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

著者	特定領域研究「日本語コーパス」総括班
ページ	1-552
発行年	2011-03-10
URL	http://doi.org/10.15084/00003344

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

平成23年3月14日、15日、16日

文部科学省科学研究費特定領域研究
「代表性を有する大規模日本語書き言葉コーパスの構築：
21世紀の日本語研究の基盤整備」
総括班
JC-G-10-02

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

2011年3月14日（月）／15日（火）／16日（水）

Program［プログラム］

3月14日(月)

10:30　**■開　会**

10:30～10:35　**■挨　拶**　影山 太郎（国立国語研究所長）

10:35～11:05　**■領域代表者総括報告**

「『日本語コーパス』の活動を終えるにあたって」 前川 喜久雄

11:05～11:30　**■ポスター発表内容紹介**

11:30～11:40　**休　憩**

11:40～14:00　**■ポスターセッション**

「日本語名詞の意味 ―日英語翻訳の観点から―」 鈴木 敏

「日本語の定型表現の読解処理 ―瞳孔径の計測データを用いて―」
梁 志鋭、阪上 辰也、古泉 隆、坂東 貴夫

「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―」
古泉 隆、梁 志鋭、阪上 辰也、坂東 貴夫、天野 修一、新實 葉子

「心理学実験とコーパスに基づく『上』の意味ネットワークの実証的研究」 徐 蓮

「複合辞『というと』『といえば』『といったら』の用法の異同に関する計量的考察」 小西 いずみ

「関係名詞としての空間的位置表現」 西口 純代

「外来語由来の造語成分『チック』について」 村中 淑子

「コーパスを用いて新語を調べる ―『スルー』を材料に―」 村中 淑子

「ポライトネスからみた『てくれる系』と『てもらう系』の使い分けに関する一考察」
ジュ・ヒョンジュ

「『かなしい、つらい、くるしい』の意味について」 加藤 恵梨

「日本語学習者のための語の用例記述に向けて
―辞書の品詞・用例から学ぶことができない語の情報―」 前坊 香菜子

「書きことばらしさの判断と測定」 井上 次夫

「書き言葉におけるダ体とデアル体の混用への考察」 徐 衛

「コーパスを用いた外来語サ変動詞の分析 ―『カットする』を例として―」 茂木 俊伸

「『ー中』の用法 ―BCCWJサブコーパス間の比較―」 新實 葉子

「BCCWJと誤用コーパスを利用した日本語作文支援に関する一考察」
八木 豊、鈴木 泰山、仁科 喜久子

「BCCWJモニター公開データに基づいた並立助詞『や』の分析」 川口 裕子

「『オノマトペ+する』の構文的特徴 ―『スル』の取りうる形式に焦点を当てて―」 黄 慧

「感情を表す動詞の考察」 韓 金柱

14:00～16:40　**■計画班研究活動・成果報告**

データ班「代表性を有する現代日本語書籍コーパスの構築」 山崎 誠

ツール班「書き言葉コーパスの自動アノテーションの研究」 松本 裕治

電子化辞書班「多様な目的に適した形態素解析システム用電子化辞書の開発」 伝 康晴

日本語学班「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」 田野村 忠温

日本語教育班「代表性を有する書き言葉コーパスを活用した日本語教育研究」 砂川 有里子

言語政策班「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」 田中 牧郎

辞書編集班「コーパスを利用した国語辞典編集法の研究」 荻野 綱男

言語処理班「代表性のあるコーパスを利用した日本語意味解析」 奥村 学

16:40　**■閉　会**

3月15日(火)

10:00　**■開　会**

10:00～11:40　**■公募班研究活動・成果報告**

日本語機能表現班「大規模階層辞書を用いた日本語機能表現解析体系の研究」 宇津呂 武仁

作文支援システム班

「バランス・コーパス利用による日本語作文支援システム『なつめ』の構築と評価」 仁科 喜久子

意見情報班「多様な文書ジャンルを対象とした意見分析コーパスの作成に関する研究」 関 洋平

日本語フレームネット班「BCCWJと意味フレームに基づく語彙・構文複合資源の構築」

小原 京子

11:40～11:50　**休　憩**

11:50～14:00　**■デモ・ポスターセッション**

「『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概要（5）

―サンプリングの最終結果―」

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』における評価表現の分布

―『日本語アプレイザル評価表現辞書（態度表現編）』を用いて―」 佐野 大樹、柏野 和佳子

「Yahoo！知恵袋の質問における修辞機能の分布 ―修辞ユニット分析を用いて―」

田中 弥生、佐野 大樹

「『現代日本語書き言葉均衡コーパス』向け外字処理ツール」 田島 孝治、高田 智和

「長単位に基づく媒体・カテゴリ間の品詞比率に関する分析」

冨士池 優美、小西 光、小椋 秀樹、小木曽 智信、小磯 花絵

「BCCWJに基づくオノマトペの品詞と意味についての分析」

宮内 佐夜香、小木曽 智信、小磯 花絵、小椋 秀樹

「Web版コーパス検索アプリケーション『中納言』のデモンストレーション」

中村 壮範、小木曽 智信

「階層的形態論情報を考慮した『現代日本語書き言葉均衡コーパス』の公開用XMLフォーマット」

小木曽 智信、間淵 洋子、前川 喜久雄

「汎用アノテーションツールSlate」 徳永 健伸、Dain Kaplan、飯田 龍

「BCCWJと関連ツールの相互運用」 狩野 芳伸、橋田 浩一

「拡張固有表現タグ付きコーパスの構築」 橋本 泰一

「BCCWJコアデータへの係り受け・並列構造のアノテーション」

浅原 正幸、岩立 将和、松本 裕治

「BCCWJに対する述語項構造と照応関係のアノテーション」 小町 守、飯田 龍

「BCCWJに基づく中・長単位解析ツール」 小澤 俊介、内元 清貴、伝 康晴

「UniDicを用いた音声認識用言語モデルの作成」 山田 篤

「作文コーパスからみる生徒の使用語彙」 鈴木 一史、棚橋 尚子、河内 昭浩

「学習データ間距離学習に基づく語義識別の性能分析」 佐々木 稔、新納 浩幸

「コーパス管理・検索ツール『茶器』」 松本 裕治、浅原 正幸、岩立 将和、森田 敏生

14:00～17:20　**■シンポジウム「日本語コーパスと外国語としての日本語研究」**

「海外の日本語教育から見た均衡コーパス―日本語教材の評価・比較・編集―」

曹 大峰（北京日本学研究センター）

「イタリア人向けの和伊辞典編纂におけるBCCWJの貢献」

カルヴェッティ・パオロ（カ・フォスカリ ヴェネツイア大学）

「副詞による括弧構造とその文脈における役割について」 アンドレイ・ベケシュ（筑波大学）

「基本動詞ハンドブック執筆へのBCCWJの利用

―辞書執筆用コーパスシステムNINJAL-LagoWordProfilerの開発―」

プラシャント・パルデシ（国立国語研究所）、赤瀬川 史朗（Lago言語研究所）

（16:00～16:20）**休　憩**

（16:20～17:20）**パネルディスカッション**

17:20　**■閉　会**

3月16日(水)

10:00　**■開　会**

10:00～11:40　**■計画班研究発表**

「多義語における意味の分布」 山崎 誠

「拡張モダリティタグ体系の設計とBCCWJへのアノテーション」
松吉 俊、佐尾 ちとせ、乾 健太郎、松本 裕治

「UniDic 2：設計と実装」 小木曽 智信、伝 康晴

「日本語研究とインターネット」 田野村 忠温

11:40～11:50　**休　憩**

11:50～14:00　**■デモ・ポスターセッション**

「複合名詞内の係り受けに着眼したアクセント変形予測の高精度化に関する実験的検討」
高野 克弥、峯松 信明

「テキストの多様性をとらえる分類指標の構築を目指して」
小磯 花絵、田中 弥生、小木曽 智信、近藤 明日子

「BCCWJを用いた語彙・文法情報のプロファイリングとその応用」 千葉 庄寿

「中学校・高校教科書の教科特徴語リストの作成 ―語彙指導の基礎資料として―」 近藤 明日子

「ジャンル別に見た特徴漢字 ―書籍のジャンルと広報紙の漢字―」 斎藤 達哉

「社会科での漢字学習事例検討 ―小学校6年生『憲』について―」 棚橋 尚子

「コーパスに基づく分類重要語彙リスト ―学校教育での活用に向けて―」 田中 牧郎

「外形で引く国語辞典への試み」 矢澤 真人

「同時共起クラスタリングを利用した大規模テキストからの動詞類語抽出」
竹内 孔一、高橋 秀幸、小林 大介

「分類器の確信度を用いた合議制による語義曖昧性解消の領域適応」 古宮 嘉那子、奥村 学

「共起語グラフのクラスタリングによる単語の多義性抽出」 鏑木 雄太、古宮 嘉那子、小谷 善行

「教師付き外れ値検出による新語義の発見」 新納 浩幸、佐々木 稔

「SemEval-2010日本語語義曖昧性解消タスク報告」
奥村 学、白井 清昭、古宮 嘉那子、横野 光

「大規模階層辞書を用いた日本語機能表現解析体系の研究」
宇津呂 武仁、鈴木 敬文、島内 蘭、阿部 佑亮、松吉 俊、土屋 雅稔

「BCCWJを利用した日本語作文支援システム『なつめ』の評価」
阿辺川 武、ホドシチェク・ボル、仁科 喜久子

「日本語フレームネットにおけるBCCWJへの意味アノテーション」
小原 京子、加藤 淳也、斎藤 博昭

「FrameSQLで見る日本語フレームネット 」 佐藤 弘明

「BCCWJを用いた語彙・構文彙の分析 ―所謂引用助詞『と』が標識する構文の場合―」
藤井 聖子

14:00～15:15　**■計画班研究発表**

「日本語教育における初級シラバスの再評価
―BCCWJにみられた『出現形の偏り』を手がかりに―」小林 ミナ

「BCCWJ複合辞辞書の仕様・開発・評価」
近藤 泰弘、坂野 収、多田 知子、岡田 純子、山元 啓史

「複数の観点から見た用例クラスタリングに基づく新語義の発見」
白井 清昭、中西 隆一郎、中村 誠

15:15～15:20　**■領域代表者挨拶**　前川 喜久雄

15:20　**■閉　会**

Contents [目次]

公募班研究活動・成果報告

デモ・ポスターセッション

シンポジウム「日本語コーパスと外国語としての日本語研究」

領域代表者総括報告

3月14日（月）　10:35〜11:05

「日本語コーパス」の活動を終えるにあたって

▶前川 喜久雄

「日本語コーパス」の活動を終えるにあたって

前川喜久雄（領域代表者：国立国語研究所言語資源研究系）[†]

At the End of the Priority-Area Program "*Japanese Corpus*"

Kikuo Maekawa(Program Supervisor. National Institute for Japanese Language and Linguistics)

１．特定領域研究「日本語コーパス」

文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（略称「日本語コーパス」)は平成 18 年 7 月に採択が内定し、同年 9 月から活動を開始した。その後、現在まで 5 年間にわたる活動を継続し、本平成 23 年 3 月末をもって 5 年間の活動を終了する。

「日本語コーパス」の主要な目標は、今後の日本語学の研究インフラとなる『現代日本語書き言葉均衡コーパス』(BCCWJ)を構築すること(研究項目 A01)と、コーパスを利用した日本語研究の可能性を開拓すること（研究項目 B01）のふたつであった。コーパスの構築は、関係者の努力によってほぼ予定どおりに進めることができた。

コーパス日本語学の可能性開拓は、もとよりコーパス構築のように着実な進展がみこめる分野ではないものの、やはり相当の進展をみることができた。過去３年間に日本語学会、英語コーパス学会、人工知能学会、日本言語学会があいついで、それぞれの機関紙や大会でコーパス関係の企画を実施したのは、直接間接に本領域の影響によるものと考えられる。

その他国内外の多くの学会、研究会やシンポジウムでコーパスに関わるテーマがとりあげられた。数えてみると、私が個人で発表したものだけで 24 件（学術誌論文 7 件、シンポジウム 7 件、国際会議招待講演 5 件、商業誌等解説論文 3 件、単行本分担執筆 2 件）ある。これは、専門的な学会発表、本領域主催の公開WSでの発表、政府審議会での参考人意見聴取、話し言葉コーパス（CSJ）に関わる講演などを除外した件数である。いま正確には把握できていないが、各研究班長による講演も多数に及んでいるから、本領域の活動は文科系の特定領域研究としては、従来になく活発だったと言ってよいだろう[1]。

２．『現代日本語書き言葉均衡コーパス』

BCCWJ の構築にあたっては本領域データ班と国立国語研究所研究開発部門言語資源グループが共同して作業にあたることが本領域の申請時から計画されていた。1 億語からなる BCCWJ のうち本領域では書籍のサンプル約 5000 万語の作成を分担し、それ以外のサンプルは国立国語研究所が分担する体制である。

国立国語研究所は 2009 年 10 月に独立行政法人としては廃止され、その業務の大部分が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に移管されたが、幸い、BCCWJ の開発は、移管後も本領域と新国語研に設置されたコーパス開発センターの共同事業として維持されることとなった。

詳しくはデータ班からの報告にゆずるが、BCCWJ のサイズは目標の 1 億語を突破してお

† kikuo@ninjal.ac.jp

[1] マスコミでの報道は 15 件であった。

り、そのうち約 6500 万語は従来利用することが困難だった書籍サンプルである。また、全サンプルに施された形態素解析の精度も目標とした 98%を上回っている。結論として領域申請書に記載したコーパス構築関係の目標は完全に達成されたと考えている。

BCCWJ のサンプルは、「KOTONOHA『現代日本語書き言葉均衡コーパス』検索デモンストレーション」サイト(http://www.kotonoha.gr.jp/demo/)において公開され、全文検索が可能となっているが、今年（2011 年）の夏までには形態素解析済データのオンライン検索環境（後述する「中納言」相当の機能を提供）と DVD によるデータ全体の公開を実施する予定である。「中納言」については、本ワークショップの期間中にも検索デモを実施するので、是非ご覧いただきたい。

３．BCCWJ の検索

本節と次節では筆者自身の興味に沿って BCCWJ の簡単な評価を試みる。まず本節では BCCWJ 以前に試みた検索の結果と BCCWJ の結果とを比較検討する。筆者は 2007 年 3 月に開催された本領域の第 1 回公開ワークショップの講演の後半で、いくつかの言語現象について、当時利用可能であったデータ（毎日新聞記事 2003 年分、国会会議録、青空文庫、『日本語話し言葉コーパス』など）を文字列検索した結果を示している。この分析はその後の拡張部分もくわえて前川(2007)として刊行されている。これを前稿と呼ぶことにする。

以下本節では前稿の分析結果と BCCWJ の分析結果との比較を試みる。BCCWJ の検索には、昨年来特定領域研究のメンバーに公開されている「中納言」を利用する。これはウェブベースの検索 GUI で、BCCWJ の短単位形態素解析済データ約 7000 万語分が検索可能である（中村・小木曽 2011）。

３．１　風景と光景

「風景」は名詞や接頭辞に続いて複合語を形成しやすいが、「光景」にはこの傾向が認められない。表 1 は前稿に掲載した調査結果（2003 年分の毎日新聞データの検索結果）と BCCWJ の検索結果を対照させている。表中の「総数」は文字列「風景」「光景」の BCCWJ 中での総出現数、「後部要素」は「原風景」「心象風景」のように複合語の後部要素としての出現数、「異なり語数」は「風景」ないし「光景」を後部要素とする複合語の異なり語数である。両コーパスから得られる結果は同一と言ってよい。

表 1.「風景」と「光景」の複合語になりやすさ

語	コーパス	総数	後部要素	（%後部要素）	異なり語数
風景	新聞 2003	954	259	(27.1)	107
	BCCWJ	2495	489	(19.6)	196
光景	新聞 2003	514	4	(0.8)	4
	BCCWJ	1912	21	(1.1)	13

３．２　動詞＋です

「昨晩、あるいは昨夜おそく、このあたりは雨が降ったです。」「収集ができるですか？」のように、動詞に「です」が後続して文末を形成する現象である。非文法的とみなされた

り、規範性が低いとみなされるが、実際に存在する。

前稿では主に青空文庫からの例を 13 件掲載した。中納言を利用して、動詞終止形直後に助動詞「です」の終止形が生じているケース（現在形）と、動詞の連用形の直後が助動詞「た」の終止形で、その直後に助動詞「です」の終止形が生じているケース（過去形）をともに検索したところ、現在形が 153 例、過去形が 154 例見つかった。「死んだですって？」「（健康の秘訣は）規則正しく寝る、食べるです。」のようなメタ言語的表現および形態素の誤解析に起因するノイズは除外してある。

検索結果の一部を以下に示す。今回の検索での新しい発見は、1)~3)のように文芸作品で役割語ないし方言として用いられている例が確認されたことであるが、4)以下のように、そのようには解釈できない例も多い。

1) 理解する方がよほど楽だ。私はつくづくそう思うですよ。（村上春樹「世界の終りとハードボイルド・ワンダーランド」）
2) 風呂の水だばわけなく汲むっす。（原田康子「海霧」）
3) 自殺でもしてくれたら、どげん清々するかと思うですよ。（佐木隆三「復讐するは我にあり」）
4) ベテランが予想以上にがんばってくれているです。（Yahoo!知恵袋）
5) ちょっとした事ですぐに起きてしまうです。（Yahoo!知恵袋）
6) 僕は逃げるのが速いから、なんとか助かったですよ。（林光「職人技を見て歩く」）
7) そのときボクは、生活すごく困ったです。（笹倉明「東京難民事件」）
8) 泣きネタはちょっと飽きたです。（Yahoo!知恵袋）
9) それでトイレがなかったという誤った俗説が広まったです。（Yahoo!知恵袋）

３．３　「起きる」「起こる」「生じる」

前稿では新聞記事（毎日新聞 2003 年分）と国会会議録を対象として、主格補語「問題が」、「事件が」と、その直後に位置する述語動詞「起きる」、「起こる」、「生じる」（ないし「生ずる」）の共起関係を分析した。表 2 にはそのうち、「問題が」の結果だけを示す。表の上部 2 行が前稿の結果であり、第 3 行は、BCCWJ に含まれる書籍サンプル（出版書籍、図書館書籍、非母集団ベストセラーの３サブコーパスの和）の集計結果である。

表 2.「問題が」と「起きる」「起こる」「生じる」の共起関係

コーパス	起きる　(%)	起こる　(%)	生じる　(%)
新聞記事(前稿)	84　(52.8)	12　(7.5)	63　(39.6)
国会会議録(前稿)	85　(20.7)	143　(34.9)	182　(44.4)
BCCWJ 書籍	113　(25.7)	114　(25.9)	213　(48.4)

表 2 を検討すると、国会会議録と BCCWJ 書籍の分布パタンが類似しており、新聞記事は独自の分布を示しているように見える。しかし、この解釈には問題がある。国会会議録は 60 年以上、BCCWJ の書籍サンプルも 30 年以上の時間幅にわたるサンプルであるのに対して、新聞記事は 2003 年という近年の日本語からのサンプルであるため、表 2 における差はジャンルの相違に起因するものではなく、日本語の時間変化の反映である可能性が

あるからである。

この可能性を検討するために、BCCWJ サンプルの書誌情報の一部として提供されている著者の生年代情報を利用した分析を実施した。結果を図 1 に示す。縦軸は各動詞の相対頻度（百分率）である。横軸はサンプルの著者の生年を 4 つのグループに分類して古い順に配置している。例えば「1930~1950」は著者の生年が 1930 年代から 1950 年代のいずれかに属することを意味しており、1930 年代には 1930 年から 1939 年までに生まれた著者が所属する。二人以上の著者によるサンプルと著者の生年代が不明のサンプルは集計から除外した。

図 1 は話者の生年代とともに「起こる」の生起率が単調に低下し、その一方「生じる」の生起率が単調に増大したことを示している。その結果、もっとも若年層の著者では、「生じる」の頻度が顕著に高く、「起こる」の頻度が顕著に低いパタンが形成されている。

この分析結果を念頭におくと、表 2 の新聞記事における「起こる」の相対頻度の低さは、新聞記事というジャンルの特徴ではなく、現代日本語の一般傾向を反映するものではないかと考えられる。

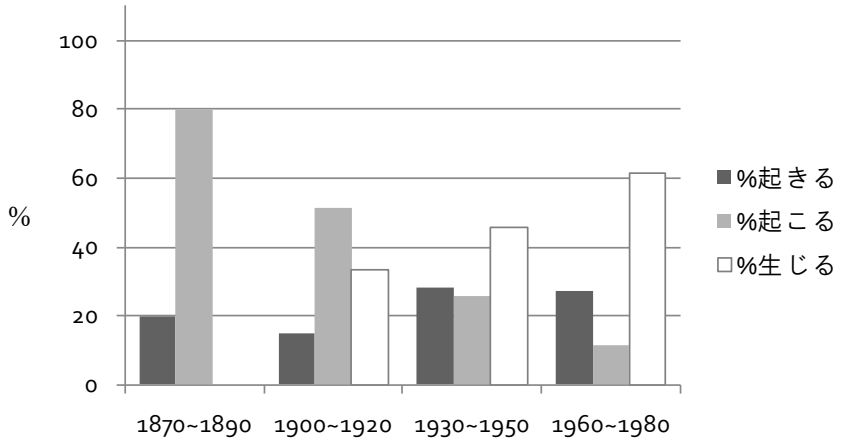


図 1. BCCWJ 書籍サンプルの著者生年代と「起きる」「起こる」「生じる」の相対頻度（主格補語は「問題が」）

４．形態素情報の活用

中納言で BCCWJ を利用すると、全文検索だけでなく、形態素情報を利用することができる。前節の検索でも形態素情報を利用しているが、本節ではもう少し積極的に形態素情報を活用した検索例を 2 件紹介する。

４．１　連体修飾における受動態

日本語の関係節では、「天井から吊るした電灯」のような能動態と「天井から吊るされた電灯」のような受動態がともに可能である。Tsunoda (2008)は、このうち受動態の占める割合が近年増加していると指摘し、独自の調査結果を示している。しかしながらこの調査は調査対象者が 8 名と限られており、また学会予稿集の分析であるためレジスター的にも非常に限定された調査にとどまっている。角田の指摘する傾向が実在するかどうかを BCCWJ を使って検証してみた。

中納言に格納されている書籍データを対象として以下の検索を行った。まず受動態過去形のデータとして、任意の動詞に助動詞「れる」ないし「られる」が後続し、その直後に助動詞の「た」の連体形が続き、その直後に任意の名詞が位置し、その直後に格助詞「が」が位置する 5-gram を検索したところ 4,002 例を得た。このデータを「V 受動た N が」と呼ぶことにする。

このサンプルの頻度を相対化するために非受動態過去形のデータも検索する必要がある。任意の動詞の直後に助動詞「た」の連体形が位置し、その直後に任意の名詞が位置し、その直後に格助詞「が」が位置している 4-gram を検索したところ 46,148 例を得た。これを「V た N が」と呼ぶ。

著者の生年代ごとに「V 受動た N が」の数値を「V 受動た N が」と「V た N が」の合計で除し、さらに 100 倍して百分率とした数値を図 2 に示す。縦軸が受動態の相対頻度の百分率、横軸が著者の生年代である。生年代の集計方法は図 1 と同一である。

図 2 には名詞に後続する格助詞が「を」の場合の結果も示した。両者は著者が若くなるほど上昇しており、角田の指摘する現象の実在性を示唆している。

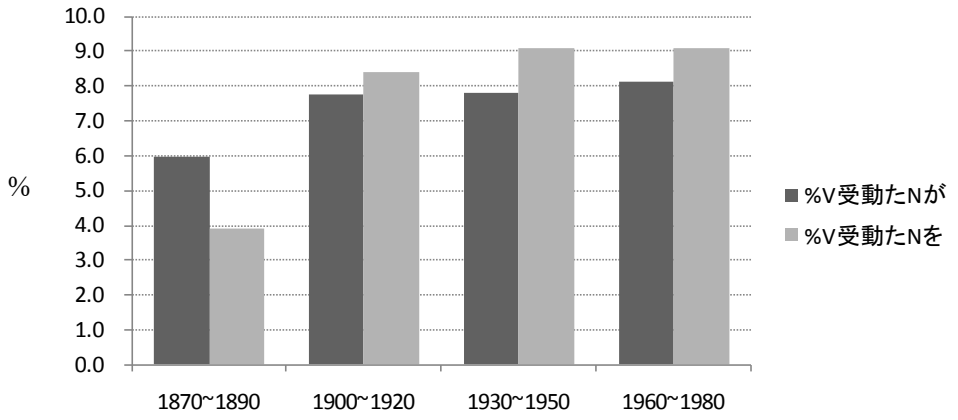


図 2. 連体修飾句における受動態の比率（過去形）

図 3 に示したのは、図 2 と同様に集計した非過去形（現在形）の結果である。任意の動詞に助動詞「れる」ないし「られる」が後続し、その直後に任意の名詞が位置し、最後に格助詞「が」ないし「を」が位置する 4-gram（「V 受動 N が」5,180 件、「V 受動 N を」2,826 件）の頻度を、非受動態サンプル、すなわち任意の動詞連体形の直後に任意の名詞が位置し、その直後に格助詞「が」ないし「を」が位置している 3-gram（「V 連体 N が」136,572 件、「V 連体 N を」82,251 件）の頻度を利用して相対化し百分率で示している。図 2 に比べると受動態の比率は低いが、やはり著者が若くなるほど受動態の比率が増加する傾向をみてとれる。

なお BCCWJ の形態素情報では、助動詞「れる」「られる」の 4 種の用法（受身、可能、自発、尊敬）を区別することができない。そのため、本節の分析では個々の用例が受動の意味で用いられているかどうかを人手で確認している。上に示した検索件数はこのような確認作業を施す以前の件数であり、例えば「V 受動た N が」の場合、4,002 件中 68 件を尊敬もしくは可能の用例と判定して除外した。この判定作業、特に受身と可能の判定にはある程度のゆれが生じることを避けられないので、本節の検索結果を読者が追試しても完全

に同一の結果が得られるとは限らない。しかし除外される用例が比較的少数にとどまっていることから、図 2,3 の全体的な傾向には影響はないものと考える。

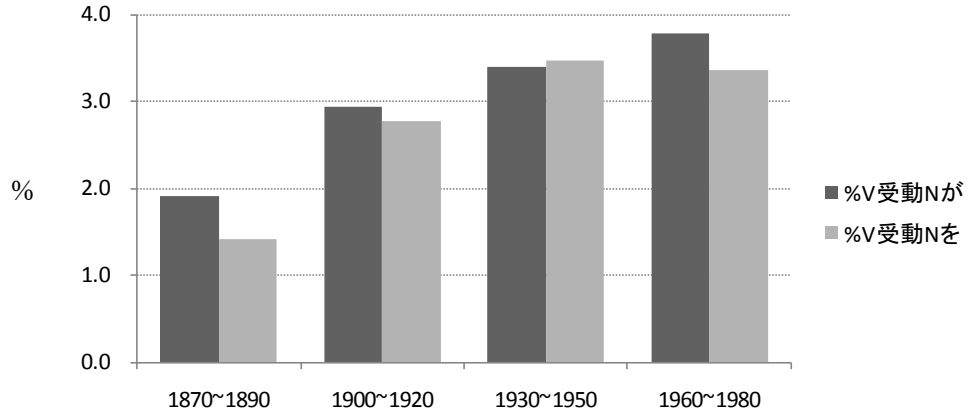


図 3．連体修飾句における受動態の比率（非過去形）

４．２　形容詞＋です

最後に述語形容詞に助動詞「です」の終止形が後続して文末を形成する現象を検索してみる。3.2 でとりあげた「動詞＋です」ほどではないが、これも規範性に欠けると意識されることの多い文末表現である。

中納言で任意の形容詞終止形に助動詞「です」の終止形が後続し、その直後に句点「。」が出現して文末を形成しているサンプルを検索したところ 4,615 例が得られた。またこの頻度を相対化するために形容詞終止形に直接句点が後続している文末を検索したところ 86,022 例が得られた。

これらの検索結果には異なり語で 423 語の形容詞が含まれていた。これらのすべてについて形容詞単独で文末を形成した場合（「A 文末」）の頻度と助動詞「です」を伴って文末を形成した場合（A です文末）の頻度を集計し、「A です文末」の相対頻度の百分率を、「A です文末」頻度÷（「A 文末」頻度＋「A です文末」頻度）×100 で計算した。

「A 文末」用例数と「A です文末」用例数の合計が 30 以上となるものだけに限って、上位 10 語と下位 10 語を示すと表 3, 4 のとおりである。いずれも第 5 列が「A です文末」の百分率である。

表 3 と表 4 の語彙項目には意味上の相違が認められる。「A です文末」を形成しやすい語彙項目（表 3）には話者の主観や心的状態を表現するもの―感情形容詞ないし情意形容詞―が多く含まれているのに対して、「A です文末」になりにくい語彙項目（表 4）には、その種の語が含まれていない。この差異と文末の形式の関係をもう少し詳しく検討してみる。

上述の検索結果には異なりで 423 語の形容詞が出現していた。それらの語を Type 1：専ら話者の心的状態についての主観的判断を表し、主格補語に「私が」「僕が」などを自然に用いることができるもの（「嬉しい」「寂しい」「悲しい」「悔しい」等）、Type 2：話者の主観的判断を表すが、主格補語に「私が」「僕が」をとるのは不自然で、「彼が」「それが」等は自然であるもの（「凄い」「きつい」「ぎこちない」「偉い」など）、Type 3：話者の主観性にはほとんど関係がなく、専ら対象の客観的属性を表す語（「丸い」「低い」「冷たい」「固い」等）の 3 タイプに分類した。Type1~3 は寺村(1982)が「感情形容詞」「感情的品定め」「性

情規定」と呼んでいるものに概略該当していると思われる。

Type1~3 の分類は筆者が単独で実施したが、その際、「不味い」(「蕎麦がまずい」～「状況がまずい」)「可笑しい」(「涙がでるほど可笑しい」～「考え方がおかしい」)のように多義性が認められる語彙項目では、人手で用例をいずれかひとつの意味(上の例であれば、「状況がまずい」「考え方がおかしい」。いずれも多数派)に絞った。図 4 は形容詞のタイプごとに「A です文末」の生起率を比較した結果である。形容詞の意味と「A です文末」の相対頻度には明らかな相関が存在している。

表 3.「A です文末」相対頻度の上位 10 語

語彙項目	(読み)	A文末	Aです文末	%Aです文末
羨ましい	(ウラヤマシイ)	42	41	49.4
しんどい	(シンドイ)	17	13	43.3
嬉しい	(ウレシイ)	313	192	38.0
美味しい	(オイシイ)	216	123	36.3
辛い	(ツライ)	151	86	36.3
怖い	(コワイ)	209	111	34.7
寂しい	(サビシイ)	65	31	32.3
きつい	(キツイ)	54	24	30.8
悔しい	(クヤシイ)	44	18	29.0
有り難い	(アリガタイ)	145	55	27.5

表 4.「A です文末」相対頻度の下位 10 語

語彙項目	(読み)	A文末	Aです文末	%Aです文末
無い	(ナイ)	49702	902	1.8
疑わしい	(ウタガワシイ)	58	1	1.7
明るい	(アカルイ)	114	2	1.7
濃い	(コイ)	69	1	1.4
鋭い	(スルドイ)	71	1	1.4
乏しい	(トボシイ)	74	1	1.3
新しい	(アタラシイ)	89	1	1.1
美しい	(ウツクシイ)	307	3	1.0
詳しい	(クワシイ)	116	1	0.9
等しい	(ヒトシイ)	167	1	0.6

ところで図 4 は、わずかではあるが Type3 にも実際に「A です文末」が生じていることを示している。この例外は何故生じるのだろうか。

すぐに思いつくのは、「A です文末」が自然とみなされるようなレジスターが存在するのではないかという仮説である。これを検討するために、BCCWJ を構成する各種サブコーパスをレジスターとみなすことにして、「A です文末」の生起率を調査してみた。結果は表 5 に示すとおりであり、Yahoo!知恵袋サブコーパスにおける生起率が圧倒的に高い。またデータを示すことは省略するが、個々の Type3 形容詞について例外のうちどれだけが知恵袋に生じているかを検討すると、平均して例外の 90%前後が Yahoo!知恵袋のサンプルであった。

暫定的な結論として、「A です文末」の生起には質的に異なるふたつの要因、すなわち個々の形容詞の意味的特性と形容詞が用いられるレジスターが関与しており、後者は前者よりも強い影響をおよぼしていると考えられる。

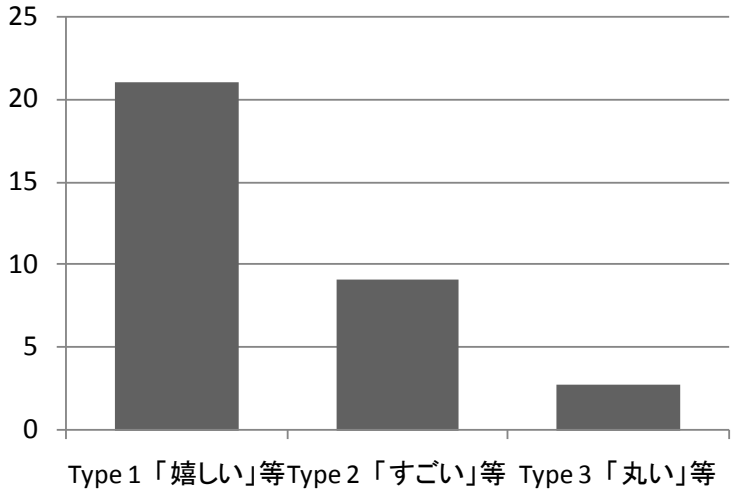


図 4．形容詞の主観性と「A です文末」の生起率（%）

表 5．レジスターによる「A です文末」生起率の変動

Subcorpus	A 文末	A です文末	%A です文末
図書館書籍	42107	281	0.7
ベストセラー	5799	37	0.6
Yahoo!知恵袋	1375	3912	74.0
白書	1712	0	0.0
出版書籍	31668	321	1.0
出版雑誌	2763	62	2.2
出版新聞	598	2	0.3

４．３　まとめ

BCCWJ に付与された形態素情報は、用例検索の精度と効率を単純な文字列検索に比べて飛躍的に向上させる。4 節に示したふたつの分析結果を得るには、それぞれ数日から 1 週間程度の時間がかかったが、その大部分は検索結果の人手チェックに費やした時間であり、検索自体は数時間で終了している。

今回とりあげたのは限られた言語現象であり、また分析自体も試行段階にとどまるものだが、総じて BCCWJ は現代日本語の語彙、文法の研究に大いに有益であるとの手応えを感じている。以上で BCCWJ の評価を終え、残る紙幅を今後の展望にあてる。

５．日本語コーパスの今後

特定領域研究はこれで活動を終えるが、コーパスを用いた日本語研究の方法を進化させ確立するのは今後に残された課題である。筆者らの研究グループは、このためにいくつかの活動を予定している。

５．１　研究成果の出版

第一に、特定領域の活動で得られた知見を学界に還元することを目的として、いわゆる講座ものの出版を計画している。幸い某書店の協力が得られることになったので、2012 年中の出版開始をめざして編集作業に着手したところである。

５．２　共同研究プロジェクト

第二に、大学共同利用機関法人となった国立国語研究所言語資源研究系の共同研究プロジェクトとして本年（平成 22 年）度から「コーパス日本語学の創生」を実施している。このプロジェクトの目標はコーパスを用いた日本語研究の促進にあり、本特定領域研究の研究項目 B01 を実質的に継承するものである。

今年度は「音声・対話グループ」と「語彙・表記・文法グループ」に分かれて共同研究会を開催しているが、来年度からは後述する「通時コーパス」プロジェクトとも協力して、コーパス日本語学に関する一種の学会機能を提供したいと考えている。一昨年から公開ワークショップに導入した一般公募によるサテライトセッションは、来年度以降は「コーパス日本語学の創生」が主催する形で継続する予定である（実は今回の公開ワークショップも、この共同研究と共催の形式をとっている）

５．３　今後のコーパス開発計画

最後に BCCWJ 公開後の大規模コーパス開発計画に触れることにする。国立国語研究所言語資源研究系では、基幹的共同研究プロジェクトのひとつとして「通時コーパスの設計」（プロジェクトリーダーは青山学院大学の近藤泰弘教授）を運営している。このプロジェクトは『日本語の史的研究に用いることができる本格的な「通時コーパス」を構築する準備段階として，コーパスの設計にかかわる諸問題について研究する』ものであるが（文言は国語研 HP より引用）、最終的には 10 年程度の時間をかけて日本語の歴史コーパスを構築する予定である。

またこれとは別に、小規模ではあるが、「近代語コーパス設計のための文献言語研究」という共同研究が田中牧郎准教授をリーダーとして運営されている。こちらは『太陽コーパス』の系譜につながるコーパスの設計を目指した研究である。私も研究会に出席したことがあるが、活発な議論が印象的であった。

以上はいずれも過去の日本語を対象としたコーパスであったが、現代語のコーパスも BCCWJ があれば、それで事足りるわけではない。単純に語数だけを問題にしても、現在 BCCWJ 全体での異なり語数（短単位）は約 22 万語であるが、そのうち約 7.5 万語は頻度が 1 である。100 以上の用例が提供できる語は約 2.5 万語にすぎず、50 例以上でも約 3.7 万語にとどまる。辞書の用例検索にはものたりない数字と言わざるをえない。

さらに単独では多数の用例が得られる語でもふたつの語の組合せになると用例が激減することがある。副助詞の「ほど」と「すら」は前者が 41,584 回、後者が 3,122 回出現しているが、「ほどすら」は BCCWJ には 1 回も出現しない。「猫の額ほどすらもない庭」のような用例が存在することが明らかであるにもかかわらずである。

このような問題を解決するための単純な解決策はコーパスの規模を飛躍的に拡大することである。前川(2007)では、2030 年頃には 100 億語規模のコーパスが出現するという予想を述べたが、BCCWJ を完成させたいま、いよいよ、このサイズの現代語コーパスの構築が現

実味を帯びた課題として感じられるようになってきた。BCCWJ を大規模コーパスと呼ぶならば、超大規模コーパスの構築にむけた準備をはじめる時期が到来しつつあるように思われる。

WWW 上のテキストデータを対象とすれば、100 億どころか 1000 億語のコーパスでも構築可能であるが、データの重複など、解決すべき問題もまた少なくない（田野村 2011 参照）。本特定領域の財産である文理融合の研究ネットワークを活用して問題の解決に努め、また BCCWJ の構築で得た知見も活かしつつ、世界に先駆けて超大規模コーパスを開発できればと考えている。

謝辞：本領域の最終公開 WS にあたり、本領域の班長、分担者、協力者の方々、ならびに本領域の活動にご理解、ご支援をいただいた領域外の方々、特に外部評価委員会の先生方に心より御礼申し上げます。

参考文献

Tsunoda, Tasaku (2008). “Predicting a future change: Relative clauses of Japanese.” In E. Verhoeven et al. (eds.) *Studies on Grammaticalization*. pp. 209-216.Berlin: Mouton de Gruyter.

田野村忠温(2011).「日本語研究とインターネット」本予稿集収載.

寺村秀夫(1982). 『日本語の意味とシンタクス I 』くろしお出版.

中村壮範・小木曽智信(2011).「Web 版コーパス検索アプリケーション「中納言」の公開」言語処理学会第 17 回年次大会発表論文.

前川喜久雄 (2007).「コーパス日本語学の可能性―大規模均衡コーパスがもたらすもの―」日本語科学, 22, pp.13-28.

ポスターセッション

3月14日（月） 11:40〜14:00

日本語名詞の意味 ―日英語翻訳の観点から―

▶鈴木 敏

日本語の定型表現の読解処理 ―瞳孔径の計測データを用いて―

▶梁 志鋭、阪上 辰也、古泉 隆、坂東 貴夫

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―

▶古泉 隆、梁 志鋭、阪上 辰也、坂東 貴夫、天野 修一、新實 葉子

心理学実験とコーパスに基づく「上」の意味ネットワークの実証的研究

▶徐 蓮

複合辞「というと」「といえば」「といったら」の用法の異同に関する計量的考察

▶小西 いずみ

関係名詞としての空間的位置表現

▶西口 純代

外来語由来の造語成分「チック」について

▶村中 淑子

コーパスを用いて新語を調べる ―「スルー」を材料に―

▶村中 淑子

ポライトネスからみた「てくれる系」と「てもらう系」の使い分けに関する一考察

▶ジュ・ヒョンジュ

「かなしい、つらい、くるしい」の意味について

▶加藤 恵梨

日本語学習者のための語の用例記述に向けて ―辞書の品詞・用例から学ぶことができない語の情報―

▶前坊 香菜子

書きことばらしさの判断と測定

▶井上 次夫

書き言葉におけるダ体とデアル体の混用への考察

▶徐 衛

コーパスを用いた外来語サ変動詞の分析 ―「カットする」を例として―

▶茂木 俊伸

「ー中」の用法 ―BCCWJサブコーパス間の比較―

▶新實 葉子

BCCWJと誤用コーパスを利用した日本語作文支援に関する一考察

▶八木 豊、鈴木 泰山、仁科 喜久子

BCCWJモニター公開データに基づいた並立助詞「や」の分析

▶川口 裕子

「オノマトペ＋する」の構文的特徴 ―「スル」の取りうる形式に焦点を当てて―

▶黄 慧

感情を表す動詞の考察

▶韓 金柱

日本語名詞の意味　－日英語翻訳の観点から－

鈴木敏（筑波大学院人文社会科学研究科）

The Meaning of Japanese Nouns: From a Point of View of J/E Translation

Satoshi Suzuki (Humanities and Social Sciences, University of Tsukuba)

0.　はじめに

(1)　a.　*The Bridges of Madison County*　　(R. J. Waller 著 1992)
　　b.　『マディソン郡の橋』　　(訳: 村松　潔 文芸春秋 1993)

日本語は一般的な複数形態素[1]と冠詞を持たない言語なので、名詞（句）の単数・複数、定・不定表現に裸名詞が使用される。英語 (1a) の 'The Bridges' が定と複数を統語的に明示するのに対し、日本語 (1b) の「橋」は単数・複数、定・不定に関して曖昧である。この論文の目的は、日本語可算名詞と英語可算名詞を比較観察し日本語名詞の持つ基本的意味を考察することである。敢えて可算名詞としたのは、質量名詞の振舞いは、その不可算性において日本語、英語に限らずあらゆる言語で共通点があるように見えるためである。日本語のように冠詞や複数形態素をもたない言語の名詞が持つ意味は、統語的な分析だけでは不十分であることを示し、日本語名詞の持つ意味について日英語翻訳の視点から考察を加える。

1.　日本語可算名詞

1.1　定・不定と単数・複数

(1b) の「橋」は一般的に個体の橋と解釈されるが、数については曖昧である。

(2)　日本語の名詞は裸で現れるときは、数に関して中立であり、単数にも複数にも解釈される…例えば、日本語の「がくせい」は、英語では "a student/students/the student/the students" に相当する…　　(水口 2004: 61)

水口が述べるように日本語の裸名詞が単数にも複数にも解釈されることは事実である。しかし、(2) の the student/the students については問題がある。

(1b) の裸名詞「橋」は (1a) の the bridges に対応し、英語の定冠詞と複数形態素の両方を表現しているようにみえる。しかし、実際には the bridges の the を十分に訳出していない。例文 (3) を見ると、それが明らかである。

[1]「たち」「ら」「ども」などの複数形態素は、英語の '-s' のように総ての可算名詞に接辞することができない。また「太郎たち」のように、太郎及びそれ以外の人の集合を表わすことから、一般的な複数形態素とは性質が異なることは広く知られている。

(3) a. The blond girl who is wearing blue jeans is my sister.
b. The blond girl wearing blue jeans is my sister
c. The blond girl in blue jeans is my sister.
d. The blond girl is my sister.
e. The girl is my sister. (Quirk et al. 1985: 245 下線は筆者)

(3a-c) の下線部は *ブルージーンズをはいた金髪の少女*、
(3d) 下線部は *金髪の少女*、と訳して問題はない。
(3e) 下線部を単に*少女*、とすると不自然である。 訳例：?*少女は私の妹だ*

(1), (3) の例が示すように the を訳出するのは名詞ではなく文脈である。読者に特定の指示物を想起させる十分な情報（文脈）がなければ the を訳出したことにはならない。(3a-c) は修飾語句によって指示物が特定されているように読者が感じるのである。

また、日本語には無生物に接辞する複数形態素が無いので、数については曖昧である。日本語可算名詞が、単数・複数を明示するために ＜数詞＋類別詞＋名詞＞形式（e.g.一人の妹、数人の妹）を取らないことは一般的（無標）である。(1b) の「橋」が単数なのか複数なのかは読者の推量にまかされている。

2. 日本語と英語の総称指示

2.1 英語総称指示と冠詞

Radden (2009) によれば、英語の総称指示（タイプ、種、クラス [2]）に使用される可算名詞の形式は、個体を表わす表現と同じ形式を用い [3]、さらに“ 総称指示を示すためだけの形式をもつ言語は無いようだ ”(ibid.: 200) とも述べている。英語の総称指示には一般的に 4 つの形式が使用される。不定単数・不定複数・定単数・定複数の 4 通りである。(Radden (2007, 2009), Krifka (1995a)) 。

(4) a. *A lion* has a bushy tail.
b. *Hedgehogs* are shy creatures.
c. *The lion* is a carnivore.
d. *The Italians* love pasta. (Radden 2009)

Radden は、総称指示に使用される 4 つの名詞句をそれぞれの形式に基づいて(4a) Representative generic 代表的総称 (4b) Proportional generic 比例的総称 (4c) Kind generic タイプ総称 (4d) Delimited generic 範囲指定総称 に分類する。それぞれの名詞句によって表現される総称指示タイプは、その形式に従って意味が異なる。

[2] 名詞が総称指示に使用される場合、その対象はタイプ、種、クラス等の名称で呼ばれる。この論文では Radden に従ってタイプを使用するが、種やクラスと同義の解釈で使用する。
[3] 英語質量名詞のタイプ表現は一般的に、 無冠詞+単数形 e.g. Oil floats. の一形式のみである (Radden 2009: 202)

2.2　日本語総称指示

一方、(5) のように日本語名詞の総称指示形式は「裸名詞」[4] 型ひとつである。

(5)　a.　ライオンは尻尾がふさふさしている
　　b.　ハリネズミは内気な動物だ
　　c.　ライオンは肉食動物だ
　　d.　イタリア人はパスタが好きだ

(6)　a.　（オスの）ライオンは尻尾がふさふさしている
　　b.　（ほとんどの）ハリネズミは内気な動物だ（中には例外もいる）
　　c.　（すべての）ライオンは（例外なく）肉食動物だ
　　d.　（イタリアに住む）イタリア人は（例外なく）パスタが好きだ

Radden (2009) によれば (4a - d) それぞれの主語名詞句の形式によって生成される英語総称文が表わす意味は (6a - d) である。(6) の例文で暗示されている (6a)（オスの）(6b)（ほとんどの）（中には例外もいる）(6c)（すべての）（例外なく）(6d)（イタリアに住む）（例外なく）は名詞句の形式から生成される。[5]

日本語総称文 (5) も、英語総称文と同様 (6) の括弧内に示した暗示を含む解釈が可能である。ただし、日本語総称文で暗示される情報内容は名詞句ではなく、文脈及び読者の世界知識に依存する。ここで一つ注意を向けるべき問題がある。日本語訳 (5a-c) に関しては、(6a-c) に含まれるカッコ内の暗示を文脈から読み取ることが可能だが、(5d) から(6d) を読みとることは難しい。

(4d) の *the Italians* をイタリア人 と訳しては正しい訳にならない。日本語名詞はその意味の中に、英語の the がもつ包括性と特定性をもたない。

(7)　a.　*The Italians* love pasta #but many/some/a few of them don't.　(Radden 2009: 221)
　　b.　*The Japanese* have felt a deep affinity with Mt. Fuji since ancient times, #but many/some/a few of them haven't.　(トラッドジャパン NHK 一部変更)

(8)　a.　*Americans* are patriotic, but of course not all of them are.
　　b.　*Americans* are patriotic, but of course some of them aren't.　(Radden 2009: 211)

(7) の *the Italians* や *the Japanese* が包括的で例外を含まないのに対し、(8a,b) の *Americans* は例外を認める表現である。

---

[4] この論文では、名詞が修飾語（句）を伴う場合も含め「裸名詞」と呼ぶ。
[5] 総称指示タイプの概念は、それぞれの名詞句が表わす個（単数・複数と定・不定）の意味と INSTANCE FOR TYPE, TYPE FOR SUBTYPE のメトニミーによって生成されるタイプの混合によって生成される（Radden 2009）

2.3　日本語の総称指示と英語裸複数形

日本語のタイプ表現には「裸名詞」が使用される。英語で直接「タイプ」を表わす表現は the＋NP 形式である。その意味で、日本語名詞が the の意味を含むという考え方もありそうである。しかし、英語の the＋NP 形式が、例外を含まない強い包括性をもつのに対し、日本語のタイプ表現は 包括性に関して柔軟である。総称指示に関する日本語裸名詞と英語の裸複数は、その表わす意味の柔軟性が非常に類似している。(9) は英語裸複数が総称指示に使用された場合の柔軟性（言語外に含まれる情報内容の多様性）を良く表わしている。

(9)　a.　Mammals give birth to live young.
　　b.　Alligators grow to attain a length of somewhere between fifteen and twenty feet.　(Most in fact perish as infants)
　　c.　Cardinals are red. (The females are a dull rust)
　　d.　Rats are bothersome to people. (Most in fact are never seen by anyone, or bother anyone)
　　e.　Bostonians are incredibly bad drivers. (Actually, only one in three is incredibly bad)
　　f.　Shoplifters are prosecuted in criminal court. (Most are never caught, much less prosecuted)

(Carlson 1977: 40)

3.　日本語名詞の可算性と質量性

日本語は可算名詞も質量名詞[6]も「裸名詞句」が項になる。この事実から、日本語名詞を全て質量名詞と捉える考え方がある。以下で、日本語名詞の可算性・質量性について考察を加える。

3.1　日本語名詞=質量名詞論

類別詞言語[7]の名詞に関する研究は数多い。中でも、その後の研究に強い影響を与えたものは Chierchia (1998) である。Chierchia は NP [+arg, -pred]タイプ言語の特徴として(10) の 4 点をあげる。

(10)　i.　Generalized bare arguments　　裸名詞が「項」になる
　　ii.　The extension of all nouns is mass　　名詞はすべて質量名詞
　　iii.　No PL　(No Plural marking)　　複数形態素をもたない
　　iv.　Generalized classifier system　　体系的な類別詞をもつ

Chierchia (1998) への反論として Cheng & Sybethma (1999) 水口(2010) 等がある。Cheng & Sybethma (1999) は、すべての言語で名詞は意味的に可算・質量の区別が認知され、英語のような言語では区別が名詞レベルで文法上（統語上）に反映され、中国語では類別詞のレベルでそれが反映される、と主張する。

---

[6] 普通名詞は一般的に可算と不可算に分類される。英語で普通名詞のタイプを論ずる場合 count./mass の表現が多く使用される。この論文ではそれに従い 可算名詞・質量名詞 と呼ぶ。

[7] Chierchia (1998) は類別詞言語という表現でなく NP [+arg, -pred] 型と呼んでいる。

Borer (2005) は Chierchia の主張をさらに進め、全ての言語において名詞が基底に持つ意味は質量であり可算性は統語的に決定されると主張する。

名詞の可算性・質量性に関するこれらの議論は、統語上の観点からのものであり、日本語のように名詞の可算性・質量性を統語的に明示する義務のない言語に関しては説明しきれない部分が残る。言い換えれば、名詞の可算性・質量性という議論は、名詞が統語上可算・質量を明示する言語についての議論と言える。次節で、Chierchia や Borer の不足点を、コーパスのデータを観察しながら証明する。

3.2　コーパスデータに見る日本語名詞の使用分布

3.2.1　可算名詞「橋」の使用分布

使用コーパス：国立国語研究所「現代日本語書き言葉均衡コーパス」モニター公開データ (2009 年度版)

検索条件：「橋」の正規表現のみを対象とし、比喩表現（例：夢のかけ橋、危ない橋を渡る)、固有名詞（例：利根大橋、本四架橋)、熟語（例：橋げた、橋梁）を除いた

使用区分を以下の６つに大別。

A. Object: 個体
B. Type: タイプ
C. Object/Type: 個体/タイプ両方の解釈が可能
D. D+NP: 決定詞＋名詞
E. Num+Cl+NP: 数詞＋類別詞＋名詞
F. Anaphora: 前方照応的用法

A. Object 型

「橋」が個体を指示する使用例。Anaphora 型と合計すると全体の 80% を占める。例文(11) の中の ささやかな橋$_2$ が Object に該当する。ちなみに 橋$_1$ が Type/Object に相当し、橋$_3$ が Anaphora に相当する。

(11) 道の両側は、変哲もないしもたやの家並みがつづいていて、少し歩くと小さな川にさしかかる。橋$_1$とも気づかず、つい通り過ぎてしまうようなささやかな橋$_2$がかかっていた。橋$_3$の右側の袂に、間口のせまい古道具屋があって、道路まで品物がはみ出していた。

（瀬戸内寂聴　『場所』　新潮社　2001）

B. Type 型

(12) しかし鋼鉄製の橋は、鋳鉄製の橋に比べて何倍も高く、良いことはわかっていても、メーカーもユーザーも手を出せなかった。　（田中孝顕監訳/Andrew Carnegie 騎虎書房 1990）

C. Type/Object 型

(13) ここだけは、渡し舟が通っておったですよ。川に橋がかかっていない時代でも、この村へは舟で行けたわけです。（荒俣宏 『陰陽師』 集英社 2002）

D. D+NP 型

(14) 桜橋を渡ろうとしていたのだ。私はまだその橋を一度も渡ったことがない。（半村良 『ぐい呑み』自選短篇集 廣済堂出版 1990）

E. Num+Cl+NP 型

(15) その合流点には高野橋と出町橋の二つの橋が架かっていて、さらにYの字の広くなる一本川に大橋が架かる。（水上勉 『在所の桜』 立風書房 1991）

F. Anaphora 型

(16) は固有名詞の前方照応型で、(17) は普通名詞の前方照応型である。

(16) あたりは薄暗くなっていた。永代橋を渡って深川に行こうと橋のたもとまでやって来た。橋を渡ろうとして、キラッ、と光ったものを見た。（峰隆一郎 『土方歳三』 徳間書店 2000）

(17) 松島の町の中を通り抜けて天竜川の橋の上に出たとき、彼は天竜川の水がいつになく濁っているのに気がついた。降雨があったのは一昨日である。上流からの濁り水が入りこんだとしても、もう澄んでいい頃であった。上流で工事でもしているのだろうか。そんなことを考えながら通りすぎようとしたとき、橋$_2$の向うからやって来る小沢銀兵衛の姿を見かけた。（新田次郎 『聖職の碑』 講談社 1976）

Table 1. 各使用区分の数と全体に占める比率

Object	Type	Type/Object	D+NP	Num+Cl+NP	Anaphora	TTL
410	51	29	90	10	297	887
46%	6%	3%	10%	1%	34%	100%

☛ Num+Cl+NP 型（数詞＋類別詞＋名詞句）の使用例は全体の 1%である。

☛ Object 型の中で、明らかに複数の橋を指していると解釈される例は 10 例未満であった。その中の 1 例を以下にあげる。

(18) 隅田川に架かる橋は、どの橋もみな幼馴染で、ただひとつ戦後にできた桜橋だけが、私には無縁の橋になっていた。（半村良 『ぐい呑み』自選短篇集 廣済堂出版 1990）

明らかに複数と解釈される例以外は、ほとんど単数と解釈される例であるが、橋の単数・複数が文法的に明示されていない限り明確ではない。読者は筆者の意図を文脈から正しく推量することが求められる。

以上で見てきたように、日本語は可算名詞が個体を指示する場合でも、それを統語的に明示することが義務的ではない。名詞の可算性・質量性は読者が文脈や世界知識から判断する。

4. まとめ

4.1 日本語名詞の持つ意味

(i) 物質を指示する場合、個体・質量両方の意味を表わす

(ii) 個体の単数・複数は文脈や世界知識から推量可能であるが、文法的に明示されない限り曖昧である

(iii) 英語定冠詞 the の意味を伝えるのは文脈であり名詞ではない。

(iv) タイプを指示する場合、英語裸複数形式と振舞いにおいて類似性を示す

(19) は川端康成「山の音」からの引用で、英訳は Edward G. Seidensticker による。原文に使用されている名詞「手拭」「煙草」「夕刊」「老眼鏡」は可算であるが、数についての言及はない。英語訳では、可算名詞の統語的な明示が必須であるために a towel, cigarettes, the evening paper, glasses と表現されている。しかし、原文の日本語は数に関して断定していないので単数にも複数にも解釈可能である。ちなみに不可算名詞の「番茶」は英語も裸形式 tea である。

(19) a. その後から、菊子が冷やした手拭と煙草などを持ってきて、また湯呑に熱い番茶をついだ。一度立って、夕刊と老眼鏡を持ってきた。

（川端康成『山の音』日本文学全集 30 川端康成集　新潮社　1959: 299）

b. Kikuko came after him with a cold towel and cigarettes and poured more tea. Then she went for his glasses and the evening paper.

(*The sound of the mountain* translated by Edward G. Seidensticker)

(19) は文脈や世界知識を援用しても、曖昧さが払拭されないケースの一例である。1 本の手拭、複数の煙草、1 個の老眼鏡、は訳者が文脈や世界知識から判断したものである。夕刊は当日の新聞であると判断した訳者が the evening paper としたのである。しかし、実際の数が明示されていないことは事実であり、あくまでも推量である。

以上、統語的な分析だけでは説明が不十分である日本語名詞の持つ意味について、日英語翻訳の観点から分析を試みた。冠詞や複数形態素をもたない言語における名詞の意味は様々な観点から分析を加えることにより、その本当の姿が見えてくると考えられる。

REFERENCES

Borer, Hagit (2005) *Structuring Sense volume 1: In Name Only*, Oxford University Press, Oxford.

Carlson, Greg (1977) "Reference to Kinds in English", Ph. D dissertation, University of Massachusetts, Amherst, *Outstanding dissertation in linguistics* (1980) Garland Publishing, NY

Cheng, Lisa Lai-Shen and Rint Sybethma (1999) "Bare and Not-So-Bare Nouns and the Structure of NP", *Linguistic Inquiry*. Vol 30, No.4 pp.509-542, MIT press

Chierchia, Gennaro (1998) "Reference to Kinds across Languages", *Natural Language Semantics* 6, 339-405.

平子義雄, 1999 『翻訳の原理』大修館書店

Krifka, Manfred et al. (1995a) "Genericity: An Introduction", Carlson, Gregory N. and Francis Jeffry Pelletier (ed.) *The Generic Book*, 1-124, The University of Chicago Press, Chicago and London

Krifka, Manfred (1995b) "Common Nouns: A Contrastive Analysis of Chinese and English", Carlson, Gregory N. and Francis Jeffry Pelletier (ed.) *The Generic Book*, 398-411, The University of Chicago Press, Chicago and London

水口志乃扶. 2004 「日本語の類別詞の特性」 水口・西光編 『類別詞の対照』61-77

水口志乃扶. 2010 「類別詞言語のものの考え方」 日本英語学会第 140 回大会予稿集 10-15

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*, Longman Group Limited

Radden, Günter and René Dirven (2007) *Cognitive English Grammar*, John Benjamins, Amsterdam /Philadelphia

Radden, Günter (2009) "Generic reference in English: A metonymic and conceptual blending analysis" Klaus-Uwe Panther, Linda L. Thornburg, and Antonio Barcelona (ed.) *Metonymy and Metaphor in Grammar*, 199-228, John Benjamins, Amsterdam/Philadelphia

コーパス：国立国語研究所「現代日本語書き言葉均衡コーパス」モニター公開データ (2009 年度版)

日本語の定型表現の読解処理　－瞳孔径の計測データを用いて－

梁志鋭（名古屋大学国際開発研究科博士後期課程）
阪上辰也（名古屋大学国際開発研究科）
古泉隆（名古屋大学教養教育院）
坂東貴夫（名古屋大学国際開発研究科博士後期課程）

Processing Japanese Formulaic Sequences in Reading: A Pupillometric Study

Chi-Yui Leung (Graduate School of International Development, Nagoya University)
Tatsuya Sakaue (Graduate School of International Development, Nagoya University)
Takashi Koizumi (Institute for Liberal Arts & Sciences, Nagoya University)
Takao Bando (Graduate School of International Development, Nagoya University)

1. 本研究の目的

本研究の目的は，日本語の定型表現としての複合動詞の読みにおいて，使用頻度による認知処理への影響があるかどうかを，瞳孔径の計測によって明らかにすることである。

２．先行研究

２．１　複合動詞

複合動詞には, 様々な種類があり, 複合動詞の分類方法は異なる（長島 1976; 野村 1977)。陳（2008: 12）は，複合動詞を「動詞連用形＋動詞」（例：笑い出す）のような「狭義の複合動詞」と，それ以外にも「名詞＋動詞」（例：名付ける）や「形容詞語幹＋動詞」（例：近寄る）などを含む「広義の複合動詞」に大別しており，これまでの複合動詞の研究において「『広義の複合動詞』の立場に立つ研究より『狭義の複合動詞』に立つ研究のほうが圧倒的に多い」としている。

本研究では,「動詞連用形＋動詞」という「狭義の複合動詞」を対象とする。

２．２　頻度効果

単語認知における頻度効果とは,「出現頻度の高い単語が出現頻度の低い単語より容易に認知される」という現象を指す（斎藤 2004: 103)。この頻度効果が特に語彙判断課題や命名課題などの視覚的単語認知課題において多く検討されており，高頻度単語の方が，低頻度単語より速く，正確に同定されると知られている（広瀬 2007)。

２．３　瞳孔径の計測

Beatty（1982)，Beatty & Lucero-Wagoner （2000)，松永（1990）によると，人間の瞳孔は音声・視覚的刺激，感情的変動，言語・認知活動などにより拡大し，そして元のサイズに収縮する。この現象から，瞳孔反応の測定が 1960 年代以降に心理学の研究に用いられてきた（松永 1990)。瞳孔のサイズの測定値としては瞳孔の直径（瞳孔径）と面積が挙げられ

るが，近年の心理学の研究では瞳孔径が測定値として用いられる傾向がある。

瞳孔反応の計測を用いる言語情報の処理に関する研究としては Just & Carpenter（1993），Hyona, Tommola, & Alaja（1995），Kuchinke et al.（2007）や Briesemeister et al.（2009）などが挙げられる。これらの実験結果によると，瞳孔反応の計測は言語課題において，Beatty（1982），Beatty & Lucero-Wagoner （2000）の指摘の通りに，言語課題の遂行における認知負荷の変化を測定するのに有効な手段として考えられる。すなわち，より難しいと判断される言語課題において瞳孔径が大きくなる傾向がある。

その中で，Kuchinke et al.（2007）は語彙判断課題を用いて，ドイツ語母語話者を対象に，ドイツ語の単語の使用頻度と感情価（emotional valence）と瞳孔径の変化について調べたところ，単語の頻度効果について，低頻度語を処理する時が高頻度語を処理する時より，瞳孔径が大きいという結果を報告している。

瞳孔径の変化はこれまで多くの研究で実証されており，有効な認知負荷の指標とされているにもかかわらず，他の手法と比べると，これまで決して大いに用いられているとは言えない。しかし，Beatty & Lucero-Wagoner（2000）は，これまで数多くの実証的証拠に基づいて考えると，認知過程における瞳孔反応がこれからの研究においても，認知負荷を測定するのに重要な指標として用いられるべきであると主張している。

３．研究課題

単語認知における使用頻度が瞳孔径に及ぼす影響について，日本語を対象とする研究は管見の限りまだない。また，瞳孔径の計測を用いた研究の中で，外国語学習者を対象とする研究は，Hyona, Tommola, & Alaja （1995）以外にまだあまり用いられていない。一方，複合動詞についても，学習者を対象に複合動詞の認知処理に焦点を当てた研究もまだ少ない。

従って，本研究は，日本語母語話者と学習者を対象に，瞳孔径の計測を用いて，日本語複合動詞の使用頻度の違いにより瞳孔径のサイズが異なるのかを研究課題として設定する。高頻度語より，低頻度語のほうが処理の困難度が高いとされるため，低頻度の複合動詞の処理において，認知負荷の指標とされる瞳孔径はより大きくなると予測される。

４．実験方法

４．１　実験参加者

日本語母語話者（NS）3 名と，中国人日本語学習者（NNS）3 名（日本語能力試験一級）である。

４．２　実験課題

実験デザインは主に Kuchinke et al.（2007）を参照し，語彙判断課題を用いた。

提示される刺激語は合計 80 個である。そのうち，正しく判断すべき日本語の複合動詞が 40 個あり，高頻度の複合動詞 20 個と低頻度の複合動詞 20 個に分けられている。残る 40 個

の刺激語は正しくないと判断すべき非単語である。すべての刺激語を 2 つのブロックに分け，ランダムに提示した。プログラムによる提示方法（図 1 参照）に関しては，最初に凝視点が 800 ミリ秒間提示され，その後，単語が最大 1700 ミリ秒間提示される。その間，まばたきせず，提示された刺激が日本語の単語であるかどうかを，ボタンを押すことによって判断するように実験参加者には指示がされた。1700 ミリ秒間の制限時間内にボタンを押すと単語が消え，再び凝視点が提示される。最初の凝視点から 2500 ミリ秒間が経過すると「^_^」というレストマークが現れるようになっている。レストマークが提示されている間に実験参加者がボタンを押すと次のトライアルが提示される。

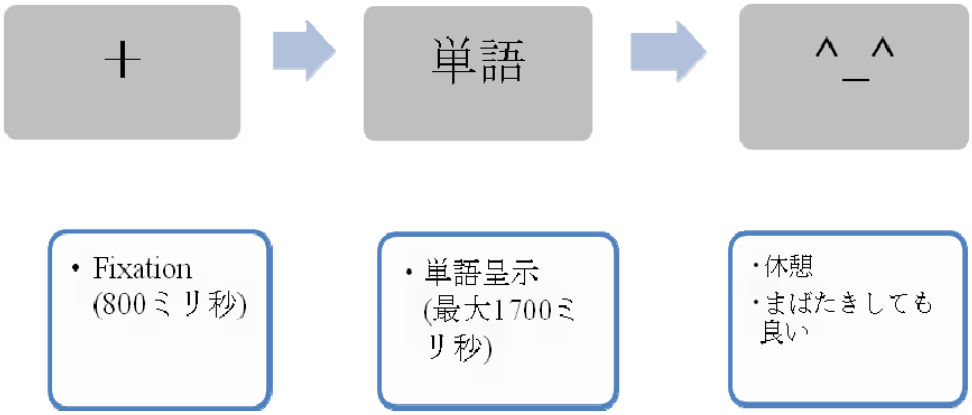


図 1　刺激の提示

また，刺激提示時の背景色は黒（RGB: 0, 0, 0）であり，文字はグレー（RGB: 120, 120, 120）であり，フォントは「新明朝」を用いた。

４．３　実験環境と手順について

実験環境（図 2 参照）と手順は以下の通りである。

① 実験参加者が，瞳孔画像を取得するための実験装置－EyeLink 1000（SR Research 社）－のカメラと刺激提示用のモニターが置かれた机の前に座る。

② 実験参加者がヘッドレスト（あご台）に顔を乗せた状態で楽な姿勢になるよう、座席の調節を行う。モニターの位置と実験参加者の目の位置との距離は約 65cm である。

③ 実験方法に慣れてもらうため，練習課題（10 題）を行う。

④ 実験課題を行う。

図 2　実験環境

４．４　刺激語

BCCWJ モニター公開データ（2009 年度版）を用い，古泉他（2011）で示された手順にしたがって複合動詞の抽出を行った。抽出された複合動詞のリストから，高頻度の複合動詞 20 個（頻度順で上位 20 個）と低頻度の複合動詞 20 個（頻度順で 444 位～494 位にある語 20 個）を抽出した。これら 2 つの単語グループの親密度・画数・モーラ数・心像性は天野・笠原・近藤（2008），天野・近藤（1999）と佐久間他（2005）を参考に統制された。一方，非単語は存在する日本語の動詞 2 つを組み合わせて作られている。

実験で用いられた複合動詞は表 1 の通りである。

表 1　実験で用いられた複合動詞

高頻度複合動詞	繰り返す，取り組む，受け取る，取り出す，立ち上がる，取り上げる，振り返る，生み出す，生み出す，差し出す，取り戻す，取り巻く，引き受ける，組み合わせる，取り扱う，思い切る，見出す，乗り出す，取り除く，乗り込む，持ち込む
低頻度複合動詞	見下ろす，抜き出す，座り込む，待ち受ける，取り調べる，出回る，追い越す，築き上げる，泣き出す，思い返す，切り裂く，盛り上がる，行き詰まる，振り払う，突き上げる，引き離す，請け負う，舞い上がる，振り回す，引っ込む

４．５　データの処理と分析

瞳孔径データはミリ秒単位で計測された。各トライアルにおける瞳孔径の計測時間帯は最初の凝視点提示の開始からレストマークが現れるまでの 2500 ミリ秒間である。そのうち，単語が提示される直前の 300 ミリ秒間と，提示後の 1700 ミリ秒間のデータポイントにおける瞳孔径の RAW データは z スコアに変換された。各実験参加者においては，各条件における z スコアをデータポイントごとに平均化し，折れ線グラフを作成した（図 3～図 8）。さらに，高頻度複合動詞・低頻度複合動詞・非単語条件について，瞳孔径の最大値を求め，時間軸にそって最大値および最大値の前後それぞれ 30 個のデータポイント（合計 61 個）

から構成される瞳孔径データの平均値と標準偏差を算出した。

また，外れ値に関しては，以下の基準で処理した。

① 正しく判断される項目のみを分析対象とした。

② 各実験参加者の反応時間(刺激語が提示されてからボタンを押すまでの時間)のデータの平均値と標準偏差を基に，平均値から標準偏差±2 倍の範囲内を分析対象とした。

③ 2000 ミリ秒間の瞳孔径データにおいて，まばたきなどの原因によって欠損値が出たトライアルは除去した。

５．結果

５．１　瞳孔径

各実験参加者の瞳孔径の変化は図 3～図 8 の通りである。図の縦軸の値は瞳孔径の z スコアであり，横軸の値は時間軸である。時間軸に表示される値は上述したように，刺激提示前の 300 ミリ秒から刺激 提示語の 1700 ミリ秒までの 2000 ミリ秒間である。また，瞳孔径のデータが z スコアに変換されたため，各条件の平均値は 0，標準偏差は 1 である。

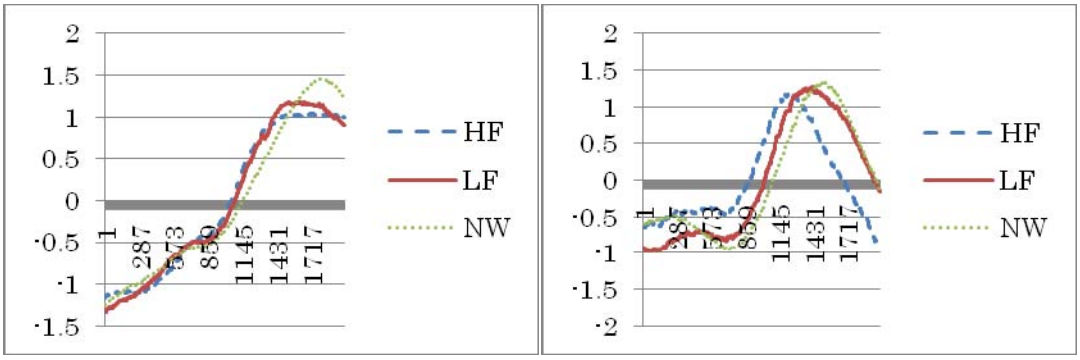


図 3（左）と図 4（右）実験参加者 NS①（左）と NS②（右）の瞳孔径の変化

（HF：高頻度条件，LF：低頻度条件，NW：非単語条件，以下同様）

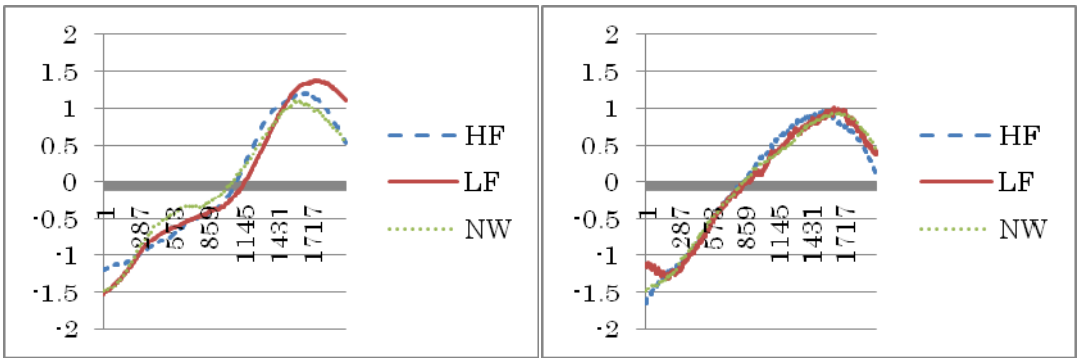


図 5（左）と図 6（右）実験参加者 NS③（左）と NNS④（右）の瞳孔径の変化

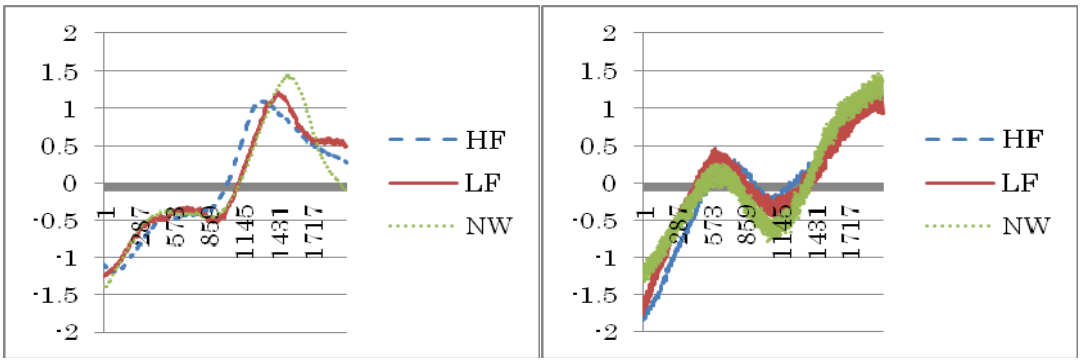


図 7（左）と図 8（右）実験参加者 NNS⑤（左）と NNS⑥（右）の瞳孔径の変化

実験者ごとの各条件における 61 個のデータポイントでの瞳孔径データの平均値と標準偏差は表 2 の通りである。

表 2　瞳孔径データ（カッコ内：標準偏差）

	高頻度	低頻度	非単語
NS①	1.0349(0.0083)	1.1673(0.0064)	1.4532(0.0060)
NS②	1.2043(0.0079)	1.3678(0.0081)	1.0699(0.0212)
NS③	1.1670(0.0203)	1.2371(0.0234)	1.3092(0.0087)
NNS④	0.9464(0.0154)	0.9673(0.0276)	0.7119(0.0284)
NNS⑤	1.0758(0.0114)	1.1712(0.0178)	1.4253(0.0179)
NNS⑥	1.2133(0.0204)	1.0607(0.0615)	1.2896(0.1039)

表 2 の結果に基づき，瞳孔径データの結果をまとめると，表 3 の通りとなる。

表 3　瞳孔径の条件間の差

実験参加者	高頻度条件と低頻度条件の比較
NS①	低頻度 ＞ 高頻度
NS②	低頻度 ＞ 高頻度
NS③	低頻度 ＞ 高頻度
NNS④	低頻度 ＞ 高頻度
NNS⑤	低頻度 ＞ 高頻度
NNS⑥	高頻度 ＞ 低頻度

５．２　反応時間と正答率

各実験参加者の 3 つの条件における反応時間と正答率は表 4 の通りである。

表 4　反応時間と正答率（カッコ内：標準偏差）

実験参加者	反応時間			正答率		
	高頻度	低頻度	非単語	高頻度	低頻度	非単語
NS①	767(127)	890(185)	1030(102)	100%	100%	95%
NS②	620(61)	730(111)	817(110)	100%	100%	95%
NS③	620(61)	730(111)	817(110)	100%	90%	95%
NNS④	936(201)	983(147)	1112(189)	95%	85%	75%
NNS⑤	653(103)	759(102)	918(76)	90%	80%	55%
NNS⑥	850(142)	954(134)	1066(144)	100%	95%	85%

６．考察

６．１　複合動詞の使用頻度の差による瞳孔径の違い

研究課題について，瞳孔径データから，特に NS の場合，低頻度条件においてより瞳孔径が大きくなる傾向が見られた。対象言語が異なるものの，Kuchinke et al.（2007）と類似した結果となった。したがって，今回の実験結果は，瞳孔径の計測が日本語の単語認知における頻度の違いによる認知負荷の差を調べるのに有効な手法であることが示唆された。

６．２　反応時間

反応時間では，多くの先行研究で示されてきたように，低頻度の単語処理が高頻度の単語処理より時間がかかるという傾向が見られた。本実験で用いられた複合動詞の項目では，頻度により処理速度が異なり，処理する際の認知負荷も異なると解釈できる。

６．３　非単語

非単語条件については，反応時間に関して，全体的に非単語の処理の方が高頻度語と低頻度語の処理よりも時間がかかった。しかし，瞳孔径データの結果ではかなり個人差が見られた。Kuchinke et al.（2007）によると，反応時間は処理の速さを反映するものの，処理の速さは必ずしも認知負荷の量を表す瞳孔径の結果とは一致しない。したがって，非単語条件に関する結果は，実験参加者による非単語の語彙判断において，処理の速度には影響しないが認知負荷には影響する要因があるためではないかと推測できる。

７．まとめ

今回の実験結果では，複合動詞の使用頻度によって瞳孔径のサイズが異なることが示された。また，母語話者と学習者の比較に関して，母語話者では 3 名の実験参加者の瞳孔径結果から一致した傾向が見られたが，学習者では 1 名の実験参加者の結果が他の実験参加者の結果とは異なった。この理由としては，複合動詞の処理における頻度効果が学習者においては母語話者ほど顕著ではない可能性が考えられるが，実験参加者が少ないため現段階では断言できない。日本語教育において，複合動詞が学習者にとって習得困難とされる

項目の 1 つとして挙げられているため（森田 1990），今後，実験参加者の人数を増やし，学習者と母語話者の複合動詞の処理における違いについてさらに注目していきたい。

文献

天野成昭・近藤公久 (1999). 『日本語の語彙特性－第 1 期書籍+CD-ROM 版』三省堂

天野成昭・笠原 要・近藤公久 (2008).『日本語の語彙特性－第 4 期書籍+ CD-ROM 版』三省堂

Beatty, J. (1982). Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin*, 91(2), pp. 276-292.

Beatty, J. & Lucero-Wagoner, B. (2000). The Pupillary System. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (2nd edition ed., pp.142-162). Cambridge: Cambridge University Press.

Briesemeister, B.B., Hofmanna, M.J., Tamma, S., Kuchinke, L., Brauna, M., & Jacobs, A.M. (2009). The pseudohomophone effect*:* Evidence for an Orthography-phonology-conflict. *Neuroscience Letters*, 455, pp. 124-128.

陳曦 (2008).『第二言語としての日本語複合動詞の習得－コーパスによる使用実態の調査を中心に－』，名古屋大学提出博士論文

広瀬雄彦 (2007).『日本語表記の心理学－単語認知における表記と頻度』北大路書房

Hyona, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quarterly Journal of Experimental Psychology,* 48A, pp. 598-612.

Just, M. A. & Carpenter, P. A. (1993). The intensity dimension of Thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology, 47*, pp. 310-339.

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－」『BCCWJモニター公開データ利用者ポスター発表予稿集』(2011年3月14日，東京)

Kuchinke et al. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychology*, 65, pp. 132-140.

松永勝也 (1990). 『瞳孔運動の心理学』ナカニシヤ出版

森田良行 (1990). 『日本語学と日本語教育』凡人社

長島善朗 (1976). 「複合動詞の構造」『日本語講座 4　日本語の語彙と表現』大修館書店, pp. 73-104.

野村雅昭 (1977). 「造語法」『岩波講座日本語 9』岩波書店

斎藤洋典 (2004). 「心的辞書」『言語の科学 3　単語と辞書』岩波書店, pp. 93-153.

佐久間尚子・伊集院睦雄・伏見貴夫・辰巳　格・田中正之・天野成昭・近藤公久 (2005).『日本語の語彙特性－第 3 期書籍+CD-ROM 版』三省堂

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—

古泉隆（名古屋大学教養教育院）†
梁志鋭（名古屋大学大学院国際開発研究科博士後期課程）
阪上辰也（名古屋大学大学院国際開発研究科）
坂東貴夫（名古屋大学大学院国際開発研究科博士後期課程）
天野修一（名古屋大学大学院国際開発研究科博士後期課程）
新實葉子（名古屋大学大学院国際開発研究科博士後期課程）

Development of Web-Based Materials for Learning Japanese Compound Verbs: Using the Frequency List Derived from BCCWJ

Takashi Koizumi (Institute for Liberal Arts & Sciences, Nagoya University)
Chi-Yui Leung (Graduate School of International Development, Nagoya University)
Tatsuya Sakaue (Graduate School of International Development, Nagoya University)
Takao Bando (Graduate School of International Development, Nagoya University)
Shuichi Amano (Graduate School of International Development, Nagoya University)
Yoko Niimi (Graduate School of International Development, Nagoya University)

１．はじめに

日本語の複合動詞（動詞連用形＋動詞）は、使用頻度が高く（森田, 1978）、また動詞単独では表しきれないことを表現するなど重要な役割を担っている（永井, 1996）。その一方で、学習者向けの教材が不十分であることが指摘されている（何, 2010; 森田, 1978）。また、実際の授業運営において、複合動詞の学習にクラス時間を十分確保することが難しいとの指摘もある（森田, 1978）。そこで本研究では、日本語大規模コーパスである BCCWJ をもとに、複合動詞の頻度リストを作成するとともに、学習者が授業外でも繰り返し学ぶことができるように Web 上で提示するツールを開発した。

２．複合動詞の頻度リスト作成

本研究では、「BCCWJ2009 モニター版コーパス」を使用し、複合動詞の候補の抽出・選定を行った。複合動詞候補の抽出にあたっては、UNIX システム上でのテキスト処理、および、形態素解析プログラムによる処理を行った。複合動詞の頻度リスト作成の主な手順は、以下の通りである。なお、検索対象としたテキストは、白書・書籍・Yahoo!知恵袋・国会会議録の 4 種類である。

1) コーパスの文字コードの変換
2) 文末と思われる箇所での改行コード挿入
3) 形態素解析プログラムでの品詞付与
4) 複合動詞候補の抽出
5) 母語話者による確認・選定

まず、以下のように nkf コマンドを実行し、文字コードの変換を行った。

† koizumi@nagoya-u.jp

```
$ nkf --ic=UTF-16LE-BOM --oc=UTF-8 bccwj.utf8
```

BCCWJ2009 モニター版コーパスに含まれるテキストの文字コードは UTF-16 であるが、後に形態素解析プログラムの MeCab を用いた処理をするために、UTF-8 への変換を行った。

続いて、文末に生じる句読点などの記号類を基準として、それらにマッチした直後に、UNIX 用の改行コード（LF）を挿入した。使用した元のテキストデータは、改行位置が必ずしも文末に統一されておらず、この点が形態素解析や複合動詞の抽出に影響を及ぼすおそれがある。そのため、以下の Perl スクリプトを実行し文末に改行を挿入することで、１行に１文という統一した形式に整形した。

```
$ perl -pe 's/(。|？|！|¥!|¥?)+/¥1¥n/g;' bccwj.utf8 > bccwj1.utf8
```

次に、形態素解析プログラムの MeCab を利用してコーパスデータに品詞を付与した。F オプションにより、出現形・品詞・基本形を出力するように指定し、その解析結果を引き継いで、grep コマンドを指定して、「動詞」という文字列を含んだ行と直後に続く１行を抽出した。

```
$ mecab -F "%m¥t%f[0,6]¥n" bccwj1.utf8 | grep -A 1 "動詞" > fu9go.1
```

その後、処理結果内に含まれる動詞以外の品詞を除去した。

```
$ grep -A 1 "動詞" fu9go.1 | grep -Ev "(助動詞|助詞|記号|名詞|接続詞|
  感動詞|副詞|形容詞|連体詞)" >fu9go.2
```

上記の処理後、cut コマンドにより、一番目のフィールド（出現形）のみを取り出した。

```
$ cut -f 1 fu9go.2 > fu9go.3
```

解析結果は、形態素１つにつき１行で出力されていることから、複合動詞が２語に分けて解析されている場合[1]は、複数行にまたがって出力されることになる。そのため、Perl のスクリプトを用いて、複合動詞を１つに連結し直した。

```
$ perl -pe 's/¥n//;' fu9go.3 | perl -pe 's/¥-+/¥n/g;' > fu9go.4
```

ここまでの一連の処理を行うと、複合動詞だけでなく、１語の単独動詞も処理結果に含まれることになる。テキスト処理で、「動詞＋動詞」の組み合わせのみを抽出するのは容易だが、その場合、「思い出す」のように形態素解析プログラムで１語として解析される複合動詞が除外されてしまう。そのため、この段階では複合動詞の候補となる項目を漏れなく

[1] 複合動詞と解釈可能な動詞でも、MeCab では１語として解析する場合と、２語に分けて解析する場合がある（例：「思い込む」は１語の動詞として解析されるが、「囲い込む」は、「囲う」と「込む」の２語の動詞に分けて解析される）。

抽出するために、1 語の動詞も処理対象に含めている。

処理のまとめとして、sort コマンドと uniq コマンドにより、頻度計算を行った。

```
$ sort fu9go.4 | uniq -c | sort -nr > fu9go.freq
```

最後の段階として、上記処理で抽出された複合動詞の候補リストを、２名の日本語母語話者がすべて確認した。リスト中にある単独動詞を除外し、さらに、「食べさせる」のような使役形や「食べさせられる」のような受身形を含んだ表現も除外した。その結果、最終的に約 8000 の項目が複合動詞として選定された。

３．Web 提示ツール

本研究では、複合動詞の頻度リストを作成するとともに、その学習を支援するツールもあわせて開発した。授業内の限られた時間で複合動詞を十分に学習することが難しいとの指摘（森田, 1978）を踏まえ、本研究では授業内で不足する学習を補うようなツールを目指した。具体的には、Web 上で複合動詞を提示し、授業外でも繰り返し学習が可能なツールを開発した。ツールの開発にあたっては以下を考慮した。

1) 特定の OS に依存せずに動作する
2) iPhone などのスマートフォンにも対応する
3) 教師が容易に表示項目を入れ替えることができる
4) 「見る・読む」ことで学習する
5) 単調にならないようにアニメーション効果を取り入れる

最近では、Mac OS など Windows 以外の OS を使用するユーザーが増えている。学習環境を広げるには、特定の OS に依存するのではなく、幅広い機種で使用できるようにすることが望まれる。また、iPhone 等のスマートフォンも普及しつつあり、これらを利用することで、学習環境をどこにでも持ち歩くことがより容易になる。そこで、本ツールは、ブラウザ上で動作し、スマートフォンを含めた様々な端末で利用できるように配慮した。図 1 は iPhone のブラウザ上で本ツールを使用した例である。

図 1：iPhone で本ツールを使用した画面

次に、教師自身が表示項目を容易に入れ替えることができるようにした[2]。複合動詞に関して何をどのように提示するかは、授業での指導法や学習者のレベルにより異なる可能性がある。教材の表示項目が変更できないと、授業の実態に即して現実的な利用が困難になる。本ツールでは、表示項目をテキストファイルに記述するようにし、Windows の「メモ帳」などのテキストエディタで容易に追加・編集することができるようにした。図 2 は本ツールのファイル構成である。Web ページの本体である HTML ファイル、Web ページの動きを制御したりアニメーション効果を与える JavaScript ファイル[3]、そして表示項目が記述されたテキストファイルで構成されている。このように、動作に関わるファイルと表示項目に関わるファイルとを切り分けることで、プログラミングなどの知識がなくとも、ワープロで文章を編集する要領で表示項目を変更できる。

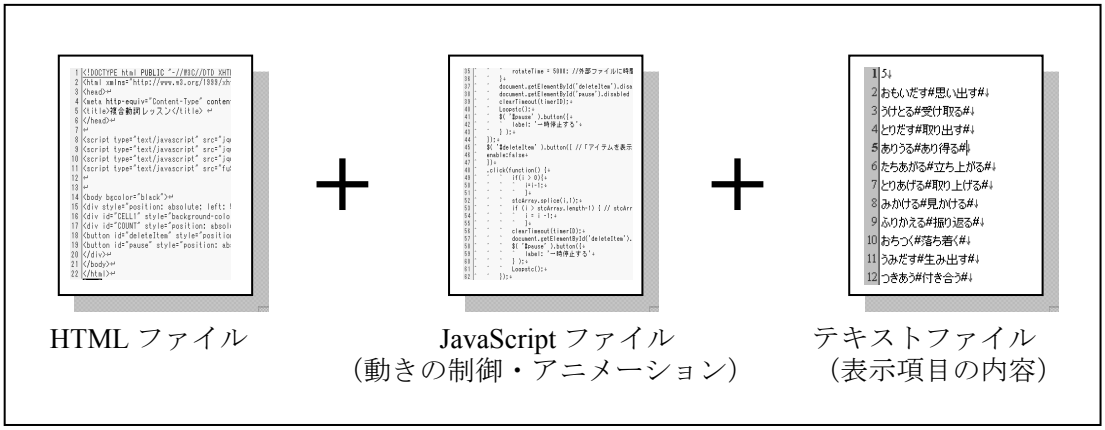


図 2：Web 提示ツールのファイル構成

また、本ツールは、用意した項目をブラウザ上で繰り返し表示することが主な機能であり、学習者は表示された項目を何度も「見て」学習することになる。本ツールでは、あえて「書く」（入力する）ことによる学習を課していないが、これはスマートフォンなどキーボードをもたない機器で文字を入力するのは比較的困難であり、提示したものを「見る」ことに限定したほうが使い勝手がよいと思われるためである。さらに学習効果という点でも、「書く」学習は行なわずとも、読んだり見たりすることで学習効果があるとの結果が報告されている（Barcroft, 2006; Webb, 2005）。文字入力を課して学習者に余分なストレスを与えるより、使い勝手を高めることで、学習への敷居を低くしている。

一方で、提示される項目を見るだけというのは単調な学習になる可能性がある。そこで、本ツールでは、単調さを軽減するために提示する際の視覚効果を工夫した。具体的には、提示項目が切り替わる際に提示ボードが回転するような視覚効果を取り入れている（図 3）。Sakaue et al. (2010)では、これと同様の視覚効果を取り入れた提示方法をアンケート調査しているが、多くの学習者がこの提示方法に対して好意的に回答している。単に文字列のみ

[2] 作成した表示項目のテキストファイルおよび本提示ツールは、Web サーバーにアップロードすることでインターネットを経由してアクセスさせることもできるが、USB メモリなどで図 2 のファイル一式を学習者に渡し、学習者のローカル PC 上で直接 HTML ファイルを起動させることもできる。ただし、iPhone の場合は、前者のインターネット経由でのアクセスに限られる。

[3] JavaScript ライブラリとして jQuery を使用した。

を切り替えて表示するより、飽きずに学習することができると思われる。

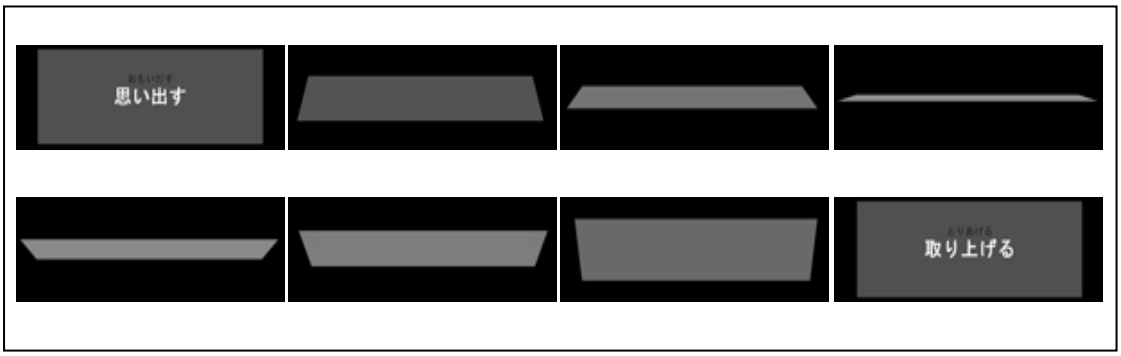


図 3：提示ボード回転の様子

４．応用例

本ツールは項目を提示することが主な機能であるが、提示の方法を工夫することで、学習者に答えを考えてもらうような問題形式で提示することも可能である。ツールに表示する項目はテキストファイルに記述されているが、１項目につき１行で記述されている。ここで例えば、偶数行に問題文を書き、次に続く奇数行にその解答を記述することで、「問題→解答」の順で提示されることになる。これにより、学習者は最初に提示された問題文を見て解答を考え、その次に提示される解答で答えを確認することができる。また、表示項目に HTML タグを記述することもできるため、タグを記述して重要な箇所の色を任意に変えることもできる[4]。

５．複合動詞リストと Web 提示ツールの公開

選定した約 8000 の複合動詞と、Web 提示ツールにその一部分を組み込んだサンプル教材を以下の URL のサイトで公開している。

http://lab.sakaue.info/wiki.cgi/iColl2010?page=Fu9Go

６．今後の課題

本研究で作成した複合動詞のリストは、「動詞の連用形＋動詞」の組み合わせを頻度順で並べたものである。しかしながら、森田(1978)が指摘するように、複合動詞の教材作成には、出現頻度を基準にするほか、原義からでは類推が難しいものや意味にずれを生じる複合動詞も抽出すること、さらには、複合動詞の中で頻繁に用いられる単独動詞を抽出し、それらを体系的に整理することも必要である。また、開発した Web 提示ツールに関しては、ツールのみで提示できる情報は非常に限られており、他の学習形態と補完し合う必要がある。本ツールでの学習が授業等での対面指導をどのように補完できるか、検討する必要がある。

[4] ツール独自の機能として、テキストファイルの文字列を<<>>で囲むと、その文字列が赤く表示されるようになっている。

文献

何志明 (2010).『現代日本語における複合動詞の組み合わせ―日本語教育の観点から』, 笠間書院.

永井鉄郎 (1996).「日本語複合動詞の教育について」『日本語教育』, 88, pp.140-151.

森田良行 (1978).「日本語の複合動詞について」『講座日本語教育』, 早稲田大学語学教育研究所, 14, pp.69-86.

Barcroft, J. (2006). Can Writing a New Word Detract from Learning It? More Negative Effects of Forced Output During Vocabulary Learning. *Second Language Research*, 22, 4, pp.487-497.

Sakaue, T., Amano, S., and Koizumi, T. (2010). *A new collocation learning tool: iColl*, Poster presented at GLoCALL 2010 at Le Meridien Kota Kinabalu, Sabah, Malaysia.

Webb, S. (2005). Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition*, 27, pp.33-52.

心理学実験とコーパスに基づく「上」の意味ネットワークの実証的研究

徐 蓮(北京外国語大学　お茶の水女子大学)

A Study on the Semantic Network of "UE" Based on Psychological Experiments and Corpus

Xu Lian（Beijing Foreign Studies University /Ochanomizu University）

1. はじめに

本稿では、認知意味論の視点から、心理学実験とコーパス考察のデータに基づいて、「上」の量的な意味ネットワークを構築し、実証的な多義語研究モデルを提案する。

1.1　先行研究とその問題点

「上」は典型的な多義語として、その意味構造に関する研究が多くの言語で盛んに行われている。(英語：William（1974)、Lakoff（1980,1987)、Dewell(1994)、Tyler (1989)、Tyler & Evans(2001)；日本語：山梨（1998)、瀬戸（1995)、谷口(2003)；中国語：沈（1999)、蓝（1999)；ロシア語：张(2001)；韓国語：赵（2002）等々)。方法論の面では、その意味構造を明らかにするにはために、多数の研究モデルがある。例えば、イメージ・スキーマに基づくモデル（Brugman(1981) 、Lindner(1982) 、Lakoff(1987),Dewell(1994) 、Tyler & Evans(2001))、拡張とスキーマに基づくモデル（Langacker(1987,1988))、現象素に基づくモデル（国広(1994，1995))、コアに基づくモデル（田中(1990))、ネットワークと現象素を統合したモデル（籾山(1994,2000)）等々が挙げられる。

多義語の意味構造研究が深まっている一方、限界が見えてくる。これまでの分析モデルは、内省に基づいたため、カテゴリーの区分、プロトタイプの認定、イメージ・スキーマのまとめと拡張プロセスの分析の厳密性に疑問があるとされている。研究の科学性と緻密性を高めようと、本稿では心理学実験とコーパスに基づく実証的研究モデルを提案する。

1.2　研究課題と研究方法

本稿では、「上」のカテゴリー構造、プロトタイプ、イメージ・スキーマ、拡張距離、拡張手段と拡張力を記述し、量的な意味ネットワーク（図 1）を構築する。

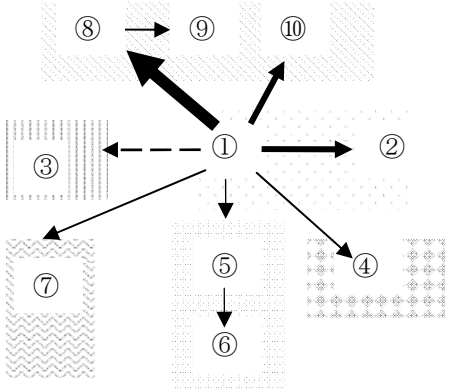


図 1　意味ネットワーク見取り図

具体的な研究課題と方法は次のようである。

（1）カテゴリー構造（網かけの部分）　カード分類実験を通じて、「上」のカテゴリー数、結合プロセス、カテゴリーの区分基準を考察する。

（2）プロトタイプ（拡張の原点①）　文の想起実験を通じて、母語話者に一番想起されやすい、しかも多く想起される用法を見つけ、プロトタイプ的意味を認定する。

（3）イメージ・スキーマ　図の想起実験を通じて、「上」のイメージを調査し、イメージ図のスキーマ要素を抽出し、イメージ・スキーマをまとめる。

（4）拡張距離（矢印の長さ）　カード分類実験のデータの多次元尺度分析の結果から、各用法の間の心理的距離を測る。

（5）拡張手段（線の形）　母語話者への非構造化インタビューに基づいて、各用法が拡張してきた手段（メタファー、メトニミー、シネクドキーなど）を分析する。

（6）拡張力（矢印の太さ）　コーパスにおけるサンプル調査を通じて、各用法の使用頻度を記述する。

以上の結果に踏まえ、「上」の意味ネットワークを明らかにする。

なお、本稿では名詞「上（うえ)」の単純語を研究対象とし、動詞（「上がる」等）と複合語（「上着」等）は除外される。

2. カテゴリー構造

類似している用法は一つのグループにまとめやすいという原理に基づいて、「上」の各用法の用例を被験者に分類させ、その結果を集計し、多次元尺度解析とクラスター分析を通じて、カテゴリーの数と結合プロセス、及びカテゴリーの区分基準を考察する。

2.1 実験

被験者　日本語成人母語話者 38 名。18～69 歳。職業と出身地は幅広くわたっている。

材料　　表 1 に記されている「上」の 22 の例文を材料とした。『大辞林』、『明鏡国語辞典』、『新明解国語辞典』、『国語大辞典』をもとにし、2 つ以上の辞書でエントリーがある用法が 11 ある。各用法の例文を 2 つずつ選んで、カード 1 枚に一文ずつ印刷した。

表 1　カード分類実験に使用する例文

意味	例文	略称
基準とする点より相対的に高い方向，または位置。	(1)煙が**上**へ昇る。（非接触の≪上方≫）	垂直 1
	(2)机の**上**に本がある。（接触の≪上部表面≫）	垂直 2
紙などを人の前に置いた時, その人から遠い方向，または位置。	(3)本文の**上**に頭注をつける。	遠近 1
	(4)一行目の**上**から三字目は何と読むのか。	遠近 2
ある物の表面。また，表面に出る方。外側。	(5)シャツの**上**にセーターを着る。	表裏 1
	(6)墨筆の**上**に朱で訂正を加える。	表裏 2
連続しているものの, 順序が先の部分。	(7)**上**に「ら」のつく言葉を言ってください。	前後 1
	(8)**上**に述べたように、復習は大切です。	前後 2
地位・身分が高い方。	(9)大阪へ出張にいくようにと**上**の方から指令が来た。	地位 1
	(10)**上**の人におべっかを使っている。	地位 2
能力・品質などが優れている方。	(11)日本語の能力は彼のほうが**上**だ。	品質 1
	(12)技術は彼の方が**上**だ。	品質 2
年齢が多い方。年長。	(13)彼女は私より三つ**上**です。	年齢 1
	(14)子供が二人いる。**上**の子は大学生で、下の子は高校生だ。	年齢 2
…である以上は。…であるからには。	(15)こうなった**上**は決行あるのみ。	条件 1
	(16)見られた**上**はしかたがない。	条件 2
…に加えて。…であるところにさらに。	(17)彼は頭がよい**上**に，実行力もある。	添加 1
	(18)叱られた**上**に罰金まで取られた。	添加 2
…したのち。…の結果として。	(19)十分調査した**上**で御返事します。	順序 1
	(20)署名・押印の**上**窓口に提出してください。	順序 2
…という観点からは。…の面では。	(21)理論の**上**ではそうだが，実際はどうか。	分野 1
	(22)数の**上**では圧倒的だ。	分野 2

手続き　　被験者に例文が書かれたカードを提示し、意味の類似性に基づいてグループを分けさせる。

2.2 結果

非計量的多次元尺度解析（Multidimensional Scaling, MDS）と階層的クラスター分析（Hierarchical Cluster Analysis，HCA）を行った。2 枚のカードが同グループに分類された頻度を基に類似性行列を作り、それを距離行列に転換し、MDS を行った。最適なグループ化を行うため、HCA の樹形図によって区分した。

2.2.1 MDS 分析の結果

MDS 分析の 3 次元解は満足のいく適合度（stress=0.17478）を示すため、3 次元解を用いる。

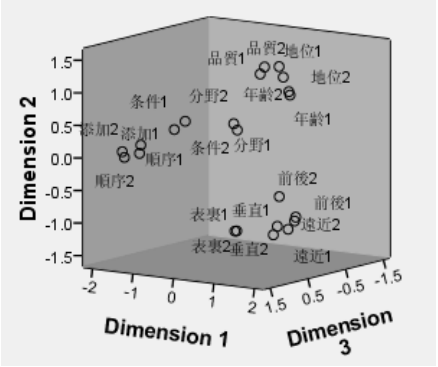


図 2　MDS 分析の結果

2.2.2 HCA 分析の結果

最適なグループ化を行うため、HCA 分析を行った。分

析オプションは ward 法を用いた。

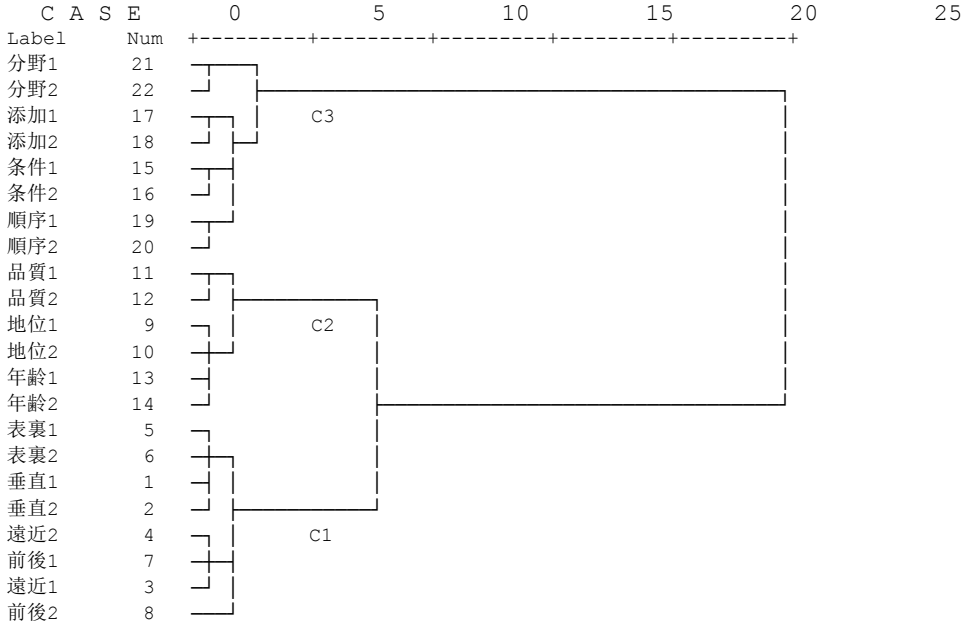


図 3　HCA 分析の結果

2.3　考察

2.3.1　カテゴリーの区分と結合プロセス

HCA 分析の結果によると、「上」の各用法は 3 つのクラスタに分けるのが適当であることが示された。それぞれの成員に基づいて、空間（C1)、程度（C2)、関係（C3）のカテゴリーと名付けた。程度のカテゴリーと空間のカテゴリーが結合してから、より心理的距離の遠い関係のカテゴリーと結合するというプロセスが示される。以下はカテゴリーの構造及び結合プロセスを図示する。

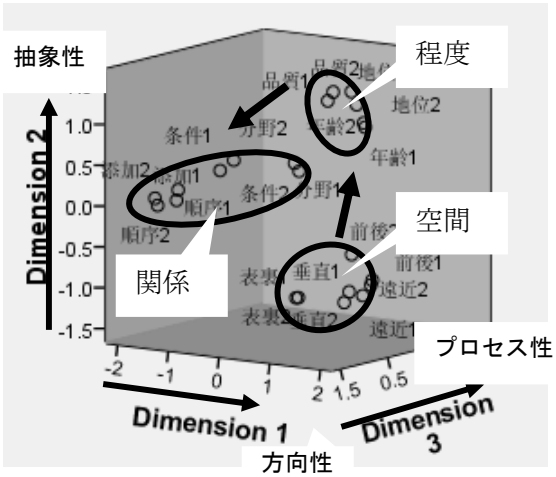


図 4　カテゴリー構造

2.3.2　カテゴリー区分の基準

MDS 分析の空間布置図の次元からカテゴリーをまとめる基準が分かる。MDS の 3 次元解は表 2 の通りである。網をかけている部分はマイナスの値。

表 2　カテゴリーの区分

カテゴリー	用法	次元 1	次元 2	次元 3
空間	垂直 1	1.0175	-1.0444	.0099
	垂直 2	1.0603	-1.1496	.1630
	遠近 1	.9568	-1.1449	-.3412
	遠近 2	1.1500	-.9423	-.3348
	表裏 1	.8283	-.9984	.9317
	表裏 2	.8246	-1.0036	.8768
	前後 1	1.0978	-1.0096	-.3604
	前後 2	-.0205	-.8337	-1.1517
程度	地位 1	1.0212	1.4030	-.0320
	地位 2	1.1114	1.2448	-.0460
	品質 1	.9193	1.3380	.3576
	品質 2	.8735	1.4184	.1945
	年齢 1	1.2125	1.0250	-.0865
	年齢 2	1.1356	.9506	-.1984
関係	条件 1	-1.9557	.1505	-.4487
	条件 2	-1.8801	.2532	-.6556
	添加 1	-1.4563	.1238	1.4619
	添加 2	-1.6725	-.0221	1.1743
	順序 1	-1.9181	.0490	.4787
	順序 2	-1.6580	-.0134	.7902
	分野 1	-1.3009	.0521	-1.4156
	分野 2	-1.3465	.1537	-1.3674

各用法の次元での分布を考察すると、その 3 つの次元とは、「方向性」「抽象性」と「プロセス性」であると考えられる。

(1)「方向性」の次元

「方向性」次元における各用法の分布は次の通りである。

表 3　「方向性」次元の分布

方向性が高い←――――――――――――――――――――――――→方向性が弱い		
程度 （年齢・地位・品質）	空間 （垂直・遠近・表裏・前後）	程度 （分野・添加・順序・条件）

方向性の高いところは上下の指示性が高く、上と下の対照がはっきり見られる。一方、方向性が弱い用法には上下の対照が失われる。

表 4　各用法の「上」と「下」の対照

用法	「上」	「下」
垂直	机の**上**に本がある。	机の**下**に靴がある。
遠近	一行目の**上**から三字目は何と読むのか。	一行目の**下**から三字目は何と読むのか。
表裏	シャツの**上**にセーターを着る。	セーターの**下**にシャツを着ている。
順序	**上**に述べたように、復習は大切です。	具体的な内容は**下**で述べます。
年齢	**上**の子は大学生だ。	**下**の子は高校生だ。
地位	**上**の人におべっかを使っている。	**下**の人がよくサボって困る。
品質	技術は彼の方が**上**だ。	技術は彼よりは**下**だ。
添加	彼は頭がよい**上**に，実行力もある。	
手順	十分調査した**上**で御返事します。	
分野	理論の**上**ではそうだが，実際はどうか。	
場合	見られた**上**はしかたがない。	

(2)「抽象性」の次元

「抽象性」次元における各用法の分布は次の通りである。空間カテゴリーの各用法は具

象的な空間関係を表し、抽象性が弱い。それに対し、抽象性の高い用法はイメージが浮びにくい抽象的なスキーマになる。

表 5　「抽象性」次元の分布

抽象性が高い←――――――――――――――――――――――――――→抽象性が弱い		
程度 （品質・地位・年齢）	関係 （条件・分野・順序・添加）	空間 （表裏・前後・遠近・垂直）

(3)「プロセス性」の次元

「プロセス性」次元における各用法の分布は次の通りである。プロセス性の高い用法は「…に加えて、…したのち」という次の段階に移るプロセスが見える。それに対し、プロセス性の弱い用法は「…において、…で」という「位置」を表す。

表 6　「プロセス性」次元の分布

プロセス性が高い←――――――――――――――――――――――――――→プロセス性が弱い										
添加	表裏	順序	品質	垂直	地位	年齢	遠近	条件	前後	分野

3. プロトタイプ

認知言語学では、多義語の各用法はプロトタイプを中心に拡張してくるのである。先行研究では、「上」プロトタイプ的意味が≪上部空間≫と規定されている（山梨 1988，瀬戸 1995)。さらに非接触の≪上方≫と接触している≪上部表面≫に分けられているとしている研究（厳 2009，徐 2009）もあるが、この 2 つのどちらかとプロトタイプを明示する研究はいまだにない。本稿では、実験を通じて、「上」プロトタイプを更に明示化する。

プロトタイプ的意味は「最も認知的際立ちが高い」用法であると指摘されている（松本 2003)。この認知的際立ちは二つの面から検定できると考えられる。

①反応時間　反応時間の一番短い用法、つまり一番想起されやすい用法がプロトタイプ的意味である。

②連想回数　一定の時間内に一番多く想起される用法がプロトタイプ的意味である。

以上の原理に踏まえ、文の想起実験を行った。

3.1　実験

被験者　日本語を母語とする大学生 33 名。

手続き　被験者に単純語「上」を含む文を思いつく順に 3 つ書いてもらう。

3.2　結果

有効回答率は 93/99（94%）である。最初に想起される文と想起される文全体を用法別に分類すると、結果は次の通りである。右の円は≪垂直≫用法の内訳を示す。

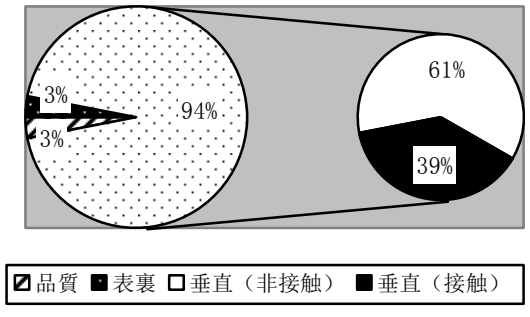


図 5　最初に想起される文の分布

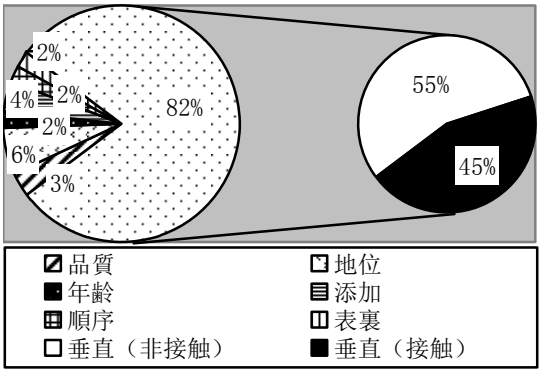


図 6　想起される文全体の分布

図 5、6 をまとめると、次の結果になる。

（1）被験者が最初に想起した文の中で、≪垂直≫用法が一番多い（94%)。その次は≪表裏≫（3%）と≪品質≫（3%）用法である。

（2）被験者が想起したすべての文の中で、≪垂直≫用法が一番多い（82%)。その他、≪地位≫（6%)、≪添加≫（4%)、≪品質≫（3%)、≪表裏≫（2%)、≪順序≫（2%)、≪

年齢≫（2%）など多数ある。
（3）≪垂直≫用法の中で、最初に想起された文の中で、非接触の≪上方≫（61%）が接触の≪上部表面≫（39%）より多い。想起文全体においても、同じ傾向が見られる。つまり、非接触（55%）＞接触（45%）。

3.3 考察

（1）≪垂直≫の用法が一番早く、しかも限られた時間内で一番多く想起されているため、プロトタイプの用法だと認定される。
（2）≪垂直≫用法の中では、≪上方≫用法が≪上部表面≫用法より早く、しかも多く想起されているため、典型度が高いことが示された。
（3）最初に想起される文は≪垂直≫用法に集中しているが、時間的に余裕があれば、他の用法も想起される。典型度の差がはっきり見られる。

4. イメージ・スキーマ

これまでの先行研究はイメージ・スキーマを内省でまとめるため、主観性が入っている点が厳密性に欠けていると思われる。その上、接触のスキーマと非接触のスキーマでは、どちらが主流かについての研究は今まだない。本稿では、イメージ図の想起実験を通じて、母語話者のイメージ・スキーマをまとめる。

被験者　日本語を母語とする大学生 33 名。18~22 歳。出身地は幅広くわたっている。

手続き　被験者に「上」のイメージを見取り図で示させる。

結果　Langacker(1987) によると、イメージ・スキーマが TR（trajector)、LM （landmark）と PATH の三つの部分からなる。イメージ図から TR、LM、PATH の 3 つの要素を抽出し、「上」のイメージ・スキーマは次の 2 種類に抽象化される。

図 7　非接触のイメージ・スキーマ

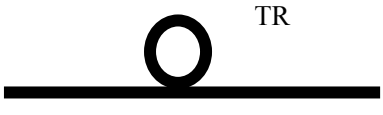


図 8　接触のイメージ・スキーマ

被験者のデータを集計した結果、非接触のイメージ・スキーマは圧倒的に多い（80%）ことが分かる。

考察　「上」には接触と非接触の 2 つのイメージ・スキーマがあり、中では、非接触のほうが優位に立っている。

5. 拡張手段

「上」の各用法はプロトタイプ的用法から、メタファーやメトニミー等の手段を通じて拡張してくる。次の分析は日本人の認知言語学研究者 3 名に確認してもらったものである。

上方→上部表面（メトニミー）——→表裏（メトニミー）—→条件（メタファー）
　　└→分野（メタファー）　　└→順序（メタファー）
　　└→添加（メタファー）

└→遠近（メトニミー）
└→前後（メトニミー）
└→品質（メタファー）
└→年齢（メタファー）
└→地位（メタファー）

図 9　拡張手段

≪上部表面≫　≪垂直≫用法において、≪上方≫がプロトタイプ的用法で、≪上部表面≫は拡張義として捉えられている。「上部表面」は「上方」、つまり「基準点より高い方

向と位置」の一部であるため、全体から部分へのメトニミーである。
≪表裏≫　「外側」の一部である≪上部表面≫から「外側」全体を表すことは部分から全体へのメトニミーである。
≪条件≫　≪表裏≫はモノの「重なる」関係を表し、≪条件≫はコトの「重なる」関係を表す。物のドメインから事のドメインへのメタファーである。
≪順序≫　≪表裏≫はモノの加える順序を表し、≪順序≫はコトをする順序であるため、物のドメインから事のドメインへのメタファーである。
≪添加≫　空間的追加関係から事物の間の関係へのメタファーである。
≪分野≫　「～の面において」はモノの上部表面から事物の「側面」へのメタファーである。
≪遠近≫≪前後≫　垂直軸から水平軸への転換である。しかし、ドメインが変わらないため、部分から部分へのメトニミーである。
≪品質≫≪年齢≫≪地位≫　空間ドメインから品質・年齢・地位ドメインへのメタファーである。

6. 拡張距離

MDS の 3 次元解（表 2）から各用法の心理的距離が計算できる。拡張の起点から終点までの心理的距離はその拡張距離とする。計算結果は表 7 で示す。

表 7　拡張距離

拡張	距離	拡張	距離	拡張	距離	拡張	距離
上方→上部表面	0.1906	表裏→条件	3.3316	上方→遠近	0.3497	上方→地位	2.3693
上部表面→表裏	0.7913	表裏→順序	2.8189	上方→前後	1.9747	上方→年齢	2.0439
上部表面→分野	3.1095	表裏→添加	2.6446	上方→品質	2.4402		

7. 拡張力

一つの用法が言語においてどれほど定着しているかは使用頻度によってある程度分かれる。頻度が高ければ高いほど、この方向への拡張が強いと言えよう。本稿では、均衡コーパスにおける使用頻度で各用法の拡張力を測る。

BCCWJ（2009 版）からランダムに 1500 のテキストを抽出し、ミニコーパスを作った。そのミニコーパスで「上」「うえ」をキーワードにして検索し、2772 の例文を得た。研究対象以外の例文（動詞、複合語、専用名詞などの例）を除き、最終的に 411 例入手した。これらの例文を日本語母語話者 3 名に用法を判断させ、各用法を使用頻度を集計した。各用法の割合は図のようである。

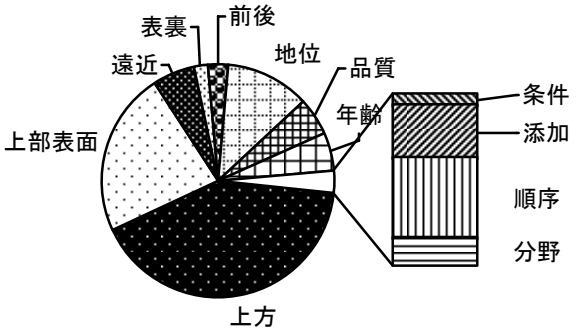
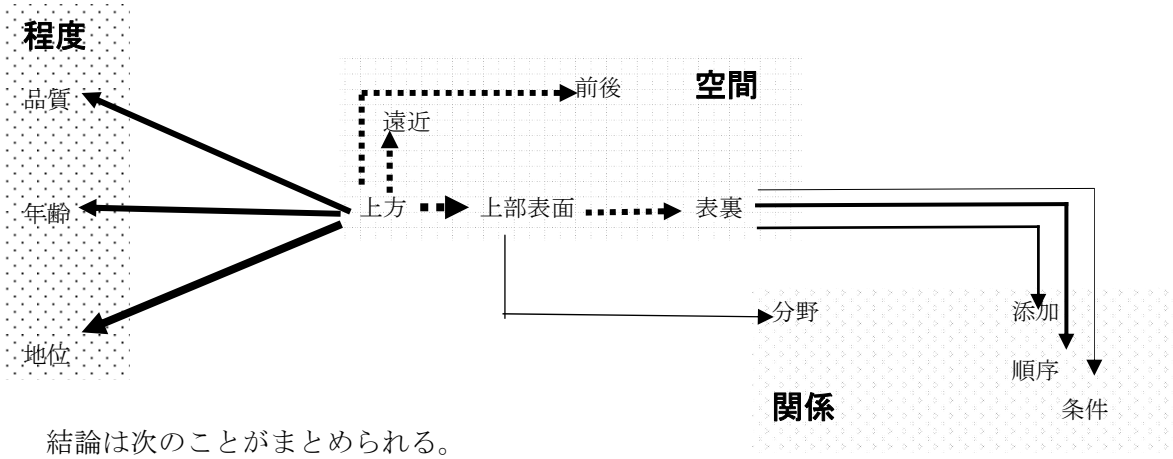


図 10　各用法の拡張力対照図

8. まとめ

以上、2~7 節の考察を通じて、「上」のカテゴリー区分、プロトタイプ、イメージ・スキーマ、拡張手段、拡張距離と拡張力を明らかにした。各要素をまとめると、「上」の意味ネ

ットワークが得られる。



結論は次のことがまとめられる。

（1）「上」の意味は空間、程度と関係の 3 つのカテゴリーに分けられている（図の網かけの部分）。方向性、抽象性とプロセス性の 3 つの基準によって、カテゴリーが区分される。

（2）「上」のプロトタイプは≪垂直≫用法の中の≪上方≫用法である。（拡張の原点）

（3）「上」は 2 つのイメージ・スキーマがある。中では、非接触のイメージ・スキーマは優位に立っている。

（4）「上」の各用法はメタファーとメトニミーによって拡張している。（実線はメタファー、点線はメトニミー）

（5）「上」の各拡張の心理的距離が大分異なり、一番近い拡張は≪上方≫から≪上部表面≫への拡張であり、一番遠い拡張は≪表裏≫から≪条件≫への拡張である。（線の長さ）

（6）「上」の各用法の拡張力にも著しい差異が見られる。一番拡張力の強い用法は≪上部表面≫の用法であり、一番弱いのは≪条件≫である。（線の太さ）

本稿では、より客観的な心理学実験とコーパス考察に基づいて、多義語の意味ネットワークを量的に構築することを試みた。この多義語の実証的研究モデルは、これまでの心的実在性が不明確な多義語の意味構造研究の弱点を補えるものであると考える。認知言語学研究の科学性と緻密性を一層高めるため、心理学・コーパス言語学の研究法を積極的に導入することの意義も実証された。

主な参考文献

1 今井むつみ（1993）「外国語学習者の語彙学習における問題点」『教育心理学研究』(3)
2 厳馥（2009）「日本語と中国語の空間表現に関する対照研究」大阪大学博士学位論文
3 徐莲等(2009) 『言語行為の認知的・語用的研究』大众文艺出版社
4 瀬戸賢一（1995)『空間のレトリック』 海鳴社.
5 田中茂範・松本曜(1997)『空間と移動の表現』研究社出版.
6 松本曜（2003）『認知意味論』大修館書店
7 籾山洋介(2000) 「多義語の複数の意味を統括するモデルと比喩」山梨正明他編『認知言語学論考 NO.1』ひつじ書房.
8 森山新（2008）『認知言語学から見た日本語格助詞の意味構造と習得』ひつじ書房
9 山梨正明（1988）『比喩と理解』東京大学出版社
10 山梨正明(1998)『認知文法論』ひつじ書房.
11 蓝纯(1999)《从认知角度看汉语的空间隐喻》 《外语教学与研究》(4)
12 沈家煊(1999)《不对称与标记论》江西教育出版社
13 Lakoff,George & Mark Johnson.（1980）*Metaphors We Live By* University of Chicago Press.
14 Lakoff,George (1987)*Women,Fire and Dangerous Things:What categories reval about the mind* University of Chicago Press
15 Langacker, Ronald W. (1987), *Foundations of Congnitive Grammar: Theoretical Prerequistites* Stanford University Press

複合辞「というと」「といえば」「といったら」の用法の異同に関する計量的考察

小西いずみ（広島大学教育学研究科）†

Quantitative Analysis of Compound Particles *"to-iuto", "to-ieba", and "to-ittara"*

Izumi Konishi (Graduate School of Education., Hiroshima University)

１．問題の所在

本発表では，「～と言う」のト・バ・タラ条件形に由来する複合辞「というと」「といえば」「といったら」，特に「X｛トイウト/トイエバ/トイッタラ｝Y」において，X がキーワード，Y がキーワードからの連想，という意味関係にあるとされるタイプを中心的な考察対象とする。例えば次のような例である（以下，BCCWJ からの引用例ではファイル名［プレーンテキストファイル，拡張子は省略］を記す。また改行位置に / を記す）。

(1)　ピラミッドの頂点というと、いま進行している大学の大学院化が連想される。
PB13_00729:48

(2)　日本酒の醍醐味といえば、やはり燗酒である。　PB14_00132

(3)　当時は鶏といったら、みんな庭を駆け回っている地鶏ですからね。　PB45_00119

藤田(2000)はこのような「連想のキーワード」をとるトイウト・トイエバについて，「どちらかというと，「～トイエバ」の方が使いやすいような気がする」とする。藤田はトイッタラについては触れていないが，発表者にとっては(1)の文でトイッタラを使うと，発話の事実的な条件としか解釈できない。

森田良行・松木正恵(1989)や日本語記述文法研究会(2009)は，文の題目（主題）を表す形式の一つとして，上のようなトイウト・トイエバ・トイッタラをあげる。また，その中には，次のような「聞き手の発話中のことばを引用して提示，その指示対象を確認・質問」する用法も含められている。

(4)　A「田中さんが結婚するらしいよ」
　　B「田中さんっていうと，あの営業部の切れ者？」（日本語記述文法研究会 2009）

上のような例について日本語記述文法研究会(2009)はトイウト・トイッタラは可，トイエバは不可とするが，発表者の内省ではトイッタラはやや不自然に感じられる。なお，グループ・ジャマシイ(1989)は，この類のトイウトは「確認の表現であり，「NGO というと，何のことですか」というような質問はしにくい」と指摘する。

このような先行研究の記述やそれと発表者の内省との不一致から，トイウト等 3 形式の異同の把握のためには，コーパスの用例の計量的な把握が有効かもしれないと考えた。

†ikonishi@hiroshima-u.ac.jp

２．本発表の目的と方法

以上のような先行研究の記述をうけ，本発表では，ひとまず，複合辞トイウト・トイエバ・トイッタラを広めに認定し，それらの用法の異同を計量的に把握する。その上で，3 形式の意味・用法の異同について，元の引用動詞句条件形「と言うと」「と言えば」「と言ったら」との連続性／異質性を手掛かりに考えたい。

トイウト・トイエバ・トイッタラの用例調査は，次のように行なった。

対象資料　『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）のうち，生産実態(出版)SC の書籍のサンプル（プレーンテキストファイル PB10_00001.txt ほか全 4,459 サンプル）。

方法　上の全 4,459 ファイルから文字列「と言うと」「と言えば」「と言ったら」「というと」「といえば」「といったら」を抽出し，動詞句「～と言う」のト・バ・タラ以外の例，言語形式自体への言及例を除いた。

当該動詞句・複合辞には「トイイマスト」「ッテイエバ」「トイヤ」等の文体的変異が存在するが，それらを含まずとも十分な用例数が得られたため，対象としなかった。なお，コアデータ（書籍以外も含む全データ）を対象とした予備調査では，くだけた文体的変異として，助詞部分がッテの形，「言えば」がイヤとなった形も抽出したが，全 92 例中「ッテイエバ」1 例・「ッテイッタラ」1 例が得られただけであった。

３．調査結果と分析

３.１　用法の分類

上の作業の結果，全 1,978 例（トイウト 857，トイエバ 977，トイッタラ 144）が得られた。ここでは，得られた用例を次のいずれかに分類する。また，以下では「X｛トイウト・トイエバ・トイッタラ｝Y」という文型における X と Y の意味関係を〈x－y〉という記法で表すことがある。

[a] **動詞句**　引用動詞句の条件形（非複合辞）としての用法。〈発話・表現－帰結〉

(5)　「わかりません。いないと言うとすぐ切ってしまいました」　PB39_00148

(6)　「作風がよい」と言えば、中国人も日本人も喜ぶだろう。　PB48_00030

(7)　「湖を見てろ」と言ったら、２時間も３時間も見続けてた　PB37_00007

[b] **問答**　〈問い－答え〉　（藤田(2000)の「自問自答的複文の前件形成」に相当）

(8)　この捻を何に使うかというと、筏を組む際にロープ代わりとして使うのである。　PB16_00127

(9)　ぼく自身はと言えば、最初からマスクもゴム手袋も使わずにいた。　PB59_00686

[c] **想起**　〈キーワード－想起対象事物（ヲ想起スル）〉（藤田(2000)の「連想のキーワード」に相当）

例は，上の(1)～(3)。Y 部分は，(1)のように想起対象事物を項とする心理動詞の場合，(2)(3)のように「想起対象事物＋ダ」の場合など，統語・形態的に性質の違うものが混

じる。詳細は後述する。

[d] **表現属性**　〈言語表現－言語表現としての属性・評価〉

(10)　養女といえば聞こえはよいが、　PB29_00461

[e] **Ｘトイエバ X**　ある事物を X と呼ぶ・表現することへの留保の態度を表す慣用表現。「X トイエバソウダ」などの変種も含む。トイエバのみ。

(11)　プライベートＩＰアドレス自体、もともとＬＡＮの内部で使うために定義されたＩＰアドレスなので、これは当然といえば当然です。　PB2n_00103

[f] **意味確認**　〈先行する発話やその一部－その表意・含意の確認〉（上の(4)に相当）

(12)　「(小西略）ところが、応対に出た奥さんによれば、おといの朝、家を出たまま戻ってこないというんですよ」/「おといの朝？」/「ええ」/「というと、宇賀神さんが殺された朝じゃないですか」　PB19_00294

[g] **恒時条件**　〈条件となる事物－その恒時・即時的帰結〉　（藤田(2000)の「恒時条件句形成」に相当）

(13)　集まるというと、あそこの嫁さんはどうだとか、あの親父はどうだとか、あの細君はこうだとか言い合い、これが子供たちに非常に悪影響を及ぼす。PB51_00005

(14)　しかし、登りくち近くにくるというと、てっぺんは見えず、道だって、けわしくなかった。　PB2n_00046

(14)のように〈恒時〉というより〈即時〉ととれる例があった。[c]とも解釈できる場合は[c]と認定した。

[h] **感嘆**　〈尺度－程度の甚だしさ〉〈事物－その動作・状態の異常性〉という関係。文全体で感嘆表現となる。

(15)　その目立つことといったらない。　PB39_00126

(16)　その沿線といったら、東京じゅうを探すくらい広大に思える。　PB29_00140

３.２　用法の分布

トイウト・トイエバ・トイッタラの [a]～[h] の用例数と%を表 1 に示す。30%以上の場合下線を付した。用法分布の差を下に列挙する。

- トイウトは [b 問答] が 50%強，トイエバは [c 想起] [b 問答] がそれぞれ 40%・30%代であり多い。[a 動詞句] はそれぞれ 2 割未満，1 割未満で，少ない。
- 一方，トイッタラは [a 動詞句] が 60%を超える。
- トイッタラは，[d 表現属性] の占める率がトイウト・トイッタラより高い。
- [e　X トイエバ X] は，先行研究の記述どおりトイエバのみ。
- [f 意味確認] は，先行研究の記述どおりトイエバの確例が見いだせない。トイッタラは 1 例あった。
- [g 恒常条件] は，トイッタラの例がなく，頻度・実数ともにトイウトのほうがトイエバより高い。ただし，3 形式を通じて，この用法が全体に占める割合は少ない。

表 1.　トイウト/トイエバ/トイッタラの用法分布

	a 動詞句	b 問答	c 想起	d 表現属性	e xトイエバx	f 意味確認	g 恒常条件	h 感嘆	計
トイウト	170	481	126	34	0	33	13	0	857
	19.8%	*56.1%*	*14.7%*	*4.0%*	*0.0%*	*3.9%*	*1.5%*	*0.0%*	
トイエバ	79	309	439	66	76	0	6	2	977
	8.1%	*31.6%*	*44.9%*	*6.8%*	*7.8%*	*0.0%*	*0.6%*	*0.2%*	
トイッタラ	91	5	9	22	0	1	0	16	144
	63.2%	*3.5%*	*6.3%*	*15.3%*	*0.0%*	*0.7%*	*0.0%*	*11.1%*	
計	340	795	574	122	76	34	19	18	1978

- [h 感嘆] は，先行研究の記述どおりトイッタラの例が多く見られたが，トイエバの例も確認できた。

[f 意味確認] にトイッタラの例があったのは，日本語記述文法研究会(2009)の例と矛盾しない。ただし，これは，小説の会話文の非共通語的変種による例である。

(17) 「角館の戸沢様の分家の門屋様と言ったら、角館から二里ほど北に行った檜木内川沿いの門屋にある、あの城でっか」 PB39_00695

[g 恒時条件] で，頻度・実例数ともにトイウトがトイエバを上回るのは，藤田(2000)の記述と矛盾しない。トイウト・トイエバともに発表者には不自然に思われる例が混じる。

(18) 姿が見えなかったというとアリバイは、まっ白というわけにはいきませんね。PB19_00288

(19) 私でも、光さんでも同じだけど、よく働きましたよ。今は時化たといえばぶらぶら遊んでいるだけだからね。 PB26_00104

[h 感嘆]のトイエバ 2 例も，発表者には不自然に感じられる。

(20) 大きな社になりますと、稲荷祭の神輿を出すところもありましたが、囃子屋台、踊り、茶番などは町内ごとにあって、初午の騒々しさといえば、また格別でした。PB12_00057: 61

(21) ひどい話ですが、私の小学生時代の悪さといえば、思い出しても寒けがするくらいであります。PB39_00112:6

容認性に個人差のある例が，書籍の書記言語表現としてわずかながらも存在することは，それなりに意味のあることだろう。例えば [h 感嘆] のトイエバの例は，トイッタラの使用が文体的にくだけたものに偏りがちなために，トイッタラ・トイエバ間に類義表現としての一種の類推意識が働いて，文体的に中立なトイエバに置き換えることによって成立したのではないだろうか。

３.３ 「想起」用法の述部による分布

[c 想起] に分類した例の「X｛トイウト/トイエバ/トイッタラ｝Y」における Y 部分には，さまざまな統語・形態構造のものがあるが，先行研究ではこのことに十分注意が払われて

いない。ここでは，Y 部分の統語・形態特徴と X との意味関係から，次のように分類する。

[1] **心理変化を表す動詞述語**　想起対象事物を表す名詞句を項とする心理動詞句。意味的には,〈キーワード X に対して,《ある事物を想起する》という心理変化が起こる〉ことを表す。次の 2 つがある。

[1a] 「－ガ・ヲ・ニ V」　＊例(1)も。

(22) 割れ目噴火といえば、現在のハワイやアイスランドで見られ、灼熱の溶岩が真っ赤なカーテン状になって続く光景が思い浮かぶ。　PB44_00308

[1b] 「－｛ト・ヨウニ｝V」

(23) たまごというと一見、壊れやすいように思われるのですが、　PB32_00050

この類は, (6)のような発話に対する心理変化を主節で述べる「と言うと」文と連続する。

[2] **心理状態を表す動詞述語・名詞述語**　想起対象事物を表す名詞句を含むが，心理変化ではなく心象を一定期間保持する状態を表している。次のタイプがある。

[2a, b]　[1a, b]のテイル形

(24) 日本ではボランティアというと、少しゆとりのある婦人の仕事だと思われているけれども、　PB24_00268

[2c] 「－（ノ・トイウ）N ガアル」（N は「印象」「イメージ」等の心象を表す名詞）

(25) からだにいい食品というと、薬より安心というイメージがあるらしく、さまざまな食品がブームになっています。　PB34_00231

[2d] 「－（ノ・トイウ）N ダ」（N は同上）

(26) これまでの外国人力士といえば、ハワイ出身、そして大型というイメージだったが、ここへきて少し雲行きが変わってきた。　PB25_00251

[3] **想起対象事物を提示する述語**　形態・統語特徴のほか，X と Y との意味関係にも考慮して細分類する。

[3a]　Y が X の属性・状態叙述。状態性述語。

(27) 昭和十六年九月六日といえば、米英に宣戦布告の三か月前です。　PB12_00198

(28) 嵯峨野というと、有名な豆腐店・森嘉があるせいか、豆腐料理を出す店が多い。PB59_00129

[3b]　X と Y が〈類概念－該当物〉という関係にあり,「－ガ挙ゲラレル」「－ガアル」「－ガ多イ」「－ガ有名ダ」など，想起対象事物が言及動作・存在・多寡・代表性等を表す述語の項として示される。

(29) マグロやイカの刺身に合う酒といったら、もう日本酒しかないではないか。PB14_00132

(30) 大阪の桜と言えば、「造幣局の通り抜け」（the passage through the cherry blossoms in the Mint Bureau's grounds

ｉｎ　ｓｐｒｉｎｇ）が有名ですが、　PB12_00047

[3c]　想起対象事物が名詞述語「－ダ」の語基として示される（ダがφやニナルの場合も。表 2-3 ではニナルの場合を内数として示す）。X と Y が〈類概念－該当物〉という関係が多いが，Y が X を象徴する事物の場合もある。　＊例(2)(3)も。

(31)　この瞬間から、「ムネオハウスといえば憲昭さん」となってしまった。　PB23_00783

[3d]　想起対象事態が Y として示されるもので，名詞述語以外。X と Y の結びつきに一般性がない場合も多い。

(32)　砂漠といえば、私たちが子どものときは童謡の『月の沙漠』をよく歌いました。　PB24_00012

(33)　「一九五五年…石神井のある家の二階で、野坂氏と久しぶりに会ったのであります。そこで二人だけで話をしたときに、彼は『きみ、律のスパイの証拠をしらんか』と言いました。『そんなものは知らんですネー。だいいち、彼はスパイなんかじゃありません』と私が答えると、彼は重ねて『ＧＨＱ関係で何か知っていないか。そういう証拠はないだろうか』と言うのです。（略）」と。/　ＧＨＱ関係というと、一九五六年に出版された司法記者団編「法務省」は、日本共産党渉外部長であった当時の伊藤律について次のように指摘した。　PB12_00193

[4] **意味作用を表す動詞述語・名詞述語**　想起対象事物が X の慣習的意味にあたる。

[4a]　「－ヲ V」　（ル形とテイル形のアスペクト対立が稀薄なので両者を区別しない。）

(34)　「大島」といえば大島紬をさし、「米沢」といえば米沢特有の銘仙をさし、「九谷」といえば九谷焼をさす。　PB17_00035

[4b]　「－（ノ・トイウ）{コト・意味}{ダ・ニナル・トスル}」

(35)　二倍以上といえばオーストラリア大陸の半分を上回る面積ということになる。　PB44_00308

[5]　**Y が省略（下の例），または，言いさし**

(36)　「コカ・コーラといえば？」―さて、何が思いついただろう？「さわやか」「炭酸でシュワシュワ」という人がいるだろう。また「アメリカ！」という人もいれば、「大企業」という人もいると思う。　PB53_00138.txt

Y と X の上の分類による 3 形式の用例数を表 2-1～2-3 に示す。これから，以下のことが分かる。

- トイウトは [1] [2] [3] がそれぞれ 25%～40%を占めているが，トイエバは [3] が 80%弱と，用例分布が偏る。トイッタラは例が少ないが [3b, c] のみである。
- トイエバは，[1] に比した [2]，[1a] に比した [1b] が，トイウトより低い。

表 2.　想起用法における用法分布　2-1：全体

	1.心理変化	2.心理状態	3.想起対象	4.意味作用	5.φ等	計
トイウト	49	32	39	4	2	126
	38.9%	*25.4%*	*31.0%*	*3.2%*	*1.6%*	
トイエバ	40	14	349	33	3	439
	9.1%	*3.2%*	*79.5%*	*7.5%*	*0.7%*	
トイッタラ	0	0	9	0	0	9
	0.0%	*0.0%*	*100.0%*	*0.0%*	*0.0%*	

2-2：[1 心理変化]と[2 心理状態]の内訳

	1.心理変化		2.心理状態			
	a. ガ等 V	b. トV	a. 1a テイル	b. 1b テイル	c. (ノ等)N ガアル	c. (ノ等)Nダ
トイウト	28	21	2	15	13	2
トイエバ	34	6	2	7	2	3

2-3：[3 想起対象事物の提示]の内訳

	a.属性・状態	b ーガ存在スル等	c.(該当物)ダ	(内. ニナル)	d.非名詞述語
トイウト	23	2	12	(6)	2
トイエバ	138	54	114	(2)	43
トイッタラ	0	2	7	(1)	0

- トイウトの [3] の内訳をみると，[3a] はそれなりの用例数があるが，[3b～d] で極端に減る。また，[3c] において「ーニナル」の占める割合が高い点に特徴がある。

４．考察および今後の課題

トイウトとトイエバとトイッタラの上のような用例分布の差は，ある程度，構成要素ト・バ・タラの意味に由来すると思われる。特にトイウトの想起用法の Y の分布の偏りは，ト文の意味から説明しやすい。

トは，基本的に前件事態と後件事態の継起的な関係を表す。前田(2009:58)は，トの中心的な意味を「後件が生起する「きっかけ」（因果関係を持つ場合もあれば，合図的なきっかけの場合もある）となる事態あるいは状況を示す」とする。この [c 想起] 用法におけるキーワード X は，まさに「合図的なきっかけ」にあたり，なかでも Y が [1 心理変化を表す動詞述語] の場合は，キーワード X をきっかけとして，継起的に後件事態〈ある事物を想起する〉が生起するという意味関係がある。トイウトの [1] の頻度が高いことは，トイウトがこのようなト条件文の意味特徴を引き継いでいることの証左と言える。また，[3c] の〈類概念ー該当物〉という関係において，Y が名詞述語として示される場合に，「ーダ」よりも「ーニナル」の例が多く見られることも，ト条件文の主節に意志・働きかけの表現が来にくいという特徴（前田 2009:65）と整合する。

一方で，[2 心理状態を表す述語] でもトイウトが安定した使用頻度を保っていることは，

トイウトがすでにト条件文の用法制限を破っていることを意味する。ト条件文で主節が状態性述語の場合は，「家に帰ると，父がテレビを見ている」のように，前件の動作により後件の状態が「発見」されるという意味関係を表す（豊田 1978，前田 2009:79-85）。しかし，トイウト文の [2] の場合，キーワード X によって Y の心理状態が「発見」されるという意味はない。

バ文の中心的な用法は，時間を超えて成り立つ事態どうしの恒常的関係や，仮定とその帰結という関係を表す場合とされるが，そのようなバの特徴が，トイエバの用法分布の偏りとどのように関わるのか，簡単には説明しにくい。しかし，バがトとは異なり，時間的前後関係だけではなく前件 X と後件 Y との因果関係や論理的関係を表すことが，想起用法のトイエバの Y が [3 想起対象事物の提示] に偏ることに関わると思われる。この観点からの考察は今後の課題となる。

トイッタラについては用例が少なかったが，全 9 例が [3b, c]，すなわち，想起対象事物を「ーガ存在スル」や「ーダ」として提示するものであった。内省によっても，[1]のような場合，「と言う」タラの動詞句としての解釈が優先されることが分かる。これはタラが事実的な条件文を表すことと関わると思われるが，トイッタラについては，よりくだけた文体での資料での用例調査も行った上で再考したい。

引用文献

日本語記述文法研究会（編）(2009)『現代日本語文法 5 第 9 部 とりたて 第 10 部 主題』くろしお出版

藤田保幸(2000)『国語引用構文の研究』和泉書院

森田良行・松木正恵(1989)『日本語表現文型』アルク

グループ・ジャマシイ(1989)『教師と学習者のための日本語文型辞典』くろしお出版

前田直子(2009)『日本語の複文　条件文と原因・理由文の記述的研究』くろしお出版

豊田豊子(1978)「発見の「と」」『日本語教育』36 号

関係名詞としての空間的位置表現

西口純代（徳島大学ソシオテクノサイエンス研究部）†

The Spatial Language as Relational Nouns

Sumiyo Nishiguchi (Center for International Cooperation in Engineering Education,
The University of Tokushima)

１．要旨

本論文は日本語の空間的位置表現に関する研究である。英語では on, in, under, between などの前置詞が位置関係を表現するのに対し、日本語では「上」「中」「隣」などの名詞に「で」「に」などの格助詞を付けて表現する。北米の Chickasaw 語でも「下」などの関係名詞が位置関係をあらわすことが報告されている (Lillehaugen and Munro 2006)。本研究では日本語の「上」などの名詞を関係名詞として扱う。「机の上」の「上」は二項述語であり、うち一項は普通名詞句「机」の指示対象によって満たされる。BCCWJ2009 の「Yahoo!知恵袋」中に 160 例見出される関係名詞の用例の多くは物理的位置関係、時間的前後関係、メタファー的拡張の解釈間で語彙的に多義、曖昧である。よって Pustejovsky (1995)の生成語彙意味論に含まれるオントロジー情報が曖昧性解消に有用であることを示す。

２．関係名詞としての空間的位置表現

英語などの言語では in, on, under, between などの前置詞が位置関係をあらわすのに対し、北米の Chickasaw 語では関係名詞が位置関係を表現する (Lillehaugen and Munro 2006)。

(1) chokka' pakna'
house top
“the top of the house (the house’s roof)”

(Lillehaugen and Munro 2006)

Pakna'は「上」を意味する関係名詞で所有格の chokka'（家）に後続する。

日本語においても「中」「上」「下」などの関係名詞が空間位置関係を表現する。

(2) 胸の前で手のひらを合わせて (4179)[1]

(3) 潜水艦の中って快適ですか? (1824)

「前」や「中」は単独では意味的に充足しない関係名詞である。「前」というのは相対的な言葉で「○○の前」というのでないと意味をなさない。その点では関係名詞の代表としてあげられる「息子」が「鈴木の息子」のように必ず誰かの息子でなければならないのと同様で

† nishiguchi@cicee.tokushima-u.ac.jp

1 括弧内の数字はBCCWJ2009 Yahoo! 知恵袋部分をChaKi.NET 1.2βで出力した際の sentence ID。

ある。(2)の「胸の前」では「胸」が関係名詞「前」の項となっている。同様に(3)の「潜水艦」が「中」の項、西山(2003)のいう「パラメータ」である。

２．１　空間表現でない関係名詞

「父」「友達」「敵」などは典型的な関係名詞と呼ばれる。父親は必ず誰かの父親、友人は誰かの友人、敵は誰かの敵としてしか意味をなさないので、father-of、friend-of、enemy-ofという関数もしくは関係を表すと考えられている。

Partee (1997) は John's brother という名詞句の意味の計算方法として、brother という語は関係名詞であるので、's であらわされているジョンとその兄弟との関係は兄弟関係を継承すると考えればよいと提案した。John's brother の場合は二個体の間の関係は兄弟関係以外には考えられないが、John's book の book は普通名詞なのでジョンと本との関係は所有、賃貸、著者などいろいろ可能で、「ジョンの持っている本」「ジョンの書いた本」など様々な解釈が可能である。

西山(2003)は「主役」「上司」などの名詞句を「非飽和名詞」と呼ぶ。どの芝居の主役か、だれの上司かを定めない限り、それらの名詞は外延を定めることができない。「優勝者」「社長」「タイトル」「目的」などはみな非飽和名詞である。

本論文では西山の「非飽和名詞」を関係名詞とみなし、さらに「扇風機の羽根の数」の「数」、「名前」などの名詞も関係名詞とみなす。普通名詞は一項述語、すなわち個体から真理値への関数であるので、関係名詞は個体から個体から真理値への関数への関数である二項述語である。

(4)　a.　[| 名前 |] = λx,y[name-of(y)(x)]

　　b.　[| 犬の名前 |] = λx[name-of(x)(εy.dog(y))]　　(590)

(4a)で「名前」は二項述語であるので「犬」が指示する個体を代入すると一項述語となる。日本語には a, the などの冠詞がないので「犬」を a dog と同様の不定名詞句として扱い、εオペレーターを使って個体を表すようにした。

２．２　関係名詞としての日本語の空間表現

本論文ではさらに普通名詞とみなされてきた空間表現を関係名詞として扱う。例えば「中」「上」「下」は二項述語、「間」はもう一つ項が入る三項述語である。

(5)　a.　[| 上 |]= λx,y[on(y)(x)]

　　b.　[$_{VP}$ [[$_{NP}$ コーヒーの上]に][ミルクを][$_{V}$ 入れる]]　(6320 から)

　　c.　[|　コーヒーの上　|] = λx[on(εy.coffee(y))(x)]

(6)　a.　[| 間 |] = λx,y,z[between(z)(y)(x)]

　　b.　[$_{PP}$[$_{NP}$[$_{NP}$ 歯と歯茎]の間]に[$_{vP}$[$_{NP}$ カスが] [$_{VP}$ たまり]] (2908)

　　c.　[| 歯と歯茎の間 |] = λx[between(εy.gum(y))(εz.tooth(z))(x)]

２．３　物理的、時間的、メタファー的位置関係の曖昧性

下の表 1 は BCCWJ2009 中の「Yahoo!知恵袋」部分における「NP_1-の NP_2」構文を ChaKi.NET 1.2βを使って抽出した 3083 例中の空間表現名詞の分布である。

表 1 BCCWJ2009 の「Yahoo!知恵袋」部分における「NP_1-の NP_2」構文 3083 例中の空間表現名詞の分布

順位	空間表現	総出合計		物理的位置			メタファー			時間関係		
		合計回数	百分率(%)	英訳	合計回数	百分率(%)	英訳	合計回数	百分率(%)	英訳	合計回数	百分率(%)
1	ほう方	54	33.75	toward	6	11.1	rather than	48	88.9			
2	なか中	34	21.25	in	21	61.8	in	13	38.2			
3	あいだ間	10	6.25	between	6	60	among	1	10	during	3	30
4	うえ上	8	5	on	5	62.5	after	1	12.5	after	2	25
5	前	6	3.75	in front of	5	83.3				before	1	16.6
6	した下	6	3.75	under	6	100						
7	上の	6	3.75	above	1	16.6	elder	5	83.3			
8	あと後	4	2.5							after	4	100
9	近く	4	2.5	near	4	100						
10	周り	3	1.875	around	3	100						
11	裏	2	1.25	on the back of	2	100						
12	下の	2	1.25				younger	2	100			
13	隣	2	1.25	next	2	100						
14	中心	1	0.625				center	1	100			
15	裏側	1	0.625	on the backside of	1	100						
16	辺り	1	0.625	around	1	100						
17	あと後の	1	0.625							after	1	100
18	直前	1	0.625	immediately before	1	100						
19	中央	1	0.625	center	1	100						
20	ふち縁	1	0.625	edge	1	100						

21	側(がわ)	1	0.625	side	1	100						
22	下(げ)	1	0.625				low	1	100			
23	左側	1	0.625	left	1	100						
24	真ん中	1	0.625	center	1	100						
25	元(もと)	1	0.625				under	1	100			
26	向こう側	1	0.625	over	1	100						
27	表(おもて)	1	0.625	top	1	100						
28	左右	1	0.625	to the both sides of	1	100						
29	側(そば)	1	0.625	beside	1	100						
30	外(そと)	1	0.625	outside	1	100						
31	後ろ	1	0.625	behind	1	100						
32	横	1	0.625	beside	1	100						
合計		160	100		76			73			11	

表 1 によれば日本語の関係名詞は三種類の読みが可能で意味的に曖昧である。物理的位置、メタファー的位置、時間的順序関係の 3 種類の読みが可能である。例えば「方」は最頻出で全空間表現中 33 パーセントを占めるが、(7a)のような比較文でメタファー的に優劣の順序として用いられる場合がそのうちの 88 パーセントを占め、(7b)のように文字通りの地理的方向を示す場合よりも多い。

(7) a. 中日より阪神の方が強い (2219)

b. 3 猫いますが皆、私の方に来ます (5177)

一方、「前」は地理的前と時間的前で曖昧である。「胸の前」(4179)の「前」は物理的位置を、「出発の前」 (4000) の「前」は時間的順序を表す。

反して「下」は文字通りの場所を表す読みしかない。

同じ漢字を使っていても「後(あと)」は時間的後だけを意味するのに対し「後(うしろ)」は文字通りのうしろのみを指す。「上」に関しても同様に「NP の上」構文の「上(うえ)」は物理的位置を表すのに用いられる。しかし抽象名詞を使って「(抽象名詞) の上」ということはできず、「(抽象名詞) 上」という複合名詞句を構成する。メタファー的上方は「上(じょう)」で、物理的上方は「上(うえ)」で表現するよう使い分けられているようである。

(8) a. ネット上でいろいろ見てたらリウマチの気があるのでは？ (3508)

b. “ネットの上でいろいろ見てたらリウマチの気があるのでは？

(9) a. コーヒーの上に、泡立てたミルク（クリーム）をふんわりと入れた飲み物 (6320)

b. *コーヒー上に泡立てたミルク（クリーム）をふんわり入れた飲み物

３．空間表現の多義性の理論的扱い

３．１形式化

そのように日本語の空間表現は多義であり、関係名詞として扱えるのであるが、実際にはどのように構成性の原理に基づいて意味計算できるのであろうか。以下に多義である空間表現の意味を定義し、「NP_1の NP_2」名詞句の意味が導き出せるようにした。

「前」

(10) a. 物理的前 [|前$_1$|] = λx,y[in-front-of(x)(y)]

b. 時間的前 [|前$_2$|] = λt,t'[before(t)(t')]

(11) a. [|胸の前|] = λy.in-front-of(εx.chest(x))(y)

b. [|出発の前|] = λe'. ∃e[before(time(e))(time(e'))&departure(e)]

「方」

(12) a. 物理的方向 [|方$_1$|] = λx,y[toward(x)(y)]

b. メタファー的方向 [|方$_2$|] = λx,y[to(x)(y)]

(13) a. 茹で上がるまで、ひたすら、鍋の中をかき回していてください。(2423)

b. 焼きっぱなしお菓子のレシピの中では、結構丁寧な説明でわかりやすいです。(2457)

「中」

(14) a. 物理的中 [|中$_1$|] = λx,y.in(x)(y)

b. メタファー的中 [|中$_2$|] = λx,y.among(x)(y)

(15) a. [|鍋の中|] = λy.in(εx.pot(x))(y)

b. [|レシピの中|] = λy.among(εx.recipe(x))(y)

「間」

(16) a. 物理的間 [|間$_1$|] = λx,y,z[between(x)(y)(z)]

b. メタファー的間 [|間$_2$|] = λx,y,z[among(x)(y)(z)]

c. 時間的間 [|間$_3$|] = λt,t'[t'= during(t)]

(17) a. 歯と歯の間辺りが茶色くなるのですか？ (2906)

b. 芸能人の間で流行っているダイエット食品 (427)

c. ここ数ヶ月の間 (3201)

３．２Generative Lexicon 理論による空間表現の曖昧性解消

生成語彙意味論 Generative Lexicon (GL) (Pustejovsky 1995)では語彙情報が豊富である。ゆえに、空間表現の曖昧性解消のための強力な道具となる。GL には意味的な項構造が入る Argument Structure 項構造, Davidsonian 的イベント項が入る Event Structure イベント構造, 付加的な語彙情報が入る Qualia Structure クオリア構造という、３つの部分がある (Davidson 1967)。

クオリア構造には四つの下位範疇をもつ情報が付加されていて、Constitutive (全体―部分関係)、Formal (オントロジー上のカテゴリー、形、色など)、 Telic (目的)、Agentive (源)という４つの下位範疇がある。Formal クオレ（クオリア）にはオントロジー情報が入っている。例えば「コーヒーの上」のコーヒーは(18)のように Formal クオレによれば液体で、その上位概念は物理的個体 a physical entity である。(19)のように「上」は物理的個体を項を項構造において選択する二項述語であるので、コーヒーを項にとり「コーヒーの上」という名詞句に合成また単一化する。関係名詞ともう一つの名詞句との素性照合が位置表現名詞の曖昧性解消に重要な役割を果たしている。

(18) コーヒー

ARGSTR = ARG1 = x: drink
D-ARG1 = y: human
D-E1 = e1: process
QUALIA = FORMAL = liquid(x)
TELIC = drink act(e1, y, x)

(19) 上

ARGSTR = ARG1 = x: physical_object
ARG2 = y: physical_object
D-E1 = e1: state
QUALIA = FORMAL = on(e1, x, y)

(20) コーヒーの上

ARGSTR = ARG1 = x: physical_object
ARG2 = y: coffee
D-E1 = e1: state
QUALIA = FORMAL = on(e1, x, y)

「上」は多義でなく単義的に物理的上方しか意味しないので(19)のようになる。それに対して「間」などの空間表現名詞は多義である。３.１ではそれぞれの意味を列挙したが、語彙意味論においては dot object すなわち二つないしは三つのタイプの融合したタイプであるということになる (Pustejovsky 1995; Carpenter 1992)。

例えば「本」は体積を持つ物体であり、情報でもあるので、物体としての本の上に手を乗せることもできるし(15a)、内容が正しいという意味で(15b)のように言うこともできる。よって物理的物体と情報の合成タイプである。

(21) book_lcp = {physical-object.information, physical-object, information}

(22) a. 本の上に手を乗せた。
b. その本は正しい。

多義である「間」を「間$_1$」「間$_2$」「間$_3$」として３種類の語彙項目を作成することも可能であるが、ここでは Pustejovsky and Anick (1988)、Pustejovsky (1995)の例に倣ってメタエントリーを作成することにする。「間」が多義であるということは、「間」の意味はその物理的、

メタファー的、時間的意味の合成したものであると考えて lexical conceptual paradigm (lcp)語彙概念パラダイムを規定する。「間」の lcp は場所的、時間的、メタファー的概念の直積である。

(23) aida.lcp = {location.time.relation, location.time, location.relation, time.relation, location, time, relation}

ここで問題となるのは、クオリア構造だけでなく、項構造にある意味的項も場所、人間、時間のどれでもあり得るということである。よって項も多重構造とする。

(24) 間

ARGSTR =	ARG1 = x: location_human_time
	ARG2 = y: location_human_time
	ARG3 = z: location_human
QUALIA =	location.time.relation_lcp
	FORMAL = between(x, y)
	FORMAL = among(x, y)
	FORMAL = during(x)

一方「前」は物理的前と時間的前で二義的に曖昧である。

(25) 前

ARGSTR =	ARG1 = x; physobj
	ARG2 = y: physobj
	E1 = e1: state
	ARG3 = E2 = e2: process
	ARG4 = E3= e3: process
QUALIA=	location・time_lcp
	FORMAL = in front of(e1, x, y)
	FORMAL = before(time(e2), time(e3))

(26) 胸

ARGSTR =	ARG1 = x: body_part
	ARG2 = y: animal
QUALIA =	FORMAL = part_of(x,y)

(27) 胸の前

ARGSTR =	1 ARG1 = x; physobj
	2 ARG2 = y: body_part
	E1 = e1: state
QUALIA =	FORMAL = in front of(e1, 1, 2)

(28) 出発

ARGSTR = ARG1= E1= e1: process
ARG2= x: human
D-ARG1 = y: location
E2 = e2: state

QUALIA = process・state_lcp
FORMAL = e1
FORMAL = ¬at(e2, x, y)
AGENTIVE = leave_act(e1, x, y)

(29) 出発の前

ARGSTR = ARG3 = E2 = e2: process
ARG4 = 1 departure
FORMAL = E3 = e3: process

QUALIA= location・time_lcp
FORMAL = before(time(e2), time(1))

４．結論

本研究は BCCWJ2009 の「Yahoo!知恵袋」部分内の「NP_1 の NP_2」構文における空間表現を抽出し、意味的分布を調べた。空間表現は関係名詞として形式化できるが、「方」「中」「間」などの頻出項目はは地理的位置関係、時間的関係、メタファー的拡張の意味の間で多義となってしまう。そこで生成語彙意味論では多重タイプとして一つの語彙項目として扱えることを示した。

文献

Carpenter, Bob (1992) *The Logic of Typed Feature Structures*, Cambridge University Press, Cambridge.

Davidson, Donald (1967) “The Logical Form of Action Sentences,” *The Logic of Decision and Action*. University of Pittsburgh Press, Pittsburgh.

西山佑司(2003). 『日本語名詞句の意味論と語用論』, ひつじ書房.

Partee, Barbara H. (1983, 1997) “Genitives: a Case Study,” *Handbook of Logic and Language*, eds. by J. van Benthem and A. ter Meulen, 464-470, Elsevier, Amsterdam.

Pustejovsky, James (1995) *The Generative Lexicon*, MIT Press, Cambridge.

Pustejovsky, James and Anick (1988) “On the Semantic Interpretation of Nominals,” *Proceedings of COLING-1988*, Budapest.

外来語由来の造語成分「チック」について

村中淑子（桃山学院大学 国際教養学部）†

Japanese Suffix "*chikku*" Derived from English "-tic"

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

１．外来語由来の造語成分について

英語から日本語に取り入れられた外来語のうち、接尾辞を含む語が日本語としてなじみ深くなると、その接尾辞相当部分が独立して、日本語の造語成分として使いこなされ、語が作り出されるようになることがある。たとえば、マヨラー、ゆっくりズム、だべりんぐ、などがそのようにして作られた語である（それぞれ、英語の接尾辞 -er,-ism,-ing が日本語の造語成分として取り入れられている)。

野村雅昭 1984 では、日本語に定着した造語単位として、「〜システム」「〜マン」などを挙げている。

森岡健二 1985 では、多くの日本人が「接尾辞がついている」と意識するであろうものとして、-ist,-er,-ing,-ic,-tion,-ism,-ty,-ous,-ly,-less,-chen [1] のついた語を挙げている。

田辺洋二 1990 では、和製英語における「語尾型」に用いられている英語の接尾辞は、「-al,-an,-ee,-er,-ful,-ing,-ism,-ist,-izer,-less,-logy,-ness,-tic,-topia,-zation」の 15 種であるとしている。

山下喜代 2007 は、国語辞典を資料として作成された「造語成分データベース」をもとに、外来語由来の造語成分を 393 例挙げている（393 例には前部分になるものと後ろ部分になるものの両方を含む)。

本発表では、造語成分「チック」のみにしぼり、その造語力を調べるために、BCCWJ の検索を行なう。原語の英語では、-ic もしくは-tic が、形容詞を作る接辞であるが、日本語における「チック」はナ形容詞を作る造語成分である。[2]

２．BCCWJ の検索結果

BCCWJ2009 を用いて、まず「チック」を検索する。BCCWJ が均衡コーパスであることから、「〜チック」という文字列そのものを数えることにも意義があると考えて、目的の造語成分チックには該当しなさそうなものも全て入れて数えている。

† tmuranaka@andrew.ac.jp

[1] -chen のみドイツ語由来である。

[2] 永田高志 1994 はファンタジックを例に挙げて「「ック」が「和製外来語の接尾辞として定着している」」と述べているが、本発表ではとりあえず「ック」は扱わないことにする。

表１　すべての「〜チック」の検索結果

プラスチック	ロマンチック	エキゾチック	エロチック	ドラマチック	バルチック	オートマチック	アスレチック	その他	計
466	117	40	33	32	29	25	25	88	855
54.5%	13.7%	4.7%	3.9%	3.7%	3.4%	2.9%	2.9%	10.3%	100.0%

このように、語源・意味を無視してすべての「〜チック」を検索すると 855 件あった。このうち約９割を、表中の８語が占めている。この８語のなかでも、ロマンチック・エロチック・ドラマチック・オートマチックの４つは、それぞれロマン・エロ・ドラマ・オートマという語形でもよく使われる。つまり「ロマンチック」は「ロマン＋チック」であると分析的に捉えやすい。これらの語のおかげで、「チック」という造語成分が取り出され、日本語の中で造語力を持つものとして働くようになったのではないかと推測される。

この８語以外の「その他」に含まれる語は、チック（症）の 11 件を除きすべて出現頻度が１ケタである。この「その他」の中に日本語において「チック」が造語力を発揮した結果できたと考えられる語が 25 件あった。以下、サブコーパスごとに年代順に列挙する。[3]

◆「Yahoo!知恵袋」

(1)関東以外にもアドマッチック天国って放送してるのでしょうか？（OC01_00616、2005）

(2)オカルトチックな演出が絶えないから（OC15_01169、2005）

(3)「ナッティー・ドリームランド」というメルヘンチックなちょっと変わった外国の漫画を（OC01_05124、2005）

(4)ビッグフィッシュ＝メルヘンチックなお父さんみなさん、さようなら（OC01_02105、2005）

(5)知恵袋の回答ボタンとブラウザの相性Ｌｕｎａｓｃａｐｅだとｆｕｎｃｔｉｏｎ　ｃｈｅｃｋｆｏｒｍが面倒チックです。一度使用すると同じタブでは効かなくなる（OC02_00359、2005）

(6) 他県の方が福井弁を真似しようとしたときに、必ず東北弁チックになるのが、すごーく気になります。（OC12_02032、2005）

(7)バラの香りがついたガムがあったのですが覚えてる方いますか？？？確か箱に入ってた記憶があります。子供ながらに高級チックで嬉しかったんですが、、、今でも売ってたら、噛みたいです。（OC08_03979、2005）

◆「国会会議録」

(8)そして、この教祖である文鮮明という男も、経歴を見ると、もうスキャンダラスチックな行動で埋められているような人物です。（OM55_00004、1998）

(9)これは合併強要路線に対する当事者たる市町村長の痛烈なやっぱり批判でもあるわけで、今どうも片山さん（※国務大臣片山虎之助）、余りムネオチックな話にならぬように、具体的なこうした市町村長のこういう声というものも対案だというふうに私は思うわけですが、最後にもう一度大臣の見解を伺っておきたいと思います。（OM65_00003、2002）

[3] サブコーパス「白書」と「書籍（ベストセラー）」には、「〜チック」の造語力が働いたと思われる例は無かった。

◆「書籍（図書館）」
(10)あーん、バッカねえもう。最悪じゃないの。どうしてそういう物語チックなことしちゃうのよ」（LBd9_00009、1989、清水義範『学問ノススメ』）
(11)あたしは、ジーパンにトレーナーという、ラフでボーイッシュな格好。ほんとうは、げろげろに少女趣味の服なんて着てみたいんだけど、なにせ、にあわないのよねー。一方、イーダは中等部のときから、乙女チック・ファッションできめてる。今日は初夏の空の色をした、ふんわりワンピース。じつに似合ってるんだ。これが。（LBf9_00082、1991 森奈津子『いつでもこの世は大霊界！』）
(12)陸奥Ａ子の魅力は単に乙女チックなところだけではありません。（LBf0_00001、1991 大塚英志『少女雑誌論』）
(13)昭和四〇年代末という時代を少女文化の視点からながめていくと、様々な〈かわいいもの〉が登場したメルヘンチックな時代と錯覚してしまう。（LBf0_00001、1991 大塚英志『少女雑誌論』）
(14)そんな不可思議なことが起こるんだったら、もっとミステリアスで芸術チックな場所で起こるのが相場ってものよねぇ（LBg9_00005、1992 東野司『消えた十二支の謎』）
(15)そのまま雪に埋もれて死ねたら、もっといいかもしれない…などと年がいもなく乙女チックな空想を楽しみながら、午後三時には事務所も閉めてしまった。（LBk9_00224、1996 山崎洋子『熟れすぎた林檎』）
(16)自民党の人々は国政から市議会レベルまで見事に小太りタヌキ顔をしているし、公明党のおとっつあんが全員揃って下ぶくれ油照りコテコテ髪というのもなにかオカルトチックで不思議な気がする。（LBk9_00096、1996 椎名誠『おろかな日々』）
(17)私が知ってる範囲では、『オール読物』『すばる』『文学界』『文芸』『小説現代』などで、年に何回か新人賞を募集しているし、もっとポルノチックな耽美小説ということなら『問題小説』『小説クラブ』『小説ノン』などでも同じような賞を出しているから、狙ってみるのもいいだろう。（LBl9_00013、1997 丸茂ジュン『耽美小説の書き方』）
(18)百歩譲ってそうとしましょう。ずいぶん変態チックな遊びみたいですけど（LBn9_00225、1999 高里椎奈『黄色い目をした猫の幸せ』）
(19)こうなると、音楽は「悟り」に近いものであることが充分に認識されるだろう。その内容が過度にメルヘンチックなのは、さておいても。（LBp7_00034、2001 鈴木淳史『クラシック批評こてんぱん』）
(20)ハムスターのおうちは、メルヘンチックなもの、便利なもの、ハムスターの健康を考えたものなどさまざまです。なかでも、メルヘンチックなものは、レンガ造り風のものや、おとぎ話に出てくるようなものなど、それだけでも室内の小物としても使えそうなものがあります。（LBq6_00025、2002 有限会社グラスウインド『ルームメートはハムスター』）
(21)有線から流れるクリスマスソング、子供のようにはしゃぐ妻、御伽の国さながらのメルヘンチックな室内、美しく装う街並み―幸福を絵に描いたようなシチュエーションが、志村を暗鬱にさせた。（LBr9_00135、2003 新堂冬樹『鬼子』）

◆「書籍（出版）」[4]

(22)政党によってはおやじ風のだじゃれやマンガチックな情景など、苦心の跡を見せた。（PB26_00141、2002 藤竹暁『ワイドショー政治は日本を救えるか』）

(23)第一場　東北の、ある片田舎 むかし、むかし、大正の中期 清純素朴でメルヘン・チックな音楽で幕上がる。 舞台は暗幕や照明でエプロン舞台の趣。（PB20_00042、2002 吉村外茂二『ランドアルツト日誌』）

(24)「女はプロセスに酔いたいところもあるんです。一人の男性に優しく抱かれるまでのプロセスをメルヘンチックに想像して、ほのぼのとした気持ちに浸る…、それが危ないムードを高めていってくれるんです」「何となく分かるな」（PB59_00215、2005 末廣圭『華の舞い』）

また、ひらがな表記「ちっく」も検索した。造語成分として働いていたのは 4 件。

(25)春彦が話していたとおり手作りのアニマルちっくな服を着ている和樹を見て、（略）。黒豹がモチーフになった手の込んだ可愛らしい服

（松岡裕太 2003『キミに可愛がられたい！』PB39_00445）

(26)ノーラさんみたいな箱入りお嬢様ちっくな人が

（鏡貴也 2004『ロマンチックはままならない』PB49_00748）

(27)東京 23 区内の、下町ちっくな場所　（Yahoo!知恵袋 2005　OC03_01054）

(28)私は文学部卒で、文学部ちっくな科目しか履修していないので

（Yahoo!知恵袋 2005　OC04_00420）

３．考察

「チック」「ちっく」の大部分は、「的」「のような」「らしい」で置き換え可能であるが、(5)「面倒チックです」と(8)「スキャンダラスチックな行動」(25)「アニマルちっくな服」はいずれも、「的」「のような」「らしい」で置き換えることができない。(5)(8)は「チック」が余剰的であり、(25)はもっと複雑である。

また、「〜チック」はややくだけた子供っぽい印象があるが、例文を見ていくと、様々な文体で使われうることが分かる。

関連して「ティック」を検索したが、既成の語ばかりであり、日本語における造語成分として働いて語を作っているものは無かった。

文献

田辺洋二 1990「和製英語の形態分類」『早稲田大学日本語研究教育センター紀要』2

永田高志 1994「和製外来語の複合語」『近畿大学文芸学部論集 文学・芸術・文化』6-1

野村雅昭 1984「語種と造語力」『日本語学』3-9

森岡健二 1985「外来語の派生語彙」『日本語学』4-9

山下喜代 2007「現代日本語の語構成要素—外来語を中心にして—」『青山学院大学文学部紀要』48

[4] ほかに、「健チック」が 1 件検索されたが、おそらく「建築」の間違いだと思われる。「『営造法式』では、三等材による間口三間の健チックは庁堂ではなく殿堂である。」（PB42_00183、2004 川端俊一郎『法隆寺のものさし』）

コーパスを用いて新語を調べる　―「スルー」を材料に―

村中淑子（桃山学院大学 国際教養学部）†

Examining a New Word with a Corpus : The Case of a Loan Word " *suruu* "(through)

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

１．はじめに

BCCWJ を用いて、新語の出現状況を調べる。

ある新しい語がいつから用いられ始め、どのように広がっていったのか、ということについて、コーパスを用いるとどのくらいの情報が得られるのか。BCCWJ が「均衡コーパス」であることから考えると、ある語の、現代日本語におけるおおよその位置づけはわかるのではないか。どの程度それがわかるのか、具体的に知ろうというのが本発表の目的である。

BCCWJ2009 に収録されたサンプルの内訳と、その年代は次の通りである。

表１　BCCWJ2009 収録サンプルの内訳（丸山 2009 の表 1 に項目追加→網掛け部分）

サブコーパス名	メディア	収録サンプル数	年代	年代の幅
生産実態（出版）サブコーパス	「書籍」	4459	2001～2005	5 年間
流通実態（図書館）サブコーパス	「書籍」	5110	1986～2005	20 年間
非母集団（特定目的）サブコーパス	「ベストセラー」	854	1976～2005	30 年間
	「白書」	1500	1976～2005	30 年間
	「Yahoo! 知恵袋」	45725	2004.11～2005.10	1 年間
	「国会会議録」	159	1976～2005	30 年間
	計	57807		

表 1 から分かるように、6 種類のサブコーパスのうち 1 つは 20 年間、3 つは 30 年間にわたる言語データから抽出されたコーパスであるので、最近の 20 年間あるいは 30 年間の中のどこかの時点で使われ始めたことばであれば、BCCWJ を検索することによって、使われ始めた時期および使用状況の一端が分かることが期待される。

本発表では、主に「スルー」を材料として調べてみることにする。

２．「スルー」について

「スルー」は、英語の through が外来語として日本語に入ってきたものである。ここで、辞書類における「スルー」の記述を確認しておく。

中規模国語辞書類およびカタカナ語辞書における「スルー」について調べたのが表２で

† tmuranaka@andrew.ac.jp

ある。単独で項目として挙げられているかどうか、複合語はあるかどうか、に注目した。[1]

表２　国語辞書等における「スルー」の記述

	スルー単独での項目化	複合語（スルーが前部要素）	複合語（スルーが後部要素）
『大辞泉　増補・新装版』第１版第３刷　1998	○	スルーチェック、スルーパス	※[2]
『広辞苑　第５版』1998	×	無し	シースルー、ドライブスルー、フォロースルー[3]
国立国語研究所編 2004『分類語彙表 増補改訂版』	×	スルーパス	ドライブスルー[4]
『コンサイスカタカナ語辞典 第３版』2005、あるいは『コンサイスカタカナ語辞典　第４版』2010[5]	○	スルーウェー、スルーストリート、スルーチェック、スルーチェックイン、スルーチケット、スルートレイン、スルーパス、スループット、スルーポケット	※[6]
『大辞林』第３版　2006	×	スルーパス、 スループット	シースルー、ドライブスルー、トランクスルー、フォロースルー、ブレークスルー、ランスルー[7]
『広辞苑　第６版』2008	×	スルーパス、 スループット	※[8]

「スルー」単独で項目化されているものについて、その語釈を見てみる。

『大辞泉』の語釈は、次の２つである。

①テニスで、ボールがネットの網の目を通り抜けて相手方のコートに落ちること。

②多く複合語の形で用い、通り抜けの、素通しの、の意を表す。「ドライブ―」「フォロー―」「シー―」

『コンサイスカタカナ語辞典』の語釈は、次の２つである。

①〔サッカー〕パスを受けるような構えをしながらボールには触れず、次の味方選手に渡すもの。〈昭〉②〔テニス〕ボールがネットの網目を抜けて相手方のコートに落ちること。〈昭〉

つまり、少なくとも現在刊行されている一般的な紙の辞書類によれば、「スルー」が単独

1 スループットとブレイクスルーは、英語では一語であるが、ここでは同列に扱っておく。
2 『大辞泉』の逆引きは出ていない。網羅的検索ができなかった。
3 『逆引き広辞苑　第５版対応』1999 による。
4 『分類語彙表 増補改訂版』データベースで検索した。
5 『コンサイスカタカナ語辞典』第３版と第４版の記述は、「スルー」に関しては、項目の立て方・語釈ともに同じであった。
6 『コンサイスカタカナ語辞典』の逆引きは出ていない。網羅的検索ができなかった。
7 『漢字引き・逆引き大辞林』1997 による。
8 『広辞苑　第６版』の逆引きは出ていない。網羅的検索ができなかった。

で語として使われることは、無いか、もしあるとしてもスポーツ用語の分野である、とされているのである。(大型日本語辞書である小学館 2001『日本国語大辞典　第２版』では、「スルー」は単独の項目としては立てられていない)。

いっぽう、ネット上の辞書『デジタル大辞泉』をみると、単独語としてのスルーが載せられており、次の５つの語釈があった。(小学館提供『デジタル大辞泉』、2010 年 7 月現在約 24 万 2,000 語を収録、http://dictionary.goo.ne.jp/leaf/jn2/120524/m0u/スルー/（2011.1.22 閲覧))

①テニスで、ボールがネットの網の目を通り抜けて相手方のコートに落ちること。②複合語の形で用い、通り抜けの、素通しの、の意を表す。「ドライブ―」「フォロー―」「シー―」③球技のパスワークで、一人飛ばしてパスをすること。「味方選手を―してゴール前のフォワードにボールを送る」④そのまま通過すること。「その信号は―して次の交差点を右折してください」⑤俗に、受け流すこと。何もしないで待っていること。「興味のない方はどうぞ―してください」

この『デジタル大辞泉』の語釈①と②は、上記の『大辞泉　増補新装版』とほぼ同一である。新しく追加されたらしい語釈③④⑤のうち、③はスポーツ用語であるが、④⑤は領域を制限しない一般用語としての使用が示されているとみられる。

以上の辞書記述から、次のようにまとめることができるだろう。「スルー」は、英語から日本語に取り入れられて以来、スポーツ用語以外の一般用語としては、単独で語として使用されることは無く、外来語複合名詞の構成要素として働いていた。しかしごく最近になって、一般用語における単独の語としての使用が、「スルーする」というサ変動詞の形で見られるようになった。紙の辞書には、2008 年あるいは 2010 年というごく最近の刊行であっても、その用法が記載されていないのは、俗語的・流行語的で定着度が低いと辞書編集者が考えているからであろうか。

このサ変動詞「スルーする」を新語と見なし、BCCWJ を用いて調べる。

３．BCCWJ における「スルー」

３.１　単純語「スルー」の出現頻度

BCCWJ2009 で「スルー」を検索した結果を、サブコーパスごとに示した。[9]

表３　BCCWJ2009 における「スルー」の検索結果

サブコーパス（メディア名）	年代	スルー（単純語）	スルー（複合語）	スルー計
「書籍」(出版)	2001～2005	1	35	36
「書籍」(図書館)	1986～2005	0	15	15
「ベストセラー」	1976～2005	0	1	1
「白書」	1976～2005	0	11	11
「Yahoo! 知恵袋」	2004.11～2005.10	22	5	27
「国会会議録」	1976～2005	0	4	4
計		23	71	94

[9] スルーを検索後、サンプルを一つずつ確認し、いわゆるゴミを取り除いた。取り除いたものは、バスルーム 121 件、カールスルーエ（もしくはカルルスルーエ）30 件、など計 187 件。３拍という短いカナ文字列の場合、ゴミが多く引っかかってしまうのであろう。

表３から分かるように、検索された「スルー」は 94 件あり、そのうち約４分の１の 23 件が単純語であった[10]。単純語の「スルー」は、１件が「書籍（出版）コーパス」（2005 年刊の小説）に見られた以外はすべて「Yahoo!知恵袋」（2004.11～2005.10）に出現していた。

また、BCCWJ2009 と共通部分の多い「現代日本語書き言葉均衡コーパス」の検索デモンストレーション版でも「スルー」を検索してみたところ、表４の通りであった。[11]

表４　「現代日本語書き言葉均衡コーパス」検索デモ版による検索結果
（BCCWJ2009 と収録サンプルが同一であると思われる部分に網掛けを付した）[12]

サンプリング対象（全て無作為抽出）	収録サンプル数	収録語数	スルー（単純語）	スルー（複合語）	スルー計
一般書籍	8821	2500 万語	1	59	60
政府刊行白書	1500	500 万語	0	11	11
過去 30 年間の国会会議録	159	500 万語	0	4	4
2005 年度版の検定教科書	412	100 万語	0	2	2
「Yahoo! 知恵袋」	45725	500 万語	22	5	27
「Yahoo! ブログ」	24027	500 万語	104	24	128
計	57807	4600 万語	127	105	232

表４から分かるように、「現代日本語書き言葉均衡コーパス」検索デモ版で検索すると、単純語の「スルー」は 127 件ある。このうち、一般書籍の１件と「Yahoo!知恵袋」の 22 件は、BCCWJ2009 で検索されたものと同一であった。また、「Yahoo!ブログ」の検索結果を見ると、出現年は 2008 年および 2009 年である。

以上のことから、次のように言えるであろう。

2005 年までの、一般書籍・政府刊行白書・国会会議録・検定教科書では、「スルー」は複合語の構成要素として使用されており、単純語としてはほぼ使われていない。

ところが 2004 年・2005 年の「Yahoo!知恵袋」では、「スルー」の単純語としての使用が、複合語の構成要素としての使用の、４倍を超えている。2008 年・2009 年の「Yahoo!ブログ」においても、単純語の「スルー」は複合語構成要素としての「スルー」の４倍以上である。また、2004 年・2005 年の「Yahoo!知恵袋」と 2008 年・2009 年の「Yahoo!ブログ」はともに約 500 万語収録のコーパスであるが、単純語の「スルー」は、後者が前者の約５倍になっている。

すなわち、各コーパスの収録語の文体的性質を無視して出現年代のみを考えれば、サ変動詞「スルーする」は、2005 年ごろまではほとんど使用されていなかったが、2005 年前後から使われ始め、その後急速に使用が増えているようだ、といえる。

---

10 「スルーして」「スルーされた」などのサ変動詞の形で検索すると 11 件であるが（「後文脈」に^[さしすせ]を入れる）、「スルーだった」「スルー？」などもサ変動詞のバリエーションであると考え、「スルー」をすべて検索したあとで、ごみ／複合語／単純語に仕分けした。

11 検索結果を確認したところ、「バスルーム」等のごみが 177 件あった。

12 網掛け部分は、サンプリング対象と収録サンプル数が BCCWJ2009 と同じであることから、同一サンプル集団であると推測される。

３.２　単純語「スルー」の意味・用法

BCCWJ2009 における単純語としての「スルー」23 件の意味用法を検討する。[13]

23 件のうち、専門語的要素の強いものが 5 件ある。「Yahoo! 知恵袋」の「ギャンブル」カテゴリに入っているもので、パチスロ用語と見なせるものである。

残り 18 件のうち、「スルー」という動作を受ける対象に人的要素が無く完全にモノであるのは 3 件。3 件のうち 1 件は、「バゲッジスルー」もしくは「スルーバゲッジ」の省略語としての「スルー」である。

(1)荷物は海外到着までスルーにしてしまい、身軽に行動すれば間に合います。(OC13_00104)

3 件の内 2 件は、通信関係の領域で信号類が「さえぎられることなく、そのまま通過する」「(通常は引っかかるはずのものが何らかの条件の下に) 引っかかることなく通り過ぎる」という意味である。

(2)来るべき東京直下型地震に備えて設置された公衆電話で、どんなに回線が混雑していても発信規制をスルーできる。　(PB59_00500　歌野晶午 2005『女王様と私』)

(3)DVD のコピーガード信号がスルーされてダビングできる。(OC02_04621)

残りの 15 件は、「スルー」という動作を受ける対象が人であるか、もしくはモノであっても背後に存在する人が意識されているものである。

意味としては、まず「(不快な存在を) 無視する」がある。(4) (5)には「無視」「シカト」といった類義語で言い換えている部分が前後の文脈にあり、「無視する」「その相手に対する反応をわざとゼロにする」という意味であることがわかる（以下、下線は本発表の筆者が付したものである)。

(4)こうやって反応する価値すらありません。人が本気で怒れば怒るほど嬉しくてしょうがない。反応があればあるほど自分が認められた気分になる。私は、この手の煽り、釣ネタは、掲示板を見なれているので完全にスルーしています。特に知恵袋は言いっぱなしですから、下劣な質問は登場しほうだいです。お怒りは１００％理解できますが、こういった相手に対する最大の反応は完全無視。論評を求める行為自体、相手の思うつぼです。(OC05_00758)

(5)数ヶ月にもわたり、同じような暴言など程度の低い言葉を書き込んできているようです。言葉を見ればその者の仕業だとすぐにわかります。そのような場合は、やっぱりそこを見ずにシカト・無視で宜しいでしょうか？　スルーが一番いいと思います。なまじ反応してしまうともっとひどい事を書いてくると思います。(OC14_01962)

(6)の「スルーしてください」は「無視してください」で置き換えられるが、不快なものとして扱われることを想定しているのではなく、ただ「反応してほしくない」「(何らかの理由・条件の下に) 相手の反応を求めていない」という意味である。

(6) （略）だと思いますか？（略）ちょっと興味があります。私は（略）かな？と思って

[13] 「Yahoo!ブログ」に現れた単純語「スルー」104 例については、「現代日本語書き言葉均衡コーパス」検索デモ版による検索の、前文脈と後文脈が 20 字ずつと少なく、意味の特定のしにくい例があるため、ここでは扱わない。

ます。興味ない方はスルーしてください。「知らない・・・」みたいな回答はやめてください。(OC06_01849)

(7)は、「反応ゼロ」の意味ではない。下線部分でそれが明らかである。つまり、「ことさらに強い反応をせず、通り一遍の反応だけで済ませて通り過ぎる」「特段の反応をせずにそのことを終わらせる」という意味で使われているようである。
(7) ご希望に添えず、残念です。でほって置いていいんじゃないですか。遠回しに返品を要求しているかもしれないし。あっさり返事して、スルーが一番です。(OC14_01236)

(8)は、相手に対してどのように反応するかではなく（下線部分から、既に反応は済んでいることがわかる)、当該事項を「忘れる」という意味で使われているようである。
(8)それはせっかちな落札者の方ですね！そんなにせっかちだと、いずれ何処かでトラブルをおこすと思います。出品者の方の対応は普通だと思います。わたしもいつも発送の連絡と一緒に、入金確認の連絡をしています。発送が２～３日後になるときは、発送予定と一緒に入金確認の連絡をしています。そんな変な落札者はスルーして、またオークションを愉しみましょう！！(OC14_02361)

(9)は、下線部分の「忘れた頃に反応が返ってくる（登録してもらえる）こともありうる」という内容と、その後ろの「そのままスルー」という内容が、対比的に扱われているので、「スルー」が、ずっと反応が返ってこないままの状態が続く、すなわち「忘れられる」という意味で使われていることがわかる。
(9) ｙａｈｏｏは申請したからといってすぐに登録してもらえるとは限らないわ。忘れた頃に、って事もあればそのままスルーなんて事も。。(OC02_00252)

(10)は、「反応が期待されるところで反応がない状態のまま、そのことが終わりになる」という意味で使われている。「見逃す」「見落とす」とも置き換え可能である。
(10) 郵便局の窓口の係は予め貼っていった切手の合計額をちゃんと見ているのでしょうか？どうしても１０円分足りなかったので他に発送もあるので窓口で追加分払おうと思っていました。（送料３９０円に対して３８０円分貼りつけ）そしたらそのままスルーされそうになったので、指摘しました。こんな場合、１０円不足で受取人に迷惑かかりますよね？
(OC14_00983)

以上をまとめると、単純語としてのサ変動詞「スルー」には、およそ次の４つの意味があると言ってよいだろう。
①（ものが）関所あるいは網のようなものにさえぎられることなく、そのまま通過する。
②（不快な存在を）無視する。相手に対する反応を意識的にゼロにする。
③ことさらに強い反応を示さず、そのことを終わりにする。通り過ぎる。
④反応がない状態が続く。忘れる。見落とす。

４つの意味に共通するのは「通り過ぎる」「あるところ・ひと・ものに、ものあるいは意識を留まらせない」といった内容である。④の意味を、③の意味の受動的な状態と捉えて、③と④を一つにまとめることも可能であろう。

このような意味をもつサ変動詞「スルーする」がなぜ一般用語として使われるようになったかについては、次の３つの要因が考えられる。

（１）スポーツ用語の、一般用語への転用。とくにサッカー用語からの流入。

（２）なじみの深い複合語からの流入。とくに、J スルーカードやスルーチェックイン、スルーバゲッジなどの交通・旅行用語。

（３）パチスロ用語からの流入。

４．他のコーパスによる検索結果

BCCWJ 以外のコーパスでもサ変動詞としての「スルー」を検索し、BCCWJ の結果と並べて示したのが表 5 である。比較参考のため、類義語と思われる「シカト」「無視」「流す」「聞き流す」「受け流す」も検索した。

表５　いくつかのコーパスにおける「スルー」と類義語の検索結果

コーパス名	データ採集年	レコード数	スルー（単純語／全数）	シカト[14]	無視	流す[15]	聞き流す・受け流す
『女性のことば・職場編』	1993	11421	0/0	0	0	0	0・0
『男性のことば・職場編』	1999～2000	11099	0/0	0	1	0	0・0
『名大会話コーパス』	2001～2003		0/0	0	19	5	1・0
BCCWJ2009	1976～2005		23/94	21	1916	※	91・50
均衡コーパス検索デモ版	～2009		127/232	21	1899	※	93・55

表 5 から、やはり 2003 年以前には「スルー」は単純語として使用されていないことが分かる。ただ、『女性のことば・職場編』『男性のことば・職場編』においては、類義語の使用もほとんど無いことから、職場という場面においては、そもそもその意味の語を使う文脈自体が稀であったとも考えられる。『名大会話コーパス』については、「無視」や「(同義語としての）流す」が見られるので、「スルー」が使われていない理由は場面的制限ではない、と考えてもよいだろう。『名大会話コーパス』の話者には大学生が多く、場面としては友人同士の気軽な雑談がほとんどであることから、「スルー」が使われていないのは流行語や俗語を避けているからではなく、「スルー」が話者のボキャブラリーにその当時は入っていなかったからであると考えられる。

---

14 シカトについては、カタカナ表記のみ検索した。BCCWJ2009 と「均衡コーパス（現代日本語書き言葉均衡コーパス）検索デモ版」の検索結果がともに 21 件であるが、サンプルを確認すると 18 件が同一で、3 件は異なるものである。

15 「流す」は、「受け流す」と同義で使われている場合のみの数を数えている。BCCWJ と「均衡コーパス検索デモ版」では「流す」の数がそれぞれ 3397、3544 と膨大であるため「受け流す」と同義であるものを数えていない。

５．おわりに

「スルー」という語を材料として BCCWJ の検索を行ない、2000 年以降に現れた新語であれば、BCCWJ を用いて、その使用の始まり、および意味用法の広まりを知ることができそうであるという感触を得た。

しかし、問題点も無い訳ではない。今回、調査の材料とした「スルー」という語が BCCWJ のサブコーパス「Yahoo!知恵袋」（および「Yahoo!ブログ」）にみられたことを、時代変化の現れであると解釈した。しかし、「Yahoo!知恵袋」「Yahoo!ブログ」はいずれも音声言語ではなく書記言語であること、しかもネット上の言語であることを重く見るならば、「スルー」が多く現れたのは 2005 年以後の新語だからではなくネット言語であるから、という解釈も成り立ちうる。今後、他の語について、あるいは他のコーパス等についても調べ、時代的要素と場面的要素について検証していきたい。

文献

現代日本語研究会編(1997)『女性のことば・職場編』ひつじ書房

現代日本語研究会編(2002)『男性のことば・職場編』ひつじ書房

丸山岳彦(2009).「サンプリング方法について」『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）解説文書

岩波書店(1998)『広辞苑　第５版』

岩波書店(1999)『逆引き広辞苑　第５版対応』

岩波書店(2008)『広辞苑　第６版』

国立国語研究所編(2004)『分類語彙表 増補改訂版』データベース

三省堂(1997)『漢字引き・逆引き大辞林』

三省堂(2006)『大辞林』第３版

三省堂編修所編(2005)『コンサイスカタカナ語辞典 第３版』

三省堂編修所編(2010)『コンサイスカタカナ語辞典 第４版』

小学館(1998)『大辞泉　増補・新装版』第１版第３刷

小学館(2001)『日本国語大辞典　第２版』

『デジタル大辞泉』小学館提供（2010 年 7 月現在、約 24 万 2,000 語を収録）
http://dictionary.goo.ne.jp/leaf/jn2/120524/m0u/スルー/（2011.1.22 閲覧）

関連 URL

特定領域「日本語コーパス」ホームページ：http://www.tokuteicorpus.jp/

KOTONOHA「現代日本語書き言葉均衡コーパス」検索デモンストレーション
：http://www.kotonoha.gr.jp/demo/

『茶漉』日本語用例・コロケーション抽出システム（一般公開版）
：http://tell.fll.purdue.edu/chakoshi/public.html

ポライトネスからみた
「てくれる系」と「てもらう系」の使い分けに関する一考察

ジュ・ヒョンジュ(韓国 高麗大学校 中日語文学科) †

A Study on the Difference in the Usage of “te kureru” Type and “te morau” Type in Japanese with a Viewpoint of Politeness Theory

JU HYUNJU (Dept. of Chinese & Japanese Languages and Literatures, Korea University)

１．はじめに

この研究は「日本語書き言葉均衡コーパス」のデータを用い、現代日本語「てくれる系」と「てもらう系」の使い分けについて分析したものである。「てくれる系」と「てもらう系」は授受補助動詞の一種で、話し手の視点や行動の方向性を持つ表現である。

(1) a. パパは絵本を読んだ。(作例[i]) 【AはVする】
b. パパは私に絵本を読んでくれた。 【AはBにVてくれる系】
c. 私はパパに絵本を読んでもらった。 【BはAにVてもらう系】

上記の(1a)を基底文とし、話し手の置かれている状況によって、(1b)もしくは(1c)の表現が想定できる。これは、(1a)の命題に対して、話し手の頭の中で、(1b)か(1c)を選択する過程を経て、発話されるとも考えられる。このように、「てくれる系」や「てもらう系」の選択は、話し手の主観的態度を表わし、これを基底文に対するモダリティ的意味機能をもっているとも解釈できよう。

この際、同じ基底文からどのような要因によって「てくれる系」と「てもらう系」を選択するのかという問題が重要となってくる。この要因について探るため、本研究では語用論的なアプローチの中、ポライトネス理論(Politeness Theory)[ii]を援用し、分析を進めたい。

ポライトネス理論におけるフェイス(face)は、発話状況において人間が持っている欲求のことで、ポジティブ・フェイス(positive face)とネガティブ・フェイス(negative face)の二種類が存在する[iii]。ポジティブ・フェイスは、他人に好かれたい、認めてほしいという欲求で、この欲求を満足させる発話としては褒めことばなどがある。ネガティブ・フェイスは、他人に邪魔されたくない、侵害されたくない欲求をいい、相手と心理的に距離をおいて発話する尊敬語の表現などがある。

本研究では「てくれる系」と「てもらう系」の両表現の使い分けの要因をポライトネス理論(Politeness Theory)のフェイスの概念を用いて考察する。話し手の意図が人間の社会的欲求をどう守っているのかが言語表現に反映されていることを前提とし、「てくれる系」と「てもらう系」の選択要因の手がかりを探ることが本研究の主な目的である。

２．先行研究の分析

２．１． ポライトネスの概念

ポライトネスは、円滑な人間関係を築き、維持するために行われる言語行動を研究する分野の一つである。Brown & Levinson(1987:55-84)は、ポライトネスの概念を相手のフェイスを害せず、またフェイスを守るための言語学的戦略(strategy)と述べる。つまり、フェイスの問題をどう取り扱うのかによって言語形式の解釈が決まる。言い換えれば、話し手は聞き手のフェイスを守りながら会話

† murasakiju@korea.ac.kr

を進めていく[iv]が、この際話し手が聞き手のどのフェイスを守るのかによって発話形態が決まるわけである。

(2) a. 二つのコップを重ねて熱いコーヒーを入れて渡したとき
「あら、そこまで考えてくれたの？ありがとう」
b. 他の人より先に帰るとき
「あのう、すみません。お先に帰らせていただきます」

(2a)の場合、話し手は相手の行動を直接言及することで、相手の欲求の中で自分の行動を認めてほしいという欲求であるポジティブ・フェイスを守っている。「てくれる」と「ありがとう」という表現で、相手の行動「考える」を通し、話し手が利益を得たというような言い方をしている。(2b)の場合、話し手は相手の邪魔されたくないという欲求であるネガティブ・フェイスを守っている。ここでは「(さ)せていただく」と「すみません」という表現で、相手の領域に邪魔しないような言い方で表現している。

本研究で取り扱うポライトネスの概念は、より丁寧な表現、より丁寧ではない表現を見るのではなく、丁寧さの意が発揮できるような根拠として、聞き手のポジティブ・フェイスとネガティブ・フェイスをどう守るのかに注目し、考察したいと思う。

２．２．　「てくれる系」vs.「てもらう系」

「てくれる系」と「てもらう系」の使い分けに関する先行研究は、ジョン(1995)、熊田(2001)、高見・加藤(2003)、金澤(2007)がある。

まず、金澤(2007)は「てくださる」と「ていただく」の表現のなか、相互交換が可能な用例を集め、実際にはどの表現がよく使われているのかについて述べている。結果は「てくださる」と「ていただく」が相互交換可能な表現のなか、「てくださる」の使用が約 14%、「ていただく」の使用が約 86%であった。その理由については、「相手となるべく直接的な関わりを持たない形で人間関係を維持してゆきたいというミーイズム的な心理が、無意識のうちに関わっているのではないかと考える(ibid:50)」と述べる[v]。しかし、「ていただく」の表現がまったく相手との関わりを持たないわけではない。また、その関わりを持たない根拠として自己中心的な心理が作用される説明にはさらに論理的な説明が必要となる。

次に、ジョン(1995)は、日本語母語話者を対象にし、親疎関係による依頼表現のパターンについてアンケート調査で分析している。「先生が本を買ってくださった」と「先生に本を買っていただいた」を比べた結果は「ていただく」の方がより多用されたことが分かる。これに関して「ていただく」表現が「てくださる」表現より丁寧な表現であると述べている(ibid:37)。しかし、多用されたことでより丁寧な表現であるという捉え方は再考の余地があると思われる。したがって、本研究では「てくれる系」と「てもらう系」の表現が使用される要因について検討する。

熊田(2001)は恩恵行為として「てくれる」と「てもらう」を、働きかけ、当然性、負担の程度、それから発話時すでに行われている出来事なのかどうか(既往未然)という 4 つの段階に分けてアンケート調査を行っている。このなか、「てもらう」は、受恵者が本動詞の行動主に行為を行うように要求することであり、ここでは受恵者に対する配慮意識が高まる[vi]と述べている。だが、ここでは、行動主に行為を要求する主体が受恵者を配慮しているかどうかを探る根拠を調べる必要がある。また、配慮意識が高まるか低まるかはどう判断するのかも課題として残る。

そして、高見・加藤(2003)は両表現は本動詞の性質や性格、参与者の有情と無情の区別による「てくれる系」と「てもらう系」の共通点と相違点について述べている。特に相違点については例(3)

からみると、(3a)の「てくれる系」は恩恵を被ることのみを表わしているのに対して、「てもらう系」は恩恵の与え手に感謝する気持ちまで表明する表現であると述べる。したがって、「洗濯物」のように無情物の場合には容認度が低く、(3b)のように非適格文となると説明している。

(3) a. 洗濯物がすぐに乾いてくれて、助かった。(高見・加藤, 2003:97)
b.*洗濯物にすぐに乾いてもらって、助かった。(ibid)

感謝表現と「てくれる系」と「てもらう系」との共起関係については、本動詞の動作主の性格によってどのような相違点が存在するのかを綿密に考察する必要があると思われる。

以上先行研究でなされてきた議論をまとめてみたが、「てくれる系」より「てもらう系」のほうが丁寧に感じ、多用されている要因について未だに論理的な根拠が示されていないことは周知の通りである。本研究では未解決であった課題を中心に、「てくれる系」と「てもらう系」の使い分けについて考察する。

３．研究方法および分析対象

「日本語書き言葉均衡コーパス[vii]」のデータから得られた「てくれる系」と「てもらう系」の例文（以下、『日本語コーパス』と略す）を分析対象とする。研究方法は、『日本語コーパス』から「てくれる系」と「てもらう系」で用例を抽出し、次の(4a)のように【A はＶする】という命題に対して、(4b)と(4c)の過程を経て、(4d-1)と(4d-2)の選択に分けられる例を調べる。

(4) a. A は本を読む。【A はＶする】
b. B は(A が本を読むの)を聞く。
c. (A が本を読むこと)は B にとって何らかの影響を受ける。
d-1. A は B に本を読んでくれた。【A は B にＶてくれる系】
d-2. B は A に本を読んでもらった。【B は A にＶてもらう系】

それから、本動詞の行動主である A を中心に、A が話し手・聞き手・第 3 の関与者の場合に分け、各々登場人物のフェイスとはどのように関わり合っているのかについて検討する。

４．「てくれる系」vs.「てもらう系」

「てくれる系」と「てもらう系」は基底文の状況を話し手がどのような意図を持って発話するかによって決定される。「てくれる系」と「てもらう系」の発話状況では「話し手」と「聞き手」の存在は不可欠である[viii]。発話に登場する人物が加わるときもあるが、ここではそれを第 3 の関与者と言う。第 3 の関与者は二人以上も集団もあり得るが、本研究では二人まで考慮に入れることにする。次の<表1>のように本動詞の動作主(A)を中心に考えてみると、タイプ1では話し手、タイプ 2 では聞き手、タイプ 3 では第 3 の関与者と考えられる。各タイプには本動詞の受け手(B)が話し手と聞き手、第 3 の関与者が入る。

タイプ 1 の例は、「*私はあなたに本を読んでくれる」「*あなたは私に本を読んでもらう」、そして「*私は彼に本を読んでくれる」「＊彼は私に本を読んでもらう」などで分かるように「てくれる系」と「てもらう系」の視点と方向性の制約のため、非適格文となる。

＜表 1＞「てくれる系」「てもらう系」表現の登場人物

	話し手	聞き手	A（本動詞の動作主）	B（本動詞の受け手）	表現成立
タイプ 1	●	●	話し手	－	成立せず
タイプ 2	●	●	聞き手	話し手 第 3 の関与者	
タイプ 3	●	●	第 3 の関与者	話し手 聞き手 第 3 の関与者Ⅱ	

タイプ 2 は、本動詞の行動主が聞き手であり、受け手は話し手の場合と第 3 の関与者の場合がある。このタイプでは基底文【A は V する】の文をまず、【A は B に V てくれる系】にしてみると、聞き手の行動においてそのまま聞き手を主体と捉えて表現している。ここで主体と表現するのは、【A は V する】文をそのまま活かし、その文に「B に」「てくれる系」を添えることで、「A の行動」をより明確に強調する。これは、A が持っている欲求で一つであるポジティブ・フェイスを守る結果として捉えられる。

そして、【B は A に V てもらう系】にすると、「てくれる系」の主体であった「A は」が、「A に」と格が下がり、「B に」であったのが「B は」と格が上がる。ここでは、格の変動のため、エネルギーが必要となる。また、主体が変わることにより、本動詞の行動主が主語でなくなることがポイントである。つまり、これは聞き手である本動詞の行動主が持っている「他人に邪魔されたくない欲求」であるネガティブ・フェイスを守る結果である。

(5) a. シェフを呼んでくれるかしら。吉野と言ってもらえば判るわ。 （『日本語コーパス』）
b. あの椅子の修理はあなたにやってもらいますよ。 （『日本語コーパス』）

「てくれる系」の例である例(5a)では、「あなたは私にシェフを呼んでくれる」のように省略されている部分には聞き手を主語と取っていることが分かる。これは、未来に起こると思われる「シェフを呼ぶ」の行為をほめるという意味が含意されている。つまり、聞き手のポジティブ・フェイスを守る表現であるといえよう。反面、(5b)の「てもらう系」は「あなたが椅子の修理をする」という命題に対し、聞き手の「邪魔されたくない欲求」を守る意味として「てもらう系」が使われ、聞き手のネガティブ・フェイスを守る表現となる。

(6) a. パパ・ママ今まで育ててくれてありがとう。 （『日本語コーパス』）
b.?今まで育ててもらってありがとう。

例(6a)は親が聞き手で、「親が私を育ててくれる」という表現が「ありがとう」という感謝表現と共起している。しかし、(6b)のように「私は親に育ててもらう」と「ありがとう」の共起関係には違和感が感じられる。(6a)では、「てくれる系」表現により、聞き手のポジティブ・フェイスを守っているため、聞き手

が褒めてほしいと思っているところを「ありがとう」という感謝表現で表現している[ix]。しかし、(6b)では、「てもらう系」表現を使い、聞き手の行動を直接言及しない方法で発話しているため、感情を直接発話する表現と相容れない面があると解釈できる。

タイプ 3 は、本動詞の行動主(A)が第 3 の関与者であり、本動詞の受け手(B)には話し手、聞き手、第 3 の関与者Ⅱの場合が想定できる。

(7) a. 私の母は、そのことを私に教えてくれたように思います。
母は私にそのことを教えてくれる。 ⇔ 私は母にそのことを教えてもらう。 (『日本語コーパス』)
b. 身近にいる家族などに教えてもらったほうが、わかりやすいかもしれません。
あなたの家族はあなたに教えてくれる。 ⇔ あなたはあなたの家族に教えてもらう。
(『日本語コーパス』)
c. 彼の親に連絡して地元に連れて帰っていってもらったほうが良いのではないかと。
彼の親は彼を連れて帰っていってくれる。 ⇔ 彼は彼の親に連れて帰っていってもらう。
(『日本語コーパス』)

(7a-c)の例で分かることは、本動詞の行動主が話し手でも聞き手でもない場合、その場にいない人物に対して、そのフェイスを守ろうとする行為は考え難いと思われる。ここでは、話し手が本動詞の行動主のフェイスを守ろうとするよりは、話し手と聞き手との心的距離が縮まれる効果があると思う。言い換えると、これは話し手と聞き手とのわれわれ意識の表出であると思われる。

(8) 悪い姿を見られたのに、やさしい人に心配してもらってありがたかった。(『日本語コーパス』)

(8)の例文で「やさしい人」は発話時にはいない人物である。この文を聞き手に発話することによって得られる効果は、話し手が感じているありがたさを一緒に感じてほしいというのがその根底に存在していると思われる。それで、先述した(6b)の例では「てもらう系」と「ありがとう」の共起関係が薄かったが、(8)の「てもらう系」と「ありがたい」という感謝表現の共起関係は成立するのである。それは、「ありがたい」と「ありがとう」という表現の相違点にもつながる。「ありがたい」は話し手の心理状態の表現であるが、「ありがとう」はその心理状態を積極的に伝達しているため、「てもらう系」よりは「てくれる系」の表現と共起しやすいのではないかと思う[x]。

以上ポライトネスのフェイスの概念を用いて、「てくれる系」と「てもらう系」の使い分けを本動詞の行動主を中心にいくつかのタイプに分けて分析を行った。両表現の使い分けにおいて、話し手が聞き手のどのフェイスをどう処理するかということが最も重要であろう。

５．おわりに

本研究では、「てくれる系」と「てもらう系」の使い分けがどのような基準で行われているのかを探るために、ポライトネスという観点から分析考察を行った。ポライトネスのフェイスの概念から「てくれる系」は聞き手の「相手に好かれたい欲求」であるポジティブ・フェイスを守っている表現であり、「て

もらう系」は聞き手の行動が話し手にとっていい意味を持つことを間接的に表わしているため、聞き手の「邪魔されたくない欲求」のネガティブ・フェイスを守っている表現であるとまとめられる。

今回は、日本語の「てくれる系」「てもらう系」の使い分けに焦点をあてて分析したが、韓国語・英語の表現との比較・対照研究を通し、日本語の表現の特徴がさらに明確に把握できると思われ、これに関しては今後の課題としたい。

文献

宇佐美まゆみ(2000).『言葉は社会を変えられる』, 明石書店, pp.204-292.

