第2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2回 コーパス日本語学ワークショップ

予稿集

2012年9月6日、7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2回 コーパス日本語学ワークショップ

予稿集

2012年9月6日(木)／7日(金)

Program [プログラム]

9月6日（木）

13:00～13:10 **■挨　拶**　前川 喜久雄

■口頭発表（1）

13:10～13:40 **BCCWJコアデータの頻度情報に基づく日本語論文・レポートライティング指導の試み**
▷堀 一成、坂尻 彰宏

13:40～14:10 **テクストの結束性に与る語彙とその機能について**
▷高崎 みどり

14:10～14:40 **句末境界音調のピッチレンジに与える要因：『日本語話し言葉コーパス』の分析**
▷五十嵐 陽介、小磯 花絵

14:40～15:00 **■ポスターの紹介**

15:00～17:00 **■ポスター発表（1）**

『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
▷小西 光、浅原 正幸、前川 喜久雄

漢字・語彙指導の根拠としてのコーパスの役割
▷河内 昭浩

「語りかけ性」を有すると判断される書きことばの表現
▷保田 祥、柏野 和佳子、立花 幸子、丸山 岳彦

中古和文における長単位の概要
▷冨士池 優美

助動詞レル・ラレルへの意味アノテーション作業経過報告
▷小山田 由紀、柏野 和佳子、前川 喜久雄

動詞語義及び意味役割付与作業システムの構築
▷上野 真幸、竹内 孔一

『太陽コーパス』にみる、動詞性名詞「報告」の使用実態
▷佐藤 佑

名詞「甲斐」の文法的性格
▷中平 詩織

中国人日本語学習者の「的」付きナ形容詞の習得に関する研究
―BCCWJ コーパス調査とアンケート調査の分析を通じて―
▷呉 雪梅

「語彙レベル」から見た近代の語彙と現代の語彙
―『太陽コーパス』と『現代日本語書き言葉均衡コーパス』を用いて―
▷田中 牧郎

通時コーパス用『中納言』：Webベースの古典語コンコーダンサー
▷小木曽 智信、中村 壮範

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―
▷土屋 智行、伝 康晴、小磯 花絵

「気持ち」の意味について
▷加藤 恵梨

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
▷田中 リベカ、川添 愛、戸次 大介

メタファー表現の生産性に対する意味の焦点と表現メディアの影響
―＜急激な増加＞や＜大量の存在＞を表す表現の場合―
▷大石 亨

書籍テキストへの文体情報付与の試み ―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―
▷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

極性反義語の用例分布とその解釈
▷服部 匡

9月7日（金）

10:00～10:20　**■ポスターの紹介**

10:20～12:20　**■ポスター発表（2）**

現代日本語書き言葉均衡コーパスに対する難易度付与
▷佐藤 理史

文末機能表現シソーラスと述部正規化システム
▷松木 久幸、佐藤 理史、駒谷 和範

論文の論理構造における分野基礎用語に関する分析
▷内山 清子

コーパス用テキストの文字校正支援ツールの設計と実装
▷堤 智昭、須永 哲矢

『現代日本語書き言葉均衡コーパス』を用いた文末表現のバリエーションの分析（2）
▷丸山 岳彦

Webデータに基づく複合動詞データベースの構築
▷山口 昌也

日本語話し言葉コーパスを用いた複合境界音調の発言継続表示機能の検討
▷小磯 花絵

文の長さ分布から見た文生成のメカニズム
▷古橋 翔

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
▷石本 祐一、小磯 花絵

Praat起動用Excelアドイン"Praat Launcher"
▷西川 賢哉

多様な話者による演技感情音声の収集と特徴の比較
▷宮島 崇浩、菊池 英明

BCCWJに含まれるウェブデータの特性について
―データ重複の諸相とBCCWJ使用上の注意点―
▷田野村 忠温

漢字四字成語の受容とその延命
▷砂岡 和子、羅 鳳珠、王 雷、姜 柄圭、松崎 実夏

日本語学習者にとって読みやすい文章について ―日本語教科書における書き換えの分析から―
▷クリスティーナ・フメリャク・寒川

段落間の類似度を利用したテクストの結束性の測定
▷山崎 誠

状態空間表現を用いた文章の特徴付け
▷馬場 康維、小森 理

近代文語論説文を対象とした濁点の自動付与アプリケーション
▷岡 照晃

12:20～13:30　**昼食・休憩**

■口頭発表（2）

13:30～14:00　**近代対訳コーパスにおける日韓語彙の諸相 ―文体の異なる対訳コーパスの比較を通して―**
▷張 元哉

14:00～14:30　**オンライン・コミュニケーション上での平均使用語彙数に関する研究**
▷荒牧 英治、増川 佐知子、森田 瑞樹、保田 祥

14:30～15:00　**『明六雑誌コーパス』の開発 ―近代語コーパスのモデルとして―**
▷近藤 明日子、小木曽 智信、須永 哲矢、田中 牧郎

15:00～15:15　**休憩**

15:15～16:15　**■全体討議**

Contents [目次]

口頭発表（1）

9月6日（木）　13:10～14:40

BCCWJ コアデータの頻度情報に基づく 日本語論文・レポートライティング指導の試み

堀 一成（大阪大学　全学教育推進機構）†
坂尻 彰宏（大阪大学　全学教育推進機構）†

Attempt to Provide Educational Support for Japanese Academic Writing, Using Frequency Information Retrieved from the BCCWJ Core Corpora Data

Kazunari Hori (Osaka University, Center for Education in Liberal Arts and Sciences)
Akihiro Sakajiri (Osaka University, Center for Education in Liberal Arts and Sciences)

１．概要

論文・レポートのライティング指導の基礎データとするため、BCCWJ・コアデータ「『代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備』総括班(2011)」より語彙頻度情報をマイニングし、指導に活用した事例を報告する。

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

今回は、このような試みの第一報として報告する。BCCWJ コアデータの白書データを選び、動詞・名詞の頻度情報を得た。一般文でも利用される頻度が高いと判断される語を、頻度上位のリストから除き、論文・レポートで用いることを推奨する用語集として受講者に提供した。実際のセミナー授業での活用の様子なども併せて報告する。

２．文章指導における言語資源活用事例と本研究の目的

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

本研究では、特に大学初年次学生のアカデミックな表現に対する知識不足に対応するための教材を開発し、かつその教材が、教員・指導者の経験や内省によるものでなく、コーパスなどの根拠情報から定量的に得られるものとすることを目的としている。それにより、受講者の教材に対する納得感を向上させたいと考えている。

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

今回の発表では、このような教材開発の最初の試みとして、BCCWJ コアデータより動詞・名詞の頻度情報を抽出し、アカデミックな文に特徴的に使われると考えられる用語のリストを作成し、大学学部初年次のライティングセミナー授業で教材として活用した事例の報告を行う。

最近では、英語ライティングのテキスト「富岡(2012)」が発行されているが、これは上記の考えに基づくものである。British National Corpus に基づき、使用頻度順に 150 の動詞・助動詞を並べ使用例文と解説文を提示したものである。また「古泉　他(2011)」では、頻度情報を教育に応用した事例が紹介されているが、単語そのものを覚えることを主眼としたものであった。

３．頻度リストの作成方法

以下に教材として提示した動詞・名詞の頻度情報を作成した手順を説明する。

（１）BCCWJ コアデータ 白書データからの情報抽出

まず、BCCWJ コアデータのうち、CSV 形式のものを Excel2010 に読み込み、各種フィルタリング処理を行った。まず、比較的長い特徴的な単語を抽出するため、長単位情報を基に選択する。品詞情報が「動詞－一般」あるいは「名詞－普通名詞－一般」となっているもののみをそれぞれ抽出した。その単語リストの出現頻度を統計ツールで計算し、頻度順に並べ替えた。

（２）一般文でも利用される頻度が高いと判断される語のフィルタリング

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

（３）人手による用語選定と整形

上記のように機械的操作によって得られたリストには、大学生のアカデミックライティングにあまり用いることのない単語も含まれているので（元データが白書であるため)、最後に報告者(両名)が実際のライティング指導資料として適当と判断する語に絞り、用語表として学習者に提供した。動詞は上位３００語、名詞は上位１７６語のリストとなっている。

４．作成データのライティング指導への活用

作成した頻度データを、報告者（坂尻）が担当するライティング指導セミナー授業で教材として提供した。受講者には、ライティングの実践において口語的な表現を避けるための一つの方法として、あるいは、表現に迷った際の判断基準の一つとして、使うことを勧めた。また、とりあえずの使い方として、使いたい表現や迷う表現を 50 音で検索して頻度を確認し、国語研の Web ツール（少納言と NLB）を使って文脈での用法や出典（ブログ等か書籍等か）を参照することを提案してみた。

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

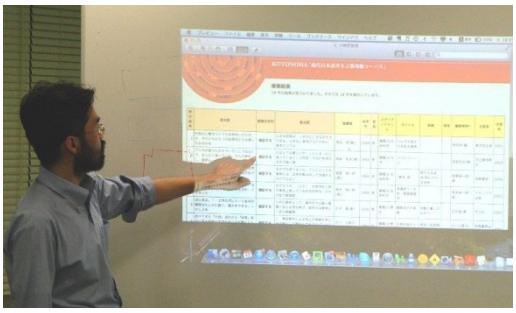


図 1 報告者（坂尻）が少納言の利用方法を担当セミナーで説明している場面

図 2　配布単語頻度データと少納言の説明を聞き、利用方法を学んでいる受講学生

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

図 3　報告者（坂尻）が NLB の操作法を説明している場面

本資料に基づくライティング授業に参加した受講学生（全員学部１年生）から意見を徴収した。７割ほどの受講生は、「(資料をライティングの際の参考に）使えそう」「使ってみたい」との意見であったが、「分からない」「使えなさそう」との意見もあった。さらに、「使い方や応用のポイントをより詳しく指導してほしい」あるいは「コーパスサイトとの併用の仕方を詳しく説明してほしい」などの意見もあったので、受講者のうち希望する者にはコーパスサイトを開いてのインストラクションを行った。その結果、説明を受けた学生は、納得しているようであった。

５．今後の展開

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

◎　対象コーパスデータの検討

今回の抽出は、試行としての時間的制約により、抽出対象を白書データのみとした。この抽出対象は語彙に偏りがあるなど、十分でないと認識している。まず、BCCWJ DVD データを本稿執筆時点で入手手続き中であり、入手後すみやかに対象データを DVD データとしたい。またアカデミックな文章に比較的近い表現（硬い文章）が多く含まれるであろうと判断し白書データジャンルを選定したが、BCCWJ の図書館データや教科書データの内、対象とすべきデータはまだ多数あると認識しており、適切な対象を追加選定したい。さらに BCCWJ だけでなく、Wikipedia 項目の内、学術的文章の参考になりうると判定できるものについては、対象としたいと考えている。

◎　特徴的な語・表現の抽出方法の改良

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

◎　作業手順のプログラム化

今回の頻度情報資料は Microsoft Excel を用い、手作業で抽出やフィルタリングを行った。今後作業対象コーパスの拡大を予定しており、できる限り早急に作業をプログラム化したいと考えている。

◎　資料インストラクション手法の改善

受講生に資料の有効活用法を説明する方法も改善が必要である。前述したとおり、受講生より、資料の活用法が良くわからないとの感想も得られている。より時間を掛け、頻度リストや関連 Web ツールを使用して、より良い文を選定する具体的な方法を、文章作成指導手順に組み込み提示したいと考えている。そのためのわかりやすい教材作りも必要だと考えている。

６．まとめ

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

謝　辞

本研究は、文部科学省　科学研究費補助金　基盤研究（B）課題番号:22320103「多言語会話文・語彙データベース構築と異文化交流におけるその活用に関する研究」（研究代表者：萬宮健策）による補助を得ている。

文　献

古泉隆、梁志鋭、阪上辰也、坂東貴夫、天野修一、新實葉子(2011)「日本語複合動詞を学ぶための Web 教材開発　－BCCWJ の頻度データに基づいて－」特定領域研究「日本語コーパス」平成２２年度公開ワークショップ（研究成果報告会）予稿集、pp.27-32

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

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

富岡龍明(2012)「コーパス活用　英語基本語を使いこなす　動詞・助動詞編」研究社

仁科喜久子　監修(2012)「日本語学習支援の構築」凡人社

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

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

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/

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

NINJAL-LWP for BCCWJ ホームページ　http://ninjal-lwp-bccwj.ninjal.ac.jp/

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

表 1　教材とした動詞頻度表の一部

国立国語研究所　書き言葉コーパス（BCCWJ コアデータ）より抽出した				
論文・レポートに役立つ動詞表現集（頻度順）				
２０１２年７月１３日　大阪大学　堀　一成、坂尻　彰宏				
動詞リスト番号	頻度	動詞	読み	頻度順位
1	779	する	スル	1
2	321	図る	ハカル	5
3	282	実施する	ジッシスル	6
4	233	推進する	スイシンスル	8
5	129	基づく	モトヅク	10
6	95	応ずる	オウズル	15
7	92	踏まえる	フマエル	16
8	91	増加する	ゾウカスル	17
9	86	活用する	カツヨウスル	19
10	83	利用する	リヨウスル	20
11	79	取り組む	トリクム	24
12	78	伴う	トモナウ	25
13	76	有する	ユウスル	28

表 2　教材とした名詞頻度表の一部

国立国語研究所　書き言葉コーパス（BCCWJ コアデータ）より抽出した				
論文・レポートに役立つ名詞表現集（頻度順）				
２０１２年７月１３日　大阪大学　堀　一成、坂尻　彰宏				
名詞リスト番号	頻度	名詞	読み	頻度順位
1	296	整備	セイビ	3
2	210	推進	スイシン	8
3	170	地域	チイキ	10
4	132	障害	ショウガイ	14
5	129	注	チュウ	16
6	126	図表	ズヒョウ	17
7	106	情報	ジョウホウ	23
8	99	支援	シエン	26
9	98	実施	ジッシ	27
10	93	資料	シリョウ	29
11	92	対象	タイショウ	30
12	89	推移	スイイ	33

テクストの結束性に与る語彙とその機能について

高崎みどり（お茶の水女子大学大学院人間文化創成科学研究科）

Characteristic Cohesive Feature of Vocabulary and its Function in the Text

Midori Takasaki (Ochanomizu University, Graduate School of Humanities and Sciences)

１．はじめに

本発表の目的は、語彙とテクストとの関係について、結束性と“談話構成語”の観点から、考察することである。

使用コーパスは、国立国語研究所「文章における語彙の分布と文章構造」プロジェクト（プロジェクトリーダー：山崎誠）作成の“学術入門書コーパス”から、『政治学入門』（阿部齊　岩波テキストブックス）、『日本外交史講義』（井上寿一　岩波テキストブックス）、『アメリカの経済　第２版』（春田素夫・鈴木直次　岩波テキストブックス）、『刑法原論』（内藤謙　岩波テキストブックス）、の４種、計 976 ページ、約 19 万 4 千字を使用した。

２．先行研究について

本発表で依拠する「結束性」と語彙との関わり、および「談話構成語」に関わる先行研究について簡単に触れる。

ハリディ／ハッサン（1997）では、「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。」（p.9）とし、「語彙的結束性」の言語体系における表示は「再叙[1]（語彙的指示の同一性）」や「コロケーション（語彙環境の類似）」によってなされる、とする。一方「文法的結束性」は、「指示、代用、省略、接続」によってなされる、とする。

次に「談話構成語」について、マッカーシー（1995）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語に注目する。「談話構成語　discourse- organizing words」と呼ばれる語で、“issue” ”problem” “dilemma”のような語がその例とされ、「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。」（p.106）と、説明されている。すなわち、テクストの中で使用されるとき、“issue（争点）”がテクストの内容のうちのどこの部分をさすのか、“dilemma（ジレンマ）”とは何と何とをさすのか、といったその語が及ぶ範囲が一種の“分節”となることの他、より大きなテクストパターン

[1] 「再叙（reiteration）」の内容は、同一語の繰り返し、同義語や近似同義語、上位語、一般名詞（people, stuff, move などのような一般的指示をもつ名詞類）、人称指示語、だという。「コロケーション（collocation）」は、「同じ語彙的環境を共有すること」を意味し、類似した文脈に現れる傾向のある 2 つの語彙項目あるいは「長い結束性の連鎖」が構築されることとなると述べる。

（問題―解決など）を示し、談話の全体像を予測させる働きを持つ、とする

３.「問題」という語についての「学術入門書コーパス」検索結果

さて、大量のコーパスを使用して、結束性について考察するときの方法として最も簡便な方法は、「再叙」のなかの同語反復に注目して、ある語を検索にかけてその出現の様相を追うという方法が考えられる。今回、談話構成語についても見たかったので、上記のマッカーシー（1995）などでもその例に挙げられている「問題」という語を選んでみた。

語彙的結束性という考え方と、談話構成語およびそれによる「分節」という概念を適用して、以下では「問題」で試行した結果を報告する。

「問題」という語そのものは、4 つの資料で計 729 回出現した。「××問題」「問題意識」「問題点」、「途上国債務問題」など、熟語や臨時一語[2]のような形で使用されている場合も含む。その中で『アメリカの経済』の第１章のコラム「“９・１１”テロの衝撃」での以下のような使われ方に注目した。少し長くなるが原文をしめす（見出し、段落における改行一字右寄せなどはそのままとした）。

例１　“９．１１”テロの衝撃

２００１年９月１１日朝，ニューヨークの世界貿易センター（ＷＴＣ），ワシントン郊外の国防総省，そしてペンシルバニア州西部での，ハイジャックされた４機の旅客機によるテロは，世界に大きな衝撃を与え，アメリカのみならず，多くの国々の政治・社会，国際政治を大きく変えることになった．経済的影響にかぎっても，直接の物的損害に加え，運輸・通信の阻害，波及する旅行・流通・生産への障害から，各種のテロ対策，さらにその後の軍事行動など，波紋は長期にわたる．ここでは，直後の経済**問題**①を紹介する．

最も直接的な打撃を受けたのは航空運輸であるが，当初の運行停止，再開後も警備強化による渋滞や旅行手控えによる旅客の減少，貨物の渋滞などにより，関連する経済活動への障害へと波及した．航空会社だけでなく航空機製造にも影響が及び，旅行業界が大打撃を受け，小売売上げも一時大きく落込んだ．保険会社は巨額の支払い**問題**②に直面し，製造業では部品調達の困難からＪＩＴ経営が弱点を露呈した．カナダ国境で陸上輸送の滞りも現れた．景気の下降はこの事件で決定的になったといってよい．しかし，マクロ経済への影響は短期的で，数週のうちに力強い回復が始まった．

もうひとつの大**問題**③は，ニューヨークの金融の中枢近くが破壊されたことによる直後の困難である．マンハッタン島南部のベライゾン（Verison）社の電話施設が使用不能になり，破壊されたＷＴＣビル内の証券会社数社と，金融市場取引決済銀行のひとつが業務不能になった．政府証券市場は翌々日まで，株式市場も翌週月曜まで開かなかった．近隣の諸機関の営業継続にとっては，Ｙ２Ｋ（２０００年**問題**④）対策でのバックアップ施設が救いの神になり，同時に危機管理の**問題**⑤点も明らかになった．

↑通信途絶による支払い連鎖切断への対策が急がれた．中央銀行当局は十分な流動性供給を行う準備があることを声明，攻撃翌日の貸出額は１日で４６０億ドルに及んだ．攻撃前の１週間の平均日額は２億ドルを下回る．その他，小切手取立て前入金扱い（フロート），海外ドル決済容易化のための外国中央銀行とのスワップ，市場再開後はオペでも，資金を供給

---

[2] 石井正彦（1999）「臨時一語と文章の凝縮」（『国語学』173　pp.91-104）等において、多く凝縮的文章においてみられる、その場限りで国語辞書にも登録されていない臨時的な語の組み合わせ、複合語をさす。

し，また，民間金融機関にも市中への柔軟な資金供給を要請して，金融メルトダウンを回避した．リスク管理上の**問題**⑥点は，バックアップサイトが被災施設に近すぎ，決済処理が過度に少数の機関に集中し，代替テレコム施設が無効（予備の複数業者がいずれもベライゾンの回線に依存），などであった．↑<『アメリカの経済』pp.36-37>

上記文章中には「問題」という語が 6 回使用されている（①～⑥）。これは同語反復で、この間には再叙としての結束性が生じている、とだけ考えれば、このテクスト引用部分を観察したことになるだろうか。これら 6 語はみな“同じ”ものであろうか。

①「問題」は「経済問題」のかたちでこれからその内容を述べることを予告して、波線部の範囲を“経済問題”として分節化している。その分節化の手掛かりとして、内部には、「打撃」「渋滞」「減少」「障害」「大打撃」「落ち込んだ」「困難」「弱点」「滞り」「下降」のように、語彙論的な関係性ではないが“（経済から見て）望ましくない状態＝問題”を上位概念とする、マイナスの意味を含んだ語彙が、テクスト内で臨時的な結束関係を結んでいる。いちおう、全部に波線が施してはあるが、実は正確に言えば、この分節化はこれらの中から、「問題」の具体的な内容となるものだけを、意味的にすくい取っているのだといえよう。

③「大問題」で、「もうひとつ」という限定がなされていることで、①の「問題」とは別の大きな問題のありかが予告され、同様に波線部の内容が分節化される。この分節内にも「困難」「不能」のようなマイナスの意味を含んだ語を、「問題」の具体的な内容としてすくい取ることができる。ここで注目すべきは、⑤⑥の「危機（リスク）管理（上）の問題点」が③「大問題」の中に部分的なもう一つの「問題」の分節として入れ込まれている（↑から↑までの部分）という、複雑な分節同士の関係になっていることである。

一方、同じ「問題」の語が反復されていても、②「問題」（支払い**問題**）と④「問題」（２０００年**問題**）は、対応するテクスト内の分節がなく、この部分だけに意味が留まっている。もし、この本の他の節や章、あるいは注など別のところに対応する分節があり、“見よ項目”や注番号が付随していれば、分節と対応しているということになり、談話構成語として働いている可能性もあるのだが。

ここで思い起こされるのは、「テクストは、結局、形式の単位ではなく、意味の単位である」（ハリディ／ハッサン　1991　p.157）という捉え方である。テクストを構成するものは、意味の分節であり、また、それらのテクストの意味を実現するべく構成された分節の組み合わせや包含関係である。そして長大な学術的テクストを読み込むには、段落ごとに“筆者の言いたいこと”をまとめるのでなく、いくつかの合図になるような語から、大小の分節を作り出し、対応させて、それらをイメージ化してプロットを構成することが必要になってくる。そのように考えると「問題」①③⑤⑥は、現に意味の分節に対応しているという合図である談話構成語として機能しているといえよう。

次に、別の角度からもう少し「問題」の出現箇所を追う。

「問題」の出現例として同書第 2 章第 4 節「製造業の国際競争力の低下」の中に「大量システムの限界」という項があって、“自動車産業を中心とする大量生産型の製造業の技術革新が遅れ、労働意欲の低下、生産効率の低下を引き起こした”という主旨で、その内容に沿った事象を色々と挙げている。その次の項でこの第 4 節の最後の項にあたる「『覇者のおごり』」の最初は

例２　「覇者のおごり」

最後に，<u>このような</u>**問題**⑦の背後にあった企業経営の**問題**⑧を指摘しておこう．（『アメリカの経済』p.59）

と始まっている。そして続いてその“企業経営の問題”が述べられていく。「このような」という指示語で前項「大量システムの限界」の全体を大きく受けて、“問題”として分節化しており、項を超えた分節化がおこっているといえる。「このような」が無ければ、どこを受けているのかすぐにはわからないので、「問題」が談話構成語として働いているのかどうかも明確にはならない。こうした指示語の機能を、仮に「談話構成補助機能」と呼ぶ。これについては後述する。

そして､指示語の談話構成補助機能によって大きく前の内容の分節を受け止めた⑦｢問題」を、続いてまた別の事柄（｢企業経営」）に収束させ、さらに別の方向に文脈を展開させようとして予告的に⑧「問題」を使用している。

ここで注意したいのは、①～⑥そして⑦⑧は、ただの「問題」という語の反復現象と片付けてはならないということである。すなわち上記で指摘したように、“問題”として「問題」の周囲にある語句を手掛かりにテクスト内の意味を分節化し、「問題」から別の「問題」へ繋げたり、さらに「問題」の中に「問題」という語を分節の入れ子のように使用したり、分節間の関係を作りつつ、事柄間の関係づけを果たし、知識の整理から理論化へ、いくつかの条件から結論への誘導へ等､学問的な思考過程形成そのままに談話構成する仕掛けとして利用されている。

こうした”合図になるような語”は、おそらく経験的には知られているのであろう。あるいは語自体でなく、使用法である場合もあって、たとえば、上記テクスト内では、辞書的に言えば「問題」という語の語義のブランチのうち、多く第一義とされる「答えを求めるための問い。解答や教えを要求する問い。質問。」(『日本国語大辞典』) ではなく、第二義以降の「批判や論争、または研究の対象となる事柄。解決しなければならない事柄」「心にとめて考えるべき事柄。注目すべき点」（いずれも『日本国語大辞典』）の方で、談話構成語として文脈を形成しているといえる。第一義は古語辞典類にも用例があるもので、現代でも「試験の問題は難しかった」のように使用されるのであり、「問題」という語がただちに談話構成語として働くとは言えない。それはあくまでも特定のテクスト内での出来事なのである。

こう見てくると、語彙論上の語同士の関係から分節が形成されるというよりも、談話構成語によって分節される範囲内に、その談話構成語を相対的上位語として、下位語同士が結ぶ結束関係が生じているとはいえないであろうか。そしてそれらの下位項目同士は、普通、語彙論的にいえば、ある品詞の上位語・同位語・下位語、また類義語等は同じ品詞であり、単純語・複合語のレベルも同じでなくてはならない、といった制限があるようだ。しかし実際にテクスト内で語どうしが結ぶ関係というのは、たとえば上記の例１で①「問題」が、「打撃」「渋滞」「減少」「障害」「大打撃」「落ち込んだ」「困難」「弱点」「滞り」「下降」といった語の表わす内容と、上位下位として関係づけられていたように、もっと自由であり、その都度的で、意外性を持ち、創造的であるといえよう。

４．「点」「動き」（名詞）「アプローチ」についての「学術入門書コーパス」検索結果

「点」「動き」（名詞）「アプローチ」の 3 語についても同様の調査をしたが紙幅の都合で略述する。

「点」については、4 資料で 539 回使用され、「焦点」「論点」「観点」「転換点」「係争点」等の熟語として使用される場合が多く、「～という点からみて」「この点、～」等の慣用語句的用法、「～のは、・・・点である」「・・・点は～である」等の文型として幅広く使用され、これらの形で、分節化に関与し、談話構成語となることも多かった。

典型的に談話構成に与る用法としては、

例３　資金循環の説明から外れるが、重要な政策問題として 2 点触れておこう。（『アメリカの経済』p.86）

として、「（１）」、「（２）」と箇条書きに、企業の健全性や住宅ローンについての課題が述べられている部分が、“重要な政策課題”と名付けられた分節となっている。

国語辞典の意味ブランチとの関連で言えば、やはり第一義の「大きさがなく位置だけをもつ図形」（『日本国語大辞典』以下同）といった幾何学的定義でなく、「さし示す事柄。箇所。」という意義が利用されている。しかしながら、第一義からの“広がりの無さ” “小ささ”が、簡単さ、あるいは絞られた集約的な感じを連想させ、論点の簡約化、集約化という談話構成を、言述のストラテジーとしたいときに使用されたのがきっかけかもしれない。

「点」については、マッカーシー（1995）の指摘に「すべての言語にそういう談話構成語があるのならば、教授／学習のプロセスにおいて、なんとか転移（transfer）を利用できないものか。他言語から、または他言語へ翻訳する場合、point（点）, argument（議論）, issue（争点）, fact（事実）のような語に直接対応する、信頼できる訳語があるのか。」（p.112）とあるように、第一義でない意味を談話構成語として使用するような場合は、翻訳語の影響もあるかもしれない。

和語「動き」は 94 回と少ないが、『アメリカ経済』に 85 回と集中し、経済に関する金利や株価、失業率、ＧＤＰなど夥しい統計データにおける数字の増減を「動き」と名付けて、その現象そのものや、現象の原因や意味を分析している範囲を分節化する。これも辞書の第一義でなく「一つに決まらないで、別の状態に変化すること。移り変わりのようす。変動。また、ひとの地位や仕事が変わること。異動。」の第二義の方が、分節を捉えやすい。

外来語「アプローチ」については、29 回しか使用されていない。専門分野の内容に関連して、というよりは、研究手法の説明等として使用されるが、他の 3 語のように、広い範囲の具体的な事象を分節化して談話構成の機能を果たすような場合は無く、この語は談話構成語としては未成熟であるのかもしれない。

以上、具体的な語を少しばかりみてきたが、既に述べたように、ハリディ／ハッサン（1997）では、「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。」（p.9）と言っている。これらの調査を通して、談話構成語と、そのカバーする分節とはその関係性において、結束関係にあると言ってもよいかと思われる。

長大なテクストを読み解くストラテジーとしては、個々の語彙のつくる結束性よりも、大きく談話構成する合図となるような語句と、それと意味的な関係を結ぶ分節との結束関係を追うことが必要かと思われる。

５．指示語の談話構成補助機能についての「学術入門書コーパス」検索結果

上記例 2 では「このように」という、前を大きく受ける指示語があるために「問題」の談話構成語としての機能がより明確になるということを述べた。指示は結束性の中で「文法的結束性」に属しているということからも、談話構成補助機能をもつことは納得できよう。

ここでは、さらに指示語の対象を広げて、不定指示ド系の指示語も合わせて見てみることとしたい。

『刑法入門』の第 3 章「犯罪現象の法的処理過程（１）」の中の１節を例にとる。

例４　しかし，実際には，１９９５年中の数字をみると，裁判が確定した者の総数は１０３万１７１６人であり，新たに行刑施設に収容されて懲役・禁錮・拘留の執行を受けた者は２万１８３２人である．だが，罰金の裁判が確定した者は，９６万７５１２人に達している．これに対して，無罪の裁判が確定した者は５２人で，簡易裁判所の略式手続による罰金・科料を含めた全事件裁判確定人員に対する無罪率は０．００５％にすぎない．有罪率は９９．９９５％に達する．通常第一審（地裁・簡裁）の公判手続による判決の場合をみても，無罪となった者は５６人で，判決人員総数６万１０２人に対する無罪率は０．０９％にとどまっている．後者の場合にも有罪率は９９．９１％に達する．

このように**罰金が多用されて，自由刑が現実に執行されることは比較的に多くないが，無罪となることは極度に少ないという解決にいたる**までに，犯罪現象の法的処理過程で，どのような**制度**がどのように**運用されている**のであろうか．本章と次章では，この**問題**を，公式統計と図２，図３（３９頁）を主な手がかりにして，制度運用の量的事実の側面を中心に検討することにしよう．（『刑法入門』p.26）

この例では「よう」を含むことで共通する「このように」「どのような」「どのように」の 3 つの指示語がかたまって出てきている。ちなみに「このような」は 4 資料合計 266 回、「このように」は 140 回、「どのような」は 81 回、「どのように」は 57 回使用されていた。

さて、「このように」は、前方のかなり広い状態部分を指し示す機能があるが、それだけで意味分節ができるほど繊細には働かない。その後の太字部分は「このように」と同格的であり、前方波線部の、多くの数字や専門用語が並ぶ記述内容を、1 つの過程として明確に端的に言い換えて意味を与え、分節化している。が、その分節化は前方状態指示「このように」がなければ、前を広く受けている保証が無く、談話構成としては、不安定なままとなる。

なお、太字部分の構造としては、談話構成語「解決（にいたる）」に連体修飾節が付されているものと考えて良いかと思う。

次の、「どのような**制度**がどのように**運用されている**のであろうか」に注目する。【ド系の指示語＋～疑問詞：か】というのは、読み手に対する働きかけ表現という面もあるが、「このような」と逆に後方に分節化が向かう機能に注目したい。この場合だと「制度」と「運用」について述べられている後方の部分までを分節化することになり、かつ後方でそれらが確実に述べられていることを保証し予告する。

加えてこの直後の「この問題」で「制度」と「運用」をひとまとめにして、「どのような～どのように～か」という疑問文を、【前方指示語「この」＋談話構成語「問題」】という名詞節に次元変換している。後方部分で述べられているということを予告し、かつ、「制度」・「運用」の分節としてその場所を特定化しているのが、下線部の後の「本章と次章」であり、

これはテクスト展開における現場指示的、メタテクスト的な表現である。そして同様に「公式統計と図２，図３（３９頁）を主な手がかりにして」は、後方の非文章テクストである図と、テクスト外の引用テクストである白書等の“公式統計”に関連付けられて、かなり大がかりに分節化することが予告されている。

すなわち不定のド系の指示語で投げかけられた語句は、その不定が特定の部分となって呼応し分節化が完了するまで、ずっとペンディングとなるという、かなり強力な談話構成機能を持っているといえるだろう。

６．終わりに

前後に意味の分節を作り、それと結束関係を結ぶ談話構成語について考察してきた。学術書の場合、専門用語がキーワードになりがちだが、「問題」のような、分節に意味付けを与えて、論をすすめる談話構成語をたどっていけば、学術的テクストを成立させているメカニズムが解明できるかもしれないと考えている。引き続き作業を続けたい。

テクストは生成過程であり、それを形成するのは意味の単位（＝分節）でしかない。その分節はしっかりとした固定的な存在でなく、現れたり消えたり、統合されたり、分岐したりする過程的な存在であるということもわかってきた。多岐にわたる複雑な大小の分節が入りこみ合って、それが思考過程を反映する形で、「結論」までもつれあっていくのである。“過程”に注目して観察していきたい。

一方、談話構成語は、読み手に対して、意味のある分節をここで作りなさいという合図になる。それどころか、談話構成語でまとめられていれば、いちいち前に戻って分節を作らなくても“前の方にはそんなことが書いてあったのか”と、まとめや要約の結果だけを受け取って先に進むことさえできるのだ。読み手にとっても、良き手掛かりとなるツールであろう。

今回、非常に長大なテクストを見たために、自分としては、今まで社説やコラム的な文章では気付かなかったことがいくつかわかったように思う。たとえば、語彙的結束性についても、ある分節の中には、語彙論的な語同士の関係による結束性よりも、その分節のもつ意味に関連する語彙が、臨時的につながって結束性の帯を作り上げているという場合をしばしば目にした。ひとつのテクストの中で、語彙の結束性が何重にも発揮され、指示・代用等の文法的結束性も加わって、たとえ何か見逃しても、分かりにくい表現があっても、他の結束性で十分カバーされ、必要な意味の構造が理解できるように用意されているのである。

本発表では、「問題」という語を中心に取り上げたが、夥しい語彙のうちの何が談話構成語として働くのか、まだまだわからない点が多い。高頻度で使用され、ある程度一般的な意味をもつ、あるいはあるテクストの中では、相対的に上位語で、一般的な意味をもつもの、ということはいえるかもしれない。今回の「問題」「点」「動き」「アプローチ」の他に「面」「角度」「情勢」のような談話構成語になりそうな語について、実際のテクストの中の使用形態と照合することも引き続き必要である。

談話構成語にともなう、主としてコ系の指示語は、代用というよりも、談話構成語が分節をつくることを助ける談話構成補助機能があることも確認した。一方、ド系の不定指示語は、「〜か」をともなうことで、テクストの後方へと向かって牽引力を発揮し、かなり強力な談話構成補助となりうることもわかった。指示語のすべてではないが、コ系（こう、こうして、この、こんな・・）やド系（どう、どうして、どの、どんな・・）が、談話構成語と組み合わさることで、より確実に談話構成の機能が実現できるのではないか、と思われる。

今回使用した、全体テクストとしての一冊の本まるごとのコーパスは、「はしがき」「あとがき」や見出し、目次、図・表、注、演習問題、索引等まで含めて観察でき、しかもきわめて大量のデータが、加工次第で一気にとれるので、テクスト分析にとっての興味深い現象が種々観察できる。コーパス言語学は、テクスト分析にとっても、今後の可能性を豊かに孕んでいる。

文献

M．A．K．ハリディ／R．ハッサン（1997）『テクストはどのように構成されるか』安藤貞雄他訳　ひつじ書房
M．A．K．ハリディ／R．ハッサン（1991）『機能文法のすすめ』筧寿雄訳　大修館書店
マッカーシー、マイケル（1995）『語学教師のための談話分析』安藤貞雄・加藤克美訳　大修館書店

＊辞典類　『日本国語大辞典』第2版　小学館

句末境界音調のピッチレンジに与える要因：『日本語話し言葉コーパス』の分析

五十嵐 陽介（広島大学）†
小磯 花絵（国立国語研究所理論・構造研究系）‡

Factors Affecting Pitch Ranges of Boundary Pitch Movements: An Analysis of the Corpus of Spontaneous Japanese

Yosuke Igarashi (Hiroshima University)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

１．はじめに

韻律句（prosodic phrase）末尾に生じる音調で発話の語用論的解釈（質問、継続、強調など）に貢献する音調を句末境界音調（Boundary Pitch Movement, BPM）という。どのようなBPMがいくつあるのかに関する研究や、BPMの機能に関する研究は古くからなされているが（川上 1963, 郡 1997）、BPMのピッチレンジ（pitch range）に関する研究はほとんどなされていない。

日本語の韻律句のピッチレンジを決定する主要な要因のひとつとして、アクセント句のピッチレンジを縮小させるダウンステップ（downstep）と呼ばれる現象が知られている（Pierrehumbert and Beckman 1988）。我々は以前の研究で（五十嵐・小磯 2012, Igarashi and Koiso 2012)、『日本語話し言葉コーパス』（前川 2004, 以降 CSJ）の分析に基づいて、BPMにダウンステップが観察されるか否かを検討した。その結果、BPMには、アクセント句の主要部（BPMを除いた部分）に観察されるダウンステップに類似した現象が観察されることが明らかになった。しかしながらこの研究では、話し方のスタイルや韻律句におけるBPMの位置などの要因が考慮されていなかった。

本研究の目的は、CSJの分析を通じて、BPMのピッチレンジに効果を与える要因を明らかにすることにある。第2節では、BPMの記述のために必要な諸概念を導入したのち、我々の以前の研究（五十嵐・小磯 2012, Igarashi and Koiso 2012）を中心に、BPMのピッチレンジに関連する研究を再検討する。第3節では用いたデータを記述する。第4節では分析結果を報告する。第5節で結果の考察を行い、結論を述べる。

２．日本語の韻律構造の記述

２.１　ダウンステップ

日本語におけるダウンステップとは、アクセント核（lexical pitch accent）が、それを含むアクセント句（以降 AP）に後続する APの基本周波数（F0）頂点（ピッチレンジの上限）を、反復的（iterative）に低下させる現象である（Pierrehumbert and Beckman 1988）。図1は「旨い飴がありました」（左）と「旨い豆がありました」（右）の音声波形と基本周波数（F0）曲線を示したものである。双方の発話とも、最初の AP（ウマ┐イ）はアクセント核を持つ有核句である。そのため2番目の AP（アメガ／マメ┐ガ）にダウンステップが生じ、F0頂点が低下する。一方、2番目の APは、左の発話ではアクセント核を持たない無核句（アメガ）であるのに対して、右の発話では有核句（マメ┐ガ）である。したがって、3番目の AP（アリマ┐シタ）にダウンステップが観察されるのは、右側の発話のみとなる。

† igarashi@hiroshima-u.ac.jp, ‡ koiso@ninjal.ac.jp

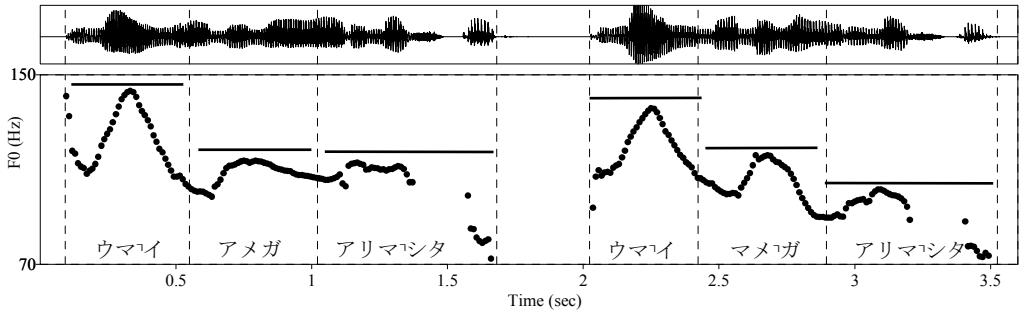


図 1　ダウンステップ.

左の発話は「旨い飴がありました」、右の発話は「旨い豆がありました」。縦の点線は AP 境界を表す。各 AP のピッチレンジの上限を水平方向の実線で示している。発話者は第 1 筆者。

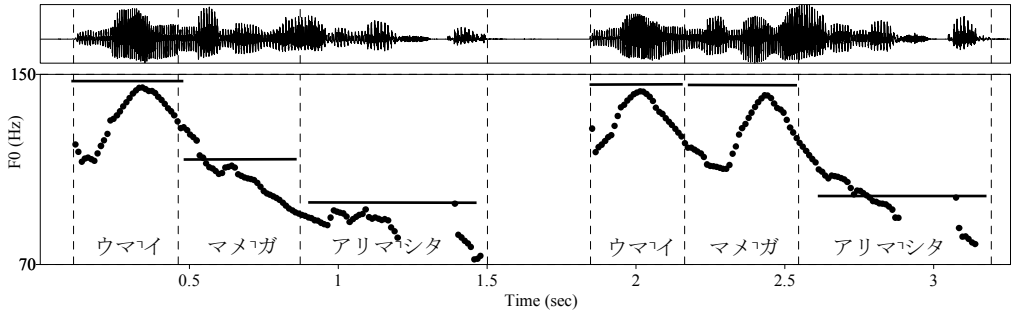


図 2　フォーカスによるダウンステップの阻止.

発話は左右とも「旨い飴がありました」。左の発話は「旨い」に、右の発話は「豆」にフォーカスが置かれている。縦の点線は AP 境界。水平方向の実線はピッチレンジの上限。発話者は第 1 筆者。

ダウンステップは、AP の F0 頂点を低下させるが、ピッチレンジの下限はあまり低下させない。このため AP の F0 変化幅は、有核句が先行することにより反復的に縮小することになる。

２.２　イントネーション句

ダウンステップの効果は、一定の統語構造やフォーカスによって阻止されることが知られている（Pierrehumbert and Beckman 1988; Venditti et al. 2008）。図 2 は「旨い豆がありました」の F0 曲線を示したものであるが、左の発話は「旨い」にフォーカスが置かれており、発話は「豆」にフォーカスが置かれている。左側の発話では 2 番目以降の AP にダウンステップが観察される。フォーカスは後続要素の F0 頂点をさらに低下させる効果があるため（post-focal prosodic subordination, cf. Venditti et al. 2008）、図 2（右）の発話と比較して、F0 頂点の低下の程度がより顕著になっている。

一方図 2（左）では、フォーカスを受けた語「豆」を含む AP（マメ┐ガ）の F0 頂点は、それに先行する AP（ウマ┐イ）の F0 頂点とほぼ同水準となっており、ダウンステップが観察されない。この現象を記述するために、Pierrehumbert and Beckman (1988)の韻律理論では、AP（マメ┐ガ）の始端に、AP より階層的に上位の韻律句の境界を仮定する。この韻律句は、CSJ が採用している日本語の韻律ラベリング体系である X-JToBI（Maekawa et al. 2002）およびその前身である J_ToBI（Venditti 2005）では、イントネーション句（intonation phrase, 以降 IP）と呼ばれおり、ダウンステップの生じる領域、あるいはピッチレンジが指定される領域として定義される。X-JToBI では、図 2 の発話に図 3 に示す韻律階層が仮定される。

図 3　図 1 の発話の韻律階層.

左側は「旨い」にフォーカスが置かれた発話、右側は「豆」にフォーカスが置かれた発話。AP より下位の韻律単位は省略してある。

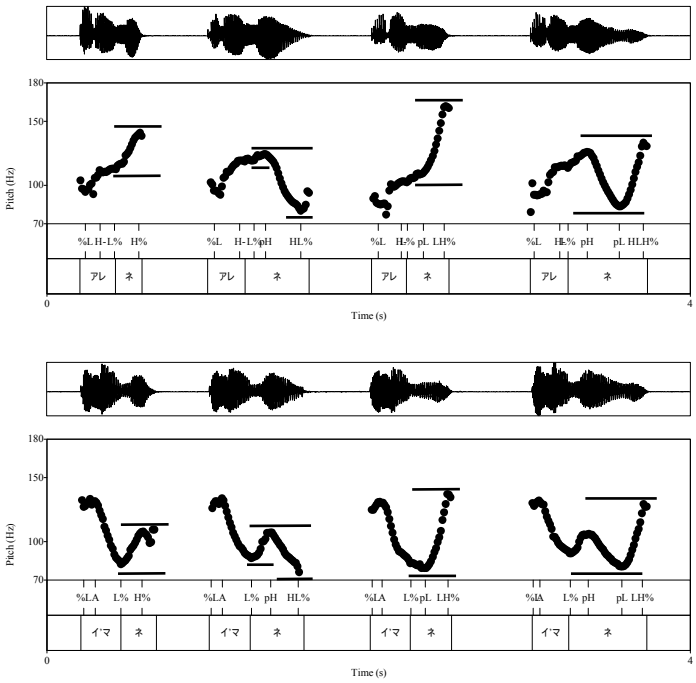


図 4　BPM.

上は無核句（アレネ）の句末に BPM を伴う発話（4 発話）、下は有核句（イ┐マネ）の句末に BPM を伴う発話（4 発話）である。BPM の種類は上下ともに左から H%, HL%, LH%, HLH%である。水平方向の実線によってピッチレンジの上限と下限を示している。HL%には下限が発話者は第 1 筆者。

２.３　句末境界音調（BPM）

日本語にどのような BPM がいくつあるかに関する研究は数多くあるが、見解の一致は得られていない（川上 1963, 郡 1997）。日本語の BPM の種類と数を確立することは重要な研究課題であるが、本研究ではこれに取り組まない。本研究では X-JToBI の枠組みに基づいて BPM を記述する。

X-JToBI では主要な BPM として図 4 に示す 4 種類が認められている。H%（上昇調 1）は F0 が単純に上昇するタイプの BPM であり、HL%（上昇下降調）は上昇の後に下降が生じるタイプの BPM である。LH%（上昇調 2）は、F0 が上昇する点は H%と同様であるが、上昇の前に低い F0 が一定時間継続する点が異なる。HLH%（上昇下降上昇調）は、上昇の後下降が生じ、その後さらに上昇が生じるタイプの BPM である（五十嵐他 2006）。

BPM は複数の F0 屈曲点によって特徴づけられるため BPM の物理的実現の分節音列上の生起時刻と F0 値を記述するためには、複数のトーンラベルが必要となる。この目的のために X-JToBI では、BPM のトーンラベルを分解し、複数の屈曲点の生起位置を記述している。この際、トーンラベルを単に分解しただけでは検索に支障をきたすので、分解されたラベ

ルの一部には補助記号を付与することにしている。4 種類の BPM それぞれの記述に用いられるトーンラベルと、そのラベルが記述する物理的なイベントは表 1 に要約されている。トーンラベルを用いた BPM のラベリング例は図 4 に示されている。

表 1 BPM ラベル

BPM	用いられるトーンラベルとそれが記述する物理的イベント
H%	L%（上昇開始点），H%（上昇終了点）
HL%	L%（上昇開始点），pH（下降開始点）， HL%（下降終了点）
LH%	L%（低 F0 区間開始点），pL（上昇開始点），LH%（上昇終了点）
HLH%	L%（上昇開始点），pH（下降開始点），pL（上昇開始点），LH%（上昇終了点）

２.４　BPM のピッチレンジ

BPM にダウンステップは観察されるのであろうか。我々はこの問いに答えるために、CSJ Core に含まれる 177 ファイル（約 39 時間）に生じる BPM のピッチレンジを計測した（五十嵐・小磯 2012, Igarashi and Koiso 2012）。その結果、BPM に先行するアクセント核の数が増えるほど BPM の F0 頂点が低下する現象（ダウンステップ）は、どの種類の BPM にも観察されることが明らかになったが、上昇幅・下降幅の縮小の有無や縮小の様相は BPM の種類ごとに異なることも明らかになった。各 BPM の結果は以下のように要約できる。

H%に関しては、先行アクセント核数の増加にしたがって、上昇の終点である F0 頂点だけでなく上昇の起点も同程度低下するため、上昇幅はほぼ一定に保たれることが分かった。一方 HL%に関しては、BPM の上昇部分と下降部分でピッチレンジ制御が異なることが明らかになった。上昇の起点は、H%の場合と同様に、先行アクセント核数の増加にしたがって反復的に低下するため、上昇幅はほぼ一定になることが明らかになった。一方、下降の終点は、アクセント核数がゼロの場合とそれ以外の場合に差が観察される以外は、ほぼ一定の値に保たれるため、下降幅は、先行アクセント核数の増加にしたがって反復的に縮小することが分かった。LH%には明確な傾向が観察されなかったが、これはこの BPM の総数が少なかったためであると考えられる。（HLH%は十分な数が得られなかったため分析の対象外とした。）

しかしながら、我々はこの研究で、話し方のスタイルや韻律句内での BPM の位置などの要因を考慮しなかった。これらの要因は BPM のピッチレンジ制御に効果を与える可能性がある。前川（2011）によると、BPM の種類の選択や韻律句内での BPM の位置は、話し方のスタイルと強く相関している。前川（2011）は CSJ の BPM の出現頻度を分析し、あらたまった発話には H%が、くだけた発話には HL%が相対的に多く出現することを明らかにした。また前川（2011）は CSJ に含まれる「学会講演」「模擬講演」等の別（以降これをレジスターと呼ぶ）を分離する変数を検討し、その変数に BPM に関連する特徴が含まれていることを明らかにした。ひとつは BPM の種類である。「学会講演」は H%（および PNLP, EUAP[1]）が多いことによって特徴付けられ、「模擬講演」は HL%が多いことによって特徴付けられる。もうひとつは BPM の生じる位置である。BPM は IP の最後に位置する AP 末に生じることが多いが、それ以前の AP の末尾に生じることもある。「学会講演」は IP 末以前に BPM が生じる現象が多いことによって特徴付けられ、「模擬講演」は、そのような現象が少ないことによって特徴付けられる。

前川（2011）の分析結果は、特定のスタイルに強く結びついたイントネーションパタンが存在していることを示している。そのパタンを規定する特徴に、これまで明らかにされてきた BPM の種類や、BPM の出現する位置だけでなく、ピッチレンジを含めた BPM の音声実現の違いが含まれている可能性を検討してみることは有益であるように思われる。

[1] PNLP と EUAP については五十嵐他（2006）参照。

３．データと計測

データとして使用したのは CSJ Core に含まれる「学会講演」70 ファイル（19 時間）と「模擬講演」107 ファイル（20 時間）である。以下、用いたデータと計測方法ついて詳述する。

３.１　分析対象

4 種類の BPM のうち LH%（286 件）と HLH%（7 件）は数が少ないので除外した。先行アクセント核数（3.2 参照）が 6 以上の BPM は除外した。AP の次末モーラ以前から上昇を開始する BPM の変種も対象から除外した。H%には、F0 上昇終了後に同水準の F0 値が持続される変種が観察されるが、今回の分析ではこの変種と通常の変種との区別はしなかった。その結果、H%（21794 件）、HL%（6820 件）の計 28614 件の BPM が分析対象となった。

３.２　先行アクセント核数の計算法

先行アクセント核数は、当該 BPM に先行するアクセント核で、かつ当該 BPM が生じた IP に含まれるアクセント核の数と定義した。たとえば、[(ナ┐オヤワ)][(ダ┐レト)（ノミ┐ヤデ）（ノ┐ンダノ)]「直也は誰と飲み屋で飲んだの？」（“[ ]”は IP 境界を“( )”は AP 境界を表す）という発話の末尾に BPM が生じた場合、その BPM の先行核数は 3 となる。

３.３　話し方のスタイル

概して CSJ の「学会講演」は比較的あらたまったスタイルに特徴付けられ、「模擬講演」は比較的くだけたスタイルに特徴付けられる（前川 2011）。本研究では「学会講演」と「模擬講演」の別、すなわちレジスターの別をスタイルを表す変数として代用する。

３.４　韻律句における BPM の位置

BPM を有する AP が IP の終末に位置するか否かを、韻律句における BPM の位置を表す変数（IP 非終末と IP 終末の 2 水準）として用いた。BPM を有する AP が IP の終末に位置するか否かは、X-JToBI の Break Indices (BI)のラベルを参照することで一義的に決定できる。

３.５　F0 値の抽出

F0 値は CSJ の XML ファイルに記録されている F0 値（Hz）を利用した。原則として X-JToBI のトーンラベルひとつにひとつの F0 値が与えられているが、母音の無声化等の理由により F0 値が与えられていない場合もある。このような欠損値は分析対象から除外した。性差や個人差の影響を最小限にするために、F0 値は談話ごとに Z スコアに変換した。

BPM のピッチレンジに関する F0 値の抽出方法は BPM の種類ごとに異なる。H%の場合は上昇の起点（「上昇起点」と呼ぶ）と上昇の終点（「頂点」と呼ぶ）の値を計測した。上昇起点として[L%]の持つ値を採用し、頂点として[H%]の持つ値を採用した（ラベル付与位置については図 4 参照）。以降、上昇起点と頂点の差を「上昇幅」と言及する。

一方 HL%の場合は、上昇部分の起点（「上昇起点」と呼ぶ）と終点（「頂点」と呼ぶ）に加えて下降部分の終点（「下降終点」と呼ぶ）を計測した。上昇起点として[L%]の持つ値を採用し、頂点として[pH]の持つ値を採用した。下降終点として[HL%]の値を採用する。以降、上昇起点と頂点の差を「上昇幅」、頂点と下降終点の差を「下降幅」と呼ぶこととする。

４．分析

４.１　先行アクセント核とスタイルの効果

はじめに、BPM のピッチレンジに対する先行アクセント核数の効果がスタイルごとに異なるか否かを検討した。各条件に該当する BPM の件数を表 2 に示す。件数が 100 未満の BPM（表中の網掛け部分）は以後の分析から除外した。H%のピッチレンジの分析結果を図 5 に、HL%のピッチレンジの分析結果を図 6 に要約する。以降、図中のエラーバーはすべて標準偏差を表す。

BPM の種類の違いに関わらず、頂点、上昇起点、下降終点が先行アクセント核数の増加にしたがって単調に低下する傾向がほぼ一貫して観察される。この傾向は学会講演、模擬講演の両レジスターに認めることができる。HL%の下降幅も両レジスターにおいて先行アクセント核数の増加にしたがって反復的に縮小する傾向が観察される。

最も顕著なレジスターの効果は上昇幅に指摘することができる。学会講演では、上昇幅は両 BPM ともに先行核数の増加にしたがって拡大する。この上昇幅の拡大は模擬講演では観察されない。模擬講演における H%の上昇幅は下降傾向にあり、HL%の上昇幅はほぼ一定に保たれる。このレジスター間の差は、先行核数の増加による頂点の低下の程度（ダウンステップの程度）にその原因を求めることができる。模擬講演では頂点も上昇起点も同程度に低下するのに対して、学会講演では上昇起点の低下の程度よりも頂点の低下の程度が緩やかである。

表 2　先行アクセント数とレジスターの別から見た BPM の件数.

		先行アクセント核数					
レジスター	BPM	0	1	2	3	4	5
学会講演	H%	2203	7102	3613	991	173	23
	HL%	214	846	479	119	38	10
模擬講演	H%	1138	3878	1995	599	102	17
	HL%	634	2544	1457	415	74	12

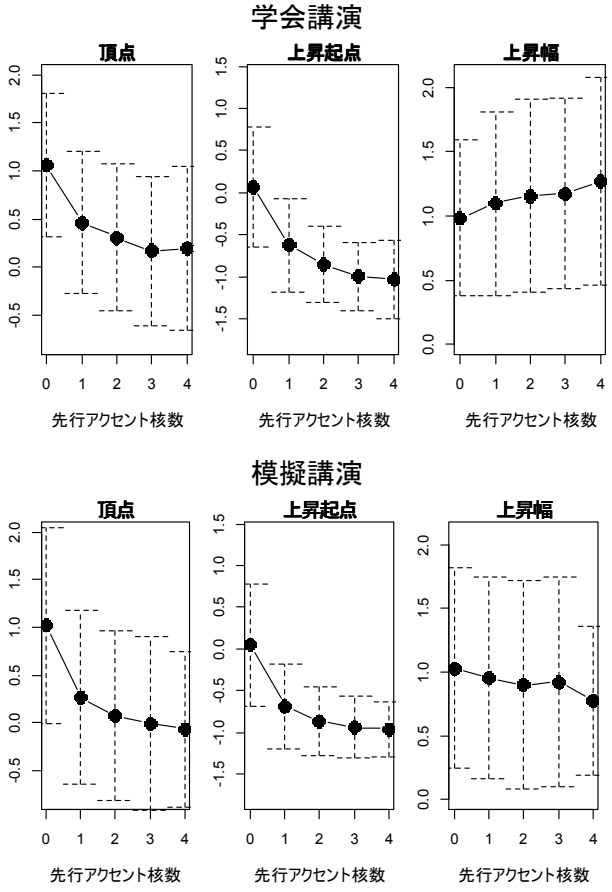


図 5　H%のピッチレンジ：レジスター別．学会講演（上）、模擬講演（下）.

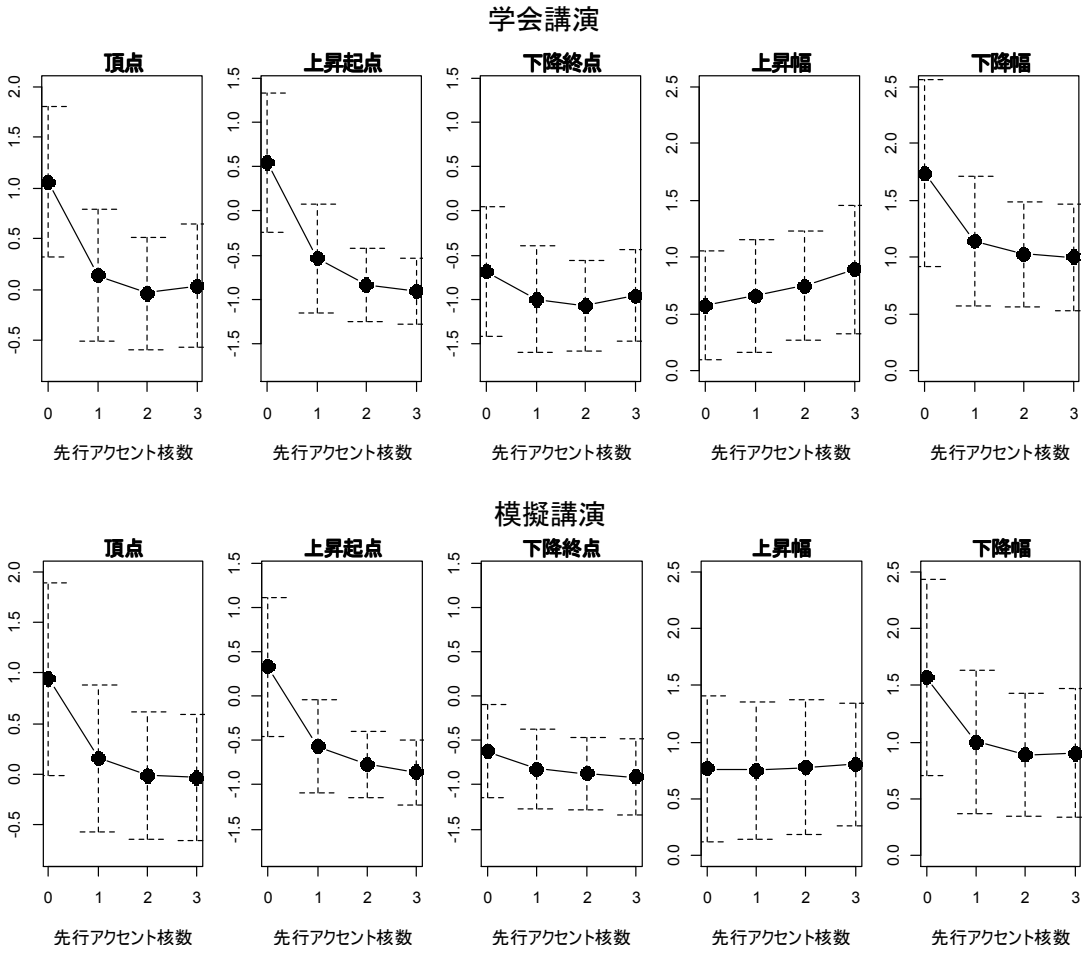


図 6　HL%のピッチレンジ：レジスター別．学会講演（上)、模擬講演（下）.

４.２　IP における BPM の位置の効果

次に IP における BPM の位置（非終末/ 終末）が BPM のピッチレンジに効果を与えるか否かを検討した。先行アクセント核数とレジスターの双方が影響することがすでに分かっているので、この 2 つの要因の効果も同時に検討した。各条件に該当する BPM の件数を表 2 に示す。件数が 100 未満の BPM は以後の分析から除外した。IP における BPM の位置の別（非終末/終末）だけが異なり、他の要因の値が同一である条件のペアは、両条件における BPM の件数が 100 件以上でない限り、以降の分析から除外した。ただしレジスターが学会講演でかつ先行核数が 0 である条件のペアは、一方の条件に該当する BPM の件数が 100 未満であるが、分析対象とした。表中の網掛け部分は分析対象外とした BPM を表す。

表 3　先行アクセント数とレジスターの別から見た BPM の件数.

			先行アクセント核数					
レジスター	BPM	IP 内の位置	0	1	2	3	4	5
学会講演	H%	非終末	1098	3061	973	207	37	7
		終末	1105	4041	2640	784	136	16
	HL%	非終末	61	151	32	10	1	0
		終末	153	695	447	109	37	10
模擬講演	H%	非終末	479	1260	335	69	8	2
		終末	659	2618	1660	530	94	15
	HL%	非終末	136	405	77	7	5	0
		終末	498	2139	1380	408	69	12

H%のピッチレンジの分析結果を図 7 に、HL%のピッチレンジの分析結果を図 8 に要約する。図から明らかなように、IP における位置の効果が観察される。「学会講演」および「模擬講演」における先行核数 0 の H%の上昇幅、「学会講演」における先行核数 0 の HL%の下降幅を除いて、すべての特徴（頂点、上昇起点、下降終端、上昇幅、下降幅）は、IP 非終末より IP 終末場合の方が低く実現される。この効果は BPM の種類の違い、レジスターの違いに関わらず観察される。

先行核数の効果およびレジスターの効果は 4.1 で認められたものと同様のものが観察される。すなわち、先行核数が増加するにしたがって、1) 両レジスターにおいて H%と HL%の頂点、上昇起点、下降終点が単調に低下し、2) 両レジスターにおいて HL%の下降幅が縮小するが、3) 模擬講演では H%, HL%の頂点と上昇起点とが同程度に低下するのに対して、学会講演では上昇起点の低下の程度よりも頂点の低下の程度が緩やかであるため、4) H%と HL%の上昇幅は学会講演においては単調に拡大するが、模擬講演においては拡大することはない。これらの効果は、IP における BPM の位置の違いに関わらず観察される。

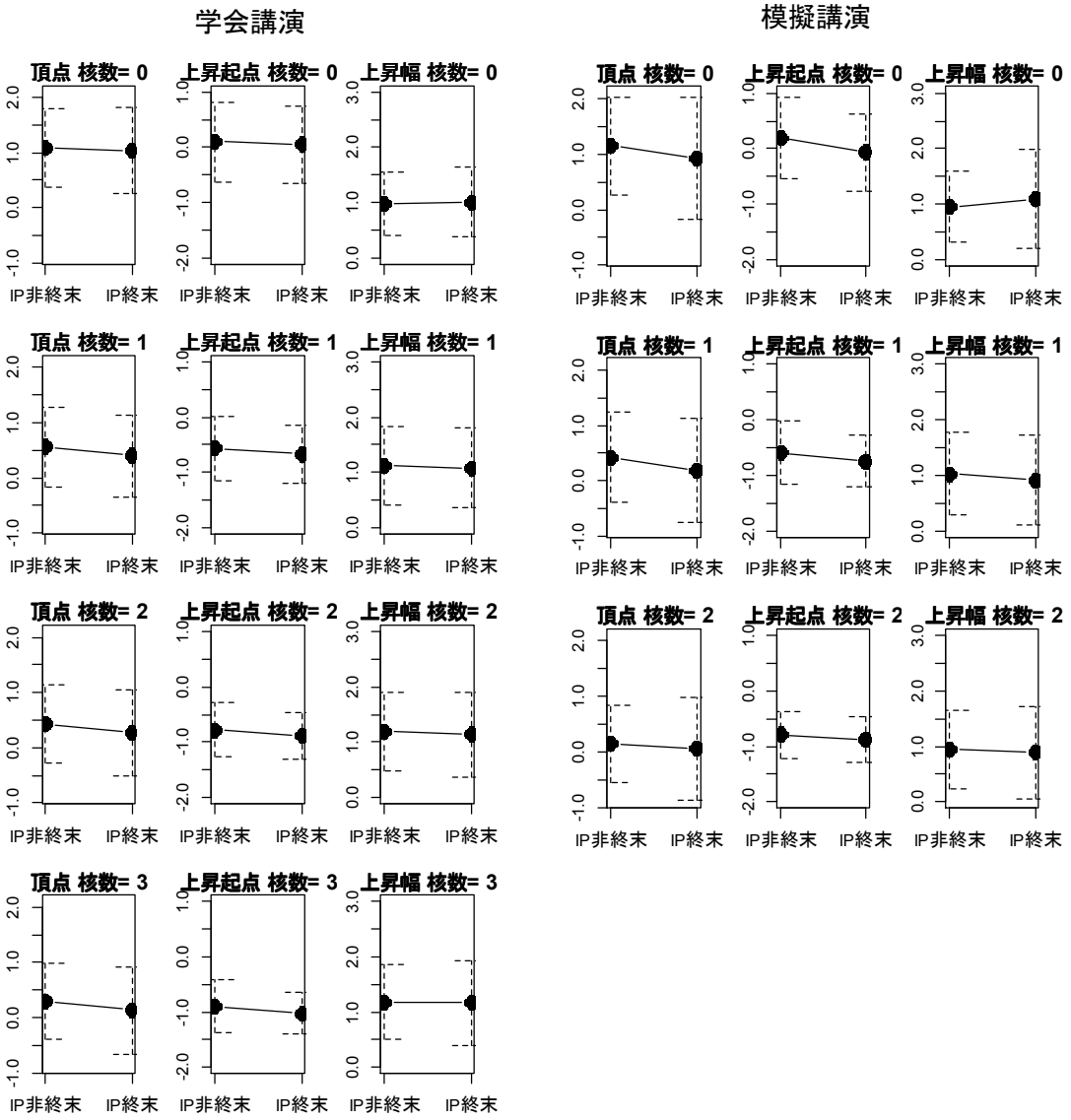


図 7　H%のピッチレンジ : スタイル別および BPM の IP 内の位置（IP 非終末/ IP 終末）別. 学会講演（左）、模擬講演（右）.「核数」は先行アクセント核数を意味する。

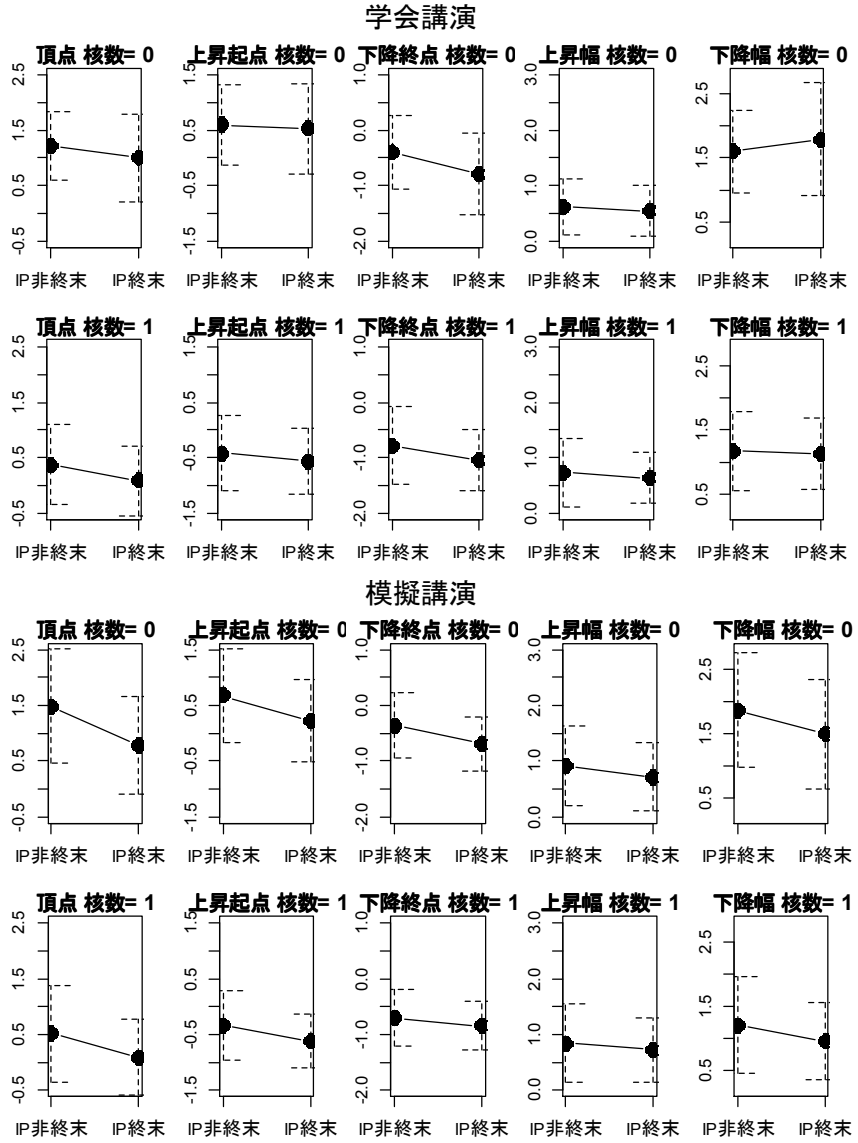


図 8　HL%のピッチレンジ：スタイル別および BPM の IP 内の位置（非終末/ 終末）別.
学会講演（上)、模擬講演（下).「核数」は先行アクセント核数を意味する。

５．考察と結論

今回の分析結果によって、話し方のスタイルと、韻律句における BPM の位置の双方が、BPM のピッチレンジに効果を与えることが明らかになった。スタイルの差は、ダウンステップの抑制として現れる。「学会講演」「模擬講演」の双方において、先行核数が増えるにつれて、BPM の上昇起点および頂点が低下するが、模擬講演では上昇基点も頂点も同程度に低下するのに対して、学会講演では上昇起点の低下の程度よりも頂点の低下の程度が緩やかである。その結果、「学会講演」では先行核数の増加とともに上昇幅が単調に拡大する。このダウンステップの抑制は、IP における BPM の位置の違いに関わらず観察される。

前川（2011）は、「学会講演」の発話者の一部は、講演前に周到な練習を重ね発話すべき内容を大部分暗記しているため、通常よりも顕著に長い発話のプランニングを行っている可能性があり、このことが特定のイントネーションパタン（具体的には IP 非終末に現れる BPM）が「学会講演」に頻出する原因となっている可能性を指摘している。BPM のダウン

ステップが抑制される理由も発話のプランニングに見つけられるかもしれない。

BPM のダウンステップが抑制されうるという事実は、日本語のピッチレンジ制御を扱う適切なモデルを構築するための示唆を与える。我々がこれまで明らかにしてきたように、BPM の頂点は、それに先行するアクセント核によって反復的に低下するが、この事実は、AP 主要部のピッチレンジ制御と BPM のピッチレンジ制御が密接に結びついていることを示している。その一方で、先行アクセント核による BPM の頂点の低下が、あるスタイルにおいて抑制されうる事実は、第 1 に BPM のピッチレンジ制御が AP 主要部のピッチレンジ制御からある程度独立していることを意味し、第 2 に BPM のピッチレンジはスタイル等の要因によって比較的自由に（そしておそらくは連続的に）変動させられうることを意味する。今後は、BPM をも組み込んだピッチレンジの理論を構築することが必要である。

韻律句における BPM の位置の効果は、（先行アクセント核数が同じにも関わらず）IP 末の AP に生じる BPM の F0 が、IP 末の AP 以前に生じる BPM の F0 より低くなるというものである。この効果はスタイルの違いに関わらず観察される。IP より大きな韻律的単位である発話の終末で F0 が局所的に低下する final lowering という現象が知られているが（前川 2011）、IP 末における BPM の F0 低下が final lowering と関連した現象なのか否かを検討するのは今後の課題である。また、統語的・意味的な切れ目の強さが BPM の出現率や音声実現に影響することが報告されているので（小磯 2012）、IP 末における BPM の F0 低下を統語的・意味的な切れ目の強さと関連付けて検討することが、BPM のピッチレンジを明らかにするための有望なアプローチのひとつとなると言えるだろう。

付　記

本研究は、国立国語研究所（言語資源研究系）基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）および萌芽・発掘型共同研究「会話の韻律機能に関する実証的研究」（リーダー：小磯花絵）による補助、および文部科学省科学研究費補助金・若手研究（B）「イントネーションの音韻論と音声学を峻別する実験手法の確立」（研究代表者：五十嵐陽介）による補助を得ています。

文　献

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

五十嵐陽介、小磯花絵（2012）「日本語話し言葉コーパスにおける句末境界音調のピッチレンジ制御」『第 1 回コーパス日本語学ワークショップ予稿集』, pp. 355-364.

川上蓁（1963）「文末などの上昇調について」『国語研究』16, pp. 25-46.

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

郡史郎 (1997)「日本語のイントネーション一型と機能」杉藤美代子(監), 国広哲弥他(編)『日本語音声 2：アクセント・イントネーション・リズムとポーズ』, pp. 169-202, 三省堂

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（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

Igarashi, Yosuke and Hanae Koiso (2012) Pitch range control of Japanese boundary pitch movements, To appear in Proceedings of Interspeech 2012, Portland, Oregon.

Maekawa, K., H. Kikuchi, Y. Igarashi and J. Venditti (2002) X-JToBI: An extended J_ToBI for spontaneous speech, *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 1545-1548, Denver, Colorado.

Pierrehumbert J. and M. Beckman (1988) *Japanese Tone Structure*, Cambridge: MIT press.

Venditti, J. (2005) The J_ToBI model of Japanese intonation, In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*.172-200, New York: Oxford Univ. Press.

Venditti, J., K. Maekawa, and M.E. Beckman (2008). 'Prominence marking in the Japanese intonation system', in S. Miyagawa, and M. Saito (eds.), *Handbook of Japanese Linguistics*, pp. 456-512, New York: Oxford University Press.

ポスター発表 (1)

9月6日 (木)　15:00～17:00

『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

小西 光[*]・浅原 正幸[†]・前川 喜久雄[‡]
（国立国語研究所 コーパス開発センター）

Temporal Information Annotation on Balanced Corpus of Contemporary Written Japanese

Hikari Konishi, Masayuki Asahara and Kikuo Maekawa
(National Institute for Japanese Language and Linguistics)

１．はじめに

情報検索や情報抽出において、テキスト中に示される事象を実時間軸上のある時区間もしくは時点に関連づけることが求められている。Web 配信されるテキスト情報に関しては、文書作成日時(Document Creation Time: DCT) が得られる場合、テキスト情報と文書作成日時とを関連づけることができる。しかしながら、文書作成日時が得られない場合や、文書に記述されている事象が起きる日時が文書作成日時と乖離する場合には他の方策が必要である。テキスト中に記述されている時間情報解析の精緻化が求められている。

時間表現抽出は、固有表現抽出の部分問題である数値表現抽出のタスクとして研究されてきた。英語においては、評価型国際会議 MUC-6 (the sixth in a series of Message Understanding Conference) (R. Grishman and B. Sundheim 1996) で、タグづけ済み共有データセットが整備され、そのデータを基に各種の系列ラベリングに基づく時間表現の切り出し手法が開発されてきた。TERN (Time Expression Recognition and Normalization) (DARPA TIDES 2004) では、時間情報の曖昧性解消・正規化がタスクとして追加され、様々な時間表現解析器が開発された。さらに、時間情報表現と事象表現とを関連づけるタグづけ基準 TimeML (J. Pustejovsky et al. 2003b) が検討され、TimeML に基づくタグつきコーパス TimeBank (J. Pustejovsky et al. 2003a) などが整備された。2007 年には、時間情報表現－事象表現間及び 2 事象表現間の時間的順序関係を推定する評価型ワークショップ SemEval-2007 におけるサブタスク TempEval (M. Verhagen et al. 2007) が開かれ、種々の時間的順序関係推定器が開発された。後継のワークショップ SemEval-2010 におけるサブタスク TempEval-2 (M. 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、フランス語を含めた 6 言語が対象となった。

一方、日本語においては IREX (Information Retrieval and Extraction Exercise) ワークショップ (IREX 実行委員会 1999) の固有表現抽出タスクの部分問題として時間情報表現抽出が定義されているのみで、時間情報の曖昧性解消・正規化に関するデータが構築されていなかった。我々は TimeML に基づいた日本語に対する時間情報タグづけ基準を定義し、時間情報の曖昧性解消・正規化を目的とした時間情報タグつきコーパスを構築した。本稿ではタグづけ基準を示すとともに、タグづけしたコーパスの詳細について示す。

２．対象とする時間情報表現

まず以下の例文を見て欲しい。

彼は 2008 年 4 月から週に 3 回ジョギングを 1 時間行ってきたが、昨日ケガをし

[*] hkonishi at ninjal.ac.jp
[†] masayu-a at ninjal.ac.jp
[‡] kikuo at ninjal.ac.jp

て走れなくなり、今朝 9 時に病院に行った。

本稿の研究対象である時間情報表現[1]は時間軸上の時点もしくは時区間を表現するテキスト中の文字列とする。時間情報表現は以下の 3 つの分類に分けられる。①日付表現(“DATE”,相当)・②時刻表現(“TIME”相当)は「2008 年 4 月」「昨日」「今朝 9 時」といった、時点及び時区間の時間軸上の位置を定義することを目的として用いられる表現である。③時間表現(“DURATION”相当)は「1 時間」といった、時間軸上の位置に焦点をあてずに時区間幅を定義することを目的として用いられる表現である。④頻度集合表現(“SET”相当)は「週に 3 回」といった、時間軸上複数の時区間を定義することを目的として用いられる表現である。

曖昧性を解消しながら時間情報表現を時間軸上の特定の区間に写像することを正規化と呼ぶ。日付・時刻表現において、表層の情報だけで正規化ができる表現と、文脈の情報を用いなければ正規化ができない表現がある。前者を定時間情報表現(fully-specified temporal expression) と呼び、後者を不定時間情報表現(underspecified temporal expression) と呼ぶ。上の例では「2008 年 4 月」が定時間情報表現であり、「昨日」「今朝 9 時」が不定時間情報表現である。時間情報表現の正規化には計算機で扱う日付や時刻を扱うための国際標準 ISO-8601 形式 [2]への変換が一般的である。しかしながら、自然言語では表現できるが、ISO-8601 形式では直接表現できない時間情報表現がある。例えば、時間表現や頻度集合表現は時間軸上不定な場合が多く ISO-8601 形式だけでは表現できず、方策が必要である。

想定する時間情報表現解析では、手がかりとしてテキストが書かれた日付・時刻を表す文書作成日時を用いることを仮定している。例えば、文書作成日時が 2008 年 9 月 1 日であれば、「昨日」は 2008 年 8 月 31 日(ISO-8601 形式では“2008-08-31”)を表し、「今朝 9 時」は 2008 年 9 月 1 日午前 9 時(同“2008-09-01T09:00”)を表す。

３．TimeML〈TIMEX3〉タグに基づいた時間情報タグづけ基準の概略

本節では日本語時間情報表現に対するタグづけ基準の概略を示す。タグづけ基準は、言語資源管理に関する国際標準 ISO/TC 37/SC 4[3]において 2009 年に採用された ISO 24617-1 (SemAF/Time)の基になっている TimeML (J. Pustejovsky et al. 2003b) ⟨TIMEX3⟩ タグの仕様に準拠している[4]。以下、日本語の例を用いながら⟨TIMEX3⟩のタグの仕様を説明する。細かな点で日本語に合うように変更しているがタグ名は⟨TIMEX3⟩をそのまま利用している。

３.１　タグづけ対象

タグづけ対象は日付表現(“DATE”)・時刻表現(“TIME”)・時間表現(“DURATION”)・頻度集合表現(“SET”)の 4 種類である。図 1 にタグづけ事例を示す。

日付表現は「一九二九年二月」「前日」のような日暦に焦点をあてた表現である。時刻表現は「午前十時ごろ」「午後六時ごろ」「昼」「九日昼」のような 1 日のうちのある時点に焦点をあてた表現である。日付表現と時刻表現の区別は時間軸上の粒度の区別でしか

---

[1] 「時間情報表現」は①「日付表現」(“DATE”)・②「時刻表現」(“TIME”)・③「時間表現」(“DURATION”)・④「頻度集合表現」(“SET”)の 4 種類を包含するものを指す。

[2] 日付や時刻を YYYY-MM-DDThh:mm:ss などといった数値と記号列で表記する標準。YYYY は年を表す 4 ケタの数字が、MM は月を表す 2 ケタの数字が、DD は日を表す 2 ケタの数字が、hh は 24 時間制で時刻を表す 2 ケタの数字が、mm は分を表す 2 ケタの数字が、ss は秒を表す 2 ケタの数字が入る。様々な略記方法が提案され、例えば「2008 年 4 月」は "2008-04" と表記する。詳細については ISO-8601 に対応する日本工業規格 JISX0301 「情報交換のためのデータ要素及び交換形式 ― 日付及び時刻の表記」参照のこと。

[3] http://www.tc37sc4.org/

[4] 2003 年の TimeML と区別するために ISO 24617-1 の基準を ISO-TimeML と呼ぶ。

ない。便宜上不定の現在を表す「今」という表現を時刻表現に分類する。時間表現は「その間」のような時間軸上の両端に焦点をあてておらず、期間を表すことに焦点をあてている表現である。頻度集合表現は「毎日」のような複数の日付・時刻・時間に焦点をあてた表現である。この分類は、解析の方便のために導入したものである。時間軸上一つもしくは複数の時点・時区間を表現するものをタグづけ対象である時間情報表現とする。

現在のタグづけ基準では⟨TIMEX3⟩タグの入れ子を許さない。日付・時刻表現の線形結合はこれを一つの日付・時刻表現として切り出す。例えば「九日昼」のように日付表現と時刻表現が連接する場合には一つの時刻表現として切り出す。時間を表す際に、開始時点と終了時点を明示している場合には、開始時点と終了時点とを別々の日付・時刻表現として切り出す。例えば「午前十時ごろから午前六時ごろまで」は一つの時間表現として切り出さず、「午前十時ごろ」と「午前六時ごろ」の二つの時刻表現として切り出す。事象が起こる期間を表すために、今後、関連する事象表現に対し、この二つの時刻表現への参照関係を付与する予定である。頻度集合表現は、文字列上できるだけ短い単位を切り出す。例えば「毎日」を頻度集合表現として切り出すが、「毎日午前十時ごろから午前六時まで」は現在のところ頻度集合表現として切り出していない。

⟨sentence type="quasi"⟩⟨TIMEX3 tid="t1" type="DATE" value="2003-10-20" valueFromSurface="2003-10-20"⟩二〇〇三年十月二十日⟨/TIMEX3⟩　⟨TIMEX3 tid="t2" type="DATE" value="2003-W43-1" valueFromSurface="XXXX-WXX-1"⟩月 曜日 ⟨/TIMEX3⟩⟨/sentence⟩ ⟨br type="automatic_original" /⟩ ⟨sentence type="quasi"⟩ ⟨TIMEX3 tid="t3" type="TIME" value="2003-10-20T17:30:XX" valueFromSurface="XXXX-XX-XXT17:30:XX"⟩ 午後五時三十分⟨/TIMEX3⟩⟨/sentence⟩ ⟨br type="automatic original" /⟩ ⟨blockEnd /⟩ ⟨paragraph⟩ ⟨sentence⟩　ステイシーはだらけた姿勢でモニターの前に陣取り、白黒の画像に見入っていた。⟨/sentence⟩ ⟨sentence⟩ 彼女は伸びをし、腕時計に目をやった。⟨/sentence⟩ ⟨sentence⟩⟨TIMEX3 tid="t4" type="DURATION" value="PT2H30M" valueFromSurface="PT2H30M"⟩二時間半⟨/TIMEX3⟩ で収穫ゼロ。⟨/sentence⟩

図 1 タグづけ例　（PB59_00001）

３.２　⟨TIMEX3⟩ の属性

⟨TIMEX3⟩ タグの属性のうち　@tid, @type, @value, @valueFromSurface, @temporalFunction , @freq, @quant, @mod を概説する。

@tid 属性は 1 文書中の各時間情報表現に付与される識別子である。各時間情報表現を一意に同定するために用い、今後同一指示、参照、事象表現との時間的順序を表す際に用いる。

@type 属性は DATE, TIME, DURATION, SET の 4 つの値を持つ。それぞれ日付表現・時刻表現・時間表現・頻度集合表現を意味する。

@value 及び @valueFromSurface 属性は時間情報表現が含意する日付・時刻・時間の値を表す。値として ISO-8601 形式を自然言語表現向けに拡張したものを用いる。このうち @value は文脈情報を用いて正規化を行った値を付与し、@valueFromSurface 属性は文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する。属性にわりあてる値の詳細を３．３節に示す。定時間情報表現は@value と@valueFromSurface の値は同じになるが、不定時間情報表現は同じになるとは限らない。

@temporalFunction 属性は true, false のいずれかの値を持ち、@valueFromSurface が文脈情報により曖昧性解消可能か否かを表す。定時間情報が得られる不定時間情報表現は true の値を持ち、その他の時間情報表現は false の値を持つ。

@freq, @quant 属性は頻度集合表現に付与される頻度情報及び量化子情報である。属性にわりあてる値の詳細を３．４節に示す。

@mod 属性は時間情報表現のモダリティを表す。例えば 「2000 年以前」をタグづけするために@mod 属性に ON_OR_BEFORE という値をわりあてることにより「以前」というモダリティを表現する。属性にわりあてる値の詳細を３．５節に示す。

作成したコーパスに対し上記属性を付与した。他の属性として、記事配信日時など特別な意味を表す時間情報表現に付与する@functionInDocument、同一指示を表す@anchorTimeID 、時間表現の開始位置と終了位置を表す@beginPoint, @endPoint、タグづけ時の問題点を自由記述する@comment がある。これらの情報は作業者が気づいた範囲で付与を行ったが完全ではない。

３.３　@value 及び @valueFromSurface

各表現に付与する @value 及び @valueFromSurface は ISO-8601 形式を基として、自然言語が表す時間情報向けに拡張したものである。ISO-8601 の標準表記では、日付・時刻表現を XXXX-XX-XXTXX:XX:XX の形で表す。

表 1 日付表現に対する @value

単位	記号	日付表現例	@value
年月日	XXXX-XX-XX	1980 年 7 月 7 日	1980-07-07
曜日	XXXX-WXX-X	水曜日	XXXX-WXX-3
季節	XXXX-{SP,SU,FA,WI}	冬	XXXX-WI
四半期	XXXX-QX	第一四半期	XXXX-Q1
年度	FYXXXX	１９９８年度	FY1998
世紀	XXXX	１１世紀	10XX
紀元前	BCXXXX	紀元前２０２年	BC0202
		４０００年前	KA4
		２億年前	MA200

表 2 曜日表現に対する @value

曜日表現例	@value
月曜日	XXXX-WXX-1
火曜日	XXXX-WXX-2
水曜日	XXXX-WXX-3
木曜日	XXXX-WXX-4
金曜日	XXXX-WXX-5
土曜日	XXXX-WXX-6
日曜日	XXXX-WXX-7
週末	XXXX-WXX-WE

日付表現に対する値の事例を表 1 に示す。自然言語向けの拡張により、ISO-8601 では表現できない季節・四半期・年度などが表現できるようになっている。曜日表現に対する値の事例を表 2 に示す。曜日表現が表す WXX の数値部分は年内の暦週の番号を表す。日本語でよく用いられる「第３水曜」のような月内の暦週の番号を表す方策がとられていない。このため独自拡張として XXXX-XX-W3-3 のように YYYY-MM のあと月内の暦週の番号を WX で表記することを許す。実際のタグづけでは @valueFromSurface には月内の暦週の番号に基づく値を、@value にはカレンダーを参照することにより ISO-8601 の標準表記 XXXX-XX-XX 形式の値をわりあてた。

表 3 時刻表現に対する @value

単位	記号	時刻表現例	@value
時刻	XXXX-XX-XXTXX:XX:XX	２００６年８月８日午前８時４５分３０秒	2006-08-08T08:45:30
時刻（略記）	TXX:XX:XX	午前８時４５分３０秒	T08:45:30
その他	XXXX-XX-XXTXX	未明 *	XXXX-XX-XXTDN
		朝	XXXX-XX-XXTMO
		昼	XXXX-XX-XXTMI
		日中	XXXX-XX-XXTDT
		午後	XXXX-XX-XXTAF
		夕方	XXXX-XX-XXTEV
		夜	XXXX-XX-XXTNI
		深夜 *	XXXX-XX-XXTMN

時刻表現に対する値の事例を表 3 に示す。自然言語向けの拡張により、「朝」「昼」「夜」などが表現できるようになっている。* が付与されている「未明」と「深夜」は日本語新聞記事に頻出したために独自に導入した値である。厳密に TimeML の⟨TIMEX3⟩ 互換にする際にはどちらも「夜」と同じく TNI をわりあてる。

表 4 時間表現に対する @value

単位	記号	時間表現例	@value
年	PnY	３年間	P3Y
月	PnM	２ヶ月	P2M
日	PnD	５日	P1D
時間	PTnH	３時間	PT3H
分	PTnM	３０分	PT30M
秒	PTnS	９秒８０	PT9.80S
週	PnW	１週間	P1W

表 5 不定な表現に対する @value

時間表現例	@value
「今」「現在」	PRESENT_REF
「近年」「以前」	PAST_REF
「今後」「将来」	FUTURE_REF

時間表現に対する値の事例を表 4 に示す。基本的に ISO-8601 の時間表現[5]と同じであり、接頭辞として P を付与し、その後に数値とともにそれぞれ年、月、日、時間、分、秒、週を表す Y, M, D, H, M , S, W を接尾辞として付与する。月(M) と分(M) を区別するために日と時間の境界に T を付与する。

「今」「近年」「今後」など不定な表現に対する値の事例を表 5 に示す。これらは全て自然言語向けに導入した値である。

頻度集合表現は上記 @value 属性を流用しながら次節に示す @freq, @quant 属性を組み合わせることによって表現される。

３.４ 頻度集合表現に対する @freq 及び @quant 属性

頻度集合表現は @value, @freq, @quant 属性を組み合わせることにより複雑な時間情報を表現する。

頻度情報を表すためには、期間を表す@value 属性とともに、@freq 属性に nX をわりあてることにより、焦点をあてている期間中に事象が n 回起こることを示す。例えば「週に２回」を表現する際には

⟨TIMEX3 type="SET" value="P1W" freq="2X"⟩週に２回⟨/TIMEX3⟩

のようにタグづけする[6]。

@quant 属性には「毎日」「毎週」「毎１０月」といった表現に EACH をわりあて、「１０日おき」「３日毎」といった表現に EVERY をわりあてる。この際@value 属性には期間を表す値だけでなく、日付・時刻を表す値が入ることがある。以下に例を示す。

⟨TIMEX3 type="SET" value="P1D" quant="EACH"⟩ 毎日⟨/TIMEX3⟩

⟨TIMEX3 type="SET" value="XXXX-10" quant="EACH"⟩ 毎１０月⟨/TIMEX3⟩

⟨TIMEX3 type="SET" value="P10D" quant="EVERY"⟩ １０日おき⟨/TIMEX3⟩

頻度集合表現は、できるだけ文字列上小さな単位で切り出しているため、現在のところ上記定義で意味論的表示に曖昧性が生じていない。例えば「毎日午前十時ごろから午前六時まで」のような表現の場合、表現全体の単位で切り出すとすると、@value, @freq, @quant 属性のみで曖昧性なく意味論的表示に落とすことは困難である。これは、時間情報表現間の時間的順序関係のタグづけにおいて今後対処していきたい。

---

[5] ISO-8601 では時間を表現するために Time interval 形式 と Duration 形式の２つがあるが、ここでは Duration 形式を用いる。

[6] 説明に不要な属性は省略して表示。以下同様。

３.５　モダリティ修飾子 @mod 属性

時間情報表現は接尾表現をともない様々なモダリティを表現する。@mod 属性は時刻、時間表現に対するモダリティ修飾子である。表 6 に取りうる値の一覧と例を示す。

日付・時刻・時間表現に共通して用いられる@mod 属性として START, MID, END, APPROX がある。例えば、「６０年代初頭」「１０月半ば」「約４０年」は

⟨TIMEX3 type="DATE" value="196X-XX-XX" mod="START"⟩ ６０年代初頭⟨/TIMEX3⟩

⟨TIMEX3 type="DATE" value="XXXX-10-XX" mod="MID"⟩ １０月半ば ⟨/TIMEX3⟩

⟨TIMEX3 type="DURATION" value="P40Y" mod="APPROX"⟩ 約４０年⟨/TIMEX3⟩

のようにタグづけする。

日付・時刻表現に対する@mod 属性として、BEFORE, AFTER, ON_OR_BEFORE, ON_OR_AFTER がある。例えば「１９９８年以前」は

⟨TIMEX3 type="DATE" value="1998" mod="BEFORE"⟩１９９８年以前⟨/TIMEX3⟩

のようにタグづけする。

時間表現に対する @mod 属性として、EQUAL_OR_LESS , EQUAL_OR_MORE, LESS_THAN, MORE_THAN がある。例えば「１０分以内」は

⟨TIMEX3 type="DURATION" value="PT10M" mod="EQUAL_OR_LESS"⟩１０分以内⟨/TIMEX3⟩

のようにタグづけする。

表 6 @mod 属性に対する値

値	定義	例
@mod=START	日付時刻表現の初期	「初め」「初頭」
@mod=MID	日付時刻表現の中期	「半ば」「中ごろ」
@mod=END	日付時刻表現の後期	「末」 「暮れ」
@mod=APPROX	近似表現	「ごろ」
@mod=BEFORE	日付時刻表現より前	「前」
@mod=AFTER	日付時刻表現より後	「過ぎ」
@mod=ON_OR_BEFORE	日付時刻表現以前	「以前」
@mod=ON_OR_AFTER	日付時刻表現以後	「以降」「以来」
@mod=EQUAL_OR_LESS	時間表現の範囲以下	「以内」
@mod=EQUAL_OR_MORE	時間表現の範囲以上	「以上」
@mod=LESS_THAN	時間表現の範囲未満	「未満」「近く」
@mod=MORE_THAN	時間表現の範囲超過	「余り」「過ぎ」

４．タグの分析

BCCWJ のコアデータ[7]の一部である、全ジャンル(OW(白書), PB(書籍), PN(新聞), OC(Yahoo! 知恵袋), PM(雑誌), OY(Yahoo! ブログ)) の部分集合“A”と比較的時間表現が多いジャンルである PN(新聞) の部分集合“B”について人手によりタグづけした。

表 7 にデータの概要を示す。表中「ファイル数」はタグづけしたファイルの数、ファイル数の下のカッコ内の数字「時間表現あり」は時間表現一つ以上含むファイルの数を表す。

まず、OC, OY などのユーザー生成コンテンツはサンプリングの長さにもよるが時間表現

[7]コアデータは、OW: 白書、PB: 書籍、PN: 新聞、OC: Yahoo! 知恵袋、PM: 雑誌、OY: Yahoo! ブログからなり、それぞれ約 5 万語単位で、タグづけすべき優先順位をもつ部分集合（A > B > C > D > E）が設定されている。

表 7 @type 属性毎の出現数と文脈による曖昧性解消可能性

ジャンル	ファイル数 （うち時間表現あり）	DATE	TIME	DURATION	SET	合計
		（文脈に曖昧性解消可能なものの数）				
OW	17 16	596 (414)	0 (0)	191 (0)	6 (0)	703
PB	25 25	209 (51)	28 (12)	105 (0)	14 (0)	356
PN	110 110	1323 (999)	193 (162)	553 (0)	41 (0)	2110
OC	518 250	341 (95)	70 (19)	184 (0)	37 (0)	632
PM	23 23	333 (108)	37 (2)	131 (0)	28 (1)	529
OY	257 198	632 (215)	161 (58)	117 (1)	22 (0)	932

が一つも含まれないものがある一方、OW, PB, PN, PM などのユーザー生成コンテンツ外のほとんどは時間表現が必ず一つ以上含まれている。OW の中で唯一、時間表現が一つも含まないサンプルは平成 16 年度の森林・林業白書であった。

ジャンル毎の文書作成日時を示すタグを除いた@type 毎のタグの出現数を“DATE”，“TIME”，“DURATION”，“SET” 表 7 右に示す。合計はジャンル毎の時間表現の合計を表す。カッコ内は、文脈より曖昧性解消が行うことができたものの数を表す。日付表現の曖昧性解消は、和暦から西暦への換算や、西暦 2 ケタ表記から西暦 4 ケタ表記への換算、さらに年が省略されている表現の文脈や文書作成日時に基づく年の補完によるものがある。ジャンル間の差異において OW(白書) が時刻表現を一つも含まないという特色がわかった。

OW は和暦西暦換算の事例が多い一方、PN は文書作成日時がメタデータ中に明示的に含まれているため、曖昧性解消が文脈によって行える事例が多かった。他のジャンルは曖昧性がない表現が多いわけではなく、曖昧性解消するに足る情報がデータ中に含まれていないことにより、具体的な時間軸上の区間を指し示すことができない事例が多かった。時刻表現の曖昧性解消は、日付が省略されている場合の日付の補完のほか、午前と午後の曖昧性解消が含まれる。時間表現で文脈により曖昧性解消される事例は、「３年くらい前」という相対的に何年に起きたかを日付表現に置き換えられるもの一例であった。集合表現で文脈により曖昧性解消される事例は、「１時間おき」という具体的に「１１月１４日の１時から６時の１時間おき」といったように時刻が特定できるもの一例であった。

ここで日付表現の曖昧性解消とは不定表現を完全に定表現に変換することだけではなく、部分的に情報を補完すること(例えば「３日」という表現に対し９月であることまでがわかるが何年であるかまではわからないので @value="XXXX-09-03"をわりあてること) も含まれる。さらに時刻表現の曖昧性解消の際、日付の情報が含まれていない場合には時刻レベルの補完のみにとどめられている場合[8]がある。実際に補完すべき時刻情報がより多くあり、時間情報表現の正規化が重要であることがわかる。

表 8 頻度集合表現の統計

	@freq=nil	@freq= n X	otherwise
@quant=nil	3	43	2
@quant=EACH	75	2	5
@quant=EVERY	18	0	0

頻度集合表現はタグづけしたテキスト中に 148 件出現した。@quant 属性、@freq 属性別

8 時刻レベルの補完のみにとどめられた時刻表現の正規化は、今後、時間的順序関係を表す〈TLINK〉タグを、文書作成日時を含めた日付表現への参照表現としてタグづけすることにより解決する。

の統計を表 8 に示す。(@quant="EACH", @freq=nil) に分類される表現（例「毎日」）が最も多く、次に(@quant=nil, @freq="nX") に分類される表現（例「１日３回」「週２度」）が多かった。その他複雑な表現として「１カ月あたり１時間」(@value="PT1H", @freq="P1M") といった表現があった。

表 9 @mod 属性の統計

@type	DATE	TIME	DURATION	SET
@mod=START	27	11	1	0
@mod=MID	5	0	2	0
@mod=END	72	0	5	1
@mod=APPROX	19	35	95	2
@mod=BEFORE	0	5	−	0
@mod=AFTER	0	6	−	0
@mod=ON OR BEFORE	7	0	−	0
@mod=ON OR AFTER	36	21	−	0
@mod=EQUAL OR LESS	−	−	16	0
@mod=EQUAL OR MORE	−	−	29	0
@mod=LESS THAN	−	−	13	0
@mod=MORE THAN	−	−	5	0

各時間情報表現に付与された @mod の統計を表 9 に示す。日付表現で多かった @mod の値は END であり、「年末」「月末」といった表現が例示される。それ以外の表現では「約」「ごろ」が付与された APPROX が多かった。日付表現では「五日以前」「４月以降」といった ON_OR_BEFORE, ON_OR_AFTER がある一方、BEFORE, AFTER は存在しなかった。時刻表現においては逆に「八時前」「５時すぎ」といった BEFORE, AFTER が出現する一方、ON_OR_BEFORE, ON_OR_AFTER はほとんど出現しなかった[9]。

５．関連研究

表 10 に関連研究を示す。

表 10 関連研究

英語に関する関連研究		
MUC-6 (R. Grishman and B. Sundheim 1996)	評価型会議	時間情報表現の切り出しのみ
(A. Setzer 2001)	タグづけ基準	時間情報表現の切り出しと正規化
TERN (DARPA TIDES 2004)	評価型会議	時間情報表現の切り出しと正規化
TimeML (J. Pustejovsky et al. 2003b)	タグづけ基準	事象間の時間的順序関係
TimeBank (J. Pustejovsky et al. 2003a)	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
(B. Boguraev and R. Kubota 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
(I. Mani 2006)	解析手法	二事象間の時間的順序関係解析
TempEval (M. 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 (M. 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
日本語に関する関連研究		
IREX (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
関根らの拡張固有表現体系 (S. Sekine et al. 2002)	タグつけ基準	時間情報表現の切り出しのみ
本論文	コーパス	時間情報表現の切り出しと正規化

英語においては、評価型国際会議 MUC-6 (R. Grishman and B. Sundheim 1996) の１タスク

[9] 数少ない ON_OR_AFTER の時刻表現は 「夜来」と「昼以降」の２事例。

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。タグづけ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 のタグづけ基準 ⟨TIMEX⟩ に対し、Setzer は時間情報表現の正規化に関するタグづけ基準を提案している(A. 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。ISO-8601 形式を拡張した @value 属性などが設計され、時間表現の正規化が自動解析対象となっている。その後、Pustejovsky らによりタグづけ基準 TimeML (J. Pustejovsky et al. 2003b) 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が付加される。これらの情報は人手でタグづけすることを目的に設計され、TimeBank (J. Pustejovsky et al. 2003a) 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(B.Boguraev and R. Kubota Ando 2005; I. 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、扱いにくいものであった(B.Boguraev and R. Kubota Ando 2006) 。2007 年に開かれた SemEval 2007 の 1 タスク TempEval(M. 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 (M. 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、スペイン語に関しても同様のデータを利用したタスクが設定された。

日本語においては、IREX (IREX 実行委員会 1999) における 1 タスクとして、固有表現抽出タスクが設定された。IREX における時間情報では、日付・時刻表現を対象にし、相対的な表現が定義に含めている。また、関根らは拡張固有表現体系(S. 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、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 2010)。日本語においては、表現の分類の体系化が進んでいるが、正規化のための研究は他言語と比べて遅れをとっている。

６　おわりに

本稿では作成している日本語時間情報タグつきテキストコーパスについて説明した。タグつきデータはタグの情報のみ github 上に公開する。BCCWJ を入手することでタグつきテキストコーパスが復元できる。

以下、今後の展望を示す。

今回作成したテキストコーパスをベンチマークとして正規化を行う日本語時間表現解析器の開発を現在進めている。作成中の解析器では、まず、表層文字列からわかる値をラティス上に展開し、セミマルコフモデルを用いて曖昧性解消を行う。解析対象表現－文書作成日時および解析対象表現－隣接時間情報表現の時間的順序関係を今回作成したタグづきコーパスを用いて機械学習器を用いて推定することにより、不定時間情報表現に対する情報補完を行う。

今後、TimeML で行われている事象表現と時間表現間の時間的順序関係(TimeML における⟨TLINK⟩) 付与を進めていきたい。そのためには、対象となる事象表現の策定、事象表現に対する分類(TimeML における EVENT@type) 、テンス・アスペクト体系の整備(同 MAKEINSTANCE@tense, MAKEINSTANCE@aspect) 、節間の関係定義(同 SLINK) など解決すべき問題は山積している。現在は事象表現を動詞に限定し、事象表現に対する分類として工藤らの動詞分類(工藤 1995, 2004)を基にした階層的ラベルを設計し付与している。階層的ラベルの上位の情報を得ることにより TimeML で定義されている

EVENT@type の 8 分類に対応する設計になっている。テンス・アスペクト体系については中村らのテンス・アスペクトの解釈(中村 2001) を参考にしてラベルを設計する予定である。今後 TimeML に準じた事象表現に対するタグづけを行い、最終目標である事象表現に対する時間情報付与の研究を進めていきたい。

謝　辞

本研究は、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) 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

B. Boguraev and R. Kubota Ando (2005). “TimeML-Compliant Text Analysis for Temporal Reasoning.” In Proc. of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05) , pp. 997–1003.

B. Boguraev and R. Kubota Ando (2006). “Analysis of TimeBank as a Resource for TimeML parsing.” In Proc. 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.

R. Grishman and B. Sundheim (1996). “Message Understanding Conference-6: a brief history.” In Proc. of the 16th International Conference on Computational Linguistics (COLING-96), pp. 466–471.

IREX 実行委員会(1999). IREX ワークショップ予稿集.

I. Mani (2006). “Machine Learning of Temporal Relations.” In Proc. of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-2006), pp. 753–760.

J. Pustejovsky et al. (2003a). “The TIMEBANK Corpus.” In Proceedings of Corpus Linguistics 2003, pp. 647–656.

J. Pustejovsky et al. (2003b). “TimeML: Robust Specification of Event and Temp oral Expressions in Text.” In Proc. of the 5th International Workshop on Computational Semantics (IWCS-5) .

S. Sekine et al. (2002). “Extended Named Entity Hierarchy.” In the proc. of the Third International Conference on Language Resources Evaluation (LREC-02) .

A. Setzer (2001). Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study . Ph.D. thesis, University of Sheffield.

M. Verhagen et al. (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” In Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 75–80.

M. Verhagen et al. (2010). “SemEval-2010 Task 13: TempEval-2.” In Proc. 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.

漢字・語彙指導の根拠としてのコーパスの役割

河内　昭浩（安田女子大学文学部）†

The Role of Corpus as Evidence for *Kanji* and Vocabulary Instruction

Akihiro Kawauchi (Faculty of letters, Yasuda Women's University)

１．はじめに

発表者は、特定領域研究「日本語コーパス」言語政策班（代表；田中牧郎氏）の一員として、コーパスの国語教育への応用を模索してきた。本発表では、国語科学習指導、特に漢字・語彙指導においてコーパスの果たす役割について述べる。常用漢字表の改定に伴い、国語教育で指導すべき漢字について、改めて検証することが求められている。ここでは、言語政策班で作成した「教科書コーパス」を用いて、指導すべき常用漢字の選定のための資料の提示を行う。

また文学教材において指導すべき語彙とその指導方法について、「現代日本語書き言葉均衡コーパス」をもとに設定された、「語彙レベル」を用いた提案を行う。

コーパスの国語教育への応用として、これまでコーパスを活用した新たな教材の開発や、ICT 教材の一つとしてコーパスを授業で活用することなどを試みてきた。それらの開発・活用は今後も継続する課題である。今回の提案は、従来の国語科の指導事項や指導方法を、コーパスを用いて評価し、その指導の根拠を与えるというものである。今回は、漢字・語彙指導に焦点を当てているが、様々な国語科の学習の「指導の根拠」として、コーパスはこれから大きな役割を果たしていくものと考えている。

２．指導すべき漢字の選定

平成 22 年 11 月、改定常用漢字表が内閣告示された。これまで 1945 字であった常用漢字は、196 字追加、5 字削除により、計 2136 字になった。改定に伴い、中学校国語科で学年別に取り扱う常用漢字の数が増加されたが、具体的に、どの漢字をいつ指導するかは定められていない。

常用漢字表の性格は、「一般の社会生活における漢字使用の目安」とされている。またその「一般の社会生活における漢字使用」とは、「義務教育における学習を終えた後、ある程度実社会や学校での生活を経た人を対象として考え」ているとされている。一方国語教育においても、子どもの生活に資する漢字指導を施すことが求められている。その上で、中学校終了時までに「常用漢字の大体を読むこと」を指導することになっている。

では漢字使用の場面において、具体的にどこまでが、国語教育の想定する子どもの現在の「生活」で、どこからが、常用漢字表の想定する将来の「一般の社会生活」なのか。社会生活の場面を区切り、その境を実証的に明らかにすることは現状では困難である。そこで次善の策として、「教科書」における漢字の読み書きを、まずは子どもの生活における漢字使用の場面と位置づけたい。教科書は子どもの日常生活（学校生活）の基幹であり、そ

† kawauchi@yasuda-u.ac.jp

の内容は、子どもを社会生活へと導く手だてである。各教科の教科書で、出現頻度の高い漢字、重要度の高い漢字を、まずは子どもの生活に必要な漢字ととらえたい。そしてその上で、今後、「現代日本語書き言葉均衡コーパス」等を用いて、「一般の社会生活」における漢字使用の状況を明らかにしていきたい。

今回はまず、全常用漢字 2136 字の、小中学校教科書における出現状況を概観する。また、教科別の常用漢字の出現頻度を示し、指導方法について提言を行う。さらに、教科書における出現順位と、常用漢字表の改定の際に使用された「漢字出現頻度表順位対照表」の順位とを比較、検討し、中学校における段階的な常用漢字の指導の、一つの指針となるデータを提示する。

尚、こうした漢字に関する考察は、言語政策班において、相澤正夫氏や棚橋尚子氏らによって進められてきた。本発表はそれらの研究より教示を仰いでいる。

以下、発表時に掲示する資料の一部とその概要を述べる。

［資料１］　小中学校教科書　常用漢字出現順位総合表

小中学校教科書における出現順位を軸に、全常用漢字 2136 字の、各教科における順位を示したものである。漢字の種類を、［教育］（学年別漢字配当表に示された漢字、1006 字）、［常用］（改定前から常用漢字表にあり、［教育］以外の漢字、934 字）、［新規］（新しく常用漢字表に加えられた漢字、196 字）に区分している。発表時に資料を提示する。

［資料２］　小中学校教科書順位　種類別出現状況表

［資料２］は、常用漢字の種類（［教育］、［常用］、［新規］）の教科書における出現傾向をつかむために作成したものである。250 字ごとの、［教育］、［常用］、［新規］漢字の割合を示している。1 位から 1000 位まででは、［教育］漢字の割合が高く、1000 位以降は、［常用］漢字の割合が高くなっている。

小中学校教科書順位	［教育］		［常用］		［新規］	
1　～　250	250	100.0%	0	0.0%	0	0.0%
251　～　500	245	98.0%	4	1.6%	1	0.4%
501　～　750	219	85.9%	34	13.3%	2	0.8%
751　～1000	159	64.9%	83	33.9%	3	1.2%
1001～1250	99	39.4%	144	57.4%	8	3.2%
1251～1500	29	11.5%	215	85.0%	9	3.6%
1501～1750	3	1.2%	228	91.2%	19	7.6%
1751～2000	2	0.7%	221	81.0%	50	18.3%
2001～2136	0	0.0%	5	4.6%	104	95.4%
	1006	47.1%	934	43.7%	196	9.2%

［資料３］　①　小中学校教科書順位上位　［常用］漢字表

② 書籍順位上位　［常用］漢字表
③ 小中学校教科書順位上位　［新規］漢字表
④ 書籍順位上位　［新規］漢字表
⑤ 小中学校教科書順位下位　［教育］漢字表
⑥ 書籍順位下位　［教育］漢字表

［資料３］は、資料２で把握した傾向をもとに、どの［常用］、［新規］漢字が出現上位となっているのか、またどの［教育］漢字が出現下位となっているのかを示したものである。また対照として、常用漢字表の改定の際に使用された、凸版印刷作成による「漢字出現頻度表順位対照表」のデータを借り、［書籍順位］として付記している。①は、教科書の上位に入る［常用］漢字を示している。色のついている漢字は、対照の書籍順位では上位には入らないことを示している。②は、書籍で上位に入る［常用］漢字を示している。①、②を見ると、教科書と書籍では、数の違いだけではなく、上位となる［常用］漢字そのものに違いがあることが分かる。③、④は、教科書、書籍で上位に入る［新規］漢字を示している。［新規］上位漢字では、書籍では上位に来ても教科書では上位に来ない漢字が多い。また⑤、⑥は教科書、書籍で下位となる漢字を示している。下位漢字では、書籍では下位に来ても教科書では比較的上位に来る漢字が多い。特に「秒」（661 位）、「磁」（553 位）、「銅」（563 位）、「径」（553 位）など、教科書ではかなり上位に入る［教育］漢字が、書籍では下位になる。（⑤、⑥については、発表時に資料を提示する。）

①小中学校教科書順位上位［常用］

小中順位	常用	*書籍順位*
264	環	*713*
390	江	*290*
430	響	*694*
452	影	*500*
506	鮮	*592*
507	震	*822*
511	項	*1154*
513	僕	*454*
517	違	*255*
535	徴	*965*
544	離	*396*
555	施	*730*
559	胞	*1239*
565	介	*376*
576	換	*924*
576	溶	*1545*
605	歳	*268*
622	彼	*53*

②書籍順位上位　［常用］

書籍順位	常用	小中順位
53	彼	622
91	郎	697
181	込	876
255	違	517
256	吉	636
268	歳	605
272	井	743
290	江	390
349	御	793
370	突	1001
376	介	565
396	離	544
423	撃	1069
429	振	826
452	奥	1147
454	僕	513
455	佐	1130
456	頼	870

③　小中学校教科書順位上位［新規］

小中順位	新規	*書籍順位*
465	岡	*532*
688	阪	*855*
743	奈	*786*
779	藤	*264*
796	韓	*860*
955	鹿	*1022*

④　書籍順位上位　［新規］

書籍順位	新規	小中順位
264	藤	779
414	誰	1651
474	俺	2064
532	岡	465
561	頃	1815
786	奈	743
855	阪	688
860	韓	796
909	弥	1006
985	那	1298

［資料４］各教科出現順位上位漢字表

［資料４］は、［国語］以外の各教科の、出現順位上位の漢字を列挙したものである。それぞれ［国語］の順位を対照させてある。一瞥して、各教科の学習内容に沿う漢字が多いことが分かる。

［資料4］ 各教科出現順位上位漢字表(一部)

数学順位	数学	国語順位	理科順位	理科	国語順位	社会順位	社会	国語順位
1	数	153	1	物	33	1	国	84
2	形	57	2	水	91	2	人	1
3	方	15	3	気	36	3	地	61
4	算	1031	4	電	334	4	日	13
5	角	539	5	化	173	5	本	17
6	分	6	6	動	54	6	大	14
7	何	55	7	調	68	7	年	34
8	式	369	8	体	60	8	生	11
9	表	20	9	変	103	9	中	12
10	長	95	10	図	159	10	業	269
11	次	52	11	酸	1144	11	調	68
12	考	22	12	地	61	12	行	31
13	求	523	13	分	6	13	市	292
14	計	362	14	大	14	14	会	44
15	図	159	15	素	604	15	自	30
16	線	285	16	流	176	16	見	7
17	面	127	17	生	11	17	学	35
18	問	115	18	質	250	18	県	276
19	点	97	19	中	12	19	世	102
20	辺	476	20	見	7	20	民	355
21	直	209	21	合	25	21	政	808

［資料５］　各教科特徴漢字　出現頻度・文例集

［資料５］は、［資料４］をもとに、国語以外の各教科での出現頻度は高いが、国語ではあまり見られない漢字とその文例をまとめたものである。こうした他教科の学習内容に深くかかわる漢字を、他教科の学習事項と連関させる形で、国語科の中で指導していくべきだと考えている。例えば、数学（算数）上位の漢字、「算」は、今年度採択の M 社の国語の教科書を見ると、小学 2 年生の「ことばの指導」のページに、「算数」という単語が挙げられているのみで、その「算数」という単語 1 語をもって、「算」という新出漢字を学ぶことになっている。それよりは、［資料５］で示したような文脈（「たし算のしかたを考えよう」など）の中で、学ぶことができたほうが、子どもたちにとってより有意義である。

また、［新規］漢字には、主に都道府県名で用いられる漢字（阪、奈など）が新たに加えられている。学年別漢字配当表に配当された漢字ではないこれらの漢字は、国語科では中学校で、新出漢字として扱うことになる。しかし一方で、新しい小学校学習指導要領社会では、［第 3 学年及び第 4 学年］の段階で、都道府県名の指導を行うことが明記されている。子どもたちが社会科で、興味関心をもって都道府県名を学ぶその時に、国語科で、都道府県名の漢字を指導することができれば、小学生の学習全体に一貫性を与えることができる。

このように、他教科の学習内容を国語科の教科書教材に意欲的に取り入れていくことによって、子どもたちの知の連関が、国語を基軸に生まれていくはずである。

［資料５］　各教科特徴漢字　出現頻度・文例集（一部）

数学	種類	配当	小中計	小計	中計	文例
算	教育	2	1619	1352	267	たし**算**のしかたを考えよう。 ひっ**算**でしましょう。（小2）
等	教育	3	566	174	392	この＞，＜のしるしを不**等**号といいます。（小3）
倍	教育	3	424	343	81	何**倍**かをもとめるときは，わり算を使います。（小3）
値	教育	4	348	52	296	この本の**値**段は，雑誌の値段の5／3倍です。（小6）

理科	種類	配当	小中計	小計	中計	文例
酸	教育	5	876	297	579	ホウ**酸**のとけかたを調べよう。（小5）
液	教育	5	569	225	344	ヨウ素**液**にひたす。（小5）
管	教育	4	312	128	184	色水を入れた試験**管**。（小4）
炭	教育	3	310	96	214	石油や石**炭**などは，便利なエネルギー（小4）
塩	教育	4	277	165	112	氷に，水と食**塩**をまぜたものをかける。（小4）

社会	種類	配当	小中計	小計	中計	文例
政	教育	5	880	134	746	大使館や**政**府観光局から資料を送ってもらう。（小5）
域	教育	6	793	129	664	工業地域を京葉工業地**域**というんだね。（小5）
権	教育	6	660	96	564	人**権**や報道被害の問題も頭に入れて，（小5）
府	教育	4	601	126	475	〔 〕は，都道**府**県をしめしています。（小3・4）

今回は、「教科書コーパス」と、常用漢字表の改定の際に使用された、「漢字出現頻度表順位対照表」のデータをもとに資料を作成した。今後は「現代日本語書き言葉均衡コーパス」を用いて、より、社会生活に必要な漢字の選定を行っていきたい。また、「教科書コーパス」のデータも更新していく予定である。さらに、他教科の学習内容を取り入れた、国語科における教材化の具体案を今後提示していくつもりである。

３．文学教材の語彙指導

教科書の文学教材には、脚注欄に、文やイラストなどで説明の施される語句（以下、「語注語句」と呼ぶ）と、語彙の学習を行うための語句（以下、「注意語句」と呼ぶ）が配置される。語注語句は、生徒の本文読解を補助する役割を持つ。逆に言えば、大人が、そのままでは子どもたちが理解できないであろう、読解のために解説が必要であろうと考える語句に注が付けられていることになる。また注意語句は、その語句を用いて類義語や対義語などの語彙の学習を行うために列記される。つまり教材中の注意語句は、社会生活を行う上で習得が必要な語句であると判断されていることになる。

それらの妥当性を判断するために、「語彙レベル」を使用する。特に、流通実態（図書館）サブコーパスの、書籍における語彙レベル（以下、「レベル_LB」と記す）と語注語句、注意語句との相関について述べる。流通実態（図書館）サブコーパスは、公共図書館のデータを基に、書き言葉の流通の実態を反映させた、「現代日本語書き言葉均衡コーパス」 を構成する一つのサブコーパスである。そこでのレベルが高いということは、社会で流通している、つまり社会生活上重要度が高い語句ととらえることができる。逆にレベルの低い語句は、社会生活上重要度が低く、子どもたちが、知らない、理解できない語句である可能性が高いと言える。従って役割から考えれば、語注語句は、レベル_LB の下位(d・e)であり、注意語句はレベル_LB の上位(a・b)であることが妥当であると考えられる。

このような観点から、教材「羅生門」及び教材「走れメロス」の語彙を分析した結果を資料として提示する。

尚、こうした文学教材の語彙の検証にはすでに、田中(2011)による中学校教科書定番教材の「少年の日の思い出」の語彙研究や、鈴木(2011)による、本発表でも扱う、中学校の「走れメロス」の語彙研究などがある。

［資料１］「羅生門」の語注語句

教材「羅生門」の語注語句を、レベル別に掲げたのが［資料１］である。語句の多くはレベル_ LB の d・e に相当するものが多い。ただ中には、「山吹」や「きりぎりす」といった一見平易と思われる語も下位にある。大人には馴染み深く感じられる動植物の語彙が、現代語の流通実態としてすでにあまり見られなくなっていることが分かる。

レベル_LB	語注語句
a	（該当なし）
b	楼　太刀　鞘
c	烏帽子　朱雀　局所　申　くさめ　やもり　しらみ　蟇
d	下人　検非違使　きりぎりす　弩　羅生(門)　太刀帯
e	旧記　洛中　狐狸　低回　築地　丹塗り　市女(笠)　山吹　甍　火桶

［資料２］「羅生門」の語彙と語彙レベルの相関

教材「羅生門」の主な語彙をレベル別に示したものが［資料２］である。下線部は、［資料１］で示した語注語句に当たる。レベル下位の語彙の中には、生徒が「知らない」などとは思いもよらない語彙が見られる。例えば「羅生門」の冒頭文の語彙レベルは以下の通りとなる。

ある日の暮れ方(e)のことである。一人の下人(d)が，羅生(e)門の下で雨やみ(e)を待っていた。

「下人」、「羅生門」は語注語句である。一方「暮れ方」、「雨やみ」に語注を付ける教科書はない。しかしレベル_LB において「下人」は d、「羅生（門）」「暮れ方」「雨やみ」は e といずれも現代語においてほとんど流通していない語彙であることが分かる。「下人」、「羅生門」と同様に、「暮れ方」、「雨やみ」も、生徒にとって「知らない」、「見たことがない」語彙である可能性が高いことになる。そのように考えると、この冒頭文は、読み馴染んだ大人（教師）が思う以上に、生徒にとっては難文である可能性がある。

また、こうした語彙レベルをもとに語注語句、注意語句の選定を行っていくことで、客観性と統一性のある、文学教材における語彙指導を行うことができるようになる。

［資料２］「羅生門」の語彙と語彙レベル

レベル_LB	羅生門の語彙
a	（漢語）人　門　匹　以上　市　地震　料　始末　習慣　気味　近所　たくさん　勿論　段　主人　当時　適当　模様　影響　けしき　次第　手段　死　道　結局　当然　積極　勇気　風　遠慮　範囲　数　人間　事実　人形　部分　一層　瞬間　感情　多分　恐怖　呼吸　記者　丁度　同時　問題　合理　自分　無言　無理　全然　意志　支配　意識　得意　満足　役人　現在　大体　意味
b	（漢語）　火事　仏像　金銀　修理　胡麻　糞　紺　格別　途方　肯定　楼　草履　死骸　好奇　老婆　悪　反感　未練　執拗　明白　生死　憎悪　成就　時分　平凡　先方　寸　　色　白髪（和語）　太刀　鞘
c	（漢語）円柱　烏帽子　飢饉　死人　点々　永年　朱雀　局所　臭気　嗅覚　木片　善悪　勝敗　眼球　円満　肉食　失望　侮蔑　気色 （和語）申　蟋蟀　大路　さびれる　打ち砕く　盗人　点々　引き取る　棄てる　棲む　飛びまわる　上の空　夕焼け　にきび　くさめ（くしゃみ）　やもり　暇　吹き抜ける　虱　墓　大股　目下　鋼　安らか　見下す　鬘
d	（漢語）　格別　衰微　余波　大儀　中段　下人　検非違使　存外 （和語）きりぎりす　雨風　床板　弩　手荒い （固有）羅生（門）　太刀帯
e	（漢語）旧記　仏具　箔　洛中　狐狸　刻限　低徊　逢着　腐爛　暫時　語弊　冷然 （和語）雨やみ　築地　暮れ方　丹塗り　市女（笠）　山吹　甍　火桶　干魚　白髪頭　喉仏

［資料３］「走れメロス」の語注語句（〇）と注意語句（●）（レベル e）

［資料３］［資料４］は、各社の中学校教科書における、教材「走れメロス」の、レベル別の語注語句と注意語句の一覧である(一部)。「走れメロス」は全社掲載の定番教材である。

しかし語彙指導の観点は必ずしも一致していない。また、同じレベルでも、語注語句とされるものと注意語句とされるものとがあり、今後その妥当性を検証していく必要がある。

	M社	K社	T社	S社	G社	レベル_LB
竹馬の友	●	●	●	●	●	e
乱心		●				e
巡邏	○	○	○	○	○	e
悪びれる		●				e
私欲	●					e
無二	●	●	●		●	e
やつばら			○			e
車軸(を流す)	○	○	○	○	○	e
喜色		●				e
信実	○	●	●	●	●	e
照覧	○	●	○	●		e
胴震い		●				e
五臓	○	●	○		○	e
残光		●				e
刑吏		○		○	○	e

［資料４］「走れメロス」の語注語句（○）と注意語句（●）（レベル a）

	M社	K社	T社	S社	G社	レベル_LB
正当	●		●			a
まさしく	●					a
精神			●			a
まさか	●					a
義務		●		●		a
徐々(に)			●			a

４．まとめ

コーパスから得られる情報を照射することで、国語科の指導に新たな展望が開けてくる。この確信を形に変えるべく、今後も提言を続けていきたいと考えている。

文献

文化審議会答申（2010）「改定常用漢字表　基本的な考え方」

文部科学省（2008）「中学校学習指導要領　第２章第１節　国語　第３学年２内容　ｳ(ｲ)」pp.20

田中牧郎（2011）「語彙レベルに基づく重要語彙リストの作成―国語施策・国語教育での活用のために―」『特定領域研究「日本語コーパス」言語政策班報告書　言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用』pp.77-88

田中牧郎（2011）「『少年の日の思い出』の語彙指導」前同、pp.195-204

鈴木一史（2011）「学習意識と語彙～実生活で生きてはたらく語彙～」『教育科学国語教育　No.736』明治図書、pp.84-87

「語りかけ性」を有すると判断される書きことばの表現

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）

Addressing Expressions in Written Japanese

Sachi Yasuda (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)

１．はじめに

我々のプロジェクトでは，『現代日本語書き言葉均衡コーパス』（BCCWJ）に収録されている図書館サブコーパス（10,551 サンプル）の書籍サンプルに，人手で文書分類の観点から情報を付与する作業を行っている（柏野・奥村 2012）．

本稿は，「語りかけ性」に着目する．書きことばではあるが，語りかけている印象を与える表現がある．これまで，どのような観点に基づいて「語りかけ性」の有無が判断されるか，複数作業者の判断が一致した 500 例ほどのサンプルを分析し，高頻度に現れた表現の抽出を試みてきた（保田ほか 2012）．しかし，低頻度であっても，判断の決め手となる表現がある．そこで本稿は，約 4,000 例の観点付与結果より，作業者が「語りかけ性」があると判断したサンプルについて分析を行った．作業者が「語りかけ性」を感じる程度の差を見ることで，「語りかけ性」の判断に関わりの強い表現を明らかにする．「語りかけ性」があると判じられたサンプルのジャンルや販売対象との関連性ついても報告する．また，作業者が判断時に記載したコメントの分類を行い，作業者が「語りかけ性」があると判じた根拠となる表現をまとめる．コメントのあったサンプルを確かめることで，表現の量や出現位置に見られる傾向も得られた．

２．「語りかけ性」について

書籍テキストの中には，著者が読者に対して直接語りかけていると解釈できる文体がある．「あなた」や「みなさん」などの呼びかけ表現や，「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など，「直接的な語り」表現と呼べるような表現が含まれるテキストを，本稿では「語りかけ性を有するテキスト」と呼ぶ（柏野 2010）．以下のようなサンプルが，「語りかけ性」があると考えられる例[1]である．

> お金を稼ぐために事業を始めるべきでないとしたら、なぜ事業を始めるのでしょうか。答えはあなたの情熱と夢にあります。お気に入りの趣味として事業を始めることを考えることができますか。それはほんの少数の人たちにしか理解できない夢です。なぜかって。まず第１に、たいていの人たちがそんなことが可能とさえ思っていないからです。そんなことは、才能があり、お金持ちで、有名でなければできないと彼らは考えているのです。（LBp3_00158『世界一わかりやすいほんとうのお金持ちになる法』）

保田ほか（2012）では，書籍サンプルからランダムに選び出した約 500 のサンプルを 1

[†] yasuda_s@ninjal.ac.jp

[1] サンプルの出典は，BCCWJ のサンプル ID と書名とで記す．

セットとし，「語りかけ性あり」として 3 人の作業者の判断が一致したサンプルの分析を行った．品詞，活用形，語彙素において，観点付与を行う作業者が共通して分類に用いている可能性の高い指標が整理されている（表１）．

また，小磯ほか（2011）は，調査者から得た評定語を指標として分析を行っている．そのとき，「書きことば的―話しことば的」という尺度に，「読み手に語りかける―語りかけの少ない」という尺度や，改まりの程度などの複数の観点が関与する可能性を示している．前掲の保田ほか（2012）でも，作業者が「話しことば的」の観点とは異なる指標で判断を行っている可能性が示唆されている．感動詞，融合（「～じゃない」「～なきゃ」など），終助詞「よ」のように，「話しことば的」と作業者が判断したサンプルで出現率が高くとも，「語りかけ性あり」と判断されたサンプルでは出現率が低いという要素があるためである．作業者による「語りかけ性」の有無の判断は，「話しことば的」との差異を含め，複雑な条件によって行われているものと考えられる．そこで，本稿は「語りかけ性」を有しているとの判断を行うにあたり，作業者が根拠として用いた表現に注目する．

表１．「語りかけ性」の有無に関わる表現（保田ほか，2012）

「語りかけ性あり」
　活用形　　：意志推量形（「～でしょう」「～だろう」など）が多い
　語彙素　　：「です（助動詞）」「ます（助動詞）」「事（名詞）」が多い
　　「あなた（代名詞）」「自分（名詞）」「ね（終助詞）」が多い
「語りかけ性なし」
　品詞・語種　：固有語が多い
　語彙素　　：「た（助動詞）」が多い

３．調査データ

本稿で扱うデータのセットは表２の通りである．

BCCWJ の図書館サブコーパスに含まれる書籍をランダムに並べ替え，6 人の作業者がアノテーションを行っている．現在までにアノテーション作業が完了しているサンプル 9,000 のうち，保田ほか（2012）で行った分析以降にアノテーションが為されたサンプル 6,000 から，テキスト部分が極端に少ないなど[2]の理由によりアノテーション対象外とされたサンプルを除く 3,814 のサンプルを分析対象とした．

「語りかけ性」についてのアノテーションにあたっては，「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」「特に語りかけ性はない」の三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（語りかけ性が）ある」は 334 サンプル（本稿で扱うサンプルの 8.76%），「どちらかといえば（語りかけ性が）ある」が 449 サンプル（本稿で扱うサンプルの 11.77%）得られた．

また，判断時に用いた表現や印象が備考欄へコメントとして記述されている場合がある．作業者によって記述量にはばらつきがあるが，「語りかけ性」に関しては作業者ごとにそれぞれの作業サンプル数の 2%～5%のコメントを得た．

表２．分析対象データ

語りかけ性の有無	とてもある	どちらかといえばある
サンプル数	334	449
語数	1,009,172	1,350,107

[2] 対談，Q&A 形式，図解，用語解説など形式的に特徴のあるサンプルは，分類対象外（非対象）とされた．作業者は，分類対象としたサンプルのみ観点付与を行っている．

４.「語りかけ性」を有すると判断される書きことばの表現

４.１　分析手法

まず，作業者が，「語りかけ性」があると判断したサンプルに含まれた表現について，特徴的に見られるはずの表現（保田ほか，2012）を確認し，「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」と判断されたそれぞれのサンプル群の差異を確かめる．

形態素解析にあたっては，MeCab 0.98＋UniDic 1.3.12 を用いた。品詞・活用形・語彙素のそれぞれの分析については，形態素解析の結果に基づく．

「語りかけ性」が「とてもある」と判断されたサンプル群に多く確認できる表現は，「語りかけ性」を明らかとするものと考えられ，「どちらかといえばある」と判断されたサンプル群にも確認できるはずである．しかし，判断レベルに差が生じる場合には，「語りかけ性」が「どちらかといえばある」と判断されたサンプルでは，判断基準とされた表現の出現頻度が低いことが予測される．

次に，「語りかけ性」を有すると判断されたサンプルの属するジャンルと販売対象の関連性を調べ，ジャンル等による特性を見る．同様に，「とてもある」「どちらかといえばある」の差異も確かめる．

さらに，作業者が自ら判断の根拠とした旨を記載したコメントを分類し，作業者が根拠とした表現が，実際のサンプル群にどのように現れているのかを調べる．

４.２　「語りかけ性」を有すると判断されたサンプルに含まれる表現

品詞と活用形，語彙素について，保田ほか（2012）で抽出された「語りかけ性あり」「語りかけ性なし」の二種類のアノテーション結果に見られた表現との対照を行い，「語りかけ性」が「とてもある」「どちらかといえばある」の差異を見る．

４．２．１　品詞

「語りかけ性あり」と判断されるサンプル群において，「語りかけ性なし」と判断されるサンプル群との有意差（有意水準 0.1%以下）の見られる品詞は，「ね」「よ」などを含む終助詞である．保田ほか（2012）では，「語りかけ性あり」サンプルでは，終助詞がサンプル中に 0.5%程度の出現率であり，「語りかけ性なし」サンプルでは 0.2%に留まる．本稿の調査においても，同程度の出現率が期待される．

本稿の調査では，「語りかけ性」が「とてもある」における終助詞の出現率は 0.69%で，「どちらかといえばある」における出現率は 0.41%であった．「語りかけ性」があると判断されるサンプルにおいては，終助詞の出現頻度が高いことが確かめられたといえる．

また，「とてもある」と「どちらかといえばある」では，終助詞の出現率に 1.7 倍の差異が見られている．そのため，「語りかけ性」があるという判断の根拠として，終助詞が利用されている可能性の高いことが考えられる．

４．２．２　活用形

「語りかけ性あり」と判断されるサンプル群には「でしょう」「だろう」などの意志推量形が 0.46%の出現率で見られ，「語りかけ性なし」とされたサンプル群における 0.24%という出現率とは有意差がある（保田ほか，2012）．

本稿の調査では，「とてもある」群における意志推量形の出現率は 0.41%，「どちらかといえばある」群では 0.31%であった．

但し，「どちらかといえばある」群では，「語りかけ性なし」とされたサンプル群の出現率に近い．意志推量形の出現頻度が低いことが「語りかけ性」があることを妨げず，「語りかけ性」があることに関わっているとは考えられる。しかし，「語りかけ性」があるという判断の根拠として積極的に用いられているとも言い切れないことが明らかとなった．

４．２．３　語彙素

「語りかけ性あり」と判じられるサンプルに有意に出現すると期待される語彙素について，本稿の調査との対照を行った．結果は表３である．

「語りかけ性」の有無に関わると考えられる語彙素は，本稿の調査データにおいても類似した出現率が見られている．また，「とてもある」と「どちらかといえばある」では，「語りかけ性がある」サンプル群に見られる表現の中でも，「とてもある」に顕著である．

但し，「自分」については，語りかけ性が「どちらかといえばある」サンプル群において，「語りかけ性がない」サンプル群とほぼ等しい出現率であり，「語りかけ性」があるという判断を妨げるのではないが，積極的に関係しているとも言い難いことがわかった．

表３．語りかけ性の有無に関わると考えられる語彙素と本稿の調査における出現率

語彙素	品詞	とてもある	どちらかといえばある	ある（保田ほか，2012）	ない（保田ほか，2012）
た	助動詞	2.03%	2.14%	1.65%	3.41%
ます	助動詞	1.69%	1.03%	1.35%	0.09%
です	助動詞	1.15%	0.64%	1.19%	0.05%
事	名詞-普通名詞-一般	0.81%	0.73%	1.05%	0.55%
自分	名詞-普通名詞-一般	0.17%	0.11%	0.25%	0.10%
ね	助詞-終助詞	0.13%	0.06%	0.09%	0.01%
よ	助詞-終助詞	0.13%	0.06%	0.04%	0.01%
貴方	代名詞	0.07%	0.03%	0.04%	0.01%

４.３　「語りかけ性」を有するサンプルと図書分類コード

ジャンル（NDC コード）と販売対象（C コード）を確認しておく．

本稿の調査対象とした「語りかけ性」を有すると判断されたサンプル群の NDC コードと C コード別の分布は，以下の図１・２の通りである．それぞれ、「語りかけ性」が「とてもある」「どちらかといえばある」とアノテーションされたサンプルをあわせたグラフである．

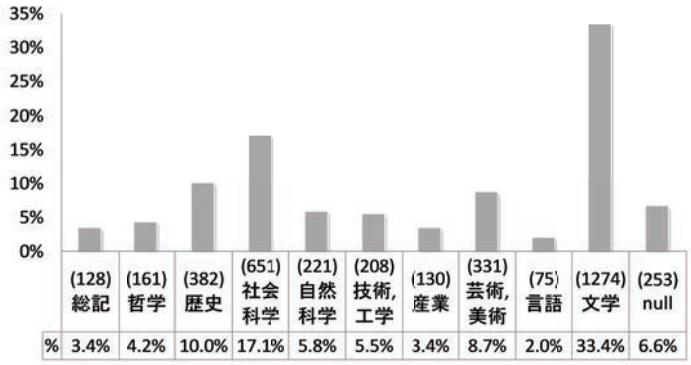


図１．語りかけ性があると判断されたサンプルの NDC コード別分布

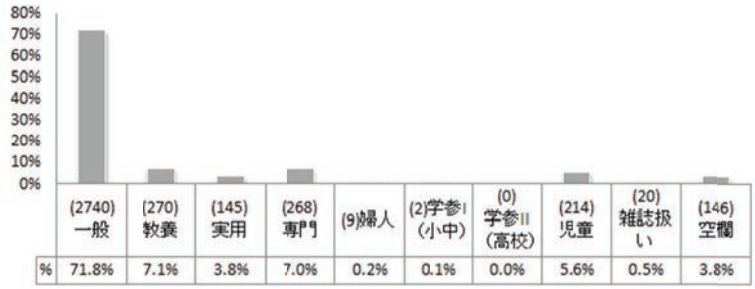


図２．語りかけ性があると判断されたサンプルの C コード別分布

図１において文学の割合が多いのは，幼児・児童向けのサンプルが含まれるためであり，図２における「児童」を販売対象とした書籍の割合と関係があることがわかる．NDC コードの分布から，全分野に「語りかけ性」が見とめられるのだといえよう．

図２の販売対象分布では，「児童」のほか「実用」が特徴的である．実用書は，いわゆるハウツーものの可能性が高いためである．

次に，「とても語りかけ性がある」と「どちらかといえば語りかけ性がある」の差異を見てみたい．選択肢毎に，図３は NDC コード，図４は C コードの分布を示している．

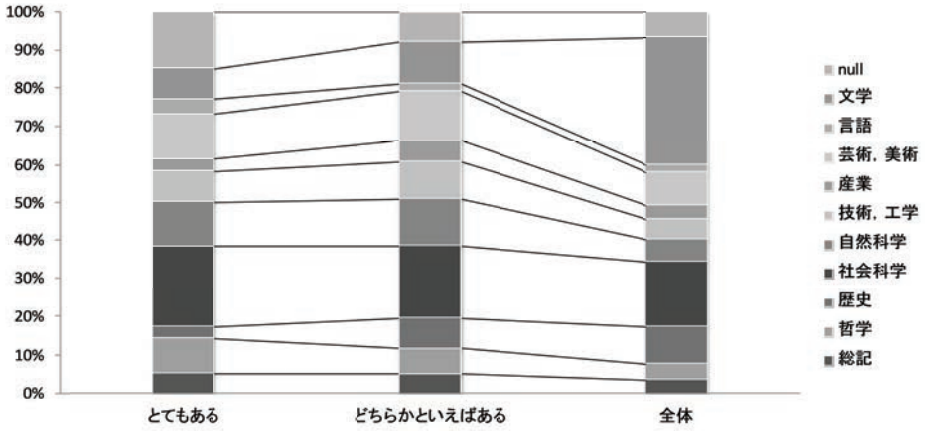


図３．語りかけ性があると判断されたサンプルの NDC コード別分布（選択肢別）

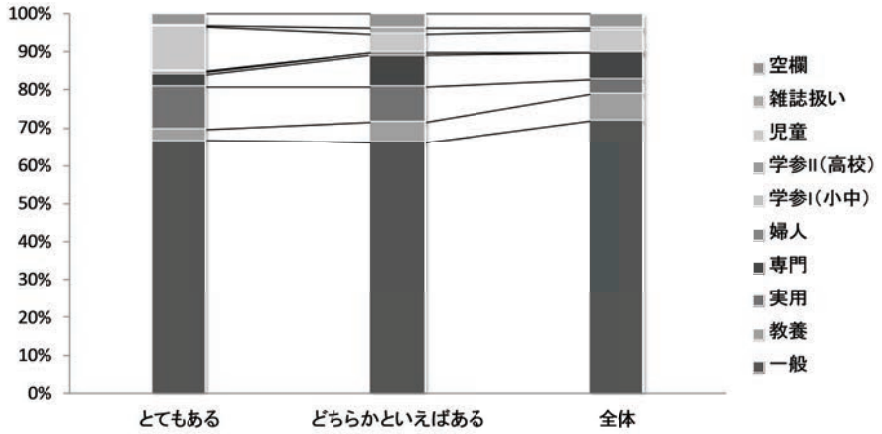


図４．語りかけ性があると判断されたサンプルの C コード別分布（選択肢別）

図３では，語りかけ性が「とてもある」サンプル群は，哲学でのみ「どちらかといえばある」に比して割合が大きいが．哲学分野には啓蒙書が含まれることから，明らかに語りかけている書籍に対して「とてもある」と判断されることが予測される．

また，図４を見ると，語りかけ性が「とてもある」サンプル群は，「どちらかといえばある」に比べ，「実用」と「児童書」の割合が多いことがわかる．従って，「教養」「専門」などでは，「どちらかといえばある」の割合が高い．作業者は語りかけ性があると判断しても，「とてもある」とまでは言い難い場合には，「どちらかといえばある」を選択するため，このような分布になると考えられる．

なお，C コードで「実用」と分類されていない書籍であっても，「語りかけ性」があると判断される書籍の種類には，いわゆるハウツーものが多いことが明らかである．「語りかけ性」があるとされた書籍は，「とてもある」で，その 53%がタイトルに「入門」「～法」「～使い方」「～術」「～ハンドブック」「～がわかる本」「～マニュアル」などが含まれ，「どち

らかといえばある」では同様に 30%,「語りかけ性」があるとされるサンプル群全体で 40%以上であった．タイトル例は以下のようなものがある．

『気持ちよく家族を介護するための本』『能力と意欲を伸ばす積極育児法』
『基礎から楽しく学べるはじめての陶芸入門』『はがきの書き方』
『資産ゼロから大成功する「魔法の粉」の使い方』『消化器ガン克服マニュアル』
『居住者のためのマンション管理の手引き』『新カラーコーディネート術』
『動物の見つけ方、教えます！』『定年後は山歩きを愉しみなさい』など

NDC コードや C コードの分類には現れないが，啓蒙書や指導書であることが示されているといえる．

４.４　「語りかけ性」に関する作業者コメント

作業者は,「語りかけ性」が「どちらかといえばある」という判断を行った際，コメントを記述する傾向がある.「とてもある」と判断されたサンプルにコメントがある場合は,「明らかにあなたに語りかけている体」などにとどまっている.「とてもある」とまでは言い難いが,「語りかけ性」があると感じた場合に「どちらかといえばある」を選択し，その判断根拠を示したものと考えられる．

挙げられた判断根拠は表４にまとめた.「昔話」は内容を示すが，コメントのあったサンプルは,「むかしむかし」で始まるいわゆる物語調のテキストである．表現としては，読み手との一体感を生じる「私たち」「われわれ」のような人称のほか，主として文末表現が注目されている．また,「多い」と感じたとされる文末表現以外でも，読み手や読み手の判断を想定した表現が現れると,「語りかけ性」があるとの印象になることがわかる．

表４.「語りかけ性」があると作業者が判断した根拠（作業者コメント）

・昔話
・ハウツーものに見られる表現：
 -希望：「～てください」「～てほしい」「～したい」
 -注意・禁止：「～ように」「～に注意」「～はいけない」
 -勧誘：「～てごらん」「～しよう」
 -可能：「～はず」「～できます」
 -評価：「～が大切」「～を要する」「～が便利」「～が狙い目」
　　（その他「よい」「悪くない」などの評価表現もコメントあり）
・人称：「私たち」「われわれ」が多い
・文末表現：
 -「のである」「わけです」「からです」「ものです」などが多い
 -読み手を想定した表現：
　「いただく」「申し上げる」「～の方（かた）」「ございます」
　「あげる」「させてください」「お聞かせすることにしよう」
 -婉曲表現：
　「～と思う（見える／感じ）」「～でしょうか」「～だろう」
 -読み手の判断を想定した表現：
　「思うかもしれません」「おわかりのことでしょう」

前述した通り,「語りかけ性」があると判断されたサンプルは，いわゆるハウツーもの書籍が 4 割を占めるため，ハウツーものに出現する表現に関するコメントが多く見られた．ハウツーもののテキスト例は以下のようなものが典型的であるといえる．

桜の花の塩漬けは、デパートにでもどこにでも売っています。あられは日本橋にぶぶづけ用の、塩だけで味をつけて炒り上げたかわいいあられを売っている店があり、これなら最適ですが、要するにしょうゆを使わない塩味のあられか、かきもちならなんでもいいのです。

ご飯のおこげを形よく包丁して、塩をふっても間に合います。ただ飯粒が乱れてとぶと、いかにも洗い流しのようで気持ちが悪いですから、焦げていない反対側のほうもちょっと焦げ目のつく程度に焼いて使いましょう。(LBr5_00039『季節を料理する』)

なお,「多い」「頻出する」という作業者の印象や，作業者が判断根拠とした表現が，実際のサンプルにどのように現れていたのかを見ておく．

表５は，文末に「多い」ことを作業者が判断根拠にしたという表現の，本稿におけるデータ中に現れた数である．また，一文が平均的な文長と仮定したとき（30 形態素）の，文末に出現する頻度を算出した.「語りかけ性」が「とてもある」サンプル群では約 12%,「どちらかといえばある」サンプル群では約 10%の文末に上り，作業者の印象は「多い」に至るものと考えられる．

また,「です」「ます」の出現率が,「とてもある」と「どちらかといえばある」の差異であるため，常体と敬体の出現率が異なっている．

表５．語りかけ性があるとの判断根拠にされた文末表現の出現率

	とてもある	文末出現率	どちらかといえばある	文末出現率
のだ	148	0.44%	556	1.24%
のである	75	0.22%	635	1.41%
のです	2,183	6.49%	1,485	3.30%
わけだ	16	0.05%	53	0.12%
わけである	2	0.01%	34	0.08%
わけです	246	0.73%	165	0.37%
からだ	23	0.07%	121	0.27%
からである	15	0.04%	93	0.21%
からです	219	0.65%	144	0.32%
ものだ	45	0.13%	93	0.21%
ものである	18	0.05%	113	0.25%
ものです	323	0.96%	240	0.53%
ことだ	37	0.11%	128	0.28%
ことである	27	0.08%	177	0.39%
ことです	460	1.37%	243	0.54%
計	3,837	11.4%	4,280	9.5%

作業者のコメントとして，語りかけている印象のある表現の量がサンプル全体のテキスト量に対して「少ない」か，あるいは「一部分」であったため，サンプル全体としては「語りかけ性」がないという判断をする旨の記載が散見されている.「一部分」には,「冒頭のみ」「文章末のみ」の追記のある場合も見られる．

すなわち，作業者が「語りかけ性」があるとの根拠に用いる表現の位置は,「文章末のみ」である場合に,「どちらかといえばある」を選択した旨が記述された例もあったが，コメントの記載されている場合には，ほぼ「文章末のみ」に語りかけている部分があったものの，全体としては「特にない」を選択したとの内容である．反対に,「冒頭のみ」であると記載されている場合には完全に「特にない」が選択されている.「文章末のみ」がサンプル全体の印象に関わる可能性はある．しかし，判断根拠とする表現が見つかっても，サンプル全体のテキスト量に対して「少ない」との印象があれば,「語りかけ性」があるとは判断してない傾向があるということがわかった．

もっとも，アノテーションの行われたサンプルは書籍の全文ではない．サンプル全体における表現の量を判断根拠とする傾向があるのならば，以降に「語りかけ性」のない文が続いている「冒頭のみ」よりも，サンプル内の文字数によっては,「語りかけ性」のない文

の量が予測できない「文章末のみ」に見つかった場合のほうが,「語りかけ性」があるとの判断を行う頻度が上がる可能性があろう.

