>	濁点・半濁点付仮名に変換した箇所		
<g>	外字（JISX0213 外）絵文字等	@type（必須） @ref（任意）	
<kana>	片仮名を平仮名に変換した箇所		
<kanbun> <kanbun/>	漢文（返読）	@type（任意）返読前 返読後等 @originalText （ 任 意 ） @id（任意）	

hi 要素　○や□で囲まれるなど外形的特徴を持った語以上の文字列を表す。囲みの人物表示は必ずしも話者になるわけではなく、機能は一定ではない。このようなものを外形的にマークアップし、@rend 属性で様態を記述する。「マアマア」等間投詞的なものは除く。

ruby 要素　文字列の右側に付され、文字・文字列の読みを表す振り仮名等を指す。@rubyText 属性内にルビ文字列を記述し、複数短単位に対して付されている場合は、先頭の短単位のみに認定し、@rubyBase に実際の対象文字列を付す。いわゆる宛漢字等も含む。

lRuby 要素　文字列の左側に付される小書き。例えば本文の方言形に対応する語を左側に記すなど、概して注釈的性質がある。右側ルビと共存する場合は、右側ルビよりも比較的対象範囲が大きい。語単位で付されることが多い。

vMark 要素　電子化に際して新たに濁点を付与した箇所に付す（踊り字の箇所は除く）。

４．４位置情報と本文外情報

表４　底本テキストの位置情報や本文外の情報を表すタグ

タグ(要素)	説明	属性
<pb/>	ページ開始（空要素）	@n（必須）
<cb/>	段開始（空要素）	@n（必須）
<lb/>	行開始（空要素）	
<info/>	本文外情報（空要素）	@originalPage（任意）@ text（任意）

```
<text textID="洒落本大成_024_京都_興斗月" series="洒落本大成#24" title="興斗月" yomi="きょとつき" year="1836" year_w="天保 7"><front><article type="序"><p><s><pb n="131"/><cb n="1"/><lb/>年<vMark>ご</vMark>ろ我勝れて河東を好め<vMark>ど</vMark>も価高きゆへうとまれて行こ<lb/>と稀也</s><s>只老留誌の類を見て鬱を散る而已なりし<vMark>が</vMark>或夜東方<lb/>に見馴ぬ光あり</s><s>これなむ興斗つきといふ<info originalPage="一オ"/></s><s>唖の川に輝<lb/>舟あり</s><s>是に乗て蚓ののたくり一冊としはりに至ま<vMark>で</vMark>自作せ<lb/>りと慢して馬鹿の底をた<odoriji originalText="ゝ">た</odoriji>く</s><s>是を名号て興斗つゐて何処ま<lb/><vMark>で</vMark>乗て行<info originalPage="一ウ"/>と云<lb/></s></p><block type="date"><s>天保七年<lb/>申<kana>の</kana>孟夏<lb/></s></block><block type="author"><s>前代未聞<lb/>武木右衛門<lb/>自序<lb/><info originalPage="二オ"/></s></block></article><titleBlock><block type="内題"><s><pb n="132"/><cb n="1"/><lb/>興斗月<lb/></s></block><block type="author"><s>武木右衛門戯作</s></block></titleBlock></front>（以下略）
```

図 4『興斗月』冒頭の形式化例（大成 29 巻 pp.131-132）

```
<body><article><p><s><ruby rubyText="おゝ">大</ruby><ruby rubyText="き">木</ruby><ruby rubyText="ど">戸</ruby>の<ruby rubyText="ちり">塵</ruby>は<ruby rubyText="みづ">水</ruby><ruby rubyText="うり">売</ruby>の<ruby rubyText="しづく">雫</ruby>にしめり<ruby rubyText="てん">天</ruby><ruby rubyText="りう">竜</ruby><ruby rubyText="じ">寺</ruby>の<ruby rubyText="かね">鐘</ruby>は<ruby rubyText="ひぐらし">蜩</ruby>の<ruby rubyText="こへ">声</ruby>にひ<odoriji originalText="ゞ">び</odoriji><lb/>く</s><s><kana><hi rend="囲み">くつわのをと</hi></kana></s><s>ちやんらん／＼</s><speech><s><speaker><hi rend="囲み">馬士二人歌</hi></speaker></s><s><verse>お<odoriji originalText="ゝ">お</odoriji>れへと<info text="上"/>な<kana>あ</kana>引い<lb/>かぬ<kana>あ</kana>う</verse></s><s><verse><kana>そ</kana><kana>れ</kana>そうだにな<kana>あ</kana>引</verse></s></speech><speech><s><speaker><hi rend="囲み"><kana>あ</kana><kana>と</kana><kana>の</kana>馬士</hi></speaker></s><s>かみ<ruby rubyText="むら">村</ruby>の
```

<kana>う</kana><ruby rubyText="ゑ">江</ruby><ruby rubyText="ご">五</ruby>右<lb/><ruby rubyText="ゑ">衛
</ruby><ruby rubyText="む">門</ruby>が<kana>あ</kana>よめ<ruby rubyText="じよう">女</ruby><kana>なあ
</kana><ruby rubyText="うみ">産</ruby><ruby rubyText="づき">月</ruby>だ<kana>あ</kana>といつけがどふだ<kana>あ
</kana></s><s>まだひり<lb/><ruby rubyText="だ">出</ruby>さねへかな<kana>あ</kana></s></speech>（中略）
<speech><s><speaker><hi rend="囲み">金</hi></speaker></s><s><kana>あ</kana><kana>い</kana>さあ。おさらば／＼</s>
</speech><s>○<ruby rubyText="なつ">夏</ruby>の<ruby rubyText="よ">夜</ruby><lb/>は。まだ<ruby rubyText="よひ">宵
</ruby>ながら。<ruby rubyText="あけ">明</ruby>ぬるを。<ruby rubyText="し">知</ruby>らせよふとて。<ruby rubyText="
からす">烏</ruby>がか<lb/>あ／＼。<ruby rubyText="かね">鐘</ruby>がごん／＼。<ruby rubyText="つき">春</ruby><ruby
rubyText="ごめ">米</ruby>屋ががつたり／＼<info originalPage="丁付なしオ"/></s> </article></body>
<back><article type="跋"><block type="section"><s><lb/>跋</s></block> <s><lb/><cb n="1"/><pb n="311"/><ruby
rubyText="すい">粋</ruby>とは<ruby rubyText="うめ">梅</ruby><ruby rubyText="ぼし">干</ruby><ruby rubyText="や">野
</ruby><ruby rubyText="ぼ">父</ruby>とは<ruby rubyText="にはとり">鶏</ruby>の名かときくやうな<ruby rubyText="し
ん">新</ruby><ruby rubyText="じゆく">宿</ruby>田舎にあや<lb/>め咲とはしほらしとぞめきの<ruby rubyText="こえ">声
</ruby><ruby rubyText="う">有</ruby><ruby rubyText="てう">頂</ruby><ruby rubyText="てん">天</ruby>にひ<odoriji
originalText="ゞ">び</odoriji>き（中略）
<s>鳴<lb/><ruby rubyText="あゝ">呼</ruby><ruby rubyText="わが">吾</ruby><ruby rubyText="とう">党</ruby>いきちよ
んの君子をしてこれにあそはしめば<ruby rubyText="すなはち">則</ruby>其<lb/><ruby rubyText="しり">尻</ruby>つま
らざるにちか<odoriji originalText="ゝ">か</odoriji>らん<ruby rubyText="ずい">随</ruby><ruby rubyText="いき">行
</ruby><ruby rubyText="さん">散</ruby><ruby rubyText="じん">人</ruby><ruby rubyText="ずい">随</ruby><ruby
rubyText="がへり">帰</ruby>の<ruby rubyText="まくら">枕</ruby><ruby rubyText="もと">上</ruby>に<ruby rubyText="ば
つ">跋</ruby>す<lb/></s> <block type="date"><s>安永乙未秋<lb/></s></block><block type="publisher"><s>新甲館蔵書
<lb/><info originalPage="丁付なしオ"/></s></block></article></back>

図 5『甲駅新話』本文・後付の例（大成 6 巻 pp.295-311）

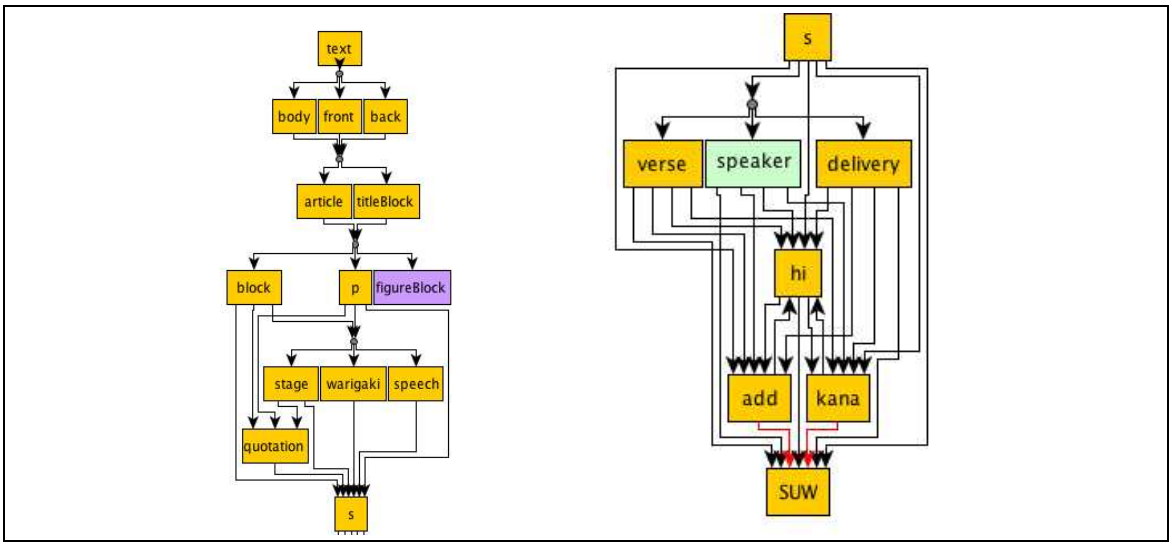


図 6「近世口語コーパス」の文書型定義図（SUW 要素まで）

５．本研究における課題

５．１　割書きの扱いをめぐって

割書や引用の前後で文を区切ると、「という」など半端な文が生じるという問題がある。一方で、テキストの構造からみれば〔会話―割書―会話―割書…〕という直感的に定式化した流れがある。このような複数の構造を厳密にカバーするのは困難である。

しかし、言語学的な利用を考慮すると、「割書・引用で文を区切る」ことは、多くの場合「発話と地の文を区別する」ことと合致する。また割書を解体するにしても、分類や単位の切れ目の認定は困難であり、洒落本においては、最大公約数的に「主本文に対する付属的な何か」を表す「割書」であるほうが利用の便を考慮しても現実的ではないかと考える。

５．２文認定と解釈の問題

根本的な問題であるが、『洒落本大成』は注釈や句読点等が付された校訂本文ではなく、形態論情報はもちろん、文区切りを与える際には、高度な文解釈が求められる。活用語の終止形と連体形が統一される時期でもあり、文末の認定にはしばしば困難が伴う。話者が長々と話すもの、道行などのように語調がかかわるものは特に難解である。文脈・活用形・ソ系指示詞などが手がかりだが、それでも不明確な箇所については強いて文区切りを付すことはしない方針である。

例を挙げる。下のように、名詞が連続する場合、並列的に述べられているのか、文が切れるのか、厳密に判断するのは難しい。

（例）店もなく揚屋もなく商ひ場といふてはうゑもなき雲天のざしきいつこさためぬ枕の数（大成 7 巻『無論里問答』p.50 下段）

また、引用周りの扱いも問題となる。発話の連続を引用の「と」等で受ける場合、「と」がどこまでをマークするのか不明確なことが多く、また一口に文と言っても、場合によっては幾重にも階層ができ、巨大な文が出来上がってしまい、著しく均質性を損なうことがある。そのため、原則直接的な引用や割書き、話者表示の前後では文を区切る。

発話内の引用については、間接引用なのか直接引用なのかが不明確な場合が多い。ただし、手紙を読み上げる箇所等があり、このような明確に直接引用とわかるものについては、文区切りを付す方針である。

このように可能な限り客観的な根拠を探るのが原則だが、細部に決定的な規則を設けるのは困難である。個別に判断し、場合によっては保留せざるを得ないのが現状である。

５．３　修辞・言葉遊びへの対応

掛詞や洒落のような言葉遊びや、文や語と区切りとは関係ない七五調などが見られる。これらは近世期に限らず歴史的作品では重要な修辞技法の一種であるが、階層構造から逸脱したパラレルなものである。これに対しては、索引等では二重に採取する方針がとられているものもあるが、コーパスの開発に利用している現状のシステムでは形態論情報の二重付与に対応していない。今後、システム拡張を含め、対応を検討していく必要がある。

（A）我搗栗といわゐそめ老もわかいも<u>よろ昆布</u>（大成 7 巻『三幅対』p.352 下段 16 行）

（B）＜知暁＞ごさまのかぎ迄預けしは＜青蜉＞<u>づけし</u>は物を思はざりけり＜几石＞<u>ざりけり</u>のちじよくをすゝぐ（大成 2 巻『穿当珍話』p.207 上段 3～5 行）

６．おわりに

文や引用の認定・解釈は、歴史的な資料をコーパス化する際の大きな課題である。また、５．３のような修辞・言葉遊びの類は、今後和歌集や歌舞伎・浄瑠璃を積極的に扱うことを考慮すると、大きな課題である。

歴史的資料を対象にコーパスを構築するにあたっては、外形と機能、言語の線条性と版面がもつ構造のバランスをとり、適切にラベルを与えていくことが重要である。その上で「何を拾いたいか、どこまで期待されているか」という利用者のニーズに沿う必要がある。

1 作品中に会話・地の文・割書き・序・後書き・手紙など、比較的多様な要素を持つ洒落本を対象に 1 つの記述モデルを確立しておくことは、「日本語歴史コーパス」全体に汎用性をもつ仕様を作る上での一つの足掛かりになると考える。

文　献

市村 太郎、河瀬 彰宏、小木曽 智信(2012)『近世口語テキストの構造化とその課題』情報処理学会研究報告　人文科学とコンピュータ研究会報告(CH96)　pp.1-8

近藤明日子、田中牧郎「『明六雑誌コーパス』の仕様」『国立国語研究所共同研究報告 12-03 近代語コーパス設計のための文献言語研究 成果報告書』pp.118-143 国立国語研究所

近藤泰弘(2012)「日本語通時コーパスの設計について」『国語研プロジェクトレビュー』3 pp.84-92 国立国語研究所

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」『国立国語研究所報 122　雑誌『太陽』による確立期現代語の研究『太陽コーパス』研究論文集』pp.1-48　博文館新社

田中牧郎、小木曽智信(2000)「総合雑誌『太陽』の本文の様態と電子化テキスト」『日本語科学』8　pp.141-152　国立国語研究所

安永尚志(1998)『国文学研究とコンピュータ』勉誠社

山口昌也、高田智和、北村雅則、間淵洋子、大島一、小林正行、西部みちる(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『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2』　文部科学省 科学研究費 特定領域研究 「日本語コーパス」データ班

洒落本大成編集委員会(1978-88)『洒落本大成』中央公論社

関連 URL

「大系本文（日本古典文学・噺本）データベース」http://base3.nijl.ac.jp/

「忍頂寺文庫洒落本データベース」http://www.let.osaka-u.ac.jp/~iikura/Ninjoji_Ono/syarebon.html

「Text Encoding Initiative」（ガイドライン P5 日本語版）
http://docsci.infon.org/stack/P5JA/index-toc.html

説話のパラレルコーパスの設計

—平安・鎌倉時代の文体変異の研究に向けて—

田中 牧郎　（国立国語研究所言語資源研究系）†

Design of a Parallel Corpus of *Setsuwa* Literature:

Toward Studies of Stylistic Variation of Heian and Kamakura Japanese

TANAKA Makiro (National Institute for Japanese Language and Linguistics)

１．日本語史における文体的変異

言語史をとらえるためのコーパスの設計においては、言語の時間的変異に加えて社会的変異をどのように反映させるかを研究することも求められる。日本語史上重要な社会的変異には、地域、階層などによる変異も考えられるが、文献資料によって最も明確に跡づけることができるのは文体による変異である。特に、現代日本語の書き言葉の源流となった和漢混淆文が確立する鎌倉時代までのそれを正しく把握することは、極めて重要なことである。図１は、鎌倉時代までの文体的変異の概略を示したものである。

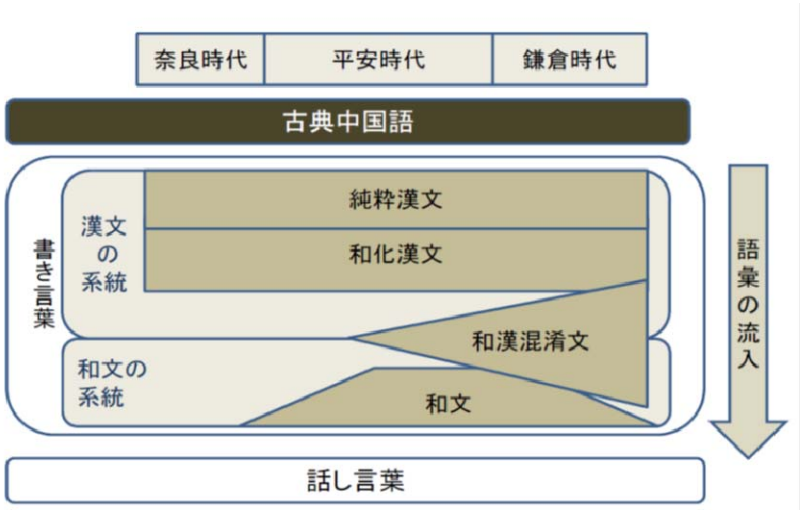


図１　日本語（鎌倉時代まで）の文体的変異

古い時代の話し言葉そのものは資料がない。書き言葉のうち、和文、和漢混淆文、和化漢文、純粋漢文を、それぞれバランスを考慮しながらコーパス化していくことが求められよう。一方で、言語史研究は、書き言葉そのものよりも話し言葉の歴史に関心を寄せることの方が多いために、まずは、話し言葉に近い書き言葉である、和文の系統を優先させることは現実的な方策であろう。また、漢文の系統の書き言葉には、古典中国語の要素も濃いため、日本語史資料としては扱いにくい面も強くある。このような背景から、国立国語研究所の「通時コーパス」（歴史コーパス）の開発においても、まずは、和文の系統のテキ

† mtanaka@ninjal.ac.jp

ストからコーパス化に着手し、順次、漢文の系統のテキストに手を広げていくことを考えている。

ところで、上記のような文体的変異をとらえることのできるコーパスを設計するには、それぞれの文体のテキストを個別に選定してコーパス化するだけでなく、文体的変異が分析しやすいように、テキスト相互を関連付ける工夫をすることも望まれよう。そのような工夫の一つとして、本稿では、同一内容が異なる文体で書かれたテキストに着目し、そのパラレルコーパスを設計することについて検討を加えたい。

２．同文説話

日本語史の資料となるテキストにおいて、同一内容が異なる文体で書かれたものは多く存在するが、特に、古典文学のジャンルの一つである「説話」は、そのようなケースが多い一群として注目される。説話は伝承を書き記した文学であるが、説話集に収録されることで後世に伝わる。その説話集には、様々な文体で書かれたものがあり、その結果、同一の説話（同話）が異なる文体で記録されることも多い。特に、平安時代から鎌倉時代の説話集に収録された説話は、文体的変異をとらえるのに好適なものが種々現存している。

日本の伝承は、仮名成立以前は純粋漢文や和化漢文で書かれ、仮名が成立した後は和文で書かれものも現れるが、その書かれた説話は説話集の文体に応じて種々の文体として実現する。一方、中国から漢文で書かれた説話が書物を通して日本に伝えられ、純粋漢文として受容されることがあるが、それが日本で訓読されたり、訓読の域を脱して和化漢文や和漢混淆文に翻訳されたりして、説話集に収録される場合もある。このような、異なる説話集に収録される同内容の説話は、「共通説話」と呼ばれる。共通説話のなかには、内容は同じでも、その表現が大きく異なり、文どうしの対応がとれないものも多い。一方、共通説話の中には、表現の共通性が高く、文レベル、語レベルで対応を取ることが可能な「同文説話」も多い。この同文説話をパラレルコーパスにすれば、文体的変異をとらえるのに好適なデータベースとなるだろう。

文体的変異の大きい同文説話を数多く収録しているのが、和漢混淆文である『今昔物語集』（平安時代末期、12 世紀前半成立）とその周辺の説話集である。『今昔物語集』との同文説話を多くおさめる説話集には様々なものがあるが、小峯（1997）などを参考に、『今昔物語集』との同文説話の話数を集計して掲げると次の通りである。

漢文系統の説話集

純粋漢文：三宝感応要略録（63 話）、冥報記（49 話）、法苑珠林（14 話）など

和化漢文：法華験記（96 話）、日本霊異記（74 話）、日本往生極楽記（32 話）など

和文系統の説話集

宇治拾遺物語（61 話）、古本説話集（28 話）、俊頼髄脳（8 話）など

同文とする認定にはゆれも多いので、上記の話数は目安にとどまるが、漢文系統（純粋漢文・和化漢文）の説話集にも、和文系統の説話集にも、『今昔物語集』との同文説話が多いことが見てとれる。『今昔物語集』と上記の説話集との同文説話を使って平安時代末期の文体的変異を明らかにしようという研究は、佐藤（1984）、山口（1984）、藤井（2003）、今野（2009）、舩城（2011）など、日本語学の分野に多くの蓄積がある。しかし、例えば和漢混淆文である『今昔物語集』が、和文と和化漢文のどちらにより近いのかということについての研究者の見解は一致していないなど、この時期の文体的変異の全体像は実はよく分かっていない。『今昔物語集』とその周辺の説話集の同文説話のパラレルコーパスを作るこ

とは、このような議論を実り多いものにすることができると思われる。それは同時に、和漢の両系統を統合するような通時コーパスの作成のための重要な一段階にもなるだろう。

３．和文説話集と『今昔物語集』の同文説話のパラレルコーパス

3.1　和文説話集と『今昔物語集』のテキスト

和文説話集のうち、『古本説話集』（平安時代末期、12 世紀前半）及び『宇治拾遺物語』（鎌倉時代前期、13 世紀前半）と、『今昔物語集』の同文説話は、直接的な継承関係によって生じたものではなく、散逸した『宇治大納言物語』（平安時代後期、11 世紀中ごろ）に収録する説話を、それぞれが採ったことによって生じたものである。先行研究によれば、『宇治大納言物語』の説話を採録する際に、『古本説話集』『宇治拾遺物語』はその表現にあまり手を加えず、『今昔物語集』は手を加えることも多かったとされている。

同文説話の一例として、『宇治拾遺物語』第 121 話と、『今昔物語集』巻 4-9 の冒頭部分を示してみよう。[1]

宇治拾遺物語　121

> 昔、天竺に一寺あり。住僧もつとも多し。達磨和尚この寺に入りて、僧どもの行ひを窺ひ見給ふに、ある坊には念仏し、経を読み、さまざまに行ふ。ある坊を見給に、八九十ばかりなる老僧の、ただ二人ゐて圍碁を打つ。

今昔物語集 巻 4－9

> 今昔、天竺に、陀楼摩和尚と申す聖人在ます。此の人、五天竺に不行至ぬ所無く行て、諸の比丘の所行を善く見て世に傳へ給ふ人也。一の寺有り、其の寺に入て比丘の有様共を伺ひ見る、寺に比久共多く住む。或る房には佛に花香を奉り或る房には経典を讀誦する比丘有り、様々に貴く行ふ事無限し。但し其の中に一の房有り、人住たる氣色無し、草深塵積れり。深く入て見れば、八十許なる老比久二人居て碁を打つ。

3.2　文と語の対照

上記の二つのテキストをそれぞれ文に区切り、文単位で対照すると、表１のようになる。『宇治拾遺物語』の４文が、『今昔物語集』の７文に対応しており、1 対 1 で対応している文はなく、複数の文が入り組んで対応している様子が見えている。

二つのテキストの対照は、語レベルにまで降りていって対応付けることも可能であり、表２と表３は、『宇治拾遺物語』の第一文と、『今昔物語集』の第一文について、それぞれ対照した作業の結果を示したものである。『宇治拾遺物語』から『今昔物語集』を見た表 2 によると、すべての語に対応語があるが、4 番目の「一寺」には、「一/の/寺」3 語が対応している。また、『今昔物語集』から『宇治拾遺物語』を見た表 3 によると、1 番めの「今」、7 番目以降の「と」「申す」「聖人」「在(まし)ます」には、対応語がない。また、5 番目の「陀楼摩」の対応語は「達磨」で表記が異なっているが、同語の「だるま」と認定する。さらに、表２と表３の範囲にはないが、対応語があっても、別の語に対応することもあり、例えば、表１の『宇治拾遺物語』の３番目の文の３つめの語「この」は、『今昔物語集』では４つめ

[1] 『宇治拾遺物語』のテキストは『新編日本古典文学全集』（小学館）による。『今昔物語集』のテキストは、巻 1～巻 10 は『日本古典文学大系（旧版）』（岩波書店）、巻 11～31 は『新編日本古典文学全集』による。『新編日本古典文学全集』のテキストは、小学館から国立国語研究所に提供されたものを用い、『日本古典文学大系（旧版）』のテキストは、国文学研究資料館の「日本古典文学本文データベース」のものを利用した。なお、『今昔物語集』のテキストは漢字片仮名交じり文であるが、片仮名を平仮名に変換してデータ化を行った。

の文の冒頭の語「其の」に対応している。

このように、文レベル、語レベルの対応のありようは複雑であり、パラレルコーパスを作る際は、何と何を対応付けるのかについて、認定にゆれが生じないような詳しい基準を作成しておく必要がある[2]。

表 1『宇治拾遺物語』121 と『今昔物語集』巻 4-9 の文の対応

文 id	宇治拾遺物語	今昔対応文	文 id	今昔物語集	宇治対応文
u001	昔、天竺に一寺あり。	k001-003	k001	今（は）昔、天竺に、陀楼摩和尚と申す聖人在ます。	u001-003
			k002	此の人、五天竺に不行至ぬ所無く行て、諸の比丘の所行を善く見て世に傳へ給ふ人也。	u#
			k003	一の寺有り。	u001
u002	住僧もつとも多し。	k004	k004	其の寺に入て比丘の有様共を伺ひ見る、寺に比久共多く住む。	u002-003
u003	達磨和尚この寺に入て、僧どもの行ひをうかゞひ見給ふに、ある坊には念仏し、経を読み、さまざまに行ふ。	k001-005	k005	或る房には佛に花香を奉り或る房には経典を讀誦する比丘有あり、様々に貴く行ふ事無限し。	u003
u004	ある坊を見給ふに、八九十ばかりなる老僧の、ただ二人ゐて圍碁を打つ。	k006-007	k006	但し其の中に一の房有り、人住たる氣色無し、草深塵積れり。	u004
			k007	深く入て見れば、八十許なる老比久に人居て碁を打つ。	u004

表 2 『宇治』から『今昔』への語の対応

文 id	語 id	宇治	今昔	対応文	対応語	分類
u001	001	昔	昔	k001	002	①同語
u001	002	天竺	天竺	k001	003	①同語
u001	003	に	に	k001	004	①同語
u001	004	一寺	一の寺	k003	001-003	②別語
u001	005	あり	有り	k003	004	①同語

表 3『今昔』から『宇治』への語の対応

文 id	語 id	今昔	宇治	対応文	対応語	分類
k001	001	今	¢	u001	#	③非対応
k001	002	昔	昔	u001	001	①同語
k001	003	天竺	天竺	u001	002	①同語
k001	004	に	に	u001	003	①同語
k001	005	陀楼摩	達磨	u003	001	①同語
k001	006	和尚	和尚	u003	002	①同語
k001	007	と	¢	u003	#	③非対応
k001	008	申す	¢	u003	#	③非対応
k001	009	聖人	¢	u003	#	③非対応
k001	010	在ます	¢	u003	#	③非対応

このような対応付けの作業については、基準を整備した上で進めていく必要があるが、その基準を策定しつつ探索的な試行作業を、次の５対の同文説話に対して実施した。デー

[2] 個々のテキストにおける単語認定は、小木曽（2012）などが示す、中古和文の短単位認定基準に従う。これとは別に、二つのテキストを対応付ける基準の策定が必要である。

タ量は、和文説話集が延べ語数で約 5500 語、『今昔物語集』が約 6700 語である。その思考結果の概要は 5 節で報告する。

宇治 31／今昔巻 23-21、宇治 91／今昔巻 5-1、宇治 102／今昔巻 14-29、宇治 137／今昔巻 4-9、宇治 187／今昔巻 31-11

４．漢文説話集と『今昔物語集』の同文説話のパラレルコーパス

4.1　漢文説話集と『今昔物語集』のテキスト

『今昔物語集』との同文説話を収録する漢文系の説話集は多様で、直接的な関係があるものと間接的な関係にとどまるものとがあり、その識別は難しい場合も多い。直接関係にない『宇治拾遺物語』と『今昔物語集』をパラレルコーパスにしたように、同文性の高い同文説話を持つ漢文説話集も、パラレルコーパス作成の対象としてよいだろう。ここでは、『日本霊異記』中巻 32 話と『今昔物語集』巻 20-22 を例に取り上げたい。この二書の関係については、『今昔物語集』が『日本霊異記』（平安時代初期、9 世紀前半）を直接の典拠として翻訳したとされている。上記のペアの各冒頭部を掲げよう[3]。

日本霊異記 中巻 32

聖武天皇世、紀伊國名草郡三上村人、為薬王寺、率引知識、息晋薬分。其薬料物、寄乎岡田村主姑女之家、作酒息利。時有斑犢。入薬王寺、常伏塔基。

今昔物語集 巻 20-22

今昔、紀伊国の名草の郡三上の村に一の寺を造て、名を薬王寺と云ふ。其後、知識を引て、諸の薬を儲て、其の寺に宜て、普く人に施しけり。而る間だ、聖武天皇の御代に、其の薬の料物を、岡田の村主と云者の姑の家に宿し置く。而るに、其の家の主其の物を酒に造て、其を人に与へて、員を増して得むと為るに、其の時に、斑なる小牛出来て、薬王寺の内に入、常に塔の本に臥す。

4.2　文と語の対照

３節で和文説話集との同文説話の対応付けを行ったのと同様の方式で、文レベルの対応と語レベルの対応を示すと、表 4、表 5、表 6 のようになる。漢文説話集を扱う際には、和文説話集の場合にはなかった問題が二点あり、その処理方法をあらかじめ決めておく必要がある。

第一点は、原文そのままの漢文テキストを扱うのではなく、訓読文テキストを扱うという点である。平安・鎌倉時代の日本における漢文は、純粋漢文は訓読されることも多く、和化漢文は日本語を表記している部分が多く、特に『今昔物語集』の説話と対応付ける際は、漢文としてではなく日本語文として扱うべきだと考えられる。表 4〜6 の『日本霊異記』のテキストは、『新編日本古典文学全集』の訓読文によったものである。訓読する際に、原文にない文字を読み添えたことによって生じた語は、括弧で括った[4]。括弧で括った原文にない語については、集計や分析の対象から外すことが必要であろう。

第二点は、語の同定の問題である。和文や和漢混淆文の場合は、語の同定は比較的容易で、どこまでを一語と認めるのか、またどう読むのかについて、迷うものは多くない。これに対して漢文は、単位や読みの同定について複数の可能性が想定できるものが多い。『今昔物語集』との対応付けということを考える場合、複数の可能性がある場合は『今昔物語

[3] 『日本霊異記』のテキストは、『新編日本古典文学全集』（小学館）による。

[4] 活用語尾など語の一部を読み添えるものについては括弧などで括ることはしなかった。

集』の表現に近いものを採るという規則を立てるのが現実的だろう。『日本霊異記』の場合は、原則として『新編日本古典文学全集』の訓読文に従い、『今昔物語集』の表現により近い認定が可能なものについては、これを修正していくことにしたい。

表 4　『日本霊異記』中巻 32 と『今昔物語集』巻 20-22 の文の対応

文 id	霊異記	今昔対応	文 id	今昔	霊異記対応
r001	聖武天皇（の）（み）世（に）、紀伊國名草郡三上（の）村（の）人、藥王寺（の）為（に）、	k001,k003	k001	今（は）昔、紀伊（の）国の名草の郡三上の村に一の寺を造て、名を薬王寺と云ふ。	r001
r002	知識（を）率引（して）、晋（く）藥分（を）息し（き）。	k002	k002	其後、知識を引て、諸の薬を儲て、其の寺に宜て、普く人に施しけり。	r002
r003	其（の）藥料（の）物（を）、岡田村主（の）姑女之家（に）寄せ乎、	k003	k003	而る間だ、聖武天皇の御代に、其の薬の料物を、岡田の村主と云者の姑の家に宿し置く。	r001,r003
r004	酒（を）作り利（を）息し（き）。	k004	k004	而るに、其の家の主其の物を酒に造て、其を人に与へて、員を増して得むと為るに、其の時に、斑なる小牛出来て、薬王寺の内に入て、常に塔の本に臥す。	r004-006
r005	時（に）斑（なる）犢有り（き）。	k004			
r006	藥王寺（に）入り、常（に）塔（の）基（に）伏せ（り）。	k004			

表 5　『霊異記』から『今昔』への語の対応

文 id	語 id	霊異記	今昔	対応文	対応語	分類
r002	001	知識	知識	k002	003	①同語
r002	002	（を）	を	k002	004	—
r002	003	率引	引	k002	005	②別語
r002	004	（し）	#	k002	#	—
r002	005	（て）	て	k002	006	—
r002	006	晋（く）	普く	k002	020	①同語
r002	007	藥分	薬	k002	009	②別語
r002	008	（を）	を	k002	010	—
r002	009	息し	施し	k002	023	
r002	010	（き）	けり	k002	024	—

表 6　『今昔』から『霊異記』への語の対応

文 id	語 id	今昔	霊異記	対応文	対応語	分類
k002	001	其	¢	r002	#	③非対応
k002	002	後	¢	r002	#	③非対応
k002	003	知識	知識	r002	001	①同語
k002	004	を	（を）	r002	002	—
k002	005	引	率引	r002	003	②別語
k002	006	て	（て）	r002	005	—
k002	007	諸	¢	r002	#	③非対応
k002	008	の	¢	r002	#	③非対応
k002	009	薬	藥分	r002	007	②別語
k002	010	を	（を）	r002	008	—
k002	013	儲	¢	r002	#	③非対応
k002	014	て	¢	r002	#	③非対応
k002	015	其の	¢	r002	#	③非対応
k002	016	寺	¢	r002	#	③非対応
k002	017	に	¢	r002	#	③非対応
k002	018	宜	¢	r002	#	③非対応

(以下略)

このように漢文説話集の場合は、和文説話集の場合以上に詳細な基準が必要になるが、その基準を作成しながら、上記のような対応付けを次の共通説話の 10 対について試行した。分量は漢文説話が延べ約 2000 語、『今昔物語集』が約 4000 語である。

法苑珠林巻 37 敬塔篇 35 施繞部 5／今昔巻 1-36、法苑珠林巻 37 敬塔篇 35 感福部 5／今昔巻 2-11、三宝感応要略中 21／今昔巻 6-39、法華験記中 49／今昔巻 12-40、法華験記上 18／今昔巻 13-02、法華験記下 89／今昔巻 14-15、法華験記下 111／今昔巻 15-44、日本霊異記上 5／今昔巻 11-23、日本霊異記中 32／今昔巻 20-22、日本往生極楽記 25／今昔巻 15-18

５．対照結果の分類と分析例

5.1 対照結果の分類

３節、４節で説明したような方法で、『今昔物語集』以外の説話集（説話集Ａ）と『今昔物語集』（説話集Ｂ）の共通説話について対照を行ったデータは、次のように分類できる[5]。

説話集Ａ（『今昔物語集』以外）の語

① 同語対応　　例：昔（宇治）　　［昔（今昔）］
② 別語対応　　例：率引（霊異記）　［引く（今昔）］
③ 非対応　　例：為（霊異記）　　［￠（今昔）］

説話集Ｂ（『今昔物語集』）の語

① 同語対応　　例：昔（今昔）　　［昔（宇治）］
② 別語対応　　例：引く（今昔）　　［率引（霊異記）］
③ 非対応　　例：今（今昔）　　［￠（宇治）］

表 2、表 3、表 5、表 6 の事例にはこの分類も書き込んだ。このうち、①同語対応は、文体が異なる説話集で同一の語が使われるものであるので、文体的変異のない語ということになる。例えば、和文説話集から『今昔物語集』を見たとき、「足」「知る」「打つ」などはほとんどすべての箇所で同じ語が対応しており、漢文説話集から『今昔物語集』を見ると、「寺」「見る」「聞く」などがほぼ全例同じ語が対応している。

一方、②別語対応は、別語が選ばれた原因は、文体が異なることにある可能性が高く、②に分類されることが多い語は、文体的変異を研究する際に、特に注目すべき語群であると考えられる。そして、③非対応は、一方の説話集で書かれ他方の説話集で書かれない理由が、文体とは別の要因（説話集の編纂目的、編者の思想など）である場合も多いと思われるが、文体が関与している可能性もあり、②に次いで注目していく必要があると思われる。

5.2　和文説話集との対照結果の分析

上記の分類結果のデータの集計と分析を進めているが、その中間報告として文体的変異の観点から特徴的だと考えられるいくつかの語について、紹介したい。まず、説話集Ａが和文説話集の場合の対照結果の方から取り上げたい。

説話集Ａ（『今昔物語集』以外）の側で、②別語対応が特に多いものに、「給(た)ぶ」「むず」「で」「囲碁」「築地」「此かる」などがある。表 7 は、これらの分類別の件数を示したものだが、「此かる」を除く５語は、『今昔物語集』の同文説話では、ほぼ決まった語（下線の

[5] このような分類は、山元・田中・近藤（2012）でも示した。本稿の A①・B①は、そのときの 2.2、A②は 2.1、A③は 1.0、B②は 2.3、B③は 3.0 に、それぞれ相当する。

語）が対応しており、それぞれ、文体的な対語関係にあるものと考えられる。

表 7　説話集Ａにおいて②別語対応が多い語の例

語	①同語	②別語	③非対応	②の『今昔物語集』での対応語
給(た)ぶ	0	5	0	給ふ 4、奉る 1
むず	1	4	0	むとす 3、なり[推定]1
で	1	3	1	ずして 3
囲碁	0	5	0	碁 5
築地	0	5	0	築垣 4、城 1
此かる	1	5	1	此く 1、此れ 1、然る 1、己等が様なる 1、夜叉の一党 1*

＊は、語単位の対応でなく、複数の語に対応するもの

文体的対語と考えられるものについて、平安時代後期（11 世紀初め）までの和文作品を対象とした『日本語歴史コーパス 平安時代編』（国立国語研究所）[6]を使って頻度調査を行うと、次のように、いずれの対においても、頻度は一方の語に極端に偏っており、平安時代において文体的な特異性を持つ語であったことが裏付けられる。

(1) 給ぶ（21）／給ふ（17868）　(2) むず（21）／むとす（218）
(3) で（1079）／ずして（31）　(4) 囲碁（0）／碁（29）
(5) 築地（9）／築垣（0）

(1)(2)(4)の３対は、平安和文作品では、『今昔物語集』に特徴的な語の方が、頻度が圧倒的に多くなっている。一方、(3)(5)の２対は、和文説話集に特徴的な語の方が、頻度が圧倒的に多くなっている。このような現象は、文体的な特異性の内実が、二つの群で異なっていることを考えさせられ、研究を深めていく必要性が強く感じられるところである。

一方、決まった語が対応していない「此かる」のように、対語をもたない文体的な特徴語が存在していることも判明するが、こうした語をどのように位置付けるべきかについても研究していかなければならない。

説話集Ｂ（『今昔物語集』）の側で②別語対応が多いものには、「間」「碁」「比丘」「古老」「暫く」「美麗」「王宮」などが指摘できる。このうち「碁」は、Ａ②に多い「囲碁」の対応語になっていたものであるが、それ以外はＡ②に多い語とは一致しない。それらをまとめた表 8 によれば、和文説話集の同文説話で特定の語（下線の語）が対応する場合もあれば、多様な語が対応する場合もある。これらの語についても、当時の他の文献での出現状況を参照しながら、文体的変異の内実について研究することが期待されよう。

次に、③非対応が特に多い語がどのようなものか見ていこう。説話集Ａ（『今昔物語集』以外）の側で③非対応になるものをあげると、「此処（ここ）」「度（たび）」「いみじ」「付ける」「候ふ」「え（副詞）」「やがて」「かな（終助詞）」「など（副助詞）」などがある。これらの現象が、文体を理由として特徴語になっているものなのか、別の理由によるものなのかについては、個々の語の事情について研究していく必要がある。同じようにして、説話集Ｂ（『今昔物語集』）の側で③非対応が多いものに注目すると、「各々」「先（さき）」「一つ」「他（た）」「更に」「忽ち」「未だ」「相（あひ）」など、多くの語がリストアップされる。これらについても、このような特徴

[6] https://maro.ninjal.ac.jp/search

が生じる背景や事情を検討していくことが求められるだろう。

表 8　説話集Ｂにおいて②別語対応が多い語の例

語	①同語	②別語	③非対応	②の和文説話集での対応語
間	2	5	2	ほど 3、が 1、に 1
碁	0	6	3	囲碁 5、この事 1
比丘	0	6	6	僧 3、住僧 1、寺僧 1、他僧 1
暫く	1	3	1	暫し 3
美麗	0	6	1	をかしげ 2、めでたし 1、美し 1 、あはれげ 1、玉 1
王宮	0	3	1	内裏 1、御内 1、公卿殿上人*1

＊は語単位の対応でなく、複数の語に対応するもの

5.3　漢文説話集との対照結果の分析

それでは、説話集Ａ（『今昔物語集』以外）が漢文説話集である場合はどうであろうか。漢文説話集の場合、３節で述べたような難しい問題がつきまとうため、データの集計と分類が不十分な段階であるが、②別語対応が特に多いものとしては、「詔（の）る」「沙門」「比丘」「故」などが挙げられる。これらのうちはじめの３語は、表９に見るように、『今昔物語集』の同文説話では、ほぼ決まった語（下線の語）が対応しており、それぞれ、文体的な対語関係にあるものと考えられる。一方、「故」は文体的対語はもたない文体的特徴語と見ることができるだろう。

表 9　説話集Ａにおいて②別語対応が多い語の例（漢文説話集）

語	①同語	②別語	③非対応	②の対応語
詔る	0	4	1	仰す 4
沙門	1	5	1	僧 3、持経者 1、海蓮 1
比丘	2	6	1	僧 5、汝 1
故	1	3	1	が為に 1、依て 1、て 1

説話集Ｂ（『今昔物語集』）の側で②別語対応が多いものに、「后」「僧」「申す」「成る」などが指摘できる。このうち「僧」は、Ａ②に多い「沙門」「比丘」の対応語になっていたものであるが、それ以外はＡ②に多い語とは一致しない。それは、表 10 のように、和文説話集の同文説話では、特定の語やある類の語が対応する場合もあれば、多様な語に対応する場合もある。

表 10　説話集Ｂにおいて②別語対応が多い語の例（漢文説話集）

語	①同語	②別語	③非対応	②の対応語
后	0	4	1	皇后 4
僧	6	9	7	比丘 5、沙門 3、我 1
申す	1	4	1	奏す 3、願ふ 1
成る	0	4	4	生む 1、得 1、作す 1、所役 1

最後に、③非対応が多いものはどういう語だろうか。まず説話集Ａ（『今昔物語集』以外）

の側では、文末の「矣」や副詞「亦」が非常に多いが、これらは、漢文説話に使われていても、『今昔物語集』では排除されるタイプの語であったと考えられる。ほかに、「其れ」「已に」「更に」などが同様のタイプとしてあがってくる。

説話集Ｂ（『今昔物語集』）の側では、「たり」「き」「は」「を」などの助詞・助動詞が多いが、これは、そもそも漢文にはない語である。自立語にも、「間（あひだ）」「給ふ」「伝ふ」など、漢文にはなく『今昔物語集』が独自に使う語は多い。こうした語の性質についても、まずは一語一語について研究し、文体的特徴を持つ事情を究明していくことが望まれよう。

６．おわりに

『今昔物語集』の同文説話を材料にして、平安時代末期の文体的変異を解明しようとする研究は、２節にあげたように従来から盛んであったが、そこに関連説話集とのパラレルコーパスを持ち込むことで、その方面の研究をいっそう活発化させ見通しのよいものにしていく効果が期待できる。それが実現すれば、平安・鎌倉時代における文体的変異に関して、従来は十分に目の行き届いていなかった、より広い範囲の研究につなげていくことができると期待される。そのような研究は、パラレルコーパスの対象にならない、この時代の多くのテキストのコーパス化に対しても有益な知見をもたらすことになると思う。

付記

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）、及び、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、研究代表者：田中牧郎）による成果の一部です。

文献

小木曽智信（2012）『和文系資料を対象とした形態素解析辞書の開発 研究成果報告書』（科研費報告書、http://dl.dropbox.com/u/73297026/report/unidic-EMJ_report2012.pdf）

小峯和明（1997）『今昔物語集の形成と構造　補訂版』笠間書院

近藤泰弘（2012）「日本語通時コーパスの設計について」（『国立国語研究所プロジェクトレビュー』国立国語研究所、3-2、pp.84-92）

今野真二（2009）『文献日本語学』港の人

佐藤武義（1984）『今昔物語集の語彙と語法』明治書院

築島裕（1963）『平安時代の漢文訓読語につきての研究』東京大学出版会

藤井俊博（2003）『今昔物語集の表現形成』和泉書院

舩城俊太郎（2011）『院政時代文章様式論考』勉誠出版

峰岸明（1986）『平安時代古記録の国語学的研究』東京大学出版会

山口仲美（1984）『平安文学の文体の研究』明治書院

山元啓史・田中牧郎・近藤泰弘（2012）「通時コーパスと言語空間論」（『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、pp.241-248）

「日本語歴史コーパス 平安時代編」先行公開版について

小木曽智信（国立国語研究所言語資源研究系）†
須永哲矢（国立国語研究所コーパス開発センター）
冨士池優美（国立国語研究所コーパス開発センター）
中村壮範（マンパワージャパン株式会社）
田中牧郎（国立国語研究所言語資源研究系）
近藤泰弘（青山学院大学文学部／国立国語研究所言語資源研究系）

On the Public Beta Release of *the Heian Period Series of the Corpus of Historical Japanese*

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Tetsuya Sunaga (National Institute for Japanese Language and Linguistics)
Yumi Fujiike (National Institute for Japanese Language and Linguistics)
Takenori Nakamura (Manpower Japan Co., Ltd.)
Makiro Tanaka (National Institute for Japanese Language and Linguistics)
Yasuhiro Kondo (Aoyama Gakuin University / National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所では「通時コーパスの設計」プロジェクトが中心となって「日本語歴史コーパス[1]」(Corpus of Historical Japanese, CHJ) の開発準備を進めてきた（近藤 2012）。今回、このうち平安時代の仮名文学作品からなる「平安時代編」のデータ整備が進んだことから、これを先行公開版として一般公開を行うこととした。「現代日本語書き言葉均衡コーパス」（BCCWJ）の公開にも用いられているウェブインターフェイス「中納言」（小木曽ほか 2011）を利用しての公開となる。

本発表では「CHJ 平安時代編」の先行公開版の内容について説明し、「CHJ 中納言」のデモンストレーションを行う。

２．「日本語歴史コーパス 平安時代編」先行公開版の概要

「CHJ 平安時代編」は、日本の代表的な古典文学作品である、平安時代の仮名文学作品をコーパス化したものである。先行公開版では、次の 10 作品のデータが利用可能である。本文はすべて、許諾を得て小学館「新編日本古典文学全集」（新編全集）を利用している[2]。

古今和歌集、土佐日記、竹取物語、伊勢物語、落窪物語、大和物語、枕草子、源氏物語、紫式部日記、和泉式部日記

収録した本文データには「中古和文 UniDic」（小木曽ほか 2012，Ogiso et al. 2012）と MeCab を用いて形態素解析を施し、その解析結果に対して人手による修正を行った。これにより、出現するすべての語に読み・品詞・活用型・活用形・語種等の形態論情報（短単位）が付与されている。

さらに、新編全集の情報を利用して本文に「本文種別」と呼ぶ情報を付与し、当該箇所が地の文なのか会話文なのか、あるいは和歌や手紙なのかといった区別がなされている。『源氏物語』では話者も表示される。

† togiso@ninjal.ac.jp

1 これまで暫定的に「通時コーパス」と呼称されていたものの正式名称。

2 コーパス化の対象は原文のみで、現代語訳等は含まない。

テキスト量

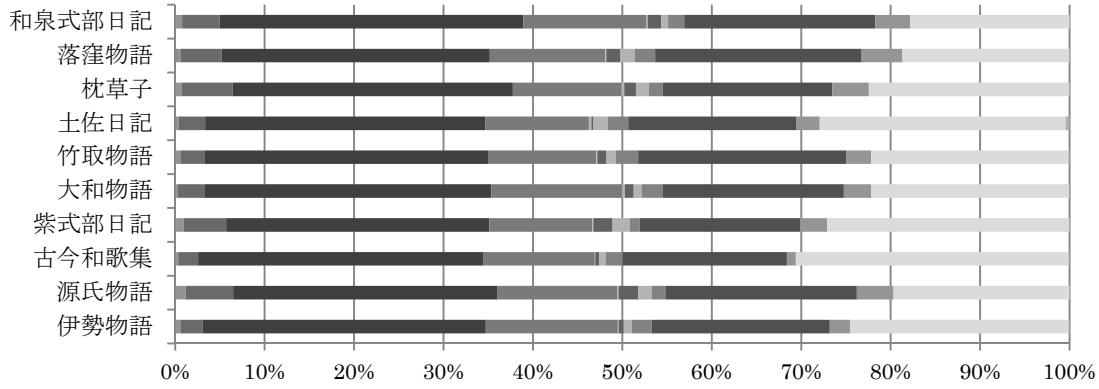
「CHJ 平安時代編」のテキストの量は、表 1 に示す通りである。全体で約 79 万語、うち 65%に近い 51 万語を源氏物語が占めている。

表 1　作品別の語数（短単位，記号を含む）

作品名	語数
伊勢物語	15894
古今和歌集	32286
和泉式部日記	12630
土佐日記	8129
大和物語	26740
枕草子	79851
源氏物語	510572
竹取物語	12583
紫式部日記	20710
落窪物語	68561
総計	**787956**

品詞・語種構成

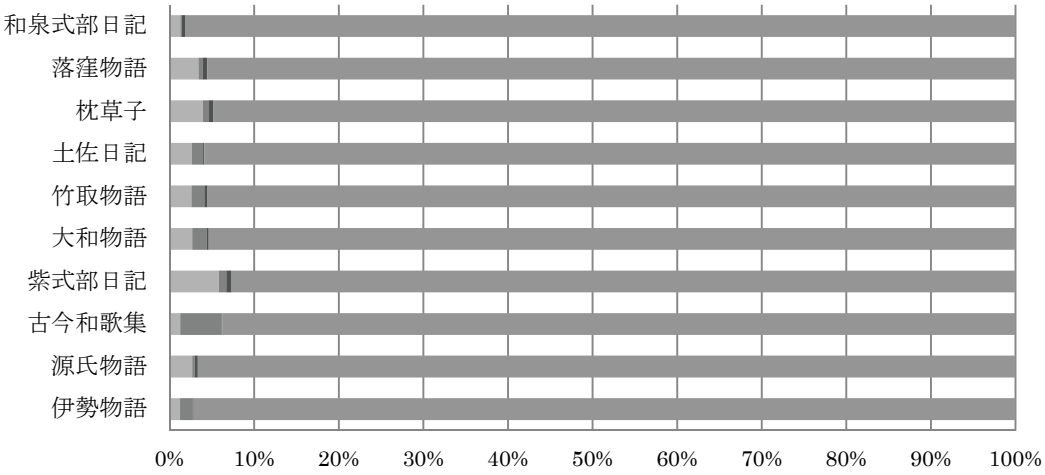
コーパスに付与された短単位の形態論情報を元に、作品ごとに品詞別の語数を集計したものが図 1 である。品詞の認定基準は『中古和文 UniDic 短単位規程集』（小椋・須永 2012）によっている。なお、以下では語数に記号を含まない。



	伊勢物語	源氏物語	古今和歌集	紫式部日記	大和物語	竹取物語	土佐日記	枕草子	落窪物語	和泉式部日記
■感動	7	334	26	15	8	7	0	51	91	16
■形状	72	4788	83	144	62	51	26	437	255	68
■形容	338	23150	688	815	692	280	195	3687	2454	452
■助詞	4325	128841	9867	5029	7327	3210	2068	20363	16175	3651
■助動	2017	58447	3875	1969	3344	1229	767	7903	6961	1476
■接続	13	449	12	19	72	12	18	184	70	21
■接頭	77	9682	142	368	224	97	16	848	862	162
■接尾	120	6675	231	332	204	112	105	950	871	84
■代名	309	6578	579	184	531	254	151	999	1209	194
■動詞	2718	93308	5687	3072	4641	2357	1239	12325	12465	2300
■副詞	308	17800	312	510	699	286	176	2657	2457	419
■名詞	3341	85827	9456	4633	5062	2238	1815	14571	10073	1916
■連体	12	14	26	3	12	11	30	9	9	0

図 1　作品別品詞構成

同様に語種別の集計を行ったものが図 2 である。どの作品でも大部分が和語であり、全体では 96%を占める。古今集は歌人の名前・官職などを含むため、固有名や漢語の割合が高くなっている。ごくわずかに現れる外来語はサンスクリット語由来の仏教語である。



	伊勢物語	源氏物語	古今和歌集	紫式部日記	大和物語	竹取物語	土佐日記	枕草子	落窪物語	和泉式部日記
■外	0	105	2	5	3	3	0	17	5	0
■漢	169	11539	388	990	607	259	174	2546	1837	136
■固	201	1205	1516	151	398	155	88	467	258	16
■混	19	1686	20	105	51	39	9	347	314	44
■和	13268	421351	29058	15841	21819	9688	6335	61607	51537	10563

図 2　作品別語種構成

形態論情報の特徴

CHJ に付与される形態論情報は、電子化辞書 UniDic（伝ほか 2007）の設計にもとづくものである。UniDic は、短単位と呼ばれる厳密な規定によって単語の区切り方が定められており、揺れが少ない斉一な単位による解析が可能になっている（小椋ほか 2011）。また、図 3 に示すように語彙素・語形・書字形・発音形という見出し語の階層構造を持っており、利用者が必要に応じて、見出し語のレベルを選択して利用することができる。図 4（次頁）は具体的な見出し語（例：何処［イズコ］）の例である。

図 3　見出し語の階層構造

「語彙素」は、異語形や異表記をまとめ上げた辞書見出し（lemma）に相当するもので、「語形」はそのうち異語形を区別したもの、「書字形」は異表記を区別したものである。「発音形」は発音を示すものであるが、CHJ においては現代における読み方を参考までに示したものに過ぎない。利用者は、表記そのものに関心があるのであれば書字形を、語形の差異に関心があるのであれば語形を、辞書見出しのレベルでまとめ上げたいのであれば語彙素を利用すればよい。

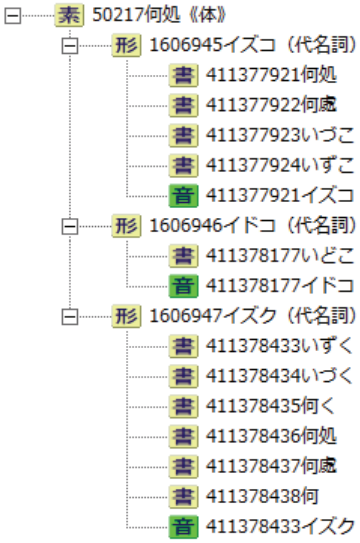


図 4　見出し語の階層構造の例（何処［イズコ］）

CHJ で利用している中古和文 UniDic の短単位は、原則として現代語と同様の基準によっており、相互に比較することができるように配慮したものである。ただし、語の歴史的変化や中古語の実態を踏まえ、時代別に異なった扱いをしている語も少なくない。たとえば、現代語では連体詞とされる「この」「その」が、中古語では代名詞「こ」「そ」と格助詞「の」に分けて数えられている。「CHJ 中納言」を用いて中古語の検索をする場合には、この短単位の規定について理解をしておく必要がある。

先行公開版データの制限

CHJ では、BCCWJ と同様、短単位だけでなく長単位の情報も付与する計画である。しかし、先行公開版で公開するデータは短単位のみである。平安時代編の完成版で、長単位のデータも公開する予定である。

また、CHJ 平安時代編が基づいている『中古和文 UniDic 短単位規程集』には完全でない部分が残されている。たとえば、複合動詞を一語と認めるか分割するかという認定基準はその例である。そのため、先行公開版では複合動詞の認定に揺れがあるなどの問題が残っている。これも完成版では統一的な基準の下に修正される予定である。

３．「日本語歴史コーパス」中納言

CHJ の公開は現在のところ、ウェブ版のコンコーダンサー「中納言」（図 5）のみで行っている。「CHJ 中納言」は、CHJ むけに若干の修正を行っているが、基本的に BCCWJ で利用されている「中納言」と同じものである。書面による申込み手続きを経ることで無償で利用できる（手続きは「日本語歴史コーパス」ホームページを参照）。

データには形態論情報が付与されているため、表層の文字列だけでなく、形態論情報を利用することで高度な検索条件の指定を行うことができる。たとえば、語彙素「読む」（終止形）を指定することで「読ま」「読み」「読む」「読め」といった各活用形を一括で検索することが可能である。また、先述の UniDic の見出し語の階層構造により、見出し語を語彙素で指定すれば、その異表記を一括検索することができる。したがって、漢字表記と仮名表記の違い、異体字や送り仮名の揺れなどを一々意識することなく検索できる。

また、たとえば品詞情報を使って「形容詞すべて」のように大きな語群を検索対象とすることもできる。形態論情報を組み合わせて、たとえば「漢語名詞」「形容詞の連体形」などの詳細な条件で検索を行うことも可能である。

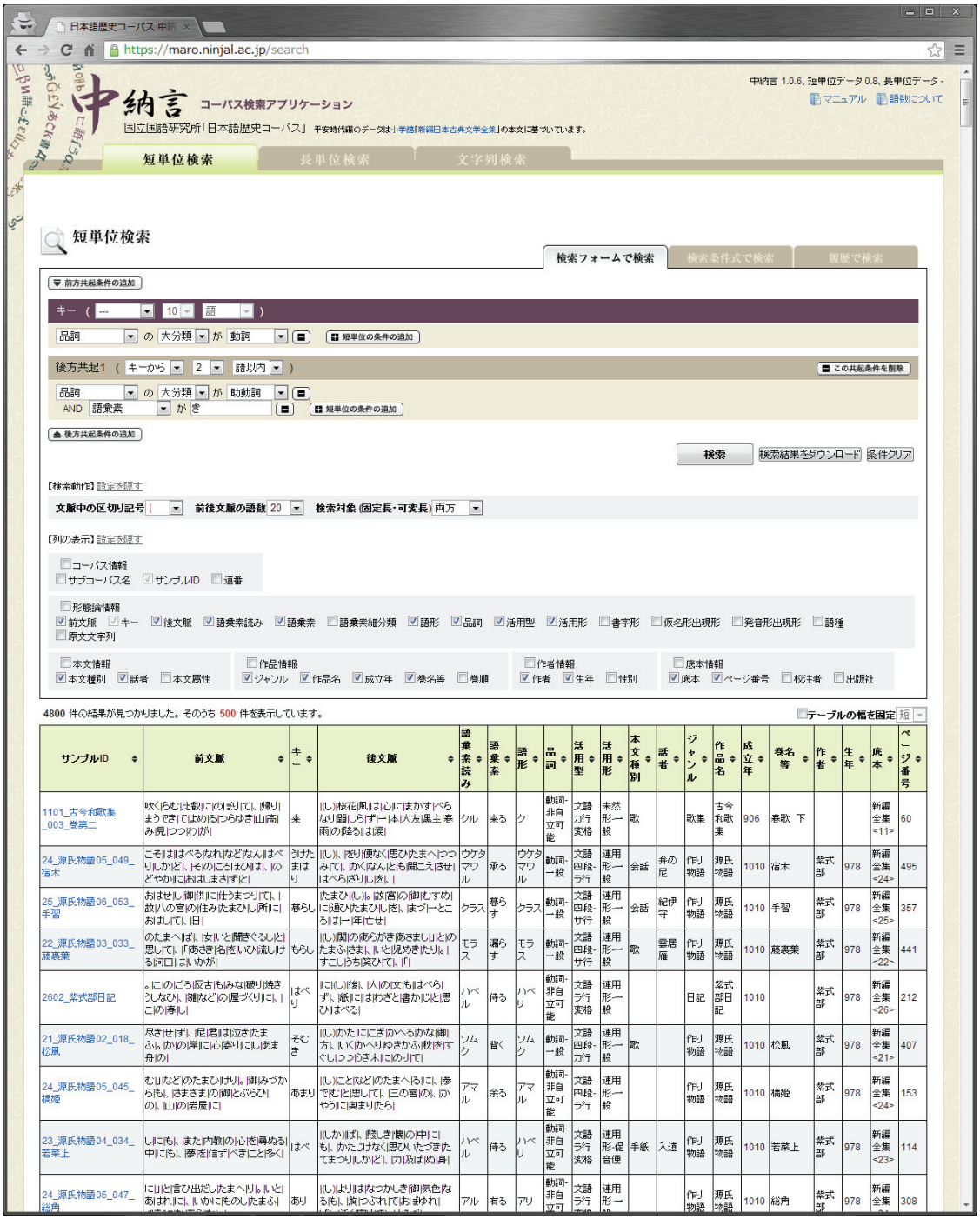


図 5　日本語歴史コーパス「中納言」検索実行画面

さらに、複数の語（最大 10 語）を組み合わせた検索も行うことができる。これにより、「特定の形容詞の連体形の後に来る名詞」であるとか、「特定の動詞に続く助動詞」、「特定の動詞の前方 5 語以内に来る“名詞＋を”」といったような、従来の索引では不可能であった検索が可能になっている。

形態論情報を使った検索以外に、「文字列検索」で表層の文字列による検索を行うこともできる。この場合にも、検索結果は形態論情報付きで表示されるため、調査したい語にどのような形態論情報が付与されているか分からない場合には、いったん文字列検索を行うことで形態論情報を確認することができる。

検索結果の項目

検索結果には、表 2 に示す項目が表示可能である。デフォルト表示が「非表示」のものは画面上のチュエックボックスをオンにすることで表示されるようになる。

「コーパス情報」は、検索結果のコーパス中の位置を示す情報である。サンプル ID と連番とで短単位の位置を一意に指定することができる。

「形態論情報」は、当該箇所の KWIC と、キーに付与されている形態論情報からなる。「キー（書字形出現形）」が実際に出現した表層形（活用変化後の形）であるのに対し「書字形」は終止形の形である。形態論情報中の「～出現形」はすべて活用変化後の形であることを示す。

表 2 検索結果表示項目

分類		順番	列名	デフォルト表示
コーパス情報		1	サブコーパス名	非表示
		2	サンプル ID	表示
		3	連番	非表示
形態論情報	KWIC	4	前文脈	表示
		5	キー（書字形出現形）	表示
		6	後文脈	表示
		7	語彙素読み	表示
		8	語彙素	表示
		9	語彙素細分類	非表示
		10	語形	表示
		11	品詞	表示
		12	活用型	表示
		13	活用形	表示
		14	書字形	非表示
		15	仮名形出現形	非表示
		16	発音形出現形	非表示
		17	語種	非表示
		18	原文文字列	非表示
本文情報		19	本文種別	表示
		20	話者	表示
		21	本文属性	非表示
作品情報		22	ジャンル	表示
		23	作品名	表示
		24	成立年	表示
		25	巻名等	表示
		26	巻順	非表示
作者情報		27	作者	表示
		28	生年	表示
		29	性別	非表示
底本情報		30	底本	表示
		31	ページ番号	表示
		32	校注者	非表示
		33	出版社	非表示

「本文情報」の「本文種別」は「会話」「手紙」「歌」「詞書」等の別である。「話者」は会話の話者表示だが、新編全集で明示されているものだけが出力され、作品によっては情報がない。「本文属性」は和歌である場合に歌番号が出力されている。

「作品情報」は当該の作品の基本的な書誌情報である。「ジャンル」には平安時代編では作り物語・日記・随筆・歌集がある。「成立年」は正確な年が不明のものは有力な説に従い、おおよその年代を記入している。「巻名等」は研究に必要と考えられる範囲で新編全集にもとづいて巻名や章段のタイトル、部立てなどを記入している。

「作者情報」は当該の作品の作者の情報である。詳細が不明のものは分かる範囲で記入している。『古今和歌集』については仮名序以外には作者情報を出力していない。

「底本情報」は CHJ 平安時代編が依拠した新編全集の情報である。「底本」は当該作品が収録された新編全集の巻数、「ページ番号」は当該箇所が現れるページ数を示す。これにより、ヒットした用例について書籍の新編全集を開いて当該箇所を確認することができる。CHJ には現代語訳や注は含まれていないため、こうした情報を確認するためには新編全集本体を参照する必要がある。

これらの情報を含む検索結果は表形式でダウンロードすることができるため、これを表計算ソフトに読み込むことで、自由に集計を行うことができる。特にピボットテーブルと呼ばれる機能を用いることで、クロス集計を自在に行うことが可能である。ダウンロードファイルには、常に表 2 の全ての項目が含まれており、さらに最終列（34 列目）に「反転前文脈」が出力される。この列は前文脈を使ってソートを行うためのもので、前文脈の文字列の並びを逆転させキーに近い文字から順に並べたものである。

先行公開版インターフェイスの制限

現在の「CHJ 中納言」では、検索対象を指定することができず、常にコーパス全体を検索することになる。したがって作品別・ジャンル別に用例数などを確認するためには、いったん検索結果をダウンロードして集計を行う必要がある。

この問題は、2013 年 3 月に予定している「中納言」のアップデートで改善される予定である。このアップデートにより、作品別・ジャンル別などの検索対象指定が可能になるほか、検索条件指定の方法などさまざまな機能が改善される予定である。

４．今後の計画とまとめ

CHJ 平安時代編の完成版は、2013 年度中の公開を予定している。完成版では、上述した制限をなくし、全ての作品について長単位を付与するほか、『更級日記』『讃岐典侍日記』等の作品を追加する予定である。

先行公開版には制限があるものの、本コーパスにより、これまでの古典語の研究手法では不可能だった検索や集計が可能になった。本コーパスが広く利用され、新しい研究成果につながることを期待したい。また、これを機にコーパス日本語学の裾野が歴史的研究の分野にまで広がり、研究がより盛んになることに期待したい。

文　献

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

小木曽智信・中村壮範・鈴木泰山・八木豊・山崎誠・前川喜久雄（2011）「コーパス検索システム「中納言」デモンストレーション」『日本語コーパス完成記念講演会予稿集』pp.43-46

小木曽智信ほか（2012）『和文系資料を対象とした形態素解析辞書の開発』科研費 基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」（課題番号 21520492）研究成果

報告書（中古和文 UniDic ホームページからダウンロード可能）
小椋秀樹・須永哲矢（2012）『中古和文 UniDic 短単位規程集』科研費 基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」(課題番号 21520492) 研究成果報告書 2 （中古和文 UniDic ホームページからダウンロード可能）
Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den and Yuji Matsumoto. (2012) UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp.911-915. Istanbul, May 2012.（http://www.lrec-conf.org/proceedings/lrec2012/pdf/906_Paper.pdf からダウンロード可能）
近藤泰弘（2012）「日本語通時コーパスの設計」『NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム 通時コーパスと日本語史研究予稿集』pp.1-10

関連 URL

日本語歴史コーパス「中納言」 http://maro.ninjal.ac.jp/
日本語歴史コーパスホームページ（国立国語研究所コーパス開発センター）
http://www.ninjal.ac.jp/corpus_center/chj
NINJAL 通時コーパスプロジェクト ホームページ http://www.historicalcorpus.jp/
中古和文 UniDic http://www2.ninjal.ac.jp/lrc/index.php?UniDic
MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.googlecode.com

ポスター発表（2） Aグループ

3月1日（金） 13:00～14:00

医学用語の選択に見られる特徴

金子 周司（京都大学大学院薬学研究科）†

Characteristics of the Choice of Japanese Medical Words in the Corpora of Scientific and Clinical Documents

Shuji Kaneko (Kyoto University Graduate School of Pharmaceutical Sciences)

１．はじめに

医療や生命科学の急激な進歩は、莫大な数の専門用語を新たに生み出している。筆者は医学系学生や研究者が英語の専門用語を学習・活用するための電子辞書の開発に 20 年来取り組んできたが、日本語については訳語として位置づけ、あまりその特性について深く考察してこなかった（金子 2006）。しかし今後、医療や教育の電子化がますます進展し、自然言語処理が医療サポートや知識発見に応用されていくことを考えると、医学用語の日本語表記について理解を深めることが必要と思われる（金子、大武 2010）。

本研究ではどのようにして和文で医学用語が選択されているかを少しでも知るために、医学文献や医薬品解説書を元にしたコーパスを構築し、専門用語を抽出した上で異表記を収集、解析した。我が国で医学用語をどのように表記するかについては、決まり事や法則があるわけではなく、研究者が独自に編み出したり、すでにある文書を参考に用語が選ばれたりしている。コーパスの解析結果からも、日本語では漢字、カタカナ、ひらがな、英語綴りなどを混在して用いることができるため、異表記が非常に多いという特徴があることがわかってきた。編集者や許認可者による修正を経た後の文書においても、医学用語の多様性は維持されている。いくつかの例を紹介して考察してみたい。

２．コーパスの概要

筆者は、出版社である株式会社羊土社の協力を得て、医学研究者が書く総説の全文コーパスを以前に構築した（金子 2006）。本研究ではそれをさらに拡張し、1996 年から 2005 年の 10 年間にわたって『実験医学』誌に発表された全総説をテキスト化することで実験医学コーパス（37.3Mbyte）とした。本コーパスについては、用語の解析目的でのみ使用できる許諾契約を締結した。また、財団法人日本医薬情報センター（通称 JAPIC）が有料で販売している医療用医薬品全 13,000 種の添付文書情報（2008 年版）について、解析目的での使用許諾を得てテキスト化し JAPIC コーパス（49.6Mbyte）とした。

解析としては、ライフサイエンス辞書に収録している 157,347 語の日本語をクエリーとして、コーパス中で一致する文字列の頻度を Perl スクリプトにより求めた。これら 2 種類のコーパスの概要を表 1 に示す。JAPIC コーパスのほうがサイズ的には大きいが医薬品固有名が多いこともあり、以下においては同規模のコーパスとして頻度（語数）の比較を行う。

表 1　本研究で用いたコーパスの概要

	実験医学	JAPIC
文字数	20,235,504	25,247,795
読点数	271,158	418,839
句点数	504,847	687,669
「など」頻度	22,897	26,997
「血管」頻度	13,146	12,345
「高い」頻度	4,265	4,087

† skaneko@pharm.kyoto-u.ac.jp

３．各コーパスの特徴

表２はそれぞれのコーパスで求めた頻度のうち、一方での値が他方の 100 倍以上であった特徴語を示している。

実験医学コーパスにおいては、最先端の成果を研究者自らが執筆していることもあり、「遺伝子」「タンパク質」「配列」といった生体分子の名称や物性を表す語が多く、「シグナル」や「ドメイン」のように専門家の間でのみ通用する jargon とも考えられるカタカナ語が多用されている点が特徴的である。

一方、JAPIC コーパスで最も頻度が高いのは「本剤」であるが、これは医薬品添付文書における主語として多用されるためである。その他にも「経口投与」「血中濃度」など添付文書における解説として専門家に注目されるべき特徴語が見られる。医薬品は同一作用機序をもつ類似薬が多いこともあり、それらの添付文書間では記述も似ている傾向がある。このことは実験医学コーパスからは 50,257 種類の語が抽出されたのに対して、JAPIC コーパスからは 36,449 語しか抽出されなかった解析結果に反映されている。

図１には各コーパスを構成している文字種の割合を示した。いずれのコーパスにおいても英数字とカタカナが 3〜4 割、漢字の割合も 3〜4 割を占めており、きわめて専門用語に満ちた文書であることがわかる。

表２　コーパスの特徴語（頻度データ）

語	実験医学	JAPIC
遺伝子	45, 676	294
タンパク質	31, 544	53
シグナル	18, 755	14
ドメイン	14, 474	7
配列	10, 640	16
本剤	35	99, 945
経口投与	90	21, 550
血中濃度	135	19, 973
既往症	2	14, 879
妊婦	11	14, 007

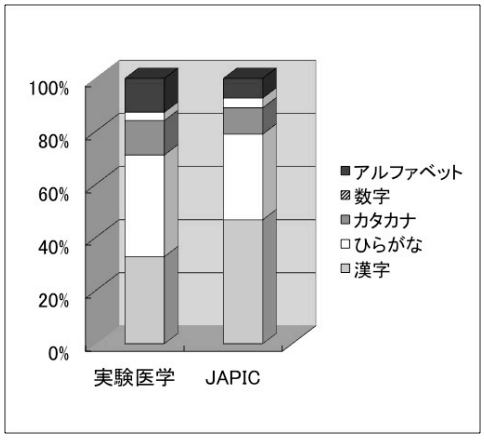


図１　コーパスを構成する文字種

４．用語の選択

以上のように構築したコーパスを用いて専門用語の頻度を見ていくと、いくつかの問題点に気づく。それらをデータと共に説明していく。

４.１.「protein」使い分けの実情

Protein とはアミノ酸がペプチド結合によって連なり、そのアミノ酸の性質や場の状況によって特異的な立体構造をとる最も重要な生体構成分子種である。英語においては protein という語の他には比較的短い鎖を指す peptide（ペプチド）という語があるが、protein に異表記は存在しない。しかしながら日本語においては protein が卵白に多く含まれることに起源をもつ「蛋白」から「蛋白質」という語を生み出し、日本医学会は用語集で「蛋白質」を推奨している。しかし文部科学省は学術用語として「タンパク質」という表記を標準としており、新聞や報道等においては「たんぱく質」という表記が多く採用されている。それぞれから「質」を除去した表記も多く用いられ、さらには「プロテイン」と表記すれば一般社会においてはサプリメントとして用いられる補助栄養食品を指すかのように微妙に使い分けられている。

ひとつの英語に対して複数のカタカナ外来語が生じることは、例えば vector が数学の世界では「ベクトル」、分子生物学や情報学では「ベクター」と書かれるように、ドイツ語由来と英語の発音に近い複数のカタカナ語が時代を異にして生じるため珍しいことではない。しかし医学用語の場合には、ひらがなや漢字での異表記まで加わって非常に多様になっている。今回、構築したコーパスにおいて調べてみた結果を表３に示すが、それぞれ編集者や許認可者の手が入った文書であるにもかかわらず、多様な表記が検出された。

表３　「protein」の日本語表記における選択

語	実験医学	JAPIC
たんぱく質	0	14
タンパク質	***31, 544***	53
タンパク	4, 330	174
蛋白質	2, 693	489
蛋白	1, 067	***4, 337***
プロテイン	801	77

実験医学コーパスで「タンパク質」が多く、JAPIC コーパスで「蛋白」が多いのは基礎医学と医療という分野間の差異であると考えられるが、より詳細に見ていくと用語はさらに精密に選択されている（表４）。

表４　「protein」の連接語に応じた選択

コーパス	連接語	タンパク質	タンパク	蛋白質	蛋白	プロテイン
実験医学	結合～	***1, 327***	199	139	29	0
	プリオン～	***53***	9	2	***54***	0
	～分解	***676***	128	18	11	0
	～キナーゼ	75	42	3	0	***489***
JAPIC	～結合	0	43	0	***1, 693***	0
	糖～	2	20	***114***	76	0

実験医学コーパスにおいて、前に「結合」や後に「分解」が連接する場合はいずれも「タンパク質」が多く用いられていた。「プリオン」との連接においては「プリオン蛋白」という表記が特異的に高い傾向が見られた。このことはカタカナ同士が連接した場合に元の語の境界が分かりづらくなることを避けている表現なのかもしれない。しかし、タンパク質をリン酸化する酵素である protein kinase を表す際には、そのままカタカナ語として「プロテインキナーゼ」が最も頻出した。

JAPIC コーパスにおいては、「蛋白結合」のように「蛋白」という表記が全般的に好んで

用いられていたが、この用語はいずれの省庁や団体も推奨している表記ではない。一部においては「糖蛋白質」のように「質」をつけた表記が集中しているケースも見られたが、これは他の類似薬で用いられた文書をそのまま流用して使っているために複製増幅効果が現れたものと推察される。

４.２．薬物のカテゴリーを表す名称

先行研究において筆者は、英語圏で発達した医学や生命科学が日本へ「輸入」された際に必ずしも専門用語を直訳するのではなく、日本人が理解しやすいように「意訳」を行ってきた実例をいくつか提示した（金子 2006)。この結果は PubMed で公開されている英語の医学文献抄録と実験医学コーパスの前身である日本語テキストを比較解析して得られたものであったが、今回、JAPIC コーパスを新たに解析することによって、これらの指摘が準公的な医薬品添付文書においても適用できることが明らかになってきた。

その一例として、表５は腫瘍の増殖に対して抑制的に作用するカテゴリーの薬物に与えられる一般的な名称を調査した結果である。この結果から、いずれのコーパスにおいても多様な表記が混在していることがわかる。専門的には「癌≠悪性腫瘍」であり「癌＝上皮細胞の（つまり一部の）悪性腫瘍」であることを加味すると、このように階層の異なる概念を同一視している現状は好ましいとは言えない。

表５　腫瘍増殖を抑制する薬物の名称

語	実験医学	JAPIC
抗癌薬	19	0
抗癌剤	***763***	32
抗がん薬	0	9
抗がん剤	3	6
制癌剤	37	3
抗腫瘍薬	13	0
抗腫瘍剤	12	52
抗悪性腫瘍薬	10	13
抗悪性腫瘍剤	1	***741***
悪性腫瘍治療薬	1	0

５．まとめ

医学用語は長らく標準化の方向性で議論されていた。しかしながら、本研究で編集者や許認可者の修正を経た文書コーパスを解析した結果、コントロールされた状況においても医学用語の多様性は失われていないことが明らかになった。実際に現場で作成される文書（例えば電子カルテや学会抄録など）はさらに多様で混沌としているであろうことは容易に想像できる。今後、医療文書の電子化などによって情報の利活用を目指す場合、このように多様な異表記に耐えうる（かつ英語表記や略記にも対応した）頑強なシソーラスを早急に整備することが必要と思われる。

文　献

金子周司 (2006)「ライフサイエンス辞書とは」情報管理, 49:1, pp.24-35.
金子周司、大武 博 (2010)「ライフサイエンス辞書からクリニカルインフォマティクスへ」情報管理, 53:9, pp.473-479.