岡本真一郎(1997).「聞き手への配慮と言語表現」『愛知学院大学文学部紀要』Vol.27, 愛知学院大学文学会, pp.23-36.

金澤裕之(2007).「『～てくださる』と『～ていただく』について」『日本語の研究』3, 日本語学会, pp.47-53.

熊田道子(2001).「待遇意識からみた『～てくれる』系表現と『～てもらう』系表現-恩恵の与え手が恩恵行為を行うことに対する配慮意識を中心に-」『国語学研究と資料』, 早稲田大学文学学術院, pp.15-28.

高見健一・加藤鉱三(2003).「受益表現の新展開 6『～てくれる/もらう』の相違」『言語』Vol.32, No.6, 大修館書店, pp.96-101.

滝浦真人(2008).『ポライトネス入門』, 研究社.

정혜경(ジョン・ヘギョン, 1995).「親疎關係에 의한 依賴表現의 使用法」『日語日文学研究』Vol.27, 韓国日語日文学会, pp.27-55

Brown, P. & Levinson, S.(1987). Politeness : Some universals in language usage. Cambridge, Cambridge University Press, pp.55-84, pp.403-426.

---

i 出典の記述がない例文は筆者による作例であり、ネイティブ・チェック済みである。

ii 本研究で取り扱うポライトネス理論は、主に Brown & Levinson(1987)で提唱された概念で、話し手が聞き手を配慮して発話した表現を意味する。先行研究では「丁寧さ」「配慮表現」など様々な用語で書かれている。しかし、「そのブラウス素敵ね」という発話をポライトネス理論では、一つのポライトネス戦略として説明しているが、これを日本語の「丁寧さ」という用語では説明しきれない面がある(宇佐美, 2000:250-252)ため、本稿では「ポライトネス」と表記する。

iii フェイスの概念について、Brown & Levinson(1987:61-63)では、「positive face」「negative face」と記述し、その定義について次のように述べている。

ⅰ) positive face ; the want of every member that his wants be desirable to at least some others.

ⅱ) negative face ; the want of every 'competent adult member' that his actions be unimpeded by others.

研究者によっては各々「積極的フェイス」「消極的フェイス」とも言う(岡本, 1997:24)が、本研究では「ポジティブ・フェイス」「ネガティブ・フェイス」とする。

iv 岡本(1997:24)では「一般的に人々は互いのフェイス維持のために協力する」と述べているが、これはフェイスを守ることとの意味につながると思われる。

v 「『送っていただき…』と表現すると頭の中に相手の姿が現れる必要はない。(中略)とりあえず、相手に直接言及しない表現形式である『送っていただき』の方が選択され易くなっているように思われる」(金澤, 2007:50-51)

vi 「恩恵の受け手が恩恵の与え手に対し、恩恵行為を要求するということは、恩恵の与え手に恩恵の受け手のための負担を負わせることになる。そのため、『働きかけ』がある場合、恩恵の受け手に対する配慮意識は高まる」(熊田, 2001:17)

vii 日本語コーパスホームページ:http://www.tokuteicorpus.jp/

viii ここでは、話し手と聞き手のコミュニケーションの状況を前提とするため、話し手の独りことばは対象外とする。

ix 「ありがとう」の感謝表現は「好かれたい・よく思われたい欲求」の表れとして表出され、「すみません」の謝罪表現は「邪魔されたくない欲求」として表現される(滝浦, 2008:28)。

x 「彼はありがたかった」は言えるが、「＊彼はありがとう」が非適格文であることにも関わるように、人間の心理を直接表出するのか、間接的に状態を表現するのかの差であると思われる。

「かなしい、つらい、くるしい」の意味について

加藤恵梨（名古屋大学留学生センター）

The Meaning of *kanashii, turai, kurushii*

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

１．　本研究の目的

本研究では「かなしみ」の感情を表すと考えられる「かなしい」「つらい」「くるしい」の意味分析を行うとともに、それらの意味の類似点および相違点を明らかにすることを目指す。

２．　「かなしい」の意味分析

2.1.　先行研究の記述とその検討

森田（1977）と飛田・浅田（1991）では「かなしい」について次のように記されている。

森田（1977：308）

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

飛田・浅田（1991：154）

心が痛んで泣きたいような様子を表す。

しかし、上のような記述は「かなしい」の類義語である「つらい」や「くるしい」にも当てはまると考えられるため、「かなしい」の意味について実例を基に再度検討する必要がある。

2.2.　「かなしい」の意味分析

2.2.1.　別義１：〈思いと異なるよくない事態に〉〈衝撃を受け、力がぬけるように感じる〉〈さま〉

(1)　（前略）それが芸能人の家庭なんかだと、十年も寝たきりだった老人が死んだときにも大泣きしている。「もう悲しくて生きる気力がなくなりました」嘘つけって。
（やっぱり私は嫌われる）

(2) 怒っているはずなのに、惨めな思いが先に立ち、悲しくて涙が溢れてきた。美由紀は嗚咽を嚙みしめて、松浦から顔を背け、窓を叩く雨を見た。　（はちまん）

例(1)は家族を失ったこと、例(2)は相手の男性の家に泊まりたいという発言が聞き入れら

れず、惨めな思いをすることを「かなしい」と表している。よって例(1)と(2)では、主体の思いとは異なる良くない事態が生じたことに「かなしい」と感じているということができる。また、例(1)では「悲しくて生きる気力がなくなりました」とあることから、「かなしい」とは主体の思いと異なる良くない事態に衝撃を受け、力が抜けるように感じるさまであるといえる。

2.2.2. 別義２：〈不純なものがまじっていないようすに〉〈衝撃を受け、心が動かされる〉〈さま〉

(3) 美的特質、純情、優しさといった心性を何も損なうことなく傍でそっと眺めていたい、自由な子供のままの無邪気さで女性を戯れさせておきたい…そういう内奥からの希求は、自我の確立を許さなかった劣悪で薄幸な環境からくる、自己主張のまるでできない人間の悲しいほど純粋な公平無私を型どっていたものと考えられる。
（細い蔓）

(4) （前略）シエラ号事件はどうか？　このスキャンダラスな事件は、調査捕鯨船団に乗り組んでいる悲しいほど誠実な彼らの罪ではないはずなのに、日本政府の捕鯨問題に対する取り組みや水産業界のモラルの低さをさらけ出し、違反とゴマカシだらけの捕鯨船団というイメージを定着させたのだ。　（クジラを捕って、考えた）

例(3)と(4)では、ある人の態度に全く不純なものがなく、心が動かされるほど純粋あるいは誠実であるさまを「かなしい」と表している。よって、「かなしい」の別義２は、不純なものがまじっていないようすに衝撃を受け、心が動かされるさまということができる。

３．　「つらい」の意味分析

3.1. 先行研究とその検討

森田（1977）と飛田・浅田（1991）では、「つらい」の意味について次のように記述されている。

森田（1977：308）

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

飛田・浅田（1991：368）

① 精神的に苦痛を感じる様子を表す。

② 冷酷で思いやりのない様子を表す。

しかし、次の例(5)のように、森田や飛田・浅田の記述では説明できない例も見られる。よって、例(5)のような「つらい」の意味記述も含め、実例を基に分析する必要がある。

(5) （前略）いわゆる「日本史上の有名人」があまり出てこないので、観ているうちに頭の中で系譜のおさらいをしたくなった。終演後も「えー、あの人はいとこ？兄弟？まあなんでもいいわ」という会話が出ており、プログラムなしではつらいだろう。この点でもえらく難しい戯曲の感じがした。
(http://www.musical-fan.org/kb/index.cgi?b=musical&c=e&id=217)

3.2. 「つらい」の意味分析

3.2.1. 別義１：〈身体に負担がかかり〉〈耐えられないと感じる〉〈さま〉

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

例(6)では目がかゆいというように、身体に負担がかかることで「つらい」と感じている。また例(7)では、体がつらくなったことで「どうしてもそこにいることができませんでした」とあるように、「つらい」とは身体に負担がかかることで耐えられないと感じるさまであるということができる。

3.2.2. 別義２：〈望ましくない事態に接し〉〈耐えらないと感じる〉〈さま〉

(8) 「(前略) つまり、好きなことをしている時は、どんなに大変でも、あまり苦にはならないけど、反対に、厭なことをしている時には、どんなに楽なことでも、とても辛く感じてしまう（後略)。」 (紅玉の火蜥蜴)
(9) マザーは貧しい人に尽くすというこの活動を、最初はたった一人で始めたと聞く。彼女は普通の人より愛の量が多すぎて、世の中に貧しさで苦しんでいる人がいることが辛くて耐えられなかったのだろう。 (インド::やっぱりノープロブレムへの旅)

例(8)では厭なことをすること、例(9)では「世の中に貧しさで苦しんでいる人がいる」ということに「つらい」と感じている。よって、ここでは主体が望んでいない事態に接することで「つらい」と感じているということができる。さらに例(9)では「辛くて耐えられなかった」とあることから、「つらい」とは望ましくない事態に接し、耐えらないと感じるさまであるということができる。

3.2.3. 別義３:〈自身の行為を成立させるのが〉〈困難であると感じる〉〈さま〉

(10) (前略) いわゆる「日本史上の有名人」があまり出てこないので、観ているうちに頭の中で系譜のおさらいをしたくなった。終演後も「えー、あの人はいとこ？兄弟？まあなんでもいいわ」という会話が出ており、プログラムなしではつらいだろう。この点でもえらく難しい戯曲の感じがした。 (＝(5))

(11) ISDN 回線を契約して 2 回線を使う場合、1 台の電話機付き FAX で 2 回線を使うことは出来るのでしょうか？ (FAX と電話機、それに家の電話と 3 つ並べるのはスペース的に辛いので・・・) (http://qanda.rakuten.ne.jp/qa2565503.html)

例(10)では「つらい」と「難しい」が同じような意味を表すものとして用いられていることから、プログラムなしで戯曲の人物関係を理解することが困難であると感じているさまを表している。続いて例(11)ではあるスペースに FAX、電話機、家の電話の 3 台を並べることが困難であると感じるさまを「つらい」と表している。よって、「つらい」の別義３は、自身の行為を成立させるのが困難であると感じるさまということができる。

４. 「くるしい」の意味分析

4.1. 先行研究とその検討

森田 (1977) と飛田・浅田 (1991) では、「くるしい」の意味は次のように記されている。

森田 (1977 : 190)

がまんできないほど肉体的または精神的に圧迫感を覚え、つらいさま。

飛田・浅田 (1991 : 214-215)

① 肉体的・精神的に苦痛である様子を表す。

② ①の精神的な苦痛の原因を財政面に限定した意味である。すなわち、財政的に困難な様子を表す。

③ ①から進んだ意味で、精神的な苦痛を感じさせるように不自然な様子を表す。

森田では、肉体的または精神的に圧迫感を覚えることによって「くるしい」と感じると指摘されているが、「くるしい」と感じる要因は「圧迫感」に限られないことが次の例から分かる。

(12) 手術は一応成功し、ガスも便も通じたのだから、あとは一枚一枚紙をはがすように回復していくはずなのに、三十九度前後の熱が続いて、だるくて苦しい。

(人は死ねばゴミになる :: 私のがんとの闘い)

次に飛田・浅田では「くるしい」の一つ目の意味を「肉体的・精神的に苦痛である様子

を表す」と記されているが、この意味記述は「つらい」にも当てはまると考えられるため、実例を基に再度「くるしい」の意味について検討する必要がある。

4.2. 「くるしい」の意味分析

4.2.1. 別義１:〈身体の異常な状態に〉〈苦痛を感じる〉〈さま〉

(13) （前略）その桐原の言い方で、聖司は口のなかの蕎麦を噴き出しそうになったが、懸命にこらえているうちに蕎麦が喉に詰まりかけ、苦しくて涙が溢れてきた。
（にぎやかな天地）

(14) 中毒症状は激しい嘔吐と下痢で相当に苦しい。しかし重態に陥ることはない。
（世紀を超えて広がる「毒」:: 気の毒・液の毒・固の毒）

例(13)では喉に異物が詰まり、呼吸を十分にすることができないことによって、例(14)では激しい嘔吐と下痢によって「くるしい」と感じている。これらは身体が異常な状態になることで苦痛を感じるさまを表している。

4.2.2. 別義２:〈追いつめられ〉〈どうすることもできないと感じる〉〈さま〉

(15) 味方飛行機の援護がまったくない裸の水上艦隊が、敵機に襲われた場合、その戦いがいかに苦しく惨憺たるものであるか、それは今次「あ」号作戦のあとにくりひろげられる比島沖海戦（レイテ沖海戦）で、いやというほど味わわされるのである。
（戦艦大和いまだ沈まず::艦橋見張員の見た世紀の海戦）

(16) 「火事場の馬鹿力」といわれる通り、人間は切羽詰ったときに本当の力が出る。苦しくてにっちもさっちもいかなくなったところで、光明が差し、救いの神のようなアイディアが不思議に湧いてくる。
（http://www.cosmoprints.co.jp/cosmonews/?p=225）

例(15)では、もう後がないというところまで追いつめられることによって「くるしい」と感じている。また、例(16)では「苦しくてにっちもさっちもいかなくなった」とあることから、「くるしい」は、追いつめられ、どうすることもできないと感じるさまを表しているということができる。

4.2.3. 別義３:〈ある事態の成立が〉〈不可能であると感じる〉〈さま〉

(17) （前略）おそらく浪本氏は、教育委員会の教科書採択の権限の否定が、理論的に苦しい、成立しない議論であることを内心よくわかっているのであろう。
（『教科書採択の真相: かくして歴史は歪められる』 p.161）

(18) しかし、客観的に見て、ライスの説明は苦しい。　　(『情報と外交』 p.153)

例(17)では「くるしい」が「成立しない」という表現と同等の意味として用いられていることから、ここでの「くるしい」は、教育委員会の教科書採択の権限を否定することといったある事態の成立が不可能であると感じるさまを表している。また例(18)では「客観的に見て苦しい」とあるように、ある事態の成立が不可能であるとみなすのは何らかの根拠に基づいたものであると考えられる。よって、「くるしい」の別義３はある事態の成立が不可能であると感じるさまということができる。

５．　「かなしい」と「つらい」の意味の類似点と相違点について

5.1.　先行研究とその検討

飛田・浅田（1991：154-155）では、「かなしい」と「つらい」の意味について次のように述べられている。

> 精神的な苦痛を表す意味で「かなしい」は「つらい」に似ているが、「つらい」は意味の範囲が広く、さまざまな感情においてたえがたいという意味を表すのに対して、「かなしい」は悲哀に限定される点が異なる。

飛田・浅田では「つらい」の方が「かなしい」よりも意味の範囲が広いと述べられている。この指摘が正しいのか、それ以外にも両語の意味に違いがないのかについて、以下で実例を基に考察する。

5.2.　考察

(19) 殺人事件の遺族は、死んだ人のことを忘れることはありませんし、命日がくれば非常に悲しい、つらい、というふうに話されます。　　(トラウマの心理学)

例(19)では「かなしい」と「つらい」が同等の意味として用いられている。また、ここでの「かなしい」は別義１で、「つらい」は別義２で説明可能であることから、「かなしい」の別義１と「つらい」の別義２が類似しているということができる。

では次に、「かなしい」の別義１と「つらい」の別義２の相違点について考察する。

(20) 毎日が大変な状況であり、何をするのか予想できないので、下校までまったく目を離せない状況が続きました。担任の私自身がどうしてよいのか分からず、先行きが分からないことからイライラがつのり、体に変調をきたしてしまいそうで、学校に行くのが大変つらい（??かなしい）時期でもありました。

(ADHD の子育て・医療・教育 :: 親と医師、教師が語る)

例(20)は「つらい」の別義２を表すが、これを「かなしい」に置き換えると不自然な表現となる。「かなしい」は自分の意志とは無関係なところで思いと異なる良くない事態が起こり、その結果を受けて生じる感情であるため、「学校に行く」といった自分の意志で望ましくない行動をとるというような場合には用いることができない。

(21) 政治家と親密な関係を持たない公務員にとって、自分の業績を政治家に見落とされることはとても辛い。政治家に評価されなければ、せっかくの仕事が昇進のポイントにならないからだ。 (巨大市場インドのすべて)

(22) 自分の腹を痛めた宗助利則が、落馬してから記憶を全く失い、父や母の顔もおぼえていないのは、悲しい (??つらい) ことであった。しかし、園城寺で新しい一人の僧として生まれかわるのは、せめてもの救いになる。 (戦国風流：：前田慶次郎)

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

「つらい」の別義２は例(21)のように、自身に何らかのダメージを受けることによって生じる感情であるということができる。そのため、例(22)のようにダメージがあまり大きくない事態に対して「つらい」というと不自然な表現となり、例(23)のように「ちょっぴり」といった副詞とも「つらい」は共起しがたい。

６．「つらい」と「くるしい」の意味の類似点と相違点について

6.1. 先行研究とその検討

森田（1977：190）では「つらい」と「くるしい」の違いについて、「くるしい」には肉体的、生理的苦痛の気持ちが強く、「つらい」には精神的苦痛の色合いが濃いと述べられている。

確かに、実例においても、「つらい」は「くるしい」よりも主体の精神的苦痛に焦点が置かれている表現であると考えられる例が見られる。

(24) 隆一は、今刀比羅さんの石段を登ったときのことを、今でも苦しく (??つらく) も楽しい記憶として、大切にしている。 (ステップ)

例(24)のように「苦しくも楽しい記憶」と表現することは可能であるが、この「くるしい」を「つらい」に置き換えると不自然な表現となる。これは、「くるしい」の意味は肉体的苦痛に焦点が置かれているため、肉体的疲労が回復すれば石段を登ったことも「楽しい記憶」となるのに対し、「つらい」の意味は精神的苦痛に焦点が置かれているため、「つらい」と「楽しい記憶」とが共起しないためである。

では、「くるしい」が精神的苦痛を表す場合に「くるしい」と「つらい」の意味が類似し

ているのかについて以下で考察する。

6.2. 考察

(25) 「火事場の馬鹿力」といわれる通り、人間は切羽詰ったときに本当の力が出る。苦しく（??つらく）てにっちもさっちもいかなくなったところで、光明が差し、救いの神のようなアイディアが不思議に湧いてくる。 （＝(16)）

例(25)は「くるしい」の別義２を表すが、これを「つらい」に置き換えると不自然な表現となるであろう。「くるしい」が表す精神的苦痛とは、追いつめられてどうすることもできないと感じるさまを表すが、「つらい」はそのような意味を表さない。

次に、「つらい」の別義３と「くるしい」の別義３の関係について考察する。

(26) ISDN 回線を契約して 2 回線を使う場合、1 台の電話機付き FAX で 2 回線を使うことは出来るのでしょうか？（FAX と電話機、それに家の電話と 3 つ並べるのはスペース的に辛い（くるしい）ので・・・） （＝(11)）

(27) しかし、客観的に見て、ライスの説明は苦しい（??つらい）。 （＝(18)）

例(26)のように「つらい」は、主体の行動の成立に関する意味を表すため、例(27)のように他者の説明が客観的に成立するか否かというような場合には用いることができない。それに対して「くるしい」は、例(26)のように主体の行動に関しても、例(27)のように他者の行動に関しても用いることができるという点で両語は異なっている。

７． 「かなしい」と「くるしい」の意味の類似点と相違点について

「かなしい」と「くるしい」については意味が類似しているとはいいがたい。

(28) 「火事場の馬鹿力」といわれる通り、人間は切羽詰ったときに本当の力が出る。苦しく（??かなしく）てにっちもさっちもいかなくなったところで、光明が差し、救いの神のようなアイディアが不思議に湧いてくる。 （＝(25)）

８． 今後の課題

今後も他の感情形容詞の意味について分析していきたいと考える。

引用文献

飛田良文・浅田秀子（1991）『現代形容詞用法辞典』東京堂出版
森田良行（1977）『基礎日本語Ｉ』角川書店

日本語学習者のための語の用例記述に向けて
－辞書の品詞・用例から学ぶことができない語の情報－

前坊香菜子（一橋大学大学院生）†

Desiderata for Dictionary Entries for Non-native Learners of Japanese: Lexical Information that cannot be Deduced from Part of Speech

Kanako Maebo (Ph.D. student, Hitotsubashi University)

１．はじめに

外国語を学習するには辞書は不可欠である。英語に関しては、コーパスによる研究の成果を反映した数多くの優れた辞書が出版されている。しかし、日本語学習辞典に目を向けると、様々なレベルの日本語学習者（以下、学習者）の使用に耐えうる辞書はみられない。学習者は、二言語辞書、あるいは国語辞典を学習辞典に代用していることが多いが、その記載内容は意味記述が中心で統語的な情報や共起語などの情報が不十分であることが多い。

例えば、語の品詞情報はほとんどの辞書に記述されており、語の使用を考える一つの助けとなっている。しかし、村木（2004）や森下（2006）は、「抜群」・「最寄り」などのように辞書では名詞扱いされているにもかかわらず統語的には名詞の特性を持っていない語があると指摘している。また、意味としては難しくない語であっても、使用するとなると単純ではないこともある。例えば、「多い」は形容詞であるため「多い人が集まる」という使い方も統語的には可能であるはずだが、実際には「多くの人が集まる」としたほうが自然な日本語である。しかし、ほとんどの辞書では、「多い」と「多く」は別の見出し語になっており、このような使用の相違は明示的に記述されていない。また、「多い」の対義語である「少ない」には「多い」のように「連用形＋の名詞」の形をとることができないが、そのような記述は辞書にはない。

母語話者は多くの実例に接する中でこのような事柄を暗黙のうちに了解していることが多いが、日本語を母語としない学習者にとっては教科書と辞書以外の情報に接することが難しいため、学習辞典ではこうした事柄を明示的に記述することが求められる。

２．調査の目的

類義語や対義語で品詞が同じ語の場合、意味の違いで使い分けを考えようとするが、実際には統語的な違いや共起する語を提示することで理解を助けることもある。そこで、前述の「多い」「少ない」を含む量を表す類義語を例として、語の統語的な特性とその統語的なパターンの中で共起する語の傾向を明らかにするために量的な調査を行う。調査対象とするのは、「大量」「少量」「多量」「たくさん」「少し」「多い」「少ない」の７語である。

３．調査方法

調査対象語の 7 語を「『現代日本語書き言葉均衡コーパス』モニター公開データ（2009年度版）」（以下、BCCWJ）から検索し、用例を採取する。用例を採取する際、かな表記と漢字表記がある場合はそれらも含めて検索した。「多い」「少ない」は活用があるため活用

† id092004@g.hit-u.ac.jp

した語形も含めて検索した。

用例を採取した後、それぞれの語の統語的な特性をみるため直後に接続する助詞「が」「を」「に」「の」の割合を調査した。この調査では語の単独でのふるまいをみるため、複合語を除いた割合を調査している。共起する語の調査では、①調査語＋助詞「が」「を」「に」「の」の直後に共起する語と、②調査語の直前でガ格とヲ格をとっている語（例「～が大量」の「～」にくる語）を対象とした。

４．調査結果

４.１「大量」「多量」「少量」

４.１.１　統語的な特性

「大量」「多量」「少量」の品詞は辞書に名詞と記されている。しかし、調査の結果、名詞であれば本来現れる統語的な特性と一致しない傾向がみられた。語の直後に接続している助詞「が」「を」「に」「の」を調査した結果を表１に示した。

表１　「が」「を」「に」「の」が接続する用例数

	大量		多量		少量	
全用例数	1527	%	260	%	314	%
が	1	0.07	0	0.00	1	0.32
を	0	0.00	0	0.00	21	6.69
に	711	46.56	106	40.77	7	2.23
の	760	49.77	147	56.54	139	44.27
助詞用例計	1472	96.40	253	97.31	168	53.50

「大量」「多量」が名詞であることからヲ格をとることができるように思えるが、用例ではヲ格をとった用例はなかった。ノ格が用例のほぼ半数を占めている。そして、「に」を接続した語形で副詞的な使用が 4 割以上を占めている。全用例の大部分がノ格と「に」を接続した語形である点が特徴的である。また、調査した助詞を接続する用例が全用例の大部分を占めている。

一方、「少量」は「大量」「多量」と品詞が同じで、意味も単純に対義語のようであるが、実際には異なったふるまいをしている。ノ格をとる割合が多いことは対義語の２語と似ている。しかし、「少量」は対義語がとっていない「を」を接続する用例が採取された。これは「塩少量を入れる」のように調味料や油など液体類の直後に「少量」が続くものである。また、「大量」「多量」とは異なり「に」と接続する用例は 7 例のみで副詞的な使用は少ない。そして、全用例のうち調査した助詞が接続している割合が対義語と比較して少ない。

４.１.２　共起する語

共起する頻度が高い語は、語の意味や用法を明確にすると考えられる。そのような語を辞書の記述に用いることは学習者にとっても有用な情報となる。

前後に共起する語を調査した結果を表２に示した。特定の語が共起する場合もあるが、

多くの場合それに類する言葉であることも多い。特定の語が少ない場合は「/」や「～類」とグループとして提示した。「大量」は、名詞は 5 例以上、動詞は 10 例以上の語、「多量」「少量」は 5 例以上の語を挙げている。

表２　前後に共起する語

	後ろに共起する語	前に共起する語
大量	に～: 生産する、発生する、ある、使う、存在する、輸入/輸出する、買う/売る の～: 水、血/血液、資金/資本、情報/データ、	～が: 商品/製品、魚介類、分泌物類(アルブミン等) ～を: 商品/製品、資金/預金、情報/データ、食品類、
多量	に～: 含む、摂取する/摂る、 の～: 水、血液/出血、金/銀、元素類、	～が: 化合物類(アミノ酸・カリウム) ～を: 元素類(酸素・水素)、化合物類(ビタミン・カルシウム)
少量	の～: 塩、出血、水、油、アルコール類、(発注) を～*: 混ぜる、取る　*2 例以上	～を: 油、(単価)

『明鏡国語辞典』を見ると、「大量」は「数量が多い」、「多量」は「分量が多い」という説明のみである。用例を見ると、「商品、資金、情報、血液」などは 2 語に共通して使われているが、「多量」では化合物類、元素類の名詞の使用が目立った。それに関連して動詞も共起している語が異なっている。「少量」は共起する語をみると、液体や調味料類が多く使用されており、動詞もそれと関係のある語の使用がみられた。

調査対象語の前後に共起する語の統語的なパターンをみると、「大量」「多量」は「～が大量／多量に…」「～を大量／多量に…」が大部分である。しかし、「少量」にはそのようなパターンはみられなかった。

４.２「たくさん」「少し」

４.２.１　統語的な特性

「たくさん」「少し」は副詞であるため助詞が接続する割合は高くなかった。調査結果を表３に示した。

表３　「が」「を」「に」の」が接続する用例数

	たくさん		少し	
全用例数	5493	%	14442	%
が	0	0.00	2	0.01
を	1	0.02	0	0.00
に	19	0.35	5	0.03
の	1345	24.49	285	1.97
助詞用例計	1365	24.85	292	2.02

副詞であるため「が」「を」が接続することはほとんどなく、「の」と接続した語形の用例が最も多かった。一般的に学習者は「の」は名詞の後に接続するものとして認識してい

るため、副詞にこのような用法があることには気づかないことが多い。辞書で副詞とされている語で「の」が接続する語には、この点を明示的に記述しておくべきであろう。

４.２.２　共起する語

「たくさん」「少し」と共起する語を表４に示した。20 以上採取されたものである。「たくさん」「少し」は副詞であるため、接続する助詞の前後の名詞だけでなく、直後に共起する動詞、形容詞も調査した。この 2 語は共起する語と関係のある特徴的な統語的なパターンはなかった。しかし、ほとんどの用例では語の動詞の直前に来ている。日本人であれば直感的にわかることであるが、学習者には知りえない情報である。副詞が動詞の直前に置かれるということをすべての副詞に書く必要はないかもしれないが、何らかの形で示したほうがよいだろう。また、「たくさん」の直後に形容詞が続く例は少なかったが、「少し」では「高い」「不安」などの形容詞を修飾した用例がみられた。

「たくさん」を『明鏡』で調べると、「数量・回数が多いさま」という意味があり、その例として「あの店にはたくさん行った」が挙げられている。BCCWJ の調査では「たくさん行く」という例はない。共起する語を見ても、「数量が多い」ことが意味の中心であるように思われる。

「少し」が「少しの～」という語形で使用されている際に共起する語として「間／あいだ」「時間」「辛抱」の 3 語が多く出現していた。そのうち『明解』では「もう少しの辛抱だ」という例が挙げられているが、典型的な例としては「少しの間」も取り上げたほうがよいだろう。

表４　前後に共起する語

	後ろに共起する語	前に共起する語
たくさん	に～：なる、ある、もつ、こうむる の～：人、回答、種類、本、情報、問題、友人	～が：人、種類、問題、回答、本、例、店、友人 ～を：水、野菜、話
	ある、いる、つく/つける、作る、できる、出る/出す、取る、見る、もらう、飲む、稼ぐ、含む、起こる、載る、残る/残す、使う、食べる、書く、知っている、読む、入る/入れる、買う、並べる/並ぶ、来る、持つ、集まる/集める	
少し	に～:する の～：間、時間、辛抱、	～が：こと、もの、話 ～を：こと、水、話、お酒、時間類、感情類、体の一部類、
	ある、おかしい、遅れる、伺う/尋ねる/聞く、話す、違う、安心する、違う、飲む、下がる/下げる、過ぎる、開ける/開く、間、休む、考える、行く、高い、困る、残る/残す、(時間が)かかる、出る、触れる、心配、先、前、大きい、遅れる、長くなる、入れる、不安、変える/変わる、歩く、離れる、	

４.３「多い」「少ない」

４.３.１　統語的な特性

「多い」「少ない」は形容詞であるため助詞が接続するとは考えにくいが、実際には助詞が使われる用例がある。表５には調査結果を示した。

表５「が」「を」「に」の」が接続する用例数

	多い		多く		少ない		少なく	
全用例数	31765	%	16621	%	7743	%	3098	%
が	474	1.49	474	2.85	1	0.01	0	0.00
を	194	0.61	185	1.11	0	0.00	0	0.00
に	87	0.27	45	0.27	14	0.18	0	0.00
の	6706	21.11	6706	40.35	1	0.01	0	0.00
助詞用例計	7461	23.49	7410	44.58	16	0.21	0	0.00

この 2 語は形容詞であるため活用した語形があるわけであるが、採取された用例の中で最も多く出現した語形は「多く」「少なく」で、各語の全用例の半数を占めていた。

「多い」は、助詞が接続している用例が採取されたが、大部分が「多く」に接続している。一方、「少ない」は助詞が後ろに接続する用例はわずかである。「に」の用例が採取されているが、これは「少ないにもかかわらず」「少ないにせよ」という形で現れている。

「多く」「少なく」は用例の半数を占めているが、「多く」は助詞が接続した用例があるのに対して、「少し」にはない。「多く」少なく」の用例を分析すると、「多く」は形容詞と副詞、名詞の使い方がそれぞれ半数であるが、「少なく」はほとんどが形容詞であった。

「多くが」は「～の多くが」という形や文や節の最初など名詞として使われていた。『明鏡』では「多く」の見出し語の中で「ある集団の中の大多数。大部分。」という意味の例文に「学生の多くが反対した」と挙げているが、この記述だけでは「多くが」名詞として使われることをすぐに読み取ることはできない。学習辞典には明示的な記述が必要である。

また、学習者に多く見られる間違いに「多い＋名詞」「多くの+名詞」がある。この点に関して用例を見ると、「多い＋名詞」では「N１が/の多い N2」「N１に多い N2」という形で、そして、「多くの N」は「多くの N が/を」の形で使われていた。

４.３.２　共起する語

本稿では用例の中で多くみられた「多くの～」「～が多い」「多く～」に共起した語を取り上げる。30 以上出現した語を示した。また、「～の多くが」というパターンもよくみられたため共起する語を調査し、10 以上現れたものを提示した。「～が少ない」20 例以上「少ない～」「少なく」は 10 例以上とした。表６にその結果を示した。

表６　前後に共起する語

	後ろに共起する語	前に共起する語
多い 多く	多くの～:　人、こと、もの、課題、患者、企業、研究、研究者、国、国民、困難、子供、時間、若者、種類、女性、場合、人間、点、読者、日本人、部分、分野、問題 多い～:　気、国、場合、人、地域、法人	～が多い:　人、ケース、場合、数、企業、問題、国、店、人数、回答、量、部分、子供、女性、～点、～者、 ～の多くが:　その、国民、人(人、～人、～員)
	多く～:　取る、ない、なる、見られる、含む、見かける、見られる、見受けられる、使う、持つ、集まる/集める、出る、存在する、用いる、	
少ない	少ない～:　傾向、子/子供、資本/資金、場合、人、数、製品、地域、量、	～が少ない:　人、機会、人口、場合、情報、数、～数、例/用例、量、時間、
	少なく～:　する、なる、	

「～が多い」「～が少ない」に共起する名詞に関しては目立った違いはみられない。しかし、「多い～」「少ない～」では共起する語の違いがみられる。「多い気がする」は頻度が高いが「少ない気がする」は3例のみである。また、「傾向」は「少ない」にはあるが「多い」には使われていない。「多い＋名詞」と「多くの＋名詞」でも統語的に違いがあるため、共起する語の傾向が異なる。辞書では２つの形を用例としては挙げられてはいるが、学習辞典では単に用例だけなく、明示的に違いを示す必要があるだろう。

副詞の用法についてであるが、用例では「多く」は様々な動詞と共起し副詞としての用法がみられるが、「少なく」は副詞の用例は少なく、共起する動詞は「する」「なる」が多かった。このような違いも語の意味と共にパターンで示せば、有用な情報となるだろう。

５．まとめ

「大量」「多量」「少量」「たくさん」「少し」「多い」「少ない」について、その統語的な情報と共起する語を調査した。各語のふるまいを見ると同じ品詞であっても異なる使用傾向が観察された。「大量」「多量」「少量」のうち「大量」「多量」は「に」と接続して副詞としての用法がみられたが、「少量」にはみられなかった。「たくさん」「少し」の統語的な特性には大きな違いはなかったが、「たくさんの」と「少しの」とでは共起する語に違いあった。また、「多い」には使用する際に注意すべき点が多いことが明らかになった。辞書では「多い」を形容詞、「多く」を副詞、名詞として見出し語を分けて説明しているが、学習者の立場から考えると、「多い」をイ形容詞として学ぶため「多く」が別の見出し語であることには気づかないだろう。学習辞典には「多く」だけでなく「多い」の中でも「多く」について記述があるほうが親切だろう。

語の使用には統語的な情報や共起する語の情報は不可欠である。現在の辞書でもよく読みこめば、それらの情報を見つけることができるかもしれないが、それでは学習者には不親切である。井上他（2006）が述べているように「表現の意味や使用上の注意事項を過不足なく記述し、「用例から察する」ことを学習者に強制しない」辞書の記述が求められる。

本調査では量的な観察が中心であったが、実際の意味記述をするのであればさらに用例を分析し、意味ごとに統語的な特徴や共起する語を抽出する必要がある。また、語を知ることには使用する場面の知識も必要となる。これらの分析を今後の課題としたい。

文献

井上優、有賀千佳子(2006).「これからの学習者用日本語辞書」日本語学, 25:7, pp.22-29.明治書院.
村木新次郎(2004).「漢語の品詞性を再考する」同志社女子大学日本語日本文学, 16, pp.1-35.
明治書院(2006).「特集　外国語学習者のための辞書」『日本語学』25:7, .
森下訓子(2006).「様態・量・程度を意味する和語系単語の統語的な特性について」同志社女子大学日本語日本文学, 18, pp.17-36.
『明鏡国語辞典』（2002-2008）大修館書店　電子辞書

書きことばらしさの判断と測定

井上次夫（小山工業高等専門学校一般科）†

On the Judgment and Measurement of Written Japanese Degree

Tsugio INOUE (Dept. General Education, Oyama National College of Technology)

１．はじめに

論説・論文等の論述文では、通常、常体で書きことば（書記言語・文字言語）が用いられる。その書きことばを「単語の文体」という観点から見ると、宮島（1972）の「日常語」「文章語」ということになり、井上（2010a）の「汎用体」「書記体」「文語体」ということになる。しかし、それら文体間の境界には必ずしも明確でない部分があり、井上の一連の研究（2009a・2009b・2010a）では国立国語研究所『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese, 以下、BCCWJ）を用い、その境界の画定に関する試みを行っている。

その後、井上（2010b）では「単語の文体」をいっそう明確に位置づけるため、それを書きことばらしさの程度として捉え直し、文体判断アンケートに基づく単語の文体値（アンケート文体値）を求めて、文体判定を試みた。そこで、本稿では井上（2010a）で提案したBCCWJ の白書コーパス及び Yahoo! 知恵袋コーパスに基づく単語の位置（極座標）を単語の文体値（コーパス文体値）として捉え直し、新たな基準による文体判定を試みる。そして、それら 2 種の文体値について検討し、単語の文体判定のあり方を考察する。

２．単語の文体の分類

表 1 は、井上（2010a）で示した「単語の文体の 5 分類」である。なお、各文体が書きことばらしさの程度に応じるように、新たに文体レベル 1～5 の項目を加えた。

表 1　単語の文体の 5 分類

	くだけた　←　うちとけた　←　普 通　→　あらたまった　→　かたい				
井上（2010a）	**卑俗体**	**口頭体**	**汎用体**	**書記体**	**文語体**
宮島（1977）	俗語	くだけた日常語	無色透明な日常語	あらたまった日常語	文章語
田中（1999）	会話的	話しことば的	一般	書きことば的	文語的
ことば	話し言葉	話し言葉	話し言葉・書き言葉	書き言葉	書き言葉
主な場面	私的会話	日常会話	公私等の別なし	公的発言・論説文	論文・詩歌
語例	あっし	あたし	わたし	わたくし	小生
	いろんな	いろいろな	さまざまな	多様な	多岐にわたる
	けど	だけど	けれども	しかし	しかしながら
	φ	なかでも	とくに	とりわけ	なかんずく
文体レベル	**1**	**2**	**3**	**4**	**5**

† inoue@oyama-ct.ac.jp

３．文体判断アンケート

井上（2010b）では、高専生及び大学生を対象に意味・用法が近似した単語群を与え、それらが単語の文体 5 分類のどのレベルに位置するかの判断を行う文体判断アンケートを実施し、そこから文体値を求めた。

３．１　高専生・大学生調査

単語の文体に関し、個人によりどのような文体判断が行われるかの実態を明らかにするため、平成 21 年度小山工業高等専門学校の 2 年生 111 人と白鴎大学「国語概説」の受講生 1～4 年生 108 人を対象に、表 2 の「単語の文体」5 分類案に基づき、調査語が 5 分類のどこに位置するかの文体判断を求めた。

表 2　「単語の文体」5 分類案

１ 卑俗体	２ 口頭体	３ 汎用体	４ 書記体	５ 文語体
低俗で野卑な感じ、主に私的な場で使用される	くだけた感じ、日常会話などで使用される	特別な感じはなく、公私の別なく使用される	あらたまった感じ、主に公的な場で使用される	高級でかたい感じ、論文・詩歌等で使用される
例）あっし いろんな	例）あたし いろいろな	例）わたし さまざまな	例）わたくし 多様な	例）小生 多岐にわたる
話し言葉　←　―　―　―　―		中　間	―　―　―　―	→　書き言葉

なお、アンケートでは高専生に 15 組 50 語、大学生に 14 組 50 語を示したが、共通する調査語は表 3 に示す 9 組 28 語であった。

調査後、各文体に書きことばらしさのレベル数値（卑俗体・1、口頭体：2、汎用体：3、書記体：4、文語体：5）を与える方法により、平均値を求めた。それを各語の「アンケート文体値」とみなし、下記の暫定的な判定式に基づき単語の文体判定を行った。

〈1〉判定式：1.0≦卑俗体＜1.8≦口頭体＜2.6≦汎用体≦3.4＜書記体≦4.2＜文語体≦5.0

表 3 では、1「言う」、3「いままで」、6「たくさん」、17「では」、19「およそ」、21「約」のように文体値が境界値と重なるもの（境界値の妥当性）、13「けれども」のように文体値が高専生（3.3）と大学生（3.6）で異なる結果、文体判定が汎用体と書記体に分かれるもの（インフォーマントの妥当性）、20「おおかた」のように書記体という判定自体に疑問が残るもの（アンケート法の妥当性）など注意が必要なものは存在するが、概ね妥当な文体判定になっているのではないかと思われる。

なお、文体値が境界値と重なる語については境界値としての最適値を求める研究が必要であり、高専生と大学生で文体判定が異なる語については一定の教養レベルの社会人を対象とする調査が必要であろう。また、文体判定そのものに疑問が残る語については、アンケートの対象者、規模、方法等に改善の余地があると同時に、コーパスに基づく実証的な方法が必要であると思われる。

表３　単語の文体判断Ⅰ（数値はアンケート文体値）

（判定式:1.0≦卑俗体＜1.8≦口頭体＜2.6≦汎用体≦3.4＜書記体≦4.2＜文語体≦5.0）

組	No.	語	高専	大学	文体	組	番	語	高専	大学	文体
①	1	言う	2.8	2.6	汎用	⑥	16	じゃあ	1.7	1.6	卑俗
	2	述べる	4.0	4.0	書記		17	では	3.2	3.4	汎用
②	3	いままで	2.6	2.6	汎用	⑦	18	だいたい	2.1	2.1	口頭
	4	これまで	3.2	3.5	汎用		19	およそ	3.3	3.4	汎用
③	5	いっぱい	1.6	1.6	卑俗		20	おおかた	3.5	3.9	書記
	6	たくさん	2.3	2.6	口頭		21	約	3.7	3.4	書記
	7	おおく	3.2	3.1	汎用	⑧	22	たまげる	1.5	1.3	卑俗
④	8	おんなじ	1.7	1.3	卑俗		23	びっくりする	2.1	2.0	口頭
	9	おなじ	2.7	2.6	汎用		24	おどろく	3.0	3.1	汎用
⑤	10	けど	1.7	1.7	卑俗		25	仰天する	3.8	3.9	書記
	11	でも	1.9	1.8	口頭		26	驚嘆する	4.6	4.7	文語
	12	しかし	3.3	3.5	汎用	⑨	27	ちょくちょく	1.5	1.4	卑俗
	13	けれども	3.3	3.6	書記		28	しばしば	3.4	3.6	書記
	14	しかしながら	4.4	4.3	文語	※文体判定は高専と大学の平均値による。					
	15	しかるに	4.7	4.8	文語						

３．２　漢検上級合格者調査

アンケート法の妥当性を高めるため、平成 22 年 12 月、漢検上級（1 級・準 1 級）合格者 32 人を対象に同様の文体判断アンケートを実施した。そこでの調査語 14 組 50 語のうち、上掲の表 3 と共通するものは③「いっぱい・たくさん・おおく」と⑦「だいたい・およそ・おおかた・約」の 2 組 7 語であった。表 4 に結果を示す。文体欄には、高専・大学・漢検それぞれの文体値に基づく文体名を順に記す。

表４　単語の文体判断Ⅱ（数値はアンケート文体値）

組	No.	語	高専	大学	漢検	文体
③	1	いっぱい	1.6	1.6	1.6	卑俗/卑俗/卑俗
	2	たくさん	2.3	2.6	2.3	口頭/汎用/口頭
	3	おおく	3.2	3.1	3.0	汎用/汎用/汎用
⑦	4	だいたい	2.1	2.1	1.9	口頭/口頭/口頭
	5	およそ	3.3	3.4	3.4	汎用/汎用/汎用
	6	おおかた	3.5	3.9	2.8	書記/書記/汎用
	7	約	3.7	3.4	4.0	書記/汎用/書記

表 4 では、1「いっぱい」、3「おおく」、4「だいたい」、5「およそ」の 4 語は高専・大学・

漢検 3 者の文体判定が一致している。しかし、2「たくさん」には口頭体と汎用体、6「おおかた」、7「約」には汎用体と書記体の幅がある。これに対しては、文体に幅がある語として認めていく考え方といずれかに決定すべきという考え方とがある。すなわち、言語の実態として文体判断に幅があることは認めざるを得ない一方で、教育・実用的立場からはいずれか 1 つに判定することは有用であると言わなければならないのである。

しかし、いま、単語の文体を相対的に捉える立場からすると、③は文体値の大小による文体式（書きことばらしさの大小を表す式）として「いっぱい＜たくさん＜おおく」のように書き表すことができ、上述の「たくさん」の文体判定の問題は保留される。⑦についても 6「おおかた」を除けば、文体式「だいたい＜およそ≦約」が成立する。なお、「おおかた」には名詞と副詞による文体差が予想されるのであり、大学生と漢検上級合格者とで文体値に 1.1 という大差があるため、後述のコーパス文体値による検討が必要である。

４．ＢＣＣＷＪ調査

４．１　ＢＣＣＷＪに基づくコーパス文体値

井上（2010a）では BCCWJ2009 年度版の白書コーパスと知恵袋コーパスを用いて単語の位置（極座標）を求めることを提案した。それに基づけば、語 W の文体値は、語 W の白書コーパスにおける出現数を x 、知恵袋コーパスにおける出現数を y とするとき、次のように表すことができる。

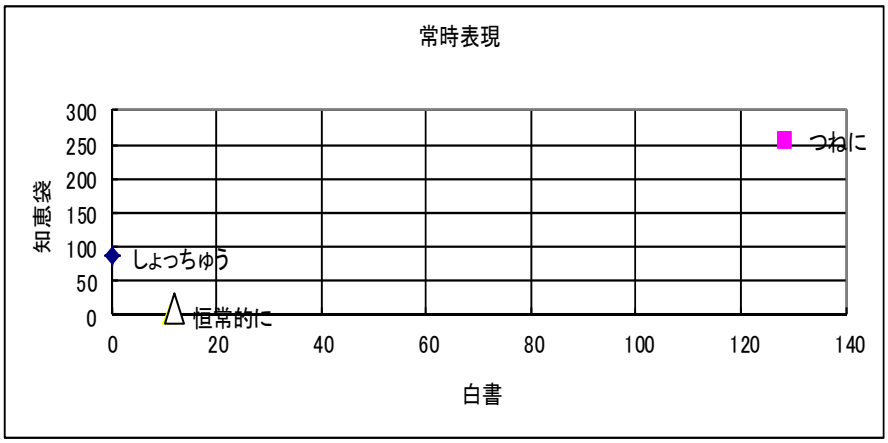
〈2〉語 W の文体値：W（a , b）　ただし、 $a = \tan\theta$,　$b = r/10$,　$r^2 = x^2 + y^2$

（a：書きことばらしさの程度, b：使用頻度）

例えば、常時表現「しょっちゅう・つねに・恒常的に」の場合、下の図表のようになり、文体式「しょっちゅう＜つねに＜恒常的に」の成立が実証される。

表５　常時表現

	語	白書	知恵袋	a	b
1	しょっちゅう	0	87	∞	8.7
2	つねに	128	257	2.0	28.7
3	恒常的に	11	1	0.1	1.1



以下、上記 a（＝tan θ）の値を「コーパス文体値」とみなし、それに基づく文体判定は暫定的に次の〈３〉を用いる[注1]。

〈３〉

0.0 ≦a＜0.2	→	E 層：レベル 5	書きことばらしさの程度：強い
0.2 ≦a＜0.6	→	D 層：レベル 4	
0.6 ≦a＜1.7	→	C 層：レベル 3	
1.7 ≦a＜5.7	→	B 層：レベル 2	
5.7 ≦a≦ ∞	→	A 層：レベル 1	弱い

これに基づけば、常時表現の文体は「しょっちゅう（A 層：レベル 1）＜つねに（B 層：レベル 2）＜恒常的に（E 層：レベル 5)」となる。なお、〈３〉の境界値の妥当性（最適な境界値の確定）については今後の課題である。

４．２　文体値と文体判定

ここでは、調査語に与えられる「アンケート文体値」と「コーパス文体値」という 2 種の文体値に基づく文体判定のあり方について具体例を用いて検討する。

(1)「いっぱい」類・「だいたい」類

高専生・大学生・漢検上級合格者へのアンケート調査語（表 4・数値は 3 者の平均）について、それぞれコーパス文体値（a）を求め文体判定を行ったものが表 6 である。

表６　単語の文体判断Ⅲ（アンケート文体とコーパス文体）

組	No.	語	コーパス (a)	レベル	アンケート	文体名
③	1	いっぱい	176.7	1	1.6	卑俗
	2	たくさん	374.5	1	2.4	口頭
	3	おおく	0.5	4	3.1	汎用
⑦	4	だいたい	85.5	1	2.0	口頭
	5	およそ	0.8	3	3.4	汎用
	6	おおかた	1.3	3	3.4	汎用
	7	約	0.2	4	3.7	書記

まず③では、コーパス文体値に基づけば「たくさん≦いっぱい＜おおく」（「≦」：文体レベルは同じだが、右項のほうが書きことばらしい）となり、これはアンケート文体値に基づく「いっぱい＜たくさん＜おおく」と一部異なる。ここでは、いずれの文体式が適当かの判断は保留し、ひとまず「いっぱい・たくさん＜おおく」（「・」：2 語は文体レベルが同じ）と表記しておくことにする。

次に⑦では、コーパス文体値に基づけば「だいたい＜おおかた≦およそ＜約」であり、これはアンケート文体の「だいたい＜おおかた・およそ＜約」とほぼ一致する。

[注1] ここでは、θ＝10°、30°、60°、80°を暫定的な境界とした。また、b の取り扱いについては保留。

(2) その他の語

表 3 の③⑦を除く他の 21 語について検討する。表 7 にはコーパス文体値（a : 書きことばらしさの程度, b : 使用頻度）とそれに基づく文体レベル、アンケート文体値（高専・大学の平均値）とそれに基づく文体名を示す。

表 7　単語の文体判断Ⅳ（アンケート文体とコーパス文体）

組	No.	語	コーパス文体値		文体レベル	アンケート文体値	文体名
			a	b			
①	1	言う	50.8	406.6	1	2.7	汎用
	2	述べる	0.1	13.0	5	4.0	書記
②	3	いままで	32.8	121.3	1	2.6	汎用
	4	これまで	0.2	91.1	4	3.4	汎用
④	5	おんなじ	∞	2.2	1	1.6	卑俗
	6	おなじ	5.3	460.2	2	2.5	汎用
⑤	7	けど	∞	6.1	1	1.7	卑俗
	8	でも	2878.0	2878.0	1	1.8	口頭
	9	しかし	1.2	135.4	3	3.4	汎用
	10	けれども	4.5	0.9	2	3.5	書記
	11	しかしながら	0.1	135.4	5	4.4	文語
	12	しかるに	0.0	0.5	5	4.8	文語
⑥	13	じゃあ	∞	26.3	1	1.7	卑俗
	14	では	1.2	2024.4	3	3.3	汎用
⑧	15	たまげる	∞	0.1	1	1.4	卑俗
	16	びっくりする	∞	1.9	1	2.1	口頭
	17	おどろく	40.0	4.0	1	3.1	汎用
	18	仰天する	∞	0.1	1	3.9	書記
	19	驚嘆する	―	―	―	4.7	文語
⑨	20	ちょくちょく	∞	1.5	1	1.4	卑俗
	21	しばしば	1.2	4.5	3	3.5	書記

表 7 からは以下の点が指摘できる。文体式は後述（5.参照）。

・文体レベル 1 の 11 語…アンケート文体値では卑俗・口頭体。ただし、1・3・17 が汎用体、18 が書記体になる。
・文体レベル 2 の 2 語…アンケート文体値では 6 は汎用体、10 は書記体になる。
・文体レベル 3 の 3 語…アンケート文体値では汎用体・書記体になる。
・文体レベル 4 の 1 語…アンケート文体値では汎用体になる。
・文体レベル 5 の 3 語…アンケート文体値では書記体・文語体になる。
・文体レベル不明 1 語…19 は BCCWJ の「白書」「Yahoo!知恵袋」コーパスに出現しない。

以上、少数の例からではあるが、文体レベルを判定する〈3〉の境界値の設定（4.1 参照）には課題が残る。一方、19 のような BCCWJ に出現しない語についてはアンケート文体値（3.1 参照）を有効に活用できると言えるだろう。

今後、基本的にはコーパス文体値に基づく文体判定をその判定法の精度を高めながら行っていく一方で、コーパスに出現しない語や出現しにくい語については被験者の年代・教養レベル、方法に留意しながら個別にアンケート文体値を求め、多くの語について適切な文体判定が行えるようにしていくことが望ましいと考える。

５．まとめ

単語の書きことばらしさの程度の判定には個人差が伴うため、その判定には幅を有するケースが存在する。そこで、アンケート文体値及びコーパス文体値に基づくそれらの単語の文体判定及び文体レベルの位置づけを図ろうとした。そして、次の結果を得た[注2]。

〈4〉　書きことばらしさの測定と判定

【コーパス文体値】			強	【アンケート文体値】		
0.0 ≦a＜0.2	→	E 層：レベル 5	▲	文語体	←	4.2＜文語体≦5.0
0.2 ≦a＜0.6	→	D 層：レベル 4	｜	書記体	←	3.4＜書記体≦4.2
0.6 ≦a＜1.7	→	C 層：レベル 3	｜	汎用体	←	2.6≦汎用体≦3.4
1.7 ≦a＜5.7	→	B 層：レベル 2	｜	口頭体	←	1.8≦口頭体＜2.6
5.7 ≦a≦ ∞	→	A 層：レベル 1	▼	卑俗体	←	1.0≦卑俗体＜1.8
			弱			

例　①言う＜述べる　②いままで＜これまで　③いっぱい・たくさん＜おおく　④おんなじ＜おなじ　⑤けど＜でも＜けれども＜しかし＜しかしながら＜しかるに　⑥じゃあ＜では　⑦だいたい＜おおかた＜およそ＜約　⑧たまげる＜びっくりする＜仰天する＜おどろく＜驚嘆する　⑨ちょくちょく＜しばしば

上記 2 種の文体値の関係は必ずしも 1 対 1 に対応するものではない。また、その境界値については暫定的なものに過ぎない。さらに、コーパス文体値の使用頻度（b）についても検討が必要である。しかしながら、それらの文体値を用いて、意味・用法が近似した語群の書きことばらしさを相対的に位置づける文体式を導くことは可能となった。

６．おわりに

従来、単語の文体は連続的、相対的なものであり、個人差が生ずるものと指摘されてきた。そこで、本稿では個人差の実態をアンケートにより明らかにしたうえで高専生・大学生・漢検上級合格者の文体判断の平均値によりそれを「アンケート文体値」とした。また、BCCWJ の白書・知恵袋コーパスにおける出現数調査により、書きことばらしさを 5 層とする文体の位置（極座標）に基づく「コーパス文体値」を示した。

これら 2 種の文体値の判定基準のうち、それぞれ境界となる数値の適切性（境界の最適値）、また 2 種の文体値の相互関係についてはなお不透明な部分も残されており今後の課題

注2　例のうち、2 種の文体値が異なる⑤「けれども」「しかし」、⑥「おおかた」「およそ」の判定はコーパス文体値によった。また、⑧については 2 種の文体値を勘案し、筆者が判定した。

であるが、それぞれ求められた文体値により単語の文体（書きことばらしさの程度）の相対的位置づけを表す文体式の導出については根拠が与えられたと言える。このことは任意の 1 語の文体値を明らかにすることに結びつくものでもあるため、今後、いっそう多くの単語群を対象にそれら文体値に基づく文体判定法の精度を高めていくことが必要である。

文献

井上次夫(2009a).「日本語コーパスに基づく『語の文体』の明確化」文部科学省科学研究費特定領域研究「日本語コーパス」『平成 20 年度公開ワークショップ サテライトセッション予稿集』pp.109-118.

井上次夫(2009b).「論説文における語の文体の適切性について」『日本語教育』141, pp.57-63.

井上次夫(2010a).「BCCWJ を用いた『語の文体』の位置づけ」文部科学省科学研究費特定領域研究「日本語コーパス」『平成 21 年度公開ワークショップ サテライトセッション予稿集』pp.91-100.

井上次夫(2010b).「単語の文体意識についてー話しことばと書きことばの区別ー」『全国大学国語教育学会第 118 回大会発表要旨集』pp.129-132.

荻野綱男(2006) .「WWW による単語の文体差の研究」『日本語学会 2006 年度秋季大会予稿集』,pp.139-146.

菊沢季生(1930).「国語位相論」『国語科学講座Ⅲ』,明治書院, pp.2-67.

後藤斉(2001).「日本語コーパス言語学と語の文体レベルに関する予備的考察」『東北大学文学研究科研究年報』pp.200-214.

島本基(1990).「語の位相」『講座日本語と日本語教育』7,明治書院, pp.298-322.

田中章夫(1978).『国語位相論』,明治図書.

田中章夫(1999).『日本語の位相と位相差』,明治図書.

徳川宗賢・宮島達夫(1980).『類義語辞典』20 版 ,東京堂出版.

前川喜久雄(2008).「話し言葉と書き言葉」『日本語学』27-5, pp.23-33.

宮島達夫(1972).『動詞の意味・用法の記述的研究』(国立国語研究所報告 43),秀英出版,pp.708-732.

宮島達夫(1977).「単語の文体的特徴」『松村明教授還暦記念国語学と国語史』,明治書院, pp.871-903.

宮島達夫(1988).「単語の文体と意味」『国語学』154, pp.78-88.

宮島達夫(2008).「文章の文体と単語の文体－国研コーパスを利用してー」『近代語研究』14, 武蔵野書院, pp.375-386.

【付記】

本稿は、科学研究費補助金（基盤研究（C））「論述文における適切な単語使用のための教材開発と指導法の研究」（研究代表者 井上次夫，課題番号 22531046，平成 22-24 年度）における研究成果の一部である。

書き言葉におけるダ体とデアル体の混用への考察

徐　衛（中国・蘇州大学外国語学院）

A Corpus-Based Study on the Incorporation of DA-Style and DEARU-Style in Written Japanese

Xu Wei（School of Foreign Languages, Soochow University, P.R. China）

キーワード：書き言葉，ダ体，デアル体，混用，コーパス

１　はじめに

現代日本語書き言葉の文体は、文末表現からデス・マス体、ダ体、デアル体に分けられる。その中でダ体とデアル体は「常体」として一括りに扱われることがあるが、混用が見られる。とくに、新聞・雑誌にはその混用の例が多いようである。例えば、

○　残念ながら、今回の解禁は、弊害への対策が間に合わないこと、各党の合意を優先したことなどから、まだまだ範囲が狭く不十分である。

メールの禁止やツイッターの自粛もそうだが、最大の問題は一般の有権者がネットを通じて選挙運動をするのを認めなかったことだろう。自分の支持する候補への投票を呼びかけることはできないのである。だが虚偽記載や中傷はともかく、支援の呼びかけのどこに問題があるのか。次の段階として、参院選後にはそうした運動を含め解禁の範囲を広げるべきである。

当面、参院選に向けて心配なのは、どんな書き込みが選挙運動に当たるのか、普通の人にはわかりにくい点だ。違反をおそれて書き込みをためらう人が出てくる。それではネットの効果も十分に表れないだろう。どんな書き込みは許され、どこからが違反なのか、わかりやすい指針づくりを急いでほしい。

（『朝日新聞』【社説】2010 年 5 月 27 日より）

文体の統一は日本語教育で強調され、規範的なものとして受け入れられているようで、議論されることが少ないが、ダ体とデアル体の混用について、日本語教育学会（2005）では、「『である体』の文章は『だ体』の文章よりも形式的で改まった印象を与えるが、文末の単調さを避けるために両者を混用する例も多い（p.358）」と書いてある。

大野早苗（2010）では、1940 年代から現在までの社説の中の社説のダ体とデアル体の使用状況を調査し、社説の文体はデアル体からダ体へと変遷したという結論を出した。それでは、社説と同じく、一般的常体を用いて書かれた現代日本語の文章の文体はどうなっているだろうか。ダ体とデアル体が実際にどこまでどのように混用しているかを明らかにさせるには、各ジャンルの文章の文体に関する実証的な研究は不可欠であろう。

本研究はデアル体が基調となっていた文章にダ体が混入する状況を考察することを目的とする。具体的には、BCCWJ モニター公開データを利用して、ジャンルによる各種類の「書籍」（生産実態サブコーパス）よりサンプルを抽出し、デアル体文章内に出現するダ体文の用例数およびデアル体文の用例数を調査し、両者の比率、性質などを考察してみたい。

２　調査の資料と対象

ここで資料として用いるのは、BCCWJ モニター公開データ（2009 年度版）の生産実態（出版）サブコーパスの中の「書籍」サンプルである。調査にあたっては、文末に出現するもののみを対象とし、デアル体はデアル、デアロウ、デアッタ、ダ体はダ、ダロウ、ダッタをカウントした。「ダロウカ、デアロウカ」はそれぞれダロウ、デアロウに含めて集計したが、文末の「デハアル、デモアル」など、取り立て助詞を介入させる形式は除外した。

まず、各サンプル中の文末「デアル」の用例数を調査した。1 哲学、2 歴史、3 社会科学、4 自然科学、5 技術・工学、6 産業、7 芸術・美術、8 言語、9 文学という 9 種類のジャンルの中で、「である」の用例数がもっとも多いサンプルをそれぞれ 10 本抽出し、デアル体が基調となっていた「書籍」サンプルを確認した。抽出されたサンプルの基本的状況は表 1 のようである。

表 1　抽出されたサンプルの基本的状況

	サンプル数	総文字数	文末「である」最高数	文末「である」最少数
1 哲学	10	99161	74	46
2 歴史	10	98039	57	37
3 社会科学	10	108135	90	61
4 自然科学	10	118026	71	41
5 技術・工学	10	100160	78	27
6 産業	10	99028	59	30
7 芸術・美術	10	93244	61	26
8 言語	10	85255	48	23
9 文学	10	108788	78	37
合計	90	909836	／	／

資料における談話の引用などに含めた用例はごく少ないが、ある場合、その用例数を除外した。文体混用の研究はある文章・書籍に限るデータの集計などの調査が意義があるので、今回は上記の資料だけに基づいて調査を行った。

３　調査の結果

調査結果を各サンプルの文末「デアル」の用例数の多少を順序として表 2 に示す。

表 2　ジャンルによる各サンプル中の調査対象の用例数

	サンプル	デアル	デアロウ	デアッタ	ダ	ダロウ	ダッタ
1 哲学	1	74	1	6	7	1	2
	2	65	5	42	1	0	1
	3	62	8	0	3	1	0
	4	56	0	18	0	2	1
	5	55	4	16	0	0	0
	6	53	5	3	9	6	1
	7	52	3	1	0	1	0
	8	50	15	4	5	0	0
	9	46	7	1	0	0	0
	10	46	0	0	0	2	3
	計 1	559	48	91	25	13	8

2 歴史	1	57	0	4	0	5	3
	2	55	5	23	0	1	1
	3	53	2	2	14	7	11
	4	51	4	14	4	3	0
	5	49	0	3	2	0	0
	6	47	1	3	0	0	22
	7	46	0	0	0	0	0
	8	41	7	5	0	1	0
	9	37	0	1	12	3	6
	10	37	0	0	11	0	0
	計 2	473	19	55	43	20	43
3 社会科学	1	90	4	4	0	1	0
	2	83	2	1	0	1	0
	3	79	6	4	0	1	0
	4	66	12	2	3	0	0
	5	66	3	0	14	1	0
	6	66	3	0	4	5	0
	7	66	2	5	9	6	1
	8	64	9	4	7	1	1
	9	63	7	3	9	5	0
	10	61	1	0	17	14	0
	計 3	704	49	23	63	35	2
4 自然科学	1	71	1	6	19	5	1
	2	62	2	2	0	3	0
	3	62	1	0	35	6	1
	4	54	2	0	0	0	0
	5	54	0	0	1	1	0
	6	47	7	0	0	0	0
	7	47	0	2	31	3	5
	8	45	6	2	11	3	0
	9	43	8	2	0	9	0
	10	41	5	3	10	0	0
	計 4	526	32	17	107	30	7
5 技術・工学	1	78	9	0	0	0	12
	2	59	0	17	8	3	21
	3	59	0	0	1	6	0
	4	45	7	11	6	3	1
	5	44	0	0	4	0	0
	6	34	1	2	4	3	4
	7	33	2	1	9	2	2
	8	32	0	7	2	0	1

	9	30	3	0	0	0	0
	10	27	4	0	0	7	1
	計 5	441	26	38	34	24	42
6 産業	1	59	1	0	6	0	1
	2	55	1	15	7	2	4
	3	48	0	1	8	0	0
	4	47	0	6	24	15	26
	5	46	2	2	0	1	0
	6	44	0	8	25	0	5
	7	40	3	3	0	2	0
	8	40	0	2	5	0	1
	9	32	1	1	20	5	12
	10	30	0	1	0	0	0
	計 6	441	8	39	95	25	49
7 芸術・美術	1	61	3	7	0	0	0
	2	45	0	0	0	6	4
	3	42	23	17	3	2	1
	4	41	1	1	10	5	0
	5	32	0	16	3	3	0
	6	31	2	2	0	0	0
	7	29	0	1	2	2	3
	8	29	0	0	5	2	6
	9	28	3	16	3	0	0
	10	26	0	2	11	4	4
	計 7	364	32	62	37	24	18
8 言語	1	48	2	1	0	1	0
	2	36	5	0	0	0	0
	3	34	0	0	0	0	0
	4	32	4	2	0	0	0
	5	31	0	1	5	0	0
	6	28	1	0	0	2	0
	7	27	0	1	0	3	0
	8	27	0	0	25	3	1
	9	24	0	6	15	1	4
	10	23	2	1	0	0	0
	計 8	310	14	12	45	10	5
9 文学	1	78	2	1	23	2	0
	2	67	15	2	0	1	0
	3	52	1	3	23	5	9
	4	46	0	7	1	1	0
	5	45	10	18	0	0	1

	6	43	4	9	13	2	3
	7	40	1	5	14	4	5
	8	38	8	3	8	7	0
	9	38	6	2	14	7	0
	10	37	5	0	0	1	10
	計 9	484	52	50	96	30	28
合　計		4302	280	387	545	211	202

次の表 3 は、デアル体用例数とダ体用例数の集計結果と両者の比率である。デアル体用例数は「デアル、デアロウ、デアッタ」用例数の集計で、ダ体用例数は「ダ、ダロウ、ダッタ」用例数の集計である。表 4 には、デアルとダ、デアロウとダロウ、デアッタとダッタの集計用例総数と両者の比率を示した。

表 3　デアル体用例数とダ体用例数の比率

	デアル体		ダ体		計
	用例数	%	用例数	%	用例数
1 哲学	698	93.8	46	6.2	744
2 歴史	547	83.8	106	16.2	653
3 社会科学	776	88.6	100	11.4	876
4 自然科学	575	80.0	144	20.0	719
5 技術･工学	505	83.5	100	16.5	605
6 産業	488	74.3	169	25.7	657
7 芸術･美術	458	85.3	79	14.7	537
8 言語	336	84.8	60	15.2	396
9 文学	586	79.2	154	20.8	740
合計	4969	83.8	958	16.2	5927

表 4　デアルとダ、デアロウとダロウ、デアッタとダッタの比率

	デアル	ダ	デアロウ	ダロウ	デアッタ	ダッタ
用例総数	4302	545	280	211	387	202
比率	100 : 13		100 : 75		100 : 52	

表 2 を見ると、デアル体が基調となっていた「書籍」の 90 サンプルにおいて、ダ体文が混入していないのが 12 サンプルで 13.3%しか占めていない。ほかの 78 サンプル（86.7%）は多かれ少なかれダ体文が混入している。ダ体文の文末形式を見ると、ダ、ダロウ、ダッタの用例総数はそれぞれ 545 例、211 例、202 例で、ダで終わる文が半数以上占め、ダロウ文とダッタ文がほぼ同じであることが分かった。それに対して、デアルで終わる文（4302 例）は圧倒的に多いが、デアロウで終わる文（280 例）はデアッタ文（387 例）より少ない。また、表 4 に示されたように、混入したダ文の総数が多いが、推量や過去・完了を表す場合、ダロウ文とダッタ文のほうが出現しやすいと言えよう。

表 3 を見ると、ジャンルによる用例数およびその比率は異なっていて、ある程度の差が見られる。ダ体用例数の比率は「哲学」（6.2%）、「社会科学」（11.4%）のほうがやや低いが、「産業」（25.7%）、「文学」（20.8%）、「自然科学」（20.0%）のほうが割合に高い。しかし、

表 2 のデータ全体から見ると、同一ジャンルの中で、「ダ、ダロウ、ダッタ」の用例数はサンプルによって多かったり少なかったりしていて、その変動に傾向性が見られない。それで、本研究の資料調査の限りではジャンルによる文体の混用に関しては容易に結論が出られないと思われ、今後一層の詳細な研究が期待される。

４　混入しやすいダ体文の性質

ダ体文（ダ、ダロウ、ダッタで終わる文）の文末形式を考察してみると、次のようにまとめられる。

①ダッタで終わる文は、文章内容によって「名詞＋ダッタ」の形式が最も多い。

②ダで終わる文は、文末に「～のだ」の形式が最も多い。全面的な統計はしていないが、「歴史」サンプルの集計 43 用例のうち、「～のだ」の用例が 21 である。また、「社会科学」サンプルの集計 63 用例のうちは 40 例、「文学」サンプルの集計 102 用例のうちは 25 例である。それ以外、出現しやすいのは「からだ/そうだ/ようだ/べきだ、ことだ/わけだ/はずだ」および「重要だ/必要だ/駄目だ」のような形容動詞の用例である。

③ダロウで終わる文は、絶対優位の文末形式はないが、「いい/よいだろう、ない/あるだろう、の/ことだろう、いえる/わかる/できるだろう」のような用例が比較的に多い。

５　おわりに

本研究は単にデアル体が基調となっていた文章にダ体がどのように混入しているかについて、コーパスよりサンプルを抽出して調査を行ってみたが、少なくとも以下のことが分かった。①明確なデアル体が基調である文章（あるいは書籍）にも、その大部分が（本研究では 86.7%）ダ体文の混入が見られる。②混入状況は文章の性質より異なるが、混入したダ文の総数は最も多いのに対して、推量や過去・完了を表す場合、ダロウ文とダッタ文のほうが出現しやすい（ダロウ文が最も出現しやすい)。③ジャンルによる用例数およびその比率は異なっていて、ある程度の差が見られるが、本研究の資料調査の限りでは容易に結論が出られない。④ダで終わる文は、「のだ」の用例がもっとも多いと同時に、「ことだ/ものだ/はずだ/わけだ」、「からだ/ようだ/そうだ/べきだ」および「重要だ/必要だ/駄目だ」のような形容動詞の用例も多く見られる。ダロウで終わる文は、「いい/よいだろう、ない/あるだろう、の/ことだろう、いえる/わかる/できるだろう」のような用例が割合に多い。

以上、コーパスを利用した部分的な調査および結果分析を述べてきたが、現代日本語書き言葉におけるダ体とデアル体の混用については、不明なところが多いだろう。例えば、ダ体が基調となっていた文章にデアル体が混入する状況はあるかどうか。ダ体かデアル体かがあいまいな文章も相当あると思われるが、その場合、文章の文体をどのように確認したらよいのか、またその混用への考察をどのようにしたらよいのか等。以上のようなダ体とデアル体の混用問題に関するより詳細な調査、分析を行うことを今後の検討課題としたい。

主な参考文献

日本語教育学会（1982）『日本語教育事典』大修館書店
田中章夫（1999）『日本語の位相と位相差』明治書院
日本語教育学会（2005）『新版 日本語教育事典』大修館書店
三牧陽子（2007）「文体差と日本語教育」『日本語教育』(134)：58-67.
大野早苗（2010）「社説の文体―デアル体からダ体へ」表現学会『表現研究』91 号

コーパスを用いた外来語サ変動詞の分析

―「カットする」を例として―

茂木俊伸（鳴門教育大学 大学院学校教育研究科）†

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

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

１．はじめに

本研究では，『現代日本語書き言葉均衡コーパス』モニター公開データ 2009 年度版（以下，単に BCCWJ と呼ぶ）のデータに基づき，サ変動詞としての外来語の使用実態について分析を行う。

以下では，まず文法的観点を取り入れた外来語研究の必要性について述べたうえで（第 2 節），BCCWJ における外来語サ変動詞の使用状況を確認する（第 3 節）。さらに事例研究として，多義的なサ変動詞「カットする」を取り上げ，語義と構文的特徴との対応という観点からこの語の詳細な記述が可能であることを示す（第 4 節）。

２．「外来語の文法」研究の必要性

外来語研究の課題として，和語や漢語と比べ，基本的な語の研究が遅れているという点が挙げられる（例えば，石野 1996，金 2006a,b，宮田 2007）。これは，類義の和語や漢語との使い分けを含めた意味的側面だけでなく，外来語が文中にどのように現れるのか，という文法的側面の分析に関しても同様である。

外来語の文法的ふるまいは，基本的に漢語と同様であり，「する」を付加すれば動詞として，「だ／な」を付加すれば形容動詞として用いられる（楳垣 1963:27）[1]。近年，大規模データを用いて，特定の品詞の外来語に焦点を当てた分析（新川・加藤 2005，森下 2007 など）も見られるようになったものの，個別の語の具体的な分析は進んでいない[2]。

一方で，外来語は，日本語教育における学習上の困難点になっているにもかかわらず，十分な手当てがなされていない，ということが指摘される（澤田 1993，加藤・新川 2002，中山ほか 2008 など）。例えば，多くの日本語教材では例文と辞書的意味（別語への言い換え）が提示されているにとどまっており，文法的側面にはあまり配慮が見られない。

このような状況において，特に構文レベルの文法的な情報を含めた外来語の記述を蓄積していくことは，日本語における外来語の実態を明らかにするという点でも，辞書編集や

† E-mail: tmogi@naruto-u.ac.jp

1 ただし，漢語が単独で副詞として用いられるのに対し，外来語には同様の用法がほとんど見られない。

2 このような「後回し」状態には，日本語の語彙における外来語の割合が比較的低いことが影響していると考えられる。例えば，基本語彙のサ変動詞（相澤 1993）における外来語の割合は 3.1%である。

教育，自然言語処理といった応用分野への基礎資料を提供するという点でも有意義であると言える。本研究は，外来語サ変動詞を取り上げて，この問題に取り組むものである。

３．BCCWJ における外来語サ変動詞

BCCWJ から外来語サ変動詞を抽出すると，延べ語数 18,094 語，異なり語数 1,421 語という結果が得られた[3]。次のページの《表 1》は，頻度上位 50 語とともに，それらの語の辞書等における記述を示したものである[4]。

このうち，頻度上位 20 語は，次のような語である。

> クリックする，チェックする，スタートする，インストールする，コピーする，コントロールする，メールする，カバーする，クリアする，カットする，イメージする，ダウンロードする，セットする，プレゼントする，リードする，アピールする，キスする，チャレンジする，アクセスする，ヒットする

下線を引いた語は，コンピュータ用語（もしくはその用法）が多く見られたものである。「クリックする」の例が突出して多い理由は，ソフトウェアのマニュアル本において操作手順の説明に繰り返し用いられているためであり，書籍の例が約 900 例を占める。また，「インストールする」「メールする」「ダウンロードする」は，Yahoo! 知恵袋の例が約 9 割を占めている。

また，《表 1》からは，外来語辞典や国語辞典では，下線の語がサ変動詞用法を持つことが明記されていない場合があることが分かる。荻野（1996）は，コーパスに現れる漢語サ変動詞で，新しい口語的用法や専門用語（用法）が辞書に記載されていない例を指摘しているが，外来語サ変動詞においても同様の傾向が見て取れる。

次に，澤田（1993）が提示している「基本外来語」等のサ変動詞のリストとの対応関係を見ると，上の 20 語と重複しないのは，下線の語に限定されることが分かる。《表 1》のように，頻度上位 50 語との重複が 39 語見られることを考え合わせても，BCCWJ の資料性あるいは時代の変化が影響していると考えられる語を除けば，おおむね基本的な外来語が抽出できているものと考えられる。

次節では，上の高頻度語から，特定の分野に偏らない一般的な動詞であり，かつ，多義的である，という条件を満たす「カットする」を分析対象として選択し，詳細な記述を試みる[5]。

---

[3] 用例の検索には，BCCWJ に同梱されている「全文検索システム『ひまわり』BCCWJ パッケージ」を使用した。検索は，「カタカナ＋「する」の活用形」という条件で行い，「カタカナ＋る」「カタカナ＋できる」の形は含めていない。また，異なり語数は，(1) 「バ／ヴァ」や長音記号の有無のような表記のゆれがある場合もそれぞれ 1 語とする，(2) 複合語（例：カット&ペーストする，パスカットする）もそれぞれ 1 語とする，(3) 混種語（例：ドタキャンする）も含める，(4) ヲ格が挿入された形の方が一般的な語（例：「サッカーする」）も含める，という形でカウントしている。

[4] 《表 1》の記号は，「○」…サ変動詞用法があることを（辞書によっては自他の記述も）明記，「△」…サ変動詞用法以外の用法（名詞等）のみ掲載，「×」…見出し語になし，ということを表す。

[5] 丸山（2010）には，『岩波国語辞典（第 7 版）』の記述とコーパスへの出現頻度とを対応させた「カバーする」の分析が見られる。

《表 1》BCCWJ における外来語サ変動詞の頻度上位語と辞典等の記述

No.	見出し語	用例数	コンサイス4 2010	岩波国語7 2009	明鏡国語 2002	佐々木 2001	澤田 1993
1	クリック	1,385	△	○(他)	○(他)	×	×
2	チェック	815	○	○(自他)	○(他)	○(他)	○
3	スタート	553	○	○(自)	○(自)	○(自)	○
4	インストール	338	○	○(他)	○(他)	×	×
5	コピー	334	○	○(他)	○(他)	○(他)	○
6	コントロール	313	○	○(他)	○(他)	○(他)	○
7	メール	304	△	△	△	×	×
8	カバー	284	○	○(他)	○(他)	○(他)	○
9	クリア	261	○	○(他)	○(他)	○(他)	○
10	カット	252	○	○(他)	○(他)	○(他)	○
11	イメージ	243	○	○(他)	○(他)	○(他)	○
12	ダウンロード	240	△	○(他)	○(他)	×	×
13	セット	220	○	○(他)	○(他)	○(他)	○
14	プレゼント	206	○	○(他)	○(他)	○(他)	○
15	リード	206	○	○(自他)	○(他)	○(自他)	○
16	アピール	205	○	○(自他)	○(自他)	○(他)	○
17	キス	202	○	○(自)	○(自)	×	○
18	チャレンジ	197	○	○(自他)	○(自)	○(自)	○
19	アクセス	194	△	○(自)	○(自)	×	×
20	ヒット	174	○	○(自)	○(自)	○(自)	○
21	リラックス	169	○	○(自)	○(自)	○(自)	○
22	アップ	167	○	○(自他)	○(自他)	○(自他)	○
23	キャンセル	167	○	○(他)	○(他)	○(他)	×
24	サポート	164	○	○(他)	○(他)	○(他)	○
25	ノック	150	○	○(他)	○(自)	×	○
26	メモ	142	○	○(他)	○(他)	○(他)	○
27	エスカレート	134	○	○(自)	○(自)	○(自)	○
28	デザイン	132	○	○(他)	○(他)	×	○
29	リンク	126	○	○(他)	○(他)	○(他)	×
30	シフト	123	○	○(自)	○(自)	×	○
31	デビュー	123	○	○(自)	○(自)	×	○
32	バックアップ	123	○	○(他)	○(他)	○(他)	○
33	ストップ	122	○	○(自他)	○(自他)	○(自他)	○
34	セックス	112	○	○(自)	△	×	×
35	クローズアップ	110	○	○(他)	○(他)	○(他)	○
36	アドバイス	108	○	○(他)	○(他)	○(自)	○
37	オープン	107	○	○(自他)	○(自他)	○(他)	○
38	マスター	104	○	○(他)	○(他)	○(他)	○
39	エッチ	101	○	○(自)	○(自)	×	×
40	バイト	92	△	△	○(自)	×	○
41	コメント	90	○	△	○(自他)	○(自)	○
42	デート	89	○	○(自)	○(自)	×	○
43	ドラッグ	89	△	△	○(自)	×	×
44	フォロー	88	○	○(他)	○(他)	○(他)	○
45	プラス	87	○	○(他)	○(他)	○(他)	○
46	キャッチ	85	○	○(他)	○(他)	○(他)	○
47	カウント	83	○	○(他)	○(他)	×	○
48	マッチ	83	○	○(自)	×	○(自)	○
49	アプローチ	81	○	○(自)	○(自)	○(自)	×
50	サイン	81	○	○(自)	○(自他)	×	○

４．事例研究―「カットする」の分析―

以下では，BCCWJ における用例をもとに「カットする」の意味分析を行い（4.1 節），さらにこの動詞がどのような成分や文末形式を伴って現れるのかを詳しく見ていく（4.2 節）。

4.1　意味的特徴

BCCWJ における「カットする」の用例を意味的に分類すると，次の《表 2》のようになる。

これらに共通する「カットする」の基本的な意味は，「ある形のものに切れ目を入れることによって，別の（より小さい）形にする」のようなものであると考えられるが，ここでは辞書等も参考にしながら，おおよその用法ごとに，語義[1]～[4]を立てた。

《表 2》「カットする」の意味分類

語義		用例数	
[1] 切る	[食べ物]を 小さく加工する／切り分ける [食べ物の皮]を 剥く	46	98
	[細長いモノ]を 短くする [薄いモノ]を 小さくする	52	
[2] 髪を切る	[(髪の)毛]を 切って整える	33	33
[3] 減らす	[映像／文章／項目]を 削除(して短縮)する	45	103
	[お金／数]を 減らす	58	
[4] さえぎる	[紫外線／光]を 遮断する	12	14
	[ボール／パス]を 取る	2	
[5] その他		4	4
	計:		252

それぞれの具体的な用例を，次に挙げる（出典はサンプル ID によって表す）。

(1)　語義[1]：切って｛小さく／短く｝する。

a. 野菜なども使う分だけ皮をむいて使う大きさにカットし、密閉容器に入れていくといい。（PB25_00290）

b. 皮をナイフで薄くカットします。果肉は食用とします。（LBm5_00033）

c. 編み終わりは、ひもの余分をカットし、内側にボンドでとめる。（LBs5_00032）

d. 先ほど受験用写真をきれいにカットするコツを質問した者です。（OC10_00401）

(2)　語義[2]：髪の毛を切って整える。

避暑地のお嬢さんらしく、帽子に隠れた髪は、学校にいる時より短くカットされている。（LBr9_00274）

(3)　語義[3]：余剰な部分について｛減らす／削る｝。

a. 読んでみると、議事録からは『あー』とか『うー』はカットされている。（PB26_00141）

b. 公務員の数は半分になり、給与は３割カット、退職金は半減、または全額カットされる。（PB53_00657）

(4)　**語義[4]：さえぎって届かなくする。**

a. 前述の通り遮光板で上方の光が綺麗にカットされるため、マルチリフレクターより暗く感じる。（OC06_03304）

b. ディフェンスしてて相手の蹴ったボールを胸でカットした、とかならまだしも。（OC14_03488）

(5)　**その他**：

a. **＜斜めに切るように打つ＞**　バレーボールのボールなどを打つときにボールの下部を手のひらでカットするようにするのと、（PB4n_00054）

b. **＜解除する＞**　中古車で、リミッターがカットされている車は売られているものなのでしょうか？（OC06_00923）

語義[2]の「髪を切る」は，語義[1]と「細長いモノを短くする」という点で共通するため，意味的には統合することも可能である。ここでは，後述のように語義[2]の例が「介在性」という特徴を持つことから独立させた。

語義[3]は，辞書類の記述では一般的に「文章やお金の一部を削る」こととされているが，(3b)のように「全額」「全部」のような表現と共起し，そのもの全体を削ることを表す場合もある。ただし，このときも，「余剰部分を全部削って，大きな単位として見た場合により小さくする」ということを表していると考えられる。

なお，多くの辞書類では専門用語として独立した語義が立てられている(4b)(5a)のようなスポーツ用語の例は，BCCWJ では少数だった。

4.2　構文的特徴

次に，4.1 節で示した語義ごとに，共起成分をはじめとする「カットする」の構文的特徴を見ていく。

4.2.1　共起成分

まず，「カットする」を述語とする文における，格成分と副詞的成分の出現を見る。

次の《表 3》は，5 例以上見られた共起成分を挙げたものである。

《表 3》「カットする」の共起成分

語義	用例数	格成分				副詞的成分	
		ヲ	道具デ	場所デ	カラ	結果	数量
[1] 切る	98	66	9		1	32	8
[2] 髪を切る	33	11	3	6		11	
[3] 減らす	103	52			5	4	16
[4] さえぎる	14	12	2				
[5] その他	4	2					
計	252	143	14	6	6	47	24

他動詞「カットする」のヲ格成分については，ハで主題化されているケースも含め，252 例中 143 例（56.7%）で同一文中に顕在している。対象がガ格成分となっている受身文が 54 例，連体修飾の主名詞になっている例が 10 例あるため，正確な意味でヲ格成分が文中に

現れない例は，45 例（17.9%）である。このうちの 20 例が語義[2]に見られるが，「私はいつも美容院でカットしている」のような例を，「髪を」が省略されていると見るべきか，自動詞構文とすべきかは，判断が難しい（ここでは便宜上，他動詞と考えた）。

この他の格成分で特徴的に現れるのは，語義[1]の道具デ格（例：ハサミでカットする），語義[2]の場所デ格（例：美容院でカットする），語義[3]のカラ格（例：賃金からカットする）である。

また，副詞的成分に目を向けると，「～を短く／一口大に／好きな形にカットする」のような結果を表す表現が語義[1]と語義[2]に，「～を少し／一部／全額カットする」のような数量や程度を表す表現が語義[3]に多く見られた。

これらの共起の状況をふまえ，10%以上の用例への出現を目安として，それぞれの語義の典型的な文型を示したものが，次の《表 4》である。

《表 4》「カットする」の語義と文型

語義		格成分	文型
[1] 切る	加工する／剥く／切り分ける	[食べ物／料理]を ([道具]で)	ヲ格 (＋道具デ格) (＋結果表現)
	短く／小さくする	[細い／薄いモノ]を ([道具]で)	
[2] 髪を切る	髪を切って整える	([(髪の)毛]を) ([場所]で)	(ヲ格) (＋場所デ格) (＋結果表現)
[3] 減らす	削除／短縮する	[映像／文章／項目]を	ヲ格 (＋数量／程度表現)
	数量を減らす	[お金／数]を	
[4] さえぎる	遮断する	[紫外線／光]を ([道具]で)	ヲ格 (＋道具デ格)
	取る	[ボール／パス]を ([身体部位]で)	

4.2.2 文末形式

次に，ヴォイスや補助動詞といった，「カットする」に後接する文末形式の特徴を見る。次の《表 5》は，5 例以上見られた文末形式を挙げたものである。

《表 5》「カットする」の文末形式

語義	用例数	受身ラレ	テイル	テシマウ	テモラウ	テイク	ヨウ
[1] 切る	98	9	4	1		2	
[2] 髪を切る	33	2	8		6	1	
[3] 減らす	103	39	16	6	1	4	5
[4] さえぎる	14	2		1			
[5] その他	4	2	2				
計	252	54	30	8	7	7	5

まず，顕著に目立つのは，語義[3]で，他の語義に比べ受身文の割合が高いという点であ

る（用例の 37.9%が受身文であり，受身文の 72.2%がこの語義）。すなわち，「映像や文章，お金の余剰な部分を減らす／削る」という出来事は，他に比べ，（意図に反して）行為を受ける立場から描かれることが多い，ということである。このことは，この語義に集中するテシマウ文（6 例）がすべて受身文の例であることからも裏付けられる。

一方で，助動詞ヨウの例に関しては，行政や経営者等の立場から出来事を描いており，ほとんどが国会会議録と経済書に見られるものである。

また，語義[2]では，恩恵を表すテモラウ文（やテクレル文）が特徴的に見られる。もともとこの語義の「カットする」には，（主語自身ではなく）主語の依頼によって他の人物（美容師等）が髪を切った場合も「私はいつもの美容院で髪をカットした」のような表現ができるという「介在性」（佐藤 2005）が特徴として見られる。テモラウ等による補助動詞構文は，この場合の依頼先を構文的に明示したものであると考えられる。

また，テイル文は，22 例が結果状態解釈（うち 16 例が受身文）と多数を占め，それ以外の解釈の例のほとんど（繰り返し・習慣解釈 6 例，継続・進行解釈 1 例）が，語義[2]に見られる（繰り返し・習慣解釈の残り 1 例は語義[1]）。語義[2]の繰り返し・習慣解釈は，「髪を切る」という語義との対応によって生じているものと考えることができる。

なお，テイク文に関しては，いずれの語義においても「徐々にその動作が進む」という漸進性を共通して表しており，特徴的なふるまいは見られない。

4.3　分析

以上の観察から，BCCWJ の「カットする」の用例からは，対象としてとる名詞の特徴だけでなく，語義ごとに共起成分や文末形式の特徴が指摘できることが分かった。

例えば，語義[2]「髪を切る」では，(a)（「髪を」が前提になっているため）ヲ格成分が現れない場合がある，(b)「美容院で」のような場所デ格をとることがある，(c)「短く」のような結果表現をとりやすい，(d) テイル形で繰り返し・習慣を表しやすい，(e) 介在性の文になるが，テモラウ等で動作主を明示することがある，といった特徴が見られる。

このように，動詞の意味・用法と構文的特徴との対応が一定の範囲で観察されるということは，形から意味が，意味から形が，ある程度予測できるということを意味する。例えば，日本語教育を目的とした場合，辞書的意味だけでなく《表 4》のような文型も同時に提示することで，ある文で使われている「カットする」がどの用法であるのかを共起成分から判断したり，特定の用法で「カットする」を使う際に，どの成分を共起させてよいかを予測したりする材料を提供できることになる。

５．おわりに

本研究の事例研究の結果は，外来語サ変動詞に関して，意味だけでなく「文の中でどのように使われるか」という文法的観点からの分析を行うことの重要性を示している。《表 1》のように国語辞典類の自他の情報にもズレが見られること，また，動詞用法辞典やコロケーション辞典では外来語の項目がきわめて限定されていることから明らかなように，まずはコーパスを用いて基礎的な記述を蓄積していくことが課題となる。

ただし，多義語であっても文型の違いが見られるとは限らないため，タイプ別の記述方法を考える必要があること，BCCWJ のような大規模コーパスでも細かな用法まで見る場合に用例が不足する可能性があること等の方法論的な問題点も検討していく必要がある。

参考文献

相澤正夫（1993）「『日本語教育のための基本語彙調査』と複合サ変動詞」『研究報告集』14，pp.281-332，国立国語研究所.

楳垣　実（1963）『日本外来語の研究』研究社出版.

石野博史（1996）「辞典における外来語の語義記述－「オープン」の場合－」『言語学林 1995-1996』，pp.273-286，三省堂.

荻野綱男（1996）「言語データとしての話者の内省・新聞 CD-ROM・国語辞典の性質－サ変動詞の認定をめぐって－」『計量国語学』20(6)，pp.233-252，計量国語学会.

加藤理恵・新川以智子（2002）「カタカナ語教育のためのカタカナ語考察 その 1－カタカナ語動詞のグループ分けと教育への提言－」『名古屋大学日本語・日本文化論集』10，pp.53-76，名古屋大学留学生センター.

金　愛蘭（2006a）「外来語「トラブル」の基本語化－20 世紀後半の新聞記事における－」『日本語の研究』2(2)，pp.18-33，日本語学会.

金　愛蘭（2006b）「新聞の基本外来語「ケース」の意味・用法－類義語「事例」「例」「場合」との比較－」『計量国語学』25(5)，pp.215-236，計量国語学会.

佐々木瑞枝(監修)（2001）『アカデミック・ジャパニーズ日本語表現ハンドブックシリーズ 5　よく使うカタカナ語』アルク.

佐藤琢三（2005）『自動詞文と他動詞文の意味論』笠間書院.

澤田田津子（1993）「日本語教育のための基本外来語について」『奈良教育大学紀要（人文・社会科学)』42(1)，pp.225-239，奈良教育大学.

中山恵利子・陣内正敬・桐生りか・三宅直子（2008）「日本語教育における「カタカナ教育」の扱われ方」『日本語教育』138，pp.83-91，日本語教育学会.

新川以智子・加藤理恵（2005）「カタカナ語形容詞のインターネット上での使用状況－カタカナ語教材作成に向けて－」『名古屋大学日本語・日本文化論集』12，pp. 141-158，名古屋大学留学生センター.

丸山直子（2010）「動詞の格情報－辞書記述とコーパス－」『特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集』，pp.65-72，特定領域研究「日本語コーパス」総括班.

宮田公治（2007）「外来語「メリット」とその類義語の意味比較－新聞を資料として－」『国立国語研究所報告 126 公共媒体の外来語－「外来語」言い換え提案を支える調査研究－』，pp.402-409，国立国語研究所.

森下訓子（2007）「洋語名詞の形容詞的用法について」『同志社女子大学大学院文学研究科紀要』7，pp.71-88，同志社女子大学大学院文学研究科.

辞典類

楳垣　実(編)（1972）『増補外来語辞典』東京堂出版.

北原保雄(編)（2002）『明鏡国語辞典』大修館書店.

小泉　保ほか(編)（1989）『日本語基本動詞用法辞典』大修館書店.

三省堂編修所(編)（2010）『コンサイスカタカナ語辞典（第 4 版)』三省堂.

西尾　実・岩淵悦太郎・水谷静夫(編)（2009）『岩波国語辞典（第 7 版)』岩波書店.

姫野昌子(監修)（2004）『研究社日本語表現活用辞典』研究社.

「−中」の用法−BCCWJ サブコーパス間の比較−

新實葉子（名古屋大学大学院生）†

Usage of the Suffix'-chû': Comparison between BCCWJ Subcorpus

Yoko Niimi (Nagoya University)

1. はじめに

本発表の目的は、事象の進行相を表す接尾辞「−中」の使用実態を「BCCWJ2009 モニター版コーパス」から得られる実例から明らかにし、特にサブコーパスごとの用法の違いを調査することである。

接尾辞「−中」は漢語・和語・外来語に接続し「〜（を）している」という意味を表す、生産性の高い接辞である。大島(2010)で『新奇用法』として言及されている、ブログの見出しなどで近年使用が広がりつつある「幸せ中」「専業主夫中」などの表現は、「−中」の生産性の高さと、「〜中」が使用される文体の特性から生じた表現である。

本研究では時間に関する接辞「−中」を時間名詞接辞とアスペクト接辞に分類した杉岡(2009)の分析に基づき、事象の進行相を表すアスペクト接辞「−中」について、『新奇用法』も対象に含め、「BCCWJ2009　モニター版コーパス」を用いて調査する。

2. 先行研究

2. 1 杉岡(2009)

杉岡(2009)は、時間をあらわす接辞「-中」を時間を表す「−中」を時間名詞接辞とアスペクト接辞の二種類に分け、両者の意味カテゴリーが異なることに起因する時間副詞との共起可能性や統語的特性の相違を説明する。

■ 意味的相違

時間名詞接辞：時間軸におけるある点（場所）を指す(−中、−前、−後)

アスペクト接辞：事象の進行相（進行中＝未完了）を表し、主語の一時的状況を表す述語を形成する。(未−、−中、−済み)

■ 共起可能な時間副詞

アスペクト接辞:「まま」「まだ」「昨夜から」「朝から」「すでに」等の副詞句と共起可能。

時間名詞接辞はこれらの語と共起できない。

■ 統語的相違

時間名詞接辞：VNP[1](1a)/NP(1b)[2]に接続し、副詞句を形成する

(1) a. [[ジョンが論文を執筆]$_{VNP}$ −中 $_{N}$] $_{NP}$ (に)

b. [[ジョンの論文の執筆]$_{NP}$ −中 $_{N}$] $_{NP}$ (に)

アスペクト接辞：VNP に接続し、述語となる(2a)/名詞修飾句となる(2b)

(2) a. ジョンが[[論文を執筆]$_{VNP}$ −中 $_{Asp}$] $_{Asp}$ だ。

b. [[論文を執筆] $_{VNP}$ −中 $_{Asp}$] $_{Asp}$ の学生

---

† niimiy@nagoya-u.jp

1 ここで用いる VNP とは、影山(1993)「動名詞」である。

2 時間名詞接辞が VNP ではない NP(出来事・時間)についた例(杉岡(2009):89)

出来事：レース中、ゲーム中、夕食中、裁判中

時間：楽器中、冬休み中、就業時間中、有効期限中

時間名詞接辞の「−中」は「−前」「−中」「−後」という時間軸上の前後関係を持ち、時間軸における一点を指し示すため「３時に」などの副詞句と交換することができる。

(3) {学生を指導中/３時}に電話が入った。 (杉岡(2009):88)

アスペクト接辞の「−中」は「未−」「−中」「−済み」という時間軸上の前後関係を持ち、事象の進行アスペクトに言及し、「進行中＝未完了」を表す。したがって、進行の意味での「ている」 と同じ意味を持ち、主語の一時的状態を表す述語を形成する。また、この述語は名詞修飾の表現を作ることもできる(2b)。

杉岡(2009)は、時間名詞接辞「-中」とアスペクト接辞「−中」が異なる意味カテゴリーに属するため、それぞれの「前接続要素+中」は異なる意味解釈と異なる選択制限を持ち、格助詞の交替や時間副詞との共起可能性が異なると述べる。

以下の例はＮＰにアスペクト接辞がついているように見えるケースである。

(4) ジョンは、今、論文の執筆中だ。

(5) ジョンは、今、会議中/休暇中/レース前だ。 (杉岡(2009):96)

(4)の「執筆」には名詞修飾句「論文の」がかかっているため「論文の執筆中」は NP である。(5)で「−中」に前接する要素は、項構造を持たないが事象を表す名詞で、(4)は(5)と平行した表現である。つまり、(4)は、「今、ジョンが「論文を執筆中」という出来事の時間軸のある地点にある」という意味を表しており、結果的に、一時的状態に言及するアスペクト接辞の用法と意味が似ると説明される。

ここで、杉岡(2009)で説明が難しいものについて考えたい。

(6) ただ今、マイクのテスト中(です)。

(6)の「−中」の前接要素「テスト」は VNP であるが、名詞修飾句「マイクの」がかかっているため「マイクのテスト」は NP であり、(4)と同様の例と考えられる。次に、「〜中」の様態を修飾する副詞句が含まれる場合を以下に挙げる。

(7) a. 心をこめて準備中

b. ?心をこめた準備中

(7a)の「心を込めて」は、「準備中」が表す事象の様態を修飾している。このような副詞句を(6)の文に加えてみる。

(8) a. ただ今、心をこめてマイクのテスト中。

b. ただ今、心をこめたマイクのテスト中

(8a)の「心をこめて」は「テスト中」が表す、「テストをしている」事象を修飾している。一方、(8b)の「心をこめた」が(8a)のように「テスト中」の事象を修飾している解釈は難しい。このことから、村岡(2009)で時間名詞接辞の用法として説明される(4)(6)のような例も時間軸上のある位置を指すのではなく、アスペクト接辞と同様に事象の進行を表していると考えられる。すなわち、(4)(6)の統語的特性は時間名詞接辞の用法だが、意味的にはアスペクト接辞であると考えられる。このような「−中」の用法を、本稿では『アスペクト接辞の非典型用法』とする。

2. 2 大島(2010)

大島(2010)は「〜ている」との比較を通して「−中」のアスペクト用法が持つ独自の意味を詳細に観察している。『新奇用法』と呼ばれる近年において使用が増加している用法にも言及しており、新奇用法が容認される条件を検証している。

大島(2010)で「−中」のアスペクト用法とされるは、「〜中。〜中だ。〜中である。」のような「〜中」が述語となる表現である。コピュラ文だけではなく、「育児中の主婦」のような連体修飾も「主婦が育児中である」という意味のものは含める。大島(2010)では「〜中、P」「〜中に、P」「〜中は、P」の用法を、「〜中」が P を時間軸に位置づける時間表現であるとしてアスペクト用法から区別し、時間表現に関しては取り扱わない。

■大島(2010)の接辞「−中」の区分[3]
時間表現：「〜中、P」「〜中に、P」「〜中は、P」
アスペクト用法：「〜中。」「〜中だ。」「〜中である。」「〜中の N」

■「〜ている」と「〜中」の比較

表 1 ：大島(2010)による「〜ている」と「〜中」の比較

意味	〜ている	〜中
動作・出来事の持続[−長期]	その事件を調べ**ている**。	その事件を調査**中**
[＋長期]	銀行に勤め**ている**。	不可
結果の持続	入院し**ている**。	入院**中**
習慣・反復	電車で会社に通っ**ている**。	不可
性状 [−可変性]	とがっ**ている**。すぐれ**ている**。	不可
[＋可変性]	流行っ**ている**。混ん**でいる**。	流行**中**。混雑**中**。
経験・回顧	一度事故を起こし**ている**。	不可
反実仮想	助けてもらわなかったら、今ごろは死ん**でいる**。	不可

表 1 は大島(2010)による「〜中」と「〜ている」との比較をまとめたものである。大島(2010)は「〜ている」との比較から「〜中」の意味特性を[＋長期][＋可変性]にまとめている。さらに「故障中」の容認可能性のゆれ[4]と、「婚約中」「?結婚中」の容認度の違いを、「〜中」のもつ「一時的状態」の意味に起因するとして[＋一時性]の意味特性を加える。最後に、持続を表す表現という意味での意味特性[＋継続性]を加え、「〜中」の意味特性を[＋T]として以下のようにまとめている。

(9)　[＋T] : [＋一時性] [−長期] [＋可変性] [＋継続性]

この[＋T]の意味解釈可能性が「〜中」表現の許容性を左右しており、新たな用法である新奇用法が生み出される際の基準もこの[＋T]である。大島(2010)では「?結婚中」の容認度が低い一方で、ウェブ検索で得られる「遠距離結婚中」「テスト結婚中」は奇妙に感じないことについて、これらの表現は「遠距離」「テスト」といった修飾語がつくことで、全体が[＋T]として解釈されるため容認されていると説明する。

■ 新奇用法 1　和語単純動詞＋中：「考え中」「休み中」「悩み中」

和語と「—中」が結び付く場合、「取調べ中」「取り扱い中」「売り出し中」など複合動詞＋中は安定した容認度が認められるが、単純動詞＋中は「話し中」「休み中」などの少数の例しか見られず、また、「田中さんの休み中に、P」(時間用法)では認められても「田中さんは休み中だ」(アスペクト用法)は言えないとされる。

北原(2005)では、若者言葉に和語単純動詞＋中が見られる「ストーリーを考え中」「旦那の態度にむかつき中」「買うか買わぬか迷い中」ことを、メールの影響であるとする。大島(2010)は KOTONOHA の検索デモンストレーションを利用し、「悩み中」を検索したところ出典が全て「Yahoo!知恵袋」であることから、ジャンルの偏りが見られる『新奇用法』だとする[5]。

■ 新奇用法 2 ブログの見出しの「〜中」: インパクト性と「〜中」の[＋T]特性の相性

大島(2010)では、インターネットのブログの見出しに多用される「ただ今病気中」「只今専業主夫中」「着席中」「パソコン死亡中」など、これまでの接尾辞「—中」の用法感覚からは違和感を覚える実例を取り上げ、見出しの持つ『ブログで次々に更新されて行く主体の一時的状態の記述を要約して示す役割』((大島(2010):136)、下線は新實)に「〜中」の[＋

[3] これらは杉岡(2009)のいう「−中」の時間名詞接辞用法とアスペクト接辞用法に対応すると考えられる。
[4] 「故障中」は違和感を与える表現としてこれまでにも注目されてきた(柏野(1993)など)
[5] 「考え中」という表現は最近のものではないと考えられ、青空文庫に 8 例あることが述べられている。

T]特性が適合するためと説明する。

大島(2010)では格関係などの統語的特徴に言及していないが、新奇用法の創出の際に基準となる意味特性[＋T]は、村岡(2009)のアスペクト接辞(2)も同様に持つと考えられる。また、本稿でいう『アスペクト接辞の非典型用法』もこの意味特性[＋T]の解釈が認められる。

3. 本発表での検索対象と想定される結果

3.1 検索対象とする「前接要素＋中」

以上の先行研究から、接辞「−中」の典型的な用法の統語的特性は杉岡(2009)で述べられたアスペクト接辞(2)のものを基本とし、意味的特性は杉岡(2009)と大島(2010)のものに従う。

すなわち、(2a)(6)のように、述部位置に現れる「〜中」と、(2b)のように「〜中の N」を「N が〜中である」と言い替えられるものが検索対象である。

(2)′ a. ジョンが[[論文を執筆]$_{VNP}$ −中 $_{Asp}$] $_{Asp}$ だ。

b. [[論文を執筆] $_{VNP}$ −中 $_{Asp}$] $_{Asp}$ の学生

(6)′ ただ今、マイクのテスト中(です)。

3.2 想定される結果

■ 各サブコーパスの特性傾向

形式的：OM(国会会議録)、OW(白書)
非形式的：BK(書籍)、OC (Yahoo!知恵袋)

「BCCWJ2009 モニター版コーパス」のサブコーパスの内訳から、OM と OW がよりフォーマルな文体であり、BK と OC がよりフォーマルさが低い文体であると考えられる。

■ 想定される使用の偏り

KOTONOHA の検索デモンストレーションを使用した大島(2010)でも、新奇用法の実例が Yahoo!知恵袋に偏って見られることが報告されているように、典型的な用法から逸脱した『アスペクト接辞の非典型用法』や『新奇用法』は、フォーマルさの低い文体に多く用いられると考えられ、話し言葉寄り・ネット上の書き言葉のサブコーパスである OC(Yahoo!知恵袋)と BK(書籍)に多く用いられる傾向があると想定される。

4. コーパス検索

本研究では「BCCWJ2009 モニター版コーパス」を使用し、接尾辞「−中」のアスペクト用法を抽出する。該当例の抽出には、検索過程の透明性と検索結果の再現性を保つため、UNIX システム上でのテキスト処理、および、形態素解析プログラムを利用した。検索対象としたサブコーパスは、Yahoo!知恵袋(以下 OC)・書籍(以下 BK)・白書(以下 OW)・国会会議録(以下 OM)の 4 種類である。

1) 文字コードの変換
2) 「中」を含む行の抽出 (Grep)
3) 「中」を含む行の形態素解析
4) 「中_名詞_接尾_副詞可能」を含む行の抽出 (Perl)
5) Excel 上での処理

1) BCCWJ2009 モニター板コーパスのテキストファイルは文字コードが UTF-16 なので、形態素解析プログラムで処理するために UTF-8 への変換を行った。

```
nkf --ic=UTF-16LE-BOM --oc=UTF-8 all_utf16 > all_utf8
```

2)「中」の文字を含む全ての行を grep コマンドで抽出する。(非該当例が大量に含まれる)

```
grep '中' all_utf8 > chu_grep
```

3) 手順 2 で得たファイルを対象に、形態素解析プログラム「MeCab」による品詞情報の付与を行う。

```
mecab -b 5000000 -F "%m_%F_[0,1,2] " chu_grep > chu_grep_mecab
```

MeCab は出力時の品詞情報のフォーマットを指定する事ができる。今回は"_"(アンダーバー)でタグをつなげ、単語の間を半角スペースでわかち書きした形式で出力するよう指定した。上記のスクリプトを実行して得られる結果の一部を以下に示す。

> 株_名詞_一般 の_助詞_連体化 売買_名詞_サ変接続 を_助詞_格助詞_一般 やっ_動詞_自立 て_助詞_接続助詞 いる_動詞_非自立 方_名詞_非自立_一般 は_助詞_係助詞 、_記号_読点 仕事_名詞_サ変接続 中_名詞_接尾_副詞可能 に_助詞_格助詞_一般 値動き_名詞_一般 を_助詞_格助詞_一般 どう_副詞_助詞類接続 やっ_動詞_自立 て_助詞_接続助詞

4) 事前に「食事中」「夏休み中」「今月中」「考え中」「ご飯中」など、検索対象とその関連表現の形態素解析を行い、上記全ての「−中」には「中_名詞_接尾_副詞可能」のタグが付与されることを確認し、「中_名詞_接尾_副詞可能」を含む行を Perl のスクリプトで抽出した。 手順 2 で用いた Grep ではなく Perl スクリプトを使用することで、「中_名詞_接尾_副詞可能」タグが 1 行に 2 件以上含まれる場合に該当業ごとタグの出現回数に応じて出力する。これは、手順 5 で KWIC 形式に整える際に、該当例 1 件ごとに 1 行表示する必要があるためである。

```
perl -ne 'while(/中_名詞_接尾_副詞可能/g) { print; }' chu_grep_mecab > chu_perl
```

5) テキストエディタを利用して Excel で読みやすい形式に整形作業を行った。「前 1 語とタグ」と「中_名詞_接尾_副詞可能」の前後にタブを区切り文字として入れ、タブを全て消去する。この結果 Excel 上で KWIC 表示することができる。図 1 はその一例である。

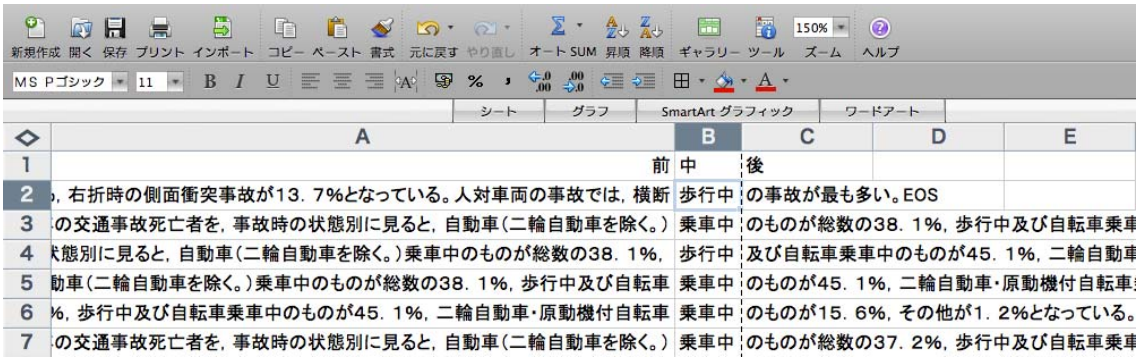


図 1　Excel 上での KWIC 表示

6) Excel 上で KWIC 形式の整理と非該当例の消去を手作業で行った。この過程で文脈を確認しながら「朝食中」「今月中」などの非該当例を排除する。

5. 検索結果

表２は「BCCWJ2009 モニター版」各サブコーパスにおける時間を表す「−中」の出現数と種類である。

表２　各サブコーパスごとの「−中」実例

	出現数	種類
OC(Yahoo!知恵袋)	1686	434
OM(国会会議録)	669	150
BK(書籍)	3640	810
OW(白書)	1065	207

以下の節で各サブコーパスにおける接辞「−中」使用の特徴を述べる。

5.1 OC(Yahoo!知恵袋)

大島(2010)でも KOTONOHA の検索で新奇用法の実例が報告されている通り、典型的な事象を表す接辞「−中」とは異なる前接要素が多く見られる。特に(10)のように、通常は「する」と結び付くことが可能な動名詞とは見なされない名詞が「−中」に前接する例が見られる点で他のサブコーパスと異なる特徴が見られる。

(10) 「知恵袋中[6]」「プール中[7]」「ネット中[8]」「迷子中[9]」「ご飯中」[10]

和語単純動詞の連用形が「−中」に前接する例はサブコーパス中で最も多く見られる。(括弧内の数字は出現数)

(11) 「考え中:13」「話し中:4」「話中:4」「悩み中:4」「探し中:3」「探し中:3」
「作り中:1」「預かり中:1」「凹み中:1」

アスペクト接辞の非典型用法は 8 例見られる。(12)に一例を挙げる。

(12) ３個中１つのプログラムのインストール中です

5.2 BK(書籍)

BK(書籍)サブコーパスはさらに、ベストセラー・生産実態(出版)・流通実態(図書館)の３つのサブコーパスに分類されるが、本研究ではそれらを一つの「書籍」コーパスとして扱う。書籍には小説の会話文も含まれており、多用な「−中」使用が見られた。

BK(書籍)サブコーパスには和語の複合動詞連用形が「−中」に前接する例がサブコーパス中で最も多く見られる。

(13) 「取調べ中:4」「取り壊し中:4」「取り込み中:4」「売り出し中:3」
「貸し出し中:2」「見習い中:2」「張込み中:2」「張り込み中:1」
「建て替え中:1」「埋め立て中:1」「切り替え中:1」「申し立て中:1」
「呼び出し中:1」「組立て中:1」「積替え中:1」「取扱い中:1」

また、外来語が「−中」の前接要素となる例も多く見られる。(14)に一例を示す。

(14) 「ダイエット中:8」「ゲーム中:8」「デート中:7」「プレー中:6」
「パトロール中:4」「ドライブ中:4」「ストライキ中:3」「ジョギング中:3」

---

6 「知恵袋中に急にログインページに切り替わりました。」より。「Yahoo!知恵袋を閲覧している最中に」の意で解釈。(時間名詞接辞)

7 「ある学校でプール中の小学児童４０人の...」より。(アスペクト接辞)

8 「最近ネット中にカタカタと音がしてフリーズしてしまいます。」より。(時間名詞接辞)

9 「自分探しの旅は終わりましたか？それともまだ迷子中？」より。(アスペクト接辞)

10 「今でも、ご飯中に箸をくわえたまま寝てしまったり」より。(時間名詞接辞)

OC(Yahoo!知恵袋)と同様に、和語単純動詞の連用形が「−中」に前接する例も見られる[11]。

(15)「話し中:7」「考え中:2」「調べ中:1」「温め中:1」

アスペクト接辞の非典型用法は 8 例見られる。(16)に一例を挙げる。

(16) ただ今物品の整理中です

5.3 OM(国会会議録)

OM(国会会議録)には、アスペクト接辞の非典型用法が見られない。また、外来語が「−中」の前接要素となる例も含まれず、OM(国会会議録)における「−中」の使用は「漢語+中」が圧倒的多数である。わずかながら、和語の複合動詞連用形が前接要素となった例が見られる。

(17)「取りまとめ中:4」「取り調べ中:3」「埋め立て中:1」

OM(国会会議録)にのみ見られる表現に「仕掛かり中:9」がある。OM(国会会議録)を対象に「仕掛か」を検索したところ、「仕掛かり路線」「仕掛かり品」「仕掛かり中」の形でしか用いられておらず、「−中」の用例のうち特殊な表現として注意が必要である。

5.4 OW(白書)

OW(白書)には、和語の単純同士連用形が前接要素となる例は見られず、和語の複合動詞連用形が前接要素となる例が「とりまとめ中:2」の 1 例見られた。外来語が前説要素となる例も「パトロール中:3」の 1 例である。アスペクト接辞の非典型用法は 2 例見られた。(18)に一例を挙げる。

(18) 現在，原案のとりまとめ中である。

6. 考察

「−中」の前接要素の種類と、『アスペクト接辞の非典型用法』『新奇用法』の使用頻度は OC(Yahoo!知恵袋)・BK(書籍)に多く見られ、OW(白書)・OM(国会会議録)は比較的少ないことが確認された。このことから、典型的な「−中」の用法はフォーマルな文体に多く用いられ、非典型的な「−中」の使用はくだけた文体に多く用いられる傾向があると言える。また、「−中」の生産性は使用される文体に大きく左右されると言える。

OC(Yahoo!知恵袋)サブコーパスから得られた「−中」の実例の中でも、(10)に挙げた「普通名詞+中」の用例には注意が必要である。統語的にアスペクト接辞と考えられ、また、意味も「事象の進行相を表し、主語の一時的状況を表す」アスペクト接辞のものである。大島(2010)では、全体として[＋T]の解釈が可能である場合に、事象の進行を表す『新奇用法』が容認されるとしている。本来、「前接要素+中」が事象の進行相を表し得たのは前接要素が動詞性(項構造を持つ・ガ格を取る・事象を叙述する)を備えている場合のみだったのが、「〜ている」との類似性と、大島(2010)で指摘された「表現の短さ」という利便性から、「前接要素+中」が事象の進行相を表すために必要な条件が緩くなりつつあると考えられる。

本稿では『アスペクト接辞の非典型用法』と称した、(6)に例示される「−中」の用法が容認される条件については深く考察していない。該当例が豊富に得られたとは言えず、十分な考察を進めるだけの具体例がない段階で議論を進めることは難しい。現段階で考えられるのは、『新奇用法』の許可が、「−中」を含む文全体の意味特性[＋T]での解釈に依拠しているように、『アスペクト接辞の非典型用法』も該当表現を含む文全体が[＋T]の意味解釈ができる場合に容認される可能性を今後検討することである。

---

[11] 「お泊まり中」「お休み中」など、名詞につく接辞「お」がついた例が多くみられる。今回はこのような例を除外したが、こういった例を「お+名詞+中」と扱うか「お+動詞の連用形+中」と扱うかに関して今後の検討が必要である。

7. おわりに

本研究では、事象の進行相を表す、生産性の高い接辞「−中」の使用実態を明らかにするため、BCCWJ2009 モニター版コーパスにおける実例を用いて調査した。

今後の課題に、本稿で『アスペクト接辞の非典型用法』とした「−中」の用法((6)のようなもの)が容認される条件の調査と、OC(Yahoo!知恵袋)サブコーパスに見られた「迷子中」のような新奇用法の分析が残る。携帯電話のメールや Twitter などの文字数が制限された場で個人が表現する機会が格段に増えた今、『新奇用法』の使用は増えつつあるように思われる。そのような『新奇用法』の成立にはどのような統語的制約・意味的制約が課せられるのか分析することで、現代における接辞「−中」の用法をより明らかにすることが望まれる。

参考文献

大島弘子 (2010). 「漢語接尾辞「〜中」によるアスペクト用法−「〜ている」との違い−」『漢語の言語学』pp.121-139. くろしお出版.

影山太郎(1993)『文法と語形成』ひつじ書房.

柏野健次(1993). 「「故障中」の意味論」『意味論から見た語法』pp.226-243. 研究社.

杉岡洋子 (2009). 「「−中」の多義性−時間をあらわす接辞をめぐる考察−」『語彙の意味と文法』pp.75-94. くろしお出版.

BCCWJ と誤用コーパスを利用した日本語作文支援に関する一考察

八木豊（株式会社ピコラボ）†
鈴木泰山（株式会社ピコラボ）
仁科喜久子（東京工業大学）

Japanese Writing Support System Using BCCWJ and Learners' Error Corpus

Yutaka YAGI (Picolab Co., Ltd.)
Taizan SUZUKI (Picolab Co., Ltd.)
Kikuko NISHINA (Tokyo Institute of Technology)

１．はじめに

仁科らは，特定領域研究「日本語コーパス」内において，作文支援システム班として日本語作文支援システム「なつめ」の開発および評価を行ってきた（仁科 2011）．「なつめ」では，BCCWJ を含む日本語コーパスから大量の共起情報を収集し，それらを効果的に日本語学習者に閲覧させることで作文支援における一定の成果をあげている．しかしながら，現在，「なつめ」のシステム内で使用しているコーパスは基本的に正しい使われ方をしている日本語として収集したもので，誤用に関する情報は含んでいない．日本語学習者の誤用を分析，タグ付けした学習者誤用コーパスを合わせて利用することができれば，異なる観点からの日本語作文支援が可能になることが予想される．そこで，我々は，学習者誤用コーパスの作成に向けて，複数の日本語教師の協力のもとで日本語学習者が犯し易い誤用の分類・定義を行い，日本語学習者が書いた作文およそ 5,000 文を収集した．現在までに，そのうちのおよそ 3,500 文に対して誤用のタグ付けを実施済みで，タグ付けした誤用に対する分析を進めている（曹 2010）．今後は，汎用アノテーションツール Slate を利用して引き続きタグ付けを実施していく予定である．

本稿では，作成中の学習者誤用コーパスに対して行った調査・分析の結果を報告するとともに，BCCWJ と学習者誤用コーパスを合わせて利用した作文支援の可能性について，語の組み合わせの誤りおよび発音・表記の誤りに対する訂正例の提示に関する内容を中心に考察を行う．

２．学習者誤用コーパスの概要

本章では，調査・分析に使用した学習者誤用コーパスの概要について説明する．

今回我々が使用した学習者誤用コーパスは，大学あるいは大学院留学生の日本語による作文を収集し，それらに対して表 1 に示す誤用種別定義にしたがって複数の日本語教師が誤用のタグ付けを実施したものである．誤用種別定義は，「ディコース」，「構文・文法」，「語彙用法」，「発音・表記」の 4 つを第 1 層とする 3 層の階層構造を成しており，下のほうの階層ほど，誤用のより詳細な内容を指し示すようになっている．これまでに，164 人の留学生から 261 テキスト，5,600 文を収集しており，作文の主な内容は日本語教育の授業の中で設定した特定のテーマについてのレポートである．今回は，その中から，誤用のタグ付けを実施済みであるおよそ 3,500 文を使用して調査・分析を行った．タグ付けを実施済みの誤用の数は全体で 5,391 件，誤用種別ごとの件数は表 1 に示す通りである．

† yagi@picolab.jp

表１　誤用種別定義

第1層	第2層	第3層	件数
ディスコース			
	論理的整合性		
		序論構成	0
		本論構成	0
		考察の構成	0
		結論構成	1
	段落接続		10
	文接続		29
	指示語		38
	スタイル(文章としての統一性・場)		448
	待遇表現		4
	文のまとまり		
		句読点 前件・後件の整合性	203
		箇条書き語の並列	3
語彙用法			
	口語・文語のスタイル		404
	語の組み合わせの誤り(違う意味になる)		124
	不適切な語の選択(類義語)		462
	母語干渉		51
	語の余剰・重複・冗長・欠落		335
	語形誤り		35
	位相		1
発音・表記			
	表記発音		
		かな	61
		カタカナ	41
		漢字	58
		促音	18
		長音	14
		濁音清音	76
		拗音撥音	0
		その他	11

第1層	第2層	第3層	件数
構文・文法			
	統語的呼応(主述のねじれ、副詞呼応)		73
	語順		26
	係り受け		
		連体修飾	18
		連用修飾	24
	品詞		
		動詞	450
		形容詞・形容動詞	133
		副詞	98
		疑問副詞・疑問詞	1
		接続詞	64
		助詞	925
		機能語・助詞相当	97
		助動詞	120
		数詞	2
		名詞(句)	434
	活用		
		自他動詞	33
		終止	23
		連体	7
		連用	64
		已然	2
		未然	7
	文法機能(テンス・アスペクト・モダリ		
		授受	8
		可能	37
		自発	5
		受け身	70
		使役	17
		テンス	90
		アスペクト	94
		モダリティ	42

学習者誤用コーパスは，タグ付けした誤用に関する情報の他に，作文を行った日本語学習者の情報として，国籍，母語，性別，日本語レベル，学習時間などを収録している．現在のところ，中国語母語話者による作文の数が圧倒的に多く，その他の母語話者による作文の数が少ないため（表２），母語による違いを定量的に分析することは難しいが，将来的には，そういった分析を進めることで母語干渉による誤用を捉え，日本語学習者の母語に応じた作文支援をすることを検討している．

また，これまでに実施してきたタグ付けの結果，表１に示す現在の誤用種別定義に冗長な部分や不備があることも認識しており，その影響でタグ付けする内容に揺れが生じてい

表２　学習者の母語

母語	人数
中国語	119
ベトナム語	15
韓国語	11
マレー語	3
スペイン語	2
その他	14

る箇所も多々存在する．現在，誤用種別定義の修正やタグ付けデータの修正を並行して進めているが（仁科 2011），本稿の調査・分析ではそういったものも含んだ状態のデータを使用している．

３．学習者誤用コーパスの分析

本章では，作成中の学習者誤用コーパスのうち以下に示す誤用種別に対して行った分析の結果とともに，BCCWJ と学習者誤用コーパスを合わせて利用した作文支援の可能性について言及する．

- 語の組み合わせの誤り
- 発音・表記

３．１　語の組み合わせの誤り

学習者誤用コーパスにおいて「語の組み合わせの誤り」としている誤用は，いわゆるコロケーションの問題である．例えば以下に示す学習者による作文では，「関係が多い」という語の組み合わせが不自然であるとして「関係が深い」に訂正されている．

例文１．だから、実は文字が単独に存在するものではなくて、社会との関係が多い→関係が深いと思います。

本稿では，こういった「語の組み合わせの誤り」のうち，日本語作文支援システム「なつめ」で収集している共起データ（阿辺川 2011）を利用することが可能な「名詞＋助詞＋動詞」の組み合わせの誤りに着目した．現在の学習者誤用コーパスには「名詞＋助詞＋動詞」の組み合わせの誤りは 117 件登録されており，表３は，誤用の列に示す「名詞＋助詞＋動詞」の組み合わせに対して式１にしたがって算出した値でそれらを昇順に並べたものから 10 件を示したものである．訂正例の列は学習者誤用コーパス上でタグ付けされた情報の一つで，それぞれの誤用に対する訂正例を示している．

$$(O(n,p,v) / O(n,p)) \times (O(n,p,v) / O(p,v)) \qquad \text{（式１）}$$

式中の n,p,v はいずれも誤用側の名詞，助詞，動詞で，O(n,p,v)は誤用側の「名詞＋助詞＋動詞」の組み合わせに対する「なつめ」共起データ内の頻度を表している．同様に，O(n,p)は「名詞＋助詞」の組み合わせに対する共起頻度，O(p,v)は「助詞＋動詞」の組み合わせに対する共起頻度を表している．「なつめ」共起データは BCCWJ を含む日本語コーパスから大量の正用を収集したものであり，前述の式１で算出した値が小さい「名詞＋助詞＋動詞」の組み合わせは，一般的な日本語の文章では使用されない表現であると考えることができる．したがって，日本語学習者の作文中で表３の誤用の列に示すような「名詞＋助詞＋動詞」の組み合わせが使用されている場合には，完全にではないがある程度の確信を持って誤用であると判定し，該当する誤用に対してタグ付けされた訂正例を訂正候補の一つとして提示することが可能となる．

反対に，前述の式１で算出した値が大きい「名詞＋助詞＋動詞」の組み合わせは，一般的な日本語の文章でも普通に使用される表現であると考えられるため，それだけでは誤用であると判断できない．文脈や作文の内容など，その他の情報を含めて別途判定する必要がある．

例文２．十年後のわたしはどうなりますか。二十年後のわたしはどうなりますか。数十年後は？わたしいつも思っています→考えています。

例えば，例文２の「私が思う」という表現では，タグ付けを行った日本語教師は，現在の学習者誤用コーパスために収集した作文が，学習者が日本語教育の授業でレポートを書いたものであるという背景を考慮したうえで「私が思う」という表現よりも「私が考える」

表３　「名詞＋助詞＋動詞」の誤用および訂正例

（式１）の値×10^6	誤用	訂正例
0.00011907	頭,に,出る	頭,に,浮かぶ
0.00015241	人,を,もたらす	人,を,生み出す
0.00029828	私,が,発達する	私,が,成長する
0.00030957	問題,に,会う	問題,に,出会う
0.00042389	手紙,を,伝える	手紙,を,届ける
0.00044457	数,を,高める	数,を,増やす
0.00055275	嘘,を,作る	嘘,を,つく／言う
0.00066743	文化,を,述べる	文化,を,表す
0.00102388	漢字,を,発見する	漢字,を,発明する／考案する
0.00107140	努力,が,含まれる	努力,が,つぎ込まれる／注ぎ込まれる

という表現が好ましいと判断している．これを「なつめ」共起データを使用して検証してみると，共起データ全体では「私が思う」は 183 件出現しており，式１にしたがって算出した値も 839.97 と表３に挙げた誤用事例よりもはるかに大きく，一般的な日本語の文章でも使用する表現であることがわかる．しかし，対象を科学技術論文（BCCWJ 以外から収集したデータ）に限定すると 1 件も出現していない．したがって，共起データからも，科学技術論文やレポートでは「私が思う」よりも「私が考える」という表現が好ましいということがいえる．

我々は，以下に示す段階を踏むことで，これと同様の誤用事例，すなわち，一般的な日本語の文章でも使用する表現ではあるが，場面や話題，立場などレジスタを考慮することで誤用になりうる事例を抽出することを検討している．

（１）　語の組み合わせの誤りの中から，「思う」と「考える」のように，誤用側の単語（思う）と訂正例側に挙げられた単語（考える）が同一概念あるいは近い概念に属している事例を抽出する（日本語 WordNet（Bond ら 2009）などの概念体系辞書を利用）

（２）　（１）で抽出した事例の誤用側の「名詞＋助詞＋動詞」の組み合わせに対して「なつめ」共起データを使用して，共起データ全体および科学技術論文などのジャンル毎に式１の値を算出する

（３）　（２）で算出したジャンル毎の値で極端に小さいものがあれば，そのジャンルの作文においては当該事例の誤用側の表現を使用することは誤用であると判断する

学習者誤用コーパスからこのような事例を自動的に抽出すれば，例えば，学習者がレポートで「私が思う」という表現を使用している場合に，それはレポートでは不適切な表現で「私が考える」のほうが好ましいということを示すことができる．

３．２　発音・表記

学習者誤用コーパスにおいて「発音・表記」としている誤用は，字形の類似や音の類似によって生じる文字レベルの誤り全般である．例えば以下に示す学習者による作文では，「コンミュニケション」を「コミュニケーション」に訂正している．

表４　編集距離の分布

編集距離	割合
1	81.36%
2	16.10%
3	1.69%
4 以上	0.85%

表５　頻度の多い編集操作

編集操作	
「ー」を削除	「ッ」を「ン」に置換
「ズ」を「ス」に置換	「で」を「て」に置換
「ー」を「ッ」に置換	「こ」を「ご」に置換
「だ」を「た」に置換	「っ」を挿入
「て」を「で」に置換	「ッ」を「ー」に置換
「っ」を削除	「す」を「ず」に置換
「と」を「ど」に置換	

例文３．先ず、メンバー同志でコンミュニケション→コミュニケーションをとることが難しかった。

我々はまず，このような「発音・表記」の誤り対して，誤用側と訂正側の文字列間の編集距離および編集操作を集計した．文字列間の編集距離とは，一方から他方を得るために必要な編集操作（削除，挿入，置換）の回数である．先ほどの例文の場合，誤用側の一つ目の「ン」を削除し，「ケ」と「シ」の間に「ー」を挿入すると訂正側の文字列と一致するので，編集距離は 2 となり，編集操作は「ン」の削除と「ー」の挿入となる．集計した結果，編集距離の分布は表４に示す通りとなった．

この結果から，日本語学習者が（一つの単語内で）表記誤りを犯す場合，およそ 97.5%は編集距離が 2 以内，編集距離 3 まで広げると 99%をカバーできることがわかる．つまり，日本語学習者の作文に表記誤りがあった場合に提示する修正候補は，表記誤りを含む文字列から編集距離が 2 ないしは 3 の範囲で探せばおおむね問題ないということになる．

次に，頻度の多かった編集操作を表５に示す．これは，日本語学習者がどのような表記誤りを犯し易いかということを示している．

以上の結果をふまえて，表記誤り e に対して正しい表記 c を見つけたい場合，e から編集距離が 3 以内の文字列の集合を修正候補とし，以下に示す確率モデルに基づいて修正候補の中から確率最大となる c を見つけることで可能ではないかと考えている．

$$\arg\max_c P(c|e) \qquad \text{（式２）}$$
$$= \arg\max_c P(e|c)\,P(c)\,/\,P(e)$$
$$= \arg\max_c P(e|c)\,P(c)$$
$$\fallingdotseq \arg\max_c \left(\prod_{a\in A} P(a)\right) P(c)$$

ただし，A は c から e への編集操作の集合で，誤りモデル $P(e|c)$では日本語学習者が犯し易い編集操作に基づいた間違い易さを，言語モデル $P(c)$では正しい表記 c の一般的な文書中での使われやすさを，それぞれ誤用コーパス，BCCWJ から算出すること想定している．

４．まとめ

本稿では，我々が作成中の学習者誤用コーパスについて概要を説明し，BCCWJ と学習者

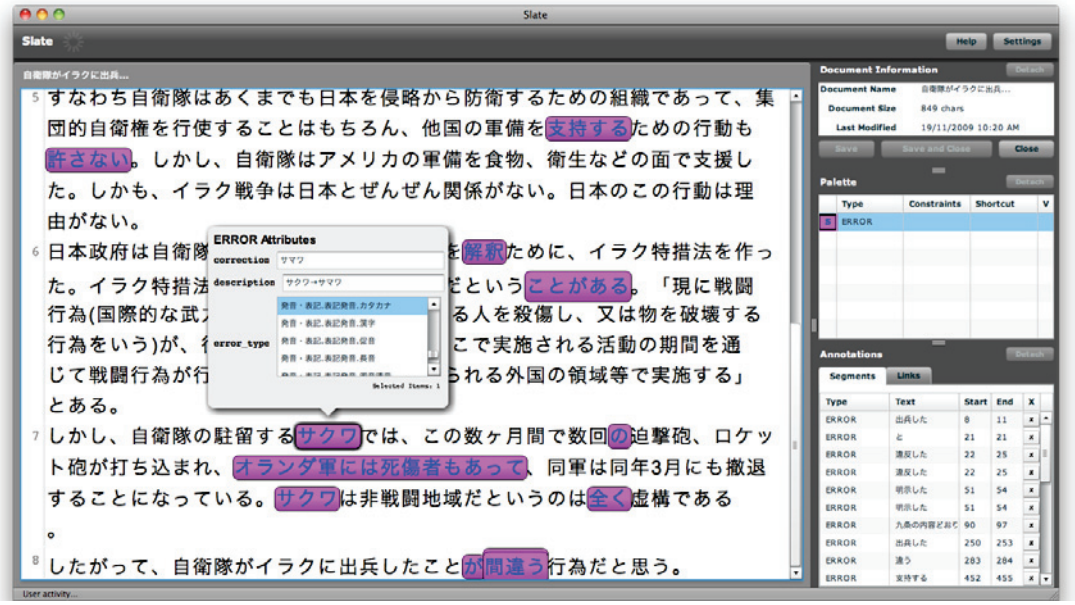


図 1　Slate による誤用タグ付け

誤用コーパスを合わせて利用した作文支援の可能性について言及した．今後は，実際の作文支援システムとして実装し，「なつめ」と連携することが目標であるが，現段階では，機械的な処理が困難である部分を人手で作業している箇所も多く，まずはその点を解決していく必要がある．

また，現状の学習者誤用コーパスには不備も多く，定量的な評価・分析を安定して行うためには質量ともに改善する必要がある．その点について，我々は，誤用種別定義の修正やタグ付けデータの修正を並行して進めている．また，今後は，特定領域研究「日本語コーパス」ツール班が開発した汎用アノテーションツール Slate（徳永 2010）を使用してタグ付けを実施することで，タグ付けの効率改善を図ることができると考えている（図 1）．

文献

仁科喜久子, 村岡貴子, 因京子, Joyce Terence Andrew, 鎌田美千子, 阿辺川武. 特研究活動・成果の総括：作文支援システム班　バランス・コーパス利用による日本語作文支援システム「なつめ」の構築と評価. 2011

阿辺川武, ホドシチェク・ボル, 仁科喜久子. 日本語コーパスを利用した作文支援システム「なつめ」の評価. 2011

曹紅荃, 黒田史彦, 八木豊, 鈴木泰山, 仁科 喜久子. 学習者作文支援システムのための誤用データベース作成－動詞の誤用分析を中心に－. 世界日語教育大会論文集, pp.1571-1～1571-9. 2010

Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. Extending the Japanese WordNet. 言語処理学会第 15 回年次大会. 2009

徳永健伸, Dain Kaplan, 飯田龍. Slate - A multi-purpose annotation tool. 情報処理学会自然言語処理研究会. Vol.NL-199. No.19. 2010. Nov

関連 URL

日本語作文支援システム「なつめ」: http://hinoki.ryu.titech.ac.jp/

BCCWJ モニター公開データに基づいた並立助詞「や」の分析

川口裕子（神戸女学院大学大学院）

The Analysis of Japanese Particle '*ya*' Based on BCCWJ Data

Yuko Kawaguchi (Kobe College, Graduate School)

1. 序

日本語並立助詞「や」は、一般的に名詞があるセットの一部として取り上げられていることを示し、すべての要素が列挙されていることを示唆する「と」とは明確に区別される。しかし、実際には、助詞の「と」や「か」に相当する意味でも用いられている。

以下は、通訳ガイド試験対策講座の模擬試験問題であるが、「や」が文脈により、「と」の意味に近い'and'や「か」の意味に近い'or'と英訳されている。「や」が、文脈によってさまざまな意味で用いられていることを示す一例である。

……特に日本の歴史について語る場合、アメリカやイギリスの旅行者は、彼ら自身の歴史との対比を聞くと喜ぶ。

In particular, when the guide talks about Japanese history, tourists from the United States and United Kingdom enjoy hearing comparisons made with their own history.

……この意味においては、浴衣は湯から上がった時や、パジャマ代わりに着るものだ、という方がより的を射ているのかもしれない。

In this sense, it might be more to the point to say that *yukata* is something people wear when getting out of the bath or instead of pajamas.

本研究では、「現代日本語書き言葉均衡コーパス」（BCCWJ）モニター公開データ（2009 年度版）上の用例を用い、並立助詞「や」の使用の実態調査を行う。特に、先行研究によれば「と」または「か」が好まれるべき場合に「や」が用いられている事例について分析し、「や」が選択される理由について考察する。

2. 並立助詞「や」

2.1 辞書に見る「や」・「と」・「か」の定義

『てにをは辞典 - A Dictionary of Japanese Particles』は、並立助詞「や」・「と」・「か」をそれぞれ以下のように定義している。

('ya':) The basic function of the particle 'ya' is similar to that of the particle 'to'. However,

whereas ‘to’ refers to specific, clearly defined things, ‘ya’ refers more to a category, and makes the things more vague.……Placed after a noun, loosely refers to it and other things that are also present or similar.

(‘to’:) The most basic function of ‘to’ is to list all nouns, noun phrases and noun clauses.

(‘ka’:) …… lists a number of choices.

その他の辞書でも、概ね「と」と「や」を似た概念として扱っているが、両者の違いは「新日本語の基礎Ⅰ」 (1990) の以下の見方とほぼ一致する。

The difference is that ***to*** enumerates, while ***ya*** refers to only two or three representative items.

しかしながら、「や」には日本語でいうと「か」に相当する‘or’の意味も含まれているとする辞典もわずかながら存在する。研究社新和英中辞典第五版は「や」を以下のように定義している。

(‘ya’:)［…と］and［…または］or…,
彼女はバッグに歯ブラシや当座の着替えを放り込むとすぐに出かけた。
Throwing a toothbrush and a change of clothes into her bag she hurried out.
君なら百万や二百万の金はすぐに用意できるだろう。
Putting together one or two million yen is nothing for someone like you.

2.2 先行研究に見る「や」

並立助詞「や」を扱った先行研究には、「と」と「や」の使い分けに関するものが圧倒的に多い。寺村 (1991)は、「と」を「全部列挙」、「や」を「一部列挙」と区別している。また、益岡・田窪 (1992)は、「と」を「総記」、「や」を「例示」とし、柏木 (2006)は「と」を「全体化」、「や」を「類化」と区別している。いずれも寺村が「二つ以上の名詞を並立的に結びつけるのに「ヤ」が用いられるときは、それらの名詞が、あるセットの具体例として、そのメンバーの一部として取り上げられていることを示す。」と説明するような見解である。

一方、「と」と「や」を使い分けず、全集合取り上げの「と」の代わりに「や」が用いられる例も研究されている。朴（2006)は、以下の例を「や」の「ぼかし機能」であるとしている。

「もう池本屋も、広島や長崎が原爆されたことを忘れとる。みんなが忘れとる。あのときの灼熱地獄―あれを忘れて、何かこのごろ、あの原爆記念の大会じゃ。あのお祭り騒ぎが、わしゃあ情けない」（黒い雨）

国広 (1967)は、大した意味もなく「や」を用い、表現を柔らげる用法は日本語表現の一つの特徴であるとの見方を示す。以下の例では、実際に調査したのが太平洋と大西洋だけであるかははっきりしないが、「や」が一つの漠然とした表現として用いられている可能性があるとしている。

> ペターソン教授といえば、深海の底土の研究で世界的に有名である。先年のアルバトロス号の深海体験では、太平洋や大西洋の深部から、ボーリングで土を採って、いろいろな研究をしている。

国広 (1967)はさらに、研究社新英和中辞典の定義と一致する「や」の用法も紹介する。

> 『や』は……いわば‘and’と‘or’の両方にまたがるものである。しかし英語では、‘and’と‘or’は対立概念であって、英語と同じ平面に立って考えれば両方にまたがる概念というものはあり得ない。「や」は全く異なった平面に属しているわけである。

以上のように、並立助詞「や」は、一般的にはあるカテゴリーの例を示す場合に用いられ、「と」・「か」とは使い分けられるものであるが、文脈により、「と」や「か」に相当する意味で用いられることもあるということがいえる。

3. BCCWJ データに基づいた並立助詞「や」の分析

3.1 本研究の目的

本研究の目的は、先行研究の外観を踏まえ、並立助詞「や」の使用の実態を調査し、分析することである。特に、先行研究によれば「と」または「か」が好まれるべき場合に「や」が用いられている事例について分析し、「や」が選択される理由について考察する。

3.2 使用するデータ

本研究では、「現代日本語書き言葉均衡コーパス」（BCCWJ）モニター公開データ（2009 年度版）上の用例を用いて調査を行う。同コーパスには、以下の表のような 57,807 のサンプルが収められている。

「BCCWJ モニター公開データ 2009」に収められたサンプルの内訳（表 1）

サブコーパス名	メディア	収録サンプル数
生産実態(出版)サブコーパス	「書籍」	4,459 サンプル
流通実態(図書館)サブコーパス	「書籍」	5,110 サンプル
非母集団(特定目的)サブコーパス	「書籍」(ベストセラー)	854 サンプル
	「白書」	1,500 サンプル
	「Yahoo!知恵袋」	45,725 サンプル
	「国会会議録」	159 サンプル
	計	57,807 サンプル

「『現代日本語書き言葉均衡コーパス』モニター公開データ(2009 年度版)
書誌情報・サンプル情報・著者情報について」より引用

本研究では、上記サンプルのうち「Yahoo!知恵袋」以外は、それぞれ約 100 分の 1 に相当するデータをランダムに抽出し、調査の対象とした（生産実態(出版)サブコーパス:45 サンプル、流通実態(図書館)サブコーパス：51 サンプル、非母集団(特定目的)サブコーパス「書籍」（ベストセラー）：9 サンプル、「白書」: 15 サンプル、「国会会議録」: 2 サンプル）。「Yahoo!知恵袋」に関しては、田野村(2009)が「無作為抽出によって収集された日本語テキストには内容上公開になじまないものが多数含まれる」と述べるように、「名誉棄損や侮辱に当たるもの」「差別的なもの、非道徳性の高いもの」「猥褻なもの」が含まれているため、データの約 200 分の 1 に相当する 225 サンプルを調査の対象とした。

3.3 分析方法

分析の事前準備として、KH Coder の KWIC コンコーダンスを利用し、対象データより「名詞 1＋助詞『や』＋名詞 2」、「名詞 1＋助詞『と』＋名詞 2」、「名詞 1＋助詞『か』＋名詞 2」の用例を抽出した。KH Coder には、形態素解析ソフト Win Cha が含まれているため、異なった品詞のデータが抽出されることはなかったが、助詞が並立助詞以外の用途で使われている例(ex.「商品投資販売業者や顧問業者と同等の行為規制が必要であり、この規定は削除すべきであるというふうに私は強く主張いたします。」など)が抽出された。そのため、抽出されたすべての例を目視検証し、必要のないものを取り除いた。その後、それぞれの用例についての検討を行った。

4. 結果と分析

4.1 結果

対象データでのそれぞれの助詞の使用頻度は以下の通りであった。

使用データより抽出された並立助詞「や」「と」「か」の使用頻度（表２）

サブコーパス名	メディア	や	と	か
生産実態(出版)サブコーパス	「書籍」	78	74	1
流通実態(図書館)サブコーパス	「書籍」	75	78	1
非母集団(特定目的)サブコーパス	「書籍」(ベストセラー)	10	9	0
	「白書」	9	12	1
	「Yahoo!知恵袋」	18	9	2
	「国会会議録」	15	9	2
	計	205	191	7

4.2　分析

4.2.1　一部列挙の「や」

辞書や先行研究でも一般的であった一部列挙の用例は、数多く見られた。

また、利用者による高山植物や盆栽用植物等の盗採、踏圧による湿原や高山のお花畑等の裸地化などの現象も近年増加の傾向にある。（非母集団「白書」）

時代が進んで中世に入ると、領邦君主権や都市共同体権力に代表される、フランク時代の王権よりも質的に一段と高度な（＝流血裁判権を行使する）公権力が成立し、それらの手による平和の維持がなされるようになったため、刑法史は新たな展開の局面を示すに至る。（流通実態「書籍」）

私たちは、テレビやインターネットによって、世界各地の苦しむ人たちの姿を、リアルタイムで、見たり聞いたりすることができる。（生産実態「書籍」）

英検試験管の資格を持った人のみですよ。中学校や高校の教師で持っている人も結構います。（非母集団「Yahoo!知恵袋」）

これらの用例は、「など」等の言葉が使われていない場合でも、一部列挙であることが理解できる。たとえば、「テレビやインターネット」という表現により、それらに類する「新聞」「ラジオ」「雑誌」などが容易に想起されるものである。

4.2.2　「や」と「と」の区別

辞書や先行研究であれば、全部列挙では「と」が好まれるが、抽出された「や」の用例の中

に、全部列挙であると考えられるものは少なかった。以下はその例である。

普段、手に取りそうもない本はあるか。どれも小説や随筆の類いで、そのどれに手がのびるか、予測はつかぬ筈だ。（流通実態「書籍」）

全部列挙か一部列挙のどちらにでも取れる例は非常に多く見受けられた。

そういう中で、現在そういう地方における官公需の中小企業や地場企業に対する発注というのはどういうぐあいになっておるのか、お調べのあれがありましたらひとつ教えていただきたいと思います。（非母集団「国会会議録」）

つい先日、ロス警察一番の弱虫コンクールというものが催され、刑事や警官のすべてが無理矢理参加させられ、このサムが見事（？）一等賞に選ばれた。（流通実態「書籍」）

また、「と」の用例には「根拠のある部分と根拠のない部分」「前期と後期」「男性と女性」などのように、列挙されている名詞自体が全部列挙であることを示唆する用例が数多く見られたが、「や」に関してはそのような用例は見られなかった。

一方で、「や」と「と」が、同様の文脈で用いられている例が抽出された。

［ページ設定］では、１ページの行数や１行の文字数を設定することができます。行数や文字数は、用紙サイズや余白、フォントのサイズに関係しています。（生産実態「書籍」）

ワードでは１行の文字数や１ページの行数が自動的に設定されています。文字数、行数を自分で設定したい場合には、［ページ設定］を使います。文字数と行数は、用紙サイズと余白の幅、そしてフオントの大きさに関係しています。（生産実態「書籍」）

多くの先行研究によれば、「と」と「や」は区別され、使い分けられるべきものであるが、実際には上記のように、明確な７理由なしに「や」が用いられている例が多くみられる。

4.2.3 「や」と「か」の区別

「か」には、「決める」「判断する」「見分ける」など選択肢があることを示唆する動詞とともに使われる用例が多くみられる。

また，これとは別に，外国メーカー又は輸入総代理店が，輸入総代理店経由の商品か並行輸入品かを見分けられるように，商品の形状，包装箱のデザイン等を変えたりしているものも見られた。（非母集団「白書」）

しかし、抽出された「や」の用例では、そのような動詞と共起関係にあるものは見られなかっ

た。以下は、「か」に近い意味で「や」が用いられていると思われる用例である。

フォントの種類やサイズを変えるなら、［文字数と行数］タブで［フォントの設定］をクリックして、詳細な設定をします。５０行に収めたいのに上限が４８行であった場合には、余白や行間を狭めればよいでしょう。（生産実態「書籍」）

第２に、不要なファイルやアプリケーションソフトは削除しましょう。
（生産実態「書籍」）

「五歳ですか…幼稚園や保育園には？」（流通実態「書籍」）

4.2.4 「や」・「と」・「か」の区別

辞書の定義では、「や」と「と」、「か」はそれぞれ異なった意味を持つが、実際にはこの３つの助詞の使い分けが曖昧な用例も見受けられた。

買い置きしておくと便利な缶詰や乾物を野菜炒めにプラス。味つけ役になったり、具になったりと大活躍。うまみとボリュームをアップしてくれます。（生産実態「書籍」）

残念ながら（切手は）換金はできません。一枚につき５円支払って、別の切手や葉書には交換できますが…（非母集団「Yahoo!知恵袋」）

1つ目の例では、野菜炒めにプラスするのは「缶詰と乾物両方」であるとも「缶詰か乾物いずれか一方」であるとも解釈することができる。このように「や」を用いて「と」を意味することも「か」を意味することも可能である。

5. 結語

本研究では、「現代日本語書き言葉均衡コーパス」（BCCWJ）モニター公開データ（2009年度版）上の用例を用い、並立助詞「や」の使用の実態調査を行った。特に、同じ並立助詞である「と」と「か」との使い分けに着目し、辞書の定義や先行研究によれば「と」や「か」が好まれるべき場合に「や」が用いられている事例を分析した。その結果、辞書や先行研究では、３つの並立助詞ははっきりと区別されるものであるにもかかわらず、その実際の使い分けは、非常に曖昧であるということがわかった。「と」や「か」が好まれる例で、「や」が用いられている例が数多いことから、並立助詞「や」には、「と」と「か」の意味も含まれていると考えることができる。先行研究では、「や」を一部列挙とし、全部列挙の「と」と明確に区別するものが圧倒的に多いが、本研究での用例を見る限り、国広 (1967)の「いわば‘and’と‘or’の両方にまたがるもの」との主張が「や」の使用の実態に合っていると思われる。朴（2006)が「ぼ

かし機能」、国広 (1967)が「表現を柔らげる用法」と説明するように、「や」と他の並立助詞をあえて使い分けずに曖昧な表現を使うことは、日本語の特質の一つであるとの見解も、実情に合った見方であると考えられる。

文献

安藤淑子（2001）「中級レベルの作文に見られる並列助詞「や」の問題点―「と」の用法の比較を通してー」　『日本語教育』108：42-50.

ハロー通訳アカデミー秋期第 5 回校内模擬試験　(2008 年 5.6 月実施)

ハロー通訳アカデミー秋期第 2 回校内模擬試験　(2008 年 2 月実施)

市川保子（1991）「並立助詞『と』と『や』に関する一考察」　『文芸言語研究 言語篇』20：61-79.筑波大学文芸・言語学系.

海外技術者研修協会（1990）「新日本語の基礎Ⅰ」東京：スリーエーネットワーク

柏木成章（2006）「『全体化』と『類化』：並立助詞論、特に『と』・『や』を中心として」『別科日本語教育 : 大東文化大学別科論集』　8：99-107.

国広　哲弥（1967）「'And'と『と・に・や・も』―日英両語語彙の比較―」『言語研究』50：34-49

Martin Collick・David P.Dutcher・田辺宗一・金子　稔（編）（2002）「新和英中辞典第 5 版」

益岡隆志・田窪行則 (1992)「基礎日本語文法―改訂版」　東京：くろしお出版.

中川裕志　武藤伸明 (1997)「並立助詞『と、や、に』の意味の形式的分析」pp.2770-2779『電子情報通信学会論文誌. D-II, 情報・システム, II-情報処理』

朴 点淑　(2006)「現代日本語における並立助詞『と』『や』」『岡山大学言語学論叢』12: 51-62. 岡山大学.

スー・A・川島 (1999)「てにをは辞典 - A Dictionary of Japanese Particles」東京：講談社インターナショナル

田野村忠温(2009)「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」日本知能学会誌 24 巻 5 号

寺村秀夫　(1991)「日本語のシンタクスと意味Ⅲ」東京：くろしお出版.

関連 URL

KH Coder (http://koichi.nihon.to/psnl/)

「オノマトペ＋する」の構文的特徴
―「スル」の取りうる形式に焦点を当てて―

黄慧（東京外国語大学大学院博士後期課程）

A Study on Japanese “Onomatopoeia+ -suru”

HUANG　HUI　(Tokyo University of Foreign studies)

1.　はじめに

本発表は、「オノマトペ＋スル」の構文的特徴に焦点を当て、「スル」の取りうる形と「オノマトペ＋スル」全体の構文的特徴との間にどのような関係があるのかについて考察することを目的とする。本発表はその中間報告段階に当たる。

2.　先行研究

「オノマトペ＋スル」の諸特徴について言及したものには、西尾(1981)、中北(1991)、影山(2005)、笹本(2007)、大塚(2009)、譙(2010)などがある。

西尾(1981)は、「オノマトペ＋スル」の動詞的用法について概観したものであり、大塚(2009)、譙(2010)はアスペクトに焦点を当てた研究である。中北(1991)はオノマトペと「スル」との結合について述べた研究である。影山(2005)は、語彙概念構造を用いて「オノマトペ＋スル」について動詞分類を行っている。オノマトペの名詞修飾について言及しているものには笹本(2007)がある。

しかし、これらの先行研究は「オノマトペ＋スル」の諸特徴を取り上げているものの、あまり数量的調査を重視していないことが問題点としてあげられる。

本発表では、大規模書き言葉均衡コーパスを用いて、数量的調査を行うことで「オノマトペ＋スル」の諸特徴を明らかにし、「スル」が取りうる形式とオノマトペの間にはどのような関係があるのかについて見ていく。

3.　研究方法・研究対象

3.1.　検索語彙の選定およびコーパスの詳細

擬音語・擬態語辞典 7 冊を用いて、この 7 冊の辞書いずれにも収録されているオノマトペ 595 語を検索対象とし、2009 年度版『現代日本語書き言葉均衡コーパス　モニター公開データ(以下 BCCWJ と称する)』を使用し、用例を収集することにする。BCCWJ は、「書籍」、「白書」、「Yahoo!知恵袋」、「国会会議録」で構成されている。

検索方法としては、「コーパス全体を収集対象に、前後文脈 100 文字まで」という設定で、595 語のオノマトペを 1 語ずつ検索し、ヒットした用例を全て収集することにした。

3.2. コーパスから収集したオノマトペの詳細

595 語のオノマトペを検索した結果、全部で 73,878 例の用例を収集することができた。そのうち，「スル」と共起した用例は 8,674 例ある。595 語のうち、「スル」と共起して現れたオノマトペの異なり語数は 305 語である。本発表では、収集した用例のうち、その用例数が 20 例以上のオノマトペ 91 語に絞って用例分析を行う。今回はこれら 91 語のオノマトペをランダムで 10 例ずつ、計 910 例について分析を行った。

4. 考察

910 例の「オノマトペ＋スル」の用例をまず、「文末述語用法」と「連体修飾用法」「その他」に分類した。以下詳細を表 1 に示す。

表 1 「オノマトペ＋スル」の用法

詳細	用例数	割合
文末述語用法	219	24%
連体修飾用法(名詞修飾)	312	34%
その他(条件節、ナガラ節、従属節等)	379	42%
合計	910	100%

「オノマトペ＋スル」は、全体的に文末述語用法として用いられるものよりも連体修飾用法として用いられるものが多いことが分かる。オノマトペは人の様子や物の状態を描写するものであるため、連体修飾用法として名詞を修飾することで属性規定していると考えられる。その他として分類されているものの中には、条件節、ナガラ節、従属節内に現れるものなどがある。影山(2005)では、「オノマトペ＋スル」の中にはスル形よりもサセル形を多用するオノマトペがあると述べている。910 例のうち、サセル形の用例は 58 例あるが別扱いすることにする。(詳細は後述する) 「オノマトペ＋スル」の全用例のうち、「スル／シタ／シテイル／シテイタ」の形をとる用例は、表 2 のようになっている。

表 2 スル／シタ／シテイル／シテイタの詳細

スルの形式	用例数	全体との割合
シタ	279	31%
シテイル	132	15%
スル	125	14%
シテイタ	48	5%
その他	326	35%
合計	910	100%

以上、収集した「オノマトペ＋スル」の用法およびスルがどのような形を取っているのかについて数量的考察を行った。4.1.から主に文末述語用法と連体修飾用法における「スル／シタ／シテイル／シテイタ／サセル／サレル」の形について考察を行う。

4.1. 文末述語用法

「オノマトペ＋スル」の文末述語用法には、「スル／シタ、シテイル／シテイタ」以外にも、「オノマトペ＋シハジメル／シテシマウ／シカケル」など様々な形式がある。さらに、文末にモダリティ形式を伴った「シタラシイ／スルベキ」などの形式も含まれており、一般的な動詞と同じ文法的特徴を持っていることが確認できる。文末述語用法の上位にあるものを次の表 3 に示す。

表 3　文末述語用法の詳細

	詳細	用例数	割合
1	シテイル	56	26%
2	シタ	43	20%
3	スル	43	20%
4	シテイタ	29	13%
5	その他	48	21%
	合計	219	100%

文末述語用法として用いられた「オノマトペ＋スル」はシテイルの用例が最も多い。「うろうろシテイル」のように動的な意味を持っているオノマトペが現在進行形として用いられる場合と「じめじめシテイル」のように物の属性を表す場合と両方に用いられるため、シテイルの用例数が多くなっていると考えられる。

文末述語用法において、シタとスルの形で現れた用例数は同じぐらいの割合を占めている。これは文末述語用法おけるシタおよびスルは、テンス・アスペクトの分化があるものが多いことに起因していると思われる。述語用法として用いられたオノマトペを金田一(1978)の下位分類に従い、分類を行った。以下、表 4 に詳細を示す。

表 4　スルの形式とオノマトペの性質

	シテイル	割合	シタ	割合	スル	割合	シテイタ	割合	合計
擬態語	17	30%	7	16%	7	16%	9	31%	40
擬容語	22	40%	12	28%	9	21%	9	31%	52
擬情語	17	30%	24	56%	27	63%	11	38%	79
合計	56	100%	43	100%	43	100%	29	100%	171

全体的に人間を主体としている擬容語と擬情語が 8 割を占めていることが分かる。それは、文末述語用法として用いられるオノマトペは人の気持ちを表す「いらいらスル／がっかりシタ」のような擬情語が多いことに影響を受けていると思われる。シタの形で文末に現れたオノマトペは、ある出来事の過去を表すものと「手の中がぬるぬるシタ」のような物の状態を表すものがある。

シテイルとシテイタの形で用いられた擬態語、擬容語、擬情語の間の割合はあまり差が無く、シタとスルの形で用いられた擬態語、擬容語、擬情語の間の割合にもあまり差がない。しかし、シテイルとシテイタ、そしてスルとシタこの両グループ内における下位分類の割合にはかなりの差がある。文末述語用法におけるスルとシタの形は、半分以上が人間の感情を表している。これらはスルで今現在と将来のどちらも表すことができ、シタの形と対立を成しているためであると言える。

4.2. 連体修飾用法

笹本(2007)は「オノマトペ＋スル」と「オノマトペ＋ノ／ナ」の形で名詞を修飾するものに焦点を当てた研究である。そのため、笹本はテンス・アスペクトの分化がない、いわゆる属性規定するものに対象を絞っている。本発表は、「オノマトペ＋スル」の諸特徴に焦点を当てたものであるため、対象を限定せずに「オノマトペ＋スル」の諸形式で名詞を修飾するもの全てを調査対象とした。

「オノマトペ＋スル」が名詞を修飾している 312 例を用いて、「オノマトペ＋スル」の諸形式と対象物との間にはどのような関係があるのかについて考察を行う。修飾される名詞を分類する際には、国立国語研究所の『分類語彙表』を用いることにする。312 例の「オノマトペ＋スル」の連体修飾用法におけるスルの諸形式は次のようになっている。

表 5　連体修飾用法とスルの諸形式(サレル形の用例数が少ないため小数点第一位まで計算)

詳細	用例数	割合
シタ	219	70.2%
スル	47	15.1%
シテイル	33	10.6%
シテイタ	6	1.9%
サセル	4	1.3%
シテクル	2	0.6%
サレル	1	0.3%
合計	312	100%

表 5 から分かるように、連体修飾用法として用いられた「オノマトペ＋スル」の諸形式はシタに集中している。シタを用いて名詞を修飾するものは 219 例中 160 例が擬態語であるこ

とが確認できた。「ごつごつシタ手／ぬるぬるシタ液体」のように、あるものの状態や性質を描写するものが多かい。つまり、ある物の状態をシタの形で属性規定する用法が非常に多いことになる。これがシタの用例数が占める割合が圧倒的に多くなった原因であると言える。「オノマトペ＋スル」の形がどのような名詞を修飾しているのか、分類語彙表を用いて下位分類を行った。その結果を次の表 6 に示す。

表 6　連体修飾用法における名詞の分類

名詞の下位分類	用例数	割合
抽象的関係	116	37%
自然物および自然現象	85	27%
人間活動　精神および行為	46	15%
生産物および用具	34	11%
人間活動の主体	31	10%
合計	312	100%

連体修飾用法として用いられた用例の中には、「抽象的関係」の名詞が最も多かったが、「抽象的関係」のなかでも「事柄」を表す「こと／もの」など形式名詞を修飾している用例が目立っている。「抽象的関係」の名詞を修飾するオノマトペは擬態語が最も多く 116 例中 62 例で半数以上を占めている。擬態語が修飾する名詞は「関係／雰囲気」などがある。

「自然物および自然現象」は、「身体」を表す名詞が含まれているため用例数が多くなっていると言える。つまり、人間の「指、手、足、頭、体、肉体、目、口」など身体名詞の特徴や様子を描写するオノマトペが非常に多いこと、そして「味／色」などを表す名詞を修飾するオノマトペが多いことで「自然物および自然現象」の占める割合が高くなっている。「自然物および自然現象」を描写するオノマトペ 85 例中、71 例がシタの形で用いられている。さらにシタで用いられた 71 例中、62 例が擬態語であることが確認できた。

シタ形、スル形、シテイル形が修飾する名詞について分類すると次の表 7 のようになっている。

表 7　修飾される名詞の下位分類

	下位分類項目	用例数	割合
シタ(219)	自然物および自然現象	71	32%
	抽象的関係	67	31%
	人間活動　精神および行為	38	17%
	生産物および用具	32	15%
	人間活動の主体	11	5%
		219	100%

シテイル(47)	抽象的関係	23	49%
	人間活動　精神および行為	8	17%
	自然物および自然現象	8	17%
	人間活動の主体	7	15%
	生産物および用具	1	2%
		47	100%
スル(33)	抽象的関係	21	64%
	人間活動の主体	9	27%
	自然物および自然現象	3	9%
		33	100%

「自然物および自然現象」の名詞はシタの形で用いられやすいことはすでに前述したとおりである。スルおよびシテイルにおいて「抽象的関係」の名詞が多いのは、擬情語で用いられたものが多く、擬情語は主に人間の感情、感覚、気持ちなどを描写するものが多いからであると推測できる。

4.3.　その他

ここでは、「オノマトペ＋サセル」と「オノマトペ＋サレル」の形を取り上げる。

サセル形で現れた用例は全部で 58 例あるが、異なり語数は 25 語ある。サセル形で現れた用例のうち、使用回数が多いオノマトペを次に示す。表 8 の 7 語以外のオノマトペはすべて 1 例ずつである。

表 8　サセル形で現れたオノマトペ

	オノマトペ	用例数	割合 (10 例中)
1	ぱちくり	9	90%
2	ぱちぱち	8	80%
3	ひらひら	8	80%
4	ばたばた	4	40%
5	きょろきょろ	3	30%
6	きらきら	3	30%
7	ぎらぎら	3	30%

影山(2005)では、これらのオノマトペは、目、口、手、足など身体名詞を修飾する際にスル形よりもサセル形を多用すると説明している。表 8 の「ぱちくりサセル」および「ぱちぱちサセル」はすべて「目」を修飾するものであり、「ばたばたサセル」はほとんど「足／手」

を修飾するものである。これらは、目であれば「ぱちくり」が用いられると言う修飾される名詞と修飾するオノマトペの間には強い結びつきがあると考えられる。このように、身体名詞を修飾するオノマトペはサセル形を多用する。サセル形で用いられたその他の用例を分析したところ、「びっくりサセル／がっかりサセル／うっとりサセル」のように感情・感覚を表すものが多かった。

次に、サレル形で現れた用例について考察を行う。サレル形で現れた用例は全部で 9 例、異なり語数は 3 語ある。「あっさりサレル／やきもきサレル」の用例は、敬語のサレル形であるため、考察対象から排除する「ちやほや」は 10 例中、7 例がサレル形で用いられている。「ちやほや」はスル形よりもサレル形を多用することは 1 つの特徴として挙げることができる。

以上、「オノマトペ＋スル」の諸形式がどのように用いられているのかについて見てきた。

5. おわりに

本稿では、「オノマトペ＋スル」の構文的特徴に焦点を当て、スルの取りうる形が「オノマトペ＋スル」の構文的特徴とどのような関係があるのかについて見てきた。91 語のオノマトペをそれぞれ 10 例ずつ抽出し、910 例の「オノマトペ＋スル」について分析を行った。

全体的に文末述語用法および連体修飾用法が占める割合はほとんど同じである。連体修飾用法が多く用いられているのは、物や人の属性を規定する用法が連体形を取っているためであると言える。文末述語用法ではシテイルが最も多く用いられているが、これはある出来事の進行形として用いられる場合と属性規定する形として用いられる場合があるため、その他の形式より占める割合が高いと言える。

連体修飾用法はシタで現れたものが約 70%を占めている。ある物の状態や様子をシタの形で属性規定する用法が非常に多いことで、シタの用例数が圧倒的に多くなったと考えられる。連体修飾用法を用いて修飾する名詞は「抽象的関係」や「自然物および自然現象」が占める割合が高いことが分かった。これは、「抽象的関係」においては、「事柄」を表す形式名詞「もの／こと」などが非常に多かったこと、そして、「自然物および自然現象」においては「身体名詞」を表す「体、目、足、手、顔、頭」などが非常に多かったことに起因している。

サセル形、サレル形で現れるオノマトペは、それぞれ特徴を持っている。サセル形でよく用いられるオノマトペは主に「目、足、手、口」などの身体名詞が多い。主にサレル形で現れるオノマトペは「ちやほや」1 語のみであるが、サレル形を多用することは 1 つの特徴として挙げることができる。

本発表は、「スル」の取りうる形と「オノマトペ＋スル」全体の構文的特徴は互いに影響し合っていることについて見てきた。しかし、スルの諸形式の提示にとどまっていること、そして、この 910 例からある程度の傾向性を見ることはできたものの、数量的に不十分であり、明確な結果を得ることができなかったことは今後の課題にしたいと思う。今後はもっと多くの用例を分析し、考察結果に加えてきたい。

参考文献

【辞書類】

浅野鶴子編　金田一春彦概説(1978)『擬音語・擬態語辞典』, 角川書店.
阿刀田稔子, 星野和子編(1995)『擬音語擬態語使い方辞典』, 創拓社.
天沼寧編(1985)『擬音語・擬態語辞典』, 東京堂出版.
小野正弘(2007)『擬音語・擬態語 4500 日本語オノマトペ辞典』, 小学館.
曹金波(2008)『標準日本語擬声語・擬態語』,中国: 大連理工大学出版社.
飛田良文, 浅田秀子編(2002)『現代擬音語・擬態語用法辞典』, 東京堂出版.
山口仲美編(2003)『暮らしのことば　擬音・擬態語辞典』, 講談社.

【文献類】

大塚望(2009)「擬音語・擬態語と「する」の結合について」, 『日本語日本文学』19: 17-36, 創価大学日本語日本文学会.
影山太郎(2005)「擬態語動詞の語彙概念構造」, 『第 2 回中日理論言語学研究会』, ハンドアウト: 1-9.
金田一春彦(1978)「概説」浅野鶴子編, 『擬音語・擬態語辞典』, 角川書店.
黄慧(2010) 「オノマトペの名詞修飾について　ー「オノマトペ＋する」と名詞修飾ー　」『Association of Teachers of Japanese』配布資料
譙燕(2010)「「オノマトペ＋スル」動詞の分類と用法」, 『日本語の擬音語・擬態語に関する研究』: 73-89, 中国北京, 学苑出版社.
田守育啓・ローレンス・スコウラップ(1999)『オノマトペー形態と意味ー』,くろしお出版.
中北美千子(1991)「擬音語･擬態語と形式動詞「する」の結合について」,『国文目白』31: 247-256, 日本女子大学国語国文学会.
西尾寅弥(1981)「「擬音語・擬態語＋する」の形式について」, 『語学と文学』20: 82-96, 群馬大学語文学会.
星野和子(2005)「擬態語の文法」, 『駒沢女子大学研究紀要』12: 185-198, 駒沢女子大学.
宮地裕(1978)「擬音語・擬態語の形態論小考」, 『国語学』115: 33-39, 国語学会.
楊淑雲(1993)「「擬態語＋する／なる」の形式について」『東北大学文学部日本語学科論集』3: 95-106, 東北大学.

【使用したコーパス】

『現代日本語書き言葉均衡コーパス(2009 年度モニター公開データ)』国立国語研究所.