５．まとめ

本稿では,「語りかけ性」についてアノテーションが行われた 3,814 のサンプルのうち,作業者によって「語りかけ性」があると判断されたサンプル群の分析を行った.

特に,「語りかけ性」が「とてもある」「どちらかといえばある」の程度の異なる判断が為されたサンプル群の対照を行った.その結果,「語りかけ性」があると判断されたサンプル群に出現率の高い表現が,「とてもある」と積極的に判断根拠とされているのかどうかを確認することができた.「とてもある」に出現率が高い表現でも,「どちらかといえばある」では「語りかけ性なし」サンプルに近い出現率となっていることがある.この場合,「語りかけ性」があるサンプルに特徴的な数値であるとしても,判断根拠とは言い難い.

また,アノテーション作業者が判断の根拠とした旨を記載したコメント欄を分析し,同時に実際のサンプルを確かめることを行った.出現率が低くとも,「語りかけ性」があると判断する根拠となる表現がどのようなものかが明らかとなった.

謝　辞

本研究は,国立国語研究所の共同研究プロジェクト「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

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』B5-6.

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.

保田祥,柏野和佳子,立花幸子,丸山岳彦(2012)「「語り性」を有する書きことばの典型例の分析」『第 1 回コーパス日本語学ワークショップ』予稿集, pp.139-146.

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/

中古和文における長単位の概要

冨士池　優美（国立国語研究所　コーパス開発センター）†

Outline of Long-unit-word in the Early Middle Japanese

Yumi Fujiike　(Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では「通時コーパスの設計」プロジェクトを中心に、歴史的な日本語のコーパス構築の準備が進められている。「通時コーパス」に格納される資料の一つに、源氏物語を中心とした中古和文がある。通時コーパスの形態論情報については、言語単位として、コーパスからの用例収集に適した「短単位」と、格納したサンプルの言語的特徴の解明に適した「長単位」の 2 種類を採用する。この 2 種類の言語単位について代表形・品詞等の情報を与える。

中古和文については、長短 2 種類の言語単位のうち短単位は中古和文 UniDic[1]で形態素解析後、人手修正をする形でデータ整備が進められており、認定基準については既に小椋・須永（2012）にまとめられている。これに対して長単位の認定基準については現在検討中の段階である。本稿では長単位の認定基準について、その概要と課題について述べる。また、長単位解析の現状について、あわせて報告する。

２．長単位の概要

「通時コーパス」中古和文の長単位は、『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ）で採用した単位を中古和文用に修正・拡張する方向で検討を進めている。この BCCWJ の長単位は、『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese、以下 CSJ）で採用した単位を書き言葉用に修正・拡張したものであり、これまでに国立国語研究所が実施してきた語彙調査における、長い単位の系列[2]に属するものである。

長単位は文節を基にした単位である。長単位の認定は、文節の認定を行った上で、各文節の内部を規則に従って自立語部分と付属語部分に分割していくという手順で行う。そのため、長単位の認定基準は、文節と長単位、二つの認定基準から成る。

本節では、文節の認定基準、長単位の認定基準、長単位の認定基準に関する BCCWJ からの変更点、長単位の長所及び認定基準に関する検討課題について述べる。以下、例文中の

† yfujiike@ninjal.ac.jp

1 中古和文 UniDic については小木曽ほか（2012）を参照。

2 これまでに国立国語研究所が実施してきた語彙調査における調査単位の概要については、小椋ほか（2011）pp.1-4（第 1 章『現代日本語書き言葉均衡コーパス』の言語単位）を参照。

文節の境界を「┃」、長単位の境界を「｜」とし、注目している境界を「‖」、切らないことを示す場合には「-」を、中でも注目している部分には「=」を用いる。また、注目している単位には下線を付す場合がある。

２．１　文節の認定基準

長単位の認定にあたっては、まず文節の認定を行う。現代語では文節は一般に付属語又は付属語連続の後で切れるが、中古和文においては、文節の切れ目に当たる付属語がないことが多いため、付属語を伴わない自立語については特に構文的な機能に着目して文節を認定することになる。文節を認定する上で問題となることの一つに、固有名、付属語を含む敬語表現、「―が～」「―つ～」「―の～」で１短単位と認める体言句、表記上分割不可能なものがある。これらについては、内部にある助詞・助動詞の後では切らないこととする。

┃源=頼朝┃　┃小野=小町┃　┃壇=ノ=浦┃　┃造ら=せ=たまふ┃
┃雁=が=音┃　┃わた=つ=うみ┃　┃天=の=川┃　┃尋常（よ=の=つね）┃

２．２　長単位の認定基準

長単位は、規定に基づいて文節を分割する、あるいはしないことによって得られた要素を１単位とする形式であり、文節を超えることはない。文節・長単位・短単位の関係を次ページの表１に示す。

以下、長単位認定基準の概要を示す。

〔１〕区切り符号は１長単位とする。

｜時めきたまふ｜あり｜けり｜。｜

〔２〕付属語は１長単位とする。

｜あら｜ぬ｜が｜、｜

ただし、「～おはす・おはします・きこゆ・たてまつる・たまふ・はべり」という形式の敬語表現の中に現れる付属語の後ろでは切らない。

｜弾か=せ=たまふ｜

〔３〕助詞・助動詞を伴わない自立語は、主語・主題、連用修飾、連体修飾の各成分の後ろで切る。

｜いと‖はしたなき‖こと‖多かれど｜

〔４〕体言に形式的な意味の「す」が直接続く場合、体言と「す」とを切り離さない。

｜心づかひ=し｜て｜　｜消息=せよ｜

〔５〕並列の関係にある語は切り離す。

｜四位‖五位｜こきまぜ｜に｜　｜あまたたび｜傾き‖あやしぶ｜

〔６〕同格の関係にある体言連続は切り離さない。

｜母=北の方｜なむ｜いにしへ｜の｜人｜の｜

〔７〕数を表す要素を含む自立語は、以下のように長単位を認定する。

(1)数を表す要素は、単位の変わり目の後ろで切る。

｜三尺‖六寸｜ばかり｜

(2)数を表す要素の前で切る。

｜長さ‖二十丈｜、｜広さ‖五丈｜

(3)数を表す要素とそれに続く体言・接辞とは切り離さない。

｜ひとつ=后腹｜

表１　文節・長単位・短単位の関係

文節	長単位	長単位 語彙素	長単位 語彙素読み	長単位品詞	短単位
B				空白	
	いづれ	何れ	イズレ	代名詞	いづれ
	の	の	ノ	助詞-格助詞	の
B	御時	御時	オオントキ	名詞-普通名詞-一般	御
					時
	に	なり	ナリ	助動詞	に
	か	か	カ	助詞-係助詞	か
	、	、		補助記号-読点	、
B	女御	女御	ニョウゴ	名詞-普通名詞-一般	女御
	、	、		補助記号-読点	、
B	更衣	更衣	コウイ	名詞-普通名詞-一般	更衣
B	あまた	数多	アマタ	副詞	あまた
B	さぶらひたまひ	侍ひ給う	サブライタマウ	動詞-一般	さぶらひ
					たまひ
	ける	けり	ケリ	助動詞	ける
B	中	中	ナカ	名詞-普通名詞-一般	中
	に	に	ニ	助詞-格助詞	に
	、	、		補助記号-読点	、
B	いと	いと	イト	副詞	いと
B	やむごとなき	やんごとなし	ヤンゴトナシ	形容詞-一般	やむごとなき
B	際	際	キワ	名詞-普通名詞-一般	際
	に	なり	ナリ	助動詞	に
	は	は	ハ	助詞-係助詞	は
B	あら	有り	アリ	動詞-一般	あら
	ぬ	ず	ズ	助動詞	ぬ
	が	が	ガ	助詞-格助詞	が
	、	、		補助記号-読点	、
B	すぐれ	優る	スグル	動詞-一般	すぐれ
	て	て	テ	助詞-接続助詞	て
B	時めきたまふ	時めき給う	トキメキタマウ	動詞-一般	時めき
					たまふ
B	あり	有り	アリ	動詞-一般	あり
	けり	けり	ケリ	助動詞	けり
	。	。		補助記号-句点	。

２．３　BCCWJ からの変更点

（１）文節認定基準における主語・主題に関する変更

助詞・助動詞を伴わない自立語に関しては、文節を「主語・主題の後ろで切る」という規定が CSJ にあった。これに対して、BCCWJ は書き言葉を対象としておりこの規定が適用されることが少なく、「持続可能な」「センス抜群の」のような漢語形状詞を述部に持つものの文節認定の判断が難しいこともあり、規定を削除していた。しかし、中古和文においては、助詞を伴わずに主語・主題を示すことが多いため、この規定が再び必要となった。

（2）並列に関する変更

CSJ では並列を切り離していたのに対して、BCCWJ では、漢語の並列を中心に、並列と見るか 1 語と見るかの判断が困難な例が頻出したため、切り離さないこととしていた。

BCCWJ：｜公正=妥当｜な｜実務慣行｜

これに対して、中古和文では漢語の並列のような問題はあまり起こらないため、並列は切り離すこととした。この規定により、複合動詞と動詞の並列が区別され、語と語との係り受けが明確になる。

中古和文：｜あまたたび｜傾き‖あやしぶ｜

２．４　長単位の長所

ここでは長単位と短単位の違いに着目して、長単位の長所を示す。短単位は、基準がわかりやすくゆれが少ないため用例収集には便利であるが、合成語を構成要素に分割してしまう場合が少なくない。短単位と比較したとき、長単位ではこのような場合にも一般的な古語辞典の見出し語に近い形式が取り出されることが大きな違いと言える。そのほか、品詞について、以下のような違いがある。

（1）品詞付与方針

短単位と長単位の品詞体系は共通であるが、品詞付与方針が異なる。短単位では可能性を考慮した品詞を付与しており、名詞-普通名詞-形状詞可能[3]等がある。これに対して長単位では文脈に即して品詞を付与する方針をとり、名詞-普通名詞-〇〇可能といった品詞は設けない。長単位末尾に位置する短単位の品詞が名詞-普通名詞-｛副詞・形状詞・サ変形状詞可能｝の場合、その用法に基づき名詞・副詞・形状詞に判別する。「哀れ」（名詞-普通名詞-形状詞可能）を例とすると、「もののあはれ知りすぐし、」の場合は名詞を、「皇子もいとあはれなる句を作りたまへるを」の場合は形状詞を付与する。

（2）合成語の扱い

短単位では最小単位の 1 回結合を最大とするのに対して、長単位では結合回数の制限なく、合成語を認める。このとき「愛敬づく」のように、合成の結果として品詞が変わることがある。また、短単位では接辞を切り離していたのに対して、長単位では接辞を含めた形式が 1 長単位となる。接辞は品詞に影響を与える。中古和文では、漢語接頭辞がほぼ用

[3] 「形状詞可能」は、名詞としても形状詞としても使われ得ることを示す。

いられないことから接頭辞による品詞の変化は見られないが、接尾辞が付加することによって品詞が変わることが多い。

｜物語-めく｜	名詞＋接尾辞	→ 動詞
｜たへ-がたし｜	動詞＋接尾辞	→ 形容詞
｜うつくし-げ｜	形容詞＋接尾辞	→ 形状詞

上に挙げたような違いにより、より精密化された品詞情報を活用することができることが長所と言えるだろう。

２．５　認定基準に関する検討課題

（１）並列の認定

先に述べたように、中古和文では並列の扱いを BCCWJ から変更して、原則として分割することとした。ここで、問題になるのが、代名詞や副詞の並列である。例えば「こなたかなたの人々など」の「こなた」「かなた」は代名詞の並列であるが、係り受けを意識した場合には「こなたかなた」全体で「人々」にかかると見るのが妥当だろう。副詞に関しては、「と」「かく」のバリエーションが問題になる。例えば「と見かう見、見けれど、」の場合は「と」「かう」をそれぞれ副詞とすることに問題はないが、「とざまかうざまにこころみきこゆるほど」となると、「とざま」「かうざま」を並列と見てそれぞれ文節認定するのか、「とざまかうざま」全体で「こころみきこゆる」にかかると見るのか、判断が難しいところである。これらについては係り受けを考慮しながら規定を整備する必要がある。

（２）複合辞・連語の扱い

BCCWJ では複合辞を付属語（助詞・助動詞相当句）として認めていた。中古和文についても、いくつかの語が複合して助詞・助動詞のような機能を持つ形式が存在する。例えば、BCCWJ で認めていた複合辞「という」と同じ語構成の「といふ」は、中古和文でも「愛宕といふ所」のように、現代語同様の助詞相当句として用いられている。共通の形式以外に、「伏籠の中に籠めたりつるものを」「山がつになりて、いたう思ひくづほれはべりし年ごろの後、こよなく衰へにてはべるものを」の「ものを」は詠嘆・逆接を表すものとして、中古和文では助詞相当句とみることができるだろう。このように複合辞らしき形式が存在することは確かであるが、ここで問題となるのはその選定である。BCCWJ では先行研究における扱いや頻度、機能的用法の割合等に基づき複合辞を選定したが、中古和文については、個別の表現に関する検討はあるものの、複合辞となる表現を集めたリスト類がないのが現状である。複合辞を認めるかどうかがまず問題であるが、認めるとした場合には複合辞選定基準の検討が大きな課題となる。

BCCWJ では複合辞のほか、連語を認めている。これは CSJ や先行研究のほか、連濁や係り受けを考慮して、全体で 1 長単位とすることが妥当と考えられる語を人手修正作業の過程で抽出したものである。中古和文についても同様に、連語を認める方針である。「知らず読み」のような付属語を中間に持つ語等、人手修正作業の過程で抽出していく。

（３）固有名の扱い

中古和文においては、女性の呼称を中心に、「朧月夜」のような一般名詞や、「明石の御方」「藤壺」「中務」のように地名・建物名・官職名を用いて人（名）を表すため、短単位の品詞情報を基に固有名の品詞認定をすることが難しい。実例に基づき、品詞付与基準を整備していくことが今後の課題となる。

３．中古和文の長単位解析の現状

長単位は規定としては文節を分割して取り出す言語単位であるが、実際には短単位を基に自動解析する。中古和文の長単位自動解析にあたっては、短単位から長単位を自動構成する解析器 Comainu ver.0.60 [4]を用いた。解析器の学習には BCCWJ のコアデータが用いられており、中古和文への特別な対応はなされていない。

現在、源氏物語桐壺巻の人手修正済み短単位データ 6500 語[5]を基に、長単位の自動解析を行い人手修正を施した。この結果、6500 短単位は 5760 長単位となった。このうち 5166 長単位（89.7%）は短単位と境界が一致しており、長単位構成要素数（1 長単位を構成する短単位数の平均）は 1.13 である。これを BCCWJ コアデータ[6]と比較してみよう。BCCWJ コアデータにおける媒体別長単位構成要素数[7]と短長境界が一致する（1 短単位から構成される）長単位の割合[8]を表 2 に示す。源氏物語桐壺巻は長単位構成要素数が BCCWJ のどの媒体よりも少なく、短長境界が一致する長単位の割合も高い。ここから、源氏物語桐壺巻には合成語が少ない様子が見てとれる。

表２　BCCWJ コアデータにおける媒体別長単位構成要素数

媒体	書籍	雑誌	新聞	白書	Yahoo!知恵袋	Yahoo!ブログ
構成要素数	1.18	1.23	1.32	1.44	1.16	1.18
短長一致(%)	86.7	84.7	79.7	74.7	87.6	―

次に、長単位自動解析の精度について見る。源氏物語帚木～花宴巻及び伊勢物語について計 72720 短単位から記号類を除く 1000 語をサンプリングチェックしたところ、精度は語彙素認定で 95.1% [9]であった。以下、精度確認の結果から長単位境界の認定に関する典型的な誤りについて、具体的にどのような事例があるのかを見ていく。

---

4 小澤ほか（2011）を参照。

5 短単位データは中古和文 UniDic により形態素解析した後、人手修正したものを用いた。

6 「コアデータ」とは、形態素解析システムや長単位自動解析器等の学習用データとして、自動解析後に人手修正を施した高精度のデータセットである。内訳は書籍・雑誌・新聞・白書それぞれ約 20 万語と Yahoo!知恵袋・Yahoo!ブログそれぞれ約 10 万語である。

7 山崎（2011）を参照。

8 冨士池（2010）を参照。

9 精度 95.1%、つまり 1000 語のうち、49 の誤りがあったことになる。内訳は境界の誤りが 28、品詞の誤りが 9、語彙素の誤りが 0、基データの短単位の誤りが 12 であった。

（1）接辞

誤）｜うち‖具し｜　→　正）｜うち=具し｜
｜なま‖いとほし｜　→　｜なま=いとほし｜
｜頼み‖がた｜　→　｜頼み=がた｜

接辞に関しては、BCCWJ と中古和文とで認定基準上の大きな違いはない。そのため、接辞を含む形式であっても大半は正しく解析されていた。その中で誤解析となった接頭辞「うち」「なま」は中古和文で新たに接頭辞とされたものである。そのほか、現代語では比較的頻度の低い「ほの」についても同様の誤解析が見られた。「頼みがた」については、接尾辞「難い」自体は現代語で用いられるが、活用形「語幹−一般」は現代語にはない形式であった。これらの接辞の誤解析は長単位解析器の学習用データに BCCWJ、つまり現代日本語書き言葉を用いているために中古和文特有の形式に対応できていないことが原因と考えられる。

（2）敬語表現

誤）｜ものし‖たまふ｜　→　正）｜ものし=たまふ｜
｜弾か‖せ‖たまふ｜　→　｜弾か=せ=たまふ｜

「弾かせたまふ」については付属語を含む形式であり、中古和文で新たに規定に追加されたものであるため、接辞と同様の学習用データの問題と考えられる。一方、「ものしたまふ」は動詞 2 語として解析されたが、「たまふ」の品詞は「動詞−非自立可能」であり、「動詞−非自立可能」を後項に取る複合動詞というのは現代語で一般的な形式である。「給う」の頻度が BCCWJ で低いことが影響を与えているとも考えられるが、この問題の原因は不明である。

（3）複合動詞

誤）｜具‖し｜　→　正）｜具=し｜
｜法気‖づき｜　→　｜法気=づき｜
｜消え‖のこり｜　→　｜消え=のこり｜

複合動詞に関しては現代語でも十分あり得る品詞構成であり、上に挙げた例については特に問題がないように見え、誤解析の原因は不明である。

現時点での中古和文の長単位解析においては、長単位解析器の学習に現代語を用いていることが誤解析の大きな要因になっているようである。品詞構成上は現代語・中古和文共通であっても、現代語における頻度が低いことが影響しているようにも見えるが、それだけでもないようである。上に挙げた誤解析の不明な点に関しては、解析結果の分析を継続的に行っていく中で、原因を見出していきたい。今後、長単位の認定基準とともに、人手修正済みデータを整備していく。中古和文データを用いて長単位解析器の学習を行うことについては、今後の課題としたい。

４．終わりに

本稿では、「通時コーパス」の中古和文で採用する長短 2 種類の言語単位のうち長単位の認定基準の概要及び現在検討中の課題について説明した。また、長単位解析の現状についても報告した。長単位の認定基準については、今後「通時コーパス」の準備作業を進めていく中で、適宜修正・追加を行っていく予定である。

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。

文　献

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」、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 1
（http://dl.dropbox.com/u/73297026/report/unidic-EMJ_report2012.pdf よりダウンロード可能）

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）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上）」、国立国語研究所内部報告書
（LR-CCG-10-05-01）

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
（https://maro.ninjal.ac.jp/Comainu/ にリンクあり）

冨士池優美（2010）「『現代日本語書き言葉均衡コーパス』における長単位の構成要素について」、日本語学会 2010 年度秋季大会予稿集、pp.237-242

山崎誠（2011）「『現代日本語書き言葉均衡コーパス』の構築と活用」、『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集（JC-G-11-01）、pp.11-20

関連 URL

中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

中・長単位解析器 Comainu　https://maro.ninjal.ac.jp/Comainu/

助動詞レル・ラレルへの意味アノテーション作業経過報告

小山田由紀（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）

An Interim Report of the Semantic Annotation of Auxiliary Verb *reru* / *rareru*

Yuki Oyamada (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept. Corpus Studies, NINJAL)
Kikuo Maekawa (Dept. Corpus Studies, NINJAL)

1. はじめに

現在、助動詞レル・ラレルに対して意味のアノテーションを行っている。本稿では、その途中経過を報告する。このアノテーションは、レル・ラレルを意味別に分析するために必須であるとともに、文体分析の指標として受動文の生起率を計算するなどの目的にとっても必要である（Biber, 2009）。

周知のとおり、レル・ラレルには受身・尊敬・可能・自発の４つの意味があり[1]、その相互関係に関する先行研究も多数発表されている。このレル・ラレルの多義性を自動分類する可能性を探るべく、今回は人手によるアノテーションを行った。また、受身・尊敬・自発・可能の４つの意味以外に、レジスターを特徴づける表現について、レル・ラレルを含む表現全体の属性とみなして分類を試みた。

2. 作業対象

今回の研究では『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）のコアデータより抽出した、助動詞レル・ラレル全 7,906 件を作業対象とした。

コアデータ全体のレル・ラレルの出現数と語数（短単位数）とをレジスター別に表したグラフを図１に示す。レル・ラレルの出現数は各レジスターとも約２割であり[2]、出現数に特徴は見られなかった。語数も同じく各レジスターとも約２割であり違いが見られなかった。

また、書き言葉と話し言葉との比較のため、BCCWJ の国会会議録と『日本語話し言葉コーパス』（以下、CSJ と略す）も作業対象とした。国会会議録は中納言で抽出した助動詞レル・ラレル全 33,320 件から 1,666 件を、CSJ は全 43,590 件から 1,676 件をランダムサンプリングした。それぞれのサンプリング件数は、図１の値から各レジスター約 1,600 件と考えて決定した。

[†] oyamada-y@ninjal.ac.jp
[1] 尾上（1998a,1998b,1999）では、レル・ラレルが後続する動詞や可能動詞などを述語とする文を「出来文」と捉え、通常の４分類に「意図成就」を加えた５分類を提唱している。
[2] OC（Yahoo!知恵袋）と OY（Yahoo!ブログ）は、同じ web のデータとして合計した。

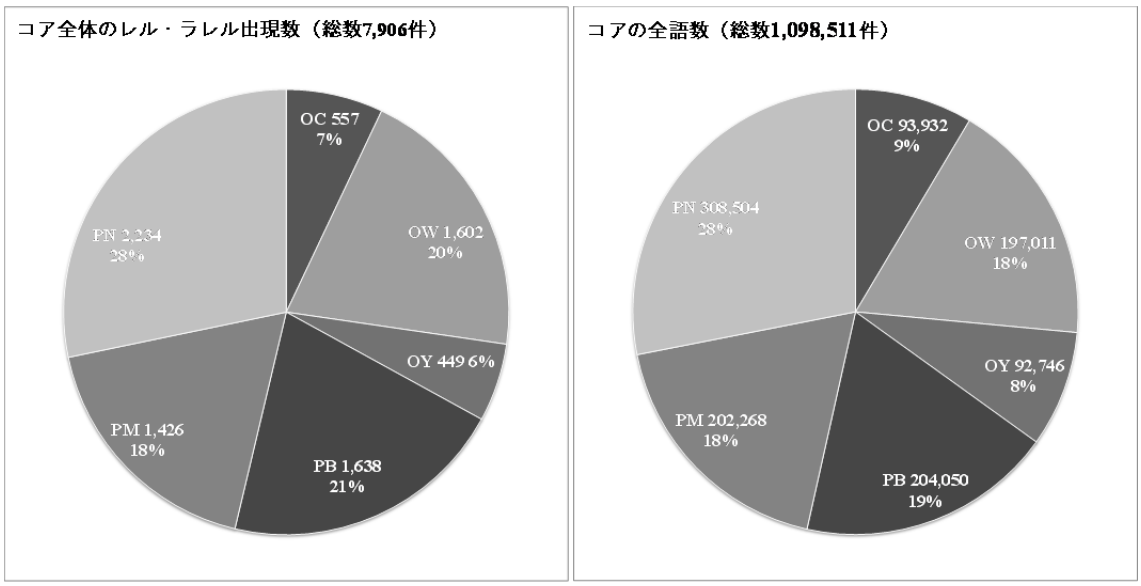


図１　BCCWJ コアデータにおけるレル・ラレル出現数と全語数

3.　作業経過

本章では、作業開始以来現在までの作業経過を時間に沿って記述する。

3.1.　BCCWJ コアデータに対する内部での試行

まず細かな基準は定めずに、BCCWJ コアデータのサンプル ID 下一桁が“1”の 764 件を対象に、筆者３名による判定を試行した。

次に、その判定結果を考慮してフローチャートを作成した。フローチャートは紙幅の関係で記載しないが、特徴としては、最初に動作主と態変換の有無を判定してからそれぞれの意味に分岐するというものであった。

また、新たに「受身（客観的）」という意味を追加した。これは「～と見られる」「～と考えられる」など、論文や白書、新聞などに使われる文体である。この種の表現についてはすでに志波(2009)が意味・構造的なタイプを分析しており、重なるところが多い。

3.2.　内部での再試行

BCCWJ コアデータのサンプル ID 下一桁が“5”の 556 件を対象に、作成したフローチャートを基づいて筆者３名で再試行した。

3.3.　外部作業者による試行

内部で判定作業済みの、BCCWJ コアデータのサンプル ID 下一桁が“5”の 556 件を対象に、筆者らを含まない作業者２名に試行を依頼した。筆者３名と判定結果が異なるものや作業者からの質問に対して、作業マニュアルを作成した筆頭著者がコメントした。

3.4.　外部作業者による判定作業

前節の外部作業者２名に、BCCWJ コアデータの残り 7,350 件の作業を依頼した。

その判定結果を考慮して判定項目を変更した。レル・ラレルの意味を４つに戻し、別項

目としてレル・ラレルを含む表現全体の特徴を判定する。詳細は次章で述べる。

そのため、BCCWJ コアデータ全 7,906 件に対して、変更後の判定項目に従って筆頭著者が判定・修正を行った。

3.5. 話し言葉データに対する判定作業

BCCWJ 国会会議録のデータ 1,666 件と CSJ データ 1,676 件を対象に、筆頭著者と外部作業者１名とで判定作業を行った。

上記作業における κ 統計値を表１に示す。この値から人手による意味判定作業の一致度は高いと言ってよいことが分かる。

表１　意味判定作業の κ 統計値

作業経過	κ 統計値
１．BCCWJ コアデータに対する内部での試行	0.683
２．内部での再試行	0.766
３．外部作業者による試行	0.671
４．外部作業者による判定作業	0.692
５．話し言葉データに対する判定作業	0.864

4. 現状のアノテーションの項目と仕様

4.1. 概要

3.4.に述べた外部作業者による判定作業では、以下の問題が見つかった。

- 作業者はフローチャートどおりには作業していないと思われる。
- 助動詞レル・ラレルに対するアノテーションなのか、レル・ラレルが付いた表現全体に対するアノテーションなのかが不明確なまま作業を進めていた。

そこでフローチャートによる作業をやめ、判定項目を再度変更した。レル・ラレルの意味を４つに戻し､別項目としてレル･ラレルを含む表現全体の特徴を判定することにした[3]。現状の判定項目を図２に示す。

図２の表現全体の特徴の３項目について説明する。

「心情誘導」は、「～に癒される」「～に惹かれる」など、“動詞＋(ラ)レル”全体で発話者の感情を表わし、自発の意味に近づく表現である。必ず意味は「受身」になる。出現数はかなり少ない。

「客観化」は、フローチャートでは新たに「受身（客観的)」として追加した意味である。「～と思われる」「～と見られる」「～と言われる」など、“動詞＋(ラ)レル”全体で発話者だけの意見でなく専門家や世論の見方として客観的な内容であるというニュアンスを持たせる表現である。意味は「受身」と「可能」のどちらかになる。

「存在確認」は、「～に傾向が見られる」「～に特徴が挙げられる」など、“動詞＋(ラ)レル”全体で第一義の動詞の意味から離れ対象の存在を表す表現である。必ず意味は「可能」になる。

[3] 態変換や影響の有無、動作主・対象・主語の特徴、結合価など、かなり細かな判定基準を設けて、実際の判定作業は行った。

「客観化」と「存在確認」のどちらも、志波(2009)がすでに指摘しており、どちらも論文や白書、新聞の文体に特徴的な表現である。

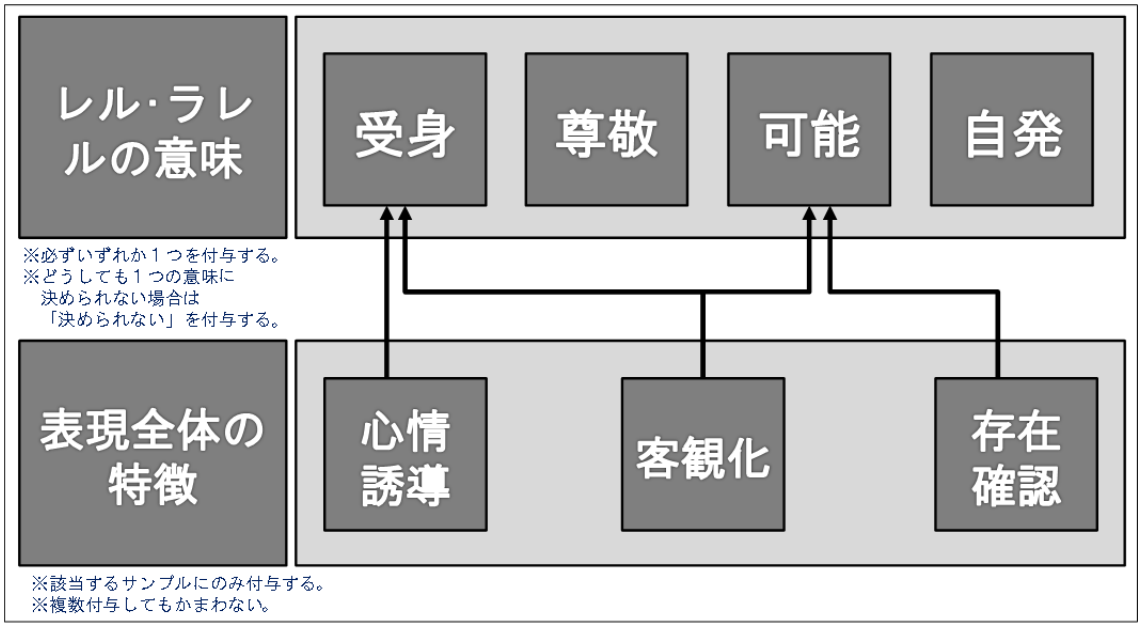


図２　現状のアノテーション項目

4.2.　客観化

論文や白書、新聞の文体に特徴的な表現である。「～と思われる」「～と見られる」「～と言われる」など、“動詞＋(ラ)レル”全体でモダリティ表現に近い働きをし、発話者だけの意見でなく専門家や世論の見方として客観的な内容であるというニュアンスを持たせる表現である。志波(2009)がすでに指摘しているところであるが、本作業での仕様を分かりやすく図示したのが図３である。

志波(2009)との大きな違いの１つは、主たる事象に対する態度の表現と捉えている点である。そのため、志波が判断型としているものについても、係り受けと関係なく、客観化として同じように扱う。例えば、図３の「捜査はお蔵入りと見られる」は、主たる事象は「捜査はお蔵入り（だ)」であり、それに対する発話者の態度が「と見られる」で表わされていると考える。

もう１つの大きな違いは、志波(2009)が分析対象としているサンプルは受身文であり[4]、筆頭著者が作成したフローチャートでも「受身（客観的)」として意味を「受身」と考えていたが、レル・ラレルの直前語彙素（動詞）の種類[5]と動作主とアスペクトの組み合わせで付与するレル・ラレルの意味が異なるよう判断基準を変えた点である。表２に示す。

基本形は通常反復・超時のアスペクトと言われる。思考系動詞の基本形を「可能」とした理由は、以下のとおりである。

- 動作主が意図していないことが重要な「自発」とは考えづらい。

[4] 「本研究は受身文を調査対象としている」と明記されている。「広義認識動詞に限り、自発・可能用法も調査対象に含めた」とあるが、合計 202 例中 25 例のみであり、多くが「受身」と捉えられている。
[5] 動詞の意味が直感的に合うようにしてはいるが、筆頭著者の恣意的分類である。

- 例えば論文などでは、思考を重ねた結論を述べる際に使われることが多く、ある程度事態は完了している。尾上(1999)の言う結果的成就を正面に出して表現する「意図成就」に近い表現と考える。

「存在確認」の意味を「可能」にしたことについても同じことが言える。

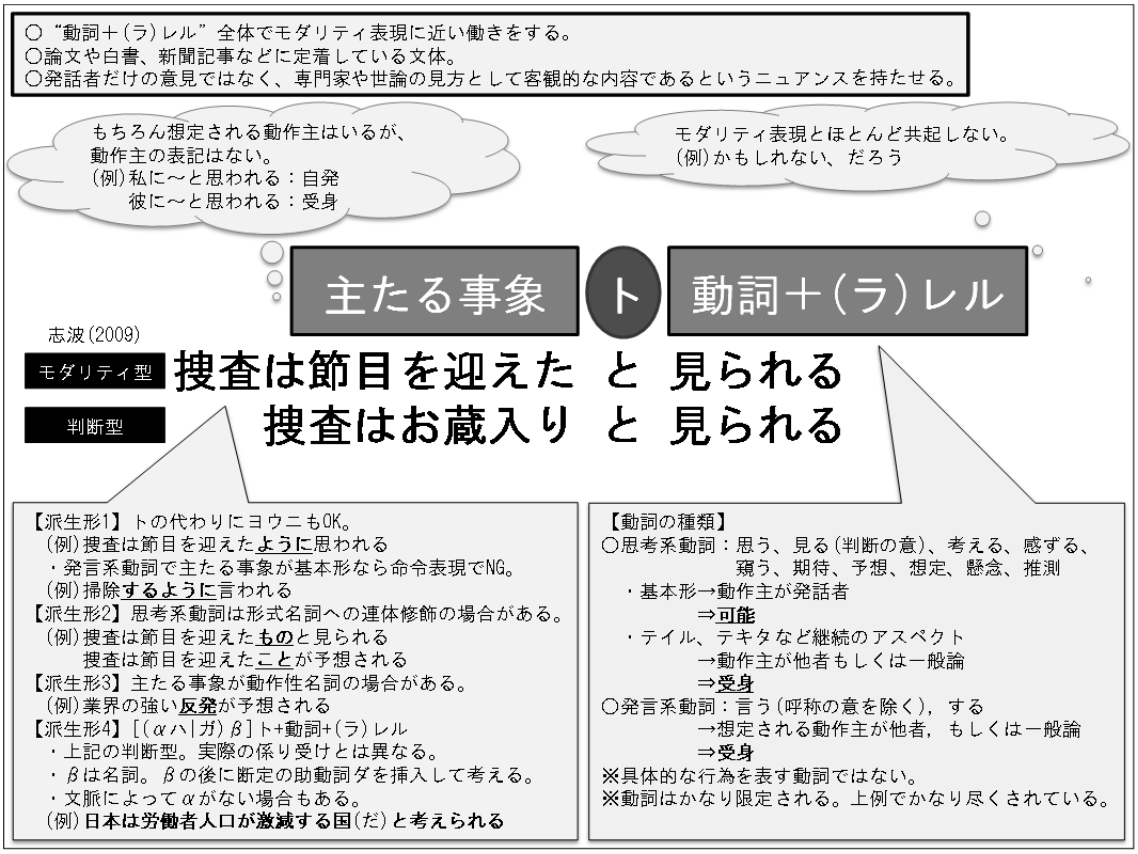


図３　客観化の判断基準

表２　客観化に付与する意味の条件

意味	直前語彙素	動作主	アスペクト	型の例
可能	思考系動詞	発話者	基本形	～と見られる
受身	思考系動詞	他者 or 一般的に	継続のアスペクト	～と見られている
受身	発言系動詞	他者 or 一般的に	－	～と言われる／～と言われている

5.　作業結果

5.1.　BCCWJ コアデータ

3.4.で述べたとおり、BCCWJ コアデータ全 7,906 件については、外部作業者による判定作業の後に判定項目の変更があったため、筆頭著者が修正を行った。修正の手順は以下のとおりである。

①　外部作業者２名の判定が一致するサンプルは、基本的に正解とする。
②　外部作業者２名の判定が異なるサンプルは、筆頭著者が改めて判定する。
③　判定項目・判定基準が変更になったものは、あたりをつけて修正する。

そのため、修正し切れていない可能性は否定できないが、現状の判定結果を表３に示す。85%強が「受身」、12%弱が「可能」である。４つのうちこの２つでほぼ占められる。

表３　BCCWJ コアデータの判定結果

	受身	尊敬	可能	自発	決められない
総計	6,771	169	944	19	3
割合	85.64%	2.14%	11.94%	0.24%	0.04%
客観化	222	0	332	0	0
存在確認	0	0	118	0	0
心情誘導	6	0	0	0	0

5.2.　話し言葉

話し言葉については、作業者それぞれの値を示す。

5.2.1. 国会会議録

国会会議録のデータ 1,666 件の判定結果を表４に示す。

特徴としては、「尊敬」の総計件数が異なることが挙げられる。国会会議録は対話形式であり、発話者や動作主をわざわざ発言しないという傾向が顕著である。そのため前後文脈から判断することになるが、発話者や動作主は何か、そもそも敬意を表しているのか不明なサンプルが多く、作業者によってその判断が分かれたのではないかと推測される。

表４　国会会議録データの判定結果

小山田＼作業者	受身	尊敬	可能	自発	決められない	総計
受身	1,043	63	18	0	0	1,124
尊敬	17	351	1	0	0	369
可能	30	2	133	1	0	166
自発	3	1	1	0	0	5
決められない	0	1	0	0	1	2
総計	1,093	418	153	1	1	1,666

5.2.2. 『日本語話し言葉コーパス』（CSJ）

CSJ データ 1,676 件の判定結果を表５に示す。判定結果自体に大きな問題はなさそうである。

注意を要する問題として、ら抜き動詞の扱いが BCCWJ と CSJ で異なることがある。例えば「見れる」は、BCCWJ では動詞「見る」の一語形という扱いになるが、CSJ では動詞「見る」と助動詞「れる」の２語となる。本作業では、ら抜き動詞は９サンプルのみであったが、理論上は BCCWJ よりも CSJ の方が「可能」が多くなる可能性がある。

表５　CSJ データの判定結果

小山田＼作業者	受身	尊敬	可能	自発	決められない	総計
受身	1,174	12	15	0	1	1,202
尊敬	7	53	1	0	0	61
可能	36	1	351	1	1	390
自発	1	0	0	1	0	2
決められない	4	2	10	0	5	21
総計	1,222	68	377	2	7	1,676

6.　考察

6.1.　レジスターに関する考察

助動詞レル・ラレルの意味判定結果（筆頭著者の判定・修正結果）をレジスター別に表にしたものが表 6 である。出現数を 100 万語（短単位）あたりの生起数に正規化して示す。

表６　レジスター別の意味判定結果

	BCCWJ								CSJ	
	OC	OW	OY	PB	PM	PN	コア全体	国会	学会講演	模擬講演
受身	4,876	6,893	3,914	6,861	6,022	6,399	6,164	4,408	5,030	4,063
尊敬	596	5	248	211	163	42	154	1,447	175	267
可能	447	1,233	668	912	836	784	859	651	2,087	809
自発	0	0	11	44	25	13	17	20	8	8
決められない	11	0	0	0	5	3	3	8	87	67
総計	5,930	8,132	4,841	8,027	7,050	7,241	7,197	6,533	7,387	5,214

CSJ には 5 つのレジスターが存在するが、ここでは用例の多い学会講演と模擬講演のみを扱う[6]。

まず、最も大きな特徴として、全体的に受身が圧倒的に多いことが分かる。

次に、レジスターによってレル・ラレルの意味の分布が異なり、その違いは対人配慮の有無が関係していると思われる。一般的に、対人配慮が多いレジスターは OC（Yahoo!知恵袋）と国会会議録であり、逆に対人配慮が少ないレジスターは OW（白書）と PN（新聞）と CSJ 講演と考えられる。CSJ 講演（特に学会講演）は不特定多数の聴衆を聞き手としているために、特定個人に対する対人配慮は生じにくいと考える。

① 「尊敬」は OC と国会会議録に多く、OW と PN に少ない。これは、「尊敬」は、対人配慮が必要なレジスターには生じやすく、対人配慮が少ないレジスターには生じにくいからだと説明できる。

② 「受身」は対人配慮が少ないレジスターに多く見られる。OW（となぜか PB[7]（書籍））に多く、OC、OY（Yahoo!ブログ）、国会会議録、CSJ に少ない。

③ 「可能」も対人配慮が少ないレジスターに多く見られる。OW と CSJ 学会講演に多く、

[6] CSJ の音声のタイプには、1 学会講演、2 模擬講演、3 対話、4 朗読と再朗読、5 その他がある。
[7] PB は小説や随筆など雑多であり、日本十進分類法（NDC）で分類すべきかもしれない。

OC に少ない。OW と学会講演に「可能」が多いのは、第４章に詳述した表現全体の特徴の「客観化」「存在確認」が多いからではないかと推測される。OC に「可能」が少ないのは、レル・ラレルよりも可能動詞を使う傾向があるからかもしれない。

6.2. 直前要素に関する考察

意味や属性の違いには、助動詞レル・ラレルの直前要素（主に動詞）が深く関わると考えられるため、BCCWJ コアデータ全 7,906 件に対する外部作業者２名の判定結果のうち判定が異なる（揺れる）サンプルについて考察を試みた。

レル・ラレルの直前語彙素を全体の個数順に表７に示す。動詞「為る」は、「為る」の直前語彙素が重要な場合がほとんどであり、「為る」の直前語彙素も表にした。どちらも紙幅の関係で全体の上位 20 位を挙げる。表現全体の特徴の「客観化」「存在確認」に関係する、思考系動詞（薄い灰色セル）と発言系動詞（濃い灰色セル）が出現数も揺れも多く、判定が難しいことが分かる。

表７　レル・ラレル直前語彙素（上位 20 位）

直前語彙素	揺れの個数	全体の個数	揺れ/全体
為る	192	3006	6.39%
行う	1	323	0.31%
言う	63	299	21.07%
見る	96	230	41.74%
考える	35	163	21.47%
呼ぶ	0	125	0.00%
開く	0	110	0.00%
使う	0	91	0.00%
求める	4	90	4.44%
思う	45	81	55.56%
知る	1	67	1.49%
認める	11	65	16.92%
居る	3	62	4.84%
得る	10	60	16.67%
含む	0	58	0.00%
作る	0	53	0.00%
書く	0	43	0.00%
置く	0	42	0.00%
付ける	1	40	2.50%
成す	0	39	0.00%

「為る」直前語彙素	揺れの個数	全体の個数	揺れ/全体
と	117	174	67.24%
開催	0	65	0.00%
期待	20	60	33.33%
実施	0	54	0.00%
化	0	44	0.00%
に	0	41	0.00%
逮捕	0	41	0.00%
設置	0	37	0.00%
発表	0	36	0.00%
予想	15	35	42.86%
注目	0	34	0.00%
指摘	0	31	0.00%
利用	1	30	3.33%
使用	0	28	0.00%
決定	0	25	0.00%
評価	0	23	0.00%
構成	0	22	0.00%
採択	0	22	0.00%
紹介	0	20	0.00%
導入	0	20	0.00%

そこで、思考系動詞は「感ずる[8]」「思う」「見る」「考える」、発言系動詞は「言う」「為る」を代表動詞として、今後の作業で揺れが生じないよう典型例を載せた表８を作成した。

[8] 「感ずる」は、上の表に載っていないが 22 位である。（揺れの個数：17 件　全体の個数 33 件　揺れ/全体：51.52%）

表８　思考系動詞・発言系動詞の典型例

レル・ラレルの意味	表現全体の特徴	レル・ラレルの直前要素				コメント
		感ずる	思う	見る	考える	
尊敬	－	〈PM41_00070〉 このときの一件は、皇后も不快に**感じられ**ていたようです。	〈OC06_02966〉 販売代理店に勤める友人のつてで車、新古試乗車を買う予定です。どのくらい値引きが可能だと**おもわれ**ますか？	〈PN2d_00022〉 新規投資はもう少し様子を**見られ**たほうが良いと思われます。	〈OC08_01893〉 運用を**考えられ**ないなら、貯金はどこも目くそ鼻くそですね。定期といえども、です。	《感ずる》《見る》 ・少数。 《考える》 ・本例のみ。おかしな表現。
自発	－	〈PB39_00023〉 どこかこの世ではない場所、海ではなくて空に近い場所で横たわっているみたいに**感じられ**た。波が全身をくすぐるように揺れ、本当に、気を失いそうなくらい気持ちがよかった。	〈PN1f_00012〉 昨秋の審査に寄せられた候補者たちの姿は頼もしく**思われ**た。			
可能	－	〈OW6X_00077〉 宿泊と食事の分離を行い，合理的で選択の自由を**感じられる**サービスの提供に取り組む動きがでてきている。				《感ずる》 ・基本的には、動作主が他者や一般論である。 ・発話者の感情の表現であるのに、自発より可能を強く感じる例がある。 〈PM12_00020〉 もちろん警察は、不幸な事故と判断した。だが私には、お前の殺意が感じられた。
可能	存在確認			〈OW6X_00032〉 また、業種によって回復時期に差が**みられる**点には留意が必要である。		・本来の知覚の意味が薄れているため、受身とは考えづらい。 ・動作主が意図していないことが重要な自発とも考えづらい。 ・“存在の知覚⇒確認⇒提示”という手順を踏んでおり、ある程度事態は完了している。動作主の意図の有無は不明だが、結果的成就を正面に出して表現する「意図成就」に近い表現と考えられる。 ・対象が具象名詞であったり、アスペクトが完了の意味であったりすると、本来の知覚の意味が強くなる。
可能	客観化		①〈OY15_07462〉 帰宅した相方。おそらくどこか（路上？）で寝ていたものと**思われる**。			・末尾はル形。 ・動作主は発話者。 ・細分化すると以下の4つ。 ①発話者の個人的意見。 専門家や世論の見方ではないが、客観的な内容であるというニュアンスを持たせようとする意図を感じる。 ②発話者自身が専門家。 思考を重ねた結果ある程度事態は完了している。結果的成就を正面に出して表現する「意図成就」に近い表現と考える。 ③専門家の見方を加味している。 ④発話者の意見だけでなく一般論でもある。 《感ずる》 ②感情の動詞でも客観化がある。 〈PB35_00013〉 地盤が悪くては意味がないからです。意外とこの段階で手抜きをしている業者が多いように感じられます。 《見る》 ②少数。本例も二重否定で特殊な例。 ③発話者自身が知覚行為をしていないことがほとんどのため違和感あり。 〈PN2e_00004〉：情報元の丸括弧による明示。 最も保守的で「ブッシュ氏の本来の信条に一番近い存在」（クリントン前大統領の報道官だったロックハート氏）とみられる。 〈PN1d_00004〉：発話者の調べではない。 調べでは、ネズミ類やイタチなどの小動物のミイラとみられる。
		②〈PB11_00021〉 目に見えない世界を扱う霊能者を低く評価する傾向があるようにも**感じられる**のですが、	②〈PN2d_00022〉 新規投資はもう少し様子を見られたほうが良いと**思われ**ます。	②〈PB48_00016〉 女たちの用語には何にでもオが付くから、これもその例の一つと**見られ**ぬことはないが、	②〈OW6X_00003〉 多様な需要に応えられる環境を土地市場にもたらすとともに、新たな土地需要を喚起していると**考えられる**	
				③〈PN3d_00019〉 死体遺棄現場の山林付近に残されたタイヤの跡は、比較的新しいタイヤでできた跡と**みられる**。		
			④〈OW6X_00234〉 一般的にインターネット利用等の機会に接しやすいと**思われる**勤務者や学生	④〈PM41_00071〉 月刊誌『ブブカ』に、ベッドの上で彼女と裸で抱き合っている写真や、明らかにタバコを吸っていると**みられる**写真が掲載されてしまったのです	④〈PN4b_00007〉 うかつに個人情報を送ると、ある日突然、ネットショッピングによる高額の請求が舞い込むといったことが**考えられる**。	
受身	客観化			〈PB39_00010〉 三村氏はためらう夫人を説得して、早々に結婚してあげたのだと**見られ**ている。	〈PM51_00213〉 例えば米カリフォルニア沖約三十キロにあるサンタ・バーバラ諸島は、米大陸と陸続きになったことはないと**考えられ**ている	・アスペクトは継続の意味。テイル・テキタなど。 ・動作主は他者もしくは一般的に。 ・一般論に多い表現。 ・思考は既に完了し、その思考内容は否定されずに継続している。
受身	－		①〈PN4a_00017〉 概念がＤＮＡという化学高分子で説明されたことから、「遺伝子決定論」の勝利は確定的と**思われ**た。ネズミの子がネズミになるのは遺伝子の配列を読めばわかる、と考えられたからだ。	①〈PN1d_00001〉 収支見込みも、当初は数年間の赤字が必至と**みられ**ていたが、	①〈PN4a_00017〉 ネズミの子がネズミになるのは遺伝子の配列を読めばわかる、と**考えられ**たからだ	・単純に事実。 ・細分化すると以下の2つ。 ①完了したイベント。 アスペクトは完了。動作主は他者もしくは一般的に。 ②典型的な受身。 他者から影響を被る。
			②〈OC08_02391〉 サインしてしまった以上、認めたと**思われ**てもしかたありません。	②〈OC14_00041〉 他人に**見られ**たら、嫌だなと言う品は「評価不要」で御願いしましょう。。。		《考える》 ・少数。

レル・ラレルの意味	表現全体の特徴	レル・ラレルの直前要素		コメント
		言う	為る	
尊敬	－	〈PM41_00070〉 美智子皇后は、『皇室の伝統は、祈りと継承です』と常々**言われ**ています。	〈0Y14_16795〉 みなさんの中で「ギクッ（◉∀◉　）」と**され**た方もいるでしょう。	
自発	－			
可能	－	〈PB48_00016〉 ゴサまたはゴサンという語を用いつづけていたという証拠は、絶対に挙げられないとは**言われ**ぬが		《言う》 ・少数。 ・発話者の発言。
可能	存在確認			
受身	客観化	①〈PN2e_00008〉 「生活館」は一時閉鎖を余儀なくされ、「億単位の損失」（関係者）を出したとも**いわれる**	①〈PN2e_00008〉 気を吐くホームセンター業界の背景には「ガーデニングとペットブームがある」（業界関係者）と**される**	・アスペクトに関係なく ・動作主は他者もしくは一般的に。 ・細分化すると以下の3つ。 ①政治ネタなど出所が怪しい情報を客観的な表現にしている。情報の発信者の丸括弧による明示が多い。 ②発話者の意見でもあるが、断言を避けている。 ③一般論。
		②〈PM41_00071〉 中村は、日本代表には欠かせない存在ですが、彼の活躍に危機感を抱いているのが中田だと**言われ**ています。	②〈PN1d_00010〉 約三億七千三百万円だけを申告、法人税約八千九百万円を脱税したと**される**	
		③〈PN3b_00016〉 二十三日は二十四節気の一つ「大暑」。１年で最も暑さが厳しい日と**いわれる**が、東日本の各地で涼しい朝を迎えた。	③〈PM41_00026〉 その間に行われた著作権侵害の被害額は、一説では数千億円にも上ると**され**ている	
受身	－		①〈PB35_00013〉 塗料には植物百％のＡＵＲＯ社のものが標準と**され**ているのです	・単純に事実。 ・細分化すると以下の3つ。 ①法律など決定されていること。 ②完了したイベント。 ・アスペクトは完了。 ③典型的な受身。 ・他者から影響を被る。
		②〈PM41_00093〉 平安時代は藤原氏の権勢と共に栄え、国家に異変がある時にはこの山が鳴動すると**いわれ**、御破裂山と恐れられるようにもなった	②〈PB22_00002〉 慶喜は賢明のほまれ高く、すぐにでも将軍が勤まると**され**た	
		③〈0C03_00375〉 今から４年ほど前に母親に「ローン組めないようにしたから」と**言われ**ました	③〈0C04_01067〉 貴方が知り合いにお金などを貸してそ知らぬ顔を**され**たらどうでしょうか？	

6.3. 結論

本作業の結果、レジスターによってレル・ラレルの意味の分布が異なり、その違いは対人配慮の有無が関係していることが明らかになった。また、意味や属性の違いにはレル・ラレルの直前要素（主に動詞）が深く関わっていることも判明した。

7. 今後の展開

いわゆる「迷惑の受身」についての分析を行う予定である。

また、本稿では言及しなかったが、機械学習もすでに試行しているので、別の機会に報告したい。最終的には BCCWJ 全体 784,935 件（コアデータ 7,906 件を含む）を対象に自動アノテーションを施したいと考えている。

参考文献

D. Biber (2009) *Register, Genre, and Style.* Cambridge Univ. Press.

庵功雄 (2001)『新しい日本語学入門 ことばのしくみを考える』

尾上圭介 (1998a)「文法を考える 5 出来文(1)」『日本語学』17 巻 7 号

尾上圭介 (1998b)「文法を考える 6 出来文(2)」『日本語学』17 巻 10 号

尾上圭介 (1999)「文法を考える 7 出来文(3)」『日本語学』18 巻 1 号

川村大 (2004)「受身・自発・可能・尊敬―動詞ラレル形の世界―」『朝倉日本語講座 6 文法Ⅱ』

志波彩子 (2009)「認識動詞の非情主語受身文―「見られる」「思われる」「言われる」「呼ばれる」を中心に―」東京外国語大学『日本研究教育年報 13』

動詞語義及び意味役割付与作業システムの構築

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 an 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て，さらに，動詞と係り関係にある単語との意味的関係を付与する述語項構造付与システムの構築を行っている．動詞語義とは文章内の動詞の意味を同定したものであり，同じ表層の動詞でも複数の意味を持つ場合が多く，既存の辞書では対応できない語義も存在する．また意味役割とは文章内の各項が文章中でどのような意味を持つかを同定したものであり，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および意味役割付与作業システムは BCCWJ の CORE の文章に対し，動詞語義およ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 上でのアプリケーションとして CakePHP[2] を用いて構築を行った．

本稿では本システムの構造や使い方を解説し，実際に 800 文程度に対して動詞語義付与を行い，作業者の行動から必要な機能の割り出しを行った．

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．

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 89 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．

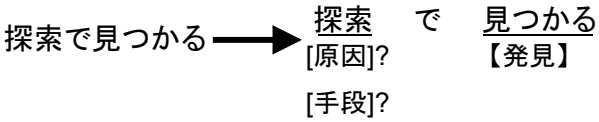


図 1　人によって判断が揺れる意味役割の例

図 1 は「探索で見つかる」という文の動詞語義及び意味役割付与例である．「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．Web 上でデータ入力，編集などができるデータベース管理ツールとしてはファイルメーカーなどがあるが，利用形態が制限されており，言語処理ツールを利用するなどといった拡張もできない．そこで，フリーでダウンロードできるデータベース管理ツールとフレームワークを用いて開発を行う．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．また Web 上でシステムを利用する場合，インターネットブラウザから作業を行うことができるためシステムのダウンロードを行う必要がない．ここで複数人で複数文に対して付与を行う場合，次のような問題が発生する．

(1) 対象となる文の数が膨大なため，付与される文が統一性をもたない．
(2) 同じ文に対して複数の付与が行われる．

まず (1) を解決するために作業対象を指定する必要がある．つまりデータベース上に存在する文すべてに作業対象か否かの状態を持たせるということである．次に (2) であるが，本システムでは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，(2) の問題を解決し，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 のなかから動詞のみを抽出し，その頻度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．

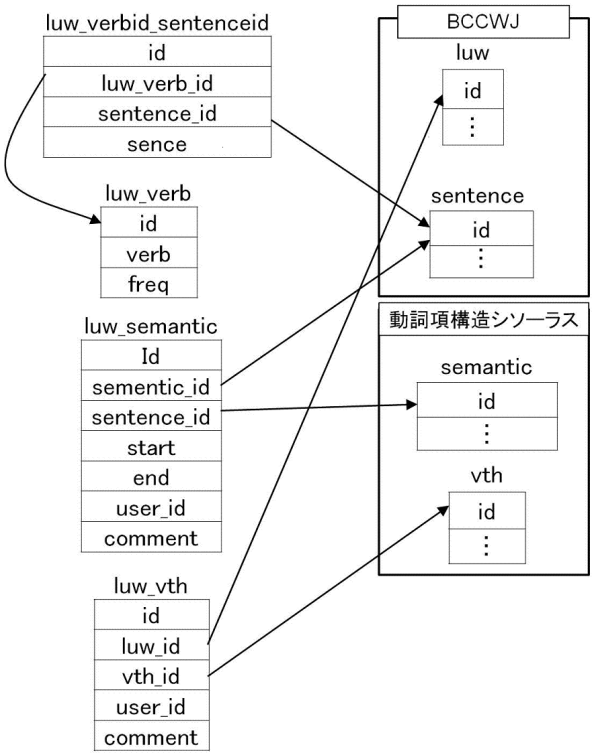


図 2　データベース構造

図 2 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．

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．文に動詞語義及び意味役割を付与する場合，はじめに動詞語義を決定する必要がある．動詞語義は動詞項構造シソーラスのデータを利用するため，動詞項構造シソーラスに語義の存在しない動詞は付与できない．また各動詞語義の例文が複数必要であるが，人が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，人が文を見て付与対象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．このため付与作業の手順としては以下の 5 つがあげられる．

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．それぞれの作業において，最適な操作環境とデータを表示する必要があるため，それぞれにおいて，作業用 Web ページを作成した

以下にそれぞれについて説明する．

A) 作業文の選択

最初の作業として，付与作業を行いたい動詞を選択する．動詞を選択すると選択された動詞を含んだ文の一覧を表示する．その中から作業対象としたい文を選択する．選択された文は作業対象となり，付与作業が可能となる．

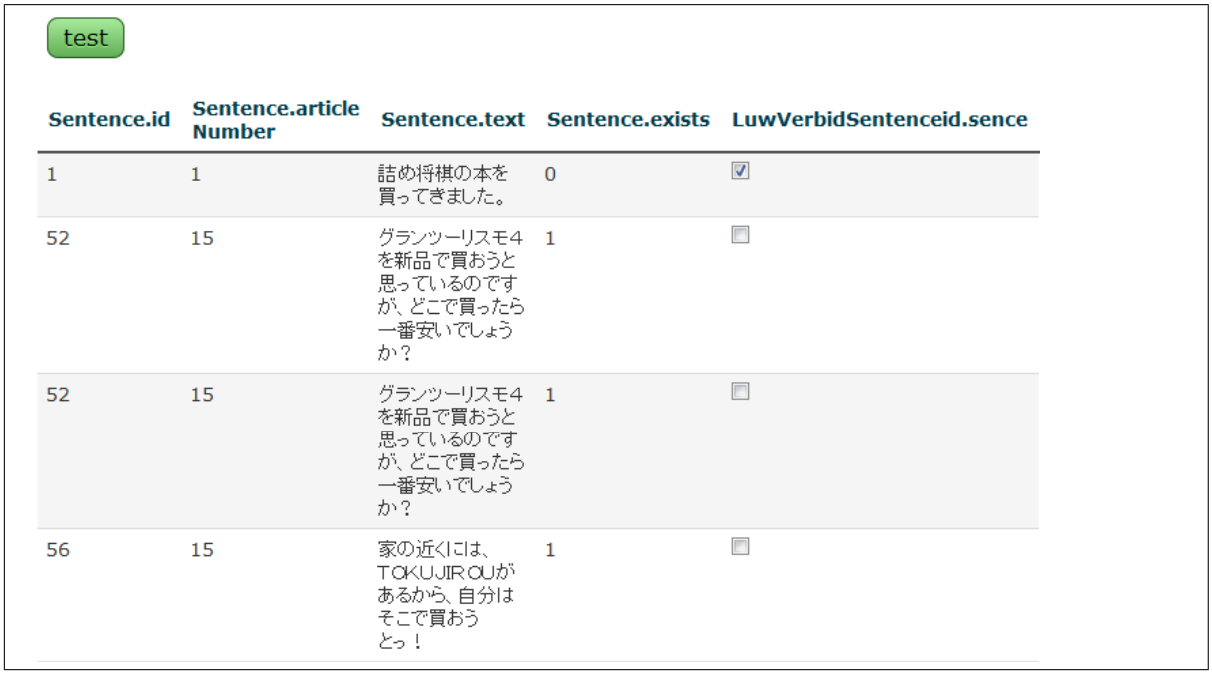


図 3　作業文の選択例

図 3 は「買う」という動詞の作業文選択画面である．「買う」という動詞を含んだ文の一覧を表示する．各文にはチェックボックスを設置しており，作業対象とする文を作業者が選択できるようにしている．チェックボックスにチェックを入れることで語義付与対象となる．またチェックを外すことで付与対象から外すこともできる．

B) 動詞語義付与

作業対象が存在する場合，語義の付与を行うことができる．付与作業を行いたい文を選択すると，短単位または長単位に分割された文を表示する．分割された文章の動詞を選択すると，選択した動詞の語義一覧を表示する．表示した語義から選択を行うことで語義の付与を行う．

図 4 は「詰将棋の本を買ってきました。」という文の「買う」に対する動詞語義の付与画面であ

図 4　動詞語義付与例

る．各語義にラジオボタンを設置しており，各語義のラジオボタンを選択することで対象となる動詞の語義を選択できる．語義一覧表示は動詞項構造シソーラスの分類のみを表示しているが，各語義の id を選択することで例文などの語義詳細を閲覧できる．

C) 動詞語義の決定

複数人の作業により動詞語義が付与された後，作業リーダーは動詞語義の決定を行う．よってこの作業では，複数人による作業の付与結果を表示し，どの付与結果を決定するかを選択するページを構築した．付与された動詞語義の一覧から意味役割の決定を行う．

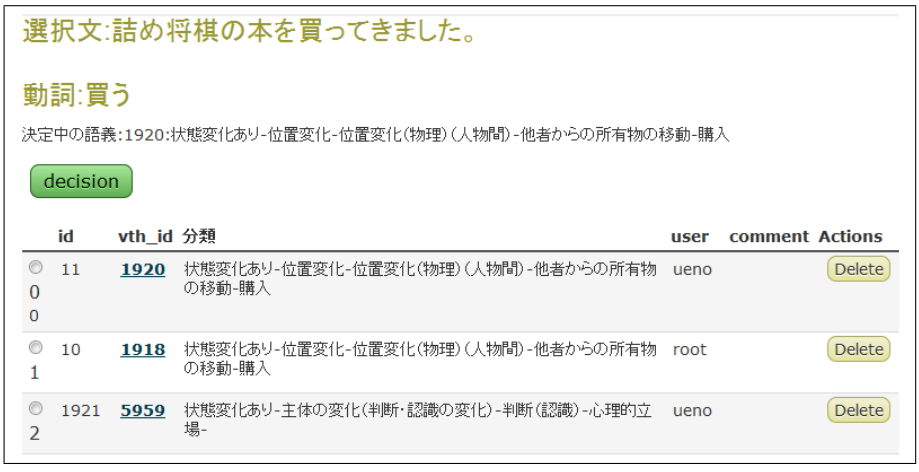


図 5　動詞語義決定の例

図 5 では「詰め将棋の本を買ってきました」という文の「買う」という動詞に対して，3 つの語義が付与されている．各語義の左側にラジオボタンを設置している．決定したい語義のラジオボタンにチェックを入れることで動詞語義を決定することができる．

D) 意味役割付与

作業対象である文の動詞語義が決定されている場合，意味役割の付与を行うことができる．付与対象となる項と意味役割の種類を選択することで意味役割の付与を行う．また付与した意味役割にコメントをつけることができる．

図 6　意味役割付与例

図 6 は「詰将棋の本を買ってきました。」という文の「買う」に対する意味役割の付与画面である．各項にチェックボックス，各意味役割にラジオボタンをそれぞれ設置している．まず対象となる項のチェックボックスにチェックを入れる．そして，意味役割一覧から付与したい意味役割のラ

ジオボタンにチェックを入れることで意味役割を付与できる．図 6 では「詰め将棋の本」に対して［対象］を付与している．

E) 意味役割の決定

複数人の作業により意味役割が付与された後，作業リーダーは意味役割の決定を行う．よってこの作業では，複数人での作業の付与結果を表示し，どの付与結果を決定するかを選択するページを構築した．付与された意味役割の一覧から意味役割の決定を行う．

id	sentence_id	luw_id	text	semantic	comment	author	Actions
3	詰め将棋の本を買ってきました。	買う	詰め将棋の本	対象		ueno	決定中 決定解除
18	詰め将棋の本を買ってきました。	買う	詰め将棋の本	結果		root	決定 Delete
19	詰め将棋の本を買ってきました。	買う	詰め将棋の	対象		root	決定 Delete

図 7　意味役割決定の例

図 7 では「詰め将棋の本を買ってきました」という文の「買う」という動詞に対して，「詰め将棋の本」に［対象］，「詰め将棋の本」に［結果］，「詰め将棋の」に［対象］という 3 つの意味役割が付与されている．決定を選択することにより，意味役割を決定状態にすることができる．また決定解除で決定状態を解除できる．決定状態にある意味役割は消去などの操作を行うことができない．図 7 では「詰め将棋の本」に対する意味役割として［対象］が決定状態にあることを示している．

4.　付与作業のユーザーインターフェースの改善

実際に本システムを使って 3 人の作業者で動詞語義付与を 800 文程度行った所以下のような意見が寄せられた．

1. 自分の作業が終わっているかどうかを簡単に確認したい．
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. 作業例文を選ぶ際に付与が終わっている文には「OK」を表示．
2. 同じ作業を次の文で行う「次の文へ」のリンクを作成．
3. 動詞語義選択の際に適合語義なしの選択肢とコメント欄を追加．

これらの改善により作業負担が減るとともに，適合語義なしの選択肢を追加したため，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．しかし今後の課題として動詞項構造シソーラスの拡張という観点から考慮した場合，適合語義なしという分類では不十分であると考えられる．改善法としては「動詞語義の追加」という動作を実装することである．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．

5. 考察

作成したシステムの動作速度についての考察を行う．本システムはサーバクライアント方式で作成しており，レンタルサーバで独立した CPU を優先して利用しているため，膨大な量のデータを利用するにもかかわらず，各作業は操作から 1，2 秒程度で動作する．しかし，付与データが増えると動作速度は低下すると考えられる．動作速度を下げる要因として，データの取得にかかる時間が長くなることがあげられるためである．さらに本システムは一つの動作に複数のテーブルを用いている動作が多い．取得するデータ数が同じであってもテーブル数が増えると動作が遅くなる．これはフレームワークの仕様であるが，データを取得する場合にテーブル選択，データ検索，データ取得という手順で行っているためである．またフレームワークの機能として，結合させたテーブルのデータを自動で取得するというものがあるが，本システムのように扱うデータが多い場合，必要なデータ以外のデータも多く取得してしまうため，動作が非常に遅くなる．このため本システムではこの自動取得の機能をなるべく用いず，必要データのみを取得している．意味役割付与動作では，フレームワークの機能で自動取得した場合に比べ，4 分の 1 程度の時間でデータを取得できた．

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．また実際にシステム利用者の意見を取り入れることでユーザーインターフェースの改善を行った．

参考文献

[1] 動詞項構造シソーラス．http://cl.it.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. 619–622, 2011.

[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. 143–146, 2010.

[8] MySQL．http://www-jp.mysql.com/.

『太陽コーパス』にみる、動詞性名詞「報告」の使用実態

佐藤 佑（東京外国語大学）

Verbal Noun "*Houkoku*（報告）" in Modern Japanese

—A Research Based on the Data of 'Taiyo Corpus'—

SATO Yu (Tokyo University of Foreign Studies)

１．はじめに

現代日本語のサ変動詞（行動する、出入りする、プレーする……）は、その大多数が語幹の名詞用法（（太郎の）行動、（店への）出入り、（一流選手の）プレー……）を持ち、それは第一義的には「動詞の表す事態をすること」（行為）の意味を表す。

発表者は、佐藤（2011b）において、サ変動詞語幹および「(身体の）動き」「(靴下の）繕い」などの動詞連用形が事態を表す諸例（両者および一部の関連形式[1]を「動詞性名詞」と総称する）の構文・意味的特徴を詳述した。その骨子は以下のようなものである。

> 動詞性名詞は、「太郎が近所の公園で遊ぶの（を見た）」「花子が来年の夏に結婚すること（を知った）」のような節単位の名詞化形式（名詞化節）に比べ、名詞句内部には限られた情報だけを表し、広範な文脈の流れを反映する「凝縮」的な表現である[2]。

名詞化節と異なり、動詞性名詞を核とする名詞句は、それ自体に多くの要素を明示することが難しい。たとえば、「??太郎の近所の公園での遊びを見た」「??花子の来年の夏の結婚を知った」というのは、いずれも著しく不自然である。これには様々な要因が関わっているが、一つには名詞句の性質上、複雑な連体修飾が難しくなること、また当該形式は文脈に強く依存し、それ自身において多くの情報を提示する必要もないということが指摘できる。

少なくとも文学作品などにみる限り、動詞性名詞が連体修飾を受ける場合、修飾要素は単一であることが圧倒的に多い（佐藤（2011b）の調査範囲中では、「いつもの平穏な生活」のように複数表れる例は約１割に過ぎない）。そして、「太郎の家出」「ビルの破壊」といった当該の動作・作用にとっての主体および直接対象を表す例をはじめ、事態の生起する時間（夜中の電話）、場所（故郷での暮らし）、原因（長旅の疲れ）、動作・作用の行われ方＝様態（静かな会話）など、様々な副次的要素が表れうる。

こうした動詞性名詞句の中でも、発話行為や授受行為を中心とする、X【人１（動作の主体）】・Y【人２（動作の向かう相手）】・Z【やりとりされるもの・ことがら（対象・内容）】の三者間の関係については、連体修飾の構造についてきわめて重要な特徴が認められる。その典型的な例の一つに、「報告」が挙げられる。

サ変動詞「報告する」は、伝達の主体 X・相手 Y・内容 Z の三者の関係を、待遇性・話題の偏りといった問題にはあまり左右されず、比較的シンプルかつ客観的に表すものと考えられる。そして、その語幹の名詞用法である「報告」は以下に挙げるように、動詞性名詞としても特筆すべき特徴を備えていると考えられる。

---

[1] 一例として、「引っ越し」（「引っ越す」とも「引っ越しする」とも対応する）など。

[2] この「凝縮」という概念は、石井（2007）において指摘された臨時一語の性質に関する考え方を、動詞性名詞の使用全般に拡張したものである。

①連体修飾構造の明確性

「X が Y に Z と報告する」との対応でいうと、「X の／からの」「Y への」「Z という／との」といった連体修飾で、参与者間の関係を適切に表し分けられる。また、特に（発話を表す動詞性名詞の特徴として）Z については「Z の／という／との……」というように、実現のパターンが多様であり、それらが各々どう使い分けられるかも問題になる。

②複合語・臨時一語の生産性

（下の③ともかかわって）「調査報告」「山田氏報告」などの複合語・臨時一語（臨時に形成される複合語）のバリエーションがきわめて豊富である。

③意味の多側面性[3]

発話行為そのものだけでなく、発話の内容・記録などにも言及しうる（cf. 彼の報告を疑う/雑誌で読む）。また、その境界は明確でないことが少なくない（cf. 敵将戦死の報告が部隊を奮い立たせた）。

本発表では、現代日本語における動詞性名詞「報告」の意味・用法の実態を明らかにし、またそのメカニズムを詳細に検討するための足がかりとして、成立期現代語における名詞「報告」の統語的・意味的特徴を検討する。具体的には、明治末～戦前までの日本語の使用実態をもっとも適切かつ容易に調査することのできる『太陽コーパス』を用い、主として上記①－③の問題を検討する。

２．データの収集

検索ツール「ひまわり」（ver. 1.3 β 05）を利用し、検索文字列を「報告」として全範囲の検索を行った[4]。なお、ルビ検索による異表記の検討も行ったが[5]、現代語と同じ「報告」以外の表記で、同様の意味の語が表れた例はないものと判断された。

検索結果を Excel にペーストした後、手作業で名詞「報告」と動詞「報告する（致す）」とを別のシートに振り分け、さらに前者は年次ごとにシートを分けた。また、行為との移行関係を中心に考察するため、「報告書」「報告者」といった、「報告」を前項とし、かつ直接行為に言及する余地のない合成語（103 例）については主たる考察対象からは除外するが、必要に応じて参照できるよう、やはり別のシートを作って保存した。

３．『太陽コーパス』における「報告」の分析

３．１．名詞「報告」のデータ概要

名詞「報告」およびそれを中心とする名詞句、あるいはそれを後項とする合成語の用例は、計 518 例が得られた（1895 年：132 例、1901 年：112 例、1909 年：104 例、1917 年：62 例、1925 年：108 例）。なお、動詞「報告する」の用例は計 251 例（60 例・55 例・50 例・39 例・47 例）と、全体としては名詞「報告」の半数未満にとどまった。

名詞「報告」が直接連体修飾を受けたもの（「日本銀行の報告」「麻に關する報告」など）は計 242 例、「報告」を後項とする複合語・派生語・臨時一語は計 136 例（「近年の決算報告」など、さらに別途連体修飾を受ける例も含む）、いずれも認められず「報告」が単体で用いられた例は計 140 例であった[6]。

---

[3] 本研究において、こうした問題について考える上で、西尾（1961）における連用形名詞の分類（特に「動作・作用そのもの」＞「イ 動作・作用そのもの（何々スルコト：泳ぎ、調べ、貸出しなど）」「ロ 動作・作用の内容（何々スルトコロノコトガラ：考え、教え、望みなど）」「ハ 動作・作用のありさま・方法・程度・具合・感じなど（金遣い（が荒い）、滑り（がいい）など）」）がきわめて重要な視座を与えている。

[4] なお、修飾要素や文中での位置などを見やすくするため、前文脈・後文脈は各 100 文字とし、さらに広範な文脈を見る必要がある場合は別途本文を参照するなどした。

[5] ルビ検索による異表記の検討方法について、詳しくは佐藤（2011a）などを参照されたい。

[6] カウントは可能な限り綿密に行うことを期したが、「毎月報告を求める」を単体扱いとした判断など、微

なお、コーパスの収録対象となっている雑誌『太陽』の特性に由来するものでもあるが、名詞「報告」が用いられる文章は、おおむね時代が古いほど文語体が多く、新しくなるにつれ口語体が主流となっている。一記事中に文語・口語が混在する等の問題（主に小説）もあり、必ずしもコーパスの「文体」タグが当該の用例の文体を正確に反映しているとは言えない場合もあるが、「報告」の用例に関してはおおむね実態を反映しているものと判断された。文体ごとの分布は、下の表 1 に示すとおりである。

表 1 文語・口語の分布（年次別）

	1895	1901	1909	1917	1925
文語	130 (98.5%)	102 (91%)	58 (55.7%)	26 (41.9%)	3 (2.8%)
口語	2 (1.5%)	10 (9%)	46 (44.3%)	36 (58.1%)	105 (97.2%)
計	132	112	104	62	108

以下、名詞「報告」が受ける連体修飾の内実（３．２.）、「報告」の語構成力（３．３.）、「報告する」行為と名詞「報告」の関係の多様性（３．４.）の順に概観する。

３．２.「報告」の連体修飾

まず、名詞「報告」がどのような連体修飾を受けているかの実情を概観する。なお、「(近年の) 決算報告」「(各般の) 事務報告」など複合語・派生語・臨時一語がさらに連体修飾を受ける諸例については、主に語構成のあり方を概観する関係上、次項で簡単に触れるにとどめる。

A. 伝達行為の主体

「報告する」主体（「X が Y に Z と報告する」の X）が表れる例（84 例）は、(1)(2)をはじめ、現代語と同様に「X の」「X からの」の形で実現するものが多い。

(1) 然れども此の流通資本の増加を以て、直ちに通貨の増加と誤認すべからず、日本銀行の報告によれば、近來我國の通貨は、増加せずして寧ろ減少の傾向あり、(1895 年 11 号／「商業」) [7]

(2) 紐育の火星委員會の會長は、去年の接近期間中、聽取記録を作り得る素人研究者達からの報告を蒐集した。(1925 年 4 号／「世界のラヂオ」)

ただし、文語体の文章には、(3)のような「X よりの報告」の例も見られる。こうした後置詞の使用は、現代語の感覚では一般的とは言いがたい。

(3) 在安平有地艦隊司令長官よりの報告によれば二十日午後二時安平沖警戒中英獨兩國の軍艦及領事館員より劉永福は十九日の夜若干の部下を以て三艘の支那舩に乘り遁走し……(1895 年 11 号／「海内彙報」)

B. 伝達行為の相手

「報告する」相手（「X が Y に Z と報告する」の Y）は、現代語と同じく「Y への」の例が見られるが、以下の(4)を含め計 3 例ときわめて少ない。

妙な部分も多少残っている。さらに客観的な条件付けの基準設定は今後の課題としたい。

[7] 用例は、後ろに括弧書きで年次と号・記事タイトル・著者名を示す。ただし、著者が不明の記事については記事タイトルまでの表示とする。

(4) マルシヤン「陛下。長い間皮を集めて漸やく作つたのでございます」／奈（マルシヤンの手を握り）「忝い。私はお前に取らすものがない、此の握手で滿足してくれ、おい監督殿政府への報告の種が出來たぞ」（1917 年 12 号／「脚本　落日（帝国劇場台本）」／佐藤紅緑）

現代語ではこの他、「Y に対する」「Y に向けた」などの形が考えられるが、該当する例は見られなかった。以下の(5)は唯一得られた「～に対する報告」の用例であるが、現代語の感覚としては「～に関する」と同じような意味（→３．２．C）になると思われる。

(5) 試みに現時の我領事の通商貿易に對する報告の如き之を歐米各國領事の報告に比較し見よ如何に彼れの報告の其の商工業に密接して實際的に、我れの報告の商工業に迂遠にして非實際的なるかよ。（1909 年 14 号／「行政税制整理問題　外務行政刷新上の四要望」／望月小太郎（談））

C. 伝達の内容

「報告する」内容（「X が Y に Z と報告する」の Z）を表す連体修飾要素としては、「Z との」「Z といふ」「Z の」といった形の引用節が見られる（(6)-(8)）他、(9)(10)のように話題の中心となる事物が後置詞を伴って表れる例も確認された（計 67 例）。いずれも現代語と共通して見られるタイプの表現である。

(6) 一月計り經て、宮御方辛じて會津へ入せらるゝのよし、覺王院も御傍にあるとの報告に、少しく愁眉を慰めたり、（1895 年 6 号／「彰義隊　下」／曳尾叟）
(7) 翌八日御幣使街道に斥候して見ると、太田の宿に敵が居ると云ふ報告ジヤから、夜中進んで見ると、未だ太田迄來ない、（1901 年 10 号／「追懐談」／川村純義（談））
(8) 工藤行幹氏より第二號議案黨則改正の説明あり、神鞭知常氏財政調査延期の報告あり、（1901 年 1 号／「海内彙報」）
(9) 蠶業の改進夫れ何れの時をか期せん。麻に關する報告は同く、 品質良好なるもの多しと雖、 間々栽培製造の方法良しからざる爲め、 色澤不良、 繊維の長短強弱其完きを得ざるものあるは遺憾なりと。（1901 年 13 号／「農業世界」／上野英三郎）
(10) 此等穿鑿のなされたる精神の、自然の結果として、此等の宗教に就ての報告は、積極的に其何を有するやに非ずして、寧ろ消極的なる其なき所の何たるをいふにすぎざりしなり。（1895 年 11 号／「ショツペンハウァー氏の支那宗教論」／Ｓ．Ｔ．生）

一方、(11)のような外の関係の修飾で、伝達内容が表される（現代語では考えにくい）表現の例も若干数見られた（ただし、「報告」を修飾する連体修飾節は、「各省から出る報告」「其向へ爲したる報告」のような内の関係のものが大半を占める）。

(11)その後數年經つて、獨逸のフライブルグ大學のデウラ・カンプ氏並びにその門弟キユウプルレ氏が動物試驗を試みて、結核の治癒状態、人體に效能ある報告に接した。時に千九百十三年であつた。（1925 年 1 号／「最近Ｘ光線療法の進歩」／宮原立太郎）

D. その他

必須項のみならず、様々な副次的要素を伴いうることも、動詞性名詞の特徴である。『太陽コーパス』上の「報告」においても、上で扱った主体（X）・相手（Y）・内容（Z）の 3 種の他にも、多様な修飾要素を伴って表れている。

(12)(13)のように報告の内容などを評価・規定する例[8]が 25 例程度見られた他、(14)(15)のように「報告」が行われた状況（時間や場所など）を表す例などは、他に 17 例（時間的

[8] その他、「數多の」など報告行為の多寡を表す例も若干数見られた。

なもの 9 例・場所的なもの 8 例）と、全体的にあまり多いとは言えない[9]。

(12) 倫敦商業會議所の劃策經營する所は實に他に比類を見ざる所にして、或は諸種の統計及び報告を蒐集頒布し、或は公會を催し或は學者及び專門家に托して講演會を設け、孜々外國貿易擴張上須要なる智識を研磨し、同時に商業上の爭論を仲裁し又政府の諮問に應じ或は政府に建議し駐外領事より緊要なる報告を徵し汎く之を頒布する等商業の發達上頗る有益の擧に出づること多し。（1901 年 3 号／「商業世界」／佐野善作；祖山鍾三）

(13) 聯合會の人々は勿論議定書に贊成であるから、内閣の政策に不服を有つてゐた。そこで報告者は、『フランスは平和に對してデクレーアし、イギリスは戰に向つてデクレーアした。』と面白い報告をした。（1925 年 12 号／「欧米雑感　国際心理の矛盾と煩悶」／塩沢昌貞）

(14) 前刻の報告中に成るべくは還暦の機會を避けたかりしことを縷々述べましたが、先生の教職に就かれましてより滿五十年になりまする其機會を採りたる方が宜しかりしならんことを最近に至り遲蒔きながら氣附きたる次第でございます。（1917 年 12 号／「菊池大麗先生」／藤沢利喜太郎）

(15) 此合同救助は克く其目的を達し、千八百九十三年三月十六日英蘭銀行の廣間に於ける報告に依れば、ベーリングの負債は一時三千三十一萬三千磅に達せしも今や僅に四百五十五萬八千八百十三磅あるのみにて其所有せし手形は毫も損失を生ぜず盡く取立濟と爲りたりとあり。（1901 年 8 号／「商業世界」／祖山鍾三；佐野善作）

なお、以下の(16)のように複数の項目が同時に表れる（この場合、内容=Z とそれに関する評価を表す形容詞の 2 項が表れている）例も見られるが、計 25 例と限られる。(17)に見られるように、現代語では複数の連体修飾要素によって（たとえば、「七年前の露國醫師大會（で）の報告」のように）実現しそうな例であっても、連用修飾の形で実現する例が多く目に止まった。

(16) 而して收支の明細なる報告は市民の前に提供せられざるべからず。（1909 年 2 号／「電車問題の根本的解決」／安倍磯雄）

(17) 今より七年前露國醫師大會の報告によれば明治三十八年より同四十二年に至る五年間に於て四萬五千以上の自殺者を生じ、明治四十一年同二年の統計によれば自殺者千人中年齢八歳以上四十歳の者四十五人、……（1917 年 5 号／「欧州大戦と露国の革命」／浮田和民）

３．３．「報告」の複合・派生・臨時一語形成

「報告」が何らかの名詞成分を前項とし、あるいは接頭辞を伴い、複合語・派生語・臨時一語を形成する例は 136 例に上った。そのうち、「報告する」行為の主体は、複合語・臨時一語として実現した例は以下の(18)の 3 例を含め計 25 例と比較的少ない。これは連体修飾の形で表れる例の約 1/3 が行為主体である（３．２．A）のとは大きく事情が異なる。

(18) 經常部臨時部共に委員長報告通りに決定し、次で乙號豫算に入り島田三郎氏より繼續費打切りの動議を提出せしも少數にて成立せず、全部委員長報告通り決定し、丙號豫算各特別會計に就ても田川氏より臺灣航路補助費削除の動議ありしも、總て委員長報告通りに決定せしも、唯だ高柳覺太郎氏は濱松鐵工所問題につき委員長の報告に滿足せず、（1909 年 4 号／「彙報」）

主体の例は「彙報」における「委員長報告」が半数近く（11 例）に上り、その他「領事報告」「審査總長報告」など、行為主体を表す前項は役職名が大半を占めた。特に固有名＋「報告」の複合語・臨時一語は特に少なく、以下の(19)(20)など 4 例が見られたのみであった。

[9] 他にも先に触れた「各省から出る報告」などの例が見られたが、紙幅の関係上詳細は割愛する。

こうした用法の自由度は、現代語に比べて高くないように見受けられる。

(19) ……殊に今日にては容易に獲得し難き東洋諸學會の數十年間の報告雜誌全部を網羅し、且稀覯書としては十六世紀版のマルコポロの紀行やモルガの費島史の原本初版の如き今日の時價一萬圓以上に登るもの數册、 或は最も珍奇なるエズイツト報告數十册を算する如き、此の如き大蒐集は到底今後不可能であらう。(1917 年 10 号／「案頭三尺」／内田魯庵)
(20) ◎震災豫防調査會報告第四號及第五號　同報告第四號は去る七月三十日發行せられたり、其目次は（一）委員臨時委員及囑託員（二）委員會……(1895 年 10 号／「科学」)

一方で、「決算報告」「現状報告」などといった、前項が「報告」の内容（cf. ３．２．C）を表す例は、計 85 例ときわめて多く見られた。また、(22)(23)のように複合語がさらに連体修飾も受けることで意味の具体化がより詳細になる例（計 53 例）は、大半がこうしたタイプに集中する。

(21) ◎駐支米國公使　ラインシユ氏は卅一日黎總統を訪問し歐洲平和會議提案報告及米支經濟連絡の件に關し長時間會談すと。(1917 年 3 号／「日誌」)
(22) 「ヘトール」の結核病竈に對する作用は如此、而して、ランデレル氏の治驗報告は如何と問ふに、千八百九十九年、獨逸伯林に於て、萬國結核撲滅會議を開けり、(1901 年 10 号／「肺結核の新療法（ランデレル氏「ヘトール」療法)」／ＸＹＺ生)
(23) そして今手元にある千通未滿の内、階級や境遇教養年齢等の點で割に雜駁なのを調査の都合上あと廻しにし、不取敢群として相當に數も纏まつた（一）東京帝大生百四十名（二）早稻田大學生百八十名（三）京都同志社大學豫科生二百十七名、合計五百十七名に就ての統計報告の極大要は次の通り。(1925 年 1 号／「愛慾世界の鳥瞰図　京都同志社大学、東京帝大、早稲田大学学生の性生活の統計的調査」／山本宣治)

その他、場所（(24)など 5 例）、時間（(25)など 2 例）、「報告」の内容などを評価する要素（(27)など 8 例）など３．２．D で見た諸例に類する複合語・臨時一語が見られた他、(26)の 1 例のみではあるが道具・手段が前項となる例も得られた。
なお、(25)などは現代語の感覚としては「前回」が「書いた」を修飾していると考えるのが自然であろうが、後述する派生語との関連上、一応複合であると考えておくことにする。

(24) ◎馮段協議　段總理は馮總統を訪ひ李開＝の陸榮廷會見報告を俟ちて約法國會臨時參議院に關する命令を同時に發する件に付協議すと。(1917 年 12 号／「日誌」)
(25) 報告は先づ友人デーンコート醫師の話から始まる。醫師は前回報告の末尾にちよつと書いた約束に從つて、その翌日再びラトランド家に往診したが、脈を取つてみるまでもなく、患者メーブル孃の病状は更に見直すところなく、益益惡化してゆく模様である。(1925 年 12 号／「長篇探偵小説　ハートの九『第六回』」／延原謙（訳)；ビ・エル・フアルジヤン（作))
(26) 然るにＺＲ３號は、毫も斯の如き難飛行をしなかつた。これは絶えず大陸との無線通信が行はれ、また海上幾多の場所に配置せられた船から、時々に天候の無線報告をうけて、天候の險惡なコースを避けるやうにして飛んだからである。1925 年 4 号／「ラヂオ漫談」／近藤生)
(27) 而して祕密會は午後四時十三分より初まり五時五十分に終り、再公開後祕密會經過の形式的報告ありて此日の議場を閉づ。(1917 年 3 号／「第三十八議会解散顛末」)

「報告」に接頭辞が上接する（派生語を形成する）例に目を向けると、計 9 例と少ない中にも（あるいは、その少なさにこそ）現代語との相違が如実に表れている。
重要な一点として、「ご（御）報告」の少なさが指摘できる。名詞「報告」に接頭辞「御」が上接した例は(28)を含めわずか 2 例、動詞「御報告致す」も 2 例（(29)含む）しかない。

(28) 大山第二軍司令官の報告　第一師團より左の電報ありたり御報告に及ぶ（1895 年 4 号／「軍事」）

(29) 又前後の事情より推斷しますれば啻に今日の記念會の間に合はざるのみならず何時出來するか殆んど無期限に延引するの懸念あると同時に男爵に於て記念品の贈呈を非常に迷惑に御思召さるゝの御意向が其後段々と益々明瞭になりましたから、委員の間の協議により斷然記念品を差上げることを見合はせることになりました。其事を御報告致しまするの同時に御賛成者御一同の事後承諾を茲に御願ひ致す次第でござります。（1917 年 12 号／「菊池大麗先生」／藤沢利喜太郎）

現代語において、目上の人物に対する「報告」行為は、名詞の場合であれ動詞の場合であれ（少なくとも当事者間における発話では）「御」を伴うことがほぼ必定であると考えられるが、『太陽コーパス』の実例においては、報告行為の主体と相手の上下関係などのあり方に関わらず、そうした例がほとんど見られない。「報告」1 語の使用だけを材料に判断するのは早計に過ぎるが、こうした待遇表現に関する現代語との異同、歴史的変遷といった問題も、動詞性名詞の研究全般において重要な意味を持つであろう。

また、以下の(30)のような接頭辞は、現代語では動詞性名詞の使用全般において実現しがたいと考えられる。1895 年・1901 年に各 1 例が見られるのみと実例は限られるが、他の動詞性名詞についても調査・検討し、より体系的に見ていく価値はあろう。

(30) 牛痘の製造に前報告の如く水牛を限り之を用う埃及牛は肩の隆起を具へざるも＝甲頗る強剛なり（1901 年 12 号／埃及の家畜／ドクトル、ヤンソン農業世界）

その他、以下の(31)(32)のような用法は現代語でも問題なく実現可能であろう。

(31) 同局長の言に曰く、 今回の損害は其區域頗る廣大にして、甚しき慘状を極めたる事は、豫め諸報告に依り粗ぼ想像を畫きて之れに臨みたるが、（1895 年 9 号／「海内彙報」）

(32) 挿圖は重もに昨明治廿七年六月東京附近地震の際の被害物を撮影或は描寫せるものを石版にしたるものにして、凡そ百三十餘枚なり、一見直に瓦飛び壁裂くるの當時を追憶せしむ、同報告の如き號を追ふに從ひ益々記事の精采を放つが如きは、斯學の爲に頗る喜ぶべき事なり。（1895 年 10 号／「科学」）

３．４．「報告」の多側面性

『太陽コーパス』における「報告」の、「報告する」事態との関係は、典型的には「行為（報告スルコト）」「内容（報告スルトコロノコトガラ）」「記録（報告サレタコトガラ）」の 3 パターンに分類される。

「行為」「内容」の諸例に関しては、現代語と大筋で変わらない用いられ方をしていると考えられる。一方、「記録」の意味については、現代語に比べて自由度が高いと見られる。

A. 「報告する」行為

「報告」の内容は問題とせず、事実として「報告」が実行されたか否かに言及する場合、名詞「報告」の動詞「報告する」との関係はもっとも直接的である。こうした意味は、機能動詞結合（cf. 村木 1991[10]）において特に明らかに表れるものである。

以下の(33)-(35)のように、現代語でも同様の表現が見られる例が多いが、(36)(37)など現代語では考えにくい機能動詞の使用も見られる。後者は特に文語の文章に集中する。

[10] 機能動詞とは、「実質的な意味を名詞にあずけて、みずからはもっぱら文法的な機能をはたす動詞」（村木 1991:203）。それらと名詞の組み合わせが、全体として動詞相当に働くもの、たとえば「さそいをかける」（≒さそう）「連絡をとる」（≒連絡する）といったものを機能動詞結合と呼ぶ。

(33) 米國シカゴ市に於いてパンコストが喉頭癌の放射線治療に關する報告を**した**中に、パリーのレガアト博士が三週間毎日持續してこの治療を始め、その經過の成績良好なることを證明してをるといふことを傳へてゐる。(1925 年 1 号「最近X光線療法の進歩」／宮原立太郎)

(34) 今や大要を悉して事務報告を**終れり**、 幸にして本會の基礎略定まり、 組織愈健全強固なるに當り、……(1909 年 5 号／「海外通信　在米日本人会報告」／牛島謹爾；久万俊泰)

(35) ……午後一時卅分再開、各地より祝電披露の後、細迫氏より委員會で決定したる議事順序の報告が**あつて**議事に入り、次の如き申合せを爲した。(1925 年 12 号「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―」／新田生)

(36) 而して三日に至り改革派は本部の名義にて新政黨組織の目的、犬養氏除名の理由を全國黨員に宣言せしに對し、六日非改革派は第二回の報告を**發する**と同時に、……(1909 年 5 号／「彙報」)

(37) 石坑鑛穴の穿掘は、其地の歴史に關し甚だ有益の報告を**與ふる**こと決して少からず。(1895 年 11 号／「地質学及び地質学者」／佐藤伝蔵)

B. 「報告する」内容・「報告した」内容の記録

名詞「報告」が「聞く」などの直接知覚行為の対象として表れることは、とりもなおさずそれが「報告されるところの内容」の意味に解釈されることと連動する。人の行為である「報告（すること)」を直接「聞く」ことはできないが、「報告（されるところのことがら)」を「聞く」ことは可能ということがその証左である[11]。

(38) 其後貴君には久留米の病院にて、御治療中との報告を**聞き**、奥様へ其報を幾度と無く申上て見ましたが、正氣を失ふ悲しさは、そりや嘘ぢや、僞説ぢやと計にて、益々嘆の積るばかり。(1895 年 8 号「夜の鶴（上)」／福地桜痴)

知覚・認識され、理解された「報告」の内容は、さらに種々の判断行為の対象として処理されることにもなる。

(39) ◎刑法及刑事訴訟法改正案　曩きに司法大臣より辯護士協會に諮問したる刑法及刑事訴訟法改正案は同協會の調査委員に於て遂に之れを否決したるが總會に於ては該委員の報告を**是認す**べき傾向ありといふ（1901 年 10 号／「海内彙報」

ただし、以下の(40)などは、直接知覚した「報告」の内容に関して「點檢する」と解釈していいか疑わしい。どちらかといえば、提出された報告書などを読み、その内容を検討していると考えた方が自然である。

(40) 斯の困難の場合に、人々互に警戒を爲したるを以て、爲めに良好なる經驗を得たるなり。其證據たる昨年の上半季に於ける、各銀行及び各會社の報告を**點檢する**に、東京に於ては十五銀行、第一銀行、三井銀行、三菱銀行、第三銀行、安田銀行等、 阪に於ては住友銀行、鴻池銀行、山口銀行、三十四銀行等、何れも皆日本銀行に對して借金を有し……(1901 年 9 号／「財政整理と民間の事業界」／小松崎吉雄)

---

11 こうした問題について、佐藤（2011b）では「行為」と「内容」の双方を「総括する」ものと主張したが、こうした構文において「内容」だけを「聞く」と考えることはできても「行為」だけを「聞く」と見ることはできない（仮にできるとすると「彼の報告は聞いたが何を言っているか聞き取れなかった」のような発話が許容されることになるが、実際はそうではない）ことから、やはり「内容」の方をより重視すべきであると思われる。

文書化されるなどして記録された「報告」の内容を表すことが明らかな名詞「報告」の例として、以下の(41)などが挙げられる。こうした例は佐藤（2011b）などでは「直接動作に関わらない」として排除されているが、動詞性名詞の周辺的な用法として注目し、そのあり方、広がり方を見ていく価値のあるものであると思われる。

(41) サトウ氏が蒐集した日本耶蘇會年報は兩三年前京都大學の藏に歸したが、 その中に西暦一五九三年三月より翌年同月に至る文祿二三年に亙る一年間の報告を收めた冊子が二部存する。（1917 年 1 号／「西洋画伝来の起源」／新村出）

上の (40) (41)は現代語でも違和感なく受け入れられると思われるが、下の(42)-(44)のように多様な動作の対象として結果物（「報告書」など）としての「報告」が表れることは難しいと考えられる。このように、『太陽』発刊時の日本語では現代語よりも「報告」単体で「結果物」の意味を表す能力が高かったであろうことが示唆される。ただし、こうした現代語で考えにくい例は 1895 年・1901 年に集中しており、1908 年以降にはほぼ見られない。

(42) 今日に當つて、敵國の我に對して此終局を如何すると云ふことに就いては、一旦媾和の使節も派遣致して本月上旬に於て廣島で兩回の面會を致しましてありますが、其意思甚だ曖昧にして未だ眞正の和を求めたるものと認めませぬに依つて、之を拒絶するの止むを得ざるに出たのであります、其大體の顛末は過日外務次官をして本院に其報告を提出致さして置きましたことに於て、明白と存じます、（1895 年 3 号／「海内彙報」）
(43) 亦之と同時に双方の會社共に自己の車輛にて運搬する時は之に對して其使用料を求めざる可からざる等の事あり依りて暫時の間は諸會社は此精算を爲す爲め、通し切符より生ずる收入は各自記帳して其報告を交換し居れり。（1901 年 2 号／「商業世界」／佐野善作；祖山鍾三）
(44) 又交換所は紛失荷物の發見所として、大に便利を與へり。其方法は、紛失又は發見せる荷物の詳細を毎日交換所に報告し、交換所は又各停車場へ其報告を配布す。（1901 年 2 号／「商業世界」／佐野善作；祖山鍾三）

一方で「報告書」の使用は計 80 例、その過半数は 1895 年・1905 年の 2 年分で占めており、「報告」との使い分けについてはさらに詳しく検討する必要がある。
また、(45) (46)など、一部現代語にはあまりない語構成（〇〇デ VN スル）も認められた。

(45) 此語は余が大正五年四月二十三日但馬國の青谿書院に池田草庵先生御贈位報告祭に參拜せしとき、其近傍の寺院に於ける草庵先生始め但馬出身の名士の墨蹟展覽會場に於て看たる所なり。（1917 年 10 号／「我国の徳育と孔子教」／高瀬武治郎）
(46)（※(19)と一部共通） モリソンの文庫の内容は精しく知らぬが、殆んど其の生涯の所得を傾注して蒐集したもので、支那を中心として印度、印度諸島、費島、朝鮮、日本等に關する各國の典籍備はらざるなく、殊に今日にては容易に獲得し難き東洋諸學會の數十年間の報告雜誌全部を網羅し、……（1917 年 10 号／「案頭三尺」／内田魯庵）

なお、上述した A・B の分類とは観点が異なるが、以下の(47)のように「報告」の意味合いが現代語と異なる（この場合、「勧告」に近いであろう）例も見られた。

(47) 又當季中銀行の請求に應じ規則第六十二條に依り過怠金を徵して取引停止の報告を取消したるもの五十六名當季中取引停止解除の請求に依り組合銀行の會議に付したるもの二十四名にして投票の結果に依り解除したるもの十五名否決したるもの九名（1901 年 9 号／「海内彙報」）

４．おわりに

以上、本発表では『太陽コーパス』のデータを利用し、近代語（成立期現代語）における名詞「報告」の統語論的特徴（３．２.）・語構成論的特徴（３．３.）・意味論的特徴（３．４.）を考察した。本発表で明らかになった、『太陽コーパス』における動詞性名詞「報告」の、現代語と特に大きく異なると考えられる特徴は以下に示すとおりである。

①連体修飾に関して

○「X よりの報告」(3)、「Z する報告」(11)など、現代語には見られない連体修飾の構造が認められる。

○「Y への報告」の例が極端に少ない。

②複合・派生・臨時一語形成に関して

○行為主体を前項とする複合・臨時一語形成の生産性が低い（特に固有名の場合）。

○名詞「御報告」・動詞「御報告する」のいずれもきわめて少ない。

③意味の多側面性に関して

○「報告」だけで「報告書」などの結果名詞と同様の意味を表す能力が高い（(42)-(44)）。

これらはいずれも現代語との異同、あるいは現代語に至る変遷といったことを考える上で重要な手がかりになると考えられるが、「現代語と異なる」部分の指摘は発表者の内省によるところが大きく、必ずしもデータに基づいた客観的な考察にはなり得ていない。

同様にコーパスを用いた考察ということを考えた場合、「現代日本語書き言葉均衡コーパス（BCCWJ）」では「報告」の検索結果が 15000 件超と膨大になり、動詞と名詞の厳密な振り分けも含め[12]、本発表で行ったように綿密な分類・分析を行うのは難しい。たとえば３．２．で実例数の少なさを指摘した「Y への報告」などは、直感的に現代語の方が広く用いられると考えられるが、傾向の差を比較する術は今後さらに検討する必要がある。

このように、総体として現代語の名詞「報告」の使用実態が十分的確に把握できているわけではないこともあり、現時点では当時の「報告」について目立った特徴を列挙するにとどまった。現代語との厳密な比較検討は今後の課題とし、その手法を模索していきたい。

文　献

石井正彦（2007）『現代日本語の複合語形成論』 ひつじ書房.

佐藤　佑（2011a）『「太陽コーパス」の入門とケーススタディ』 東京外国語大学大学院 総合国際学研究院 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」

――――（2011b）「現代日本語の事態描写に関わる動詞性名詞と名詞化節の諸相」 東京外国語大学大学院 地域文化研究科 地域文化専攻 博士論文（未公刊）.

西尾寅弥（1961）「動詞連用形の名詞化に関する一考察」 『国語学』43（pp.60-81）

村木新次郎（1991）『日本語動詞の諸相』 ひつじ書房.