関連 URL

ライフサイエンス辞書プロジェクト　http://lsd.pharm.kyoto-u.ac.jp/

日本語教育用の形容詞の語彙リストと難易度レベル

スルダノヴィッチ・イレーナ（国立国語研究所日本語教育研究・情報センター／
リュブリャーナ大学文学部）†
李在鎬（筑波大学人文社会系）

Vocabulary List of Adjectives and Levels of Difficulty for Japanese Language Education

Irena Srdanović (National Institute for Japanese Language and Linguistics/
University of Ljubljana)
Lee Jae-Ho (University of Tsukuba)

１．はじめに

大規模コーパスの構築と共に、コーパスに現れる語彙の把握ができるようになり、バランスが取れたコーパスほど、抽出された高・中・低頻度の語彙が実際に利用される語彙の実情を示す傾向が見られる。どのコーパスでもそれぞれの独自の特徴があり、その特徴は語彙分布にもあるが、コーパスが大ければ大きい、また均衡が取れていればそれだけ、語彙の分布に偏りが少なく、得られた語彙データの信頼性が高くなる。大規模コーパスにおいてサブコーパス別、いわゆるジャンル別のデータも得られるようになり、分散度（dispersity）による語彙の特徴が取り出せるようになってきた。第二言語教育においてもこのような語彙リストがよく使われるようになり、近年複数のリソースを利用して作成されている。[1]Nation（2001）によると、英語の高頻度の 2000 基本語彙がテキストの内容 70%～80%をカバーするため、学習者にまずその語彙を教えるべきという指摘がある。

従来の国語研究および日本語研究において、語彙リストの研究は様々なものあり、語彙を確定するために、語彙頻度調査の実施、専門家の判定、編者の判定、児童・生徒の理解度の調査、成人の獲得語数の調査、語の親密度の調査などの方法が用いられてきた。また、『教育基本語彙の基本的研究』のような複数の語彙リストがすでにデータベース化されている。そのうち日本語学習者を対象にしたリストの例としては、「日本語教育基本語彙データベース」（教育基本語彙の基本的研究―増補改訂版 2008）、「日本語能力試験出題基準」（国際交流基金・日本国際教育協会 1994）があげられる。大規模な現代日本語書き言葉均衡コーパス（以下 BCCWJ）などのコーパス開発と共に、コーパスを基にした日本語教育向けの語彙リストの作成が始まった。その例は、近年作成された「日本語を勉強する人のための語彙データベース」（松下 2011）および「日本語教育語彙表」（砂川 2012、李・砂川 2012）である。

日本語教育用の語彙リストはいくつかあるが、その作成方法、基にした資料の特徴・現代性などに違いがあり、どの程度収録された語彙が一致しているかについては必ずしも明らかではない。饗場（2011）が 3 種の語彙リストを調べた結果、リストごとに非共通語が多くあると明らかにした。例えば、形容詞・形容動詞を取り上げると、各語彙における共通語彙の割合は 54.8%、59.2%、90.5%である。スルダノヴィッチ（2012）がリストごと、コーパスごとに形容詞の語数を比較した結果、差異が多く見られることが分かった。本論文の目的は、形容詞を対象にした既存の日本語の語彙リストを検討し、その語彙リストにあ

† irena.srdanovic@ff.uni-lj.si

1　例えば、日本人英語学習者のための語彙リスト JACET8000（大学英語教育学会基本語改定委員会(編) 2003）は、学習者が遭遇しやすいサブコーパスと BNC コーパス頻度を対数尤度比で比較し、作成されたものである。

る形容詞と大規模コーパスから取り出せる形容詞を比較しつつ、語彙リストに把握されていない項目を検討することにある。そこから得られたデータを基にして、今後の課題としては、新たな日本語学習者用形容詞の語彙リストおよび形容詞と他の単位との組み合わせの記述を目指すことである。

２．日本語教育用の語彙リスト

以下に取り上げる語彙リストは、本研究の対象にして、それぞれのリストに現れる形容詞を検討する。

２．１『日本語能力試験出題基準』の旧語彙リスト

『日本語能力試験出題基準』（1994）の旧語彙リスト（以下「旧 JLPT 語彙リスト」）はテスト作成のために作られたものであり、教育目標のために作成されたリストではないが、日本語教育において幅広く利用されている。語彙難易度は 4 段階に分かれ、下位級の 4 級から上位級の 1 級までである。旧 JLPT 語彙リストは、作られてから 30 年以上経過しているため、語彙の変化に対応していない、カタカナ語や擬音語・擬態語などの語彙が少ないという問題点がある（李・砂川 2012）。なお，2010 年から実施されるようになった新しい日本語能力試験のために作られた「語彙リスト」は 5 段階の難易度に分けられているが，テスト運用上の理由から非公開になっている。

２．２　「日本語教育基本語彙データベース」、「教育基本語彙データベース」

「日本語教育基本語彙データベース」（以下「国研日本語語彙 DB」）は、「国語研教育基本語彙データベース」に登録した 6103 語、国立国語研究所『日本語教育基本語彙七種比較対照表』の 6195 語などの 6 種の教育基本語彙リストのデータをデータベース化したものである。データベース化された 6 種のリストは様々な方法で集められた語彙データであり、総数は 11826 語である。その詳細は国立国語研究所報告 127『教育基本語彙の基本的研究』（2009、 563 ページ）において確認することができる。

同じく国立国語研究所報告 127 に搭載されている「教育基本語彙データベース」（以下「国研国語語彙 DB」）は、7 種の教育語彙を利用したデータベースで、主に国語教育のデータをカバーしている。語数は、27234 である。データベースは小学生低学年・高学年、中学生の理解度を測定したデータに基づいて語彙難易度の 3 段階に分けている（1 は最も低い難易度）。

２．３　「日本語を勉強する人のための語彙データベース」

「日本語を勉強する人のための語彙データベース」（以下「TM 語彙リスト」）は、『現代日本語書き言葉均衡コーパス』（BCCWJ）モニター公開データ（2009 年度版）の書籍および「Yahoo 知恵袋」（約 3300 万語）を使って、松下（2010、2011）が作成した語彙リストであり、Nation（2001）の英語学習のために提案された語彙リスト作成の枠組みに基づいている。その特徴は、コーパス頻度およびサブコーパスごとの語彙分布を基にして、語彙を「一般用」、「留学生用」に分けたデータである。一般用のデータは基本 2500 語を含み、総合数は 20326 語である。留学生用のデータは 3・4 種の分野でよく使われる単語であり、科学分野別の特徴のあるデータも掲載されている（20312 語）。語彙レベルで一般人の生活を中心に考えた重要度のランクおよび Basic、Inter、Adv、H-Adv、S-Adv 五つの語彙ランクがある。語彙の見出し語には UniDic 辞書の短単位の「語彙素」を使っている。リストには旧 JLPT 語彙リストの語彙難易度、語種、品詞などの情報が含まれている。

２．４　「日本語教育語彙表」

「日本語教育語彙表」は、学習者向け辞書開発の基礎資料として開発されたもので、リストの総合数は 18010 語である。独自に開発された日本語教科書 100 冊の「日本語教科書

コーパス」およびBCCWJの2009年度版の公開データを利用し、見出し語を決定している。また見出し語に対して，日本語教育歴10年以上の教師5名が語彙の難易度を判定し、統計的に調整したデータである。見出し語にはUniDicに基づく短単位と単語N-gramによる複合語が入っている。語彙の難易度は、初級前半、初級後半、中級前半、中級後半、上級前半、上級後半の6段階に分かれている。

２．５　その他

上述した日本語教育用の語リスト以外に、話題別語彙表などがある（山内（編）2008、橋本・山内2008）。また、近年Can-doタスクに基づいた語彙表の作成が行われている。現在の段階でタスク・話題に関するデータは少ない。課題遂行能力に基づくコミュニケーションのための日本語教育のためには、山内(編)(2008)のような試みは今後も加速化されるべきであろう。

３．コーパスから取り出せる語彙リスト

上に取り上げた日本語教育語彙表とTM語彙リストは、コーパスから取り出した語彙に基づいて作成されたデータであるが、両方のリストはBCCWJの全体版が公開されていないときに作られたものである。本研究では、BCCWJの全体コーパスおよび大規模なウェブコーパスから取り出した語彙頻度リストを利用し、既存のデータと比較する。

BCCWJは総語数1億語の大規模コーパスで、次のサブコーパスで構成されている：出版書籍（PB）、出版新聞（PN）、出版雑誌（PM）、図書館書籍（LB）、また特定目的コーパスとして白書（OW）、ベストセラー（OB）、知恵袋（OC）、ブログ（OY）、法律（OL）、国会会議録（OM）、広報誌（OP）、教科書（OT）、韻文（OV）である。本研究ではMeCabとUniDicの短単位で解析されたデータを利用した。

JpWaCは4億語のウェブコーパスで、スケッチエンジンというレクシカルプロファイリングツールに載せている（スルダノヴィッチ・仁科 2008）。このコーパスは、副詞分布による13種のデータを分析した結果、均衡BCCWJコーパスの書籍のデータに最も類似しており、偏りの少ないデータであることが明らかになった（Srdanovićら2008）。コーパスはChaSen と IPADICで解析されたデータで、統一するため、取り出した形容詞のリストをUniDicで再解析する。

４．日本語教育用の形容詞の語彙リスト

本節では既存リストおよびコーパスに現れる形容詞の項目を調べ、新しい日本語教育用の形容詞語彙リストの項目として検討する。クラスター分析でコーパスごとに現れている最も高頻度の形容詞を分析し、グラフで表示した上でその形容詞の日本語教育においての利便性を議論する。

４．１ 既存の語彙リストに現れる形容詞—語数と難易度

表1は既存の語彙リストに現れる形容詞の語数を示している。

コーパス頻度と分散度に基づいたTM語彙リスト（一般基本語彙）にある形容詞は93語である。この形容詞は最も基本的な語彙で、高頻度であり、様々なサブコーパスに表れるので早い段階で導入し、学習するようにリストの作成者が推薦している。国研国語語彙DBは日本人の国語教育用のため、形容詞数は大きくなっている（460語）。他のリストは日本語教育用の語彙リストで、いわゆる日本語学習者が勉強するための語彙がカバーされており、高い数字からみると、TM一般語彙リスト（353語）、TM留学生語彙リスト（345語）、日本語語彙表（302語）、国研日本語語彙DB（236語）、旧JLPT語彙リスト（245語）で並ぶ。このうちもっとも収録語数が多いリストと収録語数が少ないリストを見ると3分の1の差が見られる。

表 1　語彙リストに現れる形容詞の語数

	旧 JLPT 語彙リスト[1]	国研 国語 語彙 DB	国研 日本語 語彙 DB	TM 語彙リスト（一般）	TM 語彙リスト 2500	TM 語彙リスト（留学）	日本語教育語彙表[1]
形容詞-一般	245	460	236	342	90	334	280
接尾辞-形容詞的	/	/	/	8	0	8	12
形容詞-非自立可能	/	/	/	3	3	3	10
合計 - 形容詞 - 語数	245	460	236	353	93	345	302
合計 - 形容詞 - %	3, 27	1, 69	2, 00	1, 74	3, 68	1, 70	1, 68
合計 - 語彙リスト - 語数	7500	27234	11826	20326	2524	20312	18011

[1] 30 語の形容詞は他の品詞と形容詞と両方分析されている語も含んだ。

更に語彙リストにおける形容詞の難易度レベルを調べた結果、図 1 に示した。

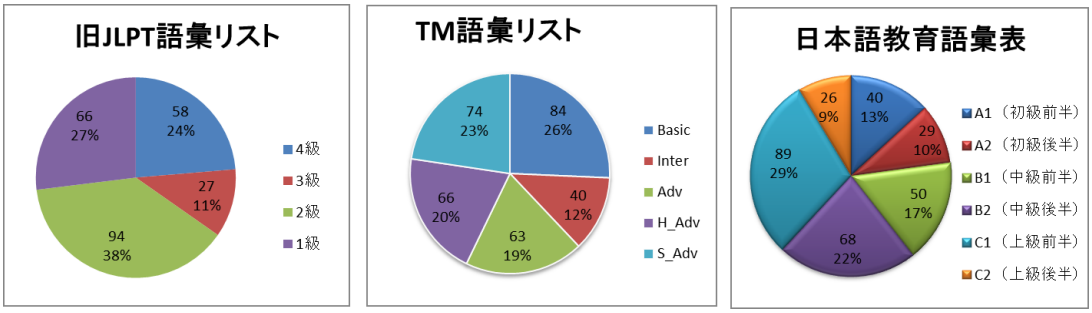


図 1　語彙リストにおける形容詞の難易度―レベルごとの形容詞の語数・割合

旧 JLPT 語彙リストでは 2 級の形容詞は 40%ぐらいの形容詞となっており、その妥当性に疑問が残る。TM 語彙リストにおける形容詞の分布がもっともバランスがとれているといえるが、初級と中級の形容詞の語数の割合を調整する必要があるかどうか検討が必要である。日本語教育語彙表は、初・中・上級ごとのバランスがあるといえるが、中級後半と上級前半の形容詞を合わせた割合が全体の半分の量になっている。一方で上級後半の語数は少なく，語彙のレベル分けに偏りがある。

４．２　コーパスに現れる・現れない形容詞

語彙リストを比較するためには統一が必要である。統一は表記統一、品詞統一、形容詞の単位の統一であり、またそれによってコーパスの頻度数も調整する必要がある。たとえば、「すごい」と「凄い」は表記が違うため別の項目としてコーパスの語彙リストに現れるとき、その頻度数も再計算する必要がある。特にコーパスのデータを比較するに当たって、利用した形態素解析ツールおよび電子化辞書によって差異が見られる。たとえば、UniDic 短単位の特徴は、語彙の表記を統一した上、短単位で解析を行う傾向がある。IPADIC の特徴は、語彙の表記を別に捉え、複合語も単位として載せる傾向がある。本研究では、BCCWJ が利用している MeCab と UniDic の短単位のデータをベースにして、コーパスのデータを揃

えた。それで、JpWaC が ChaSen と IPADIC を利用したため形容詞のリストを UniDic で再解析した際、形容詞の語数が非常に変わったため、ある程度手で直した。手直しの差異、表記の統一をしたが、短単位で分けた複合形容詞をリストから亡くさないように元のまま複合の形容詞として保存した（たとえば、「興味深い」）。表 2 では、各コーパスにおける形容詞の語数が見られ、JpWaC の場合統一前と統一後のデータを示した。前述した既存の語彙リストにおける形容詞の語数と比較すると、大きな差異があることがすぐ見てとれる。

表 2　BCCWJ と JpWaC に現れる形容詞の語数

	BCCWJ-UniDic	**JpWaC-IPADIC[1]**	**JpWaC-再計算 UniDic-手直し[2]**
形容詞-一般	789	1522	903
接尾辞-形容詞的	12	7	0
形容詞-非自立可能	6	11	3
合計―形容詞―語数	807	1540	906

[1] 元々の ChaSen-IPADIC の品詞タッグは「形容詞―非自立」と「形容詞一接尾」。
[2] 再解析の後、5 頻度までのデータを手で直して、672 語になった。4 頻度以下は形容詞として分析されたデータだけを計算した（234 語）。

BCCWJ と JpWaC の語彙リストを比較した結果、二つのコーパスに現れる形容詞の分布は、ほとんど類似していることが分かった。両者ともあまり偏りがないデータと考えられ、２種のコーパス比較で得られたデータで他の既存のデータの評価ができると考えられる。両方のコーパスに現れる、あるいは現れない形容詞を観察すると、データ処理方法の差異しか見られなかった（いくつか残った表記問題を含む）。あるコーパスリストに無い語はコーパスに無いわけではなく、そのコーパスの処理方法の結果、取り出されていないケースが多かった。ここでは、とくにコーパス語彙頻度リストの作成にあたって、形態素解析の依存性またその問題点が見られる。BCCWJ リストにあるが、JpWaC リストにない、また JpWaC リストにあるが BCCWJ リストにない項目は、大ざっぱに言って、三つに分けられる。

- 短単位の形態素解析を利用したためおよび形態素解析の誤りで現れていない形容詞
- 表記の違いがある形容詞
- 低頻度で、限られた分野の形容詞である

JpWaC に無いが、BCCWJ にある形容詞は、学習者用の語彙リストに現れない、「けばい、労労じい、露けい」などの例があげられる。

一方、BCCWJ リストにないが、JpWaC にある形容詞は、複合形容詞か表記の違いのものであり、直接コーパスを文字列か違う表記で検索した結果、BCCWJ コーパスにも現れるものである。たとえば、「興味深い」のような複合の形容詞で、UniDic の短単位で二分以上の部分に分ける。たとえば、「興味深い」は「興味」と「深い」になる。同様の問題が見られるものとして，高頻度 100 語のみを対象にした場合、「興味深い、格好いい、詰まらない、勿体ない、数多い」がある。

４．３　語彙リストに現れる・現れない形容詞

前節に取り出した「興味深い、格好いい、詰まらない、勿体ない、数多い」の形容詞は語彙リストに扱われているか、どの難易度レベルで扱われているか調査した。結果を表 3

に示す。

表 3　短単位で取り出せなかった高頻度 100 語以内の形容詞が語彙リストにあるか

	国研国語語彙 DB	国研日本語語彙 DB	旧 JLPT 語彙リスト	TM 語彙リスト	日本語教育語彙表
興味深い	-	-	-	-	中級後半
格好いい	-	ある、級なし	-	-	-
詰まらない	1	-	4	-	-
勿体ない	1	1	2	-	中級前半
数多い	-	-	-	-	中級前半

国語教育語彙リストのデータベースでも三つの形容詞が無い理由は、一般的に複合語のデータが圧倒的に少ないこと、この語彙が近年頻度が高くなったことなどが考えられる。旧 JLPT 語彙リストは、三つ、日本語教育語彙表は二つの高頻度複合形容詞を扱っていないという結果が得られた。また，TM 語彙リストは UniDic 短単位を利用しているので、対象の形容詞は語彙リスト以外だと予想できる。「詰まらない」の語彙レベルは、統一されているが、「勿体ない」の場合、低学年のレベル(1)と中級レベルが見られる。以上の形容詞の頻度と分布から、それぞれの形容詞を学習項目として語彙リストに入れることが推薦できる。

さらに、BCCWJ の高頻度の形容詞の 100 語は、日本語教育語彙リストにカバーされているかを調べた。国研日本語語彙 DB、TM 語彙リスト、日本語教育語彙表では、100 語以内の形容詞がすべてカバーされている。しかし、表記統一には問題があるとよく見られる。旧 JLPT 語彙リストには、「幅広い」という形容詞がなかった。

高頻度の 100 語以降、特に 200 語以降の形容詞がリストにあるか無いかを検討すると、2 種の均衡コーパスおよびサブコーパスにおいて同じような分布を持っている形容詞のケースは多いが、散発的に数少ない形容詞がリストにはある。たとえば、「初々しい」は旧 JLPT 語彙リストと日本語教育語彙表にあるが、「歯痒い」が無い。同じように「心強い」と「名高い」は同じ分布なのに、前者だけが語彙リストに載っている。そのため、コーパスの頻度を基にして同じような分布の形容詞は新しい項目として日本語教育用の扱いを検討する必要がる。

高頻度の 200 語・300 語の形容詞のうち、日本語教育語彙表と旧 JLPT 語彙リストの中上級・上級のものがあるが、同じような特性を持った形容詞は扱っていない。逆に、その面から BCCWJ を基にした TM 語彙リストがもっとよくカバーされている。

４．４　クラスタ分析で見られる形容詞分布

コーパス頻度をもとに、高頻度語に対する統計的な分析を以下の手順で行った。1)BCCWJ の中で合計頻度 500 以上の形容詞 158 語を抽出。2)158 語に対する JpWaC の総頻度を抽出し、分析データを作成。3)BCCWJ の合計頻度やサブコーパスでの頻度、JpWaC の総頻度を対数変換。4) 対数変換済み値をもとに、SPSS で階層的クラスタ分析と主成分分析を行った。ただし、解析データに関して、一点だけ調整した。品詞や語彙の切り方の相違により、BCCWJ には形容詞として掲載されているが、JpWaC では形容詞として認定されていない 6 語「易い、旨い、らしい、〜ぽい、さり気無い、限り無い」は分析対象から外した。最終的には、152 語を対象に分析を行った。

階層的クラスタ分析におけるオプションとして、クラスタ法は、Ward 法を使用、サンプル間の距離定義は、ユークリッド距離を使用した。クラスタ分析の分類精度を評価するため、判別分析を行った。判別分析では、階層的クラスタ分析で出力したクラスタ数を従属

変数に、対数変換後の値を独立変数にして、変数同時投入法で解析を行った。判別分析の結果、6 個のクラスタの場合 88.2%、5 個のクラスタの場合 86.8%、4 個のクラスタの場合 90.1%、3 個のクラスタの場合 89.5%の予測精度が示された。この結果を受け、152 語のデータは、4 個のクラスタとして捉えるのがもっとも適切であると判断した。

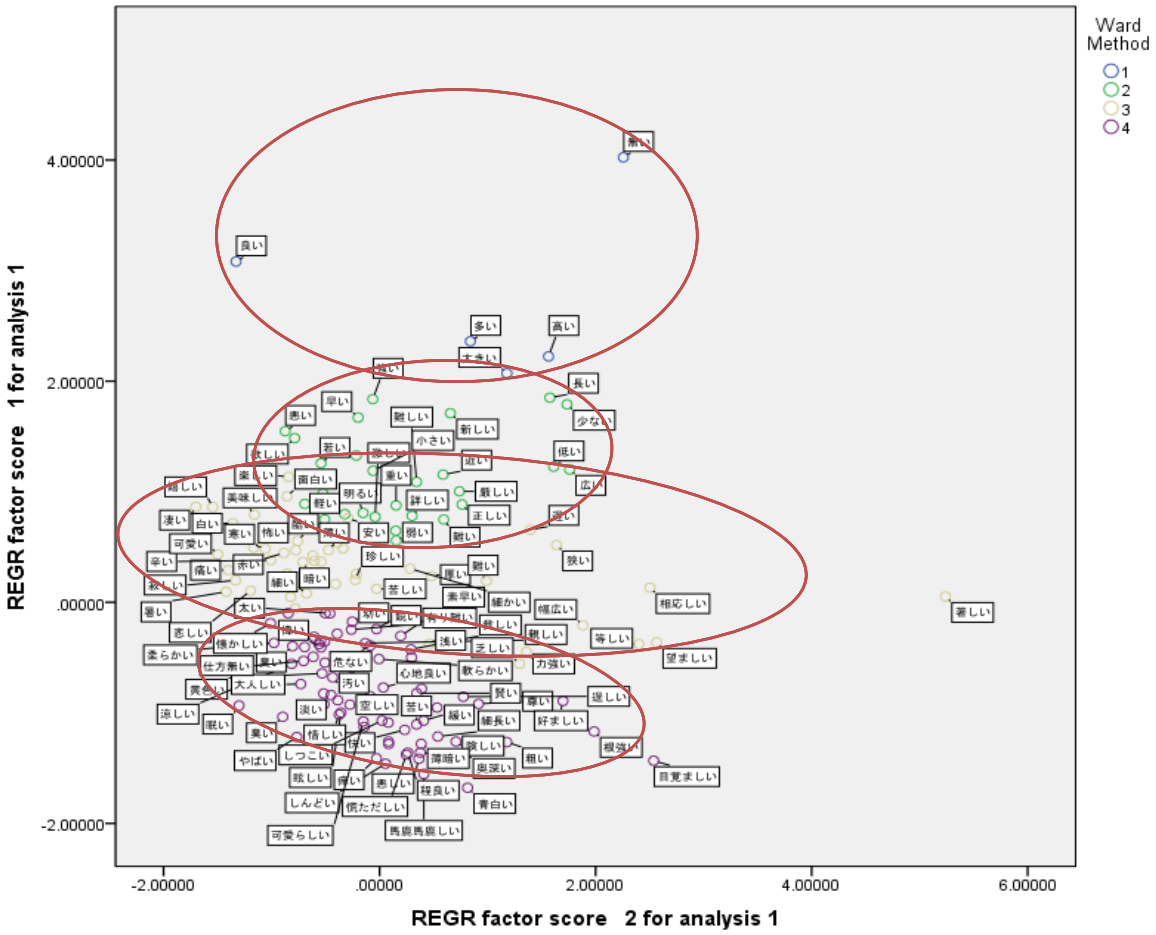


図2　第一主成分×第二主成分の得点によるサンプルの散布図

主成分分析では、クラスタ分析と同様に、対数変換済み値を使用した。第一主成分と第二主成分の合計固有値は、81.8%で、この二つの主成分で、8 割以上のデータが説明できる。また、Kaiser-Meyer-Olkin の標本妥当性の測度も 0.920 となり、説明力の高い分析であることが明らかになった。このことを踏まえ、クラスタ分析の結果と対応する形で、主成分得点をもとに、152 語の散布図を作成した。

図 2 に目立つ他のクラスタから離れている形容詞（無い、よい、著しい）は、特性を持っている語である。「無い」と「よい」は、非自立可能な形容詞であり、「著しい」は一番高頻度で特定目的白書のコーパスに現れ、偏りがある分布をもった形容詞である。既存の語彙リストでも、高い段階で教えている語である（JLPT:1 級、日本語教育語彙表：中上級、TM 語彙リスト：Inter）。

更に、各クラスタの具体例を表 4 に示す。

表4 クラスタの具体例

区分	ﾀｲﾌﾟ数	形容詞の例
ｸﾗｽﾀ1	5	無い、良い、多い、高い、大きい
ｸﾗｽﾀ2	27	弱い、短い、明るい、重い、激しい、強い、悪い、少ない、長い、早い、新しい、欲しい、深い、若い、古い、小さい、軽い、難しい、難い、正しい、低い、近い、優しい、広い、美しい、詳しい、厳しい
ｸﾗｽﾀ3	48	乏しい、力強い、浅い、等しい、幅広い、望ましい、相応しい、苦しい、厚い、著しい、恥ずかしい、青い、濃い、細い、凄い、悲しい、楽しい、細かい、美味しい、面白い、熱い、難い、嬉しい、暑い、冷たい、白い、忙しい、寂しい、可愛い、狭い、珍しい、酷い、安い、温かい、遠い、甘い、怖い、黒い、暗い、薄い、辛い、素晴らしい、痛い、可笑しい、寒い、赤い、固い、遅い
ｸﾗｽﾀ4	72	目覚ましい、馬鹿馬鹿しい、悪しい、青白い、根強い、重たい、奥深い、程良い、粗い、慌ただしい、しんどい、分厚い、険しい、醜い、尊い、快い、苦い、細長い、痒い、緩い、薄暗い、荒い、可愛らしい、眩しい、鈍い、切ない、逞しい、惜しい、空しい、好ましい、賢い、危うい、しつこい、心地良い、凄まじい、淡い、情けない、臭い、羨ましい、やばい、でかい、黄色い、辛い、涼しい、汚い、大人しい、軟らかい、貧しい、悔しい、危ない、怪しい、不味い、煩い、眠い、めでたい、素早い、臭い、柔らかい、丸い、親しい、きつい、幼い、久しい、懐かしい、有り難い、物凄い、鋭い、偉い、恐ろしい、太い、仕方無い、宜しい

各クラスタの解釈のため、BCCWJ の総出現頻度をもとに、平均値を確認した。クラスタ 1 は、平均頻度が 161732.2 となり、超高頻度の形容詞である。クラスタ 2 は平均頻度が 13993.4 となり、高頻度の形容詞と言える。クラスタ 3 は 5673.4 となり、一定量の使用が確認されるが、高頻度で初級でもよく教える語もある。クラスタ 4 は、平均頻度が 1523.9 となり、比較的に頻度も低く、難易度も高い語彙が多いことが確認された。クラスタ 1、2 は初級学習者には必須、クラスタ 3 は、初級・中級学習者に分けられ、クラスタ 4 は主に中・上級学習者向けの形容詞であるが、「黄色い」および「円い」のような普段初級で学習される語彙も現れる。クラスタで得られた結果は、サブコーパスおよびコーパスごとの頻度・分布を基にしたグルーピングで、直接語彙習得段階と結びつけにくいところもあり、他の要因を考慮に入れつつ教育のために利用可能である。また、同じような方法で 158 語以外の形容詞も分析する必要がある。

５．新しい形容詞の語彙リストに向けて

既存の語彙リストとコーパスを分析した結果、新しい形容詞の語彙リストを作成するメリットがあることが明らかになった。本研究で収集した形容詞のデータを更にデータベース化して、統一し、比較できるような表として提供することが望ましい。今回は BCCWJ およびそのサブコーパスによる形容詞の語彙頻度のリストをベースにしたが、今後、今月公開された超大規模 JpTenTen ウェブコーパスのデータを利用する予定である。また、今回の研究に UniDic の短単位を利用したが、長単位のデータが抽出できるようになったので、複合形容詞を適切に扱うために長単位にデータを揃えてデータベース化する予定である。2 節に取り上げた語彙リストにあるデータを統合し、形容詞の見出し語以外に、表記、品詞、それぞれの語彙リストにあるかどうか、その取り出した難易度レベル、コーパス・サブコーパスごとの頻度、形容詞の形式、意味分類などの情報を提供することを目標としている。

６．まとめと今後の課題

本論文では、形容詞を対象にして既存の日本語の様々な語彙リストと大規模な 2 種のコーパスから取り出した頻度リストを比較した。その結果、形容詞の語数およびカバーされた形容詞、その難易度レベルの間にキャップがあると判断された。例として取り出した既存の語彙リストにはない高頻度の形容詞は、今後の語彙リストの項目として入れる必要が

ある。

どの既存の語彙リストも全体の大規模な現代日本語均衡コーパスか大規模な現代日本語の調査を利用していないため、語彙リストの再作成が必要であると考えられる。また、コーパスデータを用いた語彙リストは、形態素解析および電子辞書の言語処理方法への依存性があるということが確認され、特に複合形容詞のデータはほとんどカバーされていない。そのため、今後 UniDic の長単位の分析結果が望まれるとことである。

謝　辞

本研究は、博報財団第 7 回「日本語海外研究者招聘事業」「日本語教育における語の共起関係」という研究（平成 24～25 年度、招聘研究員：スルダノヴィッチ・イレーナ）および「研究種目と分野：基盤研究(A)日本語教育研究課題名：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（研究代表者：砂川 有里子（筑波大学)」による支援を得ています。「教育基本語彙の基本的研究」のデータベースおよび BCCWJ のコーパスを含めて国立国語研究所が研究環境を与えてくださったことに感謝いたします。

文　献

饗場淳子（2011）「日本語教育用語彙に共通する語についての一考察」早稲田大学大学院教育学研究科紀要 18-2, pp. 275-285.

国際交流基金・(財)日本国際教育協会（1994）『日本語能力試験出題基準』凡人社

教育基本語彙の基本的研究―増補改訂版―（2008）国立国語研究所報告 127, 明治書院

砂川有里子(2012)「学習辞書編集支援データベース作成について -『学習辞書科研』プロジェクトの紹介」『日本語教育連絡会議論文集』24, pp. 164-169.

スルダノヴィッチ・イレーナ, 仁科喜久子（2008）「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 国書刊行会, pp. 59-80

スルダノヴィッチ・イレーナ（2012）「複数のデータを活用したイ形容詞と名詞のコロケーションの記述―日本語教育のための資料作成を目指して―」第 82 回 NINJAL サロン, 2012 年 11 月 27 日

大学英語教育学会基本語改定委員会(編) (2003)「大学英語教育学会基本語リスト: JACET List of 8000 Basic Words」大学英語教育学会

橋本直幸・山内博之（2008）「日本語教育のための語彙リストの作成」『日本語学（特集 :「語彙の教育」)』27-10, 明治書院, pp. 50-58.

松下達彦 (2010) 「日本語を読むために必要な語彙とは？ ー書籍とインターネットの大規模コーパスに基づく語彙リストの作成」『2010 年度日本語教育学会春季大会予稿集』pp. 335-336.

松下達彦 (2011) 「日本語を読むための語彙データベース」(The Vocabulary Database for Reading Japanese) Ver. 4.0. (http://www.geocities.jp/tatsum2003/よりダウンロード可能)

李在鎬・砂川有里子(2012)「コーパスを活用した日本語語彙表の構築」2012 年日本語教育国際研究大会（ICJLE2012）パネルセッション 日本語教育につながるコーパス研究―現状と今後の展望―（名古屋大学）

山内博之（編）（2008）『日本語教育スタンダード試案　語彙』ひつじ書房

Nation, Paul (2001) Learning vocabulary in another language. Cambridge University Press

Srdanović, Irena, Bekeš, Andrej, 仁科喜久子(2008)「複数のコーパスに見られる副詞と文末モダリティの遠隔共起関係」特定領域研究,「日本語コーパス」平成 19 年度公開ワークショップ（研究成果発表会）予稿集, pp. 223-230.

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/
日本語教育語彙表の検索システム「学習項目解析システム」http://lias.intersc.tsukuba.ac.jp/
中納言検索システム（BCCWJ）https://chunagon.ninjal.ac.jp/
スケッチエンジン検索システム（JpWac, JpTenTen）http://www.sketchengine.co.uk/
TM 語彙リスト http://www.geocities.jp/tatsum2003/
形態素解析辞書 UniDic http://download.unidic.org/twitter.com/unidic

『枕草子』長単位データを用いた相の類の分析

冨士池　優美（国立国語研究所　コーパス開発センター）†

The Adjectives and the Adverbs in *Makura-no-Soushi*: An Analysis Based on Long-unit-word

Yumi Fujiike　(Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では現在、「日本語歴史コーパス」の準備が進められている。日本語歴史コーパスの形態論情報については、言語単位として「短単位」「長単位」の 2 種類を採用し、それぞれに代表形・品詞等の情報を与える[1]。

冨士池（2012b）では『枕草子』長単位データを用いて、随想的章段・類聚的章段・日記的章段の章段分類を別ジャンルの文章と見立て、品詞比率の比較を行った。その結果、名詞率・形容詞率は類聚的章段で高く、日記的章段で低く、随想的章段はその中間となった。また、副詞率は章段分類の別なくほぼ一定であった。一般に形容詞率や副詞率といった相の類の比率は名詞率と負の相関関係が認められるものだが、負の相関関係が認められないことが確認された。これは形容詞や副詞といった相の類が『枕草子』で多用されていることの現れと考えられる。

本発表では『枕草子』の相の類に注目し、どのように用いられているのか、実態を探る。

２．問題の所在

２．１　冨士池（2012b）

まず、『枕草子』長単位データを用いて、随想的章段・類聚的章段・日記的章段の章段分類を別ジャンルの文章と見立て、品詞比率の比較を行った冨士池（2012b）について、概要を示す。

（１）調査対象

調査にあたり、準備中の『枕草子』長単位データを用いた。ここで、言語単位の概要を説明したい。日本語歴史コーパスでは言語単位として、短単位・長単位の 2 種類を採用している[2]。短単位は言語の形態的側面に着目して規定された言語単位であり、意味を持つ最小の単位（最小単位）を規定した上で、その最小単位を短単位認定規定に基づいて結合さ

† yfujiike@ninjal.ac.jp

1 短単位データについては「日本語歴史コーパス　平安時代編（先行公開版）」として、『枕草子』を含む仮名文学作品 10 作品を公開中。　http://www.ninjal.ac.jp/corpus_center/chj/

2 「日本語歴史コーパス」中古和文の単位認定基準は「現代日本語書き言葉均衡コーパス」で用いた認定基準を中古和文に対応できるように変更・拡張したものである。短単位の認定基準については小椋秀樹・須永哲矢（2012）、長単位の認定基準ほか概要については冨士池（2012a）を参照。

せる（または結合させない）ことにより、短単位を認定する。これに対して長単位は構文的側面に着目して規定された言語単位であり、文節を規定した上で、文節を長単位認定規定に基づいて自立語と付属語に分割することにより、長単位を認定する。例えば「小野−小町」「たへ−がたし」「うつくし−げ」は短単位では「−」の位置で切り離されるが、長単位では一まとまりとなる。短単位と長単位は同じ品詞体系を持つ。

章段分類の認定については、池田亀鑑（1963）の分類に基づき、「随想」「随筆」となっているものを随想的章段、「分類」となっているものを類聚的章段、「日記」となっているものを日記的章段とした[3]。これらが混在する章段もある。新編全集で「一本」とある章段については分類が不明であったため、調査対象からは除外した。

（2）調査結果

従来行われてきたテキストを特徴付ける指標として、主要な品詞比率を章段分類別に求めたほか、樺島忠夫・寿岳章子（1965）で提案された品詞比率に基づく指標である MVR（後述）を用いて、各章段分類の文体的特徴の把握を試みた。資料規模があまりに小さい章段と章段分類が混在する章段は除外し、随想的章段・類聚的章段・日記的章段のいずれかから延べ語数の多い 50 章段を取り出し、調査対象とした。50 章段で全体の約 6 割に当たる。

品詞比率

図 1 に示した名詞率・動詞率・形容詞率・副詞率は「各品詞の延べ語数／総語数」である。「Z」は随想的章段、「R」は類聚的章段、「N」は日記的章段を示し、ひげの下端が最小値、上端が最大値、箱の下端が第 1 四分位点、横線が中央値、上端が第 3 四分位点、ひげから外れた点は軽度の外れ値である。

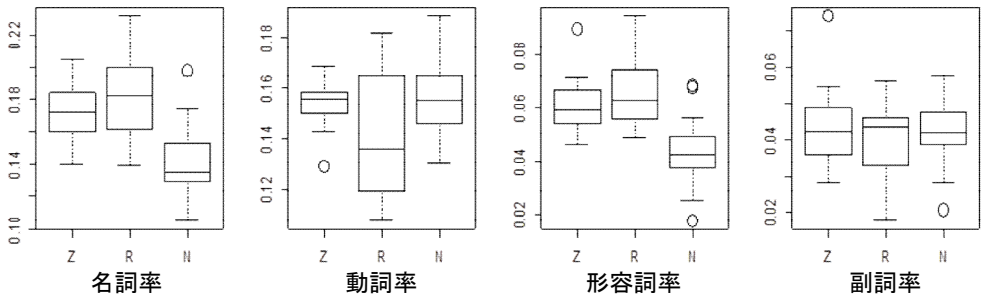


図 1　章段分類別の品詞比率

図 1 から、日記的章段の名詞率が他の章段分類と比較して低い様子が見てとれる。また、類聚的章段では章段による名詞率のばらつきが大きい。動詞率については、類聚的章段における章段によるばらつきの大きさが名詞率以上に目立つ。日記的章段の動詞率がやや高くなっている。形容詞率は名詞率と似た傾向を示しており、日記的章段の形容詞率は他の章段分類と比較して低い。副詞率はばらつきの差がややあるものの、章段分類による大きな差異は見られなかった。小磯ほか（2008）や冨士池ほか（2010）は現代書き言葉を対象

3　章段分類の認定に関しては諸説あり、定番と言えるような基準がないようである。今回の調査では、内容を重視し、池田（1963）に基づくこととした。

とした調査であるが、名詞率と動詞率・形容詞率・副詞率は負の相関にあった。これに対して冨士池（2012b）の調査結果では形容詞率が名詞率と正の相関を示しており、副詞率も負の相関は見られなかった。

MVR

名詞の比率は文章の特質を表し、名詞の比率に応じて他の品詞もある傾向を持って変化する、つまり文章のジャンルによって品詞の割合が決定されると考えられる。自立語について、品詞をその機能によって、体（名詞類）・用（動詞）・相（形容詞・形状詞・副詞・連体詞）・他[4]の四つに分類したとき、体の類と、用・相それぞれの類の関係を見るにあたり、樺島・寿岳（1965）は MVR という指標を提案した。MVR は「MVR＝100×相の類の比率／用の類の比率」の式で表される。体の類の比率（以下、名詞率とする）は、一般に要約的な文章で大きく、描写的な文章で小さいとする。また、MVR の値が大きいほどありさま描写的であり、MVR の値が小さいほど動き描写的と考えられるとし、名詞率と MVR の組み合わせから以下のような文体的特徴が見出せるとした[5]。

名詞率：大、MVR：小	要約的な文章
名詞率：小、MVR：大	ありさま描写的な文章
名詞率：小、MVR：小	動き描写的な文章

この指標を用いて、品詞比率から見る文体的特徴の把握を試みた。横軸に名詞率、縦軸に MVR を取った散布図を図 2 に示す。左下から右上にかけて、概ね日記的章段（黒）、随想的章段（白）・類聚的章段（灰）の順で並んでいる様子が見てとれる。文体的特徴としては、日記的章段は名詞率が小さく MVR も小さい「動き描写的」な文章、随想的章段は日記的章段と重なるが MVR がやや高く「ありさま描写的」な文章であり、類聚的章段は樺島・寿岳（1965）では示されていない、名詞率が大きく MVR も大きい文章ということになる。類聚的章段は名詞率が高く、動詞率が低く、形容詞率が高いということで文体的特徴を「動き描写の少ない」文章としておく。「×」は『枕草子』全体の名詞率と MVR を示すものであり、ほぼ中心に位置している。これまで『枕草子』の品詞比率と考えられてきたものは、名詞率・MVR 共にばらつきがあるものが集約された結果であると言えるだ

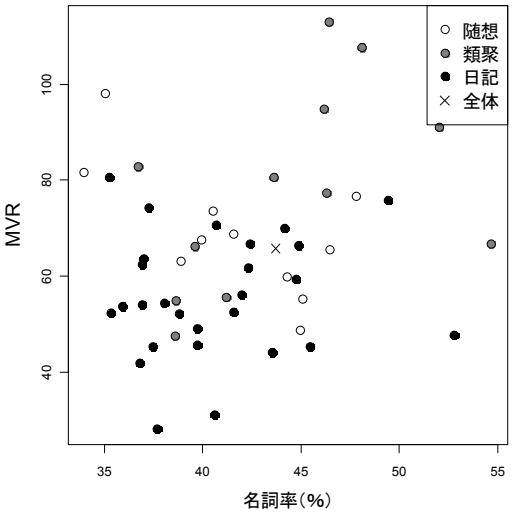


図 2　名詞率と MVR（『枕草子』章段分類別）

4 樺島忠夫（1950）の分類による。樺島・寿岳（1965）にならい、「他」については数が少ないため省略した。

5 樺島・寿岳（1965）p.30-36。単位の長さについては言及がないが、品詞の説明に「大きな区分として自立語と附属語との二種に分かれる」（p.27）とあることから、文節に基づく長い系列の単位、つまり長単位相当と推測される。

ろう。

このような『枕草子』の各章段の名詞率・MVR の分布は中古和文の中でどのような位置付けになるのだろうか。他の作品と比較を試みるため、参考程度ではあるが、図 2 に『古典対照語い表』に基づく同様の散布図を重ね合わせたものを図 3 に示す。「＊」が『古典対照語い表』に基づくもの（作品の略称を付した）で、白・灰・黒・×の点が図 2 の点である。『古典対照語い表』所収の他作品と比較すると『枕草子』は名詞率が 44.2%、MVR が 69.5 となっており、名詞率が小さく MVR が高い「ありさま描写的」な文章と見える。『蜻蛉日記』は名詞率が低く MVR もやや低めで、『枕草子』の日記的章段と同傾向に見え、似た品詞構成から成る可能性がある。その一方で類聚的章段のように MVR が高いものは『古典対照語い表』所収作品には見られなかった。各章段分類と他作品との関係については、本来、単位を揃えて慎重に検討すべき問題であり[6]、他作品の長単位データ整備後の課題としたい。

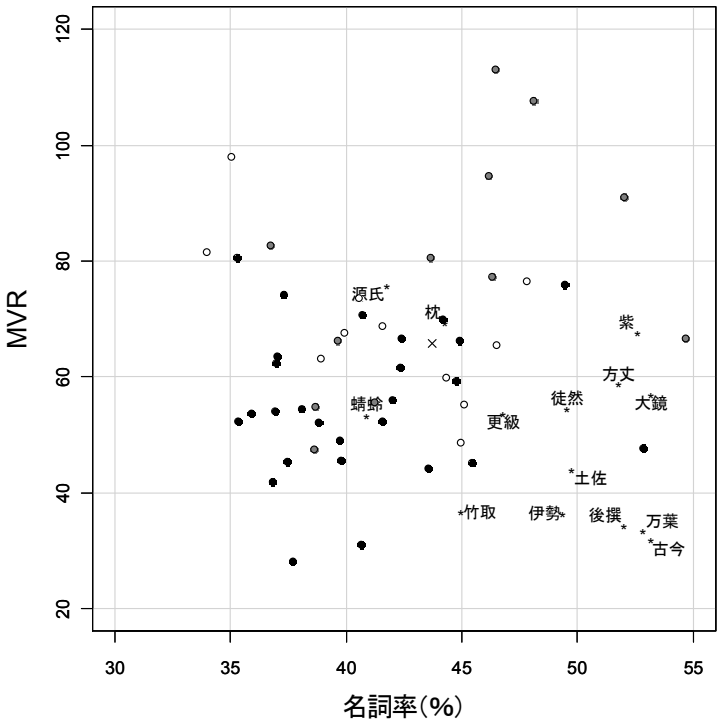


図 3　名詞率と MVR（図 2＋『古典対照語い表』）

２．２　問題の所在

『枕草子』はこれまで、名詞率に対して MVR、つまり動詞に対する形容詞類の割合が高いと考えられてきたが、冨士池（2012b）では名詞率・MVR が共に高い類聚的章段、名詞率・MVR が共に低い日記的章段、名詞率・MVR が共に中間的（他作品と比較すると MVR が高い可能性がある）随想的章段という、異なる品詞比率を持つ文章の集合体であることが明らかになった。特に類聚的章段は、『古典対照語い表』所収の他作品との比較の限りでは、他に類を見ない品詞比率に見える。MVR は動詞に対する形容詞類の割合であるが、冨士池（2012b）では動詞率は名詞率と負の相関があり、一般的な傾向を示していた。そこで、本稿では相の類（形容詞・形状詞・副詞）に注目し、どのように用いられているのか、章段分類別に見ることで実態を探る。

6 『古典対照語い表』の単位は文節に基づくものであり「長単位」に近いものであるが、一部接辞を認めていないため（「御」を冠する語はそれを除いた形式を 1 単位として認める等）、短単位と対応するものもある。

３.『枕草子』相の類の分析

３．１　調査概要

調査にあたっては『枕草子』全体の対象とした長単位データを使用する[7]。随想・類聚・日記のいずれかの要素が混在する章段を「混在」章段とし、新編全集で「一本」とある章段については「不明」とした。ここで、『枕草子』の資料規模として、各章段分類の章段数・延べ語数（長単位）を表 1 で確認しておく。

表１　各章段分類の章段数・延べ語数

	随想	類聚	日記	混在	不明	計
章段数	97	140	50	12	28	327
延べ語数	17643	15094	34604	6392	1497	75230

『枕草子』における相の類の用いられ方を見るにあたって、章段分類別に相の類の頻度・比率、名詞・動詞・形容詞・形状詞[8]・副詞の相関関係から、相の類の各品詞が各章段分類でどのように用いられているのかを見る。また、高頻度語やコレスポンデンス分析（対応）を通して、各章段分類で相の類のどのような語が用いられているのかを確認する。

３．２　調査結果

（１）品詞比率

表２　各章段分類の相の類の頻度・比率

	随想	類聚	日記	計
形容詞	907	820	1293	3435
	5.14%	5.43%	3.74%	4.57%
形状詞	236	184	246	774
	1.34%	1.22%	0.71%	1.03%
副詞	710	520	1431	2969
	4.02%	3.45%	4.14%	3.95%
計	1853	1524	2970	7178
	10.50%	10.10%	8.58%	

章段分類別相の類の頻度と比率を表 2[9]に示す。比率は「各品詞の延べ語数／総語数」で、２．１節の図 1 には表示されていない平均値を示したものである。図 1 と表 2 から日記的章段は相の類の比率が随想・類聚と比較して低く、形容詞においてそれが顕著であることがわかる。また、類聚で副詞の中央値は高いが平均比率は低く、他の章段分類と比べて副詞の比率が低い章段が多い様子がうかがえる。このように、相の類の中でも章段分類により品詞構成に差がある様子が確認できる。

（２）品詞比率の相関関係

２．１節では名詞率と MVR の相関を見たが、MVR は動詞と相の類の比率を見る指標であり、動詞及び相の類に属する各品詞の比率が集約されてしまい、前項（１）で見たような章段分類による品詞構成の差が捉えきれない。そのためここでは各品詞の相関関係を見る。図 4 に普通名詞・動詞・形容詞・形状詞・副詞それぞれの比率（自立語中の割合）の相関を示す。図 4 の点はそれぞれ随想的章段（白）・類聚的章段（灰）、日記的章段（黒）、混在章段・不明（濃い灰）である。

---

7 冨士池（2012b）は 2012 年 8 月時点の長単位データを使用したが、本稿では 2012 年 12 月時点のものを使用した。全体を対象としたため極端な結果を示す章段もある。例えば第 19 段「たちは」の場合、「たちは　たまつくり。」のみであり、自立語について見たときの品詞比率は名詞率 100%となる。

8 形状詞は形容動詞語幹に相当する。

9 計は随想・類聚・日記のほか、混在・不明を含む。

普通名詞との相関に注目すると、動詞は左上に類聚（灰）、右下に日記（黒）・随想（白）となっており、負の相関があるように見えるが、形容詞・形状詞・副詞はそれぞれ異なる様相を見せている。形容詞は類聚（灰）が負の相関を示しているが全体的に比率が高く、日記（黒）はある一定の低めの範囲に固まり、随想（白）は一部が正の相関を示しているように見える。形状詞は名詞の割合とは関係なくある一定の範囲にある。副詞は類聚（灰）が形容詞と同様に負の相関を示しており、日記（黒）・随想（白）はある一定の範囲にあるが、形状詞と比べるとばらつきが大きい。ここから、形容詞・副詞において名詞と負の相関関係が見出せなかったのは、日記（黒）・随想（白）の品詞構成によるものであることが確認できる。「猫は、上のかぎり黒くて、腹いと白き。」（第 50 段・猫は）など類聚的章段に多く、モノとその描写という性格を持ち動詞を必ずしも必要としないことから、名詞率に関わらず形容詞率が高い傾向にある。また、「冬は、いみじう寒き。夏は、世に知らず暑き。」（第 114 段・冬は[10]）、「坤元録の御屏風こそ、をかしうおぼゆれ。漢書の屏風はおぼしくぞ聞えたる。月次の御屏風もをかし。」（第 278 段・坤元録の御屏風こそ、をかしうおぼゆれ）のような比較的短い随想的章段で、名詞率と形容詞率が共に高くなっていた。その一方で日記的章段を中心に名詞率・形容詞率が共に低い章段がある。「大進生昌が家に、宮の出でさせたまふに、東の門は四足になして、それより御輿は入らせたまふ。」とはじまる第 6 段などが代表的なものであるが、日記的章段は他章段と比べて名詞率が低く、動詞率が高い傾向にある。『枕草子』中に、名詞率・形容詞率が共に高い章段と、共に低い章段が混在するため、結果として名詞率と形容詞率が負の相関を示さなかったことがわかる。

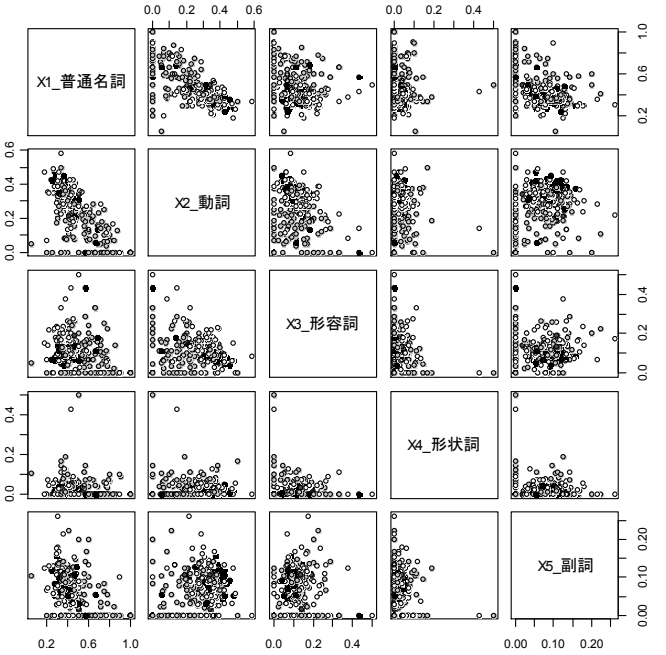


図 4　普通名詞・動詞・形容詞・形状詞・副詞の相関

（3）章段分類別頻度上位語

ここまで比率を見てきたが、ここでは各章段分類でどのような語が使われているのかを確認する。表 3 に形容詞・形状詞・副詞の章段分類別頻度上位 10 語を示す。全章段分類頻度上位 10 語に含まれない語に網掛けを付した。網掛けした語を見ると、「若し」は類聚に 7

10 池田（1963）では随想的章段と分類されていたが、「～は」型であり類聚的章段と類似の品詞構成を持つものと考えられる。章段分類には再考の余地があるものと考える。

例、日記に 13 例あり、随想にやや多い程度である。一方で「疾し」のように全 63 例中 43 例が日記と出現章段分類に偏りがあるものもあり、比率だけではなく語に関しても、章段分類による差がある様子がうかがえる。

表 3　形容詞・形状詞・副詞の章段分類別頻度上位 10 語

形容詞						形状詞						副詞					
随想	907	類聚	820	日記	1293	随想	236	類聚	184	日記	246	随想	710	類聚	520	日記	1431
可笑し	166	可笑し	93	いみじ	159	哀れ	19	哀れ	29	哀れ	16	いと	215	いと	172	いと	250
いみじ	86	無し	61	可笑し	109	清気	17	流石	8	然様	8	然	45	然	27	然	95
良し	45	いみじ	55	無し	105	鮮やか	8	可笑し気	7	漫	8	猶	39	猶	21	唯	75
無し	44	憎し	32	めでたし	66	然様	8	更	7	無下	7	少し	31	唯	19	猶	75
白し	26	めでたし	29	疾し	43	忍びやか	7	憎気	7	大き	6	唯	27	必ず	14	皆	66
めでたし	25	良し	25	怪し	37	細やか	7	清気	6	可笑し気	6	又	17	少し	14	如何で	49
憎し	18	白し	20	憎し	32	可笑し気	6	大き	4	清気	6	数多	15	又	14	え	49
近し	16	近し	18	良し	29	更	6	汚気	4	顕証	6	え	15	え	13	又	45
若し	16	多し	16	多し	27	大き	5	心殊	4	まめやか	6	皆	14	良く	11	少し	39
多し	15	怪し	14	近し	24	艶やか	5	唯	4	密か	6	まいて	13	未だ	10	など	39
高し	15					無下	5	徒然	4					皆	10		

（4）コレスポンデンス分析

ここでは、コレスポンデンス分析（対応分析）で章段分類と相の類の高頻度語との対応を確認する。相の類の頻度上位 20 語のコレスポンデンス分析結果の散布図を図 5 に示す。分析には統計分析パッケージ R の ca パッケージの ca 関数を用いた。第 1 次元の寄与率は 80.33%、第 2 次元の寄与率は 19.67%であった。図 5 では第 1 次元の正の方向に日記、負の方向には随想・類聚が布置されており、第 1 次元は日記とその他を分ける軸、第 2 次元は随想と類聚を分ける軸と見ることができる。第 1 次元を見ると、正の方向に、日記と共に「如何に」「如何で」「猶」「然」「唯」「皆」「又」「え」といった副詞、「いみじ」「怪し」「めでたし」といった形容詞が布置された。副詞はほぼ正の方向に布置されており、動き描写にあたり用いられる語と言える。例外である「いと」は程度の甚だしい様を表すものであり、主に形容詞が布置された負の方向に共に布置され、ありさま描写に用いられたと見ることができる。第 2 次元を見ると、正の方向に随想及び「可笑し」「良し」が、負の方向に類聚及び「哀れ」「憎し[11]」が布置された。「めでたし」「可笑し」「良し」「哀れ」は共に肯定的な評価を表す

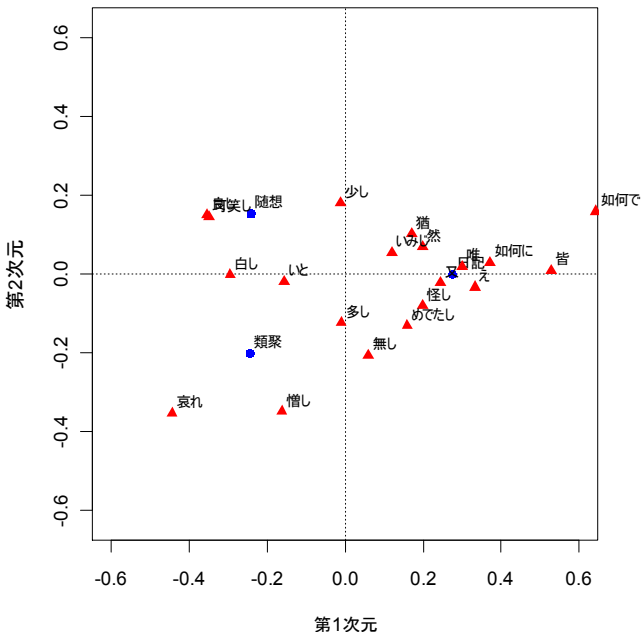


図 5　相の類頻度上位 20 語の散布

11　「憎し」は第 26 段「にくきもの」での多用が影響したものと考えられる。「形容詞＋もの」型章段は多くあるが、その形容詞が章段中に多用されることは少なく、「にくきもの」は例外と言える。

語であるが、「めでたし」は日記を、「可笑し」「良し」は随想を、「哀れ」は類聚をそれぞれ特徴付ける語と考えられる。

４．終わりに

『枕草子』長単位データを用いて品詞構成とその相関関係について章段分類別に見たところ、日記的章段で動詞率が高いこと、類聚的章段や短い随想的章段で動詞率が極端に低く名詞率・形容詞率が共に高いことが確認された。『枕草子』中に、名詞率・形容詞率が共に高い章段と、共に低い章段が混在するため、結果として名詞率と形容詞率が負の相関を示さなかったと言えるだろう。また、相の類の頻度上位語とそのコレスポンデンス分析から、日記的章段で副詞が動き描写に用いられ、随想・類聚的章段では形容詞がありさま描写に用いられたことや、肯定的な評価を表す形容詞類の中でも章段分類により差がある様子がうかがわれた。

付　記

本稿は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダーは近藤泰弘客員教授）の成果の一部である。また、用いた新編全集『枕草子』の電子テキストは、小学館から上記プロジェクトのために提供されたものである。

文　献

池田亀鑑（1963）『全講枕草子』至文堂

小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規定集」、平成 21（2009）－平成 23（2011）年度科学研究費補助金基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2 （http://dl.dropbox.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf よりダウンロード可能）

樺島忠夫（1950）「類別した品詞の比率に見られる規則性」『国語国文』24-6

樺島忠夫（1988）『日本語はどう変わるか　－語彙と文字－』岩波書店

樺島忠夫・寿岳章子（1965）『文体の科学』綜芸舎

小磯花絵・小木曽智信・小椋秀樹（2008）「短単位情報に基づくジャンル間の文体に関する分析」『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集』、pp.99-106

冨士池優美・小西光・小椋秀樹・小木曽智信・小磯花絵（2010）「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「日本語コーパス」平成 22 年度公開ワークショップ予稿集、pp. 273-280

冨士池優美（2012a）「中古和文における長単位の概要」、第 2 回コーパス日本語学ワークショップ予稿集、pp.51-58

冨士池優美（2012b）「枕草子の語彙――章段分類に注目して――」、第 101 回国語語彙史研究会（2012 年 9 月 29 日）口頭発表資料

宮島達夫編（1971）『古典対照語い表』笠間索引叢刊 4、笠間書院

接続助詞「けど」の音調と意味用法に関する予備的考察

田頭 未希（東海大学 教養教育センター）†

A Preliminary Study about Tone and Discourse Function of *Kedo*

Miki Tagashira (Foreign Language Center, Tokai University)

１．はじめに

田頭(谷口）(2012a ,b)では、話し言葉にみられる接続助詞「が」について、句末音調とその意味・用法について分析し、「言いさし」の用法では下降調、「談話主題の提示」の用法では上昇調となることが比較的多いこと、また一方で、特定の音調と用法が一対一で強く結びついているわけではなく、むしろ話し言葉では句末音調と意味・用法はある程度幅をもって対応していることを述べた。さらに、接続助詞「が」は語彙情報としては下降調をとるが、松永（2002）が指摘している「韻律句末の音調はイントネーションによって影響を受けやすいので注意が必要である」という見解を、量的分析により明らかにした。話し言葉では、句末は上昇調や上昇下降調などの音調変化を伴う方が一般的で自然であるといえる。

本稿では、接続助詞「が」とほぼ同じ意味・用法を持ち、「が」と同様に話し言葉において頻繁に使用される接続助詞「けど」「けれど」「けども」「けれども」[1]を取り上げ、その音調と意味・用法の関係について考察する。

２．目的

本研究の大きな目的は、日本語の話し言葉について韻律句末の音調と句末に表れるそれぞれの品詞の意味・用法との対応関係を体系的に記述することである。本稿では接続助詞「けど」類に注目する。

接続助詞「けど」類を扱う理由は以下である。まず、文末のイントネーションを扱った研究に比べ、いわゆる発話途中とみられる場所のイントネーションとその意味・用法・機能などとの関連を扱った研究が少ない。また、接続助詞はその本来の機能から発話末にも発話途中にも表れる品詞であるため、同じ品詞で同時に２つの生起位置の音調と意味・用法の関係をみることができる。田頭(谷口）(2012a, 2012b)で扱った「が」と同様に、複数の意味・用法を持つこと、さらに動詞や形容詞などに後続した場合、語彙情報としては「けど」類の内部では音調変化を伴わないことがあげられる。また一方で、「が」とは異なり２モーラ以上から構成された語で、実際の話し言葉ではその内部で音調変化を伴うバリエーションが「が」よりも多いと予測でき、もしバリエーションがあれば韻律句末の音調変化の現れ方の比較ができると考えられる。

† t-miki@tokai-u.jp

1 本稿では、接続助詞「けど」「けれど」「けども」「けれども」はそれぞれ異形態であるが、同じ意味・用法を持つ一つの形式と考える。本稿中に書かれている形態は基本的に他の３つのいずれに入れ替えて読んでも意味は変わらない。「けど」「けれど」「けども」「けれども」を合わせて言うときには「けど」類という表現を用いることにする。

３．分析データ

３.１　音声資料

『日本語話し言葉コーパス』（以下 CSJ）（Maekawa 2003[2]）のコアデータのうち、韻律情報が付与されている約 18 時間分（模擬講演 107 ファイル）を分析資料とした。