【使用した分類語彙表】

国立国語研究所(2004)『分類語彙表－増補改訂版－』　付録ＣＤ－ＲＯＭ

感情を表す動詞の考察

韓　金柱（東京外国語大学 大学院地域文化研究科博士後期課程）

A Study on Emotional Verbs

Han Jinzhu
(The Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies)

1. 研究の目的

感情を表す動詞には、ヲ格補語をとるグループ（以下の例（1）a.b)、ヲ格補語もニ格補語もとり得るグループ（以下の例（2）a.b)、ニ格補語をとるグループ(以下の例（3）)の３つがある。また、形態的に対応する形容詞があるかないかという観点からは２つのタイプに分けることができる。１つは、形態的に対応する形容詞があるタイプ（以下の例の（1）a（2）a、対応する形容詞は、それぞれ「懐かしむ」に対して「懐かしい」、「悲しむ」に対して「悲しい」)、もう１つは、形態的に対応する形容詞がないタイプ（以下の例の（1）b（2）b、および例の（3））である。（以下、分析の対象となる感情を表す動詞に下線を付して示す。また特に出典を示さない例文は執筆者による作例である。）

（1）a.　故郷を懐かしむ。
　　b.　彼女を呪う。
（2）a.　親友の死{を/に}悲しむ。
　　b.　彼女が失敗したこと{を/に}驚く。
（3）　　試合に飽きる。

本研究では、感情を表す動詞を、その動詞がどのような補語をとるか、また、形態的に対応する形容詞があるかないか、というこれらの２つの観点から分類し、それぞれの意味特徴について探る。さらに、ヲ格補語もニ格補語もとり得るグループについては、ヲ格補語をとる場合とニ格補語をとる場合との違いについても考察を行う。

2. 分析の対象

『類語辞典』より 1,237 語[1]、『感情表現辞典』より 645 語[2]の人の感情に関わる動詞を抽出した。そのうち、『類語辞典』と『感情表現辞典』のどちらにも収録されている 70 語の動詞を分析の対象とした。

[1] 『類語辞典』は動詞・形容詞を 100 のカテゴリーに分類している。そのうち、人の感情が関わるカテゴリーには、計 12 のカテゴリーが示されている（「感じる（感覚・感情）」「愛する・好む（愛情・愛好）」「望む・欲する（欲望）」「喜ぶ・楽しむ（歓喜）」「悲しむ・泣く（悲痛）」「悔やむ・惜しむ（後悔・哀惜）」「苦しむ・悩む（苦悩）」「困る・恥じる（困苦・恥辱）」「嫌う・憎む・怒る（険悪）」「驚く・おびえる（驚嘆・恐怖）」「思う・見込む（思考）」「敬う・信じる（尊敬・信心）」）。ここでは、それら 12 のカテゴリーに分類されているすべての「動詞」の類を抽出した。

[2] 『感情表現辞典』に載っているすべての動詞・動詞句を抽出した。『感情表現辞典』は、「喜、怒、哀、怖、恥、好、厭、昂、安、驚」という 10 類の感情を立てている。

3. 用例収集について

用例については、「『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）」（以下、BCCWJ と記す）および、ウェブ（検索エンジン google を使用）から収集し、考察を行った。BCCWJ ついては、今回は特にコーパスの検索対象を限定せず、「書籍」「白書」「国会会議録」「Yahoo!知恵袋」の全体から用例を収集した。

4. 先行研究の記述の検討

寺村（1982）は、感情を表す動詞について、「～ニ」という形の、感情の動きの誘因を表す補語をとるタイプのものと、「～ヲ」という形の、感情の向かう対象を表す補語をとるものの２つのタイプがあるとしている。しかし、感情を表す動詞は、必ずしもヲ格補語をとるものと、ニ格補語をとるものの２つに、截然と分けられるとは言えない。例えば、「悲しむ」「驚く」などは例（2）のように、ニ格補語とヲ格補語を両方とることが可能である。

清水（2007）は、心理動詞を「ニ格型心理動詞」（「ニ格誘因型」と「ニ格対象型」）「両用型心理動詞」「ヲ格型心理動詞」の計４つの型に分類している。

このように、清水（2007）では、ヲ格補語もニ格補語もとり得るという「両用型」のタイプがあることが指摘されている。また、清水（2007）は「～ト」という引用節をとる場合に着目し、以上の４つの型を以下の２つのグループに分けている。

「刺激－応答」グループ	ニ格誘因型
	両用型心理動詞（ニ格をとる場合）
「主題－説明」グループ	ニ格対象型
	両用型心理動詞（ヲ格をとる場合）
	ヲ格型心理動詞

「刺激－応答」の関係とは、ニ格で標示された名詞句の示すものが刺激となって感情が生まれ、その感情への「応答・反応」が引用節として表出されるという関係のことを指す。例えば、「南極点到達の知らせに喜一さんは『横断成功ではないが初志を貫いたのだから、よくやったとほめてやりたい』と喜ぶ」（清水（2007：29））という文の場合、「南極点到達の知らせ」が刺激となって喜びの感情が生まれ、「横断成功ではないが初志を貫いたのだから、よくやったとほめてやりたい」という応答・反応が表出されることになる。

また、「主題－説明」の関係とは、ヲ格またはニ格で標示された名詞句の示すものが感情を抱いている人にとっての「主題」となり、引用節でそれについての説明・判断が示されるという関係のことを指す。例えば、「バレンタイン監督との再会、アロマーらと並んでの入団発表……。そのすべてを『すごくうれしい』と喜んだ。」（清水（2007：29））という文の場合、監督との再会や入団発表など「そのすべて」がこの喜びの感情を抱いている人にとっての主題となり、「すごくうれしい」がその主題についての説明・判断となっているというわけである。

このように、「両用型心理動詞」については、「ヲ格」をとる場合は「主題（名詞句）－

説明（引用節）」の関係、「ニ格」をとる場合は「刺激（名詞句）－応答（引用節）」の関係を表すとしている。しかし、これは「名詞句」と「引用節」との関係を表すものであり、引用節をとらない場合（例えば、「彼女は母校の優勝に喜んだ」と「彼女は母校の優勝を喜んだ」など）において、両者の違いは何かについては考察されていない。

以上、先行研究での記述を見ると、感情を表す動詞については、以下の３つの問題点が残されていると思われる。①どの動詞がヲ格補語をとり、どの動詞がニ格補語をとるのか、またどの動詞がそのいずれをもとることができるのかという点について整理し、確認する必要がある。②ヲ格補語をとるもの、ニ格補語をとるもの、ヲ格補語もニ格補語もとり得るものの３つのタイプの意味的違いについて踏み込んで分析されていない。③ヲ格補語もニ格補語もとり得る両用型については、ヲ格補語をとる場合とニ格補語をとる場合との意味的な違いについて明らかにされていない。

5. 本稿における分析

第 2 節で抽出した 70 語の動詞を、形態的に対応する感情形容詞[3]があるものと対応する感情形容詞がないものに分類した。その結果、以下のように対応する感情形容詞がある動詞が 28 語、対応する感情形容詞がない動詞が 42 語となった。

表 1　対応する感情形容詞がある動詞（28 語）

動詞	形容詞	動詞	形容詞	動詞	形容詞
哀れむ	哀れな	恐れる	恐ろしい	懐かしむ	懐かしい
憐れむ	憐れな	悲しむ	悲しい	悩む	悩ましい
悼む	いたましい	嫌う	嫌いな	憎む	憎い
厭う	厭わしい	悔いる	悔しい	妬む	妬ましい
愛おしむ	愛おしい	悔やむ	悔しい	恥じる	恥ずかしい
忌む	忌まわしい	苦しむ	苦しい	腹立つ	腹立たしい
疎む	疎ましい	恋する	恋しい	喜ぶ	喜ばしい
恨む	恨めしい	好む	好ましい	煩う	煩わしい
羨む	羨ましい	好く	好きな		
惜しむ	惜しい	楽しむ	楽しい		

表 2　対応する感情形容詞がない動詞（42 語）

飽きる	うろたえる	困る	たまげる	戸惑う	ひがむ	むせる
呆れる	怒る	しおれる	ためらう	萎える	引かれる	もだえる
憧れる	驚く	沈む	照れる	泣く	ひるむ	やく
慌てる[4]	おののく	痺れる	動じる	嘆く	ほころぶ	揺れる

[3]本研究での「感情形容詞」とは、「私は―。」の形で第一人称名詞句を主題として、感情表出文を作ることが可能である形容詞（「悔しい」「痛い」など）のこととする。

[4] 「慌てる」に対応する形容詞「あわただしい」も本稿でいう感情形容詞には当てはまらないため、対応する感情形容詞がない動詞とする。

呻く	怯える	しょげる	尊ぶ[5]	のぼせる	惚れる	酔う
敬う	焦がれる	そねむ	ときめく	呪う[6]	むくれる	笑う

6. 対応する感情形容詞があるタイプ

韓（印刷中）では、対応する感情形容詞があるタイプについて、「～がる」との比較を視野に入れて考察を行った。そこでは、表1に示した対応する感情形容詞がある動詞28語を分析の対象とし、それらの動詞がどのような補語をとるかという観点からヲ格補語をとるタイプとヲ格補語もニ格補語をもとり得るタイプに分類し、それぞれのタイプについて分析を行った。以下、その分析結果をまとめて示す。

まず、対応する感情形容詞がある動詞は、「主体が当該の感情を心の中に抱く」ことを表すと考えられる。（ヲ格およびニ格補語に、＿＿＿を付して示す。）

ヲ格補語のみをとるタイプ（例(1a)）		主体がその補語によって表されるものに対して当該の感情を抱くことを表す（以下の例(4)）
ヲ格補語もニ格補語もとり得るタイプ（例(2a)）	ヲ格補語をとる場合	主体がその補語によって表されるものに対して当該の感情を抱くことを表す(以下の例（5）)
	ニ格補語をとる場合	主体が「～に」という補語によって表されるもの、そのものに対して当該の感情を抱くことを表すのではなく、「～に」という補語によって表されるものによって引き起こされる「状況」の中で、当該の感情を抱くことを表す（以下の例（6）（7）（8））

(4) やがて戦争が終わり、餃子や鍋貼の製法を身につけた日本人が帰還し、当時の味を懐かしみ、餃子などの中華料理を作り始めました。（『Yahoo!知恵袋』）

(5) その時、私は二十七歳でコウスケは八歳でしたから、年若く生んだとしたら私の子どもとしてもあり得ないことではなかったのです。クラス中が、コウスケの転校を悲しみました。転校していく前夜、四年生のお兄ちゃんとコウスケの二人を連れてレストランに夕食を食べにいきました。（『子どもの言葉はどこに消えた？』）。

(4) での「～を」という補語によって表される「当時の味」は主体が「懐かしい」という感情を抱く対象となっている。この場合、「懐かしむ」は、主体となる「（帰還した）日本人」が「～を」という補語によって表される対象「当時の味」を自発的に思い出し、その「当時の味」に対して、「懐かしい」という感情を抱くことを表している。(5) での「～を」という補語によって表される「コウスケの転校」も、「悲しい」という感情を抱

[5] 「尊ぶ」には対応する形容詞「尊い」がある。ただし、形容詞「尊い」は「*私は尊い。」の形で第一人称名詞句を主題として、感情表出文を作ることができない。したがって、この「尊い」は本稿でいうところの感情形容詞には当てはまらないため、「尊ぶ」には対応する「感情形容詞」がないと考える。

[6] 「呪う」に対応する形容詞「呪わしい」も本稿でいう感情形容詞には当てはまらないため、対応する感情形容詞がない動詞とする。

く対象を表している。この場合も、「悲しむ」は、主体となる「クラス中（の生徒）」が「～を」という補語によって表される対象「コウスケの転校」に対して「悲しい」という感情を抱くことを表している。

(6) 私の彼女は人から笑顔がいいねって言われますが、実際には仕事に悩み、毎日のように電話で泣いています。でも、人前では本当に笑顔です。見えないところで泣いているって結構あるもんですよ。（『Yahoo!知恵袋』）
(7) あとの者どもは舌をふるって進みかねている。すかさず飛んでくる毛野のつぶて、二人目の若党が今度は咽喉を破られて血潮を吐いてぶっ倒れた。その手並に恐れた従者ども、たまらず、ぱっと逃げ散って影も形も見えない。（『総里見八犬伝』）
(8) 社会への参加を認められなかった元患者たちは人権をふみにじられたとして国を訴え、裁判の結果、2001 年にはじめて国の責任が認められました。写真は、2001 年 5 月 11 日、熊本地方裁判所の判決に喜ぶ人びと。（『共同通信社提供』）

(6) での「仕事」、(7) での「手並」、(8) での「判決」はいずれも「～に」という補語によって表されている。これらの「～に」という補語によって表されるものは、主体が当該の感情を抱く対象を表すわけではない。例えば、(6) での「悩む」は「～に」という補語によって表される「仕事」そのものに対して「悩ましい」という感情を抱くことを表すとは考えにくい。この場合、「仕事に悩む」は、主体が「仕事」によって引き起こされる状況（例えば、人間関係の難しさ、仕事の困難さなど）の中で、「悩ましい」という感情を抱くことを表すと思われる。(7) での「手並に恐れる」は「手並」によって引き起こされる状況（例えば、自分が殺される恐怖感など）の中で「恐ろしい」という感情を抱くことを表すと思われる。(8) での「喜ぶ」も「判決」によって引き起こされる状況（国の責任が認められることを長い間待ちに待ってきたという経緯を経ての達成感）の中で、「喜び」を抱くことを表す。

7. 対応する感情形容詞がないタイプ

ここでは、まず、分析の対象とする表 2 に示した 42 語の動詞がそれぞれヲ格補語をとるかニ格補語をとるかについて確認する。ヲ格補語をとるかニ格補語をとるかについては、BCCWJ から用例を収集して確認し、さらに、3 名の母語話者にもその可否を判断してもらった。その結果、ヲ格補語のみをとる動詞が 3 語、ヲ格補語もニ格補語もとり得る動詞が 9 語、ニ格補語のみをとる動詞が 30 語であることが分かった。

7.1 ヲ格補語のみをとるタイプ

表2に示した「対応する感情形容詞がない動詞」のうち、ヲ格補語のみをとるタイプには以下のような3語の動詞がある。

敬う　　尊ぶ　　呪う

上記のような動詞は、「～を」という補語によって表されるものが感情の対象[7]となっている。例えば、以下の（9）（10）のようなものである。

（9）家庭においては家長が家族に指図をし、長屋にあっては家主が店子を差配した。指図し差配するかわりに、問題が生じたときには家長や家主が矢面に立って責任を果たした。その姿を見て、こどもは父親を敬い、店子は家主に信を置いた。
（『家族力』）

（10）美しい人をみて、みにくい自分を思い、富みさかえる人をみて、貧しく疲れた自分をなげき、世に入れられている人をみて、世に入れられない己れをかえりみて世を呪うのである。（『良寛』）

（9）での「父親を」はたとえば「父親のせいで」などと言い換えることができない。このことから、「～を」という補語によって表される「父親」は感情の対象となっていると考えられる。つまり、ここでの「敬う」は、対象となる「父親」に対する主体となる「こども」の敬いの気持ちを表す。（10）での「世を」も、「世によって」などと言い換えることができない。つまり、この場合の「呪う」も、対象となる「世」に対する主体の呪いの感情を表す。

7.2　ヲ格補語もニ格補語もとり得るタイプ

表 2 に示した動詞のうち、ヲ格補語もニ格補語もとり得るタイプには以下のような 9 語の動詞が含まれることがわかった。

呆れる　憧れる　怒る　呻く　驚く　ためらう　泣く　嘆く　笑う

7.2.1　ヲ格補語をとる場合

まず、このタイプの動詞がヲ格補語をとる場合について見てみる。この場合も 7.1 と同様、「～を」という補語によって表されるものは、主体の感情の対象を表すと思われる。例えば、以下の（11）（12）のようなものである。

（11）日本が、今、外国人労働者を受け入れないことを怒る外国人は多い。確かにかつての歴史をかえりみれば、そうした外国人たちの怒りも無理はないという気がする。（『世界の教育』）

（12）一ぺんおっこったら、それでこりる、と言うことはないのか、と言って、私を哀れむ人がある。同じことを何べんしたら分かるのだ。好い加減に眼を醒ましても好

[7]山岡（2002）では、「ヲ格」と「ニ格」のそれぞれについて、「のせいで」に置き換えることができるものを「原因格」置き換えることができないものを「対象格」としている。ここで、ヲ格補語、ニ格補語によって表されるものが、感情の対象となっているか、感情の原因となっているかについての判断は、山岡（2002）に従うことにする。ただし、ここでは、「のせいで」の他に、「（な）ので」「によって」「のために」に置き換えることができるものについても、感情の原因となっているとする。つまり、「のせいで」「（な）ので」「によって」「のために」のいずれかに置き換えることができないものは、感情の対象となっているとし、置き換えることができるものは感情の原因となっているとする。

いじゃないか、と言って、私のことを呆れる人もある。(『生きて行く私』)

(11)での「外国人労働者を受け入れないことを」、(12)での「私のことを」は、それぞれ「外国人労働者を受け入れないことによって」、「私のことのせいで」などと言い換えると不自然な文になる。このことから、いずれも「〜を」という補語によって表されるものが感情の対象となっていると考えられる。つまり、(11)での「怒る」は、主体となる人物である「外国人」の「〜を」という補語によって表されるもの「外国人労働者を受け入れないこと」に対する感情を表し、(12)での「呆れる」は、主体となる「人」の「〜を」という補語によって表されるもの「私のこと」に対する感情を表す。

7.2.2 ニ格補語をとる場合

ニ格補語をとる場合には、次の(13)(14)のように、「〜に」という補語によって表されるものは、感情の原因となっていると考えられる。

(13) 上るほどに、道の両側にポツポツと建て売りの別荘が現れ始めた。そして突然、道は二車線になり、うって変わって路面の質もよくなり、しゃれた街路灯が道の両側を飾っていた。浅川はこの変化に驚く。パシフィックランドの敷地内に入ったとたん、贅沢な装飾があちこちに顔を出したのだ。(『リング』)

(14) 決勝戦は大荒れとなった。決勝戦を見ようとつめかけた観客があまりに多く、ピッチに侵入したか、あるいはピッチいっぱいに広がったのだった。試合は翌日に順延となり、この決定に怒った観客が暴動を起こし、スタンドのひとつを焼き払った。(『南米サッカーのすべて』)

(13)での「変化に」は「変化によって」などと言い換えることが可能である。このことから「〜に」という補語によって表される「変化」は、主体の「驚き」が引き起こされる原因となっていると考えられる。つまり、この場合、「驚く」は、「この変化」によって引き起こされる主体の感情を表す。(14)での「決定に」も、「決定によって」と言い換えることが可能である。つまり、(14)での「怒る」は、「〜に」という補語によって表されるもの「決定」によって引き起こされる感情を表す。

7.3 ニ格補語のみをとるタイプ

表 2 に示した動詞のうち、ニ格補語のみをとるタイプには以下のような 30 語の動詞が含まれることがわかった。

飽きる	慌てる	うろたえる	怯える	おののく	焦がれる	困る	沈む
痺れる	しおれる	しょげる	たまげる	照れる	動じる	ときめく	
戸惑う	腹立つ	萎える	逆上せる	ひがむ	引かれる	ひるむ	
ほころぶ	惚れる	むくれる	むせる	もだえる	やく	揺れる	酔う

上記のような動詞は、心的な反応を表す動詞となっている。例えば、以下の(15)(16)

のようなものである。

（15）この頃の信長は、時間のかかる攻城戦と摑みどころのない一向宗徒のゲリラ戦とに飽きて、華々しい野戦決戦を望んでいたのだ。(『ある補佐役の生涯』)

（16）アフリカでは各地で植民地からの独立をめぐって紛争が続き、人種・民族問題も世界の随所で噴出していたのである。そのなかで、人びとは核の脅威に怯え、内乱の戦火におののき、不当な差別と虐待、貧苦と病苦にあえいでいた。

(『新・人間革命』)

（15）の「時間のかかる攻城戦と摑みどころのない一向宗徒のゲリラ戦とに」、（16）の「脅威に」は、7.2.2 と同様、それぞれ「時間のかかる攻城戦と摑みどころのない一向宗徒のゲリラ戦とによって」「脅威によって」と、「によって」を用いて言い換えることが可能である。ただし、これらの補語によって表されるものは、「飽きる」「怯える」などの心の反応を生じさせる原因となっているわけではないと思われる。この場合、「～に」という補語によって表されるものは、当該の心の反応が生じる“状況”を引き起こす原因となっているのではないだろうか。ここでは、それらの補語が表すものにより引き起こされる状況の中で、心的な反応が生じることが描写されていると考えられる。例えば、（15）での「飽きる」は「時間のかかる攻城戦と摑みどころのない一向宗徒のゲリラ戦」によって引き起こされる状況（戦争を続けたくないという拒絶感、戦争による疲労感）の中で生じる心の反応を表す。（16）での「怯える」は「核の脅威」によって引き起こされる状況（核に対する恐怖感）の中で生じる心の反応を表す。

以上、第 7 節では対応する形容詞がないタイプについて、考察を行った。その結果を以下の表にまとめる。

ヲ格補語のみをとるタイプ（例(1b)）		「～を」という補語によって表されるものに対する感情を表す
ヲ格補語もニ格補語もとり得るタイプ（例(2b)）	ヲ格補語をとる場合	「～を」という補語によって表されるものに対する感情を表す
	ニ格補語をとる場合	「～に」という補語によって表されるものによって引き起こされる感情を表す
二格補語のみをとるタイプ（例(3)）		「～に」という補語によって引き起こされる状況の中で生じる心の反応を表す

引用文献

韓金柱（2010）「接尾辞『がる』の意味・用法―様態の『そうだ』と比較して―」『東京外国語大学 大学院博士後期課程論叢 言語・地域文化研究』第 16 号 東京外国語大学大学院 pp.271-284

韓金柱（印刷中）「感情形容詞に対応する「～む」動詞について―「～がる」との比較を視野に入れて―」『東京外国語大学 大学院博士後期課程論叢 言語・地域文化研究』第 17 号 東京外国語大学大学院

柴田武・山田進・加藤安彦・籾山洋介（2008）『講談社 類語辞典』講談社

清水泰行（2007）「心理動詞の格と意味役割の対応・ずれ―『引用構文』における名詞句と引用節の意味関係から―」『日本文藝研究 58（4)』関西学院大学 pp.23-39

寺村秀夫（1982）『日本語のシンタクスと意味 I 』くろしお出版

中村明（1993）『感情表現辞典』東京堂出版

山岡政紀（2002）「感情描写動詞の語彙と文法的特徴」『日本語日本文学 12』創価大学 pp. 23-54

計画班研究活動・成果報告

3月14日（月）　14:00〜16:40

データ班「代表性を有する現代日本語書籍コーパスの構築」

▶山崎 誠

ツール班「書き言葉コーパスの自動アノテーションの研究」

▶松本 裕治

電子化辞書班「多様な目的に適した形態素解析システム用電子化辞書の開発」

▶伝 康晴

日本語学班「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」

▶田野村 忠温

日本語教育班「代表性を有する書き言葉コーパスを活用した日本語教育研究」

▶砂川 有里子

言語政策班「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」

▶田中 牧郎

辞書編集班「コーパスを利用した国語辞典編集法の研究」

▶荻野 綱男

言語処理班「代表性のあるコーパスを利用した日本語意味解析」

▶奥村 学

研究活動・成果の総括：データ班
代表性を有する現代日本語書籍コーパスの構築

山崎 誠 （班 長：国立国語研究所言語資源研究系）†
小椋 秀樹（分担者：国立国語研究所言語資源研究系）
小沼 悦（分担者：国立国語研究所コーパス開発センター）
柏野 和佳子（分担者：国立国語研究所言語資源研究系）
佐野 大樹（分担者：国立国語研究所コーパス開発センター）
高田 智和（分担者：国立国語研究所理論・構造研究系）
冨士池 優美（分担者：国立国語研究所コーパス開発センター）
間淵 洋子（分担者：国立国語研究所コーパス開発センター）
丸山 岳彦（分担者：国立国語研究所言語資源研究系）
山口 昌也（分担者：国立国語研究所言語資源研究系）

Final Progress Report: 'Data Handling' Group

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)
Hideki Ogura (National Institute for Japanese Language and Linguistics)
Etsu Onuma(National Institute for Japanese Language and Linguistics)
Wakako Kashino (National Institute for Japanese Language and Linguistics)
Motoki Sano(National Institute for Japanese Language and Linguistics)
Tomokazu Takada (National Institute for Japanese Language and Linguistics)
Yumi Fujiike(National Institute for Japanese Language and Linguistics)
Yoko Mabuchi (National Institute for Japanese Language and Linguistics)
Takehiko Maruyama(National Institute for Japanese Language and Linguistics)
Masaya Yamaguchi(National Institute for Japanese Language and Linguistics)

１．データ班の目標

データ班の目標は，『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese，以下，BCCWJ)の中核的な部分をなす「書籍コーパス」の構築である。具体的には，現在流通している書籍を対象にそれらを適切に代表する「書籍コーパス」の構築方法を開発し，その方法論に基づき，実際に 5,000 万語規模のコーパスを完成させることである。なお，この「書籍コーパス」は，国立国語研究所で構築する『現代日本語書き言葉均衡コーパス』の一部であり，全体では 1 億語規模のコーパスが構築されることになる。

上述の目標を達成するために，以下の計画を立てた。

(1)母集団を確定し，そこから無作為にサンプルを抽出する（代表性の確保）。
(2)XML（文書構造記述言語）を用いて，分析用の情報を記述する。
(3)長短２つの言語単位により解析を行う。解析の精度は 98%以上とする。
(4)電子化辞書班と連携し, 形態論情報データベースを整備拡張し, 形態素解析用辞書 UniDic の解析精度向上を図る。
(5)総括班と連携し利用許諾依頼を行う。

以下，本稿では書籍コーパスの記述を中心に，BCCWJのそのほかのサブコーパスについても記述し，BCCWJの全体像を解説する。

２．達成状況

今年度末の書籍コーパスの語数は，約6,000万語であり，計画時の目標5,000万語を上回る達成状況となった。サンプリング，電子化ともに終了している。BCCWJのそのほかのサブ

† yamazaki@ninjal.ac.jp

コーパス（以下，SC）と合わせて1億語という目標も達成している。BCCWJ全体の達成状況は表1のとおりである。

表1　BCCWJのサンプル数と語数

サブコーパス（SC）	媒体	サンプル数 (個) [1]	語数(万語)[2]
出版（生産実態）SC	書籍	11,212	2,954
	雑誌	2,483	569
	新聞	1,490	88
図書館（流通実態）SC	書籍	11,242	3,005
特定目的（非母集団）SC	白書	1,500	500
	教科書	483	120
	広報紙	355	400
	ベストセラー	1,696	447
	Yahoo!知恵袋	91,450	1,000
	Yahoo!ブログ	52,680	1,000
	韻文	253	15
	法律	348	100
	国会会議録	159	500
合計		175,351	10,698

３．BCCWJ の設計

３.１　基本方針

BCCWJ の設計に当たって考慮した基本方針は以下の 4 点である。

(1) 現代日本語の縮図となるコーパス
(2) 汎用的な目的に供するコーパス
(3) 公開可能なコーパス
(4) 既存のコーパスとの調和

３.２　BCCWJ の構成

BCCWJは性質の異なる3つのサブコーパスから構成される（図1）。この3つは，書き言葉の実態を捉える上で，異なる観点から設計されたものである。出版（生産実態）SCは，書き言葉が生み出される出版の実態に着目したもので，出版目録等により母集団を決定するものである。出版SCは2001年～2005年を対象とする。図書館（流通実態）SCは，書き言葉

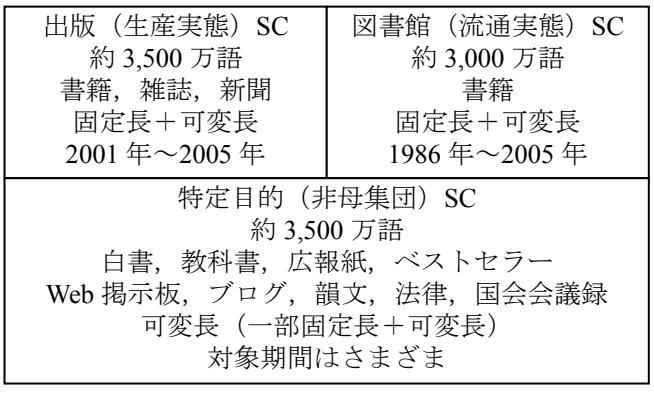


図 1　BCCWJ の構成

[1] 「サンプル数」は実際にサンプリングを行って取得したサンプルの数である。その後の著作権処理の過程で拒否になったサンプル数を含んでいる。
[2] 「語数」は短単位で数えた場合の推計値である（空白・補助記号はカウントしていない）。

が世の中に出回っている状態に着目したもので，生み出された書き言葉の需要という局面をとらえることが狙いになっている。母集団としては，図書館の所蔵目録を利用する[3]。図書館 SC は 1986 年～2005 年を対象とする。特定目的（非母集団[4]）SC は，前述の 2 つの SC では十分な量のサンプルが集まらないもの，あるいは，書き言葉の実態を把握する上で重要なデータを個別に集めるものである。特定目的 SC に収録するデータの対象期間はデータの種類によりさまざまである。

３.３　固定長サンプルと可変長サンプル

コーパスに収録する 1 サンプルの長さが長くなれば集めるサンプルの数が少なくなり，労力も少なくて済む。ただしその場合，いわゆる文脈の影響が出て語の頻度に影響が出る可能性がある（木村 1982，p.237）。また，1 サンプルの長さが一定かどうかという点も検討すべき問題である。一定の長さのサンプルは計量的な分析に向いているが，多くの場合，文章が途中で切れてしまうことになり，文脈を前提とした分析に向かない。この問題を解決するため，BCCWJ では汎用の観点から収録するサンプルの長さを 2 種類設計した。

(1)固定長サンプル

1 つのサンプルの長さを 1,000 字とするサンプル。この場合の 1,000 字には句読点などの記号類は含めない。母集団からの抽出比率が統計的な意味をもつため，語彙表や漢字表などの作成に適している。

(2)可変長サンプル

1 つのサンプルの長さを固定せず，節，章などの文章のまとまり（仮に「記事」と呼ぶ）を 1 サンプルと考えるもの。ただし，無制限に長いサンプルができるとそのサンプルの影響が強く出てしまうため，記事の上限を 1 万字とした。可変長サンプルはテキストの論理構造の把握やテキスト内での役割をもった要素の分析などに適している。

ここで触れた事項以外のサンプリングに関する詳細は，丸山ほか(2011a)(2011b)を参照されたい。

４．各サブコーパスの内容

４.１　出版 SC

出版 SC は，書籍，雑誌，新聞から構成されている。いずれも 2001 年から 2005 年の間に国内で発行されたものが対象である。

４.１.１　書籍

出版 SC の書籍は国立国会図書館の蔵書目録を電子化した「J-BISC」をもとに，以下に該当するものを除外して母集団を決定した。これらは，言語表現が主体でないなど，書き言葉コーパスの設計趣旨に照らして適切でないとして除外したものである。

- 40 ページ以下の書籍[5]
- J-BISC にページ数の記載のない書籍
- 官公庁刊行物のうち市場に流通しない書籍（内部報告書等）
- 学習試験図書
- 電子資料・地図資料など
- 漫画・写真集・図画集
- 複製，復刻など

その結果，317,117 冊，74,911,520 ページが母集団となった。これを NDC(日本十進分類法)および発行年ごとに層別し，各層ごとに何サンプルを取得すればよいかという抽出比を求めた。その際，冊数による比ではなく，推定文字数による割合を利用した。文字数の推定は，次の手順で行った。

(1)2003 年に出版された書籍 65,719 冊について，NDC ごとに判型（大きさ）別に冊数の割合を求める。

---

[3] 書き言葉の受容という局面に着目した母集団も想定できるが，母集団を適切に設定することが難しいため今回の設計には含めていない。

[4]「非母集団」という名称は，この SC に納められたデータが厳密な意味での母集団からのランダムサンプリングによらないものがあるという意味である。

[5] 1 冊全体がサンプルとして収録される可能性があるため。

(2)各判型に含まれる総冊数の 5%をランダムに選ぶ。
(3)選ばれた各冊から 5 ページをランダムに選び文字数を計測する。
(4)NDC・判型ごとに 1 ページあたりの平均文字数を求め（計測対象になっていない部分は，回帰式により推測），それに基づいて，NDC 別の総文字数を算出。
上記の推定の結果，出版 SC「書籍」の母集団は約 48 億 5 千万字と推計された。

表 2　出版 SC「書籍」の構成比

ジャンル（NDC）	冊数	冊数の割合（%）	推定文字数	推定文字数の割合（%）	取得したサンプル数	固定長サンプル語数	固定長ｻﾝﾌﾟﾙ語数による構成比（%）	可変長サンプル語数	可変長ｻﾝﾌﾟﾙ語数による構成比（%）
0 総記	11,132	3.51	1,636,414,548	3.37	363	213,529	3.24	833,197	2.82
1 哲学	18,067	5.70	2,597,610,813	5.35	610	358,824	5.44	1,490,930	5.05
2 歴史	24,624	7.76	4,301,204,340	8.86	926	544,706	8.26	2,447,545	8.29
3 社会科学	62,986	19.86	12,408,321,943	25.56	2,721	1,600,588	24.27	7,194,570	24.35
4 自然科学	28,745	9.06	5,069,594,034	10.44	1,119	658,235	9.98	2,646,734	8.96
5 工業	31,377	9.89	4,615,929,967	9.51	1,008	592,941	8.99	2,447,023	8.28
6 産業	15,332	4.83	2,196,387,437	4.52	480	282,353	4.28	1,232,742	4.17
7 芸術	25,387	8.01	3,258,432,447	6.71	728	428,235	6.49	1,809,129	6.12
8 言語	5,211	1.64	888,800,128	1.83	198	116,471	1.77	466,008	1.58
9 文学	73,716	23.25	9,341,275,486	19.24	2,557	1,504,118	22.81	7,625,880	25.81
n 記録なし	20,540	6.48	2,225,954,208	4.59	502	295,294	4.48	1,347,602	4.56
合計	317,117	100.00	48,539,925,351	100.00	11,212	6,595,294	100.00	29,541,360	100.00

表 2 に示したのは，推定文字数から得られた NDC ごとの推定文字数の割合と最終的に取得したサンプル数に基づく固定長サンプル，可変長サンプルの語数とその割合である。

４.１.２　雑誌

出版 SC の雑誌は，『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）を典拠として利用した。そこに掲載された雑誌のうち，2001 年から 2005 年の間に発行されたもので，社団法人日本雑誌協会に加盟していた出版社が発行したすべての雑誌から新聞，要覧，漫画，非日本語による定期刊行物などを除いたものが母集団である。書籍と同様の文字数調査の結果，この母集団の総文字数は 10 億 5 千万字と推計された。表 3 に推定文字数から得られた雑誌のジャンルごとの割合と，最終的に取得したサンプル数を示す。

表 3　出版 SC「雑誌」の構成比

ジャンル	タイトル数	総冊数	推定文字数	推定文字数の割合（%）	取得したサンプル数	固定長サンプル語数	固定長ｻﾝﾌﾟﾙ語数による構成比（%）	可変長サンプル語数	可変長ｻﾝﾌﾟﾙ語数による構成比（%）
1.総合	833	38,383	7,421,447,806	70.58	1,786	1,050,588	71.93	4,111,719	72.29
2.教育	163	5,456	877,875,592	8.35	193	113,529	7.77	472,600	8.31
3.政治	57	3,168	456,459,405	4.34	114	67,059	4.59	208,197	3.66
4.産業	12	599	110,640,958	1.05	25	14,706	1.01	33,200	0.58
5.工業	170	7,101	1,468,293,360	13.96	323	190,000	13.01	790,200	13.89
6.厚生	24	1,072	180,964,513	1.72	42	24,706	1.69	71,569	1.26
合計	1,259	55,779	10,515,681,634	100.00	2,483	1,460,588	100.00	5,687,485	100.00

４.１.３　新聞

出版 SC の「新聞」は，『全国新聞ガイド』（社団法人日本新聞協会発行）等を参考に，2001 年～2005 年に発行された全国紙，ブロック紙，地方紙から以下の 16 タイトルを母集団として選んだ[6]。

・全国紙：朝日新聞，毎日新聞，読売新聞，日本経済新聞，産経新聞
・ブロック紙：北海道新聞，中日新聞，西日本新聞
・地方紙：河北新報，新潟日報，京都新聞，神戸新聞，中国新聞，高知新聞，愛媛新聞，琉球新報

[6] 日本経済新聞，愛媛新聞は著作権処理の都合で採録対象から外した。

この母集団の総文字数は約 6 億 4 千万字と推計された。表 4 に推定文字数から得られた新聞のジャンルごとの割合と，最終的に取得したサンプル数を示す。

表 4　出版 SC「新聞」の構成比

ジャンル	タイトル数	総冊数	推定文字数	推定文字数の割合 (%)	取得したサンプル数	固定長サンプル語数	固定長ｻﾝﾌﾟﾙ語数による構成比 (%)	可変長サンプル語数	可変長ｻﾝﾌﾟﾙ語数による構成比 (%)
全国紙	5	15,950	2,417,622,461	37.68	550	323,529	36.91	345,956	40.02
ブロック紙	3	9,570	1,296,592,154	20.21	305	179,412	20.47	162,057	18.75
地方紙	8	24,105	2,701,855,499	42.11	635	373,529	42.62	356,351	41.23
合計	16	49,625	6,416,070,114	100.00	1,490	876,470	100.00	864,364	100.00

４.２　図書館 SC

図書館 SC は，書籍のみから構成される。書き言葉の流通実態を図書館の所蔵状況で捉えようとするものである。所蔵目録のデータは，東京都立中央図書館の作成した「ISBN 総合目録」を利用した。図書館 SC の母集団は，1986 年から 2005 年までの 20 年間に発行された書籍のうち，東京都内の 13 自治体以上で共通に所蔵されている 335,721 冊[7]，推計総文字数 47 億 9 千万字とした。これは，出版 SC「書籍」の部分と母集団からの抽出比およびサンプルサイズを揃えるためである。表 5 に推定文字数から得られた書籍のジャンルごとの割合と，最終的に取得したサンプル数を示す。

表 5　図書館 SC「書籍」の構成比

ジャンル (NDC)	冊数	冊数の割合 (%)	推定文字数	推定文字数の割合 (%)	取得したサンプル数	固定長サンプル語数	固定長ｻﾝﾌﾟﾙ語数による構成比 (%)	可変長サンプル語数	可変長ｻﾝﾌﾟﾙ語数による構成比 (%)
0 総記	7,438	2.22	1,003,528,880	2.10	249	146,471	2.21	601,669	2.00
1 哲学	15,969	4.76	2,343,849,711	4.90	560	329,412	4.98	1,466,585	4.88
2 歴史	31,436	9.36	5,010,749,621	10.47	1,133	666,471	10.08	3,056,778	10.17
3 社会科学	54,450	16.22	8,946,058,392	18.69	2,195	1,291,176	19.52	5,716,463	19.02
4 自然科学	22,674	6.75	3,028,276,363	6.33	663	390,000	5.90	1,682,878	5.60
5 工業	28,325	8.44	3,149,144,051	6.58	690	405,882	6.14	1,616,570	5.38
6 産業	12,781	3.81	1,690,150,481	3.53	380	223,529	3.38	955,392	3.18
7 芸術	29,104	8.67	4,057,291,256	8.47	897	527,647	7.98	2,167,036	7.21
8 言語	5,863	1.75	956,625,910	2.00	217	127,647	1.93	427,326	1.42
9 文学	103,279	30.76	15,485,091,056	32.34	3,765	2,214,706	33.49	11,212,003	37.31
n 記録なし	24,402	7.27	2,206,890,351	4.61	493	290,000	4.39	1,150,711	3.83
合計	335,721	100.00	47,877,656,072	100.00	11,242	6,612,941	100.00	30,053,411	100.00

４.３　特定目的 SC

４.３.１　白書

1976 年から 2005 年までの 30 年間に発行された政府系刊行物「白書」1,006 冊からランダムにサンプリングにより，1,500 サンプルを抽出したものである。

層別情報は，9 ジャンル（安全，外交，科学技術，環境，教育，経済，国土交通，農林水産，福祉），6 期（5 年ごと）である。

４.３.２　教科書

小学校・中学校・高等学校の各学習指導要領（平成 10～11 年文部省告示，平成 15 年一部改正）に基づき，2005 年度から 2007 年度に実際に使用された検定教科書を対象とした。各校種・各学年・各教科から 1 種ずつ，発行部数の多い教科書を選出した。母集団は 144 冊である。層別情報は，10 教科（国語，数学，理科，社会，外国語，技術家庭，芸術，保健体育，情報，生活）と 3 校種（小学校，中学校，高等学校）である。

[7] 出版 SC「書籍」の場合と同様，書き言葉コーパスの設計趣旨に照らして適切でない書籍を除外した数。

４.３.３　広報紙

人口構成比などを考慮し，全国から 100 の自治体（区市町村）をサンプリングし，そこで 2008 年度に発行された広報紙をデータとした。層別情報は 8 地域（北海道，東北，関東，中部，近畿，中国，四国，九州・沖縄）である。

４.３.４　ベストセラー

『出版年鑑』（出版ニュース社）および『出版指標年報』（全国出版協会出版科学研究所）のどちらかに，各年のベストセラーとして上位 20 位までに挙げられた書籍 951 冊を対象とした。層別は行っていない。

４.３.５　Yahoo!知恵袋

ヤフー株式会社より提供された「Yahoo!知恵袋」のデータ（2004 年 10 月から 2005 年 1 月にかけて投稿された 3,120,839 の質問とそれに対する複数の回答）から 91,450 サンプルを抽出した。1 つのサンプルは 1 つの質問とそれに対するベストアンサーから成る。層別情報は，15 個の大カテゴリ，82 個の中カテゴリ，279 個の小カテゴリである。

４.３.６　Yahoo!ブログ

ヤフー株式会社より提供された「Yahoo!ブログ」のデータ（2008 年 4 月 26 日から 2009 年 4 月 25 日までに投稿された 3,463,413 記事）から 1.8%を抽出した 52,680 記事を対象とする。層別情報は，15 個の大カテゴリ，54 個の中カテゴリ，316 個の小カテゴリである。これらは Yahoo!知恵袋のカテゴリーと別のものである。

４.３.７　韻文

短歌，俳句，詩を対象とする。サンプル取得のためのデータは以下のとおり。

・短歌：『現代短歌全集』（筑摩書房，2002 年刊）第 14 巻～第 17 巻
・俳句：『増補現代俳句大系』（角川書店，1980 年～1982 年刊）第 8 巻～第 15 巻
・詩：「現代詩文庫」シリーズ（思潮社，1986 年～2005 年刊）118 冊

著作権処理の結果，利用できることになった歌集 60，句集 92，詩集 101 からそれぞれ 5 万語ずつ，計 253 サンプルを取得した。

４.３.８　法律

1976 年から 2005 年までの 30 年間に公布され，2009 年時点でも施行されている 718 の法律から 348 サンプルを取得した。層別情報は，43 ジャンル，6 期（5 年ごと）である。

４.３.９　国会会議録

1976 年から 2005 年までの 30 年間における「国会会議録」を対象とした。利用したデータは，国立国会図書館より提供された第 77 国会から第 163 回国会までに開かれた 32,986 会議の会議録データである。ここから約 500 万語分に相当する 159 会議を抽出した。

５．電子化

５.１　文字入力

BCCWJ に収録したデータの原資料の多くは紙媒体である。そのため文字入力の過程が必要になる。文字入力は以下の仕様に基づいて行っている。

(1)JIS X 0213 : 2004 規格に基づき字形を詳細に区別し入力し分ける。例えば，原文に「補塡」とあった場合は「補填」ではなく，原文のまま入力する。この規格を採用した理由は，文字集合としての大きさだけでなく，印刷字体を考慮した包摂基準を持つこと，および，『太陽コーパス』などの他のコーパスとの連携を考慮したためである。この規格の運用を実際のデータで検証した高田(2009)によると，この文字集合に収まらなかった外字は延べで 704 字（0.001%），異なりで 263 字（3.77%）と報告されている。

(2)記号・改行の意味による統制を行う。例えば，「－（マイナス）」と「ー（長音符号）」のように，異なる記号が同じ意味として用いられる場合がある．このような場合，原文のまま入力すると検索が円滑にいかなくなる．そのため，原文における見え方ではなく，その意味によって入力し分ける。

(3)組文字を切り離す。㈱，㌢のようないわゆる組文字は「(株)」，「センチ」のようにすべて 1 字ずつ切り離して入力する。

５.２　XML 化

BCCWJ では，『日本語話し言葉コーパス』および『太陽コーパス』で培われたタグ付け

の経験を生かし，文書構造が的確に再現されるよう XML で記述するタグセットを用意している。研究用の付加情報は，以下の三つに類別される。タグの詳細は，山口ほか(2011)を参照されたい。

(1)文書構造情報

記事，見出し，段落，引用，文などの枠組みを用意し，テキストを構造化して表現する。これらの枠組みは，以下の目的を満たすものとして策定したものである。

・ある一定のまとまりをもつ文書（記事を想定）が有する階層性を表現できること
・媒体やジャンルによる文書構造の差異を表現できること
・文体，語彙，文法に差異の見られるような要素を区別できること

(2)文字情報

文字の読みに関するルビ，誤植などの校正注，文字集合に含まない文字や記号（外字）などの情報を付与する。

(3)サンプリング情報

サンプリング時に決定するサンプル抽出基準点（乱数による縦横交叉点から決まる 1 文字）の情報を付与する。

表 6 に主なタグとその意味を掲げる。

表 6　主なタグの一覧

種類	タグ名	内容
サンプル	sample	サンプリングによって 1 サンプルとされた文書要素
階層構造	article	同一著者による，同一テーマのひとまとまりの文書要素
	cluster	title 要素が包括する文書要素全体
	title	特定範囲の文書要素の内容を代表する記述
	list	箇条書きなど，列挙された文書要素の集まり
	paragraph	段落を表す文書要素
	sentence	文に相当する文書要素
図表	figure	付随する文書要素のある図・表・写真・絵など
	caption	図表についてのタイトルや説明
引用	citation	当該 article 要素の本文において言及される，他文献からの引用要素
	speech	発話の引用・書き起こし，心内発話の描写
	quote	当該 article 要素とは異なる著作物からの引用や，発話・心内発話の引用・描写・書き起こし
注記	noteBody	脚注，注記など本文と区別して記述される注記
その他	abstract	article 要素，または cluster 要素の概要に相当する文書要素
	authorsData	著作者表示・署名にあたる要素
	verse	詩，和歌，俳句，歌謡などの韻文
文字・表記	ruby	ルビ付き文字
	correction	原文の誤植を訂正した文字
	missingCharacter	JIS X 0213:2004 で規定されている文字以外の文字（JIS 外字）

６．形態論情報付与

６.１　言語単位の設計方針

BCCWJ で利用する言語単位は，基本的に語彙調査における調査単位の場合と同様，大量の言語データをコンピュータで処理することを念頭に，揺れの少ない規則の集合として設計した。具体的な設計の基本方針は以下の 3 点である。

(1)コーパスに基づく用例収集，媒体・ジャンルの言語的特徴の解明に適した単位を設計する。

(2)『日本語話し言葉コーパス』と互換性のある形態論情報を設計する。

(3)国立国語研究所の語彙調査における知見を活用する。

これらの基本方針に基づき整備している「形態論情報データベース」は，電子化辞書班

と共同で進めている形態素解析用電子化辞書UniDicに活かされている。形態論情報データベースの設計と実装については小木曽・中村(2011)に詳しい。

６.２　短単位と長単位

短単位は用例収集を目的として，長単位は言語的特徴の解明を目的として採用したものである。これらはいずれも『日本語話し言葉コーパス』で採用されたものである。

短単位は，形態的な面に着目したもので，形態素に相当する最小単位の1回結合までの複合語を認めるものである。長単位は文節をもとにした単位で，文節を構成する自立語と付属語とがそれぞれ長単位となる。後述のコアデータにおける短単位と長単位の数を表7に示す。短単位数を長単位数で割った，1長単位あたりの平均短単位数がYahoo!知恵袋，書籍が低く，新聞，白書が高い値になっており，データを特徴付ける値として利用できることを示唆している。

表7　コアデータの短単位数と長単位数

媒体	短単位	長単位	短単位/長単位
書籍	234,431	199,393	1.18
雑誌	245,543	200,211	1.23
新聞	360,825	273,441	1.32
白書	228,272	159,019	1.44
Yahoo!知恵袋	110,696	95,094	1.16
Yahoo!ブログ	118,305	99,985	1.18

BCCWJで利用した言語単位全般については小椋ほか(2011)に詳しい。

６.３　コアデータの作成

コアデータは，形態素解析システム等の学習用データとするため，自動形態素解析結果に人手修正を加え，99%以上の精度としたデータで，語数は短単位で約130万語，長単位で約100万語である。

７．報告書

サンプリング関係 7 冊，電子化関係 4 冊，形態論情報付与関係 8 冊（うち 2 冊は電子化辞書班と連携）計 19 冊の報告書を作成した。これらは主として書き言葉コーパス構築のノウハウを蓄積したものである。

８．BCCWJ の公開に向けて

2011 年度のデータ公開に向けて，データの最終的整備，公開形式の決定などの課題が残っている。公開に当たっては，コーパスの初心者から上級者まで幅広く利用してもらえるように配慮し，コーパス言語学の普及の一助としたい。

文　献

木村睦子(1982)「語彙の計量」『講座日本語の語彙 1 語彙原論』，pp.225-243，明治書院.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕(2011)『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）（下）』

小木曽智信・中村壮範(2011),『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版』

高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也(2009)『JIS X 0213:2004 運用の検証』

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(2011a)『『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』

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(2011b)『『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装』

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)『『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

研究活動・成果の総括：ツール班
書き言葉コーパスの自動アノテーションの研究

松本裕治　（班長：奈良先端科学技術大学院大学情報科学研究科）†
徳永健伸　（分担者：東京工業大学大学院理工学研究科）
乾健太郎　（分担者：東北大学大学院情報科学研究科）
橋田浩一　（分担者：産業技術総合研究所社会知能技術研究ラボ）
橋本泰一　（分担者：東京工業大学総合プロジェクト支援センター）
浅原正幸　（分担者：奈良先端科学技術大学院大学情報科学研究科）
小町守　（分担者：奈良先端科学技術大学院大学情報科学研究科）
飯田龍　（協力者：東京工業大学大学院理工学研究科）
大山浩美　（協力者：奈良先端科学技術大学院大学情報科学研究科）
岩立将和　（協力者：奈良先端科学技術大学院大学情報科学研究科）
森田敏生　（協力者：総和技研）
Dain Kaplan　（協力者：東京工業大学大学院理工学研究科）
狩野芳伸　（協力者：東京大学情報学環）

Final Progress Report：‘Tools and Annotation’ Group

Yuji Matsumoto　(Nara Institute of Science and Technology)
Takenobu Tokunaga　(Tokyo Institute of Technology)
Kentaro Inui　(Tohoku University)
Koiti Hasida　(Advanced Industrial Science and Technology)
Taiichi Hashimoto　(Tokyo Institute of Technology)
Masayuki Asahara　(Nara Institute of Science and Technology)
Mamoru Komachi　(Nara Institute of Science and Technology)
Ryu Iida　(Tokyo Institute of Technology)
Hiromi Oyama　(Nara Institute of Science and Technology)
Masakazu Iwatate　(Nara Institute of Science and Technology)
Toshio Morita　(Sowa Giken)
Dain Kaplan　(Tokyo Institute of Technology)
Yoshinobu Kano　(The University of Tokyo)

１．ツール班の課題と活動の概要

ツール班は，本特定領域研究で構築されるコーパスに対して，様々なアノテーションを施すための自動言語解析ツールとアノテーション支援およびコーパス利用ツールの構築を目標としてきた．日本語コーパスへのアノテーションとしては，形態素情報が最も基本的な情報であるが，その基本単位となる辞書 UniDic の開発は電子化辞書班が，そして，形態素情報のアノテーションはデータ班が担当することになっていた（プロジェクトの途中から，長単位に基づく文節情報のアノテーションもデータ班の担当となった）．ツール班では，そのレベルより上のアノテーションを担当し，そのための様々なツールの構築，および，コーパスへの具体的なアノテーション作業を実施した．

構築したツールの主なものは，自動言語解析ツールとしては，日本語係り受け解析，固有表現解析，述語項構造解析，照応・共参照解析，モダリティ解析ツールがあり，これらの解析ツールを機械学習を用いて構築するため，および，性能評価のため，それぞれに対応するタグ（アノテーション）付きコーパスを構築した．

† matsu@is.naist.jp

コーパスアノテーションの支援ツールとして，形態素，文節，係り受け解析に特化したコーパス管理ツール「茶器」，および，汎用のコーパスアノテーションツール「Slate」を構築した．また，様々なタグ付きコーパスやコーパス構築ツールを相互運用するためのツールを構築した．

次節で，ツール班のメンバーが担当したツールやデータの概要を示す．班長および分担者は当初メンバーとほとんど変わっていないが，協力者としては，プロジェクト期間を通じて様々な学生や研究者の協力を得た．協力者の詳細は過去の活動報告を参照していただくことにして，次節では最終年度の関係者のみを示す．なお，以下の各グループの活動の詳細は，本ワークショップで，口頭発表，あるいは，ポスター発表で報告する予定である．本報告書内の対応する稿を合わせて参考にされたい．

２．ツール班の開発ツール・データの概要

本プロジェクトで構築したツールやデータについて，各グループの活動の概要を示す．

松本裕治，浅原正幸，岩立将和（以上奈良先端大），森田敏生（総和技研）：係り受け解析ツールの構築，タグ付きコーパスの構築支援やタグ付きコーパスの検索等の利用を支援するためのコーパス管理ツール「茶器」を開発した．また，未解析コーパスの文境界を判定し，形態素解析器 MeCab や係り受け解析器 CaboCha を呼び出すことにより，形態素や係り受け情報のタグ付けを行う GUI を開発した．茶器については，本報告書の松本他の稿を参照のこと．

また，BCCWJ のコアデータに対して，係り受け情報と並列構造のアノテーションを行った．この作業の進捗については，本報告書内の浅原他の稿を参照のこと．

乾健太郎（東北大），松吉俊，佐尾ちとせ（奈良先端大）：モダリティ体系の設計と BCCWJ へのアノテーションを行った．言語表現における話者の態度や価値判断に関する情報付与の体系を拡張モダリティとして新たに設計した．また，その基準に基づくアノテーションを BCCWJ のコアデータに対して行った．詳細は，本報告書の松吉他の稿を参照のこと

徳永健伸，Dain Kaplan，飯田龍（東京工業大）：言語コーパスに対するアノテーションのほとんどは，ある範囲（セグメント）に対するラベルの付与，および，セグメント間の関係（リンク）の付与に帰着できる．本グループでは，セグメントとリンクに基づく汎用のコーパスアノテーションツール「Slate」を Web アプリケーションとして構築した．利用者は，セグメントやリンクの名称や性質の定義を行うことができ，利用者による柔軟なカスタマイズが可能である．詳細は，本報告書の徳永他の稿を参照のこと．

飯田龍（東京工業大），小町守（奈良先端大）：BCCWJ コアデータに対する述語項構造および照応関係のアノテーションを行った．文および文章の意味解析において，文中の述語とその項（意味上の主語，目的語など）と，照応表現（省略語を含む）とその参照先を特定することは，重要な基盤技術である．そのための基礎データとして，述語項構造，事態性名詞の項構造，および，照応関係のアノテーション規準の設定と，それに基づくタグ付きコーパスの構築を行った．詳細は，本報告書の小町他の稿を参照のこと．

橋本泰一（東京工業大）：拡張固有表現タグ付きコーパスの構築と拡張固有表現抽出ツールの開発を行った．新聞記事等の意味解析のため，人名，地名，組織名や日付，数値表現などの固有表現の認識が重要な処理とみなされている．さらに，より一般的な文章の理解のため，約 200 種類の固有表現が定義されている．BCCWJ コアデータに対して，拡張固有表現のタグ付けを行った．また，そのデータを用いて，機械学習に基づき，拡張固有表現抽出ツールを実装し，利用のための GUI を構築した．詳細については，本報告書の橋本の稿を参照のこと．

橋田浩一（産総研），狩野芳伸（東大）：ここまでに述べたように，ツール班では，コーパスに対する様々なタグ付けを行うとともに，タグ付け作業の支援やタグ付けされたコーパスを利用するためのツールを構築してきた．また，他班で構築されているコーパス利用ツールも存在する．様々なタグ付きコーパスや関連ツールがバラバラに存在することが多く，それらの間の互換性や相互運用を促進することは重要な課題である．本グループでは，タグ付きコーパスとコーパス関連ツールを相互運用するためのツールの構築を行った．詳細については，本報告書の狩野他の稿を参照のこと．

３．ツール班の活動について

ツール班では，毎年３～５回程度，全員が集まる班会議を開催し，各グループの進捗報告と連携について議論してきた．また，関連するグループ間では個別の会合を適宜開催してきた． その他，2007 年～2009 年の３年間，毎年９月に京都大学で開催された科学技術振興調整費新興分野人材養成プログラム「自然言語処理技術」セミナーにてコーパス解析ツールとコーパス管理ツールの講習を行った．また，主としてプロジェクトメンバーに対して，毎年秋にツール講習会を開催し，メンバーからのフィードバックを得た．

４．おわりに

ツール班の目標と具体的な活動内容の概要を報告した．タグ付きコーパス構築と利用のためのツールの開発は当初目標をほぼ達成したが，開発の過程において，さらに追加すべき機能等が明らかになった．詳細については，それぞれの報告書を参照されたい．BCCWJ のコアデータに対するタグ付けは，目標としていたすべてのアノテーションをコアデータ全体に施すことはできなかったが，引き続き作業を続ける予定である．最後に，本特定領域研究に関連して行った主な外部発表と受賞一覧を最近のものから順に以下に添付する．

主な研究発表一覧（査読なしの国内発表等を除く）

[学術雑誌論文]

飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 「述語項構造と照応関係のアノテーション：NAIST テキストコーパス構築の経験から」 自然言語処理, Vol.17, No.2, pp.25-50, April 2010.

Ai Azuma and Yuji Matsumoto, “A generalization of forward-backward algorithm,” *Transactions of the Japanese Society for Artificial Intelligence*, Vol.25, No.3, pp.494-503, April 2010.

渡邉陽太郎, 浅原正幸, 松本裕治. 「述語語義と意味役割の結合学習のための構造予測モデル」 人工知能学会論文誌, Vol.25, No.2, pp.252-261, January 2010.

小町守, 飯田龍, 乾健太郎, 松本裕治. 「名詞句の語彙統語パターンを用いた事態性名詞の項構造解析」. 自然言語処理, Vol.17, No.1, pp.141-159, January 2010.

吉川克正, Sebastian Riedel, 浅原正幸, 松本裕治. 「Markov Logic を利用した時間的順序関係の同時推論」. 人工知能学会論文誌, Vol.24, No.6, pp.521-530, November 2009.

Vera Sheinman, Takenobu Tokunaga. "AdjScale: Visualizing differences between adjectives for language learners," *IEICE Transaction of Information and Systems*, Vol. E92-D, No.8, pp.1542-1550, August, 2009.

岩立将和, 浅原正幸, 松本裕治. 「トーナメントモデルを用いた日本語係り受け解析」 自然言語処理, Vol.15, No.5, pp.169-185, November 2008.

Takenobu Tokunaga, Chu-Ren Huang, Yat Mei Lee. "Asian language resources: the state-of-the-art," *Language Resources and Evaluation*, Vol.42, No.2, pp.109-116, May 2008.

渡邉陽太郎, 浅原正幸, 松本裕治. 「グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類」 人工知能学会論文誌, Vol.23, No.4, pp.245-254, April 2008.

橋本泰一, 吉田恭祐, 野口正樹, 徳永健伸, 田中穂積. 「関係データベースを用いた構文木付きコーパス検索手法」 自然言語処理, Vol.14, No.4, pp.3-22, 2007.

Ryu Iida, Kentaro Inui, Yuji Matsumoto. "Zero-anaphora resolution by learning rich syntactic pattern features," *ACM Transactions on Asian Language Information Processing (TALIP),* Vol 6, Issue 4, Article 12, 2007.

Chu-Ren Huang, Takenobu Tokunaga, Sohpia Yat Mei Lee. "Asian language processing: current state-of-the-art," *Language Resources and Evaluation,* Vol.40, No.3-4, pp.203-218, March, 2006.

[国際会議発表]

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, Yuji Matsumoto. "Coreference based event-argument relation extraction on biomedical text," In *Proceedings of the Fourth International Symposium on Semantic Mining in Biomedicine (SMBM 2010)*, October 2010.

Hiromi Oyama, Yuji Matsumoto. "Automatic error detection method for Japanese case particles in Japanese language learners' writing," In *Corpus, ICT, and Language Education*, pp.235-245, Septempber 2010.

Tokunaga Takenobu, Yasuhara Masaaki, Terai Asuka, David Morris, Anja Belz. "Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving", *Proceedings of the Eighth Workshop on Asian Language Resources*, pp.38-46, August 2010.

Yotaro Watanabe, Masayuki Asahara, Yuji Matsumoto, "A structured model for joint learning of argument roles and predicate senses," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp.98-102, July 2010.

Dain Kaplan, Ryu Iida, Takenobu Tokunaga, "Annotation Process Management Revisited," *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. pp.3654-3661, May 2010.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee and Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, David Traum, "Towards an ISO Standard for Dialogue Act Annotation," *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, May 2010.

Tokunaga Takenobu. "Aspects of Language Resource Management: Creation and Utilisation," *Proceedings of the 2nd FLaReNet Forum*, February 2010.

Ai Azuma, Yuji Matsumoto. "A Generalization of Forward-backward Algorithm," In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pp.99-114, Lecture Notes in Computer Science 5781, September 2009.

Tokunaga Takenobu, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Xia Yingju, Chu-Ren Huang, Shu-Kai Hsieh, Shirai Kiyoaki. "Query Expansion using LMF-Compliant Lexical Resources," *Proceedings of the 7th Workshop on Asian Language Resources*, ACL-IJCNLP 2009, pp.145-152, August 2009.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, Yuji Matsumoto. “Jointly Identifying Temporal Relations with Markov Logic,” In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*, pp.405-413, August 2009.

Ryu Iida, Kentaro Inui, Yuji Matsumoto. “Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution,” In P*roceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*, pp.647-655, August 2009.

Yotaro Watanabe, Masayuki Asahara, Yuji Matsumoto. “Multilingual Syntactic-Semantic Dependency Parsing with Three-Stage Approximate Max-Margin Linear Models,” In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, pp.114-119, June 2009.

Vera Sheinman, Tokunaga Takenobu. "AdjScale: Differentiating between similar adjectives for language learners", *Proceedings of 1st International Conference on Computer Supported Education (CSEDU 2009)*, INSTICC Press, pp.229-235, March 2009.

Yotaro Watanabe, Masakazu Iwatate, Masayuki Asahara, Yuji Matsumoto. “A Pipeline Approach for Syntactic and Semantic Dependency Parsing,” In *Proceedings of the 12th Conference on Natural Language Learning (CoNLL-2008)*, pp.228-232, August 2008.

Masakazu Iwatate, Masayuki Asahara and Yuji Matsumoto. “Japanese dependency parsing using a tournament model,” In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pp.361-368, August 2008.

Tokunaga Takenobu, Dain Kaplan, Chu-Ren Huang, Shu-Kai Hsieh, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Kiyoaki Shirai, others. “Adapting International Standard for Asian Language Technologies,” *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, May 2008.

Masaki Noguchi, Kenta Miyoshi, Takenobu Tokunaga, Ryu Iida, Mamoru Komachi, Kentaro Inui. “Multiple Purpose Annotation using SLAT - Segment and Link-based Annotation Tool -,” *Proceedings of 2nd Linguistic Annotation Workshop*, pp.61-64, May 2008.

Neil Rubens, Vera Sheinman, Takenobu Tokunaga, Masashi Sugiyama. "Order retrieval", *Large-Scale Knowledge Resources*, Springer-Verlag. Vol.4938, LNAI, pp.310-317, March 2008.

Hiromi Oyama, Yuji Matsumoto, Masayuki Asahara, Kosuke Sakata. “Construction of an error information tagged corpus of Japanese language learners and automatic error detection,” In *Proceedings of the Computer Assisted Language Instruction Consortium*, March 2008.

Kiyoaki Shirai, Takenobu Tokunaga, Chu-Ren Huang, Shu-Kai Hsieh, Tzu-Yi Kuo, Virach Sornlertlamvanich, Thatsanee Charoenporn. “Constructing Taxonomy of Numerative Classifiers for Asian Languages,” *Proceedings of the Third International Joint Conference on Natural Language Processing*, pp.397-402, January 2008.

Ryu Iida, Mamoru Komachi, Kentaro Inui, Yuji Matsumoto. “Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations,” ACL Workshop `Linguistic Annotation Workshop', pp.132-139, 2007.

Yotaro Watanabe, Masayuki Asahara, Yuji Matsumoto. “A Graph-Based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields”, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.649-657, 2007.

Mamoru Komachi, Ryu Iida, Kentaro Inui and Yuji Matsumoto. “Learning Based Argument Structure Analysis of Event-nouns in Japanese,” *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp.120-128, 2007.

Koiti Hasida, Noriaki Izumi, Akira Mori. “CBTO: Compositional Business-Task Organization,” W3C Workshop on

Declarative Models of Distributed Web Applications, 2007.

Vera Sheinman, Neil Rubens, Tokunaga Takenobu. “Commonly Perceived Order within a Category,” Proceedings of OntoLex 2007: from text to knowledge, ISWC07 Workshop, 2007.

Vera Sheinman, Tokunaga Takenobu. “WordSets: Finding Lexically Similar Words for Second Language Acquisition,” Recent Advances in Natural Language Processing, 2007.

Ryu Iida, Kentaro Inui, Yuji Matsumoto. “Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution,” *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* pp.625-632, July 2006.

Ichikawa Hiroshi, Hakoda Keita, Hashimoto Taiichi, Tokunaga Takenobu. “Efficient sentence retrieval based on syntactic structure,” *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp.407-411, July 2006.

Tokunaga Takenobu, Virach Sornlertlamvanich, Thatsanee Charoenporn, Nicoletta Calzolari, Monica Monachini, Claudia Sonia, Chu-Ren Huang, Xia YingJu, Xia YingJu, Yu Hao, Laurent Prevot, Shirai Kiyoaki. “Infrastructure for standardization of Asian language resources,” *Proceedings of the COLING/ACL 2006* Main Conference Poster Sessions, pp.827-834, July 2006.

[学会誌解説その他の出版物]

松本裕治. 「コーパスへの自動アノテーションツールとアノテーション支援環境の構築」. 人工知能学会誌, Vol.24, No.5, pp.632-639, September 2009.

松本裕治. 「統語情報の付与」 国文学と鑑賞, Vol.74, No.1, pp.44-52, January 2009.

松本裕治, 大山浩美.「言語処理による作文支援・語彙学習への可能性について」 日本語教育「特集 作文教育のための語彙研究」, Vol.140, pp.37-47, 日本語教育学会, January 2009.

Yuji Matsumoto. “Corpus Annotation/Management Tools for the Project: Balanced Corpus of Contemporary Written Japanese,” *Large-Scale Knowledge Resources, 3rd International Conference on Large-Scale Knowledge Resources, Proceedings, LNAI4938*, pp.106-115, March 2008.

橋田浩一, 和泉憲明.「オントロジーに基づく知識の構造化と活用」 情報処理, Vol.48, No.8, pp.843-848, 2007.

浅原正幸.「自然言語処理と系列ラベリング技術」 オペレーションズ・リサーチ, Vol.52, No,11, pp.688-694, 2007.

松本裕治.「統計的統語解析」 オペレーションズ・リサーチ, Vol.52, No,11, pp.695-699, 2007.

徳永健伸.「言語処理を利用した知的情報アクセスー検索，抽出，要約，分類，QAー」 オペレーションズ・リサーチ, Vol.52, No.11, pp.713-718, 2007.

[受賞]

The Best Paper Award of the SMBM2010 (the Fourth International Symposium on Semantic Mining in Biomedicine), Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, Yuji Matsumoto, Coreference based event-argument relation extraction on Biomedical Text, 2010.

人工知能学会 2007 年度業績賞，松本裕治, 2008.

言語処理学会第 14 回年次大会優秀発表賞, 岩立将和，浅原正幸，松本裕治, “トーナメントモデルを用いた日本語係り受け解析,” 2008.

平成19年度山下記念研究賞, 飯田龍, “NAISTテキストコーパス: 述語項構造と共参照関係のアノテーション,” 2007.

言語処理学会第 13 回年次大会優秀発表賞，飯田龍，他, “日本語書き言葉を対象とした述語項構造と共参照関係のアノテーション:NAIST テキストコーパス開発の経験から,” 2007.

2007 年度日本 OSS 貢献者賞: 松本裕治, “日本語形態素解析システム「茶筌 (ChaSen)」の開発をはじめとした OSS への貢献,” 独立行政法人 情報処理推進機構(IPA), 2007.

Best Asian NLP Paper Award, COLING/ACL 2006, Ryu Iida, Kentaro Inui, Yuji Matsumoto, “Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution” 2006.

研究活動・成果の総括：電子化辞書班
多様な目的に適した形態素解析システム用電子化辞書の開発

伝 康晴　（班長：千葉大学文学部）†
峯松 信明　（分担者：東京大学大学院情報理工学系研究科）
小木曽 智信（分担者：国立国語研究所言語資源研究系）
小磯 花絵　（分担者：国立国語研究所理論・構造研究系）
山田 篤　（連携研究者：京都高度技術研究所研究部）
内元 清貴　（連携研究者：情報通信研究機構総合企画部）

Final Progress Report: 'UniDic' Group

Yasuharu Den (Faculty of Letters, Chiba University)
Nobuaki Minematsu (Graduate School of Information Science and Technology, The University of Tokyo)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)
Atsushi Yamada (Research Division, ASTEM RI)
Kiyotaka Uchimoto (Strategic Planning Department, NICT)

1. 電子化辞書班の目的と総括

本計画研究の目的は、形態素解析システム用電子化辞書 UniDic を整備・拡充することにより、（1）本研究領域が目指す大規模書き言葉コーパスの構築を支援するとともに、（2）日本語学・日本語教育学における語彙・文法調査研究、自然言語処理における構文・意味解析研究、音声情報処理におけるテキスト音声合成研究など、多様な目的に適した統合的な電子化辞書およびその利用システムを提供することにある。

2. 研究成果の総括

形態素解析辞書に関しては、以下の特徴を持つ辞書を設計・開発した(伝ほか, 2007)。

- 「短単位」という揺れがない斉一な単位で設計
- 語彙素・語形・書字形・発音形の階層構造を持ち、表記の揺れや語形の変異にかかわらず同一の見出しを与えることが可能
- アクセントや音変化の情報を付与でき、テキスト音声合成などに利用可能

辞書データベースを国立国語研究所内で構築しながら、形態素解析システム ChaSen・MeCab 用辞書を随時公開し、最終的に語彙素約 21 万語・書字形約 33 万語の規模と、品詞認定約 98.9%・語彙素認定約 98.6%（MeCab 版）の解析精度を達成できた。最終年度には、辞書データベースを XML ファイル群として記述し、ユーザがカスタマイズ可能な辞書作成環境を提供する新しい方式で UniDic2 を設計・開発した（小木曽・伝, 2011）。また、中・長単位解析システムを含む、形態素解析の後処理ツール群を作成し、多様な目的に供することができた。

以下、これらの成果について順を追って簡単に紹介する。

† den@cogsci.L.chiba-u.ac.jp

3. 形態素解析辞書

3.1. 規模

形態素解析辞書の規模は、研究開始当初、語彙素（辞書見出し）約 10 万語・書字形（表記）約 14 万語にすぎなかったが、BCCWJ のサンプルを中心とするコーパスからの見出し語追加を続けた結果、最新版では語彙素約 21 万語・書字形約 33 万語と倍以上の規模に達している（図１）。

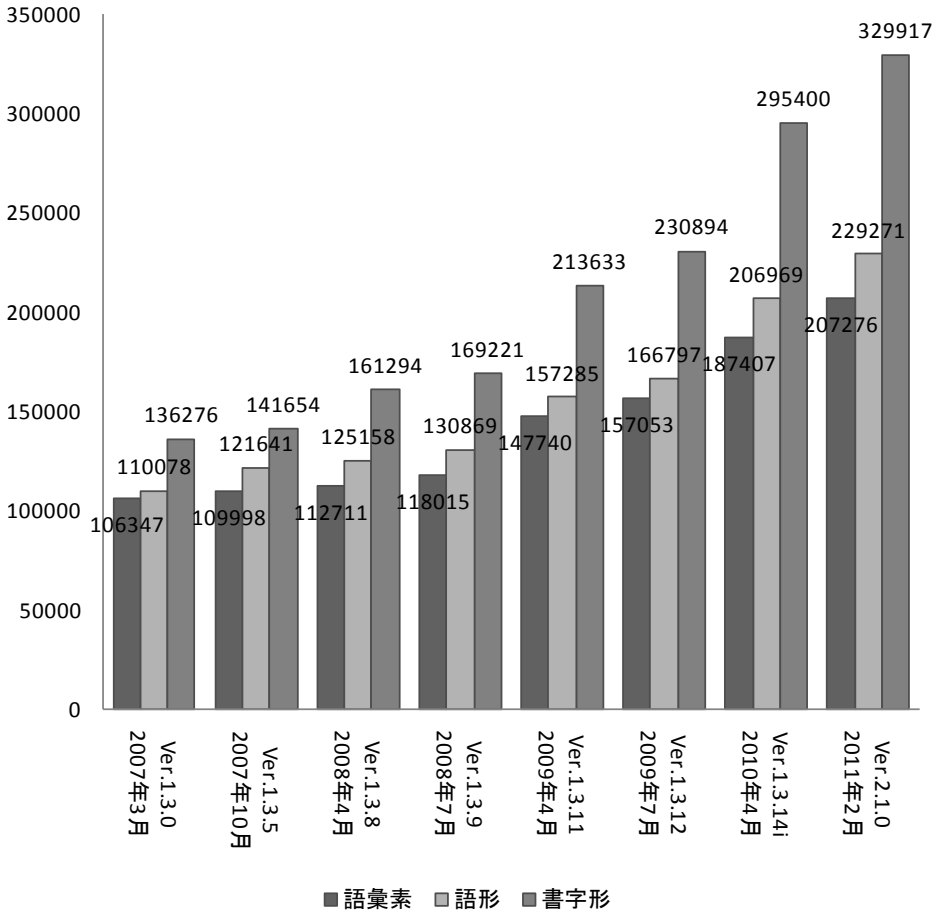
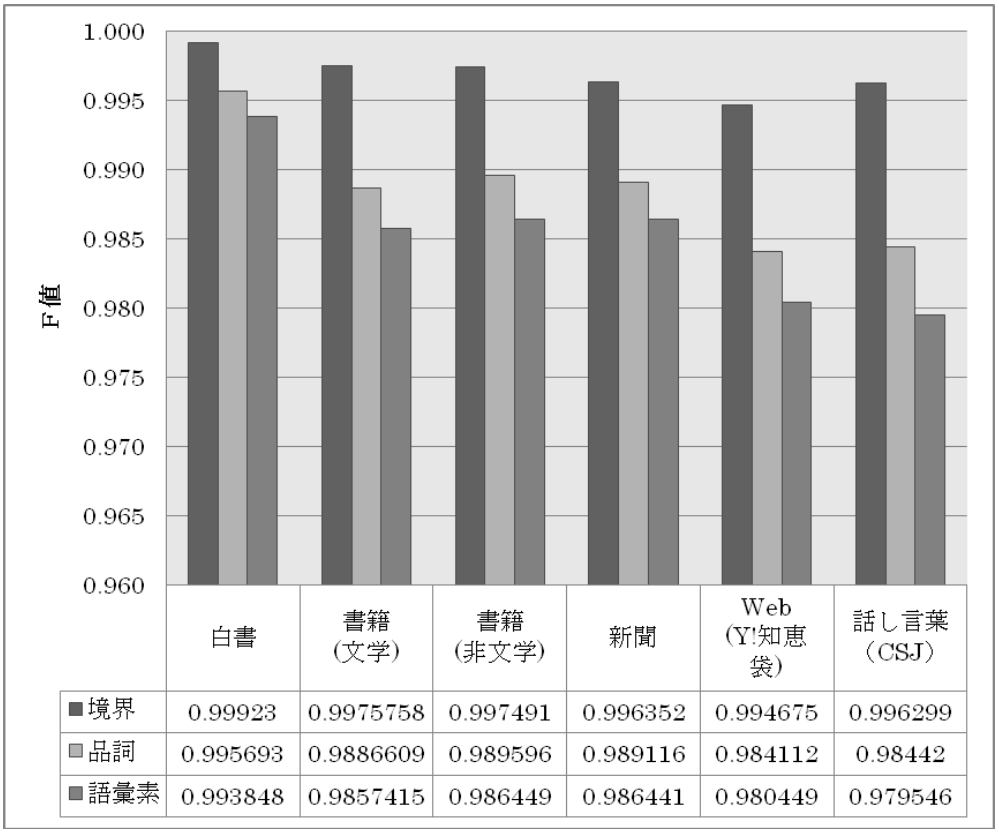


図 1：形態素解析辞書の見出し語数の推移

3.2. 解析精度

既存の形態素解析辞書では、学習データの偏りのために新聞記事以外のテキストで解析精度が大きく低下していたのに対し、BCCWJ の多様なジャンルのテキストを学習に利用した UniDic では、新聞記事のみならず文学作品や話し言葉、Web など多くの種類のテキストにおいて高い精度での解析を実現している（図 2）。また、語種を統計学習に利用したこと(Den et al., 2008)や短単位規定の改善を繰り返しながら見出し語の整備に努めてきたことも高精度化に貢献している。



	白書	書籍(文学)	書籍(非文学)	新聞	Web(Y!知恵袋)	話し言葉(CSJ)
■境界	0.99923	0.9975758	0.997491	0.996352	0.994675	0.996299
□品詞	0.995693	0.9886609	0.989596	0.989116	0.984112	0.98442
■語彙素	0.993848	0.9857415	0.986449	0.986441	0.980449	0.979546

図 2：MeCab 版 UniDic の解析精度（F 値）（ver. 1.3.12）

3.3. ユーザ

- 登録ユーザ数 ：約 3300 名
- 商用利用ライセンシー：10 社
  - 海外：Apple Computer Inc., Acapela Group, Telecom Italia
  - 国内：アイウィーヴ, ATR-Trek, イプロス, クレオ, 日経リサーチ, ネイバージャパン, ボイスシグナル

3.4. 形態素解析 GUI

UniDic を用いた形態素解析を人文系ユーザでも容易に実行できる環境を提供するため、GUI「茶まめ」を開発・公開した。

4. 辞書データベース

辞書の構築にあたっては、国立国語研究所内にデータベースサーバを立て、複数の作業者で利用してきた。この辞書データベースは語彙素・語形・書字形・発音形の階層構造を管理する利便性に優れたものであるが、一般公開している形態素解析辞書は階層構造を持たず扱いにくいものであった。そこで、辞書データベースの階層構造を XML で表現し、また活用表などを別に記述することで、より可読性・拡張性に優れた形態論辞書 UniDic2 を設計・開発した。詳細については、本 WS の研究発表 (小木曽・伝, 2011) 参照。

5. 中・長単位構成システム

UniDic では、語彙形態論研究に適した短単位に加え、音声研究に適した中単位、構文・意味研究に適した長単位という複数粒度の「語」を利用している。中・長単位は短単位解析（形態素解析）結果から中・長単位構成システムにより自動構成する。

5.1 長単位構成システム

長単位構成システムは、Uchimoto らの方法(Uchimoto & Isahara, 2007)を改良したチャンキングモデルと後処理に基づく手法を用いて作成した。統計モデルとして、チャンキングモデルには CRF と MMA を、後処理には SVM を用いた。BCCWJ の白書・書籍・新聞・雑誌・Web（Yahoo! 知恵袋）コアデータを用いて学習・評価することにより得られた最新の長単位解析の精度（F 値）は境界認定 98.9%・品詞認定 98.6%・語彙素認定 98.5%である。

5.2. 中単位構成システム

中単位構成システムは、最大全域木に基づく依存構造解析手法を用いて作成した(Uchimoto & Den, 2008)。BCCWJ の白書・書籍・新聞コアデータのうち、6,547 文に対して短単位間の係り受け情報と中単位情報を人手で付与し、システムを学習・評価した。その結果、98.3%の係り受け解析精度が得られた。さらに、短単位間の係り受け情報に基づく中単位境界同定規則を適用することにより、中単位境界を認定した。その解析精度（F 値）は 99.2%である。

5.3. 中・長単位解析ツール

これらの手法を実装し、中・長単位解析ツール Comainu を開発した（2011 年 3 月公開予定）。本ツールは長単位解析・中単位境界解析・文節解析の機能を持つ。

6. その他の成果

以上の成果以外にも、個別的なテーマに関して以下の研究成果をあげることができた。

6.1. ジャンル別辞書

各ジャンルに特化して高い精度での解析が行なえる形態素解析辞書を作成するために、ジャンル別辞書の作成を試みた。汎用の形態素解析辞書のコストと、少量の特定ジャンルのコーパスで学習したコストとを、学習によって求めた混合率で混合する（コストの線形和をとる）ことにより特定ジャンル向けの辞書を作成した。とくに、Web テキスト向けの辞書の精度向上で有効であった(小木曽・伝・渡部, 2009)。

6.2. ジャンルごとの文体比較

ジャンルによる形態論特徴を明らかにするため、短単位・長単位の形態論情報に着目したジャンルごとの文体比較研究を行なった。短単位を用いた調査(小磯ほか, 2009)では、書籍・新聞・白書・WEB データ（Yahoo!知恵袋）・国会会議録・話し言葉を対象に、品詞比率・語種比率・異なり語率などからジャンルを判別する統計モデルを構築し、80%の判別率を得た。さらに、先行研究との比較のため長単位を用いた調査(小磯ほか, 2010)を行ない、樺島

の指標 MVR（100×形容詞・形容動詞・副詞・連体詞／動詞率）がジャンルの特徴をとらえる上で不十分であることを示した。

6.3. 汎用後処理ツール

UniDic では、中・長単位構成以外にも、短単位解析結果の再解析・音変化処理・アクセント変化処理などのさまざまな後処理がある。これらの後処理を統一的に実現するための汎用後処理ツールを設計・開発した(山田・伝, 2010)。

6.4. 短単位解析結果の再解析

短単位解析結果の中に誤りの多い事例がある（「で」「に」における格助詞と助動詞「だ」連用形との混同)。これらの特定の事例に特化した統計モデルを構築し、短単位解析結果を再解析する手法を開発し、精度を向上できることを示した(中村・伝, 2008)。また、この手法を汎用後処理ツールを用いて実装した(アブドハリリ・伝, 2010)。

6.5. 音変化処理

複数の語が結合して複合語が構成されると、語頭や語末の音が変化することがある（「イチ（一)」+「ホン（本)」→「イッポン」)。まず、その典型である数詞と助数詞類の結合に伴う音変化を処理するツール ChaOne を開発した(山田, 2007)。そこでは、辞書中に記述された語頭・語末変化に関わる制約を用いて発音形を選択するという方式をとった。これに対して、処理対象を一般の連濁や促音化に広げ、汎用後処理ツール(山田・伝, 2010)と統計モデルを用いる方式を新たに検討した(山田, 2010)。新方式でも数詞・助数詞類に関しては ChaOne と同等の性能が得られているが、一般の連濁・促音化では検討課題が残されている。

6.6. アクセント変化処理

複数の語が結合して複合語が構成されると、構成語のアクセント型から変化することがある（「テ’ンキ（天気)」+「ヨホー（予報)」→「テンキヨ’ホー」)。このようなアクセント変化を処理するため、単独ラベラーによるアクセント句境界とアクセント核位置を付与したコーパス（約 7,200 文）を作成し、これを学習データとする統計モデルを各種検討した。CRF によるモデルを用い、アクセント句境界推定で 93.6%の精度（F 値）を得(Minematsu et al., 2007)、その結果を受けたアクセント核推定で 94.1%の精度を得た(印南ほか, 2008)。これらの手法は長い複合名詞に関する精度が劣るため、中単位構成に相当する係り受け処理を導入しアクセント句境界推定に利用したところ、その結果を受けたアクセント核推定で 4〜8 短単位からなる複合名詞に対する精度を 77.7%から 85.7%まで改善することができた(高野ほか, 2011)。

文献

アブドレイム・アブドハリリ・伝康晴 (2010). 汎用後処理ツールを用いた短単位解析結果の再解析　特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集, pp. 141-144.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007). コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用　日本語科学, 22, pp. 101-123.

Den, Y., Nakamura, J., Ogiso, T., & Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the 6th Conference on International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco, pp. 1019-1024.

印南圭祐・渡辺美知子・峯松信明・広瀬啓吉 (2008). CRF に基づくアクセント変形予測モデルにおけるエラー解析　言語処理学会第 14 回年次大会発表論文, pp. 969−972.

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

小磯花絵・小木曽智信・小椋秀樹・宮内佐夜香 (2010). 長単位情報に基づくジャンル間の文体に関する分析　特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集, pp. 183-190.

Minematsu, N., Kuroiwa, R., Hirose, K., & Watanabe, M. (2007). CRF-based statistical learning of Japanese accent sandhi for developing Japanese text-to-speech synthesis systems. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*, Bonn, Germany, pp. 148-153.

中村純平・伝康晴 (2008). 形態素解析誤りの多い助詞・助動詞の再解析　言語処理学会第 14 回年次大会発表論文集, pp.73-76.

小木曽智信・伝康晴 (2011). UniDic2：設計と実装　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.

小木曽智信・伝康晴・渡部涼子 (2009). ジャンル別 UniDic 作成の試み　特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集, pp.17-22.

高野克弥・清水信哉・峯松信明・広瀬啓吉 (2011). アクセント句境界推定におけるチャンカー出力の効果的利用に関する実験的検討　日本音響学会 2011 年春季研究発表会講演論文集.

Uchimoto, K. & Den, Y. (2008). Word-level dependency-structure annotation to Corpus of Spontaneous Japanese and its application, In *Proceedings of the 6th Conference on International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco, pp. 3118-3122.

Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*. Hyderabad, India, pp. 1731-1737.

山田篤 (2007). 数字列への読み付与―NumTrans と ChaOne―　特定領域研究「日本語コーパス」平成 19 年度全体会議予稿集, pp. 85-90.

山田篤・渡部涼子・小木曽智信 (2010). 汎用後処理ツールを用いた音変化処理の評価　特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集, pp. 145-150.

山田篤・伝康晴 (2010). UniDic 汎用後処理ツールの設計と実装　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. 23-28.

関連 URL

UniDic 公開ホームページ：http://download.unidic.org/

研究活動・成果の総括：日本語学班
コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発

田野村忠温（班長：大阪大学大学院文学研究科）
服部匡（分担者：同志社女子大学表象文化学部）
杉本武（分担者：筑波大学大学院人文社会科学研究科）
石井正彦（分担者：大阪大学大学院文学研究科）

Final Progress Report: 'Japanese Linguistics' Group

Tadaharu Tanomura (Osaka University)
Tadasu Hattori (Doshisha Women's College of Liberal Arts)
Takeshi Sugimoto (University of Tsukuba)
Masahiko Ishii (Osaka University)

１．日本語学班の研究目的

日本語学班は以下のことを目的として研究活動を行った。

(1) コーパスの利用による日本語研究の深化発展

a. コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発

b. 学界に対するコーパス利用の啓蒙・普及

(2) コーパス構築へのフィードバック

『現代日本語書き言葉均衡コーパス(BCCWJ)』構築へのフィードバック

２．日本語学班の主な研究活動

日本語学班として行った主な活動は以下の通りである。班員が個別に行った活動は除いている。

（2006年度）

・平成18年度全体会議に参加、国立国語研究所、2006年9月9～10日
・日本語学班研究会を開催、千里朝日阪急ビル、2006年9月24日
・日本語学班研究会を開催、千里朝日阪急ビル、2007年1月14日
・筑波大学対照言語研究プロジェクトシンポジウム「対話する言語学――日本語と諸外国語の対照的分析による発見と創出――」を共催、筑波大学、2007年2月17～18日
・平成18年度公開ワークショップに参加、時事通信ホール、2007年3月17～18日
・平成18年度研究成果報告書『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅰ』を刊行、2007年3月31日

（2007年度）

・日本語学班研究会を開催、学術総合センター、2007年4月23日
・日本語学班研究会を開催、千里朝日阪急ビル、2007年8月1日

・『コーパス日本語学ガイドブック』を刊行、2007年9月7日
・平成19年度全体会議に参加、北陸先端科学技術大学院大学、2007年9月7～8日
・日本語学班Webサイト(http://www.tokuteicorpus.jp/team/)を作成・公開、2007年9月22日
・辞書編集班拡大班会議（コロケーション研究会）を共催、ホテルスワ（つくば市)、2007年11月10～11日
・日本語学班研究会を開催、大阪大学東京オフィス、2007年12月2日
・日本語教育班拡大班会議（複合辞研究会）を共催、筑波大学、2007年12月16日
・日本語学班研究会を開催、千里朝日阪急ビル、2008年2月17日
・平成19年度公開ワークショップに参加、時事通信ホール・国立国語研究所、2008年3月15～16日

（2008年度）
・平成19年度研究成果報告書『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅱ』を刊行、2008年4月30日
・日本語学班研究会を開催、国立情報学研究所、2008年5月19日
・日本語学班研究会を開催、大阪大学、2008年8月4日
・「日本語用例検索」サイト(http://www.tokuteicorpus.jp/team/jpling/kwic/)を作成・公開、2008年8月18日
・平成20年度全体会議に参加、奈良県新公会堂、2008年9月6～7日
・辞書編集班拡大班会議を共催、ホテルスワ、2008年11月23～24日
・日本語学班研究会を開催、国立情報学研究所、2008年12月15日
・日本語教育班拡大班会議（複合辞研究会）を共催、筑波大学、2009年2月22日
・日本語学班研究会を開催、大阪大学、2009年3月1日
・平成20年度公開ワークショップに参加、東京工業大学、2009年3月14～16日
・平成20年度研究成果報告書『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅲ』を刊行、2009年3月31日

（2009年度）
・日本語学班研究会を開催、国立情報学研究所、2009年5月25日
・日本語学班研究会を開催、大阪大学、2009年8月28日
・平成21年度全体会議に参加、国立国語研究所、2009年9月5～6日
・辞書編集班拡大班会議を共催、ホテルスワ、2009年11月14～15日
・日本語学班研究会を開催、国立情報学研究所、2009年11月30日
・日本語教育班拡大班会議（複合辞研究会）を共催、筑波大学、2010年2月13日
・日本語学班研究会を開催、大阪大学、2010年2月22日
・平成21年度公開ワークショップに参加、東京工業大学、2010年3月14～16日
・平成21年度研究成果報告書『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅳ』を刊行、2010年3月31日

（2010年度）
・日本語学班研究会を開催、国立情報学研究所、2010年5月31日
・平成22年度全体会議に参加、国立国語研究所、2010年8月30～31日
・日本語学班研究会を開催、大阪大学、2010年9月6日
・「コーパス言語学ワークショップ」を開催、高麗大学校（韓国・ソウル市)、2010年10月10日
・平成18年度～平成22年度研究成果報告書『コーパス日本語学の新展開』を刊行、2010年10月20日
・日本語学会2010年度秋季大会にてワークショップ「コーパス日本語学の新展開――コーパスと方法論

の多様化――」を開催、愛知大学、2010年10月23日
・辞書編集班拡大班会議を共催、ホテルスワ、2010年11月13～14日
・「コーパス日本語学セミナー」を開催、台湾大学（台湾・台北市）、2010年12月4日
・日本語教育班拡大班会議（複合辞研究会）を共催、筑波大学、2011年1月29日
・日本語学班研究会を開催、大阪大学、2011年2月21日
・平成22年度公開ワークショップに参加、時事通信ホール、2011年3月14～16日

３．日本語学班の刊行物等

日本語学班による刊行物は以下の通りである。

・『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅰ』平成18年度研究成果報告書、2007年3月31日
・『コーパス日本語学ガイドブック』、平成19年度研究成果報告書、2007年9月7日
・『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅱ』平成19年度研究成果報告書、2008年4月30日
・『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅲ』平成20年度研究成果報告書、2009年3月31日
・『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅳ』平成21年度研究成果報告書、2010年3月31日
・『コーパス日本語学の新展開』平成18年度～平成22年度研究成果報告書（最終報告書）、2010年10月20日

また、日本語学班の Web サイトを開設・運用した（URL は下記の通り）。主な内容は、刊行物の案内とその訂正・更新情報、および、後述の日本語用例検索ページである。

http://www.tokuteicorpus.jp/team/jpling/

４．日本語学班の研究成果の概要１……目的(1a)

本稿冒頭に記した２種３類の目的のうち、まず日本語学班の中心的な目的である(1a)、すなわち、コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発に関わる研究成果の概要を述べる。

日本語研究の精密化と新しい研究領域・手法の開発という２つの課題は排他的な関係にはないが、研究成果を便宜上その２つに分類して述べる。その詳細およびここで省略した研究成果については最終報告書『コーパス日本語学の新展開』その他の刊行物を参照願いたい。

４．１　日本語研究の精密化

コーパスから得られる豊富な用例を利用することにより、従来少数の用例や内省に依存して行われてきた種類の日本語研究を精密化し、より実証的なものにすることができる。日本語の文法や意味に関わるさまざまな問題を題材としてその可能性を検討したが、その

主なものは複合辞の研究と意味分析という２つの見出しのもとにまとめることができる。

・複合辞の研究

複合助詞には、「に関する」と「に関しての」のように連体用法において２通りの語形を持つもの、「に対して」と「に対し」、「ために」と「ため」のように連用用法において２通りの語形を持つものがある。このような、各用法で複数の語形を持つ複合助詞「によって」「において」「に関して」「に対して」「に際して」「ために」について、『現代日本語書き言葉均衡コーパス(BCCWJ)』等のコーパスにおいて、いずれの語形が用いられているか調査した。その結果、「によって」は、意味用法によって異なりはあるものの、連体用法において圧倒的に「による」の形が用いられること、「に際して」は、他の複合助詞と異なり、連体用法において圧倒的に「に際しての」の形が用いられること、「に対して」は、接続助詞用法の場合「に対して」「に対し」のいずれも用いられるのに対して、格助詞用法の場合「に対して」が好まれること、「ために」は、目的用法の場合「ために」が、理由用法の場合「ため」が好まれることなどを明らかにした。さらに、このような語形の偏りは、複合助詞の品詞性（格助詞的か接続助詞的か）の違いによるものであることを示した。（杉本）

日本語のコピュラの形式と分布の問題は日本語文法の根幹に位置する基礎的な問題でありながら、これまで研究の関心が向けられることすらなかった。2006 年にこの問題に関する主として内省に基づく考察の結果を発表したが、コピュラ形式の分布はその一部にゆれの側面があり、内省ではその様相を正確に判断することができない。本研究では、数種類の大規模なコーパスを用いて、連体的な位置におけるコピュラ形式「な」と「の」の使い分けや、引用の「と」の直前におけるコピュラの潜在の様相について部分的な調査・分析を行った。ただし、この問題については時間その他の制約により満足の行く調査ができなかったので、今後より包括的な考察を期したい。（田野村）

・意味分析

接続助詞のように使われる「代わり（に）」は興味深い意味的二重性を示す。これは単なる意味上の問題ではなく、その背後には文法的な性格の違いがあると考えられるが、ともあれその意味的二重性はその区別が不明瞭なケースがあることもあって問題の本質が正確に捉えにくい。2008 年に作成した巨大な Web コーパスを利用し、そこから得られる多様な用例の観察・分析を通して、「代わり」の意味的二重性について立ち入った考察を行った。（田野村）

「移動先」、「留学先」、「投資先」などのように目標地点や相手方などを表す「漢語動名詞＋‘先’」の種類と用法について、特に新聞記事データベースの実例観察を行った。その結果、「動名詞＋‘先’」の中に、ある種の多義の解釈を許すものがあることが分かった。例えば、「受注先」は「受注者」を指すことと「被受注者（発注者）」を指すこととがあり、「引用先」は「引用主体（の文書）」を指すことと「被引用者（引用の出典）」を指すことがある。多様な「～先」の意味分析を行った上で、多義が生じるメカニズムの解明を行っ

た。（服部）

４．２　新しい研究領域・手法の開発

コーパス固有の特性を生かすことにより、旧来の言語研究資料では実現できなかった様々な種類の研究への道が拓かれる。4.1 で述べた日本語研究の精密化以上に力を入れてその可能性を追求した。その主な研究成果を、現代日本語の通時変化、コロケーション・呼応、マルチメディア・コーパス、統計的手法という分類に従って述べる。

・現代日本語の通時変化

国会会議録は、1947 年から今日に至る日本語の話しことばの変化の様相をうかがい知るのに有効に使える膨大な資料である。これを国立国会図書館の Web サイトから取得してコーパスとして用いることにより、過去 60 年間における日本語の変化を明らかにする試みを行った。

まず、「属する＞属す」や「論ずる＞論じる」に代表される一字漢語サ変動詞の活用の変化とそのゆれの様相については、1990 年前後の新聞記事に基づく調査・分析の結果を過去に発表したことがあったが、国会会議録から得られる用例を分析することにより、特定の一時期だけの日本語の観察からは見えてこない、活用の変化のより詳細な様相と動向を明らかにすることができた。ほかにも異形態や各種の慣用句形の選択に関わる通時変化の様相を観察・分析した。（田野村）

また、「可能性」のように尺度的な属性を表わす名詞とその属性の値の大きさ・小ささを表わす形容詞とが構成する「可能性が大きい／多い／高い／強い／濃い」のようなコロケーションについて、その頻度の通時的な推移を調査した。その結果、「～性」「～率」「～度」のような名詞では、「高い」が主として用いられる方向へと変化が進行中の可能性があることが分かった。さらに、発話年代と発話者の出生年代という２つの変数に注目し、補助動詞のイルとオルの選択、人を主語とする存在文でのイル（オル）とアルの選択、「～的｛ナ／ノ｝」の選択、「まする」の使用、副詞と否定辞との共起率などについて分析を行った。その結果、どの現象でも、同じ年代に出生した話者でも発話年代が進むにつれて使用傾向が変化する現象が見られた。これは、発話年代での区分に基づく分析では知りえなかった事実である。（服部）

現代語の通時コーパスとして、20 世紀後半の新聞コラムのコーパスを作成し、現代語の通時的研究を試みると同時に、通時コーパスの分析に関する問題点の探索を継続した。新聞のコラムを対象としたのは、テクストタイプが同じである、各年の延べ語数がほぼ等しくなる、話題が適当に分散していて特定の語彙に集中しない、という理由のほかに、各年の書き手が原則として同一個人であり、言語変化に関する書き手の影響を検討するのに好適だからである。作成したコーパスは、毎日新聞のコラム「余録」欄の、1950・60・70・80・90・2000 年の各１年分、計６年分を収めたプレイン・コーパスで、データ量（「茶筌」の処理結果）は（固有名詞・数・記号・助詞・助動詞を除く）実質的な単語で各年７～８

万語（延べ）である。これにより、20 世紀後半で増減する単語の抽出や、単語使用における「著作年代の差」と「書き手の違い」との影響関係の分析などを行った。（石井）

科学技術振興機構（JST）が運営する科学技術文献情報のオンライン・データベースサービス“JDreamⅡ”を現代語の通時コーパスとみなして、専門用語の借用／翻訳の選択がどのようになされるのかを、事例研究により試行的に検討した。“JDreamⅡ”には 1976 年以降の文献情報データが大量に蓄積されており（約 4,800 万件）、それを経年的に調べることで、専門家が原語に接触して借用／翻訳を主体的に選択した段階をとりだすことができると見込まれる。“JDreamⅡ”の（医学分野を除く）「科学技術全般ファイル」を使って、コンピュータ関連分野の“ubiquitous”という英単語（原語）が日本語の標題でどのように表現されているかを調べた結果、外国論文の訳題では翻訳（「遍在」等）を選択し、日本語論文の原題では借用（「ユビキタス」）を選択する傾向が確認でき、借用が「自国語に訳出する」というより「自国語で案出する」という位相の中で行われることが示唆された。（石井）

・コロケーション・呼応

従来、日本語研究においてコロケーションは基本的なキーワードとして広く認知されてこなかったが、昨今辞書編集や日本語教育への応用の観点から興味を引くようになってきた。日本語コーパスからコロケーション情報を抽出するには、英語の場合とはまた違った手法が必要となる。2008 年に作成した巨大な Web コーパスを用いて、日本語コーパスからのコロケーション情報の抽出手法を検討し、それによって得られた情報に基づいて日本語コロケーション辞典の数項目を試作し、また、用言の文法的性格の分析や類義的な複合辞の違いの分析などを行った。また、複合的な性質を持つコロケーションの概念に着目する意義について、実例に即して考察した。（田野村）

従来「呼応」と呼ばれた現象について、通時的・共時的の両面からコーパスを用いた分析を行った。通時的な面では、「全然」と「全く」の２つの副詞について分析したところ、「全く」に関しては、否定辞とよく共起する方向に変化しており「違う」との共起率も高まってきていることが分かった。また共時的な面では、否定形の述語とよく共起する副詞について分析し、従来知られていなかった傾向をいくつか発見した。（服部）

・マルチメディア・コーパス

コーパスを用いた言語研究を、話しことばを基本とする「言語使用」の研究に拡大・発展させていくためには、言語形式が検索できるというだけではなく、実際の発話場面における映像と音声を参照して、その使用にかかわる各種の情報をも同時に得ることのできる「マルチメディア・コーパス」が必要になる。その試作版として、国立国語研究所「テレビ放送の語彙調査」（1989 年 4～6 月）のＮＨＫ総合・教育テレビのデータを収めた「ＮＨＫコーパス」と、大阪の 6 放送局 7 チャンネルが放送した 23 種類の対談番組（2009 年 3～8 月、106 回分の放送）を収めた「対談番組コーパス」とを作成した。どちらも、特定の単語を指定すれば、それを含む文が発話された際の映像・実音声が再生できるようにした

ものである。これらを用いて、特定の言語形式・表現と非言語行動との関係（思考動詞「思う」が一人称主語の述語として発話される際の話し手の視線、擬音語･擬態語の発話時の身振りの有無、指示詞（直示用法）の発話時の指差しの共起など）の分析や、（言語使用としての）談話と映像との関係を談話分析の手法を用いて検討することなどを試みた。（石井）

・統計的手法

コーパスを用いた発見的研究の新たな可能性を探るため、程度副詞と述語の間の共起関係を、新聞記事のデータに多変量解析（因子分析）の手法を適用することによって分析し、それに基づいてそれぞれの程度副詞の特徴づけを行うことを試みた。それぞれの程度副詞と述語との共起の有無のデータを基に、程度副詞の共起傾向に関わる因子を抽出し、因子得点と述語の意味的（文体的）特性を参照しながら意義付けを試み、内省に基づく従来の記述と対照した。その結果、共起例の有無という単純なデータに基づく方法でもおよそ従来の記述と合致する結果が導かれることが分かり、この方法を他のまだ十分記述されていない現象の分析に利用しうる可能性が示された。（服部）

コーパス日本語学において「探索的データ解析」（Exploratory Data Analysis）という統計手法が有効な分析ツールとなることを、これまでの計量的日本語研究の成果・知見を検証・追試することによって確認し、その上で、探索的データ解析が用意する一連の手法のうちのどれが、日本語についてのどのような調査・研究に利用可能であるのかを、独自に用意した中学校歴史教科書や新聞コラムのコーパスを試料として明らかにした。とくに、探索的データ解析の手法のうち、日本語研究において有効と考えられる 10 の手法（幹葉表示、数値要約と平行箱型図、データのならし、ヒンジ散布度一定化のための再表現、抵抗直線、蛇行箱型図、ルートグラム、二元分析（中央値精錬法）、リジット解析、ロジット変換）をとりあげ、具体的な日本語研究への適用事例とともに、その利用法を紹介・解説した。（石井）

５．日本語学班の研究成果の概要２……目的(1b)

目的(1b)、すなわち、学界に対するコーパス利用の啓蒙・普及に関わる主な研究成果は以下の通りである。

2007 年には、コーパス初心者向けの『コーパス日本語学ガイドブック』を作成・刊行した。内容は、電子データの扱いになじみの薄い日本語研究者が自分もやってみようという気になるよう、日本語テキストを自前で処理するための初歩的なノウハウを解説したものである。本書は希望の言語研究者に無料で配布している（研究期間終了接近に伴い、2010 年 12 月より配布対象の範囲を広げた）。

2010 年 10 月には、日本語学会の 2010 年度秋季大会（愛知大学）でワークショップ「コーパス日本語学の新展開――コーパスと方法論の多様化――」を開催し、日本語学班の研究成果を発表し、参加者と質疑応答、議論を交わした。また、国外への情報発信として、同年 10 月、12 月に、それぞれ韓国（高麗大学校）、台湾（台湾大学）でコーパス日本語学

のワークショップ、セミナーを開催し、現地の日本語研究者に対して日本語のコーパスをめぐる状況を紹介するとともに研究成果を披露したほか、研究期間全体を通じて各班員がアジアやヨーロッパの各地において講演や学会発表を行った。

その他、日本語研究者が手持ちの日本語テキストをそのまま使って簡便に KWIC 索引を生成することのできるソフトウェアを作成・公開した。『コーパス日本語学ガイドブック』の刊行を機に作成し、その添付 CD-ROM に収録したものであるが、その後の改訂版をマニュアルとともに日本語学班の Web サイトでダウンロードできるようにしてある。それを Web に移植して青空文庫所収の数千作品から日本語の用例を自由に検索できるようにした日本語用例検索ページ（日本語学班 Web サイト内に設置）は、国内外の日本語研究者・学習者および一般の人々によって日々頻繁に利用されている。

６．日本語学班の研究成果の概要３……目的(2)

目的(2)、すなわち、『現代日本語書き言葉均衡コーパス(BCCWJ)』構築へのフィードバックとして行った活動の概要は以下の通りである。

当初の段階のコーパス仕様説明会などで意見や要望を述べたほか、2008 年と 2009 年に作成・配布された『BCCWJ 領域内公開データ』をその都度具体的な事例研究に適用して試用するとともにコーパス自体の内容を調査し、改善・検討を要する問題点を研究会、領域のメーリングリスト、『人工知能学会誌』の特集記事などで指摘・発言した。

また、『BCCWJ 領域内公開データ』（および、領域外の研究者のための『BCCWJ モニター公開データ』）に添付して配布されたコーパス検索ソフトウェア『ひまわり』、そして、その後公開された Web ベースのコーパス検索ソフトウェア『中納言』について、試用に基づく改善の提案を述べた。

研究活動・成果の総括：日本語教育班
代表性を有する書き言葉コーパスを活用した日本語教育研究

砂川 有里子　（班　長：筑波大学大学院人文社会科学研究科）†
井上 優　（分担者：国立国語研究所言語対照研究系）
小林 ミナ　（分担者：早稲田大学大学院日本語教育研究科）
滝沢 直宏　（分担者：名古屋大学大学院国際開発研究科）
投野 由紀夫　（分担者：東京外国語大学大学院総合国際学研究院）
山内 博之　（分担者：実践女子大学文学部）
千葉 庄寿　（連携研究者：麗澤大学外国語学部）
橋本 直幸　（連携研究者：首都大学東京オープンユニバーシティー）
奥川 育子　（協力者：筑波大学大学院人文社会科学研究科）
小西 円　（協力者：早稲田大学大学院日本語教育研究科）
清水 由貴子　（協力者：早稲田大学日本語教育研究センター）
曹 大峰　（協力者：北京日本学研究センター）
本田 ゆかり　（協力者：東京外国語大学大学院地域文化研究科）

Final Progress Report: 'Japanese Language Education' Group

Yuriko Sunakawa　(University of Tsukuba)
Masaru Inoue　(National Institute for Japanese Language and Linguistics)
Mina Kobayashi　(Waseda University)
Naohiro Takizawa　(Nagoya University)
Yukio Tono　(Tokyo University of Foreign Studies)
Hiroyuki Yamauchi　(Jissen Women's University)
Shoju Chiba　(Reitaku University)
Naoyuki Hashimoto　(Tokyo Metropolitan University)
Ikuko Okugawa　(University of Tsukuba)
Madoka Konishi　(Waseda University)
Yukiko Shimizu　(Waseda University)
Dafeng Cao　(The Beijing Center for Japanese Studies)
Yukari Honda　(Tokyo University of Foreign Studies)

１．日本語教育班の活動目的と課題

日本語教育班は，「現代日本語書き言葉均衡コーパスを日本語教育に活用する方法の開発」を目的とする。従来教師の経験と勘にもとづいて作成されていた日本語教科書，教材，シラバスについて，客観的な言語データにもとづいて考えるべき部分を見極め，コーパスを日本語教育に活用する方法について検討するため，以下の 3 つの課題を設定して活動を行った。

†sunakawa@sakura.cc.tsukuba.ac.jp

・課題１「日本語教材コーパスの作成と分析」
・課題２「書き言葉均衡コーパスを活用した日本語教材作成法の開発」
・課題３「日本語教育のためのコーパス活用ツールの開発」

２．各課題の活動成果

課題１「日本語教材コーパスの作成と分析」

・初級・中級向け日本語総合教科書および読解教材 48 種について，テキストの内容を「課題」「解説」「例文」「練習問題」などに種別してタグ付けを行った。研究領域内で利用するための著作権処理を行っている。（井上）
・日本語教材コーパスを用いた語彙表を試作し，BCCWJ の語彙情報との比較をおこなった。（井上・千葉）
・BCCWJ の評価と試用を目標に，日本語教材の分析と作成に向けた多種のコーパス（BCCWJ を含む）の比較利用とその可能性の検証を行った。（曹）

課題２「書き言葉均衡コーパスを活用した日本語教材作成法の開発」

＜日本語教育のためのコロケーション研究＞

・BCCWJ(2009)と Google 刊行の N-gram ファイルを利用し，コロケーションと拡大コロケーション（lexical bundle）の抽出と，これまでに作成したリストの精緻化を行った。それをもとに品詞ごとのコロケーションリストと拡大コロケーションのリストを作成した。（滝沢）
・BCCWJ(2008)と BCCWJ(2009)の結果を比較対照させることにより，コーパスサイズとコロケーション認定の統計値の妥当性との関連について検討した。（滝沢）

＜文法項目・語彙項目に関する「原型シラバス」の作成＞

・言語政策班が作成した検定教科書コーパスを用いた文型調査を行い，年少者に対する日本語教育の文型指導に資する調査報告を行った。（砂川・清水・奥川）
・BCCWJ 領域内公開データを用いた機能語調査をもとに複合辞の類義語研究を行った。（砂川・清水・奥川）
・BCCWJ 領域内公開データ(2009 年度版)を使い，『日本語能力試験出題基準』１・２級の機能語を中心とする 439 項目についてジャンル別の頻度調査を行い，「BCCWJ による機能語データベース（スタンドアロン版）」を作成した。（砂川・清水・奥川・千葉）
・初級文型の「原型シラバス」作成のためのケーススタディを行い，発表した。具体的には，「い形容詞」「義務の表現」「原因・理由の表現」「否定疑問（のではないか）」「伝聞表現」について分析することによって，「原型シラバス」に盛り込むべき内容について提案した。（小林・小西）
・橋本(2008)で提案した話題別語彙リストの改訂作業を行なった。対数尤度比を用いて BCCWJ に出現する語を 100 の話題に分類し，それらを話題別語彙リストに付加していった。（山内・橋本）
・KY コーパス・親密度・BCCWJ の三者を突き合わせることにより，「使用」という観点から見た場合の実質語の難易度に関する分析を行なった。その結果，具体物を表す名詞の使用は，「難易度」ではなく，「親密度」から説明することができ，抽象

概念を表す名詞の使用は，主に「難易度」で説明することが可能であることが明らかになった。（山内・橋本）
・BCCWJ(2009)を使って語彙頻度分布調査に基づく日本語学習語彙表を作成した。（本田）

課題３「日本語教育のためのコーパス活用ツールの開発」

・BCCWJ 領域内公開データのサブコーパス 10 種より無作為抽出された 2000 語×100 ファイル，計 20 万語の語彙頻度・分布統計を算出した。（投野・本田）
・構築中の 2000 語単位のファイルおよび長単位データをもとにスケッチエンジンへの実装を試みた。（投野）
・BCCWJ を用いた語彙統計情報データベースの公開にむけ，パッケージ化＆ドキュメント作成を進めた。（千葉）
・語彙情報データベースの Web サービス化に関する実験とツールの試作を行った。（千葉）
・現在利用できる形態素・係り受け解析ツールを用いて解析したコーパスから語彙パターンを柔軟に検索するための検索式（言語）の仕様の策定と検索ツールの試作を行った。（千葉）

３．班会議・研究会の活動報告

３．１　班会議の実施

年に 3 回～6 回の班会議を開催し，活動についての審議，研究発表，ワークショップ等を行った。

2006 年度

第 1 回	2006 年 7 月 31 日（月）	早稲田大学西早稲田キャンパス
第 2 回	2006 年 8 月 22 日（火）	東京大学駒場キャンパス
第 3 回	2006 年 10 月 1 日（日）	早稲田大学西早稲田キャンパス
第 4 回	2006 年 11 月 3 日（日）	国立国語研究所
第 5 回	2006 年 12 月 17 日（日）	早稲田大学西早稲田キャンパス
第 6 回	2007 年 2 月 11 日（日）	早稲田大学西早稲田キャンパス

2007 年度

第 1 回	2007 年 4 月 15 日（日）	早稲田大学西早稲田キャンパス
第 2 回	2007 年 6 月 1 日（金）	奈良先端科学技術大学（データ班と合同）
第 3 回	2007 年 7 月 15 日（日）	早稲田大学西早稲田キャンパス
第 4 回	2007 年 9 月 30 日（日）	早稲田大学西早稲田キャンパス
第 5 回	2007 年 12 月 2 日（日）	早稲田大学西早稲田キャンパス

2008 年度

第 1 回	2008 年 4 月 12 日（土）	早稲田大学早稲田キャンパス
第 2 回	2008 年 7 月 6 日（日）	早稲田大学早稲田キャンパス
第 3 回	2008 年 7 月 30 日（水）	早稲田大学早稲田キャンパス
第 4 回	2008 年 12 月 21 日（日）	早稲田大学早稲田キャンパス

2009 年度

第 1 回	2009 年 5 月 9 日（土）	早稲田大学早稲田キャンパス
第 2 回	2009 年 6 月 7 日（日）	国際情報学研究所学術総合センター
第 3 回	2010 年 1 月 23 日（土）	早稲田大学早稲田キャンパス

2010 年度

第 1 回	2010 年 5 月 8 日（土）	早稲田大学 26 号館
第 2 回	2010 年 8 月 30 日（月）	国立国語研究所
第 3 回	2011 年 3 月 21 日（月）	筑波大学総合研究 A 棟

３．２　複合辞研究会

データ班，日本語学班，辞書編集班と合同で日本語研究，日本語教育，辞書編集などで多くの課題を抱えた複合辞をテーマとする研究会を平成 19 年度より年 1 回開催した。

第 1 回	2007 年 12 月 16 日（日）	筑波大学総合研究 A 棟
第 2 回	2009 年 12 月 22 日（日）	筑波大学総合研究 A 棟
第 3 回	2010 年 2 月 13 日（土）	筑波大学総合研究 A 棟
第 4 回	2011 年 1 月 29 日（土）	筑波大学総合研究 A 棟

４．平成 22 年度の発表実績

最終年度の平成 22 年度は，研究成果の発表に加えて，コーパスを日本語教育に活用する研究と教育実践を活性化させることを目指した啓蒙的な活動に力を入れた。

４．１　口頭発表・講演

- 2010 年 5 月 29 日　ウズベキスタン日本語教育セミナーワークショップ「コーパスを活用した日本語教育研究」，タシケント日本文化センター（砂川）
- 2010 年 6 月 12 日　講演「日本語教育語彙リストの構築にむけて」，ワークショップ「日本語教育とコーパス」麗澤大学（橋本）
- 2010 年 6 月 12 日　講演「現代日本語書き言葉均衡コーパス（BCCWJ）を利用した，日本語教材の評価」，ワークショップ「日本語教育とコーパス」麗澤大学（千葉）
- 2010 年 7 月 31 日　世界日本語教育大会研究発表「内容と能力を重視した日本語教材の開発－多文化理解のための日本語教育をめざして－」，国立政治大学（曹）
- 2010 年 8 月 1 日　世界日本語教育大会研究発表「素性情報を利用した，解析済み日本語コーパスからの語彙パターンの抽出」，国立政治大学（千葉）
- 2010 年 8 月 1 日　世界日本語教育大会パネルセッション「これからの日本語教育養成課程に求められるもの　アジア 6 地域の大学間交流活動を通した連携と今後の課題」「中国における日本語教員養成とその課題」国立政治大学（曹）
- 2010 年 8 月 31 日　特定領域「日本語コーパス」全体会議ポスター発表「異なるジャンルにおける伝聞表現」，国立国語研究所（小西）
- 2010 年 8 月 31 日　特定領域「日本語コーパス」全体会議研究発表「実質語の難易度について」，国立国語研究所（山内・橋本）

・2010 年 10 月 16 日　研修会「コーパスを利用した言語研究の成果と初級文法シラバス」，愛媛大学城北キャンパス（小林）
・2010 年 11 月 14 日　「原型シラバス構築のための基礎作業」，特定領域「日本語コーパス」辞書編集班拡大班会議（小林）
・2010 年 11 月 14 日　「「検定教科書コーパス」を用いた定型表現の調査報告－小学校「算数」と中学校「数学」に出現する定型的な条件表現について－」，特定領域「日本語コーパス」辞書編集班拡大班会議（砂川・清水・奥川・近藤）
・2010 年 11 月 20 日　講演「学習者コーパスの使用方法の一例－ＫＹコーパスからわかったこと－」，国立国語研究所共同研究プロジェクト「学習者の言語環境と日本語の習得過程に関する研究」第 2 回公開研究会「日本語教育研究における学習者コーパスの役割」，国立国語研究所（山内）
・2011 年 1 月 24 日　「言語の慣習性とコーパス」，大阪大学文学研究科講演（滝沢）
・2011 年 1 月 27 日　「コーパスによる日本語の分析」，2010 年度南山大学人間文化研究科言語科学専攻講演会（滝沢）
・2011 年 2 月 28 日　「言語の慣習性とコーパス」，DePaul University, Department of Modern Languages 講演（滝沢）
・2011 年 3 月 9 日　言語処理学会第 17 回年次大会ポスター発表「大規模均衡コーパスを利用した語彙・文法情報の評価とその応用」，豊橋技術科学大学（千葉）
・2011 年 3 月 15 日　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）シンポジウム「日本語コーパスと外国語としての日本語研究」での発表「海外の日本語教育から見た均衡コーパス－日本語教材の評価・比較・編集－」，時事通信ホール（曹）
・2011 年 3 月 16 日　特定領域「日本語コーパス」平成 22 年度公開ワークショップ研究発表「日本語教育における初級シラバスの再評価－BCCWJ にみられた「出現形の偏り」を手がかりに－」，時事通信ホール（小林）
・2011 年 3 月 16 日　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ポスター発表「BCCWJ を用いた語彙･文法情報のプロファイリングとその応用」，時事通信ホール（千葉）

４．２　パネルセッション・国際フォーラム

・2010 年 7 月 31 日　パネルセッション『日本語教育におけるコーパスの活用』世界日本語教育大会，国立政治大学（砂川・井上・橋本・小林・奥川）
・2010 年 10 月 16 日　北京日本学研究センター30 周年記念大会パネルセッション「日本語学と日本語教育におけるコーパスの応用」，北京日本学研究センター（徐・山崎・砂川・曹・朱・施）
・2010 年 12 月 11 日～12 日　筑波大学との共催による国際フォーラム「日本語学習辞書の開発と日本語研究」，筑波大学国際会議室（砂川・曹・杉本・矢澤）
・2011 年 3 月 18 日　パネルセッション「「大規模日本語書き言葉コーパス」と日本語教育」，早稲田大学日本語教育学会 2011 年春季大会，早稲田大学早稲田キャンパス（小林・山崎・投野・奥川・小西・清水・近藤）

４．３　ワークショップ

・2010 年 6 月 12 日　麗澤大学との共催によるワークショップ「日本語教育とコーパス」
　　麗澤大学（橋本・千葉）
・2010 年 7 月 8 日・9 日　ワークショップ「エディターを使ったコーパス検索（1）」
　　筑波大学（滝沢）
・2011 年 3 月 7 日・8 日　ワークショップ「エディターを使ったコーパス検索（2）」
　　筑波大学（滝沢）

４．４　発表論文

・平成 22 年度の発表論文数：20（成果 DVD「特定領域研究「日本語コーパス」研究成果報告」所収）

５．総括班ＤＶＤ掲載予定の成果物

・平成 18 年度～22 年度の研究成果報告書
・コロケーションおよび拡大コロケーションのファイル（滝沢）
・『言語研究のためのテキスト処理入門』（滝沢）
・BCCWJ による機能語データベース（スタンドアロン版）（砂川・清水・奥川・千葉）
・BCCWJ サブコーパス 10 種の評価用語彙頻度・分布統計（各サブコーパス 2000 語×100 ファイル＝20 万語）（投野・本田）
・BCCWJ を用いた初級シラバスの再評価（小林・小西）
・BCCWJ による語彙情報データベース（sqlite3），サンプル分析ツールとマニュアル（千葉）

研究活動・成果の総括：言語政策班
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

田中　牧郎（班　長：国立国語研究所言語資源研究系）†
相澤　正夫（分担者：国立国語研究所時空間変異研究系）
斎藤　達哉（分担者：専修大学文学部）
棚橋　尚子（分担者：奈良教育大学教育学部）
近藤明日子（連携研究者：国立国語研究所コーパス開発センター）
河内　昭浩（協力者：群馬県立館林高等学校）
鈴木　一史（協力者：東京大学教育学部付属中等教育学校）

Final Progress Report: 'Language Policy' Group

TANAKA Makiro　（National Institute for Japanese Language and Linguistics）
AIZAWA Masao　（National Institute for Japanese Language and Linguistics）
SAITO Tatsuya　（Senshu University）
TANAHASHI Hisako（Nara University of Education）
KONDO Asuko　（National Institute for Japanese Language and Linguistics）
KAWAUCHI Akihiro　（Tatebayashi High School）
SUZUKI Kazufumi（Tokyo University Secondary Education School）

１．研究活動の概要

言語政策班では、これまでコーパスがほとんど使われていなかった、国語政策と国語教育の分野にコーパスを導入することを目的に、二つの分野の諸課題で共通して問題になることが多い語彙と漢字を取り上げて、研究を進めてきた。まず、小学校・中学校・高等学校で使われている教科書の全文コーパスを作成し、BCCWJ とあわせてコーパスに基づく語彙表・漢字表を作成する研究を行った。そして、作成された語彙表・漢字表を通してコーパスを使うことで、国語政策と国語教育の諸課題にどのように取り組んで行くことができるかについて研究を進めた。

５年間を通して班会議での研究発表と討議を活動の中心に据え、成果がまとまったものから順次、学会や論文誌などで発表し、折り返し時点（2008 年 9 月）と最終時点（2011 年 2 月）に報告書を発行し、最終の報告書と同時に語彙表・漢字表等のデータを公開する CD-ROM を配布した。最終の報告書と CD-ROM の内容を、特定領域成果 DVD にそのまま収録した。また、応用分野を扱っていることを考慮し、国語政策と国語教育の実践的場面に直接貢献する成果を上げることにも努めた。

２．語彙表・漢字表の作成

2.1　語彙表・漢字表の概要

言語政策班では BCCWJ や教科書コーパスを用いて、多種多様な語彙表・漢字表を試作しつつ研究を進めたが、最終成果物として公開した語彙表・漢字表は、以下に記す五種類である。公開した語彙表・漢字表は、国語政策や国語教育の分野で応用の範囲が広く、今後の研究の基盤となり得ると考えたものである。他にも、もっと具体的な課題に対応するための語彙表・漢字表も色々と作成したが、それらは、各研究者の論文や著書等に付随する

† mtanaka@ninjal.ac.jp

ものとして今後公開していくことを考えている。

2.2 教科書コーパス語彙表

小学校・中学校・高等学校の全教科全学年一種ずつの全文を対象とした「教科書コーパス」（言語政策班が作成）[1]に用いられている全語彙の語彙表（異なり語数 50,329、延べ語数 2,518,486）。UniDicの短単位で解析し、小学校教科書は人手により修正（中学校・高等学校教科書は無修正）したものを集計。校種・学年・教科別の頻度（度数・出現率）、初出学年、BCCWJの図書館サブコーパス（LB）と比較した際の教科や校種ごとの特徴度（対数尤度比に基づく）などを収録している。

2.3 BCCWJ 主要コーパス語彙表

『現代日本語書き言葉均衡コーパス』（BCCWJ）のサブコーパスのうち、図書館書籍（固定長、LB_FL）、出版書籍（固定長、PB_FL）、雑誌（固定長、PM_FL）、新聞（固定長、PN_FL）、Yahoo!知恵袋（可変長、OC_VL）、Yahoo!ブログ（可変長、OY_VL）の六つのサブコーパスの語彙頻度（度数・使用率・サンプル数）を収録した語彙表（2010 年 12 月 9 日現在のデータに基づく。以下同じ）。各サブコーパスに対してカバー率の基準を適用し 5 つに区画した語彙レベルを設定。区画の基準とレベル別語数は、表１・表２の通りである。

表 1　レベル分けの基準

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　各サブコーパスの延べ語数・異なり語数（レベル別）

	LB_FL		PB_FL		PM_FL	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
全体	3,938,696	86,002	3,903,395	82,784	896,988	45,900
レベル a	3,074,655	4,177	3,045,639	3,842	700,831	4,336
レベル b	395,994	6,330	391,312	5,609	92,353	5,293
レベル c	242,911	11,595	239,221	10,506	51,085	7,493
レベル d	118,642	14,176	124,601	14,290	37,925	13,984
レベル e	106,494	49,724	102,622	48,537	14,794	14,794

	PN_FL		OC_VL		OY_VL	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
全体	624,020	35,727	2,762,864	49,809	6,127,125	76,823
レベル a	486,976	3,420	2,155,871	2,071	4,779,106	3,441
レベル b	63,784	4,045	275,758	2,776	617,945	4,724
レベル c	40,018	6,941	165,957	5,122	372,114	8,406
レベル d	20,607	8,686	83,349	7,062	181,482	10,285
レベル e	12,635	12,635	81,929	32,778	176,478	49,967

[1] 「教科書コーパス」の全体は、著作権者との合意により非公開だが、一部を BCCWJ 非母集団サブコーパスの教科書サンプルとして公開する（すでに検索デモンストレーションサイトでは公開している）。

2.4　学校・社会対照語彙表

「教科書コーパス語彙表」に収録した中学校・高等学校分の語彙頻度（教科別度数）や初出学年と、「BCCWJ 主要コーパス語彙表」に収録した語彙レベルとを比較対照し、教科や媒体の特徴語を見やすく表示し、中学校・高等学校において、社会で現に使われている実態を踏まえた語彙教育のあり方を考える際の参照資料となるように再編した語彙表である。

表３は、「学校・社会対照語彙表」の語彙素読みが「カ」の部分の冒頭部について、語彙素情報、初出学年、全教科の度数、主要４教科の度数、BCCWJ 主要サブコーパスの語彙レベル、特徴媒体、特徴教科についての情報だけを抜き出して示したものである。

表３　「学校・社会対照語彙表」（部分）

語彙素情報				教科書						BCCWJ						特徴語	
読み	語彙素	語種	品　詞	初出学年	全教科	国語	数学	理科	社会	LB	PB	PM	PN	OC	OY	特徴媒体	特徴教科
カ	カ	固	名詞-固有名詞-人名-姓	中	4	2	0	0	2	d	c	d	c	e	d		
カ	下	漢	接頭辞	小_前	9	3	0	2	0	b	b	a	c	c	b	書	
カ	下	漢	接尾辞-名詞的-副詞可能	中	326	16	2	43	245	a	a	a	a	b	a	教	社
カ	化	漢	接尾辞-名詞的-サ変可能	小_後	4326	192	28	904	2306	a	a	a	a	a	a	教	理社技情
カ	価	漢	接尾辞-名詞的-一般	中	48	0	0	17	0	d	b	d	e	c	c	教	理技
カ	価	漢	接尾辞-名詞的-助数詞	高	62	0	0	57	0	e	d			e	e	教	理技
カ	加	漢	名詞-普通名詞-一般	高	5	0	0	0	5	c	c	d	d	e	c		
カ	可	漢	名詞-普通名詞-形状詞可能	中	17	4	2	0	2	b	b	a	c	b	b	書	
カ	夏	漢	名詞-普通名詞-一般	小_前	0	0	0	0	0	e	e				e		
カ	家	漢	接尾辞-名詞的-一般	小_前	818	178	2	21	389	a	a	a	a	a	a	書	芸

2.5　教科特徴語リスト

「教科書コーパス」（中学校・高等学校分）のうち主要学習部分のみ [2]の語彙を長い単位で解析し、語彙頻度を図書館書籍（LB_FL）のそれと比較して特徴度を算出し（対数尤度比に基づく）、特徴語の候補を抽出した後、人手により不適切なものを除外したリストである。表４は、中学校の理科の特徴語を五十音順に配列したものの冒頭部分である。レベル_LBはカバー率によって区画した図書館書籍（LB）の語彙レベルである [3]。主要学習部分だけを対象にしていること、長い単位を採用していることの二点で、2.4 の「学校・社会対照語彙表」に収録する特徴語よりも、実践で使いやすい語彙リストになっている。この語彙リストの詳細は、近藤（2011）を参照してほしい。

[2] 「教科書コーパス」は、本文の性質別にマークアップが施されており、主要学習部分だけを取り出して解析対象にすることなどができる。

[3] 「教科特徴語リスト」の場合は長い単位によっているので、2.3 で記した「BCCWJ 主要コーパス語彙表」のレベル分けとは別の基準が必要になる。あらためて基準を設定した。

表４　「教科特徴語リスト」（中学・理科）の冒頭

語彙素読み	語彙素	語種	品詞	度数	特徴度	度数_LB	レベル_LB
アクエイキョウ	悪影響	漢+漢	合成名詞	5	22.75	36	Ⅲ
アサガタ	朝方	和	名詞-普通名詞-副詞可能	3	21.14	5	Ⅴ
アツガミ	厚紙	和	名詞-普通名詞-一般	4	27.88	7	Ⅳ
アツリョク	圧力	漢	名詞-普通名詞-一般	9	30.56	123	Ⅲ
アメ	雨	和	名詞-普通名詞-一般	14	24.68	494	Ⅰ
アラワス	表わす	和	動詞-一般	43	129.57	729	Ⅰ
アルカリセイ	アルカリ性	外+漢	合成名詞	12	101.75	7	Ⅳ
アルミニウム	アルミニウム	外	名詞-普通名詞-一般	4	23.81	13	Ⅳ
アンザンガン	安山岩	漢+漢	合成名詞	3	24.96	2	Ⅴ
イオウ	硫黄	漢	名詞-普通名詞-一般	9	64.37	14	Ⅳ

2.6　NDC ジャンル別漢字出現頻度表

BCCWJ の出版書籍（固定長、PB_FL）を用いて、NDC（日本十進分類法）に基づくジャンルごとに、漢字の度数、度数%、累積度数%、総合順位（全ジャンルでの順位）との差、順位差ランク、サンプル数、サンプルカバー率などを収録した漢字表である。表 5 は、自然科学のジャンルの漢字について、度数上位 10 字の情報を示したものである。この漢字表については、3.3 で再度言及する。

表５　NDC ジャンル別漢字出現頻度表の「自然科学」の上位 10 字

順位	漢字	度数	度数%	累積度数%	殊別	配当学年	総合順位	順位差	順位差ランク	サンプル数	サンプルカバー率	サンプルカバー率（全体）	サンプルカバー率差
1	人	1944	0.894%	0.894%	常用	1	1	0	F	464	75.4%	81.3%	-5.9%
2	一	1861	0.856%	1.750%	常用	1	2	0	F	504	82.0%	87.9%	-6.0%
3	生	1848	0.850%	2.600%	常用	1	8	5	E	471	76.6%	67.6%	9.0%
4	分	1831	0.842%	3.443%	常用	2	6	2	E	507	82.4%	74.0%	8.4%
5	性	1567	0.721%	4.163%	常用	5	43	38	E	403	65.5%	39.8%	25.7%
6	的	1529	0.703%	4.867%	常用	4	7	1	E	456	74.1%	61.8%	12.3%
7	大	1455	0.669%	5.536%	常用	1	3	-4	G	490	79.7%	77.4%	2.3%
8	体	1377	0.633%	6.169%	常用	2	38	30	E	417	67.8%	51.8%	16.0%
9	者	1374	0.632%	6.801%	常用	3	16	7	E	337	54.8%	55.1%	-0.3%
10	療	1302	0.599%	7.400%	常用		363	353	D	242	39.3%	6.0%	33.3%

３．国語政策でのコーパスの活用

3.1　国語政策の分野へのコーパス導入の概要

国語政策は、公共的な場で用いられる言語について、多くの人が依拠しやすい規範を定めていこうと、漢字を中心に具体策がとられてきたが、基本的に、言語の実態を把握した

上で、問題ある部分に手当てをしていくという考え方があったと言ってよい。この言語の実態把握について従来は、個々の政策課題ごとにその都度調査が行われてきたが、現代語の書き言葉を代表する BCCWJ を用いれば、より適切な実態把握を効果的に行うことができるようになると思われる。加えて、UniDic により均質な形態素解析が実現されれば、従来は実態把握が不十分だった語彙についての具体策を手がけることが可能になる。そして、語彙情報を整理した上で漢字を見ることで、漢字政策への新しい視座を持つことも期待できる。このような展望で研究を進め、難解用語に関わる語彙政策課題にデータを提供することと、漢字政策に新しい視点を加えることなどに、具体的な成果が得られた。

3.2 難解用語の抽出と序列化

BCCWJ を一般的な語彙が反映したコーパスと扱い、医療専門用語が反映した医療分野のコーパス（後述する「病院の言葉」プロジェクトで作成）と語彙頻度を比較し、専門用語の抽出を試みた。医療分野のコーパスについて、医療専門家を読者とする文章を集めた専門家向けコーパスと、患者など一般人を読者とする非専門家向けコーパスとに分け、相互に語彙頻度を比較し、専門度による序列化を試みた。さらに、非専門家向けコーパスにおける語彙頻度（度数・記事数）を用いて、非専門家にとっての重要度で序列化を試みた。この作業によって、専門家から非専門家に対して重要事項を伝える際に、難解な専門用語が情報伝達の障害になっている「難解用語の言語問題」（田中・相澤 2010）に対応する基礎データが作成できた。このデータを、別に行った質問調査などと突き合わせて検証したところ、目的を達している面とまだ不十分な面との両面があることが明らかになった。詳細は、田中・近藤（2011）に記した。

なお、作業結果のデータの一部は、この言語問題に取り組み「『病院の言葉』を分かりやすくする提案」[4]を行った、国立国語研究所「病院の言葉」委員会に提供し、具体的な語彙政策に貢献した。

3.3 漢字政策への新しい視点

2010 年に内閣告示となった新しい「常用漢字表」は、文化審議会国語分科会において、書籍・新聞・ウェブ等の漢字頻度調査に基づいた議論を経て決められたものである。その議論の過程で、国立国語研究所から文化審議会に提出した、問題となった漢字の訓や語の表記のデータ（相澤・小椋・斎藤 2008）は、BCCWJ を用いて言語政策班の研究者が作成したものであり、やはり具体的な政策に貢献した例である。

今回の文化審議会での本格的な検討には BCCWJ の完成が間に合わなかったが、今後の漢字政策には検討の開始時から BCCWJ を活用することが想定でき、これまでには十分議論できなかった視点からの多角的な検討が可能になると見込まれる。例えば、表 5 に示した「ジャンル別漢字出現頻度表」を分析して斎藤（2011）が明らかにした、ジャンルによって漢字の出現順位に大きな差があるという事実は、規範となる漢字表にどの漢字を入れどの漢字を除外するかの具体的議論に役立つだろう。また、固有名詞での使われやすさを網羅的に把握したデータに基づけば、人名用漢字・常用漢字表・学年別漢字配当など主要な漢字政策の議論を活性化させる可能性が大きいことを、相澤（2011）は問題提起している。

[4] 国立国語研究所「病院の言葉」委員会（2009）および http://www.ninjal.ac.jp/byoin/。

４．　国語教育でのコーパスの活用

4.1　国語教育の分野へのコーパス導入の概要

国語教育においては、従来は漢字には体系化された規範や指導手順があったが、語彙は話題になることは多かったが十分に体系化されたそれがなかった。BCCWJ と UniDic はこの状況を大きく変え、語彙教育を体系化させていくことが期待できる。そして、語彙教育の進展は、漢字教育の見直しにつながっていくのではないかと考えられる。こうした展望のもと、現代の書き言葉の実態と教科書の実態を反映した語彙表を作成し、これを基盤に置いて、語彙教育のあり方を考える研究と、国語の授業実践をどのように工夫するかの研究を行った。

4.2　語彙教育のあり方の研究

国語教育で語彙を扱うには、学習の基本となる語彙や習得の目標となる語彙を分類したりした規範的な語彙リストがあるのが望ましい。従来もそれはあったが専門家の見識に基づいて語彙が選定されており、客観的な選定根拠は示されていない[5]。「BCCWJ主要コーパス語彙表」に収録した六種のサブコーパスの語彙レベルを指標に、国語教育における重要語彙リストをどのように作っていくかについて研究を行った（田中 2011a）。その結果、図書館書籍（LB）の語彙は、一般的な語彙のありようを反映し、文章語をよく取り込み、語彙の基本的部分が安定していることから、その語彙レベルを重要語彙リストの指標として扱うことができるという見通しを得た。これに、専門的な語彙を取り除く指標として図書館書籍（LB）のサンプル数、及び、平易な語彙を取り除く指標としてYahoo!知恵袋（OC）の語彙レベルを判断材料に加えれば、重要語彙リストとしての妥当性は高まると考えられる。

学校における語彙教育は国語科のみではなく、各教科の専門教育の中でも行われているが、両者を連携させる取り組みが求められ、その際にコーパスや語彙表は有力なツールになると予想される。そこで、教科書コーパス（中学校・高校分）に基づく教科特徴語リストを利用して、教科教育で取り上げられる語彙を特定し、それを国語科における語彙教育と関連づける方法を検討した（田中 2011b）。例えば、「抵抗」「圧力」「反発」「摩擦」は、いずれも理科と社会科の特徴語になっており、生徒は、理科で自然科学の専門用語としての原理や現象を学習し、社会科で一般用語としての意味や概念に頻繁に触れることが見て取れる。生徒の概念形成を確実なものにし語彙力を高めていくには、こうした学習実態を踏まえて適切な指導を行っていく手順を研究していくことが必要だと思われる。表 3 に一部を示した「学校・社会対照語彙表」はその材料を豊富に提示している。

今回の研究で、語彙レベルや教科特徴語など語彙の基本情報がデータベース化されたことは、様々な新しい研究を生んでいくと思われる。その一つの方向として、鈴木（2011）は、中学生・高校生の作文で使われている語彙が、どのようなレベルの語彙なのか、どの

[5] 阪本（1984）などが代表的なもの。専門家選定による語彙リストの価値は高いが、語の選定の根拠となる客観的データがあると議論は活性化するだろう。

ような教科と関わりが深いのかについて分析し、学年の進行に伴って生徒の使用語彙がどのように変わっていくのかをとらえようとしている。

4.3　国語の授業実践の工夫

コーパスや語彙表・漢字表は、教師や生徒が日々実践している授業活動にも直接役立てることができる。河内（2011）は作文の授業にコーパスを用いた実践例を報告している。それによれば、テーマ型作文の課題に取り上げられる語彙について、教科書コーパスやBCCWJ を検索して共起語句を抽出して生徒に見せることで、発想の補助資料として効果が上がったという。書籍・新聞・インターネットを使って作文の材料を集める場合とは別の効果が、コーパスによる材料集めにはあると言えそうである。

また、棚橋（2011）は、小学校の教育漢字の一部が、教科書コーパスの特定の教科に偏って出現する事実を踏まえ、そうした漢字そのものの学習を国語科ではなく当該教科で行う可能性について研究し、６年生の社会科の授業で「憲」の字を学習する実践を報告したものである。従来の漢字指導が、字形や音訓に傾きがちであったものを、その漢字が使われる語彙や概念の学習とともに指導されるべきことを具体的に示している。

4.4　国語教師対象のワークショップの開催

国語教育の分野にコーパスを普及するには、国語教師を中心とする国語教育の実践家自らがコーパスを使った様々な工夫を重ねるようになることが望まれる。そうした機運を作るために、国語教師・教材開発者等を対象とする「国語教育とコーパス」をテーマとしたワークショップを３回開催し、4.2、4.3 で述べた研究などを紹介し、参加者にコーパスを利用する体験をしてもらった。総じて、参加者がコーパスに抱く潜在的な期待は大きく、教育活動の具体的場面を想定した教師用利用ツールを開発することなどで、ニーズを喚起していく可能性は大きいと感じられた。

５．今後に向けて

コーパスになじみのない分野でコーパスを使う研究は、模索と試行錯誤の繰り返しであった。研究成果は十分に完結したものにはならなかったが、国語政策や国語教育において従来から議論されていた問題に対して、現在使われている語彙や漢字の実態や、現在教えられている語彙や概念の実態を映し出したデータベースを整備することで、確かな議論に導いたり新しい論点を加えたりすることが、十分に可能であるという見通しを得ることはできた。こうした実態把握に基づく議論を経た上であれば、政策課題や教育実践に取り組む具体的局面で、当事者が直接コーパスを使う効果的な手順を構築していくこともできるのではないかと考えられる。言語政策班の研究を出発点として、より応用的な研究に踏み出す段階が来ていると考えられる。

文献

相澤正夫（2011）「漢字政策に役立つ漢字表のあり方―固有名に使われる漢字の検討のために―」（田中・相澤・斎藤ほか 2011 に収録）

相澤正夫・小椋秀樹・斎藤達哉（2008）「『現代日本語書き言葉均衡コーパス』に基づく漢字音訓一覧表」（文化審議会国語分科会漢字小委員会漢字ワーキンググループ用資料）

河内昭浩（2011）「『テーマ型作文』の語彙―『テーマ語彙集』の活用による―」（田中・相澤・斎藤ほか 2011 に収録）

国立国語研究所「病院の言葉」委員会(2009) 『病院の言葉を分かりやすく―工夫の提案―』（勁草書房）

近藤明日子（2011）「中学校・高校教科書の教科特徴語リストの作成」（田中・相澤・斎藤ほか 2011 に収録）

阪本一郎（1984）『新教育基本語彙』（学芸図書）

斎藤達哉（2011）「BCCWJ による『NDC ジャンル別漢字出現頻度表』の分析」（田中・相澤・斎藤ほか 2011 に収録）

鈴木一史（2011）「作文コーパスからみる生徒の使用語彙」（田中・相澤・斎藤ほか 2011 に収録）

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）『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』（特定領域研究「日本語コーパス」言語政策班）

田中牧郎・相澤正夫（2010）「難解用語の言語問題への具体的対応―「外来語」と「病院の言葉」を分かりやすくする提案―」（『社会言語科学』13-1）

田中牧郎・近藤明日子（2011）「難解用語の抽出と序列化におけるコーパスの利用―医療用語を例に―」（田中・相澤・斎藤ほか 2011 に収録）

田中牧郎（2011a）「語彙レベルに基づく重要語彙リストの作成―国語政策・国語教育での活用のために―」（田中・相澤・斎藤ほか 2011 に収録）

田中牧郎（2011b）「『分類重要語彙リスト』の作成による教科教育と語彙教育の関連づけ」（田中・相澤・斎藤ほか 2011 に収録）

棚橋尚子（2011）「他教科における漢字指導実践―社会科（小学校第 6 学年）の事例―」（田中・相澤・斎藤ほか 2011 に収録）

研究活動・成果の総括：辞書編集班
コーパスを利用した国語辞典編集法の研究

荻野綱男（班長：日本大学文理学部）†
近藤泰弘（分担者：青山学院大学文学部/国立国語研究所言語資源研究系）
矢澤真人（分担者：筑波大学大学院人文社会科学研究科）
丸山直子（分担者：東京女子大学現代教養学部）

Final Progress Report: 'Dictionary Compilation' Group

Tsunao Ogino (College of Hum. and Sci., Nihon University)
Yasuhiro Kondo (College of Literature, Aoyama Gakuin University/
National Institute for Japanese Language and Linguistics)
Makoto Yazawa (Graduate School of Hum. and Soc. Sci., Tsukuba University)
Naoko Maruyama (School of Arts and Sci., Tokyo Woman's Christian University)

１．辞書編集班の研究目的

辞書編集班は、全体として、コーパスを用いた辞書編集の方法を研究する。

既存のコーパスを利用してこのような課題を追求する一方、特定領域研究の進行とともに利用に供される新しいコーパスによる用例分析を行い、新開発コーパスの特性を明らかにするとともに、それが辞書編集にどのように役立つかを明らかにする。

国語辞書のどういう部分の記述にコーパスを活かすかという点では、主な分担課題を四つ設定して、それぞれを分担者が研究することにする。

２．辞書編集班の全体としての活動

辞書編集班では、分担者ごとに四つの小グループに分かれて、それぞれの研究を進めるとともに、班全体として研究進捗状況を確認しあい、今後の研究方向を考えるための会議を開催してきた。班会議は、2006 年度３回、2007 年度～2010 年度は４回開催した。2010 年度は、５月２４日、８月２３日、１１月１３日～１４日、２月２１日に開催した。このうち、１１月は、「コーパスと辞書」というタイトルの元、他班からの研究発表も含めて「拡大班会議」ということで開催した。

３．今年度の研究進捗状況

３．１　コロケーション辞書の概念設計と試作（荻野綱男・荻野孝野）

今年度は Google N-gram データからのコロケーション抽出を試みた。

３．１．１　コロケーションを抽出するための資料

コロケーションとして、当面「名詞＋助詞＋動詞」を考える。コロケーションの記述に際して、BCCWJ はデータ量が小さくて、これを資料としたのでは十分な記述ができない。WWW 内の日本語資料が、量的に膨大なので、使い物になるようだ（前年度までの結論）。

検索エンジンをそのまま利用するやり方では、用例が（当面は）1000 例までしか得られ

†ogino@chs.nihon-u.ac.jp

ない。1000 例中に含まれるコロケーションでも、特定のものに集中することがあるので、実例を広くカバーしているかどうか、心配である。

一方、WWW 内の日本語資料の大部分をダウンロードし、係り受けを解析し、動詞ごとに「名詞＋助詞」を整理したものとして「黒崎-河原データ」がある。こちらは (1) 頻度が付いている (2) かなり網羅している（ようだ） というメリットがある一方、 (1) 文脈がないので、意味が汲み取れないようなものがある (2) 単位認定の間違いなどがある（全部扱っているといえるか自信が持てない） (3) 機械的に係り受け解析をするための道具として作られているので、深い解析には向いていないというデメリットもある。

そこで、今年度は Google の n-gram データ（工藤拓、 賀沢秀人著、「Web 日本語 N グラム第１版」、言語資源協会発行）を利用することにした。

これは、添付文書によれば次のようなものである。「N グラムは一般に公開されている日本語の Web ページで Google がクロールしたものから抽出されている。ただし、閲覧に特別な認証が必要なページや、meta タグに noarchive,noindex 等が指定されているページは対象に入っていない。抽出対象となった文数は約 200 億文で、出現頻度 20 回以上の１～７グラムを収録している。」

３．１．２　Google の n-gram データの利用方法

パソコンでデータを利用するために、以下の３段階の事前準備を行った。

[1] UTF-8 からシフトＪＩＳにコード変換

ただし、変換できない一部文字が「?」になる。

[2] 改行コードの変換

行末には 16 進で 0A が入っている（LF だけということ）。これを CR+LF に置き換えた。

[3] 出現頻度の直前の「タブ」を「半角空白＋タブ」に置き換え

こうしておかないと、単語間の区切りが認識されず、ソートプログラムがうまく動かない。

３．１．３　7-gram からコロケーションを見ていく

ある程度の文脈がないと、意味がわからず、コロケーションが抽出できないと考え、7-gram のデータでコロケーションを見てみた。検索エンジンによる方法よりも幅広いデータの抽出が可能であった。

しかし、一方では、7-gram データでコロケーションがうまく抽出できない場合もあった。たとえば「を担ぐ」を指定した場合 92 例しか得られない。以下では、５例だけを示す。

貞清 氏（57）を 担ぐ	27
で 神輿（ みこし ）を 担ぐ	33
（ 元 または 験 ）を 担ぐ	30
と 民主党（ 日 ）を 担ぐ	21
ある 「 本社 神輿 」 を 担ぐ	35

単純なコロケーション（長い文脈が不要で、「名詞＋助詞＋動詞」だけで十分）の場合は、7-gram だとかえって文脈がばらついて、それぞれの頻度が 20 未満になり、コロケーション抽出がうまくいかないと考えられる。つまり、7-gram が（文脈が長いという意味で）記述に適しているわけではないということがわかった。

３．１．４　3-gram データでコロケーションを見ていく

大量のコロケーションの整理を行うことができた。「～を動詞」の場合の「～」の部分にくる名詞として、7-gram よりもはるかに大きなバラエティが得られ、辞書記述の資料として適しているように思えた。

しかし、3-gram では、うまくコロケーション記述ができない場合がある。たとえば、Google の単語の認識のミスなどにより、単語の一部が独立した単語のように抽出されている例がある。５例ほど示す。

い を 倒す　　20　「ばたふらい」などの途中
う を 倒す　　28　「してんのう」などの途中
お を 倒す　　53　「ただお」などの途中
か を 倒す　　1165　「いづれか」などの途中
き を 倒す　　39　「ばっかんき」などの途中

これらは、検索エンジンで確認すると、ミスの理由がわかるが、辞書記述のためのデータとしては問題である。

また、複数形だけなど名詞や数詞がないので意味をなさない場合もある。こちらも５例ほど示す。

がた を 倒す　21
くらい を 倒す　22
くん を 倒す　177
さま を 倒す　215
さん を 倒す　2089

その他、文脈が短すぎて意味がうまく読み取れないようなものがかなりあった。

３．１．５　3-gram、4-gram、5-gram を混合した処理

7-gram では、文脈が長すぎて、十分な用例数が確保できない。

3-gram では、文脈が短すぎて、文脈がわからなくなる。単語の一部が切り出されてしまうことも問題である。

そこで、石川慎一郎氏のアイディアにより、3-gram、4-gram、5-gram を混合した処理を行うことにした。

これは、3-gram のデータの先頭に「## 」を付け、4-gram のデータの先頭に「# 」を付け、5-gram のデータと一緒にして文脈の終わりから初めにかけての「逆順ソート」を行う。すると、3-gram データが並ぶのが基本となるが、高頻度の 4-gram や 5-gram がその後に続けて並ぶ形になる。つまり、3-gram では文脈が不足しがちなところを 4-gram や 5-gram で補うことができる。

たとえば、

パイプ を 挟む　　103

という用例の直後に

燃料 パイプ を 挟む　20

があり、3-gram では、単に「パイプ」とあるだけで、たばこの道具かと思うが、そういうのもあるかもしれないが、ものを通す管のようなものという場合が多いことがわかる。

もう１例示そう。

具 を 挟む　　1051

# <S> 具 を 挟む	44
# 、 具 を 挟む	102
# から 具 を 挟む	28
て から 具 を 挟む	26
# が 具 を 挟む	77
ラーメン が 具 を 挟む	77
# て 具 を 挟む	65
# で 具 を 挟む	50
# な 具 を 挟む	75
好き な 具 を 挟む	61
# に 具 を 挟む	242
パン に 具 を 挟む	88
間 に 具 を 挟む	59
中 に 具 を 挟む	25

というわけで、3-gram だけでは「具 を 挟む」だけで、意味がわかりにくいが、直後にさまざまな 4-gram や 5-gram が位置するので、意味の誤解は少ない。

試行錯誤の結果、どうも、この方法が優れているようだということになった。

３．１．６　n-gram データの整理結果

n-gram データから抽出したコロケーション情報については、成果 DVD 中に資料を含めた。これらは検討途中の生データに過ぎないが、3-gram から 5-gram を一緒に扱うことで、適当な量の、かつ適当な長さの文脈を集めることができ、かなり能率がよいことがわかった。

コロケーション抽出は、WWWのような大規模データを使わないとうまくいかないように思われる。

とはいえ、このような整理は基本的に人手による作業なので、時間と人件費がかかってしまい、思うような記述の進展はむずかしかった。

十分な検討ができる資料はあるのだが、検討に時間がかかるということである。

コロケーションの記述には、言語データの整備もさることながら、その先の人手による整理・チェックにかなりの手間がかかることがわかった。その意味で、実用的なコロケーションの記述はきわめてむずかしいものである。

なお、以上の研究と並行して、格助詞パターンの試験的記述も行っている。こちらは、2010 年度にデータの範囲を広げ、「雑誌、教科書、読み物、新聞」について作業を行ない、データ種別と格助詞パターンの関係を検討したものである。（協力者＝荻野孝野）

３．２　日本語複合辞の研究（近藤泰弘・山元啓史・坂野収・多田知子・岡田純子）

本年度、複合辞グループは、5 年間のまとめとして、「BCCWJ 複合辞辞書」を完成した。本グループでは、BCCWJ を辞書作成の立場から評価するという辞書編集班の目標を達成するため、当初から「文法的辞書」を作成するという目標を掲げた。そして、当面の作業として適切だと思われた BCCWJ におけるすべての複合辞のリストを作成することを試みた。複合辞研究は近年非常に盛んになっており、多くの先駆的な業績がある。その中において、

新たにリストを作ることは、屋上屋を架すことにならないかとも危惧されたが、従来の複合辞リストを改めて見直し、用例から帰納してリストを作成することによって、新規の複合辞を多く収めたものを作ることができた。今回作成したものを「BCCWJ 複合辞辞書」（Ver.1.0）と呼称する。その作業はすべて辞書編集班・複合辞グループに属する、近藤・坂野・多田・岡田・山元の 5 名が行った。なお、青山学院大学院の近藤の 2006 年度から 2010 年度の演習参加者からは有益なアドバイスを得た。

「BCCWJ 複合辞辞書」は、エクセル版および PDF による印刷版として作成した。その内容は、表 1 に記したようになっており、文法機能、意味範疇などの基本的素性から、前接語などの文法情報、そして BCCWJ 内の用例、先行文献での有無などを知ることができる。ただし、印刷版は現在のところ簡略版であり、見出しと基本素性および用例のみを収載している。

また、本作業を行うにあたって様々な言語学的な問題が発生したが、それらについては、複合辞研究会などで発表を行った他、学術論文として刊行したものもある。

さらには、本辞書を用いてどのような研究が可能であるかという、複合辞辞書の評価という問題にも取り組み、日本語教育の観点からの問題点を摘出した。

今回は BCCWJ2008 および 2009 を用いて複合辞辞書を作成したが、BCCWJ が完成した時にはそれによって再度リストを訂正することを考えている。それはバージョン 1.1 ということになるだろう。また、今回はエクセル版、印刷版の二種類を作成したが、電子書籍としての PDF 版なども作成を予定している。iPad や SONYReader などで利用しやすい形態を考えたい。

また、複合辞辞書を作成するにあたっては、従来の品詞分類の限界を痛感した。単純辞機能語の場合も含め、今後、品詞分類の再考を行う必要があると考えている。複合辞リストを利用した自然言語処理的な研究も必要である。さらには、単純辞機能語を補充することで、BCCWJ 機能語辞書を作成することも必要かもしれない。これについては、他の班の作業に協力して行うことも可能だろう。様々な課題はあるものの、本年度は、5 年間のまとめとして「BCCWJ 複合辞辞書」を作成することができたことは成果であった。

表１＜「複合辞辞書」記載内容＞

記載項目	記載内容
項番	「小見出し」対応の複合辞番号。全部で 925 項番。順序は昇順（あ・い・う・え・お順）。
大見出し	表現形式（言葉）が異なる複合辞
中見出し	「大見出し」に、係り/副・助詞が単純に挿入された辞や、同一表現ながら文法機能が異なるものを、下位の「中見出し」とした。
小見出し	「大見出し」と同一表現、同一文法機能でも、意味範疇の異なるものを「小見出し」として分類した。それに、「中見出し」項目を加えて、表現/文法機能/意味範疇が異なるものを、すべて「小見出し」で記載。
構成組成	「小見出し」表現を形態素に分解したもの。
意味範疇	複合辞の意味・機能

前接		複合辞の前にくる文法形態
文法機能		複合辞の相当品詞
解説・用法		意味範疇や用法の補足説明など必要により記入
備考		当該複合辞の関連表現など、必要により記入
機能的用法の用例	用例1	「BK（書籍）」からの引用
	用例2	「PM（雑誌）・PN（新聞）・WR（白書）」のいずれかからの引用
	用例3	「YC（Yahoo! 知恵袋）・YB（Yahoo! ブログ）・MD（国会会議録）・TB（検定教科書）」のいずれかからの引用
	用例4	上記3分類の各々から引用することを原則としたが、見つからない場合（用例「なし」）、他分類から用例を追加
内容的用法	用例	「辞（機能語）」ではなく「内容語」としての用例
先行文献（左記文献に記載あれば〇印）		森田・松木（1989）『日本語表現文型』（アルク）
		国研（山崎・他）（2001）『現代複合辞用例集』（国立国語研究所）
		グループ・ジャマシイ（1998）『日本語文型辞典』（くろしお出版）
		松吉・佐藤（2008）『日本語機能表現辞典「つつじ」』（HTML 版）

３．３　コーパスを利用した国語辞典の記述内容の検証（矢澤真人・橋本修・楊ソルラン）

３．３．１　グループの研究目的

現在刊行されている国語辞典の記述内容の妥当性について、コーパスを利用して検証し、適切な辞書記述について考える。

本年度は、国語辞典における動詞の自他の認定に関する問題と辞書の記述内容の検証に関わる研究、および辞書の記述と社会的規範性に関わる研究を進めた。

前者については、従来記述が不十分であった「ヲ」格と「ノコトヲ」格との対比的研究を行うとともに、国語辞典において記述がかなり揺れる漢語サ変動詞の自他について、「自他両用漢語動詞辞典」の完成を図った。

後者に関しては、和語の形式化と漢字表記の実態に関わる研究を進めた。

これらの研究成果については、班内会議・公開発表会で発表するとともに、講演会やシンポジウムなどを通じて、広く一般への公開を図った。

矢澤真人「検証辞書・辞典　国語辞典の情報を活用する」『月刊国語教育研究（日本国語教育学会），461 号，pp28-31，2010/09

口頭発表・講演

矢澤真人「国語辞典から日本語を考える」（つくば市立吾妻中学校平成２２年度「生きる力・未来講座」，2010/12/17），つくば市立吾妻中学校

矢澤真人「外形から引く日本語辞典への試み」（平成 22 年度筑波大学国際連携プロジェクト企画国際研究フォーラム「日本語学習辞書の開発と日本語研究」），2010/12/12，筑波大学

矢澤真人「日本語変換システムと国語辞典」（語彙・辞書研究会第３８回研究発表会「シンポジウム・コンピュータを用いた日本語研究における辞書の役割」），2010/11/20，新宿ＮＳビル南３０８会議室
矢澤真人「コーパスを利用した言語教育」（ 北京師範大学「中日の言語研究・言語教育」シンポジウム・パネルディスカッション「コーパスを利用した言語研究・言語教育」），2010/10/17，中国・北京師範大学
矢澤真人「言語変化と日本語教育」（韓国日語教育学会招請講演），2010/05/15，韓国建国大学校

３．４　辞書記述のためのコーパス利用（丸山直子・星野和子）

３．４．１　グループの研究目的

コーパスを辞書記述に役立てる方法や、辞書記述に役立つコーパスの性質の検討を行う。

語義記述（語釈）、例文掲載、その他（類義語・反義語の記述ほか）辞書記述に必要な項目を洗い出し、実際にコーパスをもとに辞書記述を行うことで、どの項目にコーパスを利用することができるかを検討する。さらに、辞書記述に役に立つコーパスとはどういうものかについても検討する。

当グループは、動詞の格情報と、オノマトペの意味・用法を中心に、コーパスを利用した辞書記述を試みてきた。動詞の格情報については東京女子大学の丸山直子、オノマトペについては元駒澤女子大学の星野和子が担当して分析を進めた。

３．４．２　動詞の格情報について

動詞の格情報について、その格に立つ名詞、表層格、深層格、動詞の語義のかかわりを、六つの型に整理し、それぞれについて、該当する動詞の一覧を示し、代表的な語について、辞書の記述（『岩波国語辞典』『新明解国語辞典』『日本語新辞典』『明鏡国語辞典』）とコーパス（ＢＣＣＷＪ、新潮文庫１００冊、朝日新聞 2005）における現れ方を調査した。以下２０語について、コーパス情報付きの辞書を試作した。

（Ａ）同じ名詞に複数の（異なる）格助詞がつくもの
　（Ａ－１）両者を殆ど同じように使うことができる（意味関係（深層格）が同じあるいは類似の）もの
　　ニ／デ‥‥「驚く」「しびれる」／「満ちる」／「勝つ」「つまずく」／「終わる」
　　ニ／ヲ‥‥「欠席する」「信頼する」「納得する」
　　ニ／トシテ‥‥「迎える」「雇う」
　（Ａ－２）意味関係（深層格）が異なるもの‥‥「置く」「掲げる」
（Ｂ）格助詞は同じ（つまり表層格は同じ）だが、意味関係（深層格）が異なるもの
　（Ｂ－１）格助詞の前に来る名詞が同じもの
　　（Ｂ－１ａ）動詞の語義が異なり、名詞と動詞の意味関係が異なるもの
　　（Ｂ－１ｂ）動詞の語義は同じだが、名詞と動詞の意味関係が異なるもの
　（Ｂ－２）格助詞の前に来る名詞が異なるもの

　　（Ｂ－２ａ）動詞の語義が異なり、名詞と動詞の意味関係が異なるもの
　　ヲ‥‥「カバーする」「蹴る」「振る」
　　ニ‥‥「はまる」

ガ‥‥「ひらめく」
（Ｂ－２ｂ）動詞の語義は同じだが、名詞と動詞の意味関係が異なるもの
「推薦する」「当選する」

３．４．３　オノマトペについて

2010 年度はハ行のオノマトペに関し、先ず、二種類の小型国語辞典（『明鏡国語辞典』『岩波国語辞典第七版』）の見出し語と記述された語義を比較した。二つの辞書の見出し語は必ずしも同じではなく、語の採用の仕方にその辞書の特徴の一つが現われていると思われた。語義記述の仕方にも辞書の違いが現われている。しかし、どの語義を第一番目に置くかということに関してはどちらの辞書にも統一的な見解がないように思われた。

次に、BCCWJ の LB で二つの辞書が採用した語と採用してはいないが関連する語を検索し、LB の用例数と二つの辞書の見出し語との関連を見た。見出し語は LB の用例数の多少と全く関連しないというのが結論である。『新潮文庫の百冊』『青空文庫』にはある語が LB にはない、あるいは、その逆ということも認められた。従って、辞書は各自がそれぞれ独自にコーパスを持っており、それに基づいて何らかの観点から語の採用や語義記述を行っていると結論付けた。

上記の考察をもとに、主に LB の用例によってハ行オノマトペ約 90 語の語義区分と記述を「ハ行オノマトペの語義区分試案」として試みた。取り上げた語は（1）二つの辞書が見出し語としているものと、(2) 辞書になくても LB で用例数の比較的多いものである。語義の記述に関しては、先ず、その語の指す主体と述語に着目して語義区分を行い、用例数が最多のものを第一義として最初に記し、以下、用例数順に記述した。この記述順ではオノマトペの語義の派生の仕方は不明である。また、具体的な事物についての表現と比喩表現をどう仕分けるかという問題も残る。指す主体に重点があるオノマトペと、修飾する述語に意味の重点があるオノマトペのあることも分ってきた。小型国語辞典におけるオノマトペの記述はどうあるべきか今後の課題として残っている。

BCCWJ の評価としては、基本的な動詞の格情報については、BCCWJ からある程度網羅的に用例を集めることができるが、オノマトペの方は、BCCWJ では十分に収集できないと言えそうである。

研究活動・成果の総括：言語処理班
代表性のあるコーパスを利用した日本語意味解析

奥村 学 (班長: 東京工業大学)[1]
白井 清昭 (分担者: 北陸先端科学技術大学院大学)
竹内 孔一 (分担者: 岡山大学)
新納浩幸 (分担者: 茨城大学)
佐々木稔 (分担者: 茨城大学)
中村 誠 (分担者: 北陸先端科学技術大学院大学)
高村大也 (分担者: 東京工業大学)
杉山一成 (協力者: 東京工業大学)
古宮嘉那子 (協力者: 東京農工大学)
九岡 佑介 (協力者: 北陸先端科学技術大学院大学)
田中 博貴 (協力者: 北陸先端科学技術大学院大学)
中西 隆一郎 (協力者: 北陸先端科学技術大学院大学)
高橋 秀幸 (協力者: 岡山大学)
小林 大介 (協力者: 岡山大学)

Final Progress Report: 'Natural Language Processing' Group

Manabu Okumura (Tokyo Institute of Technology)
Kiyoaki Shirai (Japan Advanced Institute of Science and Technology)
Koichi Takeuchi (Okayama University)
Hiroyuki Shinnou (Ibaraki University)
Minoru Sasaki (Ibaraki University)
Makoto Nakamura (Japan Advanced Institute of Science and Technology)
Hiroya Takamura (Tokyo Institute of Technology)
Kazunari Sugiyama (Tokyo Institute of Technology)
Kanako Komiya (Tokyo University of Agriculture and Technology)
Yûsuke Kuoka (Japan Advanced Institute of Science and Technology)
Hiroki Tanaka (Japan Advanced Institute of Science and Technology)
Ryûichirô Nakanishi (Japan Advanced Institute of Science and Technology)
Hideyuki Takahashi (Okayama University)
Daisuke Kobayashi (Okayama University)

1 研究目的

日本語を対象にした言語処理研究では，形態素解析，構文解析について研究が進み，高精度なツールの開発も行われてきており，それらのツールが日本語学，日本語教育など他の研究分野でも広く利用されるようになってきている．その一方で，意味解析については依然研究が遅れており，一般に利用可能なツールの開発レベルにまで解析精度が到達していない．また，代表性のあるコーパスを用いた言語処理研究は，これまでそのようなコーパスが存在しなかったため，日本語に関してはまったく行われてこなかったと言って良い．そこで本研究課題では，研究項目 A で構築する代表性のあるコーパスを用いた実証研究を行っている．具体的には，以下の 3 つを柱とした日本語意味解析手法の開発を行っている．

[1] oku@pi.titech.ac.jp

1) 機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築

2) 単語の新語義，新用法の自動発見手法の開発

3) 語彙概念構造に基づく動詞の意味構造の自動構築法の開発と語彙概念付与システムの開発

3 つの柱のうち，2) 単語の新語義，新用法の自動発見手法の開発については新たな研究計画を平成 21 年度より追加した．新たに追加する研究はコーパス中の特異な用例を検出する手法の開発である．単語の特異な用例は，その単語の使われ方を調査する上で有用である．また特異な用例を検出・排除することで，用例集を精度良く分析することが可能となる．またコーパス内の特異な用例の有無を調べることで，そのコーパスの一般性や特殊性も考察できる．2) では当初，コーパス中の単語の用例集合をクラスタリングし，同じ意味を持つクラスタを作成した上で新語義を発見する手法を構想していた．しかし，この手法では，一定量同じ意味の用例が出現するまでクラスタが構成できず，したがって，新語義を発見できないという問題点があった．そのため，上述した特異な用例検出手法により，ごく少数の特異な用例しか出現していない時点でも新語義を発見できる手法を開発することで，2) で当初構想していた手法を補完し，新語義発見手法の完成度を増すことを狙っている．

2 機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築

東京工業大学の研究グループでは以下の 4 つを柱に研究を進めてきた．

- 語義タグ付コーパスの構築,
- BCCWJ を用いた新しい語義曖昧性解消タスク,
- 半教師ありクラスタリング手法の開発と, 多義性解消への適用,
- 代表性のあるコーパスを用いた語義曖昧性解消.

以下順に概要を述べる．

2.1 語義タグ付コーパスの構築

データ班から公開されているコアデータに対して, 岩波国語辞典中の語義の区分に基づき, 人手で語義を付与する作業を行った．語義付与対象単語は, 岩波国語辞典中に見出し語がある単語で, かつ, 複数の語義を持ち, 品詞が, 名詞, 動詞, 形容詞, 副詞であるものとした．

過去のタグ付コーパス構築例にならい (白井, 2003), タグ付けの際, 辞典中に該当の語義が見当たらない場合「該当なし」という判断を許し, また, 最下層の語義のどれかでは判断できない場合, より上位のラベルを付与することを許している．さらに, 一部のデータについては, 複数の作業者による「ゆれ」の度合を計るため, 一致度の調査を行っている．1000 語分 (白書 600 語，Yahoo!知恵袋 400 語) について一致度を計算したところ，平均 Kappa 値が，白書では 0.704， Yahoo!知恵袋では 0.652 だった.「該当なし」の場合，大辞林をひき，該当する語釈文があれば，それを明記し，該当するものがなければ，作業者自身が考えた語釈文を記載してもらうようしている．

2.2 BCCWJ を用いた新しい語義曖昧性解消タスク

語義曖昧性解消に関する評価型ワークショップである Semeval-2 (`http://semeval2.fbk.eu/Semeval2.html`) に BCCWJ を用いた語義曖昧性解消の評価型タスクを提案し，採択された．国内外の合計 10 グループが参加表明をしていたが，2010 年の 3 月から 4 月にかけて行われた formal run では最終的に，領域内の 2 グループと領域外の 2 グループ (海外からの 1 グループを含む) が結果を提出した．

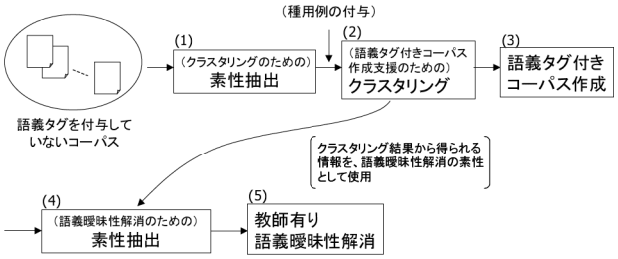


図 1: 語義曖昧性解消手法の構成

タスクの詳細，データ，評価手法，参加システムの概要，結果等については，task description paper である (Okumura et al., 2010) を参照していただきたい．また，タスクの web ページ (`http://lr-www.pi.titech.ac.jp/wsd.html`) も参照していただきたい．

このタスクの参加者も含め，現時点では合計で領域内 5 グループ，領域外の国内 4 グループ，領域外の海外 4 グループが，構築した語義タグ付コーパスを利用していることになる．

2.3 半教師ありクラスタリング手法の開発と，多義性解消への適用

我々は, クラスタリング時に, 教師情報を部分的に利用する, 半教師ありクラスタリング手法を開発している. 半教師ありクラスタリングでは, 用例対が同じ語義に対するものである, あるいは, 異なる語義に対するものである, ある用例がある語義に対するものである, 等の事実を既知のものとしてシステムに与えることで, より精度の高いクラスタリングを実現する.

語義タグ付けの支援において利用できるだけでなく, 半教師ありクラスタリングは, 以下の点においても利用できると我々は考えている.

- 新語義候補の検出,
- 多義性解消システムの性能改善.

従来のクラスタリング手法に比べ, 高精度のクラスタリングが実現できることから, より精度の高い新語義候補検出が期待できる. また, 用例のクラスタリング結果の情報を利用することで, より性能の良い多義性解消手法が実現できる.

具体的には，図 1 に示すように，(1) 語義タグを付与していないコーパスから，クラスタリングのための素性を抽出する．(2) 抽出した素性を用いて，クラスタリングを行う．このクラスタリング時に，種用例として語義タグが付与された用例を与え，複数の種用例と，その種用例間の関係を考慮した制約を導入し，重心の変動を抑えることに着目した半教師ありクラスタリングを適用する (Sugiyama and Okumura, 2007)．(3) 語義タグ付きコーパスの作成は，生成された各クラスタが類似する用例である情報を利用して，作業者が用例に語義タグを付与することで行う．(4) 語義タグを付与したコーパスから，語義曖昧性解消を行なうための素性を抽出するとともに，(2) のクラスタリング結果から得られる情報も素性として抽出し，(5) 教師ありの語義曖昧性解消を行う．詳細は杉山の報告 (Sugiyama and Okumura (2009)) を参照していただきたい．

2.4 代表性のあるコーパスを用いた語義曖昧性解消

代表性のあるコーパス中には, 複数のジャンルのテキストが混在していることになる. したがって, コーパスは, いくつかのジャンルごとのサブコーパスに分割できることになる. この時，単語によっては，サブコーパスごとに，出現する語義の頻度分布が異なる場合が存在する．すると, あるジャンルのテキスト中の用例を対象に語義曖昧性解消しようとする時，同一ジャンルのサブコーパスを学習に利用するのが良さそうであるのは言うまでもないが，それ以外に，コーパス中のどのサブコーパスをどのように学習に利用するのが良いのかは自明な問題ではない．これはある種の領域適応 (domain

adaptation) の問題であるが，これまでのように単一ジャンルのテキスト (たとえば，新聞データ) を利用していた場合にはさほど顕在化していない問題である．

複数のジャンルのテキストに対する語義タグ付コーパスが徐々に構築できてきており，昨年度語義曖昧性解消における領域適応に関する研究に着手し，ソース (適応元) データとターゲット (適応先) データの性質により，ソースデータとターゲットデータの組み合わせごとに効果的な領域適応手法が異なることが分かっている．そのため，ソースデータとターゲットデータの組み合わせごとに効果的な領域適応手法を自動的に選択する手法の開発を行っている．領域適応手法の自動選択は，ソースデータとターゲットデータの性質に関する情報を元に決定木学習を用いて行う．自動的に選択された領域適応手法を用いることで，語義曖昧性解消の性能が有意に向上することが確認されている．詳細は古宮の報告 (古宮嘉那子・奥村学 (2010)) を参照していただきたい．

3　コーパスからの新語義の発見

北陸先端科学技術大学院大学の研究グループでは，コーパスから単語の新しい意味・用法を発見する研究に取り組んできた．我々が提案する手法の概要を図 2 に示す．

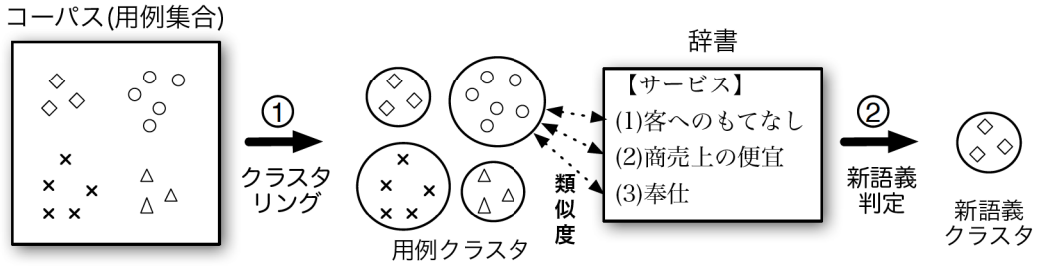


図 2: 提案手法の概要

まず，対象単語を含む用例をコーパスから収集し，同じ語義を持つ用例をまとめたクラスタを作成する (用例クラスタリング; 図 2 ①)．用例は以下の 4 種類の特徴ベクトルで表現する．対象語の直前または直後に現われる単語を素性とする隣接ベクトル，対象語の周辺に出現する単語を素性とする文脈ベクトル，対象語と二次共起 (間接共起) する単語を素性とする連想ベクトル，テキストのトピックを素性とするトピックベクトルである．クラスタリングの際には，これらの特徴ベクトルのいずれかの類似度が高い用例をまとめてクラスタを作成する．1 種類の特徴ベクトルを用いる先行研究 (Schütze, 1998) と比べ，複数の特徴ベクトルを同時に用いることで，語の類似性を様々な観点から評価できる点に提案手法の特徴がある (中西他, 2011)．

次に，用例クラスタが新語義を持つ用例の集合であるかを判定する (新語義判定; 図 2 ②)．クラスタ集合を $\mathcal{C} = \{C_1, \cdots, C_n\}$，辞書で定義されている語義の集合を $\mathcal{S} = \{S_1, \cdots, S_m, NS\}$ とし (NS は新語義を表わす)，用例クラスタと語義を対応付けるマッピング関数 $M : \mathcal{C} \rightarrow \mathcal{S}$ を決める．この際，(1) M で対応付けられたクラスタと辞書の語義の類似度が高く，(2) 辞書のどの語義とも類似度が低いクラスタは新語義 NS に対応させ，(3) 似ているクラスタは同じ語義に対応付ける場合に高くなるようなマッピング関数のスコアを定義し，それが最大となる M を 1 つ選択する．選択された M において NS に対応付けられたクラスタを新語義の用例を集めたクラスタとして出力する．

SemEval-2 日本語タスク (Okumura et al., 2010) のデータを用いて提案手法の評価実験を行った．用例クラスタリングについては，2 つ以上の用例をまとめたクラスタにおいて，同じ語義を持つ用例が占める割合の平均値 (AP) は 0.857 となり，比較的良好な結果が得られた．また，1 種類の特徴ベクトルを用いる手法に比べて AP の値が約 4%改善したことから，複数の特徴ベクトルを利用することの有効性を確認した．一方，新語義判定の F 値は 0.63 となり，まだ改善の余地が大きいことがわかった．ただし，ベースラインとして，用例クラスタと語義の類似度を計算し，それが閾値以下の場

合に新語義と判定する手法を試したところ，F 値は 0.54 であった．したがって，提案手法は単純に用例クラスタと語義の類似度を測って新語義か否かを判定する手法よりも優れていることがわかった．

4　特異用例の検出

茨城大学の研究グループが言語処理班で活動した平成 21 年度と平成 22 年度の研究成果を述べる．

基本的に「特異用例の検出」というタスクに取り組んできた．特異用例とは対象単語の少し変わった用法をもつ用例のことである．このような用例を検出することで，語義識別に対する質の高い訓練データを作成することができる．また特異用例はその対象単語の言語的性質を調べる際にも役立つ．さらに特異用例の存在の有無により，利用しているコーパスの均一性や代表性も評価できる．

アプローチとしては，特異用例を用例集合内の外れ値と見なし，データマイニング分野の外れ値検出の手法を利用することである．外れ値検出の手法は多岐にわたるが，大きく分類するとデータの生成に確率モデルを用いるものと用いないものに分けられる (山西健司, 2009)．確率モデルを用いた場合，データの生成確率が得られるので，その確率が低いデータを外れ値とすればよい．このアプローチでは，いかに適切な確率モデルを導入できるかが鍵となる．確率モデルを用いない手法としては Local Outlier Factor (LOF)(M. M. Breunig and H-P. Kriegel and R. T. Ng and J. Sander, 2000) や One Class SVM(Larry M. Manevitz and Malik Yousef, 2002) が代表的である．LOF は密度ベースの手法であり，概略，データの近傍の密度を利用することで，そのデータの外れ値の度合いを測り，その値によって外れ値を検出する．One Class SVM は ν–SVM を利用した外れ値検出の手法である．すべてのデータは $+1$ のクラスに属し，原点のみが -1 のクラスに属するとして，ν–SVM を使って 2 つのクラスを分離する超平面を求める．その結果，-1 のクラス側に属するデータを外れ値とする．

平成 21 年度は LOF と One Class SVM を組み合わせて特異用例を検出する手法を提案した (Shinnou and Sasaki, 2010)．概略，LOF の出力と One Class SVM の出力の交わりを出力とする手法である．コーパスを「白書」，対象単語を Semeval-2 で用いられた名詞に設定した実験では，提案手法により妥当な特異用例が検出できた．またコーパスの種類を変えることで，検出される特異用例が多様になることや，コーパスの規模を拡大することで，検出される特異用例が減少することなども確認できた．ただし特異用例は定義が曖昧であり，しかも特異用例かどうかを主観的に判断することも難しいため，手法の評価について課題が残った．

そこで平成 22 年度は，検出対象を特異用例の 1 つである新語義の用例に限定した．新語義の用例とは対象単語の語義が辞書に記載されている以外の語義として用いられている用例のことであり，単に新語義と呼ぶ場合もある．検出対象が新語義であっても，上記手法がそのまま利用できる．しかし新語義検出は，一般の外れ値検出とは異なり，教師データが利用できるという枠組みで考えた方が自然である．ここでの教師データとは一部の用例について，その対象単語の語義が付与されたものである．この教師データを利用して，検出の精度を高める手法を提案した (新納浩幸・佐々木稔, 2011)．

具体的には，提案手法は 2 つの検出手法からなる．第 1 の手法は LOF を教師付きの枠組みに拡張したものである．第 2 の手法は，教師データから語義識別の分類器を学習し，各データの語義を推定する．推定された語義のクラスターとデータとの距離関係から外れ値かどうかを判定する．提案手法では第 1 の手法により外れ値の候補を取り出し，第 2 の手法でその候補を選別する．

提案手法の有効性を確認するために，2 つの実験を行った．人工的に作ったデータに対するものと，SemEval-2 の Japanese WSD タスク (Okumura et al., 2010) のデータに対するものである．SemEval-2 の Japanese WSD タスクは通常の語義識別のタスクであるが，識別する語義の対象に新語義を含めている点に大きな特徴がある．このためこのタスクの訓練データを教師データとして利用して，テストデータから新語義を検出するという設定で実験が行える．2 つの実験を通して，外れ値検出に教師データを利用する効果が確認できた．また通常の語義識別を行い，識別の信頼度から新語義を判定するアプローチでは新語義の検出が困難であることも示した．

また平成 21 年度と 22 年度を通して，用例間の距離学習についての研究も平行して行った．クラスタリングや語義識別等のタスクの精度は，用例間の適切な距離測定に大きく依存する．特異用例の検出に関しても同様である．距離学習では，教師データを利用して用例間の距離をより適切に設定する．用例は特徴ベクトルで表現されるが，対象単語が同じ語義を持つ場合，それら特徴ベクトルを近づけ，異なる語義を持つものは離すように特徴ベクトルの移動処理を行った後に距離を測定する．タスクを語義識別に設定し，距離学習手法の Neighborhood Component Analysis (NCA)，Local Fisher Discriminant Analysis (LFDA) を試したが，前者は目的関数が局所解に収束し，後者は識別平面の形状は変化しないため距離学習を行う効果は現れなかった．そこで Large Margin Nearest Neighbor (LMNN)(Weinberger and Saul, 2009) を利用することを提案した (佐々木稔・新納浩幸, 2011)．LMNN は目的関数が大域解に収束することが保証され，座標軸変換ではなく，データ移動による距離学習を行うために，学習の効果が期待される．SemEval-2 の Japanese WSD タスクでは，従来の SVM による語義識別と比較して，LMNN を用いた場合に 0.7%の精度向上が行えた．また LMNN により移動された特徴ベクトル間の位置関係を調べることで，新語義を検出できる可能性がある．

5　同時クラスタリングを利用した動詞類義語獲得

5.1　はじめに

岡山大学の研究グループでは項構造レベルの動詞辞書を人手で構築するために大規模テキストから自動的に語義を抽出する手法の開発を行ってきた．動詞項構造辞書とは動詞の類語を概念としてシソーラス形式でまとめ，さらに動詞の使用例について項の意味役割まで付与した事例とリンクさせたデータである．既に人手による構築で，4425 語 (7473 語義) の動詞に対して例文付きで辞書を構築している (Takeuchi et al. (2010)) がさらなる拡張を行うために半自動でテキストから辞書知識を構築する手法について検討を行ってきた．その結果，動詞類語を獲得する手法を新たに提案することができた．また現状では意味役割まで付与した事例をテキストから獲得するためのツールおよび評価コーパスの開発までおこなっている．以下では主要な成果と現状についてまとめる．

5.2　動詞類義語獲得

動詞の類語をテキスト中の係り受け関係から獲得する手法として，同時クラスタリングがベクトルベースの動詞だけのクラスタリングに対して有効であることを明らかにした (Takeuchi and Takahashi (2009))．さらにその評価実験において新聞記事と日本語均衡コーパスを利用した場合の動詞類義語の獲得精度についても調べ，均衡コーパスの方がより幅広く動詞の類義語を獲得できることを示した．このことから均衡コーパスが辞書構築に有効であることを実験で示した．

さらに，クラスタリングに対して正解である動詞類義語を最初に与えることでよりよくクラスタが獲得できないかという課題について取り組み，同時共起クラスタリングについて半教師ありの枠組みの導入方法について検討をおこなった (竹内・高橋 (2010))．また，現在では他の同時クラスタリング方法として Weighted kernel k-means との比較も行っているが，現状では同時クラスタリングの精度が上回っている (竹内他 (2010))．

5.3　意味役割ラベル付与システムの構築

意味役割とは動詞の係り関係にある項の役割を整理した物で例えば「雇う」の動詞に対しては「太郎は (Agent) 彼女を (Theme) 税理士として (As) 雇った [所属]」のようになる．ここで () 内が意味役割ラベルを示し [] は動詞の語義を表す．こうした意味に踏み込むことで同じ語義を共有する「太郎が税理士を (As) 雇用した」の場合,「を格」であっても雇われる時の役割を示していることを獲得できる．これに関して現在，規則ベースの意味役割付与システムを構築している (竹内他 (2009))．さらに，動詞語義と意味役割の事例を付加したコーパスの構築を行っている (竹内・森本 (2009))．

5.4 まとめ

岡山大学研究グループによる成果および現状を簡単にまとめた．大規模テキストから動詞類義語を獲得するには同時共起クラスタリングが有効であることを示した．また，均衡コーパスが辞書構築において効果的であることを動詞類語獲得実験で示した．さらに意味役割ラベルを付与した事例の獲得について現在タグ付きコーパスを構築しながら規則ベースの付与システムの構築を進めている．

文献

Larry M. Manevitz and Malik Yousef (2002). "One-class SVMs for document classification," *Journal of Machine Learning Research*, Vol. 2, pp. 139–154.

M. M. Breunig and H-P. Kriegel and R. T. Ng and J. Sander (2000). "LOF: Identifying Density-Based Local Outliers," in *ACM SIGMOD 2000 International Conference on Management of Data*, pp. 93–104.

中西隆一郎、白井清昭、中村誠 (2011). 「複数の観点から定義された用例間類似度に基づく語義識別」，言語処理学会第 15 回年次大会発表論文集.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2010). "SemEval-2010 Task: Japanese WSD," in *Proceedings of SemEval-2010*, pp. 69–74.

Hinrich Schütze (1998). "Automatic Word Sense Discrimination," *Computational Linguistics*, Vol. 24, No. 1, pp. 97–123.

Hiroyuki Shinnou and Minoru Sasaki (2010). "Detection of Peculiar Examples using LOF and One Class SVM," in *LREC-2010*.

白井清昭 (2003). 「SENSEVAL-2 日本語辞書タスク」，自然言語処理，10 巻 3 号，pp.3–24.

Kazunari Sugiyama and Manabu Okumura (2007). "Personal Name Disambiguation in Web Search Results Based on a Semi-Supervised Clustering Approach," in *Proc. of the 10th International Conference on Asian Digital Libraries (ICADL'07)*, pp. 250–256.

Kazunari Sugiyama and Manabu Okumura (2009). "Semi-supervised Clustering for Word Instances and Its Effect on Word Sense Disambiguation," in *Proc. of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), Lecture Notes in Computer Science (LNCS), Springer-Verlag, Vol.5449*, pp. 266–279.

Koichi Takeuchi and Hideyuki Takahashi (2009). "Co-clustering with Recursive Elimination for Verb Synonym Extraction from Large Text Corpus," *IEICE Transactions on Information and Systems*, Vol. E92-D, pp. 2334–2340.

Koichi Takeuchi, Kentaro Inui, Nao Takeuchi, and Atsushi Fujita (2010). "A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings," in *The 8th Workshop on Asian Language Resources*, pp. 1–8.

Kilian Q. Weinberger and Lawrence K. Saul (2009). "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *The Journal of Machine Learning Research*, Vol. 10, pp. 207–244, June.

古宮嘉那子、奥村学 (2010). 「語義曖昧性解消のための領域適応手法の自動選択」，情報処理学会自然言語処理研究会，198-6.

佐々木稔、新納浩幸 (2011).「距離学習に基づく語義識別の性能分析」，言語処理学会第 17 回年次大会.

山西健司 (2009). データマイニングによる異常検知，共立出版.

新納浩幸、佐々木稔 (2011).「教師付き外れ値検出による新語義の発見」，言語処理学会第 17 回年次大会.

竹内孔一、高橋秀幸 (2010).「同時共起クラスタリングを利用した動詞辞書構築」，日本語平成 *21* 年度公開ワークショップ *2010* 年 *3* 月 *15-16* 日.

竹内孔一、森本真衣子 (2009).「動詞項構造シソーラスに基づく動詞語義ならびに意味役割付与データの構築」，% sl 電子情報通信学会言語理解とコミュニケーション研究会，pp.13–18.

竹内孔一、土山傑、守屋将人、森安祐樹 (2009).「類似した動作や状況を検索するための意味役割及び動詞語義付与システムの構築」，電子情報通信学会，言語理解とコミュニケーション研究会, NLC-2009-33，pp.1–6.

竹内孔一、高橋秀幸、小林大介 (2010).「グラフに基づくクラスタリングによる動詞類義語の獲得」，電子情報通信学会，言語理解とコミュニケーション研究会, NLC-2010-11，pp.13–18.

公募班研究活動・成果報告

3月15日（火）　10:00〜11:40

日本語機能表現班「大規模階層辞書を用いた日本語機能表現解析体系の研究」

▶宇津呂 武仁

作文支援システム班「バランス・コーパス利用による日本語作文支援システム『なつめ』の構築と評価」

▶仁科 喜久子

意見情報班「多様な文書ジャンルを対象とした意見分析コーパスの作成に関する研究」

▶関 洋平

日本語フレームネット班「BCCWJと意味フレームに基づく語彙・構文複合資源の構築」

▶小原 京子

研究活動・成果の総括：日本語機能表現班
大規模階層辞書を用いた日本語機能表現解析体系の研究

宇津呂武仁　(班長：筑波大学大学院システム情報工学研究科)*
鈴木敬文　(協力者：筑波大学大学院システム情報工学研究科)
島内蘭　(協力者：筑波大学大学院システム情報工学研究科)
阿部佑亮　(協力者：筑波大学大学院システム情報工学研究科)
松吉俊　(協力者：奈良先端科学技術大学院大学情報科学研究科)
土屋雅稔　(協力者：豊橋技術科学大学情報メディア基盤センター)

Final Progress Report:
'Japanese Functional Expressions' Group

Takehito Utsuro　(University of Tsukuba)
Takafumi Suzuki　(University of Tsukuba)
Ran Shimanouchi　(University of Tsukuba)
Yusuke Abe　(University of Tsukuba)
Suguru Matsuyoshi　(Nara Institute of Science and Technology)
Masatoshi Tsuchiya　(Toyohashi University of Technology)

1. はじめに

機能表現[1]とは,「にあたって」や「をめぐって」のように，2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現である．一方，この機能表現に対して，それと同一表記をとり，内容的な意味をもつ表現が存在することがある．例えば，文 (1) と文 (2) には「にあたって」という表記の表現が共通して現れている．

(1) 出発する にあたって，荷物をチェックした.

(2) ボールは壁 にあたって，跳ね返った.

文 (1) では，下線部はひとかたまりとなって,「機会が来たのに当面して」という機能的な意味で用いられている．それに対して，文 (2) では，下線部に含まれている動詞「あたる」は，動詞「あたる」本来の内容的な意味で用いられている．このような表現においては，機能的な意味で用いられている場合と，内容的な意味で用いられている場合とを識別する必要がある．

我々は，このような日本語機能表現の解析の課題に対して，これまでに，国立国語研「現代語複合辞用例集」 [国研 01] に収録されている 125 機能表現の異表記を展開した 300 表現について，新聞記事中の用例に対して機能的用法・内容的用法を判別した用例データベース [土屋 06] を作成・公開した．また，機能的・内容的用法の自動判別ツールを作成し，係り受け解析ツールとの統合により，複合辞としての機能的用法を考慮した係り受け解析を実現した [注連 07]．また，日本語機能表現の全表記を網羅した辞書として，日本語機能表現の全表記約 17,000 を網羅的に収録した「つつじ」 [松吉 07, 松吉 08][2] が公開されたのを受けて，17,000 表現全てを対象とした機能的・内容的用法の判定方式を提案した [長坂 08].

*utsuro @ iit.tsukuba.ac.jp

[1]機能表現は，複数形態素からなる複合辞と一つの形態素からなる機能語から構成されるが，本稿では，複合辞と同等の意味で機能表現という用語を用いる．

[2]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 機能表現辞書の 9 つの階層

階層		分類数	表現数		
			合計 (L^9 表現数)	助動詞型以外	助動詞型
L^1	見出し語	—	341 (488)	281	207
L^2	意味	45/128/199	435 (488)	281	207
L^3	派生 (格助詞型，接続助詞型，連体助詞型，接続詞型，助動詞型，形式名詞型，とりたて詞型，提題助詞型)	8	555	348	207
L^4	機能語の交替	—	774	492	282
L^5	音韻的変化	38	1,187	633	554
L^6	とりたて詞の挿入	18	1,810	659	1151
L^7	活用	—	6,870	659	6211
L^8	「です/ます」の有無	2	9,722	895	8827
L^9	表記のゆれ	—	16,801	1360	15411

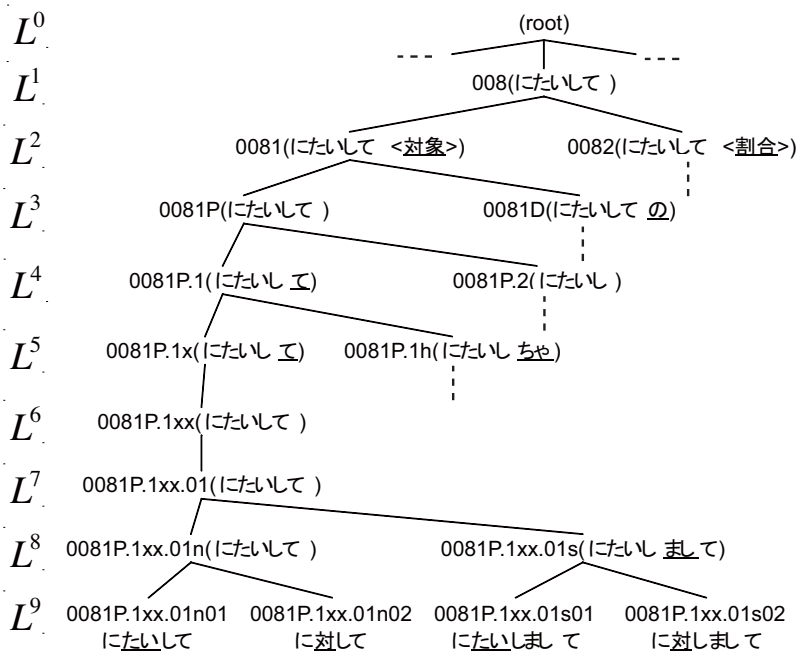


図 1: 機能表現辞書階層構造の一部

本研究では，この提案方式をふまえて，日本語機能表現の全表記約 17,000 を網羅的に収録した辞書「つつじ」の階層的構造および言語学的特性を活用して，網羅的な日本語機能表現の解析，および，日本語機能表現の集約的翻訳の枠組みを実現した.

2. 日本語機能表現一覧「つつじ」

代表的な機能表現の規模を超えて機能表現の表記を網羅的に列挙した辞書を設計・編纂することを目的として，日本語機能表現一覧「つつじ」 [松吉 07] が編纂された.「機能表現一覧」においては，日本語における機能表現の表記を網羅することを目的として，機能表現の構成要素の組み合わせとして，機能表現の異形を階層的に収録している．表 1 および図 1 に示すように，全体としては，形態に基づいて，全機能表現の表記の集合が 9 つの階層構造によって構成されている．階層の上位に

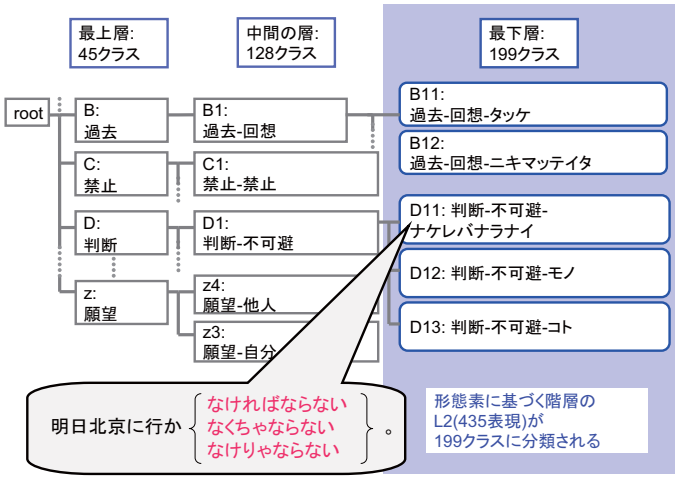


図 2: 日本語機能表現一覧「つつじ」: 意味的等価クラスの一部

は，341 種類の機能表現を見出し語として配置し，意味の違い，機能表現末尾の活用，機能表現の各構成要素の音韻的変化，とりたて詞の挿入，口語表現・敬語表現の言い換えなどによる異形として，16,801 表現が収録されている．また，機能表現の意味的な分類は，図 2 に示す 3 階層の体系によって構成されている [松吉 08]．この階層の最下層に位置する全 199 個の各意味的等価クラスに属する機能表現は，一定の文脈のもとで言い換え可能であるとされている．また，機能表現の文体については，常体，敬体，口語体，堅い文体の 4 種類の文体を区別して，各表現に付与している．

3. 派生関係及び用例を利用した日本語機能表現の解析

3.1 概要

[鈴木 10, 鈴木 11] において,「つつじ」の階層性を利用し，階層において下位に位置する機能表現 (以下，派生的表現) について，用法が類似するより上位の表現 (以下，代表的表現) の用例を参照して，用法判定を行う方式を提案した．

階層の上位に位置する代表的表現は，L^4 階層相当の 1,000 表現程度の規模とする [長坂 08]．そして,「機能表現一覧」において，代表的表現を除く表現を派生的表現と定義する．ただし，代表的表現を選定する際には，以下の制約を課す．

- 機能表現の語頭の無声・有声の制約により前接する活用語の活用型が制限される場合は，この制限を保持する．
- 機能表現の仮名表記・漢字表記の違いを保持する．
- 助動詞型の機能表現の場合には，活用形を保持する．

この提案方式においては，前後の形態素の品詞が代表・派生間において不変の場合には，代表的表現と派生的表現の間で用法の傾向に相関がある，という点に注目する．さらに，前後の形態素品詞に加え，代表的表現と派生的表現の間で，機能表現の表記を構成する形態素列の品詞パターンの間に派生関係があるという特性を利用する．提案方式に基づいて，派生的表現の用法の分析を行った結果，代表的表現の表記の用法判定済み用例集合 (約 38,000 例) を参照して，派生的表現の表記の用法判定を行うことにより，80%以上の用例の用法を正しく判定できることが分かった．

3.2 派生的な表現の解析方式

以下では，代表的表現の表記の用法判定済用例集合 S_c^{tr} を参照して，派生的表現の表記の用法判定を行う方式について述べる．

3.2.1 機能表現表記照合個所の表現形式

まず，一文中で，機能表現表記と文字列照合する個所を $e = \langle f, l, r \rangle$ (ただし，f は機能表現表記，l は機能表現表記の先頭の文字位置，r は末尾の文字位置) によって表現する[3]．このとき，評価用の文において機能表現表記 f_{ts} と照合した個所を $e_{ts} = \langle f_{ts}, l_{ts}, r_{ts} \rangle$ とし，e_{ts} に前接する形態素を m_{-1}^{ts}，後接する形態素を m_{+1}^{ts} とする．一般には，f_{ts} の可能性としては，派生的表現の表記 f_d の場合，および，代表的表現の表記 f_c の場合の二通りが考えられる．ここで，f_{ts} が派生的表現 f_d の場合には，f_d の代表的表現 f'_c の用例が，用法判定済用例集合 S_c^{tr} 中の機能表現表記照合個所の一つ $e_{tr} = \langle f'_c, l_{tr}, r_{tr} \rangle$ となる．一方，f_{ts} が代表的表現 f_c の場合には，f_c 自身の用例が，用法判定済用例集合 S_c^{tr} 中の機能表現表記照合個所の一つ $e_{tr} = \langle f_c, l_{tr}, r_{tr} \rangle$ となる．いずれの場合も，e_{tr} に前接する形態素を m_{-1}^{tr}，後接する形態素を m_{+1}^{tr} とする．

ここで，次節の解析手順においては，評価用の文における用法判定対象個所の単位として，相互に重複して連続する複数の機能表現表記から構成される列をひとまとめとして，機能表現表記列の用法判定を一括して行う．具体的には，評価用の文において，連続する 2 個の機能表現表記の文字列のうちの少なくとも一部が重複するような機能表現表記列 $E = e_i, \ldots, e_k$ (すなわち，機能表現表記列 $E = e_i, \ldots, e_k$ 中における連続する任意の 2 個の機能表現表記の組 $e_j,\ e_{j+1}$ において表記の文字列の少なくとも一部が重複する: $l_j < l_{j+1} < r_j < r_{j+1}$) をひとまとめとする．

3.2.2 解析手順

まず，評価用の文における用法判定の単位である機能表現表記列 $E = e_i, \ldots, e_k$ に対して，以下の条件**「前後形態素が類似する用法判定済用例の存在」**の成否を判定する．

> **「前後形態素が類似する用法判定済用例の存在」**
> $E = e_i, \ldots, e_k$ 中で，少なくとも一つの機能表現表記照合個所 $e_{ts} = \langle f_{ts}, l_{ts}, r_{ts} \rangle$ に対して，機能表現表記 f_{ts} に対応する機能表現表記照合個所 e_{tr} が用法判定済用例集合 S_c^{tr} 中に存在する．さらに，前接形態素 m_{-1}^{ts} と m_{-1}^{tr}，および，後接形態素 m_{+1}^{ts} と m_{+1}^{tr} の間で，それぞれ，品詞大分類[4] が一致する．

そして，この成否に応じて，下記の手順 (I) もしくは (II) を行う．

(I) **「前後形態素が類似する用法判定済用例の存在」**が成り立たない場合，機能表現表記列 $E = e_i, \ldots, e_k$ 中の全ての機能表現表記が内容的用法であると判定して終了する．

(II) **「前後形態素が類似する用法判定済用例の存在」**が成り立つ場合，以下を行う．

(II-i) 条件「機能表現表記列 E において，最長の表記となる照合個所 e_{ts} がただ一つである．さらに，e_{ts} に対して，用法判定済用例集合 S_c^{tr} 中の対応する機能表現表記照合個所 e_{tr} (複数個所の場合もあり得る) を参照することにより，e_{tr} に対する用法判定結果 l_{tr} が一意に決まる.」が成り立つならば，機能表現表記列 E に対して,「e_{ts} の用法は l_{tr}，E 中のその他の機能表現表記の用法は内容的用法」を採用して終了する．その他の場合は，(II-ii) を行う．

[3]ただし，機能表現表記 f としては,「機能表現一覧」 [松吉 07] における一文字表記の機能語は除外する．
[4]IPAdic (`http://sourceforge.jp/projects/ipadic/`) を用いる．

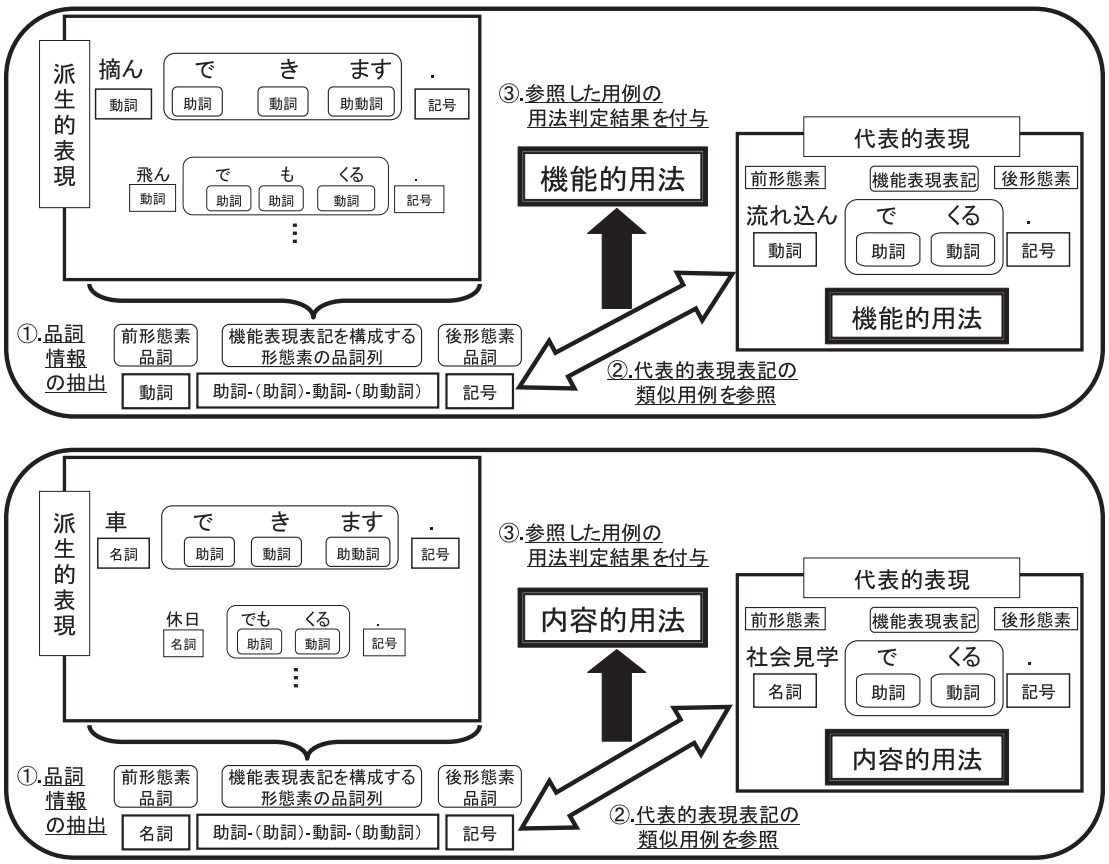


図 3: 模式図: 「代表的表現の表記の用例」を参照して「派生的表現の表記の用例」の用法を判定

(II-ii) 条件**「前後形態素が類似する用法判定済用例の存在」**において,「前後形態素の品詞大分類の一致」の代わりに「前後形態素の品詞細分類が一定以上の基準で類似する」を課し，(II-i) と同様の手順を行う．機能表現表記列 E に対する用法判定結果が一意に決まらない場合には，(II-iii) を行う．

(II-iii) 条件**「前後形態素が類似する用法判定済用例の存在」**において,「前後形態素の品詞大分類の一致」の代わりに「機能表現表記を構成する形態素の品詞列が一定以上の基準で類似する」を課し，(II-i) と同様の手順を行う．機能表現表記列 E に対する用法判定結果が一意に決まらない場合には,「不正解」と判定し終了する．

以上の手順にしたがって，派生的表現の表記の用法が機能的用法であると判定した例の模式図を図 3 上半分に，内容的用法であると判定した例の模式図を図 3 下半分に，それぞれ示す．

3.4 評価

代表的表現の表記の用法判定済用例としては，毎日新聞 1995 年の 1 年分から収集して人手で機能表現表記の用法判定を行った約 38,000 用例を参照することとする．評価対象としては，同じく毎日新聞 1995 年の 1 年分のうち，機能的用法と内容的用法として適度な割合で新聞記事内に出現する代表的表現に対して，用例数が 10 例以上となる派生的表現を中心に収集した 1,882 用例，及び，機能的用法に偏って新聞記事内に出現する代表的表現に対して，用例数が 50 例未満となる派生的表現を中心に収集した 916 用例の計 2,798 用例 (243 表現) を評価対象とする．

表 2: 派生関係及び用例を利用した機能表現の解析: 評価結果

(a) 代表的表現の用例を参照する手法

類型		割合 (%)	
「3.2.2 節の手順 (II)」前後の形態素の品詞もしくは機能表現表記を構成する形態素の品詞列の条件を満たす代表的表現の用法判定結果を採用し正解		71.6	
「3.2.2 節の手順 (I)」前後の形態素の品詞が一致する代表的表現が存在しないため，内容的用法と判定し正解		10.9	
不正解	適切な作例をすることにより正解可能	13.2	17.5
	作例しても正解不可能	4.3	
合計		100	

(b)「代表的表現の用例+左・右接続接続情報を参照する手法

類型		割合 (%)	
「3.2.2 節の手順 (II)」において用法判定済用例の一つとして左・右接続情報を追加して正解		77.0	
「3.2.2 節の手順 (I)」により正解		10.0	
不正解	適切な作例をすることにより正解可能	8.1	
	作例しても正解不可能	4.9	13.0
合計		100	

評価結果を表 2(a) に示す．また，3.2.2 節のいずれかの手順における判定結果が「不正解」となる場合について，代表的表現の適切な用例を作成して用法判定済用例集合 S_c^{tr} に追加した場合に，正解可能か否かの分析を行った結果も併せて示す．この結果から分かるように,「適切な用例の作例なしで正解」となる割合は約 82%，作例を許す場合は約 95%である．

また，表 2(b) には，用法判定済用例集合 S_c^{tr} に対して，用法判定済用例の一つとして，左・右接続情報 [松吉 07, 松吉 08] を追加した場合の評価結果を示す．左・右接続情報とは，機能表現表記の用法が機能的用法である場合の情報である．左接続情報は，直前に接続可能な形態素の情報を示しており，右接続情報[5]は，機能表現表記を構成する末尾の形態素の情報を示したものである．これらは「機能表現一覧」 [松吉 07] において，各機能表現ごとに定義されており，53 種類の左接続情報，および，51 種類の右接続情報が掲載されている．これらの左・右接続情報を追加した場合,「適切な用例の作例なしで正解」となる割合は，約 87%に改善する．

4. 日本語機能表現の集約的翻訳

日本語には 16,000 種類以上の機能表現の異形が存在する．従来の機械翻訳ソフトは，日本語機能表現の異形に対して個別に訳語を割り当てる手法を用いていると考えられるが，この手法では全ての異形を網羅することが困難である．そのため，日本語入力文中に，翻訳規則が未定義の機能表現の異形が存在した場合に，その表現を正しく翻訳できないという問題を抱えていた．そこで，本研究では，日本語機能表現の異形を網羅的に機械翻訳するために，類似する意味を持つ日本語機能表現を予め 1 つのクラスにまとめ，各クラスに対して 1 つの集約的な翻訳規則を作成する手法を提案した．機能表現の意味クラスとしては,「つつじ」の意味的等価クラス (199 クラス) を用いた．

以上の考え方に基づき，[坂本 09b, Sakamoto09a] では，日本語学習者向けの機能表現用例集 [グループ・ジャマシイ 98]，及び新聞記事テキストから，各意味的等価クラスに含まれる機能表現が出現した例文が十分な数収集できた 91 クラスを対象として，それらの機能表現の集約的英訳可能性を検証した．その結果，49 クラスについては，1 クラスに対して 1 規則で英訳可能となったが，その他の

[5]右接続情報に加えて，IPAdic を用いて形態素解析を行った場合の形態素列の情報を参照することにより，機能表現表記の直後に接続可能な形態素の情報が得られる．

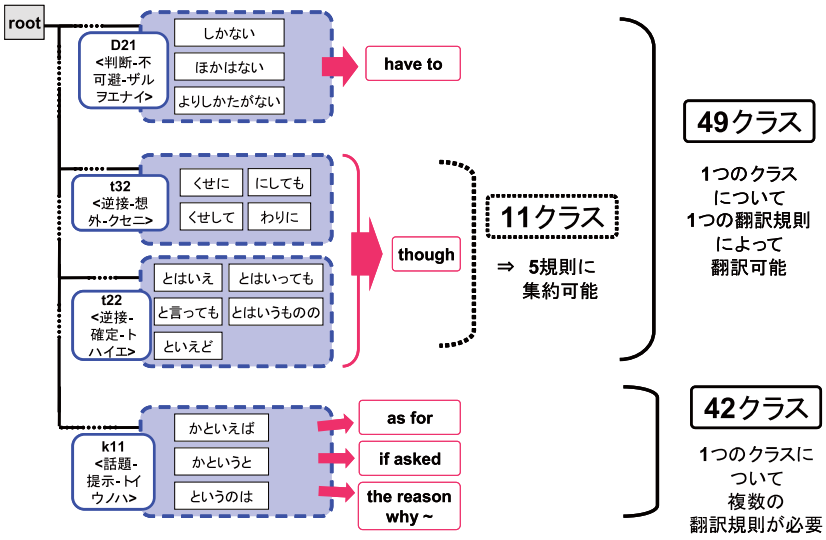


図 4: 集約的英訳可能性に基づく意味的等価クラスの粒度の再編

42 クラスについては，1 クラスに対して複数の英訳規則が必要であることが明らかになった (図 4). 同様に，[劉 10] においては，日本語学習者向けの機能表現用例集 [グループ・ジャマシイ 98] の中国語訳を日中対訳コーパスとして利用し，日本語機能表現の集約的中国語訳規則の作成・評価を行った. 一方，[Nagasaka10, 島内 11] では，日英対訳特許文を対象として，日本語機能表現の集約的英訳規則の作成および評価を行った．この研究では，NTCIR-7 の特許翻訳タスクで配布された 1,798,571 件の日英対訳特許文対に対して統計的機械翻訳モデルを適用することによりフレーズテーブルを学習し，日英対訳機能表現対を獲得するために用いた．特許文の場合は，使用される機能表現の意味範囲が狭く，その種類も少ないので，翻訳規則作成が容易である点が大きな利点となる．対象として，「つつじ」の 199 意味的等価クラスの中で，91 意味的等価クラスに属する日本語機能表現について，翻訳規則を作成し，その中の意味的等価クラス 12 個に属する日本語機能表現について評価を行なった結果，96.6%の正解率を得ることが出来た.

5. おわりに

本研究では,「機能表現一覧」の階層性を利用し，階層において下位に位置する派生的表現について，用法が類似するより上位の代表的表現の用例を参照して，用法判定を行う手法を実現した．また,「つつじ」の意味的等価クラスを利用して，日本語機能表現の集約的翻訳を実現した．今後の課題としては，以下が挙げられる．まず，提案方式では，代表的表現の用例，派生的表現の用例のいずれについても，できるだけ多くの用法の用例を収集し，用法判定結果を付与した用例集合を蓄積することが性能改善の鍵を握る．そこでは，大規模な未解析テキストコーパスを情報源として，機能表現表記の前後の形態素の品詞のバリエーションをできるだけ多く収集し，サンプリングして用法判定結果を付与することが最も効果的である．また，新聞記事を対象として構築した代表的表現の表記の用法判定済み用例集合を参照して，新聞記事以外の多様なジャンルのテキスト中の機能表現表記の用法判定を行うタスクにおいて，提案方式の有効性を評価する必要がある．さらに，機能表現を考慮した文解析を高度化する目的においては，機能表現の検出・係り受け解析と格構造解析を統合する方式を確立することが必要である．情報抽出・テキストマイニング・評判抽出・質問応答・含意認識等の応用の観点からは，機能表現が担う多様なアスペクト・モダリティの同定が不可欠であり，これまでの研究成果 (例えば，[江口 10]) をふまえて，多方面の応用における発展が期待される．一方，日本語機能表現の集約的翻訳を組み込んだ機械翻訳手法を実現するためには，多義性を持った機

能表現の意味的曖昧性を解消する方式の確立が不可欠である．

参考文献

[グループ・ジャマシイ 98] グループ・ジャマシイ（編）：教師と学習者のための日本語文型辞典, くろしお出版 (1998).

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[劉 10] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊：意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析, 言語処理学会第 16 回年次大会論文集, pp. 194–197 (2010).

[松吉 07] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[江口 10] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治：モダリティ，真偽情報，価値情報を統合した拡張モダリティ解析, 言語処理学会第 16 回年次大会論文集, pp. 852–855 (2010).