関連 URL

佐藤（2011a） Web 公開版
http://cblle.tufs.ac.jp/assets/files/publications/handbooks_06/index.pdf

[12] たとえば「彼にこのプロジェクトは難しい局面にあると報告はしたが理解してもらえたかはわからない」（動詞）と「帰省して両親に結婚の報告はしてきたが、あまりいい顔をされなかった」（名詞）の相違は、前後要素の形態素情報などによって機械的に峻別することが困難である。

名詞「甲斐」の文法的性格

中平詩織（九州大学大学院人文科学研究科）†

Grammatical Properties of Japanese Noun 'Kai'

Shiori Nakahira (KYUSHU UNIVERSITY Graduate School of Humanities)

0　はじめに

本稿で扱う「甲斐」とは、以下の例に表れるもののことである。

(1)　「そうではない。ここに住んでいても、わしなどは頭痛持ちで、数年来、苦しんでいるよ」　「それでは湯浴みなど、やるかいがない」　正興と友人はそういって笑った。［LBp2_00046］

(2)　私が濡れるのはいいけど、配布物が濡れては使い物にならんなる(ママ)。手提げを抱きしめながら傘を差して配布した。がんばった甲斐あって明日は一日ゆっくりできる。［OY14_08062］

(3)　稲場と静子は少年課の人間に相談。努力の甲斐あって、朋美は更生の兆しを見せるが、稲場が「積木くずし」を出版したことでさらなる家族崩壊を招き…。［PM51_00688］

(4)　さつまいもやかぼちゃとか　いろんな野菜がサイコロ状に切ってある　サラダ。コリコリと　かなり触感があって、食べ甲斐がある。。［OY15_07063］

上記の例は全て「甲斐」の前に何らかの要素が接続している。(1)(2)は連体修飾節相当、(3)は名詞修飾、(4)は動詞連用形と、様々な形式に名詞「甲斐」が接続し、また、その直後には「ある」「ない」といった述語が続く。名詞は主語や目的語になりうる品詞であるが、本稿で扱う「甲斐」は、文中に表れる際には文の主語や目的語・補語など自由に立つことはできず、上記のような文末（節末）に限られるといった制限がみられる。また、この「甲斐がある／ない」は一種の文法的な機能をも持っているようである。

本稿では、名詞「甲斐」の述語・前接要素はどのような語であるか、また「甲斐」はどのような文章で表れるかを「中納言」の検索結果から観察する。

1　調査方法

調査には「中納言」を用いた。名詞「甲斐」のみで検索すると、先に(1)(2)(3)(4)で述べた形式以外に、山梨県の地名であったり、名字・名前など人名相当のものであったりする固有名詞なども含まれてしまう。本稿で対象とするのは、固有名詞ではない、文法的に機

† nhs.093004@gmail.com

能する「甲斐」であるため、これでは結果が膨大になり、調査を行うには不適切である。

しかしそれら固有名詞とは別のふるまいを見せる「甲斐」の例を観察すると、以下の条件が見えてきた。

(5)　・「甲斐」の前に何らかの要素を持つ
　　（ア）　連体修飾節（「甲斐」の直前節末の動詞テンス形式はル）
　　（イ）　連体修飾節（「甲斐」の直前節末の動詞テンス形式はタ）
　　（ウ）　名詞＋助詞「ノ」
　　（エ）　動詞連用形[1]
　　・「甲斐」の後に続く述語が「アル」「ナイ」になる

地名・人名などの固有名詞とは異なり、文法的機能をもっていると考えられる名詞「甲斐」は(5)のいずれかの条件にあてはまる。このため、「中納言」で単語に付されている形態論情報を利用して、前接要素の形式ごと[2]に再度検索を行った。検索式は以下の通り。

A：【連体修飾】ル＋甲斐

キー: (品詞 LIKE "動詞%" AND 活用形 LIKE "連体形%") AND 後方共起: 語彙素 = "甲斐" ON 1 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="50" AND tglKugiri="" AND tglFixVariable="2"

B：【連体修飾】タ＋甲斐

キー: (品詞 LIKE "動詞%" AND 活用形 LIKE "連用形%") AND 後方共起: (品詞 LIKE "助動詞%" AND 語彙素読み = "タ" AND 語彙素 = "た") ON 1 WORDS FROM キー AND 後方共起: (品詞 LIKE "名詞%" AND 語彙素 = "甲斐") ON 2 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="50" AND tglKugiri="" AND tglFixVariable="2"

C：名詞＋ノ＋甲斐

キー: 品詞 = "名詞-普通名詞-サ変可能" AND 後方共起: (品詞 LIKE "助詞%" AND 語彙素 = "の") ON 1 WORDS FROM キー AND 後方共起: (品詞 LIKE "名詞%" AND 語彙素 = "甲斐") ON 2 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="50" AND tglKugiri="" AND tglFixVariable="2";

D：動詞連用形＋甲斐

キー: (品詞 LIKE "動詞%" AND 活用形 LIKE "連用形%") AND 後方共起: 語彙素 = "甲斐" ON 1 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="50" AND tglKugiri="" AND tglFixVariable="2"

---

1 動詞連用形に接続する「甲斐」は、しばしばひらがな表記で「がい」と書かれた例が見られた。この「がい」という形式は複合語になる際の連濁だと考えられ、（エ）の「甲斐」は動詞連用形に接続して名詞化する接尾辞だと考えられる。この場合も述語は「アル」もしくは「ナイ」になっており、名詞「甲斐」は述語への制限を有する。

2 「甲斐」に接続する連体修飾節末のテンスに注目すると、それぞれ(1)「ル甲斐」(2)「タ甲斐」の形式となっている。この「ル／タ」は連体修飾節内のテンスとしては同等の構造としてみるべきである。しかし、「中納言」の形態論情報では「ル」は〈動詞・普通形〉に、「タ」は〈動詞連用形・助動詞タ〉に分析される。検索の上では別の構造となってしまうため、別の検索式で検索を行った。

２　「甲斐」がとる述語

「甲斐」が接続する述語は「アル」もしくは「ナイ」が表れる。「甲斐」だけが独立して用いられることはなく、「甲斐がある」／「甲斐がない」といった形式で固定化されて用いられている。

「甲斐」の述語に制限があることについては寺村（1977）に同様の記述がある。また、須永 (2011) では、中古和文における語認定に、語と語の結び付きの強さ（コロケーション強度）を測る論文があり、「かひあり」「かひなし」を一語と認めるか否かという例が挙げられている。本稿で対象とするのは、現代日本語における「甲斐」のふるまいであるが、「甲斐」と存在を表わす述語間に、通時的にも共起関係が見られることの示唆になると考えられる。

３　「甲斐」の前接要素

「甲斐」の前接要素はどのようなものがあるか。注 2 でも先に述べたとおり、(1)や(2)は動詞の連体形で、連体修飾節内のテンスの対立ととれる。これらは連体修飾節として「甲斐」にかかる形式である。

(3)は名詞に助詞がついたものであるが、ここに表れる名詞は「〜する」をつけると動詞としても用いられるサ変動詞に限られる。連体節内の助詞「ノ」が用いられているが、これらは全て「〜シタ」に置き換えても意味が通るため、名詞＋「ノ」も連体修飾節に準じるものであると考えられる。

(4)は動詞連用形に「甲斐」が接続する。上記の 3 例とは構造が異なり、「甲斐」自体もひらがな表記で「がい」になっている例がしばしば見られる。これは和語の複合語化に表れる連濁であり、動詞連用形に「甲斐」が接続して一語となっていると分析できる。このとき、「甲斐」は名詞化接尾辞と呼ぶことができ、上記の 3 形式とは性質を異にする。

（6）　ガンバレ、ニッポン！そして、櫻井君、ＪＵＭＰのみなさんも応援しがいがありましたよね！　［OY04_01074］

「中納言」の形態論情報には、接辞としての「甲斐」と被連体修飾語としての「甲斐」の区別はなく、どちらも名詞として情報が付されるが、用例を観察すると二つの用法があることがわかる。

これらの 4 つの前接要素のうち、一番多く用いられているのは連体修飾節の「タ」であった。（タ：167 例／ル：24 例／名詞ノ：70 例／／動詞連用形：62 例）

「甲斐」の用例の特徴として、複文中に多く表れることも指摘すべき点である。
主節・従属節のどちらにも「甲斐がアル」の形式が見られるが、特に連体修飾節の節末テンス形式が「タ」の形式（B）と名詞＋ノの形式（C）は、「甲斐あって」と従属節で用いられるときの前接要素に例が多い。

(7)　このヘルパーさんと、メーカーの担当営業社員さんの努力の甲斐あって、最近では販売実績が上向き始めてきたということですね。　［OY01_01119］（従属節）

この複文を考える際、「甲斐」が接続している節としていない節とで二つの事態があることが指摘できる。(7)は〔ヘルパーと社員が努力した〕ことにより〔販売実績が上向き始めた〕一種の因果関係とも言える二つの事態が見られる。「甲斐」が接続する節は必ず先行し、その節の事態が起きた結果、別の事態が生じるという構造になる。

これは単文に表れる「甲斐」においても、文脈上で因果関係にあたる事態が存在する。(8)では〔早朝からがんばった〕ことにより〔一位の表彰台に乗る〕事態が生じている。単文の場合はその文以外の文脈で因果関係における事態を補完しており、「甲斐」にかかる連体修飾節が、別の事態を導いたという形式を表わす。

(8)　念願の１位の表彰台に乗ることができて、本当に、去年のリベンジを見事にやってのけたというわけです。寒かったし、早朝から大変でしたが、がんばってきた甲斐はありました。　［OY15_17923］

連体修飾節の「タ」が多い理由は、完了した事態が因果関係の理由となることが多いためであろう[3]。

評価表現の一部に存在の述語形式「アル／ナイ」が使われることは珍しくはない[4]。また、名詞が述部を含んで文末形式化しているという特徴は、井島 (1998) の「組立モダリティ表現」に類似している。井島氏の論文は「[犯人が都内に潜伏している］可能性がある」という文の中の「可能性」といった語に対しての考察である。文末で使用される時に限り、「可能性」にかかる連体修飾節が主題相当となり、後ろの主語に当たる名詞と述語の組立モダリティによって価値判断を下すという説明がなされている。「甲斐」も連体修飾節を必須とし、述語は「アル／ナイ」と存否を表わす。「甲斐がある」は 2 つの事態の因果関係を表わす評価表現のひとつである。

４　「甲斐」が表れる資料

サブコーパスごとの「甲斐」が使用されている数の多寡を確認する。

「甲斐」は書籍（書籍・図書館ともに）・ブログサブコーパスに多く用例が見られる。出版（書籍 76 例・図書館 97 例）とブログ（90 例）は群を抜いて使用数が多く、次いで雑誌（22 例）・ベストセラー（20 例）と続く[5]。

[3] 接続要素・複文で表れる「甲斐」の機能は拙稿(2009)で考察を行った。
[4] 評価のモダリティの一つとして、高梨(2010)では肯否の対立関係「必要がある」「必要がない」を取り上げている。これらは「「ことはない」「こともない」「までもない」に比べ文法化の度合いがかなり低い(p134)」と述べられている。
[5] サブコーパスごとの使用数分布のグラフは 5 章。

しかし、上述したもの以外のサブコーパスでは「甲斐」の使用数はすべて 10 例未満となっている。新聞・広報誌・白書等のサブコーパスには見られない。

これらの違いは「甲斐がある」という表現が口語的であること、また評価性を持つ表現であることが関係していると考えられる。前章で確認したが、「甲斐」は二つの事態の因果関係を表わす。このため前後の文脈を長くとれる小説やブログなどには多く用例が見られるようになり、字数の制限がある新聞・広報誌には用いられにくい。また、小説などの会話文に用いられたり感想を述べる部分は、書き言葉均衡コーパスのなかでも口語性が高いといえるため、その特徴が反映されたと考えられる。

５　「価値」との比較

「甲斐」と統語的に類似する「価値」について比較する。

「価値」も連体修飾節が接続し、文末で用いられる際は次のように存在述語「アル」もしくは「ナイ」を伴う。また、「価値がある」全体で固定化され、連体修飾節で表現された評価を行う形式となっている。

(9)　a. それはなんですかと尋ねると、「『茅の輪』です。きょうは大きな茅の輪をくぐる七夕祭りをやっているんです」と。これは一見の価値ありと喜び勇んで、天満宮と七夕祭りの関係を探るべく鳥居に向かった。　［PM21_00718］

b.リアップのシャンプーとコンディショナーを使用してますが、このシャンプーとコンディショナーが本当にいいんです。一寸高めですが、試す価値あり。発毛剤まで使えとはいいませんが。　［OC09_02675］

名詞「価値」には今まで見てきた「甲斐」との共通点がいくつか見られるが、「甲斐」との違いはどこにあるだろうか。　1 章で確認した「甲斐」と同じ検索式で検索語を「価値」に入れ替え、検索を行った。「価値」には連用形に接続する形式はないため、A,B,C のみを検索した。また、述語が「アル／ナイ」になっていないものは今回の考察対象外として除外した。

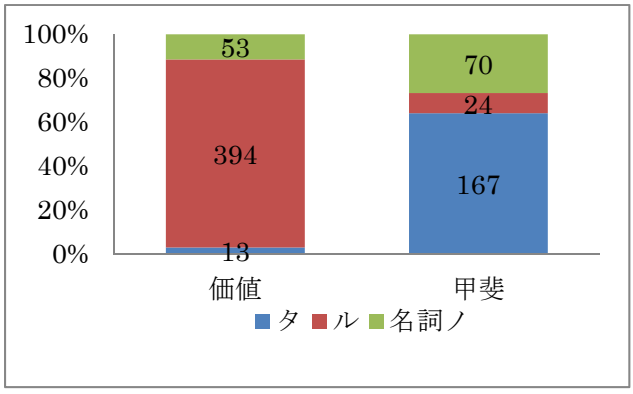


図 1　「甲斐」と「価値」の前接要素の割合

図 1 は「甲斐」「価値」にかかる連体修飾節末のテンス形式と、名詞＋「ノ」形式の割合をみたものである。

「甲斐」は連体修飾の節末テンス「タ」に接続する形式が最も多いが、「価値」では対象的に「タ」が最も少なく、「ル」の形式が最も多く、次いで名詞＋「ノ」がみられた。この名詞＋「ノ」は「一読」「一見」などが多く見られ、また節末「ル」の文末形式も「〜テミル価値あり」という形式が多かった。「試す」という動詞などからも、意味的な共通性が見受けられる。

(10)　五十代以上の場合は「失敗したらやり直し・再就職ともにより厳しい」という点を考慮し、借金を清算した上で再就職の見込みがある人なら、検討してみる価値ありと言えるでしょう。　［LBr5_00042］

「甲斐」は因果関係を表わすことも可能であるため、複文中に表れることもあったが、「価値」は複文中では現れず、単文にしか例が見られない。

「価値」は評価を表わすといっても、(9b)のように何かを行うことに関しての評価を行うものである。そのため、「一読」や「〜テミル」といった〔何かを試す〕表現が接続しやすいと考えられる。試した後の感想などは必須のものではないため、複文には表れない。

次に、「甲斐」と「価値」のサブコーパスごとの使用数の比較を確認する。

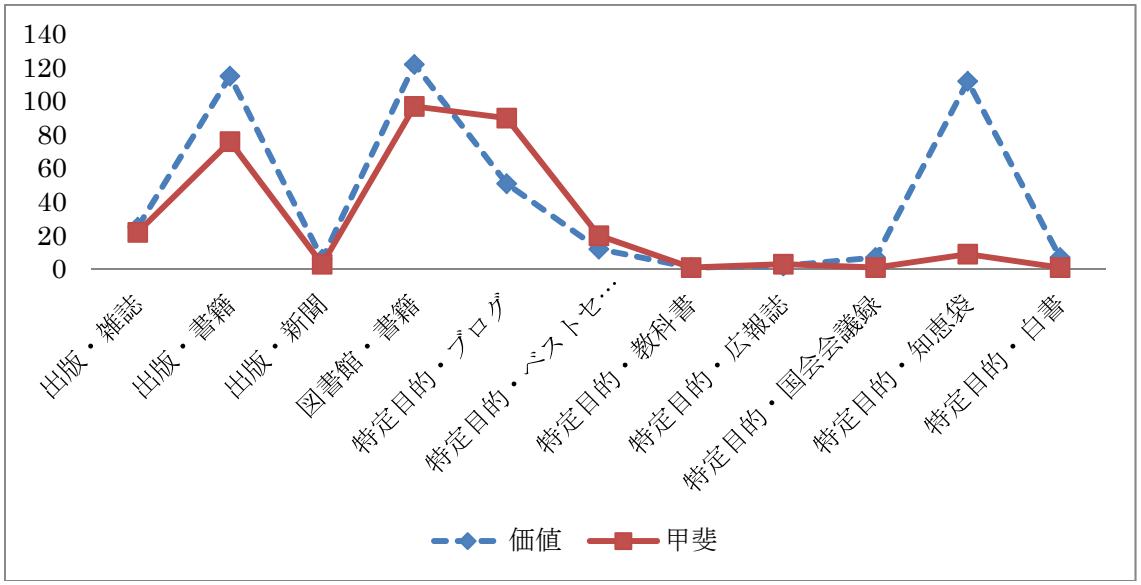


図 2　サブコーパスごとの使用数分布

使用数に違いはあるものの、「甲斐」と「価値」のサブコーパスごとの使用度合いがわかる。2 つの語は、書籍に用例が多いこと、新聞・教科書・広報誌・国会会議録・白書に用例が少ないことは共通している。しかし、同じインターネット資料であるブログと知恵袋に

おいて語ごとの使用数の偏りが見られるのが特徴的である。ブログには「甲斐」が用いられることが多く、知恵袋には「価値」の用例が多い。

これは、「甲斐がある」「価値がある」の用いられる文脈の違いがサブコーパスの性格の違いに反映されているといえる。

先に述べたように、「価値がある」は〔何かを試す〕ときに用いられる。Yahoo!知恵袋は利用者が質問を行い、別の利用者が回答をしていくサービスである。評判を問うたり使う前に確認をしたい、もしくはおすすめのものを教えてほしいといった質問に対して、推薦する文言に個人の評価「一見の価値あり」などを付け加えるため、「価値」の用例が多いのだと考えられる。次の(11)は知恵袋サブコーパスからの例で、「〜教えてください。」までが質問、「切れなくなったハサミも〜」からが回答となる。

（11）　切れなくなった爪切りを再び復活させる方法があれば教えてください。切れなくなったハサミも、アルミホイルをくしゃくしゃしたものを何度か切ると復活します。爪きりでも試してみる価値はあります。　［OC08_02122］

ブログは感想を書くためにも用いられるが、日記として利用されることが多いサービスである。因果関係や前後の文脈を入れることが可能なため、「甲斐」が指向されると考えられる。

６　おわりに

本稿では名詞「甲斐」が持つ文法的性格について考察を行った。以下に「甲斐」の特徴をまとめる。

（12）
- 「甲斐」は存在表現「アル」「ナイ」と共起する。
- 前接要素は連体修飾節（ル形・タ形）／名詞＋助詞「ノ」／動詞連用形がある。このうち、動詞連用形のときは「甲斐」は接尾辞相当となる。
- 「甲斐がアル／ナイ」は二つの事態の因果関係を評価する表現である。

本稿では意味の考察で連体修飾節相当のもののみを取り扱ったが、動詞連用形に接続する接尾辞「甲斐」の機能も考察が残されている。これは別稿に譲る。

また、「価値」を検索した際、「価値がある」といった形式より「価値あり」と助詞を省略し、言い切りの文末になるものがほとんどであった。このような「〜〜の価値あり」といった定型表現との関わりについても考察の余地があるだろう。

「甲斐」も「価値」も存否の述語をとるとしたが、「アル」のほうが用例がやや多かった。これに対しては明確な分析ができていないため、今後の課題としたい。

参考文献

井島正博 (1998) 「組立モダリティ表現」『東京大学国語研究室創設百周年記念国語研究論集』汲古書院

影山太郎 (1993) 『文法と語形成』ひつじ書房

須永哲也 (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

高梨信乃 (2010) 『評価のモダリティ　現代日本語における記述的研究』くろしお出版

寺村秀夫 (1977) 「連体修飾のシンタクスと意味―その 3―」寺村 1992 所収

――――― (1992) 『寺村秀夫論文集 I 』くろしお出版

中平詩織 (2009) 「形式名詞「甲斐」の意味・構造に関する一考察」『福岡大学日本語日本文学』19

『現代日本語書き言葉均衡コーパス』（略称 BCCWJ）検索ツール

短単位検索 Web アプリケーション「中納言」URL: http://chunagon.ninjal.ac.jp/search

中国人日本語学習者の「的」付きナ形容詞の習得に関する研究
—BCCWJ コーパス調査とアンケート調査の分析を通じて—

呉　雪梅（一橋大学言語社会研究科）

Chinese Learners' Uses of Na-Adjective Ending in "-teki": Based on Analysis of BCCWJ and Questionnaire Surveys

Wu　Xuemei (Hitotsubashi University Graduate School of Language and Society)

１．要旨

日本語では明治以降、接尾辞の「的」付きナ形容詞が大量に作られてきた。以前は、藤居（1957）のように「的」が濫用されているという指摘もあったが（藤居（1957））、現在では、「的」付き形容詞の使用が定着している。小出（2004）ではさらに新しい意味でも使われているという指摘がある（小出（2004））。一般的に、「的」付きナ形容詞の形成パターンは、「漢語＋的」「外来語＋的」「和語＋的」「混種語＋的」という 4 つである。王娟他（2001）によると、その中では「漢語＋的」というパターンが最も多いと言われている。

接尾辞「的」の意味は、遠藤（1984）によって、以下の 4 つに分類されている。

a.「～に関する」、「～について」:「科学的な説明」「教育的な立場」
b.「～のような性質を有する」:「代表的な書物」「一般的な知識」
c.「～の状態にある」、「～らしい」:「圧倒的な勝利」「貴族的な顔」
d.「～としての/である」:「職業的な軍人」「物的な証拠」

一方、中国語にも助詞“的”がある。中国語の助詞“的”は複雑な働きを持っている。中には、“健康的生活”（健康的な生活）、“一般的说法”（一般的な言い方）のように、形容詞と名詞の中に置かれ、性質を表す機能がある。

『現代漢語詞典　第５版』（商務印刷局出版）及び『大辞林 第三版』により、中国語の“的”と日本語の「的」の意味と用法について、以下のように整理をした（次ページ表 1 参照）。

中国語の“的”は、助詞として、日本語の接尾辞の「的」と同じように、名詞の後ろに置いて使われているが、名詞に付いて、修飾機能のある独立語となる機能はないと考えられる。

それぞれの用法は違うが、両者は形が似ており、意味が対応して、同じである場合が少なくない。中国と日本は同じ漢字圏であり、このような「漢語＋的」ナ形容詞の習得には、中国人学習者にとって、メリットがあると思われる一方、曹・仁科（2006）で指摘されているように、中国人日本語学習者の「的」の過剰使用及び「的」の脱落などの誤用もよく見られる。

本研究は、日中両国語を対照する観点から、「的」付きナ形容詞と非「的」付きナ形容詞について学習者の誤用の特徴を分析し、指導方法の探索を試みるものである。

表 1 中国語の「的」と日本語の「的」の比較

中国語の“的”
＊助詞、非独立語 **1.形容詞と中心語の中に置かれ、二語の修飾関係を表す：** “幸福的生活”（「幸福な生活）、“漂亮的衣服”（「綺麗な服」）など **２．名詞と名詞の間に置かれ、所属関係を表す（日本語の「の」に相当する）：** 「我的书」（私の本） “这本书，是你的吗？”「この本は、君のですか？」 **３．特定の人や、職業を指す：** “男的”（「男の人」）、“送报的”（「新聞配達員」） **４．述語の後に置かれ、動作主や、時間、場所などを強調する。この用法は、過去の状況に限る。** “谁买的书？”（誰が買った本なの？」） “他是昨天进的城。”（彼は昨日町に来た。） **５．句末に置かれ、肯定のニュアンスを強調する。** “这件事儿我是知道的。”（この件について、私が知っていた。」） **６．数詞の真ん中に置かれ、「かける」と「プラス」の意味である。** “五米的三米”（「五メートルかける三メートル」）、 “两个的三个”（「二つプラス三つ」）

日本語の「的」
＊接尾辞、非独立語 **１．名詞およびそれに準する語に付いて、形容動詞の語幹を作る。** ァ主に物や人を表す名詞について、それそのものではないが、それに似た性質を持っていること表す。～よう。～ふう。 例：「母親的な存在」、「スーパーマン的な働き」 ィ主に抽象的な事柄を表す漢語に付いて、その状態にあることを表す。 例：「印象的な光景」、「積極的な行動」 ゥ物事の分野・方面などを表す漢語について、その観点や側面から見て、という意を表す。 例：「事務的な配慮」、「学問的に間違っている」 **２．人の名前・行為・職業などを表す語、またはその一部に付いて、それに対する軽蔑や親しみの気持ちを表す。** 例：「泥的」＝「泥棒」、「正的」＝正雄・正子など

２．先行研究

「的」付きナ形容詞と非「的」付きナ形容詞について、言語対照や、意味論、量的な分析など、多くの視点から、数多くの先行研究がある（遠藤（1984）、堀口（1992）、南雲（1993）、丸山（1993）、望月（2010）など）が、「的」付きナ形容詞の習得指導に関する論文は、多くはない（梁（2005）、曹・仁科（2006）、望月（2010）など）。望月（2010）及び曹・仁科（2006）は、学習者の書いた作文をデータ化し、そこから誤用の分析をした上で、指導方法を提案したものである。しかし、これらの提案が役に立つかどうかを検証した先行研究は、筆者の知る限り、まだないようである。

「的」と一緒に共起する語幹の判断基準は様々な見解があり、統一されていない。言語コーパスを用いて、客観的に検証し、整理する必要があると思われる。

３． 事前調査

事前調査として、以下のような方法でコーパスの検索結果と母語話者及び学習者のアンケート結果と比較した。

（1）『日本語能力試験出題基準』(2006 年 3 月版）に掲載されている N2 レベルまでの二字漢語を調査対象（1802 語）とする。

（2）すべての二字漢語が語幹として接尾辞「的」を付けて使えるかどうかを国立国語研究所の『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と言う）で検証する。検証の結果は次ページの図 1～3 に示した。

（3）抽出した二字漢語が「的」と一緒に使えるかどうかを母語話者と中国人日本語学習者に答えてもらった。

（4）BCCWJ で検証した結果と、母語話者と学習者の答を対照し、それぞれの答の違いや、学習者の誤用とその特徴（どの意味で間違えやすいのか、どの文脈で間違えているのか）を分析し、その誤用を防ぐ指導方法を探る。BCCWJ と母語話者、学習者の比較結果は図 4～図 8 に示した。

事前調査は、2012 年 6 月に日本語母語話者 5 名（男 2、女 3）と中国人日本語学習者 10 名[1]（男 4、女 6）へのアンケート調査を実施した。アンケート調査の内容は、抽出した 1802 個の二字漢語が「的」と使えるかどうかを被験者に判断してもらうということである。
調査量が多すぎるため、一定した正確な結果が出にくいことが分かった。したがって、より自然な結果が出やすい事前調査をもう一回行う必要があると考えられる[2]。今回は、一回目の事前調査で出たデータその分析を紹介する。

４． 事前調査で得たデータについて

4.1 BCCWJ での検索結果：

「的」が付く漢語は、697 語（38.68%）であった。一方、「的」が付かない漢語は、990 語（54.94%）であった（この中には、「的」が付かないナ形容詞 132 語（6.60%）が含まれている)。なお、「条件付き」というのは、「国家主義的」のように、語幹の前に名詞を加えて、「○○＋語幹」＋「的」というような形で「的」と共起している例を指す。

		数			割合
語幹の品詞とその割合	名詞	1712 95.00%	名詞・普通名詞・一般	946	52.25%
			名詞・普通名詞・サ変可能	593	32.91%
			名詞・普通名詞・副詞可能	90	5.00%
			名詞・普通名詞・形状詞可能	76	4.22%
			名詞・普通名詞・サ変形状詞可能	14	0.78%
			名詞・普通名詞・助数詞可能	5	0.28%
			固有名詞・地名・一般	1	0.06%
	副詞			19	1.05%
	形状詞			77	4.27%
	延べ語数：1821　　異なり語数：1802				

図 1　BCCWJ の検索結果

[1] 10 人の学習者はすべて超上級レベルである。
[2] 第二回の調査は現在実施中である。

「的」と共起できない語とその割合		
品詞	数	割合
ナ形容詞	132	7.33%
名詞	851	47.23%
名詞・サ変可能	244	13.54%
副詞	6	0.33%
その他	13	0.72%

図 2 「的」と共起しない語

語幹の品詞とその割合		
品詞	数	割合
名詞	1141	63.32%
名詞・サ変可能	508	28.19%
ナ形容詞	132	7.33%
副詞	17	0.94%
そのほか	13	0.72%
延べ語数：1811		
異なり語数：1802		

図 3 語幹の品詞

4.2 母語話者と学習者の比較結果

A.1082 語の中で、「的」と一緒に使えるとされた判断率。

母語話者：864 語（33.83%）

学習者： 743 語（41.21%）

B.選択した語幹の一致率

母語話者： 58 語（3.22%） 図 4 参照

学習者： 6 語（0.33%） 図 5 参照

母語話者と学習者が一致した語：4 語（政治、国際、世界、意味）

	語数	割合
総語数	1802	100%
全員一致	58	3.22%
4 個一致	97	5.38%
3 個一致	100	5.55%
2 個一致	187	10.38%
一個一致	394	21.86%
合計	864	47.89%

図 4 母語話者の一致率

	語数	割合
総語数	1802	100%
全員一致	6	0.33%
9 個一致	11	0.61%
8 個一致	22	1.22%
7 個一致	24	1.33%
6 個一致	32	1.78%
5 個一致	30	1.66%
4 個一致	59	3.27%
3 個一致	86	4.775
2 個一致	137	7.60%
1 個一致	338	18.76%
合計	745	41.34%

図 5 学習者の一致率

図 4 と図 5 から、母語話者や学習者にもかかわらず、「的」と一緒に使える語幹の選択基準がそれぞれであり、一致率が極めて低いということが分かった。

4.3 「的」の付かないナ形容詞（132 語）について（図 6）[3]

「的」の付かないナ形容詞の場合、母語話者の正解率が高く、一方、学習者のほうが色々間違えやすいことが見て取れる。指導方法として、まずは、これらの「的」の付かないナ形容詞を整理して学習者に教える必要があると思われる。

4.4 上位 100 位語幹の特徴とは？（図 7）[4]

「的」が付く二字漢語は、BCCWJ の検索で、用例の多い上位 100 位の語幹の特徴について、現段階、以下のようなことが分かった。

A. すべて名詞である。その中では、サ変可能なものが 18 語、サ変かつ形状詞になれるのは 1 語（「直接」）、副詞および接続詞として使われるのが 1 語（「一方」）、接続詞として使われるのは 1 語であった（「絶対」）。

B.上位 100 位の語幹は、母語話者と学習者の一致率が高い。

4.5 「的」と共起しない語幹について（図 8）[5]

BBCWJ を利用し、「的」と共起しない二字漢語は図 8 の通りである。しかし、母語話者と学習者もこのような語を語幹として迷ったり、選んだりしている。この傾向を見て、いくつかの疑問があった。

---

[3] 図 6 は 132 語から 50 語を抽出したものである。

[4] 50 個を抽出したものである。データの整理は未完成である。

[5] 元々「的」が付かない語彙は 984 個であり、「～な」形のナ形容詞を除いて、残り 865 語であり、その中から 50 語を抽出したものである。

A. このような語は、どのような特徴をしているのか。

このような語は、ほとんど名詞であり、中には、サ変動詞になれる語もある。他には、副詞も含まれている。このような語の性質をみると、変化を表さない、さらに「性質」を表すこともない。そのため、「的」と共起できないと考えられる。

サ変可能の名詞については、「的」と共起できる名詞もあれば（例えば、「比較」、「代表」である。図７を参照）、図８の中にある「演奏」、「引退」という語は、同じくサ変動詞になれるが、「的」と共起できない。「的」と共起できるサ変可能の名詞はどのような特徴をしているのかをこれからの課題として分析を深めるつもりである。

B. 今回の調査で、学習者からフィードバックをもらった。中国人の学習者には、「的」という語は、とても抽象的な意味であり、話し言葉より、書き言葉特に論文を書くときに使うという意見があった。ほかに、「的」がいつ付くかということについて、多くの学習者は「なんとなく感覚で」と答えた。母語話者と学習者がどのような基準でこのような語幹を選んだのか。たとえば、どのような場面や、文脈を考えながら、その語幹を選んだのかについて、調査したい。今後の調査には、被験者にインタービュする工夫が必要と思われる。

５． まとめ

本稿では、現在進めている修士論文の構想と実際行われた事前調査の結果を紹介した。

接尾辞「的」に関する研究は、数十年前から行われてきた。最初は、「的」の濫用問題として取り扱われていたが、後には「的」の意味や、「的」の量的特徴、「的」の日中対照研究などの観点から数多くの研究がなされている。そして、「的」の習得問題にも関心が集まっている。しかし、言語コーパスのない時代には、データの分析方法は非常に限られていた。「的」付きナ形容詞と非「的」付きナ形容詞の特徴について、遠藤（1984）、王娟他（2001）、原（1986）、堀口（1992）などいつくかの先行研究はあったが、はっきりできた先行研究はいまだにないと言えるだろう。

今回の事前調査で、学習者だけではなく、母語話者にも「的」付きナ形容詞の誤用が見られる。「的」付きナ形容詞の使用は個人的な部分もあり、さらに「的」の使い方はますます変化し、「的」が付く語幹への容許度も上がっていくので、はっきりそのルールを出すのは難しいと思われる。

調査対象を絞って、言語コーパスを利用し、科学的な分析方法である程度のルールをまとめ、頻度の高い語彙に対して、その特徴を整理し、自分なりの指導方法を生み出して、学習者の習得に役に立つことを期待している。

漢語	級	形容詞	母語話者					中納言	中国人学習者									
			A	B	C	D	E		F	G	H	I	J	K	L	M	N	O
率直	2	率直な						×		○					○	○		
卑怯	2	卑怯な					○	×										
冷静	2	冷静な						×		○			○			○	○	
立派	4	立派な						×										
乱暴	2	乱暴な					○	×	○				○			○		
余計	2	余計な						×										
幼稚	2	幼稚な						×		○								
陽気	2	陽気な						×		○								
容易	2	容易な						×		○								
愉快	2	愉快な						×										
有利	2	有利な						×		○			○					
有名	4	有名な						×										
有能	2	有能な						×		○								
優秀	2	優秀な						×									○	
厄介	2	厄介な						×										
面倒	2	面倒な	△				○	×										
迷惑	2	迷惑な						×	○	○								
明確	2	明確な					○	×		○		△				○	○	
無理	3	無理な						×										
豊富	2	豊富な						×										
膨大	2	膨大な						×										
便利	4	便利な						×										
平凡	2	平凡な						×								○		
平気	2	平気な						×										
不利	2	不利な						×									○	
不満	2	不満な					△	×									○	
不便	3	不便な						×										
不平	2	不平な	△				△	×					○					
普通	3	普通な						×										
不正	2	不正な		○				×						○				
無事	2	無事に					○	×										
不幸	2	不幸な					△	×										
不潔	2	不潔な						×										
複雑	3	複雑な						×									○	
不運	2	不運な					○	×										
微妙	2	微妙な					○	×										
必要	3	必要な						×		○							○	
莫大	2	莫大な					○	×										
馬鹿	2	馬鹿な						×										
呑気	2	呑気な						×										
熱心	3	熱心な						×										
独特	2	独特な						×	○									
得意	2	得意な						×										
透明	2	透明な						×										
当然	2	当然な					△	×									○	
天然	2	天然な					○	×					○		○		○	
適度	2	適度な						×										
適当	3	適当な						×		△								

図 6 「的」の付かないナ形容詞への判断

（凡例）○：「的」が付く／△：はっきり判断できない／×：「的」が付かない。以下の表も同じ

漢語	級	例文数	母語話者					中納言	中国人学習者									
			A	B	C	D	E		F	G	H	I	J	K	L	M	N	O
具体	2	10288	○	○	○		○	○	○		○	○	○	○	○		○	○
基本	2	9802	○	○	○	○	○	○	○			△	○		○	○	○	○
一般	2	6348	○			○	○	○	○	○		○	○	○	○		○	○
社会	3	5622	○	○	○	○	○	○	○	○	○	○		○	○	○	○	
比較	2	4094	○	○	○	○		○	○			○	○	○		○		○
経済	3	3950	○	○		○	○	○	○	○	○	○	○	○	○	○	○	○
個人	2	3376	○	○	○	○	○	○	○	○	○	○	○		○	○	○	○
効果	2	3277	○	○	○	○	○	○	○	○		○	○			○	○	○
精神	2	3178	○	○	○		○	○	○	○	○	○	○	○		○		○
政治	3	3154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
最終	2	2955	○	○	○	○	○	○	○	○		○	○				○	
国際	3	2775	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
歴史	3	2690	○	○	○			○	○	○	○	○			○		○	○
世界	3	1936	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
代表	2	1828	○	○			△	○	○			○	○		○	○		△
主義	2	1745	○			○		△										
合理	2	1715	○	○	○		○	○	○	○		○	○	○	○	○	○	
長期	2	1700	○	○	○	○	○	○	○			○		○	○	○	○	○
科学	3	1683	○	○		○	○	○	○		○	○	○		○	○	○	○
技術	3	1679	○	○	○	○	○	○	○	○	○	○	○		○	○	○	○
自動	2	1671	○	○		○	○	○	○		○	○	○	○	○	○	○	○
一時	2	1668	○				○	○	○			○	○		○	○	○	
徹底	2	1563	○	○		○	○	○	○	○		○					○	
現実	2	1513	○	○	○	○	○	○			○	○	○	○		○	○	○
決定	2	1468	○	○	○	○	○	○	○		○	○	○				○	○
魅力	2	1431	○	○			△	○	○					○	○	○	○	○
全国	2	1424	○		○		○	○	○		○	○	○		○	○	○	
計画	3	1423	○	○	○	○	○	○	○	○	○	○	○	○	○	○		○
定期	2	1422	○					○	○	○					○	○		
一方	2	1339				○		○	○	○	○		○		○		○	
結果	2	1309	○	○	○		○	○	○		○	○	△		○	○	○	○
専門	3	1268	○	○	○	○	○	○		○	○	○	○	○	○		○	○
典型	2	1260	○		○		○	○	○	○	○		○			○	○	○
文化	3	1234	○	○	○	○	△	○	○	○	○	○			○	○	○	
全体	2	1210	○	○	○		○	○	○	○	○		○		○			
日常	2	1169	○	○		○		○	○	○	○		○				○	
心理	2	1136	○	○	○		○	○	○				○	△		○		○
直接	2	1117	○	○		○	○	○	○						○		○	○
絶対	2	1074	○	○	○		○	○					○		○	○	○	
物理	2	1018	○	○	○			○	○			○	○		○			
宗教	2	1002	○	○	△	○	○	○	○	△		○				○	○	○
時間	4	923	○	○		△	○	○		○	○	○	○		○	○	○	○
理想	2	906	○	○	○			○	○	○	○	○	○		○	○	○	○
対照	2	875	○		○	○	○	○	○		○	○	○	○	○	○	○	△
中心	2	862	○	○			○	○	○			○	○	○	○			
継続	2	843	○	○	○			○					○			○	○	△
基礎	2	842	○			○	○	○	○	○	○	△	○		○	○	○	○
安定	2	823	○	○	○		○	○	○	○					○			
人間	2	822	○	○	○			○	△									
印象	2	821	○	○	○	○	○	○		○		△	○		○		○	○

図 7 上位の語幹と選択の分布

漢語	級	母語話者					中納言	中国人学習者									
		A	B	C	D	E		F	G	H	I	J	K	L	M	N	O
挨拶	3						×			○							
愛情	2						×										△
合図	2					△	×										
握手	2						×										
圧縮	2						×										
安心	3						×			○					○		
案内	3						×										
委員	2						×										
以下	3						×										
以外	3					△	×										
育児	2					○	×					△					
以後	2						×										
以降	2						×										
医師	2						×										△
医者	4						×								○	○	
以上	3						×										
椅子	4					○	×										
以前	2						×										
一々	2						×										
一応	2						×										
一段	2						×					○					
一度	3						×		○	○							
一部	2	△					×		○			○			○	○	
一流	2						×		○			△				○	
一種	2						×										
一瞬	2						×					△				○	
一緒	4					○	×										
一生	2						×				△	○					
一旦	2	△					×										
一定	2						×	○	○							○	
移動	2						×					○					
以内	3					△	×										
違反	2						×										
衣服	2	△					×										
以来	2						×										
依頼	2						×			○							
引退	2						×										
引用	2						×										
引力	2					△	×								○		
有無	2						×										
運転	3	△				○	×								○		
英文	2						×			○							
英和	2						×										
液体	2						×	△				△		○			△
宴会	2					△	×										
延期	2						×										
演説	2						×										
演奏	2						×										
遠足	2						×										
煙突	2						×										

図 8「的」の付かない語幹と選択の分布

参考文献

宇佐見英美子（2001）「接尾辞『～的』について」　『津田塾大学紀要』33. PP.239-261 津田塾大学紀要委員会
遠藤織枝（1984）「接尾語『的』の意味と用法」『日本語教育』53：PP. 125-138
王娟・曲志强・林伸一（2001）「『的』付きナ形容詞と非『的』ナ形容詞の分類と意味の特徴」　『山口国文石井大教授退官記念特集号』PP. 124-146、山口大学文理学部国語国文文学会
謝貴蘭（1993）「日本語の『的』と中国語の『的』について」『日本語教学国際研討論文集』東呉大学　PP. 257-270
曹红荃・仁科喜久子（2006）「中国人が学習者の作文誤用例から見る共起表現の習得―名詞と形容詞及び形容動詞の共起表現について―」　PP. 70-79
『日本語教育』130 号　特集コーパスと日本語教育――現状と課題――
小出慶一（2004）「接辞「的」の新しい用法――「的には」という用法について――」『群馬県立女子大学国文学研究』24, PP. 1-14、群馬県立女子大学国語国文学会
『日本語能力試験出題基準』　2006 年 3 月　　凡人社出版
望月通子（2010）「日本語学習者と母語話者の『的』の使用研究―コーパスを利用して」『2010 世界日本語教育大会論文集・予稿集』398　PP.1-10
原由起子（1986）「―的―中国語との比較からー」『日本語学』5 巻 3 号　PP. 73-80
堀口　和吉（1992）「助辞「～的」の受容」『山辺道』第 36 号, PP.59-76　天理大学国語国文文学会
藤居信雄（1957）「的ということば」　『言語生活』71：PP. 71-76
藤居信雄（1961）「的の意味」　『言語生活』119　PP. 80-83
丸山千歌（1996）「英語の接尾辞-tic の訳語「的」について――『中央公論』1962 年 11 号の場合―」PP. 15-42『ICU 日本語教育研究センター紀要　6』国際基督教大学日本語教育研究センター
南雲千歌（1993）「現代日本語の「的」について――雑誌『中央公論』1992 年 11 月号の場合」PP. 72-98　『ICU 日本語教育センター紀要　3』国際基督教大学日本語教育研究センター
梁賢後（2005）「韓国人学習者の漢語系形容動詞の習得に関する研究――中国人学習者との比較を中心に――」日本語教育学会秋季大会　金沢大学　　PP.103-114
山下喜代（1999）「字音接尾辞「的」について」『日本語研究と日本語教育』森田良行教授古稀記念論文集刊行会編　PP. 24-38

関連 URL

国立国語研究所『現代日本語書き言葉均衡コーパス』
https://chunagon.ninjal.ac.jp/

「語彙レベル」から見た近代の語彙と現代の語彙
—『太陽コーパス』と『現代日本語書き言葉均衡コーパス』を用いて—

田中 牧郎　（国立国語研究所言語資源研究系）†

Vocabulary Level in Modern and Contemporary Japanese:
Based on Analyses of "Taiyo Corpus" and "Balanced Corpus of Contemporary Written Japanese"

TANAKA Makiro (National Institute for Japanese Language and Linguistics)

１．背景と目的

近代から現代にかけて、日本語の語彙には大きな変化があったが、その変遷の具体的過程は明らかになっていない。こうした研究は、通時的なコーパスを作って記述することによって進めることが望まれよう。国立国語研究所が公開した『太陽コーパス』は、明治時代後期から大正時代を対象とするコーパスで、『現代日本語書き言葉均衡コーパス』は現代を対象とするコーパスである。それぞれが扱う時代の間に昭和時代が欠落しているなど、近代から現代への変遷をとらえるには、この二つのコーパスでは不足だが、明治後期から大正期と現代とを対照することで、近代語から現代語への通時的研究への見通しをつけていくことも可能ではないかと思われる。

語彙の変遷を扱うには、語彙の全体を把握しつつ、変化しない部分と変化する部分とをより分け、変化した部分についてその背景や事情を考察していくことが重要だと思う。そのような研究のためには、語彙を計量的に扱う方法が有効であり、コーパスに付与された形態論情報を用いることが考えられる。『現代日本語書き言葉均衡コーパス』には、形態素解析辞書 UniDic によって形態論情報が付与されているが、『太陽コーパス』にはこれがない。しかし、近時、近代語テキストにも近代語用の UniDic を整備し形態素解析を適用する研究が進んできているので、試験的に『太陽コーパス』に形態論解析を付与したデータを用いて、『現代日本語書き言葉均衡コーパス』のそれと対照してみたい。その際、二つのコーパスの語彙を対照するための枠組みとして、使用頻度に基づいて語彙を階級に分ける「語彙レベル」を用いる。

２．「図書館書籍サブコーパス」と『太陽コーパス』の語彙レベル

『現代日本語書き言葉均衡コーパス』（BCCWJ）は、多様な媒体のサブコーパスから構成されることを特徴としている。筆者らは、特定領域研究「日本語コーパス」言語政策班の研究の一つとして、BCCWJ のサブコーパスのうち６種について語彙調査を行い、「BCCWJ 主要コーパス語彙表」作成し、公開した（田中・近藤 2011）。この語彙表には、６種のサブコーパスごとに、出現した語彙すべての度数、使用率、使用サンプル数、そして語彙レベルの情報が一覧できるようにしたものである。このうち、「語彙レベル」とは、語彙を度数の高いものから順に並べ、上位の語から度数を累積していき、その累積度数が延べ語数の

† mtanaka@ninjal.ac.jp

何パーセントを占めるかという累積使用率（カバー率）によって、５段階に分けたものである。田中・近藤（2011）から、レベルを区画する基準（表１）と、「図書館書籍サブコーパス」（固定長）について五段階に分類した語数（表２）を下に示す。なお、UniDic で付与される品詞情報のうち、助詞・助動詞・記号類・未知語等は対象外としている。

表１　語彙レベルとカバー率

語彙レベル	カバー率（累積使用率）
a	0 - 78%
b	- 88%
c	-94%
d	-97%
e	-100%

表２　図書館書籍における語彙レベルごとの語数

語彙レベル	延べ語数	異なり語数
全体	3,938,696	86,002
a	3,074,655	4,177
b	395,994	6,330
c	242,911	11,595
d	118,642	14,176
e	106,494	49,724

BCCWJ のサブコーパスうち、幅広いジャンルについて書かれていてよく読まれている媒体は何かと言えば、「図書館書籍」であろう。なぜなら、「図書館書籍」は、公共図書館の多くに共通して所蔵されているものからサンプリングされたものだからである。そこで、本稿でも、この図書館書籍の語彙レベルを、現代語の語彙レベルを代表しているものと扱うことにした。

『太陽コーパス』は、1895（明治 28）年から 1928（昭和 3）年まで刊行された総合雑誌『太陽』を対象とするコーパスで、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年分の全文がおさめられている。一資料のみが対象だが、この雑誌が当時よく読まれ、幅広い層の著者が広範なジャンルの記事を書いていることから、当時の書き言葉をある程度代表できるものと考えられる。

表３　『太陽コーパス』における語彙レベルごとの語数

語彙レベル	延べ語数	異なり語数
全体	5,384,879	74,089
a	4,199,256	3,476
b	539,920	4,835
c	326,250	8,834
d	161,045	10,177
e	158,408	47,267

この『太陽コーパス』についても、上記と同じカバー率の基準で語彙レベルに分け各語にレベル情報を付与した。その結果を示すと、表３のようになる。表２と表３を比べると、延べ語数は『太陽コーパス』の方がかなり多いが、異なり語数は「図書館書籍」（固定長）の方が多い。二つのコーパスの語彙のありようは異なっていることが見て取れる。

3. 近代の語彙と現代の語彙の対照

3.1　近代は基本的レベルで現代は周辺的レベルの語彙

語彙レベルによって、近代の語彙と現代の語彙を対照した先行研究に、近藤・小木曽（2009）があり、語種構成比率など全体的観点から考察が行われている。本稿では、個別の語にまで焦点を当ててみたい。とくに、近代と現代とで語彙レベルに非常に大きな変動があるものに注目したい。

まず、『太陽コーパス』ではレベル a となっていて、近代では最も基本的なレベルの語彙でありながら、BCCWJ の「図書館書籍」（固定長）では、最も周辺的なレベル e のものを見ていこう[1]。表 4 は、品詞と語種によって分類して、そのすべて（固有名詞を除く）を示したものである。品詞は UniDic の品詞情報をもとに四種に統合した。表４を見ると、名詞-一般の漢語が最も多く、動詞・形容詞・形状詞・副詞には和語や混種語も多くなっている。

表４　『太陽コーパス』でレベル a、「図書館書籍サブコーパス」でレベル e の語彙

品詞	和語	漢語	外来語	混種語
名詞-一般	謂（いい）、魚（うお）、件（くだん）、差し支え、灯し火、一つ	医、位地、一円、一斑、英仏、各省、気球、旭日、現時、効、公債、工兵、国運、国人、三位、爾後、時々、寺内、授、償金、諸種、正貨、政略、責、智、智識、勅令、体（てい）、逓信、徳義、土人、内国、博文、万人、弊、俸給、本邦、命、約、列国	耶蘇	
名詞-サ変等可能		円満、協商、建議、出入、兌換、騰貴		
動詞・形容詞・形状詞・副詞	豈、如何で、怒（いか）る、え、阿る、如此し、希（こいねが）う、然（さ）り、須らく、宜（むべ）、縦し、悪（わろ）い	爾来、超然、畢竟		合（がっ）する、記する、毫も、失する、製する、着する、徴する、変ずる、弁ずる
接辞		該、口、主		

3.2　近代は周辺的レベルで現代は基本的レベルの語彙

次に、『太陽コーパス』でレベル e で最も周辺的なところにあったもので、「図書館書籍」（固定長）では最も基本的なレベル a にある語彙について、一覧にしてみよう（表 5）。表５では、名詞-一般と名詞-サ変等可能の漢語と外来語が特に多くなっており、名詞-一般や動詞・形容詞・形状詞・副詞では和語も比較的多くなっている。

[1] 『太陽コーパス』でレベル a で、「図書館サブコーパス」で全く出現しない語彙を見ることも必要だが、本研究に用いた現代語用の UniDic と近代語用の UniDic とでは、語彙素や品詞の認定が異なる場合があり、機械処理による語彙の対照だけでは、十分な照合ができない部分が残るため、ここではそのタイプは扱わなかった。

表５　『太陽コーパス』でレベル e、「図書館書籍サブコーパス」でレベル a の語彙

品詞	和語	漢語	外来語	混種語
名詞－一般	生き物、兎、餌、親指、方々、側、切っ掛け、気配、塵（ごみ）、汁、食べ物、玉葱、所、主、鼠、初（はつ）、湖	一連、衛星、映像、円、課題、観点、気温、機構、基地、基盤、業績、芸能、現地、高校、次元、視点、時点、集落、衝撃、世代、体制、土（ど）、濃度、便（びん）、複数、米軍、訳（やく）、要因、理念、路線	エンジン、カー、カード、カメラ、カレー、クラス、クリーム、グループ、ケーキ、コース、コスト、サイド、サラリーマン、シーン、ショー、スクール、スタイル、スピード、スペース、タクシー、チーズ、チャンス、テープ、テーマ、トマト、ドラマ、ニュース、パワー、ビジネス、フィルム、プラン、プロセス、ママ、メンバー、リズム、ルール、レストラン、レベル、ワイン	大勢（おおぜい）、生地（きじ）、地元、職場、夕食
名詞－サ変等可能	幾ら、子育て、そう、生（なま）	移行、依存、運営、活性、規制、強調、結成、研修、検討、構想、構築、個別。志向、取材、出演、出土、制作、正常、駐車、直前、定着、定年、展示、統合、導入、入手、入力、配慮、飛行、普段、優先、予感、冷蔵	アップ、オーケー、カット、コントロール、ショック、スイッチ、スタート、セット、ソフト、チェック、デザイン、テスト、バック、バランス、ブルー、プラス、プレゼント、マイナス、ラン	
動詞・形容詞・形状詞・副詞・感動詞	否（いや）、思い込む、組み合わせる、差し出す、そろそろ、辿り着く、次々、取り組む、入（はい）り込む、放る、恵まれる、さす、よし	一体、公的、直接、不可欠		
接辞	形（なり）、焼き	刊、手、人（じん）、帯、道		

文　献

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

近藤明日子・小木曽智信（2009）「語種を観点とした近代語と現代語の語彙の比較―形態素解析辞書「近代文語 UniDic」「UniDic」を用いて―」言語処理学会第 15 回年次大会

田中牧郎（2010）「雑誌コーパスでとらえる明治・大正期の漢語の変動」（『国際学術研究集会・漢字漢語研究の新次元 予稿集』）

田中牧郎・近藤明日子（2011）「BCCWJ 主要コーパス語彙表」田中・相澤ほか（2011 所収）

田中牧郎・相澤正夫・斎藤達哉・棚橋尚子・近藤明日子・河内昭浩・鈴木一史・平山允子（2011）『言語政策に役立つ、コーパスを用いた語彙表・漢字表の作成と活用』特定領域「日本語コーパス」言語政策班成果報告書

通時コーパス用『中納言』:
Web ベースの古典語コンコーダンサー

小木曽 智信（国立国語研究所言語資源研究系）†
中村 壮範（マンパワージャパン株式会社）

Chunagon for the NINJAL Diachronic Corpus: a Web-based Concordancer of Classical Japanese

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Takenori Nakamura (Manpower Japan Co., Ltd.)

１．はじめに

国立国語研究所の共同研究プロジェクト「通時コーパスの設計」[1]では、日本語の歴史的資料をコーパス化するための研究が行われている。その一環として、先行して整備が進んでいる一部のデータを格納した Web アプリケーション・通時コーパス用『中納言』の共同研究者向けの公開を開始した。通時コーパス用『中納言』は『現代日本語書き言葉均衡コーパス』（以下 BCCWJ とする）の公開にあたって開発された Web ベースのコンコーダンサー『中納言』に若干の機能拡張を行い、『源氏物語』などの通時コーパスの一部のデータを格納したものである。これにより、一般の古典研究者にも使いやすいインターフェイスを用いて通時コーパスを利用することが可能になった。本発表では通時コーパス用『中納言』と現在利用可能なデータについて紹介する。

２．通時コーパス用『中納言』の概要

『中納言』はコーパスに付与された形態論情報を用いて高度な検索を行うことが可能な Web アプリケーションである。検索条件指定の自由度が高く、複数の語を組み合わせて、詳細な条件指定を行うことができる。検索結果はキーとなる語の形態論情報、サンプルの書誌情報とともに KWIC 形式で一覧表示されるほか、表形式のテキストデータとしてダウンロードして利用することもできる。

今回、準備中の通時コーパスデータをこのシステムに格納して利用することを可能にした（次ページ図１）。BCCWJ では「短単位」と「長単位」の二つの異なるサイズの形態論情報を用いることができたが、通時コーパスのデータでは今のところ長単位の整備が進んでいないため、通時コーパス用『中納言』では短単位だけが利用可能となっている。

通時コーパス用『中納言』では、上記『中納言』の機能に加えて BCCWJ にはなかったテキストに関する情報が利用できるようになっている。一つは「本文種別」と呼んでいる情報で、小学館新編古典文学全集に付けられている情報を元に、「会話」「手紙」「歌」「詞書」とそれ以外（地の文）の別が、個々の語について付与されている。さらに「話者情報」として、「会話」についてはその話者、「手紙」については書き手、「歌」については歌番号などの情報が付与されている。

† togiso@ninjal.ac.jp

[1] http://historicalcorpus.jp

図 1　通時コーパス用「中納言」検索実行画面

３．収録データ

現在、通時コーパス用『中納言』に格納され利用可能になっているデータは表 1 に示した 13 作品、約 87 万語である。これらの作品は、後述する「中古和文 UniDic」による自動形態素解析結果をもとに、すべて一度は人手による修正・チェックを経たものである（ただし、一部データの抜き取り調査によると現時点での精度はおおむね 98%程度であると思われる）。国語研究所の通時コーパスは現在のところ設計の途上であり、ごく一部のデータが作成されているに過ぎない。それでも、『源氏物語』をはじめとする中古の主要な古典文学作品をカバーしている。

このうち『竹取物語』『伊勢物語』『土佐日記』『大和物語』『枕草子』『源氏物語』の 6 作品は、小学館の新編日本古典文学全集に基づくデータであり、最終的な通時コーパスにも同じテキストが用いられる予定である。残る 7 作品は、入手しやすいデータをもとに形態素解析の学習用データとしたものや、研究・試験用に作成したものであり、参考データにとどまる。

表 1　収録データ（2012 年 6 月 30 日現在）

作品名	語数（短単位）	備考
竹取物語	12583	小学館・新編日本古典文学全集
伊勢物語	15900	
土佐日記	8113	
大和物語	26733	
枕草子	79879	
源氏物語	510714	
古今仮名序	3107	その他 （「中古和文 UniDic」学習用データ）
紫式部日記	20346	
大鏡	82796	
更級日記	16652	
方丈記	4191	
徒然草	41675	
恋路ゆかしき大将	44819	
計	*867508*	

４．形態論情報

通時コーパスのデータは、BCCWJ と同様に、形態素解析技術を用いて全ての本文テキストに単語の切れ目、読み、品詞、活用などの形態論情報を付与している。形態素解析のための辞書は、BCCWJ の構築に用いられた「UniDic」をもとに、中古和文を解析できるように語彙を増補しパラメータを調整した「中古和文 UniDic」を用いている。「中古和文 UniDic」は、未知語のないテキストであればおおむね 96～97%程度の精度で解析を行うことが可能になっている[2]。

「中古和文 UniDic」が付与する形態論情報は、BCCWJ と同様の「短単位」を採用し、中古語であっても現代語とできるかぎり基準を揃え、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。この中古和文用の短単位の規定は、小椋・須永（2012）にまとめられている[3]。

通時コーパス用『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。

５．検索方法

『中納言』に格納されているデータ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「給う」（終止形）を指定することで「給う」「給は」「給ひ」「給ふ」などの各活用形を一括で検索することが可能である。また、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。

[2]「中古和文 UniDic」は次のサイトで報告書 PDF と共に一般公開している（無償）。
http://www2.ninjal.ac.jp/lrc/index.php?UniDic

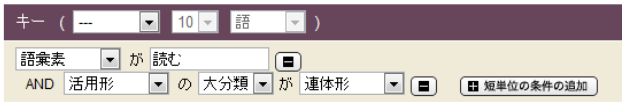
[3]『中古和文 UniDic 短単位規程集』の PDF ファイルも上記ウェブサイトで一般公開している。

５．１　検索条件の指定

具体的には、形態論情報を使った検索では、次のコントロールで検索条件を設定する。

「選択」で条件指定する属性（「語彙素」「出現書字形」など）を選び、右の空欄でその中身を指定する。

「短単位の条件の追加」ボタンで一つの単位について詳細な条件指定を追加できる。次の例では、語彙素が「読む」でかつ、活用形が「連体形」の例を検索している。（活用形など選択肢が決まっているものはドロップダウンメニューから選択する）



さらに、複数の単位を組み合わせて指定することもできる。「前方共起条件の追加」ボタンでキーの前方に出現する単位を指定、「後方共起条件の追加」ボタンでキーの後方に出現する単位を指定する。共起条件は前方後方合わせて最大 10 個まで追加できる。

共起位置は、「キーから」または「文頭から」を基準として、「n 語」または「n 語以内」のように、細かく指定する事ができる。次の例は、「美しい」の連体形の直後（後方 1 語）に来る名詞を検索したものである。

『中納言』では、形態論情報を使った検索以外に「文字列検索」によって表層の文字列にもとづく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で形態論情報を確認すると便利である。

なお、このようにして画面上で指定した検索条件は、システムが解釈できる「検索条件式」に変換されたのち、検索が実行される。この検索条件式は、検索履歴として自動的でサーバー上に記録されるほか、画面上で編集をして再検索に利用することが可能になっている。

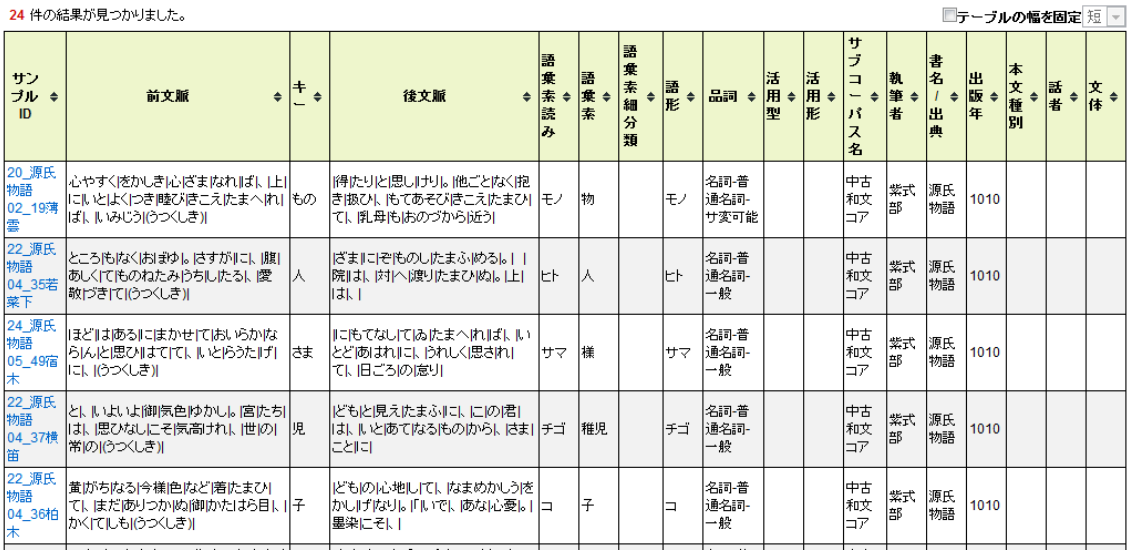
たとえば上記の前方共起 1 を利用した検索例は、次の検索条件式で表される。

```
キー: 品詞 LIKE "名詞%" AND 前方共起: (語彙素 = "美しい" AND 活用形 LIKE "連体形%") ON 1 WORDS FROM キー IN core="true" OR core="false" WITH OPTIONS unit="1" AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"
```

この条件式を控えておくことにより、『中納言』のユーザーであれば全く同じ検索を再現することができる。

５．２　検索実行とダウンロード

検索の実行には画面上の「検索」ボタンをクリックする。これにより，画面下部に図 2 のような検索結果が表示される。ただし、画面上に表示される用例数は 500 例までとなっている。これ以上の用例を確認する場合には結果をダウンロードする必要がある。



24 件の結果が見つかりました。　テーブルの幅を固定 短

サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	語彙素細分類	語形	品詞	活用型	活用形	サブコーパス名	執筆者	書名／出典	出版年	本文種別	話者	文体
20_源氏物語 02_19薄雲	心やすく|をかしき|心|ざまなれ|ば|、上|に|い|と|よく|つき|睦び|きこえ|たまへ|れ|ば|、|いみじう|(うつくしき)|	もの	|物		モノ	名詞-普通名詞-サ変可能			中古和文コア	紫式部	源氏物語	1010			
22_源氏物語 04_35若菜下	ところ|もなく|おほゆ|。|さすが|に、|腹|あしく|て|ものねたみ|うち|し|たる|、|愛敬づき|て|(うつくしき)|	人	|ざま|に|ぞ|ものしたまふ|める|。|　|院|は、|対|へ|渡りたまひぬ|。上|は、|	ヒト	人		ヒト	名詞-普通名詞-一般			中古和文コア	紫式部	源氏物語	1010			
24_源氏物語 05_49宿木	ほど|は|ある|に|まかせ|て|おいらかな|ら|ん|と|思ひ|はて|て|、|いと|らうた|げ|に、|(うつくしき)|	さま	|に|も|て|な|し|て|ゐ|たまへ|れ|ば|、|いと|ど|あはれ|に、|うれしく|思さ|れ|て、旧ごろ|の|恋り|	サマ	様		サマ	名詞-普通名詞-一般			中古和文コア	紫式部	源氏物語	1010			
22_源氏物語 04_37横笛	と、|いよいよ|御|気色|ゆかし。|宮たち|は、|思ひなし|に|そ|気高け|れ、|世|の| 常|の|(うつくしき)|	児	|ども|と|見え|たまふ|に、|に|の|君|は、|いと|あてなる|ものから、|はま|ことに|	チゴ	稚児		チゴ	名詞-普通名詞-一般			中古和文コア	紫式部	源氏物語	1010			
22_源氏物語 04_36柏木	黄|がち|なる|今様|色など|着たまひ|て、|まだ|ありつかぬ|御かたはら目|、|かく|て|し|も|(うつくしき)|	子	|ども|の|心地し|て|、なまめかしう|を|かし|げ|なり。|「|い|で|、|あな|心憂。| 墨染|に|そ|、|	コ	子		コ	名詞-普通名詞-一般			中古和文コア	紫式部	源氏物語	1010			

図 2　検索結果の表示

この検索結果の各列は、【列の表示】のチェックボックスにより表示の ON／OFF を切り替えることができる。検索結果のサンプル ID をクリックすることにより、当該位置周辺（前後 30 単位）に付けられている形態論情報を確認することができる（図 3）。

close

18_枕草子	138570	。		。		補助記号-句点				記号	。	中古和文	1		1	22861.9
18_枕草子	138580	いみじう	イミジイ	いみじい		形容詞-一般	文語形容詞-シク	連用形-ウ音便	イミジュー	和	いみじう	中古和文	1		1	22865.6
18_枕草子	138590	うつくしき	ウツクシイ	美しい		形容詞-一般	文語形容詞-シク	連体形-一般	ウツクシキ	和	うつくしき	中古和文	1		1	22870.5
18_枕草子	138600	ちご	チゴ	稚児		名詞-普通名詞-一般			チゴ	和	ちご	中古和文	1		1	22872.8
18_枕草子	138610	の	ノ	の		助詞-格助詞			ノ	和	の	中古和文	1		1	22873.9
18_枕草子	138620	いちご	イチゴ	苺		名詞-普通名詞-一般			イチゴ	和	いちご	中古和文	1		1	22876.7
18_枕草子	138630	など	ナド	など		助詞-副助詞			ナド	和	など	中古和文	1		1	22878.8
18_枕草子	138640	食ひ	クウ	食う		動詞-一般	文語四段-ハ行	連用形-一般	クイ	和	食ひ	中古和文	1		1	22880.8
18_枕草子	138650	たる	タリ	たり	完了	助動詞	文語助動詞-タリ-完了	連体形-一般	タル	和	たる	中古和文	1		1	22882.8
18_枕草子	138660	。		。		補助記号-句点				記号	。	中古和文	1		1	22883.9

図 3　キー周辺の形態論情報の表示

「検索」ボタンの代わりに「検索結果をダウンロード」ボタンをクリックすることによって、検索結果と検索条件式をテキストデータとしてダウンロードすることができる。データは zip 圧縮されており、アーカイブ中のファイル kwic.txt が検索結果のデータ（タブ区切りの表形式テキスト、文字コードは UTF-8）、summary.txt が検索条件式のデータとなっている。

６．検索例

検索例として、完了の助動詞「つ」「ぬ」に上接する動詞のリストを検索する例を示す（図4）。ここでは、助動詞の前方 2 語以内に来る動詞をキーとして検索結果を取得している。

図 4　助動詞「つ」の上接動詞の検索条件指定

この条件での検索結果の表示は図 5 のようになる。

2485 件の結果が見つかりました。そのうち 500 件を表示しています。　テーブルの幅を固定 短

サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	語彙素細分類	語形	品詞	活用型	活用形	サブコーパス名	執筆者	書名／出典	出版年	本文種別	話者	文体
1203_大和物語	て|山|に|入り|に|け り|。|さて|おこせたり|ける|。|から く|して|思ひわする|恋し|さ|を|うたて|	鳴き	|(つる)|うぐひす|の|声|恋し|、|さて|も|君|わすれ|け り|かし|うぐひす|の|鳴く|を り|のみ|や|思ひ|づ|べき|	ナク	鳴く		ナク	動詞-一般	文語四段-カ行	連用形-一般	中古和文コア		大和物語	951	歌		
1101_古今和歌集	とぶらは|む|もみぢ葉の|散り|て|つもれ|る|わが|やど|に|たれ|を|まつ虫|にこそ|鳴く|らむ|ひぐらし|の|	鳴き	|(つる)|な へ|に|日|は|隔れ|ぬ|と|思へ|ば|山|の|陰|に|ぞ|あり|ける|ひぐらし|の|鳴く|	ナク	鳴く		ナク	動詞-一般	文語四段-カ行	連用形-一般	中古和文コア	(一二七人)	古今和歌集	906	歌	歌番号(204)	
22_源氏物語 04_35若菜下	し|て|、|言は|む|方|なく|おぼゆ|。| 年ごろ|、|まめ事|に|も|あだ事|に|も|召しまつはし|、|参り|	馴れ	|(つ)|も の|を|、|人|より|は|にまやかに|思しとどめたる|御|気色|の|あはれ|になつかしき|を|、|あさまし く|	ナレル	慣れる		ナル	動詞-非自立可能	文語下二段-ラ行	連用形-一般	中古和文コア	紫式部	源氏物語	1010			
23_源氏物語 06_52蜻蛉	も|知ら|ね|ば|、|身|を|投げ|たまへ|らん|とも|思ひ|も|寄ら|ず|、|鬼|や|	食ひ	|(つ)|らん|、|狐|め く|もの|や|とり|もて|去ぬ|らん|、|いと|昔|物語|の|あやしき|もの|の|事|	クウ	食う		クウ	動詞-一般	文語四段-ハ行	連用形-一般	中古和文コア	紫式部	源氏物語	1010			
20_源氏物語 02_13明石	い かで|思ふ|さま|に|て|見|たてまつら|む|と|、|年月|を|頼み過ぐし|、| 今|や|思ひかなふ|と|にそ|	頼み	|き こえ|(つれ)|、|心苦しき|こと|を|も|、|もの|の|は じめ|に|みる|かな|と|嘆く|を|見る|に|	タノム	頼む		タノム	動詞-一般	文語四段-マ行	連用形-一般	中古和文コア	紫式部	源氏物語	1010	会話		
22_源氏物語 04_39夕霧	つくり|たまふ|ほど|と|、|三条|殿|、|限りな|め り|と|、|さしも|や|は|と|にそ|かつ|は|	頼み	|(つれ)|、|まめ人|の|心|変る|は|なごりなく|なむ|と|聞き|し|は|まことな り|け り|、|と|世|を|	タノム	頼む		タノム	動詞-一般	文語四段-マ行	連用形-一般	中古和文コア	紫式部	源氏物語	1010			

図 5　助動詞「つ」の上接動詞の検索結果

上記の画面は「つ」の上接動詞を検索するものだが、検索条件式中で括弧と OR 演算子を用いることで二つの助動詞を一度に検索するようにすることもできる。

```
キー: 品詞 LIKE "動詞%" AND 後方共起: ((語彙素 = "つ" OR 語彙素 = "ぬ" ) AND 品詞 LIKE "助動詞%") WITHIN 2 WORDS FROM キー IN core="true" OR core="false" WITH OPTIONS unit="1" AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"
```

６．１　検索結果の集計

このような検索によって取得したデータをダウンロードすることで、アプリケーションに読み込んでさまざまな処理を施し、集計やデータの分析に利用することができる。たとえば、Microsoft Excel のピボットテーブル機能を利用することで、高度な集計処理を容易に実現することができる。

図 6 は、上記の条件で検索した助動詞「つ」と「ぬ」の上接動詞のリストを Excel に読み込んで助動詞の情報を加えたものである。

<table>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th><th>F</th><th>G</th><th>H</th><th>I</th><th>J</th><th>K</th><th>L</th><th>M</th></tr>
<tr><td>1</td><td>助動詞</td><td>サンプルID</td><td>連番</td><td>前文脈</td><td>キー</td><td>後文脈</td><td>語彙素読み</td><td>語彙素</td><td>語彙素細分</td><td>語形</td><td>品詞</td><td>活用型</td><td>活用形</td></tr>
<tr><td>2</td><td>つ</td><td>24_源氏物語05_49宿木</td><td>31810</td><td>人|を|むなしく|見なし|きこえ|</td><td>たまう</td><td>|(て)|し|後|思ふ||に|は|、|帝|の</td><td>タマウ</td><td>給う</td><td>尊敬</td><td>タマウ</td><td>動詞-非</td><td>文語四段-</td><td>連用形-ウ</td></tr>
<tr><td>3</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>94060</td><td>時|たち離れ|たてまつら|ず|</td><td>馴れ</td><td>|きこえ|(つる)|人||に||にはか||に</td><td>ナレル</td><td>慣れる</td><td></td><td>ナル</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>4</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>94070</td><td>ち離れ|たてまつら|ず|馴れ|</td><td>きこえ</td><td>|(つる)|人||に||にはか||に|別れ|</td><td>キコエル</td><td>聞こえる</td><td></td><td>キコユ</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>5</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>4100</td><td>|ため||に||は|、|いま|見出で|</td><td>たまひ</td><td>|(て)|む|と|思ふ|も|口惜し|や|。</td><td>タマウ</td><td>給う</td><td>尊敬</td><td>タマウ</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>6</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>7840</td><td>、|「|病||に|より|て|位|を|返し|</td><td>たてまつり</td><td>|(て)|し|を|、|いよいよ|老|の|</td><td>タテマツル</td><td>奉る</td><td></td><td>タテマツル</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>7</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>75020</td><td>、|阿闍梨|ものせよ|と|言ひ|</td><td>やり</td><td>|(つる)||は|」|と|のたまふ||に|、|</td><td>ヤル</td><td>遣る</td><td></td><td>ヤル</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>8</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>22570</td><td>|出で立ち|て|、|夢|の|心地|</td><td>し</td><td>|(つる)|嘆き|も|さめ||に||けり|。|</td><td>スル</td><td>為る</td><td></td><td>ス</td><td>動詞-非</td><td>文語サ行変</td><td>連用形-一</td></tr>
<tr><td>9</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>39170</td><td>五節|を|思し忘れ|ず|、|また|</td><td>見</td><td>|(て)|しが|な|と|心||に|かけ|た</td><td>ミル</td><td>見る</td><td></td><td>ミル</td><td>動詞-非</td><td>文語上一段</td><td>連用形-一</td></tr>
<tr><td>10</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>53640</td><td>|近き|惟光|うけたまはり|や|</td><td>し</td><td>|(つ)|らむ|、|さる|召し|も|や|と|</td><td>スル</td><td>為る</td><td></td><td>ス</td><td>動詞-非</td><td>文語サ行変</td><td>連用形-一</td></tr>
<tr><td>11</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>75410</td><td>|ものせ|させ|たまふ|こと|や|</td><td>はべり</td><td>|(つ)|らん|」|、|「|さる|こと|も|な</td><td>ハベル</td><td>侍る</td><td></td><td>ハベリ</td><td>動詞-非</td><td>文語ラ行変</td><td>連用形-一</td></tr>
<tr><td>12</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>60460</td><td>も|聞こえあはせ|人||に|思ひ|</td><td>きこえ</td><td>|(つる)|を|、|かく|思しなり||に|</td><td>キコエル</td><td>聞こえる</td><td></td><td>キコユ</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>13</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>62060</td><td>|たてまつら|む|と|こそ|思ひ|</td><td>たまへ</td><td>|(つれ)|」|とて|も|、|消え入り|</td><td>タマウ</td><td>給う</td><td>謙譲</td><td>タマウ</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>14</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>74610</td><td>いかで|さやか||に|御|容貌|を|</td><td>見</td><td>|(て)|しが|な|、|と|思す|も|、|う</td><td>ミル</td><td>見る</td><td></td><td>ミル</td><td>動詞-非</td><td>文語上一段</td><td>連用形-一</td></tr>
<tr><td>15</td><td>つ</td><td>20_源氏物語02_14澪標</td><td>82360</td><td>ら|ぬ|やう||に|て|ここ||に|渡し|</td><td>たてまつり</td><td>|(て)|む|と|思す|。|女|君||に|も|</td><td>タテマツル</td><td>奉る</td><td></td><td>タテマツル</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>16</td><td>つ</td><td>21_源氏物語03_24胡蝶</td><td>20000</td><td>、|父|大臣||に|も|知ら|せ|や|</td><td>し</td><td>|(て)|まし|など|、|思しよる|をり</td><td>スル</td><td>為る</td><td></td><td>ス</td><td>動詞-非</td><td>文語サ行変</td><td>連用形-一</td></tr>
<tr><td>17</td><td>つ</td><td>21_源氏物語03_24胡蝶</td><td>35950</td><td>し|の||ほど|も|、|見あらはし|</td><td>はて</td><td>|たまひ|(て)|む|や|」|など|、|う</td><td>ハテル</td><td>果てる</td><td></td><td>ハツ</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>18</td><td>つ</td><td>21_源氏物語03_24胡蝶</td><td>35960</td><td>|ほど|も|、|見あらはし|はて|</td><td>たまひ</td><td>|(て)|む|や|」|など|、|うち|語ら</td><td>タマウ</td><td>給う</td><td>尊敬</td><td>タマウ</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>19</td><td>つ</td><td>21_源氏物語03_24胡蝶</td><td>39350</td><td>|思しより|て|、|「|もの|の|心|</td><td>得</td><td>|(つ)|べく||は|ものし|たまふ|め</td><td>エル</td><td>得る</td><td></td><td>ウ</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>20</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>61530</td><td>弱く|て|、|昼|も|空|を|のみ|</td><td>見</td><td>|(つる)|もの|を|、|いとほし|と|</td><td>ミル</td><td>見る</td><td></td><td>ミル</td><td>動詞-非</td><td>文語上一段</td><td>連用形-一</td></tr>
<tr><td>21</td><td>つ</td><td>21_源氏物語03_24胡蝶</td><td>43350</td><td>たき||に|、|慰む|こと|なく|て|</td><td>過ぎ</td><td>|(つる)|年ごろ|を|、|かく|て|見|</td><td>スギル</td><td>過ぎる</td><td></td><td>スグ</td><td>動詞-非</td><td>文語上二段</td><td>連用形-一</td></tr>
<tr><td>22</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>530</td><td>|本意|深き|を|、|后の宮|の|</td><td>おはしまし</td><td>|(つる)||ほど||は|、|よろづ|憚り|</td><td>オワシマス</td><td>おわします</td><td></td><td>オワシマス</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>23</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>8590</td><td>々|なやま|せ|たまふ|こと|も|</td><td>あり</td><td>|(つれ)|ど|、|いと|かく|うち||は</td><td>アル</td><td>有る</td><td></td><td>アリ</td><td>動詞-非</td><td>文語ラ行変</td><td>連用形-一</td></tr>
<tr><td>24</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>8940</td><td>わし|たり|。|御|位|を|去ら|せ|</td><td>たまひ</td><td>|(つれ)|ど|、|なほ|そ|の|世||に|</td><td>タマウ</td><td>給う</td><td>尊敬</td><td>タマウ</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>25</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>12280</td><td>御|遺言|違へ|ず|仕うまつり|</td><td>おき</td><td>|(て)|しか||ば|、|かく|末|の|世|</td><td>オク</td><td>置く</td><td></td><td>オク</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>26</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>80090</td><td>れ||ば|、|など|て|乗り添ひ|て|</td><td>行か</td><td>|ざり|(つ)|らん|、|生きかへり|た</td><td>イク</td><td>行く</td><td></td><td>イク</td><td>動詞-非</td><td>文語四段-</td><td>未然形-一</td></tr>
<tr><td>27</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>21190</td><td>ら|ん|こと|をば|見|隠し|教へ|</td><td>きこえ</td><td>|(つ)|べから|む|人|の|うしろや</td><td>キコエル</td><td>聞こえる</td><td></td><td>キコユ</td><td>動詞-非</td><td>文語下二段</td><td>連用形-一</td></tr>
<tr><td>28</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>22340</td><td>権|中|納言|の|朝臣|の|独り|</td><td>あり</td><td>|(つる)||ほど||に|、|うち|かすめ|</td><td>アル</td><td>有る</td><td></td><td>アリ</td><td>動詞-非</td><td>文語ラ行変</td><td>連用形-一</td></tr>
<tr><td>29</td><td>つ</td><td>20_源氏物語01_04夕顔</td><td>12410</td><td>し|め|たそかれ||に||ほのぼの|</td><td>見</td><td>|(つる)|花|の|夕顔|あり|つる|</td><td>ミル</td><td>見る</td><td></td><td>ミル</td><td>動詞-非</td><td>文語上一段</td><td>連用形-一</td></tr>
<tr><td>30</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>31340</td><td>|かならず|うけひき|申さ|せ|</td><td>たまひ</td><td>|(て)|む|、|年ごろ|の|御|本意|</td><td>タマウ</td><td>給う</td><td>尊敬</td><td>タマウ</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>31</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>32830</td><td>|世|の|中|を|御|心|と|過ぐし|</td><td>たまひ</td><td>|(つ)|べき|も|おはします|べか|</td><td>タマウ</td><td>給う</td><td>尊敬</td><td>タマウ</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>32</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>45030</td><td>させ|たまへ|り|し|御|気色|を|</td><td>見</td><td>|たてまつり|(て)|しか||ば|、|お</td><td>ミル</td><td>見る</td><td></td><td>ミル</td><td>動詞-非</td><td>文語上一段</td><td>連用形-一</td></tr>
<tr><td>33</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>45040</td><td>|たまへ|り|し|御|気色|を|見|</td><td>たてまつり</td><td>|(て)|しか||ば|、|おのづから|、|</td><td>タテマツル</td><td>奉る</td><td></td><td>タテマツル</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>34</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>45340</td><td>あらじ|かし|と|心ときめき|も|</td><td>し</td><td>|(つ)|べけれ|ど|、|女|君|の|、|</td><td>スル</td><td>為る</td><td></td><td>ス</td><td>動詞-非</td><td>文語サ行変</td><td>連用形-一</td></tr>
<tr><td>35</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>49210</td><td>と|かく|とりわき|て|聞きおき|</td><td>たてまつり</td><td>|(て)|む|をば|、|こと||に|こそ||は</td><td>タテマツル</td><td>奉る</td><td></td><td>タテマツル</td><td>動詞-非</td><td>文語四段-</td><td>連用形-一</td></tr>
<tr><td>36</td><td>つ</td><td>22_源氏物語04_34若菜上</td><td>68780</td><td>し、|権|中|納言|など|の|独り|</td><td>ものし</td><td>|(つる)||ほど||に|、|進み寄る|べ</td><td>モノス</td><td>ものす</td><td></td><td>モノス</td><td>動詞-非</td><td>文語サ行変</td><td>連用形-一</td></tr>
</table>

図 6　助動詞「つ」の上接動詞の検索結果（一部）

このデータをピボットテーブルで集計することで、助動詞別に高頻度な上接動詞をリストアップすることができる（図 7）。

助動詞	つ		助動詞	ぬ
行ラベル	データの個数 / キー		行ラベル	データの個数
給う	284		成る	726
侍る	123		給う	725
見る	116		侍る	245
為る	107		有る	137
思う	103		出でる	134
有る	90		過ぎる	134
聞こえる	83		止む	108
奉る	70		経る	95
過ぐす	46		失せる	86
言う	44		入る	84
見える	38		果てる	83
思す	31		返る	72
初める	29		来る	65
遣る	27		為る	61
捨てる	26		参る	60
申す	25		立つ	57
宣う	23		渡る	55
果てる	22		暮れる	52
聞く	21		忘れる	44
成す	21		消える	42
来る	19		絶える	41
参る	19		更ける	39
置く	19		知る	38
取る	19		明ける	38
許す	18		寝る	38
得る	15		出で来る	37
侍ふ	13		罷出づ	35
渡る	13		泣く	33
変える	13		亡くなる	32
止す	13		降る	32
留める	13		隠れる	32
渡す	12		思う	31
覚える	12		散る	31
下ろす	12		止まる	29
おわします	12		思す	27
明かす	12		居る	26
暮らす	11		別れる	26
ものす	11		おわす	26
投げる	10		おわします	26
失う	10		上る	26
おわす	10		見える	22

図 7　助動詞「つ」「ぬ」の高頻度の上接動詞（一部）

７．おわりに

検索例で見たように、従来であれば大変な労力と時間を要していた検索・集計作業を、極めて簡単に行うことが可能になった。また、従来では不可能であった高度な組み合わせ検索が可能になった。今後、通時コーパス用『中納言』を用いて、単に研究を省力化するだけでなく、これまで不可能であった新次元の古典語研究がなされ、有益な研究成果が生み出されることに期待したい。

現在、通時コーパス用『中納言』の公開範囲は共同研究者の一部に限定しているが、今後は公開範囲を拡大していく予定である。

文　献

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（http://dl.dropbox.com/u/73297026/report/unidic-EMJ_report2012.pdf 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（http://dl.dropbox.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf 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）

関連 URL

コーパス検索アプリケーション『中納言』（BCCWJ）　http://chunagon.ninjal.ac.jp/

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

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—

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

Towards automatic transformation between different transcript conventions: Aspect of intonation

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、書き言葉コーパスのめざましい発展がある一方、話し言葉コーパスは、大規模なものを開発するには音声収録・転記という初期段階での負担が大きく、独話を中心とする『日本語話し言葉コーパス』(CSJ) を除いて大規模なものはほとんど存在しない。特に、会話コーパスについては、各研究プロジェクトによる小規模なレベルのものしか存在しない。

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問題を解決することを目標として立ち上げられた。既存のコーパスの共有化に際しては、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。

本研究の目的は、CSJ 方式のような言語学志向の転記と、会話分析方式のような相互行為志向の転記との相互変換について検討をおこなうことである。特に、会話分析方式の音調マーカーと相関する言語・音響的な情報について考察をおこなう。具体的には、CSJ 方式と会話分析方式の双方で転記・アノテーションされた会話コーパスを用いて、CSJ 方式の韻律ラベルと会話分析方式の音調マーカーの対応関係を分析する。さらに、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変換をおこなうにあたって必要な言語・音響特徴を検討する。

2. 方法

2.1 データ

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 分をもちいた。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。

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

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（完全版/簡略版）が提供される。X-JToBI では、アクセント句の末尾に句末境界音調が付与される。具体的には、(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%）、(5) 上昇下降上昇調（L%HLH%）の合計 5 種類が認定される。ただし、上昇下降上昇調は本データには出現しなかった。

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 2 名（X 氏、Y 氏）によっておこなわれた。X 氏は chiba0232 を、Y 氏は chiba0432 を

転記した。2012 年現在で、X 氏は約 6 年、Y 氏は約 5 年の会話分析経験を有する。

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おこなっている。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おこなっている。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。

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おこなっている。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。

2.3 言語・音響特徴

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変換するため、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、分析に用いた[*1]。

■言語特徴

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）

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） 長い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） 長い発話単位中で末尾から何番目のアクセント句か（対数値）

[*1] 音響特徴として、アクセント句の平均 F0・句末単語の平均 F0・句末単語の最小 F0・アクセント句の平均パワー・句末単語の平均パワー・句末単語の平均モーラ長も抽出したが、他の特徴との相関が高いため用いなかった。

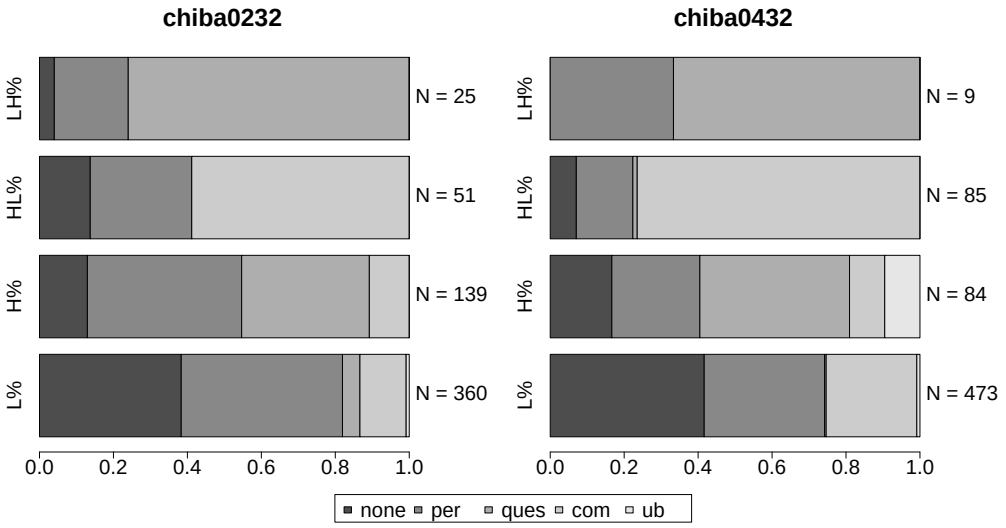


図 3　句末境界音調と音調マーカーの対応

2.4　分析手順

たんなるアクセント句境界に音調マーカーが付与されていることはほとんどなかったため、ピッチレンジのリセットを伴わないアクセント句末の L% は分析対象外とした。まず、分析対象のアクセント句に付与された句末境界音調（L%、H%、HL%、LH%）と音調マーカーとの対応を調べた。次に、（データ数の少ない LH% を除く）句末境界音調ごとに、各言語・音響特徴が音調マーカーによってどのように異なるか調べた。さらに、言語・音響特徴から音調マーカーを予測する多変量モデルを構築し、予測精度と各特徴の貢献度を検討した。

会話分析方式の 2 名の転記者の経験や方略の違いによる影響を検討するため、すべての分析は 2 つのデータ（chiba0232 と chiba0432）に対して別々におこなった。

3.　結果

3.1　韻律ラベルと音調マーカーの対応

CSJ 方式の句末境界音調と会話分析方式の音調マーカーの対応関係を図 3 に示す。

L% のアクセント句は、4 割近くが会話分析方式による音調マーキングがされておらず、残りの 6 割はピリオドまたはコンマが大半を占めていた。また、chiba0232 ではピリオドの割合が多く 4 割を占めていたが、chiba0432 ではピリオドとコンマがそれぞれ 2 割程度であった。H% のアクセント句は、クエスチョンが 4 割近くを占めているが、その他の音調も多く、特に chiba0232 ではピリオドが全体の 4 割を占めていた。HL% のアクセント句は、6〜7 割をコンマが占めており、残りはピリオドの音調か、マーキングがされていないものであった。LH% のアクセント句は総度数が少ないが、およそ 7 割をクエスチョンが占めており、残りの 3 割のほとんどをピリオドが占めていた。

平坦音調を表すアンダーバーは、全データ中、出現数が極めて少なかった。また、L% と

HL% におけるクエスチョン、さらに LH% のアクセント句自体も出現数が少なかった。そのため、これらの事例は以降の分析では対象に含めなかった。

3.2 各言語・音響特徴との関係

■言語特徴 L%、H%、HL% におけるそれぞれの音調マーカーと言語特徴との関係を調べたところ、いくつかの音調マーカーに強く関連する品詞が見られた。

chiba0232 の L% と H% では、末尾単語の品詞（lastPOS）が、音調マーカーなしで終助詞が少なく、その他の助詞が多いという傾向があった。ピリオドではこれと逆の傾向であり、特に H% のピリオドで終助詞の比率が高かった（60% 以上）。HL% でも、ピリオドで終助詞の比率が極めて高く（90% 以上）、一方、コンマでは接続助詞の比率が比較的高かった（40% 以上。他の音調マーカーには出現せず）。また HL% では、次末単語の品詞（penultPOS）でも、助動詞や接続助詞が他の音調マーカーよりも多く出現するという傾向があった。

chiba0432 では、品詞の分布に関する傾向は chiba0232 ほどはっきりとしなかった。ただし、HL% のピリオドで終助詞の比率はやはり高かった（80% 弱）。

■音響特徴 L%、H%、HL% におけるそれぞれの音調マーカーと音響特徴との関係を図 4・5 に示す。

まず L% では、音調マーカーなしのアクセント句は、アクセント句の最小 F0（f0MinAP）が全般的に高く、最終抽出可能 F0 点の位置（lastF0Loc）が句末に近い傾向にあった。ピリオドでは、f0MinAP に加えて、句末単語の最大パワー（f0MaxWord）が低かった。また、lastF0Loc がより早期にあらわれる傾向があった。コンマでは、lastF0Loc がピリオドと音調マーカーなしの中間であった。これらの傾向は chiba0232 と chiba0432 に共通してみられた。

H% では chiba0232 と chiba0432 で違いがみられた。音調マーカーなしのアクセント句に関しては、chiba0232 ではアクセント句の最大 F0 値（f0MaxAP）が高いのに対し、chiba0432 ではむしろ f0MaxAP が低かった。他にも、chiba0232 では句末単語の最大パワー（pwrMaxWord）や最終抽出可能 F0 点の位置（lastF0Loc）に他の音調マーカーとの違いがみられたものの、chiba0432 では同様の傾向は観察されなかった。音調マーカーが付与されているアクセント句に関しては、ピリオドで句末単語の最大 F0（f0MaxWord）が相対的に低い点が両データの共通点として挙げられるが、それ以外で共通する傾向は観察されなかった。

HL% では、chiba0232 の lastF0loc が音調マーカーなしのときにアクセント句のより末尾に位置しているが、chiba0432 では特に大きな差はみられなかった。ピリオドでは、句末単語の最大 F0（f0MaxWord）やアクセント句の平均モーラ長（amdAP）が他の音調マーカーより大きく、この傾向は chiba0432 でより顕著であった。コンマは、アクセント句の最小 F0（f0MinAP）が chiba0432 で低いことが確認された。

■その他の特徴 発話中でのアクセント句の位置に関する特徴を調べたところ、ピリオドやクエスチョンの大半（80〜100%）は発話末のアクセント句であった。

3.3 多変量モデル

ここまで、多くの言語・音響特徴が音調マーカーによって異なることを示した。本節では、これらの言語・音響特徴から音調マーカーを予測する多変量モデルを構築し、予測精度と各特

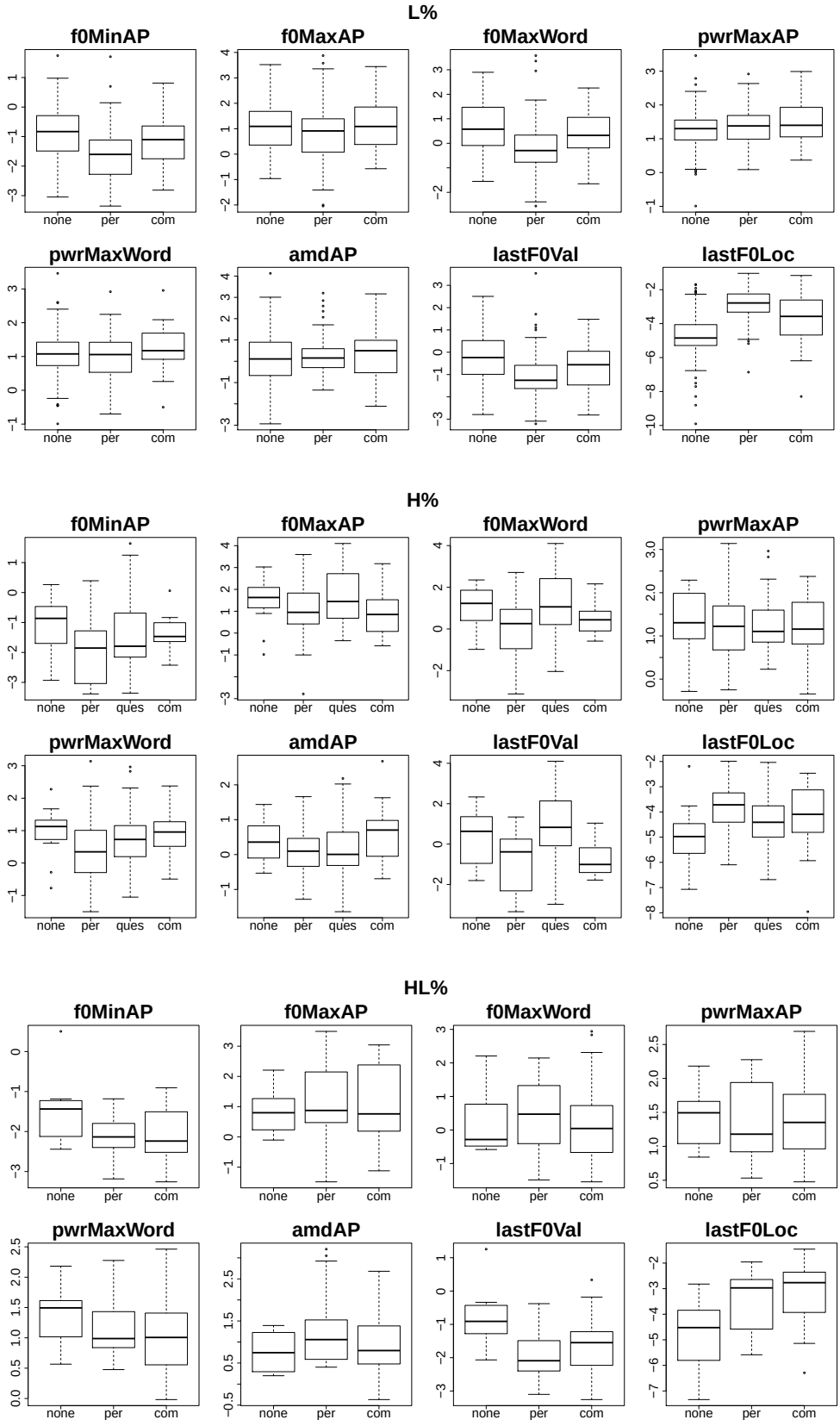


図 4　音調マーカーと音響特徴（chiba0232）

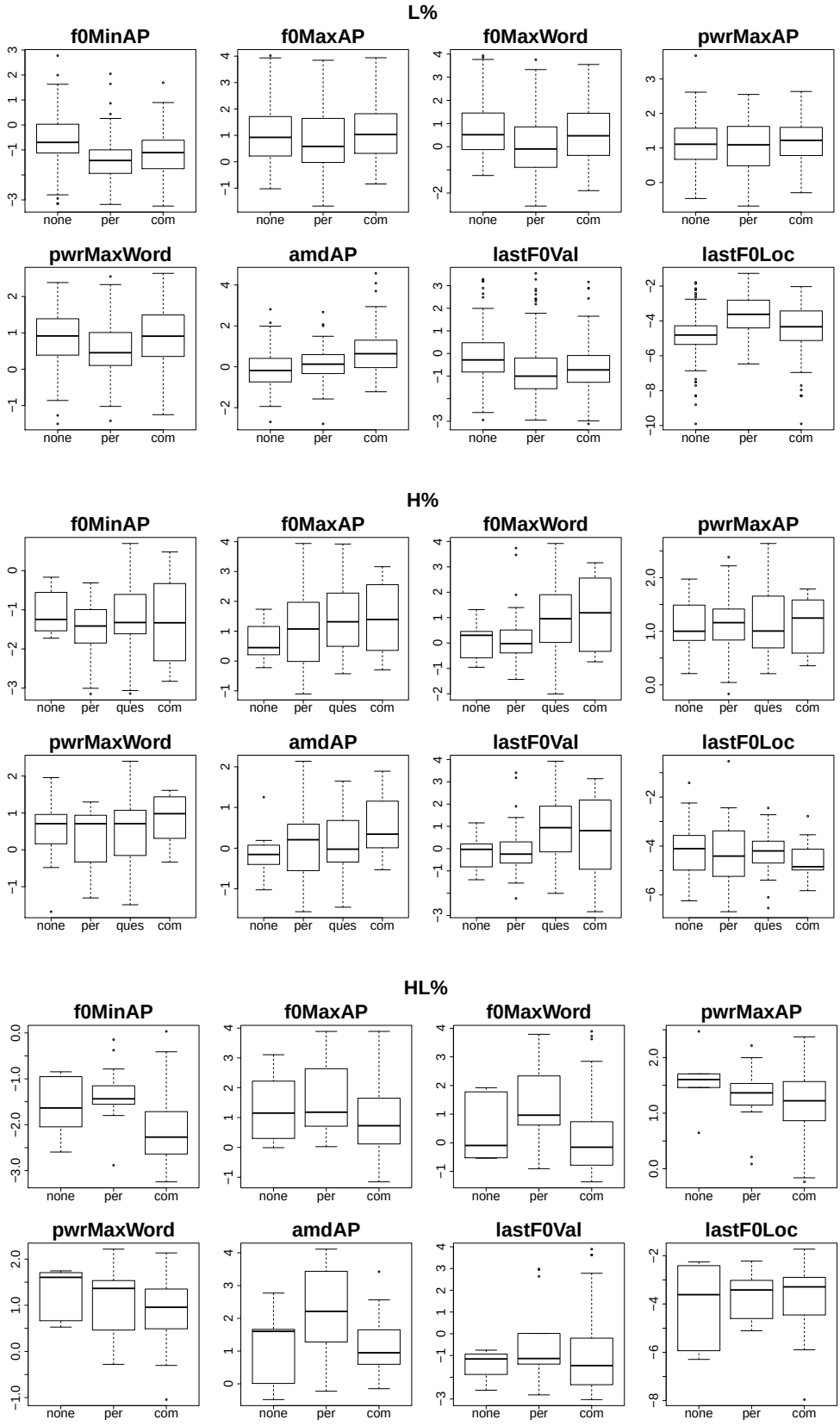


図 5　音調マーカーと音響特徴（chiba0432）

表 1　ランダムフォレスト法による予測精度（OOB 推測精度）（chiba0232）

L%　（正解率 = 76.5%）

	予測値		
観測値	none	per	com
none	**107**	13	1
per	9	**89**	3
com	20	15	**3**

H%　（正解率 = 62.1%）

	予測値			
観測値	none	per	ques	com
none	**8**	1	3	4
per	0	**34**	10	0
ques	0	13	**30**	0
com	5	7	1	**0**

HL%　（正解率 = 58.0%）

	予測値		
観測値	none	per	com
none	**1**	1	5
per	0	**3**	10
com	1	4	**25**

表 2　ランダムフォレスト法による予測精度（OOB 推測精度）（chiba0432）

L%　（正解率 = 73.4%）

	予測値		
観測値	none	per	com
none	**154**	18	17
per	11	**113**	2
com	50	16	**48**

H%　（正解率 = 33.3%）

	予測値			
観測値	none	per	ques	com
none	**2**	3	8	0
per	1	**3**	15	1
ques	4	8	**20**	2
com	1	2	5	**0**

HL%　（正解率 = 83.3%）

	予測値		
観測値	none	per	com
none	**0**	1	5
per	0	**5**	8
com	0	0	**65**

徴の貢献度を検討する。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。

それぞれの会話データと句末境界音調に対する OOB（out-of-bag）推測による予測精度を表 1・2 に示す。chiba0432 の H% に対して予測精度がかなり落ちるが、それ以外は 60〜80% の精度を示している[*2]。

各説明変数の貢献を重要度順に図 6・7 に示す。いずれのデータ・句末境界音調においても発話末尾からの位置（`revLoc`）の貢献が大きい。これはピリオドやクエスチョンの大半が発話末のアクセント句であるためである。それ以外は、データ・句末境界音調ごとに多少異なる。chiba0232 では、末尾単語の品詞（`lastPOS`）や次末単語の品詞（`penultPOS`）といった言語特徴の貢献が総じて高い。これに対して、chiba0432 では、HL% を除いて言語特徴の貢献はあまり大きくなかった。音響特徴では、句末単語の最大 F0（`f0MaxWord`）やアクセント句の最小 F0（`f0MinAP`）が HL% 以外でやや大きな貢献を示した。また、`lastF0Val` や `lastF0Loc` といった最終抽出可能 F0 点に関する特徴も HL% 以外で比較的大きな貢献があった。chiba0432 では、アクセント句の平均モーラ長（`amdAP`）も H% を除いて貢献が高かった。

4. 議論

3 節の結果から、一部の句末境界音調を除いて、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比較的高い精度で変換できることが分かった。いずれのデータについても、多

[*2] ランダムフォレスト法は非線形なモデルなので、句末境界音調ごとに別々にモデルを構築するのではなく、句末境界音調を説明変数に加えてデータ全体を一括してモデル化することも可能である。この方法による OOB 推測精度は、chiba0232 と chiba0432 に対してそれぞれ 72.5% と 71.9% であった。

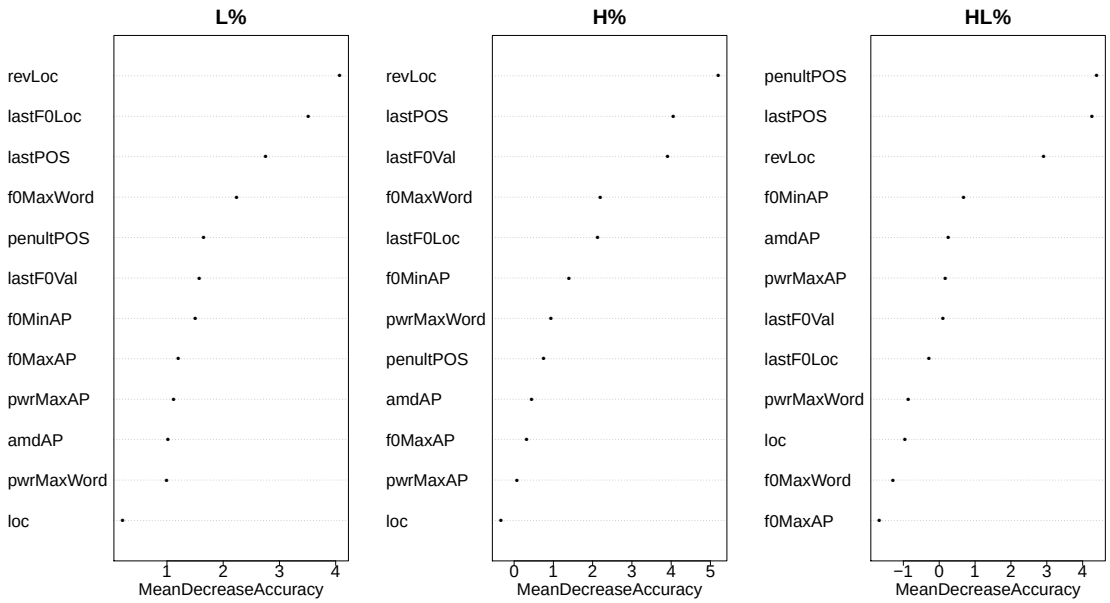


図 6　説明変数の重要度（chiba0232）

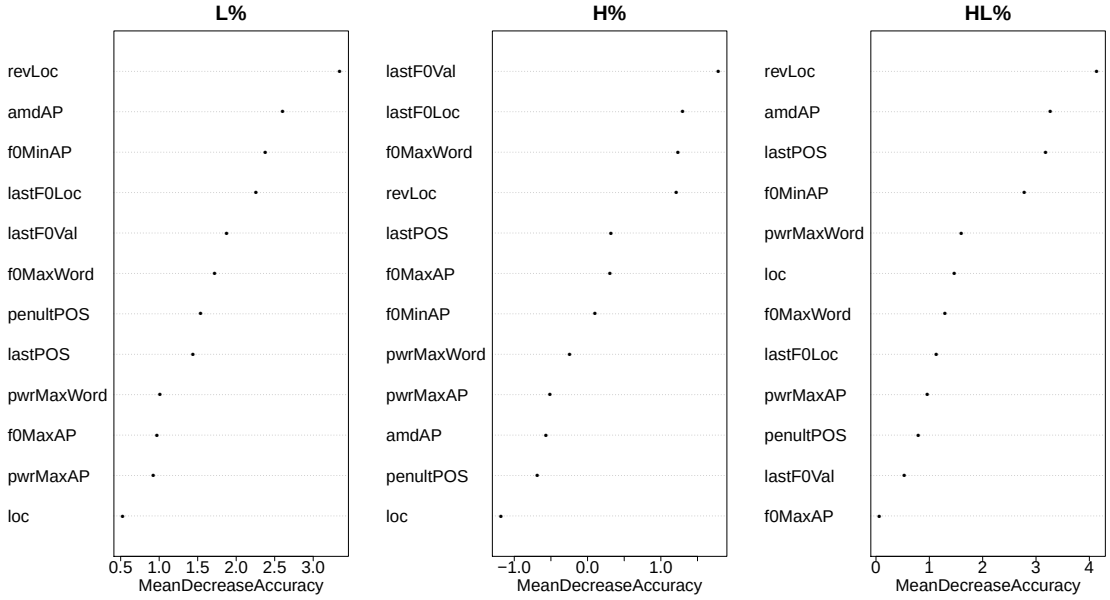


図 7　説明変数の重要度（chiba0432）

変量モデルで貢献の大きな音響特徴は、句末単語の最大 F0 やアクセント句の最小 F0、それに最終抽出可能 F0 点に関する特徴であった。最後の点は特に興味深い。最終抽出可能 F0 点は、アクセント句/句末単語中の F0 の最大値や最小値よりも句末境界音調の特徴をより正確に表しており、たとえば H% では、この F0 値が大きいとクエスチョンになりやすい。よって、この特徴は、X-JToBI 体系では区別できない疑問上昇調と強調上昇調などを区別する手掛かりになるかもしれない。また、この F0 点が早めに生起するということは、句末に無声化母音（「です」

「ます」の末尾の/u/など）が出現したり、パワーの弱まりがあったりすることを示しており、何らかの音調マーカーが付与される場合（発話末であることが多い）の指標となっている。

さらに、多変量モデルに貢献する言語・音響特徴が 2 つの会話データで違いがみられることも分かった。chiba0232 では言語特徴の貢献が総じて高く、一方、chiba0432 では音響特徴の貢献が相対的に高かった。この理由として、両データの話者の言語・音響的な特徴の違いも可能性として考えられるが、転記者が音調マーキングをおこなう際の方略として音響特徴を重視するか、言語特徴を重視するかという点に左右されていると考えられる。

この点を確認するために、chiba0232 を担当した X 氏と chiba0432 を担当した Y 氏に対して事後インタビューをおこない、どのような方略で音調マーキングをおこなったか尋ねた。両者とも韻律のみで判断するよう心掛けていると回答したが、全体的な方略に大きな違いがあった。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 (Chafe 1994, 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。そのため、Y 氏のほうがより音響特徴を重視した音調マーキングになったのではないかと思われる。一方、X 氏は特にそのような方略はとっていなかった。X 氏は会話分析により傾倒しており、韻律に集中しつつも、統語や行為の側面も重視していたのではないかと思われる。今後、転記者の方略の違いにどのように対応するか検討したい。

謝辞　会話分析方式の転記を作成していただいた黒嶋智美・横森大輔の両氏に感謝します。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.

Chafe, Wallace (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: Chicago University Press.

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., 東京.

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.

関連 URL

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

「気持ち」の意味について

加藤恵梨（名古屋大学留学生センター）

A Semantic Analysis of *kimochi*

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

１．はじめに

『現代日本語書き言葉均衡コーパス』をもとに[1]、「気持ち」という語の意味を分析し、語の本質的な意味である意義素を記述することを目指す[2]。また、「気持ち」が複数の意味を有する場合、別義間の関連性について比喩を用いて説明する。さらに、「気持ち」の類義語である「気分」との意味の違いについても考察する。

２．先行研究の記述とその検討

２.１　先行研究の記述

先行研究において、「気持ち」の意味がどのように記述されているのかを概観し、それらの記述を検討する。

森田（1989: 429）は、「気持ち」は「特に物事に接して感ずる心の状態である。したがって、心にある状態を与える対象か場面が条件として存在する」と述べ、「一汗流していい気持ちだ」「気持ちのいい朝」「私に対する気持ちを聞かせてください」「面接試験の前は気持ちが落ち着かない」といった例を挙げている。また、「こちらにそのような気分を催させる対象にも用いる」と述べ、その例として「とても気持ちのいい方ね」「気持ち悪い格好の虫」などを挙げている。さらに、「ある環境や状況・立場に置かれた者としての心の様子にも用いる」と述べ、「患者の気持ちになって介抱しよう」「気持ちを引き締めてかかる」「これからは気持ちを入れ替えて勉強しろ」といった例を挙げている。加えて、「肉体条件に起因する快・不快の感覚も気持ちの問題と考える」と述べ、「船に揺られて気持ちが悪い」「凝っている肩を揉んでもらって、とても気持ちがいい」といった例を挙げている。

次に『講談社類語辞典』は、次の四つの意味を記述している。

- いろいろなことを知覚することによってその人が感じる心の状態。
  「～のよい音楽が流れている」「彼の言動は人をいやな～にさせる」　(p.130)
- ある物事に対して生じる、心の状態や思い。
  「もう少し彼女の～を思いやるべきだ」　(p.214)
- ある物事に対する考えや思い。
  「あんなことを言う彼の～が理解できない」(p.226)

---

[1] 例文の後に、作者と出典のみが記されている引用例は、『現代日本語書き言葉均衡コーパス』から引用したものである。

[2] 意義素とは、「ある語がいろいろの具体的な場面・文脈で示す細かな意味のゆれを取り除いたあとに残る核的な意味のこと」（国広 1997: 12）である。具体的な文脈で用いられた語が、文脈ごとに微妙に意味が異なることは事実である。しかし、国広（1997: 174）が述べているように、「文脈の影響を受けて違って見える語義を細かく追求して行けば、多義はいくらでも数を増す」ことになり、文脈の影響によって違って見える語義を細かく記述するのではなく、語の本質的な意味、すなわち意義素を記述する必要がある。

・心持ち[3]。
「～右に寄ってください」「ねじを～緩める」　（p.1450）

２.２　先行研究の記述の検討

森田と『講談社類語辞典』の記述には大きな違いはないと言える。

森田の記述に「こちらにそのような気分を催させる対象にも用いる」とあることから、「気持ち」は「気分」と意味が類似しているということが確認できる。しかし、「気持ち」の意味を「気分」を介して説明したのでは、「気持ち」の意味、あるいは両語の意味の違いが明らかにされているとは言いがたいため、より明確な記述をする必要があると考えられる。

次に『講談社類語辞典』では、「気持ち」の一つ目の意味（＝「いろいろなことを知覚することによってその人が感じる心の状態」）の例として「彼の言動は人をいやな気持ちにさせる」が挙げられている。しかし、この例は二つ目の意味（＝「ある物事に対して生じる、心の状態や思い」）でも説明できそうであり、一つ目の意味と二つ目の意味記述にどのような違いがあるのかが明確に示されているとは言いがたい。よって、一つ目の意味と二つ目の意味の違いをより明確に記述する必要があると考えられる。

以下では、先行研究の記述とその検討をふまえ、「気持ち」の意味を分析する。

３．比喩について

ある語が多義語[4]である場合、複数の意味の関連性について考察する必要がある。その関連づけを考える際に重要な役割を果たすと考えられるのが、「メタファー」「メトニミー」「シネクドキー」という比喩である。

籾山（2009, 2010）は「メタファー」「メトニミー」「シネクドキー」を次のように定義している。

> メタファー
> ２つの事物・概念の何らかの「類似性（similarity）」に基づいて、本来は一方の事物・概念を表す形式を用いて、他方の事物・概念を表すという比喩　（籾山 2010: 35）

> メトニミー
> ２つの事物の外界における「隣接性（contiguity）」、さらに広く２つの事物・概念の思考内、概念上の「関連性」に基づいて、一方の事物・概念を表す形式を用いて、他方の事物・概念を表す比喩　（籾山 2010: 44）

> シネクドキー
> 本来はより一般的な意味を持つ形式を用いて、より特殊な意味を表す、あるいは逆に、本来はより特殊な意味を持つ形式を用いて、より一般的な意味を表すという比喩
> （籾山 2009：28）

本稿では籾山の「メタファー」「メトニミー」「シネクドキー」の定義に従い、これらの比喩を用いて「気持ち」が有する複数の意味の関連性について考察する。

---

[3] 俗語（日常卑近な話し言葉として用いられ、あらたまった場では用いにくいと感じられる語）であると記されている。

[4] 國廣（1982: 97）は「多義語」について、「『多義語』（polysemic word）とは、同一の音形に、意味的に何らかの関連を持つふたつ以上の意味が結び付いている語を言う」と記述している。

４．意味分析

本稿では、「気持ち」に４つの多義的別義を認め、それぞれの意味を記述する[5]。また、分析の最後に別義間の関連性について述べる。

４．１　別義１：〈身体に受けた刺激によって生じる〉〈心の状態〉

(1) 内視鏡室に入る。まず肩に注射をうたれる。つぎにのどを麻痺させる麻酔薬を口に含むのだが、この麻酔薬がなんともまずくて気持ちが悪い。閉口した。おそらくまちがって飲みこんだりしないようにという配慮なのだろう。
（佐藤宏明『精神病棟の中で』）

(2) ほどよくかゆいところをかく刺激は、だれにとってもとても気持ちのよいものです。そのため、かくという行為につい浸ってしまいたくなります。
（福冨雅康『アトピーはかいて治そう :: 薬もお金も時間もかけない』）

(3) タオルを水で絞って、首と胸を拭う。幸いここでは井戸水を使っているので、水は冷たくて気持ちがいい。　（柴田よしき、宮本陽吉『R-0 amour :: 長編ホラー小説』）

例(1)の文中にある「麻酔薬がなんともまずくて気持ちが悪い」という表現は、まずいと感じるような刺激を舌に受け、そのことによって好ましくない心の状態にあるということを表していると考えられる。よって、例(1)の「気持ち」は、舌に受けた刺激によって生じる心の状態であると言うことができる。次に例(2)では、「かゆいところをかく刺激」は「気持ちのよいもの」とあることから、ここでの「気持ち」は、皮膚をかくという刺激によって生じた、好ましい心の状態を表していると考えられる。さらに例(3)は、水で首と胸を拭うことで「冷たくて気持ちがいい」と述べている。ここでの「気持ち」は、首や胸に受けた水の冷たさによって生じた、好ましい心の状態を表していると考えられる。よって、例(1)から(3)の「気持ち」は、身体に受けた刺激によって生じる心の状態を表していると考えられる。

以上から、「気持ち」の別義１は〈身体に受けた刺激によって生じる〉〈心の状態〉と記述することができる[6]。

４．２　別義２：〈ある物事によって引き起こされる〉〈心の状態〉

(4) （前略）上部には幅広のゴムを画びょうで留めて、スリッパ差しに。まだスペースがあれば、収納ポケットをつけて靴ベラや印鑑セットを収納する。このようにすると、狭い玄関でも、いつもスッキリ片づいて気持ちがいい。
（平成暮らしの研究会『家事そんなやり方じゃダメダメ！ :: もっと手ぎわよく済ませるために』）

(5) 『鬼平犯科帳』で有名な鬼平こと長谷川平蔵宣以は、天明七年（一七八七）より火付盗賊改めを拝命した実在の人物である。中村吉右衛門扮する平蔵が、自ら江戸市中を探索し、盗賊を捕らえ判決を申し渡す姿は、テレビで見ていても気持ちがいい。
（山本博文『サムライの掟』）

(6) 若い頃というのは、きまりきった毎日が蜿々と続くことに耐えられない気持ちになるものだ。荒業よりも日常のささいな仕事の繰り返しに耐える方が悟りに通じる、と言った僧侶もいる。　（原京加『霧の記憶』）

---

[5] 例文中、直接の分析対象となっている箇所は二重下線＝で示し、それ以外の問題となる箇所は下線＿で示す。ただし、例文が短く、該当箇所が明白である場合は、下線の処理を施さない。また、例文の文頭に付された「??」は、その表現が非文ではないが、容認度が低いことを示す。

[6] 各語の意味あるいは意味特徴は、〈　〉で括って示す。

まず例(4)を見ると、文中に「狭い玄関でも、いつもスッキリ片づいて気持ちがいい」とあることから、ここでの「気持ち」は、玄関がスッキリと片づいていることによって引き起こされる、好ましい心の状態を表していると考えられる。続いて例(5)は、平蔵が「自ら江戸市中を探索し、盗賊を捕らえ判決を申し渡す姿」をテレビで見て、「気持ちがいい」と感じている。ここでの「気持ち」は、ある人の好ましい行動や姿を見ることによって引き起こされる心の状態を表していると考えられる。さらに例(6)では、「きまりきった毎日が蜿々と続くことに耐えられない気持ちになる」とあることから、ここでの「気持ち」は、きまりきった毎日が蜿々と続くということによって引き起こされる、耐えられないという心の状態を表していると考えられる。よって、例(4)から(6)の「気持ち」は、ある物事によって引き起こされる心の状態を表していると言うことができる。

以上から、「気持ち」の別義２は〈ある物事によって引き起こされる〉〈心の状態〉と記述することができる。

４．３　別義３：〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉

(7) （前略）自分の体が、どういう状況かわからず、ただ不安を募らせていた私の気持ちをくんで、能動的な聞き方をしてくれた看護婦さんに涙が出るほどのうれしさを感じました。　（中井喜美子、親業訓練協会、近藤千恵『看護ふれあい学講座 :: 具体例で学ぶコミュニケーション訓練』）

(8) 「ありがとよ、おふみさん」
肝煎連中が気持ちのこもったねぎらいをくれた。　（山本一力『あかね空』）

(9) 私はこの時、次女もまた、父親がいなくなったことを寂しがっていると悟り、転居しようと決心した。転居はその学期の終わりに設定した。不思議なもので、母親が気持ちを変えると子どもたちもそれをすんなり受け入れる。
（神庭靖子『今どきのママ＆キッズ :: おかあさんのための児童精神医学』）

まず例(7)の「気持ち」は、自分の体に対して、どういう状況かわからず、不安な思いを抱いているという心の状態を表していると考えられる。続いて例(8)の「気持ち」は、ある女性に対して「ありがと」という思いを抱いている心の状態を表していると考えられる。さらに例(9)は、夫の転勤に伴わず、夫とは離れて今の家に住み続けようと思っていたが、娘たちが父親がいないことを寂しがるため、転居しようと決心したことを「気持ちを変える」と表現している。このことから、ここでの「気持ち」は、夫に伴って転居するかどうかという問題に対して、転居せずにここにいようという思いを抱いている心の状態を表していると考えられる。よって例(7)から(9)の「気持ち」は、ある物事に対して何らかの思いを抱いている心の状態を表していると言うことができる。

以上から、「気持ち」の別義３は〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉と記述することができる。

４．４　別義４：〈（多くは副詞的に用いられ）ほんの少し〉

(10) 「ことし十五になりました。十五と言えば、そろそろクリームや白粉の一つも欲しい年ごろ、そこでこうして呼び出して、気持ちばかりの小遣いをやろうという情けない父親であります」。　（久世光彦『謎の母』）

(11) 「気持ち右に寄って下さい」　（『講談社類語辞典』、p.1450）

(12) 「ねじを気持ち緩める」　（『講談社類語辞典』、p.1450）

まず例(10)の「気持ち」は、父親が娘に渡す小遣いの額が、ほんの少しであることを表し

ている。同様に、例(11)と(12)の「気持ち」においても、ほんの少し右に寄る、あるいはねじをほんの少し緩めるということを表していると考えられる。このことから、例(10)から(12)の「気持ち」は「ほんの少し」という意味を表していると言える。また、ここでの「気持ち」は例(11)や(12)のように副詞的に用いられることが多い。

以上から、「気持ち」の別義４は〈(多くは副詞的に用いられ) ほんの少し〉と記述することができる。

４．５　別義間の関連性について

四つの別義間の関連性について考察する。

まず、「気持ち」の別義１（＝〈身体に受けた刺激によって生じる〉〈心の状態〉）と別義２（＝〈ある物事によって引き起こされる〉〈心の状態〉）の関係について考察する。別義１と２は〈心の状態〉という共通の意味を有している。また、別義１から２は、身体を通じた経験から、より心理的な経験へと意味が拡張していると考えられるため、別義２は別義１からメタファーによって成り立っていると言うことができる。

続いて、別義２と別義３（＝〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉）の関係について考察する。別義３の〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉というのは、ある物事によってある心の状態が引き起こされ（＝別義２）、その結果として、ある物事に対して何らかの思いを抱くような心の状態が生じると考えられる。よって、別義２と３には因果関係が認められ、別義３は別義２からメトニミーによって成り立っていると言うことができる。

さらに、別義３と別義４（＝〈(多くは副詞的に用いられ) ほんの少し〉）の関係について考察する。別義４は、別義３の〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉という意味が、〈ほんの少し〉というより狭い意味に限定して用いられていると考えられる。よって、別義４は別義３からシネクドキーによって成り立っていると言うことができる。

５．「気持ち」と「気分」の意味の違いについて

分析結果をもとに、「気持ち」と意味が類似していると考えられる「気分」との意味の違いについて考察する。

『使い方の分かる類語例解辞典』は、「気持ち」と「気分」の意味の違いについて、「気持ち」は「気分」よりも「感情や考えている内容を具体的に表わすことが多い」とし、「『どうしても大学に行きたい気持ち』とはいうが、『大学に行きたい気分』とは普通いわない」と述べている[7]（p.230）。

『使い方の分かる類語例解辞典』が挙げている「気持ち」の例（「『どうしても大学に行きたい気持ち』とはいうが、『大学に行きたい気分』とは普通いわない」）から、『使い方の分かる類語例解辞典』は、本稿の「気持ち」の別義３（＝〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉）と「気分」の意味の違いについて説明していると考えられる。

次の例(13)の「気持ち」は別義３を表すが、「気分」に置き換えると不自然な表現となる。このことから、「気持ち」の別義３と「気分」には互換性がないと考えられる。

(13) 私はこの時、次女もまた、父親がいなくなったことを寂しがっていると悟り、転居しようと決心した。転居はその学期の終わりに設定した。不思議なもので、母親が気持ち（??気分）を変えると子どもたちもそれをすんなり受け入れる。　（＝(9)）

[7] 『使い方の分かる類語例解辞典』では、「気持ち」「心持ち」「心地」「気分」という四語の意味の類似点・相違点について考察されている。

しかし、次の例(14)の「気分」は「気持ち」に置き換えても、その語を含む文の意味が大きく異ならない。

(14) 体じゅうを蚊に刺された痒みと、朝日の眩しさでようやく目醒めたものの、昨日にも増した宿酔いで歩行もままならない。体も冷えきっていて胸がむかつき、気分（気持ち）が悪い。醜く歪んだ顔は幽鬼のようだ。
（塚本青史『霍去病 :: 麒麟竜彗星譚』）

例(14)の「気分」と「気持ち」は、体も冷えきり、胸がむかつくといった体の不調によって生じる心の状態を表している。ここでの「気持ち」は本稿で言う別義１（＝〈身体に受けた刺激によって生じる〉〈心の状態〉）を表すことから、「気持ち」の別義１と「気分」に互換性があると考えられる。

一方で、次の例(15)の「気持ち」は別義１を表すが、「気分」に置き換えると不自然な表現となる。

(15) ほどよくかゆいところをかく刺激は、だれにとってもとても気持ち（??気分）のよいものです。そのため、かくという行為につい浸ってしまいたくなります。（＝(2)）

例(15)では文中に「かゆいところをかく」とあるように、ここでの「気持ち」は、外部から皮膚に刺激を受けることによって、好ましいと感じる心の状態を表している。例(15)の「気持ち」を「気分」に置き換えると不自然な表現となるのは、「気分」は例(15)のように外部からの刺激によって生じる心の状態ではなく、例(14)のように体の内部の状態によって生じる心の状態を表すからであると考えられる。

以上から、「気持ち」の別義１と「気分」の意味が類似しているが、「気持ち」が表す〈身体に受けた刺激〉というのは、身体内外からの刺激を表すのに対し、「気分」は身体内部の刺激によって生じる心の状態であるという点が異なると言うことができる。

また、次の例(16)の「気持ち」は別義２（＝〈ある物事によって引き起こされる〉〈心の状態〉）を表すが、「気分」に置き換えても、その語を含む文の意味が大きく異ならないと考えられる。

(16) （前略）上部には幅広のゴムを画びょうで留めて、スリッパ差しに。まだスペースがあれば、収納ポケットをつけて靴ベラや印鑑セットを収納する。このようにすると、狭い玄関でも、いつもスッキリ片づいて気持ち（気分）がいい。　（＝(4)）

例(16)の「気持ち」と「気分」は、玄関がスッキリと片づいていることによって引き起こされる心の状態を表している。このことから、「気持ち」の別義２と「気分」の意味に互換性があると考えられる。

しかし、次の例(17)と(18)の「気分」を「気持ち」に置き換えると不自然な表現となる。

(17) （前略）女性客は実用としてのバッグを買うために並んでいるのではなく、エルメスを持って歩くことで得られる気分（??気持ち）のよさを買っているのである。はっきり言ってあんまりきれいじゃない女の人がエルメスを抱えていると、あいつあれで美人だと思って道歩いてるんだ、などと男は揶揄するが、本人はエルメスを抱えて気分（??気持ち）がいいから、男がこっちを向いていると思っているかもしれない。
つまり消費とはそういうものなのだ。物語なのである。そういう物語を毎日毎日、

自分で演出し、自分で時間を使って生きている。道行く女の人がどういうお化粧してどういう心の動きをしているのかは他人には窺うことは不可能だが、男も女もみんな各人がそれぞれ考えて生きている。　（猪瀬直樹『ラストチャンス』）

(18) 例えば、お正月などの時期はただでも気分（??気持ち）がいい、おめでたい。そして仕事も休み、その上、楽しみにしていた仲間と過ごすことができる…。こんな気分（??気持ち）が上向いているときは、人間はまた凄いエネルギーを出すことができるのです。　（松本祐『毎日楽しい氣の暮らし』）

例(17)は、「エルメスを持って歩くことで得られる」心の状態や、「エルメスを抱え」ることによって得られる心の状態を「気分」と表しているが、「気分」を「気持ち」に置き換えると不自然な表現となる。文中で、「気分」が「心の動き」という表現に言い換えられているように、ここでの「気分」は、エルメスのカバンを持って歩くことで、女性の心理状態が、普段の心理状態よりも良い方向へと変化した状態にあることを表していると考えられる。

次に例(18)を見ると、例(18)においても「気分」を「気持ち」に置き換えると不自然な表現となる。文中に「お正月などの時期はただでも気分がいい」とあるように、ここでの「気分」は、「気分がいい」と感じる具体的な要因がなくても、お正月の雰囲気といった漠然とした要因によって引き起こされる心の状態を表している。しかし、「気持ちいい」は具体的な要因によって引き起こされる心の状態であり、「ただでも」といった漠然とした理由によって引き起こされる心の状態を表すことはできない。さらに、文中に「気分が上向いている」とあることから、ここでの「気分」も例(17)のように、普段の心理状態よりも良い方向へと変化した状態にあることを表していると考えられる。しかし、「気持ち」は、普段の心理状態よりも良い方向あるいは悪い方向に変化した状態にあることを表すことができないという点で両者は異なっている。

以上から、「気持ち」の別義２と「気分」の意味が類似しているが、「気持ち」は「気分」のように具体的な要因がなくても、漠然とした理由によって引き起こされる心の状態を表すことができないという点に加え、「気持ち」は「気分」のように普段の心理状態よりも良い方向あるいは悪い方向へと変化した状態にあるということを表すことができないという点で両者は異なると言うことができる。

６．まとめ

本稿では「気持ち」に四つの多義的別義を認め、次のように記述した。

別義１:〈身体に受けた刺激によって生じる〉〈心の状態〉
別義２:〈ある物事によって引き起こされる〉〈心の状態〉
別義３:〈ある物事に対して何らかの思いを抱いている〉〈心の状態〉
別義４:〈(多くは副詞的に用いられ）ほんの少し〉

また、別義間の関連性については、別義２は別義１からメタファーによって成り立っており、別義３は別義２からメトニミーによって成り立っており、別義４は別義３からシネクドキーによって成り立っているということを述べた。

今後の課題としては、「気分」の意味を明確に記述すること、また「気持ち」や「気分」の類義語である「心地」や「心持ち」の意味、および四語の意味の違いについて考察する必要があると考える。

文　献

國廣哲彌（1982）『意味論の方法』、大修館書店
国広哲弥（1997）『理想の国語辞典』、大修館書店
柴田武、山田進、加藤安彦、籾山洋介（編）（2008）『講談社類語辞典』、講談社
小学館辞典編集部（編）（2003）『使い方の分かる類語例解辞典　新装版』、小学館
籾山洋介（2009）『日本語表現で学ぶ入門からの認知言語学』、研究社
籾山洋介（2010）『認知言語学入門』、研究社
森田良行（1989）『基礎日本語辞典』、角川学芸出版

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

田中リベカ（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（お茶の水女子大学大学院人間文化創成科学研究科／国立情報学研究所）

MCN Corpus:

Methodology on the Modality Annotation based on Linguistic Tests

Ribeka Tanaka (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

自然言語処理においては近年、文の表層的な形式から得られる浅い情報にとどまらず、その興味の対象が深い意味に移行してきている。言語の意味に関しては、母語話者間で理解が共有されており、その構造には何らかの法則性があることが示唆される。このため、そのような話者間に共通の意味の認識が反映されているデータの需要が今後ますます高まることが予想される。

しかし、意味を対象としたリソースを作成する方法論は確立していない。意味アノテーションにおいては、言語表現の多義性解消や用法の特定などにおいてアノテータの一貫した判断が求められるため、ガイドラインにおいて判断基準を明確に提示する必要がある。ガイドラインがこのような判断基準や手段の提示を欠いている場合や、提示に失敗している場合、アノテータ間での判断の不一致を引き起こし、本来共有しているはずの意味への認識が正しく反映されない恐れがある。このような判断の不一致を避けるために、しばしばアノテーション作業を複数人でなく一人で行うという方針がとられることがある。しかしこのような方針においては、作成されたリソースが一人のアノテータの主観によるものではないとは言い切れない。またその判断の正当性を第三者が検査するのも困難である場合が多い。

そこで田中ら(2012)[2]では、言語学的テストを用いた意味アノテーションの方法論を提案した。ここで言う「言語学的テスト」とは、理論言語学の理論構築および検証に用いられるテストで、文や文の一部の容認性や適切性を判定するものである。言語学的テストは、容認性や適切性を左右する条件を特定するために、検証したい部分以外の条件をほぼ同じにした二つ以上の文からなる群として提示されることが多い。

本論文では以下、第二節で田中ら(2012)[2]における意味アノテーションの方法論を紹介する。特に、意味アノテーションの作業を一種の「分類タスク」と考えた場合に、適切な言語学的テストの有無が分類を決定する際の判断の一貫性を左右することを実例を交えて論じる。第三節では、アノテーションに有効な言語学的テストの設計についての一般的な

方法論を提案する。第四節でその適用例を見た後、第五節では言語学的テストにおけるアノテータの判断の一致度に関して筆者らが行った調査について述べ、実際にガイドラインを作成する際に注意すべき技術的な問題について考察する。そして最後に、本手法のもつ言語学的な意義について論じる。

なお、本論文において「意味アノテーション」の具体例として用いるのは、「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）[1]）に基づくアノテーション作業である。このガイドラインは、様相表現・条件表現・否定表現など、言語情報の確実性に影響する表現とそのスコープ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。

2 コーパスアノテーションにおける言語学的テスト

意味アノテーションに限らず、コーパスに対するアノテーション作業の多くは、スキーマの設定者があらかじめ設定したラベルをアノテータがテキストの一部に対して付与していくものである。これはある意味、実際のテキスト中の表現を用意されたカテゴリに分類していく作業であるととらえることもできる。アノテーションガイドラインにはそのような分類を行う際の判断基準についての指示が多かれ少なかれ含まれるわけであるが、その細やかさ・適切さのレベルは様々である。

たとえば、テキストに現れる「(と）いう」という表現のうち、他人による認識あるいは主張を表すものにのみアノテーションをしたいとする。このとき、ガイドラインには以下のようにアノテーション対象のカテゴリの説明のみを提示し、テキストに現れる個々の「(と）いう」の出現がそのカテゴリに属するものかどうかの判断はアノテータに委ねる、という方法が一般的である。

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）

しかしこのような方法は、どのような基準で「他人の報告する事柄」と見なせば良いのかということや、そもそもどのような表現を「命題」とみなすかなどについても明確には定められていないという点で、表現を分類する基準や詳しい知識をアノテータが持っていることを前提としていると考えられる。このため、母語話者であっても専門的な知識を有していないアノテータには判定が困難である。

これよりもう少し細やかな指示としては、以下のように例文を示した上で、実際の表現がその例文と同じ用法で使われているかどうかを判断させるような方法がある。

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：今冬のインフルエンザの流行は全国的に遅れているという。

上のように例文を提示した場合、ほぼ同じ構造の文に関してはアノテータは専門的な知識を要さずに判定することができる。しかし、表現の形式が変わったり、少しでもニュアンスが異なるような表現に遭遇したりすると途端に判断が困難になる。上の例は「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」にある記述を内容を変えずに再編集したものであるが、筆者らが実際にアノテーション作業を行ったところ、アノテーションの可否が判断出来ない、又はアノテーションの結果が一致しない例に直面した。そのような例の一部を以下に示す。

1. 呼吸困難に陥る可能性があるという。
2. 「彼を絶対に許さない」と言う。
3. 太郎が責任をとるべきという人はどうかしている。
4. 太郎は結婚したという話だ。
5. インスリンというホルモン。
6. 車がガタガタという。
7. お前という人間が信じられなくなった。
8. サプリというサプリは試した。

1.については、アノテーションガイドラインに記載されている例文と非常によく似ていることから、「他人の認識を表す『(と)いう』」としてアノテーションできると想像がつく。他方、2.-8.については決定的な判断基準がない。例文と似ているかどうかをどのような点に着目して判断するのかについて明らかにされていないため、判断基準がアノテータの主観的な理解に委ねられてしまうのである。筆者らの行ったアノテーション作業においては、特に2.3.4.の表現についてアノテータ間でアノテーション可否の判断が大きく分かれた。特に困難であったのは、3.4.の判断である。これらはどちらも名詞句を修飾していて見た目には同じ構造をしており、一見するとガイドラインの例文とは異なる形をとっている。しかしガイドラインの設計者は、3.は他人の認識としてアノテーション可能であり、4.はアノテーション不可能であることを意図している。このように、例文ベースの指示では例文との類似度の判断は困難であり、アノテータ間でも判断が一致しにくい。

そこで筆者らは、田中ら(2012)[2]において言語学的テストを導入し、ガイドラインを以下のように改善した。

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：今冬のインフルエンザの流行は全国的に遅れているという。
テスト１：「(と）述べる（述べられる)」に置き換えても意味が変化しない。
テスト２：名詞句を修飾する場合「～との」に置き換え不可

テスト１，２はともに表現の「置き換え」に基づく言語学的テストである。これらのテストを先に挙げた例に適用すると、結果は以下のようになる。

1. （テスト１）呼吸困難に陥る可能性があると述べる（述べられる)。/(テスト２：適用不可)
2. （テスト１）「彼を絶対に許さない」と述べる。/(テスト２：適用不可)
3. （テスト１）太郎が責任をとるべきと述べる人はどうかしている。/(テスト２)*太郎が責任をとるべきとの人はどうかしている。
4. （テスト１)*太郎は結婚したと述べる（述べられる）話だ。/(テスト２）太郎が結婚し

たとの話だ。

5. (テスト１)*インスリンと述べる（述べられる）ホルモン。/(テスト２)*インスリンとのホルモン。
6. (テスト１)*車がガタガタと述べる（述べられる)。/(テスト２：適用不可)
7. (テスト１)*お前と述べる（述べられる）人間が信じられなくなった。/(テスト２)*お前との人間が信じられなくなった。
8. (テスト１)*サプリと述べる（述べられる）サプリは試した。/(テスト２)*サプリとのサプリは試した。

上のように、実際に該当箇所の表現を置き換えた文について考え、別の表現に置き換え可能かどうかを判定させる言語学的テストを導入すると、このテストについて複数のアノテータのYES/NO判定は一致しやすく、結果としてアノテーション可否の判断の一致度も向上した。実際に、テスト１、テスト２のいずれかでNOの判定がでた 4.-8. の表現についてはアノテーション不可能であると判断できるようになった。また、言語学的テスト導入前には困難であった 3. と 4. の区別も、それぞれのテストで異なる結果が出たことで明確になった。

また、筆者らが上記のテストを用いて行った実際のアノテーション作業において、注目すべき現象がみられた。「置き換えが可能」という判断よりも「置き換えが不可能」という判断の方がアノテータ間で一致しやすかったのである。このことから、ガイドラインにおいては、「置き換え可能ならばアノテーション対象表現である」という指示よりも「置き換え不可能ならばアノテーション対象表現ではない」という指示を用いる方が、一貫性のあるアノテーション結果を得るのに有効であるとの結論に至った。

３ 言語学的テストの設計方針

前節で述べた作業は筆者らが行った具体的なアノテーション作業であるが、アノテーションガイドラインを設計するに当たっての重要な示唆を含んでいると考えられる。本節ではこれらの考察を元に、理想的なガイドラインを作成するにはどのようなテストを設計すべきかを論じる。

ここで、ある表現Ｅに対して、その用法として$A_1, A_2, \cdots, A_n$のｎ個の分類を考える。ここで、実際のテキストに表現Ｅの出現Ｅ’があったとき、このＥ’がｎ個の分類中どこに属するか（つまりＥ’が表現Ｅのどの用法になっているか）を判定するテストを考える。

表現の出現Ｅ’が特定の分類A_iに属するかどうかの判断基準となるテストを、ここでは「個別テスト」と呼ぶ。また分類$A_1, A_2, \cdots, A_n$の個別テストを集めたもの全体を「テストセット」と呼ぶこととし、以下において個別テストとテストセットの２段階における性質を考える。

３．１ 個別テストの構築

まず個別テストの内容については、別の表現に実際に置き換えた上で「意味が変わるかどうか」「文自体が意味不明になるかどうか」をYES/NOで判定する形式や、「活用しているか」「コト節を受けているか」など表層上明らかな性質を条件に設定するのが望ましい。前節で見たように、「この表現と同じ用法か」という形式をとった場合、判断は個々人の独自の解釈に左右される。しかし置き換えた上で意味が変わるかどうかの判定は言語直感に訴えるものであり、母語話者を対象とする限りは専門的な知識を要求しないと考えられる。

無論、テストを作成する際にはどのような語に置き換えるよう指示するかについても慎重にならねばならない。たとえば慣用的に使われている表現を用いると、必ずしも本質的ではない点で混乱を生む可能性がある。細心の注意を払ってもなお、設計者が意図しない解釈がされた場合には個別テストを修正する必要があるが、これに関しては後述する。

前節で述べたように、「置き換えが可能」という判断よりも「置き換えが不可能」という判断の方がアノテータ間で一致がしやすいという現象が見られた。この性質をテストに反映するために可能な個別テストのパターンは次の２通りが考えられる。

1)　表現Ｅ’を表現e_iに置き換え不可能ならば、表現Ｅ’は分類A_iに属さない。
2)　表現Ｅ’を表現e_iに置き換え不可能ならば、表現Ｅ’は分類A_iに属する。

本手法ではこのうち、1）の形式のテストを採用する。このような形式を満たすテストを作成するためには、それぞれの分類A_iについて、

3)　表現Ｅ’が分類A_iに属するならば、表現Ｅ’を表現e_iに置き換え可能である。

といえるような表現e_iを見つけ、3）の対偶をとれば良い。これにより、1）の形式のテストが得られる。仮に 2) の形式のテストを採用すると、テストを作成するためには、上と同様の方針に基づいてそれぞれの分類A_iについて

4)　表現Ｅ’が分類A_iに属さないならば、表現Ｅ’を表現e_iに置き換え可能である。

といえるような表現e_iを探すことになる。しかし、A_i以外の分類に属するすべての表現と置き換え可能であるような表現e_iを探すことは、対象が広すぎるため非常に困難である。これに対して 3) を満たすような表現e_iについては、ある特定の分類A_iについてその用法がもつ性質に着目して探せば良いため、対象が絞られるだけでなく言語学においてなされた用法の分析の成果を有効に活用することが可能である。以上を踏まえて、本手法では 1) の形式を個別テストの基本的な型とし、これを「ネガティブテスト」と呼ぶ。適切なネガティブテストには、その表現に置き換えられなかったら特定の分類に属さないのだと言い切れるような強い条件を用い、判断に迷うような紛らわしい例を避けることが重要である。

なお、「置き換え不可能である」という判断がアノテータ間で一致することの背景については、次のように考えられる。「置き換え可能」であるという判断は、ニュアンスの変化を許すか、規範的な日本語のみを許すかなど、どの程度なら置き換え可能とみなすかについて個人差がある。また、自分が発話する表現と理解できる表現にも差があり、日常生活ではそれらの程度の異なる言語事象が混在しているため、「置き換え可能か」と問われた場合にはどこを基準にするべきかが明確に定まらない。しかし一方、「置き換え不可能」である表現はコミュニケーションにおいて使用できない（正しく意味をなさない）表現である。このため、本来置き換え不可能な表現を強引に置き換えた文を提示されると、母語話者は言語直感によってその文が正しくないこと、すなわち「置き換え不可能」であることを判定することが可能である。逆にいうと、テストを作成する際には母語話者なら誰もが「置き換え不可能」と判断するであろう表現を意図的に選ぶことが重要である。

３．２ テストセットの構成

上のような方針で個別テストを設計することを前提とした上で、テストセットとしてどのような性質を満たすべきなのかを考える。

一つの個別テストにおいては、表現e_iに置き換え不可だった場合は分類A_iに属さないと言えるが、置き換え不可だと言い切れないと判断された場合はその分類に属すとも属さないとも判定されない。これは、表現e_iに置き換え不可能であるということが、分類A_iに属さないことの十分条件ではあるが、必要条件にはなっていないことによる。つまりネガティブテストは単独では分類を決定しない。そこで、このような個別のネガティブテストを全ての分類について作成する。個々の個別テストでは「置き換え不可能なので、この分類には属さない」又は「置き換え不可能ではないので、この分類に属さないとは言い切れない」の２パターンの判定がなされうる。「この分類には属さない」の判定が出た場合、表現の出現Ｅ’がその分類に属す可能性はなくなるので分類先の候補から除くことができる。また、「この分類に属さないとは言い切れない」という判定が出た場合には、その時点ではこの分類に属すか否かは判定できないので分類先の候補として残す。ここで、網羅的な分類と適切なネガティブテストから構成された理想的なテストセットを作成すると、全ての分類の個別テストを試した結果、分類先の候補は１つしか残らないはずである。この残った１つの候補は、個別テストにおいて「置き換え不可能ではないので、この分類に属さないとは言い切れない」と判定された唯一の分類であり、すなわち表現の出現Ｅ’の正しい分類先であるといえる。このように網羅的な分類と各々の分類についての適切な個別テストからなるテストセットを構成し、消去法により表現の出現Ｅ’の分類先を特定する。

テストセットを構成する際には、任意の２つの分類A_i, A_jの個別テストを比較したときに一方が他方を含まないようにする必要がある。仮にA_iの個別テストが全てA_jの個別テストに含まれていると、「分類A_jに属さないとは言い切れない」と判定された場合には必ず分類A_iについても同様の判定がなされ、必然的に分類先が１つに特定されない。無論、複数の個別テストが同じ内容であることも同様の理由により避けねばならない。ガイドラインの設計者は、異なる用法として分類を分けるならばその根拠となる性質をテストに反映しなくてはならない。

テストセットにおいて、個々の個別テストは独立に振舞う。決定木のように、あるテストの判定が次のテストを適用する条件になっているわけではなく、どの個別テストから適用するかの順によらず分類先は決定する。そのため、一部徐々に条件を狭めていくような状況を作りたいときには、個別テストで意図的に実現する必要がある。例えば、分類A_iに分類される出現は表現e_a, e_bの両方に置き換え可能なのに対し、分類A_jに分類される出現は表現e_aには置き換え可能だが表現e_bには置き換え不可能であるという状況を考える。このとき、２つの分類の個別テストはそれぞれ次のように作成される。

【分類A_iの個別テスト】

_ 表現Ｅ’を表現e_aに置き換え不可能ならば、この分類ではない。

_ 表現Ｅ’を表現e_bに置き換え不可能ならば、この分類ではない。

【分類A_jの個別テスト】

_ 表現Ｅ’を表現e_aに置き換え不可能ならば、この分類ではない。

_ 表現Ｅ’を表現e_bに置き換え不可能ならば、この分類である(可能性がある)。

ここで、下線部のように「この分類である」という表現を意図的に用いることにより、e_aに置き換え可能であるという共通の条件のもと、e_bに置き換え可能か否かでどちらの

分類に入るかが決定するような形式を実現することができる。(なお下線部のような表現は、先述したように条件が見つけにくいため個別テストの基本的な型としてはふさわしくないと考えられるが、このように十分対象が狭まったと考えられる状況下では避ける理由はない。)また、「可能性がある」としたのは個別テストの独立性を保つためである。分類A_jの個別テストとしては、「表現E'を表現e_bに置き換え不可能」という条件のみからは、まだ「この分類である」と決定できないことは明らかである。一方、「e_aに置き換え不可能」というテストとは切り離して考えたいので、このような表現を用いるのが適切だと考えられる。「可能性がある」としておくことで、他のケースと同様、消去法の後に分類が決定することになる。

このような形式で個別テストを作成することは一見冗長であり、個別テストを互いに独立なものとしたことの弊害にも見える。しかし以下で述べるようにこの方法論においてはテストの修正が重要であるため、個別テストが互いに独立であることによって一部の個別テストの変更がそれ以外の箇所には影響せず、修正を容易にしているということは大きなメリットである。

３．３ テストの修正

本手法では、網羅的な分類と適切な個別テストでテストセットが構成されると、理想的には消去法で分類先が１つ特定される。しかし、ガイドラインの設計者が最初から用法を網羅的に全て把握しているわけではなく、また設計者にとって明らかな個別テストであっても、実際に使用してみるとアノテータにとっては紛らわしいこともある。このため分類先が１つに特定できないということが起きた場合には、アノテータは分類先が特定出来なかったことをガイドライン設計者に報告し、設計者はテストセットの修正を行う必要がある。

消去法の結果、候補が複数残った場合

消去法を行った結果、分類先の候補が１つだけ残ることが理想であるが、複数残ってしまう場合がありうる。これは本来その表現が分類されるべき分類の個別テスト以外にも、「置き換え可能である」との判定が出てしまった個別テストが存在したことを意味する。置き換える表現の選択が悪かったことが原因であるため、意図せず「置き換え可能である」の判定がされてしまった分類の個別テストの内容のみを修正する。この際、それ以外の分類に対しては「置き換え不可能である」との判断が正しくなされたのは事実であるので、修正をする必要はない。またスキーマの設計者が意図する本来の分類先についても、「置き換え可能である」の判断がされたことは自然な現象であるので修正をする必要はない。

消去法の結果、候補が１つも残らなかった場合

全ての個別テストで、「置き換え不可能である」という判定が出てしまったときには候補が１つも残らないことがある。このとき、２つの原因が考えうる。

１つは、本来の分類先の個別テストが適切でなかったため、ガイドラインの設計者が「置き換え可能である」として設定したにも関わらずアノテータが「置き換え不可能である」と判断した場合である。このような場合には、「置き換え可能である」の判定が出るように本来の分類先の個別テストを修正する。

もう１つは、その表現の出現が本質的にどの分類にも属さない用法であった場合である。この場合は分類が網羅的でなかったことが原因であるため、ガイドライン設計者は新しい分類を増やしその個別テストを作成する。

なお他の可能性として、消去法の結果候補が１つに絞られ分類が決定されたが、それが

ガイドライン設計者の意図とは異なる分類である場合がありうる。このような場合には、アノテータによる問題点の報告がなされないためテストの修正はなされない。このことは、複数のアノテータが同じテキストに対してアノテーションをして結果を比較するような過程をとらない限り、避け得ないことである。しかし、これはガイドライン設計者の意図する分類の個別テスト、およびアノテータによる分類先の個別テストの二箇所において設計の誤りがある場合にのみ起こることである。また、それ以外の場合には上に述べたように、網羅的な分類と適切な個別テストが最初から得られていなくても、ある程度適切なテストセットであるならば実際のアノテーション作業を通して問題箇所がピンポイントに浮き彫りになり、理想的なテストセットへと修正していくことが可能である。

４ 適用例

前節の考察に基づき、適切なネガティブセット、および表現の意味の網羅的な分類を含むテストセットの例を表１に示す。ここでは、「（と）いう」をその意味に応じて８つのカテゴリに分け、テキストにおけるこの表現の個々の出現例がどのカテゴリに属すかを判定するためのネガティブテストを８つ用意している。

表１：テストセット構築例

iu	1	いう	言う	MODAL	hearsay	話者にとって真偽が未知、確実性判断なし	動作主が明示されず、「世間一般」「人々」「専門機関あるいは情報ソース」である。「～によると」と共起することが多い。 ソースが「世間一般」の場合は、「いう」を「いわれる」に置き換えてもほとんど意味が変化しない。	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。 世界には自分と同じ顔の人間が７人はいるという。 東京で一番おいしいという焼肉屋へ行ってみた。	Negative test5:「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）の場合はこのカテゴリではない。	[命題(S)]と　いう
iu	2	いう	言う	MODAL	hearsay	話者にとって真偽が未知、確実性判断なし	「述べる（述べられる）」「主張する（主張される）」「報告する（報告される）」と同義。命題をとる。	花子は、彼を絶対に許さないという。 太郎が責任をとるべきという人は、どうかしている。	Negative test1：「述べる（述べられる）」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。 Negative test3：過去形「（と）いった」に置き換え不可な場合は、このカテゴリではない。	[動作主(NP)]が［命題(S)]と　いう。
iu	3	いう		ー			「呼ばれる」「名づけられた」と同義。	山田さんという人が来た。 血糖値を下げるのは、インスリンというホルモンだ。	Negative test4:「呼ばれる」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。 Negative test8:「という～」を除けない、あるいは除いて意味が変化する場合はこのカテゴリではない。	[NP]という
iu	4	いう		ー			「音や声を発する」の意味。副詞、あるいは「～と」を伴って擬音語あるいはそれに類する表現をとる。	あの人はぶつぶつ（と）言っている。 車がガタガタ（と）いう。	Negative test3：過去形「（と）いった」に置き換え不可な場合は、このカテゴリではない。 Negative test6:副詞、あるいは「～と」を伴って擬音語あるいはそれに類する表現をとっていない場合は、このカテゴリではない。 補助的テスト：このカテゴリに属するものは、「ぶつぶつ」「ガタガタ」のような擬音語的な表現を取り去ると意味をなさない。（十分条件） *あの人は言っている。（比較：あの人はぶつぶつと言っている。） *車がいう。（比較：車がガタガタいう。）	[擬音語]という
iu	5	いう		ー			「話」「噂」「見解」などを修飾する。	太郎が結婚したという話だ。 時期尚早であるという見解を示した。	Negative test2：「との」に置き換え不可な場合は、このカテゴリではない。 Negative test3：過去形「（と）いった」に置き換え不可な場合は、このカテゴリではない。	「命題(S)]という話/噂/見解・・
iu	6	いう		ー			名詞句を修飾。補足的な意味を付け加える機能を持つ。	お前という人間が信じられなくなった。 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。 今日という日を忘れないようにしよう。 それが男というものだ。	Negative test8：「という～」を除けない、あるいは除いて意味が変化する場合はこのカテゴリではない。	[NP]という
iu	7	いう		ー			前後に同じ形の名詞をとる。	サブリというサブリは試した。 医者という医者には相談したものの、どうにもならなかった。	Negative test4:「呼ばれる」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。 Negative test7:前後に同じ形の名詞をとっていない場合は、このカテゴリではない。	[NP]という
iu	8	いう		ー				太郎の友人であるという人物が現れた。		「命題(S)]という[NP]

分類１から７に関しては、どの２つの分類のテストを比較しても互いに異なるように設計した。ただ一箇所、分類３と分類６では一方が他方を含むようになってしまっているため、この分類６に別の個別テストを増やすなどして修正することが必要である。

また、分類８の個別テストは空欄になっているが、これは分類８がごく最近見つかった分類であることによる。ネガティブテストを適用した結果、どの分類についても「この分類ではない」という判定が出たため、新しく追加されたものである。今後この欄に新しい

個別テストを追加する予定である。

「(と) いう」に関してこれが完全に網羅的な分類であるかは現時点では不明であるが、この分類のみに関して言えば、明確な方針のもと修正を行った後、ネガティブテストの組み合わせによってどのカテゴリに属するかを判断することが可能である。

５ 言語学的テストにおけるアノテータの判断一致度の調査

筆者らは本手法に基づき、ネガティブテストにおけるアノテータの判断の一致度を調査した。調査は、大学1年生～大学院生を含む日本人学生 40 人程度を対象に行った。被験者には用法の分類などの詳細については一切明らかにせず、「(と)いう」が出現する文を提示し、「(と)いう」の箇所を別の表現 e に置き換えると意味が変化するか、という形式の問いに対してYES/NO/わからない、で回答するようにした。

この調査結果については現在分析中である。本手法を用いる際、どういう状況ならアノテータの判断が一致したと言って良いのか、また実際に何人以上の回答が意図に合わなかった場合にテストを修正するのかについて、調査結果を元に仮説を構築する予定である。

調査を行う中で、特に言語学的テスト（ネガティブテスト）の提示方法の改良について、いくつかのノウハウが得られた。

まず、「表現Ｅ’を別の表現 e に置き換えると意味が変化する」と問うた場合には被験者による YES/NO の判断の一致度は悪かったが、実際に表現Ｅ’を表現 e に置き換えた文を提示し、「この文は置き換える前の文と比較して意味が変化する」と問う形式に変更すると一致度が良くなった。このことから、一部の被験者は置き換え可能かどうかを判定する際、誤って都合の良い置き換えをしてしまうことが示唆された。

また、文の外見上自明な特徴について問う設問も交ぜたところ、YES/NO で答えず「わからない」という回答をする被験者が多かった。あまりにも自明な事柄を問うと、設問者の意図について必要以上に意識してしまうためであると考えられる。

提示する文が短いと、置き換え後の意味が変化するかどうかの判断が困難となり、判断が一致しにくかった。一方、前後の一文程度を併せて提示すると、提示しないときと比較してアノテータの判断は収束した。実際のアノテーション作業ではどの程度前後の文を読むかはアノテータによって差が生じることが考えられるため、意味の変化を考える際に文脈をどの程度考慮するかについては個人差が生じる恐れがある。

今後ガイドラインの改良を進めていくにあたっても、これらの要因について考慮していく必要があると考えている。

６ 言語学的な意義

本手法は、自然言語処理におけるコーパスの意味アノテーションのための手法であるが、この手法を通して得られるテストセットは言語学的にも意義があるものであると考えられる。

通常、表現の用法を全て列挙することは容易ではない。一般に国語辞典には表現の意味の用法が記載されているが、筆者らの行ったアノテーション作業の結果によれば、実際の言語現象においては辞典に記載されている用法よりも更に細かく意味が区別されていることが明らかである。また、「（と）いう」のような複合語については、「と」と「いう」の意味を単純に合わせたものではない独自の意味をもつにも関わらず、辞書の項目にないことも多い。したがって本手法を通して表現の網羅的で重複のない分類が得られること、複

合語のような表現に対してもその意味が分析されることは大きな長所である。

また、本手法においては第三者が分類の正当性を確認できるという点も重要である。用法の分類や表現を区別する根拠はガイドライン設計者によって考えられたものではあるが、その正当性は常にアノテータによって確認される。アノテータの判断が一致し、ガイドライン設計者が意図した分類に分類されるということを以って、正しい分類であることが裏付けられる。逆にアノテータの判断が一致しなかった場合には、設計者は自分の考えた分類や個別テストを再検討することになり、設計者が誤った分析をしていた場合でも修正が行われる。ガイドライン設計者の主観ではなく、母語話者が共通して持つ言語の意味への客観的な認識が、用法の分類と各個別テストに反映されるのである。

7 おわりに

以上、意味アノテーションにおける言語学的テストの利用、およびその際の方法論について論じた。

上で述べた方法論に基づき、「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」に含まれるものを中心に、340 種類の様相表現、13 種類の否定表現、48 種類の条件表現の網羅的な分類とネガティブテストの構築を行い、ガイドラインの改良を行う予定である。

謝辞

国立情報学研究所人工頭脳プロジェクト「ロボットは東大に入れるか」NLP コア定例ミーティングにて、メンバーの方々に貴重なコメントをいただいた。

文献

[1]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.