３.２　韻律句末の音調

本稿での「韻律句」の意味を定義しておく。イントネーションの物理的変化量として基本周波数を考え、時間軸に沿って示される音調の変化のうち、冒頭の上昇から始まり発話末にかけて下がっていく基本周波数で示されるひとつの山のまとまりを「韻律句」と呼ぶ（Pierrehumbert and Beckman 1988）。韻律句にはIntonation Phrase[3]（以下IP）とAccentual Phrase（以下AP）の2つがある。音調の連鎖という意味では、東京方言では、ひとつのアクセント句は、「相対的に低いピッチ（%L）で始まった後すぐに上昇し（H-）、アクセント核[4]があればそこで下降し（H*+L）、最後もまた低く終わる（L%）」という基本周波数の一連の変化からなる（五十嵐他 2008）。

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

４．接続助詞「けど」類

４.１　音調の型

『明解日本語アクセント辞典』（1997）によると、接続助詞「けど」類について語彙的に与えられている音調は「け」から「ど」にかけてアクセントを持つが、基本的には「けど」類単独で使用されることは少なく、動詞や形容詞、名詞に後続して用いられるため、次のように説明できる[6]。形容詞の場合も以下の説明の動詞の場合と同様に、起伏式形容詞の場合には形容詞の型を変えないで低く下がってつき、平板式形容詞の場合には最後の拍を変え、低く下がってつく[7]。ここでは便宜上、前接要素と比べ、低く下がる音のみを下線付きで表記する。

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

---

2 CSJ の概要について説明している論文のひとつである。

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

4 語彙的に指定されたアクセントを意味する。なお、この注釈は筆者が加筆したもので、五十嵐他(2008)は本文カギ括弧の表現である。%L、H-、H*+L などは CSJ で採用されている韻律ラベリング X-JToBI で使われる記号である。

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

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

7 平板式形容詞の場合は、形容詞の最後の拍を低く変え、低く下がってつく。

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

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

４.２　用法

接続助詞「けど」類は先行研究[8]により次のような用法[9]を持つことが指摘されている（森田 1980, 渡辺 2000、永田・大浜 2001 他）。例は（a）先行研究からの引用、または筆者による作例と（b）CSJ から取り出した例（鍵括弧にデータの Talk ID）を示す。（当該要素の太字表記、句読点位置と推定される箇所でのスペースは CSJ の転記にて分かち書きされた箇所を示す。）

（1）談話主題の導入：話題の移行、主題の提示する。
（a）お借りした本ですけど　とても面白かったです
（b）えーっと　まず　中学三年の　ま　高校入試の　頃に　ちょっと遡るんですけど　その時は　まー[S00M0065]

（2）逆接・対比：前出の文脈と相反する事項を述べる。統語的には取り立ての「は」「も」が用いられることや対照的な叙述が表現される。
（a）あの映画は　前半は面白かったけど　後半は退屈した
雨が降ったけど　運動会は行われた（永田・大浜 2001 より）
（b）普通だったら　その　立ち入り禁止区間を　入ったところにお咎めの　言葉　一言ぐらい　言うと　思うんですけど　それも　全く　言わずに[S01F0183]

（3）並列・累加：二つの事柄を並べる。
（a）彼は　走るのも速いけど　泳ぐのもうまい
（b）凄く　頭が　良くて　で　凄い　おとなしいんですけど　超面白い　感じ　で[S02F0094]

（4）挿入：補足説明を付け加える。「けど」節がなくても、前後の文意が通じる。
（a）この前貸した本を明日　もし無理だったら明後日でもいいんだけど　返してくれる？（永田・大浜 2001）
（b）そんな　ことも　ありましたし　娘と　二人で　毎日　あの　猫の　こと　書いて　あのー　夏目漱石じゃないけど　あのー　猫の小説でも書けると　いいねなんて[S01F1522]

（5）前置き：後続する事項を補足したり、後件の解釈を阻害する要因を排除す

[8] 「けれども」の用法を４つに分類するもの（三枝 2007）、６つに分類するもの（森田 1980, 永田・大浜 2001）など研究者によって必ずしも一致しているわけではなく、また分類された用法の語も少しずつ異なっていることに注意が必要である。本稿では森田や永田らの研究に用いられている分類を基に、「逆接」と「対比」の分類には曖昧な場合がある（渡部 2000）を反映させ、６つの用法に分類した。

[9] 定義は、のものを筆者により短くまとめている。例は永田・大浜（2001）より。

るために置く。「挿入」との違いは、「前置き」は「けど」節が前後の文意を理解するために必要である点、「談話主題の導入」との違いは、「談話主題の導入」がそれ以前までの話題との関係性の有無であるのに対し、「前置き」は「けど」節と後続節との関連が重要であるといえる。

(a) ちょっとMさんにききたいんですけどね　要するに　いま　この少年法の議論になってる　ま　いろんな　あのー　驚くような事件ですよね　その事件が頻発してくると　何か社会はおかしいんじゃないかとあの・・・(永田・大浜 2001)

(b) まず　理系は　全部　却下という　形に　なってますので　文系で　考えて　それで　今度は　経済学部とか　ありますけど　これはやっぱり　うん　経済だから　計算を　するんだろうと　やっぱり算数が　出てくるんじゃないかと[S01M0225]

(6) 言い切りの回避・言いさし：話し手の主張を弱める働きをする。対話では次に会話が続くことを話し手が意識しているサインとして働く。または、最後まで述べずに、話題が他のことへと移行してしまうような場合もある。

(a) あの人はとてもいい方だと思いますけど

(b) 面白い　授業で　それで　何とか　続けてこれた　面も　あるのかなと　で　歴史は　あんまり　得意じゃなかったから　きっとそれが　良かったんじゃないかと　思うんですけど　で　まー　そうやって　予備校に　通っていて　で　[S01M0225]

模擬講演データにみられた接続助詞「けど」類の例を上記 6 つの用法に分類することを試みた。当該要素の前後の転記、時間にして数秒から数十秒間に相当する箇所を読み、前後の文脈から筆者が判断し分類[10]を行った。

５．結果

分析データ全体では接続助詞「けど」は 1937 例あった。そのうちランダムに約半数を抽出し、1019 例について用法の分類を試みた。従ってこれが本稿での分析データ総数である。内訳を表１に示す。形態としては「けれども」「けど」の使用頻度が多いことが分かる。その他としては、「け」「げ」「けお」「けよ」「けれとも」「けろ」が含まれる。1019 例のうち、4.2 節で示した用法の分類において、判断に迷ったものが 45 例あった。その 45 例を除いた 974 例をもとに以下の分析結果と考察を行うことにする。

974 例の句末音調の割合の分布を表２に示す。全体としては、HL、つまり上昇下降調が約半数を占めている。韻律句末の音調変化は、前接要素のアクセントの位置と句末までの距離が関係している可能性が高いため、「けど」類を構成モーラ数[11]で分け（表３）、句末音調との関係を調べた。結果を表４[12]に示す。H つまり上昇調は４モーラの場合に特徴的に多く、L つまり下降調は４モーラで特徴的に少ないといえる。句末音調の型は韻律句末からの距離が影響している点が指摘できる。

次に用法について考える。用法別の例数を表５に示す。用法として最も頻度が高かった

---

10 分類に関して、判断に迷ったものや判断できなかったものについては、今回の分析ではすべて除外した。

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

12「け」などの １モーラの 3 例と、LHL の音調の型 4 例は分割表には入っていない。

表１　分析データの内訳

	けれども	けーども	けども	けれど	けーど	けど	その他	全体
データ全体	708	148	370	44	6	644	17	1937
分析データ	414	88	179	13	5	313	7	1019

表２　句末音調別の例数

	度数
L	202 (21%)
HL	451 (46%)
H	317 (33%)
LHL	4 (0%)
合計	974

表３　モーラ数別の例数

	度数
4モーラ	471 (49%)
3モーラ	190 (20%)
2モーラ	310 (32%)
1モーラ	3 (0%)
合計	974

表４　「けど」類構成モーラ数と音調

度数 列% 行%	H	HL	L	
２モーラ	82	148	79	309
	25.87	32.82	39.70	
	26.54	47.90	25.57	
３モーラ	50	89	49	188
	15.77	19.73	24.62	
	26.60	47.34	26.06	
４モーラ	185	214	71	470
	58.36	47.45	35.68	
	39.36	45.53	15.11	
	317	451	199	967

表５　用法別の例数

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

のは「挿入」の用法で、用例の約半数をしめていた。次が「逆接･対比」の用法で、この２つの用法で約８割ということになる。表６は句末音調と用法の関係を示す。いずれの用法も上昇調や上昇下降調となる割合が、下降調よりも高い。「逆接･対比」、「前置き」、「挿入」、「談話主題の導入」、「並列･累加」の用法では上昇下降調が高く、特に、「談話主題の導入」では６割がこの音調をとっている。また用例数は少ないが、「言い切りの回避」の用法では上昇調が高いことが分かる。さらに、どの用法でどの形態が使われやすいのかについては、「談話主題の導入」の場合には「けれども」の使用が半数以上を占め、また「累加･並列」

では「けども」と「けれども」で８割以上を占めているのが特徴としてあげられる。それ以外の用法では「けども」「けれども」に「けど」が加わり、３つの形態が使用されていることが分かる。

６．考察

表１に「けど」類の形態の使用頻度を示したが、もちろん話者によっては「けれども」の使用が圧倒的に他の形態よりも多い人もいれば、「けど」の使用頻度が高い人もおり、話者のくせがはっきりと現れている人がみられる。一方で、例えば「談話主題の導入」の場合には「けれども」を必ず使用するというように、用法と形態にある程度の決まりのようなものを持って話している話者もみられた。

音調に関しては、上昇下降調の頻度が高かったが、それにはいくつか理由が考えられる。まず、「けど」の場合、もともと「け」から「ど」にかけての下降を語彙情報として持って

表６　用法と音調

度数 列% 行%	H	HL	L	
逆接･対比	87	134	55	276
	27.44	29.71	27.23	
	31.52	48.55	19.93	
言い切りの回避	14	8	6	28
	4.42	1.77	2.97	
	50.00	28.57	21.43	
前置き	28	35	14	77
	8.83	7.76	6.93	
	36.36	45.45	18.18	
挿入	164	226	118	508
	51.74	50.11	58.42	
	32.28	44.49	23.23	
談話主題の導入	20	43	7	70
	6.31	9.53	3.47	
	28.57	61.43	10.00	
並列･累加	4	5	2	11
	1.26	1.11	0.99	
	36.36	45.45	18.18	
	317	451	202	970

いる。また、この上昇下降調には韻律句末の２モーラにまたがっての上昇下降と、句末の最終モーラ内での上昇下降も含まれ、バリエーションが多いことも上げられる。「けど」類は最大で４モーラの長さのものが前接要素について発話されることから、前接要素自体の長さやアクセントの位置などとの関係から韻律句末での語彙情報としての音調が顕在化されやすくなったと考えられる。「けれども」は「ど」と「も」の間でもアクセントのような下降をつけて発音される場合も考えられ、これらが上昇下降調の割合を高めている原因にあげられる。「けど」類と同様の意味･用法を持つ接続助詞「が」は圧倒的に上昇調をとる割合が高かった（田頭（谷口）2012b）点とも比較でき興味深いといえる。

用法については、「挿入の用法」が最も頻度が高かったが、「けど」類は他の接続助詞に比べ、節の結びつきに論理性が弱い点（三枝 2007）が指摘されており、それゆえ「けど」類節がなくても文全体の意味には影響を与えない付け足し的な説明や挿入が話し言葉では多用されていると考えられる。また「言い切り回避用法」は、その使われ方として終助詞の「ね」や「よ」に似ており、文末についてモダリティに相当するような意味も担っているといえる。尾谷（2003）はこのような使われ方をする「けど」類を一種のポライトネスマーカーとしての機能を持つと指摘しているが、音声面からも上昇調を伴うことで、より丁寧さを加えているといえる。

用法と音調の関係については、ある特定の用法と音調が強く結ぶつき、この用法の場合にはこの音調をとるという関係性はみられない。ゆるやかな傾向として、例えば、「談話主題の導入」の場合には上昇下降調が用いられやすいなどを指摘することはできる。この傾向は接続助詞「が」でもみられた（田頭（谷口）2012a,b）。

類似の用法を持ちながら、音調とのゆるやかな対応関係には違いがみられる「けど」類と「が」の比較については今後の課題としたい。

７．まとめ

『日本語話し言葉コーパス』を利用し、接続助詞「けど」類の音調と意味･用法、それらの関係について分析を行った。接続助詞「が」と同様に、「けど」類でも話し言葉では上昇下降調や上昇調をとる頻度が高いことが量的分析より明らかになった。音調と意味･用法との関係については、「談話主題の導入」として「けど」類が使われた場合に上昇下降調が用いられやすいという傾向はみられたものの、全体として、音調と意味･用法はゆるやかに対応していることが指摘できる。

参考文献

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

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

尾谷昌則（2003）p「主体化に関する一考察：接続詞「けど」の場合」『日本認知言語学会論文集』第 3 巻、pp. 85–95. （http://www.i.hosei.ac.jp/odani/kedo_JCLA3.pdf よりダウンロード可能）

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

田頭(谷口)未希（2012a）「接続助詞「が」の音調と意味用法 –『日本語話し言葉コーパス』の分析を通して–」第一回コーパス日本語学ワークショップ発表資料

永田良太･大浜るい子（2001）「接続助詞ケドの往訪間の関係について –発話場面に着目し

て–」『日本語教育』110、pp.62-71、日本語教育学会
森田良行（1980）『基礎日本語２　–意味と使い方–』、角川書店
渡辺学（2000）「逆接表現の記述と体系　ケド･ワリニ･クセニをめぐって」『現代日本語研究』７、大阪大学大学院
Maekawa, K. (2003) Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition.* 5-8. Tokyo.
Miki Tagashira-Taniguchi (2012b)　*Tone and Function on /ga/ in Japanese*. Workshop on Innovation and Applications in Speech Technology (IAST) in Dublin
Pierrehumbert, B. and M. Beckman (1988) *Japanese Tone Structure*. Cambridge, MA: MIT Press.

学習者が犯す誤用の要因・背景からみる日本語作文支援

八木 豊（株式会社ピコラボ）[1]
ホドシチェク・ボル（東京工業大学）
阿辺川 武（国立情報学研究所）
仁科 喜久子（東京工業大学）

Relevance of Learners' Errors in the Development of a Japanese Writing Support System

Yutaka YAGI (Picolab Co., Ltd.)
Bor Hodošček (Tokyo Institute of Technology)
Takeshi ABEKAWA (National Institute of Informatics)
Kikuko NISHINA (Tokyo Institute of Technology)

１．はじめに

近年、国立国語研究所による「現代日本語書き言葉均衡コーパス」（以後 BCCWJ）をはじめとする日本語大規模コーパスの開発が進展し、オンラインのコーパス検索ツールとしての「中納言」、「少納言」、「NINJAL-LWP for BCCWJ」によって、特定の語の頻度や共起関係、文法的な振る舞いなどを知ることができるようになり、日本語の研究者には大きな恩恵をもたらした。また日本語教育の分野でも、日本語教育の研究者や教師によるこれらのコーパスやツールを利用した教育方法や教材開発の動きがみられるようになってきた。仁科他(2011)、Hodošček 他(2011)による日本語作文支援システム「なつめ」[2]の開発もその一つであり、文書作成時に表現したい共起語の検索と例文参照を可能にした。しかしながら、このシステムは上級レベルの一部の学習者を除いて、利用するには困難な点が多い。例えば、単語の表記を正しく習得してないと検索できない、提示される例文は学習者の日本語能力に応じたレベルに絞り込まれていないなどの問題があるためである。そこで、さらに広範囲の学習者にも容易に利用できるシステムを目指し、学習者作文コーパス「なたね」[3]を構築し、そこに見られる学習者の犯しやすい誤用を分析し、その誤用の要因や背景を知ることで、学習者が入力した文の誤用を自動的に指摘して修正案を示すシステムを最終目標とすることとした。

２．学習者作文コーパス「なたね」

「なたね」は、我々が独自に収集した学習者作文に対して日本語教師による添削を行った誤用タグ付きデータである。誤用タグは大きく「誤用の対象」、「誤用の内容」、「誤用の要因・背景」という 3 つの視点から構成しており、さらにそれぞれを 3 階層に細分類することで、全体として約 70 種類を定義している（曹他(2012)）。2012 年 12 月現在、大学院や大学あるいは語学学校に在籍する 192 人の日本語学習者による 285 作文（総文字数 205,520

[1] yagi@picolab.jp
[2] 日本語作文支援システム「なつめ」 http://hinoki.ryu.titech.ac.jp/natsume/
[3] 学習者作文コーパス「なたね」 http://hinoki.ryu.titech.ac.jp/natane/

表1 母語別「学習者数」

母語	男性	女性	性別未入力	計
中国語	50	43	22	115
マラーティー語	6	23	7	36
ベトナム語	6		7	13
韓国語	6	1	4	11
スペイン語	2			2
マレー語	1			1
スロベニア語	1			1
ハンガリー語	1			1
タイ語			1	1
母語未入力	1		10	11
計	74	67	51	192

表2 母語別「作文数」

母語	男性	女性	性別未入力	計
中国語	62	64	26	152
マラーティー語	6	23	7	36
ベトナム語	18		9	27
韓国語	24	3	7	34
スペイン語	2			2
マレー語	8			8
スロベニア語	7			7
ハンガリー語	1			1
タイ語			1	1
母語未入力	5		12	17
計	133	90	62	285

字）に含まれる約 6,500 箇所の誤用に対しておよそ 9,000 件の誤用タグを付与して公開している[4]。収集した作文は、PC 入力と手書きの区別、辞書使用の有無や時間制限などのコントロールを行っておらず、作文のテーマも、自己紹介からエッセイ風のものまで様々である。作文データそのもの以外に、性別、国籍、母語、学習歴、日本語能力（日本語能力試験のレベルや日本語教師による主観評価）といった学習者のメタ情報も可能な範囲で併せて収集しており、作文を公開するにあたっては、複数の日本語教師の協力のもとに本人の承諾を得ることができた情報のみを公開している。「なたね」における母語別の学習者数および作文数を表１、表２に示す。作文を収集できる環境が限られていることから現状では中国語を母語とする学習者が多く、全体の半分以上を占めている。

３．「誤用の要因・背景」の分析

本章では、「なたね」に付与した誤用タグのうち「誤用の要因・背景」に着目して、学習者が犯しやすい誤りの傾向、学習者の母語や日本語能力といったメタ情報との関連について分析を行う。表３は「誤用の要因・背景」に含まれる誤用タグの頻度を母語別に集計した結果である。表見出しのアルファベットは学習者の母語を表しており（脚注参照）、それぞれの列がその母語における誤用タグの頻度、右端の列が「なたね」全体の頻度である。以降では、「誤用の要因・背景」に含まれる誤用タグの項目「類似」「母語干渉」「レジスター」を取り上げ、順を追って説明する。

３．１．類似

類似した語句との混同が要因となっている誤用が該当し、類似している内容に応じて、意味の類似、字形の類似、音の類似の 3 つに下位分類している。それぞれについて代表的な誤用例を以下に挙げる。矢印の左側の下線部が誤用箇所、矢印の右側の斜体が日本語教師による訂正例で、末尾の括弧内には学習者の母語を記した。

【意味の類似】成長についてだんだん深く<u>了解</u>→*理解*できた。（中国語）
【字形の類似】<u>公島</u>→*広島*と東京とおきなわを見たいです。（マラーティー語）
【音の類似】これは私のしょうらいの<u>ゆうめい</u>→*ゆめ*です。（マラーティー語）

意味の類似では、特に日本語で用いられるある漢語の意味が中国とは異なる意味で用い

[4] 総文字数には句読点やその他の補助記号も含む。ただし、現在もメンテナンスを継続しており、Web サイト上での表示はここで挙げた数値と一致しないことがある。

表3 誤用の要因・背景[5]

項目	zh	mr	vi	ko	es	ms	sl	hu	th	未	計
意味の類似	38	141	11	32		2	7		2	10	243
字形の類似	2	47	1	4		2	1				57
音の類似	7	110	1	10		3	2		1		134
母語干渉	45	6	1	5				1			58
レジスター	384	12	8	46	9		2	4		18	483
文体の不統一	411	21	10	10	9			3		14	478
その他	12	3	1	3						2	21
計	899	340	33	110	18	7	12	8	3	44	1474

られることがしばしばある。例えば日本語で「理解」と表現する場合に、中国語では「了解」と表現することができる。このような場合、学習者は日本語のコンテクストの中に母語の意味と合致する語を挿入してしまう。日本語において「理解」と「了解」は意味的に類似してはいるが使い分けが必要であることから「(漢語の) 意味の類似」という誤用タグを付与している。この例は、中国語からの母語干渉と重なるものである。

類似に関する誤用の中で字形の類似や音の類似では、マラーティー語を母語とする学習者の誤用が著しく多くなっている。これは、マラーティー語では、作文を収集した多くが日本語レベル初級の学習者で平仮名・片仮名の読み書きも不十分であることに加えて、原則としてパソコンなどを使用せず、手書きの作文を収集したことで余計に顕著な傾向が現れたためといえる。実際は字形の類似と音の類似は相互的であり、どちらによるものかの判定は困難である。例えば、マラーティー語話者の作文中に「首で走いて→*道を歩きながら*」という誤用がある。「首」は「道」という字形の誤り、「走」は「歩」の字形の誤りである。直接学習者にインタビューできないため判定は推測によることになるが、音声では「みち」「あるいて」と認識していると思われる。上級者で日本語の音声を正確に習得していない場合があっても、漢字表記では音声習得の不正確さは顕在化しないが、非漢字圏初級学習者は、仮名表記をすることで音の類似による誤用が顕著になっている。

その他の母語については中上級の学習者で構成されており、字形の類似や音の類似による誤用はほとんど見られなくなる。意味の類似による誤用については、日本語レベルが上がっても中国語や韓国語といった漢字圏の学習者を中心に散見されることと対照的である。

３．２．母語干渉

中国語を母語とする学習者による熟語の誤用など、学習者の母語の影響に因ると考えられる誤用が該当する。類似の場合と同様に、代表的な誤用例を以下に挙げる。

【母語干渉】十月一日午後、わたしたちは 4 時の火車→*汽車*に乗って、…(中国語)
【母語干渉】この場合は更生された→*更生した*人間ならば例外にしたいと思う。(韓国語)

母語干渉は、コーパス全体でも 58 件と少ないうえに、そのうちのおよそ 8 割は中国語を母語とする学習者による漢字選択の誤りである。これは、中国語を母語とする学習者の割合が多いこともあるが、日本語教師が添削する際に母語干渉であると判断できる内容は、漢

[5] zh：中国語、mr：マラーティー語、vi：ベトナム語、ko：韓国語、es：スペイン語、ms：マレー語、sl：スロベニア語、hu：ハンガリー語、th：タイ語、未：母語未入力

字圏の学習者による漢字選択の誤りに限定されやすいためではないかと考える。その他は、前述の 2 つ目に挙げた誤用例のように、韓国語を母語とする学習者が自動詞に「される」をつける誤りが 2 件ほど含まれている以外に、母語干渉と一概には言えないものも含まれており、今後、タグ付けした日本語教師への確認および必要ならば修正を行う予定である。

３．３．レジスター

機能文法では言語表現の異なりを「社会的な拘束力をもつ言語学上の規範」における言語使用域の変異即ち「レジスター」と呼び、Halliday(2004)はレジスター機能として次の 3 項目を挙げている。

（1）コミュニケーションの目的と主題に関わる「フィールド」（Field of discourse）
（2）コミュニケーションを行うための手段に関わる「モード」（Mode of discourse）
（3）コミュニケーションパートナー同士の関係に関わる「テナー」（Tenor of discourse）

書き手と話し手がどのような関係で、どのようコンテクストのもとで発話するかによって、それぞれ異なる語彙・文法項目で記述されることを示すものである。

学習者作文においては、授業で提出するレポート内で話し言葉を使用しているなど「場」にそぐわない表現全般がレジスターの誤りに該当する。現時点でレジスターに関する誤りのタグは 483 件あるが、「話し言葉と書き言葉」の違いによるものが大部分である。類似の場合と同様に、代表的な誤用例を以下に挙げる。

【レジスター1】少子化のせいで→*ために*、これから日本人の労働者がだんだん→*次第に*少なくなります。（ベトナム語）
【レジスター2】文章を読んでいるとき、とても苦しいですね。ときどき意味はちゃんと→*十分に*理解できないこともありますよ。（中国語）
【レジスター3】女性たちも経済的に力を持ち始め、徐徐に平等に向けての運動をやり始めた→*始めた*。（韓国語）

レジスター1 の例では、理由や原因を示す接続表現に主観的な意味を含む「せい」を用いている。アカデミックな文章では判断に感情表現を含ませるのは不適切であり、レジスターの誤りと判断される。「だんだん」は話し言葉であるため、書きことばの表現に修正案が示されている。

レジスター2 の例は、初級会話で学習した終助詞が使用されている作文である。日本語の終助詞は、コミュニケーション相手の同意を求めるために有効な表現であるが、アカデミックな文章ではこの種の表現を使用しないことを習得していない例である。

レジスター3 の例は、「（運動を）やり始める」という動詞が話し言葉のなかでもくだけた表現となっている。他にも次のようなくだけた表現の作文がみられる。これらの表現は、初級教科書でも現れないものであり、日本留学後のコミュニケーションを通して教室外で習得した表現と推測される。

【レジスター4】「しかし、伝統的な習慣とか→*など*では女性が不平等な目に会うことがいまだにも多く残っている。」（韓国語）
【レジスター5】くじ引きで日本語クラスに入り日本語を勉強し始りました。うちの→*私たちの*クラスで 10 人がアメリカの大学に入学して、他の 20 人は全部日本に来ました。（中

国語）

レジスターの誤りは、前述の類似の誤りとは反対に初級の学習者であるマラーティー語母語話者にはほとんどみられなかった。これは、初級学習者がレジスターを使い分けるに至っていない点にある。初級で教えられる語彙および教材の構成からみると、おおむね話し言葉が優先的に導入される。そのため、日本語教師のほうで初級学習者に対してはそこまでチェックせず表記の誤りなどその他の添削を優先するということが、日本語教師へのインタビューから明らかになった。

レジスターが問題になるのは、このようなシラバスで学んできた学習者が中級から上級に至った段階で、レポートなどのアカデミックな文章を書く必要性が生じる場合である。アカデミックな文章では、学習者は話し言葉と書き言葉を区別して書き分けなければならないほか、作文全体を通して文体の統一も図らねばならない。次の例は、作文中での文体の不統一による誤用である。作文全体の中で文末の「真の鍵でしょう」の部分のみが丁寧体となっている。「である」の推量形がわからないために「でしょう」にしたと推測できる誤用例が、他の学習者の作文にも散見する。

【文体の不統一】現状がかえない、どうしも真の先進国にならない。女性の社会進出は先進国に真の鍵でしょう→*であろう*。（中国語）

以上のようなレジスターの不整合としての誤用例は、話し言葉による会話場面を中心とする初級の教材での学習内容を習得した後で、文章を書く段階に入って、書き言葉のレジスターの知識が不足しているためと考えられる。現時点では、このような区別をレジスターの異なりとして体系的に教える教材はほとんどなく、アカデミックな表現が必要な上級レベルの学習者に対する教材やコースウエアへの対応が十分でないと推測できる。

４．まとめと今後の課題

本稿では、作文を支援するシステムを上級者のみでなく広範囲の学習者にも容易に利用できるシステムを目指し、学習者作文コーパス「なたね」を構築し、自動校正システムを最終目標として、そこに見られる学習者の犯しやすい誤用を分析し、その誤用の要因や背景を考察した。

誤用の要因と背景を分析するために、「なたね」に収録されている「意味の類似」「字形の類似」「音の類似」「母語干渉」「レジスター」の誤用例を観察し、考察した結果、以下のような結論を得た。

(1) 「字形の類似」「音の類似」による誤りは、非漢字圏初級学習者の例に多く見られた。語の表記と音声理解は相互的なものであり、どの母語の学習者にも誤った理解はあるが、特に非漢字圏初級学習者は漢字表記にハンディキャップがあるため、仮名表記を使用することで音声理解の誤りが顕在化していると考えられる。

(2) 「意味の類似」による誤りの中で漢字圏学習者によるものは、母語における漢語の意味と日本語における意味の異同によって誤ることがあり、母語干渉の影響もあると考えられる。

(3) 「母語干渉」は、語の意味の類似によるものが多く見られ、構文的なものもわずかであるが見られた。

(4) 「レジスター」の誤用は、初級レベルではほとんどタグが付けられていない。その理由は、初級学習者の語彙、表現の学習範囲が話し言葉中心であり、レジスターの違いを示すバリエーションがないことから、誤用としてタグを付けられないためである。一方、上級者では、話し言葉によって学んだ日本語の知識で、アカデミックな文章を書く段階になって、レジスターの知識が不十分であるために、不適切な表現が散見されることになる。文体の不統一についても文法的な知識の不足が影響している部分があると考えられる。

上級学習者は初級で学んだ話し言葉に加えて、アカデミックな書き言葉、さらに高度なフォーマルな話し言葉、手紙などのフォーマルな書き言葉表現など様々なバリエーションを習得する必要が生じてくる。これらの表現を教室の授業だけで学ぶには、時間的制限もあり、習熟することは困難である。

我々の今後の課題としては、さらに学習者データを追加し、不適切な表現を分析することで、学習者に必要な適切な文章表現の提示を可能にするシステムを目指す必要がある。

謝辞

本研究は、文部科学省科学研究費補助金基盤研究（C）「日本語作文支援システムで考慮すべき学習者属性情報と提示項目の分析研究」(研究代表者：阿辺川武、研究期間：2012 年 4 月～2015 年 3 月）および同補助金挑戦的萌芽研究「日本語学習者誤用コーパスを利用した作文システムの開発」（研究代表者：仁科喜久子、研究期間：2010 年 4 月～2013 年 3 月）による助成を得て実施しています。

参考文献

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.

Hodošček Bor、阿辺川武、Bekeš Andrej、仁科喜久子（2011）「レポート作成のための共起表現産出支援―作文支援ツール「なつめ」の使用効果―」専門日本語教育研究 13 号、pp.33-40.

曹紅荃、八木豊、黒田史彦、仁科喜久子（2012）「学習者コーパス「なたね」の構築と応用の可能性」第 5 回「日本語教育とコンピュータ」国際会議（Castel/J）

Halliday M.A.K. and C.M.I.M. Matthiessen (2004). An Introduction to Functional Grammar. 3rd ed. London: Arnold

仁科喜久子監修（2012）「日本語学習支援の構築　言語教育・コーパス・システム開発」凡人社

八木豊、ホドシチェク・ボル、仁科喜久子（2012）「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－」第 1 回コーパス日本語学ワークショップ予稿集、pp.315-320.

近代女性向け雑誌記事における一人称代名詞の分析
—形態論情報付き『近代女性雑誌コーパス』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 Published in the Modern Women's Magazines: An Analysis of Morphologically Annotated *Modern Women's Magazines Corpus*

KONDO, Asuko (Center for Corpus Development, NINJAL)

１．はじめに

これまでの近代語の一人称代名詞に関する研究では、小説の会話部分、落語速記、口語文典といった話し言葉的性質の強い口語文を主に分析の対象としてきた[1]。一方で、雑誌『太陽』（1895～1928 刊）に基づく『太陽コーパス』（国立国語研究所（編）、2005）を用いた分析から、当時、一人称代名詞は話し言葉的性質の強い口語文のみに出現するのではなく、例えば論説文のような、書き言葉的性質の強い文章にも多く出現し、その使用実態は話し言葉的性質の強い口語文のそれとは大きく異なることも明らかにされつつある（近藤、2008・2011・2012）。本稿では、『太陽コーパス』の比較資料として作成された『近代女性雑誌コーパス』（国立国語研究所、2006）に形態論情報を新たに付与したデータを利用して、明治後期から大正期にかけて刊行された女性向け雑誌に出現する一人称代名詞を網羅的に抽出し、分析対象とする。そして、一人称代名詞の語形と文章の種類との対応関係や、語形と著者性別・文体との対応関係の点から、文語文体から口語文体へ文体が大きく変化した時期における雑誌記事での一人称代名詞の使用実態について、その一部を明らかにする。

２．形態論情報付き『近代女性雑誌コーパス』

『近代女性雑誌コーパス』は、明治後期から大正期にかけて刊行された女性向け雑誌に基づくコーパスである。『女学雑誌』（1894～1895 年刊行分 31 冊）、『女学世界』（1909 年刊行分 6 冊）、『婦人倶楽部』（1925 年刊行分 3 冊）の計 40 冊 1362 記事が収録されており、『太陽コーパス』と比較させながら、当時の女性が読んでいた書き言葉の実態を把握することが可能な資料となっている（田中、2006）。この『近代女性雑誌コーパス』を利用することで、雑誌『太陽』の主な読者層からは外れていた女性を対象とする雑誌記事での一人称代名詞の使用実態の一部が明らかになると考える。また、『太陽コーパス』の掲載記事のほとんどが男性の著したものであるのに対し、『近代女性雑誌コーパス』の掲載記事は女性の著したものが多く含まれる。よって、『近代女性雑誌コーパス』から当時の一人称代名詞使用の男女差について明らかになることも期待される。

ただし、公開されている『近代女性雑誌コーパス』には、形態素解析による形態論情報が付与されていない。これはコーパス開発当時、近代語を含めた古い時代の日本語資料に

† kondo@ninjal.ac.jp

[1] ある程度の年代にわたる複数の資料を対象に複数の一人称代名詞の分析を行った先行研究として、岡田（1998）・祁（2006a）（2006b）・那須（1986）・房（2004）などがある。

ついて、実用的な精度で形態素解析することが実現されていなかったためである。しかし、近年になり近代の文語論説文を対象とする形態素解析辞書「近代文語 UniDic」（小木曽、2009）や旧仮名遣いの口語文を対象とする形態素解析辞書（小木曽、2012）が開発されるなど、近代語の資料も実用的な精度で形態素解析できる環境が整ってきた。国立国語研究所の形態論情報データベース（小木曽・中村、2011）には、これらの辞書を用いて形態素解析したデータが格納されている。本稿ではこのデータベースの 2013 年 1 月時点のデータに基づき、分析・考察を行う。

なお、形態論情報から得られる『近代女性雑誌コーパス』の延べ語数（記号類除く）は 1,265,905 語である。

３．一人称代名詞の抽出

漢文・欧文部分[2]を除いた『近代女性雑誌コーパス』全体から、次の①～③の手順で一人称代名詞を抽出した。

① 形態論情報で品詞が「代名詞」となっている見出し語を抽出する。
② ①で抽出された見出し語のうち、一人称代名詞として用いられるものを選択する[3]。
③ ②で選択した見出し語について、コーパスのルビ情報や文脈を参照し、解析誤りの修正を手作業で行う 。また、一人称代名詞の表記に用いられる主な文字列と同じ出現形をとる代名詞以外の見出し語についても、一人称代名詞と新たに見なされるものは修正を行う。さらに、出雲（2004）で明治の女性の論説的文章に用いられる一人称代名詞として挙げられている語形のうち、コーパスの形態論情報では代名詞とされていなかった「妹（まい）」「小妹（しょうまい）」について、新たに文字列検索を行い、一人称代名詞と見なされるものは修正を行った。

この手順により抽出された一人称代名詞の語形とそれぞれの粗頻度と出現記事数を示したものが表 1 である。1 記事に複数の語形が出現する場合、出現記事数はそれぞれの語形で重複してカウントした（表 2 以下も同様）。27 種類の語形が得られ、一人称代名詞全体の粗頻度は 4,896 語となった。近藤（2012）の『太陽コーパ

表１　一人称代名詞の粗頻度

語形	粗頻度	出現記事数
わたし	2,025	182
わたくし	930	206
余	614	111
僕	432	52
吾人	201	66
おれ	106	30
わらわ	96	12
我々	80	40
余輩	80	15
それがし	57	18
わし	52	14
妾(しょう)	45	16
あたし	42	8
我輩(わがはい)	41	29
拙者	33	3
妹(まい)	16	2
てまえ	10	7
おいら	8	5
わなみ	6	1
おら	4	4
あたい	5	4
小妹(しょうまい)	5	2
やつがれ	2	2
わて	2	1
あっし	1	1
うら	1	1
わい	1	1
一人称代名詞全体	4,894	555

[2] 漢文・欧文部分の抽出は、コーパスの引用タグの種別属性値を使って行った。『女性雑誌コーパス』の XML タグの仕様は『太陽コーパス』のものに準拠する。『太陽コーパス』の XML タグの仕様については田中（2005）を参照のこと。

[3] 「われ」のように、一人称代名詞だけでなく二人称代名詞・反射指示代名詞といった他の用法でも多く用いられる見出し語については本稿では考察対象外とした。

ス』での抽出結果と比較すると、『太陽コーパス』に出現せず『女性雑誌コーパス』のみに出現する語形として「妾・妹・てまえ・あたい・小妹・やつがれ・わい」の 7 語形があげられる。そのうち「妾・妹・あたい・小妹」の 4 語形は女性専用と言える一人称代名詞であり、女性向け雑誌ならではの出現傾向となっている。

4．文章の種類と一人称代名詞との対応関係

次に、表 1 でとりあげた 27 語形について、文章の種類との対応関係について見ていく。ここでいう文章の種類とは、その書き言葉・話し言葉的性質の強弱により分類するものである。まず、書き言葉的性質の強い文章の代表として「非文学作品である記事の地の文」を選ぶ。『女性雑誌コーパス』の扱う時期は言文一致の完成する時期に一致し、地の文も文語文体から口語文体に大きく変化する。このことから、「非文学作品である記事の地の文」を文語文体のものと口語文体のものにさらに分けることにする。反対に話し言葉的性質の強い文章の代表として、「文学作品の会話部分」を選ぶ。本稿の分析の観点とする文章の種類を改めてあげると、（ア）文語地の文（非文学）、（イ）口語地の文（非文学）、（ウ）口語会話（文学）の 3 種類となる。文章の種類別の本文の抽出は、コーパスに XML タグによって付与された各種情報に基づいて行った[4]。

この 3 種の文章の種類ごとに、各語形の粗頻度、出現記事数、およびコーパス中で該当種類の文章を含む全記事数を示したものが表 2 である。値が 0 の場合、空欄で示した（表 3 以下も同様）。（　）内の

表 2　一人称代名詞の文章の種類別粗頻度・出現記事数

語形	文語地の文		口語地の文		口語会話	
	粗頻度	出現記事数	粗頻度	出現記事数	粗頻度	出現記事数
わたし	2	1 (0.2)	457	67 (25.3)	462	52 (29.4)
わたくし	13	3 (0.5)	399	60 (22.6)	89	27 (15.3)
余	258	52 (9.0)	6	4 (1.5)	4	3 (1.7)
僕	1	1 (0.2)	15	7 (2.6)	289	29 (16.4)
吾人	163	46 (8.0)	7	4 (1.5)		
おれ			1	1 (0.4)	75	19 (10.7)
わらわ	2	1 (0.2)	1	1 (0.4)		
我々	7	5 (0.9)	25	12 (4.5)	25	8 (4.5)
余輩	60	10 (1.7)				
それがし	46	13 (2.3)				
わし					45	8 (4.5)
妾(しょう)	6	4 (0.7)				
あたし			2	1 (0.4)	12	6 (3.4)
我輩(わがはい)	26	17 (3.0)	3	3 (1.1)		
拙者					33	3 (1.7)
妹(まい)	16	2 (0.3)				
てまえ			1	1 (0.4)	6	4 (2.3)
おいら					7	4 (2.3)
わなみ						
おら					3	3 (1.7)
あたい					4	3 (1.7)
小妹(しょうまい)	1	1 (0.2)				
やつがれ						
わて					2	1 (0.6)
あっし						
うら						
わい					1	1 (0.6)
コーパス全体記事数		575 (100.0)		265 (100.0)		177 (100.0)

[4] 地の文は、引用タグによってマークアップされていない部分とした。会話部分は、種別属性値が「会話」の引用タグによってマークアップされている部分とした。文学作品・非文学作品の区別は、記事タグのジャンル属性値の NDC 番号の左 2 桁が 91～99 および 9X となっているタグでマークアップされている部分を文学作品とし、それ以外の部分を非文学作品とすることで行った。文語・口語の区別は、記事タグおよび引用タグの文体属性値に基づき行った。

値はコーパス全体記事数に対する該当語形の出現記事数の割合（単位%）である。

語形と文章の種類との対応関係を見るために、表 2 の（ ）内の値を用いてコレスポンデンス分析[5]を行った。粗頻度ではなく出現記事数に基づく値を用いるのは、一人称代名詞は特定の記事に集中して用いられる傾向があり、粗頻度を用いた分析ではその特定の記事の傾向に分析結果が左右されると考えたためである。分析結果から第1次元（寄与率80.26%）と第 2 次元（寄与率 19.74%）の得点を 2 次元空間上に布置したものが図 1 である。

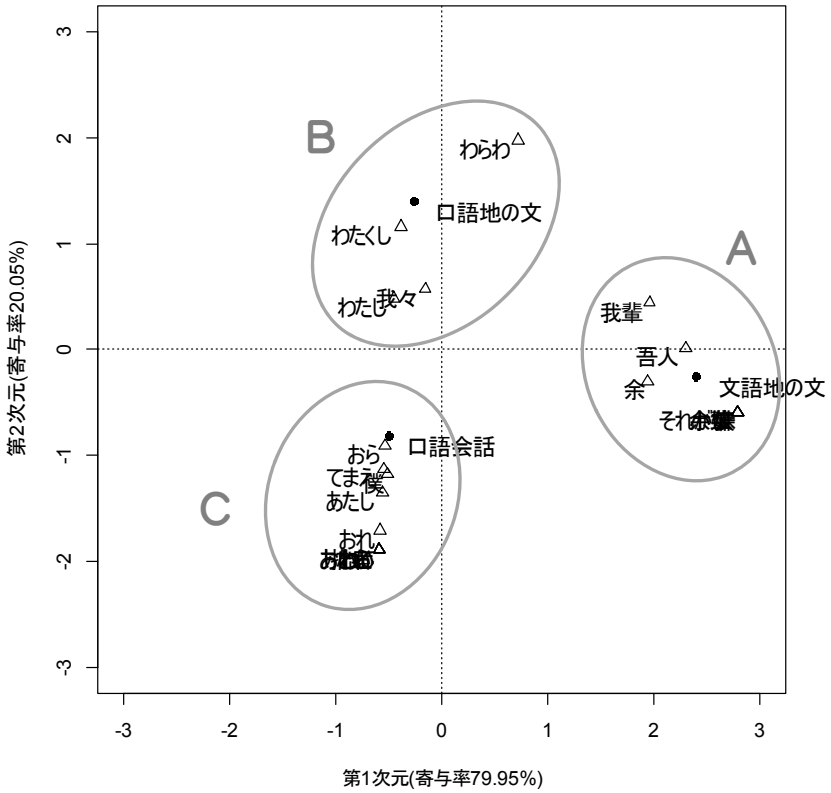


図 1　一人称代名詞の語形と文章の種類の散布図

図 1 から、一人称代名詞の語形と文章の種類との間に強い対応関係が確認でき、文章の種類との親疎関係から語形を次の A～C の 3 グループに分けることができる。文語地の文に近い A グループの語形が最も書き言葉的性質が強く、B・C の順に書き言葉的性質が弱く話し言葉的性質が強くなり、口語会話に近い C グループの語形が最も話し言葉的性質が強いと見なされる。

A　「吾人・妾・小妹・それがし・妹・余・余輩・我輩」
　…文語地の文で主に用いられ、口語地の文・口語会話ではほとんど用いられない
B　「わたくし・わたし・わらわ・我々」
　…口語地の文・口語会話で主に用いられ、文語地の文ではほとんど用いられない
C　「あたい・あたし・おいら・おら・おれ・拙者・てまえ・僕・わい・わし・わて」
　…口語会話で主に用いられ、文語地の文・口語地の文ではほとんど用いられない

地の文での一人称代名詞の使用実態は口語会話とは大きく異なることになり、一人称代

[5] 分析には統計分析ソフト R の MASS パッケージの corresp 関数を用いた。

名詞全般について考察するためには、分析対象として口語会話だけでなく地の文も含める必要があることが改めて確認された。

５．文語地の文の一人称代名詞

ここからは、地の文に出現する一人称代名詞について特に取り上げ、その使用実態についてより詳しく分析・考察する。まず、文語地の文に出現する一人称代名詞を取り上げる。

その前提として、コーパス全体における文語地の文を持つ非文学記事（以下、「文語記事」）の実態を見ていく。表 3 は、文語記事の数量を刊行年・著者性別・記事文体ごとに示したものである。1894 年刊行分の記事については 1895 年刊行分と併せて集計した。著者性別はコーパスに同梱されている著者リスト（authors.xml）に挙げられている著者について、その著者名や記事内容から判断し、「男」「女」「不明」の 3 種に分類した。無署名および複数著者の記事はすべて著者性別を「不明」とした。記事文体とは記事中で使用されている文末辞の種類によって分類するものを言う。ここでは文末辞に「候（そうろ）う」を用いる候文か否かという観点から記事文体を分類した。地の文に動詞「候う（ソウロウ）」が 1 回以上出現する文語記事 35 記事のうち、記事の一部分のみが候文になっているものや前代の著作であるものを除いた 13 記事を「候文」とし、それ以外の記事を「非候文」に分類した。

表 3　刊行年・著者性別・記事文体別の文語記事数

	1895		1909		1925		通年	
	非候文	候文	非候文	候文	非候文	候文	非候文	候文
男	243	4	1				244	4
女	6	3	6	3			12	6
不明	246	3	59		1		306	3
小計	495	10	66	3	1		562	13
合計	505		69		1		575	

刊行年ごとの合計記事数の経年変化を見ると、1895 年に 505 記事あった文語記事の数は年を追うごとに急激に減少し、1925 年に至って 1 記事しかない。これは、文語文体から口語文体へ文体の基調が大きく変化した当時の書き言葉のありようと相関した変化と言える。著者性別ごとに記事数の経年変化を見ると、1895 年は男性が著者の記事がほとんどを占め、女性が著者の記事はわずかであるのが、1909 年は男性が 1 記事に対し女性が 9 記事と記事数の多寡が逆転していることが分かる。

また、記事文体と著者性別との対応関係を見るために、通年での値による男女別・記事文体別の 2×2 クロス表でフィッシャーの正確確率検定[6]を行ったところ p=.0000 となり、1% 水準で非候文は男性が有意に多く、候文は女性が有意に多いことが確認された。つまり、女性は「候う」を用いて読み手への配慮を示す文体を多用する傾向があることになる。

以上の文語記事全体の傾向を踏まえ、文語地の文に出現する一人称代名詞について見ていく。著者が男性または女性の記事について刊行年・記事文体別に一人称代名詞の語形ごとの出現記事数を示したものが表 4 である。1925 年の文語記事には一人称代名詞が出現しなかったので表中に載せていない。

[6] 検定には R の fisher.test 関数を用いた。

表 4　文語地の文における一人称代名詞の著者性別・刊行年・記事文体別の出現記事数

男性

グループ	語形	1895		1909		通年		合計
		非候文	候文	非候文	候文	非候文	候文	
A	余	34				34		34
	吾人	28				28		28
	我輩	11				11		11
	余輩	9				9		9
	それがし	9				9		9
B	我々	2				2		2
C	僕	1				1		1

女性

グループ	語形	1895		1909		通年		合計
		非候文	候文	非候文	候文	非候文	候文	
A	妾		1				1	1
	小妹		1				1	1
	妹		1				1	1
	余	1				1		1
B	我々	1			1	1	1	2
	わたくし		1		1		2	2

まず男性による記事を見ると、非候文のものにのみ一人称代名詞が出現する。この傾向は、表 3 にある男性による記事の文体別数と相関した結果と考えられる。出現する語形は A グループの「余・吾人・我輩・余輩・それがし」を中心とし、B グループの「我々」もわずかに出現する。以下に用例をあげる。

(1) 余は先般園照マクレーン事件裁判の判決書を英國に申遣はし置きしを以到着次第貴社にも贈る可し（1984 年 29 号「園照女に関して　其一」佐伯理一郎）[7]

(2) 何［なに］を以［もつ］て然［し］か言［い］ふと問［と］ふものあらば、吾［ご］人［じん］は先［ま］づ「女［ぢよ］學［がく］生［せい］が家［いへ］を成［な］して評［ひやう］判［ばん］あしき所以［ゆゑん］、何處［いづこ］にありや」と反［はん］問［もん］せん。（1894 年 31 号「女生徒の卒業と婚嫁」巌本善治）

なお、C グループの「僕」が 1 記事に出現するが、これは小松（1999）に言う漢文で使用される聞き手への敬意の強い「僕」に通じるものであり、当時の口語会話に出現する「僕」とは性質の異なるものである。

次に女性による記事を見ると、まず「妾・小妹・妹」という女性専用の語形が出現する点が大きな特徴としてあげられる。一方で、「余」のように主に男性が用いるとされる語形を女性も用いる場合があることも分かる。以下に例をあげる。

(3) 小妹事茲に貴欄を愛讀せらるゝ御婦人方に對し數言の祝辭を呈し度御受納の上貴白に御掲載被下候はゞ幸甚に存じ候。（1895 年 2 号「米国ハリス夫人の寄書」フロラ、ビー、ハリス）

(4) 本篇の批評につきては、未だ一冊にまとまりて出板にならぬ先きより、既に諸大家の評やかましければ更に予の拙評を加ふるの要を見ず。（1895 年 8 号「新刊書」磯松まつ子）

また、男性による記事とは異なり、A グループの「妾・小妹・妹」や B グループの「わたくし」のように候文の記事に出現する語形が認められる。この傾向は、表 3 にある女性による記事の文体別数と相関した結果である可能性がある。ただし、女性による記事や候

[7] 用例の引用に際し、［　］内にルビを示し、末尾の（　）内に刊行年・号数・記事題名・記事著者を示す。

文の記事の数量自体が少ないため、明確なことは言えない

語形別に見ると、A グループのうち「余・吾人・我輩・余輩・それがし」は非候文にのみ出現するのに対し、同じ A グループでも「妾・小妹・妹」は候文にのみ出現する。B グループでは「我々」が非候文のみに出現し、「わたくし」は候文のみに出現する。同じグループ内でも記事文体に対応した使い分けがあったと見られるが、これについても、出現記事数の少ない語形については明確なことは言えない。文語地の文での語形と著者性別と記事文体との対応関係については、今後対象資料を広げて分析を重ねる必要があるだろう[8]。

６．口語地の文の一人称代名詞

次に、口語地の文での一人称代名詞の使用実態についてより詳しく見ていく。

その前提として、コーパス全体における口語地の文を持つ非文学記事（以下、「口語記事」）の実態を見るため、その数量を刊行年・著者性別・記事文体ごとに表 5 に示す。刊行年と著者性別の分類については表 3 と同様の処理をした。記事文体は、ここでは常体か敬体かという観点から分類した。助動詞「です」「ます」の出現回数が 0 回の記事の文体を「常体」、助動詞「です」「ます」の出現回数が 1 回以上で、かつ動詞「御座る（ゴザル）」＋助動詞「ます」の出現回数が 0 回の記事の文体を「敬体」、動詞「御座る（ゴザル）」＋助動詞「ます」の出現回数が 1 回以上の記事の文体を「敬体（ございます）」として分類した。

表 5　刊行年・著者性別・記事文体別の口語記事数

	1895			1909			1925			通年		
	常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)
男	1		1	11	16		11	21	4	23	37	5
女				8	15	47		9	9	8	24	56
不明		1		19	9	16	16	45	6	35	55	22
小計	1	1	1	38	40	63	27	75	19	66	116	83
合計	3			141			121			265		

刊行年ごとの合計記事数の経年変化を見ると、1895 年には記事数 3 と同年の文語記事数 505 に比較してもわずかであったのが、1909 年には 141 と急増し、同年の文語記事数 69 を上回るまでになる。1925 年は 121 と 1909 年より記事数が減少するが、同年の文語記事数は 1 であり、非文学記事のほとんどは口語記事であることになる[9]。文語文体から口語文体へ文体の基調が大きく変化した当時の書き言葉のありようと相関する変化と言える。著者性別ごとに記事数の経年変化を見ると、1895 年は女性が著者の記事がないが、1909 年では男性 27 記事に対して女性 70 記事と女性による記事数のほうが多くなる。しかし、1925 年には男性 36 記事に対して女性 18 記事と男性による記事数のほうが多くなる。

また、記事文体と著者性別との対応関係を見るために、通年での値による記事文体別・男女別の 3×2 クロス表に χ^2 検定[10]による多重比較（ボンフェローニ法）を行ったところ、「常体」対「敬体」で p=.0925、「常体」対「敬体（ございます）」で p=.0000、「敬体」対「敬体（ございます）」で p=.0000 となり、1%水準で常体および敬体は男性で有意に多く、敬体

[8] 出雲（2004）は、清水紫琴・若松賤子らの論説的文章に出現する一人称代名詞について主に『女学雑誌』で調査し、「吾人・余・余輩・わらわ・妾・小妹・妹・われ・わなみ・わたし・わたくし」等の出現を報告する。ただし、各語形の候文・非候文文語・口語別の頻度は示されていない。

[9] 田中（2006）によれば、著作権の事情から 1925 年分は 1909 年分より公開できた分量が少なく、さらに 1925 年は 1909 年より文学ジャンルの占める割合が高くなっている。非文学の口語記事数が 1909 年より 1925 年で少なくなるのはこれらのことが主に影響していると考えられる。

[10] 検定には R の chisq.test 関数を用いた。

（ございます）は女性で有意に多いことが確認された。つまり、口語記事において女性は男性より読み手に配慮した丁重な文体を用いていることになる。文語記事で女性が候文を多用する傾向と併せて見れば、当時の女性は文語・口語を問わず、丁重さといった読み手への配慮がより強く表せる文体を選択する傾向にあったことになる。

以上の口語記事全体の傾向を踏まえ、口語地の文に出現する一人称代名詞について見ていく。著者が男性または女性の記事について刊行年・記事文体別に一人称代名詞の語形ごとの出現記事数を示したものが表 6 である。

表 6　口語地の文における一人称代名詞の著者性別・刊行年・記事文体別の出現記事数

男性

グループ	語形	1909			1925			通年			合計
		常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)	
A	吾人		1			3			4		4
	余	1	1					1	1		2
	我輩		2						2		2
B	わたくし	5	7					5	7		12
	わたし	2	3		2	13	1	4	16	1	21
	我々	2			3	4		5	4		9
C	僕	2	1		1	1		3	2		5

女性

グループ	語形	1909			1925			通年			合計
		常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)	常体	敬体	敬体(ございます)	
A	我輩		1						1		1
B	わたくし	1	6	29		1		1	7	29	37
	わたし	4	8	20		5	2	4	13	22	39
	我々	1						1			1
C	てまえ			1						1	1

まず、男性による記事を見ると、一人称代名詞の出現する記事は敬体が過半を占め、常体がそれに続き、敬体（ございます）はほとんどない。この傾向は、表 5 にある男性による記事の文体別数に相関した結果と考えられる。出現する語形は B グループの「わたくし・わたし・我々」を中心とし、A グループの「吾人・余・我輩」や C グループの「僕」も少数出現する。以下に例をあげる。

(5) まづ私［わたし］は、根［こん］本［ぽん］として、新［しん］夫［ふう］婦［ふ］は、當［たう］然［ぜん］別［べつ］居［きよ］すべきだと云［い］ふ説［せつ］を採［と］るものです。(1925 年 6 号「当然別居すべきもの」千葉亀雄)

(6) 先［せん］日［じつ］余［よ］が大［おほ］隈［くま］伯［はく］に遇［あ］つた時［とき］、伯［はく］は女［ぢよ］史［し］の事［こと］をヒドくほめて、吾［わが］黨［たう］の女［ぢよ］傑［けつ］ぢやと云［い］つてゐられたが、此［この］分［ぶん］では壽［じゆ］命［みやう］も伯［はく］同［どう］様［やう］百二十五歳［さい］迄［まで］は確［たし］かゝも知［し］れぬ。(1909 年 8 号「実践女学校」河岡潮風)

(7) 僕［ぼく］を女性［をんな］苛［いぢ］めの鬼［おに］でゞもあるやうに思［おも］ふてる女［ぢよ］學［がく］生［せい］諸［しよ］君［くん］もあるか知［し］らんが、女［をんな］に對［たい］しては却々［なか〜〜］親［しん］切［せつ］な男［をとこ］で女［ぢよ］學［がつ］校［かう］に教［けう］師［し］たること實［じつ］に十四［よ］年［ねん］、女［ぢよ］子［し］教［けう］育［いく］にかけては隨［ずい］分［ぶん］

の古［ふる］狸［だぬき］である。（1909 年 5 号「卒業生への注文」青柳有美）

次に女性による記事を見ると、一人称代名詞の出現する記事は敬体（ございます）が過半を占め、敬体がそれに続き、常体はわずかである。この傾向は、表 5 にある女性による記事の文体別数に相関した結果と考えられる。出現する語形は B グループの「わたくし・わたし」を中心とし、B グループの「我々」、C グループの「てまえ」もわずかに出現する。以下に例をあげる。

(8) 私［わたくし］は女［ぢよ］子［し］の獨［どく］身［しん］主［しゆ］義［ぎ］を絶［ぜつ］對［たい］に御［お］止［と］め申［まを］し上［あ］げますが自［じ］分［ぶん］は此［この］境［きやう］遇［ぐう］を天［てん］職［しよく］と感［かん］謝［しや］して及［およ］ぶ限［かぎ］り働［はたら］くつもりで御［ご］座［ざ］います。（1909 年 5 号「私の実行する精力主義」嘉悦孝子）

(9) 次に夫［をつと］も私［わたし］も至つてその讀［どく］書［しよ］家［か］ぢやありませんが、而も愛［あい］書［しよ］家［か］の方［はう］で、小［ちひ］さい建［たて］物［もの］ですが、裏［うら］に圖［と］書［しよ］館［くわん］もム［ござ］います。（1909 年 10 号「私の豪奢世界」すみれ女史）

(10) 堪［たま］らなくなつて男［だん］子［し］は帽［ぼう］を振［ふ］りわれ〜〜は手［て］にせる白［しろ］いハンカチーフを上［うへ］に捧［さゝ］げた。（1909 年 10 号「夏の新橋驛」秋雨）

なお、A グループの「我輩」の例は「我輩は下女である」という記事題名中に用いられたものであり、該当記事本文では「わたし・わたくし」が出現することから、例外的なものと見なしてよい。

語形別に見ると、A グループの「吾人・余・我輩」と B グループの「我々」と C グループの「僕」は常体および敬体の記事に出現し、敬体（ございます）の記事には出現しない。ただし、これらの語形の出現する文脈を確認すると、男性による記事のうち、「余」の出現する敬体の 1 記事、「吾人」の出現する敬体の 2 記事、「我輩」の出現する敬体の 2 記事、「僕」の出現するに 1 記事ついては、記事文体は敬体に分類したものの記事中の一部を占める常体の文中にこれらの語形が出現していることがわかった。また、女性による記事に出現する「我輩」については、上述のとおり例外的なものであった。整理すると、A グループの「余・我輩」は常体の記事のみに出現し、A グループの「吾人」と B グループの「我々」と C グループの「僕」は常体および敬体の記事に出現していることになる。一方 B グループの「わたくし・わたし」は常体・敬体の記事だけでなく敬体（ございます）の記事にも出現する。書き言葉的性質の強い A グループは常体・敬体に出現し、A グループよりも話し言葉的性質の強い B グループは敬体（ございます）に出現する傾向にあるとおおよそ言えるが、B・C グループにありながら敬体（ございます）に出現しない「我々・僕」のような語形もあり、語形の持つ書き言葉・話し言葉的性質の強弱と記事文体との間に対応関係があるとは単純には言い切れない。同じグループに属する語形であっても、さらにその内部で記事文体に対応した使い分けがあったと考えるのが穏当であろう。一人称代名詞の語形と記事文体との対応関係については、今後、調査対象資料を広げてさらに分析を深めたい。

7．おわりに

以上、形態論情報付き『近代女性雑誌コーパス』を用いて、主に非文学記事の地の文での一人称代名詞の使用実態について分析・考察した。記事著者の性別と記事文体との間には対応関係あることがまず確認でき、さらに記事文体に応じて一人称代名詞が使い分けら

れている傾向が見られた。実際の記事の執筆においては、一人称代名詞の選択より先に、著者性別に応じた記事文体の選択があったと考えるほうが自然であるならば、一人称代名詞の出現傾向は記事文体の出現傾向に強く影響を受け、その記事文体の出現傾向は著者性別の傾向に影響を受けていることになる。当時の記事地の文の一人称代名詞の使用実態を解明するためには、その前提として当時の著者性別ごとの記事文体の使用実態を把握することが重要であることが、本稿の考察から確認された。

記事文体と一人称代名詞の語形との対応関係については、『近代女性雑誌コーパス』のテキスト量が分析に十分ではなかったことや記事文体の認定方法に一層の工夫が必要であったことなどから、本稿では精緻な考察までたどりつけなかった。これらの問題点について改善をはかり、さらに考察を進めていきたい。

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

出雲朝子（2004）「女性の文章と近代」『日本語学』23:7、pp.27-37

岡田賢二（1998）「明治期の東京語における人称代名詞の研究―明治・大正期の落語の速記本にあらわれた一、二人称代名詞―」『埼玉大学国語教育論叢』2、pp.34-58

小木曽智信（2009）『科学研究費補助金研究成果報告書　近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（http://www2.ninjal.ac.jp/lrc/index.php?UniDic よりダウンロード可）

小木曽智信（2012）「旧仮名遣いの口語文を対象とした形態素解析辞書」『じんもんこん 2012 論文集』2012:7、pp.25-32

小木曽智信、中村壮範（2011）『特定領域研究「日本語コーパス」平成 22 年度研究成果報告書　『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装　改訂版』（JC-U-10-01）

祁福鼎（2006a）「明治時代語における自称詞の使用実態と使用規範について」『文学研究論集』24、pp.45-61

祁福鼎（2006b）「明治時代語における自称詞の推移と位相について」『明治大学日本文学』32、pp.95(1)-78(18)

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

小松寿雄（1999）「キミ・ボク対使用補考」『学苑』705、pp.68-77

近藤明日子（2008）「近代語における一人称代名詞「よ」「わがはい」―『太陽コーパス』を資料として―」『社会言語科学』11:1、pp.116-124

近藤明日子（2011）「『太陽コーパス』に見る一人称代名詞「吾人（ごじん）」―「余（よ）」との比較から―」『近代語研究』16、pp.63-80

近藤明日子（2012）「単語情報付き『太陽コーパス』を用いた一人称代名詞の分析」『日本語学会 2012 年度秋季大会予稿集』pp.229-234

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社、pp.1-48

田中牧郎（2006）「『近代女性雑誌コーパス』の概要」『日本学術振興会科学研究費補助金研究成果報告書 基盤研究(B) 「20 世紀初期総合雑誌コーパス」の構築による確立期現代語の高精度な記述』pp.55-62（http://www.ninjal.ac.jp/corpus_center/cmj/doc/19w-mag-summary.pdf よりダウンロード可）

那須小代美（1986）「三遊亭円朝の人情噺における人称代名詞の考察」『国文研究』32、pp.38-52

房極哲（2004）「近代語における一、二人称代名詞の変遷について」『日本文化學報』21、pp.1-15

関連 URL

R	http://www.r-project.org/
『近代女性雑誌コーパス』	http://www.ninjal.ac.jp/corpus_center/cmj/woman-mag/
近代文語 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic

『虎明本狂言集』コーパスの構造化　－仕様と事例の検討－

小林 正行（群馬大学 教育学部）
市村 太郎（国立国語研究所 コーパス開発センター）

Structuring the Corpus of *Toraakira-bon Kyogen*

Masayuki Kobayashi (Gunma University)
Taro Ichimura (National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所「通時コーパス」プロジェクトの一環として検討されている『虎明本狂言』の電子化について、資料の電子化に際し、いかなる要素を認定し、どのように構造化するのが適切かについて検討し、モデルを示す。

狂言テキストは演劇資料であり、台詞と卜書きから成る台本本文を中心とし、さらに舞台外の要素として注釈が付されることがある。底本である大塚光信編『大蔵虎明能狂言集翻刻註解』（2006，清文堂）は原資料に付された情報をよく残したまま活字化し、さらに原本にはない要素を付加している。

本発表では、多様なテキストの段階を持つ『虎明本狂言集』のタグセットや処理方針を示し、いくつかの例を提示する。

なお本発表では便宜上『大蔵虎明能狂言集　翻刻註解』を「底本」と呼ぶこととする。

２．『虎明本狂言集』コーパス化の意義

狂言は、中世から近世にかけての言語資料として重要な位置を占めている。登場人物が多彩で身分関係が明確であること、対話劇の形で進行し場面・状況が明確であることから、口語資料としての価値は極めて高い。

狂言資料の中でも『虎明本』は、寛永 19 年（1642）大蔵流十三世宗家大蔵弥太郎虎明の手による大蔵流の祖本である。本狂言 237 曲を収めており、狂言の類別や詞章の整備された台本として、質・量とも第一級の資料である。その詞章には、中世、室町時代の言葉を伝承していると見られる点、書写当時である近世初期の日常語の影響を受けたと思われる点、舞台言語として整理され固定化・類型化する兆候が見られる点がある。狂言史上の位置を踏まえ、他の台本との比較ということが不可欠であるが、注釈書や総索引が整備され、中世から近世の言語資料として広く利用されてきている。

しかし、刊行されている『大蔵虎明本狂言集総索引』は、狂言の類別に合わせた 8 分冊の形をとっており、単語認定の基準にばらつきがある。一定の基準でアノテーションされた形態論情報付コーパスの完成は、狂言の言語の研究だけにとどまらず、中世から近世初期にかけての言語研究に大きな成果をもたらす。

３．コーパスの設計方針

本研究では、コーパスの主な利用者として、言語研究者を想定する。そのため、言語的に重要な、文と短単位（ほぼ語に相当）が基本的な単位となる。

底本は、『大蔵虎明能狂言集　翻刻註解』上下巻を用いる。最新の活字本文であり、注記・ミセケチ等原本の情報を反映させることに配慮されており、また読みの指示など、詳細な注記がなされている。

本研究では、そのような底本の状況をできるだけ反映しつつ、単なる文字列の電子化ではなく、どこで得られた、どういう要素の、どのような性質を持つ語の表記体であるという情報が付された用例の一覧を、短時間で取り出せるようなコーパスを目指している。

そのため、底本内の各文書要素について XML を用いて記述し、国語研が作成した『太陽コーパス』の仕様や BCCWJ の仕様、『明六雑誌コーパス』の仕様を継承しながら、TEI P5 を参考に必要なタグを選択・追加し、構造化する。市村・河瀬・小木曽(2012)では、洒落本コーパスも含め、「近世口語テキスト」として、共通の基礎的な構造化案を示したが、本発表では、さらに実際の作業の過程で現れた問題を基に、タグ仕様を再設定する。

構造化されたデータには、さらに品詞情報や活用形等、形態素レベルで情報を付与する。なお、各演目はそれぞれ作品としては独立しているため、１演目を１テキストとする。

４．狂言テキストの構造とタグセット

狂言テキストは、台本本文を中心とし、その前後にはしばしば注釈が付される（図 1)。

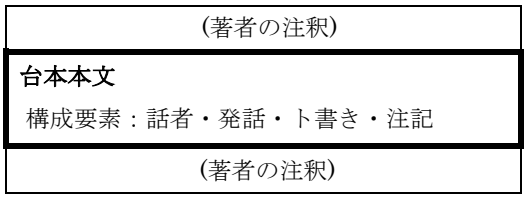


図 1　狂言テキストの構造概略

各々独立した演目ではあるが、全体として筆者は同一であり、形式や言語的状況は比較的安定している。

台本であるため、読み物とは異なり、序や後書がつかず注釈が多くなる。また当然台詞とト書きが中心となる。会話文に付記される話者の表示は、原著者によるものと、校注者によるものがあり、会話文の前後や合間にト書きが付される。

本コーパスと並行して「通時コーパス」プロジェクトでは『洒落本大成』のコーパスの設計も進められているが、文書の構造を比較すると、話者・会話文と割書きで主に構成される洒落本とはある程度の類似性があるといえる。そのため、洒落本大成コーパスの仕様との共通化を図り、基本的には共通のタグセットで表現する。

一方で、台詞やト書き、本文校訂や書き入れ・注釈という、舞台台本である狂言ならではの要素については、新たに要素を設定し、運用も改める。

以下、各要素について詳説する。(なお DTD については、同日発表の「洒落本コーパスの構造化」にある図を参照されたい。)

４．１文書の構造に関する要素

表 1　文書の構造に関するタグ（太線は階層上の大きな切れ目）

タグ（要素）	説明	属性
<text>	作品（演目）全体、作品のシリーズ・タイトル等を開始タグ内に記述	@textID（必須） @series シリーズ名（必須） @title　作品タイトル（必須） @yomi 作品名の読み（任意） @year　西暦成立年（必須） @year_w 和暦成立年（必須）
<front>	前付け部分 （狂言の場合は原則<titleBlock>のみ）	
<body>	主本文	

<article>	記事	@type（任意）
<titleBlock>	<article>レベルでのタイトル等の記述	
<p>	タイトルや注釈等を除く本文の塊	
<block>	<p>で記述された本文とは区別されるタイトル・注釈等のブロック要素	@type（必須）
<s>	文	
<SUW>	短単位	（多岐にわたるため省略）

文書構造に関する基本的な要素は洒落本と共通である。テキスト全体を表す<text>と、それを構成する<front><body>から成る。作品に関する情報は、属性値で<text>内に記述する。

さらに、内部は<titleBlock>と<article>に分割され、さらに本文に<p>、注釈に<block>を付す。これらはさらに<s>に分割され、文は形態論情報を記述した<SUW>に分割される。

なお作品ごとに序文・後書きが付されることはないため、実質台詞・ト書き・注釈による大きな<body>と、タイトルのみの<front>で構成される。また作品内に小見出し等はなく、洒落本等に比べればシンプルな文書構造といえる。

article 要素　前付・後付を除いた中心的本文は、小見出し等を伴う複数の要素から成ることがあり、このような階層の要素を表すものとして、<article>を用いる。狂言の各作品には小見出し等は見られず、実質テキストのタイトルを除くすべての部分が該当する。

p 要素　<article>内の本文の塊全体で付与する。視覚上、また内容上いわゆる段落を認定するのは困難である。本研究では「主たる本文かそれ以外か」に重点をおいている。

block 要素　視覚上また構成上、明らかに主本文の塊と区別される要素を表す。type 属性で、タイトル・著者・日付・注釈等の別を記述する。

狂言では、演目内的な会話文とト書きを「主本文の塊」と見、演目外的な、また追加的な情報を付加する注釈を「本文の塊と区別される要素」と見る。

titleBlock 要素　テキストのタイトル箇所に付与する。狂言単独で見たときには<block>タグのみでも事足りるが、共通仕様をめざす洒落本では内題が現れることがあり、<article>と同階層でマークアップされるため、それに合わせて本要素を付与する。

s 要素　すべてのテキストは文に分割される。ただしいわゆる「文」とは完全に同一ではなく、発話や割書の区切りでも切る。なお、<s>が<s>を含むような階層性は認めない。