[長坂 08] 長坂泰治, 宇津呂武仁, 土屋雅稔：大規模日本語機能表現辞書の階層性を利用した機能表現検出, 言語処理学会第 14 回年次大会論文集, pp. 837–840 (2008).

[Nagasaka10] Nagasaka, T., Shimanouchi, R., Sakamoto, A., Suzuki, T., Morishita, Y., Utsuro, T. and Matsuyoshi, S.: Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Translation Rule Acquisition from Parallel Patent Sentences, *Proc. 7th LREC*, pp. 1778–1785 (2010).

[Sakamoto09a] Sakamoto, A., Nagasaka, T., Utsuro, T. and Matsuyoshi, S.: Identifying and Utilizing the Class of Monosemous Japanese Functional Expressions in Machine Translation, *Proc. 23rd PACLIC*, pp. 803–810 (2009).

[坂本 09b] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集, pp. 654–657 (2009).

[島内 11] 島内蘭, 阿部佑亮, 鈴木敬文, 宇津呂武仁, 松吉俊：特許文における日本語機能表現の集約的英訳規則の作成と評価, 言語処理学会第 17 回年次大会論文集 (2011).

[注連 07] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史：日本語機能表現の自動検出と統計的係り受け解析への応用, 自然言語処理, Vol. 14, No. 5, pp. 167–197 (2007).

[鈴木 10] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係を利用した日本語機能表現の解析, 情報処理学会研究報告, Vol. 2010, No. (2010–NL–199) (2010).

[鈴木 11] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係および用例を利用した日本語機能表現の解析, 言語処理学会第 17 回年次大会論文集 (2011).

[土屋 06] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一：日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).

研究活動・成果の総括：作文支援システム班
バランス・コーパス利用による日本語作文支援システム「なつめ」の構築と評価

仁科喜久子　（班長：東京工業大学留学生センター）†
村岡貴子　（連携研究者：大阪大学留学生センター）
因京子　（連携研究者：日本赤十字九州国際看護大学）
Joyce Terence Andrew（連携研究者：多摩大学 グローバルスタディーズ学部）
鎌田美千子　（連携研究者：宇都宮大学留学生センター）
阿辺川武　（連携研究者：国立情報学研究所連想情報学研究開発センター）

Final Progress Report: 'Writing System Support' Group

Kikuko Nishina　(International Student Center, Tokyo Institute of Technology)
Takako Muraoka　(International Student Center, Osaka University)
Kyoko Chinami　(The Japanese Red Cross Kyushu International College of Nursing)
Joyce Terence Andrew　(School of Global Studies, Tama University)
Michiko Kamada　(International Student Center, Utsunomiya University)
Takeshi Abekawa　(Research and Development Center, National Institute of Informatics)

1．最終年度の目標

研究の最終年度に当たり BCCWJ を利用した日本語作文支援システムを構築し、それとともに BCCWJ の評価をすることを最終目標とした。「作文支援システム班」は公募班として

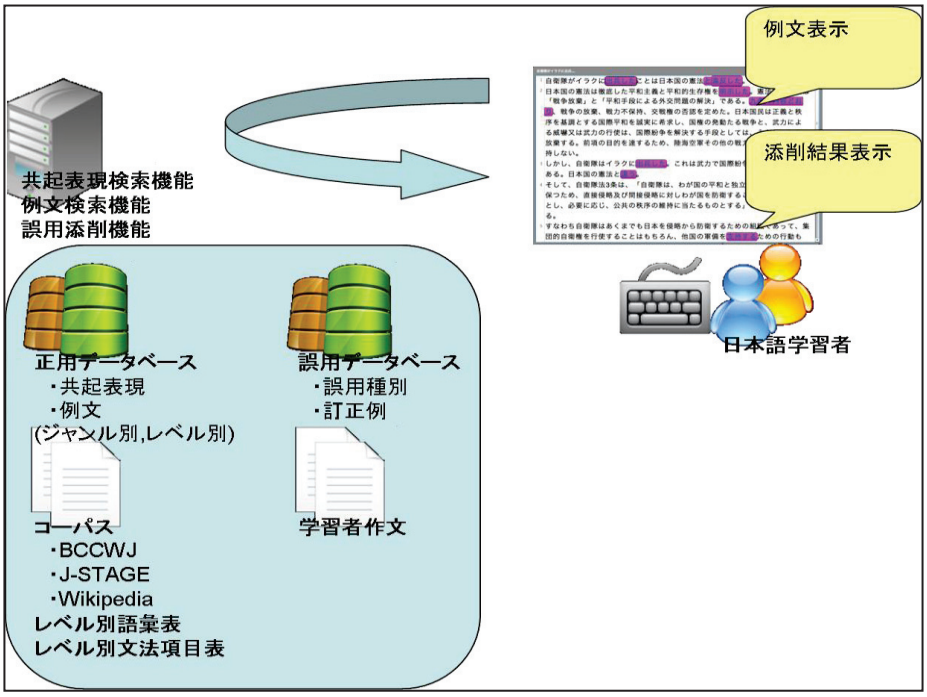


図１　作文支援システム「なつめ」の完成図

† knishina@ryu.titech.ac.jp

2 期にわたり本特定研究に参加し、作文支援システム「なつめ」の開発を行ってきた。

当初からの目標は理系留学生が論文やレポート作成をすることを可能にするシステム構築にあった。そのため学習者が論文を書くために必要とされる論文スキーマの実証的な分析、BCCWJ の内容、表記法を分析するとともに、分析に必要な科学技術論文などのコーパスの追加、学習者誤用コーパス収集および分析を行った。科学技術論文というジャンル、領域特有の表現形式を示すレジスター概念の重要性、BCCWJ を含むコーパスを評価すること、また、学習者にとっての困難な点を把握するための誤用分析の重要性を認識し、誤用コーパスのタグセットの整備を行った。また誤用作文のタグ付けの効率化のため、ツール班で開発中の「SLATE」を利用して大量のタグ付データを得ることで機械学習への可能性を検討した。さらに BCCWJ における表記の分布に関する認知言語学的考察、BCCWJ の評価-科学技術論文との対照などの研究も行ったが、本稿では次の項目に焦点を当てて特定研究の 4 年間にわたる研究の最終報告として以下順次報告する。1)科学技術論文の収集とコーパスの整備　2）インターフェースにおける共起の提示方法の改善　3)誤用検索インターフェースの改善と SLATE の利用　4)「なつめ」の評価実験

２．科学技術論文の収集とコーパスの整理

理系論文作成支援を目指すためには、均衡コーパスとともに科学技術系の論文コーパスの参照が必須となる。J-STAGE 上には多数の学会の電子化された論文が掲載されており、一部は無料公開もされている。しかしながら、システム上に例文表示をするためには掲載許諾が必要となる。そこで本領域前川代表を通して「土木工学会」、「日本医科大学会論文集」などの使用許諾の手続きを得た。さらに本班独自に「自然言語処理」、「電気学会」「日本化学会」の使用許諾も得た。その中から pdf をテキスト化して使用可能にしたものが表 1 の科学技術コーパス一覧である。現時点で論文数 662 件 5,184,350 文字であるが、継続して許可申請を行っている。また参照データとして、Wikipedia1,014,070,429 文字をデータベース化している。

表 1　収集した科学技術論文コーパス一覧

コーパス	論文数	文字数
自然言語処理	254	2, 663, 587
土木工学会	34	231, 685
日本医科大学医学会雑誌	28	109, 619
電気学会論文誌	162	1, 053, 056
日本化学会誌	184	1, 126, 403
合計	662	5, 184, 350

３．インターフェースにおける共起の提示方法の改善

３．１ ジャンルごとのコーパスのグループ化

「なつめ」では BCCWJ の他に、学会誌、Wikipedia をシステムのコーパスとして利用している。現時点で「なつめ」β版として公開しているサイトで表示するコーパスのグループを次のように変更した。

変更前：BCCWJ のサブ・コーパスとして流通書籍、ベストセラー、Yahoo!知恵袋、国会会

議録、検定教科書、白書、Yahoo!ブログ、生産書籍、雑誌、新聞、Wikipedia、自然言語処理、科学技術論文

変更後：BCCWJ 書籍（流通書籍、ベストセラー、生産書籍、雑誌）Yahoo!知恵袋、国会会議録、検定教科書、白書、Yahoo!ブログ、新聞、Wikipedia、科学技術論文

すなわち、BCCWJ の複数のジャンルを再構成し、「流通書籍、生産書籍、ベストセラー、雑誌」を一つのグループとみなした。また論文誌「自然言語処理」を新たに収集した J-STAGE の論文誌と併せて「科学技術論文」としてまとめた。Wikipedia の量が膨大であり、参照コーパスとして利用できると考えて加えた。

３．２　ジャンル別共起語の表示

図 2 は、「なつめ」の共起検索画面のフォトショットである。名詞「実験」を入力し、その右にあるリストの提示方法として「頻度、Dice 係数、Tscore、MI 値」など共起の指標を入力後、類義語を選択した表示結果を選択すると、この値により語が順に提示される。利用者が使用する意図あるいは目的によってこれらの指標を選択することができる。さらにその後画面右下の類義語を選択し、「を」の下欄の「行う」を選んだ結果である。格助詞「が、を、に、で、から、より、と」の 8 種が表示されており、それぞれの格助詞の下欄には頻度順に語が表示されている。またそれぞれの語の左側にある四角形は幅の大小で、頻度の高低を示している。図 2 の場合は、「を」「行う」が最も頻度の高い共起語であることを示している。同一ジャンル中で共起の度合いが高いものは数字がピンク色の背景色で特に低いものは空色の背景色で示される。図 2 では「実験を＊」についての「ジャンル別頻度」では、「を行う」の左の数字がピンク色で示され、科学技術文においては「実験」は「を行う」の共起頻度が高いことがわかる。さらに、そこで示される数字の部分をクリッ

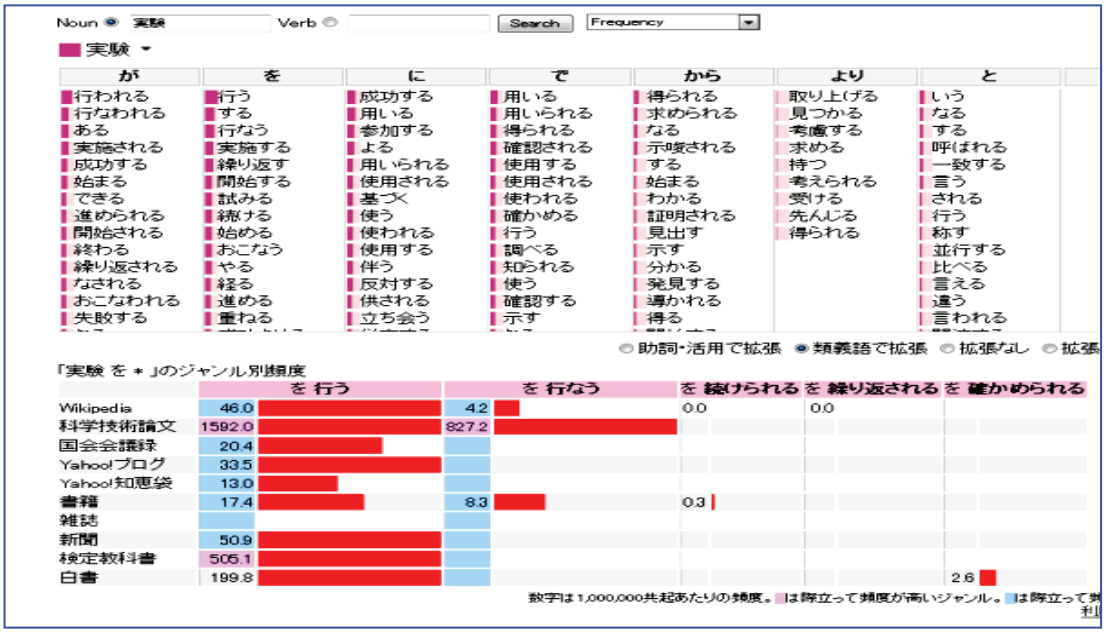


図 2「なつめ」の共起検索画面

クすると例文が 5 例表示されるように設定されている。下記はその一例である。

例文 1 : そこで, 次のような課題について実証実験を行った. 出典:安永 尚志「自然言語処理 (2/20): 「国文学作品のテキストデータ記述ルールについて」」 1996.

３．３　共起語の 2 語以上の比較

「なつめ」の「名詞」あるいは「動詞」のフレームにそれぞれ 4 語まで、語を入力してそれぞれの共起の度合いを比較することができる。例えば「実験、試験、テスト」を入力すると図 3 のように 3 語が 3 色の背景色(赤色、青色、緑色)で示される。「実験、試験、テスト」がそれぞれ四角形の大小幅で示される。「を行う」がどの語とも強く共起することが分かった。一方、「試験」と「テスト」は「受ける」が強い共起を示すが、「実験」は「受ける」とは共起が弱いことがわかる。一方、「検証する」は「実験」でのみ共起が見られ、その他は共起が見られない。このような共起の強弱を知ることで、学習者が語の選択のヒントを得ることが期待される。

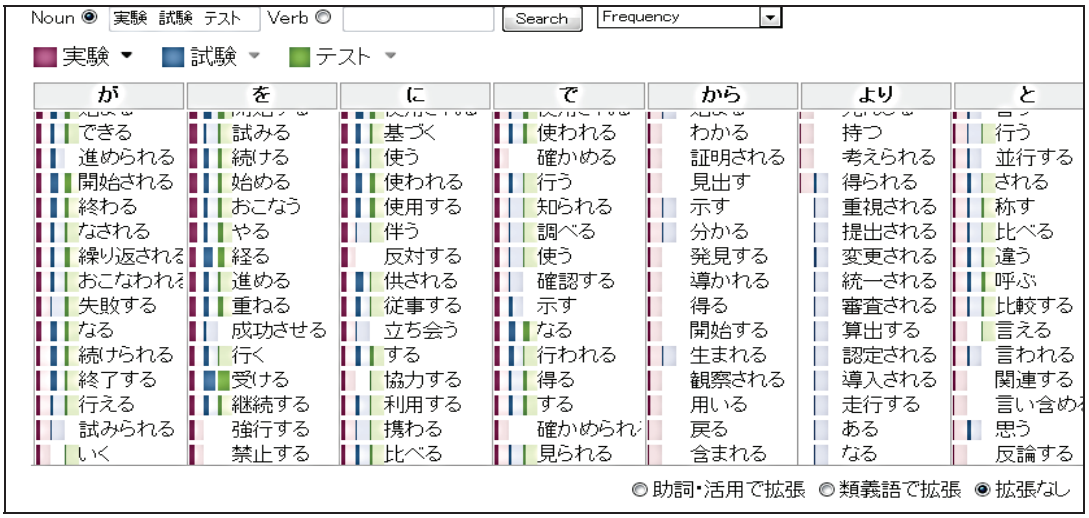


図 3　3 語の共起を比較する表示画面

３．４　レジスターを意識した語の選択

「社会的な拘束力をもつ言語学上の規範」というレジスターの概念をもとに Halliday らは機能文法として次の三つの言語の使用域による変異を提示している。

(1) コミュニケーションの目的と主題に関わる「フィールド」(Field of discourse)

(2) コミュニケーションを行うための手段に関わる「モード」(Mode of discourse)

(3) コミュニケーションパートナー同士の関係に関わる「テナー」(Tenor of discourse)

我々のコーパスにおいては「フィールド」としては, 学会での論文、政府機関の報告書である白書、情報伝達のための新聞記事、日常生活の話題を扱う知恵袋、個人的な気持ちを述べるブログなどの変異が考えられる。「モード」としては、書籍、雑誌などの文字媒体、対面の議論を記録した国会議事録などでさまざまな変異が見られる。テレビや映画などの映像と音声を含む話し言葉の変異も考えられるが、本コーパスは含んでいない。「テナー」

としては話者(書き手)と聞き手(読み手)の関係で、講義、講演を含む独話、対話、会話など話者と聞き手の複数話者、これらのディスコース参加者の社会的、個人的関係、特性（男女、年齢、社会的地位)、場のフォーマリティの程度などが含まれるが、「書き言葉コーパス」は一人の著者による多数への一方的な伝達が中心といえる。科学技術論文は専門的分野において専門家がその分野の読者を対象に論述するものといえる。次章では、誤用コーパスを用いて、学習者が作文する際これらのレジスターの使い分けが定着していないための語の不適切な選択について述べる。

４．誤用検索インターフェースの改善と SLATE の利用

４．１　学習者作文コーパスの収集と分析

作文支援ツール構築のためには、BCCWJ、科学技術論文などの正用データとともに学習者の誤用の傾向を知るための誤用データベースを利用する必要である。学習者が文を入力したときに、その入力文の妥当性を評価し、誤った場合にシステムが正用例を示すというプロセスが必要となる。例えば学習者が「影響をあげた」と入力したとする。誤用コーパスには「漢字は日本の文化に大きな影響をあげた」と例があるが、正用例で「影響をあげる」という例は見られない。このことから学習者の入力は誤用の可能性が高く、正用コーパスから「影響を与える」という共起を修正候補とし、この共起を含む例文を提示することができ、学習者は与えられたヒントによって適切な文を作成することが可能になる。

現在に至るまで作文支援班では収集した大学および大学院留学生の日本語作文をもとに、データベースを構築している。現時点で 5,000 文程度整理されており、誤用タグを付与したものは 3,500 文程度となった（曹他 2010)。また、これらの作文の執筆者である日本語学習者の情報を整備し、学習者データベースとした。119 名の中国語母語話者をはじめとする 10 カ国 164 名分の作文が 5,391 件収録されている。情報の内容は学習者に年齢、身分、日本語学習歴、母語、専門分野、作文学習の経験などが記述されている。

学習者作文に対するタグセットを付与するためにEXCEL上に作業者が書き込んできたが、タグをつけるために表の上下左右を移動しながら誤用の種類を判断する作業による誤りは多くなり、効率が悪い。そこで、この困難さを除くために、SLATE を用いることとした。これにより効率的なアノテーションができ、機械学習に向けての準備が可能になる。これを機に従来の 3 層構造の不備を修正して、表 2〜表 6 のような改善案を提案した。
表 2 で誤用の枠組みとして、それぞれ視点の異なる分類方法として、誤用の対象、誤用の内容、誤用の要因あるいは背景、誤用判断に必要な情報の 4 項目を設定する。表 3 では誤用の対象を示しており、音素・文字レベルから段落レベルまでの言語単位における誤用の形態を扱う。表 4 は、同じく音素・文字レベルからディスコースに至るまでの内容に関する誤用を扱うことを示す。表 5 はこれらの誤用が学習者の意識の何に起因しているかを記述する項目である。表 6 は誤用の情報の階層を示している。

タグをつけるに当たり、明らかに「誤り」と考えられるものと、「誤りとは言えないが不自然」なものの 2 種類があることがわかる。表 3 の「送り仮名、音、正書法からの逸脱、

表 2　誤用データベースの枠組み

誤用の対象
誤用の内容
誤用の要因／背景
誤用判断に必要な情報の階層

表 3　誤用の対象

論理的整合性の欠如
段落の欠落・余剰
接続
指示語
主述のねじれ
語句の欠落・余剰
活用
文法機能
語の共起
語の選択
正書法からの逸脱
送り仮名
文字種
音
異なる文字種による表記を推奨
その他

表 4　誤用の内容

論理構成要素	序論
	本論
	結論
段落	
文	
節・句	
品詞	
句読点	

表 5　誤用の要因・背景

母語干渉
意味の類似
字形の類似
同音異字
レジスター

表 6　誤用判断に必要な情報の階層

談話
意味
構文
形態素

文法機能、活用、主述のねじれ」などは明確に誤りと判定できる場合が多い。一方、「語の選択、語句の欠落・余剰、段落の欠落・余剰」などは、後者の「不自然」といわるものも含まれ機械的に判断するのはむずかしい。このことから前者は客観的な判断と量的な採取も可能であり、機械学習による自動タグ付けの期待が持てる。後者については更なる方法の検討が必要となる。

５．学習者実験による「なつめ」と BCCWJ の評価

５．１　学習者実験の実施

3 章で述べた「なつめ」における共起検索の有効性について、以下の学習者評価実験を試みた。本実験は、2010 年 7 月に行った予備実験に改良を加えた第 2 回目の実験であり、基本方針は第 1 回目とほぼ同様である。

期日：2010 年 11 月 24 日～30 日

実験対象：理系学部 1 年生、2 年生 40 名　大学予備教育研修生　4 名

学部生は日本語能力試験 1 級合格者である。大学予備教育研修生は、参考のためにデータを取ったものであり、日本語学習歴は前半 6 か月の集中日本語教育を韓国で終えて、日本での後半 6 か月の教育を開始して間もなくの大学入学前の予備学生である。

実験場所：教室およびコンピュータのアクセス可能な場所

実験手順：実験協力者に論文らしい文および文章を作成することを課題とする。

作題文：それぞれ 1 級から 4 級および級外までの語彙がほぼ均等になるように配置され、論文のためには書き換えが必要な項目が均等に含まれる問題用紙を A､B の 2 種類準備する。

実施方法：(1)クラスの学生を SPOT（Simple Performance-Oriented Test;小林他（1995））の点数などから成績が均等になるように 1 グループ、2 グループに 2 分類し、1 グループに問題文 A、2 グループに問題文 B を配布し、「できるだけ論文らしい表現になるように」と指示し筆記で解答させた。所用時間は 60 分程度であり、電子辞書の使用は許可した。

(2)次に授業後に「なつめ」β 上にある共起検索サイトを利用して、1、2 グループに対して電子化テキスト問題文を交換して課した。即ち、1 グループに問題文 B、2 グループに問題文 A を課した。筆記テストと同様にできるだけ論文らしい文を作成すること、その際できるだけシステム上の「類義語」を参照し、「科学技術文」など論文に近いコーパスに高い頻度があるものを選択するようにと指示した。

採点方法：1 文中に 1 セットないし 2 セットの共起ペアの検索が可能になるように設定した。そこで出題者が期待する共起セットが科学的な表現に置き換える共起セット 23 セット各 2 点とし、ツールを利用した場合も、筆記の場合も同じ基準で採点した。一方では、「なつめ」で検索が可能ではない語句について評価することを試みた。科学技術論文のレジスターとして必要と思われる副詞、形容詞、形容動詞、文末モダリティなどの表現項目を 20 項目挿入し、それらの書き換えがされた場合は 1 点を加点することとした。また書き換えるべき語は適切に選択しているが、書き換えた語がやや不自然と思われる場合は 0.5 点加算することとした。

５．２　実験結果の考察

筆記受験者 44 名、そのうち「なつめ」の検索実験に参加した者 40 名について上記の評価方法によって得点を集計した。A、B それぞれのグループの得点を集計し、筆記実験のグループ別平均値の差を検定した結果､A､B 間の有意差は認められず(t=-0.1454､自由度=38、p 値=0.8851)､ほぼ均一レベルの学習者であり､問題文の難易度の差もないことがわかった。そこで A、B グループを合わせた 40 名分の筆記と「なつめ」検索による解答をみると、受験者全体で筆記の最高得点は 59 点、「なつめ」検索による最高得点は 62 点であった。さらに筆記と「なつめ」による試験の平均得点を検定した結果、１％水準で「なつめ」を使用した場合の得点が有意に上がったことがわかった（t=7.2269、p 値<0.001)。また筆記試験結果と「なつめ」による解答結果を筆記得点順に上位群 35 点以上（8 名)、中位群 16 点以上（20 名)、下位群 15 点以下（12 名）に分けて比較する。この中で研修生が上位群に 1 名、

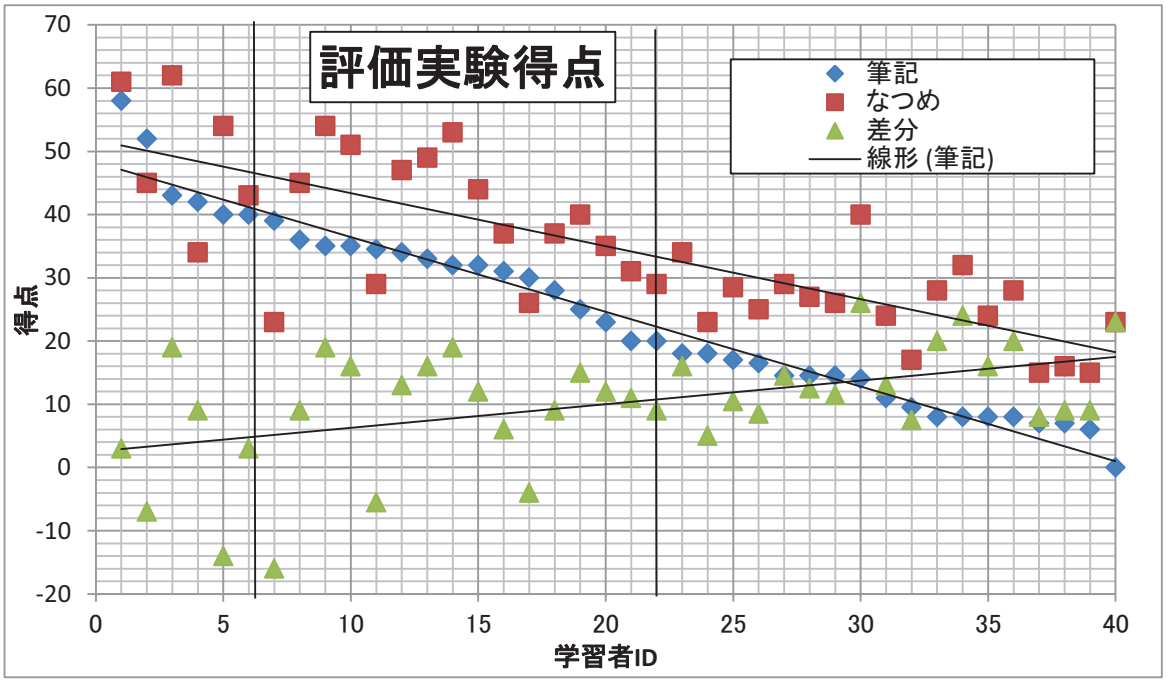


図 4　「なつめ」と筆記による評価実験結果の得点分布

表 7　レベル別の得点結果

	上位群	中位群	下位群
対象学習者	8 名	20 名	12 名
筆記得点（平均点）	58～36（43.75）	35～16.5（26.78）	14.5～0（9.29）
なつめ得点(平均点)	62～23※(45.88)	54～23（37.36）	40～15（24.57）
差分範囲(平均点)	19～-16（0.75）	19～-5.5（10.41）	26～7.5（15.29）

中位群 1 名、下位群 2 名が含まれ、学習期間が短いにもかかわらず上位に入っているものもいることがわかった。上位群は「なつめ」を使用した効果の差分が小さく、中位群、下位群の順で差分は大きくなっていることがわかる。上位群の 3 名は筆記の結果の方が「なつめ」を利用したものより高く、中位群でも同様に筆記の方が「なつめ」検索より高いものが 2 名いる。本実験において「なつめ」の得点が低い原因としていくつか考えられる。

その理由の一つは先述した共起以外の項目で、副詞、名詞、形容詞、形容動詞、モダリティを含む文末表現のレジスター項目を挿入しており、上位群はそれらの項目を適切に書き換えたことで、得点をしている。また面談の結果、「なつめ」の操作が困難な学習者とそうでない学習者がいることがわかり、操作になじめない学習者は習得している知識を用いて筆記で解答する方が検索に要する時間より速いことが考えられる。

５．３　評価実験のまとめ

「なつめ」を利用した学習者実験を行うことで日本語コーパスおよびその他のコーパスの有効性、インターフェースの良否の評価をした。学部留学生を中心とする日本語能力試験 1 級合格者を中心に 40 名の実験協力者を得て、筆記試験と「なつめ」を用いた両方の実

験を行った結果、特に日本語能力試験 1 級合格者の中でも中位群、下位群において「なつめ」利用の効果が高いことが明らかになった。

学習者のコメントとして「名詞と動詞の組み合わせだけでなく、形容詞や副詞などの使用も含まれるなら、もっとよい」、「適当な名詞を分かっていればそれに対して適した動詞を見つけることができるので便利だ」「もし元の言葉がよく使う言葉なら、この言葉を探しやすくて、直しやすい」などの記述がある。短所として「ある語から同一品詞の類義語を直接探しにくい。」などの記述があった。

上位群に効果が高くない理由としては、名詞・動詞の共起以外の様々な語の共起、レジスターのヴァリエーションが現時点で利用可能でないことが一要因と考えられ、今後の改良点として留意する。本実験においてはアカデミックな文章における書き手の意識を観察することができた。実験協力者は学部 1 年生が大多数であり、レポートなどの課題においてアカデミックな文章が要求されていることは知っているが、具体的な語法は習得していない状況が観察された。ここに「レジスター」という視点を導入すれば、教授項目のかなり部分が整備されることが予測される。

６．全体のまとめと今後の課題

作文支援班では、外国人留学生の論文作成支援をするシステムを開発するため、BCCWJ のコーパスを再分類し、新たに論文コーパスを収集した。それとともに誤用データベース開発の必要性も述べ、開発を進めた。特定目的の作文支援ということから文章における各レジスターによる表現の異なりに注目し、コーパスをもとに共起検索と例文表示を中心に開発を行い、誤用データベースから得られる学習者の誤用確率情報を利用することで、作文支援システムとしてユニークな「なつめ」の実現を試みた。さらにコーパスの質およびシステムの効果を評価するために共起に焦点を当てた学習者実験を行った。その結果、日本語能力が上位にある学習者において「なつめ」を利用した方が筆記より得点が有意に高いことが明らかになった。一方、誤用データベースを構築するために SLATE を利用してアノテーションが正確で早く付与できるように試み、機械学習による誤用データベース作成の可能性を確認した。それと並行して教師用として利用可能な検索システムも開発した。

今後の計画として均衡コーパスとしての BCCWJ と誤用コーパスを併用し、さらに語彙ネットワークを利用して意味的なサポートも可能にし、確率的な手法によって学習者が自らの作文の妥当性を知ることができるシステムへと発展させることを考えている。論文作成の外に、「手紙」「メール」「報道記事」などの分野でのレジスターによる記述方法にも展開できる可能性が考えられる。「なつめ」をさらに実用的なシステムとするためには、目的別の正用コーパスの整備、機械学習を考慮に入れた誤用コーパスの拡張、名詞動詞以外の形容詞、副詞などの共起検索項目の追加改善が必要であり、本特定研究終了後も継続して改善する予定である。

文献

曹紅荃・仁科喜久子「語彙のネットワークとグルーピングを利用した日本語の語彙学習と指導」

Journal of Yanbian University Social Science, No42, pp55-60, Dec, 2009.
阿辺川武・Hodošček Bor ・仁科喜久子「日本語作文支援システム「なつめ」における共起語検索方法の改訂」、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予稿集 pp243-244、Mar 2010.
市川保子「日本語誤用例文小辞典」(1997)
阿辺川武・Hodošček Bor ・仁科喜久子(2010b)「日本語作文支援システム「なつめ」―利用者の視点―」特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集 pp243-244
小野正樹・小林典子・長谷川守寿 (2009, 2010)『コロケーションで増やす表現 Vol. 1, 2』くろしお出版
小林典子・フォード丹羽順子・山元啓史 (1995)「『日本語能力簡易試験 (SPOT)』の得点分布. 傾向」『筑波大学留学生センター日本語教育論集』第 10 号: 107-120. 4.
仁科喜久子(2010) 日本語コーパスに基づいた日本語学習支援システムにおける語の提示, 語彙・辞書研究会 第３８回研究発表会, 語彙・辞書研究会, pp. 9-16
Biber, Douglas (1988) Variation across Speech and Writing. Cambridge University Press
Biber, Douglas and Susan Conrad (2009). Register, Genre, and Style. Cambridge: Cambridge Textbooks
Halliday, and Matthiessen (2004). An Introduction to Functional Grammar. 3d ed. London: Arnold
Hodošček Bor (2010) Development of a Register-based Writing Assistance System for Academic Japanese （Master thesis at Tokyo Institute of Technology）
Irena Srdanovic Erjavec, Tomaž Erjavec, Adam Kilgarriff (2008) A web corpus and word sketches for Japanese. Journal of NLP. pp.1-22

関連 URL

東京工業大学留学生センター仁科研究室で開発している日本語学習支援システム「あすなろ」「なつめ」http://hinoki.ryu.titech.ac.jp/

学会発表

Andrej Bekeš. “The role of Adverb based structure in Japanese discourse: The case of conditional adverb *moshi* ”, 23th Paris Meeting on East Asian Linguistics, 2010.
曹 紅荃・黒田史彦・八木豊・鈴木泰山・仁科 喜久子「学習者作文支援システムのための誤用データベース作成-動詞の誤用分析を中心に-」, pp. 1571-1～1571-9, 世界日語教育大会論文集, 主催：国立政治大学 台湾 2010.
村岡貴子・因京子・仁科喜久子「専門日本語ライティング能力の獲得を目指す日本語テキスト分析タスク活動を通じたスキーマ形成」pp. 13250-0～13250-9 世界日本語教育大会論文集, 主催：国立政治大学, 台湾, 2010.
鎌田美千子「文体の違いに対する日本語学習者のパラフレーズ-具体例からの抽象化に着目して-」pp. 11361-1～1136. 7, 世界日本語教育大会論文集, 主催：国立政治大学, 台湾, 2010
Joyce, Terry, Hodošček, Bor & Nishina, Kikuko(accepted). “Orthographic representation within the Japanese writing system”, Presentation at "Units of language - units of writing" 7th International Workshop on Writing Systems and Literacy, Paris, 2010.

研究活動・成果の総括：意見情報班
多様な文書ジャンルを対象とした
意見分析コーパスの作成に関する研究

関　洋平　（班長：筑波大学大学院 図書館情報メディア研究科）†
神門　典子　（分担者：国立情報学研究所 情報社会相関研究系）
佐野　大樹　（連携研究者：国立国語研究所 コーパス開発センター）
柏野　和佳子（連携研究者：国立国語研究所 言語資源研究系）
稲垣　陽一　（協力者：きざしカンパニー）
栗山　和子　（協力者：白百合女子大学 文学部）

Final Progress Report: 'Opinion Information' Group

Yohei Seki　(University of Tsukuba)
Noriko Kando　(National Institute of Informatics)
Motoki Sano　(National Institute for Japanese Language and Linguistics)
Wakako Kashino　(National Institute for Japanese Language and Linguistics)
Yoichi Inagaki　(Kizasi Company, Inc.)
Kazuko Kuriyama　(Shirayuri College)

1. 意見情報班の目標

ここ数年，Webなどの大量の電子化テキストに現れる意見情報を抽出し，集約や可視化を行うことで，世論調査や評判分析といった応用を実現する研究が進んでいる（Pang&Lee, 2008; 大塚ら 2007)。対象となる文書ジャンルは，個人が自身の体験や意見を記述するブログやTwitter，Web上のレビューサイト，報道機関が発信するニュースなどであり，商品や映画の評判分析，トレンド分析，政策や選挙のための情報分析，世論調査などについて実用化が進められている[1]。なお，一口に意見情報といってもその特徴はさまざまであり，文書ジャンル（例：新聞，ブログ，SNS・QAサイトなどのコミュニティサイト，雑誌，会議録等）やドメイン（政治，映画，商品，恋愛相談等）に応じて，文書中で使用される概念や語彙の傾向は異なり（Gliozzo et al., 2009），その話題についての意見情報の傾向も異なる（Blitzer et al., 2007; 関ら，2010)。たとえば，新聞記事やニュースサイトでは，以下のような引用意見が頻出する。

(1) ネット関連業界では「ヤフーも、どこかと合併するのではないか」という観測も消えないが、A氏は「インターネットで大切なのは規模ではない」と否定的だ。

(1)のように，周囲や他者（特に専門家）の意見を引用することは新聞記事では多いが，個人の意見や体験に焦点を置いたブログや，他者とのやり取りに重点を置いたコミュニティサイトのような文書ジャンルでは，必ずしも多く出現しない。

また，抽出しても応用の目的にそぐわない意見情報も存在する。たとえば，中国における段ボール肉まんの事件について，日本での世論を調査する際に，

(2) その情報は正しかったのだろう。

という意見がブログ中に出現したとする。(2)の文は，直前の「当初の報道によれば、番組

† yohei@slis.tsukuba.ac.jp
[1] http://www.sentimentsymposium.com

はタレコミ情報を出発点にして、店主に商談を持ちかけて撮影を実行したとされていた。」という文に対する確認の意見とする。しかし，たとえば（2）の文を肯定的な評価意見と判定して，世論調査の集計に利用すれば，分析結果には誤った情報が混入する。

昨年度，意見情報班では，国立国語研究所の製作する現代日本語書き言葉均衡コーパス（BCCWJ）の中から Yahoo!知恵袋と書籍，それ以外に NTCIR 意見分析コーパスから新聞記事，また，ICWSM2009 コーパスからブログを主な対象とすることで，さまざまな文書ジャンルやドメインを対象として意見情報の分析を進め，傾向の違いを整理し，情報アクセスに着目した応用を進める上で，必要となる意見情報を明らかにした。

今年度の前半は，これらの意見情報の特徴の違いを判別することを目的として，アプレイザル理論（Martin&White, 2005）に基づく英語の辞書（Bloom et al., 2007）から，日本語のアプレイザル辞書を構築し，意見抽出における効果について検証を行った。検証に当たっては, NTCIR-8 多言語意見分析タスク（Seki et al., 2010）の評価用データセットを利用した。

今年度の後半は，口語表現・会話表現に焦点を当てて，Yahoo!ブログと国会会議録を対象として，ある程度の規模の意見分析コーパスを作成した。

本論文の構成は以下の通り。2 章では，意見情報の差異を分析するためのアプレイザル理論の関連研究を紹介する。3 章では，今年度構築したコーパスの概要について紹介する。4 章では，構築したコーパスの分析結果について説明する。5 章で結論についてまとめる。

2. 関連研究

2.1 アプレイザル理論

アプレイザル理論（Martin&White, 2005）は，システミック文法の対人メタ機能（interpersonal meta-function）を，談話意味論（discourse semantics）の観点から整理した体系である。Martin&Whiteは，テキスト中に現れる対人メタ機能の意味は，仮想的な読者（putative reader）に対する感情や対話であるという信念に基づき，appraisal, negotiation, involvement の 3 つのシステムから構成されるとし，appraisal は，態度評価（attitude），形勢・やり取り（engagement），程度評価（graduation）の 3 つのシステムから構成されるとした。このうち態度評価は，感情（emotion），倫理（ethics），美学（aesthetics）の区別に基づき，主体が表明する感情（affect），人間の振舞に対する規範や世評に基づく評価（judgment），事物や事象等に対する観照や価値に基づく評価（appreciation）の 3 つに分類される。以下で特徴を説明する。[2]

1　主体が表明する感情（affect，感情）

第 1 のタイプは，主体が表明する感情の態度評価で，心理状態を記述する動詞，属性形容詞，叙述形容詞，形容詞に関連した副詞などで表現される。下位タイプの要素と具体例を以下に示す。

- 切望・敬遠（dis/inclination）：要求する，切望する，～たい/用心深い，恐れ（て～しない）
- 幸福・不幸（un/happiness）：笑う，愛する/泣く，かなしい，嫌悪する
- 安心・不安（in/security）：信頼する，任せる/驚く，心配する
- 満足・不満（dis/satisfaction）：充実した/怒る

2　人間の振舞に対する規範や世評に基づく評価（judgment，規範・世評）

第 2 のタイプは，人間の振舞に対する社会規範や世評に基づく態度評価で，肯定・否定

[2] 用語の和訳は，第 1 著者と第 3 著者で協議したものに準拠する。

の両面から，以下の下位タイプが指定されている。この意見情報は，政策や選挙のための情報分析，世論調査で重要な役割を果たすと考えられる。肯定・否定の基準は，ドメインに依存する。

- 通常・特別（normality，特殊性）：自然な，ラッキーな /奇妙な，風変わりな
- 有能・無能（capacity，有能さ）：強力な，健全な，成熟した/弱い，愚かな，鈍い
- 頑強・軽薄（tenacity，信頼性）：勇敢な，信頼に足る，忠実な/軽率な，せっかちな
- 真実・不実（veracity，正直さ）：正直な，信憑性のある，率直な/だます，嘘つきの
- 倫理・邪悪（propriety, 倫理的是非）：公正な，遵法精神のある，思慮深い /邪悪な，残酷な

3　事物や事象等に対する観照や価値に基づく評価（appreciation，反応・構成・価値）

第 3 のタイプは，事物や事象等に対する観照や価値に基づく態度評価で，肯定・否定の両面から，以下の下位タイプが指定されている。この意見情報は，商品の評判分析などで重要な役割を果たすと考えられる。肯定・否定の判断基準と用語の選択は，ドメインに強く依存する。

- 衝撃・退屈（impact，衝撃性）：目立つ，刺激的な，強烈な/うんざり，単調な
- 魅力・嫌悪（quality，質感）：華麗な，美しい，魅惑的な/不愉快な，グロテスク
- 調和・混乱（balance，調和性）：均整のとれた，一貫，すらっとした/むら，矛盾，ずさん
- 明瞭・複雑（composition，複雑さ）：純粋，わかりやすい，正確/飾り立て，わかりにくい
- 有用・無用や有害（valuation，価値）：鋭い，革新的な/つまらない，浅はかな

アプレイザル理論の関連研究として，Somasundaran ら（2007）では，電子掲示板とニュースを対象として，Sentiment（肯定・否定の傾向をもつ意見）と Arguing（議論）を分類し，質問応答に利用している。Argamon ら（2007）は，態度評価の下位タイプの自動分類に取り組んでいる。佐野（2010a, 2010b）は，ブログを対象としてアプレイザル理論に基づく評価語彙と評価対象の関係などの分析を進めている。

3.　構築した意見分析コーパス

3.1　対象コーパスの概要

意見情報班では，今年度は，現代日本語書き言葉均衡コーパス（BCCWJ）の中から，Yahoo! ブログのコアデータと，国会会議録の本会議を対象として文書を選択した。選択した文書のデータサイズを**表 1** に示す。

表 1　作成した意見分析コーパス

文書ジャンル	ソース	文書数	総文数	内容	作成年度
Yahoo!知恵袋	BCCWJ	251	1,924	コアデータのうち，主要7カテゴリ	平成21年度
書籍		10	407	評論，随筆など	平成21年度
Yahoo!ブログ		**471**	**6,944**	**コアデータすべて**	**平成22年度**
国会会議録		**14**	**5,812**	**本会議録**	**平成22年度**
新聞	NTCIR	780	21,391	NTCIR-6, 7 MOAT	平成21年度
Yahoo!知恵袋	API	16	118	アプリケーション用途	平成21年度
ブログ	ICWSM	80	2294	アプリケーション用途	平成21年度

Yahoo! ブログについては，15 カテゴリのすべてのコアデータを対象とし，国会会議録については，衆参両議院の本会議録として収録されているすべてのデータを対象とした。

3.2 アノテーション属性

コーパスのアノテーション属性は，平成 21 年度の方針に従い，以下の要素とする。アノテーションの単位は，文または 1 文に含まれる複数の意見とする。

- 意見情報：意見性，極性，態度評価，形成・やり取り評価，発話行為タイプ，意見保有者，意見対象
- 体験情報：体験性，体験主，体験タイプ，意見誘発体験

また，今年度分のデータを対象とした新規の属性として，以下の要素を追加した。

- 意見対象タイプ：意見対象について，拡張固有表現（関根ら，2007）の第 1 階層を手がかりとして，タイプラベルを付与した。
- カテゴリ適合性（Yahoo!ブログのみ）：
  Yahoo!ブログでは，Yahoo!側で設定したカテゴリと異なる内容のブログが多い。この点を判別するために，カテゴリ適合性の属性を追加し，カテゴリに適合した内容の文とそうでない文を判別した。
- 会話タイプ（国会会議録のみ）：
  国会会議録では，対話相手とのやり取りの分類が重要である。本コーパスでは，会話タイプを定義し，質問，回答，議事進行，呼びかけ，その他を区別した。

また，2 名の判定者間での判定者間一致率（**κ** 係数）を確認した結果を表 2 に示す。

表 2 Yahoo!ブログ，国会会議録の判定者間一致率（κ 係数）

文書ジャンル	意見情報						体験情報				その他	
	意見性	極性	態度評価	やり取り	発話行為	意見対象タイプ	体験性	体験タイプ	体験主	意見誘発	カテゴリ適合性	会話タイプ
Yahoo!ブログ	0.8934	0.7904	0.6106	0.6499	0.8280	0.6662	0.7281	0.6787	0.6248	0.5969	0.6444	
国会会議録	0.9171	0.8745	0.6569	0.6365	0.6560	0.6350	0.9002	0.7615	0.8205	1.0000		0.8174

4. 分析結果と考察

4.1 Yahoo!ブログの分析結果

Yahoo!ブログを対象として，態度評価の分布状況をカテゴリ別にまとめた結果を，表 3 に示す。15%以上の割合を示した態度評価の値は太字で強調している。昨年度も Yahoo!知恵袋のアノテーション結果に基づき議論したように，政治カテゴリには規範・世評が多いなど，カテゴリ（ドメイン）に依存して意見の分布が異なる状況が把握できる。

表 3 Yahoo!ブログにおけるカテゴリ別態度評価の分布

カテゴリ	感情の詳細					規範・世評の詳細						反応・校正・価値の詳細					
	切望/敬遠	幸せ/不幸	安心/不安	満足/不満	合計	通常/特別	有能/無能	頑強/軽薄	真実/不実	倫理/邪悪	合計	衝撃/退屈	魅力/嫌悪	調和/混乱	明瞭/複雑	有用/無用	合計
ビジネスと経済	9.29	5.00	10.71	12.86	**37.86**	2.14	0.71	7.14	0.00	0.71	10.71	3.57	5.00	4.29	3.57	**35.00**	**51.43**
コンピュータとインターネット	0.00	**27.59**	0.00	3.45	**31.03**	3.45	3.45	**24.14**	0.00	0.00	**31.03**	0.00	6.90	0.00	3.45	**31.03**	**41.38**
生活と文化	12.54	9.83	5.08	13.90	**41.36**	2.71	1.02	4.75	0.00	0.00	8.47	12.20	10.17	3.05	4.07	**18.98**	**48.47**
エンターテインメント	11.51	7.93	3.32	6.65	**29.41**	1.79	2.56	5.37	0.26	0.00	9.97	**21.99**	**20.46**	3.58	4.60	8.95	**59.59**
家庭と住まい	10.26	9.40	5.56	6.41	**31.62**	0.85	1.71	8.97	0.43	0.00	11.97	**28.21**	18.80	0.85	0.85	5.56	**54.27**
政治	10.78	2.99	2.40	13.17	**29.34**	4.19	8.38	**17.37**	3.59	2.40	**35.93**	1.80	2.99	3.59	**19.16**	1.80	**29.34**
健康と医学	**20.69**	3.45	9.20	12.64	**45.98**	4.60	0.00	0.00	0.00	0.00	4.60	1.15	9.20	5.75	10.34	**17.24**	**43.68**
学校と教育	10.14	1.45	1.45	10.14	**23.19**	2.90	1.45	5.80	0.00	0.00	10.14	**26.09**	14.49	2.90	1.45	**17.39**	**62.32**
科学	0.00	7.41	14.81	14.81	**37.04**	0.00	0.00	0.00	0.00	0.00	0.00	**25.93**	**25.93**	0.00	0.00	11.11	**62.96**
出会い	0.00	0.00	0.00	**18.75**	**18.75**	0.00	0.00	12.50	0.00	0.00	12.50	**37.50**	**25.00**	6.25	0.00	0.00	**68.75**
地域	5.53	5.53	7.04	5.53	**23.62**	0.50	0.50	4.52	0.00	1.01	6.53	**35.68**	13.57	0.50	1.51	**16.58**	**67.84**
芸術と人文	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**33.33**	0.00	0.00	**66.67**	**100.00**
Yahoo!サービス	12.89	5.15	2.06	10.82	**30.93**	1.03	3.61	6.70	0.00	0.00	11.34	**22.68**	**15.98**	4.12	2.58	12.37	**57.73**
趣味とスポーツ	9.07	4.72	3.80	7.54	**25.12**	1.16	0.12	1.04	0.06	0.37	2.76	5.02	10.78	14.46	**36.58**	3.19	**70.04**
特集（趣味とスポーツ・競馬）	13.32	1.82	0.97	3.03	**19.13**	0.00	0.00	0.00	0.00	0.24	0.24	2.54	10.29	**19.85**	**45.16**	2.30	**80.15**
マクロ平均	8.40	6.15	4.43	9.31	**28.29**	1.69	1.57	6.55	0.29	0.31	10.41	14.96	14.86	4.61	8.89	**16.54**	**59.86**

さらに，頻出する上位 4 つの態度評価と意見対象タイプの組み合わせについて，表 4 に示す。こちらも，コンピュータ，エンターテイメント，政治などドメインに特徴的な意見

を把握することができる。

表 4 カテゴリ別に頻出する態度評価と意見対象タイプの組み合わせ

カテゴリ	意見数	頻度1位		頻度2位		頻度3位		頻度4位	
		態度評価:意見対象タイプ	頻度	態度評価:意見対象タイプ	頻度	態度評価:意見対象タイプ	頻度	態度評価:意見対象タイプ	頻度
ビジネスと経済	140	(評価)有用/無用:製品名	25	(評価)有用/無用:組織名	6	(評価)有用/無用:行為概念	6	安心/不安:抽象概念	5
コンピュータとインターネット	29	(評価)有用/無用:製品名	9	幸せ/不幸:イベント名	6	頑強/軽薄:人名	5	幸せ/不幸:製品名	2
生活と文化	295	(評価)有用/無用:製品名	31	(反応)魅力/嫌悪:製品名	17	満足/不満:製品名	13	幸せ/不幸:製品名	12
エンターテインメント	391	(反応)衝撃/退屈:製品名	31	(反応)魅力/嫌悪:人名	31	(反応)魅力/嫌悪:製品名	29	(反応)衝撃/退屈:人名	27
家庭と住まい	234	(反応)衝撃/退屈:自然物名	27	(反応)魅力/嫌悪:自然物名	21	(反応)衝撃/退屈:人名	20	(反応)衝撃/退屈:製品名	14
政治	167	頑強/軽薄:人名	19	(構成)明瞭/複雑:抽象概念	12	有能/無能:人名	9	満足/不満:人名	8
健康と医学	87	(評価)有用/無用:製品名	8	切望/敬遠:行為概念	6	切望/敬遠:自然物名	4	(構成)明瞭/複雑:製品名	4
学校と教育	69	(反応)衝撃/退屈:人名	7	(反応)魅力/嫌悪:人名	5	(評価)有用/無用:行為概念	4	頑強/軽薄:人名	4
科学	27	満足/不満:自然物名	4	(反応)魅力/嫌悪:製品名	3	(評価)有用/無用:製品名	3	(反応)衝撃/退屈:自然物名	3
出会い	16	(反応)魅力/嫌悪:自然物名	4	(反応)衝撃/退屈:自然物名	4	頑強/軽薄:製品名	1	満足/不満:自然物名	1
地域	199	(反応)衝撃/退屈:製品名	22	(評価)有用/無用:製品名	18	(反応)衝撃/退屈:施設名	15	(反応)魅力/嫌悪:施設名	10
芸術と人文	3	(評価)有用/無用:イベント名	2	(反応)魅力/嫌悪:製品名	1				
Yahoo!サービス	194	(反応)魅力/嫌悪:製品名	15	(反応)衝撃/退屈:抽象概念	11	(評価)有用/無用:製品名	10	(反応)衝撃/退屈:製品名	8
趣味とスポーツ	1632	(構成)明瞭/複雑:製品名	173	(構成)明瞭/複雑:人名	133	(構成)明瞭/複雑:行為概念	79	(構成)調和/混乱:製品名	68
特集(趣味とスポーツ・競馬)	826	(構成)明瞭/複雑:製品名	172	(構成)明瞭/複雑:行為概念	84	(構成)調和/混乱:製品名	53	(反応)魅力/嫌悪:製品名	42

4.2 国会会議録の分析結果

国会会議録を対象として，アノテートした意見情報の分布状況を表 5 に示す。

表 5 国会会議録を対象とした態度表現の分布

カテゴリ	感情の詳細					規範・世評の詳細						反応・校正・価値の詳細					
	切望/敬遠	幸せ/不幸	安心/不安	満足/不満	合計	通常/特別	有能/無能	頑強/軽薄	真実/不実	倫理/邪悪	合計	衝撃/退屈	魅力/嫌悪	調和/混乱	明瞭/複雑	有用/無用	合計
1期'76-80衆議院本会議	**21.35**	5.62	3.37	1.12	**31.46**	3.37	3.37	2.25	2.25	0.00	11.24	5.62	5.62	13.48	2.25	**26.97**	**53.93**
1期'76-80参議院本会議	**33.11**	0.00	6.14	1.37	**40.61**	2.39	4.10	1.71	1.71	0.00	9.90	3.07	4.78	12.63	5.80	**19.11**	**45.39**
2期'81-85衆議院本会議	**22.54**	4.23	2.82	0.00	**29.58**	2.82	2.82	2.11	0.70	2.82	11.27	2.82	2.11	4.23	7.75	**34.51**	**51.41**
2期'81-85参議院本会議	**41.86**	4.42	2.33	0.00	**48.60**	1.86	4.42	7.44	1.16	2.56	**17.44**	3.49	2.79	4.88	4.65	14.42	**30.23**
3期'86-90衆議院本会議	**52.63**	**21.05**	0.00	0.00	**73.68**	0.00	10.53	0.00	0.00	0.00	10.53	**15.79**	0.00	0.00	0.00	0.00	**15.79**
3期'86-90参議院本会議	**40.81**	2.35	3.63	0.43	**47.22**	3.42	2.35	3.21	1.28	0.85	11.11	5.56	3.85	5.98	4.06	**20.51**	**39.96**
4期'91-95衆議院本会議	**34.24**	1.55	1.93	0.77	**38.49**	11.61	3.48	5.22	3.29	1.16	**24.76**	3.68	2.13	3.87	6.58	**19.15**	**35.40**
4期'91-95参議院本会議	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**66.67**	**66.67**
5期'96-00衆議院本会議	**34.96**	1.36	1.36	1.45	**39.13**	1.72	4.44	4.17	4.53	2.36	**17.21**	4.98	1.63	6.52	2.72	**21.20**	**37.05**
5期'96-00参議院本会議	**15.38**	0.59	3.25	2.37	**21.60**	2.07	8.58	7.40	6.80	5.03	**29.88**	5.33	0.59	5.92	4.14	**23.67**	**39.64**
6期'00-05衆議院本会議	**32.43**	0.85	6.54	5.12	**44.95**	2.56	3.70	3.98	1.56	1.14	12.94	1.71	0.71	7.82	4.13	**19.49**	**33.85**
6期'00-05参議院本会議	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**73.68**	**73.68**
衆議院本会議平均	**33.02**	5.78	2.67	1.41	**42.88**	3.68	4.72	2.96	2.06	1.25	14.66	5.76	2.03	5.99	3.90	**20.22**	**37.90**
参議院本会議平均	**21.86**	1.23	2.56	0.69	**26.34**	1.62	3.24	3.29	1.83	1.41	11.39	2.91	2.00	4.90	3.11	**36.34**	**49.26**
マクロ平均	**27.44**	3.50	2.61	1.05	**34.61**	2.65	3.98	3.12	1.94	1.33	13.02	4.34	2.02	5.44	3.51	**28.28**	**43.58**

この結果からわかることとして，国会会議録（本会議）の意見は全般に“切望・敬遠”ならびに“有用・無用”が頻出する傾向が把握できる。さらに，頻出する上位 4 つの態度評価と意見対象タイプの組み合わせについて，表 6 に示す。

表 6 国会会議録を対象とした態度評価と意見対象タイプの分布

カテゴリ	意見数	頻度1位		頻度2位		頻度3位		頻度4位	
		態度評価:意見対象タイプ	頻度	態度評価:意見対象タイプ	頻度	態度評価:意見対象タイプ	頻度	態度評価:意見対象タイプ	頻度
1期'76-80衆議院本会議	89	(評価)有用/無用:抽象概念	14	切望/敬遠:抽象概念	9	(構成)調和/混乱:抽象概念	7	切望/敬遠:行為概念	7
1期'76-80参議院本会議	293	切望/敬遠:抽象概念	50	切望/敬遠:行為概念	41	(構成)調和/混乱:抽象概念	37	切望/敬遠:製品名	32
2期'81-85衆議院本会議	142	(評価)有用/無用:製品名	36	(評価)有用/無用:行為概念	18	切望/敬遠:行為概念	16	切望/敬遠:製品名	11
2期'81-85参議院本会議	430	切望/敬遠:行為概念	119	切望/敬遠:抽象概念	63	(評価)有用/無用:行為概念	40	(評価)有用/無用:抽象概念	30
3期'86-90衆議院本会議	19	幸せ/不幸:行為概念	4	切望/敬遠:行為概念	4	有能/無能:人名	2	(反応)衝撃/退屈:行為概念	2
3期'86-90参議院本会議	468	切望/敬遠:抽象概念	98	切望/敬遠:行為概念	74	(評価)有用/無用:抽象概念	42	(評価)有用/無用:行為概念	38
4期'91-95衆議院本会議	517	切望/敬遠:行為概念	101	切望/敬遠:抽象概念	63	(評価)有用/無用:行為概念	50	(評価)有用/無用:製品名	34
4期'91-95参議院本会議	3	(評価)有用/無用:製品名	2						
5期'96-00衆議院本会議	1104	切望/敬遠:行為概念	182	(評価)有用/無用:製品名	158	切望/敬遠:抽象概念	147	切望/敬遠:製品名	76
5期'96-00参議院本会議	338	(評価)有用/無用:製品名	52	切望/敬遠:抽象概念	22	有能/無能:人名	16	頑強/軽薄:人名	16
6期'00-05衆議院本会議	703	切望/敬遠:製品名	84	(評価)有用/無用:製品名	79	切望/敬遠:抽象概念	70	切望/敬遠:行為概念	57
6期'00-05参議院本会議	19	(評価)有用/無用:製品名	14						

以上から，国会の本会議においては，“抽象概念”，“行為概念”，“製品名”などが，意見の対象とされていることがわかる。

5. おわりに

Yahoo!ブログ，国会会議録など，意見情報を多く含む文書ジャンルを対象として，昨年度の経験などを反映したアノテーションを実現し，コーパスを作成した。その結果，前年度に引き続き，Yahoo!ブログについては，ドメインに特徴的な意見の傾向を把握することができた。また，国会の本会議によく利用される態度評価も明らかとなった。

引き続き，今年度前半に作成した言語横断アプレイザル辞書や，連携研究者の方で開発を進めている評価表現辞書（佐野，2011a, 2011b）を活用し，意見分析システムの開発を続けていく予定である。

謝辞

この研究の一部は，科学研究費補助金特定領域研究（課題番号 21011003），基盤研究 B（課題番号 21300029），若手研究 (B) (課題番号 21700268)，ならびに筑波大学図書館情報メディア研究科プロジェクト研究の助成を受けて遂行された。

参考文献

S. Argamon, K. Bloom, A. Esuli, F. Sebastiani (2007) “Automatically determining attitude type and force for sentiment analysis.” *Proc. of 3rd Language and Technology Conference*, Poznan, Poland.

J. Blitzer, M. Dredze, and F. Pereira (2007) “Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification.” *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 440-447.

K. Bloom, N. Garg, and S. Argamon (2007) “Extracting Appraisal Expressions.” *Proc. of the Human Language Technology Conf. of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2007)*, Rochester New York, USA, pp.308-315.

A. Gliozzo and C. Strapparava (2009) *Semantic Domains in Computational Linguistics*, Springer-Verlag.

J. R. Martin and P. R. R. White (2005) *The Language of Evaluation: Appraisal in English*, Palgrave Macmillan.

Bo Pang and Lillian Lee (2008) *Opinion Mining and Sentiment Analysis*, Now Publishers.

Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando (2010) “Overview of Multilingual Opinion Analysis Task at NTCIR-8 -A Step Toward Cross Lingual Opinion Analysis.” *Proc. of the Eighth NTCIR Workshop Meeting*, NII, Japan, pp.209-220.

S. Somasundaran, T. Wilson, J. Wiebe, V. Sttoyanov (2007) “QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in Online Discussions and the News”, *Proc. of the International Conference on Weblogs and Social Media (ICWSM) 2007*, Boulder, Colorado, USA.

大塚裕子，乾孝司，奥村学（2007）『意見分析エンジン』，コロナ社.

佐野大樹（2010a）「ブログにおける評価情報の分類と体系化～アプレイザル理論を用いて～」，電子情報通信学会 第 1 回集合知シンポジウム.

佐野大樹（2010b）「評価表現に基づくブログ分類の試み～アプレイザル理論を用いて～」，言語処理学会第 16 回年次大会.

佐野大樹（2011a）「日本語における評価表現の分類体系～アプレイザル理論をベースに ～」，電子情報通信学会 第 2 回集合知シンポジウム.

佐野大樹(2011b)「患者の語りにおける感情表現の使用傾向」, 第 27 回社会言語科学会研究大会.

関根聡, 竹内康介（2007）「拡張固有表現オントロジー」，言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」, pp.23-26, 2007.

関洋平，神門典子，稲垣陽一，栗山和子（2010）「新聞記事とコミュニティ QA を対象とした詳細な意見分析コーパスの作成と分析」，情報処理学会第 97 回情報学基礎研究会・第 195 回自然言語処理研究会合同研究会.

研究活動・成果の総括：日本語フレームネット班
BCCWJ と意味フレームに基づく語彙・構文複合資源の構築

小原京子（班長：慶應義塾大学理工学部）†
斎藤博昭（分担者：慶應義塾大学理工学部）
藤井聖子（分担者：東京大学総合文化研究科）
佐藤弘明（分担者：専修大学商学部）

Final Progress Report: 'Japanese FrameNet' Group

Kyoko Ohara (Faculty of Science and Technology, Keio University)
Hiroaki Saito (Faculty of Science and Technology, Keio University)
Seiko Fujii (Graduate School of Arts and Sciences, The University of Tokyo)
Hiroaki Sato (School of Commerce, Senshu University)

１．研究目的・研究体制

日本語フレームネット班の目的は、フレーム意味論の意味論的枠組みとコーパスからの用例に基づくオンライン語彙資源日本語フレームネット（JFN）を構築していくことにより、「代表性を有する大規模日本語書き言葉コーパス」（BCCWJ）の均衡性・代表性を評価・確認していくことであった。

日本語フレームネット班は第一期（2007 年 4 月～2009 年 3 月）公募班、第二期（2009 年 4 月～2011 年 3 月）公募班として四年間特定領域研究「日本語コーパス」に参加した。それぞれの概要は以下のとおりである。

- 第一期（2007 年 4 月～2009 年 3 月）
  - ・研究課題
    「フレーム意味論とコーパスデータに基づく日本語語彙情報資源『日本語フレームネット』の構築」
  - ・研究目的
    日本語フレームネットの理論的・方法論的モデルの構築
  - ・班長
    斎藤博昭 （慶應義塾大学理工学部） 全体の総括、言語処理
  - ・研究分担者
    藤井聖子 （東京大学総合文化研究科 ） 言語分析
    小原京子 （慶應義塾大学理工学部） 言語分析、アノテーション
- 第二期（2009 年 4 月～2011 年 3 月）
  - ・研究課題
    「BCCWJ と意味フレームに基づく語彙・構文複合資源の構築」
  - ・研究目的
    日本語フレームネットを、語彙の意味情報に加え構文の意味情報をも含む語彙・構文

† ohara@hc.cc.keio.ac.jp

複合資源へと発展させていくためのパイロットスタディ

・班長

小原京子	（慶應義塾大学理工学部）	全体の総括、言語分析、アノテーション

・研究分担者

斎藤博昭	（慶應義塾大学理工学部）	言語処理
藤井聖子	（東京大学総合文化研究科 ）	言語分析
佐藤弘明	（専修大学商学部）	言語処理

第一期の研究目的に関しては、BCCWJ データをアノテーション対象とし、ツールを開発し、アノテーション手順を確立することにより、日本語フレームネットの理論的・方法論的モデルを構築することができた。そして、FrameSQL 上で語彙アノテーションデータを公開することができた。第二期の「語彙の意味情報に加え構文の意味情報をも含む語彙・構文複合資源へと発展させていくためのパイロットスタディ」では、語彙素以外に文の意味に寄与する要素を洗い出し、構文の意味アノテーションへの道筋を整えた。

２．活動内容

日本語フレームネット班ではこれまで主に、1）アノテーション、2）ツールの開発、3）自然言語処理への応用、4）BCCWJ の評価を行ってきた。1）のアノテーションに関しては、語彙アノテーションと全文テキストアノテーションという、二つのモードでフレーム意味論に基づく意味アノテーションを行った。語彙アノテーションでは BCCWJ モニター公開データ 2008 年度版を、全文テキストアノテーションでは BCCWJ コアデータを対象とした。

2）のツールに関しては、FrameSQL と呼ばれるツールで語彙アノテーション・データが検索閲覧できるようにした。さらに、この FrameSQL 上で、日本語フレームネットデータから対応する動詞項構造シソーラス・データを参照できるようにした（佐藤 2010, Sato 2010）。また、全文テキストアノテーション Web Report と、語義分析やアノテーション対象文の選定の際に用いる JFN-KWIC という検索システムを新たに開発した（小原他 2011, Saito et al. 2008, 曽根・小原・斎藤 2010, 曽根・斎藤・小原 2010）。

3）の自然言語処理への応用については、Tagami et al. (2009) などで成果を公表した。

4）の BCCWJ の評価については、その成果を藤井・上垣 (2008)、藤井・内田 (2009) などで発表してきた。

以下では、上記 1) の BCCWJ への意味アノテーション作業結果について、特に日本語固有の意味フレームと構文の意味アノテーションの観点から述べる。第 3 節は語彙アノテーション、第 4 節は全文テキストアノテーションについて報告する。第 5 節は構文の意味アノテーションのパイロットスタディについて報告する。

３．語彙アノテーション

語彙アノテーションとは、語彙項目ごとに BCCWJ の中からアノテーション対象例文を選び、タグ付けしていくモードである（小原 2008a）。日本語フレームネット班では具体的に以下の四つを行った：1）動詞・形容詞・副詞・事態性名詞の、BCCWJ における出現頻

度の高い語彙素（lexeme）から順にアノテーション対象を決める；2）その語彙素が喚起する意味フレーム（言語の発話や理解の際に必要となる、体系的知識構造）を同定する。実際には、英語フレームネット上で英語語彙分析のために既に定義された意味フレームの中から当該日本語語彙素に適用できるものを探す。英語フレームネット上の既成の意味フレームの中に適切なものがない場合には、i) 英語語彙分析に必要であるにもかかわらず、英語フレームネット (http://framenet.icsi.berkeley.edu/) 上で今現在はまだ定義されていないだけなのか、あるいは ii) 英語語彙分析に不要なため英語フレームネット上で定義されていないが、日本語語彙の意味分析のためには新たに定義が必要な意味フレームなのか、を検討する； 3) JFN-KWIC コンコーダンサーを用いて BCCWJ から当該語彙素を検索し、さらにその中から当該意味フレームに関与する例文のみを選別する。さらに、共起語や結合価パターンを考慮しつつ、アノテーション対象とする例文を選定・抽出する；4）抽出した文に、JFNDesktop アノテーションツールを用いて意味フレームに基づく意味情報（フレーム要素。意味フレームの部分を成す意味的要素。Cf. 意味役割）[1]、統語情報、文法情報などのタグを付与する、の四つのプロセスである。

日本語フレームネット・プロジェクトの主要目的の一つは、英語語彙分析のために英語フレームネットで定義された意味フレームが類型論的に異なる日本語の語彙意味記述にどこまで適しているのかを検証していくことである。そこで、語彙アノテーション作業においても、日本語固有の意味フレームやフレーム要素を定義する必要があるのか、あるとすればどのような意味フレームとフレーム要素かを検討してきた。次節で見るように、日本語フレームネットで意味フレームを独自に定義する必要に迫られたケースはこれまでにさほど多くはなかった。むしろ、日英両言語の語彙に関与する意味フレームの違いとは、個々の意味フレームのレベルではなく、体系的な違い、すなわち意味フレームの立て方に関する違いであることがわかってきた。

たとえば、 (1a) のように「散る」という語に関与する意味フレームを同定するケースを考えてみる。まず、この自動詞「散る」に対応する他動詞「散らす」に関与する意味フレームとしては、scatter という英語の対応語を手掛かりに `Dispersal` フレーム（「動作主が起点（狭い空間）から終点（広い空間）へと個体の集合体をばらまく」）が同定できる (1b) 。しかしながら、既存の意味フレームのうち自動詞「散る」に関与していると考えられるのは物の移動に関する一般的なフレーム `Motion` フレーム（「個体が起点を出発し、経路を通ることにより終点に行く」）のみである (1a) 。

(1) a. 桜の花びらが散る$_{\texttt{Motion}}$。

b. 桜の花びらを散らす$_{\texttt{Dispersal}}$。

[1] フレーム意味論ならびにフレームネットにおけるフレーム要素とは、それぞれの意味フレームに依存した具体的な意味的要素である。これに対し、VerbNet (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html) では汎用的な意味役割を付与している。VerbNet の汎用的意味役割を用いて FrameNet と PropBank (http://verbs.colorado.edu/~mpalmer/projects/ace.html) の「意味役割」の汎化を行った研究に松井他（2010）がある。汎用的意味役割の課題については、上代日本語コーパスへの意味アノテーションに基づくホーン（2011）を参照されたい。

他の自動詞「刑死する」とそれに対応する他動詞「処刑する」の場合も同様である (2) 。まず、他動詞「処刑する」に関与する意味フレームとしては、execute という英語の対応語を手掛かりに Execution フレーム（「処刑者がある行為を理由にある個人に罰として死を与える」）が同定できる (2b) 。ところが、既存の意味フレームのうち自動詞「刑死する」に関与していると考えられるのは、死に関する最も一般的なフレーム Death フレーム（「ある参与者の死についての描写」）のみである (2a) 。

(2) a. 死刑囚が刑死する$_{\text{Death}}$。
b. 死刑囚を処刑する$_{\text{Execution}}$。

英語の語彙意味分析のために定義された既存の意味フレームでは、動詞の自他交替に関しては他動詞的観点から定義された意味フレームが比較的多く、自動詞的観点から定義されたものは少ない。上で見た Dispersal フレームと Execution フレームも他動詞的観点から定義されたものである。動詞の自他交替に関して、他動詞的観点と自動詞的観点のそれぞれから意味フレームが定義されているケースはほとんどなく、例外は Detaching フレーム（「動作主がある物を他の物から外す情景」）と Becoming_detached フレーム（「ある物が他の物から外れる情景」）の対くらいである[2]。Filling フレーム（「容器をある物でいっぱいにする」）と Fullness フレーム（「容器がある内容物でいっぱいになっている状態」）も一見それぞれ他動詞的観点と自動詞的観点から定義された意味フレームの対のように見えるが、実際にはそうではない。Filling フレームが他動詞的観点から定義されているのに対し、Fullness フレームの方は自動詞的観点から定義されているのに加えて状態相についての意味フレームでもある。したがって、Filling フレームは他動詞「満たす」に関連していると考えられるが、Fullness フレームは自動詞「満ちる」に関与する意味フレームではない。「満ちる」の語彙的アスペクト（動作相、Aktionsart）は開始相であり、状態相ではないからである[3]。日本語動詞は一般に語彙的アスペクトが開始相のものが多い。

このように既存の意味フレームは当初英語の語彙意味分析のために定義されたものなので、日本語の語彙意味分析に用いようとすると、定義の一部に含まれ前提とされている視点（他動詞的・自動詞的）や語彙的アスペクト（状態相・開始相）が日本語の語彙のそれらとずれていることが多々あることがわかった。今後の語彙アノテーション作業、特にフレーム同定作業ならびにフレーム定義作業においても十分この点を考慮する必要がある。

４．全文テキストアノテーション

全文テキストアノテーションとは、特定のサンプルテキスト内の全ての文の、意味フレ

[2] ちなみに、意味フレーム名の冒頭に “Becoming_”と付くものは自動詞的観点から定義された意味フレームであるが、2011 年 2 月 7 日現在 1034 ある意味フレームのうち、Becoming_a_member フレーム、 Becoming_ aware フレーム、Becoming_ detached フレーム、Becoming_dry フレーム、Becoming_separated フレームの 5 フレームのみである。

[3] 「外す」関連では、前述の Detaching フレーム（他動詞的観点）と Becoming_detached フレーム（自動詞的観点・開始相）に加えて、Being_detached フレーム（状態相）も定義されている。

ームを喚起(evoke)する全ての語彙項目に対してタグ付けをしていくモードを指す。BCCWJ コアデータ書籍ジャンル（総数 84 サンプル）の固有表現以外の語彙項目を対象とした。さらに、全文テキストアノテーション結果を基に、英語フレームネットで既に定義された意味フレームがどこまで日本語の語彙記述にも有効であったかを調べた（cf. 小原 2011）。

各語彙項目の喚起する意味フレームを同定する際の手順は以下の通りである。まず、英語フレームネット上で英語語彙分析のために既に定義された意味フレームの中から、当該日本語語彙項目に適用できるものを探す。もし英語フレームネット上の既成の意味フレームの中に適切なものがない場合には、i) 英語語彙分析に必要だが、英語フレームネット上で今現在はまだ定義されていないだけなのか、あるいは ii) 英語語彙分析に不要なため英語フレームネット上で定義されていないが、日本語語彙の意味分析のためには新たに定義が必要な意味フレームなのか、を検討する[4]。

このようにして、英語フレームネット上で既に定義された意味フレームのうちどの程度が BCCWJ コアデータ書籍ジャンルの語彙記述にも使うことができたかを調べた。その結果、書籍ジャンルのサンプルにおける英語フレームネット上の意味フレームの適合率は平均 82 パーセントであった。適合率の算出に当たっては、異なり語（type）数ではなく延べ語（token）数を用いた。

適切な意味フレームが英語フレームネットで既に定義された意味フレームの中に見つからず、意味フレーム名が付与できなかった語彙項目を品詞別にまとめたのが(3)である。括弧内の数字は 179 語の異なり語のうち、その品詞に分類されるものの数を示す。

(3) 意味フレーム名が付与できなかった語彙項目（品詞別）（括弧内は異なり語数）

a. 形容詞（1 語）

あらい

b. 接続詞（4 語）

だから、しかし、ならば、すなわち

c. 形容動詞（13 語）

好意的、当然、一方的、文字通り、圧倒的、地理的、分野的、のろま、順調、簡単、凄絶、徹底的、科学的

d. 動詞・動詞句（12 語）

悪びれる、過ごす、占める、建ち並ぶ、向きあう、する、潜む、遊ぶ、さし出す、向く、間に合う、気をつける

e. 副詞（21 語）

実際のところ、もちろん、必ずしも、一切、しっかり、ギリギリ、一般に、たとえば、半ば、最も、実際、言い換えれば、はじめて、例えば、多分、単に、少なくとも、たまたま、つかつか、代々、もともと

f. 事態性名詞（22 語）

営業、制御、通行、参照、捨象、仲介、紹介、関連、生活、転勤、出血、観闇、闇遊

[4] ドイツの SALSA プロジェクトにおいても同様の方針がとられている (http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index) 。

び、体験、経験、解剖、出版、棚割、お使い、刑死、牢死、埋葬

g. 名詞（106 語）

支店、肉声、セールス、手づくり、基準、盤、常識、神、精神、精神構造、神霊、海魚、借り、玩具、行き、モノ、聖書、単位、一方、核、スケール、印税、こと、社会、民間、産業主義、かけがえ、ホームレス、毛布、巣、諺、育ち、魚、白豚、国語、紙、体質、良性、悪性、細胞、割合、組織、元気、ぬいぐるみ、犬、ペット、ワン、畳、ソファ、抱っこ、声、心、音楽、闇、画面、モデル、自然、事象、法則、原理、現象、遊び、弾力性、バネ、授業、内容、呪力、レーダー、先、共通、点、民俗、文化、港、岸壁、博士、一説、自分、例、あて字、死体、神仏、内臓、川獺、罪人、前例、構成、ユニット、障子、週別、襖紙、同週、ポスト、門、通い、暮らし、志士、墓、夢、屍骸、土、野犬、墓地、侠客、情況、筋肉

意味フレーム名が付与できなかったこれらの語彙項目のほとんどは上記の i) の「英語語彙分析のためにも必要と考えられるが、フレームネット上で今現在はまだ定義されていない」ケースと考えられ、ii)の「英語語彙分析に不要なため英語フレームネットでは定義されていないが、日本語語彙の意味分析には新たに定義が必要と考えられる意味フレーム」はごく少数にとどまった。すなわち、異なり語 179 語のうち、日本語フレームネットで独自に意味フレームを定義する必要があると考えられた語彙項目は、(3)で下線を付けた 4 語（「畳」、「障子」、「襖紙」、「侠客」の名詞）のみであった。

i) の、「英語語彙分析にも必要と考えられるが、フレームネット上で今現在はまだ定義されていない」ため意味フレームが付与できなかった語彙項目の中には、接続詞(3b)や副詞(3e)が含まれている。これは、英語フレームネットでは副詞や接続詞のアノテーションがまださほど進んでいないことが原因と考えられる。さらに、副詞(3e)の中には、「実際のところ」、「実際」、「もちろん」、「言い換えれば」などのメタテキスト的な文副詞が多く含まれている。これらの語彙項目の意味や用法は文の談話的状況と密接に関係しているので、これらが喚起する意味フレームやそのフレーム要素を意味論的に定義するのは容易ではないであろうと予測できる。一方、意味フレームが付与できなかった形容動詞(3c)、動詞・動詞句(3d)、事態性名詞(3f)、名詞(3g)が関与する意味分野は多岐にわたることがわかった。

５．構文の意味アノテーションに向けて

日本語フレームネットの理論的バックボーンであるフレーム意味論とはそもそも、文全体の意味と関連づけて文中に現れる語彙項目の意味を理解しようとするものである。しかしながら、文の意味とはそこに現れる語彙項目の意味の和のみで成り立っているのではない。語彙項目以外の様々なレベルの「構文」の意味も文全体の意味に関与している。このような文の意味のとらえ方は、日本語フレームネットのもう一つの理論的枠組みである構文文法にも共通している (Fillmore & Baker 2009, 小原 2010)。実際に今現在英語フレームネット・プロジェクトでは、フレーム意味論と構文文法に基づき、語彙の意味と構文の意味の統一的な記述のための枠組みを検討しているところである。

このような背景から、日本語フレームネット班の公募班第二期（2009 年 4 月～2011 年 3 月）の目的は、「日本語フレームネットを、語彙の意味情報に加え構文の意味情報をも含む

語彙・構文複合資源へと発展させていくためのパイロットスタディ」とした (cf. 小原 2008b, Ohara 2008) 。そして、全文テキストアノテーション作業の過程で、語彙素以外に文全体の意味に寄与している要素にどのようなものがあるかを調査した。その結果、いわゆる短単位の基本語、複合名詞、複合動詞以外で文全体の意味に貢献するものとしては、大きく分けて以下の四種類があることがわかった(4)。

(4) a. 支援動詞構文[5]

例：問題にする、気になる、電話がかかる、声をかける、気をつける、耳にする

b. 複合辞・助詞相当句[6]

例：という、として、に関して、にとって、にわたって、によると、というと、といって、ところによると、の通り、など。

c. 複合辞・助動詞相当句

例：ことがない、ていく、ほどである、とする、た方がいい、とされている、てみる、からだ、のようだ、というのだ、方がいい、ばいい、ことになる、というわけだ、ものでもない、ように言う、という、たらいい、はずである、による、かもしれない、なくてはならない、ことがある、のことだ

d. 定型的な表現

例：よりももっと…、しか…ない、のは…のことだ、方がより…、というのは…である、なかには…がある

これらの四種類の表現の BCCWJ コアデータ書籍ジャンルからの例を、それらが喚起する意味フレーム名とともに (5) に挙げる。下線部分が冒頭に名称を記した意味フレームを喚起する要素 (Frame Evoking Element (FEE)) である。

(5) a. 支援動詞構文

`Perception_experience` フレーム：

尾張出身の人ならば、反対せられそうなものと気をつけているが、まだそういう話も耳にしないので、単に一説として自分の知っていることだけを並べてみる。(「毎日の言葉」)

b. 複合辞・助詞相当句

---

[5] ここでは「支援動詞構文」を広くとらえ、動詞ではなく名詞が第一義的な意味フレーム喚起要素（Frame Evoking Element: FEE）であるような構文を指す(cf. 藤井・上垣 2008, 上垣・藤井 2008)。動詞の項となる名詞は事態性名詞以外でもよく（例：「気」、「耳」など）また名詞は本動詞の直接目的語以外のものも含める（「中心にする」、「気になる」、「電話がかかる」など)。

[6] (4b) と (4c) で下線を付けた表現は、BCCWJ で今現在認められている助詞相当句 21 語、助動詞相当句 39 語の中に含まれているものである（富士池他 2008, 富士池他 2010)。全文テキストアノテーション作業では BCCWJ コアデータ書籍ジャンルに出現するフレーム喚起要素 (FEE) はすべてアノテーション対象としているため、BCCWJ 上での出現頻度を考慮に入れ選定された助詞相当句と助動詞相当句の集合に含まれないものもアノテーション対象となっている。

Attributed_information フレーム：

その報告によると、ＯＥＣＤに加盟している先進諸国三十カ国においては、製造業の生産高は二倍になるが、製造業に従事する労働者の割合は多くても十パーセント、少ない国では二パーセントにまで激減すると予想しています。(「教養教育は進化する」)

c. 複合辞・助動詞相当句

Attempt_suation フレーム：

日本的な神霊常識も、嘘ではありませんが、そういうものにこだわるよりも、もっと大きい、全世界的な神をしっかりつかまえる方法をとればいいのです。(「とこしえの命を得るために」)

Unattributed_information フレーム：

自宅横には親方が経営していた文化住宅「玉ノ海荘」もあるという。(「尼崎相撲ものがたり」)

d. 定型的な表現

Inclusion フレーム：

ほとんどのホクロ・黒アザは良性ですが、なかには悪性のものがあります。(「医師による切らない『赤アザ・赤ら顔(浮きでた青い血管)』の最新治療」)

以上から、日本語フレームネットを語彙意味情報と構文意味情報の両者を含む語彙・構文複合資源へと発展させていくためには、今後は特に上記 (4d) のタイプ、すなわち定型的な表現を網羅的に収集し、それらの表現の意味と文全体の意味との関連を明らかにしていく必要があることが今期のパイロットスタディで明らかとなった。

６．まとめ

日本語フレームネット班の公募班としての四年間の研究活動と成果について、主にBCCWJ への意味アノテーションの観点から報告した。特に、英語フレームネット上の意味フレームの日本語語彙意味記述への適用性について語彙アノテーションと全文テキストアノテーションの両方から考察した。

第一期には、BCCWJ データをアノテーション対象とし、ツールを開発し、アノテーション手順を確立することにより、目的であった日本語フレームネットの理論的・方法論的モデルを構築することができた。第二期には、「語彙の意味情報に加え構文の意味情報をも含む語彙・構文複合資源へと発展させていくためのパイロットスタディ」において、語彙素以外に文の意味に寄与する要素を洗い出し、構文の意味アノテーションへの道筋を整えることができた。

付記

日本語フレームネット班究活動中に多大なるご協力をいただいた研究協力者の方々、特に木越壽子氏、田上隼人氏、久保谷俊太氏、曽根孝明氏、李陽氏、加藤淳也氏、並びに前木香織氏とアレクサンドル・カバッシュ氏に御礼申し上げる。

文献

上垣渉、藤井聖子 (2008).「日本語支援動詞構文におけるイディオム性と規則性」『言語処理学会第 14 回年次大会予稿集』pp.845-848.

小原京子 (2008a).「日本語フレームネットのアノテーション体系」,『特定領域研究「日本語コーパス」平成 19 年度公開ワークショップ（研究成果報告会）予稿集』 pp. 203-210.

小原京子 (2008b).「日本語フレームネットにおける語彙と構文の意味 : パラレルコーパスの比較対照分析から」,『言語処理学会第 14 回年次大会予稿集』pp.857-860.
（http://jfn.st.hc.keio.ac.jp/publications.html よりダウンロード可能）

小原京子 (2010).「フレームネットにみるフレーム意味論と構文文法」日本英文学会関西支部　第 5 回大会シンポジウム『構文文法の現在と未来』大阪市立大学, 2010 年 12 月 18 日.

小原京子 (2011).「日本語フレームネットの全文テキストアノテーション : BCCWJ への意味フレーム付与の試み」,『言語処理学会第 17 回年次大会予稿集』

小原京子、加藤淳也、斎藤博昭 (2011).「日本語フレームネットにおける BCCWJ への意味アノテーション」, 『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』

佐藤弘明 (2010).「FrameSQL で利用する日本語フレームネット」,『特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』pp. 143-146.

曽根孝明、小原京子、斎藤博昭 (2010).「『現代日本語書き言葉均衡コーパス』を対象とした全文検索システム」,『言語処理学会第 16 回年次大会発表論文集』pp. 506-509.

曽根孝明、斎藤博昭、小原京子 (2010).「『現代日本語書き言葉均衡コーパス』を対象とした全文検索システム JFN-KWIC 」,『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. 125-130.

藤井聖子、上垣渉 (2008).「支援動詞構文における事態性名詞と動詞との項共有と連結性―『日本語コーパス』を用いた分析―」日本言語学会第１３６回大会.

藤井聖子、内田諭 (2009).「フレーム間関係を用いた日英語の語彙分析 ―「伝達」「判断」フレームの場合 ―」,『言語処理学会第 15 回年次大会予稿集』

富士池優美、小椋秀樹、小木曽智信、小磯花絵、内元清貴 (2008).「『現代日本語書き言葉均衡コーパス』における長単位の概要」『特定領域研究「日本語コーパス」平成１９年度公開ワークショップ（研究成果報告会）予稿集』pp. 51-58.

富士池優美、小西光、小椋秀樹、小木曽智信、小磯花絵 (2010).「『現代日本語書き言葉均衡コーパス』長単位情報に基づく予備的分析」『特定領域「日本語コーパス」平成２２年度全体会議予稿集』pp. 101-108.

ホーン、スティーブン・ライト(2011).「上代日本語のコーパスにおける意味役割の付与」国立国語研究所. 2011 年 1 月 20 日.

松井優一郎、岡崎直観、辻井潤一 (2010).「自動意味役割付与における意味役割の汎化」『自然言語処理』pp.59-89. Vol.17, No. 4.

Fillmore, Charles J. and Collin Baker (2010). “A frames approach to semantic analysis.” In Heine, Bernd and Heiko Narrog (Eds.) *The Oxford Handbook of Linguistic Analysis*. pp. 313-339. Oxford University Press.

Ohara, Kyoko Hirose (2008). “Lexicon, Grammar, and Multilinguality in the Japanese FrameNet.” *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC2008). (http://jfn.st.hc.keio.ac.jp/publications.html よりダウンロード可能)

Saito, Hiroaki, Shunta, Kuboya, Takaaki, Sone, Hayato, Tagami, Kyoko, Ohara (2008). “The Japanese FrameNet Software Tools.” *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC2008). (http://jfn.st.hc.keio.ac.jp/publications.html よりダウンロード可能)

Sato, Hiroaki (2010). “How FrameSQL Shows the Japanese FrameNet Data.” *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010) , USB メモリー収録のためページ番号無し.

Tagami, Hayato, Shinsuke Hizuka, and Hiroaki Saito (2009). “Automatic Semantic Role Labeling based on Japanese FrameNet - Progress Report -.” *Proceedings of Conference of the Pacific Association for Computational Linguistics* (PACLING2009), pp.181-186. (http://jfn.st.hc.keio.ac.jp/publications.html よりダウンロード可能)

関連 URL

日本語フレームネットホームページ：http://jfn.st.hc.keio.ac.jp/ja/index.html

FrameSQL ホームページ： http://sato.fm.senshu-u.ac.jp/jfn23/notes/index2.html

デモ・ポスターセッション

3月15日（火） 11:50〜14:00

『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概要（5）―サンプリングの最終結果―
▶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』における評価表現の分布
―「日本語アプレイザル評価表現辞書（態度表現編）」を用いて―
▶佐野 大樹、柏野 和佳子

Yahoo！知恵袋の質問における修辞機能の分布 ―修辞ユニット分析を用いて―
▶田中 弥生、佐野 大樹

『現代日本語書き言葉均衡コーパス』向け外字処理ツール
▶田島 孝治、高田 智和

長単位に基づく媒体・カテゴリ間の品詞比率に関する分析
▶冨士池 優美、小西 光、小椋 秀樹、小木曽 智信、小磯 花絵

BCCWJに基づくオノマトペの品詞と意味についての分析
▶宮内 佐夜香、小木曽 智信、小磯 花絵、小椋 秀樹

Web版コーパス検索アプリケーション「中納言」のデモンストレーション
▶中村 壮範、小木曽 智信

階層的形態論情報を考慮した『現代日本語書き言葉均衡コーパス』の公開用XMLフォーマット
▶小木曽 智信、間淵 洋子、前川 喜久雄

汎用アノテーションツールSlate
▶徳永 健伸、Dain Kaplan、飯田 龍

BCCWJと関連ツールの相互運用
▶狩野 芳伸、橋田 浩一

拡張固有表現タグ付きコーパスの構築
▶橋本 泰一

BCCWJコアデータへの係り受け・並列構造のアノテーション
▶浅原 正幸、岩立 将和、松本 裕治

BCCWJに対する述語項構造と照応関係のアノテーション
▶小町 守、飯田 龍

BCCWJに基づく中・長単位解析ツール
▶小澤 俊介、内元 清貴、伝 康晴

UniDicを用いた音声認識用言語モデルの作成
▶山田 篤

作文コーパスからみる生徒の使用語彙
▶鈴木 一史、棚橋 尚子、河内 昭浩

学習データ間距離学習に基づく語義識別の性能分析
▶佐々木 稔、新納 浩幸

コーパス管理・検索ツール「茶器」
▶松本 裕治、浅原 正幸、岩立 将和、森田 敏生

『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概要 (5)
—サンプリングの最終結果—

丸山 岳彦　（データ班分担者：国立国語研究所 言語資源研究系）†
山崎 誠　（データ班班長：　国立国語研究所 言語資源研究系）
柏野 和佳子　（データ班分担者：国立国語研究所 言語資源研究系）
佐野 大樹　（データ班分担者：国立国語研究所 コーパス開発センター）
秋元 祐哉　（データ班協力者：国立国語研究所 コーパス開発センター）
稲益 佐知子　（データ班協力者：マンパワー・ジャパン株式会社）
田中 弥生　（データ班協力者：国立国語研究所 コーパス開発センター）
大矢内 夢子　（データ班協力者：国立国語研究所 コーパス開発センター）

Outline of Sampling Method in the Balanced Corpus of Contemporary Written Japanese (5) : Final Result of Sampling in BCCWJ

Takehiko Maruyama　(National Institute for Japanese Language and Linguistics)
Makoto Yamazaki　(National Institute for Japanese Language and Linguistics)
Wakako Kashino　(National Institute for Japanese Language and Linguistics)
Motoki Sano　(National Institute for Japanese Language and Linguistics)
Masaki Akimoto　(National Institute for Japanese Language and Linguistics)
Sachiko Inamasu　(Manpower Japan)
Yayoi Tanaka　(National Institute for Japanese Language and Linguistics)
Yumeko Oyauchi　(National Institute for Japanese Language and Linguistics)

1　導入

2006 年度に『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ と記す）』の構築が開始されてから、5 年が経過した。この間、我々のグループ（サンプリングサブグループ）では、BCCWJ を構成する 3 つのサブコーパス（以下 SC と記す）「出版 SC」「図書館 SC」「特定目的 SC」の設計、およびサンプリングの実作業を担当してきた。2011 年 1 月現在、当初の設計方針に基づいて継続してきたサンプリング作業はすべて完了している。

本稿では、これまでに報告してきたサンプリングの進捗状況 (丸山ほか, 2007, 2008, 2009, 2010) の総括として、BCCWJ におけるサンプリングの最終結果を示す。また、これまで詳細を報告してこなかった「特定目的 SC」の設計とサンプリングの最終結果についても示す。さらに、サンプリングと並行して進めてきた「書誌情報データ」の設計と実装についても報告する。

2　「出版 SC」「図書館 SC」の設計とサンプリングの最終結果

2.1　「出版 SC」「図書館 SC」の設計

「出版 SC」「図書館 SC」の設計については、これまでの進捗状況報告、および報告書 (丸山・秋元, 2007, 2008; 柏野ほか, 2009) の中で繰り返し述べてきた。その要点を、以下に挙げる。

†maruyama@ninjal.ac.jp

- 「出版 SC」は、2001 年から 2005 年までに国内で発行された書籍・雑誌・新聞を対象とし、そこに含まれる総文字数（推計 65,471,677,099 文字）によって母集団を定義する。
- 「図書館 SC」は、1986 年から 2005 年までに国内で発行された書籍のうち、東京都内 13 自治体以上の公立図書館で共通に所蔵されている書籍を対象とし、そこに含まれる総文字数（推計 47,877,656,072 文字）によって母集団を定義する。
- 母集団を「ジャンル」「発行年」によって層別し、層別ランダムサンプリングを実施する。
- 母集団の中からランダムに指定された 1 文字を「サンプル抽出基準点」とし、そこから 1,000 文字の範囲を「固定長サンプル」として、その点を含む章や節のまとまりを「可変長サンプル」として、それぞれ取得する。
- 「出版 SC」の固定長サンプルを 1,000 万語取得することを基準として、各層に含まれる文字数の比例割当により、各層から取得するサンプル数を定める。

上記の方針に基づき、取得するサンプル数とそこから得られる固定長サンプル・可変長サンプルの語数を、表 1 のように試算した。この際、可変長サンプルの平均文字数を、書籍で 3,900 文字、雑誌で 3,000 文字、新聞で 1,000 文字と仮定した。また、1 語は 1.7 文字で構成されると仮定した。

表 1: 出版 SC・図書館 SC の設計（取得サンプル数・取得語数）

SC	メディア	サンプル数	固定長サンプル語数	可変長サンプル語数
出版 SC	書籍	12,604	7,414,118	28,915,059
	雑誌	2,730	1,605,882	4,817,647
	新聞	1,666	980,000	980,000
	合計	17,000	10,000,000	34,712,706
図書館 SC	書籍	12,604	7,414,118	28,915,059

この試算により、出版 SC では約 3,500 万語、図書館 SC では約 3,000 万語が取得できることになり、特定目的 SC の約 3,500 万語と合計して、BCCWJ 全体を構成する語数である「1 億語」を達成することができると見積もった。

2.2 作業の進捗に伴う設計の見直し

2006 年度からサンプリングの設計を開始し、以降 5 年間、ランダムに選ばれた書籍・雑誌・新聞を入手してサンプルを抽出する作業を継続した。この結果が電子テキスト化され、サンプルの数が蓄積されることにより、可変長サンプルの平均文字数について正確な見積もりが得られるようになった。これによると、可変長サンプルの平均文字数は、書籍で平均 4,534 文字、雑誌で平均 3,873 文字、新聞で平均 980 文字となり、新聞を除いて当初の見積もりを上回る結果となった。このため、設計通りに出版 SC で 17,000 サンプル、図書館 SC で 12,604 サンプルを取得すると、可変長サンプル全体の語数が大幅に増大してしまう見込みとなった。そこで、当初の設計の 80%が達成されていることを最低条件として、当初に見積もった取得サンプル数を下方修正した。

また、当初から予想されたことであるが、著作権処理の過程において著作権者から利用を拒否する旨の回答が来たため、公開することができなくなったサンプルが多数生じた。そこで、サンプリング作業の進捗にあわせて各層の「許諾率」を計算し、許諾率の低い層からは当初の計画より多めにサンプルを取得するよう調整しながら作業を進めた。

2.3 サンプリング作業の完了と最終結果

サンプリング作業の完了が近づくにつれて、当初に設計した構成比になるべく近似するように、各層から取得するサンプル数を細かく調整した。書籍・雑誌・新聞のメディア別、ジャンル別、発行年別に層を分けた上で、各層の構成比、および許諾率を計算し、当初の設計から不足している層には必要な数のサンプルを補填した。全体の構成比を見極めながら微調整を進め、2010 年 5 月をもって、当初に設計した構成比に可能な限り近似させた形で、サンプリング作業を完了することができた。

サンプリングの最終結果から、表 1 に相当する部分のみを示すと、表 2 のようになる。

表 2: 出版 SC・図書館 SC のサンプリング結果（取得サンプル数・取得語数）

SC	メディア	サンプル数	固定長サンプル語数	可変長サンプル語数
出版 SC	書籍	11,212	6,595,294	29,541,361
		(89.0%)	(89.0%)	(102.2%)
	雑誌	2,483	1,460,588	5,687,485
		(91.0%)	(90.9%)	(118.0%)
	新聞	1,490	876,471	864,364
		(89.4%)	(89.4%)	(88.1%)
	合計	15,185	8,932,353	36,093,211
		(89.3%)	(89.3%)	(104.0%)
図書館 SC	書籍	11,242	6,612,941	30,053,412
		(89.2%)	(89.2%)	(103.9%)

※ 下段は当初の設計に対する達成率

最終的に取得したサンプル数は、当初の設計に対して、出版 SC の書籍で 89.0%、雑誌で 91.0%、新聞で 89.4%、図書館 SC の書籍で 89.3%、という結果になった。全体の取得サンプル数を算出する基準とした、「出版 SC」の固定長サンプルを 1,000 万語取得するという点については、最終的には 89.3%の約 893 万語となった。図書館 SC においても、89.2%というほぼ同等の結果となった。一方、可変長サンプルの語数は、当初の設計に対して、出版 SC の書籍で 102.2%、雑誌で 118.0%、新聞で 88.1%、図書館 SC の書籍で 103.9%という結果になり、新聞のみ設計を下回ったものの、全体的には当初の設計を上回る語数が得られた。

サンプリングの設計時におけるサンプル数と語数の試算、およびその最終結果ついて、出版 SC・図書館 SC のジャンル別に、表 3、4 に示す。表中の「S」は「サンプル」を表わす。実際には「ジャンル」だけでなく「発行年」も含めた層別が実施されているが、ここでは省略する。サンプリングの最終的な結果の詳細については、最終報告書 (丸山ほか, 2011) を参照していただきたい。

2.4 著作権処理と公開サンプル数

先述のとおり、取得した全サンプルのうち、公開対象となるのは著作権処理を経て公開可能と判断されたもののみであり、表 2 に示したすべてのサンプルが公開されるわけではない。したがって、公開サンプル数は表 2 の数値を下回ることになる。特に雑誌については、一定量のサンプルを取得した後、特定の出版社が出版した雑誌のすべてについて利用を拒否する旨の連絡が来たケースもあった。雑誌の達成率が他のメディアに比べて若干高いのは、その分を補正したことが理由である。

現在、著作権処理の最終的な結果を取りまとめている段階であるため、最終的に公開されるサンプル数は確定していない。メディア・ジャンル・発行年によって層別された各層の公開サンプル数は、近日中に確定する予定である。

表 3: サンプリングの設計時におけるサンプル数と語数の試算、およびその最終結果（出版 SC）

メディア	ジャンル	設計時					最終結果				
		S 数	固定長 S 語数	構成比	可変長 S 平均文字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均文字数	可変長 S 語数
書籍	0. 総記	425	250,000	2.5%	3,900	975,000	363	213,529	2.4%	3,902	833,197
	1. 哲学	674	396,471	4.0%	3,900	1,546,235	610	358,824	4.0%	4,155	1,490,930
	2. 歴史	1,117	657,059	6.6%	3,900	2,562,529	926	544,706	6.1%	4,493	2,447,545
	3. 社会科学	3,222	1,895,294	19.0%	3,900	7,391,647	2,721	1,600,588	17.9%	4,495	7,194,570
	4. 自然科学	1,316	774,118	7.7%	3,900	3,019,059	1,119	658,235	7.4%	4,021	2,646,734
	5. 技術工学	1,199	705,294	7.1%	3,900	2,750,647	1,008	592,941	6.6%	4,127	2,447,023
	6. 産業	570	335,294	3.4%	3,900	1,307,647	480	282,353	3.2%	4,366	1,232,742
	7. 芸術	846	497,647	5.0%	3,900	1,940,824	728	428,235	4.8%	4,225	1,809,129
	8. 言語	231	135,882	1.4%	3,900	529,941	198	116,471	1.3%	4,001	466,008
	9. 文学	2,426	1,427,059	14.3%	3,900	5,565,529	2,557	1,504,118	16.8%	5,070	7,625,880
	n. 記録なし	578	340,000	3.4%	3,900	1,326,000	502	295,294	3.3%	4,564	1,347,602
	小計	12,604	7,414,118	74.1%	—	28,915,059	11,212	6,595,294	73.8%	—	29,541,361
雑誌	1. 総合	1,927	1,133,529	11.3%	3,000	3,400,588	1,786	1,050,588	11.8%	3,914	4,111,719
	2. 教育	228	134,118	1.3%	3,000	402,353	193	113,529	1.3%	4,163	472,600
	3. 政治	119	70,000	0.7%	3,000	210,000	114	67,059	0.8%	3,105	208,197
	4. 産業	29	17,059	0.2%	3,000	51,176	25	14,706	0.2%	2,258	33,200
	5. 工業	381	224,118	2.2%	3,000	672,353	323	190,000	2.1%	4,159	790,200
	6. 厚生	47	27,647	0.3%	3,000	82,941	42	24,706	0.3%	2,897	71,569
	小計	2,730	1,606,471	16.1%	—	4,819,412	2,483	1,460,588	16.4%	—	5,687,485
新聞	全国紙	628	369,412	3.7%	1,000	369,412	550	323,529	3.6%	1,069	345,956
	ブロック紙	337	198,235	2.0%	1,000	198,235	305	179,412	2.0%	903	162,057
	地方紙	702	412,941	4.1%	1,000	412,941	635	373,529	4.2%	954	356,351
	小計	1,666	980,588	9.8%	—	980,588	1,490	876,471	9.8%	—	864,364
合計		17,000	10,000,000	100%	—	34,715,059	15,185	8,932,353	100%	—	36,093,211

表 4: サンプリングの設計時におけるサンプル数と語数の試算、およびその最終結果（図書館 SC）

メディア	ジャンル	設計時					最終結果				
		S 数	固定長 S 語数	構成比	可変長 S 平均文字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均文字数	可変長 S 語数
書籍	0. 総記	263	154,706	2.1%	3,900	603,353	249	146,471	2.2%	4,108	601,669
	1. 哲学	617	362,941	4.9%	3,900	1,415,471	560	329,412	5.0%	4,452	1,466,585
	2. 歴史	1,321	777,059	10.5%	3,900	3,030,529	1,133	666,471	10.1%	4,587	3,056,778
	3. 社会科学	2,356	1,385,882	18.7%	3,900	5,404,941	2,195	1,291,176	19.5%	4,427	5,716,463
	4. 自然科学	797	468,824	6.3%	3,900	1,828,412	663	390,000	5.9%	4,315	1,682,878
	5. 技術工学	828	487,059	6.6%	3,900	1,899,529	690	405,882	6.1%	3,983	1,616,570
	6. 産業	444	261,176	3.5%	3,900	1,018,588	380	223,529	3.4%	4,274	955,392
	7. 芸術	1,070	629,412	8.5%	3,900	2,454,706	897	527,647	8.0%	4,107	2,167,036
	8. 言語	252	148,235	2.0%	3,900	578,118	217	127,647	1.9%	3,348	427,326
	9. 文学	4,076	2,397,647	32.3%	3,900	9,350,824	3,765	2,214,706	33.5%	5,063	11,212,003
	n. 記録なし	583	342,941	4.6%	3,900	1,337,471	493	290,000	4.4%	3,968	1,150,711
合計		12,607	7,415,882	100%	—	28,921,941	11,242	6,612,941	100%	—	30,053,412

3 「特定目的 SC」の設計とサンプリングの最終結果

3.1 「特定目的 SC」に収録されるメディア

「特定目的 SC」は、「出版 SC」「図書館 SC」の母集団には入らないけれども，書き言葉の研究を遂行する上で必要と思われる種類の書き言葉を収める SC である。「特定目的 SC」に含まれるメディアは、「白書」「教科書」「広報紙」「ベストセラー」「Yahoo!知恵袋」「Yahoo!ブログ」「韻文」「法律」「国会会議録」の 9 種類である。

このうち、「白書」「教科書」「広報紙」「法律」は公的な性格の強い書き言葉であり、これらの分析により言語政策に関わる基礎資料を提供することが期待できる。「ベストセラー」はあらゆる書籍の中で特に多くの人に読まれたものであり、出版の実態を反映する「出版 SC」の書籍、流通の実態を反映する「図書館 SC」の書籍に対して、一般読者に受容された実態を反映する資料として考えることができる。「Yahoo!知恵袋」「Yahoo!ブログ」は一般人が書いたウェブ上の書き言葉であり、そこに見られる文字遣い・言葉遣いを収集することにより、ウェブ上の書き言葉が持つさまざまな変異のありさまを捉えることができる。「韻文」は、短歌・俳句・詩という、通常の書き言葉（いわゆる文章）とは異なるスタイルを持つ書き言葉であり、現代日本語の書き言葉の重要な一部を構成するものとして採録することにした。「国会会議録」は、国会における会議での発言を書き起こしたテキストである。そもそも書き言葉として執筆されたテキストではないものの、「会議録」自体は書き言葉の一種であることから、書き言葉のバリエーションの 1 つとして採録することにした。

3.2 「特定目的 SC」の設計とサンプリングの結果

「特定目的 SC」では、「出版 SC」や「図書館 SC」とは異なり、サンプルの取得対象（母集団）は定められているものの、すべてのメディアにおいてそれらが数量的に定義されているわけではない。また、その性格上、母集団の数量的な実態を忠実に反映するようなサンプリングは必ずしも実施されているわけではない。さらに、「白書」を除いて、取得されているのは可変長サンプルのみである。

また、「出版 SC」や「図書館 SC」ではサンプルの取得元（原本）はすべて印刷物であったが、「特定目的 SC」のうち「Yahoo!知恵袋」「Yahoo!ブログ」「法律」「国会会議録」については、既存の電子データからサンプルを取得した。

「特定目的 SC」に収録されたメディアの種類と、その対象期間、取得対象、取得したサンプル数、取得した語数について、表 5 に示す。なお、語数は推計値である。また、今後の著作権処理の進捗などの事情により、最終的な公開サンプル数は変動する可能性がある。

表 5: 「特定目的 SC」の構成

メディア	対象期間	取得対象	S 数	可変長 S 語数	取得元の媒体
白書	1976 年–2005 年	1,006 冊	1,500	500 万語	印刷物
教科書	2005 年–2007 年	145 冊	483	120 万語	印刷物
広報紙	2008 年	100 自治体	355	400 万語	印刷物
ベストセラー	1976 年–2005 年	951 冊	1,408	371 万語	印刷物
Yahoo!知恵袋	2004 年–2005 年	3,120,839 質問	91,450	1,000 万語	電子データ
Yahoo!ブログ	2008 年–2009 年	3,463,413 記事	52,680	1,000 万語	電子データ
韻文	1986 年–2005 年	130 冊	253	15 万語	印刷物
法律	1976 年–2005 年	718 法律	348	100 万語	電子データ
国会会議録	1976 年–2005 年	32,925 会議	159	500 万語	電子データ

以下、各メディアごとの設計とサンプリングの経過について概略を示す。

白書： 対象を「1976 年から 2005 年までの 30 年間に発行されたすべての白書」と定め、官報の記載などから合計 40 タイトル、1,006 冊の白書を特定した。取得語数は、全体で 500 万語とした。1976 年から 2005 年までを 5 年刻みで 6 期に分割し、各期から 250 サンプルずつ、計 1,500 サンプルを取得することにした。各期に含まれる全ページに対してランダムに優先順位を割り振り、順位の高い順から指定されたページを開け、そこに印刷されている文章を一定の手続きにより取得した。

また、白書の各タイトルを、9 種類のジャンルに分類した。結果、1,500 サンプルのジャンルごとの内訳は、「安全」が 24.3%、「外交」が 8.3%、「科学技術」が 8.1%、「環境」が 6.7%、「教育」が 1.6%、「経済」が 19.7%、「国土交通」が 11.8%、「農林水産」が 7.3%、「福祉」が 12.1%となった。

教科書： 対象を「2005 年度に実際に使用された検定教科書」と定め、小学校・中学校・高校の各学年・各教科からできるだけ発行部数の多い順に 1 種ずつの教科書を選出し、145 冊の教科書を特定した。この教科書に印刷されている総文字数を推計したところ、7,859,456 文字という結果を得た。この文字数を、「教科（10 分類）」と「学校（3 分類：小・中・高）」で区分した層に比例割当し、各層の構成比を定めた。各層に含まれる全ページに対してランダムに優先順位を割り振り、順位の高い順から指定されたページを開け、そこに印刷されている文章を一定の手続きにより取得した。

取得された 483 サンプルの教科ごとの内訳は、「国語」が 15.1%、「数学」が 8.3%、「理科」が 20.3%、「社会」が 23.6%、「外国語」が 5.2%、「技術家庭」が 4.1%、「芸術」が 15.7%、「保健体育」が 5.4%、「情報」が 1.7%、「生活」が 0.6%となった。また、学校別の内訳は、小学校が 23.8%、中学校が 17.0%、高校が 59.2%となった。

広報紙： 対象を「地方自治体で 2008 年に発行された広報紙」と定めた。人口比などを考慮して全国から 100 の自治体（区市町村）をサンプリングし、そこで発行された広報紙を入手した。1 自治体から 6 万字程度を取得することとし、発行された広報紙をランダムに選んでその全文を取得した。

取得された 355 サンプルの地域ごとの内訳は、北海道が 5.9%、東北地方が 6.8%、関東地方が 32.7%、中部地方（北陸・甲信越）が 9.9%、中部地方（東海）が 9.6%、近畿地方が 18.6%、中国地方が 4.8%、四国地方が 3.9%、九州地方が 7.9%となった。

ベストセラー： 対象を「1976 年から 2005 年の各年において、『出版年鑑』または『出版指標年報』のどちらかにベストセラーとして上位 20 位までに挙げられた書籍」とした。調査の結果、合計 951 冊の書籍を同定した。これらの原本をできる限り入手し、各冊から 2 サンプルずつを取得した。2 か所のページをランダムに指定して、そこに印刷されている文章を一定の手続きにより抽出した。

Yahoo!知恵袋: ヤフー株式会社より提供された、Web 上のナレッジコミュニティサービス「Yahoo!知恵袋」における Q&A 形式のデータを対象とした。元データには 2004 年 10 月から 2005 年 10 月にかけて投稿された 3,120,839 の質問と、それに対する複数の回答が含まれていた。また、すべての質問は 15 個の大カテゴリ、82 個の中カテゴリ、279 個の小カテゴリに分類されていた。

取得語数は全体で 1,000 万語とし、元データから 91,450 サンプルを取得した。1 サンプルは、1 つの質問と「ベストアンサー」と呼ばれる 1 つの回答の組で構成した。全体の構成比は、小カテゴリに含まれる質問数の比を取得するサンプル数に比例割当して決定した。

Yahoo!ブログ： ヤフー株式会社より提供された、「Yahoo!ブログ」の記事データを対象とした。元データには、2008 年 4 月 26 日から 2009 年 4 月 25 日までに投稿された 3,493,413 記事が含まれていた。これらの記事は、1,000 記事以上の投稿があるブログであること、抽出時点で 1ヶ月以上掲載されていること、全角で 21 文字以上の本文を持つ記事であること、という条件を満たすものである。

取得語数は全体で 1,000 万語とした。元データを投稿の日時順にソートし、全体の 1.8%を等間隔抽出によって抽出し、52,680 サンプルを取得した。さらに、抽出した記事中に含まれる個人情報などに対しては、伏せ字処理を実施した。

韻文： 「韻文」として、短歌・俳句・詩の 3 種類を取得することにした。短歌は『現代短歌全集』（筑摩書房、2002 年刊）の第 14 巻〜第 17 巻を、俳句は『増補現代俳句大系』（角川書店、1980 年〜1982 年刊）の第 8 巻〜第 15 巻を、詩は「現代詩文庫」シリーズ（思潮社、1986 年〜2005 年刊）の 118 冊を、それぞれ対象とした。なお、『現代短歌全集』は昭和 34 年から昭和 63 年の間に発表された歌集、『増補現代俳句大系』は昭和 25 年から昭和 54 年の間に発表された句集を集めたものである。

著作権処理の結果、60 の歌集、92 の句集、101 の詩集を利用できることになった。短歌・俳句・詩でそれぞれ 5 万語ずつを取得することとし、各作品からほぼ等量ずつのサンプルを抽出した。

法律： 1976 年から 2005 年までの間に公布され、2009 年 9 月の時点でも施行されている法律を対象とした。Web 上の「法令データ提供システム」（`http://law.e-gov.go.jp/`）から 718 法律をダウンロードし、対象データとした。ここから 100 万語を取得することとし、公布年（1976 年から 2005 年まで）を 5 年刻みで 6 期に分割して各期から 30 万文字を取得することにした。各期の全法律からランダムに 1 文字を指定し、その文字を含む一定の範囲（1 万字以内）を取得した。

取得された 348 サンプルを、法務省『日本現行法規』に基づく 50 のジャンルに分類した。上位の内訳を示すと、「金融・保険」が 11.5%、「民事」が 10.3%、「行政組織」が 6.3%、「国税」が 5.2%、「産業通則」が 5.2%、「厚生」が 4.9%、「社会福祉」が 4.3%などとなった。

国会会議録： 1976 年から 2005 年までの 30 年間に開催された国会における会議録を対象とした。Web 上の「国会会議録検索システム」（`http://kokkai.ndl.go.jp/`）で公開されているデータのうち、第 77 回国会から第 163 回国会までに開かれた 32,986 会議の会議録データを国立国会図書館より受領し、これらを対象データとした。このうち、両院協議会で開かれた 61 会議、発言部分の文字数が 1,000 文字以下の 6,401 会議、第 77 回国会のうち 1975 年に開催された 33 会議は除外した。

開催年（1976 年から 2005 年まで）を 5 年刻みで 6 期に分割し、さらに 2 種の開催院（「衆議院」「参議院」）、4 種の会議種別（「常任委員会」「特別委員会」「本会議」「その他」）による合計 48 層に対象データを層別した。1 サンプルの範囲を 1 会議とし、取得語数は 500 万語とした。48 の各層に含まれる発言文字数を比例割当し、各層から取得するサンプル数を算出した。結果、対象データ全体から 159 サンプルを取得した。

4 書誌情報データの設計と実装

5 年間に渡って継続してきたサンプリングの実作業と並行して、取得した個々のサンプルの出自に関する情報を取りまとめ、データベース化する作業も行なってきた。このデータベースを「書誌情報データ」と呼ぶ。書誌情報データに格納されている情報の主な要素を、表 6 に示す[1]。

設計当初は書籍の書誌情報を整備することを主たる目的としており、国立国会図書館の蔵書目録データを用いてその初期値を定めた。その後、BCCWJ に格納される全メディアに関して、このデータ構造を用いて書誌情報を整備することにした。メディアの種類が増えるにしたがって、「ジャンル」列の数を増設するなど部分的な拡張はしたものの、基本的な設計を変更することはなく、表 6 に示したデータ構造によってすべての書誌情報を表現した。16 列で構成される書誌情報データをカンマ区切りで出力した例を、図 1 に示す。1. は書籍、2. は雑誌、3. は白書、4. は広報紙、5. は国会会議録の書誌情報の例である。

[1]実際の書誌情報データには、表 6 に示した以外にも、サンプリングの結果に関する情報や著作権者の情報などが含まれる。

表 6: 書誌情報データに格納される情報

0. サンプル ID （Sample_ID）	サンプルに対して付された ID
1. 書誌 ID （Bib_ID）	原本に対して付された ID
2. タイトル （Title）	原本のタイトル
3. 副題 （Subtitle）	原本の副題
4. 巻号 （Number）	原本の巻号
5. 責任表示 （Bib_author）	原本の責任表示（著者、編者、監修者など）
6. 出版者 （Publisher）	原本の出版者（出版社）
7. 出版年 （Year）	原本の出版年
8. ISBN（ISBN）	原本に付された ISBN（国際標準図書番号）
9. 判型 （Size）	原本のサイズ
10. ページ数 （Pages）	原本のページ数
11. ジャンル (1) （Genre_1）	原本のジャンルに関する情報 (1)
12. ジャンル (2) （Genre_2）	原本のジャンルに関する情報 (2)
13. ジャンル (3) （Genre_3）	原本のジャンルに関する情報 (3)
14. ジャンル (4) （Genre_4）	原本のジャンルに関する情報 (4)
15. 責任表示 ID （Bib_author_ID）	原本の責任表示に対応する ID

1. PB18_00030，PB_20164203，日英対照動詞の意味と構文，，，影山太郎|編，大修館書店，2001，4469244597，21cm，317，8 言語，835，3082，，00111247
2. PM11_00712，PM_00010110，ＡＥＲＡ（アエラ），，2001 年 10 号，，朝日新聞社，2001,，A4 変型判，80，1 総合，，，週刊，
3. OW1X_00137，WR_00000005，わが外交の近況，昭和 53 年版，，外務省，大蔵省印刷局，1978,,,,外交，，，，
4. OP44_00001，PR_14212017，広報あつぎ,，2008 年 17 号,，神奈川県厚木市，2008,,,，関東地方，神奈川県，，，
5. OM65_00010，MD_02050010，国会会議録,，第 154 回国会,，，2002,,,,参議院，常任委員会，予算委員会，，

図 1: 書誌情報データの例

BCCWJ の利用者は、この書誌情報データを参照することにより、BCCWJ を構成するすべてのサンプルの出自と属性を知ることができる。特に、発行年やジャンル情報など、層別に用いた情報を参照することにより、BCCWJ を検索した結果の分布が母集団のどの層に属する現象なのかを知ることができる。厳密な手順で取得された大量のサンプルを、その書誌情報と関連付けて利用することにより、コーパスの分析結果が現代日本語書き言葉のどの位相に位置づけられるものであるかを明確にすることができるわけである。このような利点は、例えば Web をコーパスとして用いる方法論では得ることのできないものであり、均衡コーパスとしての BCCWJ が持つ意義を最大限に特徴付けるものであると言える。

5　まとめ

以上、本稿では、過去 5 年間に渡って継続してきたサンプリング作業の最終結果について報告した。また、BCCWJ に格納された各サンプルの出自を知るための書誌情報を提供する「書誌情報データ」の設計と実装について述べた。

2006 年度の作業開始以降、母集団となる総文字数を定義し、サンプル構成比を算出し、その設計をできるだけ忠実に実現するべくサンプリングに努めてきた結果、固定長サンプルのサイズ変更などはあったものの、最終的にはほぼ当初の設計通りの結果を得ることができた。これまで世界中で構築されてきた均衡コーパスを広く見渡してみても、コーパスの設計からサンプリングの最終結果に至る過程がここまで詳らかになった事例はなかったと思われる。均衡コーパスの設計、および構築手順の妥当性を評価する上で、これらの情報を蓄積し開示することが極めて重要であるという点を指摘して、5 年間に渡るサンプリングの最終報告としたい。

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）による補助を得ています。また、BCCWJ のサンプリング作業では、5 年間に渡り、以下の各機関・各社より多大なご協力をいただきました。記して感謝申し上げます。

国立国会図書館、日本図書館協会、立川市図書館、東京都立中央図書館、
東京都立多摩図書館、東京都立日比谷図書館、八王子市図書館、横浜市中央図書館、
埼玉県立久喜図書館、埼玉県立浦和図書館、埼玉県立熊谷図書館、大阪市立中央図書館、
一橋大学附属図書館、自治大学校図書室、湘北短期大学図書館、学習研究社、小学館、
ヤフー株式会社　（順不同）

文献

柏野和佳子、丸山岳彦、稲益佐知子、田中弥生、秋元祐哉、佐野大樹、大矢内夢子、山崎誠. (2009).『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例. 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-01) 特定領域研究「日本語コーパス」データ班.

丸山岳彦、柏野和佳子、山崎誠、稲益佐知子、秋元祐哉、佐野大樹、田中弥生、大矢内夢子. (2011).『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装. 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-10-02) 特定領域研究「日本語コーパス」データ班.

丸山岳彦、秋元祐哉. (2007).『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 —現代日本語書き言葉の文字数調査—. 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-02) 特定領域研究「日本語コーパス」データ班.

丸山岳彦、秋元祐哉. (2008).『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 (2) —コーパスの設計とサンプルの無作為抽出法—. 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書 (JC-D-07-01) 特定領域研究「日本語コーパス」データ班.

丸山岳彦、柏野和佳子、山崎誠、佐野大樹、秋元祐哉、稲益佐知子、吉田谷幸宏. (2007).『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概要. 特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集 (pp. 79–88).

丸山岳彦、柏野和佳子、山崎誠、佐野大樹、秋元祐哉、稲益佐知子、田中弥生. (2008).『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概要 (2) —流通実態サブコーパスの設計—. 特定領域「日本語コーパス」平成 19 年度公開ワークショップ（研究成果報告会）予稿集 (pp. 37–46).

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. (2009).『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概要 (3) —代表性を実現するためのサンプリング手法—. 特定領域「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集 (pp. 33–42).

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. (2010).『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概要 (4) —コーパスの設計とサンプリングの実際—. 特定領域「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. 37–46).

『現代日本語書き言葉均衡コーパス』における評価表現の分布
―「日本語アプレイザル評価表現辞書（態度表現編）」を用いて―

佐野 大樹　（データ班分担者・意見情報班協力者：国立国語研究所コーパス開発センター）†
柏野 和佳子　（データ班分担者・意見情報班協力者：国立国語研究所コーパス開発センター）

Frequency Distribution of the Evaluative Expressions in the Balanced Corpus of Contemporary Written Japanese: An Exploration Employing the Appraisal Dictionary of Japanese (vol.I attitude)

Motoki Sano　(National Institute for Japanese Language and Linguistics)
Wakako Kashino　(National Institute for Japanese Language and Linguistics)

1　はじめに

「評価」は価値観や規範の構築・保持・変更・破壊を施行するための社会システムとして捉えることができるが、それを言語として表象する具体的な手段の 1 つは感情表現を用いて評価対象への肯定的・否定的態度を示すことである。感情・感情表現の研究は言語学的立場 (中村, 1979) や心理学的立場 (Plutchik, 1960) から行われてきたが、感情の感受の仕方は個人、もしくは、コミュニティに固有のものであるという考えもあり、社会における感情表現の使用傾向について調査したものはほとんどない。このため、例えば患者の語りなど、特定のテクストタイプにおける感情表現の使用傾向が一般的な傾向とどの程度類似した、もしくは、異なるものなのか検討することは難しかった。

そこで本稿では、代表性を有する大規模コーパスである『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を社会における感情の表象の一側面を映す “鏡” と考え、BCCWJ において感情表現がどのように分布しているか調査し、社会における感情表現の使用傾向の特徴について検討する。

2　方法

2.1　使用データ

2.1.1　サンプルの選択

使用データには、「『現代日本語書き言葉均衡コーパス』領域内公開データ（2009 年度版）」（以下、BCCWJ2009）に収録されている生産実態（出版）サブコーパス (Publication SubCorpus 以下、PSC) の書籍データ固定長サンプルのうち日本十進分類法 (以下、NDC。BCCWJ2009 の「Genre1」)・日本図書コード（以下、C コード。BCCWJ2009 の「Genre3」)・生年情報 (BCCWJ2009 の「birthyear」) が付与されているものを用いた。書籍・PSC・固定長サンプルを使用した理由は以下の通りである。

書籍 BCCWJ2009 には、書籍・新聞・雑誌・ブログ等様々な媒体のテクストが収録されている。このうち書籍を使用したのは、書籍サンプルには NDC ・C コード・生年情報などが付与されており、感情表現の分布について調べる際に有用な書誌情報が充実しているためである。

PSC 書籍データには PSC と流通実態（図書館）サブコーパス（Library SubCorpus 以下、LSC）があるが、PSC を選択したのは、刊行年に関する条件を除けば PSC のほうが LSC に比べて、サンプルの取得対象の範囲が広いためである。LSC は、東京都内 13 自治体以上の公立図書館に共通に所蔵されている書籍を母集団（推計 47,877,656,072 文字）としてサンプリングを行ったのに対して、PSC は 2001 年〜2005 年

† toki.sano@ninjal.ac.jp

に刊行された全ての書籍を母集団（推計 65,471,677,099 文字）としてサンプリングを行った (丸山岳彦ほか, 2011)。この性質上、LSC に比べて PSC は、専門的な書籍や社会にあまり流通していない書籍など幅広い種類の書籍を含む。

固定長 サンプルの種類には固定長と可変長サンプルがある。固定長サンプルは、母集団の中からランダムに選ばれた 1 文字を基準（サンプル抽出基準点）として、1,000 文字の範囲を取り出すのに対して、可変長サンプルでは「章」や「節」などのまとまりを取り出す。可変長サンプルは文章構造の分析などを、固定長サンプルは頻度調査など統計的分析を念頭に設計されたものである (丸山・秋元, 2007)。本研究では感情表現の分布について調査するため、固定長サンプルを用いた。但し、BCCWJ2009 に収録されている固定長サンプルには 1 文字目と 1,000 文字目を含む文も収録されているため[1]、本研究では、語数が 800 語を超える 99 サンプル（800 語〜4,656 語）は除外することにした。

2.1.2 書誌情報

調査に用いた書誌情報は、NDC・C コード一桁目（販売対象）・生年情報である。NDC は J-BISC（国立国会図書館蔵書目録）、C コードは出版社が付与したものである。各タイプの延べ語数・異なり語数・サンプル数を表 1 に示す[2]。

表 1: 延べ語数・異なり語数・サンプル数

タイプ	延べ語数	異なり語数	サンプル数
NDC0 総記	94,274	9,297	154
NDC1 哲学	217,074	14,318	340
NDC2 歴史	293,451	22,305	442
NDC3 社会科学	910,648	27,552	1,394
NDC4 自然科学	233,045	13,809	369
NDC5 技術・工学	202,303	14,918	325
NDC6 産業	131,479	10,950	207
NDC7 芸術・美術	172,726	14,575	275
NDC8 言語	63,362	7,445	100
NDC9 文学	981,213	34,920	1,525
C1_0 一般	2,382,414	53,199	3,711
C1_1 教養	174,301	13,764	262
C1_2 実用	183,289	12,101	290
C1_3 専門	499,247	21,164	769
C1_5 婦人	4,393	1,087	7
C1_6 学参 I（小中）	1,197	361	2
C1_8 児童	40,380	4,671	67
C1_9 雑誌扱い	14,354	2,966	23
生年_1830 年代	692	288	1
生年_1850 年代	1,268	461	2
生年_1860 年代	2,043	670	3
生年_1870 年代	6,042	1,544	9
生年_1880 年代	8,894	2,358	13
生年_1890 年代	16,768	3,025	26
生年_1900 年代	33,169	5,177	50
生年_1910 年代	60,601	8,193	92
生年_1920 年代	284,856	20,959	432
生年_1930 年代	601,461	29,105	924
生年_1940 年代	812,161	32,677	1,258
生年_1950 年代	754,625	30,259	1,184
生年_1960 年代	543,077	24,892	859
生年_1970 年代	154,828	12,974	247
生年_1980 年代	18,473	3,014	30
生年_1990 年代	617	279	1

[1] 各固定長サンプルから厳密に 1,000 文字抽出するには、「sampling 要素」(山口昌也ほか, 2008) を用いて抽出する必要がある。但し、文字単位での範囲指定であるから、語を分断する場合がある。そこで、本研究では使用データに含めるサンプルを語数によって限定することにした。

[2] 空白・記号は除く。語数の計測には MeCab0.98 と UniDic1.3.12 を使用した。

2.2 感情表現の分析

感情表現の分析には、現在筆者が構築している『日本語アプレイザル評価表現辞書（態度表現編）』(佐野, 2011a, 2011b, 2011c) の〈内評価〉というカテゴリを用いた。この辞書は約 10,000 語の評価表現を図 1 の体系に則し分類したもので、〈内評価〉には、感情表現が 1,482 件（見出し語）[3]収録されている。この分類体系では、品詞を問わず感情表現を分類できるよう設計しており、また、中村 (1979) と異なり評価極性を問わず分類が可能なカテゴリを設けてある。以下、〈内評価〉の分類について簡単に説明する。

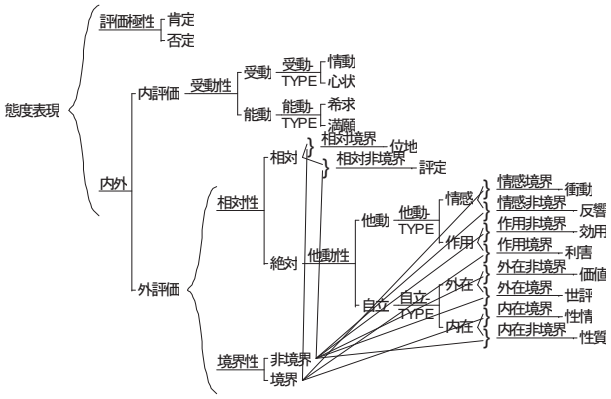


図 1: 態度表現の分類体系

2.2.1 〈内評価〉の下位カテゴリ:〈受動〉と〈能動〉

〈内評価〉には、「楽しむ」「憤慨する」などある対象への評価者の感情、もしくは、「笑う」「泣く」など感情を表す行為を示す表現を掲載してある。これらの表現には、(a) 対象を感受している・した結果としてわき起こるものと (b) 評価者の精神世界に対象を位置づけるものがある。(a) が〈受動〉、(b) が〈能動〉に該当する。例えば (1) の「安心」は〈受動〉に、 (2) の「愛慕」は〈能動〉に該当する。

(1) 彼はその言葉を聞いて 安心 した（『ラグビーボールを抱きしめて』天宮一大 ）
(2) 愛慕 してやまない文豪トルストイをロシアに訪れ…（『国木田独歩空知川の岸辺で』岩井洋）

2.2.2 〈受動〉の下位カテゴリ：〈情動〉と〈心状〉

〈受動〉はさらに、(c) 喜怒哀楽や感動・高揚感など突発的な感情を示すものと (d) 心身の安定性・安全性・危険性を感受したことを示すものがある。(c) が〈情動〉（全 212 件[4]）、(d) が〈心状〉（全 411 件）に該当する。〈情動〉・〈心状〉の下位カテゴリと例を表 2 に示す 。

表 2: 〈情動〉・〈心状〉の下位カテゴリと例

〈情動〉のカテゴリ	例
肯〈楽しさ・愉快さ〉	楽しむ・感興・嬉々・愉快・興ずる
肯〈喜び・幸せ〉	喜び・喜悦・欣喜・歓心・歓喜・悦
否〈怒り〉	怒る・お冠・業腹・煮え繰り返る
否〈悲しさ〉	悲しむ・哀絶・感嘆・物悲しい

〈心状〉のカテゴリ	例
肯〈安心〉	安心・安息・安堵・ほっと・休まる
肯〈壮快・安定〉	癒える・落ち着く・清々・慰む
否〈恐怖〉	臆する・脅える・寒心・ぎょっと
否〈心配・不安〉	案ずる・懸念・不安・心配・心細い

[3] 語義にすると 1,656 件である。
[4] 語義によってカテゴリが変わる見出し語もあるため、カテゴリ間で重複する表現もある。

2.2.3 〈能動〉の下位カテゴリ：〈希求〉と〈満願〉

〈能動〉はさらに、(e) 評価者の趣向と評価対象を照合して愛情・希望・欲求等を示すものと、(f) 目的の達成度や評価者がもつ規範と評価対象を照合して満足度を示すものがある。(e) が〈希求〉（全 427 件）、(f) が〈満願〉（全 478 件）に該当する。〈希求〉・〈満願〉の下位カテゴリと例を表 3 に示す。

表 3: 〈希求〉・〈満願〉の下位カテゴリと例

〈希求〉のカテゴリ	例
肯〈愛情〉	愛する・恋しい・慕う・愛でる
肯〈欲求〉	心待ち・懇請・切望・懐かしむ
否〈恨み・憎しみ〉	恨む・怨ずる・憎しみ・面憎い
否〈嫌う・疎む〉	嫌う・嫌気・疎意・嫌悪・厭忌
〈満願〉のカテゴリ	例
肯〈満足〉	甘心・自足・満悦・満足・堪能
肯〈感謝〉	有難い・感恩・感謝・謝意・随喜
否〈後悔・悔しさ〉	悔いる・心残り・懲りる・思い残す
否〈不平・不満〉	不服・ぶすぶす・不満足・慊焉

感情表現の分類は主観的になりやすいが、『日本語アプレイザル評価表現辞書（態度表現編）』では岩波国語辞典（第 5 版）に掲載された表現から人手で網羅的に感情表現を抽出し、各カテゴリに属する表現をリスト化しており、これを用いることで感情表現の分類であっても一定の客観性を保つことができると考えた。

2.3 分布の調査方法

2.3.1 カバー率

『日本語アプレイザル評価表現辞書（態度表現編）』の〈内評価〉に掲載されている表現がどの程度使用されているか、使用データ全体と 2.1.2 で述べた書誌情報のタイプごとに調べた。〈内評価〉のカテゴリごとの登録語数、及び、タイプごとに総語数が異なるため、以下の式を比較に用いることにした。

カバー率 ＝ 使用された感情表現の異なり語数÷当該の〈内評価〉のカテゴリに含まれる登録語数
÷当該のタイプ全体における異なり語数× 10,000

2.3.2 出現率

『日本語アプレイザル評価表現辞書（態度表現編）』の〈内評価〉に掲載されている表現の出現率を使用データ全体と 2.1.2 で述べた書誌情報のタイプごとに調べた。出現率の計測には、以下の式を用いた。

出現率 ＝ 当該の〈内評価〉のカテゴリに含まれる表現の使用度数÷当該タイプ全体における延べ語数× 100

2.3.3 〈内評価〉のカテゴリと NDC・販売対象・生年情報との対応

〈内評価〉のカテゴリと 2.1.2 で述べた書誌情報のタイプとの対応関係について総合的に調べるためコレスポンデンス分析を行った。

3 結果と考察

3.1 使用データ全体における傾向

使用データ全体における〈内評価〉のカテゴリごとのカバー率と出現率を図 2 に示す[5]。カバー率に関しては、〈内評価〉のカテゴリの間に顕著な違いは見られないが、出現率では〈能動〉の下位カ

[5] 使用データに含まれる語のうち『日本語アプレイザル評価表現辞書（態度表現編）』に掲載されているもので、使用度数 30 以上のものに関して、語義によって評価表現になる場合とならない場合があるものは調査対象から除外した。また、語義によって〈内評価〉のカテゴリが異なるものも除外した。両者合わせて 65 件である。

テゴリである〈希求〉・〈満願〉の値が高い。データ全体の傾向としては、対象を感受している・した結果としてわき起こる感情を示す〈受動〉よりも、評価者の精神世界に対象を位置づける〈能動〉のほうが多用される傾向にある。書籍では、執筆している時間と実際に感情を感受する時間に差があることが多いため、突発的な感情を示す表現を多く含む〈受動〉よりも〈能動〉の出現率が高いのではないかと考える[6]。

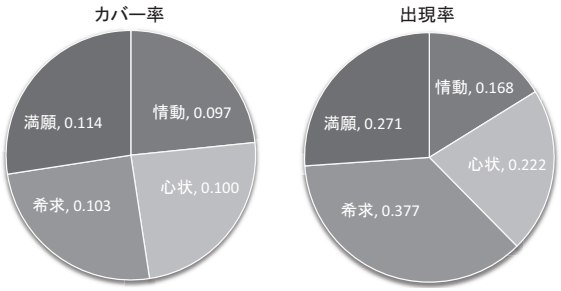


図 2: 使用データ全体におけるカバー率・出現率

3.2 タイプ別の傾向

3.2.1 NDC

NDC 別のカバー率と出現率を図 3 に示す。カバー率をみると、〈満願〉のカバー率はどのタイプでも高いが、〈心状〉のカバー率には差がある。4 自然科学においては〈心状〉が最も高い。医療に関連する書籍で〈苦しみ・痛み・つらさ・消沈〉に関する表現や〈安心・壮快・安定〉に関する多様な表現が用いられていた。一方で、5 技術. 工学、6 産業、7 芸術. 美術では〈心状〉が最も低い。また、〈情動〉のカバー率も NDC によって差があり、7 芸術. 美術でのカバー率が最も高い。骨董品や楽器等の創作・演奏に関連する書籍で〈楽しさ・愉快さ〉に関する表現や〈喜び・幸せ〉に関する多様な表現が用いられていた。一方、2 歴史、3 社会科学、 4 自然科学などでは低い。

一方、出現率をみると、〈希求〉が全体的に高い。特に、1 哲学、9 文学、 7 芸術. 美術で高く、これらのタイプでは〈愛情〉に関する表現が多用されていた。4 自然科学で〈心状〉の値が高いのは、カバー率の場合と同様の理由であった。

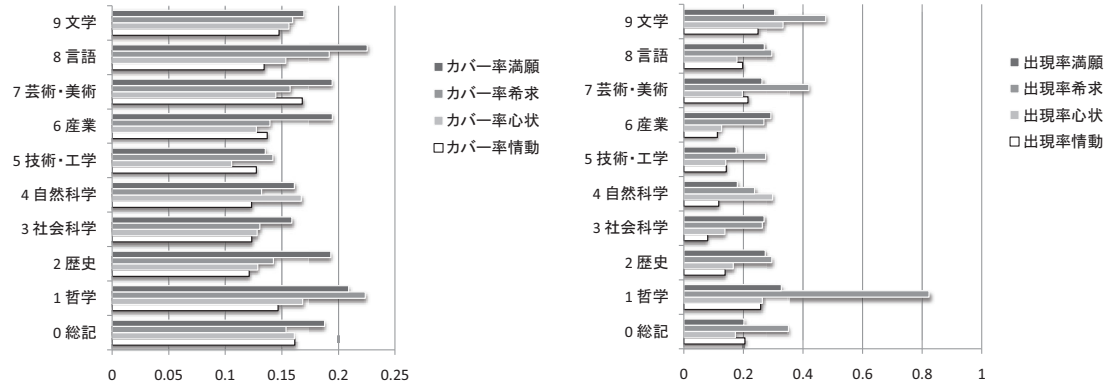


図 3: NDC 別カバー率・出現率

[6] 補足となるが、使用データにおいて 929 件の感情表現が用いられており、使用度数としては 32,249 件あった。しかし、使用度数の 9 割をカバーするのに必要な表現の数は 294 件であった。評価表現辞書は、登録語数よりも登録語に対してどのような情報が付与されているかが評価の上で重要であると思われる。

3.2.2　販売対象

販売対象別のカバー率と出現率を図 4 に示す[7]。カバー率をみると、サンプル数が比較的多い 0〜3 までは、いずれも〈情動〉〈心状〉〈希求〉〈満願〉の順にカバー率が高くなっていく傾向がみられるが、出現率をみると販売対象によって〈内評価〉のカテゴリに違いがみられる。特に、0 一般と 3 専門では、3 専門で出現率が全体的に低いだけでなく、〈受動〉に該当する〈情動〉・〈心状〉の出現率と〈能動〉に該当する〈希求〉・〈満願〉の出現率に顕著な違いがみられる。0 一般においては、〈希求〉の値が他に比べ高いものの、全てのカテゴリにおいて約 0.2 以上の出現率がある。一方、3 専門では〈受動〉に該当するカテゴリの値が小さく、特に〈情動〉は 0.1 にも満たない。〈情動〉は喜怒哀楽や感動・高揚感など突発的で一時的な感情を扱う表現が多いため、固定化した知識を扱うことが多い 3 専門では、あまり使用されないのではないかと思われる。

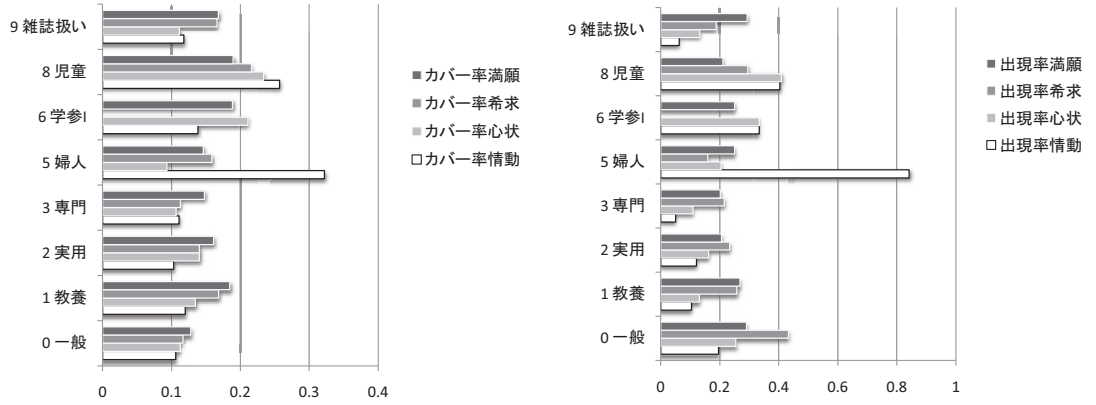


図 4: 販売対象別カバー率・出現率

3.2.3　生年

生年別のカバー率と出現率を図 5 に示す[8]。カバー率をみると、どのカテゴリも凹状の変化を示している。出現率では〈満願〉が減少傾向にあるものの、他のカテゴリでは 1940 年代まで減少し、1950 年代から増加傾向にある。このことは、1900 年代の初めと終わりに生まれた著者は使用する感情表現が多様であり、かつ、感情表現を多用する傾向にあるが、1940 年代前後に生まれた著者は感情表現を用いることを抑圧される傾向にあることを示すものと考えられる。

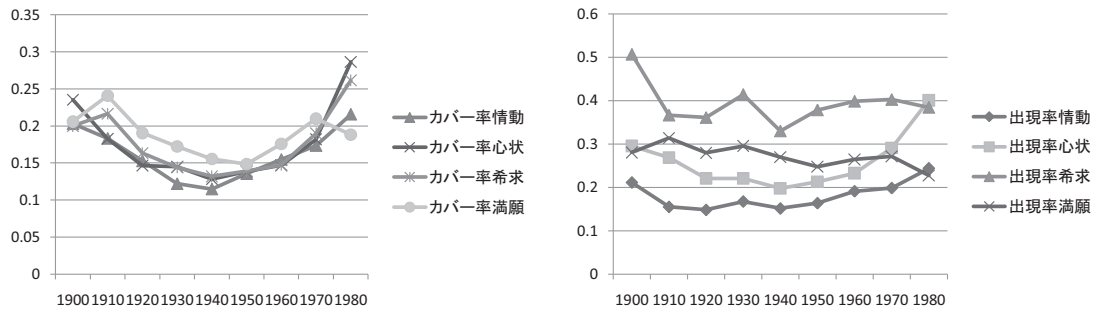


図 5: 生年別カバー率・出現率

[7] 5 婦人・6 学参 I・9 雑誌扱いはサンプル数が 30 以下と少ないが、他にこのような調査を報告したものがほとんどないこともあり、参考までにデータを提示することにした。
[8] 販売対象の場合と異なり、ここでは通時的変化を見るため、サンプル数が 30 より小さい年代は図に含めていない。

また、1900 年代初めは〈能動〉のカテゴリの出現率が高いのに対して、1900 年代終わりでは〈受動〉のカテゴリの出現率が高くなってきている。特に〈心状〉は 1940 年代から出現率が 0.2 も増加している。〈能動〉と〈受動〉の違いを考慮すると、1900 年代初めに生まれた著者は、評価者が評価対象を位置づけ、能動的な役割を果たすことが多いのに対して、1900 年代終わりでは、評価者が評価対象によって感情を左右され、受動的な役割を果たすことが多くなってきている可能性がある。幼少期における経験が感情の表現方法に影響を与えるとすれば、戦前・戦後では感情の感受の仕方に違いがあるのかもしれない。

感情の感受の仕方は個人によって異なるものだと考えがちだが、社会動向を個人経験の集合と捉え、この集合を代表性を有する大規模コーパスが映すとすれば、生年による感情表現の使用傾向の違いは、感情の感受の仕方は社会動向によっても影響されるものであることを示唆するものと捉えられるのではないだろうか。

3.3 〈内評価〉のカテゴリと各タイプの対応関係

コレスポンデンス分析を用いて〈内評価〉の下位カテゴリと上述した全てのタイプの対応関係について分析した結果を図 6 に示す[9]。

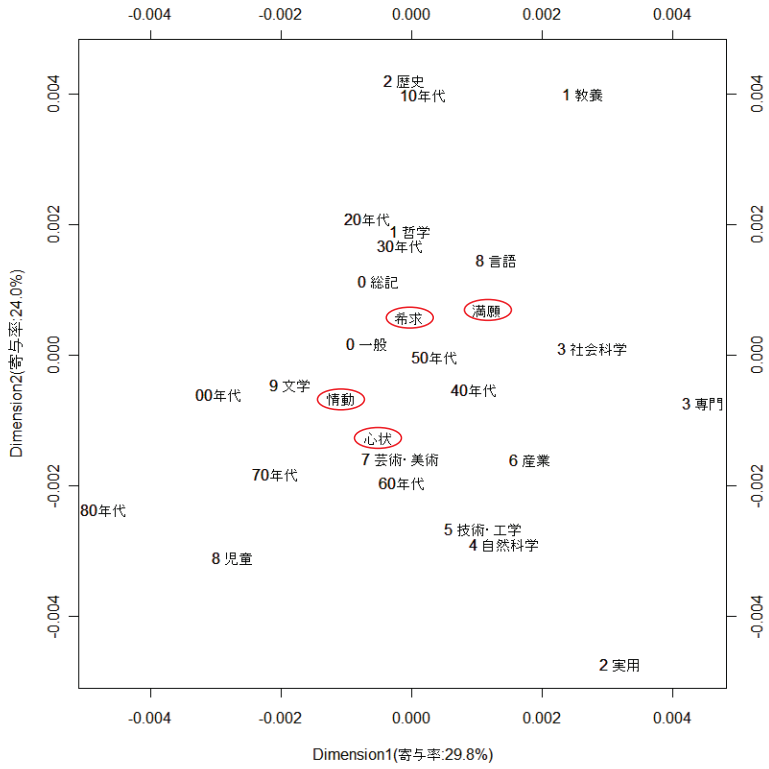


図 6: コレスポンデンス分析の結果

次元 1 の正方向に〈満願〉が、負方向に〈情動〉が、ほぼ中央に〈希求〉・〈心状〉が布置されていることから、正方向に布置されるタイプほど〈満願〉を使用する傾向が強く、負方向に布置されるタイプほど〈情動〉を使用する傾向が強いと考えられる。また、次元 2 の正方向に〈満願〉・〈希求〉、負方向に〈心状〉・〈情動〉が布置されていることから、正方向に布置されるタイプほど〈能動〉を使用する傾向が強く、負方向に布置されるタイプほど〈受動〉を使用する傾向が強いと考えられる。以

[9] サンプル数が 30 より小さいタイプのデータは除外している。

上を踏まえると、図 6 より相対的にみて以下の傾向が認められる。

〈満願〉が多いタイプ 3 社会科学、6 産業、3 専門、2 実用、1 教養
〈情動〉が多いタイプ 9 文学、8 児童、1980 年代、1900 年代、1970 年代

〈満願〉が多いタイプには、産業・専門など実用性が高い NDC や販売対象がある。一方〈情動〉が多いタイプには、文学・児童など創造性が高い NDC や販売対象がある。また、様々な感情表現を多用する傾向がある 1900 年代初め・終わりの年代がある。

〈能動〉が多いタイプ 2 歴史、1 哲学、8 言語、0 総記、1 教養、1910 年代、1920 年代、1930 年代
〈受動〉が多いタイプ 4 自然科学、5 技術. 工学、7 芸術. 美術、6 産業、2 実用、8 児童、1980 年代、1970 年代、1960 年代

〈能動〉が多いタイプには、哲学・言語など概念的な事象を扱うもの、及び、1900 年代初めのものがある。一方、〈受動〉が多いタイプには、物理的事象や生産物を扱うもの、及び、1900 年代終わりのものがある。

感情の感受の仕方は、個人・コミュニティにおける内的要因によって影響を受けるだけでなく、実用性が求められる分野か、創造性が求められる分野か、扱う事象が概念的か、物理的か、また、生年はいつか、など、外的要因によっても影響を受けるものだと考えられる。

4 まとめと今後の展望

本研究では、BCCWJ2009 に付与された書誌情報と『日本語アプレイザル評価表現辞典（態度表現編）』を用いて、感情表現の分布について調べた。今後は、今回調査した感情表現の使用傾向の特徴を踏まえた上で、テクストタイプごとの感情表現の使用傾向について調査し、レジスターを記述していくなかで、個人や社会の動向の変化を感情表現を含めた評価表現の使用傾向から捉えられるシステムを構築することができればと考えている。

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）・若手研究（B）「日本語「書き言葉らしさ・話し言葉らしさ」測定法の設計」（平成 21〜23 年度、代表者：佐野大樹）による補助を得ています。

文献

丸山岳彦、秋元祐哉. (2007). 『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 —現代日本語書き言葉の文字数調査—. 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-02) 特定領域研究「日本語コーパス」データ班.

丸山岳彦、山崎誠、柏野和佳子、佐野大樹、秋元祐哉、稲益佐知子、田中弥生、大矢内夢子. (2011). 『現代日本語書き言葉均衡コーパス』におけるサンプリングの概要 (5)—サンプリングの最終結果—. 本予稿集所収.

中村明. (1979). 感情表現辞典. 六興出版.

Plutchik, R. (1960). The multifactor-analytic theory of emotion. *Journal of Psychology*, *50*, 153-171.

佐野大樹. (2011a). 『日本語アプレイザル評価表現辞書（態度表現編）』の構築—評価の多様性を捉えるための言語資源の開発—. 言語処理学会第 17 回年次大会発表論文集 掲載予定.

佐野大樹. (2011b). 患者の語りにおける感情表現の使用傾向—『アプレイザル評価表現辞書（態度表現編）』を用いた乳がん患者・前立腺がん患者の語りの分析—. 第 27 回社会言語科学会研究大会発表論文集 掲載予定.

佐野大樹. (2011c). 日本語における評価表現の分類〜アプレイザル理論をベースに〜. 信学技報, 109(390), NLC2010–33 (pp. 19–24).

山口昌也、高田智和、北村雅則、間淵洋子、小林正行、西部みちる. (2008). 『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.2.0. 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書特定領域研究「日本語コーパス」データ班.

Yahoo!知恵袋の質問における修辞機能の分布
―修辞ユニット分析を用いて―

田中 弥生（データ班協力者：国立国語研究所コーパス開発センター）†
佐野 大樹（データ班分担者：国立国語研究所コーパス開発センター）

Variation of Rhetorical Function in Yahoo!Chiebukuro: An application of Rhetorical Unit Analysis to texts on the Web

Yayoi Tanaka (National Institute for Japanese Language and Linguistics)
Motoki Sano (National Institute for Japanese Language and Linguistics)

１．はじめに

Web上のコミュニケーションの分析にはさまざまなものがある[1]。例えばQ&Aサイトについて、決まった解の有無によって質問タイプを分類するもの(三浦・川浦 2008)、情報検索型・社会調査型に分類するもの(栗山・神門 2009)、factoid型－non-factoid型分類を質問応答システムに応用する研究(田村ほか 2008)、結束性の分析から対話性を検討するもの(田中 2010)などがある。しかし、質問や回答の一般性・抽象性に焦点をあてた分析は行われていない。本研究は、元来話し言葉の談話分析手法である修辞ユニット分析（Rhetorical Unit Analysis 以下、RUA）をWeb上のコミュニケーションに適用する研究の一環として、Q&Aサイトの化粧品などに関する質問に適用し、質問の脱文脈化の程度における分布を調べるものである。

以下、２．でRUAの概要を述べ、３．で分析対象と分析方法を説明し、４．で分析結果と考察、５．でまとめと今後の課題を述べる。

２． RUAとは

RUAは、発話機能、中核要素、現象定位の３つをメッセージ[2]単位で認定して修辞機能を特定するもので、その結果として脱文脈化の程度を知ることができる。脱文脈化言語は「一般化された要素の習慣的・恒久的な行動や状態について表現する言語」、文脈化言語は「物質的状況に存在する要素の現在の行動や状況について表現する言語」である（Cloran 1999)。表１は佐野(2010b)の修辞機能の特定表に脱文脈化指数（脱文脈化の程度の表わし方）を合わせて示したものである．脱文脈化指数の数値が大きいものほど脱文脈化の程度が高く、小さいものほど文脈化の程度が高い[3]。

佐野(2010b)で述べたように、RUAはテクストの意味単位を特定するための手法だが、その過程においてメッセージの修辞機能の種類の特定にも利用できる。英語においては母子会話、教師と生徒の会話等の分析に活用され、知識伝達の分析に有用な枠組みと考えられているが、日本語に適用した研究は、日本語教育への適用を検討した佐野(2010b)、書き言

† yayoi@ninjal.ac.jp

1 本稿は、電子情報通信学会 言語理解とコミュニケーション研究会第２回集合知シンポジウムにて「Yahoo!知恵袋における質問の修辞ユニット分析－脱文脈化-文脈化の程度による分類－」として口頭発表した内容を発展させたものである。

2 選択体系機能言語理論の意味層（semantics）における最小単位で、基本的には節によって表される。

3 脱文脈化指数の詳細は佐野・小磯(2011)参照。

葉への適用性を検討した佐野・小磯(2011)などがあるものの、まだ少ない。

表 1　修辞機能の特定と脱文脈化指数(佐野 2010b に加筆)

<table>
<tr><td colspan="3" rowspan="4"></td><td colspan="6">発話機能</td></tr>
<tr><td rowspan="3">提言</td><td colspan="5">命題</td></tr>
<tr><td colspan="5">現象定位</td></tr>
<tr><td colspan="2">現在</td><td rowspan="2">過去</td><td colspan="2">未来</td><td rowspan="2">仮定</td></tr>
<tr><td>非習慣的
・一時的</td><td>習慣的
・恒久</td><td>意図的</td><td>非意図的</td></tr>
<tr><td rowspan="4">中核要素</td><td rowspan="2">状況内</td><td>参加</td><td>[1]行動</td><td rowspan="2">[2]実況</td><td>[7]自己記述</td><td rowspan="2">[3]
状況内回想</td><td>[4]計画</td><td rowspan="2">[5]
状況内予想</td><td rowspan="2">[6]
状況内推測</td></tr>
<tr><td>非参加</td><td rowspan="3">n/a</td><td>[8]観測</td><td rowspan="2">[4]計画</td></tr>
<tr><td colspan="2">状況外</td><td>[9]報告</td><td>[13]説明</td><td rowspan="2">[10]
状況外回想</td><td colspan="2" rowspan="2">[11]予測</td><td rowspan="2">[12]推量</td></tr>
<tr><td colspan="2">定言</td><td>n/a</td><td>[14]一般化</td></tr>
</table>

「n/a」は該当なし
背景が灰色の部分が修辞機能の種類
[ ]内は脱文脈化指数

３．Q&A サイトへの RUA の適用基準

３.１.分析対象

分析には、「『現代日本語書き言葉均衡コーパス』領域内公開データ（2009 年度版）」に収録されている「Yahoo!知恵袋」データを利用した。中カテゴリ「コスメ、美容」の中で質問文 1 つから成る「質問」90 投稿のうち、国立国語研究所(1960、1963)で示されている「要求表現」の「質問的表現」が使用されている 59 投稿を分析対象とした[4]。RUAでは、当該のメッセージだけではなく他のメッセージの照応なども考慮して中核要素と現象定位を認定するが、本研究では、複雑さを排除するため、照応を考慮する必要のない単一の質問文から成る「質問」投稿を対象とした。

３.２.分析手順

分析の手順は、1.分析対象のメッセージと種類の認定、2.発話機能の認定、3.中核要素の認定、4.現象定位の認定、5.修辞機能の特定と脱文脈化指数の確認、である。 1 から 4 によって発話機能、中核要素、現象定位が認定されれば、それらの組み合わせによって 5 の修辞機能と脱文脈化指数が決まる[5]。

３.３.分析対象のメッセージと種類の認定

RUAでは、テクストをメッセージという単位に分割し、メッセージの種類を認定する[6]。メッセージは、「位置付けpositioning」「拘束bound」「自由free」のいずれかに分類し、さらに「拘束」は「拘束：意味的従属」と「拘束：形式的従属」に分類する。「位置付け」は、挨拶・定型句・フィラーなど、述部を含まない節のみによって構成されるものである。「拘束：意味的従属」は、従属するメッセージの状況（時間・場所・原因・結果等）を説明するもので、従属しているメッセージの一部と考える。「拘束：形式的従属」は、意味的には並列の関係であるが、時制（過去）などの側面で、従属するメッセージに形式的に依存するものである。「自由」は独立して時制やムードなどを表わすものである。RUAでは、これ

[4] 59 投稿で使用されている「質問的表現」は、「ですか」「ますか」「でしょうか」「ませんか」「でしたか」「ましたか」「ました？」である。
[5] RUA の詳細は、佐野(2010b)、佐野・小磯(2011)で述べた。
[6] メッセージの種類の認定については、佐野(2010a)参照。

らのうちの「自由」と「拘束：形式的従属」について、修辞機能の認定を行う。

(1)は「自由」の例である。(2)は、【】[7]内の「T字型カミソリで剃っていると」が原因を示しているため「拘束：意味的従属」で、それ以外の部分が「自由」である。また、(3)の、「こういう長い連休でずっと家にいる方」は「髭そったり」「髪をセットしたり」に関わるが便宜上aに含め、a.とb.は並列となる。(3)a.は「拘束：形式的従属」、(3)b.が「自由」である。

(1) ヒリヒリしみる化粧水って、自分に合わないってことでしょうか？

(2) 男性のヒゲは、【T字型カミソリで剃っていると】濃くなるんですか？

(3) a.こういう長い連休でずっと家にいる方、髭そったり、
b.髪をセットしたりしますか？

メッセージの種類を認定した結果、本研究の分析対象メッセージは表２のとおりで、３.４.以降の認定を行う対象は、「自由」59 メッセージ、「拘束：形式的従属」1 メッセージの合計 60 メッセージとなる。

表２ メッセージの種類と数

メッセージ総数	71
「自由」	59
「拘束：形式的従属」	1
「拘束：意味的従属」	11
「位置付け」	0

３.４.発話機能の認定

メッセージの種類を認定した後、発話機能を認定する[8]。発話機能は、「提言」か「命題」のどちらかに分類する。「提言」は品物・行為の交換に関するメッセージ、「命題」は情報の交換に関するメッセージが該当する。例えば、(4)は「お塩を取る」という行為を要求する「提言」で、(5)は「おいしいお塩」についての情報を要求する「命題」である[9]。

(4) お塩を取っていただけますか？

(5) おいしいお塩を教えていただけますか？

上述の(1)(2)(3)の発話機能は、いずれも「命題」である。

３.５.中核要素の認定

次に中核要素を認定する。中核要素はメッセージの中心となる要素で、基本的には主語によって表現される。中核要素はまず「状況内要素」「状況外要素」「定言要素」のいずれかに分類し、「状況内要素」はさらに「参加要素」「非参加要素」に分類する（図１参照）。

３.５.１.状況内要素

「状況内要素」はコミュニケーションが行われている場面に存在するものが該当し、そのコミュニケーションの当事者が「参加要素」、その場面に存在するが参加していない人や事象が「非参加要素」となる。Q&A サイトは対面や電話のように空間や時間を共有しておらず、サイト上の「質問」と「回答」そのものがコミュニケーション空間であるため、「参加要素」には質問投稿者、回答投稿者、および回答する可能性のある閲覧者が該当すると考えられる。

---

7 【】部分は、メッセージの種類が「拘束：意味的従属」であることを示す。

8 発話機能については、Halliday & Matthiessen(2004)参照。

9 (4)(5)は筆者の作例である。断りがないものについては Yahoo!知恵袋データからの引用である。

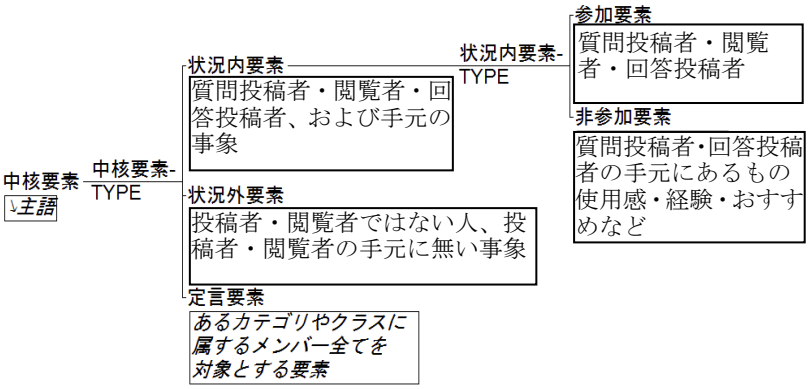


図１　Yahoo!知恵袋「コスメ、美容」における中核要素の分類基準

A)　状況内：参加要素

「状況内：参加要素」には、基本的には一人称、二人称が該当する。(6)では「みなさん」が該当する。また、(7)では「あなたは」又は「みなさんは」が省略され、(8)では「私は」が省略されていると考えられる。

(6)　朝のぼさぼさの髪を**みなさん**[10]どうやってセットしてますか？

(7)　**φ**[11]クリスマス限定キット何を買う予定ですか？

(8)　虫除けスプレーと日焼け止めクリーム、**φ**どういう順番で使えば良いでしょうか。

B)　状況内：非参加要素

「状況内：非参加要素」は、質問投稿者・回答投稿者の手元にあるものが該当する。また、話題が「コスメ、美容」であるため、使用感・おすすめなど投稿者の経験に関わるものについても、「状況内：非参加要素」と判断した。

(9)　**ヒリヒリしみる化粧水って、**自分に合わないってことでしょうか？

３.５.２. 状況外要素

「状況外要素」は、「投稿者でも閲覧者でもない人、投稿者・閲覧者の手元にはない事象」である。

(10) **ゲランのチェリーブロッサムと、チェリーブロッサムグリッターは**同じ香りなんでしょうか？

(11) **鼻の頭の黒いぶつぶつを上手に取る方法は**ありませんか。

３.５.３. 定言要素

「定言要素」は、「あるカテゴリやクラスに属するメンバー全てを対象とする要素」である。(12)は化粧水の中の収れん化粧水という種類全体について聞いているもので、「定言要素」となる。

(12) **収れん化粧水とは**どのようなものですか？

３.６. 現象定位の認定

現象定位はメッセージが伝達されている時（Time of speaking 以下、Ts）を基準として、

10　中核要素を太字ゴシック体で示す。
11　中核要素（主語）が省略されている場合はφとする。

メッセージによって表現されている出来事がいつ起こったかを示す要素である。現象定位はまず「現在」「過去」「未来」「仮定」に分類し、「現在」はさらに「非習慣的・一時的」「習慣的・恒久」に、「未来」は「意図的」「非意図的」に分類する（図２参照)。「現在」と「過去」は Ts において出来事がすでに起こったもの、「未来」と「仮定」は Ts では起こってないものが該当する。

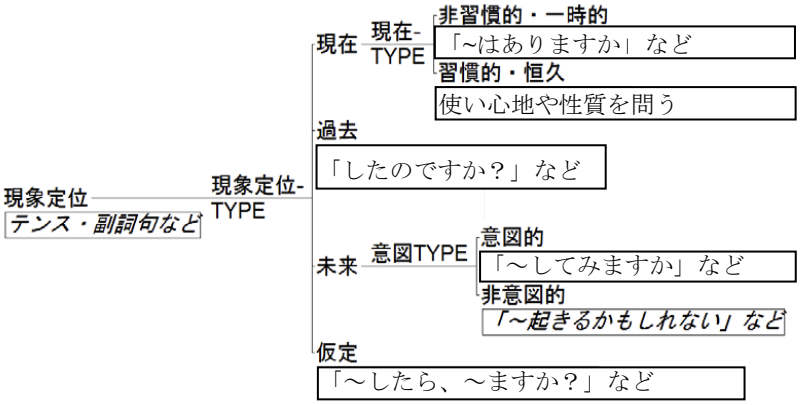


図２ Yahoo!知恵袋「コスメ、美容」における現象定位の分類基準

３.６.１.現在

A) 現在：習慣的・恒久

現象定位の「現在」を「非習慣的・一時的」と「習慣的・恒久」に分類する基準は習慣性と恒久性である。習慣を聞いている場合には、「現在：習慣的・恒久」と認定する。例えば、「いつも」「毎～」などの表現があるか、あるいは無い場合には挿入できるか否かが指標となる。例えば前掲の(6)や(8)が該当する。また、物事の定義や変わらない性質を聞いているものが恒久に該当する。例えば前掲の(12)が該当する。

B) 現在：非習慣的・一時的

一方、習慣性・恒久性について述べていないものは、「現在：非習慣的・一時的」となる。例えば前掲の(10)(11)が該当する。

３.６.２.過去

Ts より前に起こったことを聞いたり表したりしている場合は現象定位が「過去」となる。

(13) **ジョンソンから新発売のボディーソープ「ソフトウォッシュ」を使ってみた方、**どんな感じでしたか？[12]

３.６.３.仮定

「仮定」は、「A が生じた場合、B が起こる」という因果関係を持つものが含まれる。(14)は「まめに塗りなおした場合」に「SPF の高くないもので海辺にいても大丈夫である」という因果関係を聞いているため、現象定位は「仮定」である。

(14)【まめに塗り直せば、】【ＳＰＦの高くないもので海辺にいても】φ大丈夫でしょうか

３.６.４.未来

Ts では起こってないことを聞いたり述べていて「仮定」でない場合は「未来」である。「未

[12] 現象定位をイタリック体で示す。

来」は意図できる行動・現象か、できない行動・現象かによって 2 つに分類される。

A)　未来：意図的

意図出来る行動・現象の場合は「未来：意図的」である。前掲の(7)が該当する。

B)　未来：非意図的

意図出来ない行動・現象の場合は「未来：非意図的」である。

(15) **ちふれ化粧品の福袋は**出るのでしょうか？

３.７.修辞機能の特定と脱文脈化指数の確認

発話機能と中核要素と現象定位の組み合わせから、修辞機能が決定される。前掲の表 1 より、(16)は、発話機能は「命題」、中核要素が「収れん化粧水とは」で「定言要素」、現象定位が「どのようなものですか」で「現在：非習慣的・一時的」であるので、修辞機能は「一般化」で、脱文脈化指数は[14]となる。あるカテゴリについてその性質を聞いているので「一般化」という修辞機能になるのである。

(16) **収れん化粧水とは**どのようなものですか？（(12)再掲）

また、(17)は、発話機能は「命題」、中核要素は「ちふれ化粧品の福袋は」で「状況外要素」、現象定位が「出るのでしょうか」で「未来：非意図的」のため、修辞機能は「予測」で脱文脈化指数は[11]となる。コミュニケーションが行われている場にない、まだ起こっていないことについて聞いているので「予測」となる。

(17) **ちふれ化粧品の福袋は**出るのでしょうか？（(15)再掲）

さらに、(18)は、発話機能は「命題」、中核要素は省略されている「あなたは」あるいは「みなさんは」で「状況内：参加要素」、現象定位は「何を買う予定ですか」で「未来：意図的」のため、修辞機能は「計画」で脱文脈化指数は[4]である。コミュニケーションの当事者のこれからの行動について聞いているので、「計画」となる。

(18) **φ**クリスマス限定キット何を買う予定ですか？（(7)再掲）

４．分析結果と考察

分析の結果得られた修辞機能の分布を、空間的距離のレベルと時間的距離のレベルの 2 つの側面から図３(a)に示す[13]。[ ]内は脱文脈化指数である。縦軸は空間的距離のレベルで、中核要素を示している。中核要素がコミュニケーションの参加者の場合には空間的距離のレベルが低く、中核要素がコミュニケーションの場から離れていれば空間的距離のレベルは高くなる。一方横軸は時間的距離のレベルで、現象定位を示している。投稿者の経験や一時的な行動を示すものは時間的距離のレベルが低く、まだ起こっていないことや恒久的なことについては時間的距離のレベルが高くなる。例えば、中核要素が「定言要素」で現象定位が「現在：習慣的・恒久」のメッセージは右上に付置され、その修辞機能は「一般化」、脱文脈化指数は[14]である。また、円の大きさは出現数を示している。図３(b)に示したように、右上が脱文脈化の程度が高い修辞機能、左下が脱文脈化の程度が低い修辞機能、左上から右下の範囲は脱文脈化が中程度の修辞機能という分布になっている。

---

[13] 空間的距離のレベルと時間的距離のレベルの詳細については、佐野(2010b)を参照のこと。

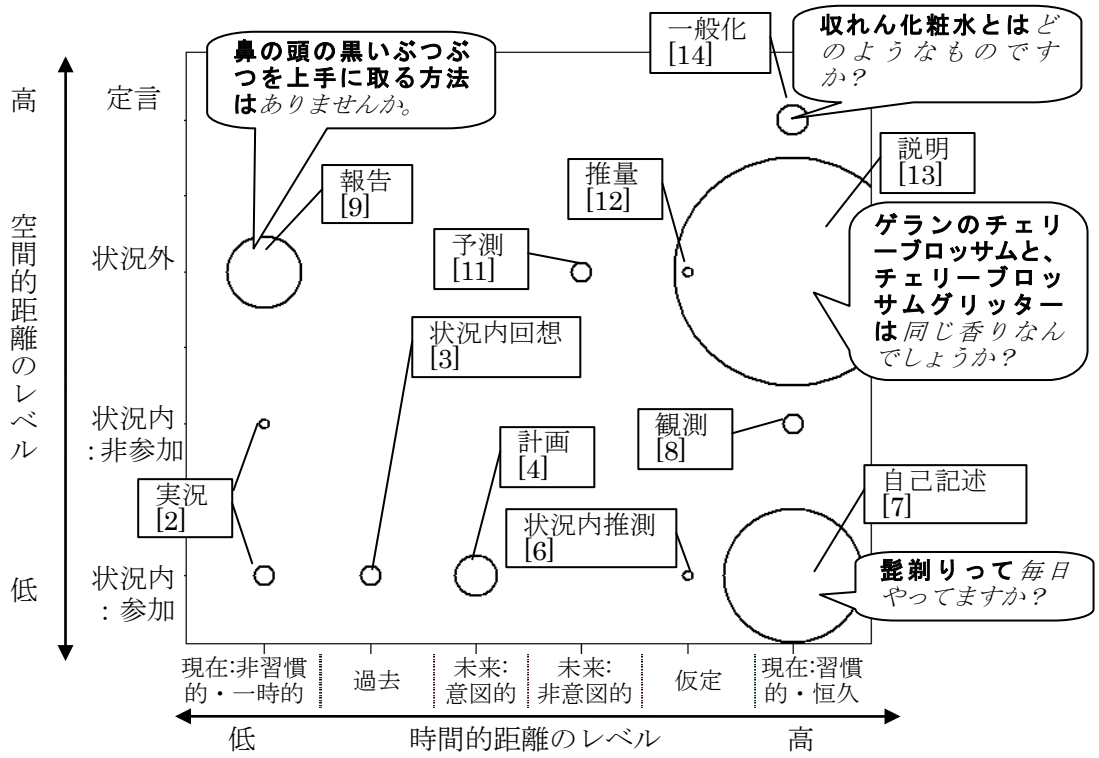


図３(a)　空間的距離と時間的距離からの Yahoo!知恵袋「コスメ、美容」の修辞機能の分布

図３(a)より、Yahoo!知恵袋の「コスメ、美容」では、一般的・抽象的な質問から個人的・具体的な質問まで、修辞機能が分布していることが明らかになった。図３(a)を図３(b)とあわせてみることによって、本研究の分析対象データにおいては脱文脈化の程度が高い修辞機能がより多く使われていることがうかがえるが、修辞機能や脱文脈化の程度の量的な分析については分析対象数を増やし、今後検討したい。

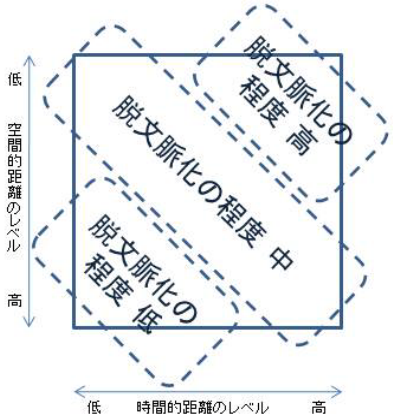


図３(b)　脱文脈化程度の分布

修辞機能や脱文脈化の程度を示す指数は、Web での検索時にも有効に利用できるのではないかと考える。例えば化粧品について質問するために Q&A サイトを利用する場合、例(13)であげた「ジョンソンから新発売のボディーソープ「ソフトウォッシュ」を使ってみた方、どんな感じでしたか？」のように具体的な商品についての実際に使っている人の感想を聞いてみたいという場合もあれば、あるメーカーから発売されているボディソープの種類を知りたい場合や、また、ボディソープそのものの性質や特徴、どう使うものなのかといった一般的なことを聞きたい場合もあるだろう。しかし、ただ「ボディソープ」というキーワードで検索したのでは、これらの情報が混在して抽出される。そこで、Q&A サイトへの投稿に修辞機能や脱文脈化指数を付与することができれば、キーワードとあわせて指定して検索することによって、感想などの個人的な情報、あるいは一般的な性質など、利用者が必要とする情報を抽出しやすくなると考えられる。

５．まとめ

本研究は、話し言葉の談話分析手法であったRUAをWeb上のテクストであるYahoo!知恵袋の「コスメ、美容」に関する質問に適用し、脱文脈化の観点からの分布を確認した。その結果、Yahoo!知恵袋ではおススメを聞いたり、今後の予定や計画を聞いたりする、脱文脈化程度が低い個人的・具体的な質問から、化粧品の性質のような一般的・汎用的な脱文脈程度が高いものまで、さまざまな質問がなされていることが明らかになった。

今後は、本研究の分析対象に含めた「質問的表現」以外の質問や、質問文以外へのRUAの適用を検討していく予定である。また、投稿内の修辞機能の展開が「質問」と「回答」でどのように対応しているかを分析することによって、Q&Aサイトの修辞機能や脱文脈化の程度の特徴を明らかにしたい。さらにYahoo!知恵袋以外のQ&Aサイトとの比較を行い、サイトの違いによる修辞機能や脱文脈化の程度の差の有無を明らかにしたいと考える。また各要素の認定の自動化に向けたリストの整備を進め、大規模データへの適用の可能性について検討していきたいと考えている。

謝辞　本研究は、文部科学省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）、及び、科学研究費補助金（基盤C）「書き言葉コーパスに基づくテキスト分類尺度の探索的研究」（平成21年度～23年度、研究課題番号：21520493、研究代表者：小磯花絵）による補助を得ています。本研究では、『現代日本語書き言葉均衡コーパス』領域内公開データ（2009年度版）に含まれる「Yahoo!知恵袋」のデータを利用させていただきました。記して感謝の意を表します。

文献

Cloran, C. (1999). Contexts for learning. In Christie, F. (ed.) *Pedagogy and the Shaping of Consciousness*, London: Cassell, 31-65.

Halliday, M.A.K. & Matthiessen. C. (2004).*An Introduction to Functional Grammar (3rd ed.)* London: Arnold.

栗山和子・神門典子(2009). Q&Aサイトにおける質問と回答の分析. 情報処理学会研究報告 Vol.2009, No.2, DBS-NO.148(19).

国立国語研究所. (1960).話しことばの文型 1 －対話資料による研究. 秀英出版.

――――――――. (1963).話しことばの文型 2 －独和資料による研究. 秀英出版.

佐野大樹(2010a).日本語における修辞ユニット分析の方法と手順 ver.0.1.1－選択体系機能言語理論（システミック理論）における談話分析－（修辞機能編）. http://researchmap.jp/systemists/資料公開/ 閲覧日 2011.1.19

――――― (2010b).選択体系機能言語理論を基底とする 特定目的のための作文指導方法について ―修辞ユニットの概念から見たテクストの専門性―. 専門日本語教育研究, 12, 19-26.

佐野大樹・小磯花絵(2011). 現代日本語書き言葉における修辞ユニット分析の適用性の検証－「書き言葉らしさ・話し言葉らしさ」と脱文脈化言語・文脈化言語の関係－. 機能言語学研究, 6, 掲載頁未定.

田中弥生. (2010). Q&Aサイトの「質問－回答」における結束性－省略の特徴分析－. 信学技報, 109(390), NLC2009-34, 7-12.

田村元秀・村上仁一・徳久雅人・池原悟. (2007). Web検索エンジンを用いたWhy型質問応答システムに関する研究(特許分類). 情報処理学会研究報告. 自然言語処理研究会報告, 2008(4), 15-21.

三浦麻子・川浦康至. (2008). 人はなぜ知識共有コミュニティに参加するのか：質問行動と回答行動の分析. 社会心理学研究, 23(3), 233-245.

『現代日本語書き言葉均衡コーパス』向け外字処理ツール

田島　孝治（データ班協力者：国立国語研究所理論・構造研究系）†
高田　智和（データ班分担者：国立国語研究所理論・構造研究系）

Uncoded Character Processing Tool for BCCWJ

TAJIMA, Koji　(National Institute for Japanese Language and Linguistics)
TAKADA, Tomokazu　(National Institute for Japanese Language and Linguistics)

１．はじめに

現代日本語を対象とした１億語規模の『現代日本語書き言葉均衡コーパス』(略称BCCWJ)の構築が，国立国語研究所において進められている。大規模コーパスは，自然言語処理研究だけでなく，日本語教育や，辞書編纂，言語政策など様々な分野において活用が期待されている（前川 2008)。BCCWJ の主要部分は，既存の電子化テキストを集積するのではなく，紙媒体で流通している書籍・雑誌・新聞などの言語資料に対し「サンプリング」，「電子化」の大きく二つの手順を行うことで構築される。

BCCWJ 構築における電子化の過程では，言語資料に存在する文字の字形を一定の粒度で記述できるように，文字セットとして JIS X 0213 を利用している。従来，言語資料の電子化には国内文字コード規格 JIS X 0208 が用いられてきた。しかし，言語資料には文字セットに存在しない文字（いわゆる外字）や異なる字形の文字（異体字）が登場することも多く，その扱いが問題となっていた。BCCWJ は JIS X 0213 文字セットを利用し，UTF-16 でエンコードされたデータファイルを作成することで，規模と質の両面で文字の記述性と再現性を向上させている。

しかしながら，多量の文献の中には文字セットで扱える範囲を超えた，算術記号や日本語以外の言語で利用する文字などが存在する。これらの文字を電子データ上で処理できるようにするには，特別な処理が必要である。しかし，資料の電子化過程においては，文字セットで表現できない文字は全て「**=**」として入力されているため，多量のファイルの中から「**=**」を含むファイルを探し，それぞれに文字に対する情報を追記していく必要がある。さらに，入力者によっては文字セットに対する知識の不足から，本来文字セットに含まれる文字を外字と判断し「**=**」を入力してしまう場合もある。

このため，多量の電子テキストから，「**=**」が入力された箇所を抽出し，対応する原資料を即座に表示して修正作業を行うことができる作業用ツールが必要である。

本稿では，まず BCCWJ の外字処理の概要について述べる。次に，外字処理を効率的に行うために開発した外字処理ツールについて特徴と機能を述べる。最後に，BCCWJ の構築時に開発ツールを利用した実例を示すと共に，ツールの利用によって得られた知見を述べる。

２．BCCWJ における外字処理の概要

BCCWJ で利用可能な文字セットは，JIS X 0123 文字セットから半角英数字など一部を除いて定義されており，11,176 種類の文字からなる。この 11,176 文字の集合を本稿では BCCWJ 文字セットと呼ぶことにする。

† t-koji@tela.cs.tuat.ac.jp

BCCWJ のデータファイルは XML 形式であり，テキストは UTF-16LE でエンコードされている。BCCWJ 文字セットの文字は，このエンコードを使えば表現できる。しかしながら，サンプリング対象となる新聞や書籍などの文献は，特殊な記号や絵文字，BCCWJ 文字セットに含まれない漢字を含んでいる場合がある。このような文字を本稿では外字と呼ぶ。

BCCWJ において，外字は XML 中のタグを用いて表現することになっている。図 2.1 に外字を XML タグで表した例を示す。原文を入力する際に外字は「=」としておき，その後，この文字に対して分析処理を行う。分析処理によって，=が表す外字が漢字やローマ字などの「文字」と判断できる場合には missingCharacter タグを記述する。一方，=が表す外字が☺や☎などの「絵」だった場合には，image タグを記述する。また，外字として入力されていた=が，実際には BCCWJ 文字セットで表現可能な文字という場合がある。この場合，外字であるという入力自体が間違いであるとし，適切な文字に変更する。

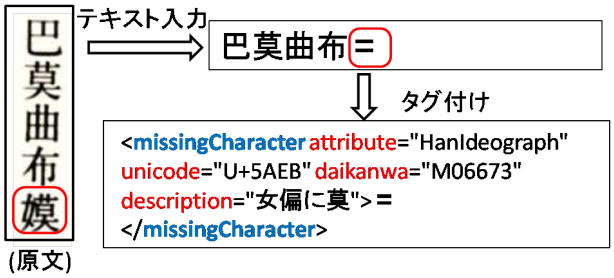


図 2.1　BCCWJ における外字のタグによる表現

外字の中でも「文字」を表現する際に用いる missingCharacter タグは次の属性を持つ。

- attribute (必須)：　文字種を表す属性で，次の値を取る。

属性名	文字種	属性名	文字種
HanIdeograph	漢字	Latin	ラテン文字
Hiragana	平仮名	Greek	ギリシア文字
Katakana	カタカナ	OldHanzi	古代中国文字
RomanNumeral	ローマ数字	etc	その他

- unicode (必須)：Unicode4.0 の 16 進コード

  書式は，4～5 桁の Unicode の先頭に「U+」を加えた 6～7 桁の文字列で，U+*****とする。Unicode で表現できない文字の場合は U+FFFD（REPLACEMENT CHARACTER）を記述する。

- daikanwa (任意)：　『大漢和辞典』の親字番号

  5 桁の諸橋番号の先頭に「M」を加えた 6 桁の文字列で表す。attribute が HanIdeograph となる場合は必須項目とする。『大漢和辞典』にない漢字は「M99999」を記述する。

- ref (任意)：　管理番号

  Unicode あるいは『大漢和辞典』にない文字の場合に記入する管理番号で，4 桁の管理番号の先頭に「KC」を加えた 6 桁の文字列とする。

- description (任意)：　字体記述，属性記述など任意の覚書き

文字に対する情報を任意の文字列で記入する。

外字の中で「絵」を表現する際に用いる image タグは次の属性を持つ。

- no (必須)： 出現番号

  サンプル内での出現番号を整数で記述する。

- description (任意)： 形状記述，(missingCharacter と同様の用途)

  文字に対する情報を任意の文字列で記入する。

３．外字処理支援ツール GETAWAY

３．１　ツールの概要

コーパスの作成にあたり製作した外字処理支援ツールを「GETAWAY」と名付ける。GETAWAY は，BCCWJ 構築における **=** 文字の処理を支援するためのソフトウェアである。本ツールは BCCWJ 文字セット内の文字かどうかの判断をユーザが効率的に行うと共に，その結果を元のファイルに簡単に反映させる機能を持つ。指定したフォルダに含まれる XML またはテキストファイルの解析を行い，ファイル中の **=** 文字の位置を抽出する。対象とするファイルは UTF-16LE でエンコードされている必要がある。その後，**=** で表現された文字に関する情報をユーザが入力できる GUI を提示する。ユーザは GUI を用いて **=** 文字に対する情報を入力する。入力結果から **=** 文字が BCCWJ 文字セット内の文字と判断できた場合は漢字などの文字に，外字の場合にはタグに置換する。置換処理時には，元ファイルの上書きを避けるため，"_replaced"を元のフォルダ名に追加した新たなフォルダを作り，その中に置換後のファイルを設置する。

表 3.1　ツールに含まれるファイルとその内容

ファイル名	内容
getaway2.00.jar	ツール本体
getarep.ini	環境設定ファイル
Unicode_heisei070731.csv	Unicode・JIS 面区点・平成明朝グリフの対応表
大漢和設定.csv	大漢和辞典の漢字番号と巻番号の対応表

３．２　GETAWAY に含まれるファイル

本ツールは**表 3.1** に示すファイルから成る。これらのファイルはすべて同一のフォルダに展開しておく必要がある。なお，対応表ファイルの作成にあたっては，「汎用電子情報交換環境整備プログラム」の成果物を利用している。

３．３　GETAWAY の使用方法

GETAWAY は GUI を用いて **=** 文字に対する処理を行う。**=** に対する情報は CSV ファイルにより保存する。なお CSV が完成した後には CUI から一斉置換を行うことも可能である。ここでは GUI の使い方について詳細に述べる。

GUI を利用したツールの操作手順を次に述べる。GETAWAY を通常起動した場合には，GUI により動作し，「ファイル」メニューのみの Window が開く。メニューからは以下の処理が実行できる。

(1) **=** を抽出して開く

フォルダ内に入ったテキストファイルを解析し，**=** を抽出する。

その後，タグ付け作業用の一時ファイルを作成し，目視作業を開始する。

(2) ファイルを開く

タグ付け作業の一時ファイルを開く。

(3) 上書き保存

タグ付け作業の一時ファイルを上書きして保存する。

(4) 名前を付けて保存

タグ付け作業の一時ファイルを別名で保存する。

(5) 選択結果で置換

= を含むテキストファイルに，ツールで付与したタグを追加し置換する。

(6) 終了

プログラムを終了する。

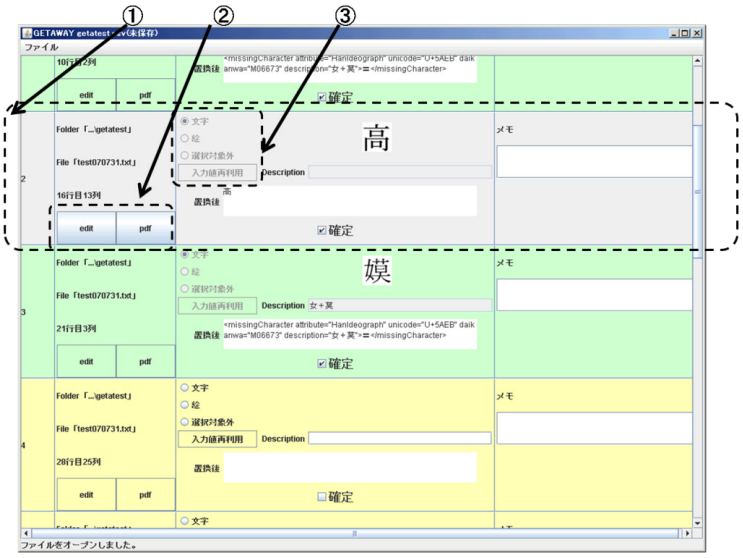


図 3.1　GETAWAY のユーザインタフェース

ユーザは次の手順で **=** の処理を行う。まず，メニューから「**=** を抽出して開く」を選択し，タグ付けを行う XML やテキストファイルが入ったフォルダを指定する。フォルダの指定時には参照用のダイアログを用いることもできる。フォルダの指定後，**=** 文字の抽出処理が行われ，図 3.1 のような結果が表示される。ファイルに含まれる 1 つの **=** に対し，①の一行が対応している。②の pdf ボタンを押すことで，原資料の PDF が自動的に表示される。また edit を押すと，原資料をテキスト化したファイルが表示される。**=** となっている文字が何かを確認した後に，ユーザは③のラジオボタンを押してタグなどの文字に関する情報の入力を行う。情報の入力処理は **=** が表す対象が「文字」か「絵」かにより手順が異なる。対象が「文字」の場合の GUI の遷移を図 3.2 に示す。ユーザは③のラジオボタンから「文字」を選択する。すると Unicode の入力を求めるダイアログが開く。原典を参照し，辞書などを使って文字を調べる。そして，文字を表現した Unicode があれば，このダイアログに入力する。一方 Unicode で表現できない場合には，FFFD を入力する。入力した Unicode により，その後の動作が変化する。文字が JIS X 0213 で表現可能で BCCWJ 文字セットに含まれている場合，グリフが表示され，意図した文字になっているかユーザに確認を促す。そして，置換後の結果は，文字そのものでありタグは作られない。一方，入力された Unicode が表わす文字が BCCWJ 文字セットに含まれない場合には，ツールは **=** が表す文字を外字と判断する。この場合は大漢和番号の入力を求めるダイアログが表示される。大漢和辞典に

掲載された文字の場合には大漢和番号を入力するとグリフが表示される。大漢和辞典に存在しない文字の場合は「無し」を選択する。大漢和辞典の掲載の有無にかかわらず description として文字の詳細を記述することが可能である。外字と判断された場合には，置換後の結果を表すテキストエリアに missingCharacter タグが表示される。

どちらに判断した場合も共に，置換後の結果を表す内容を確認し「確定」のチェックボックスをクリックすれば，一つの＝処理が完了となる。完了後の行は，BCCWJ 文字セットに含まれる文字の場合には灰色，外字の場合には緑色となり，視覚的に処理完了がわかりやすくなっている。

＝が表す対象が「絵」の場合には，③のラジオボタンから「絵」を選択する。すると description のみを入力するダイアログが開く。任意の注釈を入力して OK を押すことで image タグが生成される。

ある文字を表す＝が，特定のテキストファイルに多数存在する場合，入力値再利用ボタンを使うと効率的に処理できる。このボタンを押すと，過去に入力した＝の ID（ツールが割り当てた数値）を入力するダイアログが開く。ここに ID を入力すると過去に設定した値がコピーされ，文字番号や description を改めて入力する必要がない。

抽出された全ての＝に対して，タグ付けが完了したら，メニューより「選択結果で置換」を選択し，実行する。すると元のテキストファイル中の＝文字を一斉置換し，その結果を新たなファイルとして保存することができる。＝を処理中に一時的に本ツールを終了する場合には，メニューから「上書き保存」または「名前を付けて保存」を選択すれば途中結果を CSV 形式のファイルとして保存することができる。

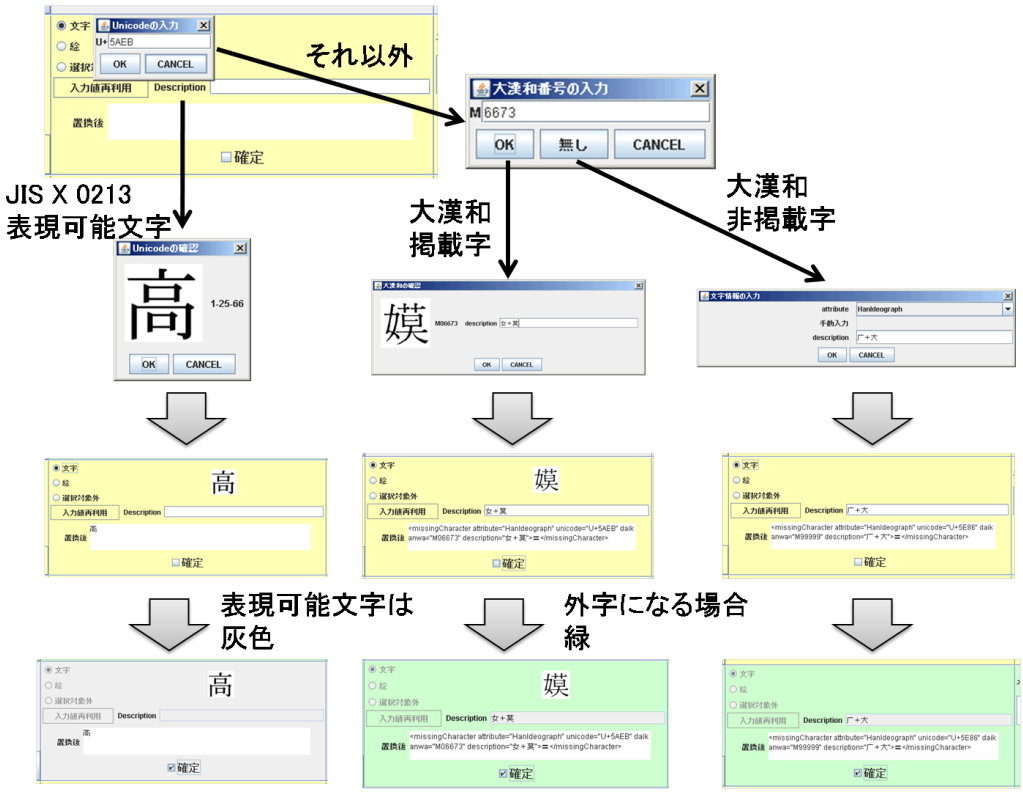


図 3.2　文字に対するタグ付け手順

３．４　設定ファイルの詳細

環境設定ファイルには，Unicode・JIS 面区点・平成明朝グリフの対応表のパス，PDF リ

ーダーのパス，テキストエディタのパス，データの一時保存先を記述する。環境設定ファイルの標準ファイル名は「getarep.ini」である。

３．５　ツールの実装と動作環境

本ツールは，プログラミング言語 Java を用いて製作した。このため，実行時には Java の VM が必要である。動作確認は，Windows XP 上で Sun Microsystems の Java 環境のランタイム（Java(TM) SE Runtime Environment (build 1.6.0_11-b03)）を用いて行っている。Linux 上でも Sun Microsystems の JRE 環境を構築することで利用可能である。ただし，これ以外の JavaVM を利用した場合には正しく動作しない恐れがあるため注意を要する。

４．BCCWJ 構築におけるツールの利用結果

2009 年に納品された BCCWJ のデータに対し，本ツールにより外字処理を行った。納品データは 3,997 ファイル，約 1,900 万文字であった。GETAWAY を使って**=**を抽出し，処理した結果を表 4.1 に示す。初めから入力されていた**=**の個数は 3,831 個であったが，外字処理により 2,998 字を BCCWJ 文字セット内のキャラクタに置き換えることができた。BCCWJ の文字セット内のキャラクタに置き換えられた文字は，多くが電話マーク（727 個）やハートマーク（187 個）などの絵文字であった。また，「艸」「裄」のように，特定のファイルのみで頻出する外字も多く，本ツールに搭載した過去の入力を再利用できる機能が有効に働くことが分かった。

一方で，電話マークやハートマークなどを，１ファイルに 100 個以上含むデータも存在した。このようなサンプルでは，過去の入力を再利用できたとしても入力の手間が大きい。この場合，入力環境の対応により，JIS X 0213 の文字が入力できない場合でも，初めから**=**以外の代用可能な文字を入力するなどの対策を取るべきである。

表 4.1　2009 年納品分データにおける=**の内訳**

分類	BCCWJ 文字セット内の文字		BCCWJ 文字セット外の文字		画像
区分	X0208	X0213	大漢和辞典掲載字	その他	-
個数	165	2,833	40	379	414
処理	**=** → キャラクタ		**=**→missingCharacter タグ		**=** → image タグ

付記

本ツールの作成には，「汎用電子情報交換環境整備プログラム」（経済産業省委託，平成 14 年度～平成 20 年度）の成果物を利用しました。

文献

前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉コーパス』の開発」 日本語の研究, 4:1, pp.82-95.

山口昌也，高田智和，北村雅則，間淵洋子，小林正行，西部みちる(2008) 「『現代日本語書き言葉コーパス』における電子化フォーマット ver2.0（国立国語研究所内部報告書 LR-CCG-07-03）」国立国語研究所

高田智和，小林正行，間淵洋子，大島一，西部みちる，山口昌也(2009) 「JIS X 0213:2004 運用の検証（国立国語研究所内部報告書 LR-CCG-09-01）」国立国語研究所.