[2]田中リベカ、小池恵里子、戸次大介、川添愛(2012)「言語学的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.

メタファー表現の生産性に対する意味の焦点と表現メディアの影響
—＜急激な増加＞や＜大量の存在＞を表す表現の場合—

大石 亨（明星大学情報学部）

Influence of Semantic Focus and Registers on the Productivity of Metaphorical Expressions –The Case of the Verbs Representing "Rapid Increase" or "Existence in Large Quantity"

Akira Oishi (Meisei University)

１．はじめに

以下の（１）から（５）の表現は，何らかの事物（嫌なものであることが多い）が急激に増加したり，その結果として大量に存在したりすることを表すメタファー表現である。

（１） 人口が爆発する
（２） 外来語が氾濫する
（３） 拝金主義が蔓延する
（４） 批判が噴出する
（５） 赤字が膨れ上がる

ここで，（１）の「爆発」を「破裂」や「炸裂」に置き換えたり，（２）を「外来語で水浸しになる」などと言い換えたりすることはできない。このように，概念メタファー理論(Lakoff and Johnson 1980, 1999; Lakoff 1993)の予測に反して，メタファー表現の産出が保守的であり，使用語彙が制限されていることは「まだら問題」と呼ばれて研究者の注目を集めている(Grady 1997; Clausner & Croft 1997; 黒田 2005; 鍋島 2007, 2011; 松本 2007)。

松本(2007)は，語におけるメタファー的意味の実現に関して見られるギャップ（概念メタファーの不適用）に対する説明として，１）目標領域における対応物の不在による写像の不成立，および，２）語義的な経済性を求める傾向の二つを挙げている。前者は身体部位詞の物体部位詞への意味拡張に対して，後者は類義動詞におけるさまざまなメタファー的意味の実現に対する説明として与えられているものである。本稿は，類義動詞を考察対象とするので，２）語義的な経済性を求める傾向が関わってくることになるが，これは，次のような制約としてまとめられている。

（６）**語義的経済性の制約**：概念間の対応関係がある時，ある語がそれに基づくメタファー的意味を実現させることができるのは，より適切な語（過剰指定がより少ない表現）がなく，かつ，同じ意味を表すものとして他の語が定着していない場合のみである。(松本 2007, 82 ページ)

この制約は，一般的現象としてメタファー表現の保守性を捉えようとしたもので，その限りで異論はない。しかし，具体的な意味の過剰指定の内容や語の定着性の程度については，少数の用例のみに基づいて論じられており，その説明が正しいかどうか，実際の言語使用に基づいて検証する必要がある。そのために，本研究では，（１）から（５）のように，＜急激な増加＞や＜大量の存在＞を比喩的に表すいくつかの動詞を取り上げ，それらの類義動詞が大規模コーパスの中でメタファー表現として用いられている割合を，表現メディアごとに調査した。本稿における表現メディアとは，「日本語書き言葉均衡コーパス(BCCWJ)」のサブコーパスのことである。また，調査対象語彙は表 1 のとおりである。

表 1　調査対象語彙と出現頻度

カテゴリー	調査対象語彙	出現頻度		MLR
		リテラル（L）	メタファー（M）	（M／L）
爆発系	爆破する	135	4	0.03
	破裂する	188	58	0.31
	爆発する	439	489	1.11
	炸裂する	71	80	1.13
氾濫系	冠水する	31	0	0.00
	浸水する	95	0	0.00
	溢れ出す	88	73	0.83
	溢れ出る	93	108	1.16
	氾濫する	79	144	1.82
繁殖系	繁殖する	300	10	0.03
	繁茂する	82	3	0.04
	増殖する	312	107	0.34
	蔓延る	21	174	8.29
	蔓延する	7	262	37.43
噴出系	噴火する	73	5	0.07
	沸騰する	537	64	0.12
	噴き出る	79	10	0.13
	噴き出す	502	106	0.21
	噴出する	172	145	0.84
	沸き上がる	14	117	8.36
	沸き起こる	10	249	24.90
膨張系	膨張する	223	113	0.51
	肥大する	67	43	0.64
	膨らむ	1074	819	0.76
	膨れ上がる	171	283	1.65
総計		4863	3466	0.71

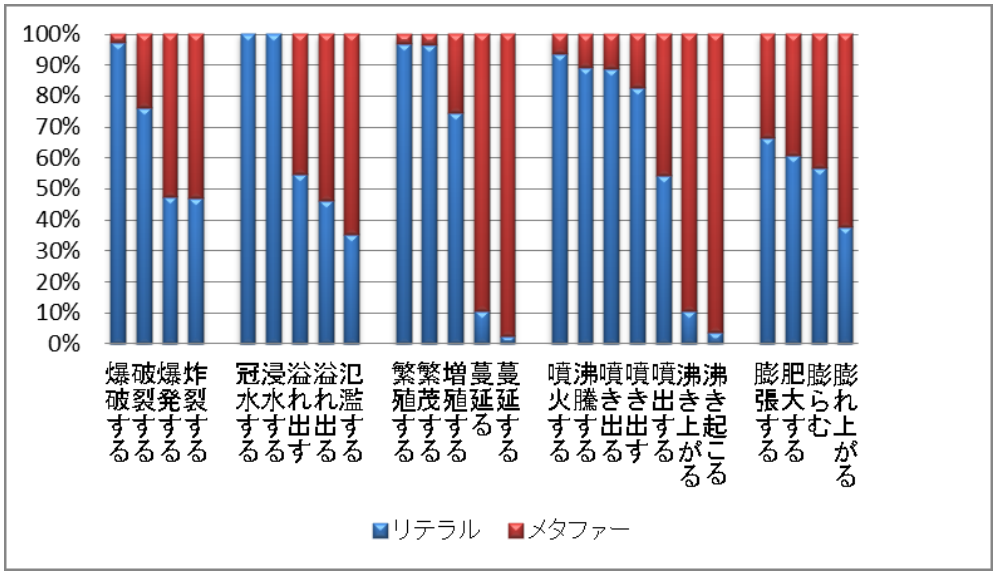


図 1　調査対象語彙のメタファー表現と字義的表現の出現比率

用例の抽出には BCCWJ 検索サイト「中納言」を用い，ダウンロードした検索結果を SortKWIC(田野村 2012)によって Excel に収めたものを，手作業によりメタファー表現と字義的な表現に分類した[1]。表 1 の第 3 列が字義的な意味での出現頻度，第 4 列がメタファー表現としての出現頻度である。第 5 列は，メタファー表現の頻度の字義的な表現の頻度に対する割合(Metaphor/Literal Ratio: MLR)を表している。また，図 1 はこの比率を 100%積み上げ棒グラフで表したものである。図 1 から明らかなように，字義的には類似した事態を表すこれらの類義語には，メタファー表現の使用比率に極端なばらつきが見られる。以下では，このようなメタファー表現の生産性の違いが，メタファーによって置き換えられる対象が図地分化における図でなければならないという制約（メタファーの対象焦点化制約）および印象形成力や字義的用法の頻度など，複数の要因によってもたらされていることを論じる。さらに，調査の中で見出されたパターンに基づいて過去の研究事例を見直してみると，そこで与えられている意味の過剰指定の内容とは異なる説明が可能であることを示す。

２．メタファー表現の生産性

本節では，表 1 のカテゴリーごとに語彙のメタファー表現の生産性の違いを具体的に見るとともに，その原因を考察する。同時に，メタファー表現の出現する表現メディアにどのような偏りがあるかについても述べる。

２.１　爆発系語彙

爆発系では，「爆破する」＜「破裂する」＜「爆発する」＜「炸裂する」の順にメタファー表現の比率が高くなっている。「爆破する」がメタファーとして用いられているのは，139 例中 4 例のみであり，「破裂する」のメタファー表現は 58 例あるが，「癇癪（玉）が破裂する」という固定表現(12 例)と，「心臓が破裂しそう」という誇張表現(8 例)など，定型表現が目立つ。一方，「爆発する」は，「怒り」や「不満」等，押さえつけられていた感情が一気に表出する状況を中心として，「人口」，「人気」，「打線」，「アフロヘア」等にも広く用いられており，「炸裂する」は，「パンチ」「拳」などの攻撃を中心に，「音響」や「叫び」，「親バカ」など，多様な語彙と共起している。

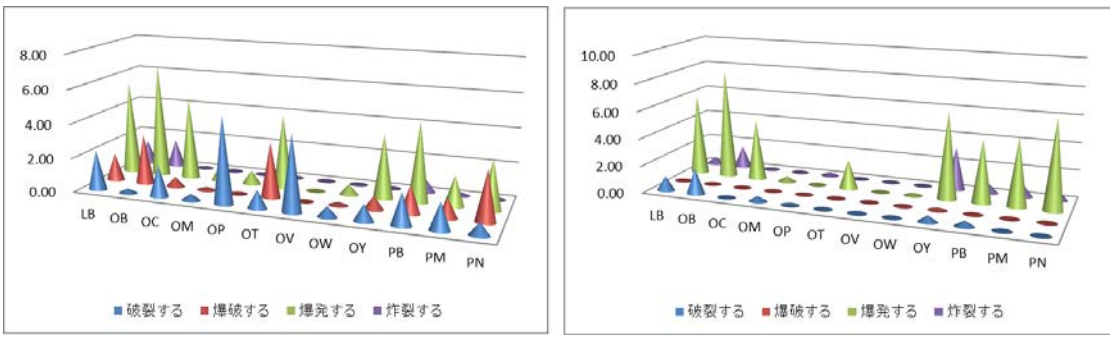


図 2　爆発系語彙の表現メディア別字義的表現(左)とメタファー表現(右)の調整頻度

図 2 に，表現メディア別の 100 万語あたりの調整頻度(PERMIL)を字義的な用法とメタファー用法に分けて示す。横軸の LB，OB 等は BCCWJ のサブコーパスにつけられた記号（表

[1] ある用例がメタファー表現であると認定する基準は，その用例で述べられている対象がモノであるかどうかによる。ここで，「モノ」とは五感のすべてに訴える対象を指す。水は目に見え，叩けば音がするし，触れることも味わい匂いを嗅ぐこともできるので，モノである。一方，光や音はモノではない。したがって，水の様態を表す語彙を光や音に用いればメタファー表現となる(大石 2006)。しかし，モノの爆発に伴って発する光や音に対して爆発系の語彙を用いる場合はメトニミーであり，モノの存在を前提とせず音の発生だけを表現するメタファー表現とは区別される。

2 参照）である。左側のグラフが表す字義的な用法では，「破裂する」「爆破する」「爆発する」は広い範囲の表現メディアで用いられているのに対し，右側のグラフが表すメタファー表現では「爆発する」のみが広く用いられ，「炸裂する」は Yahoo！ブログ(OY)で特徴的に(32 例)用いられていることがわかる。

字義的な用法を観察すると，「破裂する」は広報誌(OP)で「水道管」に多用されるのをはじめとして，薄い膜状のもので覆われている容器の破損を，「爆破する」は，教科書(OT)や新聞(PN)で，「ビル」や「列車」，「大岩」など，重量感を持つ大きな自然物や大規模な構築物について用いられている例が大多数を占めている。これらの，比喩的に使われることが少ない動詞は，爆発物が勢いよく外部に放出されることよりもむしろ，容器の損壊による機能の喪失や，原形の消滅による障害物の除去などを表す文で用いられている。一方，「爆発する」は，電球から超新星まで多様な物体と共起しており，爆発物そのものに焦点が当たっている。また，「炸裂する」は爆発に伴う強烈な音や光による感覚的なショックに焦点があり，相手に対するインパクトを表すメタファー表現として新しく用いられ始めた語彙といえよう。以上のように，メタファーによって表現される対象や知覚状況に焦点が当たっていない語彙は，メタファー表現に用いられにくいという制約を，本稿では「**メタファーの対象焦点化制約**」と呼ぶ。

２.２　氾濫系語彙

氾濫系では，「浸水する」と「冠水する」が，メタファー表現としてはまったく用いられていない。これらは，浸水の深さや水につかった対象の面積に焦点がある語彙であり，メタファーの対象となるべき水を図と考えたときに，地にあたるものに焦点を当てる語彙である。したがって，前節で述べた「爆破する」や「破裂する」と同様，「メタファーの対象焦点化制約」によって水がメタファーによる置き換えの対象となる表現として使用されることがブロックされる。これに対し，「溢れ出る」と「溢れ出す」 は容器に対する水の動きの様態を表し，「氾濫する」は堤防という決められた枠を越えて横溢する水およびその影響に焦点がある。「溢れ出す」「溢れ出る」が「感情」や「気」，「思い」に対して用いられることが多いのに対し，「氾濫する」は「情報」「広告」「雑誌」「言葉」等に対して批判的な含意を伴って用いられる。水のメタファーが様態ごとに使用語彙の制限を受けることは，大石(2006)で詳述した。感情は水の湧出に，情報は広範囲な流通に焦点を当てる語彙を使用するのである。

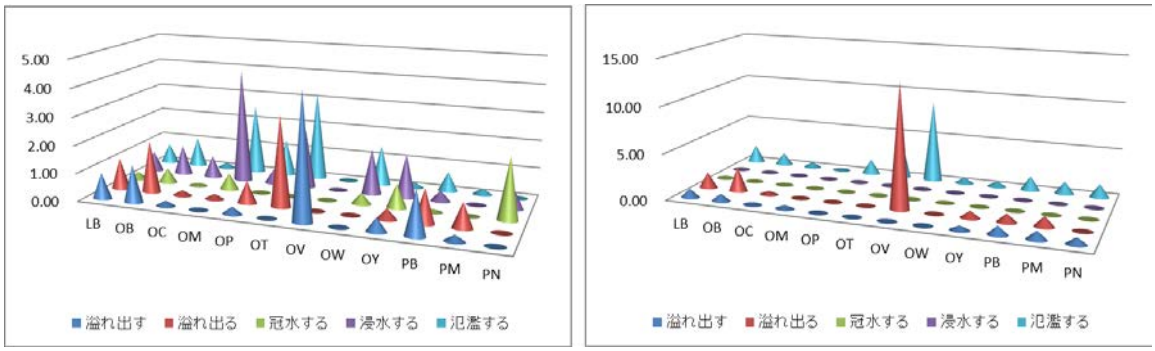


図 3　氾濫系語彙の表現メディア別字義的表現(左)とメタファー表現(右)の調整頻度

図 3 は，表現メディアごとの出現状況である。ここでも，字義的には広い範囲のメディアに用いられている語彙が，メタファー表現では「氾濫する」以外は比較的限られたメディアでのみ用いられていることがわかる。グラフでは，韻文(OV)で「溢れ出る」と「氾濫する」が飛びぬけて用いられているように見えるが，粗頻度はそれぞれ 3 例と 2 例であり，韻文の総語数の少なさが調整頻度の値を突出させ，相対的に他のメディアの山が低く見えている（縦軸目盛に注意）。当然ながら，粗頻度では書籍(LB,PB)の値が高い。

２.３　繁殖系語彙

繁殖系では，比喩的に用いられることが非常に少ない「繁茂する」「繁殖する」に対して，「蔓延する」「蔓延る」の使用はほとんどメタファーであるという明確な対比が見られた。「増殖する」はその中間である。「繁茂」は植物に，「繁殖」は動物や細菌に用いられるという使い分けがあるが，いずれも旺盛な生命力を表現するのに対し，「蔓延する」や「蔓延る」には，病気や風潮が気づかないうちに広まってしまっているという否定的な含意が含まれている。植物の葉が茂るのは日当たりの良いところであるが，日陰では光を求めて蔓ばかりが伸びる。この明暗の違いが両者の評価的意味の根底にあり，そこに雑草の根絶し難さが伴って非常に強い嫌悪感をもたらす。この**印象形成力**がメタファー表現の使用を後押ししたものと考えられる。「増殖する」が抽象物に用いられるときも，コンピュータウイルスや不良債権など，否定的な含意を持つことが多い。このカテゴリーの語彙は，すべて対象の増加を図として取り上げるものであるから，メタファーの対象焦点化制約では説明できない。ここでは表現の持つ印象形成力が，メタファー表現の生産性に影響していると考えておく。

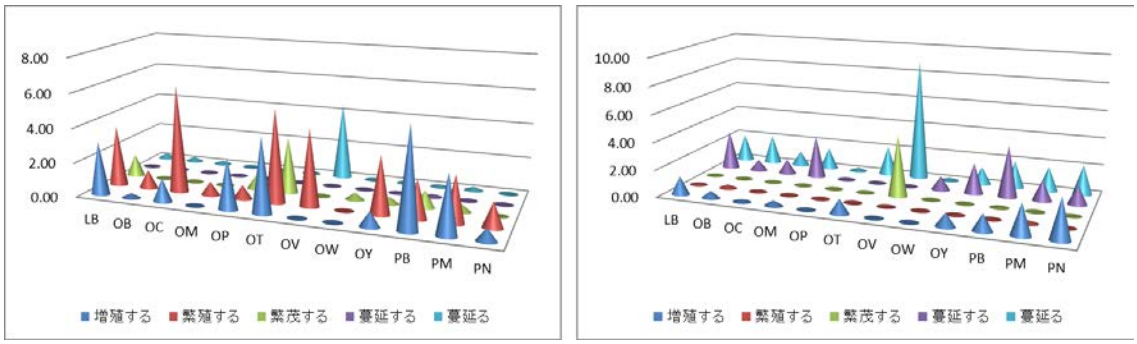


図 4　繁殖系語彙の表現メディア別字義的表現(左)とメタファー表現(右)の調整頻度

図 4 は，表現メディア別の調整頻度であるが，ここでも，「繁殖する」・「繁茂する」と，「蔓延する」・「蔓延る」が，ちょうど左右のグラフで相補的に分布していることがわかる。また，韻文(OV)の粗頻度は，「蔓延る」のメタファー表現が 2 例，「繁茂する」のメタファー表現は 1 例のみであり，前節と同様割り引いて捉える必要がある。

２.４　噴出系語彙

噴出系では，「噴火する」が比喩的用法の割合が最も低い。「噴火する」の字義的用法における共起語のほとんどは火山であり，これは噴出物にとっては容器に対応するものである。したがって，メタファーの対象焦点化制約によって説明できる。

「沸騰する」は「議論」や「世論」に，「噴き出る」は「欲望」「いとしさ」などに，「噴き出す」は「感情」「怒り」「不満」などにそれぞれ用いられているが，「沸騰する」は「水」「お湯」と，「噴き出る」と「噴き出す」は「汗」や「血」と共起する字義的用法の頻度が非常に高いために，メタファー表現の比率は相対的に低くなっている。

これに対し，「噴出する」は「問題」「矛盾」「批判」「議論」「怒り」「不満」といった複数の領域に属する語彙の共起頻度が高く，「溶岩」や「火砕流」といった字義的用法の頻度と肩を並べる勢いである。さらに，「沸き上がる」「沸き起こる」は，どちらも「拍手」「声」「歓声」「どよめき」という音声や，「気持ち」「思い」「感情」「衝動」のような感情，「議論」「疑問」「批判」など，家族的類似性を持つ複数のカテゴリーのメタファー表現が圧倒的多数を占めており，主に「雲」と共起する字義的用法を凌駕している。このように，**字義的用法の頻度と，メタファー用法の多様性**はトレードオフの関係にある。

これは，言語によるコミュニケーションにおける二つの相矛盾する要求のせめぎあいによって説明することが可能である。メタファー表現をはじめとする創造的な表現を使用す

るということは，常に聞き手が意味不明に陥るという危険性をもたらすことになる。字義的な表現にあらたな抽象的な意味を付与することによって意味的な曖昧性が増えるからである。一方，曖昧さのない形でコミュニケーションを図ろうとすると，異なる意味には異なる表現を用いる必要があるとともに，表現の固定化をもたらすことになる。このように，表現の創造性と透明性には本来的なトレードオフがあり，少数の形式で異なる意味を表す形式の節約原理と，異なる意味には異なる形式を用いるという単純原理の二つがせめぎあうことになる。メタファー表現を使うと創造性は上がるが透明性は減るので，慣用化やコロケーションの固定化と表現の使い分けによって透明性を確保する必要があるのである。「沸き上がる」や「沸き起こる」，「噴出する」のように，字義的な用法が「雲」や「溶岩流」など狭い範囲に限定されており，使用頻度もそれほど高くなければ，その表現をメタファー表現に使用しても，曖昧性の出現する機会がそれほど増えることはない。逆に，「噴き出る」「噴き出す」のように字義的な用法の頻度が高ければ高いほど，意味が曖昧になる可能性もまた高まると考えられる。

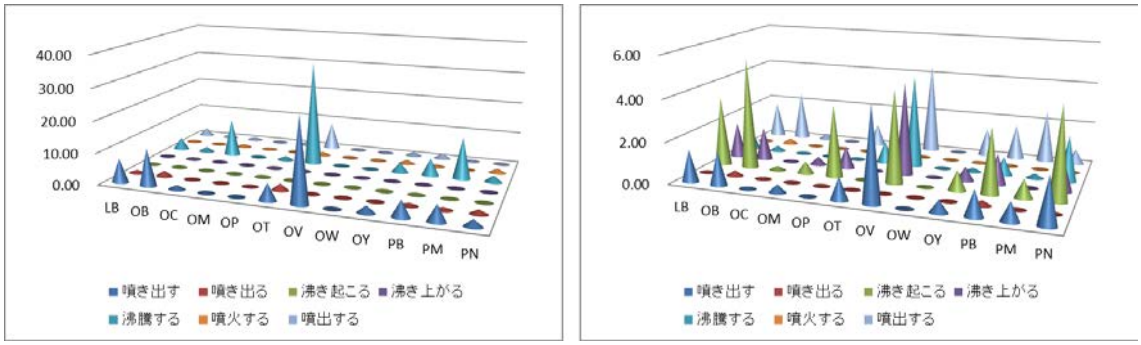


図 5　噴出系語彙の表現メディア別字義的表現(左)とメタファー表現(右)の調整頻度

図 5 は，噴出系語彙の表現メディア別の調整頻度である。左側の字義的用法のグラフは，「沸騰する」が教科書(OT)で多用されており(31 例)，韻文(OV)での「噴き出す」(6 例)とともに PERMIL の値を突出させたために，全体的にフラットに見えてしまっているが，「噴き出す」「噴き出る」「沸騰する」の列が盛り上がっている。これに対し，右側のメタファー表現では，「沸き起こる」「沸き上がる」「噴出する」の列が立ち上がっていることがわかる。

２.５　膨張系語彙

膨張系語彙のメタファー表現では，「肥大する」が「幻想」「欲望」「自我」「自意識」など，比較的狭い範囲の想念に用いられているのに対し，「膨らむ」は，「夢」「希望」「期待」「イメージ」「想像」「妄想」など肯定的な想念と「借金」「赤字」「損失」「債務」など，否定的な意味を持つ経済的概念に用いられている。また，「膨れ上がる」は「人口」「数」「群衆」「規模」「額」など，一般的な数の増大を表すことが多いが，「膨らむ」と同様，「借金」「債務」「赤字」「損失」「支出」などの否定的な経済的概念にも用いられている。これに加えて，「怒り」「欲望」「不安」「恐怖」「感情」「思い」など，どちらかといえば負の意味を持つ感情にも用いられている。最後に，「膨張する」は「医療費」「経費」など経済的費用と「欲望」「感情」などの感情や「人口」「東京」などの都市の拡大に用いられている。このように，字義的には同じような意味を表す４つの語が，メタファー用法でも経済と感情，想念という同様の領域で用いられているが，評価的な意味と表現スタイルにおいて微妙に使い分けられている。

図 6 は，膨張系語彙の調整頻度を表現メディア別に示したものである。ここでは，「膨らむ」の使用頻度が非常に高いために，他の語彙の使用状況がわかりにくくなっている。目立つのは新聞(PN)と雑誌(PM)における「膨らむ」のメタファー用法であるが，「膨らむ」は字義的にもメタファー表現にも広い範囲で用いられており，もっとも基本的な和語基礎

語彙であることを物語っている。「膨張」「肥大」は漢語であり，「膨れ上がる」は複合語である。このような語種と語構成は，全体的な使用頻度に重大な影響を与えているが，メタファー表現の多様性という点では，大きな違いは見られない。むしろ，複合語である「膨れ上がる」がメタファー表現の使用比率が高い。英語においても，メタファー表現に句動詞が用いられることが多いことが Deignan (2005)によって指摘されているが，それと並行する現象といえる。これも，複合語によってメタファー使用の適用範囲を使い分けることによって，曖昧性を減じていると考えることができる。

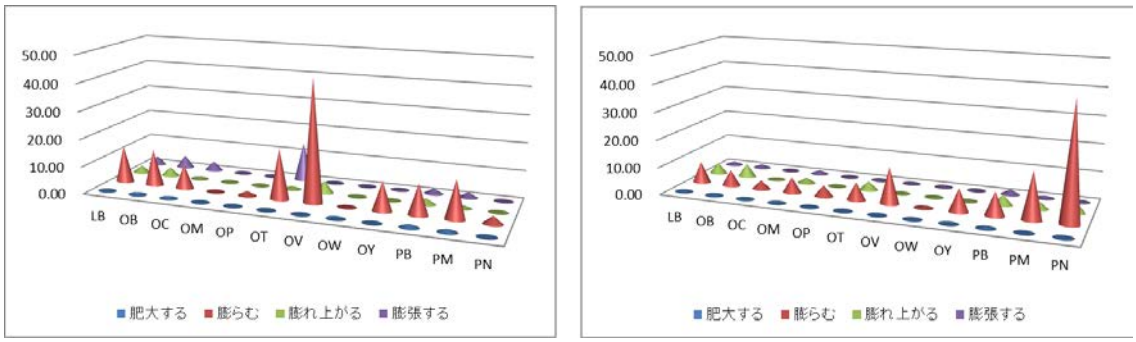


図 6　膨張系語彙の表現メディア別字義的表現(左)とメタファー表現(右)の調整頻度

以上のように，類義語間のメタファー表現の生産性および多様性の違いは，語彙使用状況における焦点の違い，評価的な含意がもたらす印象形成力，字義的な用法の使用範囲と生起頻度など，複数の要因によってもたらされている状況が明らかになった。

３．表現メディア別使用状況の集計

表 2 は，調査対象全語彙の表現メディアごとの粗頻度と 100 万語あたりの調整頻度(PER MIL)およびメタファー表現の字義的な表現に対する割合 (Metaphor/Literal Ratio: MLR) を示したものである。表 2 によると，メタファー表現の調整頻度は新聞(74.44)と韻文(71.02)で高く，法律(0)と白書(3.69)で低い値となっている。しかし，各メディアに出現するメタファー表現のタイプはまったく異なっている。新聞や雑誌では，「夢が膨らむ」のような，トークン頻度の高い定型表現が多用されているのに対し，「女優魂が炸裂する」のような新規なメタファー表現は，小説やブログなど，私的な言説にトークン頻度は低いが，多様に，したがってタイプ頻度は高く出現している。後者は，いわば突然変異のように生まれ，ほとんどは淘汰されて消え去るものであろうが，まれに社会に受け入れられたものが慣用化し，前者のように広く用いられるようになると考えられる。

また，国会会議録や新聞でメタファー・リテラル比(MLR)が 1 を超えており，メタファー表現が字義的表現より多く用いられていることを表している。これは，やや意外な印象を受けるが，全メタファー表現の出現頻度のうち約 4 分の 1 を占めている「膨らむ」が，これらのメディアでは経済的な場面で多く用いられていることが大きな原因である。逆に，MLR が低いのは，Yahoo!知恵袋(0.28)，教科書(0.20)と法律(0)であり，これらのメディアでは，字義的な現象を述べることが，メタファー的な用法よりも多いことが，少なくとも本稿で取り上げた語彙については確認された。

４．松本(2007)との違い

2.1 節と 2.2 節で提示した「メタファーの対象焦点化制約」は，一見すると，第 1 節で述べた松本 (2007)の意味の過剰指定の裏返しにすぎないと考えられるかもしれない。しかし，具体例を考察することによって両者の違いを明確にすることができる。

松本(2007)では，（７）の例を挙げて，メタファーの適用における「漏れる」と「漏る」の違いを取り上げている(65 ページ，例文番号は変更)。

表 2　サブコーパス別出現状況

サブコーパス名（記号）		図書館･書籍(LB)	ベストセラー(OB)	Yahoo! 知恵袋(OC)	国会会議録(OM)	広報誌(OP)	教科書(OT)	韻文(OV)
総語数		30,377,866	3,742,261	10,256,877	5,102,469	3,755,161	928,448	225,273
粗頻度	ﾒﾀﾌｧｰ	1257	166	131	86	45	21	16
	ﾘﾃﾗﾙ	1776	216	468	59	66	106	21
	MLR	0.71	0.77	0.28	1.46	0.68	0.20	0.76
調整頻度	ﾒﾀﾌｧｰ	41.38	44.36	12.77	16.85	11.98	22.62	71.02
	ﾘﾃﾗﾙ	58.46	57.72	45.63	11.56	17.58	114.17	93.22
サブコーパス名（記号）		白書(OW)	Yahoo! ブログ(OY)	法律(OL)	出版･書籍(PB)	雑誌(PM)	新聞(PN)	計
総語数		4,882,812	10,194,143	1,079,146	28,552,283	4,444,492	1,370,233	104,911,464
粗頻度	ﾒﾀﾌｧｰ	18	299	0	1145	180	102	3466
	ﾘﾃﾗﾙ	32	343	0	1518	228	30	4863
	MLR	0.56	0.87	0.00	0.75	0.79	3.40	0.71
調整頻度	ﾒﾀﾌｧｰ	3.69	29.33	0.00	40.10	40.50	74.44	33.04
	ﾘﾃﾗﾙ	6.55	33.65	0.00	53.17	51.30	21.89	46.35

（7）a. 水が｛漏れる／漏る｝。
　　b. 秘密が｛漏れる／*漏る｝。

この違いに対する説明として，松本(2007)は，〈引力による下方向への移動〉という意味が，「漏る」に存在するために，それが過剰指定となって，「漏る」のメタファー的意味を阻止していると論じている。

> 「『漏る』と『漏れる』において重要なのは，『漏る』が引力による下方向への流動体の移動（したたるような移動）に限られるのに対し，『漏れる』にはそのような限定がない，という点である。（中略）『漏る』に見られる〈引力による下方向への移動〉という側面は，流動体の漏出と情報の漏洩との間の対応関係には見られない要素である。この要素の存在ゆえに，『漏る』はこのメタファー的意味を持てないのだ，ということである。」(松本(2007)，66 ページ)

しかし，コーパス中で，「漏る」という動詞が最も多く用いられているのは，「雨漏り」を表す状況である。「屋根が漏る」「天井が漏る」「この茶碗は漏る」という表現が象徴しているように，「漏る」という動詞が焦点を当てているのは，漏り出る水ではなく，容器にあたるものの瑕疵である[2]。たとえ「水が漏る」という表現がされている場合であっても，それは容器が使い物にならないことや，修理しなければならないという状況で用いられているのである。したがって，メタファーによって〈情報〉と置き換えられるべき「水」には二次的な注意しか与えられないために，「漏る」がこのメタファーには用いられないという「メタファーの対象焦点化制約」による自然な説明が可能である。

意味の過剰指定とは，メタファーには不要な意味が字義的な表現に含まれているということであるが，どのような字義的表現にもなんらかのメタファー表現とは異なる意味は含まれているはずである。「メタファーの対象焦点化制約」のほうは，焦点がずれているということであるから，説明の前提としてフレーム意味論的な意味の枠組みを想定している。

---

2 この点については，松本(2007)にも，「『尿が漏る』といえばおむつの問題であるのに対し，『尿が漏れる』の方は失尿の場合にも使える。」(85 ページ，注 7)という指摘がある（下線は本稿著者による）。

その分，適用可能性は減ることになるが，過剰な意味の無制限な認定に歯止めをかけることができるという利点がある。

松本(2007)の語義的経済性の原則は，意味の過剰指定に加えて，他の語の定着性が別の語のメタファー表現を妨げることも述べている。この例としては，「食べる」と「食う」や，「だく」と「いだく」などの例が挙げられている。これらは，メタファー表現と字義的表現を語彙的に使い分けることであるから，2.4 節で述べた創造性と透明性のトレードオフと同趣旨の説明であると考えられる。

しかし，その中にも定着性だけではなく，メタファーの対象焦点化制約によって，より自然に説明できるものが含まれている。それは，（８）に挙げる「歩く」と「歩む」の違いである。(79 ページ，例文番号は変更)

（８）a. 駅まで｛歩く／*歩む｝。
　　b. 孤高の人生を｛*歩く／歩む｝。
　　c. ｛学問／信仰｝の道を｛??歩く／歩む｝。

この例は，二つの類義動詞間で，使い分けがよりはっきり分化している例として取り上げられているものである。ここでの説明は，次のようなものである。

「多くの話者にとって『歩む』は物理的な移動の意味を持たず，メタファー的意味のみを持つ。一方，『歩く』は物理的移動に使われ，メタファー的意味に解釈するのは難しい。（中略）『歩く』と『歩む』の間に，過剰指定の差があるとは考えられない。一番自然な説明は，メタファー専門の語として確立している『歩む』が，『歩く』のメタファー的意味の実現を阻止しているというものである。」(松本(2007)，79 ページ)

しかし，「歩く」と「歩む」の間に本当に意味の違いはないのだろうか。答えは，同論文の中に書かれている。松本(2007)は，「歩む」が死んだメタファーではないことを証明するために，「歩む」が「生きる」などと異なり，「一歩一歩」など，歩行の様態を表す副詞と自然に共起することを述べている。

（９）　人生を一歩一歩踏みしめながら｛歩んでいく／?生きていく｝。
(松本(2007)，79 ページ，例文番号は変更)

すなわち，「歩む」は移動全体ではなく，一歩一歩に焦点が当てられている語なのである。「人生を歩む」というメタファー表現が，抽象的な移動ではなく経験を一つ一つ重ねていくことを表すものであり，移動全体に対する一歩一歩が地と図の関係になっているとすれば，「一歩一歩」に焦点がある「歩む」がこのメタファーに用いられ，移動全体すなわち地に焦点がある「歩く」は用いられないことが，メタファーの対象焦点化制約によって自然に説明される。なお，『精選版　日本国語大辞典』（小学館，2006）の「歩む」の語誌には「類義語『あるく』『ありく』が，足の動作にとどまらぬ移動全体を表すのに対し，『あゆむ』は，一歩一歩足を進めていく動作に焦点がある。」とあり，万葉集や源氏物語では字義的な意味の用例がある。また，「歩く」の語誌には，同様の説明に加えて，「『歩む』が目標を定めた確実な進行であるのに対し，『あるく』『ありく』は散漫で拡散的な移動を表す。」とあり，やはり万葉集の例が挙げられている。したがって，「歩む」が「歩く」のメタファー的意味の実現を阻止しているというよりは，それぞれの字義的な意味の持つ違いによって，よりふさわしい意味へと使い分けがなされてきたというほうがふさわしく，むしろ「歩む」が字義的に使われなくなったことが曖昧性を減じ透明性を高める役割を果たしたと考えるべきであろう。

同じことは，「だく」と「いだく」，「食べる」と「食う」についても言える。『精選版　日

本国語大辞典』の「いだく」の語誌には，「(1)同意語の「むだく」「うだく」「いだく」「だく」の先後関係は，「むだく」が奈良時代から平安初期，「うだく」が平安初期から鎌倉時代頃，「いだく」が平安初期から現代，「だく」が平安中期から現代，という順になる。(2)ダクが①の意味（字義的な意味，大石注）で勢力を拡大していくのに伴って，イダクは次第に③の意味（心の中にある考えや感情を持つこと＝メタファー的意味，大石注）に限定されるようになり，現在に至る。」とあり，「食う」の語誌には，「(1)上代では口にくわえる意での用例が多く，「食」の意にはハムが用いられた。(2)平安時代には，和文脈にクフ，漢文脈にクラフが用いられ，待遇表現としてのタブ（後にタブルを経てタベル）も登場する。(3)室町時代には，クラフが軽卑語，クフが平常語となり，タブルも丁寧語としての用法から平常語に近づいて行った。」とある。いずれも古くから存在する語の字義的な意味が，新しい語によって取って代わられ，古い語はメタファー的意味に特化する方向に変化してきたことを示している。この際，メタファー的意味が新しい表現によって担われないことで，メタファー表現の持つ本来的な曖昧性が抑えられ，意味の透明性が増しているのである。認知的メタファー理論にも以上のような歴史的な視点を取り入れることで，より深い理解が得られるものと期待される。

文　献

Bybee, Joan L. (2010)　*Language, Usage and Cognition*　Cambridge University Press.

Clausner, T. C. & W. Croft (1997) “Productivity and schematicity in metaphors.” *Cognitive Science* 21: pp.247-282.

Deignan, A. (2005) *Metaphor and Corpus Linguistics* John Benjamins B. V.（渡辺秀樹, 大森文子, 加野まきみ，小塚良孝訳(2010)『コーパスを活用した認知言語学』大修館書店）

Grady, J. (1997) “THEORIES ARE BULDINGS revisited.” *Cognitive Linguistics* 8: pp.267-290.

Lakoff, G. (1993) “The contemporary theory of metaphor.” In Andrew Ortony, ed., *Metaphor and Thought*, 2nd ed., pp. 202-251. Cambridge University Press.

Lakoff, G. and M. Johnson (1980) *Metaphors We Live By*, University of Chicago Press. (渡部昇一, 楠瀬淳三,下谷和幸訳 (1986) 『レトリックと人生』大修館書店)

Lakoff, George and Mark Johnson (1999) *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, Basic Books. (計見一雄訳 (2004) 『肉中の哲学: 肉体を具有したマインドが西洋の思考に挑戦する』哲学書房)

黒田航(2005) 「概念メタファーの体系性，生産性はどの程度か: 被害の発生に関係するメタファー成立基盤の記述を通じて」，日本語学，24(6), pp.38-57, 明治書院.

国立国語研究所(2011) 「現代日本語書き言葉均衡コーパス』利用の手引き　第 1.0 版」, 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.

田野村忠温(2012) 「「少納言」「中納言」検索結果活用ツール」，第 1 回コーパス日本語学ワークショップ予稿集，pp.9-11.

鍋島弘治朗(2007) 「黒田の疑問に答える--認知言語学からの回答」,日本語学 26(3), pp.54-71, 明治書院.

鍋島弘治朗(2011) 『日本語のメタファー』，くろしお出版.

大石亨(2006) 「「水のメタファー」再考―コーパスを用いた概念メタファー分析の試み―」，日本認知言語学会論文集第 6 巻(JCLA 6)，pp.277-287.

松本曜(2007) 「語におけるメタファー的意味の実現とその制約」，山梨正明他編，『認知言語学論考』，No.6，pp.49-93,ひつじ書房.

関連 URL

コーパス検索アプリケーション「中納言」　https://chunagon.ninjal.ac.jp/login

書籍テキストへの文体情報付与の試み
—『現代日本語書き言葉均衡コーパス』の収録書籍を対象に—

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）

Annotation of Writing Styles of the Book Samples in the Balanced Corpus of Contemporary Written Japanese

Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Sachi Yasuda (Center for Corpus Development, NINJAL)
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

『現代日本語書き言葉均衡コーパス』(BCCWJ)1の図書館サブコーパスには，書籍テキストが 10,551 サンプル収録されている。大規模な書籍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は，種々の書籍テキストをさまざまな観点から分類できることが望ましい(EAGLES1996)。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はすでにいくつか行われている（小磯ほか 2008，間淵ほか 2010，小磯ほか 2011）。国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」において，書籍テキストを所望の目的で分類するために，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検証とを行っている（柏野・奥村 2012，保田ほか 2012）。

本稿では，はじめにアノテーション作業の概要を述べる。そして，現時点でのアノテー

* waka@ninjal.ac.jp

1 詳細は http://www.tokuteicorpus.jp/。

ションの途中経過報告として，すでに分類指標を付与した 3,494 テキストの分類結果の内訳と，そこから得られた典型例及び，文体の特徴を支える言語的特徴について述べる。そして，本作業を通し，図書館サブコーパスにどのような文体をもつテキストがどのように分布して収録されているのかを把握することができることを示す。

２．アノテーション作業

２.１　分類指標の設計

BCCWJ に収録されている書籍サンプルには, NDC（日本十進分類法）によるジャンルや, C コード（日本図書コード）による販売対象，発売形態，また，著者情報，形態論情報などが付与されており，それらを利用して，半自動的に種々の観点から分類することは可能である。しかしながら，EAGLES(1996)がコーパスへ付与することが望ましいと挙げる，(A)対象読者に想定される読解レベル（難易度），(B)テキストの作成意図，(C)さまざまな文体情報の 3 種に関する情報は C コード以外には与えられておらず，それらの観点によるテキストの分類や抽出は困難である。そこで，(A)を補う「専門度」，(B)を補う「客観度」，(C)を補う「硬度」「くだけ度」「語りかけ性度」という，あわせて 5 つの分類指標を新たに設計した。EAGLES(1996)でコーパスに備えることが望ましいと議論されている「文体情報」とは，形式性，親疎性，口語性に関わる文体情報だと言える。よって，その形式性，親疎性を問うものとして「硬度」と「くだけ度」の指標を，口語性を問うものとして「語りかけ性度」という指標を設けた。

(a) 専門度

テキストの専門度を想定読者のスケールで測ることとし, 次の 5 段階の選択肢を設けた。

1 専門家向き
　読む前提に高度な専門知識が必要なもの
　それを仕事にしているような人向きのもの
2 やや専門的な一般向き
　読む前提に多少の専門知識が必要なもの
3 一般向き
　特に専門的な内容ではないもの
専門的な内容であっても，読む前提に専門的知識を特に必要とせず，一般向きに書かれているもの
4 中高生向き
　中高生向きに書かれているもの
　専門性の有無にかかわらず，中高生でも読めそうなもの
5 小学生・幼児向き
　明らかに小学生や幼児向きとして書かれているもの

このときの「専門性」はテキストを理解する上での「高度な知識の必要性」の有無と考える。たとえば，パソコン関係では，技術者や研究者向きであれば「専門家向き」であり，やや高度な知識を必要とするパソコンを趣味にしている人向きであれば「やや専門的な一般向き」であり，家庭向きであれば「一般向き」と考える。

(b) 客観度

テキストの書き手の意図を捉えるための指標を検討した。「論説，随筆，報告文，紀行文，手順書・・・」といったような体系的な分類案を作成し，それに基づいた指標の付与が理想であるが，指標の設計やその判断にかかる負荷が大きいことが予測されるため，今回の試行作業ではそのようは指標は設けなかった。代わりに，書き手の態度には「客観的」か「主観的」かの区別があると考え，その判断付与を行うこととした。次の 4 段階の選択肢を設けた。

1 とても客観的

2 どちらかといえば客観的

3 どちらかといえば主観的

4 とても主観的

ここで「客観的」とは，主に，事実，観察，論証などが述べてあるもの。誰が読んでも納得できる妥当性が高いもの。「主観的」とは，主に，経験や感想などが述べてあるもの。妥当性は筆者の自由と定義する。なお，これはノンフィクションと判断をしたテキストについてのみ付与する。

(c) 硬度

テキストの文体の形式性，親疎性を捉えるために「硬いか軟らかいか」を判断することとした。「硬い」とは，かしこまっている感じ，堅苦しい感じであり，「軟らかい」とは，かしこまっていない感じ，親しみやすい感じである。次の 4 段階の選択肢を設けた。

1 とても硬い

2 どちらかといえば硬い

3 どちらかといえば軟らかい

4 とても軟らかい

(d) くだけ度

テキストの文体の形式性，親疎性を捉えるためのもう一つの指標として，さらに「くだけているか」を問うこととした。「くだけているか」の逆には「改まっているか」を想定するが，「改まっているか」の度合いは問いにくいと考えた。よって，ここでは「くだけている」度合いを問う，次の 3 段階の選択肢を設けた。

1 とてもくだけている

2 どちらかといえばくだけている

3 くだけていない（＝改まっている）

(e) 語りかけ性度

テキストの文体の口語性を問うものとして「語りかけ性度」という指標を設けた。口語性の高いテキストを「語りかけ性がある」ものと捉えることとした。たとえば，「あなた」や「みなさん」などの呼びかけ表現や，「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など，読み手に直接的に語りかけているような表現があるものを，「語りかけ性がある」ものと考える。それら「語りかけ性」の度合を問う選択肢として次の 3 段階のものを設けた。

1 とても語りかけ性がある

2 どちらかといえば語りかけ性がある

3 特に語りかけ性はない

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。また，アノテーション作業は，次のとおり二段階で進めている。

[作業対象と内容]

- 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サンプル全体を範囲とする。1テキストの平均はおよそ3,000語。
- 作業ファイル：サンプルを取得した書籍の紙面コピーの電子化ファイルを参照する。

●作業量：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：
①形式による判定を行う。構造的に単純なテキストタイプ（例：章節構造）であれば細分類の対象とする[2]。
②細分類をする。「専門度，客観度，硬度，くだけ度，語りかけ性度」の分類指標を付与する。

[作業の一段階目]
●目的:人手付与の作業上の問題点の検討,典型例の抽出,分類指標の検証及び基準の検討。
●態勢：判断のゆれを検証するために同一サンプルを作業者3人で判定。
●判断：付与すべき指標の種類についてごく簡単な説明があるのみ。
●付与済テキスト数：3,324テキストを判定し，細分類を付与したのは2,672テキスト。

[作業の二段階目（途中）]
●目的：全10,551サンプルへの付与。
●態勢：1サンプル1人以上の判定と機械判定の相互参照。
●判断：判断事例付きマニュアルを参照。
●付与済テキスト数：本稿作成時，細分類の付与済みは，現時点3,494テキスト。

３．アノテーション作業結果

３.１　分類指標の付与結果

細分類付与済みの 3,494 テキストについて，その付与結果の内訳を表 1 に示す。なお，客観度はノンフィクションのみ対象としているため，現時点の付与済みテキスト数は 2,314 である。

表 1 分類指標の付与結果 (3,494 テキスト)

専門度	テキスト数・(%)	客観度	テキスト数・(%)	硬度	テキスト数・(%)	くだけ度	テキスト数・(%)	語りかけ性度	テキスト数・(%)
専門家向き	91(3%)	とても客観的	427(18%)	とても硬い	207(6%)	とてもくだけている	186(5%)	とても語りかけ性がある	326(9%)
やや専門的な一般向き	345(10%)	どちらかといえば客観的	904(39%)	どちらかといえば硬い	1135(32%)	どちらかといえばくだけている	1020(29%)	どちらかといえば語りかけ性がある	572(16%)
一般向き	2685(77%)	どちらかといえば主観的	623(27%)	どちらかといえば軟らかい	1834(52%)	くだけていない	2288(65%)	特に語りかけ性はない	2596(74%)
中高生向き	195(6%)	とても主観的	360(16%)	とても軟らかい	318(9%)				
小学生・幼児向き	178(5%)								

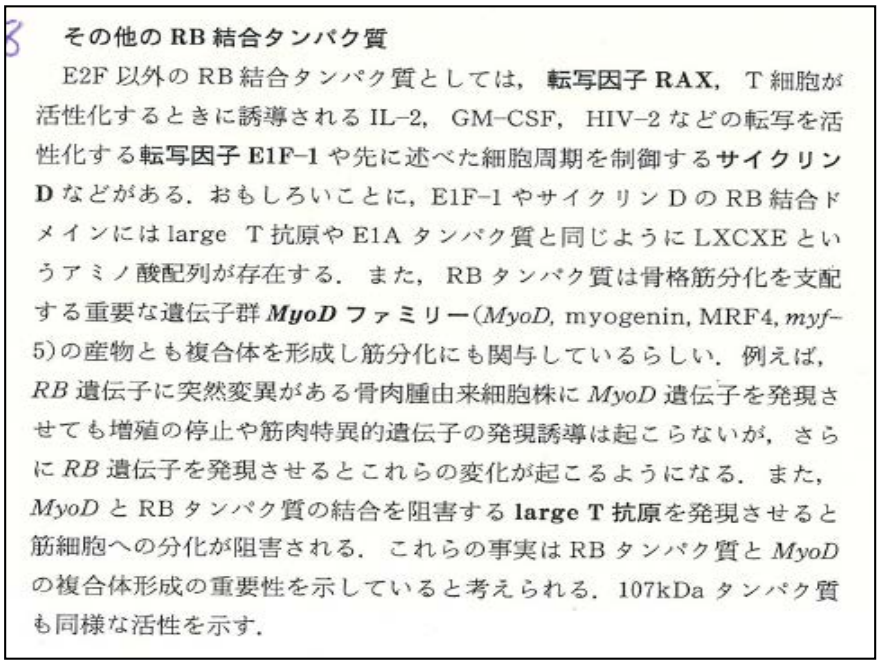
専門度は「一般向き」が多い。客観度も真ん中あたりが多いが，そのうち，「どちらかといえば客観的」の方が多い。硬度はさらに真ん中あたりが多いが，そのうち，「どちらかといえば軟らかい」の方が多い。くだけ度は約 3 分の 1 近く「どちらかといえばくだけている」である。語りかけ性度は「とても」と「どちらかといえば」をあわせて約 4 分の 1 である。

[2] 対象外とした形式が特徴的なテキスト（例：対談，Q&A 形式，図解，用語解説）については，一定量が分類されてから細分類を検討する予定でいる。

３.２　分類の典型例

ここでは, [作業の二段階目]で使用したマニュアルに掲載した典型例（サンプルの出典は, BCCWJのサンプルIDと書名とで記す）を示す。

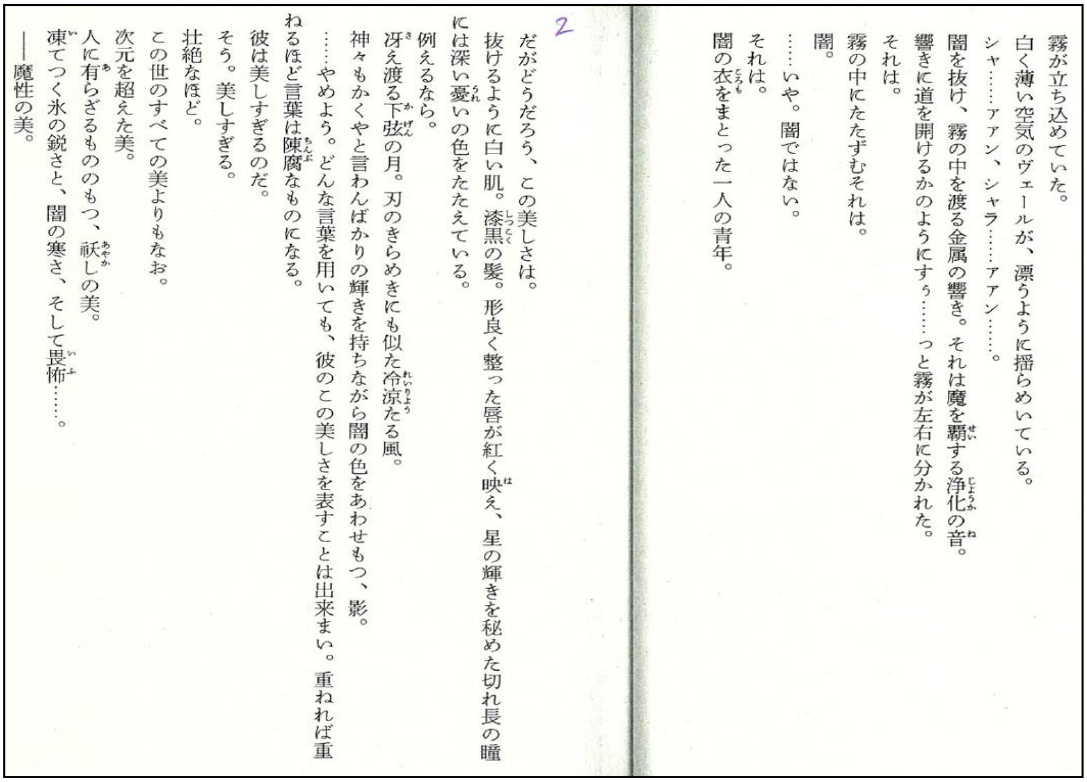
(1)専門度：1 専門家向き　　（LBi4_00021『がんと遺伝子』）



その他の RB 結合タンパク質

E2F 以外の RB 結合タンパク質としては，**転写因子 RAX**，T 細胞が活性化するときに誘導される IL-2，GM-CSF，HIV-2 などの転写を活性化する**転写因子 E1F-1** や先に述べた細胞周期を制御する**サイクリン D** などがある．おもしろいことに，E1F-1 やサイクリン D の RB 結合ドメインには large T 抗原や E1A タンパク質と同じように LXCXE というアミノ酸配列が存在する．また，RB タンパク質は骨格筋分化を支配する重要な遺伝子群 ***MyoD* ファミリー**(*MyoD*, myogenin, MRF4, *myf*-5)の産物とも複合体を形成し筋分化にも関与しているらしい．例えば，*RB* 遺伝子に突然変異がある骨肉腫由来細胞株に *MyoD* 遺伝子を発現させても増殖の停止や筋肉特異的遺伝子の発現誘導は起こらないが，さらに *RB* 遺伝子を発現させるとこれらの変化が起こるようになる．また，*MyoD* と RB タンパク質の結合を阻害する **large T 抗原**を発現させると筋細胞への分化が阻害される．これらの事実は RB タンパク質と *MyoD* の複合体形成の重要性を示していると考えられる．107kDa タンパク質も同様な活性を示す．

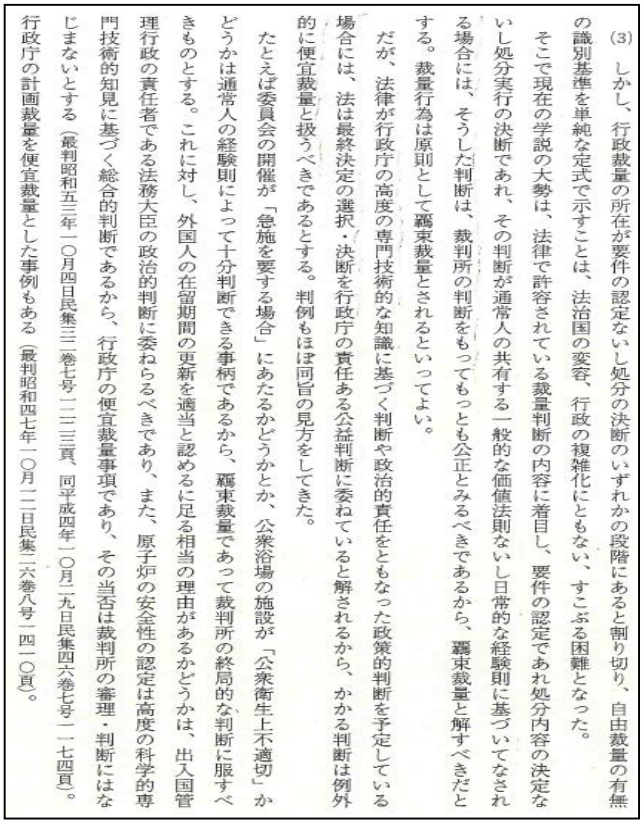
(2)専門度：4 中高生向き　　（LBf9_00090『超魔炎獄変』）



霧が立ち込めていた。
白く薄い空気のヴェールが、漂うように揺らめいている。
シャ……ァァン、シャラ……ァァン……。
闇を抜け、霧の中を渡る金属の響き。それは魔を覇する浄化の音。
響きに道を開けるかのようにすぅ……っと霧が左右に分かれた。
それは。
霧の中にたたずむそれは。
闇。
……いや。闇ではない。
それは。
闇の衣をまとった一人の青年。

だがどうだろう、この美しさは。
抜けるように白い肌。漆黒の髪。形良く整った唇が紅く映え、星の輝きを秘めた切れ長の瞳には深い憂いの色をたたえている。
例えるなら。
冴え渡る下弦の月。刃のきらめきにも似た冷涼たる風。
神々もかくやと言わんばかりの輝きを持ちながら闇の色をあわせもつ、影。
……やめよう。どんな言葉を用いても、彼のこの美しさを表すことは出来まい。重ねれば重ねるほど言葉は陳腐なものになる。
彼は美しすぎるのだ。
そう。美しすぎる。
壮絶なほど。
この世のすべての美よりもなお。
次元を超えた美。
人に有らざるもののもつ、祇しの美。
凍てつく氷の鋭さと、闇の寒さ、そして畏怖……。
――魔性の美。

(3)客観度：1 とても客観的
（LBo3_00158『行政法要論』）



(3) しかし、行政裁量の所在が要件の認定ないし処分の決断のいずれかの段階にあると割り切り、自由裁量の有無の識別基準を単純な定式で示すことは、法治国の変容、行政の複雑化にともない、すこぶる困難となった。

そこで現在の学説の大勢は、法律で許容されている裁量判断の内容に着目し、要件の認定であれ処分内容の決定ないし処分実行の決断であれ、その判断が通常人の共有する一般的な価値法則ないし日常的な経験則に基づいてなされる場合には、そうした判断は、裁判所の判断をもってもっとも公正とみるべきであるから、羈束裁量と解すべきだとする。裁量行為は原則として羈束裁量とされるといってよい。

だが、法律が行政庁の高度の専門技術的な知識に基づく判断や政治的責任をともなった政策的判断を予定している場合には、法は最終決定の選択・決断を行政庁の責任ある公益判断に委ねていると解されるから、かかる判断は例外的に便宜裁量と扱うべきであるとする。判例もほぼ同旨の見方をしてきた。

たとえば委員会の開催が「急施を要する場合」にあたるかどうかか、公衆浴場の施設が「公衆衛生上不適切」かどうかは通常人の経験則によって十分判断できる事柄であるから、羈束裁量であって裁判所の終局的な判断に服すべきものとする。これに対し、外国人の在留期間の更新を適当と認めるに足る相当の理由があるかどうかは、出入国管理行政の責任者である法務大臣の政治的判断に委ねらるべきであり、また、原子炉の安全性の認定は高度の科学的専門技術的知見に基づく総合的判断であるから、行政庁の便宜裁量事項であり、その当否は裁判所の審理・判断にはなじまないとする（最判昭和五三年一〇月四日民集三二巻七号一三三頁、同平成四年一〇月二九日民集四六巻七号一一七四頁）。行政庁の計画裁量を便宜裁量とした事例もある（最判昭和四七年一〇月二日民集二六巻八号一四一〇頁）。

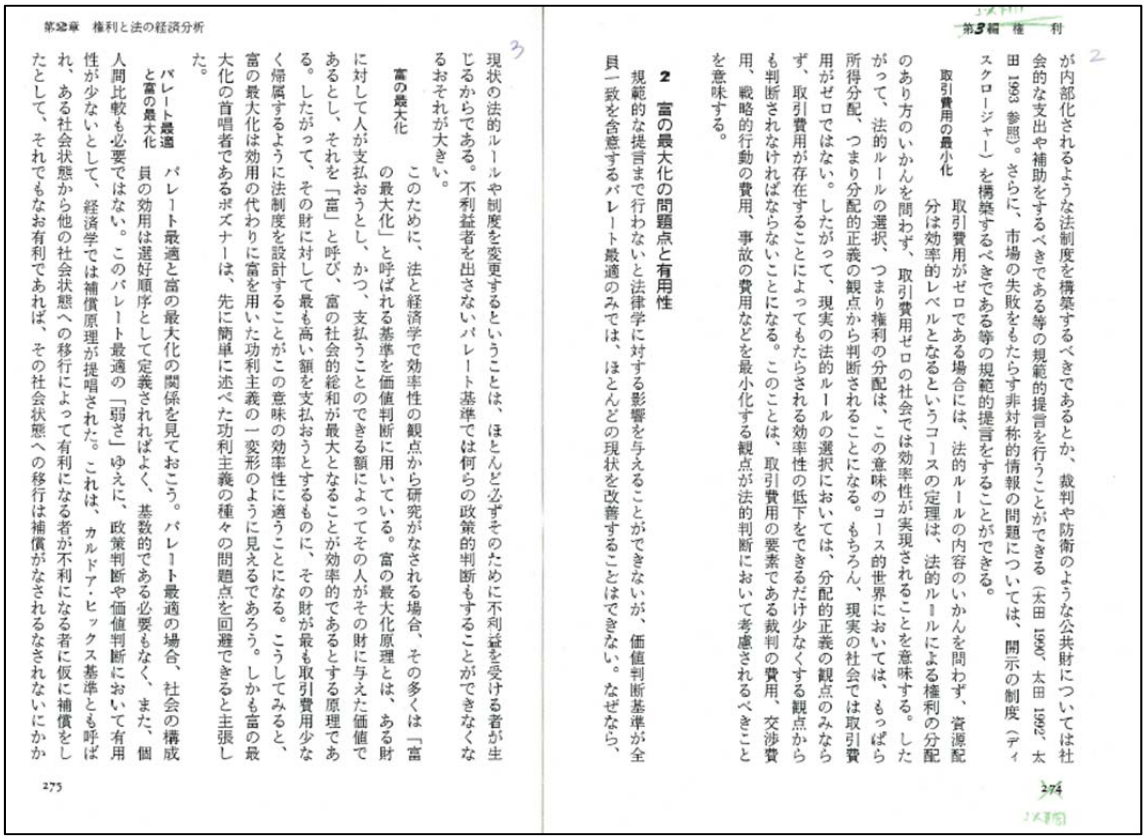
(4)客観度：4 とても主観的
（LBo3_00132『教師をめざす若者たち』）

人形は、私と木下さんを大胆にさせてくれました。人形を持った私たちは子供たちのなかに自然に入っていけました。教壇から降り、人形を片手にして、子供たちと触れ合ったのです。子供たちは人形に触れ、私も人形を通して子供たちの顔や体に触れていきました。私と子供たちの間に、大きな橋が架かったのです。言葉が互いに通じ合わないからこそ、見えてくるものがたくさんありました。言葉は、嘘をつくこともできるし、自分を隠すこともできます。しかし、心は嘘をつけないことを実感したのです。

どんなに上手な言葉を使っても、思っていないことを発すれば、子供に伝わらない。どんなに下手な言葉でも、心から伝えたいという愛情があれば、伝わるものであるということを信じることができました。この実感は日本でも通じる「教育の原則」であると思いました。

二日目、子供たちと綿花摘みを一緒にしました。敦煌の子供たちの手は「仕事をしている手」でした。その表情を持った手が「一緒に綿花を摘もうね」と私の手を引いてくれたとき、何かが私に伝わってきました。小さなその手から、人間としての強さが伝わってきました。それは、この子供たちが、それまでに培ってきた力だと感じました。

(5)硬度：1 とても硬い（LBi3_00033『現代法社会学入門』）



第3章 権利

が内部化されるような法制度を構築するべきであるとか、裁判や防衛のような公共財については社会的な支出や補助をするべきである等の規範的提言を行うことができる（太田 1993、太田 1995、太田 1983 参照）。さらに、市場の失敗をもたらす非対称的情報の問題については、開示の制度（ディスクロージャー）を構築するべきである等の規範的提言をすることができる。

取引費用の最小化　取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。したがって、現実の法的ルールの選択においては、分配的正義の観点のみならず、取引費用が存在することによってもたらされる効率性の低下をできるだけ少なくする観点からも判断されなければならないことになる。このことは、取引費用の要素である裁判の費用、交渉費用、戦略的行動の費用、事故の費用などを最小化する観点が法的判断において考慮されるべきことを意味する。

2 富の最大化の問題点と有用性

規範的な提言まで行わないと法律学に対する影響を与えることができないが、価値判断基準が全員一致を含意するパレート最適のみでは、ほとんどの現状を改善することはできない。なぜなら、

274

第2章 権利と法の経済分析

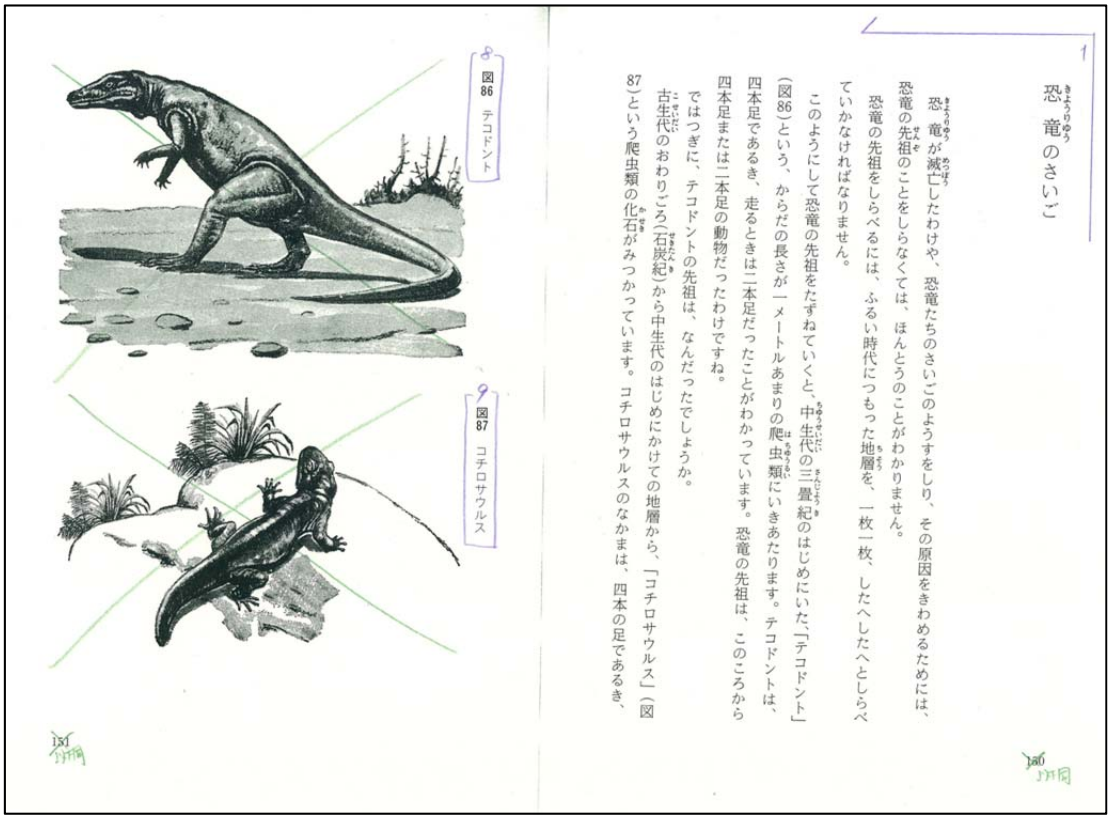
現状の法的ルールや制度を変更するということは、ほとんど必ずそのために不利益を受ける者が生じるからである。不利益者を出さないパレート基準では何らの政策的判断もすることができなくなるおそれが大きい。

富の最大化　このために、法と経済学で効率性の観点から研究がなされる場合、その多くは「富の最大化」と呼ばれる基準を価値判断に用いている。富の最大化原理とは、ある財に対して人が支払おうとし、かつ、支払うことのできる額によってその人がその財に与えた価値であるとし、それを「富」と呼び、富の社会的総和が最大となることが効率的であるとする原理である。したがって、その財に対して最も高い額を支払おうとするものに、その財が最も取引費用少なく帰属するように法制度を設計することがこの意味の効率性に適うことになる。こうしてみると、富の最大化は効用の代わりに富を用いた功利主義の一変形のように見えるであろう。しかも富の最大化の首唱者であるポズナーは、先に簡単に述べた功利主義の種々の問題点を回避できると主張した。

パレート最適と富の最大化　パレート最適と富の最大化の関係を見ておこう。パレート最適の場合、社会の構成員の効用は選好順序として定義されればよく、基数的である必要もなく、また、個人間比較も必要ではない。このパレート最適の「弱さ」ゆえに、政策判断や価値判断において有用性が少ないとして、経済学では補償原理が提唱された。これは、カルドア・ヒックス基準とも呼ばれ、ある社会状態から他の社会状態への移行によって有利になる者が不利になる者に仮に補償をしたとして、それでもなお有利であれば、その社会状態への移行は補償がなされるなされないにかか

273

(6)硬度：4 とても軟らかい（LBa4_00010『恐竜の世界をたずねて』）



恐竜のさいご

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへしたへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三畳紀のはじめにいた、「テコドント」（図86）という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

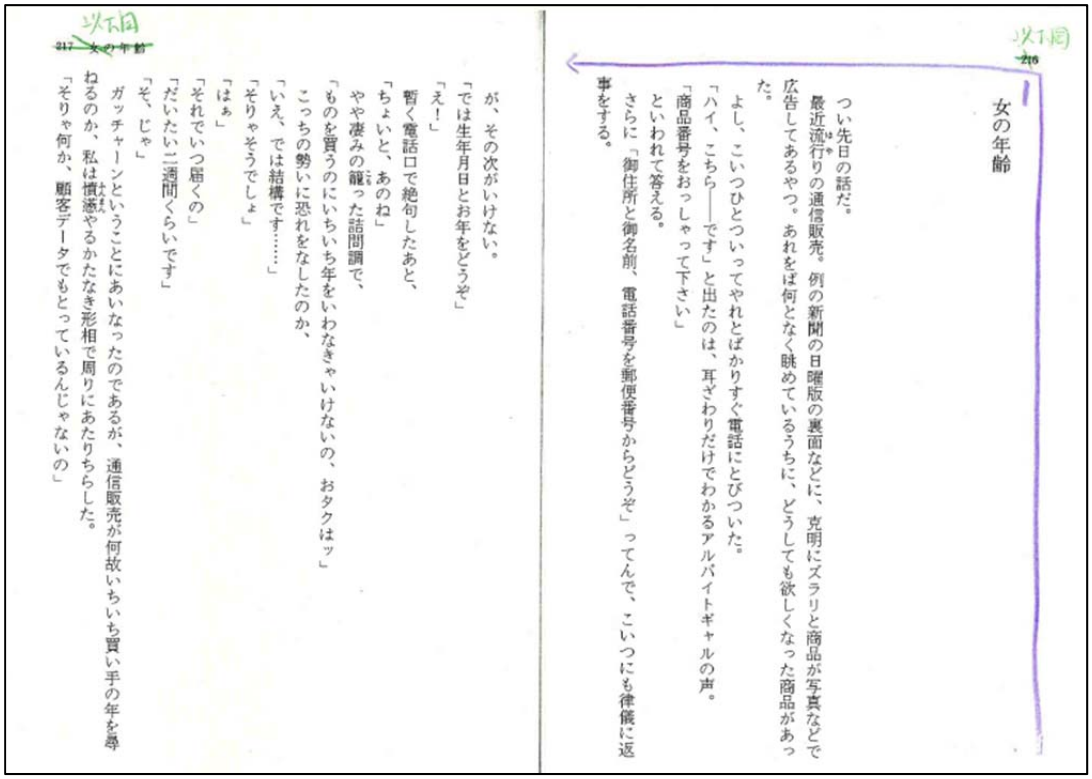
ではつぎに、テコドントの先祖は、なんだったでしょうか。

古生代のおわりごろ（石炭紀）から中生代のはじめにかけての地層から、「コチロサウルス」（図87）という爬虫類の化石がみつかっています。コチロサウルスのなかまは、四本の足であるき、

図86 テコドント

図87 コチロサウルス

(7) 1 とてもくだけている（LBf9_00067『男はオイ！女はハイ…』）



女の年齢

つい先日の話だ。

最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをば何となく眺めているうちに、どうしても欲しくなった商品があった。

よし、こいつひとついってやれとばかりすぐ電話にとびついた。

「ハイ、こちら――です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。

「商品番号をおっしゃって下さい」

といわれて答える。

さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

が、その次がいけない。

「では生年月日とお年をどうぞ」

「え！」

暫く電話口で絶句したあと、

「ちょいと、あのね」

やや凄みの籠った詰問調で、

「ものを買うのにいちいち年をいわなきゃいけないの、おタクはッ」

こっちの勢いに恐れをなしたのか、

「いえ、では結構です……」

「そりゃそうでしょ」

「はぁ」

「それでいつ届くの」

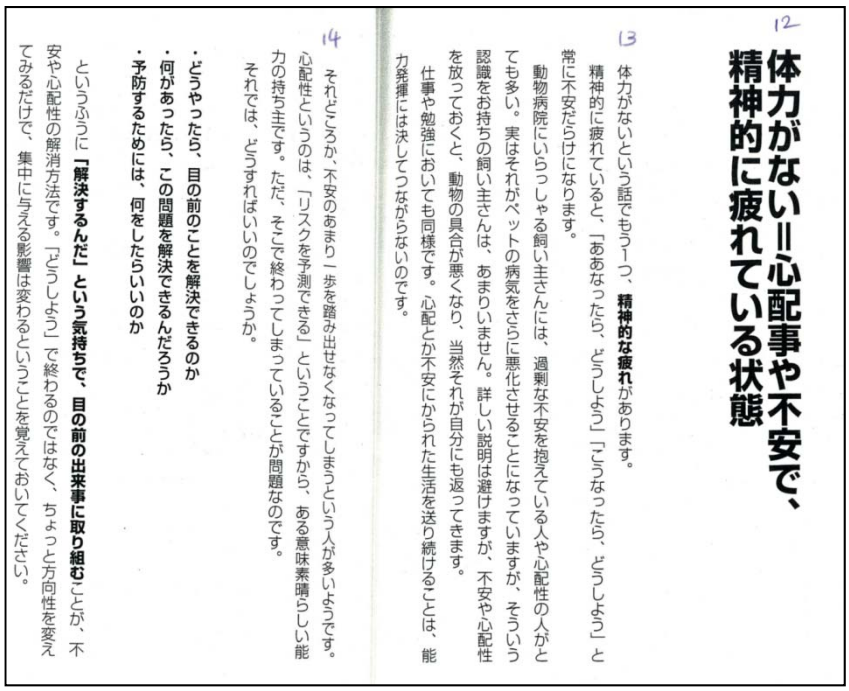
「だいたい二週間くらいです」

「そ、じゃ」

ガッチャーンということにあいなったのであるが、通信販売が何故いちいち買い手の年を尋ねるのか、私は憤懣やるかたなき形相で周りにあたりちらした。

「そりゃ何か、顧客データでもとっているんじゃないの」

(8) 1 とても語りかけ性がある（LBt1_00013『5 分間集中力トレーニング』）



12
体力がない＝心配事や不安で、精神的に疲れている状態

13
体力がないという話でもう一つ、**精神的な疲れ**があります。
精神的に疲れていると、「ああなったら、どうしよう」「こうなったら、どうしよう」と常に不安だらけになります。
動物病院にいらっしゃる飼い主さんには、過剰な不安を抱えている人や心配性の人がとても多い。実はそれがペットの病気をさらに悪化させることになっていますが、そういう認識をお持ちの飼い主さんは、あまりいません。詳しい説明は避けますが、不安や心配性を放っておくと、動物の具合が悪くなり、当然それが自分にも返ってきます。
仕事や勉強においても同様です。心配とか不安にかられた生活を送り続けることは、能力発揮には決してつながらないのです。

14
それどころか、不安のあまり一歩を踏み出せなくなってしまうという人が多いようです。
心配性というのは、「リスクを予測できる」ということですから、ある意味素晴らしい能力の持ち主です。ただ、そこで終わってしまっていることが問題なのです。
それでは、どうすればいいのでしょうか。

・どうやったら、目の前のことを解決できるのか
・何があったら、この問題を解決できるんだろうか
・予防するためには、何をしたらいいのか

というふうに**「解決するんだ」という気持ちで、目の前の出来事に取り組む**ことが、不安や心配性の解消方法です。「どうしよう」で終わるのではなく、ちょっと方向性を変えてみるだけで、集中に与える影響は変わるということを覚えておいてください。

3.3 文体の特徴を支える言語的特徴

5 つの分類指標のうち，硬度，くだけ度，語りかけ度を判定する際に手掛かりとなりそうな言語的特徴を，これまでの作業で得られている典型例の範囲で分析した。判定基準とするにはさらに分析が必要であるが，現段階に捉えられた特徴を以下に述べる。

硬度判定の参考情報を次の表 2 に示す。

表 2 硬度判定の参考情報

	硬い	軟らかい
和語：漢語（平均は 6：2）	5:3	7:1
語彙	難解	平易
	新密度の低い語（学術・専門用語）	接頭辞「御（お・ご）」
	感動詞ほとんどなし	副助詞（～か，たり，や，まで等）
	数詞が多い	
	文語助動詞（べし）	
副詞	いかに，より等のかしこまった語	バリエーションが豊富
	出現頻度は平均より低い	
接続詞	ないし，いかに等のかしこまった語	
	使用率は平均より少し高め	
文末	終助詞はほとんどなし	です・ます
		「である」はほとんどなし
主語・述語	抽象物の主語＋受動態の述語	
その他	疑問・回答が対応	記号が多い
	断定・定義を表す文	

くだけた印象を与えるサンプルに特徴的な点を以下に列挙する。

- 述語省略など，文法的に破格の文がある
- 一人称が主語の文が多い
- 平易な語に加え，俗語がある
- 音変化（拗音化，撥音化など）の語がある
- オノマトペが多い
- 感覚や感情表現が多い
- 回答のない，いいっぱなしの疑問文がある
- 終助詞（ね，よ等）がある
- 「～だ。」で終わる文が平均より多め
- 外来語が多い
- 副詞はバリエーションが豊富で，出現頻度が平均より高い

最後に，語りかけ性度判定の参考情報を表 3 に示す。

表 3 語りかけ性度判定の参考情報

	語りかけ性がある	語りかけ性はない
語種		固有名詞（人名・地名など）が多い
語	「あなた，みなさん」などの呼びかけ	
	「自分」が多い	
文末	「ます」／終助詞「ね」が多い	「た」が多い
	意志推量「だろう，でしょう」などが多い	
	体言化（「～のです，～ということだ」など）が多い	

４．おわりに

BCCWJ に収録する書籍コーパスの有効活用を可能とするための分類指標の人手付与作業の概要と，作業の途中経過を報告した。

人手判定の一方で，表層的な情報を利用した機械判定を試みている。Support Vector Regression によるランキングを行ったところ，機械判定と人手判定の相関は見られた。中でも，専門度，硬度，くだけ度の相関は高かった。現在，機械判定と人手判定のずれを対照させ，人手判定の見直しを行っている。さらに人手判定の基準を明確にすることで，機械判定の精度向上も目指したい。

今年度がプロジェクトの最終年度である。最終成果として，分類指標の明確な基準を示すとともに, BCCWJ の図書館サブコーパスに収録される 10,551 サンプルの全てに分類指標を付与し，コーパスの研究や教育の利用価値を高めることを目指す。

さらに文体的な特徴を支える言語表現の分析を進め，辞書記述への応用を考えている。

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものです。また，BCCWJ の構築は，文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度，領域代表者：前川喜久雄）による補助を得たものです。

文　献

EAGLES. 1996. EAGLES Preliminary recommendation on Text Typology, *EAGLES Document EAG–TCWG–TTYP/P*, Version of Jun 1996. (http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html)

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』B5-6.

小磯花絵，小木曽智信，小椋秀樹，冨士池優美，宮内佐夜香(2008)「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』pp.192-195.

小磯花絵，田中弥生，小木曽智信，近藤明日子(2011)「評定実験に基づくテキスト分類尺度の体系化の試み」『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集，pp.47-52.

間淵洋子，柏野和佳子，山口昌也，高田智和(2010)「コーパスを用いたテキスト分類指標の検討—BCCWJ の文書構造情報分析を中心に—」『言語処理学会第 16 回年次大会予稿集』PA1-11.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012)「「語り性」を有する書きことばの典型例の分析」『第 1 回コーパス日本語学ワークショップ』予稿集， pp.139-146.

極性反義語の用例分布とその解釈

The Distribution of Polar Antonyms in Corpora and its Implications

服部匡(同志社女子大学表象文化学部)
Tadasu Hattori (Doshisha Women's College of Liberal Arts)

1. はじめに

極性反義語(polar antonym)とは、「大きい-小さい」のような対関係にある語である(Cruse(1986))。対のそれぞれを大値語、小値語と呼ぶことにする。これらの語対は、基本的用法では、例えば、(1)には必ずしも前提がないが(2)には対象が小さいという前提があるといった非対称性を示す。これらの語はまた、基本的用法の他にさまざまな抽象的・比喩的用法を持っている。

(1) どのくらいの大きさ?
(2) どのくらいの小ささ?

服部(2011a,b,2012)では、程度的な側面を持つ名詞と共起する場合の量的形容詞類(多くは極性反義語にあたる)の種類別の出現傾向を分析したが、本発表では、観点を変えて、(基本的用法での)極性反義関係を軸とした用例分布の観察・分析を行う。関連した問題は、西尾(1972)、久島(2001,2002)、鍋島(2011)などで扱われている。

本発表では(3)の各形容詞対[1]をとりあげ、それぞれ、大値語・小値語の用例数比率が、主語名詞の種類によってどのような相違を示すかを観察分析する。比較のため、存在表現の「ある-ない」、および、「濃厚(ダ)-希薄(ダ)」についても観察する。

(3) 大きい-小さい、高い-低い、強い-弱い、多い-少ない、濃い-薄い、広い-狭い、深い-浅い、長い-短い、太い-細い

コーパスとしては、新聞記事のデータ[2](テキストとして約 6GB)を使用する。その理由は、データのサイズが大きいことと、抽象的な名詞の出現頻度が高いことである。

2. 分析手順

分析の手順は次のとおりである。各新聞記事データに MeCab(0.994)と電子辞書 UniDic(1.3.12) による形態素(短単位)解析を施し、プログラムと手作業によって(4)に当たる用例を調査対象として抽出する。

(4) **名詞類＋{が/は}＋形容詞(終止連体形)**
ただし当該名詞類を終端とする名詞句が形容詞と主述関係を構成している例に限る。
名詞類とは、名詞・名詞性接尾辞の短単位要素を指し形式的なものも含む。

こうして得られた用例数を、名詞類・形容詞の組により集計する。詳細なデータは、服部(準備中)を参照されたい。

---

1 「多い」では(1)の場合「多さ」ではなく「数」などが用いられる。また「薄い」は「濃い」の他に「厚い」等とも、また、「高い」は「低い」の他に「安い」とも反義関係を持つ。
2 毎日新聞(1999-2005 年)、読売新聞(1987-1991 年、2000-2008 年)、朝日新聞(1988-1998 年)の記事データ。各新聞社の許諾を得て研究用に利用している。

3. 極性反義語の用例分布

各形容詞対との合計共起用例数の多い名詞類を 50 ずつ取り出し[3]、各名詞類に対する、大値語・小値語の用例数の比率を示すと以下の各図のようになる。数値は、大値語の比率(大値語用例数/(小値語用例数+大値語用例数))である。分布の傾向により、タイプに分けて示す。

多くの形容詞対では大値語との共起に偏る名詞類が多いが、「薄い-濃い(厚い)」の対では、逆に、小値語との共起に偏る名詞類が多い。また、「広い-狭い」と「太い細い」は中間的であり、大値語と小値語に関して対称的に近い分布を示す。

（A） 大値語傾斜型

多くの形容詞対はこのタイプの分布を示す。このタイプでは、大値語とのみ、または主に大値語と共起する名詞類が大部分である。たとえば「小さい」ものは注意を引きにくく「大きい」ものは目立つというように、極性反義語(の、少なくとも、基本的用法)では大値語の方が目立つ特性を表わしていることを反映したものと思われる。

中でも、特に「強い-弱い」と「深い-浅い」では大値語との共起例しかない名詞類の比率が高い。偶然ではなく語の特徴として大値語としか結びつかないものが多いようである(図1・図 2)。「強い」での当該名詞は、多く、心情や感じに関わるものである[4]。

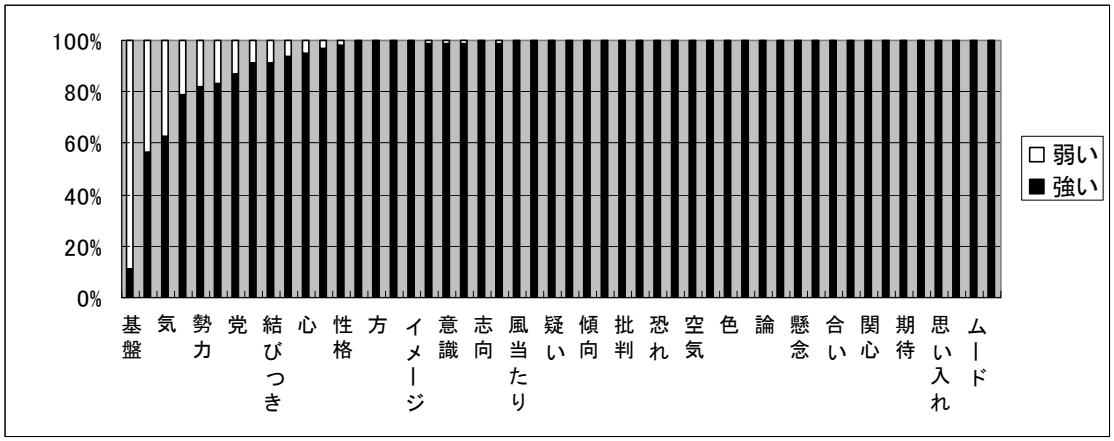


図 1 「強い/弱い」と各名詞類との共起傾向(用例数：11,594 から 285 まで)

名詞類(左から)：基盤, 力, 気, 毒性, 勢力, 風, 党, 面, 結びつき, 意欲, 心, 影響, 性格, 性, 方, 感, イメージ, 印象, 意識, 気持ち, 志向, 抵抗, 風当たり, 見通し, 疑い, 声, 傾向, 思い, 批判, 反発, 恐れ, 意見, 空気, 要望, 色, 不満, 論, 色彩, 懸念, 側面, 合い, 要素, 関心, 反対, 期待, 不安, 思い入れ, 甘み, ムード, 気分

[3] ただし用例数の少ない形容詞対については、50 よりやや少ない数まで。

[4] この場合、「強くある」という表現が存在する。「強い」(の一部)は、「多い」と同様、一種の存在表現とみなしうる可能性がある。

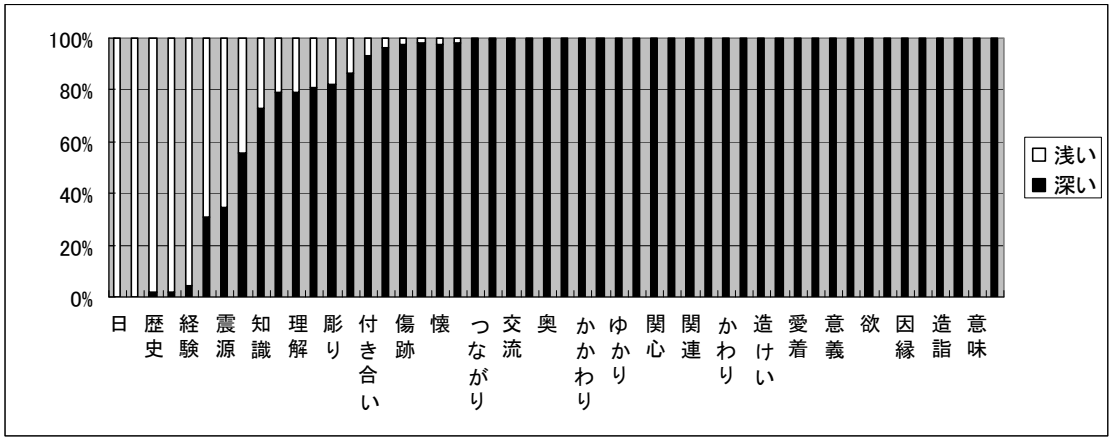


図 2　「深い/浅い」と各名詞類との共起傾向(用例数：2,212 から 31 まで)

名詞類：日, キャリア, 歴史, 眠り, 経験, 底, 震源, 水深, 知識, 傷, 理解, 方, 彫り, 奥行き, 付き合い, 根, 傷跡, 溝, 懐, 雪, つながり, なじみ, 交流, 関係, 奥, 縁, かかわり, 悩み, ゆかり, 感, 関心, 思い, 関連, 親交, かわり[5], 思い入れ, 造けい, 苦悩, 愛着, 結びつき, 意義, 感慨, 欲, 罪, 因縁, 悲しみ, 造詣, 愛情, 意味, 印象

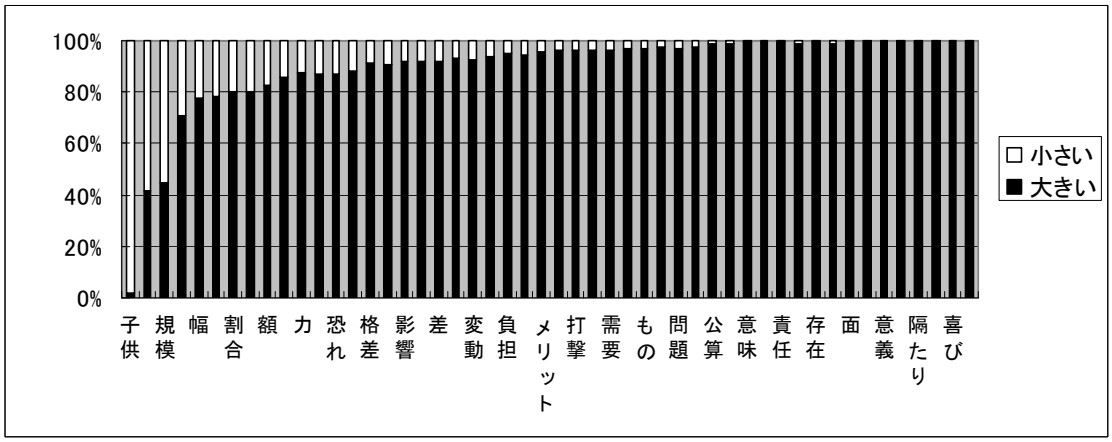


図 3　「大きい/小さい」と各名詞類との共起傾向(用例数：9,236 から 273 まで)

名詞類：子供,体,規模,声,幅,スケール,割合,率,額,性,力,余地,恐れ,リスク,格差,比重,影響,効果,差,危険,変動,被害,負担,部分,メリット,感,打撃,ダメージ,需要,不安,もの,衝撃,問題,方,公算,こと,意味,役割,責任,ショック,存在,違い,面,期待,意義,ところ,隔たり,功績,喜び,反響

5　「かかわり」の誤解析であるがそのままにしておく。

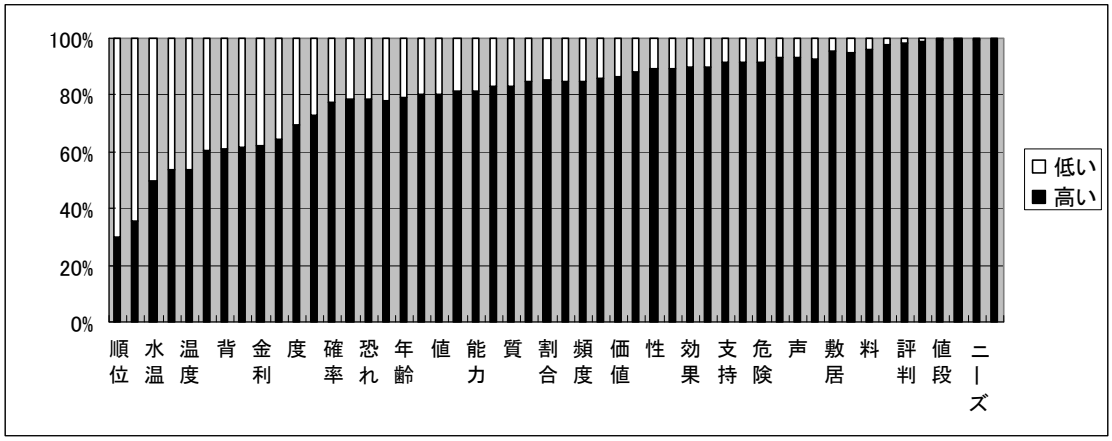


図 4　「高い/低い」と各名詞類との共起傾向(用例数：53,525 から 335 まで)

名詞類 : 順位, 度合い, 水温, 意識, 温度, 気温, 背, 水準, 金利, 率, 度, 濃度, 確率, コスト, 恐れ, 精度, 年齢, レベル, 値, 比率, 能力, リスク, 質, 関心, 割合, 力, 頻度, 評価, 価値, 倍率, 性, 方, 効果, 価格, 支持, ハードル, 危険, 需要, 声, 比重, 敷居, 費, 料, 期待, 評判, 人気, 値段, 料金, ニーズ, 呼び声

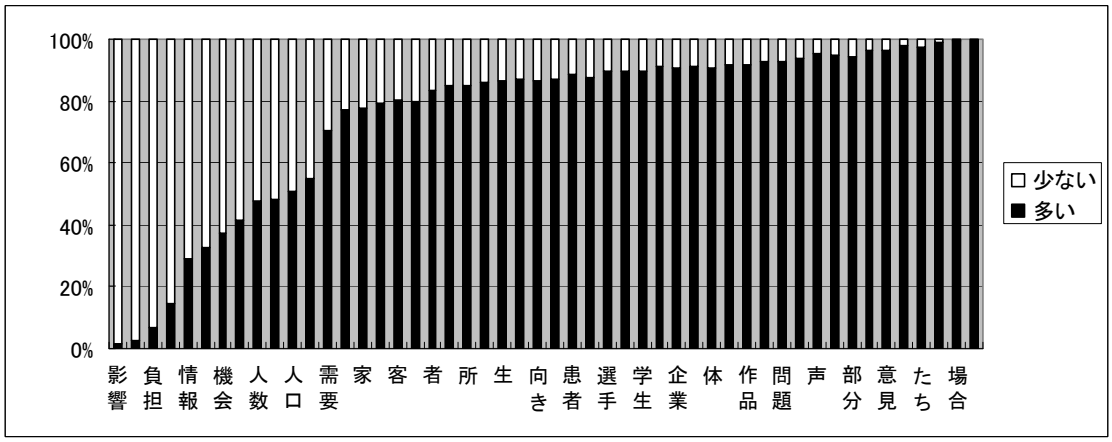


図 5　「多い/少ない」と各名詞類との共起傾向(用例数：32,762 から 746 まで)

名詞類 : 影響, 性, 負担, 時間, 情報, 人通り, 機会, 額, 人数, 数, 人口, 量, 需要, 例, 家, 店, 客, 利用, 者, 議員, 所, 国, 生, 人, 向き, 生徒, 患者, 若者, 選手, 女性, 学生, もの, 企業, 子, 体, 方, 作品, ところ, 問題, こと, 声, 日, 部分, ケース, 意見, ファン, たち, 点, 場合, 課題

「多い-少ない」に関しては、対象を数え上げる場合は「多い」との共起が多く、抽象的な量を表す場合は「少ない」との共起が多いように思われる(図 5)。

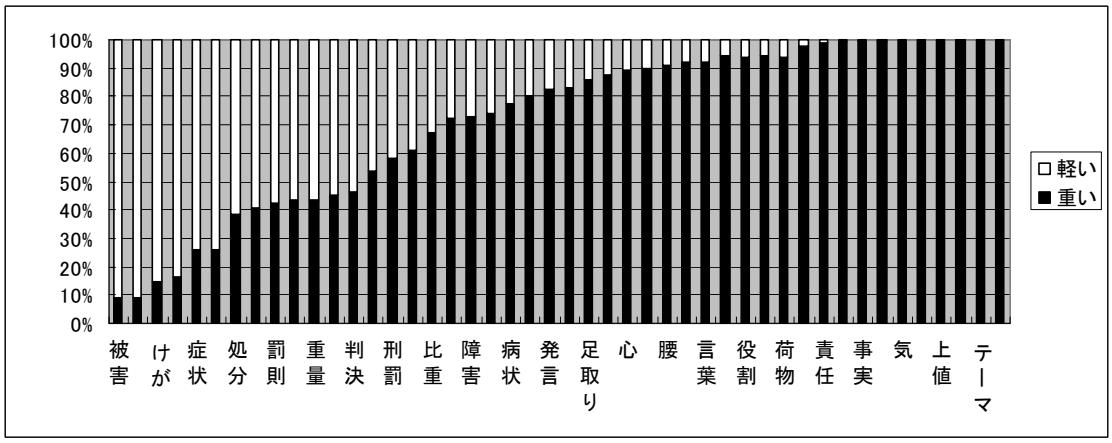


図 6　「重い/軽い」と各名詞類との共起傾向(用例数：2,212 から 8 まで)

名詞類 : 被害, 作用, けが, 手, 症状, 程度, 処分, 度, 罰則, 体重, 重量, 量刑, 判決, 刑, 刑罰, 体, 比重, 負担, 障害, もの, 病状, 足, 発言, 方, 足取り, こと, 心, 罪, 腰, 口, 言葉, 頭, 役割, 結果, 荷物, 点, 責任, 荷, 事実, 課題, 気, 意味, 上値, 責務, テーマ, 使命

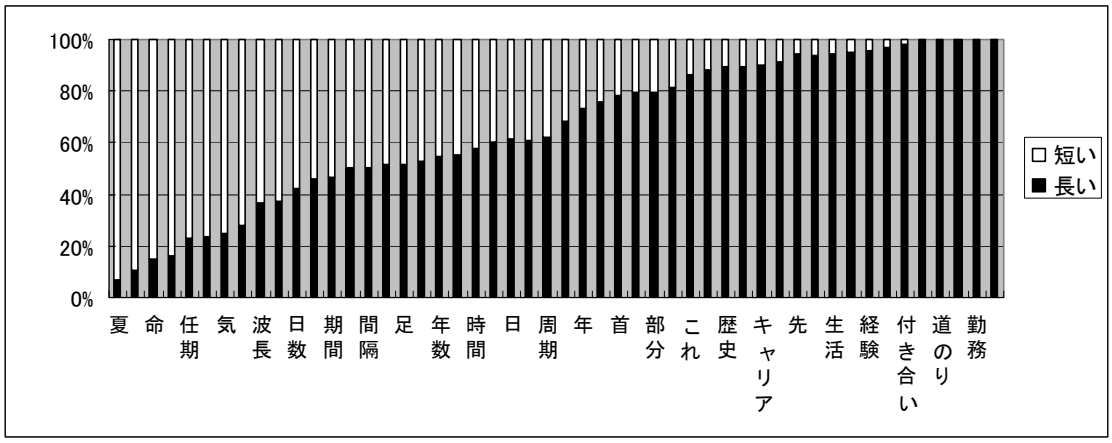


図 7　「長い/短い」と各名詞類との共起傾向(用例数：2,370 から 19 まで)

名詞類：夏, 余命, 命, 工期, 任期, さ, 気, 丈, 波長, 期限, 日数, 距離, 期間, 寿命, 間隔, 間, 足, 時期, 年数, 手足, 時間, 人生, 日, 髪, 周期, 期, 年, 尾, 首, 夜, 部分, 年間, これ, 歴, 歴史, シーズン, キャリア, それ, 先, 方, 生活, 私, 経験, ほう, 付き合い, 畑, 道のり, 息, 勤務, 暮らし

（B）　均衡型

「広い−狭い」、「太い−細い」の 2 つの形容詞対がこの型に属する。大値語との共起に偏る名詞類と小値語との共起に偏る名詞類の数が拮抗している。また、特に「広い−狭い」では、名詞類により大値語の共起比率が 0%から 100%まで連続的に分布している(図 8)。「太い−細い」は、「食が細い」などの慣用句以外では抽象的・比喩的に用いられることがあまりなく、いわば汎用性に欠ける。

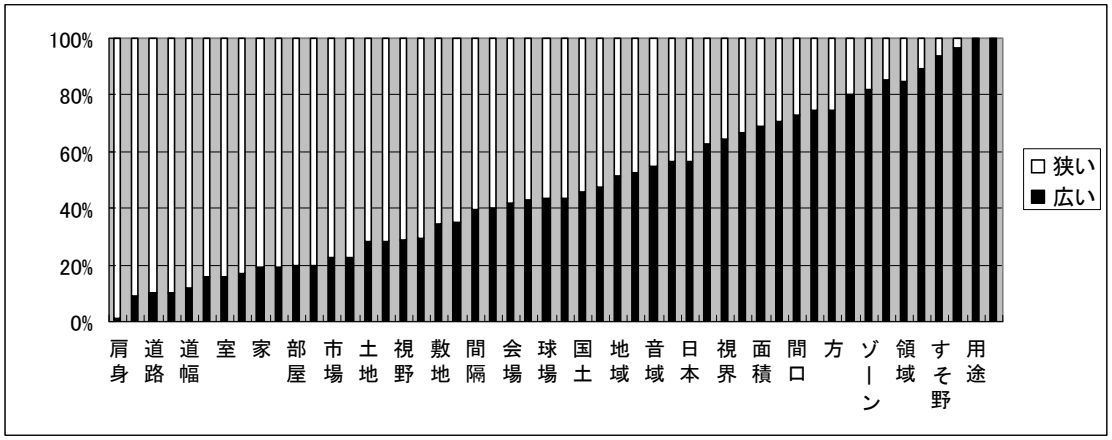


図 8　「広い/狭い」と各名詞類との共起傾向(用例数：1,255 から 31 まで)

名詞類：肩身, 通路, 道路, 住宅, 道幅, 道, 室, 場所, 家, グラウンド, 部屋, 門戸, 市場, 地球, 土地, 歩道, 視野, スペース, 敷地, 川幅, 間隔, エリア, 会場, 場, 球場, 庭, 国土, 心, 地域, 世間, 音域, 空間, 日本, 区, 視界, 世界, 面積, 幅, 間口, 空, 方, 対象, ゾーン, 範囲, 領域, 分野, すそ野, 層, 用途, 顔

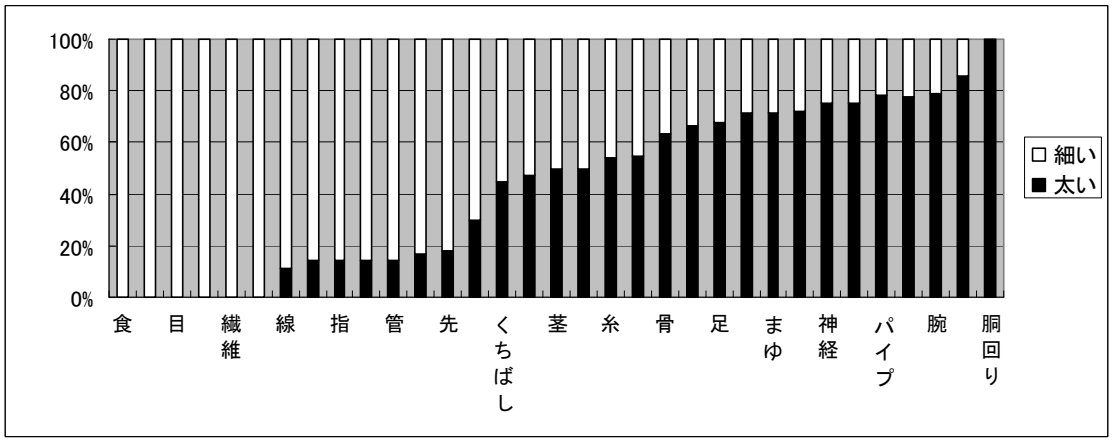


図 9　「太い/細い」と各名詞類との共起傾向(用例数：177 から 7 まで)

名詞類：食, 体, 目, 道, 繊維, 攻め, 線, 両端, 指, 幅, 管, 血管, 先, 声, くちばし, 部分, 茎, 首, 糸, 直径, 骨, ウエスト, 足, 柱, まゆ, 幹, 神経, 脚, パイプ, 方, 腕, 回り, 胴回り

（C） 小値語傾斜型

「薄い–濃い(厚い)」[6]のみが属する。小値語の「薄い」としか共起用例のない名詞類が多い。それには「望み、見込み」、「意欲、関心、意識」、「関わり、縁、なじみ、関与」など意味的に類似性のあるグループがいくつかある。

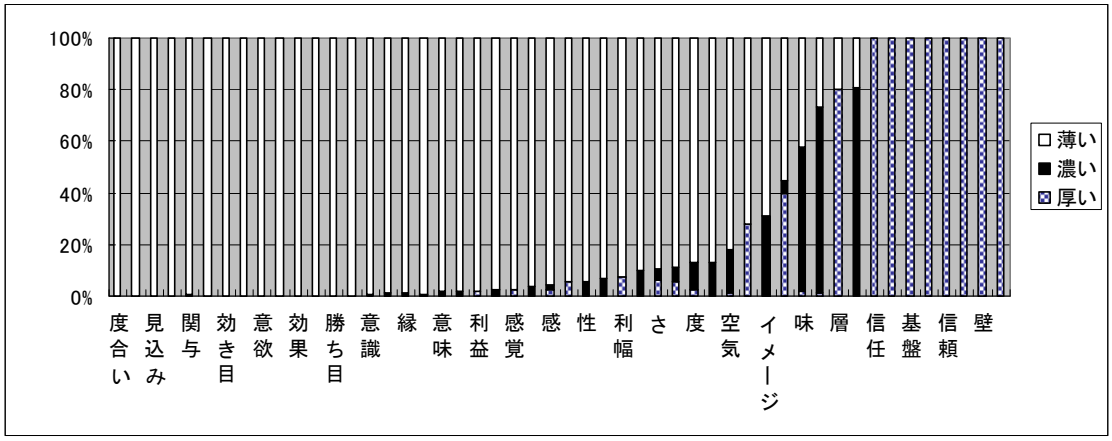


図 10　「濃い+厚い/薄い」と各名詞類との共起傾向(用例数：2,694 から 69 まで)

名詞類 : 度合い, 根拠, 見込み, 望み, 関与, かかわり, 効き目, 髪, 意欲, 関心, 効果, 意義, 勝ち目, なじみ, 意識, 関連, 縁, 認識, 意味, 実感, 利益, 魅力, 感覚, 見通し, 感, 期待, 性, 影, 利幅, 印象, さ, 関係, 度, つながり, 空気, 心, イメージ, 皮, 味, 中身, 層, 色, 信任, 支持, 基盤, 信望, 信頼, 人望, 壁, 人情

（D） その他

比較のため、存在表現の「ある」「ない」の分布を示すと次のようになる。「長い」「短い」などと比較的似た分布であることがわかる。

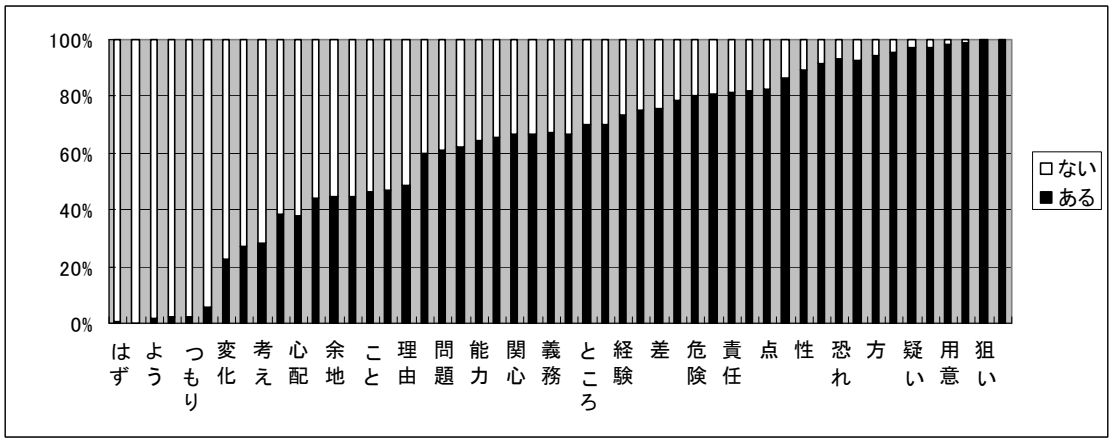


図 11　「ある/ない」と各名詞類との共起傾向(用例数：172,906 から 6,363 まで)

名詞類 : はず, 変わり, よう, 仕方, つもり, わけ, 変化, 影響, 考え, 余裕, 心配, 例, 余地, 方法, こと, 意味, 理由, 関係, 問題, 力, 能力, 動き, 関心, もの, 義務, 自信, ところ, 不安, 経験, さ, 差, 違い, 危険, 感, 責任, 必要, 点, 効果, 性, 事情, 恐れ, 人気, 方, 声, 疑い, 面, 用意, 限界, 狙い, 経緯

[6] 他に「篤い」と表記されるものなどが若干あると思われるが、考慮していない。

また、「薄い-濃い」との比較のため「濃厚-希薄」の分布を示すと次のようになる。「薄い」の場合とは大きく異なり、大値語に傾斜した分布であることが分かる。

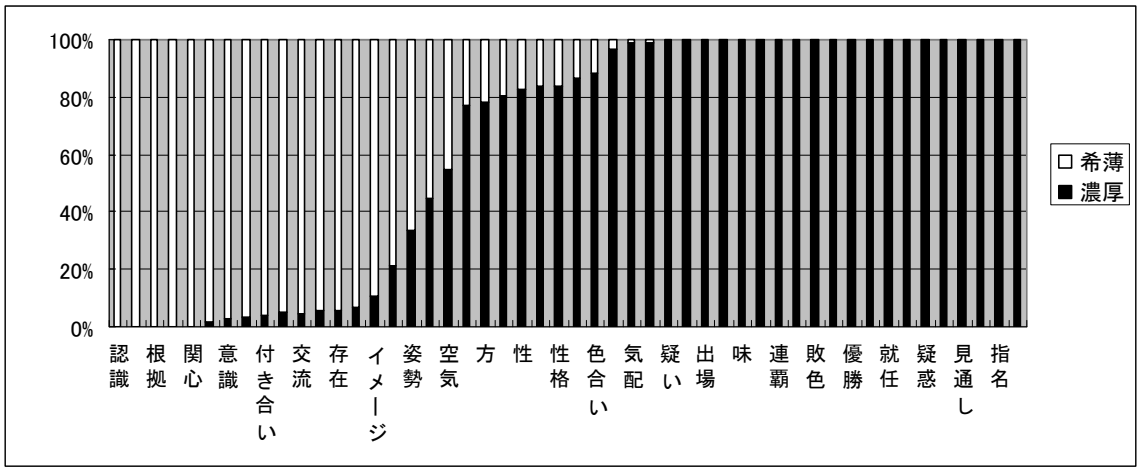


図 12　「濃厚/希薄」と各名詞類との共起傾向(用例数：403 から 13 まで)

名詞類 : 認識, 感覚, 根拠, 実感, 関心, つながり, 意識, 視点, 付き合い, 感, 交流, 関係, 存在, さ, イメージ, かかわり, 姿勢, 印象, 空気, 要素, 方, 関与, 性, 雰囲気, 性格, 色, 色合い, 影響, 気配, 色彩, 疑い, こと, 出場, 先発, 味, 起用, 連覇, 入り, 敗色, 形跡, 優勝, 移籍, 就任, 線, 疑惑, 容疑, 見通し, 打ち, 指名, 留任

また共起名詞類を眺めると、「濃い」と「濃厚」では相違がある。「優勝が濃厚」のような言い方(「優勝の可能性が濃厚」の意)は、「濃い」には対応する用例が見えない。

4.　名詞類に対する複数形容詞対の共起パターン

ここで観点を変える。ある名詞類に対して複数の形容詞対がよく共起する場合、各形容詞対に関して大値語/小値語との共起傾向がおよそ一様である場合と、一つ(複数)の形容詞対のみ特異な傾向を示す場合とがある。表 1 を見られたい。

例えば、「疑い」や「色」は、主な共起形容詞対で一様に大値語との共起が多い。一方、「(～)性」、「(～)感」、「恐れ」などは、全体に大値語との共起が多い中で「濃い-薄い」と「多い-少ない」の対のみ小値語に偏る。同様に、「関連」では「薄い」が、「影響」では「少ない」が、例外的に高い共起度を示す。

反対に、「見込み」では、主な共起形容詞対で小値語との共起が多いが、「強い-弱い」が例外になる。

表１　名詞類と形容詞対の共起パターン

	疑い	色	性	感	恐れ	関連	影響	見込み
多い	0	**55**	45	1	12	**28**	32	0
少ない	0	18	***1816***	***247***	***302***	4	***2000***	**93**
高い	**78**	0	**47888**	**152**	**346**	9	5	30
低い	4	0	5670	31	95	10	28	**44**
大きい	**11**	0	**4488**	**628**	**364**	0	**7903**	6
小さい	0	0	745	26	56	0	680	**13**
深い	0	**13**	6	**229**	0	**168**	0	0
浅い	0	2	1	0	0	0	0	0
強い	**7087**	**827**	**11494**	**4291**	**1405**	**53**	**407**	***29***
弱い	0	2	102	26	0	2	13	0
濃い	**708**	**901**	72	10	1	3	**39**	0
薄い	16	235	***1257***	***561***	***9***	***231***	24	**227**

鍋島(2011)は、「可能性に関する用語」(と鍋島がみなすもの)と「濃い-薄い」などとの共起関係(内省による)をメタファー的基盤の観点から分析している。可能性に関する用語のうち「見込み・期待」など良い可能性を表わすものは「薄い」系とのみ共起し、「疑い・敗色」など悪い可能性を表わすものは「濃い」系とのみ共起するとの観察が示され、共起しない組合せは《善は白・悪は黒》というメタファーとの衝突によるとの説明が提案される。

この説明は興味深いものであり、たしかに「黒い疑惑」などの言い方の存在からも、「悪は黒」というメタファーが「疑い」と「濃い」の共起に関係していることは考えられる。

しかし、鍋島の説明の有効性には疑問の余地がある。まず、「見込み」/「疑い」は、「濃い-薄い」に関してのみ共起用例が(それぞれ)小値語/大値語の方に偏るのではなく全体的に小値語/大値語との共起に偏るのである。「見込み」という語はその値が小さい局面で用いられやすい語であり、「疑い」はその逆と思われる。また悪い事態の生起可能性に関わる語であっても、「恐れ」のように「濃い」とあまり共起しない語も存在する。

5.　まとめ

基本的用法での極性反義関係を軸として、形容詞対と名詞の共起用例の分布を観察・分析し、多くの形容詞対では大値語とよく共起する名詞類が多い中、「薄い-濃い(厚い)」はその逆であること、「広い-狭い」では大値語と小値語の分布が対称的であることなどを発見した。また、ある名詞類と共起する形容詞対に対して、大値語/小値語との共起傾向が一定である場合と、一つ(複数)の形容詞対のみ特異な傾向を示す場合とがあることが分かった。「強い」「少ない」「薄い」などが例外になることが多いが、その理由の説明は、今後の課題である。

文　献

Cruse, Alan(1986)　*Lexical Semantics*, Cambridge University Press.
久島茂(2001)　『〈物〉と〈場所〉の対立―知覚語葉の意味体系―』くろしお出版.
久島茂(2002)『《物》と《場所》の意味論　「大きい」とはどういうこと』　くろしお出版.
鍋島弘治朗(2011)　『日本語のメタファー』　くろしお出版.
西尾寅弥(1972)『形容詞の意味・用法の記述的研究』秀英出版.
服部匡(2011a)「程度的な側面を持つ名詞とそれを量る形容詞類との共起関係―通時的研究―」　『言語研究』140 号,　pp.89-116.
服部匡(2011b)　「名詞と尺度的形容詞類の共起傾向の推移―国会会議録のデータから―」　『同志社女子大学学術研究年報』62 号, pp.113-141
服部匡(2012)　「名詞と尺度的形容詞類の共起頻度の推移―国会会議録のデータから―」　『同志社女子大学大学院文学研究科紀要』12 号,　pp.1-11.
服部匡(準備中)　「極性反義語の用例分布」

本研究は、学術研究助成基金助成金(基盤研究(C)「有無・量的大小・増減・出現消滅の述語の総合的研究」、課題番号 23520479)および、国立国語研究所共同研究プロジェクト「コーパス日本語学の創成」による研究成果である。

ポスター発表 (2)

9月7日 (金)　10:20～12:20

現代日本語書き言葉均衡コーパスに対する難易度付与

佐藤 理史　（名古屋大学大学院工学研究科）

Assigning Readability Score to Every Text in BCCWJ

Satoshi Sato　(Graduate School of Engineering, Nagoya University)

1　はじめに

我々は、これまで、『現代日本語書き言葉均衡コーパス (BCCWJ)』[1] の各サンプルテキストに難易度を付与すべく、準備を行なってきた。我々が開発した難易度測定システム『帯 2(obi2)』[1] は、あらかじめ用意した、難易度の規準となるコーパス (規準コーパス) に基づき、与えられたテキストの難易度を決定するシステムである。このシステムでは、まず、規準コーパスに含まれる、それぞれの難易度に対応するサブコーパスに対し、言語モデル (文字 bigram モデル) を作成する。次に、未知のテキストに対して、それぞれの言語モデルに対する尤度を計算し、最大の尤度をとる言語モデルに対応する難易度を出力する[2]。このように、本システムでは、規準コーパスが難易度スケールを規定する。

『帯 2』が最初に提供した難易度スケールは、13 段階の学年区分 (小学 1 年から高校 3 年、および、大学) に対応する obi2/T13 である。このスケールの規準コーパスには、各学年の教科書から抽出したテキストで構成する「教科書コーパス」[2] を用いている。これとは別に、均衡コーパスから規準コーパスを作成する方法を考案し、BCCWJ のモニター公開データ (2009 年度版) に含まれる書籍データ 10,423 サンプルから、相対的難易度を表す難易度スケール obi2/B9 を作成した [3]。その後、31 名の被験者実験の結果との比較により、obi2/B9 の難易判定能力が、平均的な人間と同程度であることを明らかにした [4]。

今回、2012 年 1 月に 2 枚組 DVD として正式にリリースされた『現代日本語書き言葉均衡コーパス』(以下、BCCWJ リリース版) に含まれる、書籍の可変長サンプル 18,000 件を用いて、難易度スケール obi2/B9 を再構成した。さらに、再構成した obi2/B9 を用いて、BCCWJ リリース版に含まれる全サンプルに難易度を付与した。本稿では、これらの内容について報告する。

2　難易度スケール obi2/B9 の再構成

本節では、難易度スケール obi2/B9 の再構成について、前回 [3] との差分を中心に述べる。

2.1　使用したサンプル

今回の obi2/B9 の再構成には、BCCWJ リリース版に含まれる、出版サブコーパスの書籍 (PB) および図書館サブコーパスの書籍 (LB) の可変長サンプル、全 20,688 サンプルのうち、有効 bigram 数が多い 18,000 サンプルを利用した。有効 bigram とは、『帯 2』が難易度計算のために使用する可能性がある文字 bigram (連続する 2 文字) のことで、2 つの文字は、いずれも、ひらがな、カタカナ、JIS 漢字第 1 水準のいずれかである。使用するサンプル数は、前回の反省から、1000 で割り切れる数とした。最も有効 bigram 数が少ないサンプルは、1,074 個の有効 bigram を含む。