SUW 要素　短単位（おおよそ語に相当）を表す。すべての文は短単位に分割される。本研究での基本的な単位である。語彙素・語形・書字形・活用型・活用形・発音形等語に関する多くの情報が、属性で記述される。開発中の「近世口語 UniDic」による解析結果を人手で修正して付与する。

<text textID="虎明本狂言_034_大名_入間川" series="虎明本狂言・大名狂言之類#34" title="入間川" year="1642" year_w="寛永 19"><front><titleBlock><block type="title"><s><pb n="170"/><lb/>入間川</s></block></titleBlock></front>
<body><article> <speech><s><speaker value="大名"/><lb/>「罷出たる者<kana>は</kana>、<hi rend="傍線">東</hi>国
<info text="「はるかおん国ともなのる"/>にかくれもなひ大名です、</s><s><ruby resp="annotator" rubyText="訴">そ
</ruby><ruby resp="annotator" rubyText="訟">せう</ruby>の事有て永々<lb/>在京仕る処に、<ruby resp="annotator"
rubyText="あん">安</ruby><ruby resp="annotator" rubyText="ど">堵</ruby>の<ruby resp="annotator" rubyText="み">御
</ruby><ruby resp="annotator" rubyText="げう">教</ruby><ruby resp="annotator" rubyText="しよ">書</ruby>をいた<odoriji
originalText="ゞ">だ</odoriji>き、殊にお<ruby resp="annotator" rubyText="いとま">暇</ruby>を下された程に、急でく<lb/>
だらふと存る、</s><s>太郎くわじやあるか</s></speech><speech><s><speaker value="太郎冠者"/>「お前に
</s></speech><speech><s><speaker value="大名"/>「いそひで<ruby resp="annotator" rubyText="立">た</ruby><lb/>て
</s></speech><speech><s><speaker value="太郎冠者"/>「御<ruby resp="annotator" rubyText="機">き</ruby><ruby
resp="annotator" rubyText="嫌">げん</ruby>が<ruby resp="annotator" rubyText="良">よ</ruby>う御ざある
</s></speech><speech><s><speaker value="大名"/>「その事よ、</s><s>訴<corr type="erratum" originalText="詔"
resp="annotator">訟</corr>こと<ruby resp="annotator" rubyText="悉">／＼</ruby>く<lb/><ruby resp="annotator" rubyText="
安">あん</ruby><ruby resp="annotator" rubyText="堵">ど</ruby>し、おいとまを下された<kana>は</kana>
</s></speech><speech><s><speaker value="太郎冠者"/>「やれ／＼それ<kana>は</kana>めでたひ事で<lb/>御ざる、</s>

図 2　作品冒頭部分の形式化例（上巻『入間川』p.170）

<speech><s><speaker value="入間"/>「扨<kana>は</kana><hi rend="傍線">いるま</hi>やうをのけ<lb/>てか
</s></speech><speech><s><speaker value="大名"/>「中／＼</s></speech><stage><s><add>二度云て三度め程に
</add></s></stage><speech><s><speaker value="入間"/>「<ruby resp="annotator" rubyText="存">ぞん</ruby>じも<ruby
resp="annotator" rubyText="寄">よ</ruby>らぬに、色々の物を<ruby resp="annotator" rubyText="貰">もら</ruby>ふて、うれ
<lb/>しうなひと申事<vMark>が</vMark>ござらふぞ、</s><s>身にあまつてかたじけなふ御ざる</s></speech><stage><s>
「と云て<lb/>いた<odoriji originalText="ゞ">だ</odoriji>く</s></stage><speech><s><speaker value="大名"/>「身にあまつて
<ruby resp="annotator" rubyText="かたじけない">忝</ruby>とおしやる<kana>は</kana>、うれしうなひといふ事じ<lb/>や
程に、こちへお<ruby resp="annotator" rubyText="返">かや</ruby>しやれ</s></speech><stage><s>「と云て皆とりかへす
</s></stage><speech><s><speaker value="入間"/>「あのたらしが、<pb n="176"/><lb/>やるまひぞ／＼
</s></speech><stage><s>「と云ておいいるなり</s><s>「太郎くわじや<kana>は</kana>、太刀を主にわたしてひ<lb/>つこ
む</s></stage> <block type="注釈"><s><lb/>「私<kana>に</kana>云、右つめのこと<kana>は</kana>何共がてんのゆき
かたき事也、</s><s>然共<hi rend="傍線">いるま</hi>やうをのけて<lb/>といふ<kana>は</kana>、のけいでと云事を、ま
こと<odoriji originalText="ゝ">と</odoriji>心得ていふたによつて取かへした<kana>は</kana>こと<kana>は</kana><lb/>
りなり<info originalPage=""/></s></block><block type="注釈"><s><lb/>一</s><s>いる<corr type="omission"
resp="annotator">ま</corr>やうをのけてと云<kana>は</kana>、のけ<add>い</add>でと云ことじやにと云へ
<vMark><kana>ば</kana></vMark>よくきこへ候へ共そ<lb/>れにて<kana>は</kana>人がしるによつていわぬがよき也
</s></block></article></body></text>

図 3　作品末尾の形式化例（上巻『入間川』pp.175-176）

キー	語彙素	出現形発音形	品詞	解析活用型	活用形
「	「		補助記号-括弧開		
罷	罷る	マカリ	動詞-一般	文語四段-ラ行	連用形-一般
出	出でる	イデ	動詞-一般	文語下二段-ダ行	連用形-一般
たる	たり	タル	助動詞	文語助動詞-タリ-完了	連体形-一般
者	者	モノ	名詞-普通名詞-一般		
は	は	ワ	助詞-係助詞		
、	、		補助記号-読点		
東国	東国	トーゴク	名詞-普通名詞-一般		
に	に	ニ	助詞-格助詞		
かくれ	隠れ	カクレ	名詞-普通名詞-一般		
も	も	モ	助詞-係助詞		
なひ	無い	ナイ	形容詞-非自立可能	形容詞	連体形-一般
大名	大名	ダイミョー	名詞-普通名詞-一般		
です	です	デス	助動詞	助動詞-デス	終止形-一般
、	、		補助記号-読点		
そせう	訴訟	ソショー	名詞-普通名詞-サ変可能		
の	の	ノ	助詞-格助詞		
事	事	コト	名詞-普通名詞-一般		
有	有る	アリ	動詞-非自立可能	文語ラ行変格	連用形-一般
て	て	テ	助詞-接続助詞		
永々	長々	ナガナガ	副詞		
在京	在京	ザイキョー	名詞-普通名詞-サ変可能		
仕る	仕る	ツカマツル	動詞-一般	文語四段-ラ行	連体形-一般
処	所	トコロ	名詞-普通名詞-副詞可能		
に	に	ニ	助詞-格助詞		
、	、		補助記号-読点		
安堵	安堵	アンド	名詞-普通名詞-サ変可能		
の	の	ノ	助詞-格助詞		
御	御	ミ	接頭辞		
教書	教書	ギョーショ	名詞-普通名詞-一般		
を	を	オ	助詞-格助詞		
いただき	頂く	イタダキ	動詞-非自立可能	文語四段-カ行	連用形-一般
、	、		補助記号-読点		
殊に	殊に	コトニ	副詞		
お	御	オ	接頭辞		
暇	暇	イトマ	名詞-普通名詞-一般		
を	を	オ	助詞-格助詞		
下さ	下す	クダサ	動詞-一般	文語四段-サ行	未然形-一般
れ	れる	レ	助動詞	助動詞-レル	連用形-一般
た	た	タ	助動詞	助動詞-タ	連体形-一般
程	程	ホド	名詞-普通名詞-副詞可能		
に	に	ニ	助詞-格助詞		
、	、		補助記号-読点		
急	急ぐ	イソイ	動詞-一般	文語四段-ガ行	連用形-イ音便
で	で	デ	助詞-格助詞		
くだらふ	下る	クダロー	動詞-一般	文語四段-ラ行	意志推量形
と	と	ト	助詞-格助詞		
存る	存ずる	ぞんずる	動詞-一般	文語サ行変格	連体形-一般
、	、		補助記号-読点		
太郎	太郎	タロー	名詞-普通名詞-一般		
くわじや	冠者	カジャ	名詞-普通名詞-一般		
ある	有る	アル	動詞-非自立可能	文語ラ行変格	連体形-一般
か	か	カ	助詞-終助詞		
「	「		補助記号-括弧開		
お	御	オ	接頭辞		
前	前	マエ	名詞-普通名詞-副詞可能		
に	に	ニ	助詞-格助詞		

図 4　短単位解析済みデータの例（一部項目省略・上巻『入間川』p.170）

４．２　文・語の機能に関する要素

表 2　文・語の機能に関する要素

タグ（要素）	説明	属性	
<speech>	会話	@source（任意） @type（任意）	
<quotation>	①単純な発話以外の引用要素 ②ト書き内の台詞指示等	@source（任意） @type（任意）	
<stage>	ト書き		↑文以上
<speaker> <speaker/>	話者 （校注者付記の場合は空要素）	@value（任意）	↓文未満
<delivery>	発話等のスタイルの表示		
<verse>	韻文		

speech 要素　1 回的な会話文の連続を表す。<speaker>を発話の内部に認定し、一体として扱う。会話文内に話者が示されていない場合には@source 属性で話者を可能な限り記述する。また、底本では、紙面上（）付で校注者により話者が示されことも多く、それについては空要素とし、@value 属性で話者を記述する。

quotation 要素　手紙や和歌等、単純な会話文以外の引用要素を表す。@type 属性でどのような種の引用かを、@source 属性で出典を記述する。

　また、しばしば現れるト書き内の台詞指示等は、本来階層は異なるが、本要素で記述する。その場合、基本的に話者表示がないため、@source 属性で話者を記述する。

stage 要素　本文内的な要素としてト書きを表す。狂言は舞台演劇であり、台詞とト書きが比較的明確に分かれるのが特徴である。ト書きは時に内容としては本文外的な挿入的なものあり、この点注釈と重なるのだが、会話と会話を割って、または会話に付属して述べている点において、本文の塊の外側に付される注釈ほどの独立性はない（つまり階層的には別の次元のもの）と見る。そのため、内容が注釈的なト書きであっても、会話に割って入る以上は、本文内的な要素であって、<block>扱いはしない。

speaker 要素　会話文に付属する、小書き等で記される話者の表示である。底本では原作者による話者表示のほか、校注者が補った（　）付の話者がある。これらは、原作者の表示と区別するため空要素とし、@value 属性内に記述する。

delivery 要素　台詞の内部には、話者だけでなくその台詞のスタイルを小書き等で記してある場合がある。狂言においては、「舞がけり」など、散文資料の発話に比べ細かく台詞指示がなされており、重要な要素である。

verse 要素　韻文は、歌・舞等について、文未満の単位で付与する。

４．３　語・文字単位で外形・機能等を表す要素

『虎明本狂言集』は、筆者による本文修正や、テキストの追加・削除の指示が頻繁に行われるという点で特徴的である。

また、底本には『洒落本大成』とは異なり、校注者による誤りの指摘（ママ注）や、校注者が追加した振り仮名等があって、原資料の筆者の指示と、校注者による情報との 2 段階で記述し分ける必要がある。

そのため、「洒落本コーパス」の仕様に比べ、本文校訂に関する記述が詳細である。

表 3　語・文字単位で外形等を表す要素

タグ(要素)	説明	属性	
<hi>	文字列（語）に対する装飾	@rend（必須）	↑短単位以上
<lRuby>	左ルビ	@rubyText（必須） @rubyBase（任意） @resp（任意）	↓短単位未満
<ruby >	ルビ	@rubyText（必須） @rubyBase（任意） @resp（任意）	
<odoriji>	踊り字を開いた文字	@originalText（必須）	
<gap/>	抹消・破損等で判読できない文字の存在（空要素）		
<corr> <corr/>	本文修正	@type（必須） @originalText（任意） @resp（任意）	
<unclear>	推読された文字	@originalText（任意） @type（任意）	
<vMark>	濁点付仮名に変換した箇所		
<g>	外字	@type（必須） @ref（任意）	
<kana>	片仮名を平仮名に変換した箇所		
<add>	著者によって追加されたテキスト		
<kanbun> <kanbun/>	漢文（返読）	@type（任意）返読前 返読後 @originalText（任意） @id（任意）	

hi 要素　傍線が付される、小書きされるなど、外形的特徴を持った文字列（語）を表す。狂言では固有名詞に傍線が引かれるケースがあるが、必ずしも機能は一定ではない。

ruby 要素　文字列の右側に付され、文字・文字列の読み等を表す振り仮名等を指す。@rubyText 属性内にルビ文字列が記述される。右側漢字傍記も含む。

凡例によると、原資料に付されている振り仮名・漢字傍記(A)については＜＞が付されており、校注者によって新たに付されたもの(B)には何も付されていない旨の記述がある。そのため、@resp 属性で校注者により付与されたものを区別する。

(A)　<ruby rubyText="〈ソサノヲ〉">素盞烏</ruby>（上巻『ゑびす大黒』p.6）

(B)　<ruby resp="annotator" rubyText="戯">ざれ</ruby>事（上巻『連歌毗沙門』p.10）

lRuby 要素　文字列に沿って小書きされる文字は、右側の振り仮名だけでなく、左側に付されることもある。rubyText 属性内にルビ文字列が記述される。

corr 要素　本文テキスト修正箇所であり、文字単位で付す。狂言の場合、本文テキストの正誤にかかわる指示としては、ミセケチ等による筆者の校訂箇所と、ママ注によって校注者が誤りを指摘している箇所の 2 種があり、特徴的なものと言える。これらは区別すべきものであり、また原文を確認できる形にすることが重要である。一方で、形態論情報を付すことを考慮すると、本文としては「きれいな本文」であることが望ましい。

まず@type 属性で誤字（erratum）・衍字（excess）・脱落（omission）の別を付し、本文

は修正後の形とするが、@originalText 属性で元のテキストを記述する。また@resp 属性で筆者（writer）の指示によるものか、校注者（annotator）の指摘によるものかを記述する。

校注者の指摘するママ注には、いかなる誤りかが頭注に明記されず、推測困難な場合がある。そのような箇所は修正せず、@type 属性で「修正なし」と記述する。

```
<s>今日<ruby resp="annotator" rubyText="最">さい</ruby><ruby resp="annotator" rubyText="上">じやう</ruby>吉<corr originalText="日日" type="excess" resp="annotator">日</corr>でござる<lb/>により、聟殿のおいでなされうずるとのおこ<corr type="omission" resp="annotator">と</corr>じや</s>
```

図 5 ママ注の形式化例（上巻『鶏聟』p.353　下線は筆者）

vMark 要素　底本にはなく、電子化に際して新たに濁点を付与した箇所に付与する。ただし踊り字箇所はもとのテキストを属性値に記録するため、タグ付け対象とはしない。

add 要素　筆者による傍記や「〇」符号等によって、挿入指示がなされた本文に付与する。文字単位から複数文単位まで多岐にわたり、強調表示も含む。文を超える単位で挿入指示がなされる場合は文単位で付与し、短単位未満の場合は、文字単位で付与する。

```
<speech><s><speaker value="えびす"/><delivery>かたり</delivery></s><s>「夫<hi rend="傍線">ゑびす</hi>三郎殿といつ<vMark><kana>ぱ</kana></vMark>、<hi rend="傍線"><ruby resp="annotator" rubyText="伊弉諾">いざな<vMark>ぎ</vMark></ruby></hi><hi rend="傍線"><ruby resp="annotator" rubyText="伊弉冉">いざな<kana>み</kana></ruby></hi>の<ruby resp="annotator" rubyText="尊">みこと</ruby>、あ<info originalPage=""/>ま<lb/>の岩くらの<ruby resp="annotator" rubyText="苔">こけ</ruby><ruby resp="annotator" rubyText="蓆">むしろ</ruby>にて、<hi rend="傍線">男</hi><hi rend="傍線">女</hi>の<ruby resp="annotator" rubyText="語">かた</ruby>らひをなし、日神月神、<ruby resp="annotator" rubyText="蛭">ひる</ruby><ruby resp="annotator" rubyText="子">こ</ruby><lb/><ruby resp="annotator" rubyText="素盞烏">そさのお</ruby>の御子をまうけ給ふ、<hi rend="傍線">ひるこ</hi>と<kana>は</kana>某が事、</s><s><info text="○"/><add><info text="○"/><kana><hi rend="傍線">天照太神</hi>より三番めのをと／＼成<kana>れ</kana><vMark>ば</vMark>とて、<hi rend="傍線">西の宮</hi>の<hi rend="傍線">ゑびす三郎</hi>殿といは<odoriji originalText="ゝ">は</odoriji>れ</kana></add>うぢ<info text="氏〈ウジ〉"/><corr originalText="し" type="erratum" resp="writer">す</corr><corr originalText="ゆ" type="excess" resp="writer"/><ruby rubyText="〈図性〉">じやう</ruby><info text="種〈シユ〉 姓〈ジヤウ〉"/><ruby resp="annotator" rubyText="誰">たれ</ruby><lb/>にか<ruby resp="annotator" rubyText="劣">おと</ruby>りたまふべき、</s><s>なんぼういみじき<ruby resp="annotator" rubyText="位">くらい</ruby>にて<kana>は</kana>なきか、よく／＼<lb/><ruby resp="annotator" rubyText="信">しん</ruby><ruby resp="annotator" rubyText="仰">がう</ruby>せよ、</s><s><ruby resp="annotator" rubyText="楽">たのし</ruby>うなさうずるぞ<lb/></s></speech>
```

図 6 複雑な注記・本文訂正等の形式化例（上巻『ゑびす大黒』p.4　下線は筆者）

４．４位置情報と本文外情報

表 4　底本テキストの位置情報を表すタグ

タグ(要素)	説明	属性
<pb/>	ページ開始（空要素）	@n（必須）
<cb/>	段開始（空要素）	@n（必須）
<lb/>	行開始（空要素）	
<info/>	本文外情報（空要素）	@originalPage（任意）@text（任意） @type（任意）　@originalText（任意）

info 要素　本文外の情報を空要素**<info/>**で表す。底本には影印の改ページが付されており、その位置情報を@originalPage 属性で記述する。また注などの傍記が本文脇に付されることがあるが、本文外に相当する傍記・注記等は@text 属性で記述する。

５．コーパス化に向けての課題

５．１　本文認定と読み順の確定

『虎明本狂言集』には、筆者・校注者による本文に関わる多くの校訂や注記の情報があり、複雑な状況を呈している箇所もある。上欄や本文末に付された挿入指示については、場所が指定されていない場合、内容によって適切な挿入箇所を定めなければならない。

また、同じ傍記であっても、本文に追加する要素、本文を訂正する要素、注記と多様で、また必ずしも校注者による言及があるわけではなく、個別に検討・判断する必要がある。

何が本文で何が本文でないか、また、どの順で読むべきか等は、基本的なことではあるが、本文を決めなければならないコーパスにおいて大きな課題である。

５．２　解釈の問題

舞台台本であるため会話の切れ目が比較的わかりやすく、また底本には校注者の詳細な注が付されているため、近世散文資料等に比べれば文認定は容易である。しかし「。」によって文が区切られているわけではなく、また間接引用か直接引用かがはっきりせず、文認定が困難な箇所は存在する。文認定や現代語訳が行われていない資料を扱う際の共通の課題である。

また、２で述べたように、言語的な状況は中世語に近いとされるため、濁点付与に関しては、タグを付与するとはいえ慎重を期する必要がある。例えば、現代では濁音で発音されるものでも、清音で読んでおくべきものがしばしばみられ（「かがやく」―「かかやく」など）、これらは『日葡辞書』等の記載を参照するなどし、個別に検討する必要がある。

６．おわりに

『虎明本狂言集』では、主に版本が主体である洒落本とは異なり、筆者の本文に対する校訂や補遺・補入が随所に見られる。底本においてもそれがよく反映されており、また校注者によって追加された要素も多く見られる。今後、本文校訂に関する多様な要素を持つ資料を対象とするにあたって、文書構造としてどのレベルまで想定し、記述するのかが課題となる。本研究で言えば、傍記等の中での振り仮名等を構造化するのは現状では困難であり、このようなものの扱いをどうするのかは今後の課題である。

また、さまざまなレベルで出現する補遺・補入の類の扱いは、階層構造を前提とする XML を用いる以上、資料ごとに検討され続けなくてはならない課題であろう。

日本語史資料として、狂言はもちろん、浄瑠璃・歌舞伎等の舞台資料は極めて重要である。本研究での検討は、「日本語歴史コーパス」構築に向けて、これら舞台作品を含めた仕様を作る上での足掛かりになると考える。

文　献

市村 太郎、河瀬 彰宏、小木曽 智信(2012)『近世口語テキストの構造化とその課題』情報処理学会研究報告　人文科学とコンピュータ研究会報告(CH96) pp.1-8

大塚光信編(2006)『大蔵虎明能狂言集　翻刻註解』清文堂

北原保雄、村上昭子、鬼山信行、小川栄一、山崎誠、吉見孝夫、土屋博映、大倉浩編(1983-1989)『大蔵虎明本狂言集総索引』1-8　武蔵野書院

近藤明日子、田中牧郎「『明六雑誌コーパス』の仕様」『国立国語研究所共同研究報告 12-03 近代語コーパス設計のための文献言語研究 成果報告書』pp.118-143　国立国語研究所

近藤泰弘(2012)「日本語通時コーパスの設計について」『国語研プロジェクトレビュー』3 pp.84-92 国立国語研究所

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」『国立国語研究所報 122　雑誌『太陽』による確立期現代語の研究　『太陽コーパス』研究論文集』pp.1-48　博文館新社

田中牧郎、小木曽智信(2000)「総合雑誌『太陽』の本文の様態と電子化テキスト」『日本語科学』8　pp.141-152　国立国語研究所

安永尚志(1998)『国文学研究とコンピュータ』勉誠社

山口昌也、高田智和、北村雅則、間淵洋子、大島一、小林正行、西部みちる(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『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2』　文部科学省 科学研究費 特定領域研究 「日本語コーパス」データ班

関連 URL

「Text Encoding Initiative」（ガイドライン P5 日本語版）
http://docsci.infon.org/stack/P5JA/index-toc.html

自発発話におけるイントネーション句単位の F0 変動の特徴

石本 祐一（国立情報学研究所 情報学プリンシプル研究系/音声メディアグループ）†
小磯 花絵（国立国語研究所 理論・構造研究系）

F0 Characteristics at the Level of Intonational Phrase in Spontaneous Japanese

Yuichi Ishimoto (Principles of Informatics Research Division/Speech Media Group, NII)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

1. はじめに

小磯・石本（2012）および石本・小磯（2012）では，自発音声の発話の韻律的特徴を探るため，イントネーション句（IP）を単位として発話の基本周波数（F0）の変動を調べ，発話（文）内に強い節境界が存在する場合と存在しない場合において，

- 発話中に強い統語的節境界が存在しない場合（一つの節で発話が構成される場合）
  1. IP 全体の F0 最大値・最小値が発話内で徐々に下降する
  2. 発話の長さ（IP 数）に関わらず，IP はほぼ一定の高さで始まり一定の高さで終わる（つまり発話の長さによって F0 下降の傾きが異なる）
  3. ただし発話が長い場合，若干高い F0 で発話を開始する
  4. 発話末に著しい F0 下降がみられる（final lowering）
- 発話中に強い統語的節境界が存在する場合（二つ以上の節で発話が構成される場合）
  1. 発話中の節境界で IP の F0 下降傾向が途切れてリセットされる
  2. F0 下降のリセット時, IP の F0 最小値は発話末のレベルにまで達せず, final lowering に相当する著しい F0 下降はみられない
  3. F0 下降のリセット後，IP の F0 最大値は発話冒頭のレベルにまで達することもあればそれより低いこともある

という傾向がみられることを指摘した。

Pierrehumbert & Beckman (1988) は, アクセント句 (AP) や IP より上位に「発話 (Utterance)」という単位を設定し，その範囲で 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，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）が生じるとしている。上述の F0 下降傾向は declination のようにも見えるが，発話の長さによって傾きが異なることから，単純に declination とみなすことはできない。また発話末を特徴付ける final lowering が観察されない発話中の強い統語的な節境界において，F0 下降傾向がリセットされることからも，「発話」に見られる F0 declination とは解釈しづらい。

† ishimoto@nii.ac.jp

一方 Kawahara and Shinya（2008）は，発話とダウンステップの生じる領域（上記 IP に相当）との間に統語的な節に相当する別の韻律階層が日本語にも存在することを，読み上げ音声を対象とした分析に基づき主張している。F0 下降傾向が節の範囲で観察されるという我々の結果はこの説を支持するものとも解釈できる。しかし，強い節境界で必ず F0 下降のリセットが生じているわけではないこと，強い節境界以外の場所でも F0 下降のリセットが生じる可能性があることなどを考えると，F0 下降を指定する領域として，発話と IP の間に節に相当する単位を設けると結論付けることは拙速であり，さまざまな要因も検討する必要がある。

その一つとして，息などの生理的な要因が挙げられよう。F0 の下降は，一息で発話される範囲において残りの息の量との関係で生じる可能性があるが，意味的なまとまりを示す節境界と息継ぎの位置がある程度一致していたために，結果として統語的な節境界で F0 下降のリセットが観察されたとも考えられる。そこで本研究では，節境界だけではなく，息継ぎ（によって生じるポーズ）や発話時間などを含む物理的な特徴量を総合的に考慮した分析を行い，発話中の IP 単位の F0 変動に関わる要因について多角的に検討する。

2. データ

2.1 コーパス

分析には『日本語話し言葉コーパス（*Corpus of Spontaneous Japanese*:以下 CSJ）』（前川 2004）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。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，CSJ 全体にあたる 661 時間の音声のうち，「コア」と呼ばれるデータ範囲の中から学会講演 70（約 19 時間）・模擬講演 107（約 20 時間）を分析対象とした。

2.2 節単位

CSJ に付与されている節単位情報（丸山ほか 2006）を分析に利用する。節単位は原則「節（clause）」の境界によって得られる文法的・意味的なまとまりを持った単位であり，節境界の構造的な切れ目の大きさの観点から以下の 3 つに分類される。

絶対境界（Absolute boundary）　いわゆる文末に相当する境界。
強境界（Strong boundary）　後続の節に対する従属度の低い，切れ目の度合が強い節境界。
弱境界（Weak boundary）　後続の節に対する従属度の高い，切れ目の度合が弱い節境界。

これらの境界は形態素解析結果に基づき自動で判別され，人手による修正・操作を経た上で，絶対境界，強境界のいずれかで区切られる単位が節単位と認定されている。

2.3 発話・イントネーション句・アクセント句

Pierrehumbert & Beckman（1988）の韻律理論では，「発話－イントネーション句 (IP) －アクセント句 (AP)」という韻律的階層構造が仮定される。AP は，第 1 モーラから第 2 モーラ付近にかけての F0 の上昇と句末への緩やかな下降を有し，かつアクセント核による下降を最大ひとつ持ちうる単位と定義される。IP は AP の上位階層に位置し，AP のピッチレンジを指定す

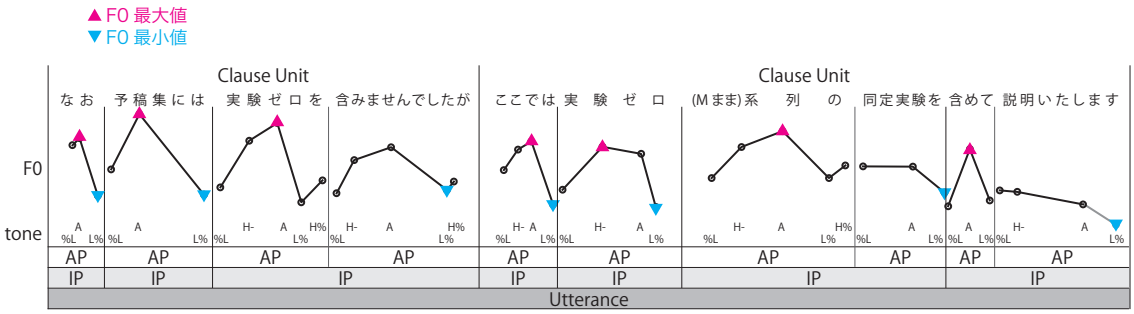


図 1　韻律的階層構造

る単位と定義される。アクセント核が引き起こす後続 AP のピッチレンジの縮小効果は，IP の範囲で観察される。発話は上述の通り，IP の上位階層に位置し，F0 declination が観察され，かつその末尾で final lowering が生じる範囲と定義される。以上をまとめると，各階層で F0 は，

- AP の領域では，山状に上昇後下降する
- IP の領域では，AP 単位で低下する（ダウンステップ）
- 発話の領域では，発話末に向かって低下し，発話末で急激に低下する

といった変動を示す。この韻律的階層構造を図 1 に示す。

本研究では IP を単位とした発話の F0 変動の特徴を探る。CSJ にはラベリングスキーム X-JToBI（五十嵐ほか 2006）に基づき韻律情報が付与されているが，この中に，韻律境界の切れ目の強さに関する情報として Break Index（BI）が存在する。BI=2 は AP 境界，BI=3 は IP 境界，BI=F はフィラー境界，BI=D は言い淀み境界に対応する。ここでは，BI=3 で区切られる範囲を IP と認定し，フィラー，言い淀み部分を除いて分析に用いた。ただし，フィラーを狭んでダウンステップが続く場合はフィラーを内包する形で IP を認定した。

このように IP を認定した上で，IP の F0 特徴量として X-JToBI に基づく Tone 情報から，

F0 最大値　IP 頭の AP の句頭音調（H-）あるいはアクセント核（A）のうち高い方の F0 値
F0 最小値　IP 末尾の AP の下降音調（L%）の F0 値

を求めた。また，分析において性差・個人差の影響を小さくするために，F0 値は話者ごとの平均 F0・標準偏差によって Z スコアへ変換して用いる。

韻律的な発話の境界情報は X-JToBI では示されないため，本研究では発話に相当する単位を，前節に示した絶対境界によって区切られる区間とする[*1]。分析対象となる発話数は 9,885 となった。

3.　IP の F0 最大値・最小値と IP 前後のポーズの関係

3.1　方法

本節ではまず，IP 前後のポーズと F0 最大値・最小値との関係について調べる。

1 節でもまとめたように，小磯・石本（2012）および石本・小磯 (2012) では，

1. IP の F0 最大値・最小値が発話末へ向けて徐々に低下する

[*1] 明示的な文末表現が置かれるもののほか，「と文末」や「体言止め」なども含む。

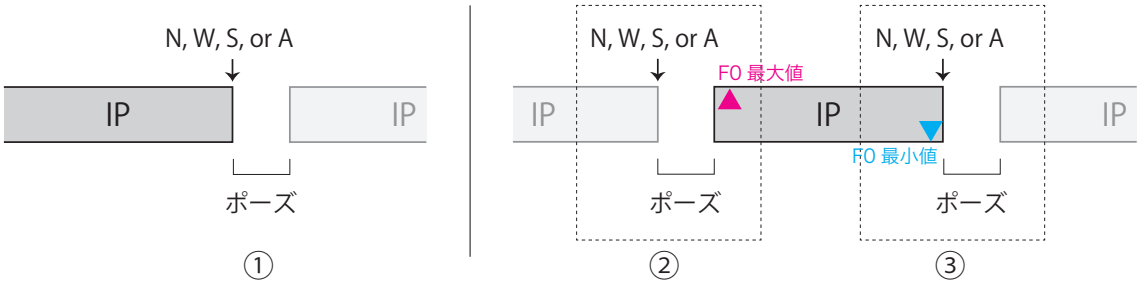


図 2　IP 前後のポーズと節境界（非境界:N，弱境界:W，強境界:S，絶対境界:A）

表 1　IP 前後の節境界と IP 数（非境界:N，弱境界:W，強境界:S，絶対境界:A）

	IP 後				
IP 前	N	W	S	A	合計
N	23706	4326	1841	920	30793
W	2793	545	239	141	3718
S	1263	119	40	14	1436
A	1227	135	50	26	1438
合計	28989	5125	2170	1101	37385

2. 発話中に強い節境界が見られる場合，IP の F0 下降はそこでリセットされる

という傾向が観察された。このように，先の分析結果からは統語的な境界である強境界において下降のリセットという急激な F0 変動が起こっているようにみえるが，実際は統語的な境界が主な要因ではなく，強境界（および絶対境界）で生じやすい息継ぎやポーズ（空白期間）が F0 下降の途切れの要因となっている可能性が考えられる。

そこで IP の前後のポーズにより F0 最大値・最小値に違いが現れるか分析を行う。具体的には，IP 末の節境界とその直後のポーズの関係（図 2 ①）を確認した上で，IP 前のポーズと F0 最大値（図 2 ②）および IP 後のポーズと F0 最小値（図 2 ③）の関係を調べる。節境界の種類は，2.2 節に示す「絶対境界」「強境界」「弱境界」およびこれら節境界の存在しない「非境界」の 4 種とする。なお，CSJ には息継ぎのタイミングは記録されていないため，ポーズをその代替的指標として取り上げる。

ポーズの長さが 2 秒以下の IP に限定したところ，分析対象となる IP の数は 37,385 であった。IP 前後の節境界の種類と IP 数を表 1 に示す。

3.2　結果と考察

IP 末の節境界の種類とその直後のポーズ長との関係について，図 3 に示す。分散分析の結果，節境界の違いは有意であった（$F(3.000,3394.846)=1123.324$, $p<0.001$）。Tukey 法による多重比較によれば，すべての節境界間に有意差があった。予想していたように，統語的な切れ目が強いほどポーズが長くなるという傾向が顕著に見られる。

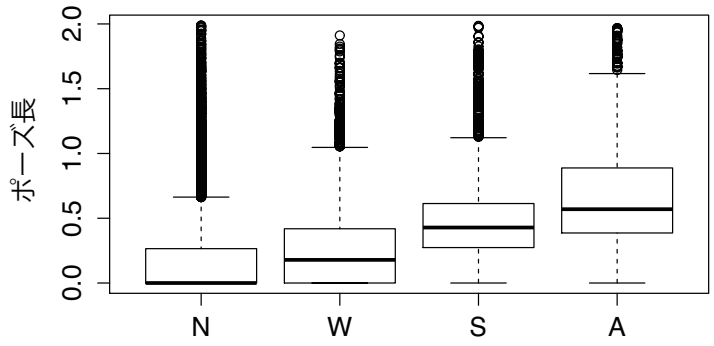


図 3　IP 末の節境界の種類とその直後のポーズ長（非境界:N，弱境界:W，強境界:S，絶対境界:A）

次に IP 前のポーズの長さと IP の F0 最大値（図 2 ②）の散布図を，IP 直前の節境界の種類ごとに図 4 に示す。ポーズが F0 下降のリセットと関わるならば，節境界の種類に関わらず，ある程度の長さのポーズが置かれると，そこで F0 下降はリセットされ，直後の IP の最大値は高くなることが予想される。しかし図 4 からわかるように，非境界と強境界ではその傾向は全く見られず，また弱境界と絶対境界についてもわずかにその傾向がうかがえる程度である。

IP 後のポーズの長さと IP の F0 最小値（図 2 ③）の散布図を図 5 に示す。IP 単位の F0 下降がポーズの位置で止まるのであれば，ポーズの前の F0 最小値はポーズがない場合よりも低くなるはずである。図 5 からわかるように，確かに統語境界の種類によらずこの傾向は観察されるものの，その効果は決して大きくはない。

このようにポーズの長さだけに着目すると，IP の F0 値に与える影響は部分的かつ微小であることが分かった。

4.　IP の F0 最大値・最小値と発話の時間パラメータの関係

4.1　方法

本節では，ポーズ以外の発話の物理量も取り上げ，F0 最大値・最小値との関係を調べる。具体的には，発話の時間に関わるパラメータとして，

- IP 直前のポーズ長
- IP 直後のポーズ長
- 当該 IP を含む発話の長さ
- 当該 IP を含む節単位の長さ
- IP の長さ
- IP の開始時刻から発話末までの時間
- IP の開始時刻から節単位末までの時間

を話者ごとに Z スコアに変換して説明変数とし，F0 最大値および F0 最小値を応答変数とする一般化線形混合モデルを構築した。なお，話者の違いによる影響を誤差項としてモデルに加えてある。分析対象は前節の分析と同じ IP である。F0 最大値については IP 直前の節境界の種類で，F0 最小値は IP 直後の節境界の種類でグループ分けを行った上でモデルに当てはめた。

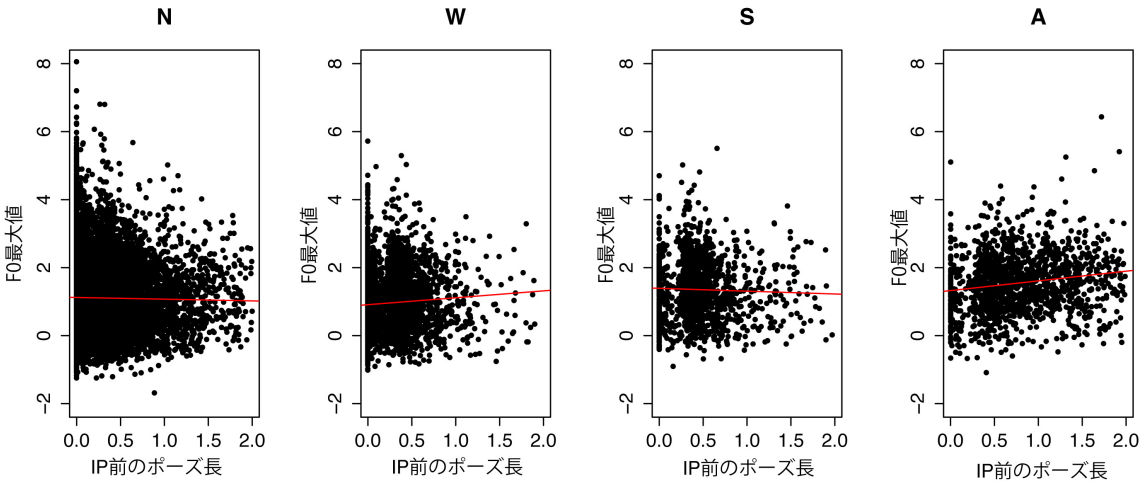


図 4　IP 直前のポーズの長さと F0 最大値（図上部の N, W, S, A は IP 直前の節境界）

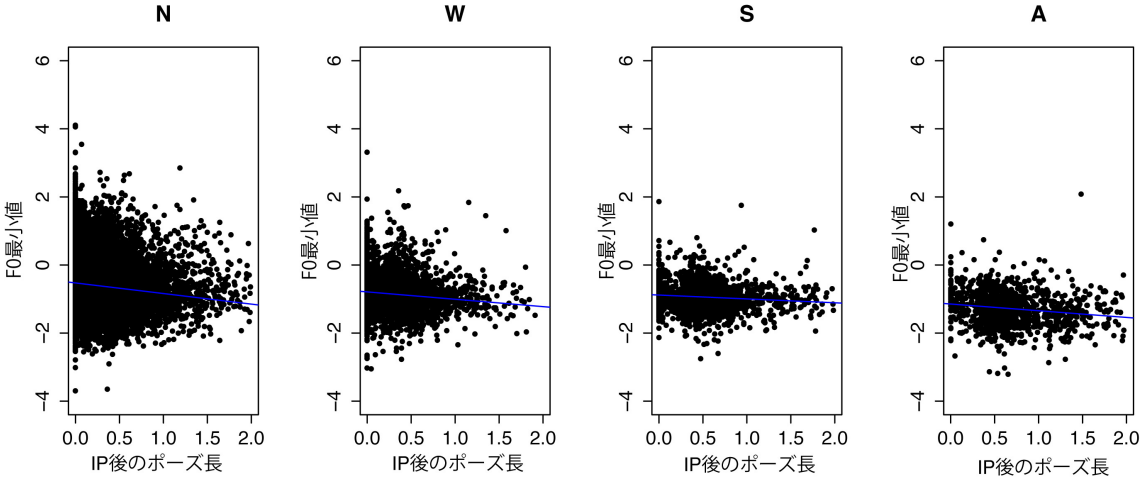


図 5　IP 直後のポーズの長さと F0 最小値（図上部の N, W, S, A は IP 直後の節境界）

4.2　結果と考察

4.2.1　IP 直前の節境界と F0 最大値

IP 直前の節境界ごとの F0 最大値に対する結果を表 2 に示す[*2]。なお，直前が強境界および絶対境界の場合の節単位末までの時間は節単位長に等しく，加えて絶対境界の場合は発話末までの時間と発話長も等しくなるため，それぞれ除外している。

表 2 から，すべての節境界に共通して，IP 長が長いほど当該 IP の F0 最大値は高くなる傾向にあることが分かる。この結果は，IP 内の AP のダウンステップで F0 が下がりすぎないよう，長い IP の場合は IP 冒頭を高めの F0 で始めるといった調整が，場所を問わず共通して見られることを意味する。

[*2] 一般化線形混合モデルの構築には R の `lme4` パッケージに収録されている `lmer` 関数を用いた。また，p 値は `languageR` パッケージの `pvals.fnc` 関数で算出した。

表 2　一般化線形混合モデルの説明変数（応答変数: F0 最大値）

	IP 直前が非境界 (N)					IP 直前が弱境界 (W)				
	Estimate	Std. Error	t 値			Estimate	Std. Error	t 値		
(Intercept)	1.012	0.018	56.99	p<0.01	**	0.909	0.022	40.86	p<0.01	**
前ポーズ長	0.012	0.006	2.03	p=0.04	*	0.053	0.015	3.43	p<0.01	**
後ポーズ長	-0.074	0.006	-12.65	p<0.01	**	-0.115	0.016	-7.03	p<0.01	**
発話長	-0.045	0.008	-5.84	p<0.01	**	-0.010	0.020	-0.50	p=0.61	
節単位長	-0.197	0.008	-24.76	p<0.01	**	-0.053	0.022	-2.43	p=0.02	*
IP 長	0.282	0.006	51.33	p<0.01	**	0.285	0.016	17.53	p<0.01	**
発話末まで	0.048	0.008	6.11	p<0.01	**	-0.004	0.021	-0.19	p=0.85	
節単位末まで	0.221	0.008	28.06	p<0.01	**	0.124	0.022	5.55	p<0.01	**
	IP 直前が強境界 (S)					IP 直前が絶対境界 (A)				
	Estimate	Std. Error	t 値			Estimate	Std. Error	t 値		
(Intercept)	1.316	0.038	34.69	p<0.01	**	1.32	0.051	26.07	p<0.01	**
前ポーズ長	0.019	0.020	0.92	p=0.36		0.104	0.014	7.154	p<0.01	**
後ポーズ長	-0.043	0.028	-1.56	p=0.12		-0.040	0.025	-1.58	p=0.11	
発話長	-0.032	0.032	-1.00	p=0.32		0.064	0.029	2.18	p=0.03	*
節単位長	0.044	0.030	1.44	p=0.15		-0.001	0.028	-0.04	p=0.97	
IP 長	0.317	0.023	13.95	p<0.01	**	0.283	0.021	13.55	p<0.01	**
発話末まで	0.037	0.035	1.06	p=0.29						
節単位末まで										

**: 1% 水準，*: 5% 水準，+: 10% 水準

IP 直前が非境界の場合，上記に加えて次の傾向が見られる。まず，節単位末までの時間が長いほど，つまり節の前部に位置するほど，F0 最大値は高い傾向にある。この結果は，非境界の多くで単純に F0 が下降する傾向が見られることを意味する。発話末までの時間も同じく有意であることから，発話の残りの時間が長いほど F0 が高いという，発話全体を通した declination の存在もうかがえる。節単位長や発話長の影響もみられるが，節単位や発話が長ければそれだけ IP 単位の下降の頻度が増すため，F0 が下がり切らないよう下がり幅（下降の傾き）を調整していることを表していると考えられる。また，IP 直前のポーズが長いと F0 最大値が高くなるという傾向も見られる。この結果は，非境界の多くは下降傾向を見せるものの，長いポーズが見られる場合には，そこで IP 単位の F0 下降がリセットされる可能性があることを示唆する。

IP 直前が弱境界の場合も，非境界とほぼ同様の傾向が見られるが，非境界で有意であった発話長・発話末までの時間の効果は見られない。非境界も弱境界も，長いポーズのあとでは F0 下降はリセットされる傾向が示唆されるが，図 3 に示したように，そもそも非境界ではポーズが置かれることは少なく（中央値 0），結果として F0 下降がリセットされることも相対的に少ないと考えられる。その結果，統計的な観点から，発話全体を通した declination の存在が非境界にのみ見られたものと推測される。

IP 直前が強境界の場合（当該 IP が節単位冒頭に位置する場合），非境界や弱境界とは異なり，その上位階層に位置する単位（強境界の場合は発話）の長さや単位末までの時間の影響は

表 3　一般化線形混合モデルの説明変数（応答変数: F0 最小値）

	IP 直後が非境界 (N)					IP 直後が弱境界 (W)				
	Estimate	Std. Error	t 値			Estimate	Std. Error	t 値		
(Intercept)	-0.582	0.015	-39.72	p<0.01	**	-0.787	0.017	-45.63	p<0.01	**
前ポーズ長	0.023	0.004	6.22	p<0.01	**	0.017	0.007	2.31	p=0.02	*
後ポーズ長	-0.089	0.005	-17.10	p<0.01	**	-0.063	0.008	-7.74	p<0.01	**
発話長	-0.022	0.006	-3.74	p<0.01	**	-0.010	0.010	-0.94	p=0.35	
節単位長	-0.092	0.006	-15.07	p<0.01	**	-0.029	0.010	-2.80	p<0.01	**
IP 長	-0.159	0.004	-35.69	p<0.01	**	-0.060	0.007	-8.19	p<0.01	**
発話末まで	0.035	0.006	5.72	p<0.01	**	0.024	0.010	2.29	p=0.02	*
節単位末まで	0.111	0.006	17.92	p<0.01	**	0.039	0.011	3.65	p<0.01	**
	IP 直後が強境界 (S)					IP 直後が絶対境界 (A)				
	Estimate	Std. Error	t 値			Estimate	Std. Error	t 値		
(Intercept)	-0.873	0.025	-35.57	p<0.01	**	-1.177	0.039	-30.49	p<0.01	**
前ポーズ長	0.004	0.009	0.43	p=0.67		0.0007	0.015	0.05	p=0.96	
後ポーズ長	-0.042	0.008	-5.40	p<0.01	**	-0.045	0.011	-4.13	p<0.01	**
発話長	-0.010	0.012	-0.82	p=0.41		-0.020	0.019	-1.06	p=0.29	
節単位長	-0.019	0.011	-1.68	p=0.09	+	0.0004	0.018	0.02	p=0.98	
IP 長	-0.034	0.008	-4.34	p<0.01	**	0.006	0.013	0.50	p=0.61	
発話末まで	0.013	0.011	1.20	p=0.23						
節単位末まで										

**: 1% 水準，*: 5% 水準，+: 10% 水準

見られないことから，強境界をはさんで F0 は単純に下降せずリセットする傾向にあること，またそのリセット後の高さが発話の長さや発話中の位置に因らないことを意味する。また強境界では IP 直前のポーズ長の影響は見られない。つまり，強境界ではポーズの長さに関わらず F0 下降のリセットが生じる可能性が高いことを意味する。

IP 直前が絶対境界の場合（当該 IP が発話冒頭に位置する場合），発話長が長いほど F0 最大値が高い傾向にあることから，長い発話単位の場合，末尾で F0 が下がり切らないよう発話の冒頭をより高い F0 で開始する傾向にあることが示唆される。また絶対境界の場合，強境界とは異なり IP 直前のポーズ長の影響が見られる。発話末では通常ポーズが置かれるが，まれに次の発話冒頭（接続詞など）までポーズを置かず連続して発話することがある。そのような場合，F0 下降は発話冒頭要素まで続き，そのあとにリセットが生じる可能性もある。あるいは，上述の通りより長い発話をする場合はより高い F0 で発話を開始するが，長い発話のプランニングのためにポーズが相対的に長くなった結果，ポーズ長と発話冒頭の F0 最大値との間に相関が見られた可能性もある。

4.2.2　IP 直後の節境界と F0 最小値

IP 直後の節境界ごとの F0 最小値に対する結果を表 3 に示す。IP 直後の節境界が絶対境界／強境界の場合とは，当該 IP は発話／節単位の末尾に位置することを意味する。なお，直後が強境界および絶対境界の場合の節単位末までの時間は IP 長に等しく，加えて絶対境界の場合

は発話末までの時間も IP 長に等しくなるため，それぞれ除外している。

結果から，絶対境界以外では IP 長の影響がみられ，IP が長いほど末尾の F0 は低い値となることがわかる。前節の分析で，IP 内の AP のダウンステップで F0 が下がりすぎないよう，長い IP では冒頭を高めの F0 で始めるといった調整がとられる可能性を指摘したが，その調整だけで IP 末尾の F0 値が一定に納まるわけではなく，長い IP では F0 がより低い位置まで低下することを意味する。絶対境界直前（発話末）で IP 長の影響が見られないのは，発話の F0 の下限が固定的であるという小磯・石本（2012）の結果と整合的である。当該話者の発話のピッチレンジの下限よりも低くなることはなく，下げ止まりの状態になるということであろう。

IP 直後が非境界と弱境界の場合，IP 長と後続ポーズ以外の結果は，前節の F0 最大値と同じ傾向を示しており，同様の考察を導くことができる。つまり，非境界および弱境界ではポーズによって F0 下降のリセットが生じることはあるものの，総じて F0 下降が続く位置と言える。

IP 直後が強境界の場合（当該 IP が節単位末尾に位置する場合），前節の F0 最大値と同様に発話末までの時間の影響がないことから，強境界をはさんで F0 は単純に下降せずリセットする傾向にあることが示唆される。また節単位長が長くなると F0 最小値はより低くなるという傾向も見られる。強境界の場合，リセット直前の F0 最小値は発話末ほどは下がらず，また値も一定しないが（小磯・石本 2012），リセット直前の F0 最小値には節単位の長さが関わる可能性が示唆される。

IP 直後が絶対境界の場合（当該 IP が発話末尾に位置する場合），ポーズ長以外の時間のパラメータの影響を全く受けていないことから，前述のように発話末の F0 の下限はかなり固定的であることが分かる。

最後に，すべての節境界に共通して IP 直後のポーズ長が有意であり，直後に長いポーズがあると F0 最小値は低くなる傾向にあることがわかる。強境界の場合，上述の通り節単位長が長いほどリセット直前の F0 最小値はより低くなる傾向にある。仮に長い節単位ほど末尾での息の残量が少なくなり，直後により長い息継ぎの時間（ポーズ）を置く傾向があるとするならば，節単位の長さを媒介に IP 直後のポーズ長と直前の F0 最小値との間に相関が見られたと考えることができる。その他の境界の場合も，同様の可能性が考えられる。

5. おわりに

本稿では，小磯・石本（2012）および石本・小磯（2012）で観察された IP 単位の F0 下降現象と強境界での下降のリセットが統語的な要因だけに依存しているのかを調べるため，息継ぎ（ポーズ長）や発話時間などの物理的な特徴量を含めた総合的な分析を行った。その結果，小磯・石本（2012）および石本・小磯（2012）で観察された各種傾向が改めて統計的に確認されたほか，新たに次のことが明らかになった。

1. 非境界・弱境界では通常，発話（節単位）末に向けて F0 が低下する傾向が見られるが，長いポーズが置かれると，ポーズ前で F0 が低下しポーズ後に高い F0 で始まる「F0 下降のリセット」が生じる可能性がある。
2. 強境界では，明らかに単純な F0 下降とは異なる F0 変動をしており，F0 下降のリセットが起きていると考えられるが，F0 最大値にはポーズの影響がみられないことから，

ポーズの長さによらず F0 下降のリセットが生じている可能性が高い。強境界のように発話内容の意味的なまとまりの強い区切りでは，ポーズの有無といった生理的要因に関わらずリセットが生じることが示唆される。

3. 節境界で F0 下降のリセットが生じる場合，リセット直前の F0 最小値は発話末ほどは下がらずまた値も一定しないが，この F0 最小値に節単位長が関わる可能性がある。
4. 発話が長いほど発話冒頭の F0 最大値が高くなるのと同様に，IP が長いほど IP 冒頭の F0 最大値が高くなることから，発話・IP という大小異なる単位で，発話前にプランニングをして各単位中で F0 が必要以上に下がりきらないような調整が行われている可能性がある。

このように，発話に見られる IP 単位の F0 下降やそのリセットに関して多くの事実が徐々に明らかになってきたが，発話中に見られる F0 下降のリセットを引き起こす要因や発話全体に見られる declination との関係などについてはまだ解明されていない。この点については今後の課題としたい。

参 考 文 献

小磯花絵，石本祐一 (2012) 「日本語話し言葉コーパスを用いた「発話」の韻律的特徴の分析 –イントネーション句を切り口として–」第 1 回コーパス日本語学ワークショップ予稿集, pp. 167–176.

石本祐一，小磯花絵 (2012) 「日本語話し言葉コーパスを用いた統語境界におけるイントネーション句変動の分析」第 2 回コーパス日本語学ワークショップ予稿集, pp. 239–246.

Pierrehumbert, Janet B. and Mary E. Beckman(1988) *Japanese tone structure*, Cambridge: MIT Press.

Kawahara, Shigeto and Takahiro Shinya (2008) "The intonational of gapping and coordination in Japanese: evidence for intonational phrase and utterance," *Phonetica*, 65, pp. 62–105.

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.

小磯花絵，伝康晴，前川喜久雄 (2012) 「『日本語話し言葉コーパス』RDB の構築」第 1 回コーパス日本語学ワークショップ予稿集, pp. 393–400.

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.

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

※ 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。

日本語話者の英語発話にみられる日本語の音節構造と母音の無声化との関係 - Japanese AESOP コーパスの分析から

近藤眞理子（早稲田大学国際教養学部）
鍔木元（早稲田大学国際情報通信研究科）

Relationship between Syllable Structure and Vowel Devoicing in Japanese Speakers' English –Analysis of Japanese AESOP Corpus

Mariko Kondo (SILS and LASS, Waseda University)
Hajime Tsubaki (GITS and LASS, Waseda University)

１．はじめに

本研究は２００８年よりアジア各国の研究機関との共同プロジェクトとして進行しているアジア言語話者の英語発話コーパス構築プロジェクト（AESOP： Asian English Speech cOrpus Project）(Meng et al., 2009; Vischeglia et al., 2009等) の日本語話者の英語の音声発話のデータをもとに、日本語の音韻特性を考察したものである。AESOPの日本語話者の発話データは現在約１６０人分のデータが集められ、現在も収集中である(Kondo, 2012)。臨界期を過ぎた外国語学習者には、通常母語の音韻特性が音声知覚と産出の両面で現れる（Lenneberg, 1964; Patkowski, 1989）。また第一言語の音韻特性は音素やフレーズなど、様々な音韻単位および、リズムやイントネーションなどの韻律面で顕著に現れ、どの音韻単位における間違いも韻律の乱れもすべてコミュニケーションにとって重要である。個々の音素の発音の正確さは単語の判別には重要であるが、第二言語の発話の流暢さの判定評価により大きく影響を及ぼすのは、韻律面の正確さである (Anderson-Hsieh et al., 1992)。実際のコミュニケーションでは、韻律が意味の強調やフレーズ境界、統語構造、スピーチアクト、また話者の感情や態度などの伝達をつかさどっている(Prince et al., 1991; Hirschberg, 2002; Grice & Bauman, 2007)。したがって第一言語が第二言語に韻律上間違った影響を与えると、意味や発話の意図、感情の理解に誤解を生じさせ、意思の疎通の妨げとなる。しかし、第二言語の韻律習得の重要性とは裏腹に、これまで第二言語の音声習得の研究は、音素などの個々の音の習得が中心となっていたので (Jilka, 2007)、AESOPの発話コーパスでは、英語の韻律習得を主な研究目的の一つとして構成した。したがってAESOPコーパスを用いて日本語話者の英語発話の韻律上の問題を検証することで、韻律特性を含む日本語の音声特性を浮き彫りにすることができる。

今回の分析では、日本語話者の英語の発音の間違いの中から、母音の挿入に焦点を当て、音節構造と母音の無声化現象について考察し、日本語の発話リズムの特性を検証する。

２．手法

今回の分析には、すでに収録済みの日本語母語話者約 160 人の英語発話データのうち、分析が済んでいる関東方言話者５０人分の ”The North Wind and the Sun” (International Phonetic Association, 1999) の読み上げ文を対象とした。被験者は日本語を母語とする大学生、大学院生である。英語のレベルは、日本の大学レベルで英語教育の経験がある英語教員（英語母語話者三名、日本語母語話者五名）に、発話の自然さ、流暢さ、音の正確さ、英語らしさ全体に関して主観的に評価をしてもらった。判定の基準は、レベル１（英語として理解が困難）；レベル２（英語として発音は良くない）；レベル３（ごく平均的日本語話者の英語の発音）；レベル４（英語としてとても自然で上手な発音）；レベル５（英語母語

話者レベルの発音）で、各レベルの中間値 (.5) を設け、9 段階評価とした。50 人の被験者の判定レベルの内訳は表 1 の通りである。

表１　被験者の英語の発音レベル主観評価判定基準と各レベルごとの被験者数

レベル	1	1.5	2	2.5	3	3.5	4	4.5	5
判定基準	very poor		poor		average		good		very good
人数	0	2	10	11	10	11	5	1	0

音声データに対して、隠れマルコフモデル(HMM)に基づいた音響モデル、HTK モジュール（URL 参照）及び単語辞書を用いた自動音素アラインメントを実施した。本アラインメントは分析周期 10ms、窓幅 20ms で行われ、ARPABET 表記（URL 参照）の音素列が出力結果となる。従来、アラインメントに使用される単語辞書は、標準アメリカ英語の発話コーパス TIMIT（URL 参照）に準拠しているため、日本語訛を含む英語発話データのアラインメントには必ずしも適しておらず、正確なアラインメントができないデータも存在する。そのため、本アラインメントにおいては、”The North Wind and the Sun”に出現する全ての単語について、日本語母語話者の英語発話を想定した音素列パターンを単語辞書データに追加し、日本語訛を含む英語音素列もほぼ正確に抽出できるようにした（表 2)。自動アラインメントに手修正を加えた結果をもとに、日本語母語話者の英語の発音がどのように異なるか検証した。

表２　単語発話の追加音素列パターン　“*stronger*” の例

英語母語話者の発音	[stɹɑŋgɚ]	ARPABET
日本語話者の予測発音	[stɹɑŋgɚ]	s t r aa ng g er
	[stɹɔŋgəɹ]	s t r ao ng g er
	[stɹɔŋgə]	s t r ao ng g ah
	[stɔɾɔŋgaː]	s t ao r ao ng g aa
	[sʊtɔɾɔŋgaː]	s uh t ao r ao ng g aa
	[stlɑŋgɚ]	s t l aa ng g er
	[stlɔŋgəɹ]	s t l ao ng g er
	[stlɔŋgə]	s t l ao ng g ah

３．結果と考察

３．１　音素の逸脱

自動アラインメントの結果、発音のモデルとした標準アメリカ英語から逸脱した発音と分析されたものが 2,480 例抽出された（図１）。個別の例で一番多かったものは/r/の脱落で、この例のほとんどは音節の尾子音の/r/を発音せず、/r/に先行する音節核の母音を伸ばすことにより長母音として発音している例であるが、これは今回自動アラインメントの規範の発音として使った標準アメリカ英語が rhothic アクセントであるために、規範とは異なる発音とされているのであって、母音に後続する尾子音の/r/の脱落は、必ずしも間違いではない。したがって、/r/の脱落を除いた日本語話者の英語の発話特徴を分析した。結果は、従来から

指摘されている日本語話者の英語の発音の問題点、つまり母音の音質、母音の挿入、子音の音質、子音の脱落、子音の挿入などを映し出している。例えば/l/ → /r/, /b/ → /v/ , /s/ → /θ/ など、日本語にない子音音素を日本語にある音声的に近い別の子音で代用している例は1,216 例あった（表３）。子音の音素では、日本語にない/l/が/r/([ɹ], [ɻ], [ɾ]) に, /v/が/b/に, /θ/が/s/に, /ð/が[dz], [dz̪], [dʒ], [z]などの日本語の音素または異音に置き換わっている例が非常に多く見られた。また英語の発音主観評価値とモデル発音からの逸脱とみなされた音素の数には強い相関がみられた（R=－0.454; p<0.005）。

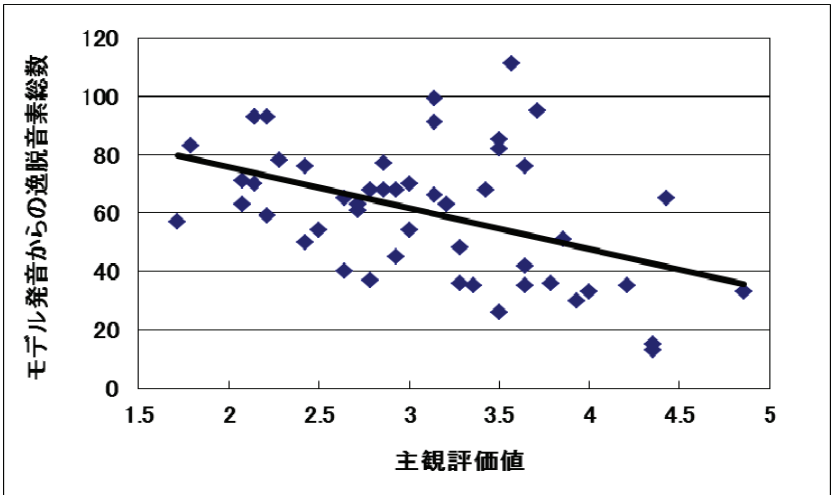


図１　モデル発音からの逸脱とみなされた音素の数と英語の発音主観評価値との関係

表３ 他の音素に置き換わった子音の総数 1,216 例中の
主な英語子音音素の代用

英語音素	代用された子音	サンプル数
/l/	/r/ ([ɹ], [ɻ], [ɾ])	203
/r/	/l/	64
/ð/	[dz], [dz̪], [dʒ], [z]	142
/v/	/b/	24
/θ/	/s/	78
/r/	脱落	546

３．２　母音の逸脱

子音の間違いよりも多かったのが母音の問題で、母音の音質にかかわるものが 1,346 例、母音の挿入が 463 例あった。母音の音質に関しては、/ə/ の調音位置が中央でなく、 [a] (105 例）、[ʌ]（292 例）、[ɔ]（55 例）、[i, ɪ]（30 例）、[e]（18 例） などと調音周辺部の母音として調音されているものが多かった。英語の /ə/ は、強勢の置かれていない弱音節に起きるが、日本語話者が日本語にはない /ə/ を、日本語の五母音のどれかに分類し、強母音として発音していることが伺える（第４節参照）。その他、多く見られた母音の間違いは/ʌ/

（281 例）で[a]と認識された例が 164 例、/æ/ の間違いが 107 例で、うち[a]（52 例）か[ʌ]（44 例）と認識されたものが殆どで、続いて /ɔ/ の間違いが 94 例で[a]（61 例）と[ʌ]（29 例）として認識されたものが大半であった。

３．３　母音の挿入と音節構造

母音の間違いで非常に多かったのが、母音の挿入で 463 例が抽出された（図２）。母音の挿入は日本語と英語の音節構造の違いからくるものである（表４）。英語の子音の連続の環境（/$C_1C_2(C_3)$/）において、実際の英語の音節構造にかかわらず、日本語話者は母音を挿入することにより/C_1V.C_2V.(C_3V)/と軽音節/CV/に分析することに起因している。第二言語の音節構造に第一言語の音節構造では許容されない構造がある場合、話者は様々な発音のストラテジーを試みるが、日本語母語話者が第二言語で子音の連続に接した場合、子音を省略したり融合させ新しい子音で代用するのではなく、子音間に母音を挿入し音節構造を/CV/に再構成するという手段をとる。また日本語の音節構造は基本的に開音節で、閉音節はモーラ子音 /N/ と /Q/ が音節末に来るとき以外は起きない。したがって、第二言語の語末が閉音節のときも、語末に母音を挿入することで/-CV/という新しい音節を構成する。

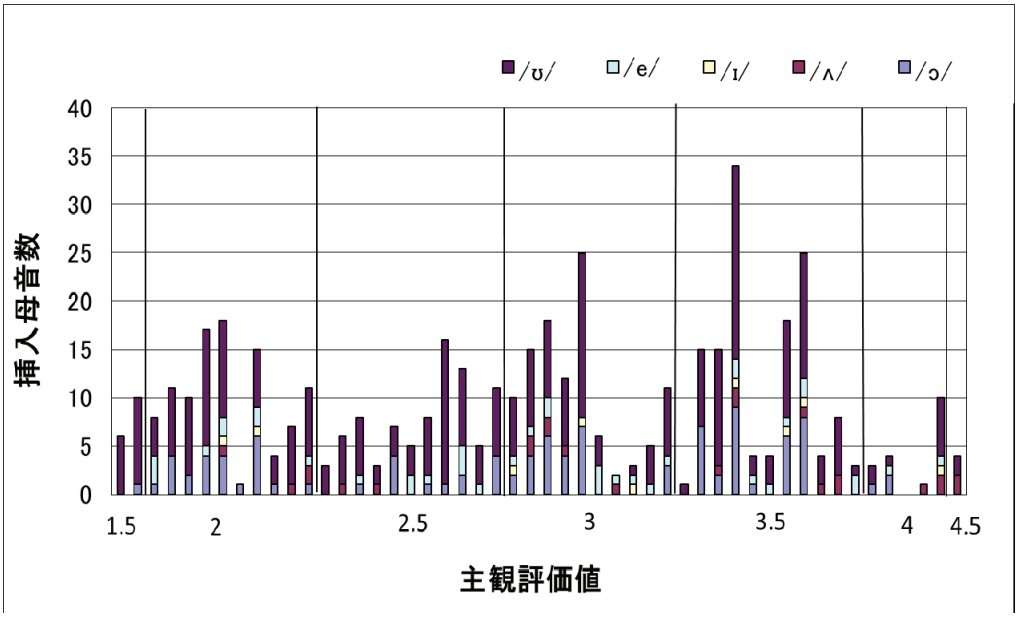


図２　英語レベル別挿入母音数と母音の種類

表４　日本語と英語の音節構造の比較

日本語	英語
(C)(j)V(V)(C)(C)	(C)(C)(C)V(C)(C)(C)

AESOP コーパスの ”The North Wind and the Sun” には、子音の連続が起きる単語および語末に子音が来る単語が数多くあるが、すべての母音の挿入例が見られた（表５）。

他に wind [ʊ̲ɪndɔ], warm [ʊ̲ɔːmʊ] にみられるように、/w＋ 母音/ のときで、母音が/a/ 以外のとき、/w/が子音/w/でなく母音/u/となり、/wɪn-/ , /wɔːm/という閉音節が、/ʊ.ɪn-/, /ʊ.ɔː.mʊ/ と後続母音とは別の音節に構造が変わっている例も見られた。

表５ Japanese AESOP “The North Wind and the Sun” の挿入母音の例

(a) 連続子音間（下線の母音が挿入母音）

disputing	[dɪsʊ̲pju:tɪŋ]	*attempt*	[ɑtempʊ̲tə̲]	*warmly*	[waʊmʊ̲li:]
wrapped	[rɑpʊ̲t]	*cloak*	[kʊ̲raʊk]	*stronger*	[sʊ̲tə̲rəŋgɑ:]
obliged	[əbʊ̲raidʒɪ̲d]	*fold*	[fourʊ̲d]	*immediately*	[ɪmedɪətə̲ri:]
agreed	[ʌgʊ̲ri:d]	*blue*	[bʊ̲lu:]		
succeed	[sɑkʊ̲si:d]	*closely*	[kʊ̲rowsʊ̲ri:]		

(b) 語末（二重下線部が挿入母音）

first	[fa:stə̲]	*succeed*	[sɑkʊ̲ʃi:də̲]	*came*	[keɪmʊ̲]
wrapped	[ræptə̲]	*agreed*	[ʌgʊ̲li:də̲]	*warm*	[wɑ:mʊ̲]
attempt	[ɑtempʊ̲tə̲]	*take*	[teɪkʊ̲]	*his*	[hidzʊ̲]
last	[rɑstʊ̲]	*took*	[tʊkʊ̲]	*was*	[wɑzʊ̲]
obliged	[ʌbʊ̲raɪdʒidʊ̲]	*up*	[ʌpʊ̲]	*as*	[ædzʊ̲]
should	[ʃʊdʊ̲]	*confess*	[kʌnfesʊ̲]	*along*	[ɑrəŋgʊ̲]
fold	[faʊldʊ̲]	*off*	[ɔfʊ̲]	*disputing*	[dɪsʊ̲pju:tɪŋʊ̲]
around	[ɑraʊndʊ̲]	*gave*	[ge:vʊ̲]	*making*	[meɪkɪŋʊ̲]

挿入された母音は弱母音の[ɔ], [ʊ]が圧倒的に多い（図２）。日本語への外来語での挿入母音の規則は、歯茎破裂音/t, d/ の後は/o/, 破擦音/tʃ, dʒ/ の後は /i/, その他の子音の後は/u/である（Shinohara, 2004）。しかし実際の英語の発話ではいわゆるカタカナ語で英語を発音しているわけではなくとも、日本語の/CV/の音節構造を根底に英語を発話してしまい、はっきりとではないが子音の連続の間に無意識に/t, d/の後に[ɔ]、/tʃ, dʒ/の後に[ɪ]、その他の子音のあとには[ʊ]が入ってしまうと考えられる。特に、一番例が多かった[ʊ] の挿入は、話者の英語レベルとの負の相関がみられなかった（R=－0.205; p>0.1）(図３)。[ɔ]、[ʊ]は日本語の母音/o/と/u/の実際の音質に近いことを考えると、両母音が実際によく挿入されるのは納得がいく。

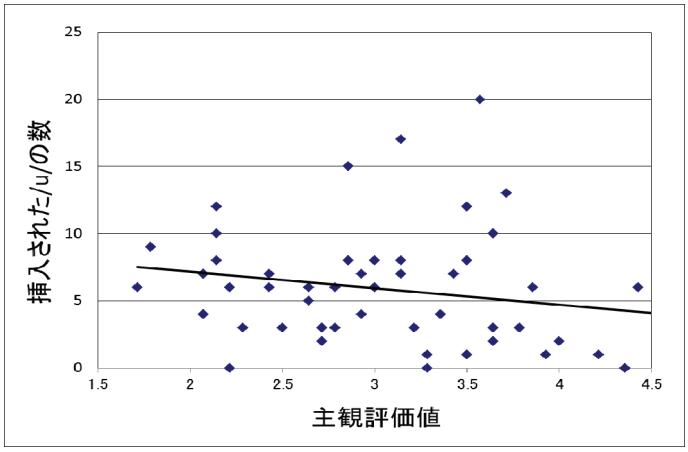


図３ 英語レベル別の子音間または語末の/u/の挿入

さらに、高母音/u/は/i/とともに、標準語を含む東日本を中心とする多くの方言で無声子音間、または無声子音が先行する発話末では無声化する。 “The North Wind and the Sun” のデータでは無声子音の連続が少なからずあり、うち高母音/u/が挿入される可能性がありかつその/u/が無声化される可能性のある環境があるのは、*attempt* /ateNpu̥to/, *disputing* /disu̥pjuutiNgu/, *first* /faasu̥to/, *last* /rasu̥to/, *stronger* /su̥toroNgaa/, *succeed* /saku̥siido/, *wrapped* /raQpu̥to/であるが、*first*t と *last* 以外の *disputing* [dɪsʊ̥pju:tɪŋʊ] (5 例)、 attempt [ətempʊ̥tɔ] (10 例)、 stronger [sʊ̥tɔraŋgər] (20 例)、 succeed [səkʊ̥si:d] (8 例)、 wrapped [ræpʊ̥tɔ] (6 例)と典型的な無声化環境で/u/に準ずる母音の挿入（下線部）が少なからずみられた。無声化環境での母音/u/の挿入は、音節構造と音節内での位置にかかわらず見られた（音節末/dɪs**u̲**.pju:t.ɪŋ**u̲**/, /sək**u̲**.si:d/、頭子音 /s**u̲**trɑŋ.ə-(r)/, 尾子音 /ə.tempu̲t/, /ræp**u̲**t/）。この傾向は英語のレベルが高い話者にも少なからずみられた。日本語話者にとって/(C)V/という日本語の絶対的な音節構造があり、それが第二言語の発話においても強く作用しており、必ずしも英語の音節構造では音素列を分析していないであろうことが推測される。今回 ”*first*” と ”*last*” の尾子音の /-st/ では /u/ の挿入例が見られなかったが、歯茎摩擦音が先行し、無声破裂音が後続する環境での無声化率が一般的に高いことから（武田&桑原, 1987; 吉田&匂坂, 1990; Maekawa & Kikuchi, 2005）、無声化が特に起きやすい環境であることが要因であろう。同じ/st/の連続でも ”stronger” では挿入母音の例が多くみられることから、シラブル内での位置や、二連続子音と三連続子音での挿入母音の生起率が異なるのかは、これからの検討課題である。

これらの結果は、日本語話者の英語発話で、ストレスが置かれていない音節で母音の弱化が起きにくいという研究報告（Lee et al., 2006; Kondo, 2008; Sugahara, 2009）や、日本語発話においては単一の無声化環境での高母音の無声化率はほぼ 100%近いという研究報告（Kondo, 2005）を考慮すると、日本語話者の日本語及び第二言語の知覚と発話、特にリズムにおいて、如何に /CV/ という基本の音節構造またはモーラが強く影響しているかがわかる。また、この音節構造にかかわる問題は、英語レベルの高い話者にもみられる問題であることから、Mazuka et al. (2011) の研究結果にみられるよう音節構造は第一言語習得の根幹をなし、第二言語習得のあらゆる面に影響を与えていると考えられる。

４．まとめ

日本語話者の英語発話の分析から、日本語話者にとって日本語の音素にない英語の音の発話は母音も子音も難しいが、個々の音素の発音の問題は、大方英語のレベルが上がるにつれて解消されていく。しかし、レベルが上がっても根深く残る問題の一つが子音間と語末の閉音節の後の母音の挿入である。母音の挿入は、/CV/ を基本とする日本語の音節構造を保とうとすることから起きていると推測されるが、この影響は単なる音節構造という抽象的な音韻理論の問題にとどまらず、英語発話に際し、フットを基本とした強勢リズムや強勢リズムに付随して起きる弱音節での母音の弱化などの発話リズム全体の問題にかかわってくる。反面、日本語話者の英語発話を考察することで、日本語話者にとっての音素の認識や、韻律上重要な単位は何かなどの、日本語の音韻に関する根本的な問題の答えが見える。Japanese AESOP の分析がさらに進んでいけば、日本語話者の英語発話の問題だけでなく、日本語そのものの特性の解明にもつながっていくであろう。

謝辞

本研究は、文部科学省科学研究費補助金基盤研究（B）「第一言語の韻律特性が日本語学習者の音声知覚・生成に及ぼす影響の解明」（平成 22～25 年度、研究代表者：近藤眞理子）による補助を得ています。

参考文献

Anderson-Hsieh, Janet, Ruth Johnson and Kenneth Koehler (1992) “The Relationship between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentals, Prosody and Syllable Structure”, *Language Learning*, 42:4.