長単位に基づく媒体・カテゴリ間の品詞比率に関する分析

冨士池優美 （データ班分担者：国立国語研究所コーパス開発センター）†
小西　光　（電子化辞書班協力者：国立国語研究所コーパス開発センター）
小椋　秀樹 （データ班分担者：国立国語研究所言語資源研究系）
小木曽智信 （電子化辞書班分担者：国立国語研究所言語資源研究系）
小磯　花絵 （電子化辞書班分担者：国立国語研究所理論・構造研究系）

Analysis of LUW-Based POS Ratio: In Relation to Media and Category

Yumi Fujiike (National Institute for Japanese Language and Linguistics)
Hikari Konishi (National Institute for Japanese Language and Linguistics)
Hideki Ogura (National Institute for Japanese Language and Linguistics)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)

１．はじめに

『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）には「コア」[1]と呼ばれるデータセットがあり，自動解析結果を人手修正した精度の高い「短単位」「長単位」情報が提供される。各サンプルには文章の内容を表すカテゴリ情報[2]が付与されている。

本発表では，長単位情報を利用し，中央官庁刊行の白書，書籍，新聞，雑誌， Yahoo!知恵袋（以下，知恵袋），Yahoo!ブログ（以下，ブログ）のコアデータを対象に品詞比率を調査し，サンプルの掲載媒体とカテゴリ情報の二つの観点から，文体との関係について検討する。

２．長単位の概要

長単位は，構文的な機能に着目し，文章の言語的特徴の解明を目的とした言語単位である。

長単位では「国立国語研究所」「品詞比率」「分析する」のような複合語を 1 単位として認める。「を」「だ」のような付属語は単独で長単位とするのが原則であるが，「における」「ている」のような複合辞も付属語として 1 長単位としている[3]。

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ず，その用法に基づき名詞・形状詞・副詞に判別する。「結果」（名詞-普通名詞-副詞可能）を例とすると，「これらの結果に基づき」の場合は名詞を，「結果、様々な社会問題が発生し」の場合は副詞を付与する。

† yfujiike@ninjal.ac.jp

1 「コアデータ」の設計については小椋ほか（2009）を参照。

2 カテゴリ情報は, BCCWJ において「ジャンル情報」として付与されている。詳細については丸山（2009）を参照。

3 認定基準の詳細については小椋ほか（2011）を参照。

３．調査の目的と対象

長単位に基づく品詞比率に関する研究は，これまでに小磯ほか（2009），冨士池ほか（2010）がある。小磯ほか（2009）は白書・新聞記事・社説・小説と講演を対象とした調査，冨士池ほか（2010）は白書・書籍・新聞の長単位コアデータを対象とした調査である。これらで分析対象となった長単位データは，名詞-普通名詞-〇〇可能が未判別であり，2 節で挙げた「結果」の例にはどちらも名詞が付与されている。

今回は，名詞・形状詞・副詞の判別を行い精密化したデータを利用し，対象となる媒体を増やして分析を行う。調査対象は白書・書籍・新聞・雑誌・知恵袋・ブログの 6 媒体から成る長単位コアデータである。表 1 に長単位コアデータの延べ語数を示す。資料規模の参考として，短単位延べ語数をあわせて示した。

表１　長単位コアデータ延べ語数

	白書	書籍	新聞	雑誌	知恵袋	ブログ
長単位	159019	199393	273441	200211	95094	99985
短単位	228272	234431	360825	245543	110696	118305

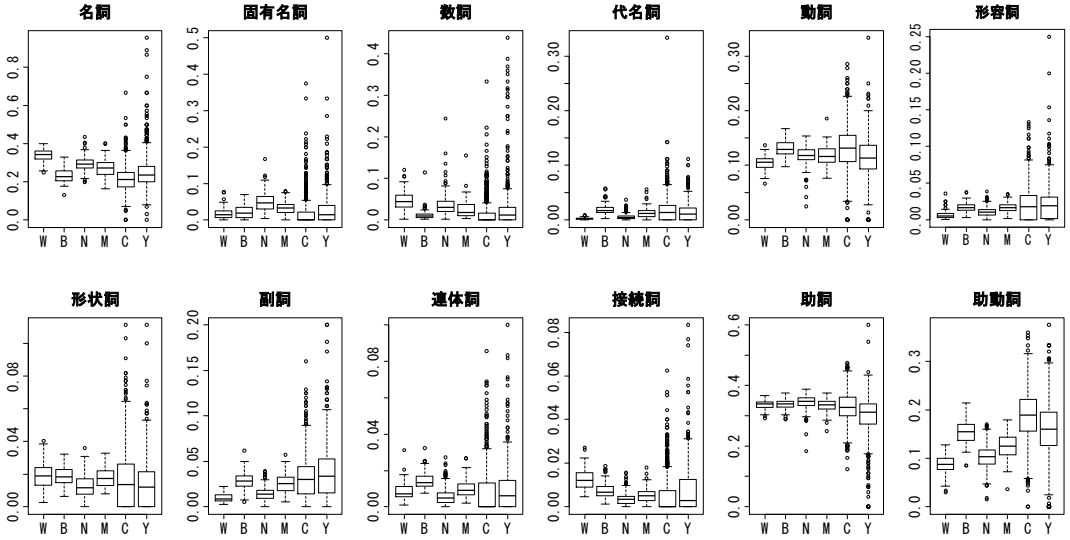
まず，掲載媒体という観点から品詞比率を調査し，文体との関連について考察する。ここで品詞比率に媒体差があった場合に，各媒体に含まれるカテゴリの偏りが擬似的に媒体差として現れた可能性もある。そこで，文章の内容を表す「カテゴリ情報」を観点として加え，あるカテゴリに限定して同様の調査を行い，全体の場合と同様の媒体差が現れるかを確認する。また，カテゴリの文体に与える影響についても検討する。

４．品詞比率

４．１　媒体別品詞比率

品詞比率（空白・記号・補助記号・URL 類を除く，延べ語数）を媒体別に示したものが図 1 である。「名詞」は固有名詞・数詞を除いたものである。

名詞率は媒体差が大きく現れており，知恵袋，書籍，ブログ，雑誌，新聞，白書の順に高くなっている。動詞率は，知恵袋・書籍と比較して，新聞・雑誌・ブログ・白書で比率が低くなっており，名詞率とほぼ負の相関にある。形容詞・副詞といった相の類の比率も動詞率と同様に名詞率とほぼ負の相関にあるが，ブログに関しては相対的に形容詞率が低く，副詞率が高くなっている。相の類の中で，連体詞率は形容詞率や副詞率と似た傾向を示すものの，知恵袋での比率が低い。また，形状詞率は相の類のほかの品詞とは傾向が異なり，媒体差が小さい。固有名詞率は新聞・雑誌で高く，数詞率は白書・新聞で高いのに対して書籍・知恵袋で低く，代名詞率は書籍で比率が高くなっており，これらは媒体ごとの内容の特徴が反映されたものと考えられる。助詞率は媒体差が小さく，ブログでやや比率が低くなっている。助動詞率は白書，新聞，雑誌，書籍，ブログ，知恵袋の順で高く，動詞や相の類と同様に，名詞率とほぼ負の相関にある。また，知恵袋とブログは他媒体と比較してサンプルの分散が大きく，品詞の別なく，極端に比率の高いサンプルがあることがわかる。



W：白書，B：書籍，N：新聞，M：雑誌，C：知恵袋，Y：ブログ

図１　品詞比率

冨士池ほか（2010）と今回の結果を比較してみよう。書籍の形状詞率，新聞の副詞率がより高くなり，白書は変化が小さかった。これらは，用法に基づき名詞・形状詞・副詞の判別をした結果，媒体の特徴がより明確になったものと考えられる。

相の類の比率は名詞率とほぼ負の相関関係を持つが，形状詞率のみ傾向が異なることを先に述べた。白書の名詞率は他媒体より際立って高く，負の相関関係がある相の類の一つである形状詞率は他媒体よりも低くなることが予想されるが，実際には白書の形状詞率は他の媒体と同程度であり，予想より高い。冨士池ほか（2010）では形状詞的接尾辞「的」の頻度が白書で高いことを示したが，名詞の形状詞化により形状詞率を高めている可能性がある。

４．２　媒体と文体の関連

ここで，体・用・相の三つの類の相関を見たい。樺島・寿岳（1965）は「100×相の類の比率／用の類の比率」で求められる MVR という指標を提案し，MVR と名詞の比率との組み合わせから，名詞の比率が大きく MVR が小さければ要約的な文章，名詞の比率が小さく MVR が大きければありさま描写的な文章, 名詞の比率が小さく MVR も小さければ動き描写的な文章と考えられるとしている。

図 2-1 は，体の類に対する MVR の分布である。体の類の比率（%）を x 軸，MVR を y 軸にとっている。

図から，知恵袋とブログは他の媒体と比較して分散が大きく，体の類に対する MVR の分布が類似していることがわかる。体の類の比率が小さく MVR が小さい，動き描写的なものを中心とし，体の類の比率が小さく MVR が大きい，ありさま描写的なものも多いが，体の類の比率が極端に大きい，要約的なものも少ないながら観察され，文体が一様でないことがわかる。

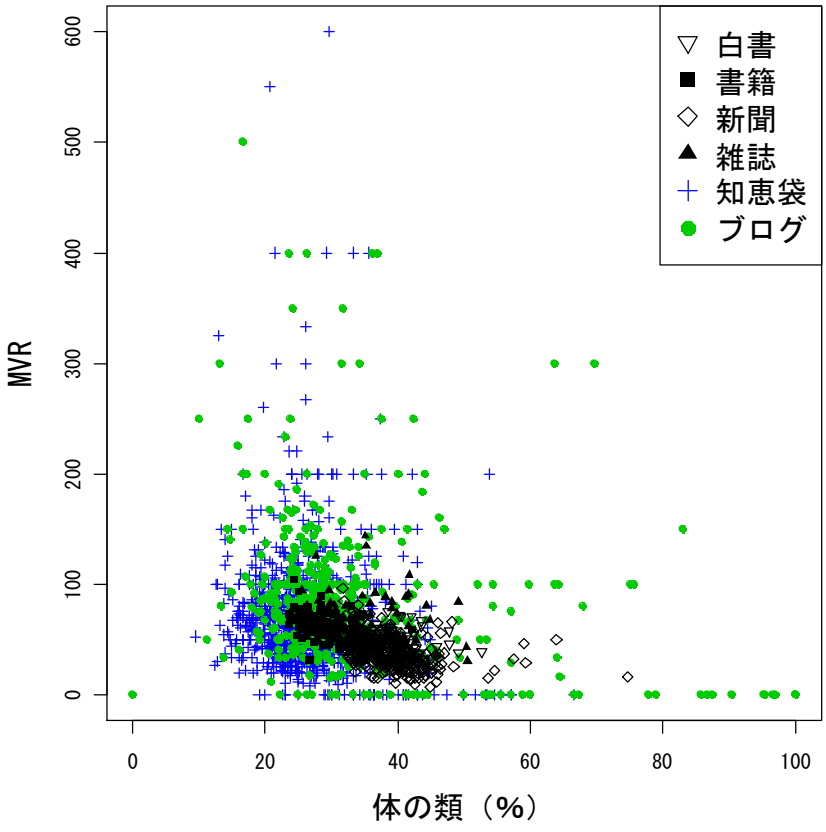


図 2-1　体の類に対する MVR の分布

サンプルの例を以下に示す。

体：大，MVR：小

ＪＮＢ→郵便局口座への振込み手数料はいくらでしたっけ？？？ ジャパンネットバンクの「郵貯Ｗｅｂ送金」のことですよね？振込手数料は２９４円です。

（知恵袋，体：62.9%，MVR：0）

◆サッカーＪ２第３６節　コンサドーレ札幌―湘南ベルマーレ　２７日午後２時、札幌ドーム（豊平区羊ケ丘１）。前売りＳＳ席４２００円、Ｓ指定席３７００円、ＳＡゾーン席３千円（小中学生千円）、ＳＢゾーン２５００円（同８００円）、Ｂ自由席２千円（同６００円）。当日券各２００円増し。北海道フットボールクラブ☎０１１・７５０・２９３６

（北海道新聞，体：87.2%，MVR：12.9）

体：小，MVR：大

クレヨンしんちゃんをみたのですが しんちゃんは、ぴちぴちおねいさんが大好きですが実際の５歳児もしんちゃんのようにおねいさんが好きなのでしょうか 嫌いではないでしょう。むしろ大好きでしょう。ただ、しんちゃんみたいに、積極的かどうか疑問です男の子だから、しょうがないと言ってしまえばそれまでですが・・・

（知恵袋，体：30.7%，MVR：600.0）

体：小，MVR：小

ニンニクが臭いというのは消化して食道の中から出てくる臭いですよね？
そのものをこんがりあぶると香ばしいのですが臭いという人がいます。にんにくの臭いは、消化して、血液の流れに乗り、肺にたどり着くのです。そして、呼吸とともに臭いがでてくるのです。

（知恵袋，体：20.9%，MVR：50.0）

次に白書・書籍・新聞・雑誌が集中する部分を見てみよう。図 2-1 のうち，知恵袋とブログを除いたものが図 2-2 である。概ね，書籍，雑誌，新聞，白書の順に体の類の比率が大きくなり，これに従い MVR が小さくなるのが見てとれる。要約的な文章と考えられるものに新聞と白書があり，新聞には極端に体の類の比率が高いものがある。これに対して，書籍・雑誌はありさま描写的な方向に分布している。特に雑誌はありさま描写的な傾向が強いが，要約的なものもあり，分散が大きい。

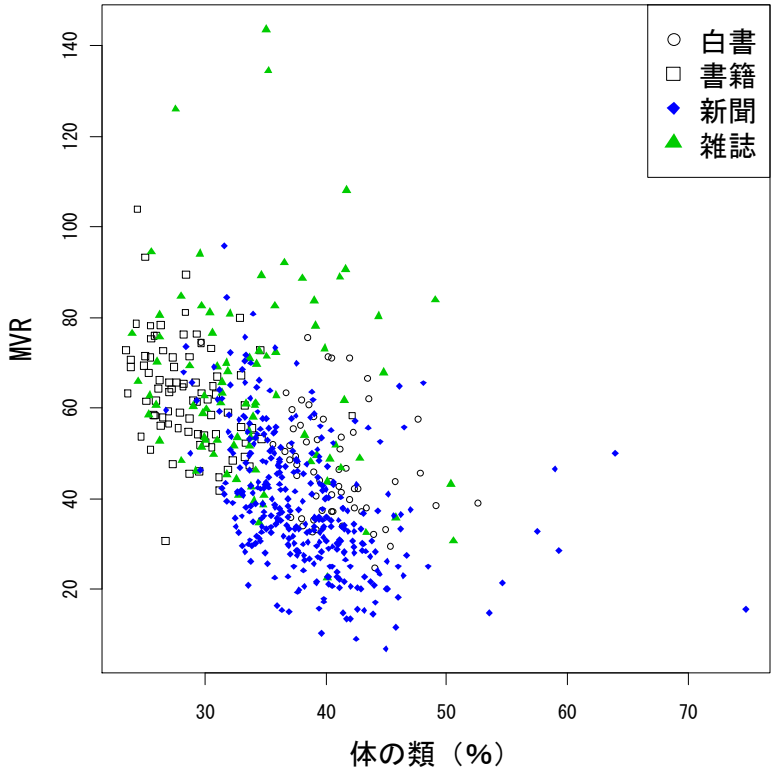


図 2-2　体の類に対する MVR の分布（知恵袋・ブログを除く）

サンプルの例を以下に示す。

MVR：大

カジュアルもキレイめもお手のもの！　かわいらしさが残るカジュアルなスタイリングは、マネしたいポイントがいっぱい。
ｓｃｅｎｅ　デート
デートの日は絶対ミニスカ！　フワフワファーで女のコらしく

（雑誌 CanCam，体：41.5%，MVR：136.2）

４．３　カテゴリ・媒体と文体の関連

BCCWJ では，各サンプルに文章の内容を表すカテゴリ情報が付与されている。具体的には，書籍の日本十進分類表（NDC），雑誌における『雑誌新聞総かたろぐ』の「分野」，知恵袋の質問が投稿されたカテゴリ名，ブログが投稿されたカテゴリ名である。白書についてはタイトルの内容に応じて国立国語研究所で独自に分類したものが付与され，新聞については内容を表す情報ではなく，配達エリア（全国紙・ブロック紙・地方紙）の別が示されている[4]。

品詞比率と文体の関連の研究においては，観点として形式（新聞の記事・社説，小説，短歌・俳句等）を設定することが多く，国立国語研究所の語彙調査では媒体（放送，雑誌等）を観点とするが，カテゴリ（内容）も文体に影響している可能性がある。4.2 節まで見てきた媒体差についても，各媒体に含まれるカテゴリの偏りが擬似的に媒体差として現れた可能性もある。そこで，カテゴリを限定した上で同じ分析を行い，全体の場合と比較し，4.2 節と同様の媒体差が現れるかを確認する。また，カテゴリの文体に与える影響についても検討する。

各媒体に共通するカテゴリとして，書籍の日本十進分類法 3 番台（社会科学）を中心に，雑誌・白書・知恵袋・ブログについてはその下位分類（社会科学，政治，法律，経済，財政，統計，社会，教育，風俗習慣・民俗学・ 民族学，国防・軍事）[5]と共通・類似した名称を持つもの[6]を選定し， 5 媒体から表 2 にあるカテゴリを対象とした。新聞については，カテゴリ情報である「配達エリア」からは内容が判別できないため，除外した。

表 2　調査対象カテゴリ情報

媒体	カテゴリ情報	サンプル数
書籍	**社会科学**	18
雑誌	政治・経済・商業	7
白書	安全	30
	外交	
	教育	
	経済	
知恵袋	ビジネス、経済とお金	78
	ニュース、政治、国際情勢	
ブログ	ビジネスと経済	29
	政治	

4 ここで選定したカテゴリ情報とは BCCWJ の「ジャンル(1)」を言う。

5 NDC 新訂 9 版分類表（2 次区分表）による。

6 雑誌には「教育・学芸」のカテゴリがあるが，これは文芸雑誌の小説・批評に付与されているため，除外した。また，知恵袋の「子育てと学校」，ブログの「学校と教育」は，主に勉強法や学校生活などに関するものであり，社会科学系とは言い難いため，除外した。

これらの社会科学系サンプルについて，体の類に対する用・相の類の割合の関連を見るために，体の類に対する MVR の分布を図 3 [7]に示す。

図 3 を見ると，全体の傾向を示した図 2-1 と同種の傾向，例えば白書は体の類の比率が高く，知恵袋とブログは体の類の比率が小さく，MVR が小さいものもあれば大きいものもあり，分散が大きいことが見てとれる。このことから，4.2 節で見た媒体差は，各媒体に含まれるカテゴリの偏りが擬似的に現れたものではないことがわかる。

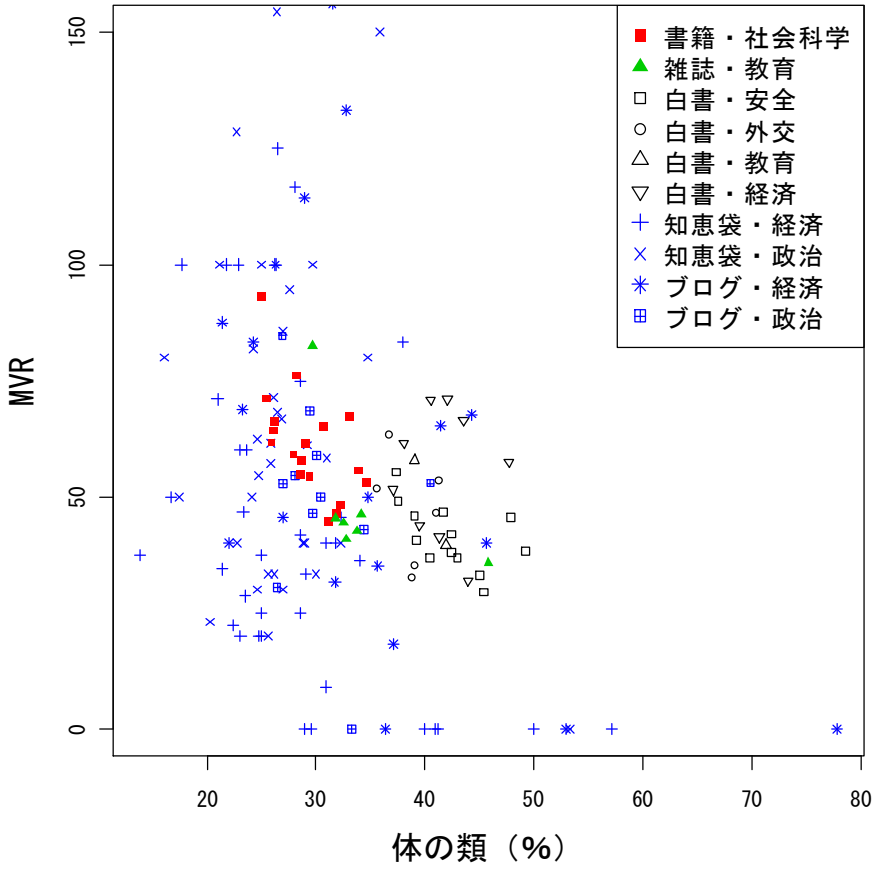


図 3　体の類に対する MVR の分布（社会科学系）

細かく見ると，白書・知恵袋については全体の場合とほぼ同様の分布となっているが，雑誌については全体とは若干異なる傾向が見られた。社会科学系に限定した結果，体の類の比率はほぼ差がなかったが，MVR は全体で 22.5 から 143.7 まで分布していたものが社会科学系では 35.8 から 82.5 とやや低い位置での分布となっており，形容詞・副詞の類が抑制されている傾向が見られた。書籍についても同様に，MVR が小さく，相の類が抑制される傾向が見られた。ブログは，カテゴリ情報が経済か政治かによって若干傾向が異なる。経済に関するブログ記事は全体の場合とほぼ同様の分布，つまり分散が大きくなっているのに対して，政治に関するブログ記事は雑誌や書籍と同様の位置に分布しており，MVR が小さく，相の類が抑制される傾向が見られた。

[7] 図 3 は y 軸を 0 から 150 までに取っているが，150 から 600 の間に知恵袋 7 サンプル，ブログ 1 サンプルが分布している。

このように社会科学というカテゴリでは，全体と比べた場合に MVR が相対的に小さくなる，つまり形容詞・副詞の類が抑制される傾向が，雑誌・書籍・ブログ（政治）に共通して見られることから，媒体とは別にカテゴリが品詞比率に影響を与えている可能性のあることが示唆される。白書・知恵袋・ブログ（経済）などではこの影響は観察されなかったが，白書については行政報告書という媒体自体の制約が強いことに起因している可能性がある。ブログ（経済）には，個人的な経済状況の相談など，社会科学系ではないサンプルが含まれていることに起因していると考えられる。知恵袋もブログ（経済）同様に，カテゴリ名の「お金」「ニュース」に該当する社会科学系ではないサンプルが含まれている。

５．まとめ

媒体別の品詞比率と，サンプルを社会科学系に絞った場合の品詞比率から，①名詞と動詞・形容詞・副詞・助動詞の比率はほぼ負の相関関係にある，②白書・新聞は書籍・雑誌と比較して要約的，雑誌はありさま描写的な傾向が強く，知恵袋・ブログは品詞比率の分散が非常に大きい，③カテゴリを限定しても媒体による品詞比率の差が見られる，つまり媒体差は各媒体に含まれるカテゴリの偏りが擬似的に現れたものではない一方で，媒体によってはカテゴリを限定することで全体と異なる傾向が共通して見出されたことから，カテゴリが品詞比率に影響を与えている可能性があるということがわかった。

参考文献

小椋秀樹ほか（2009）　「『現代日本語書き言葉均衡コーパス』における形態論情報付与作業の進捗状況」『特定領域「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』pp.57-64

小椋秀樹ほか（2011）　国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集　第 4 版』

樺島忠夫・寿岳章子（1965）　『文体の科学』（綜芸社）

小磯花絵ほか（2009）　「長単位情報に基づくジャンル間の文体に関する分析」『特定領域「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.183-190

冨士池優美ほか（2010）　「『現代日本語書き言葉均衡コーパス』長単位情報に基づく予備的分析」　『特定領域「日本語コーパス」平成 22 年度全体会議予稿集』pp.101-108

丸山岳彦（2009）　「『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）サンプリング方法について」（『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）DVD 所収）

ＢＣＣＷＪに基づくオノマトペの品詞と意味についての分析

宮内佐夜香（データ班協力者：国立国語研究所コーパス開発センター）†
小木曽智信（電子化辞書班分担者：国立国語研究所言語資源研究系）
小磯　花絵（電子化辞書班分担者：国立国語研究所理論・構造研究系）
小椋　秀樹（データ班分担者：国立国語研究所言語資源研究系）

Part-of-Speech and Meaning of Onomatopoeic Word: Analysis Using BCCWJ

Sayaka Miyauchi (National Institute for Japanese Language and Linguistics)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)
Hideki Ogura (National Institute for Japanese Language and Linguistics)

１　はじめに

日本語のオノマトペは品詞論的に多様なふるまいをすることが多くの先行研究において言及されているが，その品詞性の判断基準にはさまざまな議論があり，また個々の語の特性によるふるまいの差が大きいことから考察の範囲を制限した論考が多く，これまで大きな傾向を把握するには至らなかったことが指摘される。そこで宮内ほか(2011)では，(1)『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）を利用してオノマトペの用例を広く収集する，(2)品詞性判断の指標を明確にするために後接要素という形態的な情報のみを用いる，(3)多量のデータを統計的手法で分析する，という方法によって，オノマトペ全体の品詞論的な分類を試みた。本研究では品詞性の分類を行った上で，各分類の語群がどのような意味傾向を持つのかを数量的に調査し，品詞性と意味的特徴との関連を確認する。

２　先行研究

後接要素を指標として意味との関連を論じた研究としては，宮地(1978)がある。宮地(1978)は副詞のうち擬音語・擬態語について「ゼロ型」「に型」「（と）する型」「と型」と 4 種に整理し，各種の意味を考察しており，形態論的な類別が有効なことを示している。

佐々木(1986)は，擬態語について「ト語尾・ニ語尾が下接するか否かで分類調査」を行ったもので，「ニ語尾」は「静的意味合」であり，「ト語尾」の多くは「動的意味合」を持つこと，「ト語尾」を取るものの中には「サ変動詞の語幹」としても用いられる「静的意味合」を持つものがあることなど，「動的」「静的」という概念でそれぞれの意味を論じている。

また田守(1993)には「する」の付加する形式について「人間の心理状態を記述する「擬情語」は《中略》例外なく動詞組み入れが可能である」(p.45)と述べられている。

加藤・坂口(1996)は上記の諸研究を踏まえた上で，「後接成分」の表れ方による分類を試みている。副詞的な用法以外についても言及されている点が注目され，その中で「に」と「だ」が「結果の状態」を表すものとして「同様の働き」であるという指摘がある。

本研究では以上のような先行研究を踏まえた上で，多量の用例に基づいた数量的な分析を行い，個別の検討からは明確になりにくかった全体的な傾向の把握を目的とする。

† smiyauchi@ninjal.ac.jp

３　方法

以下３.１～３.３に述べる後接要素を指標とした分類の方法は宮内ほか(2011)に同じである。３.４にはその結果を踏まえた意味的な傾向の調査方法を述べる。

３.１　分析データ

分析データにはBCCWJ内のサンプルのうち，白書，新聞，雑誌，書籍，Webデータ（Yahoo!知恵袋・Yahoo!ブログ）を使用した[1]。BCCWJには短単位・長単位の２種の言語単位に基づき形態論情報が付与される（形態論情報の詳細は小椋ほか2011参照）。今回はこのうち短単位を言語単位として用例を収集した。今回利用したデータは，短単位を採用した形態素解析用電子化辞書UniDicを用いてMeCabによって自動解析されたものを元に，一部に人手修正を加えたものである。サンプル数と延べ語数は表１のとおりである。

表１：分析データの内訳

コーパスの種類	サンプル数	延べ語数
白書	1,529	5,000,185
新聞	1,489	1,391,029
雑誌	2,439	5,525,606
書籍（生産実態）	10,277	29,289,845
書籍（流通実態）	10,640	30,669,726
書籍（ベストセラー）	1,516	4,002,216
Yahoo!知恵袋	45,725	5,190,722
Yahoo!ブログ	52,680	10,453,668
計	126,295	91,522,997

３.２　用例の収集

分析対象は一般に擬音語・擬態語とされるものを広く対象として選定した。「かくかく」「がっしり」等の様態を表すもの，「とんとん」「からん」等の音を表すものの他，和語の実質的な意味との関連も考えられる「つやつや」「ねじねじ」「ひんやり」などの語も対象とした。漢語系の「凛々」「燦々」等は対象外とした。短単位においては，オノマトペは単体で１語とし「と」「に」などの後接要素は原則的に別語として切り離す。用例収集においてはオノマトペとそれに続く後接要素を分析対象として抽出した[2]。ただし特殊な語形を除くために，１サンプルにしか現れない語は除外した。

次に後接要素については，表２のような要素を指標とし，これに当たらない用例は分析対象外とした。対象外のものは文末での使用（例：よく寝てすっきり。），複合語の構成要素（例：しっとり感），間に他の成分を挟んで用言を修飾する副詞用法等である。

以上のような基準により抽出された用例は異なり1,719語，延べ127,458語であった。例として総計上位語の後接要素別の頻度を表３に示す。

表２：オノマトペ後接要素一覧（付・短単位の品詞情報）

後接要素ラベル	内訳	短単位の品詞情報
ト	出現形が「と」のもの	助詞-格助詞「と」(引用含む)
ニ	出現形が「に」のもの	助詞-格助詞「に」・助動詞「だ」連用形-ニ
ナ	出現形が「な」のもの	助動詞「だ」連体形-一般
ダ	出現形が「だ」「だっ(た)」・「です」	助動詞「だ」終止形・連用形-促音便・助動詞「です」
デ	出現形が「で」のもの	助詞-格助詞「で」・助動詞「だ」連用形-一般
ノ	出現形が「の」のもの	助詞-格助詞「の」
スル	動詞「する」	動詞「する」
用言	動詞(「する」以外)・形容詞・形容動詞	動詞・形容詞・形状詞
格助詞	「と」「に」「で」「の」を除く格助詞	助詞-格助詞

[1] オノマトペを広く収集するため多種の媒体を選択したもので，本研究ではこれらの媒体間の違いには特に言及しない。

[2] 短単位では「と」を含めて１語の副詞として情報を付与しているものもある（「じっと」「どっと」「ぞっと」「ぼうっと」等）。こうした副詞は分析対象外とした。

表 3：オノマトペ後接要素別頻度表（総計上位語）

	ト	用言	スル	格助詞	ダ	ナ	デ	ノ	ニ	計
はっきり	2,747	2,560	2,979	0	0	1	6	0	0	8,293
ゆっくり	3,745	1,876	404	1	49	19	52	11	24	6,181
たっぷり	583	1,165	39	2	48	49	78	617	160	2,741
どんどん	114	2,329	12	2	0	0	3	2	1	2,463
すっかり	23	2,312	2	1	0	0	1	0	0	2,339
さっ	1,652	11	4	1	2	1	0	0	0	1,671
…	…	…	…	…	…	…	…	…	…	…
総計	**63,670**	**30,723**	**21,028**	**669**	**1,938**	**640**	**1,339**	**3,749**	**3,702**	**127,458**

３.３　後接要素に基づく分析

表３のようなオノマトペの後接要素別頻度表に対して対応分析を行ない，オノマトペの各語と後接要素の相対的な位置関係を求めた。分析にはRのMASSパッケージの中のcorresp関数を用いた。さらに距離の近いオノマトペをグループ化するために，対応分析で得られた各オノマトペのスコア（2次元）を対象にクラスター分析を行なった。これにはhclust関数を用い，ユークリッド距離＋群平均法によって分析した。

３.４　意味的傾向の調査

３.３の手法によって分類した各グループの所属語の意味的な傾向の分析をするために，今回は『分類語彙表　増補改訂版』（国語研究所 2004。以下『分類語彙表』）を利用した。ただし，短単位と『分類語彙表』の見出し語は必ずしも一致せず，またそもそも『分類語彙表』に採録されていない語も多くある。そのため，今回は『分類語彙表』との一致が見られた語のみを意味分析の対象とした。また，本研究では後接要素として扱う「(っ）と」を付加することで『分類語彙表』の見出し語と一致する語が多くあるため（「さっ（と)」等)，それらについてはその情報を利用し，分析対象とした。

４　品詞性に基づく分類

宮内ほか(2011)で行った品詞性に基づく分類について以下に記述する。対応分析とクラスター分析の結果から，オノマトペについて６つのグループを抽出した。所属する語の頻度上位10語と各グループの所属語の総語数（延べ・異なり）を表 4に示した。また各グループの後部要素の頻度を表 5に，頻度1位の語の後部要素の頻度を表 6に示した。さらに対応分析の結果から，代表語として表 4に示した上位10語と後接要素の相対的位置を表示したものが図 1，所属語全ての位置のみを図 1 と同じく後接要素とともに表示したものが図2である。以下，上記分析で分けられたオノマトペの各グループと後接要素の位置の対応を，表 5に示した後接要素の頻度を参照しながら見ていく。

［グループ１］は〈ト〉を中心に位置している。所属語は1音の「ば（っと)」2音の「ちら（っと)」など通常「ト」を付加するタイプの語が所属している。その他「ゆったり」「ちらちら」等単独でも連用修飾成分となる語が多いが，こうした〈ト〉の有無どちらも可能なものであっても頻度上〈ト〉付加に偏る語群が，ここに所属していると言えるだろう。

［グループ２］は主に〈用言〉周辺に位置しており，単独で連用修飾成分となる性質の強い語群と言えるが，〈ト〉や〈スル〉の頻度も高い。連用成分として多様な後接要素を取り得る一群と考えられる。

［グループ３］は〈スル〉周辺に位置しており，サ変動詞として使用される頻度が高い

表 4：オノマトペ・６グループ頻度上位 10 語と所属語数

	グループ1		グループ2		グループ3		グループ4		グループ5		グループ6		総計
	ゆっくり	6,181	はっきり	8,293	すっきり	1,490	たっぷり	2,741	ばらばら	1,044	びしょびしょ	62	
	さっ	1,671	どんどん	2,463	どきどき	872	ぴったり	1,659	ぎりぎり	800	こてんぱん	22	
	ゆったり	1,102	すっかり	2,339	がっかり	777	さらさら	463	ぼろぼろ	593	べろんべろん	10	
	ばっ	1,092	ぼんやり	1,551	にこにこ	674	ふわふわ	366	びかびか	339	けちょんけちょん	7	
	ぐっ	1,018	さっぱり	1,315	うんざり	628	がたがた	326	どろどろ	260	べこべこ	5	
	ちら	894	のんびり	1,265	わくわく	559	ばりばり	291	からから	210	こちんこちん	3	
	ふっ	699	じっくり	1,119	ごろごろ	504	ばっちり	277	くたくた	181	ぐちょぐちょ	2	
	ちらり	691	にっこり	824	うろうろ	491	ごちゃごちゃ	251	ぐちゃぐちゃ	178	ぽっかぽか	1	
	すっ	675	そろそろ	803	にやにや	386	ちょい	249	くしゃくしゃ	177	ねじねじ	1	
	ぴん	675	きらきら	698	ぐずぐず	319	つるつる	235	とろとろ	146			
上位10語計		14,698		20,670		6,700		6,858		3,928		113	52,967
総語数・延べ		54,004		43,330		12,216		11,860		5,935		113	127,458
総語数・異なり		1,015		355		95		155		90		9	1,719

表 5：オノマトペ６分類の後接要素　総計

	グループ1	グループ2	グループ3	グループ4	グループ5	グループ6	計
ト	45059	13157	1786	3021	647	0	63,670
用言	6194	19577	1166	3410	375	1	30,723
スル	1687	9340	8691	1048	262	0	21,028
格助詞	212	143	72	203	39	0	669
ダ	189	302	258	699	483	7	1,938
ナ	50	86	50	299	154	1	640
デ	207	201	56	468	402	5	1,339
ノ	249	294	101	1814	1282	9	3,749
ニ	157	230	36	898	2291	90	3,702
計	54004	43330	12216	11860	5935	113	127458

表 6：頻度１位語の後接要素

	1	2	3	4	5	6
	ゆっくり	はっきり	すっきり	たっぷり	ばらばら	びしょびしょ
ト	3745	2747	357	583	88	0
用言	1876	2560	225	1165	13	1
スル	404	2979	881	39	2	0
格助詞	1	0	2	2	0	0
ダ	49	0	12	48	102	6
ナ	19	1	3	49	70	1
デ	52	6	2	78	98	4
ノ	11	0	7	617	95	6
ニ	24	0	1	160	576	44
総計	6181	8293	1490	2741	1044	62

ものが所属している。

［グループ４］は〈格助詞〉に近い位置にある。〈格助詞〉は全用例で 0.5%程度だが，このグループでは 1.7%と相対的に高い頻度となっており，特徴的である。後接要素の全体の傾向は〈ト〉や〈用言〉の頻度が特に高いが，〈ダ〉〈ナ〉〈デ〉〈ノ〉の後接する頻度も［グループ１，２］などより高くなっていることが指摘される。この一群は副詞的にも形容動詞的にも，名詞的にも働く多機能な語群であると言える。他に〈格助詞〉が現れるのは主に［グループ１，２］で，概ね副詞的なグループに所属していることが分かる。

［グループ５］は〈ダ〉〈ナ〉〈デ〉〈ノ〉から〈ニ〉周辺にかけて位置する。［グループ６］とは異なり〈用言〉〈ト〉の頻度が比較的少ない。この一群は［グループ４］同様に多機能とも言えるが，より形容動詞的な性質の強い語が所属していると言える。

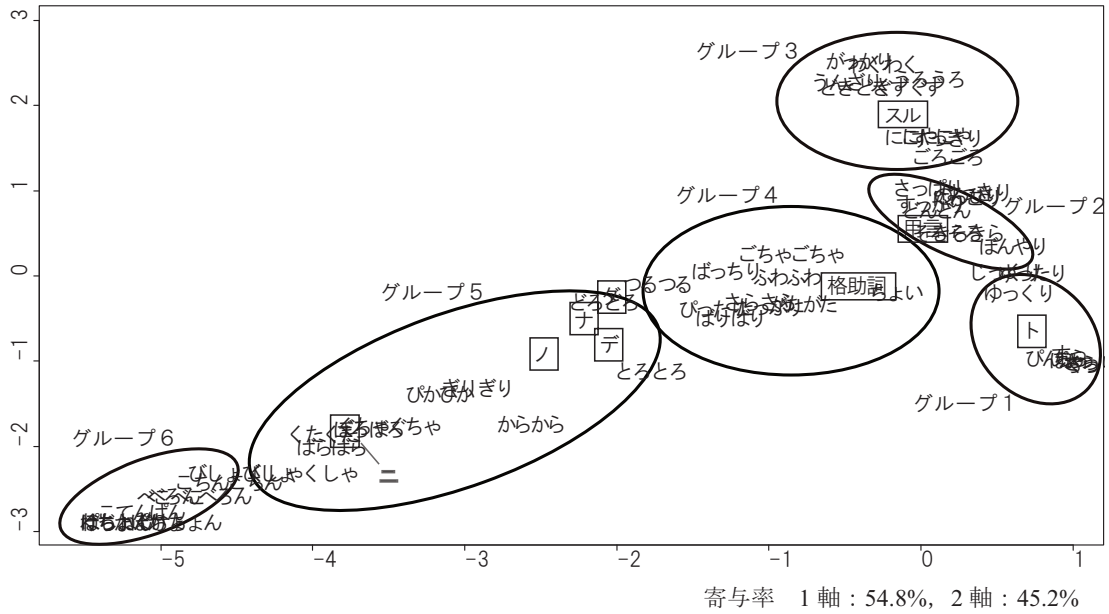


図 1：各グループ頻度上位 10 語・後接要素の相対的位置

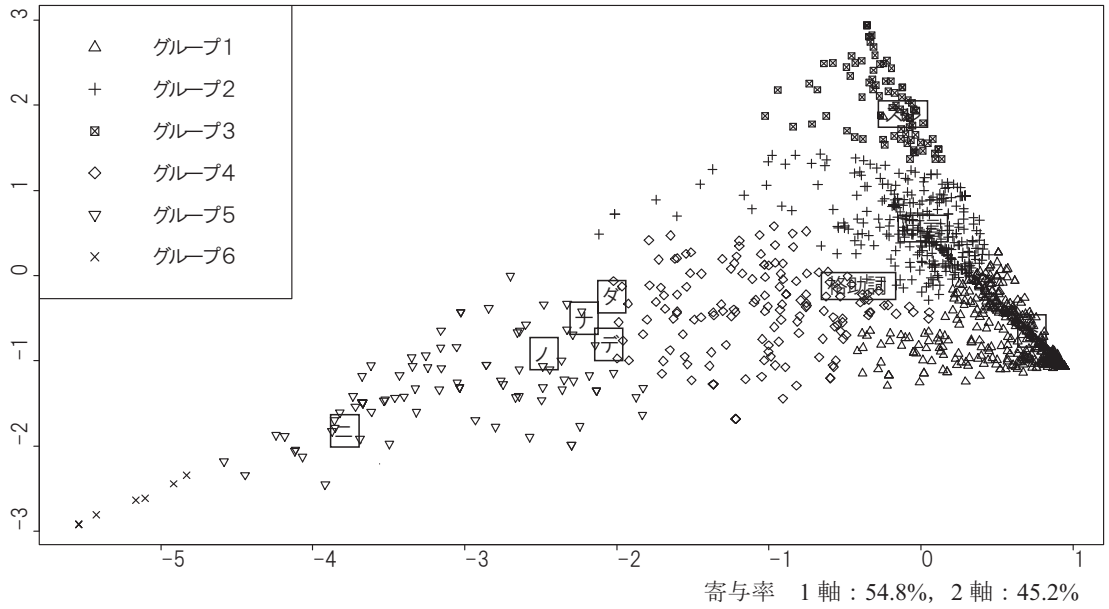


図 2：各グループ所属語と後接要素の相対的位置

［グループ６］は全体に低頻度語であり，他の分類との用例数の差が大きいが，これはほとんどの場合〈ニ〉を後接して用いられる語が所属している。

以上から宮内ほか(2011)ではオノマトペの語群の品詞性を表 7 のようにまとめた。

表 7：オノマトペの品詞性分類

グループ	品詞性の分類	異なり語数
1	ト付加型副詞・名詞	1015語
2	副詞（単独）・ト付加型副詞・スル動詞・名詞	355語
3	スル動詞	95語
4	多機能型（副詞的・名詞）	155語
5	多機能型（形容動詞的）	90語
6	ニ付加型副詞	9語

５　『分類語彙表』との対応

５.１　所属語の一般性

３.４に述べた方法で『分類語彙表』と今回の調査語の対応を確認した結果を，グループごとに表 8 に示す。調査語計 1719 語（異なり）中 825 語が『分類語彙表』の見出し語と一致した。また，例えば「さっぱり」が「関係-量-限度」「関係-量-一般・全体・部分」「活動-心-快・喜び」「活動-生活-文化・歴史・風俗」「自然-自然-味」など複数の分類項目に記載されている等，一つの語形が複数の意味を持つことがあるため，対応する『分類語彙表』の見出し語数（③）が一致した調査語数（②）よりも多くなっている。その語数も示した。

表 8：調査語と『分類語彙表』見出し語との対応

	①所属語数（異なり）	②一致した語数	③対応した見出し語数	④カバー率（②/①）	品詞性
グループ1	1015	402	563	39.6%	ト付加型副詞・名詞
グループ2	355	222	354	62.5%	副詞（単独）・ト付加型副詞・スル動詞・名詞
グループ3	95	71	121	74.7%	スル動詞
グループ4	155	84	140	54.2%	多機能型（副詞的・名詞）
グループ5	90	41	65	45.6%	多機能型（形容動詞的）
グループ6	9	5	5	55.6%	ニ付加型副詞
計	1719	825	1248	48.0%	

さて，調査語（①）のうち『分類語彙表』の見出し語と一致した語（②）の割合をカバー率と称して表 8 に示した（④）。カバー率が最も高いのは［グループ３］のスル動詞で，次いで［グループ２］の副詞全般が高い。最も低いのは［グループ１］のト付加型副詞である。『分類語彙表』の見出し語を日常的に使用される一般的な語彙と考えるならば[3]，スル動詞化する語群（「うとうとする」「すっきりする」等）や「と」付加なしに副詞として用いられる語群（「ばんばん叩く」「はっきり見える」等）には，オノマトペの中でも一般化した固定的なものが多く所属していると言える。対してト付加型副詞の語群には，個々の具体的な現象を表現するような比較的臨時性の強い語形（「ぼわーんと響く」「ぽぽんと投げる」等[4]）が多く含まれていることが，カバー率の低さに現れていると言えそうである。［グループ１］の所属語数の多さも臨時性と関連付けられるだろう。

５.２　各グループの意味的傾向

次に『分類語彙表』の項目の情報を参照して，各グループ所属語の傾向を確認する。項目のうち部門と中項目から表 9 のようなラベルを設定し，ラベル別の調査語の頻度と比率を表 10 に示した。中項目についても各グループ内の比率を図 3 に折れ線グラフで示した。

［グループ１］のト付加型副詞は「関係」に偏っており，中項目で言えば「作用」の比率が高い。これは「ぱっ（と）」「さ（っと）」等「動き」を表す語が分類されたものであり，「ト型」は「動的」という佐々木(1986)の指摘に合致する。［グループ３］のスル動詞は「活動」の比率が高く，中項目は「心」が 43%を占めている。田守(1993)の「人間の心理状態を記述する「擬情語」」が多くスル動詞化するという指摘と合致するものである。これらは従来の指摘が数量的にも確認されたものと言える。

［グループ２］は副詞（単独），ト付加型副詞，スル動詞等，副詞的という共通性はある

[3] 『分類語彙表』まえがき「２．収録した語句について」参照。

[4] これらは同時にオノマトペそのものの形態として「と」なしでは副詞として働かない語形である（*ぼわーん響く）。こうした形態と臨時性の関連も考えるべきだが，今回は言及しない。

表 9：意味に関するラベル
（『分類語彙表』を参照）

ラベル	『分類語彙表』 類・部門	分類番号
関係	相の類・ 抽象的関係	3.1
活動	相の類・ 人間活動 ―精神及び行為	3.3
自然	相の類・ 自然物及び自然現象	3.5
鳴き声	その他の類	4.5(動物の鳴き声)
感動	その他の類	4.30(感動), 4.31(判断)
体言	体の類	1.1(関係), 1.4(生産物), 1.5(自然)

表 10：各グループ所属語の
『分類語彙表』部門別頻度・比率

グループ	関係	活動	自然	鳴き声	感動	体言	総計	品詞性
1	233 41.4%	134 23.8%	186 33.0%	7 1.2%	1 0.2%	2 0.4%	563 100%	ト付加型副詞
2	118 33.3%	116 32.8%	106 29.9%	4 1.1%	6 1.7%	4 1.1%	354 100%	副詞
3	22 18.2%	67 55.4%	31 25.6%	1 0.8%	0 0.0%	0 0.0%	121 100%	スル動詞
4	47 33.6%	21 15.0%	63 45.0%	6 4.3%	2 1.4%	1 0.7%	140 100%	多機能(副詞的)
5	14 21.5%	12 18.5%	37 56.9%	0 0.0%	1 1.5%	1 1.5%	65 100%	多機能(形動的)
6	2 40.0%	1 20.0%	2 40.0%	0 0.0%	0 0.0%	0 0.0%	5 100%	ニ付加型副詞
総計	436 34.9%	351 28.1%	425 34.1%	18 1.4%	10 0.8%	8 0.6%	1248 100%	

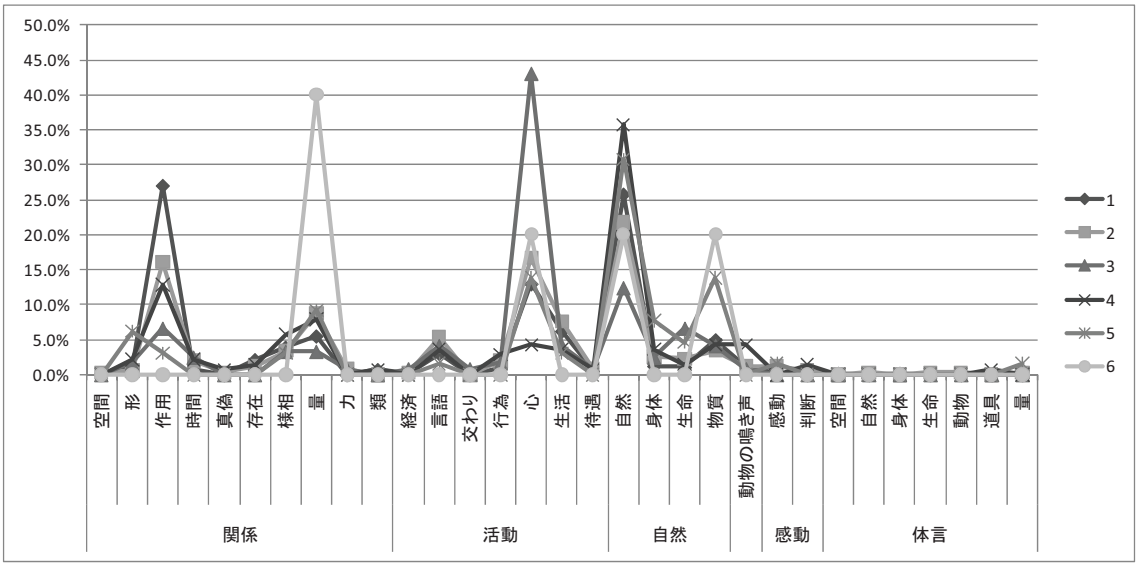


図 3　『分類語彙表』中項目の各グループ内の比率

が多様なタイプの所属する語群である。ここでは「関係」「活動」「自然」の比率に大きな差はない。例えば上位 10 語内の「さっぱり」が５.１に例として挙げたように多様な意味を持っている点が特徴的である。また［グループ４］は形容動詞的なものの他に副詞（単独）等を多く含み，名詞的な用法も含む多機能な語群である。ここでは「自然」と「関係」の比率がともに高い。例として上位 10 語内の「がたがた」を見ると，「関係-様相-調和・混乱」「活動-言語-言語活動」「自然-自然-音」「自然-生命-生理・病気」という４つの項目に挙がっており，［グループ２］と同様多義の語が頻度上位語となっていることが分かる。このように品詞的に多様なタイプが所属するグループは，意味的にも多様な様相を示している。この傾向は本研究でオノマトペを形態のみでまとめあげて分析を行なったことにも起因すると思われる。意味の異なるオノマトペを別語として扱った上で品詞性を検討することで，また異なる傾向が得られる可能性がある。

形容動詞的な多機能型の［グループ５］は「自然」がもっとも高い比率となっており，同じ多機能型ながら［グループ４］と比べて意味的に偏った傾向が得られる。上位 10 語のうち中項目の「自然」の語としては「ぴかぴか（光)」「どろどろ（材質)」等が当たる。「物

質」は「からから（乾湿)」「ぐちゃぐちゃ（乾湿)」等である。他「関係」の比率も比較的高いが，これには他のグループより「形」の語（「くしゃくしゃ」「くたくた」等）が多いことが指摘される。他，「ぎりぎり（関係-量)」「くたくた（活動-心)」などがある。「物質」や「形」，「材質」などの語が比較的多く含まれていることから，この語群には具体物の様態に対して用いる語が多く所属していると言えるのではないだろうか。ト付加型副詞などが「動的」な修飾を担うのに対して，「静的」な語群と言えそうである。［グループ６］は頻度が低い語群だが，個々の語を確認すると，「けちょんけちょん」「こてんぱん」が「関係-量」，「べろんべろん」が「活動-心」，「こちんこちん」が「自然-材質」，「びしょびしょ」が「自然-物質」で，［グループ５］の特徴と類似する。２グループの傾向の合致から，ニ付加型の特徴の「静的」は広く形容動詞的な語群の特徴として関連付けられると言える。

６　おわりに

以上，後接要素に基づいて分類したオノマトペの各グループを『分類語彙表』と対応させたことで主に明らかになったことは，次のようにまとめられる。

○スル動詞や単独で副詞として働く語群は日常的に使われる固定的で一般的な語形である。対して，ト付加型副詞の語群には臨時的な語形が多く含まれると言える。

○「動的」と「静的」の違いはト付加型副詞の語群と形容動詞的語群にそれぞれ対応する。

○多様な品詞を含む語群には，意味的にも多様性が認められる。純粋に形態的な指標のみで分類したことによって発生した現象でもあり，より詳細な分析を期すべきである。

今回の手法では『分類語彙表』の見出し語にない調査語は扱えなかったが，そうした語形の詳細も実例に基づき解明する必要がある。また，より詳細に意味を記述するためにはやはりオノマトペの被修飾成分を考慮すべきである。今後の課題としたい。

付記　本研究は，文部科学省科学研究費特定領域研究「日本語コーパス」による補助を得たものである。

参考文献

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

加藤久雄，坂口昌子(1996)「後接成分とオノマトペの性質について」『奈良教育大学紀要　人文・社会科学』，45:1，pp.1-11

国立国語研究所(2004)『分類語彙表　増補改訂版』大日本図書

佐々木文彦(1986)「擬態語類の語尾について」『松村明教授古希記念 国語研究論集』明治書院，pp.723-736

田守育啓(1993)「日本語オノマトペの統語範疇」『オノマトピア：擬音・擬態語の楽園』筧寿雄・田守育啓編　勁草書房，pp.17-75

星野和子(2005)「擬態語の文法」『駒沢女子大学研究紀要』，12，pp.185-198

宮内佐夜香，小木曽智信，小磯花絵，小椋秀樹(2011)「『現代日本語書き言葉均衡コーパス』に基づくオノマトペの分析―品詞性の検討を中心に―」『言語処理学会第 17 回年次大会発表論文集』

宮地裕(1978)「擬音語・擬態語の形態論小考」『国語学』，115，pp.33-39

楊淑雲(2008)「擬態語の語尾とその後接成分について」『国語学研究』，47，pp.44-53

Web 版コーパス検索アプリケーション「中納言」のデモンストレーション

中村　壮範（データ班協力者　　：マンパワージャパン株式会社）†
小木曽智信（電子化辞書班分担者：国立国語研究所言語資源研究系）

Demonstration of Online Concordancer 'Chunagon'

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

1.はじめに

『現代日本語書き言葉均衡コーパス』（以下、BCCWJ とする。）は 2011 年中に一般公開を開始する予定であるが、公開形式の一つに Web オンラインサービスが予定されている。

これまでに公開されているBCCWJの検索ツールは「BCCWJ検索デモンストレーションサイト」や全文検索システム「ひまわり」など、表層の文字列を対象としたものであった。しかし、BCCWJには電子化辞書班・データ班が連携して開発を進めている形態素解析辞書UniDic[1]により形態論情報が付与されることになっているため、その情報を利用した検索ができるようになれば、表層の文字列にとらわれず、見出し語や品詞などを基に用例を収集することが可能になり、コーパスを利用する上で有益である。そこで、品詞などの短単位情報を検索条件に指定して検索を行うことができるWeb検索アプリケーション「中納言」を開発した。「中納言」はBCCWJの人手修正済みコーパスの作成などに用いているコーパス修正ツール「大納言」を基にし、検索機能に特化して、インターフェイスをWeb用に改めたものである（小木曽・中村 2011）。

2.中納言の特徴

中納言の画面を図 1 に示す。中納言の主な特徴は以下の通りである。

1) Web アプリケーションであるため、インターネットが利用できる環境と標準的なブラウザがあれば、特別なソフトをインストールすることなく利用することができる。
2) 「短単位検索」「文字列検索」の 2 種類の検索方法を提供している。「短単位検索」とは BCCWJ に付与された短単位情報について条件を指定して検索を行う機能、「文字列検索」とは検索条件に文字列や正規表現を使用してテキストデータの検索を行う機能である。
3) 「短単位検索」では共起条件を指定することができる。
4) 検索結果として、文脈、品詞などの短単位情報のほか、サンプルのタイトルや著者などの情報を表示することができる。
5) 検索結果は、タブ区切りテキスト形式でダウンロードすることができる。

† tnakamura@ninjal.ac.jp
[1] UniDic については伝ほか(2007)を、短単位については小椋・小磯ほか(2011)を参照。

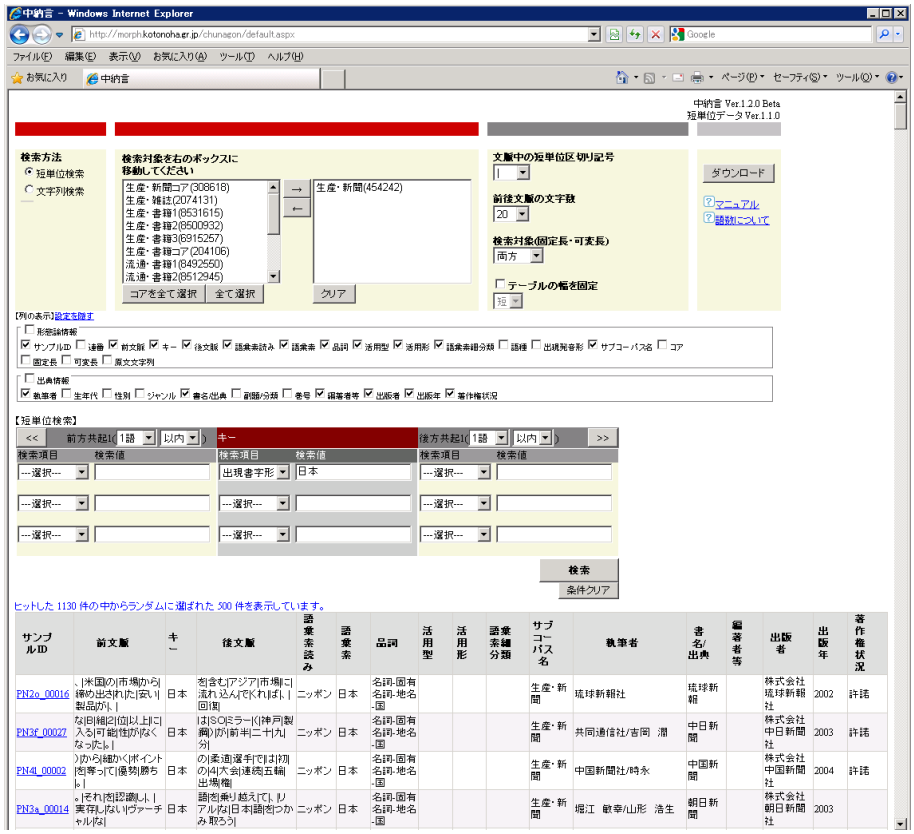


図 1　中納言の画面（短単位検索）

3.中納言の検索機能

2 節で述べた「短単位検索」と「文字列検索」について、以下で詳細を解説する。

3.1.検索時の指定項目

中納言の画面上部に表示される操作画面（図 2）では「検索方法（短単位検索・文字列検索）」「検索対象サブコーパス」「文脈の文字数」「文脈内の短単位区切り記号」「検索対象（固定長・可変長）」などが指定できる（BCCWJ の設計の詳細は山崎（2007）参照）。

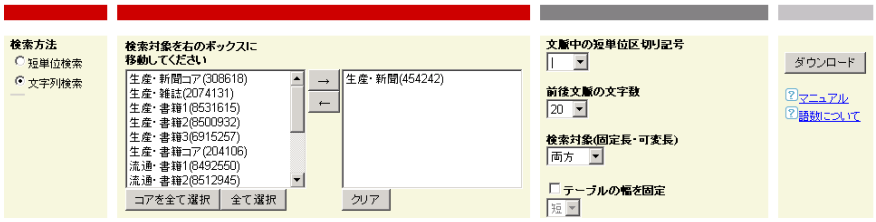


図 2　検索時の指定項目

3.2.短単位検索

BCCWJ のデータには形態素解析辞書 UniDic による形態論情報が付与されている。UniDic では、表記が異なっても同じ語であれば、一つの見出し語にまとめるという方針を取り、語を階層化した形で辞書登録している。この階層の最上位を語彙素と呼んでおり、この語彙素の下に語形、更に語形の下に書字形という階層が設けられている。

短単位検索では、この情報を生かした柔軟な検索条件指定が可能になっている（図 3）。

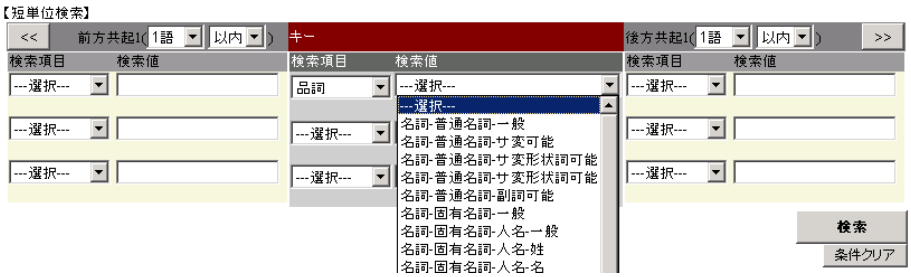


図 3　短単位検索

A 検索項目指定

検索項目はドロップダウンにより選択することができる。選択肢には「出現書字形」「品詞」「語彙素」「語彙素読み」「活用形」「活用型」がある。

B 検索値指定

検索項目に「出現書字形」「語彙素」「語彙素読み」を指定した場合には検索値をテキストボックスに直接入力し、「品詞」「活用型」「活用形」を指定した場合には、検索値を指定するテキストボックスがドロップダウンリストに変化するため、そこから選択する。選択肢が表示されるので、ユーザーが UniDic の品詞体系を完全に把握している必要はない。

C 共起範囲指定

キーとなる短単位の前方・後方それぞれ 1～5 語まで、またはキーとなる短単位を含む文の文頭から文末までを共起範囲として指定して検索することができる。共起語についても、上記 1)、2)に示した検索条件を指定できる。

UniDic による形態論情報を用いることができるため、図 3 の「短単位検索」の検索項目指定で「語彙素」または「語彙素読み」を指定することによって、検索語の異語形や異表記形を網羅的に検索することができる。例えば、検索条件で検索項目を「語彙素」、検索値を「矢張り」と指定することで、「やはり」「やっぱり」「やっぱ」「やっぱし」「矢張り」など、「矢張り」という語彙素見出しを持つ全ての語形、及びその語形見出しを持つ全ての書字形を検索することが可能である。

3.3.文字列検索

中納言のもうひとつの検索方法に「文字列検索」がある。文字列検索では検索したい文字列を指定することで短単位の境界を意識せずに文字列を全文検索することができる。したがって、短単位の区切りが分からない場合に、まずは文字列検索によって短単位の区切りを調べ、次に行う短単位検索での語の検索条件指定を行いやすくする、といった短単位検索の補助的な使い方をすることができる。

4.検索条件の保存と再利用

検索条件の保存や検索結果の再現を容易にするために、検索条件の指定方法を記述する簡易言語（X-CQL）を規定して、この形式による検索条件のエクスポート・インポートを可能にするよう準備を行っている。記述には XML 形式を用いている。記述例として「助動

詞「らしい」(接尾辞ではない)が名詞を連体修飾する用例」を抽出するための X-CQL を以下に示す（記述方法は検討中のものであり今後変更される可能性がある）。

```
<x-cql application="中納言" version="1.0.1">
<corpus selected="PB OB LB"/>
  <condition0 品詞="助動詞%" 語彙素読み="ラシイ" 活用形="連体形" />
  <condition1 品詞="名詞%" />
</x-cql>
```

X-CQL による記述により中納言の画面上で行える条件指定をすべて記述することができる。中納言の画面上で検索条件を入力すると、X-CQL が画面上に表示されるため、ユーザーはこれをテキストファイルにコピー&ペーストすることで検索条件の保存を行うことができる。したがって、ユーザーが直接 X-CQL を記述する必要はないが、テキストエディタなどを使用してユーザーが独自に記述することもできる。いったん保存した条件は、中納言の短単位検索モードの画面上で取り込んで検索条件に反映させることができる。

5.おわりに

以上、中納言の詳細について述べた。中納言は 2009 年 9 月下旬から特定領域研究「日本語コーパス」のメンバーに対して公開している。現時点での検索対象となるデータは「BCCWJ 領域内公開データ（2009 年度版）DVD」の XML データ約 8000 万語である。1 億語以上の本格的な公開は 2011 年 7 月頃を予定している。

参考文献

小木曽智信・中村壮範（2011）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版』（LR-CCG-10-06）

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

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007)「コーパス日本語学のための言語資源－形態素解析用電子化辞書の開発とその応用－」『日本語科学』、 vol.22、 pp.101-123.

山崎誠（2007）「『現代日本語書き言葉均衡コーパス』の基本設計について」『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』、pp.127-136.

小木曽智信・中村壮範（2010）「「現代日本語書き言葉均衡コーパス」のための形態論情報データベースについて」『第 16 回公開シンポジウム「人文科学とデータベース」論文集』、pp.45-52.

関連 URL

KOTONOHA 検索デモンストレーションサイト　http://www.kotonoha.gr.jp/demo/

全文検索システム『ひまわり』（国立国語研究所「言語データベースとソフトウェア」）
http://www2.ninjal.ac.jp/lrc

階層的形態論情報を考慮した『現代日本語書き言葉均衡コーパス』の公開用 XML フォーマット

小木曽智信（電子化辞書班分担者：国立国語研究所言語資源研究系）†
間淵　洋子（データ班分担者：国立国語研究所コーパス開発センター）
前川喜久雄（総括班班長：国立国語研究所言語資源研究系）

A New XML format of the BCCWJ that Enables Hierarchical Representation of Morphological Information

Toshinobu Ogiso　(National Institute for Japanese Language and Linguistics)
Yoko Mabuchi　(National Institute for Japanese Language and Linguistics)
Kikuo Maekawa　(National Institute for Japanese Language and Linguistics)

１．はじめに

これまで、BCCWJ の XML フォーマットとしては、テキストに文書構造を単純にマークアップした形式を提案し公開を行ってきた。この形式は文字列に依拠した利用に対しては十分な対応が可能であったが、BCCWJ の形態論情報を埋め込んで利用するためには不十分な点があった。本発表では、この文字列ベースの XML フォーマットをもとにして、言語構造を一定程度反映させた新しい XML フォーマットを提案する。これにより、短単位・長単位をはじめとする形態論情報や、文を単位とする情報などの言語構造に関わる情報を付与することを可能にする。

２．文字ベースの XML とその問題点

2.1.　文字ベースの XML のタグセット

BCCWJ では、ランダムサンプリングによって採集したサンプルから、長さを 1000 字に固定した固定長サンプルと、節や章など文章の意味上のまとまりをとりだした可変長サンプルの 2 種類を作成している。固定長と可変長のサンプルは別個に取得するのではなく、同一のサンプリングポイントから、2 通りの方法によって重複部分を持つ形で作成している。

各々のサンプルは、XML 形式で表 1 に示すタグを用いてマークアップを施される（山口ほか 2008）。マークアップにあたっては単語等の切れ目は意識していない。

なお、文を示す sentence タグは、入れ子構造を許しており、大きな文の中に複数の文が含み込まれることがある。

2.2.　BCCWJ の形態論情報

一方、BCCWJ では、すべてのサンプルに対して形態論情報の付与が行われる。形態素解析辞書 UniDic の解析結果に基づく短単位と、これを組み上げた長単位の二つの単位による情報が付与される（伝ほか 2007，小椋ほか 2010）。

短単位は、単位の認定、品詞や見出しの付与方法などについて詳細な規定を定めた言語単位である。和語の場合は単純語または単純語 2 語の結合を 1 短単位とし、漢語の場合は二字漢語までを 1 単位とするものである。助詞等の付属語や記号も 1 単位となる。次の例文の「/」が短単位境界である。

† togiso@ninjal.ac.jp

/国立/国語/研究/所/で/研究/し/て/いる

一方、長単位は、この短単位を組み上げたもので、文節から付属語を取り去ったものが長単位に相当する。付属語は原則として短単位単独で長単位となるが、複合辞として認定した「ている」などは1長単位となる。また、「研究し」は漢語サ変動詞として1長単位にまとめられる。次の例文の「/」が長単位境界である。

/国立国語研究所/で/研究し/ている/

したがって、短単位・長単位・文節は入れ子の構造を取る。文節はこれが連なって文を構成するし、短単位は文字から構成されるから、BCCWJ の形態論情報は、結局次のような言語単位の階層構造の中に位置づけられることになる。

文章／文／文節／長単位／短単位／文字

XSLT などを用いて形態論情報を活用するためには、この階層構造・包含関係がそのまま XML フォーマットに反映されることが望ましい。

表 1　文字ベースの XML フォーマットの主なタグ

種類	タグ	説明
サンプル	sample	サンプリングによって 1 サンプルとされた文章の範囲
	sampling	サンプリングポイントに関する情報
階層構造 (文書構造)	article	同一著者による、同一テーマのひとまとまりの文章
	title	ある範囲の文章の内容を代表する記述。章の題、新聞の見出しなど
	cluster	title 要素がまとめる文章の範囲
	list	箇条書きや名詞句の羅列など、列挙された要素
	paragraph	段落に相当する文の集まり
	sentence	文に相当する語の集まり
図　表 (文書構造)	figure	図・表・写真・絵など
	caption	図表等についてのタイトルや説明
引　用 (文書構造)	citation	当該 article 要素とは異なる著作物からの引用
	speech	発話や心内発話の引用・書き起こし
	quote	行内における引用・発話表現
注　記 (文書構造)	noteBody	脚注、後注など、本文と区別して記述される注記
その他 (文書構造)	abstract	article 要素、または cluster 要素の概要に相当する要素
	verse	詩、和歌、俳句、歌謡などの韻文
文字・表記	ruby	ルビ付き文字
	correction	原文の誤植を訂正した文字
	missingCharacter	規定の文字集合に含まれない文字 (JIS 外字)

2.3. 文字ベースの XML と形態論情報の齟齬

2.1.で示した文字ベースの XML フォーマットは、2.2.で示した言語単位の階層構造ときれいに対応しない場合がある。ruby タグはその典型的な例である。

ルビ（ふりがな）は、次の 1)～5)のように単語中の一部分の文字に対してつけられる場合から、一文に対して一つのルビが対応するようなものまで様々なものが存在する。BCCWJ の ruby タグは原則として単漢字に対するルビとして付与されているが、熟字訓などでは複数の文字にまたがることになる。

1) 語彙（い） （短単位よりも短いルビ）
2) 時雨（しぐれ） （短単位と一致するルビ）
3) 喜望峰（ケープタウン） （短単位よりも長いルビ）
4) 新しい芸術（アール・ヌーヴォー） （長単位よりも長いルビ）
5) 達者でな（アスタ・ラ・ビスタ） （文全体にかかるルビ）

文字ベースの XML では上記のような例は単純に範囲内の文字列を ruby タグで囲み、ルビ文字を rubyText 属性の値としてきた。これらが短単位の形態論情報タグ（SUW）とともにマークアップされるとき、例 1)のように短単位よりも短い ruby は SUW の子要素とならざるを得ない。一方、例 3)～5)では、逆に ruby は SUW の親要素となるほかない。

1a) <SUW>語<ruby rubyText="い">彙</ruby></SUW>
2a) <SUW><ruby rubyText="しぐれ">時雨</ruby></SUW> （または）
<ruby rubyText="しぐれ"><SUW>時雨</SUW></ruby>
3a) <ruby rubyText="ケープタウン"><SUW>喜望</SUW><SUW>峰</SUW></ruby>
4a) <ruby rubyText="アール・ヌーヴォー"><SUW>新しい</SUW><SUW>芸術</SUW>
</ruby>
5a) <ruby rubyText="アスタ・ラ・ビスタ"><SUW>達者</SUW><SUW>で</SUW><SUW>
な</SUW></ruby>

ここでは省略するが、長単位タグ（LUW）を考えるときには、関係はさらに複雑なものとなる。そして、このままでは形態論情報との上下関係が定まらず、利用上不便を来すこととなる。

このほかに引用タグ（quote）も短単位と齟齬を来す場合がある。引用文では、ときに用言の活用語尾の一部分だけが引用され、残りが地の文で補われる場合がある。

6) <quote>「解剖後厚く弔」</quote>うべしという指示

このとき、短単位「弔う」は quote の終了タグを越えることになる。ただし、これは単にタグだけの問題ではない。引用符（“」”）が短単位内に入り込んでいるため、この文字までが問題となる。

2.4. 文認定の問題

文（sentence）の認定をめぐっては、sentence タグの入れ子が認められているという問題がある。たとえば、次のように文中に引用がある場合には、全体を sentence で囲みつつ、引用部分も sentence でマークアップされている。

7) <sentence>驚きながらそう誤魔化した構治の言葉に、<quote>「<sentence>落ちた、
落ちたって言わないでよ。</sentence><sentence type="quasi">結構辛がってる

んだから</sentence>」</quote>言って夕美子は目を伏せ、(中略) うつむいている。
</sentence>

複雑な構造をとる文の場合、そのいずれもが文として認められるという点で、このマークアップにも積極的な意味がある。

しかし、(1) 上位の sentence がきわめて長くなる場合がある (2) 形態素解析などの解析ツールの入力となる「文」を定めがたい (3) データを文番号で管理できない、などのデメリットがある。

文についてはまた、sentence タグが付与されない環境が生じているという問題がある。verseLine は詩歌の行を示すタグであるが、これが用いられる場合には、文（sentence）の認定を行わず、原文の改行位置を基準にそのまま verseLine としてきた。そのため、verseLine は sentence を親に持たない特殊な要素となっており、また、原文の状況によっては形態論情報の単位と齟齬を来す可能性がある。

8) <verseLine>霊山の</verseLine>
<verseLine>誓いも深き</verseLine>
<verseLine>君ら西</verseLine>
<verseLine>我ら東と</verseLine>
<verseLine>白馬も雄々しく</verseLine>

2.5. 固定長と可変長の問題

2.1.で触れたとおり、文字ベースの XML では、固定長と可変長を別の XML ファイルとして扱っていた。文字列を対象とした調査を行う場合には別ファイルとなっていることが望ましい場合も多いが、データに対して新たな情報を付与する場合には問題となる。たとえば、自動で付与された形態論情報に対して人手で修正を施す場合には、重複部分について二度手間が生じるほか、同一箇所に異なる形態論情報が付与される可能性が生じる。

したがって、特に形態論情報を付与する場合には、固定長と可変長を統合した形式をソースとし、そこから固定長・可変長の二つの情報が取得できるようにすることが望ましい。

３．形態論情報付き XML フォーマット

3.1. 基本方針

以上のような問題点を踏まえ、新しい形態論情報付きの XML フォーマットは、これまでの XML との互換性をできる限り確保しつつ，言語構造と齟齬を来す要素について修正を行うこととした。さらに、新フォーマットから旧フォーマットへは自動変換できるように設計している。

3.2 階層構造

2.2.で示した形態論情報の階層構造に、表 1 のタグを納めるならば、図 1（次ページ）のような階層が考えられる（網掛けはすべてのテキストに必須の要素）。

この構造に照らせば、文字ベースの XML フォーマットの問題は、文（sentence）タグの階層が一律に付与されていないことが第一の問題である。そして、ruby が上の階層に飛び出したり、quote が下の階層を侵犯したりすることで、形態論情報と齟齬を来しタグの交叉を招くのが第二の問題だということになる。

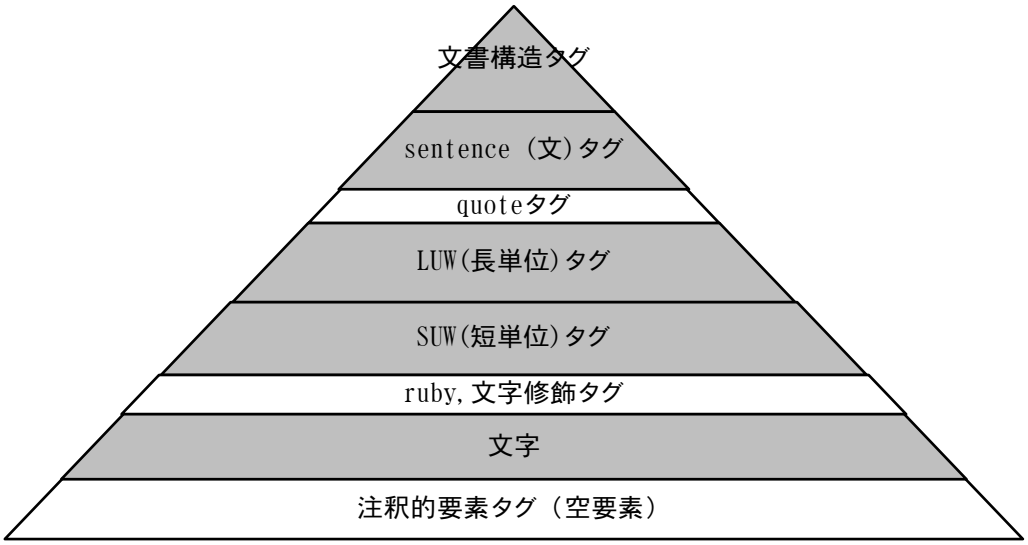


図 1 形態論情報付き XML フォーマットの階層構造

3.3. 変更点

これらの問題点を解消するため、新しい XML フォーマットでは固定長・可変長を統合した XML をベースとして、各タグについて次のような対処を行った。

A. 文（sentence）

文タグの階層を整備するために、sentence の入れ子を認めることをやめ、上位の文は superSentence として文書構造タグの一種とした。下位の sentence はそのまま残し、superSentence の一部分を新たに sentence で囲み type= " fragment "とした。

```
7') <superSentence>
  <sentence type="fragment">驚きながらそう誤魔化した構治の言葉に、
  </sentence>
  <quote><sentence>「落ちた、落ちたって言わないでよ。</sentence>
  <sentence type="quasi">結構辛がってるんだから」</sentence></quote>
  <sentence type="fragment">言って夕美子は目を伏せ、（中略）うつむいている。
  </sentence></superSentence>
```

また、verseLine については改行位置を空要素の verseLine タグとして残しつつ、文相当の範囲を新たに sentence で囲み、type= " verse "とした。

```
8') <sentence type="verse">霊山の<verseLine/>誓いも深き<verseLine/>君ら
  西<verseLine/>我ら東と<verseLine/>白馬も雄々しく<verseLine/>
  </sentence>
```

これにより、すべての短単位はいずれかの sentence に属することとなり、サンプルは sentence の集合としても捉えられることとなった。

B. ルビ（ruby）

短単位を越えるルビについては、先頭の短単位を ruby タグで囲み、そのタグの属性値として本来のルビ範囲のテキストを保持することとした。これにより、元の状態に戻すことを可能にすると同時に、複数単位に渡る特殊なルビを容易に取り出すことを可能にしている。

3a') <SUW><ruby rubyText="ケープタウン" rubyBase="喜望峰">喜望</ruby>
</SUW><SUW>峰</SUW>

4a') <SUW><ruby rubyText="アール・ヌーヴォー" rubyBase="新しい芸術">新しい
</ruby></SUW><SUW>芸術</SUW>

C. 引用（quote）

短単位を分断する引用については、引用符のテキストを移動し、元の場所に空要素タグを残すことで対処した。

6') <quote>「解剖後厚く弔<move type="original" text="」"/>う<move type="modify">」</move></quote>べしという指示

これにより短単位 SUW で引用符（“」”）を含まない「弔う」を囲むことが可能になると同時に、quote と SUW の交叉も解消される。

D. 注釈的要素タグの空要素化

これらのタグ以外に、元のタグセットの仕様では、本来ならば本文テキストとして扱うべきでない文字列がそのまま残されている場合があった。たとえば注釈タグ（noteBody）関連のタグがその一つである。

9) 国際ルールに反しない形でタイド化を行っている<noteMarker>（注１）</noteMarker>

これについては次のような空要素タグに仕様を変更することで問題を解消している。

9') 国際ルールに反しない形でタイド化を行っている<noteMarker text="（注１）"/>

このような空要素化処理をする場合、属性値に入れられるテキスト部分にタグが用いられていることがある。

10) <noteMarker><enclosedCharacter description="○">６６
</enclosedCharacter><noteMarker>

これはタグ表記が必要な丸付き数字を含むテキストだが、この場合には次のような記法によって info 属性に元の情報を保持できるようにした。

10') <noteMarker text="６６" info="enclosedCharacter:description=○"/>

3.4. 新形式のサンプル

以上のような変更を加え、階層化された形態論情報を付与した XML のサンプルを【付録】として掲げる（次ページ）。

3.5. 文字ベースの XML との互換性

3.3.で示した変更点は、原則として元の情報を保持したまま、形態論情報との併存を図ったものである。したがって、この形態論情報付き XML フォーマットから、文字ベースの XML フォーマットに変換することが可能である。可変長・固定長の文字ベースの XML フォーマットは、今後も引き続き提供される予定である。

４．おわりに

以上、BCCWJ の新しい形態論情報付き XML フォーマットについて述べた。XML でマークアップされた文書構造と、長短二つの形態論情報とを同時に有効利用可能なこの形式を利用することにより、コーパスがさらに活用されることを期待する。

文献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22 pp.101-123

前川 喜久雄（2008）「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」日本語の研究 4-1

小椋秀樹・小磯花絵・冨士池優美・宮内左夜香・小西光・原裕（2011）国立国語研究所内部報告書「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）」 LR-CCG-10-05-01/02

山口昌也・高田智和・北村雅則・間淵洋子・小林正行・西部みちる（2008）「『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.2.0」

【付録】新 XML フォーマットサンプル

（PB23_00027 の一部，形態論情報の ID は省略した。）

```
<sentence>
<LUW l_orthToken="日本" l_lForm="ニッポン" l_lemma="日本" l_pos="名詞-固有名詞-地名-国" l_origText="日本"><SUW orthToken="日本" lForm="ニッポン" lemma="日本" pos="名詞-固有名詞-地名-国" form="ニッポン" pronToken="ニッポン" wType="固" BOS="True">日本</SUW></LUW>
<LUW l_orthToken="に" l_lForm="ニ" l_lemma="に" l_pos="助詞-格助詞" l_origText="に"><SUW orthToken="に" lForm="ニ" lemma="に" pos="助詞-格助詞" form="ニ" pronToken="ニ" wType="和">に</SUW></LUW>
<LUW l_orthToken="あっ" l_lForm="アル" l_lemma="有る" l_pos="動詞-一般" l_cType="五段-ラ行" l_cForm="連用形-促音便" l_origText="あっ"><SUW orthToken="あっ" lForm="アル" lemma="有る" pos="動詞-非自立可能" form="アル" cType="五段-ラ行" cForm="連用形-促音便" pronToken="アッ" wType="和">あっ</SUW></LUW>
<LUW l_orthToken="た" l_lForm="タ" l_lemma="た" l_pos="助動詞" l_cType="助動詞-タ" l_cForm="連体形-一般" l_origText="た"><SUW orthToken="た" lForm="タ" lemma="た" pos="助動詞" form="タ" cType="助動詞-タ" cForm="連体形-一般" pronToken="タ" wType="和">た</SUW></LUW>
<LUW l_orthToken="の" l_lForm="ノ" l_lemma="の" l_pos="助詞-準体助詞" l_origText="の"><SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-準体助詞" form="ノ" pronToken="ノ" wType="和">の</SUW></LUW>
<LUW l_orthToken="は" l_lForm="ハ" l_lemma="は" l_pos="助詞-係助詞" l_origText="は"><SUW orthToken="は" lForm="ハ" lemma="は" pos="助詞-係助詞" form="ハ" pronToken="ワ" wType="和">は</SUW></LUW>
<LUW l_orthToken="、" l_lForm="" l_lemma="、" l_pos="補助記号-読点" l_origText="、"><SUW orthToken="、" lForm="" lemma="、" pos="補助記号-読点" form="" pronToken="" wType="記号">、</SUW></LUW>
<LUW l_orthToken="むしろ" l_lForm="ムシロ" l_lemma="寧ろ" l_pos="副詞" l_origText="むしろ"><SUW orthToken="むしろ" lForm="ムシロ" lemma="寧ろ" pos="副詞" form="ムシロ" pronToken="ムシロ" wType="和">むしろ</SUW></LUW>
<LUW l_orthToken="、" l_lForm="" l_lemma="、" l_pos="補助記号-読点" l_origText="、"><SUW orthToken="、" lForm="" lemma="、" pos="補助記号-読点" form="" pronToken="" wType="記号">、</SUW></LUW>
<quote>
<LUW l_orthToken="「" l_lForm="" l_lemma="「" l_pos="補助記号-括弧開" l_origText="「"><SUW orthToken="「" lForm="" lemma="「" pos="補助記号-括弧開" form="" pronToken="" wType="記号">「</SUW></LUW>
<LUW l_orthToken="葦牙" l_lForm="アシカビ" l_lemma="葦牙" l_pos="名詞-普通名詞-一般" l_origText="葦牙"><SUW orthToken="葦牙" lForm="アシカビ" lemma="葦牙" pos="名詞-普通名詞-一般" form="アシカビ" pronToken="アシカビ" wType="和"><ruby rubyText="あし">葦</ruby><ruby rubyText="かび">牙</ruby></SUW></LUW>
<LUW l_orthToken="の" l_lForm="ノ" l_lemma="の" l_pos="助詞-格助詞" l_origText="の"><SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" form="ノ" pronToken="ノ" wType="和">の</SUW></LUW>
<LUW l_orthToken="ごとく" l_lForm="ゴトシ" l_lemma="ごとし" l_pos="助動詞" l_cType="文語助動詞-ゴトシ" l_cForm="連用形-一般" l_origText="ごとく"><SUW orthToken="ごとく" lForm="ゴトシ" lemma="ごとし" pos="助動詞" form="ゴトシ" cType="文語助動詞-ゴトシ" cForm="連用形-一般" pronToken="ゴトク" wType="和">ごとく</SUW></LUW>
<LUW l_orthToken="萌え出る" l_lForm="モエイズ" l_lemma="萌え出ず" l_pos="動詞-一般" l_cType="文語下二段-ダ行" l_cForm="連体形-一般" l_origText="萌え出る"><SUW orthToken="萌え出る" lForm="モエイデル"
```

lemma="萌え出でる" pos="動詞-一般" form="モエイズ" cType="文語下二段-ダ行" cForm="連体形-一般" pronToken="モエイズル" wType="和">萌え出る</SUW></LUW>
<LUW l_orthToken="」" l_lForm="" l_lemma="」" l_pos="補助記号-括弧閉" l_origText="」"><SUW orthToken="」" lForm="" lemma="」" pos="補助記号-括弧閉" form="" pronToken="" wType="記号">」</SUW></LUW>
</quote>
<LUW l_orthToken="と" l_lForm="ト" l_lemma="と" l_pos="助詞-格助詞" l_origText="と"><SUW orthToken="と" lForm="ト" lemma="と" pos="助詞-格助詞" form="ト" pronToken="ト" wType="和">と</SUW></LUW>
<LUW l_orthToken="いう" l_lForm="イウ" l_lemma="言う" l_pos="動詞-一般" l_cType="五段-ワア行-イウ" l_cForm="連体形-一般" l_origText="いう"><SUW orthToken="いう" lForm="イウ" lemma="言う" pos="動詞-一般" form="イウ" cType="五段-ワア行" cForm="連体形-一般" pronToken="イウ" wType="和">いう</SUW></LUW>
<LUW l_orthToken="、" l_lForm="" l_lemma="、" l_pos="補助記号-読点" l_origText="、"><SUW orthToken="、" lForm="" lemma="、" pos="補助記号-読点" form="" pronToken="" wType="記号">、</SUW></LUW>
<LUW l_orthToken="いま" l_lForm="イマ" l_lemma="今" l_pos="名詞-普通名詞-一般" l_origText="いま"><SUW orthToken="いま" lForm="イマ" lemma="今" pos="名詞-普通名詞-副詞可能" form="イマ" pronToken="イマ" wType="和">いま</SUW></LUW>
<LUW l_orthToken="の" l_lForm="ノ" l_lemma="の" l_pos="助詞-格助詞" l_origText="の"><SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" form="ノ" pronToken="ノ" wType="和">の</SUW></LUW>
<LUW l_orthToken="植物繁茂" l_lForm="ショクブツハンモ" l_lemma="植物繁茂" l_pos="名詞-普通名詞-一般" l_origText="植物繁茂"><SUW orthToken="植物" lForm="ショクブツ" lemma="植物" pos="名詞-普通名詞-一般" form="ショクブツ" pronToken="ショクブツ" wType="漢">**<ruby rubyText="プロリフェレーション" rubyBase="植物繁茂">植物</ruby>**</SUW><SUW orthToken="繁茂" lForm="ハンモ" lemma="繁茂" pos="名詞-普通名詞-サ変可能" form="ハンモ" pronToken="ハンモ" wType="漢">繁茂</SUW></LUW>
<LUW l_orthToken="の" l_lForm="ノ" l_lemma="の" l_pos="助詞-格助詞" l_origText="の"><SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" form="ノ" pronToken="ノ" wType="和">の</SUW></LUW>
<LUW l_orthToken="モデル" l_lForm="モデル" l_lemma="モデル" l_pos="名詞-普通名詞-一般" l_origText="モデル"><SUW orthToken="モデル" lForm="モデル" lemma="モデル" subLemma="model" pos="名詞-普通名詞-一般" form="モデル" pronToken="モデル" wType="外">モデル</SUW></LUW>
<LUW l_orthToken="に" l_lForm="ニ" l_lemma="に" l_pos="助詞-格助詞" l_origText="に"><SUW orthToken="に" lForm="ニ" lemma="に" pos="助詞-格助詞" form="ニ" pronToken="ニ" wType="和">に</SUW></LUW>
<LUW l_orthToken="近い" l_lForm="チカイ" l_lemma="近い" l_pos="形容詞-一般" l_cType="形容詞" l_cForm="連体形-一般" l_origText="近い"><SUW orthToken="近い" lForm="チカイ" lemma="近い" pos="形容詞-一般" form="チカイ" cType="形容詞" cForm="連体形-一般" pronToken="チカイ" wType="和">近い</SUW></LUW>
<LUW l_orthToken="考え方" l_lForm="カンガエカタ" l_lemma="考え方" l_pos="名詞-普通名詞-一般" l_origText="考え方"><SUW orthToken="考え" lForm="カンガエル" lemma="考える" pos="動詞-一般" form="カンガエル" cType="下一段-ア行" cForm="連用形-一般" pronToken="カンガエ" wType="和">考え</SUW><SUW orthToken="方" lForm="カタ" lemma="方" pos="接尾辞-名詞的-一般" form="カタ" pronToken="カタ" wType="和">方</SUW></LUW>
<LUW l_orthToken="。" l_lForm="" l_lemma="。" l_pos="補助記号-句点" l_origText="。"><SUW orthToken="。" lForm="" lemma="。" pos="補助記号-句点" form="" pronToken="" wType="記号">。</SUW></LUW>
</sentence><br type="automatic_original"/>
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<speaker>
<sentence type="quasi">
<LUW l_orthToken="大江" l_lForm="オオエ" l_lemma="大江" l_pos="名詞-固有名詞-人名-姓" l_origText="大江"><SUW orthToken="大江" lForm="オオエ" lemma="オオエ" pos="名詞-固有名詞-人名-姓" form="オオエ" pronToken="オーエ" wType="固" BOS="True">大江</SUW></LUW>
</sentence><br type="automatic_original"/>
</speaker>

汎用アノテーションツール Slate

徳永健伸　(ツール班分担者：東京工業大学大学院理工学研究科)[1]
Dain Kaplan　(ツール班協力者：東京工業大学大学院理工学研究科)
飯田龍　(ツール班協力者：東京工業大学大学院理工学研究科)

Multi-purpose Annotation Tool Slate

Tokunaga Takenobu　(Department of Computer Science, Tokyo Institute of Technology)
Dain Kaplan　(Department of Computer Science, Tokyo Institute of Technology)
Iida Ryu　(Department of Computer Science, Tokyo Institute of Technology)

1 はじめに

これまでに様々なプロジェクトを通して，様々なコーパスが作成されており，効率よく，信頼性の高いコーパスを作るために専用のコーパス作成ツールも数多く作られてきた．プロジェクトの主眼がコーパスの作成に置かれることが多いことから，これらのコーパス作成ツールは作成するコーパスに特化し，必ずしも汎用性が高くない．したがって，既存のツールを別のコーパス作成のために再利用することが困難であった．たとえば，Serengeti (Stührenberg et al., 2007) はテキスト中の共参照関係のアノテーションのために開発されたツールであり，共参照関係をアノテーションするためには適しているかもしれないが，これを評価対象とその属性をアノテーションして評判分析のためのコーパスの構築するために利用することは困難である．

一般にソフトウェア・システムを作成するにはコストがかかるため，同じような機能を持つツールをコーパスごとに作成するよりは，汎用のツールを用い，ツールの作成のための資源をコーパス作成に割り当てる方が望ましい．また，コーパス作成の初期の段階ではアノテーションの方法が必ずしも厳格に定義できていることは少なく (Marcus et al., 1993)，アノテーションの仕様の変更にともなってツールの変更も必要となる可能性もある．このような観点からも汎用的で柔軟なコーパス作成ツールが望まれている．これまでにも汎用のアノテーションツールを開発する試みはいくつかあったが (Orăsan, 2003; Cunningham et al., 2002; Mueller and Strube, 2001)，必ずしも成功しているとはいえない．

Dipper らはコーパス作成のためのアノテーション・ツールを汎用性に関する以下の 7 つの観点から分類している．(1) 扱うデータの多様性，(2) 多層のアノテーション，(3) アノテーションの多様性，(4) 簡便さ，(5) カスタム化可能性，(6) 品質の保証，(7) 相互変換可能性．我々もこれらの特に (1) から (5) の観点を重視して汎用性の高いアノテーション・ツール SLAT (Segment and Link-based Annotation Tool) (Noguchi et al., 2008) の開発を進めてきた．しかし，これらの観点はいずれも，アノテーションの対象となるコーパス中心の観点であり，コーパス作成の過程をどのように管理するかという視点が欠けている．コーパスの規模はますます大きくなり (Davies, 2009)，またアノテーションされる情報もますます複雑化・多様化している．今日ではすでにアノテーションされた既存のコーパスにさらに別の情報をアノテーションするような多層的なアノテーションが一般的となっている (Miltsakaki et al., 2004; Iida et al., 2007). 前述の汎用ツールはこのような多層的なアノテーションを扱うことは考えていない．以上のような背景から，アノテーションの汎用性を重視した SLAT を拡張し，コーパス作成過程の管理まで視野に入れた枠組を開発し，Slate (SLAT Enhanced) として実装を進めている．本稿では，まず，コーパス作成過程において必要となる機能を洗い出し，それを基礎として設計した枠組と Slate のインタフェースについて紹介する．

[1] take@cl.cs.titech.ac.jp

2 コーパス作成に求められるもの

コーパスに対する要求は，量的な拡大と同時に，互いに関連した様々な情報を重層的に付与するといった質的な拡大も高まっている．この要求を満たすためには，アノテーション・ツール自身の柔軟性やインターフェースの洗練が必要であることはもちろん，複数のアノテーションの間の関係や作成プロセスに関わるアノテータやデータの管理までも視野に入れる必要がある．ここでは，前述の Dipper らの考察に加え，より広い視野からコーパス作成においてシステムが支援すべき項目を検討し，以下の項目を洗い出した．

(1) ユーザ管理・役割管理
ある程度の規模のコーパス作成には複数の人間が関与するのが普通である．これらのユーザはコーパス全体の設計やアノテーションすべき情報を設計する管理者と実際のアノテーション作業をおこなうアノテータに大別することができる．システムはこれらのユーザ種別とその権限を管理できなければならない．

(2) タスクの割り当てと進捗管理
管理者はアノテータにタスクを割り当て，アノテータの進捗を管理する．進捗によってはタスクを別のアノテータに割り振ることも必要かもしれない．

(3) 新しいタスクの生成
管理者は新しいアノテーション・タスクを容易に生成できなければならない．

(4) タスクの修正
アノテーション・タスクの初期の段階では，アノテーションの設計が流動的でアノテーションの設計自身の修正が発生することがある．このような場合，システムはアノテーション・タスクの修正に柔軟に対応できなければならない．

(5) アノテーションの分析・統合
複数のアノテータがコーパスを分割してアノテーションすることを考えると，同じテキストに異なるアノテータがアノテーションした際の一致率などの分析や，テキストを分割してアノテーションした際の結合などを支援する必要がある．

(6) 版管理
アノテーションの設計変更によりコーパスに複数の版ができる可能性がある．いわゆる版管理の機能が必要である．

(7) 多層アノテーション
すでにアノテーションされたコーパスにさらに別の情報をアノテーションするような多層的なアノテーションが最近では多く試みられている．このためには，現在作業中のアノテーションから既にアノテーションされた情報を参照するなどの機能が必要となる．

(8) 他システムとの連携
アノテーションの種類によっては，自動的にアノテーションをおこない，それを人手で修正した方が効率がよい場合もある．これを実現するためには，他のシステムの出力を柔軟に受け入れるなどの機構が必要である．

(9) 入出力の拡張性
種々の入出力フォーマットに対応できる必要がある．

(10) 多言語処理
依然としてその量においては英語のコーパスが主流であるが，現在では様々な言語のコーパスが作成されるようになってきた．多言語への対応は必須である．

3 枠組みの概要

前節で述べた項目は大別すると，枠組みの問題 (1)〜 (7) と実装の問題 (8)〜(10) に分類することができる．図 1 に我々の枠組みの概要を UML で記述したものを示す．この枠組みで直接的に対応しているのは前述の項目のうち (1)〜(4) と (7) である．

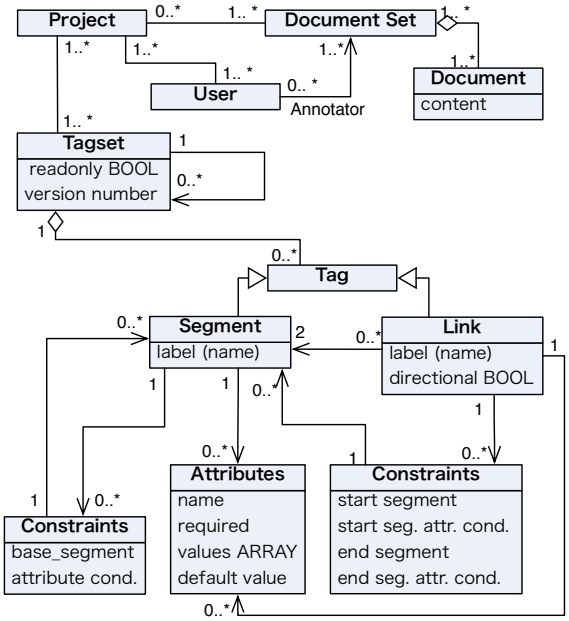


図 1: 枠組みの UML 表現 (簡略版)

文字列から成る Document は，内容を格納するための最小の要素である．複数の Document をまとめて Document set を構成できる．たとえば，ひとりのアノテータがタグ付けする Document のまとまりを Document set として定義するのは自然な使い方であろう．

Project はアノテーション・タスクを定義するもので，ひとつ以上の Document set を参照できる．Document set は Project とは独立に定義されているので，複数の Project から同じ Document set を参照することもできる．この構造によって多層的なアノテーションを自然に実現できる．

User (アノテータ) はひとつ以上の Project に属することができ，Project ごとに異なる Document set のタグ付けをおこなうことができる．また，Project はひとつ以上の Tagset を使うことができるので，たとえば，形態素レベルの Tagset と同時に統語レベルの Tagset を用い，多層的なアノテーションをおこなうことができる．

Tagset には高橋らが提案した Segment と Link (高橋, 2006) の基本要素が含まれる．Segment はどのような Attribute や Constraint を持つかを規定することによって定義し，実際のアノテーションによってそのインスタンスが作成され，label，Attribute の値が設定される．

大規模で複雑なコーパスの構築には複数のアノテータが関与することになる．つまり，ひとつの Project には複数の User が関連付けられる．User には種々の権限が付与されるのが普通であり，その管理も支援対象として考慮すべきであろう．User は，Project を企画し監督する立場の管理者と，指示にしたがって Document に対してアノテーションをおこなう作業者に大別できると考えられる．管理者は Project で用いる Tagset を決めたり，Document set をシステムにアップロードし，作業者に割り当てる．場合によっては作業の途中で Tagset に修正を施すかもしれない．一方，作業者にはこのような権限はないのが普通であるが，タグ付けのインタフェースにおける「見え」をカスタマイズすることは許されるかもしれない．これらの複雑な権限管理は Project と User の属性として定義することができる．

すでに述べたように我々の枠組では，ある Document set に対して複数の Tagset を関連付けることができるので，多層的なアノテーションを容易に実現できる．逆に，ひとつの Tagset を複数の Document set に関連付けることができるので，異なる Document のセグメント間にリンクを付与することも可能である．これはたとえば，複数文書とそれらの複数文書から生成された要約の間の関係を記述するのに有用である．

我々の枠組ではアノテーションの定義をユーザがアノテーション対象とは独立におこなえることが最大の特徴である．これによって，既存のアノテーションを利用したアノテーションを定義すれば多層的なアノテーションを実現できる．多層のアノテーションを容易にするために，定義したアノテーションには，制約を課すことができる．このような制約は不用意なアノテーションの誤りを防ぐのにも役立つ．たとえば，品詞のアノテーションを含むコーパスに，固有表現のアノテーション，さらには共参照のアノテーションを多層的におこなう場合を考えよう．固有表現の情報を付与するのは品詞として名詞の情報が付与されているものに限定する，あるいは共参照の情報を付与するのは固有表現の特定のクラスのものに限定するなどの制約を定義すれば，人為的なミスを防ぐことができる．

図 1 の枠組を使ってアノテーション作業がどのように管理できるか例を用いて説明する．Penn Treebank (Marcus et al., 1993) のように形態素・統語情報を付与するプロジェクトを考えよう．たとえば，これを Project A とする．Project A では Tagset X を使うものとしよう．次に Project A の成果の上にさらに PropBank (Kingsbury and Palmer, 2002) のような述語-項構造を付与することを考える．これを Project B として，Tagset Y を使うものとする．これらのプロジェクトをまったく独立におこなうと，2 つのアノテーションの間で情報を付与する対象に矛盾をきたす可能性がある．我々の枠組では，Tagset X と Y の間に適切な制約を記述することにより，不用意なミスを防ぐことができる．この場合，Project B は Tagset Y と同時に Tagset X も参照することになる．

4 Slate

SLAT では Web ブラウザベースのインタフェースでセグメントとリンクによるアノテーションを提供していたが，Slate (Segment and Link-based Annotation Tool Enhanced) では，User，Project，Document set などの管理も含めた実装となっている．また，SLAT では Javascript を用いて実装していたために，動作速度が遅いという問題が指摘されていたが，Slate では Adobe Flash を用いて全面的に実装をやり直したために，動作が高速化できた．また，Adobe Air 環境を用いて Web ブラウザとは独立のアプリケーションとしても動作できる．

コーパス作成作業の流れの概略は以下のようになる．

管理者の作業

- Project の作成
- Tagset の定義
- アノテーションする文書のアップロード
- Document set の作成
- Users (アノテータ・アカウント) の作成
- Project の定義 (User，Tagset，Document set の関連付け)
- アノテーション作業の割当

アノテータの作業

- 割当られた作業の選択
- アノテーション作業の開始

図 2 に Slate の管理画面の例を示す．これは新規プロジェクトを作成する画面で，タブを切り換え User，Tagset，Document set を指定することにより，これらを関連付けることができる．アノテータは図 3 に示すインターフェース画面からアノテーション作業をおこなう．基本的なインターフェース

は SLAT を踏襲しているが，タグ・セットの表示，タグ一覧，文書情報などを右側に集約し，これらをサイズ変更が可能なペインとして用意した点が大きく異なる．

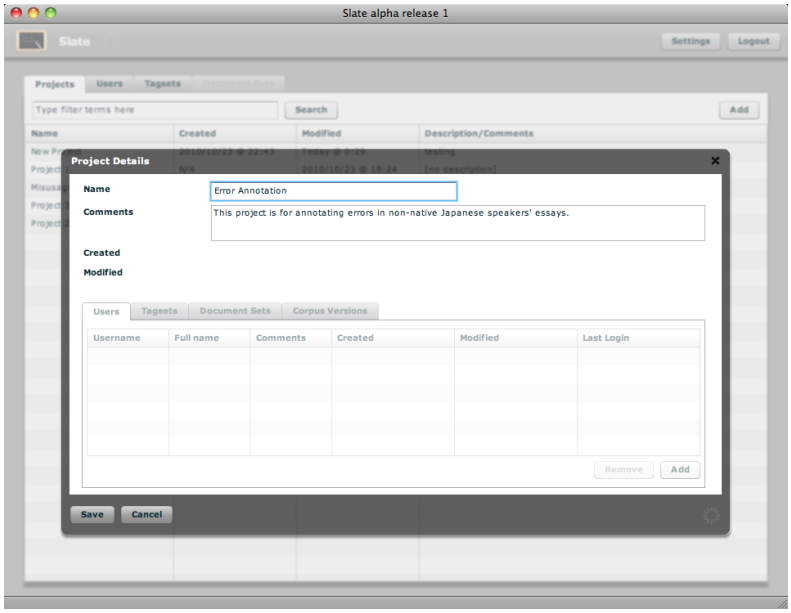


図 2: Slate の管理画面の例 (プロジェクトの作成)

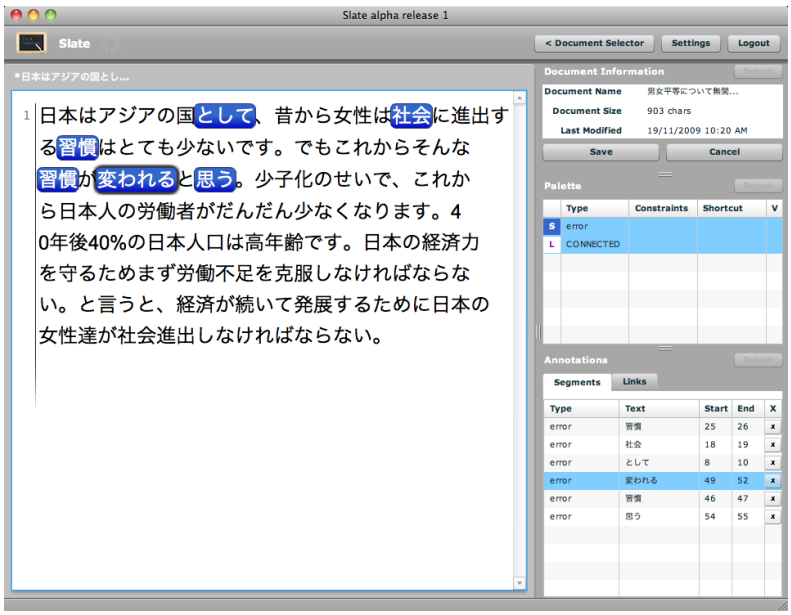


図 3: アノテーション・インタフェース

5 おわりに

本稿では，アノテーション・タスクの複雑な構造をアノテーション・タスクに関わる実体の関係を関係–実体モデルでモデル化する枠組みを提案した．この枠組によれば，これらの実体とその関係の定義によりアノテーションを定義することができる．また，近年のアノテーションの主流である多層的なアノテーションも自然に扱うことができる．この枠組の実装として Slate を紹介した．

文献

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an architecture for development of robust hlt applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+) design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1989–1993.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2237–2240.

Christoph Mueller and Michael Strube. 2001. MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50.

Masaki Noguchi, Kenta Miyoshi, Takenobu Tokunaga, Ryu Iida, Mamoru Komachi, and Kentaro Inui. 2008. Multiple purpose annotation using SLAT – Segment and link-based annotation tool. In *Proceedings of 2nd Linguistic Annotation Workshop*, pages 61–64.

Constantin Orăsan. 2003. PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop of Discourse and Dialogue*, pages 39–43.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Alexander Mehler, and Irene Cramer. 2007. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the Linguistic Annotation Workshop*, pages 140–147.

高橋 哲朗，乾 健太郎. 2006. アノテーションツール “Tagrin” の紹介. 言語処理学会第 12 回年次大会発表論文集, pages 228–231.

BCCWJ と関連ツールの相互運用

狩野　芳伸（ツール班協力者：東京大学 情報学環）†
橋田　浩一（ツール班分担者：産業技術総合研究所 社会知能技術研究ラボ）

Interoperability of BCCWJ and Related Tools

Yoshinobu Kano (Interfaculty Initiative in Information Studies, the University of Tokyo)
Koichi Hashida (Social Intelligence Technology Research Laboratory, the National Institute of Advanced Industrial Science and Technology)

１．はじめに

近年、一般に利用可能な日本語の言語資源(注釈つきコーパスやソフトウェアツール)が充実しつつある。基盤的な言語資源の場合、再利用や組み合わせが頻繁に必要となり、相互運用性が求められる。しかしながら、言語資源は異なる組織・異なる時期において開発され、互換性がないことが多い。そこで、科研費特定領域研究「日本語コーパス」の成果を中心に、日本語言語資源間の相互運用を UIMA/U-Compare に準拠して実現した。データの処理やソフトウェアツールの実行の相互連携を極力自動化して容易にしている。これら互換化実装そのものに加え、実装の方法を調達仕様書のテンプレートとして公開し、多くの利用者による多様な資源の互換化作業を省力化することにより、仕様の明確な資源の相互運用が自然に普及するように配慮した。

２．背景

２.１ UIMA

UIMA (Unstructured Information Management Architecture)とは、非構造化データのための相互運用性を提供するフレームワークである。UIMA は国際標準化団体 OASIS により承認された国際標準であり、その実装は Apache UIMA としてオープンソースで公開されている。学術・産業の双方で利用者は増加しており、特に自然言語処理で広く利用され、UIMA 互換言語資源は数多く公開されている。

UIMA は XML 記述によるプログラミング言語非依存のメタデータ定義をさまざまなレベルで提供しており、API は Java および C++のものが存在する。その設計はコンポーネント志向というべきもので、あらゆる処理は UIMA コンポーネントからなる workflow の実行により行われる。コンポーネントは入れ子にでき、子の実行順序はプログラマブルで動的な制御も可能であるため、理論的にはほとんどあらゆる順序が実現できる。コンポーネントはウェブサービスとしても展開でき、ローカルサービスと混在させて透過的に実行できる。大規模データに対応した並列処理も可能である。

UIMA コンポーネントは CAS と呼ばれる構造体を受け取り、必要に応じ処理結果等を追加して返す。CAS は生テキストと付加データ(以下アノテーションと呼ぶ)から成り、アノテーションは相互に参照可能なため任意のグラフ構造を表現できる。アノテーションは明示的に型付けされる。Type system と呼ばれる型階層の定義は XML で記述され、開発者が提供する必要がある。アノテーションは生テキスト中の文字位置を用いてテキストと関連づけ

† kano@is.s.u-tokyo.ac.jp

することを想定している。これは stand-off annotation style と呼ばれ、XML のような in-line style よりも重層的なデータの取り扱いが容易である。

２.２ U-Compare

UIMA は優れたフレームワークであるが、あくまでフレームワークであるため、コンポーネントそのものや型定義は標準としては提供されない。また、基本的に開発者を想定しており、一般ユーザにとって UIMA を用いたシステムを構築するのはそれほど容易ではない。

我々はそういった UIMA の不足点を補い必要な機能を実装した統合自然言語処理システム U-Compare を開発しオープンソースライセンス(LGPL)で公開している。U-Compare の狙いの一つは、ユーザや開発者の負担を軽くするために、ツールへのアクセス・組み合わせ・実行を相互運用性に基づき徹底して自動化することとである。もう一つの狙いは、単に既存のものをつなげる以上に、比較・評価・解析・視覚化といった自然言語処理で必要とされる豊富な機能を UIMA 準拠で提供することにある。U-Compare はおおまかには platform 部分と互換 component 群に分かれる。

互換 component 群はすべて U-Compare type system に互換であり、入出力条件さえ満たせば単に実行順序を指定するだけで実行できることを保証している。Component 群は以下に述べる platform とは独立に使用可能である。U-Compare ではこれまで提供していたのは英語の言語資源のみであった。

platform 部分は、任意の UIMA コンポーネントに対し、workflow 作成 GUI、比較・評価機能、結果の視覚化など自然言語処理に必要な様々な機能を提供している。これらの機能は汎用であり、以下に述べる日本語資源であってもそのまま利用できる。

platform を通じて U-Compare の component を利用する場合、platform のみならず component についても、インストール・更新・実行が自動的に行われる。既存の component を組み合わせて使うだけであれば、プログラミングは全く不要である。一方でコマンドラインモードも用意されており、GUI で作成した workflow をコマンドラインツールとして実行するといったことも可能である。

既存ツールを UIMA 対応にする場合、多くは入出力形式の変換が必要で、これを wrapping と呼ぶ。通常は UIMA の Java または C++ API を用いる。

他の言語で実装されたツールの場合や、ソースコードの改変にコストがかかる場合のために、我々は標準入出力経由でオリジナルのツールとやりとりし UIMA コンポーネント化する native tool wrapper を開発した。この wrapper はツールを実行するプロセスのリカバリ機能も備えている。これを用いれば開発者はフォーマット変換部分のみを実装すればよい。標準入出力経由であれば言語非依存であるため、Java でのコーディングをしなくともよいように、スクリプト言語で扱いやすい標準形式も用意した。これにより変換処理を好みの言語で行うことができる。このように標準入出力を用いると、実行時にプロセスが分離されエラーが波及しないことと、入出力形式さえ同じならば指定するコマンド名を変えるだけで wrap するツールを変更できるという利点もある。

３．BCCWJと関連ツールの相互運用

既存の言語資源の相互運用を考える場合は、以下のような点を考慮する必要がある。

まず、相互運用という場合、たいていはなんらかの変換処理が必要になる。そのときに、

表現形式が変わってもオリジナルの情報を失うことがないよう、復元するのに十分な情報を保持すべきである。

次に、stand-off style を用いるからには、もともとのテキストをそのまま保持し、ツールの処理結果は極力アノテーションの側で扱うのが望ましい。ほとんどのツールは本質的に元テキストに改変を加えることはないからである。テキスト部分を不変とすれば、各ツールの影響の範囲が明確になると同時に、本来不要な前・後処理も考える必要がなくなり、入出力条件が簡素化される。たとえば、改行や空白はツールによっては無視されたり特殊な扱いを受けたりするが、こういった特殊文字もオリジナルのまま保存するようにする。

また、ユーザ・開発者からみたデータ構造のわかりやすさも考慮する必要があると考える。たとえば形態素列を取得したい場合、形態素型のオブジェクトを順に反復処理するのがわかりやすいが、コーパスによっては形態素的なものが平行して二種類ついていることもある。適切に型と型階層を定義することで、プログラムからの読み込みを容易にすることが必要であろう。

UIMA におけるデータ処理の単位は常に CAS であるが、CAS がテキストのどの部分を保持するべきかは特に定められていない。分散処理の可能性など処理効率を考えれば処理単位は小さいほどよいが、UIMA 枠内で複数の CAS にまたがる依存関係を扱うのは困難であるため、依存関係が閉じた最小のテキスト領域を CAS に対応させるのが妥当である。

UIMA コンポーネントを作成する際は、その機能単位に注意する必要がある。UIMA はあくまで枠組みでありどのような実装も可能であるが、コンポーネントが再利用を前提にしたブロックであると考えると、既存ツールをそのままコンポーネントにするのは適切でないことも多い。再利用という観点からすると、より小さな機能単位に分割したほうが理論的には再利用の可能性が高まる。しかし、小さすぎると入出力条件が複雑になりがちで、組み合わせての再利用がかえって難しくなる。我々は、入出力条件が入出力 type の静的リストで表現できる機能単位であるという条件下で、極力小さな機能単位に分割し、コンポーネント化している。

type と type system の定義はこれらの点を考慮しつつ、対象とする言語資源で必要な概念を十分表せるよう定義しておく必要がある。

こういった点を踏まえて、我々は日本語言語資源のうち再利用される可能性の高い基盤的な注釈付きコーパスやツール群を UIMA コンポーネント化した。その実装は U-Compare ウェブサイトでオープンソースで公開されている。U-Compare platform を通じた利用であれば、他のコンポーネントの利用と同じくマウス操作だけで完結する。

３.１ Type Systemの設計

全体に共通して用いられる type として、文境界を表す `Sentence`、形態素を表す `Morpheme`、係り受けを表す `Dependency`、述語項構造を表す `FrameRelation`、モダリティ情報を表す `Modality` を定義した。それぞれオリジナルの情報を復元できるよう必要なフィールドセットが定義されている。特定の言語資源で用いられる type については以下各言語資源の項で触れる。

また、id 値による XML タグ間参照や、in-line スタイルで間接的に表現されていたアノテーション間の関係を、直接的にリンクとして表現している。たとえば UIMA Java API を使用して `Dependency` から係り先 `Morpheme` を取得するには、`getTarget` メソッドを呼び出せばよい。

３.２ BCCWJ

代表性を有する大規模日本語書き言葉コーパス(以下では「日本語コーパス」という)は、国立国語研究所を中心に開発されている 1 億語を超える規模の均衡コーパスである。そのうち一部のテキストについては、形態素・係り受け・述語項構造・モダリティ等の情報が付加されている。形態素および拡張モダリティ情報についてはXML形式で表現されており、これを読み込める BCCWJReader を開発した。係り受けについては後述の Cabocha と同形式であり、述語項構造については別途 Reader を開発した。

３.３ GDAコーパス

GDA (Global Document Annotation, 大域文書修飾)とは、統語的依存関係、代名詞等の照応、共参照、多義語の語義など、広汎な言語情報を XML で表現可能なフォーマットである。代表的な GDA 形式のコーパスとしては、毎日新聞 3000 記事に対するアノテーションが GSK より公開されている。

GDA のアノテーションは他に比べ非常に細密であり、タグの種類も膨大である。形態素は他と同等であるため `Morpheme` を用いたが、他の情報については GDA 向けに別途 type をいくつか定義し、大まかなタグ種以外は文字列フィールドとしてタグ種名を保持した。

GDA の特徴の一つは、人間のアノテーション作業をサポートするために、(復元可能な)省略を許すなど工夫がされている点にある。UIMA 読み込み後の使用は機械的なものが主であるため、こうした省略は復元して格納した。また、GDA における統語的な構造は句構造であり、深い入れ子になっているが、その親子関係は in-line style で間接的に表現されている。これを UIMA 側に読み込む際は明示的な親子関係を抽出し表現した。親子関係は U-Compare の木構造表示コンポーネントと組み合わせれば視覚化することもできる。

３.４ 京都大学テキストコーパス (Ver.4)

京都大学テキストコーパスは、毎日新聞記事計約 4 万文に対して形態素・構文情報を付与したもので、うち 5,000 文に対しては、格関係、照応・省略関係、共参照の情報が付与されている。前者は一般的な非交差係り受け情報であるが、後者はそれと平行して関係情報が追加されており、形態素境界も必ずしも一致しない。我々はこれらすべての情報を読み込める KyotoCorpusReader を開発し、それぞれを別個の系列として格納するようにした。

３.５ 形態素解析ツールChasen

Chasen は NAIST 松本研究室で開発された、品詞付けを含む形態素解析ツールである。我々は native tool wrapper を用いて ChasenWrapper コンポーネントを作成した。Chasen は品詞以外にも基本形や読みの判定を行うため、出力される情報はすべて `Morpheme` のフィールドに格納した。

Chasen はルールベースで文境界を判定し、その結果を(デフォルトでは)改行として出力する。stand-off style という観点からは、元テキストに修正を加えずあくまで stand-off ポジションで表現したいので、文境界は `Sentence` により明示的に出力し、後段の処理で文境界を分ける必要があるときはこれを用いて wrapper 内で適宜処理することとした。

３.６ 係り受け解析ツールCabocha

Cabocha は Support Vector Machine を用いた係り受け解析器である。我々は、Cabocha の入出力指定オプションを固定した上で、native tool wrapper を用いて、`Morpheme` をうけとり `Dependency` を返す UIMA コンポーネント CabochaWrapper を作成した。上述のように内部的に `Sentence` を用いて入力を区切ったうえで処理している。

３.７ アノテーションツールChaki

Chaki (茶器)は NAIST 松本研究室で開発されているアノテーションツールで、特に形態素や係り受け関係の編集が容易に行えるよう設計されている。Chaki では拡張 Cabocha 形式ファイルでのインポートおよびエクスポートが可能であり、この形式で読み書きをする UIMA コンポーネント ChakiReader および ChakiWriter を開発した。拡張 Cabocha 形式では、Cabocha の扱う形態素と係り受け情報に加え、`Group`, `Link`, `Segment` によるタグを用いた拡張がなされており、日本語コーパスに含まれている交差した係り受けを表現できる。これら三種の情報を表せるよう type system を拡張した。

今後、Chaki は東京工業大学徳永研究室で開発されているアノテーションツール Slate とも互換化される予定である。

３.８ 中納言

中納言は国立国語研究所で開発されたコーパス検索システムで、Web アプリケーションとして公開されている。中納言での検索は形態素解析を前提としており、独自の入力形式がある。我々は UIMA 側の `Morpheme` 列を中納言で読み込み可能な形式のファイルに変換し保存する UIMA コンポーネントとして ChunagonWrapper を開発した。

４．実装と調達仕様書のテンプレート化

我々の目的は、即座に利用可能な互換コンポーネントを提供するのに加え、第三者が新たな互換コンポーネントを容易に作成できるようにすることにある。

前述の互換化された言語資源群は、日本語の自然言語処理においてもっともよく使われるであろう形式やデータタイプをカバーしている。また、英語を含め一般に言語処理でよく用いられる形式の多くについて、読み書きできるコンポーネントを U-Compare から配布している。これらの実装をテンプレートとして再利用すれば、大概の言語資源互換化作業はごく一部の修正で済む。

現在の type system で対応しきれていない type が必要なときは、新たに定義する必要がある。意味的な互換性を保つためには type の互換性が必要であるため、第三者が新たに定義する際は必要に応じて我々も型設計作業をサポートしたいと考えている。

さらに、互換化したい既存言語資源の形式が明確に定義されているのであれば、実装作業は比較的単純であり、簡単な調達仕様書を書けば発注も容易であるので、そのテンプレートを用意した。

５．おわりに

我々は BCCWJ を中心とした日本語言語資源について、国際標準 UIMA に準拠して互換コンポーネント群を作成公開した。これらの組み合わせと実行にプログラミング作業は不

要である。これらは単独で利用可能であるが、任意の UIMA コンポーネントに対応した言語処理システム U-Compare にも統合し、U-Compare の提供する様々な機能と共に簡単に言語資源を使用できるようにした。また、ソースコードと調達仕様書をテンプレートとして公開し、新たな互換コンポーネント作成の際の省力化を図っている。今後はさらなる日本語言語資源の追加や、基盤機能の拡張を予定している。

謝辞

本研究の一部は、科学研究費補助金特定領域研究「日本語コーパス」および基盤研究 C(21500130)の助成を受けて行われた。作業において多大なご協力をいただいた「日本語コーパス」ツール班と関係者の皆様方、特に松本裕治氏(NAIST)、森田敏生氏(総和技研)、山崎誠氏(国語研究所)、中村壮範氏(国語研究所)、徳永健伸氏(東京工業大学)、Dain Kaplan 氏(東京工業大学)、乾健太郎氏(東北大学)、小町守氏(NAIST)、松吉俊氏(NAIST)には深謝申し上げたい。

関連URL

特定領域「日本語コーパス」ホームページ：http://www.tokuteicorpus.jp/

U-Compare 日本語ページ： http://u-compare.org/japanese.html

U-Compare, UIMA-based integrated NLP system： http://u-compare.org/

Apache UIMA：http://uima.apache.org/

アノテーションツール茶器(Chaki)： http://sourceforge.jp/projects/chaki/

コーパス検索アプリケーション「中納言」：http://morph.kotonoha.gr.jp/chunagon/

GDA (大域文書修飾)：http://www.i-content.org/gda/

GSK 新聞記事GDAコーパス 2004：http://www.gsk.or.jp/catalog/GSK2009-B/catalog.html

形態素解析器Chasen：http://chasen-legacy.sourceforge.jp/

日本語係り受け解析器Cabocha：http://chasen.org/~taku/software/cabocha/

京都大学テキストコーパス：http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

拡張固有表現タグ付きコーパスの構築

橋本泰一 (ツール班分担者：東京工業大学 総合プロジェクト支援センター)†

Constructing Extended Named Entity Annotated Corpora

Taiichi Hashimoto (Research Project Support Center, Tokyo Institute of Technology)

1．はじめに

「机」「椅子」「空」「愛」といった一般的な概念を表す表現ではなく，「夏目漱石」「東京オリンピック」「日本」などの物，イベントや考え方を表す言語表現を固有表現と呼ぶ．固有表現は，質問応答，情報抽出，機械翻訳，テキストマイニングなどの応用技術に用いられる自然言語処理におけて重要な基礎知識である．日本語においては，評価型ワークショップ IREX において，新聞記事をベースに固有表現タグ付きコーパス（CRL 固有表現データ[*1]）が構築され，そのコーパスをもとに日本語における固有表現認識に関する研究が進んだ．そして，様々な固有表現認識手法 [山田 07, 山田 04, 浅原 04, 中野 04, 渡辺 04, 笹野 08, 土屋 08] が提案されてきた．

IREX で定義された固有表現の種類は，組織名，人名，地名，日付表現，時間表現，金額表現，割合表現，固有物名の 8 種類である．この 8 種類のタグを毎日新聞記事（1,174 記事数）に付与したコーパスが CRL 固有表現データである．しかし，このコーパスを利用して開発された固有表現認識器を，質問応答システム，情報抽出システムやテキストマイニングに利用しようとしても実際に認識できる固有表現の種類が少なく，新聞以外の分野の文書に対する精度も十分満足のいくレベルではない．

そこで，新たな固有表現の定義として，「関根の拡張固有表現階層」（以下，拡張固有表現）[*2]が提案された．「関根の拡張固有表現階層」は，MUC(Message Understanding Conference) プロジェクト [Gri96]，IREX プロジェクト [Sek00]，ACE(Automatic Content Extraction) プロジェクト[*3]の各定義をもとに，関根が拡張を行った固有表現の定義 [Sek08, Sek04, Sek02] である．拡張固有表現の大きな特徴は，固有表現の種類の豊富さである．MUC では，組織名，人名，地名，日付表現，時間表現，金額表現，割合表現の 7 種類，IREX では，MUC の 7 種類に固有物名を加えた 8 種類を固有表現として定義している．一方，拡張固有表現（バージョン 7.1.0）では，物やイベントなどの「名前」「時間表現」「数値表現」を最初の階層とし，最大 4 階層で構成され，200 種類の固有表現の定義が定められている．これは様々な自然言語処理技術への応用を考慮し，新聞記事や百科事典などに見られる概念や単語を考慮していることに起因する．細かい分類を活かして，Wikipedia から固有表現辞書を獲得する手法についても研究が進んでいる．[渡邉 08, 杉原 09]

これまで，200 種類の固有表現を定義した「関根の拡張固有表現階層」をもとに，白書，書籍，Yahoo!知恵袋，新聞，雑誌の各コアデータ，白書と毎日新聞に対して，固有表現タグを付与してきた．また，白書，書籍，Yahoo!知恵袋の 3 コアデータに対して，機械学習アルゴリズムの一つである CRF をベースに固有表現抽出に関する評価実験を行い，精度が約 80%，再現率が約 46%，F 値が約 60% という結果を得た．[橋本 10] 本稿では，現在タグ付け作業を行っているブログコアデータへの経過について報告する．

† hashimoto.t.ab@m.titech.ac.jp

*1 http://nlp.cs.nyu.edu/irex/index-j.html

*2 http://nlp.cs.nyu.edu/ene/

*3 http://www.itl.nist.gov/iad/mig/tests/ace/

表 1　拡張固有表現タグ付きコーパスの概要（2011 年 2 月 7 日時点）

	文書数	総文字数	平均文字数	形態素数	平均形態素数	表現数		平均表現数	達成度(%)
						異なり	のべ		
白書コア	62	351,649	5671.8	228,651	3687.9	5,797	11,089	178.9	100
書籍コア	83	369,391	4450.5	-	-	5,485	13,683	164.9	100
知恵袋コア	938	179,345	191.2	110,649	118.0	3,449	5,407	5.8	100
新聞コア	340	563,562	1480.1	-	-	-	38,103	100.3	100
雑誌コア	88	399,264	4537.1	-	-	8,346	19,708	224.0	100
ブログコア	300	91,366	304.6	-	-	-	3,585	12.0	63.7
毎日新聞	8,584	3,643,361	424.4	-	-	63,545	252,763	29.4	-
白書	400	2,340,364	5850.9	-	-	23,857	74,203	185.5	-
CRL	1,174	593,763	505.8	-	-	7,153	19,254	16.4	-

2．拡張固有表現タグ付きコーパス

平成 19 年度は，毎日新聞および白書に対し，拡張固有表現 (Version 7.1.0) の定義に則ってタグ付けを行った [橋本 08]．毎日新聞は 8,584 記事に対し，のべ 252,763 個，異なり 79,632 個のタグを付与し，白書は 400 文書に対し，のべ 74,203 個，異なり 23,857 個のタグを付与した．これまで利用されていた研究に用いられていた CRL 固有表現データは，毎日新聞 (1,174 記事，のべタグ数 19,254 個，異なりタグ数 7,153 個) にタグ付けされたものであった．平成 20 年度は，白書，書籍，Yahoo!知恵袋各コアデータに対してタグ付けを，平成 21 年度は，新聞，雑誌の 2 コアデータに対してタグ付けを行った．平成 22 年度は，書籍コアデータの置き換え（2 文書），雑誌コアデータの追加（9 文書，全 88 文書）を行った．加えて，新たにブログコアデータへのタグ付け作業を行っている．これまでに作成した拡張固有表現タグ付きコーパスの概要（2011 年 2 月 7 日時点）を表 1 に示す．従来，固有表現タグ付きコーパスとして利用されている CRL 固有表現データに比べ，文字数で約 13 倍の文書に対してタグ付きコーパスを構築した．

白書コアデータ（62 文書，総形態素数 228,651）に対して，のべ 11,819 個，異なり 5,276 個の固有表現が，書籍コアデータ (83 文書) に対して，のべ 13,683 個，異なり 4,884 個の固有表現が，Yahoo!知恵袋コアデータ (938 文書，総形態素数 110,649) に対して，のべ 5,609 個，異なり 3,270 個の固有表現が付与された．新聞コアデータ (340 文書) に付与されたタグは，のべ 38,103 個，雑誌コアデータ (88 文書) に付与されたタグは，のべ 18,250 個，異なり 8,346 個であった．2011 年 2 月 7 日時点では，300 文書のブログに対してタグ付けが終了しており，のべ 3,585 個のタグが付与されている．ブログコア全体の約 63.7% が終了している．

書籍コア，新聞コア，雑誌コア，ブログコアの形態素数が明記されていないのは，コアデータに対して人手により付与された正確な形態素情報が入手できなかったためである．また，新聞コア，ブログコアの固有表現の異なり数が記載されていないのは，個人情報保護のために一部の固有表現に対して伏せ字処理が行われているため正確な数を計算できなかったためである．

2．拡張固有表現解析ツール

このコーパスを用いて固有表現抽出ツールを作成した．その概要図を図 1 に示す．

固有表現認識手法として，機械学習アルゴリズムの一つである Conditional Random Fields（CRF）を用

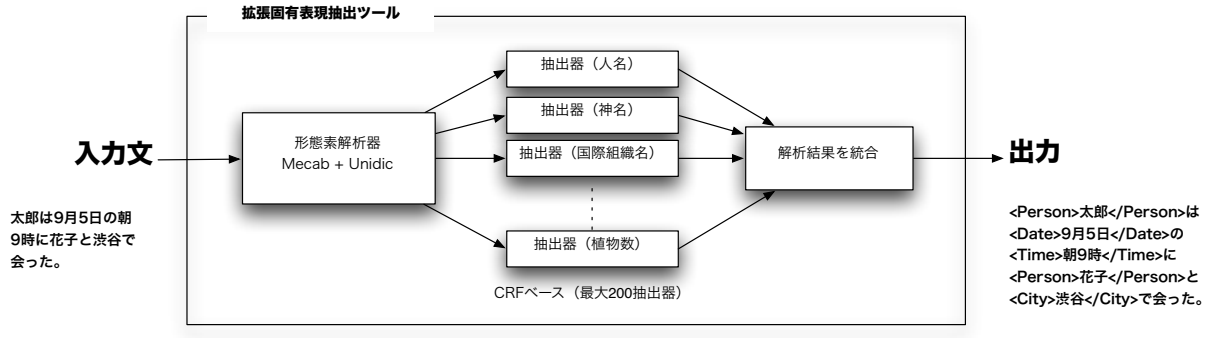


図 1　拡張固有表現抽出ツール概要図

いた手法を，同定するチャンクタグの方式は IOB2 を採用した．入力はコアデータに人手付与された形態素および品詞とし，学習および認識で用いた素性は，該当形態素および単語，前後 2 形態素および単語，前 2 拡張固有表現タグを利用する．認識方向は文頭から文末に向かって行い，各拡張固有表現のタグごとに学習と認識を行う．認識された拡張固有表現タグが互いに入れ子や交差する場合も存在する．

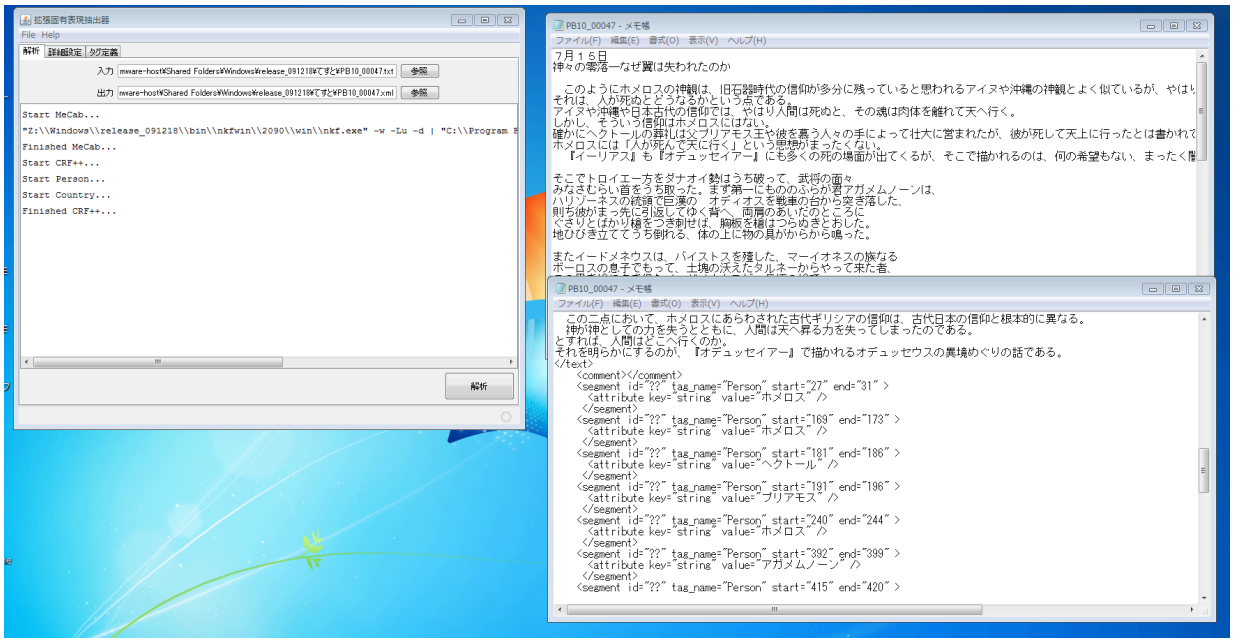
評価実験では，白書，書籍，Yahoo!知恵袋 3 コアデータ，コア以外の白書，毎日新聞の 5 種類を用い，10 分割交差検定により評価し，精度が約 80%，再現率が約 46%，F 値が約 60% という結果を報告した [橋本 10]．コア以外の白書および毎日新聞は，Unidic（バージョン 1.3.12）[伝 09] と Mecab（バージョン 0.98）を用いて形態素解析を行った．本ツールも評価実験に用いたコーパスにより学習モデルを作成しているため，同程度の認識精度を示すと考えられる．

ツール本体は，プログラミング言語 Ruby をベースにコマンドライン用のツールとして作成されている．しかし，ユーザの多くは Windows を利用していることを考慮し，プログラミング言語 Java により，マルチプラットフォームに対応した GUI インターフェースを作成した．この GUI インターフェースをもちいて，拡張固有表現を抽出した文書とタグの種類を指定し簡便に抽出器を実行することができる．(図 2)

3．おわりに

本稿では，様々なジャンルの固有表現タグ付きコーパスの構築に向けて，拡張固有表現抽出ツールについて報告した．「関根の拡張固有表現階層」の定義 (Version 7.1.0) に則って，2011 年 2 月 7 日時点で白書 (62 文書)，書籍 (49 文書)，Yahoo!知恵袋 (600 文書)，新聞 (340 文書)，雑誌 (88 文書)，ブログ (300 文書) のタグ付けされたコーパスを作成した．機械学習アルゴリズムの一つである CRF をベースとした固有表現抽出器を作成した．評価実験により，精度が約 80%，再現率が約 46%，F 値が約 60% という評価結果を報告しており，本ツールはほぼ同程度の認識精度があることが期待される．また，Windows ユーザの利用を考慮し，簡便な GUI インターフェースを作成した．

今後の課題として，タグ付け作業の効率化が急務であると感じた．特に，200 種類もの他種類のタグ付けを行うためには，タグ付けツールのユーザインタフェースの改良とタグ付け結果の一貫性のチェックを効率化する必要がある．また，拡張固有表現認識精度の向上のためのアルゴリズムの改良，ツールの高速化が挙げられる．


拡張固有表現抽出器
File Help
解析
詳細設定
タグ定義
入力
出力
参照
Start MeCab...
Finished MeCab...
Start CRF++...
Start Person...
Start Country...
Finished CRF++...
解析
PB10_00047 - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
7月１５日
神々の零落―なぜ翼は失われたのか
PB10_00047 - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
</text>

<attribute key="string" value="ホメロス" />

<attribute key="string" value="ホメロス" />

<attribute key="string" value="ヘクトール" />

<attribute key="string" value="プリアモス" />

<attribute key="string" value="ホメロス" />

<attribute key="string" value="アガメムノーン" />

図 2　GUI インターフェース スクリーンショット

謝辞

本実験を実施するにあたり，ニューヨーク大学の関根聡氏には，毎日新聞記事への拡張固有表現タグデータのご提供，およびタグ修正作業に対する多大なる助言をいただきました．ここに，心より感謝の意を表します．

参考文献

[Gri96] Grishman, R. and B. Sundheim: Message Understanding Conference - 6: A Brief History, in *COLING-96*, 1996.
[Sek00] Sekine, S. and H. Isahar: IREX: IR and IE Evaluation project in Japanese, in *LREC2000*, pp. 1977–1980, 2000.
[Sek02] Sekine, S., K. Sudo, and C. Nobata: Extended Named Entity Hierarchy, in *LREC2002*, 2002.
[Sek04] Sekine, S. and C. Nobata: Definition, Dictionary and Tagger for Extended Named Entities, in *In Proceedings of the Forth International Conference on Language Resources and Evaluation*, 2004.
[Sek08] Sekine, S.: Extended Named Entity Ontology with Attribute Information, in *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2008.
[橋本 08] 橋本, 乾, 村上：拡張固有表現タグ付きコーパスの構築, 情報処理学会自然言語処理研究会 (2008-NL-188), 2008.
[橋本 10] 橋本, 中村：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, 2010.
[笹野 08] 笹野, 黒橋：大域的情報を用いた日本語固有表現認識, 情報処理学会論文誌, Vol. 49, No. 11, pp. 3765–3776, 2008.
[山田 04] 山田, 工藤, 松本：Support Vector Machine を用いた日本語固有表現抽出, 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2004.
[山田 07] 山田：Shift-Reduce 法に基づく日本語固有表現抽出, 情報処理学会自然言語処理研究会 (NL-179-3), pp. 13–18, 2007.
[杉原 09] 杉原, 増市, 梅基, 鷹合：Wikipedia カテゴリ階層構造の固有名詞分類実験における効果, 情報処理学会自然言語処理研究会 (2009-NL-189), pp. 57–64, 2009.
[浅原 04] 浅原, 松本：日本語固有表現抽出におけるわかち書き問題の解決, 情報処理学会論文誌, Vol. 45, No. 5, 2004.
[中野 04] 中野, 平井：日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, Vol. 45, No. 3, 2004.
[伝 09] 伝：多様な目的に適した形態素解析システム用電子化辞書, 人工知能学会誌, Vol. 24, No. 5, pp. 640–646, 2009.
[渡辺 04] 渡辺, 桝井, 福本：固有表現抽出ツールＮＥｘＴの精緻化とユーザビリティの向上, 言語処理学会第 10 回年次大会, 2004.
[渡邉 08] 渡邉, 浅原, 松本：グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類, 人工知能学会論文誌, Vol. 23, No. 4, pp. 245–254, 2008.
[土屋 08] 土屋, 肥田, 中川：非頻出語に対して頑健な日本語固有表現の抽出, 情報処理学会自然言語処理研究会, pp. 1–6, 2008.

BCCWJ コアデータへの係り受け・並列構造アノテーション

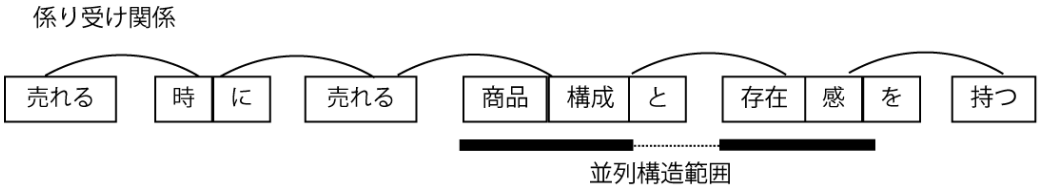
浅原　正幸（ツール班分担者：奈良先端科学技術大学院大学）†
岩立　将和（ツール班協力者：奈良先端科学技術大学院大学）
松本　裕治（ツール班班長：奈良先端科学技術大学院大学）

Annotation of Dependency and Coordinated Structure on the BCCWJ Core Data

Masayuki Asahara　(Nara Institute of Science and Technology)
Masakazu Iwatate　(Nara Institute of Science and Technology)
Yuji Matsumoto　(Nara Institute of Science and Technology)

１．はじめに

本稿では BCCWJ コアデータに対する係り受け・並列構造アノテーションについて述べる。係り受け構造は、主述の関係、修飾-被修飾の関係、並列・同格の関係、接続-被接続の関係を文節単位に付与する。並列・同格構造に対しては、文節単位の関係づけのみならず、形態素単位（国語研短単位）に対応する範囲を付与する。図１に係り受け関係と並列構造の例を示す。図１中の四角１マスは国語研短単位をなし、連結されている単位が文節をなす。文節間の係り受け関係は弧で示し、基本的に最右文節を根とする木をなす。形態素単位に付与される並列構造は、並列句の範囲を太下線で示し、その対応関係を点線で示す。



[出典：ＰＢ４６＿００００６６]

図１　係り受け関係と並列構造範囲

以下では、２節で係り受け・並列構造の付与基準を先行研究である京都大学テキストコーパスの係り受けアノテーションと対比しながら説明する。３節では、実際のアノテーションの工程とともに２０１１年２月７日時点での進捗について示す。

２．係り受け関係、並列・同格構造アノテーション基準

係り受け構造が付与された日本語のコーパスに「京都大学テキストコーパス」（以下「京大コーパス」）がある。この先行研究の基準（黒橋 2000）と同様に活用形や付属語の文法的働きに従い、主述の関係、修飾-被修飾の関係、並列・同格の関係、接続-被接続の関係に対して、文節間の係り受け関係を付与する。以下では京大コーパスの基準と BCCWJ の基準とで、特に異なる点について示す。

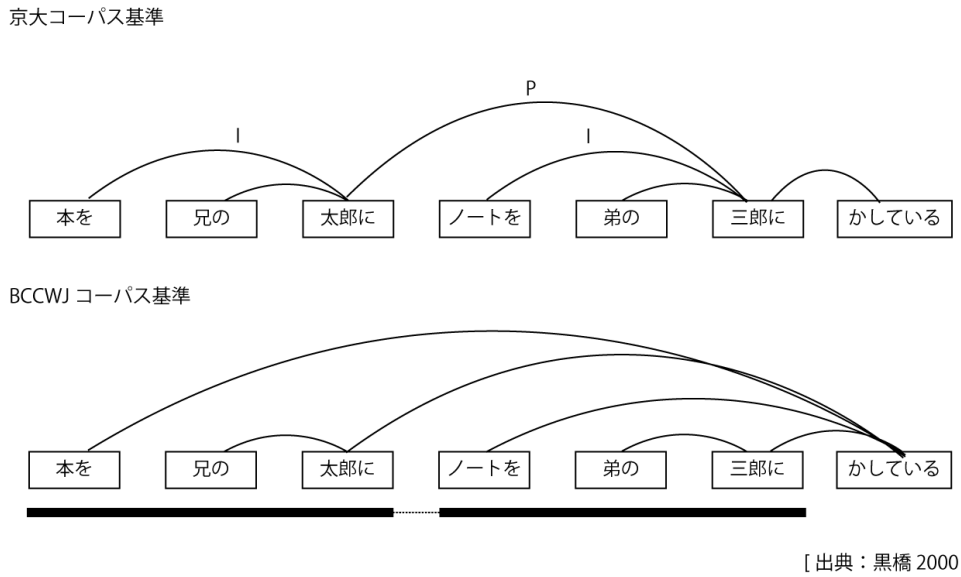
† masayu-a@is.naist.jp

２.１　並列・同格構造の範囲と係り受け関係種別

京大コーパスでは「並列関係（Ｐラベル)」・「部分並列内の関係（Ｉラベル)」・「同格関係（Ａラベル)」・「通常の係り受け関係（Ｄラベル)」の４種類を区別して係り受け関係を付与している。この係り受け関係の種別は並列・同格関係を区別するために導入されている。これに対し、BCCWJ では、係り受け関係に種別を設けない。BCCWJ においては、別途、並列・同格構造の範囲を付与することによってこの区別を廃止する。

京大コーパスにおいて部分並列内の係り受け関係については、非交差制約を遵守するために、木構造上もっとも近い祖先に係り先を移動する手続きを行い、その移動を意味するためにＩラベルが用いられていた。BCCWJ においては、並列・同格構造を範囲で対応関係とともに示し、係り先は真の係り先に係ける。この方法により、部分並列の情報を係り受け関係のラベルとして保持することを回避する。

図２に２つの基準における係り受け関係基準の対比について示す。尚、以降の図では形態素境界と京大コーパス基準におけるＤラベルは省略する。

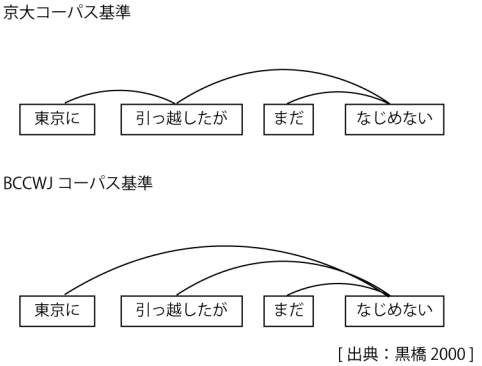


［出典：黒橋 2000］

図２　京大コーパスにおける関係ラベルと BCCWJ における並列構造の範囲付与

２.２　主語以外の格要素と従属節・主節の述語との関係

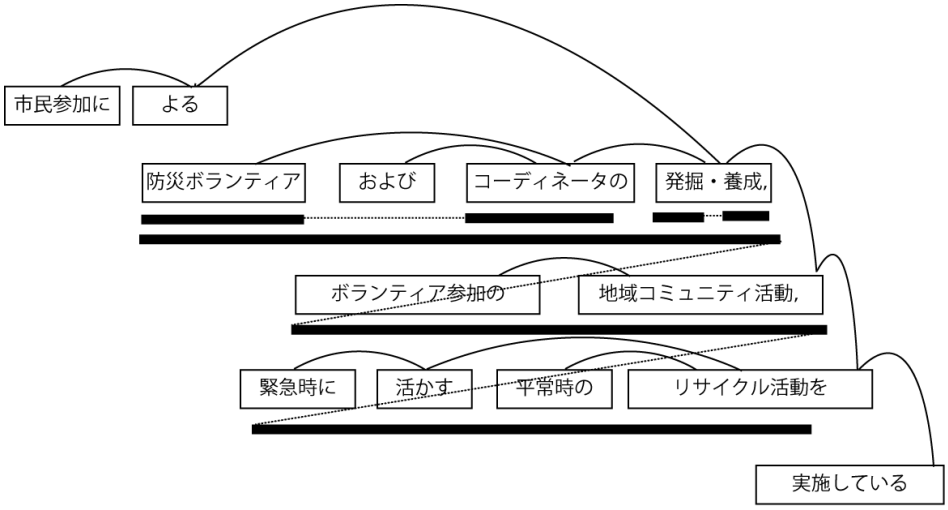
京大コーパスでは、主語以外の格要素（ヲ格、ニ格など）は、複数節の述語全てに係る場合に、最左従属節に係けるとし、その他の節の格要素は省略されたとみなす。これに対し、BCCWJ では係りうる主節（最右節）の述語に係ける。図３に２つの基準間の対比を示す。図中「東京に」は京大コーパスでは近い述語「引っ越したが」に係けるが、BCCWJ コーパス基準では遠い述語「なじめない」に係ける。



［出典：黒橋 2000］

図３　複数の述語に係る格要素

２.３　複数構成句を持つ並列構造・並列句の入れ子・接続表現・構成句に係る要素



[出典：OW６X＿００００５６]

図４　並列構造に対する各種アノテーション

並列構造は係り受け構造の中で例外的な構造を多く含む。ここでは図４の例文により、並列構造に関連する様々な現象に対してどのようにアノテーションを行うかを説明する。

・３つ以上の並列：例文中「防災ボランティアおよびコーディネータの発掘・養成」と「ボランティア参加の地域コミュニティ活動」と「緊急時に活かす平常時のリサイクル活動」が並列をなし、図のように並列構造を構成する句（以下「構成句」と呼ぶ）の範囲をタグ付けする。この例では、各構成句は、最後の構成句のすぐ右にある「を」を格助詞として共有する。この場合、係り受け関係は隣の構成句に係けることとする。

・並列構造の入れ子：並列構造の範囲については入れ子を許す。先の３つの並列構造の最左要素「防災ボランティアおよびコーディネータの発掘・養成､」中には、「防災ボランティア」-「コーディネータ」と「発掘」-「養成」の２つの別の並列構造を含むが図のように並列構造の範囲を付与する。

・接続表現：接続表現は並列構造の構成句内の要素とせず、構成句間の要素とする。接続表現が１つ以上の文節をなす場合、右隣接構成句の最右文節に係ける。図の例において､「および」は「コーディネータ」に係ける。

・並列構造の構成句に係る要素：並列構造の左から、並列構造内の構成句全てに係る要素は、並列構造の最左構成句内の係けるべき要素に係ける。図の例において、「(市民参加に)よる」は「発掘・養成､」に係ける。

２.４　その他

その他､BCCWJ の係り受け･並列アノテーションにおいて､以下のような基準を定める：

・テ形・述語並列：テ形などの述語並列は並列としない。

・～から～まで：静的な述語に対して範囲を表す場合は「～から」は「～まで」に係ける。動的な述語に対して経路を表す場合には「～から」も「～まで」も述語に係ける。

・非交差制約の廃止：部分並列の中の関係を廃止したために係り受け関係が交差しうるが

交差を許す。

３．作業工程と進捗

本節では実際の作業工程を示すとともに進捗について報告する。全工程を図５に示す。事前工程として国語研によるコアデータサンプリング・書き起こし・短単位形態論情報付与（文境界付与を含む）・文節境界付与がある。文節境界付与は計画当初ツール班で対応予定であったが、長単位形態論情報と文節単位が深く関連することから国語研で付与していただくことになった。短単位形態論情報付与と文節境界付与との間に時間差があり、またコアデータの分野（OW, PB, PN, OC, PM, OY の６種類）ごとにさらに時間差がある。このことから短単位形態論情報が付与されたデータから順に「並列・同格範囲付与」「係り受け修正（１次）」・「係り受け修正（２次）」・「リリース作業」の４つの工程に分割して作業を順に進める。係り受け修正作業は、解析器を構成してその結果を修正するが、構成する解析器の訓練元データの違いにより２回行う。１次作業では京大コーパスから訓練して構成した解析器の出力を多人数で修正し、大量のデータの粗い誤り修正を目的とする。２次作業では１次作業で構成されたデータから訓練して構成した解析器を構成し、１次作業の結果と解析器の結果の齟齬を少人数で修正することによりアノテーションの一貫性を担保することを目的とする。

以下では各工程の具体的な手続きを説明し、その後進捗について報告する。

３.１　並列・同格範囲付与

６種類のコアデータに対し、短単位形態論情報が付与された順に並列・同格範囲を付与する。図５のフローチャート中では[データ①]に対して[人手作業（あ）]を行い[データ③]を生成する部分である。

具体的な作業はスプレッドシートソフトウェア上に短単位形態論情報を読み込んで形態素単位に BIO(並列構造)・bio(同格構造) を付与することにより行う（図６Ｅ列参照)。B および b はそれぞれ並列および同格の構成句の開始形態素を示す。I および i は構成句の開始形態素以外の形態素を示す。O および o は構成句間にある形態素（主に接続表現）を示す。構成句を構成しない形態素と構成句間外にある形態素には何も付与しない。作業者に注意を促すために図６Ｄ列１３行のように、接続表現についてはあらかじめ注釈を付けておく。またアノテーションは入れ子を許す。

アノテーション作業は英語に対する並列構造アノテーション作業の経験を持つ作業者１名が全て行う。当初アノテーション支援環境の構築を試みたが、作業者がスプレッドシートソフトウェアの扱いに十分慣れていたために既存のソフトウェアを用いた。

並列・同格範囲のアノテーション作業時には、文節境界を示しておらず係り受け構造を意識せずに作業する。次に説明する係り受け修正（１次）において、係り受け構造と並列・同格範囲を照らし合わせて齟齬がある場合には、齟齬に気づいた時点で修正を行う。

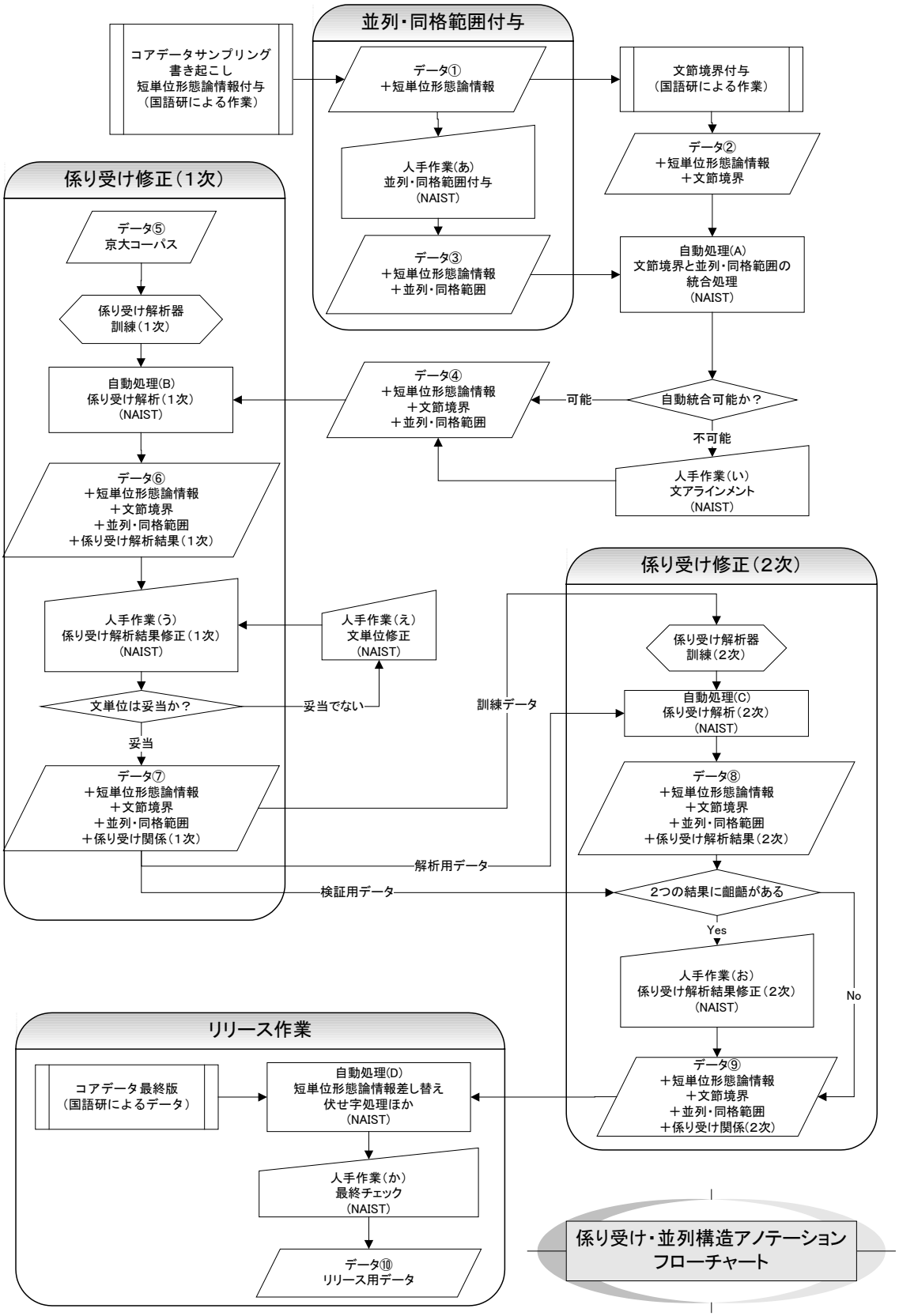


図５　係り受け・並列構造アノテーションフローチャート

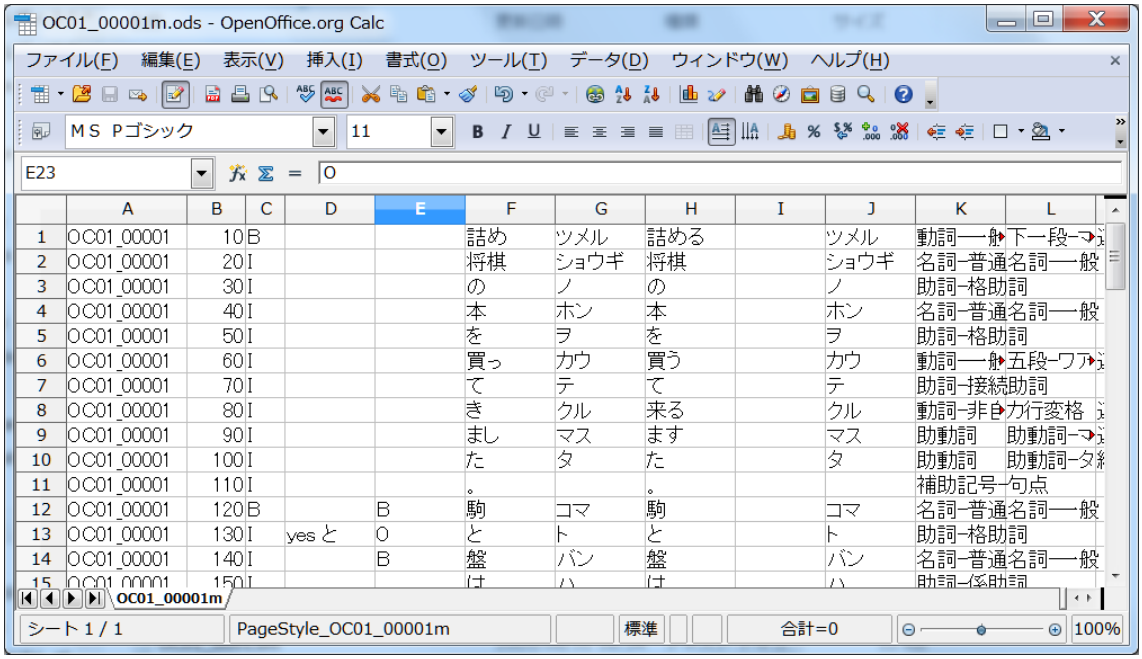


図６　並列・同格アノテーション

３.２　係り受け修正（１次）

並列・同格範囲付与[人手作業（あ）]と国語研による文節境界付与作業が終わったデータより順に係り受け修正（１次）（図５左枠内）を行う。

この作業を行うための準備作業として、並列・同格範囲を付与したデータ③と文節境界を付与したデータ②を統合する必要がある。簡易プログラムを作成し表層形をチェックしながら２つのアノテーションを自動統合する処理[自動処理(A)]を実施したが、並列・同格構造付与の前後で書き起こし誤りや位取り記数法の変更などによる表層形のずれが生じ、自動統合できないデータが多々あった。これに対し、人手で文アラインメントを取る作業[人手作業（い）]を行い、データ④を作成した。

次に係り受け解析器の構成を行う。京大コーパスに対して、係り受け情報を保持したまま、形態素解析用辞書 UniDic-1.3.12 と形態素解析器 MeCab-0.98 を用いて短単位形態論情報を付与し、データ②のうち OW, PB データで訓練した条件付確率場に基づく文節区切り器を用いて国語研文節単位を付与することにより、データ⑤を作成した。このデータ⑤中、２つの文節単位が真部分集合になっていないようなデータは、単純な係り受け木操作で対応が取れないために削除し、このデータを用いて係り受け解析器を訓練した。係り受け解析器として、係り受け関係の交差を認めないトーナメントモデル（岩立 2008）を用いる。学習器は３次の多項式カーネルを用いた Support Vector Machines を用いる。

訓練した係り受け解析器を用い、データ④を解析し[自動処理(B)]データ⑥を得る。このデータ⑥の文節境界、係り受け関係、並列・同格範囲を人手で見て確認して修正を行う[人手作業（う）]。修正作業には ChaKi.NET 1.3～1.5 を用いる。この修正作業は現在のところ７人で並列して行っている。

人手による修正作業中に、小数点やＵＲＬ中のピリオドなどで文が分割され、妥当でな

い文単位が定義されることが多々あった。文単位修正機能が ChaKi.NET 1.5 以降に実装され、別途文単位修正作業[人手作業（え）]を行った。ChaKi.NET 1.4 以前では文単位修正ができなかったために、文単位に誤りがある文の ID を管理しておき、ChaKi.NET 1.5 が公開され次第、作業者２人により修正作業を行った。

上記工程を経て、係り受け修正（１次）が完了したデータ⑦が完成する。

３.３　係り受け修正（２次）

係り受け修正（１次）が完了したデータ⑦は、随時作業者間で基準の統制を行っているとはいえ、一貫性を保持することは困難である。表１に OW データ作業時の作業者５人の間のアノテーション一致率を示す。

表１　OW データに対する作業者間のアノテーション一致率（%）

	文節（文単位）	並列範囲（文単位）	同格範囲（文単位）	係り受け（文単位）	係り受け（文節単位）
最大値	100.0	94.4	96.4	62.5	94.5
平均値	98.4	81.2	94.1	58.5	92.2
最小値	95.2	75.9	87.5	48.0	89.9

京大コーパスで評価される一般的な係り受け解析器の精度 92%前後（岩立 2008）と比較しても一致率が低いことがわかる。そこで２回目の係り受け修正作業（図５右枠）を行う。

具体的には、データ⑦を用いて係り受け解析器を構成し、交差検定することによりアノテーションの誤りを検出する[自動処理(C)]。元のデータ⑦と新たに構成した係り受け解析器の出力データ⑧とを比較する。係り受け解析器として、前述のトーナメントモデルを用いて非交差制約ありと非交差制約なしの解析器の２つを導入する。学習器として効率性のため２次の多項式カーネルに基づくオンライン学習器 opal (Yoshinaga 2009) を用いる。OW データに対して訓練データ・テストデータに全く同じデータを用いるという設定で行った係り受け一致率を表２に示す。

表２　OW データに対する１次修正データと解析器出力の一致率（%）

	係り受け（文単位）	係り受け（文節単位）
非交差制約あり	64.2	92.8
非交差制約なし	60.0	91.9

１次修正データと解析器出力が異なる文に対して、手作業で再修正を行う[人手作業(お)]。再修正作業は、アノテーション基準の一貫性を重視し、少人数で行う。このようにして２回の人手修正を経たデータ⑨を得る。

３.４　リリース作業

２回の人手修正を経たデータ⑨とは別に、国語研において形態論情報などの修正作業が行われ再度表層形などに齟齬が出現すると考える。またリリース時にはコアデータの一部に伏せ字処理が行われる。再度形態論情報などを自動統合し[自動処理(D)]、完全に自動処理できない部分について人手でチェックを行い[人手作業(か)]リリース用データ⑩を生成する。リリース用データは ChaKi.NET で読み込める .db ファイルと、国語研が作成する形

態論情報付き XML データに対するスタンドオフアノテーションの２形式を作成する。

３.５　進捗

２０１１年２月７日現在、図５中「並列・同格範囲付与」および「係り受け修正（１次)」の作業を行っている。各データの人手作業の進捗率を表３に示す。

表３　「並列・同格範囲付与」および「係り受け修正（１次)」の進捗率

	人手作業（あ） 並列・同格範囲付与	人手作業（い） 文アラインメント	人手作業（う） 係り受け解析結果修正	人手作業（え） 文単位修正
OW	100%	100%	100%	100%
PB	100%	100%	95%	77%
PM	100%	79%	16%	12%
PN	100%	100%	10%	0%
OC	100%	100%	50%	0%
OY	53%	0%	0%	0%

今後の予定であるが、２０１１年３月末までに、全データに対して[人手作業（あ）（い）]を、OW, PB, OC, PM に対して[人手作業（う）（え）]を、完了させる（全体の 70%相当)。残りのデータは（全体の 30%相当）に対する[人手作業（う）（え）]は、２０１１年４月以降も引き続き行い、約 400~500 時間・人の作業量を想定している。「係り受け修正（２次)」は全体の文の約 40%に対して必要であり、アノテーションの齟齬の類型化（浅原 2010）を行いながら、経験と専門知識を持つ人員により効率的に進める。

４．おわりに

本稿では BCCWJ コアデータに対する係り受け・並列アノテーション作業について現状を報告した。早期の全データ公開に向けて引き続き作業を進める所存である。

文献

Naoki Yoshinaga and Masaru Kitsuregawa(2009) “Polynomial to Linear: Efficient Classification with Conjunctive Features**”.** In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1542-1551.

浅原正幸、岩立将和、松本裕治(2010). 「BCCWJ コアデータに対する係り受け・並列構造アノテーション～進捗と課題～」特定領域「日本語コーパス」平成 21 年度公開ワークショップ予稿集, 2010 年 3 月 13 日.

岩立将和、浅原正幸、松本裕治(2008). 「トーナメントモデルを用いた日本語係り受け解析」自然言語処理, Vol. 15, No. 5, pp.169-185.

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

参考ＵＲＬ

京都大学テキストコーパス Version 4.0 Web ページ：
http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html.

BCCWJ に対する
述語項構造と照応関係のアノテーション

小町 守（ツール班分担者：奈良先端科学技術大学院大学情報科学研究科）*
飯田 龍 （ツール班協力者：東京工業大学大学院情報理工学研究科）

Annotating Predicate-Argument Structure and Anaphoric Relations to BCCWJ

Mamoru Komachi (Nara Institute of Science and Technology)
Ryu Iida (Tokyo Institute of Technology)

1 はじめに

形態素解析・統語解析の研究が成熟しつつある中、意味・談話解析の研究も近年発展を遂げている。たとえば、Gildea and Jurafsky (2002) は英語の述語に対する深層格解析のタスクを意味役割付与 (Semantic role labeling) と呼び、Fillmore and Baker (2000) に基づいた自動解析手法を提案した。また、PropBank (Palmer et al., 2005) は意味役割を大規模にアノテートした初めてのコーパスである。

これらの自動解析手法は CoNLL (Conference on Computational Natural Language Learning) の共通評価タスクによって取り上げられ（2004–2005, 2008–2009）、様々な解析手法が検討された。そして、本年開催される CoNLL 2011 の共通評価タスクでは、PropBank を含む英中アラビア語の OntoNotes コーパス (Hovy et al., 2006) を用いた共参照解析タスクが設定されるなど、意味解析を要素技術とする次の基盤技術の研究も盛んになってきた。また、述語項構造は情報抽出 (Harabagiu et al., 2005; Surdeanu et al., 2003) など広く応用先のある要素技術である。

一方、動詞や形容詞以外にも事態を表す名詞（事態性名詞と呼ぶ）があることが知られており、Meyers et al. (2004a) は NomBank (Meyers et al., 2004c,b) コーパスを作成した。日本語におけるゼロ照応（省略）解析は事態性名詞の解析タスクと類似しており、英語の事態性名詞の項構造解析を取り扱った (Gerber and Chai, 2010) が ACL でベストペーパーを飾るなど、事態性名詞の解析も注目されている。

日本語においては京都テキストコーパス 4.0 (Kawahara et al., 2002) が形態素情報・統語情報に加え、「関係タグ」と呼ばれる共参照や照応関係も含んださまざまな情報を付与しているほか、GDA コーパス (Hasida, 2005) も agent や theme などの意味役割や共参照の情報が付与されている。また、NAIST テキストコーパス (飯田他, 2010) には、京都テキストコーパス 3.0 に対し、照応・共参照・述語項構造の情報が付与されている。解析済みブログコーパス『KNB コーパス』(橋本他, 2009) はブログ記事を対象に京都テキストコーパス同様の格・省略・照応情報を付与したものである。

我々は乾・飯田 (2009) が述べるように、網羅性の高くかつ多様な分野のテキストに重層的な意味情報を付与することを目的とし、BCCWJ に照応と述語項構造のアノテーションを行なった。タグ付与基準に関しては (飯田他, 2010) で述べた NAIST テキストコーパスの仕様を踏襲し、語彙概念構造辞書 (竹内, 2004) を参照してアノテートした。

* komachi@is.naist.jp

表 1: BCCWJ の述語項構造のアノテーションにおいて助動詞とした語

	パターン	例
補助動詞系	V てあげる、V てもらう、など	代筆してあげる、食べてもらう
可能	V できる、V れる、V られる	説得できる、行かれる、見られる
受け身	V れる、V られる	騙される、認められる
使役	V せる、V させる	喋らせる、論じさせる
願望	V たい	食べたい

2 BCCWJ に対するアノテーション

2.1 述語項構造のアノテーション

述語項構造については、述語の基本形にその項となる表現を表層格（ガ格、ヲ格、ニ格）レベルでタグ付与する。

我々は BCCWJ に対して述語項構造のアノテーションをする際、NAIST テキストコーパス 1.5β(飯田他, 2010) のタグ付与基準[1] に従ってタグ付与を行なった。アノテーションマニュアルは随時ウェブページ上[2]で更新している。

述語の格要素に関しては、agent や theme などの意味役割（深層格）、PropBank で付与されている ARG0 や ARG1 といった意味役割相当のラベルなどが考えられるが、我々は NAIST テキストコーパスに従い、表層格レベルから格交替だけを原形に戻し、タグ付与を行なうことにした。たとえば、

(1) 太郎は次郎にカレーを**食べさせた**。

という文で、京都テキストコーパス 4.0 では述語「食べさせた」に対して＜ガ=太郎, ヲ=カレー, ニ=次郎＞という格構造でタグ付与するのに対し、我々は述語「食べ」に対し＜ガ=次郎, ヲ=カレー＞という項構造でタグを付与する。ただ、この場合使役者と述語の間に関係を認定することができないため、格要素を増やす助動詞に対して＜追加ガ（ニ）＞というタグを付与する[3]。つまり、助動詞「させた」に対し、＜追加ガ=太郎＞のようにタグを付与する。今回認定した助動詞のリストを表 1 にまとめる。一方、

(2) 太郎はカレーが好きだ。

のような動詞や可能動詞を含む二重主語構文においては、＜ハ＞と＜ガ＞を用いてタグを付与した。上記の場合、述語「好き」に対して＜ハ=太郎, ガ=カレー＞となる。

また、NAIST テキストコーパスのタグ付与において、FrameNet や NOMLEX (Macleod et al., 1997, 1998) のような述語項構造を示す辞書を用いなかったことによる反省を踏まえ、必須格と任意格の曖昧性がある場合、青山他 (2007); 大西他 (2008) および語彙概念構造辞書[4] (竹内, 2004) の知識を参考にし、必須格かどうかの判定を行なった。

必須格が曖昧な事例として、たとえば

(3) 私は彼に搾り取られた。

[1] http://cl.naist.jp/~ryu-i/coreference_tag.html
[2] https://sites.google.com/site/naistcorpus/predicate_tag
[3] 格の交替あるいは格の追加がない単語には「助動詞」タグを付与しない
[4] http://cl.it.okayama-u.ac.jp/rsc/lcs

という文があり、語彙概念構造辞書ではカラ格が必須格として「私カラ彼ガ搾り取る」として記述されているが、現在ガヲニの 3 格のみを必須格としてアノテーションしているため、述語「絞り取る」に対して＜ガ=彼＞、そして助動詞「れる」に対して＜ガ=私＞としてアノテーションする。同様に、

(4) 私は今日来なくてよいと言われた。

という文に対しても、語彙概念構造辞書ではニ格は必須格ではないとされているが、後継の動詞項構造シソーラス 0.902 [5] (竹内他, 2007) では必須格とされており、必須格かどうかの判定に辞書を使うと辞書の揺れによってアノテーションが左右されてしまうため、述語項構造を揺れなくつけるためには辞書の精錬も必要である。

2.2 事態性名詞のアノテーション

事態性名詞についても、述語と同様に表層格レベルで項を付与する。
事態性名詞については、モノを指す表現にも項を付与する。
事態性名詞のうち、モノを指す表現には、どの種類かタグ付与する。

冒頭で述べたように、我々は述語だけではなく事態性名詞に関しても項構造を付与した。たとえば、

(5) 太郎の採用は不当だ。

という文に関して、述語[6]「不当だ」のガ格になっている事態性名詞採用のヲ格として「太郎」をアノテートする。また、

(6) 太郎の料理はまずい。

という文で、料理は「太郎ガ料理スル」という事態を表すとともに、太郎が料理した結果物を指している。このように、事態性名詞の中には体系的に事態とモノの両方を指しうる種類のものがあるため、飯田他 (2010) で提案したように、「内容/結果物」「モノ」「役割」「ズレ」という 4 種類の分類を行い、モノを指しうる事態性名詞にアノテートした。

2.3 照応関係のアノテーション

また、我々は BCCWJ に対して照応関係のアノテーションを行なった。**照応**とは代名詞や指示詞などの照応詞によって他の表現を指す機能のことを指す。照応と似た関係として、**共参照**がある。共参照とは、2 つの表現が可能世界において同一の実態を指す機能のことを言う。たとえば、

(7) 太郎は iPhone を買った。彼はずっと欲しがっていたのだ。

において、太郎と彼は照応関係かつ共参照関係である。一方、

(8) 太郎は iPad を買った。次郎もそれ を買った。

において、iPad と それ は照応関係にあるが共参照関係にはない。

Mitkov (2002) によると、前者のように照応関係かつ共参照関係にある場合は identity-of-reference anaphora (IRA)、後者のように照応関係にあるが共参照関係にない場合を identity-of-sense anaphora (ISA) と呼ぶ。

[5] http://cl.it.okayama-u.ac.jp/rsc/data/index.html
[6] 動詞、形容詞、「名詞 + だ」を述語と認定する

表 2: BCCWJ の 4 ジャンルのコアデータに対する照応・述語項構造アノテーションの進捗

		記事数	文数	単語数	コアデータ
PN	（新聞）	478	5,730	127,077	A まで完了
PB	（書籍）	55	4,691	113,399	A まで完了
OW	（白書）	30	2,414	100,396	A まで完了
OC	（知恵袋）	938	6,402	103,188	B まで完了

今回 BCCWJ においてタグ付与を行なったのは、NAIST コーパス (飯田他, 2010) と同様、以下の基準に従う。

照応関係については、IRA の関係のみを対象として照応の関係を認定する。

また、NAIST コーパスには bridging reference や間接照応の情報 (Inoue et al., 2010) が付与されているが、BCCWJ においては付与していない。NAIST コーパスとの比較のためにこれらの情報をつけることは今後の課題である。

2.4 アノテーションの進捗

2011 年 2 月 7 日現在、BCCWJ に対して完了している述語項構造と照応関係のアノテーションの進捗について表 2 で示す。単語数は UniDic 1.3.12[7] を用いて MeCab 0.98[8] で自動解析した結果なので、人手解析結果と一致しない。また、雑誌コアおよびブログコアについては未着手である。

このうち新聞コアデータに関しては、複数の作業者間で一致率を見た。用いたデータは新聞記事 9 記事（90 文，1,653 語）である。一致率を求める手順は飯田他 (2010) と同様に、一方の作業者のタグ付与の結果を正解、他方の作業結果をシステムの出力とみなし、再現率と精度で評価して表 3 に示した。このうち作業者 A は NAIST テキストコーパスのアノテーションに従事した熟練の作業者、作業者 B,C は自然言語処理分野の大学院生である。

作業者 A-B 間の一致率を見ると、飯田他 (2010) が報告しているように、それぞれのタグ付与は多くの場合 8 割を超える品質で作業ができている。一方、A-C 間の一致率を見ると、多くの場合 A-B 間の一致率に比べて精度・再現率ともに低い。作業者 C に聞き取り調査をしてみると、いつ辞書を参照してよいかが分からない、という意見が得られた。これは、作業者 A-C 間ではガ格と比較してヲ格、ニ格のタグ付与の一致率が低いことからも、項構造の決定の際の辞書参照に任意性があることを示す。今後は辞書引きをアノテーションツールの機能として組み込むなど、辞書引きの不統一をなくすことで解決できると考えられる。

また、述語に比べて事態性名詞のラベル付与の一致率が低い原因の一つも、既存の動詞項構造シソーラスなどの辞書が述語を中心に作られており、事態性名詞の項構造のアノテーションに必ずしも助けにならない、という点も挙げられる。NomBank のように、事態性名詞に関する項構造のリソース (小町他, 2010) も、動詞項構造シソーラス同様構築していく必要がある。

[7] https://www.tokuteicorpus.jp/dist/index.php
[8] http://mecab.sourceforge.net/

表 3: 作業者間の新聞コアデータにおける一致率（精度と再現率）

	作業者 A-B 間				作業者 A-C 間			
	精度		再現率		精度		再現率	
述語	87.1	(155/178)	89.6	(155/173)	80.1	(133/166)	76.9	(133/173)
ガ格	80.4	(123/153)	100.0	(123/123)	96.2	(101/105)	100.0	(101/101)
ヲ格	96.3	(77/80)	98.7	(77/80)	85.3	(58/68)	96.7	(58/60)
ニ格	82.1	(23/28)	88.5	(23/26)	90.5	(19/21)	59.4	(19/32)
事態	93.9	(93/99)	84.5	(93/110)	72.9	(70/96)	63.6	(70/110)
ガ格	83.5	(76/91)	100.0	(76/76)	80.0	(32/40)	100.0	(32/32)
ヲ格	59.0	(13/22)	100.0	(13/13)	52.4	(11/21)	73.3	(11/15)
ニ格	77.8	(7/9)	100.0	(7/7)	50.0	(4/8)	100.0	(4/4)

3 照応と述語項構造のアノテーション

京都テキストコーパスや NAIST テキストコーパスでは新聞記事の分野しかカバーできなかったが、BCCWJ において、基本的なアノテーション方針は NAIST テキストコーパスの基準を踏襲しつつ、さまざまな分野に照応と述語項構造のアノテーションを拡充することができた。また、動詞の項構造辞書を用いることで、述語項構造関係のアノテーションの支援を図り、一定の改善を見ることができた。

OntoNotes コーパス (Hovy et al., 2006) は BCCWJ のように様々なジャンルのテキストに対し、統語構造と述語項構造、語義と共参照の情報をアノテートしたものである。彼らの目標は、機械学習の訓練データとして使うために、90%の一致率でアノテートをする、というものである。BCCWJ や NAIST コーパスのアノテーション基準においても、全ての項構造を 90%という高い一致率でアノテートすることができない。さらなる仕様の洗練と、Slate (徳永他, 2010) のようなアノテーションツールを組み合わせ、質の高いコーパスを作ることで、様々な応用の可能性が開けるであろう。

今後は作成されたコーパスを用いて述語項構造と照応関係の同時学習や、文書全体の大域的情報を用いた全体最適化、新聞記事で訓練されたモデルからの転移学習など、大規模で多様な分野のコーパスが整備されることで、解析技術の発展も期待できる。コーパスや辞書の拡充を続けつつ、意味解析の実用化に取り組んでいきたい。

参考文献

Fillmore, Charles J. and Collin F. Baker (2000) “FrameNet: Frame semantics meets the corpus,” in *Proceedings of the 74th Annual Meeting of the Linguistic Society of America*.

Gerber, Matthew and Joyce Y. Chai (2010) “Beyond NomBank: a Study of Implicit Arguments for Nominal Predicates,” in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1583–1592.

Gildea, Daniel and Daniel Jurafsky (2002) “Automatic Labeling of Semantic Roles,” *Computational Linguistics*, Vol. 28, No. 3, pp. 245-288.

Harabagiu, Sanda, Cosmin Adrian Bejan, and Paul Morarescu (2005) “Shallow Semantics for Relation Extraction,” in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI ’05)*, pp. 1061–1066.

Hasida, Koiti (2005) 「GDA 日本語アノテーションマニュアル 草稿 第 0.74 版」. `http://i-content.org/gda/tagman.html`.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel (2006) "OntoNotes: The 90% Solution," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 57–60.

Inoue, Naoya, Ryu Iida, Kentaro Inui, and Yuji Matsumoto (2010) "Resolving Direct and Indirect Anaphora for Japanese Definite Noun Phrases," *Journal of Natural Language Processing*, Vol. 17, No. 1, pp. 221–246.

Kawahara, Daisuke, Sadao Kurohashi, and Koiti Hasida (2002) "Construction of a Japanese Relevance-tagged Corpus," in *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pp. 2008–2013.

Macleod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves (1998) "NOMLEX: A Lexicon of Nominalizations," in *Proceedings of Euralex98*, pp. 187–193.

Macleod, Cathrine, Adam Meyers, Ralph Grishman, Leslie Barret, and Ruth Reeves (1997) "Designing a Dictionary of Derived Nominals," in *Proceedings of Recent Advances in Natural Language Processing*, pp. 142–151.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman (2004a) "Annotating Noun Argument Structure for NomBank," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pp. 803–806.

Meyers, Adam, Ruth Reeves, and Catherine Macleod (2004b) "NP-External Arguments: A Study of Argument Sharing in English," in *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pp. 96-103.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman (2004c) "The NomBank Project: An Interim Report," in *Proceedings of the HLT/NAACL 2004 Workshop Frontiers in Corpus Annotation*, pp. 24–31.

Mitkov, Ruslan ed. (2002) *Anaphora Resolution*, Studies in Language and Linguistics: Peason Education.

Palmer, Martha, Paul Kingsbury, and Daniel Gildea (2005) "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106.

Surdeanu, Mihai, Sanda Harabagiu, John Williams, and Paul Aaseth (2003) "Using Predicate-Argument Structures for Information Extraction," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8–15.

乾健太郎・飯田龍 (2009) 「日本語書き言葉コーパスへの重層的意味情報付与〜照応・共参照，述語項構造，モダリティ，談話関係〜」,『科研費特定研究「日本語コーパス」平成 21 年度全体会議予稿集』.

橋本力・河原大輔・黒橋禎夫・新里圭司 (2009) 「構文・照応・評判情報つきブログコーパスの構築」,『言語処理学会第 15 回年次大会論文集』，614–617 頁. http://nlp.kuee.kyoto-u.ac.jp/~hasimoto/KNBC_v1.0_090925.tar.bz2 よりダウンロード可能.

小町守・飯田龍・乾健太郎・松本裕治 (2010) 「名詞句の語彙統語パターンを用いた事態性名詞の項構造解析」,『自然言語処理』，第 17 巻，第 1 号，141–159 頁.

青山桜子・阿部修也・大西良明・乾健太郎・松本裕治 (2007) 「事態間関係の獲得のための動詞語釈文の構造化」,『言語処理学会第 13 回年次大会論文集』，286–289 頁.

大西良明・乾健太郎・松本裕治 (2008) 「事態間関係知識の整備と含意文生成への応用」,『言語処理学会第 14 回年次大会論文集』，1152–1155 頁.

竹内孔一 (2004) 「語彙概念構造による動詞辞書の作成」,『言語処理学会第 10 回年次大会論文集』，576–579 頁.

竹内孔一・乾健太郎・藤田篤・竹内奈央 (2007) 「語彙概念構造に基づく事態上位オントロジーの構築」,『言語処理学会第 13 回年次大会論文集』，859–862 頁.

徳永健伸・Dain Kaplan・飯田龍 (2010) 「汎用アノテーションツール Slate」,『情報処理学会研究報告. 自然言語処理研究会』，第 2010-NL-199 巻，1–10 頁.

飯田龍・小町守・井之上直也・乾健太郎・松本裕治 (2010) 「述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から」,『自然言語処理』，第 17 巻，第 2 号，25–50 頁.

BCCWJ に基づく中・長単位解析ツール

小澤俊介（電子化辞書班協力者：名古屋大学大学院情報科学研究科）†
内元清貴（電子化辞書班連携研究者：情報通信研究機構総合企画部）
伝康晴（電子化辞書班班長：千葉大学文学部）

Middle and Long Unit Word Analysis System Based on the BCCWJ

Shunsuke Kozawa (Graduate School of Information Science, Nagoya University)
Kiyotaka Uchimoto (Strategic Planning Department, NICT)
Yasuharu Den (Faculty of letters, Chiba University)

１．はじめに

文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」の電子化辞書班では、語彙形態論研究に適した短単位、音声研究に適した中単位、構文・意味研究に適した長単位という複数粒度の「語」を高精度（98%以上）で自動構成するシステムを提供することを目的のひとつとしている。その各単位の例を図 1 と図 2 に挙げる。図 1 は「固有名詞仮名表記に関して論文を三本執筆した。」という文における短単位、中単位、長単位の関係を、図 2 は短単位と長単位の関係を表している。長単位は中単位を、中単位は短単位をそれぞれ結合することにより構成できる。例えば、「固有名詞仮名表記」という長単位は「固有名詞」「仮名表記」という二つの中単位から成るとともに、さらに「固有」「名詞」「仮名」「表記」のように分割した四つの短単位から成る。本稿では、中・長単位を自動構成する方法とそのツール Comainu について述べる。

文	固有名詞仮名表記に関して論文を三本執筆した。																							
文節	固有名詞仮名表記に関して									論文を			三本		執筆した。									
長単位	固有名詞仮名表記				に関して					論文		を	三本		執筆し			た	。					
中単位	固有名詞		仮名表記		に関して					論文		を	三本		執筆し			た	。					
短単位	固有	名詞	仮名	表記	に	関し	て			論文		を	三	本	執筆	し		た	。					

図 1：短単位、中単位、長単位の例

2. 長単位解析

2.1. チャンキングモデルと後処理に基づく長単位解析

長単位は、短単位列を入力とし、以下に述べるチャンキングモデルと後処理に基づく手法により認定する。長単位を認定するという問題は、長単位を構成する短単位のそれぞれに下記の四つのラベルのうちいずれかを付与する問題に置き換えることができる。これらのラベルの尤もらしさを推定するモデルを**チャンキングモデル**と呼ぶ。これは Uchimoto らの方法（Uchimoto & Isahara, 2007）におけるラベルの定義を次のように改良したものである。

† kozawa@el.itc.nagoya-u.ac.jp

短単位							ラベル	長単位					
書字形	語彙素読み	語彙素	発音形	品詞	活用型	活用形		書字形	語彙素読み	語彙素	品詞	活用型	活用形
固有	コユウ	固有	コユー	名詞-普通名詞-形状詞可能			B	固有名詞仮名表記	コユウメイシカナヒョウキ	固有名詞仮名表記	名詞-普通名詞-一般		
名詞	メイシ	名詞	メイシ	名詞-普通名詞-一般			I						
仮名	カナ	仮名	カナ	名詞-普通名詞-一般			I						
表記	ヒョウキ	表記	ヒョーキ	名詞-普通名詞-サ変可能			I						
に	ニ	に	ニ	助詞-格助詞			B	に関して	ニカンシテ	に関して	助詞-格助詞		
関し	カンスル	関する	カンシ	動詞-一般	サ行変格	連用形-一般	I						
て	テ	て	テ	助詞-接続助詞			I						
論文	ロンブン	論文	ロンブン	名詞-普通名詞-一般			Ba	論文	ロンブン	論文	名詞-普通名詞-一般		
を	ヲ	を	ヲ	助詞-格助詞			Ba	を	ヲ	を	助動-格助詞		
三	サン	三	サン	名詞-数詞			B	三本	サンホン	三本	名詞-数詞		
本	ホン	本	ホン	接尾辞-名詞的-助数詞			I						
執筆	シッピツ	執筆	シッピツ	名詞-普通名詞-サ変可能			B	執筆し	シッピツスル	執筆為る	動詞-一般	サ行変格	連用形-一般
し	シ	為る	スル	動詞-非自立可能	サ行変格	連用形-一般	I						
た	タ	た	タ	助動詞	助動詞-タ	終止形-一般	Ba	た	タ	た	助動詞	助動詞-タ	終止形-一般
。		。		補助記号-句点			Ba	。		。	補助記号-句点		

図 2：短単位と長単位の例

Ba　1短単位のみで長単位を構成し、かつ、その品詞、活用型、活用形が長単位のものと一致する。

Ia　長単位を構成する短単位のうち末尾の要素で、かつ、その品詞、活用型、活用形が

長単位のものと一致する。

B　長単位を構成する短単位のうち先頭の要素で、かつ、その品詞、活用型、活用形のいずれかが長単位のものと一致しない。

I　長単位を構成する短単位のうち先頭以外の要素で、かつ、その品詞、活用型、活用形のいずれかが長単位のものと一致しない。

これは、長単位を構成する末尾の短単位の品詞、活用型、活用形が長単位のものと一致する場合に付与されるラベルは「Ba」か「Ia」、そうでない場合は、長単位を構成する先頭の要素に付与されるラベルは「B」、長単位を構成する先頭以外の要素に付与されるラベルは「I」であることを意味する。したがって、このモデルにより、単位境界だけでなく、品詞、活用型、活用形の情報も得られる。例えば、図 2 の短単位には、「ラベル」の列に示されるようなラベルが付与される。これらのラベルを正しく推定できれば、「Ba」あるいは「Ia」が付与された短単位から品詞、活用型、活用形が得られる。図 2 は、「論文」などの長単位については品詞、活用型、活用形も得られることを表わしている。

チャンキングモデルの素性としては、着目する短単位とその前後 2 短単位、あわせて 5 短単位について、以下の情報を利用する。

- 書字形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形、語種
- 階層化された素性（例えば「名詞-普通名詞-一般」）に対して、上位階層で汎化した素性（例えば「名詞」「名詞-普通名詞」）を利用する。
- 句読点や中点などの補助記号の場合、前後 1 短単位の品詞を「名詞-普通名詞-一般」「名詞-普通名詞-サ変可能」「名詞-普通名詞-副詞可能」「名詞-数詞」「その他」の 5 クラスにまとめ、素性として利用する。
- 辞書素性

  学習データ中の長単位から、2 短単位以上からなる助詞・助動詞について、前接する短単位の品詞・活用型・活用形と後接する短単位の品詞を含め（例えば、図 3)、辞書を作成した。辞書に含まれる長単位を構成する短単位列か否かの情報を素性として利用する。

この他、BCCWJ では①などの丸付き数字では長単位境界が区切れるため、丸付き数字か否かに関する素性も利用している。

一方、「執筆為る」などの長単位については品詞がこれらを構成する短単位「執筆」「為る」のどちらとも異なるため、各短単位には「B」あるいは「I」のラベルしか付与されない。この場合は、ラベルを正しく推定できたとしても品詞は得られず、単位境界の情報のみが得られることになるため、その長単位に対し、後処理として次に述べる品詞推定モデル及び活用型・活用形推定モデルを適用することにより、最も尤もらしい品詞、活用型、活用形を推定する。品詞推定モデルは、長単位を構成する短単位列が与えられると、学習データに現れた品詞を候補としてその品詞候補すべてについて尤もらしさを計算するモデルである。

			動詞	サ行変格	連用形
て	テ	て	助詞-接続助詞		
いる	イル	居る	動詞-非自立可能	上一段-ア行	終止形-一般
			助詞		

図 3：助動詞辞書の要素例

ただし、助詞と助動詞については長単位を構成する短単位列が複合辞と一致している場合のみ品詞候補とし、それ以外の場合には、助詞と助動詞を除くすべての品詞候補から最尤の品詞を出力する。複合辞と一致しているかどうかは、予め用意した複合辞辞書との文字列マッチングにより自動判定する。素性としては、着目している長単位とその前後の長単位、あわせて 3 長単位について、各長単位を構成する短単位の情報を用いる。具体的には、各長単位を構成する短単位について、先頭から 2 短単位と末尾から 2 短単位に着目し、各短単位に関する書字形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形、及び、階層化された素性に対して上位階層で汎化した情報を素性として用いる。長単位が 1 短単位からなる場合は、先頭から 2 短単位目の情報は与えられなかったもの（NULL）として扱う。例えば、図 2 の「執筆し」では、「三本」「執筆し」「た」の 3 長単位に対し、「三|本|執筆|し|た|NULL」（先頭から各 2 短単位）、及び、「NULL|た|し|執筆|本|三」（末尾から各 2 短単位）の各短単位に関する情報を素性として用いる。活用型推定モデル、及び、活用形推定モデルは、推定するカテゴリが品詞ではなくそれぞれ活用型、活用形となる点、及び、動的素性を用いる点を除いて品詞推定モデルと同様である。動的素性としては、活用型推定モデルでは着目している長単位の品詞（自動解析時は品詞推定モデルにより自動推定した品詞）を、活用形推定モデルでは着目している長単位の品詞と活用型（自動解析時は品詞推定モデル、活用型推定モデルによりそれぞれ自動推定した品詞と活用型）を用いる。

長単位の語彙素読み・語彙素表記は、基本的に短単位の語形と語形代表表記をそれぞれ結合することで生成する。ただし、語彙素読みでは長単位末尾の短単位が活用語の場合は語形基本形を結合し、語彙素表記では長単位の末尾以外の短単位が活用語の場合は語形代表表記出現形を結合する。また、短単位の品詞が人名・地名になっている部分は短単位書字形を用いる。複合辞辞書に登録されている複合辞については、辞書引きにより読みと表記の情報を得る。

2.2. 実験と考察

2.1 節に述べた手法を用いて実験を行った。チャンキングモデルの学習と適用には、Yamcha と CRF++、MMA を用いた。Yamcha は SVM に基づく汎用チャンカーであり、カーネルは多項式カーネル（べき指数 3）を採用した。解析方向は文末側から文頭側とし、多クラスへの拡張は one-versus-rest 法を用いた。CRF++は CRF に基づく汎用チャンカーであり、MMA（Kruengkrai et al., 2009）は MIRA に基づく形態素解析システムである。後処理には SVM を用いた。BCCWJ の白書・書籍・新聞・雑誌・Web（Yahoo! 知恵袋）コアデータのうち、27,610 文（白書 : 5,216 文/205,150 短単位、書籍 : 8,288 文/212,878 短単位、新聞 : 14,101 文/326,402 短単位、雑誌 : 10,800 文/218,636 短単位、Web : 5,725 文/99,917 短単位）でモデ

ルを学習し、3,069 文（白書 : 580 文/23,127 短単位、書籍 : 921 文/21,656 短単位、新聞 : 1567 文/34,425 短単位、雑誌 : 1,200 文/26,911 短単位、Web : 637 文/10,835 短単位）で評価した。

表 1 に長単位解析の解析精度[1]を示す。白書・書籍・新聞・雑誌・Webのいずれに対しても、98%超の正解率となっている。また、CRF・MIRAのいずれのモデルを用いた場合でも、語彙素認定において 98%を超える精度が得られている。

表 1：長単位解析システムの解析精度

モデル		白書	書籍	新聞	雑誌	Web	全て
CRF	境界認定	99.3	99.0	98.9	98.7	98.4	98.9
	品詞認定	99.1	98.8	98.6	98.4	98.3	98.6
	語彙素認定	99.0	98.6	98.6	98.4	98.3	98.6
MIRA	境界認定	99.3	98.9	98.9	98.7	98.5	98.9
	品詞認定	99.0	98.7	98.5	98.4	98.4	98.6
	語彙素認定	99.0	98.6	98.5	98.3	98.4	98.5

3. 中単位解析

中単位は語の内部構造に従った単位であり、長単位を超えない範囲で、直接的な係り受け関係を持つ、隣接する短単位同士を結合したものとして定義できる。中単位は、長単位を入力とし、以下に述べる短単位間の係り受け解析と中単位境界同定ルールにより認定する。例えば、図 4 の 4 短単位から構成される長単位「固有名詞仮名表記」では、「固有名詞」と「仮名表記」の 2 つの中単位が生成される。

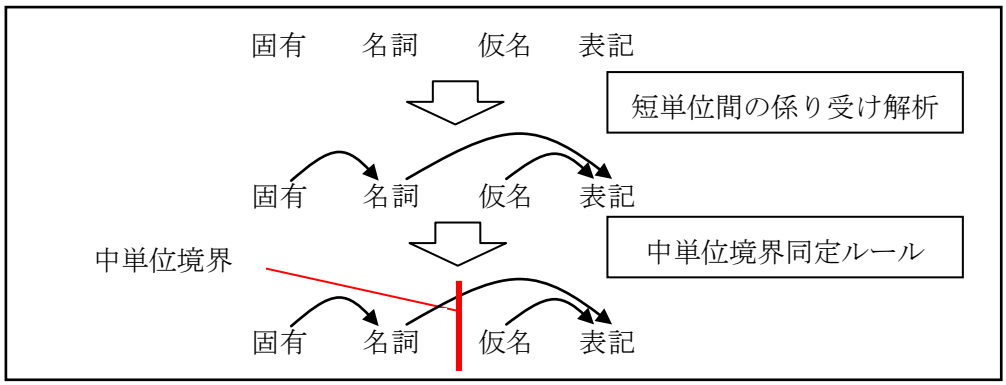


図 4：中単位解析の例

3.1. 短単位間の係り受け解析

最大全域木に基づく依存構造解析手法（McDonald et al., 2005）を短単位間の係り受け解析に適用した（Uchimoto & Den, 2008）。学習および解析には MST Parser を用いた。first order

[1] SVM の学習に時間がかかったため、予稿集には間に合いませんでした。最新の結果はポスター発表をご参照ください。

の素性を採用した。短単位間の係り受け解析の素性には、以下の情報を利用する。

- 書字形出現形、語彙素表記、品詞、活用型、活用形
- 階層化された素性に対して、上位階層で汎化した素性

BCCWJ の白書・書籍・新聞コアデータのうち、6,547 文（白書：2,068 文/81,867 短単位、書籍：3,308 文/82,175 短単位、新聞：1,171 文/29,161 短単位）に対して、短単位間の係り受け情報と中単位情報を付与したデータを対象に学習、評価した。学習および評価は 10 分割交差検定により行った。このときの係り受け解析精度を表 2 に示す。白書・新聞に対しては約 98%、書籍に対しては 99%超の正解率となっている。ただし、長単位を構成する短単位数が増加するにつれて精度が低下するため、長い長単位について今後改善が必要である。

表 2：短単位間の係り受け解析精度

長単位を構成する短単位数	白書	書籍	新聞	全て
2 短単位以上	97.9	99.4	97.7	98.3
3 短単位以上	97.9	97.7	94.9	96.3
4 短単位以上	94.0	96.2	92.4	94.0
5 短単位以上	92.3	95.7	91.5	92.5

3.2. 中単位境界同定ルールによる中単位の認定

短単位間の係り受け情報に基づく中単位境界同定ルールにより、中単位境界を認定する。以下にルールを示す。

1. 長単位を超えない範囲で、順次係り受けの語は繋げる。
2. 語をまたいだ係り受けは区切る。
3. 補助記号は前後の形態素と区切る。

ただし、例外として以下のルールを設ける。

- 長単位の品詞が形状詞の場合
  - 語をまたいだ係り受けの場合も区切らず、一つの中単位とする。
- 長単位の品詞が名詞の場合
  - 短単位が以下の接頭辞の場合は区切る。
    各、計、現、全、非、約
  - 係り受けが並列の場合、並列の形態素同士は区切る。但し、並列の形態素の品詞が接頭辞の場合は、区切らない。
  - 後続する短単位列が「名詞+接尾辞」であり、接尾辞に係る場合は区切らない。
  - 後続する短単位列が「接頭辞+名詞」であり、名詞に係る場合は区切らない。

3.2 節で述べた短単位間の係り受け情報を自動付与したデータを対象に、中単位認定を行った。表 3 に中単位解析の解析性能を示す。白書・書籍・新聞のいずれに対しても、F 値で 98％～99％となっている。しかし、白書と新聞では長単位を構成する短単位数の増加に伴

う性能低下が著しい。これは次に述べる理由から短単位間の係り受け解析による影響と考えられる。

表 3：中単位解析の解析性能（F 値）

長単位を構成する短単位数	白書	書籍	新聞	全て
全て	98.5	99.8	98.9	99.2
2 短単位以上	95.2	99.2	96.3	96.6
3 短単位以上	91.3	97.0	91.6	92.3
4 短単位以上	83.1	95.6	84.2	85.0
5 短単位以上	77.2	95.8	85.3	81.7

中単位境界同定ルールの性能を調べるため、正解の係り受け情報を用いて中単位境界解析を行った。その結果、長単位を構成する短単位数がいずれの場合であっても性能は 99%を超えており、中単位境界同定ルールは十分な性能を保持していることが分かった。このことから、中単位境界解析の性能を上げるには、短単位間係り受け解析の性能を上げる必要があると言える。

4. 中・長単位解析ツール Comainu

2 章と 3 章で説明した手法を実装することにより、中・長単位解析ツール Comainu を作成した。本ツールは以下の機能を持つ。

- 長単位解析
  平文または短単位列を入力すると、長単位を付与した短単位列を出力することができる。平文が入力された場合、Chasen もしくは Mecab により形態素解析を行った後に長単位解析を行う。長単位解析のチャンキングモデルには SVM と CRF、MIRA のいずれかを用いることができる。
- 中単位境界解析
  平文または短単位列もしくは長単位情報を付与された短単位列を入力すると、中・長単位を付与した短単位列を出力することができる。平文が入力された場合には形態素解析と長単位解析、短単位列が入力された場合には長単位解析を行った後に中単位境界解析を行う。
- 文節境界解析
  平文または短単位列を入力すると、文節境界を付与した短単位列を出力することができる。平文が入力された場合、形態素解析を行った後に文節境界解析を行う。

平文や短単位列の直接入力だけでなくファイル入力にも対応している。解析結果をファイルに保存することも可能である。

図 6 に Comainu による中・長単位解析の実行例を示す。図 6 の例では、短単位列を入力とし、MIRA を用いて学習したチャンキングモデルによる長単位解析及び中単位解析を実

行し、中・長単位が付与された短単位列を出力している。出力の 29～34 列目はそれぞれ長単位の書字形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形を表し、出力の 35 列目は短単位間の係り受け情報、36 列目は中単位番号、37 列目は中単位の書字形出現形を表す。

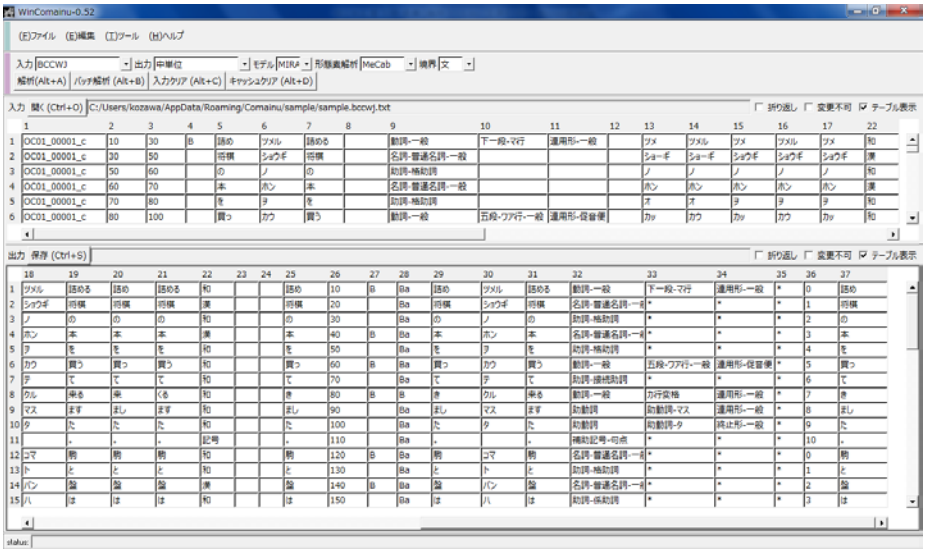


図 6： Comainu による中・長単位解析の実行例

5. まとめ

本稿では、チャンキングモデル及び後処理に基づく長単位解析手法、及び、短単位間係り受け解析及び中単位境界同定ルールに基づく中単位解析手法について述べ、それらを実装したシステムである中・長単位解析ツール Comainu を紹介した。BCCWJ コアを用いた実験では、境界認定 98.9%、品詞認定 98.6%、語彙素認定 98.6%の長単位解析精度、及び、99.2%の中単位解析精度を得た。今後、本ツールを公開する予定である。

文献

Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., & Isahara, H. (2009). An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. In *Proc. of ACL-IJCNLP* 2009.