2.2　難易度スケール B9 の構成手順

難易度スケール obi2/B9 の構成手順は、おおむね、前回の構成手順 [3] を踏襲した。具体的手順を以下に示す。

[1] http://kotoba.nuee.nagoya-u.ac.jp
[2] 実際の計算は、スムージングを適用するため、もう少し複雑である。

表 1: Stanine

stanine	1	2	3	4	5	6	7	8	9
割合	4%	7%	12%	17%	20%	17%	12%	7%	4%
範囲	0–4%	4–11%	11–23%	23–40%	40–60%	60–77%	77–89%	89–96%	96–100%

(1) 難易度付きコーパス B_0 を用意する。
(2) $i = 1$
(3) 難易度付きコーパス B_{i-1} を用いて、与えられた 2 つのテキストの難易を判定する比較器を構成する。
(4) 構成した比較器を用いて B_{i-1} に含まれるサンプルを難易順にソートする。
(5) ソート結果に基づき、B_{i-1} の各サンプル (テキスト) に 9 段階の難易度を付与する。難易度の決定には、stanine (後述) を用いる。
(6) こうして得られた 9 段階の難易度コーパスから、その一部 (各難易度に対して同数のサンプル) を取り出す。これを規準コーパス C_i として、obi2/B9$_i$ を構成する。
(7) obi2/B9$_i$ を用いて、コーパス B_{i-1} の各テキストに難易度を付与する。この結果として得られる難易度付きコーパスを B_i とする。
(8) i=$i+1$ として、(3) へ。

実際には、この手続きを、$i = 7$ まで実行した。上記の手続きに出てくる stanine は、ソートされたリストの各要素に、表 1 に示す割合に従って、1 から 9 の整数を割り当てる方法である。整数は、難易度が高いものほど値が大きくなる方向で割り当てる。

今回の obi/b9 の再構成においては、一貫して、前述の 18,000 サンプルを用いた。すなわち、上記の B_i は、すべて同一の 18,000 サンプルから構成されている。添字の違いは、各サンプルに付与されている難易度の値の違いのみである。

以下に、上記の手続きの各ステップの詳細を示す。

難易度付きコーパス B_0 の作成 18,000 サンプルに対して、前回作成した (すでに存在する) obi2/B9 を用いて、9 段階の難易度を付与した。これを、難易度付きコーパス B_0 として使用した。

比較器の構成 前回と同一の手順で、比較器を SVM として構成した。ただし、頻度パラメータ f_c (SVM の属性として使用する、bigram の最低頻度) は、$f_c = 100$ を用いた。

コーパスのソートと難易度付与 コーパスのソートは、18,000 サンプルを 18 個のパーティションに分け、パーティションごとにソートする方法を採用した[3]。stanine に基づく難易度付与も、パーティションごとに行なう。この結果、9 段階の難易度が付与された 1,000 サンプルが、18 パーティション分、得られる。

規準コーパスの作成と難易度スケールの構成 1 つのパーティションから、1 つの難易度に対して、40 サンプルを選ぶ[4]。この結果、各難易度に対して 720 サンプル、全部で 6,480 サンプルが選ばれる。これを規準コーパス C_i として obi2/B9$_i$ を構成する。このとき、難易度の測定に使用する bigram の最低頻度 f を $f = 100$ とした。

[3] 今回使用した計算機では、このソートに約 1 週間かかる。
[4] 難易度 1 および 9 の割合は 4%なので、1 つのパーティションに 40 サンプルずつしかない。

表 2: 難易度付きコーパスの変化

	R	RMSE	s_0	s_1
B_0–B_1	0.974	0.504	0.763	0.997
B_1–B_2	0.985	0.350	0.888	0.998
B_2–B_3	0.987	0.326	0.899	0.999
B_3–B_4	0.986	0.334	0.898	0.998
B_4–B_5	0.987	0.315	0.911	0.998
B_5–B_6	0.988	0.304	0.914	0.998
B_6–B_7	0.989	0.295	0.919	0.998
B_7–$B_{7(f=75)}$	0.997	0.165	0.975	0.999
B_0–$B_{7(f=75)}$	0.958	0.631	0.666	0.987

コーパスに対する難易度の再付与 こうして得られた obi2/B9$_i$ を用いて、18,000 サンプルに難易度を付与し直す。この結果、新たな難易度付きコーパス B_i が得られる。

2.3 難易度付きコーパスの変化と 2 分割交差検定の変化

ここでは、上記の手続きの繰り返しの過程で、難易度付きコーパス B_i の各難易度がどのように変化したかに着目する。この変化は、2 つのコーパスの各難易度の値を、同じサンプルに対する添字が同一になるように並べたのち、2 つの列 $A = \{a_j\}$ と $B = \{b_j\}$ $(j = 1, 2, \ldots, n)$ を比較することによって得られる。この比較には、次の 4 つの指標を利用する。

$$\text{相関係数 } R(A,B) = \frac{\sum_{j=1}^{n}(a_j - \bar{a})(b_j - \bar{b})}{\sqrt{\sum_{j=1}^{n}(a_j - \bar{a})^2}\sqrt{\sum_{j=1}^{n}(b_j - \bar{b})^2}} \tag{1}$$

$$\text{root mean square error } RMSE(A,B) = \sqrt{\frac{(a_j - b_j)^2}{n}} \tag{2}$$

$$\text{一致率 } s_0(A,B) = \frac{1}{n}\sum_{j=1}^{n} d(a_j, b_j, 0) \tag{3}$$

$$\text{差 1 を許容した一致率 } s_1(A,B) = \frac{1}{n}\sum_{j=1}^{n} d(a_j, b_j, 1) \tag{4}$$

ここで、関数 $d(a_j, b_j, v)$ は、a_j と b_j の差が v 以下かどうか調べる関数で、次のように定義する。

$$d(a,b,v) = \begin{cases} 1 & \text{when } |a-b| \le v \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

上記の 4 つの指標のうち、RMSE を除く 3 つの指標は、列 A と B が似ているほど大きな値をとる。RMSE は、逆に、列 A と B が似ているほど小さな値をとる。

表 2 に、コーパス B_{i-1} と B_i の変化を上記の 4 つの指標で計測した結果を示す。この表より、コーパス B_{i-1} と B_i の差は、かなり小さいことがわかる。B_0 と B_1 の差が最も大きいが、それでも相関係数 $R = 0.974$、一致率 $s_0 = 76.3\%$、差 1 を許容した一致率 $s_1 = 99.7\%$ である。最後の 2 つのコーパス B_6 と B_7 の一致率 s_0 は、91.9%である。これは、最後の 1 回のループにおいて難易度の値が変化したサンプルは、全体の 8.1%にすぎないことを意味する。

表 3 に、C_i を規準コーパスとする obi2/B9$_i$ の 2 分割交差検定の結果を示す。この表より、2 分割交差検定の評価値の変化は小さいことがわかる。ただ、$i = 1$ と $i = 7$ の評価値を比較すると、いず

表 3: obi2/B9$_i$ の 2 分割交差検定

obi2/B9$_i$	R	RMSE	s_0	s_1
1	0.976	0.568	0.700	0.994
2	0.978	0.544	0.725	0.994
3	0.976	0.564	0.711	0.992
4	0.978	0.542	0.718	0.996
5	0.977	0.551	0.721	0.994
6	0.977	0.554	0.714	0.995
7	0.978	0.542	0.724	0.996
7($f = 75$)	0.978	0.538	0.726	0.996

れの評価値の向上している。このことは、前述の手順による繰返しにより、より一貫した難易度が付与される方向に動いていることを、間接的に示している。

2.4 最終的に採用した難易度スケール

最終的に採用した難易度スケールは、規準コーパス C_7 から作成した。ただし、難易度の測定に使用する bigram の最低頻度 f の値は、$f = 100$ ではなく、$f = 75$ を採用した。表 3 の最終行に、$f = 75$ の場合の 2 分割交差検定の評価値を示す。この表に示すように、$f = 75$ の評価値は、$f = 100$ と比べ、RMSE と s_0 において、若干良い。

なお、表 2 に示すように、18,000 サンプルに対して、$f = 75$ の obi2/B9$_7$ で難易度を付与した場合 ($B_7(f = 75)$) と、$f = 100$ の obi2/B9$_7$ で難易度を付与した場合 (B_7) の差は、非常に小さい。事実、97.5%のサンプルで、難易度の値は同一である。

最終的に採用した、新しい obi2/B9 は、33,360 種類の有効 bigram に基づいて難易度を決定する。なお、以前の obi2/B9 が利用する有効 bigram は、29,017 種類であった。

表 2 の最終行に、18,000 サンプルに、以前の obi2/B9 によって難易度を付与した場合 (B_0) と、最終的に採用した新たな obi2/B9 によって難易度を付与した場合 ($B_7(f = 75)$) の差を示す。この表の他の値と比較すると、これら 2 つの難易度付きコーパスの差は大きいように見えるが、実際には、66.6%のサンプルで難易度の値が一致しており、値に差があるものも、そのほとんど (98.7%) が ± 1 の範囲にある。つまり、BCCWJ のリリース版を用いて作り直した obi2/B9 は、以前の obi2/B9 とそれほど大きな差はない。

2.5 被験者実験との比較

以前行なった被験者実験との比較 [4] と全く同じことを、新たに構成した obi2/B9 を用いて行なった。この文献 [4] の表 2 の末尾に、新たな obi2/B9 のデータ (nB9) を追加したものを表 4 に示す。3 つの指標 d_0, d_c, d_i の値は、以前の B9 (oB9) より減少している。これらの値は、いずれも人間の多数派の回答とのハミング距離 (差の数) を表しており、その値の減少は、人間の多数派の回答に、より近づいていることを意味する。

3 BCCWJ に対する難易度付与

新たに構成した obi2/B9 を用いて、BCCWJ リリース版に含まれる全サンプルに対して、難易度を付与した。

3.1 固定長サンプルに対する難易度付与

表 5 に、固定長サンプルに対する難易度付与の結果を示す。図 1 は、白書 (OW) を除く各レジスタの難易度分布を、棒グラフ化したものである。このグラフの縦軸は、全体に対する割合を示してお

表 4: 被験者実験との比較 (120 ビット 1 対比較コード)

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	d_0	d_c	d_i
m	111000	101001	110000	111011	111111	001111	000011	101001	000110	000100	111011	000100	100001	000001	000000	111111	000100	010100	101001	000001			
ci	−+?+−+	−?+−+	++??++	−+−++−	+−−+−+	−+?+−+	−−+?++	−?+−++	−−+?+	??+?++	+++−??	−+−+?+	?−+−++	+−++??	+−−+−+	?++?+?	++−?++	++−+−	−++−+−	?+++??	120	96	62
p_{01}	.0...	.....	....1.	.....	.....	.....	.....	.....	.....	1..0..	.....	.....	.....	1....0	.....	0.....	...0..	.....	....1.	....1.	11	2	2
p_{02}	.0...	.1....	.....	.....	.....	...0..	....0.	.1....	.....	.....	....0.	.....	.....	....1.	1.....	0	.....	.....	.....	....1.0	11	4	4
p_{03}	.....	.1...0	.1....	.....	.....	...0..	...1..	.0....	.1....	.....	....0.	.....	0	.....	.....	00....	...0..	.....	.....	.....	12	6	6
p_{04}	.0....	.10..0	...1..	.....	0....0	.....	.....	.1....	.....	.....	00.0..	.....	.....	....1.	.....	0.....	.....	.....	.....	11.1.0	17	8	7
p_{05}	.....	.....	..11..	.....	.....	..00..	...1.0	.1....	.....	1..0..	.....	.....	.....	.....	.....	.....	...0.1	.....	.....	1.....	12	3	3
p_{06}	1	.1...0	..11..	.....	.....	.1....	...1..	.1....	.....	.1....	....0.	.....	.....	1.1.1.	.....	1.....1	0....0	.....	.1....	....1.	18	7	7
p_{07}	.....	.....	..11..	.....	.....	0.1....	...1..	.....	.....	.....	0..1..	.1..1.	.....	....0.	.....	.....	...0..	1.....	.....	1.1.1.	15	7	7
p_{08}	.....	..0...	.0...1	.....	.....	.0..0	1.1.0.	..0...	....0.	11....	000...	...0..	0.....	...1.0	11....	000...	.1..1.	.....	0.0...	....1.	30	22	18
p_{09}	.00...	.1...0	.0....	.0..0.	.....	.1....	....0.	..0...	.....	.1....	....0.	.1....	0.....	..1.1.	...1..	.....	.....	.....	....1.	.....	18	12	12
p_{10}	.....	..0...	..1..1	.....	.....	.0..0	.11..	.....	....0.	11.0..	...1..	...0..	0.....	.111.	1.....	...00.	.....	.....	0.0...	11.1.0	26	13	12
p_{11}	.0....	....1.	..1111	0..0.	.....	..0...	...1..	.1....	.....	11....	.....	.0.1.0	0.....	0	...1..	...00.	.....	.....	.0....	...110	25	11	11
p_{12}	.01..	0...1.	.....	.....	.....	.....	.....	.....	.....	.....	...1..	...0..	.....	..1.1	.....	...000	.....	.....	.....	1.1...	17	7	6
p_{13}	...1..	.1....	.....	.....	.....	.....	00	.1...0	.1..0.	.....	....0.	.....	0	.....	.....	.....	.0....	.....	.....	1....0	13	7	7
p_{14}	.....	..0...	...1..	.0....	.....	.0.....	111...	.1....	....0.	.....	....0.	.1....	.....	....1.	.....	.....	.....	.....	.....	0	14	8	6
p_{15}	.0....	.1...0	.....	...1..	0	11....	...100	.0..0	.....	.....	...1.0	....1.	0.....	.....	...1..	.0.0..	...0..	.....	.0....	1.1...	23	14	13
p_{16}	.....	.....	00...1	.....	.....	.0..0	1...00	.1....	.....	1.0..	00	...0..	0.....	.....	11....	...0..	.....	.....	.....	.1.1.	20	12	10
p_{17}	.0....	.1....	.0....	.....	.....	.....	.....	....1.	.1....	.....	....0.	....1.	.....	.....	....1.	.....	.....	.....	....1.	.....	9	5	5
p_{18}	.....	.1...0	..1...	.....	.....	....0.	.....	.1....	.....	.....	.....	....1.	.....	....1.	....1.	.....	....1.	.0.0..	....1.	1.1.1.	16	8	7
p_{19}	.....	.1....	0.1...	...1..	0.....	.....	...1..	1...0	.....	1..0..	...1..	.0.1.0	0.....	.1.1.	1.....	.00000	...0..	.....	1....	0	25	12	12
p_{20}	.....	.....	0..11.	....0.	0.....	.....	.0000	111.00	..0...	.11...	11....	0001..	.1..1.	.....	...1.1	1...1	0....0	...0..	.....	...1.0	33	22	17
p_{21}	.....	.....	00.1..	...1..	.....	.....	.....	1	.....	...0..	...1..	....1.	.....	.1.1.	.....	.....	.....	.0....	.....	1....0	14	7	7
p_{22}	.....	..0..0	..11.	....0.	.....	110000	...1..	..0...	.....	11....	.....	....1.	.....	0	.1.1..	...00.	...0..	.....	.....	.....	24	13	10
p_{23}	.....	.1....	..1..1	.....	.....	...0..	...1..	.1....	....0.	.1....	.....	.....	.....	.....	.....	0	.....	.....	....1.	0	11	3	3
p_{24}	.00...	....1.	..11..	.....	0...0	.....	..11..	1.1.0	....0.	...0..	.0..0.	.1..1.	.....	0	.1....	0.....	.....	1....	.....	..111.	24	12	11
p_{25}	..11.	..0...	..1...	.....	000.00	11...0	...1..	.....	....0.	...0..	...1..	...0.1	0.....	0	1.1.11	000...	.....	.0....	.....	1.1.1.	31	21	14
p_{26}	.....	..0...	.0....	.....	0	..0..0	.....	.1....	.....	...0..	.....	.....	0.....	1.....	1....1	0....0	.1....	.....	.....	.....	14	8	8
p_{27}	.00...	.1...0	00...1	.....	0.....	.0....	...11.	.1...0	1.....	11....	...1..	.....	.....	.....	.....	000..0	...0..	.....	.....	...11.	26	12	12
p_{28}	.....	.....	..1...	.....	.....	.0..0	.....	.0....	.....	...0..	.....	....1.	0.....	.....	.....	...000	...0..	.0....	.....	1.....	14	5	5
p_{29}	.00...	.....	..11..	.....	.....	.....	.....	.....	....0.	11....	...1.0	.....	.....	..1.1.	.....	...0..	.....	.0....	.....	11...0	16	5	5
p_{30}	.....	.1....	..11..	.....	0.....	..00..	...1..	.0....	....0.	.....	0..0..	.1....	0...0	1...0	.....	0...0	...0..	1.....	.....	1....0	22	8	8
p_{31}	...1..	.1...0	.0.....	....0.	.....	.....	.....	.....	.....	.....	...1.0	...0.1	0.....	111..0	1....1	...000	...011	1.....	.....	.....	23	16	15
av.																					18.8	9.7	8.7
oB9	.....	.?....	.0...?	....0?	0.....	.?....	....?.	.....	....?.	?..0..	....0.	.1....	0....0	.....	.....	....?.	.....	.?..?.	.....	....?.	14.5	8.5	8.0
nB9	.?...1	.?....	.0...?	....?.	0.....	.?....	..??..	.?....	.....	....?.	....0.	.1....	0....?	....?.	.....	.....	.....	.....	....?.	....?.	12.5	7.5	7.5

表 5: 固定長サンプルに対する難易度付与結果

サブコーパス-媒体	1	2	3	4	5	6	7	8	9	計	平均	分散
出版-書籍 (PB)	321	567	1299	1433	1976	1284	1298	1205	734	10117	5.34	4.43
出版-雑誌 (PM)	97	61	227	555	622	172	181	75	6	1996	4.60	2.46
出版-新聞 (PN)	5	0	20	218	488	392	260	88	2	1473	5.62	1.41
図書館-書籍 (LB)	536	835	1874	1920	2404	1533	911	315	223	10551	4.51	3.28
特殊目的-白書 (OW)	0	0	0	0	2	8	143	935	412	1500	8.16	0.38
書籍 (PB+LB)	857	1402	3173	3353	4380	2817	2209	1520	957	20668	4.92	4.01
stanine (S9)	4	7	12	17	20	17	12	7	4	100	5.00	3.84

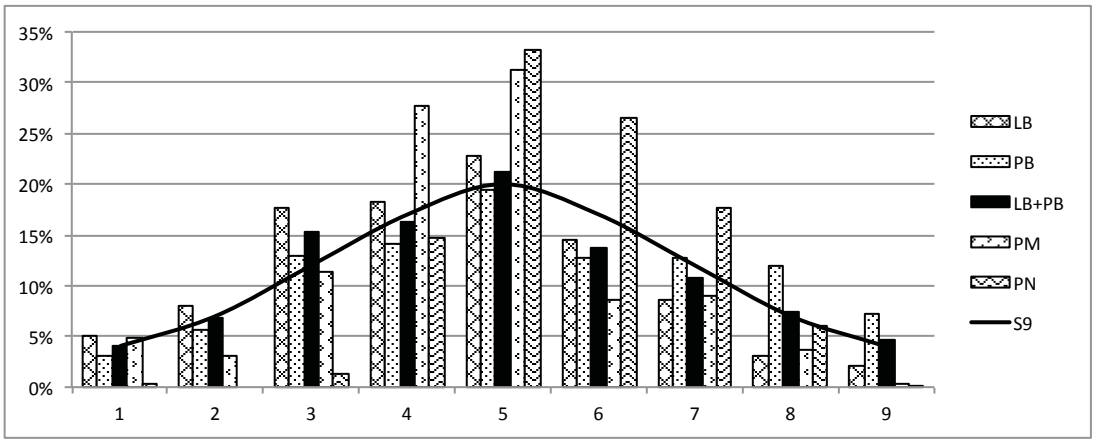


図 1: 固定長サンプル (LB, PB, PB+LB, PM, PN) の難易度分布

り、曲線 (S9) は、stanine の分布を示している。

このグラフより、書籍全体 (PB+LB) の難易度分布は、ほぼ、stanine の分布に従っていることがわかる。これは、難易度分布が、ほぼ設計通りとなっていることを示している。出版-書籍 (PB) と図書館-書籍 (LB) の 2 つのレジスタを比較すると、図書館-書籍 (LB) の方が難易度の平均が低く、分散も小さい。つまり、「図書館サブコーパスの書籍 (LB) は、出版サブコーパスの書籍 (PB) に比べ、難易度が低い方に偏っている」ということである。図書館-書籍 (LB) の母集団は、都内 13 自治体以上の図書館が共通して所蔵している書籍であるのに対し、出版-書籍 (PB) の母集団は、2001 年から 2005 年の間に国内で出版された書籍である [5]。2 つのレジスタの難易度の分布の違いは、このような母集団の違いによるものと考えるのが妥当である。この点については、後で再度、議論する。

雑誌 (PM) および新聞 (PN) は、書籍 (PB や LB) と比較する分散が小さい。つまり、比較的難易度が揃っている。雑誌 (PM) のサンプルの 59%は、難易度が 4 または 5 である。一方、新聞 (PN) は、サンプルの 60%が難易度が 5 または 6 であり、サンプルの 92%が難易度 4 から 7 の範囲である。これらの難易度の集中は、新聞・雑誌の編集において、難易度がコントロールされていることの現れと考えることができる。なお、雑誌 (PM) と新聞 (PN) では、雑誌の方が難易度の平均が低く、分散が大きい。このことは、我々が、日常生活において感じている印象と一致する。

図 1 のグラフには、白書 (OW) を含めなかった。白書のほとんどのサンプルは、難易度 7 から 9 であり、難易度の平均値が非常に高く、かつ、分散は小さい。

3.2 可変長サンプルに対する難易度付与

固定長サンプルと同様に、すべての可変長サンプルに対しても難易度を付与した。ただし、可変長サンプルの長さはさまざまで、1000 字未満のサンプルや、有効 bigram を一つも含まないサンプルも

表 6: 可変長サンプルに対する難易度付与結果 (有効 bigram 数が 300 以上のサンプル)

サブコーパス-媒体	1	2	3	4	5	6	7	8	9	計	平均	分散
出版-書籍 (PB)	288	534	1341	1369	1933	1290	1251	1181	717	9904	5.34	4.38
出版-雑誌 (PM)	100	58	209	559	606	166	171	65	4	1938	4.57	2.40
出版-新聞 (PN)	6	5	25	211	333	247	177	80	13	1097	5.54	1.84
図書館-書籍 (LB)	490	727	1993	1844	2396	1519	881	305	215	10370	4.53	3.19
白書 (OW)	0	0	0	0	2	11	111	969	397	1490	8.17	0.35
教科書 (OT)	47	2	7	22	85	69	125	29	0	386	5.46	4.07
広報紙 (OP)	0	0	0	0	27	90	162	73	2	354	6.81	0.75
ベストセラー (OB)	24	144	405	359	280	111	35	6	3	1367	3.91	1.75
Yahoo!知恵袋 (OC)	145	290	2945	2995	1596	309	290	113	111	8794	4.01	1.75
Yahoo!ブログ (OY)	887	1753	3825	3421	2063	561	421	142	44	13117	3.64	2.18
韻文 (OV)	69	16	104	58	5	0	0	0	0	252	2.66	1.35
法律 (OL)	0	0	0	0	0	0	0	0	334	334	9.00	0.00
国会会議録 (OM)	0	0	0	0	53	54	19	12	21	159	6.33	1.83
書籍 (PB+LB)	778	1261	3334	3213	4329	2809	2132	1486	932	20274	4.93	3.94

サブコーパスを明示していないものは、特殊目的サブンコーパスに属する。

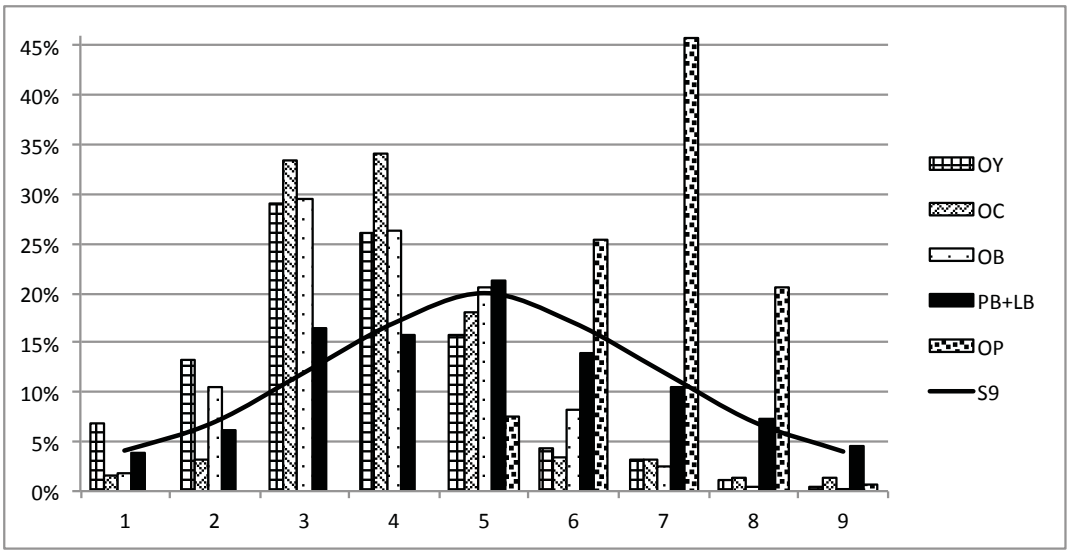


図 2: 可変長サンプル (OY, OC, OB, PB+LB, OP) の難易度分布

存在する[5]。『帯 2』が難易度を安定して決定するためには、少なくとも数百の有効 bigram が必要である。そのため、難易度分布の分析では、難易度を付与したすべてのサンプルを対象とするのではなく、サンプルサイズに対して適当な閾値を設定し、閾値以上のサンプルのみを対象とするのが適切である。

表 6 に、有効 bigram 数が 300 以上のサンプルにおける難易度分布を示す。可変長サンプルのレジスタのうち、固定長サンプルに含まれる出版-書籍 (PB)、雑誌 (PM)、新聞 (PN)、図書館-書籍 (LB)、白書 (OW) については、固定長サンプルと同じ傾向であるので、以下では議論しない。また、韻文 (OV)、法律 (OL)、国会議事録 (OM)、教科書 (OT) の各レジスタは、その特殊性を勘案し、ここでは議論の対象としない。以下では、広報紙 (OP)、ベストセラー (OB)、Yahoo!知恵袋 (OC)、Yahoo!ブログ (OY) について議論する。図 2 に、これらのレジスタの難易度分布の棒グラフを示す。なお、比較のため、書籍 (PB+LB) の難易度分布も合わせて示した。

[5] このような場合、『帯 2』は難易度判定不能として、0 を出力する。

Yahoo!知恵袋 (OC) と Yahoo!ブログ (OY) は、短いサンプルが多く、前者は 9.6% (8794/91445)、後者は 24.9% (13117/52680) のみが、分析対象となっている。分析対象となったサンプルの半数以上が難易度 3 または 4 となる。この 2 つのレジスタの中では、Yahoo!ブログ (OY) の方が難易度の平均は低く、分散は大きい。

ベストセラー (OB) は、書籍全体 (PB+LB) と比較して、かなりやさしい方に偏る。ピークは難易度 3、これに難易度 4 を含めると、全体の 56%に達する。分散は 1.75 で、図書館-書籍 (LB) の分散 3.19 と比べて、かなり小さい。

広報紙 (OP) のピークは、難易度 7 であり、全体の 46%を占める。難易度 6 から 9 のサンプルを合わせると、全体の 91%を占める。ピークが難易度 8 にある白書 (OW) よりはやさしいが、新聞 (PN) と比較すると、平均難易度は高く、分散は小さい。

3.3 レジスタの難易度序列

固定長サンプルおよび可変長サンプルに付与された難易度の分布を総合し、検討の対象としたレジスタ群に難易度の序列を付けると、次のようになろう。

$$\mathrm{OY} < \mathrm{OC} < \mathrm{OB} < \mathrm{LB} < \mathrm{PM} < (\mathrm{PB+LB}) < \mathrm{PB} < \mathrm{PN} < \mathrm{OP} < \mathrm{OW}$$

この結果は、我々が日常的に感じている印象とほとんど矛盾しない。少しだけ意外なのは、広報紙 (OP) の難易度の高さであろうか。この点については、さらに調査を進める必要があるが、広報紙に含まれる、比較的長い漢字連続 (たとえば、「札幌市危機管理対策室」、「緊急輸送道路沿道建築物」、「医療保険年金課高齢者医療係」) が、obi2/B9 難易度を上げている可能性がある。

相対難易度の規準として、我々は、前回より、書籍 (PB+LB) のサンプルを用いている。これは、BCCWJ において、書籍のサンプルが、生産および流通の両側面において適切にサンプリングされており、均衡コーパスとして最もふさわしいという判断に基づく選択であった。今回の BCCWJ の全サンプルに対する難易度付与の結果、書籍 (PB+LB) の難易度の分散は、他のレジスタと比較して、相対的に大きいことが明らかになった。すなわち、難易度の値に多様性があるという観点においても、書籍レジスタを相対的難易度の規準としてを用いるという選択は、妥当である。

3.4 ジャンル別の難易度分布

BCCWJ の書誌情報データには、ジャンル情報が含まれている。ここでは、出版-書籍 (PB)、図書館-書籍 (LB)、ベストセラー (OB) の 3 つのレジスタに対し、ジャンル別に難易度を集計した。日本十進分類法 (NDC) の第 1 次区分に対する難易度の平均値および分散を表 7 に示す。この表の「割合」は、その区分の全体に占める割合を示す。なお、これらは、有効 bigram が 300 以上の可変長サンプルに対する集計結果である。

先に示したように、この 3 つのレジスタにおける難易度の序列は、PB>LB>OB である。表 7 に示すように、平均難易度が高い「3 社会科学」の割合は PB が最も大きく、逆に、平均難易度が低い「9 文学」の割合は OB が最も大きい。このことが、全体の平均難易度が PB>LB>OB となる、主に要因である。

しかし、その一方で、各区分別に平均難易度を比較すると、7 つの区分で、PB>LB>OB という序列が観察される。つまり、ジャンルを固定した場合でも、一般的傾向として、PB>LB>OB という難易度の序列が観察される。

BCCWJ の書誌情報データには、日本十進分類法 (NDC) の他に、「C コード (図書分類コード)」が付与されている。このコードの左から 1 桁目は「販売対象コード」で対象読者を表す。2 桁目は「発行形態コード」で、発行形態を表す。これらのコード別の難易度の平均値と分散を、表 8 および表 9

表 7: NDC 別の難易度分布 (PB, LB, OB)

	出版-書籍 (PB)			図書館-書籍 (LB)			ベストセラー (OB)			
	平均	分散	割合	平均	分散	割合	平均	分散	割合	P>L>O
0 総記	5.91	3.66	(0.032)	5.35	2.83	(0.023)	4.60	2.99	(0.029)	√
1 哲学	5.07	1.18	(0.054)	5.20	1.08	(0.049)	4.55	0.93	(0.102)	
2 歴史	5.10	1.27	(0.086)	5.10	1.05	(0.103)	4.46	0.60	(0.046)	
3 社会科学	6.76	3.13	(0.253)	5.84	2.82	(0.197)	5.29	1.88	(0.127)	√
4 自然科学	6.40	2.76	(0.103)	5.60	1.95	(0.061)	4.96	1.44	(0.020)	√
5 技術・工学	5.92	5.78	(0.091)	5.05	5.07	(0.061)	3.58	3.31	(0.031)	√
6 産業	6.06	3.35	(0.044)	5.46	2.57	(0.033)	4.67	0.89	(0.011)	√
7 芸術・美術	4.71	1.53	(0.064)	4.50	1.45	(0.078)	3.25	1.31	(0.070)	√
8 言語	5.44	1.71	(0.018)	5.02	1.55	(0.018)	4.88	0.36	(0.012)	√
9 文学	3.15	1.06	(0.214)	3.27	1.02	(0.333)	3.43	0.98	(0.530)	
分類なし	4.73	5.22	(0.043)	2.51	4.18	(0.043)	3.41	1.48	(0.021)	–
計	5.34	4.38	(1.000)	4.53	3.19	(1.000)	3.91	1.75	(1.000)	

表 8: 販売対象コード別の難易度分布 (PB, LB, OB)

	出版-書籍 (PB)			図書館-書籍 (LB)			ベストセラー (OB)			
	平均	分散	割合	平均	分散	割合	平均	分散	割合	P>L>O
0 一般	4.42	2.62	(0.544)	4.31	2.32	(0.737)	3.82	1.59	(0.640)	√
1 教養	6.30	2.56	(0.044)	5.74	2.51	(0.068)	5.42	2.66	(0.014)	√
2 実用	6.80	4.95	(0.059)	5.50	4.74	(0.041)	4.88	3.36	(0.006)	√
3 専門	7.38	1.58	(0.197)	6.78	1.93	(0.062)	7.00	1.00	(0.003)	
5 婦人	3.25	5.49	(0.002)	2.48	2.85	(0.004)				
6 学参 I(小中)	3.20	3.36	(0.001)	4.62	5.48	(0.001)				
7 学参 II(高校)	6.50	1.25	(0.000)	5.00	0.00	(0.000)				
8 児童	2.10	2.27	(0.016)	1.90	1.89	(0.037)				
9 (雑誌扱い)	4.71	4.56	(0.034)	4.17	4.89	(0.007)				
コードなし	5.80	3.70	(0.103)	4.64	2.72	(0.043)	3.99	1.79	(0.337)	–

に示す。これらの分類においても、一般的傾向として、PB>LB>OB という難易度の序列が観察される。

以上の調査結果を総合すると、「出版-書籍 (PB)> 図書館-書籍 (LB)> ベストセラー (OB)」という難易度の序列は、内容、対象読者、発行形態を固定しても、かなり一般的に成り立つと判断することができる。これら 3 つのレジスタの母集団は、それぞれ、

出版-書籍 (PB) 2001 年から 2005 年の間に国内で出版された書籍

図書館-書籍 (LB) 都内 13 自治体以上の図書館が共通して所蔵している書籍

ベストセラー (OB) 1976 年から 2005 年までの 30 年間において、『出版年鑑』および『出版指標年報』のいずれかに、各年のベストセラーとして上記 20 位までに挙げられた書籍 951 冊。

である [1, 5]。つまり、「多くの人々に読まれている」という観点では、PB<LB<OB という順序となる。これは、PB>LB>OB という難易度序列とは、ちょうど反対になる。これらのことから、「多くの人々に読まれる書籍は、難易度の低いものに偏る」という帰結を導くことができよう。

4 おわりに

本稿では、BCCWJ リリース版を用いた obi2/B9 の再構成と、再構成した obi2/B9 を用いた BCCWJ リリース版全サンプルへの難易度付与について報告した。今回構成した難易度スケール B9 を含む『帯 2』システム、および、BCCWJ リリース版の全サンプルに対する難易度データは、準備ができ次第、http://kotoba.nuee.nagoya-u.ac.jp において公開する予定である。

表 9: 発行形態コード別の難易度分布 (PB, LB, OB)

	出版-書籍 (PB)			図書館-書籍 (LB)			ベストセラー (OB)			
	平均	分散	割合	平均	分散	割合	平均	分散	割合	P>L>O
0 単行本	5.76	4.09	(0.584)	4.86	3.08	(0.532)	3.93	1.83	(0.489)	√
1 文庫	3.55	1.59	(0.135)	3.44	1.41	(0.208)	2.50	0.25	(0.001)	√
2 新書	4.17	3.10	(0.060)	4.62	3.04	(0.079)	3.75	1.60	(0.124)	
3 全集・双書	5.92	4.30	(0.078)	4.85	4.14	(0.123)	3.83	0.44	(0.039)	√
4 ムック・その他	4.71	4.63	(0.034)	4.20	4.67	(0.008)				
5 事・辞典	6.12	5.03	(0.002)	5.42	3.58	(0.005)	6.50	0.25	(0.001)	
6 図鑑	3.00	4.50	(0.001)	3.53	5.31	(0.002)				
7 絵本	2.71	2.49	(0.001)	2.00	1.71	(0.001)				
9 コミック	4.36	2.05	(0.001)	2.50	0.25	(0.000)	2.50	0.25	(0.009)	
コードなし	5.80	3.70	(0.103)	4.64	2.72	(0.043)	3.99	1.79	(0.337)	–

謝辞　本論文の 3.4 節 (ジャンル別の難易度分布) は、国立国語研究所の丸山岳彦氏の助言に基づくものである。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。本研究の一部は、国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものである。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] 国立国語研究所コーパス開発センター. 『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.

[2] 松吉俊, 近藤陽介, 橋口千尋, 佐藤理史. 全教科を収録対象とした日本語教科書コーパスの構築. 言語処理学会第 14 回年次大会発表論文集, pp. 520–523, 2008.

[3] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

[4] 佐藤理史, 柏野和佳子. テキストの難易度に対する人間の判断と機械の判断. 第 1 回コーパス日本語学ワークショップ予稿集, pp. 195–202, 2012.

[5] 丸山岳彦, 山崎誠, 柏野和佳子, 佐野大樹, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子. 『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用. 国立国語研究所内部報告書, LR-CCG-10-01, 国立国語研究所, 2011.

文末機能表現シソーラスと述部正規化システム

松木 久幸 (名古屋大学大学院 工学研究科) †
佐藤 理史 (名古屋大学大学院 工学研究科) ‡
駒谷 和範 (名古屋大学大学院 工学研究科) ††

A Thesaurus of Predicate Functional Expressions and its Application to Predicate Standardization

Hisayuki Matsuki (Graduate School of Engineering, Nagoya University)
Satoshi Sato (Graduate School of Engineering, Nagoya University)
Kazunori Komatani (Graduate School of Engineering, Nagoya University)

1 はじめに

日本語には、文末の用言に接続し多様な意味を表す機能表現が多数存在する。これらの表現を**文末機能表現**と呼ぶ。文末機能表現には、ほぼ同じ意味を表す表現が複数存在する。このため、テキストマイニングにおける意見の集約など、文の意味の同一性の判定が必要となるタスクにおいては、同一の意味を持つとみなす文末機能表現を、一つの代表表現に置き換える**正規化** が必要となる [1]。たとえば、あるタスクでは、〈依頼〉の意味[1]を表す「〜テください」「〜テほしい」「〜テくれるか」を、すべて一つの代表表現（たとえば、「〜テください」）に正規化することが求められる。

このような正規化は、それぞれの文末機能表現に代表表現を定義することによって実現することができると考えられる。しかしながら、実際には、問題はそれほど単純ではない。第一に、文末機能表現の単位が問題となる。たとえば、「〜テもらう」と「〜たい」は、それぞれ単独で文末の用言に接続する一方で、結合した「〜テもらいたい」という形でも文末の用言に接続する。単独で用言に接続する場合は、それぞれ〈受益〉、〈希望〉の意味を表すが、結合した場合は〈願望〉の意味となる。このため、「〜テもらう」と「〜たい」に代表表現を定義するだけでは不十分であり、「〜テもらいたい」にも代表表現を定義する必要がある。この問題を解決するために、我々は、用言直後から文末までの長い単位を文末機能表現と捉え、この単位を見出し語とする辞書（シソーラス）を作成する。すなわち、「〜テもらう」と「〜たい」だけを見出し語とするのではなく、「〜テもらいたい」も見出し語に含める。なお、本論文では、「〜テもらう」と「〜たい」のような短い単位を、文末機能表現の構成要素と呼ぶことがある。

第二の問題は、どの範囲の表現を一つの代表表現に正規化すべきかが、応用タスクに依存して定まるという点である。たとえば、事実か推量かの判定だけが必要となるタスクでは、「〜にちがいない」と「〜かもしれない」は、どちらも〈推量〉の意味を表す「〜だろう」に正規化することができよう。一方、推量の確信度も考慮したいタスクでは、これらの表現を一つの代表表現に正規化するのは、明らかに不適切である。この応用タスク依存性の問題を解決するために、我々は、文末機能表現を意味分類した辞書（文末機能表現シソーラス）に代表表現を直接定義するのではなく、ユーザー定義に従って、この辞書から正規化用の辞書を生成するアプローチを採用する。

図１に、作成したシステムの全体像を示す。この図に示す通り、作成したシステムは、３つのサブシステムから構成される。以下、本論文では、これらのサブシステムについて説明する。

†h_matuki@nuee.nagoya-u.ac.jp
‡ssato@nuee.nagoya-u.ac.jp
††komatanii@nuee.nagoya-u.ac.jp
[1]本論文では、機能表現の意味を表す際に、記号 '〈〉' を用いる。

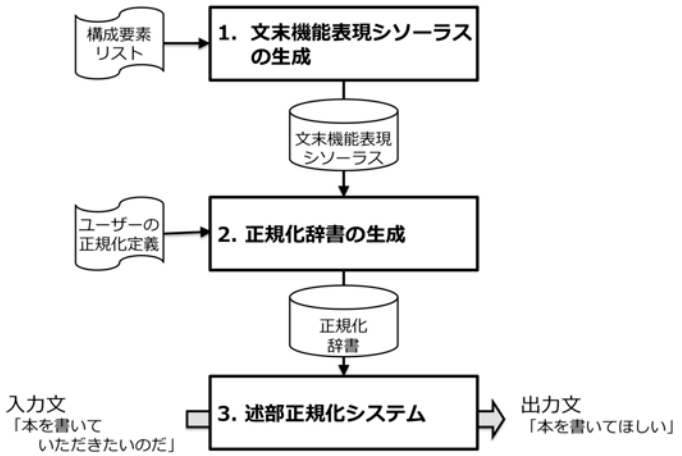


図 1: 本研究の全体像

2　文末機能表現シソーラスの生成

本節では、文末機能表現シソーラスの生成について述べる。このシソーラスの見出し語リストには、以前の研究 [5] に基づいて、作成した文末機能表現のリストを使用する。このリストは、人手で定義した 165 種類の構成要素から合成されている。この見出し語リストに意味表現を付与し、文末機能表現シソーラスを作成する。

2.1　意味体系

文献 [2, 3, 4] を参考に、本シソーラスで使用する意味体系を設計した。この意味体系は木構造の形式をとる（表 1）。この木構造の葉ノードが、意味を表現する最小構成要素であり、これを**意味ラベル**と呼ぶ。意味ラベルは、根ノードからのパス表現で表す。以下に、意味ラベルの例を示す。

(1) 〈/モダリティ/表現類型/意志/つもりだ〉

(2) 〈/モダリティ/表現類型/意志/意志形〉

例 (1) は、機能表現「つもりだ」に付与される意味ラベルで、その意味が、最上位レベルでは〈モダリティ〉に区分され、第２レベルでは〈表現類型〉、第３レベルでは〈意志〉、最下位レベルでは〈つもりだ〉に区分されることを表している。意味体系の最下位レベルのほとんどは、具体的な機能表現（リテラル）に対応する。

例 (2) は、活用形に対応する意味ラベルの例である。このような意味ラベルは、7 種類あり、それらは、タ形、意志形、命令形、推量形、省略意志形、文語命令形、省略推量形に対応する[2]。なお、リテラルと活用形以外の最下位レベルには、〈デス列〉、〈可能形〉、〈否定形〉、〈謙譲語〉の４つがある。

本論文では、以降、意味ラベルの最下位レベルを除いた部分を、略記表現（表 1 の最左列）で表す。たとえば、〈/ヴォイス/可能/ことができる〉を、〈可能/ことができる〉と表す。また、最下位レベルを省略し、〈可能〉と書くこともある。

本シソーラスでは、この意味ラベルを組み合わせて、文末機能表現の意味を表現する。すなわち、文末機能表現の意味表現は、意味ラベル列となる。なお、意味ラベルの接続は、記号 '=' で表す。たとえば、〈開始/テいる〉と〈必然性/はずだ〉の接続は、〈開始/テいる=必然性/はずだ〉と表す。

2.2　意味表現の付与

見出し語に対する意味表現の付与は、以下の手順で行う。

1. 文末機能表現の構成要素に対して、人手で意味表現を定義する。

[2] これ以外の活用形は、意味を持たないものとして扱う。

表 1: 意味体系

上位レベルの略記表現	意味ラベル 上位レベル	最下位レベル
〈 可能 〉	/ヴォイス/可能/	ことができる, こともできる, Rうる, 可能形
〈 受身 〉	/ヴォイス/受身/	(ら) れる
〈 使役 〉	/ヴォイス/使役/	(さ) せる
〈 過去 〉	/テンス/過去/	タ形
〈 否定 〉	/肯否/否定/	ぬ, ない, テない, Rない, 否定形
〈 開始 〉	/アスペクト/開始/	Rはじめる, Rだす, Rかける
〈 継続 〉	/アスペクト/継続/	ている, てる, ていく, てくる, Rつづける, Rつつある
〈 終了-無意志 〉	/アスペクト/終了/無意志/	Rおわる, Rやむ, Rあがる
〈 終了-有意志 〉	/アスペクト/終了/有意志/	Rおえる, Rあげる
〈 完遂 〉	/アスペクト/完遂/	Rきる, Rぬく, Rつくす, Rとおす, テしまう
〈 場面 〉	/アスペクト/場面/	ところだ, ばかりだ
〈 残存 〉	/アスペクト/残存/	テおく, テある
〈 意志 〉	/モダリティ/表現類型/意志/	テみる, テみせる, つもりだ, つもりがある, つもりもある, 意志形, 省略意志形
〈 希望 〉	/モダリティ/表現類型/希望/	Rたい
〈 願望 〉	/モダリティ/表現類型/願望/	テほしい
〈 依頼 〉	/モダリティ/表現類型/依頼/	テくれるか
〈 命令 〉	/モダリティ/表現類型/命令/	命令形, 文語命令形
〈 疑問 〉	/モダリティ/表現類型/疑問/	か, かい, かしら
〈 推量 〉	/モダリティ/真偽判断/推量/	だろう, 推量形, 省略推量形
〈 可能性 〉	/モダリティ/真偽判断/蓋然性/可能性/	かもしれない, ことがある, こともある
〈 必然性 〉	/モダリティ/真偽判断/蓋然性/必然性/	はずだ, にちがいない
〈 証拠性-観察 〉	/モダリティ/真偽判断/証拠性/観察/	Rそうだ, ようだ, みたいだ
〈 証拠性-伝聞 〉	/モダリティ/真偽判断/証拠性/伝聞/	らしい, そうだ, という, とのことだ
〈 必要 〉	/モダリティ/価値判断/必要/	ざるをえない
〈 適当 〉	/モダリティ/価値判断/適当/	バよい, バいい, タラいい, ほうがいい, べきだ
〈 許容 〉	/モダリティ/価値判断/許容/	テよい, テいい
〈 非許容 〉	/モダリティ/価値判断/非許容/	テハだめだ, タラだめだ, テハならない, タラいけない
〈 説明 〉	/モダリティ/説明/	のだ, ことだ, ものだ, わけだ
〈 丁寧さ 〉	/モダリティ/伝達/丁寧さ/	です, ます, Rなさる, Rくださる, デス列, 謙譲語
〈 態度 〉	/モダリティ/伝達/態度/	かよ, よ, ね, ぜ, ぞ,っけ,っけね,って,ってね, とか, な, なあ, ね, よ, よね, よねえ, わ, わね, わよ
〈 内-授与 〉	/その他/授受/内授与/	テあげる, テやる
〈 他-授与 〉	/その他/授受/他授与/	テくれる
〈 受益 〉	/その他/授受/受益/	テもらう
〈 難易-易 〉	/その他/難易/易しい/	Rやすい, Rよい, Rいい
〈 難易-難 〉	/その他/難易/難しい/	Rがたい, Rにくい, Rづらい
〈 不履行 〉	/その他/不履行/	Rかねる, Rしぶる, Rわすれる, Rそこなう, Rそんじる, Rそびれる
〈 語彙的 〉	/その他/語彙的/	Rすぎる, とする, Rなおす, Rかえす, Rがちだ, Rあう
〈 動詞化 〉	/その他/動詞化/	Rなる, Rする, Rある
〈NONE〉	/その他/NONE/	バならない, バいけない

2. 構成要素の列に対し、意味表現を機械的に合成する。
3. 直前の用言の活用形がもつ意味表現を、文末機能表現の意味表現に追加する。
4. 非構成的意味をもつ構成要素列の意味表現を書き換える。

2.2.1 構成要素に対する意味表現の定義

まず、見出し語リストの合成に用いた 165 種類の構成要素に対し、人手で意味表現を定義した。このうち、120 個の構成要素に対応する意味表現は、意味体系の設計の段階で、意味ラベルとして取り込まれているため、その対応は自明である。

残り 45 個の構成要素に対しては、意味ラベルの列を定義した（表 2)。この表の I は、否定に関わるグループで、たとえば、「ことがない」の意味表現を、〈 可能性=否定 〉と定義した。II は、可能形のグループで、たとえば、「Rだせる」の意味表現を〈 開始=可能 〉と定義した。III は、謙譲語のグループで、たとえば、「テいただく」の意味表現を、〈 受益=丁寧さ 〉と定義した。XI は、接尾辞や終

表 2: 構成要素に定義した意味ラベル列

	意味ラベル列	構成要素の例	数	意味ラベル列	構成要素の例	数
Ⅰ	継続=否定	Rつつない	3	可能性=否定	ことがない	2
	必然性=否定	はずがない	2	説明=否定	わけがない	1
	意志=否定	まい	3			
Ⅱ	開始=可能	Rだせる	1	完遂=可能	Rきれる	5
	終了-無意志=可能	Rおわれる	2	残存=可能	テおける	1
	内-授与=可能	テやれる	1	受益=可能	テもらえる	1
	受益=丁寧さ=可能	テいただける	1	不履行=可能	Rおとせる	1
	語彙的=可能	Rあえる	2			
Ⅲ	内-授与=丁寧さ	テさしあげる	1	他-授与=丁寧さ	テくださる	1
	受益=丁寧さ	テいただく	1			
Ⅸ	受身=否定	(ら) れない	2	受身=必要	(ら) れざるをえない	1
	受身=丁寧さ	(ら) れます	1	使役=否定	(さ) せない	2
	使役=必要	(さ) せざるをえない	1	使役=丁寧さ	(さ) せます	1
	使役=受身=否定	(さ) せ (ら) れない	2	使役=受身=丁寧さ	(さ) せ (ら) れます	1
	使役=受身	(さ) せ (ら) れる	1	使役=受身=必要	(さ) せ (ら) れざるをえない	1
	疑問=態度	かね	3			

助詞の意味表現からなるグループで、たとえば、「(ら) れます」の意味表現を、⟨ 受身=丁寧さ ⟩ と定義した。

2.2.2 構成要素の列に対する意味ラベルの合成

先に述べたように、シソーラスの見出し語は、構成要素を接続することによって生成される。この構成要素の接続と同時に、意味表現の合成を行う。以下に、構成要素 L の直後に構成要素 R を接続する手順を示す。なお、活用形を k_i、構成要素または活用形 X の意味表現を $s(X)$ と表す。

1. 辞書形として与えられる L に対し、可能な活用形 k_i に対応した L_i を作成する。この際、L_i の意味表現 $s(L_i)$ を、$s(L_i) = s(L) + s(k_i)$ により合成する。
2. L_i の直後に R を接続できる場合、それらを接続した列 L_iR を生成する。この際、L_iR の意味表現を、$s(L_iR) = s(L_i) + s(R)$ により合成する。

なお、この手順で、構成要素の接続性は、L_i が R に対して定義される左接続条件[3]を満たし、かつ、L_iR のいずれかの表記が『現代日本語書き言葉均衡コーパス』に出現する場合に、接続できると判定する。

例として、「テいる (L)」の直後に「のだ (R)」を接続する場合を考える。手順 1 で、L に対して、終止形「テいる (L_1)」、テ形「テいて (L_2)」、タ形「テいた (L_3)」などを生成する。ここで、終止形やテ形は、対応する意味表現をもたないので、$s(k_i)$ は空となり、$s(L_1) = s(L_2) = s(L) =$⟨ 継続 ⟩となる。一方、タ形は、意味表現 $s(k_3) =$⟨ 過去 ⟩をもつので、$s(L_3) =$⟨ 継続=過去 ⟩となる。この後、直後に R(「のだ」) が接続可能な L_1 と L_3 に対して手順 2 が適用され、「テいるのだ (L_1R)」と「テいたのだ (L_3R)」が生成される。この際、$s(L_1R) =$⟨ 継続=説明 ⟩、$s(L_3R) =$⟨ 継続=過去=説明 ⟩が合成される。

見出し語生成は、このような構成要素の接続を繰り返すことによって実現される。すなわち、まず、構成要素 A と構成要素 B を接続することで、見出し語 AB を生成する。次に、こうして生成した見出し語 AB を新たな構成要素とし、さらに構成要素 C を接続することで、見出し語 ABC を生成する。このようにして、より長い文末機能表現が生成される。

[3]各構成要素には、その構成要素がどのような構成要素に接続可能であるかの制約が記述されている。これを、左接続条件と呼ぶ。

2.2.3 直前の用言の活用形がもつ意味表現

すでに述べたように、活用形のいくつかは、それ自身が意味をもつ。たとえば、「書こう（意志形）」は、⟨ 意志 ⟩という意味をもつ。一方、文末機能表現「つもりだ」は、やはり、⟨ 意志 ⟩という意味をもつ。しかしながら、「書こう」は、明示的な文末機能表現をもたないため、このままでは、「書こう」を「書くつもりだ」に正規化することができない。

この問題に解決するために、本シソーラスでは、文末用言の活用形を、表層形をもたない擬似的な文末機能表現とみなし、それに相当するエントリを設定する方法を採用する。具体的には、意味をもつ活用形 k それぞれに対し、表記が空文字列（λ）で、意味表現を $s(k)$ とするエントリを設定する。このエントリは、活用形 k にのみ接続可能とする。このようなエントリを設定することにより、たとえば、文末用言「書こう」は、用言「書く」の意志形に、意味表現⟨ 意志 ⟩をもつ表記 λ の文末機能表現が接続していると解釈することが可能となり、その結果、「書こう」を「書くつもりだ」に正規化することが可能となる。

上記で述べた擬似的な文末機能表現の導入と整合させるために、シソーラスの各エントリには、直前用言の活用形 k の意味情報を取り込む。具体的には、2.2.2 節で生成されたエントリの意味表現の左側に、直前用言の活用形の意味情報を追加する。たとえば、エントリ「のだ（⟨ 説明 ⟩）」に対して、タ形（⟨ 過去 ⟩）に接続する場合の意味表現⟨ 説明=過去 ⟩を合成する。この結果、 1 つのエントリは、直前用言の活用形に依存した、複数の意味表現をもつこととなる。

2.2.4 意味表現の書き換え

文末機能表現が非合成的意味をもつ場合、2.2.2 節や 2.2.3 節で述べた意味表現の合成法では、適切な意味表現を生成することができない。たとえば、「テもらいたい」の意味表現は、その構成要素から⟨ 受益=希望 ⟩と合成されるが、適切な意味表現は⟨ 願望 ⟩である。このような、意味の非合成性に対処するために、合成して得られた意味表現に書き換えルールを適用し、適切な意味表現へと変換する。現在までに、定義した書き換えルールを表 3 に示す。

この表に示すように、書き換えルールは、大きく 2 種類に分類される。I は、特定の文末機能表現に対するルール群である。たとえば、ルール I(1) は、文末機能表現「R なさる」の命令形「R なさい」の意味表現を⟨ 命令 ⟩へと書き換えるためのルールである。これに対して、II は、意味ラベルの上位レベルのみを参照するルール群である。たとえば、ルール II(3) は、⟨ 否定 ⟩の直後に⟨ 疑問 ⟩がくる場合、この部分を⟨ 疑問 ⟩へと書き換える。このルールによって、「ないかしら」「テないか」などの意味表現が⟨ 疑問 ⟩に書き換えられる。

書き換えルールの適用は、以下の手順で行う。ここで、書き換え対象の意味表現を $S = s_n \cdots s_0$、書き換え後の意味表現を T と表し、ルールは、k 個の意味ラベルの列を、l 個の意味ラベルの列に書き換えるものとする。

1. $i = 0$ とする。
2. 書き換え対象の意味表現 $s_{i+k-1} \cdots s_i (0 \leq i \leq n,\ 1 \leq k)$ に適用可能なルールをすべて求める。
3. 適用可能なルールが存在した場合、コストが最小のルールを適用し、書き換え後の意味表現を T の先頭に追加する。$i = i + k$ とする。
4. 適用可能なルールが存在しなかった場合、s_i を T の先頭に追加し、$i = i + 1$ とする。
5. $i \leq n$ の場合、手順 2 に戻る。

たとえば、構成要素から合成される「テおいてくれるか」の意味表現は、⟨ 残存=他-授与=疑問 ⟩である。$i = 0$ で、ルール II(1) が適用可能であり、⟨ 他-授与=疑問 ⟩が⟨ 依頼 ⟩に書き換えられる。この結果、「テおいてくれるか」の意味表現は、⟨ 残存=依頼 ⟩となる。

表 3: 現在までに定義した書き換えルール

		書き換え前	書き換え後	コスト	適用する表現の例
I	(1)	*/Rなさる = */命令形	⟨ 命令 ⟩	3	Rなさい
	(2)	*/テやる = */謙譲語 = */命令形	⟨ 依頼 ⟩	3	テください
	(3)	*/Rかねる = */ない	⟨ 可能性 ⟩	4	Rかねない
	(4)	説明/* = */Rする	⟨ 意志 ⟩	4	ようにする
	(5)	否定/* = */バならない	⟨ 必要 ⟩	4	なければならない
		否定/* = */バいけない	⟨ 必要 ⟩	4	なければいけない
	(6)	意志/* = */とする	⟨ 意志 ⟩	4	ウとする
II	(1)	他-授与/* = 疑問/*	⟨ 依頼 ⟩	4	テくれるか
		他-授与/* = 丁寧さ/* = 疑問/*	⟨ 依頼=丁寧さ ⟩	4	テくださるか
		受益/* = 可能/* = 疑問/*	⟨ 依頼 ⟩	4	もらえるか
		受益/* = 丁寧さ/* = 可能/* = 疑問/*	⟨ 依頼=丁寧さ ⟩	4	テいただけるか
	(2)	受益/* = 希望/*	⟨ 願望 ⟩	4	テもらいたい
		受益/* = 丁寧さ/* = 希望/*	⟨ 願望=丁寧さ ⟩	4	テいただきたい
	(3)	否定/* = 疑問/*	⟨ 疑問 ⟩	6	ないか
	(4)	意志/* = 意志/*	⟨ 意志 ⟩	6	テみよう (「テみる」の意志形)
		推量/* = 推量/*	⟨ 推量 ⟩	6	でしょう (「だろう」の推量形)

注意：ルール中の記号 '*' は、任意の上位クラス、あるいは、最下位クラスにマッチすることを表す。

表 4: 作成したシソーラスのエントリの例

見出し /準用言/Rはじめる–/連用形/テ形=/準用言/ている–/辞書形	**ID** X010=X510
表記 はじめている, 始めている	**頻度** 364,559
活用型 /動詞/母音動詞	**活用形** /辞書形
左接続条件 /準用言/連用形	**意味表現** ⟨ 開始=継続 ⟩
見出し /助動詞/のだ–/デス列/終止形=/終助詞/か–/辞書形	**ID** Y200=Z010
表記 のですか, んですか	**頻度** 13996,15076
活用型 /無活用型	**活用形** /辞書形
左接続条件 /助動詞/のだ/タ形接続; /助動詞/のだ/基本形接続	**意味表現** ⟨ 過去=説明=丁寧さ=疑問 ⟩; ⟨ 説明=丁寧さ=疑問 ⟩
見出し /用言活用形/意志形	**ID** P020
表記 λ	**頻度** 1
活用型 –	**活用形** –
左接続条件 /用言活用形/意志形接続	**意味表現** ⟨ 意志 ⟩

2.3 作成したシソーラス

以上の手順で、45,948 エントリからなるシソーラスを作成した。エントリの例を表 4 に示す。この表に示すように、1 つのエントリは**見出し**、**ID**、**表記**、**頻度**、**活用型**、**活用形**、**左接続条件**、**意味表現**の 8 つの要素からなる。

これらの要素のうち、**見出し**と **ID** は、そのエントリに固有な識別子である。**表記**は、そのエントリの出現形を示す。出現形が複数存在する場合は、それらが列挙される。**頻度**は、表記に記述された出現形が、『現代日本語書き言葉均衡コーパス』に出現する回数を示す。**活用型**と**活用形**は、そのエントリの活用情報を表す。

左接続条件は、そのエントリが接続可能な用言に対する制約情報を表現している。具体的には、接続可能な用言の集合を規定するラベル（左接続キー）を列挙する。たとえば、助動詞「のだ」に対して使用される左接続キー "/助動詞/のだ/タ形接続" は、動詞・形容詞のタ形に接続可能であることを示す。

意味表現には、2.2.3 節で述べたように、直前用言の活用形に依存した複数の意味ラベル列が列挙される。これらは、左接続条件で列挙される左接続キーに対応する。

2.4 活用形展開シソーラス

前節で述べたシソーラスのエントリのうち、活用するエントリの活用形は、すべて辞書形である。それに対して、実際の文の文末は、タ形や意志形などの活用形をとりうる。これに対応するために、

先に述べたシソーラスから、文末になりうるすべての活用形を生成（展開）したシソーラスを自動生成する。この結果、45,948 エントリのシソーラスから、140,233 エントリの活用形展開シソーラスが得られた。

3　正規化辞書の生成

述部正規化システムは、前節で述べた活用形展開シソーラスをそのまま使用するのではなく、ユーザーの正規化定義に基づいて意味表現を書き換えることによって生成される**正規化辞書**を使用する。正規化辞書生成のための意味表現の書き換えは、２段階に分けて行う。なお、述部正規化システムでは、代表表現は頻度に基づいて自動的に決定するため、正規化辞書では、代表表現を明示的には定義しない。

3.1　意味ラベル単体に対する整理・集約

第１段階の書き換えでは、意味ラベルの整理・集約を行う。ユーザーは、文末機能表現シソーラスのそれぞれの意味ラベルに対して、どのような書き換えを行うかを定義する。システムは、この定義に従って、シソーラスのエントリの意味表現を書き換える。

意味ラベルの書き換えには、次の３種類がある。

1. シソーラスの意味ラベルをそのまま使用する（書き換えない）。
2. 意味ラベルを削除する。（つまり、ある種の意味情報を無視する）
3. 新たな意味ラベルへと書き換える。（つまり、シソーラスが採用している意味体系に含まれない、独自の意味区分を導入する。）

シソーラスの意味ラベルの種類は、リテラルや活用形に対応する最下位レベルを除くと、39 種類である。この粒度で正規化したいのであれば、意味ラベルを書き換える必要はない。一方、より大きな粒度で正規化を行いたい場合は、特定の意味ラベルを削除するか、あるいは、複数の意味ラベルを統合する新しい意味ラベルを導入することで、これを実現する。意味ラベルの書き換えを指示するユーザー定義の例を、表５に示す。

ユーザー定義は、書き換えるべき意味ラベルの集合と、それに対する書き換えコマンドで構成される。意味ラベルの集合は、根ノードからのパス表現（あるいは、最下位レベルを省略した略記表現）で記述する。たとえば、定義 (7) のパス表現⟨/モダリティ/真偽判断/⟩は、このパスと前方一致する意味ラベルの集合、すなわち、⟨ 推量 ⟩、⟨ 可能性 ⟩、⟨ 必然性 ⟩、⟨ 証拠性-観察 ⟩、⟨ 証拠性-伝聞 ⟩からなる集合を意味する。

書き換えコマンドには、次の３種類がある。コマンド ‘+’ は、意味ラベルを書き換えないことを指示する。コマンド ‘−’ は、書き換え対象の意味ラベルを削除することを指示する。‘+’ と ‘−’ 以外のコマンドは、書き換え対象の意味ラベルを、そのコマンド（意味ラベル）へと書き換えることを指示する。たとえば、定義 (3) は、⟨ 説明 ⟩を削除することを指示する。この定義に基づき、「テいるのだ」の意味表現⟨ 継続=説明 ⟩が⟨ 継続 ⟩に書き換えられる。一方、定義 (5) と (6) は、それぞれ⟨ 願望 ⟩と⟨ 依頼 ⟩を⟨ 要求 ⟩に書き換えることを指示する。これらの定義に基づき、「テほしい⟨ 願望 ⟩」や「テくれるか⟨ 依頼 ⟩」が、いずれも⟨ 要求 ⟩に書き換えられる。

3.2　意味ラベル列に対する書き換え

第２段階の書き換えでは、整理・集約後の意味ラベルの列に対して、さらなる書き換えを行う。先の第１段階の書き換えは、意味ラベル単体に対して定義される。これに対して、第２段階の書き換えでは、隣接する意味ラベルを考慮して、書き換えるか否かを指定することができる。表６に、ユーザーの定義例と適用例を示す。

この表に示すように、ユーザーは、書き換え前の意味ラベル列、書き換え後の意味ラベル列、コス

表 5: 意味ラベル単体に対する定義の例

	もとの意味ラベル	コマンド		もとの意味ラベル	コマンド
(1)	⟨ 開始 ⟩	+	(2)	⟨ 継続 ⟩	+
(3)	⟨ 説明 ⟩	−	(4)	⟨/モダリティ/伝達/*⟩	−
(5)	⟨ 願望 ⟩	要求	(6)	⟨ 依頼 ⟩	要求
(7)	⟨/モダリティ/真偽判断/⟩	推測			

表 6: 意味ラベル列に対する定義の例

	書き換え前	書き換え後	コスト	適用例
(1)	⟨ 開始 = 意志 ⟩	⟨ 意志 ⟩	4	Rはじめよう
(2)	⟨ 終了 = 過去 ⟩	⟨ 過去 ⟩	4	テしまった
(3)	⟨ 開始 = 継続 ⟩	⟨ 継続 ⟩	6	Rはじめている
(4)	⟨ 継続 = 過去 ⟩	⟨ 過去 ⟩	6	テいた

トの３つ組を記述する。たとえば、定義 (3) は、⟨ 開始=継続 ⟩を⟨ 継続 ⟩に書き換えることを指示している。システムは、2.2.4 節と同様の方法で、この定義に基づき、意味ラベル列を書き換える。たとえば、定義 (3) に基づき、「Rはじめている」の意味表現⟨ 開始=継続 ⟩が⟨ 継続 ⟩に書き換えられる。その結果、「Rはじめている」は、「テいる⟨ 継続 ⟩」と同じグループに分類されることになる。

4 述部正規化システム

述部正規化システムは、前節で述べた正規化辞書に基づいて、文末機能表現を同定し、正規化する。本システムの構成を図 2 に示す。この図に示すように、本システムは、**文末機能表現同定システム**と**言い換えシステム**の２つのサブシステムから構成される。

4.1 文末機能表現同定システム

文末機能表現同定システムは、形態素解析済の文を受けとり、その文の文末機能表現を同定し、意味表現（意味ラベル列）を付与する。入力形態素列を $m_n \cdots m_0$(添字は文末に近いほど小さい)、正規化辞書のエントリを e とするとき、本システムは、以下の２つの条件を満たす形態素列 $m_i \cdots m_0 (-1 \leq i < n)$ のうち最大の i をとる形態素列[4]を、文末機能表現 e と同定する。

条件 1 形態素列 $m_i \cdots m_0$ の表記が、正規化辞書のエントリ e の表記と一致する。
条件 2 形態素 m_{i+1} は、エントリ e の左接続条件（左接続キー）を満たす。

なお、このような条件を満たす形態素列 $m_i \cdots m_0$ が存在しなかった場合は、入力文中に文末機能表現は存在しないと判定する。文末機能表現が同定された場合は、その形態素列に、対応するエントリ e の、条件を満たした左接続キーに対する意味表現を付与して出力する。

たとえば、入力形態素列「本/を/書き/始めて/いる/に/ちがいない」に対し、上記の２つの条件を満たすエントリとして、「にちがいない」、「テいるにちがいない」、「始めているにちがいない」の３つが見つかる。これらのうち、最も長い（最大の i をとる）「始めているにちがいない」を、文末機能表現と同定し、その意味表現⟨ 継続=推測 ⟩を付加して出力する。

4.2 文末機能表現言い換えシステム

入力文中に文末機能表現が同定された場合、それを代表表現に書き換えて出力する。この手順を以下に示す。なお、意味表現 s に対して、s に含まれるそれぞれの意味ラベルの最下位レベルを除いた

[4] $i = -1$ の場合は、空文字列 λ とする。

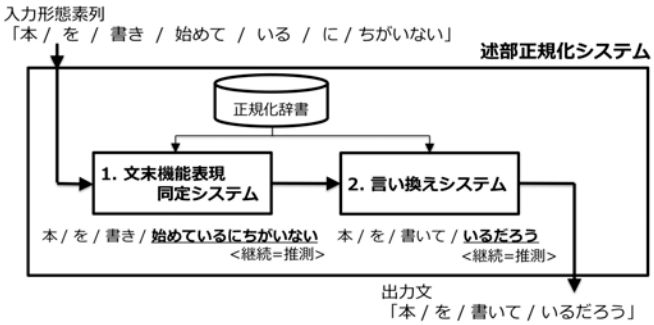


図 2: 述部正規化システムの全体像

ものを $g(s)$ と表す。

1. 代表表現の決定
   同定された文末機能表現 e に付与された意味表現 $s(e)$ とし、e の直前の形態素を m とするとき、次の 2 つの条件を満たす正規化辞書エントリ e' のうち、頻度が最も高い表記をもつエントリを代表表現 e_r とする。

   条件 1 エントリ e' のある左接続キーに対応する意味表現を $s(e')$ とするとき、$g(s(e')) = g(s(e))$ である。

   条件 2 用言 m は、条件 1 の左接続キーが要求する活用形をとりうる。

2. 代表表現への置換
   文末機能表現 e を代表表現 e_r に置き換える。このとき、必要があれば、用言 m の活用形を、e_r が接続できる形に変更する。

先に示したように、入力形態素列「本/を/書き/始めて/いる/に/ちがいない」に対して、「始めているにちがいない⟨ 継続=推測 ⟩」が同定される。意味表現⟨ 継続=推測 ⟩をもち、かつ、用言「書く」に接続可能なエントリとして、「R つつあるはずだ」「テいるだろう」などがみつかり、これらのうち、頻度が最も高い「テいるだろう」が代表表現として選ばれる。最終的に、「書く」の活用形が連用形からテ形に変更され、代表表現を接続した「本を書いているだろう」が生成される。

4.3 正規化の例

以下に、作例に対するシステムの出力を示す。なお、(1) には文末機能表現同定システムの出力を、(2) には言い換えシステムの出力を示し、同定された、あるいは、言い換えによって得られた文末機能表現を、下線部で示す。

例 1 「窓 / を / 開けて / いただき / たい / の / です」
(1) 窓 / を / 開けて / いただきたいのです⟨ 要求 ⟩
(2) 窓 / を / 開けて / ほしい⟨ 要求 ⟩

例 2 「対立 / 関係 / が / 生じ / かね / ない / の / かも / しれ / ない」
(1) 対立 / 関係 / が / 生じ / かねないのかもしれない⟨ 可能性 ⟩
(2) 対立 / 関係 / が / 生じる / かもしれない⟨ 可能性 ⟩

例 3 「席 / を / 譲って / くれ / ない / かしら」
(1) 席 / で / 譲って / くれないかしら⟨ 要求 ⟩
(2) 席 / で / 譲って / くれるか⟨ 要求 ⟩

5 おわりに

本論文では、文末機能表現シソーラスと、それを利用した述部正規化システムについて述べた。現時点の状況は、全体の枠組みがほぼ完成したところである。今後、意味体系の見直しや、シソーラスの意味表現書き換えルールを整備することにより、シソーラスをより良いものとすると同時に、述部正規化システムの性能向上を目指す予定である。

謝辞 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。本研究では，現代日本語書き言葉均衡コーパス DVD 版を利用した。

参考文献

[1] T. Izumi, K. Imamura, G. Kikui, and S. Sato. Standardizing Complex Functional Expressions in Japanese Predicates: Applying Theoretically-Based Paraphrasing Rules. In *23rd International Conference on Computational Linguistics*, p. 64, 2010.

[2] 日本語記述文法研究会（編）. 現代日本語文法 4 第 8 部 モダリティ. くろしお出版, 2003.

[3] 日本語記述文法研究会（編）. 現代日本語文法 3 第 5 部 アスペクト 第 6 部 テンス 第 7 部 肯否. くろしお出版, 2007.

[4] 益岡隆志. 日本語モダリティ探究. くろしお出版, 2007.

[5] 松木久幸, 佐藤理史, 駒谷和範. 文末機能表現シソーラスの編纂に向けて–文末機能表現の網羅的生成–. 言語処理学会 第 18 回年次大会 発表論文集, 2012.

論文の論理構造における分野基礎用語に関する分析

内山 清子（国立情報学研究所）†

An Analysis of Domain-Specific Introductory Terms in Logical Structure of Scholarly Papers

Kiyoko Uchiyama (National Institute of Informatics)

１．はじめに

学術論文には分野で使われる専門用語や、著者が自分の研究を特徴づけるために作り出す独自の専門的な複合語などが数多く含まれる。これらの語は、分野の初心者にとって初めて遭遇する用語であり、その用語の意味を理解した上で論文を読み進めることが必要となる。しかし、分野初心者にとって、専門用語はすべて未知の語であり、どの語が重要な語であり最初に学ぶべき用語であるのか、また対象論文の研究内容の手がかり語となる用語であるのかなどの区別ができない。こうした専門用語に対して優先度を示すことにより、分野初心者が論文を読んで理解するための支援になるのではないかと考えた。そこで、本研究では、対象分野において最初に必ず学ばなければならない語、その分野における基礎的・必須である専門用語を分野基礎用語と呼び、分野基礎用語の選定方法を検討し、論文の論理構造における出現分布について分析を行う。

２．関連研究と分野基礎用語の位置づけ

従来、分野の用語（専門用語）については、専門性や重要性といった指標や関連用語収集などのテーマで研究がおこなわれてきた。まず、専門度を推定する研究として、専門外の人に対して専門用語を使わずに平易な用語に置き換えるために、専門外の人から見て比較的専門的な用語か、かなり専門的な用語かの 2 段階に分けたものがある。次に用語の重要性については、複合語を構成している単語の種類や隣接する単語の数をベースにして用語らしさとしての重要性を計算する手法が提案されてきた。また、関連用語収集として、複数の書籍に共通する用語をシードワードに設定して、その用語から関連する用語を自動的に収集する研究が行われた。この研究におけるシードワードは、本研究における分野基礎用語と一部一致している。

本研究において、論文を理解するために効率的な用語として分野基礎用語を位置づけるために、分野基礎用語から始まり専門性・難易度が高い用語に至る学習段階を想定し、自分の知識と目標レベルに応じた以下の 4 段階の知識・学習レベルを設定した。

(1) 一般、大学学部生、他の研究分野の研究者
(2) 大学学部生（その分野を専門に学びたい学生）
(3) 大学院修士（修士論文テーマ探し）
(4) 大学院博士、研究者（博士論文、研究論文テーマ探し）

まず、第一段階の一般、大学学部生、他の研究分野の研究者に対しては、分野知識を持っていないことを前提として、分野の全体的な概略を説明した解説文や理解しやすい教科書などに掲載されている用語を提示することが有効であると考える。次は学部 3 年生を想定して、卒業論文をまとめるために必要な分野の成り立ちも含めた詳細な概要を把握する必要がある。この段階では分野でよく利用される用語の理解を深めることが重要となる。第 3 段階は、大学院修士の学生が自分の修士論文のテーマを探すために、その分野の最新動向も踏まえて、興味のあるトピックに関する論文を読む必要性が出てくる。この段階では、論文を読むために、よく使われる用語に関連した専門性の高い用語を学ぶ。最後の段階では、大学院博士課程の学生や研究者として、過去の詳細な研究成果も含めた狭く深い

† kiyoko_at_nii.ac.jp

情報が重要となってくる。この段階では、分野の中の特定のトピックに対する専門家が使っている専門性と難易度の高い知識を持っていることが前提となる。本論文では、このような4つの知識・学習段階を考えた中で、分野初心者に重要な最初のレベル（1 と 2）に必要な用語を分野基礎用語と位置付ける。

３．分野基礎用語の選定

分野基礎用語を抽出する対象分野を自然言語処理とした。これまで実験的に自然言語処理の研究者一名に、分野基礎性の定義を説明した上で、重要な自然言語処理用語を 308 語選定し第 2 章で説明した 4 段階に分類してもらった。内訳は 1 レベルが 20 用語、2 レベルが 186 用語、3 レベルが 89 用語、レベル 4 が 13 用語である。この正解セットを用いた自動抽出手法として、一般コーパス（毎日新聞）と専門コーパス（情報処理学会自然言語処理研究会で発表された論文）を比較して、対数尤度比、カイ二乗値、イエーツ補正カイ二乗値等の各尺度の平均精度や、C-Value による用語らしさの検定をして実験を行ってきた。

結果的に出現頻度に基づいて分野基礎用語を自動的に抽出することは難しく、精度が低かった。また分析結果から、抽出時のスコアランキングで基礎性の度合いをつけることが現実的ではないことがわかり、正解セット自体を再検討することにした。理想的な選定方法としては、専門家に分野基礎用語を選定してもらい、多くの専門家が共通して選定した用語は分野基礎用語であると決定することが考えられる。しかし、専門家の意見を数多く集めることが難しいため、専門家の判断と同等であると見なせる客観的な基準を検討した。

そこで分野基礎用語を抽出する対象として、教科書、事典、論文の 3 種類を用意した。用語は、形態素解析を行い品詞が名詞あるいは名詞の連続であるものを抽出した。この 3 種類とも専門家が執筆したものであるため、これらのリソースから抽出した用語は複数の専門家の判断と同等であると考えられる。詳細は以下の通りである。

(1) 教科書 :「自然言語処理」分野の日本語の教科書 39 冊の目次に出現する用語（異なり語数 694 語）
(2) 事典 :「言語処理学事典」の目次に出現する用語（異なり語数 463 語）
(3) 論文 : 情報処理学会自然言語処理研究会で発表された論文のタイトル、抄録、キーワードに含まれる用語（異なり語数 13493 語）

教科書と事典の目次に出現する用語に着目した理由として、目次は初心者にもわかりやすい表題および学んでほしい用語を必ず著者が選定する、つまり著者が考える分野基礎用語は目次に含まれると考えたためである。この 3 種類のリソースに共通して出現する用語は 90 語であり、この 90 語を分野基礎用語と選定した。

４．論文の論理構造における分野基礎用語

論文の論理構造において、分野基礎用語がどのような出現パターンを示すのかを調べた。本論文における論理構造とは、「抄録」、「はじめに」、「関連研究」といった論文を構成している章に関連している意味のあるまとまりのことを指している。分析対象の論文コーパスは、分野基礎用語の選定時に利用した論文とは異なり、情報処理学会の論文誌に掲載された自然言語処理分野の論文の中から抄録で「実験」、「評価」、「精度」、精度の数値「%」などを含んでいる 100 論文を選んで論文コーパスとした。実験を扱った論文に絞ったのは、論理構造が比較的わかりやすく、論文の流れもある程度パターン化できるのではないかと仮定したためである。本論文では、論理構造の要素を「抄録」「はじめに」「実験」「関連研究」「おわりに」「その他」の 6 種類に分けた。「その他」は多くの場合、「関連研究」の記述の後から、「実験」記述の前までのまとまりを指している。

分析対象の論文コーパスを論理構造の要素に分割し、それぞれの要素の中における分野基礎用語の出現傾向を分析した。表 1 に出現頻度 100 以上の用語について、論理構造別の出現頻度を示す。なお、ある用語が別の用語の部分文字列となっている場合は（「文字」「文字列」など）、重複している数を差し引いて数えている。

表 1　論文の論理構造における分野基礎用語の出現頻度

	抄録	はじめに	実験	関連研究	おわりに	その他	合計
意味	54	231	360	93	49	561	1348
コーパス	64	160	448	79	59	330	810
品詞	33	116	339	28	34	361	550
辞書	30	103	310	43	36	239	522
日本語	40	136	182	45	38	225	441
生成	19	80	186	46	36	324	367
未知語	15	50	167	22	20	160	274
知識	28	101	88	16	37	105	270
言い換え	17	93	99	25	29	185	263
形態素解析	25	60	131	14	20	122	250
文字	7	26	89	24	9	65	220
シソーラス	9	39	89	34	16	39	187
アルゴリズム	16	43	73	14	17	252	163
照応	7	36	65	40	6	68	154
固有表現	5	14	108	4	7	91	138
形態素	6	27	87	2	10	124	132
文字列	8	22	82	13	4	186	129
クラスタリング	7	30	66	14	7	23	124
語義	7	35	57	7	16	45	122
機械学習	16	43	25	15	13	34	112
構文解析	7	45	35	13	9	46	109
機械翻訳	14	51	22	13	8	27	108
言語処理	16	59	9	12	10	31	106
決定木	11	34	40	11	9	74	105
言語モデル	10	19	53	12	10	70	104

最も出現頻度が高い「意味」は、一般的な文章にも使われる単語であるため、用語と見なすことが難しいが、実際に出現している文を読むと、「意味」が他の分野基礎用語と共起して出現するなど、重要な役割を果たしていることがわかった。自然言語処理において「意味」を理解することが目的でもあるため、本論文では用語と扱うことに意義があると考える。このように表 1 のリストを見ると、分野初心者でも意味がわかるような「品詞」「辞書」「文字」などの単語が並んでいる。これらは分野基礎用語の定義である、「必ず学ばなければならない語、その分野における基礎的・必須である専門用語」という基準からはずれることになる。しかし、これらの単語は、研究の背景など導入部分を記述するためには必須の語、および重要な手がかり語の役割をはたしていることがわかった。

次に、分野基礎用語が出現する文が全体のどのくらいの割合を占めているのかを調べ、表 2 に示す。分野基礎用語が一つの文に複数出現することもあるため、文単位での傾向を分析した。その結果、「抄録」、「はじめに」の論理構造の要素では、全体の半分以上を占めていることがわかった。次いで「おわりに」「関連研究」の要素で 4 割以上に分野基礎用語が含まれている。これは分野基礎用語の 90 語のうち頻度 0 を除いた 74 語が、「抄録」や「はじめに」などの論文の重要な部分を説明する文章に半分以上含まれるということになる。この結果を見ると、「抄録」や「はじめに」に多く出現する用語が分野基礎用語なのではないかと予測されるが、これまで行ってきた実験では「抄録」の中で高頻度な用語が、分野基礎用語にはなっていなかった。今回はこれまでと正解セットや分析対象コーパスが異な

っているため、単純に比較することはできない。しかし、今回の対象コーパスが論文誌に採択された実験論文であるため、論理構造がはっきりしていることや、用語の使い方や表現も推敲を重ねるなど、質の高い文章であることから、分野基礎用語の出現傾向が特徴的になったのだと考えられる。

これまで、分野基礎用語は分野特有の専門用語で、分野初心者がその分野を理解する上で必ず学ばなければならない用語と考えていた。しかし、客観的な指標による分野基礎用語の選定および実際の論文中に出現する傾向を分析すると、必ずしもその用語自体を学ぶ必要はなく、むしろその用語が手がかり語となって周辺の用語との関連により、その分野の理解を深める役割を果たしていた。つまり、分野基礎用語をベースとして、周辺用語との関連を示してあげることにより、分野初心者への論文理解を手助けすることができるのではないかと考えられる。

表２　論文の論理構造における分野基礎性用語を含む文の割合

	文数	分野基礎性用語を含む文数	割合
抄録	656	362	0.552
はじめに	2448	1284	0.525
実験	8931	2701	0.302
関連研究	1222	542	0.444
おわりに	805	394	0.489
その他	11965	3439	0.287
合計	26027	8722	0.376

５．まとめ

本論文では、その分野で必ず学ぶべき用語や手がかり語となる分野基礎用語の選定基準と、実際の論文における出現パターンの分析を行った。選定の基準は、多くの専門家が執筆した本や事典の目次、論文のタイトル、抄録、キーワードの中から共通して出現するものとした。この客観的な基準に従って抽出した分野基礎用語が論文の論理構造の要素別に出現する頻度に基づいて分析を行った。

分析の結果から、今後は分野基礎用語が出現する文が研究のどのような内容を表現しているのか（研究の背景、動機、既存研究の比較など）をさらに詳しく分析し、分野基礎用語と共起する用語との文法的関係（主語、目的語、補語、修飾語など）と意味的関係（目的、手法、対象など）を付与するなど、論文の内容理解の支援をする表現方法を検討していく。

文　献

中川裕志、森辰則、湯本紘彰(2003)「出現頻度と連接頻度に基づく専門用語抽出」、自然言語処理、Vol.10 No.1、pp.27-4

佐々木靖弘、佐藤理史、宇津呂武仁(2006)「関連用語収集問題とその解法自然言語処理」、Vol.13 No.3、pp.151-175

千田恭子、篠原靖志、奥村学(2005)「技術成果を効果的に伝える表題作成支援手法：開発と評価」、情報処理学会論文誌、Vol.46 No.11、pp.2728-2743

内山清子(2010)「専門用語の分野基礎性に関する一考察」、情報処理学会自然言語処理研究会報告、2010-NL-199(15), pp.1-6

Kiyoko Uchiyama(2011)、「A Study for Identifying Domain-Specific Introductory Terms in Research Papers」、Proceeding of the 9th Terminology and Artificial Intelligence、pp.147-150

自然言語処理学会、『言語処理学事典』(2010)、共立出版株式会社

コーパス用テキストの文字校正支援ツールの設計と実装

堤　智昭（東京農工大学）†
須永　哲矢（国立国語研究所）††

Design and Implementation of the Support Tool for the Proofreading of Corpus Text

Tomoaki Tsutsumi(Tokyo University of Agriculture and Technology)
Tetsuya Sunaga (National Institute for Japanese Language and Linguistics)

1. はじめに

国立国語研究所では「近代語コーパス」の構築が構想されており、これが実現した場合、近代の活字資料が言語研究目的で電子化されていくことになる。近代の活字資料を電子化しコンピュータにテキストとして入力する場合、近代の活字には入力仕様で規定された符号化文字集合に存在しない文字や符号化文字集合での文字と字形差がある文字が多く存在する。そのため言語研究用として保証できる質の電子テキストを得るためには、一度入力した電子テキストに対して入力仕様に定められた符号化集合に収まっているか等を確認する校正作業は必須である。具体的には、原文と照らし合わせて入力文字を確認し「=」として入力するか、存在する文字に包摂して入力するという対応が考えられる。また、電子化する時に、原文資料に対してどのような改変を行ったかといったメモ情報などを本文データに影響を与えないように付与する必要がある。このようなテキスト入力とその文字校正作業は非常に煩雑であり、人の手のみでこれらの作業を行った場合、時間がかかり校正漏れ等作業ミスが発生する可能性が高い。そこで、本研究ではコーパス用テキストの文字校正作業の作業ミスを減らし、高効率化するための作業支援ツールを設計、作成した。本稿では、明治時代の学術誌『明六雑誌』でのツール適用例を紹介する。

1.1　言語研究用電子テキストでの文字処理

紙媒体の活字資料を電子テキストに写し取ってコーパスを構築する場合、誤入力等を見つけ出す、通常の意味での「校正」ははもちろん必要であるが、電子テキスト化に当たってはさらに、「文字の表現の仕方そのものの仕様統一を図る」という、別種の校正が必要になってくる。近代の活字資料、『明六雑誌』（1874～1875）を例にとると、近代の活字では図１のような字形差が見られる。

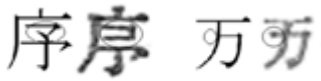


図１　『明六雑誌』に出現する「序」「万」の字形（右側）

これらの字形差に対して、入力作業者によっては「序」「万」を入力するか、外字として「=」を入力するかの揺れが生じうる。また、第一次入力段階では差異の存在そのものを見落としている可能性もある。入力対象とする原資料のどの文字に、通用字形との差異が

† t_tsu77@ninjal.ac.jp
††tsunaga@ninjal.ac.jp

あるかをあらかじめ知ることはできないため、一次入力時点で問題となる文字を経験として洗い出したうえで処理方針を確定し、次の校正段階で確認、統一化を図るということになる。

1.2　言語研究用という目的に応じた文字処理方針と、そのための作業

「近代語コーパス」における文字処理方針の概要を紹介する。

1.2.1　拡張包摂

図１のような差異に関して、「序」「万」を入力すべきか、「**=**」とすべきかは、その電子テキストの使用目的による。言語研究用のテキストとしては、読めること、語が語として取り出せることが望ましいため、「**=**」はなるべく少なく、読める文字として「序」「万」として表示されるテキストの方が、有用性が高い。JIS 漢字では、「漢字の字体の包摂規準」を定めており、包摂規準の範囲内の差異であれば、同一の符号位置の文字として処理することができる（図２）。

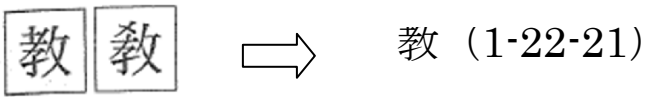


図２　JIS 包摂規準の例

図１のような差異を包摂してよいことを示す包摂規準は設定されてはいないが、既存の包摂規準と照らして、包摂規準の拡大解釈ないし拡張で、同一字とみなしてよいものと判断し、「序」「万」を入力する。

1.2.2　別字代用

図３は『明六雑誌』に出現する、「すう」と読む字である。これはコーパスで使用する文字集合になく、電子的に表現できない。「すう」に当たる字としては「吸」があるが、図３とは字形差が大きすぎ、同一字とみなせるような差異ではないため、「包摂」という処理は当たらない。このような場合にも、研究資料としての有用性を考え、「**=**」にはせず、同訓の通用字（ここでは「吸」）で代用することとする。

図３　『明六雑誌』出現漢字（「すう」）

1.2.3　文字処理情報の付与

上記「拡張包摂」「別字代用」といった処理は、コーパス構築作業上、使用目的から要請された臨時的な処理であり、文字処理一般に通用している処理方針とは言えない。そのため、そのような処理をした文字に関しては、ただ文字を入力するだけでなく、タグの形で処理内容の情報を残しておくことが望ましい。

【例】

図１→＜包摂＞序＜/包摂＞

図３→＜外字　代用＝”1” unicode=”564F”＞吸＜/代用＞

=　→＜外字　代用＝”0”＞**=**＜/外字＞

1.3　校正支援ツールの必要性

コーパス化にあたってどの程度の差異までを同一字とみなす（包摂する）か、また、どの字をどの字で代用するかといった指針は、原資料全体を見渡してからでないと確定させることができない。そのため、一次入力作業中に、処理上問題となる文字を洗い出し、次の工程である校正時に統一を図るという順序になる。

校正において、「包摂」「代用」の処理を統一的にしつつ、見落としの可能性まで含め確認作業を行うというのは非常に煩雑であり、作業上のミスも生じやすい。ここでの文字処理作業で、特に難しいのは以下の２点である。

（１）該当文字だけでなく、多岐にわたる情報をタグの形で付与したい。

「包摂」「代用」の処理を経て入力された文字に関しては、逐一その旨をタグとして記入しなければならない。また、研究資料としての有用性を考え、原資料はどのような字であったかの情報も残すためには、unicode で表現可能な文字に関しては unicode 番号を記入しておくなど、タグ内に様々な形で注記をせねばならない。

（２）一括変換はできず、原資料との目視確認が必要である。

原資料に使用されている活字が全て均質であるとは限らない。そのため、注意すべき字が確定し、その字に対する処理が決まったとしても、一括変換はできない。

図４　『明六雑誌』における「敵」活字字形

『明六雑誌』には、図４（A）のように、通用字「敵」とは異なり、右側が「欠」となっている活字が出現する（U+6B52 で表現可能。ただし「通時コーパス」では使用しない文字コード)。しかし、全ての「敵」が（A）の活字で表現されているわけではなく、より通用字に近い（B）の活字も出現する。このため、一次入力されたテキストの「敵」のすべてを、例えば「＜代用　Unicode=”6B52”＞敵＜/代用＞」などと一括で変換するわけにはいかず、言語資料としての質を鑑みるならば、一文字ずつ原資料と照合を行いながら確認していかなければならない。

（１）（２）の作業は極めて煩雑であり、手作業では多大な時間を要する上に、ミスも生じやすい。そこで、要確認文字の原資料照合・目視確認を援助し、「包摂」「代用」などの処理方針に従ってのテキスト上の文字置き換え、およびタグ情報付与を行いやすくする支援ツールを設計・実装することで、作業の効率化を図った。

近代の活字資料の電子化作業工程と、本ツールの位置づけを図５に示す。

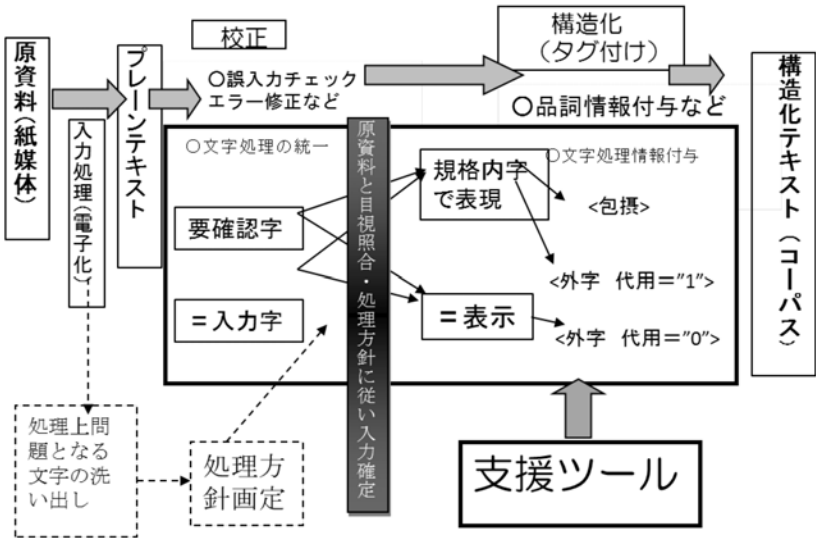


図５　作業工程とツールの位置づけ

2. システム概要

本ツールは、UTF-8 で記述された XML ファイルを読み込み、対象文字を抽出し効率的な校正作業を支援するツールである。今回は、コーパス用テキストの文字校正作業を行う作業者を対象ユーザとし、操作しやすい GUI をもつツールを設計した。本ツールのもつ主な校正用機能は以下の３つである。

（1）対象文字の抽出・確認

xml 形式で記述された文章データから、確認対象となる文字を検索しリスト形式で表示する。また、テキストデータの該当箇所、及び原文 PDF の該当箇所を自動で表示し、目視照合支援も行う。

（2）対象文字の置き換え

抽出した対象文字に対して置き換えを行う。書き換える文字の入力方法は、IME を用いた日本語入力に加え、unicode 番号から入力することも可能とした。同じ文字が複数個抽出され、同一の書き換えを何度も行う場合は、記憶した内容で自動的に置き換えることも可能とした。

（3）メモ機能

「記号」「合字」といった文字種類や unicode 番号等のメモを、xml タグの属性として記述する。

これら 3 つの校正用機能に加え、本ツールの評価実験、管理を効率化するためにネットワークを利用した情報収集、保守機能を設計、実装した。

3. 設計・実装

3.1 GUI 設計

本ツールのメイン画面例を図 6 に示す。メイン画面は次の 3 つから校正される。

① 設定やファイルの読み書きを行うメニューバー
② 校正対象の文字をリスト表示する画面（以下、リスト画面）
③ 読み込んだファイルをテキスト形式で表示する画面（以下、テキスト画面）

校正作業は、主に②のリストを操作して行う。リストは操作が行われると、行の色が白から茶色に変更される。これにより、編集済みか未編集かが判断しやすくなる。また、最終更新時刻も記録する。③のテキスト画面では、対象の XML タグを赤文字で強調し、それ以外を黒文字で表示している。②のリストと③のテキスト画面は連携しており、リストからある行を選択すると、③のテキスト画面でも選択した行のタグがある箇所に自動でスクロールする。

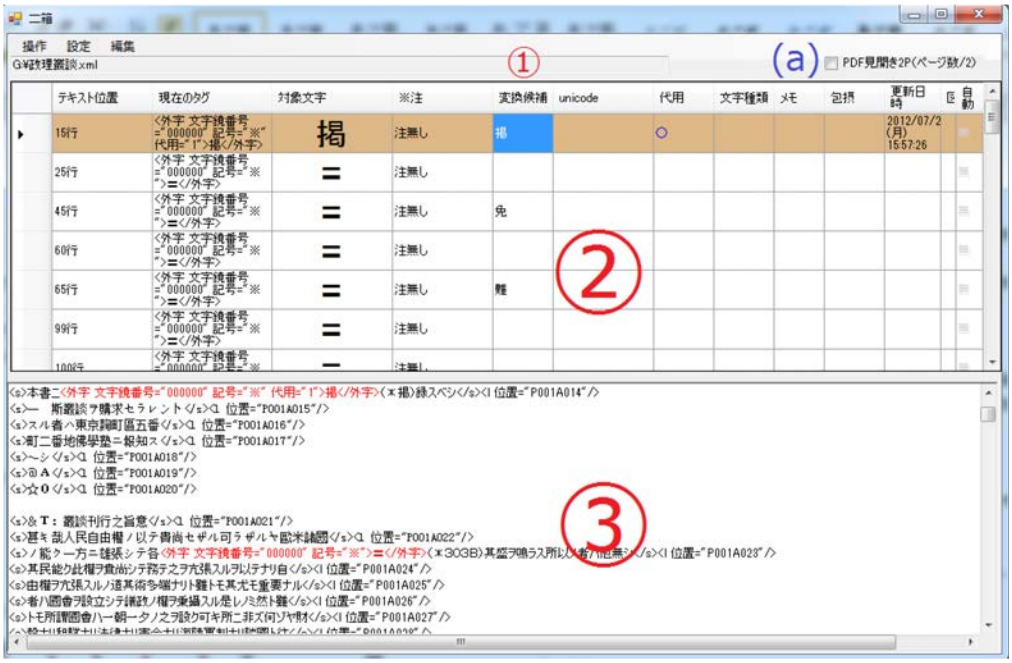


図 6 ：メイン画面

3.2 目視確認機能

本ツールでは、3.1 で示したように校正対象の XML データをテキスト画面に表示している。それに加えて、原文 PDF を表示することで目視確認を補助する。原文 PDF は、リスト画面の『現在のタグ』セルをダブルクリックして表示する。XML データには予め、行ごとに原文の何ページに存在するかという情報が記述してあるものとし、この機能はその情報を読み取り実行する。PDF を表示する時には、対象の文字があるページの番号を XML データから取得し、表示する。原文 PDF の作成方式によっては、PDF のページ１ページにつき原文の見開き２ページで作成されている場合も考えられる。その場合、原文のページと原文 PDF のページにずれが生じる問題が考えられる。そこで、図６の(a)で示したチェックボックスを作成し、チェック時には見開き２ページとしてページ計算を行うことで対応した。自動で原文 PDF を開くためには、PDF データのあるフォルダの場所を、設定→詳細設定から設定する必要がある。

3.3 対象文字のリスト化

本ツールでは、＜外字＞タグ、及び＜確認＞タグのついた文字を検索し昇順リスト形式で表示する。例としては「<外字 記号=″※″>=</外字>」のようなものがあげられる。この場合=が対象文字となる。リストの列は表１のように設計した。

表１：リスト設計

行名	内容
テキスト位置	対象テキストが、読み込んだ XML ファイルの何行目にあるかを表示する
現在のタグ	対象の XML タグを表示する
対象文字	対象の文字を表示する
※注	タグの後に注がついている場合、その内容を表示する。
変換候補	タグの後に（＊文字）という記述があった場合、＊以下の文字を変換候補文字として表示する
Unicode	Unicode を入力することができる。ここに入力された値は XML の属性値に記述される
代用	校正を行った文字が代用されたものであるか否かを表示する。値は「○」「×」の２種類である。ここに入力された値は XML の属性値に記述される。包摂行の値とは排他である。
文字種類	校正を行った文字の種類を表示する。文字種類は「記号」「合字」「漢字」「カナ」「絵文字」の 5 種類とした。ここに入力された値は XML の属性値に記述される
メモ	残しておきたいメモを表示する。ここに入力された値は XML の属性値に記述される
包摂	校正を行った文字が、漢字包摂基準をもとに包摂されたか否かを表示する。値は「○」「×」の２種類である。ここに入力された値は XL の属性値に記述される。代用行の値とは排他である。
更新日時	その行が最後に操作された時刻を表示する。形式は「西暦/月/日（曜日）時：分：秒」である。
図	対象文字のフォントを表示する。表示には、フォントデータを個別に用意する必要がある。
自動	自動置き換え機能の対象であるか否かを表示する。自動置き換え機能については後述する。

3.4 校正機能
3.4.1 対象文字の置き換え
（1）直接入力
リストの中から、対象文字セルを指定してキーボードから直接文字を入力することができる。
（2）unicode 入力
『対象文字』セルをダブルクリックすると、図 7 に示すような画面が表示され unicode を指定して文字を入力することができる。たとえば、「U+3042」と入力すると対象文字列に「あ」が表示される。ここで入力した unicode は『unicode』セルにも自動で入力され「現在のタグ」セルに反映される。入力できるコードの範囲は、外部ファイルで指定可能である。外部ファイルでは、一行に一文字ずつ「人　¥t　U+4EBA」のように「文字　¥t　unicode」（¥t はタブを表す）という形式で利用可能文字を指定する。このファイルに記述されていない文字コードが入力された場合は、図 8 のように入力不可能であることを表示し、『対象文字』セル、『unicode』セル、『現在のタグ』セルへの入力は行われない。

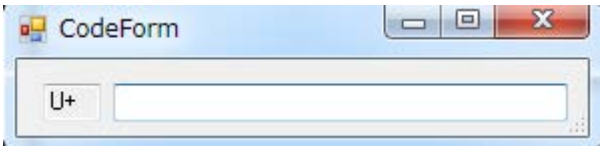


図 7 ：unicode 入力画面

図 8 ：unicode 入力範囲警告

（3）変換候補入力
　変換候補文字がある場合、『変換候補』セルをダブルクリックすると変換候補文字が『対象文字』セルに自動入力される。

3.4.2　付属情報入力
（1）unicode
『unicode』セルに値を入力すると、現在のタグの属性値に「unicode=”unicode セルの値”」の形式で入力される。
（2）代用
代用セルを左クリックすると、「○」「×」が自動入力される。1 クリックごとに○×は入れ替わる。代用セルの値は、現在のタグの属性値に「代用=”0or1”」の形式で入力される。○ならば 1、×ならば 0 が入力される。
（3）文字種類
『文字種類』セルも『代用』セルと同様に左クリックすると値が自動入力される。1 クリックごとに値は入れ替わる。値は「記号」「合字」「漢字」「カナ」「絵文字」の 5 種類

である。文字種類セルの値は、現在のタグの属性値に「moji=”値”」の形式で入力される。
（４）メモ
『メモ』セルを指定して文字をキーボードから直接入力することができる。このセルには任意の値を入力することができる。『メモ』セルの値は、現在のタグの属性値に「memo=”値”」の形式で入力される。
（５）包摂
『包摂』セルも、『代用』セルと同様に左クリックすると「○」「×」が自動入力され、１クリックごとに○×が入れ替わる。

3.4.3 校正補助機能
（１）戻る、進む機能
校正作業中に変更した内容を元に戻したい時のために、校正によるデータの変更を全て記録することで操作を一つ戻る、または戻した操作を一つ進める機能を実装した。データの変更は、変更されたセル、変更前の値、変更後の値、変更した時刻の４つのデータを１セットとし、リスト形式で保存する。
（２）自動入力機能
一つのファイル内に、同一の校正対象となる文字が複数存在する場合が考えられる。そのような場合、一度入力した内容を何度も入力することになる。そこで、作業効率化のため対象の自動入力機能を実装した。自動入力を行う文字は外部ファイルに保存可能とし、「餐¥n 〈確認 代用=″1″ memo=″沃＋食″〉」のように「対象文字 ¥n 置き換え後のタグ」（¥nは改行を表す。¥r¥n にも対応する）という形式で記述する。リストの校正対象全てに対して自動入力を行う方法と、選択した一つの行に対してのみ自動入力を行う方法の２種類を実装した。自動入力は、外部ファイルのデータと対象文字とを比較し、同一の文字だった場合現在のタグセルと対象文字セルの値の置き換えを行う。ただし、同一の文字でも前後の文脈や目視確認の結果によって校正対応が変わる場合も考えられるため、置き換え実行前に置き換えるかどうかの確認をする画面を表示する。また、自動置き換えの対象となるか否かはリスト表示の「自動」列に「無」「有」で表示される。

3.4.4 データ保存
本ツールでは、３種類の保存方式を実装した。
（１）中途保存
作業を一時中断したい場合のため、編集中の XML ファイルを書き換えることなく作業内容のみを保存する。それにより、中断前とまったく同じ状態で作業を再開することができる。作業内容保存データは txt 形式で保存される。
（２）確定保存
校正作業が完了し、保存する場合のための保存方式である。本ツールを用いて校正した内容を XML ファイルに適用し保存する。
（３）タグ消去保存
全ての作業が終わり、本ツールで利用している校正用タグを全て消去したい場合のための保存方式である。編集している XML ファイルから、本ツールの校正用タグである＜外字＞タグ、及び＜確認＞タグを全て消去し、保存する。

3.5 ネットワーク機能
本ツールは情報収集及びメンテナンスを行うために、ネットワークを介してサーバと通信を行う。サーバには FTP サーバを利用した。
（１）作業情報収集機能
本ツールの有効性を示すため、行った校正作業内容を記録する必要がある。校正作業は、複数人で行う為記録した校正内容の収集を効率的に行いたい。そこで、本ツールでは記録

した校正内容をネットワーク経由で自動的にサーバに送信する。リスト表示する画面へ行った作業を保存し、保存データは「リストの行，列，変更前のセルの値，変更後のセルの値，時刻」といった「,」区切りの CSV 形式で記述される。保存された CSV ファイルは、ツールごとに割り振られた ID、作業を行なっている XML ファイル名、及び時刻で管理を行う。データの送信には FTP を用いた。また、ネットワークに接続できない環境下での作業も想定し、送信データと同様のデータをローカルフォルダに保存する機能も実装した。

（２）自動更新機能

（１）でも示した通り、本ツールを用いた校正作業は複数人で作業を行う。ツールの更新時にはタグの仕様変更など、全作業者で共有するべき更新が行われる場合も考えられるため、全員で同じツールを使用することが必要となる。しかし、作業者全員が常にツールの更新を確認することは非常に手間がかかる。そこで、本ツールではツール起動時に、ネットワーク経由でツールの自動更新を行う機能を実装した。データの送受信には FTP を用いた。ツール更新時には、更新内容等の連絡事項を表示することで、作業者全員での情報共有を可能とした。

4. まとめ

今回、コーパス用テキストの校正作業を効率化するための作業支援ツールを設計、製作した。本ツールでは、対象文字の抽出・確認、対象文字の置き換え、メモといった作業を効率化するための機能を、簡単な操作で行える GUI と共に実装した。さらに、校正データの保存方式を３種類用意し、作業の状態に応じて柔軟に対応できるようにした。また、ネットワークを介したツールの更新や、有効性を示すための作業情報収集の効率化を図った。今後は、現在行なっている校正作業の作業情報を収集、解析し本ツールの有効性の確認を行う予定である。

文　献

田島孝治、高田智和（2010）「JIS　X　0213 文字セット運用のための文字処理支援ツール」特定領域研究「日本語コーパス」、pp.77-84 平成 21 年度公開ワークショップ予稿集

須永哲矢、堤智昭、高田智和(2011)「明治前期雑誌の異体漢字と文字コード−『明六雑誌』を事例として−」、pp.381-388、じんもんこん 2011 論文集

須永哲矢、堤智昭、高田智和(2012)「明治前期の漢字活字とＪＩＳ漢字包摂規準－『明六雑誌』活字字形への、包摂規準適用実験－」第 95 回人文科学とコンピュータ研究発表会

『現代日本語書き言葉均衡コーパス』を用いた文末表現のバリエーションの分析 (2)

丸山 岳彦　（国立国語研究所 言語資源研究系）†

Analyzing Variation of Sentence Final Expressions in the BCCWJ (2)

Takehiko Maruyama　(Dept. Corpus Studies, NINJAL)

1　はじめに

本稿では、『現代日本語書き言葉均衡コーパス（以下、BCCWJ と記す）』を利用した現代日本語文法研究および社会言語学的な研究の試みとして、レジスターごとに特徴的に観察される文末表現のバリエーションについて分析を行なう。ここでの分析の前提となる丸山 (2012) では、BCCWJ に含まれる 12 種類のレジスターのテキストを分析し、各レジスターに見られる典型的な文末表現を抽出した。その続編となる本稿では、これらの文末表現が実際にどのような形で用いられているのかについて、単語（短単位）N-gram を用いた集計結果をもとに分析を行なう。

2　BCCWJ の各レジスターに頻出する文末表現

本稿での議論の前提として、丸山 (2012) で示した分析結果を示す。丸山 (2012) では、BCCWJ に含まれる 12 種類のレジスターから約 20,000 文ずつを取り出し、文末位置から逆向きに文字単位の N-gram を抽出して、そこに頻出する文末表現を分析した。その上で、レジスターごとに異なる文末表現が特徴的に観察されることを指摘し、テキストが持つ機能や定形性という特徴が、頻出する文末表現に影響を与えていることを指摘した。表 1 に、5gram 以上に現れた文末表現を集計し、レジスターごとに出現率の高い文末表現をリスト化した結果を示す。なお、文末の「。」は省略してある。ゴシック体は、表 1 中、そのレジスターにしか現れていない文末表現を表す。

表 1: BCCWJ の各レジスターに頻出する文末表現（5gram 以上、上位 10 位）

出版書籍		図書館書籍		雑誌		新聞		白書		教科書	
のである	1.85%	のである	2.11%	ています	1.27%	している	2.32%	っている	9.91%	ましょう	4.08%
ています	1.73%	なかった	1.62%	している	1.00%	っている	1.43%	している	9.08%	**てみよう**	2.54%
っている	1.41%	っている	1.48%	っている	0.94%	れている	1.04%	**なっている**	6.28%	ています	2.09%
している	1.40%	している	1.18%	れている	0.90%	なかった	0.92%	れている	5.41%	れている	2.04%
なかった	1.34%	であった	1.18%	なかった	0.65%	**になった**	0.74%	**となっている**	5.09%	っている	1.57%
れている	1.21%	ています	1.17%	のである	0.65%	していた	0.67%	されている	2.73%	している	1.50%
あります	1.00%	っていた	0.97%	あります	0.61%	**たという**	0.58%	となった	2.29%	**みましょう**	1.37%
であった	0.96%	れている	0.95%	されている	0.46%	ています	0.54%	のである	2.03%	**てみましょう**	1.27%
りません	0.83%	たのである	0.78%	**でしょう**	0.46%	となった	0.49%	**めている**	1.99%	あります	1.25%
なります	0.73%	あります	0.74%	ください	0.45%	**るという**	0.47%	**ものである**	1.90%	りました	0.98%

広報紙		Yahoo!知恵袋		Yahoo!ブログ		法律		国会会議録	
ください	9.07%	**しょうか？**	5.46%	ています	1.72%	**ならない**	21.55%	**ございます**	14.08%
ています	8.46%	**でしょうか？**	5.44%	**いました**	1.48%	**なければならない**	18.79%	**でございます**	12.35%
しています	3.59%	**のでしょうか？**	2.79%	りました	1.21%	**ができる**	13.79%	**おります**	10.98%
てください	2.76%	思います	2.77%	**きました**	1.15%	**ことができる**	13.78%	**ております**	10.84%
あります	2.66%	ています	2.77%	しました	1.11%	**ることができる**	11.95%	思います	8.47%
しました	2.49%	ください	2.75%	思います	0.82%	**しなければならない**	10.62%	あります	7.76%
れました	2.30%	と思います	2.58%	と思います	0.73%	**ものとする**	8.12%	と思います	7.16%
なります	2.17%	てください	2.44%	あります	0.64%	**することができる**	6.88%	**であります**	6.87%
してください	2.05%	りません	2.17%	**ていました**	0.61%	**るものとする**	6.36%	**いと思います**	4.18%
ましょう	1.94%	あります	1.60%	**っています**	0.57%	**準用する**	4.60%	**けでございます**	4.16%

† maruyama@ninjal.ac.jp

3　本稿の目的、分析の対象および方法

丸山 (2012) での分析結果を受け、本稿では、各レジスターで特徴的に見られる文末表現について、単語（短単位）単位による N-gram を抽出し、各文末表現が実際にどのような形で用いられているかを分析する。分析対象とする文末表現は、表 1 の結果を参考にして、図 1 に示す 13 種類を選び、4 つのグループに分類した。

デス形：です。 / でした。 / でしょう。
マス形：ます。 / ました。 / ましょう。 / ません。
質問・依頼形：か。 / か？ / ください。
その他：ている。 / できる。 / ならない。

図 1: 分析対象とする文末表現（13 種類）

BCCWJ 検索サイト「中納言[1]」を用いて、対象とする文末表現を全てのレジスタを対象に検索し、結果をダウンロードして、13 個の KWIC データを得た。検索式の例を図 2 に示す。

```
キー: （書字形出現形 = "です" AND 品詞 LIKE "助動詞%"） AND 後方共起: （書字形出現形 = "。"
AND 品詞 LIKE "補助記号-句点%"） ON 1 WORDS FROM キー IN (subcorpusName="出版・書籍" AND
core="true") OR (subcorpusName="出 版・書 籍" AND core="false") WITH OPTIONS unit="1"
AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"
```

図 2: 中納言の検索式の例

次に、各 KWIC データを 11 種類のレジスターごとに分割し、合計 143 個の分析用データを得た。ここで分析する 11 種類のレジスターを、図 3 に挙げる。BCCWJ に格納されている全データのうち、「ベストセラー（OB）」「韻文（OV）」は、分析上の都合から、除外した。

出版 SC：書籍（PB）、雑誌（PM）、新聞（PN）
図書館 SC：書籍（LB）
特定目的 SC：Yahoo!知恵袋（OC）、Yahoo!ブログ（OY）、白書（OW）、
国会会議録（OM）、広報紙（OP）、教科書（OT）、法律（OL）

図 3: 分析対象とするレジスター

さらに、143 個の分析用データから、文末表現から逆向きに単語（短単位）N-gram を抽出した。N-gram の抽出には、田野村忠温氏（大阪大学）が開発・公開している「BNAnalyzer [2]」を利用した。

「BNAnalyzer」は、「中納言」の検索結果（KWIC データ）を入力として、検索キーの前後文脈の N-gram の一覧（1gram〜8gram）を Excel ファイルとして出力するツールである（田野村, 2012）。処理結果として出力される Excel ファイルの例を、図 4 に示す。

[1] https://chunagon.ninjal.ac.jp/
[2] http://www.tanomura.com/research/BNAnalyzer/

	A	B	C	D	E	F
1	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
2	て(2162)	して(676)	てみて(221)	してみて(79)	ないようにして(28)	ないように注意して(11)
3	で(195)	みて(243)	にして(141)	ようにして(77)	ので、注意して(16)	ので、注意して(7)
4	ご覧(26)	ないで(121)	注意して(85)	を参照して(62)	を参考にして(14)	／ふたつ／３つ(6)
5	_(23)	おいて(63)	参照して(63)	考えてみて(29)	ように注意して(12)	ますので注意して(6)
6	を(16)	考えて(59)	ておいて(60)	に注意して(26)	想像してみて(10)	を想像してみて(6)
7	許し(13)	あげて(58)	てあげて(52)	参考にして(25)	にしてみて(9)	足して４で割って(6)
8	待ち(12)	せて(55)	しないで(41)	で注意して(16)	の見積書をご覧(9)	。_「ちょっと待って(5)
9	注意(11)	見て(54)	をして(35)	覚えておいて(16)	を考えてみて(9)	いうことを忘れないで(5)
10	安心(10)	教えて(47)	と考えて(32)	気をつけて(15)	_「ちょっと待って(8)	五十九を参照して(5)
11	ごらん(9)	やって(30)	させて(30)	試してみて(14)	ことを忘れないで(8)	参考にしてみて(5)
12	せ(9)	いて(29)	を見て(28)	をクリックして(13)	するようにして(8)	くらいお灸をして(4)
13	参照(7)	行って(27)	を教えて(25)	を選択して(13)	ボタンをクリックして(8)	ことを知っておいて(4)
14	聞き(7)	待って(27)	ていて(24)	してあげて(11)	で、注意して(7)	医師の診察を受けて(4)
15	つ(6)	言って(24)	になって(23)	を忘れないで(11)	ページを参照して(7)	楽しみにしていて(4)
16	入り(6)	なって(23)	確認して(22)	見積書をご覧(10)	、気をつけて(6)	気を悪くしないで(4)
17	休み(5)	つけて(22)	ちょっと待って(20)	「ちょっと待って(9)	」を参照して(6)	細心の注意を払って(4)

図 4: BNAnalyzer の処理結果の例（LB 「ください。」に前接する要素）

143 個の分析用データを BNAnalyzer で処理し、Excel ファイルを出力させた。そのうち、前文脈の N-gram を示すシート「前文脈の N-gram」のみを取り出し、1gram から 8gram までの各結果をそれぞれ抽出した。N-gram の表現と出現頻度をタブ区切りで配列し直し、全体で合計 282,171 行の頻度付き N-gram の一覧を得た。集計結果の例を図 5 に示す。以下では、この N-gram データを用いて分析を行なう。

	media	EOS	gram	string	freq
1	LB	kudasai	2gram	して	676
2	LB	kudasai	2gram	みて	243
3	LB	kudasai	3gram	てみて	221
4	LB	kudasai	3gram	にして	141
5	LB	kudasai	2gram	ないで	121
6	LB	kudasai	4gram	してみて	79
7	LB	kudasai	4gram	ようにして	77
8	LB	kudasai	3gram	注意して	65
9	LB	kudasai	3gram	参照して	63
10	LB	kudasai	2gram	おいて	63
11	LB	kudasai	4gram	を参照して	62
12	LB	kudasai	3gram	ておいて	60

図 5: N-gram の集計結果の例

4　分析 1：総文数に占める各文末表現の比率

分析の 1 点目として、分析対象とする 13 種類の文末表現が、各レジスターにおける総文数に占める割合について見る。BCCWJ に格納されている「文」の数をどう捉えるかについては複数の解釈の可能性があり得るが、ここでは、BCCWJ-DVD 版に格納されている「文書構造タグ」において、属性なしの <sentence> タグで囲まれている範囲（すなわち、句点類で終わる「文」に相当する範囲）を収集し、その数を総文数とした。各レジスターごとの総文数を表 2 に、13 種類の文末表現が各レジスターの総文数に占める割合を図 6 に、それぞれ示す。

表 2: 各レジスターの総文数

SC	レジスター	総文数
出版 SC	書籍 (PB)	1,087,715
	雑誌 (PM)	197,069
	新聞 (PN)	54,932
図書館 SC	書籍 (LB)	1,276,651
特定目的 SC	白書 (OW)	95,267
	教科書 (OT)	39,966
	広報紙 (OP)	97,454
	Yahoo!知恵袋 (OC)	582,862
	Yahoo!ブログ (OY)	487,167
	法律 (OL)	17,637
	国会会議録 (OM)	116,022
合計		4,223,682

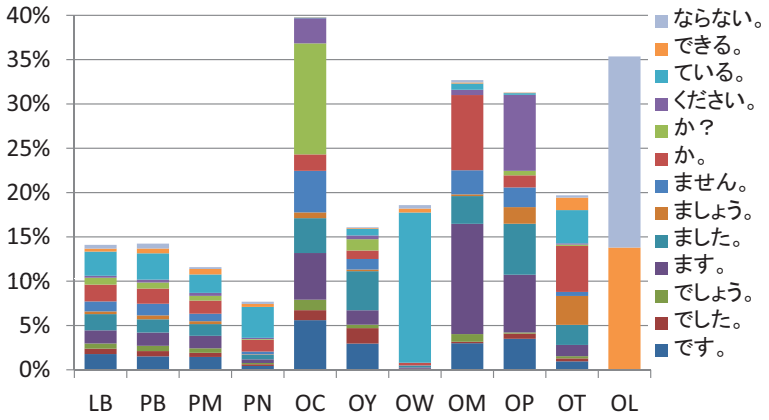


図 6: 各レジスターの総文数に占める各文末表現の比率

以下では、図 6 から読み取れるいくつかの特徴について論じる。

13 種類の文末表現が総文数に占める割合　13 種類の文末表現の合計が総文数に占める割合を各レジスターごとに見てみると、Yahoo!知恵袋 (OC)、法律 (OL)、国会会議録 (OM)、広報紙 (OP) が 30%を超えているのに対して、新聞 (PN)、雑誌 (PM)、書籍 (LB、PB) は 15%を下回っていることが分かる。丸山 (2012) で示した結果と同様、前者のレジスターに含まれるテキストが比較的少ない種類の文末表現によって構成されているのに対して、後者のレジスターに含まれるテキストには多様な種類の文末表現が現れていると見ることができる。

図書館 SC、出版 SC における文末表現　図書館 SC の書籍 (LB)、出版 SC の書籍 (PB)、雑誌 (PM)、新聞 (PN) の結果を見ると、どのレジスターもほぼ同じような分布を示していることが分かる。大半が普通体で書かれる新聞のみ、デス形・マス形の比率が低く、「ている。」の比率が高い、という違いが観察されるものの、出版される書き言葉の文末表現は、ほぼこのような分布を示すのであろう。

特的目的 SC における文末表現　一方、特定目的 SC では、レジスターごとに分布が大きく異なっている様子が見て取れる。特徴的な個所を挙げてみると、以下のようになるだろう。

- Yahoo!知恵袋 (OC) では、「か？」の比率が顕著に高い。これは、質問と回答の組で構成される Yahoo!知恵袋において、疑問符「？」を付けて質問をする表現が多用されるためと考えられる。
- 白書 (OW) では、「ている。」の比率が顕著に高い。これは、国内外の現況について客観的に記述・報告を行なう文体の中で、「ている。」が好まれていることに起因すると考えられる。
- 国会会議録 (OM) では、「ます。」「か。」の比率が顕著に高い。前者は、会議中「ございます。」という文末が多用されるため、後者は、「その理由はなぜですか。」「可能な条件があるのかどうか。」など質問をしたり疑義を呈したりする場合に「か。」が多用されるためと考えられる。
- 広報紙 (OP) では、「ください。」の比率が顕著に高い。「郵便往復はがきでお申し込みください。」「納税通知書をご覧ください。」のように、市民に対する広報記事の中で「ください。」が多用されているためと考えられる。
- 教科書 (OT) では、「ましょう。」「か。」の比率が顕著に高い。それぞれ、「あてはまる数を書きましょう。」「それぞれ何個含まれているか。」のような問題文の中で、多用されていると考えられる。

- 法律 (OL) では、「ならない。」「できる。」のみで、総文数の 35%を占める。丸山 (2012) でも述べたように、法律はある行為を命令・禁止したり、ある権利を保障したりすることを明示的かつ曖昧性のないように述べるためのテキストであり、そのことを示すための文末表現が特徴的に表れているものと考えることができる。

5　分析 2：各文末表現の N-gram と出現率

分析の 2 点目として、13 種類の文末表現が実際にどのような形で用いられているのかについて、N-gram の集計結果を用いて分析する。ここでは、図 5 に示したような N-gram の一覧の中から、3 グラム（3 短単位）以上の連接を分析の対象とする。各 N-gram の出現数をそのメディアの総文数で割って出現率を求め、出現率で降順ソートすることにより、出現しやすい文末表現の用いられ方を集計した。以下、4 グループ・13 種類の文末表現ごとに、出現率で上位 5 位までの結果を示す。なお、例えば「(OM 3)」は、「国会会議録で現れた 3gram の例」であることを表わす。

デス形

表 3: 「です。」の前接要素

と思うの (OM 3)	293 (0.25%)
と思うん (OM 3)	258 (0.22%)
ているわけ (OM 3)	220 (0.19%)
次のとおり (OP 3)	63 (0.06%)
は次のとおり (OP 4)	49 (0.05%)

表 4: 「でした。」の前接要素

ありません (OC 3)	440 (0.08%)
ありません (OY 3)	245 (0.05%)
ありません (LB 3)	519 (0.04%)
ありません (PB 3)	435 (0.04%)
いません (OC 3)	209 (0.04%)

表 5: 「でしょう。」の前接要素

ているわけ (OM 3)	43 (0.04%)
ているの (OT 3)	10 (0.03%)
が、いかが (OM 3)	27 (0.02%)
ているの (OC 3)	124 (0.02%)
ですが、いかが (OM 4)	20 (0.02%)

「です。」の結果の上位 5 位は国会会議録 (OM) と広報紙 (OP) が占めており、特に国会会議録の「思う (の | ん) です。」という形の比率が高い。一方「でした。」の結果を見ると、Yahoo!知恵袋 (OC)、Yahoo!ブログ (OY)、図書館書籍（LB)、出版書籍（PB) という 4 つのレジスターで「ありませんでした。」という形が上位 4 位を占めている。つまり、「でした。」という文末表現は、幅広いレジスターにおいて「ありませんでした。」という形で多用されていると言える。また、「でしょう。」の結果を見ると、国会会議録に現れた「ているわけでしょう。」「いかがでしょう。」、教科書 (OP)・Yahoo!知恵袋に現れた「ているのでしょう。」が上位を占めている。このうち「ているわけでしょう。」以外は、基本的に問いかけとして用いられる文末表現であると言える。

マス形

表 6: 「ます。」の前接要素

わけでござい (OM 3)	1268 (1.09%)
たいと思い (OM 3)	1132 (0.98%)
してい (OP 3)	645 (0.66%)
考えており (OM 3)	632 (0.54%)
お願いし (OC 3)	1428 (0.24%)

表 7: 「ました。」の前接要素

開催され (OP 3)	172 (0.18%)
が行われ (OP 3)	129 (0.13%)
ことにし (OT 3)	46 (0.12%)
してい (OP 3)	94 (0.10%)
と言われ (OC 3)	553 (0.09%)

「ます。」の結果では、国会会議録が多く現れている点が目立つ。「わけでございます。」「たいと思います。」「考えております。」などは、デスマス体で話される国会答弁の中で、多用される文末表現

表 8: 「ましょう。」の前接要素

してみ (OT 3)	73 (0.18%)
考えてみ (OT 3)	69 (0.17%)
調べてみ (OT 3)	60 (0.15%)
ようにし (OP 3)	120 (0.12%)
ないようにし (OP 4)	67 (0.07%)

表 9: 「ません。」の前接要素

かもしれ (OC 3)	1594 (0.27%)
ではござい (OM 3)	297 (0.26%)
ではあり (OC 3)	1321 (0.23%)
ではあり (PB 3)	1717 (0.16%)
なければなり (OP 3)	151 (0.15%)

であると言える。一方、「ました。」の結果では、広報紙が多く現れている。「開催されました。」「が行われました。」のように、区市町村で開催されたイベントを報告するような記事が多く含まれていることを示唆する。また、「ましょう。」の結果では、教科書 (OT)、広報紙が上位 5 位を占めている。教科書では「してみましょう。」「考えてみましょう。」「調べてみましょう。」という形で、読み手である児童・生徒に対する指示が表わされている。一方、広報紙では「(ない) ようにしましょう。」という形で、読み手である区市町村民への呼びかけが行なわれている。

依頼・質問形

表 10: 「か。」の前接要素

みません (OP 3)	463 (0.48%)
てみません (OP 4)	421 (0.43%)
ではない (OM 3)	454 (0.39%)
じゃないです (OM 3)	403 (0.35%)
ありません (OM 3)	379 (0.33%)

表 11: 「か？」の前接要素

なのでしょう (OC 3)	2055 (0.35%)
ないのでしょう (OC 3)	1290 (0.22%)
なんです (OC 3)	1253 (0.21%)
みません (OP 3)	91 (0.09%)
てみません (OP 4)	80 (0.08%)

表 12: 「ください。」の前接要素

を教えて (OC 3)	2533 (0.43%)
提出して (OP 3)	268 (0.28%)
てみて (OC 3)	1252 (0.21%)
はお問い合わせ (OP 3)	200 (0.21%)
をして (OP 3)	196 (0.20%)

「か。」の上位 5 位は、広報紙と国会会議録が占めている。一方、「か？」の上位 5 位は、すべて Yahoo!知恵袋が占めている。広報紙では、区市町村民に呼び掛ける表現として「みませんか。」「みませんか？」の両方が用いられているが、数の上では疑問符よりも句点の方が好まれているようである。一方、Yahoo!知恵袋では質問を投稿する際に「なのでしょうか？」「ないのでしょうか？」「なんですか？」のような形が好んで用いられていると言える。また、「ください。」の上位 5 位は、Yahoo!知恵袋および広報紙が占めている。「教えてください。」「提出してください。」のように、読者や区市町村民に依頼をする表現として用いられている。

その他

「ている。」の上位 3 位は白書 (OW) が占めている。「こととしている。」「となっている。」という文末表現による客観的な描写が、白書で多用されていると言える。一方、「できる。」の上位 5 位は、すべて「〜することができる。」という形が占めている。「動詞+ことができる。」という文末表現が、コロケーションとして用いられていることが示唆される。「できる。」と「ならない。」の上位 3 位は、いずれも法律 (OL) が占めている。分析 1 でも述べたように、ある行為を命令・禁止したり、ある権利を保障したりすることを述べるために、これらの文末表現が多用されていると考えることができる。

表 13: 「ている。」の前接要素

こととし (OW 3)	642 (0.66%)
）となっ (OW 3)	378 (0.39%)
%となっ (OW 3)	344 (0.35%)
」と話し (PN 3)	56 (0.10%)
と考えられ (OT 3)	30 (0.08%)

表 14: 「できる。」の前接要素

することが (OL 3)	1175 (6.66%)
命ずることが (OL 3)	279 (1.58%)
を命ずることが (OL 4)	277 (1.57%)
することが (OT 3)	105 (0.26%)
することが (OW 3)	88 (0.09%)

表 15: 「ならない。」の前接要素

しなければ (OL 3)	1850 (10.49%)
しては (OL 3)	255 (1.45%)
届け出なければ (OL 3)	242 (1.37%)
しなければ (PB 3)	1296 (0.12%)
しなければ (OW 3)	66 (0.07%)

6 分析 3：長い一致を持つ文末表現

最後に、BNAnalyzer が出力する最長の 8gram の例について見てみよう。各レジスターにおいて、8gram の文末表現が総文数に占める比率の高い順に上位 3 位までを抽出した。結果を表 16 に示す。

表 16: 8gram の文末表現（各レジスター上位 3 位まで）

図書館書籍	ていたのではないだろうか。	39 (0.03 ‰)
	しているのではないだろうか。	13 (0.01 ‰)
	について次のように述べている。	12 (0.01 ‰)
出版書籍	といっても過言ではありません。	26 (0.02 ‰)
	ていたのではないだろうか。	24 (0.02 ‰)
	しているのではないだろうか。	18 (0.02 ‰)
雑誌	中から１つ選んでお答えください。	7 (0.04 ‰)
	に満足した？何歳ですか？	6 (0.03 ‰)
	条件をＣＨＥＣＫ！確定申告は必要か？	4 (0.02 ‰)
新聞	郵送かファクスで資料をお寄せください。	4 (0.07 ‰)
	の連絡、資料の返却はしません。	3 (0.05 ‰)
	。落選の方へは連絡しません。	2 (0.04 ‰)
Yahoo!知恵袋	教えてください。よろしくお願いします。	76 (0.13 ‰)
	にはどうしたらいいのでしょうか？	25 (0.04 ‰)
	するにはどうしたらいいですか？	23 (0.04 ‰)
Yahoo!ブログ	ブログは重たいのでこちらからご覧ください。	80 (0.16 ‰)
	心ある対応をよろしくお願い致します。	38 (0.08 ‰)
	と作品の内容は一切関係ありません。	23 (0.05 ‰)
白書	五入のため，合計は百にならない。	21 (0.22 ‰)
	四捨五入のため合計は百にならない。	12 (0.12 ‰)
	．５%）の順となっている。	12 (0.12 ‰)
国会会議録	ますが、御異議ございませんか。	87 (0.75 ‰)
	ますが、御異議ありませんか。	68 (0.59 ‰)
	たいというふうに考えております。	53 (0.46 ‰)
広報紙	※当日、直接会場へお越しください。	49 (0.50 ‰)
	について考えてみませんか。	21 (0.22 ‰)
	ています。詳しくはお問い合わせください。	19 (0.19 ‰)
教科書	１人ぶんは何まいになりますか。	7 (0.18 ‰)
	と，何人に分けられますか。	7 (0.18 ‰)
	どのようにかかわっているのだろうか。	5 (0.13 ‰)
法律	をとるべきことを命ずることができる。	54 (3.06 ‰)
	の物件を検査させることができる。	41 (2.32 ‰)
	、その旨を公示しなければならない。	33 (1.87 ‰)

以下、表 16 に見られる特徴的な点を指摘する。まず、出現率が群を抜いて高いのが法律である。「をとるべきことを命ずることができる。」が 54 回、「の物件を検査させることができる。」が 41 回出現している。これらは、法律文における定形的な言い回しとして用いられている表現と考えられる。同様に、国会会議録の「ますが、御異議 (あり | ござい) ませんか。」「たいというふうに考えております。」も比較的高い出現率を示している。これらも、国会における発言での定形的な言い回し（特に前者は議長や委員長による発言）であると考えてよい。

広報紙では「※当日、直接会場へお越しください。」「について考えてみませんか。」が頻出している文末表現となっている。このうち前者について調べてみると、49 例中 46 例が『広報はままつ』からの例であった。すなわち、特定の広報紙において定形的に用いられている表現であると言える。一方、後者については、さまざまな広報紙が出典となっていた。また、白書の「四捨五入のため合計は百にならない。」という表現は、すべて『中小企業白書』からの例であった。これも、特定の白書において定形的に用いられている表現と言える。

Yahoo!知恵袋の「教えてください。よろしくお願いします。」「にはどうしたらいいのでしょうか?」「するにはどうしたらいいですか？」なども比較的高い出現率を示している。これらは質問を投稿する際の「型」のようなものがユーザー間で共有されているのだろう。

最後に、Yahoo!ブログについて見ると、「ブログは重たいのでこちらからご覧ください。」が 80 回、「心ある対応をよろしくお願い致します。」が 38 回などとなっている。これは、田野村 (2012) が指摘する、同一データが重複して出現している例であると思われる。前者は、宣伝用ブログ記事に含まれるリンクへ誘導するために貼り付けられる文言である。また、後者について調べてみると、「☆乱筆乱文は文学的素養不足の為お許し下さい。今後とも心ある対応をよろしくお願い致します。」という 2 文を、記事の末尾に常に記している書き手によるものであった。

このように見てくると、書き言葉には、（典型的には小説のように）書き手が新たな文章表現を創出する場合だけでなく、ある固定的な文章表現が繰り返し利用される場合も少なからずあることが分かる。後者には、あるレジスターにおいて定形的な表現が好まれて多用されているという場合と、機械的に複製された文章表現が繰り返し出現しているという場合とがあることになる。しかしながら、両者の違いを厳然と区別することは簡単ではない。

そのような重複の問題をノイズとして回避したいという立場がある一方で、例えば、機械的に複製された文章表現が繰り返し出現するのがブログという書き言葉の実態である、という見方もできる。このような問題をどのように捉えるかは、分析を実施する個人に任せられていると言ってよい。

7　まとめ

BCCWJ の 12 種類のレジスターに含まれるテキストを N-gram によって分析し、各レジスターに見られる典型的な文末表現を抽出した。この結果、特に特定目的 SC において、いくつかの文末表現が選考的に多用されていることを明らかにした。

参考文献

丸山岳彦 (2012). 「『現代日本語書き言葉均衡コーパス』を用いた文末表現のバリエーションの分析」. 『言語処理学会 第 18 回年次大会 発表論文集』, pp. 591–594. 言語処理学会.