Grice, Martine and Stefan Bauman (2007) “An Introduction to Intonation –Functions and Models,” in *Non-native Prosody: Phonetic Description and Teaching Practice*, Trouvain, J. and Ulrike, G. (eds), pp. 25-52, Berlin: Mouton de Gruyter.

Hirschberg, Julia (2002) “Communication and Prosody: Functional Aspects of Prosody,” Speech Communication, Volume 36, Issues 1-2, pp. 31-43.

International Phonetic Association (1999) Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet, Cambridge: Cambridge University Press.

Jilka, Matthias (2007) “Different Manifestations and Perceptions of Foreign Accent in Intonation,” in *Non-native Prosody: Phonetic Description and Teaching Practice*, Trouvain, Jurgen and Ulrike, Gut (eds), pp. 77-96, Berlin: Mouton de Gruyter.

Kondo, Mariko (2005) “Syllable Structure and its Acoustic Effects on Vowels in Devoicing Environments”, *Voicing in Japanese*, Van De Weijer, Jeroen M., Kensuke Nanjo and Tetsuo Nishihara (Eds.). pp.205-228.

Kondo, Mariko (2009) “Is Acquisition of L2 Phonemes Difficult?; Production of English Stress by Japanese Speakers”, *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference (GASLA 2009)*, Melissa Bowles, M., Ioin, T. Montrul, S. and Tremblay, A. (eds.). Somerville, MA: Cascadilla *Proc.* Project. 105-112.

Kondo, Mariko (2012) “Design and Analysis of Asian English Speech Corpus - How to Elicit L1 Phonology in L2 English Data”, In Tono, Yukio, Yuji Kawaguchi and Makoto Minegishi (Eds.). Vol. IV: *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins. 251-278.

Lee, Borim, Susan G. Guion, and Tetsuo Harada (2006) “Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals”, *Studies in Second Language Acquisition*, 28, 487-513.

Lenneberg, Eric (1967) *Biological Foundations of Language*. New York: John Wiley & Sons.

Maekawa, Kikuo and Hideaki Kikuchi (2005) “Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report”, *Voicing in Japanese*, Van De Weijer, Jeroen M., Kensuke Nanjo and Tetsuo Nishihara (Eds.). pp.205-228.

Mazuka, Reiko, Cao, Yvonne, Dupoux, Emmanuel, and Christophe, Anne (2011) "The development of phonological illusion: A cross-linguistic study with Japanese and French infants", *Developmental Science*, 14 (4), pp.693-699.

Meng, Helen, Chiu-yu Tseng, Mariko Kondo, Alissa Harrison and Tanya Visceglia (2009) “Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper”, *Proceedings of 2009*

INTERSPEECH (Brighton, UK, 6-10 September 2009). 1715-1718.

Patkowski, Mark (1989) Age and accent in a second language: a reply to James E. Flege. *Applied Linguistics* 11: 73-89.

Price, Patti, Mari Ostendorf, Stefanie Shattuck-Hufnagel and Cynthia Fong (1991) “The Use of Prosody in Syntactic Disambiguation," *Journal of the Acoustical Society of America*, 90(6), pp. 2956-2970.

Shinohara, Shigeko (2004) “Emergence of Universal Grammar in Foreign Word Adaptations”, *Constraints in Phonological Acquisition*. Kager, René, Joe Pater, and Wim Zonneveld (Eds.). Cambridge: Cambridge University Press. 292-320.

Sugahara, Mariko (2009) “Secondary stress vowels in American English: The target undershoot of F1 and F2 formant values”, *Proceedings of the 16th International Congress of Phonetic Sciences. Saarbrucken, Germany*. 633-636.

武田一哉、桑原尚夫（1987）「母音無声化の要因分析と予測手法の検討」日本音響学会　1987年度秋季研究発表会公演論文集、105－106.

Visceglia, Tanya, Chiu-yu Tseng, Mariko Kondo, Helen Meng, and Yoshinori Sagisaka (2009) “Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project)”, *2009 Oriental COCOSDA* (Beijing, China, 10-12 August, 2009).52-57.

吉田夏也・匂坂芳典（1990）「母音無声化の要因分析」ATRテクニカルレポート(TR-I-0159), 1-9.

ARPABET: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogID=LDC93S1

The HTK modules: http://htk.eng.cam.ac.uk/

TIMIT: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

『日本語話し言葉コーパス』における韻律単位の認定基準について

小磯 花絵（国立国語研究所理論・構造研究系）†
前川 喜久雄（国語研究所言語資源研究系）
五十嵐 陽介（広島大学）

Criteria for Intonational Unit Identification: The Case of *the Corpus of Spontaneous Japanese*

Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)
Kikuo Maekawa (Dept. Corpus Studies, NINJAL)
Yosuke Igarashi (Hiroshima University)

1. はじめに

『日本語話し言葉コーパス』（*Corpus of Spontaneous Japanese*, 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 650 時間の日本語自発音声からなるデータベースである（前川 2004,2006）。2004 年に公開を開始して以降，幅広い領域で利用されており，データを修正，追加しながら，第 2 刷（2008 年），第 3 刷（2011 年）と版を重ねている。

現在，CSJ のうち多種多様な研究用付加情報が付されたコアと呼ばれるデータ範囲（約 45 時間，50 万語）を対象に，各種情報を相互に関連付けて表現した RDB（小磯ほか 2012）を構築しており，近日中の公開を目指している。CSJ-RDB 版は，原則として CSJ 第 3 刷に含まれる XML 文書に表現されている単位や各種情報を反映したものであるが，新たにアクセント句とイントネーション句という 2 種類の韻律単位を明示的に設けることとした。

CSJ が採用している韻律ラベリング体系 X–JToBI（Maekawa et al. 2002; 五十嵐ほか 2006）およびその前身である J_ToBI（Venditti 1995, 2005）では，アクセント句やイントネーション句は単位としては明示的に表現されず，韻律境界の切れ目の強さを表す情報（Break Index，以下 BI 情報）によって間接的に表される。概ね `BI=1` は語境界，`BI=2` 類はアクセント句[*1]の境界（同時に語境界），`BI=3` はイントネーション句[*2]の境界（同時に語とアクセント句の境界）に

† koiso@ninjal.ac.jp

[*1] アクセント句は，句頭第 1 モーラから第 2 モーラ付近にかけての F0 の上昇と句末への緩やかな下降を有し，かつアクセント核による下降を最大ひとつ持ちうる単位と定義される。なお X–JToBI では，アクセント句末にポーズや上昇調などの複合境界音調（BPM）が出現する場合，`BI=2` と `BI=3` の中間値として，`BI=2+p`，`BI=2+b`，`BI=2+bp` を新規に導入している。

[*2] イントネーション句は，アクセント句の上位階層に位置し，アクセント句のピッチレンジを指定する単位と定義される。先行アクセント句と比較してピッチレンジが拡大した場合，そこにイントネーション句境界があるとみなされる。アクセント核が引き起こす後続アクセント句のピッチレンジの縮小効果は，イントネーション句の終端で阻止されることになる。なお，J_ToBI が立脚する Pierrehumbert & Beckman (1988) の音韻理論では，アクセント句より階層的に上位の単位として「中間句 (intermediate phrase)」と「発話 (utterance)」の二つが認められているが，J_ToBI における「イントネーション句」はこれらを融合させた単位である。

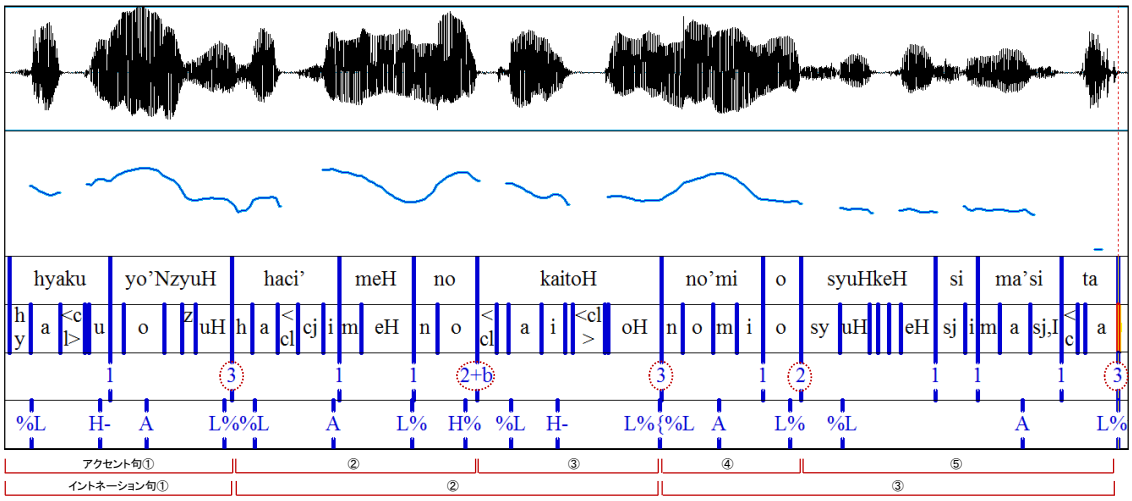


図 1　非流暢現象が存在しない発話例「１４８名の回答のみを集計しました」

相当することから，アクセント句は両端を BI=2 か BI=3 で区切られる範囲として，イントネーション句は両端を BI=3 で区切られる範囲として，認定することができる。図 1 に，X–JToBI に基づくトーンと BI 情報，およびそれに基づき同定したアクセント句とイントネーション句の範囲の例を示す。

このように流暢に発話されている部分では，BI 情報から単純に韻律単位を同定することができる。しかし CSJ が対象とするような自発性の高い発話には，図 2 にあるように，「アノ」「エット」「ンー」などのフィラーや，「わた 私が」のような言いよどみが頻出する。J_ToBI が立脚する音韻理論（Pierrehumbert & Beckman 1988）は朗読音声に基づき構築されたものであるため，このような自発性の高い話し言葉に特徴的な現象を十分に記述できないという問題があった。X–JToBI（eXtended J_ToBI）はこの種の現象への積極的な対応が検討され様々な拡張がなされたが，その結果，体系が複雑化し，アクセント句・イントネーション句の認定が単純には行かなくなった。例えば図 2 にあるフィラーを一つとっても，それを独立したアクセント句とみなすべきか，それとも先行あるいは後続するアクセント句の一部とみなすべきかはすぐには決まらない。

CSJ の第 1 刷，第 2 刷では，各種情報を統合した XML 文書の中で，単語（短単位）の属性としてアクセント句の ID 情報（APID）が表現されているが，時間的な制約もあり，非流暢現象を含めて BI 情報からアクセント句を汲み上げる方法を十分に検討することができなかった。CSJ-RDB 版を作成するにあたり，これら非流暢現象の扱いを中心にアクセント句・イントネーション句の認定基準を改めて検討し，また X–JToBI の認定基準についても一部変更を施すこととなった。そこで本稿では，検討の対象となったフィラー・言いよどみの扱いを中心に韻律単位の認定基準について報告する。

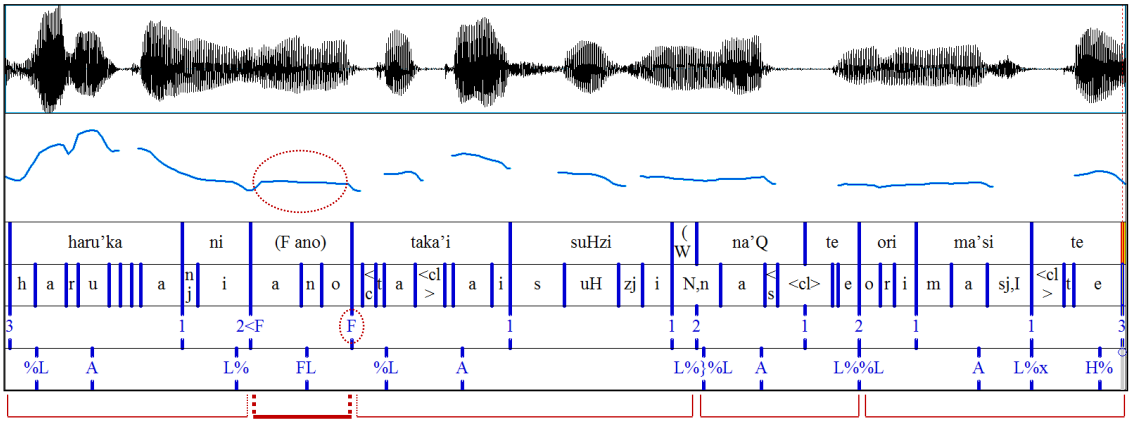


図 2　非流暢現象（フィラー）が存在する発話例「遥かに あの 高い数字になっておりまして」

2.　韻律単位の認定基準

2.1　フィラーの扱い

■対象とするフィラー　CSJ の転記テキストでは，語彙的・機能的な観点からフィラーが認定されるのに対し（小磯ほか 2006），X–JToBI では韻律的な観点からフィラーが定義される（五十嵐ほか 2006）。原則として，転記基準でフィラーと認定されるもののうち，フィラーの主観的長さが 1 モーラのもの，あるいは，句頭の上昇が認められずかつアクセント類似の局所的なピッチの下降が認められないものだけが，X–JToBI ではフィラーとみなされる。また接続詞の「で」など，転記では機能的観点からフィラーとは認定されないものであっても，長さが 1 モーラで韻律的にフィラーと似た形で現れるものは，X–JToBI ではフィラーとみなされる。

■フィラーに関わる BI 値の認定基準の概要　フィラーは次の方針でその前後の BI 値が認定されている。これはフィラーが無音区間と類似した機能を持つとし，フィラーを透過的に（存在しないものとして）扱うという立場で定めたものである[*3]。

1. フィラーに先行するアクセント句（図 2 の例では「遥かに」）の BI 値は，フィラーを透過してそれに後続するアクセント句（「高い数字に」）との関係で `BI=2` か `3` を判断した上で，フィラーの始端を示す（フィラーが後続することを示す）「`<F`」を付与し，`BI=2<F,3<F` とする。
2. フィラーの終端には「`F`」を単独で付与する。透過要素のため後続するアクセント句との韻律的接続関係は特定しない。

■フィラーに関わる韻律単位の認定方針

アクセント句：このようにフィラーは透過要素として扱われるため，現在付与されている BI 値から，フィラーを先行あるいは後続するアクセント句の一部に含めるべきか，それとも独立したアクセント句と認定すべきかは決まらない。そこで，無理にどちらかのアクセント句に含

[*3] そのような仮定の実証的根拠としては，例えば前川（2012）がある。

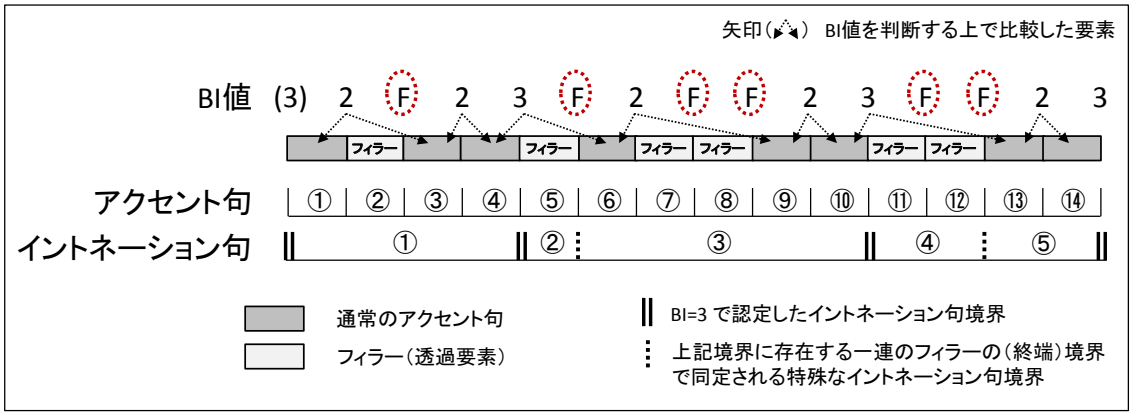


図 3　フィラーが関わる場合のアクセント句・イントネーション句の認定例

めるのではなく，独立した特殊なアクセント句（フィラー句）と認定する。連続して出現するフィラーはそれぞれ別のアクセント句とする。

イントネーション句：フィラーを透過要素とみなして BI=3 の位置でイントネーション句を認定した上で，句境界に存在するフィラー（連鎖する場合はその全体）は無理に前後のイントネーション句には含めず独立した特殊なイントネーション句として，またイントネーション句の内部に存在するフィラーは当該イントネーション句に含める形で認定する。

以上の方針で認定したアクセント句・イントネーション句の認定例（模式図）を図 3 に示す。

なお，対話データに限定して，「ウン」「フーン」「ンー」等の応答表現やあいづちの一部に対し BI 値を F2 としているが，韻律単位の認定基準については F と同じとする。

2.2　言いよどみの扱い

■対象とする言いよどみ　韻律ラベリングにおいて問題となるのは，「わた 私が」のように，言いよどみに伴い語の断片などが生じ，かつその境界（当該要素の前側あるいは後側境界，ないしその両方）に何らかの韻律上の不連続性が知覚される場合である。

言いよどみ部の後側境界に韻律的不連続が感じられる場合とは，アクセント句の終端を特徴付ける句末への緩やかな下降が見られない事例に相当する。「すじ 推定したものです」や「典型的もんだ 典型的事例は」のように，言いよどみに伴い発話を途中で言いやめたケースに典型的に見られる（図 4（A）参照）。一方，言いよどみ部の前側境界に韻律的不連続性が認められる場合とは，アクセント句の始端を特徴付ける句頭第 1 モーラから第 2 モーラ付近にかけての F0 の上昇が見られない事例に相当する。「聞き分け易さが の 評価が」のように機能語を言い直すケースに典型的に見られる[*4]（図 4（B）参照）。「相互関係を い 示します」のような短い言いよどみの場合には，句頭の上昇も句末の下降も観察されないことがあり，その場合は言いよどみ部の両側に韻律的不連続性があるとみなされる（図 4（C）参照）。なお語彙的・統

[*4] 韻律的に不連続性が観察されるのは，言い直された要素（上記例では「が」）ではなく，後続の言い直した要素（下線部の「の」）である。なお，「前足を から」のように，2 モーラ語以上の機能語で言い直した場合，言いよどみではなく通常のアクセント句として扱われる。

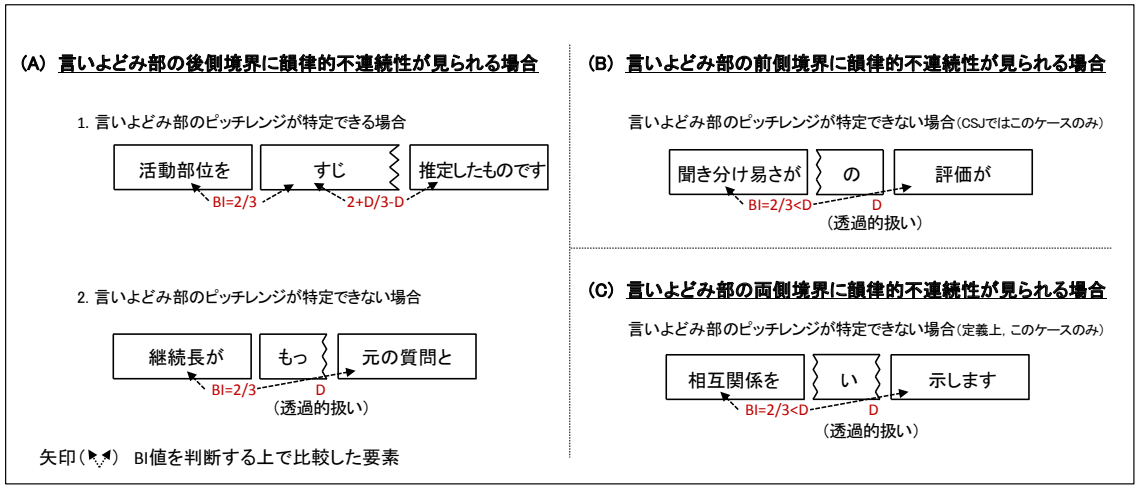


図 4　言いよどみのタイプと BI 値

合的な言いよどみであっても，韻律上の不連続性が全く認められない場合には，通常の語の場合と同様に扱われる。

■言いよどみに関わる BI 値の認定基準の概要　韻律的不連続性の観察される言いよどみについては，次の方針で前後の BI 値が認定される[*5]。

1. 言いよどみ部に先行するアクセント句の BI 値は，言いよどみ部のピッチレンジが特定できる場合[*6]，言いよどみ部と先行アクセント句との比較で BI=2 か 3 を判断するのに対し（図 4（A1）），言いよどみ部のピッチレンジが特定できない場合は，フィラーと同様に言いよどみ部を透過要素とみなし，言いよどみ部の前後のアクセント句との比較で BI=2 か 3 を判断する（図 4（A2）（B）（C））。
2. 言いよどみ部の前側境界に韻律的不連続性が認められる場合，先行アクセント句の BI 値に，始端の韻律的不連続性の後続を示す「<D」を付与する（図 4（B）（C））。
3. 言いよどみ部のピッチレンジが特定できない場合，透過要素とみなし，言いよどみ部と後続するアクセント句との韻律的接続関係は特定せず，言いよどみ部の終端を示す BI=D のみを付与する。一方，言いよどみ部のピッチレンジが特定できる場合，言いよどみ部と後続アクセント句とを比較し，イントネーション句境界が認められる場合は BI=3-D を，認められない場合は BI=2+D を付与する[*7]。

[*5] 第 2 刷までの基準では，図 4（B）に示す始端境界にのみ韻律的不連続性が見られる場合，「聞き分け易さが（BI=<D）の（BI=2）評価が（BI=3）」のように，語断片の先行要素に BI=<D を，言い直し部の終端に BI=2,3 を付与していた。この方針は，言いよどみ部を先行アクセント句の一部とみなす立場と言える。しかし実例を観察すると，言いよどみが複数繰り返されたりフィラーが連鎖したりと，先行要素との間に強い韻律的不連続性を感じることが多い。そこで第 3 刷では，図 4（B）に示すように，言いよどみ部の先行要素に BI=2,3 を，言いよどみ部の終端に BI=D を付与するよう，仕様を変更した。これにより，後述のアクセント句認定基準に従うと，言いよどみ部の前までが一つのアクセント句と認定され，言いよどみ部は独立した単位とみなされる。

[*6] 言いよどみ部のピッチレンジが特定できる場合とは，言いよどみ部に句頭上昇か句末下降のいずれかが観察され，かつ，句頭音調（トーン記号 H-）かアクセント核（トーン記号 A）のいずれかが見られる場合である。トーンの詳細は五十嵐ほか（2006）を参照のこと。

[*7] 第 3 刷までの基準では，言いよどみ部の終端には「D」を単独で付与し，言いよどみ部と後続するアクセント句と

■言いよどみに関わる韻律単位の認定方針

アクセント句：ピッチレンジが特定できる言いよどみ部は，不完全な単位ながらも通常のアクセント句と同様の手続きで BI=2 か 3 かの判断がなされるため，独立したアクセント句とする。一方，ピッチレンジの特定できない言いよどみ部は透過要素とみなすため，フィラーと同じ扱いとし，独立した特殊なアクセント句とする。結果，ピッチレンジの特定の有無に関わらず，言いよどみ部は独立したアクセント句と認定される。

イントネーション句：ピッチレンジが特定できない言いよどみ部はフィラーと同じ扱いとする。つまり，透過要素とみなしてイントネーション句を認定した上で，その境界に存在する言いよどみ部（連鎖する場合はその全体）は独立した特殊なイントネーション句として，イントネーション句の内部に存在する言いよどみ部は当該イントネーション句に含める形で認定する。一方，ピッチレンジが特定できる言いよどみ部は，通常のアクセント句と同様の基準に従い，BI=3,3-D の場合にイントネーション句境界があるとみなす。

以上の方針で認定したアクセント句・イントネーション句の認定例（模式図）を図 5 に示す。

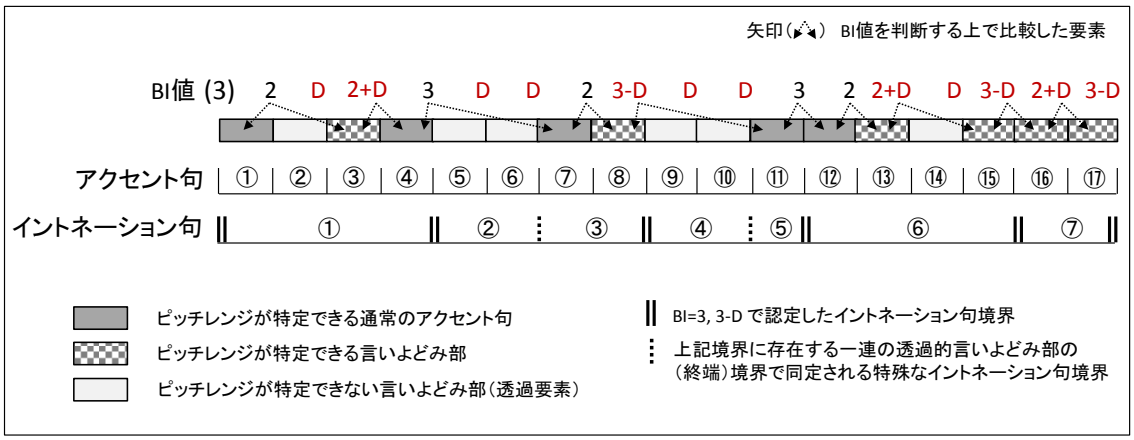


図 5　言いよどみが関わる場合のアクセント句・イントネーション句の認定例

2.3　アクセント句・イントネーション句の認定基準

以上，フィラーと言いよどみごとに，問題の所在と認定方針を示した。これらの検討をふまえ，一般化したした形で韻律単位の認定基準をまとめる。

アクセント句の認定基準は極めて単純に，次のようにまとめることができる。

アクセント句　BI 値が 2，3，F，F2，D，2+D，3-D のいずれかで区切られる単位

の韻律的接続関係は一切認定されていなかった。しかしこの基準では，ピッチレンジが特定できる言いよどみの場合（図 4（A1）），後続アクセント句との韻律的接続関係が一切認定されないことになる。これではイントネーション句の認定に問題が生じるため，（A1）の事例を対象に，後続アクセント句とのピッチレンジを比較し，BI=2 か 3 の判断を人手で行うこととした。なお，タグ BI=D+2,D-3 は暫定的なものであり，今後，韻律ラベリングの最新版を公開する際に変更する可能性がある。

イントネーション句の基準をまとめる前に，簡単に整理しておこう。ピッチレンジの特定できる言いよどみ（BI=2+D,3-D）は，通常のアクセント句（BI=2,3）と同じ扱いとなる。一方，フィラー（BI=F,F2）とピッチレンジの特定できない言いよどみ（BI=D）は，いずれも透過要素として扱われる。以上をふまえ，イントネーション句の認定基準を以下のようにまとめる。

イントネーション句

(1) BI=3,3-D の位置にイントネーション句の境界があるとみなす

(2) (1) で認定されたイントネーション句の境界に存在する透過要素（BI=F,F2,D の要素，連鎖する場合はその全体）は，前後とは独立したイントネーション句とする

補足 イントネーション句の内部に存在する透過要素は当該イントネーション句に含める

図 6 に，この基準に従い認定したイントネーション句の例を，フィラーと言いよどみが連続するケースを含めて示す。なおアクセント句は各セルに相当するため省略する。

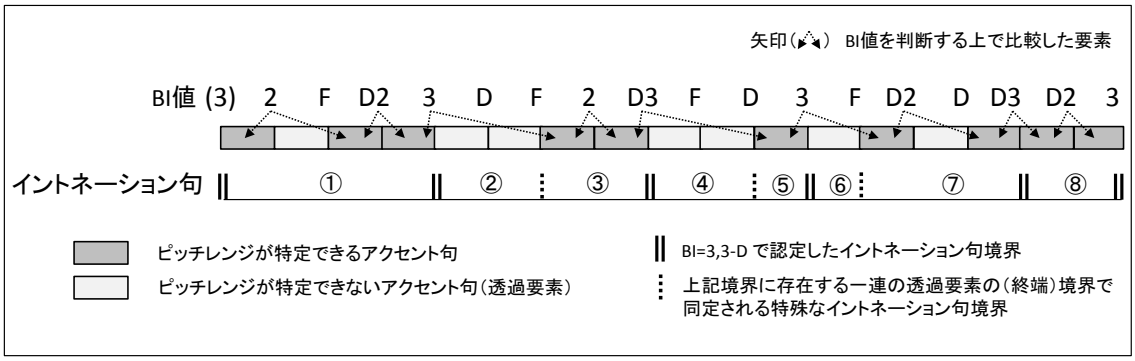


図 6 イントネーション句の認定例

3. おわりに

非流暢現象の音韻論的なトーン指定の仕組みや韻律構造に与える影響については，CSJ を対象とした前川（2012）などの研究でようやく検討が始まった状況である。そのため，純粋に理論的な観点から認定基準を定めることが難しいケースも少なからず存在する。しかしこのような状況だからこそ，アクセント句・イントネーション句を含む CSJ-RDB 版を公開することに意味があると言えよう。今後，非流暢現象の韻律特徴についての研究が進展するにつれて，認定基準もより堅固な基礎に立つことができるようになると期待される。

参 考 文 献

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

菊池英明・塚原渉（2006） 「XML 文書」『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. 455–526.

小磯花絵・西川賢哉・間淵洋子（2006） 「転記テキスト」『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. 23–132.

小磯花絵・伝康晴・前川喜久雄（2012） 「『日本語話し言葉コーパス』RDB の構築」『第 1 回コーパス日本語学ワークショップ予稿集』, pp. 393–400.

前川喜久雄（2004） 「『日本語話し言葉コーパス』の概要」『日本語科学』, 15, pp. 111–133.

前川喜久雄（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.

前川喜久雄（2012） 「自発音声中のフィラーの特性に関する予備的分析:位置と高さの分析」『第 26 回日本音声学会全国大会予稿集』, pp.115–120.

Maekawa, Kikuo, Hideaki Kikuchi, Yosuke Igarashi & Jennifer Venditti (2002) “X–JToBI: An extended J_ToBI for spontaneous speech“, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, pp.1545–1548.

Pierrehumbert, Janet & Mary Beckman（1988） *Japanese tone structure*, Cambridge: The MIT Press.

Venditti, Jennifer（1997） Japanese ToBI Labelling Guidelines, In K. Ainsworth-Darnell M. D’Imperio (eds.) Papers from the Linguistics Laboratory, *Ohio State University Working Papers in Linguistics* 50, pp.127–162. （First distributed in 1995 at a web document）

音声言語コーパスにおける speaking style の評定と分布
—転記テキストに着目して—

沈　睿（早稲田大学人間科学学術院）†
菊池　英明（早稲田大学人間科学学術院）‡

Rating and Distribution of Speaking Style in Speech Corpora
-Focusing on Speech Transcriptions-

Raymond SHEN (Faculty of Human Sciences, Waseda University)
Hideaki KIKUCHI (Faculty of Human Sciences, Waseda University)

１．はじめに

郡（郡 2006）によれば，speaking style（口調，発話様式）は個別言語の特徴記述に必要不可欠である．故に，一つの言語を習得する際，その言語の speaking style の習得も重要だと思われる．しかし，現在の外国語教育の分野では，文法や語彙などの内容が優先されており，speaking style に関する内容が十分に取り込まれていない．一方，コンピュータ技術の進歩に伴い，大規模コーパスや CALL システムなどの外国語教育への応用に関する研究が盛んに行われているため，大量の話し言葉を収録した音声言語コーパスを speaking style の習得にも活用できるのではないかと考えられる．その際，まず音声言語コーパスの speaking style の判別をできるようにしなければならない．

近年，speaking style は音声研究の分野で注目されてきたが，speaking style の定義に関しては，半世紀以来，明確な定義が提案されていない．当初，Uhlmann は speaking style を「特定利用のための口頭あるいは書かれた表現」として定義した（Uhlmann 1964）．この定義に基づけば，speaking style は話し言葉と書き言葉の両方で重要な指標と言える． Eskenazi はデータに基づいて speaking style を定義することを提案した(Eskenazi 1993)．本稿では主に Eskenazi の提案を用いるため，後述する．

音声研究において，speaking style は研究目的に応じて定義され，様々な分野で研究されている．音声の物理特性や音響特徴に焦点を絞った研究が多いが，発話内容や言語的特徴に注目する研究はまだ少ない．本研究は外国語教育への応用を想定しているため， 話し言葉の音響や韻律の側面から改善をしやすい発話内容や言語的特徴が記述される転記テキストに着目する．なお，従来の自然言語処理の分野で，書き言葉に対する文体・ジャンルの判別や著者推定などの研究は多く行われている．品詞率，語種率と形態素パタンを特徴量とした方法が有効であることが実証されたため，本研究もそれらの特徴量を speaking style の判別に用いることを想定し，話し言葉の転記テキストに着目して speaking style の定量化とモデルの構築を試みる．

２．手法

本稿では，Eskenazi が提案した speaking style を表現する 3 尺度を用いる．Eskenazi はデータに基づいて(data-driven)speaking style を定義することを提案した(Eskenazi 1993)．Eskenazi は，人間のコミュニケーションは，あるチャンネルを通じて，情報(message)が話し手から聞き手へ伝達することであり，speaking style を定義する際，この情報の伝達過程を考慮することが必要であると主張した．Eskenazi によれば，speaking style は「明瞭さ」(Intelligibility-oriented, 以降 I とする)，「親しさ」(Familiarity, 以降 F とする)，「社会階層」(Social strata, 以降 C とする)の 3 尺度で定義できる．Eskenazi によれば，「明瞭さ」は話し手の発話内容の明瞭さの度合いであり， 情報の読み取りやすさ・伝達内容の理解しやすさ

† raymondshenrui@gmail.com
‡ kikuchi@waseda.jp

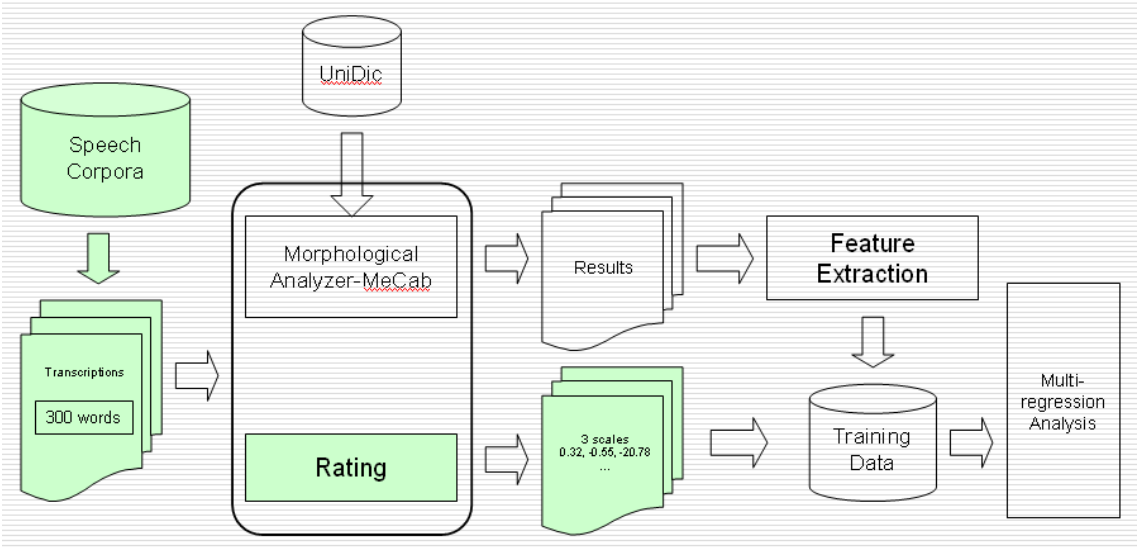


図 1 本研究の流れと本稿の位置づけ(緑の部分)
Figure 1 process in this study (the green part is mentioned in this paper)

や，読み取りの困難さ・伝達内容の理解の困難さを示す． 発話者が意図的に発話の明瞭さをコントロールしている場合も含む．「親しさ」は 話し手と聞き手との親しさにより変化する表現様式の度合いであり， 家族同士の親しい会話や，お互いの言語や文化を全く知らない外国人同士の親しくない会話などにあらわれる発話様式を示す．「社会階層」 は発話者の発話内容の教養の度合いであり，口語的な，砕けた，下流的な表現（社会階層が低い）や，洗練された，上流的な表現（社会階層が高い）を示す． 話し手と聞き手の背景や会話の文脈によって変化する場合もある．

本稿の流れについて図 1 を用いて説明する．なお，本稿は緑の部分を紹介する．まず，speaking style の異なる様々な音声言語コーパスから 6 コーパス（Speech Corpora）をバランスよく選出する．続いて 6 コーパス（カテゴリ）から 10 サンプルずつ音声の転記テキスト(Transcriptions)を選出し，speaking style の最も安定する部分と思われる最中部の約 300 字程度のテキスト(300 words)を抽出する．なお，本研究では，サンプルごとの speaking style の集積をサンプルが属するコーパスの speaking style とみなす．続いて抽出したテキストに対し，上述の speaking style の 3 尺度を用いて評定実験を行う(Rating)．本稿はここまで紹介するが，さらに，Mecab を用いて抽出したテキストに対して形態素解析を行い(Results)，品詞率，語種率，形態素パタンを特徴量として抽出する(Feature Extraction)．評定実験で得られた結果の平均を求め，3 尺度の学習データにする(3-scales)．最後に R の lm 関数を用い，線形重回帰分析のステップワイズ変数選択（変数増減法）(Multi-regression Analysis)で，3 尺度において，それぞれの判別モデルを求める．

３．評定実験

本章では，評定実験の詳細について述べる．

３.１ 評定者

本評定は，大学生男女 22 名の評定者による．

３.２ 刺激

本実験の刺激に音声言語コーパス内の転記テキストを使用する．

３.２.１ 音声言語コーパス

なるべく多様な speaking style を含む音声言語コーパスを使用するために，実験で使用する音声の転記テキストを以下の 6 種類の音声コーパス(カテゴリ)から選出した．

(1)日本語話し言葉コーパス(前川ら 2000)-講演（CSJ1 と呼ぶ）

日本語話し言葉コーパス(the Corpus of Spontaneous Japanese, CSJ)は，日本語の自発音声を大量に集めて多くの研究用情報を付加した，質・量ともに世界最高水準の話し言葉研究用のデータベースである．本研究では，CSJ に収録された speaking style の中でも，特に学会発表及び模擬講演発表（ひとつのテーマに関してのモノローグ）をまとめて扱う．

⑵日本語話し言葉コーパス-インタビュー（CSJ2 と呼ぶ）

⑴と同じく CSJ から選出した，インタビュー形式の対話である．講演音声と対話音声の speaking style は大いに違うと思われるので，今回の実験目的を考慮し，別のカテゴリとした．なお，インタビュアーとインタビュイーとの両方のチャネルの音声を使用した．

⑶千葉大地図課題対話コーパス(堀内ら 1999)（MAPTASK と呼ぶ）

地図を用いて課題を遂行するための対話コーパスである．

⑷新入生対話コーパス（FDC と呼ぶ）

大学の研究室に所属して 1 ヶ月の大学生同士の間の自由対話を収録したコーパスである．本コーパスは初対面の二者の対話音声が，時間経過および二者の親密性の向上とともにどのように変化するのかを調べることを目的としている．

⑸車載環境における質問応答の対話コーパス(宮澤ら 2010)（AUTO と呼ぶ）

本コーパスは，模擬車内環境でドライビングゲームをプレイしたドライバー役被験者と，同乗してナビゲーションを行ったナビゲーター役被験者に対して，走行実験終了後に，実験中の動画を見せながら感想やナビゲーションの的確さをインタビューした際の対話音声である．インタビューはドライバー，ナビゲーターそれぞれに対して実施した．なお，ドライバーとナビゲーターとの両方のチャネルの音声を使用した．

⑹旅行についての対話コーパス(岩野ら 1997)（TRAVEL と呼ぶ）

旅行の計画について，面識のある二人の研究室メンバーの間で交わされた自由対話を収録したコーパスである．

３.２.２ 転記テキストの加工

上述の 6 種類のカテゴリから 10 個ずつ合計 60 個の音声サンプルを無作為で選出する．Speaking style の最も安定する部分を抽出するため，上記の各音声に付随する転記テキスト中部より約 300 字のテキストを切り出す．なるべく発話の内容の影響を避け，発話様式や口調だけで評定してもらうため，テキストの名詞（代名詞は除く）の部分を全て「○○」に自動変換した（図 2 に参照）．

３.３ 評定方法

本評定は SD 法を用いる．一つのテキストを読んだ後，3 尺度のそれぞれについて 7 段階で評定してもらう．「明瞭さ」に関して，不明瞭の場合 1，明瞭の場合 7，「親しさ」に関して，親しい場合 1，親しくない場合 7，「社会階層」に関して，低い場合 1，高い場合 7 とする．評定はインターネット上のアンケートサイトを介して行う．評定の前に，尺度についての詳細説明をよく読むように指示した．

４．結果と考察

3 章で述べた評定実験によって得られた結果を本章で述べる．

まず，3 尺度の相関係数を求めた．明瞭さ I と親しさ F の相関係数が.26，明瞭さ I と社会階層 C が.48，親しさ F と社会階層 C が.56 である．今回の評定実験の結果によると，3 尺度が必ずしも独立ではないことが分かった．

さらに，22 名の評定者の評定結果の平均を各サンプルの得点として図 2(X 軸が明瞭さ I，Y 軸が親しさ F，Z 軸は社会階層 C)に示したように 3 尺度空間上にプロットした結果，刺激テキストがコーパスの収録条件や収録環境などの特徴により分かれる仮説とほぼ一致だと分かった．例えば，図 2 の赤い点が旅行についての対話コーパス（TRAVEL）のサンプルを示し，3 尺度空間上に不明瞭・親しい・口語的な位置に集まっている．TRAVEL コーパス内の音声が研究室メンバー同士の間の旅行計画についての自由対話なので，予想される分布

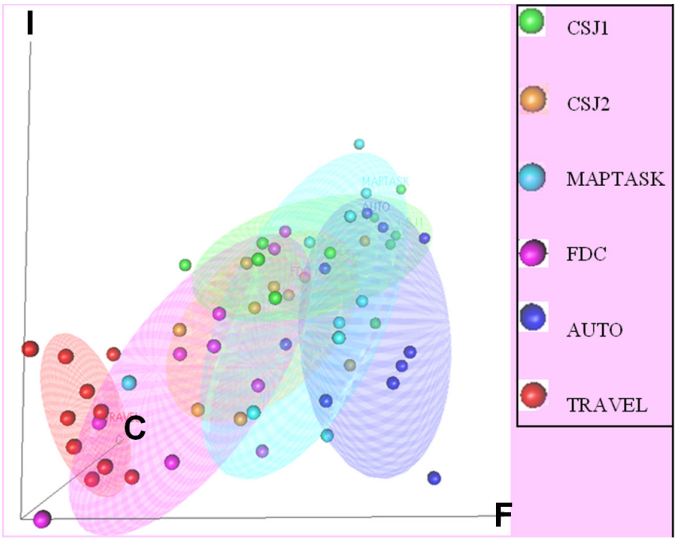


図 2 刺激テキストの 3 尺度空間上の分布
Figure 2 the distribution of text stimuli on the space of 3-scales of speaking style

と一致すると言える.

５．まとめ

本稿では，従来の書き言葉に対する文体やジャンル判別の手法を話し言葉における speaking style の定量化とモデル化に用いる前に， speaking style を「明瞭さ」(Intelligibility-oriented),「親しさ」(Familiarity),「社会階層」(Social strata)の 3 尺度を用いて音声言語コーパスからサンプリングした音声の転記テキストに対する評定実験を行い，音声コーパスの speaking style の分布を考察した．その結果，刺激テキストがコーパスの収録条件や収録環境などの特徴により分かれる仮説通り 3 尺度空間上で分かれることが分かった.

今後の方針として，従来の自然言語処理の手法を用い，speaking style の 3 尺度それぞれの判別モデルを構築し，学習者や教師に speaking style の自動判別サービスを提供しようと考えている.

文　献

Eskenazi, M. (1993)「Trends in Speaking style Research」 Keynote speech, Proceedings Eurospeech'93, Berlin.
Uhlmann, A.M. (1964) 「Meyers Neues Lexikon」 VEB Bibliograhisches Institut Leipzig, ausgabe in acht bänden edition.
岩野裕利, 杉田洋介, 松永美穂, 白井克彦(1997)「対面および非対面における対話の違い-頭の振りの役割分析」 音声言語情報処理研究報告, Vol. 15-29, pp.105-112.
郡史郎(2006)「日本語の『口調』にはどんな種類があるか」 音声研究, Vol. 10-3, pp.52-68.
堀内靖雄, 中野有紀子, 小磯花絵, 石崎雅人, 鈴木浩之, 岡田美智男, 仲真紀子, 土屋俊, 市川熹(1999)「日本語地図課題対話コーパスの設計と特徴」 人工知能学会誌, Vol.14-2, pp.261-272.
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, Vol.4-2, pp.51-61.
宮澤幸希, 影谷卓也, 沈睿, 菊池英明, 小川義人, 端千尋, 太田克己, 保泉秀明, 三田村健(2010)「自動車運転環境下におけるユーザーの受諾行動を促すシステム提案の検討」 人工知能学会誌, Vol.25-6, pp.723-732.

ポスター発表（2） Bグループ

3月1日（金） 14:00～15:00

ブラウザベースの動詞語義及び意味役割付与作業システム

上野 真幸（岡山大学大学院自然科学研究科）

竹内 孔一（岡山大学大学院自然科学研究科）

Construction of a Browser-based Annotation Tool for Verb Meanings and Semantic Role Labels

Masayuki Ueno(Graduate School of Natural Science and Technology, Okayama University)

Koichi Takeuchi(Graduate School of Natural Science and Technology, Okayama University)

1. はじめに

文中の動詞の語義を同定し，さらに，動詞と係り関係にある単語との意味的関係を人手で付与する述語項構造付与システムの構築を行った．

動詞語義概念とは動詞がとりえる１つの異なる意味について共通属性としてまとめたもので，例えば「走る」なら「電車が走る」「閃光が走る」などがそれぞれ異なる語義概念であり，１つの語義概念「電車が走る」の場合なら，【移動】であり，「移動する」「歩く」などを同じ動詞の類語として扱う単位である．また意味役割とは述語と係り元との関係を表したものであり，人により判断の揺れるものが存在するため信頼性の高いアノテーションを行うことは困難である．本稿で構築した動詞語義及び意味役割付与作業システムは BCCWJ の CORE の文章に対し，動詞語義及び意味役割を人の手で付与する際の補助を目的としたシステムである．動詞語義と意味役割については無料で公開されている動詞項構造シソーラス [1] のものを用いる．また複数人による付与を可能とするため，Web ブラウザベースで構築を行う．そのために CakePHP[2] を用いて構築を行った．

本システムの構造とインターフェースについて解説し，実際に動詞語義付与及び意味役割付与を行い，テキスト入力による付与と比べてどの程度作業効率が上がったか，そして各作業者間の付与の一致率を報告する．

2. 関連研究

関連研究として FrameNet[3] や Slate[4] が挙げられる．FrameNet はフレーム意味論に基づいて構築された英語の語彙辞書である．FrameNet はある語を想起させる意味概念を frame として仮定し，意味概念を持つ単語を結びつけることで構築されている．日本語においても日本語 FrameNet[5] の開発が進んでいる．また佐藤 [6] によって frameNet のデータを Web ブラウザ上で検索表示できるソフトウェアである FrameSQL が開発された．単語列ベースで詳細な付与が可能である一方で，システム自身は公開されていない．Slate は徳永ら [7] によって開発された汎用アノテーションツールであり，多様なアノテーションに対応できる．クライアントサーバ型のアプリケーションであり，複数の作業者でアノテーションを行うことができる．汎用性の高いツールであるため本研究で扱う動詞語義及び意味役割の付与もできるが，本研究で提案する動詞語義及び意味役割は種類が多いため，Slate で実行するには煩雑になり適していない．また，本システムの特徴である，複数人作業者の付与結果の記録と決定を行うことができない．

3. 動詞語義及び意味役割付与作業システム

本研究で扱う動詞語義及び意味役割付与作業システムとは，文章から動詞を選択し，選択した動詞の語義決定を行い，意味役割の付与を行うものである．システムの構成，利用している言語資源，及びシステムの特徴を述べる．また実際の作業の観点から必要と感じた機能についても説明する．

3.1. 人手による動詞語義及び意味役割付与の問題点

動詞語義及び意味役割の人手による付与を行う際の問題点を整理する．本システムで扱う意味役割は 95 種類，動詞語義は動詞ごとに複数存在する．動詞語義及び意味役割は種類が多く人手で記述を行う場合の処理は非常に煩雑なものとなる．既存のツールで本システムで扱う動詞語義及び意味役割の付与を人手で行った場合，膨大な時間を要し，間違いも発生しやすい．また動詞語義及び意味役割は人によって判断の揺れるものが存在し，信頼性の高いデータを作成するのは困難である．例えば「探索で見つかる」という文の動詞語義及び意味役割付与を考えると，「見つかる」は【発見】という語義であるが，一方「探索」は，「探索」という [原因] で【発見】したのか，「探索」という [手段] で【発見】したのか人によって判断の異なるところである．

3.2. 動詞語義及び意味役割付与作業システムの設計

前節で述べたように，動詞語義及び意味役割付与は煩雑で人によって判断が異なることがあるため，一人で作業を行うのは好ましくない．よって複数人で作業を行うために Web 上での閲覧・作業を実現する．動詞語義などは大量のデータ検索を行うのため，利用するデータベースは検索が高速である MySQL[8] を利用する．フレームワークは高速開発に適している CakePHP を利用する．本システムでは複数人による付与をひとつに集約するために，動詞語義及び意味役割が複数付与された場合，付与事例を考慮して，最終的に 1 人の作業者（作業リーダー）が判断して決定を行う．これにより，1 人で付与した結果よりも信頼性の高いデータが得られると考える．

3.3. データベース構築

本システムは BCCWJ コーパスの文章データ，LUW 単位でのデータ及び動詞項構造シソーラスを用いる．さらに LUW から動詞のみを抽出し，その頻度（出現回数）を格納したテーブル，付与された動詞語義及び意味役割のテーブルを作成した．

図 1 の左側は本システムで構築したデータベース図である．luw と sentence は BCCWJ コーパスの CORE の luw と文章データをそれぞれテーブルとしたものである．semantic と vth は動詞項構造シソーラスの意味役割と動詞語義をそれぞれ格納したものである．`luw_verb` は luw テーブル内に出現した動詞とその回数（freq）を格納している．`luw_verbid_sentenceid` は `luw_verb` テーブルの動詞がどの文で使われているかと作業状態か否か（sence）を示したテーブルである．`luw_semantic` 及び `luw_vth` は付与された意味役割及び動詞語義を格納するテーブルである．図 1 の右側は「詰め将棋の本を買ってきました」という文の「買う」に対し，語義【購入】と係り元である「詰め将棋の本を」に意味役割 [対象] を付与した例である．

3.4. 動詞語義および意味役割付与の手順

本システムの付与作業の手順を説明する．意味役割は動詞と項との意味的関係であり，動詞は複数の語義を持つため，語義により意味関係が異なり，語義を先に決定する必要がある．また各動詞語義の例文が複数必要であるが，作業者が付与対象となる文を見なければ動詞語義を決定するのは困難であるため，作業者が BCCWJ の文を見て作業対象となる文を決める必要がある．よって動詞語義付与を行う前に作業文の選択を行う必要がある．そして作業対象に選ばれた文に対して，動

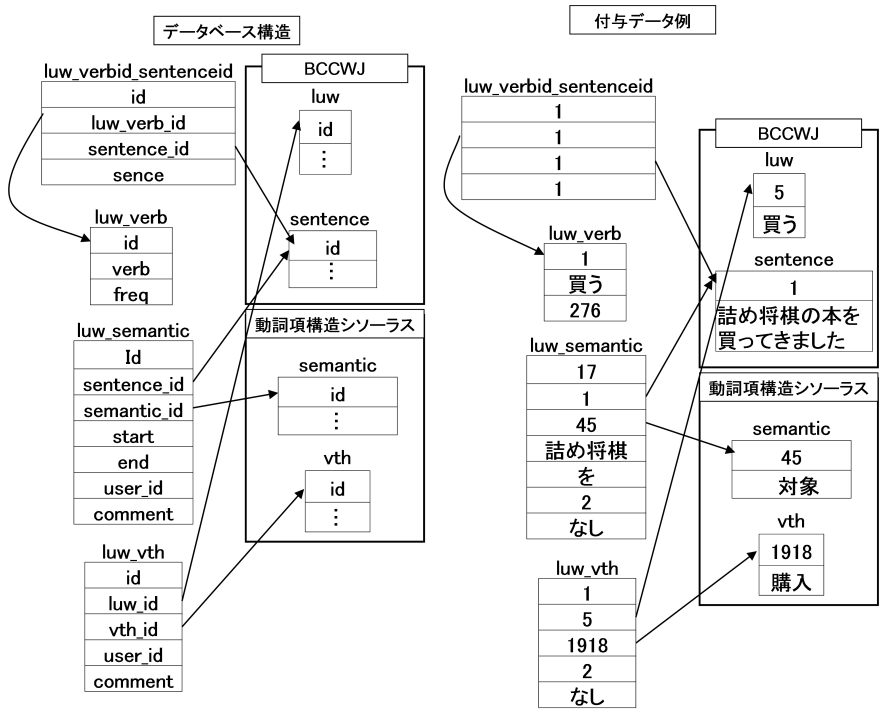


図 1　データベース構造と付与データ例

詞語義付与及び意味役割付与を行った後，付与された動詞語義及び意味役割から作業リーダーが決定を行う．以上の観点から，付与作業の手順は以下の 5 つに集約できる．

A) 作業文の選択.
B) 動詞語義付与.
C) 動詞語義の決定.
D) 意味役割付与.
E) 意味役割の決定.

この順に各作業者が作業を行う．それぞれに作業用 Web ページを作成した．基本作業画面は前回 [9] で述べたのでここでは割愛する．

3.5.　作業効率向上のための機能

実際に本システムを使って 5 ヶ月間作業者を雇い，付与作業を行った所以下のような意見が寄せられた.

1. 前回自分がどこの作業をしていたのか覚えていない.
2. 次の作業文にすすむときにページの移動が多い.
3. 当てはまる動詞語義，意味役割がない.

作業の負担を減らすためにこれらの意見を参考に改善を行った．具体的には以下のようにシステムを改善した.

1. 前回作業（語義付与，意味役割付与）を行った文へのリンクを作成.
2. 同じ作業を次の作業文で行う「次の文へ」のリンクを作成.

3. 動詞語義，意味役割選択の際に適合語義，意味役割なしの選択肢とコメント欄を追加.
4. 動詞語義追加の機能を作成.

これらの改善により作業負担が減るとともに，動詞項構造シソーラスの拡張にもつながると考えられる．作業者が適合する語義なしと判断した場合，語義を追加し適合語義を作成することにより，動詞項構造シソーラスを拡張できる．この作業では一般作業者が語義を判断し，それを確認した作業リーダーが新たな語義を追加するといった流れになると考えられる.

4. 作業結果

2012 年 9 月から 2013 年 1 月 25 日の間で，本システムを用いて 4 人の作業者を雇い付与を行った．作業者は 3 人の学部学生 (文系 2 名理系 1 名) と 1 人の大学院生であり，作業期間や意味役割分類の経験等はバラバラである．分類経験が無い作業者に対しては分類についての説明を行った．各作業者の動詞語義と意味役割付与数を表 1 に示す.

表 1　動詞語義及び意味役割付与数

	作業者 A	作業者 B	作業者 C	作業者 D
語義付与数	2324	6770	2220	6376
意味役割付与数	696	12346	3558	13766
意味役割付与文数	295	5672	1578	6755

各作業者間で付与数に差があるのは作業期間に差があるためである．2012 年 9 月から作業を開始した作業者 B と作業者 C においては，約 6000 文に対して語義と意味役割を付与している．各作業者間の一致率を表 2 に示す.

表 2　各作業者間の一致率

	語義一致率	意味役割一致率
作業者 A，B	92%（984/1064）	56%（177/312）
作業者 A，C	91%（932/1023）	68%（209/306）
作業者 A，D	82%（714/869）	65%（457/696）
作業者 B，C	81%（900/1101）	63%（920/1458）
作業者 B，D	84%（4647/5523）	65%（7781/11830）
作業者 C，D	82%（951/1151）	54%（1704/3129）

この一致率は両作業者が付与を行った文のみを対象としている．この結果では動詞語義は 80% 以上の一致率を記録している．動詞語義は一動詞につき約 3〜4 つの語義から選択する形式であり，作業者は他作業者の付与結果を見ることができないため，高い一致率であるといえる．意味役割は 95 種類の意味役割から選択するため，全体的に語義に比べ低い結果となっている．現在もこの作業は継続中であり，意味役割の決定作業が進んでいないため，作業リーダーとの一致率は不明である．これらの一致率は本システムで算出することができる．そのため作業が進んでいるリアルタイムでの一致率を常に見ることができる．一致率閲覧画面の例を図 2 に示す.

図 2 は作業者 A，D の一致率表示画面である．両作業者が作業した文の数や一致した数を表示している．語義平均選択肢数は両作業者が付与した語義の選択肢数の平均値であり，語義付与の曖昧性を示している．位置が一致した意味役割数は選択された意味役割の種類は違うが，選んだ係り

語義	2324
語義が無いと思ったもの	131(0.056368330464716)
二人が語義を付与した動詞数	869
二人が一致した語義数	714/869(0.82163406214039)
語義選択肢平均数	2.9747899159664(2124/714)
付与された意味役割数	696
二人が意味役割を付与した動詞数	295
位置が一致した意味役割数	573/696(0.82327586206897)
一致した意味役割数	457/696(0.6566091954023)

図 2　付与数と一致率の表示例

先が同じものの数を示している．この画面は計算量が多いため他の動作に比べると表示が遅い．

5.　考察

付与作業の効率について考察を行う．今回の付与作業での付与速度は，9 月時点での作業者 2 名の付与データを 100 件ほど人手で見たところ，1 語義付与約 15～30 秒程度で行われている．本システムを用いずに語義をテキストで入力する場合慣れた作業者でも 1～2 分程度要するためかなりの効率が上がったといえる．効率が上がった主な要因として以下の点が挙げられる．

(1) 語義や意味役割を選択するだけで文字入力する必要が無い．
(2) 次の作業に移る際の手間が少ない．

(1) はシステムの補助をなしにこの付与作業を行う場合，例えば「買う」に対する語義であれば，【状態変化あり-位置変化-位置変化（物理）（人物間）-他者からの所有物の移動-購入-彼女が高級品を専門店から買う】等を入力しなければならないため非常に手間がかかる．これらの入力を大きく減らしたことが作業の効率が上がった大きな要因であると考えられる．(2) は作業の量が非常に多いため，各作業間の画面クリックが例え一つでも多い場合，作業の際ストレスがかかるとの意見が作業者から寄せられた．この点を改善し，なるべく作業間の動作を減らした結果，作業の効率上昇につながったと考えられる．

6.　まとめ

本研究では文章の意味を同定するために，動詞語義および意味役割の付与を補助するシステム構築を行った．データベースとフレームワークを用いて，Web 上で作業，閲覧を実現した．この手法により複数人による付与作業が可能となり，インターネットブラウザさえあればシステムのダウンロードを必要とせず，複数人で 1 つの文に対する付与を実現することで，データの信頼性向上を図った．また 5 ヶ月間 4 名の作業者に作業をを行ってもらい，語義では 80%，意味役割では 60% 程度の一致を得た．

参考文献

[1] 動詞項構造シソーラス. http://cl.cs.okayama-u.ac.jp/rsc/data/index.html.

[2] CakePHP. http://cakephp.jp/.

[3] FrameNet. http://framenet.icsi.berkeley.edu/.

[4] Slate. http://www.cl.cs.titech.ac.jp/slate/.

[5] Japanese FrameNet. http://jfn.st.hc.keio.ac.jp/ja/index.html.

[6] 佐藤弘明. FrameSQL で利用する日本語フレームネット. 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. 143–146, 2010.

[7] Dain Kaplan, 飯田龍, 徳永健伸. 汎用アノテーションツール Slate. 言語処理学会 第 17 回年次大会 発表論文集, pp. 619–622, 2011.

[8] MySQL. http://www-jp.mysql.com/.

[9] 上野真幸, 竹内孔一. 動詞語義及び意味役割付与作業システムの構築. 第 2 回日本語コーパスワークショップ, NL-207-15, pp. 69–76, 2012.