McDonald, R., Crammer, K., & Pereira, F. (2005). “Online Large-Margin Training of Dependency Parsers”, *In Proc. of ACL 2005*, pp. 91-98.

Uchimoto, K. & Den, Y. (2008). Word-level Dependency-structure Annotation to Corpus of Spontaneous Japanese and Its Application, In *Proc. of LREC* 2008.

Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. In *Proc. of the 20th IJCA I*, pp. 1731-1737.

UniDic を用いた音声認識用言語モデルの作成

山田　篤（電子化辞書班連携研究者：京都高度技術研究所研究部）†

Constructing Language Model for Speech Recognition using UniDic

Atsushi Yamada (Research Div., ASTEM RI/Kyoto)

１．はじめに

大語彙連続音声認識では、統計的言語モデルを利用して大語彙の音声認識を行う。言語モデルの構築にはCMU-Cambridge統計的言語モデルツールキットやPalmkit[1]等を用いるが、この際に大量の学習用テキストが必要になる。ここで用いる学習用テキストは、語が空白で区切られ、クラス名が付加された以下のような形式のものである。

言語+名詞　モデル+名詞　を+助詞　構築+名詞　する+動詞　。+記号

このような学習用テキストを統計処理し、高頻度語彙リストを構築し、それらの語彙に制限した言語モデルと認識用辞書を構築するため、学習用テキストには、認識対象として必要な語彙が含まれていなければならない。このとき、語の単位として斉一性が保たれていることが重要である。また、認識用辞書を構築するためには、語の読み（発音）が正しく得られることも必要となる。

本稿では、UniDic 及びその関連ツールを用いて、言語モデル構築のための学習用テキストを作成する方法について報告する。

２．学習用テキストにかかる要件

学習用テキストにかかる第一の要件として、語の単位の斉一性がある。特定の語が、ある場合は切り出されたり、別の語の一部に含まれたりすると、出現頻度の計量や N-gram の学習に悪影響を及ぼす。

第二の要件として、表記が統一されていることがあげられる。この代表的なものに数字の取り扱いがある。「２０１１年」と「二〇一一年」を別々に数えてしまうと、語彙のカバー率が低下する。また、認識用辞書との関連では、これは「ニセンジューイチネン」という読み（発音）と対応していてほしい。「蓋然」を漢字で表記するか、かなで表記するかの違いも、これらが学習用テキスト内で混在していると、カバー率が低下する原因となる。

第三の要件として、可能な読み（発音）との対応がある。「日本」に対し「ニホン」という読みしかふられていないとすると、「ニッポン」という発音に対しては認識できなくなる。それぞれの読みが別々にふられていても不十分で、どちらの読みも可能であることが示されている必要がある。

３．UniDic を用いた対応

第一の要件に対しては、UniDic が斉一性のある短単位に基づき構築されているため、形態素解析用の辞書として UniDic を用いることで解決できる。

第二の要件のうち、表記の揺れの問題については、UniDic において書字形が異なる同一

† yamada@astem.or.jp

1 http://palmkit.sourceforge.net/

の語彙素として表現されているものについては、書字形ではなく語彙素を見ることにより、その同一性が取得できる。たとえば「蓋然性」を解析すると、

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
蓋然	ガイゼン	ガイゼン	蓋然	*	名詞-普通名詞-一般	*	*
性	セー	セイ	性	*	接尾辞-名詞的-一般	*	*

「がい然性」を解析すると、

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
がい然	ガイゼン	ガイゼン	蓋然	*	名詞-普通名詞-一般	*	*
性	セー	セイ	性	*	接尾辞-名詞的-一般	*	*

となり、これらが同一の語彙素であることがわかる。

数字の取り扱いについては、UniDic のみでは解決ができない。たとえば UniDic で「２０１１年」を解析すると、

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
２	ニ	ニ	二	*	名詞-数詞	*	*
０	レー	レイ	零	*	名詞-数詞	*	*
１	イチ	イチ	一	*	名詞-数詞	*	*
１	イチ	イチ	一	*	名詞-数詞	*	*
年	ネン	ネン	年	*	接尾辞-名詞的-助数詞	*	*

「二〇一一年」を解析すると、

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
二	ニ	ニ	二	*	名詞-数詞	*	*
〇	レー	レイ	零	*	名詞-数詞	*	*
一	イチ	イチ	一	*	名詞-数詞	*	*
一	イチ	イチ	一	*	名詞-数詞	*	*
年	ネン	ネン	年	*	接尾辞-名詞的-助数詞	*	*

となり、「２」と「二」が同一の語彙素「二」の異なる書字形であることは得られるが、「ニセンジューイチネン」という読みとは対応がつかない。このためには、数字列の前処理ツールである NumTrans を用いて、「２０１１年」を「二千十一年」に書き換えてから、形態素解析にかける必要がある。

第三の要件については、UniDicでは同一語に対する発音形の違いとして表現されるため、1 つの語に対して複数の発音形が存在していれば、それらの読み方が可能であることがわかる。ただし、UniDicを用いた形態素解析結果では、いずれか 1 つの発音形しか出力されないため[2]、何らかの後処理が必要となる。

UniDic 短単位辞書では、書字形出現形、発音形出現形、語彙素読み、語彙素表記、語彙

[2] これまで ChaSen 版の UniDic では発音形の併記出力を行っていたが、これは廃止される予定である。

素細分類、品詞、活用型、活用形の 8 つが基本 8 属性となり、辞書中のエントリはこれら 8 属性によって一意に定まる。たとえば、助数詞「本」の基本 8 属性は

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
本	ホン	ホン	本	*	接尾辞-名詞的-助数詞	*	*

のようになる。このうち、発音形出現形を除く 7 属性をキーとして辞書検索を行い、発音形出現形を再取得すると、次のようになる。

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
本	ホン/ボン/ポン	ホン	本	*	接尾辞-名詞的-助数詞	*	*

この場合、3 つのエントリがみつかるため、発音形出現形に 3 つの値が併記されている。

ただし、「今日」に対する「キョー」「コンニチ」のように語彙素読みが異なる場合もある。

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
今日	キョー	キョウ	今日	*	名詞-普通名詞-副詞可能	*	*

書字形出現形	発音形出現形	語彙素読み	語彙素表記	語彙素細分類	品詞	活用型	活用形
今日	コンニチ	コンニチ	今日	*	名詞-普通名詞-副詞可能	*	*

これを取得するためには、発音形出現形、語彙素読みを除く 6 属性で辞書検索を行う必要がある。さらに、固有名詞と普通名詞等で品詞が異なる場合、活用型が異なる場合等も考えられるが、現実的には 6 属性までにとどめるべきであろう。

４．UniDic を用いた処理手順

以上をまとめると、UniDic を用いた言語モデル構築のための学習用テキストの作成手順は以下のようになる。

1. 数字列の前処理
2. UniDic を用いた形態素解析
3. 読み併記のための後処理

数字列の前処理には NumTrans を用いる。なお、現在の NumTrans は置換前の元のテキストも出力するために XML 形式の出力を採用しているため、置換テキストのみを出力するモードを新たに追加した。

形態素解析には MeCab 版の UniDic を用いる。前処理で出力されたテキストをそのまま MeCab で処理する。このために、入力をあらかじめ 1 文 1 行の形式にしておくことが望ましい。

後処理には、汎用後処理ツールで採用されているSQLite3[3]形式の辞書を用いることとし、汎用後処理ツールの辞書引き部分のみを抜き出したツールを新たに作成した。本ツールで設定できる項目は以下のとおりである。

・dicdb:　# SQLite3 で作成した辞書ファイル

[3] http://www.sqlite.org/download.html

・table:　# SQLite3 のテーブル名
・input:　# 入力フォーマットを表す属性列
・output:　# 出力フォーマットを表す属性列
・sql:　# input のうち辞書検索に用いる属性列

基本的には、input, output は基本 8 属性、sql は発音形出現形、語彙素読みを除いた 6 属性という使い方をするが、常に語彙素読みを除くと、品詞によっては、大量の発音形が取得されてしまう可能性がある。選択的に sql を変更する機能については、その必要性も含めて、現在検討中である。

最後に、後処理結果を変換して、語が空白で区切られ、クラス名が付加された学習用テキストを得る。辞書引きツールの output で区切り記号を変更可能にし、この形式を指定できるようにすれば、この変換は不要になる。このときに、書字形ではなく、語彙素を語として出力すれば、表記の揺れも統一できる。

なお、従来の言語モデル構築では、大量のデータを扱うために、少しでも容量を圧縮する必要から、品詞名を ID 化する等の措置をとっていた。今回は形態素解析後に辞書引きを行うため、MeCab の品詞 ID 出力マクロ (%h) は使えないので、もしも品詞名の ID 化を行うとすれば、品詞名-ID 変換用の後処理ツールを作るか、別途辞書ファイルを用意して、辞書引きツールに出力させる方法が考えられる。

５．おわりに

本報告では、UniDic を用いた大語彙連続音声認識用言語モデルの構築方法について述べた。言語モデルの構築のためには、大量の学習用テキストが必要となるため、大量のテキストデータを自動で解析し、必要な情報を付与する仕組みが必要となる。UniDic 及びその関連ツールを用いることで、この作業が容易になることを期待している。このとき、大量のデータを必要とするため、それまでに解析したテキストと新たに解析したもので、語の認定基準等が一致していることが重要である。BCCWJ や日本語話し言葉コーパス (CSJ) を学習用テキストとして用いることも考えられるが、CSJ と BCCWJ の間の短単位の仕様変更にどのように対応すればよいかは課題として残されている。

文献

P.R. Clarkson and R. Rosenfeld. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. *Proceedings ESCA Eurospeech*.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵. (2007). コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用. 日本語科学、*22* (pp.101-123).

山田篤. (2007). 数字列への読み付与---NumTrans と ChaOne---. 特定領域研究「日本語コーパス」平成 19 年度全体会議予稿集 (pp. 85-90).

山田篤・伝康晴. (2010). UniDic 汎用後処理ツールの設計と実装. 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.23-28).

作文コーパスからみる生徒の使用語彙

鈴木一史　（言語政策班協力者：東京大学教育学部附属中等教育学校）†
棚橋 尚子（言語政策班分担者：奈良教育大学教育学部）
河内昭浩（言語政策班協力者：群馬県立館林高等学校）

Students' Vocabulary in the"Sakubun *Corpus*"

Suzuki Kazufuni　(The University of Tokyo Secondary School attached the Faculty of Education)
Tanahashi Hisako(Faculty of Education,Nara University of Education)
Kawauchi Akihiro（Tatebayashi high-school）

１．はじめに

本稿は、特定領域研究「日本語コーパス」言語政策班報告書（JC-P-10-01）「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」の一部である。その中の鈴木（2011）の報告をポスター発表するための資料として、再掲したものである。

２．調査データ

本校（東京大学教育学部附属中等教育学校）は中等教育学校であるために、生徒は中学 1 年（1 年生）から高校 3 年（6 年生）まで、一貫した教育がおこなわれている。1 学年は、1 クラス 40 人で 3 クラスあり、120 人が在籍している。また、男女の比率は 50:50 である。国語の教員は 6 人いて、担当分野や担当学年が固定しているわけではない。2-2-2-制のカリキュラムを取っているため、2 年ごとに担任が変わる。国語科すべての教員が 6 年間で 1 度は 1 人の生徒と接するため､特定の学年が特定の教員の影響を強く受けることが少ない。このような状況の中で、同時期に共通の課題で作文を書かせデータベース化することは、各学年の特徴をとらえるには最適であると考え、この調査を行った。

調査対象は中学 1 年から高校 5 年。調査は、平成 21 年 1 月に、冬休みの宿題として課した。課題は、「年末・年始の行事をふまえて、文化について述べよ」である。原稿用紙を配付し、字数は 800 字程度とした。手書きで提出させた。

時間を区切った提出物ではないので、他文献やネットからの引用も考えられ、分析時に考慮すべき課題である。しかし、手書き提出にしたために、調べたとしても最終的に自分で書きなおす必要があり、その点では生徒自身の言葉であるととらえてよいだろう。

手書き作文の解析方法は以下の手順で行った。手書き文字を電子データ化した。文章表現上どのような間違いを犯しやすいか分析するため、表記ミスなどもそのまま入力。次に全文検索ソフトで検索できるように、タグ付けを行う。解析ソフトは「茶まめ」と形態素解析辞書 Unidic1.3.12 を使用した。これにより、文章を単語に区切り、品詞情報、語種情報が付与される。

5 年生の作文データ分析に、レベルと初出学年、特徴教科のデータを関連付けると表 1 のような語彙表ができる。これは 5 年生のデータの一部分である。「年末・年始…」という課題から、「正月」という語が何度も使われていることが分かる。

† suzuki-j@hs.p-u-tokyo.ac.jp

表１　生徒使用語彙表例

語彙素読み	語彙素	語彙素のカウント	品詞	語種	レベル_LB_FL	初出学年	特徴教科
ジョウ	城	1	接尾辞-名詞的-一般	漢	a	小後	
ジョウ	場	2	接尾辞-名詞的-一般	漢	a	小前	保
ジョウ	状	146	接尾辞-名詞的-一般	漢	a	小前	理芸
ショウカイ	紹介	3	名詞-普通名詞-サ変可能	漢	a	小前	国外
ショウガイ	障害	1	名詞-普通名詞-サ変可能	漢	a	小後	技保
ショウガク	小学	3	名詞-普通名詞-一般	漢	a	小前	
ショウガツ	正月	160	名詞-普通名詞-一般	漢	a	小前	
ショウギョウ	商業	4	名詞-普通名詞-一般	漢	a	小後	社
ジョウキョウ	状況	3	名詞-普通名詞-一般	漢	a	中	保情
ショウゲキ	衝撃	1	名詞-普通名詞-一般	漢	a	中	
ジョウケン	条件	1	名詞-普通名詞-一般	漢	a	小後	数理社技保情
ショウコ	証拠	1	名詞-普通名詞-一般	漢	a	小前	
ショウゴ	正午	1	名詞-普通名詞-副詞可能	漢	b	小前	
ショウシ	少子	3	名詞-普通名詞-一般	漢	c	中	社技

３．作文コーパスの分析

３．１　データ概要

作文コーパスの全データは以下のとおりである。「1 年」が 1 年生全体の作文データであり、以下同様に学年が上がっていく。「ALL」は全ての作文の合計である。学年により人数に異なりがあるのは、宿題として課したために、提出した人数による。人数にばらつきがあるために、これからのデータ分析は主に「異なり語数」によって行う。

表２　作文コーパス概要

	人数	延べ語数	一人当たり	異なり語数	異なり語数/延べ語数	句点の数	一文の平均語数
1年	118	64,285	544.8	3310	5.15%	2729	23.6
2年	103	54,745	531.5	3488	6.37%	2412	22.7
3年	66	37,205	563.7	2765	7.43%	1183	31.4
4年	84	45,976	547.3	3396	7.39%	1859	24.7
5年	81	47,707	589.0	3559	7.46%	1964	24.3
ALL	452	249,918	552.9	7389	2.96%	10147	24.6

一人あたりの語数は、531.5 語から 589.0 語であり、全体では 552.9 語である。800 字程度の作文を課し、おおよそ 550 語程度で作文を仕上げてきたことになる。

異なり語数は、2765 語から 3559 語である。3 年生は人数も少なく、また学年の特質も影響していると思われるために、以下の分析データでも 3 年生のみ特殊な状態が生じているように見える。1 学年当たりおおよそ 3000 語強の語彙で文章を書いていることがうかがえる。しかし、全体をみると 7000 語を超えている。これは、学年によって重なっている語があるために単純な合計ではない。しかし逆に、学年別の分析よりはるかに多いということは、多様な語彙を感じさせる。

次に、異なり語数を延べ語数で割ることによって、語彙の広がりを調べる。1 年生では 5.15%であるが、徐々に数値が上がり、5 年では 7.46%である。同じ年末年始について書く

場合でも、1 年生はお年玉など発想が同様になる傾向があるが、学年が上がると、多様な語彙を使って年末年始について表現していることがうかがえる。語彙の多様さは発想の多様さにもつながり、同じ年末年始の過ごし方でも、捉え方が多様になってきていると考えられる。

一文の長さは、20 字強。学年が上がると、若干長くなっているが、それほど大きな差はない。つまり、一文の長さが文章の巧拙に影響するわけではないようである。

3．2　語種分析

作文コーパスの語彙を異なり語数で学年・語種別にカウントしたものが以下の表 2 である。学年が上がるに従って抽象的な言葉、つまり漢語を多く使うようになっているのではないかということである。

表 3　作文コーパス語種分析

語種	和語	漢語	外来語	記号	固有	不明	混成
1年	50.4%	36.4%	4.4%	1.3%	3.7%	0.5%	2.8%
2年	49.9%	37.1%	3.8%	1.2%	4.6%	0.5%	2.4%
3年	51.1%	37.4%	4.3%	1.2%	3.0%	0.4%	2.4%
4年	47.7%	39.4%	4.7%	1.1%	3.5%	0.4%	2.8%
5年	46.5%	41.2%	4.5%	1.2%	3.3%	0.3%	2.8%
ALL	44.1%	40.0%	5.9%	1.1%	5.0%	0.4%	2.7%

表 2 から、和語の割合がかなり減ってきているのに対し、漢語の割合は増えているように思う。そこで、抽象的な言葉の比率を調べるために、「和語」と「漢語・外来語」について、グラフ化した。

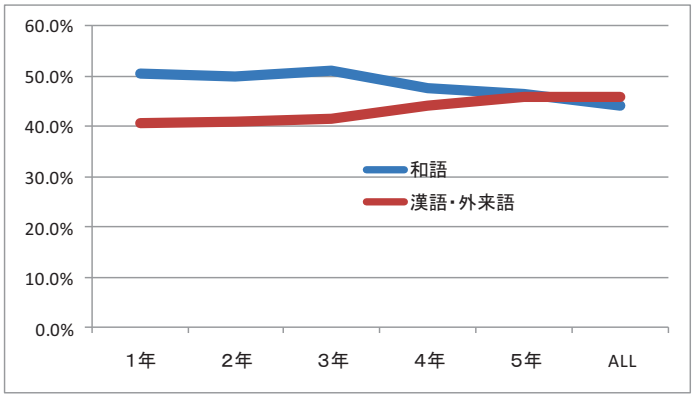


図 1　和語と漢語・外来語の使用変化

4 年生までは和語の使用率が高いが、5 年生で和語と漢語・外来語の比率がほぼ同じになる。そして、1 年から 5 年まで徐々に和語の比率が下がっていき、反対に漢語・外来語の比率は上がっている。ここから、学年が上がるにつれて段々と抽象的な言葉や概念を表す言葉が増えていることがうかがえる。

ALL では完全に逆転している。つまり、各学年で和語については重なる言葉が多く、漢語については、異なる言葉を使っているということである。たとえば、「私」や「降る」などはどの学年でも使うが、「酒宴」や「儀礼」などは、個別に使う生徒がいるということである。

３．３　レベル分析

作文コーパスの語彙表と「学校・社会対照語彙表」を関係づけ、レベルデータを付与し、それぞれについて異なり語数の数をカウントしたものが表 4 であり、それをグラフ化したものが、図 2 である。

表 4　学年別レベル分け

レベル	a	b	c	d	e
1年	48.0%	18.6%	13.4%	8.1%	11.9%
2年	47.5%	18.8%	13.6%	7.7%	12.5%
3年	50.7%	18.0%	13.5%	7.0%	10.8%
4年	48.1%	18.9%	13.5%	7.6%	11.9%
5年	48.9%	18.3%	13.5%	7.3%	12.0%
ALL	36.1%	22.1%	17.4%	9.6%	14.8%

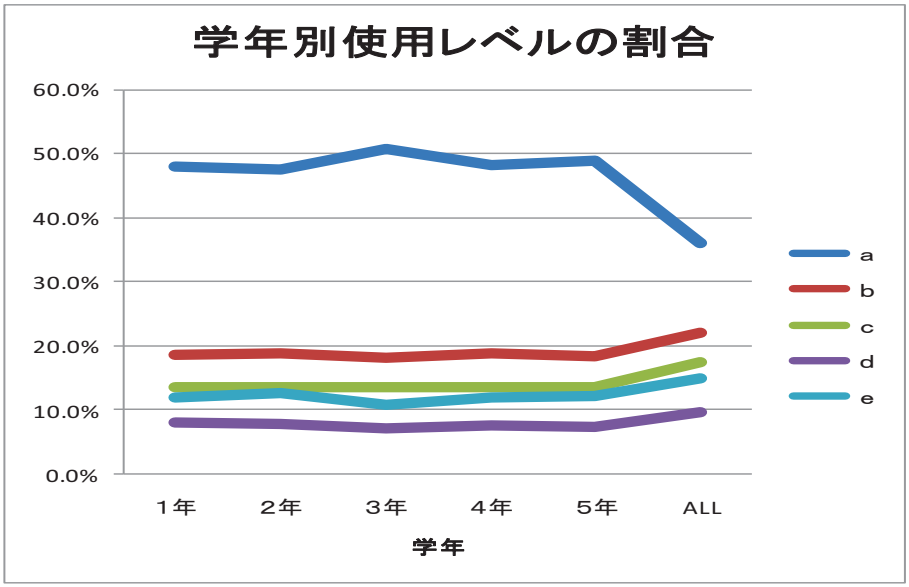


図 2　学年別使用レベルの割合

全ての学年でａが 50%近く使われ、文章を成り立たせている。ａからｅの使用率は 1 年から 5 年まで、ほとんど同じである。文章の特性として、ａが減るから文章の発達が見られ、ｄｅを使用すると高度な文章であるとはいえないようである。

しかし、全体（ALL）を分析すると、ａの数値が格段に下がり、その他の全てのレベルが確実に上昇している。これは、ａという基礎的な言葉は、全ての学年で同様に出現するために､1年から5年までほとんどａの語彙は増えていかないことが示されている｡つまり、学年を問わず重複語彙が多いということである。その反面、レベルｂｃｄｅの語彙は、重複はするものの、いろいろな言葉を使っていることが分かる。全学年で同じテーマで作文を書かせていることを考えると、同じテーマや題材に対しても、様々な語彙で考えて書いていることがうかがえる。これは、異なり語数が、1 年から 5 年までほとんど同じ数であったものが、前学年で見ると倍以上になっていることと合わせて考えると、異なり語数の伸長は主にレベルｂｃｄｅを中心に起こっていることが分かる。また、語種変化から考えると、その語彙は漢語や外来語のレベルの高い語が使用語彙として増えているようである。

３．４　初出学年分析

次に、作文コーパスの使用語彙がいつ学習したものであるか、すべきものであるかについて、初出学年のデータを踏まえて考察する。

表５　生徒使用語彙の初出学年

初出学年	小学校前期	小学校後期	中学校	高校
1年	42.9%	28.2%	18.6%	10.3%
2年	44.1%	27.3%	18.4%	10.2%
3年	45.7%	26.7%	17.9%	9.6%
4年	41.3%	27.4%	20.2%	11.1%
5年	39.7%	27.5%	20.8%	11.9%
ALL	32.4%	29.2%	24.1%	14.3%

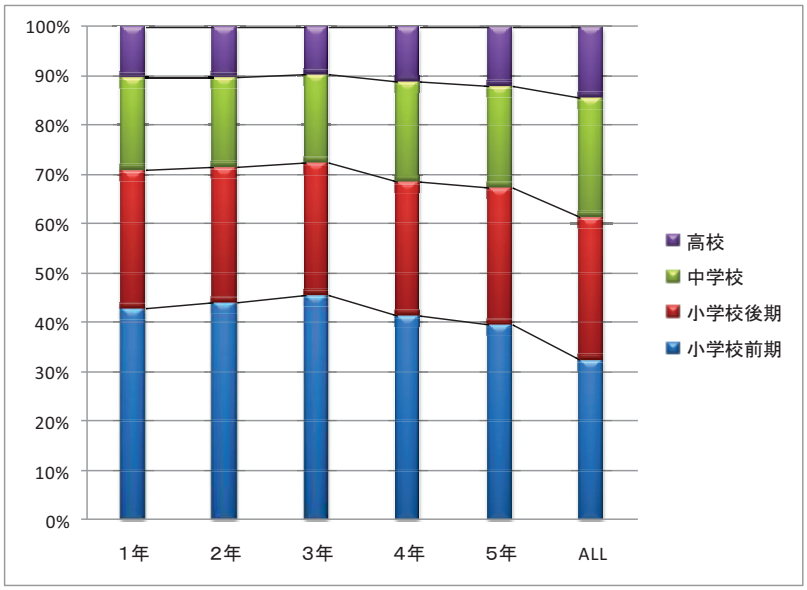


図３　生徒使用語彙の初出学年推移

学年を追ってみてみると、小学校前期で習う言葉は 1 年から 3 年まで上昇し、その後低くなっている。これは教科書に出てきた言葉が、自分の使用語彙になるまでにはしばらくかかることを意味している。しかし、小学校前期の語彙は、基礎的な語彙であり使用頻度の高い語彙であると考えられる。全学年データで小学校前期語彙は 2312 語、そのうちレベル a は 1592 語、約 70%である。前項でみたように、レベル a は重複が多い。それでも 3 年生までは使用語彙が増えていることが分かる。逆に 4 年生からは減っていく。基礎的な語彙よりも新しく獲得した語彙を使う傾向がみられる。そして、全体では小学校前期の使用語彙は 32.4%。これは全体のレベル a の使用率 36.1%に近い。

小学校出現語彙に比べて増えているのが、中学・高校で習う語彙である。習った言葉を徐々に自分のものとして使用していく様子がうかがえる。

３．５　教科特徴語分析

最後に、教科書に特徴的な語彙との比較をする。これは、一般の書籍に比べて、教科書に特徴的な語彙ということは、学校で習うことで身につけて、しかも使えるようになってきていることを意味する。基本的に言葉は国語で習い、それを使う練習をしながら、使用

語彙が増えていくと考えられる。また、そうであるからこそ、国語科の授業の中で漢字練習などが多くなされることとなっている。他教科で漢字練習や語彙テストなどはあまり行われていないだろう。

これを検討するために、教科特徴語と使用語彙との関係を調べたものが表 6 である。教科特徴語については、複数教科にわたって特徴度が高いものは除き、各教科単独で数値が高く有意な語彙をカウントした。

表 6　使用語彙の教科特徴語の割合

教科特徴	国語	社会	数学	理科	英語	技術	芸術	保健	情報
1年	3.4%	6.0%	1.0%	2.2%	1.7%	4.1%	2.6%	0.9%	0.8%
2年	3.6%	6.2%	1.0%	2.6%	1.7%	3.5%	3.1%	1.5%	0.9%
3年	3.4%	5.6%	1.3%	1.8%	1.7%	3.6%	2.9%	1.4%	1.2%
4年	3.0%	6.3%	1.1%	2.0%	1.7%	3.5%	2.5%	1.6%	0.8%
5年	6.4%	3.1%	1.1%	2.4%	1.7%	2.8%	2.5%	1.5%	1.0%
ALL	3.0%	6.4%	0.8%	2.5%	1.4%	3.1%	2.5%	1.4%	1.1%

表 6 の全学年データをグラフにしたものが、図 4 であるが、その際に、語彙全体にどれだけ教科特徴語が含まれているのかを比較するために、横に並べてグラフ化した。全学年で使用されている語彙の中で、各教科の特徴語の割合を示したものが図 4 の ALL である。

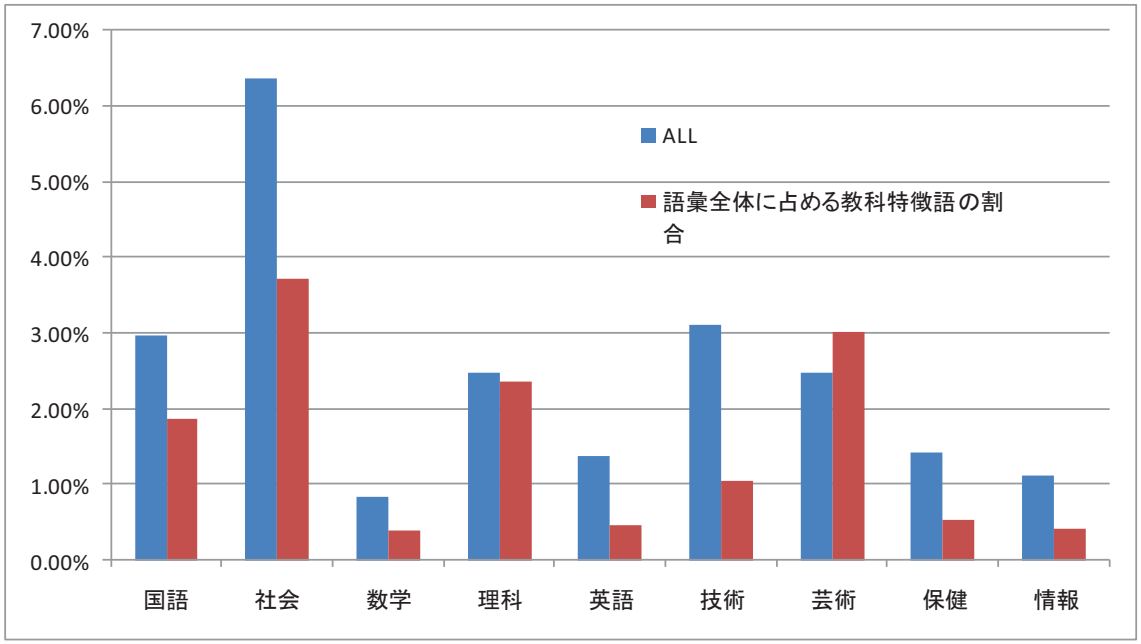


図 4　生徒使用語彙と語彙全体に占める教科特徴語比較

この図をみると、圧倒的に社会の特徴語が多い。社会科で習った言葉を使って文章を書いていることが分かる。この結果は、課題が「文化について」というテーマと評論文のような文種であったことも影響していると考えられる。しかし、語彙全体に占める割合の倍近い語彙を使用しているということは、評論文などを書く力は、社会科的語彙の拡充が必要であることが分かる。また、社会科での語彙の拡充により、それが概念や考える力などを形成し、文章力につながるのではないか。社会科のテストなど、別の方面からの検証も必要である。

次に多いのが、技術科である。技術科は「技術・家庭」科であり、おせちなどの食文化にかかわることを学習するために、このような結果になっていると考えられる。具体的な語は以下のものである。(全学年データで、技術科だけに特徴的なレベル e の語彙)
田作り、取り分ける、酢の物、挟む、魚肉、取り皿、干し柿、レトルト、八宝、ごまめ、きと、グラタン、満たす、汲み、である。

このことから、語彙の学習は国語科だけで行うのでは、なかなか広がって行かないことが分かる。生徒は様々な学習を通して言葉を広げ、生活の中で生かしながら、自分の使用語彙として定着させている。したがって、語彙の拡充は国語科の教科書の中だけで完結してしまうのではなく、様々な教科や生活と結びつけながら行うことで、考えも広がっていく。そのような観点でこれからの学習方法の開発を行う必要がある。

次に、学年ごとの変化を調べるために、図 5 を作製した。

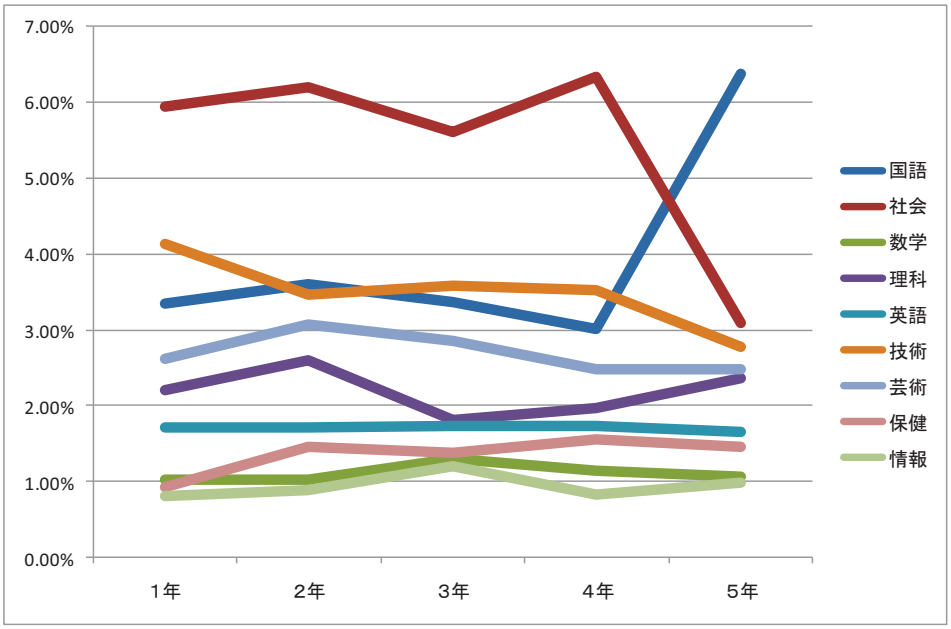


図 5　各教科特徴語の学年推移

1 年から 4 年までは、社会科や技術科の特徴語彙が多いなど、全学年で見たときと同様の傾向がみられる。特筆すべきは 5 年生の国語と社会の関係である。5 年生では、国語特徴語彙が飛びぬけて多くなり、社会特徴語彙が一気に減っている。全体を 100 としているので、片方が増えれば片方が減るということになるが、他の教科はそれほど大きな変動がないので、この二つの教科の特性が影響したものと考えられる。

表 7 は国語にのみ特徴的な語彙の中で、全てのレベル e を抜き出したものである。これらの言葉は、一般にあまり頻度が高くないが、教科書には特徴的に出てくる語であり、しかも使用語彙となっている語である。5 年生になってこれらの言葉が特徴的に出てくるということは、5 年生の使用語彙に関してはかなり国語科に負うところが大きく、学習によって身につけていく様子がうかがえる。

表 7　生徒使用語彙*国語特徴語*レベル e

語彙素読み	語彙素	語彙素のカウント	品詞	語種	レベル_LB_FL	初出学年	特徴教科
ナナクサ	七草	13	名詞-普通名詞-一般	和	e	中	国
モチヨル	持ち寄る	3	動詞-一般	和	e	小前	国
アヤツル	操る	1	動詞-一般	和	e	高	国
ジュウタイ	重態	1	名詞-普通名詞-一般	漢	e	高	国
トシノクレ	年の暮れ	1	名詞-普通名詞-一般	和	e	高	国
ハレガマシイ	晴れがましい	1	形容詞-一般	和	e	中	国
ヒモジイ	ひもじい	1	形容詞-一般	混	e	中	国
ホームルーム	ホームルーム	1	名詞-普通名詞-一般	外	e	中	国
ミンブ	民部	1	名詞-普通名詞-一般	漢	e	高	国
モギ	裳着	1	名詞-普通名詞-一般	和	e	高	国
シュウ	終	1	接頭辞	漢	e	中	国

国語科のカリキュラムとして、4 年生までが必修単位「国語総合」を履修することになっており、5 年生から「現代文」や「古典」などを選択することになっている。このことを考えると、5 年生での使用語彙として抽象度も上がっていることがうかがえる。国語科が担う責任は大きい。

4 年生までの必修単位では、全教科との連携や生活や生きることに結び付いた学習が望まれ、5 年からは一般的にはそれほど多く目にすることがないような抽象的な語彙を積極的に学習の中に取り入れていく必要があり、そのことが学習者にとって知識を活性化し、使用語彙の広がりへと結びついていっている。

4．まとめ

語彙レベルや教科特徴語など、生徒の表現語彙を分析する際にそれらを使用することで、生徒の使用語彙の発達や伸長がとらえられた。課題として、この使用語彙を、どのようにしたら効果的に伸ばせるかという教育プログラムの問題である。それには、具体的な一つ一つの語彙についての分析が必要であろう。使用語彙の拡充が生徒の概念形成を助け伸ばすことになる。

文献

鈴木一史（2011）「作文コーパスからみる生徒の使用語彙」　『特定領域研究「日本語コーパス」言語政策班報告書　』

近藤明日子（2009）「中学校教科書の教科特徴語の抽出と考察―『現代日本語書き言葉均衡コーパス』の語彙との比較から―」『特定領域研究「日本語コーパス」平成 19 年度公開ワークショップ（研究成果報告会）予稿集』

田中牧郎（2008）「教科書コーパスの設計」『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』特定領域研究「日本語コーパス」言語政策班中間報告書

田中牧郎（2009）「言語政策に役立つ、コーパスを用いた語彙表・漢字表などの作成と活用」『人工知能学会誌』24-5

田中牧郎・近藤明日子・河内昭浩・鈴木一史・棚橋尚子（2010）「＜学校の語彙＞と＜社会の語彙＞―「教科書コーパス」と「流通実態サブコーパス」の比較―」『特定領域研究「日本語コーパス」平成 22 年度全体会予稿集』

学習データ間距離学習に基づく語義識別の性能分析

佐々木稔（言語処理班分担者：茨城大学工学部）†
新納浩幸（言語処理班分担者：茨城大学工学部）‡

Word Sense Discrimination Based on Distance Metric Learning from Training Documents

Minoru Sasaki (Faculty of Engineering, Ibaraki University)
Hiroyuki Shinnou (Faculty of Engineering, Ibaraki University)

１．はじめに

ある単語が含まれる用例文集合に対して、語義別に用例文を分類することは本格的な意味解析を行う上で、非常に有用なデータセットの構築への可能性が広がる。例えば、語義別に分類された用例文集合が存在すれば、語義ごとに周辺の共起語を分析することで語義識別モデルを作成し、単語の意味を特定するための分類器を作ることができる。また、動詞についての格フレームを容易に自動構築することや語義ごとに項目分けをしたシソーラスを容易に構築することなどが可能となる。このようなシソーラスを構築するためには、単語に対する既存の語義識別能力を更に向上させることが不可欠である。単語が辞書中のどの意味区分に該当するのかを高い精度で識別することができれば、語義識別モデルを構築することに向けた学習データとしての利用や、意味を調べたい利用者に分かりやすい用例文を提供することへの利用などが可能となる。

語義識別システムは一般的に分類問題として定式化され、教師あり学習手法が用いられる。正解の語義が割り振られた用例文集合を教師データとし、その集合より語義を識別する分類モデルを構築する。この識別モデルに対して語義が不明な用例文を与え、各語義の中で最も相応しい語義を自動的に選択する。このとき、単語と共起する特徴を比較可能な形式に変換するために、頻度などを要素とするベクトルとして表現する。これにより、Support Vector Machine(SVM) (Cortes 1995) などといった教師あり学習手法を利用することが可能となる。

本稿では、既存の語義識別手法に対して更なる識別精度の改善を目的とするために、用例間距離学習手法を利用した語義識別モデルの構築について検討を行う。一般的にベクトル空間モデルを基本とした語義識別は、ある単語について同じ語義を持つ場合にはその単語の周辺において共起する単語の出現傾向が類似していると言われる。また、異なる語義で単語を使う場合には、一方の語義と比較して異なる単語が出現する傾向にある。距離学習手法は同じ語義を持つ特徴ベクトルの点集合は近い場所に集め、異なる語義を持つ点は遠い場所に離すことで、より語義識別しやすい特徴ベクトルを獲得する。

今回の報告では、最適な位置関係を得るために座標軸を変換する距離学習手法である Local Fisher Discriminant Analysis(LFDA) (Sugiyama 2006) (Sugiyama 2007)、Semi-Supervised Local Fisher Discriminant Analysis(SELF) (Sugiyama 2010) を利用する場合と、データの移動を行いデータ間の最適な位置関係を求める距離学習手法である Neighborhood Component

† msasaki@mx.ibaraki.ac.jp
‡ shinnou@mx.ibaraki.ac.jp

Analysis(NCA) と Large Margin Nearest Neighbor(LMNN) を利用する場合について語義識別実験を行った結果を示す。

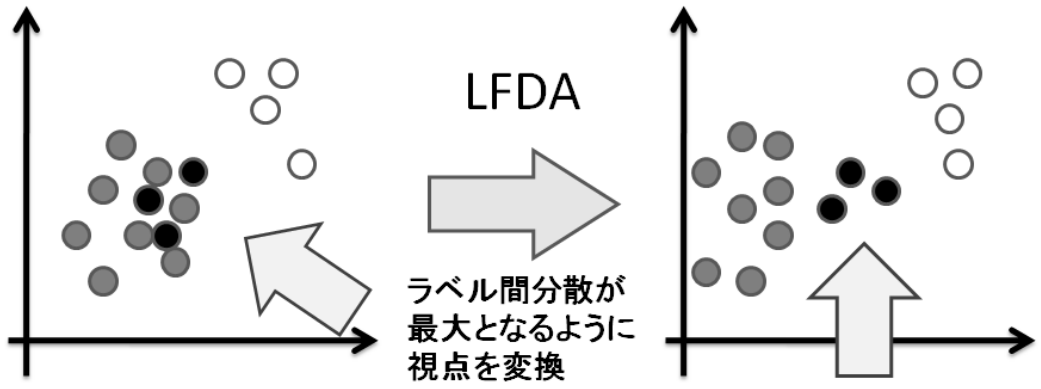


図 1　Local Fisher Discriminant Analysis による距離学習

２．ラベルによるデータ間距離の学習手法

教師データによる距離学習手法は、ラベル付きデータ集合に対して各データ間の距離をラベルに応じて変化させ、データ集合の最適な位置を求めるものである。同じラベルを持つデータ間の距離は短く、異なるラベルを持つデータ間の距離は遠くなるように変換を行う。その際、距離学習の方法には大きく分けて、座標軸変換による学習とデータ移動による学習という 2 つの種類が存在する。本節ではこれら 2 つの学習手法の概要について説明する。

２．１　座標軸変換による距離学習

距離学習の方法で座標軸変換を利用することは、データ分析などでは一般的な方法としてよく利用される。これはラベル間の関係を調整するために、各ラベルに対してラベル内分散が最小、ラベル間分散が最大となるように、座標軸を回転させて最適なデータの位置関係を求める手法である。この考え方を利用した分析手法で代表的なものは、Local Fisher Discriminant Analysis(LFDA) (図 1) である (Sugiyama 2006) (Sugiyama 2007)。LFDA ではスパースな行列に対して一般化固有値を計算することができない場合があるため、主成分分析を組み合わせた Semi-Supervised Local Fisher Discriminant Analysis(SELF) も存在する (Sugiyama 2010)。

この手法はデータの可視化や分析をする場合において、全データの位置関係を調べるときに有効な手段となる。しかし、この手法を利用して未知データの識別を行う場合は問題が生じる。ラベルに応じてデータが移動する訳ではなく、座標軸が回転されているため、SVM などで識別平面を求めると、同じ形の識別平面が回転して存在することになる。これにより、未知データを識別しても精度はほとんど変化しない結果となる[1]。従って、未知データに対してラベルの識別を行う際には、座標軸変換による距離学習と SVM などの識別平面による分類手法との組合せは適していない事がわかる。

---

1 座標軸の回転をする際に次元縮退が同時に行われるため、その分に対応する少しの精度変化は存在する。

２．２　データ移動による距離学習

距離学習の別の方法として、データ移動による手法も存在する。これは、座標軸を回転させてラベル間のデータ関係を最もよく表現する変換を行うのではなく、データそのものをラベルに応じて移動させることで最適なデータの位置関係を求める手法である。この考え方を利用した分析手法として、Neighborhood Component Analysis(NCA) (Goldberger 2005) と Large Margin Nearest Neighbor(LMNN) (Weinberger 2009) が存在する。

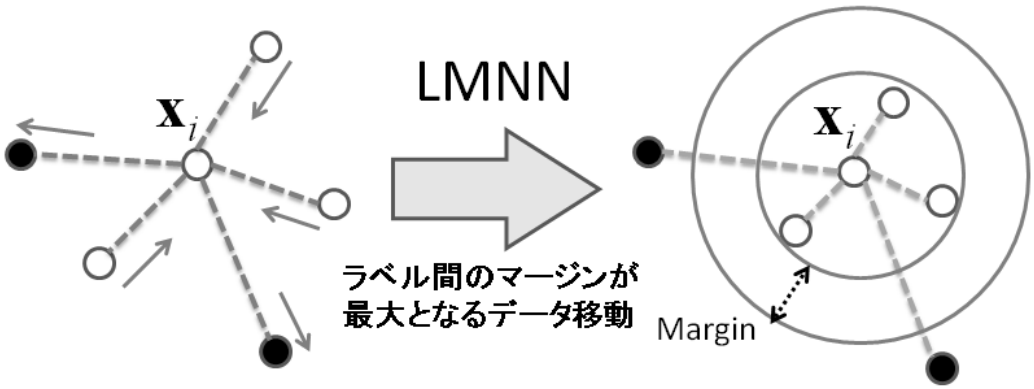


図２　Large Margin Nearest Neighbor による距離学習

これらの手法は共にデータ間のマハラノビス距離を最適化するもので、それぞれの手法において設定した目的関数に対して最適な変換行列を求める。例えば、n 個の D 次元ベクトル $\mathbf{x}_i\ (i = 1, \bullet\bullet\bullet, n)$ と各ベクトルに対応するラベル $c_i\ (i = 1, \bullet\bullet\bullet, n)$ を考えたとき、2 つのベクトル $\mathbf{x}_i$ と $\mathbf{x}_j$ のマハラノビス距離は $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j)^{\mathrm{T}}(\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)$ となる。ここで、行列 $\mathbf{M}$ は、$\mathbf{M} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ を表し、これらの距離学習手法はこの行列 $\mathbf{M}$ を求めることが目的である。

２．２．１　NCA の目的関数

NCA は 2 つのデータ $\mathbf{x}_i$ と $\mathbf{x}_j$ の近さを表す尺度 p_{ij} を以下の式で表す。

$$p_{ij} = \frac{\exp\left(-\left\|Ax_i - Ax_j\right\|^2\right)}{\sum_{k \neq i}\exp\left(-\left\|Ax_i - Ax_k\right\|^2\right)}$$

この類似度を利用して、データ $\mathbf{x}_i$ に対して同じラベルを持つデータについて総和を求めたものが $\mathbf{x}_i$ の重要度となる。

$$p_i = \sum_{j \in C_i} p_{ij}, \quad C_i = \left\{ j \mid c_i = c_j \right\}$$

NCA の目的関数は、この重要度 p_i をすべてのデータについての和を最大化することで、

最終的にそのときの変換行列 **A** を求める。しかし、この目的関数は局所解に収束する可能性があるため、探索を行って収束したとしてもそれが大域的な最適解ではない場合がある。

２．２．２　LMNN の目的関数

LMNN は図 2 に示すように、データ $\mathbf{x}_i$ に近い指定した数の同じラベルのデータは近くに移動し、異なるラベルのデータはマージンが最大となるように移動する。このとき、近傍に存在するデータを表すフラグ行列 η を定義し、データ $\mathbf{x}_j$ が $\mathbf{x}_i$ の近傍にある場合に $\eta_{ij} = 1$、近傍にない場合は $\eta_{ij} = 0$ とする。このとき、目的関数となるコスト関数は以下のように定義され、この関数を最小とする変換行列 **A** を求める。

このコスト関数の第 1 項は同じラベルについての距離関係を表し、第 2 項は異なるラベルについての距離関係を表している。この関数を半正定値計画問題として最適解を求める。

３．距離学習を用いた語義識別手法

前節において紹介した距離学習手法を用いて語義識別を行う概要を説明する。

３．１　特徴抽出

語義の判別を行う単語を含む一文に対して、それと共起する単語を抽出する。本稿における語義識別手法では、学習データ、テストデータ共に形態素解析を利用して名詞と動詞を特徴として抽出することとする。この共起単語についての頻度を要素とするベクトルを作成し、距離学習と語義識別に使用する。

３．２　距離学習とモデル構築

学習データに対して、距離学習手法を利用して語義識別モデルを構築する。本稿では、NCA、および、LMNN を利用して距離学習を行い、語義識別モデルに適用するためのデータに変換する。変換されたデータ集合に対して、NCA では SVM を利用して識別平面を求め、語義識別を行うためのモデルを構築する。また、LMNN では最近傍法を利用して、テストデータに最も近い学習データのラベルを判定結果として出力する。

３．３　語義の識別

構築した識別モデルに対して、語義を調べたいテストデータを入力し、自動的に語義の識別を行う。このとき語義の数が 3 個以上存在する場合は、SVM と LMNN では識別方法が異なる。

SVM を利用する場合は、one-versus-rest 方式で各語義について繰り返し識別を行い、語義の識別をする必要がある。LMNN の場合は、One Nearest Neighbor(1-NN) 方式で、最も近い学習データの語義を識別結果とするため、繰り返し識別する必要はない。

４．実験

NCA、LMNN などの距離学習手法を利用した語義識別手法の精度を評価するために識別実験を行った。本節では、語義識別実験の概要を説明する。

４．１　データ

本実験で使用するデータは、Semeval2010 日本語 WSD タスクで課題として公開されたデータを利用する。これは 50 語の対象単語が指定され、その各単語についてそれを含む文を共起データとして使用する。共起データである文の数は学習データ、テストデータにおいて各 50 文用意され、学習データには対象単語の語義ラベルが付与されている。

表 1　各手法による語義識別の結果(1)

単語	1NN のみ	SVM のみ	SELF+SVM	NCA+SVM	LMNN+1NN
現場	30	39	39	37	29
場所	44	48	48	48	48
取る	13	13	13	13	14
乗る	27	25	25	20	27
会う	28	33	33	33	33
前	24	31	31	29	27
子供	26	18	18	21	26
関係	39	39	39	39	39
教える	15	9	9	9	13
勧める	20	16	16	16	27
社会	40	43	43	43	42
する	18	21	21	23	20
電話	31	28	28	35	33
やる	46	47	47	47	47
意味	26	27	27	23	26
あげる	15	18	18	18	17
出す	18	14	14	17	26

４．２　評価方法

テストデータに対する語義識別結果を評価するために、50 件のデータに対し、距離学習を行わずに SVM で識別、NCA で距離学習を行い SVM で識別、LMNN で距離学習を行い 1-NN で識別した各実験について正解数の比較を行う。また、各単語の正解数の比較だけではなく、全テストデータにおける各手法の正解率を平均的な精度として評価を行う。

５．　実験結果と考察

５．１　テストデータによる識別

各手法に対する実験結果を表 1～3 に示す。NCA を利用した場合は、9 単語について精度が向上したものの、10 単語は精度が下がり、残りの 31 単語は変化なしの結果となった。全体的には性能改善の傾向が見られず、更なる改良が必要な結果となった。その方法として、学習データ用例文数の拡充、特徴抽出手法の改善、および、射影する次元数の最適化が考えられる。

LMNN を利用した場合は、SVM のみを利用する場合と比較して、精度が 68.9% から 69.6% と若干向上する結果が得られた。これより、NCA や LFDA、SELF を利用するよりも

表 2　各手法による語義識別の結果(2)

単語	1NN のみ	SVM のみ	SELF+SVM	NCA+SVM	LMNN+1NN
生きる	47	47	47	47	47
経済	47	49	49	49	49
良い	24	12	12	15	23
他	50	50	50	50	50
開く	45	45	45	45	45
もの	44	44	44	44	44
強い	43	46	46	46	45
求める	39	38	38	38	39
技術	39	42	42	42	41
与える	21	29	29	28	25
市場	14	35	35	34	20
立つ	18	26	26	22	16
手	41	39	39	39	40
考える	49	49	49	49	49
見える	19	26	26	23	23
一	45	46	46	46	46
入れる	28	36	36	36	34

高い精度で識別可能なモデルの構築をすることができると考えられる。また、NCA では少ない学習データで距離学習を行っていたために局所解に収束し、識別精度が下がる傾向があったが、LMNN を利用し大域解を得るための変換行列を求めることで、識別精度が向上することも確認することが可能である。

５．２　距離学習の効果

従来法としてよく使われるSVMに基づく語義識別ではデータ間の関連性などといったより深い分析作業に手間がかかる。しかし、距離学習に基づく語義識別ではこの作業を簡単に分析することが可能となる。まず、SVM とは異なり、1-NN を利用することでテストデータに対して最も近い学習データを特定することができる。テストデータに対して、最も近い学習データの選ばれる傾向を分析した結果、LMNN を利用した場合は 3 つ程度の特定の学習データのみで語義を識別する傾向があった。その中には単語数の少ない短い文が選ばれることが多かったが、どのような内容の文が識別に使われやすいのかなど、より深い分析は今後の課題として進めていく予定である。

また、SVM では識別する場合は、one-versus-rest 方式で繰り返し識別が行われる。このとき、3 つ以上語義がある場合は、テストテータと各ラベルの最短距離を比較することが難

表 3　各手法による語義識別の結果(3)

単語	1NN のみ	SVM のみ	SELF+SVM	NCA+SVM	LMNN+1NN
場合	42	43	43	43	45
早い	31	26	26	27	28
出る	22	30	30	30	28
入る	20	25	25	26	34
はじめ	38	30	30	33	44
情報	39	40	42	37	32
大きい	45	47	47	47	47
見る	39	40	40	40	40
可能	23	28	28	28	30
持つ	30	34	34	34	29
時間	43	44	44	42	44
文化	46	49	49	49	49
始める	39	39	39	40	39
認める	39	35	35	35	39
相手	41	41	41	41	40
高い	26	43	43	43	43
全体	0.6544	0.6888	0.6896	0.6876	0.6964

しい。LMNN では各ラベルとの最短距離を計算することが可能であるため、テストデータの識別しやすさを分析するには非常に有効な手段となる。また、新語義とみなされるデータの位置関係を調査する際の手段としても有効であると考えられる。

６．おわりに

本稿では、既存の語義識別手法に対して更なる識別精度の改善を目的とするために、用例間距離学習手法を利用した分類モデルの構築について検討した。その結果、LMNN を利用した語義識別手法を利用することで、従来よく利用される SVM よりも高い精度で識別することが可能であることを示した。また、LMNN を利用した場合は、3 つ程度の特定の学習データのみで語義を識別する傾向や 3 つ以上の語義を持つ場合の各語義間の関係を調べる上で有効な手段であることが分かった。今後の課題としては、教師データを利用した座標軸変換のより効果的な利用方法を考え、語義識別性能の改善を行う予定である。

参考文献

Corinna Cortes and Vladimir Vapnik (1995). “Support-vector networks”, Journal of Machine Learning, 20(3), pp. 273–297.

Jacob Goldberger, Sam Roweis, Geoff Hinton and Ruslan Salakhutdinov (2004). “Neighborhood Component Analysis”, Proceedings of Advances of Neural Information Processing (NIPS), pp.513-520.

Masashi Sugiyama (2006). “Local fisher discriminant analysis for supervised dimensionality reduction”, Proceedings of the 23rd international conference on Machine learning (ICML06), pp.905–912.

Masashi Sugiyama (2007). “Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis”. Journal of Machine Learning Research, vol. 8, pp.1027–1061.

Masashi Sugiyama, Tsuyoshi Idé, Shinichi Nakajima and Jun Sese (2010). “Semi-supervised local fisher discriminant analysis for dimensionality reduction”. Machine Learning, vol.78, pp.35–61.

Kilian Q. Weinberger and Lawrence K. Saul (2009). “Distance metric learning for large margin nearest neighbor classification”, Journal of Machine Learning Research, vol.10, pp.207–244.

コーパス管理・検索ツール「茶器」

松本裕治（ツール班班長：奈良先端科学技術大学院大学情報科学研究科）†
浅原正幸（ツール班分担者：奈良先端科学技術大学院大学情報科学研究科）
岩立将和（ツール班協力者：奈良先端科学技術大学院大学情報科学研究科）
森田敏生（ツール班協力者：総和技研）

ChaKi: Annotated Corpus Management and Search Tool

Yuji Matsumoto (Nara Institute of Science and Technology)
Masayuki Asahara (Nara Institute of Science and Technology)
Masakazu Iwatate (Nara Institute of Science and Technology)
Toshio Morita (Sowa Research Co., Ltd.)

１．はじめに

本プロジェクトのツール班では，日本語コーパスに対する種々の言語情報の自動タグ付け，および，タグ付け支援ツールの開発を行ってきた．「茶器」は，主に，形態素，文節の分かち書き，および，文節間の係り受け情報のタグ付けが施されたコーパスを格納，タグ付け支援，コーパス利用のための種々の機能を提供することを目的として開発されたツールである．プロジェクト中で何度か仕様の更新や再実装を行い，当初計画してきた機能を一通り完成することができた．本稿では，現在の茶器の主な機能について概説する．茶器の最新版とオンラインマニュアルが，本稿の最後に示す URL より入手可能である．

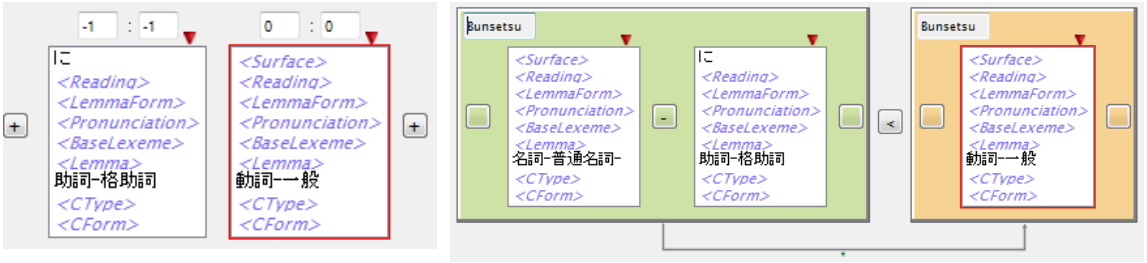
２．タグ付きコーパス管理・検索ツール「茶器」の基本機能

日本語コーパスプロジェクトでは，データ班がコーパスの構築を，電子化辞書班が日本語辞書(UniDic)の構築とコアデータ(コーパス全体から選択された 100 万語規模のコーパス)への(短単位に基づく)形態素情報のタグ付けと(長単位に基づく)文節情報のタグ付けを担当している．ツール班では，形態素解析以上の様々なアノテーションのための自動言語解析ツールと解析済みコーパスの管理ツールを開発しているが，茶器は，その中でも，形態素，文節，係り受け情報を付与されたコーパスの構築支援，および，検索等の利用環境の提供に特化したツールである．茶器の基本機能を以下にまとめる．

1. タグ付きコーパスのデータベースへの格納：形態素，文節，文節係り受け解析，あるいは，その一部が施された解析済みコーパスを関係データベースへ格納する．データベースシステムとして，MySQL，SQLite，PostgreSQL など様々な関係データベースシステムが利用可能である．個人ユーザによるパソコンでの利用の場合には，SQLite を用いれば，データベースが一つのファイルとして格納されるため，データベースのやりとりがファイル単位で行えるようになった．データベースをサーバに格納し，茶器をクライアントとしてネットワーク経由でコーパスにアクセスすることも可能であり，MySQL はそのような用途で用いるのに向いている．茶器は，日本語だけを対象にしたシステムではなく，他に，中国語，英語等多言語の品詞タグ付きコーパス，単語係り受け解析済みコーパスを取り扱うことができる．
2. 検索機能：文字列，形態素列，および，文節係り受け構造を用いた検索要求を発行す

† matsu@is.naist.jp

るインタフェースを提供している．複数のコーパスを指定して同時に検索することも可能である．文字列検索では，簡易型の正規表現を用いることができる．形態素列検索では，形態素が持つ任意の情報（出現形，読み，発音，原形，品詞，活用型，活用形等）を指定した検索が可能である．係り受け解析については，文節内の形態素情報および文節間の係り受け構造を指定し，それを部分構造として含む文を検索できる．下の右図が単語（列）検索の実行，左図が係り受け構造の検索の実行例である．

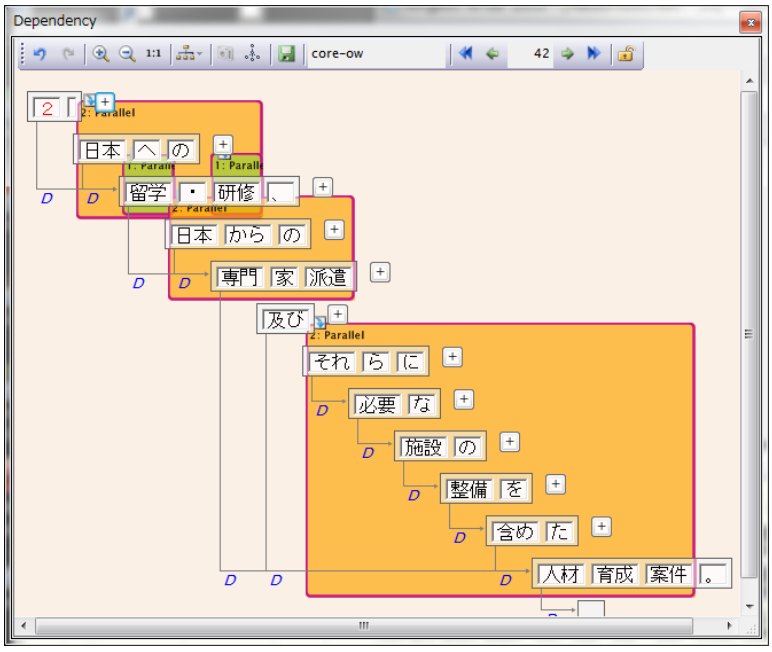


3. 検索結果の表示：検索は文単位で行われ，検索結果は，KWIC(key word in context)形式で１文１行の形で表示される．KWIC パネル上で表示される各形態素には，その形態素がもつ情報のうち二つ（例えば，出現形と品詞）を選択して 2 行に表示することができる．また，詳細な形態素情報を表示するパネルがあり，マウス位置の形態素がもつ情報を表示する．検索要求に適合する形態素のみを検索したいという要求のために，単語リスト(WordList)という検索オプションがあり，検索条件に合致した単語の一覧表示する機能がある．単語の表示を，例えば，原形に限定することにより，活用形などを無視した出現頻度を表示することができる．検索結果は，Excel ファイルとして出力することもできる．また，文の係り受け構造を表示する DependencyPanel を備えており，文節間の係り受け構造を表示する．
下図は，上で示した係り受け検索を行った検索結果の KWIC 表示例である．

Corpus	Doc	Char	Sen	Left	Center	Right
core-ow	12	175	834	に より 、 東西 対立 の 時代 に	比し て	、 主要 国 間 の 協調 が 図ら れる よう に なり 、
core-ow	12	417	838	理 を 中心 と する 国連 の 役割 に	対する 期待 が 高まる と とも に	、 国際 社会 が 対応 を 迫ら れ
core-ow	12	417	838	ー 国 内 に おけ る 紛争 へ と	変わっ た 結果	、 国連 に よる 平和 維持 活動 の 任務 は 、
core-ow	12	669	840	よ 、 軍事 部門 及び 文民 部門 に	及ぶ	様々 な 機能 を 成功 さ せ て 活動 を 終了 し た 。
core-ow	12	1193	848	の ため の 方策 に つい て 議論 を	行っ	て き た 。
core-ow	12	1515	853	、 7 後方 支援 拡充 と 本部 に	よる	支援 能力 の 強化 。

4. 統計情報の取得と表示：検索された文集合に対して，KWIC の中心語とその前後に現れる形態素の出現回数の表示を行う．また，頻度以外に，中心語と前後に出現する語との相互情報量などの統計情報を計算し表示する．また，検索された文集合に含まれる頻出単語系列で利用者が指定した条件（出現頻度，系列の長さ，系列に含まれるギャップの最大値など）を満たすものをすべて列挙する機能を提供する．
5. タグ付け作業支援およびタグ付け誤り修正：未解析のコーパスに対して，形態素や係り受け情報などのタグ付け作業を行うことや，解析済みコーパスに含まれる解析誤りを修正するインタフェースを提供する．解析誤りを発見した場合に，検索機能を用いて同様の誤りを含む箇所をコーパスから網羅的に検索し，それに対して誤りを修正するインタフェースを提供している．係り受け解析木の表示パネル DependencyPanel は，

係り受け木を表示するだけでなく，形態素区切りや品詞等の形態素情報の修正，文節区切りの修正，係り受け構造の修正を行うことができる．下の２つの図は，上は，形態素列の表示と形態素情報の修正を行う際の表示（形態素表示モード），下は，係り受け解析木の表示（対角表示モード）の例である．係り受け木には，並列構造のアノテーションを重ねることができる．図では，緑色で示したボックス(1. Parallel)，および橙色のボックス(1. Paralle)の２種類の並列構造のアノテーションが施されている．



6. 並列・同格構造，埋め込み構造のアノテーション機能：上の図で示した並列構造は，言語解析における大きな問題であり，また，係り受け構造のみを用いて表現することが困難な現象である．現在は，並列構造の範囲を示すアノテーションと係り受け構造のアノテーションを独立に行っている．茶器では，上図のように DependencyPanel を用いて，係り受け構造だけでなく，並列構造の範囲指定を行うことが可能になっている．同じ機能を用いて，同格関係にある形態素列に関するアノテーションも行うことができる．図１に並列構造と同格関係のアノテーションの例を（水平表示モードで）示す．３つの要素からなる並列構造と，それを１つの要素として含む同格構造を示している．並列，同格いずれの要素も任意の形態素列であり，文節切りとは独立に範囲指定が可能である．それぞれの構造の要素（形態素列）は，マウス操作によって範囲指定する．

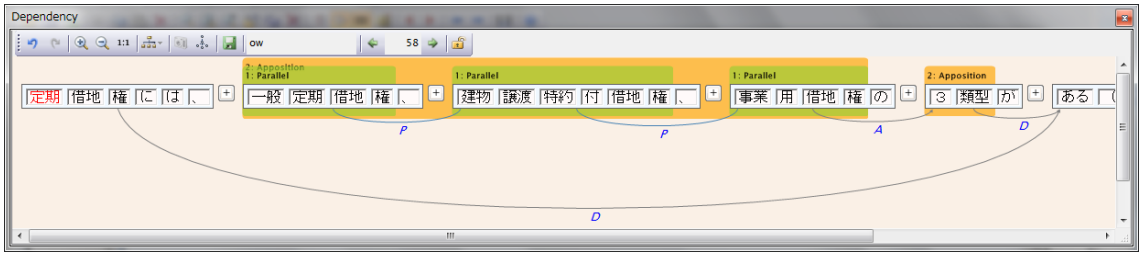


図１．並列構造と同格構造のアノテーション

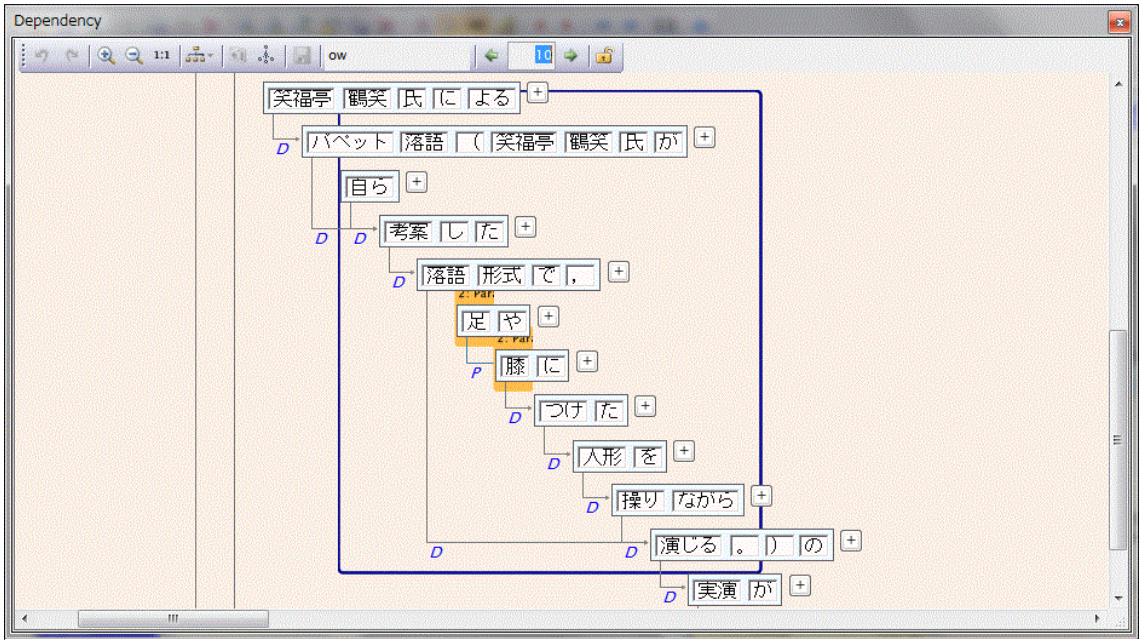


図２．埋め込み構造内部のアノテーション

図２は，埋め込み構造の例である．この文の丸括弧で囲まれた部分は，「パペット落語」の説明文であり，この文全体にとっては，これは「パペット落語の」という文節の一部でしかない．しかし，その内部は独立した文となっている．茶器では，このように内部の係り受けのアノテーションを外部と同様に行うことができる．外側の文から見れば，この埋め込み構造は一つの文節の要素とみなすことができる．

３．今年度の茶器の公開に関する活動

茶器の公開，配布について，今年度は，以下の活動を行った．

- ツール講習会（１１月２０日，キャンパスイノベーションセンター，東京田町）：主に領域内メンバーを対象として，茶器に関する講習会を行った．日本語コーパス外部公開版に含まれる形態素情報付きのコアデータの SQLite ファイルを配布し，茶器の基本機能に関する講習を行った．また，UniDic を用いた形態素解析器 MeCab，および，係り受け解析器 CaboCha，および，これらの解析ツールを呼び出して適用するためのインタフェースソフトを配布し，任意のコーパスを自動解析して，その結果を茶器に取り込む方法について講習した．

４．あとがき

タグ付きコーパス管理・検索ツール茶器の基本機能と今年度の活動について述べた．Microsoft .NET framework 上での実装を行い，当初予定していた機能の実装が完了した．MySQL だけでなく SQLite 等の種々の関係データベースシステムが利用できるようになったため，コーパスの実装の容易さや検索効率が大幅に改善された．

当初は予定していなかったが開発の段階で明らかになってきた問題点をいくつか列挙しておく．一つはコーパスそのものの修正である．コーパスの整備はコーパス班が担当し，コーパスのオリジナルの文字列の修正は行わないとの前提でプロジェクトを開始したため，茶器では文字データの修正は考慮していない．しかし，実際には個人研究者が管理しているコーパス等ではそのような事態は起こりうる．また，学習者データや誤りを含むコーパスの修正に関しては，オリジナルの文字（単語）列と修正後の文字（単語）列の両方を管理する必要がある．このような機能の拡張が今後必要になるだろう．

茶器は，一つのコーパスを一つのファイルとして管理しており，文書の構造は単なる文の列と考えている．段落のような部分構造の概念や文書間の関係など，具体的な用途に応じて，構造化文書や複数の文書データを関連付けて扱う機能の設計と実装が望まれる．複数文書の関連については，上でも触れた学習者の誤り修正コーパスや多言語のパラレルコーパスの扱いにも関係するので，様々な利用形態を考慮した仕様設計が必要と考えられる．

関連 URL

「茶器」配布用ページ：http://sourceforge.jp/projects/chaki/

このページより茶器の最新版(ChaKi.NET)をダウンロードすることができる．また，インストール方法とオンラインマニュアルがこのページからのリンクにより参照できる．

シンポジウム「日本語コーパスと外国語としての日本語研究」

3月15日（火） 14:00〜17:20

海外の日本語教育から見た均衡コーパス ―日本語教材の評価・比較・編集―

▶曹 大峰（北京日本学研究センター）

イタリア人向けの和伊辞典編纂におけるBCCWJの貢献

▶カルヴェッティ・パオロ（カ・フォスカリ ヴェネツイア大学）

副詞による括弧構造とその文脈における役割について

▶アンドレイ・ペケシュ（筑波大学）

基本動詞ハンドブック執筆へのBCCWJの利用
―辞書執筆用コーパスシステムNINJAL-LagoWordProfilerの開発―

▶プラシャント・パルデシ（国立国語研究所）、赤瀬川 史朗（Lago言語研究所）

海外の日本語教育からみた均衡コーパス
—日本語教材の評価･比較･編集—

曹大峰（北京日本学研究センター）

The BCCWJ from the Viewpoint of Overseas Japanese Education: Evaluation, Comparison and Compilation of Japanese Textbooks

Cao Dafeng (Beijing Centre for Japanese Studies)

代表性を有する日本語コーパス、即ち日本語均衡コーパスの構築は、海外の日本語研究と日本語教育に従事する人々には長年の夢として待ち望まれていた。それが BCCWJ という文部科学省科学研究費特定領域研究による大規模な書き言葉均衡コーパスとして実現されているのは、実に喜ばしいことであり、その応用は今後海外へと大きく広がっていくものと期待される。

本発表では同プロジェクト日本語教育班の活動に参加して得られた筆者の見識と試用事例に基いて、海外の日本語教育という側面から BCCWJ の役割を描き出してみるとともに、日本語教材の評価･比較･編集にその応用の可能性を探ってみたい。

１．海外の日本語教育の課題から

国際交流基金『2009 年海外日本語教育機関調査』の結果によれば、海外の日本語教育は 2006 年より 3 年間で 671,941 人と 22.5%増加し、2009 年には 133 か国・地域（125 か国と 8 地域）で行なわれており、学習者数は 3,651,761 人に達したという。また、教育上の問題点として最も多く挙げられたのは「教材不足」、2003 年と 2006 年の調査結果と連続して教材の開発と作成支援は目立つ課題として提示されていた。しかし、海外においては種々の事情でその課題を解決することは容易なことではない。特によい教材を開発するには、質のよい日本語資源を確保することは前提条件であり、これまではその収集と選定には様々な困難があった。

BCCWJ は、表 1 のように多様なジャンルと文体を備えた大規模な日本語書き言葉均衡コーパスなので、その完成と公開により、これからは海外でも質と量の両面から代表性をもつ日本語コーパスを入手することができるようになり、日本語教材の開発には初めて 1 億語規模の日本語資源が提供されることになる。

表 1　BCCWJ のジャンルと文体（前川 2007、丸山 2009）

<table>
<tr><th>目的</th><th>媒体種類</th><th>予定規模
（語数）</th><th>年代</th><th>内容
ジャンル</th><th>文体</th></tr>
<tr><td rowspan="3">出版</td><td>新聞</td><td>440 万</td><td rowspan="3">2001-05 年</td><td>16(32(32))</td><td rowspan="2">日常的な書き言葉</td></tr>
<tr><td>雑誌</td><td>880 万</td><td>6(912)</td></tr>
<tr><td>書籍</td><td>3080 万</td><td rowspan="2">10</td><td rowspan="2">多分野にわたる書籍文体</td></tr>
<tr><td>流通</td><td>書籍</td><td>3500 万</td><td>1976-05 年</td></tr>
<tr><td rowspan="6">特定
目的</td><td>ベストセラー</td><td rowspan="5">3000 万</td><td rowspan="5">1975-05 年
2001-05 年</td><td></td><td></td></tr>
<tr><td>検定教科書</td><td>10</td><td>多科目の教科書文体</td></tr>
<tr><td>白書</td><td>9</td><td>改まった公文体</td></tr>
<tr><td>Yahoo!知恵袋</td><td>14(59(130))</td><td rowspan="3">書き言葉の周辺文体
(WEB 文体･話言葉の記録文体)</td></tr>
<tr><td>Yahoo!ブログ</td><td>15(53(285))</td></tr>
<tr><td>国会会議録</td><td>2(4)</td></tr>
</table>

最近の第二言語習得研究と外国語教育学では、質のよい学習材料とその適切なインプットが目標言語の習得効果を確実に上げられると主張し実証されている。BCCWJ は現代日本語を中心に豊富な種類・ジャンルと文体をもつ日本語資源なので、質のよい学習材料を提供する基盤的な役割が大きい。これまでの海外の日本語教材は環境と条件に限られて、その内容やジャンルなどには時代遅れ・現実離れ・実態外れの問題がしばしば指摘されているが、これからは BCCWJ の活用により、海外の［教材不足］の問題が質の向上とともに改善されるばかりではなく、基本語彙・基本文型・共起関係など日本語教育スタンダードやシラバス整備の現実化・充実化・能率化も図れるものと期待されている。

２．基盤的コーパスとしての役割

コーパスの開発が多様な種類にわたって進められている中、言語教育または外国語教育の実践と研究に役立つためのコーパス応用研究は新しい領域として形成しつつある。曹 2010 では日本語教育に利用可能なコーパスの種類を、表 2 のように大きく学習資源コーパスと教育資源コーパスに分けて整理してみた。

表 2　日本語教育に利用可能なコーパス

学習資源コーパス	教育資源コーパス
均衡コーパス	基準コーパス
並列コーパス	参照コーパス
WEB コーパス	評価コーパス
表現コーパス	教材コーパス
用例コーパス	講義コーパス
訳例コーパス	学習者コーパス

まず、学習資源コーパスには、均衡コーパス、並列コーパス、WEB コーパスのような現実にある言語資源をそのままコーパス化したものと、表現コーパス、用例コーパス、訳例コーパスのような学習のために抽出・整理を施したものがある。前者は言語の使用実態の多側面を反映しており、「データ駆動型学習(DDL)」に最適だが、初級の学習者には内容が雑多で難しい一面がある。そこで、後者のような特定の表現や用例または翻訳例を抽出し学習のための配慮や加工を施した「学習コーパス」が開発され、出版物として市販されるまでになり、注目される話題となった。ただ、このようなコーパスでは、文脈離れになりがちな一面もあり、場面性を保つ工夫が必要であろう。

また、教育資源コーパスには、基準コーパス、参照コーパス、評価コーパスのような言語教育のための基準や参照枠や評価問題などをコーパス化したものと、教材コーパス、講義コーパス、学習者コーパスのような言語教育の内容や講義と学習者の実態をコーパス化したものがある。前者は教育や学習のスタンダードやシラバスを示すものとして現実性･客観性･有効性を要するものであるが、その更新とコーパス化がしばしば遅れがちである。後者は教育と学習の実態を記録したものとして、実態の把握と分析に役立つものであるが、その開発は個々の研究者によるものなので、今後相互比較や評価に必要な基準が求められると思われる。

上述のように、日本語教育には多様な種類のコーパスを利用することが出来るのであるが、その中で、均衡コーパスとしての BCCWJ は基盤的な役割を果たすものといえよう。並列コーパスは対訳関係を持つ多言語情報、WEB コーパスは WEB の言語情報を提供するというのに対して、均衡コーパスは特定の言語を代表するような言語情報を提供するものなので、日本語教育には最も基本的な学習資源コーパスとして、他のコーパスとは図 1 のような関係をなすものと考えられる。

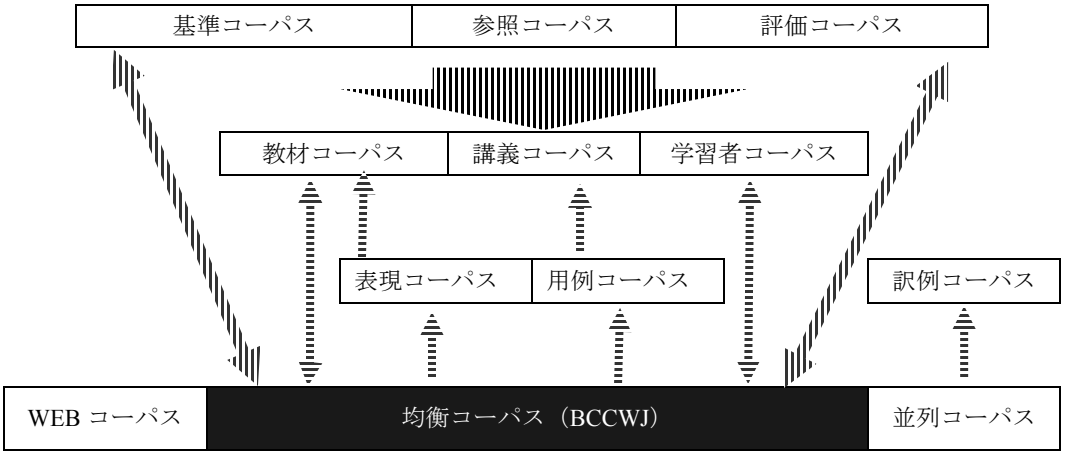


図 1　基盤的コーパスとしての BCCWJ

3．日本語教材の評価･比較･編集

海外の日本語教育ではその環境と条件により教材の役割が大きい。BCCWJ はよい教材を開発するための豊富な日本語資源を用意してくれているので、すでに開発した教材を比較し評価したり、これから開発する教材に質のよい材料を選定するのに活用できる。

ここにまず教科書コーパスとの照応に試用した事例を報告する。文体の学習項目である丁寧体の述語否定形式について、新聞コーパスによる田野村(1994)と日常会話コーパスによる小林(2005)の調査では、日本語教材における「ません系」の絶対化問題が指摘され、中国の日本語教材コーパスによる曹 2010 の調査でその問題が実証されたが、今回、BCCWJによる照応を通して、さらに多ジャンル間の差異と日本の日本語教材[1]や検定教科書の状況を観察することができた。表 3 には照応結果が示されている。

表 3　丁寧体の述語否定形式に関する日本語教材･BCCWJ･日常会話の照応

件数 / 文類	ません系/ないです系						
	教材(中)	教材(日)	検定教科書	新聞	雑誌	国会議事録	日常会話(小林 05)
名だ	104/18	180/4	15/(5)	16/6	146/165	621/633	6/74
ナ形	29/5	39/0	5/0	2/1	12/11	15/3	1/7
イ形	134/10	105/57	13/0	3/0	25/18	19/8	4/38
動詞	2500/5	2573/28	276/0	240/16	1446/169	8713/277	178/278
計	2767/38	2897/89	309/5	261/23	1629/363	9368/921	189/397
%	99/1	97/3	98/2	92/8	82/18	91/9	32/68

上表の結果から分るように、「ません系」の絶対化現象は海外ばかりではなく、日本の日本語教材と検定教科書にも見られることであり、その原因究明と対策は待たれる課題であろう。一方、BCCWJ においては、新聞では田野村(1994)の調査結果（92.8/7.2）とほぼ一致しているものの、ジャンル別の差異は面白い。国会議事録は話し言葉の記録なのだが、日常会話より「ないです系」の使用は遥かに少なく、むしろ、新聞の状況に近い。それに対して、雑誌のほうでは「ないです系」は倍ほどに使われるようになっている。

[1] 日本語教育班で構築された日本語教材コーパス（日本語教材 38 種収録・担当：井上優）を利用させていただいた。

このように、ジャンル別の使用状況を調べられるのは BCCWJ の大きな特徴といえよう。これまでの日本語教材には日本語の実態が必ずしも的確に反映されているとは言えず、今後は BCCWJ のような均衡コーパスによる評価と補足が期待されることであろう。

また、複数の教材を比較する基準として BCCWJ を利用することもできる。表 4 は最も生産性が高い「出す」型複合動詞を取り上げ、BCCWJ における頻度データ(石川 2010)をもとに、中国の教材コーパスにある四種の教材を比較した結果であるが、「出す」の V2 率とそのカナ表記率は BCCWJ より低いこと、教材間に差異があることが観察され、今後の改善課題が示唆されることであろう。

表 4　「出す」型複合動詞に関する BCCWJ と教材間比較

		V1	V2(漢字)	V2(仮名)	V2 率
BCCWJ		3327	5490	1262	66.99
教材(中)	SW	24	5	0	17.24
	BW	4	6	0	60
	BD	6	8	0	57.14
	DW	5	6	5	64.7

表 5　文型「～に～がいる」の使用状況(BCCWJ)

共起＼ジャンル		Yahoo! 知恵袋	書籍	新聞雑誌	検定教科書	計
名詞に	場所	29	15	1	2	47
	地名	3	3	1	1	8
	代名詞	1	1	0	0	2
	位置	24	25	2	2	53
	人(抽象)	23	8	2	1	34
	人(具体)	41	11	2	0	54
	計	121	63	8	6	198
名詞が	人(具体)	6	8	1	1	16
	人(抽象)	114	47	5	3	169
	動物	1	6	1	2	10
	組織･国	0	2	1	0	3
	計	121	63	8	6	198

日本語教材の編集に際しては、実際の使用状況と場面に即して文型を導入するように、初級で学ぶ文型「～に～がいる」の使用状況を、表 5 のように試しに BCCWJ で記述してみたところ、次の２パターンとジャンル別の差異が見えてきた。

(1)位置≧場所>地名>場所指示**に** 人(抽象)>人(具体)>動物>組織**が** います

(2)人(具体)>人(抽象)**に** 人(抽象)>人(具体)>動物**が** います

また、それに対して「～は～にいる」という文型は次の1パターンしか見えず、その単純性が示唆されることになる。

1 人称>人>動物は 場所≧地名≧場所指示>位置>抽象場所に います

この結果はすでに教材の編集に生かされているが、今後も応用実践を続けていきたい。

文　献

国際交流基金(2010).『2009 年海外日本語教育機関調査』結果（速報値）

曹　大峰(2010).「教科書コーパスと日本語教育」『日本語学研究』第 27 輯　韓国日本語学会

石川慎一郎(2010).「『現代日本語書き言葉均衡コーパス』（BCCWJ）における複合動詞『～出す』の量的分析」統計数理研究所研究レポート,238

丸山岳彦(2009).「『現代日本語書き言葉均衡コーパス』領域内公開データ（2009 年度版）書誌情報・サンプル情報・著者情報について」

前川喜久雄(2007).「特定領域研究『日本語コーパス』―目標, 進捗状況, そして夢―」特定領域研究「日本語コーパス」平成 18 年度公開ワークショップ予稿集

小林ミナ(2005).「日常会話にあらわれた『ません』と『ないです』」『日本語教育』125 号

田野村忠温(1994).「丁寧体の述語否定形の選択に関する計量的調査－『～ません』と『～ないです』－」大阪外国語大学論集,第 11 号

イタリア人向けの和伊辞典編纂における BCCWJ の貢献

カルヴェッティ・パオロ（カ・フォスカリ　ヴェネツイア大学）

The Contribution of BCCWJ in the Editing of a Japanese-Italian Dictionary for Italian Readers

Paolo Calvetti (Dept. of Asian and North African Studies - Ca' Foscari University Venice)

0. はじめに

1990 年ローマにある国立アジア・アフリカ研究所（当時名称：中近東・東アジア研究所）とナポリ大学「オリエンターレ」（当時名称：国立ナポリ東洋大学）の共同研究のもとで「和伊辞典」編纂計画が発足したが、その実現は今日に至るまで難航してきた。２０年も経っているが、辞典の発行はまだ遠いようで、経済的な理由などでその確かな見通しはまだ立っていない。しかし、イタリアで日本語を勉強している学生の人数の増加が著しく、日本語を仕事で使う人の数は無視できないものであるため、イタリア人向けの和伊辞典の必要性が痛感されている。とは言え、困難な研究環境の中で、その計画の実現に向かい編集作業を続けている少数人のチームがある。7 万語程度の見出し語、用例の豊富な辞典を目指し、今まで出版された和伊辞典と比べれば（本稿の最後の付録を参照されたい）「大」辞典を編集しようとしている。[1]

1. 和伊・伊和対訳辞典の現状

歴史的な背景や出版界の事情の相違により和伊・伊和対訳辞典編集の業績と歴史はイタリアと日本とではかなり違う。本題から脱線してしまう危険があるので、ここではその詳細は省略するが、結論から言うと日本人の辞書編纂研究者の業績の方が多く、商業出版社との関わり合いも刺激的であると思われ、その結果「使える」辞書が編集され、市販されているのである。電子辞書の普及につれても、その本体に搭載されている機種もあれば、フラッシュカードの形で伊和・和伊のバージョンも入手できる。イタリア人の学生もそのようなものを実際には利用している。

さて、イタリアの辞書編纂を歴史的に振り返る意味でも、明治時代から出版された和伊・伊和辞典のリストをまとめてみると、次のようなことになる。

日本で編集された辞典

- *1876 曲木如長　『仏伊和三国通語』績文社*
- *1936 井上静一　『伊太利語辞典』第一書房 (増訂版 1942)*
- 1938 吉田弥邦　藤堂高紹　『伊日辞典』伊日辞典刊行会
- 1963 下位英一　坂本鉄男　『イタリア語小辞典』大学書林
- 1964 野上素一　『新伊和辞典』白水社　(増訂版 1981)
- 1982 武田正實　『現代和伊熟語辞典』日外アソシエーツ
- 1982/1998 高橋久　『和伊辞典』イタリア書房 (ポケット版 1998)
- 1983 池田康，他　『伊和中辞典』 小学館 (第２版　1999)

[1] Paolo Calvetti,"Perché un nuovo dizionario Giapponese-Italiano", Luisa Bienati, Matilde Mastrangelo 編, *Un'isola in Levante. Saggi sul Giappone in onore di Adriana Boscaro* 所載, 2010, Napoli, ScriptaWeb 出版社, pp. 389-403.

•1987 下位英一　『和伊辞典』大学書林
•1988 坂本鉄男　『和伊辞典』白水社
•1994 西川一郎　『和伊中辞典』小学館 (西川一郎・和田忠彦　監修　第２版　2008)
•2001 郡史郎，池田康　『ポケットプログレッシブ伊和・和伊辞典』小学館

イタリアで編集された辞典

•1910 Chimenz S., *Piccolo dizionario italiano-giapponese*, Hoepli
•1912 Balbi B., *Piccolo vocabolario manuale italo-giapponese*, Hoepli (第２版 1939)
•1940 Scalise G., *Dizionario Italiano-Giapponese*, Società editrice Don Bosco
•1978 Nishikawa I., Dizionario giapponese-italiano dei termini fondamentali, 国際交流基金・日本文化会館
•1992 Scalise M., Mizuguchi A., *Dizionario Giapponese. Italiano-Giapponese, Giapponese-Italiano*, Vallardi
•2000 Kimura A., Hashigata K., *Dizionario giapponese. Giapponese-Italiano, Italiano-Giapponese*, De Agostini
•2006 Marino S., Enomoto Y., *Dizionario giapponese-italiano, italiano-giapponese*, Zanichelli
•2006 Borriello G., Petrella D., *Giapponese: dizionario per immagini*, Vallardi

日本語教育制度、特に大学制度によってイタリア語は、第２外国語としてしか選択されないことがあるので、ほかの外国語（中国語、韓国語、そして、英語、ドイツ語、フランス語）の対訳辞典と比べれば、出版社は商業的な発想からイタリア語辞典にはそれほど力を注いでいないようである。にもかかわらず、日本でのイタリア語関係の辞典編集の活動はイタリアと比較すれば著しいと言えよう。また、残念なことに、近年イタリアで伝統のある出版社からは発刊された辞典でさえも、ほとんど使い物にならないのが現状である。

2. イタリア人向けの日本語対訳辞典の必要性

イタリアで１９８０年代の後半から大学を中心に日本語教育のブームが起こり、その時まで外国としてエキゾチックなことばとされていた日本語は、数多くの学生の興味を引き、数人、数十人の学生のクラスが一気に１００人台を超えることになったのである。日本語教育自体のあり方、教授法の再検討、新しいアプローチの教科書の必要などがその日本語ブームに伴って生じたわけである。

現在ヨーロッパの中で日本語の学生数で英国、フランス、ドイツに続いて４位[2]を占めているイタリアでは、日本語と関係のある仕事をしている人の数は不明ではあるが、翻訳、通訳、観光産業、貿易関係などに携わっている人たちはわずかながらもコンスタントに増加していると思われる。年に１０冊以上の日本の小説も翻訳され、出版されている[3]。この

[2] 国際交流基金編「2009 年海外日本語教育機関調査　速報値」http://www.jpf.go.jp/j/japanese/survey/result/dl/news_2009_02.pdf
[3] Luisa Bienati, Paola Scrolavezza 共著, “La narrativa giapponese moderna e contemporanea”, Venezia 2009 によると 2001 年から 2008 年に渡って、108 冊の近現代文学の小説が出版された。

ような人たちは皆、多種多様の対訳辞典に頼りながら自分の職業に携わっていると思われる。

しかしながら、イタリア人の読者、辞典の利用者のために編集された頼りになるまともな和伊辞典は存在しないと言っても過言ではない。

3. イタリア人向けの和伊辞典の特徴

上述したように現在市販されている和伊辞典の構成や目的は、イタリア語を母国語としている利用者のニーズに背反するか、あるいはそのニーズに応える資格と内容を持っていないといえる。というのは、今日本で出版された和伊辞典は主に日本人がイタリア語で表現したいときに日本語からイタリア語への、広義での翻訳の補助的な手段として工夫され、編集されたものである。対訳辞典というものは双方向的なものではなく、しかも、特定の利用者（和伊辞典の場合、日本人）のニーズに応えるために構成、編集されているのである。例えば、「試験」の項目には「試験に合格する」という用例があるとしたら、同義の「試験に受かる」、「試験に通る」などはないのに対して用例に当たるイタリア語の翻訳は三つもの文がある：「superare [passare] un esame, essere promosso agli esami; vincere un concorso」。

なお、その文法的な注釈やシンタックスの記述などはイタリア語に関するもので、見出し語の品詞や日本語そのものについては大抵の場合、何も書かれていない。これは、辞典の利用者は日本語の母語話者であることが前提となっているからであり、そのような情報を必要としていないと考えられているからである。逆に、イタリア人の日本語の学習者や仕事で（研究、翻訳等）、また、趣味で日本語を読んだり、聴解しようとしたりする人にとっては省略されているこの種の情報は欠かせないものである。もちろん、イタリア人利用者が日本人向けの和伊辞典を丁寧に読んで、用例などを厳密に調べればその大切な情報が得られることもあろう（事実上、選択もないため、イタリア人は日本で編集され、出版された和伊辞典を使う）が、その情報は辞典編纂計画の意図には必ずしも含まれていないし、偶然に存在していても潜在的なものにすぎない。

陳腐な例ではあるが、POS（品詞）の表記は辞典の受動的な知識獲得の役割だけでなく、その能動的な役割にも役に立つ。例えば、「感動」という見出し語を調べれば、名詞、サ変として使えることは記述されていないので、知らない読者は（その見出し語を調べた読者はその単語については知識がないはずであるが）動詞としても使えることが分からないし、せっかく以前知らなかった単語を調べたのに、その言葉についての新しい知識・情報が限られてしまい、その正しい運用もできないかもしれない。動詞の場合では動詞の自他の区別もされていないこともあって、二カ国語間の非対称性のため、対訳の文章だけでは見出し語の性格が分からない[4]。

また、用例に関して考えてみると、利用者がイタリア人なら、和伊辞典に紹介されている用例は日本語のある概念、ある表現を表すための代表性、および、模範性のある文章として見なしていると想像できる。すなわち、それぞれの用例は日本語のサンプルであり、自分が直面する文章にも表れうる一例として扱われるということである。しかし、実際には現在市販されている日本語が含まれている対訳辞典の多数には（和英、和仏などの辞典も同様であるが）、日本人向けのものであり、それぞれの起点言語（日本語）の用例文が単

[4] 見つかる・見つける（「探し出す」の意味で）の対はそれぞれ自動詞と他動詞であるが、辞典に掲載されている用例のイタリア語の対訳ではイタリア語で他動詞の形で表れることが多い。そのため、日本語母語話者ではない利用者が翻訳からは自他の区別が分からない。

なる目標言語（和伊辞典の場合はイタリア語）において、適切に、妥当性のある表現を発せられるような手段に過ぎず、必ずしも「自然な」日本語ではない。

例えば、単文の用例に限るが、「写真」という見出し語の場合は「写真を撮る」という用例はあるが、「写真を写す」（またはネットなどでよく使われている「写真する」）のような例はない。その理由はこの三つの表現は同義として扱われ、同じ意味を表しているので、すべて列挙する必要はなく、逆に唯一の「写真を撮る」という用例に、目標言語であるイタリア語の三つもの異なる表現が記載され、文字通りの翻訳も注釈の形で付き加えられている「fare una fotografia, fotografare; scattare una fotografia（シャッターを切る）」[5]。

言い換えれば、用例の文は日本人にとってはイタリア語の文へのただの導入であるが、イタリア人はその同じ文を日本語の一つの代表的、模範的な文章として扱い、そして、それに対するイタリア語の文を翻訳として解釈しがちである。

また、日本の場合に限らないが、調べれば、違う出版社、違う編集者の同類の対訳辞典、または、同じ出版社の違う対訳辞典（和伊、和英、和西など）に掲載されている用例はよく似ている、場合によっては同一であることが分かる場合も存在する。それは、今述べたように、日本語の用例は実際使用された文のサンプリングではなく、抽象的で出版社の辞書編纂部に用意されている「用例レパートリー」のようなものから選ばれるためである。

さらに、非日本語母語話者が対訳辞典を利用する際、無視できないもう一つの重要な点がある。それは言語位相である。言うまでもなく、そのようなことも日本人向けの対訳辞典は配慮しておらず、「亡くなる」「死ぬ」「死亡する」「死去する」「くたばる」などは大抵目標言語に同じ対訳があり、たまには日本人の読者のためを考えて、対訳語の言語位相に関しての注釈があるだけである。これも辞典利用者からすると対象言語の解釈（理解・翻訳）のためだけでなく、言語運用、発話のためにも大切な要素であり、発話のフィードバックの重要な情報でもある。

もう一つ、対訳辞典は模範的、規範的な役割も担っていると考えられるので、日本語の場合はその正書法、特に漢字表記の変種に関しても非日本語母語話者にこれらに関する示唆を提供するのが妥当と思われる。対訳辞典だけでなく、国語辞典にも類義・同音語の表記（例えば、ワカル：「分かる」「解る」「判る」やトオル：「通る」「徹る」「透る」など）の使い分けもはっきり記述されていないこともある。漢字の伝統を強調する社会にはその情報も無視できないもう一つの要素であるといえよう。

4. 辞典編纂における『現代日本語書き言葉均衡コーパス』（BCCWJ）の可能性

国立国語研究所の協力、および、提供をいただき、『現代日本語書き言葉均衡コーパス』（BCCWJ）を実験的に『和伊大辞典』の編集に使い始めた。用例収集が最初の目的であったが、共起関係・コロケーションの観察や分析も編纂の作業に大いに役立つと分かり、言語位相の区別なども大量コーパスのおかげで辞典の計画において指摘した特徴に合った答えも見つけることができた。

3. で述べたように対訳辞典の編纂に必要とされているいくつかの点について BCCWJ の貢献を以下に取り上げてみたいと思う。

[5] 西川一郎編集・和田和彦監修『小学館和伊中辞典』, 東京, 小学館, 2008 年第 2 版.

4. 1.

『和伊大辞典』に収録されるべき日本語の見出し語とそれと関連する用例、つまり、辞典に収録したい「日本語」は厳密に定義されていないが、漠然として「現代の一般の日本人が日常使う（読む・書く・聞く・話す）時に必要とする日本語」とする。もちろんその「一般の日本人」の像はあまりにも不確定でもあるし、同じ「日本人」でも場合によって新聞をただ読むこともあれば、仕事で専門的な（その分野に関わっている話者でなければ理解しがたい程度の）文章を書くこともある。実用的に、かつ、具体的に言えば、読者・利用者の視点から考えるとこの辞典には極端なサブジャンルや極端な専門用語を除いて、日本語を理解したいときに読解・聴解の有効な手段になりうる「日本語の代表性を有する言語」を収集したいと思う。現代日本語に限るので、一世代、Rey-Debove 氏が言うように語彙の分野で 60 年間程度使用されたことばを収録しておきたいと思う[6]。しかし、廃語でも現代で歴史用語として、あるいは、文学の作品で使われている文語的な語彙も収録し、発話に利用されていないが文章の引用で登場するような単語も辞典に載せる予定である。

そのような前提のもとで BCCWJ の構成を考えて（その均衡と代表性については既にいろいろ議論されてきたが[7]）、従来の対訳辞典に収集されている用例と比較すれば辞書編纂（特に対訳辞典）の水準を相当に上昇させると思う。

まず、De Mauro 氏が言う「lingua dell'uso」（（実際）使用言語）がサンプリングの形で提供されているに違いない[8]。その構成には任意的、主観的な選択も行われている（ジャンルの割合、収集されていないジャンル、等）と指摘されるかもしれないが、辞典編集の具体的な作業の中では、見出し語に関連する BCCWJ より得られる用例は多様なコンテストによるものであることによって、見出し語として収録されている単語の意味範囲をほとんど網羅している。辞典の用例として適切でない文（長すぎる、前後関係がなければ完成文でも意味が分かりにくい、事実の背景を知らなければ意味が通じない、など）もあるが、それは「自然言語」と「見本言語」との差の問題である。従って、編集の際、用例の再編集（長さ、従属節の多い複文などの調整）が必要になってくるが、現在の段階では、多くの場合では BCCWJ で用例の収集がそれぞれの項目を完成させている。

コーパス言語学の視点から言えば、BCCWJ に基づくことによって、伝統的な辞書編纂学と違って、辞典に収録したい単語の意味を抽象的に抽出しないで、むしろそれを様々な文脈・場面（コ・テクスト、コンテクスト）で分析が可能になり、その意味を釈義的（パラフレーズ的）に解釈し、その目標言語に違った形での釈義（翻訳）も与えることができる。しかも、「中納言」のソフトで（特にそのためだけに開発されているわけでもないが）コーパスから得られる文章をただの例文としてではなく、その中から探し出す単語を体系的に考察することも出来、他の言語要素との意味関係の研究も出来ると思う[9]。

4. 2.

3. に触れたように、原則として対訳辞典には POS が記述されていないが、非日本語母語

---

[6] 60 年間とは「synchronie pratique」（実用的共時）の概念によるものである。Rey-Debove Josette, "Le domain du dictionnaire", *Langages*, XIX, 1970, pp. 3-34.

[7] 山崎誠,「代表性を有する現代日本語書き言葉コーパスの設計」, 国立国語研究所(2006)所載, pp.63-70.

[8] Tullio De Mauro, "Introduzione al Grande Dizionario Italiano dell'Uso", *Grande Dizionario Italiano dell'Uso* 第一巻に所載, VII-XLII 頁, Torino, UTET 出版社 1997.

[9] コーパス言語学と辞書編纂学との関係については Wolfgang Teubert,"Corpus Linguistics and Lexicography", *International Journal of Corpus Linguistics*, 6 (Special Issue), 2001, pp. 125-153 を参照されたい。

話者には大切な情報である。「中納言」のようなソフトでBCCWJを利用すると、一貫性のあるPOSの記述があり、辞典の編集の具体的な作業に携わっている研究員には大変ありがたい機能である。もちろん、イタリアで使われている日本語の文法用語はBCCWJのコーパスにタグされている品詞と必ずしも一致するのではないが、その品詞の分類と名称は一定しているので、それぞれの翻訳や調整も出来るのである。一方、日本の国語辞典の品詞分類や用語にも相違がないわけでもないので、対訳辞典の編集中にその整理は必要でもあるし、目標言語との関係を考えた上で実用的な文法の記述用語を独自に使用する必要が生まれてくると思う。

4. 3.

非日本語母語話者にとっては文の統語関係、特に名詞と助詞、動詞の活用形と名詞との意味関係（連体修飾節など)、述語と助詞との関係の変種・選択などの記述は非常に大切である。対訳辞典は、言うまでもなく、文法書ではないが、提供されている用例の文についての補助的な説明がないと用例の効果自体が衰えてしまうと思われる。しかも、対訳辞典を利用する読者の一部は学習者であることを考えれば、辞典の教育的な価値も無視できないものである。というのは、教科書で習ったことと、実際には発せられる言葉とは一致しないこともあるからである。例えば、願望を表す助動詞「タイ」にかかる名詞につく助詞は教科書では規範的には「が」とされ、「を」は補助的な選択として紹介する教科書が多い。しかし、量的だけ見てもBCCWJで前者の出現は1471件に対して後者は9386件なのである（Nガ動詞-タイ　vs　Nヲ動詞-タイ)。また、移動動詞（飛ぶ、歩く、走る、など）とその移動が行われている場所を表す名詞に必ず「を」という助詞が付くと言われているが、実際には意味合いによって違うようなので、大量データのコーパスを頼りにすれば、統計的にどの形の方が代表的であるかということも指摘できるし、実用的な注釈も付け加えられる。

1. じつはな、みょうなおねがいでまいりましたが、おたくのヘリコプターを、ちょっと一日、わたしどもの町の上**で**飛ばせていただきたいのですが。
2. ストレッチしながら草がむしれていったり、木が削れていったり、家が建っていったりした方がずっと面白い。ただただベルトコンベアの上**で**歩いているよりは。
3. 大きな公園**で**歩こう。hinaを、好きなように歩かせてあげたかったのでcohiの幼稚園の間に、大きな公園に行ってきました。
4. 最短の道**で**歩くのであれば、古代東海道の丁字路を左に曲がって、すぐ右折するとＪＲ吉原駅に辿り着きます。
5. 坂道**で**歩く速度が落ちるうえ、みんな記念撮影するので混雑もピークに達する。
6. 子供たちが歓声を上げながらその橋の上**を**走ったり自転車に乗ったりしていた。
7. 彼が僕の脇を擦り抜けて、パンの並んでいる棚のほうへ歩きだしたとき、僕は、白河庭園**で**走って逃げていったあの中年男の、身体の調子が万全ではないような、ちょっとふらつくような足取りを、そこに見たのだ。
8. ドイツ**で**走っている車も、日本を走っているものに較べ、随分とズッシリとした走りをしている。
9. こうしてダブルオーは、人類が造った二足歩行ロボットで、初めて公の前**で**走っ

たロボットとなったのだ[10]。

量的には確かに「N　ヲ　ハシル」の方が「N　デ　ハシル」の出現より多いが、大量コーパスによる用例のより精密な選択が出来、それぞれの意味上の違いの説明も辞典の中で出来ると思われる。

なお、(試験に)「ウカル」の意味で使う移動動詞の「トオル」もその目的語に「に」という助詞が付くとされて教科書に表れる文型であるが、６人の日本語講師のネイティブに伺ったところ、その方が正しいという人が多いことが分かった[11]。助詞を抜いてコーパスから選んだ文を提出したら、「に」を入れたがっている傾向がはっきりとあらわれたのである。しかし、その用法は現実には違うと BCCWJ による分析では立証が出来る。

また、(試験に)「受かる」と（法律を国会で)「立法させる」の意味での既述の「通る」とそれに係る名詞に付く助詞との関係をもう一つの例としてあげてみたい。

多くの対訳辞典には「試験**に**通る」と「国会**を**通る」、それぞれ「に」と「を」とにしている。しかし、下記の例はその「原則」に反しているが、６人のインフォーマントのほとんどが「原則」に従った回答を出したのである[12]。

1. 大学病院で実際に診療に当たっているのは教授や助教授ではなく、学生か卒業まもない歯医者の卵です。**国家試験を通れば**、学生でも歯科医としての資格を持っていますから、いくらでも診療できます。
2. さて，最後に，これから弁護士を目指す人にとって１番の関心事は，「司法試験に合格できるか」を除けば，「果たして自分は弁護士になってきちんと食べていけるであろうか」あるいは，「苦労して**試験を通った**は良いが，それなりの見返りは得られるであろうか」というところであろう。
3. 労働者派遣法の改正案で日雇いは原則禁止になりそうだ。今回の案が秋の臨時**国会に通れば**日雇い労働者らは困ってしまうだろうな。
4. 今まで著作権侵害の動画と音楽をアップロードする行為が違法だったのを、それに加えてダウンロードも違法にするものです。法案はできていて、**国会に通れば**施行される見込みです。

上述の６人の日本人のインフォーマントは外国人に「正しい」日本語を教えているという強い意識もあるかもしれないが、言語の規範とその実態との相違や言語使用のヴァリエーションを考察が出来るのも大量コーパスの長所である。

このような場合には対訳辞典にその文型の変種による意味・用途の違いを指示しなければならないと思う。コーパスによるシンタックスと意味との関わり合いの分析により辞典に載せられる言語の実用の貴重な情報も得られるのである。

上述のように対訳辞典の日本語の例文はその目標言語への誘導手段に過ぎないので一つずつの例は十分とされているが、非日本語母語話者の読者にはたりないと思われる。日本人のインフォーマントに「通る」の例文を提供したら、「優秀な成績で司法試験を通ってき

[10] 3. と 4. の文は BCCWJ 外の例です。それぞれ(http://cohinata146.blog90.fc2.com/blog-entry-192.html)と(http://www7b.biglobe.ne.jp/~fujisan60679/umi01.html)より抽出した文である。

[11] 「試験にトオル」（試験にウカル）と「予選をトオル」（ある段階を通過して次の段階へ進む）での「に」と「を」の出現には混交の傾向があるかもしれない。

[12] 1. と 2. には 6 人も「に」を選択し、3. と 4. に 6 人の中 5 人も「を」を選んだ。

た人が多い」のようなこの文では他の「通る」の文と違って「を」の助詞があることによって「通過」的な過程の意味合いが表れ、すなわち試験という段階を通って、次の段階へ進むというふうな解釈が行われた。意識的に固定された「試験に」というパターンも数多くの例文を考察することによって新しい結論を生み出したのである。辞典の用例収集にも議論・討論のもう一つのヒントを与えてくれた。

最後に、言語位相のことについても簡単に述べたいと思う。あらゆる外国語に触れる際に把握しにくい要素の中で言語位相は確かにその一つのである。外国語を習った人には同じ経験があると思う。初めて耳にした、目にした言葉は誰に対しても、どんな場面でも使えるか使えないかと躊躇することがある。酒場で知人から冗談交じりに言われて、記憶に残ったことばを翌日学会や改まった場で使ってみたら、横目でにらまれ「外人だから」かろうじて許された経験がない人は少ないと思う。「ご馳走になる」、「食事する」、「ご飯を食べる」、「めしを食う」、などはみなイタリア語に「mangiare」に訳されては非日本語母語話者の辞典利用者にあまり役に立たない。BCCWJ の利用ソフトの「中納言」だけではその言語位相が指摘されないが、出典の提示と広い文脈の観察でどのような場面、どのようなジャンルで使用されているか見当が付き、辞典に読者用のタグも加えられるのである。

5. 将来の課題と均衡コーパスの進展

コーパス自体よりもその検索ソフトと関係がある課題であるが、品詞のタグをより細かく出来れば共起関係をもっと簡単に、精確に調べることが出来る。たとえば、4. 3. に述べたように移動動詞に係る名詞は品詞としては名詞であるが、その移動が行われる場所を指しているので特別なステータスの名詞である。すなわち、「場所名詞」に付く「助詞」と一定の「動詞」をソフトで指定することが出来たら、上述の項目について豊富な分析が出来ると思う。今更、この話は無意味かもしれないが、日本語のシンタクスに合ったパラメータも将来的に加えることが可能になればより高度な分析も出来るようになると思う。

なお、辞書編集に携わっている者としては、やはりコーパスのジャンルの均衡を改めて考える必要があるのではないかと思うこともある。というのは見出し語に備える用例をコーパスから抽出する文を探すに当たって、たまにジャンル、内容に偏った例が表れているような気がすることもあるからである。BCCWJ のコーパス構築の過程では中立的な原理に基づいているそうであり、「流通」というサブコーパスの要素も考慮することによって、よく読まれている書籍が選択されてくるのである[13]。

そのせいかもしれないが、場合によって特別なニュアンスもない語彙素を探したらその文章のジャンルが限られてしまう現象が起こるようなのである。例えば、「スボマル」（窄まる）の用例を調べたところ、49 件の中 25 件も性的な描写に関する内容の文が抽出され、特段に道徳問題にならないにせよ、辞典には容易に使えるものではないと思う。コーパスの分析を始める以前、特定の意味範囲でも使用されうると想像できる単語（サシイレル、アイブ、ウチマタ、など）の使用の「偏った」例の割合も無視できないものである。コーパスの構築は、客観的な基準に従って行われたが、その類の書籍や内容・テーマが用例上の比重が高いことも興味深いものである。データの母集団の扱い方は中立的で客観的である以上、その結果の内容を価値観で分析して判断することはないが、なぜそのような内容がよく抽出されるか、本当に一般的な読書傾向を反映しているか、または、語彙のレベル

[13] 前川喜久雄,「特定領域研究『日本語コーパス』のめざすもの」,『日本語コーパス全体会議総括班報告』, 2006.9.9, pp. 1-8.

で調べられた単語の代表的な用途を示しているか、と言うような課題が残されている。厳密な意味でのコロケーション・共起関係よりも広義でのコテクストの分析にも役立つものでありながらも、コーパスの代表性の側面にも改めてスポットを浴びせる必要があるのではないかと思う。

コーパス言語学は、計量言語学と違って研究の対象である言語を単なるデータに還元させるものではないとすれば[14]、言語使用の「環境」にもその研究範囲を広げることも出来ると思う。BCCWJ の精細なデータ（出典、著者、サブコーパス、年代など）により、そのような「環境」の分析と記述も可能になると思われる。

最後になるが、大量均衡コーパスとその検索ソフトが出来たため、より精密な言語の分析、言語使用の考察、また豊富な用例のサンプリングも今まで考えられなかったほどの円滑さでできるようになり、海外などで行われている対訳辞典のプロジェクトの進歩を強く支援することになったということをここで証言したい。また、国立国語研究所で KOTONOHA のプロジェクトが完成したら、一層の発展が可能になるものと期待している。

[14] Wolfgang Teubert,“Corpus Linguistics and Lexicography”, *International Journal of Corpus Linguistics*, 6 (Special Issue), 2001, p.129.

uetsukeru【植え付ける】[uꜜetsɯ̥kᵝɾɯ] *v.t.*2 **I.** ① piantare ¶植木を～¶道の両側に花の種を植え付けたり、木を植え付けたりして、道を美しくするような仕事を始めた。Sono iniziati i lavori per abbellire la strada mettendo a dimora semi di fiori e piantando alberi ai suoi lati.② inoculare (un virus, battere) 病原菌を～ inoculare un germe patogeno ¶ツツガムシ病原菌を植え付けられた患者の中で死亡している者があったらしい。Sembra che alcuni pazienti sottoposti all'inoculazione del battere dell'*orientia tsutsugamushi* siano deceduti.**II.** inculcare; instillare; radicare ¶イメージを植え付ける inculcare uno stereotipo, radicare un'immagine¶彼は息子に友人の失敗をひそかに喜ぶような歪んだ競争心を植えつけた。 Ha inculcato al figlio un distorto spirito di competizione tale da farlo godere dei fallimenti degli amici | 自分は自力ではなにもできないという思いを子どもの心に植えつけています。Ha radicato nei figli l'idea che non riescano a far nulla con le proprie forze. | 1972年代の二回の石油危機と、頻発した公害・環境問題とは、人々の心に資源・エネルギーの有限感を植え付けた。Le due crisi petrolifere degli anni '70, e i problemi di danni da inquinamento ambientale, hanno minato nella gente la convinzione dell'illimitatezza delle risorse energetiche.

図１　『和伊大辞典』草稿

うえつける　植え付ける　**1**《植える》piantare; 《移植する》trapiantare, mettere a dimora ◇植え付け piantatura㊛;《移植》trapianto㊚, messa㊛ a dimora ¶ぶどうを植えつける piantare un terreno a vigna
2《思想などを》seminare, infondere, inculcare ¶彼は労働者のあいだに不満の種を植えつけた. Ha seminato lo scontento fra i lavoratori.

図２　小学館『和伊中辞典』2008[2]

うえつける【植え付ける】 Ⅰ❶〔植物を〕plant ¶畑にトマトの苗を植え付けた He *planted* tomato seedlings in the field. ❷〔病原菌などを〕¶チフス菌をねずみに植え付ける *inoculate* rats with typhus
Ⅱ〔心に刻み付ける〕plant; implant; fix; root (► fix, root は受け身で使われることが多い) ¶彼のその日の行動は彼女の心に強い不信の念を植え付けた His conduct that day *planted* [*implanted* / *instilled*] a strong distrust in her heart. / After what he did that day, a strong distrust of him *was rooted* [*fixed*] in her heart.

図３　小学館『プログレッシブ和英中辞典』2009[3]

副詞による括弧構造とその文脈における役割について

アンドレイ・ベケシュ（筑波大学）

On Adverb Based Bracket Structures and Their Role in the Context

Andrej Bekeš (University of Tsukuba)

０．はじめに

日本語において、副詞がいわゆる助動詞、接続助詞、取り立助詞を含む様々な要素と体系的に共起するという現象が以前から研究者の注意を引いてきた（南 1993、工藤 2000、Narrog 2009 などを参照）。このような遠隔共起によってできた構造（以降、括弧構造、 Bekeš 2008）は今まで主としてその意味的な側面が論じられてきたが、括弧構造が文脈に生起する動機、または本来表している意味以外にも文脈において何らかの働きを果たしているかどうかなどはあまり論じられていないようである。

本研究では、推量モダリティ、条件モダリティ、そして限定の取り立てと関わっている副詞による括弧構造に焦点を当て、その文脈における役割を曖昧性の減少、そして談話マーカー（DM）としての働きという観点から検証していく。

１．副詞による括弧構造

１．１　括弧構造の定義

Bekeš (2008) などで論じたように、次のような構造、即ち要素が体系的に共起し、括弧のように、開ける要素と括弧を閉じる要素が呼応するものを、括弧構造という。括弧構造の例としては、数学に限らず、筆記言語において自然発生的にできた様々な「括弧」記号（例えば（…）、［...］、【...】等）や句読点（例えば日本語の「」、『』、英語のいわゆる smart quotes “...”、スペイン語の　¿…? など）がそうである。

また、句読点に類似するが、言語の要素から構成されている括弧構造もしばしば見かけられる。そのよい例は副詞である。副詞は日本語に限らず、様々な言語で多様な要素と共起し、括弧構造を作り出すのである。まず（１）で日本語の例を幾つか挙げる。

（１）a　推量

… どうやら後手が余しているらしい。　BCCWJ （新聞）

… たぶん日本公開の際も同様だろう。　BCCWJ （新聞）

b　アスペクト

… もう夜の闇がたまっていた。　BCCWJ （新聞）

同様に、類型論的に近い韓国語にも似た現象が見られる。（２）では Seo et al. (2006) による推量モダリティを表す副詞 *ama* （多分）と文末表現 *eulgeosi-da* との共起の例をあげる。

（２）　*Bi-ga o- ass-euni　geu-neun ama usan-eul gajyeoga-ass-eulgeosi-da.*

雨が 来 たので　彼 は　多分 傘 を　取っ　た <未来><助詞>.

また、類型論的に異なっている中国語でも、「已経 --了（*yijing--le*）」、「已経--過（*yijing--V-guo*）」のように、時態副詞と助詞、助動詞との体系的共起が見られる (cf. Smith and Erbaugh 2005)。

１．２　推量副詞、条件副詞、限定副詞による括弧構造

推量副詞による括弧構造の例は１．１．の（１）ａで挙げた。工藤 (2000) は、書き言葉の豊富な資料分析に基づき、推量副詞の共起を論じている。

条件副詞では、「もし」、「たとえ」、「かりに」など、様々なものがあり、「たら」、「(r)e-ば」（以降、「ば」と表す）、「と」、「ても」、「なら」などの要素と共起するが、本研究では、コーパスでの頻度からみて、考察を「もし」と、「たら」、「ば」との共起に依る括弧構造に限定する（下記の（３）ａを参照）。

限定副詞も「ただ」、「たんに」、「ひたすら」、「もっぱら」などと多様であり、限定の取り立て助詞の「だけ」、「しか」、「ばかり」と共起する。本研究では考察をコーパスでの頻度がもっとも高い「ただ」と「だけ」の共起による括弧構造に限る（下記の（３）ｂを参照）。

（３）ａ　**条件**
もし機会があったらと食い下がり….　BCCWJ （新聞）
ｂ　**限定**
方舟は、動力が無くて、ただ浮いているだけ。　BCCWJ （新聞）

２．データと仮説

２．１　データ

括弧構造による曖昧性の減少という文脈効果、それに談話マーカーとしての働きは両方とも、時間的制約の中で行われている話し言葉に先に現れやすく、時間的制約が少ない、より保守的でありがちの書き言葉では現れにくいと予想できる。両者を比較するために、利用するデータには、BCCWJの新聞サブコーパス（以降「BCCWJ新聞」、国立国語研究所 2009）に加え、日本語話し言葉コーパスの講義・講演のサブコーパスCSJ−A（国立国語研究所 2005）、それに親しい人同士による即興的会話コーパスのNUJCC（大曽他 2003）という、二つの話し言葉コーパスを用いる。

２．２　括弧構造の特徴

副詞による括弧構造を含め、括弧構造に見られる「開きの要素」と「締めくくりの要素」の共起関係は、工藤 (2000) が論じているように、一種の呼応関係であるが、義務的ではないので、準文法的という。基本的には確率論的と考えた方が適切であろう。例えば推量副詞と文末表現の場合、その共起が一定の確立で体系的に起こっているということが、書き言葉（工藤 2000、Srdanović Erjavec et al. 2008）、および話し言葉（Bekeš 2008）の様々なコーパスで裏づけられている。

有標性。まず、コーパスにおける括弧構造と、副詞だけか、または副詞が呼応する単独の要素だけ分布頻度を見よう。副詞による括弧構造が、単独のモダリティ表現または取り立て助詞より著しく少ないということが分かる（表１ａでは[ただ-だけ]/[だけ]の割合、表１ｂでは[もし-たら]/[たら]の割合、[もし-ば]/[ば]の割合の列を参照）。同じ結果は推量モダリティの場合にも見られる（Bekeš 2008 参照）。Halliday (1991) によれば、同等な役割を果たし

二つの表現形式の場合、著しく少ない方が確率論的な意味で有標であるという。

表１ａ　諸コーパスにおける「ただ」と「だけ」の分布

コーパス	推定サイズ（語の数）	ただ の頻度	だけ の頻度	ただ-だけ の頻度	[ただ-だけ]/[だけ] の割合	「ただ」に続く「だけ」の経験的確立
BCCWJ新聞	1.000.000語	91	872	15	0.017	0.165
CSJ−A	35.000語	23	463	9	0.019	0.281
NUJCC	700.000語	39	966	11	0.011	0.220

表１ｂ　諸コーパスにおける「もし」と「たら」及び「ば」の分布

コーパス	サイズ：語の推定数	もし の頻度	たら の頻度	ば の頻度	もし-たら の頻度	もし-ば の頻度	[もし-たら]/[たら] の割合	[もし-ば]／[ば] の割合	「もし」に続く「たら」の経験的確立	「もし」に続く「ば」の経験的確立
BCCWJ新聞	1,000,000	31	344	1264	15	6	0.044	0.005	0.48	0.19
CSJ−A	35,000	52	208	560	4	22	0.020	0.04	0.08	0.42
NUJCC	700,000	117	4273	1477	66	12	0.015	0.008	0.56	0.10

表１ａ，ｂからは括弧構造のもう一つの確立論的な側面が見られる。副詞と呼応する要素は副詞が出現すれば、必ず出てくるのではなく、コーパスによって異なっているが、偶然による共起に比べて一定の高い確率で共起する（表１ａの「ただ」に続く「だけ」の経験的確立、表１ｂの「もし」に続く「たら」の経験的確立、「もし」に続く「ば」の経験的確立の列を参照)。講義・講演のコーパスであるCSJ-Aでは限定を表す「ただ」—「だけ」の共起、そして、条件を表す「もし」—「ば」の共起が高い確率で起こっている。一方、ジャンルとして正反対のはずである、BCCWJの新聞サブコーパスとNUJCC会話コーパスでは、「もし」—「たら」の共起の高い確率が目立つ。Bekeš (2008) で報告されているように、推量副詞によるモダリティ表現との共起にも同様な現象が見られる。

有標制は括弧構造の意味的余剰性からも推測できる。それぞれの副詞と呼応する表現との意味的余情性は異なっているが、例えば高い確立で観察できる推量副詞とモダリティ表現の体系的な共起関係（上述の工藤 2000、Srdanović Erjavec et al. 2008、Bekeš 2008 を参照）がかなりの意味的余剰性を示唆していると考える。このような意味的余剰性は Grice (1975) に基づいて、括弧構造の有標性として解釈可能である。例えば、表１ｂに見られる「もし」と「たら」、「もし」と「ば」の確率の非常に高い共起において、副詞と呼応要素とが意味的に重複する部分があり、括弧構造が単独の条件表現より余剰的であるといえる。

２．３　仮説

仮説１。前節から、括弧構造がその有標性のため、文脈において囲んでいる分節に対してなんらかの特異な、談話マーカー的な働きをするということが予測される。

仮説２。副詞と高い確率で共起要素と副詞との組み合わせが一端括弧構造として認定されれば、常識的な意味での括弧が成立する。このことにより、条件、限定などのスコープがより明示的に与えられるようになる。ここからは、長い括弧の場合、括弧構造の使用が文脈における曖昧性の減少となんらかの関わりがあるという仮説を立てることができる。

３．　仮説の検証

３．１　三つのコーパスにおける括弧の長さの分布

NUJCC、CSJ-AとBCCWJ新聞という三つのコーパスで、括弧のタイプの長さの分布を分析した結果を下記の表２にまとめた。

表２　NUJCC、CSJ-AとBCCWJ新聞における括弧タイプとの長さの分布

	NUJCC	CSJ-A	BCCWJ新聞	NUJCC	CSJ-A	BCCWJ新聞	NUJCC	CSJ-A	BCCWJ新聞
括弧の長さ（形態素数）	もし-たら 頻度	もし-たら 頻度	もし-たら 頻度	もし-ば 頻度	もし-ば 頻度	もし-ば 頻度	ただ-だけ 頻度	ただ-だけ 頻度	ただ-だけ 頻度
＞20	1（.2%）	0（0%）	0	0（0%）	**・3（14%）**	0	0（0%）	0（0%）	0
16-20	2（.3%）	0（0%）	0	0（0%）	**・5（25%）**	0	0（0%）	0（0%）	1（6.7%）
11-15	4（.6%）	0（0%）	0	1（10%）	2（ 9%）	**・1（17%）**	0（0%）	0（0%）	2（13.3%）
6-10	**・18（27%）**	**・2（50%）**	**・5（33.3%）**	**・5（45%）**	0（ 0%）	**・2（33%）**	**・7（41%）**	**・3（33%）**	**・3（ 20%）**
1-5	**・41（62%）**	**・2（50%）**	**・10（66.7%）**	**・5（45%）**	**・12（54%）**	**・3（50%）**	**・10（59%）**	**・6（67%）**	**・9（ 60%）**
合計頻度	66（100%）	4（100%）	15（100%）	11（100%）	22（100%）	6	17（100%）	9（100%）	15（100%）
メジアン	**4.5**	**5.5**	**5**	**6.5**	**5.0**	**5.5**	**5.0**	**4.5**	**4**

表２で分かるように、どの副詞による括弧でも、その分布は短い方に偏り、全ての括弧の長さのメジアンが4形態素～6.5形態素の範囲内に収まっている。さらに、メジアンだけでなく、実際の括弧構造の分布も、CSJ-Aにおける「もし-ば」を除いて、すべて、ほぼ10形態素以下という短い括弧に偏っている。短い括弧構造の場合はその認定が曖昧である恐れがないので、NUJCCでも、BCCWJ新聞でも、短い括弧構造が用いられるのは何らかの特異な状況の現れであると推測できる。NUJCCのような即興的会話では難しい事柄を題材とする会話が少ないので、曖昧性の減少と関わるような長い括弧の必要性は低い。その上、新聞は書き言葉であり、曖昧と感じる箇所も再確認が可能なので、曖昧性を積極的に減少するストラテジーの必要性が認められにくいと思われる。そうであるならば、表１の分析結果で観察できるように、長い括弧は不要である。表２で唯一の例外はCSJ-Aというモノローグの講義・講演のサブコーパスである。正確な情報伝達が要求されているCSJ-Aのデータでは、内容の難易度が新聞の平均的な記事より高いだけでなく、口頭で発せられるゆえに再確認が不可能なので、仮説２に基づいて曖昧性の減少ストラテジーが必要になると予測できる。実際、CSJ-Aの場合のみ、「もし-たら」よりは主観性が低く、客観的な論理関係を表すのに相応しいという「もし-ば」が長い括弧構造を作る例が目立つ。実際、「もし-ば」の22例中12例は長さが5形態素以下で極めて会話的である。一方、8例は長さが全て15形態素を越えているもので、さらに残りの2例も長さが10-15形態素の間で中間的性格を帯びている。この長い括弧群の使用が曖昧性の減少説による上記の予測と一致していると考えられる。なお、括弧構造と曖昧性の減少との関わりを思わせる現象は、推量副詞の場合、即興的会話における聞き手の相槌などの干渉のタイミングと括弧構造の生起確立との相関においても観察されている（Bekeš 2008, Ch.5 を参照）。

３．２　「ただ」と「もし」による括弧とその文脈に見られる用法

本節では、NUJCC、CSJ-A、BCCWJ新聞における「ただ」と「もし」による括弧とその文脈における用法を検証する。「ただ」と「もし」による括弧の文脈上の振る舞いを観察するためには各々のコーパスで、実例を、括弧をキーに用いて、先行文脈と後続文脈がそれぞれ400字とうい範囲でKWIC検索をかけ、抽出した。括弧と前後文脈との関わりを主題の持続および主題の展開という視点（砂川 2005）、そして Mann & Thompson (1988) による argumentative structure の視点から分析を試みた。ただし、本研究のこの段階では、少ないデータで全体的な傾向を把握するため、細部に至る分析を集約するという、より粗雑な記述的な分析カテゴリを利用することにした。

「ただ-だけ」による括弧の分析。分析の結果は下記の表３にまとめた。「ただ-だけ」の使用は大きく「導入」、「主張・結論」、「主張・対比」、そして解釈困難な曖昧な用法に分け

られる。すべてのコーパスで頻度がもっとも多いのは「導入」である。「導入」では、「ただ-だけ」による括弧は、段落または短い文章の初めに現れ、「ただ」の限定取立て助詞としての対比機能に依存しながら、対比項と関わっている内容が後続文脈に導入される。さらに、単独の「だけ」と異なって、「導入」の分布が段落の初め、そして「主張・結論」の分布が段落の終わりという特異な位置に偏るということは、限定副詞「ただ」と「だけ」との組み合わせの有標性によると考える。

一方、「主張・対比」はNUJCCとBCCWJ新聞のみで現れるが、いずれの場合も、相対頻度は低い。このような分布は、NUJCCでは即興的会話に見られる時間的制約、またBCCWJ新聞では多様なジャンルの寄せ集めであることによると推測できる。

さらに、CSJ-Aでは、曖昧な例を除けば全ての用法が「導入」に集中している。原因は説明的ジャンルである講演・講義が必要とするレトリックにあると考える。

曖昧なケースはBCCWJ新聞では少ない(7%弱)。これとは対照的に、NUJCCとCSJ-Aではそれぞれ35%と25%で、大きな割合を占めている。その原因は時間的制限であると考える。

表３　NUJCC、CSJ-AとBCCWJ新聞における「ただ-だけ」による括弧と文脈における働き

コーパス　→		**NUJCC**	**CSJ-A**	**BCCWJ新聞**
文脈における働き・特徴		**頻度(%)**	**頻度(%)**	**頻度(%)**
導入	主題についての主張を伴う談話への要素導入	6 (35%)	9 (75%)	5 (33.4%)
主張・結論	先行文脈に依存する重要な論点の主張・結論	2 (12%)	---	5 (33.4%)
主張・対比	主張の焦点（先行文脈の要点)、 同時に後続文脈と 対比	1 (　6%)	---	2 (13.3%)
	先行文脈との対比の焦点	2 (12%)	---	2 (13.3%)
曖昧な例	「ただ」が接続詞か福祉課の 判断が曖昧、括弧の役割が曖昧	6 (35%)	3 (25%)	1 (　6.6%)
	TOTAL	**17** (100%)	**12** (100%)	**15** (100%)

「もし」と「たら」、「ば」による括弧の分析。「ただーだけ」の場合と同様な方法で、「もし-たら」、「もし-ば」の括弧をキーに、前後に幅400字の文脈幅で三つのコーパスから抽出したデータをさらにコーパス本体のテクストと照らし合わせながら分析した。分析の結果は下記の表４にまとめた。

表４　NUJCC、CSJ-AとBCCWJ新聞における「もし」による括弧と文脈における働き

コーパス　→		**NUJCC**		**CSJ-A**		**BCCWJ新聞**	
	括弧　→	もし-たら	もし-ば	もし-たら	もし-ば	もし-たら	もし-ば
↓　文脈における働き・特徴　↓		**頻度(%)**	**頻度(%)**	**頻度(%)**	**頻度(%)**	**頻度(%)**	**頻度(%)**
導入	文脈における重要な状況	33 (50%)	3 (27%)	1 (25%)	---		---
	談話への新しい主題導入	14 (21%)	2 (18%)	2 (50%)	---	4(24%)	2(33%)
論理的派生	完全な叙述としての実現	---	---	1 (25%)	10 (45%)		---
論拠の 重要部分	構成要素に焦点が当てられた 条件としての実現	---	---	---	4 (18%)		---
	先方照応としての実現	---	---	---	3 (14%)		---
主張・結論	先行文脈に依存する重要な 論点の主張・結論	---	---	---	2 (　9%)	10(67%)	4(67%)
結論の提供	会話相手への協力として	4 (　6%)	3 (27%)	---	---		---
引用	引用内に出現	9 (14%)	1 (10%)	---	---	1(7%)	(導入と重複)
定型表現	先方照応としての実現	---	---	---	3 (14%)		---
曖昧な例	役割が不明確	6 (　9%)	2 (18%)	---	---		---
	TOTAL	**66** (100%)	**11** (100%)	**4** (100%)	**22** (100%)	**15**(100%)	**6**(100%)

分析結果をNUJCC、CSJ−A、BCCWJ新聞の順で見ていこう。まず目立つのはコーパスによる「もし-たら」と「もし-ば」の用法の頻度の違いである。主観性が前面に出てきやすい親しい人たちの個人的な即興的な会話であるNUJCCでは「もし-たら」が圧倒的に多い。多様なサブジャンルで様々な話題を取り扱うBCCWJ新聞でも「もし-たら」が「もし-ば」より多いが、その差はNUJCCに見られほど大きくない。一方、学術講義・講演であるCSJ-Aでは、より客観性が高い「もし-ば」の用法が圧倒的な多数を占めている。

それぞれのコーパスにおける用法の分布にもさらに興味深い違いが観察できる。まずNUJCCでは「もし-たら」と「もし-ば」の分布における二極化が明確に現れている。「もし-たら」は「導入」の2タイプに圧倒的に集中しながら、「引用」、そして、共同発話的な性格を帯びている「結論の提供」にも現れる。即ち、NUJCCのような即興的な会話ではやり取りの結論的部分が条件副詞による括弧構造の形で共同発話として現れることがある。なお、「もし-ば」の総数は「もし-たら」の十分の一にしか及ばないが、その用法には「もし-たら」と類似した二極的分布が見られる。

一方、CSJ-Aでは、「もし-たら」と「もし-ば」の分布は対照的である。「もし-たら」の総数は「もし-ば」の五分の一弱にしか及ばないが、その四分の三が「導入」に集中し、のこりの四分の一（1例）は「論理的派生」に用いられている。また、22例もある「もし-ば」は「導入」では用いられておらず、最も多く用いられているのは「「論理的派生」（半分弱）、及び「論拠の重要部分」の両タイプ（三分の一弱）である。その他、「主張・結論」（一割弱）及び「定型表現」（約七分の一）にも用いられている。

続いて、BCCWJではNUJCCと類似して、「もし-たら」と「もし-ば」の分布において、同様な傾向が観察できる。分布は「導入」（「もし-たら」24%、「もし-ば」33%）と「主張・結論」（「もし-たら」、「もし-ば」両方67%）とほぼ二極的である。「もし-たら」の一例はまた引用として提示されおり、「引用」として分類される。

最後に、単独の「だけ」、「たら」、「ば」の分布は、「ただ-だけ」、「もし-たら」及び「もし-ば」の括弧構造と異なり、これらの括弧構造のように、「導入」、「主張・結論」などの特定の用法・働きへの著しい偏りは観察されていない。むしろ全ての用法・働きにおいて相対的に均等に分布されているようである。単独の「だけ」、「たら」、「ば」の分布はそれぞれのコーパス本体のテクストにおいて確認し、本稿では敢えてデータとして提示しなかった。

４．まとめ

観察してきた各々の括弧構造の分布の違いは偶然ではなく、恐らく検証した三つのコーパスのデータが発せられた場面および状況と関わっているものであると見なす事ができる。ある言語表現を理解するためには、その用法の、特定の場面・状況のタイプとの関連づけが必要である。同じような場面・状況で発せられたデータを集めたコーパスはその数もサイズも限られている。本研究で対象としたコーパスも例外ではない。観察した実例が少ないため、それをさらに幾つかのカテゴリーに分けようとすると、各々のカテゴリの例が極端に少なくなるここが避けられない。そのため、本研究で扱ったデータにおいて、統計的な有意義性をある程度主張できるのはNUJCCのデータだけである。従って、この段階では、本研究は質的、発見的及び記述的な範囲内に止まらざるを得ない。

また、３．２節で見てきたように、三つのコーパスにおいて観察してきた括弧構造の分布の偏りに比べて、単独の「だけ」、「たら」、「ば」の分布は特定の働き・用法にさほど偏らず、

より汎用的に用いられているようである。この単独形式の汎用性は無標の現れであると見なせる。これに対して、限定的に特異な状況で用いられがちの括弧構造は有標と見なすことができる。ここからは、第２節で括弧構造の有標性に対して別の根拠で立てた仮説の妥当性が確認されたということが言えるであろう。

各々のコーパス、各々の括弧構造で観察した分布の偏りは短絡的に談話マーカー的な性格と関連づけるのは困難であり、この段階ではそれぞれのジャンルにおけるストラテジーとしての解釈が妥当であろう。ただし、NUJCCに見られる「導入」（71%）に偏る「もし-たら」の用法は、ストラテジーから、談話マーカーのへの第一歩として捉えることも不可能ではなかろう。

最後であるが、KOTONOHA計画全体及びBCCWJでは、日本語の研究者には今までなかった研究法をその根底から変える可能性が与えられた。言語現象を広い文脈で観察するアプローチを取った本研究から見れば、今後の発展として望みたいのは、コーパス構築が均衡コーパスの範囲を超え、ジャンル別の多量均質データのコーパス構築を新たな目標とすることである。

参考文献

Bekeš, Andrej (2008). *Text and Boundary: A Sideways Glance at Textual Phenomena in Japanese* (Razprave FF), Ljubljana: ZIFF, 150pp.

Fischer, Kerstin (2006). “Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume”. In Fischer, Kerstin (ed.) .

Fischer, Kerstin (ed.) (2006). *Approaches to discourse particles*. Amsterdam: Elsevier.

Grice, H. Paul (1975). ”Logic and Conversation”. In Peter Cole and Jerry L. Morgan (eds.) *Syntax and Semantics*, Vol. 3, Speech Acts, pp. 41–58. New York: Academic Press.

Halliday, Michael A.K. (1991). “Corpus studies and probabilistic grammar”. In: K. Aijmer & B. Altenberg (eds.) *English corpus linguistics*, pp. 30-43. London: Longman.

工藤 浩 (2000)「副詞と文の陳述的なタイプ」、仁田義雄・益岡隆志編『日本語の文法３ モダリティ』(161-234)、岩波書店、東京。

前川喜久雄（2006)「特定領域研究『日本語コーパス』のめざすもの」 「日本語コーパス」全体会議総括班報告(2006.09.09) p.1-8 http://www2.ninjal.ac.jp/kikuo/tokutei_H18_1.pdf

Mann, William C. and Sandra A. Thompson (1988). “Rhetorical Structure Theory: Toward a functional theory of text organization.” *Text* 8 (3): 243-281.

南 不二男 (1993).『日本語文法の輪郭』、大修館書店、東京。

Narrog, Heiko (2009). *Modality in Japanese: the layered structure of the clause and hierarchies of functional categories*. Amsterdam: John Benjamins.

Seo, Young Ae, Sang Kyu Park, and Key Sun Choi (2006). “Structural Disambiguation of Korean Adverbs based on Correlative Relation and Morphological Context”. *ETRI Journal*, Vol. 28. No. 6, pp. 803-806. (Accessed at http://etrij.etri.re.kr/Cyber/BrowseAbstract.jsp?vol=28&pg=803, July 2010).

Smith, Carlota S. and Erbaugh, Mary S. (2005). "Temporal interpretation in Mandarin Chinese", *Linguistics.* Volume 43, Issue 4, pp. 713-756.

Srdanović Erjavec, Irena, Andrej Bekeš and Kikuko Nishina (2008). “Distant collocations between suppositional adverbs and clause-final modality forms in Japanese language corpora”. In Takenobu

Tokunaga and Antonio Ortega (eds.) *Large-scale knowledge resources : construction and application ; Third International Conference on Large-Scale Knowledge Resources, LKR 2008, Tokyo, Japan, March 3-5, 2008 ; proceedings*, (Lecture notes in computer science, 4938), Berlin: Springer, pp. 252–266.

砂川由里子 (2005). 『文法と談話の接点―日本語の談話におけるシュアぢ展開機能の研究』、くろしお出版、東京。

スルダノヴィッチ・イレーナ、ベケシュ・アンドレ、仁科喜久子 (2009)「コーパスに基づいた語彙シラバス作成に向けて--推量的副詞と文末モダリティの共起を中心にして」、『日本語教育』、142：pp. 69-79。

コーパス資料

国立国語研究所 (2009). BCCWJ（新聞サブコーパス）「『現代日本語書き言葉均衡コーパス』領域内公開データ（2009年度版）、国立国語研究所研究開発部門言語資源グループ、2009年7月31日。1,237のサンプル、2.132.121字。

国立国語研究所 (2005, 2007). “Corpus of Spontaneous Japanese”「日本語話し言葉コーパス」(CSJ)の講演、講義サブコーパスCSJ-A. 試し版、ひまわり HPからダウンロード。87.766字。大曾美惠子 (2003). 名大会話コーパス (NUJCC)科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成13年度～15年度、研究代表者：大曾美惠子）データ量：約101時間、1.874.245字。

謝辞

この研究の一部は国際交流基金のフェローシップによって支えらている。同基金、フェローとしての受け入れ先を提供した東京工業大学、暖かく受け入れて下さった同大学留学生センターの仁科喜久子教授に厚く感謝申し上げる。

基本動詞ハンドブック執筆への BCCWJ の利用
ー辞書執筆用コーパスシステム NINJAL-LagoWordProfiler の開発ー

プラシャント・パルデシ（国立国語研究所言語対照研究系）
赤瀬川 史朗（Lago 言語研究所）

Using BCCWJ for the Compilation of the Handbook of Usage of Basic Verbs in Japanese: The Development of NINJAL-LagoWordProfiler for Dictionary Making

Prashant Pardeshi (National Institute for Japanese Language and Linguistics)
Shiro Akasegawa (Lago Institute of Language)

1. 日本語学習者用基本動詞用法ハンドブックの作成プロジェクト

コミュニケーションの基本単位となる文の骨格を決める重要な要素の一つが述語としての動詞である。日本語を外国語として学ぶ学習者にとって、日本語の運用能力を向上させるために、使用頻度の高い基本動詞の体系的な学習が不可欠である。国立国語研究所では日本語研究の成果を日本語教育に応用する目的で本稿の筆頭著者がリーダーを務める共同研究プロジェクト「日本語学習者用基本動詞用法ハンドブックの作成」が 2009 年 10 月から実施されている。本プロジェクトでは、基本動詞の全体像、つまりその統語的振舞い（格枠組み、受動形の有無、アスペクト的な特徴など）、意味拡張（意味ネットワーク）、自他の対をなすカウンターパートおよび類義語との対比等々、を把握することが効率的な学習に必要なものであり、さらに、日本語の体系だけでなく、母語の体系と日本語の体系間の類似点や相違点を理解することは学習効果を最大限に引き伸ばすことに役立つと位置付けられている。そこで本プロジェクトの目標は、言語学、日本語学、日本語教育学、対照言語学、第二言語習得研究、辞書編纂学、認知言語学、コーパス言語学などといった様々な研究分野の最新の知見を取り入れ、世界の日本語学習者の体系的かつ効率的な学習に役立つ「日本語学習者用基本動詞用法ハンドブック」のプロトタイプを開発し、それに基づいて、日・中、日・韓、日・英、日・マラーティー語の試作版を作成することである。

本プロジェクトで開発を目指す「日本語学習者用基本動詞用法ハンドブック」のプロトタイプの最大の特色はコーパスを利用することによる見出しの執筆である。具体的には、見出し語の語義、語義頻度、統語的および意味的な共起環境などといった情報を、自然に使われている日本語データの集成である『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）から抽出し、客観的な分析を参考にしながら見出し執筆を行う。以下本プロジェクトで使用される辞書執筆用コーパスシステム NINJAL-LagoWordProfiler（以下、NINJAL-LWP）の BCCWJ への実装の過程について報告し、NINJAL-LWP の辞書執筆用への応用例をいくつか紹介する。

2. 辞書執筆用コーパスシステム LagoWordProfiler の開発の経緯

LWP は赤瀬川が辞書執筆・編集用に 2005 年から開発を続けているコーパスシステムである。ブラウザ上で動作するウェブアプリケーションである。しかし、その構成はシンプルで、HTML ファイルと XML ファイルの集合体である。そのため、ウェブ上での利用のほか、ファイルをローカルに置いてオフラインで利用することも可能になっている。対応言語は日本語のほか、英語などの欧米語に対応している。また、日英、英日などのパラレルコーパスにも対応可能である。

HTML ファイルと XML ファイルの集合体であるのは、LWP ではリアルタイムの検索を行わないからである。LWP ではあらかじめ頻度、コロケーション、用例などのデータを抽出し、それらの情報を XML ファイルに記録する。ユーザはブラウザでその結果を順にたどりながら文字通りブラウズする。これまでコーパスは検索するものであったが、LWP ではコーパスは検索の対象ではなく、ブラウジングの対象である。言葉を換えれば、LWP はコーパスを「検索する」ためのツールではなく、コーパスを「読む」ためのツールといってよい。

では、なぜ「読む」ためのツールにする必要があるのか。それは辞書執筆・編集と深いかかわりがある。辞書制作はまさに時間との闘いである。コーパスの登場によって、辞書制作は劇的な変化を遂げた。それまでは執筆者が長年にわたって蓄積してきた用例カードに頼っていたが、コーパスがそれに取って代わった。しかし、いいことずくめではなかった。大量の情報に埋もれることなく、コーパスからいかに有益な情報を取り出すかという新たな課題が生まれた。

辞書制作でコンコーダンサーなどの汎用ツールを利用すると、それまで以上に制作に時間がかかってしまうという皮肉な結果にもなりかねない。実際、コーパスを利用した初期の辞書編集ではこうした失敗が数多くあった。時間的制約というのは、辞書制作の現場では重要なファクターになる。LWP では、執筆者や編集者は検索した結果をたどるだけでよいので、検索結果を待つという時間の無駄を最大限省くことができる。

さらに辞書執筆・編集でコーパスを活用する上でもう一つ重要な点がある。記述の対象となる見出し語の振舞いをコーパスを利用して網羅的に示すことである。人はどうしても珍しいものに目が向く。言語現象においても、珍しい語や用法に興味が向かいがちである。過去の用例カードにはそうした視点から作られたものが多い。だが、辞書ではむしろ慣用的な語や用法を示すことにその第一義がある。コーパスを辞書制作に利用するためには、そうした一般的な言葉の実態を余すことなく客観的に示すツールが不可欠になる。

LWP はこうした辞書制作の現場が要求する「時間的制約」と「一覧性」という二つの要件を満たす形で開発を進めてきた。すでに複数の英和辞書や和英辞書のプロジェクトで活用してきた実績がある。今回の日本語の基本動詞ハンドブックプロジェクトは日本語辞書での初めての LWP の利用になる。

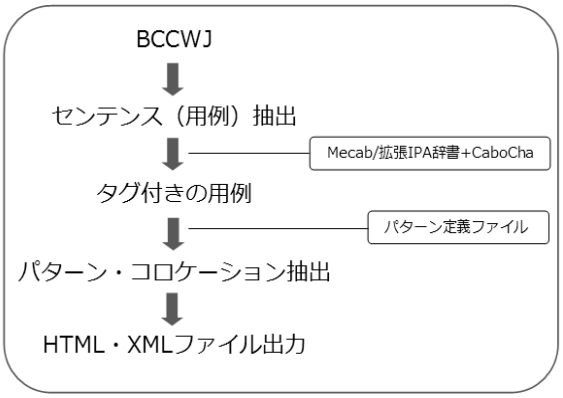


図 1　LWP+BCCWJ の開発の流れ

3. NINJAL-LWP の開発の流れ

今回の BCCWJ の LWP 実装にあたっては、領域内公開データ（2009 年版）を使用した。本予稿執筆時点では開発中のため、以後述べることは最終的には一部変更される可能性があることをあらかじめお断りしておきたい。

3.1. センテンス抽出

全体の開発の流れを示したのが図 1 である。LWP では用例はセンテンス単位で示されるため、まずコーパスからセンテンスを抽出するのが最初の作業になる。何をもってセンテンスと認定するかはいろいろな尺度があるが、今回は、用例という観点から見てふさわしくないと考えられるもの、具体的には記事の見出しや図表の解説、数値などを列挙したデータなどは抽出対象から排除した。

また、今回は流通実態サブコーパス、非母集団サブコーパス、生産実態サブコーパスのうち、可変長の XML ファイルとして提供されているデータから speech タグで囲まれた発話部分を地の文とは区別して収集し、地の文と発話文での差異を調査できるように工夫した。各サブコーパスから抽出したセンテンス（用例）数を表 1 に示す。

表 1　サブコーパスごとのセンテンス（用例）数

サブコーパス	ファイル数	本システムでの記号		センテンス数	語数	センテンス当たりの語数
流通実態SC, 非母集団SC, 生産実態SC	23,642	地の文	WT	2,143,944	55,990,804	26
		会話文	SP	390,800	8,772,271	22
国会会議録	159	MD		116,823	5,046,656	43
Yahoo!ブログ	12,700	YB		134,574	2,440,539	18
Yahoo!知恵袋	45,725	YC		300,197	5,402,259	18
		合計		**3,086,338**	**77,652,529**	**25**

3.2 タグの付与

次に 3.1 で抽出した用例に対して、形態素解析と係り受け解析を行なった。解析器と辞書の選定については様々な角度から検討した。重要なファクターとなったのは、代表表記の問題と文字コードの問題であった。これら二つのハードルを越えるために最終的に選択した方法は、IPA 辞書に新たに代表表記の情報を加え、未収録語も相当数追加した上で(以下、この辞書を拡張 IPA 辞書と呼ぶ)、MeCab+CaboCha で解析させる方法である。以下、この二つの問題点について述べる。

3.2.1 代表表記

LWPでは、見出し語ごとに抽出した情報をまとめてパターンやコロケーションを分類して表示する仕組みになっている。そのため、さまざまな書字形や語形を一つの見出し語、つまり代表表記[1]に集約することはきわめて重要な意味を持つ。

英語ではこうした作業のことを lemmatization（レマ化）と呼んでいる。例えば、赤瀬川が

[1] UniDic では語彙素が代表表記に相当する。

作成した英語のレマ化用の辞書では、動詞の cancel は以下のように、@cancel_VB（@は見出し語の cancel という意味）にまとめている。（それぞれ_VB は動詞の原形、_VBP は現在形、_VBZ は三単現、_VBD は過去形、_VBN は過去分詞、_VBG は現在分詞を表す。）

cancel_VB, cancel_VBP, cancels_VBZ, canceled_VBD, canceled_VBN, canceling_VBG, cancelled_VBD, cancelled_VBN, cancelling_VBG => @cancel_VB

動詞 cancel はイギリス英語とアメリカ英語で語尾の l を重ねるか重ねないかの違いがあるので、原形も含めて、9 種類の形が一つの見出し語にまとめられることになる。

では、日本語ではどうか。似た意味を持つ「取り消す」を例にとってみよう。次は拡張 IPA 辞書に代表表記を追加するために、今回作成した動詞の基本データ（レキシコン）である。各フィールドは「｜」によって区切られている。

取り消す｜五段・サ行｜トリケス｜トリケス｜とりけす,とり消す,取りけす,取り消す,取消す｜｜｜可能動詞:取り消せる

先頭のフィールドが代表表記で、5 番目のフィールドが書字形である。書字形は 5 種類あるがすべて基本形で示されている。基本形以外の 6 種類の活用形も含めれば、全体で 5×7=35 種類の形が一つの見出し語「取り消す」にまとめられることになる。これを見ても、英語と比べて、日本語ではずいぶんとレマ化に手間がかかることが分かる。

3.2.2 文字コード

当初、形態素解析と係り受け解析は、代表表記に対応したJUMAN+KNPで行なう計画であった。だが、実際に作業を始めてみると、大きな問題にぶつかった。文字コードである。JUMANの文字コードはWindowsではShift-JIS[2]に固定されている。ところが、BCCWJの書籍を中心としたデータではUnicode文字が数多く使われている。JUMANで解析するとこうした文字は文字化けを起こす。それに比べて、Web上のYahoo!知恵袋やYahoo!ブログのデータではUnicode文字はほんの一部に限られたため、大きな問題とはならなかった。

一方、MeCabで使われているIPA辞書は文字コードを指定してコンパイルできるため、UTF-8 のエンコーディングを使用すれば、上記のJUMANのような問題は避けることできる。だが、もともとUnicode以前の文字コードで作成された辞書であるため、UTF-8 に変換しても当然ながらUnicode文字の単語は収録されていない。文字種が豊富な書籍データに対応するために新たに多数の語を登録した。全収録数は、オリジナルのIPA辞書 [3]の収録数（約 39 万語）の約 2 倍に相当する 79 万 9 千項目となった。

3.3 パターン・コロケーションの抽出

次に、3.2 で作成した品詞情報と係り受け情報の付いた用例データをSQLデータベースに

[2] Linux では EUC-JP に固定されている。

[3] IPA 辞書（mecab-ipadic-2.7.0-20070801）。

登録し、同時に、コーパスで使用されている全動詞（一般の動詞とサ変動詞[4]）の頻度表を作成した。LWPでは、パターンとコロケーションの抽出する際に、あらかじめ品詞ごとにパターン定義ファイルを用意しておき、その抽出条件をもとに見出し語ごとに抽出する仕組みになっている。今回のプロジェクトでは、基本動詞ハンドブックの執筆用に動詞の定義ファイルを大幅に拡張して、動詞の振舞いをくまなく調査できるようにした。

3.3.1 定義ファイル

定義ファイルは独自のフォーマットで記述する。以下は「名詞＋について＋動詞」のパターンを定義した箇所である。フィールドは「｜」で区切っているが、実際のデータではタブ区切りである。

B036｜名詞+複合助詞 <<｜…について%(TARGET)%｜%(POS=名詞;CHECKLINK)% %(基本形=について;POS=助詞)% %(TARGET;GETLINK)%

先頭から順に、「パターン ID」、「グループ名」、「パターン名」、「パターン」を表す。最後のパターン・フィールドで、抽出する条件を定義する。「%(…)%」で 1 語を表す。%(POS=名詞)%は任意の名詞を表し、後に続く CHECKLINK はその名詞の係り先の情報を抽出するための指示である。次の%(基本形=について;POS=助詞)%は、複合助詞の「について」を表している。最後の%(TARGET)%は抽出する見出し語、ここでは動詞の定義ファイルなので、抽出する動詞を表す。後に続く GETLINK は、先ほどの名詞の係り先がこの動詞かどうかチェックするための指示である。

LWP では、この CHECKLINK と GETLINK を組み合わせることで、係り受け情報をパターン抽出に最大限に利用している。例えば、「考える」という動詞を抽出する場合、このパターン一つで「就職について考える」、「就職についてじっくり考える」のどちらも抽出できるようになる。

今回作成した動詞用の定義ファイルにはこうしたパターンが約 200 件ほど登録されている。つまり、一つの動詞について 200 種類の検索を行なっているのと同じ計算になる。これをすべての動詞（今回の NINJAL-LWP では見出し語単位で約 18,000 種の動詞があった）について繰り返しながら抽出作業を進めていく。

3.4 HTML・XML ファイル出力

3.3 で抽出したデータは見出し語ごとのテキストファイルとして保存されている。LWP では、パターンやコロケーションの頻度情報を表示するパネルが 2 列と用例を表示するパネルが 1 列ある。この最終段階では、頻度やコロケーションや用例などの情報を保存した XML ファイルを見出し

図 2　LWP の画面構成

[4] サ変可能名詞にサ変動詞「する」が後続するパターンを指す。

語ごとに出力し、システムが最終的に完成する。

4. NINJAL-LWP の実際

では、実際にこのシステム構成や特長を説明する。

4.1 画面構成

図 2 はLWPの画面構成である。見出し語パネル、頻度パネル、用例パネルの 3 つの部分から構成される [5]。

見出し語パネルには、見出し語が 50 音順に並ぶ。動詞の部分をクリックすると、右側にその動詞の頻度とコロケーションと用例の詳細な内容が表示される。頻度パネルは 2 列でなっており、左側はパターンの大まかな分類、右側にはそれぞれのパターンのコロケーションが示される。用例パネルには、右側の頻度パネルで選択したコロケーションの用例が示される。

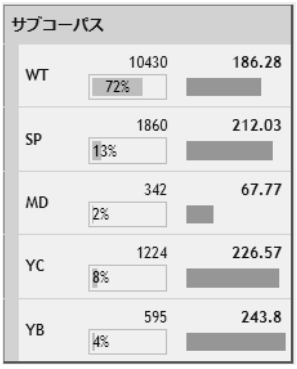


図 3　サブコーパスの分布

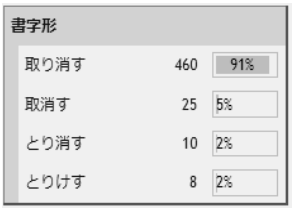


図 4　書字形の比率

4.2 基本情報の表示

3.3 で作成した動詞頻度表には、動詞のさまざまな頻度情報が含まれている。サブコーパスごとの頻度のほか、書字形ごとの頻度、さらにサブコーパスごとの書字形の頻度、活用形の頻度などがある。

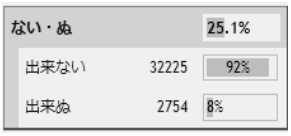


図 5　否定表現の比率

こうして解析されたデータは頻度パネルにある基本情報の画面で確認できる。図 3 は動詞「読む」のサブコーパスごとの頻度と全コーパスに占める割合（左）と 100 万語当たりの頻度（PMW）（右）を表している（WT などの記号については表 1 を参照）。これを見ると、サブコーパスによって使用頻度のばらつきがあることが分かる。

また、図 4 は、先ほど例に出した「取り消す」の書字形の割合を示したものである。「取り消す」という表記が 9 割を占めていることが分かる。さらに書字形については、サブコーパスごとの頻度や割合も表示される。

この他、動詞が「れる・られる」「せる・させる」「ない・ぬ」に接続する頻度の割合も示され、その動詞が受動・使役表現や否定形としてどの程度使われているのかをおおよその傾向を知る目安となる。図 5 では、動詞「出来る」の否定表現の割合が 25.1%であることを示している。うち、92%が「出来ない」の形であることが分かる。

図 6　パターン頻度

4.3 パターンの頻度

頻度パネルには、LWP が提供する一番重要な情報が集約されている。定義フ

[5] 図 2 は LWP+BCCWJ ではなく、青空文庫で作成したデモ版。

ァイルによって抽出したパターン・コロケーションの頻度情報がこの画面に表示される。

図 6 は動詞「動かす」のパターン頻度情報である。頻度情報の表示には「頻度順」（左）と「グループ」（右）の 2 種類があり、タブによって切り替えができる。

頻度順では、パターンの頻度の高いものから順に表示される。グループ別では、動詞の前に「名詞＋助詞」が来るパターン（体を動かす）、「名詞」後続するパターン（動かす力）、動詞を修飾する副詞（少し動かす）、他の動詞との共起（動かしていく）など、グループごとの頻度が参照できる。

また、今回新たに「名詞＋複合助詞＋動詞」というパターンも一部追加した。複合辞については、近年、自然言語処理の分野でも研究が盛んになっている（土屋 2006）。本来の内容語の意味で使われる場合もあるため、単に複合助詞を形態素解析辞書に追加するという単純な方法ではかえって処理の精度を損なってしまう。共起する名詞の属性を特定するための規模の大きい学習用コーパスの必要性を強く感じる。今回の「名詞＋複合助詞＋動詞」パターンでは、機能語として使われる確率の高いものに限定して定義ファイルに追加し、精度と実用性の両立を図った。

4.4 コロケーション（共起語）の頻度

頻度パネルの右側には、それぞれのパターンに含まれる共起語のリストが表示される。リストには共起語の単純頻度のほか、MI スコアと log ダイス（Lexical Computing Ltd. 2007）が表示される。ソートができるので、複数の統計値を比較しながら、共起関係の強さを客観的に確認することができる。コロケーションを示した箇所をクリックすると、該当する BCCWJ の用例が右の用例パネルに表示される。

4.5 用例の確認

辞書の執筆では、各種の頻度情報によって全体像を把握できたとしても、やはり最後は用例で確認したい場合が少なくない。LWP では、作例をつくるときの参考になるように、図 7 のように、用例はセンテンスの短い順に並べている。

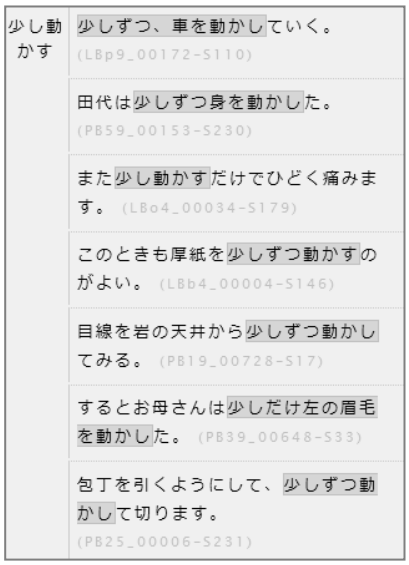


図 7　用例表示

4.6 文脈の確認

それぞれの用例の最後には出典が表示される[6]。出典箇所をクリックすると、前後の文脈を示すダイアログが表示される。談話分析をはじめ、代名詞の指示対象を確認したり、特殊な表現が使用されている理由を探ったりする場合など、利用する場面は多い。

4.7 シノニム比較機能

二つの動詞の振舞いの違いを多角的に調べるのがシノニム比較機能である。図 8 は、ヲ

[6] 図 7 は開発中のものでファイル名が表示されている。

格で「走る」と「駆ける」の違いを調べたものである。数値はlogダイスを表し、数値の大きいものほど動詞との結びつきが強いことを示している[7]。

名詞を～	走る	駆ける
千里を～	8.4	-
道を～	7.8	3.1
廊下を～	6.4	5.4
階段を～	4.6	9.1
脳裏を～	-	7.6

図８　シノニム比較機能

5 まとめ

以上、基本動詞ハンドブック執筆用の NINJAL-LWP システムの概要について述べてきた。辞書制作の現場では、時間的制約からコーパスを利用する際は見出し語の共起情報の一覧性が要求されること、また、そうした要件を満たす形で LWP を開発してきたことを示した。

また、実際の開発では、一つの見出し語に集約するためのレマ化の作業の重要性を明らかにした。2007 年に公開された UniDic は語彙素と語形・書字形を対応させる形で代表表記に対応しているため、当初は MeCab/UniDic での解析を考えていた。しかし、現状では CaboCha には未対応のため、今回は断念せざるを得なかった。今後、MeCab/UniDic+CaboCha という組み合わせで解析できるようになれば、さらに精度の高い抽出が可能になってくると期待している。

本システムは 2011 年 3 月に完成し、4 月以降、執筆の現場での運用が始まる。今後は、執筆者からのフィードバックを十分活かして、さらなるシステムの改善につなげていきたい。

また、LWP を辞書執筆以外の用途にも広げて、冒頭でも述べたようにコーパスを「読む」ためのツールとしての活用にも取り組むつもりである。そうした活動を通して、日本語の言語研究や教育に少しでも寄与できればと考えている。

文献

Erjavec, I., Erjavec, T., Kilgarriff, A. (2008). A web corpus and word sketches for Japanese 『自然言語処理』, 15:2, pp.137-159.

国立国語研究所(2001).『現代語複合辞用例集』 国立国語研究所.

国立国語研究所(2008). 『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. 2.0（国語研究所内部報告書）.

砂川有里子、駒田聡、下田美津子、他(1998). 『日本語文型辞典』くろしお出版.

土屋雅稔、宇津呂武仁、松吉俊、他(2006). 「日本語複合辞用例データベースの作成と分析」『情報処理学会論文誌』, 47: 6, pp.1728-1741.

伝康晴、峯松信明、小木曽智信、他(2007). 「コーパス日本語学のための言語資源－形態素解析用電子化辞書の開発とその応用－」『日本語科学』, 22, pp.101-122.

Lexical Computing Ltd. (2007). Statistics used in the Sketch Engine. http://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf

関連 URL

日本語学習者用基本動詞用法ハンドブック作成プロジェクト: http://www.ninjal.ac.jp/research/project/bc/youhoujiten/

[7] 開発中のため、数値は Sketch Engine の JpWaC による。

計画班研究発表

3月16日（水）　10:00〜11:40

多義語における意味の分布

▶山崎 誠

拡張モダリティタグ体系の設計とBCCWJへのアノテーション

▶松吉 俊、佐尾 ちとせ、乾 健太郎、松本 裕治

UniDic 2：設計と実装

▶小木曽 智信、伝 康晴

日本語研究とインターネット

▶田野村 忠温

多義語における意味の分布

山崎 誠（データ班班長：国立国語研究所 言語資源研究系）†

Distribution of Senses in Polysemy

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

１．はじめに

本研究は多義語の使用実態のうち量的な分布の実態を調査したものである。具体的には，(1)原義と比喩的意味との割合はどのようになっているか，(2)多義語を構成するそれぞれの意味が品詞が転成した場合にも維持されているかどうか，(3)活用形などの語形により意味の現れ方が異なるかどうかについてケーススタディ的に調査したものである。

２．多義性の研究

多義語の意味の量的な分布や，語形と意味との関連は，辞書の意味記述において「(まれ)」や「(多くは～の形で)」などの形で用法の多寡を示す情報として現れている。しかし，具体的にどれくらいの量的な分布になっているかは具体的な調査を俟たないと分からない。

形容詞に関しては，橋本・青山(1992)の用法調査を受けて，宮島(1993)，丹保(1997)が連用用法と連体用法における意味の現れ方の違いを指摘している。李ら(2007)の調査は，動詞のル形，タ形，テイル形などに多義の使用傾向の違いが現れるかどうか分析したもので本稿のアプローチに近い。また，文章中での多義の出現傾向については Gale et als.(1992)，山崎(2010)など結束性に基づく分析がある。

３．データと処理方法

本稿で使用したデータは『現代日本語書き言葉均衡コーパス』（以下，BCCWJ と略す。）である。検索は Web 版アプリケーション「中納言」を利用した。収録データ及び語数は以下のとおりである[1]。

データ	語数
書籍（生産）	2,415 万語
書籍（流通）	2,908 万語
雑誌	207 万語
新聞	76 万語
白書	489 万語
ベストセラー	368 万語
Yahoo!知恵袋	516 万語

「中納言」で検索される結果は短単位であるため，基本的に他の語と複合語を形成しない用法が抽出され，複合語のうち最小単位の 1 回結合に該当するものは抽出されない。ただし，複合語のうち，接辞と結び付くものや文法的複合動詞を形成するものは接辞や複合動詞後項が切り出されるため，前項要素として抽出される。今回の分析はそのような条件の

† yamazaki@kokken.go.jp

[1] 語数は短単位で数えたもの。空白・記号・補助記号は含んでいない。短単位とは，ほぼ形態素に相当するところの最小単位の 1 回結合までを許す言語単位である。詳細は小椋ほか(2010)を参照。

もとで行ったものである。

４．原義と比喩の割合

表 1 は，多義語のうち，原義と比喩的意味との関係が捉えやすいものを任意に選んで頻度を調査したもので，比喩の割合の高い順に掲出した。

表 1 原義と比喩の割合

	頻度（割合%）		出現サンプル数（割合%）		1 サンプル当たりの平均頻度	
語	原義	比喩	原義	比喩	原義	比喩
割り切る	0(0%)	244(100%)	0(0%)	225(100%)	0	1.08
履き違える	0(0%)	29(100%)	0(0%)	27(100%)	0	1.07
続投	2(7.1%)	26(92.9%)	2(9.1%)	20(90.9%)	1	1.3
射止める	12(20.3%)	47(79.7%)	10(17.5%)	47(82.5%)	1.2	1
追い風	23(21.7%)	83(78.3%)	15(20.6%)	58(79.5%)	1.53	1.43
降板	8(22.2%)	28(77.8%)	7(25%)	21(75%)	1.14	1.33
不透明	72(25.7%)	208(74.3%)	38(17.8%)	175(82.2%)	1.90	1.19
受け皿	52(26.5%)	144(73.5%)	38(27.7%)	99(72.3%)	1.37	1.46
逆風	19(29.2%)	46(70.8%)	14(28.6%)	35(71.4%)	1.36	1.31
割り切れる	44(30.8%)	99(69.2%)	18(15.9%)	95(84.1%)	2.44	1.04
落とし穴	69(31.5%)	150(68.5%)	22(14.8%)	127(85.2%)	3.14	1.18
脱線	49(35.5%)	89(64.5%)	37(33.6%)	73(66.4%)	1.32	1.22
ハードル	88(37.6%)	146(62.4%)	8(6.3%)	119(93.7%)	11	1.22
燃え尽きる	63(45.7%)	75(54.4%)	54(55.1%)	44(44.9%)	1.17	1.71
抱え込む	182(46.0%)	214(54.0%)	166(46.1%)	194(53.9%)	1.10	1.10
浮かび上がる*	151(50.3%)	149(49.7%)	145(50.4%)	143(49.7%)	1.04	1.04
綱引き	35(52.2%)	32(47.8%)	23(45.1%)	28(549%)	1.52	1.14
握り潰す	39(54.2%)	33(45.8%)	34(53.1%)	30(46.9%)	1.15	1.1
ウイルス*	163(54.3%)	137(45.7%)	87(48.9%)	91(51.1%)	1.87	1.51
焦げ付く	35(59.3%)	24(40.7%)	29(60.4%)	19(39.6%)	1.21	1.26
揉み消す	55(57.3%)	41(42.7%)	52(58.4%)	37(41.6%)	1.06	1.11
代名詞	193(59.4%)	132(40.6%)	50(29.2%)	121(70.8%)	3.86	1.09
処方箋	127(62.3%)	77(37.8%)	72(54.6%)	60(45.5%)	1.86	1.28
冷え込む	69(63.9%)	39(36.1%)	67(64.4%)	37(35.6%)	1.03	1.05
操り人形	19(65.5%)	10(34.5%)	17(63.0%)	10(37.0%)	1.12	1
曲がり角	88(70.4%)	37(29.6%)	73(67.6%)	35(32.4%)	1.21	1.06
壁*	270(90%)	30(10%)	253(89.7%)	29(10.3%)	1.07	1.03
引き出し	479(97.8%)	11(2.2%)	314(96.6%)	11(3.4%)	1.53	1
黒船	115(99.1%)	1(0.9%)	61(98.4%)	1(1.6%)	1.89	1

*「浮かび上がる」「ウイルス」「壁」については，300 例をランダムに選んで調査した。

表 1 から，当然のことであるが原義と比喩の割合はさまざまであることが分かる。辞書上は原義の用法が認められても BCCWJ では観察されない語があったが，だからといって，原義の用法を「まれ」とするのは早計であろう。それは用例数が十分でないばかりでなく，

「続投」の例のように，原義の用法が特定の分野だけで用いられる場合，その分野をカバーしていないデータには現れてこないからである。

表 1 には，1 サンプル当たりの平均頻度を示した。それによると，原義と比喩とで差が認められるものがある。「割り切れる」「落とし穴」「ハードル」「代名詞」は原義の平均頻度が比喩の 2 倍以上ある。これらのうち，「割り切れる」「ハードル」「代名詞」はそれぞれ，原義においては使用分野が限定されるものであることから，特定のサンプルに集中的に現れる傾向があると推察される。表 1 におけるそれぞれの語の平均頻度の平均は，原義で 1.78，比喩で 1.19 であり，比喩で用いられたほうが特定のサンプルへの集中度が低いことが分かる。これは比喩としての用法の方が使用分野の限定がされにくくなっていることを示すものであろう。

５．品詞の転成

本稿では和語の動詞とその連用形転成名詞との関係について扱う。一般に，和語の動詞が多義語である場合，その転成名詞よりも多義語を構成する意味の数が多いと推測される。例えば，『明鏡国語辞典』第 2 版（以下，『明鏡』と略す。）で見ると，「当たる」には自動詞として 17 個，他動詞として 2 個の意味があるが，その名詞形である「当たり」の意味は 9 個（他に造語成分として 2 個）である。このことを踏まえ，本稿では動詞と名詞の意味の出現状況の把握がしやすいよう，それぞれの多義を構成する意味が並行的であるペアを選び考察することにした。

また，本稿では全体的な使用傾向を探るため，検索結果が 200 例を超える場合は，ランダムに選んだ 200 例を分析の対象とした。なお，意味分類の際，誤解析や分類不能の例は対象から外し，正味 200 例を抽出している。

５.１　「戦う」と「戦い」

「戦う」「戦い」は，『明鏡』では動詞の③④の意味が名詞③に合併した形で対応しているため，名詞に合わせて，動詞③④を 1 つの意味として扱った。

動詞「戦う」[2]

①武力を用いて争う。戦争する。交戦する。

②競技・選挙などで優劣を競う。勝負を争う。競争する。

③自分の利益や権利などを守ったり獲得したりするために争う。闘争する。

④身に降りかかる困難な誘惑などを乗り越えようとする。闘争する。

名詞「戦い」

①たたかうこと。戦争。戦闘。

②競争。競技。試合。

③抗争。闘争。「貧苦との―」「労使の―」

表 2 「戦う」「戦い」の意味の分布

意味	動詞	(%)	名詞	(%)
(1)戦争	146	(73.0)	145	(72.5)
(2)競争	19	(9.5)	28	(14.0)
(3)闘争	35	(17.5)	27	(13.5)
合計	200	(100.0)	200	(100.0)

[2] 以下，語釈は『明鏡』からの引用である。意味の理解に必要な場合は例文も挙げた。

表 3 「戦う」「戦い」の意味の分布（データ別）

	動詞			名詞		
データ	(1)戦争	(2)競争	(3)闘争	(1)戦争	(2)競争	(3)闘争
書籍	137	15	33	135	25	25
雑誌	1	3	0	4	1	0
新聞	0	1	0	1	1	2
白書	0	0	0	1	0	0
Yahoo!知恵袋	8	0	2	4	1	0
合計	146	19	35	145	28	27

表 2 からは，動詞，名詞とも意味の分布には差がないことが確認される。同様に表 3 からもデータごとの出現状況は動詞と名詞とでほとんど同じであることが見て取れる。

５.２ 「潤う」と「潤い」

動詞「潤う」とその名詞形「潤い」には全く平行する以下の 3 つの意味が認められる。

動詞「潤う」

①ほどよく水けを帯び（て生き生きす）る。適度に湿（しめ）る。

②恵みを受けて、経済的なゆとりができる。金銭的に豊かになる。

③心にうるおいが与えられる。

名詞「潤い」

①湿りけ。水け。

②経済的なゆとり。

③しっとりとした情趣や精神的な豊かさ。

BCCWJ における意味の分布は表 4 に示すとおりである。

表 4 「潤う」「潤い」の意味の分布

意味	動詞	(%)	名詞	(%)
(1)物理的	61	(45.2)	84	(42.0)
(2)経済的	59	(43.7)	5	(2.5)
(3)精神的	15	(11.1)	111	(55.5)
合計	135	(100.0)	200	(100.0)

表 5 「潤う」「潤い」の意味の分布（データ別）

	動詞			名詞		
データ	(1)物理的	(2)経済的	(3)精神的	(1)物理的	(2)経済的	(3)精神的
書籍	31	50	10	32	5	41
雑誌	23	2	2	50	0	1
新聞	0	0	0	0	0	0
白書	0	0	3	1	0	68
Yahoo!知恵袋	7	7	0	1	0	1
合計	61	59	15	84	5	111

表 4 からは，動詞では「(3)精神的に潤う」意味が少なく，名詞の場合は「(2)経済的な潤い」の意味が少ないことが分かる。意味の分布では動詞と名詞には並行的でない関係が見

いだされる。名詞の「(3)精神的」の用例を見ると，「潤いある」「潤いのある」が 54 例を占めている。これらの形は抽出した中では 1 例を除き「(3)精神的」の意味で使われていた。

表 5 はデータ別に分布を見たものである[3]。ここでは各データの違いが見て取れる。雑誌では，動詞でも名詞でもほとんどが「(1)物理的」の意味であること，白書は，動詞でも名詞でも「(3)精神的」の意味がほとんどであることが分かる。

５.３　「扱う」と「扱い」

「扱う」「扱い」もほぼ平行した形で意味が対応している事例である。

動詞「扱う」

①物などを手で動かしたり、有効に使ったりする。また、道具・機械などを操作する。取り扱う。

②ある一定のしかたで他の人を遇する。

③あるものを（仕事として）取り上げてそれを処理する。また、ある問題やテーマとして取り上げる。

④《「～として―」などの形で》それに相応するものをみなして、物事などを処理する。

名詞「扱い」

①物や道具・機械などを扱うこと。扱い方。取り扱い。

②人を遇すること。もてなし。あしらい。待遇。対応。応対。

③物事を処理すること。処理法。

④それに相応するものとして、～として扱うこと。

表 6 によると，動詞では「(3)処理」の意味が半数近くを占めるのに対して，名詞では「(4)相応」の意味がほぼ半数を占める。これは，名詞のほうに「子供扱い」「特別扱い」などの例が多かったためである。名詞の「(2)待遇」の意味では，57 例中 27 例が「扱いを」という連続として現れていることが特徴的である[4]。

表 6 「扱う」「扱い」の意味の分布

意味	動詞	(%)	名詞	(%)
(1)操作	54	(27.0)	19	(9.5)
(2)待遇	31	(15.5)	57	(28.5)
(3)処理	95	(47.5)	27	(13.5)
(4)相応	20	(10.0)	97	(48.5)
合計	200	(100.0)	200	(100.0)

表 7 「扱う」「扱い」の意味の分布（データ別）

	動詞				名詞			
データ	(1)操作	(2)待遇	(3)処理	(4)相応	(1)操作	(2)待遇	(3)処理	(4)相応
書籍	44	27	81	19	16	51	24	72
雑誌	2	1	4	0	1	2	0	4
新聞	0	0	0	1	1	1	0	1
白書	0	0	4	0	0	0	2	1
Yahoo!知恵袋	8	3	6	0	1	3	1	19
合計	54	31	95	20	19	57	27	97

[3] 書籍（生産）・書籍（流通）・ベストセラーをまとめて「書籍」としている。以下の表も同じ。

[4] このうち 21 例は「扱いを受ける」の形である。

表 7 から，データごとの出現状況では，「扱う」「扱い」ともに，書籍が圧倒的で，そのほかのデータにはあまり出現していないことが分かる。このことは，「扱う」「扱い」とほぼ同じ多義構成である「取り扱う」「取り扱い」が，白書での出現が相対的に多いことと対照的である[5]。

６． 活用形による違い

小林(2008)では，白書に現れた形容詞について活用形ごとの頻度及び肯定・否定，過去・非過去の頻度を挙げ，日本語教育の立場からシラバスに対する提言を行っている。本稿では，活用形ないしは語形と意味とが頻度の上でどのような対応を示すのか観察する。

６.１ 「甘い」

「甘い」は，『明鏡』の①～③（味・香り），④（甘美），⑤～⑪（厳しくない）の３つに分類した[6]。表 8 は活用形ごとの意味の分布である。

表 8 「甘い」の活用形ごとの分布

意味	全体	語幹	終止形	連体形	連用形	連用形－促音便
(1)味・香り	89	17	10	46	14	2
(2)甘美	39	3	1	25	10	0
(3)厳しくない	72	17	7	23	20	5
合計	200	37	18	94	44	7

「甘い」については，『明鏡』の注記に，以下のような記述がある。

「④〔多く連体形で〕心がとろけるように快い。また，愛情こまやかにうちとけている。甘美だ。」

この④は上記の表 8 の(2)に対応する。そこで，(2)の意味の出現状況を見てみると，39 例のうち，約 2/3 が連体形で現れていることが分かる。上の指摘は概ね適正と言えよう。

ここで注目したいのは連用形促音便の形である。連用形促音便は抽出した 200 例ではすべて「甘かった」という語形として実現している。検索結果全体では「甘かった」が 111 例出現しているが，これらの意味分布を見てみると表 9 のようになる。表 9 からは「甘かった」は 80%の割合で「(3)厳しくない」という意味に偏っていることが分かる。

表 9 「甘かった」の意味の分布

意味	用例数	(%)
(1)味・香り	20	(18.0)
(2)甘美	2	(1.8)
(3)厳しくない	89	(80.2)
合計	111	(100.0)

また，連用形の「甘く」は，「(3)厳しくない」の意味の約半数が「甘く見る」という慣用表現であること，「(2)甘美」の例は 10 例中 8 例が書籍（生産）の成人向け描写に使われているという偏りが見られた（残り 2 例は書籍（流通）とベストセラーが 1 例ずつ)。書籍（生産）は，出版リストに基づき選定されたサンプルであるため，このような成人向け書籍も含まれている。この点については，サブコーパスの違いが用法の差として現れる例として

[5] 白書の用例数は「取り扱う」で 201 例（全体 1,080 例），「取り扱い」で 588 例（全体 1,930 例）である。

[6] 具体的な語釈は字数の関係で省略する。

注意したい。

「甘い」の派生語として「甘すぎる」「甘め」を取り上げる。「甘すぎる」は形と意味の対応がきれいに分かれていることが特徴的である。表 10 は検索結果に出現したすべての「甘すぎる」69 例のうち，用例数の多い語形及びその関連の語形 59 例を抜き出して調べたものである。

表 10　「甘すぎる」の語形と意味の分布[7]

意味	甘すぎず	甘すぎる	甘すぎた	甘すぎて	甘すぎない
(1)味・香り	9	0	0	6	5
(2)甘美	0	0	0	1	3
(3)厳しくない	0	22	8	3	2
合計	9	22	8	10	10

表 10 からは，「甘すぎず」が「(1)味・香り」の意味だけに偏り，「甘すぎた」「甘すぎる」が「(3)厳しくない」の意味だけに偏ることが分かる。

「甘い」に程度を表す接辞「め」が付いた形，「甘め」は，表 11 にあるように「(1)味・香り」の意味が多い。「甘い」程度を表すのであるから，全体の傾向に一致することが期待されるが，圧倒的に「(1)味・香り」の意味に偏っている。

表 11　「甘め」の意味の分布

意味	甘め
(1)味・香り	27
(2)甘美	3
(3)厳しくない	3
合計	33

６.２　「戦う」

5.1 で挙げた「戦う」には，同時を表す接続助詞に続く「戦いながら」「戦ってきた」の形に意味の偏りが見られた。「戦う」は全体では 7 割が「(1)戦争」の意味であるのに対して，「戦いながら」の形は約 7 割が「(3)闘争」（「病魔と闘いながら」「偏見と闘いながら」など）の意味，「戦ってきた」も全体の傾向と比べて「(3)闘争」の意味が多くなっていることが分かる。

表 12　「戦いながら」「戦ってきた」の意味の分布

意味	戦いながら	(%)	戦ってきた	(%)	「戦う」全体での%
(1)戦争	14	(23.0)	35	(45.5)	73.0
(2)競争	3	(4.9)	9	(11.7)	9.5
(3)闘争	44	(72.1)	33	(42.9)	17.5
合計	61	(100.0)	77	(100.0)	100.0

６.３　「叫ぶ」

動詞「叫ぶ」には，『明鏡』では次の 2 つの意味が挙がっている。

①大声を出して言う。大声を発する。

②強く主張する。声高に訴える。

[7] 「甘すぎない」には，「甘すぎなくて」「甘すぎません」を含む。

BCCWJ で 200 例を抽出して観察したところ，「(2)主張」の意味は約 10%ほどであった。しかし，未然形「叫ば」に接続する形「叫ばず」「叫ばない」「叫ばれる」については，「(2)主張」がほとんどを占めることが分かった。

表 13 「叫ぶ」の未然形の意味の分布[8]

意味	叫ばず	叫ばない	叫ばれる
(1)大声	11	13	12
(2)主張	1	0	143
合計	12	13	155

7． おわりに

多義性のある和語の動詞とその名詞形とでは意味に平行性が認められてもそれらの出現頻度の傾向は異なるものがあることが分かった。

また，「甘い」の例のように活用形と意味との対応関係についても偏りが認められる場合があった。

さらに，「甘すぎる」「甘め」「闘いながら」「叫ばれる」のように特定の語形が一つの意味に集中しやすい例があることも分かった。

このような現象は，データにおける素材や場面の多寡に依存しているのだろうか。そうであれば，内容や話題が異なるデータで観察すると本稿の結果とは異なる傾向が確認されるだろう。あるいは，意味的な属性と関連する特徴として捉えられるものなのか，今後類似の例などを対象にしてさらに考察を進めたい。

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）による補助を得た。また，本研究は，国立国語研究所の共同研究プロジェクト「テキスト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。

参考文献

William A. Gale, Kenneth W. Church, and David Yarowsky(1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pp.233-237, Harriman, NY.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕(2010)『『現代日本語書き言葉均衡コーパス』形態論規程集第 3 版』（LR-CCG-09-12），国立国語研究所

小林ミナ(2008)「『白書』にあらわれたイ形容詞」，『代表性を有する書き言葉コーパスを活用した日本語教育研究』平成 19 年度研究成果報告書，pp.19-28.

丹保健一(1997)「形容詞の連体,連用,終止用法の出現頻度と意味との関連性をめぐって:「高い」「広い」「寂しい」を例として」，「三重大学教育学部研究紀要 人文・社会科学」48，pp.9-18.

橋本三奈子・青山文啓(1992)「形容詞の三つの用法:終止、連体、連用」，「計量国語学」18-5，pp.201-214.

宮島達夫(1993)「形容詞の語形と用法」，「計量国語学」19-2，pp.94-104.

山崎誠(2010)テキストにおける多義語の意味実現の傾向，計量国語学会第 54 回大会,予稿集 pp.25-30.

李在鎬・鈴木幸平・永田由香・黒田航・井佐原均(2007)「動詞「流れる」の語形と意味の問題をめぐって」，「計量国語学」26-2，p.64-74.