田野村忠温 (2012).「BCCWJ に含まれるウェブデータの特性について—データ重複の諸相と BCCWJ 使用上の注意点—」. 『第 2 回 コーパス日本語学ワークショップ 予稿集』.

Webデータに基づく複合動詞データベースの構築

山口昌也 (国立国語研究所言語資源研究系) †

Constructing a Japanese Compound Verb Database Based on Web Pages

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

1 はじめに

本稿では，日本語の複合動詞データベースの構築について述べる。対象とする複合動詞は,「切り倒す」「持ち上げる」といった,「動詞（連用形）＋動詞」タイプの複合動詞である。

本データベースを構築する目的は，複合動詞とそれを構成する動詞（以後，構成動詞）との関係を，大量の用例に基づいて分析することである。本研究で特にターゲットするのは，複合動詞と構成動詞の格要素の対応関係である。

例えば，複合動詞「切り倒す」「打ち破る」のヲ格について，構成動詞との対応関係を見てみる。まず,（E1）「切り倒す」の「木を」の場合は,「切る」（E1a）「倒す」（E1b）ともに適格である。一方,（E2）「打ち破る」の「記録を」の場合は,「打つ」（E2a）が不適格である。複合動詞と構成動詞の関係を調べるためには，以上のようなテストを多数行うことにより,「切り倒す」のヲ格は両方の構成動詞のヲ格と対応関係があり,「打ち破る」のヲ格は「破る」のヲ格とだけ関係があると判断する。

E1 太郎が木を切り倒す　　**E2** 太郎が記録を打ち破る
E1a 太郎が木を切る　　**E2a** * 太郎が記録を打つ
E1b 太郎が木を倒す　　**E2b** 太郎が記録を破る

このような格要素の対応関係分析は，複合動詞と構成動詞間の格支配関係分析（山本 1984 など）や LCS による意味構造の記述（影山 1993，由本 2005 など）といった，より複雑な分析の基礎となる。しかし，分析は内省によることが多く，網羅的・客観的に分析結果を検証することが困難である。

また，現状では，客観的な分析をするための資料が十分整備されていない。例えば，言語学的な資料としては,『複合動詞資料集』（野村・石井 1987）や『合成語のためのデータベース』（山下 2007）,『複合動詞リスト』（姫野 1999）などが作成されているが，語構成，連接頻度情報といった語自体の情報が主体であり，用例や格要素の情報は収録されていない。自然言語処理の分野でも，形態素解析システムの辞書の中に複合動詞が登録されている。しかし，網羅的に登録されておらず，形態素解析システムでコーパスを解析しても，すぐには複合動詞用の資料として活用できず，解析後に複合動詞を再認定する必要がある。

以上の背景のもと，上記の内省のプロセスを，大量の実例に基づいて行うための複合動詞データベースを構築する。データベースには，複合動詞とその構成動詞を収録する。収録の可否は，用例が一定量収集できるか否かにより決定する。用例は，多様性，および，収集コストを鑑みて，Web から収集する。個々の動詞は，用例，格要素の情報を保持する。また，複合動詞の場合は，語構成の情報をあわせ持つ。

この後の節では，構築する複合動詞データベースの構造の設計について説明した後，Web データから半自動的に複合動詞データベースを構築する方法を示す。さらに，構築された複合動詞データベースの内容を概観する。最後に，複合動詞データベースを使った分析例を示す。

なお，本研究は，国立国語研究所の共同研究プロジェクト「文脈情報に基づく複合的言語要素の合成的意味記述に関する研究」*の一環として行っている。

†http://www2.ninjal.ac.jp/masaya
*http://www.ninjal.ac.jp/research/project/c/bunmyaku/

2　データベースの構築

2.1　設計方針

次の三つの設計方針のもと，複合動詞データベースを構築した。

(a) 対象とする複合動詞は，いわゆる「語彙的複合動詞」(影山 1993) とする
(b) 一定数以上の用例を収集可能な複合動詞を収録する
(c) Web データの特性に合わせた，用例・格要素情報の作成を行う

(a) により，対象とする複合動詞を限定する。影山 (1993) では，生成文法の見地から,「動詞（連用形）＋動詞」タイプの複合動詞を，語彙的複合動詞と統語的複合動詞に分類している。このうち，語彙的複合動詞は，語彙部門で派生され，レキシコンに記述されるタイプの複合動詞である。統語的複合動詞は，統語的複合動詞は統語部門で形成される。統語的複合動詞の前項・後項動詞は,「食べ始める」（＝食べるのを始める）「使い慣れる」（＝使うのに慣れる）というように補文的な関係を持ち，意味的に透過的な合成が行われる。本研究で統語的複合動詞を扱わないのは，その性質上，前項動詞と複合動詞との格関係が明らかだからである。

(b) は，データベースの利用目的である,「複合動詞と構成動詞との関係を，大量の用例に基づいて分析する」ことを達成するために設けた。このことは，Web の使用実態を反映した複合動詞を収録することにもつながる。

(c) は，通常の書籍にはない，Web 特有の性質に対応するために設けた。例えば，Web データには，同一表現の繰返し（例：掲示板におけるスレッドのタイトル）や，ページ単位での引用など，用例・格要素の頻度を計測する際に「ノイズ」となる要素が存在する。これらの影響を軽減しつつ，用例・格要素を収集する。

2.2　Web コーパスの構築と収録動詞の選定

データベースに収録する複合動詞，構成動詞の選定手順は，次のとおりである。この方法の特徴は，個々の複合動詞，構成動詞ごとに Web コーパスを構築し，収録できる用例の量を確認しつつ，収録動詞を選定していくことである。

(1) 『複合動詞資料集』から，複合動詞の構成要素として多用される動詞上位 10 語を選択し,「種動詞」とする。そして，Baroni・Bernardini (2004) の方法で，それぞれ個別に Web コーパスを作成する。個々の Web コーパスのサイズは，10000 ページ（前項の動詞用に連用形で 5000 ページ，後項の動詞用に終止形で 5000 ページ）である。
(2) 作成した Web コーパス中のデータを形態素解析した後，種動詞＋動詞，動詞＋種動詞のパターンを抽出し，複合動詞候補の頻度表を作る。
(3) 複合動詞候補のうち，一定数以上（今回は頻度 5 以上）出現した複合動詞候補を人手で確認し，個別に Web コーパス（サイズは 2000 ページ）を作成する。さらに，2.3 節の「用例の抽出と格解析」を行ない，最終的にデータベースに収録するかを決定する。
(4) 収録が決定した複合動詞の構成動詞を種動詞として，再帰的に (1)〜(4) を繰り返す。例えば，複合動詞「切り捨てる」の場合，後項動詞の「捨てる」が種動詞となる。

2.3　用例の抽出と格解析

前節までの処理で，動詞ごとの Web コーパスが構築される。これらの Web コーパスを対象に，用例の抽出と格解析を行う。なお，用例の抽出と格解析は，個別の Web コーパスごとに行うものであり，統合した Web コーパスに対して，実施するわけではない。

用例抽出は，収集対象の動詞を含む「文」を単位とする。「文」の区切りは，句読点，空白文字を用いた。収集対象の動詞を含むか否かは，形態素解析した結果に基づいて判断する。用例を 100 例以

上収集できた動詞は収録対象とし，個々の用例に対して構文解析，および，格解析を行う。この結果をもとに，用例に格要素情報（格助詞と格要素のペア）を付与する。なお，形態素解析には JUMAN (ver.6.0)，構文解析・格解析には KNP (ver.3.01) を用いた。

なお，設計方針 (c) に対応するため，収録動詞を決定する際の用例数の計測と，データベースに対する検索処理（3.2 参照）の際に，次の処理を行なっている。

- 用例の出現頻度は，出現ページ数として計測する。ただし，まったく同一の用例は複数のページに出現していたとしても，重複して計測しない。また，データベースへの重複登録も行わない。
- 格要素の名詞の出現頻度も，出現するページ数（以後，出現ページ数）として計測する。例えば，同一の Web ページ内で「畑でトマトを作る」「太郎がトマトを作る」という用例が出現したとしても，ガ格の格要素としての「太郎」は，頻度 1 である。

3 構築されたデータベース

3.1 収録データ

2 節の方法を用いて，複合動詞データベースを構築した。収集した動詞は，複合動詞 3399 語，構成動詞 1075 語である†。各動詞ごとに収集された用例は，複合動詞が平均 1088.4 文（異なりページ数 784.8 ページ），構成動詞が平均 7839.1 文（異なりページ数 2922.8 ページ）となった。複合動詞のうち，用例数が 1000 以上収集できたものは，1839 動詞であった。用例数の分布をヒストグラムにした結果を図 1，2 に示す。

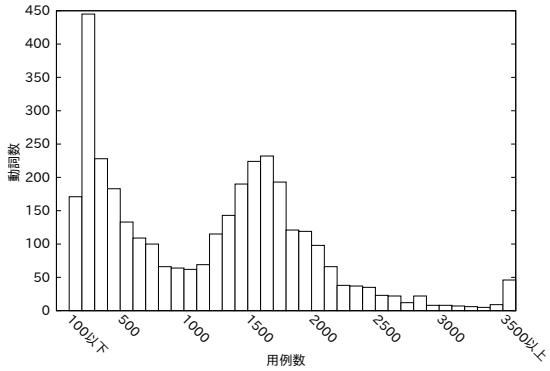


図 1: 収集された用例数の分布（複合動詞）

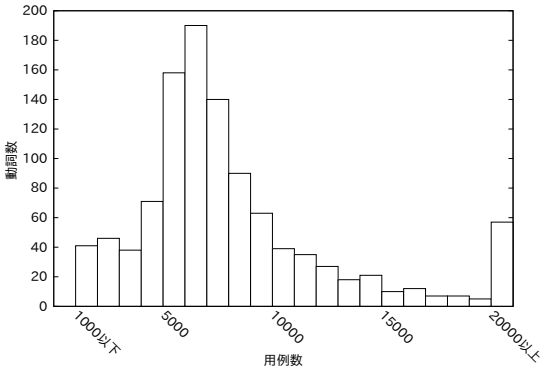


図 2: 収集された用例数の分布（構成動詞）

次に，構成動詞の内訳は，異なりで 999 語となった。前項・後項動詞を個別に見てみると，前項動詞が異なりで 719 語，後項動詞が異なりで 578 語である。使用頻度上位 5 位を，それぞれ表 1，2，3 に示す。

表 1: 構成動詞上位 5 位（前項）

表記	読み	度数
見る	みる	73
引く	ひく	67
取る	とる	60
打つ	うつ	58
突く	つく	48

表 2: 構成動詞上位 5 位（後項）

表記	読み	度数
込む	こむ	228
上げる	あげる	130
出す	だす	122
上がる	あがる	70
取る	とる	65

表 3: 構成動詞上位 5 位（前後）

表記	読み	度数
込む	こむ	231
上げる	あげる	131
出す	だす	126
取る	とる	125
見る	みる	80

†本稿執筆時点では構築途中のため，今後，増減する可能性がある

3.2　データベースの機能

構築したデータベースは，言語研究者を始めとした一般の利用者が手軽に利用できるように，検索用のサイト‡を試験的に公開している。ここでは，その機能の一部を紹介する。

複合動詞検索　読み，もしくは，表記を指定して，複合動詞を検索する。構成動詞を指定した場合は，当該の構成動詞の他，構成動詞を含む複合動詞の一覧が検索される。一覧から詳細に調べたい動詞を選択すると，当該動詞と構成動詞との「重複度」（山口 2012），および，格要素一覧（後述）などが表示される。図 3 は,「言い当てる」を検索した例である。

重複度は，二つの動詞の格要素が重複して使用される度合いを表す。図 3 の例では,「言い当てる」と「言う」の重複度が 46%となっている。これは,「言い当てる」のヲ格の格要素のうち,「言う」のヲ格で使用されている格要素が 46%あることを示している。このように，重複度により，複合動詞と構成動詞との関係を客観的に捉えることができる。

格要素閲覧　検索した動詞の格要素を，格ごとに一覧表にして表示する。図 3（下段の表）が,「探し出す」の格要素一覧である。それぞれの格要素の右には，出現したページ数が付与される。格要素は，ページ数で降順に表示され，格も格要素のページ総数が多い順に左から右に表示される。

複合動詞と構成動詞との関係分析を支援する機能として，構成動詞でも使用される格要素を色分けして表示することができる。また，それぞれの格要素には,（後述の）「用例閲覧」機能用のリンクがあり，容易に用例を閲覧できるようになっている。

用例閲覧　用例を文単位で表示する。各用例には，出典情報として，取得元の Web ページの URL が併記される。

重複度/格パターン

	ヲ	修飾	ガ	デ	ニ	時間	総ページ数
言い当てる	631 p	146 p	36 p	42 p	10 p	11 p	1129 p
言う	46%	90%	100%	21%	70%	64%	3429 p
当てる	44%	85%	75%	57%	70%	64%	2050 p

格要素一覧　（重複：v1　v2　v1,2　off）

ヲ		修飾		ガ		デ		ニ		時間	
こと	59	正確に	34	人	12	見た	17	とき	4	時	4
名前	46	ズバリ	32	者	9	一言	9	こと	4	瞬時	4
未来	27	ずばり	15	それ	6	発	7	前	3	あと	3
本質	23	見事に	13	これ	5	見る	5	夜明け	3		
犯人	20	的確に	7	方	3	聞いた	4				

図 3: データベースの検索結果例

4　活用例：格要素の重複度による複合動詞・構成動詞の分析

4.1　分析のねらい

構築した複合動詞データベースの活用例として，重複度を用いて，複合動詞・構成動詞間の関係を分析してみる。

複合動詞と構成動詞間の関係を記述する場合，従来の分析では，構成動詞の性質がどのように複合動詞に継承されるか，十分に記述されていない。例えば，山本 (1984) では，複合動詞と構成動詞とが格支配構造上の関係を持つか否かによって，複合動詞を 4 種類に分類している。また，由本 (2005) の LCS（Lexical Conceptual Structure）による分析では，構成動詞の LCS が複合動詞の LCS の一

‡http://csd.ninjal.ac.jp/comp/

部に組み込まれる形で記述されている。これらの分析では，構成動詞の性質が複合動詞にそのまま継承されるように見える。

それでは，複合動詞と構成動詞に密接な関係があると考えられる「投げ捨てる」の場合,「投げ捨てる」のヲ格の格要素は,「投げる」のヲ格の格要素としても使えるだろうか？　逆に，前項動詞の意味が希薄な「打ち捨てる」のヲ格の格要素は,「打つ」でもヲ格の格要素とならないだろうか？ この疑問を重複度を用いて検証することが，本節の分析のねらいである。

4.2　分析対象の動詞と重複度

ここでは，問題を簡略化するために，対象とする動詞を，後項動詞に「込む」を持つ複合動詞とし，対象とする格はヲ格とする。また，格解析などの誤りによるノイズを軽減するため，対象とする動詞には，(a) 複合動詞，構成動詞ともに 1000 例以上の用例を持つこと，(b) ヲ格の格要素を 50 例以上持つこと，を条件として加えた。この結果，対象となる複合動詞は，99 語となった。

分析対象の複合動詞の重複度を複合動詞データベースに基づいて計算し，重複度の昇順にプロットした結果を図 4 に示す。

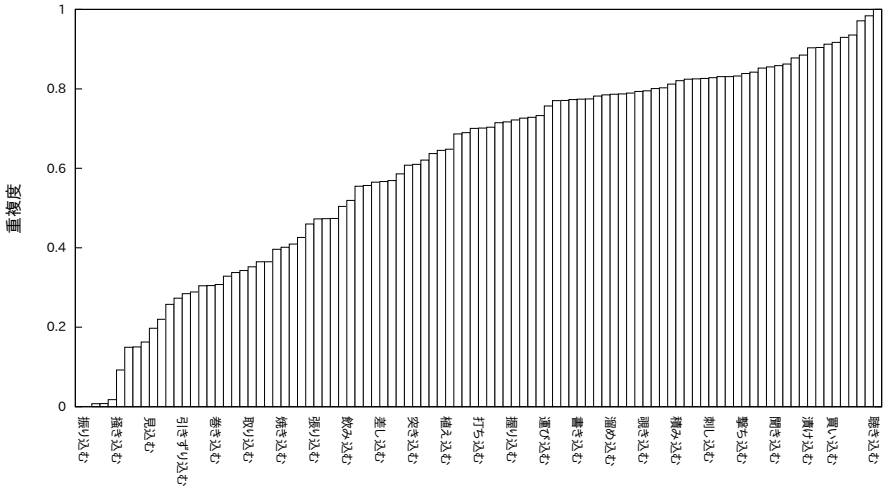


図 4: 重複度の分布

4.3　重複度による複合動詞・構成動詞間の関係の分類

既存の分析を極端に解釈すれば，重複度は，0 と 1 の両端周辺に集中するはずである。しかし，図 4 のとおり，幅広く分布している。この要因を探るために，重複度によって，複合動詞・構成動詞間の関係を次の四つに分類し，考察した。

継承　重複度が 1.0 に近い場合である。この場合，複合動詞の格要素は，構成動詞とほぼ一致する。一致しない格要素については，データ不足が原因と考えられる。今回，重複度 0.9 以上の動詞は 8 例あった。上位の 3 語（括弧内は重複度）は,「聞き込む」(1.0)「着込む」(0.97)「塗り込む」(0.94) である。このうち，一致しなかった格要素を見てみると,「レインウェアを 着込む」「溶剤を 塗り込む」というように，意味的には構成動詞の格要素としても問題ないものだった。このような不一致は，収集する用例数を増やせば，減少するものと考えられる。

別義　継承とは逆に，重複度が 0 に近い場合である。この場合，複合動詞と構成動詞との意味的な関係が少ない，別義と考えられる複合動詞であった。重複度の下位 3 語は,「振り込む」(0.0)「擦り込む」(0.01)「申し込む」(0.01) である。なお，わずかながら一致した格要素は，次のように，限定された格要素だけであった。

- 「傷口にアルカリ性の 墨を 擦り込む」(http://b.z-z.jp/thbbs.cgi/kangofu/331/)
- 「その 旨を フロントまたは管理室に申し込む」(http://www.mm-m.ne.jp/glkumiai/kiyaku_2.html)

派生　継承と別義が混在することにより，重複度が減少する場合である。「織り込む」(0.29) と「追い込む」(0.56) の例を次に示す。いずれの例共に，右側が継承，左側が別義の関係にある。このように，別義に相当する格要素が，重複度の減少の要因となっている。

糸を織り込む ⟺ 糸を織る　　最新の情報を織り込む ⟺ * 情報を織る
魚を追い込む ⟺ 魚を追う　　内閣を総辞職に追い込む ⟺ * 内閣を総辞職に追う

分布変化　格要素の生起確率の分布が，複合動詞と構成動詞で大きく異なる場合である。まず，実際の例として,「流し込む」の格要素を見てみよう。表 4 は，複合動詞となった時に生起確率が上昇した格要素，表 5 は下降した格要素である。前者は上位 10 語のうち，重複度を減少させた 5 語（つまり，頻度 0 だった語）を挙げている。後者は，上位 5 語である。

表 4: 生起率が上昇した格要素

格要素	複合動詞	構成動詞	増減
モルタル	0.014	0.000	0.014
樹脂	0.012	0.000	0.012
ビール	0.012	0.000	0.012
金属	0.011	0.000	0.011
セメント	0.011	0.000	0.011

表 5: 生起率が下降した格要素

格要素	複合動詞	構成動詞	増減
血	0.002	0.068	-0.066
水	0.043	0.110	-0.067
電流	0.009	0.112	-0.103
情報	0.001	0.118	-0.117
涙	0.001	0.219	-0.218

表 4 に挙げた格要素は，今回，構成動詞「流す」のヲ格の格要素としては出現しなかったが，どの格要素も意味的に「流す」の格要素となりうる。これは，重複度の面から見ると,「継承」と同じ状況である。しかし,「継承」の場合と異なるのは，格要素の生起確率に，系統的な分布変化が見受けられる点である。「流し込む」の例で言えば，モルタルなど，何かの材料となるものや，アルコール飲料などの変化が大きかった。これは，複合動詞と構成動詞との関係を記述する上で，重要な情報になる。今後，生起率が下降した格要素（表 5）とあわせて，分析を進める予定である。

5　おわりに

本稿では，Web データに基づいて，日本語複合動詞のデータベースを半自動的に構築する方法を示し，実際に構築した結果を紹介した。さらに，重複度の面から複合動詞と構成動詞の関係分析を行ない，生起確率の分布変化が両者の関係を記述する際に重要になることを示した。現時点で収録されている複合動詞は 3399 語であり，そのうち，用例数が 1000 以上の複合動詞が 1839 語ある。今後は，複合動詞と構成動詞との関係を分析する過程で，データベースの質的な改善を図る予定である。

参考文献

山本清隆 (1984) 複合動詞の格支配，都大論究，Vol.21，pp.32–49

影山太郎 (1993) 文法と語形成，ひつじ書房

由本陽子 (2005)『複合動詞・派生動詞の意味と統語』，ひつじ書房，pp.110–129

野村雅昭，石井正彦 (1987) 複合動詞資料集，科研費特定研究 (1) 言語データの収集と処理の研究

山下喜代 (2007) 日本語教育のための合成語のデータベース構築とその分析，科学研究費補助金 研究成果報告書

姫野昌子 (1999)『複合動詞の構造と意味用法』，ひつじ書房，pp.245–260

M. Baroni and S. Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004

山口昌也 (2012) 複合動詞と構成要素動詞の格要素の対応関係分析，言語処理学会第 18 回年次大会予稿集

日本語話し言葉コーパスを用いた
複合境界音調の発言継続表示機能の検討

小磯 花絵（国立国語研究所理論・構造研究系）†

Continuation Function of Boundary Pitch Movements in the Corpus of Spontaneous Japanese

Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

1. はじめに

文（発話）の末尾に上昇調など様々な音調が見られるように，文の内部にも様々な音調が現れる。上昇調や上昇下降調など，単純に下降せず上昇成分を伴う音調はこの典型と言えよう。郡（1996,2003）は，文中（特に文節末）に生じる上昇調や上昇下降調は大きな意味の区切りに生じることが多く，またそのあとにポーズが置かれることも多いことから，これらの音調には，文の意味の区切りを明確にすることで伝達効果を高め，ポーズがあっても発言がまだ続くことを示す機能（以下「発言継続表示機能」）があると指摘している。発言継続表示機能は，会話において，話し手が発言の継続を示しターンを維持する機能につながることを考えると，これらの音調の基本的な役割と考えることができる。しかし少なくとも日本語については，上昇調，上昇下降調の持つ発言継続表示機能に関し，十分なデータに基づく実証的な研究はあまり見られない。

そこで本研究では，『日本語話し言葉コーパス (*Corpus of Spontaneous Japanese*: CSJ)』（前川 2004,2006）を対象に，上昇調，上昇下降調の出現傾向を主に統語構造との関係から調査することによって，これらの音調の発言継続表示機能について検討することを目的とする。小磯（2012）では，CSJ コアの独話と対話の比較を通して上昇調，上昇下降調の特徴について検討したが，CSJ コアの対話には一部付与されていないアノテーション情報が存在するため分析が制限された。そこで本研究では，CSJ コアの独話に限定した上で，付与されているアノテーション情報のうち，主に節単位情報と係り受け構造情報を活用しながら，統語構造と上昇調，上昇下降調の出現傾向との関係について多角的に調査をする。この結果を踏まえ，上昇調，上昇下降調の持つ発言継続表示機能について検討する。

2. 方法

2.1 データ

分析には CSJ のコアデータのうち，学会講演 70 ファイル（約 19 時間），模擬講演 107 ファイル（約 20 時間）を用いた。模擬講演とは一般話者による主に個人的な内容に関するスピー

† koiso@ninjal.ac.jp

チのことである。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）を用いた。

2.2 韻律情報

CSJ コアに付与されている X-JToBI に基づく韻律情報（五十嵐ほか 2006）を用いて句末境界音調の情報を抽出した。句末境界音調には，下降調 L%（対象ファイル中 116997 件）に加え，複合境界音調（Boundary Pitch Movement: BPM）として，単純な上昇調 L%H%（29599 件），上昇前に一定期間低ピッチが見られる上昇調２ L%LH%（349 件），上昇下降調 L%HL%（9901 件），上昇下降上昇調 L%HLH%（8 件）があるが，上昇調２と上昇下降上昇調は頻度が低いため分析対象外とした。後述するように，本分析では文（に相当する単位）の内部の句末境界音調を対象とするため，いわゆる疑問の上昇調は分析対象外となる。文の途中に生じる半疑問についても対象外とした。また，語断片など言い淀みに付与された句末境界音調も分析から除外した。

X-JToBI では，韻律境界の切れ目の強さ（Break Index, 以下 BI）に関する情報が原則 1〜3 の整数によって表現される。概ね BI=1 は語境界，BI=2 はアクセント句（AP）境界，BI=3 はイントネーション句（IP）境界に相当し，この順に韻律上の切れ目の強さが増す。上記句末境界音調は原則として BI=2 以上の境界に付与される。BPM が AP 境界（正確には IP 終端以外の AP の境界）に生じた場合，BI=2+b として BI=2 と BI=3 の中間の切れ目の強さとされる。同様に AP にポーズが後続した場合も BI=2+p（BPM を伴う場合は BI=2+bp）として BI=2, 3 の中間とみなされる。分析では，IP 終端以外の AP 境界（BI=2+b，以下単純に AP 末と記す）と IP 境界（BI=3）という BPM の出現位置の違いについても検討する。

2.3 統語情報

文法的・意味的な切れ目の大きさの指標として，CSJ コアに付与されている節単位情報（丸山ほか 2006）を用いた。節単位とは，原則「節 (clause)」の境界によって得られる文法的・意味的なまとまりを持った単位である。境界の切れ目の大きさの観点から以下の三つに分類される。

絶対境界　いわゆる文末に相当する境界

強境界　後続の節に対する従属度の低い，つまり切れ目の度合が大きい節境界

弱境界　後続の節に対する従属度の高い，つまり切れ目の度合が小さい節境界

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，原則，絶対境界か強境界のいずれかで区切られる単位が節単位と認定される。

本研究で着目するのは，文に相当する絶対境界の内部の文法的・意味的な切れ目の大きさと句末境界音調との関係であるため，絶対境界（文末相当）以外の箇所に出現した句末境界音調を分析対象とし，対応する節境界の種類（「強境界」・「弱境界」・「非境界（上記三種の節境界がない位置）」）との関係を見た。自動判別後の人手修正の結果，節境界クラスに変更があったもの（倒置などの理由で強境界が非境界に変更となったものなど）は分析対象外とした。また弱境界のうち感動詞，フィラー文，接続詞と判断されたものも分析対象外とした。

また，節単位を範囲とした文節単位の係り受け構造情報（内元ほか 2004）を用いて計算される係り先の距離も，節単位内の統語的な切れ目の大きさの指標の一つとして分析に用いた。

表 1　節境界クラスごとの句末境界音調の出現頻度・出現率

	模擬講演			学会講演			講演全体		
	強境界	弱境界	非境界	強境界	弱境界	非境界	強境界	弱境界	非境界
下降調	889	6120	49071	257	3212	43928	1146	9332	92999
	25.2%	61.9%	86.4%	12.8%	48.2%	74.9%	20.7%	56.4%	80.5%
上昇調	1163	2081	4994	1273	3000	13106	2436	5081	18100
	33.0%	21.0%	8.8%	63.5%	45.0%	22.3%	44.1%	30.7%	15.7%
上昇下降調	1471	1692	2733	474	449	1653	1945	2141	4386
	41.8%	17.1%	4.8%	23.7%	6.7%	2.8%	35.2%	12.9%	3.8%

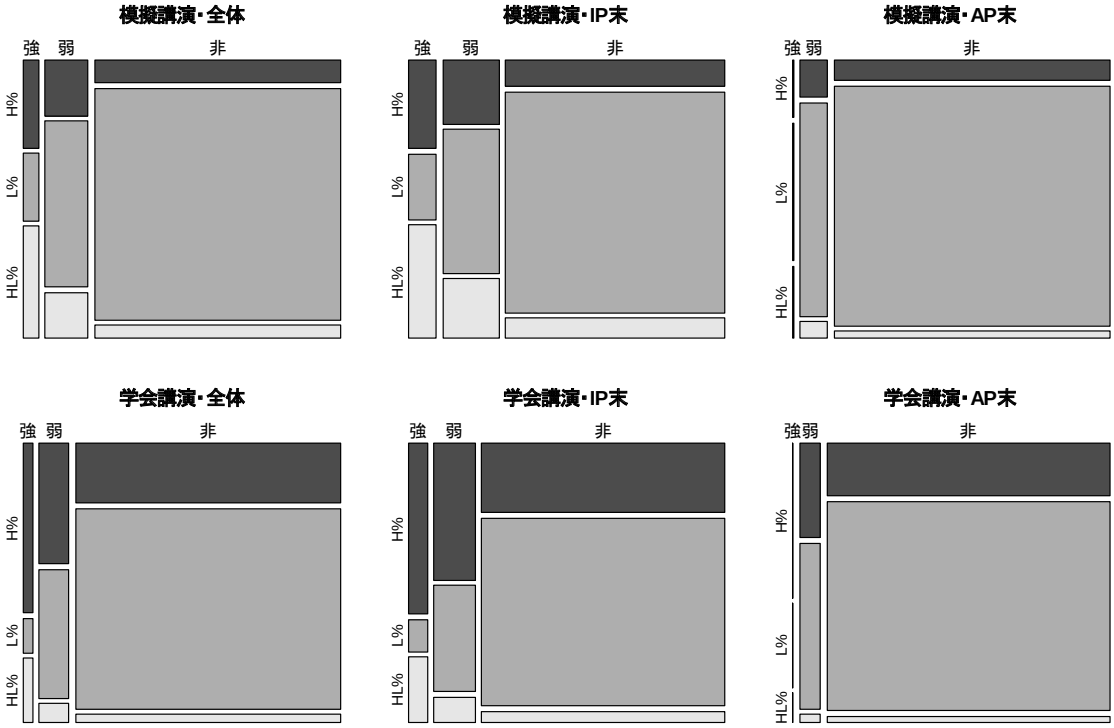


図 1　節境界クラスごとの句末境界音調の出現傾向（強：強境界，弱：弱境界，非：非節境界。L%：下降調，H%：上昇調，HL%：上昇下降調。各棒の太さの違いは節境界クラスの頻度の比率を反映。）

AP と終端が一致する文節（つまり終端に句末境界音調が生じる文節）を対象に，当該文節が「**被験者に**／ 提示した」のように直後の文節に係る場合は距離を 1，「**これらの音を**／それぞれ／三回ずつ／被験者に／ 提示した」のように 4 つ先の文節に係る場合は距離を 4 とした。ここでは単純に係り先の距離が遠い方がより統語的な切れ目は大きいとみなす。係り先の距離の分析では係り先情報のない文節は対象外とした。

3.　結果

3.1　統語的・意味的な切れ目の強さと上昇調・上昇下降調の出現率

■節境界クラス：　学会講演と模擬講演は話し方のスタイルに差があることから（前川 2011），分けて分析した。表 1 と図 1（左端「全体」参照）に，節境界クラスごとの句末境界音調の出

表 2　AP 末・IP 末別にみた節境界クラスごとの句末境界音調の出現頻度・出現率

	模擬講演						学会講演					
	IP 末			AP 末			IP 末			AP 末		
	強境界	弱境界	非境界	強境界	弱境界	非境界	強境界	弱境界	非境界	強境界	弱境界	非境界
下降調	839	3631	21698	50	2489	27373	231	1626	17031	26	1586	26897
	24.5%	53.6%	82.6%	51.5%	79.8%	89.7%	12.0%	39.6%	70.2%	31.3%	62.0%	78.1%
上昇調	1142	1640	2637	21	441	2357	1225	2100	6323	48	900	6783
	33.3%	24.2%	10.0%	21.6%	14.1%	7.7%	63.8%	51.2%	26.1%	57.8%	35.2%	19.7%
上昇下降調	1445	1504	1944	26	188	789	465	377	911	9	72	742
	42.2%	22.2%	7.4%	26.8%	6.0%	2.6%	24.2%	9.2%	3.8%	10.8%	2.8%	2.2%

現頻度・出現率を示す。模擬講演と学会講演のいずれにおいても，統語的・意味的な切れ目が大きくなるほど（つまり非境界＜弱境界＜強境界の順に），上昇調，上昇下降調ともに出現率が高くなる傾向が見られる。

次に，句末境界音調の出現位置が IP 末（BI=3）の場合と AP 末（BI=2，2+b）の場合に分けて節境界クラスごとの句末境界音調の出現頻度を求めた。結果を表 1 と図 1（中央「IP 末」・右端「AP 末」参照）に示す。

図 1 から分かるように，強い文法的・意味的境界である強境界では，そもそも強い韻律境界である IP 末となることが多く AP 末は極めて少ない傾向にあるが，このような偏りはあるものの，AP 末，IP 末の内部で見た場合，先に見た傾向，つまり統語的・意味的な切れ目が大きくなるほど上昇調，上昇下降調の出現率がそれぞれ高くなるという傾向が，模擬講演，学会講演の両者に観察される。

■係り先の距離：　弱境界，非境界を対象に，係り先の距離と句末境界音調との関係を調べた。結果を図 2（左二列「全体」）に，IP 末と AP 末に分けた結果を図 3 に示す。また，係り先の距離が 1 の場合（直後に係る場合）と 2 以上の場合（係り先が離れている場合）に分けて，各音調の頻度を IP 末・AP 末，弱境界・非境界ごとに求めた結果を表 3 に示す。

図 2（左二列「全体」）および図 3 から，いずれの条件においても，係り先の距離が遠くなるほど上昇調，上昇下降調ともに出現率が高くなる傾向が見てとれる。この傾向は特に距離が 1 から 4，5 の範囲で強い。模擬講演では，弱境界，非境界ともに上昇調，上昇下降調の出現率はほぼ同程度であるが，学会講演については上昇調の使用が目立ち，相対的に上昇下降調の比率は低くなっている。しかし低頻度ながらも，他と同様，係り先の距離が遠くなるほど出現率が高くなる傾向が確認できる。また表 3 から，IP 末・AP 末，弱境界・非境界のいずれにおいても，直後に係る場合（距離が 1 の場合）に上昇調，上昇下降調があまり出現しないことが頻度の面からも分かる。この傾向は特に上昇下降調に強く見られる。

係り先の距離の遠近の違いがある特定の文法構造と関わり，それが句末境界音調の選択に影響している可能性も考えられる。そこで，条件を揃えるため，弱境界からは頻度の高い「テ節」を，非境界からは格助詞「の」を伴い連体修飾格となる事例を選択し，改めて係り先の距離と各音調の出現率との関係を調べた。「テ節」の結果を図 2（右二列上段「テ節」）に，格助詞「の」の結果を図 2（右二列下段「格助詞の」）にそれぞれ示す。全体の場合と同様，係り先

表 3　係り先の距離が 1 の場合と 2 以上の場合の各音調の出現頻度

		IP 末				AP 末			
		弱境界		非境界		弱境界		非境界	
		距離 1	2 以上	距離 1	2 以上	距離 1	2 以上	距離 1	2 以上
模擬講演	下降調	1011	2454	7397	9429	1597	721	13849	6321
	上昇調	171	1412	633	1729	206	204	1054	709
	上昇下降調	78	1403	273	1578	26	155	257	484
学会講演	下降調	510	1043	5508	7390	1065	407	13895	5022
	上昇調	156	1907	1395	4467	459	413	3789	2114
	上昇下降調	12	361	101	793	21	49	445	269

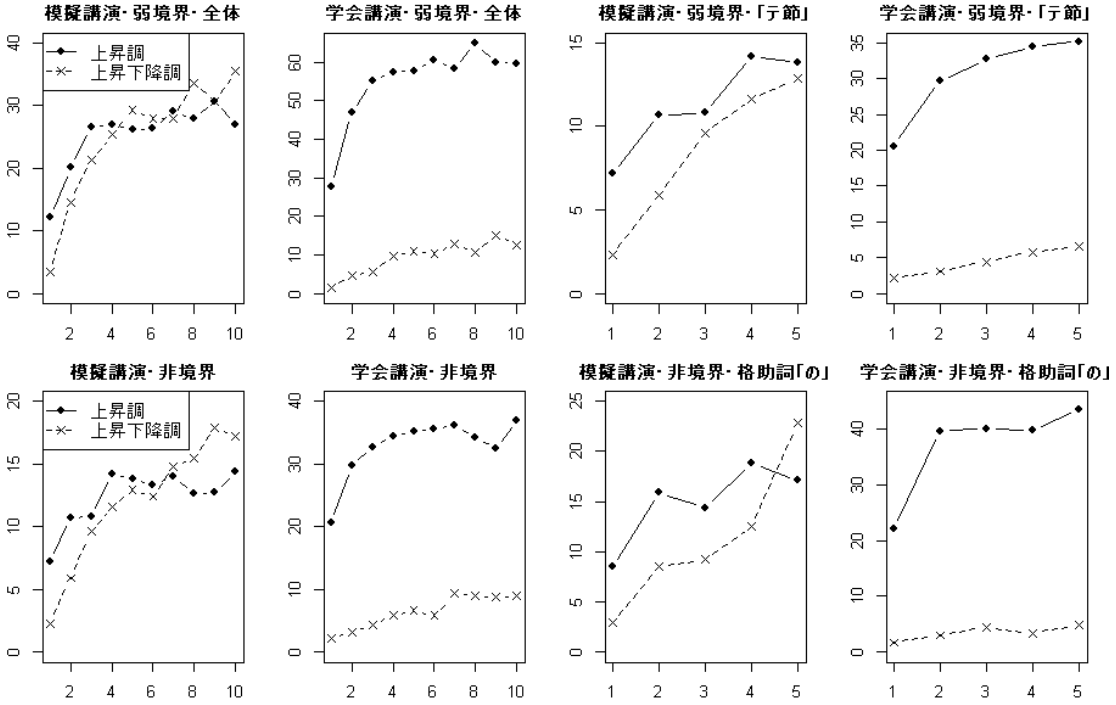


図 2　係り先の距離と上昇調・上昇下降調の出現率との関係

の距離が遠くなるほど出現率が高くなるという傾向が見られる。これらの事例についても IP 末と AP 末に分けた上で同様の分析をしたが，いずれにおいても全体と同じ傾向を示した（スペースの都合で図は省略する)。

■挿入節，引用節，連体節：　ここでは，以上の分析で対象外とした次のような箇所（具体例の下線部）に焦点を当てる。

引用節　例：{苦労してる けども 頑張ろうよ} っていうことでね

連体節　例：また {見かけ上は連用形名詞です が 合成語の省略形である} ものもありました

挿入節　例：色んなパターンを {ここに書いてある数字は頻度です が} 沢山集めてみました

いずれも本来は強境界の位置であるが，引用節内，連体節内，挿入節内に現れているため，CSJ の節単位情報では強い統語境界（節境界）とはみなされない箇所である。上昇調・上昇下降調

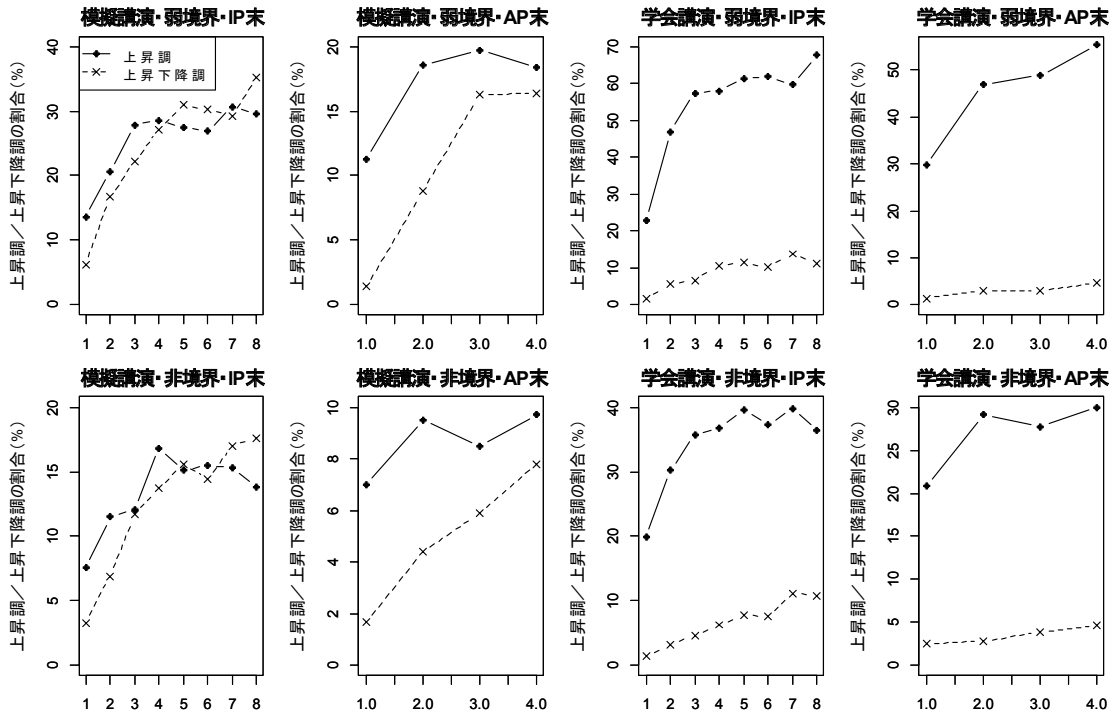


図 3　係り先の距離と上昇調・上昇下降調の出現率との関係—IP 末・AP 末別—

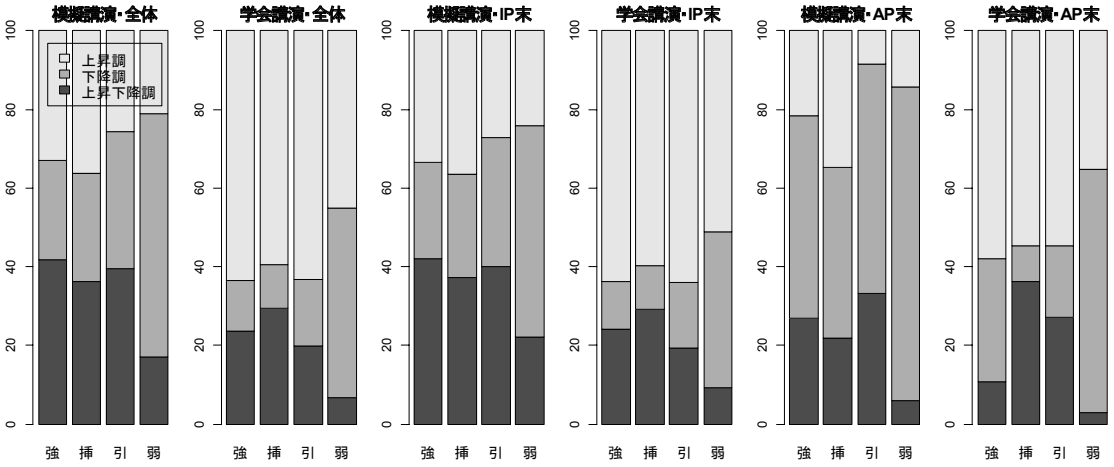


図 4　挿入節，引用節・連体節における上昇調・上昇下降調の出現率（強：強境界，挿：挿入節，引：引用節・連体節，弱：弱境界）

の出現率を図 4 に示す。引用節と連体節は構造の類似性から結果をまとめて示す。また参考のために強境界と弱境界の結果も合わせて示す。図から，引用節・連体節，挿入節のいずれも，強境界と同程度に上昇調，上昇下降調の出現率が高いことが分かる。

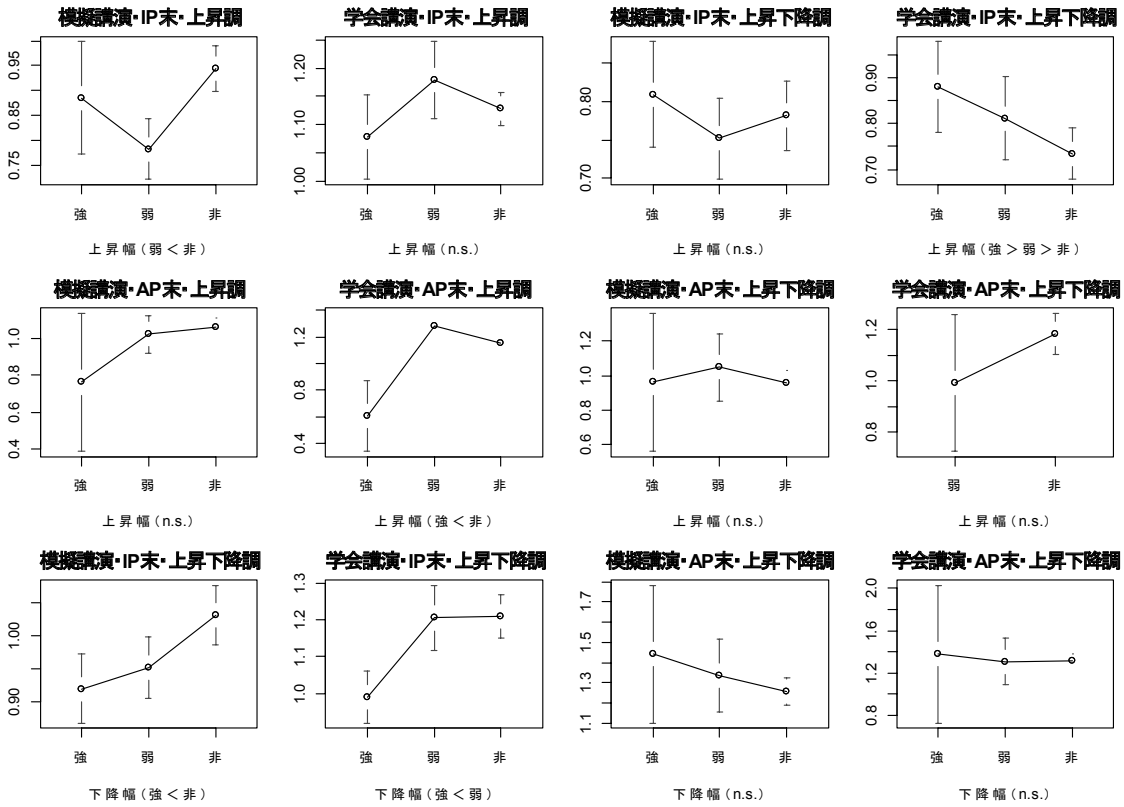


図 5　節境界クラスごとの上昇調・上昇下降調の上昇幅・下降幅（先行アクセント核数を 1 に限定。5% 水準で有意差が見られる箇所を括弧内に記載。）

3.2　統語的・意味的な切れ目の強さと上昇調・上昇下降調の大きさ

統語的・意味的な切れ目の大きさが，上昇調，上昇下降調の継続長や F0 の振幅に影響するかを検討した。統語的・意味的な切れ目が大きくなるほど，上昇調や上昇下降調をより高く，より長く発話するといったように，音調の大きさを変化させている可能性があるためである。本稿では，スペースの都合で次のパラメータに限定して結果を報告する。なお F0 値はファイル毎に正規化してから分析に用いた。

上昇幅　下降位置の F0 値と上昇成分の終端位置の F0 値の差（両音調）

下降幅　上昇成分の終端位置の F0 値と最終下降位置の F0 値の差（上昇下降調のみ）

■節境界クラス：　BPM のピッチレンジには，同一 IP 内の先行するアクセント核の数が影響する可能性があるため（Igarashi & Koiso 2012），ここでは先行アクセント核数が 1 のものに限定して，節境界クラスごとに各パラメータの平均を求めた。結果を図 5 に示す。図から，節境界の切れ目の大きさに応じて単調に上昇ないし下降する傾向を学会講演と模擬講演の両方に共通して示すものは，IP 末に生じた上昇下降調の下降幅だけであることが分かる。この傾向は AP 末には見られない。

■係り先の距離：　次に，同様に先行アクセント各を 1 に限定した上で，係り先の距離ごとに各パラメータの平均を求めた。上昇幅の結果を図 6 に，下降幅の結果を図 7 に示す（係り先の

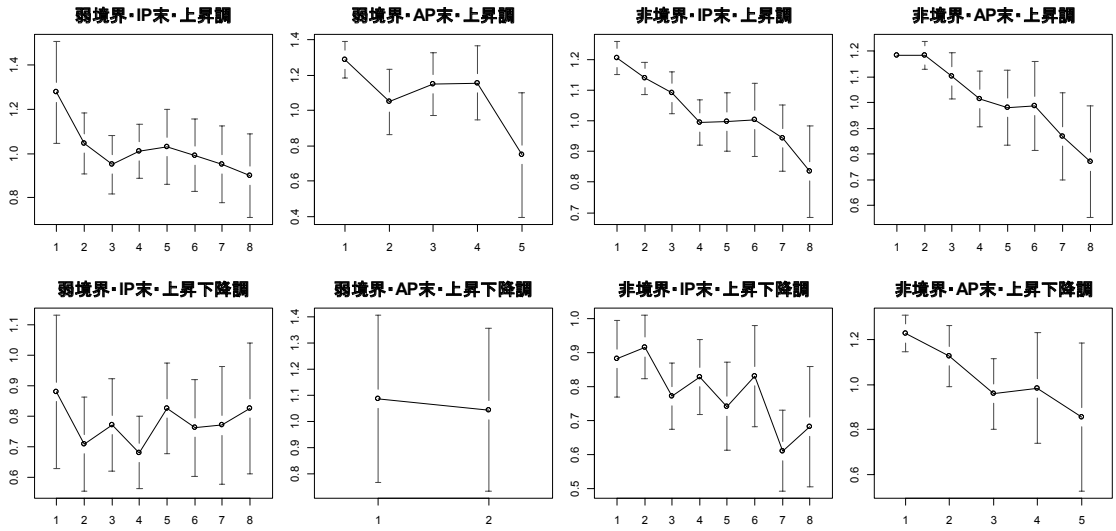


図 6　係り先の距離と上昇調・上昇下降調の上昇幅との関係

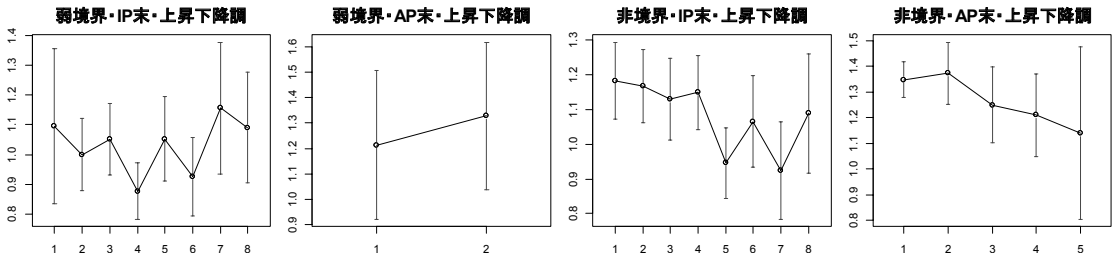


図 7　係り先の距離と上昇下降調の下降幅との関係

各距離が頻度 20 以上の結果に限定して図示)。非境界については，IP 末・AP 末の上昇調・上昇下降調の上昇幅および上昇下降調の下降幅に，係り先の距離が遠くなるほど上昇幅が徐々に小さくなる傾向が見られる。この傾向は特に上昇調の上昇幅にきれいに見られる。前節の分析から，係り先の距離が遠くなるほど上昇調，上昇下降調の頻度は増すが，BPM の大きさは逆に小さくなるという傾向である。

4. 考察

3.1 節の分析から，「非節境界＜弱境界＜強境界」のように統語的・意味的な切れ目が大きくなるほど，上昇調，上昇下降調ともに出現率が高くなる傾向が見られること，またこの傾向が，学会講演・模擬講演といったレジスターの違いに関わらず[*1]，また BPM の出現する位置である IP 末と AP 末の違い（ピッチレンジのリセットの有無の違いによる韻律境界の大きさの違い）に関わらず[*2]観察されることが分かった。以上の結果は，上昇調，上昇下降調が統語的・

[*1] 細かくみると，学会講演では，節単位の種類や IP 末・AP 末の違いを問わず上昇調が多く見られることが分かる。学会講演で上昇調が多いことは前川（2010）の報告にある通りである。

[*2] いずれの節クラスにおいても，相対的に強い韻律境界である IP 末の方が上昇調，上昇下降調の出現率は高い傾向にあることが分かる。これは，BPM がもともと強い韻律境界である IP 境界に出現しやすいことによる。

意味的に大きな切れ目に出現しやすいことを意味する。特に強境界では，BPM が極めて高い確率（模擬講演：75%，学会講演：87%）で生じている。小磯（2012）の結果では対話でも同程度に生じており，独話・対話の種類を問わず観察される傾向のようである。

その一方で，相対的により弱い統語境界である弱境界や非境界でも少なからず BPM が出現しており（弱境界：模擬 38% 学会 52%，非境界：模擬 14% 学会 20%），この位置で何故生じるのかという疑問が生じる。この点については，係り先の距離の結果が一つの回答を与えると考えられる。弱境界，非境界では，係り先の距離が遠くなるほど，上昇調，上昇下降調ともに出現率が高くなる傾向が見られた。同じ弱境界あるいは非境界であっても，「**被験者に**／提示した」のように直後の文節に係る場合と比べ，「**これらの音を**／それぞれ／三回ずつ／被験者に／ 提示した」のようにより先の文節に係る場合の方が，より統語的な切れ目は大きくなる。弱境界や非境界の BPM は，係り先の距離が 2 以上の統語的により大きな切れ目に出現しやすいということである[*3]。

以上の結果は，これらの音調が統語的・意味的に大きな切れ目に出現しやすいことを示しはするものの，大きな切れ目で「まだ発話（文）が継続する」ことを表示する機能があることを保証するものではない。ここで次の二つの点に着目したい。

一つは文末に相当する絶対境界（文末相当）における BPM の出現率の低さである。仮にこれらの音調が，統語的・意味的に大きな切れ目に出現しやすいだけで発言継続表示機能がないとするならば，絶対境界でも BPM は頻出する可能性がある。しかし，絶対境界における BPM の出現率は決して高くない（模擬講演：上昇調 20.7% 上昇下降調 1.3%，学会講演：上昇調 7.0% 上昇下降調 0.2%）。上昇調の結果に疑問上昇調が含まれていることを考えれば，強境界と比べて格段に低いと言ってよいだろう。

もう一つは挿入節，引用節，連体節の結果である。「が節」や「けれども節」など，もともと強境界である位置（ただし挿入節，引用節，連体節内のため非節境界と認定された位置）を対象に BPM の出現傾向を見たところ，強境界と同程度に BPM が出現することが示された。引用節の事例「{苦労してる けども 頑張ろうよ} っていうことでね」を考えると，「けども節」は引用のスコープにあるが，BPM によってこの位置に大きな切れ目があること**だけ**が表示されたのでは解釈に支障をきたしうる（下記左参照）。これらの音調に継続性の表示機能があると考える方が自然である（下記右参照）。

大きな切れ目だけをマーク	継続性をマーク
～けれども／…する	{ ～けれども ＋ …する }
～けれども／…する ということだ	{ ～けれども ＋ …する } ということだ

勿論，強境界と同程度に BPM が出現するということは，仮に BPM に発言継続表示機能を認めたとしても引用内か否かの区別にはならないことは上記右の模式図からも明らかである。しかし少なくとも大きな切れ目**だけ**をマークするという仮説の矛盾は指摘できよう。

[*3] 非境界で係り先の距離が 1 という，統語的な切れ目がより弱く，かつ韻律境界としても弱い AP 末にだけ，特に BPM が多く出現するという傾向も見られる。前川（2011）は，AP 末に BPM が存在しながらも，ピッチレンジのリセットが生じず IP 末とならない発話が学会講演に多いことを指摘し，学会講演の一部の話者が，講演前に周到な練習を重ね発話すべき内容を大部分暗記していることがこの特殊なパターンの頻出要因となっている可能性があるとしている。この影響が，特にこの条件に顕著に現れている可能性もある。今後の検討課題とする。

最後に，上昇調・上昇下降調の大きさと統語的・意味的な切れ目の強さとの関係について言及する。節境界クラスと係り先の距離という二つの観点から，統語的・意味的な切れ目の強さと上昇調・上昇下降調の大きさとの関係を見たが，少なくとも統語的・意味的な切れ目が大きくなるほど上昇調・上昇下降調の上昇幅・下降幅が大きくなる（より強調して発話される）という関係は見られず，むしろその逆の傾向が部分的に観察されるに留まった。BPM の振幅は，統語的・意味的な切れ目の強さと単純に相関するのではなく，その他の要因も関わっている可能性がある。この点については今後の検討課題とする。

付記：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）および基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）による成果である。

参 考 文 献

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.

Igarashi, Y. & H. Koiso （2012） “Pitch range control of Japanese boundary pitch movements”, *Proc. of InterSpeech 2012*.

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」『日本語話し言葉コーパス』DVD 付属マニュアル.
http://www.ninjal.ac.jp/csj/manu-f/dependency.pdf

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.

小磯花絵（2012） 「独話と対話における句末音調の比較－『日本語話し言葉コーパス』を用いて－」『社会言語科学会第 30 回大会発表論文集』.

郡史郎（1996） 「音声の特徴から見た文」『日本語学』15 (9) pp.60–70.

郡史郎 (2003) 「イントネーション」『朝倉日本語講座 音声 3 音声・音韻』（上野編）朝倉書店, pp.109–131.

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（2011） 『コーパスを利用した自発音声の研究』（博士論文）.

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

文の長さ分布から見た文生成のメカニズム

古橋　翔（東北大学大学院理学研究科）†

Mechanism of Writing Sentences Based on the Distribution of Sentence Lengths

Sho Furuhashi (Graduate School of Science, Tohoku University)

１．はじめに

言語学の一分野である計量文献学では、文を構成する文字や単語を数え、最頻値、平均値や分布型などから作者の文体を特徴づけてきた。

日本語における文の長さ（文長）に関する研究に、長さの単位を文字とし、文長分布が対数正規分布であると示した安本(1958)と新井(2001)、対数正規分布とガンマ分布であると報告した佐々木(1976)の先行研究がある。一方で、長さの単位を形態素とし、文長分布が Hyper Pascal 分布と報告した石井と石井(2007)の研究がある。

現象の仕組みを理解する上で、実験や観測により得られたデータの分布型を再現するモデルを考えることは、重要である。上記の先行研究においても文長分布の生成モデルが挙げられている。佐々木は、対数正規性を生む一例として Kaptyen のアナログマシンを例に Multiplicative な確率過程を挙げ、ガンマ分布に対しては、文の構成要素の長さが指数分布に従うから、文の構成要素が合わさった文長はガンマ分布に従うのではないかと考察している。石田らは、G. Altmann (1988)の Hyper Pascal 分布の生成モデルを挙げている。

本研究では、先行研究で挙げられた日本語の文長分布型から、日本語文の生成メカニズムを解明しようと試みた。研究に当たり、佐々木が提示したガンマ分布の生成モデルに注目する。佐々木は考察の中で、

「述べようとする『思想』あるいは『事柄』をこまかく分割したとき、その分布は指数分布に従う。そして、文の長さは、その指数分布に従うものが、いくつかくっついたものと考えることができる。従って、文の長さの分布は、『ガンマ分布』に従う。」

と述べている。また、くっつくべき個数 m を定数と考えることに若干の問題があるとし、m がある分布に従うとするならば、「複合分布」を考えなければならないと考察している。

本研究では、「述べようとする『思想』あるいは『事柄』」が文を構成する句や節に対応すると考え、佐々木の考察により文長の分布型が説明できるかどうか調べた。

２．文長の分布型

先行研究で取り上げられた分布型を説明する。

２.１対数正規分布

対数正規分布は、

$$f_{\mathrm{LN}}(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}(x>0),$$

で定義される連続型確率分布である。パラメータ μ は $\ln x$ の平均値で、σ^2 は $\ln x$ の分散で

† furuhashi@cmpt.phys.tohoku.ac.jp

ある。変数 x の対数 $\ln x$ が正規分布に従うので、$f_{\mathrm{LN}}(x)$ は長い裾をもつ分布である。

２.２ガンマ分布

ガンマ分布は、

$$f_G(x) = \frac{x^{k-1}}{\Gamma(k)\theta^k}\exp(-\frac{x}{\theta})(x \geq 0),$$
$$= 0(x < 0),$$

で定義される連続型確率分布である。パラメータは、形状母数 k と尺度母数 θ の二つである。特徴として、長い裾が挙げられる。確率変数 $X_1, X_2, \cdots, X_\alpha$ が指数分布 $\beta^{-1}\exp(-x/\beta)$ に従うならば、$X_1 + X_2 + \cdots + X_\alpha$ はガンマ分布（$k = \alpha$ 、$\theta = \beta$）に従う。

３．文の構造

日本語の文構造は、文節間の係り受け関係を表した依存構造木で表現できる。依存構造木は、文節をノードとして係り元から係り先へ矢印を張り構築され、係り受けは一般的に非循環性が仮定されているので、文全体の係り受け関係はツリーとなる。図１は「友人の太郎は次郎が持っている本を花子に渡した」という文の依存構造木である。依存構造木のルート「渡した」に対して、「友人の太郎は」は動作主を表す句、「次郎が持っている本を」は動作の対象を表す句、そして「花子に」は動作の方向を表す句である。従って、リーフからルートの子ノードまでの部分は、ルートに対する主格や対格等に対応する句である。本研究では、このようなノードの集合を枝と呼ぶとする。

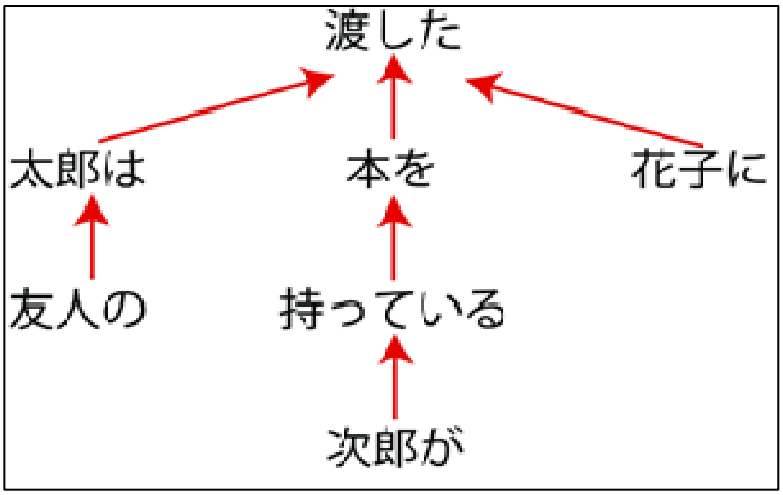


図 1 依存構造木。リンクの向きは、係り受けの向きを表す。

４調査方法とサンプル

本研究の調査方法と用いたサンプルを説明する。

４.１調査方法

本研究は、枝に着目して依存構造木を統計的に解析する。依存構造木の構成単位は文節なので、長さの単位は文節とする。まず始めに、サンプルの文長の分布型を調べる。分布型は、先行研究で挙げられていたガンマ分布と対数正規分布に限定し、赤池情報量規準（AIC）で判断する。次に、一文当たりの枝の数の分布をみる。枝の数が一定であればガンマ分布の生成モデルに当てはまるが、実際は佐々木が指摘したように一定ではないと思わ

れる。さらに、枝の長さ分布を調べ、指数分布となっているか確かめる。そして、一文中の枝の長さに相関があるかどうか調べ、各枝の独立性を確認する。

４.２サンプル

本研究では、京都大学大学院情報学研究科黒橋・河原研究室が提供している京都大学テキストコーパス Version 4.0 と、現代日本語書き言葉均衡コーパス（BCCWJ）を用いた。

京都大学テキストコーパスは、Web ページからダウンロードできるパッケージを用いた。パッケージに含まれているのは、形態素・構文・関係の付加情報のみであり、テキストは含まれていない。京都コーパスを本来の形に変換するためには、毎日新聞 1995 年版 CD-ROM が必要である。しかしながら、本研究で必要なのは係り受け情報なので、毎日新聞 1995 年版 CD-ROM は使用しなかった。総文数は 38397 である。

```
<?xml version="1.0" encoding="UTF-16"?>
<sample sampleID="OW1X_00000" version="20070814" type="variableLength">
<article articleID="OW1X_00000_V001" isWholeArticle="false">
<titleBlock>
<title><sentence type="quasi"> 第２節　内外均衡の背景 </sentence><br type="automatic_original"/></title>
</titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは、それぞれがバラバラに生じてきたわ
けではない。　</sentence><sentence> 以下では、それらの動きの重要な背景として、・・・
</paragraph>
```

図 2　BCCWJ の XML

BCCWJ は、C－XML フォルダ下にあるサブコーパス LB（図書館サブコーパス「書籍」）、OB（特定目的サブコーパス「ベストセラー」）、PB（出版サブコーパス「書籍」）と PN（出版サブコーパス「新聞」）に収蔵されている XML ファイルを用いた。

XML ファイルから文を取り出す方法を示す。まず、XML ファイル中の記事（article タグで挟まれた部分）を、BCCWJ に添付されている記事情報データに基づき実著者の人名 ID ごとに分ける。記事の中には別の ID を持つ記事が含まれているものがあるが、このような記事は除外する。次に、実著者ごとに振り分けた記事から、文として sentence タグで挟まれた部分を、その sentence タグの階層情報とともに取り出す。sentence タグの階層情報とは、article タグから sentence タグに至るまでに通るタグの集合である。例えば、図 2 では、「第二節　内外均衡の背景」のタグの階層情報は “article, titleBlock, title, sentence”、「５３年度中にみられた内外均衡回復に向けての動きは、それぞれがバラバラに生じてきたわけではない。」のタグの階層情報は “article, paragraph, sentence” となる。一部の文は中に別の sentence タグを含んでいる。その場合は、一番外側の sentence タグに従う。また、speech タグが閉じた直後には「と、言った。」が多いが、本研究ではこれを文とは認めない。よって、speech タグ直後の sentence タグは除外した。

更に、取り出した文を以下の条件に従い選別する。

(ア) 階層情報が article, paragraph と sentence タグのみで構成されている。
(イ) sentence タグで挟まれた部分に、ruby と sampling 以外のタグが含まれていない。
(ウ) sentence タグの種類が quasi ではない。
(エ)　▼◇〒@※◆■…＝などの記号が含まれない。
(オ) ”)”、”＜” が含まれない。

(ア) は、会話文、箇条書きやキャプション等を除いた地の文を対象とするため、(イ) は選出基準を簡潔にするため、(ウ) は、文に準ずるものではなく、きちんとした文を対象とす

るため、(エ)と(オ)は、係り受け解析の精度を高めるためである。

係り受け解析には、形態素解析に MeCab 0.98（辞書は MeCab-Ipadic）を用いた日本語係り受け解析器 CaboCha 0.60 pre4（TinySVM と YamCha なし）を使用した。

５．結果

京都大学テキストコーパスから得られた結果を示す。

５.１文長分布

文長分布を図３に示す。文長の平均は 9.7 で標準偏差は 5.3 である。最尤法により推定したパラメータ値は $k = 3.45$，$\theta = 2.81$，$\mu = 2.12$，$\sigma^2 = 0.34$ である。また、AIC は、対数正規分布の場合 229744、ガンマ分布の場合 227865 であり、ほぼ同じ値であり、どちらか一方と断言できない。

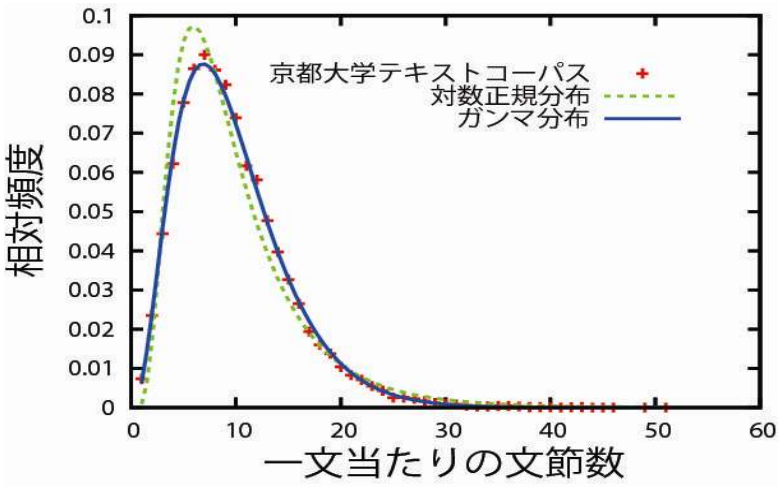


図 3　文長分布

５.２枝の数分布

図４は枝の数分布である。

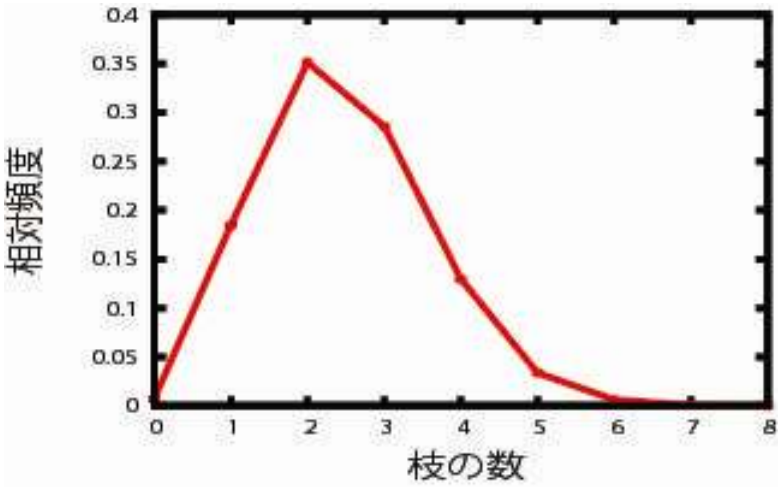


図 4　枝の数分布

平均 2.48、分散 1.22、最頻値は 2 である。ガンマ分布の生成モデルのような定数に枝の数はならないが、狭い範囲に集中している。

５.３枝の長さ分布

枝の長さの累積分布を図５に示す。図４より、依存構造木の枝の数は一定ではないので、枝の数ごとに分布をとった。

分布をみると、枝の数が２の場合全体的に指数分布に従うが、それ以外は枝が短い領域で指数分布からずれている。枝の数が１の分布は、枝の数が２の分布より上、枝の数が３以上の分布は、枝の数が２の分布より下にあり、累積分布では長さ１で相対頻度の累積が必ず１になる点を考慮すると、枝の数が多くなるにつれて、短い枝の割合が大きくなることがわかる。このような変形はあるものの、指数分布のような分布型がみられる。

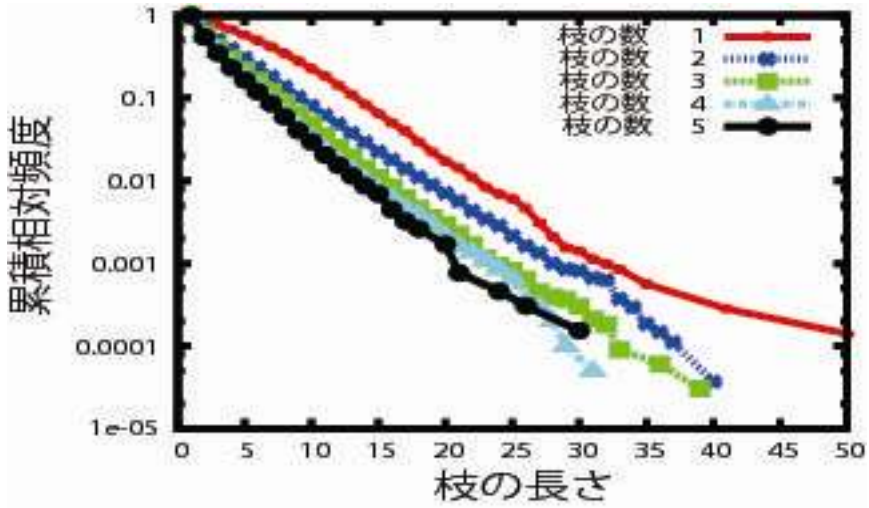


図 5　依存構造木を構成する枝の長さの累積分布。
依存構造木に含まれる枝の数ごとにプロットした。

５.４枝の長さ相関

ガンマ分布の生成モデルでは、確率変数は互いに独立である。文を構成する枝の長さが独立（無相関）かどうかピアソンの相関係数により調べた。

ピアソンの相関係数は、次のようにして求める。まず各文の枝をラベル付けする。具体的には、枝に含まれるルートの子ノードに対して文頭からの出現順位付けを行った。例えば図１の依存構造木の場合、ルート「渡した」の子ノード「太郎は」、「本を」、「花子に」はこの順番に文中に現れるので、一番目の枝は「友人の太郎は」、二番目の枝は「次郎が持っている本を」、三番目の枝は「花子に」である。文 k の枝の数を b_k 、 j 番目の枝の長さを s_{kj} とし、枝の数 x の文に対する i 番目と j 番目の枝の長さの相関を表すピアソンの相関係数 $r_{ij}(x)$ は、

$$r_{ij}(x)=\frac{\sum_k (s_{ki}-\overline{s_i(x)})(s_{kj}-\overline{s_j(x)})\delta_{b_k,x}}{\sqrt{\left(\sum_k (s_{ki}-\overline{s_i(x)})^2\delta_{b_k,x}\right)\left(\sum_k (s_{kj}-\overline{s_j(x)})^2\delta_{b_k,x}\right)}},$$

となる。但し、

$$\overline{s_i(x)} = \frac{\sum_k s_{ki}\delta_{b_k,x}}{\sum_k \delta_{b_k,x}},$$

である。相関係数は枝の数で区別して計算した。理由は枝の長さが枝の数に依存する可能性を排除するためである。表 1 にピアソンの相関係数を示す。相関係数の絶対値は 0.1 程度なので、長さに相関は見られない

表 1　ピアソンの相関係数。左から枝の数が 2（文の数 13487)，3（文の数 10972)、4（文の数 4982）である。

	i=1	i=2
j=1	1.0	-0.15
j=2		1.0

	i=1	i=2	i=3
j=1	1.0	-0.12	-0.099
j=2		1.0	-0.049
j=3			1.0

	i=1	i=2	i=3	i=4
j=1	1.0	-0.11	-0.10	-0.074
j=2		1.0	-0.046	-0.044
j=3			1.0	0.11
j=4				1.0

６．作者別に見た場合

京都大学テキストコーパスで得られた結果が他のコーパスで得られるか確かめるために、BCCWJ を用いて同様の解析をした。PN から、毎日新聞社(人名 ID: 258099、総文数: 1323)、朝日新聞社(人名 ID: 256908、総文数: 1798)、読売新聞東京本社(人名 ID: 259670、総文数: 1871)と中日新聞社(人名 ID: 263664、総文数: 1234)、LB、OB と PB から、赤川　次郎(人名 ID: 873、総文数: 3410)、 森村　誠一(人名 ID: 47302、総文数: 3150)、西村　京太郎(人名 ID: 55252、総文数: 1468)と司馬　遼太郎(人名 ID: 70104、総文数: 4532)を対象とした。
以下に、文長分布、枝の数分布、枝の長さの累積分布を示す。枝のサイズ相関はまだ調べていないので結果を示せない。

６．１文長分布

図６に、新聞社と作家それぞれの文長分布を示す。ともに、京都大学テキストコーパス（図３）同様に裾が伸びた分布をしている。これらの分布が、対数正規分布とガンマ分布どちらに当てはまるかを AIC により評価する。表２は、最尤法で推定した対数正規分布とガンマ分布のパラメータ値であり、表３は AIC の値である。

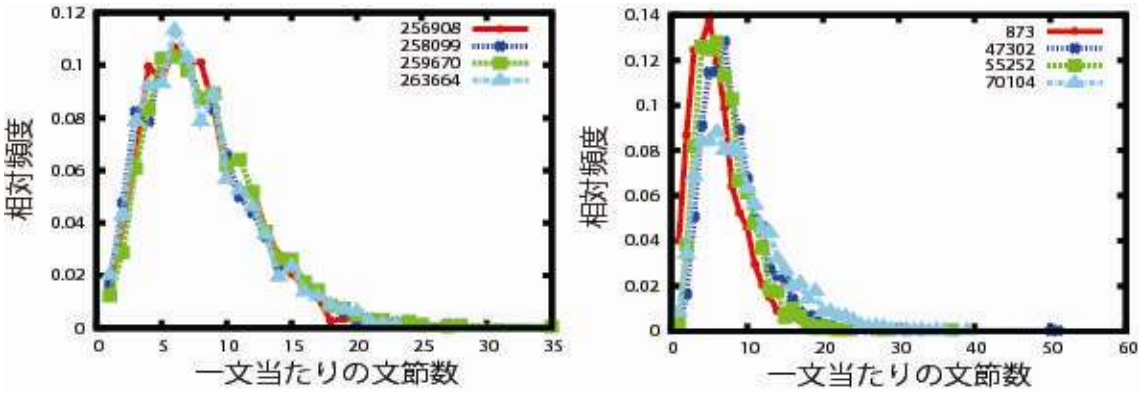


図 6　文長分布。左のグラフが新聞社。右のグラフが作家。

表 2　最尤法によるパラメータ値

	毎日新聞社	朝日新聞社	読売新聞東京本社	中日新聞社	赤川次郎	森村誠一	西村京太郎	内田康夫
k	3.29	3.62	3.60	3.29	3.02	4.41	4.30	4.90
θ	2.38	2.17	2.32	2.36	1.99	1.85	1.65	1.72
μ	1.90	1.92	1.98	1.89	1.62	1.98	1.84	2.02
σ^2	0.361	0.327	0.324	0.364	0.386	0.241	0.253	0.220

表 3　AIC の値

	毎日新聞社	朝日新聞社	読売新聞東京本社	中日新聞社	赤川次郎	森村誠一	西村京太郎	内田康夫
ガンマ分布	7342	9844	10484	6828	17334	17000	7544	7146
対数正規分布	7438	9985	10602	7438	17483	16956	7556	7158

最尤法で推定したパラメータ値は、京都大学テキストコーパスの値と大きな違いは見られない。また、AIC の値をみると、ガンマ分布と対数正規分布間に大きな違いは見られない。したがって、本研究で用いる新聞社と作家の文長分布は、京都大学テキストコーパスの文長分布と似た特徴を有している。

６．２枝の数分布

図７は枝の数分布である。図４の京都大学テキストコーパスと比較すると、全体的に同じ分布型をしている。

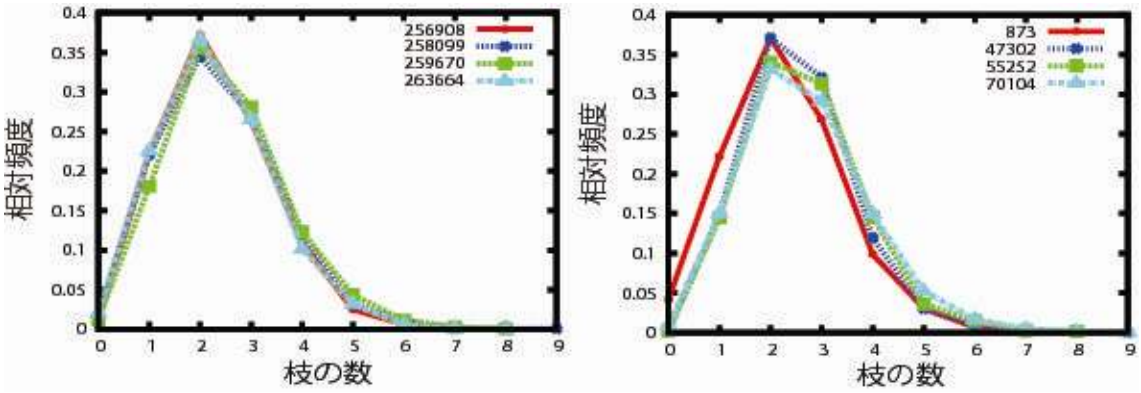


図 7　枝の数分布。左のグラフが新聞社。右のグラフが作家。

６．３ 枝の長さ分布

新聞社から読売新聞東京本社（人名 ID259670）、作家から森村誠一（人名 ID 47302）を選んで、枝の長さ分布をとった結果を図８に示す。京都大学テキストコーパス（図５）と同じ特徴を持った分布型をしている。

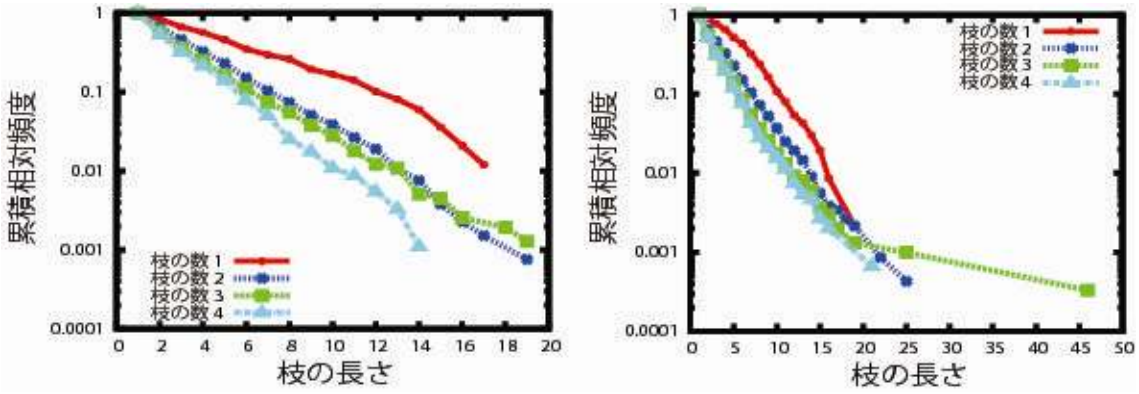


図 8　文を構成する枝の長さ分布。各文の依存構造木に含まれる枝の数ごとにプロット。左のグラフが読売新聞東京本社（人名 ID259670）、右のグラフが森村誠一（人名 ID47302）。

以上から、小説家や他の新聞社でも、京都大学テキストコーパスで得られた結果と非常に類似している。よって、文生成に対する佐々木の考察は一般的に当てはまると推測できる。

7．まとめ

本研究では、係り受け関係と文長に注目して、日本語文の生成メカニズムを調べた。依存構造木のルートに該当する文節に係る句や節の長さが、互いに独立で指数分布に従うことが分かった。ただし、句や節の数は一定ではないので、単純なガンマ分布ではなく複合分布となる。この結果は、佐々木の考察と一致するものであった。

今後の方針として、日本語以外で依存文法に従い同様の解析を行い、文生成における共通点と相違点を明らかにすることが挙げられる。依存文法による解析情報をもつコーパスは多くの言語で公開されている。しかしながら、依存文法に基づいた文構造がツリーにならない、サンプルの量が少ないなど問題がある。このような問題をいかに克服するかが、まず取り組むべき課題である。

文　献

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

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

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

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

Gabriel Altmann (1988) Verteilungen der Satzlängen. *Glottometrika*, 9, pp. 147-170.

関連 URL

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

MeCab　http://mecab.sourceforge.net/

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

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

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）

Prosodic Changes of Intonation Phrases at Syntactic Boundaries in the Corpus of 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

Pierrehumbert & Beckman（1988）は，アクセント句（以下 AP）やイントネーション句（以下 IP）より上位に「発話（Utterance）」という単位を設定している。発話は，F0 declination（発話に要する時間の関数として単純に F0 が低下する現象）が見られる範囲であり，その末尾で final lowering（平叙文末尾で F0 が局所的に下降し発話の終了を示す現象）が生じるとされる。しかし，自発音声の発話の特徴を分析した研究は，Maekawa（2010）などごく少数に限られ，その実態はまだ明らかにはなっていない。

小磯・石本（2012）は，自発音声の発話の特徴を探るため，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を対象に，IP を単位として F0 の変動を調べ，次の傾向が見られることを指摘した。

1. 発話中に強い節境界が存在しない場合（一つの節で発話が終了する場合）
   （a）IP 全体の最大値・最小値が発話内で徐々に下降する
   （b）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）
   （c）ただし発話が長い場合，若干高い F0 で発話を開始する
   （d）発話末に final lowering に相当する著しい下降が見られる
2. 発話中に強い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）
   （a）発話中の節境界で IP の下降傾向はリセットされる
   （b）リセット時，IP の最小値は発話末のレベルにまで達しない
   （c）final lowering に相当する著しい下降も見られない
   （d）リセット後，IP の最大値は発話冒頭のレベルに戻ることもあれば戻らないこともある

この結果は，Pierrehumbert and Beckman の指摘する「発話－ IP － AP」という韻律的階層構造に対し，発話と IP の間に別の単位が存在する可能性を示唆するものであり，「発話」とダ

† ishimoto@nii.ac.jp

ウンステップの生じる領域である「Major phrase（以下 MaP)」の間に「Intonational Phrase（以下 IntP，上記 IP と異なる点に注意)」を設ける説を支持するものと考えられる。

日本語の IntP については十分な検討がなされておらずその存在も十分に検証されていないが，そのような中 Kawahara and Shinya（2008）は，読み上げ音声を対象とした分析に基づき，句頭の initial rise の幅が「MaP ＜ IntP（に相当すると想定する節）＜発話」の順に大きくなること，また final lowering が発話と IntP にのみ観察され，かつその程度が「IntP ＜発話」の順に大きくなることなどから，日本語にも発話と MaP の間に IntP が存在する可能性があることを指摘している。

そこで本研究では，Pierrehumbert and Beckman の枠組みで言うところの IP と発話の間にもう一階層あると仮定し（暫定的に上記略語を用いて IntP と称す)，その存在の可否について検討する。特に Kawahara らの分析を参考に，各階層の単位の冒頭あるいは末尾の F0 の特徴に着目して分析を行い，Kawahara らが読み上げ発話を元にその存在を示した IntP が自発発話においても観察されるかを検討する。

2. データ

2.1 発話単位

分析に用いた『日本語話し言葉コーパス（*Corpus of Spontaneous Japanese*:以下 CSJ)』（前川 2004）は自発性の高いモノローグを中心に構成された話し言葉コーパスであり，学会における口頭発表（以下「学会講演」）と，一般話者による主に個人的な内容に関するスピーチ（以下「模擬講演」）を主対象としている。CSJ 全体は 661 時間の音声から構成されるが，本研究ではこのうち「コア」と呼ばれるデータ範囲の中から学会講演 70（約 29 時間）・模擬講演 107（約 20 時間）を分析対象とした。実際の分析には CSJ 第 3 刷に基づき作成された RDB（小磯ほか 2012）を用いる。

発話に相当する単位として，CSJ に付与されている節単位情報を利用した。節単位情報は原則「節（clause)」の境界によって得られる文法的・意味的なまとまりを持った単位であり，CSJ において構文・談話レベルの情報を付与するための基本単位として設計されたものである（丸山ほか 2006)。節単位は，節境界の構造的な切れ目の大きさの観点から以下の 3 つに分類される。

絶対境界（Absolute boundary）　いわゆる文末に相当する境界。明示的な文末表現の直後。
強境界（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る。

本研究では，発話に相当する単位を絶対境界によって区切られる区間とする。また，さしあたり IntP が強境界で区切られる単位と強い関連を持つと仮定した上で，強境界で現れる特徴を絶対境界（発話境界）や非節境界（IP 境界）と比較することで，IntP の存在について検討する。なお，話し言葉では強境界にあたる「〜けれども」「〜が」などの接続表現に final lowering などが生じて発話の終わりとなることもあるが，本研究の分析対象である講演のような発話では

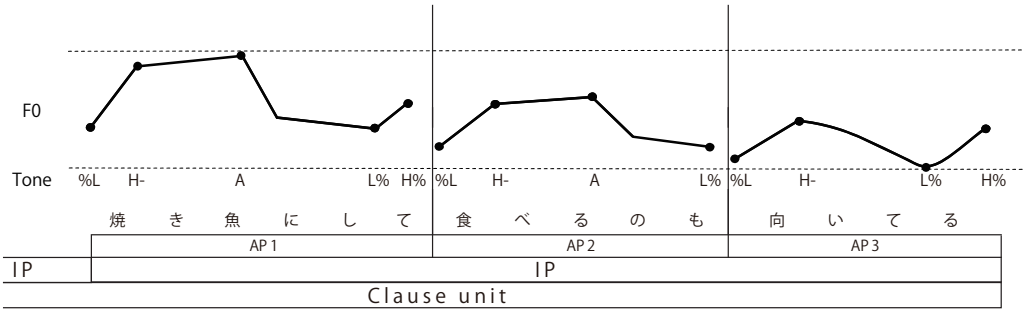


図 1　IP と AP の構造と Tone 情報

少数であることから統計的に誤差として扱われる。

2.2　イントネーション句とアクセント句

本研究では IP を単位に発話の韻律的特徴を探る。IP は，アクセント核が後続するアクセント句（AP）の F0 ピークを反復的に低下させるダウンステップの生じる領域と定義され，IP 境界でピッチレンジのリセットが生じる。節単位内にはひとつ以上の IP が含まれ，IP 内にはひとつ以上の AP が含まれることになる。IP と AP の関係を図 1 に示す。

CSJ にはラベリングスキーム X-JToBI（五十嵐ほか 2006）に基づき韻律情報が付与されているが，この中に，Break Index（BI）という，韻律境界の切れ目の強さに関する情報が存在する。BI=2 は AP 境界，BI=3 は IP 境界，BI=F はフィラー境界，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，X-JToBI に基づいた Tone 情報から，各分析において IP の特徴としての F0 値を求める。

3.　分析 1

3.1　方法

本節では，図 2 に示すように，発話内（絶対境界内）に強い統語境界である強境界を二つもつ場合，つまり発話に三つの節単位（CU）が存在する場合を対象に，発話の冒頭（CU1 の冒頭）と節単位の冒頭（CU2・CU3 の冒頭），発話の末尾（CU3 の末尾）と節単位の末尾（CU1・CU2 の末尾）の F0 の特徴をそれぞれ比較する。

この分析は，発話と IntP の冒頭・末尾の F0 変動の比較を視野に入れたものである。1 節で言及したように，Kawahara らは，句頭の initial rise の幅が「MaP ＜ IntP ＜発話」の順に大きくなること，また final lowering が発話と IntP にのみ観察され，かつその程度が「IntP ＜発話」の順に大きくなることを指摘している。そこで，発話と節単位の冒頭・末尾に上記傾向が見られるかを中心に検討する。

対象となる発話は 322 発話であった。求めた F0 は

F0 開始値　節単位頭の IP の最初の AP の句頭境界音調（%L）の F0 値

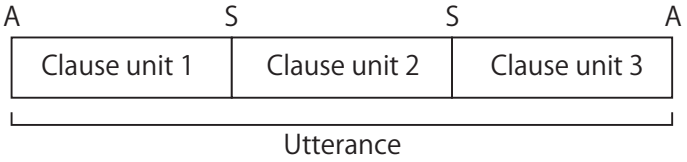


図 2　分析 1 の対象となる発話の構造（絶対境界:A，強境界:S）

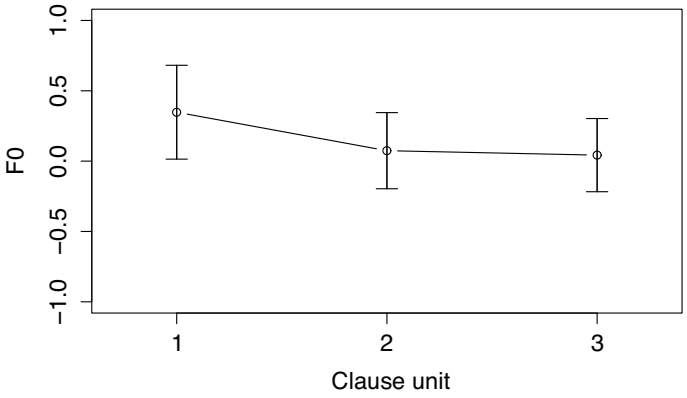


図 3　F0 開始値

F0 最大値　節単位頭の IP の最初の 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

initial rise　F0 最大値と F0 開始値の差

F0 最小値　節単位末尾の AP の下降音調（L%）の F0 値

である。ただし，発話末は無声化していることが多く，発話末の F0 最小値が抽出できたのは 64 発話であった。

3.2　結果と考察

各節単位の F0 開始値を図 3 に示す。この F0 開始値に対して分散分析を行ったところ，節単位による違いが有意であった（$F(2,315)=46.46$, $p< .01$）。Tukey 法による多重比較を行った結果，CU1 の F0 開始値が CU2，CU3 よりも有意に大きかった。すなわち，発話の開始において高い F0 から話し始めているといえる。

次に各節単位の最初の IP の F0 最大値を図 4 に示す。この F0 最大値に対して分散分析を行ったところ，節単位による違いが有意であった（$F(2,213)=7.54$, $p< .01$）。Tukey 法による多重比較を行った結果，CU1 の F0 最大値が CU2，CU3 よりも有意に大きかった。すなわち，発話の開始では F0 開始値と同様に句頭の上昇あるいはアクセントの F0 も高くなることが示された。

上記の F0 開始値と F0 最大値から求めた initial rise を図 5 に示す。initial rise に対して分散分析を行ったところ，1% 水準で有意とはならなかった。すなわち，Kawahara らが示したような発話冒頭の initial rise の拡大が本データではみられなかった。

最後に，各節単位末の IP の F0 最小値を図 6 に示す。この F0 最小値に対して分散分析を行ったところ，節単位による違いが有意であった（$F(2,83)=31.07$, $p< .01$）。Tukey 法による多

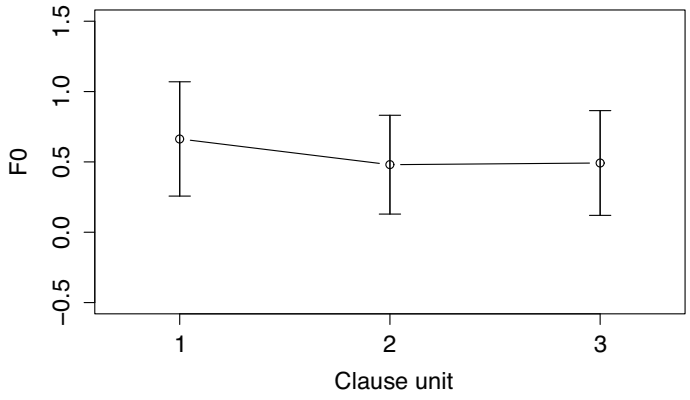


図 4　F0 最大値

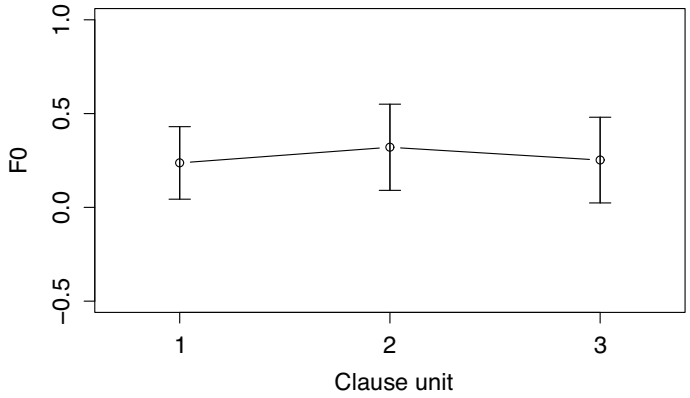


図 5　節単位頭の initial rise

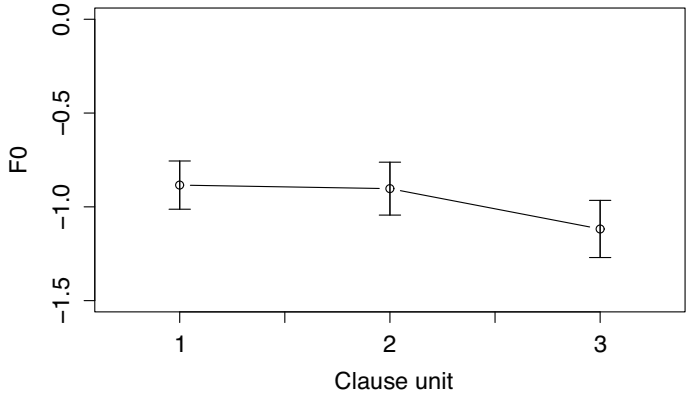


図 6　節単位末の F0 最小値

重比較を行った結果，最後の節単位 CU3 の F0 最小値が，CU1，CU2 よりも有意に小さかった。すなわち，強い統語境界直前に比べて発話末の F0 が大きく低下していることを示している。

以上をまとめると，(1) 発話冒頭では，発話中の節単位冒頭と比べ，AP がより高い位置で開始し，また AP の最大値も高い，(2) この傾向が同程度に生じるため，Kawahara らの指摘する initial rise の傾向は観察されない，(3) 発話の末尾で final lowering とみられる強い F0 の

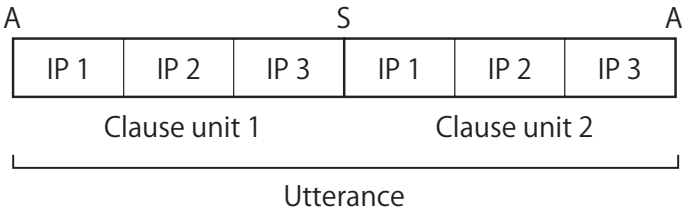


図 7　分析 2 の対象となる発話の構造（絶対境界:A，強境界:S）

下降が見られるが，発話中の節単位末尾には観察されない，となる。

節単位に含まれる IP は，節単位冒頭から末尾にかけて IP の F0 最大値・最小値ともに下降し，発話末あるいは発話中の強い節境界でその下降はリセットされる傾向にあるが（小磯・石本 2012），上述の結果は，発話中の強い統語境界では，発話末ほど F0 は下がりきることはなく，またリセット後も，F0 の高さが必ずしも発話の開始時と同程度まで戻らないことを意味する。ここで特徴的なことは，F0 最大値も F0 最小値も発話の冒頭もしくは末尾でのみ顕著な変化を示しているという点である。

この結果は Kawahara らの指摘とは必ずしも一致しないが，いずれにせよ発話と節単位の冒頭・末尾の特徴に違いがあるというものであり，IntP の存在を示唆する結果と言える。initial rise に関する Kawahara らの知見との相違は，発話頭における発話内容のバリエーションの多さに起因するとも考えられるが，詳細は今後の検討課題である。

4.　分析 2

4.1　方法

前節の分析で，発話の末尾にのみ final lowering が観察される可能性が示唆された。しかし前節の分析は発話・節単位末の F0 最小値を比較したに過ぎない。そこで本節ではこの点をより詳細に調べる。具体的には，図 7 に示すように，内部にひとつの強い統語境界をもつ発話を対象とし，それぞれ 3 つの IP をもつ節単位に限定して各 IP 末の F0 の違いをみる。条件を満たし，分析対象となったのは 70 発話であった。

求めた F0 は

F0 開始値　IP 内の最終 AP の句頭境界音調（%L）の F0 値
F0 最大値　IP 内の最終 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 内の最終 AP の下降音調（L%）の F0 値

である。これらは図 1 における AP3 の Tone 情報に等しい。

4.2　結果と考察

各 IP 末の AP の F0 開始値を図 8 に示す。IP 内の最終 AP の F0 開始値については，強境界前の節単位 CU1 内も，絶対境界前の CU2 も，同じように下降傾向を示しており，また F0 の値もほぼ同じである。

次に各 IP 末の AP の F0 最大値を図 9 に示す。強境界前の CU1 ではすべての IP に渡って緩やかに低下しているが，絶対境界直前の IP では F0 最大値が急激に小さくなっていることが分

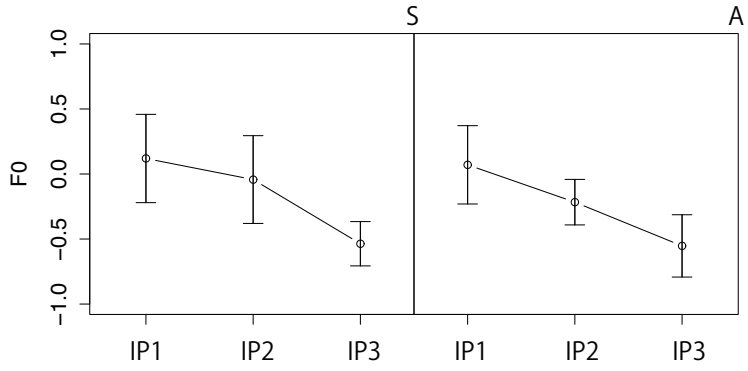


図 8　IP 内の最終 AP の F0 開始値

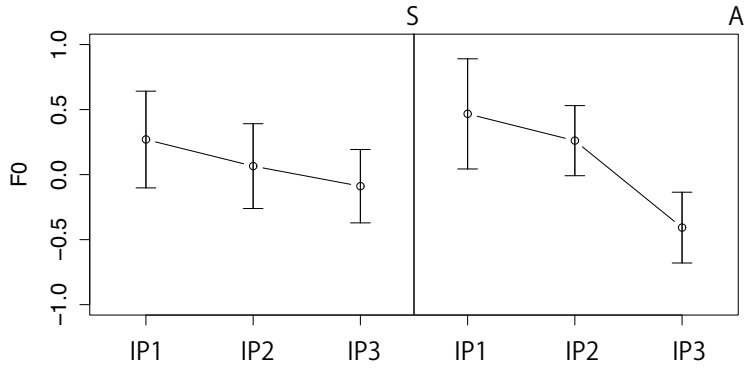


図 9　IP 内の最終 AP の F0 最大値

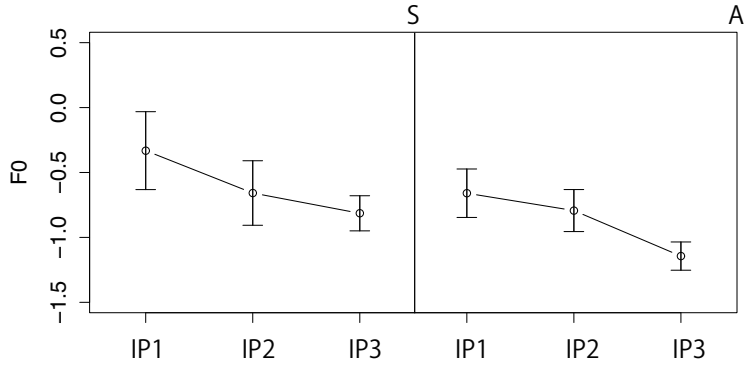


図 10　IP 内の最終 AP の F0 最小値

かる。

最後に各 IP 末の AP の F0 最小値を図 10 に示す。F0 最小値は発話内で全体的に発話末へ向かって緩やかに低下しており，発話末である絶対境界前で急激に低下する傾向が見られる。

絶対境界直前で最終 AP の最小値がその前に比べて急激に低下するのは発話末の final lowering の影響と考えられるが，AP の F0 最大値も急激に低下するということは，final lowering が最終 AP の末尾だけでなく AP の広い範囲に渡って見られることを意味するものであり，Maekawa（2010）の指摘と一致する。また最終 AP の F0 最大値の急激な低下は，Kawahara らの報告する発話末の動詞の F0 ピークの低下の結果とも整合的である。

一方，今回の分析でも強境界直前ではこのような急激な下降は見られず，最終 AP にわたる final lowering が発話末を特徴付ける要素であることが指摘できる。また，強い統語境界で F0 開始値，F0 最大値の下降が大きくリセットされていることから，このような統語境界の位置では IP よりも大きな韻律構造が存在することがうかがえる。以上の結果は，IP と発話の間に IntP のような中間的単位が介在することを示唆するものといえる。

5. おわりに

本稿では，小磯・石本（2012）の結果にもとづき，イントネーション句と発話の間に中間的韻律単位が介在し，かつこの韻律単位が強い統語境界で区切られる単位と強い関連を持つと仮定した上で，各階層の単位の冒頭あるいは末尾の F0 の特徴に着目して分析を行い，その存在の可否について検討した。その結果，分析 1 では，イントネーション句の F0 最大値・F0 最小値が発話の冒頭もしくは末尾でのみ顕著な変化を示し強い統語境界では変化は認められないことがわかった。また分析 2 では，発話中の強い統語境界で F0 開始値・F0 最大値の下降が大きくリセットされること，発話末に対応する絶対境界直前のアクセント句では F0 最大値・F0 最小値の急激な低下がみられるが，発話中の強境界前では急激な低下がみられないことがわかった。これらは，「発話－イントネーション句－アクセント句」という韻律的階層構造に対し，発話とイントネーション句の間に中間的韻律単位が存在することを示唆している。

参 考 文 献

Pierrehumbert, Janet B. and Mary E. Beckman(1988) *Japanese tone structure*, Cambridge: MIT Press.

Maekawa, Kikuo (2010) “Final lowering and boundary pitch movements in spontaneous Japanese,” Proc. DiSS-LPSS Joint Workshop, pp. 47–50.

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.

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。

Praat 起動用 Excel アドイン “Praat Launcher”

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）†

Praat Launcher: An Excel Addin for “doing phonetics by computer”

Ken’ya Nishikawa (Lab. for Language Development, RIKEN Brain Science Institute)

1. はじめに

Praat (Boersma and Weenink 2012) は、豊富な機能を持つ音声分析用ソフトウェアである。最近では、少量の音声データの音響分析にとどまらず、音声コーパスの構築作業（アノテーション作業）においても Praat が使用されている。筆者の関係している範囲では、『理研母子会話コーパス』(R_JMICC)（Mazuka, Igarashi, and Nishikawa 2006, 五十嵐、馬塚 2006）、『日本語話し言葉コーパス』(CSJ)（前川 2006）の構築作業で Praat が用いられており、音声に関する研究用付加情報が TextGrid ファイル（Praat 用アノテーション形式ファイル）で管理されている。

コーパスアノテーション作業に Praat が有用であることは少なくとも著者の経験からは明らかだが、問題点もある。それは、ある程度の規模のファイルセットを扱おうとすると、操作が繁雑になることである。例えば CSJ では、コアと呼ばれるサブセット（約 44 時間）について、201 個の TextGrid ファイル（1 ファイル平均約 13 分）が提供されているが、これらを Praat で扱おうとすると、実際の作業の前に、(i) 201 個のファイルの中から特定の TextGrid ファイル（とそれに対応する音声ファイル）を手動で開き、さらに (ii) 平均約 13 分の長さの講演の中で特定の箇所に手動で移動する、という操作を行なわなければならない。これらは、Praat を使う限りいわば「当たり前」の操作ではあるが、コーパスサイズが大きくなると、作業者にとってかなりの負担となる。

この問題を解消するため、Praat 起動ツール “Praat Launcher” を作成した。Praat Launcher は Microsoft Excel のアドインとして実現されており、Excel ワークシート上で実行されると、次の処理を自動的に行なう：

1. Praat を起動する
2. ワークシートで指定されているファイル（音声・TextGrid ファイル）を読み込む
3. 読み込まれたデータを Praat Editor で開く
4. ワークシートで指定されている区間（または点）を表示する

このツールを用いることにより、作業者はわずらわしい操作から解放され、アノテーション作業に集中することができる。Excel 2003/2007/2010 (for Windows), 2004/2011 (for Macintosh)

† nisi@brain.riken.jp

で動作する[*1]。以下のサイトから入手可能である：

http://language.world.coocan.jp/scripts/?PraatLauncher

Praat Launcher は、R_JMICC および CSJ 第三刷の構築作業で実際に使用されている。また、音声コーパスの構築に限らず、公開済みコーパスの閲覧、小規模の読み上げ音声の分析、実験刺激音声の管理などにおいても有用だと思われる。

本稿では、Praat Launcher の使用法を紹介する[*2]。

2. 使用法

Praat Launcher を実行するには、以下の手順を踏む[*3]：

1. データシートの作成
2. 設定シートの作成
3. Praat 起動コマンドの実行

以下、順に説明する。

2.1 データシートの作成

データシートとは、特定の列 (column) にファイル名（ベース名）とタイムスタンプが記されたワークシートのことである。このワークシート上から Praat Launcher が実行される。Praat Launcher 自身にはデータシートを作成する仕組みは備わっていないので、別途作成する必要がある[*4]。例を図 1 に示す。これは、CSJ から「(けど) も」を検索し加工したものである。

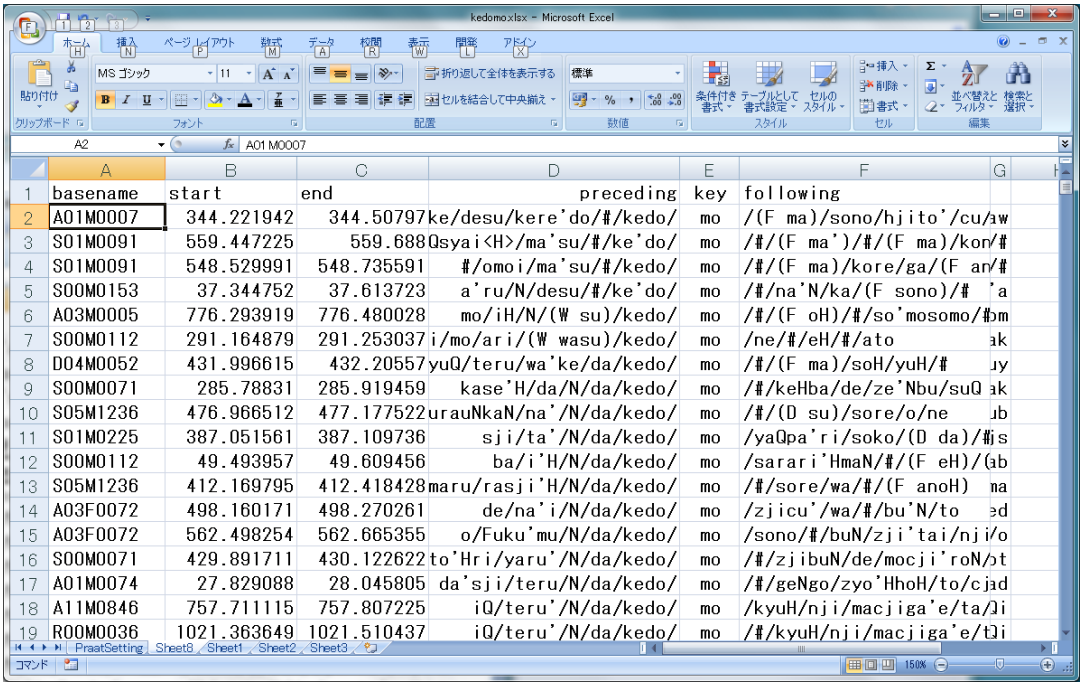
図 1 では、A 列にファイル名、B 列および C 列にタイムスタンプ（それぞれ「も」の開始時刻、終了時刻）が記されている。D 列以降には、検索対象文字列「も」が文脈付きで（いわゆる KWIC 形式で）表現されている。ファイル名・タイムスタンプ以外の情報は、Praat Launcher の実行においては必要とはされないが、作業者がラベルを確認したり、それらを基に Excel の

[*1] Excel 2008 (for Macintosh) では動作しない。これは、Praat Launcher の記述に用いたプログラミング言語 (Visual Basic for Applications; VBA) が Excel 2008 に搭載されていないためである。なお、次バージョンである Excel 2011 では幸いなことに VBA が復活している。

[*2] 紙面の都合上、本稿では Praat Launcher の基本的な機能の紹介にとどめざるをえない。機能の詳細やセットアップ方法については、上記サイトにある文書を参照されたい。

[*3] 以下の説明は、Praat Launcher, Version 2.1.1（2012 年 8 月 6 日）に基づく。このバージョンには大きな変更が加えられているので、それ以前のバージョン (Version 1.5.x) をお使いの方は注意されたい。

[*4] 図 1 に挙げた例は、CSJ TextGrid ファイルをもとに Praat スクリプトで作成したものである。TextGrid はテキストファイルなので、Perl 等のスクリプト言語を使って加工することも可能である。また、数が少なければ、手で直接ワークシートに入力してもよい。公開済みのコーパスであれば、他の手段も考えられる。例えば CSJ では、各種情報を統合して表現した XML 文書が提供されているので（菊池、塚原 2006 参照）、XSLT を使って Praat Launcher 用のデータを作成することができる（XSLT を使えば、検索条件を細かく指定できて便利である；もっとも、XSLT による情報の抽出・加工は、多くの人文系研究者 — その中には本稿の筆者も含まれる — にとって敷居が高いようであるが）。また、現在国立国語研究所で構築中の CSJ-RDB（小磯、伝、前川 2012 参照）は、もともとテーブルの形で管理されているため、SQL による検索で、そのまま Praat Launcher 用データシートとして使える出力結果を得ることができる（RDB+SQL は XML+XSLT と比べてはるかにとっつきやすいという利点もある）。



	A	B	C	D	E	F
1	basename	start	end	preceding	key	following
2	A01M0007	344.221942	344.50797	ke/desu/kere'do/#/kedo/	mo	/(F ma)/sono/hjito'/cu/aw
3	S01M0091	559.447225	559.688	Qsyai<H>/ma'su/#/ke'do/	mo	/#/(F ma')/#/(F ma)/kon/#
4	S01M0091	548.529991	548.735591	#/omoi/ma'su/#/ke'do/	mo	/#/(F ma)/kore/ga/(F an/#
5	S00M0153	37.344752	37.613723	a'ru/N/desu/#/ke'do/	mo	/#/na'N/ka/(F sono)/# 'a
6	A03M0005	776.293919	776.480028	mo/iH/N/(W su)/kedo/	mo	/#/(F oH)/#/so'mosomo/#m
7	S00M0112	291.164879	291.253037	i/mo/ari/(W wasu)/kedo/	mo	/ne/#/eH/#/ato ak
8	D04M0052	431.996615	432.20557	yuQ/teru/wa'ke/da/kedo/	mo	/#/(F ma)/soH/yuH/# uy
9	S00M0071	285.78831	285.919459	kase'H/da/N/da/kedo/	mo	/#/keHba/de/ze'Nbu/suQ ak
10	S05M1236	476.966512	477.177522	urauNkaN/na'/N/da/kedo/	mo	/#/(D su)/sore/o/ne ub
11	S01M0225	387.051561	387.109736	sji/ta'/N/da/kedo/	mo	/yaQpa'ri/soko/(D da)/#s
12	S00M0112	49.493957	49.609456	ba/i'H/N/da/kedo/	mo	/sarari'HmaN/#/(F eH)/(ab
13	S05M1236	412.169795	412.418428	maru/rasji'H/N/da/kedo/	mo	/#/sore/wa/#/(F anoH) ma
14	A03F0072	498.160171	498.270261	de/na'i/N/da/kedo/	mo	/zjicu'/wa/#/bu'N/to ed
15	A03F0072	562.498254	562.665355	o/Fuku'mu/N/da/kedo/	mo	/sono/#/buN/zji'tai/nji/o
16	S00M0071	429.891711	430.122622	to'Hri/yaru'/N/da/kedo/	mo	/#/zjibuN/de/mocji'roN/ot
17	A01M0074	27.829088	28.045805	da'sji/teru/N/da/kedo/	mo	/#/geNgo/zyo'HhoH/to/cjad
18	A11M0846	757.711115	757.807225	iQ/teru'/N/da/kedo/	mo	/kyuH/nji/macjiga'e/ta/Ji
19	R00M0036	1021.363649	1021.510437	iQ/teru'/N/da/kedo/	mo	/#/kyuH/nji/macjiga'e/tJi

図 1　データシート例（CSJ「（けど）も」検索結果の一部）

「並べ替え」や「フィルタ」などの操作を行なったりすることができるようになるので、用意しておくと便利である。

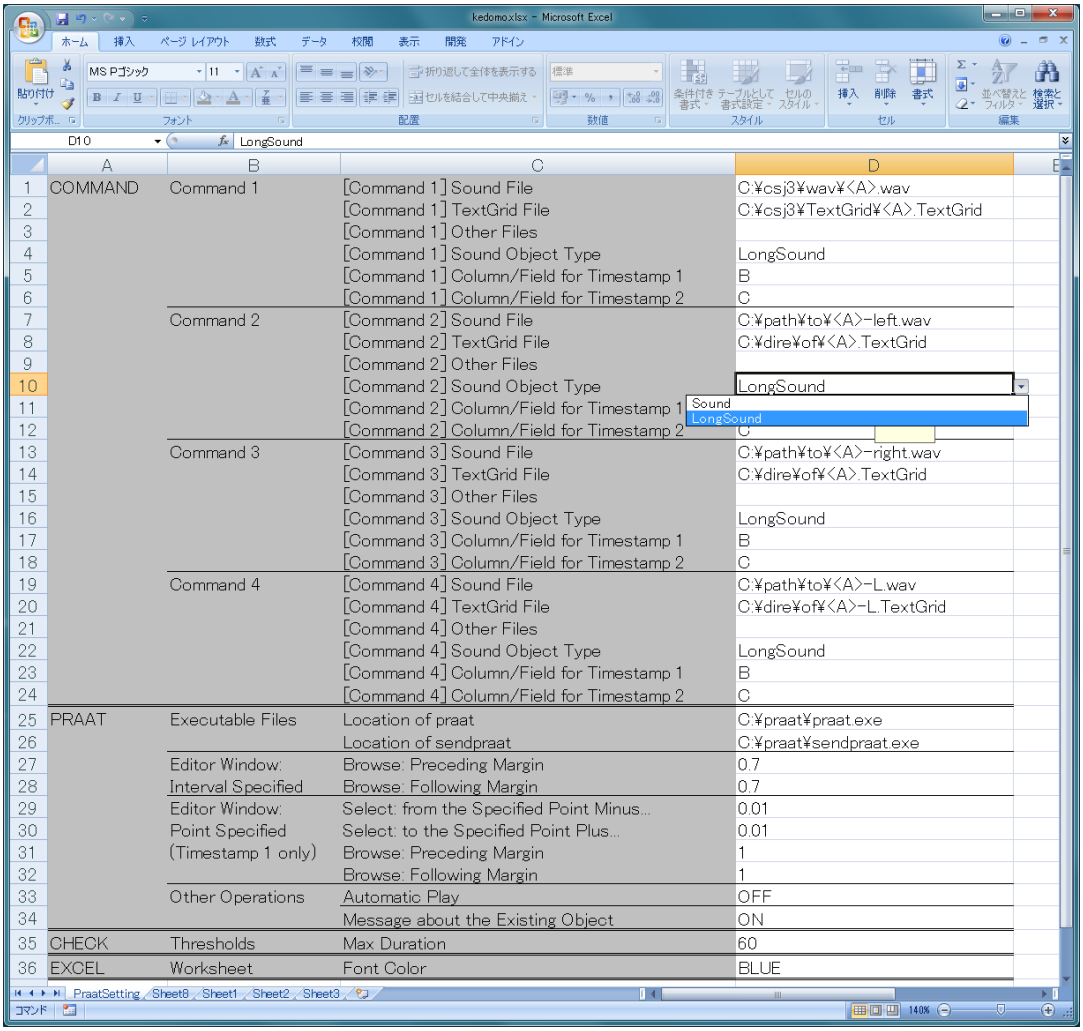
図 1 にはタイムスタンプ列が二つあるが（開始時刻・終了時刻）、一つだけでも Praat Launcher の実行は可能である（この場合、区間ではなく点が指定されることになる）。ただし、少なくとも一つは必要である[*5]。

2.2　設定シートの作成

設定シートは、起動に関する各種パラメータ（対象とする音声・TextGrid ファイル、データシートにおけるタイムスタンプ列、Praat 実行ファイルなど）を指定するためのワークシートである。“PraatSetting” というシート名を持つ。アドインメニューの “Setting” → “Create” を実行することで作成される。例を図 2 に示す。設定シート作成後、同シート D 列の値を自身の環境に合わせて変更する。一部の項目については、プルダウンメニューからの値の入力が可能である。

設定シートの前半（1-24 行）は起動コマンドに関わるパラメータである。Praat Launcher には起動コマンドが 4 つ用意されており (Command 1-4)、一部のパラメータに関してそれぞれに異なる値を設定できる。ただし、使用しない起動コマンドについては設定の必要はない（または、他の起動コマンドと同じ値を設定してもよい）。起動コマンドのパラメータについて説

[*5] タイムスタンプ値を欠くデータシートから Praat Launcher を実行したい場合、値を “0” とする列を作成し、それをタイムスタンプ列とする。



	A	B	C	D
1	COMMAND	Command 1	[Command 1] Sound File	C:¥csj3¥wav¥<A>.wav
2			[Command 1] TextGrid File	C:¥csj3¥TextGrid¥<A>.TextGrid
3			[Command 1] Other Files	
4			[Command 1] Sound Object Type	LongSound
5			[Command 1] Column/Field for Timestamp 1	B
6			[Command 1] Column/Field for Timestamp 2	C
7		Command 2	[Command 2] Sound File	C:¥path¥to¥<A>-left.wav
8			[Command 2] TextGrid File	C:¥dire¥of¥<A>.TextGrid
9			[Command 2] Other Files	
10			[Command 2] Sound Object Type	LongSound
11			[Command 2] Column/Field for Timestamp 1	Sound / LongSound
12			[Command 2] Column/Field for Timestamp 2	C
13		Command 3	[Command 3] Sound File	C:¥path¥to¥<A>-right.wav
14			[Command 3] TextGrid File	C:¥dire¥of¥<A>.TextGrid
15			[Command 3] Other Files	
16			[Command 3] Sound Object Type	LongSound
17			[Command 3] Column/Field for Timestamp 1	B
18			[Command 3] Column/Field for Timestamp 2	C
19		Command 4	[Command 4] Sound File	C:¥path¥to¥<A>-L.wav
20			[Command 4] TextGrid File	C:¥dire¥of¥<A>-L.TextGrid
21			[Command 4] Other Files	
22			[Command 4] Sound Object Type	LongSound
23			[Command 4] Column/Field for Timestamp 1	B
24			[Command 4] Column/Field for Timestamp 2	C
25	PRAAT	Executable Files	Location of praat	C:¥praat¥praat.exe
26			Location of sendpraat	C:¥praat¥sendpraat.exe
27		Editor Window:	Browse: Preceding Margin	0.7
28		Interval Specified	Browse: Following Margin	0.7
29		Editor Window:	Select: from the Specified Point Minus...	0.01
30		Point Specified	Select: to the Specified Point Plus...	0.01
31		(Timestamp 1 only)	Browse: Preceding Margin	1
32			Browse: Following Margin	1
33		Other Operations	Automatic Play	OFF
34			Message about the Existing Object	ON
35	CHECK	Thresholds	Max Duration	60
36	EXCEL	Worksheet	Font Color	BLUE

図 2　設定シート例

明する[*6]。

- Sound File, TextGrid File, Other Files
  - 音声ファイル・TextGrid ファイル・その他のファイル（Pitch ファイル・スクリプトファイルなど）をパス・拡張子を含めて指定する。
  - 指定の際、“<列>” という表記法でデータシートカレント行のセルの値を表現することができる。例えば “C:\path\to\<A>.wav” は、カレント行 A 列のセルの値が “A01F0055” であれば “C:\path\to\A01F0055.wav” と解釈され、“S07M0833” であれば “C:\path\to\S07M0833.wav” と解釈される。
    - “C:\path\to\<A>-L.wav” のように指定することで、データシート上のセルの値とは異なるファイル名を持つファイルを開くこともできる（この場合、例

[*6] それ以外のパラメータについては説明を省略する。起動コマンドの各パラメータと “Location of praat”, “Location of sendpraat” に適切な値が設定されていれば、Praat Launcher は動作するはずである。

えば “C:\path\to\A01F0055-L.wav” と解釈される)。
    - ◇ “C:\path\to\<A>-<E>.wav” のように “<列>” 表記を組み合わせて使用することも可能である。
  - Other Files の値は空であってもよい。
  - Sound File と TextGrid File のうち、どちらか一方のファイルだけを開きたいときは、もう一方の値を空にしておく。
- Sound Object Type
  - 音声ファイルをどのタイプの Object として Praat に読み込むかを指定する。値は “Sound” か “LongSound” のいずれかである。CSJ で提供されている .wav ファイルのような長めの音声ファイルであれば、“LongSound” を選択する。
- Column/Field for Timestamp 1, Column/Field for Timestamp 2
  - データシートにおいてタイムスタンプ（開始時刻・終了時刻）が記されている列を指定する。データシートにタイムスタンプ列が一つしか存在しない場合には、両方に同じ値（タイムスタンプが記されている列）を指定する。

設定シートに関して補足しておく：

- 設定シートは今扱っているデータシートと同じワークブックに作成する必要がある (Praat Launcher は現在のワークブックに属する設定シートしか参照しない[*7])。
- 設定シートは省略可能である。Praat Launcher は、現在のワークブック内に設定シートがあればそのシートの値を、なければソースコードに定義されているデフォルト値を用いて Praat を起動する（設定シートに最初に現れる値が、ソースコードで定義されているデフォルト値である)。ソースコード上のデフォルト値を自身の環境に合致するよう書き換えれば[*8]、設定シートを作成しなくても Praat を起動できるようになり便利である[*9]。

2.3 Praat 起動コマンドの実行

Praat の起動はデータシートから行なう。データシート上の任意の箇所にセルポインタを合わせ、アドインメニュー “Praat”（図 3)、右クリックメニュー（図 4)、ショートカットキーのいずれかからから起動コマンド (Command 1-4) を実行する（筆者はもっぱらショートカットキーを用いている[*10])。すると、その行のセルの値（と設定シート D 列の値）が読み込まれ、Praat が起動する。起動後、その行のフォント色が変更される。

[*7] この仕様により、現在のデータシートを通常の Excel ワークブック形式 (.xls, .xlsx) で保存すれば、設定シートを含めて保存されるので、次回からは特別な操作なしに作業を再開できる。

[*8] プログラミング言語に関する若干の知識があれば、ソースコードの書き換えはそれほど難しくはないと思われる。詳細については、上掲のサイトを参照されたい。

[*9] R_JMICC 構築作業および CSJ 第三刷構築作業においては、ソースコード内の起動パラメータデフォルト値を作業環境に合わせて書き換えた。

[*10] ショートカットキーの割り当てはアドインメニュー（図 3）に記してあるので、適宜参照されたい。ショートカットキーは “C-S-□” という形に統一されているが、これは「Ctrl キーと Shift キーを押しながら□キーを押す」という意味である。

図 3　アドインメニュー内起動コマンド

図 4　右クリックメニュー内起動コマンド

Praat Launcher 実行例を図 5 に示す（データシート 17 行目で実行）。次の二点を確認していただきたい。

- データシート 17 行目 A 列に指定されているファイル (`A01M0074.{wav/TextGrid}`) が開かれていること
- データシート 17 行目 B 列に指定されているタイムスタンプ (27.829088) から、17 行目 C 列に指定されているタイムスタンプ (28.045805) までが Praat TextGridEditor 上で選択されていること

3.　終わりに

Praat Launcher の使用法を一通り紹介した。Praat Launcher が Praat の補助ツールとして、直接・間接に音声研究に寄与するものとなれば幸いである。

付記

- Praat Launcher は無保証です。これによって生じた損害に対して作者は一切の責任を負いません。
- Praat Launcher の改変・再配布は自由に — 作者の許可なしに — 行なうことができます。ただし、改変したものを配布する場合には、オリジナルの作者の名前を保持した上で、改変したことを明示してください。
- Praat Launcher の仕様は予告なしに変更されることがあります。

謝辞

Praat Launcher の作成にあたり、五十嵐陽介、小磯花絵、宇都木昭の各氏から貴重なコメントを頂戴した。Praat Launcher 開発版のヘビーユーザーであることを強要された理研「コーパスチーム」のメンバー — 西海枝洋子、伊藤直子、小西隆之、渡辺和希の諸氏 — からは、多くのフィードバックをいただいた。小西光氏には Praat Launcher 改良の重要なヒントをご示唆いただいた。Praat Launcher 関連文書の整備にあたっては、Andrew Martin、西海枝洋子両氏の助言を得た。ここに記して感謝する。Last, but not the least, I would like to thank Paul Boersma and David Weenink, the authors of Praat.