個人用コーパスの作成とアノテーションを支援する環境の実現

山口昌也 (国立国語研究所言語資源研究系) †

Implimentation of an Environment for Personal Corpus Construction and Annotation

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

1 はじめに

本稿では，個人の研究者がコーパスを構築し，アノテーションするのを支援するための環境について述べる。本環境は，全文検索システム『ひまわり』(山口・田中 2005) を機能拡張することにより実現する。

現在，Web データをはじめとして，大規模な電子的テキストを容易に入手できるため，個人の研究者でもコーパスを構築できるようになっている。また，コーパスを構築したり，活用したりするためのシステムとしても，Web ベースのコーパス検索ツール『中納言』(小木曽ら 2011)，コーパス管理ツール『茶器』(松本ら 2006)，テキストマイニングツール KHCorder(樋口 2003) など，様々なツールが利用できる。ただし，収集したデータを用いて，研究者自らがコーパスを構築し，言語研究用に利用することを考えた場合，いくつかの技術上の障害がある。本稿では，コーパス構築時のデータの統一性と，コーパス構築後のアノテーションに焦点を絞って，研究者を支援する方法を考える。

2 支援の方法

2.1 全体的な支援の流れ

図 1 に全体的な支援の流れを示す。提案する支援方法は，図の左側のコーパス構築時の支援（図 1 左）と，構築後の追加的なアノテーション支援（図 1 右）に分けられる。

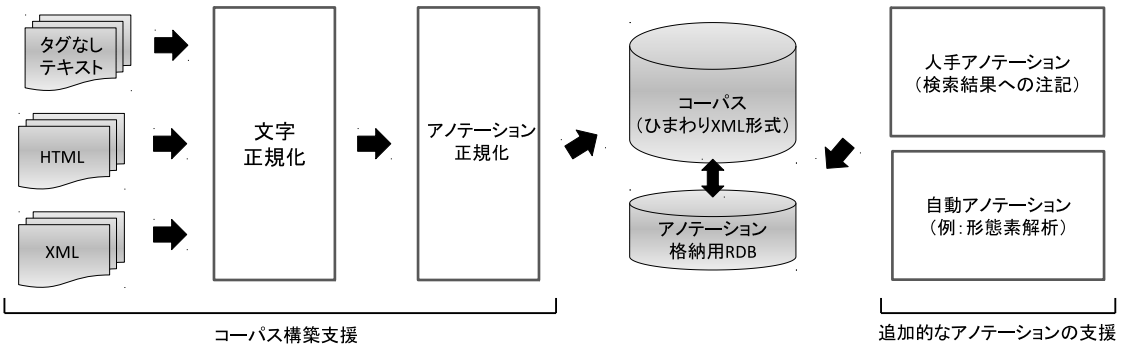


図 1: 全体的な支援の流れ

コーパスの構築時の支援としては，複数の電子テキストを統合する際の (1) 文字の正規化，(2) アノテーションの正規化，の 2 点を行う。原資料とするテキストの形式は，タグなしテキスト，HTML，XML とする。コーパス構築支援の結果，生成されるコーパスは，統一された XML 形式（ひまわり XML 形式）のファイルとなる。

また，コーパス構築後の支援として，追加的なアノテーションの支援を行う。支援の方法は，2 種類ある。一つは，形態素解析システムなどの自然言語処理システムの解析結果のアノテーション支援である。もう一つは，検索結果に注記などを加える，人手のアノテーションの支援である。

†http://www2.ninjal.ac.jp/masaya

2.2 コーパス構築支援

2.2.1 文字の正規化

いわゆる半角・全角数字のように，同一の文字が電子的に別の文字として扱われると，検索漏れの原因となる。ここでは，文字の正規化として，符号化方式の正規化と文字コードポイントの正規化を行う。

まず，符号化方式の正規化についてである。現在，電子化テキストの符号化には，Shift JIS，UTF-8，EUC など，多くの符号化方式が用いられる。これらをひまわり XML 形式データで用いている UTF-16 に統一する。なお，この際，原資料の符号化方式は，自動推定している。

符号化方式の正規化に加えて，文字のコードポイントの正規化を行う。具体的には，複数のコードポイントに割り当てられている文字や，合成を含む文字を単一のコードポイントに統合する。前者の例として，いわゆる半角，全角の英数字を挙げる。後者の例としては，テ＋゛⇒デのような，半角の仮名文字と濁点の合成がある。

コードポイントの正規化の方法としては，二つのオプションを用意している。一つは，変換テーブルを用いるもので，拡張版の『ひまわり』には，半角英数文字（JIS X 0201 のラテン文字用図形文字）から全角文字への変換テーブルが付属している。もう一つは，Unicode の規格として規定されている正規化形式 (NFKC, Normalization Form Compatibility Composition) に基づく正規化 (Davis and Whistler 2012) である。

2.2.2 アノテーション形式の正規化

前述のとおり，原資料のファイル形式は，タグなしテキスト，HTML，XML の 3 種類である。『ひまわり』は，個々の形式ごとに変換規則を定め，それぞれの形式からひまわり XML 形式に変換する。ファイル形式ごとに変換方法は，次のとおりである。なお，アノテーション形式の正規化については，従来から公開している，ひまわり XML 形式データの作成支援ツール『えだまめ』の機能に拡張を加えたものとなっている。

タグなしテキスト　タグなしテキストは，一般的な統一規格に基づいたタグでアノテーションされていないが，独自形式のタグでアノテーションされている場合がある。例えば，青空文庫の「テキストファイル」形式では，ルビが「五月雨《さみだれ》」のような形式で記述されている。

このような独自形式のタグを，利用者が定義した文字列置換規則により，XML タグへ変換する。文字列置換規則は，指定された正規表現にマッチした文字列を指定された文字列に置き換える。

HTML　HTML でアノテーションされたテキストは，XHTML へ変換したのちに，指定された XSLT スタイルシートにより，ひまわり XML 形式に変換する。標準で添付しているスタイルシートは，青空文庫の XHTML ファイルから書誌情報などを取得できるようになっている。

XML　XML でアノテーションされたテキストは，指定された XSLT スタイルシートにより，ひまわり XML 形式に変換する。

2.3 アノテーション支援

2.3.1 概要

ここで言うアノテーション支援とは，コーパス構築後に行う，追加的なアノテーションの支援である。そのため，コーパス自体には変更を加えず，stand off でアノテーションするよう設計した。アノテーション結果は，コーパス中の位置情報とともにリレーショナルデータベースに格納される。なお，リレーショナルデータベースエンジン自体も Java 言語で記述されており[1]，拡張された『ひまわり』に標準で同梱している。したがって，構築したコーパスとアノテーション結果を再配布することも容易である。

[1]オープンソースのリレーショナルデータベースエンジン H2 を利用している。

想定するアノテーションとして，(1) コーパスを形態素解析システムや構文解析システムで解析した結果を，コーパスに追加的に自動アノテーションする，(2) 検索結果に注記を人手アノテーションする，ことを考慮した。以下の節では，それぞれの支援方法について，詳しく説明する。

2.3.2 自動アノテーション

タグづけされたテキストに対して，形態素解析や構文解析を行うには，タグを取り除くなどの前処理を行う必要がある。また，解析後も，元々タグづけされていたアノテーションと解析した結果とを統一して検索できるようにする必要がある。

そこで，自動アノテーションの支援では，このような処理を自動的に行うようにする。現時点では，形態素解析システム Mecab と JUMAN の解析結果を扱えるようになっている。データベースに登録する情報は，コーパス中の当該文字列をマークアップし，形態素解析結果に含まれる品詞，活用などの情報を属性として付与するのと同等の情報である。『ひまわり』は，それらの属性をコーパス中の XML のタグの属性と同様に検索することができる。

2.3.3 人手アノテーション

コーパスを利用する際，検索の結果に注記をつけておきたい場合がある。例えば，用例を分類するような場合である。図 2 は,「掘り出す」の用例を語義分けしている例である。

利用者は,「掘り出す」の用例を検索し，それぞれの用例の「メモ 1 」「メモ 2 」欄に注記を加える。「保存」ボタンを押すタイミングで，注記がリレーショナルデータベースに記録される。

注記の形式は 2 種類あり,「メモ 1 」欄が自由記述,「メモ 2 」欄が選択記述となっている。注記欄は追加することも可能である。また，選択記述の項目は，利用者が定義できる。自由記述欄では，選択範囲に値を一括指定するなど，効率的な入力ができるようになっている。これにより，検索した内容をまとめてデータベースへ記録しておく，などの応用が可能である。

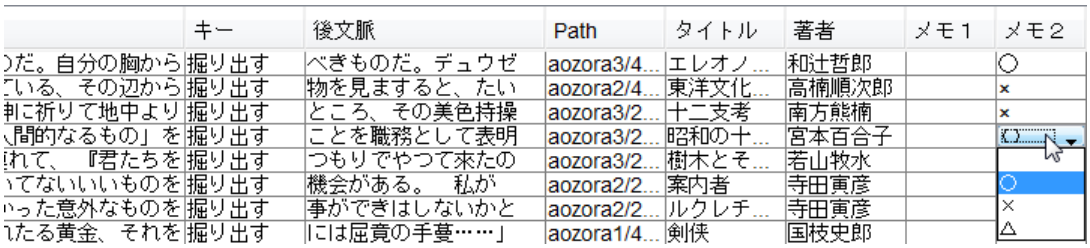


図 2: 人手アノテーションの例

3 実行結果例

以上の支援方法を実装し，拡張版の『ひまわり』として一般に公開している[2]。また，支援機能を利用した例として，大量の「青空文庫」作品を統合し，ひまわり XML 形式に変換した結果 (「青空文庫パッケージ」) も公開している。ここでは，提案した支援手法の実行例として,「青空文庫パッケージ」の構築方法と,「青空文庫パッケージ」への形態素解析結果のアノテーションについて紹介する。

3.1 「青空文庫パッケージ」の構築

「青空文庫パッケージ」は，コーパス構築支援機能を用いて,「青空文庫」で公開されている 10667 作品をひまわり XML 形式のデータに変換したものである。手順は，次のとおりである。

[2] α 版（ver.1.5a02）という位置づけで，2012-10-11 に公開した。なお，このバージョンでは，文字コードポイントの正規化部分は，まだ含まれていない。

(1) 収録対象の作品を選択する。今回は，青空文庫のサイトで公開されている「作家別作品一覧拡充版」から，次の条件に合致する作品を選択した。なお，底本が複数ある作品は,「文字遣い種別」が新字，新仮名の作品を優先している。

- 著作権が切れていること
- XHTML 版が存在すること
- 『ひまわり』用にインポートに成功すること

(2) 「作家別作品一覧拡充版」には，個々の作品の URL が記載されている。その情報をもとに，Web ページのダウンロードツール wget で XHTML 版のファイルを一括ダウンロードした。

(3) ダウンロードしたファイルをコーパス構築支援機能を用いて，ひまわり XML 形式に変換した。ただし，全作品を一括して変換すると，メモリ不足の問題が発生するため，三つに分割し，検索時にそれらを一括検索するようにしている。

3.2 「青空文庫パッケージ」へのアノテーション

「青空文庫パッケージ」に対して，形態素解析結果のアノテーションを行った。使用した形態素解析システムは，MeCab(ver.0.994, IPADIC) である。アノテーションに要した時間は，Ubuntu 12.04 上 (CPU: Intel Xeon E5520 2.27GHz, Memory: 8GB) で約 15 時間かかった。解析結果の形態素数は，85325922 であった。体系的な検索速度の計測は行なっていないが，基本形「投げる」を含む用例を検索したところ，約 23 秒で 4130 例を得ることができた。

4 おわりに

本稿では，個人の研究者がコーパスを構築し，アノテーションするのを支援することを目的として，文字・アノテーションの正規化，自動・手動アノテーションに関する支援方法を提案した。また，全文検索システム『ひまわり』を機能拡張することにより，提案手法を実現した。

参考文献

山口昌也, 田中牧郎 (2005) 「構造化された言語資料に対する全文検索システムの設計と実現」，自然言語処理 vol.12, No.4, pp.55–77

小木曽智信，中村壮範，鈴木泰山，八木豊，山崎誠，前川喜久雄 (2011) 「コーパス検索システム「中納言」デモンストレーション」，日本語コーパス完成記念講演会予稿集，pp.43–46

松本裕治，浅原正幸，橋本喜代太，投野由紀夫，大谷朗，森田敏生 (2006) 「タグ付きコーパス管理/検索ツール『茶器』」，言語処理学会第 12 回年次大会論文集，pp.460–463

樋口耕一 (2003)「コンピュータ・コーディングの実践 —漱石『こころ』を用いたチュートリアル—」，年報人間科学 24，pp.193–214

Mark Davis, Ken Whistler Eds. (2012) *Unicode Normalization Forms, Unicode Technical Reports UAX 15*, http://unicode.org/reports/tr15/

関連 URL

『ひまわり』『えだまめ』 http://www2.ninjal.ac.jp/lrc

『中納言』 https://chunagon.ninjal.ac.jp/

『茶器』http://sourceforge.jp/projects/chaki/

KHCorder http://khc.sourceforge.net/

H2 http://www.h2database.com/html/main.html

Mecab http://mecab.googlecode.com/svn/trunk/mecab/doc/

JUMAN http://nlp.ist.i.kyoto-u.ac.jp/index.phpJUMAN

wget http://www.gnu.org/software/wget/

『現代日本語書き言葉均衡コーパス』に対する時間表現・事象表現間の時間的順序関係アノテーション

保田 祥 (国立国語研究所コーパス開発センター)
小西 光 (国立国語研究所コーパス開発センター)
浅原 正幸 (国立国語研究所コーパス開発センター)
今田 水穂 (国立国語研究所コーパス開発センター)
前川 喜久雄 (国立国語研究所言語資源研究系/コーパス開発センター)

Temporal Ordering Annotation on the Balanced Corpus of Contemporary Written Japanese

Sachi Yasuda (Center for Corpus Development, NINJAL)
Hikari Konishi (Center for Corpus Development, NINJAL)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Mizuho Imada (Center for Corpus Development, NINJAL)
Kikuo Maekawa (Dept. Corpus Studies/Center for Corpus Development, NINJAL)

1. はじめに

情報抽出や文書要約の分野において、情報の可視化を目的として、テキスト中に出現する事象表現を時間軸上に写像することが行われている。事象表現を時間軸上に写像するためには、テキスト中に出現する時間表現の正規化（時間軸への写像）のみならず、対象となる「文書作成日時と事象表現」や「時間表現と事象表現」、「二つの事象表現」間の時間的順序関係を付与することが必要になる。

テキスト中の時間情報表現を分析する研究は日本語以外の言語で進んでおり、時間表現の文字列の切り出しや正規化のみならず、時間表現と事象表現の関連付けなどが行われている。表 1 に英語もしくは日本語を対象とした時間情報表現に関連する研究を示す。以下、まず英語の時間情報表現に関する代表的な研究を俯瞰する。

英語においては、評価型国際会議 MUC-6 (Grishman and Sundheim (1996)) の一タスク固有表現抽出の中に時間情報表現の抽出が含まれていた。MUC-6 で定義されている時間情報表現タグ ⟨TIMEX⟩ は日付表現 (@type="DATE") と 時刻表現 (@type="TIME") からなる。アノテーション対象は絶対的な日付・時刻を表す表現にのみ限定され、”last year” などといった相対的な日付・時刻表現は含まれていない。この MUC-6 のアノテーション基準 ⟨TIMEX⟩ に対し、Setzer は時間情報表現の正規化に関するアノテーション基準を提案している (Setzer (2001))。評価型国際会議 TERN (DARPA TIDES (2004)) では、時間情報表現検出に特化したタスクを設定している。TERN で定義された時間表現情報タグ ⟨TIMEX2⟩ は、相対的な日付・時刻表現、時間表現や頻度集合表現が検出対象として追加されている。時間表現の正規化情報を記述する

表 1　関連研究

英語の時間情報表現に関する関連研究		
MUC-6 (Grishman and Sundheim (1996))	評価型会議	時間情報表現の切り出しのみ
Setzer (2001)	基準	時間情報表現の切り出しと正規化
TERN ⟨TIMEX2⟩ タグ (DARPA TIDES (2004))	評価型会議	時間情報表現の切り出しと正規化
TimeML ⟨TIMEX3⟩ タグ (Pustejovsky et al. (2003))	基準	時間情報表現の切り出しと正規化
TimeML ⟨TLINK⟩ タグ	基準	事象間の時間的順序関係
TimeBank (Pustejovsky et al. (2003))	コーパス	TimeML 基準によるタグつきコーパス
Aquaint TimeML Corpus	コーパス	TimeML 基準によるタグつきコーパス
Boguraev and Ando (2005)	解析手法	時間情報表現-事象間の時間的順序関係解析
Mani (2006)	解析手法	二事象間の時間的順序関係解析
TempEval (Verhagen et al. (2007))	評価型会議	時間情報表現-事象間/二事象間の時間的順序関係解析
TempEval-2 (Verhagen et al. (2010))	評価型会議	時間情報表現-事象間/二事象間の時間的順序関係解析
TempEval-3 (2013 年開催)	評価型会議	時間情報表現-事象間/二事象間の時間的順序関係解析
日本語の時間情報表現に関する関連研究		
IREX (実行委員会 (1999))	評価型会議	時間情報表現の切り出しのみ
拡張固有表現体系 (Sekine et al. (2002))	基準	時間情報表現の切り出しのみ
橋本・中村 (2010)	コーパス	時間情報表現の切り出しのみ
小西ほか (2012)	基準/コーパス	時間情報表現の切り出しと正規化
本研究	基準/コーパス	時間情報表現-事象間/二事象間の時間的順序関係

ISO-8601 形式を拡張した @value 属性などが設計され、こちらも自動解析対象となっている。その後、Pustejovsky らによりアノテーション基準 TimeML (Pustejovsky et al. (2003)) が提案されている。その中では、TERN で用いられている ⟨TIMEX2⟩ を拡張した ⟨TIMEX3⟩ が提案され、さらに時間情報表現と事象表現の時間的順序関係を関連づけるための情報 ⟨TLINK⟩ が付加される。これらの情報は人手でアノテーションすることを目的に設計され、TimeBank (Pustejovsky et al. (2003)) や Aquaint TimeML Corpus などの人手によるタグつきコーパスの整備が行われた。これらのコーパスに基づく時間情報表現の自動解析 (Boguraev and Ando (2005); Mani (2006)) が試みられたが、タグの情報に不整合があったり、付与されている時間的順序関係ラベルに偏りがあったりなど扱いにくいものであった (Boguraev and Ando (2006))。2007 年に開かれた SemEval 2007 の 一タスク TempEval (Verhagen et al. (2007)) では、時間的順序関係のラベルを簡略化し、人手で見直したデータによる時間的順序関係同定のタスクが行われた。このタスクでは、時間表現に対して正規化された @value 属性などが付与されており、事象表現の時間的順序関係同定に利用してよい。TempEval-2 (Verhagen et al. (2010)) では、英語だけでなく、イタリア語、スペイン語、中国語、韓国語に関しても同様なデータを利用したタスクが設定された。2013 年に開かれる SemEval-2013 のサブタスク TempEval-3 では、データの規模を大きくした英語、スペイン語が対象となっている。

次に日本語の時間情報表現に関する研究を示す。日本語においては、IREX (実行委員会 (1999)) の 一タスクとして、固有表現抽出タスクが設定された。IREX の時間情報では、日付・時刻表現を対象にし、相対的な表現が定義に含まれている。関根らは拡張固有表現体系 (Sekine et al. (2002)) を提案し、辞書/オントロジやコーパスの作成などを行っており、『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese; 以下 “BCCWJ”) にも同じ体系の拡張固有表現タグが付与されている (橋本・中村 (2010))。小西らは (小西ほか (2012); 小西ほか (2013)) TimeML に基づく ⟨TIMEX3⟩ 相当のタグを BCCWJ の一部に付与し、時間情報表現の正規化を行った。

本研究では、BCCWJ の一部に対し時間情報表現と事象表現の時間的順序関係を付与するた

めに、事象表現の切り出しと分類を行った。さらに、Allen の時区間論理に基づきテキストに出現する時間表現と事象表現の時間的順序関係を付与した。以下ではアノテーション基準を示し、得られたデータの傾向について報告する。

2. アノテーション基準

2.1 アノテーション作業の概要

アノテーション作業対象は新聞サブコーパス 54 ファイル（部分集合 A）とする。先行研究における時間情報表現の正規化作業により、時間情報表現は ⟨TIMEX3⟩ タグにより切り出され、時間関連の情報が与えられている。時間情報表現の正規化作業については詳細については文献 (小西ほか (2012); 小西ほか (2013)) を参照されたい。

アノテーション作業は、まず、事象表現の境界を認定し ⟨EVENT⟩ タグを付与し、⟨EVENT⟩ の属性として、事象表現の分類を表す @class 属性を付与する。次に限定された「文書作成日時と事象表現」や「時間表現と事象表現」、「二つの事象表現」間の組み合わせに対して、時間的順序関係を付与する。以下では、それぞれの作業の基準について示す。

2.2 事象表現の認定とクラス分類

時間的順序関係のアノテーションを行うために、事象表現か否か、また事象表現が時間軸上の具体的な特定の範囲で生起したものか否かの判断が必要となる。また事象構造が動作なのか状態なのかといった識別が必要になる。このため国語研長単位の動詞、形容詞を中心に事象表現か否かの判別を行い、事象表現とみなしたものについては ⟨EVENT⟩ タグで切り出す作業を行う。また事象表現として切り出す際に国語研長単位が適さない場合には長くする方向で修正を行う。切り出された ⟨EVENT⟩ タグに対して @class 属性にその事象表現の特性を付与する。このアノテーションスキーマは TimeML に準じている。

⟨EVENT⟩ タグを付与するか否かの判断基準として、文書作成日時もしくは他の事象表現との時間的順序関係が定義できるかどうかを重要視する。何らかの変化を含む事象表現ではなく、恒常的あるいは一般的なことをいっていると考えられ得る事象表現においては、時間的順序関係のアノテーションは不可能であるため、⟨EVENT⟩ タグは付与しない。

⟨EVENT⟩ タグを付与しない例 (**太字**が注目している表現)

クラブの運営について 1 票を**持っている**わけではない。
国際会議 57 件を**含め** 2,111 件。火を**使わない**調理法。

連体修飾節中の動詞が、一般的（Generic）と判断される場合 ⟨EVENT⟩ タグを付与しない。

⟨EVENT⟩ タグを付与しない例：連体修飾 (**太字**が注目している表現)

旅の安全を**守る**道祖神。オリーブ畑に**囲まれた**レストラン。

副詞的用法や慣用的な場合も時間的順序関係がつけがたいため、⟨EVENT⟩ タグを付与しない。

〈EVENT〉タグを付与しない例：副詞的表現や慣用表現 (**太字**が注目している表現)

やむを**得ない**。 **相次いで**出している。
なりふり**構わぬ**販売攻勢。

文脈によっては、「ある」「なる」「する」などの動詞も、一般的なことを述べているため時間的順序関係がつけがたい場合がある。この場合、〈EVENT〉タグ付与しない。

〈EVENT〉タグを付与しない例：「ある」「なる」「する」(**太字**が注目している表現)

～のためこの名が**ある**。～が基本と**なる**。
～を原則と**する**。

時間的順序関係が確認できる事象構造には〈EVENT〉タグを付与したうえで、@class 属性を付与する。@class 属性は、OCCURRENCE、REPORTING、PERCEPTION、ASPECTUAL、I_ACTION、I_STATE、STATE の 7 種に分類される。

- OCCURRENCE 項に事象を取らない事象表現一般
- REPORTING 項に事象を取る表現活動動詞に相当するもの
- PERCEPTION 項に事象を取る認識・知覚動詞に相当するもの
- ASPECTUAL 項に事象を取るアスペクトを表出するもの
- I_ACTION 項に事象を取る遂行動詞に相当するもの
- I_STATE 項に事象を取る思考・感情動詞に相当するもの
- STATE 静態動詞のうち時間表現に直接関連するもの

一般の事象表現は OCCURRENCE にあたる。静態動詞は STATE に分類されるため、STATE にしないもので、物 (Thing) を項とする事象表現はすべて OCCURRENCE とする。残りの 5 種類は、事物ではなく、事象を導入する場合にのみ用いる。

以下にそれぞれの例を挙げる。

■OCCURRENCE：事象表現一般　何かが起こった、変化した、発生したなどの一般的な事象構造は、OCCURRENCE とする。すなわち、事象ではなく物 (Thing) を項とし、静態動詞ではない場合は、すべて OCCURRENCE とする。無意志的（状態・位置）変化動詞や非意志的（現象一般）動詞もこれに含まれる。また、過程 (PROCESS) を示す動詞（例：「住む」）も、OCCURRENCE とみなすこととする。

〈EVENT〉@OCCURRENCE の例

湿地や干潟、河原などが埋め立てで〈EVENT〉減った〈/EVENT〉東京湾。
裸地を好むコアジサシに〈EVENT〉嫌われた〈/EVENT〉か、巣は一つだけ。
ニュース写真として〈EVENT〉掲載させていただく〈/EVENT〉ことがあります。
経常利益は数億円単位の黒字に〈EVENT〉なる〈/EVENT〉。
メニューに〈EVENT〉挑戦した〈/EVENT〉。

■REPORTING：表現活動動詞　表現活動動詞が、事象に関する発言や告知などをはじめ、概ね「～と」を用いた引用を行う場合などで、REPORTING に分類する。なお、「～を」が用いられている場合は項が物事（Thing）であるため、OCCURRENCE となる。表現活動動詞には、言う・報

告する・告げる・説明する・陳述する・指摘する・伝えるなどが含まれる。新聞サブコーパスでは、文末の「という」が概ねこれにあたるが、名詞化された事象（例：「報道がいうには～」）に用いられるなどで、時間的順序関係と結びつかない場合はそもそもタグをつけない。

〈EVENT〉@REPORTING の例（**太字**が注目している項）

大学院でのこうした取り組みは**初めてと** 〈EVENT〉 いう 〈/EVENT〉。
～どうかと 〈EVENT〉 提言する 〈/EVENT〉。

■PERCEPTION：認識・知覚動詞　認識動詞や知覚動詞で、主に事象に関する物理的な知覚が、「～の」などによる体言化によって導入される場合などは、PERCEPTION に分類する。但し、項が Thing であるときは、OCCURRENCE である（例「ホスピスという言葉を初めて聞いた」）。見る・観察する・見かける・眺める・聞く・聴く・耳にする・睨む・探る・感じるなどが含まれる。

〈EVENT〉@PERCEPTION の例（**太字**が注目している項）

母親が炊飯器でおでんを**作ったのを** 〈EVENT〉 見て 〈/EVENT〉、

なお、文脈により、物理的な知覚を導入するのでない場合が多く、新聞サブコーパスにおいては出現が少ない。

〈EVENT〉@PERCEPTION としない例 （**太字**が注目している項）

［個人名］に［**内容**］**について** 〈EVENT〉 聞いた 〈/EVENT〉。（インタビューであるため、OCCURRENCE）
A を B と 〈EVENT〉 見る 〈/EVENT〉。（判断であるため、OCCURRENCE や I_STATE）

■ASPECTUAL：アスペクト動詞　事象のアスペクト（相）を示す動詞が、事象を導入している場合はこれにあたる。明示的に記述されている場合に限定する。そのため、接頭語などの造語成分（例：「再」＋動詞による「再団結する」「再開発する」など、「終」「開」による「終演する・開幕する」など）を含む動詞については、ASPECTUAL に含めない。

アスペクトを明示的に表す動詞は、以下のようなものがある。

1. Initiation：始める・始まる
2. Reinitiation：再開する
3. Termination：終える・止める・終わる・中止する・停止する・あきらめる・放置する
4. Culmination：やり終える・完成させる
5. Continuation：続ける・続行する・持続する・維持する・やり通す・～し続ける・保つ

〈EVENT〉@ASPECTUAL の例 （**太字**が注目している項）

トーナメントは、日本時間 10 日夜に第 1 日が 〈EVENT〉 始まる 〈/EVENT〉。
［個人名］が勝てば **3 連覇に** 〈EVENT〉 続く 〈/EVENT〉 偉業達成。
二年目も引き続き**好調を** 〈EVENT〉 維持したい 〈/EVENT〉。
とろ火状態を 〈EVENT〉 保つ 〈/EVENT〉。

■I_ACTION（Intensional Action）内包的な動作　明示された事象の導入を行う（項とする）遂行動詞は I_ACTION と分類する。遂行しない場合は後述する I_STATE として区別を行う。また、イベントが助詞によって分割されている場合の後半部（例：「連絡をとる」「明らかにする」

など）は I_ACTION と考える。次の I_STATE との差別として、挑む・予防する・遅らせる・依頼する・要求する・説得する・約束する・決定する・提案するなど、遂行性のある動詞がこれにあたる。また、同様に、REPORTING との差別として、宣言する・主張する・申し出る・断定するなど、PERCEPTION との差別として、調査する・精査するなどが I_ACTION にあたる。

〈EVENT〉@I_ACTION の例（**太字**が注目している項）

女性が**受け入れられるべきかと** 〈EVENT〉 問われれ 〈/EVENT〉 ば、イエスだ。
再建を国際社会全体で 〈EVENT〉 取り組む 〈/EVENT〉 契機。
支払えないケースが 〈EVENT〉 出ている 〈/EVENT〉。
［個人名］は速い**転がりを** 〈EVENT〉 確かめていた 〈/EVENT〉。

■I_STATE（Intensional States）内包的な静態動詞　事象を導入する（項とする）が、事象を遂行しない動詞は I_STATE とする。代替・候補が言及されるなどの状態の導入が主となる。主に思考動詞や感情動詞がこれにあたり、信じる・思う・望む・欲する・期待する・計画するなどの思考動詞のほか、恐れる・心配する・悩むなどの感情動詞、また、遂行のない動詞として、求める・〜しようとする・〜したがるなど、〜できる・〜できないなども含まれる。

〈EVENT〉@I_STATE の例（**太字**が注目している項）

連覇を 〈EVENT〉 狙う 〈/EVENT〉。**生活が** 〈EVENT〉 できる 〈/EVENT〉。
未現像でも 〈EVENT〉 構いません 〈/EVENT〉。
よく見ていてくれたと 〈EVENT〉 感謝する 〈/EVENT〉。（遂行性がないため、I_ACTION ではない）

■STATE 事象とみなす静態動詞、形容詞　時間的順序関係と直接かかわらない場合、文書作成時間に従属しない場合には、〈EVENT〉 タグをつけないが、以下の種類の静態動詞（工藤 (1995)）と形容詞について、時間と関わる場合に限り、〈EVENT〉@STATE とする。

1. 存在動詞：ある・いる・存在する・点在するなど
2. 空間的配置動詞：そびえている・ひしめきあっている・面している・隣接しているなど
3. 関係動詞：値する・あたる・あてはまる・相当する・意味する・示す・適するなど
4. 特性動詞：甘すぎる・大きすぎる・泳げる・話せる・似合うなど

〈EVENT〉@STATE の例

マネジャーに就任する意向が 〈EVENT〉 ない 〈/EVENT〉 ことを明らかにした。（存在）
東京湾岸でも生活 〈EVENT〉 できる 〈/EVENT〉 環境さえあれば。（特性動詞。この場合、「生活ができる」であれば I_STATE）
彼女のようにモノをはっきり 〈EVENT〉 言える 〈/EVENT〉 ことがこれからは大切だ。（特性動詞）
おいしく 〈EVENT〉 食べられます 〈/EVENT〉。（特性動詞）

2.3　時間的順序関係の認定

〈EVENT〉 タグにより認定した事象表現に対して、時間的順序関係を認定する。表 2 に示す、Allen の範囲代数 (Allen (1983)) に基づくラベル（13 種類）を付与する。

尚、二つの事象表現が during/equal/contains の 3 つの時間的順序関係にある場合、部分事象の関係か全く同一の事象の関係であることがありうる。そのような場合には表 3 の三つのラベ

表 2　Allen の範囲代数に基づく時間的順序関係ラベル

ラベル	意味
after	時間・事象表現 A が事象表現 B より後に起こる
met-by	時間・事象表現 A が事象表現 B の直後に起こる
overlapped-by	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点は B の開始点より後、A の終了点は B の終了点より後である
finishes	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点は B の開始点より後、A の終了点と B の終了点は同時である
during	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点は B の開始点より後、A の終了点は B の終了点より前である
started-by	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点と B の開始点は同時、A の終了点は B の終了点より後である
equal	時間・事象表現 A と事象表現 B の時間的重なりが完全に一致する
starts	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点と B の開始点は同時、A の終了点は B の終了点より前である
contains	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点は B の開始点より前、A の終了点は B の終了点より後である
finished-by	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点は B の開始点より前、A の終了点と B の終了点は同時である
overlaps	時間・事象表現 A と事象表現 B の間に時間的な重なりがあるが、A の開始点は B の開始点より前、A の終了点は B の終了点より前である
meets	時間・事象表現 A が事象表現 B の直前に起こる
before	時間・事象表現 A が事象表現 B より前に起こる

表 3　事象-部分事象間関係を表現するラベル

ラベル	時間的順序	意味
is_included	(during 相当)	事象表現 A は事象表現 B の一部（部分事象：subevent）である （時間的な重なりがあり、A の開始点は B の開始点より後、A の終了点は B の終了点より前である） 例えば、「卵を 割る」は「オムライスを 作る」と is_included の関係にある。
identity	(equal 相当)	事象表現 A と事象表現 B は全く同じ事象を示す（言い換え） 例えば、「オムライスを 作る」と「オムライスを 料理する」は identity の関係にある。
includes	(contains 相当)	事象表現 B は事象表現 A の一部（部分事象：subevent）である （時間的な重なりがあり、A の開始点は B の開始点より前、A の終了点は B の終了点より後である） 例えば、「オムライスを 作る」は「卵を 割る」と includes の関係にある。

ルを付与する。

計 13+3 種類のラベルをまとめると図 4 のようになる。このほかにテキストの情報だけでは全く時間的順序関係がわからない場合に付与するラベルとして “vague” を利用する。

これらの計 13+3+1 種類のラベルを「文書作成日時と事象表現の順序関係 (“DCT” と呼ぶ)」「同一文内の時間表現と事象表現間順序関係 (“T2E” と呼ぶ)」「隣接事象表現間順序関係 (“E2E” と呼ぶ)」「隣接文の末尾の事象表現間順序関係 (“MATRIX” と呼ぶ)」の 4 種類の表現対について付与する。

尚、アノテーション作業に際し、以下の点に注意した。

- 時間は基本的に区間としてアノテーションを行う。1 秒でも区間とする。
- 事象は瞬間動詞は点とし、それ以外は区間を考慮する。
- 状態動詞などで開始点・終了点がわかりにくいものは、前工程の ⟨EVENT⟩ タグの認定時で排除されているべきだが、わかりにくい場合には作業者の理解にゆだねる。

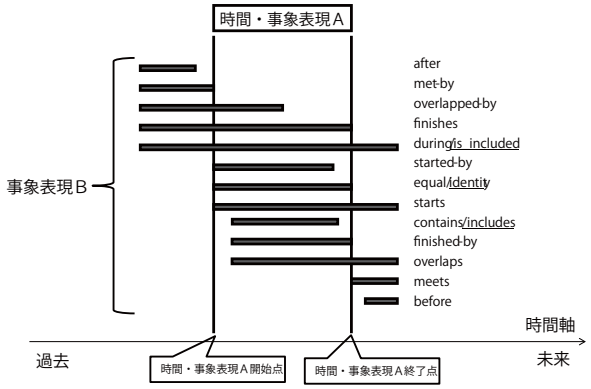


表 4　時間的順序関係ラベル一覧

表 5　時間情報表現の分布

時間表現分類	(⟨TIMEX3⟩@type)	件数
文書作成日時	(DATE)	54
日付表現	(DATE)	727
時刻表現	(TIME)	107
時間表現	(DURATION)	291
頻度・集合表現	(SET)	19
合計		1198

表 6　事象表現の分布

	⟨EVENT⟩@class	件数
項に事象を取らない事象表現	OCCURRENCE	2367
項に事象を取る事象表現	(5 種類全て)	(1291)
	REPORTING	126
	PERCEPTION	27
	ASPECTUAL	63
	I_ACTION	880
	I_STATE	195
静態表現で事象表現として認めるもの	STATE	181
合計		3839

3. アノテーション情報の分析

3.1 事象表現の認定とクラス分類

時間的順序関係を行う前に、時間情報表現と事象表現の範囲を切り出す必要がある。時間情報表現については先行研究によりなされており、今回対象新聞サブコーパス 54 ファイル上の分布は表 5 のようになっている。事象表現の認定とクラス分類については、作業者二人と監督者一人と助言者一人で議論しながら作業を行った。クラス分類を含めて 75%-80% の一致率がコンスタントに得られるまで作業者二人が同一ファイルを作業し、基準が固まった時点で分担して作業を行った。分布は表 6 のとおり。

3.2 時間的順序関係の認定

作業者三人により時間的順序関係認定作業を開始した。計 13+3+1 種類のラベルを「文書作成日時と事象表現の順序関係 (“DCT”)」「同一文内の時間表現と事象表現間順序関係 (“T2E”)」「隣接事象表現間順序関係 (“E2E”)」「隣接文の末尾の事象表現間順序関係 (“MATRIX”)」の 4 種類の表現対に対して付与した。

以下、作業者三人分の作業結果を示し、考察する。表 7 が 13+3+1 種類のラベルと 4 種類の表現対ごとに集計したものである。∩ の三つの数字は、三人の作業者が何件その関係を認定したかを示す。括弧右 “=” 以下の数字はその中で三人で一致した件数を示す。まず、ラベルの件数として、始点・終点の一致を必要としない “after”, “during”, “contains”, “before” の頻度が多かった。始点・終点のいずれかの一致を必要とするラベルのうち、もっとも多いものは時間軸上の完全の一致を示す “equal” であった。また “vague” についても複数の作業者が認定し、314 件一致しているところから、文脈を用いても時間的順序関係が推定できないものが少なからずあることがわかる。

表 8 に 4 種類の関係ごとの一致率を集計したものを示す。一致率の評価基準として、「ラベル

表 7　時間的順序関係ラベルの評価

ラベル	DCT	T2E	E2E	MATRIX	全て
関係数	3839	2188	2972	1245	10244
after	2352∩2326∩2133=1961	396∩441∩432=315	627∩631∩639=432	292∩284∩277=198	3667∩3682∩3481=2906
met-by	0∩0∩0=0	5∩10∩2=2	18∩12∩3=2	7∩3∩2=1	30∩25∩7=5
overlapped-by	11∩5∩4=2	59∩52∩42=20	3∩3∩2=0	0∩0∩1=0	73∩60∩49=22
finishes	2∩8∩1=0	10∩1∩11=0	5∩8∩5=1	1∩0∩0=0	18∩17∩17=1
during	449∩424∩650=217	105∩100∩113=62	206∩139∩225=67	112∩86∩134=43	872∩749∩1122=389
started-by	1∩0∩0=0	9∩2∩8=0	3∩14∩6=2	0∩3∩0=0	13∩19∩14=2
equal	1∩17∩0=0	37∩70∩51=19	263∩412∩307=154	62∩140∩90=29	363∩639∩448=202
starts	2∩0∩0=0	30∩9∩14=2	6∩16∩2=0	0∩1∩1=0	38∩26∩17=2
contains	164∩85∩144=63	830∩853∩868=671	299∩292∩344=117	148∩152∩188=64	1441∩1382∩1544=915
finished-by	0∩0∩0=0	3∩3∩0=0	6∩7∩6=0	1∩3∩0=0	10∩13∩6=0
overlaps	2∩2∩4=1	75∩84∩70=32	6∩27∩5=0	1∩4∩3=0	84∩117∩82=33
meets	1∩13∩0=0	25∩26∩2=2	88∩88∩32=22	9∩15∩0=0	123∩142∩34=24
before	739∩767∩746=572	389∩360∩383=288	1058∩994∩1098=713	418∩436∩422=294	2604∩2557∩2649=1867
is_included	0∩0∩0=0	0∩0∩0=0	19∩2∩6=1	6∩0∩1=0	25∩2∩7=1
identity	0∩0∩0=0	0∩0∩1=0	11∩7∩24=2	16∩5∩15=2	27∩12∩40=4
includes	0∩0∩0=0	0∩0∩0=0	27∩10∩2=1	18∩2∩0=0	45∩12∩2=1
vague	115∩191∩157=38	212∩177∩191=100	327∩309∩265=128	154∩111∩111=48	808∩788∩724=314

作業者 A の認定数 ∩ 作業者 B の認定数 ∩ 作業者 C の認定数=三者で一致した件数

表 8　時間的順序関係ラベルの一致率の分析

一致率評価基準	DCT	T2E	E2E	MATRIX	全て
関係数	3839	2188	2972	1245	10244
ラベル 13+3+1	2854(0.743)	1513(0.691)	1642(0.552)	679(0.545)	6688(0.653)
ラベル 13+1	2854(0.743)	1513(0.691)	1667(0.561)	697(0.560)	6731(0.657)
ラベル 5+1	2873(0.748)	1605(0.734)	1862(0.627)	776(0.623)	7116(0.695)
ラベル 3+1	2880(0.750)	1644(0.751)	1884(0.634)	780(0.627)	7188(0.702)

三人の作業者が一致したラベル数 (一致率)

13+3+1 種類を区別するもの (ラベル 13+3+1)」「部分集合であるか否かを区別せず、ラベル 13+1 種類を区別するもの (ラベル 13+1)」「TempEval で用いられているラベル 5+1 種類 (“BEFORE”, “BEFORE-OR-OVERLAP”, “OVERLAP”, “OVERLAP-OR-AFTER”, “AFTER”, “VAGUE”) に縮退するもの (ラベル 5+1)」「ラベル 3+1 種類 (“BEFORE”, “OVERLAP”,“AFTER”, “VAGUE”) に縮退するもの (ラベル 3+1)」の 4 種類を用いる。まず、もっとも厳しい一致率評価基準 (ラベル 13+3+1) でも 65.3% の三人の一致率 (Cohen’s kappa 0.733) であった。我々の手法では事象構造の認定については複数人で合議的に行い、その後限られた関係について時間的順序関係アノテーションを行っているが、事象構造の認定と関係対に対する関係タグ付与作業を同時に行っている英語のデータ TimeBank 1.2 における ⟨TLINK⟩ の一致度 (関係対の認定の一致率 55% と一致した関係対に対する関係タグの一致率 77%) と比較しても遜色ないレベルだと考える。4 種類の関係については、 “DCT” が最も一致率が高く、次に “T2E” が高かった。これは片方が時間情報表現である場合に、時間情報表現側の時間軸上の絶対位置が推定しやすいことによるからだと考える。一致率評価基準について始点・終点の境界値一致の認定を緩和することで、 “E2E”, “MATRIX” の関係は若干一致率があがることから、被験者間で事象構造の時間的な境界値にずれが生じていることがわかる。

4. おわりに

本稿では『現代日本語書き言葉均衡コーパス』に対する時間的順序関係アノテーションについて現状を報告した。時間的順序関係を付与する事象表現の認定にあたり、時間軸上の性質や取りうる項が事象である場合に他の事象表現にどのような影響を与えるのかに基づいて、事象表現を 7 種類に分類した。認定した事象表現に対し時間的順序関係を付与し、一致率などについて報告した。

謝辞　本研究を行うにあたり、助言いただきました日本ＩＢＭの吉川克正氏、アノテーションに従事していただいた方々に感謝いたします。本研究は文科省科研費特定領域研究「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」、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

小西光・浅原正幸・前川喜久雄 (2012).「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」　第 2 回コーパス日本語学ワークショップ発表論文集.

小西光・浅原正幸・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」　言語処理学会第 19 回年次大会発表論文集.

Allen, J. (1983). “Maintaining knowledge about temporal intervals.” *Communications of the ACM*.

Boguraev, B., and R. Kubota Ando (2005). “TimeML-Compliant Text Analysis for Temporal Reasoning.” *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, pp. 997–1003.

Boguraev, B., and R. Kubota Ando (2006). “Analysis of TimeBank as a Resource for TimeML parsing.” *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*.

DARPA TIDES (2004). *The TERN evaluation plan; time expression recognition and normalization*. Working papers, TERN Evaluation Workshop.

Grishman, R., and B. Sundheim (1996). “Message Understanding Conference-6: a brief history.” *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pp. 466–471.

Mani, I. (2006). “Machine Learning of Temporal Relations.” *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-2006)*, pp. 753–760.

Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, B. Sundheim, L. Ferro, M. Lazo, I. Mani, and D. Radev (2003). “The TIMEBANK Corpus.” *Proceedings of Corpus Linguistics 2003*, pp. 647–656.

Pustejovsky, J., J. Casta no, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz (2003). “TimeML: Robust Specification of Event and Temporal Expressions in Text.” *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*.

Sekine, S., K. Sudo, and C. Nobata (2002). “Extended Named Entity Hierarchy.” *The Third International Conference on Language Resources Evaluation (LREC-02)*.

Setzer, A. (2001). “Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study.” Unpublished doctoral dissertation, University of Sheffield.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Kats, and J. Pustejovsky (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 75–80.

Verhagen, M., R. Saurí, T. Caselli, and J. Pustejovsky (2010). “SemEval-2010 Task 13: TempEval-2.” *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pp. 57–62.

工藤真由美 (1995).『アスペクト・テンス体系とテクスト - 現代日本語の時間の表現-』　ひつじ書房.

工藤真由美 (2004).『日本語のアスペクト・テンス・ムード体系標準語研究を超えて』　ひつじ書房.

中村ちどり (2001).『日本語の時間表現』　くろしお出版.

橋本泰一・中村俊一 (2010).「拡張固有表現タグ付きコーパスの構築–白書，書籍，Yahoo! 知恵袋コアデータ–」　言語処理学会第 16 回年次大会発表論文集, pp. 916–919.

実行委員会，IREX (1999).『IREX ワークショップ予稿集』.

『理研母子会話コーパス（R-JMICC)』構築の試みと研究成果 — 対乳児自発音声における 日本語特有の韻律的・分節的特徴の解明を目指して —

西海枝洋子（理化学研究所 脳科学総合研究センター 言語発達研究チーム）†
渡辺和希（筑波大学大学院）
小西隆之（理化学研究所 脳科学総合研究センター 言語発達研究チーム）
伊藤直子（筑波大学大学院）
金礪愛（早稲田大学人間科学学術院）
五十嵐陽介（広島大学文学研究科文学部）
宮澤幸希（理化学研究所 脳科学総合研究センター 言語発達研究チーム）
西川賢哉（理化学研究所 脳科学総合研究センター 言語発達研究チーム）
馬塚れい子（理化学研究所 脳科学総合研究センター 言語発達研究チーム）

Riken Japanese Mother Infant Conversation Corpus (R-JMICC) — Compilation and Recent Findings of Japanese-Specific Prosodic and Segmental Characteristics in Infant-Directed Speech —

Yoko Saikachi (Lab. for Language Development, RIKEN Brain Science Institute)
Kazuki Watanabe (Graduate school of University of Tsukuba)
Takayuki Konishi (Lab. for Language Development, RIKEN Brain Science Institute)
Naoko Ito (Graduate school of University of Tsukuba)
Ai Kanato (Waseda University)
Yosuke Igarashi (Hiroshima University)
Koki Miyazawa (Lab. for Language Development, RIKEN Brain Science Institute)
Ken'ya Nishikawa (Lab. for Language Development, RIKEN Brain Science Institute)
Reiko Mazuka (Lab. for Language Development, RIKEN Brain Science Institute)

1. はじめに

大人は乳児に語りかける際、韻律的・分節的特徴を強調した独特な話し方をする（対乳児音声、Infant-Directed Speech; IDS）。全体的に声が高く、ゆっくりとして、ピッチの変動幅が大きいなどの IDS の特徴は、多言語において報告されており(Fernald et al. 1989, Soderstrom 2007)、その普遍性と言語獲得における重要性が示唆されてきた。しかし、日本語は他の言語とは異なる独特な音韻体系、そして韻律構造を持つ。その特異性を考慮した上で、対乳児音声の音響的特徴を定量的に分析し、明らかにした研究はほとんど行われてこなかった。

そこで本研究チームでは、ここ数十年の間に開発が進められてきた音声コーパスの構築技術を応用し、『理研母子会話コーパス（R-JMICC）』(Mazuka, Igarashi, and Nishikawa 2006, 五十嵐, 馬塚 2006)の構築を進めている。このコーパスは、2005 年に収録を行った 18-24 ヶ月の乳児を持つ母親 22 名による自発的な会話音声データ（対乳児音声および対成人音声）

† ysaikachi@brain.riken.jp

と、その 4 年後の 2009 年に収録を行った同じ母親による音声データ（対乳児音声、対成人音声および読み上げ音声）で構成されており、『日本語話し言葉コーパス（CSJ）』（前川 2004, 2006）にほぼ準拠した、形態論情報、分節音情報、および韻律情報が付与されている。これらの付加情報を活用することにより、韻律句の構成やピッチアクセント、句末音調といった日本語の韻律特性を考慮しながら、自発音声における韻律的・分節的特徴を分析することが可能となった。

本稿では、まずコーパスの概要を説明した後に（2 節）、コーパス分析に基づく最近の研究成果の具体例として、日本語対乳児音声における 1) ピッチレンジの拡大という韻律的強調の局所性（Igarashi and Mazuka 2008, Igarashi et al. 2009）（3 節）と、2) 長短母音の持続時間長の分布特性 (Bion et al. in press)（4 節）を紹介する。

2. データ

2.1. 『理研母子会話コーパス』

『理研母子会話コーパス（R-JMICC）』は、2005 年に収録されたデータと、2009 年に収録されたデータで構成される（表 1 参照）。2005 年収録のデータは、会話形式の自発音声（以下 05 セット）、2009 年収録のデータは、会話形式の自発音声（以下 09 セット）と、文章の読み上げ音声（以下 ATR セット）から構成される。

表 1　『理研母子会話コーパス』の概要

収録年 （参加者数）	発話スタイル （名称）		内容 (収録時間/文の総数)	総時間	語数
2005 年 (22 人)	自発音声 (05 セット)	IDS	絵本を見ながらの会話 （15 分）	11 時間	49340
			玩具で遊びながらの会話 (15 分)		
		ADS	育児に関する会話 （10 分）	3 時間	25012
2009 年 (20 人)	自発音声 (09 セット)	IDS	絵本を見ながらの会話 （10 分）	7 時間	33280
			玩具で遊びながらの会話 (10 分)		
		ADS	育児に関する会話 （10 分）	3 時間	25901
	読み上げ音声 （ATR セット）		ATR 音素バランス 503 文 A セット (50 文)	4 時間	14538

2.1.1 参加者

05 セットの参加者は、母親 22 名（25-43 歳、平均年齢 33.0, SD±3.6）とその子供である。母親は全て関東地方（東京・神奈川・埼玉・千葉）出身であり、標準日本語を話す。子供の月齢は 18-24 カ月（平均 20.4, SD±2.7 カ月）であった。2009 年には、2005 年の収録に参加した親子 22 組中 20 組が再度参加した。

2.1.2 音声の収録環境

理化学研究所言語発達研究チーム内の防音室で、母親にヘッドセット型コンデンサマイク(CROWN, CM-312A)を装着してもらい、一組ずつ録音を行った。また、直接の分析対象とはしていないが、2005 年にはコンデンサマイク(Behringer, B-5)をテーブル上に配置し、2009 年には子どもにもヘッドセット型ダイナミックマイク(SHURE, SM10A)を装着してもらい、子供の発話も収録した。音声は、DAT(TASCAM, DA-P1)を用いて収録した(44.1 kHz、16 ビット)。

2.1.3 収録内容

05 セットの内容は 3 種類である。まず、IDS として 1)絵本を見ながら母親が子供に話しかける音声と、2)玩具で遊びながら子供に話しかける母親の音声を収録した。次に、対成人発話(Adult-Directed Speech; ADS)として、3)同じ母親による成人の実験者(同年代かつ子育て中の女性）との会話音声も収録した。ADS は、育児に関する内容が多かった。2009 年には、05 セットと同一のタスクを用いた自発音声(09 セット)に加え、母親による読み上げ文(ATR セット)の収録を行った。

2.2. 研究用付加情報

R-JMICC には、書き起こしテキスト（転記テキスト)、形態論情報（単語境界や、品詞、および活用形についての情報)、分節音情報、韻律情報といった様々な研究用付加情報が付与されている（図 1 参照）(Mazuka, Igarashi, and Nishikawa 2006) [1]。付加情報は、概ね CSJ（国立国語研究所（編）2006）に準拠している。

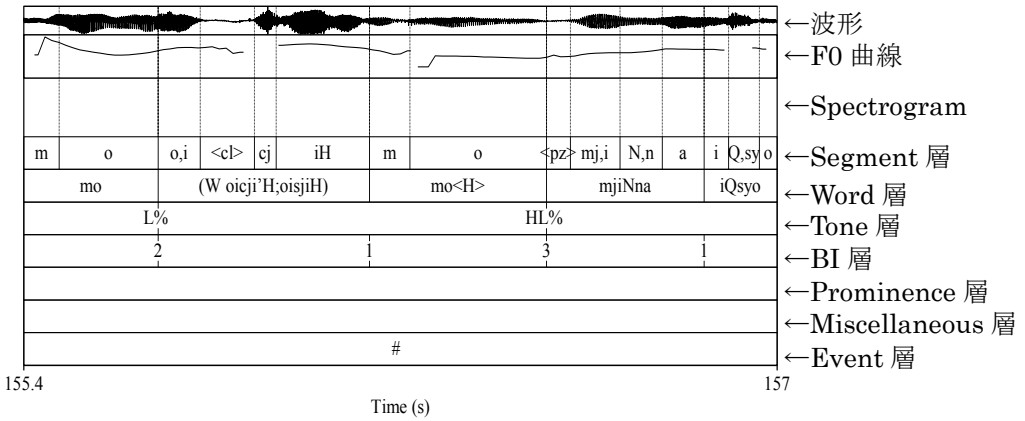


図 1　研究用付加情報の一例

(a) 音切り・書き起こし：発話を200msec以上のポーズで区切り（音切り）、区切られた単位（転記基本単位；IPU）ごとに、聞き取れる範囲で忠実に発話を記す。談話・音声現象（フィラー、語断片、歌声、対成人音声等）を記述するためのタグもこの段階で付与する。

(b) 形態論情報付与：短単位（辞書の見出し語に相当）および長単位（複合語を一つと扱う）という二種類の形態論的単位を認定し、それぞれの単位に品詞などの付加情報を付与する。

(c) 分節音情報および韻律情報付与：分節音情報として、子音・母音の種類とその境界位置に関する情報を、韻律情報として、アクセント・イントネーション（韻律句境界のレベル、句末音調の種類等）に関する情報を記す。韻律ラベリングスキームは、CSJで採用されているX-JToBI方式の体系を一部改訂したものである（五十嵐、馬塚 2006)。

3. 日本語対乳児自発音声における韻律的特徴：ピッチレンジ拡大の局所性

3.1 研究の背景および目的

[1] 05 セットは、一連のアノテーション作業が終了しており、3 節および 4 節で紹介する研究成果は、05 セットの解析結果に基づくものである。ATR セットも一通りアノテーション作業が終了している。09 セットは、(a)-(b)の作業が終了しており、現在 (c)の分節音情報および韻律情報の付与を進めている。

これまで多くの研究により、ピッチレンジの拡大という韻律的強調が IDS の主要な特徴とされてきたが、日本語にはそのような強調が存在しないのではないかと示唆されてきた（Fernald et al. 1989）。五十嵐らは、日本語特有の韻律的強調の実態を明らかにするために、アクセント、ダウンステップ、句末音調などの日本語に特徴的な韻律構造による影響を考慮した上で、R-JMICC(05 セット)におけるピッチレンジの分析を行った（Igarashi and Mazuka 2008, Igarashi et al. 2009）。

3.2　日本語の韻律構造

3.2.1　韻律句の特徴

R-JMICC のラベリングスキームである X-JoBI では、アクセント句（Accentual Phrase, 以下 AP）とイントネーション句(Intonational Phrase, 以下 IP）の 2 種類の韻律句が仮定されている。AP は、句頭で上昇し、その後句末にかけて段々と下がり、低く終わるというようなピッチ曲線で特徴づけられる（Venditti 2005）。AP がアクセントを含む有核句の場合、アクセントによる急激なピッチの下降があるため、無核句と比較すると、AP の最低ピッチが引き下げられ、ピッチレンジが拡大する。

AP より階層的に上位に位置づけられる IP は、アクセント核が、後続する AP の最大ピッチを反復的に低下させるダウンステップという音韻現象が生じる領域と定義され、IP 境界でピッチリセット（前の文脈とは独立した新たなピッチレンジの設定）が生じる。IP 内のアクセントの数が増えると、ダウンステップの影響により一番目の AP の最大ピッチが高くなり、そのため IP のピッチレンジが拡大すると考えられる。

3.2.2　句末複合境界音調（Boundary Pitch Movement; BPM）

AP の終端には、下降調だけではなく、上昇調や上昇下降調などの局所的音調が生じる。句末複合境界音調（Boundary Pitch Movement, 以下 BPM）と呼ばれるこの音調は、質問、強調、継続など、語用論的な意味あるいは発話意図の伝達に重要な役割を果たしており、R-JMICC では、主に H%（上昇調 1)、LH%（上昇調 2)、HL%(上昇下降調)によって表現されている（Igarashi and Mazuka 2008, Igarashi et al. 2009）（図 2)。

韻律句の BPM 以外の場所（以下、主要部と呼ぶ）のピッチパターンは、アクセントやダウンステップ等、語彙情報によって規定されているため、韻律的強調による変化の度合いは限定的と考えられる。一方、韻律句末に生じる BPM はそのような制約がなく、発話の意図などを表現するためにピッチの特徴を強調しやすい場所であると考えられる。

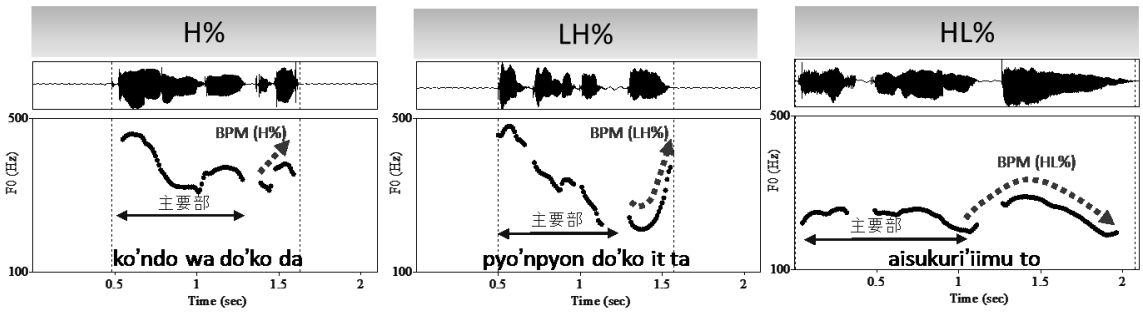


図 2　BPM および主要部のピッチ曲線

3.3　コーパス分析[2]

[2] IDS データは、絵本を読みながらの会話音声のみ分析対象としている。また、22 名中 1 名の母親のデータは声質に問題があるため分析対象外としている。

3.3.1 Utterance(発話)[3]全体の特徴

五十嵐らはまず、Utterance 全体を対象とした分析を行った(図 3)。その結果、ピッチの最大値・平均値・最小値は、ADS と比較して IDS では有意に高くなっていたが、IDS におけるピッチレンジの拡大は観察されず、先行研究（Fernald et al. 1989）と同様の結果が得られた。

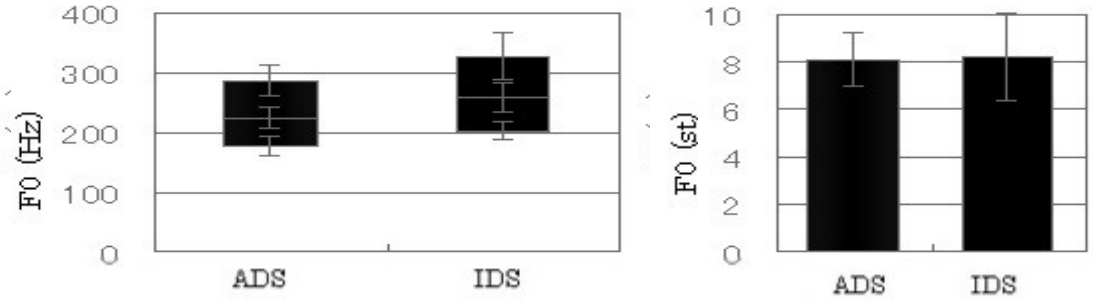


図 3　Utterance 全体のピッチの特徴。左図：ピッチの最大値（図の上限値）・平均値（図の中央値）・最小値（図の下限値)。右図：ピッチレンジ。エラーバーは標準偏差を示す。

3.3.2 BPM および主要部の特徴

ピッチレンジの拡大を Utterance 全体で測定するのではなく、主要部と BPM に分けて分析を行った。まず、BPM の相対頻度を分析したところ、H%と LH%は、IDS でより頻繁に現れ、逆に、HL%は、ADS でより頻繁に観察された(図 4)。

次に、BPM 毎にピッチの特徴量を分析したところ、IDS では、全ての BPM の種類で、ピッチの最大値および平均値が ADS よりも有意に高く、ピッチレンジが有意に拡大していた(図 5、図 6)。ピッチの最小値については、H%と LH%では、IDS のほうが有意に高かったが、HL%では発話者間に有意差はなかった。

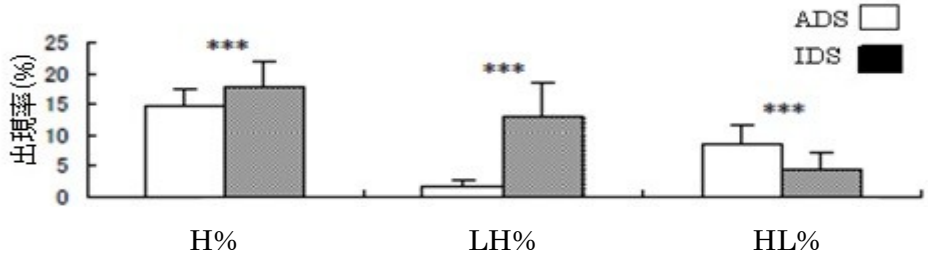


図 4　BPM の出現率

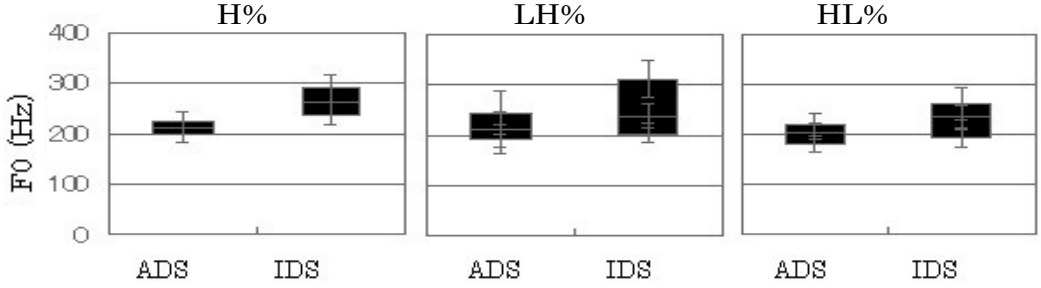


図 5 BPM のピッチの特徴：最大値・平均値・最小値

[3] ここでは、Utterance（発話）を、200msec 以上のポーズが後続する IP 境界で区切られる単位と定義している。

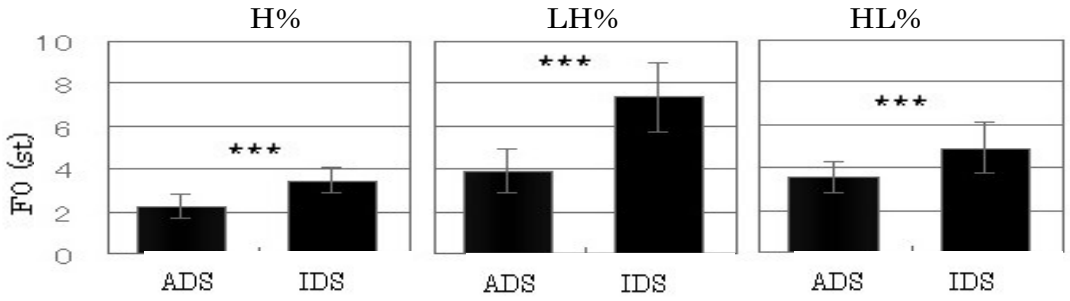


図 6　BPM におけるピッチレンジ

次に、主要部のピッチの特徴量を比較したところ、ピッチの最大値・平均値・最小値は、ADS と比較して IDS では有意に高かったが、ピッチレンジは ADS が IDS と比較して有意に大きかった(図 7)。

つまり、Utterance を主要部と BPM に分けて分析すると、IDS におけるピッチレンジの拡大は韻律句末の BPM でのみ顕著に現れ、主要部では IDS よりも ADS のほうが、ピッチレンジが大きいことが明らかになった。

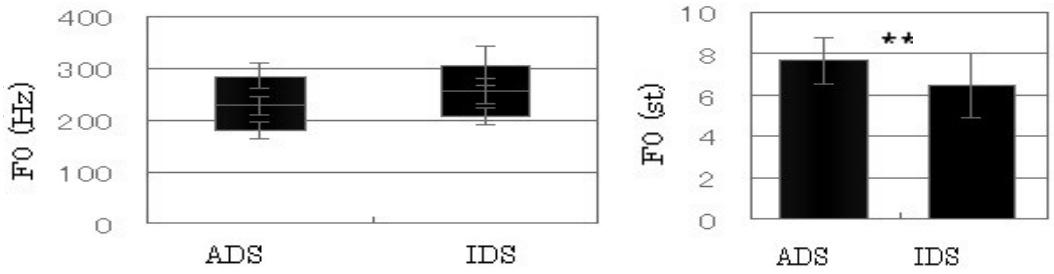


図 7 主要部のピッチの特徴：最大値・平均値・最小値（左図)、ピッチレンジ（右図）

3.3.3　IP の長さを考慮した場合の主要部におけるピッチレンジ

最後に、主要部におけるピッチレンジを、IP の長さを考慮して分析した結果を紹介する。まず、持続時間長(msec)、モーラ数、短単位数、IP の数、AP の数、アクセントの数といった項目について計測したところ、全てにおいて、ADS のほうが IDS より長かった。

次に、IP の長さ（アクセント数）を統一して ADS と IDS のピッチレンジを比較した（図 8)。 その結果、アクセント数が 1、2、3 個では、IDS は ADS と比べて、ピッチレンジが有意に大きかったが、0 個および 4 個の場合はレジスター間に有意な差は認められなかった。

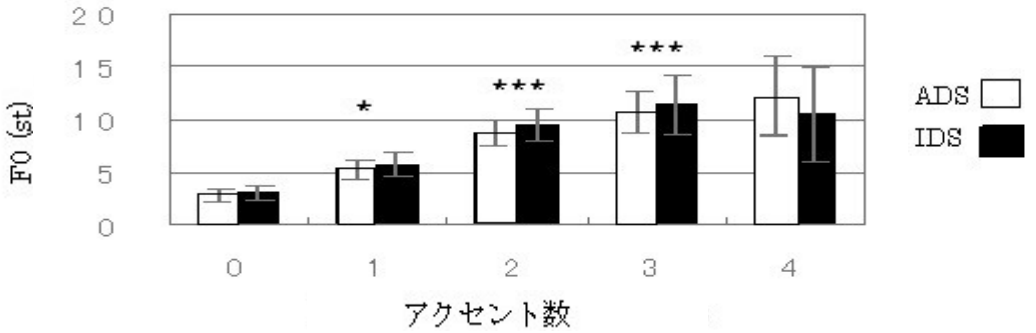


図 8 IP の長さ（アクセント数）を統一した際のピッチレンジ

つまり、IP の主要部のピッチレンジは主にアクセントの数によって決められていること、そして主要部で ADS の平均ピッチレンジが大きいのは、ADS では、より多くのアクセントが含まれているためであり、このことが IDS におけるピッチレンジの拡大を分かりづらくしている理由であることが明らかになった。

3.4 ピッチの特徴に関するまとめ

五十嵐らは、韻律情報が付与されている R-JMICC を使用して、日本語特有の韻律構造を考慮した分析を行うことにより、日本語の対乳児自発音声では、ピッチレンジの拡大という韻律的特徴の強調が、韻律句末の BPM に局所的に現れることを明らかにした。五十嵐らの研究の大きな意義は、ここ数十年で研究が積み重ねられてきた韻律の音韻構造 (Ladd 1996, Pierrerhumbert and Beckman 1988)に基づく分析を行うことによって、対乳児音声における韻律的特徴の強調が、全ての言語で同じように現れているのではなく、個別言語独自の韻律構造の枠内で表出するものであることを示したことである。今後、多様な言語で解析が進むことにより、個々の言語特有の対乳児音声の特徴だけではなく、構造的な違いが大きい言語間における共通した特徴を明らかにすることが出来れば、音声言語獲得における言語普遍性の解明への一歩となることが期待される。

4. 日本語の IDS における長短母音の出現頻度分布特性

本節では、長短母音という日本語特有の音韻対立に焦点をあてて、対乳児自発音声における音韻の出現頻度分布特性を分析した Bion らによる研究(Bion et al. in press)を紹介する。

4.1 研究の背景

日本語では、「床」/toko/と「渡航」/tokoo/のように、母音の長短が語の意味を区別するが、英語のような言語では母音の長短のみで語の意味を区別することはない。では、乳児は、一体どのようにして母語特有の音韻的な対立構造を習得しているのであろうか。乳児にとっての主要な入力音声である対乳児音声には、何か手がかりとなるような情報が含まれているのだろうか。

Werker らは、日本語を母語とする母親による読み上げ音声の分析に基づき、日本語では長短母音間に母音長の確かな差が存在し、音韻的な長さの指標として使用することが出来ることを指摘している(Werker et al. 2007)。しかし、この結果は、無意味単語対を用いた、発話内容を予め統制している読み上げ音声の収録に基づく結果であり、多様な実在語を豊富に含む自発的な日本語対乳児音声においても、同様の結果が得られるのかどうかは明らかになっていない。

また、乳児のように、長母音、短母音というカテゴリーが存在すること自体をまだ認識していない場合、入力音声における音韻カテゴリーの出現頻度分布特性が、音韻の獲得に重要な役割を果たしていることが、統計学習モデルを用いた研究 (Vallabha et al. 2007)や、乳児を対象とした実験的研究 (Maye et al. 2002)によって指摘されている。このような研究では、入力音声の特徴として、1) 異なる音韻カテゴリーが入力音声に同頻度で存在すること、そして 2)出現頻度分布が二つの山を持つような双峰性であること、という二つの条件を前提にしているが、日本語の対乳児自発音声においても、前提とされている頻度分布特性が存在しているのかどうかは確かめられていない。

4.2 研究の目的

Bion らの研究の目的は、母語の音韻構造を習得するために重要とされている特徴が、実際の入力音声において存在しているのかどうか、R-JMICC(05 セット)における日本語の長

短母音の出現頻度分布特性を分析し検証することである。

4.3　コーパス分析

4.3.1　平均値の比較

Bion らはまず、自発的な日本語対乳児発話音声においても、長短母音間に確実な音響的な違いが存在するのかどうかを検討するために、日本語の各母音(a,e,i,o,u)について、短母音と長母音の長さの平均を比較した。図 9 に示すように、母音の種類に関係なく長母音は短母音よりも長く、その差は統計的に有意であった。

次に、Werker et al.(2007)に従って、母音のカテゴリーが従属因子、母音長が独立因子、そして話者がランダム効果とするロジスティック回帰分析を行った。母音の持続時間長に基づいて母音のカテゴリー（短母音　vs. 長母音）を予想するモデルは有意であった。22 名それぞれの話者においても、同じような結果が得られた。

このように、大量の母音を含む自然発話の収録においても、Werker et al. (2007)と同様に、日本語の長短母音の長さには、確かな差があることが確認された。

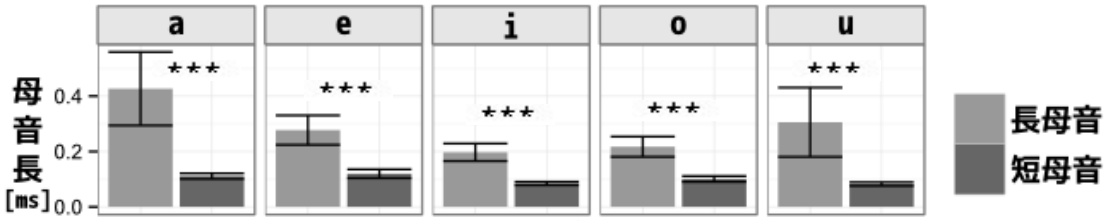


図 9　長短母音の持続時間長：被験者間平均および標準偏差

4.3.2　出現頻度分布特性の分析

次に、単純な出現頻度分布モデルによって、日本語における音韻的な長さの習得を説明することが可能なのかどうかを探るために、対乳児発話データにおける母音の頻度分布を分析した（図 10)。その結果、コーパスに現れている母音のほとんどが短母音であること（全体の 94%)、そしてそれぞれの母音について、入力音声における長短母音の分布が完全に重複していることが分かった。22 名それぞれの話者においても、同様の結果が確認された。

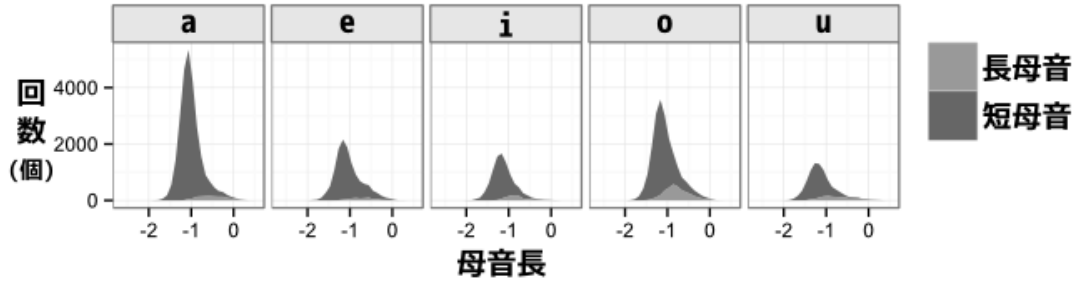


図 10 長短母音の出現頻度分布

つまり、実際の日本語の入力音声では、ほとんどの母音が短母音であり、もともとの出現頻度の偏りがあるため、長短母音を合わせた母音長の出現頻度分布は、音韻カテゴリー習得モデルが前提としたような双峰性ではなく単峰性であることが確認された。

4.4　長短母音の分析に関するまとめ

Bion らは、初期音韻獲得モデルの前提条件とされてきた入力音声における出現頻度分布の特性を検討するために、日本語の対乳児自発音声における長短母音の分析を行った。その結果、日本語の長短母音には、持続時間長に顕著な差があるが、出現頻度の偏りがあるため、出現頻度分布による音韻習得モデルのアルゴリズムのみでは習得は困難なことが示唆された。そのため、頻度分布以外にも韻律情報や語彙情報など、他の情報を考慮に入れなければ、日本語では母音長が語の意味の違いを区別しているということを、習得するのは難しいと考えられる。実際、日本語を母語として学んでいる乳児は、音韻的な母音長に対する感度を示すのに時間がかかることが明らかになっている(Mugitani et al. 2009, Sato, Sogabe, and Mazuka 2010)。最近では、音声・音韻知覚の発達過程を語彙獲得との関係で捉えようとするモデルも提唱されており (Feldman, Griffiths, and Morgan 2009)、今後の研究として、異なる習得モデルによる仮説を、入力音声における実際の語彙や音響的指標の分布によって検証していくことが重要である。

5. おわりに

本稿では、対乳児自発音声コーパス構築の取り組みと、最近の研究成果について紹介した。本コーパスのように、統制の加えられていない対乳児自発音声を、ある程度の量、収録し、詳細なアノテーションを行うことにより、日本語特有の音韻・韻律構造を考慮しながら、音声的特徴を分析することが可能になった。本稿で紹介したように、R-JMICC(05 セット)の解析によって、今までの研究では見落とされていた日本語特有の韻律的強調の実態の解明（Igarashi and Mazuka 2008, Igarashi et al. 2009)、そして母語固有の音韻獲得モデルとして重要視されてきた音韻の出現頻度分布特性の検証(Bion et al. in press)に貢献してきた。

このような研究成果は、主に 18-24 ヶ月の乳幼児に語りかけた母親の発話データ（05 セット）の分析に基づくものである。対乳児発話音声は、語りかける子供の年齢の変化や言語発達とともに、その音響的特徴もさまざまな次元において次第に変化していくことが知られている(Kitamura et al. 2001)。今後、09 セットのラベリングを進め、05 セットと 09 セットの解析結果と組み合わせることで、言語発達に伴って対乳児音声がどのように変化していくのかを明らかにしていくことが可能になる。また、ATR セットの音声ラベルを合わせて解析し、読み上げ音声との比較分析を行うことで、対乳児発話および対成人発話という異なる自発音声スタイルの特徴を、より相対的に捉える事が可能になると考えられる。

謝辞

現在進行中のコーパス構築作業において韻律ラベリングに関する貴重な助言を頂いている宇都木昭氏、そして原稿準備にご協力いただいた浅井拓也氏に感謝いたします。本研究は、文部科学省科学研究費補助金・基盤研究(C)21610028（研究代表者：馬塚れい子）および若手研究(B)23720223（研究代表者：西海枝洋子）の補助を得ています。

参考文献

Bion, Ricardo, Koki Miyazawa, Hideaki Kikuchi, and Reiko Mazuka (in press) “Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech”, *PLoS ONE*.