[8] 「叫ばない」には，「叫ばなかった」「叫ばなく（ても，とも，等）」「叫ばなければ」を含む

拡張モダリティタグ体系の設計と BCCWJ へのアノテーション

松吉　俊　（ツール班協力者：奈良先端科学技術大学院大学情報科学研究科）†
佐尾ちとせ（ツール班協力者：奈良先端科学技術大学院大学情報科学研究科）
乾　健太郎（ツール班分担者：東北大学大学院情報科学研究科）
松本　裕治（ツール班班長：奈良先端科学技術大学院大学情報科学研究科）

Design of an Annotation System of Extended Modality of Events and Its Appliation to BCCWJ

Suguru Matsuyoshi　(Nara Institute of Science and Technology)
Chitose Sao　(Nara Institute of Science and Technology)
Kentaro Inui　(Tohoku University)
Yuji Matsumoto　(Nara Institute of Science and Technology)

１．はじめに

言語学から言語処理研究にいたる様々な基礎・応用分野において、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を有効に利用するためには、コーパスに対して様々な情報を付与する（タグ付けする）必要がある。本稿では、文の意味レベルのタグ付けとしてツール班が実施した拡張モダリティのアノテーションについて報告する。

一般に、文章に記述される情報は、単純な命題のみではなく、そこには、命題に対する情報発信者の主観的な態度も記述される。例えば、次の文 i)、ii)、iii)からは、それぞれその次に記述したような書き手の態度を読み取ることができる。

i) 携帯電話に使用するＳＤカードを購入したいのですが、どの位の容量のがいいのですか？　（出典：OC02_03906）
→　ある命題（「書き手がＳＤカードを購入するコト」）が成立することを望んでいる

ii) 地域や状況によってあり方も多様化しているようだ。　（出典：PN1a_00002）
→　ある命題(「地域や状況によってあり方が多様化するコト」)が成立しているであろうことを推量している

iii) 紙幣を預かるときは、特に目と耳の確認を怠らないようにしてください。　（出典：PB26_00004）
→　ある命題(「あなたが目と耳の確認を怠るコト」)が成立することを否定的に評価し、受け手にそれが実現しないように働きかける

命題に対するこのような態度は、言語学においてモダリティと呼ばれ、現在も多くの研究者によって活発に研究が続けられている（例えば、Palmer(2001)、仁田ら(2003)、益岡(2007)など)。

言語処理において、与えられた文章から情報を抽出するにあたり、個々の事象に対して、その述語と項構造を認識するだけではなく、書き手が表明している態度や真偽判断、価値判断などの情報も解析し、その解析結果に基づいて情報を整理することは重要である。なぜならば、文章に記述されている事象が、実際に成立した事実であるのか、それとも、成立しなかったことであるのか、もしくは、書き手がその成立を望んでいるだけであるのか、といったことを自動的に認識することは、質問応答や情報抽出などの応用分野にとって必

† matuyosi@is.naist.jp

須の技術の一つであるからである。これまでに、コーパスに対するアノテーションとして、語の分かち書きと品詞、句や固有表現のチャンキング、文節係り受け、句構造、述語項構造、照応などがタグ付けされてきた。我々は、これらの情報の上に、意味レベルの情報として、事象のモダリティとその周辺情報をタグ付けする。モダリティの情報をタグ付けしたコーパスを基盤として、事象のモダリティを高い精度で自動的に解析するシステムを構築することを目指す。

本稿では、我々が独自に設計した拡張モダリティ（モダリティとその周辺情報）の体系と、それに基づいて実施した BCCWJ へのタグ付けの進捗について述べる。

２．関連研究

言語学において、用語も含めて統一した見解は存在しないようであるが、モダリティは、おおよそ、次のように分類される（以下は、仁田ら(2003)と益岡(2007)による）。

- 真偽判断のモダリティ：断定か、推量かを表す。
- 価値判断のモダリティ：必要か、許可できるかを表す。
- 表現類型のモダリティ：叙述、意志、行為要求、勧誘、疑問、感嘆のいずれかの態度を表す。
- 丁寧さのモダリティ　：普通体か、丁寧体かを表す。
- 伝達態度のモダリティ：聞き手の存在に対する話し手の意識のありようを表す。
- 説明のモダリティ　　：先行文脈との関係づけを表す。

本研究では、１つの事象に対するモダリティに着目するため、他の事象との関係づけを表す説明のモダリティは扱わない。また、他のモダリティに比べ、言語処理において重要度が低いと思われるので、丁寧さのモダリティと伝達態度のモダリティも扱わないこととした。

独自のモダリティの体系を設計するにあたり、表現類型、真偽判断、価値判断のモダリティ、および、その周辺情報のアノテーションを扱う関連研究を調査した。調査結果を表１に示す。

表１　本研究および関連研究が扱うモダリティとその周辺情報

	表現類型	真偽判断	価値判断	肯否極性	真偽アスペクト	態度表明者	時間	仮想性
本研究	○	○	○	○	○	○	○	○
Light ら(2004)	×	○	×	×	×	×	×	×
Rubin ら(2005)	○	○	×	×	×	○	○	×
TimeML (2006)	○	○	×	○	×	×	○	○
Prasad ら(2006)	○	○	×	○	×	○	×	×
Saurí ら(2007)	×	○	×	○	×	○	×	×
Medlock ら(2007)	×	○	×	×	×	×	×	×
Szarvas ら(2008)	×	○	×	○	×	×	×	×
原ら(2008)	○	○	×	○	○	○	○	○
FactBank (2009)	○	○	×	○	×	○	○	○
川添ら(2010)	×	○	×	○	×	○	○	○

モダリティとその周辺情報をタグ付けするための体系、および、その自動解析に関する研究は、近年、主に英語や日本語を対象として進められており、言語処理分野の研究のほか、主に真偽判断のモダリティを対象とした生物医学分野の研究（Light ら(2004)、Medlock ら(2007)、Szarvas ら(2008)）も存在することが分かった。

３．拡張モダリティの設計

３.１　事象

我々のアノテーション対象は、文章に存在するすべての事象のモダリティである。ここで、事象とは、行為、出来事、状態の総称であり、本研究では、益岡(2007)に従い、事象にヴォイス（受動態、使役態、可能態）を含める。

３.２　拡張モダリティの項目

本研究では、文章に存在する事象に対する、以下の６種類の項目をまとめて、事象の拡張モダリティと呼ぶ。

- 態度表明者：対象とする事象の成否の判断や、他者への働きかけや問いかけをしている人物、もしくは、団体。
- 相対時：態度表明時から見た、対象事象の相対的な時間関係。過去・現在のことであるのか、それとも、未来のことであるのかを表す。
- 仮想：仮定された条件の有無。仮想世界の話であるのか、そうでないのかを表す。
- 態度：叙述、意志、働きかけ、問いかけなどの伝達的態度。前章で述べた表現類型のモダリティに相当する。
- 真偽判断：態度表明者による対象事象の真偽判断。対象事象が成立か不成立かを、確信度とともに表す。前章で述べた真偽判断のモダリティに加え、肯否極性とアスペクトの情報を取り扱う。
- 価値判断：態度表明者による対象事象の価値判断。対象事象の成立が望ましいことであるかどうかを表す。

それぞれの項目に対するラベルの一覧を表２の左側に示す。上の<態度>、<真偽判断>、<価値判断>の組が、前章で述べた、表現類型、真偽判断、価値判断のモダリティにほぼ相当する。仁田ら(2003)によると、価値判断のモダリティは、基本的意味の面から、「必要」、「許可・許容」、「不必要」、「不許可・非許容」の４つに分類される。本研究では、これらの意味を独自の体系により<態度>に記述し、態度表明者が事象の成立を望ましいと判断しているのか、それとも、望ましくないと判断しているのかを<価値判断>に記述する。このような記法をとることにより、<態度>、<真偽判断>、<価値判断>に関して、表現力があり、かつ、見やすいラベル体系が構築できたと思われる。

残りの３つの項目は、事象の事実性をより明確に記述するために導入したものである。<態度表明者>は、Wiebe ら(2005)が提案した「態度表明者の入れ子構造」により、態度を表明する人物や情報源を詳細に表す。ここで、“wr:筆者”ラベルは態度表明者が書き手であることを、“wr:筆者_arb:不特定”ラベルは態度表明者が不特定の個人や集団であると書き手が述べていることを表す。<相対時>において、真偽の判断が推量の場合に、「事象成立時が未来であるため、その真偽が定まっていない」のか、それとも、「事象成立時は未来ではなく、態度表明者が事象の真偽を確認していないだけである」のかを明示する。<仮想>は、通常の事象であるのか、それとも、条件が存在するなど、仮想的な事象であるのかを明示する。

表２　拡張モダリティの項目とラベル、および、現在のラベルの分布

		OC	OW	PN	PB
文数		6,404	5,835	16,433	9,869
形態素数		110,649	228,651	360,814	234,540
事象候補数		31,528(−%)	78,596(−%)	103,824(−%)	67,521(−%)
事象候補数（タグ付け済み）		26,592(−%)	22,497(−%)	13,561(−%)	16,385(−%)
事象数（タグ付け済み）		14,089(100%)	7,733(100%)	8,819(100%)	9,466(100%)
項目	**ラベル**				
態度表明者	wr:筆者	13,757(98%)	7,320(95%)	8,149(93%)	8,155(86%)
	wr:筆者_arb:不特定	112(1%)	88(1%)	33(0%)	86(1%)
	（その他）	220(1%)	325(4%)	637(7%)	1,225(13%)
相対時	非未来	11,972(85%)	6,214(80%)	7,726(88%)	8,164(86%)
	未来	2,117(15%)	1,519(20%)	1,093(12%)	1,302(14%)
仮想	0	12,445(88%)	7,348(95%)	8,484(96%)	8,388(88%)
	条件	1,167(8%)	290(4%)	242(3%)	724(8%)
	帰結	477(4%)	95(1%)	93(1%)	354(4%)
態度	叙述	11,146(79%)	6,440(83%)	7,923(90%)	8,236(87%)
	意志	314(2%)	754(10%)	280(3%)	394(4%)
	欲求	293(2%)	44(1%)	180(2%)	150(2%)
	働きかけ-直接	496(4%)	40(1%)	41(1%)	85(1%)
	働きかけ-間接	458(3%)	385(5%)	268(3%)	236(3%)
	働きかけ-勧誘	13(0%)	0(0%)	1(0%)	20(0%)
	許可	28(0%)	35(0%)	27(0%)	29(0%)
	問いかけ	1,341(10%)	35(0%)	99(1%)	316(3%)
真偽判断	成立	9,192(65%)	5,672(73%)	6,888(78%)	6,600(70%)
	不成立	985(7%)	188(3%)	671(8%)	919(10%)
	不成立から成立	74(1%)	18(0%)	11(0%)	58(1%)
	成立から不成立	34(0%)	7(0%)	3(0%)	31(0%)
	高確率	874(6%)	930(12%)	508(6%)	804(8%)
	低確率	143(1%)	72(1%)	88(1%)	154(2%)
	低確率から高確率	18(0%)	83(1%)	22(0%)	20(0%)
	高確率から低確率	11(0%)	18(0%)	6(0%)	4(0%)
	0	2,758(20%)	745(10%)	622(7%)	876(9%)
価値判断	0	12,337(88%)	6,458(84%)	8,014(91%)	8,465(89%)
	ポジティブ	1,462(10%)	1,196(15%)	685(8%)	818(9%)
	ネガティブ	290(2%)	79(1%)	120(1%)	183(2%)

事象に対する拡張モダリティ（“<態度表明者>, …, <価値判断>”）の例を以下に示す（下線が対象事象。二重下線は対象事象の核となる述語を示す。）。

iv) 化学専攻は昨年、文部科学省の「二十一世紀ＣＯＥ（センター・オブ・エクセレンス）プログラム」に選ばれた。　（出典：PN3b_00001）
→ “wr:筆者, 非未来, 0, 叙述, 成立, 0”

v) ただ、オファーがあってもやらないでしょうね。　（出典：OC06_01997）
→ “wr:筆者, 未来, 帰結, 意志, 低確率, ネガティブ”

vi) 今後の定期借地権の役割、活用方向などを踏まえ、より利用しやすい制度となるよう、こうした課題について検討し改善を図っていくことが重要である。　（出典：OW6X_00003）
→ “wr:筆者, 未来, 0, 働きかけ-間接, 0, ポジティブ”

vii) 以前はメールはアウトルックに入っていましたがこのメッセージがでてはいらなくなりました　（出典：OC02_01068）
→ “wr:筆者, 非未来, 0, 叙述, 成立から不成立, 0”

viii) 徹先輩が、今度またあの月浜珈琲店に行ってみよう、と言う。　（出典：PB59_00003）
→ “wr:筆者_5680:徹, 未来, 0, 働きかけ-勧誘, 0, ポジティブ”　（“5680”は形態素 ID）

ix) 初歩的な質問ですが、研修医って一番若くて何歳でなれますか？　（出典：OC04_00001）
→ “wr:筆者, 非未来, 0, 問いかけ, 0, 0”

４．拡張モダリティアノテーション

４.１　概要

前章で説明した拡張モダリティの情報を、拡張モダリティタグとして BCCWJ 内の事象に付与する。今回我々が対象としたのは、BCCWJ コアデータ内の次の 4 ジャンルである。

Yahoo!知恵袋(OC)、白書(OW)、新聞(PN)、書籍(PB)

これらにおけるコーパス内の文数と形態素数を表２の上部に示す。

コーパス内の事象に拡張モダリティタグを付与するためには、その前準備としてコーパス内のすべての事象を見つけ出す必要がある。この事象認識は、通常、述語項構造解析によって実行される。我々のアノテーション作業開始時において、BCCWJ コアデータには、述語項構造（≒事象）の情報が付与されていなかったため、我々は独自に事象を認識することにした。本来ならば、与えられた文章に存在する個々の事象の範囲を特定して、それを明確にマークアップするべきではあるが、現在のところ、これを高い精度で自動的に実現することは困難であり、そのほとんどを人手で行うとすると、かなりのコストがかかる。文章において、ほとんどすべての事象は、それぞれ、1 つの述語を核として表現されるため、その述語を事象の代表形態素(列)として用いることができると考えられる。そこで、本研究では、事象の範囲を明確にマークアップすることはせず、述語に対して拡張モダリティタグを付与することで、その述語を核として持つ事象にそのタグを付与したと見なすこととした。例えば、上の例文 viii)において、「行っ」という述語に対して拡張モダリティタグを付与することで、それを核として持つ事象「今度またあの月浜珈琲店に行くコト」にそのタグを付与したと見なす。

BCCWJコアデータはXMLによって表現されている。我々は、このXMLに、事象の拡張モダリティを表現する<eme:event>要素を追加する。この<eme:event>要素の例を図１に示す。<eme:event>要素は、BCCWJにおける<sentence>要素の直接の子要素として記述する。各事象に対して、それぞれ１つの<eme:event>要素を用意するので、<sentence>要素の下に、その文に含まれる事象の数だけ<eme:event>要素が存在することになる[1]。我々のアノテーション

```
<sentence>
<SUW … />教え
    :
<SUW … />。
<eme:event eme:morphIDs="660" eme:orthTokens="教え" eme:source="wr:筆者" eme:time="非未来"
 eme:conditional="条件" eme:pmtype="叙述" eme:actuality="成立" eme:evaluation="0"
 eme:pseudo="限定修飾" eme:lastupdate="20100531" />
<eme:event eme:morphIDs="710" eme:orthTokens="いる" eme:source="wr:筆者" eme:time="非未来"
 eme:conditional="条件" eme:pmtype="叙述" eme:actuality="成立" eme:evaluation="0"
 eme:lastupdate="20091111" />
<eme:event eme:morphIDs="740" eme:orthTokens="心配" eme:source="wr:筆者" eme:time="未来"
 eme:conditional="帰結" eme:pmtype="働きかけ-間接" eme:actuality="0" eme:evaluation="ネガティブ"
 eme:lastupdate="20101220" />
</sentence>
```

はXMLの要素を単純に追加するだけであるので、他の様々な情報に対するアノテーションと衝突する可能性はほとんどないと思われる。

図１ BCCWJ 内の拡張モダリティタグ

４.２ 事象認識

本研究では、事象の核となる述語を表す品詞として、主に、動詞、形容詞、形状詞、名詞-普通名詞-サ変可能/形状詞可能を用いた。さらに、名詞述語による事象をもれなく抽出するために、後続形態素列に基づく述語抽出規則を作成して用いた。これらの品詞や規則により抽出されるのは事象候補の述語のリストであり、その中には、事象の述語だけでなく、「(て) いる」や「(て)あげ」のような補助動詞や、「(に) よる (と)」や「(とは) 言っ (ても)」のような複合辞の一部、名詞述語でない名詞などが含まれる。本研究では、事象でない事例に対して、「対象外」という補足欄に“機能語”や“名詞”などを人手で記述し、通常の事象と区別する。

表２の上部における「事象候補数 (タグ付け済み)」と「事象数 (タグ付け済み)」の差分が、これまでに見つかった、事象でない事例の数である。おおよそ、その１／３が機能語 (補助動詞や複合辞の一部)、１／３が名詞述語でない名詞、そして、残りの１／３が限定修飾の事例である。限定修飾については、４.４節で述べる。

４.３ 現状

タグ付け作業は、主に１人の作業者が行っている。作業者には、事象候補が含まれる文全体 (表層形の列と基本形の列) と核となる述語の位置を提示する。タグ付けにかかる時間は、1,000 事象候補あたりおよそ５時間である。

本論文執筆時点における、拡張モダリティの各項目に対するラベルの分布を表２の下部に示す。タグ付け済み事象数は 40,107 であり、このうち、OC の 14,089 事象に対しては、

[1] 正確には、同じ事象に対しても入れ子の態度表明者ごとに<eme:event>要素を用意するので、<eme:event>要素のほうが多いこともある。

我々が試作したモダリティ自動解析システムの解析結果をフィードバックさせ、それを参照しながらのタグ見直し作業を数回行い、タグの質を向上させている。

タグ付け開始後しばらくの間は、OW、PN、PB に関しては、述語に後続する特徴的な形態素列ごとにタグを付与していたため、表２のラベル分布はコーパス全体の分布を正確に反映していない可能性がある。しかしながら、それぞれの項目において、おそらく、全体の 70%ー90%の事例を占めるラベルが存在することは確かであり、言語処理においては、残りの 10%ー30%の事例に関して、そのラベルを正確に判定することが重要となる。

４.４　現在の体系の問題点

多数の事例に対してタグ付けすることにより、事象の認め方に関して２つ、拡張モダリティのラベルに関して２つ問題があることが明らかになった。

- 事象の認め方の問題１：限定修飾
  これまで限定修飾の事例は対象外としていたが、文 x)の事象「テレホンカードを使うコト」のように、拡張モダリティタグを付与すべき事例が見つかった。質問応答や含意認識などの応用を考慮すると、この例以外にも限定修飾の事例に対して拡張モダリティタグを付与すべきであると思われる。今後は、補足欄「対象外」に“限定修飾”と記述しつつ、拡張モダリティの情報も付与する。

x)　使っていないテレホンカードが百枚くらいあります。　（出典：OC03_00001）
→　“wr:筆者, 非未来, 0, 叙述, 不成立, 0”

- 事象の認め方の問題２：事象とも解釈できる機能語
  文 xi)の「(ーと）思う」や、「(ーを）図る」、「(~~をう~~ 期待（する)」、「(気が）する」のように、別事象のモダリティ表現であるのか、それとも、独立の事象であるのかその判断が悩ましい事例が存在することが分かった。今後は、補足欄「対象外」に“機能語-事象可能”と記述しつつ、限定修飾の場合と同様に、これらに対しても拡張モダリティの情報を付与する。

xi)　どのアニメでもやってるし多少は仕方ないと思う。　（出典：OC01_05000）
→　“wr:筆者, 非未来, 0, 叙述, 成立, 0”

- ラベルの問題１：実時間軸の外の“非未来”
  現在の体系では、<相対時>のラベルは２種類であるため、文 xii)のような脱時間的な一般事象と、文 xi)のような個別事象を区別できない。

xii)　（ア）　責任裁定は，公害に係る被害についての損害賠償を請求する者の申請に基づいて，裁定委員会が公開の期日を開いて当事者に陳述させ，証拠調べ，事実の調査などを行って事実を認定し，その認定した事実に基づいて裁定するものである。　（出典：OW6X_00008）
→　“wr:筆者, 非未来, 0, 叙述, 成立, 0”

- ラベルの問題２：態度非表明の<真偽判断>=“0”
  文 xiii)のような、そもそも態度を表明していない事例と、「 ーかどうか分からない」の使用など、真偽不明（“叙述, 0, 0”）を明示的に表明している事例を区別できない。

xiii)　押し込むのはちょっとだけで良いんです。　（出典：OC01_02382）
→　“wr:筆者, 非未来, 0, 叙述, 0, 0”

５．おわりに

本稿では、文章に存在する事象のモダリティおよびその周辺情報を表現する拡張モダリ

ティの体系について述べ、この体系に基づいて BCCWJ へアノテーションする方法、および、その現状について報告した。このアノテーション結果の拡張モダリティタグ付与コーパスの最新情報は、原稿末尾の URL にて公開予定である。

今後は、構築したコーパスと機械学習手法を用いて事象の拡張モダリティを自動的に解析するシステムの開発と精度向上に取り組むつもりである。

文献

原一夫、乾健太郎(2008).「事態抽出のための事実性解析」情報処理学会研究報告 2008-FI-89, 2008-NL-183, pp.75-80.

川添愛、齊藤学、片岡喜代子、崔栄殊、戸次大介(2010).「言語情報の確実性アノテーションのための様相表現の分類」九州大学言語学論集, 31, pp.109-129.

Marc Light, Xin Ying Qiu and Padmini Srinivasan (2004). "The language of bioscience: Facts, speculations, and statements in between." In *BioLink 2004 workshop on linking biological literature, ontologies and databases*, pp.17-24.

益岡隆志(2007). 『日本語モダリティ探究』くろしお出版.

Ben Medlock and Ted Briscoe (2007). "Weakly supervised learning for hedge classification in scientific literature." In *the 45th Annual Meeting of the Association of Computational Linguistics*, pp.992-999.

仁田義雄（代表），日本語記述文法研究会（編）(2003). 『現代日本語文法 4』くろしお出版.

Frank Robert Palmer (2001). 『*Mood and Modality Second edition*』Cambridge University Press.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi and Bonnie Webber (2006). "Annotating attribution in the Penn discourse treebank." In *the COLING/ACL Workshop on Sentiment and Subjectivity in Text*, pp.31-38.

Victoria Rubin, Elizabeth Liddy and Noriko Kando (2005). "Chapter 7: Certainty Identification in Texts: Categorization Model and Manual Tagging Result*" Computing Attitude and Affect in Text: Theory and Applications*, Springer, pp.61-74.

Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer and James Pustejovsky (2006). "TimeML Annotation Guidelines Version 1.2.1." http://www.timeml.org/site/publications/timeMLdocs/annguide 1.2.1.pdf

Roser Saurí and James Pustejovsky (2007). "Determining modality and factuality for text entailment." In *the International Conference on Semantic Computing*, pp.509-516.

Roser Saurí and James Pustejovsky (2009). "Factbank: a corpus annotated with event factuality." In *Language Resources and Evaluation*.

György Szarvas, Veronika Vincze, Richárd Farkas and János Csirik (2008). "The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts." In *the Workshop on Current Trends in Biomedical Natural Language Processing*, pp.38-45.

Janyce Wiebe, Theresa Wilson and Claire Cardie (2005). "Annotating expressions of opinions and emotions in language." In *Language Resources and Evaluation 39 issue 2-3*, pp.165-210.

関連 URL

拡張モダリティタグ付与コーパス：http://www.cl.ecei.tohoku.ac.jp/resources/modality/

拡張モダリティタグ付与コーパス作成の作業基準書：
http://www.cl.ecei.tohoku.ac.jp/resources/modality/manual.pdf

UniDic2: 設計と実装

小木曽 智信（電子化辞書班分担者：国立国語研究所言語資源系）†
伝 康晴（電子化辞書班班長：千葉大学文学部）

UniDic2: Design and Implementation

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Yasuharu Den (Faculty of Letters, Chiba University)

1. はじめに

電子化辞書班で開発を行っている UniDic は、これまで MeCab・ChaSen 用の形態素解析辞書として公開されており、『現代日本語書き言葉均衡コーパス』（BCCWJ）のサンプルへの形態論情報の付与に用いられている。その一方で、UniDic は当初より形態素解析辞書以外の利用を視野に入れた設計がなされている（伝ほか 2007）。多様な情報が付加された UniDic の見出し語は、言語研究や自然言語処理、音声処理等の各方面で利用が見込まれる価値の高いデータである。UniDic のこうした可能性を引き出すためには、言語資源として利用しやすい形式でデータを提供する必要がある。そこで電子化辞書班では、UniDic の見出し階層構造を反映した XML 形式により、言語資源としての電子化辞書 UniDic2 を公開することとした。あわせて、XML 形式の形態論情報からユーザが必要に応じて拡張した形態素解析辞書を作成することのできる UniDic Tools を公開する。

2. 言語資源としての UniDic

UniDic は斉一な単位による解析を実現するために、見出し語の認定を厳密なルールによって定めた「短単位」を見出し語に採用している（小椋ほか 2011）。さらに、柔軟な見出し語付与を可能にするために、語彙素・語形・書字形・発音形の階層構造（図 1）を持たせ、表記の揺れや語形の変異にかかわらず同一の見出しを与えることを可能にしている。

図 1 UniDic の階層構造

UniDic の見出し語を管理する形態論情報データベース上では、この階層構造をそのままテーブル構造に反映させ、各表を関連づけて見出し語を格納している（小木曽ほか 2011）。それぞれの表には 2011 年 1 月現在、語彙素約 21 万項目、語形約 23 万項目、書字形約 33 万項目の見出し語を収録している。

† togiso@ninjal.ac.jp

UniDic2 では、この階層構造をそのまま反映した XML 形式（付録 参照）で提供する。これにより、形態素解析辞書以外での利用を行いやすくするとともに、

・利用者が各階層に新たな見出し語を追加すること

・利用者が各階層の見出し語に情報を自由に付与すること

を可能にする。

2.1. 語彙表の展開

辞書管理システム上の「語彙素」「語形」「書字形」「発音形」の 4 つの表は、語頭・語末変化表、活用表によって出現形まで展開される（図 2）。見出し語を出現形まで展開した表を（変化形展開）語彙表と呼んでいる。

図 2　語彙表生成の流れ

語頭変化は、連濁のように語頭が変化するもの（亀［語頭変化型=カ濁］：カメ，ガメ）、語末変化は促音化のように語末が変化するもの（三角［語末変化型=ク促］：サンカク，サンカッ）である。これらの変化は、語形につけられた語頭・語末変化型の属性と、各変化型がどのような語形を派生するかを記述した語頭・語末変化表とを組み合わせることによって生成される。活用についても同様に、語形レベルの活用型と活用表から、各活用形が生成される。

UniDic 2 では、辞書管理システムから、基本形語彙表（lexBaseCore.xml）、語頭変化表（iFormCore.xml）、語末変化表（fFormCore.xml）、活用表（inflCore.xml）を XML 形式で出力する。たとえば、基本形語彙表は以下のようである。

```
<Lemma lemma="同じく" lForm="オナジク" class="相" goshu="和">
  <Form formBase="オナジク" formOrthBase="同じく" pos="副詞">
    <Orth orthBase="おなじく" kanaBase="オナジク" />
    <Orth orthBase="同じく" kanaBase="オナジク" />
    <Pron pronBase="オナジク" />
  </Form>
  <Form formBase="オナジュウ" formOrthBase="同じゅう" pos="副詞">
    <Orth orthBase="同じゅう" kanaBase="オナジュウ" />
    <Pron pronBase="オナジュー" />
  </Form>
</Lemma>
```

辞書管理システムの外部でこれらの表を組み合わせて変化形展開語彙表を生成するため、ユーザが基本形語彙表に手を加えることで、見出し語の追加や不要な見出し語の削除を行うことができる。

なお、ある活用型の中で特定の語のみに存在する活用形（「歩っ（た）」）やカタカナ書きされた活用形（「アツイ」）など、通常の変化・活用表展開では出力できない変化・活用形

については、特殊変化・活用形として変化・活用後の形を基本語彙表に直接記述することにより対応している（付録の AltOrth）。

2.2. 形態素解析辞書の生成と付加情報の付与

語彙表をユーザの手元で XML ファイルから生成する場合、形態素解析システム用の辞書を作成するためには、統計モデルのコスト情報を後から語彙表に付与する必要がある。そのため、UniDic 2 ではコアとなる基本 8 属性（語彙素読み・語彙素表記・語彙素細分類・品詞・活用型・活用形・書字形・発音形）を持つ語彙表を基本辞書として、これに対して各種の情報を後から付加することを可能にしている。この方法により、形態素解析辞書のコスト情報を付加できるようにするだけでなく、たとえば分類語彙表番号のような意味に関わる情報（語彙素レベル）、各種漢字表との関わりなどの表記に関する情報（書字形レベル）、アクセントなどの音声情報（発音形レベル）などを追加していくことができる。これらの付加情報は、見出し語階層の各レベルで情報記述を行えば、語彙表中の対応する見出し語に情報が付加される（4 節参照）。

3. UniDic Tools

UniDic2 では、XML ファイル群で記述された形態論辞書から辞書データベースを作成したり、作成した辞書ベースをさまざまな形態で利用したりするためのツール群を UniDic Tools として提供している。UniDic Tools は Linux システム上のみで動作し、以下のツール群を含む。

1. XML ファイル群で記述された形態論辞書から辞書データベースを作成するツール
2. 辞書データベースから形態素解析システム用辞書を生成するツール
3. 辞書データベースを検索するツール
4. 辞書データベースから情報を取得し、形態素解析済みテキストに付加するツール

以下では、これらのツール群について概説する。

3.1. 辞書データベースの作成

XML ファイル群で記述された形態論辞書から SQLite のデータベースファイルを作成できる。この処理は標準的な `configure && make` コマンドで実行できる。辞書データベースの作成にかかる時間はノート PC（Core2 Duo/3.06GHz）で 5 分程度である。

3.2. 形態素解析辞書の作成

辞書データベースから形態素解析システム MeCab 用の辞書を作成できる。設定ファイルを切り替えることで、さまざまな情報を含む形態素解析辞書を作成することができる。

3.3. 辞書データベースの検索

辞書データベースを検索するための CUI および GUI ベースのツールを提供している。

CUI 版検索ツールは Perl スクリプトで実装され、Linux および Windows のコマンドシェルで動作する。たとえば、以下のような検索が可能である。

```
# 語彙素が「に」であるすべてのエントリーを検索
> lemma="に"
ニ|に|*|助詞-接続助詞|*|*|に|に|ニ|ニ|和
ニ|に|*|助詞-格助詞|*|*|に|に|ニ|ニ|和
ニ|に|*|助詞-格助詞|*|*|ニ|ニ|ニ|ニ|和
ニ|に|*|助詞-格助詞|*|*|にぃ|にぃ|ニー|ニー|和
ニ|に|*|助詞-格助詞|*|*|にー|にー|ニー|ニー|和
ニ|に|*|助詞-格助詞|*|*|にゃ|にゃ|ニャ|ニャ|和
ニ|に|*|助詞-格助詞|*|*|にゃぁ|にゃぁ|ニャー|ニャー|和
ニ|に|*|助詞-格助詞|*|*|にゃあ|にゃあ|ニャー|ニャー|和
ニ|に|*|助詞-格助詞|*|*|ん|ん|ン|ン|和
```

```
# 語彙素が「痛い」のエントリーで活用形が「連用形-一般」のものを検索
> lemma="痛い" and cForm="連用形-一般"
イタイ|痛い|*|形容詞-一般|形容詞|連用形-一般|いたい|いたく|イタイ|イタク|和
イタイ|痛い|*|形容詞-一般|形容詞|連用形-一般|イタい|イタく|イタイ|イタク|和
イタイ|痛い|*|形容詞-一般|形容詞|連用形-一般|痛い|痛く|イタイ|イタク|和
```

一方、GUI 版検索ツールは Windows でのみ動作し、図 3 のようなインターフェースを通じて、辞書データベースを検索できる。

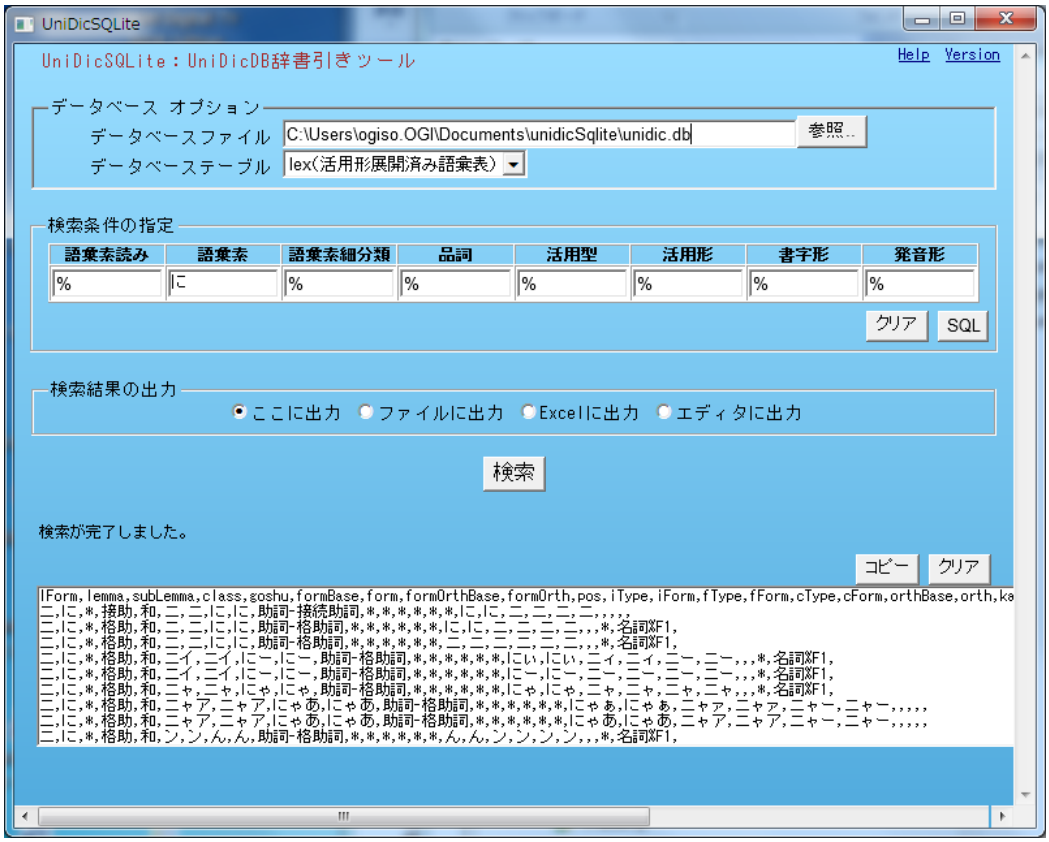


図 3　GUI 版辞書データベース検索ツール

3.4. 形態素解析済みテキストへの付加情報付与

UniDic2 では、品詞・活用型・活用形などの基本的な形態論情報以外にも様々な情報を利用できる。UniDic Tools は、基本情報のみからなる形態素解析済みテキストに対して、様々な付加情報を付与するプログラムを提供している。たとえば、以下のような形態素解析済みテキストに対して、

```
chamame B コーパス コーパス コーパス コーパス 名詞-普通名詞-一般
chamame I 言語　ゲンゴ ゲンゴ 言語 名詞-普通名詞-一般
chamame I 学 ガク ガク 学 接尾辞-名詞的-一般
chamame I を オ ヲ を 助詞-格助詞
chamame I 勉強 ベンキョー ベンキョウ 勉強 名詞-普通名詞-サ変可能
chamame I し シ スル 為る 動詞-非自立可能 サ行変格 連用形-一般
chamame I て テ テ て 助詞-接続助詞
chamame I い イ イル 居る 動詞-非自立可能 上一段-ア行 連用形-一般
chamame I ます マス マス ます 助動詞 助動詞-マス 終止形-一般
```

語種を各行末に追記し、以下のような出力テキストを得ることができる。

```
chamame B コーパス コーパス コーパス コーパス 名詞-普通名詞-一般　外
chamame I 言語　ゲンゴ ゲンゴ 言語 名詞-普通名詞-一般　漢
chamame I 学 ガク ガク 学 接尾辞-名詞的-一般　漢
chamame I を オ ヲ を 助詞-格助詞　和
chamame I 勉強 ベンキョー ベンキョウ 勉強 名詞-普通名詞-サ変可能　漢
chamame I し シ スル 為る 動詞-非自立可能 サ行変格 連用形-一般 和
chamame I て テ テ て 助詞-接続助詞　和
chamame I い イ イル 居る 動詞-非自立可能 上一段-ア行 連用形-一般 和
chamame I ます マス マス ます 助動詞 助動詞-マス 終止形-一般 和
```

入力テキスト中の属性列の並びや出力テキストに含める属性列の並びなどは定義ファイル中に指定でき、柔軟な利用が可能である。

4. 設定ファイル

4.1. 概要

UniDic2 の最大の特徴は、設定ファイルを切り替えることで、さまざまな情報を含む辞書データベースを作成できることである。たとえば、形態素解析システムを動作させる上で必要最小限の情報のみを含む辞書を作成することもできるし、アクセント型を付与する後処理システムで参照するためのアクセント情報を含む辞書を作成することもできる。さらに、ユーザが独自に語彙を追加したり、付加情報を追加したりすることもできる。

図 4 に XML ファイル群で記述された形態論辞書から辞書データベースを作成する過程の一例を図示する。この例では、基本情報（基本形語彙表・語頭変化表・語末変化表・活用表）を記述した XML ファイル群と、付加情報（語彙表発音形付加情報・活用表発音形付加情報）を記述した XML ファイル群とを辞書データベースに読み込み、基本形語彙表（lexBase）と変化形を展開処理した語彙表（lex）を作成する。さらに、別途記述された単語コスト情報ファイルから作成される単語コスト表（lexCost）と変化形展開語彙表

とを結合して CSV ファイルに出力することで、形態素解析辞書（lex.csv）を作成する。この一連の過程が 3.1 項・3.2 項のツールにより全自動で実行される。

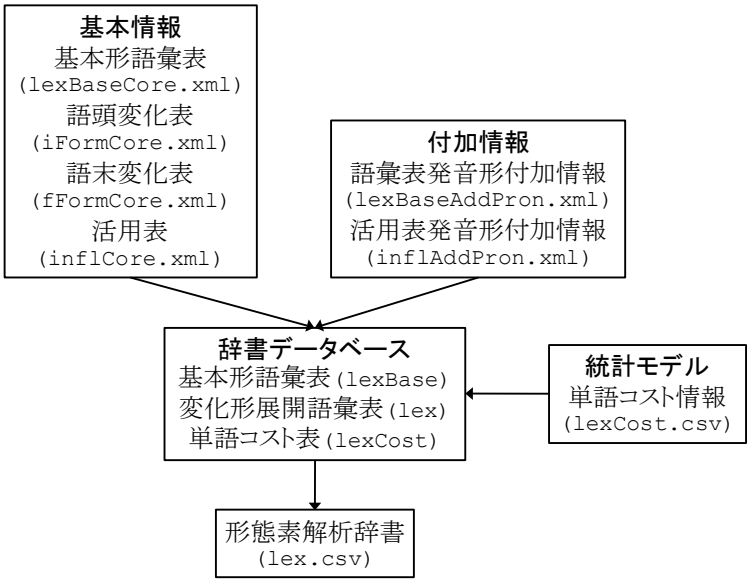


図 4　辞書データベースの作成過程

設定ファイルは、基本情報や付加情報を記述した XML ファイル群の名前や付加情報として定義される属性の名前などを定義する。以下、設定ファイルの仕様を説明する。

4.2. 最小設定

形態素解析システム MeCab を動作させる上で必要最小限の情報のみを含む辞書データベースを作成するには、以下のような設定ファイルを用いる。

```
xmldir=xml_samples                # 形態論辞書 XML ファイル群があるディレクトリ
lexBaseCore=lexBaseCore.xml       # 基本形語彙表を記述した XML ファイル
iFormCore=iFormCore.xml # 語頭変化表を記述した XML ファイル
fFormCore=fFormCore.xml # 語末変化表を記述した XML ファイル
inflCore=inflCore.xml   # 活用表を記述した XML ファイル
```

4.3. 拡張設定

最小設定の内容に加え、発音形付加情報を辞書データベースに含めるには、設定ファイルに以下の定義を追加する。

```
lexBaseAddPron=lexBaseAddPron.xml/aType:t,aConType:t
        # 語彙表発音形付加情報を記述した XML ファイルとそこでの属性の名前と型
inflAddPron=inflAddPron.xml/aModType:t
        # 活用表発音形付加情報を記述した XML ファイルとそこでの属性の名前と型
```

ここでは、lexBaseAddPron.xml 中でテキスト型の属性 aType（アクセント型）と aConType（アクセント結合型）が記述され、inflAddPron.xml 中でテキスト型の属性

aModType（アクセント修飾型）が記述されていることを宣言している。属性の型としては、テキスト型（t）・整数型（i）・実数型（r）が利用できる。

4.4. ユーザ定義辞書の利用

システム辞書の語彙をユーザが独自に定義した語彙で拡張することもできる。以下の設定ファイルでは、基本形語彙表を記述する XML ファイルとして、userCore.xml を追加で指定し、それらに対する発音形付加情報が userAddPron.xml に記述されていることを宣言している。

```
lexBaseCore=lexBaseCore.xml,userCore.xml
        # 基本形語彙表は 2 つの XML ファイルで記述
lexBaseAddPron=lexBaseAddPron.xml,userAddPron.xml/aType:t,aConType:t
        # それぞれの基本形語彙表に対する語彙表発音形付加情報ファイル
```

同様の仕組みにより、lexBaseCore で指定する XML ファイルの組み合わせをさまざまに変更することで、話し言葉用・近代語用・ブログ解析用など、目的に応じてカスタマイズされた語彙表を作成することが可能になる。

4.5. ユーザ定義付加情報の利用

語彙の拡張に加えて、付加情報も拡張できる。以下の設定ファイルでは、発音形付加情報として新たに整数型属性 nMorae（モーラ数）を指定する場合である。

```
lexBaseAddPron1=lexBaseAddPron.xml,userAddPron.xml/aType:t,aConType:t
lexBaseAddPron2=lexBaseAddPron2.xml,userAddPron2.xml/nMorae:i
lexBaseAddPron="$lexBaseAddPron1;$lexBaseAddPron2"
        # 2 種類の語彙表発音形付加情報を宣言し、; 区切りで連結
```

このように、語彙・付加情報をユーザが拡張できることで、極めてカスタマイズ性に富んだ辞書データベース（および形態素解析辞書）作成機能を提供することができる。

5. おわりに

UniDic2 により UniDic は今まで以上に利用しやすい言語資源として公開されることとなった。これにより UniDic のもつ可能性が引き出され、さらに広く多様な目的で利用されるようになることを期待したい。

文献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22 pp.101-123

小椋秀樹・小磯花絵・冨士池優美・宮内左夜香・小西光・原裕（2011）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）』LR-CCG-10-05-01/02

小木曽智信・中村壮範（2011） 国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版』LR-CCG-10-06

付録　基本語彙表 XML（lexBaseCore.xml）サンプル

```
<Lemma lemma="熱い" lForm="アツイ" class="相" goshu="和">
  <Form formBase="アッツイ" formOrthBase="あっつい" pos="形容詞-一般" cType="形容詞" subCType="ツイ">
    <Orth orthBase="あっつい" kanaBase="アッツイ" cTypeOrth="かな" />
    <Pron pronBase="アッツイ" />
  </Form>
  <Form formBase="アツイ" formOrthBase="熱い" pos="形容詞-一般" cType="形容詞" subCType="ツイ">
    <Orth orthBase="あつい" kanaBase="アツイ" cTypeOrth="かな" />
    <Orth orthBase="アツい" kanaBase="アツイ" cTypeOrth="一般" />
    <Orth orthBase="熱い" kanaBase="アツイ" cTypeOrth="一般" />
    <Orth orthBase="アツイ" kanaBase="アツイ" cTypeOrth="一般">
      <AltOrth orth="アツイ" kana="アツイ" cForm="終止形-一般" subCForm="一般" />
      <AltOrth orth="アツイ" kana="アツイ" cForm="連体形-一般" subCForm="一般" />
    </Orth>
    <Pron pronBase="アツイ" />
  </Form>
  <Form formBase="アツーイ" formOrthBase="熱ーい" pos="形容詞-一般" cType="形容詞" subCType="ーイ">
    <Orth orthBase="アツーイ" kanaBase="アツーイ">
      <AltOrth orth="アツーイ" kana="アツーイ" cForm="連体形-一般" />
      <AltOrth orth="アツーイ" kana="アツーイ" cForm="終止形-一般" />
      <AltOrth orth="アツーク" kana="アツーク" cForm="連用形-一般" />
    </Orth>
    <Pron pronBase="アツーイ" />
  </Form>
</Lemma>
<Lemma lemma="開ける" lForm="アケル" class="用" goshu="和">
  <Form formBase="アケル" formOrthBase="開ける" pos="動詞-一般" cType="下一段-カ行">
    <Orth orthBase="あける" kanaBase="アケル" />
    <Orth orthBase="開ける" kanaBase="アケル" />
    <Orth orthBase="開ケる" kanaBase="アケル" />
    <Pron pronBase="アケル" />
  </Form>
  <Form formBase="アケレル" formOrthBase="開けれる" pos="動詞-一般" cType="下一段-ラ行" subCType="一般">
    <Orth orthBase="開けれる" kanaBase="アケレル" />
    <Pron pronBase="アケレル" />
  </Form>
</Lemma>
<Lemma lemma="同じく" lForm="オナジク" class="相" goshu="和">
  <Form formBase="オナジク" formOrthBase="同じく" pos="副詞">
    <Orth orthBase="おなじく" kanaBase="オナジク" />
    <Orth orthBase="同じく" kanaBase="オナジク" />
    <Pron pronBase="オナジク" />
  </Form>
  <Form formBase="オナジュウ" formOrthBase="同じゅう" pos="副詞">
    <Orth orthBase="同じゅう" kanaBase="オナジュウ" />
    <Pron pronBase="オナジュー" />
  </Form>
</Lemma>
<Lemma lemma="亀" lForm="カメ" class="体" goshu="和">
  <Form formBase="カメ" formOrthBase="亀" pos="名詞-普通名詞-一般" iType="カ濁">
    <Orth orthBase="かめ" kanaBase="カメ" iTypeOrth="かな" />
    <Orth orthBase="カメ" kanaBase="カメ" iTypeOrth="カナ" />
    <Orth orthBase="亀" kanaBase="カメ" iTypeOrth="一般" />
    <Pron pronBase="カメ" />
  </Form>
</Lemma>
<Lemma lemma="十" lForm="トオ" class="数" goshu="和">
  <Form formBase="トオ" formOrthBase="十" pos="名詞-数詞" fType="オ長削">
    <Orth orthBase="とお" kanaBase="トオ" fTypeOrth="かな" />
    <Orth orthBase="十" kanaBase="トオ" fTypeOrth="一般" />
    <Pron pronBase="トー" />
  </Form>
</Lemma>
```

日本語研究とインターネット

田野村忠温（日本語学班班長：大阪大学大学院文学研究科）

The Study of Japanese and the Internet

Tadaharu Tanomura (Osaka University)

１．日本語研究資料としてのインターネット文書

日本語研究資料として見たインターネット文書の特徴と魅力は、大量性と質的多様性の２点にある。

１．１　インターネット文書の大量性

インターネット文書を総体として見れば、利用可能な他のいかなる電子文書よりもはるかに大規模であり、しかも、それが年々成長を続けている。

拙論(2008,2009a)で述べた推計によれば、Yahoo!JAPAN（以後適宜Yahoo!と略記）が収集しサーチエンジンの検索対象としている日本語文書の量は2008年夏の時点で26兆字程度である。これは平均的な小説単行本換算で約１億３千万冊の量に相当する。これにより、インターネット文書には通常のコーパスには少数しか含まれない用例が大量に含まれること、そして、通常のコーパスにはまったく含まれない用例が含まれることを期待し得ることになる。

各種の日本語文書のサイズの比較を（表１）に示す。これは拙論(2009b)に掲載した表を簡略化したものである。

（表１）各種日本語文書のサイズ比較

日本語データの種類	字数　(データ量)	小説との比率
小説単行本	20万字	1
新潮文庫の100冊	2,000万字　(40メガバイト)	100
新聞記事１年分	6,000万字　(120メガバイト)	300
BCCWJ	2.8億字　(1億語)	1,400
Yahoo!知恵袋ベータ版データ	16億字　(3.2ギガバイト)	8,000
国会会議録のデータ	35億字　(7ギガバイト)	18,000
拙作Webコーパス	750億字　(150ギガバイト)	375,000
Web上の日本語文書（Yahoo!収集分）	26兆字　(52テラバイト)	130,000,000

インターネット上にある日本語文書の総量は上述の値をはるかに上回るものと思われるが、その具体的な量をネットワーク経由で知ることは実際上のみならず原理的にも不可能である。すなわち、インターネット上でリンクされたすべての文書を取得することは実際上不可能であり（あまりに大量でしかも常に変動しているため)、また、そもそも万人が自由にアクセスできない文書も存在する（例えば、公然とリンクされていない文書、ユーザー認証を経て初めてアクセスできる文書、検索システムなどを通じて部分的にのみアクセ

スできる文書がある）ことから、通信可能な状態に置かれた文書の全体を把握することは原理的にも不可能である。

１．２　インターネット文書の質的多様性

インターネット文書を性質の面から見れば、文学作品や新聞などの出版物に基づく従来のコーパスよりもはるかに多様である。特に、インターネット上での個人的な情報発信の普及により、インターネット文書は規範に制約されない日本語の様相、生きた日本語の姿の観察を可能にする。このことは単なる流行語や俗語に対する雑学的関心に応えてくれるだけではなく、言語変化の様相や原因を考える材料をも提供してくれる。

動詞「あじわう」の第３拍「わ」が「あ」に変化する現象がある。いくつかの述語形についてGoogleで検索したときのヒット件数を（表２）に示す（各キーワードに " " を加えたフレーズ検索による)。表では、「あじわ」「あじあ」を含む対応形式のうち相対的にヒット件数の多いほうをグレー地にしてある。この統計から、「わ」から「あ」への変化はこの動詞において一様に起きているわけではなく、「わわ」という同音連続を避けるという動機のもたらす結果であることが推測できる。

（表２）「あじわ～」と「あじあ～」（Google検索、2011年1月）

「あじわ～」のヒット件数		「あじあ～」のヒット件数		「あ」の比率
あじわわせ	9,000	あじあわせ	19,500	68%
あじわわない	2,340	あじあわない	7,380	76%
あじわわれ	248	あじあわれ	5,230	95%
あじわわされ	1,120	あじあわされ	3,720	77%
あじわい	275,000	あじあい	2,500	1%
あじわった	118,000	あじあった	10,400	8%
あじわう	40,200	あじあう	5,210	11%
あじわおう	27,800	あじあおう	608	2%
あじわえば	4,970	あじあえば	1,930	28%

同じ調査を約11年前にサーチエンジンAltaVistaを使って行ったことがあり、そのときにも同様の結論を得た（拙論(2000a,2003)）。今回の調査では当時の調査に比べてヒット件数が全体に２～３桁増えており、上記の推測をいっそう確からしいものとしている。動詞「にぎわう」においても同様に、「にぎあわせ」「にぎあわない」のヒット件数が「にぎわわせ」「にぎわわない」のそれを上回っている。

この同音連続の回避という動機は語形変化の原因となるだけでなく、日本語の文法的な表現のあり方にも影響を与えている（拙論(2000a)）。

２．サーチエンジン

日本語の研究や観察にインターネット文書を役立てる手軽な方法は、1.2でも実例を示した、サーチエンジンの利用である。サーチエンジンを使えば、特定の表現のインターネット文書における用例を容易に取得することができ、また、表示されるヒット件数を通して表現の使用傾向を知ることができる。[1] 今やインターネットを使う日本語研究者なら誰しも、

[1] サーチエンジンのヒット件数に着目して日本語表現の使用傾向を考える可能性を最初に公に指摘したの

気になる日本語の表現をサーチエンジンで検索してみた経験があるのではなかろうか。

２．１　日本語研究の手段としてのサーチエンジンの問題点

しかし、サーチエンジンによることばの調査は簡便で有用ではあるが、それを学問的な研究の手段として見ると問題点が少なくない（拙論(2000b)）。インターネット文書には誤りが多いという文書の内容に関わる問題点を別としても、サーチエンジンでは意図通りの正確な検索ができないという仕様上の制約があり――正規表現は使えず、また、あいまい検索の機能により、求めていない表現の用例までサーチエンジンは拾ってくる――、加えて、サーチエンジンの示すヒット件数が必ずしもあてにならないという信頼性の問題がある。

サーチエンジンのヒット件数の信頼性に関しては、論理的不整合と時間変動という２種類の問題点がある（拙論(2008,2009a)）。

(1) ヒット件数の論理的不整合

ヒット件数の論理的不整合の単純な例としては、「A AND B」と「B AND A」（あるいは、「A OR B」と「B OR A」）という条件で検索したときのヒット件数が大幅に食い違うという問題が以前はあったが、このたび何種類かの表現で試してみた限りではその問題は確認できなかった。しかし、複合的な条件での検索のヒット件数の信頼性に依然として重大な問題があることは、（表３）に示すGoogleでの検索結果からも明らかである。複合条件での検索結果のうち、A、B単独のヒット件数が正しいと仮定したときに期待し得るヒット件数の範囲を大きく逸脱した部分をグレー地にしてある。[2]

（表３）AND、ORを含む複合条件での検索（Google検索、2011年1月）

A	B	"A" のヒット件数	"B" のヒット件数	"A" AND "B" のヒット件数	"A" OR "B" のヒット件数
犬	猫	65,500,000	143,000,000	19,800,000	217,000,000
男	女	359,000,000	489,000,000	3,480,000,000	7,530,000,000
大人	子ども	92,700,000	28,900,000	38,200,000	595,000,000
東京	大阪	220,000,000	104,000,000	87,900,000	311,000,000
とても	すごく	85,600,000	36,000,000	76,900,000	397,000,000
暑い	寒い	15,400,000	30,900,000	3,780,000	56,000,000
日本語	コーパス	596,000,000	279,000	1,030,000	600,000,000

Yahoo!はGoogleに比べればこうした論理的不整合が少ないというのが拙論(2008)での観察によって得た結論であった。

(2) ヒット件数の時間変動

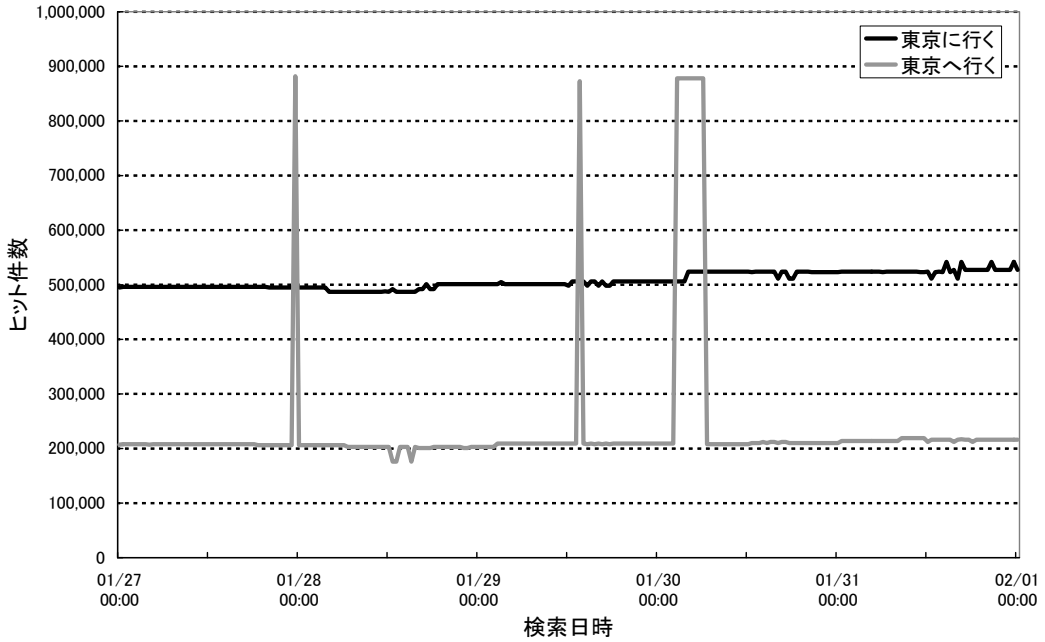
ヒット件数の時間変動に関しても深刻な問題がある。[3] その具体例として、2008年1月の数日間における「東京に行く」「東京へ行く」のGoogleヒット件数の推移を示す図を拙論

は岡島(1997)である。

[2] ヒット件数の論理的不整合の一部は 2.1.2 で見るヒット件数の時間変動の結果である可能性がある。しかし、ヒット件数の論理的不整合をすべて時間変動に帰することはおそらく不可能である。

[3] ヒット件数の時間変動については、田中(2003)、荻野(2004) 、荻野他(2007)に指摘と具体例の記述がある。拙論(2008,2009a)はそれをより詳しく調査してみたものである。

(2008)から引用する。



（図１）Google検索でのヒット件数の推移――「東京に行く」「東京へ行く」

この図では安定しているかのように見える「東京に行く」のヒット件数さえ、実は長期的に見れば極端に不安定なグラフの一部を成すものであったことについては拙論(2009a)で報告した。

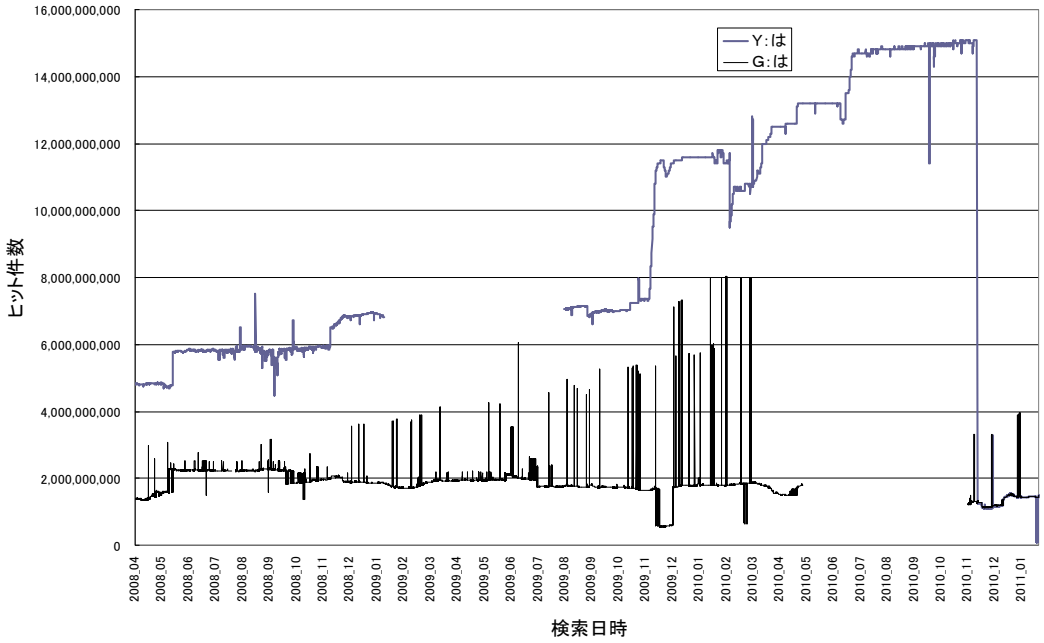
２．２　最近の状況――GoogleとYahoo!JAPANの技術提携とその影響

拙論(2008,2009a)での分析の結論は、Yahoo!のほうがGoogleよりデータ量が多く（Googleの3～4倍程度)、ヒット件数も安定している（論理的不整合、時間変動ともに少ない)、したがって、日本語表現の用例を探したり使用頻度を調べたりする目的にはYahoo!が適しているというものであった。

しかし、その相対評価は今から４か月前に意味を失った。2010年7月27日にGoogleはYahoo!JAPANに対する検索技術のライセンス提供を発表し、その数か月後にYahoo!は従前のサーチエンジンからGoogleのそれへの切り替えを行った。発表者が2008年以来継続的に行っている調査における検索ヒット件数の記録によれば、その切り替えは2010年11月17日(水)の午前10時に行われた（±30分の誤差の可能性あり)。ただし、サーチエンジンの返す検索結果はサーチエンジンへのアクセス条件に依存しており、したがって切り替えがその日時に一斉に行われたわけではない可能性はある。

技術提携の結果、GoogleとYahoo!の検索内容は非常に近いものになった。今や両者のヒット件数もページのランク付けもほとんど同じである（ただし、Yahoo!がGoogleサーチエンジンの提供する情報を明らかに意図的に操作していると見られる場合も現に観察された)。このことは、従来Yahoo!のサーチエンジンが持っていたGoogleに対する量的、質的な優位性を失ったことを意味する。すなわち、Yahoo!のヒット件数は従来の数分の１に減少し、また、Yahoo!もGoogle検索のヒット件数の論理的不整合と時間変動という２つの問題を抱え込むことになった。

2008年から2011年にかけての、GoogleとYahoo!での「は」というキーワードによる検索ヒット件数の推移の様子を（図２）に示す。増加を続けていたYahoo!のヒット件数が2010年11月を境に一挙に落ち込み、以後はGoogleのヒット件数とほぼ一致する値を示すようになった。(Google、Yahoo!のグラフのそれぞれに数か月線が途切れているところがあるのは、サーチエンジンの仕様変更への迅速な対処を怠ったためである。)



（図２）GoogleとYahoo!のヒット件数の推移――2008～2011年

３．Webコーパス

検索仕様上の制約が大きく、表示されるヒット件数の信頼性にも深刻な問題があるというサーチエンジンの難点を回避してインターネット文書を日本語研究に利用するには、文書に直接アクセスできるようにするしかない。それには、インターネット文書を大量に取得してローカルなデータとする――すなわち、手元のパソコン上に置く――のが最も手っ取り早く確実な方法である。そのようにインターネット文書を大量に収集・集積したデータをWebコーパスと呼ぶ。

３．１　Web コーパスの作成とその手順

発表者は2008年に約１千万件の日本語インターネット文書から成るWebコーパスを作成し、以来、日本語のコロケーション情報の抽出、形態変化の考察のための異形態分布の調査、生起頻度の高くない機能表現の意味分析などに利用してきた(拙論(2009b,2009c,2009d,2010))。[4]

Webコーパスのサイズは、先に示した（表１）にある通り、字数にして約750億字、ファイルサイズ（Shift-JIS）にして約150ギガバイト、平均的な小説単行本の40万冊弱に相当する。時間さえかければより巨大なWebコーパスを作ることができるが、現段階のパソコンで研究に利用するには上記の量がほぼ上限と判断した。

[4] 服部(刊行予定)によれば同氏も2010年にWebコーパスを作成されたとのことである。

Webコーパスは概略次のような手順で作成した。

①大量のキーワードないしキーワードの組のリストを作成
②個々のキーワード（の組）をサーチエンジンに与えて検索
③検索によって得られる最初の（最大100個の）URLが指す文書を取得
④文書の文字コードを統一し、HTMLタグなどの不要な情報を除去

実際の処理に際しては考慮を要するいくつかの問題に直面し、可能な限り適宜対処した。

①のキーワード（の組）のリストの作成については、そのサイズと内容が問題となる。サイズに関しては、1つのキーワードによる検索で得られるURLは最大100個しか利用しないので（③を参照）、1千万件の文書を収集するには少なくとも10万個のキーワード（の組）が必要であることになる。事前の試行段階では日本語の基礎的な語彙をキーワードとしていたが、それではとうてい足りないことにすぐに気付いた。そこで、主として、まず多種多様の文書を収集し、それを機械的に分解することによって多数のキーワードを得るという方法を取った。それと平行して、いくつかの特定の分野に特有の用語や表現のリストを手作業で用意する方法も行った。2通りの方法を使ったのは、前者では内容的に偏りの少ないコーパス、後者では偏りの多いコーパスが得られるものと期待されたからである。実際、コーパスを使って語句を検索してみるだけでも、そのねらいが達成されていることが確かめられる。2通りの方法で作ったWebコーパスの量の比率は2：1、それぞれ約100ギガバイト、約50ギガバイトのファイルサイズである。

③の段階に関して対処を要すると思われた問題の1つは、サーチエンジンは一部のWebサイトのページを優先的に表示する傾向があることである。その中でも特に優先度の高いWikipediaのページは収集の対象から除外することにした。

約150ギガバイトのコーパスの作成に要した所要時間は記録によれば11日である。勤務先の研究室と自宅のネットワークの両方を使って作業を進めた。いずれの環境でも複数の文書を並行的に取得した。その同時取得数を増やせば所要時間をさらに短縮できたはずであるが、ネットワークに過度の負担をかけるのを避けるために、過大な数の同時取得は控えた。それでも勤務先では研究室のデータ通信量が異常に多いことに気付いたネットワーク管理者によって通信を強制的に遮断されたことがあり、それ以後は研究室での文書取得は平日の日中は停止する形に変更した。

３．２　インターネット文書におけるデータ重複の問題

Webコーパス作成・使用の経験上、日本語研究資料としてのWebコーパスの質を低下させる最大の要因はデータの重複である（拙論(2010)）。インターネット上にはさまざまな経緯で同一の文章や文・句が複製されて出現する。完全に同一の文書が複製されて別の場所に置かれている場合以外にも、実質的に同じ文書が一部変形・加工された形で存在する、同一のブログ記事のタイトルや広告文が多数のページに出現する、掲示板などで他者の発言が引用される（そして、それがさらに引用される）などいろいろな場合がある。

Webコーパスの作成に際してそうしたデータの重複を理想的な形で排除することは現実に不可能である。と言うよりそもそも、データの重複が望ましくないことは明白でも、それをどのような基準で排除すれば理想的な処置と言えるのかという問いに対する自明の回答は存在しないと言うべきであろう。

理想には程遠いが現実に行える妥協策の1つは、同一のURLの文書を重複して収集することは避けるというものであろう。もっとも、物理的にも同一の文書が異なるURLによっ

て示されることもあるので、表面的なURLの同一性に基づく検査では同一文書の重複が見逃される可能性はある。

発表者がWebコーパスを作成したときにはデータ重複の問題に対する認識が浅く、重複排除のための努力を十分に払わなかった。作成済みのWebコーパスから重複を減らす努力をすることも可能であるが、むしろ将来新規にWebコーパスを作り直すことがあればそのときの優先課題としたいと考えている。

３．３　Webコーパスへの形態素情報などの付与

発表者の作成したWebコーパスは、インターネット文書のURLとその文書が文字情報の形で含むテキストとから成る対の集合である。各文書を得るときに使用した検索キーワード（の組）は利用価値が乏しいとの判断からコーパスに含めていないが、処理過程の記録（ログ）を用いて復元することは可能である。

「Webコーパスは単なるテキストなのか、それとも形態素解析を施しているのか」と尋ねられることがあるが、この問いはあまり実質的な意味を持たない。形態素解析ソフトを利用してテキストに形態素情報を付加するだけなら何らむずかしいところはなく、それに必要とされるのは十分な外部記憶装置の容量と処理に要する時間だけである。形態素解析を手作業で行うとか、機械的な解析結果を手作業で修正するといったことになれば話は違ってくるが、Webコーパスの規模上そうした処理は現実に不可能である。

ともあれ、今のところWebコーパスはもっぱら単なるテキストとして作成し、形態的な情報が必要となる分析の局面においてその都度必要な箇所だけを形態素解析ソフトを呼び出して解析するという方法で利用している。

３．４　Webコーパスの著作権など

「Webコーパスは譲渡・配布できるのか」という問いについては、現行の著作権法上は残念ながら否と言わざるを得ない。Webコーパスの複製・配布は、他者の著作物である１千万件の文書をコピーして配布することを意味する。実際、譲渡・配布以前に、そもそもWebコーパスを作成すること自体についても厳しく言えば法的に問題があり得るものと思われる。

Webコーパス自体でなく、それを生成するソフトウェアの配布にはおそらく問題がない。もし可能であれば、発表者がWebコーパスを作成したときのソフトウェアを整理・統合し、より容易に使える単一のソフトウェアにして公開したいと考えている。

文献

岡島昭浩(1997)「インターネットで調べる」『日本語学』第16巻第12号

荻野綱男(2004)「各種検索エンジンの実態と特徴」『日本語学』第23巻第2号

荻野綱男(2006)「検索エンジンGoogleの使い方とWWWコーパスによる日本語研究」城生佰太郎博士還暦記念論文集編集委員会編『実験音声学と一般言語学』（東京堂出版）

荻野綱男・末永絵梨・下重秋弓・三好亜萌(2007)「WWWの検索による日本語研究(2)」『東京女子大学日本文学』第103号

田中ゆかり(2003)「ネット検索は日本語の研究に有用か」『日本語学』第22巻第5号（2003年4月臨時増刊号『コーパス言語学』）

田野村忠温(2000a)「現代日本語資料としてのインターネット」『大阪外国語大学における情報処理教育・研究の高度化』、大阪外国語大学　［『電子資料と日本語研究』（私家版、2002年）に再録］

田野村忠温(2000b)「電子メディアで用例を探す――インターネットの場合――」『日本語学』第19巻第6号
田野村忠温(2003)「コーパスによる文法の研究」『日本語学』第22巻第5号（2003年4月臨時増刊号『コーパス言語学』）
田野村忠温(2008)「日本語研究の観点からのサーチエンジンの比較評価――Yahoo!とGoogleの比較を中心に――」『計量国語学』第26巻第5号
田野村忠温(2009a)「日本語研究の観点からのサーチエンジンの評価・続――検索ヒット件数の時間変動のその後とWeb文書量の推計の修正――」『計量国語学』第26巻第8号
田野村忠温(2009b)「コーパスからのコロケーション情報抽出――分析手法の検討とコロケーション辞典項目の試作――」『阪大日本語研究』21（大阪大学大学院文学研究科日本語学講座）
田野村忠温(2009c)「サ変動詞の活用のゆれについて・続――大規模な電子資料の利用による分析の精密化――」『日本語科学』第25号
田野村忠温(2009d)「『代わり』の分析試論――巨大なWebコーパスに基づく考察――」田野村忠温・服部匡・杉本武・石井正彦『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発Ⅲ」（文部科学省科学研究費補助金特定領域研究「日本語コーパス」日本語学班）
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』第138号
服部匡(刊行予定)「ウェブを利用した研究例」『講座　ＩＴと日本語研究６　コーパスとしてのウェブ』（明治書院）

デモ・ポスターセッション

3月16日（水） 11:50〜14:00

複合名詞内の係り受けに着眼したアクセント変形予測の高精度化に関する実験的検討
▶高野 克弥、峯松 信明

テキストの多様性をとらえる分類指標の構築を目指して
▶小磯 花絵、田中 弥生、小木曽 智信、近藤 明日子

BCCWJを用いた語彙・文法情報のプロファイリングとその応用
▶千葉 庄寿

中学校・高校教科書の教科特徴語リストの作成 ―語彙指導の基礎資料として―
▶近藤 明日子

ジャンル別に見た特徴漢字 ―書籍のジャンルと広報紙の漢字―
▶斎藤 達哉

社会科での漢字学習事例検討 ―小学校6年生「憲」について―
▶棚橋 尚子

コーパスに基づく分類重要語彙リスト ―学校教育での活用に向けて―
▶田中 牧郎

外形で引く国語辞典への試み
▶矢澤 真人

同時共起クラスタリングを利用した大規模テキストからの動詞類語抽出
▶竹内 孔一、高橋 秀幸、小林 大介

分類器の確信度を用いた合議制による語義曖昧性解消の領域適応
▶古宮 嘉那子、奥村 学

共起語グラフのクラスタリングによる単語の多義性抽出
▶鏑木 雄太、古宮 嘉那子、小谷 善行

教師付き外れ値検出による新語義の発見
▶新納 浩幸、佐々木 稔

SemEval-2010日本語語義曖昧性解消タスク報告
▶奥村 学、白井 清昭、古宮 嘉那子、横野 光

大規模階層辞書を用いた日本語機能表現解析体系の研究 （pp.207〜214参照）
▶宇津呂 武仁、鈴木 敬文、島内 蘭、阿部 佑亮、松吉 俊、土屋 雅稔

BCCWJを利用した日本語作文支援システム「なつめ」の評価
▶阿辺川 武、ホドシチェク・ボル、仁科 喜久子

日本語フレームネットにおけるBCCWJへの意味アノテーション
▶小原 京子、加藤 淳也、斎藤 博昭

FrameSQLで見る日本語フレームネット
▶佐藤 弘明

BCCWJを用いた語彙・構文彙の分析 ―所謂引用助詞「と」が標識する構文の場合―
▶藤井 聖子

複合名詞内の係り受けに着眼したアクセント変形予測の高精度化に関する実験的検討

高野 克称（電子化辞書班協力者：東京大学大学院工学系研究科）†
峯松 信明（電子化辞書班分担者：東京大学大学院情報理工学系研究科）‡

An Experimental Study on Improving the Performance of Accent Sandhi Prediction Using the Internal Structure of Compound Nouns

Katsuya Takano (The University of Tokyo)†
Nobuaki Minematsu (The University of Tokyo)‡

1 はじめに

日本語テキスト音声合成システムにおいて，文中のアクセント核位置を適切に推定することは，自然な読み上げ音声出力に対する必要条件である。日本語の場合，各々の単語は固有のアクセント核位置情報を持つが，それらの単語を連結し，文中で発声するとアクセント核の位置は頻繁に変化する（アクセント結合）。我々は，入力文に対する形態素解析結果を用いて定義される各種素性を入力とし，CRF（Conditional Random Fields）を用いたアクセント核位置の予測を試みてきた [1]。しかし従来の研究では，複合名詞に関して十分な精度が得られていなかった [2, 3]。原因の一つとして，長い複合名詞は内部に複数のアクセント句を有することがあるが，複合名詞内のアクセント句境界位置の推定が困難である点が上げられている。本稿では，4〜8 形態素の長い複合名詞に対して係り受け解析を行い，そこから得られる（複合名詞内）アクセント句境界を利用することで精度向上を検討した [4]。

2 CRF によるアクセント核位置変化予測

JNAS 中の文章に対して形態素解析システム Chasen（Unidic1.3.8 使用）による解析結果に,「アクセント句境界」と「アクセント核位置」の情報を単独ラベラに付与させ，学習データ（6,753 文）と評価データ（527 文）を用意した [1]。素性としては，[基本形／基本形読み／書字形／品詞／活用型][品詞][活用型][活用形][モーラ数] [単独発声アクセント型][組み合わせ素性][単独型種類ラベル][アクセント句内の相対位置] [特定位置のモーラ] 等を用い，CRF によるアクセント核位置の推定を行った。アクセント句単位での正解率が 92.6%であるのに対し，複合名詞部分の正解率は 89.0%であった [3]。以降，特に長い複合名詞を対象としたアクセント核位置推定の精度向上を狙う。

† takano@gavo.u-tokyo.ac.jp, ‡ mine@gavo.t.u-tokyo.ac.jp

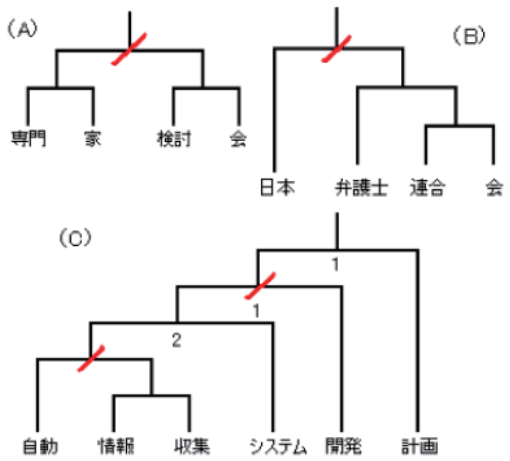


図 1: 複合名詞に対する三種類の係り受け

3 複合語名詞のアクセント核位置推定

3.1 複合名詞データの準備

4 から 8 つの形態素からなる複合名詞を 800 語（割合は，4：50%，5：25%，6：20%，7：4%，8：1%）を新聞記事より抽出した。Mecab を用いて形態素解析を行い,「単独発音アクセント核位置」を Unidic1.3.12 より取得し，連続発声時のアクセント核位置情報を人手でラベリングした。これを，学習データ 650 語と評価データ 150 語に分割して使用した。

3.2 CRF によるアクセント核位置推定

CRF を用いてアクセント核位置を推定すると，形態素単位で 77.7 %の精度が得られた。素性としては，基本的には第 2 節で用いた素性を用いている。第 2 節における複合名詞部分の精度と比べて精度が極端に低いのは，より長い（4〜8 形態素）複合名詞を用いているためである。

3.3 複合名詞に対する係り受け解析の利用

4 つ以上の形態素から形成される複合名詞は，アクセント核を複数所持することがある。そこで，複合名詞に適切なアクセント句境界を設けることが精度向上へと繋がると期待される。実際に日本人が複合名詞を朗読する際は，通常 2，3 単語を一つのアクセント句として認識することが多いことが報告されている [5]。そこで，本稿では [5] の方法に倣い，複合名詞内の係り受け解析によって得られる木構造（図 1 参照）を基に，複合名詞のアクセント句境界推定を試みた。なお [5] 同様に，複合名詞内の係り受け解析は人手で行っている。中単位解析器を使った自動解析結果の利用については後ほど述べる。人手によって導出された係り受け関係（木構造）に対し，下記の手順でアクセント句境界を推定した [5]。図 1 で斜線で示されているのが，最終的に推定されたアクセント句境界である。

1. 最上位節点から始めて，順に下位節点を走査し，4 単語以上の単語列が（アクセント句境界の挿入によって）無くなるまで 2. の作業を繰り返す。

表 1: 各形態素間の結合力

前方単語	後方単語	結合力
非用言性名詞	非用言性名詞	3
用言性名詞	非用言性名詞	2
名詞	用言性名詞	1
名詞	接尾辞	4

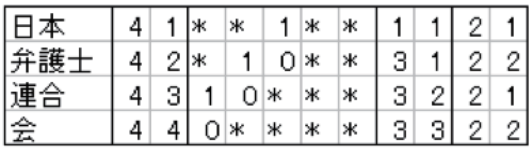


図 2: 係り受け情報やアクセント句境界情報の素性化

2. 木構造の最上位節点が，A) 両端の枝が木構造（複数の形態素から形成される纏り）の場合，B) 左枝が単語で右枝が木構造の場合，C) 左枝が木構造で右枝が単語の場合のいずれであるかを判断し，以下の処理を行う。

   A) の場合，最上位節点に境界を設定する。

   B) の場合，最上位節点に境界を設定する。

   C) の場合，最上位から 3 つの節点までが左枝が木構造で右枝が単語の時，表 1 に従って，それぞれの形態素間の結合力を求める。ただし，この途中で A) や B) の構造が現れた時は，その節点に境界を設定する。3 つの結合力が得られたところで，前方から相対的に結合力が弱い場所を探し，その節点に境界を設定する。

3.4 アクセント核位置推定実験

木構造（図 1）の様子を素性化し，第 3.2 節で用いている素性に追加した。図 2 の左から 3 番目の枠内に木構造を 1/0 で表現している。この結果，アクセント核の推定精度は 81.2% と向上した。

次に，木構造から得られるアクセント句境界を第 3.2 節の素性に追加した。図 2 の左から 4 番目の枠内であり,「日本」と「弁護士連合会」に分割されることを意味する。この場合，アクセント核の推定精度は 85.7%へと向上した。

3.5 考察

以上の実験は，複合名詞に対して人手で係り受け解析を行わせた場合の結果である。特定領域研究「日本語コーパス」電子化辞書班により提供された中単位解析器を用いて係り受けを自動抽出し，これより定義される境界情報を，第 3.4 節と同様に素性化して実験を行ったところ，79.1%の向上しか観測されなかった。例えば，図 2 の一番右の枠に示すように，この解析器は「日本弁護士連合会」に対して,「日本弁護士」と「連合会」に（不適切な）分割

を行ってる。長い複合名詞に対して，より高精度にその内部的な係り受けを抽出するツールの開発が望まれる。

4　まとめ

CRF を用いた従来のアクセント核位置推定において，高い精度を示すのが困難であった複合名詞に焦点をあて，係受けに基づくアクセント句境界推定を行うことで，複合名詞に対する精度向上の可能性を示すことができた。しかしながら，アクセント句境界推定に必要な係り受け解析を高精度に行うツールがまだ存在しておらず（文を対象とした係り受け解析器は存在するが，複合名詞を対象としたものは無い），今後，これらのツールの開発が望まれる。

参考文献

[1] 黒岩，峯松，伝，広瀬，“大規模アクセントラベリングコーパスの構築とそれに基づくハイブリッド型アクセント結合処理”，電子情報通信学会音声研究会，SP2006-174, pp.31-36 (2007)

[2] 印南，渡辺，峯松，広瀬，“CRF に基づくアクセント変形予測モデルにおけるエラー解析”，言語処理学会年次大会発表論文集，pp.969-972 (2008)

[3] 印南，渡辺，峯松，広瀬，“規則処理を参考にした CRF によるアクセント結合モデル”，日本音響学会春季講演論文集，2-P-13，pp.473–476 (2009)

[4] 高野，清水，峯松，広瀬，“複合名詞内アクセント句境界を用いた アクセント結合予測の高精度化に関する実験的検討”，日本音響学会春季講演論文集 (2011)

[5] 藤石, 宮崎，“日本語複合語構造解析に基づく複合語アクセント句の自動抽出法”，情報処理学会第 49 回全国大会，2–51 (1994)

テキストの多様性をとらえる分類指標の構築を目指して

小磯　花絵（電子化辞書班分担者：国立国語研究所理論・構造研究系）†
田中　弥生（データ班協力者：国立国語研究所コーパス開発センター）
小木曽智信（電子化辞書班分担者：国立国語研究所言語資源研究系）
近藤明日子（言語政策班連携研究者：国立国語研究所コーパス開発センター）

Towards the Construction of the Classification Criteria for the Variation in Written Texts

Hanae Koiso (National Institute for Japanese Language and Linguistics)
Yayoi Tanaka (National Institute for Japanese Language and Linguistics)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Asuko Kondo (National Institute for Japanese Language and Linguistics)

１．はじめに

『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）には，書籍や新聞，雑誌，中央官庁刊行の白書，Web 上のデータ（Yahoo!知恵袋・Yahoo!ブログ）など，さまざまな媒体・ジャンルのデータが含まれており，BCCWJ を使うことによって，例えば，書籍と新聞，インターネット上のテキストの文体がどことなく異なるという直観を，定量的分析を通して言語特徴の差として具体的に捉えることが容易に行えるようになった。

例えば，小磯ほか (2008a, 2008b, 2009, 2010) や宮内 (2009) はそうした取り組みの一つである。小磯ほか (2008a, 2008b) では，書籍・新聞・白書を取り上げ，テキストに含まれる各品詞率（名詞率や動詞率など），語種率（漢語率や和語率など），異なり語率，文の長さなどに着目してジャンル毎の比較を行った結果，多くの素性に関して媒体毎に異なる傾向が見られること，また線形判別分析を用いてこれらの素性から当該テキストのジャンルを推定するモデルを構築した結果，leave-1-out 交差検証において約 94.4%という高い精度でジャンルの判別が行えることが分かった。小磯ほか (2009) では，上記３ジャンルに WEB データ（Yahoo!知恵袋），国会会議録，『日本語話し言葉コーパス』から学会講演と模擬講演を加えて同様の分析を行い，小磯ほか (2008a, 2008b) ほどではないが 79.9%という比較的高い判別率を得ている。これらの結果は，媒体やジャンル毎に特徴的な言葉の使い方が存在することを示唆するものである。しかしその一方で，書き言葉の多様性は，新聞・書籍・インターネットといった媒体やジャンルの違いでは捉えきれない広がりがあることも事実である。

例えば次の二つのテキストを見てみよう。いずれも行政白書の「環境」に関するテキストである。白書は全般的に，名詞や内容語などの使用率が高い傾向にあり（小磯ほか 2008a, 2008b），Halliday (1985)により文章の複雑さの指標として提案された語彙密度を用いて白書と書籍（文学）の比較を行った佐野・丸山 (2008) の結果と合わせると，白書は文章としてより複雑でフォーマルな性質を有するということになる。

† koiso@ninjal.ac.jp

【例１】また、林齢の高い人工林における適切な密度管理、公益的機能の低下した保安林を複層林へ誘導・造成するなど、育成複層林施業、長伐期施業等により二酸化炭素を長期にわたって固定し得る森林づくりを推進するとともに、育成に長期間を要する広葉樹の特性に応じた保育を進めるなどの適切な整備や針広混交林化を推進する。（ID:OW6X_00029）

【例２】ひとえに循環型社会といっても、その言葉から想像される社会は様々です。昨年の循環型社会白書では、循環型社会のイメージとして 3 つのシナリオを示しました。（ID:OW6X_00011）

これらの例を見てみると，確かに行政報告書ということもありいずれもフォーマルなテキストではあるが，例 1 は文も長く構文も複雑で，短い文の続く例 2 と比べると，文章としてより複雑で堅苦しいといった印象を受ける。このような印象の違いは，媒体やジャンル，主題の違いでは捉えきれないものであり，これらとは異なる次元の指標の体系化が求められる。

この種の類型化・体系化の試みは，文体研究や理論研究の中で古くから行われており，様々な観点や指標が提案されてきた。例えば永野 (1968) は，日本語の文章文体研究で取り上げられてきた分類の視点を，「機能」「内容」「形式」という 3 つの分類基準とその具体的設定項目に整理し，体系化を試みている。また英語を中心とした研究では，Halliday が状況のコンテクストとして，「活動領域（field）」「役割関係（tenor）」「伝達様式（mode）」の 3 つを提案している（Halliday and Hassan (1985) など）。言語表現自体を直接分類するものではなく，言語選択に影響を与えるものとして状況のコンテクストを体系化したものであるが，書き言葉・話し言葉を体系的にとらえる上で重要な指標と言える。

しかしこれらの観点や指標によって，多種多様な書き言葉が具体的にどのように，またどの程度妥当に分類できるのかといったことを実証的に評価した取り組みは，少なくとも日本語の研究を見る限りあまり行われていない。そこで著者等は，従来指摘されてきた文章を評価・分類する指標を参考に八つの分類尺度を構成した上で，400 のテキストを対象に被験者に 5 段階で評定してもらうという実験を実施したが，少なくとも評定結果からは，多様なテキストを有効に分類することはできなかった。

そこで，理論的側面から類型化・体系化を試みるという方向を離れ，まず人が種々のテキストを読んだ際に感じる印象を表わす表現を調査し，実際のテキストに対して評定実験を行った上で，分類指標を探索的に体系化することを試みる。本報告ではこのうち，今までに実施した (1) テキストの印象を評定する表現の収集調査，(2) 評定実験，(3) 評定結果の分析に基づく分類指標試作版の構築について報告する。

２．評定語の収集

一般の人がテキストを描写・評価する際に用いる様々な表現を収集することを目的に，調査者に 40 のテキストを読んでもらい，それぞれのテキストから受ける印象を記述してもらうという調査を実施した。以下に調査の具体的手続きと収集した評定語の概略を記す。

２．１　方法

資料：BCCWJ のうち自動解析結果を人手修正した精度の高い「短単位」「長単位」情報が付されたコア（小椋ほか 2010）から，新聞（5 サンプル），小説以外の書籍（10 サンプル），

雑誌（10 サンプル），行政白書（4 サンプル），Yahoo!ブログ（11 サンプル），計 40 サンプルを選んだ。小説には複数の人物の会話文が多く含まれている可能性が高く，テキストから受ける印象を一意に決めづらいことが多いため対象外とした。同様の理由で引用の多いサンプルは対象外とした。

各サンプルのサイズは約 300 文字（300 文字を越えて最初に現れる文の文末まで）とした。文字数で区切っているため，必ずしも内容的にまとまった単位にはなっておらず，話題の途中で始まったり，あるいは途中で切れたりしているものもある。

調査者：7 名の調査者（男性 1 名，女性 6 名）が調査に参加した。

手続き：調査者には二つのテキストの対が一つの頁に記された調査票 20 頁（20 対 40 サンプル）が渡された。サンプルの組合せや出現の順番はランダムに決定し被験者毎に異なる。

調査者は二つのテキストを読んだ上で，まずそれぞれのテキストから受ける印象やそれを描写する表現を思いつく限り自由に記述してもらった。次に，二つのテキストに共通する印象，対立する印象があれば，それを記してもらった。対立する印象については，例えば「硬い－柔らかい」のように，対語形式とするよう依頼した。テキストを比較して共通する印象，対立する印象を考えることで，個々のテキストを読むだけでは思いつかないような評定語が出てくる可能性があると考え，このような手続きを踏んだ。

なお調査者には，テキストの内容に対する印象ではなく，テキストの表現や文体から受ける印象を記すよう具体例を挙げて注意を促した。例えば殺人に関するテキストを読んで「怖い」と評価するのではなく，殺人に関するリアルな文章を読んで「臨場感あふれる」と評価するように，といった指示である。

２．２　収集した評定語

前節に記した手続きの結果，調査者あたりおよそ異なりで 80～180，7 名全員で 441 の評価語が抽出された。その上で，次の手続きに従い評定語対を作成した。

まず，「練られた」「良く練られている」「文章が練られた」のような類似した表現の評価語をまとめ上げた上で，「練られた」「推敲された」「整理された」のように類似したカテゴリーのものをまとめ上げた。複数の類似した表現によって一つのカテゴリーを構成しないもの（例えば「こびるような」など）は，テキストの印象を表す表現として典型的ではないと判断し除いた。また，文章の内容や著者自身に関する印象と解釈される，あるいは解釈される危険性の高いもの（例えば「器の大きい」など）も省いた。その上で，各カテゴリーの中から対語形式の評定尺度を作成した場合に一番自然で曖昧性の少ないものを選択した。例えば「簡潔性」に関する評価語では，「簡潔な」「削ぎ落とされた感じ」「無駄のない」「煩雑な」「ごちゃごちゃした」「冗長な」「簡潔－冗長な」「すっきりした－ごちゃごちゃとした」「まとまり感のある－まとまり感のない」などが分類されたが，この中から対語にした場合に表現として一番自然だと考えられる「簡潔な－冗長な」を選択した。

以上の手続きにより評定語の整理をした結果，次の 20 の評定語対が構成された。

- 改まった－くだけた
- 型にはまった－個性的な
- よく練られた－練られていない
- 整然とした－雑然とした
- 簡潔な－冗長な
- 自然な－わざとらしい
- 直接的な－婉曲的な
- 客観的な－主観的な
- 臨場感のある－臨場感のない
- 具体的な－抽象的な
- 読み手に語りかける－語りかけの少ない
- 書きことば的－話しことば的
- 相手の理解を配慮した－相手の理解を無視した
- 重い－軽い
- 暗い－明るい
- 冷静な－興奮した
- 硬い－柔らかい
- めりはりのある－単調な
- テンポのよい－テンポの悪い
- 親しみやすい－とっつきにくい

３　評定実験

３．１　方法

安定した尺度を構成するため，得られた 20 の評定語対をもとに，テキストに対する印象について SD 法による 5 段階評定実験を実施した。テキストサンプルとして，前節に示した評定語の収集調査と同じ 40 サンプルを用いた。評定語の収集調査とは異なる 3 名の被験者（男性 2 名，女性 1 名）が実験に参加した。被験者には，テキストを熟読した上で，20 の評定尺度に基づき 5 段階で評定してもらった。サンプルは被験者毎にランダムに配置した。また評定尺度は適宜左右を反転させた。本番に先立ち練習問題として 5 サンプルを評定してもらった。被験者には，前節の評定語の収集調査と同様，テキストの内容に対する印象ではなく，テキストの表現や文体から受ける印象に従って評定してもらうよう，複数の具体例を挙げて指示した。

３．２　結果

得られた評定結果を対象に，被験者毎に各評定尺度毎の分散を求めたところ，「具体的な－抽象的な」と「直接的な－婉曲的な」については，いずれの被験者においても分散が小さく評定値が偏っていることが分かった。この偏りは，単純に評定対象としたサンプルに偏りがあったためとも考えられるが（例えば「婉曲的」なテキストがほとんどなかったために「直接的」に寄ってしまった，など），評定尺度の設定自体に問題があった可能性も十分にあるため，分析対象外とした。

上記を除く 18 の評定尺度を対象に，因子分析（最尤法，バリマックス回転）を行った。事前に主成分分析を行い，固有値などから 3 因子が一つの目安であると判断した（累積寄与率 80.2%）。因子分析の結果，因子負荷量がいずれも小さい値しか示さない評定尺度や複数の因子にまたがって高い因子負荷量を持つ評定尺度は削除した。これを繰り返した結果，最終的に 12 の評定尺度が残った。各尺度の因子負荷量を表 1 に示す。

表１　因子分析の結果－因子負荷量－

	因子１	因子２	因子３
改まった－くだけた	**0.945**		-0.279
硬い－柔らかい	**0.919**	-0.147	-0.180
重い－軽い	**0.908**	-0.108	-0.307
型にはまった－個性的な	**0.895**		-0.129
書きことば的－話しことば的	**0.893**	-0.144	-0.217
冷静な－興奮した	**0.837**	0.323	-0.207
暗い－明るい	**0.707**	-0.181	-0.470
簡潔な－冗長な	0.120	**0.831**	
整然とした－雑然とした	0.124	**0.817**	0.362
自然な－わざとらしい	-0.325	**0.746**	0.397
めりはりのある－単調な	-0.241	0.285	**0.802**
テンポのよい－テンポの悪い	-0.197	0.208	**0.767**
寄与率	48.40%	18.60%	15.80%

因子３の結果から見ていこう。因子３については，「めりはりのある－単調な」と「テンポのよい－テンポの悪い」など，抑揚や速さ感に関する尺度が相対的に高い正の負荷を示しており，文章の抑揚・リズムに関する因子であると解釈することができる。一方，因子２は，「簡潔な－冗長な」「整然とした－雑然とした」「自然な－わざとらしい」が相対的に高い正の負荷を示しており，文構成の明晰性に関わる因子であると解釈することができる。

一方，因子 1 は七つもの尺度が高い正の負荷を示している。このうち「改まった－くだけた」や「硬い－柔らかい」は文章のスタイルに関するものであり，そのスタイルの違いにより，軽重，明暗，動静などの印象が派生したと考えると，これは文章のスタイルに関する因子と解釈することができる。しかし「書きことば的－話しことば的」については，確かに「話し言葉的」といった場合に口語調のくだけた印象が喚起されることから，スタイルとの関連性があると考えられる一方で，「話しことば的」であっても改まり度の高い文章が容易に想像つくことからも分かるように，必ずしも同種の尺度ではない。

次に挙げる二つのテキストは，それぞれ「改まった，かつ，話し言葉的」「くだけた，かつ，書き言葉的」と判断されたサンプルの一部を抜粋したものである。

【例３】では週別に販売数の推移を追い、どのように商品構成を変え、売れ筋の棚割と在庫を変え、販売数を伸ばしたかを**見てみよう**。(ID:PB46_00066)

【例４】お会計は現金で**おやぢ**とやり取りします。麺・だし・天ぷらといたって**フツー**ですが安心して食べられます。(ID:OY14_04336)

例 3 は決してくだけた表現は用いられていないが，読み手に対する働き掛けの表現が含

まれており，これが「話しことば的」という印象に影響を与えたと考えられる。実際，因子分析の過程で落とされたが「読み手に語りかける－語りかけの少ない」という評定尺度で語りかけの程度が高いと判断されたものである。一方例 4 は，その種の働き掛けはないが（語りかけの程度が低いと判定），「おやぢ」や「フツー」といった表記の仕方がくだけた印象を与えていると考えられる。しかし次の例のように，くだけた表現が多く使用されるものは，読み手への働き掛けの有無に関わらず話しことば的と判断されることも多い。

【例５】汗かいて帰ってきたら、シャワー浴びてすっきりして冷蔵庫から冷たいモノ出して、、、っていうのが超天国な気分。さすがに今日は、冷たいモノ食べたい気分なのでそうめんにしちゃいました。（ID:OY01_00848）

このように「書き言葉的－話し言葉的」という尺度は，改まりの程度や読み手に対する働き掛けの程度など，複数の観点が関与する多義的な意味合いをもつ尺度である可能性がある。今回のデータではスタイルに関わる因子と強い関係を示したが，今後サンプルのバリエーションを増やして慎重に検討する必要がある。

３．３　テキストの分類

前節までの分析で得られた三つの因子を用いて暫定的にテキストを分類し，どのような種類のテキストが分類されるかを概観する。ここでは単純に，各因子毎に，それと強く関係する評定尺度の平均を取るという方法で代表値を算出した。以下に具体例を示す。

例 6 と例 7 はいずれもスタイルと文構成の明瞭性が高いサンプルであるが，例 6 は抑揚・リズム性が低く，例 7 は高いという点において異なる。例 6（行政白書），例 7（新聞）ともに，過去に実施した事柄・過去の出来事を伝達しているものだが，例 6 は全ての文が過去形で変化がないのに対し，例 7 は時制も過去形・現在形と変化があり，またアスペクトや伝聞形式が使用されるなど，バリエーションに富んでいることが分かる。次に挙げる例 8・例 9 も抑揚・リズム性の高いサンプルであるが，やはり文末文体の変化や読み手に対する働き掛けなどが見られる。また例 6 では文の途中を省略したため分かりづらいが全般的に長い文が続くのに対し，後者は比較的短い文によって構成される。例 7 と同カテゴリーのテキストを概観すると，文長が相対的に短い，あるいは長い文と短い文が混在するものが多い傾向が見られる。このような違いが抑揚・リズムの違いに結びついた可能性がある。

【例６　スタイル:高，文構成明晰性:高，抑揚リズム:低】

「森林・林業・木材産業分野の研究・技術開発戦略」及び「林木育種戦略」に基づき、≪省略≫効率的かつ効果的に**推進した**。独立行政法人森林総合研究所及び独立行政法人林木育種センターにおいては、≪省略≫研究・技術開発等を**実施した**。また、研究・技術開発等の実施に当たっては、≪省略≫評価と見直しを**行った**。（1）試験研究の効果的推進　森林・林業・木材産業分野の研究・技術開発戦略に基づき、試験研究の効果的・効率的推進を**図った**。　（ID:OW6X_00007）

【例７　スタイル:高，文構成明晰性:高，抑揚リズム:高】

一方、公立高一年の少女（十六）＝殺人予備容疑で逮捕＝は、犯行に使うための文化包丁を学校帰りに百円ショップで**購入していた**。価格や切れ味などの点で二人の凶器に隔たりがあり、河内長野署捜査本部は、それぞれの殺意に関連があるとみて**調べている**。調べでは、少年は犯行の三日前の先月二十九日午後一時ごろ、自宅に近いホームセンターに一人で訪れ、刃渡り約二十センチの刺し身包丁を購入していた**という**。（ID:PN3d_00013）

次に挙げる例８と例９は，共にスタイルが低く抑揚・リズム性の高いサンプルであるが，前者は文構成の明瞭性が高いのに対し後者は低いという違いが見られる。例 8 は書籍のサンプル，例 9 はブログのサンプルである。被験者にはサンプルの出典に関する情報は一切与えられていない。同じスタイルが低めで抑揚・リズム性の高いテキストでも，執筆・出版過程でおそらく十分な推敲がなされたテキストと，相対的に十分な推敲がなされずその日の出来事を思い付く順に記したテキストでは，文構成の明晰性の観点でその差が出てくるということであろう。

【例８　スタイル:低，文構成明晰性:高，抑揚リズム:高】

こう言いますと、たいていは会場からどっと笑い声が起きます。 おそらく、ご自分はそこまでひどくはないと思いながらも、みなさん多少は身に覚えがあることだからでしょう。それだけ世の母親というものは、わが子の欠点をリストアップすることに熱心で、わが子の悪い点に関しては権威でいらっしゃいます。だから、お子さんの悪いところをあげてくださいなどと言おうものなら、それこそいくらでも出てきて際限がありません。（ID:PB23_00051）

【例９　スタイル:低，文構成明晰性:低，抑揚リズム:高】

汗かいて帰ってきたら、シャワー浴びてすっきりして冷蔵庫から冷たいモノ出して、、、っていうのが超天国な気分。さすがに今日は、冷たいモノ食べたい気分なのでそうめんにしちゃいました。でも、それだけだとなんなので　夏野菜たっぷりサラダと激辛のフライドチキン。ＣＭでレッドホットでしたっけ、やってるの見て「美味しそう?」と。。。米国に市販のメニューをそっくりに作るなんちゃって本があるというのですがフライドチキンも載ってます。（ID:OY01_00848）

４．おわりに

本研究では，一般に人がテキストを読んだ際に感じる印象を表わす表現を調査し，実際のテキストに対して評定実験を行った上で，分類指標を探索的に体系化することを試みた。分析の結果，「スタイル」「文構成の明晰性」「抑揚・リズム性」という三つの因子が抽出された。これらの因子に従い暫定的にテキストを分類してその違いを概観したところ，たしかにテキストの多様性（の一面）を捉えており，少なくとも全く妥当性・有効性のない指標ではないことはうかがわれたものの，例えば「スタイル」の因子に，必ずしもそれだけでは捉えきれない「書きことば的－話しことば的」が含まれるなどの問題も見られた。勿論，今回評定の対象としたサンプルは 40 とかなり少ないため，サンプル数やバリエーションを増やした上で検討し直す必要があるが，今回の一連の研究を振り替えって率直に感想

を述べるならば，この方法でテキストの分類指標を作成することは果たして可能なのかという疑念も残る。

本研究はもともと『日本語話し言葉コーパス』のために設計・付与された印象評定データ（山住ほか 2005）に着想を得て始めたものである。話し言葉では，話している内容や言葉の選び方だけではなく，声の抑揚や声質，発話速度，間の取り方など，声や話し方に関する様々な特徴も聞き手の印象形成に強く影響を与えており，またそもそも前者より後者の方が話し言葉における聞き手への印象形成に強く影響しているという指摘もある（トラッドギル 1975)。今回の書き言葉を対象とした実験では,後者の声や話し方に関する特徴(に相当するとも考えられる紙面での文字のサイズやフォント，配置の仕方など）は含まれておらず，あくまで選択された言語表現だけが評価の対象となる。このような情報だけで読み手の印象を個人内・個人間で安定的に取り出すことは容易ではない可能性もある。勿論，一足飛びに不可能という結論を得るのではなく，このような問題を念頭に置きつつ，今回得られた結果を慎重に検討し，必要に応じて追加実験をするなどして，この方法でどこまでテキスト分類尺度の構築が可能かを追求して行きたい。

文献

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕(2010)『『現代日本語書き言葉均衡コーパス』形態論情報規定集第３版』国立国語研究所内部報告書.

小磯花絵・小木曽智信・小椋秀樹・冨士池優美・宮内佐夜香(2008a)「『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・小木曽智信・小椋秀樹(2008b)「短単位情報に基づくジャンル間の文体に関する分析」『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集』pp. 99-106.

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

小磯花絵・小木曽智信・小椋秀樹・宮内佐夜香(2010)「長単位情報に基づくジャンル間の文体に関する分析」『特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ(研究成果報告会)予稿集』pp. 183-190.

佐野大樹・丸山岳彦（2008)「システミック文法に基づく書きことばの複雑さ測定 －日本語大規模コーパスを用いた語彙密度計測－」『言語処理学会第 14 回年次大会予稿集』, pp. 1097-1100.

トラッドギル，ピーター（1975)『言語と社会』(土田滋訳)，岩波書店.

永野賢(1968)「文章の分類論」森岡健二ほか（編)『作文講座 4 文章の理論』, 明治書院.

宮内佐夜香・小木曽智信・小椋秀樹・小磯花絵(2009)「BCCWJ における接続表現形式とジャンル別の文体的特徴の関連について」『特定領域研究「日本語コーパス」平成 21 年度全体会議予稿集』pp. 99-106.

山住賢司・籠宮隆之・槙洋一・前川喜久雄（2005)「講演音声の印象評定尺度」,『日本音響学会誌』61 巻 6 号, pp. 303-311.

Halliday(1990) Some grammatical probalems in scientific English, *Annual Review of Applied Linguistics*, **6**, pp. 13-37.

Halliday and Hassan(1985) *Language, Context and Text: A Social Semiotic Perspective*, Deakin University Press.

BCCWJ を用いた語彙・文法情報のプロファイリングとその応用

千葉庄寿（日本語教育班連携研究者：麗澤大学外国語学部）†

Doing Lexical and Grammatical Profiling with BCCWJ

Shoju Chiba (Faculty of Foreign Studies, Reitaku University)

１．BCCWJ をもちいた語彙・文法情報の評価

英語コーパス言語学の初期の展開において，辞書学をはじめとする語彙研究への関心が重要な役割を果たしたことが知られている(Biber *et al*.1998)。日本の英語教育においても，基本語リストの作成(大学英語教育学会基本語改訂委員会編 2003)や英和辞典の編纂などに語彙教育への大規模コーパスの応用事例をみることができる。

2011 年に公開される予定の『現代日本語書き言葉均衡コーパス』(BCCWJ)は，サンプリング手法を用いて収録するサンプルに(少なくとも部分的に)統計的な代表性をもたせた大規模な「均衡コーパス」である(前川 2007:14; 丸山 2009:129)。このような「書き言葉のサンプル」たる設計思想をもつコーパスは「サンプルコーパス」(sample corpus, 齊藤ほか 2005[2]: 23)とも呼ばれ，後藤(2003: 8-9)が述べる，言語研究用に設計された「最狭義のコーパス」の最右翼の候補として，日本語の研究において未だ立ち後れている大規模コーパスを活用した定量的な語彙研究に画期的な活路を開くことが期待できる。

具体的には 2 つの活用方法が考えられよう。第一に,「書き言葉のサンプル」である BCCWJ そのものを分析し，さまざまな場面・用途に応用できる語彙データを得ることができる。実際に，BCCWJ の応用と評価を目的とした BCCWJ の研究班(研究項目 B01)のいくつかが BCCWJ の定量的な語彙研究に取り組んでいる。例えば，言語政策班は「国語政策や国語教育に役立つさまざまな語彙表を作成していくための基盤として，分野ごとの特徴度の設定と，頻度に基づく語彙レベルの設定，という二つの作業を行う」(田中 2009: 666)ことをその主要な任務としている (*cf.* 前川 2006: 1-2)。

一方，BCCWJ の利用価値は BCCWJ そのものの語彙の研究にとどまるものではない。「書き言葉のサンプル」としての BCCWJ との比較を通じ，他のコーパスデータの語彙特徴を測ることもできる。このようなコーパス間の比較の手法は BCCWJ のプロジェクトでも議論されている。近藤(2008)は，対数尤度比を指標として用い，形態素解析されたデータを用いた語彙を計量した特徴語抽出の手法を用いて教科書の語彙特徴を分析している。同様の手法を用いて，日本語教育班でも BCCWJ に基づいた日本語教育のための語彙リストの作成を試みている(橋本ほか 2008; 山内編 2008)が，こちらは BCCWJ の書籍データを話題の内容に応じた小規模なサブセットに分割し，個々の話題データの語彙特徴を抽出するものである。

本稿が射程とするのは後者であり，BCCWJ を短単位辞書 UniDic (伝ほか 2007)を用いて解析し作成した語彙情報データベースに基づき，BCCWJ の語彙・文法情報と他のコーパス(テクスト)の語彙・文法情報との比較を手軽に行うシステムの構築を報告する(本システムの公開情報については論文末を参照)。本ポスターではまた，BCCWJ の語彙情報データベースを利用した語彙・文法情報の分析ツールを用い，日本語教育における教材の開発と評価への

† schiba@reitaku-u.ac.jp

活用事例を紹介するとともに，現在の課題と今後の展望を述べる。さらに，BCCWJ によるテクストの評価についてのより広範な応用の可能性についても議論し，語彙・文法に関する信頼できる量的情報を将来どのように活用できるかを模索したい。

日本語教材に語彙情報を付与する試みとして，これまで「リーディング チュウ太」(川村 2000)や「あすなろ」(仁科 2000)などの優れた日本語読解学習支援システムが構築されてきている(各サービスの URL は論文末を参照)。しかし，これらいずれも教材テクストの分析による語彙頻度や「日本語能力試験」の語彙レベルなどの情報は考慮するものの，大規模サンプルコーパスの語彙・文法情報を活用するには至っていない。また，コロケーション情報に基づく語彙分析に利用できるオンラインツールとして日本語用例・コロケーション抽出システム「茶漉」(深田 2007)があるが，教材データなど自前のデータの分析目的に簡便に利用することはできない。

２．語彙情報データベースと語彙・文法情報分析ツール

本稿で構築した BCCWJ の語彙情報データベースは，動作が軽く，インストールおよびデータベースファイルの扱いが簡単なパブリック・ドメイン [1] の関係データベースエンジン(RDBMS)である SQLite 3.7.x で構築する。

語彙情報データベースは簡便を期し，単独出現する短単位の語彙素のテーブルと 2 グラム bigram の頻度に関するテーブルの 2 種類について BCCWJ の語彙情報を収録している。前者については UniDic の短単位の語彙素と品詞のペアを「レマ」lemma (「レンマ」とも)としてインデックスを作成した。後者は隣り合う 2 つの短単位のレマのペアについてインデックスを作成している。

分析にあたっては，分析対象のデータを UniDic で事前に解析する必要がある。本システムの利用に際しては，Windows 環境で手軽に利用できる UniDic の解析フロントエンドである「茶まめ」を使って分析対象のファイルを解析し，結果をファイルに出力しておく。BCCWJ の語彙情報データベースと分析データの解析に同じ解析環境(同一バージョン，同一環境設定の UniDic)を使うことにより，出力結果をシームレスに対応させ，齟齬なく評価することができる。

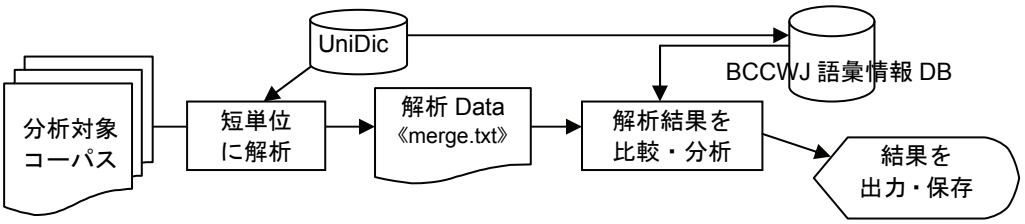


図 語彙・文法情報分析システムの概略

BCCWJ の語彙情報データベースを使い，分析対象となる語彙・文法情報を分析するツールをオフライン用のスクリプトと Web サーバ上で動作する CGI システムとしてそれぞれ Perl で構築した。Perl は ActivePerl (製品 URL は論文末を参照)の 5.8 以降のバージョンであれば，標準 [2] で SQLite が利用できるようになっている。

---

[1] http://www.sqlite.org/copyright.html

[2] 通常，Perl は CPAN (Comprehensive Perl Archive Network) を通じてモジュールを入手す

語彙・文法情報分析ツールが実装している分析手法と指標は以下の 3 種類である。

1. 語彙頻度：分析対象に現れた短単位の語彙素と品詞のペア(レマ)について BCCWJ と分析対象のコーパスの頻度を検索し，両者の数値を対数尤度比(Log-Likelihood Ratio，以下 LLR, *cf*. Kilgarriff 2001; 近藤 2008)で比較する。
2. 2 グラムの頻度：隣り合う 2 つの短単位の基本形と品詞のペアについて LLR で比較する。
3. コロケーションの計量：隣り合う 2 つの短単位の基本形と品詞のペアについて，各短単位の出現頻度と共起頻度を元に MI スコア，*t* スコアを算出し，比較する。

現在のバージョンではデータベースのサイズの問題で活用型情報による分析は行わず，語彙素情報と品詞情報のみを扱っている。

分析するコーパスが複数のファイルからなる場合には，「茶まめ」を使い解析結果を単一ファイルに出力(merge)しておくことにより，各分析ツールの結果に分析対象のコーパスデータの文書数をもとに各語彙情報の出現割合を出力する。これにより，例えば，専門用語の偏りなど，該当する用語がどの程度偏って出現しているかどうかを確認できる。

さらに，上記分析ツールは BCCWJ 全体の頻度に加え，BK (書籍)，OW (白書)，OM (国会議事録)，OC (Yahoo!知恵袋)の 4 つのサブコーパスについて，それぞれの頻度情報・出現割合の情報を出力できる(どの数値を出力するかはオプションで指定することができる)。その結果，分析対象のコーパスについて，レマの頻度，2 グラムのレマの頻度，2 グラムのコロケーション情報を BCCWJ の 5 種の集合(コーパス全体または 4 つのサブコーパス)と比較できる。

３．今後の開発・応用の方向性

他のコーパスを比較・評価するための資料としての均衡コーパスの有効性を論じる場合，以下のような基本的な問いに答える必要がある。

- どのようなサイズのコーパスデータでもその語彙的特徴を適切に比較できるか。
- どのような指標がコーパス間の比較に適するか。
- 機能語と内容語のような，出現頻度の大きく異なる語彙に同一の統計指標が適用できるか。
- どのような情報を組み合わせることで最も効果的に語彙情報を読み取ることが可能か。
- どのようなインターフェースを使うことでユーザが語彙分析を手際よく進められるか。
- BCCWJ との比較により得られた語彙・文法特徴をどのように応用できるか。

これらの問いに対する答えは，大小さまざまなコーパスを BCCWJ と比較対照しながら模索していく必要がある。本稿はこれら語彙・文法情報のプロファイリング(profiling)の手法とその活用方法の研究に取り組むための出発点と位置づけることができよう。

る。ActivePerl の場合，PPM (Perl Package Manager)を用いることで SQLite の動作に必要なモジュール(DBI, DBD::SQLite)の導入状況を確認し，必要に応じて簡単に追加・更新することができる。詳細は分析ツールに付属するマニュアルを参照されたい。

なお，本稿では短単位情報のみを扱う語彙・文法情報分析システムの構築を報告したが，BCCWJ はその言語単位として，検索や分析の目的に応じ長単位と短単位を使い分けることを当初から想定しており(伝ほか 2007)，教育等の目的には短単位よりも長単位のほうがふさわしい場合が多い(*cf.* 山内 2009)。現在，長単位の仕様はほぼ固まってきており(小椋ほか 2010[3])，今後長単位情報を付与したコーパスが普及していくものと考えられる。

文献

小椋秀樹ほか (2010[3]), 『「現代日本語書き言葉均衡コーパス」形態論情報規程集』(第 3 版, 特定領域「日本語コーパス」データ班研究成果報告書 JC-D-09-02).

川村よし子 (2000), 「インターネット時代に対応した読解教育」(『新世紀之日語教学研究国際会議論文集』), 東呉大学, 中華民国, pp. 347-365. (http://language.tiu.ac.jp/taiwan.pdf)

後藤斉 (2003), 「言語理論と言語資料―コーパスとコーパス以外のデータ―」『日本語学』22/5: 6-15.

近藤明日子 (2008), 「特徴度の設定」(特定領域「日本語コーパス」言語政策班中間報告書 JC-P-08-01), pp. 13-16.

齊藤俊雄ほか(編) (2005[2]), 『英語コーパス言語学：基礎と実践』(改訂新版), 研究社.

大学英語教育学会基本語改訂委員会(編) (2003)『大学英語教育学会基本語リスト JACET List of 8000 Basic Words』大学英語教育学会.

田中牧郎 (2008), 「語彙レベルの設定」(特定領域「日本語コーパス」言語政策班中間報告書 JC-P-08-01), pp. 7-12.

田中牧郎 (2009), 「言語政策に役立つ，コーパスを用いた語彙表・漢字表などの作成と活用」『人工知能学会誌』24/5: 665-672.

深田淳 (2007), 「日本語用例・コロケーション抽出システム『茶漉』」『日本語科学』22: 161-172.

伝康晴ほか (2007),「コーパス日本語学のための言語資源―形態素解析用電子化辞書の開発とその応用」『日本語科学』22: 101-123.

仁科喜久子 (2000), 「オンライン教材『あすなろ』プロジェクト」『東工大留学生センター年報』5: 43-45.

橋本直幸, 山内博之 (2008), 「日本語教育のための語彙リストの作成」『日本語学』27/10, 50-58.

前川喜久雄 (2006), 「特定領域研究『日本語コーパス』のめざすもの」(特定領域「日本語コーパス」平成 18 年度全体会議予稿集), pp.1-8. (http://www2.ninjal.ac.jp/kikuo/tokutei_H18_1.pdf)

前川喜久雄 (2007), 「コーパス日本語学の可能性―大規模均衡コーパスがもたらすもの―」『日本語科学』22: 13-28.

丸山岳彦 (2009), 「日本語コーパスの現状」『国文学解釈と鑑賞』74/1: 122-130.

山内博之 (2008), 「形態素解析に関する提案―日本語教育の視点から―」(特定領域「日本語コーパス」日本語教育班研究成果報告書 JC-E-07-01), pp. 84-93.

山内博之(編) (2008), 『日本語教育スタンダード試案 語彙』ひつじ書房.

Biber, Douglas *et al*. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Kilgarriff, Adam (2001), “Comparing corpora,” *International Journal of Corpus Linguistics*. 6/1: 1-37.

関連 URL

「あすなろ」(日本語読解学習支援システム)：http://hinoki.ryu.titech.ac.jp/asunaro/index-j.php

「茶漉」(日本語用例・コロケーション抽出システム)：http://tell.fll.purdue.edu/chakoshi-wiki/

「リーディング チュウ太」(日本語読解学習支援システム)：http://language.tiu.ac.jp/

ActivePerl ダウンロード(ActiveState 社)：http://www.activestate.com/activeperl/downloads

SQLite (関係データベース)：http://www.sqlite.org/

UniDic (形態素解析辞書)：http://www.tokuteicorpus.jp/dist/

本稿で構築した語彙情報分析システムの公開情報

日本語教育班ホームページを参照されたい。URL: http://www.tokuteicorpus.jp/g_teaching/

中学校・高校教科書の教科特徴語リストの作成
―語彙指導の基礎資料として―

近藤 明日子（言語政策班連携研究者：国立国語研究所コーパス開発センター）†

Lists of the Subject-Specific Vocabulary in the Textbooks for Japanese Junior High Schools and Senior High Schools: The Basic Material for Teaching Vocabulary

KONDO Asuko (National Institute for Japanese Language and Linguistics)

１．はじめに

中央教育審議会（2008）に「各教科等における言語活動の充実は、今回の学習指導要領の改訂において各教科等を貫く重要な改善の視点である」とあるように、2011 年度より順次実施される新学習指導要領では、国語科だけでなく各教科での「言語活動の充実」が強調されており、今後その重要性が一層増すことが予想される。

各教科での言語活動の充実のためには、まず各教科で必要とされる言語のありようを特定する必要があるが、その一つの方策として、各教科の教科特徴語のリスト化が考えられる。教科特徴語とは、各教科において特徴的に高頻度に使用される語彙のことであり、そのリスト化は、各教科で特に必要とされる語彙の特定、ひいては重点的な指導が望まれる語彙の特定に他ならない。本稿筆者はこれまで、コーパスを利用した教科特徴語の抽出を試行してきたが（近藤、2008a；2008b；2009）、それを踏まえ、今回、中学校・高校教科書の教科特徴語リストを完成させ公開するに至った。本稿では、そのリストの作成方法とリストの構成の概要について報告する。

２．教科特徴語リストの作成方針

教科書特徴語リストの作成にあたっては、語彙指導の基本的な資料となり、かつ語彙指導の現場で実用性のあるものとするため、以下の方針に基づくこととする。

２．１．リスト作成対象の校種

中学校や高校のある教科の教科書といった、特定分野の特徴語を抽出する方法として、その特定分野のコーパス（対象コーパス）とその比較対象とするコーパス（参照コーパス）を用意し、参照コーパスよりも対象コーパスで偏って高頻度に出現する語を特徴語として抽出する方法がある。本稿もこの方法に従って各教科の特徴語の抽出を行う。

本稿で教科特徴語とするものは、「児童・生徒が日常生活で触れる書き言葉と比較して、当該教科の教科書で特に偏って高頻度に出現する語」である。これを抽出するためには、対象コーパスとして児童・生徒が学校で使用する教科書の言葉を代表するコーパスを、参照コーパスとして児童・生徒が日常生活で触れる書き言葉を代表するコーパスを用意する必要がある。

そこで、まず対象コーパスとして、特定領域研究「日本語コーパス」言語政策班が構築

† kondo@ninjal.ac.jp

した「教科書コーパス」を利用する。「教科書コーパス」は 2005 年度に小学校・中学校・高等学校で用いられた検定教科書（各学年・各教科 1 種ずつ）を対象とした全文コーパスであり、児童・生徒が学校で使用する教科書の言葉を代表するものとして設計されたコーパスである[1]。この「教科書コーパス」を校種・教科別に分割し、それぞれを対象コーパスとすることができる。

一方、参照コーパスとして利用できるものに、『現代日本語書き言葉均衡コーパス』（BCCWJ）[2]の図書館サブコーパスがある。これは、公共図書館の収蔵図書を母集団とみなしたコーパスで、ある程度広い範囲に流通したことが確認されている書き言葉を代表するものとして設計されているものである。これを「児童・生徒が日常生活で触れる書き言葉」に相当するものと想定し、参照コーパスとすることが考えられる。ただし、一口に「児童・生徒」とは言っても年齢の幅は広い。年齢の高い高校生ならば、その日常生活は社会人のそれに近く、BCCWJ 図書館サブコーパスを「日常生活で触れる書き言葉」の代表とすることに妥当性もあろう。しかし、低年齢になればなるほど、「日常生活で触れる書き言葉」と BCCWJ 図書館サブコーパスとの間には乖離が生じ、小学生ともなれば BCCWJ 図書館サブコーパスを「日常生活で触れる書き言葉」の代表とすることにはかなり無理がある。よって、本稿では BCCWJ 図書館サブコーパスを「日常生活で触れる書き言葉」の代表として想定できるのは、高校生および中学生までとし、BCCWJ 図書館サブコーパスを参照コーパスとする方法での教科特徴語の抽出もまた高校生および中学生の使用する教科書からのみ行うこととする。

２．２．教科特徴語抽出の対象とする教科書の文書要素

対象コーパスすなわち中学校・高校の各教科の教科書の内部は、その機能から「主要学習部分」「補助」「発展」「図表」「注」「引用」の文書要素に分けることができる（近藤、2010）。これらの文書要素のうち、教科書の主幹となるのは、各単元で学習する主な内容を説明する機能を持つ「主要学習部分」要素である。基本的なリスト作成のために、この「主要学習部分」要素に出現する語彙から教科特徴語を抽出する。よって、対象コーパスは中学校・高校の各教科の教科書全体ではなく、「主要学習部分」要素の部分とする。

２．３．語の単位

リスト作成に際し、コーパスの形態素解析は形態素解析辞書 UniDic[3]を使って行う。短単位[4]を解析単位とする UniDic を用いた解析では、例えば「全自動洗濯機」のような合成名詞は「全／自動／洗濯／機」のように短い単位に分割される。本稿で抽出する教科特徴語は、語彙指導での利用を目的としている面からも、また合成名詞の多い専門用語を多く含むという面からも、合成名詞は切らずに一つの単位としたままのほうがよいと考える。そこで、UniDic による解析結果から、一定の条件をみたす短単位連続を合成名詞に近似するものと見なし、1 単位として再構成し、語の単位として利用する。

２．４．収録語の選定

教科特徴語は、対象コーパスに出現する各語について、対象コーパスでの度数と参照コ

[1] 「教科書コーパス」の詳細については、田中・近藤・平山（2011）を参照のこと。
[2] http://www.ninjal.ac.jp/kotonoha/
[3] http://download.unidic.org/
[4] 短単位の規定の詳細は小椋・小磯・冨士池・宮内・原（2010）を参照のこと。

一パスでの度数から算出される特徴度（後述）の値に基づき抽出する。しかし、特徴度がある水準より大きい語すべてを教科特徴語として機械的にリストに収録するにはいろいろ問題がある。例えば、形態素解析や単位再構成の誤りによって実際にはコーパスに出現しない語が教科特徴語として抽出される場合がある。このようなものは手作業での修正が必要である。また、教科特徴語として抽出された語のなかには、教科書の特定の題材に由来するゆえに、当該教科において一般性の低いものが混在する。こうした語は基本的なリストでの重要度は低いと考え、本稿で作成するリストからは除外したい。そのためには、教科での一般性が低いか否かを人によって判断する必要がある。

そのような人手によるリストの整備を経て得られる教科特徴語リストでも、その掲載語数が多すぎては実用的なリストとは言えないであろう。よって、校種・教科ごとに作成するリストに掲載する語数に上限を設ける。

３．リストの作成方法

ここでは、２で述べた方針によるリスト作成の具体的な方法について説明する。

３．１．使用コーパス

まず、対象コーパスとして利用する「教科書コーパス」[5]から中学校・高校教科書部分を取り出し、それぞれを「教科書コーパス」での教科分類に従い、「国語・数学・理科・社会・外国語・技術家庭・芸術・保健体育・情報」の 9 種（中学は情報がないので 8 種）のコーパスに分割する。この計 17 種のコーパスから、citation・figureBlock・noteBody・skippedSpan・skippedBlock・supplement の各要素を除外して得られる「主要学習部分」要素のみを対象コーパスとして用意する。

次に、参照コーパスとして、BCCWJ 図書館サブコーパスに収録予定の固定長サンプル（LB_FL）、計 10,640 サンプル[6]を用意する。

３．２．形態素解析と同語異語判別

17 種の対象コーパスと 1 種の参照コーパスは、形態素解析辞書 UniDic（MeCab 版）の最新版（非公開）を用いて形態素解析する[7]。ただし、合成名詞は切らずに一つの単位とするために、解析結果から、一定の条件をみたす短単位連続[8]を合成名詞に近似するものと見なし、1 単位として再構成する[9]。

短単位および再構成された単位の同語異語判別は、UniDic によって付与される属性[10]のうち「語彙素読み」「語彙素」「語彙素細分類」「語種」「品詞」「活用型」を用い、これらの値

[5] 2010 年 12 月 9 日版（非公開）を利用する。
[6] 2010 年 12 月 9 日版（非公開）を利用する。
[7] この形態素解析結果は国立国語研究所内の BCCWJ のための形態論情報データベースで管理されている。このデータベースの 2010 年 12 月 9 日時点のデータを利用する。データの利用にあたっては、特定領域研究「日本語コーパス」データ班の協力を得た。
[8] UniDic によって付与される品詞属性値が「名詞-普通名詞」「名詞-固有名詞」「接頭辞」「接尾辞」「形状詞-一般」「形状詞-タリ」のいずれかで始まる短単位が複数連続するものを 1 単位として再構成する。ただし、その短単位連続の先頭に品詞属性値が「接尾辞」で始まる単位が位置する場合と、末尾に品詞属性値が「接頭辞」で始まる単位が位置する場合は、それぞれを短単位連続から切り出し、別の単位と認定する。
[9]合成名詞等を長いまま 1 単位とする言語単位として、BCCWJ で採用されている長単位がある。将来的には、教科特徴語の抽出で長単位による形態素解析結果を利用することを予定している。長単位の規程の詳細は小椋・小磯・冨士池・宮内・原（2010）を参照のこと。
[10] UniDic の付与する属性の詳細について、UniDic 同梱のマニュアルを参照のこと。

がすべて一致するものを同語と見なし、一つの見出し語のもとにまとめる。再構成した単位の場合は、「語彙素読み」「語彙素」属性の値には、構成前の短単位の属性値を結合したものを用いる[11]。「語彙素細分類」「語種」「活用型」属性の値には、構成前の短単位の値を+で結合したものを用いる。「品詞」属性の値は、構成前の短単位が「接頭辞」のみからなるものは「接頭辞」、「接尾辞」のみからなるものは「接尾辞」、最後尾の短単位が活用型「形容詞」の接尾辞からなるものは「形容詞」、それ以外は「合成名詞」とする。

このようにして得られた対象コーパス・参照コーパスの語彙のうち、UniDic の付与する品詞属性の値が「名詞・代名詞・形状詞・連体詞・副詞・接続詞・感動詞・動詞・形容詞・接頭辞・接尾辞」のいずれかで始まる見出し語と、短単位連続から再構成した単位からなる見出し語を対象に教科特徴語の抽出を行う（助詞・助動詞・記号類は対象外とする）。これらの見出し語の異なり語数・延べ語数を示すと表 1 のようになる。

表 1 対象コーパス・参照コーパスの異なり語数・延べ語数

			異なり語数	延べ語数
対象コーパス	中学校	国語	6,113	30,806
		数学	970	13,012
		理科	2,463	17,703
		社会	7,607	38,746
		外国語	975	3,135
		技術家庭	2,650	12,685
		芸術	4,025	17,092
		保健体育	1,668	6,740
	高校	国語	6,716	33,683
		数学	2,430	46,181
		理科	15,408	192,616
		社会	35,787	293,080
		外国語	1,398	3,674
		技術家庭	7,202	42,766
		芸術	10,648	54,976
		保健体育	3,755	19,123
		情報	3,855	28,295
参照コーパス			256,186	3,452,846

３．３．特徴度の算出

次に、各対象コーパスの語彙について、対象コーパスでの度数と参照コーパスでの度数を比較し、対象コーパスに偏って高頻度に出現する程度（以下、「特徴度」）を数値化する。特徴度の指標とする統計値として、対数尤度比（log-likelehood ratio、G^2）を用いる。対数尤度比は、英語学において特徴語の抽出のための指標として一定の評価を得ているものである（石川、2008、p.99）。

対象コーパスに出現する語 W の対数尤度比は、次の式[1]によって求めることができる（Kilgarriff、2001）。

$$G^2=2(alna+blnb+clnc+dlnd-(a+b)ln(a+b)-(a+c)ln(a+c)-(b+d)ln(b+d)-(c+d)ln(c+d)+(a+b+c+d)ln(a+b+c+d)) \quad \cdots[1]$$ [12]

a：対象コーパスでの語 W の度数
b：参照コーパスでの語 W の度数
c：対象コーパスの延べ語数－a
d：参照コーパスの延べ語数－b

[11] 「語彙素読み」の値については、合成名詞となることで連濁等の語形変化が起きるものは、短単位の「語彙素読み」を単純に結合した値を用いるのではなく、合成名詞として一般的なものに修正したものを用いる。例えば、「カザン（火山）」と「ハイ（灰）」で構成される合成名詞の「語彙素読み」は「カザンハイ」ではなく「カザンバイ」に修正した値を用いる。また、「語彙素」の値については、再構成前の短単位では人名・地名等の固有名詞の多くが片仮名表記となるが、合成名詞の「語彙素」では文脈での表記に修正した値を用いる。例えば、「シュシ」と「学（ガク）」で構成される合成名詞の「語彙素」は「シュシ学」ではなく「朱子学」に修正した値を用いる。

[12] ln は自然対数を表す。また、a または b が 0 の場合、alna または blnb が 0 と見なし対数尤度比を算出する（高見、2003）。

さらに、単語 W の対象コーパスでの使用率が参照コーパスでの使用率より低い場合 (ad-bc<0 の場合)、対数尤度比に-1 を乗じる補正（内山・中條・山本・井佐原､2004)を行った値を単語Wの特徴度とする。特徴度は、単語 W が参照コーパスに比べて対象コーパスでより高頻度に出現する場合、正の値をとり、高頻度に出現する偏りの程度が大きいほど大きい値をとる。

表 2 リスト収録候補の見出し語の異なり語数･延べ語数

		異なり語数	延べ語数
中学校	国語	739	13,957
	数学	480	10,023
	理科	718	10,472
	社会	1,357	19,474
	外国語	291	1,552
	技術家庭	784	7,231
	芸術	757	7,481
	保健体育	444	3,754
高校	国語	1,085	14,998
	数学	1,013	35,606
	理科	5,178	138,826
	社会	5,209	172,077
	外国語	301	1,125
	技術家庭	1,359	24,437
	芸術	1,831	26,736
	保健体育	686	10,173
	情報	1,064	18,830

３．４．リスト収録語の選定

まず、特徴度が 10.83 より大きい語（p<.001）を有意に偏って高頻度に出現する語と見なし[13]、各校種・教科の教科特徴語リストの収録候補とする。候補となった見出し語の異なり語数・延べ語数を表 2 に示す。

次に、このリスト収録候補の見出し語を一覧し、解析の誤りを含むことが疑われるものについて、出現した文脈を確認する。その結果、1 つの見出し語に所属するすべての用例が同一の見出し語に修正されることが確認された場合、修正後の見出し語をリスト収録候補に追加し、修正前の見出し語は削除する。例えば、見出し語「【常傭漢字表】(合成名詞)[14]」に所属する用例はすべて正しくは「【常用漢字表】(合成名詞)」に所属するべきものであった。この場合、リスト収録候補から「【常傭漢字表】(合成名詞)」を削除し、代わりに「【常用漢字表】(合成名詞)」を追加する。しかし、1 つの見出し語に所属する用例が複数の見出し語に分かれて修正される場合は、修正後の見出し語をリスト収録候補に追加することはせず、修正前の見出し語の削除のみおこなう。これは、修正後の見出し語の特徴度が 10.83 より大きいかどうか推定が困難なためである。例えば、「【はあ】(感動詞-一般)」に所属する用例は、正しくは「【ハ】(記号-一般)」「【は】(助詞-係助詞)」のいずれかに所属するべきものであった。この場合、リスト収録候補から「【はあ】(感動詞-一般)」の削除のみ行い、「【ハ】(記号-一般)」「【は】(助詞-係助詞)」の追加は行わない。そして、このリスト修正作業で新たに追加された見出し語が既存の見出し語と重複する場合は、特徴度がもっとも高い見出し語を残し、その他のものは削除する[15]。

この修正作業を経たリスト収録候補から、次の①～⑤の条件にあてはまる見出し語を削除する。

① 上述の修正作業の結果、品詞が助詞・助動詞・記号類等のリスト対象外のものとなるもの
② 数詞
③ 人名・地名
　例：【ボブ】、【横田選手】、【北海道】、【アフリカ大陸】

[13] 特徴度の有意水準とその臨界値は高見（2003）を参照した。
[14] 以下、見出し語は【　】内に語彙素、(　) 内に品詞を示して表記する。
[15] この際、度数を合計する等の調整は行わなかった。

④　合成名詞に近似するものとして再構成された単位のうち、語とは見なせないもの
例：【其々平行】、【全て合同】、【殆ど自覚症状】

⑤　引用文や例文の題材に由来するなど、当該教科での一般性が低いと考えられるもの
例：【金魚】、【菊人形】、【洗面台】（以上、中学の国語の特徴語の収録候補）、【大仏】、【格闘ゲーム】、【ハンバーガーショップ】（以上、中学の英語の特徴語の収録候補）

４．リストの収録語数と教科書の語彙カバー率

以上の整備を終えたリスト収録候補の見出し語数が 500 語を超える校種・教科については、特徴度降順上位 500 位までの語をリストに収録する。この結果、校種・教科別の教科特徴語リストに収録される見出し語の異なり語数・延べ語数を表 3 に示す。また、表 1 に示した校種・教科別の対象コーパス全体での異なり語数・延べ語数に対する、表 3 の割合（語彙カバー率）を示したものが表 4 である。

表 3　教科特徴語リストの見出し語の異なり語数・延べ語数

		異なり語数	延べ語数
中学校	国語	500	11,351
	数学	381	7,741
	理科	500	8,502
	社会	500	13,443
	外国語	109	750
	技術家庭	500	6,804
	芸術	506	6,188
	保健体育	434	3,584
高校	国語	511	9,824
	数学	500	26,202
	理科	500	78,289
	社会	500	86,499
	外国語	29	168
	技術家庭	531	18,565
	芸術	500	17,456
	保健体育	500	8,977
	情報	503	12,973

表 4　教科書全体の語数に対する教科特徴語リストの見出し語の異なり語数・延べ語数の割合（語彙カバー率）

		異なり語数	延べ語数
中学校	国語	8.2%	36.8%
	数学	39.3%	59.5%
	理科	20.3%	48.0%
	社会	6.6%	34.7%
	外国語	11.2%	23.9%
	技術家庭	18.9%	53.6%
	芸術	12.6%	36.2%
	保健体育	26.0%	53.2%
高校	国語	7.6%	29.2%
	数学	20.6%	56.7%
	理科	3.2%	40.6%
	社会	1.4%	29.5%
	外国語	2.1%	4.6%
	技術家庭	7.4%	43.4%
	芸術	4.7%	31.8%
	保健体育	13.3%	46.9%
	情報	13.0%	45.8%

表 4 からわかるように、語彙カバー率は校種・教科により差がある。リスト収録語数に上限を設けたことにより、対象コーパス全体での語数が多い校種・教科については語彙カバー率が低くなっている。また、外国語ではリスト収録候補の語数が少なかったことに加え、収録候補から主に前述の①⑤の条件により多くの語が削除されたため、リスト収録語数そのものが少なくなり、結果、語彙カバー率も低くなっている。

５．リストの構成

教科特徴語リストは校種・教科ごとに計 17 種作成する。ファイル形式は Microsoft 社の表計算ソフト Excel の Excel 97-Excel 2003 ブック形式（.xls）とする。ファイルは、田中・相澤ほか（2011）の付録 CD-ROM や特定領域研究「日本語コーパス」研究成果報告 DVD に

収録し公開する。また、後日インターネット上でも公開する予定である。表 5 に例として中学校の数学の教科特徴語リストの一部を示す。

表 5 中学校の数学の教科特徴語リスト（一部）

ID	語彙素読み	語彙素	語種	品詞	活用型	説明	度数	特徴度	度数_LB_FL	レベル_LB_FL
1	アタイ	値	和	名詞-普通名詞-サ変可能			75	507.06	217	Ⅱ
2	アツマリ	集まり	和	名詞-普通名詞-一般			7	34.43	57	Ⅲ
3	アテハマル	当て嵌まる	和	動詞-一般	五段-ラ行-一般		7	33.15	63	Ⅲ
4	アラワス	表わす	和	動詞-一般	五段-サ行		107	561.85	729	Ⅰ
5	アラワス	表わす	和	動詞-一般	下一段-サ行	「表せる」	5	32.92	16	Ⅳ
6	アル	或る	和	連体詞			19	14.69	1796	Ⅰ
7	イコウ	移項	漢	名詞-普通名詞-サ変可能			5	55.85	0	
8	イチ	位置	漢	名詞-普通名詞-サ変可能			20	49.09	638	Ⅰ
9	イッテイ	一定	漢	名詞-普通名詞-サ変形状詞可能			13	41.08	278	Ⅱ
10	イッパン	一般	漢	名詞-普通名詞-一般			18	50.28	473	Ⅰ
11	イドウ	移動	漢	名詞-普通名詞-サ変可能			12	35.45	288	Ⅱ
12	イフゴウ	異符号	漢+漢	合成名詞			3	33.51	0	
13	イレカエル	入れ替える	和	動詞-一般	下一段-ア行		4	20.04	31	Ⅳ
14	イロイロ	色々	和	形状詞-一般			36	132.59	577	Ⅰ
15	インスウ	因数	漢	名詞-普通名詞-一般			5	55.85	0	
16	インスウブン	因数分解	漢+漢	合成名詞			10	100.91	2	Ⅴ
17	ウエ	上	和	名詞-普通名詞-副詞可能			73	120.33	3783	Ⅰ
18	ウヘン	右辺	漢	名詞-普通名詞-一般			4	21.6	25	Ⅳ
19	ウラガエス	裏返す	和	動詞-一般	五段-サ行		3	12.94	34	Ⅲ
20	エイカク	鋭角	漢	名詞-普通名詞-一般			3	22.96	5	Ⅴ

すべてのリストが表 5 同様の構成をとる。リストの各列の詳細については、ファイルに添付する説明文書を参照してほしいが、最右列の「レベル_LB_FL」についてここで説明する。「レベル_LB_FL」は LB_FL に出現する見出し語について、その度数降順の累積使用率により Ⅰ～Ⅴの 5 段階に分けたものである[16]。レベルの設定基準と各レベルに所属する見出し語の LB_FL における度数および異なり語数を表 6 に示す。

表 6 LB_FL におけるレベルの設定

レベル	累積使用率	度数	異なり語数
Ⅰ	0 ～ 60%	158,650 ～ 348	1,108
Ⅱ	～ 70%	347 ～ 126	1,689
Ⅲ	～ 80%	125 ～ 33	5,778
Ⅳ	～ 90%	32 ～ 6	28,236
Ⅴ	～ 100%	5 ～ 1	219,375

レベルⅠに所属する見出し語は LB_FL での度数が高く、レベルⅤになるほど度数が低くなる。つまり、レベルは「生徒が日常生活で触れる書き言葉において、頻繁に使われる程度」、言い換えれば「生徒の日常生活での馴染み度」を示す指標ともなるものである[17]。この「馴染み度」の情報は、例えば、生徒にとって馴染みのない低レベルの語を優先的に指導する等、語彙指導での重要度を判断する際に利用することもできるであろう。

６．おわりに

以上、中学校・高校教科書から校種・教科別に教科特徴語を抽出し、整備した上で教科特徴語リストを作成した。コーパスという大規模な言語資源を活用し、客観的な指標に基

[16] 累積使用率による語彙のレベル分けについては田中（2011）を参照した。

[17] なお、「レベル_LB_FL」が空値の見出し語は、LB_FL での度数が 0 の語である。これはレベルⅤの語よりもさらに「馴染み度」の低い語と言える。

づき抽出された教科特徴語候補に、人手による選別作業を加えることで、基本的かつ実用的なリストが作成できたと考える。しかし、選別方法の妥当性について考えるべき点は少なくなく、リストに収録すべき語が漏れていたり収録するにふさわしくない語が残っていたりすることもあろう。そういった点の改善や、本リストを語彙指導の現場でどのように活用するかといった検討については今後の課題としたい。

文献

石川慎一郎（2008）『英語コーパスと言語教育』　大修館書店

内山将夫、中條清美、山本英子、井佐原均（2004）「英語教育のための分野特徴単語の選定尺度の比較」, 自然言語処理, 11-3, pp.165-197.
（http://www2.nict.go.jp/x/x161/members/mutiyama/pdf/chara.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、原裕（2010）『特定領域研究「日本語コーパス」平成 21 年度研究成果報告書　『現代日本語書き言葉均衡コーパス』形態論情報規定集　第 3 版』.

近藤明日子（2008a）「中学校教科書の教科別特徴語の抽出　―理科を例として―」, 特定領域研究「日本語コーパス」平成 19 年度公開ワークショップ（研究成果報告会）予稿集, pp.181-186.

近藤明日子（2008b）「中学校教科書の教科別特徴語の抽出」, 特定領域研究「日本語コーパス」言語政策班中間報告書　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.111-114.

近藤明日子（2009）「中学校教科書の教科特徴語の抽出と考察　―『現代日本書き言葉均衡コーパス』の語彙との比較から―」, 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集, pp.117-122.

近藤明日子（2010）「検定教科書の語彙分析　―主要学習部分とその他の部分との比較から―」, 特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集, pp.209-216.

高見敏子（2003）「「高級紙語」と「大衆紙語」の corpus-driven な特定法」, （北海道大学）大学院国際広報メディア研究科・言語文化部紀要, 44, pp.73-105.
（http://www.hucc.hokudai.ac.jp/~p16537/papers/Takami_2003_MLC.pdf よりダウンロード可能）

田中牧郎（2011）「語彙レベルに基づく重要語彙リストの作成―国語政策・国語教育での活用のために―」, 田中、相澤ほか（2011）, pp.77-87.

田中牧郎、近藤明日子、平山允子（2011）「教科書コーパス」, 田中、相澤ほか（2011）, pp.7-54

田中牧郎、相澤正夫、斎藤達哉、棚橋尚子、近藤明日子、小椋秀樹、鈴木一史、河内昭浩、平山允子（2011）『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』（JC-P-10-11）

中央教育審議会（2008）「幼稚園、小学校、中学校、高等学校及び特別支援学校の学習指導要領等の改善について（答申）」
（http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2009/05/12/1216828_1.pdf よりダウンロード可能）

Adam Kilgarriff（2001）"Comparing corpora" *International Journal of Corpus Linguistics*, 6-1, pp.1-37.
（http://www.kilgarriff.co.uk/Publications/2001-K-CompCorpIJCL.pdf よりダウンロード可能）

ジャンル別に見た特徴漢字
—書籍のジャンルと広報紙の漢字—

斎藤　達哉（言語政策班分担者：専修大学文学部）†

Difference of the Distribution of Kanji due to Genre and Region

Tatsuya Saito　(School of Literature, Senshu University)

１．本発表の概要

本発表は，漢字について，それが使われたの文章の《内容》や《流通地域》よって，どのように異なっているのかを調査した結果の一端を報告するものである。

２．使用コーパス

本調査では，次の二つのデータを使用した。

第 1 のデータは，BCCWJ 領域内公開データ（2009 年度版）に収められたデータのうち，生産実態サブコーパス書籍（PB）の固定長サンプル（PB_FL）である。以下で，「書籍データ」又は「書籍」と称するのは，この PB 固定長サンプルのことである。

第 2 のデータは，全国 100 市区町村の広報紙のテキストデータである。市区町村の抽出は，「国語に関する世論調査」の地点抽出方法に倣い，人口比を考慮しつつ全国から 100 市区町村を選んだ。さらに，その市区町村が 2008 年に発行した広報紙を対象にして，広報紙 1 タイトルあたり 6 万文字を満たすだけの号を無作為に選んで入力した（採録した広報紙は表 1（本稿末尾）に示す)。以下では，これを「広報紙サンプル」又は「広報紙」と呼ぶ。

３．データの「ジャンル」と「地域」

《内容》の区分は，第 1 のデータである書籍サンプルのジャンルの別を利用し，ジャンルごとの漢字出現の差を調査した。ここでのジャンルとは，NDC（日本十進分類法）の第 1 次区分「類」を利用した区分である。書籍サンプルの各 XML ファイルの場合，ファイル名の左から 4 桁目に NDC の「類」を利用した分類が示されている（丸山(2009)）。

「0」＝総記　「1」＝哲学　「2」＝歴史　「3」＝社会科学
「4」＝自然科学　「5」＝技術・工学　「6」＝産業　「7」＝芸術・美術
「8」＝言語　「9」＝文学　「n」＝分類なし

表 2 では，書籍について，ジャンル別にサンプル数・漢字字種数・漢字数を示した。

《流通地域》の区分は，第 2 のデータである広報紙サンプルの発行された地方の別を利用し，地方ごとの漢字出現の差を調査した。ここでの地方とは，以下のものである。

「A」＝北海道地方　「B」＝東北地方　「C」＝関東地方
「D」＝中部地方　「E」＝近畿地方　「F」＝中国地方
「G」＝四国地方　「H」＝九州・沖縄地方

表 3 では，広報紙について，地方別にサンプル数・漢字字種数・漢字数を示した。

† tsaito@isc.senshu-u.ac.jp

表 2　書籍（PB）のジャンル別漢字数

	サンプル数	漢字字種数	漢字字数
0　総記	251	2,242	74,530
1　哲学	536	2,889	171,681
2　歴史	682	3,479	269,514
3　社会科学	2,267	3,446	898,053
4　自然科学	615	2,523	217,397
5　技術・工学	618	2,596	202,657
6　産業	334	2,353	119,593
7　芸術・美術	524	2,875	158,780
8　言語	153	2,252	50,833
9　文学	2,244	4,000	648,111
n　分類なし	262	2,084	67,420
ＰＢ合計	8,486	4,815	2,878,569

表 3　広報紙の地方別漢字数

	サンプル数	漢字字種数	漢字字数
A　北海道地方	21	1,844	136,800
B　東北地方	24	2,280	250,233
C　関東地方	116	2,671	957,379
D　中部地方	62	2,601	518,936
E　近畿地方	73	2,532	539,419
F　中国地方	17	2,228	194,237
G　四国地方	14	1,954	114,096
H　九州・沖縄地方	28	2,338	285,902
広報紙合計	355	3,306	2,997,002

４．ジャンルごとの特徴漢字

書籍サンプル中の漢字の出現状況に基づき，ジャンルごとの特徴漢字について考える。

４.１　サンプルカバー率で比較する

斎藤(2011)では，書籍のジャンルごとの差を検討した結果，［サンプルカバー率］を比べることを提案した。

単純に頻度数の高さを比べた場合，話題によって，1 サンプル中に同じ漢字が大量に出現することがある。例えば，雛人形について解説したサンプル中では漢字「雛」が大量に出現するということがあった。そこで，サンプルの内容に左右されにくい集計方法として，ある漢字が何%のサンプルに出現したかを示す［サンプルカバー率］を書籍全体と各ジャンルとで算出し，比較することにした。

４.２　書籍全体でのサンプルカバー率の高い漢字（上位 100）

次の漢字は，書籍全体で見たときにサンプルカバー率が上位のものである。

一(87.9%)　人(81.3%)　大(77.4%)　出(74.5%)　分(74.0%)　上(71.7%)　中(71.4%)
見(68.7%)　生(67.6%)　行(66.3%)　間(64.3%)　自(63.7%)　本(62.2%)　時(62.2%)
的(61.8%)　合(61.3%)　事(60.9%)　手(59.3%)　日(59.2%)　思(58.6%)　前(57.6%)
方(57.5%)　場(57.3%)　年(57.2%)　言(56.0%)　後(55.4%)　者(55.1%)　気(53.5%)
二(53.0%)　入(52.4%)　目(51.8%)　体(51.8%)　実(51.2%)　同(50.6%)　立(50.5%)
理(49.3%)　当(48.2%)　部(47.6%)　会(47.0%)　動(47.0%)　下(46.9%)　度(46.5%)
対(45.9%)　子(45.8%)　要(45.7%)　用(45.4%)　力(45.1%)　最(44.5%)　心(44.3%)
意(43.7%)　何(43.6%)　明(43.5%)　持(43.2%)　通(42.5%)　多(42.2%)　家(42.1%)
関(42.0%)　三(41.8%)　取(41.8%)　現(41.4%)　全(41.3%)　考(41.3%)　成(41.2%)
物(41.1%)　来(41.0%)　内(40.8%)　定(40.4%)　知(40.3%)　高(40.2%)　以(39.9%)
長(39.8%)　国(39.8%)　性(39.8%)　発(39.3%)　作(38.6%)　地(38.2%)　十(38.2%)
話(37.9%)　今(36.9%)　代(36.8%)　化(36.2%)　私(36.1%)　不(36.1%)　外(36.0%)
小(35.7%)　先(35.6%)　所(34.8%)　必(34.4%)　面(34.3%)　少(34.3%)　向(34.3%)
変(34.3%)　身(34.1%)　学(34.0%)　主(33.8%)　表(33.5%)　法(32.8%)　月(32.7%)
書(32.6%)　重(32.3%)

４.３　ジャンルごとのサンプルカバー率の高い漢字（サンプルカバー率差 10.0%以上）

ある漢字について，あるジャンルにおけるサンプルカバー率からサンプル全体におけるサンプルカバー率を引くと，［サンプルカバー率差］が算出できる（詳しくは斎藤(2011)参照)。この［サンプルカバー率差］が，あるジャンルにおいて高い場合，そのジャンルにおいて全体においてよりも広い範囲に分布していることになる。［サンプルカバー率差］の高い漢字は，そのジャンルでの「特徴漢字」と考えられるのではないだろうか。

以下，各ジャンルの［サンプルカバー率差］が 10.0%以上の漢字を上位から順に示す。

0　総記

作 使 用 数 書 示 設 利 報 選 文 初 字 読 違

1　哲学

世 感 神 心 教 自 意 考 聖 在 身 言 想 的 思 仏 界 性 人 命 愛 何 私 生 語
相 観 霊

2　歴史

五 三 国 九 地 六 二 七 八 四 日 代 年 山 東 西 月 戦 北 政 史 田 世 民 名
時 川 近 後 十 和 長 新 土 道 軍 天 朝 治 百 建 平 家 本 大 中 町 前 文 島
南 府 都 古 住 所 帰 在 野 木 城 海 市 勢 小 周 台 記 紀 公 武 諸

3　社会科学

業 社 定 的 対 要 基 関 制 会 法 資 経 務 係 保 支 問 規 権 題 成 民 金 第
点 利 決 者 等 価 政 済 義 方 主 期 産 必 提 例 員 有 用 年 意 以 条 設 発
限 活 認 企 現 当 営 得 理 特 担 任 実

4　自然科学

療 病 患 症 医 性 状 血 質 化 量 科 多 薬 体 診 発 効 学 治 果 因 害 経 酸
液 態 胞 感 障 重 検 類 健 活 常 的 剤 臓 細 疾 種 法 期 増 素 起 康 院 脳

5　技術・工学

使 用 製 設 料 加 量 工 品 材 入 図 作 造 切 技 定 電 器 構

6　産業

本 見 多 高 人 手 的 以 代 発 方 大 話 部 間 家 仕 気 付 状 来 増 題 全 現
先 決 然 面 所 続 前 直 国 水 意 特 主 数 備 相 結 下 初 近 少 可 自 重 会
無 理 時 考 点 供 進 原 総 今 向 指 上 際 持 細 食 規 解 明 場 独 最 過 構
不 起 知 予 例 設 次 返 温 環 楽 減 接 与 保 管 積 毎 帰 義

7　芸術・美術

楽 感 演 選 曲 作 戦 打 勝 描

8　言語

語 文 意 味 表 詞 使 言 書 名 英 字 代 象 音 方 単 形 異 際 説 考 例 点 違
読 学 明 古 葉 現 本 感 句 辞 記 様 世 的

9　文学

声 彼 顔 話 僕 屋 気 男 女 母 言 死 笑 思 十 郎 何 見 口 俺 夜 手 私 二 美
兵 子 父 軍 殺 首 家 太 葉 返 息 真 誰 王 飛 歩 食 隊 着 石 警 聞 胸 頭

n　分類なし

使 選 入 画 色

いずれも，現行の「常用漢字表」（2010 年 11 月 30 日・内閣告示）における常用漢字である（「9　文学」で枠囲いした俺・誰は 2010 年の改定で新たに加わった）。各ジャンルの特徴漢字の傾向について，正確な分析はその漢字を使用した語[1]を示した上でないとできないが，以下ではおおよその傾向について触れておくことにする。

「0　総記」「n　分類なし」は，内容が多岐にわたるジャンルといえるが，共通して「使」が入っている。

「1　哲学」では，「感」「心」「意」「想」「思」などの心的事項に関連すると思われる漢字，「神」「聖」「仏」などの宗教に関連すると思われる漢字が見られる。

「2　歴史」では，「五」「三」「六」「二」「七」「八」「四」のように数詞を構成すると思われる漢字が目立っている。

「3　社会科学」の特徴漢字は，経済，政治，法律などの分野で使用が高くなるもので，社会人となって生活する上で常用中の常用とも位置付けられよう。

「4　自然科学」では，「療」「病」「患」「症」「医」などに始まり，医療用語と関連すると思われる漢字が目立っている。

「5　技術・工学」では，「使」「用」「製」「設」「加」「作」「作」などの動作を表わす語と関連すると思われる漢字が目立つ（「使」は「0　総記」「n　分類なし」にも見られる）。

「6　産業」は漢字の特徴がつかみにくい。

「7　芸術・美術」では，「楽」「演」「曲」などの音楽と関連すると思われる漢字が目立っている。

「8　言語」は「語」「文」「意」「詞」などが見られる。「9　文学」では「彼」「僕」「俺」「私」「誰」「男」「女」「母」「子」「父」などの人を呼び表わすのに関連する漢字が目立つ。

以上のように，各ジャンルの特徴漢字は，字種の多少の重なりはあるものの，おおよその傾向として差が見られることが分かる。

５．地方ごとの特徴漢字

広報紙サンプル中の漢字の出現状況に基づき，ジャンルごとの特徴漢字について考える。

５．１　全国的に使用された漢字（全 355 サンプルで集計した上位 100 位まで）

次の漢字は，広報紙全体で見たときに，サンプルカバー率が上位のものである。

ア　全 355 サンプルで出現した漢字　50 字（サンプルカバー率 100%）

員 加 会 開 学 活 間 関 気 健 行 合 参 子 市 施 事 者 手 出 所 上 場 心 人
図 成 生 全 体 対 大 知 地 中 通 動 内 日 入 年 費 分 平 保 民 役 用 要 理

イ　354 サンプルで出現した漢字　32 字（サンプルカバー率 99.7%）

域 育 一 化 館 金 月 後 交 口 康 高 催 在 支 時 自 実 受 住 書 水 設 前 定
付 物 方 本 問 予 料

ウ　353 サンプルで出現した漢字　17 字（サンプルカバー率 99.4%）

以 回 各 期 記 議 業 区 見 護 公 込 小 新 申 同 話

エ　352 サンプルで出現した漢字　20 字（サンプルカバー率 99.4%）

援 下 楽 校 今 歳 持 集 象 進 第 度 土 当 道 必 報 務 無 連

[1] これについては，ポスターで紹介する。

５．２ 全国に比べ，地方で使用が高くなる漢字（サンプルカバー率差 30%以上）

次の漢字は，各地方においてサンプルカバー率が高くなっている漢字である。

A　北海道地方

幌 札 穂 靴 珠 雪 鉄 丁 曙

B　東北地方

沼 齋 走 仙 沢 披 昔 郷 陸 五 鈴 露 咲 距 泉 踊 著 卓 兄 牛 暗 七 菜 旭 煮 客 里 香 武 雄 姿 眺 尽 離 巻 八 弟 競 荘 菊 弱 祖 桜 阿 荒

C　関東地方

なし（最高で 15.8%の「塚」「往」）

D　中部地方

なし（最高で 20.9%の「走」）

E　近畿地方

阪 府

F　中国地方

債 伯 祥 還 彫 償 扶 契 頃 御 宝 賢 妻 耕 浅 刻 玲 庄 宏 佳 易 製 括 踏 宛 郷 蔵 洋 勉 冨 眺 賃 掛 軟 牛 尾 簡 慶 簿 吉 景 諸 陀 江 也 去 毅 遺 佐 哲 譲 排 慎 倍

G　四国地方

俗 礼 句 俳 片 封 亡 油 澄 炎 漁 抗 之 雲 徳 寿 短 禁 茂 抑 四 両 馬 桑 戻 窪 涼 副

H　九州・沖縄地方

州 九 宣 派 港 遣 夢 突 候 岳 城 逆 祈 牟 畑 挙 沿 鹿 帰 寛 肥 漁

広報紙では，固有名詞（地名等）を除外せずに，集計を行っている。そのため，上記には，「幌」「札」（北海道地方），「仙」（東北地方），「阪」「府」（近畿地方），「州」「九」（九州・沖縄地方）などが入っている。その一方で，地名に用いられない漢字も入っている。

後者の漢字が，実際にどのような使で使用されているかを検討する必要がある。[2]

６．ポスター発表に当たって

４.３では，文章の《内容》によってどのような特徴漢字が見られるかを，書籍のジャンルを頼りにして抽出することを試みた。

５.２では，地域向けの情報媒体である市区町村の広報紙を頼りにして，《流通地域》によってどのような特徴漢字が見られるかを，地方別に抽出することを試みた。

ポスター発表では，それらの漢字によってどのような語が表記されているのかも含めて紹介し，考察を行う。

[2] これについては，ポスターで紹介する。

表1　対象とした広報紙

	広報紙タイトル（2008年発行）	採録巻次（発行年月）	都道府県	発行市区町村
1	広報さっぽろ（ひがし区民のページ）	574(1月10日)，575(2月10日)，576(3月10日)，577(4月10日)，578(4月10日)，579(6月10日)，(7月10日)，581(8月10日)	北海道	札幌市／東区
2	広報さっぽろ（ていね区民のページ）	577(4月10日)，578(4月10日)，579(6月10日)，580(7月10日)，581(8月10日)，582(9月10日)，583(10月10日)	北海道	札幌市／手稲区
3	広報とまこまい	1660(3月1日)，1661(4月1日)	北海道	苫小牧市
4	広報のぼりべつ	690(4月1日)，691(4月1日)	北海道	登別市
5	広報おとふけ	880(4月25日)，881(4月25日)	北海道	音更町
6	広報ひろさき	55(6月1日)，56(6月15日)，57(7月1日)	青森県	弘前市
7	広報遠野	37(7月1日)，38(8月1日)，39(9月1日)	岩手県	遠野市
8	仙台市政だより（青葉区版）	1642(8月1日)，1643(9月1日)	宮城県	仙台市／青葉区
9	広報わたり	504(9月1日)，505(10月1日)，506(11月1日)，507(12月1日)	宮城県	亘理町
10	広報あきた	1682(10月3日)，1683(10月17日)	秋田県	秋田市
11	市報かみのやま	1272(11月1日)，1274(12月1日)，1273(11月15日)，1275(12月15日)	山形県	上山市
12	広報こおりやま	573(1月1日)，584(12月1日)	福島県	郡山市
13	広報あいづみさと	26(1月1日)，27(2月1日)，28(3月1日)，29(4月1日)	福島県	会津美里町
14	広報つちうら	972(2月1日)，974(3月1日)，976(4月1日)，973(2月16日)，975(3月16日)	茨城県	土浦市
15	広報とりで	1000(3月1日)，1002(4月1日)，1004(4月1日)，1006(6月1日)	茨城県	取手市
16	広報つくばみらい	25(4月17日)，26(4月15日)，27(6月19日)	茨城県	つくばみらい市
17	広報あしかがみ	1369(4月1日)，1370(4月15日)，1371(6月1日)，1372(6月15日)	栃木県	足利市
18	広報なすしおばら	83(6月5日)，84(6月20日)，95(12月5日)	栃木県	那須塩原市
19	広報みぶ	585(2月23日)，587(4月23日)，589(6月23日)，590(7月23日)	栃木県	壬生町
20	広報まえばし	1369(8月1日)，1370(8月15日)，1371(9月1日)	群馬県	前橋市
21	広報おおた	121(9月1日)，124(10月1日)，127(11月1日)，130(12月1日)	群馬県	太田市
22	広報しぶかわ	63(10月1日)，64(10月15日)，65(11月1日)	群馬県	渋川市
23	市報さいたま（浦和区版）	81(1月1日)，82(2月1日)，83(3月1日)，84(4月1日)，85(4月1日)，86(6月1日)，87(7月1日)，88(8月1日)，89(9月1日)，90(10月1日)，91(11月1日)，92(12月1日)，93(1月1日)	埼玉県	さいたま市／浦和区
24	広報ところざわ	1030(1月1日)，1031(2月1日)	埼玉県	所沢市
25	広報あげお	886(1月1日)，887(2月1日)	埼玉県	上尾市
26	広報くりはし	436(2月5日)，437(3月5日)，438(4月5日)	埼玉県	栗橋町
27	ちば市政だより（若葉区版）	192(3月1日)，193(4月1日)	千葉県	千葉市／若葉区
28	広報いちかわ	1277(4月5日)，1278(4月12日)，1279(4月19日)，1280(4月26日)	千葉県	市川市
29	広報きさらづ	649(4月1日)，650(6月1日)，651(7月1日)，652(8月1日)	千葉県	木更津市
30	広報そでがうら	695(6月1日)，697(7月1日)，699(8月1日)	千葉県	袖ヶ浦市
31	広報しんじゅく	1894(7月25日)，1896(8月15日)，1897(8月25日)	東京都	新宿区
32	こうとう区報	1575(8月1日)，1576(8月11日)，1577(8月21日)	東京都	江東区
33	区のお知らせ「せたがや」	1299(9月1日)，1300(9月15日)，1302(10月1日)	東京都	世田谷区
34	広報としま	1392(6月5日)，1409(11月25日)，1410(12月5日)，1412(12月25日)	東京都	豊島区
35	ねりま区報	1485(11月11日)，1486(11月21日)，1487(12月1日)	東京都	練馬区
36	市報むさしの	1833(12月1日)，1834(12月15日)	東京都	武蔵野市
37	広報まちだ	1529(1月1日)，1562(12月1日)，1563(12月11日)，1564(12月21日)	東京都	町田市
38	市報きよせ	1017(2月1日)，1018(2月15日)，1019(3月1日)，1020(3月15日)	東京都	清瀬市
39	広報よこまは（鶴見区版）	128(6月1日)，129(7月1日)	神奈川県	横浜市
40	広報よこはま（磯子区版）	126(4月1日)，127(4月1日)，128(6月1日)，129(7月1日)	神奈川県	横浜市／磯子区
41	広報よこはま（瀬谷区版）	127(4月1日)，128(6月1日)，129(7月1日)，130(8月1日)	神奈川県	横浜市／瀬谷区
42	市政だより（中原区版）	936(1月1日)，938(2月1日)，946(6月1日)，948(7月1日)，950(8月1日)	神奈川県	川崎市／中原区
43	広報かまくら	1103(7月1日)，1104(7月15日)，1105(8月1日)	神奈川県	鎌倉市
44	広報ずし	779(8月1日)，780(9月1日)	神奈川県	逗子市
45	広報あつぎ	1046(9月1日)，1048(10月1日)，1050(11月1日)，1052(12月1日)	神奈川県	厚木市
46	南区役所だより「みなみ風」	18(1月6日)，19(1月20日)，20(2月03日)，21(2月17日)，36(10月5日)，37(10月19日)，38(11月2日)，39(11月16日)	新潟県	新潟市南区
47	広報とおかまち	87(11月10日)，88(11月25日)，89(12月10日)	新潟県	十日町市
48	たかおか「市民と市政」	27(1月1日)，37(11月1日)，38(12月1日)	富山県	高岡市
49	広報わじま	24(2007年12月25日)，25(2月1日)，26(3月03日)	石川県	輪島市
50	市政広報ふくい	1277(1月25日)，1278(2月10日)，1280(3月10日)，1284(4月10日)	福井県	福井市
51	広報こうふ	628(4月1日)，629(6月1日)	山梨県	甲府市
52	広報南アルプス	61(4月1日)，62(4月1日)，63(6月1日)，64(7月1日)，65(8月1日)	山梨県	南アルプス市
53	広報うえだ	51(4月1日)，54(6月16日)，56(7月16日)	長野県	上田市
54	広報みなみみのわ	414(6月1日)，415(7月1日)，416(8月1日)，417(9月1日)，418(10月1日)	長野県	南箕輪村
55	広報たじみすとTajimist	2141(7月1日)，2143(8月1日)，2147(10月1日)，2149(11月1日)	岐阜県	多治見市
56	広報ひだ	54(8月19日)，56(9月17日)，57(10月17日)	岐阜県	飛騨市
57	広報しずおか「静岡気分」	131(9月1日)，132(9月15日)	静岡県	静岡市
58	広報はままつ（浜北区版）	1254(1月20日)，1272(10月20日)，1274(11月20日)	静岡県	浜松市浜北区
59	広報かけがわ	65(1月1日)，84(11月1日)，86(12月1日)	静岡県	掛川市
60	広報いず	55(10月1日)，57(12月1日)	静岡県	伊豆市
61	広報なごや（守山区版）	721(1月1日)，722(2月1日)，723(3月1日)，724(4月1日)，725(4月1日)，726(6月1日)	愛知県	名古屋市／守山区
62	広報とよた	1196(1月1日)，1197(1月15日)，1198(2月1日)，1199(2月15日)	愛知県	豊田市
63	広報とうごう	437(8月1日)，438(9月1日)	愛知県	東郷町
64	広報いせ	30(4月1日)，31(4月1日)，32(6月1日)	三重県	伊勢市
65	広報かめやま	76(4月1日)，78(6月1日)，80(7月1日)，81(7月16日)	三重県	亀山市
66	広報くさつ	975(4月15日)，976(4月1日)，977(4月15日)，987(11月1日)，988(11月15日)	滋賀県	草津市
67	広報あいこうか	73(7月1日)，75(8月1日)，77(9月1日)	滋賀県	甲賀市
68	うきょう	145(1月15日)，146(2月15日)，147(3月15日)，152(8月15日)，153(9月15日)，154(10月15日)，155(11月15日)，156(12月15日)	京都府	京都市右京区
69	宇治市政だより	1540(9月1日)，1541(9月11日)，1542(9月21日)，1543(10月1日)，1544(10月11日)	京都府	宇治市
70	広報京丹波	27(1月15日)，35(9月15日)，36(10月15日)，37(11月15日)，38(12月15日)	京都府	京丹波町

(表 1　続き)

	広報紙タイトル（2008年発行）	採録巻次（発行年月）	都道府県	発行市区町村
71	天王寺区広報紙	139(1月15日)，140(2月15日)，141(3月15日)，142(4月15日)，149(11月15日)，150(12月15日)	大阪府	大阪市天王寺区
72	区民だより「よどがわ」	139(1月15日)，140(2月15日)，141(3月15日)，142(4月15日)，143(4月15日)，144(6月15日)，145(7月15日)	大阪府	大阪市淀川区
73	堺区広報「堺」	76(2月1日)，77(3月1日)，78(4月1日)，79(4月1日)，80(6月1日)，81(7月1日)，82(8月1日)，83(9月1日)，84(10月1日)	大阪府	堺市
74	広報ひらかた	1125(2月1日)	大阪府	枚方市
75	広報紙「もみじだより」	689(1月1日)，690(2月1日)	大阪府	箕面市
76	区民広報紙なだ	(4月1日)	兵庫県	神戸市／灘区
77	広報ひめじ	951(4月)，952(4月)，955(8月)	兵庫県	姫路市
78	広報「町から町へ」	1450(7月1日)，1451(7月15日)，1452(8月1日)	奈良県	天理市
79	広報いこま	648(7月15日)，653(12月15日)	奈良県	生駒市
80	市報わかやま	761(8月1日)，762(9月1日)，763(10月1日)	和歌山県	和歌山市
81	広報田辺	40(9月1日)，41(10月1日)	和歌山県	田辺市
82	広報よなご	43(10月1日)，44(11月1日)，45(12月1日)	鳥取県	米子市
83	市報松江	44(11月1日)，45(12月1日)	島根県	松江市
84	広報やかげ	454(1月7日)，455(2月18日)，465(12月17日)	岡山県	矢掛町
85	広報ひろしま	1345(1月1日)，1347(2月1日)，1349(3月1日)	広島県	広島市／佐伯区
86	広報おのみち	917(2月12日)，918(3月10日)	広島県	尾道市
87	広報ひかり	83(3月10日)，85(4月10日)，87(4月10日)，93(8月10日)	山口県	光市
88	広報いしい	141(1月27日)，142(4月15日)，143(7月15日)，144(9月15日)，145(11月15日)	徳島県	石井町
89	広報たかまつ	1400(11月1日)，1401(11月15日)，1402(12月1日)	香川県	高松市
90	広報今治	81(6月1日)，82(6月15日)，83(7月1日)	愛媛県	今治市
91	広報土佐	487(7月1日)，488(8月1日)，489(9月1日)	高知県	土佐市
92	広報くるめ	1210(8月1日)，1211(8月15日)，1212(9月1日)	福岡県	久留米市
93	広報なかがわ	516(1月1日)，524(9月1日)，525(10月1日)，527(12月1日)	福岡県	那珂川町
94	広報有田	32(10月1日)，33(11月1日)，34(12月1日)	佐賀県	有田町
95	広報ながさき	694(11月1日)，695(12月1日)	長崎県	長崎市
96	市政だより天草	42(1月1日)，44(2月1日)，46(3月1日)，64(12月1日)	熊本県	天草市
97	市報べっぷ	1548(1月1日)，1549(2月1日)，1550(3月1日)	大分県	別府市
98	広報ひゅうが	633(2月1日)，634(3月1日)，635(4月1日)	宮崎県	日向市
99	広報きりしま	48(1月1日)，49(2月1日)，51(3月1日)，53(4月1日)，54(4月1日)	鹿児島市	霧島市
100	名護市広報「市民のひろば」	439(6月1日)，440(7月1日)，441(8月1日)	沖縄県	名護市

文献

丸山岳彦(2009), 「『現代日本語書き言葉均衡コーパス』領域内公開データ(2009 年度版)書誌情報・サンプル情報・著者情報について」『現代日本語書き言葉均衡コーパス』領域内公開データ(2009 年度版)DVD-R.

斎藤達哉(2011),「BCCWJ による「NDC ジャンル別漢字出現頻度表」の分析」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.127-142, 国立国語研究所

本稿では，科学研究費補助金若手研究(B)「公共情報媒体としての広報紙を対象とした表記法の在り方に関する調査研究」（課題番号 21720170）による成果の一部も使用している。

社会科での漢字学習事例検討
—小学校 6 年生「憲」について—

棚橋尚子（言語政策班分担者：奈良教育大学教育学部）†

Learning of Kanji in Social Studies:
The Case of "憲" in the Sixth Grader

Hisako Tanahashi　(Nara University of Education)

１．はじめに

漢字は，戦前戦後一貫して国語科において指導してきた。しかしながら，国語科での指導には実際のところ限界も多い。もっとも大きな点は，国語科の教材配列が学習指導要領に即した内容を有する文学作品や説明的文章を中心に組み立てられていて，漢字および漢字熟語は教材の採用状況によって偶然に提示される点にある。かつて輿水（1971）の示した読解の「基本的指導過程」における漢字学習の位置づけの特徴の一つには，「漢字を，文章を読む前に取り出して教えてしまわない―センテンス・メソッドの原則を守る。」とあるが，[1]国語科教科書において，その立場を貫こうとするとかなり無理な事態が引き起こされる。輿水らの提案後，学校教育は過激な受験戦争，「落ちこぼれ」の増加などの問題から「ゆとり」を重視する方向に傾き，学習指導要領が改訂されるたびに「教育内容の精選」，「厳選」，と指導事項が縮小される方向に向かった。国語について言えば，単元数が少なくなり，それとともに国語科教科書の総語彙数も少なくなった。漢字の場合も，従前は読み替え漢字の音訓もすべて教科書本文で提示していたA社教科書では，低学年を除いて，一つの漢字が提示されると音訓すべての読みを提示するようになった。つまり，教科書語彙数の減少から読み替え漢字を含む漢字語句の提示が困難になったのである。このような実態のある国語科の授業で漢字を生活に即した形で教えるのは，その設定そのものに問題があると言える。

そこで，教科書コーパスから得られた教科における漢字提示頻度の実態を基に国語科以外の教科において漢字指導をおこなうことを提案する。教科書は日本のすべての児童が手にする書籍であり，家庭や学校の蔵書環境に関係なく手にすることができる。さらに，国語科以外の教科書においては，国語科よりも提示される漢字，および漢字語句が多く存在する。次ページの表 1，表 2 は第 6 学年に配当された漢字の社会科，理科の教科書における頻度の上位 20 位である。この範囲の漢字であれば，国語科より特に社会科は圧倒的に頻出する状態にあると理解できる。

† tanahasi@nara-edu.ac.jp

1　棚橋（1998）の調査では，教師の多くは読解単元の前に漢字を取り出して指導する実態がある。

【表 1】　第 6 学年の漢字頻度―国語 - 社会

No	字種	全教科	国語	社会_
1	権	104	4	95
2	皇	81	3	78
3	憲	75	2	73
4	幕	66	2	64
5	将	60	5	52
6	城	69	9	51
7	域	90	4	50
8	遺	68	23	42
9	展	77	8	38
10	聖	38	1	37
11	障	38	3	30
12	貴	29	4	25
13	忠	26	1	25
14	敬	26	2	24
15	探	46	10	23
16	割	92	12	22
17	源	44	5	21
18	済	19	3	16
19	衆	21	1	16
20	革	16	1	15

【表 2】　第 6 学年の漢字頻度―国語 - 理科

No	字種	全教科	国語	理科
1	層	83	1	78
2	磁	68	2	64
3	灰	51	3	48
4	吸	80	10	28
5	砂	39	5	24
6	肺	33	2	23
7	呼	50	10	22
8	蒸	24	2	20
9	臓	44	1	19
10	割	92	12	17
11	針	70	3	11
12	骨	23	9	10
13	域	90	4	9
14	宇	17	3	8
15	筋	16	3	8
16	宙	17	3	8
17	棒	26	2	8
18	株	9	1	7
19	収	17	3	6
20	卵	14	3	5

【表 3】　第 6 学年社会科における提示語句

熟語および漢字	全教科	国語	社会
権利	38	3	34
人権	22	1	19
権	14	0	12
主権	9	0	9
民権	9	0	9
執権	3	0	3
法権	3	0	3
権限	2	0	2
参政権	2	0	2
権力	1	0	1
実権	1	0	1

一方，表３は，社会科において提示頻度が一位であった「権」について，教科書上にどのような語句が使用されているかを示したものである。表を見ると，国語科では「権利」と「人権」の二つの語句のみの提示であるのに対して，社会科の異なり語数は 11 になる。

社会科で提示された 11 語句のうち，学習指導要領上に明示された語句は，権利―「日本国憲法は，国家の理想，天皇の地位，国民としての権利及び義務など国家や国民生活の基本を定めていること。」，

主権―「我が国の政治の働きについて，次のことを調査したり資料を活用したりして調べ，国民主権と関連付けて政治は国民生活の安定と向上を図るために大切な働きをしていること，現在の我が国の民主政治は日本国憲法の基本的な考え方に基づいていることを考えるようにする。」参政権―「また，イの「国民としての権利及び義務」については，参政権，納税の義務などを取り上げること。」の三語句であり，これらは社会科の指導内容の中核にかかわる語句だと言える。そして，このような語句については社会科での漢字学習が妥当である。その理由は，まず「見慣れ」の機会をより多く持てること，さらに漢字の学習が教科内容の理解を強化すると考えられることによる。そのような立場から本研究では，小学校 6 学年児童を対象にした実践を行った。

2. 社会科における「憲」の学習

2.1 実践の概要

実際の授業は平成 22 年 12 月 15 日に第 6 学年の社会科で実施した。私が学習指導案を書き，担当教師と打ち合わせをおこなったのち実践に移った。

2.1.1 授業者等の情報

対象クラス　奈良県御所市立御所小学校 6 年 3 組（29 名）

授業者　　伊藤輔教諭

2.1.2 単元と目標

単元　新しい憲法ができた（大阪書籍）

目標　1. 日本国憲法の内容を理解し，戦後すぐに制定されたことに対し考えを持つことができる。

2.「憲」の意味を理解し，漢字の定着を図る。

2.2 実践の記録

以下，漢字学習に関係のある部分の授業記録を載せる。発言は教師（T），児童（C）とも意味を崩さない程度に書きまとめた。下線部分は特に漢字指導に関する部分である。

T：覚えていますか，みなさん。それで，少し思い出してほしいんねんけど，憲法ってどんなもんでしたか？

C：ルール。

C：この国のルールで，一番強いルール。

A

T：一番強いルール。おおっ，最強のルール。なるほど。そういう考えかたもできる。実は一番基になる法律なんですよと，いうことです。で，なぜ憲法が基になる法律なのか，分かる？これはね，とっても大事なことやねんけど，「憲」ていう字が，この漢字自体に，「基になる」っていう意味があるからなんです。

T:プリント渡します。まず，名前書いてや。

T：(「憲」の成り立ちを見せて)これなんていう漢字か分かります？

C：憲法の「憲」。

T：おぉ，その通り。この「憲」っていう字の昔の形は，こういう形になってん。ちなみに，はい，この部分は，何表しているか分かる？

C：目。

T：目やね，目のことを表していますよ。では，この部分は。

C：心。

T：そうやねん。これ心を指します。これね，ふたが目と心をおおっているんです。人間ってさぁ，何かを考えるときに，目とか心で考えるやんか。それを，ルールっていうもので隠してしまうっていうことや。なっ，憲法の「憲」っていうのはそういうふうに形ができているんですよ。まずそれを確認しますね。その次に，「憲」っていう字をまず書いてもらいたいと思います。はい。手，挙げて。順番にいくよ。大きく書くよ。筆順確認します。先生に合わせていってください。はい，いくよ。いち，にい，さん…（教師と一緒に筆順を空書する。）はい。筆順確認できたら，ワークシートに書いてごらん。

T：読みかたは，「けん」やな。意味，ていねいな字で書いていこうな。一番基になるという意味です。

T：はい，それではですね，書けた人は顔をあげてください。それでは，日本国憲法の実際の中身についてみていきたいと思います。では，まず教科書 112 ページから確認していきたいと思います。

（社会科の指導が続く。）

T：はい。日本国憲法の三原則といわれるものがあります。三つ大事なことがあるんです。その三つ大事なことが，これから先，読んでもらう部分にはいっています。それがどこか分かったら，線をひきながら聞いてください。はい。では，ここのところ読んでくれる人？

（音読）

T はい。では，そこのところ言ってみてください。

C：国民が主権者となり，天皇は国や国民のまとまりの象徴であるとされました。

C：すべての国民の基本的人権を保障。

C：軍隊を持たず永久に戦争をしないこと。

T：まず，この三つを書いてください。丁寧に書いてね。

T：書けましたか？書けた人は顔をあげてください。最初のほうにも少し話したんですが，1945 年 8 月 15 日に戦争が終わりました。で，そこからすぐに，日本の憲法が 46 年に決まり，47 年には始まっていきました。なぜこんなにすぐに日本国憲法を作らないといけなかったのか。一番基になるという意味があるんやったらそんなに簡単に決めていいのかな。（中略）三原則，国民主権，基本的人権の尊重，戦争放棄もしくは平和主義。これは結局，何が言いたかったのか。

C：戦争に負けた。

C：日本が平和になる。

T：うん，きっとそうやろうね。日本が平和になりましたよ，これからはもう大丈夫ですよってことを早く言わなきゃあかんかったわけやな。でもそのためにはちゃんと考えなあかんこともたくさんあったやろうけど，それが日本の国民だけでよかったんですか？―まわりの全世界の人にも平和になったってことをアピールしなければいけなかったんですよ。また詳しくは，今度説明しますが，サンフランシスコ講和会議とか，そこによって日本が独立したこととか，国際連合に加盟したこととかっていうのは，結局日本が平和になったからできたことで，それがもし平和になってなかったら，世界は認めてくれないっ

ていうことなんです。

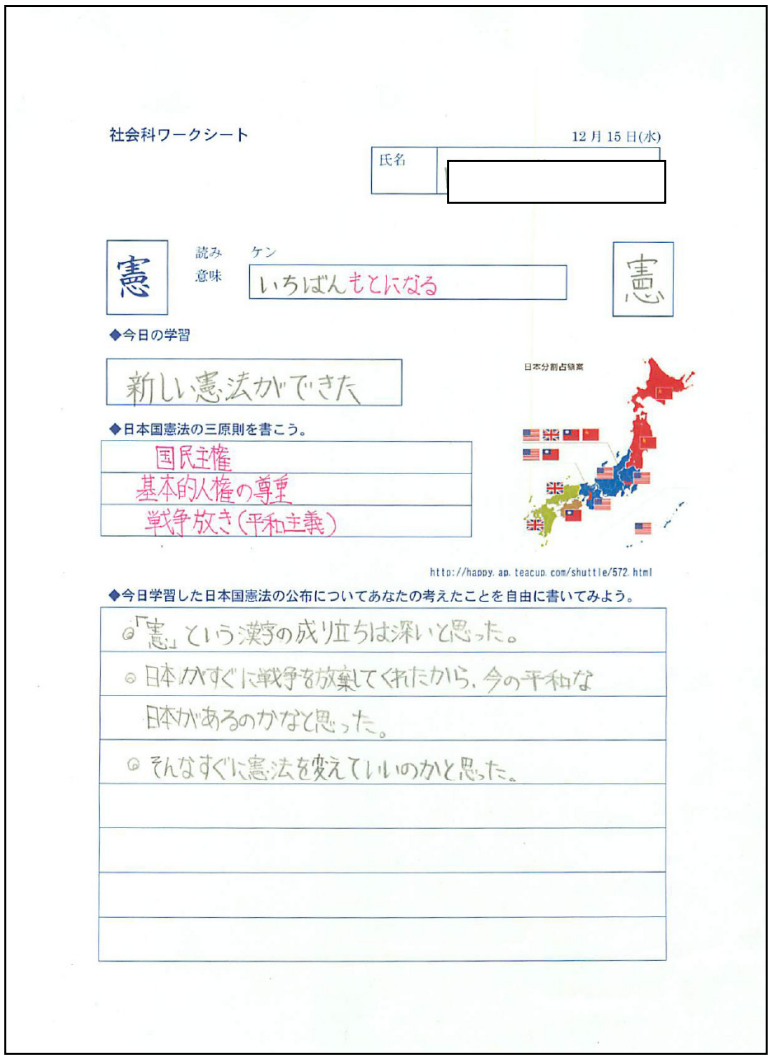
B

T：では，ワークシートの下のほうの「今日学習した日本国憲法の公布についてあなたの考えたことを自由に書いてみよう」をしてみましょう。

2.3　漢字指導場面

授業記録の中には，特に 2 か所漢字指導を目的とした場面があった。それは，A と B の囲みの部分である。A の部分でおこなった指導は，通常の国語の時間における漢字指導の内容とほぼ同じであり，漢字の成り立ち，読み，字形と筆順，意味を確認し，一度練習をさせている。授業を構想する際に，意味の確認が本時の学習の中心事項である「憲法」の意味理解につながるように考慮しており，その点は担当教師が効果的に授業を進めうまくいったと考える。

また，B の部分は，「憲法」を必然的に書く場を与えるねらいで設定した学習であり，次ページに掲載するワークシートの下半分を使って自分の考えを書かせるようにした。憲法の公布について書くことになり，自然に「憲」の字が習熟できることを目的としている。漢字そのものがまだ理解できていなければ，ワークシートの上部を見ればその漢字が表示されている。さらに，この実践では，A の部分も B の部分もワークシートを活用しながら授業が進めていけるようにした。


社会科ワークシート

12 月 15 日(水)

氏名

憲

読み　ケン

意味　いちばんもとになる

憲

◆今日の学習

新しい憲法ができた

◆日本国憲法の三原則を書こう。

国民主権
基本的人権の尊重
戦争放さ（平和主義）

日本分割占領案

http://happy.ap.teacup.com/shuttle/572.html

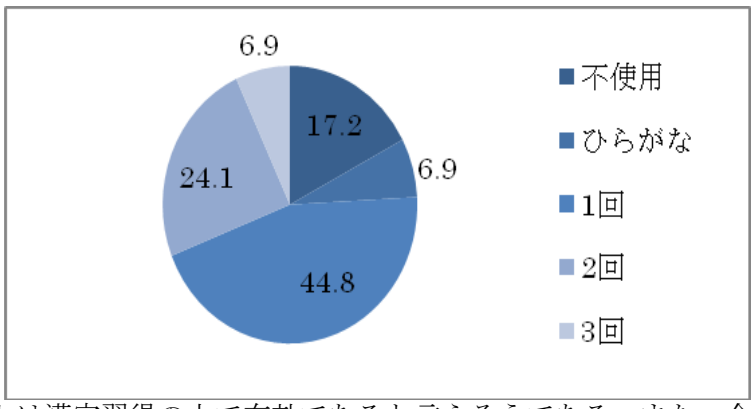
◆今日学習した日本国憲法の公布についてあなたの考えたことを自由に書いてみよう。

①「憲」という漢字の成り立ちは深いと思った。
② 日本がすぐに戦争を放棄してくれたから，今の平和な日本があるのかなと思った。
③ そんなすぐに憲法を変えていいのかと思った。

【資料 1】ワークシートの実際

ワークシートの自由記述欄について「憲法」と漢字で書いているかについては以下のグラフに示す実態であった。

【グラフ 1】ワークシートの自由記述部分での「憲」の使用（%）



この結果を見る限りおおむね児童は漢字を使って「憲法」と記述をしていることが理解できる。今回は，対照実験を実施しておらず，確定的なことは述べられないが，教科で中心となる語句と漢字指導を結びつけていくことは漢字習得の上で有効であると言えそうである。また，今回の授業では，内容と授業進度の関係で自由記述にかける時間が少ししか取れなかった。この点が改善できれば，さらに自然な形での漢字練習ができると考えられる。

3. 授業に対する児童の反応

3.1 アンケート概要

本実践では，授業後，その効果をみるために以下のようなアンケートを実施した[2]。

◆今日の授業では、社会科の中で漢字の勉強をしました。このことについてみなさんの意見を聞かせてください。

（1）漢字の勉強をすることで授業の内容はよく分かりましたか。当てはまるものの番号に○をつけてください。

①よく分かった。　１１人　②まあまあ分かった。　９人

③それほど分からなかった。　６人　④分からなかった。　２人

（2）国語以外の教科で漢字の勉強をすることについてどのように思いますか。自分の考えを書いてください。

◆あなたは漢字の勉強が好きですか。当てはまるものの番号に○をつけてください。

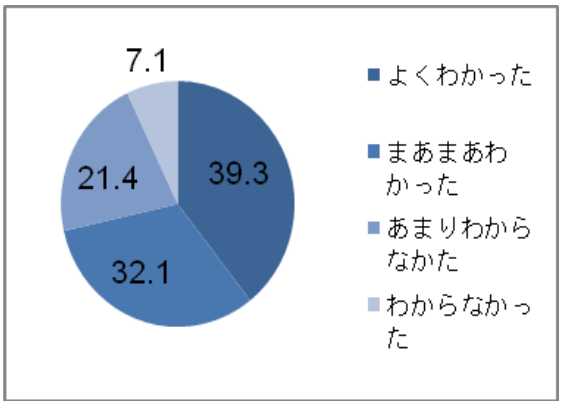
①　とても好き。　５人　②まあまあ好き。　９人

②　あまり好きではない。　８人　④きらい。　６人

[2] 実際のアンケートは総ルビ表記とした。

3.2　漢字学習と内容理解との関係

【グラフ 2】漢字学習と内容理解（%）



本アンケートの実施にあたって一人の児童が，「正直に書いていいの。」という発言をした。それに対し，指導の教員も参観していた私も「もちろん，そのほうがいいです。」と念を押したため，本アンケートには児童の実感が反映されているものと考える。「よくわかった」「まあまあわかった」と答えた児童は全体の 7 割である。この数値を高いとみるか低いとみるかは前述のように対照的な調査ができていないので明言はできないが，漢字を学んだことが内容理解の一助にはなっていると判断してもよさそうである。しかし，児童の自由記述を検討すると，そこにはかなりの抵抗感があることが分かる。以下は児童の自由記述の全体である。否定的な感想の部分に下線を付す。

① 漢字以外の勉強をしてほしいと思った。
② 色々な漢字がわかっておもしろいと思う。
③ 例えば算数の問題文の意、国語の文章の意図がわかりやすくなった。
④ 社会、楽しみにしてたから、ふつうに社会やりたかった。漢字、あんま好きじゃないです。社会は社会でやりたかったです。あんま、こうゆうのやりたくないです。
⑤ 不思議な感じがするけど、新鮮だし知識も増えるし楽しい。
⑥ かんじのべんきょうはきらいなのでしたくないです。社会はノートをたくさんかきたいです。もうしたくないです。
⑦ 算数の時間にやってもらえるとうれしいです。
⑧ あまり国語以外でやることはないのでよかったです。
⑨ 漢字は好きではないのであまりこのまない。
⑩ それはそれであるいみいいとおもいました。でも漢字は好きじゃないのでちょっといやでした。めんどくさくはありません。
⑪ めんどっちい、おとろしい、おもしろい、漢字うっとうしい。
⑫ 社会の時間に漢字の勉強するなんていいきかいだったと思う。
⑬ 初めは、社会やのになんでやろと思ったのに、「憲」という漢字の意味をしって、すごくよかったです。
⑭ 国語以外も好きだと思います。でも、算数はちょっと苦手です。
⑮ 社会の時間だから社会がしたかった。
⑯ 国語以外かんじをやるなんて、びっくりしました。

⑰ あまりなかったのでじゃっかん変だった。おかしいと思う。
⑱ 難しすぎて分からないけど好き。
⑲ 社会でよく出てくる漢字をくわしく勉強することで、その単語の意味が分かって良いと思いました。
⑳ すごと思う。
㉑ いわかんがあった。
㉒ たまに社会の時間にそう言うこともいいと思う。
㉓ 別にいいと思います。その漢字を習うことによって、その教科につながるなら、いいと思いました。
㉔ 別にいいと思った。社会がきらいだから。
㉕ へんだと思った。ふつうかんじのべんきょうで、かんじをするから。

否定的な感想には，二つの類型があった。一つは，漢字そのものの学習が嫌いだということである。これについてはアンケート項目の（2）に関連項目があるが，漢字嫌いの児童が学級の約半数に及ぶことがわかる。もう一つは社会科で漢字を学習することについての「違和感」である。この調査とは別に教員対象のアンケート調査[3]も行ったが，むしろ，その結果より学習の主体者である児童自身のほうが，国語科以外の漢字学習に抵抗感が高いことが印象的であった。

4. おわりに

今回，実際に社会科の中で漢字学習をおこなう授業の実践を検討することで，教科内容の理解に資する学習，漢字の習熟が期待できる学習が組織できる手ごたえを感じた。今回の実践は，研究の位置づけの中では，予備的なものだと言えるが，できるだけ早く学年別漢字配当表を教科配当することを考案し，小学校全体の教科カリキュラムを漢字習得の観点から見直していきたいと考える。最後に実践に協力いただいた奈良県御所市立御所小学校と伊藤輔氏に謝意を表する。

文　　献

興水実（1971）「漢字学習指導改造の視点と課題」国語教育研究所編『漢字の読み書き分離学習』，明治図書，p14

[3] 発表で詳述する。

コーパスに基づく分類重要語彙リスト
—学校教育での活用に向けて—

田中　牧郎（言語政策班班長：国立国語研究所言語資源研究系）†

A List of Important Words Grouped by Meaning and Subject: Towards the Use of Corpus in School Education

TANAKA Makiro　（National Institute for Japanese Language and Linguistics）

１．言語政策班が公開した語彙表

言語政策班では、最終成果物として四種の語彙表を公開した。

（１）教科書コーパス語彙表

言語政策班が作成した教科書コーパス（小学校・中学校・高校の全教科全学年一種ずつの全文コーパス）[1]に用いられている語彙を対象に、語彙素・品詞・語種などの情報について、校種・学年・教科別の頻度、BCCWJの図書館サブコーパスと比較した際の特徴度などを収録した語彙表。

（２）BCCWJ 主要コーパス語彙表

BCCWJ のうち、図書館書籍（固定長）、出版書籍（固定長）、雑誌（固定長）、新聞（固定長）、Yahoo!知恵袋（可変長）、Yahoo!ブログ（可変長）の 6 つのサブコーパスの語彙頻度を収録した語彙表。各サブコーパスに対してカバー率の基準を適用して、5 つに区画した語彙レベルを設定。

（３）学校・社会対照語彙表

「教科書コーパス語彙表」に収録した語彙頻度・特徴度と、「BCCWJ 主要コーパス語彙表」に収録した語彙頻度・語彙レベルとを比較対照し、中学校・高等学校において、社会で必要な語彙を考慮した教育を考える基礎資料となるように再編した語彙表。

（４）教科特徴語リスト

中学校・高等学校の、長い単位での教科特徴語のリスト。「教科書コーパス」のうち主要学習部分のみを対象にした教科別の語彙頻度と、図書館サブコーパスの語彙頻度を比較して教科特徴語の候補とした後、人手により確認し教科特徴語と認めてよいものだけを残したもの。

それぞれ、国語政策・国語教育の様々な局面で役立つ語彙表だと思うが、いずれも基礎資料として利用されることを想定して作成したものである。実際の活用場面では、これらの語彙表をもとにさらに研究を進め、より具体的な語彙リストを作り、語彙表からコーパスそのものを参照するような作業が求められるだろう。本稿では、その具体例の一つとして、中学校・高等学校における各教科の専門教育と国語科での語彙教育を連携させるために、重要語彙のリストを作って語彙分析を行いながら研究する方向について考えてみたい。

† mtanaka@ninjal.ac.jp

1 「教科書コーパス」の全体は、著作権者との合意による非公開だが、一部を BCCWJ 非母集団サブコーパスの教科書サンプルとして公開の予定（BCCWJ 検索デモサイトでは既に公開中）。

なお、本稿は、言語政策班報告書である田中ほか（2011）において、田中（2011a）・同（2011b）として発表したものの要点をまとめたものである。

２．BCCWJ による重要語彙リスト

「BCCWJ主要コーパス語彙表」と「学校・社会対照語彙表」には、BCCWJの主要な六つのサブコーパスの語彙レベルを収録した。これは表１に示すカバー率を基準に区画したものであり、これによって、規模の異なるサブコーパスであっても相互に均質なレベル分けが可能になり、多様な性質を持つサブコーパス間の語彙レベルを比較対照することによって、語彙の性質を分析したり分類したりするのに役立つと考えられる。このカバー率によって区画した各レベルに配される語数をまとめたものが表2 である。

表１　カバー率による語彙レベルの設定

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　各サブコーパスの延べ語数・異なり語数

	LB_FL		PB_FL		PM_FL	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
全体	3,938,696	86,002	3,903,395	82,784	896,988	45,900
レベル a	3,074,655	4,177	3,045,639	3,842	700,831	4,336
レベル b	395,994	6,330	391,312	5,609	92,353	5,293
レベル c	242,911	11,595	239,221	10,506	51,085	7,493
レベル d	118,642	14,176	124,601	14,290	37,925	13,984
レベル e	106,494	49,724	102,622	48,537	14,794	14,794

	PN_FL		OC_VL		OY_VL	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
全体	624,020	35,727	2,762,864	49,809	6,127,125	76,823
レベル a	486,976	3,420	2,155,871	2,071	4,779,106	3,441
レベル b	63,784	4,045	275,758	2,776	617,945	4,724
レベル c	40,018	6,941	165,957	5,122	372,114	8,406
レベル d	20,607	8,686	83,349	7,062	181,482	10,285
レベル e	12,635	12,635	81,929	32,778	176,478	49,967

田中（2011a）では、六種のサブコーパスの語彙についてレベルと語種の観点から比較し、また、各サブコーパスのレベル a の特有語を分析し、語彙の特徴を導き出したが、その要点をまとめると表３のようになる。この特徴から、学校教育における規範となるような重要語彙リストを作る際に、第一に活用すべきものは、LB（図書館書籍）の語彙レベルであると結論づけた。

表３　各サブコーパスの語彙の特徴

サブコーパス	語彙の特徴
LB（図書館書籍）	一般的な語彙のありようを反映し、文章語をよく取り込み、語彙の基本的な部分が安定している。
PB（出版書籍）	文章語的な語彙、実用的な語彙を多く含み、語彙の周辺的な部分には、新しい語彙・感覚的な語彙も取り込んでいる。
PM（雑誌）	語彙の基本的な部分にまで、新しい語彙・感覚的な語彙を取り込んでいる。また、特定の話題の語彙に偏る傾向もある。
PN（新聞）	文章語的な語彙が特に多く、語彙の基本的な部分にまで、それを取り込んでいる。時事的な語彙も多く含む。
OC（知恵袋）	日常的な語彙が多くを占め、新しい語彙・感覚的な語彙も入り込んでいる。媒体としての語彙の特徴は乏しい。
OY（ブログ）	日常的な語彙が多くを占め、新しい語彙・感覚的な語彙も入り込んでいる。日記や語りに特徴的な語彙も目立つ。

図書館書籍（LB）の語彙レベルをもとに重要語彙リストを作る場合、ａ～ｅに区切った語彙レベルのある水準までを「重要語彙」として区画できると便利である。どこで線引きをするかを決めることは難しいが、今回は作業仮説として、レベルａ・ｂ・ｃの範囲である約 22,000 語を、主として中学生を想定した重要語彙の範囲と扱うことにした[2]。

３．教科書コーパスと BCCWJ 図書館サブコーパスの比較による教科特徴語リスト

「教科書コーパス語彙表」は、語彙頻度について図書館書籍（LB）と比較した際の各教科の特徴度（対数尤度比による）を収録している。その特徴度が一定水準以上のものを教科特徴語と扱い、その情報を「学校・社会対照語彙表」に掲出した[3]。その情報をもとに、各教科の特徴語の語数と、当該教科の語彙全体の中での特徴語の比率を算出すると表４のようになる。重要語彙としたレベルa・b・cの範囲と語彙全体とに分けて掲げた

この表によれば、まず全体に、レベルａ・ｂ・ｃの範囲よりも語彙全体の方が特徴語の比率が高いことが確認できる。レベル ａ・ｂ・ｃ の範囲について教科ごとに見ると、特徴語の語数では、社会・理科・技術家庭の順に多く、数学・外国語・情報の順で少ない。同じく特徴語の比率で見ると、情報・理科・技術家庭・保健体育・数学・社会で高く、国語・外国

[2] この約 22, 000 語のリストを具体的に見ていくと、中学生に対してわざわざ取り立てて学習させるまでもないほど平易と感じられる語彙や、反対に、中学生にはまだ不必要ではないかと思われるような専門的な語彙が混じっていることに気付かれる。そこで、例えば平易な語彙を排除するために、知恵袋（OC）の語彙レベルがａ・ｂのものは重要語彙リストから除外し、また、専門性の高い語彙を排除するために、図書館書籍（LB）のサンプル数が 3 以下のものを重要語彙リストから除外することなどが考えられる。田中（2011a）でそれらを試行したところ約 17,000 語となり、その内容を確認すると重要語彙リストとしての妥当性は高まると考えられた。こうした、重要語彙リストの検証や評価を行い、教育現場で実際に使える語彙リストに仕上げていくことは今後の課題である。本稿では、この作業を経る前の約 22,000 語の範囲を重要語彙と扱う。

[3] 教科特徴語の性格や、「学校・社会対照語彙表」作成の考え方については、田中ほか（2010a）、田中（2010b）に述べた。

語・芸術で低い。

表４　教科特徴語の語数と比率

	国語	数学	理科	社会	外国語	技術家庭	芸術	保健体育	情報
特徴語の数（レベルａ・ｂ・ｃ）	1349	664	1878	2838	732	1642	1348	1019	932
特徴語の比率（レベルａ・ｂ・ｃ）	9.9%	20.0%	22.0%	19.9%	15.2%	21.7%	15.8%	20.7%	25.1%
特徴語の数（全体）	3134	1152	4333	5713	1159	2521	4482	1412	1298
特徴語の比率（全体）	14.4%	27.4%	30.7%	21.8%	20.4%	24.5%	31.9%	24.3%	29.9%

重要語彙で教科特徴語になっているものは、その教科で触れる機会が特に多いと考えられるので、生徒の語彙習得は教科学習とともに進むのではないかと考えられる。また、複数教科で特徴語になっているものは、教科間の連携をはかることで語彙習得の効果は向上するのではないかと思われる。教科学習の中で行われる語彙習得について研究することは、知識の獲得や概念の形成・確立といったことと関連づけて語彙学習や語彙教育のあり方を考えることにつながっていくだろう。

４.「分類重要語彙リスト」の作成

語彙教育の問題を知識の獲得や概念の形成と関連づけて考えるには、意味という視点が重要になる。「学校・社会対照語彙表」には、語彙素に対して『分類語彙表 増補改訂版』（国立国語研究所）の番号が関連づけてある[4]。この番号を利用して、意味の視点を取り入れた語彙分析を進めていきたい。『分類語彙表 増補改訂版』は、多義語には複数の番号を与えているが、これを別々の語に分けて語を認定したものも、「学校・社会対照語彙表（分類語彙表番号分割）」として公開している。２で見た重要語彙約 22,000 語は、この分割版によれば、約 31,000 語になる。

この語彙表から、教科の特徴語と媒体の特徴語[5]の情報を見やすい形で抽出し、図書館書籍（LB）の語彙レベルがａ・ｂ・ｃの範囲（本稿で重要語とする範囲）について、『分類語彙表 増補改訂版』の番号順に配列したものを、「分類重要語彙リスト」と称して作成する。この語彙リストを用いることで、学校教育における語彙について様々な角度から分析することが可能になる。

「分類重要語彙リスト」に収録される語彙について、『分類語彙表 増補改訂版』の大分類別に教科特徴語の語数を示すと表５のようになる。この表によれば、例えば、「1.1 体の類_抽象的関係」では理科・社会の特徴語がほぼ同数の最多で技術家庭がこれに次ぐこと、「1.2 人間活動の主体」「1.3 人間活動_精神および行為」では社会の特徴語が際立って多いこと、「1.4 生産物および用具」「1.5 自然物および自然現象」では技術家庭と理科の特徴語が特に多いことなどが見て取れる。意味分野と教科特徴語との対応を観察していくことで

[4] この関連づけ作業は完全なものではなく、完全な関連づけのためには解決を要する問題が残されている。
[5] 図書館書籍（固定長）全体と教科書全体の語彙頻度を比較して、どちらか一方に有意に偏っているものを媒体特徴語と扱った。

教科教育と語彙との関わりについて具体的な検討ができるようになるが、さらに細分された「中項目」の枠組みで見ていくと、より踏み込んだ研究が可能になる。「1.1 体の類・抽象的関係」に属する中項目「1.14 力」を例に、そのことを見ていこう。

表５　『分類語彙表 増補改訂版』の大分類別の教科特徴語（語数）

		全体	教科特徴語								
			国語	数学	理科	社会	外国語	技術家庭	芸術	保健体育	情報
1.1	体の類_抽象的関係	5500	382	497	948	947	229	633	453	396	464
1.2	人間活動の主体	2321	168	27	50	488	89	76	115	78	41
1.3	人間活動-精神および行為	6282	620	153	333	1051	361	590	421	406	437
1.4	生産物および用具	2051	115	61	219	157	63	304	196	84	92
1.5	自然物および自然現象	2160	113	22	519	188	40	282	146	214	37
2.1	用の類_抽象的関係	2214	107	119	248	212	57	173	122	131	89
2.3	用の類_精神および行為	2416	165	112	150	240	114	210	138	159	96
2.5	用の類_自然現象	340	12	7	35	13	2	19	18	16	3
3.1	相の類_抽象的関係	1482	63	64	120	133	44	91	91	73	55
3.3	相の類_精神および行為	986	30	9	31	44	24	39	64	38	24
3.5	相の類_自然現象	334	7	1	29	6	3	17	23	10	4
4	その他	273	12	17	14	7	7	1	8	5	13
	小計	26359	1794	1089	2696	3486	1033	2435	1795	1610	1355
	該当なし	4334	187	49	170	397	111	91	218	55	66
	合計	30693	1981	1138	2866	3883	1144	2526	2013	1665	1421

５．分類重要語彙リストを使った教科指導と語彙指導の関連づけ

表６は、「分類重要語彙リスト」の「1.1400 力」「1.1401 弾力・動力・圧力など」の部分を示したものである。「教科特徴語」だけでなく、教科書と書籍（図書館書籍・LB）を比較して抽出した「媒体特徴語」、教科特徴語にも媒体特徴語にもならない「無特徴」の語の情報も掲げた。これによれば、「1.1400 力」には、教科特徴語や教科書特徴語が少なく書籍特徴語が多いこと、「1.1401 弾力・動力・圧力など」には、教科特徴語や教科書特徴語が多く、書籍特徴語が少ないことが明らかである。「1.1400 力」の語彙は、教科学習以外で習得されるものが多く、「1.1401 弾力・動力・圧力など」の語彙は教科学習を通して習得されるものが多いことが推測される。また、教科で扱われる語が少ない「1.1400 力」の語彙においても、最も重要度の高いレベル a の語彙では、「力」「エネルギー」の２語が多くの教科で特徴語となっており、こうした根本概念については、教科学習を通して生徒の身についていくのではないかということをうかがわせる。

表６で「1.1401 弾力・動力・圧力など」の部分を見ると、理科に 9 語、社会に 13 語もの

特徴語があり、そのうち 7 語（抵抗・圧力・反発・摩擦・動力・電力・火力）は両科目に共通している。これらの語彙について少し詳しく考察してみよう。

表６　「分類重要語彙リスト」の「1.1400 力」「1.1401 弾力・動力・圧力など」の部分

分類項目		レベル	教科特徴語									媒体特徴語		無特徴
			国語	数学	理科	社会	外国語	技術家庭	芸術	保健体育	情報	教科書	書籍	
1.1400	力	a	—	—	力、エネルギー	エネルギー、実力	—	エネルギー	—	力、エネルギー	—	力、エネルギー	パワー、強力、実力、強烈	—
		b	—	—	—	主力	—	—	—	全力	—	—	威力、迫力、全力	自力、無力
		c	—	—	強、弱	総力、弱体	—	—	—	—	—	強、総力、弱	パンチ、魔力、非力、非力、脆弱、痛烈、激烈、熾烈	無敵、最強、他力、余力
1.1401	弾力・動力・圧力など	a	—	入力	抵抗、入力、圧力	抵抗、圧力	—	抵抗、入力	—	抵抗、血圧	入力	抵抗、入力、圧力	血圧	—
		b	—	気圧	反発、摩擦、重力、動力、電力、気圧	反発、摩擦、動力、電力、圧迫	—	動力、電力	—	圧迫	出力	反発、摩擦、重力、動力、電力、出力、圧迫、気圧	—	—
		c	—	—	引力、火力、高圧、電圧	人力、火力	—	弾力、火力、電圧	弾力		電圧	引力、火力、高圧、電圧	応力、テンション、眼圧	—

「1.1401 弾力・動力・圧力など」において、理科と社会に共通する特徴語７語のうち、やや具体性の強い「動力」「電力」「火力」を除いた、抽象概念を表す４語について、国語辞典（『デジタル大辞泉』小学館）の記述と、理科・社会の教科書の主な用例を対応させて一覧にすると、表７のようになる。

「抵抗」を例に取れば、理科の教科書では３番目または４番目の語義に相当する使われ方をしており、社会の教科書では１番目の語義にあたる使われ方をしている。前者の語義は自然科学の専門用語を基盤にもっており、後者の語義は一般用語に基づくものだと考えられる。専門用語の＜流体中の物体が流れから受ける反対方向への力＞という意味は、一般用語の＜進もうとする自身が社会や他人から受ける反対方向の力＞という意味と有機的

に関連づけられ、「抵抗」の意味・概念が学習者の中に確立していけば、この語についての語彙力は確かなものになるに違いない。表７の「圧力」「反発」「摩擦」においても、理科の学習で専門用語としての語義を、社会の学種で一般用語としての語義（の一部）を学習する機会が多いことが分かる。

表７　理科と社会に共通の特徴４語の意味と教科書の用例

	国語辞典（デジタル大辞泉）の用例	理科の主な用例	社会の主な用例
抵抗	１ 外部から加わる力に対して、はむかうこと。さからうこと。「権力に―する」「大手資本の進出に地元の商店会が―する」		独立に抵抗するフランス人、抵抗権
	２ すなおに受け入れがたい気持ち。反発する気持ち。「相手の態度に―を感じる」「一人で入るには―がある」		
	３ 流体中を運動する物体が流れから受ける、運動方向と逆向きの力。	空気の抵抗、水の抵抗	
	４ 「電気抵抗」の略。	電気抵抗	
圧力	１ 押さえつける力。気体・液体または固体が、ある面を境にして、その両側から垂直に押し合う力。単位はパスカルのほか、アト、水銀柱メートルなどを用いる。	空気に圧力を加える、重力による圧力、温度や圧力の変化	
	２ 威圧して服従させようとする力。「大国の―に屈する」		欧米列強の圧力、裁判所に圧力をかけ、圧力団体
反発	１ 他人の言動などを受け入れないで、強く否定すること。また、その気持ち。「―を買う」「運命に―する」		親米政権への反発、教皇庁の搾取に反発する、民衆の反発
	２ 他からの力をはねかえすこと。はねかえること。「磁石の同じ極どうしは―する」	髪の毛が反発しあって逆立つ、磁石を置くと反発力をうみ、反発係数	
	３ 値下がりしていた相場が一転して値上がりすること。⇔反落。		
摩擦	１ 物と物とがすれ合うこと。また、こすり合わせること。「肌を―して暖をとる」「乾布―」		
	２ 人間の社会関係で、二者の間に意見や感情の食い違いによって起こる、不一致・不和・抵抗・紛争など。軋轢(あつれき)。「貿易―」		競争と摩擦、貿易摩擦、日米摩擦
	３ 互いに接触している二つの物体のうち、一方が運動しようとするとき、または運動しつつあるとき、その接触面に運動を妨げようとする力が働く現象。また、その力。相対速度により運動摩擦・静止摩擦に、運動状態により滑り摩擦・転がり摩擦などに分けられる。	燃焼・摩擦・気体の圧縮、摩擦力、摩擦熱	

これらの語彙の指導においては、教科学習での扱われ方を踏まえて行われるのが、効果的だと思われる。従来は、教科間での語彙の比較対照があまり行われていなかったため、こうした配慮や工夫を考える機会も少なかったが、コーパスに基づく「分類重要語彙リスト」ができたことによって、教科指導と語彙指導を関連づけて、語彙をめぐる教育のあり方を論じていくことが期待されよう。

６．今後に向けて

BCCWJ と教科書コーパスを基盤に置いた「分類重要語彙リスト」は、語彙についての教育の問題を考える材料を豊富に提示してくれている。広範囲の語彙のレベル・意味・教科・媒体などへの分布を概観する方法でも、特定の意味分野や個別の語に対象を絞り込んで用例も参照しつつ踏み込んで分析する方法でも、従来では望めなかった研究を進めていくことができそうである。まずは、教育上のさまざまな具体的課題への検討において、語彙リストを参照し、コーパスの用例を検索しつつ作業を進めていく試みを重ねていくことが求められている段階だと思われる。課題の性質に応じて、必要になる語彙リストの形式は異なったものになる場合もあるだろう。国語政策や国語教育の分野にコーパスを利用するのは、始まったばかりである。

文献

田中牧郎（2011a）「語彙レベルに基づく重要語彙リストの作成 ―国語政策・国語教育での活用のために―」（田中ほか2011所収）

田中牧郎（2011b）「『分類重要語彙リスト』の作成による教科教育と語彙教育の関連づけ」（田中ほか 2011 所収）

田中牧郎・近藤明日子・河内昭浩・鈴木一史・棚橋尚子（2010a）「＜学校の語彙＞と＜社会の語彙＞―「教科書コーパス」と「流通実態サブコーパス」の比較―」（『特定領域研究「日本語コーパス」平成 22 年度全体会予稿集』

田中牧郎・近藤明日子・河内昭浩・鈴木一史・棚橋尚子（2010b）「『学校・社会対照語彙表』の作成と活用」（『日本語学会 2010 年度秋季大会予稿集』）

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）『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』（特定領域研究「日本語コーパス」言語政策班）

外形で引く国語辞典への試み

矢澤真人（辞書編集班分担者：筑波大学大学院人文社会科学研究科）†

How to Look up Words without Semantical and Grammatical Information

Makoto Yazawa (Graduate School of Hum. and Soc. Sci., University of Tsukuba)

０．はじめに

国語辞典編集班（筑波大学グループ）では、コーパス調査を元にした国語辞典の記述内容の検証を進め、現場と直結した研究を進めてきた。このうち、漢語動詞の自他に関わる調査・研究ついては、動詞の直前形態に注目した分析を行うとともに、「自他両用漢語動詞辞典」の作成を進めた（成果ＤＶＤ所収）。

一方、国語辞典の障碍についての考察を進め、利用者の観点から見た国語辞典の意味分類や記述の妥当性や理解のしやすさについても考察を進めた。今回の発表では、この一環として、見出し語およびブランチへの導引について報告する。

１．国語辞典の障碍

書籍型国語辞典には、１）読みの障碍、２）分節と整形の障碍、３）五十音配列の障碍、４）ブランチ選択の障碍などがあり、現時点での電子版国語辞典には、５）表示情報量の障碍がある。また、国語辞典の編集側や利用者にも、６）「情報」ではなく「書籍」を売買する意識、７）「理想的利用者」と「現実の利用者」の乖離といった障碍が見られる。そもそも、８）国語辞典を引くという作業自体が文書の作成や閲覧という本来の作業を中断させる障碍であり、日本語変換システムらブラウザなどを通して、文書作成・閲覧と辞書引きがシームレス化されることが望まれる（矢澤2007・2009b）。

これらの障碍のうち、１）と３）は、すでに電子版国語辞典で解消され、２）も自動解析の手法の進展により、電子化された文書の中の語句に対する辞書引きではかなり対処できる可能性がある。現在の電子版国語辞典では、５）と相まって４）の障碍はむしろ拡大しており、緊急に対処する必要がある。意識面の障碍についても、情報化社会の進展につれて、徐々に６）は希薄になりつつあるが、依然として７）・８）は残されている。

電子版の辞典が一般化する現在、利用者の使用実態を踏まえて、電子版国語辞典の特徴に合った情報の構造化と提示を進めることが必要である。文書作成・閲覧と辞典との融合も、一部では進められているが、現時点では、分離した従来型の利用形態を前提に、これらの障碍に対処することになる。

今回の発表では、形態や位置情報など外形そのものの情報から、国語辞典に立項されている見出し語へ導く「外形インデックス」と、意味でブランチを分ける「意味優先ブランチ分け」ではなく、外形と用例の型を優先させてブランチを構成する「外形・文型優先ブランチ分け」の試みを紹介する。先に「思う」を例に、より外形的な特徴を優先させたブランチ分けの試案を示した（矢澤2010）。『明鏡国語辞典第二版』では、この研究成果の一部が取り入れられ、「思う」の記述の修正が図られた。併せて、出現位置の情報から項目へ導く、外形インデックス「『な』の種類と用法」も取り入れられている。

２．見出し語・ブランチへの誘導

２．１ 見出し語への誘導

規範性を重要視する書籍型の国語辞典では、「ったく」「んなこと」「でしょ」のような文頭の省略形・融

† myazawa@lingua.tsukuba.ac.jp

合形は立項されないことが多い。空見出しを立てることも書籍型では制約される。これらの解釈に、自立語／付属語のような文法知識はかえって混乱の元となる。文法的知識にあまり依存しない誘導が必要になる。

1) 「な」の出現位置と用例によるインデックスの試み
○絶対的位置情報
「文頭」「文と文の間」「文末」「文節末」
○相対的位置情報
「体言の前」「『の』の前」「『さ』や『よう』の前」
○言いかえ
「そんな」「それならば」「なさい」「の・である」
○話し言葉の省略形・融合形のカバー
「(ん)な」「なので」「なら」「ってな」

２．２　ブランチへの誘導

しばしば、定義用語を限定して語釈をなすことが言われるが、外国語学習者用ならばまだしも、母語話者への辞典ではそれがわかりやすいとは限らない。研究者は、格体制を正確に反映した記述を推奨するが、一般の利用者はそれほど格体系を意識していないし、実際の例では、すべての格が揃って現れるわけでもなければ、どこに現れているのか判別しにくいものが多い。理論的な格体系を元にした記述を表面に出すよりは、直前形式に注目した分類の方が実際的である(矢澤 2009a)。研究者の目から見た「正確さ」一辺倒の記述ではなく、利用者の「わかりやすさ」を測って記述することが望ましい(矢澤 2008)。

2) 「思う」の直前形式に注目したブランチ分けの試み
○直前ト格を取る用法
①ア)「と思ったら」感知　　イ)感じる・情感
②ア)判断　　イ)「と思われる」　　ウ)モダリティの客体化
○直前ヲ格を取る用法
③「～時のことを思う」④「人を思う 」⑤「思わせる」
○直前に「ト」も「ヲ」も取らない自動詞的な用法
⑥「今にして思えば」「思えば思うほど」

２．３　当日のポスター発表

ポスター発表では、「な」の外形インデックスと、「打つ」の場面別インデックス、前回のポスター発表で提案した「思う」の外形・文型優先ブランチ分けの修正版などを例に、意味や文法情報にあまり頼らないで、見出し語やブランチに誘導する試みを示す。

文献

矢澤真人(2007)「国語辞書の障碍について」(平成 18 年度報告書)
矢澤真人(2008)「国語辞典のブランチについて」(平成 19 年度報告書)
矢澤真人(2009a)「国語辞典のブランチ分けと意味記述」(平成 20 年度報告書)
矢澤真人(2009b)『辞書を知る』(新「ことば」シリーズ 22)
矢澤真人(2010)「文型と語釈」『筑波日本語研究』14 号

同時共起クラスタリングを利用した大規模テキストからの動詞類語抽出

竹内 孔一 (言語処理班分担者: 岡山大学大学院)[1]
高橋 秀幸 (言語処理班協力者: 岡山大学大学院)
小林 大介 (言語処理班協力者: 岡山大学工学部)

Extracting Verb Synonyms Based on Co-clustering Approach

Koichi Takeuchi (Graduate School, Okayama University)
Hideyuki Takahashi (Graduate School, Okayama University)
Daisuke Kobayashi (Faculty of Engineering, Okayama University)

1 岡山大学研究グループ

1.1 はじめに

岡山大学研究グループでは項構造レベルの動詞辞書を人手で構築するために大規模テキストから自動的に語義を抽出する手法の開発を行ってきた．動詞項構造辞書とは動詞の類語を概念としてシソーラス形式でまとめ，さらに動詞の使用例について項の意味役割まで付与した事例とリンクさせたデータである．既に人手による構築で，4425 語 (7473 語義) の動詞に対して例文付きで辞書を構築している (Takeuchi et al. (2010)) がさらなる拡張を行うために半自動でテキストから辞書知識を構築する手法について検討を行ってきた．その結果，動詞類語を獲得する方法として，動詞と名詞の係り関係をグラフ化して同時にクラスタを獲得する同時クラスタリングに着目し，Aizawa (2002) が提案した手法を改善することで高い精度が得られることを明らかにした (Takeuchi and Takahashi (2009))．

本報告ではさらなる改善として半教師あり同時共起クラスタリング (竹内・高橋 (2010)) において，制約と収束条件との組み合わせ検討を中心に報告する．また近年提案されたグラフベースの同時クラスタリングが可能な Weighted kernel k-means 法 (以下 WKK 法) との比較をより正確に行ったので結果を報告する．

2 半教師あり同時クラスタリング

本研究で開発を進めている同時クラスタリング (Co-Clustering with Recursive Elimination (CCRE)) は Aizawa (2002) が提案した手法に多義語を繰り返し初期クラスタから獲得することで，より多くの動詞類義語集合を対応する名詞集合とともに獲得できる (Takeuchi and Takahashi (2009))．半教師あり同時クラスタリングはクラスタリングの際，制約としてクラスタに残したい語を指定して，なるべく最終クラスタに残るようにする方法である．詳細は文献 (竹内・高橋 (2010)) に譲るとして，ポイントだけを説明する．

同時クラスタリングの入力は動詞と係り関係にある名詞 (助詞付き) の 2 部グラフである．これはテキストコーパスを係り受け解析した格フレームから獲得する．つまり，動詞の類語を獲得することは，この 2 部グラフの中から緊密な係り関係にある動詞集合と名詞集合 (部分グラフ) を獲得するタスクとなる．部分グラフが緊密かどうかは情報量を利用して判定する．具体的には式 (1) が正であれば有効なクラスタとする．

式 (1) はクラスタ候補 (S_T, S_D) の情報量 (右辺の第 1 項) についてクラスタにしない場合の情報量 (右辺の第 2 項) との差を計算している．t_i と d_j はそれぞれ初期クラスタ内の動詞と名詞 (助詞付き) の要素を表す[2]．

$$\delta I(S_T, S_D) = P(S_T, S_D) \log \frac{P(S_T, S_D)}{P(S_T)P(S_D)} - \sum_{t_i \in S_T} \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (1)$$

[1] koichi@cl.cs.okayama-u.ac.jp
[2] 式の導出については Aizawa (2002) を参照．

問題は，式 (1) が正になる候補を 2 部グラフからいかに探索するかである．2 部グラフの全部分グラフに対して式 (1) を計算するのは時間がかかりすぎるので，初期クラスタから寄与度の低いノードを式 (2) と式 (3) で評価して削除する．前者が動詞ノード (t_i)，後者が名詞ノード (助詞付き)(d_j) に対する評価式である．値が低いノードほど対象とするクラスタに対する価値が低いと考えられるため，削除の対象とする．

$$\delta I(t_i, S_D) = \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (2)$$

$$\delta I(S_T, d_j) = \sum_{t_i \in S_T} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (3)$$

つまり得られるクラスタはこの削除の順序に応じて変わることから，例えば複数の動詞が同じ分類に所属することがわかっている場合，そうした動詞を削除せずに残すような形で初期クラスタを収束させれば既知の情報をクラスタ生成に利用できる．また，本手法は同時クラスタリングであるので，動詞だけでなく，名詞も同様に制約として入れることができる．そこでこうした残すべきノードに対しては式 (2) と式 (3) の値を大きくすることで収束の際，クラスタ候補に指定ノードが残るようにする．これにより半教師あり同時クラスタリングを行うことができる[3]．

本稿では，動詞集合と名詞集合の両方を初期制約して緩い形で与え，収束させる際の条件を変えることで，どの程度良いクラスタが獲得できるかを明らかにする．既知の動詞分類として動詞項構造シソーラス (Takeuchi et al. (2010)) を利用する．動詞項構造シソーラスには最大 5 階層の動詞分類と動詞要素が登録されており，動詞と共起する名詞は例文として記述されている．ただし，例文は各語義で 1 文しかないため名詞の事例が不足する．そこで分類語彙表を利用して動詞と共起する名詞と同じ概念 (もっとも深い分類を利用) に属する名詞を拡張として利用した．

3 クラスタ収束実験

本類義語獲得実験では，収束条件を変えることでどの程度動詞類義語が獲得できるかを評価する．まず入力データの構築について述べ，次に評価方法について述べる．最後に実験の設定と結果を示す．

3.1 入力データ

入力コーパスとして BCCWJ2008 の白書を対象として CaboCha[4]で係り受け解析して得られた格フレームから，動詞と係り関係にある名詞と助詞を 2 部グラフとして登録する．これを入力データとする．

3.2 評価方法

クラスタの精度を評価するために人手で構築されている動詞の分類辞書である動詞項構造シソーラスと比較を行う．通常のクラスタであればどの分類に出力されたか要素数を見て判定する必要があるが，制約付きの場合，各クラスタの目標となる分類は制約を与えた動詞分類であるのでどの分類に属すべきか指定できる．さらに，制約付き CCRE では同じ制約でも複数のクラスタを出力するが，これらはすべて目標となる分類にマージしてクラスタの評価をすれば良い．

そこで評価としてはまずクラスタ内の正解要素数を数える場合下記のようにする．

(1) 出力クラスタを与えた制約に基づく動詞分類ごにマージする．

(2) マージしたクラスタの中で，目標となる動詞項構造シソーラスの分類と比較して正解要素数を数える．ただし，この時，2 要素以上のみを正解とする．

[3]もっと強い制約として残すべき要素を消さないという指定もできるが予備実験でかなり精度が悪くなったため選択しなかった．この理由としては指定した動詞の語義に対応する例文がコーパスに無い場合，簡単に収束が破綻してしまい，有効なクラスタが得られなかったと考えられる．

[4]http://chasen.org/∼taku/software/cabocha/.