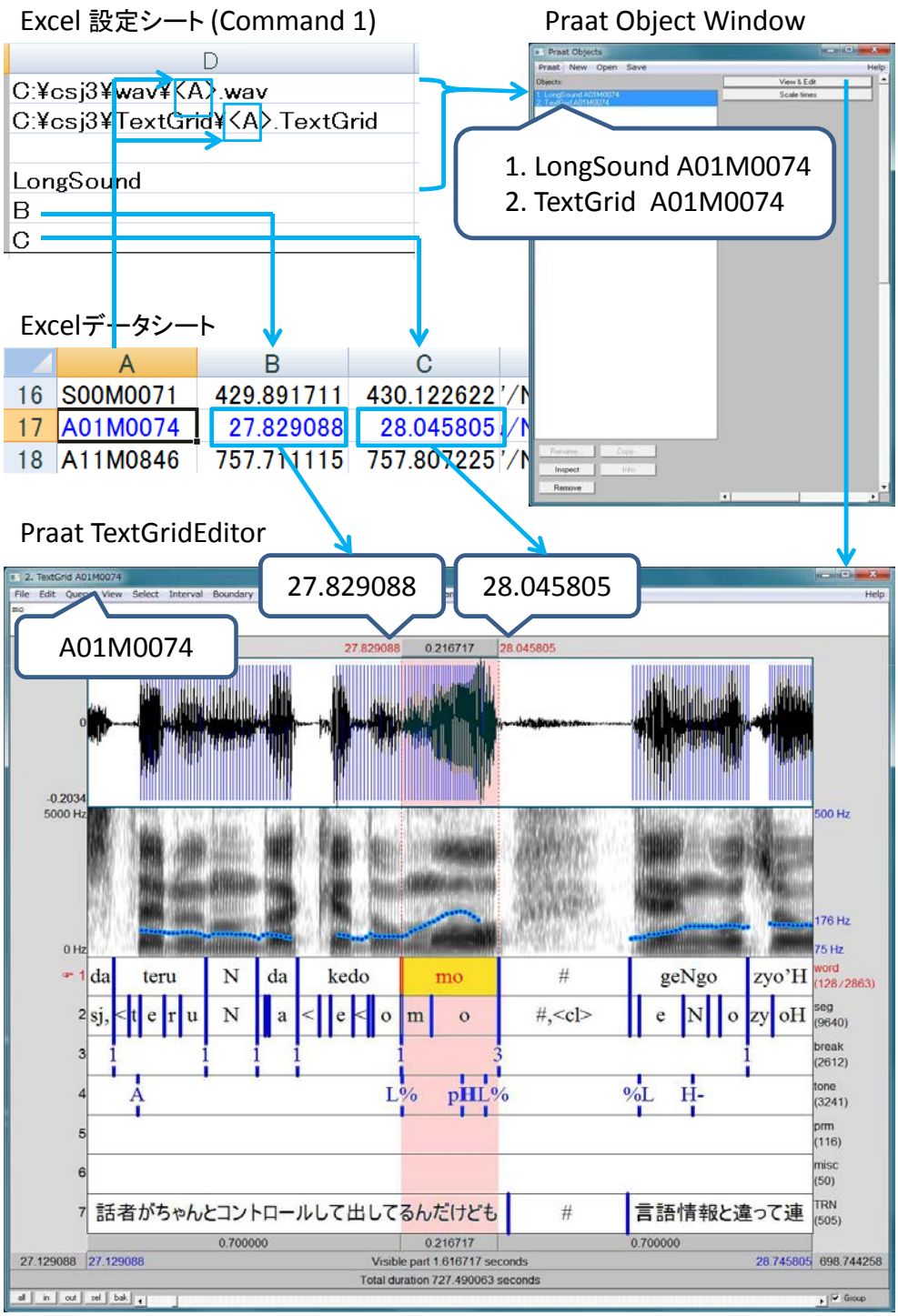


図 5　Praat Launcher 実行例

参考文献

Boersma, Paul and David Weenink (2012) Praat: doing phonetics by computer [Computer program]. Version 5.3.20, retrieved 28 June 2012 from http://www.praat.org/

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ー理研日本語母子会話コーパスー」、電子情報通信学会技術研究報告 106(443), pp. 31-35.

菊池英明、塚原渉 (2006)「第 8 章 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集 124）. pp. 455-526.（http://www.ninjal.ac.jp/csj/doc/k-report/ よりダウンロード可能）

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.（http://www.ninjal.ac.jp/event/project-meeting/m-2011/jclws01/JCLWorkshop_no1_papers/JCLWorkshop2012_53.pdf よりダウンロード可能）

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集 124）. pp. 1-22.（http://www.ninjal.ac.jp/csj/doc/k-report/ よりダウンロード可能）

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. TL, 思考と言語 106(165), pp. 11-15.

多様な話者による演技感情音声の収集と特徴の比較

宮島 崇浩（早稲田大学人間総合研究センター）†
菊池 英明（早稲田大学人間科学学術院）

Collection of Acted-Emotional Speech Using Various Actors and Comparison of Their Acoustical Features

Takahiro Miyajima (Advanced Research Center for Human Sciences, Waseda University)
Hideaki Kikuchi (Faculty of Human Sciences, Waseda University)

１．はじめに

音声研究は、非言語的情報を高度に取り扱うことが求められる段階に至っている。この潮流を受け、近年の音声コーパス研究では表現豊かな音声の収集方法が大きく着目されている(Erickson(2005))。近年は自発音声の収集が主流であるが(Campbell(2005))、演技音声を用いれば、表現豊かな自発音声の収集において制御が難しいとされる録音品質および表現の多様性の確保や、様々なメディア・職業において出現しうる幅広い音声表現の獲得の可能性がある。我々はこれまでに、様々なコンテキストを設定した「台本」を設計し、これに基づいて収録した声優による演技感情音声の音響的・心理的特徴を分析してきた(Miyajima(2012))。そして、提案手法による演技音声は、従来の演技音声と比べて自然性や表現の多様性が向上することが示唆された。本稿では、声優・映像系俳優・舞台系俳優の男女各１名による演技感情音声を比較し、話者・発話内容毎の演技音声の特徴を考察する。

２．多様な音声表現コーパス（SEN コーパス）

我々は、演技者に与える刺激（台本）を工夫することで、表現豊かな音声の収集を試みてきた。この手法によって得た音声資料群を多様な音声表現コーパス（通称：SEN コーパス）と呼称する。図１に収録のコンセプト、表１に工夫した刺激の例を示す。この刺激は、演劇論(安藤(2002))や音声のコミュニケーションモデル(Scherer(2003))を考慮し、構成（図１における『台本のフォーマット』）を決めたものである。この構成に基づき、具体的なインスタンス（内容）を用意し、音声を収録するまでが音声資料生成のスキームとなる。理想的には、より多様な心理パラメータ・音響パラメータ・刺激の内容のバリエーションを確保することで、利便性の高いコーパス構築が可能となると考えた。

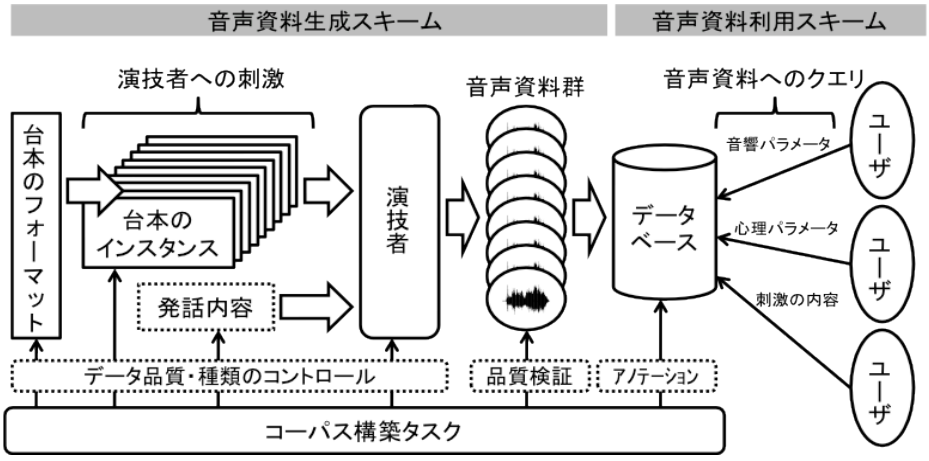


図１　提案手法のコンセプト

†miyajima@toki.waseda.jp

表 1　台本のフォーマットとインスタンスの例

台本のフォーマット		台本のインスタンスの例
共通の情報	発話時の場所・状況	ドラマ、学生寮のホール
	発話者と聞き手の関係	同じ学校、寮に住む親しい友人たち
話し手が持つ聞き手の情報	年齢・性別	17〜18 歳、男
	職業・続柄	高校生
	人物像	騒々しく、騒いでいる最中
話し手自身の情報	年齢・性別	17 歳、女
	職業・続柄	高校生
	人物像	男子校に入った、男のフリをした女 冷静だが、恋愛に関して鈍感
	発声時の背景	周りの騒ぎっぷりを呆れてみるように

表 2　収録済み音声一覧

収録 ID/ フォーマット	台本作成方法/ 発話内容	台本数（音声数）/ 演技者
SEN-1 パターン(1)	TV 番組から主観的に作成（パイロット版） 「ああ、そうですか」	304(100) 女性 1 名
SEN-2 パターン(1)	SEN-1 の項目組み合わせ＋感情表現を明示的に追加 「ああ、そうですか」	280(1400) 女性 5 名
SEN-3 パターン(2)	パーソナリティが変化するよう独自の仕様を作成 「よろしくお願いします」	1000(1000) 男女各 1 名
SEN-4 パターン(3)	SEN-1 の「発話時の背景」を長文化かつ体系化 「いたい」「からい」「いそがしい」など 6 種類	2400(7200) 男女各 3 名
Typical 基本感情語	基本感情語をそのまま刺激として使用 「ああ、そうですか」	80(480) 女性 3 名

2．1　これまでの収録

本手法を用いてこれまでに 3 回の収録を実施してきた（表 2 の SEN-1〜SEN-3)。本稿では，SEN-1 および SEN-2 の概観について触れたのち，最新の収録である SEN-4 の詳細について述べる。SEN-4 は、SEN の拡張として、複数の演技者カテゴリ（声優・舞台系俳優・映像系俳優）および複数の発話（6 種類）を用意している。そこで、それらの違いによる効果の差異を確認する（4 章)。なお、表 2 における Typical は、従来研究で用いられてきたように、基本感情語をキーに発話した感情演技音声の模倣として作成したものである。

2．2　SEN-1 および SEN-2 の特徴

SEN-1 はパイロット版であり、SEN-2 はパイロット版からの話者の拡張および台本インスタンスの改善を施したものである。SEN-1 では、Typical との自然性および心理的・音響的多様性の確認を主に考察し、SEN-2 では話者ごとの効果の違いについて主に考察している。なお、SEN-2 における話者の拡張では、演技者カテゴリを考慮せず、声優経験あるいは舞台経験のある女性 5 名を主観的に選んでおり、SEN-4 ではそれを体系化した点で異なっている（3 章で詳述)。

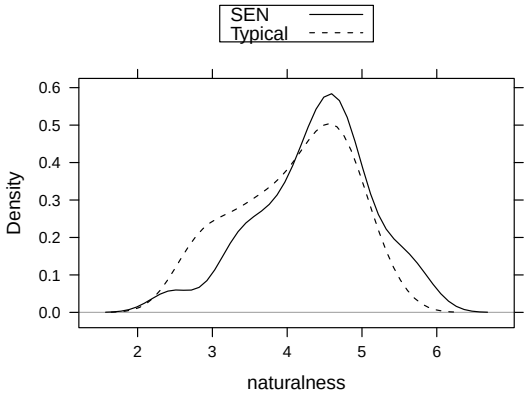


図 2　SEN-1 の自然性評価

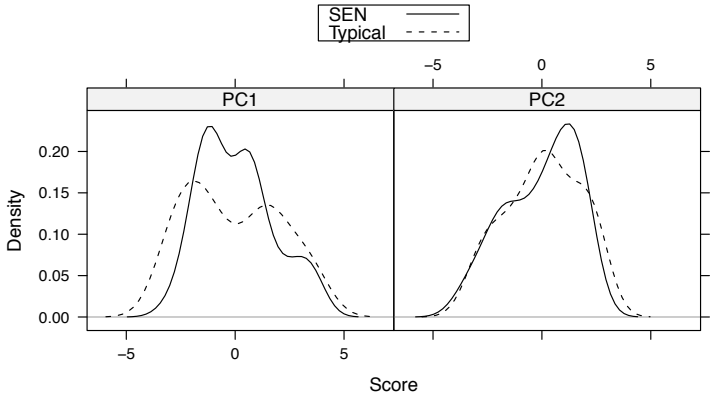


図 3　SEN-1 の心理的特徴の評価（PC1：感情価、PC2：覚醒）

２．２．１　自然性の確認

SEN-1 と Typical の各音声データ群の中から、各 50 個をランダムに抽出し、自然性の印象評定を実施した。評価は、「非常に不自然(1)～どちらでもない(4)～非常に自然(7)」の 7 段階を設定した。図 2 は、評価値に対するカーネル密度プロットである。カーネルは Gaussian カーネル、バンド幅は Silverman の手法により決定した（次節のカーネル密度プロットも同様）。評価値 4 を中心として、自然性の値が低い部分では Typical の密度が全体的に高く、高い部分では SEN-1 の密度が全体的に高いことが分かる。また、各データ群に対して t 検定を実施したところ、p=0.06 で有意傾向であることを確認した。以上より、本手法によって収集する音声は Typical と比べると自然性が高くなる傾向があることが読み取れる。

２．２．２　心理的特徴の確認

自然性同様、SEN-1 と Typical 各 50 個のデータに対して心理的特徴の測定を試みた。心理的特徴の測定方法として、森山(1999)による「音声による感情表現語」を評価語とした印象評定を実施した。感情表現語は、「怒り・喜び・皮肉・恐れ・悲しい・驚き・こび・穏やか・おかしい」の 9 語からなる。それぞれの評価語に対して、「全く当てはまらない(1)~どちらでもない(4)~非常に当てはまる(7)」の 7 段階を設定した。評価者は、大学生の男女各 6 名を選んだ。図 3 は、評価結果に対して主成分分析を実施し、第一主成分(PC1)と第二主成分(PC2)のカーネル密度プロットを表示したものである（第二主成分までの累積寄与率 73%）。第一主成分は感情価（valence）、第二主成分は覚醒(arousal)と解釈した。覚醒の観点では SEN-1 と Typical はほぼ同様であるが、感情価の観点ではいわば相互補完の関係にあることが分かる。SEN-1 は感情価の強度が強い音声表現が少ないため（明確に感情表現を指定することをできるだけ避けたため）、SEN-2 では感情表現を加えた台本を設定して音声を収録した。

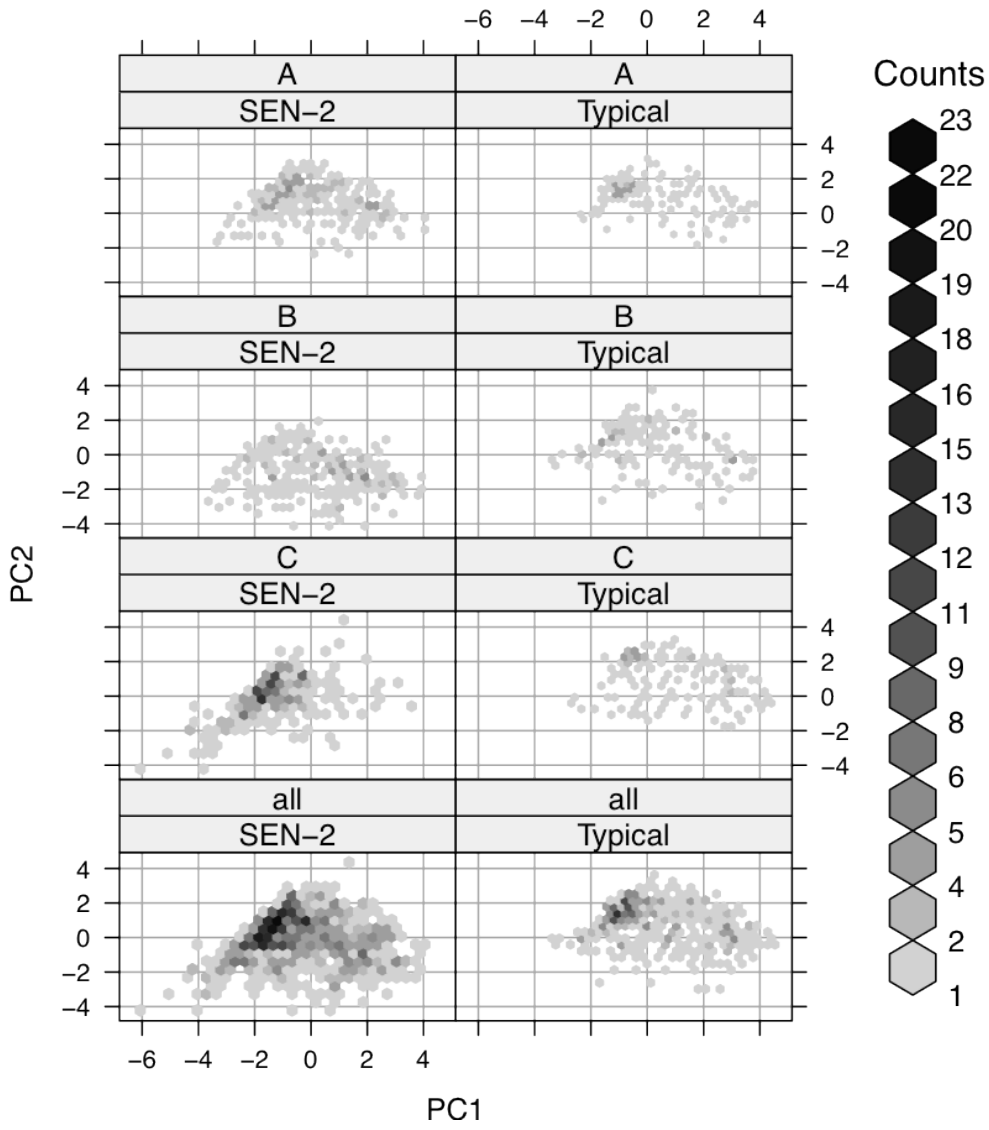


図 4　SEN-2 の音響的特徴の概観（女性 3 名分）

２．２．３　音響的特徴の確認

図 4 は、SEN-2（1 話者につき 280 個）と Typical（1 話者につき 160 個）の音響的特徴を比較したプロット図である。このプロット図では、Carr(2011)の手法により、散布図は大きさが変化する六角形の色の濃度によって密度が表現されている。音響的特徴として、一発話全体に対する F0 関係 5 種（平均、最大、最小、レンジ、標準偏差）および発話長関係 3 種（発話長、平均モーラ長、合計ポーズ長）を選び、主成分分析を行った。累積寄与率は、第二主成分までで 61%である（ここでは、概観を確認するため第二主成分までを示す）。第一主成分は発話内の F0 の変動、第二主成分は F0 の高さや発話長の成分が強く現れている。発話者は、SEN-2 と Typical の両方で収録を行った女性話者 3 名である。Typical は話者ごとの違いがほとんど見られないのに対し、SEN-2 は話者ごとによる違いが大きいことが分かる。つまり、SEN で採用した、演技に必要な多様な情報を刺激として提示するという手法は、話者ごとに表現の拡大の傾向が異なるという結果を導くことになり、これは先に述べた演劇論(安藤(2002))における、熟達者はひとつの脚本解釈から多様な演技計画を立てるという考察と合致していると考えられる。また、3 名分全体(all)で見ると、SEN-2 の音響的特徴は Typical から範囲がほぼ相似形を保ったまま拡大されていることが分かる。

３．多様な発話者カテゴリ・発話内容による SEN コーパスの拡張（SEN-4）

３．１　SEN-4 の収録の目的

SEN-1 および SEN-2 の心理的・音響的分析によって、本手法の適用による自然性の向上および心理的・音響的多様性向上の可能性が示唆された。とくに、２．２．３節で示した、SEN-2 の音響的特徴において話者ごとの違いが顕著に確認されたことは､本研究における要点だと考えられる。我々は、次の課題として、演技者のカテゴリや発話内容の違いによる多様性の広がりの差異について調べることにした。この目的のため、SEN−4 では、発話者（演技者）および発話内容を精密に選定することにし、さらに、表現豊かな発話を促すため、台本の改善も試みた。

３．２　発話者の選定

３．２．１　発話者カテゴリの考察

これまでの収録では、声優や舞台俳優としてのキャリアを持つ人物を主観的に選定してきた。しかし、SEN-3 までの収録した音声を聴取し考察するなかで、同じ声による演技のプロフェッショナルでも表現方法の違いはかなり大きく、本手法における音声表現の多様性に大きく左右することが分かってきた。そこで、いくつかの仮説を考えた。

> 仮説１：声のみによる表現と複数モダリティによる表現では表現方法が異なる
> 仮説２：聴衆者をどのように意識して演技するかで表現方法が異なる

これらの仮説に基づいて､我々は声を用いたプロフェッショナルの演技者のカテゴリ（発話者カテゴリ）として、次の区分を定義した：

- 声優：声のみを用いた表現者、観客は意識しない
- 舞台系俳優：複数モダリティを用いた表現者、観客を意識する
- 映像系俳優：複数モダリティを用いた表現者、観客を意識しない

さらに、3 カテゴリに関して、次の仮説を考えた。

> 仮説３：声の表現の多様性は、映像系俳優、舞台系俳優、声優の順に高くなる
> 仮説４：声の表現の自然性は、声優、舞台系俳優、映像系俳優の順に高くなる

本稿ではまず、仮説 3 における音響的多様性の確認をおこなう。声優は、声のみで多様な表現を求められるため、最も多様性が高いと想像できる。また、舞台系俳優と映像系俳優は、観客に対して演技することを意識するかどうかで、表現の度合いの大きさに差が出る可能性がある。例えば、舞台系俳優は観客を意識した演技をすることに対して、映像系俳優は観客を意識せず、日常生活に近い演技を行っていると考えられる。なお、観客の存在を意識して演技する点の重要性は、先行研究(安藤(2002))で述べられている。

以上のように、キャリアに基づく演技手法の質の違いにより、本手法による同一の手続きに従った演技でも、表現の広がりの差異を生み出すことが期待できる。また、その差異は、本研究の目的である、表現の多様性の獲得に強く関連するであろう。

さらに、男女による表現の違いも検討する。これまでの収録では主に女性を中心に収録を実施してきたが、基本周波数の特性の違いを考慮すると、女性の音響的な表現力が豊かである可能性が高い。これを検証するため、SEN-4 では、各発話者カテゴリについて、男女を別カテゴリとして扱い、計 6 カテゴリの話者を選定することにした。

表 3　SEN-4 における発話内容の詳細

セット	発話内容	極性	音声単語親密度	アクセント型	末尾の音素列
セット１	うまい	Positive	6.312	2	/ai/
３モーラ	からい	Neutral	6.250		
形容詞	いたい	Negative	6.312		
セット２	すばらしい	Positive	6.344	4	/asii/
５モーラ	いそがしい	Neutral	6.156		
形容詞	むずかしい	Negative	6.031		

３．２．２　オーディションの実施

演技者の選定は、オーディション形式を採用することにした。手続きとして、12 通りの台本サンプルを用意し、発話を録音してもらったうえ返送してもらうよう依頼した。台本サンプルの形式は、SEN-4 で用いた台本（表 1 の「発話時の背景」を長文化したもの）と同一であり、発話内容は「ああ」と「そうですか」の 2 種類（各 6 通り）を用意した。「ああ」を選んだ理由は、極めて短い発話でも表現力が豊かであるかどうかを確認するためであり、「そうですか」を選んだ理由は、既存の SEN のデータと比較しながら多様性を演出できる人物かどうか検討するためである。

依頼に際しては、声優・舞台系俳優・映像系俳優というカテゴライズを良く理解でき、かつ各カテゴリに対して人脈を持つ人物を仲介人として起用した。仲介人経由で、声優は特定の事務所、舞台系俳優は特定の劇団、映像系俳優は特定の映画監督に依頼し、本研究の主旨を伝えたうえで、適正があると判断したもらった人物を各複数名選定してもらった。

最終的には、サンプル音声を聴き比べ、筆者らの主観で表現力の多様性が同程度であると思われる人物を、各カテゴリから男女各一名選定した。

３．３　発話内容のデザイン

発話内容は、SEN-3 までは中立的な内容であることを条件としてきた（「ああ、そうですか」「よろしくお願いします」）が、２．２．２節で述べた SEN-1 の心理的特徴の傾向の考察から、提案手法は感情価の強度の意識的なコントロールが重要であると考えられるため、発話内容の（感情価の）極性の違いが多様性に及ぼす影響の検証、さらに発話内容の極性毎による音響パラメータと心理パラメータの差異の検証が必要であると考えた。

また、発話の長さによっても多様性に違いが生まれる可能性がある。理想としては、どのような長さでも同様の多様性が確保できることが望ましいが、これまでの収録の過程で、複数の演技者より、短すぎる発話はバリエーションのある演技は難しいという意見が得られているため、これも合わせて検証できるような発話内容を選択することにした。

これらの観点に従って、発話内容の極性および長さを複数用意することにした。極性は、先行研究(小林(2005), 東山(2008))で提案された極性評価が記された辞書に従い、Positive/Negative/Neutral をそれぞれ選ぶことにした。また、発話の長さに関しては、モーラ数およびアクセント型を揃えることにした。さらに、全ての発話において、BPM などの重要な韻律情報の比較ができるよう、末尾の音韻が統一されるようにした。その他、音声単語親密度(天野(1999))や品詞の統一なども考慮した。

以上の条件に従い、図 4 に示した 6 種類の発話を収録することにした。

表 4　変更前後の「発話時の背景」の例

SEN-1, SEN-2	まだ見ぬヒーローをカッコいいと妄想し、憧れるように
SEN-4	学校から帰る電車の中。A は、手帳とにらめっこしながら今後数週間は一日も空かずに予定が埋まっていることを改めて確認する。それを見た瞬間、先を考えると過酷・・・と思いながら出た一言。

表 5　インスタンスに付与した属性一覧

属性	内容	個数
話し手の年代	10, 20, 30–40, 50–60 代	各 25 通り（計 100 通り）
聞き手の年代	上記＋なし（独り言）	話し手と同じ年代の場合：10 通り 異なる場合：4 通り、独り言：3 通り
聞き手の性別	同姓、異性	聞き手の各年代で半数ずつ（5 通り、2 通り）

３．４　台本の改善

３．４．１　日常性と非日常性

SEN-1 では、台本のフォーマットに従い、インスタンスを TV 番組からランダムに抽出するという手法を取った。また、SEN-2 では SEN-1 から抽出した項目の組み合わせで台本のインスタンスを拡張した。

SEN-1 および SEN-2 での問題点のひとつとして、Typical に比べて自然性は向上したと考えられるものの、やはり非現実的な場面、すなわちアニメーション等の創作物やバラエティ番組で聴くことができるような、やや偏った表現が多いと思われる点であった。我々の目的は、そのような非現実的な場面での表現も、我々が日常生活下で聴取できるような表現も全て収集できる手続きを確立することにある。

そこで、台本のインスタンス生成の際、意識的に「日常的な」インスタンス、「非日常的な」インスタンスの両方を生成することにした。具体的には、一つの発話につき、日常的な場面を意識した台本を 100 本、非日常的な場面を意識した台本を 100 本用意することにした。非日常的な場面とは、SEN-1 で用いたような、テレビ番組でのワンシーンの再現であり、日常的な場面とは我々が日常生活下で遭遇するワンシーンの再現である。これらを今回は完全に創作で作成した。また、日常性や非日常性をより強調するため、次節で示すように発話時の背景の長文化を施した。台本の原案は発話内容 1 種につき各 200 本の合計 1200 本であるが、発話者の性別による違和感の無いよう改変し合計 2400 本を用意した。

３．４．２　台本フォーマット「発話時の背景」の長文化

SEN-4 の台本フォーマットは SEN-1 と同様であるが、発話時の背景を長文化し、ショートストーリー仕立てにした。これは、これまでの収録で、発話時の背景の情報が少ないために強引に表現を広げようとし、不自然な音声表現を招く可能性があるという指摘を受けてのことである。変更前後の文章の例を表 4 に示す。長文化することで、より自然な表現、また、発話者カテゴリ毎に異なる多様な表現を導くことが目的である。

３．４．３　インスタンスの各項目への属性の事前付与

SEN-4 では、インスタンス作成は完全ランダムではなく、日常・非日常の各 100 本について、話し手の年代・聞き手の年代および性別の組み合わせ（属性）を創作前に指定し、音声データへのクエリとして利用できるようにした。詳細を表 5 に示す。

３．５　収録

国立国語研究所内のマルチメディアスタジオに、パソコンおよびUSBオーディオデバイス(Roland UA- 25)を設置した。マイクはヘッドセットタイプ(SHURE WH20XMR)を採用した。サンプリング周波数は44,100Hz、ビットレートは16bitに設定した。また、バックアップとしてリニアPCMレコーダ(SONY PCM-M10)による録音を同時に行った。さらに、将来的な検証のため、デジタルビデオカメラによる撮影を行った。

収録は、約1ヶ月にわたり断続的に実施した。1名あたり1200発話の収録を4回に分け、1回あたりの時間は発話者の任意で休憩を挟みながら3時間程度であった。刺激の提示はスタジオ内に設置したディスプレイを用いて行った。提示順は全員ランダムであったが、日常的な台本から始めたほうが演技しやすいとのコメントを受け、そのようにコントロールを行った。以上の手続きで、合計7200音声の収録を完了した。

４．収録した音声の音響的特徴の概要

SEN-4の7200個の音声について、２．２．３節で用いた音響的特徴を算出し、主成分分析を試みた。第一主成分はF0関連、第二主成分は発話長関連の成分が強く含まれている。第二主成分までで累積寄与率は72%であった。図5に、3モーラ、5モーラの単語ごと、極性ごと、そして話者ごとの計36通りの音響特徴量の密度プロット（各200個）を示す。

４．１　発話長ごとの分布の傾向

全体の傾向を確認する限り、3モーラ発話と比較して5モーラ発話は分布が狭い傾向が見られた。仮説では、5モーラ発話のほうが表現が豊かになると考えていたが、今回の主成分分析で用いた特徴量で判断する限りでは、3モーラ発話のほうが表現が豊かであるような結果が得られた。これらの特徴量は、発話全体に対して計算したもっとも単純な算出方法で求めたものであったので、短い発話ではそのような単純な特徴量で表現の多様性が演出された可能性がある。モーラ長による表現力の違いをより精密に検証するためには、BPMなど局所的な音響特徴量の確認が必要である。

４．２　極性ごとの分布の傾向

すべての話者に関して、極性ごとに大きな差異は見られなかった。本手法が必ずしもこの傾向を必ず保証することではないが、少なくとも今回用意した台本では、極性に左右されず、多様な表現を生み出すことが出来たと考えられる。

４．３　話者カテゴリごとおよび性別の違いによる分布の傾向

発話者カテゴリおよび性別による表現力の比較であるが、これも仮説と反し、声優と舞台系俳優については男性のほうが音響的多様性が高いことが確認された。特に女性声優については、収録中の表現の主観的印象は非常に幅広かったので、筆者らにとって意外な結果であった。一因として、この分析では用いていない音声パワーを良くコントロールして多様な表現を再現していたので、このような結果になったのだと推測される。

４．４　考察

今回起用した6名の話者が、各発話者カテゴリを代表する特徴を持つとは限らないが、全員が発話内容6種類のいずれもほぼ同様の分布を示したことは興味深い結果である。少なくとも、音響的多様性において、発話内容による影響は比較的小さいと考えられる。

また、男性映像俳優（MAM）の分布がいずれの発話においても一番小さい分布を示しているが、台本に対する整合性は他の話者カテゴリと同程度だと思われた。映像俳優は、非常に微妙な声の表現の際で台本の状況を再現しており、興味深いデータが得られている。

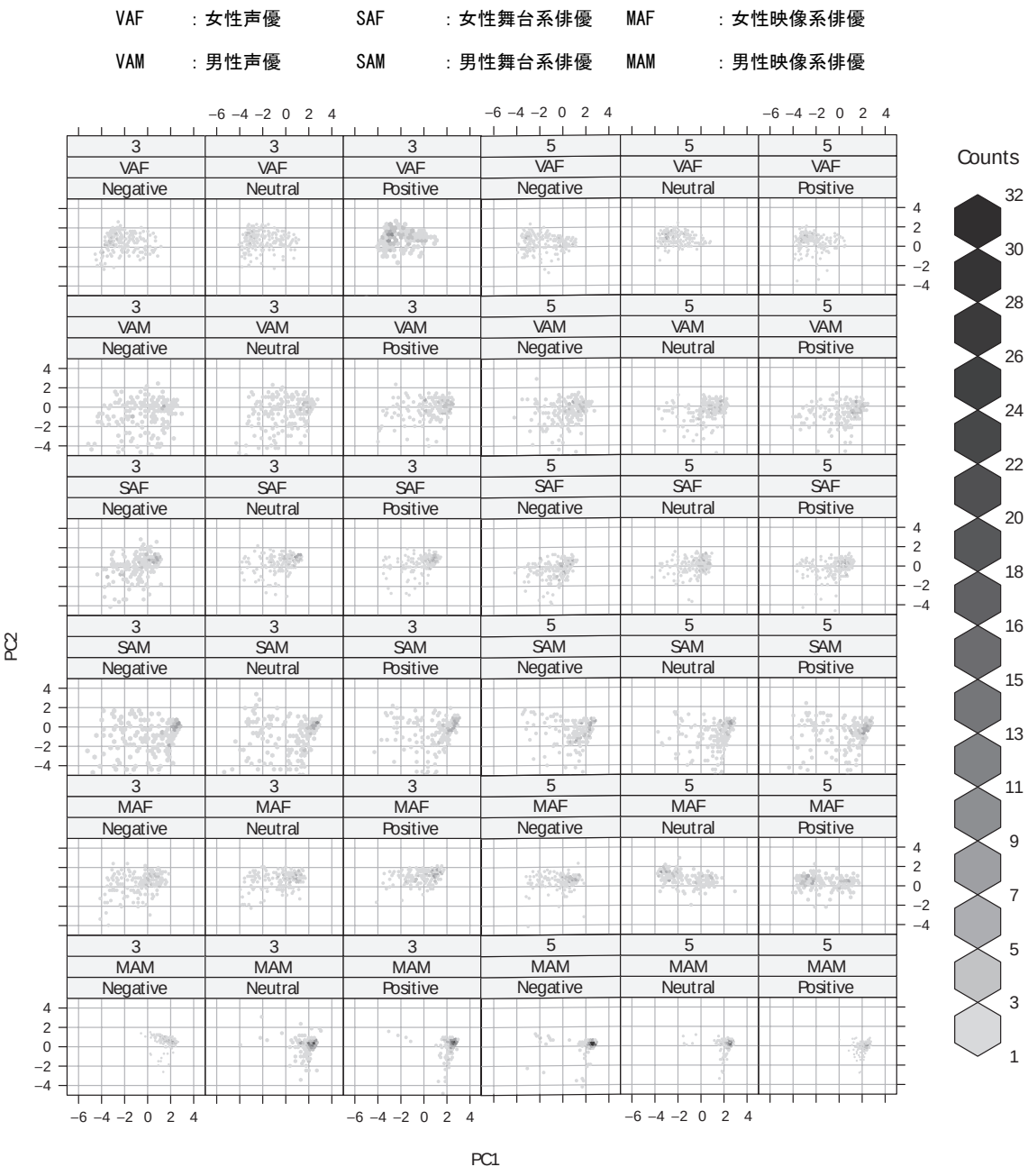


図 5　SEN-4 の音響的特徴の第一、第二主成分の密度プロット（各 200 個）

５．まとめ

我々が提案してきた多様な音声表現コーパスの構築手法に従い、多様な話者カテゴリ・多様な発話内容の収集を大規模に実施した。そして、各発話に対するシンプルな音響特徴量を用いて音声表現の分布の比較を行ったところ、いずれの話者カテゴリにおいても、発話内容の極性が分布に影響していないことが確認された。また、モーラ長が 5 モーラよりも 3 モーラのほうが広い分布を示した。これらの傾向は、選択した音響特徴の影響を強く受けていると考えられる。また、話者カテゴリ毎の特徴に関しては、我々の仮説と異なる結果が得られた。いずれに関しても、多様な心理的・音響的特徴による比較を通した追加検討が必要である。また、各演技者カテゴリの人数を増やすことで、結果がより明確になることが期待できるほか、本稿中で示した複数の仮説の検証が可能となるだろう。

謝　辞

SEN-4 の収録は国立国語研究所(言語資源研究系)基幹型共同研究「コーパス日本語学の創成」(リーダー:前川喜久雄)による成果である。また、演技者の選定や収録に関して、アーティスト(美術家・演出家)の河村美雪氏に多大なご協力を頂いた。最後に、台本に関して、早稲田大学人間科学部の菊池梨佳子氏が 1200 本もの原案を創作してくれた。ここに記して感謝の意を表する。

文　献

D. Erickson (2005). “Expressive speech: Production, Perception and Application to Speech Synthesis” Acoust. Sci. & Tech., vol.4, no.26, pp.317-325.

N. Campbell (2005). “Developments in corpus-based speech synthesis: approaching natural conversational speech” IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.376-383.

T. Miyajima, H. Kikuchi, K. Shirai, and S. Okawa (2012). “Method for Collection of Acted Speech Using Various Situation Scripts” Proc. of LREC 2012, pp.1179-1182.

安藤花恵(2002)「演技の熟達化脚本の読み取りから演技計画、演技遂行まで」心理学研究、Vol.74、No.7、pp.373-379.

K.R. Scherer (2003) “Vocal communication of emotion:a review of research paradigms” Speech Communication, vol.40, pp.227-256.

森山剛、斎藤英雄、小沢慎治 (1999) 「音声における感情表現語と感情表現パラメータの対応付け」信学論, vol.J82-D-2、no.4、pp.703-711.

D. Carr, N. Lewin-Koh, and M. Maechler (2011) “Hexbin: hexagonal Binning Routines” R package version 1.26.0.

小林のぞみ、乾健太郎、松本裕治、他 (2005)「意見抽出のための評価表現の収集」自然言語処理、Vol.12、No.2、pp.203-222.

東山昌彦、乾健太郎、松本裕治 (2008)「述語の選択選好性に着目した名詞評価極性の獲得」言語処理学会第 14 回年次大会論文集、pp.584-587.

天野成昭，他(編・著) (1999) 『日本語の語彙特性』、三省堂.

BCCWJに含まれるウェブデータの特性について
――データ重複の諸相とBCCWJ使用上の注意点――

田野村忠温（大阪大学大学院文学研究科）

On Certain Properties of the Web-Based Subcorpora of BCCWJ

Tadaharu Tanomura (Osaka University)

１．はじめに

インターネット上に存在する日本語文書は膨大かつ多様で、言語研究資料としてきわめて大きな価値と魅力を有する。そのインターネット文書の短所と言えば、一般に予想されやすいのは特殊な言葉遣いの出現や書き誤りの多さといった点であろうが、実際に最も問題となるのはむしろ同一データの重複出現である（拙論(2010)）。インターネット上にはさまざまな事情で同一の文章、段落、文、句が複製されて繰り返し現れる。

インターネット文書は「現代日本語書き言葉均衡コーパス(BCCWJ)」にも２つのサブコーパス「Yahoo!知恵袋」「Yahoo!ブログ」として収められ、両者を合わせてBCCWJ全体の約２割の分量を占めている。そしてこのたび、それらのサブコーパスにおいてもデータ重複の問題が思いのほか深刻であることが明らかとなった。

以下では、発表者が問題の認識に至った経緯と、２つのサブコーパスにおけるデータ重複の様相に関する調査・分析の結果について述べる。

２　BCCWJ N-gram分析ツール BNAnalyzer

２.１　ツールの概要

発表者は過日、BCCWJの検索結果を用いて語句のコロケーションを調べるための簡易な分析ツールBNAnalyzer（Windows上で作動）を作成し、次のところで公開した。

http://www.tanomura.com/research/BNAnalyzer/

BNAnalyzerは、BCCWJ検索サイト「中納言」での検索結果をもとに、検索語の前後にどのような表現がよく現れるかを分析する。具体的には、検索語の前後文脈のN-gram（N個の短単位または長単位の連続）の一覧を作成し、エクセルで表示する。

２.２　インストール方法および使用法

上記URLのページの説明に従ってBNAnalyzerをインストールすると、デスクトップ上にアイコンが２つ作られる。それぞれを単純版、circumcollocate版と呼ぶ。

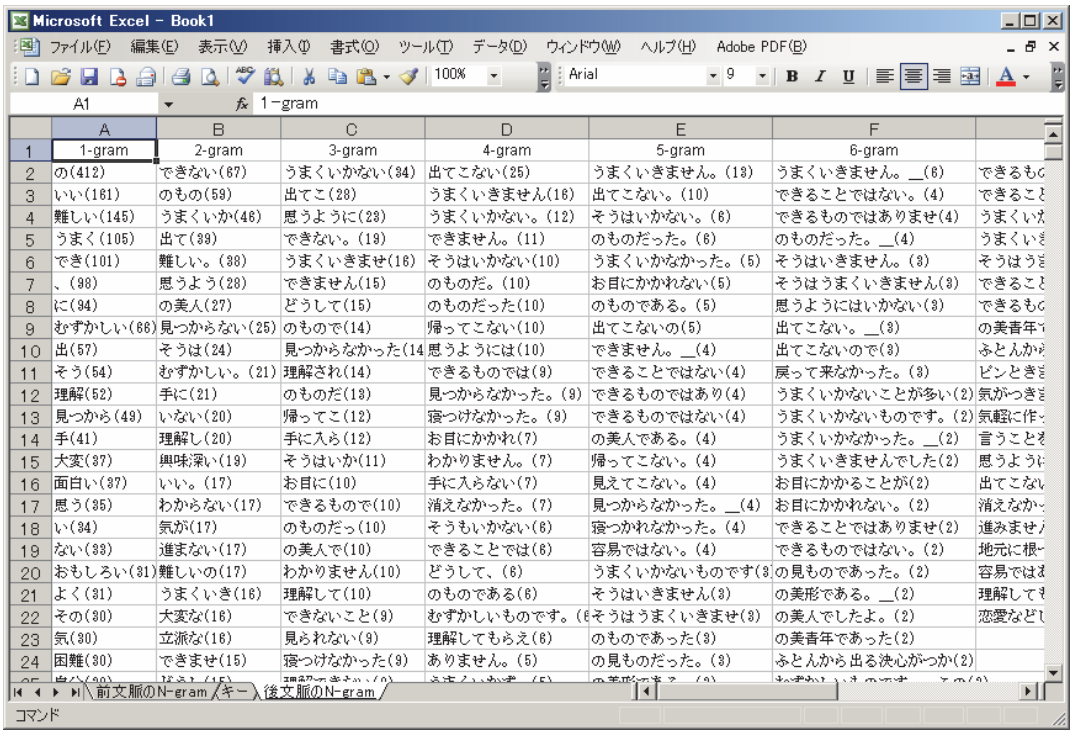
図１　BNAnalyzerのデスクトップアイコン

BNAnalyzerを使ってBCCWJの検索結果からN-gramの一覧を得るには次のようにする。

1) 中納言で［検索結果のダウンロード］によって検索結果（zipファイル）を取得
2) zipファイルをBNAnalyzerのアイコンの上にドラッグ＆ドロップ

これによりエクセルの新しいブックが開かれ、単純版の場合は、検索語（キー）の前文脈の末尾および後文脈の冒頭のN-gram（N＝1～8）が頻度順に表示される。

次に示すのは「なかなか」の検索結果に基づく後文脈のN-gramの一覧である。



	A	B	C	D	E	F
1	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
2	の(412)	できない(67)	うまくいかない(34)	出てこない(25)	うまくいきません。(13)	うまくいきません。_(6)
3	いい(161)	のもの(59)	出てこ(28)	うまくいきません(16)	出てこない。(10)	できることではない。(4)
4	難しい(145)	うまくいか(46)	思うように(23)	うまくいかない。(12)	そうはいかない。(6)	できるものではありませ(4)
5	うまく(105)	出て(39)	できない。(19)	できません。(11)	のものだった。(6)	のものだった。_(4)
6	でき(101)	難しい。(38)	うまくいきませ(16)	そうはいかない(10)	うまくいかなかった。(5)	そうはいきません。(3)
7	、(98)	思うよう(28)	できません(15)	のものだ。(10)	お目にかかれない(5)	そうはうまくいきません(3)
8	に(94)	の美人(27)	どうして(15)	のものだった(10)	のものである。(5)	思うようにはいかない(3)
9	むずかしい(66)	見つからない(25)	のもので(14)	帰ってこない(10)	出てこないの(5)	出てこない。_(3)
10	出(57)	そうは(24)	見つからなかった(14	思うようには(10)	できません。_(4)	出てこないので(3)
11	そう(54)	むずかしい。(21)	理解され(14)	できるものでは(9)	できることではない(4)	戻って来なかった。(3)
12	理解(52)	手に(21)	のものだ(13)	見つからなかった。(9)	できるものではあり(4)	うまくいかないことが多い(2)
13	見つから(49)	いない(20)	帰ってこ(12)	寝つけなかった。(9)	できるものではない(4)	うまくいかないものです。(2)
14	手(41)	理解し(20)	手に入ら(12)	お目にかかれ(7)	の美人である。(4)	うまくいかなかった。_(2)
15	大変(37)	興味深い(19)	そうはいか(11)	わかりません。(7)	帰ってこない。(4)	うまくいきませんでした(2)
16	面白い(37)	いい。(17)	お目に(10)	手に入らない(7)	見えてこない。(4)	お目にかかることが(2)
17	思う(35)	わからない(17)	できるもので(10)	消えなかった。(7)	見つからなかった。_(4)	お目にかかれない。(2)
18	い(34)	気が(17)	のものだっ(10)	そうもいかない(6)	寝つかれなかった。(4)	できることではありませ(2)
19	ない(33)	進まない(17)	の美人で(10)	できることでは(6)	容易ではない。(4)	できるものではない。(2)
20	おもしろい(31)	難しいの(17)	わかりません(10)	どうして、(6)	うまくいかないものです(3	の見ものであった。(2)
21	よく(31)	うまくいき(16)	理解して(10)	のものである(6)	そうはいきません(3)	の美形である。_(2)
22	その(30)	大変な(16)	できないこと(9)	むずかしいものです。(6	そうはうまくいきませ(3)	の美人でしたよ。(2)
23	気(30)	立派な(16)	見られない(9)	理解してもらえ(6)	のものであった(3)	の美青年であった(2)
24	困難(30)	できませ(15)	寝つけなかった(9)	ありません。(5)	の見ものだった。(3)	ふとんから出る決心がつか(2)

図２　「なかなか」の後文脈のN-gram

この一覧は、「なかなか」の後には「の」「いい」「難しい」（1-gram）、「でき-ない」「の-もの」「うまく-いか」（2-gram）、「うまく-いか-ない」「出-て-こ」「思う-よう-に」（3-gram）などの表現がよく現れることを示している。

circumcollocate版を使えば、例えば語彙素「惜しむ」の検索結果から、「寸暇を惜しんで」「努力を惜しまない」「骨身を惜しまず」のように「惜しむ」を前後からはさむように現れる２表現の慣習的な組合せがあることを知ることができる（circumcollocateの用語・概念については拙論(2010)を参照）。

以上のようなN-gramの頻度情報は、語句のコロケーションを分析する際の手がかりとなり得る。

３　BNAnalyzerによる不自然な分析結果とその原因

３.１　不自然な分析結果

さて、BNAnalyzerを作ったあと、その動作確認のために随意に選んだ少数の検索語の検索結果を分析してみると、一見しておかしいと分かる結果が得られるケースが非常に多いことが分かった。次に２つの例を示す。

	E	F	G	H
1	5-gram	6-gram	7-gram	8-gram
2	見つけたいならＹａｈｏｏ！(25)	見つけたいならＹａｈｏｏ！縁結び(	見つけたいならＹａｈｏｏ！縁結び__(	見つけたいならＹａｈｏｏ！縁結び__］(25)
3	わかった。__「(10)	はわからなかった。__(4)	ダウンロードできます。◆__メルマガ(	ダウンロードできます。◆__メルマガやっ(4)
4	はわからなかった。(9)	ダウンロードできます。◆__(4)	駄目になるるスキルでは(4)	駄目になるるスキルではなく(4)
5	は答えなかった。(8)	ファーストメールを送りました(4)	売り切れてしまうそうなので(4)	売り切れてしまうそうなので、(4)
6	忘れてしまいます。(7)	食卓に出す。〈４(4)	戻ってきた。__「(4)	治療を受けたいときなど）医療(3)
7	戻ってきた。(7)	駄目になるるスキルで(4)	いっぱいになってしまいます。(3)	食卓に出す。〈４人分(3)

図３　「すぐに」の後文脈のN-gram

	D	E	F	G	H
1	4-gram	5-gram	6-gram	7-gram	8-gram
2	」といった(54)	に殉じて自分は(47)	気持ちに殉じて自分は(47)	の気持ちに殉じて自分は(47)	あなたの気持ちに殉じて自分は(47)
3	」などという(53)	のことを自分は(34)	すべてのことを自分は(29)	関するすべてのことを自分は(25)	に関するすべてのことを自分は(25)
4	次のような(53)	の気持ちに自分は(22)	あなたの気持ちに自分は(22)	あなたに自分はその成果を(11)	てあなたに自分はその成果を(11)
5	〉」という(51)	、次のような(20)	に自分はその成果を(11)	の気持ちに殉じて自分が(9)	あなたの気持ちに殉じて自分が(9)
6	ない」という(50)	ている」という(19)	気持ちに殉じて自分が(9)	___【松下幸之助日々の(8)	_____【松下幸之助日々の(8)
7	殉じて自分は(47)	」というような(17)	__【松下幸之助日々の(8)	_と言うように「その(8)	くれればあなたの気持ちに自分は(8)

図４　「言葉」の前文脈のN-gram

いずれの例においても、とうてい一般性を持つとは考えられない「見つけたいならYahoo!縁結び」とか「あなたの気持ちに殉じて」といったN-gramがリストの上位を席巻している。

３.２　原因の調査

上の２例のうち、前者（図３）はYahoo!知恵袋かYahoo!ブログのいずれかのサブコーパスに原因がある可能性が高い。後者（図４）についてはこの分析結果だけからでは事情は分からない。

そこで、「BCCWJ-DVD版」――DVD媒体で頒布されているBCCWJ――とインターネットを用いて調べてみたところ、真相は次の通りであった。

まず、図３に見る「見つけたいならYahoo!縁結び」という高頻度N-gramは、ブログ記事の書き手が書いたものではなく、ヤフー株式会社の運営する出会い系サイトの宣伝用ブログ「そろそろ恋愛しませんか？」（http://blogs.yahoo.co.jp/yjpartnerblog/）の記事に自動的に張られるリンクのタイトル「[結婚相手をすぐに見つけたいならYahoo!縁結び]」の一部であった。リンクの実例は例えばインターネット上のYahoo!ブログの記事http://blogs.yahoo.co.jp/yjpartnerblog/archive/2008/11/11で見ることができる。この記事はBCCWJにサンプルID OY14_29479として収録されている。

図４にある「殉じて」の奇異な用法を含む多数のN-gramはいずれも「世界で一番、誰よりも愛してる人へ」と題されたブログ（http://blogs.yahoo.co.jp/geoburgher/）に頻出するものであった。BCCWJにはこのブログから「殉じて」を含む記事が36件取られており、そこには「殉じて」が計243回も現れ、BCCWJ全体に含まれる「殉じて」計264例の実に92%を占めている。図４に見る、「自分」を含むほかのN-gramも同じブログに現れるものであった。

このように、不自然な分析結果の原因は、主にBCCWJのYahoo!ブログサブコーパスにおける同一データの重複出現にあることが判明した。

インターネットに高頻度で現れる一般性の低い表現は、人間がその都度書いたわけではなく、機械的に生成または複製されたものに過ぎない。言語研究上そのようなものを通常の言語データと同列に扱ってはならないことは明らかであるが、BCCWJ評価の観点から引き続き問われるべきは、コーパスにどのような重複がどれくらい含まれているのかという問題である。

以下においては、BCCWJのYahoo!ブログおよびYahoo!知恵袋の各サブコーパスの特性に

ついて、データ重複の問題を中心に観察する。

４　Yahoo!ブログの特性

４.１　サンプルの完全一致

データの重複と一口に言っても、事例ごとに程度の差がある。まず、サンプル全体が完全に一致するものについて見る。

Yahoo!ブログサブコーパスには52,680件のサンプルが収められているが、調べてみると、うち410件のサンプルについては完全に一致するサンプルが別に存在することが分かった。

410件のサンプルをテキストの同一性に基づいて分類すれば103種類になる。そのうち、もっとも長いものは次の２サンプルで、2,863字の長さである。紙幅の関係で冒頭部分だけを示す。

サンプルID：**OY11_03448＝OY11_03449**

もぅ要らない！ＣＩＡ結成の架空自民党。日本統治体制は８１６年間「亀卜政治」海外では１５０年前～日本の自衛隊員レベルで認識。狂言的天皇と現政府の相互に「責任を取れない劣等感」を【弱点】として毎回軍戦略の【標的】に！米海・陸軍戦略プランに必ず記される。＜ＣＩＡエージェント岸信介が＜内政密告１回１０億円報酬金を貰いつづけ＜／＜５５年自民党結成。＜ＣＩＡエージェント岸信介党内に３５億円ばらまき＜／＜５７年首相就任。＜５８年岸内閣の弟左藤栄作蔵相は「国政選挙資金」を＜／＜米国大使館を通じて無心。飼い主へ甘え切った岸と佐藤。岸らの「借金の肩代わり」に国民の血税が宛てがわれ＜／＜基地の米軍施設や住居の給水高熱費はタダへ！＜恵まれた環境の広範囲にわたる地域を治外法権の基地へ。＜墜落事故が発生しても「現政府は一切対応無し」へ。＜アメリカが何時でも公開すると言う記録を＜現政府は国民へ「密約」と呼ばせ、わざわざ「密約は無かった」と言う。更に年金記録を抹消し平気でいた現政府はなるほど当たり前！。（後略）

逆に、最も短いのは次の４サンプルで、記号・空白を含めて21字の長さである。

サンプルID：**OY01_00197＝OY01_00544＝OY15_00161＝OY15_01346**

オークション　＞　チケット、金券、宿泊予約

もし空白を文字に含めなければ最短は次の２サンプルで、長さは0字である。ここでは空白を「□」で示している。

サンプルID：**OY13_04823＝OY14_33122**

□□□□□□□□□□□□□□□□□□□□□□□□□□

重複の回数の最も多いのは次のサンプルで、23件のサンプルの内容が一致している。

サンプルID：**OY04_00021＝OY04_00397＝OY04_00410＝OY04_00851＝OY04_01012＝OY04_01588＝OY04_02241＝OY04_02813＝OY04_03116＝OY04_03304＝OY04_03669＝OY04_03826＝OY04_04435＝OY04_04634＝OY04_07415＝OY13_00195＝OY15_00223＝OY15_00745＝OY15_08458＝OY15_08930＝OY15_10185＝OY15_11643＝OY15_13183**

…の記事を更新しました。本部はこちら↓ＣＩＡ☆こちら映画中央情報局です

以上の例に含まれる語句――例えば、「もぅ要らない」「チケット、金券」「映画中央情報局」――を中納言の文字列検索で検索してみれば、異なるサンプルIDを持つまったく同じ“用例”が複数表示され、データの重複を容易に確かめることができる。これは以後の重複の事例についても同様である。

参考までに、完全一致するサンプルのIDと字数を――紙幅の関係で字数が500字以上のサンプルに限定して――示せば表１の通りである。

表１　完全一致のサンプル一覧（部分）

サンプルID	字数
OY14_29957＝OY14_34602	540
OY05_02386＝OY05_02853	593
OY11_00268＝OY14_10909	626
OY02_00210＝OY02_00212	646
OY03_02330＝OY03_03987	855
OY01_00055＝OY11_00113	928
OY01_00506＝OY01_00579＝OY02_00125	947
OY01_00180＝OY01_00214＝OY01_00260＝OY01_00552＝OY01_00642＝OY11_00260＝OY11_00385＝OY11_01547＝OY11_01553	1038
OY04_01476＝OY04_01665	1,172
OY01_01375＝OY01_02646	1,365
OY14_40012＝OY14_43821	2,467
OY14_49176＝OY14_49589	2,467
OY11_03448＝OY11_03449	2,863

なお、ここでの調査は、BCCWJ-DVD版のLUWディレクトリにあるタブ区切り形式データより復元したテキスト（文字コードはShift_JISに変換）によった。他の方法で調査すれば統計（特に字数）に多少の違いが生じる可能性がある。

４.２　サンプルの部分一致

サンプルが部分的に一致するものには、完全一致に近いものから小部分の一致にすぎないものまでさまざまな段階がある。また、一致・不一致のあり方も事例によって異なる。

部分的な一致を含むサンプルの数は完全一致のそれをはるかに上回るが、類似は程度問題であるので、部分一致を含むサンプルの範囲を明確にすることは原理的に不可能である。また、大量のテキストから部分一致を検出することには現実の処理上限界がある。

ここでは、BCCWJを使用するうえで特に深刻な問題を引き起こす可能性のある、データ一致の程度のはなはだしい事例を２つ見る。

第１の事例は、「犬幼稚園Ｂｕｄｄｙ Ｄｏｇ」のブログ（http://blogs.yahoo.co.jp/lovedog111222）から取られた以下のサンプル38件である（一部に完全一致のサンプルを含む）。

OY05_00030, OY05_00066, OY05_00447, OY05_00486, OY05_00699, OY05_01030, OY05_01096, OY05_01213, OY05_01840, OY05_02017, OY05_02075, OY05_02130, OY05_02232, OY05_02252, OY05_02286, OY05_02335, OY05_02386, OY05_02461, OY05_02834, OY05_02853, OY05_03041, OY05_03152, OY05_03182, OY05_03316, OY05_03593, OY05_03641, OY05_03759, OY05_04364, OY05_04503, OY05_04574, OY05_04712, OY05_04887, OY05_04944, OY05_05154, OY05_05424, OY05_06006, OY05_06217, OY05_06422

これらのサンプルにおいては、１文ないし数文のまとまりが、異なる組み合わせや順序において現れる。例えば、次はOY05_02853＝OY05_02386のテキストであるが、

サンプルID：**OY05_02853＝OY05_02386**

犬幼稚園Ｂｕｄｄｙ Ｄｏｇに愛犬を預ける飼い主さん。送り迎えの際、愛犬たちがじゃれあう広場でお茶をしつつ、その間に情報交換タイムが始まります。しつけや健康管理の話はもちろん、去勢手術や避妊手術について、詳しい説明や報告をしたり、不安な事を質問したり。フードやおやつの選び方・与え方、犬が喜ぶおもちゃや、留守中に便利なグッズについてなど、話題は多岐にわたります。犬の幼稚園「Ｂｕｄｄｙ Ｄｏｇ」は自由登校システムのため、毎回少しずつ違うメンバーが顔を合わせて、情報は増える一方。みんな子（犬）育て真っ最中で、お互いに相談もしやすいようです。仔犬は本来、

目を輝かせ好奇心旺盛・天真爛漫で元気すぎる程です！落ち着きがない・無関心、無反応・それは仔犬の本質ではありません。犬達は犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇで仲良くじゃれあったり、時にはおもちゃを取り合ってみたり・・遊び疲れて寄り添って眠っていたり・・愛くるしい表情をいっぱい見せてくれます。その姿は本当に純粋で愛しい程です。『犬の社会性』を身につけることが、将来に良い子になる秘訣。「三つ子の魂百までも」は、人間も犬も一緒なんですね。犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇは、仔犬にとって世界を広める第一歩でもあるわけです。犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇは、きっとあなたと愛犬の間に新しい発見と更なる楽しみをもたらしてくれるはずです。お気軽にご相談ください。

同じブログから取られた他のサンプルOY05_02286＝OY05_02232においては、このテキストの後ろに短いテキストが付け加えられている。また、サンプルOY05_04364では冒頭に短いテキストが付け加えられ、かつ、「三つ子の魂百までも」以下のテキストが削除されている。

上で色分けして示したテキストの各部分は、このブログにおいてしばしばテキストの構成要素として――ときには、わずかに修正された形で――繰り返し現れる。それを“要素テキスト”と呼ぶことにすれば、次に示すサンプルでは、上記テキストの４つの要素テキストが使われ、かつ、波下線を施した別の要素テキストが付け加えられた形になっている。また、最初の要素テキストの冒頭には「仔」が加えられている。

サンプルID：**OY05_02461**

仔犬達は犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇで仲良くじゃれあったり、時にはおもちゃを取り合ってみたり・・遊び疲れて寄り添って眠っていたり・・愛くるしい表情をいっぱい見せてくれます。その姿は本当に純粋で愛しい程です。『犬の社会性』を身につけることが、将来に良い子になる秘訣。「三つ子の魂百までも」は、人間も犬も一緒なんですね。犬幼稚園　Ｂｕｄｄｙ　Ｄｏｇは、仔犬にとって世界を広める第一歩でもあるわけです。犬幼稚園は犬をしつけるのではなく、犬とじゃれあうことにより社会性を形成する。家族以外の人と接することにより人への信頼・服従を確立する。飼い主は飼い主として必要な知識を学んでいただく所です。犬幼稚園　ＢｕｄｄｙＤｏｇはきっとあなたと愛犬の間に新しい発見と更なる楽しみをもたらしてくれるはずです。お気軽にご相談ください。

程度のはなはだしい部分一致の第２の事例は、先に見た「世界で一番、誰よりも愛してる人へ」（http://blogs.yahoo.co.jp/geoburgher/）のブログの記事である。このブログからはBCCWJに少なくとも次の36件のサンプルが取られている。

OY14_04922, OY14_07061, OY14_09969, OY14_14614, OY14_15863, OY14_17848, OY14_22996, OY14_32704, OY14_33189, OY14_34427, OY14_34837, OY14_34962, OY14_35316, OY14_36252, OY14_38580, OY14_38725, OY14_40012, OY14_42324, OY14_42502, OY14_42913, OY14_43821, OY14_44162, OY14_44932, OY14_45460, OY14_46975, OY14_47226, OY14_47461, OY14_49176, OY14_49589, OY14_50563, OY14_51147, OY14_51273, OY14_53034, OY14_53047, OY14_53827, OY14_54064

このうちの１件のサンプルOY14_53034の冒頭の２文――第２の句点までの内容――だけを示せば次の通りである。

サンプルID：**OY14_53034**

（３からつづく）今夜から明日にかけて一部の地域で雨に注意する必要があり今日は昨日よりさむいのでさむさにも注意する必要があり花粉症も明日は少ないとなっているけれどまだ終息しないようで黄砂についても気をつけなければならないようでインフルエンザもまだかなりはやっていて個人的な経験で申し訳がないのだけどこの時期も風邪を引きやすく油断がならないからどうかさむさによる悪影響や花粉症による悪影響や雨や急な強い雨による悪影響やインフルエンザや風邪による悪影響が絶対に絶対に絶対に絶対に絶対に絶対に絶対に防がれ絶対にあなたが暖かい健康でゆとりのある毎日を過ごしていてほしいとほんとうにほんとうにほんとうにほんとうにほんとうにほんとうにほんとうに強くおもう。永久に絶対にどんな場合も現実としても可能性としても当為としてもあらゆるすべてのことについてあなたの気持ちに殉じてあなたのためになるように自分は言葉だけじゃなくて実証されているようにどんな犠牲を払っても自分の命を犠牲にしても自分のみがすべての責

任を負って果たして絶対にどんなことでも解決してなんでもできればとおもうのでどうかなにかあればどんなことでもどんな方法でも伝えてくれればとおもう。（後略）

このブログの記事には全般に際立った特徴があり、きわめて類似度が高く特異な印象を与える長文が繰り返し現れる。かと言って完全な機械生成物というわけでもなく、上の引用に見るようにその時節の話題が添えられていたりもする。こうした記事がどのような方法で作られているかは不明であるが、このブログでは重複の多い長文記事が短い時間間隔で続々と掲載されていることも注意を引く。[1]

４.３　その他の問題

上で最後に見た例には、「絶対に絶対に絶対に絶対に～」「ほんとうにほんとうにほんとうにほんとうに～」という同一語句の極端な反復も含まれていた。Yahoo!ブログサブコーパスには同類の事例がほかにも見られる。反復回数の最も多いのは、「眠い」を744回反復した次のブログ記事である。

サンプルID：**OY14_38757**

眠い
眠い
眠い
（中略）
眠い
眠い
眠い眠いねm（黙れ勉強すると必ず眠くなる僕・・・ってかみんな勉強してるとだんだん眠くなると思うよだってずっと座ってたら血の巡りが遅くなる（のか？）じゃんま、どうでもいいけどね（Σ

機械翻訳によって得られた日本語文で構成されたと見られる次のようなブログ記事もある。この記事の省略部分には作成上の不手際による同一の機械翻訳記事の重複も含まれている。

サンプルID：**OY04_02243**

★ハン・ジェソク韓流スターらしいね！チェンカイコー、チャン・ツイィと慈善バザー雪道　[ニュースには朴世縁記者] タレントハン・ジェソクが中国四川省地震災害民たちに温情の手助けを伝えた. ハン・ジェソクは　１９日中国北京ソピテルワンダベイジングホテルで進行された　'たいてい, 中アジアスター中国四川省災害民助け合い慈善バザー会'に参加した. 中国人口福祉基金回主催で開かれた今度バザー会は去る　５月１２日中国四川省で発生した地震で苦痛を経験している子供たちの児童教育基金用意のために企画された公式バザー会でこの日ハン・ジェソクは映画監督チェンカイゴ, ワールドスターザングツイ, シンガポール俳優ウィルリアムなど世界的なスターたちと肩を並べて意味を一緒にした. 前作　ＭＢＣ　'イブのすべて', ＳＢＳ　'硝子くつ'　などで中華圏私の寒流スターとして立地を押し堅めているハン・ジェソクは　"四川省大震災で疾病とひもじさに苦しむ子供たちを見ていつも気が重かった"と　"良い主旨の行事に同参するようになって光栄だ. 使い捨て行事ではない持続的な福祉活動を通じて小さな力でも四川省子供たちと災害民たちの希望を探すのに助けになりたい"と念をおした.（後略）

---

[1] 丸山・柏野・田中(2011)によれば、Yahoo!ブログサブコーパスの作成においては、Yahoo!ブログの元データからサンプルを取得する際、"抽出時点で1,000記事以上あるブログからの記事" であることが条件の１つとされた。発表者の憶測によれば、その方針がサブコーパスにおけるデータの重複を増やす原因となった可能性がある。なぜならば、毎日欠かさず記事を１件書いても1,000件の記事を書くには３年かかり、内容のある記事を個人が人手で書くとすれば相当の勤勉を要する。しかし、機械的な記事生成や定型文の使用に頼る場合には――そしてまた、各種業者の宣伝を目的としたブログのように複数の人間が記事を掲載するブログにおいては――1,000件達成のハードルは格段に低くなる。「世界で一番、誰よりも愛してる人へ」のブログには、本稿を執筆している2012年7月15日の午前（0時～12時）だけでも40件の長文記事が掲載されている。

また、少納言・中納言による検索における“文境界無視”とでも呼ぶべき問題がある。これは特にウェブデータの場合に限ったことではないが、ブログ記事には句点を使わないものが多く、そのような場合、文の境界がないかのように扱われる。

サンプルID：**OY11_06057**

今日も晴れたので散歩に松本城のお堀に二羽の白鳥が仲良く泳いでいたお城もなんか寒そうお城の堀越しに北アルプスの常念岳が白く映えていた緑町まで歩き上土の「花月」ホテルでカレーライスを食べた帰りは縄手通りを通って帰宅　少し欲張って歩いたので左脚の脹脛が痛くなった午後からは曇りになり寒くなった

この記事は実際のブログでは次のように表示される（http://blogs.yahoo.co.jp/sinnshuu28/archive/2008/12/8）。

今日も晴れたので散歩に [見出し]

松本城のお堀に二羽の白鳥が仲良く泳いでいた
　[白鳥の写真]
お城もなんか寒そう
　[松本城の写真]
お城の堀越しに北アルプスの常念岳が白く映えていた
　[常念岳の写真]
緑町まで歩き上土の「花月」ホテルでカレーライスを食べた
　[カレーの写真]
帰りは縄手通りを通って帰宅　少し欲張って歩いたので左脚の脹脛が痛くなった
午後からは曇りになり寒くなった

この例の場合、中納言の短単位検索ではサンプル全体が長い１文として扱われ、「泳い-で-い-た-お-城」や「食べ-た-帰り」といった短単位の連続の“用例”があるものとして処理される。中納言の長単位検索や少納言による検索の場合も同様である。

これはBCCWJにおけるテキスト構造に関わる情報の付与の仕方に関係している。一般論として、行が句点以外の文字で終わる場合、文がそのまま次の行に続く可能性と、文が句点なしでそこで終わる可能性とがある。発表者の断片的な確認によればBCCWJの情報付与では両者の可能性が区別されておらず、その結果として“文境界無視”の現象が生じるものと見られる。

ほかにも細かい問題はあるが、日本語研究上特に問題となり得ると思われるYahoo!ブログサブコーパスの特性としてこれまでに気付いたのは以上である。

５　Yahoo!知恵袋の特性

５.１　サンプルの完全一致

Yahoo!知恵袋サブコーパスにはYahoo!ブログサブコーパスに見られるほどのデータの重複はないが、サンプルの完全一致が１例だけあった。

サンプルID：**OC08_00706＝OC08_05640**

国勢調査って何のためにするの？国勢調査の人口は、議員定数や地方交付税算定の基準など、法定人口として利用されます。また、男女・年齢別人口・産業別帯・高齢者のいる世帯などの統計は、国や市町村の社会福祉雇用政策、環境設備政策、号際対策などの行政資料として利用されます。

Yahoo!知恵袋は、利用者どうしの知識の交換を目的とする問答形式の掲示板である。中納言では質問と回答の境界が考慮されず、問答が上記のような一続きのテキストとして扱われるが、上のサンプルの場合、冒頭の「国勢調査って何のためにするの？」が質問で、残

りの部分が回答である。また、Yahoo!知恵袋のサイトで１つの質問に対して複数の回答があった場合は、BCCWJにはそのうち「ベストアンサー」とされたものだけが回答として収録されている（丸山・柏野・田中(2011)）。

それにしても、問答の完全一致がなぜ生じたのか。上記のようにそれなりの長さを持つ問答が一字一句違わず２度繰り返されることは常識的に考えがたい。

発表者の推測をあえて通俗風に表現すれば、これはYahoo!知恵袋を運営するヤフー株式会社が“自作自演”の問答を半ば不手際で２度掲載してしまったものである。

そのことを理解するには、BCCWJのYahoo!知恵袋サブコーパスのデータが何物であるかを知る必要がある。BCCWJのマニュアルには正確な記載がないが、BCCWJに収録されているのは、実は“Yahoo!知恵袋”のデータではなく、その準備段階の“Yahoo!知恵袋ベータ版”のデータである。

Yahoo!知恵袋は2005年11月7日に正式運用を開始したが、その約１年半前の2004年4月7日にはその試験段階としてYahoo!知恵袋ベータ版の運用が始められた。[2] Yahoo!知恵袋ベータ版では必ずしも実際の利用者が問答のやり取りをしたわけではなく、一部の質問や回答はヤフー株式会社によって用意されたのであった。[3]

現行のYahoo!知恵袋のサイトにはYahoo!知恵袋ベータ版の時期の問答も併せて掲載されており、当該の２サンプルは今も次の異なる問答として参照することができる。

http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q116243661
（質問日時：2005/9/23 10:27:12、解決日時：2005/9/23 10:35:34、回答数：1）
http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q136197063
（質問日時：2005/9/26 13:42:59、解決日時：2005/9/27 09:38:14、回答数：7）

Yahoo!知恵袋の正式運用開始を間近に控えた2005年9月にあって、１度目の掲載は問答の模範例を示すために行われ、その３日後の２度目の掲載は複数回答からのベストアンサーの選び出しの実験あるいは例示の目的で行われた――ただし、同一の問答を使ったために重複が生じてしまった――といったところかと推測される。

以上、本発表の目的の中心を外れるが、“Yahoo!知恵袋”サブコーパスに収められたデータが正確には“Yahoo!知恵袋ベータ版”のデータであることを明らかにする目的も兼ねて、サンプルの完全一致の生じた背景に関する推測を述べた。

５.２　サンプルの部分一致

サンプルの部分一致についても、Yahoo!知恵袋サブコーパスにはYahoo!ブログサブコーパスにおけるような極端なデータの重複はない。ただし、相手の発言をコピーして引用するインターネット掲示板の慣習がデータの重複を多数生じている。

サンプルID：**OC02_07077**
ＤＶＤをプレーヤーに入れたところ、全く動きません。以前正常に作動したプレーヤーもどれも動かなくなりました。原因は何でしょうか。ちなみに６月中旬にＮｏｒｔｏｎを更新しています。それ以降に使えなくなりました。＞正常に作動したプレーヤーもどれもどれもって、何台あるのですか？ウィルスではないですか。スキャンして下さい。

[2] それぞれの日付は http://chiebukuro.yahoo.co.jp/docs/whats2004.html、http://chiebukuro.yahoo.co.jp/docs/whats2005.html による。
[3] より詳しくは http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1011740930 などを参照。

また、この問答はYahoo!知恵袋のサイトでは次のように表示されるものであるが（http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q135001663）、

質問：
ＤＶＤをプレーヤーに入れたところ、全く動きません。
以前正常に作動したプレーヤーもどれも動かなくなりました。
原因は何でしょうか。
ちなみに６月中旬にＮｏｒｔｏｎを更新しています。
それ以降に使えなくなりました。

回答：
＞正常に作動したプレーヤーもどれも
どれもって、何台あるのですか？
ウィルスではないですか。スキャンして下さい。

4.3で見た“文境界無視”の事例と共通の理由により、中納言による検索では「どれ-も-どれ」という短単位連続の“用例”があるものとして扱われる。

６　おわりに――BCCWJ使用上の教訓

BCCWJの検索結果に基づくN-gram分析ツールの作成が契機となって明らかになった、BCCWJのウェブデータに含まれるデータ重複の問題ほかについて粗い調査・分析を行ってみた。

Yahoo!ブログサブコーパスにおけるデータの重複は、用例の頻度に着目する研究にしばしば破壊的な影響を与える。Yahoo!ブログサブコーパスのサンプル52,680件に完全一致の相手を持つサンプルが410件含まれるというのは一見小さな比率のようでもあるが、上で見た「殉じて」に関わる事例などが示す通り、全体の用例数が少ないときにデータ重複による“用例”が多数あれば、分析は致命的にゆがんだものとなってしまう。そして、サンプルの部分一致は完全一致よりはるかに多いので、問題は410件のサンプルにとどまらない。Yahoo!知恵袋サブコーパスも相対的に軽度ながら同様の問題をはらむ。機械生成や単なる複製などによる表現の“用例”をそれと知らずに通常の用例と同列に扱ってしまうことのないよう十分な注意が必要である。

今回の調査・分析から得られたBCCWJ使用上の教訓を最後に一般的な形にまとめておく。均衡性の考慮から明確な基準に基づいて収集された出版物（特に書籍）のデータに、それとは採録の手順が異なるだけでなく言語的に異質でデータの重複や信頼性の問題も大きいYahoo!ブログ、Yahoo!知恵袋のデータを単純に加えて使うのは、“均衡”コーパスのまっとうな用法ではない。中納言では検索対象――少納言ではメディア／ジャンル――を目的に応じて適切に指定したうえで検索しなければならない。そして付言すれば、話しことばの書き起こしであり、やはり異質性の高い国会会議録のデータも、“書き言葉”コーパスの他の部分との不用意な併用は避けるべきであろう。

文 献

田野村忠温(2010)「日本語コーパスとコロケーション――辞書記述への応用の可能性――」『言語研究』138, pp.1-23.

丸山岳彦・柏野和佳子・田中牧郎(2011)「第３章　サンプリング」『「現代日本語書き言葉均衡コーパス」利用の手引』第1.0版（BCCWJ-DVD版所収のPDF文書）, pp.21-38. 国立国語研究所コーパス開発センター.

漢字四字成語の受容とその延命

砂岡和子（早稲田大学政治経済学術院）†
羅鳳珠（台湾元智大学中国語文学系）††
王雷（北京大学計算語言学研究所, 北京大学英語系）†††
姜柄圭（韓国西江大学中文系）††††
松崎実夏（千駄ヶ谷日本語教育研究所）†††††

Adoption of the Chinese Four-Character Idioms and their Surviving in the Contemporary Written Japanese

Sunaoka Kazuko（School of Political Science and Economics,Waseda University）
Lo Feng Ju（Dept. of Chinese Linguistics &Literature,Yuan Ze University）
Wang Lei (Dept. of Institute of Computational Linguistics& English, Peking University)
Kang Byeong Kwu (Dept. of Chinese Culture, Sogang University)
Matuzaki Mika(Sendagaya Japanese Institute)

１．概要

筆者らは国際共同プロジェクト事業の一環として、中台韓日英マルチリンガル成語辞典を編纂している[1]。中台韓日それぞれの地域を代表する現代語コーパスに基づき、最大公約数の常用成語を選定後、言語文化情報を付与し、英語訳を加え、各言語間のマルチ検索と各地域のコーパスとの閲覧リンク可能な Web 版成語辞典の公開を目指している。本文は主に日本研究チームが、日本国立国語研究所「現代日本語書き言葉均衡コーパス（BCCWJ）」の検索ツール「中納言」を利用し日中同形成語の抽出作業を行った作業経過に基づき、現代日本語に於ける漢字四字成語の受容とその延命状況について報告する。

２．中日韓英マルチリンガル成語辞典の構築

四字成語は東アジア漢字圏の智恵と文化を凝縮する表現形式のひとつであり、簡潔で口調の良い言語形式と相まって古典文学から新聞記事まで旺盛な生命力を維持している。現代中国語で常用される成語は 2000-3000 語に上るとされる[劉長征他（2007）]。長らく漢字圏にあった韓国、日本では四字成語を含め漢字語句の使用頻度が低下しているとはいえ、故事成語やことわざの愛好に見られるように、根強い延命力がある[砂岡和子他（2011b）]。

†ksunaoka@gmail.com ††gefjulo@saturn.yzu.edu.tw †††wangleics@pku.edu.cn
††††kbg4335@gmail.com †††††songqi.shixia@gmail.com

[1] 台灣蔣經國國際學術交流基金會「歷代語言知識庫建置計畫」研究代表：台灣元智大學羅鳳珠〔網路展書讀〕http://cls.hs.yzu.edu.tw に概要公開

中台韓日英マルチリンガル成語辞典は、中国古典籍資源を母体とする東アジア歴代語言データベースを基に研究や教育に共同利用する目的で構築中の多国籍プロジェクトである。

成語辞典の編纂作業は以下 4 段階で進行する[砂岡和子,羅鳳珠（2011a）参照]。

［第一段階］成語辞典の母資源となる中台成語資源の構築。大陸中国の成語母集団は北京大学計算言語学研究所開発の「成語知識庫」で、同所の「現代漢語語法信息詞典」所収の成語約 5000 条を核に、「人民日報」ならびに『成語大全』『学生成語詞典』など中国出版の代表的成語辞書 5 冊から 3 万余項目の現代中国語成語を収集し、これに頻度情報、語法属性、品詞、異形、類語、反義語、典故、釈義など 16 種の言語属性を記したコーパスである[A 北大三萬詞簡體成語庫］[2010 王雷］[図１]。このうちもっともポピュラーな中国成語 1000 語は書籍出版された[A’’北大 1000］[2012 王雷]。これを台湾で使われる繁体字に変換校正し[A’北大三萬詞繁體成語庫]、台湾中央研究院「平衡語料庫[バランスコーパス]」[B 台灣中央研究院平衡語料成語庫]、ならびに台湾教育部公布の成語データから台湾香港地域の常用成語約 1 万語をリストアップした［C 台灣教育部成語詞條]。最後に［A］［B］［C]資源間で重複率の高い成語約 3000 語を選定し、これ中国語母集団の中核的成語資源データとした[D 三千核心成語］[2]。

［第二段階]は漢字受容圏の成語資源構築と、母集団資源との共通成語の選出作業である。日韓成語の選定に当たっては、それぞれの母語の表現形式や受容条件を反映するよう、各地域を代表する現代語コーパスから漢字成語を抽出し、上掲[D]資源との照合作業を進めた。日本語の常用成語は名古屋大学佐藤理史研究室作成「基本慣用句五種対照表」所収の成語選定基準を参照し、日本国立国語研究所「現代日本語書き言葉均衡コーパス（BCCWJ)」［E]用のオンラインコンコーダンサである「少納言」「中納言」を利用し、上掲[D 三千核心成語]とのマッチング作業に当たった。韓国は国立国語院「世宗コーパス」から韓国語の成語を抽出する[F]。

［第三段階]は[D]の成語に典故、例文などの百科知識、およびその使用頻度、分布密度、難度レベル、語法属性、品詞機能などデジタル言語情報を付与する。台湾元智大学[網路展書讀]にはすでに「三国演義」「水滸伝」「紅楼夢」など歴代文学資料が構築されており、これら白話小説中の成語を抽出してマルチ辞書に加える。図２に成語知識庫構築図を示す。

［第四段階]で英語、韓国語、日本語の翻訳をつけ、各言語間のマルチ検索と各地域コーパスとリンクして閲覧可能な Web 版成語辞典として公開する。

３．現代日本語常用漢字成語の選定

既述のように日本チームは日本語常用成語の選定あたり、日本語基本慣用句データベースとして定評のある名古屋大学「基本慣用句五種対照表（以後「基本慣用句表」)」所収の成語を被検出語とした。BCCWJ コーパスには成語を抽出するための言語標識が無い。BCCWJ の DVD 媒体でマッチングしようと試行錯誤したが、異形の多さに頓挫した。

「基本慣用句表」は 1982―2006 年にかけ日本で出版された辞書から慣用句を抽出し、「慣

[2] ［劉長征他（2007）]に拠ると現代中国語の常用成語数は 2000-3000 とされる。

序号	成语	级别	全拼音	近义	反义	直译	释义	出处	例句	频度	来源
13325	艰苦奋斗	1	jian1 ku3fen	艰苦创业、	花天酒地	to work arduous	艰苦：艰难困苦。不怕艰难困苦，进行英勇顽强	毛泽东《中国革命战争的战	我们要"/"是我们民族的传统美德	33066	yu_i
37616	自力更生	1	zi4li4geng1	自食其力、	仰人鼻息、	regeneration throug	自力：依靠自己的力量；更生：重新获得生命。指	孙中山《中国问题的真解决	我们需要"/我们"克服了一个又一	30981	yu_i
25002	实事求是	1	shi2shi4qiu	脚踏实地	有名无实、	be practical and rea	求：研究。指从实际情况出发，不夸大，不缩小	东汉·班固《汉书·河间献王刘	你要"地说/说话"/"的态度/"有助	26373	yu_i
21005	千方百计	1	qian1 fang1	想方设法	无计可施、	by every possible w	想尽或用尽一切办法	宋朱熹《朱子类语·卷三十五	各单位"开辟财源/"增加收入	22341	yu_i
22276	全心全意	1	quan2xin1 q	一心一意、	三心二意	with all one's h	全：整个。投入全部精力，一点没有保留	茅盾《鼓吹集》："体验生活的时候，就应该全心全意生活，把我		20159	yu_i
13286	坚定不移	1	jian1 ding4b	坚持不懈	动摇不定、	to insist on firmly	移：改变。稳定坚强，毫不动摇。形容主张、观	宋司马光《资治通鉴·唐文宗开成五年》："推心委任，坚定不移，		16005	yu_i
33841	因地制宜	1	yin1 di4zhi4	因势利导	一成不变	to take actions	因：随顺；制：制定；宜：适当。根据各地的具体情	汉·赵晔《吴越春秋·阖闾内传》："夫筑城郭，立仓库，因地制宜，		14410	yu_i
6926	独立自主	1	du2li4zi4zh	自食其力、	仰人鼻息、	be one's own mas	独立：不依赖外力，不受外界束缚；自主：自己作	茅盾《子夜》十六："最初是买办，然后是独立自主的老板，然后		11271	yu_i
3912	成千上万	1	cheng2qian	千千万万、	寥寥无几、	thousands upon t	成：达到一定数量；上：达到一定程度。形容数量	清·文康《儿女英雄传》第30回："他看着那乌克斋、邓和公这班人		8886	yu_i
13284	坚持不懈	1	jian1 chi2bu	锲而不舍、	半途而废	to persist to th	懈：放松。坚持到底，毫不松懈	《清史稿·刘体重传》："遇大雨	"地努力，一定能学到很多东西/"的	8837	yu_i
35820	战无不胜	1	zhan4wu2b	攻无不克、	一触即溃、	invincible	形容强大无比，可以战胜一切	西汉·刘向《战国策·齐策二》："战无不胜而不知止者，身且死，爵		8820	yu_i
11295	轰轰烈烈	1		烈烈轰轰、	冷冷清清、	be dynamic and vigo	轰轰：象声词，形容巨大声响；烈烈：火焰盛燃的	宋文天祥《沁园春·题张许双	孙中山《革命原起》："而黄花岗七十	8154	cydq
21273	前所未有	1	qian2suo3w	前无古人、	史不绝书、	unprecedented	从来没有过的	宋·徐度《却扫编》下卷："而邓枢密洵武以少保领院而不兼节钺		6876	yu_i
11897	欢欣鼓舞	1	huan1 xin1 g	兴高采烈	黯然销魂	be great overjoyed a	欢欣：快乐而兴奋；鼓舞：振奋。形容非常高兴	宋苏轼《上知府王龙图书》："自公始至，释其重荷……是故莫不		6684	yu_i
26806	滔天罪行	1				One's crimes cry al	滔天：漫天。比喻罪恶极大。			6467	szcy
8232	丰富多彩	1	feng1 fu4du	丰富多采、	单调乏味	be rich and varied	内容丰富，花色繁多	刘绍棠《人民是母亲》："我的家乡的风光景色和风土人情，将更		5900	yu_i
3674	朝气蓬勃	1	zhao1 qi4pe	生气勃勃、	老气横秋、	be full of vigor	朝：早上；蓬勃：旺盛的样子。形容充满了生命和	邓小平《中国共产党第十二次全国代表大会开幕词》："党的最高		5811	yu_i
18803	明目张胆	1	ming2mu4z	明火执杖	鬼鬼祟祟	to open one's eyes	明目：睁开眼睛；张胆：放开胆量。形容有胆识，	《晋书·王敦传》："今日之事，	这个人"地干坏事/不敢"/"的活动，	5641	yu_i
33879	阴谋诡计	1	yin1 mou2g	诡计多端	光明正大	intrigues and plots	阴谋：诡秘的计谋。指暗地里策划坏的害人的	曾朴《孽海花》："大家如能个	这个人搞"/"失败了	5585	yu_i
33486	以身作则	1	yi3shen1 zu	言传身教、	以身试法	to set an example b	身：自身；则：准则，榜样。以自己的行动做出榜	春秋·鲁·孔子《论语·子路》："	领导要"/"很重要/家长"	5239	yu_i
32734	一如既往	1	yi1 ru2ji4wa	始终如一、	一反常态、	to continue as alwa	一：完全；如：像；既往：从前。指态度没有变化，	苏叔阳《第二次握手》："清末	我们应该"努力奋斗/我们"共同奋	5106	yu_i
35999	针锋相对	1	zhen1 feng1	针锋相投	逆来顺受	(in action, stra	针锋：针尖。针尖对针尖。比喻双方策略、论点	宋释道原《景德传灯录》第25卷："夫一切问答，如针锋相投，无		4998	yu_i
24609	声势浩大	1	sheng1 shi4	大张旗鼓、	无声无息、	powerful and dynam	声势：声威和气势；浩：广大。声威和气势非常	明·施耐庵《水浒传》："如今宋江领兵围城，声势浩大，不可抵敌		4987	yu_i
37449	卓有成效	1	zhuo2you3	立竿见影、	无济于事、	with outstanding su	卓：卓越。有突出的成绩和效果	明·王守仁《申行十家牌法》："若巡访劝谕著有成效者，县官备礼		4805	yu_i
2686	不屈不挠	1	bu4qu1 bu4	坚忍不拔、	卑躬屈膝、	indomitable	屈：屈服；挠：弯曲。不屈服。形容不畏强暴和困	东汉·班固《汉书·叙传下》："	面对毒刑拷打，他"/"的精神/要"地	4695	yu_i
726	百家争鸣	1	bai3jia1 zhe	各抒己见、	噤若寒蝉、	contention of a hun	百家：原指先秦时代各种思想流派，后指各种政	东汉·班固《汉书·艺文志》："凡诸子百家，……蜂出并作，各引一		4484	yu_i
35547	再接再厉	1	zai4jie1 zai4	再接再砺	得过且过	to make persistent	再：继续；接：接触；厉：磨快。指公鸡相斗，每次	唐·韩愈《斗鸡联句》："一喷一	你们一定要"再创奇迹/希望你们"	4438	yu_i
30707	欣欣向荣	1	xin1 xin1 xia	朝气蓬勃、	日暮途穷、	(flowers in spring) t	欣欣：草木茂盛的样子；荣：茂盛。形容草木长得	晋·陶潜《归去来辞》："木欣欣	你们的事业"真令人高兴/"的景象	4356	yu_i
5803	当务之急	1	dang1 wu4z	燃眉之急、	一拖再拖、	urgent matter; p	当：当前；务：应该做的事。当前急切应办的要事	战国·邹·孟轲《孟子·尽心上》："知者无不知也，当务之为急。"		4196	yu_i

図 1「北大三萬詞簡體成語庫」部分

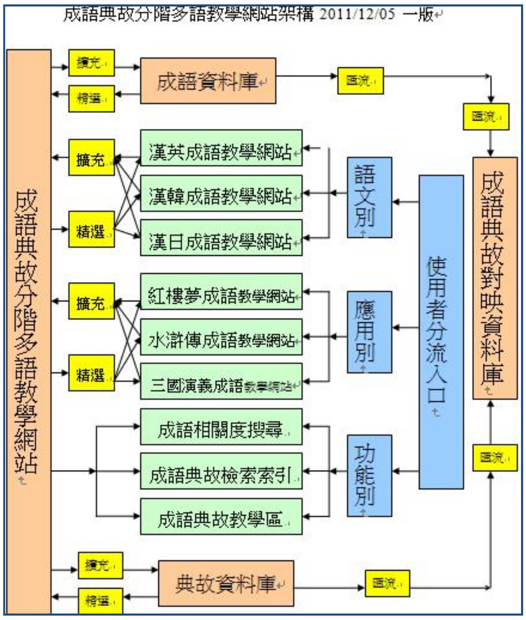


図 2 成語コーパス構築図

用句 1918 語」「ことわざ 344 語」「四字熟語 265 語」「故事成語 63 語」「その他不明 1038 語」に分類､計 3628 の慣用的表現を対照表としたデータベースである｡大半は“我に返る”“わき目も振らず”“目を丸くする”“弱音を吐く”など和語の「慣用句」で、漢字成語よりこちらのほうが多い[図 3 日本名大和語枠参照]。

我々はまず「基本慣用句」中、四字漢語を多く含む「四字熟語」と「故事成語」の計 328 語を被検索対象語とし､「少納言」「中納言」で上掲[D 三千核心成語]とのマッチングを開始した。[D]には中台母集団成語データベース[A] [B] [C]間で重複率の高い成語約 3000 語に対し、重複回数に応じ 4 レベルに分けてある。最初に[D] 中の重複回数 4-3 回の上位中国成語 2005 条について BCCWJ でマッチングした結果、ヒットした日本成語は僅か 211 語、全体の約 10%に留まった。

ヒット率低迷の原因は、日本語の漢字四字成語が少ないことと、日中成語の使用文字や語形の異なりのためである。成語に限らず日本語の四字漢語は二字＋二字の緩い連合形式から成るものが全体の 90 パーセント以上を占め、一語としての凝縮力が弱い[3]。中国の成語は“自相矛盾”“完璧歸趙”“画蛇添足”“杞人憂天”“不可思議”のように四字が典型であるが、日本語ではそれぞれ“矛盾”“完璧”“蛇足”“杞憂”“不思議”のように基幹二字もしくは三字となる。四字でも“三顧の礼（中国語原文は“茅廬三顧”)”“背水の陣（同“背水一戰”)”のように平仮名混じりや、“枚挙にいとまがない（中国語は“不勝枚舉（不可枚

[3] ［野村雅昭（1974）］の調査は約 19 万新聞語彙の四字漢語全体についてであるが, 日本語四字成語にも当てはまる特徴と言える.

舉、不遑枚舉とも)”など日本語の統語法に則して形を変えた成語例は、それこそ枚挙にいとまがない。このため日本語の四字成語は形式上、異形や類語が極めて多く[1982 宮地裕]、マッチングの障害となる。韓国語の成語も同様の傾向を示す[2011 姜柄圭]。

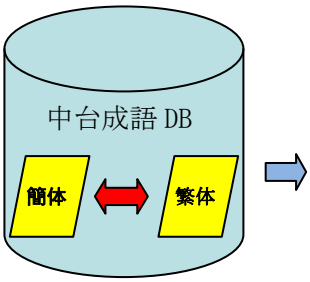
日中同形で抽出できる成語数には限界があるため、次に「基本慣用句」の日本語慣用句を中国語に翻訳し、中台成語データベースの［A］[B]［C］をダイレクトにマッチングした。成語の中訳に当たって各種日中成語辞書や、北京大学が開発する類義語検索サイト[関連URL 参照]で類似表現から適訳を検出した。すでに BCCWJ とのマッチングで抽出済みの 211 語と併せ、最終的に「基本慣用句」所収全 3628 句のうち 1108 語の中日成語ペアを選定した。

４.「少納言」「中納言」による日本語常用漢字成語の抽出

われわれは BCCWJ 現代日本語コーパスを他言語資源とのマッチングデータとしても利用し、「少納言」「中納言」を多言語検索ツールに利用した。以下、「少納言」による日中漢字成語検索と、BCCWJ によるマッチングの過程をチャートにまとめる[4]。実際に多用したのは「少納言」より高度な検索機能を備える「中納言」のほうで、「少納言」の 500 項までの検索結果表示制限も無く、複雑な結果を得ることができる。ただ「中納言」は日本語を理解しない外国人には利用申請手続きが難しいため、アクセスが簡単な「少納言」での利用で代表する。

４-１　前処理

マッチング作業用に整理加工したデータを準備する。中台成語資源[D 三千核心成語]は、データを中国大陸の簡体字と台湾香港地区の繁体字で併記し、どちらの字体でもできる。BCCWJ データ、「少納言」「中納言」とも繁体字であれば検索成功率が高いため、簡体字成語の場合は[D]で対応する繁体字同形語を取得する。このほかデータには各資源の[重複回数]や原データのコード番号、頻度情報などが記され、随時データの復元処理を担保する。図 3 は中台成語 DB[データベース]から簡体字形成語“新陈代谢”とその繁体字形“新陳代謝”を特定し、字体変換する場面。



1	0 (序號不要更動)	簡体	繁体 (紅色是補上的繁體詞表)	重疊次數	北大1000編号	北大3000編号	台灣教育部成語詞條編号	三千核心成語編号	三千核心成語頻率
576	575	与众不同	與眾不同	3	0	402	3547	342	34
577	576	取而代之	取而代之	3	0	342	2108	292	33
578	577	新陈代谢	新陳代謝	3	0	347	1279	294	33
579	578	一模一样	一模一樣	3	0	398	2097	338	33
580	579	应有尽有	應有盡有	3	0	273	1631	229	32

図 3

４-２　検索作業

繁体字の成語を「少納言」「中納言」で検索。それぞれ検索条件を指定できる[図 4]。

[4] チャート原作者は沈睿(早稲田大学人間科学学術院助手)、原文は中国語。

図 4　繁体文字列の検索場面（左図「少納言」、右図「中納言」）

4－3　検索結果を表示する[図 5]

図 5　「少納言」“新陳代謝”の検索結果

4-4　「少納言」「中納言」による中日漢字成語マッチングの課題

大規模コーパスは特殊文字を含む現地語の文字コード体系を使用する場合が多く、中日漢字のフォントや字体の異なりが、解析器の誤認の原因となる。それ以上に中台日韓の成語の表現形式の異なりはマッチングの最大の障害である。機能面では、「少納言」「中納言」は串刺検索機能や英語 I/F がない。その他の地域の成語コーパスも単言語で記述されており、現時点で検索の実行者は各言語使用者、もしくは当該言語の理解者に限定される。多言語間のマルチ検索にはまだ高いハードルがある。

5．BCCWJ 利用で知る現代日本語漢字四字成語の受容と変容

現代中国語の場合、四字成語の出現率は 5 万語中で 8%弱、1 万語では 0.2%程度という。このうち頻度 100 回の常用成語は 2200 条で全体の 85%を占め、1000 回以上の高頻度成語は 110 条で 2 割弱に過ぎない[劉長征他（2007）]。準拠するデータにより出現頻度の差も大きい。中国成語資源[A]［B］［C］と世宗コーパス[F]、名大「基本慣用句」[G]のトップ 20 に共通な語は“成千上万”“自由自在”“流言蜚語”“優柔不断”など極少数に留まる[表 1]。

現代日本語の四字漢字の成語数は、前掲名大「基本慣用句」では 100 語に満たないが、二漢字、三漢字、もしくはひらがな交じりの広義の漢字成語は 300 語前後を収録する。これを[少納言]で検索すると約 200 語のヒットがあり、その出典ジャンルは、書籍の文学、社会科学、歴史が 9 割を占め[5][表 2]、成語の継承に文学、歴史関連の書籍が不可欠な言語資源で

[5]「完璧」のヒット数は 1836 件と多いが，「少納言」は 500 件までしか検索結果が表示されず出現ジャンルの特定ができないため「検索不能」とした.

あることが分かる。中国の成語は古い典故をもつ語が多いが、現代成語には近世のものも多く、[D 三千核心成語] の出典は時代別に以下の順位となる。

宋（16%）、春秋戰國（15%）、清（14%）、明（11%）、漢（11%）、唐（11%）、元（7%）、南北朝（5%）、當代（4%）、晉（3%）、以下、三國、五代十國、金、隋、十六國、周が各 1%未満。対して日本語には“大同小異”“四面楚歌”“呉越同舟”“ 朝令暮改”“異口同音”など、春秋戰國、漢代に起源をもつ故事成語が多く残り、歴史的文化遺産を継承している。

北大3000[D]	頻度	台灣教育[B]	頻度	台灣平衡語料[C]	頻度	韓国世宗[F]	頻度	日本名大漢字[G]	頻度	日本名大和語	頻度
艰苦奋斗	33066	一毛不拔	27	不知不覺(D)	86	전전긍긍(戰戰兢兢)	98	老若男女	98	輪を掛ける	6639
自力更生	30981	掩耳盜鈴	27	不可思議(VH)	79	명실상부(名實相符)	93	臨機応変	93	我を忘れる	6634
实事求是	26373	一丘之貉	27	理所當然(VH)	74	유언비어(流言蜚語)	81	流言飛語	81	横車を押す	6416
千方百计	22341	囫圇吞棗	27	莫名其妙(VH)	53	자유자재(自由自在)	80	立身出世	80	我に返る	6288
全心全意	20159	班門弄斧	27	前所未有(VH)	52	자초지종(自初至終)	80	油断大敵	80	訳は無い	6252
坚定不移	16005	朝三暮四	27	截然不同(VH)	50	자유분방(自由奔放)	79	優柔不断	79	わき目も振らず	6248
因地制宜	14410	竭澤而漁	27	自然而然(D)	49	이목구비(耳目口鼻)	74	無味乾燥	74	路頭に迷う	6225
独立自主	11271	天衣無縫	27	脫穎而出(VH)	43	동서고금(東西古今)	71	無我夢中	71	目を見張る	6191
成千上万	8886	杞人憂天	27	迫不及待(D)	43	시시각각(時時刻刻)	70	三日坊主	70	目を丸くする	6182
坚持不懈	8837	水落石出	26	不約而同(D)	43	동서남북(東西南北)	69	満場一致	69	目を細くする	6175
战无不胜	8820	名落孫山	26	層出不窮(VH)	41	방방곡곡(坊坊曲曲)	69	本末転倒	69	目を通す	6147
轰轰烈烈	8154	胸有成竹	26	無可奈何(VI)	41	비일비재(非一非再)	68	傍若無人	68	世を去る	6135
前所未有	6876	指鹿為馬	26	依依不捨(VH)	39	의기양양(意氣揚揚)	68	暴飲暴食	68	目を回す	6134
欢欣鼓舞	6684	舉一反三	26	自由自在(VH)	39	신진대사(新陳代謝)	68	不眠不休	68	弱音を吐く	6128
滔天罪行	6467	江郎才盡	26	總而言之(Dk)	38	좌지우지(左之右之)	67	百発百中	67	目を白黒させる	6123
丰富多彩	5900	青出於藍	26	成千上萬(Nega)	36	우유부단(優柔不斷)	66	和洋折衷	66	寄ると触ると	6121
朝气蓬勃	5811	口若懸河	25	不遺餘力(VI)	35	각양각색(各樣各色)	66	理路整然と	66	目を凝らす	6110
明目张胆	5641	驚弓之鳥	25	與眾不同(VH)	34	일사불란(一絲不亂)	63	理路整然	63	目を配る	6103
阴谋诡计	5585	含沙射影	25	大街小巷(Nc)	33	유명무실(有名無實)	63	竜頭蛇尾	63	目を掛ける	6096
以身作则	5239	一敗塗地	25	取而代之(VB)	33	요지부동(搖之不動)	63	八方美人	63	目を疑う	6078

表 1

検索語列	学生不知順位（全328語中）	BCCWJ(中納言)出現回数
多士済済	2	20
特筆大書	3	9
蟷螂の斧	4	10
大所高所	5	53
豪放磊落	7	21
衆議一決	8	2
古希	10	50
立錐の余地も無い	10	19
四角四面	16	26
四分五裂	16	10
一木一草	20	13
出処進退	20	10
切歯扼腕	23	3
一言半句	23	11
青雲の志	26	9
前門の虎、後門の狼	26	6
故事来歴	26	5
得手勝手	26	6
人事不省	26	8
青息吐息	31	22

表 2

ジャンル	成語数	%	ヒット数	%
書籍文学	93	47	1346	65%
書籍社会科学	35	18	359	17%
書籍歴史	24	12	140	7%
Yahoo!ブログ	11	6	28	1%
書籍言語	11	6	35	2%
書籍芸術	7	4	13	1%
書籍哲学	7	4	48	2%
Yahoo!知恵袋	6	3	14	1%
国会会議録	4	2	31	1%
書籍自然科学	3	2	7	0%
雑誌総合	3	2	9	0%
教科書	2	1	3	0%
書籍工学	2	1	30	1%
韻文	1	1	1	0%
新聞	1	1	1	0%
書籍総記	1	1	10	0%
計	211	100	2075	100
検索不能	1		1836(*500)	

表 3

中日同形成語のマッチング作業に先立ち、2010 年と 2011 年にかけ日本人大学生計 53 名を対象に名大「基本慣用句」の認知度調査を行なった[6]。漢字成語を含め慣用句が現代日本

[6] 第二外国語として中国語を履修中の政治経済学専攻の学部生対象、名大「(日本語) 基本慣用表」所収計 3628 語から選んだ漢字成語由来の 328 語を一覧表に作成、瞬時に大体の意味が理解できるか振り分け方

の若年層にどの程度受容されているのか知るためである。結果は、“多士済済”“蟷螂の斧”“豪放磊落”“大所高所”“同病相哀れむ”“衆議一決”“古希”“苦心惨澹”“立錐の余地も無い”“功成り名遂げる”“粗製乱造”“四分五裂”などの日本語の成語、慣用句を大学生の80%以上が理解できないことが判明した。“灯火親しむべき候”“老いては子に従え”“蛍雪”“晴耕雨読”“粗製乱造”“薄利多売”“一利一害”は半数が知らない。現代青年層の成語認知度は親世代の 50-60 歳代に比べ半減している[砂岡和子他（2012）]。

若年世代の漢字成語離れはこのまま加速し続けるのであろうか。成語認知度に差があるのは、記憶に留めやすい成語と忘れやすい成語の別があることを示唆している。両者の原因を明らかにできれば現代日本語中の漢字成語の延命を支援できよう。以下、日中成語の統語法や語用法の違いとコロケーションの特色について BCCWJ 検索を通し見てゆく。

孤立語系統に属する中国語は現在に至るまで一漢字の生産性が高く、意味結合による統辞の自由度が高い。中国語の四字成語は意合結合の縮図であり、日本語や韓国語など膠着語系言語に見られない柔軟な句組織が可能である。［表 4]は上掲[北大 1000] 所収の成語をその内部構造から 7 分類し、それぞれの出現比率を示したものである[7]。連合関係に加え、主述、動詞補語、動詞目的語、動詞目的補語関係など多用な結合をもつことが分かる。

調査した 328 語は“呉越同舟”のような中国語伝来の成語（62%）と、“一所懸命”“油断大敵”“絶体絶命”など日本製の慣用句（38%）に分類できる。大多数が知っていると回答した成語は“一喜一憂”“十人十色”“春夏秋冬”“自由自在”“自暴自棄”など後者の慣用句に多い。後者は慣用句の内部統語構造が比較的単純で、並列関係（“春夏秋冬”“自由自在”“自暴自棄”）、もしくは連合関係（“一喜一憂”“十人十色”）など、漢語文法を知らなくとも推理可能な四字句が多い。対して“呉越同舟”｛主語（呉越）が述語（舟を同じくする)｝や“異口同音”｛動詞＋目的語（口を異にしながら）動詞＋目的語（音[声]を異にする)｝は典故や漢語の構造を知らないと意味の理解が難しい。

中国成語内部構造	出現比率	語例
連合関係	44%	不学无术、打草惊蛇
主述関係	23%	草木皆兵、枯木逢春
修飾関係	15%	孜孜不倦、一衣带水
動詞補語関係	4%	重于泰山、无动于衷
非四字成语	8%	三十六计，走为上计
動詞目的語関係	5%	蹉跎岁月、饱经风霜
動詞目的補語関係	2%	问道于盲、入木三分

表 4

主要語法機能	出現比率
動詞性成語	72.7%
名詞性成語	11.6%
形容詞性成語	5.3%
副詞性成語	4.7%
分類詞性成語	2.6%
修飾性成語	1.7%
補語性成語	1.4%

表 5

四字成語の語用(品詞)機能にも違いがある。[表 5]は「現代漢語語法信息詞典」所収の

式で選択させた。回答所要時間はパソコン使用で 10-15 分、紙媒体で 20-30 分と、どちらも短時間で回収した。詳しくは砂岡和子,羅鳳珠(2011b)参照。

[7] 王雷（2010）原表に砂岡が一部修正を加え作成。

成語の語法機能を頻度準に並べたもので、中国語の成語が動詞として振舞うケースが圧倒的に多いことが分かる［姜柄圭(2012)］。

対して日本語の四字漢語は単語としての統辞機能に極めて制約があり、二字漢語や三字漢語と比べサ変動詞や形容動詞の語幹になることが少ないとされる。[野村雅昭(1974)］。例えば以下の四字成語は中国語では動詞として機能するが、BCCWJ の日本語用例はともに名詞性用法である。

・新陳代謝/「この辺、新陳代謝がうまくいってないんだろうなあ」「司令部、何やってるんだろ」1996 筒井康隆(著)最後の伝令,新潮社

・厚顔無恥/二人ともぼくの厚顔無恥には頭をふった。

2004 フリードリヒ・グラウザー(著) 種村季弘(訳)『外人部隊』国書刊行会

・粉骨砕身/それがし劉芳亮、商洛山中の危機を救うために粉骨砕身をいといません

1992 姚雪垠(著)陳舜臣, 陳謙臣(訳)『叛旗』徳間書店

ただし BCCWJ の用例を調べると必ずしもそうとは言い切れない。日本語成語のレキシコン分析のため、NINJAL-LWP for BCCWJ (NLB) ver.1.00 を利用し[8]、現代日本語で比較的出現頻度の高い漢字四字 24 語について、そのコロケーションをサンプル分析した結果、体言用法に劣らず用言用法が健在で、かつその統語パターンは体言用法のそれより少数精鋭で安定している[表 7]。［表 6］は (NLB) の「総合頻度」と「最多統語比率」の値が特に高い成語につき、ぞれぞれ体言法用 3 語、用言用法 6 語のコーパス情報を引用したものである。限られた資料ではあるが、安定した成語の類型には次の 2 タイプの特徴を認めることができよう。

1、統語パターンが一定の成語：“起死回生”“一網打尽”“四面楚歌”“虎視眈々”など

2、新しい統語パターンを派生する成語；“喜怒哀楽”“新陳代謝”“厚顔無恥”など

語形	総合頻度	パターンの種類	最多パターン	最多パターン頻度比率	最多パターン中の最多用法	同左頻度比率	最大MI用例	MI値	最多出現ジャンル	同左頻度比率
起死回生	32	7	起死回生+助詞	90.6	起死回生+の+名詞	53.1	の霊薬	18.03	書籍地の文	0.63
喜怒哀楽	68	14	喜怒哀楽+助詞	89.7	喜怒哀楽+の+名詞	26.5	の情	14.10	書籍地の文	1.33
新陳代謝	118	17	新陳代謝+助詞	95.8	新陳代謝を	51.7	をする	14.58	yahoo!知恵袋	4.87
一網打尽	26	3	一網打尽+助詞	88.5	一網打尽に…	92.3	にする	16.38	書籍会話文	0.9
一朝一夕	68	5	一朝一夕+助詞	98.5	一朝一夕に	85.3	～にできる	7.63	書籍会話文	1.25
四面楚歌	26	6	四面楚歌+助詞	80.8	四面楚歌に…	26.9	に陥る	13.40	国会会議録	0.6
虎視眈々	29	3	虎視眈々+助詞	93.5	虎視眈々と	90.3	と狙う	13.58	書籍地の文	0.5
正々堂々	55	5	正々堂々+助詞	83.6	正々堂々と	74.5	と勝ち上がる	15.68	書籍会話文	2.69
意気揚々	74	7	意気揚々+助詞	93.2	意気揚々と…	56.8	と拡張する	15.01	書籍地の文	1.6

表 6

		総合頻度	頻度計	パターンの種類		例	ジャンル
体言的用法	12例	68～3(平均頻度17)	209	14～2(平均パターン7種)	90.6	起死回生+の+名詞	書籍地の文
用言的用法	12例	74～5(平均頻度29)	347	11～6(平均パターン7種)	93.5	虎視眈々と	書籍地の文

表 7

[8] http://ninjal-lwp-bccwj.ninjal.ac.jp/

1、 は“起死回生（の霊薬)”“一網打尽（にする)”“虎視眈々（と狙う)”のように、内部構造が比較的複雑な漢字四字成語がより長い統語表現と結合した形式が多い。漢字四字は日本語表現と不適合を起こしやすいが、より長いコロケーションに包摂されることで安定した日本語表現に収まる。定型慣用表現単位の獲得は漢字四字成語の保持と延命に有効であり、四字漢字語であっても動詞や形容動詞の語幹になることが保証される。

2、 は“喜怒哀楽”“新陳代謝”“厚顔無恥”のように並列構造など比較的単純な四漢字句で、新しい文脈の中で表現形式の拡張が容易なため、延命が可能と言えよう。

表 2 は前掲「基本慣用句」328 語の認知度テストで、学生が知らない慣用句（成語）トップ 20 語[9]を[中納言]で検索した結果である。“大所高所”53 件、“古希”50 件、“四角四面”26 件、“青息吐息”22 件、“豪放磊落”21 件、“多士済済”20 件を始め、現代日本語中にかなりの用例数があり、出版年も比較的新しい。コーパスに多数出現するにも拘らず、これらの成語の認知度が低い理由に、現代社会の家族関係や習慣の変化の影響に加え、意味を想起しにくい漢字（古希＝古稀が正字、士、済、磊、落、所）を含むことが考えられる。そこで学生自身に自分が知らない成語や慣用句を［少納言]で検索体験してもらったところ、身近な現代日本語中に確かにこれらの成語が出現することを、実テキスト付きで確かめることができ、印象に深く刻まれた[10]。漢字四字成語の延命支援にも BCCWJ は役立つ。

６．まとめ

BCCWJ を利用した多言語成語辞典の構築の実際とその活用について紹介を行った。「少納言」「中納言」の利用によってデータの電子的検索と収集分析の利便性がより向上し、他言語とのマッチングデータとしても有用である。言語コーパスは共時的な言語情報だけでなく通時的な流通状況を知ることができる言葉の履歴書である。多地域のコーパスを連携すれば言語間の接触と変容を正確なデータで把握可能となり、対照言語学や比較文化の研究、翻訳をはじめ、語学教育の科学性と効率向上に寄与するに違いない。

謝　辞

本研究は平成 22-24 年度文部科学省科学研究費補助金〔基盤 (C) 課題番号 : 22520445 : 研究代表者砂岡和子〕および 2010-2013 年度台灣蔣經國國際學術交流基金會（課題番号 : RG013-D-09 :「歴代語言知識庫建置計畫」: 研究代表台灣元智大學羅鳳珠)による補助を得ている。

文　献

野村雅昭(1974)「四字漢語の構造」『国立国語研究所報告 54』『秀英出版』pp36～80

[9] 「学生が知らない」ほど数字が若い、上位 31 位から BCCWJ でヒットしない語を除き、同位があるため 20 語となった。

[10] 検索作業の感想を授業用チャットに書き込んでいる。

宮地裕編(1982)『慣用句の意味と用法『明治書院』
石井正彦(2001)「文章における臨時一語化の諸形式－新聞の四字漢語の場合－」『現代日本語研究』8, pp.1-34
村木新次郎(2002)「四字熟語の品詞性を問う」『日本語学と言語学』明治書院, pp.123-135
砂岡和子,羅鳳珠(2011a)「「歴代語言知識庫」發展的中日韓漢字成語教學模式」第四屆華語文教學國際研討會発表, 台湾銘傳大學 2011 年 3 月 11~12 日
砂岡和子,羅鳳珠(2011b)「日本語常用漢字熟語の選好変化と自然言語処理」日本語言処理学会 (NLP2011) 論文集 CDROM,E3-2
砂岡和子,羅鳳珠,王雷,姜柄圭(2012)「BCCWJ で知る東アジア漢字圏四字成語の受容と変容」 言語処理学会第 18 回年次大会 (NLP2012) 論文集 pp.899-902

劉長征&秦鵬(2007)「基于中国主流报纸动态流通语料库的成语使用情况调查」『語言文字應用』No3
王雷,俞士汶,朱學鋒,李芸(2010)「面向中文資訊處理的成語知識庫的建設與應用」『第五屆文學與資訊科技國際會議』台灣亞洲大學主辦, 2010 年 1 月 22~23 日
姜柄圭(2011) 「漢韓成語的語法結構對比分析及教學策略」第四屆華語文教學國際研討會発表,台湾銘傳大學 2011 年 3 月 11~12 日
王雷(2012)『中国成语 1000』北京大学出版社

関連 URL

国立国語研究所『現代日本語書き言葉均衡コーパス (BCCWJ)』
「中納言」https://chunagon.ninjal.ac.jp/
「少納言」http://shonagon.ninjal.ac.jp/
NINJAL-LWP for BCCWJ (NLB) ver.1.00 http://ninjal-lwp-bccwj.ninjal.ac.jp/
Open MWE for Japanese multiword expressions (MWEs)
http://openmwe.sourceforge.jp/pukiwiki-j/index.php?Idioms#s0e1b7f3
佐籐理史(2007)「基本慣用句五種対照表の作成」NLP178, pp.1～6、『情報処理学会研究報告』 (http://sslab.nuee.nagoya-u.ac.jp/index.html よりダウンロード可能)
台灣中央研究院平衡語料成語庫：http://db1x.sinica.edu.tw/kiwi/mkiwi/
韓国国立国語研究院世宗コーパス： http://kkma-sc.snu.ac.kr
台灣元智大学[網路展書讀]羅鳳珠： http://cls.hs.yzu.edu.tw
台灣中央研究院[全球華語文數位教與學資源中心]http://elearning.ling.sinica.edu.tw/
北京大学信息科学技术学院计算语言学研究所〔语义相关度检索平台〕吴云芳：http://klcl.pku.edu.cn:8008/seek/index.php
成語知識庫計画；砂岡和子 http: //www.f.waseda.jpksunaokacorpuschengyuzhishiku.pdf

日本語学習者にとって読みやすい文章について
― 日本語教科書における書き換えの分析から ―

クリスティーナ・フメリャク・寒川（リュブリャーナ大学文学部）†

On the Readability of Texts for Japanese Language Learners
-- From an Analysis of Text Adaptations in a Japanese Textbook

Kristina Hmeljak Sangawa (University of Ljubljana, Faculty of Arts)

１．はじめに

リーダビリティー（読みやすさ）の判定は、弱読者のための文章を執筆・選別するときに役に立つ判定で、日本語母語話者のための読みやすさに関する研究は近年大きな成果をあげている。しかし、母語話者にとって読みやすいものは、必ずしも日本語学習者にとっても読みやすいとは限らない。そこで本稿では、日本語学習者にとっての読みやすさを探るために、中級読解教科書に掲載された文章と、その原文からなる小規模コーパスを分析し、教科書著者が文章を読みやすくするために、どのようなストラテジーを使い、テキスト内のどの要素を言い換えたかを考察する。

２．読みやすさに関する研究

日本語母語話者のための読みやすさに関する研究は､英語のリーダビリティー研究(Flesch 1948 など) の影響も受け、文章の測定可能な要素（文の長さ、高頻度語彙の割合など）から、その文章の難しさを判定する難易度算定公式を開発してきた(森岡 1952、阪本 1964、建石ら 1988、佐野・丸山 2008、柴崎ら 2008、柴崎・玉岡 2011 など)。柴崎らのリーダビリティー測定公式を用いたシステムはインターネット上公開されており[1]、実用化されている。近年、コーパスに基づいた統計的な言語モデルを用いたテキスト難易度推定システム（近藤ら 2008､佐藤 2011）も構築、公開されており[2]、日本語母語話者（特に若年者）にとって、特定の文章がどのぐらい読みやすいか、日本の学校のどの学年の文章に近いかを測ることが簡単にできるようになった。しかし、Hmeljak-Sangawa(2009)において指摘されているように、これらの測定・推定システムは、特にやさしい文章を測定する際、その表記、特に文章に含まれている教育漢字とひらがなの割合に大きく左右されるもので、標準表記で書かれている日本語学習者用のテキスト、特に初級・中級用の文章レベルを測定・推定するには限界がある。

日本語学習者にとっての読みやすさに関しても、これまでに多くの研究が行われてきたが、研究の焦点は文章の一側面に限定され、その一側面における文章難易度を測定・推定したものが多い。例えば、川村(1999)、川村ら(2008)、北村ら(2009)などの一連の研究は語彙に着目し、日本語能力試験出題規準、語彙親密度などの語彙難易尺度を用い、その尺度にそって文章に含まれる語彙の割合を基に文章の読みやすさを判定する。難構文に着目した研究もあり、水嶋ら(2011)は中止法，名詞修飾節，主格省略の検出を、竹井ら(2005)はゼロ代名詞（語句の省略）の検出と明示化を、中村ら(2012)は主格省略文の検出を提案している。また、日本語能力試験読解問題における日本語能力試験出題規準の語彙含有率を測定した研究(柴崎・李 2010)、および学習辞書のための分かりやすい用例を選別するために、語彙、文型、構文（連体修飾説、複雑な係り受け構造など）の難易度測定システムを提案した研究（小林ら 2007、水野ら 2008、吉橋ら 2007、Hazelbeck ら 2009) もあげられるが、日本語学習者にとっての読みやすさ、読みにくさの要因はまだ総合的に十分明らかにされて

† kristina.hmeljak@ff.uni-lj.si

1 http://readability.nagaokaut.ac.jp/readability

2 http://kotoba.nuee.nagoya-u.ac.jp/sc/obi2/

いない。

そこで本研究では、日本語学習者にとって難読と思われる要素を探るために、日本語ベテランである日本語教科書の著者が自らの日本語教育の経験に基づき書き換えた文章を分析・調査した。

３．データ

今回の調査で使用したデータは、産能短期大学日本語教育研究室編の『日本語を楽しく読む本・中級』におさめられている２種類の文章、すなわち、想定されている中級日本語学習者のために書き換えられた読解用のテキスト（以下これを書き換え文と記述）と、その原文である。この教科書を調査対象にしたのは、原文と書き換えた文章の両方を掲載した数少ない資料だからである。また、中級の教科書に着眼したのは、初級教科書の読み物には学習者用の書き下ろし（語彙・文型・構文を極端に制限した文章で、場合によって母語話者に違和感さえ与える文章)、上級教科書の読み物には日本語母語話者用に書かれた生教材が多く使われるのに対し、中級教科書では、日本語母語話者のために書かれた文章を日本語学習者に読みやすいように書き換えられた文章がよく利用されるからである。このような日本語学習者用の書き換え文と母語話者用の原文の比較からは、日本語学習者に特有な難読要素が得られると考え、その比較を行った。

調査した教科書は９つの章に分かれており、以下の読み物が掲載されている。

表１：分析資料の概略

章	テキスト題名	文字数		出典
		書換文	原文	
1	禁酒	449	352	『月刊アサヒ』1989 年 6 月号、朝日新聞社
2	いつもとちがう	682	527	『ポケットジョーク９トラベル』植松黎編訳、角川書店
3	賢い農夫／百姓	1327	1285	矢崎源九郎『世界の民話』社会思想社
5	振り向き賃	526	909	深田祐介『ちょっといい旅いい話』ジャルパック出版
6	ママ、手が凍るんだよ 行康君、手すり変わるよ	707 829	702 922	朝日新聞 1985 年 2 月 20 日朝刊 朝日新聞 1985 年 3 月 1 日朝刊
7	みんなって何人？	1277	2640	斉藤勇『人間関係の分解図』誠信書房
8	握手１ 握手２	1147 851	2408 1199	阿部謹也『逆光のなかの中世』日本エディタースクール 樋口清之『日本の風俗の謎』大和書房
9	趣味	1219	1972	守屋毅『日本文明 77 の鍵』（梅悼忠夫編）創元社
	合計：	9014	12916	

教科書の前書きによると、450～600 時間ぐらい勉強し 3500～5000 語程度の語彙を学習している中級レベルの日本語学習者を想定し、自作の語彙リストに基づいて 3500 語程度のレベルを標準として、原作を書き換えたということである。語彙以外の書き換えの基準について言及していないということからは、この教科書の著者は、多くのリーダビリティー研究と同様に、語彙が読みやすさの最も重要な要素だと考えていることが窺える。

４．分析方法

原文と書き換え文を電子化し、JDiff X (Matsumoto 2010)というソフトウェアを用いてテキストの対を比較した上で、検出された相違を表計算ソフトに入力し、それぞれの書き換えの種類、レベル（語彙、構文、談話など）と書き換え内容の記述を付与した。一つの文の中で複数の書き換え（相違）が確認された場合、それぞれの書き換えを別項目として

記述した。例えば、第 1 章の読み物「禁酒」には次の原文(1)と書き換え文(2)があった。

(1-原文)「亡くなった飲み友だちと約束してね、僕が飲みに行くときは、必ずオレの分も注文して飲んでくれという遺言を実践しているんだ」

(2-書き換え文)「先週、僕の親友が亡くなったんだが、彼が亡くなる前に約束してね」

「はあ」

「僕が飲みに行く時は、必ず、彼の分も注文して飲むということになったんだ。それで、その約束を実行しているってわけさ」

上記の原文と書き換え文を比較した際、次の相違（書き換え項目）が確認された。

・統語構造の単純化：「亡くなった飲み友だち」という連体修飾節から「親友が亡くなった」という単文への書き換え、「オレの分も注文して飲んでくれという遺言」という修飾節から「…彼の分も注文して飲むということになったんだ。…その約束を…」という文への分解

・談話構造の単純化：「僕が飲みに行くときは、必ずオレの分も注文して飲んでくれという遺言を[僕が]実践している」という短い複文の中で、主語が「僕」から「オレ」、そしてまた「僕」へと変わることを避け、「僕」への統一

・談話構造の明示化：会話場面を意識させる相づち「はあ」の挿入

・意味的な明示化：「先週」の挿入による時間設定の明示化、「飲み友だち／親友」の属性を明示化する「僕の」の追加

・語彙的な単純化：口語的な代名詞「オレ」の代わりに「彼」、低頻度語彙の「飲み友だち」の代わりにより広い意味の「親友」、「遺言を実践」の代わりに「約束を実行」への書き換え

・文型の明示化：「しているんだ」の中で、広い用法範囲をもつ「んだ」の代わりに、説明に使う「ってわけ」を使い「しているってわけさ」への書き換え

・文字レベルにおける明示化：単語の境界が明確でないひらがなの連鎖「行くときは」から、初級レベルで習得される漢字「時」への書き換え

・句読点による統語構造（境界点）の明示化：「必ず」という副詞の後に読点の挿入

このように原文と書き換え文の相違を記述した結果、815 項目の相違リストが得られた

５．書き換え項目の分析

以上のような分析によって得られた相違項目は、三つのストラテジー、すなわち単純化、明示化、標準化に分類できる。これらのストラテジーは、複数レベルで確認された。

５.１　単純化

このストラテジーはもっとも多く確認され、96 の削除を含み、472 の単純化の例が確認された。

・文字表記においては、低頻度漢字から平仮名への書き換え（尋ねた→たずねた、土産→おみやげ、嫌がります→いやがります、石鹸→せっけん、挨拶→あいさつ、など）

・語彙においては、低頻度内容語の書き換え（誰しも→多くの人が、奇怪／珍奇な→珍しい、 育種学→植物学、事柄→こと、食糧→食べ物、受容する→受け入れる、大衆化し→広がり、一覧表→リスト、かたっぱしから→何でも、絆→関係、帰宅した→家に帰った、など）または定義や説明の追加（古典→昔の文学作品、利己的遺伝子の乗り物→利己的遺伝子［ふりがな：セルフィッシュジーン］の乗り物［ふりがな：ヴィークル］）；また、機能語の書き換え（のみ→だけ、かならずしも…というわけではない→ばかりではない、…にせよ…にせよ→例えば、すら→も、…V られっぱなしで→その間…V られていて、など）

・文法においては、口語的な文型、または硬い口調から、初級で学習される文型への書き換え（会社を出→会社を出て、言わずに→言わないで、同じく→同じように、みせねばならない→みせなければならない、など）

・統語においては、長い文、複文の分解が多く見られた（彼女はあわてて手の甲をぬぐって私に右手の甲をさし出し、お互いの右手の甲を合わせて握手の代わりとしたことがあ

った。→彼女は急いで手の甲をふいて右手の甲を差し出した。お互いの右手の甲を合わせて握手の代わりにしたことがあった。）（この実験は、人が集団の圧力を感じ、集団に従おう、あるいは従わざるをえないというような気持ちになるのは、三人以上であることを明らかにしています。→この実験から、次のようなことがわかります。私たちが、集団の圧力を感じて他の人と同じ意見を言うようになるのは、３人以上の人が同じことを言った時であること）

・談話構造において主題移行を避けた主題の統一（王さまは…、大臣たちが集まったとき、王さまは…→…お城に帰ると、大臣たちを集めて、王さまはこう言いました）

単純化の極端な例である削除も、さまざまなレベルにおいて確認された。

・モダリティー表現の削除（どうも土産が足りない→おみやげが足りない）

・アスペクト形式の削除（捨ててしまいます→捨てます）

・意味上、付随的な情報の削除（林野庁業務部販売推進室長羽賀正雄さん→林野庁業務部の羽賀正雄さん、越後からやってくるたび東京の街なかで→東京に出てくると、法則の基礎知識をもっていた→法則を知っていた、など）

５.２　明示化

明示化の例も多く、203 の例がさまざまなレベルで確認された。

・意味的な明示化、例えば設定の明示（ある男が、バーのカウンターにすわった。→ある男が、バーのカウンターにすわった。はじめての男だった）、より具体的な表現の使用（とうとう→次の日、連絡をとった→電話をした、こいつ→この客、やっている→経営している）

・漢字表記による曖昧さ解消と単語境界の明示化（いっている→言っている、もっている→持っている、よんで→呼んで、など）

・境界の明示化も多く（70 項目）、もっとも多いのが句読点や括弧の挿入による単語境界の明示化（二十年間もうちのスープ→もう 20 年も、うちのスープ；男はそれからもときどき→男は、それからも、ときどき；実際手は→実際、手は；三人というのはみんななのです→３人というのは「みんな」なのです）、または句読点挿入による文節境界の明示化（いったいどうしたというんですか？→いったい、どうなさったんですか）、次に多いのが句読点の挿入や段落分割による統語構造の明示化

・統語的な明示化は、省略されていた主語等の挿入（手袋は、滑ってしまうから、つけてはいかれない→手袋をすると滑ってしまうから、手袋をすることはできない）、自動詞から動作主語が明確な他動詞（一杯目が終わったら→一杯目をお飲みになったら）、「の」を使った複合名詞の分割（現代日本→現代の日本、遺伝法則→遺伝の法則、など）

・結束性の明示化、指示詞の追加（その結果が図です→その結果が、上の図です）

５.３　標準化

３つめのストラテジーは、単純化に類似するストラテジーだが、低頻度要素（語彙・文型）を高頻度要素に書き換えた単純化項目のに対し、標準化と分類した項目は、初級で学習しない口語的、または硬い、あるいは特殊な文体要素を、中性的な要素に書き換えた例である。

・表記において、標準的な句読点や括弧への書き換え（ですか？→ですか。;《皆様の...いただきます》→『みなさんの...いただきます』）

・時制の統一（過去形で書かれた文章の中で：されるのは→されたのは）

・文体の統一（すべての文を「です・ます体」、または全ての文を普通体へ）

・口語表現や方言などの特殊な文体から標準表現への書き換え（飲まなくったって→飲まなくても、話さん→話さない、ようがす→わかりました、など）

・漢数字からアラビア数字への統一（百枚→100 枚、十二時→12 時、十七世紀→17 世紀　六、など）；これは、原文の縦書きから教科書の横書きへの変換によるものと思われる。

５.４　その他

数例においては、書き換えの目的が確認できなかった。例えば、タイプミスと思われる

例（追っかけてきます→追っかけてきてます)、非標準的な句読点（そうですか。わかりました。→そうですか、わかりました。）のような書き換えが数例あった。

５.５　結果のまとめ

このように、書き換えを分析した結果、さまざまなレベルにおいて、複数のストラテジーが確認された。それぞれの書き換えの例をレベルでまとめると、表２のようになり、 ストラテジー別にまとめると、表３のようになる。

表１　レベル別書き換え例の数と割合

レベル	例数	割合
表記	175	21%
句読点	56	7%
内容語	193	24%
機能語	44	5%
語彙＋統語	167	20%
活用形	20	2%
モダリティー	18	2%
統語	42	5%
意味	45	6%
結束性	10	1%
談話	33	4%
文体	12	1%
合計	815	100%

表３　ストラテジー別書き換え例の数と割合

ストラテジー	例数	割合
単純化（その内、削除の例が 96)	472	58%
明示化	203	25%
標準化（その内、数字表記統一例が 80)	128	16%
その他	12	1%
合計	815	100%

６．考察

中級日本語教科書において学習者用に書き換えられた文章をその原文と比較した結果、単純化の例、特に教科書の前書きにも明記されており、先行研究でも頻繁に取り上げられている低頻度語彙から高頻度語彙への置き換えという単純化の例は多かったが、他のストラテジー、他のレベルにおける書き換えも多く確認され、それらの書き換えは日本語学習者にとって難読だと思われる要素を暗示している。

７．まとめ

本研究では、日本語母語話者のリーダビリティー研究に対して、日本語学習者にとっての読みやすさの要因を探るために、中級日本語教科書のために書き換えられた文章と、その原文となる日本語母語話者用の文章を比較し、その相違を分析した。その結果、母語話者にとっても読みにくいとしばしば指摘される低頻度語彙、長くて複雑な文が書き換えられたことが確認され、この要素に関しては母語話者と日本語学習者の共通難点が認められたが、同時に、日本語母語話者の幼児にとって簡易であろうと思われる口語的な要素、漢

語からカタカナ語への書き換えなど、日本語学習者に特有な制限に起因すると思われる書き換えの例も確認された。これは、日本語学習者用のリーダビリティー研究の必要性を暗示している。今回の分析は 1 冊の教科書における書き換えに限定したので、より一般的に有効な日本語学習者のための読みやすさの要因と、リーダビリティー判定への応用の可能性については、更なる研究が必要である。

文　献

Flesch, R. (1948) A new readability yardstick, *Journal of Applied Psychology* 32:3, pp. 221-233.

Hmeljak-Sangawa, Kristina (2009) A Corpus for Readability Measurement for Non-Native Learners of Japanese, *Technical report of IEICE. Thought and language*, 109:84, pp.19-24.

Hazelbeck, Gregory and Saito, Hiroaki. 2009: A Corpus-based E-learning System for Japanese Vocabulary, *Information and Media Technologies* 4:4. pp.1104-1128.

川村よし子(1999)「語彙チェッカーを用いた読解テキストの分析」『講座日本語教育 34』pp.1-22

川村よし子、北村達也、冨岡洋介、林真一(2008)「単語親密度と頻度情報を活用した難易度判定システム」『日本語教育方法研究会誌』15:1.

北村達也、冨岡洋介、川村よし子(2009)「IDF を用いた単語レベル判定システムの構築と検証」『日本語教育方法研究会誌』16:1、pp. 52-53.

近藤陽介、松吉俊、佐藤理史(2008)「教科書コーパスを用いた日本語テキストの難易度推定」『言語処理学会第 14 回年次大会発表論文集』pp. 63-66.

Matsumoto, Satoshi (2010) JDiff X Document Comparison Plug-in for Jedit X. (http://www.artman21.com/en/jdiff_x/よりダウンロード可能)

水嶋博志、田聖也、北村達也、川村よし子(2011)「学習者にとって難解な構文の自動検出」『日本語教育方法研究会誌』18:1、pp. 64-65.