五十嵐陽介、菊池英明、前川喜久雄 (2006)「第 7 章韻律情報」『日本語話し言葉コーパスの構築法』(国立国語研究所報告書 124), pp.347-453.

五十嵐陽介、馬塚れい子(2006)「母親特有の話し方(マザリーズ)は大人の日本語とどう違うか−理研日本語母子会話コーパス−」電子情報通信学会技術研究報告,106:43, pp.31-35.

Igarashi, Yosuke and Reiko Mazuka (2008) “Exaggerated prosody in infant-directed speech?

Intonational phonological analysis of Japanese infant-directed speech", *BUCLD 32: Proceedings of the 32nd annual Boston University Conference on Language Development*, pp.177-188.

Igarashi, Yosuke, Ken'ya Nishikawa, Kuniyoshi Tanaka and Reiko Mazuka (2009) "Pitch range expansion in Japanese infant-directed speech", *The 4th International Workshop on the Interface between Prosody and Information Structure* (Jan. 11, 2009, Otsu, Shiga, Japan).

Feldman, Naomi, Thomas Griffiths, and James Morgan (2009) "Learning phonetic categories by learning a lexicon", *Proceedings of the 31st Annual Conference of the Cognitive Science Society,* Amsterdam, pp.2208-2213.

Fernald, Anne, Traute Taeschner, Judy Dunn et al.(1989) "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants", *Journal of Child Language*, 16:3, pp.477-501.

Kitamura, Christine, Chayada Thanavishuth, Dennis Burnham, and Sudaporn Luksaneeyanawin (2001) "Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language", *Infant Behavior and Development,* 24:4, pp.372-392.

国立国語研究所編(2006)『日本語話し言葉コーパスの構築法』(国立国語研究所報告書 124).

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)「第 1 章概説」『日本語話し言葉コーパスの構築法』(国立国語研究所報告書 124), pp.1-22.

Ladd, Robert (1996) *Intonational Phonology*, Cambridge: Cambridge Univ. Press.

Maye, Jessica, Janet Werker, and LouAnn Gerken (2002) "Infant sensitivity to distributional information can affect phonetic discrimination", *Cognition,* 82:3, pp.B101-B111.

Mazuka, Reiko, Yosuke Igarashi, and Ken'ya Nishikawa (2006) "Input for learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus", 電子情報通信学会技術研究報告, 106:165, pp.11-15.

Mugitani, Ryoko, Ferran Pons, Laurel Fais et al. (2009) "Perception of vowel length by Japanese- and English-learning infants", *Developmental Psychology,* 45:1, pp.236-247.

Pierrehumbert, Janet and Mary Beckman (1988) *Japanese Tone Structure,* Cambridge: MIT Press.

Sato, Yutaka, Yuko Sogabe, and Reiko Mazuka. (2010) "Discrimination of phonemic vowel length by Japanese infants", *Developmental Psychology,* 46:1, pp.106-119.

Soderstrom, Melanie (2007) "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants", *Developmental Review,* 27, pp. 501-532.

Vallabha, Gautam, James McClelland, Ferran Pons et al. (2007) "Unsupervised learning of vowel categories from infant-directed speech", *Proceedings of the National Academy of Sciences,* 104:33, pp.13273-13278.

Venditti, Jennifer (2005) "The J_ToBI model of Japanese intonation", Jun, S. -A. (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing,* New York: Oxford Univ. Press, pp.172-200.

Werker, Janet, Feran Pons, Christiane Dietrich et al. (2007) "Infant-directed speech supports phonetic category learning in English and Japanese", *Cognition,* 103:1, pp.147-162.

中学・高校における地学教科書の文体比較
−学年の進行に伴う文体的特徴の変化−

浅石 卓真（東京大学大学院教育学研究科）†

Comparative Analysis of the Writing Styles of Earth Science Textbooks in Junior-High and High Schools: Changes in the Textual Characteristics according to School Year

Takuma Asaishi (Graduate School of Education, the University of Tokyo)

１．はじめに

本研究では、中学・高校の地学教科書（「理科（第２分野）上巻」「理科（第２分野）下巻」「地学Ⅰ」「地学Ⅱ」）を対象として、学年の進行に応じた文体的特徴の変化を明らかにする。発表者は学年、科目、時代に応じた中・高理科教科書の文体を記述的に明らかにする研究を進めており、本稿はその事例分析の一つである。

従来の教科書研究はその殆どが内容分析だが、生徒が内容を学習する際の手掛かりは教科書の文章表現であることから、その形式的な特徴（文体）を明らかにすることも重要な課題である。教科書の文体は学年や科目により大きく異なるが、学年に応じた文体は各教科の内容を段階的に学習していく上で特に重要である。

近年の文体分析では、因子分析や主成分分析の手法が用いられることが多い。例えば安本・本多（1981）は 100 人の現代作家の作品について、15 の文体的特徴を調査して因子分析を行い、得られた３つの因子により作家を８つに類型化している。また陳（2005）は、新聞、週刊誌、教科書について 25 の文体的特徴を調査して主成分分析を行い、同様に類型化した結果、教科書の文体は新聞と比較的近いことなどを指摘している。これらの手法は、多数の著者や多様なテキスト・タイプの中での文体相互の類似性や、個々の文体を最も特徴付ける文体的特徴を特定するのには有効である。しかし文体を記述的に明らかにする場合には、それぞれの文体的特徴の内訳を詳しく観察する必要がある。

以上を踏まえて本研究では、中学と高校の地学教科書を事例として、特に学年の進行に伴う文体的特徴の変化を明らかにすることを目的とする。そのために、（１）トークン比（TTR）、（２）平均文長、（３）接続詞の数、（４）指示語の数、という４つの文体的特徴に着目して、中学・高校の４つの教科書の文体を比較する。

本稿の構成は以下の通りである。第２節では文体の捉え方と文体的特徴の種類を整理した上で、本研究で調査する文体的特徴について述べる。第３節ではデータと分析手続きを説明する。第４節では、それぞれの文体的特徴に関する比較分析の結果を示す。最後に第５節では得られた結果をまとめると共に、今後の研究の方向性について述べる。

２．分析枠組み

２.１　文体の捉え方と文体的特徴の種類

文体の捉え方は研究ごとに大きく異なる。例えば著者推定や真贋判定では文体を書き手の特性と捉えているし、新聞や週刊誌などの文体比較ではそれぞれのテキスト・タイプの特性と捉えている。しかしこれらの多様な文体観も、文章の表現上の性格を他と対比的に捉えた特殊性を問題にしている点では共通している（中村, 2011）。

また、文体を生み出す原因である文体的特徴にも様々なものがあるが（cf. 石田ほか, 2004; 金・村上, 2003）、それらは以下のようにまとめられる。

† asaishi@p.u-tokyo.ac.jp

１．頻度：（例）比喩表現の数、疑問文の数、色彩語の数
２．比率：（例）品詞別の比率、語種別の比率、単文の／重文／複文の比率
３．長さ：（例）単語長、文長、段落長
４．構文特性：（例）係り受けの距離
５．分布特性：（例）トークン比、Yule の K、Simpson の D

２.２　本研究で調査する文体的特徴

本研究では、学年の進行に伴い大きく変化する文体的特徴として、トークン比（TTR）、平均文長、接続詞の数、指示語の数を調査する。これらは特に、学年の進行に伴う内容の複雑さを反映した文体的特徴である。一般に理数教育では、学年の進行に伴い内容は複雑になる。教科書の内容を、個々の概念が教科として体系化されたものと捉えたとき、内容の複雑さには（１）多様な概念が含まれていること、（２）それらの多くが関連付けられていること、という２つの側面が考えられる[1]。

語彙のトークン比（TTR）は、これらのうち前者の側面、すなわち概念の多様性を反映した文体的特徴である。TTR は以下の式で定義される。

$$TTR = \frac{V(N)}{N}$$

ただし、N は延べ語数、V(N)は異なり語数を表す。TTR は語彙の豊富さの指標であり（金, 2009）、ここでは TTR が高いほど（語彙が豊富なほど）、それらによってより多様な概念集合が表現されていると仮定する。

一方で平均文長は後者の側面、すなわち関連付けられる概念の多さを反映した文体的特徴である。文長は一文中に含まれる単語の数であり、文長が長いほど多くの概念が何らかの形で関連付けられていると仮定する。例えば、以下の２つの文

「地層が堆積した年代（地質年代）は、古いものから古生代、中生代、新生代に分けられている。」

「地球の歴史は地層の層序にもとづいて、先カンブリア時代と顕生代に分かれ、顕生代はさらに古生代（5 億 4300 万年前ー2 億 22100 万年前）、中生代（2 億 5100 万年前ー6500 万年前）、新生代（6500 万年前ー現在）に 3 分される。」

を比較すると、後者の方が分類や言い換えを通じて多くの概念が関連付けられている。

接続詞と指示語の数も、後者の側面を反映した文体的特徴である。これは、概念同士は文内だけでなく、文同士の連接関係を通しても関連付けられるためである。市川（1978）は文をつなぐ形式として（１）前後の文相互を直接・論理的につなぐ形式、（２）前文の内容を後文に持ち込んで前後の内容をつなぐ形式、（３）その他の形式（１と２の中間的なもの）、の３つに類型化している。各々はさらにいくつかに細分類されるが、接続詞の使用は１の典型であり、指示語の使用は２の典型である。これらのうち接続詞は、市川（1978）に従うと以下のように分類される。

１．順接：（例）だから、それで、したがって、
２．逆接：（例）けれども、しかし、だが
３．添加：（例）かつ、また、それに、および
４．対比：（例）一方、あるいは、または
５．転換：（例）さて、ところで、では
６．同列：（例）例えば、すなわち、つまり
７．補足：（例）なぜなら、ただし、なお、ちなみに

一方で指示語は指示対象により以下のように分類される（馬場, 2011）。

１．事物：これ、それ、あれ、どれ

[1] 従ってここでは、実験手順の複雑さや問題解法の複雑さなどは考慮しない。

２． 場所：ここ、そこ、あそこ、どこ
３． 方向：こちら、そちら、あちら、どちら
４． 態様：こう、こんなに、そう、そんなに、ああ、あんなに、どう、どんなに
５． 名詞修飾：この、こんな、その、そんな、あの、あんな、どの、どんな

３．データと分析手続き

３.１　データ

以下の４つの文部科学省検定済み教科書を分析対象とした。

・ 中学 理科（第２分野）上巻
・ 中学 理科（第２分野）下巻
・ 高校 地学Ⅰ
・ 高校 地学Ⅱ

中学の教科書には東京書籍、高校の地学教科書には啓林館のものを用いた。それぞれの教科書の本文部分を抽出してデータとしたが[2]、中学の理科教科書には地学分野を扱った単元のほかに生物分野を扱った単元も含まれているため、生物分野の単元はデータから除外した。表１に、教科書別のデータ量を示す。

表１　データの量

	段落数	文数	延べ語数	異なり語数	文字数
理科（２分野）上	41	129	3391	589	5638
理科（２分野）下	91	231	6066	789	9709
地学Ⅰ	442	1395	35435	3460	59899
地学Ⅱ	446	1550	41179	4086	70845

３.２　分析手続き

各教科書のデータを ChaSen で形態素解析した後[3]、４つの文体的特徴を以下の手順で比較した。まず、TTR については、以下の手順で比較した。

１．データの延べ語数と異なり語数をもとに TTR を算出する。ただし、低頻度事象が事象全体の大部分を占める場合、出現頻度分布に基づく殆どの要約統計量は標本量に応じて値が系統的に変化する（Tweedie & Baayen, 1998; 影浦, 2000）ため、そのままでは教科書同士を比較できない。そこで、100 語おきに（つまり、N=100, 200, 300..と変えていく）ランダムサンプルを抽出して TTR を算出する試行を 1000 回繰り返し、それぞれの N での平均により比較する。

２．一定の延べ語数（ここでは 1000 語とする）中の各品詞の異なり語数を比較する。

文長については、以下の手順で比較した。

１．句点（。？！のいずれか）を区切りとして文を抽出し、一文中に含まれる単語数の平均を比較する。

２．文を構成している各品詞の頻度を比較する。

接続詞と指示語については、以下の手順で比較した。

１．接続詞については、ChaSen で接続詞と判定された形態素を抽出して、前述した７種類の比率を比較する。

２．指示語については、コ系、ソ系、ア系、ド系の指示語を全て抽出して、前述した５種類の比率を比較する。

それぞれの文体的特徴について、まず中学と高校の平均で比較し、さらに４教科書全てで比較することで、学年の進行に伴う文体的特徴の変化を観察する。

---

[2] 「太陽運動」のような複合名詞は一単語とした。また、独立した数式は除外した。

[3] ChaSen で「未知語」と判定されてしまう固有名詞などは、手作業で ipadic に追加した。

４．分析結果

４.１　TTR

図１に、延べ語数 N を 100 語から 2000 語に調整した時の TTR の値を示す[4]。図１から、いずれの N でみても中学より高校の TTR が高いことが分かる。また、４つの教科書を比較すると、中学では「理科（２分野）上巻」よりも「理科（２分野）下巻」の TTR が高い場合が多く、高校では特に N=500 までは「地学Ⅱ」より「地学Ⅰ」の TTR が高い。しかし、中学・高校内での差は中学・高校間の差と比べて非常に小さい。これらの結果は以下のようにまとめられる（［］の中は差が小さいことを示す）。

［理科（２分野）上 ＜ 理科（２分野）下］＜［地学Ⅱ ＜ 地学Ⅰ］

これは、高校教科書の方が中学教科書より文章中に多様な語彙が含まれており、多様な概念集合が表されていることを示している。

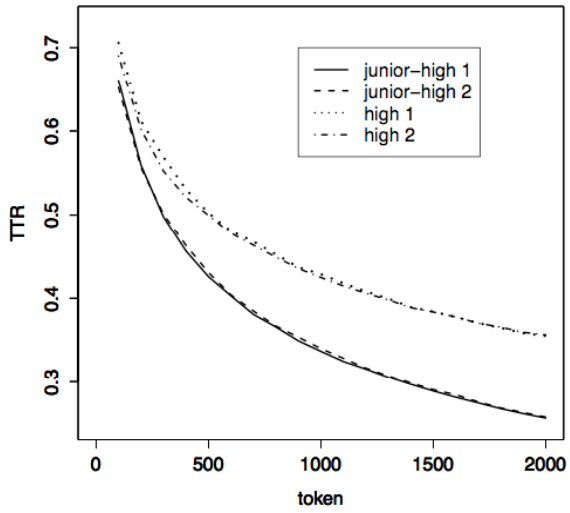


図１　TTR

N=1000 における異なり語の品詞別内訳を表２に示す[5]。表２から、いずれの教科書でも異なり語の多くは名詞であることが分かる。中学と高校の平均を比較すると、中学から高校にかけて名詞の異なり語数は大きく増加するのに対して、動詞や形容詞など他の品詞は殆ど変わらないかむしろ減少している。増加する名詞には以下のようなものが含まれていた。（１）「楕円銀河」「棒渦巻き銀河」「不規則銀河」のような複合名詞。これらの多くは、中学で説明された概念の下位概念を表している。例えば上の３つの複合名詞は、中学教科書で説明される「銀河」の下位概念を表している。これは、対象への認識の解像度の高まりを反映していると考えられる。（２）「1013」「760」「1946」のような数字や「hPa」「mmHg」のような助数詞。これは、中学での「大きい」「長い」のような形容詞が定量的な表現に置き換わるためであり、対象を認識する際の客観性の高まりを反映していると考えられる。（３）「磐梯山」「神奈川県」「アラビア海」のような固有名詞。これは、中学より高校の方が、抽象的な概念を多くの具体例と共に説明するようになるためと考えら

---

[4] 図１の凡例で、junior-high 1, junior-high 2, high 1, high 2 は、それぞれ「理科（２分野）上」、「理科（２分野）下」、「地学Ⅰ」、「地学Ⅱ」を表す。

[5] 他に、語種や文字種による分析も可能だが、ここでは最も解釈が容易な品詞に着目する。

れる。また、地名よりは少ないが、「チャンドラセカール」「オッペンハイマー」のような人名の固有名詞も増加しており、これは中学から高校にかけて科学史的内容が増えているためと考えられる。

また、４つの教科書を比較すると、学年の進行に伴い名詞の異なり数が一貫して増加していることが分かる。中学の「理科（２分野）上巻」より「理科（２分野）下巻」の名詞の異なり数が多いのは、特に数字と助数詞の増加が大きく、これは下巻の単元「天気とその変化」「地球と宇宙」で、気象や天体の状態・運動を定量的に説明しているためと考えられる。一方で「地学Ⅰ」より「地学Ⅱ」の名詞の異なり数が多いのは、特に地域を表す固有名詞の増加が大きい。これは地学Ⅰで基本的な概念を説明してから、地学Ⅱでは世界や日本の具体的な地形・地質を説明しているためと考えられる。

表２　N=1000 に含まれる異なり語数（主要な品詞のみ）

	名詞	動詞	形容詞	副詞	接続詞	助詞	助動詞	記号
理科（２分野）上	166.9	60.1	17.0	8.3	4.9	24.2	6.4	5.5
理科（２分野）下	183.4	53.1	12.4	7.6	3.7	23.3	6.7	5.1
中学平均	175.2	56.6	14.7	8.0	4.3	23.8	6.6	5.3
地学Ⅰ	265.3	53.0	13.6	8.3	4.4	23.2	5.7	9.6
地学Ⅱ	271.8	48.5	12.5	7.6	4.1	23.2	5.4	11.7
高校平均	268.6	50.8	13.1	8.0	4.3	23.2	5.6	10.7

４.２　平均文長

表３に各教科書の平均文長を示す。表３から、高校より中学の教科書の平均文長が長いことが分かる。また４つの教科書を比較すると、「理科（２分野）上巻」から「地学Ⅰ」までは学年の進行に伴い平均文長は短くなるが、「地学Ⅱ」の平均文長は４つの中で最も長い。これらは以下のようにまとめられる。

地学Ⅰ ＜理科（２分野）下 ＜ 理科（２分野）上 ＜ 地学Ⅱ

これは、中学よりも高校の教科書の方が、文章が要約的になるためと考えられる。例えば、中学の教科書での「同じくらいの大きさ」や「まばらに含まれる」などの表現は、高校では「同規模」「点在する」のような端的な表現に置き換えられる。また、指示語を使い前文の句や節の反復を避けることでも文長は短くなる。一方で「地学Ⅱ」の平均文長が最も大きいのは、これらの影響以上に、例示、引用、比較などを通じて多くの概念同士が関連付けられているためと考えられる。

表３　平均文長

	平均文長
理科（２分野）上	26.29
理科（２分野）下	26.26
中学平均	26.28
地学Ⅰ	25.47
地学Ⅱ	26.64
高校平均	26.06

一文中に出現する単語の内訳を表４に示す。表４から、いずれの教科書でも名詞または助詞の出現頻度が最も大きく、動詞または記号と続くことが分かる。中学と高校の平均を比べると、名詞と助動詞のみ高校が多いが、その他の殆どの品詞は中学が多い。

表４　一文中に含まれる単語数（主な品詞のみ）

	名詞	動詞	形容詞	副詞	接続詞	助詞	助動詞	記号
理科（２分野）上	7.78	3.95	0.78	0.24	0.19	8.09	1.08	3.60
理科（２分野）下	8.69	3.38	0.55	0.23	0.16	8.07	1.17	3.39
中学平均	8.24	3.67	0.67	0.24	0.18	8.08	1.13	3.50
地学Ⅰ	8.80	3.29	0.48	0.24	0.16	7.59	1.28	2.90
地学Ⅱ	9.24	3.31	0.44	0.23	0.15	8.08	1.40	3.11
高校平均	9.02	3.30	0.46	0.24	0.16	7.84	1.34	3.01

特に出現頻度の大きい名詞、助詞、記号について、中学から高校にかけての変化分を観察した結果、以下のことが分かった。（１）名詞は固有名詞やサ変名詞などいずれの下位分類でも増加している。（２）並立助詞や読点は中学の方が多い。これは、中学の方が「北上高地と阿武隈高地」「アークトゥルスやシリウスのような」「水星、金星、火星、土星」のように例示や分類が多いためと考えられる。（３）連体助詞「の」は高校の方が多い。これは、高校の方が「２つのプレート」「イギリスのハーシェル」のような連体修飾語として、個々の概念が詳細に定義されているためと考えられる。（４）括弧記号は高校の方が多い。これは高校の方が「古生代前半（約５億年前)」「二酸化炭素（CO2)」のような言い換えが多いためと考えられる。

また、４つの教科書を比較すると、学年の進行に伴い名詞の数は一貫して増加しており、助詞や記号の中では連体助詞「の」が一貫して増加し、読点の数は一貫して減少していた。これらは、学年の進行とともに例示や分類による説明は段階的に少なくなる一方で、個々の概念はより詳細に定義されるようになるためと考えられる。

最後に、名詞、助詞、記号以外のいくつかの品詞について検討する。動詞や形容詞の数が学年の進行に伴い減少するのは、「斜面が崩れる」のように動詞を含む表現が「斜面崩壊」のように名詞化したり、「大きい」「高い」のような形容詞が「0.03mm の」「1600 万 K の」のような連体修飾語になったり「高温」「巨大化」のように名詞化するためと考えられる。また、接続詞の数は学年の進行とともに減少する。これは、文同士を論理的につなぐ表現が少なくなるわけではなく、接続詞が担う機能が他の接続表現（接続助詞など）や一部の文末表現（「ーが原因である。」「ーからである。」）など、多様な表現で担われるようになるためと考えられる。

４.３　接続詞の数

接続詞の内訳を表５に示す。表５から、最も多い接続詞はいずれの教科書でも「また」「それに」など添加の接続詞であることが分かる。また、中学と高校の平均を比較すると、順接、逆接、添加、補足の比率は高校が大きく、対比、転換、同列の比率は中学が大きい。これらのうち、順接と逆接は一般に２つの事柄を論理的に結びつける接続表現であることから、中学から高校にかけては接続詞の中でもとりわけ文同士を論理的につなぐ接続詞が増えている。このことは、中学よりも高校の方が、現象の記述に留まらず、その発生原因から結果に至るまでのメカニズムを分析して説明しているためと考えられる。補足の接続詞は高校では少数ながら使われているが、中学では全く使われていない。これらには「なお」「ただし」が含まれていた。

逆に転換の接続詞は、中学では約５％を占めているが、高校では全く使われていない。これらは前の内容と別個の内容を導く接続表現であり、高校では前文と別な内容を示す文とをつなぐ接続詞は使われない（何らかのつながりを持たせる接続詞を使用する）ことを示している。また、中学の同列の接続詞はその全てが「例えば」であり、対比の接続詞の多くは「一方」である。これらは、中学では高校よりも例示や対照による説明が多いためと考えられる。

さらに４つの教科書を比較すると、いずれの接続詞でも学年の進行に伴う一貫した変化は見られないが、殆どの接続詞は、中学・高校の内部だけで見れば学年の進行に伴い増加・低下のいずれかで変化している。つまり、順接と逆接の接続詞の比率は「理科（２分野）上巻」より「理科（２分野）下巻」が大きく、「地学Ⅰ」より「地学Ⅱ」が大きい。逆に添加、対比、同列の接続詞の比率は「理科（２分野）下巻」より「理科（２分野）上巻」が大きく、「地学Ⅱ」より「地学Ⅰ」が大きい。これらの結果は、中学・高校の内部でも、前半では例示や対照を重視しているが、後半では個別の事例に適用できる法則を導くために分析を重視しているためと考えられる。

表５　接続詞の内訳

	順接	逆接	添加	対比	転換	同列	補足
理科（２分野）上	1（4.3%）	2（8.0%）	10（43.0%）	4（17.0%）	1（4.0%）	5（21.0%）	0（0.0%）
理科（２分野）下	6（15.8%）	7（18.0%）	14（36.0%）	6（15.0%）	3（7.0%）	2（5.0%）	0（0.0%）
中学平均	3.5（10.1%）	4.5（13.0%）	12.0（39.5%）	5.0（16.0%）	2.0（5.5%）	3.5（13.0%）	0.0（0.0%）
地学Ⅰ	21（9.5%）	30（13.0%）	97（43.0%）	44（19.0%）	0（0.0%）	27（12.0%）	3（1.0%）
地学Ⅱ	43（18.6%）	36（15.0%）	95（41.0%）	30（12.0%）	0（0.0%）	24（10.0%）	3（1.0%）
高校平均	32.0（14.1%）	33.0（14.0%）	96.0（42.0%）	37.0（15.5%）	0.0（0.0%）	25.5（11.0%）	3.0（1.0%）

４.４　指示語の数

指示語の内訳を表６に示す[6]。表６から、いずれの教科書でも指示語全体の約７割は「この」「その」などの名詞修飾であること、方向と態様の指示語は殆ど使われていないことが分かる。また、場所の指示語はいずれの教科書でも「ここでは・・について説明する」のように章や節全体を指示するものが含まれていた。

中学と高校の平均を比較すると、事物と態様の比率は高校が大きいが、場所、方向、名詞修飾の比率は中学が大きい。また、これらの中では事物と名詞修飾の変化が大きく、事物の比率は中学から高校にかけて増加し、逆に名詞修飾の比率は減少している。中学の教科書における名詞修飾は、「このような・・を」「その例として・・」といった例示表現の一部であることが多い。これは、中学で「例えば」を含む同列の接続詞の比率が高いという結果と整合的である。一方、高校で事物の指示語が増えるのは、「それを・・という。」のような定義文で使われる以外に「これは・・ためである。」「これにより・・」のような因果関係を示す表現の一部で使われている。これらは現象のメカニズムを分析している部分と考えられ、順接・逆接の接続詞の比率が高校で高いという結果と整合的である。また、名詞修飾についても例示表現以外に「このように」「その結果」など、まとめの表現で使われると考えられる。

なお指示語については４つ全ての教科書を比較すると、名詞修飾の比率が学年の進行に伴い減少している以外は、一貫した傾向は観察されない。

表６　指示語の内訳

	事物	場所	方向	態様	名詞修飾
理科（２分野）上	6（8.6%）	5（7.1%）	0（0.0%）	0（0.0%）	51（72.9%）
理科（２分野）下	14（15.2%）	1（1.1%）	1（1.1%）	0（0.0%）	64（69.6%）
中学平均	10.0（11.9%）	3.0（4.1%）	0.5（0.6%）	0.0（0.0%）	57.5（71.3%）
地学Ⅰ	113（20.3%）	15（2.7%）	4（0.7%）	2（0.4%）	375（67.3%）
地学Ⅱ	122（18.3%）	31（4.7%）	1（0.2%）	4（0.6%）	443（66.6%）
高校平均	117.5（19.3%）	23.0（3.7%）	2.5（0.5%）	3.0（0.5%）	409.0（67.0%）

6 なお、指示語は殆どがコ系かソ系であり、ア系は殆ど見られなかった。

５．おわりに

５.１　まとめ

本研究では、学年の進行に伴う文体的特徴の変化を明らかにすることを目的として、中学・高校の地学教科書の４つの文体的特徴を比較した。主な結果は以下の通りである。

１．TTR は中学より高校で大きく、中学・高校内での差は非常に小さい。異なり語の内訳を観察したときに最も多く含まれていたのは名詞であり、名詞の異なり語数は学年の進行に伴い一貫して増加する。増加分の名詞には複合名詞、固有名詞、数字・助数詞が多く含まれている。

２．平均文長は高校よりも中学で長いが、４つの教科書を比較すると「地学Ⅱ」が最も長い。文中に含まれる単語の内訳をみると名詞と助詞が最も多く、名詞の数は学年の進行に伴い一貫して増加する。さらに各品詞の内訳を見ると、連体助詞の数が学年の進行に伴い一貫して増加するのに対し、読点の数は減少している。

３．接続詞の内訳を見ると、順接、逆接、添加、補足の比率は高校で大きく、対比、転換、同列の比率は中学で大きい。４つの教科書で比較すると一貫した変化は観察できないが、順接と逆接の比率は中学・高校の前半で使われる教科書で大きく、添加、対比、同列の比率は後半で使われる教科書で大きい。

４．指示語の内訳を見ると、事物と態様の指示語の比率は高校で大きく、場所、方向、名詞修飾の比率は中学で大きい。特に中学から高校にかけて事物の比率が増加して名詞修飾の比率が減少しており、４つの教科書で比較しても学年の進行に伴い名詞修飾の比率は一貫して減少する。

これらは、学年の進行に伴い内容が複雑化するのに応じて、より多様な概念が説明されると共に、個々の概念は詳細に定義され、分析的な説明が重視されるためと考えられる。

５.２　今後の課題

最後に、今後の研究の方向性について述べる。まず、本研究の結果を一般化するためには、他科目（高校の物理、化学、生物、および中学の理科（１分野））についても同様の分析を行い、本研究の結果と比較する必要がある。また、学年だけではなく、科目や時代が異なる教科書の文体的特徴を比較することで、内容の一貫性や論理性などを反映した文体の側面を明らかできると考える。最後に、本研究では教科書の文体への影響要因として教科内容を想定しているが、想定される読者や著者による影響についても今後検討していきたいと考えている。

文　献

石田栄美、安形輝、野末道子ほか(2004)「文体からみた学術的文献の特徴分析」三田図書館・情報学会研究大会発表論文集、pp.33-36.

市川孝(1978)『国語教育のための文章論概説』教育出版.

影浦峡(2000)『計量情報学』丸善.

金明哲(2009)「計量文献学」『計量国語学事典』、p.238-248、朝倉書店.

金明哲、村上征勝(2003)「文章の統計分析とは」『言語と真理の統計：ことばと行動の確率モデルによる分析』、pp.3-57、岩波書店.

陳志文(2005)「新聞、週刊誌、高校教科書に見られる文体の類型と特性—主成分分析法を通して—」日本語文法、5 巻 1 号.

中村明(2011)「文体」『日本語文章・文体・表現事典』、p.126、朝倉書店.

馬場俊臣(2011)「指示表現」『日本語文章・文体・表現事典』、pp.104-105、朝倉書店.

安本美典、本田正久(1981)『因子分析法』培風館.

Tweedie. F.J. & Baayen R.H. (1998). How variable may a constant be?: Measures of lexical richness in perspective. *Computers and the Humanities*, Vol. 32, No. 5, pp.323-352.

web 公開予定文法用例検索システム「日本語文法項目用例文データベース『はごろも』」のレベル付けと学習者コーパスの比較

堀 恵子（東洋大学文学部）
江田すみれ（日本女子大学文学部）

Comparison of the Levels of Grammatical Items between Leaners' Corpus and "HAGOROMO", a Web-based Searching System of Grammatical Items

Keiko Hori (Faculty of Literature, Toyo University)
Sumire Goda (Faculty of Humanities, Japan Women's University)

１．はじめに

筆者らは，日本語教育の教師支援を目的として，文法項目の用例文を，複数のコーパスから抽出し，web 上で公開する「日本語文法項目用例文データベース『はごろも』」（以下，「はごろも」）を作成している。文法項目数は 1,884 項目である。文法項目には主観判定によって 6 段階のレベル付けを行った（堀ほか 2012）。今後はこの段階付けが妥当か検証する必要がある。本発表では，条件表現のバとタラを含む文法項目を取り上げ，学習者コーパスにおける文法項目の使用レベルと「はごろも」の主観判定によるレベルとを比較し，「はごろも」のレベル付けの妥当性を論じる。

２.「はごろも」の概要

２.１　研究の背景

日本語教育において，長くシラバスの拠り所となってきたのは，旧日本語能力試験の出題基準（以下，旧出題基準）である。1994 年に公開され，2002 年の改訂を経て，受験する学習者向けの授業だけでなく，教科書をはじめとした教材，種々の試験，研究における難易度の目安としても広く用いられてきた。しかし，日本語能力試験は 2010 年に改定され，出題基準は非公開となった。そこで，教育現場，特に海外の日本語教師の間には，何を教えるべきか不安が広がっていると言われている。

その旧出題基準の文法項目は，江田・小西（2008），堀ほか（2009），砂川ほか（2011）によると，3，4 級，1 級項目の頻度調査の結果，あまり使われない項目があると指摘されている。そのため，新日本語能力試験の出題基準が今後も公開されない場合に旧出題基準を使いつづけることは，学習者のコミュニケーション能力養成のために適切ではないと言えるであろう。

また近年，日本語教育における文法教育を見直そうという気運が高まっている。言語教育の目的は学習者のコミュニケーションを向上させることであるが，日本語学研究の成果に依拠しすぎて，日本語教育として必要でないものまで教えてきたのではないかと指摘されている（野田 2005）。しかしながら，ではどの項目が教えられるべきかについては，個々の項目について，どの用法が初級には不要かという提案は見られる（田中 2005 など）ものの，包括的な初級文法シラバスの提案には至っていない。

そこで，英語における van Ek and Trim（1991）の機能項目一覧のような，学習者のコミュニケーション能力を向上させることに役立つ文法項目一覧の作成が望まれる。しかし，現在学習者が多様化していることから，学習者のニーズも多様化しており，日本語教師が向きあう学習者にとって必要な項目も一様ではない。したがって現場に立つ教師自身が，文

法項目をよく理解し，取捨選択する目を持つことが望まれる。しかしながら，特に海外においては日本語の用例に充分に触れる機会が少ない場合もある。そこで，文法項目の用例や頻度情報，難易度の情報を与える項目一覧の作成が望まれる。

以上のことから，筆者らは主に日本語教師支援を目的として，初級から上級までの包括的な文法項目の一覧を作成し，複数のコーパスから用例を抽出し，海外においても利用しやすいように web 上で検索できるシステムを構築することとし, 2010 年から活動を始めた。

２.２　文法項目の選定

文法項目は，これまで広く用いられてきた文献の 2 編以上に取り上げられている項目を中心に選択した。用いた文献と選択の理由は下記の通りである。

A.「旧出題基準」: 初級から上級までの項目を網羅し，これまで多くの教育現場，研究で参照されてきた。

B.『日本語文型辞典(以下，文型辞典)』: 多くの文型，複合辞などを見出し語として採用し，意味用法を解説している。国内外の日本語教師にも広く参考にされている。

C.『現代語の助詞・助動詞(以下，助詞・助動詞)』: 助詞，助動詞を包括的に扱っている。

D.『日本語表現文型(以下，表現文型)』: 複合辞を包括的に扱っている。

E.『現代語複合辞用例集(以下，複合辞)』: 複合辞を包括的に扱い，複合辞のカテゴリー分けが合理的である。

以上の文献を参考にして，2012 年 3 月に 1,884 項目を選定し，名称を「はごろも」文法表とした。表 1 は，上記文献と，「はごろも」文法表との共通項目の数である。

表 1　参考文献と「はごろも」文法表との共通項目数（堀ほか　2012 より）

参考文献	旧出題基準	文型辞典	助詞・助動詞	表現文型	複合辞
共通項目数	881	1, 468	425	549	350

「はごろも」文法表では，1 つの形式に対し複数の用法，機能も取り上げる。例えば，「ている」は，旧出題基準では，「動作の継続」と「結果の状態」の例文だけが取り上げられていた。「はごろも」文法表では上記の文献を参考に，日本語教育において重要と思われる下記の 6 つを取り上げることにした。

(1)　動作作用の継続　　「私は今本をよんでいます。」
(2)　結果の状態　　「まどがしまっています。」
(3)　繰り返しの行為　　「ここでは，過去に何度も事故が起こっている」
(4)　経験　　「彼は 1 ヶ月前に会社を辞めている」
(5)　恒常的な状態　　「道が曲がっている」
(6)　完了　　「子供が大学に入るころには，父親はもう定年退職しているだろう」

２.３　用例の収集とデータベースの概要

用例は，話し言葉，書き言葉の複数のコーパスから抽出する。対象コーパスは下記の通りである。

⑴書き言葉

「日英新聞記事対応付けデータ（JENAAD)」／「ブログデータ（京都大学・NTT による)」京都大学情報学研究科--NTT コミュニケーション科学基礎研究所 共同研究ユニットによる　http://nlp.kuee.kyoto-u.ac.jp/kuntt/（以下，ブログ）／「白書」／「CASTEL/J CD-ROM V1.5 」日本語教育支援システム研究会（以下，CASTEL/J）／「日本語教科書」

⑵話し言葉

「日本語会話データベース」平成 8 - 10 年度文部省科学研究費補助特定領域研究「人文科学とコンピュータ」公募研究（「日本語会話データベースの構築と談話分析」 研究代表者　上村隆一）の成果による （以下,「上村コーパス」)／「宇都宮大学 パラ言語情報研

究向け音声対話データベース（UUDB)」／「名大会話コーパス」科学研究費基盤研究（B）（2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」(平成 13 年度～15 年度，研究代表者：大曾美恵子）／「BTS による多言語話し言葉日本語会話 1」宇佐美まゆみ監修(2005)『BTS による多言語話し言葉コーパス-日本語会話 1』東京外国語大学大学院地域文化研究科 21 世紀 COE プロジェクト「言語運用を基盤とする言語情報学拠点」

用例の抽出と，データベースの作成は，共同研究者とともに行った。まず，複数のコーパスから，形態素解析に基づく下処理を行った上で，文字列と品詞情報を組み合わせて用例の候補を抽出する。それを目視で精査して，用例文データベースを作成する。そして web 上の検索システムによって提供するようにする。

用例文の公開に先立ち，文法表と次節で述べる文法項目のレベルを付与した文解析システムをすでに公開している。テキストを入力すると，該当する文法項目と語彙項目について，その候補とレベルを表示する。結果は文法項目と語彙項目に分けて出力され，集計もできる。また，解析結果は CSV と HTML の 2 通りの方法で保存できる。

２.４　文法項目のレベル付け

文法項目のリストがあっても，難易度が分からなければ，どのレベルの学習者に教えるべきか分からない。そこで文法項目に対してレベル付けを行った（堀ほか 2012)。

付与するレベルは，日本語能力試験のような特定の試験に合わせるのではなく，初級，中級，上級という一般的に行われているレベル感をもとにし，さらにそれを細かくした「初級」[1]「初中級」「中級」「中上級」「上級」「超級」の 6 段階とした。CEFR が 6 段階であることも参考にしたが，CEFR の能力記述文に合わせてレベルを位置づけたのではなく，評定者である日本語教育経験者の判断によって 6 段階に分ける主観判定とした。

評定は，評定者に文法項目，文法カテゴリー，用法・機能，典型的な例文を記した一覧表を示し，他の評定者の結果が分からない状態で判定させた。評定者は日本語教育経験が 10 年以上の日本語教師 7 名である。調査期間は 2011 年 12 月から 2012 年 2 月 1 日までであった。

評定の結果，7 名のうち 5 名の評定について，平均値との間で中程度の一致が見られたため，5 名の評定結果を採用し，その平均をもってレベルを決定した。

評定の結果得られた各レベルの文法項目数は，表 2 の通りである。

表２　6 段階レベルごとの項目数

「初級」	「初中級」	「中級」	「中上級」	「上級」	「超級」
155	196	351	592	523	67

３．レベル付けの検証

今後は主観判定によって得られたレベルが妥当なものであるかどうかを検証していかなければならない。そのためには，母語話者の産出データについて，頻度情報，使用されるジャンル，レジスターの調査を行い，それとともに，外的基準となるテスト成績などのレベル付けがある学習者の産出データとの比較などの方法が考えられる。それぞれの項目についてこれらの調査を行うのは膨大な時間がかかることが予想される。

本発表では，その手始めとして，条件形式バとタラを含む項目について，口頭能力試験である OPI[2]の判定が付与されている学習者データを調査し，OPI 判定と「はごろも」文法表のレベル分けを比較する。

---

1 「はごろも」文法表に付与したレベル分けを，一般に行われている初級，中級，上級の概念と区別するため，本稿では「 」に入れて示す。

2 OPI(Oral Proficiency Interview)は，最長 30 分でテスターが与えた質問やタスクに受験者が応答し，それを録音してレベルを判定するという形で口頭能力を測るものである。(鎌田 2006)

日本語の条件形式は複数あり，それぞれが単独で条件文を作るが，各形式が複数の用法を持っている。また条件文以外にも，他の形式と結びついた複合辞や，文末との呼応のある文型として仮定，条件，反事実，前置きなどの意味を表す。そして，これらの項目はふつう初級後半から導入されるが，学習者の産出は中級のはじめから見られ，項目によっては上級になってようやく出現するものまで難易度は異なっている（堀 2005，2007）。

「はごろも」文法表においても，条件形式を含む項目は「初中級」から「超級」まで幅広く分布している。そこで，学習者データにおいてはどのように分布しているかを調査し，比較する対象として適切であると判断した。

３.１　調査対象項目について

「はごろも」文法表には条件形式バを含む項目は 57 項目，タラを含む項目は 36 項目ある[3]（表３）。

表３　「はごろも」文法表における条件形式バとタラを含む項目

	バを含む項目	タラを含む項目
「初中級」	ば+未実現，ば+条件，ば+問いかけ，	たら＋仮定，もし～たら，もしも～たら
「中級」	そういえば，とすれば，によれば，ば+くり返し・習慣，ば+意志・希望，ば～た/～ていた，ば～だろう，ばいい	たら＋勧誘・勧め，たら＋前置き，そうしたら，たら～た，たら～だろう，たらいい，たらだめ，たらどう，としたら，もしかしたら～か，もしかしたら～かもしれない
「中上級」	いちど～ば，かといえば，かとおもえば，からいえば，からすれば，からみれば，かりに～とすれば，かりに～ば，これいじょう～ば，さえ～ば，だとすれば，てからでなければ，てみれば，とすれば，なければ～た，なければ～ない，なぜかといえば～からだ，なぜならば～からだ，にしてみれば，にすれば，ば+勧誘・勧め，ば+立場・観点，ば+前置き，ば+働きかけ，ば～ところだ(った)，ば～のだが，ば～のに，ば～はずだ，ば～ほど，もあれば～もある，るところによれば，疑問詞+～ば～のか，数量詞+も～ば	いちど～たら，かとおもったら，からみたら，かりに～たら，かりに～としたら，たらどう～か，たらどんなに～か，たらよかった，ったら，といったら，とおもったら，となったら，にしたら，疑問詞+～たら～のか，数量詞+も～たら
「上級」	～も～ば～も～，かとおもえば～も，すればいいものを，とあれば，といえば，といえば～が，といえば～ぐらいのことだ，とおもえば，となれば，ともなれば，ばこそ，ばそれまでだ	(そう)だとしたら，たら～で，たら～ところだ，たら～はずだ，ったらない，といったらありはしない，といったらありゃしない，といったらない，とおもったら，ときたら，となったら，によったら，疑問詞+～たら～のか
「超級」	かとなれば	（項目なし）

３.２　KY コーパスにおける用例の抽出

KY コーパスとは，英語，中国語，韓国語の被験者の数がそれぞれ初級 5 名，中級 10 名，上級 10 名，超級 5 名の合計 30 名ずつの OPI データである(鎌田 2006)。上級は，現在は上・中・下の 3 つの下位レベルに分かれているが，収集された当時は上級と，上級の上の 2 つのみであった。そこで，全体で 9 の段階に分かれている。

[3] バ，タラを含む項目のうち，「なければならない」のように，条件と無関係の機能を持つものや，「いわば」のように独立した接続詞となっているものなどは除く。ただし，「といったら」のように前置き表現が話題を提示する機能を持つものは含めた。

KY コーパスのテキストから，表 3 に示したバ，タラを含む表現を抽出した。

３.３　結果

その結果，591 例が抽出された（表４）。自ら言い直しをしている場合は，あとの発言を採用した。下の例１は，下線部の後言い直しをしているため，使用例としては採用しない。

例１）それでちょっと，すいませんが，〈はい〉予約をキャンセルしていただければ，いただきたいんですが。(KA01)

誤用は，条件形式を使用していても，意味，機能が不適切な場合は，誤用とした。例 2 は，話題を提示する複合辞として使用しているが，後ろには話題に取り上げたものを説明する完全な形の文が来なければならない。しかし，この例は「花と言えば吉野」のような慣用表現に見られるように完全な形ではなく，すわりの悪い文となっている。そこで，誤用とした。

例２）Ｔさんの年齢とぴったりと思いますね，まあ，性格と言えば，うん，優しい男ですよ。(CA02)

表４　KY コーパスに見られる条件形式バとタラを含む使用数

	「初中級」（たら仮定，ば+未実現のことがら，ば条件）	「中級」（そういえば，たら～た，たら～だろう，ば+意志・希望，ば～た/～ていた，ばいい，ばくり返し・習慣，もしかしたら～か）	「中上級」（からいえば，からみれば，さえ～ば，てみれば，といえば，なければ～ない，なぜかといえば～からだ，ば+働きかけ，ば～ほど，ば前置き，ば立場・観点，疑問詞+～たら～のか，ば～のに）	「上級」（～も～ば～も～，といえば，といえば～が）	総計
IL	5	1	0	0	6
IM	27	12	1	0	40
IH	40	6	4	0	50
A	51	17	6	1	75
AH	116	58	36	4	214
S	141	40	21	4	206
総計	380	134	68	9	591

３.４　考察

表 4 を見ると，使用数は，総計においても，各レベルの使用数においてもおおむね OPI レベルが上がるにしたがって数が増えている。またその出現は，「初中級」「中級」の項目から，OPI レベルが上がるにしたがって「中上級」「上級」の項目へと広がりが見られ，使用数も増えている。

このことから，6 段階レベル付けは，少なくとも本稿の調査の範囲内では妥当であると言えよう。

今後は検証する項目を増やし，母語話者の産出データとの比較も行っていく。

４．まとめと今後の課題

本稿では，「はごろも」について紹介するとともに，文法項目に付与した 6 段階レベルの妥当性を検証する試みとして，条件形式バとタラを含む項目を取り上げて考察した。学習者のレベルが確定している KY コーパスにおける使用数は，「はごろも」文法表の「初中級」と「中級」レベルの項目から，「中上級」と「上級」の項目へと広がりが見られ，レベルが上がるにしたがって使用数が増えていた。その結果，本稿で取り上げた項目のうち使用が見られた項目については，レベル分けが妥当であると言えよう。

今後の課題として３点上げる。第１に，本稿で扱った項目は話し言葉だけでなく，書き言葉において使用される項目も含んでいる。したがって，今回使用が見られなかった項目

についても，直ちに使用が少ないと結論づけることはできない。今後は，話し言葉についてはより大きなコーパスを対象として調査する必要がある。また，書き言葉の調査も必要である。第２に，話し言葉のコーパスに関して，堀（2012）は，KY コーパスは口頭能力試験であることから，試験官が能力の上限を見極める目的で言語能力を超えた「突き上げ」と呼ばれる質問をするため，被験者が緊張を強いられることと，被験者が誤りを恐れて難しい項目を使用しない非用が見られる可能性とを指摘している。したがって，学習者の産出データとしては，口頭能力試験以外のデータがより望ましい。第３に，各文法項目について，母語話者データを調査して，母語話者の使用頻度と比較する必要がある。

今後は調査項目と調査対象を広げ，母語話者コーパスとの比較も行い，統計的に 6 段階のレベルの妥当性を検証していくことが課題である。

謝　辞

本研究は，文部科学省科学研究費補助金基盤研究（C）「 日本語教育のためのコーパスに基づく文法項目データベース構築と検索システムの公開」課題番号：22520538,（代表：堀恵子）の助成を受けている。

文　献

鎌田修(2006)「KY コーパスと日本語教育研究」『日本語教育』130, pp.42-51.日本語教育学会

グループジャマシイ(1998)『教師と学習者のための日本語文型辞典』くろしお出版.

江田すみれ、小西円（2008）「３種類のコーパスを用いた 3 級 4 級文法項目の使用頻度調査」『日本女子大学　紀要　文学部』57,　pp.1-28.

国際交流基金・日本国際教育協会(2002)『日本語能力試験出題基準【改訂版】』凡人社.

国立国語研究所(1951)『現代語の助詞・助動詞−用法と実例−』秀英出版.

国立国語研究所(2001)『現代語複合辞用例集』国立国語研究所.

砂川有里子、清水由貴子、奥川育子、千葉庄寿（2011）BCCWJ による機能語データベース（スタンドアロン版）（Ver. 0.9.1b）特定領域研究「日本語コーパス」研究成果報告書

田中真理（2005）「学習者の習得を考慮した日本語教育文法」野田尚史編（2005）『コミュニケーションのための日本語教育文法』pp.63-82.くろしお出版.

野田尚史編（2005）『コミュニケーションのための日本語教育文法』くろしお出版.

堀恵子(2005)「日本語条件表現の習得過程－中級学習者に対する縦断的インタビューから-」『日本語教育方法研究会誌』12:1，pp.36-37. 日本語教育方法研究会

堀恵子（2007）「日本語条件文の文末制約習得に及ぼす母語の影響－タイ語・英語・韓国語・中国語話者を対象とした文法性判断テストから－」『麗澤大学紀要』84，pp.101-126. 麗澤大学

堀恵子(2012)「習得過程研究における学習者コーパスの制約−OPI コーパスとインタビューの比較から−」『人間科学総合研究所紀要』14，pp.95-118.東洋大学人間科学総合研究所

堀恵子、荒川みどり、小池恵己子、小林佳代子(2009)「日本語能力試験出題基準の＜機能語＞を対象としたコーパス調査−目標言語使用領域での課題遂行に必要な項目を検証する−」『2009 年度日本語教育学会春季大会予稿集』pp.194-199.

堀恵子、李在鎬、砂川有里子、今井新悟、江田すみれ（2012）「文法項目の主観判定による 6 段階レベルづけとその応用」2012 年日本語教育国際研究大会ポスター発表

森田良行、松木正恵（1989）『日本語表現文型』アルク.

van Ek, J.A. & Trim, J. L. M. （1991）*Threshold 1990*. Cambridge: Cambridge University Press.

関連 URL

学習項目解析システム　http://lias.intersc.tsukuba.ac.jp/

日本語学習者の作文におけるエラータイプの自動分類へ向けて

大山 浩美　（奈良先端科学技術大学院大学 情報科学研究科）
小町 守　（奈良先端科学技術大学院大学 情報科学研究科）†
藤野 拓也　（奈良先端科学技術大学院大学 情報科学研究科）
松本 裕治　（奈良先端科学技術大学院大学 情報科学研究科）

Towards Automatic Error Type Classification on Japanese Language Learners' Writings

Hiromi Oyama, Mamoru Komachi, Takuya Fujino, Yuji Matsumoto
（Graduate School of Information Science, Nara Institute of Science and Technology）

1　はじめに

現在、様々な種類の学習者コーパスが収集され、学生へのフィードバック、教材研究、教授法の見直しへ役立てるための言語教育の調査研究に利用されている。学習者コーパスは、大学等の日本語教育機関などで収集されるのみならず、SNS を利用したウェブ上での言語学習者のための添削サービス（Lang-8[1]）もあり、膨大な数の学習者の作文を集めることが可能である。しかし、学習者が生み出した文は新聞や書籍などのコーパスと異なり正用だけでなく誤用も含まれており、そのままで分析をするのは難しく、調査分析のための前処理が必要となってくる。その処理には、半自動の誤り検出・訂正とエラータイプの分類、実際の人手によるエラータグ付与作業などが含まれる。

言語学習者にとってどの部分が誤用かを知ることは重要であるが、なぜ誤用なのか、どんな誤用なのか、という理由を知ることはさらに重要である。そのためエラータイプを分類するというタスクは学習者コーパス整備のための一つの重要なタスクであると考えられる。

現在のコーパスは、それぞれの機関でそれぞれの研究目的のために収集されているため、エラータグの設計方法は様々である。また、人手で付与されるため、エラータグつきのコーパスは少なく、タグ判定も不安定になりがちになる。エラータイプ自動分類は、人手でしか行えなかったエラー判定を一部自動化し、人手の負担、コストを軽減しようというものである。しかし、それのみならず、現在の様々なエラータグの再現性、妥当性を検証するのにも役立てる。そういった点が検証できれば、これからのエラータグ設計に役立てることもできる。

上記のような理由から、本研究は、大規模なエラータグつけコーパスがないことやエラータグの付け方がコーパスごとに異なるなどのエラータグ付与作業の困難さを一部支援することを目指し、誤用コーパスから機械的にエラータグ付与のための自動的な処理を行う手法を検討する。

具体的にはアノテータが修正した箇所のエラータイプを当てるタスクに取り組み、エラータグが付与された NAIST 誤用コーパス [17] のタグ付与基準に基づき、機械学習を用いた多クラス分類によるエラータイプ分類実験を行った。

†{hiromi-o,komachi,takuya-fu,matsu}@is.naist.jp
[1] http://www.lang-8.com

実験は、2 種類に分けられる。1 つ目は、NAIST 誤用コーパス内でトレーニングとテストを行い、どれくらいエラータイプを当てられるかというタスク（トレーニングデータ：NAIST 誤用コーパス、テストデータ：NAIST 誤用コーパス）と、もう 1 つは、まったく別のエラータグの付与されていない日本語学習者コーパス（Lang-8）（トレーニングデータ：NAIST 誤用コーパス、テストデータ：Lang-8）でどれだけ当てられるかというタスクである。

NAIST 誤用コーパス内で行った実験では、正解率は 77.1%であった。Lang-8 をテストデータとして行った実験では 83.3%であった。「動詞」や「表記」の誤りに比べて「形容詞」や「名詞」のエラータイプの自動分類が難しいことが分かり、エラータイプが正しく判定されなかった理由は、実験素性の選択の問題、形態素解析による問題、タグ範囲の設定による問題と分析した。

2　関連研究

現在、存在している日本語学習者コーパスは、大阪大学の寺村コーパス [18]（データ総数 4,601 文中 3,131 文がエラータグ付け済み）、名古屋大学の学習者コーパス [21]（756 ファイル）、東京外国語大学の「オンライン日本語誤用コーパス辞典」[2]（エラータグつき作文 40 ファイル）、筑波大学の「日本語学習者作文コーパス」[22][3]（540 ファイル）などがある。本研究もこれらの研究と同じく学習者コーパスにエラータイプを付与することを目的としているが、これらの研究がすべて人手でエラータイプを付与しているのに対し、我々は半自動化で付与することを目的としている点が異なる。

学習者コーパスにおける整備作業には、データ収集、タグ設計、誤り検出・訂正、エラータイプ分類が主に考えられ、それらに関する研究が進められている。

その中で誤り判定においては、英語のスペルミスの誤り訂正 ([13])、英語の名詞の可算性（数えられる名詞）、不可算性（数えられない名詞）の誤り検出・訂正研究 ([2], [23])、前置詞の誤り検出・訂正に関する研究 ([3], [4], [5], [12], [6])、冠詞誤り検出・訂正に関する研究 ([7], [5], [6], [14]) などがある。日本語を対象とする研究では、格助詞を対象とした研究が多い ([15], [9], [16],[24], [11])。さらに、エラータイプに特に着目せずに文を誤用文と正用文とに分類する研究もある ([10], [19])。

エラータイプ分類においては、 [1] が英語学習者の作文において実験を行っている。データは Cambridge Learner Corpus（CLC）[4]で、75 の母語数、2 千万語から成る。そのうち、5 百万語分にエラータグが付与されている [8]。多クラス分類モデルで、最大エントロピーを用い、15 クラスのエラータイプ（余剰、不足、変更、スペルミス、動詞の時制など）に分ける実験を行っている。本稿とは、学習者の書いた誤用文と正用文を多クラス分類モデルで分類するという点が共通している。しかし、日本語学習者の作文におけるエラータイプ分類実験はまだ見られない。

3　NAIST 誤用コーパスにおけるエラータグのアノテーション

本稿におけるエラータイプ分類実験は、我々が作成した NAIST 誤用コーパス [17] におけるエラータイプの自動分類に取り組んだので、まず NAIST 誤用コーパスの概要について述

[2] http://cblle.tufs.ac.jp/llc/ja_wrong/index.php?m=default
[3] http://www34.atwiki.jp/jccorpus/
[4] http://www.cambridge.org/elt/corpus/clc.htm

表 1: NAIST 誤用コーパスのエラータグ集計結果（上位 10 位まで）
VN=ベトナム/TH=タイ/CN=中国/ML=マレーシア/MN=モンゴル/KH=カンボジア/KR=韓国/SG=シンガポール

	VN	TH	CN	ML	MN	KH	KR	SG	全体的な割合
語彙選択	35.0	27.0	17.2	22.8	29.2	12.8	25.2	23.8	24.1
助詞	21.8	23.1	20.6	24.2	22.1	17.4	17.3	30.6	22.1
表記	9.8	10.1	19.8	16.9	12.7	33.6	15.5	6.8	15.7
動詞	13.8	15.3	16.8	12.1	14.2	15.9	14.6	10.2	14.1
成句	6.2	7.0	2.6	7.3	5.2	1.7	3.4	4.9	4.8
文体	1.7	1.2	2.3	6.0	4.1	6.1	3.1	6.3	3.9
名詞	2.5	2.6	3.5	1.4	3.4	2.0	4.4	2.9	2.8
文全体	2.0	2.6	1.2	3.4	0.7	1.4	2.4	2.4	2.0
形容詞	2.0	0.9	2.6	1.5	1.9	1.7	1.5	1.5	1.7
語順	1.0	1.3	1.2	0.3	0.4	1.2	0.6	0.0	0.8

べる[5]。「作文対訳 DB」の中で添削が施してある 313 名の作文中の誤用部分にタグを付与し、様々な情報を補完した [17]。ファイル数は 313、総文字数は 191,994 字となっている。本稿は「作文対訳 DB」に対しアノテーションを行なっているため、データは共通しているが、「作文対訳 DB」には誤用の種類がアノテーションされていない。

NAIST 誤用コーパスのエラータグの上位 10 位までを表 1 に示す[6]。誤用で多いのは、「語彙選択」、「助詞」、「動詞」、「表記」「成句」、「名詞」、「形容詞」に関する誤用である。「語彙選択」は、ふさわしい語彙を選べなかった（「国民」を「人民」）誤りである。「助詞」は、助詞が抜けている（不足）、不必要な助詞がつけられている（余剰）、誤った助詞が入れられている（変更）などがサブカテゴリーに含まれる。「動詞」カテゴリーの誤用には、サブカテゴリーとして動詞の活用、自動詞か他動詞か、受け身の誤り、テンスアスペクトの誤りなどが含まれる。「表記」には、ひらがな、かたかな、漢字に関する誤りなどが入る。「成句」は、きまったフレーズ（「〜たり〜たり」など）がうまく使えなかった誤りを含む。「名詞」には文を名詞化するときの「の」と「こと」を使い誤ったものや品詞選択誤りなど、「形容詞」には活用など、詳細に分類されている。「文全体」は、文がすべて書き換えられている誤り事例である。「文体」は、「です・ます体」に関する誤りで、「語順」は、語彙の出現する位置に関する誤りである。

4　多クラス分類によるエラータイプ分類

実験の概要を図 1 に示す。まず、正用文と誤用文が対になっており、エラータイプに分類されているコーパスを利用し、正用文と誤用文とを取り出す。上記の 2 種類の文から、対応する誤用箇所（x）と、正用箇所（y）、エラータイプ（t）の 3 つ組からなる事例（x,y,t）を取り出す。

それらを取る際に、正用文と誤用文において動的計画法によるマッチングを用いて置換対

[5]アノテーションについて詳しくは [17] を参照されたい

[6]その他の誤用には、「接続」「モダリティ」、「コロケーション」、「助動詞」、「指示詞」、「体言修飾」、「否定」、「副詞」、「仮定、条件」、「代名詞」、「その他　文法事項に関する誤り」などがある。詳しくは [17] を参照されたい

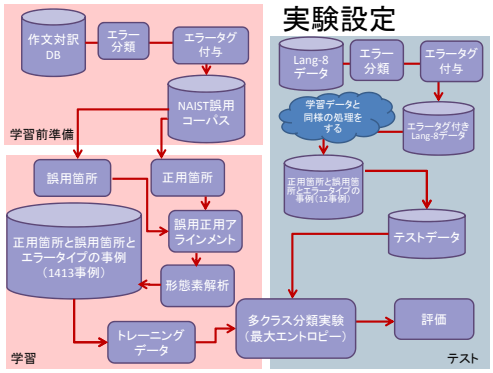


図 1: 実験概要

の抽出を行った。誤用文と正用文をそれぞれ文頭から 1 文字ずつ読み進めていき、誤用文には存在するが正用文には存在しない文字列を挿入箇所としてみなした。さらに、正用文には存在するが誤用文には存在しない文字列を削除箇所としてみなした。今回は挿入箇所と削除箇所が連続した部分を置換のペアとして抽出した。

例えば、「誰 に（誤）→ で（正） もたばこを吸う権利がある」という文から事例をとりだす。ラベルとして「助詞」の誤用（t）と、さらに、誤用箇所「に」（x）と、正用箇所「で」（y）という事例が 1,413 個取り出される。素性には、誤用箇所そのものに加えて誤用箇所周辺の文脈、正用箇所そのものに加えて正用箇所周辺の文脈、およびそれらの組み合わせなどを用いることができる。1 つの文に別々の誤用が 2 つ以上ある場合は、1 誤用につき 1 事例として取り出した。それらの事例をトレーニングし、多クラス分類器として機械学習の「最大エントロピー法[7]」を利用し、5 つの多クラス分類を試みた。

4.1 データ

NAIST 誤用コーパスから正用文と誤用文のペアを任意に取り出し、トレーニングデータとして使用した。テストデータは、NAIST 誤用コーパスから任意に取り出した文（実験 1）と Lang-8 から 12 文を取り出し、エラータイプに分類したもの（実験 2）である。Lang-8 はもともとエラータグ付与されていないため、任意に 12 文を取り出し、エラータイプを付与した。

実験に使うエラータイプは、表 1 の「形容詞」、「動詞」、「助詞」、「名詞」、「表記」のサブカテゴリーを含んだ大カテゴリー 5 つ に限定した。それぞれの例を表 2 に示す。

4.2 素性

素性は、誤用箇所の表層、正用箇所の表層、正用箇所の形態素解析結果を使用した。さらに、unidic 辞書を使った MeCab 形態素解析器[8] を利用し、形態素情報を素性として取り入れた [20]。上記の例文から、誤用箇所の「に」、それが対応する箇所の「で」、その形態素解析

[7] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html
[8] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

表 2: 誤用文例

エラータイプ	誤用文例
形容詞	発音が大変 __ だけではない（誤）→大変 な だけではない（正）
名詞	趣味は映画を見る の です（誤）→見る こと です（正）
表記	年ぱ の人（誤）→ 年配 の人（正）
助詞	英語 を わかる（誤）→英語 が わかる（正）
動詞	手紙を書 き ない（誤）→書 か ない（正）

結果である「助詞」が素性となる。

4.3 評価尺度

評価尺度として再現率、適合率、F 値、正解率を利用した。再現率は、対象とする各エラータイプの中で、正しく分類されたエラータイプを指し、適合率は、システムがあるエラータイプだと分類したもののうち正解を当てた率である。F 値はそれらの調和平均を表している。正解率とは、すべての事例の中で、正しいエラータイプに属すると判定されたもの (True Positives) とこの誤用文はエラータイプに属さないと正しく判定されたもの (True Negatives) との割合である。

$$\text{再現率} = \frac{\text{正しく分類された事例数}}{\text{各エラータイプの全事例数}} \times 100 \quad (1)$$

$$\text{適合率} = \frac{\text{正しく分類された事例数}}{\text{システムがあるエラータイプだと分類した事例数}} \times 100 \quad (2)$$

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

$$\text{正解率} = \frac{\text{正しく分類された事例数}}{\text{すべての事例}} \times 100 \quad (4)$$

5 NAIST 誤用コーパスに対するエラータイプ分類実験結果

まず、一つ目の実験、NAIST 誤用コーパス内でエラータイプ分類はどのくらいできるのかを検証するために分割交差検定（10 分割）を行った。正解率は、全体で 77.14%であった。各エラータイプにおける適合率、再現率、F 値を表 3 に示す。これを見ると、「表記の誤用」また、「動詞の誤用」が安定して分類されていることがわかる。最も分類されにくいのが「形容詞の誤用」である。

表 4 には、ある分割内での誤り事例を取り出し、エラー分析を行った結果を示している。システムが誤って分類した理由は、「素性の問題」、「プログラムミス」、「形態素解析による問題」、「タグつけ時のミス」、「タグ範囲における問題」などに分けられた。以下に数の多い 3 つの問題について詳しく述べる。

5.1 素性の問題

表記の間違いに多く見られたのは、現在の素性で分類するのが難しい問題であった。表記の間違いは一般的に書き間違い（spelling errors）を意味している。しかし、例文「結婚式は、

表 3: NAIST 誤用コーパスの 10 分割交差検定結果

エラータイプ	適合率	再現率	F 値
助詞	65.7%	64.0%	62.1%
動詞	76.0%	85.2%	80.3%
表記	84.3%	82.8%	83.5%
名詞	53.6%	36.2%	42.2%
形容詞	31.7%	24.1%	26.5%

主に結婚式場やホテルや おおきい ビルがあります」で見られるように、漢字で書くべきところを書いていない誤りも含んでいる。さらに、送り仮名（嫌らう（誤）→嫌う（正））や漢字ひらがな混合（かぶル（誤）→かぶる（正））などの事例も「表記」の誤用なのであるが、システムは「表記」だと正しく分類しなかった。今回のエラータイプ分類実験には、サブカテゴリーを含んだ大分類のタグを使用している。分類が大きければ大きいほど、エラータイプ分類の揺れがないと考えたためと、サブカテゴリーまで自動分類するのは難しいと考えたためである。しかし、例えば「表記」のように主に「書き間違い」があるグループに「送り仮名」の問題や「漢字ひらがな混合」の問題など、様々な問題が混在している場合、それぞれの事例数が足りないのが問題であることは十分に考えられる。

5.2 形態素解析による問題

今回の実験では、学習者の作文は誤りを含むため形態素解析が難しいと考え、誤用文の品詞情報は利用しなかった。そのため、例えば、「私が入学できた大学は学期の 始めの（誤）→ 始まる（正）前の 3 日間入学式を行った」という文において、「始まる - 動詞」の結果は利用できるが、「始めの」の形態素解析結果を素性に組み込まなかった。誤りが含まれていたとしても、「始め-名詞」という素性が利用できれば、「名詞」の誤りと分類される可能性はある。誤用文の形態素解析結果を利用すればこのような誤分類は解決できる可能性がある。

5.3 タグ範囲における問題

タグ付与作業を行う時、正用文と誤用文で、単語同士が 1 対 1 に一致しない場合、どこまでにタグを振るのかということが問題となる。例えば、「おいのりのあとで、家で私の家族はいっしょに朝食を食べ るものです（誤）→ます（正）」という文の場合、「食べ」までは正用文でも誤用文でも同じなのでそこにはタグを振らずに、異なる部分だけに振るとする。そうすると、エラータイプを分類するための問題となる箇所は「もの」であるが、その前後をどうするのかという問題がある。現在は、異なる部分をすべてタグで囲んでいる。そうしなければ変更されている箇所がわからない。しかし、そのすべてにタグを振ると誤用の理由となるべき箇所がわからなくなるジレンマがある。学習者は自由に誤用を犯す。教師によるその添削方法も様々である。そのため、エラータグは、単語間の対応になっていない場合が多い。エラータグを設定するときのエラーの範囲の決定の仕方には、議論の余地が十分にある。

表 4: NAIST 誤用コーパスでのエラー分析結果

	システムの問題	タグ範囲における問題	形態素解析による問題	プログラムミス	タグつけ時のミス	エラー全事例
表記	11	0	1	0	2	14
動詞	2	2	2	4	1	11
名詞	0	7	1	0	0	8
形容詞	0	1	2	1	1	5
助詞	2	0	0	0	0	2
計	15	10	6	5	4	40

6 Lang-8 に対するエラータイプ分類実験結果

Lang-8 をテストデータとして使用した場合、システムはテスト文 12 文において 10 文の分類に成功した。

表 5 は、テスト文中のどのような文においてシステムが分類に成功し、どんな文において分類ミスをしているかを表している。傾向は NAIST 誤用コーパスを用いた実験と同様で、「表記」、「助詞」は、すべての文で分類に成功しており、判定しやすい事が分かる。「形容詞」や「名詞」で分類に失敗している。テスト文 1 において、「形容詞」の誤用を「動詞」と判定している。文 1 のように、「形容詞」や「名詞」は、訂正されるときに「小さい—形容詞」「の—助詞」から「少し—副詞」「の—助詞」というように、別の品詞になってしまう。文 6 においても「名詞」を「動詞」と判定している。この際にも「その場合—名詞」から「そう— 副詞 なれ— 動詞-非自立可能ば— 助詞-接続助詞」となり、「表記」のようにほとんどが名詞から名詞への変更となる場合とは異なる。

今回は、素性として単語の表層と形態素解析結果しか使っていないため、形態素を超えた修正がないエラータイプのほうが分類されやすかったのではないかと考えられる。

7 おわりに

本稿では、エラータイプの自動分類に向けた実験を試みた。現在、収集されている様々な学習者コーパスのエラータグ付与支援ができるように NAIST 誤用コーパスとは別コーパスである Lang-8 でのテストを行った。結果、NAIST 誤用コーパスにおいても Lang-8 コーパスにおいても 80%前後の正確率を得ることができた。「助詞」、「動詞」、「表記」の誤用などは、事例数が多いことからか、十分な判定結果が観測された。しかし、「名詞」や「形容詞」などのエラータイプは分類するのが難しかった。理由として、タグの範囲が必ずしも単語対単語の 1 対 1 の対応をせず、形態素解析の情報が素性に取り込めない点などが考えられる。これは、今後の素性選択に活かしたい。

また、事例数が少ないことも自動分類を難しくした要因である。次回の実験ではトレーニング、テストともにデータを増やすつもりである。

表 5: Lang-8 のテスト文結果

結果	エラータイプ	システム出力結果	誤用文例
×	形容詞	動詞	私たちはうちに小さい時間いました（誤）→少しの時間（正）
○	表記	表記	こんな宴会・飲み会は日本語で話すたみにいいチャンスだ（誤）→ため（正）
○	表記	表記	ゆべ２３：００に友達が到着しました（誤）→ゆうべ（正）
○	表記	表記	座の手前にきれいなちゅーりっぷを見ました（誤）→チューリップ（正）
○	表記	表記	せのもんはなんですか。（誤）→せんもん（専門）（正）
×	名詞	動詞	その場合はいいと思う（誤）→そうなれば（正）
○	名詞	名詞	私のお気に入りの趣味は読み__です（誤）→読むこと（正）
○	助詞	助詞	今日最後までにできたらはいいと思います（誤）→最後まで（正）
○	助詞	助詞	もし面接のときこのような質問を出て来るかもしれません（誤）→（正）質問が
○	助詞	助詞	わけがなく、微笑みながら雪が降るのを眺めていました（誤）→わけもなく（正）
○	助詞	助詞	普通に学校はあまり日本語で喋らないです。（誤）→普通は（正）
○	動詞	動詞	今晩のパーティは本当の宴会と思よ（誤）→思います（正）

参考文献

[1] Swanson B. and Yamangil E. Correction detection and error type selection as an esl educational aid. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 357–361, 2012.

[2] C. Brockett, W.B. Dolan, and M. Gamon. Correting ESL Errors Using Phrasal SMT Techniques. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 249–256, 2006.

[3] M. Chodorow, J. Tetreault, and N-R. Han. Detection of grammatical errors involving prepositions. *Proceedings of the 4th ACL–SIGSEM Workshop on Prepositions*, pp. 45–50, 2007.

[4] R. De Felice and S.G. Pulman. Automatically acquiring models of prepositional use. *Proceedings of the 4th ACL–SIGSEM Workshop on Prepositions*, pp. 45–50, 2007.

[5] R. De Felice and S.G. Pulman. A classifier-based approach to preposition and determiner error correction in L2. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pp. 169–176, 2008.

[6] M. Gamon, J. Gao, C. Brockett, A. Klementiev, W.B. Dolan, D. Belenko, and L. Vanderwende. Using contextual speller techniques and language modelling for ESL error correction. *Proceedings of the 3rd International Joint Conference on Computational Linguistics (IJCNLP 2008)*, 2008.

[7] N. R. Han, M. Chodorow, and C. Leacock. Detection errors in english article usage by non-native speakers. *Natural Language Engineering*, Vol. 12(2), pp. 115–129, 2006.

[8] D. Nicholls. The Cambridge Learner Corpus–error coding and analysis for lexicography and ELT. In Archer et al., editor, *Proceedings of the Corpus Linguistics 2003 Conference(CL2003)*, pp. 572–581. 2003.

[9] H. Oyama, Y. Matsumoto, M.Asahara, and K.Sakata. Construction of an error information tagged corpus of japanese language learners and automatic error detection. *Proceedings of the Computer Assisted Language Instruction Consortium*, 2008.

[10] G. Sun, X. Liu, G. Cong, M. Zhou, Z. Xiong, J. Lee, and C.Y. Lin. Detecting Erroneous Sentences using Automatically Mined Sequential Patterns. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 81–88, 2007.

[11] H. Suzuki and K. Toutanova. Learning to Predict Case Makers in Japanese. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1049–1056, 2006.

[12] J. Tetreault and M. Chodorow. The Ups and Downs of Preposition Error Detection in ESL Writing. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pp. 865–872, 2008.

[13] A. Wilcox-O'Hearn, G. Hirst, and A. Budanitsky. Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. *Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)(Lecture Notes in Computer Science Vol.4919)*, pp. 605–616, 2008.

[14] X. Yi, J. Gao, and W.B. Dolan. A web based English proofing system for ESL users. *Proceedings of the 3rd International Joint Conference on Computational Linguistics (IJCNLP 2008)*, pp. 619–624, 2008.

[15] 大木環美, 大山浩美, 北内啓, 末永高志, 松本裕治. 非日本語母国話者の作成するシステム開発文書を対象とした助詞の誤用判定. 言語処理学会第 17 回年次大会, pp. 1047–1050, 2011.

[16] 今枝恒治, 河合敦夫, 石川裕司, 永田亮, 桝井文人. 日本語学習者の作文における格助詞の誤り検出と訂正. 情報処理学会研究報告 コンピュータと教育研究会報告, pp. 39–46, 2003.

[17] 大山浩美, 小町守, 松本裕治. 日本語学習者の作文における誤用タグつきコーパスの構築について—NAIST 誤用コーパスの開発—. 第一回テキストアノテーションワークショップ, 2012.

[18] 寺村秀夫. 外国人学習者の日本語誤用例集接続詞・副詞. 大阪大学：データベース版、国立国語研究所, 1990.

[19] 水本智也, 小町守, 松本裕治. 大規模添削コーパスを用いた統計的機械翻訳手法による日本語誤り訂正. 言語処理学会第 17 回年次大会, pp. 1095–1098, 2011.

[20] 藤野拓也, 水本智也, 小町守, 永田昌明, 松本裕治. 日本語学習者の作文の誤り訂正に向けた単語分割. 言語処理学会第 18 回年次大会, pp. 26–29, 2012.

[21] 大曽美恵子, 杉浦正利, 市川保子, 奥村学, 小森早江子, 白井英俊, 滝沢直宏, 外池俊幸. 日本語学習者の作文コーパス：電子化による共有資源化. 言語処理学会第 3 回年次大会論文集, pp. 131–145, 1997.

[22] 李在鎬, 林●（火偏に玄）情, 宮岡弥生, 柴崎秀子. 言語処理の技術を利用したタグ付き日本語学習者コーパスの構築. 2012 年度日本語教育学会春季大会予稿集, 2012.

[23] 永田亮, 若菜崇宏, 河合敦夫, 森広浩一郎, 桝井文人, 井須尚紀. 可算／不可算名詞の判定に基づいた英文の誤り検出. 電子情報通信学会論文誌 J89-D(8), pp. 1777–1790, 2006.

[24] 南保亮太, 乙武北斗, 荒木健治. 文節内の特徴を用いた日本語助詞誤りの自動検出・校正. 情報処理学会研究報告 自然言語処理研究報告, pp. 107–112, 2007.

会話分析方式への転記変換における データ間・個人間のゆれに関する分析

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

Towards Automatic Transformation into CA Transcript Conventions: Differences of Transcript Strategy between Data and between Transcribers

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

1. はじめに

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

既存のコーパスを共有するにあたって、転記方式の不統一や基本情報アノテーションの欠如といった問題がある。伝ほか (2012) は、本プロジェクトのメンバーが有する 10 数種のコーパスで用いられている転記方式を調査し、それらが CSJ 方式と会話分析方式に概ね大別できることを示した。土屋ほか (2012) は、CSJ 方式の韻律ラベルから会話分析方式の音調マーカーへの自動変換を試み、会話データ/転記者ごとに言語特徴を重視するか、音響特徴を重視するかという転記方略の違いが存在し、この方略の違いに対処する必要があると論じた。そこでは、CSJ 方式の言語・音響情報から会話分析方式の音調マーカーを予測するための多変量モデルを構築することで、音調マーカーの予測に貢献する言語・音響特徴を分析した。その結果、予測に貢献する言語・音響特徴が会話データ/転記者ごとに異なることが分かった。たとえば、あるデータ/転記者では、アクセント句の末尾単語や次末単語の品詞といった言語的な特徴が予測に強く貢献していたが、もう一方のデータ/転記者では、アクセント句末の F0 値などの音響的な特徴のほうが貢献していた。しかし、この研究では、2 人の転記者がそれぞれ 1 つずつのデータしか転記しておらず、上記の違いが会話データ自体の質的な違いに起因するのか、転記者の一般的な転記方略の違いに起因するのかがわからない。

本研究の目的は、この転記方略のゆれが、データ間・個人間でどのように現れるのかをより詳細に検討することである。そのため、音調マーカーを予測するための言語・音響特徴の貢献

† ttsuchiya@ninjal.ac.jp

度の違いが、会話データの特徴によって生じるのか、あるいは転記者による転記方略の違いによって生じるのかを明らかにする。具体的には、土屋ほか (2012) で用いた 2 人の転記者による各 1 つずつの会話データに加え、これら 2 つの会話データをもう 1 人の転記者によって転記したデータも比較することで、データ間および個人間による転記方略の違いを分析する。

2. 方法

2.1 談話資料

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

2.2 転記・アノテーション

2.2.1 CSJ 方式

CSJ 方式の転記テキストの例を図 1 に記す。この例には参考のために句末境界音調が記されているが、実際のアノテーションでは、これらは別ファイルとして用意され、時間情報などを利用し相互にリンクがとれる形で蓄積されている。

CSJ 方式では、X-JToBI (五十嵐ほか 2006) に基づく韻律情報が提供され、アクセント句の末尾に句末境界音調が付与される。本データでは、(1) 下降調（L%）に加え、複合境界音調として、(2) 単純な上昇調（L%H%）、(3) 上昇前に一定期間低ピッチが見られる上昇調（L%LH%）、(4) 上昇下降調（L%HL%）の 4 種類を区別した。下降調 L% は、複合境界音調が生じないアクセント句末に付与される音調であり、必ずしも明示的な下降が生じているわけではない。この点において、会話分析方式のピリオドとは若干異なる。また上昇調 L%H%、L%LH% も、疑問上昇調だけでなく強調上昇調なども含まれており、クエスチョンとは必ずしも一致しない。

2.2.2 会話分析方式

会話分析方式の転記テキストの例を図 2 に示す。会話分析方式の転記に使われる種々の転記シンボルのうち、本研究では、ピリオド（per）‘.’、クエスチョン（ques）‘?’、コンマ（com）‘,’、アンダーバー（ub）‘_’ の 4 つの音調マーカーに注目した。これらのマーカーはそれぞれ下降・上昇・継続・平坦の音調を表す。

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

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

CSJ 方式

```
281.7240 283.4033 B: 松下にじゃねえ（L%）松下だろう（H%）
283.4775 283.8650 B: 〈笑〉
284.3166 285.3986 C: あれ（L%）もともと一緒なの（L%）
285.5721 286.2149 B: もともと一緒（L%）
285.7004 286.1680 A: そうだよ（L%）
```

図 1　CSJ 方式の転記テキスト（括弧内の句末境界音調は別ファイル）

会話分析方式

```
B: 松下にじゃねえ，松下だろう::.
   (0.5)
C: あれ，もともと一緒なの?
A: そう[だよ.
B:     [もともと一緒.
```

図 2　会話分析方式の転記テキスト

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

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

2.3　言語・音響特徴

CSJ の言語・音響情報から会話分析方式の音調マーカーを予測する多変量モデルでは、分析対象アクセント句から以下の言語・韻律特徴を抽出し用いた[*1]。

■言語特徴

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

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

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

■音響特徴

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

[*1] 音響特徴として他の特徴も抽出したが、これらの特徴との相関が高いため用いなかった。

アクセント句の最大 F0（f0MaxAP）　アクセント句中の F0 の最大値（標準化得点）
句末単語の最大 F0（f0MaxWord）　末尾単語中の F0 の最大値（標準化得点）
アクセント句の最大パワー（pwrMaxAP）　アクセント句中のパワーの最大値（標準化得点）
句末単語の最大パワー（pwrMaxWord）　末尾単語中のパワーの最大値（標準化得点）
アクセント句の平均モーラ長（amdAP）　アクセント句の継続時間をモーラ数で除したもの（標準化得点）
最終抽出可能 F0 点の値（lastF0Val）　アクセント句中で最後に抽出できた F0 点の値（標準化得点）
最終抽出可能 F0 点の位置（lastF0Loc）　上記 F0 点の句末から計った時間（対数値）

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

■その他の特徴　以上に加え、アクセント句自体の位置に関する以下の特徴を用いた。
発話冒頭からの位置（loc）　発話単位中で先頭から何番目のアクセント句か（対数値）
発話末尾からの位置（revLoc）　発話単位中で末尾から何番目のアクセント句か（対数値）

2.4　分析手順

データ間・個人間での転記方略のゆれを図 3 のように比較した。まず、データ内・個人間の比較として、chiba0232 では X 氏と Z 氏による転記の音調マーカーを、chiba0432 では Y 氏と Z 氏による転記の音調マーカーをアクセント句単位で整列し比較した（実線矢印）。音調マーカーが付与されていないアクセント句は none のラベルを与えた。

次に、個人内・データ間の比較として、Z 氏によって転記された chiba0232 と chiba0432 の音調マーカーを比較した（破線矢印）。この比較のために、CSJ 方式の転記から会話分析方式の転記へ変換する多変量モデルを一方のデータを学習データとして構築し、他方のデータの音調マーカーの予測結果と人手による音調マーカーとを比較した。多変量モデルとしてランダムフォレスト法 (Breiman 2001) を用い、統計解析ソフト R 言語の randomForest パッケージを使ってモデルを構築した（mtry = 4 とした）。アクセント句単位でデータを分節化し、2.3 で述べた言語・音響特徴を説明変数に用いた。

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

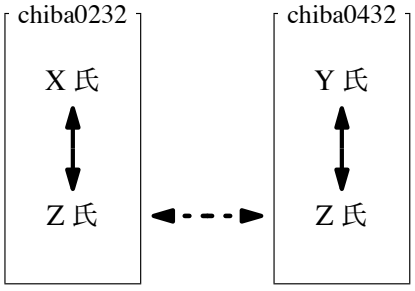


図 3　データ間・個人間比較

表 1　データ内・個人間のゆれ

chiba0232　（一致率 = 76.0%、κ = .66）

	Z 氏				
X 氏	none	per	ques	com	ub
none	**130**	12	4	2	1
per	9	**140**	15	0	0
ques	2	9	**58**	0	1
com	33	20	2	**26**	0
ub	0	1	1	0	**1**

chiba0432　（一致率 = 69.9%、κ = .58）

	Z 氏				
Y 氏	none	per	ques	com	ub
none	**184**	8	0	0	0
per	30	**126**	1	3	0
ques	5	11	**29**	1	0
com	92	14	1	**89**	5
ub	0	13	0	0	**0**

表 2　個人内・データ間のゆれ（転記者 = Z 氏）

学習 = chiba0432、比較 = chiba0232
（正解率 = 72.8%、κ = .58）

	観測値				
予測値	none	per	ques	com	ub
none	**134**	10	3	2	1
per	28	**164**	58	3	2
ques	0	3	**19**	0	0
com	12	5	0	**23**	0
ub	0	0	0	0	**0**

学習 = chiba0232、比較 = chiba0432
（正解率 = 74.8%、κ = .59）

	観測値				
予測値	none	per	ques	com	ub
none	**272**	22	6	35	5
per	34	**129**	13	12	0
ques	3	18	**12**	1	0
com	2	3	0	**45**	0
ub	0	0	0	0	**0**

3.　結果

3.1　データ内・個人間でのゆれ

chiba0232 における X 氏と Z 氏による音調マーカーの対応、および chiba0432 における Y 氏と Z 氏による音調マーカーの対応を表 1 に示す。全般的に、X 氏と Z 氏のほうが一致が高く（一致率 = 76.0%、κ = .66）、Y 氏と Z 氏では一致がより低かった（一致率 = 69.9%、κ = .58）。件数の少ないアンダーバー（ub）を別にすると、X 氏/Y 氏がコンマ（com）を付与している箇所で Z 氏がそうしていない例が多く見られた。また、chiba0432 では、Z 氏が音調マーカーを付与していない個所（none）に Y 氏がピリオド（per）やコンマ（com）を付与している例も多く見られた。

3.2　個人内・データ間でのゆれ

Z 氏による chiba0232 と chiba0432 の転記方略の違いを、多変量モデルによる予測結果と人手ラベルとの対応によって調べた。chiba0432 を学習データとして構築した多変量モデルによる chiba0232 の予測結果と Z 氏による人手ラベルとの対応、および chiba0232 を学習データとして構築した多変量モデルによる chiba0432 の予測結果と Z 氏による人手ラベルとの対応を表 2 に示す。正解率（一致率）はいずれも比較的高いが（chiba0232：72.8%、chiba0432：74.8%）、いずれの方向の予測でも、クエスチョン（ques）をピリオド（per）と誤って予測する例が多く、chiba0432 ではコンマ（com）をラベルなし（none）と誤って予測する例も多かった。

表 3　個人内・データ間のゆれ（転記者 = Z 氏）（句末境界音調ごと）

学習 = chiba0432、テスト = chiba0232
L%　（正解率 = 82.3%、κ = .67）

	観測値				
予測値	none	per	ques	com	ub
none	**118**	8	1	0	1
per	17	**109**	15	1	1
ques	0	0	**0**	0	0
com	5	0	0	**0**	0
ub	0	0	0	0	**0**

学習 = chiba0232、テスト = chiba0432
L%　（正解率 = 80.5%、κ = .59）

	観測値				
予測値	none	per	ques	com	ub
none	**254**	20	1	20	5
per	25	**96**	1	5	0
ques	1	7	**0**	0	0
com	0	0	0	**0**	0
ub	0	0	0	0	**0**

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

	観測値				
予測値	none	per	ques	com	ub
none	**15**	2	1	2	0
per	6	**41**	33	0	0
ques	0	1	**11**	0	0
com	4	0	0	**0**	0
ub	0	0	0	0	**0**

学習 = chiba0232、テスト = chiba0432
H%　（正解率 = 54.2%、κ = .32）

	観測値				
予測値	none	per	ques	com	ub
none	**15**	0	5	5	0
per	6	**21**	9	0	0
ques	2	11	**9**	0	0
com	0	0	0	**0**	0
ub	0	0	0	0	**0**

この傾向をさらに詳しく見るため、予測結果と人手ラベルとの対応をアクセント句の句末境界音調ごとに細分化した（表 3；件数の少ない HL% と LH% は除く）。いずれの方向の予測でも、L% と比べて H% での正解率がかなり低かった（chiba0232：82.3% vs. 57.8%、chiba0432：80.5% vs. 54.2%）。とくに、H% の典型的な機能であるクエスチョン（ques）を正しく予測できていない例が多かった。

3.3　言語・音響特徴の貢献度

音調マーカーの予測に貢献する言語・音響特徴を図 4 に示す。

まず、データ内・個人間で比較すると（列方向）、chiba0232 の X 氏と Z 氏の比較（左列）では、上位 6 位までの特徴（revLoc, tone, lastF0Loc, lastPOS, lastF0Val, f0MaxWord）の相対的な貢献度が同じであった。一方、chiba0432 の Y 氏と Z 氏の比較（右列）では、上位 7 位までに含まれる特徴（revLoc, tone, lastF0Loc, lastF0Val, f0MinAP, amdAP, lastPOS）は同じであるが、その順序は異なった。Y 氏は amdAP をより重視する傾向があり、Z 氏は lastPOS をより重視する傾向があった。

次に、個人内・データ間で比較すると（下段行方向）、Z 氏の chiba0232 と chiba0432 の比較において、上位 7 位までに含まれる特徴のうち 6 個（revLoc, tone, f0MinAP, lastF0Loc, lastF0Val, lastPOS）は共通していたが、その順序は異なった。chiba0232 では lastF0Loc や lastF0Val をより重視する傾向があり、chiba0432 では lastPOS をより重視する傾向があった。また、両者に共通しない特徴として、chiba0232 では f0MaxWord が、chiba0432 では amdAP が重視されていた。

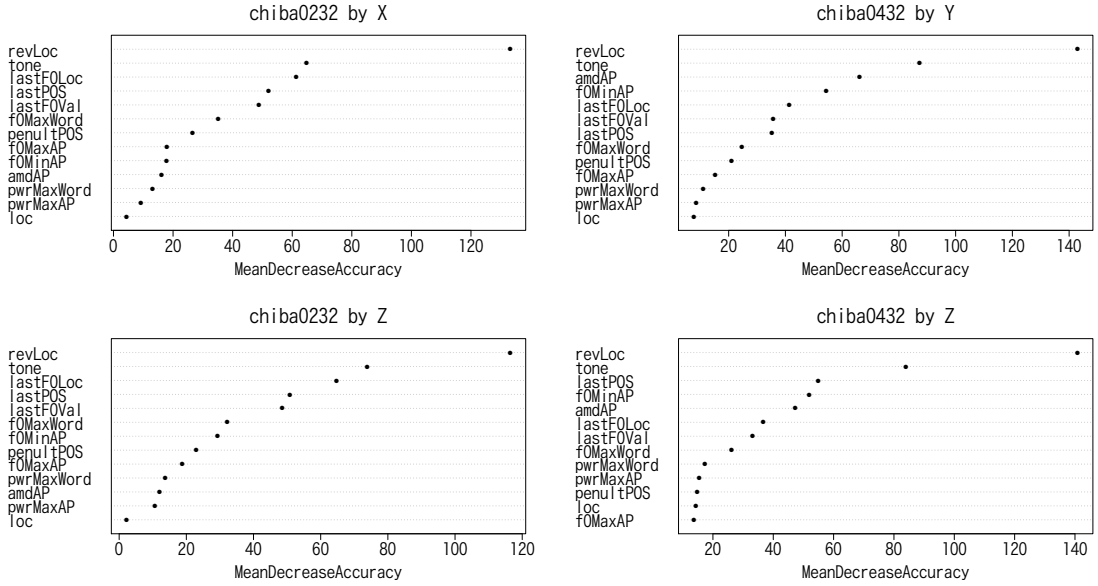


図 4　言語・音響特徴の重要度

4. 議論

転記者の音調マーキングの全般的な傾向として、Z 氏が X 氏、Y 氏よりも音調マーカーを付与していない箇所が多い。chiba0232 のアクセント句全体における none の比率は、X 氏が 31.9% であったのに対して、Z 氏は 37.3% と少し高かった。同様に chiba0432 では、Y 氏の 31.4% に対して、Z 氏は 50.8% に対して音調マーカーを付与しなかった。とくに Y 氏と Z 氏の違いが大きいが、この点は Y 氏の音調マーキングの手順に起因するものと思われる。Y 氏はカリフォルニア大学サンタバーバラ校で Du Bois 流の転記記法を学んだ経験があり、その影響から、まず最初に転記テキスト中にイントネーションユニット (Du Bois et al. 1993) を同定し、イントネーションユニット末ごとに音調マーカーを付与するという方略を採用していた (土屋ほか 2012)。このため、Z 氏よりも音調マーカーを多く付与する傾向があったものと思われる。

また、転記者間で異なる音調マーカーを付与している例も散見された。大きな違いとして、Z 氏が X 氏、Y 氏よりもコンマ（com）を付与する箇所が少ないという点が挙げられる。chiba0232 でコンマ（com）を付与した個所は X 氏の 81 箇所に対して Z 氏は 28 箇所と少なく、同様に chiba0432 では Y 氏の 201 箇所に対して Z 氏は 93 箇所と少ない。このように、どの箇所を継続音調とするかについては、転記者ごとの違いが強く現れると考えられる。

以上のデータ内・個人間のゆれに関しては、音調マーカーを付与する際に重視する言語・音響特徴の違いという点に一部原因を求めることができる。X 氏や Z 氏は lastPOS のような言語特徴をより重視していたのに対して、Y 氏は amdAP のような音響特徴をより重視していた。このことは上に述べた音調マーキングの手順の違いの現れと思われ、結果として個人間のゆれの一因となっているようである。ただし、X 氏と Z 氏は重視する特徴が似通っており、両者の継続音調のマーキングの違いは他の何らかの要因に起因するものと思われる。

さらに、個人内・データ間の違いとして 2 つの会話データ間での Z 氏の音調マーキングを比較すると、音調マーカーの予測に貢献する言語・音響特徴は全般的には共通するものの、細部では一部異なっていた。とくに chiba0432 では amdAP がより重視されており、この傾向は同じデータに対する Y 氏の転記方略と共通していた。このことから、音調マーキングを行なう際に重視する特徴はデータごとに異なると考えられる。これには、会話データの話者の言語的な特徴が関係していると思われる。chiba0432 では、話者の一人は関西方言を話しており、継続や下降の音調が他の話者と異なっている。また、事後インタビューによると、Z 氏は軽微な上昇の後に大きな下降が起きる箇所にも継続音調のマーカーを付与していると答えているが、chiba0432 の話者の一人は軽微な上昇の直後にとくに大きな下降を用いることが多いという点などから、発話の終了や継続を示す話者の発話方略にも影響を受けていると考えられる。

最後に、句末境界音調ごとの多変量モデルの予測精度は、L% と比べて H% でかなり低かった。とくに、H% の典型的な機能であるクエスチョンを正しく予測できていない例が多かった。CSJ 方式で採用している X-JToBI の上昇調 L%H% には疑問上昇調だけでなく強調上昇調なども含まれており、ここから会話分析方式のクエスチョンを抽出するには現状の言語・音響特徴だけでは難しいことがわかった。今後、韻律ラベリング方法も含め再検討する必要があろう。

以上のように、CSJ 方式から会話分析方式への転記変換において、データ間・個人間でさまざまなゆれが存在することがわかった。とくに、このゆれは句末境界音調や音調マーカーごとに大きく異なっていた。今後は、転記方略の違いをモデル化し、予測精度を上げる必要がある。

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

参考文献

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

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

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

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino (1993). “Outline of discourse transcription.” Jane A. Edwards, and Martin D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Lawrence Erlbaum. pp. 45–89.

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

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

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

関連 URL

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

日本語における否定の焦点アノテーション

松吉 俊 (山梨大学大学院医学工学総合研究部)†
大槻 諒 (山梨大学工学部)
福本 文代 (山梨大学大学院医学工学総合研究部)

Annotation of Focus of Negation in Japanese Text

Suguru Matsuyoshi (Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi)
Ryo Otsuki (Faculty of Engineering, University of Yamanashi)
Fumiyo Fukumoto (Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi)

1. はじめに

「誰がいつどこで何をする」という、主に述語項構造で表現される**事象**の末尾に、「ない」や「ん」、「ず」などの語が付くと、いわゆる否定文となる。否定文では、一般に、その事象が成立しないことが表現される。否定文において、否定の働きが及ぶ範囲を**スコープ**、その中で特に否定される部分を**焦点** (フォーカス) と呼ぶ (日本語記述文法研究会 2007)。本論文では、日本語における否定の焦点をアノテーションする枠組みを提案し、現在構築中である、否定の焦点コーパスについて報告する。

『現代日本語書き言葉均衡コーパス』(BCCWJ)[*1]から抽出した、否定のスコープと焦点の例を以下に示す。ここでは、注目している否定を表す表現を太字にしており、そのスコープを角括弧で囲み、焦点の語句に下線を付している[*2]。

(1) だが、[学校での 子どもの様子はわから] **ない**から、それだけでうれしい。 [PN1a_00002]
(2) [十七日まで 選手にも協会関係者にも明かさ] **ない**。 [PN2f_00002]
(3) [力を出し切って 敗れた] **わけではない**。 [PN2f_00003]
(4) [ＷＨＯは] 五月十八日、ジュネーブで開いた総会で台湾の総会へのオブザーバー参加問題を議題としないことを決め、[オブザーバー参加を認め] **なかっ**た。 [PN4g_00001]

文 (1) は、家庭訪問を受けた母親の発言の一部である。「ない」のスコープは、「学校での子どもの様子はわから」で表現される事象である。家庭での子どもの様子は分かると考えられるので、焦点は「学校での」とするのが妥当であると思われる。

文 (2) は、最終登録選手に関しての監督の発言の一部である。「ない」のスコープは、「十七日まで選手にも協会関係者にも明かさ」で表現される事象である。十七日かそれ以降に登録選手を明かすことが期待できるので、焦点は「十七日まで」と考える。

† sugurum at yamanashi. ac. jp
*1 http://www.ninjal.ac.jp/corpus_center/bccwj/
*2 例文の後の “PN” から始まる文字列は、その例文を抽出した BCCWJ 内のファイル名を表す。

文 (3) は、試合に敗れた選手に関する報道記事の一部である。否定の複合辞「わけではない」のスコープは「力を出し切って敗れた」であり、否定の焦点は「力を出し切って」であると解釈できる。

文 (4) は、WHO 総会に関する報道記事の一部である。「なかっ」のスコープは、「WHO はオブザーバー参加を認め」で表現される事象である。この例文においては、(前後の文脈を考慮しても、) スコープの中に特に否定される部分はないように思われる。本研究では、このような場合に、「なかっ」の焦点は、無しとせず、便宜上、スコープ全体であると考える。紙面が煩雑になるのを避けるため、焦点がスコープ全体である場合には、例文に下線を付けない。

英語や日本語を対象として、否定とその焦点について言及した言語学的研究に、加藤ほか (2010) や日本語記述文法研究会 (2007) がある。

これまでに、否定のスコープや焦点を対象としたアノテーションコーパスがいくつか構築されている。BioScope (Vincze et al. 2008) は、生医学分野における英語文章を対象として、否定のスコープをアノテーションしたコーパスであり、自然言語処理の分野において、これを利用して否定のスコープを自動解析する研究 (Morante et al. 2008, Li et al. 2010) が進められている。川添ほか (2011) は、日本語の新聞を対象として否定のスコープのアノテーションを進めている。否定のスコープを対象としたこれらの研究に比べ、否定の焦点を対象とした研究は少ない。Blanco and Moldovan (2011a) は、PropBank (Babko-Malaya 2005) における述語-項関係の上に否定の焦点をアノテーションし、コーパスを構築した。これを利用して否定の焦点を自動検出する研究 (Blanco and Moldovan 2011b, Rosenberg and Bergler 2012) も行われている。日本語では、松吉ほか (2010) が、拡張モダリティの 1 項目として、コーパスにおいて否定の焦点を扱っているが、この研究で実際にアノテーションした事例は非常に少ない。

本論文は、以下のように構成される。まず、2 章において、否定の焦点アノテーションの基本指針について述べる。続く 3 章で、与えられた日本語文章に否定の焦点をアノテーションする枠組みを説明する。4 章で、現在構築中である、2 つのジャンルのコーパスについて報告する。5 章はまとめである。

2. 否定の焦点アノテーションの基本指針

文章に存在する否定を検出し、その焦点にラベルを付け、コーパスを構築する。言語学的利用のみでなく、自然言語処理への応用も考慮して、アノテーションの基本指針を定める。

2.1 焦点の部分を除いた事象が成立すること

否定の焦点がスコープ全体でない場合、スコープの事象が成立しないことだけでなく、焦点の部分を除いた事象は成立することが推測できる (日本語記述文法研究会 2007, Blanco and Moldovan 2011a)。例えば、1 章の文 (3) において、「力を出し切って敗れた」ことは否定されるが、「力を出し切って」の部分に否定の焦点があることが分かれば、「敗れた」ことは成立することが推測できる。同様に、1 章の文 (2) において、「十七日まで」の部分に否定の焦点があることが分かれば、監督はずっと明かさないのではなく、十七日かそれ以降に「選手にも協会関係者にも明かす」ことが成立することが推測できる。Blanco and Moldovan (2011a) は、こ

の考え方に基づき、否定の焦点がスコープ全体でない場合の表現法を提案し、アノテーションコーパスを構築した。我々も、基本指針の 1 つとしてこの考え方を取り入れる。

2.2 否定要素

本論文では、文中において否定を表す表現のことを**否定要素**と呼ぶ。本研究では、次の 3 種類の語群をまとめたものを否定要素と定める。

否定辞 助動詞「ない」と「ず」、接尾辞「ない」、接頭辞「非」、「不」、「無」、「未」、「反」、「異」

非存在の内容語 形容詞「無い」、名詞「無し」

否定を表す複合辞 「のではない」、「わけではない」、「わけにはいかない」など

否定辞のみでなく、非存在の内容語まで含める理由は、「無い」は、存在の内容語「ある」の丁寧な否定「ありません」と同等と思われるからである。否定辞「ん」が使用されている「ありません」は対象とし、「無い」は内容語なので対象としないのは、不合理であると思われる。

言語学の文献 ((森田・松木 1989) など) において、否定を表す複合辞とされる表現は、1 形態素の否定辞と異なる性質を持つと思われるので、区別して扱う。

接頭辞「非」や「不」は、直後の語を否定する働きを持つのみであり、これらに対して焦点を判断する必要はないと思われがちである。しかしながら、次の例のように、「ない」や「ん」と同様に、接頭辞もスコープの一部に焦点を持つことがあるので、対象とした。

(5) 九十年代の「失われた十年」ではっきりしたのは、[もはや 民間まかせでは 過剰債務処理は] **不** [可能] ということだ。 [PN1b_00004]

これは、前の文脈から、過剰債務処理には政府の介入が必要であることが読み取れる例であり、否定の焦点は「民間まかせでは」であると考える。

2.3 否定要素としない語句

否定辞か非存在の内容語を含む 2 形態素以上の慣用表現は、全体を 1 語とし、焦点判断の対象としないこととする。これらの慣用表現は、大きく分けると、次の 2 種類からなる。

複合語 「物足り ない」、「仕方が ない」、「思わ ず」など

否定以外の意味を持つ複合辞 「なけれ ばなら ない」、「かもしれませ ん」、「だけで なく」など

上記の複合語に相当するかどうかは、次の 2 点から判断する。

- 肯定形 (例えば、「仕方がない」に対する「仕方がある」) が、通常、使用されるか
- 国語辞典 (松村・小学館『大辞泉』編集部 1998, 西尾ほか 2000) に見出しが立っているか

複合辞であるかどうかの判断は、言語学の文献 ((森田・松木 1989) など) を参考にし、前節で述べたように、否定を表す複合辞とされる表現は、否定要素として扱う。

助動詞「ない」か接尾辞「ない」、もしくは、形容詞「無い」を使った単純な否定表現に言い換えられない否定の接頭辞は、否定要素とはしない。例えば、「不 十分」は、「十分でない」ことであるので、焦点判断の対象とする。一方、「不 気味」は、「気味が悪い」ことであり、「気味がない」や「気味でない」に言い換えられないので、対象としない。

2.4 否定要素と呼応する程度・頻度の副詞

以下の例文のように、否定要素に呼応する、程度の副詞や頻度の副詞が用いられることがある。ここでは、注目している否定要素を太字にし、程度の副詞や頻度の副詞に下線を付ける。

(6) ボールを回すくらいで、そんなに ハードな練習じゃ**なかっ**た。 [PN2f_00002]

(7) 市街地では、街灯やライトアップによる“光害”で夜空の星が なかなか 見え**ない**。 [PN2g_00004]

(8) 価格は 1 万円前後で、「いつもは ぜいたくでき**ない**けれど、お正月くらい、という方が多いようです」。 [PN3b_00004]

文 (6) で述べられていることは、「全くハードな練習ではなかった」ことではなく、ハードな練習ではあったが、その程度が想定されるよりも高くなかったということである。同様に、文 (7) では、星は全く見えないのではなく、見える程度や頻度が低いということが述べられている。文 (8) の該当箇所は、いわゆる部分否定であり、「ぜいたくできる」ことが全く成り立たないわけではなく、たまには成り立つことが読み取れる。

本研究では、便宜上、呼応するこのような副詞を否定の焦点とみなす。自然言語処理における含意認識 (Dagan et al. 2005) というタスクにおいては、程度や頻度はともかく、ある事象が成立するかどうかを計算機で自動的に判定することが求められる。例えば、上の文 (7) から、「市街地で夜空の星が見える」ことが成立するかどうかを判定することが問われる。文 (7) において、「なかなか」に否定の焦点があることが分かれば、2.1 節の考え方を用いて、機械的に、「市街地では、街灯やライトアップによる“光害”で夜空の星が見える」[*3]ことが導出でき、「見える」と正しく答えることができる。

否定と呼応する、「そんなに」や「なかなか」のような程度・頻度の副詞は、厳密には、否定の焦点ではなく、含意認識で利用したいのならば、別の枠組みを用意すべきであるかもしれない。しかしながら、これらの副詞は、「多くは (持てない)」や「速くは (走れない)」のような形容詞連用形 +「は」や、「頻繁には (通えない)」のような形状詞 +「には」と同様に用いられる。このような形容詞や形状詞を否定の焦点として扱うことは自然であることから、本研究では、これらに連続するものとして、上の副詞も否定の焦点とみなす。

含意認識への応用という観点から、「全然」や「絶対に」、「決して」のような完全否定を表す副詞は、否定の焦点とはみなさないこととする。

3. 否定の焦点アノテーションの枠組み

この章では、まず、否定の焦点を判断する基準について述べる。そして、否定要素とその焦点に対して定めたアノテーション項目と、そこに付与するラベルについて説明する。

3.1 否定の焦点の判断基準

1 章で述べたように、否定要素によって特に否定される部分が否定の焦点である。これを安定して判断するために、2.1 節の考え方に基づいて、我々は次のような判断基準を定めた。

1. ある文の否定の焦点を判断する時には、その文だけでなく、周りの文脈も広く参照する
2. 対象とする文から、一部の表現と否定要素を除外した事象を生成する。その事象が成立

[*3] もちろん、この単純な書き換えは正しくない。人手で正確に書き換えると、「市街地では、街灯やライトアップによる“光害”があっても、夜空の星は、程度や頻度はともかくとして、見える」のようになるであろう。しかしながら、含意認識というタスクにおいて、上の問いに答えるためには、ここまで複雑な書き換えは必要でない。

することが推測できれば、除外した表現の部分を否定の焦点と判断する

3. 解釈に複数の可能性が考えられる場合は、否定の焦点はスコープ全体であるとする
    - 例えば、一部に焦点があると考えることもできるし、スコープ全体が焦点であると考えることもできる場合
    - 例えば、A という部分に焦点があると解釈することもできるし、B という部分に焦点があると解釈することもできる場合

基準 3. は、判断する人間の思い込みを最大限排除するために設けたものである。

3.2 項目とラベル

否定要素に対して、以下のアノテーション項目を定める。

表層文字列　文に出現した否定要素の表層文字列。出現形で記述する

形態素 ID　否定要素の形態素の ID

品詞　助動詞、接尾辞、接頭辞、形容詞、名詞、否定複合辞のいずれか (2.2 節参照)

最終更新日　“YYYYMMDD” という形式で記述された最終更新日

プログラムを用意すれば、これらは自動付与が可能である。ただし、形態素解析辞書 UniDic[*4] では、助動詞ではない「ない」は、すべて「形容詞, 非自立可能」と解析されるため、これらを半自動的に「形容詞」と「接尾辞」に分類する必要がある。

否定の焦点に対して、以下のアノテーション項目を定める。

代表表層文字列　焦点の表層文字列。ただし、後述する代表形態素のみを記述する

代表形態素 ID　焦点の代表形態素の ID

項・節の種類　ガ格、ヲ格、デ格、副詞、ノの項、ナの項、テ節、ト節など、焦点の統語的分類。複数記述可

とりたて詞の有無　「しか」や、数量語に付く「も」が存在するか

意味分類　制限-時間、制限-場所、制限-対象、付加-連用修飾、付加-連体修飾、付加-アスペクトなど、意味解釈に基づいた、否定されている語句の分類

判断の根拠　その箇所を焦点であると判断するに至った根拠。自由記述

手がかり語句　文章中に存在する、焦点判断の手がかりとなった語句。複数記述可

コーパスにおいて否定の焦点は代表 1 形態素にラベル付けする。このように決めた理由は、否定の焦点の自動検出システムを評価する際に、正解とシステムの出力の比較が容易になるからである。代表 1 形態素は、次のように定める。

- 内容語
- 複合語の場合、接尾辞を除く末尾の語
- 修飾語が存在する場合、それが係る末尾の語

表層的な格助詞や接続助詞などに基づく分類が、「項・節の種類」であり、焦点の語句が表す意味に基づく分類が、「意味分類」である。例えば、「意味分類」の “制限-場所” は、場所を表す語句に否定の焦点があり、そこではない場所をうまく選べば、対象事象が成立することを表す。「意味分類」の “付加-連用修飾” は、程度の副詞や頻度の副詞に対して付与する。

[*4] https://www.tokuteicorpus.jp/dist/

次の例のように、とりたて詞「しか」は、必ず否定要素と共起する。

(9) [普段は 決まったものしか 料理し] **ない**ので、おけいこ感覚で。 [PN3b_00004]

「しか」が付く項が否定の焦点となり、文に述べたこの場合には事象は成立するが、これ以外の場合には成立しないことを表す。「しか」が存在する事例には、2.1 節の考え方を適用できないので、特別なマークを付けて、「しか」が存在することを明示する。

数量語に付く「も」が否定要素と共起すると、「その概数には届かない」という意味と、「書き手はそれを少ない・低いと捉えている」ことが表現される (日本語記述文法研究会 2009)。これは、累加の「も」にはない性質である。例を以下に示す。

(10) [出場者ランキングの 二十位にも 入ってい] **なかっ**た [2 年生・高平慎士] が、晴れの舞台で堂々と高校 3 傑入り。 [PN1e_00003]

2.1 節の考え方の適用外ではないが、自然言語処理における評判分析・感情解析タスクに有用であると思われるので、特別なマークを付けて、数量語に付く「も」が存在することを明示する。

3.3 データ構造

否定の焦点コーパスは、図 1 のような XML によって表現する。この図は、1 章の文 (1) に対するアノテーション結果である。XML において、<wsb:negation> 要素を用いて否定要素の情報を記述し、<wsb:focus> 要素と <wsb:description> 要素、<wsb:clue> 要素を用いて否定の焦点の情報を記述する。ここで、“wsb” は、植物のわさびから名付けた名前空間である。

<wsb:negation> 要素は、1 文もしくは文の断片を表す <sentence> 要素の直接の子要素として記述する。<wsb:negation> 要素の属性に、前節で述べたアノテーション項目を記述する。<wsb:focus> 要素は、<wsb:negation> 要素の直接の子要素として記述する。否定の焦点がスコープ全体である場合は、<wsb:focus> 要素に対して、1 という値を記述した@wsb:scope 属性のみを指定する。否定の焦点がスコープの一部である場合、<wsb:description> 要素と <wsb:clue> 要素を、<wsb:focus> 要素の直接の子要素として記述する。アノテーション項目の多くは、<wsb:focus> 要素の属性に記述するが、「判断の根拠」は <wsb:description> 要素として、「手がかり語句」は <wsb:clue> 要素として記述する。

3.4 否定のスコープ

本来ならば、否定の焦点をアノテーションする前に、否定のスコープを明示的にアノテーションすべきである。既存の述語項構造解析の技術を用いれば、ある程度は自動的に否定のスコープを認識することができるが、対象が整った文章でない場合、人間による修正作業が多く発生する。本研究では、人的コストの関係から、否定のスコープをアノテーションしない。人間が否定の焦点を判断する時には、対象となる否定要素のスコープを目で確認するに留める。

4. 否定の焦点コーパス

前章で説明したアノテーションの枠組みに基づき、次の 2 つのテキストデータを対象として、否定の焦点コーパスの構築を進めている。

```
   :
 </superSentence>
 <sentence wsb:sID="73">
  <wsb:negation wsb:orthtoken="ない" wsb:morphID="13250" wsb:POS="助動詞"
wsb:lastupdate="20130112">
    <wsb:focus wsb:orthToken="学校" wsb:morphID="13170" wsb:argTypes="ノの項; デ格"
wsb:class="制限-場所" wsb:NumOfCandidates="pl">
      <wsb:description>家での様子は分かる</wsb:description>
      <wsb:clue wsb:sID="62" wsb:orthTokens="主婦" wsb:morphIDs="12020" />
    </wsb:focus>
  </wsb:negation>
  <LUW B="S" SL="vf" l_lemma="だが" l_lForm="ダガ" l_wType="和" l_pos="接続詞"
l_formBase="ダガ">
    <SUW orderID="13140" lemmaID="22916" lemma="だ" lForm="ダ" wType="和" pos="助動詞"
cType="助動詞-ダ" cForm="終止形-一般" formBase="ダ" orthBase="だ" pron="ダ"
start="19490" end="19500">だ
    </SUW>
     :
```

図 1　コーパスにおける XML ファイルの例 [PN1a_00002]

1. 楽天データ[*5]の楽天トラベル: レビューデータ
2. BCCWJ におけるコアデータ内の新聞 (PN)

4.1　楽天トラベル: レビューデータ

楽天トラベル: レビューデータのうち、先行研究において小池ほか (2012) が使用したものと同じレビュー集合を対象とした。彼らは、まず、宿泊施設に対するレビュー数の分布を調査し、90% 以上の宿泊施設はレビュー数が 1 から 58 の範囲にあることを明らかにした。そして、その結果に基づき、レビュー数が 10 から 58 の範囲の宿泊施設の全体から、無作為に 40 の宿泊施設を抽出し、独自の文分割規則により半自動的にそのレビュー集合を文分割した。

このコーパスには、5,178 文が含まれており、形態素の品詞情報のみに基づいて抽出した否定要素の候補は、1,246 個であった。以下、このコーパスを「レビュー」と表記する。

4.2　BCCWJ コアデータの新聞

BCCWJ 全体の約 1/100 のデータがコアデータに指定されており、このデータは、その他の部分と比較して高い精度で解析が施されている。コアデータの一部に言語学的情報を付与する場合、国立国語研究所が定めたファイル優先順位に従うことが推奨される。我々は、コアデータ内の新聞 340 ファイルのうち、優先順位が 1 から 54 までの “A” グループを対象とした。

このコーパスには、1 文もしくは文の断片を表す、XML の <sentence> 要素が 2,708 個含まれており、否定要素の候補は、406 個であった。以下、このコーパスを「新聞」と表記する。

4.3　コーパスの分析

2 つのコーパス「レビュー」と「新聞」における、否定要素候補の分布を表 1 に示す。2 つのコーパスにおいて、否定要素はそれぞれ 1,023 個と 304 個であり、いずれのコーパスでも、

[*5] http://travel.rakuten.co.jp/

表 1　否定要素候補の分布

	レビュー	新聞	計
助動詞	637	173	810
接尾辞	116	33	149
接頭辞	19	34	53
形容詞	211	53	264
名詞	28	6	34
否定複合辞	12	5	17
(上記小計)	(1,023)	(304)	(1,327)
複合語	94	30	124
その他複合辞	121	72	193
解析誤り	8	0	8
(上記小計)	(223)	(102)	(325)
計	1,246	406	1,652

表 2　スコープ全体でない焦点の分布

	レビュー	新聞	計
副詞	141	18	159
ガ格	30	5	35
ヲ格	7	6	13
ニ格	49	11	60
デ格	17	6	23
マデ格	5	4	9
カラ格	3	2	5
ト格	3	1	4
その他の格	1	2	3
ノの項	20	7	27
連体の述語	8	8	16
接頭辞「全」	1	0	1
テ節	1	2	3
ト節	1	0	1
アスペクト	14	0	14
計	301	72	373

助動詞「ない」と「ず」が全体の過半数を占めることが分かる。

2 つのコーパスにおいて、否定の焦点がスコープ全体でないものは、それぞれ 301 個と 72 個であった。「レビュー」では、29%(301/1,023) の否定要素が、「新聞」では、24%(72/304) の否定要素が、スコープの一部に焦点を持つことが分かる。これらの焦点の「項・節の種類」の分布を表 2 に示す。図 1 に例示されるような、ある格と “ノの項” が同時に付与されている事例は、この表では、“ノの項” として集計した。「レビュー」には、焦点が副詞である否定要素が多いことが分かる。「新聞」のデータ数が少ないので、確定的なことは言えないが、どの格が焦点になりやすいかも、2 つのコーパスで異なる傾向があるようである。

焦点である部分に付いていたとりたて詞の数を表 3 に示す。2 つのコーパスを合わせ、35%(129/373) の焦点に何らかのとりたて詞が付いていたことが分かる。とりたて詞「は」は、焦点である箇所の手がかりとして利用できそうに見えるが、「は」は、特に主題を表す「は」として、スコープ全体が焦点である事例にも多く出現するので、注意が必要である。3.2 節で述べたように、スコープの中に「しか」が付く項が存在する場合、それが否定の焦点となる。

焦点の語句が表す意味に基づく分類結果を表 4 に示す。「レビュー」には、焦点が副詞である否定要素が多いため、“付加-連用修飾” が多いことが見て取れる。「レビュー」は宿泊施設のレビュー集合であるので、場所を表す語句に否定の焦点がある “制限-場所” が、「新聞」に比べ、著しく多いことが分かる。

4.4　アノテーション作業

独自のプログラムにより、3.3 節で説明した XML 形式のファイルから、人間が見やすい HTML 形式に自動変換可能である。作業者は、ブラウザー上で HTML ファイルを確認しながら、テキストエディターにおいて XML ファイルを更新する。作業にかかる時間は、100 個の

表 3　焦点に付いていたとりたて詞

	レビュー	新聞	計
「は」	66	13	79
「しか」	34	7	41
「も」	7	1	8
「だけ」	0	1	1
計	107	22	129

表 4　焦点の意味分類結果

	レビュー	新聞	計
制限-動作主	13	5	18
制限-対象	27	12	39
制限-時間	10	9	19
制限-場所	40	3	43
制限-数量	10	5	15
制限-範囲	43	12	55
付加-連用修飾	125	15	140
付加-連体修飾	19	11	30
付加-アスペクト	14	0	14
計	301	72	373

否定要素候補に対して 3 時間程度である。

2 人の作業者が独立に「新聞」に対してアノテーション作業を行い、2 人の作業結果において焦点の場所がどれほど一致するかを調査した。全 304 個の否定要素のうち、103 個が不一致であったが、2 時間ほど 2 人で議論することにより、これらの不一致を解消することができた。不一致の主な原因は、以下の 3 点であった。

- スコープが明示されていないことによる勘違い
- 作業者のうち 1 名は、広く文脈を参照していなかった
- とりたて詞「だけ」が持つ限定の意味に引っ張られた

5. おわりに

本論文では、日本語における否定の焦点をアノテーションする枠組みを提案し、現在構築を進めている否定の焦点コーパスについて報告した。今後は、BCCWJ の新聞以外のレジスタに対してもアノテーション作業を進めることを考えている。このコーパスを用いて試作した、否定の焦点自動検出システムについては、別稿 (大槻ほか 2013) を参照されたい。

構築したコーパスは、BCCWJ および楽天データとの差分形式で、論文末に示す URL にて無償で一般公開する予定である。

謝　辞

本研究の一部は、科研費若手研究 (B)「高精度モダリティ解析のための言語資源構築に関する研究」(課題番号: 23700176、代表: 松吉俊) の支援を受けている。

文　献

日本語記述文法研究会 (編) (2007).『現代日本語文法 3』　くろしお出版.

日本語記述文法研究会 (編) (2009).『現代日本語文法 5』　くろしお出版.

Babko-Malaya, Olga (2005). *PropBank Annotation Guidelines.*, ACE (Automatic Content Extraction) Program. http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf

Blanco, Eduardo, and Dan Moldovan (2011a). “Semantic Representation of Negation Using Focus Detection.” *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 581–589.

Blanco, Eduardo, and Dan Moldovan (2011b). “Some Issues on Detecting Negation from Text.” *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pp. 228–233.

Dagan, Ido, Oren Glickman, and Bernardo Magnini (2005). “The PASCAL Recognising Textual Entailment Challenge.” *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Li, Junhui, Guodong Zhou, Hongling Wang, and Qiaoming Zhu (2010). “Learning the Scope of Negation via Shallow Semantic Parsing.” *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 671–679.

Morante, Roser, Anthony Liekens, and Walter Daelemans (2008). “Learning the Scope of Negation in Biomedical Texts.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 715–724.

Rosenberg, Sabine, and Sabine Bergler (2012). “UConcordia: CLaC Negation Focus Detection at *Sem 2012.” *Proceedings of the First Joint Conference on Lexical and Computational Semantics: SemEval'12*, pp. 294–300.

Vincze, Veronika, György Szarvas, Richárd Farkas, Gy*ddotor*gy Móra, and János Csirik (2008). “The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes.” *BMC Bioinformatics*, pp. 1–9.

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.

小池惇爾、松吉俊、福本文代 (2012).「評価視点別レビュー要約のための重要文候補抽出」 言語処理学会第 18 回年次大会論文集, pp. 1188–1191.

松吉俊、江口萌、佐尾ちとせ、村上浩司、乾健太郎、松本裕治 (2010).「テキスト情報分析のための判断情報アノテーション」 電子情報通信学会論文誌. D, 情報・システム, 93:6, pp. 705–713.

松村明、小学館『大辞泉』編集部 (編) (1998).『大辞泉（増補・新装)』 小学館.

森田良行、松木正恵 (1989).『日本語表現文型用例中心・複合辞の意味と用法』 アルク.

西尾実、岩淵悦太郎、水谷静夫 (編) (2000).『岩波国語辞典第六版』 岩波書店.

川添愛、齊藤学、片岡喜代子、崔栄殊、戸次大介 (2011).『言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4』, Technical Report of Department of Information Science Ochanomizu University.

大槻諒、松吉俊、福本文代 (2013).「否定の焦点コーパスの構築と自動検出器の試作」 言語処理学会第 19 回年次大会論文集 C6-1.

関連 URL

否定の焦点コーパス: `http://cl.cs.yamanashi.ac.jp/nldata/negation/`

聞き手行動としての日本語あいづち表現の分析：転記情報とコーディングによる発話連鎖パターンの認定

吉田悦子（三重大学人文学部）†

Detecting Patterns of Sequences by Coding Scheme and Transcribed Utterance Information: An Analysis of Japanese Reactive Tokens as Non-primary Speaker's Role

Etsuko Yoshida (Faculty of Humanities, Law and Economics, Mie University)

１．はじめに

あいづちは、話し手が発話権を保持する状態で、聞き手が産出する表現形式であり、相手の話をきいていることを示すシグナルである。本稿では、狭義のあいづち表現（非言語形式を含まない）を対象に、日本語の地図課題対話の転記テキストを基にして、音声やピッチ等の情報を部分的に追加転記し、ムーブ(move) による構造（Carletta et al.: 1997）を利用した機能分析をおこなう。このコーディングに従うと、あいづち表現は、典型的に認定(acknowledge)として一様に表記される傾向があることがわかる。この認定ムーブと他のムーブとのつながりを分析することで、発話連鎖のプロセスや相互行為の質的な違いを記述することが可能となる。その結果、あいづち表現が、どんなパターンの発話連鎖で出現し、対話進行においてどのような聞き手役割を演じているのかについて一定のパターンを明らかにすることができるのではないかと考えられる。本論では、先行研究の概観と対話データの概要および分析方法を示し、予備的な分析結果と分析例を示して、考察する。

２．あいづち表現の形式と機能

あいづち表現とは、「話し手が発話権を行使している間に聞き手が送る短い表現」(メイナード 1993：58）という定義を採用する。いわゆるあいづち詞として認定される言語形式は多様であり、「入出力制御系感動詞」として分類される（田窪・金水 1997)。英語、日本語、中国語のあいづち表現を比較した Clancy et al. (1996)では、'reactive token'という用語を導入し、backchannel を超えて表現形式を拡張し、backchannels、reactive expressions、collaborative finishes、repetitions、resumptive openers の 5 つのタイプに分類した。それぞれの頻度と会話分析的手法により分析し、発話連鎖におけるあいづち表現の役割を検討している。[1] この分類を参照し、吉田・高梨・伝（2009）におけるあいづち表現の認定基準および出現位置と形態の問題についての議論に基づき、Den et al.(2012)では、response tokens と名付けたあいづち表現について 2 段階のアノテーションを設定し、形態と直前の発話との関係性を分析、さらにあいづち表現を誘引する話し手の発話情報にも注目した。ここでの形式的分類は 6 タイプ（Responsive interjections, Expressive interjections, Lexical reactive expressions, Evaluative expressions, (Partial) repetitions, (Collaborative) completion）である。あいづち表現として利用される形式のうち、両者に共通している形式を含めると共に、本研究で扱う対話データが

†tantan@human.mie-u.ac.jp

1 Gardner(2001)はこの分類を詳細に検討し、大浜・西村(2005)は日英それぞれの形式的分類について批判的な解説を加えている。

課題対話であることから、出現しにくいと考えられる形式を除き、以下の 4 つの形式を原則として採用する。なお、非言語形式のあいづち（うなずきや笑い）は含めない：

（1）「うん」「はい」「へえ」「そう」などのあいづち詞[2]　および語彙的応答
（2）繰り返し（話し手の語句を反復する。部分反復を含む）
（3）言い換え（別の表現に言い換える）
（4）先取り発話（話し手の後続発話を、聞き手が先取って完結させる）

あいづち表現は、対話における分析の単位である発話の認定によっては、単位にも、発話断片（牧本、柏岡、キャンベル 2007;　伝ほか 2008）にも生起する。聞き手ターンとして独立している場合と話し手がターンを取っている間に重複して生起する場合もある。[3]

いずれにしても、あいづち表現は「聞き手行動」として位置づけられ、2 つのタイプの関連性について分析する。 話し手への反応としての「聞き手的行動タイプ」と局所的に話し手役割を担う「話し手的行動タイプ」があると考えられる（伝 2009）[4]。課題遂行対話には達成すべき明確な対話目的があり、役割分担が明らかであるとはいえ、自然な発話に見られるのと類似の現象が認められ、特徴的なパターンを認めた上で、一般化することは可能であろう。

3. 対話データ

3.1　対話データの概要

対話データとしては、地図課題対話による日本語と英語のパラレルコーパス各 8 対話が用意されているが、本稿では日本語 1 対話分について予備分析をおこなった。地図課題は 2 人の実験参加者により共同で達成される課題である。2 人の実験参加者に課される課題は、相手の地図が見えないように向かい合い、お互いに会話を交わしながら、情報提供者(information giver: 以下 G と記載)の地図上の経路を情報追随者(information follower: 以下 F と記載)の地図上に再現することである。なお、G の地図には経路以外に出発地点と目標地点、そしていくつかの目標物が描かれている。一方、F の地図には、出発地点と目標物だけが描かれており、目標地点と経路は描かれていない。2 つの地図は完全に同一ではなく、その違いについては課題の遂行過程で解決すべき問題の 1 つとされる（詳細については吉田 2002 を参照）。

3.2　分析方法

この課題では話し手と聞き手の役割がかなり明確に区別されており、G を中心に対話が進行することが予測される。このため、聞き手の役割は G の情報を理解し、その進行を助け、自分の課題を達成することが期待される。課題の達成は両者にとっての対話目的であり、必要な情報は相互にやりとりされ、共有される。

---

2 話し手の後続発話を、聞き手があいづちで先取るような先取りあいづちもここに含める。ただし、「うんうん」という繰り返しは一回のあいづちとみなす。

3 グラウンディング（基盤化）の形成において、あいづちの役割は、「対話の中で交換される情報が、対話参加者の間で共有されていく相互理解の形成過程」と みなされ、あいづちや繰り返しなどの対話調整機能の役割と機能が議論されている（Pickering and Garrod 2004）。

4 あいづち表現による聞き手ターンは、発話権を取得しているとはみなさないので、あいづちの拡張としてとらえて差し支えないと考える。

今回、二つの分析方法でおこなう。まず、G と F それぞれの相互行為である「あいづち表現」について、2 で示したような 4 つの形式で分類する。次に、こうしたあいづち表現を介した語り手と聞き手とのやりとりがどのような交換パターンをとるかを観察するために、機能的側面をムーブ(moves)の連鎖から考察する。発話機能を分析する方法として、対話のムーブ構造（Sinclair and Coulthard 1975、Sinclair and Coulthard 1992、石崎、伝 2001）を基盤とし、とくに、地図課題対話に特化した発話機能タグとして考案されたアノテーションであるムーブを考案したコーディングを利用する(Carletta et al.: 1997)。ムーブとは対話の局所的な構造をとらえ、ターンの下位分類として位置づけられる。ムーブには発話単位として統一された単位はなく、相互行為の基本的なユニットである conversational games を支える下位分類である。地図課題対話の場合、ムーブは initiation, response, preparation の 3 つのカテゴリーに分類され、以下のようにさらに下位分類される(Carletta et al.: 1997)：

The initiation moves（開始）の分類：
(1) Instruct move：指示
(2) Explain move：説明
(3) Align move：次の発話に移ってよいか、説明に同意しているかを確認する質問
(4) Check move：発話内容や情報について確認する質問
(5) Query-yn move：yes / no が答えになる質問 （yes, no が省略されていてもよい）
(6) Query-w move：wh で始まる質問

The response move（応答）の分類：
(7) Acknowledge move：理解したことを示す返答
(8) Reply-y move：yes / no の質問に対する yes の返答
(9) Reply-n move：yes / no の質問に対する no の返答
(10)Reply-w move：yes /no の質問以外の質問に対する返答
(11)Clarify move：ある種の質問に対する返答
The preparation move（準備）の分類：
(12)Ready move：dialogue game が終わり次の dialogue game の始まりを告げるもの

このうち、あいづちに関与するムーブは（7）の認定ムーブに該当する。このムーブは An ACKNOWLEDGE move is a verbal response which minimally shows that the speaker has heard the move to which it responds, and often also demonstrates that the move was understood and accepted.’(Carletta et al.: 1997)と定義され、次のような英語の事例が示されている。

(1) G: Ehm. If you … you’re heading southwards. (Instruct)
F: *Mmhmm.* (Acknowledge)

(2) G: Do you have a stone circle at the bottom? (Query-yn)
F: No. (Reply-n)
G: *No, you don’t.* (Acknowledge)

(1) は指示に対する認定、(2)は否定の返答に対する認定および補足（Follow-up）とみなされる。こうした認定ムーブは、いずれもあいづち表現としてみなすことができ、日本語においても観察される。

４．分析結果

まず、図 1 のように対話参与者別のあいづち表現の分布を見てみる。予想通り、F のあいづち表現が優勢であり、G のあいづち表現は全体の 4 分の 1 にとどまっている。表現の多様

さの点からもFは、40例の繰り返しを使用し、言い換えと先取り発話を各4例使用している。

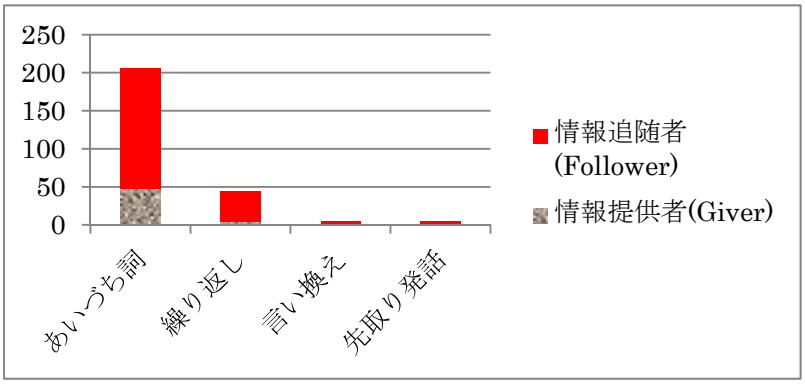


図1　対話参与者別のあいづち表現の分布

すでに述べたように、GとFの聞き手役割の比重から見れば、あいづち表現がFに多用されていることは当然の結果であるといえる。あいづち表現を実現する発話機能が認定ムーブであることは3.2で確認した。では、こうした認定ムーブはどんな発話連鎖において現れてくるのだろうか。Fのあいづち表現はどんな先行発話のあとに出現し、どんなムーブの連鎖が認められるだろうか。関連して、Gのあいづち表現はターンの交替とかかわるのか、その場合どんなムーブの連鎖が認められるだろうか。以下の分析例を見てみよう。[5]

(3)

→G:ここからあ<400+>ん<400+>もうすこしまだしたにいくんですけどお: ,+ (Instruct)
→F:+はい (Acknowledge)
G:まっすぐう<230>じゃなくて<400+>ちょっとだ<290>けかーぶするような
さっ*きからつずいていくと: , (Explain)
F:　　*ん<400+>ん (Acknowledge)
G:ゆるやかなえすじ*みたいになるぐらい: , (Explain)
F:　　*んっ (Acknowledge)
G:*えすまではいかないですけど: ,+ (Explain)
F:*はい<400+>+はい (Acknowledge)
G:まっすぐじゃなく<400+>ちょっとだけかーぶしているかんじで: ,+ (Explain)
F:+はい (Acknowledge)
G:もうすこししたのほうまで (Explain)
→F:っとこれはあのすたーとちてんからそのー<310>*つぎにいく (Check)
→G:　　*はい (Acknowledge)
F:ところまで: , (Check)
G:はい (Acknowledge)
F:あのせいかくにるーとをあのかかなかきゃいけないんですか<400+> (Check)
→　あの<400+>せん<400+>あ<400+>で: ,わたしがはじめにいたところは: , (Explain)
→G:はい (Acknowledge)
F:んっこれ<400+>おんなじですよねかいてあるえは (Check)
G:*は: ,

[5] コーパスの転記テキストは、千葉大学で開発された書き起こしツールにより作成されている。各発話単位は時間情報を利用して400ミリ秒以上のポーズにより区切られているが、紙面の都合上、この間を<400+>と記す。(2)の例のように<xxx>で示される数値は400ミリ秒未満のポーズの長さである。ムーブ認定できない部分は空欄としている。

F:*どちらもち*ず (Explain)
G: *ちがうー{*ざっし[?]} (Explain)
F: *ちがう? (Acknowledge)
F: えすじで<330>*なんか (Explain)
G: *えすえすまでいかないんですけ*ど: , (Explain)
F: *んっうん<400+>えっつぎに<300>せんをひくところまで: , (Explain)
G:はい+ (Acknowledge)
→F:+なんか<400+>つりばしみたいに<400+>なってます: ? (Check)
G:あっつりばしは: ,*そのすたーとちてんからみたら: ,*あ (Explain)
→F: *はい<400+> *はい (Acknowledge)
→F:*ひだりっかわですよ*ね (Acknowledge)
G:*ほく<400+> *ほくせい?<400+>ですよ (Explain)

G は指示ムーブで導入し、説明ムーブに移行する一方、F はあいづち詞による認定ムーブのやりとりが続く(「指示/説明―認定」)が、この連鎖がとぎれるのは、F の確認ムーブが始まるところである。G はこれにあいづち詞で応え、「確認―認定」の連鎖がある。F はこの後、説明ムーブに切り替え、G とのやりとり（「説明―認定」)の後、確認ムーブを取り、G は説明ムーブに戻る（「確認―説明」)。その後、F はあいづち詞と先取り発話「ひだりっかわですよね」による認定ムーブに戻る(「説明―認定」)。

このように、F は確認ムーブをきっかけとして説明ムーブに移行し、局所的な話し手行動を取っている。その間、G は説明を中断し、あいづちによる認定ムーブを取り、局所的な聞き手の役割を演じている。指示ムーブや説明ムーブが先行する場合、その反応としてあいづちを誘う認定ムーブが出現しやすい。こうした連鎖は、対話が進行している印象を与えている。一方で、確認ムーブは、説明ムーブを要求すると同時に、話者交替のきっかけとして機能することも示唆された。この連鎖には、問題の所在を確認し、修復を試み、解決に向かおうという流れが認められる。そして、説明ムーブに納得できた時点で、認定ムーブが出現することになるだろう。

５．まとめ

本稿では、典型的に認定(acknowledge)ムーブとして表記されるあいづち表現について分類し、その発話機能について予備的な分析をおこない、他のムーブとの連鎖において考察した。ムーブの移行や、対話の進行や修復にかかわる発話連鎖のプロセスがなんらかの相互行為のパターンを成していることが垣間見えてきた。しかしながら、ムーブの認定自体が、形式だけで割り切れるものではなく、正確な分析にはより詳細な発話情報が必要である(指示と説明の区別)。異なるタイプのあいづち表現が、どんなパターンの発話連鎖で出現し、対話進行においてどのような聞き手役割を演じているのかについては、こうした発話情報なしには明らかにすることはできないだろう。転記方法とデータの解釈の問題も未解決であり、今後の課題としたい。

謝　辞

本研究は、科学研究費補助金基盤 C「共参照関係を利用した対話プロセス研究と対話型言語教材の開発」(H 22～24 年度　課題番号 2252039　研究代表者 : 吉田悦子) の一部である。本稿は，国立国語研究所共同研究プロジェクト「多様な様式を網羅した会話コーパスの共有化」（プロジェクトリーダー　伝康晴）の共同研究発表会での発表に基づき、内容を補足

追加した。グループリーダーおよびメンバーより有益なコメントを多く頂いた。ここに記して感謝する。

文　献

Carletta, J., Isard, A., Isard, S., Kowtko, J.C., Doherty-Sneddon, G. and Anderson, A.H. (1997)‘The reliability of a dialogue structure coding scheme,’ *Computational Linguistics*, 23, pp.13-31.

Clancy,P.M., Thompson, S.A., Suzuki, R.,&Tao, H. (1996). ‘The conversational use of reactive tokens in English, Japanese, and Mandarin’. *Journal of Pragmati*cs, 26, pp.355-387.

伝康晴、小磯花絵、丸山岳彦、前川喜久雄、高梨克也、榎本美香、吉田奈央. (2008) 「対話研究にふさわしい発話単位の認定に向けて」人工知能学会研究会資料, *SIG-SLUD-A802*, pp.27-32.

伝康晴（2009）「聞き手行動の認知科学に必要なもの」『認知科学』16 (4), pp.475-480. (Dec.2009)

Den, Y., Koiso, H., Takanashi, K., & Yoshida, N. (2012). ‘Annotation of response tokens and their triggering expressions in Japanese multi-party conversations’. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)* (pp. 1332-1337). Istanbul, Turkey.

Gardner, R. (2001). *When listeners talk*. Amsterdam: John Benjamins.

石崎雅人・伝康晴(2001)『談話と対話』東京大学出版会

牧本慎平、柏岡秀紀、ニック・キャンベル(2007)「実対話コーパスに対する意味単位・構造情報のタグ付け」『言語処理学会第 13 回年次大会発表論文集』pp.764-767。

大浜るい子・西村史子（2005）「日英のターン交替と相づち使用の実相-日本人学生とニュージーランド学生の比較を通してー」『社会言語科学』第 7 巻第 2 号.pp.78-87.

Pickering, Martin J. and Simon Garrod (2004) ‘Toward a mechanistic psychology of dialogue,’ *Behavioral and Brain Sciences*, 27, pp.169-226.

泉子・K・メイナード.（1993）『会話分析』くろしお出版.

Sinclair, J. M. & Coulthard R. M. (1975). *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press.

Sinclair, J.McH. and Coulthard, R.M. (1992). ‘Towards an analysis of discourse’. In R.M.Coulthard (ed.), *Advances in spoken discourse analysis* (pp.1-34). Routledge.

田窪行則。金水敏（1997）「応答詞・感動詞の談話的機能」音声文法研究会（編）『文法と音声』(pp.257-279). くろしお出版.

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）第 19 号, pp.241-249.

吉田奈央・高梨克也・伝康晴(2009)「対話におけるあいづち表現の認定とその問題点について」『言語処理学会第 15 回年次大会発表論文集』pp.430-433.

書　名　第3回 コーパス日本語学ワークショップ予稿集
発行日　平成25年2月25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）