森岡健二(1952)「『読みやすさ』の基礎的研究」昭和２６年度国立国語研究所年報、pp. 91-108.

中村慶太、北村達也、川村よし子(2012)「検索エンジンを用いた主格省略文の自動判定」『日本語教育方法研究会誌』19:1.

阪本一郎(1964)「文の長さの比重の測定法−Readability 研究の試み」『読書科学』 8:1、pp. 2-6.

産能短期大学日本語教育研究室編(1991)『日本語を学ぶ人たちのための 日本語を楽しく読む本・中級』産能短期大学国際交流センター／凡人社

佐野大樹、丸山武彦(2008)「システミック文法に基づく書きことばの複雑さ測定」『言語処理学会第 14 回年次大会発表論文集』pp. 1097-1100.

佐藤理史(2011)「均衡コーパスを規範とするテキスト難易度測定」『情報処理学会論文誌』52:4、pp. 1777-1789.

柴崎秀子、玉岡賀津雄、山本和英、加納満、原信一郎、李在鎬(2008)「日本語コーパスを応用した文章の難易度測定の研究」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予稿集、pp. 125-130.

柴崎秀子、玉岡賀津雄(2010)「国語教科書を基にした小・中学校の文章難易学年判定式の構築」『日本教育工学会論文誌』 33:4、pp. 449-458.

柴崎秀子、李在鎬(2010)「日本語能力試験読解問題を土台にした文章の難易尺度の構築―日本語教育リーダビリティーの基礎研究―」『日本語教育学会春季大会予稿集』pp. 1-5.

建石由佳、小野芳彦、山田尚勇(1988)「日本文の読みやすさの評価式」情報処理学会研究報告 1988-HI-018/. pp. 1-8.

Yamura-Takei, Mitsuko, Aizawa, Teruaki, and Fujiwara, Miho (2005) Diversity of zeros in Japanese discourse: a corpus analysis and a tool for language teachers. In: *Proceedings of Pacific Associations for Computational Linguistics (PACLING 2005)*, pp. 358-367.

吉橋健治、傅亮、仁科喜久子(2007)「学習者に合わせた例文表示ツール」Proceedings of CASTEL-J in Hawaii 2007、pp. 223-226.

段落間の類似度を利用したテクストの結束性の測定

山崎　誠（国立国語研究所言語資源研究系）†

Measurement of Textual Cohesion Using Similarity between Paragraphs

Makoto Yamazaki (Dept. Corpus Studies,NINJAL)

１．はじめに

テクストの結束性（cohesion）は、一貫性（coherence）とともにテクストの基本的な性質とされる重要な概念である。とくに結束性は Halliday&Hasan（1976）以来、多くの研究が行われている。本稿は、テクストを構成する段落間の類似度を使ってテクストの結束性を計量的に明らかにしようとするものである。

２．テクストにおける結束性とその現れ方

結束性とは、文章をひとつの統一体としてまとめあげるために必要な性質のひとつとされる。結束性について最初に詳細に研究を行ったのは Halliday&Hasan(1976)である。それによると、結束性について次のように紹介されている。

「結束性が生じるのは，談話のある要素の解釈（INTERPRITATION）が別の要素の解釈に依存する場合である。一方を効果的に解釈するためには他方に頼らなければならないという意味で，一方は他方を前提（PRESUPPOSE）とする。こういうことが生じるとき，結束関係が成立する。その結果，前提語と被前提語という　2 つの要素が、少なくとも潜在的には，統合されて 1 つのテクストになるのである。」（邦訳 p.5）

また、結束性には文法的結束性と語彙的結束性とがあり、前者の手段として「指示」「代用」「省略」「接続」が、後者には「再叙（reiteration）」と「コロケーション」がある。再叙には以下の 4 つのタイプがある。

(a)同一語（繰り返し）

(b)同義語（または近似同義語）

(c)上位語

(d)一般語

平林（2003:72）によれば、日本人高校生の英作文（1 作文あたり平均約 100 語、10 文）の分析から 1 作文あたり平均 12.87 個の結束数が現れ、その内訳は、指示 4.9 個、接続 3.62 個、語彙的結束性 3.36 個、代用 1.00 個、省略 0 個だという[1]。

Károly（2002:162）では、英語の作文においては、(a)の同一語の繰り返しよりは(b)〜(d)を合わせた「異なる語の繰り返し」の方が多く用いられるという報告がある。しかし、本稿では、同義語（類義語）や上位語の判断を自動的に行うことが難しいため、(a)の同一語の繰り返しのみを観察対象とする。

† yamazaki@ninjal.ac.jp

1 作文における結束性の割合はひとつの目安となるが、今回例として取り上げた白書のデータでは指示(1 個)、接続(1 個)よりも語彙的結束性(181 回:繰り返し使用された語における繰り返しの数）のほうが多く、やや異なる出現状況であった。

３．データと方法

３．１ データ

本発表では、2011 年 12 月にリリースされた『現代日本語書き言葉均衡コーパス』の DVD 版を使用した。Disk1 の CORE/M-XML フォルダに含まれる白書の xml ファイル 62 個が対象である。これらの xml ファイルは可変長サンプルと固定長サンプルを統合したもので、短単位、長単位の形態論情報のタグのほか可変長部分には文章構造のタグを含んでいる[2]。本稿では結束性を適切に観察するため、文章構造のタグを含まない固定長部分を対象外とし、可変長部分のみを用いる。

対象となるデータには文章構造のタグとして<paragraph>が使われており、このタグで囲まれた部分を一つの段落としてテクストの構成を考えることにする。なお、<title>のタグで囲まれた見出し部分や注記、図表のキャプションなどは対象外とした。

３．２ 結束性の測定方法

結束性の測定方法は 2 節で挙げた語彙的結束性のタイプのうち(a)の同一語の繰り返しを利用する。各段落をひとつの語彙とみなし、語彙の類似度を利用して、各段落間の関連を観察する。

語彙の類似度を表す主な指標には C（宮島（1970））と D（水谷（1980））とがあるが、本稿では D（以降、単に「類似度」と言う）を用いる。この類似度は、非対称的であることに特徴があり、語彙 a の語彙 b に対する類似度と語彙 b の語彙 a に対する類似度がそれぞれ別の値をとることができる。見方を変えれば語彙の「依存度」（水谷（1980:147））と考えることもできるものであり、順を追って構成されるテクストの内容の関係をさぐる上で適切な指標となるものである。語彙 a の語彙 b に対する類似度は次の式で表される[3]。

$$Da|b \stackrel{def}{=} \sum_{M \in V_a \cap V_b} P_a(M) = \frac{1}{N_a} \sum_{M \in V_b} F_a(M)$$

V_x：表現域 x の上の語彙
$P_x(M)$：見出し語 M の表現域 x における使用率
N_x：表現域 x の延べ語数
$F_X(M)$：表現域 x での見出し語 M の使用度数

類似度の測定にあたっては、短単位を用い、品詞が空白、補助記号、助詞・助動詞であるものを除外した。これらは結束性を表しているわけではないからである。同様に、結束性への貢献が低い語群についても除外した。この語群の候補として田中（1973）の「無性格語」を利用した。無性格語とは、田中によれば「これらの単語は，どんな文章にも現われるようなものであって，ある特定な文章や文献の性格とか特徴とかを反映することは，ほとんどない。いわば無性格な語群であろう。」（田中（1973:157））とされるものである。田中（1973）では 108 語が無性格語としたリストアップされている。本稿ではこの無性格語を短単位での実現形に合わせて適宜修正して用いた[4]。また、田中（1973）の趣旨を汲み、リストに上がっていない数詞についても無性格語として処理した。

---

[2] タグの詳細については小木曽ほか（2011）を参照。
[3] 式は水谷（1983）より引用。
[4] 本稿で用いた無性格語のリストを付表に掲げた。

４．分析例

４．１ データ

表 1 は『平成 16 年度文部科学白書』（サンプル ID:OW6X_00000）の可変長部分を段落に分けて示したものである[5]。見出し部分は本文とは別立てにして、その見出しが及ぶ範囲が明らかになるように示した。このサンプルは、接続詞が第 5 段落の「さらに」の 1 つのみ、指示詞が第 3 段落の「これ（まで)」の 1 つのみであり、文法的結束性が少ないという特徴を持っている。

表 1　白書データ（OW6X_00000）の構造

見出し1	見出し2	見出し3	段落番号	テクスト
1　日本文化の発信による国際文化交流の推進	(1)文化庁文化交流使事業	①　文化庁文化交流使事業	P1	文化庁文化交流使事業は，芸術家，文化人等，文化に携わる人々に，一定期間「文化交流使」として世界の人々の日本文化への理解の深化や，日本と外国の文化人のネットワークの形成・強化につながる活動を展開してもらうことを目的として，平成15年度から始めた事業です。
			P2	「文化交流使」の活動には，(ⅰ)日本在住の芸術家，文化人が海外に一定期間滞在し，日本の文化に関する講演，講習や実演などを行う「海外派遣型」，(ⅱ)海外在住の日本文化に深い知見を持つ芸術家，文化人が，講演，講習，現地メディアへの投稿，出演等を行う「現地滞在者型」，(ⅲ)講演等で来日する諸外国の著名な芸術家が，日本滞在期間を利用して学校などを訪問して実演・講演等を行う「来日芸術家型」の三つの類型があります。
			P3	平成16年度は，「海外派遣型」文化交流使として11名，「現地滞在型」文化交流使として4名，「来日芸術家型」文化交流使として4組の指名を行いました。重要無形文化財保持者，写真家や音楽家など様々な分野で活躍中の方々の活動を通じて，日本文化のこれまで紹介されていなかった一面や，日本文化になじみの薄かった国や地域での日本文化の紹介などの活動を行っています。
		②　文化庁文化交流使活動報告会	P4	平成15年度に文化庁文化交流使として海外で活動した人々による報告会を，東京国立博物館平成館大講堂にて開催しました。
			P5	笑福亭鶴笑氏(落語家)，田中千世子氏(映画評論家)，バロン吉本氏(漫画家)，三浦尚之氏(福島学院大学教授)，渡辺洋一氏(和太鼓奏者)の5名が活動報告を行うとともに，国際文化交流について討論し，さらに笑福亭鶴笑氏によるパペット落語(笑福亭鶴笑氏が自ら考案した落語形式で，足や膝につけた人形を操りながら演じる。)の実演が行われました。
	(2)国際文化フォーラムの開催		P6	「国際文化フォーラム」は，国際的に著名な国内外の芸術家・文化人などを招聘し，座談会，講演などの形式により，世界の文化芸術の最新の諸相や動向について語り合ってもらうことを目的として，平成15年度から開始した事業です。
			P7	平成16年度も15年度に引き続き，11月に関西地区で，「文化の多様性」の共通テーマの下に，「国際情勢における『文化の多様性』の意義」，「シルクロードと仏教文化」などについて話し合い，世界に向け，文化のメッセージを強く発信しました。
	(3)国際芸術見本市		P8	舞台芸術のブース設置や実演を行うことにより，国内外の劇場関係者に，我が国の新進の舞台芸術作品などを紹介する国際芸術見本市を，平成16年8月に東京芸術劇場にて開催しました。文化庁から国際舞台芸術交流センターへ委任を行い，実演デモンストレーション(ショーケース(＊))を制作しました。

[5] 該当箇所は文部科学省のホームページでも確認することができる。URL は次のとおり。
http://www.mext.go.jp/b_menu/hakusho/html/hpab200401/hpab200401_2_277.html

４．２ 各段落間の類似度

４．２．１ 全体の傾向

OW6X_00000 を構成する 8 個の段落相互の類似度を表 2 に挙げた（同一段落どうしの類似度は必ず 1 になるので除く）。値は 0.0364～0.5614 の間に分布し、平均値は 0.280 である。類似度の分布の様子を図 1 に示した[6]。

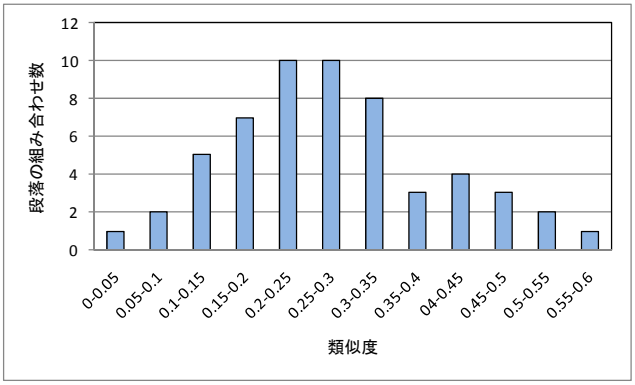


図 1 類似度の分布

表 2 はすべての段落間の類似度であるが、ここからある段落 a から他の段落 b への類似度において、対象とする相手方の段落 b との類似度が最も高い段落（表の太字のセル）がほとんど第 1 段落（P1）～第 3 段落（P3）に集中しており、このサンプルは前方の段落に依存する傾向があることが見て取れる。

表 2 段落間の類似度

	P1	P2	P3	P4	P5	P6	P7	P8	平均
P1		**0.5128**	0.4103	0.4359	0.2821	0.4615	0.2564	0.3077	0.0440
P2	0.3906		**0.5156**	0.1719	0.25	0.3125	0.0781	0.2344	0.0335
P3	0.4118	**0.5686**		0.3529	0.3333	0.2549	0.1765	0.3333	0.0476
P4	**0.5**	0.3	0.45		0.25	0.25	0.25	0.4	0.0571
P5	0.12	**0.18**	0.16	0.08		0.12	0.04	0.12	0.0171
P6	**0.4815**	0.3333	0.2963	0.1852	0.2593		0.2963	0.3333	0.0476
P7	0.2857	0.1429	0.25	0.25	0.1786	**0.3214**		0.2143	0.0306
P8	0.2368	0.2895	**0.3158**	0.1579	0.2105	0.2632	0.1053		0.1429
平均	0.0338	0.0414	0.0451	0.0226	0.0301	0.0376	0.0150	0.1429	

注

(1)縦の系列の段落が横の系列の段落に対してとる類似度の表。例えば、P3 の P2 に対する類似度は 0.5686（この値は P2 への P3 からの類似度と解することもできる）。

(2)太字は、当該段落から他の段落への類似度のうちもっとも値が高いもの。P4 の段落を例にとると、P4 の横の列（0.5,03,0.45,0.25,0.25,0.25,0.4）の中でいちばん高い値の 0.5 になる。

(3)下線は、他の段落から当該段落への類似度のうちもっとも値が高いもの。P5 の段落を例にとると、P5 の縦の列（0.2821,0.25,0.3333,0.25,0.2593,0.1786,0.2105）の中でいちばん高い値の 0.3333 になる。

[6] 例えば、階級 0.1-0.15 は 0.1 より大きく 0.15 以下であることを示す。そのほかも同様。

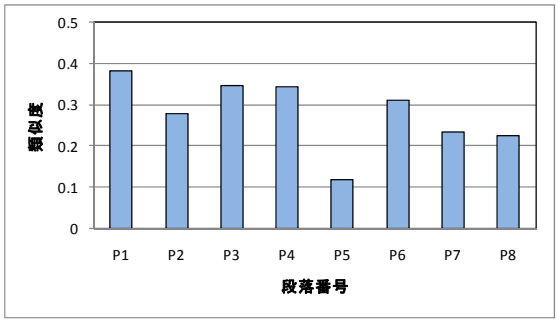


図 2 他段落への類似度

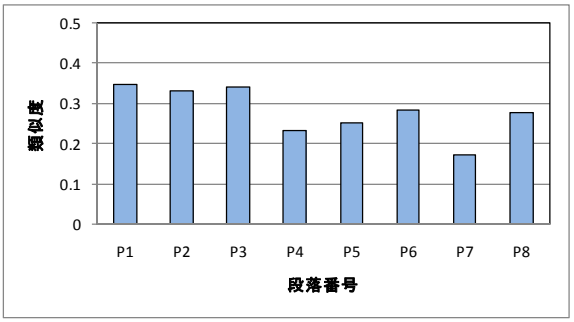


図 3 他段落からの類似度

図 2 は、ある段落の他の段落に対する類似度の平均、図 3 はある段落の他の段落からの類似度の平均である。他段落からの類似度の平均は全体的に同じような値を示しているが、他段落への類似度の平均では、第 5 段落の値がほかとくらべて低くなっていることが分かる。このことは第 5 段落と他の段落とで共通して用いられる語について、第 5 段落での使用度数は少ないが、他の段落での使用度数が多いことを示唆する。例えば、第 5 段落と、(第 5 段落への類似度がもっと高い) 第 3 段落の場合は「行う、家 (か)、活動、交流、文化、名 (めい)」の 6 語が共通して現れる語であるが、「文化」は第 5 段落に 1 回しか使用されていないのに対して第 3 段落では 7 回使用されている。同様に「交流」は 1 回に対して 3 回、「行う、名 (めい)」はそれぞれ 1 回に対して 2 回であった。この共通出現語の使用の不均衡が類似度の非対称性に影響していると考えられる。このことを踏まえて第 5 段落と第 3 段落を比較してみると、テクストの表層的な構造では第 5 段落は第 4 段落に従属するものであるが、語の分布状況から見ると第 3 段落とも関係が深いことになる。

４．２．２ 類似度からみた全体の構成

図 4 は、段落間の類似度の平均値 0.280 の 1.5 倍 (0.420) 以上の値を持つ段落の組み合わせを図示したものである。図の見方は、例えば P1→P2 であれば P1 の P2 に対する類似度が高かったことを表している。両矢印は相互に類似度の値が高かったことを示す。図 4 から、第 1 段落から第 4 段落までは結束性が強いことが伺える。一番多くの矢印が出入りしている第 1 段落がこのテクストの中心的な位置を占めていると言えよう。

第 1 段落と相互に類似度が高い 2 つの段落 (第 4 段落と第 6 段落) のうち第 6 段落は、「～は，・・・てもらうことを目的として，平成１５年度から開始した事業です。」という文構造であり、第 1 段落と骨格は同じである。このような「形式の類似性」も結束性に貢献していると考えられる。

この中では、第 5、第 7、第 8 段落が比較的独立性が高く他と分離されているが、後述のように第 5 段落は第 4 段落を具体的に展開したものであり、第 7 段落も事情は同じである。この関係は今回の分析からは読み取れない。

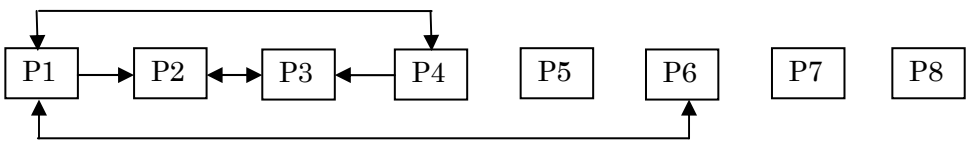


図 4 類似度からみたテクストの構成

４．２．３ 直前・直後の段落との類似度

類似度の値を利用して連続する段落間の切れ続きについて考察する。山崎（2012）では直前の段落への類似度よりも直後の段落への類似度の方が大きいところが内容的な切れ目であり表層的なテクストの構成とも一致する場合が多いと指摘しているが、このサンプルではどうであろうか。結果を図 5 に示す。

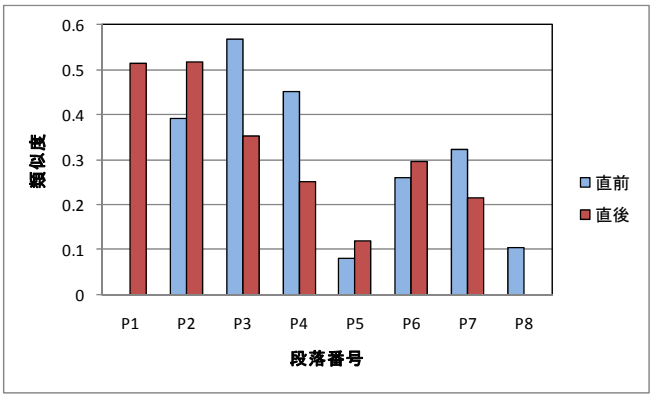


図 5 直前・直後の段落との類似度

図 5 から第 2 段落、第 5 段落、第 6 段落が直後への類似度が高いことが分かる。表層的には第 2 段落は第 1 段落の続きであり、第 1 段落の内容を具体化しているものであるが、具体的な内容が多くなったために第 1 段落への類似度が相対的に低くなったものと思われる。同様の関係は第 5 段落にも見られる。第 5 段落も直前の第 4 段落の内容を具体化したものであるが、第 4 段落が第 5 段落の短いまとめ的な内容であることから直前の段落への類似度が相対的に低くなったものである。本稿の類似度の測定は同一語かどうかによっているので、このようなものごとを具体化して述べるようなつながりについては感度が弱い。

なお、第 8 段落は後続の段落がなく、上記の方法では観察できないが、直前の段落との類似度がかなり低いことからここも内容的な切れ目に相当する可能性が高いと思われる。

５．まとめと今後の課題

本稿では段落間の非対称的類似度を利用して、テクストの結束性のようすを概観した。今回扱ったデータは白書のサンプル 1 つのみであったが、すべての段落間の組み合わせを観察することにより、どの段落とどの段落とが関係が深いのか結束性の一端を伺うことができた。また、隣接した段落以外にも結束性の高い段落があり、それらの関係を利用したテクストの構成の分析への発展の可能性を示唆した。

本稿で利用した「無性格語」のリストは雑誌九十種調査の結果から作られたもので、異なるレジスターの分析に耐えるかどうかは検証が必要であろう[7]。例えばリストには固有名詞「日本」が含まれているが、白書の分析には「日本」は重要な話題として必要な語であり、必ずしも無性格とは言えないだろう。

今後の予定としては、指示詞や接続詞などのほかの結束性を表す手段との関連も視野に入れて語彙的結束性の現れ方を総合的に記述したいと考えている。

---

7 今回使用したサンプルについては無性格語を排除しなくてもほとんど同じ結果であったが、どのような場合にこのリストが有効かは確認が必要である。

付表 本稿で用いた「無性格語」

語彙素	語彙素読み	品詞
余り	アマリ	副詞
余り	アマリ	形状詞-一般
有る	アル	動詞-非自立可能
言う	イウ	動詞-一般
行く	イク	動詞-一般
一	イチ	名詞-数詞
今	イマ	名詞-普通名詞-副詞可能
居る	イル	動詞-非自立可能
内	ウチ	名詞-普通名詞-副詞可能
円	エン	名詞-普通名詞-助数詞可能
円	エン	名詞-普通名詞-一般
御	オ	接頭辞
多い	オオイ	形容詞-一般
大きい	オオキイ	形容詞-一般
大きな	オオキナ	連体詞
置く	オク	動詞-非自立可能
於く	オク	動詞-一般
同じ	オナジ	形状詞-一般
同じ	オナジ	連体詞
思う	オモウ	動詞-一般
居る	オル	動詞-非自立可能
会	カイ	名詞-普通名詞-一般
方	カタ	接尾辞-名詞的-一般
方	カタ	名詞-普通名詞-助数詞可能
月	ガツ	名詞-普通名詞-助数詞可能
彼	カレ	代名詞
考える	カンガエル	動詞-一般
聞く	キク	動詞-一般
九	キュウ	名詞-数詞
位	クライ	名詞-普通名詞-副詞可能
来る	クル	動詞-非自立可能
五	ゴ	名詞-数詞
こう	コウ	副詞
五十	ゴジュウ	名詞-数詞
事	コト	名詞-普通名詞-一般
此の	コノ	連体詞
此れ	コレ	代名詞
三	サン	名詞-数詞
さん	サン	接尾辞-名詞的-一般
三十	サンジュウ	名詞-数詞
氏	シ	接尾辞-名詞的-一般
氏	シ	名詞-普通名詞-一般
四	シ	名詞-数詞
然し	シカシ	接続詞
七	ナナ	名詞-数詞
七	シチ	名詞-数詞
自分	ジブン	名詞-普通名詞-一般
仕舞う	シマウ	動詞-非自立可能
者	シャ	接尾辞-名詞的-一般
知る	シル	動詞-一般
十	ジュウ	名詞-数詞
十	トオ	名詞-数詞
為る	スル	動詞-非自立可能
生活	セイカツ	名詞-普通名詞-サ変可能

語彙素	語彙素読み	品詞
千	セン	名詞-数詞
そう	ソウ	副詞
そう	ソウ	名詞-助動詞語幹
そう	ソウ	形状詞-助動詞語幹
そして	ソシテ	接続詞
其の	ソノ	連体詞
其れ	ソレ	代名詞
第	ダイ	接頭辞
対する	タイスル	動詞-一般
出す	ダス	動詞-非自立可能
達	タチ	接尾辞-名詞的-一般
為	タメ	名詞-普通名詞-副詞可能
つく	ツク	動詞-一般
付く	ツク	動詞-非自立可能
強い	ツヨイ	形容詞-一般
的	テキ	接尾辞-形状詞的
出来る	デキル	動詞-非自立可能
出る	デル	動詞-一般
度	ド	名詞-普通名詞-助数詞可能
どう	ドウ	副詞
時	トキ	名詞-普通名詞-副詞可能
所	トコロ	名詞-普通名詞-副詞可能
所	トコロ	名詞-普通名詞-一般
共	トモ	名詞-普通名詞-一般
共	トモ	接尾辞-名詞的-副詞可能
取る	トル	動詞-一般
無い	ナイ	形容詞-非自立可能
中	ナカ	名詞-普通名詞-副詞可能
何	ナニ	代名詞
何	ナン	名詞-数詞
成る	ナル	動詞-非自立可能
二	ニ	名詞-数詞
日	ニチ	名詞-普通名詞-助数詞可能
二十	ニジュウ	名詞-数詞
日本	ニッポン	名詞-固有名詞-地名-国
人	ニン	接尾辞-名詞的-一般
年	ネン	名詞-普通名詞-助数詞可能
はいる	ハイル	動詞-一般
場合	バアイ	名詞-普通名詞-副詞可能
八	ハチ	名詞-数詞
日	ヒ	名詞-普通名詞-副詞可能
人	ヒト	名詞-普通名詞-一般
一	ヒト	名詞-数詞
ひとり	ヒトリ	名詞-普通名詞-副詞可能
百	ヒャク	名詞-数詞
方	ホウ	名詞-普通名詞-一般
僕	ボク	代名詞
程	ホド	名詞-普通名詞-副詞可能
前	マエ	名詞-普通名詞-副詞可能
また	マタ	接続詞
また	マタ	副詞
万	マン	名詞-数詞
見る	ミル	動詞-非自立可能
目	メ	名詞-普通名詞-一般

語彙素	語彙素読み	品詞
目	メ	接尾辞-名詞的-一般
持つ	モツ	動詞-一般
物	モノ	名詞-普通名詞-一般
物	モノ	名詞-普通名詞-サ変可能
問題	モンダイ	名詞-普通名詞-一般
遣る	ヤル	動詞-非自立可能
行く	ユク	→行く(イク)
良い	ヨイ	形容詞-非自立可能
様	ヨウ	形状詞-助動詞語幹
因る	ヨル	動詞-一般

語彙素	語彙素読み	品詞
四	ヨン	名詞-数詞
四十	ヨンジュウ	名詞-数詞
等	ラ	接尾辞-名詞的-一般
零	レイ	名詞-数詞
六	ロク	名詞-数詞
分かる	ワカル	動詞-一般
訳	ワケ	名詞-普通名詞-一般
私	ワタクシ	代名詞
私	ワタシ	代名詞

付表の注

「余り」「円」「同じ」「方（かた）」「七」「十」「そう」「所」「共」「何」「まだ」「目」「物」「私」については、短単位での品詞が複数に渡っているため、該当するものを列挙した。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「テキストにおける語彙の分布と文章構造」による研究成果の一部である。データとして利用した BCCCWJ 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）

Károly, Krisztina.(2002) *Lexical Repetition in Text.* Peter Lang.

小木曽智信、間淵洋子、前川喜久雄(2011)「『現代日本語書き言葉均衡コーパス』における形態論情報付き XML フォーマット」、言語処理学会第 17 回年次大会予稿集、pp.352-355.

田中章夫(1973)「自動抄録処理におけるキー・ワードの性格」『電子計算機による国語研究 V』秀英出版、pp.141-184.

平林健治(2003)「日本人初級学習者の英文ライティングの結束性の視点からみた分析」、愛知新城大谷短期大学研究紀要、2-4、pp.67-76.

水谷静夫(1980)「用語類似度による歌謡曲仕分『湯の町エレジー』『上海帰りのリル』及びその周辺」、計量国語学、12(4)、pp.145-161.

水谷静夫(1983)『朝倉日本語新講座 2 語彙』朝倉書店

宮島達夫(1970)「語いの類似度」、国語学、82、pp.42-64.

山崎誠(2012)「共起語率の分布からみるテキストの語彙的特徴」、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、pp.221.226.

状態空間表現を用いた文章の特徴付け

馬場康維（統計数理研究所）
小森　理（統計数理研究所）

Feature Extraction of Sentence Structure Based on State Space Representation Model

Yasumasa Baba (The Institute of Statistical Mathematics)
Osamu Komori (The Institute of Statistical Mathematics)

１．はじめに

文章は文の連なりからなる。文は一連の語や記号の系列で成り立っている。そこで、語や記号の連なりを“状態”の系列とみなし状態間の推移で文を表現することにより文章の構造を表現し様々な分析に使えないかというのがこの研究の発端である。

“状態”の定義は分析の対象・目的によって変わる。形態素解析を利用してテキストデータを品詞の系列で表現する場合には、名詞、動詞、助詞などの品詞が状態に対応する。文の構造を抽出するには名詞句、動詞句といった品詞の結合した状態を用いた方が文の構造の分析には適している。より詳細な構造を分析の対象にするならば、名詞を名詞の種別に分割した状態を考えるというように状態の分割も必要である。さらに文章全体を構造化してとらえるには段落の状態を考慮する必要がある。このように“状態”は分析の場面、場面に応じて定義される。

この報告では、文章構造のモデル化の基礎的な研究として文を状態空間で表現し時系列としてとらえる試みについて述べる。ここで用いたデータは国立国語研究所共同研究プロジェクト「文章における語彙の分布と文章構造」により作成されたテキストデータの一部である。文章のあるいは文の解析にはまず文法的なモデルを用意し単語の意味を考慮するというような方法があるが、ここでは、データから得られる情報をもとに文の構造的な把握をするというプロセスによって文章構造のモデル化を図る。

２．品詞による表現

品詞状態による表現の例を示す。用いたデータは、近藤和敬、“ヒルベルトの数学における公理的方法からカヴァイエスの概念の哲学へ ”(以下、近藤論文と呼ぶ)をテキスト化し形態素解析を行って得られた品詞データである。テキストデータには、段落のタグがついており、“論文のタイトル＋著者名＋所属”は一つの段落として扱われている。表１は形態素解析の結果を示している。最も基礎的なこのタイプのデータを時系列的に表現しただけでも文の特徴が見いだせる。形態素解析の品詞のカテゴリーが異なった状態になるように表２のように数値を対応させた。この数値は便宜上割り振ったもので何らかの最適化をしたものではない。この数値を割り振られた品詞の状態空間を用いて文章の一部を表現してみると図１、図２のようになる。図１は、近藤論文の“タイトル＋著者名＋所属部分”である。図２は最初の文の時系列である。図１には句読点がないこと、動詞＋句点で終わっていないこと等、タイトルであることが類推できる特徴が存在する。一方、図２では文末が動詞＋句点という典型的な連結で終わっている。このことからも、状態空間表示により時系列を表現することで、文の特徴抽出が可能になることが推察される。

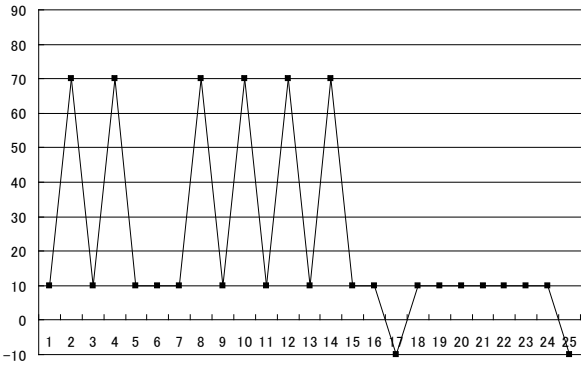


図１　形態素解析（品詞）による
タイトル部分の時系列表現

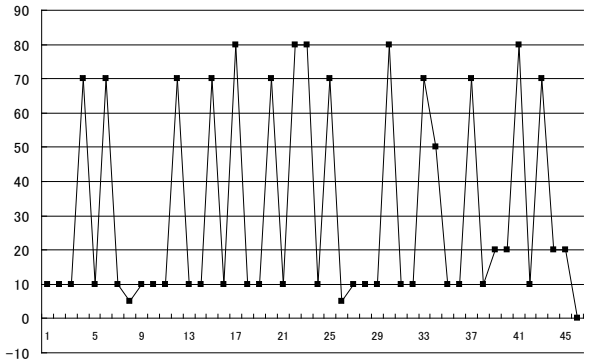


図２　形態素解析（品詞）による文の
時系列表現

表２　品詞等の状態表現

品詞等	状　態
句点	0
読点	5
名詞	10
動詞	20
形容詞	30
副詞	40
連体詞	50
接続詞	60
助詞	70
助動詞	80
接頭詞	90
その他	-10

表１　文の形態素解析

文字	品詞	品詞（詳細）	文節ID
数学	名詞	一般	1
基礎	名詞	一般	1
論	名詞	接尾	1
の	助詞	連体化	1
論争	名詞	サ変接続	2
の	助詞	連体化	2
結果	名詞	副詞可能	3
、	記号	読点	0
数学	名詞	一般	1
的	名詞	接尾	1
認識	名詞	サ変接続	1
の	助詞	連体化	1
確実	名詞	形容動詞語幹	2
性	名詞	接尾	2
の	助詞	連体化	2
アプリオリ	名詞	一般	3
な	助動詞	*	3
基礎	名詞	一般	4
付け	名詞	接尾	4
が	助詞	格助詞	4
不可能	名詞	形容動詞語幹	5
で	助動詞	*	5
ある	助動詞	*	5
こと	名詞	非自立	6
から	助詞	格助詞	6
、	記号	読点	0
合理	名詞	一般	1
論	名詞	接尾	1
的	名詞	接尾	1
な	助動詞	*	1
認識	名詞	サ変接続	2
論	名詞	接尾	2
は	助詞	係助詞	2
その	連体詞	*	3
説得	名詞	サ変接続	4
力	名詞	接尾	4
を	助詞	格助詞	4
半減	名詞	サ変接続	5
さ	動詞	自立	5
せ	動詞	接尾	5
た	助動詞	*	5
よう	名詞	非自立	5
に	助詞	副詞化	5
思わ	動詞	自立	6
れる	動詞	接尾	6
。	記号	句点	-1

表３　句等による状態表現

句等	状態
句読点	0
名詞句	10
動詞句	20
形容詞	30
副詞	40
連体詞	50
接続詞	60
その他	−10

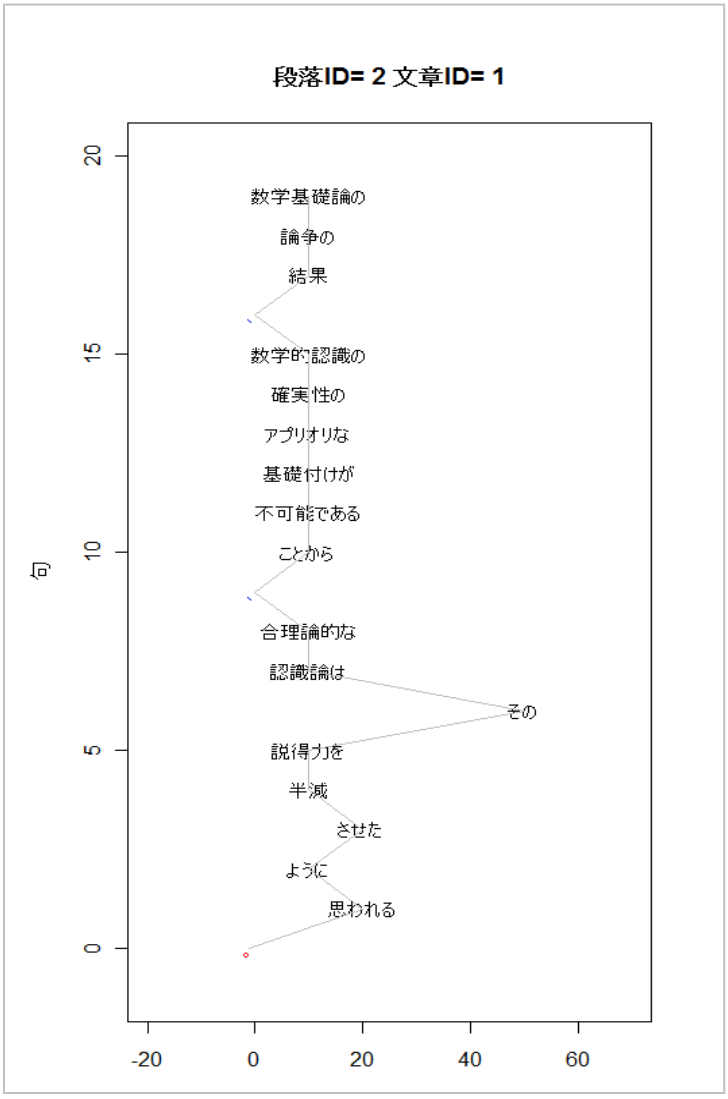


図３　助詞の違う要素表現（句）による文の時系列表現

３．句による表現

品詞による時系列を観察すると、数学＋基礎＋論のように名詞のみが連続する場合はそれらを一つの単語（便宜上名詞と呼ぶ）として扱うことが可能な場合がほとんどであることが分かる。さらに、例えば、“数学基礎論＋の”が次の“論争”を修飾している。名詞＋助詞で一つのかたまりと考える方が文の構造の表現には便利である。即ち、品詞による状態表現を特定の結びつきを示す品詞と品詞の状態に縮約した方が構造解析には都合がよい。そこで、品詞で表現した状態を縮約し助詞を中心にまとめた状態で文を表現したものが図３である。

このように助詞を中心にして状態をまとめてみると助詞の種類によりそれぞれの役割があることが分かる。そこで、名詞＋助詞を一つの状態とみなし、近藤論文のデータから推移行列を作ったものが表４である。表中、“名詞の”や “名詞を”はそれぞれ、名詞＋助詞（の）や名詞＋助詞（を）を表している。つまり名詞と助詞の連結ごとに状態を割り振ったこと

になる。なお句等はその他を除いて、出現頻度の多い順に並べてある。

表４　助詞による状態表現の推移確率(%)（近藤論文）

	名詞の	動詞	読点	句点	名詞を	名詞に	名詞が	名詞な	名詞は	名詞	その他
名詞の	18.6	0.5	0	0	13.6	7.1	11.3	7.3	6.8	1.3	34.2
動詞	9.7	0	7.2	26.2	5.5	9.4	7.5	2.8	7.2	5	20.1
読点	26.2	1.5	0	0	7.2	6.3	6.6	9.6	3.9	3	35.4
句点	11.1	0	0	0	3.6	3.6	8	4.4	12.9	9.3	46.4
名詞を	3.5	41.4	7.1	0	0	10.1	0.5	3.5	0.5	3.5	29.6
名詞に	4.3	41.7	7.5	0	1.6	3.2	1.6	6.4	0	2.7	30.5
名詞が	5.6	26	6.8	0	4	9	0	4.5	0	1.7	43.1
名詞な	18.3	0	1.8	0	20.7	3.7	9.8	2.4	7.3	0.6	34.9
名詞は	9.9	3.5	47.2	0	7	4.2	0.7	5.6	0.7	0.7	20.3
名詞	1.1	15.1	21.5	34.4	0	0	0	0	0	2.2	26.1

表５　句等の出現頻度
（近藤論文）

	度数	割合(%)
名詞の	382	11.7
動詞	362	11.1
読点	332	10.2
句点	226	6.9
名詞を	198	6.1
名詞に	187	5.7
名詞が	177	5.4
名詞な	164	5
名詞は	142	4.3
名詞	93	2.8
その他	1003	28.6

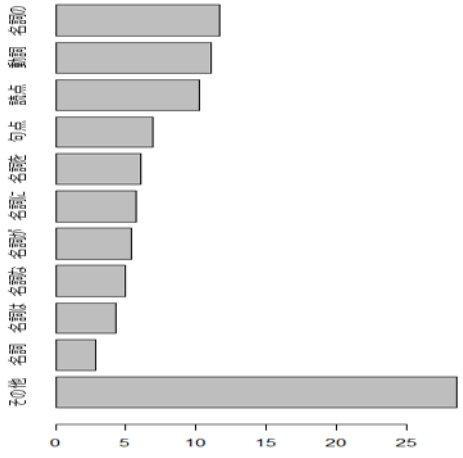


図４　句等の出現頻度（近藤論文）

４．論文の特徴の比較

比較のためにもう一つの論文データを用いて集計を行った。比較に用いたデータは、横地徳広“認識論的転回の地平を求めてーハイデガーとカント『純粋理性批判』”（以下、横地論文と呼ぶ）である。推移行列を表６に、各句等の状態の出現頻度を表７に示してある。この推移行列の状態も表４と同様に出現頻度の順に並べてある。

表４、表５、表６、表７を見ることにより、２つの論文の表現の比較が可能になる。

表６　助詞による状態表現の推移確率(%)（横地論文）

	読点	動詞	名詞の	名詞を	句点	名詞に	名詞は	名詞	名詞が	動詞て	その他
読点	0	0.4	22.9	8.2	0	6.5	9.5	4.8	8.2	0.4	38.7
動詞	10.2	0	6.6	8	28.8	8.4	8.4	8.8	1.8	0	18.4
名詞の	0	1.1	4.4	22	0	11	6.6	0.5	10.4	0	42.9
名詞を	0.6	32.9	1.9	0	0	13.3	3.2	5.1	0.6	13.9	28.1
句点	0	0	5.3	6.1	0	3	23.5	8.3	0.8	0.8	52.7
名詞に	7.6	35.6	1.7	3.4	0	1.7	0.8	3.4	3.4	15.3	26.4
名詞は	31.9	1.7	15.5	8.6	0	5.2	0	2.6	3.4	0	31.5
名詞	41	12	0	0	15.7	1.2	2.4	1.2	0	4.8	21.6
名詞が	5.1	23.1	9	6.4	0	7.7	0	1.3	0	11.5	36.1
動詞て	12	48	4	2.7	0	1.3	1.3	0	1.3	4	25.1

表７　句等の出現頻度
（横地論文）

	度数	割合(%)
読点	231	11.4
動詞	226	11.2
名詞の	182	9
名詞を	158	7.8
句点	133	6.6
名詞に	118	5.8
名詞は	116	5.7
名詞	83	4.1
名詞が	78	3.9
動詞て	75	3.7
その他	624	27.6

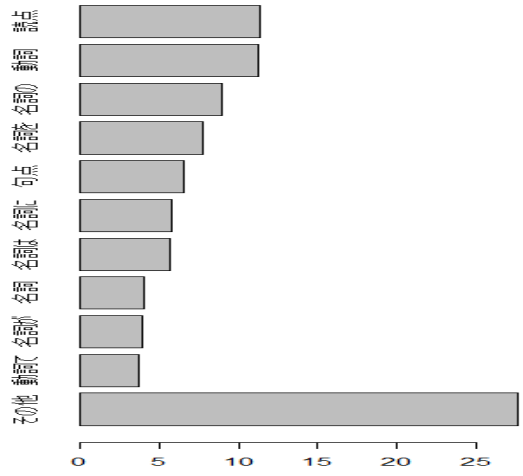


図５　句等の出現頻度（横地論文）

いずれの場合でも、動詞の次に続くのは句点である確率が高く、句点の次には文章のはじまりである名詞が続く確率が高い等の傾向が存在していることが分かる。これらの中で主語の役割を担うものは主に“名詞は”と“名詞が”であるが，句点からの推移確率が２つの論文で大きく異なることも分かる。文章のスタイルの違いが推移行列に反映されていると言え

る。ここでは，２つの論文のみを比較したが、多くの論文について推移行列の比較あるいは頻度の比較を行うことにより文献間の距離が算出できる。したがってそこから文献の類型化ができるであろう。

５．おわりに

品詞や句による状態空間表現について状態の縮約のプロセスを示した。これまでの文章の構文解析は，表４，表５，表６，表７に示した単語の出現割合または推移確率に注目したものが多かった。これはいわゆる静的な文章解析である。文章の構造理解には動的な要素も重要であり，図３に示したような時系列的な表現と解析が必要である。名文と呼ばれる文章には読者に訴えるリズムがあり，これが文章の内容理解を深める。この動的要素が色濃く出るものが詩また音楽の世界の歌詞である。今回の試みはまだ試行錯誤の段階であるが，今後大量のテキストデータの分析をすることによって、文章の背後にある時系列的な状態の推移の様々なパターンの把握と類型化を試みたい。また大量データの処理にあたり，機械学習に適した構造モデルを構築することも考えている。

参考文献

近藤和敬（2009）「ヒルベルトの数学における公理的方法からカヴァイエスの概念の哲学へ」哲学, Vol.60, pp.169-184.

田中章夫（1974）「句のエントロピーに基づく構文合成」言葉の研究第 5 集, pp.125-146.

中野洋（1974）「自動項分解析の構想」言葉の研究第 5 集, pp.147-157.

横地徳広（2005）「認識論的転回の地平を求めてーハイデガーとカント『純粋理性批判』」哲学, Vol.56, pp.270-282.

データの出処

国立国語研究所共同研究プロジェクト「文章における語彙の分布と文章構造」（チームリーダー山崎誠）

近代文語論説文を対象とした濁点の自動付与アプリケーション

岡 照晃 （奈良先端科学技術大学院大学）[†1]

Application of Automatic Labeling of *Dakuten* for Near-Modern Literary Style of Japanese

Teruaki Oka （Nara Institute of Science and Technology）

1 はじめに

日本語で初めての大規模均衡コーパスとなる**現代日本語書き言葉均衡コーパス**（**BCCWJ**）[1] が昨年公開された．これにより今，コーパスを利用した日本語研究が急速に増えつつある．

しかし，平安時代や明治時代といった古い時代の資料（**歴史的資料**）のコーパスは，現代語のコーパスほど整備が進んでいない．そのため，日本語研究の大きな位置を占める歴史的研究に，コーパスを用いることは未だ難しい．

歴史コーパスの整備が進まない原因の一つとして，歴史コーパスの整備に必要な**校訂**の作業コストが高いことが挙げられる．コーパス整備の中で必要とされる校訂作業は，歴史的資料の記述中にある**正書法に一致しない表記**をコーパスユーザにとって扱い易い形に修正し，可読性・検索性を上げることである．校訂の作業は現在，すべて人手で行われている．しかし，専門家にしか行えないため，作業人員の確保が大きな課題となっている．また，作業対象が膨大であるため，作業を完了するまでにも時間がかかる．

そこで本研究では，統計的機械学習手法を用い，歴史的資料の校訂作業を自動化することを最終的な目的とする．これにより，誰でも簡単に低コストかつ大規模に校訂作業を実施することが可能になると考えられる．そしてその第 1 段階として，校訂作業の中から**濁点付与**を取り上げ，自動化に取り組んだ [3, 4, 5].

本論文では，文献 [3, 4, 5] で提案した手法を実装した濁点の自動付与アプリケーションの紹介を行う．本アプリケーションはモデルファイルの学習に近代語のコーパス**太陽コーパス** [6] を利用し，対象の文体を近代文語論説文に限定している．しかし，濁点付与に用いるモデルを変更することで，中古和文や近世の資料にも容易に適応可能となっている．

2 校訂作業における濁点付与

歴史的資料の記述中では，図 1 に例示したように，**濁点が付いていることが期待される仮名文字に，濁点が付いていない**ことがよくある．図 1 では例えば,「欲せさる（ホッセザル）」の「さ（ザ）」から濁点が脱落している．本論文では，このような濁点の抜け落ちた文字のことを**濁点無表記文字**と呼ぶ.

表 1 を見ると，明治初期の資料でも濁音の仮名文字の約 83%（368/446）（総文字数の約 4%（368/8,423））が濁点無表記で書かれている．濁点無表記の文字をそのまま残しておくと，歴史コーパスのユーザにとっては読み辛く，検索にも不便である．そのため，濁点無表

[†1] teruaki-o[at]is.naist.jp

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

（『太陽』 1925 年 2 号 p.64 より抜粋）

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

表 1: 濁点と濁音の統計．**明六雑誌**[1]第 1 号（2011 年 12 月段階のデータ，総文字数:8,423）中に含まれる濁点文字と濁点を付けることが可能な文字の中から，実際の発音が清音の文字数と濁音の文字数の内訳を調査した．

表記＼発音	濁音（ガ，ザ，ダ，…）	清音（カ，サ，タ，…）	計
濁点文字（が，ざ，だ，…）	78	0	78
濁点を付けることが可能な文字（か，さ，た，…）	368 （濁点無表記文字）	1,273	1,641
計	446	1,273	

記文字を**濁点文字**（濁点付きの文字）に置き換えるという校訂作業が必要になる．この作業を**濁点付与**といい，現在はすべて人手で実施されている．

ただし，濁点付与のような校訂の作業を人手で行うのは非常にコストが高い．特に作業者が専門家に限られてしまうことが問題である．校訂の作業を行うためには，まず対象となる資料を読んで理解しなくてはならない．しかし，歴史的資料の中では現代とは語彙が異なる上，表記や語法の面でも多様であり，読解には専門的な知識が必要である．作業の信頼性を保証するためにも，作業者は歴史的資料の専門家に限定される．だが，そういった専門家を集めることは難しく，作業人員の確保が大きな課題となっている．

作業対象となる資料の量が膨大であることも問題である．例えば，国立国語研究所が構築した太陽コーパスは総文字数約 1,450 万文字の規模[2]である．これに対し，熟練した作業者でも 23 ページ分の資料（約 1 万 6,000 文字）の濁点付与に 1 日掛かりで取り組む必要がある[3]．上述の通り，作業者を大量に確保することは困難であり，人海戦術を取ることができない．そのため，大量の資料を少数人で少しずつ整備していくしかなく，整備を終えるまでに多大な時間が必要となっている．

また濁点付与を人手で行う場合，熟練した作業者であっても表 2 に挙げているような単純なミスをおかすことがよくある．濁点を付与すべき文字を見逃してしまうこともよくある．実際，2 人の作業者がそれぞれに行なった濁点付与の結果を比較してみたところ，濁点を付けた個所の一致率は 84%であり，一致しなかった原因のほとんどが濁点の付け逃しや，表 2 に示したような単純なミスによるものであった．また，コーパス整備の初期段階で 1 人の作業者が実施した濁点付与の結果を完成後のコーパスと見比べた．すると作業者が単独で濁点付与を行なった結果では，濁点無表記文字のおよそ 7%を見逃していたことが分かった．

国立国語研究所では太陽コーパスに引き続き，更なる近代語コーパスの整備が進められて

[1] 1874 年（明治 7 年）～1875 年（明治 8 年）の間に明六社から発行された啓蒙雑誌．全 43 号．
[2] 2005 年 11 月段階のデータ．
[3] 1 日の総作業時間を 5～6 時間とした場合．

表 2: 実際に人手で濁点のアノテーションを行なった場合のミスの例.

ミスのタイプ	正解のアノテーション	実際に作業者が行なったアノテーション
濁点文字に置き換える際の濁点無表記文字の消し忘れ	わが御ためめむぼくなければわたり給事はなし	わが御ためめむぼほくなければはわたり給事はなし
濁点を付与する際の濁点挿入位置の間違い	我が゛外交をして卑屈の平和にあらず゛、	我が゛外交をして卑屈の平和にあら゛す、

図 2: AYTC～にごり ONLY～（ブラウザ内実行時）

いる．しかし，校訂作業の負担から，現場ではその自動化を望む声が上がっている．

そこで本研究では，統計的機械学習の手法を用いて歴史的資料の校訂作業を完全自動化することを最終的な目的とする．これは従来行われてこなかった新しい試みである．歴史的資料の校訂はこれまですべて人手で行われてきたが，作業者不足と作業対象の膨大さから実施に高いコストを必要とした．また人手の作業にミスはつきものである．しかしその作業を計算機に行わせることで，大規模な校訂も低コストで実現でき，単純なミスもなくすことができると考えられる．

自動校訂技術開発の第一段階として，濁点付与の作業を取り上げ，自動化に取り組んだ [3, 4, 5]．最初の目標として濁点付与を取り扱ったのは，濁点付与のタスクが「濁点が抜け落ちた文字に濁点を補う」と明確で取り組みやすかったためである．またタスクが単純な割に，校訂処理の中での必要性は高い．これはコーパスの可読性や検索性向上の観点から，濁点の抜け落ちている問題が無視できないためである．

本論文では，文献 [3, 4, 5] で提案した手法を実装した濁点自動付与アプリケーション **AYTC～にごり ONLY～**（図 2）について述べる．手法の詳細等については文献 [3, 4, 5] に預け，ここではアプリケーションの使い方について述べる．

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

図 3: 濁点付与の対象となる文字.

3 AYTC〜にごり ONLY〜とは

AYTC は，現在開発中の歴史的資料自動校訂ツールの総称である．本論文で述べる**にごり ONLY バージョン**では，先行開発した濁点自動付与モジュール**にごり**を先行実装している．今回の実装では，対象の文体として近代文語論説文のみを扱い，濁点付与の対象となる文字は平仮名と，繰り返し記号である**くの字点**に限っている（図 3）．ただし以下のような処理をアプリケーション内部で行うことによって，明六雑誌のような漢字片仮名交じり文への対応も行なっている．

1. 入力文章中の片仮名文字をすべて平仮名文字に置換.
2. 濁点の自動付与を実行.
3. 濁点を付与した文章の中の平仮名文字を再度すべて片仮名文字に置換し，出力する.

くの字点の表記には，「／＼ [U+3033 U+3035]，／〝＼ [U+3034 U+3035]」と「く [U+3031]，ぐ [U+3032]」のみを認め，「〜〜」や「〜〵」のような表記はくの字点とはみなさない.

本アプリケーション利用の際は，実装手法による濁点付与の精度が 100%ではないことに注意が必要である．濁点付与の性能については 6 章もしくは文献 [3, 4, 5] を参照してほしい.

AYTC〜にごり ONLY〜（以下，単に AYTC）の開発に当たって，多くの文系研究者が本アプリケーションを利用可能なよう，以下の要件を満たすものを目指した.

1. 幅広い PC 環境で動作可能.
2. 利用者が自分の PC にインストールせずとも利用可能．もしくはインストールが簡単.
3. 操作方法が単純明快.

要件 1 と 2 を満たすため，AYTC は Silverlight アプリケーション[4]として開発を行なった．Silverlight アプリケーションはオンライン環境のブラウザ上で動作するアプリケーションであり，アプリケーション自体をユーザの PC にインストールすることなく利用できる．また Silverlight アプリケーションの動作には，Microsoft が提供する Web ブラウザ用プラグイン Silverlight[5]をインストールする必要があるが，Windows OS の PC には購入時からプリインストールされていることもあり，PC の操作に不慣れな者にも敷居は低いと考えられる．プリインストールされていない場合でも，下記のサイトから簡単にインストールすることができる．ブラウザには Microsoft の Internet Explorer だけでなく，Firefox や Chrome，Safari が対応している．また Silverlight アプリケーションは Windows OS だけでなく，Mac OS 上でも動作可能である.

Silverlight アプリケーションは基本的にオンライン環境のブラウザ上で動作する（ブラウザ内実行）．しかし，アプリケーションを各ユーザの PC へインストールすることで，通常

[4]Silverlight のバージョンは 5.
[5]http://www.microsoft.com/ja-jp/silverlight/

図 4: メニューバー．通常，各項目の背景は黒だが，選択時には赤色に塗られる．

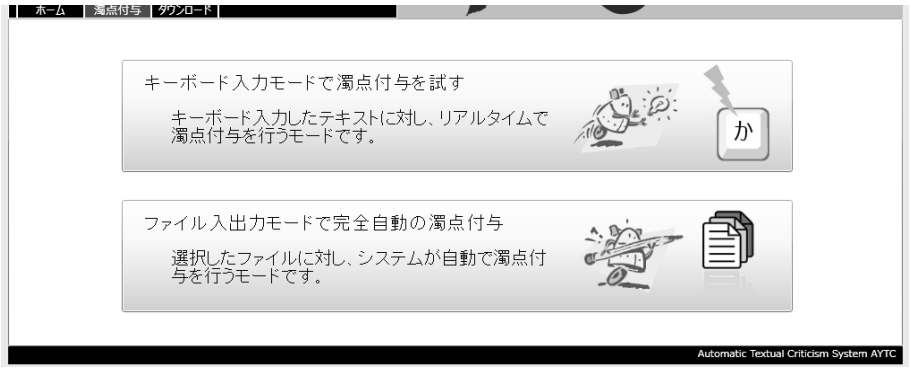


図 5: 濁点付与モード選択画面．

のデスクトップアプリケーションと同様，ブラウザを介さずオフライン環境での動作が可能になる（ブラウザ外実行）．AYTC もブラウザ外実行を念頭に入れて開発を行なっている．インストールには単純な 1 ステップの操作しか必要とせず，面倒な設定等は一切必要ない．

要件 3 を満たすために，AYTC では単純かつ直観的な操作方法を実現している．複雑な操作は一切必要ない．以下の 4 章と 5 章では，AYTC の操作方法について述べる．

4 AYTC による濁点の自動付与

メニューバー（図 4）の項目「濁点付与」をクリックすることで，濁点の自動付与を行うための画面に切り替わる（図 5）．この画面でまず，濁点の自動付与を行うためのモードを選択する．AYTC では，以下の 2 種類のモードで濁点の自動付与を行うことができる．

・キーボード入力モード

・ファイル入出力モード

キーボード入力モードでは，ユーザがキーボードやクリップボードから入力した文章に対し，AYTC がリアルタイムで濁点自動付与を行なっていく．これに対し，ファイル入出力モードでは，指定されたテキストファイルに対して，濁点自動付与が一括で行われる．

以下に，この 2 種類のモードの使い方を述べる．

4.1 キーボード入力モード

図 5 の画面で「キーボード入力モードで濁点付与を試す」をクリックすると，キーボード入力モードの画面（図 6）に切り替わる．このモードでは，入力用テキストボックス（図 6 左）に入力された文章に対し，リアルタイムに濁点自動付与が行われる．濁点付与を行なった文章は出力用テキストボックス（図 6 右）に表示される．この時，濁点を付与した文字は赤字で印字される（デフォルトは黒字）．

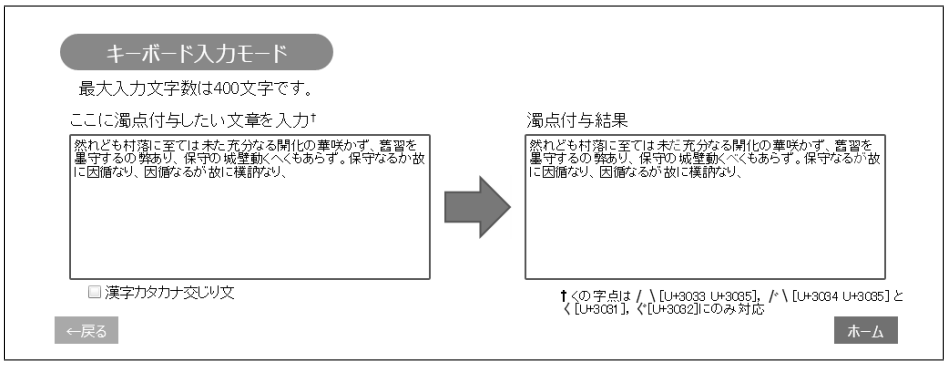


図 6: キーボード入力モード

入力用テキストボックスに入力可能な最大文字数は 400 文字に設定してある．また，入力用テキストボックスの下にあるチェックボックスにチェックを入れることで，漢字片仮名交じり文の入力にも対応する．出力用テキストボックスでは，くの字点が「く [U+3031]，ぐ [U+3032]」に統一表記される．

4.2 ファイル入出力モード

図 5 の画面で「ファイル入出力モードで完全自動の濁点付与」をクリックすると，ファイル入出力モードとなる．ファイル入出力モードでは，以下の順番で入出力ファイルを選択する．

1. 濁点付与を行いたい文章が打ち込まれた「入力ファイル」の選択
2. 濁点付与を行なった後の文章を書き出すための「出力ファイル」の選択

ファイル選択後に表示される「濁点付与」ボタンをクリックすることで，入力ファイルに対する濁点の自動付与結果を出力ファイルに得ることができる．

4.2.1 入力ファイルの選択

図 5 の画面で「ファイル入出力モードで完全自動の濁点付与」をクリックすると，入力ファイルの選択画面（図 7）に切り替わる．この画面ではまず,「参照」ボタンをクリックし，表示されたファイル選択ダイアログから濁点付与を行いたいテキストファイルを指定する．入力ファイルの文字コードは UTF-8 のみに対応しているが，BOM の有無や改行コードには依らない．参照ボタンの左上にあるチェックボックスにチェックを入れることで，漢字片仮名交じり文の入力にも対応する．

入力ファイルにはテキストファイルの他にも，太陽コーパスと同形式の XML ファイルを指定することができる．XML ファイルはファイル識別子.xml で自動認識される．AYTC は XML ファイルの記事・引用タグの属性「文体」で文語文と口語文を区別し，文語文にのみ濁点付与を実行する．

もし，ファイル読み込み時に不具合が生じた場合には，図 8 のようなエラーメッセージが表示される．ファイルの読み込みが問題なく行えると，画面右下の「次へ」ボタンが有効になる．

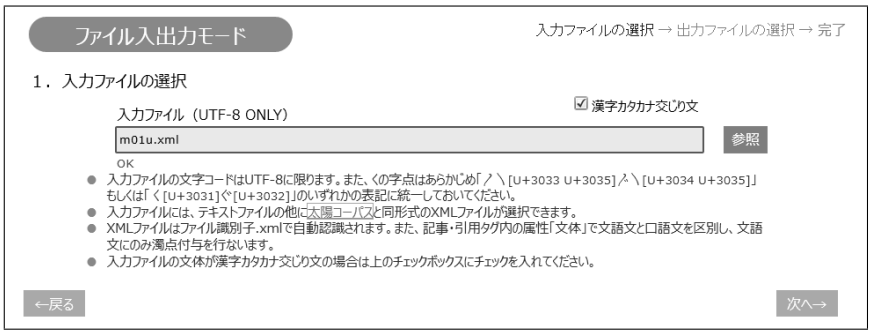


図 7: ファイル入出力モード：入力ファイル選択画面.

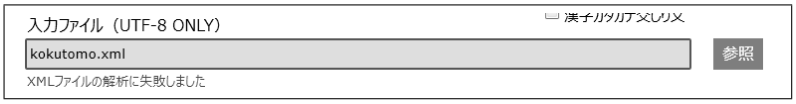


図 8: 入力ファイルの読み込みエラー.

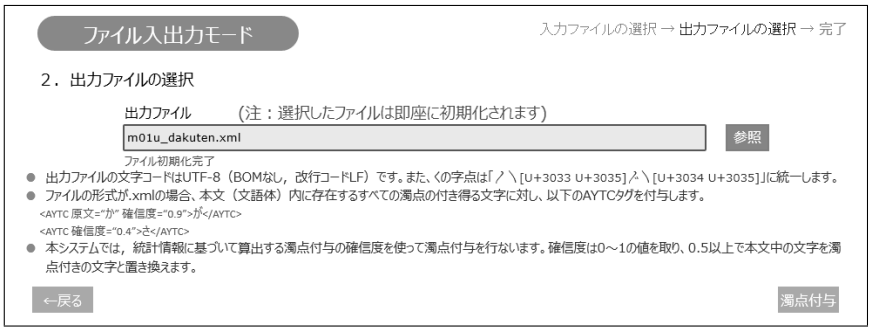


図 9: ファイル入出力モード：出力ファイル選択画面.

4.2.2　出力ファイルの選択

入力ファイル選択画面で「次へ」ボタンをクリックすると，出力ファイル選択画面（図 9）に切り替わる．ここでも前画面と同様,「参照」ボタンをクリックし，表示されたファイル選択ダイアログから，濁点付与結果を書き出したいファイルを指定する．この際，選択されたファイルは直ちに初期化され，空のファイルになってしまうことに注意が必要である．出力ファイルの文字コードは UTF-8（BOM なし，改行コード：LF）で固定となっている．また，くの字点の表記も「／＼ [U+3033 U+3035]，／″ ＼ [U+3034 U+3035]」に統一される．

AYTC には，ファイルのコンバート機能は搭載しておらず，出力ファイルの形式は入力ファイルの形式に準じたものとなる．そのため例えば，テキストファイルでの入力を XML ファイルとして出力することはできない．

ファイルの形式が XML の場合，AYTC は文語記事（もしくは文語引用）の本文中に存在するすべての濁点付与対象文字（図 3）に対し，以下の AYTC タグを付与する．

・濁点付与を行なった場合：〈AYTC 原文=”か” 確信度=”0.9”〉が〈/AYTC〉

– 本文は濁点文字に置き換え，濁点の付いていない元の文字をタグ内の属性「原文」に残す.

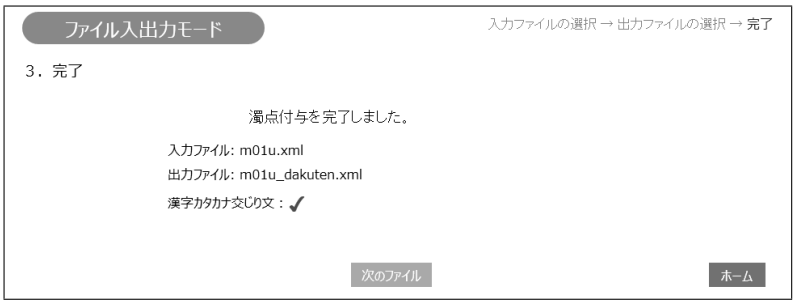


図 10: ファイル入出力モード：濁点付与完了画面.

図 11: インストール画面.

・濁点を付与しなかった場合：〈AYTC 確信度=”0.4”〉か〈/AYTC〉

– AYTC タグを付けるだけで，本文への変更は行わない．また，タグ内に属性「原文」を含まない．

AYTC では，統計情報に基づいて算出する**濁点付与の確信度**を使って各文字に濁点付与を行なっている．確信度は 0〜1 の値を取り，0 に近いほど濁点を付与することへの確信が低く，反対に 1 に近いほど確信が高い事を表している．AYTC では，確信度 0.5 以上で本文中の文字を濁点文字と置き換える処理を行う．各文字に対する確信度は，AYTC タグ内に保持してあるので，タグを見れば各文字がどのくらいの確信をもって濁点を付けたり付けなかったりしたかを確認できる．

出力ファイルを選択後，画面右下の「濁点付与」ボタンが有効になる.「濁点付与」ボタンをクリックすることで，濁点の自動付与が実行され，濁点付与完了画面（図 10）が表示される．

将来的には，この出力ファイルの選択画面と完了画面の間に，濁点付与の対象となっている文章の年代や文体を選択できる画面を挿入する予定である．本アプリケーションが対象とする文体は現在，近代文語論説文だけだが，濁点付与に使うモデル[6]を変更するだけで，中古や近世，近代の口語文の濁点付与にも容易に適応できる．そのため今後のアップデートでは，モデルを自由に選択できる機能を追加することを考えている．

5　AYTC のインストール

メニューバー（図 4）から項目「ダウンロード」をクリックすることでダウンロード用画面へ移動する．このとき，ユーザの PC に AYTC がイントール済みか否かによって，以下の 2

[6]機械学習において,「解決したい問題を数値化する方法」を「モデル」という.

図 12: インストールダイアログ.

表 3: 濁点付与の性能評価.

評価用コーパス	適合率.[%]	再現率.[%]	F 値
SUN-TEST	70.6	96.0	81.4
NF-TEST	95.8	98.0	96.9
M6-TEST	94.5	98.2	96.3

種類の画面のうち一方が表示される.

・ ユーザの PC へ AYTC が未インストールの場合：AYTC のインストール画面（図 11）

・ ユーザの PC へ AYTC がインストール済みの場合：インストール済みを通知する画面

5.1 AYTC のインストール方法

AYTC のインストール画面（図 11）中央の「AYTC をインストール」ボタンをクリックすると，図 12 のようなダイアログが表示される．ここで適宜チェックボックスにチェックを入れ,「OK」ボタンをクリックするだけで，ユーザの PC へ AYTC が自動的にインストールされる.

ブラウザ上で稼働していたアプリケーションがそのままインストールされるので，ブラウザ外実行でも，ブラウザ内実行と全く同じ感覚でアプリケーションを使用できる.

6 AYTC による濁点付与の性能評価

AYTC の濁点付与の性能評価実験を行なった．未校訂の近代文語論説文を対象とし，濁点付与の適合率と再現率を調べた．実験設定及び評価に使用したコーパスは文献 [5] と同じであるため，詳しくはそちらを参照してほしい.

ただし，文献 [5] とは異なり，濁点付与の確信度を 0〜1 の値として得るために，分類器の学習には LIBLINEAR[7]（バージョン 1.8）のロジスティック回帰を使用した [2]．またブラウザ上でも軽快に動作できるよう，学習時に L1 正則化を使い，モデルのパラメータの数を抑えた．[8]

結果を表 3 示す．詳細な結果の考察やエラー分析については，文献 [3, 4, 5] を参照してほしい.

[7] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

[8] モデルは皆パラメータを持ち，一般に，パラメータの数が多いほど精密なモデルであるが，モデルのサイズは大きくなる.

7 おわりに

本論文では，近代文語論説文を対象とした濁点の自動付与アプリケーション AYTC の紹介を行なった．AYTC は Silverlight アプリケーションとして開発を行い，幅広い環境での動作と，簡単かつ直感的な操作を実現している．

今後の課題として，近世の資料や中古和文といった近代文語論説文以外の文体への対応が挙げられる．また，濁点付与以外の校訂作業（e.g., 送り仮名の正規化）の自動化にも今後取り組んでいく予定である．

謝辞

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

文献

[1] Kikuo Maekawa (2008)「Balanced Corpus of Contemporary Written Japanese」In *Proceedings of The 6th Workshop on Asian Language Resources* (ALR 2008), pp.101-102.

[2] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin (2008)「LIBLINEAR: A Library for Large Linear Classification」 *Journal of Machine Learning Research*, 9, pp. 1871-1874.

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

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

[5] 岡照晃（2012）「統計的機械学習による歴史的資料への濁点の自動付与」第 1 回コーパス日本語学ワークショップ予稿集, pp. 13-22.

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

口頭発表（2）

9月7日（金）　13:30～15:00

近代対訳コーパスにおける日韓語彙の諸相
-文体の異なる対訳コーパスの比較を通して-

張元哉（韓国・啓明大学人文学部）†

Some Phenomena of Japanese and Korean Vocabularies in Modern Parallel Corpora
-Through Comparison of Parallel Corpora in Literary and Colloquial Styles-

Wonjae Chang(Keimyung University)

１．はじめに

現代日韓の語彙（特に漢語）は、日中や韓中よりもその類似性（同形率）が高いことが知られている。これは、近代以降に韓国が語彙交流の相手を主に中国から日本に変えたことによるものであろう。実際、現代日韓の語彙の同形率の高さ（約90%前後）が近代にさかのぼるとそれほどの高さ(約60%)ではないこと、また日韓の語彙交流が始まった19世紀末以降、現代語に向かって同形率や日本製漢語が同様の増加曲線を描くことなどがわかっている （張2000、2003a）。

このように近代以降の新漢語の増加と輸入は現代日韓の語彙形成に影響を与えており、それは単語レベルだけではなく、語彙や表現のレベルにも及んでいたと考えられている。そうすると、これまで知られている現代日韓の語彙の量的な構成の類似点や相違点は近代語にはどうであったかという疑問が浮かぶ。

そこで、本稿では、近現代的な語彙的要素が混在し、語彙交流の初期である、19世末・20世紀初期の文体の異なる日韓対訳コーパスの比較を通して、語彙量、語種、品詞、語構成などの観点から近代日韓の語彙の諸相を探ることを目的とする。

ただし、今回の調査は、資料の制約と限られた調査量の問題などもあって、近代日韓の語彙の様子が十分に捉えられるまでには至っていない。また計画されていた調査がまだ完了していない途中報告であることも断っておきたい。

２．調査資料と調査方法

近代の日本語と韓国語という二国間の語彙の様子を探るために、注意しなければならないのが、資料の時期、文体（ジャンル）と調査基準であろう。まず調査資料について見てみよう。

２．１　調査資料

日本語と韓国語の語彙を対照するには、資料の時期や文体という条件が同質なものでなければならないことはいうまでもない。しかし、近代という時期は、日韓ともに文体やジャンルの概念が確立しておらず、近代を文体の実験期とまで言うほどである。また、時期の問題も、たとえば日韓の新聞や雑誌の語彙を調査するにしても時期的に資料の有無やずれの問題が生じるなど、同質の調査はそう簡単ではない。これらの問題を十分に考慮しないと、調査結果が日韓の違いなのか、時期的な違いなのか、文体の違いなのかがよくわからなくなる恐れがあるからである。

それで、本稿では、近代という時期をこれまで指摘されてきた日韓の語彙研究の時期の区切りを踏まえて1910年代前後までにし、文体やジャンルの問題を解決するために日韓の対訳資料を使うことにする。近代における日韓対訳資料は、文体を考慮し、金秉喆（1975）[1]、李漢燮（1985）、李建志（2000）などの研究を参考にした。

† Chang_wonjae@hotmail.com

[1] 金秉喆（1975）によると1895年から1910年までに刊行された韓国翻訳文学の79作品のうち日本語からの

２．１．１　文語体の資料[2]

日本の資料：『西洋事情』（福沢諭吉、1866-1870）
韓国の資料：『西遊見聞』（兪吉濬、1895)

『西遊見聞』(以下、K西遊)は、 福沢諭吉のもとで就学した韓国最初の留学生である兪吉濬の啓蒙書であり、それ以前とは違ったハングル漢字混用体として後代の文体に大きな影響を与えたものである。『西遊見聞』の構成と内容は『西洋事情』(以下、J 西洋)と似ており、全20編の中には著者が著述した部分と『西洋事情』から翻訳した部分があるという(李漢燮1985)。本研究では翻訳された『西洋事情』の日本語と翻訳した『西遊見聞』の韓国語を調査対象とする。以下は、そのリストである。

『西洋事情』	『西遊見聞』
初編巻之一「政治」	第五編「政府의治制」
外編巻之二「政府の職分」	第六編「政府의職分」
初編巻之一「兵制」	第十三編「泰西軍制의来歴」
初編巻之一「病院」	第十七編「病院」
初編巻之一「博物館」	第十七編「博物館及博物園」
初編巻之一「蒸気機関」	第十八編「蒸気機関」

ただし、本稿での調査報告は、上のすべての編までは調査できておらず、「外編巻之二「政府の職分」- 第六編「政府의職分」」を除いた範囲での結果であることを言っておく。今後調査できていない編はもちろん、同じ文体に属する別の言語作品も調査する予定である。

２．１．２　口語体の資料

日本の資料：『文七元結』（幕末～明治初期？）
韓国の資料：『東閣寒梅』(1911)

『東閣寒梅』(以下、K東閣)は、当時翻訳家として活躍した玄公廉の作品であり、日本語と韓国語がパラレルに書かれている特徴がある。落語・人情本の『文七元結』(以下、J 文七)を原作に翻案（地名と人名だけを変えている）したものであるが、日本語の部分は『文七元結』の口演速記をもとにしているが、どの速記かはまだ不明のようである（詳細は李建志2000を参照）。

本調査では、『東閣寒梅』(1911)の日本語と韓国語（『日鮮語新小説 東閣寒梅 』(玄公廉、文明社)を対象とし、発表者が電子化したものを使うことにする。

２．２　調査方法

２．２．１　調査単位の設定

本調査の調査単位は、日本語と韓国語において同じ基準で、比較的に調査者のゆれの少ない調査単位であること、本稿の調査目的の一つである語構成の調査ができること、本調査の結果と先行調査が比較できることという条件として設定した。長い単位としては、日本語は文節、韓国語は語節（文節と同様）単位を採用する。切り分けられた文節（語節）から日本語は助詞・助動詞・記号を、韓国語は助詞・語尾・記号を取り除いた部分を１単位語とする。

直接翻訳や重訳された作品が54作品があるという。

[2] 『西洋事情』は、岡島昭浩氏のホームページ（http://www.ne.jp/asahi/nihongo/okajima/bungaku.htm、黄美静氏の作成）から、『西遊見聞』は李漢燮氏のホームページ（http://nihon.korea.ac.kr/、リンク切れ）からダウンロードした。

２．２．２　調査単位の原則

1単位の長さと幅の詳細な原則は、基本的に国立国語研究所（1987）『雑誌用語の変遷』の「長い単位」に従い、韓国語も該当するものはこれによる。この細則と異なるものや韓国語において注意が必要と思われる語は、以下のとおりである。

①文末の「ものか、ものだ、わけだ、ことだ、ところだ」などの下線の語は、1単位語として扱い、実質名詞と別見出し語にする。

②「お店、お客」などの「お～」は、１単位語とし、「店、客」と統合しない。

③「（～ては）いけない、いかない、ならない」は、１単位語とし形容詞にする。韓国語の「안된다, 못하다, 못되다」も1単位語とする。その他は「副詞（안, 못）+動詞」に切る。

④「お～する」は、１単位語とする。これに対応する、韓国語の尊敬の接辞の시(shi)が付いた「가시다」も１単位語とし、「가다」と統合しない。

⑤「を以って、における、において、について」などは、本動詞と別の見出し語にする。韓国語も同様に扱う。

⑥「急に、楽に」などは、形容動詞の語尾「に」として扱うが、「誠に、実に」などは、辞書に見出し語として載っている場合、１単位語とし副詞にする。

⑦「ている、てある、てみる、てしまう」などの補助動詞は、本動詞と別見出し語にする。韓国語も同様である。

⑧韓国語の「하다」は、「する」の意味と、引用（言う）の意味として別見出し語にし、補助動詞の場合は、⑦と同様に扱う。

⑨日韓の品詞の違いによるものは、そのまま尊重する。例えば、「～たい」（助動詞）-「싶다」（形容詞）については、前者(日本語)は助動詞なので扱わないが、後者(韓国語)は扱う。

⑩単位語としての判断材料とした辞書は、日本語は『広辞苑』（第6版、2008）『新明解国語辞典』（第6版、2005）、韓国語は『ヨンセ韓国語辞典』（1998）『標準国語大辞典』（1999）である。

３．日韓の語彙量の対照

３．１　近代における文体別の日韓の異なり語数・延べ語数

『J西洋』・『K西遊』と『J文七』・『K東閣』の異なり語数と延べ語数は以下のとおりである。

表1 近代における日韓対訳コーパスの異なり語数・延べ語数

	J 西洋	K 西遊	J 文七	K 東閣
異なり	973	1194	1281	1257
延べ	1948	1988	3628	3848

上の表１からわかることは、文語体においては異なり語数･延べ語数ともに韓国のほうが多い反面、口語体においてはそうではないということである。ただ、口語体の語数から見てわずかな違いなので、あまり日韓の相違がないかもしれない。語数の違いについては、中野洋（1976：21）では「翻訳される側とする側とでは延べ語数もかわろう。される側よりする側の方が延べ語数が多くなるかもしれない。」という指摘があり、表１の語数の違いもそのように解釈できるかもしれないが、李鐘洙（2010）の聖書の調査（韓国：旧訳1911-日本：文語訳1917）や張元哉（2004）の日韓対訳新聞の調査はそれぞれ第3国語と韓国語が原典であり、韓国語の語数が日本語より多い結果になっている。このことから考えると、語数の多少は必ずしも翻訳の方向によるものではないようである。

近代の口語体において『Ｊ文七』の異なりが若干多いが、「お～、～様(さん)、ある・

御座る、する・致す、お～になる、お～する（致す）、お～なさる（下さる）」などの待遇表現が韓国語より多いと考えられる。

３．２　使用率順異なり語数の累積使用率の対照

では、文体別に使用率順異なり語数に対する累積使用率を見てみることにする。それを図示したのが以下の図1である。

図1からわかることは、今回の狭い調査範囲でのことではあるが、口語体より文語体において日韓の累積使用率の違いが目立ち、韓国語のほうが日本語より累積使用率が低いことである。累積使用率は高頻度群の使用率が影響を与えているので、日韓の高頻度群における文体的な違いを綿密に考察する必要があると思われる。

文語体における累積使用率が韓国のほうが低いことは、現代の日韓対訳新聞の調査でも同様な傾向であるが、近代語よりその格差は小さいこと、β単位の短い単位の調査では高頻度群では韓国語の方が累積使用率が高くなる（張元哉2003b）ことから高頻度群での時代的変化の様子を捉える必要もあると思われる。

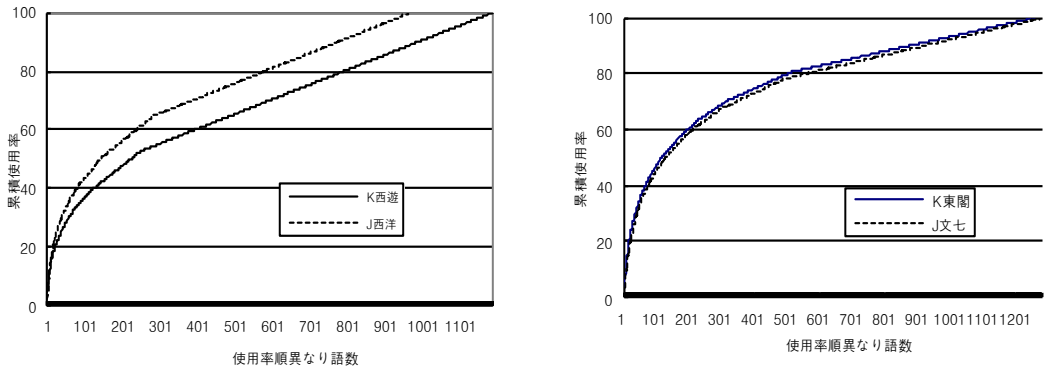


図1 近代における文体別の日韓の累積使用率

４．日韓の語種構成の対照

文体別に語種に分けて集計したのが以下の表2と図2である（延べ語数）。語種の略称は、和語・固有語（固）、漢語（漢）、外来語（外）、混種語（混）である。

表2と図2を見ると、まず、文体によって近代日韓の語種構成の相違が相当異なることがわかる。口語体では、日韓の語種構成がかなり類似していて、混種語が韓国語に多い（2倍ほど）ぐらいであるのに対して、文語体では、日韓の語種構成の相違が激しく、日本語は固有語、韓国語は漢語と混種語が非常に多い傾向が目立つ。韓国語の固有語は、1.7%を占めるに過ぎず、固有語をできるだけ排除した様子がうかがえる。

近代の文語体における日韓の語種構成の相違が、現代語の日韓の語種の違いと類似している（張元哉2004）が、近代語のような格差はそれほど大きくない。このことは語種における近代から現代への変化として捉えられる。

図3は、近代と現代における日本語の語種の比率から韓国語の語種の比率を引いた数値を図示したものである。たとえば、次のようになる。

-近代の固有語：日本語53.5%-韓国語1.7%＝51.8%

-近代の漢語：日本語38.0%-韓国語65.3%＝-27.0%

つまり、近代語から現代語に向かって日韓の語種構成の類似性が高まったということである。このような現代語への変化の裏づけとして日本語では国語研究所（1987）、韓国語では韓榮均（2009）の調査結果をみると次のようになる。

表2　近代における文体別の日韓の語種構成

	固	漢	外	混	計
J 西洋	1043	741	21	143	1948
%	53.5	38.0	1.1	7.3	100.0
K 西遊	33	1299	1	655	1988
%	1.7	65.3	0.1	32.9	100.0

	固	漢	外	混	計
J 文七	2936	536	11	145	3628
%	80.9	14.8	0.3	4.0	100.0
K 東閣	3019	507	1	312	3848
%	78.5	13.2	0.0	8.1	100.0

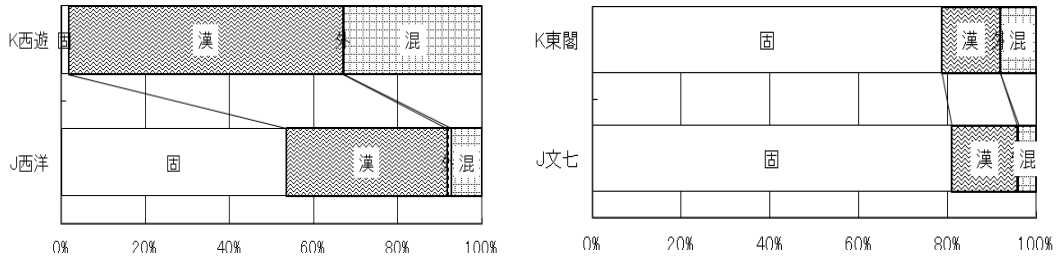


図2　近代における文体別の日韓の語種構成

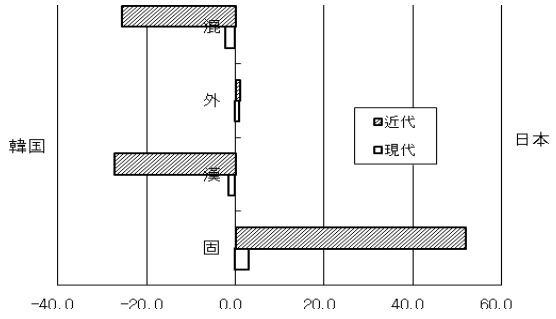


図3　近代と現代における日韓の語種の比率の差

まず、国語研究所（1987）は長い単位での「中央公論」の雑誌を調査したもので、和語と漢語は年によって増減の変動があるものの、一定の傾向は見られないこと、外来語と混種語は増えていく傾向が見られると指摘している。一方、韓榮均（2009）は文節単位での調査であり、19世紀末からの漢字ハングル混じり文の新聞の論説を調査したもので以下のような調査結果を示している。

表3　１９世紀以降の韓国新聞の論説における語種の変化（%）

語種	品詞	1890年代	1909年	1920年代	1930年代
漢語	代名詞	2.04	3.71	1.48	0.31
	連体詞	3.70	5.25	2.81	1.63
	一字名詞	11.82	9.11	3.42	4.40
混種語	一字漢語+hada	10.81	12.5	5.17	1.90

表３からいくつかの品詞の漢語と混種語が減少していることがわかる。

以上の先行論文の調査結果をまとめると、日本語は混種語が増加し、韓国語は漢語と混種語が減少していることから、近代語から現代語への日韓の語種構成の変化の様子が推測できる。

５．日韓の品詞構成の対照

５．１　文体における品詞構成

品詞分類の基準は、国立国語研究所（1964）『分類語彙表』に従う。文体別の品詞構成は以下のとおりである。

表4　近代における文体別の日韓の品詞構成

	体言	用言	相言	他	計
J 西洋	1147	537	215	48	1948
	58.9	27.6	11.0	2.5	100.0
K 西遊	1182	503	268	29	1988
	59.5	25.3	13.5	1.5	100.0

	体言	用言	相言	他	計
J 文七	1651	1154	646	177	3628
	45.5	31.8	17.8	4.9	100.0
K 東閣	1600	1213	858	168	3848
	41.6	31.5	22.3	4.4	100.0

表4からわかるように口語体のほうが文語体より比較的に日韓の品詞構成の違いが目立ち、口語体では日本語は体言類が多く、韓国語は相言類が多いことがわかる。これは、日本語が名詞志向表現、韓国語が動詞志向表現を好むことの現われであろうか（林八龍1995、金恩愛2003など）。すなわち派生名詞、複合名詞、動作性名詞や動名詞が主語、目的語、修飾語、述語に立つ際に、日本語の名詞が韓国語の動詞・形容詞になりやすい傾向があるということである。

たとえば、今回の調査資料からの例を見てみると、「不思議さ→異常である、夫婦別れをしない→夫婦が別別に別れなくて、若者が→若い人、心配を→心配して、お帰りを→帰ってください」などである（日本語→韓国語の日本語直訳）。また、これらの名詞が文語体よりは口語体によく出現する(？)ために口語体により品詞構成の相違が現れるのであろうか。

以下では日韓の口語体の相違が見られる体言類、相言類を中心に考察を行うが、比較のために文語体も合わせてデータをあげることにする。

５．２　文体における日韓の体言類と相言類の内訳

文体別に体言類と相言類の内訳を示すと以下のようである。

＜文語体における体言類と相言類の内訳＞

体言類	J西洋	K西遊
名詞	1039	1087
代名詞	35	18
動名詞[3]	15	58

相言類	J西洋	K西遊
形(形動)	69	91
副詞	89	83
連体詞	57	94

＜口語体における体言類と相言類の内訳＞

体言類	J文七	K東閣
名詞	1260	1342
代名詞	248	190
動名詞	110	45

相言類	J文七	K東閣
形(形動)	224	285
副詞	320	433
連体詞	102	139

口語体において「J文七」に体言類が多いのは、「代名詞」と「動名詞」の多さによるものであり、「K東閣」に相言類が多いのは、「形(形動)」「副詞」「連体詞」の多さによるものであることがわかる。文語体では、口語体に比べ量的に日韓の違いは見られないものの、同様の傾向が見てとれる。但し、動名詞は「K西遊」に多く、副詞は日韓ともにほぼ同じ量である。

近代韓国における文語体の動名詞の多さは、当時のハングル漢字混じり文（漢文直訳）の特徴であり、典型的な名詞用法の語というよりも述語部分に現れる語が多く、「漢語

[3] ここでの動名詞とは、日本語は動詞の連用形の転成名詞やそれを含む合成語、韓国語は動詞にㅁ(m)，ㄱ(ki)の名詞化接尾辞をつけた語である。

+함+이라」（21回）という形式のものが多い。この使い方が韓国の現代語ではあまり使われなくなっていることを考えると、 動名詞は、現代語の文語体と口語体においても日本語の方が多いのではないかと推測される。

続いて、文体別に体言･相言類における各品詞の異なり語数と用例を見てみると、表5のようになる

表5　文体における日韓の体言･相言類の異なり語数と用例

品詞		文語体		口語体	
		J西洋	K西遊	J文七	K東閣
体言	代名詞	**3語**: 此れ(30)、私(3)、此処(2)	**1語**: 此(18)	**30語**:何(40)、其れ(38)、御前(36)、此れ(26)	**24語**:나(47)、너(32)、무엇(18)、이것(13)、그것(12)
	動名詞	**11語**：競り売り(2)、妨げ(2)、妨げ(2)	**43語**:有함(3)、謂함(2)、各伸함	**75語**:使い(5)、御払い(5)、取り返し(4)	**34語**:보기(3)、하시기(3)、돌아오기(2)
相言	形容詞	**27語**:無い(27)、ごとし(9)、多い(4)	**71語**:無하다(14)、같다(2)、過하다(2)	**88語**:様(27)、無い(34)、有り難い(12)	**127語**:없다(29)、같다(21)、그렇다(10)
	連体詞	**5語**: 其の(42)、此の(10)、大いなる(2)	**3語**: 其(80)、彼(1)	**19語**:此の(35)、其の(28)、彼の(7)	**15語**：그(49)、이(47)、그런(11)
	副詞	－	－	**142語**:どうも(31)、マア(19)、どう(17)	**171語**：참(48)、좀(18)、참말(14)

()の数字は頻度

表5の異なる語数からもすでに述べた傾向とほぼ一致しており、日韓の各品詞における語の種類にもその違いが現れている。

ここで注目すべきことは、文語体も口語体も日本語は代名詞（これ、それ）が多く、韓国語は連体詞（이（この）、그（その））が多いことであり、また連体詞においては特に日本語の「その」より韓国語の「그」が多いことである。

前者は、日本語の『J文七』では代名詞（それ）で現れるものが、韓国語の『K東閣』では「이（この）、그（その）+名詞（人、時間、場所）」に対訳される例（14例）があるからである。

J文七：それがマアあんなに大きくなったんだものね。（八）
K東閣：그애기가참그러케커젓스닛가응(その子供が本当にそんなに大きくなったんだものね。訳は筆者、直訳)

後者は、指示詞の用法を大きく現場指示と非現場指示に分けた場合、非現場指示における日本語はコ・ソ・ア系が現れる反面、韓国語は이（こ）、그（そ）系しか現れない（宋晩翼1991）からではないかと思われる。コ系と이（こ）系は日韓同様に対応しているのでソ・ア系とユ（そ）系の違いであろう。上の例からすると、波線の部分の「あんな」が「그렇다」（そうだ）に対応するようなものであろう。

６．日韓の語構成の対照

近代日韓の文語体と口語体の語構成の構成は以下のとおりであり､2次結合以上のものは、最終結合を見て判断した。また、「お～する（いたす）」（23語）はここでは複合動詞に入れた。

表6　近代における文体別の日韓の語構成

	単純	派生	複合	計
J 西洋	1650	182	116	1948
	84.7	9.3	6.0	100.0
K 西遊	1158	772	58	1988
	58.2	38.8	2.9	100.0

	単純	派生	複合	計
J 文七	3014	338	277	3628
	83.1	9.3	7.6	100.0
K 東閣	3189	378	272	3848
	82.9	9.8	7.1	100.0

表6からわかることは、第一に、文語体に日韓の語構成の相違が認められ、日本語は単純語と複合語が多く、韓国語は派生語が多いということである。第二に、日本語は文体における語構成の比率がほぼ同じであるが、韓国語は文体によってかなり異なっていて、文語体に派生語の比率が多いということである。

では、文体ごとに複合語と派生語についてもう少し詳しく見ていくことにする。

６．１　文体別の複合語の内訳

複合語のうち、主な品詞を文体ごとにあげると以下のとおりである。

・文語体の複合語の内訳

J西洋：①名詞99回（73語）：共和政治(5)、人々(5)、蒸気機関(5)、立君独裁(3)、禽獣魚虫(2)、西洋諸国(2)、自主任意(2)、為替問屋(2)、フランス病院(2)、各々、鎧兜、見世物、工作貿易、貴族名家

K西遊：①名詞58回（46語）：蒸氣機關(4)、歐洲諸國(2)、禽獸蟲魚(2)、都下人民(2)、自由任意(2)、天下各國(2)、泰西各國(2)、熊虎獅犀、各其各目、古今物産、歐洲全幅、宮闕近處、飢寒疾苦、每兩六分

・口語体の複合語の内訳

J文七：①名詞71回（54語）：親子(3)、見ず知らず(2)、小間物(2)、二親(2)、行く末(2)、縞柄(2)、家々
②動名詞46回（31語）：取り返し(4)、勤め奉公(3)、身寄り(3)、心持ち(3)、人通り(3)、夫婦別れ(2)
③動詞141回（107語）：受け取る(6)、申し上げる(6)、飛び込む(5)、行き立つ(3)、しやがる(3)、見捨てる(2)

K東閣：①名詞112回（73語）：계집아해(-兒孩)(8)、큰돈(5)、고공사리(雇工-)(4)、품속(4)、거짓말(3)
②動名詞：2回（2語）：다다러오기、없어졌음
③動詞128回（65語）：돌아가다(9)、돌아오다(9)、가져오다(6)、지나가다(6)、내놓다(5)、들어가다(5)

文語体における『J西洋』に複合語が多いのは、複合名詞によるもので、『K西遊』より異なる語数も多く、そのほとんどが漢語である。一方、口語体における複合語の比率は表6からだとほぼ同じであるが、その内訳を見ると、『K東閣』には複合名詞が多く、『J文七』には複合動名詞、複合動詞が多い。

複合名詞の場合は、文語体では日本語に漢語が多い傾向にあるのに対して、口語体では韓国語に固有語と混種語の複合名詞が多いことがみられて面白い。また、口語体の『J文七』に複合動名詞や複合動詞が多いことは、5節で述べた動名詞と動詞の日韓の多少の傾向と前者は一致し、後者は複合において日本語の動詞が多くなるということになる。

これまで日韓の語構成についての調査があまり行われていなかったので、今回のデータにおける日韓の特徴が当時の言語現象を反映していて一般化できるかということと、現代語との違いなどについては、まだわかっていない。今後多くの調査が必要であるゆえんである。

６．２　文体における派生語の結合パターン

まず、文語体の派生語である。『K西遊』に派生語が非常に多く、以下に派生語における品詞と語種の関係をみることにする（○：和語、●：漢語、◎：外来語）。

＜文語体の派生語＞

	J 西洋	K西遊	
名詞	109	114	
(●+●)+●	30	48	J:世界中、動物園、天主教　K：世界上、動物園、基本形
●+(●+●)	5	7	J:新発明、全世界、総病院　K:大都會、大旨趣、新造物
(◎)+●	5	0	J:フランス帝
動詞	73	473	
(●+●)+○	63	269	J:一変する、発明する、改正する　K:收聚하다、發用하다
(●)+○	2	202	J:議する、害する　K:至하다、爲하다、作하다
○+(○)	3	0	J:相攻める、相戦う
副詞	0	25	
(●+●)+○	0	14	K:是故로、如此히、輕蔑히
(●)+○	0	9	K:順히、重히、輕히
形容詞	0	87	
(●+●)+○	0	56	K:口虛하다、輕便하다、過度하다
(●)+○	0	29	K:無하다、近하다、難하다

文語体における『K西遊』の派生語の多さは、動詞や形容詞の「一字漢字・二字漢字+hada」によるものであることがわかる。これは、4節の語種構成の特徴とつながるものであり、「漢字+hada」による語の表れは当時の文語体の特徴でもある。この造語法は韓国語の現代語に向けて減少する（4節）。名詞における各パターンは大差ない。

＜口語体の派生語＞

	J 文七	K東閣	
接頭辞:	144	10	
名詞			
お(ご)+名詞	112	0	
○+（○）	52	0	J:御金、御店
○+（○+○）	6	0	J:お屋敷
●+（●）	8	0	J:御礼、御縁
○+（●）	7	0	J:御客、御宅
お(ご)+動名詞	28	0	J:御払い御詫び
動詞			
お(ご)+動詞	14	0	J:御言う御出づ

	J 文七	K東閣	
接尾辞:	86	279	
名詞			
語基+さま(さん)	31	5	J:娘さん K:주인님
語基+さ・ki	6	46	J:可愛さ K:보기
(○)+○	7	14	J:奴等 K:그뭄께
(●+●)+●	11	7	J:年小者 K:輕薄兒
(●+●)+○	6	6	J:女郎屋 K:雇工꾼
動詞			
語基+する+hada	21	161	J:反抗するK:생각하다
形容詞			
語基+hada	0	32	K:未安하다

次は口語体の派生語の日韓の傾向である。口語体の混種語は、全体的には日韓の差が見られなかったが、詳しく見ると、『J文七』 は「接頭辞+語基」が多く、『K東閣』は「語基+接尾辞」が多い。日本語の「接頭辞+語基」には待遇性の接頭辞に固有語が付いたパターンが圧倒的に多く文語体ではあまり現れないパターンであり、韓国語の「語基+接尾辞」は文体の違いとは関係なく語基にhadaが付いた動詞と形容詞が多い。

ところで、『K東閣』に名詞化の接尾辞「기（ki）」が付いたパターンが多くみられるが、すでに述べた日本語に（複合）動名詞が多いことと、その数から考え合わせると、日本語の動名詞は、単純語の動詞の連用形とそれを含んだ複合語のパターンが多いことになる。

7．おわりに

以上のように近代語における日韓の語彙をいくつかの観点から眺めてきたのであるが、調査を行うに当たって、資料の選定や調査単位の長さと幅の問題などについて悩み、限界を多く感じた。

資料の選定については、現代語のようにジャンルや文体が確立されていないことはもちろん、当時期の日韓における資料の多少や時期のずれの問題で同質の日韓の語彙の比較に注意を要するところが多くあったからである。

調査単位については、二言語間の違いを明らかにするためには同じ調査単位の設定が必要であり、できる限り日韓の調査単位の設定の違いが調査結果に及ばないように注意を払ったつもりであるが、今後検討すべき問題もいくつかあろうかと思っている。

近代語における日韓の語彙研究は、語彙交流の研究を除けばあまり行われておらず、計量的な語彙の分析はいうまでもない。現代語より比較的日韓の類似性が低かった、近代語の日韓の語彙の特徴を明らかにすることは、近代以前の日韓の語彙の様子がわかることであり、また、現代語への変化を捉えられることにもなるのである。

これまで十分に構築されていない近代以降の日韓のコーパスを作成しつつ、近代語の共時的研究と、現代語への通時的研究を進めていきたい。

参考文献

韓榮均（2009）「文体 現代性 判別의 語彙的 準拠와 그 変化-1890년대～1930년대 논설문의 한자어 사용양상을 중심으로-」『口訣研究』23

金恩愛（2003）「日本語の名詞志向構造（nominal-oriented structure）と韓国語の動詞志向構造（verbal-oriented structure）」『朝鮮学報』188

金秉喆（1975）『韓国近代翻訳文学史研究』乙酉文化社

国立国語研究所（1964）『分類語彙表』秀英出版

国立国語研究所（1987）『雑誌用語の変遷』（国立国語研究所報告89）秀英出版

宋晩翼（1991）「日本語教育のための日韓指示詞の対照研究 - 「コ・ソ・ア」と「이・그・저」との用法について - 」『日本語教育』75

田中牧朗（2010）「雑誌コーパスでとらえる明治･大正期の漢語の変動」『国際学術研究集会漢字漢語研究の新次元予稿集』

張元哉（2000）「19世紀末の韓国語における日本製漢語-日韓同形漢語の視点から-」『日本語科学』8

----（2003a）「現代日韓両国語における漢語の形成と語彙交流」『国語学』54：3

----（2003b）「現代日韓語彙の対照研究-対訳コーパスを資料に-」『日本学報』(韓国日本学会)55：1

----（2004）「조사단위의 길이와 현대 한일어휘」『日本学報』(韓国日本学会)61：1

中野洋（1976）「「星の王子さま」6か国語版の語彙論的研究」『計量国語学』79

飛田良文（1973）「現代漢語の源流」『言語生活』259

李漢燮（1985）「『西遊見聞』の漢字語について-日本から入った語を中心に-」『国語学』141

李建志（2000）「海を渡った人情噺-朝鮮開化期の文学『東閣寒梅』と「文七元結」-」『江戸の文事』ぺりかん社

李鐘洙（2010）『韓・日 翻訳 聖書의 語彙 比較 研究-1900年 以後 刊行된 「로마서」를 中心으로』韓南大学大学院韓日語文学比較学科博士論文

林八龍（1995）「日本語と韓国語における表現構造の対照考察 - 日本語の名詞表現と韓国語の動詞表現を中心として - 」宮地裕・敦子先生古希記念論文刊行会（編）『宮地裕・敦子先生古希記念論集　日本語の研究』明治書院

손세모돌(1996)『국어보조용언연구』한국문화사

オンライン・コミュニケーション上での平均使用語彙数に関する研究

荒牧 英治（東京大学 知の構造化センター／JST さきがけ）†
増川 佐知子 （東京大学 知の構造化センター）
森田 瑞樹 （東京大学 知の構造化センター／医薬基盤研究所）
保田 祥（東京大学 知の構造化センター）

Study on Average Japanese Vocabulary for Online Communication

Eiji Aramaki (University of Tokyo)

１．はじめに

語彙数は，言語学分野にてもっとも熱心な関心が払われていた問題の１つであり，これに答えるため，数々の語彙調査がなされてきた．例えば，辞書の見出しを用いた調査(森岡健二 1951)やマスメディア（雑誌やテレビ）に出現する語の調査(国立国語研究所 1999; 玉村文郎 2002)が行われてきた．しかし，これらの調査の多くは，読者が理解できる語彙（理解語彙）の調査であり，使用された語彙（使用語彙）は行われていない．使用語彙を調査するのが困難であるのは様々な理由があるが，第一に膨大な調査コストがかかることが大きい．例えば，鶴岡調査(国立国語研究所 2012)では，被験者に対して 24 時間録音または速記を行ったが，このような実験を遂行するのに必要なコストは膨大なものになり，僅か 3 人の被験者のみの調査にとどまる．また，被験者不足の問題とは別に，サンプリングするという行為そのものが，平常時の発話とは異なる環境を実験参加者に強いる恐れがある．このため，「日本人の平均使用語彙量がどのくらいのものか，いっこうにわからない」(林四郎 1971)と言われてきた．

そこで，本研究ではウェブ上のオンライン・コミュニケーションのデータに注目する．オンライン・コミュニケーションのデータを用いれば，個人に紐付いたテキスト・データを大量に入手可能である．また，調査を後ろ向きに行う（過去のデータを用いる）ことにより，調査バイアスのないデータを得ることができる．このデータを用いて調査が可能な課題は数多くあるが，本研究では，このデータを用いて語彙数調査を試みる．

本研究ではネット上の発言を利用して，10 万人という大規模な人数で，対象者が実際に使用した語彙（使用語彙）を調査する．この結果，平均 7,000 語の語彙がオンライン・コミュニケーションで用いられてきたことが分かった．

２．材料

使用語彙調査を行うためには，誰がどんな発言したか，発言者とその使用語彙の紐付けが必要である．しかし，これが保証された Web 上のリソースは少ない．例えば，ブログは複数の執筆者により記述されることがある．また，大規模な引用が起こりうる．そこで，本研究では代表的なオンライン・コミュニケーションツールである Twitter に注目した. Twitter はユーザ数も 1,400 万人（2011 年 1 月）と多く，また，文字数制限のため大規模な引用がない．

本研究では以下の基準で約 10 万人の継続的な発言を得た．クローラの限界のため，各個人の発言について網羅性はなく，取得できていない発言もあるが，発言の取得に語彙のバイアスはなく，また，発言数から使用語彙を推定するため，語彙調査にあたっての問題はない．統計を以下に示す．

eiji.aramaki@gmail.com

- **データ期間**：2009/11/3 から 2010/3/25 の 143 日間（約 5 カ月間）
- **ユーザ数**：約 10 万人（99,964 人）
- **ユーザ抽出条件**：
  - 毎月 5 ツイート以上投稿している．(継続的な発言)
  - 総発言語数が 5000 以上.
  - 最初の 100 ツイート中に「の」が含まれている．(日本語使用者に限定)これは非日本語使用者を除くため行った.
- **全ツイート数**：約 2.5 億ツイート（253,482,784 ツイート）
- **全形態素数**: 約 43 億語（4,258,707,255 語）.

なお，形態素解析には juman7.0 (Kurohashi, Nakamura et al. 1994)を使用した．本研究では，この解析器が出力した形態素の単位を語とみなす.

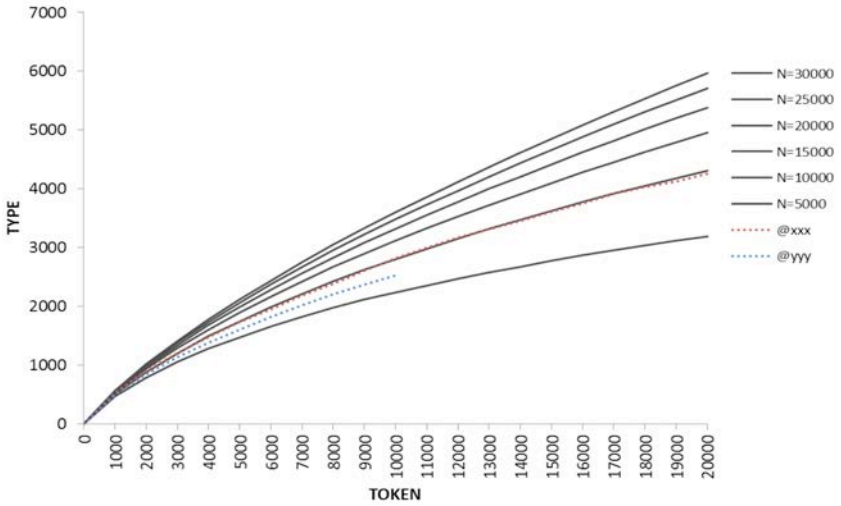


図 1: 語彙数 N（N=5000..30000）のタイプ・トークン曲線

X 軸はトークン数を示す．Y 軸はタイプ数を示す．実線は理論曲線で，語彙数（N）で区別されている．どの曲線も、トークンが多くなるにつれ，タイプの増加が鈍くなる．同じトークンに対しては，より大きな語彙数をもっていた場合ほど，タイプ数が多いと推定される．逆に，少ない語彙数では，比較的少量のトークンでタイプが飽和してしまう．破線は実際のユーザの例を示している。@xxx は N=10000,@yyy は N=7000 と推定された.

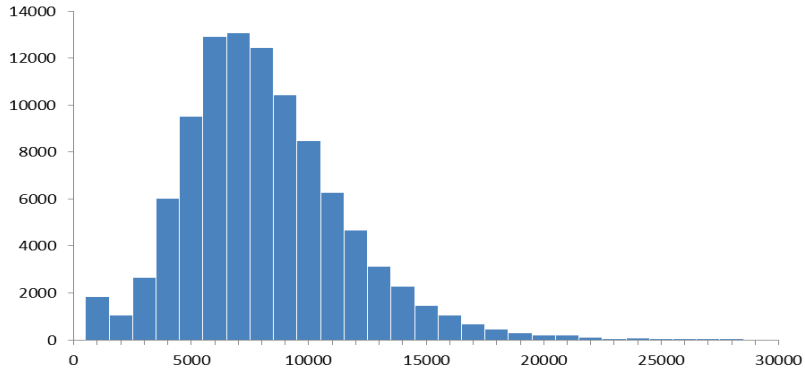


図 2: ユーザ数と語彙数．X 軸は語彙数，Y 軸はその語彙数を持つユーザ数を示す

３．方法

ある人がどれくらいの語彙をもっているかは，十分な長期間観察を行い，どのような語を使ったかを調べればよい．しかし，数万と言われる語彙すべてが使われるためには，気の遠くなるような長期間の観察が必要となってしまう．さらに，どれだけ観察しても，対象者のすべての語彙を観察したという保証を得ることはできず，調査を終了するタイミングが分からない．

そこで，本研究では，一定期間にユーザが発話した語数から，潜在的な語彙数を推測する．この推測はユーザがジップ則(Zipf 1935; Zipf 1949)に従って語を発生していると仮定することで可能になる．例えば，あるユーザが 10,000 語の語彙を持っていると仮定する．個人の発言がジップ則に従っているならば，1 位の語は全発言の 10.2%を占めるはずであり，2 位の語は 5.1%を占めるはずである．この場合，この対象者の発言を延べ 1,000 語集めた段階（以降，延べ語数をトークンと呼ぶ）で，期待される語の異なり（以降，タイプと呼ぶ）はおよそ 509 語である．逆に言えば，1,000 トークン集めて 509 タイプを得たならば，その人の語彙数は 10,000 語であると推測できる．

実際の様子を図 1 に示す．X トークンごとに Y タイプが観測されるはずという期待値を潜在語彙数（N=5,000…30,000）ごとにプロットしている．語彙数（N）は 1,000 きざみに 50 本を用意している．以降，この曲線をタイプ・トークン曲線と呼ぶ．タイプ・トークン曲線は語彙数 N が大きくなれば，傾きが急になる．これは，巨大な語彙を持っているならば，いくら観測しても，次から次へと新しいタイプが観測できるからである．逆に，語彙が少ないならば観測早々にして（すなわち，小さいトークンで）あらゆるタイプが出尽くし，曲線は飽和してしまう．

ここで，実際にコーパスから抽出したユーザのタイプ・トークン曲線を図中の点線で示す．ユーザ@xxx については実験の 5 ヶ月で 18,000 トークンが観測され(X 軸の値)，4,000 タイプが得られている（Y 軸の値）． このユーザのタイプ・トークン曲線ともっとも近い語彙数の曲線は N=10,000 であり，ここからこのユーザが潜在的に語彙数 10,000 を持つと推測できる．このようにして，12 万人すべての語彙数を推定することができる．

４．結果

推定された語彙数のヒストグラムを図 2 に示す．最も多い語彙数（最頻値）は 7,000 語であった．この 7,000 語という値は先行研究で推定されてきた理解語彙 40,000 語(林四郎 1971)と比較すると大きな格差がある．これは，普段，理解できる語彙のおよそ 1/6 のみしか使用していないことになり，なぜ，このようなギャップが生じるか，今後のさらなる研究が必要である．

４．１ 調査の限界

本研究は語彙数や語のユーザ数といった集計が困難な統計を得ることができるものの，以下の限界がある：

- **【形態素単位での集計のバイアス】**本研究の調査はすべて形態素の単位での集計であり，複合名詞は，ばらして集計されている．このため，先行研究の語彙数とずれる可能性がある．
- **【ユーザのバイアス】**オンライン・コミュニケーションに参加しているユーザは日本語話者の一部であり，偏った集団から語彙を採取している可能性がある．実際に，本研究で扱ったサービス「Twitter」では， 30%近くのユーザ東京に集中し，かつ，20 代のユーザを多いとされている(リンクシェア・ジャパン株式会社 2010)．このバイアスが語彙を実際よりも小さくしている可能性がある．

- **【環境のバイアス】**オンライン・コミュニケーションという環境事態が，語彙を変化させている可能性がある．例えば，キーボード／スマートフォンでの入力が語彙に影響している可能性がある．

上記をはじめ，材料／集計方法により様々なバイアスがあるが，先行研究の方法においてもバイアスは存在した．例えば，辞書の見出し語の調査では，選定した辞書の影響を大きく受けるであろうし，少人数の被験者を用いる場合も，日本人の均一なサンプルである保証はない．したがって，本調査が特に信頼できない調査という根拠にはならない．むしろ，膨大な人数の調査により，統計的には正確である可能性がある．

４．２ 応用可能性

本研究結果は様々に応用することができる．例えば，日本語会話辞書や旅行会話辞書など実用的な言語リソースに含まれる語彙は，実際に使用しない語彙を多く含んでいる事になる．よって，今後，旅行会話辞書の大幅なコンパクト化や，ユーザ数の多い語彙を優先して学習することで効率的な語彙習得なども可能となる．

謝　辞

本研究は，JST 戦略的創造研究推進事業（さきがけタイプ）「情報環境と人」及び，科研費補助金(若手研究 A) (挑戦的萌芽)による．本論文を書くにあたって有益な議論をいただいた東京大学医学部附属病院篠原恵美子氏に謹んで感謝の意を表する．

文　献

Kurohashi, S., T. Nakamura, et al. (1994). Improvements of Japanese Morphological Analyzer JUMAN. The International Workshop on Sharable Natural Language Resources.

Zipf, G. K. (1935). The Psychobiology of Language, Houghton-Mifflin.

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort, Addison-Wesley.

リンクシェア・ジャパン株式会社. (2010). "Twitter 利用実態調査."

国立国語研究所（1999). 高頻度語彙から見たテレビ放送語彙の特徴, 大日本図書.

国立国語研究所. (2012). "鶴岡調査."

林四郎（1971). "語彙調査と基本語彙." 国立国語研究所報告 **39**.

森岡健二（1951). "義務教育終了者に対する語彙調査の試み." 国立国語研究所年報 **2**: pp.95-107.

玉村文郎（2002). NAFL Institute 日本語教師養成通信講座８ 日本語の語彙・意味, アルク.

『明六雑誌コーパス』の開発　―近代語コーパスのモデルとして―

近藤 明日子　（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）††
須永 哲矢　（国立国語研究所コーパス開発センター）†††
田中 牧郎　（国立国語研究所言語資源研究系）††††

Development of *Meiroku Zasshi Corpus*: A Model of Modern Japanese Corpora

KONDO Asuko　(National Institute for Japanese Language and Linguistics)
OGISO Toshinobu　(National Institute for Japanese Language and Linguistics)
SUNAGA Tetsuya　(National Institute for Japanese Language and Linguistics)
TANAKA Makiro　(National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所共同研究プロジェクト（独創・発展型）「近代語コーパス設計のための文献言語研究」（プロジェクトリーダー：田中牧郎、プロジェクト略称「近代語コーパス」）では、同研究所共同研究プロジェクト（基幹型）「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）が対象とする古代から近世までの通時コーパスと、2011 年に公開された現代語の大規模均衡コーパスである『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）との間をつなぐ位置にある、近代語のコーパスのあり方について研究を行ってきた。そして、その成果を活かし、今後の本格的な近代語コーパス構築に向けてのモデルとして『明六雑誌コーパス』を開発中であり、2012 年秋の公開を予定している。本発表では、この『明六雑誌コーパス』の概要を紹介する[1]。

２．『明六雑誌コーパス』以前の近代語のコーパス

これまでに公開された大規模な近代語のコーパスとして国立国語研究所（編）(2005)『太陽コーパス』がある。『太陽コーパス』は、1895（明治 28）年から 1928（昭和 3）年にかけて博文館から刊行された総合雑誌『太陽』の 5 カ年分 60 冊の全文を電子テキスト化し、そこにマークアップ言語である XML を用いて文書構造・文字・表記等に関する情報を付与するという、当時、近代語のコーパスとして画期的な構造を持つものであった。この『太陽コーパス』により、近代語のコーパスのあり方の方向性が示されたが、その公開から 7 年以上が経過し、その本文の電子テキスト化の方法や XML タグによって付与する情報の種類やその付与方法については、今日の研究成果や技術水準から見ると不十分な部分も多く、これをそのまま今後の本格的な近代語コーパスのモデルとするには不足の感は否めない。

† kondo@ninjal.ac.jp
†† togiso@ninjal.ac.jp
††† tsunaga@ninjal.ac.jp
†††† mtanaka@ninjal.ac.jp

[1] 『明六雑誌コーパス』は現在開発中であり、本発表で紹介する事項等は公開時までに変更される可能性がある。

そこで、「近代語コーパス」プロジェクトでは、国立国語研究所において作成済みの近代語資料の電子テキストの一つである『明六雑誌』を対象とし、『太陽コーパス』の構造を引き継ぎつつ、本プロジェクトでの最新の研究成果を盛り込んだ形でコーパス化を行い、これからの近代語コーパスのモデルとして示すことにした。それが、本発表で紹介する『明六雑誌コーパス』である。

３．『明六雑誌コーパス』の概要

『明六雑誌コーパス』は、1884（明治 7）年から 1885（明治 8）年にかけて刊行された『明六雑誌』全 43 号の全文[2]を対象とするコーパスである。『明六雑誌』は当時の洋学者によって結成された学術団体である明六社の機関誌で、社中での討論や演説を掲載した、日本の学術雑誌の先駆けの一つとされるものである。近代語の実態を知る上で重要度の高い資料であり、コーパス化の望まれるものの一つと言える。

『明六雑誌コーパス』は、『太陽コーパス』の構造を引き継ぎ、電子テキスト化した本文データに、XML を用いて文書構造・形態論・文字・表記等に関する情報を付与する構造となっている。『明六雑誌コーパス』で用いられる主な XML タグをあげたものが表 1 である。

表 1　『明六雑誌コーパス』の XML タグ（一部）

タグ名	説明
雑誌	雑誌 1 号分を表す。
記事	記事を表す。
div	雑誌タイトル等を表す。
paragraph	段落を表す。
block	記事タイトル・記事著者・小見出し等を表す。
figureBlock	図表を表す。
割書	割書された文字列を表す。
引用	当該記事以外の文献や会話・心話等の引用を表す。
sentence	文を表す。
SUW	語（短単位）を表す。
ruby	振り仮名を表す。
外字	文字集合 JIS X 0213 外の文字を表す。
包摂	拡張包摂基準により包摂した文字を表す。
注	誤植・濁点脱落を校訂したことを表す。
小書	小書きされた仮名文字を表す。
敬意欠字	皇室への敬意を示す欠字を表す。
踊字	踊字を表す。
pb	原本の改ページ位置を表す。
lb	原本の強制改行位置を表す。

４．『明六雑誌コーパス』の特徴

次に、『明六雑誌コーパス』の特徴について、特に『太陽コーパス』からの改良という

[2] 表紙、目次、識語、奥付、図表中の文字列は除く。

観点から説明する。

４．１．文字集合 JIS X 0213 の採用

『明六雑誌コーパス』では、本文の電子テキスト化において JIS X 0213:2004 に準拠した文字集合を採用した。『太陽コーパス』の開発中に JIS X 0213:2000 が制定されたが、当時はまだ実装が限られており普及が見通せなかったため、『太陽コーパス』では文字集合としての採用は見送られ、代わりに JIS X 0208:1997 が採用された（田中、2005）。しかし、その後の JIS X 0213 の普及により、BCCWJ のように JIS X 0213 に準拠した文字集合を用いるコーパスも登場した。これを受けて『明六雑誌コーパス』でも JIS X 0213:2004 を採用することにしたものである。ただし、JIS X 0213:2004 そのものを文字集合とするのではなく、JIS 包摂規準の例外にあたる漢字[3]や動作環境によっては適切に表示されない問題のある漢字[4]については使用しないこととした。

これにより、非漢字では例えば合字「㌧」（コト）や繰り返し符号「〳〵」「〳〵゛」「〲」、漢字では第 3・第 4 水準漢字の電子化が可能となった。須永・堤・高田（2011）では、『明六雑誌』の本文中の漢字延べ 137,897 字の内訳について表 2 のように報告しており、ここから、JIS X 0213 の採用により、延べ 201 字の第 3・第 4 水準漢字の電子化が新たに可能となったことがわかる。

表 2　JIS X 0213 文字集合と『明六雑誌コーパス』漢字

文字区分	延べ字数
JIS X 0213	135,797
第 1 水準漢字	117,643
第 2 水準漢字	17,953
第 3 水準漢字	118
第 4 水準漢字	83
外字	2,100
計	137,897

なお、JIS X 0213 を用いてもなお外字となる漢字延べ 2,100 字については、言語研究での実用性を鑑み、包摂規準の拡張や別字での代用を行うことでできるだけ電子化する方針をとり、最終的に外字として〓表示になるものは延べ 31 字までに減らした（須永・堤・高田、2011）。

４．２．形態論情報の付与

BCCWJ で現代語の形態素解析辞書「UniDic」を用いて形態論情報の付与がなされたように、現代語のコーパスでは、形態素解析の技術を用い、文字列を語に分割し、語彙素（見出し語に相当）や品詞といった形態論情報を言語研究への利用に適した形で高精度に付与することが可能となっている。近代語の資料では、現代語以上に語形や表記のバリエーションが豊富であることを考えると、近代語のコーパスでは、現代語のコーパス以上に形態

[3] 康熙別掲字（104 字）と UCS 互換字（10 字）。

[4] CJK 統合漢字拡張Ｂに符号位置が割り当てられる文字（302 字）。

論情報の付与の必要度が高いと言える。しかし、近代語を含めた古い時代の日本語資料について、実用的な精度で形態素解析することは『太陽コーパス』開発時には実現されておらず、『太陽コーパス』では形態論情報の付与はなされなかった。それが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）が開発され、近代語のコーパスに形態論情報を付与する環境が整備されつつある。『明六雑誌コーパス』では、その最新の技術を活かし、「近代文語 UniDic」を用いて本文を形態素解析した後、人手修正を加えたものを SUW（Short-Unit Word）要素として付与した。解析に用いた「近代文語 UniDic」は、①ゆれの少ない斉一な単位である「短単位」を解析単位とする、②表記の揺れや語形の変異にかかわらない見出し語が付与される、③和語・漢語・外来語・混種語といった語種情報が付与される、といった日本語研究に適した特長を持つ。また、「近代文語 UniDic」の構造は開発の基盤となった現代語用「UniDic」と共通するため、それぞれの UniDic で解析したコーパスを利用した通時的研究も可能となる。

『明六雑誌コーパス』の SUW 要素の例

```
<SUW orthToken="洋字" lForm="ヨウジ" lemma="洋字" pos="名詞-普通名詞-一般" form="ヨウジ" pronToken="ヨージ" wType="漢" start="100" end="120" orderID="80" section="v">洋字</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" wType="和" start="120" end="130" orderID="90" section="v">を</SUW>
<SUW orthToken="以" lForm="モツ" lemma="持つ" pos="動詞-一般" form="モツ" cType="文語四段-タ行" cForm="連用形-促音便" pronToken="モッ" wType="和" start="130" end="140" orderID="100" section="v">以</SUW>
<SUW orthToken="て" lForm="テ" lemma="て" pos="助詞-接続助詞" form="テ" pronToken="テ" wType="和" start="140" end="150" orderID="110" section="v">て</SUW>
```

SUW 要素を利用して『明六雑誌コーパス』の語数を概観すると、延べ語数約 18 万語（記号類除く）、異なり語数（語彙素レベル）約 1 万 2 千語となる。

４．３．文情報の付与

近代語のコーパスの開発では、形態論情報以外にも言語の階層的構造に適切に対応した情報の付与が望まれる。例えば、文という単位について情報の付与は、近代語のコーパスにおいても優先度の高い事項と考えられる。現代語の文字資料であれば、文の末尾は、多くの場合「。」「！」等の記号や論理改行によって明示され、それに基づきコンピュータが自動的に文認定をすることも可能であり、BCCWJ でもそのように文認定が行われている（山口・高田・北村・他、2011、pp.136-138）。しかし、まだ句読法の確立していない時期の近代語の資料では、文末を示す記号を全く用いない文章や、「、」を句点・読点の両義に用いる文章など、様々な形態があり、コンピュータによる自動的な文認定が今日に至るまで実現していない。そのため、『太陽コーパス』では文という単位に関する情報の付与を事実上諦め、代わりに「、」と「。」を手がかりとして自動的に認定された単位を s 要素として情報付けを行った。しかし、結果的に s 要素は言語研究で利用しづらい情報となってしまっている。

『太陽コーパス』の s 要素の例

```
<s>詩仙堂　仝村に在り、</s><s>石川丈山の山荘なり、</s><s>遺物少からず、</s><s>
```

堂の四壁に、</s><s>漢、</s><s>晋、</s><s>唐、</s><s>宋の詩人三十六輩の像を畫き、</s><s>（狩野尚信の筆）自から其人の詩を書きて掲げたり、</s>

『明六雑誌コーパス』ではこの問題を克服するため、人手による文認定を行い、sentence 要素として情報付けを行った。これにより、文という単位に基づいたコーパス利用が可能となっている。

『明六雑誌コーパス』の sentence 要素の例

<sentence>　其他語格の若きは後日の成功を待つべし</sentence><sentence>右聊か愚考を陳じ諸先生の可否を請ふ</sentence><sentence>敢て採用を望むにあらずと雖ども諸先生幸に電覽を賜はゞ幸甚</sentence>

４．４．コーパスの公開形式

『明六雑誌コーパス』の公開形式は次の 3 種類を予定している。

① XML ファイル
② 形態論情報タブ区切りデータ
③ 「ひまわり」版

①はコーパスの全情報（本文データと XML タグによる付加情報）を納めたデータである。②は、①から特に形態論情報（SUW 要素）を抽出し表形式に整形したデータで、語彙研究等への活用を想定し作成した。③は XML 文書全文検索システム「ひまわり」に搭載した形式で、GUI による簡便な検索が可能であり、コーパス利用の幅が広がることを期待して作成した。「ひまわり」は当初『太陽コーパス』のための検索システムとして開発されたものであるが、『明六雑誌コーパス』への適用にあたっては、本文データに対する文字列検索だけでなく、語彙素・語彙素読み・品詞・語種といった形態論情報に対する検索にも対応した形に整備した（図 1、稿末）。形式②③での公開は、『明六雑誌コーパス』に付与した形態論情報を活かした新規の試みとなる。

５．おわりに

以上、『明六雑誌コーパス』の概要および特徴について紹介した。従来の近代語のコーパスでは不十分であった点を改良し、今後の近代語コーパス構築に向けての一つのモデルを示すことができたと考える。

ただし、当然のことながら、本格的な近代語コーパス構築にあたっては、まだ克服すべき課題も多い。『明六雑誌コーパス』では学術雑誌という一媒体・一ジャンルのコーパス化のモデルを示し得たに過ぎず、新聞・小説等の他媒体・他ジャンルの資料のコーパス化の方法については、別途検討が必要である。タグ付けする情報の種類や付与方法については、BCCWJ や今後開発が予定されている古代から近世までの通時コーパスとの連続性を考慮しつつ、更に改変していく必要があろう。また、『明六雑誌コーパス』最大の特徴である形態論情報についても、「短単位」による SUW 要素だけでなく、より長い単位である「長単位（Long-Unit Word）」による LUW 要素の付与（BCCWJ では実現している）も目指していく必要がある。さらに、『明六雑誌コーパス』はそれほどデータ量が多くなかったため、文の認定等、全面的に人手に頼る作業も可能であったが、大規模なコーパス構築に際しては、そうした作業の自動化について研究を進める必要が生じるかもしれない。これらの残された課題を一つ一つ克服し、今後の近代語コーパスの開発の実現に着実につなげていきたい。

文　献

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://dl.dropbox.com/u/73297026/report/unidic-MLJ_report2009.pdf よりダウンロード可能）

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

須永哲矢、堤智昭、高田智和（2011）「明治前期雑誌の異体漢字と文字コード―『明六雑誌』を事例として―」人文科学とコンピュータシンポジウム論文集、2011:8、pp.381-388

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―、pp.1-48、博文館新社

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 年度研究成果報告書　『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)

関連 URL

「近代語コーパス」プロジェクト	http://www.ninjal.ac.jp/research/project/b/kindaigo/
近代文語 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic
ひまわり	http://www2.ninjal.ac.jp/lrc/index.php
UniDic	http://download.unidic.org

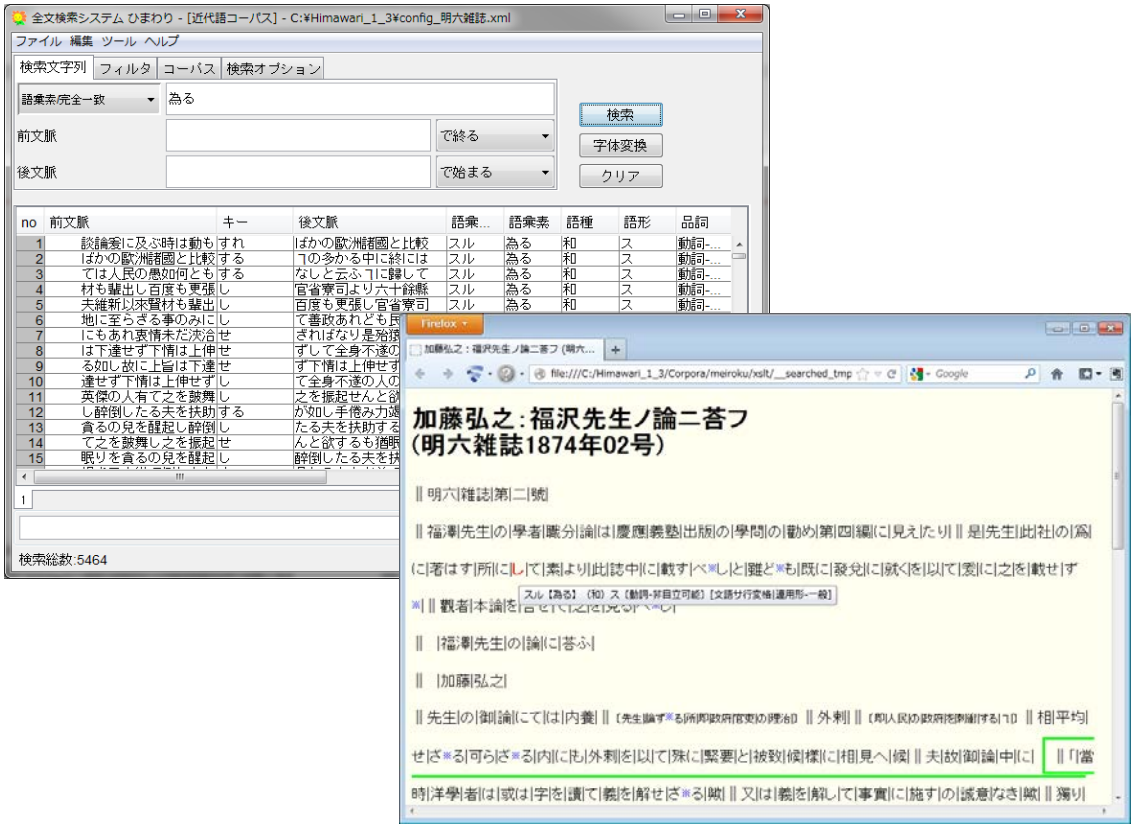


図１　「ひまわり」による『明六雑誌コーパス』の検索画面（左）および文脈確認画面（右）

書　名　第2回 コーパス日本語学ワークショップ予稿集
発行日　平成24年9月1日
発行者　国立国語研究所 言語資源研究系・コーパス開発センター
http://www.ninjal.ac.jp/organization/chart/03/
http://www.ninjal.ac.jp/organization/chart/06/
連絡先　〒190-8561　東京都立川市緑町10-2
大学共同利用機関法人 人間文化研究機構 国立国語研究所コーパス開発センター内
電　話　042-540-4300（代表）