評価方法としてはクラスタ要素の適合率と再現率を用いる．適合率は出力したクラスタの要素数をベースに数える．一方で，再現率はすでにある動詞分類の要素のうちいくつ出力したかであるので，動詞の種類数である．それぞれの式は下記の通りである．

$$(\text{適合率}) = \frac{\text{2 要素以上で最大で辞書動詞グループに属する要素数の総和 (延べ数)}}{\text{全出力クラスタの要素数 (延べ数)}} \quad (4)$$

$$(\text{再現率}) = \frac{\text{2 要素以上で最大で辞書動詞グループに属する要素数の総和 (種類数)}}{\text{(辞書の要素数 − コーパスに未出現の動詞数)(種類数)}} \quad (5)$$

3.3 類義語獲得実験結果と収束条件の考察

クラスタ収束条件を与えた制約に依存して変更した場合の類義語獲得実験を行う．具体的には初期クラスタから要素を削除していく際，コーパスに例文が無い場合，制約として指定している要素も削除されていく．そこで収束条件として，(A) 制約要素がいくつまで残っていれば良いかと (B) 制約以外の要素がいくつ残れば良いか，の 2 点に整理する．これらの組合せで類義語獲得実験を行った．結果を表 1 に示す．表 1 では例えば 1 行目では，与えた制約のうち動詞 2 語以上，名詞 2 語以上が

表 1: 収束の条件を変えた場合の適合率と再現率 (BCCWJ2008 白書)

(A) 制約の要素数 (以上)		(B) 他要素数 (以上)		適合率	再現率
動詞	名詞	動詞	名詞		
2	2	3	3	0.142(2097/14718)	0.258(2097/8120)
2	1	3	3	0.083(3302/39759)	0.406(3300/8120)
2	0	3	3	0.027(4176/152384)	0.514(4170/8120)
2	2	1	1	0.114(4278/37621)	0.526(4275/8120)
2	1	1	1	0.062(5246/84383)	0.645(5239/8120)
2	0	1	1	0.013(5845/446869)	0.719(5836/8120)
1	2	3	3	0.071(2640/37284)	0.325(2640/8120)
1	1	3	3	0.036(3833/105984)	0.472(3830/8120)
1	0	3	3	0.010(4636/443055)	0.570(4630/8120)
1	2	1	1	0.069(4669/67810)	0.575(4666/8120)
1	1	1	1	0.034(5560/165651)	0.684(5552/8120)
1	0	1	1	0.008(6056/778959)	0.745(6047/8120)

クラスタ内に残ることを条件とし (収束条件 (A) に対応)，さらに，制約として与えていない他の要素として動詞と名詞がともに 3 語以上 (収束条件 (B) に対応) であることを示している．この場合に獲得できたクラスタの適合率と再現率を右側に示している．

表 1 からまず名詞の制約をいれると適合率が上昇する一方で，再現率が下降することがわかる．また適合率は動詞，名詞それぞれの制約 (A) を少なくすると約半減している．名詞の制約を入れない場合では適合率は 1/3 から 1/4 程度に下がってしまう．また制約 (B) の影響を見ると，3 個の要素を残す場合と 1 個の要素を残す場合では適合率は約 2 割減に留まっている一方で，再現率は約 1.4 から 1.5 倍程度向上する．よって制約 (B) では 1 個の要素を残す方法が有利であることがわかる．

上記の結果から制約は動詞だけでなく，名詞もいれることが有効であること，収束条件として制約の対象ではない要素が動詞，名詞共に各 1 個残るようにした場合，適合率および再現率がバランス良く向上することがわかる．

4 WKK 法との比較

Weighted kernel k-means 法はベクトルによる入力データを kernel 空間で重み付きの k-means として分類する枠組みである．興味深いのはが WKK 法とグラフを cut することによりクラスタリン

グするグラフベースのクラスタリングと数式的に等価であることを証明している点である (Dhillon et al. (2007))．よって WKK 法は重み付きグラフを入力とすることができる．既に竹内他 (2010) で動詞類語獲得方法について提案したが，利用したソフトウェア (graclus1.2) の乱数初期化が不完全で正しい数値が得られていなかった．そこで，不備を修正し 5 回平均を行った上で動詞類語獲得を行い，制約無しの場合の CCRE と比較する．

対象とするコーパスは BCCWJ2008 白書で，クラスタ数 K の値は 50 から 4000 まで 100 ステップで変えてもっとも高い値を選ぶ．評価は同様に適合率と再現率で調和平均による値で K の最高値を選んだ．この結果を表 2 に示す．

表 2: CCRE と WKK の適合率と再現率 (BCCWJ2008 白書)

	適合率	再現率
CCRE	0.324(2042/6295)	0.118(960/8120)
WKK(K=300)	0.139(615.6/4462)	0.076(614.9/8120)

表 2 中で WKK 法の分子が小数点なのは複数回の平均精度であるためである．表から CCRE が適合率だけでなく，再現率も高いことがわかる．しかしながら，現段階では WKK はハードクラスタリングであるため多義性の高い動詞類語抽出には弱いが，WKK も制約を入れることが可能であることからソフトクラスタリングを導入できる可能性があり，同じ枠組みで評価できる可能性がある．今後，こうした比較も行っていきたい．

5 まとめ

動詞辞書構築支援として，動詞の類語を獲得する手法を発展させてきた．本研究では同時共起クラスタリングを複数回行う手法を提案し，大規模テキストデータから動詞集合と名詞集合の対を精度よく獲得できることを示した．本稿では制約付きクラスタリングについての収束条件の最適化について検討を行い，再現率が高く適合率を大きく損なわない設定があることを明らかにした．また，新たな同時クラスタリングの手法の一つである WKK との比較もすすめ，現在の精度をを比較した．今後，動詞語義だけでなく，例文獲得や意味役割を付与したデータの構築を進める予定である．

文献

Akiko Aizawa (2002) “A method of Cluster-Based Indexing of Textual Data,” in *Proceedings of COLING 2002*, pp. 1–7.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis (2007) “Weighted Graph Cuts without Eigenvectors: A Multilevel Approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 11, pp. 1944–1957.

Koichi Takeuchi and Hideyuki Takahashi (2009) “Co-clustering with Recursive Elimination for Verb Synonym Extraction from Large Text Corpus,” *IEICE Transactions on Information and Systems*, Vol. E92-D, pp. 2334–2340.

Koichi Takeuchi, Kentaro Inui, Nao Takeuchi, and Atsushi Fujita (2010) “A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings,” in *The 8th Workshop on Asian Language Resources*, pp. 1–8.

竹内孔一、高橋秀幸 (2010)「同時共起クラスタリングを利用した動詞辞書構築」，日本語平成 *21* 年度公開ワークショップ *2010* 年 *3* 月 *15-16* 日．

竹内孔一、高橋秀幸、小林大介 (2010)「グラフに基づくクラスタリングによる動詞類義語の獲得」，電子情報通信学会，言語理解とコミュニケーション研究会, NLC-2010-11，pp.13–18．

分類器の確信度を用いた合議制による語義曖昧性解消の領域適応

古宮嘉那子（言語処理班協力者：東京農工大学　工学研究院）†
奥村学（言語処理班班長：東京工業大学　精密工学研究所）

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

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

１．はじめに

通常，機械学習とは，新聞データを用いて新聞用の分類器を学習するなど，ドメイン A のデータを用いてドメイン A 用の分類器を学習するものであった．しかし一方，ドメイン B についての分類器を学習したいのに，ドメイン A のデータにしかラベルがついていないことがあり得る．このとき，ドメイン A（ソースドメイン）のデータによって分類器を学習し，ドメイン B（ターゲットドメイン）のデータに適応することを考える．これが領域適応であり，さまざまな手法が研究されている．図 1 はソースドメインを新聞，ターゲットドメインを小説にした際の領域適応の様子を示している．

語義曖昧性解消（WSD: Word Sense Disambiguation）の領域適応の手法はさまざまあるが，我々は用例によって適切な手法は異なると考えた．本稿では，少量のターゲットデータにラベル付けして学習を行う方式と，他のコーパスを訓練事例に加える方式を使って二つの分類器を学習し，学習された分類器の出力する確信度の高い方の答えを採用することにより，分類の精度を向上させる手法を示す．

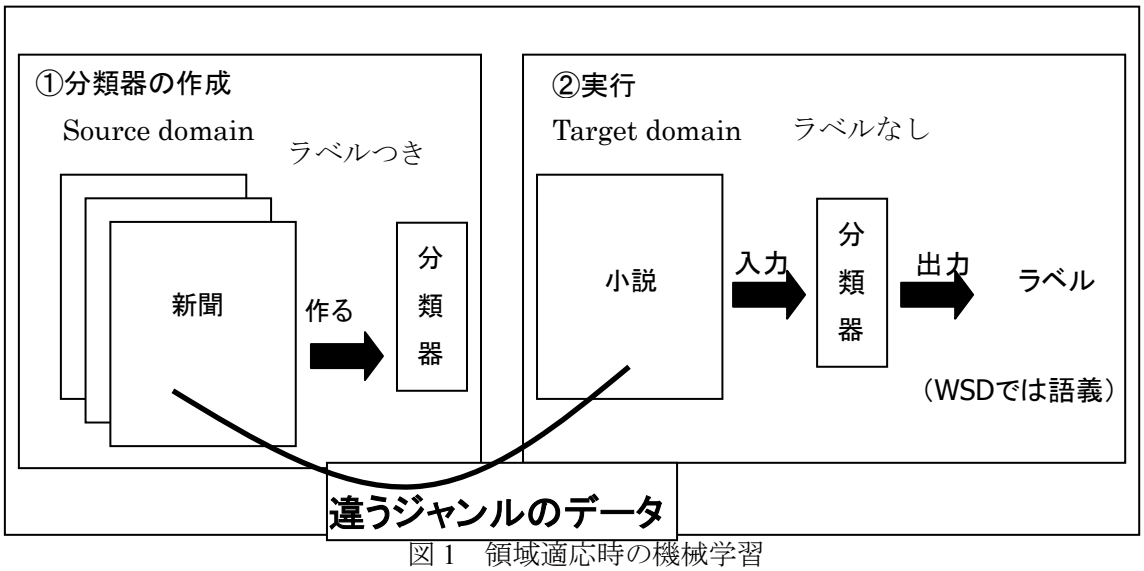


図 1　領域適応時の機械学習

２．関連研究

領域適応は，学習に使用する情報により，supervised，semi-supervised，unsupervised の三種に分けられる．まず supervised の領域適応は，多量なラベルつきのソースデータに加え，

†kkomiya@cc.tuat.ac.jp

量のラベルつきのターゲットデータを用いて学習を行うもので，訓練事例としてソースデータまたは少量のターゲットデータだけを利用する場合よりも，分類器を改良することを目指す．次の semi-supervised の領域適応は，多量なラベルつきのソースデータに加え，多量なラベルなしのターゲットデータを利用し，訓練事例としてソースデータだけを利用する場合よりも，分類器を改良することを目指す．また，最後の unsupervised の領域適応は，ラベルつきのソースデータで学習後，ターゲットデータで実行する．本研究で扱うのは，supervised の領域適応である．

領域適応の研究は自然言語処理の分野の内外においてさまざまなされており，supervised のものには（Chan and Ng (2006)），（Daumé III(2007)），（Jiang and Zhai (2007)）などがある．また，共学習を用いた適応に関する研究に（Tur (2009)）がある．

（Tur (2009)）は co-training において適応を行った co-adaptation の研究である．boosting による線形補完により適応を行い，両方の分類器においてエラー率が低下したことを報告している．

本稿では，分類器の確信度により領域適応に用いる手法を選択する手法について述べる．これに関連した研究として（張本ら (2010)）や（Asch and Daelemans (2010)），（古宮，奥村 (2010)）がある．（張本ら (2010)）は，構文解析において，分野間距離をはかり，より適切なコーパスを利用して領域適応を行えるようにした．また，（Asch and Daelemans (2010)）は，構文解析において，自動的にタグ付けされたコーパスを用いて，ソースデータとターゲットデータの類似度から性能を予測できることを示した．これらの研究では，領域間の距離からソースデータとして利用できるコーパスを選択するという立場をとっているが，（古宮，奥村 (2010)）はソースデータとターゲットデータの性質から，適切な領域適応手法を自動選択するという立場をとった．

本研究では，分類器の確信度から，用例ごとに手法を選択する．

３．用例ごとの領域適応手法の自動選択

（古宮，奥村 (2010)） において，我々は WSD のための領域適応において，ターゲットデータやソースデータの性質により，ソースデータ/ターゲットデータ/単語の組み合わせごとに最も効果的な領域適応手法が異なることを示した．

本稿では，ソースデータ/ターゲットデータ/単語の組み合わせだけではなく，一例一例，用例ごとに効果的な手法が異なると仮定する．そのため，以下のように用例ごとに領域適応の手法を選択する．

(1) 複数の手法により分類器を学習する．
(2) 用例ごとに，複数の手法による分類器の確信度を比較する．
(3) 分類器の確信度の最も高い手法による結果を採用する．

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

４．実験

４.１WSD のための領域適応手法

WSD のための領域適応手法として，本研究では以下に示す二つ(Target Only，Random Sampling) を用いる．

Random sampling：ランダムに選んだ少量のターゲットデータの用例にラベル付けしたものとソースデータの両方を訓練事例にする．

Target Only：ソースデータを用いず，ランダムに選んだ少量のターゲットデータにラベル付けしたものだけを訓練事例にする．

なお，使用するターゲットデータは常に 10 件とした．分類器としてはマルチクラス対応の SVM（libsvm）(Chang and Lin (2001))を使用した．また，libsvm の確率として出力される分類の確からしさを確信度として用いた．本実験では，分類器を二つ学習したため，合議の際には二つのうちより高い確信度である分類器の結果を採用する．カーネルは予備実験の結果，線形カーネルが最も高い正解率を示したため，これを採用した．また，学習の素性には，以下の 17 素性を用いた．

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

分類語彙表の分類コードには（国立国語研究所 (1964)）を使用した．

また，実験は五分割交差検定を用いた．Random Sampling の場合には，ソースデータの 4/5（ソースデータの黒い部分）に加え，ターゲットデータの 4/5（ターゲットデータの白の部分と斜線の部分）から 10 件（白い部分）を訓練事例とする．テストデータは，ターゲットデータの残りの 1/5（灰色の部分）である．この様子を図 2 に示す．

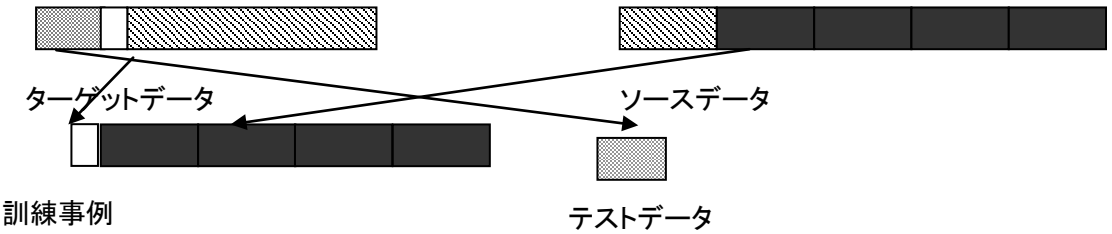


図 2　領域適応の五分割交差検定

４.２実験データ

実験には，現代日本語書き言葉均衡コーパス（BCCWJ コーパス）(Kikuo Maekawa (2008)) の白書のデータとYahoo! 知恵袋のデータ，またRWC コーパスの毎日新聞コーパス（Hashida et al. (1998)）の三つのデータを利用し，ひとつの単語につきソースデータとターゲットデータを変えることで，全部で6通りの領域適応を行った．これらのデータには岩波国語辞典（西尾ら (1994)）の語義が付与されている．これらのコーパス中の多義語のうち，ソースデータおよびターゲットデータ中に存在する用例がともに50 用例以上の単語を実験対象とした．単語の異なり数は，白書⇔ Yahoo!知恵袋 : 24　白書⇔新聞 : 22　Yahoo! 知恵袋⇔新聞 : 26 であり，全体で28 単語となった．それぞれの領域における単語ごとの最小，最大，平均用例数を表1 に示す．

また，実験には岩波国語辞典の小分類の語義を採用した．語義数ごとの単語の内訳は，2 語義 :「場合」,「自分」，3 語義 :「事業」,「情報」,「地方」,「社会」,「思う」,「子供」，4 語義 :「分かる」,「考える」，5 語義 :「含む」,「使う」,「技術」，6 語義 :「関係」,「時間」,「一般」,「現在」,「作る」，7 語義 :「今」，8 語義 :「前」，10 語義 :「持つ」，11 語義 :「進む」，12 語義 :「見る」，14 語義 :「入る」，16 語義 :「言う」，21 語義 :「出す」，22 語義 :「手」,「出る」である．

表 1　それぞれの領域における単語ごとの最小，最大，平均用例数

コーパスの種類	最小	最多	平均
BCCWJ 白書	58	7610	2074.5
BCCWJ Yahoo！知恵袋	82	13976	2300.43
RWC 新聞	50	374	164.46

５．結果

表 2 に全体の適応手法別の実験結果を示す．また，表 3 にコーパスと適応手法別の実験結果を示す．

表 2　全体の適応手法別の実験結果

領域適応手法	正解率
Random Sampling	79.85%
Target Only	79.66%
確信度による合議	**83.49%**

これらの表で，コーパスごとに一番高い正解率を太字で示した．またその値を二番目に高い正解率と比較した際，0.05 水準で有意である場合にはその値に下線を引いた．

表 3　コーパスと適応手法別の実験結果

ソースデータ	Yahoo！知恵袋	Yahoo！知恵袋	白書	白書	新聞	新聞
ターゲットデータ	白書	新聞	Yahoo！知恵袋	新聞	Yahoo！知恵袋	白書
領域適応手法	正解率					
Random Sampling	87.21%	**73.95%**	**83.97%**	72.09%	76.61%	72.66%
Target Only	88.35%	66.46%	75.74%	67.75%	74.46%	84.57%
確信度による合議	**88.54%**	72.80%	83.03%	**72.48%**	**78.10%**	**87.81%**

６．考察

表 3 から，Yahoo！知恵袋をソースデータとして新聞をターゲットデータとした領域適応と，白書をソースデーとしてタ Yahoo！知恵袋をターゲットデータとした領域適応を除いた 4 方向の領域適応において，提案手法である分類器の確信度を用いた合議が最も高い正解率を示すことが分かる．また，表 2 から，全ての方向の領域適応の平均をとった場合には，提案手法である分類器の確信度を用いた合議が最も高い正解率を示し，その値は二番目に高い正解率を示した Random Sampling の結果と比べて有意差が認められたことが分かる．これらのことから，本手法はどのようなコーパスの組み合わせに対しても有効であるわけではないが，一般的に有効な手法であると言えるだろう．

本稿では，Target Only と Random Sampling の二つの手法だけを比較し，この二つのうちより確信度の高い手法による分類器の分類結果を採用した．比較対象の分類手法が変わったとき，また増えた場合の提案手法の有効性の検証は今後の課題である．

７．まとめ

分類のターゲットとなるドメインとは異なるドメインのデータを利用して分類器をつくり，ターゲットドメインのデータに適応することを領域適応といい，近年さまざまな手法が研究されている．語義曖昧性解消（WSD: Word Sense Disambiguation）の領域適応の手法はさまざまあるが，我々は用例によって適切な手法は異なると考えた．本稿では，少量のターゲットデータにラベル付けして学習を行う方式と，他のコーパスを訓練事例に加える方式を使って二つの分類器を学習し，学習された分類器の出力する確信度の高い方の答えを採用することにより，分類の精度を向上させる手法を示した．自動的に選択された手法を用いて領域適応を行うことで，もともとの手法を一括的に使った時に比べ，WSD の平均正解率が有意に向上した．

文献

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

Yee Seng Chan and Hwee Tou Ng (2006). "Estimating Class Priors in Domain Adaptation for Word Sense

Disambiguation." *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp89-96.

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

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

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

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

Gokhan Tur (2009). “Co-adaptation: Adaptive Co-training for Semi-supervised Learning”. In *Proceedings of the IEEE International Conferenceon Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, pp. 3721{3724.

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

古宮嘉那子, 奥村学 (2010). “語義曖昧性解消のための領域適応手法の自動選択”. 情報処理学会研究報 Vol. 2010-NL-198, No. 5, pp. 1{6.

国立国語研究所 (1964). “分類語彙表”. 秀英出版.

西尾実, 岩淵悦太郎, 水谷静夫 (1994). “岩波国語辞典第五版”. 岩波書店.

共起語グラフのクラスタリングによる単語の多義性抽出

鏑木　雄太（言語処理班協力者：東京農工大学工学部情報工学科)†
古宮嘉那子（言語処理班協力者：東京農工大学工学研究院先端情報科学部門）
小谷　善行（言語処理班協力者：東京農工大学工学研究院先端情報科学部門）

Extraction of the Ambiguity of Words Based on the Clustering of Co-occurrence Graphs

Yuta Kaburagi (Depertment of Computer and Information Sciecnces,Tokyo University of Agriculture and Technology)
Kanako Komiya (Institute of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institute of Engineering, Tokyo University of Agriculture and Technology)

1, はじめに

自然言語処理では、多義語が文中に出現したとき、それがどのような意味で使われたのかを推定する、語義曖昧性解消（Word Sense Disambiguation）の研究が盛んに行われてきた。しかし、語義曖昧性解消には、辞書に定義されていない語義（新語義）は正しく判定できないという問題がある。この事態を受け、辞書にない語義を抽出する新語義発見（Word Sense Induction）の研究が数多く行われてきた。

本研究は、新語義発見に加え、これまで辞書にはなかったような未知語に対しても多義性を発見するため、多義性の抽出を行う。多くの研究によって、単語の語義を判別するために共起情報が有効であったという知見から、ある単語の多義性と語義の発見および推定に、単語の共起情報を用いる。また、グラフクラスタリングを用いることで、その多義性を抽出した。以下 2 章では関連研究について、3 章では共起語から多義性を抽出する方法について、4 章では多義性を抽出する具体的処理について、5 章では実験の概要について、6 章で実験結果に対して評価と考察を行う。

2. 関連研究

グラフ構造を用いた自然言語処理の研究は多く行われている。グラフ構造を用いている代表的なものに概念辞書である WordNet[2] がある。WordNet を辞書としたグラフベースの語義曖昧性解消に関する研究[3] も行われている。また、大規模なテキストコーパスから名詞共起情報を用いて多義性を抽出する研究[5]も行われている。本研究はこれに近いものであるが、名詞以外の品詞の共起情報を用いている点、既知の語義を共起した単語群を基に人手で列挙している点が異なる。

3. 共起語から多義性を抽出する方法

本研究は、「多義語は語義ごとに文中で共起しやすい単語が異なる」という考えに基づいている。例として「ジャケット」という多義語の場合を考える。「ジャケット」は、「上着の一種」と「レコード・本などを包む覆い・装丁」という語義を持っている。「ジャケット」が「上着の一種」の語義で出現したテキストでは「着る」や「洋服」という単語やこれら

†50006268014@st.tuat.ac.jp

共起語と関連する単語が出現する。「ジャケット」が「レコード・本などを包む覆い・装丁」の語義で出現したテキストでは「CD」や「レーベル」という単語やこれら共起語と関連する単語が出現する。共起関係を調べていくと、「ジャケット」の共起語の中では、＜服飾＞に関連する単語群と＜装丁＞に関連する単語群がそれぞれの単語群内で多く共起していることが予想される。単語群を共起関係によってまとめると、図 1 のようなイメージになる。

「ジャケット」と「＜服飾＞関連語」及び「＜装丁＞関連語」が多く共起していることから、「ジャケット」には「服飾に関係がある語義」と「装丁に関係がある語義」の少なくとも二つの語義があると推定できる。共起語同士の共起関係を用いて、語義に対応する単語の集合を自動生成することで、単語の多義性を発見できると考えた。手法としては、共起関係の表現にグラフ構造を用い、単語集合の生成にグラフクラスタリングを用いる。

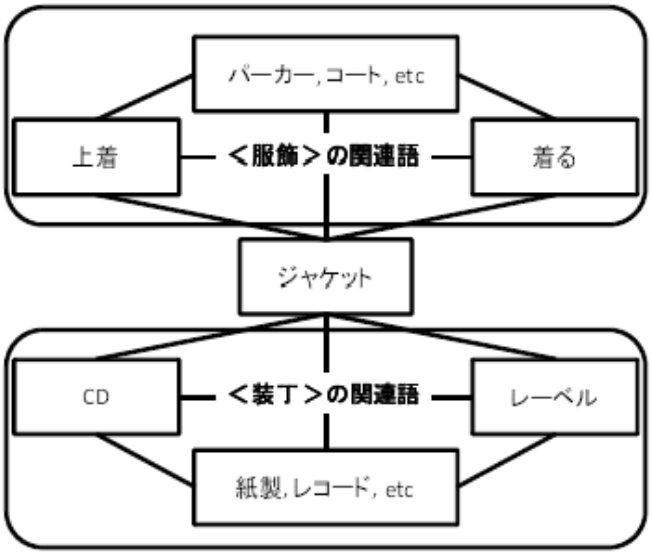


図 1: 「ジャケット」を中心とした共起語の集合

4. 共起語グラフのクラスタリングシステムの実現

クラスタリングシステムは以下の 3 ステップの処理によって実現される。

1. 共起語グラフの生成
2. クラスタリングによる多義性抽出
3. クラスタリング結果の単語群の出力

以下、節ごとにこれらの説明を行う。

4.1 共起語グラフ生成の具体的処理

4.1.1 共起語の選択方法

本研究において共起とは、「同一の文中に出現すること」と定義する。任意の二単語が一文で共起したとき、その二単語は一回共起したと数える。共起の対象とする単語は、名詞（形容動詞、サ変動詞を含む）、動詞、形容詞の各自立語とした。既存の研究[5] では、語彙統語パターンを用いて並列関係にある名詞を対象としていたが、名詞に対して動詞や形容詞が語義クラスタリングや語義推定の手助けになると考えた。名詞、動詞、形容詞のみ

を抽出するために、事前に形態素解析を行ったコーパスを品詞情報を元にフィルタリングし、表記ゆれを考慮して対象単語の用言はすべて原形に変換した。

4.1.2 共起語グラフの生成方法

共起関係をコーパス全てにおいて調べ、その情報を基にグラフを生成する。以降この共起関係を表現したグラフのことを共起語グラフと定義する。共起語グラフにおいては、1種類の単語は1つのノードによって表現され、共起関係はエッジによって表現される。ノードには単語の出現回数、エッジには両端の単語（ノード）の共起回数を保存する。例として、「太郎がりんごを食べた。」という文が現れたとき、図 2 のような共起語グラフを生成する。

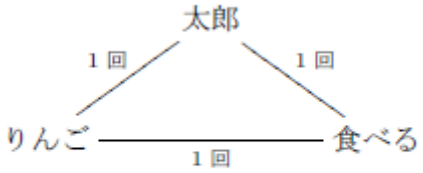


図 2: 共起語グラフの例

4.2 クラスタリングによる多義性抽出

4.2.1 対象とする多義語を中心とした共起語による部分グラフの生成

多義性を抽出したい単語を一つ選び、その単語と直接共起した単語を全て列挙する。列挙した単語に対応するノードとノード間を結ぶエッジにより、グラフの一部分を抽出する。このグラフを以降では部分グラフと呼ぶ。この部分グラフには、ターゲット単語に対応するノードとそれに繋がっているエッジは含まない。結果、図 3 のようなグラフとなる。図 3 において実線は部分グラフに含めるエッジ、点線は共起語グラフに存在するが部分グラフに含めないエッジである。以降の処理はこの部分グラフを対象として行う。

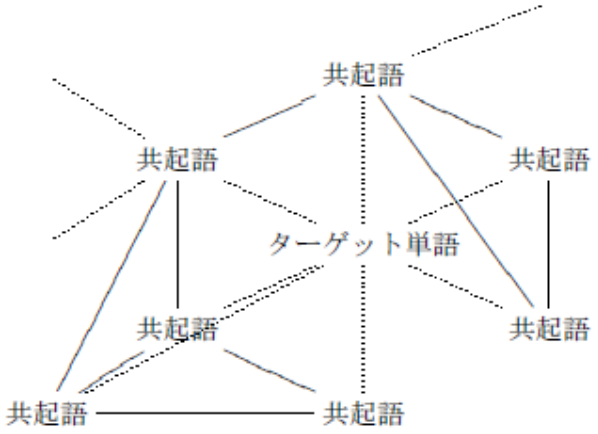


図 3: 部分グラフのイメージ

4.2.2 部分グラフエッジの重み計算方法

グラフクラスタリングを行うにあたっての部分グラフ内のエッジの重みを設定する。重みとして単語の共起回数から求めた遷移確率を用いると、コーパス内で出現確率が高い単

語に向かう遷移確率が大きくなる。その結果、クラスタリング中に複数のクラスタが頻出共起語を中心に接続され、一つのクラスタとなってしまう恐れがある。本手法では、頻出共起語の影響を抑えるためにエッジの重みには自己相互情報量を用いる。自己相互情報量 I(x, y) は式(1) で表わされる。

$$I(x,y) = \log \frac{P(x,y)P(*,*)}{P(x,*)P(*,y)} \tag{1}$$

式(1) において、P(x; y) は、グラフにおける単語 x, y の共起回数、* はグラフに存在するすべての単語を意味する。本研究では、共起グラフ全体における自己相互情報量と部分グラフにおける自己相互情報量の両方を用いる。部分グラフにおける自己相互情報量 I(x, y|part) は式(2) のように定義する。

$$I(x,y|part) = \log \frac{P(x,y|part)P(*,*|part)}{P(x,*|part)P(*,y|part)} \tag{2}$$

式(2) における part は、部分グラフに含まれることを指す。式(1) と式(2) の積を確率として正規化した値を部分グラフのエッジの重みとして用いる。ただし、式(1) と式(2) いずれかの式の値が負となった場合は、重みを 0 とした。

4.2.3 グラフクラスタリングアルゴリズム

部分グラフに適用するグラフクラスタリング手法として。マルコフクラスタリングアルゴリズム[4] を用いた。マルコフクラスタリングアルゴリズムは、グラフエッジの重みを遷移確率としてグラフ内をランダムウォークすることでクラスタリングを行う手法である。グラフの各ノードに自己ループを追加したものグラフにおける遷移確率行列 M に inflation と expansion を繰り返すことでクラスタリングを再現することができる。inflation と expansion の各定義式を次に示す。

expansion

$$M = M^2 \tag{3}$$

inflation

$$M = \Gamma_r(M) \tag{4}$$

$$\Gamma_r(M)_{pq} = (M_{pq})^r / \sum_{i=1}^{k} {M_{iq}}^r \tag{5}$$

両式における"="は代入、M_{pq} は行列 M の要素(p, q)、r は inflation パラメータ(r > 1) である。式(4) における行列 $\Gamma_r(M)$ の各要素は式(5) によって計算される。inflation を繰り返すことで、相対的に低い遷移確率はより低く、相対的に高い遷移確率はより高く更新されていく。行列遷移確率行列 M が収束するまで、inflation と expansion を繰り返し計算する。収束した行列は、いくつかの小さなグラフの遷移確率行列が一つの大きな行列内に表現されている。

4.3 クラスタリング結果の単語群の出力

クラスタリングの結果生成されたクラスタを表すグラフの遷移確率行列を基にクラスタごとに単語群を出力する。計算結果の遷移確率行列をグラフにすると例えば図 4 のようなグラフとなる。

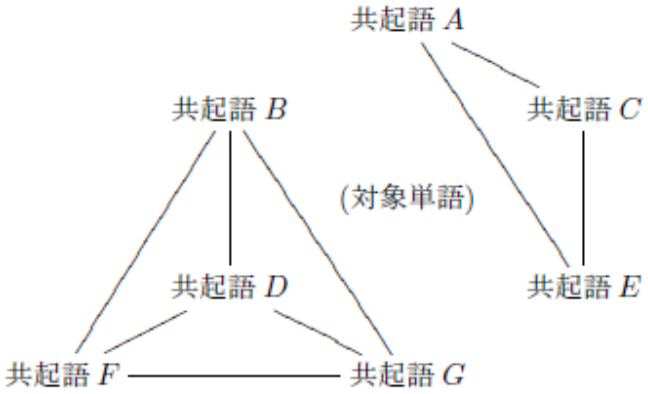


図 4: クラスタリング結果の部分グラフのイメージ

分割されたグラフに一つに含まれている単語群を一つのクラスタとする。図 4 を例にすると、単語集合は{ 共起語 A, 共起語 C, 共起語 E } と{ 共起語 B, 共起語 D, 共起語 F, 共起語 G } の二つの集合が生成される。以降では、この単語集合を語義クラスタと呼ぶ。

5. 多義語の抽出実験

5.1 共通の実験環境

言語資源はBCCWJ[1] のYahoo! 知恵袋コーパスを用いた。Yahoo! 知恵袋コーパスは、Yahoo! 知恵袋 [1]に投稿された質問とベストアンサーの組 1500 件のからなるコーパスである。今回行った二つの実験では、形態素解析によって得られている品詞情報を用いたが、語義データは用いなかった。マルコフクラスタリングアルゴリズムにおけるinflation パラメータr は 1:25 とした。部分グラフは、共起語間の正規化前の重みが 5 以下のエッジを切断し、対象単語との自己相互情報量が 3 以上の単語を含む部分グラフを用いた。

部分グラフ A 及び B のクラスタリングの結果生成した語義クラスタに含まれる単語に対しては、対象単語との自己相互情報量が 5 以上の単語のみを出力し、語義クラスタに現れる単語の差をできるだけ少なくした。

5.2 辞書を用いない多義性抽出実験の概要

辞書を用いない多義性抽出実験について述べる。コーパスから生成した共起語グラフに対してクラスタリングを行い、多義性抽出を試みる。多義性抽出対象単語として、コーパス内に 50 回以上出現する単語のうち、岩波国語辞典で複数の語義が定義されている 10 種類の名詞を選んだ。

[1] http://chiebukuro.yahoo.co.jp/

表 1：対象単語の一覧

アルバム，クラス，ケース，コーナー，ジャケット，ソース，ノート，バイト，ボール，レース

5.3 辞書を用いた多義性抽出実験の概要

辞書を用いない多義性抽出では、既知の語義であるかどうかを判断することが難しい。そこで、語義クラスタと語義を同定し、新語義を推定する手がかりとして、辞書の定義文と用例文を用いる。辞書定義文と用例文は岩波国語辞典第五版[6] を用いた。辞書の定義文と用例文の定義文内対象語と共起関係を共起語グラフに追加することで、辞書内単語が語義クラスタの生成にどの程度影響を与えるかを検証する。多義性抽出対象単語として、コーパス内に 50 回以上出現する単語のうち、岩波国語辞典で複数の語義が定義されている 4 種類の名詞を選んだ。

表 2：対象単語の一覧

ジャケット，ソース，ノート，レース

6. 評価と考察

6.1 評価方法

各多義語に対してシステムを適用した結果得られた語義クラスタに、人手で正解と不正解のラベルを付ける。正解かどうかの判断は、次のように定める。

- 各語義クラスタ内の各単語に対して一つずつ人手でタグづけをする。タグづけ方法は次のように定義した。
  - 岩波国語辞典第五版[6] を参考に対象多義語の語義を列挙する。
  - 「一つの単語が多義語の語義一つに対応する」場合、単語にその「語義に対応するタグを付ける。タグの種類数は辞書で定義されている語義の数と同じである。
  - 「一つの単語が複数の語義と対応しうる」または「多義語と単語間の意味関係が不明である」場合、「語義なしタグ」を付ける。
- 語義クラスタ内単語のうち 80 %以上に同じ「語義に対応するタグ」が付けられた場合、クラスタは「一語義を示す」とし、その語義に対応する正解クラスタとする。複数のクラスタが同じ語義の正解クラスタになることもできる。
- 同じ「語義に対応するタグ」の割合が 80 %を下回った場合は、クラスタは「複数語義を示す」とする。

このように付けた正解を基にクラスタリング結果を評価する。評価には次の尺度を用いる。

$$\text{適合率 (1)} = \frac{\text{「一語義を示す」語義クラスタ数}}{\text{生成した全ての語義クラスタ数}} \tag{6}$$

$$\text{適合率 (2)} = \frac{\text{一つ以上の対応するクラスタが存在した語義数}}{\text{生成した全ての語義クラスタ数}} \tag{7}$$

$$\text{再現率} = \frac{\text{一つ以上の対応するクラスタが存在した語義数}}{\text{岩波国語辞典第五版で定義されている語義数}} \tag{8}$$

$$F\text{ 値} = \frac{2 \times \text{適合率 (2)} \times \text{再現率}}{\text{適合率 (2)} + \text{再現率}} \tag{9}$$

ベースラインの値には、自己相互情報量が閾値 5 以下のエッジを切断したグラフにおけ

る語義クラスタを用いた。

6.2 辞書を用いない多義性抽出結果

実験の結果、各適合率、再現率、F 値は表 3 のようになった。

表 3：辞書を用いない多義性抽出の実験結果

手法	適合率(1)	適合率(2)	再現率	F 値
提案手法	0.11	0.04	0.42	0.07
エッジカットのみ	0.15	0.06	0.45	0.10

辞書を用いない多義性抽出実験の結果得られた語義クラスタのうち、「ソース」のクラスタリング結果の一部を表 4 に示す。

表 4：「ソース」の多義性抽出結果の一部

ID	クラスタ内の単語	推定語義
1	につめる, 濃い茶, 絡ませる, 水溶き, マーボー, タレ, ぬく, はさみ,茄子, 煮詰める, 鉄板, 白っぽい, ぬき, とろみ, もやし, エビ, 味噌,たれ, 黒っぽい, 色っぽい, 緑色, ナス, 本場, サラダ油, 固める, 好み,長ネギ, 胡椒, 片栗粉, インスタント, にんにく, 小麦粉, 香川, ひき肉, カニ, 分量, 寒天, 溶く, アレンジ, 焼肉, 鶏, コショウ	調味料
2	非公開, 秘密, ドラッグ, ウィルス, 完璧	情報源
3	手短, ＡＰ通信, 貝柱, 漂流, 種子島, 刻む, 漁師, アサリ, 沸かす, 沖, 茹でる, 天ぷら, ゆでる, 独特,	(複数語義)

まず、一番目の語義クラスタに注目すると、食べ物や料理に関連する単語が多くクラスタ分けされている。多くの単語に「調味料」に対応するタグが付与されたことで、語義クラスタには「調味料」の語義がついた。

次に、二番目の語義クラスタに注目すると、「非公開」や「秘密」といった「情報」に関連する単語がクラスタ分けされている。含まれている単語の数は一番目のクラスタよりも少ないが、四つの単語に「情報源」に対応するタグが付与されたため、この語義クラスタには「情報源」の語義がついた。

最後に三番目の語義クラスタに注目する。語義クラスタ内の単語のカテゴリにばらつきがある。「ＡＰ通信」は「情報源」に対応すべき単語と考えられ、「茹（ゆ）でる」は、「調味料」に対応すべき単語と考えられる。コーパス上での共起情報を調べてみると、「ＡＰ通信」が「ソース」と共起したとき、同時に「漁師」と共起していた。また、「漁師」は「アサリ」などの魚介類と共起しやすいため、クラスタリングの結果、「漁師」が二つのクラスタを一つのクラスタにまとめる橋渡しをする役目を果たしていた。

6.3 辞書を用いた多義性抽出結果

辞書を用いずに多義性抽出実験を行った結果を示す。実験の結果、各適合率、再現率、F 値、平均クラスタ数は表 5 のようになった。

表 5：辞書を用いた多義性抽出の実験結果

手法	適合率(1)	適合率(2)	再現率	F 値	平均クラスタ数
辞書あり	0.20	0.10	0.58	0.18	13.5
辞書なし	0.05	0.05	0.33	0.08	14.5

辞書の定義文を用いたことで、辞書を用いないクラスタリングに比べて、クラスタリングの結果が向上した。平均クラスタ数が減少したことで、語義に対する過分割が減っていることがわかった。

7. おわりに

本研究では、コーパスの共起関係をクラスタリングすることで単語の多義性抽出を行った。多義語は、語義によって共起しやすい単語が異なることから、共起しやすい単語同士をまとめることで、ある単語に多義性があるかどうかを抽出した。本研究では、コーパスの単語共起情報をグラフ構造を用いて表現し、グラフをクラスタリングするによって共起語のクラスタリングを行った。今回提案した共起語選択手法は、いずれの場合でも語義クラスタに辞書語義をコンピュータに自動推定されることができなかった。事前に辞書の語義情報を極力使わずに単語の多義性を抽出することで新語義の発見が期待できるためであった。

本研究では語義を人手で付与したが、語義クラスタ内の単語のから、＜音楽＞や＜服飾＞といった抽象的な意味カテゴリを自動推定することができれば、より客観的に語義を付与することができると考えられる。また、既知の語義の参考として、岩波国語辞典第五版を用いたが、岩波国語辞典以外の辞書における語義分類も存在する。今回新語義として判定した語義が、他の辞書の分類では既知の語義であるという可能性がある。

謝辞

データを提供していただいた東京工業大学奥村研究室に深く感謝する。

文献

[1] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources* (ALR), pp. 101-102, 2008.

[2] G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM,* Vol. 38, No. 11, pp. 39-41, 1995.

[3] R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation.In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.

[4] Stijn van Dongen. Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000.

[5] 田淵史郎, 鍜治伸裕, 吉永直樹. 大規模コーパスからの語義のマイニング. 日本データベース学会論文誌,Vol. 8, No. 1, pp. 77-82, 2009-06.

[6] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典第五版. 岩波書店, 1994.

教師付き外れ値検出による新語義の発見

新納 浩幸 (言語処理班分担者: 茨城大学 工学部) [1]
佐々木 稔 (言語処理班分担者: 茨城大学 工学部)

Detection of New Word Senses Using Supervised Outlier Detection

Hiroyuki Shinnou (Ibaraki University, Faculty of Engineering)
Minoru Sasaki (Ibaraki University, Faculty of Engineering)

1 はじめに

本論文では対象単語の用例集合から、その単語の語義が新語義（辞書に未記載の語義）となっている用例を検出する手法を提案する。

新語義の発見は語義識別問題に対する訓練データを作成したり、辞書を構築する際に有用である。また新語義の用例はしばしば書き誤りとなっているので、誤り検出としても利用できる。ここでのアプローチの基本は、新語義の用例が用例集合中の外れ値になると考え、データマイニング分野の外れ値検出の手法を利用することである。ただし外れ値検出のタスクは教師なしの枠組みになるが、新語義検出という本タスクの性質を考慮すると、一部のデータ（用例）にラベル（対象単語の語義）が付与されているという枠組みで考える方が適切である。そのため本論文では一部のデータにラベルがついているという教師付きの枠組みで外れ値検出を行う。

提案手法は 2 つの検出手法からなる。第 1 の手法は従来の外れ値検出手法である Local Outlier Factor (LOF)(Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander, 2000) を教師付きの枠組みに拡張したものである。第 2 の手法は、教師データから語義識別の分類器を学習し、各データの語義を推定する。推定された語義のクラスターとデータとの距離関係から外れ値かどうかを判定する。提案手法では第 1 の手法により外れ値の候補を取り出し、第 2 の手法でその候補を選別する。

提案手法の有効性を確認するために、2 つの実験を行った。人工的に作ったデータに対するものと、SemEval-2 の Japanese WSD タスク (Manabu Okumura and Kiyoaki Shirai and Kanako Komiya and Hikaru Yokono, 2010) のデータに対するものである。SemEval-2 の Japanese WSD タスクは通常の語義識別のタスクであるが、識別する語義の対象に新語義を含めている点に大きな特徴がある。このためこのタスクの訓練データを教師データとして利用して、テストデータから新語義を検出するという設定で実験が行える。

2 つの実験を通して、外れ値検出に教師データを利用する効果が確認できた。今後の課題としてはパラメータの設定法がある。本手法ではパラメータが 3 つ存在し、これらの値が結果に大きく影響する。またタスクに応じて適切な値が異なる。このため適切な設定方法が必要である。

2 教師付き外れ値検出

2.1 従来の外れ値検出手法

外れ値検出の手法は多岐にわたるが、おおまかに分類するとデータの生成に確率モデルを用いるものと用いないものに分けられる (山西健司, 2009)。確率モデルを用いた場合、データの生成確率が得られるので、その確率が低いデータを外れ値とすればよい。このアプローチでは、いかに適切な確率モデルを構築できるかが鍵となる。確率モデルを用いない手法としては LOF(Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander, 2000) と One Class SVM(B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson, 2001) が代表的である。

[1] shinnou@mx.ibaraki.ac.jp

2.1.1 LOF

LOF はデータの近傍の密度を利用することで、そのデータの外れ値の度合いを測り、その値によって外れ値を検出する。

LOF におけるデータ $x \in D$ の外れ値の度合いを $LOF(x)$ と表記する。ここで D はデータ全体の集合である。$LOF(x)$ を定義するために、いくつかの式を定義しておく。まず $kdist(x)$ は x に対する k 距離 と呼ばれる値で、以下の条件を満たすデータ $o \in D$ との距離 $d(x,o)$ として定義される。

1. 少なくとも k 個のデータ $o' \in D \setminus \{x\}$ に対して$d(x,o') \leq d(x,o)$が成立する。
2. 高々$k-1$ 個のデータ $o' \in D \setminus \{x\}$ に対して$d(x,o') < d(x,o)$が成立する。

直感的には、上記のデータ o はデータ x からの k 番目に近いデータとなる。データ x から同じ距離を持つデータが複数存在する場合を考慮して、上記のようなテクニカルな定義になっている。

次に $kdist(x)$ を利用して、$N_k(x)$、$rd_k(x,y)$ 及び $lrd_k(x)$ を以下のように定義する。

$$N_k(x) = \{y \in D \setminus \{x\} | d(x,y) \leq kdist(x)\}$$

$$rd_k(x,y) = \max\{d(x,y), kdist(y)\}$$

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} rd_k(x,y)}.$$

これらの式を用いて、$LOF(x)$ は以下で定義される。

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}$$

また LOF ではパラメータとして k が存在する。本論文では $k = 4$ を用いている。

2.1.2 One Class SVM

One Class SVM は ν–SVM(B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson, 2001) を利用した外れ値検出の手法である。すべてのデータは $+1$ のクラスに属し、原点のみが -1 のクラスに属するとして、ν–SVM を使って 2 つのクラスを分離する超平面を求める。原点はすべての点に対して類似度が 0 となるために、外れ値とみなせる。また ν–SVM はソフトマージンを利用するので、-1 のクラス側に属するデータを外れ値と判定する。

One Class SVM を利用する際には、用いるカーネル関数やどの程度のマージンの誤りを認めるかのパラメータの設定が結果に大きく作用する。本論文の実験では One Class SVM のプログラムとして libsvm [2] を用いた。カーネルは線形カーネルを利用し、マージンの誤りはパラメータ n に対応するが、$n = 0.02$ で固定した。

2.2 外れ値検出と新語義検出

一般に外れ値検出のタスクでは外れ値の客観的定義が不可能である[3]。これは外れ値にラベルをつける意味がないことを示している。なぜなら仮にあるデータが外れ値であり、その外れ値にラベルをつけることができたとしても、他の外れ値がそのラベル付きの外れ値と類似している保証がないからである。また検出元となるデータ集合は、ほぼすべて正常値である。仮にデータにラベルをつけるとすれば、正常値のラベルだけになり、教師データに意味はない。これらのことから外れ値検出の手法は教師なしの枠組みにならざるをえない。

[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[3] もしも定義できるのであれば、その定義にあったデータを取り出せばよいだけなので, タスクとしての意味はなくなる。

しかし新語義を外れ値と見なした新語義検出のタスクの場合、一般の外れ値検出とは異なった 2 つの特徴がある。1 つは外れ値の定義が明確である点である。ここでの外れ値は新語義の用例であるが、新語義とは辞書に記載されていない語義である、というように明確に定義できる。もう 1 つは正常値のデータは語義のクラスターに分割されるという点である。しかもクラスターの数も明確である。一方、通常の外れ値検出では正常値の集合がクラスターに分割されるのか、されるとしてもいくつのクラスターに分割されるのかは不明である。

ここではこれらの特徴を利用して外れ値検出を行う。具体的には検出元となる対象単語の用例集の一部に、対象単語の語義のラベルを付与し、その設定のもとで外れ値検出を行う。

2.3 語義識別問題としての新語義検出

対象単語の用例集の一部に対象単語の語義のラベルを付与した場合、帰納学習の手法を利用して語義識別を行う分類器を学習することができる。この分類器の識別の信頼度を利用して新語義の検出を行える可能性がある。ただしこのような分類器の識別の信頼度を利用する方法では新語義の発見は困難である。この点を簡単に注記しておく。

基本的に帰納学習で得られる分類器は、入力されるデータが与えられたクラスのいずれかに属することを仮定しており、その仮定の下で識別精度を高めることを目指している。例えば、SVM では分離平面だけが問題であり、クラスターの構造を考慮しない。そのため図 1 のような状況では、データ a とデータ b の識別の信頼度は分離平面までの距離が同じであるため等しいが、明らかに外れ値の度合いはデータ b の方が高い。

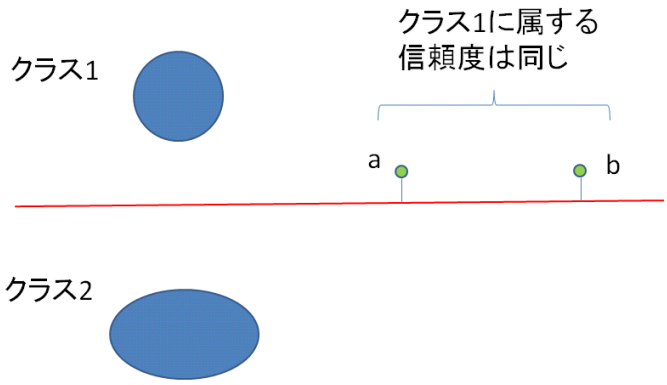


図 1: 識別の信頼度と外れ値の度合い

3 提案手法

ここでは外れ値（新語義の用例）を検出するために、2 つの手法を提案し、それらを組み合わせる。第 1 の手法は LOF を教師データを利用するように拡張したものであり、第 2 の手法は教師データから分類器を作成し、データのクラスを識別し、データと識別されたクラス間の距離関係から外れ値かどうかを判定するものである。第 1 の手法で外れ値の候補を取り出し、第 2 の手法でそれらを選別する。

3.1 教師データ付き LOF

教師データを LOF で利用するには単純に教師データをテストデータに加えればよい。しかしその場合、教師データからも外れ値が検出される可能性がある。

ここでは教師データを $k+1$ 倍してからテストデータに加えてデータセットを作り、そのデータセットに対して LOF を適用する。ただし k は LOF における $kdist$ で使われる k である。

LOF の場合、訓練データ x を $k+1$ 倍すると $kdist(x)=0$ となり、訓練データ x が外れ値として検出されることはなくなる。さらにテストデータ y と訓練データ x との距離が小さいと、その訓練データ x は $k+1$ 個存在するために、テストデータ y の密度も高まり、外れ値としては検出されなくなる。

3.2 クラス推定とクラスとの距離関係

教師付き LOF の場合、ラベル（語義）の種類による区別はない。何らかのラベルが付与されていれば、すべて正常値という扱いになる。ここでは教師データのラベルの種類を利用することを考える。

外れ値検出では、クラスタの分布がわかれば、外れ値かどうかの判断は閾値の問題だけになる。例えばクラスタの分布が多次元正規分布であれば、マハラノビスの距離からデータとクラスタ間の距離が測れるので、それによって外れ値の識別が可能になる。しかし教師データを用いたとしてもクラスタの分布の推定は困難なことが多い。

ここではクラスタの分布を仮定せずに、データがそのクラスタに対して外れ値になるかどうかを判定する。

まず教師データからクラスを識別する分類器を学習する。データ x に対してその分類器を用いて、そのデータのクラス A を推定する。次に A の中でデータ x に最も近いデータ $y \in A$ を見つけ、x と y 間の距離 $d(x,y)$ と y と A の重心 $\bar{A}$ 間の距離 $d(y,\bar{A})$ を測る (図 2 参照)。これらの比 r を求めて、ある値 r_0 以上のものを外れ値と判断する。

$$r = \frac{d(x,y)}{d(y,\bar{A})}$$

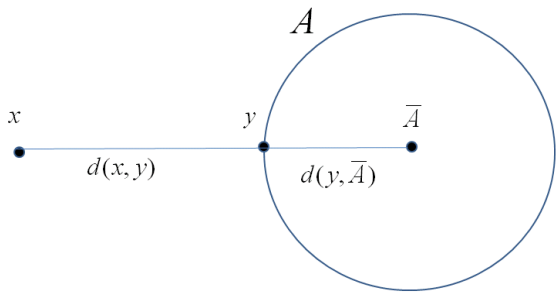


図 2: データとクラスタ間の距離関係

4 実験

ここでは提案手法の有効性を確認するために、人工的なデータと現実のデータを用いる。現実のデータは SemEval-2 の Japanese WSD タスクで使われたデータである。

4.1 人工データによる実験

3 つの 5 次元正規分布のモデルを作り、それぞれのモデルから 200 個のデータを生成する。5 次元正規分布の確率密度関数は以下である。

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu_c)\right\}$$

ここで Σ は分散共分散行列あり、ここでは各次元は独立、各次元の分散は異なる以下のモデルを利用した。

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5^2 \end{pmatrix}$$

また平均と分散は 0 以上 100 以下の値からランダムに取り出した。

各モデルから作られた 200 個のデータのうち 20 個を教師データする。また作成された 600 個の全データ内の最大値 max と最小値 min を求め、$[min, max]$ の範囲の一様分布から 5 次元の点を 20 個作成する。これが外れ値ある。以上よりクラス数は 3 個、教師データは 60 個、検出対象のデータは 560 個、うち外れ値は 20 個となる。

これらのデータに対して、LOF、One Class SVM (OCS)、それらの積を出力するもの (LOF+OCS) (Hiroyuki Shinnou and Minoru Sasaki, 2010)、教師付き LOF (S-LOF)、本手法の結果を以下に示す。ただし LOF では LOF 値の大きなもの上位 20 個を取り出すことにする。また本手法を使う際には $r_0 = 1$ とした。

表 1: 人工データに対する実験結果

手法	抽出数	正解数	F 値
LOF	20	12	0.600
OCS	30	12	0.480
LOF+OCS	8	5	0.357
S-LOF	20	12	0.600
本手法	10	10	0.667

また識別の信頼度による外れ値の検出も試みた。今、クラスは A、B、C の 3 つあるので、A か A 以外、B か B 以外、C か C 以外を識別する SVM を 3 つ学習し、各 SVM の結果が A 以外、B 以外、C 以外となった場合に、そのデータを外れ値とすることにした。この場合、検出数は 174 、正解数は 3 となり、検出の F 値は 0.031 であった。なおここで学習された SVM は外れ値を除いたテストデータ 540 個に対する正解率は 100% であり、識別の精度がよくても外れ値の検出は困難であることがわかる。

4.2 SemEval-2 Japanese WSD タスクのデータによる実験

SemEval-2 は語義曖昧性解消に関する評価型の国際会議であり、いくつかのタスクが設定されている。Japanese WSD はその中の 1 つである。通常の日本語の語義識別のタスクであるが、最も特徴的な点は、識別結果に新語義というカテゴリを含めている点である。つまりテストデータの中には設定された語義のどれでもないという答えがありえる。そのため、このタスクで用意された訓練データとテストデータを用いることで、教師付きの枠組みでの新語義の検出手法の評価が可能である。

Japanese WSD の語義識別の対象の単語は 50 単語である。この中で「可能」「入る」は教師データ内に新語義の用例があるので、それらを外して、残り 48 単語を実験対象とした。各単語を以下に示す。

名詞　21 単語
　相手、意味、関係、技術、経済、現場、子供、時間、市場、社会、情報、手、
　電話、場合、はじめ、場所、一、文化、ほか、前、もの

動詞　22 単語
　会う、あげる、与える、生きる、入れる、教える、考える、勧める、する、
　出す、立つ、出る、とる、乗る、始める、開く、見える、認める、見る、
　持つ、求める、やる

形容詞　5 単語
　大きい、高い、強い、早い、良い

新語義は「意味」で 1 用例、「手」で 3 用例、「前」で 7 用例、「求める」で 1 用例、「あげる」で 2 用例、「はじめる」で 2 用例の計 16 用例存在する。これらが検出の正解となる。以下にこの 16 用例を示す。

1. … 意識の開きが、ある意味で、科学技術と社会に関する …
2 … 医業収益等は手入力 …
3. … 本部での集約も手入力、…
4. … 経理コンピュータへの予算入力も手入力で …
5. … ランチ＝前十一時半～後３時。
6. … 二十四日火、前十時～後７時 …
7. … 来年３月二十日木までの前十時～後十時、…
8. … 大沢悠里のゆうゆうワイド（ＴＢＳ＝前８・三十）…
9. … 三十日水までの前十一時半～後２時半、…
10. … 十九日土～十二月二十日火、前十一時半～後２時半、…
11. … 前十時半と後６時、本館１階正面口で …
12. … インフラ不安に要因を求め、…
13. 国を挙げて緑化を進めた。
14. 国をあげて緑化に取り組んだシンガポールは、…
15. 16 動作などをあらわす「はじめる・はじまる」は「初」でなく、「始める・始まる」と書きます。

実験の結果を以下に示す。LOF では LOF 値の大きなもの上位 5 個を取り出すことにする。また本手法を使う際には $r_0 = 3$ とした。

表 2: SemEval-2 データに対する実験結果

手法	抽出数	正解数	F 値
LOF	240	0	0.000
OCS	1150	3	0.005
LOF+OCS	83	0	0.000
S-LOF	240	3	0.023
本手法	36	2	0.077

5 考察

人工データに対する実験結果は以下の点を示している。

1. 教師なしの LOF や One Class SVM でも、ある程度の検出は可能である。
2. LOF に教師データを利用する効果は少ない。
3. 推定クラスとの距離を測る本手法のフィルターは有効である。

ただし SemEval-2 のデータに対する実験結果を見ると、(1) や (2) は逆になる。人工データはデータの生成が単純なモデルで表現できている。このような場合は教師なしの手法でもうまくいくが、SemEval-2 のデータのようにデータの生成が複雑、つまり正常値のクラスターが複雑な形状をしている場合は、教師なしの手法は有効に働かず、教師データを利用する効果が高い。実験では、教師データを利用することで抽出できなかった新語義も抽出できている。また LOF では以下の用例が検出されている。

(a) 地盤が悪くては 意味 がないからです。
(b) ご主人に対してだけ対策してもあまり 意味 ないですよ。

一見、悪くない検出であるが、実は (a) は教師データの一つなので、(b) を検出するのは避けなければならない。教師付き LOF では、この問題を避けることができている。

ただし、SemEval-2 のデータではデータ数が少なく、しかも教師データとテストデータがほぼ同じ数存在するという不自然な状況のために LOF において教師データの利用の効果が生じたとも考えられる。

(3) については SemEval-2 のデータに対する実験でも確認できた。本来、正常値の適切な生成モデルやクラスター形状が推定できれば、外れ値検出を精度良く行えるため、教師データからそれらを推定するアプローチは有効である。

本手法の誤検出の原因について述べる。1 つは書き誤りに近いものである。例えば、以下は助詞が抜けていると見なすこともできる。

(c) 私が 子供 産んだ頃は、
(d) 忙しいでしょうから、お 時間 あるとき、

書き誤りは検出されてもしかたないし、この類の検出は有益性もあり問題は少ない。

他の誤検出の原因はいくつかあるが、複合語の認識の問題が大きい。名詞の語義識別の場合、対象単語が複合語の一部になっていれば、前後の単語の情報は語義識別の上での大きな情報となる。このため特異な複合語が検出されることが多い。検出された複合語が実際に専門性の高い用語である場合もあり、そのような場合には意味のある検出とも見なせるが、現在は複合語を単なる名詞連続で認識しているために以下のような検出が散見される。

(e) そんな 時間 必要ないけど、
(f) 給食費の未納がものすご〜く多い学校 現場 です。
(g) 加入 電話 サービスの基本料は

(e) は助詞が抜けて複合語と誤認識している。(f) や (g) などは専門用語か一般用語かの判断とも関わり、ここで行っているような単純な処理では解決は難しい。

本手法の未検出の原因としては、突き詰めれば、用例間の距離の測定方法に帰着される。ある新語義の用例と他の正常値の用例との距離がある程度、離れていたとしても、正常値の用例間の距離もそ

の程度は離れているという状況である。これは動詞や形容詞における検出では顕著である。この解決は語義識別の場合と同じであり、語義識別の精度向上の試みが本研究に応用できると考えている。これは今後の課題である。

もう一つ本手法の課題を述べておく。本手法では 3 つのパラメータが存在する。LOF における k-距離の k、LOF 値の上位いくつまでを候補に取るか、及び r_0 の値である。これらの値が異なると検出結果は全く異なってしまう。ここでの実験では予備実験を行い、適切そうな値を見積もって設定している。これらのパラメータはタスクに応じて、最適な値は異なるはずであり、これらパラメータの適切な設定方法も今後の課題である。

6 おわりに

本論文では対象単語の用例集合から、その単語の語義が新語義となっている用例を検出する手法を提案した。基本的に新語義の用例を用例集合中の外れ値と考え、外れ値検出の手法を利用する。ただし従来の外れ値検出では教師なしの枠組みであるが、ここではタスクの性質を考え、教師付きの枠組みで行った。

提案手法は教師データを利用した手法である。人工的なデータや SemEval-2 の Japanese WSD タスクのデータを用いた実験により、提案手法の効果を示した。

提案手法には 3 つのパラメータが存在するので、それらを適切に設定する方法を考案することと、語義識別の精度を向上させる工夫を本研究に利用することが今後の課題である。

文献

B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson (2001) "Estimating the support of a high-dimensional distribution," *Neural Computation*, Vol. 13, No. 7, pp. 1443–1471.

Hiroyuki Shinnou and Minoru Sasaki (2010) "Detection of Peculiar Examples using LOF and One Class SVM," in *LREC-2010*.

Manabu Okumura and Kiyoaki Shirai and Kanako Komiya and Hikaru Yokono (2010) "SemEval-2010 Task: Japanese WSD," in *The 5th International Workshop on Semantic Evaluation*, pp. 69–74.

Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander (2000) "LOF: Identifying Density-Based Local Outliers," in *ACM SIGMOD 2000*, pp. 93–104.

山西健司 (2009) データマイニングによる異常検知，共立出版.

SemEval-2010 日本語語義曖昧性解消タスク報告

奥村 学 (言語処理班班長: 東京工業大学 精密工学研究所) †
白井 清昭 (言語処理班分担者: 北陸先端科学技術大学院大学 情報科学研究科)
古宮嘉那子 (言語処理班協力者: 東京農工大学 工学研究院)
横野 光 (東京工業大学 精密工学研究所)

Report on the SemEval-2010 Japanese WSD Task

Manabu Okumura (Tokyo Institute of Technology)
Kiyoaki Shirai (Japan Advanced Institute of Science and Technology)
Kanako Komiya (Tokyo University of Agriculture and Technology)
Hikaru Yokono (Tokyo Institute of Technology)

1 はじめに

語義曖昧性解消は，意味解析技術の一つとして，古くから自然言語処理分野で研究が進められている技術である．語義曖昧性解消では，複数の語義をもつ単語を対象に，与えられた文脈中で，辞書中のその単語の語義区分に基づき, どの語義で用いられているかを自動判定する．この技術の水準向上を目的とした評価型ワークショップが過去何度か開催されている (Senseval-1/2/3, SemEval-2007[1]). その中では，様々な言語における語義曖昧性解消タスクが設定されてきており，また，最近では，あらかじめ決められた語義区分を仮定することなく，与えられた用例集合をクラスタリング等することにより，単語の語義区分を同定するタスクも設定されていたりする．

語義曖昧性解消に関するこの評価型ワークショップは 3 年に 1 度のペースで開催されており，最も最近では Semeval-2 のワークショップが 2010 年に開催された (`http://semeval2.fbk.eu/Semeval2.html`)．この Semeval-2 に我々は後述する 2 つの特徴を持つ語義曖昧性解消の評価型タスクを提案し，無事採択された．本稿では，このタスクの報告を行う. なお，タスクの詳細，データ，評価手法，参加システムの概要，結果等については，task description paper である [1] を参照していただきたい.

2 代表性のある語義タグ付コーパスの構築

2006 年 9 月にスタートした，文部科学省科学研究費補助金特定領域研究「日本語コーパス」プロジェクト (`http://www.tokuteicorpus.jp/`) では，現代日本語書き言葉の大規模な均衡コーパス (「現代日本語書き言葉均衡コーパス」，BCCWJ; Balanced Corpus of Contemporary Written Japanese と呼ばれている) を構築するとともに，それを活用した研究によりコーパスを評価することを目指している．この中で我々は現在，代表性のある語義タグ付コーパスの構築を行っている．領域内で公開されているコアデータ (BCCWJ を構成するように，サンプリングされたデータ) に対して, 岩波国語辞典中の語義の区分に基づき, 人手で語義を付与する作業を行っている. 過去のタグ付コーパス構築例にならい [2], タグ付けの際, 辞典中に該当の語義が見当たらない場合「該当なし」という判断を許し, また, 最下層の語義のどれかでは判断できない場合, より上位のラベルを付与することを許している．「該当なし」の場合，大辞林をひき，該当する語釈文があれば，それを明記し，該当するものがなければ，作業者自身が考えた語釈文を記載してもらうようしている．

日本語の語義タグ付コーパスには，EDR コーパス (20 万文)，RWC コーパス (3000 記事) があるが，いずれも代表性のあるコーパスを元にしていない．海外では，代表性のあるコーパスの上にタグ付けを行うことで，代表性のある語義タグ付コーパスの構築が進んでおり，日本語においての構築は急務であると考える．

†oku@pi.titech.ac.jp

[1] `http://nlp.cs.swarthmore.edu/semeval/index.php`,
`http://www.senseval.org/senseval3/`,
`http://193.133.140.102/senseval2/`,
`http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/index.html`.

3 BCCWJ を用いた新しい語義曖昧性解消タスク

このコーパスが利用できるようになると，以下のような特徴を持つ語義曖昧性解消 (WSD; Word Sense Disambiguation) の評価型タスクが設定できる．

1. 日本語の語義タグ付コーパスはこれまですべて新聞データを元にしていたが，日本語で最初の代表性のある語義タグ付コーパスを用いた WSD タスクとなる．
2. これまでの WSD タスクでは，あらかじめ仮定した辞書中の語義セットから語義を選択する必要があったが，実際には辞書中に該当する語義がない用例も多数存在する．そのような，辞書中に語義がない用例も対象とする，初めての WSD タスクとなる．

3.1 代表性のあるコーパスを用いた語義曖昧性解消

代表性のあるコーパス中には, 複数のジャンルのテキストが混在していることになる. したがって, コーパスは, いくつかのジャンルごとのサブコーパスに分割できることになる. 近年の語義曖昧性解消研究では, 訓練コーパスを用いて分類器を学習し, その分類器により語義を同定する (ある単語の出現がその単語の語義のうちどの語義の出現であるか分類する) 手法が採用されることが多く, また, より良い性能を得られている. この時，単語によっては，サブコーパスごとに，出現する語義の頻度分布が異なる場合が存在する．すると, あるジャンルのテキスト中の用例を対象に語義曖昧性解消しようとする時，同一ジャンルのサブコーパスを学習に利用するのが良さそうであるのは言うまでもないが，それ以外に，コーパス中のどのサブコーパスをどのように学習に利用するのが良いのかは自明な問題ではない．これはある種の領域適応 (domain adaptation) の問題であるが，これまでのように単一ジャンルのテキスト (たとえば，新聞データ) を利用していた場合にはさほど顕在化していない問題である．

3.2 新語義の発見

従来の語義曖昧性解消では，単語の語義を辞書などによってあらかじめ定義し，これらの語義の中からテキスト中の単語に対する適切な意味を選択する．ところが，単語の意味は年月とともに変化し，新しい語義や用法も日々生まれている．そのため，単語の語義をあらかじめ定義するのは必ずしも適切であるとは言えない．そこで，ある用例における単語の意味が既存の辞書に定義された意味に該当するのか，あるいは辞書の意味のいずれにも該当しない新語義なのかを判定することにより，単語の新語義を発見するという必要が生じる．

3.3 課題設定

課題の詳細は以下の通りである．なお，Semeval-2 傘下のタスクであるため，公式なタスク定義はすべて英語で記述されている．以下の英語の部分は，タスクの web ページ (`http://lr-www.pi.titech.ac.jp/wsd.html`) 中の記述の抜粋である．

Task description:

> This task can be considered an extension of SENSEVAL-2 JAPANESE LEXICAL SAMPLE Monolingual dictionary-based task. Word senses are defined according to the Iwanami Kokugo Jiten, a Japanese dictionary published by Iwanami Shoten. Please refer to that task for reference.
>
> **Input:** Test documents with marked target words from the BCCWJ corpus, where the genre of documents is also provided, because of their diversity. Examples include books, newspaper articles, white papers, blogs, magazines, and documents from a Q&A site on the WWW.
>
> **Output:** The sense ID of each target word in the Iwanami Kokugo Jiten if the sense is in the dictionary. If systems find that the sense is not in the dictionary, say ‘new sense.’

The evaluation methodology:

> Organizers will return the evaluation in two ways:
>
> a. evaluating the outputted sense IDs, assuming the 'new sense' as another sense ID. The outputted sense IDs will be compared to the given gold standard word senses, and the usual precision measure for supervised word sense disambiguation systems will be computed using the standard SENSEVAL scorer. The Iwanami Kokugo Jiten has three levels for sense IDs, and we use the middle-level sense in the task. Therefore, we can call the scoring in the task 'middle-grained scoring.'
> b. evaluating the ability of finding the instances of new senses, assuming the task as classifying each instance into a 'known sense' or 'new sense' class. The outputted sense IDs (same as in a.) will be compared to the given gold standard word senses, and the usual accuracy for binary classification will be computed, assuming all sense IDs in the dictionary are in the 'known sense' class.

The availability of the resources:

> The Iwanami Kokugo Jiten will be available soon from GSK (`http://www.gsk.or.jp/`). A corpus annotated with sense IDs will also be distributed as training data. Each article will be assigned its genre code. Participants in this task are required to submit a copyright agreement form to the National Institute of Japanese Language.

4 訓練/テストデータ

訓練データとしては，コアデータ中，白書，書籍，新聞の 3 ジャンルのデータを，テストデータとしては，コアデータ中，白書，書籍，新聞，Yahoo!知恵袋の 4 ジャンルのデータをそれぞれ用いた．テストデータにおいて，語義曖昧性解消の対象とした単語は，合計 50 語 (名詞 22 語，動詞 23 語，形容詞 5 語) である．各対象単語についてそれぞれ 50 用例，合計 2,500 用例を評価に用いた．言うまでもないが，訓練データ中とテストデータ中の用例には重なりは存在しない．

5 参加システム

国内外の 10 グループが参加登録を行ったが，最終的には，4 グループからの 9 システムが結果を提出した．なお，このタスクの参加グループも含め，現時点までで合計で領域内 5 グループ，領域外の国内 4 グループ，領域外の海外 4 グループが，このデータを利用している．

6 結果

参考までに参加システムの評価結果を表 1, 2 に示す．語義曖昧性解消に対するベースラインシステムは，これまでに語義曖昧性解消に用いられてきている種々の素性を用い，SVM により教師あり学習を行った，比較的強力なシステムである． 新語義に対する用例が比較的少なかったため (テストデータ 2,500 用例中 39 用例)，新語義発見は非常に難しいタスクとなっている．その結果，どのシステムもあまりよい性能を示せていないことが分かる．

7 おわりに

本稿では，SemEval-2 日本語語義曖昧性解消タスクについて報告した．本タスクで用いたデータは，タスクオーガナイザに連絡をとり,「コアデータに関する誓約書」を提出することにより利用可能となる．よく多くの研究者が本データを用いて研究を行ってくれることを期待している．

文献

[1] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. SemEval-2010 task: Japanese WSD. In *Proceedings of SemEval-2010*, pages 69–74, 2010.

[2] 白井 清昭. Senseval-2 日本語辞書タスク. **自然言語処理**, 10(3):3–24, 2003.

表 1: 結果: 語義曖昧性解消

	精度
Baseline	0.7528
HIT-1	0.6612
JAIST-1	0.6864
JAIST-2	0.7476
JAIST-3	0.7208
MSS-1	0.6404
MSS-2	0.6384
MSS-3	0.6604
RALI-1	0.7592
RALI-2	0.7636

表 2: 結果: 新語義発見

	正解率	精度	再現率
Baseline	0.9844	-	0
HIT-1	0.9132	0.0297	0.0769
JAIST-1	0.9512	0.0337	0.0769
JAIST-2	0.9872	1	0.1795
JAIST-3	0.9532	0.0851	0.2051
MSS-1	0.9416	0.1409	0.5385
MSS-2	0.9384	0.1338	0.5385
MSS-3	0.9652	0.2333	0.5385
RALI-1	0.9864	0.7778	0.1795
RALI-2	0.9872	0.8182	0.2308

BCCWJ を利用した日本語作文支援システム「なつめ」の評価

阿辺川　武（作文支援班連携研究者：国立情報学研究所）†
ホドシチェク・ボル（作文支援班協力者：東京工業大学）
仁科喜久子（作文支援班班長　：東京工業大学）

The Evaluation of Japanese Writing Support System "*Natsume*" Using BCCWJ

Takeshi Abekawa (National Institute of Informatics)
Bor Hodoscek (Tokyo Institute of Technology)
Kikuko Nishina (Tokyo Institute of Technology)

1. はじめに

我々は、日本語を学習する留学生を対象にした作文支援システム「なつめ」を開発している。中級者以上の学習者にとって、ある語に対して意味的に共起する語を想起することはできるが、執筆するジャンルに応じて適切な語を選択することはまだまだ難しい。そこで共起語のジャンル別表示を中心に機能の拡張を実施した結果、ユーザの入力した語に対し、BCCWJ で定義されたジャンルごとに共起する語の比較が可能になった。研究の最終年度に当たり、BCCWJ を利用した日本語作文支援システムの評価および、システムからみた BCCWJ の評価を最終目標とした。本稿では、最初に前回の報告以後に実装された複数語の入力、類義語表示、入力サジェストの各機能の説明を行い、次に留学生を対象に、構築したシステムを利用した論文執筆を想定した書き換え実験の結果を報告する。

2. 日本語作文支援システム「なつめ」について

「なつめ」は、日本語を学習する留学生向けの作文支援システムであり、BCCWJ のデータから名詞と動詞、副詞と動詞といった語の共起情報を検索、表示する機能を持つ。従来は、利用者が知りたい語を入力し、その語と共起する語をリストとして表示するだけであったが、BCCWJ のジャンル情報を有効に活用するにあたり、図１のようにジャンルごとの共起分布をグラフとして表示できるようにした(阿辺川ら 2010a)。

この結果、例えば一般書籍では高頻度で共起するが、論文ではほとんど共起しない語、またその逆の傾向を持つ語などが一覧できるようになり、日本語の執筆に慣れてない留学生にとって、ジャンルに適切な語彙を選択することが容易になる。本システムでは、「名詞・格助詞・動詞」三つ組共起の頻度を取得するデータとして BCCWJ、および独自で収集した科学技術論文と Wikipedia を使用している。これらのデータは共起頻度の計算とともに、ユーザに例文を提示するため、著作権処理の済んだ公開可能なものに限定している。

3. 拡張機能

本節では前回の報告から、拡張され追加された機能について説明する。

3. 1 複数語指定

† abekawa@nii.ac.jp

図 1 「なつめ」のスクリーンショット

学習者が入力する語を複数指定できるようになった。学習者は、システムの「名詞」または「動詞」のフォームに 4 語までの語を入力し、それぞれの語に対する共起の度合いを比較することができる。例えば図 1 は、入力語として「走る 疾走する 走行する」を指定したものである。それぞれの共起が異なる背景色で示され、その度合いが濃い色の大小幅で示される。背景色にあたる箇所が空欄の場合は、その共起はコーパス中に存在しないことを示す。このような提示法により、複数の指定語に対する複数の共起語の共起度合いを一度に閲覧することが可能になった。複数語が入力できる既存の共起語表示システムとして、Sketch Engine (Kilgarriff ら 2004)が挙げられるが、2語までの入力しかできず、3語以上の語を並べて表示することができない。

3. 2 類義語の提示

学習者が「名詞」または「動詞」に語を入力し、その下の「類義語」ボタンを押すと、その入力した語に対する類義語がリストで表示される。複数の語を入力したときは、それぞれの語を考慮しての類義語となる。例えば「思い」が入力語のときは類義語として「想い 感慨 激情 念 いら立ち」が表示され、「思い 想い」と複数の語を指定したときは「感慨 恋心 激情 郷愁 憎しみ」が類義語として表示される。

3. 3 入力語サジェスト

「名詞」または「動詞」のフォームに語を入力する際に、1 文字入力するごとに候補が表示される(図 2)。漢字・仮名だけでなくアルファベットを用いたローマ字入力にも対応する。サジェスト候補は、現時点までに入力文字列を接頭語とする語を頻度順にソートしたときの上位 10 語である。動詞の

場合、語幹を入力すれば異なる活用形を持つ語や複合動詞がサジェストされ、基本動詞以外の動詞があることに気づく。また韓国語を母語とする人のように日本語を音で覚えている学習者にとっては、漢字を入力することなく、ローマ字で音のまま入力できるので便利である。

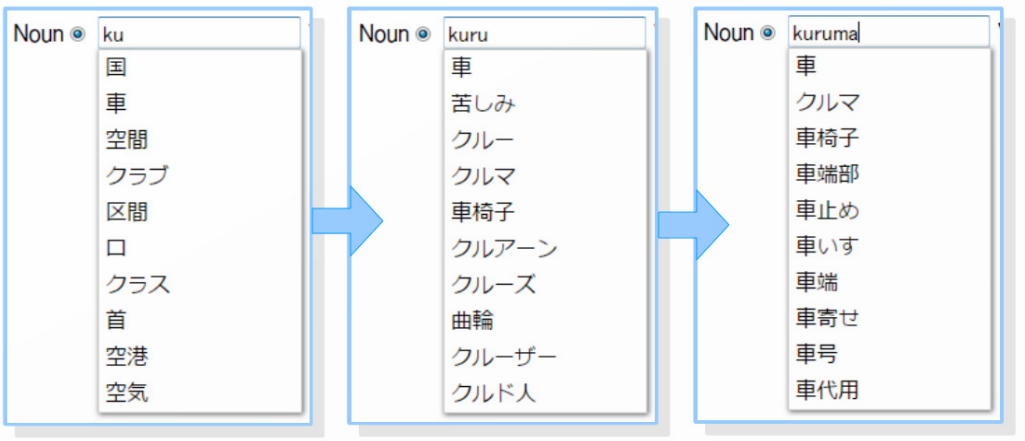


図 2　入力サジェストの例

3．4　共起語リストの並び順

共起語リストの表示順は、初期値では頻度の降順となっており、学習者にとっては、この頻度順で十分であると考えている。しかし、日本語の辞書を作成する研究者などから、特徴的に共起する語が知りたいなどの要望があり、本システムでは以下の尺度によるランキングを選択することができるようにした。Dice 係数、T スコア、Jaccard 係数、対数尤度比、χ 二乗係数、相互情報量。また、入力フォームで複数語指定のときは、ソート軸の基準となる語を指定可能とした。

4．被験者評価実験

ジャンル別表示に対応した共起検索機能の有効性を検証するために、次のような被験者評価実験を試みた。本実験は、2010 年7月に行った予備実験(阿辺川ら 2010b)に改良を加えた第2回目の実験であり、基本方針は予備実験とほぼ同様である。

4．1　実験方法

「なつめ」における共起検索の有効性について、学習者評価実験を試みた。被験者は、理系学部 1 年生、2 年生 40 名で、学部生は日本語能力試験 1 級保持者である。実験は与えられた文および文章を論文調に書き換える課題を設定した。作題文は、それぞれ 1 級から 4 級および級外までの語彙がほぼ均等になるように配置し、論文のためには書き換えが必要な項目が均等に含まれる問題セットを A、B の 2 種類準備した。

最初に被験者を 2 グループに分割し、片方のグループには問題 A を、もう片方のグループには問題 B を配布し、「できるだけ論文らしい表現になるように」と指示し筆記で解答させた。所用時間は 60 分程度であり、電子辞書の使用は許可とした。次に「なつめ」を利用して、2 グループに対して問題を交換して課した。指示は筆記テストと同様にできるだけ論文らしい文を作成するように指示した。その際、できるだけシステム上の「類義語」を参照し、「科学技術文」など論文に近いコーパスに高い頻度があるものを選択するように指示した。

採点方法は、「なつめ」で検索可能な共起語が正しく書き換えられた箇所を配点 2 点(23 箇所)、

科学技術論文のレジスターとして必要と思われる副詞、形容詞、文末モダリティなどの表現項目の問題を配点 1 点(20 箇所)、計 66 点満点とした。

4．2　実験結果

被験者 40 名について上記の評価方法によって得点を集計した。A、B それぞれのグループの得点を集計し、筆記実験のグループ別平均値の差を検定した結果、A、B 間の有意差は認められず(p 値=0.8851)、ほぼ均一レベルの学習者に対して問題文の難易度の差がないことがわかった。

図3のグラフは、システムを使用しない筆記テスト時の得点の高い順に被験者を並べたときの、筆記得点、なつめシステム利用時の得点、筆記→なつめの得点の差分を示したものである。斜行する3本の直線は、それぞれの得点の回帰直線である。また、表1は、筆記試験の得点により上位、中位、下位の 3 群に分けたときの、「なつめ」利用時における得点の増減を平均して集計した結果である。この結果、筆記上位群よりも、筆記下位群の方が「なつめ」利用時における得点の増加点数が大きいことから、日本語能力試験 1 級合格者の中でも中位群、下位群において「なつめ」利用の効果が高いことが明らかになった。

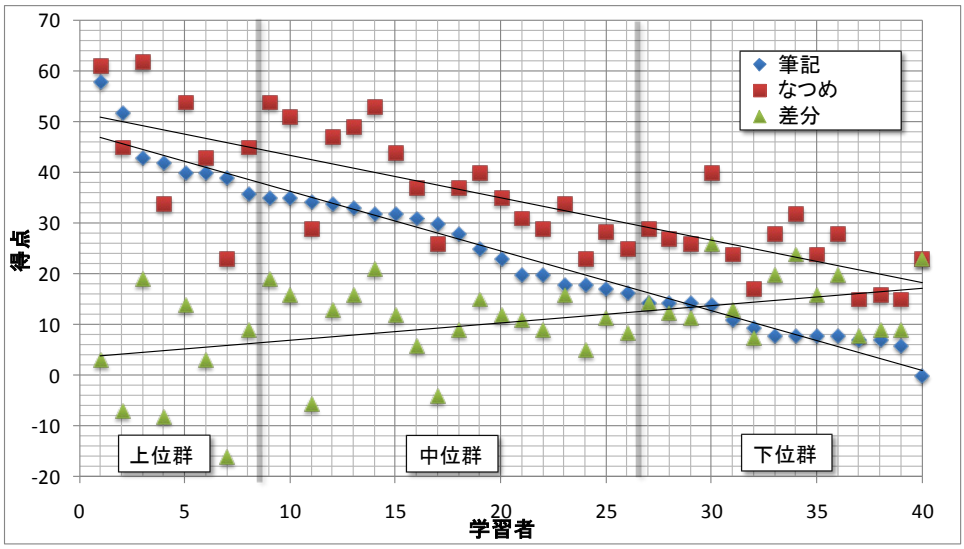


図 3　被験者の各得点分布

表１　筆記上位、中位、下位群ごとの各得点の平均点

	上位群	中位群	下位群
対象被験者	8 名	20 名	12 名
筆記得点 平均点	58～36 43.75	35～16.5 26.78	14.5～0 9.29
なつめ得点 平均点	62～23 45.88	54～23 37.36	40～15 24.57
差分範囲 差分平均点	19～-16 0.75	19～-5.5 10.41	26～7.5 15.29

次に、被験者の母語により、得点の傾向に差があるかどうかを検証した。表2は、被験者を中国語母語話者、韓国語母語話者、その他の言語（インドネシア、タイ、モンゴルなど）の話者の3群に分けて得点の平均値を集計したものである。全体の傾向として韓国語母語話者が筆記では得点が低く、その他の母語話者の得点が高い。一方、システム利用による得点の上昇度は韓国語母語話者の方が高いことから、筆記下位群の方が「なつめ」の利用の効果が高いことがわかる。また中国語母語話者は、その他の母語話者よりも平均点が低い上に差分平均点も低い。このことから中国語母語話者にとって「なつめ」システムは、それ以外の母語話者よりも使い勝手が悪い可能性が考えられる。その理由のひとつとして、漢字がわかる中国語母語話者にとって、日本語の漢字を母国語の意味で判断してしまい、日本語の適切な共起関係を把握していないことがあると思われる。今後は被験者の解答を詳細に分析し、これらの理由の分析を行いたい。

表２　母語別の各得点の平均点

	中国語	韓国語	その他
対象被験者	20 名	11 名	9 名
筆記平均点	23.2	18.9	30.6
なつめ平均点	31.8	33.1	40.7
差分平均点	9.5	14.2	9.8

４．３ 問題文の分析

本節では、評価実験で用いた問題文および模範解答文に対し、実際に「なつめ」システムを用いて共起事例が検索できるかどうかを検証した結果を報告する。最初に先ほどの被験者実験で用いた問題を日本語母語話者2名に筆記で解いてもらい模範解答文とした。次に問題文、模範解答文から書き換えの対象となる共起部分を抽出し、「なつめ」を用いて共起事例を検索した（書き換え対象となる共起部分は1つの問題文に対し複数箇所ある場合がある）。BCCWJ および、我々が独自に収集した科学技術論文に対し、問題文・模範解答文の共起対が検索可能かどうかを表3に掲載する。

表３　コーパス別の共起事例の有無

問題数	問題文		模範解答文		問題共起対例	解答共起対例
	BCCWJ	科技論文	BCCWJ	科技論文		
8	有り	なし	有り	有り	考えを言う	考えを述べる
5	有り	有り	有り	有り	論文を書く	論文を執筆する
5	有り	なし	有り	なし	被害が来る	被害をもたらす
5	なし	なし	有り	有り	影響をもらう	影響を受ける
4	なし	なし	有り	なし	関係が悪くなる	関係が悪化する
2	有り	なし	なし	なし	ごみが捨てられる	ごみが投機される
2	なし	なし	なし	なし	仕様を比べる	仕様を比較する

表3より、模範解答文の共起対が「なつめ」システムで検索できない問題文が数文あり、これらの問題に正解した被験者は、自分の実力により正解していたことになる。BCCWJ と科学技術論文を比較すると、問題文で出現する共起対、模範解答文で出現する共起対ともに、科学技術論文にある共起は必ずBCCWJ に存在することがわかった。論文を執筆するにあたり、専門用語や分野特有の言い回しについてはその分野の論文コーパスを用いる必要はあるが、今回のような論文調に必要となる基本的な共起対は BCCWJ があれば十分であることが予想される結果となった。

5. まとめ

本稿では、日本語作文支援システム「なつめ」の機能拡張および BCCWJ を用いた評価実験について報告した。今後の機能拡張については、今回の実験結果を踏まえ、共起語の意味を表示できるようにしたいと考えている。また利用者からの声の多い「形容詞 名詞」や「副詞 モダリティー」などの共起についても今後システムに搭載していきたい。また、今回の被験者実験の解答について、問題ごとの難易度や、解答にかかった時間などの観点から分析を行っていきたいと考えている。

文献

阿辺川武、Hodoscek Bor、仁科喜久子(2010a)、「日本語作文支援システム「なつめ」における共起語検索方法の改訂」、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予稿集 pp243-244、Mar 2010.

阿辺川武、HodoscekBor、仁科喜久子(2010b)、「日本語作文支援システム「なつめ」―利用者の視点―」特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集 pp243-244、 Aug 2010.

赤瀬川史朗(2010). Wordprofiler、 http://www。lagoinst。com/WordProfiler/. 2010.

Adam Kilgarriff、 Pavel Rychly、 Jan Pomikalek (2004). The sketch engine. In: Williams G、 Vessier S、 editors. Proceedings of the Eleventh EURALEX. 2004.

日本語フレームネットにおける BCCWJ への意味アノテーション

小原　京子（日本語フレームネット班班長：慶應義塾大学理工学部）†
加藤　淳也（日本語フレームネット班協力者：慶應義塾大学理工学研究科）
斎藤　博昭（日本語フレームネット班分担者：慶應義塾大学理工学部）

Full Text Annotation of BCCWJ in Japanese FrameNet

Kyoko Hirose Ohara　(Faculty of Science and Technology, Keio University)
Junya Kato　(Graduate School of Science and Technology, Keio University)
Hiroaki Saito　(Faculty of Science and Technology, Keio University)

１．はじめに

本稿では日本語フレームネット（略称 JFN）班における、「現代日本語書き言葉均衡コーパス」(BCCWJ)への意味アノテーション、つまり意味フレーム名の付与作業について報告する（http://jfn.st.hc.keio.ac.jp/）。日本語フレームネットでは、BCCWJ モニター公開データを対象に、テキスト内に出現する自立語すべてへの意味フレーム名の付与（全文テキストアノテーション）を行った。本稿では、BCCWJ の「書籍」ジャンルのテキストへのアノテーション作業を中心に、1) 英語フレームネット[1]（略称 FN）上の意味フレーム定義の適合率、2) 日本語固有の意味フレーム定義の必要性、3) アノテータ間の意味フレーム名付与の一致率について述べる。 英語・日本語フレームネットの枠組みに基づく意味フレーム名付与済みコーパスは、意味タグ付きコーパスとして情報検索・テキスト要約などの自然言語処理アプリケーションに利用されることが期待される。

フレームネット・プロジェクトでは、フレーム意味論とコーパスデータに基づき英語のオンライン語彙情報資源を構築中である（http://framenet.icsi.berkeley.edu/, Fillmore & Baker 2010）。日本語フレームネット・プロジェクトは 2002 年から始まった日本語語彙情報資源構築プロジェクトで、フレームネット・プロジェクトとの連携のもとに進められている（Ohara & Sato 2010, Tagami et al. 2009, cf. Hasegawa et al. 2010）。フレームネットの手法で、コーパスデータを用いて語の意味・用法の分析を行い、オンライン日本語語彙情報資源の雛型を構築している。英語語彙分析のためにフレームネットで定義された意味フレームが類型論的に異なる日本語の語彙意味記述にどこまで適しているのかを検討するのが主な目的の一つである。

本稿の構成は以下のとおりである。まず、次節で日本語フレームネットにおける全文テキストアノテーション、すなわち BCCWJ への意味フレーム名付与作業の概要について述べた後、第 3 節で全文テキストアノテーション結果閲覧のためのツール、全文テキストアノテーション Web Report を紹介する。第 4 節では英語フレームネット上で英語語彙の意味分析のために定義された意味フレームがどこまで日本語テキストのアノテーションに適用できたかを、適合率の観点から報告する。それを踏まえ、第 5 節では日本語固有の意味フレ

† ohara@hc.cc.keio.ac.jp

1 正式名称は FrameNet であるが、本稿では日本語フレームネットと比較して議論する際に必要に応じて FrameNet を「フレームネット」ではなく「英語フレームネット」と表記することにする。オンライン語彙資源構築にフレームネット同様の枠組み・手法を用い、フレームネット・プロジェクトと共同研究を行っているプロジェクトとしては、日本語フレームネット・プロジェクトの他に、スペイン語フレームネット・プロジェクト(http://gemini.uab.es:9080/SFNsite)やドイツ語フレームネット・プロジェクト(http://gframenet.gmc.utexas.edu/)がある。

ームとして新たに日本語フレームネット上で定義が必要な意味フレームについて考察する。第 6 節ではアノテータ間の意味フレーム名付与作業の一致率について述べる。

２．日本語フレームネットにおける全文テキストアノテーションと BCCWJ

日本語フレームネットでは語彙項目アノテーションと全文テキストアノテーションという二つのモードで BCCWJ へのタグ付けを行ってきた。語彙項目アノテーションとは、語彙項目ごとに BCCWJ の中からアノテーション対象とする例文を選びそれらの例文に対してタグ付けするモードである。これに対して全文テキストアノテーションとは、特定のサンプルテキスト内の全ての文の、意味フレーム（言語の発話や理解の際に必要となる、体系的知識構造）を喚起（evoke）する全ての語彙項目に対してタグ付けするモードを指す。これまで語彙アノテーションでは BCCWJ モニター公開データ 2008 年度版を、全文テキストアノテーションでは BCCWJ コアデータ（人手で形態素解析結果を修正した、各ジャンルのサンプルのサブセット）を対象に分析とアノテーションを行ってきた。

全文テキストアノテーションでは、テキスト内のすべての文の、意味フレームを喚起するすべての語彙項目に対してアノテーションを行う。固有表現以外の語彙項目が対象である。本稿では、BCCWJ コアデータ書籍ジャンルの各サンプル（総数 84 ファイル）の冒頭 10 行の意味フレーム喚起語への意味フレーム名付与結果について論じる。

全文テキストアノテーションを BCCWJ コアデータのサンプルごとに施すことのメリットとしては以下が挙げられる。まず、フレーム意味論に基づく意味タグ付きコーパスが作成できる。また、BCCWJ のサンプルごとに、意味フレーム（すなわち語義）の分布や、結合価パターン、ゼロ代名詞の分布などを詳細に調べることができる。将来的には BCCWJ コアデータに対する他の体系に基づくアノテーションと比較・統合することも可能となる。

３．全文テキストアノテーション Web Report

全文テキストアノテーション作業は､語彙アノテーション作業同様に JFNDesktop という、英語フレームネット用に開発されたアノテーションツールを移植・日本語化したツールを用いて行っている。図 1 は、JFNDesktop 上の全文テキストアノテーションモードでアノテーション作業を行っているところである。

アノテーション結果閲覧ツールに関しては、全文テキストアノテーション結果閲覧ツール（全文テキストアノテーション Web Report）は、語彙アノテーション結果閲覧ツール（語彙アノテーション Web Report）とは別に開発した。図 2 は全文テキストアノテーション Web Report のトップページである。BCCWJ コアデータ・テキストのうち、冒頭 10 行の全文テキストアノテーションが終了したものが表示されている。この画面でアノテーション結果を閲覧したいテキスト名をクリックすると、そのテキストのアノテーション結果が表示される（図 3)。図 3 は BCCWJ コアデータの書籍ジャンル内のテキストへの全文テキストアノテーション結果を表示したものである。青字で表示された語彙項目に対して付与された意味フレーム名の名称がその語彙項目の右下に表示されている。ちなみに意味フレーム名は英語フレームネットにおける意味フレーム名と同じものを用いており、英語で表示されている。自立語のうち青字で表示されていないものは、全文テキストアノテーション対象外（代名詞、固有表現など）のもののほか、まだ該当する意味フレームが英語フレームネットで未定義のものが含まれる。

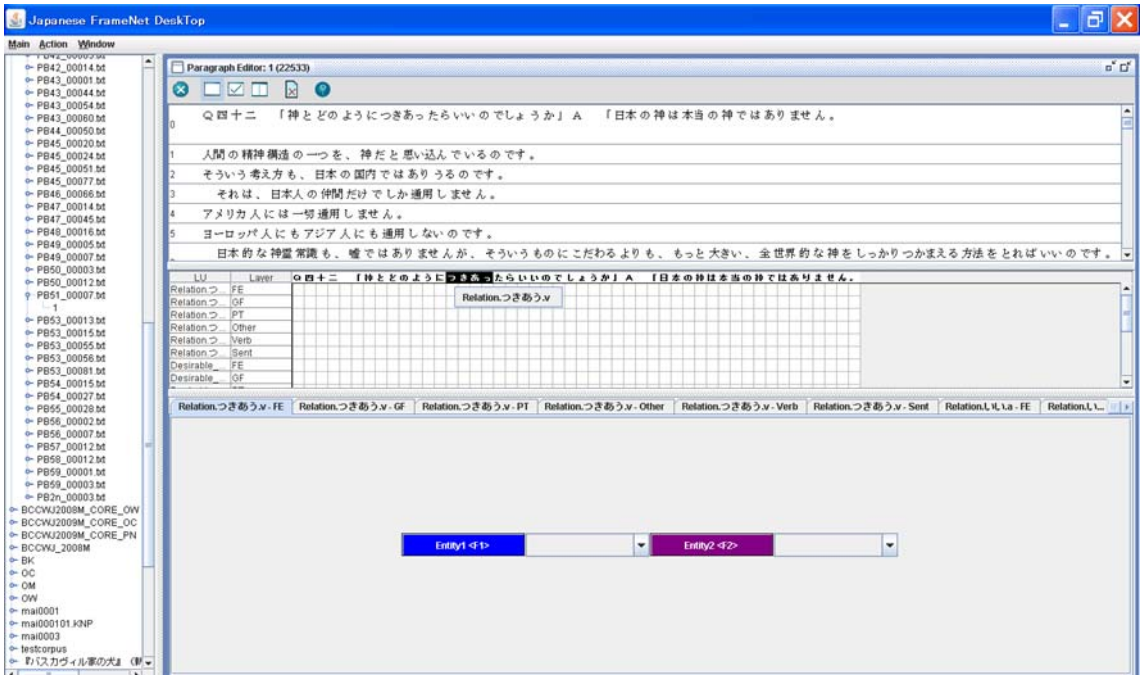


図１　JFNDesktop 上での全文テキストアノテーション作業画面

全文テキストアノテーションWeb Report_ver.2011.01.30

BCCWJ2008M_CORE_BK

- PB55_00028.txt (弁理士が答える知って得する知的財産権Q&A)
- PB40_00035.txt (宝石函)
- PB36_00006.txt (〝田舎〟社長の成功経営術)
- PB40_00003.txt (SEとして生き抜くワザ)
- PB51_00007.txt (とこしえの命を得るために)
- PB58_00012.txt (語源を楽しむ)
- PB49_00005.txt (螢の舟)
- PB11_00006.txt (ひとりの小さなおともだちが)
- PB45_00051.txt ([新編]住居論)
- PB50_00003.txt (ザ・エージェント)
- PB37_00050.txt (尼崎相撲ものがたり)
- PB43_00060.txt (企業の社会的責任)
- PB53_00015.txt (教養教育は進化する)
- PB25_00063.txt (いえづくりをしながら考えたこと。)
- PB39_00009.txt (五十メートルの戦記)
- PB43_00001.txt (授業力)
- PB54_00015.txt (医師による切らない「赤アザ・赤ら顔(浮きでた青い血管)」の最新治 療)
- PB59_00003.txt (天の前庭)
- PB26_00043.txt (犬と話ができる！)
- PB12_00001.txt (闇を歩く)
- PB33_00037.txt (子どもの感性がつくる理科授業)

図２　全文テキストアノテーション Web Report

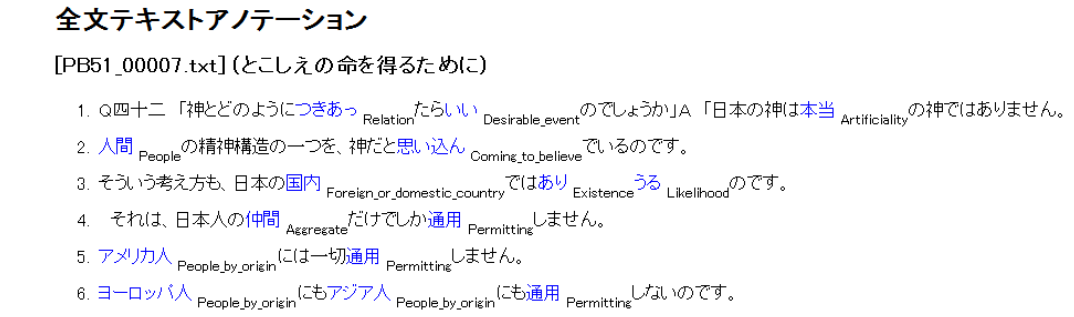


図３ 全文テキストアノテーション結果表示画面

４．英語フレームネット上の意味フレームの適合率

日本語フレームネット班では、まず英語フレームネットの英語語彙分析のための意味フレーム定義が日本語語彙分析にも適用できるかを検討し、英語フレームネット上に適切な意味フレームが存在しない場合には、i) 英語フレームネット上でたまたま未定義なだけなのか、ii) 英語の語彙分析には不要だが日本語語彙の意味分析には必要な意味フレームなのか、を考察している。

この方針を全文テキストアノテーションにも適用し、英語フレームネット上の意味フレームがどの程度 BCCWJ コアデータ書籍ジャンル上の語彙記述に用いることができたかを調べた。その結果、書籍ジャンルのサンプルにおける英語フレームネットの意味フレームの適合率は平均 82 パーセントであった。適合率の算出に当たっては、異なり語（type）ではなく延べ語（token）を用いた。

BCCWJ コアデータ書籍ジャンルのサンプルにはフィクションとノンフィクションの両方が含まれるが、概してノンフィクションの方がフィクションより適合率が低かった。ノンフィクションで平均 81 パーセントにとどまったのに対し、フィクションでは平均 90 パーセントであった。

５．日本語固有の意味フレーム

前節でふれたように、サンプル上に出現する日本語の語彙項目の意味を表すのに適切な意味フレームが英語フレームネット上に見つからなかった場合、i) 英語の語彙分析にも必要だが英語フレームネット上でまだ定義されていないだけなのか、ii) 英語の語彙分析には不要だが日本語語彙の意味分析には必要な意味フレームなのか、を検討した。その結果、適切な意味フレームが英語フレームネット上で見つからないケースのほとんどは i) であり、ii) は稀であることがわかった。異なり語 192 語のうち、ii) に該当するのは 4 語（「畳」、「障子」、「襖紙」、「侠客」）のみにとどまった。i) の中には、「実際のところ」、「もちろん」、「もっとも」などの文副詞、「だから」、「しかし」、「ならば」などの接続詞が含まれていた。英語フレームネットでは副詞や接続詞のアノテーションがまだ進んでいないことが原因と考えられる。

６．アノテータ間の意味フレーム名付与の一致率

複数アノテータが付与した意味フレーム名がどれだけ一致しているかを調べた。全文テキストアノテーション作業においては、まず、第一段階として、通常主に技術翻訳に従事しているプロの翻訳者に BCCWJ のサンプル上の日本語語句の文脈を考慮した英訳を考え

てもらい、その英語語句を英語フレームネットデータベースで検索し、元の日本語語句にふさわしい意味フレーム名を同定してもらった。第二段階では、日本語フレームネットの語彙アノテーション作業経験が 1 年以上のアノテータに第一段階の翻訳者によるアノテーション結果を検討してもらった。さらに第三段階では、筆者が最終的な意味フレーム名の同定を行った。その結果、第一段階と第三段階とでは意味フレーム名の一致率が平均 58 パーセント、第二段階と第三段階とでは一致率は平均 67 パーセントであった。このように複数アノテータが付与した意味フレーム名の一致率が比較的低いことは、日本語フレームネットによる意味フレーム名付与作業がかなり高度であることを示唆している。また、意味フレーム同定に当たって英語フレームネットのデータに照らし合わせる必要があることも関係していると考えられる。

7．おわりに

以上、本稿では日本語フレームネット班における BCCWJ コアデータ書籍ジャンルへの意味フレーム名の付与作業について報告した。英語フレームネット上の意味フレームの適合率については平均 82 パーセントであった。さらに、今現在までのアノテーション作業においては日本語の語彙意味分析のために固有の意味フレームを定義しなければならないケースはさほど見当たらなかった。今後も日本語固有の意味フレームとはどのようなものかについて検討していく必要がある。また、アノテータ間の意味フレーム名付与一致率を向上させるにはどうすればよいのかも考えていくべきである。

付記

本稿は、『言語処理学会第 17 回年次大会予稿集』に掲載した小原(2011)の一部を書き改めたものである。本稿で報告した全文テキストアノテーション作業にあたり、多大なるご協力をいただいた日本語フレームネット班研究協力者の木越壽子氏、李陽氏、並びに前木香織氏とアレクサンドル・カバッシュ氏に御礼申し上げる。

主要文献

小原京子(2011) 「日本語フレームネットの全文テキストアノテーション：BCCWJ への意味フレーム付与の試み」，言語処理学会第 17 回年次大会予稿集.

Fillmore, Charles J. and Collin Baker (2010). “A frames approach to semantic analysis.” In Heine, Bernd and Heiko Narrog (Eds.) *The Oxford Handbook of Linguistic Analysis*. pp.313-339. Oxford University Press.

Hasegawa, Yoko, Russell Lee-Goldman, Kyoko Hirose Ohara, Seiko Fujii, and Charles J. Fillmore (2010). “On expressing measurement and comparison in English and Japanese.” In Boas, Hans C. (Ed,) *Contrastive Studies in Construction Grammar.* pp.169–200. Amsterdam: John Benjamins Publishing.

Ohara, Kyoko Hirose and Hiroaki Sato (2010). “Investigating Japanese FrameNet Data with FrameSQL.” Sixth International Conference on Construction Grammar (ICCG-6). Charles University, Prague, Czech Republic. September 5th, 2010.

Tagami, Hayato, Shinsuke Hizuka, and Hiroaki Saito (2009). “Automatic Semantic Role Labeling based on Japanese FrameNet - Progress Report -.” *Proceedings of Conference of the Pacific Association for*

Computational Linguistics (PACLING2009), Hokkaido, pp.181-186.
(http://jfn.st.hc.keio.ac.jp/publications.html よりダウンロード可能)

関連 URL

日本語フレームネットホームページ：http://jfn.st.hc.keio.ac.jp/ja/index.html

FrameSQLで見る日本語フレームネット

佐藤弘明（日本語フレームネット班分担者：専修大学商学部）†

Browsing Japanese FrameNet with FrameSQL

Hiroaki Sato　(School of commerce, Senshu University)

１．研究成果DVDのデータとFrameSQL

日本語フレームネット班[1]は、『現代日本語書き言葉均衡コーパス』（BCCWJ）[2]から抽出した用例に意味情報の入力を行ってきた。入力作業は、語彙項目ごとにBCCWJの中から例文を選びタグ付けしていく語彙項目アノテーションと、特定のサンプルテキスト内のすべての語彙項目に対してタグ付けを行う全文テキストアノテーションの2つモードで行われた。語彙項目アノテーションの入力結果は、特定領域「日本語コーパス」研究成果DVDの中の日本語フレームネット班データファイル(以下、JFN成果データ)に収録した。

FrameSQL[3]は、JFN成果データの検索をWebブラウザから行うソフトウェアである。FrameSQLは、JFN成果データを検索対象として利用するだけではなく、検索メニューの作成にも利用している。本稿では、JFN成果データに含まれるそれぞれのファイルの内容を紹介し、それらがFrameSQLの中でどのように利用されているかを述べる。

２．JFN成果データの5つのファイル

JFN成果データは(1a-e)の5つのファイルから構成される。

(1)　a. jfndata100816.xml
　　b. annotationV1_3.dtd
　　c. jfn_frame_def100816.zip
　　d. FECOLORS100816.txt
　　e. furigana_hepburn100816.txt

(1a)が主要データファイルであり、アノテーション結果をXML形式で収録している。アノテーションの中で特に重要なフレーム要素（いわゆる「意味役割」）は、(2)のように入力されている。

(2)
```
<layer ID="1" name="FE" rank="1">
    <labels>
      <label name="Goods" ID="453" start="31" end="50" />
      <label name="Time" ID="1747" start="0" end="27" />
    </labels>
</layer>
```

† thc0408@gmail.com

<label>タグのname属性にフレーム要素の値が入る。それぞれのフレーム要素が文のどの語句に与えられるかを示すために、start属性とend属性の値に語句の文中での開始文字位置と終了文字位置が入力されている。

アノテーションは、(2)にもある<layer>タグによって4層に分けられており、その1つの層でフレーム要素のデータが入力されている。他の3つ層には、主語や目的語などの文法機能が入力されている層、名詞句や動詞句など句タイプが入力されている層、助詞の種類が入力されている層とがある。

(1b)は(1a)の文書構造を定義しているDTDファイルである。FrameSQLは(1b)のファイル自体は利用していない。

(1a)で入力されたそれぞれの語彙項目は、それが喚起するフレーム（背景基盤）と結びつけられている。フレームとフレーム要素の定義は、フレームごとに分かれたhtmlファイルの中で記述されており、すべてのhtmlファイルは1つのファイル(1c)にまとめて納められている。(1c)は、英語フレームネット研究組織[4]が作成したファイルに日本語の用例などを加えて作成しているが、日本語特有のフレームとフレーム要素は日本語フレームネット班が独自に定義している。

３．FrameSQLの特色

FrameSQLは、(1a)に含まれるすべてのフレーム要素、文法機能、句タイプを取り出し、それらから検索メニューを作成している。検索メニューではフレーム要素、文法機能、句タイプの3つの層のそれぞれの要素が画面上に表示される。ユーザーは、そのメニューから検索項目を選べば、(1a)のデータ構造を理解していなくても(1a)の検索を行うことができる。

日本語フレームネット班は、多くのフレーム要素を定義して意味情報の入力を行っている。そのため、データの入力結果を理解するには、それぞれのフレーム要素の定義を理解しなければならない。FrameSQLは、(1a)で入力されているフレーム要素を対応する(1c)の定義ファイルにリンクさせている。そのため、ユーザーは用例の中に現れるフレーム要素名をクリックすると、その定義が即座に参照できる。

FrameSQLでは、用例に現れるそれぞれのフレーム要素を人間に見やすく表示するために、(1d)で指定されている特定の色で表示する。(1e)は、(1a)に登録されている語彙項目とその読みをローマ字で表記したリストである。FrameSQLには、このリストを利用して語彙項目の検索をローマ字入力でも行う検索画面があるため、日本語が十分に理解できない海外の研究者でも(1a)の検索が行える。

関連URL

[1] 日本語フレームネットホームページ：http://jfn.st.hc.keio.ac.jp/ja/index.html

[2] 特定領域「日本語コーパス」ホームページ：http://www.tokuteicorpus.jp/

[3] FrameSQLホームページ： http://sato.fm.senshu-u.ac.jp/jfn23/notes/index2.html

[4] 英語フレームネットホームページ： http://framenet.icsi.berkeley.edu/

BCCWJを用いた語彙・構文彙の分析
―所謂引用助詞「と」が標識する構文の場合―

藤井 聖子（日本語フレームネット班分担者：東京大学大学院総合文化研究科）†

An Analysis of Lexicon and 'Constructicon' Using BCCWJ: The Case of Quotative TO Constructions

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

１．はじめに

国立国語研究所による『現代日本語書き言葉均衡コーパス』の開発・構築により、日本語の大規模均衡コーパスを活用した分析が増々活性化されつつある。本稿では、『現代日本語書き言葉均衡コーパス』(BCCWJ 2009 領域内公開版/一部2008領域内公開版と一致；国立国語研究所）を用いて、所謂引用助詞「と」が標識する構文の意味・用法を、構文の多層性とそれらの構文に参与する語彙という観点から分析してきた研究の概要・一部を報告する。

本研究の目的は、多機能性をもつ引用助詞「と」が標識する構文（「引用ト構文」）に関して、コーパスに基づき、引用ト節(句)と共起する語彙群と構文を示し、引用ト節(句)が参与する多層的構文の類型・繋がりを明らかにし、引用ト節(句)を喚起する語彙群とともに、構文群を体系的に記述することである。この目的のために、フレーム意味論・フレームネットにおける分析手法や概念に依拠した分析を行った。このような分析が、コーパスからの抽出文に複数層のアノテーションを付して電子資源化する日本語フレームネット構築プロセスにおいて、既存の日本語分析の知見では未だ明確なアノテーション付与基準が得られていない言語現象の問題に対する必要不可欠な分析でもある。

本分析を、文法的機能語が標識する構文に関する語彙・構文の統合的分析の一事例として提示し、コーパスに基づく「構文彙」・語彙の統合的分析と記述資源の構築を目指す意義を示し、その共同目的に向けての雛形一分析とする。「構文彙」は、Fillmore (2006, etc.) が構想を示し構築を提言した構文知識の集合体'Constructicon'の著者による邦訳である。[1]

２．背景と本研究でのアプローチ・観点

日本語の引用に関しては、日本語学において多くの優れた研究が展開してきた(砂川 1987, 1988, 1989 等; 藤田 1986, 1988, 2000 等；他)。その中で、所謂引用助詞「と」の用法が多岐に渡ることが指摘され分析され（国立国語研究所 1951, 他)、特に山崎(1993)に、用法の体系的分類に関して非常に示唆に富む再整理と提案・提示がある。本研究も、山崎(1993)の洞察に依拠するところが大きい。さらに、引用節が主節述部を伴わずに使用される現象についても、特に話し言葉の研究で活発に分析されてきた（Okamoto 1996, 加藤 1998, 2008, 2010, S. Suzuki 1995, 1996, R. Suzuki 1999, Fujii 2002, 2006, 山崎 1996, 等)。

本研究では、述部を伴って使用される所謂引用助詞トの用法に焦点を絞り、ト標識節(句)の用法を、ト標識の節(又は句)を受ける述部の述語の特徴、および、ト節(句)の述部との関係、という観点で分析する。助詞「と」自体が単独で様々な意味機能を担うとみなすので

† sfujii@boz.c.u-toko.ac.jp

[1] 'Constructicon'に関しては Fillmore (2009), Fillmore, Lee-Goldman & Rhodes (2010)等参照。

はなく、助詞トが参与し標識する（さらにト標識節(句)が参与する）「構文」(「引用ト構文」と呼ぶ）を分析の標的とし、ともに参与して構文を特徴付ける主幹語彙群を明らかにするとともに、構文の意味的・形式的特徴を構文類型ごとに浮き彫りにすることを目指す。これらの意味で、「構文と語彙」の分析・記述を目指す立場をとる。

ト標識節の構文上の捉え方に関しては、「言う」「思う」などの動詞の文構成必須要素・補文と捉える捉え方（仁田 1982 等国内の研究；海外での研究の多く）がある一方、藤田(2000,1986)は、「副詞的な成分である」と捉えるべきものであるという貴重な示唆に富む提言をしている。英語などの補文を構成する that 節と対応しうるト節の一部の用法に基づいて日本語の引用ト節を捉えることには根本的な問題があるという警鐘でもある。本研究も、この藤田(2000,1986)の重要な知見を鑑み踏襲しつつ、多様な構文に参与する引用ト節(句)を一様に捉えるのではなく、引用ト構文の多層性と広がりを分析し明示することの重要性を踏まえ、後節で提示する【内の関係】と【外付けの関係】とを峻別しつつ分析を進めた。

【内の関係】用法と【外付けの関係】用法とを峻別することは、構文の内部構造と構文の多様性・広がりを捉えるために重要であるだけでなく、コーパス分析の構想や手法を多角的にし、ぞれぞれの用法に適した分析を行うために必要である。具体的に言えば、【内の関係】用法と【外付けの関係】用法それぞれに対して、大規模コーパスから該当データを抽出するデータ処理の方法が異なり、データを用いた分析の手法や目的・内容が異なる。

３．本分析に使用した BCCWJ サブコーパス

本稿で報告する分析において分析対象とした『現代日本語書き言葉均衡コーパス 2009/2008 領域内公開版』(BCCWJ)の８つのサブコーパス（白書、書籍７サブコーパス)、および、それぞれのサブコーパスにおける総語数、引用トの生起数、それらから手作業で抽出した引用ト構文の外付け用法の用例数を、以下表１に挙げる。

表１．分析対象としたBCCWJのサブコーパス、総語数、引用ト構文の生起数、外付け用法用例数 [2]

サブコーパス	コーパス語数	引用ト構文生起数	外付け用法 生起数
白書	5,000,000	17,513	2
書籍:文学 （文学全体の約72%）	6,324,175	84,149	272
書籍:言語	398,497	6,244	27
書籍:技術工学	1,115,821	11,733	45
書籍:自然科学	1,074,332	13,684	58
書籍:哲学	1,403,199	19,623	57
書籍:歴史	2,141,841	27,846	118
書籍:総記	521,436	6,515	25
合計	17,979,301	187,307	604

『BCCWJ 領域内公開版』収録のテキストファイルを入力データとし、テキストファイルを形態素解析した上で、使用コーパス全体から検索ツールで該当構文を含む文をすべて抽出した。処理には、形態素解析ツール MeCab (http://mecab.sourceforge.net/,工藤拓, 松本裕治)、及び、検索ツール ChaKi (http://chasen.naist.jp/hiki/ChaKi/,松本裕治、等)を用いた。右欄「外付け用法」用例は、中欄データ全体（引用助詞トを含む文を検索した出力データ）から、データ全文を一例ずつ(人間)読解により識別し手作業で一例ずつ抽出した。

[2] ChaKi による出力データに対して、人手で形態素の誤解析削除とコーディングを加えたが、表１の中欄の引用ト生起数には、共格や接続助詞などト形態素の誤解析検出トークンも多少含まれている。

４．所謂引用の助詞トが参与するト構文の構文類型（作業分類）と基軸指標

本稿で報告するコーパス分析では、所謂引用の助詞「と」の用法・分類に関する国立国語研究所(1951)、藤田(1986)、特に山崎 (1993)の洞察・提案を鑑み、引用ト構文の用法の大別を以下の大分類とした。表２に示す作業分類に基づき、手作業でコーディングをした。

本大分類では、以下二つの分類軸を概念的に想定している。

４．１　分類軸１：ト節(句)の主述部との関係【内の関係】vs.【外付けの関係】

引用ト節(句)は、主節の文構成要素・主動詞の項構造の項として位置付けられる用法【内の関係】(例：1)と、主節の項構造の外に位置付けられる副詞節的用法【外付けの関係】(例：3)とがある。また、本研究のコーパスデータの分析では、これら両者の中間に位置づけられるべき用法も認められた（例：２）。

(1) Q1.【内の関係】:　「言語表現補語引用型」
- 分かったと言った。・分かったと思った。・分かったと書いた。分かったと喜んだ。
- いっときお互いにいい思いをしたと考えれば、あきらめがつくわな。
- 最近はそういう手術は難しくないと聞きましたが…

(2) Q1-2. 中間的用法：「言語表現引用同一事態型」
- 分かったとうなずいた。　・分かったと首を縦に振った。
- 「お先に失礼します」と頭をさげた。
- 若ハゲの男も「こんばんは」とウインクする

(3) Q2.【外付けの関係】:「言語表現外付け引用事態型」
- 分かったとバナナを手渡した。　・分かったと飛び出て行った。
- インフルエンザの疑いがあってはいけないと、クリニックにいきインフルエンザ検査をしました。
- ポールは「体のなかを清浄する」と、グリーンティを愛飲しているのだ。
- 「…　一二粒はありますやろ」とダンボール箱に掌を入れる。

４．２　分類軸２：狭義の引用用法と狭義の引用ではない用法

（紙幅制限のため、本稿では、狭義の引用ではない用法の分類 P1-P3 の例示を割愛する。）

表２．　引用ト構文の用法・構文類型に関する、本分析における作業分類

引用「ト」構文の用法大別		構文の類型	ト標識引用部	述部	第一段階コーディング	述部の例	
Q. 引用の用法							
	Q1.	**言語表現補語引用型**	言語表現の内容	言語表現を内容とする事態	発話思考		
					発話思考:発話	いう、話す	主張、回答、説明
					発話思考:思考	思う、考える	判断、予想
					感情	驚く、喜ぶ	心配、後悔
Q1とQ2との中間	Q1-2.	**言語表現引用同一事態型**	言語表現の内容	言語表現を内容とする事態	行為:感情	笑う、泣く	微笑み、号泣
			言語表現の内容		行為:言動	うなずく	（電話、メールはQ1とした）
	Q2,	**言語表現外付引用事態型**	言語表現の内容	言語表現を内容としない事態	事態:外付け	様々な述語（立つ、手渡す、等）	様々
	Q3.	**非言語引用事態型**	非言語的表現の内容	非言語的表現を内容とする事態	事態:他	様々な述語	様々
P. 狭義の引用ではない用法							
	P1.	**変化結果提示型**	非言語表現	変化結果	結果:変化結果	化す、なる	変化
				移動結果	移動:移動結果	進む、向かう	移動、
	P2.	**尺度属性型**	非言語表現	尺度・測定・比較	尺度関連	高まる、増える	増加、減少、低迷
	P3.	**列挙型**	非言語表現		列挙	続く、並ぶ	
参考:形態素解析(MeCab)において一貫して「引用助詞」と誤解析された形態素					誤:共格		
		頻繁に「引用助詞」と誤解析された形態素			誤:接続助詞		

５．　フレーム意味論・フレームネット[3]における【内の関係】【外付けの関係】の捉え方

本研究では、主要フレーム喚起語が何かをまず捉え、さらに、語彙または構文が喚起する各々の意味フレームにおいて引用ト節(句)がどのようなフレーム要素として機能するかを分析した。「喚起される意味フレームにおいて引用ト節(句)がどのようなフレーム要素として機能するか」という観点での分析は、フレーム要素の二つの分類軸で行った。

５．１　フレーム要素のアノテーション（その１）

一つは、参与項目のフレームにおける意味役割（一般的な「意味役割」に相当するものであるが、フレーム意味論/フレームネットにおいては、各々の意味フレームにおいて想定され定義されるきめ細かな事態構成要素の意味）の記述である。この要素意味役割のアノテーションでは、【内の関係】では、引用ト節(句)が MESSAGE, CONTENT, DECISION, LABEL, REASON 等のフレーム要素として機能することが認められた。

５．２　フレーム要素のアノテーション（その２）

もう一方は、該当フレーム要素がフレーム喚起語や喚起されるフレームとどのような関係にあるかに関する分析である。フレームネットでは、この観点において、三種類のフレーム要素の関与レベルを設定している：(i) Core frame elements (FE) コア;　(ii) peripheral FE 周辺的;　(iii) extrathematic FE 主題外フレーム要素 である。フレーム喚起語に直接喚起され、喚起されるフレーム事態に内在的に参与するフレーム要素が Core FE または peripheral FE である。中でも Core FE は、喚起されるフレーム事態において概念的に必須の参与要素であり（常にではないが）多くの場合、統語的にも選定される要素である(項構造の必須項)。一方、extrathematic FE は、フレーム喚起語に直接喚起される要素ではなく、様々なフレームと共起し、該当フレームとは別の事態を導入しつつ、該当フレームの参与者や状況を補足的に描写するフレーム要素である(Ruppenhofer et al. 2006: 135-136)。FrameNet において、extrathematic FE は多くの副詞節や副詞句の位置づけに用いられている。

５．３　フレーム要素としての引用ト節/句のバリエーション

本研究では、このフレーム要素の喚起フレームとの関わり方(Core FE, peripheral FE, extrathematic FE)という観点で、引用ト構文の【内の関係】と【外付けの関係】との峻別を捉え直し、引用ト節(句)が文全体・構文全体の中で、(主節の)フレーム喚起語や喚起されるフレームとどのような関係にあるかを、コーパスデータから抽出した用例で分析した。

5.3.1　Core Frame Element　コアフレーム要素

【内の関係】の引用ト節(句)は、多くの場合、フレーム喚起語が喚起するフレームの主要フレーム要素 Core FE として位置づけられる。その Core FE の代表例が、「話す」「述べる」「唱える」等が喚起する Statement フレームにおける MESSAGE フレーム要素である。同様に Communication フレームの下位フレームである Communication_manner フレームや Request フレームにおいても、ト節(句)が MESSAGE フレーム要素となる。その他、「喜ぶ」「驚く」「困る」等が喚起する Experiencer_subj フレームにおける CONTENT フレーム要素；「決める」等が喚起する Deciding フレームにおける DECISION フレーム要素、

---

[3]　フレーム意味論、およびフレームネットに関して、紙幅制限のため本稿で解説を含むことができないが、藤井&小原(2003), 小原(2006), 小原他(2005a, 2005b), 本特定領域研究 2008 年度公開ワークショップ予稿集(国立国語研究所)における拙著等を参照されたい。

等である。同じ【内の関係】用法における引用ト節(句)でも、フレーム喚起語が異なれば、喚起されるフレームも異なり、そのフレームにおいて喚起されト節(句)で言語化されるフレーム要素は、このように様々な意味役割を担っている。

5.3.2 Extrathematic Frame Element 主題外フレーム要素

一方、副詞節的【外付けの関係】では、ト標識引用節を extrathematic FE として位置づけることができる。副詞節的引用節は、意味的に主節に関与しているが、主節の内在的構成要素ではなく、主節述語によって要請されたフレーム構成要素ではない。

(4) 「やれるだけやってみます」ともうすぐ早朝点呼の始まる寮室へ戻っていった。

5.3.3 中間的用法

しかし、抽出したコーパスデータを吟味すると、副詞節的用法の中で、extrathematic FE とみなしてよい用例の他に、中間的なもの（Q1-2)「言語表現引用同一事態型」に多々遭遇した。 (5)の例は、主節の述部が喚起するフレーム事態とは別の事態をト引用部が導入しているわけではなく、主節の事態と同一事態を表象しており、発話言語表現を補足同格的に添えることにより、主節の提示する行為の様態描写を精巧化している。従って、このような中間的な Q1-2 でのト標識節は、extrathematic FE ではなく、peripheral FE（ここでは MANNER や CIRCUMSTANCE フレーム要素）と位置付けられる。

(5)・美女が「ニイハオ」と**出迎えて**くれる。 ・「お先に失礼します」と**頭をさげた**。
・「よし、それでよい」と**合格点をつけた**はずだ。

5.3.4 他のPeripheral Frame Elements 周辺的フレーム要素

主動詞の喚起するフレームに直接参与するフレームであるという点で【内の関係】と捉えられる用法の中には、core FEというより、peripheral FEと位置づけられる場合も多々ある。

(6) P2. 財産被害が、19.7%と多くなっている。

(7) P2. 睡眠時間がハイティーンでは男子で１時間11分と短くなっている。

(6)(7)において、尺度構文を構築する語彙「多い」「短い」によって喚起される Gradable_attributeフレームにおいて、「19.7%と」「１時間11分と」は、REFERENCE_POINT フレーム要素として、「多い」「短い」に関する具体的な値を指定する。この尺度構文において、主語（「財産被害」「睡眠時間」）ENTITYフレーム要素はCore FEであるが、それに任意的に参与して具体的な値を指定する「19.7%と」「１時間11分と」REFERENCE_POINTは peripheral FE と位置付けられる。（これらの分析に関しては、Fujii 2009参照。）

5.3.5 引用ト節(句)のバリエーション

以上、引用ト節/句が、参与する構文によって、３種類のフレーム要素 —Core FE, peripheral FE, extrathematic FE — のいずれにもなりうることを明らかにした。このことからも、引用ト節(句)の機能をそれが参与するそれぞれの構文の中で捉え記述することが不可欠であり、文法的機能語トの記述に構文的アプローチが必要であるといえる。[4]

[4] フレームネット構築においては、有限のアノテーション値を用いて一貫したアノテーションを付与する必要があり、すべての言語現象に忠実かつ正確なアノテーション付与をめざしつつも、実装段階では仕様・範疇判断になるのも、データベース作成上の実情である。しかし、肝心なのは、「実際のコーパス分析に基づいて、コーパス用例の精査分析とその分析の記述を重ねた上で、コーパスアノテーションの範疇判断に関する最適な決定に到達することができる」という(Fillmore氏率いるフレームネットの)基本姿勢である。

６．共起語彙群と共起語彙が喚起するフレームの分析：【内の関係】Q1(一部Q1-2)の場合

述部のフレーム喚起語が喚起するフレームに関して、白書、書籍文学、書籍言語、書籍技術工学４つのコーパスでの、ト節(句)と共起する動詞・事態性名詞それぞれ頻度上位 100 語（延べ 800 語）が喚起する意味フレームを分析した（図１の下位フレームを参照)。

さらに喚起するフレームごとにフレーム喚起語彙の分類語彙表を作成した。その主要フレームとその高頻度フレーム喚起語の一部を、表３【動詞】と表４【事態性名詞】に示す。

表３ 引用ト構文【内の関係】Q1(一部 Q1-2)におけるフレーム喚起語【動詞】とそのフレーム

Statement Communication
言う, 記す, 仰せる, 言い張る, 語る, 述べる, 書く, 唱える, 唱う, おしゃる, 付け加える, 申し上げる, 表わす, 話す', みとめる
Questioning Communication(uses)
問う, 聞く, 尋ねる
Request Communication(uses)
求める, 命じる, 聞く, 誘う, 頼む
Becoming_aware Perception
気がつく, 気付く;
Desiring Emotions,Experiencer_subj (uses)
願う, 望む
Experiencer_sub Emotions (Uses)
喜ぶ, 驚く, 困る, 怒る
Expectation Awareness
見込む
Facial_expression Body_movement(uses)
笑う
Coming_to_believe Event, Mental_activity(Uses)
悟る, 学ぶ
Cogitation Mental_activity (Uses)
考える, 捉える, みえる, 考える, 思う, 存じる、する
Awareness Mental_activity,Information (Uses)
分かる, 解す, 信じる, 知る
Deciding Intentionally_act(Uses) > Event
決める, 思い込む, 定める

表４ 引用ト構文【内の関係】Q1(一部 Q1-2)におけるフレーム喚起語【事態性名詞】とそのフレーム

Statement Communication
コメント, ノート, 記載, 記述, 口述, 警告, 言明, 公言, 主張, 紹介, 説明, 宣言, 断言, 注意, 提言, 発言, 発表, 否定, 報告, 報道, 明言, 注記
Questioning Communication (uses)
質問
Request Communication (uses)
依頼, 命令, 提案, 注文, お願い
Becoming_aware Perception
認識
Facial_expressionBody_movement(uses)
苦笑
Judgement_communication Judgement/Statement 批判, 批評
Desiring Emmotions, Experiencer_subj (uses) 願い, 望み
Experiencer_sub Emotions (Uses)
後悔、反省、納得、意識、恐怖、憤慨、賛嘆、びっくり
Expectation Awareness
見込み, 要請, 期待, 先回り'anticipate',
Coming_to_believe Event, Mental_activity(Uses)
推察, 推測, 推理, 断定, 実感
Cogitation Mental_activity (Uses)
察し, 思案, 思慮
Awareness Mental_activity,
理解, 認識 **Information** (Uses)
Deciding Intentionally_act (Uses)>
決意, 決定, 判定, 決心 **>Event**

７．フレーム間関係による共起語彙群の体系付け

ト構文の主なフレーム喚起語群とそのフレームをより体系的に捉えるために、継承(Inheritance)・使用(Using)・サブフレーム(Subframe)というフレーム間関係により上位フレームを同定し、フレーム喚起語が喚起するフレームからのスケールアップで、(最)上位フレームを同定した。図１は、このフレーム間関係の分析結果を示した要約図であり、Q1 の用法に関して喚起フレームがどの上位フレームに集約されるかを示す図である。図１の上層部に示すとおり、Information and Topic (>Communication), Reciprocality, Emotions > Judgment, Event >Intentionally_act, Mental_activity, Attempting_scenario, Perceptions, Evidence, 等の上記フレームに集約されている。意味フレームごとにグループ化した語彙群を、フレーム間関係に基づいてさらに上位フレームに集約していくと、多様なフレームを喚起する多様な語彙が自然クラスを構成していることが分かった。

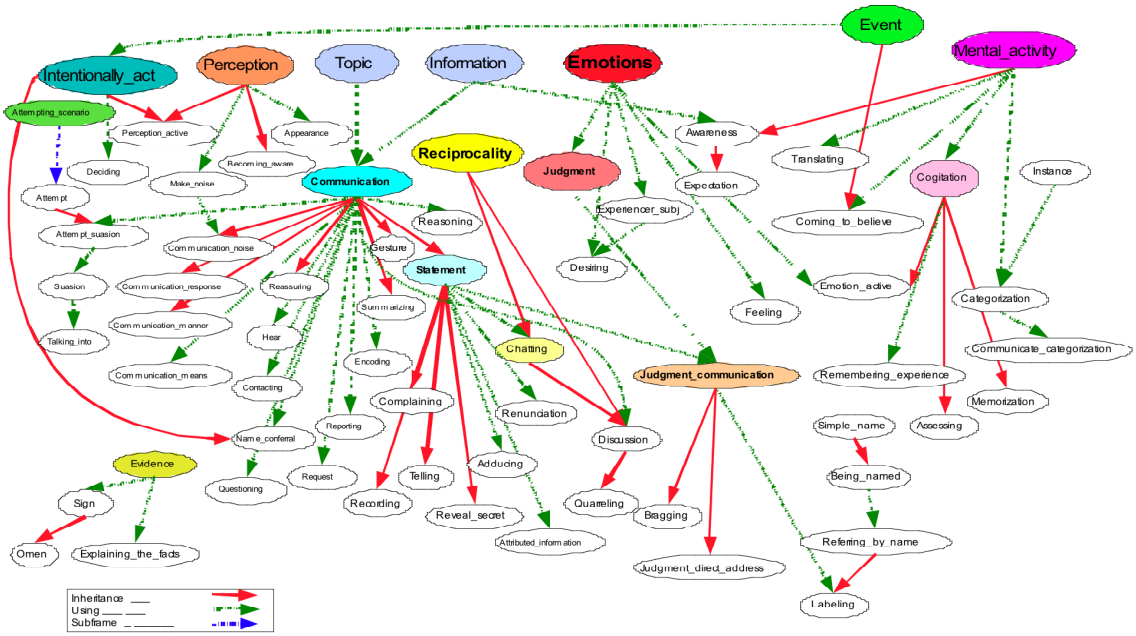


図１．引用ト構文が喚起するフレーム(BCCWJ の本分析で認められたもの)の広がりとフレーム間関係：引用部が言語表現の内容、述部が発話・思考・心的活動・判断・感情に関わるフレームを喚起する用法のみ

しかし、参与する構文タイプが異なる語彙群は、図 1 に含まれない。例えば、P2「尺度関連」は多くの用例が Change_position_on_a_scale (例: 高まる)フレームを喚起するが、これらは図１のフレーム群から離れた意味領域に別の自然クラスを構成する。

以上の【内の関係】用法に共起する語彙群に対して、副詞節的【外付け用法】は主節の述語に関しては自然クラスを構成しない。これらの多様性から、語彙分析・語彙記述自体を参与する構文タイプを考察し、構文タイプごとに行う必要があることが再確認される。

８．副詞節的【外付けの関係】用法の分析

上述の【内の関係】の場合と異なり、【外付けの関係】副詞節的用法に関しては、引用ト節(句)と共起する語彙自体を分析・記述し構築することには前者ほどの意味がなく、構文自体の構文的意味・統語的特徴・語用論的特徴を明らかにする必要がある。

(i) 引用ト標識節(引用部)の主観者(主体者)が、主節の述部の(意味的)主語と一致している；

(ii) 引用部は、(i)で同定されるト節引用部の主体者(即ち、主節述部の行為・事態の主語)の、その行為・事態における発話・心情を表出しており、主節述部の表わす行為の動機を表出することが多い。

３節の表１[最右列]で示したとおり、生起数 187,307 の引用ト構文の用例から(人間)読解・手作業で副詞節的外付け用法を 604 用例抽出し、すべての用例に関して上記作業仮説に基づくコーパス分析を行った。紙幅制限のため、その結果報告と論述は他稿で提示する。

９．おわりに

本研究では、引用ト節(句)が参与する構文の多層性を分析・記述するために、文法的機能語トが標識する構文に参与する語彙群を、コーパスに基づき、参与する構文タイプごとに分析してきた。さらに、構文類型やその語彙群相互の繋がり・体系を明確にするために、フレーム要素やフレーム間関係を分析した。

文法的機能語と共起する語彙群を分析・記述する際、参与する構文タイプごとに考察する必要があり、コーパスに基づく「構文彙」の分析と構築が有用である。第４回国際構文

理論学会(於東京大学駒場)での基調講演において Fillmore (2006)が Lexicon に準ずる ‘Constructicon’構築を提言し(本研究で ‘Constructicon’を「構文彙」と称することにした)、英語においては過去 20 年以上に渡る構文理論での構文分析の蓄積に基づき、構築が始動している。しかし、Fillmore 氏も強調するように、構文彙構築は一つ一つの構文に関する綿密かつ正確な実証的分析・記述の礎が蓄積されて初めて可能になるものであり、理論的・方法論的構想だけで実現可能な内容ではない。特に言語間での枠組みの移植も自動的に前提とすることはできない。本稿で報告した語彙的・構文的アプローチでの引用ト節(句)の分析は、そのような構文彙構築に向けて、語彙情報資源とともに、語彙分析と統合した構文分析に基づく構文彙資源をどのように構築するかを例示し議論するための一事例である。

謝辞

1998 年に国立国語研究所内で引用の日英対照研究に着手した際、井上優氏から貴重な御助言をいただき、同研究所報告書で報告された山崎誠氏の論文(1993)に重要な御教示をいただいた。山崎(1993)の卓越した洞察により、それ以前の日本語観察に道筋と展望を与えていただいた。両氏に深謝する。さらに、2002 年以降日本語フレームネット（代表：小原京子氏）の伝達・判断等関連領域を鈴木亮子氏と共同で担当し分析する中で、鈴木氏との議論・共同分析が大変有益であった。同時に FrameNet・日本語フレームネットの枠組みでの手法に重要な動機をいただいた。FrameNet のメンバー（特に Charles Fillmore 氏, Collin Baker 氏, Michael Ellsworth 氏, Russell Lee-Goldman 氏)、及び、日本語フレームネットの共同研究メンバー（本研究班 小原京子氏、斎藤博昭氏、佐藤弘明氏、他）に記して感謝の意を表する。『BCCWJ2008/2009 領域内版』のデータ処理・整理作業においては、科学研究費補助金(本特定領域および藤井)の支援により、内田諭氏と鈴木陽子氏(東京大学大学院言語情報科学専攻)と平山仁美氏(同教養学部)に御協力・補助をいただいた。深謝する。

文献

荻野孝野、小林正博、伊佐原均（2003).『日本語動詞の結合価』. 三省堂.
小原京子（2006).「フレーム意味論と日本語フレームネット」『日本語学』Vol.25. No.6, pp. 40-52.
小原京子、大堀壽夫、鈴木亮子、藤井聖子、斎藤博昭、石崎俊（2005a).「日本語フレームネット:意味タグ付きコーパスの試み」『言語処理学会第 11 回年次大会 大会論文集』
小原京子、石崎俊、大堀壽夫、斎藤博昭、鈴木亮子、藤井聖子（2005b).「日本語フレームネット概要」『日本認知言語学会論文集第 5 巻(JCLA 5)』, 613-616.
加藤陽子（2008).『話し言葉における引用の研究』東京大学大学院総合文化研究科言語情報科学 博士論文.
加藤陽子（2010).『話し言葉における引用表現―引用標識に注目して』くろしお出版.
国立国語研究所（1951）.『現代語の助詞・助動詞』国立国語研究所.
鈴木亮子（2005).「評価を伴う伝達動詞:『ほめる』・『しかる』・『おこる』の分析」『JCLA 5 巻』, 629-632.
砂川有里子(1987).「引用文の構造と機能―引用文の 3 つの類型について―」『文藝言語研究 言語篇』13 筑波大学文芸・言語学系 pp.73-91.
砂川有里子(1989).「引用と話法」『講座 日本語と日本語教育』4 明治書院 pp.355-387.
藤井聖子, 小原京子（2003).「フレーム意味論とフレームネット」,『英語青年』 Vol.14. No.6.
藤井聖子（2005).「日本語フレームネットにおける「伝達」領域での分析」『JCLA 5 巻』, 625-628.
藤井聖子, 上垣渉（2008).「支援動詞構文における事態性名詞と動詞との項共有と連結性:『日本語コーパス』を用いた分析」『日本言語学会第 136 回大会予稿集』, pp. 432-437.
藤田保幸（1986).「文中引用句「ート 」による「引用」を整理する」宮地裕(編)『論集日本語研究（一）現代編』明治書院.
藤田保幸(2000).『国語引用構文の研究』 和泉書院.
山崎誠（1993）.「引用の助詞「と」の用法を再整理する」『国立国語研究所報告 105 研究報告集14』, pp. 1-29. 国立国語研究所.
山崎誠(1996).「引用・伝聞のッテの用法」『国立国語研究所報告 研究報告集17』国立国語研究所 pp.1-22
Baker, Collin. (2006). “Frame Semantics in Operation: The FrameNet Lexicon as an Implementation of Frame Semantics.” In *The Fourth International Conference on Construction Grammar Plenary Lectures.* pp.34-43.
Fillmore, Charles. J. (2002). Varieties of Support Constructions. A plenary lecture given at the Second International Conferenceon Construction Grammar, Helsinki.
Fillmore, Charles J. (2009). Words, Grammar and Language Understanding. Frames and Constructions Conference.
Fillmore, Charles J, Russell Lee-Goldman, & Russell Rhodes. (2010). The FrameNet Constructicon. In Boas, Hans, C & Ivan A. Sag (eds.), *Sign-Based Construction Grammar,* 283-347. Stanford: CSIL.
Fontenelle, Thierry. (Ed.). (2003). Special Issue: FrameNet and Frame Semantics. *International Journal of Lexicography.* Vol.16, Special Issue 3, Oxford, Oxford University Press.
Fujii, Seiko. (2009). Capturing constructional polysemy via frames and frame-to-frame relations: A lexical analysis of quotative TO constructions. *PACLING2009: The proceedings of the 11th Conference of the Pacific Association for Computational Linguistics.*
Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Chris R. Johnson, and Jan Scheffczyk. (2006/2010). *FrameNet II: Extended Theory and Practice*.(http://framenet.icsi.berkeley.edu/book/book.pdf)

関連 URL

「日本語フレームネット」ホームページ：http://jfn.st.hc.keio.ac.jp/ja/index.html
FrameNet ホームページ：http://framenet.icsi.berkeley.edu/

計画班研究発表

3月16日（水）　14:00〜15:15

日本語教育における初級シラバスの再評価 ―BCCWJにみられた「出現形の偏り」を手がかりに―

▶小林 ミナ

BCCWJ複合辞辞書の仕様・開発・評価

▶近藤 泰弘、坂野 収、多田 知子、岡田 純子、山元 啓史

複数の観点から見た用例クラスタリングに基づく新語義の発見

▶白井 清昭、中西 隆一郎、中村 誠

日本語教育における初級シラバスの再評価
―BCCWJ にみられた「出現形の偏り」を手がかりに―

小林ミナ（日本語教育班分担者：早稲田大学大学院日本語教育研究科）†

Reconsideration of the Elementary Syllabus of Japanese as a Foreign Language

Mina Kobayashi　(Waseda University, Graduate school of Japanese Applied Linguistics)

１．はじめに

『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）を始めとする大規模なコーパスの普及によって，使用実態の側面からの言語記述が可能になった。その研究成果を参照することにより，実際の言語運用が十分に反映されていないなど，日本語教育における現行の初級文法シラバスは，必ずしも十全ではないことが指摘されている（小林 2005a, 2005b, 2009，深田，大曾 2007）。

本発表では，BCCWJ を用いて日本語教育における初級シラバスを再評価する。具体的には，２つのサブコーパス（「白書」および「Yahoo!知恵袋」）で使われたイ形容詞を「出現形の偏り」からみることにより，初級シラバスを見直すことを目的とする。

本発表の結論を先取りして言うと，大きく次の３点になる。

（１）イ形容詞の４つの活用形（後述）が使われる頻度には大きな偏りがある。その傾向は「白書」と「Yahoo!知恵袋」でほぼ同じである。

（２）「白書」でも「Yahoo!知恵袋」でも，少数の初級レベルのイ形容詞が，繰り返し何度も使われている。しかし，「白書」と「Yahoo!知恵袋」では，どの語が好まれるかが異なる。

（３）「白書」と「Yahoo!知恵袋」では，どのようなパターンが好まれるかが異なる。

２．経緯と背景

いまの日本語教育では，イ形容詞をはじめとする用言の活用規則は，すべて初級の早い段階で導入，練習される。たとえば，『日本語能力試験出題基準[改訂版]』には，次のような記述が見られる。

（４）　日本語教育における主な文法事項といえは，
・構文 / 文型
・活用
・助詞・助動詞・接辞など（いわゆる文法的な<機能語>の類）の用法
の三つをあげることかできるだろう。
３・４級レベルでは，このいずれもきわめて重要な学習事項である。
（『日本語能力試験出題基準[改訂版]』p.147）

この指針を承けて，ほぼすべての初級日本語教科書で，動詞，イ形容詞，ナ形容詞といった用言の活用形が提示され，授業では，それらの活用形が十分に定着するまで，さまざまな練習が何度も繰り返して行われる。

† minakob@waseda.jp

小林 (2008) では,「白書」コーパスを用いて，イ形容詞の活用形の出現状況を観察した。なお，ここでいうイ形容詞の活用形とは，次の４つの形をいう。

表１　イ形容詞の活用形

	非過去	過去
肯定	−い 例：小さい，楽しい…	−かった 例：小さかった，楽しかった…
否定	−くない 例：小さくない，楽しくない…	−くなかった 例：小さくなかった，楽しくなかった…

その結果，これらの活用形が使われる頻度には大きな偏りがあることがわかった。それは次のようなものであった。

表２　活用形ごとの使用頻度

	非過去	過去	合計
肯定	1527(82.7%)	238(12.9%)	1765(95.6%)
否定	79(4.3%)	2(0.1%)	81(4.4%)
合計	1606(87.0%)	240(13.0%)	1846(100.0%)

ただ，表２のもととなるデータは，2006 年 10 月時点で領域内で公開されていた「白書」コーパスのプレーンテキスト全文を，「ひまわり」（20061016 同梱版）を用いた文字列検索により得られたもので，用例数も 1846 例に過ぎなかった。また，主節末，かつ，活用形の直後に句点（。）を伴ってあらわれた用例という限られたものであった。その後，検索のためのインターフェイス「中納言」が開発されたことにより，品詞による検索が可能になった。そこで，この偏りが「白書」に特徴的なものかどうかを複数のコーパス，さらに多量のデータに基づいて再度検証し，あわせて，定性的な側面からも考察したい。

３．使用データと検索方法

本発表で用いたデータは，BCCWJ における以下のサブコーパス（短単位データ Ver.1.1.0）である。

サブコーパス	語数（短単位，可変長）
「白書」	4526465 語
「Yahoo!知恵袋」	5068071 語

検索は，「中納言」（Ver.1.2.0 Beta）の短単位検索機能で，［検索項目］→［品詞］，［検索値］→［形容詞］で抽出を行った後，次のような連用修飾，連体修飾（下線部）を除いた。

（５）まちづくりの中に市民団体やＮＰＯの活動をうまく取り入れている。
（OW6X_00099，67190）

（６）蒟蒻を乾燥させあまく味付けしてあるやつです。　（OC08_01016，1080）

（７）雇用機会均等法やその指針の趣旨に則り，好ましくない事態も発生している。
（OW4X_00161，19880）

（８）音質は付属のヘッドホンが良くないことを除けばいいですね。(OC02_04630, 460)

４．結果

３．の検索方法で得られたイ形容詞の用例は，「白書」3998 例，「Yahoo!知恵袋」45088 例である。以下の記述は，これらの用例に基づくものである。

４.１ 用例数の比較

「白書」と「Yahoo!知恵袋」のコーパス規模と用例数は，以下の通りである。

サブコーパス	語数（短単位，可変長）	用例数
「白書」	4526465 語	3998 例
「Yahoo!知恵袋」	5068071 語	45088 例

「白書」と「Yahoo!知恵袋」は，コーパス自体の規模（短単位）はそれほど大きく違わない。しかし，抽出されたイ形容詞の用例数には 10 倍以上の開きがあった。

４.２ 活用形ごとの使用頻度および割合

「白書」と「Yahoo!知恵袋」で使われたイ形容詞を，活用形ごとにまとめた。表３，４として示す。

表３「白書」で使われたイ形容詞

	非過去	過去	計
肯定	3580	297	3877
	89.5%	7.4%	97.0%
否定	119	2	121
	3.0%	0.0%	3.0%
計	3699	299	3998
	92.5%	7.5%	100.0%

表４「知恵袋」で使われたイ形容詞

	非過去	過去	計
肯定	41764	2271	44035
	92.6%	5.0%	97.7%
否定	992	61	1053
	2.2%	0.1%	2.3%
計	42756	2332	45088
	94.8%	5.2%	100.0%

表３，４から，次のことがわかる。

（９） ４つの活用形が使われる頻度には大きな偏りがある。その傾向は「白書」と「Yahoo!知恵袋」でほぼ同じである。

（10） 「白書」も「Yahoo!知恵袋」も，「非過去・肯定」がもっとも多い。次いで，「過去・肯定」「非過去・否定」の順で，「過去・否定」は，ほとんど使われない。

４.３ 使われたイ形容詞

「白書」と「Yahoo!知恵袋」では，それぞれ原形の異なりで 73 語，280 語のイ形容詞が使われていた。そのうち，原形の頻度順で上位 10 位までをまとめたものを，表５，６として示す（空欄は，用例数がゼロであったもの)。

表５ 「白書」で使われたイ形容詞上位 10 語

原形	頻度合計	累積頻度（%）	非過去肯定	非過去否定	過去肯定	過去否定
	3998		3580	119	297	2
1 ナイ	1122	1122 28.06%	1012		110	
2 オオイ	834	1956 48.92%	783	4	47	

3	タカイ	470	2426	60.68%	442	9	19	
4	オオキイ	410	2836	70.94%	368	4	38	
5	スクナイ	244	3080	77.04%	153	79	10	2
6	ヒクイ	152	3232	80.84%	138	1	13	
7	ヨイ	138	3370	84.29%	130	2	6	
8	イチジルシイ	97	3467	86.72%	86		11	
9	チイサイ	94	3561	89.07%	80	2	12	
10	ツヨイ	82	3643	91.12%	72	1	9	

表６　「Yahoo!知恵袋」で使われたイ形容詞上位 10 語

	原形	頻度合計	累積頻度（%）		非過去肯定	非過去否定	過去肯定	過去否定
		45088			41764	992	2271	61
1	ナイ	14766	14766	32.74%	14142	19	605	
2	ヨイ	13529	28295	62.75%	12598	313	599	19
3	オオイ	1882	30177	66.92%	1775	22	83	2
4	ホシイ	1313	31490	69.83%	1216	52	41	4
5	ワルイ	1051	32541	72.16%	903	65	81	2
6	オイシイ	668	33209	73.64%	585	23	57	3
7	タカイ	666	33875	75.12%	622	19	23	2
8	ウレシイ	501	34376	76.23%	429	16	54	2
9	ムズカシイ	495	34871	77.33%	468	13	14	
10	オカシイ	439	35310	78.30%	349	87	3	

表５，６から，次のことがわかる。

(12)　「白書」では，「ナイ」「オオイ」の２語で約半数（48.92%）を占める。上位 10 語で全体の９割以上（91.12%）を占めている。

(13)　「Yahoo!知恵袋」では，「ナイ」「ヨイ」の２語で２／３弱（62.75%）を占める。上位 10 語で全体の約８割（78.30%）を占めている。

(14)　「白書」と「Yahoo!知恵袋」の上位 10 語のうち，共通するのは「ナイ」「オオイ」「タカイ」「ヨイ」の４語である。

なお，それぞれの上位 10 語のうち，「イチジルシイ」（「白書」８位）は『日本語能力試験出題基準[改訂版]』で「１級（学習時間 900 時間程度）」の上級語彙，「オカシイ」（「Yahoo!知恵袋」10 位）は「２級　（同 600 時間程度）」の中級語彙とされているが，それ以外はすべて「４級（同 150 時間程度）」の初級語彙とされているものである。

４.４　特徴的なパターン

あまり使われない「過去・肯定」「非過去・否定」「過去・否定」の３つの活用形について，「白書」と「Yahoo!知恵袋」のそれぞれで特徴的なパターンを探った。その結果，「白書」においては「も少なくない」，「Yahoo!知恵袋」においては「て／でほしくない」「て／でもおかしくない」という３つのパターンを抽出することができた。

４.４.１　「白書」における「も少なくない」

「少ない」というイ形容詞は，「白書」全体で 244 例使われていた。これは，原形頻度

順で５位ではあるが，全体に占める割合は 6.1%であり，それほど大きいものではない。しかし，４つの活用形がすべて使われていたのは「少ない」１語だけであり，この点がまず特徴的である。また，「少なくない」は，「非過去・否定」119 例のうち 79 例であり，否定形での使用が多いことが目立つ。さらに，このうち 53 例が，以下のような「も少なくない」というパターンであった。

（15）しかし，これらの地域は，野菜等の生鮮食料品の供給や緑地の保全という面では，依然として重要な役割を果たしており，立地の特性を生かしつつ，農業生産の増大を図っている地域も少なくない。（OW1X_00239，3240）

（16）以上，中小卸売業を取り巻く環境変化に対する個々の商店の対応の方向についてみてきたが，経営資源に限りのある中小卸売店が単独で対応していくためには困難な場合も少なくないと思われ，共同化，組織化による対応は今後とも必要であると考えられる。（OW2X_00168，1170）

４.４.２ 「Yahoo!知恵袋」における「V て（で）ほしくない」

「ほしい」というイ形容詞は，「Yahoo!知恵袋」全体で 1313 例使われていた。1313 例のうち「非過去・否定」の「ほしくない」は 52 例であったが， その 75.0%にあたる 39 例が以下のような「V て（で）ほしくない」というパターンであった。

（17）そういう真剣な姿勢に水をさしてほしくないです。（OC09_05516，610）

（18）私は見て欲しくないから水着も着ません。（OC15_00541，470）

（19）男なら細かいことを何度も言ってほしくないと思いませんか〜？！（OC09_02309，790）

４.４.３ 「Yahoo!知恵袋」における「V ても（でも）おかしくない」

「おかしい」というイ形容詞は，「Yahoo!知恵袋」全体で 439 例使われていた。439 例のうち「非過去・否定」の「おかしくない」は 87 例であったが， その 31.0%にあたる 27 例が以下のような「V て（で）ほしくない」というパターンであった。

（20）もうサンダルはいてもおかしくないですよね？（OC09_00518，70）

（21）三十七ｗになったらいつ陣痛が来てもおかしくないと言うし生まれてもいい時期ですが、小さめだと言われているのでできれば三十九ｗまで、もたせたいと思っています。（OC10_01882，1290）

５．おわりに

以上，本発表では，BCCWJ の２つのサブコーパス（「白書」および「Yahoo!知恵袋」）で使われたイ形容詞を，「出現形の偏り」からみてきた。

イ形容詞の４つの活用形が使われる頻度には大きな偏りがあった。そして，その偏りは，「白書」と「Yahoo!知恵袋」でほぼ同じ傾向を示していた。どちらも，「非過去・肯定」がもっとも多く，この１つの活用形だけで全体の約９割を占めていた。また，「過去・否定」は，ほとんど使われていなかった。

「白書」も「Yahoo!知恵袋」も，限られたイ形容詞が繰り返し何度も使われており，そのほとんどは，『日本語能力試験出題基準［改訂版］』で「４級（学習時間 150 時間程度）」の初級語彙とされているものであった。しかし，好まれる語は「白書」と「Yahoo!知恵袋」で異なっていた。

特徴的なパターンとしては，「白書」においては「も少なくない」，「Yahoo!知恵袋」においては「て／でほしくない」「て／でもおかしくない」という３つを抽出することができ

た。特筆すべきは，この３つのパターンが『日本語能力試験出題基準［改訂版］』の文法リスト，語彙リストのいずれにも項目としてあげられていないことである。

本発表での調査によっても，日本語教育における現行の初級文法シラバスが必ずしも十全ではないことが明らかになった。それは，用言の活用形をすべて初級レベルで取りあげ，聞けて，話せて，読めて，書けるようになるまで練習するという目標設定そのものへの疑問にもつながる。

また，さらに考えるべきは，どのような視点で「初級シラバス」を組み立てるべきかであろう。日本語教育においては，学習者の多様化，学習ニーズの変化等に伴い，一律の文法シラバスではなく，学習者ごとのニーズにきめ細やかに対応できる柔軟な文法シラバスが求められている（野田（編）2005)。「白書」のようなジャンルの日本語を優先的に学びたい学習者もいれば，「Yahoo!知恵袋」のようなジャンルの日本語を必要とする学習者もいる。それらの学習者に対して，一律な「初級シラバス」を立てることへの疑問である。そのためには，多様な日本語使用の実態を丹念に観察し，記述していく作業を続けていかなければいけない。

文献

小林ミナ(2005a).「コミュニケーションに役立つ日本語教育文法」野田尚史編『コミュニケーションのための日本語教育文法』，くろしお出版，21-41.

小林ミナ(2005b).「日常会話にあらわれた「ません」と「ないです」」『日本語教育』第 125 号，日本語教育学会，9-17.

小林ミナ(2008).「『白書』にあらわれたイ形容詞」，『代表性を有する書き言葉コーパスを活用した日本語教育研究』平成 19 年度研究成果報告書，pp. 19-28.

小林ミナ(2009).「「基本的な文法項目」とは何か」，小林ミナ，日比谷潤子（編著）『日本語教育の過去・現在・未来「第５巻文法」』，凡人社，40-61.

野田尚史（編）(2005).『コミュニケーションのための日本語教育文法』，くろしお出版

深田淳，大曾美恵子(2007).「「茶漉」で見る日常会話」，『CASTEL-J in Hawaii2007 Proceedings』，125-128

BCCWJ 複合辞辞書の仕様・開発・評価

近藤泰弘（辞書編集班分担者：青山学院大学/国立国語研究所）†
坂野　収 (辞書編集班協力者：青山学院大学)
多田知子 (辞書編集班協力者：青山学院大学)
岡田純子 (辞書編集班協力者：青山学院大学)
山元啓史 (辞書編集班協力者：東京工業大学)

Development and Evaluation of the Dictionary of Japanese Compound Functional Expressions Based on BCCWJ

Yasuhiro Kondo (Aoyama Gakuin University / NINJAL)
Osamu Banno (Aoyama Gakuin University)
Tomoko Tada (Aoyama Gakuin Univerisity)
Junko Okada (Aoyama Gakuin Univerisity
Hilofumi Yamamoto (Tokyo Institute of Technology)

1　はじめに

本グループでは、BCCWJ を辞書作成の立場から評価するという辞書編集班の目標を達成するため、当初から「文法的辞書」を作成するという目標を掲げた。そして、当面の作業として適切だと思われた BCCWJ におけるすべての複合辞のリストを作成することを試みた。、従来の複合辞リストを改めて見直し、用例から帰納してリストを作成することによって、新規の複合辞を多く収めたものを作ることができた。本稿では、今回作成したものを「BCCWJ 複合辞辞書」(Ver.1.0) と呼称する。作成の手順は以下に記すとおりであるが、その作業はすべて辞書編集班・複合辞グループに属する、近藤・坂野・多田・岡田・山元の 5 名が行った。なお、青山学院大学院の近藤の 2006 年度から 2010 年度の演習参加者からは有益なアドバイスを得た。

また、従来の複合辞研究、特に、辞書の中に参考として当該語の有無を記述させていただいた、国立国語研究所『複合辞用例集』、森田・松木『日本語表現文型』、グループ・ジャマシイ『日本語文型辞典』、松吉・佐藤『つつじ：日本語機能表現辞書』にはひじょうな恩恵を受けた。また、藤田・山崎『複合辞研究の現在』、そしてその流れの中にある複合辞研究会の活動からも多くを得ている。合わせて感謝したい。

また、本辞書の改訂版については、`http://www.japanese.gr.jp` で配布を予定している（現在は準備中）。BCCWJ が完成した後に開設したい。

† yhkondo@cl.aoyama.ac.jp

以下に、今回の複合辞辞書作成の概要について解説する。

2　複合辞の定義

複合辞は、複数の形態素が結合してできたものであるが、文法的には機能語であるため、機能語としての性格と複合語としての性格をともに持っている。特に今回問題としたのは、接続詞の取り扱いである。そもそも複合辞は、自立語であるいわゆる内容語（名詞・動詞等）が、付属語である機能語（助詞・助動詞）へと変化するのである。しかし、「それで」など、接続詞の多くは複数の形態素が結合した複合形式であるが、接続詞自体は、自立語である。文法論的には、接続詞は、接続助詞と副詞との中間に位置するものであり、文と文を接続する文法的機能をも持っていることは明らかである。したがって、自立語という点を重視すれば複合辞には入れるべきではないが、機能語の側面を重視すれば、複合辞としてもよい部分もある。従来の研究ではどちらの扱いもあるが、今回の研究では、接続詞も複合辞としてリストアップすることとした。これについては、グループの多田 (2010ab) の論考があるので、参照されたい。

なお、「それが」（接続詞）のように、「代名詞 + 格助詞」が接続詞になるときには、機能語化という面を重要視すれば、文法化が起きていると言える。それに対して「だが」（接続詞）の場合は、「・・・だが」という接続助詞（付属語）から変化したものであるので、接続詞が自立語であるという側面を重視すれば、文法化ではなく、その逆の「脱文法化」が起きていると解釈可能である。このように、接続詞の複合辞としての性格付けには複雑な問題が残っており今後の課題となる。

同じくもうひとつの問題として、「急いで行かなくては。」の「ては」のように助動詞的な働きを持つものの活用はしない一連の複合辞がある。これらについては、特に「文末辞」という品詞（文法範疇）として分類することにした。この点については、グループの坂野 (2010) の論考があるので参照されたい。

以上のように、今回は、複合辞を広く解釈したこともあり、また、帰納的方法により網羅を目指したこともありで、従来のいかなる複合辞リストよりも単純にその収納数は増えていることに注意されたい。

3　複合辞を網羅するための方法論

今回は BCCWJ という均衡コーパスが存在しており、その評価をすることが必要であるため、その中にある複合辞を網羅することになる。単純に形態素解析をして KWIC 索引を作るだけではうまくいかないので、形態素解析を行った後、短単位が複数結合した N-gram を作成し、2 グラムから 5 グラムまでの結合をリストアップした。そして、その中で単純頻度が高いもの、T スコアや MI スコアの高いものなどを調査し、複合辞の候補リストを作成し、それらから各種のチェックリストによって手作業で抜き出した。最初は T スコア等でかなり自動化が可能かと考えていたが、意外にむずかしく、単純頻度によるリストからの手作業の部分が大きかった。この作業については、坂野・多田・岡田の尽力があった。

4　複合辞辞書の解説

複合辞辞書は、配布される DVD 報告書に含まれるエクセル版の「複合辞一覧表」シートが本体である。その内容にの詳細については付属の Readme.doc を参照されたい。概略は次の表 1 のとおりである。

今回の複合辞辞書は、あくまで人間が読むためのものであり、したがって、派生形や表記の異動については最低レベルの記述しかしていない。これについては「つつじ」のように複数の階層を設けて記述すれば機械処理に対応できるのであるが、それは今回の目的ではないため、異表記についての階層は作らなかった。基本的には、語形（大見出し）、その形態的文法的変異（中見出し）、文法機能による分類（小見出し）の 3 階層を作った。見出し語の情報以外には、文法機能、意味範疇、前接要素、構成、用例を掲出した。また、参考として、機能語ではなく内容語としての用法もあげた。たとえば、「・・活動の上で欠かせない」の「上で」は機能語（複合辞）であるが、「椅子の上で寝る。」の「上で」は内容語としての用例であり、複合辞ではない。この後者のようなものを参考までにあげた。また、冒頭に記した 4 文献における語の有無を参考までに掲載した。また、付属として、エクセル版の「複合辞分類表」シートもつけた。これは、大見出し・中見出し・小見出しを一覧し、文法機能によって大きく分類整理したものである。複合辞辞書（エクセル版）以外に、複合辞辞書（印刷版）も作成した。印刷版では、小見出し 925 個を見出し語として掲出してある。印刷版の TeX によるフォーマットは山元が行った。

次節では、複合辞辞書開発の評価のひとつとして、本研究で開発された辞書の見出しが日本語教育の教科書においてどのように分布しているのかを考察する。

5　評価

均衡コーパス BCCWJ から、前節のようにして得られた複合辞は、ある特定目的のテキストにおいてはどのような分布を見せるのだろうか。均衡コーパスによれば、複合辞の抽出時の網羅性はある程度保証され、教育項目の選定作業を進めたり、教育項目の適切性の評価の尺度として用いることできよう。均衡コーパスにしたがって集められた複合辞は、日本語教育の入門期から上級まで、どの学習段階に分布して提示されているのだろうか。本節では、BCCWJ 複合辞辞書の小見出しが、各種日本語教科書データベースを用いて、初級、中級のどの段階で扱われているのかを調査し、検討する。

5.1　目的

実際の日本語教科書では、複合辞はどのように扱われているのだろうか。均衡コーパスから抽出した複合辞の見出し語は、初級、中級のいずれ教科書で扱われているのだろうか。またどんな複合辞が扱われているのだろうか。さらに、教科書では扱われていないものとはどのような特徴をもつ

表 1　複合辞辞書のフィールドの説明

記載項目	記載内容
項番	「小見出し」対応の複合辞番号。全部で 925 項番。順序は昇順（あいうえお順）。
大見出し	表現形式（言葉）が異なる複合辞
中見出し	「大見出し」に、係り/副・助詞が単純に挿入された辞や、同一表現ながら文法機能が異なるものを、下位の「中見出し」とした。
小見出し	「大見出し」と同一表現、同一文法機能でも、意味範疇の異なるものを「小見出し」として分類した。それに、「中見出し」項目を加えて、表現/文法機能/意味範疇が異なるものを、すべて「小見出し」で記載。
構成組成	「小見出し」表現を形態素に分解したもの。
意味範疇	複合辞の意味・機能
前接	複合辞が接続する（の前にくる）文法形態
文法機能	複合辞の相当品詞
解説・用法	意味範疇や用法の補足説明など必要により記入
備考	当該複合辞の関連表現など、必要により記入
機能的用法の用例	各ジャンルからの引用
内容的用法の用例	「辞」ではなく「内容語」としての用例
先行文献 （先行文献に記載あれば〇印）	森田・松木（1989）『日本語表現文型』（アルク）
	国研（山崎・他）（2001）『現代複合辞用例集』（国立国語研究所）
	グループ・ジャマシイ（1998）『日本語文型辞典』（くろしお出版）
	松吉・佐藤（2008）『日本語機能表現辞典「つつじ」』(HTML 版）

複合辞なのだろうか。以上の観点から日本語教育で扱われている BCCWJ 複合辞辞典の範囲について分析する。

5.2　方法

検索語群として、BCCWJ 複合辞辞書の小見出し（680 種：形のみで下位分類は問わない）を用いる。

日本語教科書は、初級（24 種）、中級（16 種）の各課本文のみを対象とする。各課に付けられた練習問題は、教科書毎に取扱いが異なるため分析対象としない。分析に用いた各教科書は表 2、3 のとおり。各教科書は、光学的文字読み取り装置で、電子化され、データベース化された。データ

表 2　初級 24 種類の日本語教科書リスト

管理番号	教科書名	開発元
B01	Learn Japanese vol. I	University of Maryland College
B02	Learn Japanese vol. II	University of Maryland College
B03	Learn Japanese vol. III	University of Maryland College
B04	Learn Japanese vol. IV	University of Maryland College
B05	An Introduction to Modern Japanese	O. Mizutani, N. Mizutani
B06	Intensive Course in Japanese –Elementary–	対外日本語教育振興会
B07	Beginning Japanese Part I	B. Block, W. Cornyn, I. Dyen
B08	Beginning Japanese Part II	B. Block, W. Cornyn, I. Dyen
B09	日本語初歩	国際交流基金
B10	外国学生用日本語教科書—初級—	早稲田大学語学研究所
B11	日本語 I	国際学友会日本語学校
B12	Nihongo no Kiso I	海外技術者研修協会
B13	Japanese –A Basic Course–	A. Alfonso, K. Niimi
B14	Buisiness Japanese	日産自動車国際課
B15	生活日本語 I	文化庁
B16	生活日本語 II	文化庁
B17	Situational Functional Japanese	筑波ランゲージグループ
B18	Basic Kanji Book Vol. I,II	加納, 清水, 竹中, 石井
B19	ようこそ Yookoso	當作 靖彦
B20	日本語 I	東京外国語大学附属日本語学校教材開発研究協議会
B21	初級日本語	東京外国語大学附属日本語学校
B22	文化初級日本語 I	文化外国語専門学校日本語科
B23	文化初級日本語 II	文化外国語専門学校日本語科
B24	Japanese for Today	Gakken

ベースは、1 文（文頭あるいは句点から句点まで）を 1 レコードとして、ページ数、文数、本文、練習などのタグが付与された。本来ならば、複合辞の連語的な使われ方が取り扱われているかどうかが、見たいが、そのためには人手にて意味判別を 1 行ずつ行わなければならないため、今回は、複合辞の形が出現するかどうかのみを見た‡。

手順としては、まず、複合辞辞書の見出しのリスト 680 を取り出し、その 680 のいずれが初級教科書データベースに出現したのかを調べる。つぎに、同じく 680 の複合辞のいずれが中級教科書に出現したのかを調べる。その上で、680 の複合辞のうち、初級までの複合辞、中級までの複合辞、そして、初中級の教科書で出現しなかった複合辞の 3 群に分類し、3 群の特性について、分析、考察を行う。

5.3　結果

初級教科書、中級教科書について、複合辞辞書の小見出しの出現の有無を調べたところ、表 4 のようになった。

‡ 頻度が低い時には、複合辞ではない可能性が高いが多数回出現する時は複合辞である可能性も含まれよう。

表 3　中級 16 種類の日本語教科書リスト

管理番号	教科書名	開発元
I01	INTERMEDIATE JAPANESE VOLUME 1	大阪外国語大学留学生別科
I02	INTERMEDIATE JAPANESE VOLUME 2	大阪外国語大学留学生別科
I03	日本語 II	東京外国語大学附属日本語学校教材開発研究協議会
I04	中級日本語	東京外国語大学留学生日本語教育センター
I05	現代日本語コース中級 I	名古屋大学総合言語センター日本語学科
I06	日本語　中級 I	東海大学留学生別科
I07	日本語 II	国際学友会日本語学校
I08	日本語中級 I	国際交流基金
I09	日本語中級 II	国際交流基金
I10	ちょっと ひとこと	朝日カルチャーセンター (佐々木倫子)
I11	日本語でビジネス会話 中級編〈本文冊〉	日米会話学院日本語研修所
I12	文化中級日本語 I	文化外国語専門学校
I13	文化中級日本語 II	文化外国語専門学校
I14	中級から学ぶ日本語	荒井, 太田, 大藪, 亀田, 木川, 長田, 松田
I15	総合日本語中級	水谷信子
I16	日本語中級 J301 －基礎から中級へ－英語版	土岐, 関, 平高, 新内, 鶴尾

表 4　分析対象の複合辞の総数と初級教科書、中級教科書で出現した複合辞の数

内訳	複合辞の数	割合	計算方法
初級 24 教科書	271 (a)	39.9%	(a)/(全体)
初級 24 以外	409 (b)	60.1%	(b)/(全体)
中級 16 教科書	430 (c)	63.2%	(c)/(全体)
中級 16 以外	250 (d)	36.8%	(d)/(全体)
複合辞辞典全体	680 (全体)	100.0%	

5.3.1　初級の範囲で見られたもの

初級教科書で扱われた複合辞の数は 271 である。初級で扱われた複合辞は、「なければならない」「かもしれない」「なければならない」のようなイディオムや「おかげさまで」のような挨拶で使われる決まり文句的な表現である。小さな単位「お／かげ／さま／で」「なけ／れ／ば／なら／ない」に分割すると意味が見えにくいので、1 つの表現としてまとめて扱かわれるものである。「(〜た) ほうがよい」のように比較の表現を教える時には定番といったものも見られる。表 5 の左欄で見られるものはいずれも初級教科書の早い時期に提示され、基礎の基礎と呼ばれるものである。

5.3.2　中級の範囲で見られたもの

中級教科書で見られた複合辞の数は 430 である。これは複合辞辞典 (680) の 63.2% に当たり、250(36.8%) の複合辞が中級の教科書において扱われなかったことがわかる。中級教科書で見られた 430 の中には、初級ですでに出現したものもあるので、中級教科書のみには、どのような複合辞があるかを見るために、中級教科書に見られた複合辞のリスト (430) と初級教科書に見られた複合辞のリスト (271) の差分を見る必要がある。

その結果、中級のみに見られる複合辞は、173 (全体から見ると、およそ 25%) であった。わずか

表 5　初級教科書 24 種に見られる上位 20 の複合辞とその出現頻度と中級 16 教科書のみに見られる上位 20 の複合辞とその出現頻度（複合辞文字列長 3 文字以上のみを対象とした）

順位	初級 24 教科書	出現頻度	中級 16 教科書	出現頻度
1	てください	655	したがって	27
2	ている	314	といった	24
3	がいい	225	一方で	14
4	ないで	210	その結果	13
5	をして	188	を通して	11
6	それで	167	ながらも	10
7	それから	156	ねばならない	9
8	ように	143	どころか	9
9	てもいい	125	といえる	8
10	という	113	に従って	7
11	なければ	110	要するに	6
12	てから	110	に応じ	6
13	だから	98	にもかかわらず	6
14	ないか	95	と思うと	6
15	それでは	93	ためだ	6
16	なくて	90	を問わず	5
17	ほうがいい	89	べきだ	5
18	について	83	としたら	5
19	ところで	81	その一方で	5
20	うちに	71	その一方	5

に初級のみに 14 の複合辞が見られたが、おおむね中級は初級の複合辞を包含するものと考えてよいだろう。ことになる。全体 680 のうち中級でほぼ 6 割、初級でほぼ 4 割、そして中級はほぼ初級を包含するので、中級教科書まででカバーされない複合辞は 4 割弱である。

表 5 の右欄は、中級教科書に見られる複合辞、頻度上位 20 である。初級よりやや書きことば的な感じのするものが中級では見られる。また、論理展開を記述するため文接続に関わる表現も見られる。

一般的に中級では、話しことば的な表現に加えて書きことば的な表現の導入やより長めの文章が取り扱われる。初級で学ぶ日常会話的な表現を前提に、その次の段階として、従来の書きことば的な要素の加わったものを取り扱いつつ、文や段落をつないで、すでに述べたことに説明を加えたり、対比したりするための複合辞が扱われている。

5.3.3　初級および中級教科書に出現しなかった複合辞

中級教科書まででカバーされない複合辞はほぼ 40% であるが、その中身はどのようなものだろう。

表 6 に中級教科書までに出現しなかった複合辞を一覧にした。一覧に見られる傾向としては、1. 助詞に相当する語句、2. 文と文の接続に関連する語句、が多く、3. 書きことばに見られる話しことば調の語句、4. 話しことばによく見られ、かつ、古語を含む語句、が見られる。

1. 助詞に相当する語句。

「〜について」「〜に関して」のような助詞に相当する語句で、やや論述文に見られる表現

の類。

たとえば、210) に至っては、211) に至り、212) に次いで、213) に従い、214) に乗じ、215) に乗じて、216) に先立ち、217) に先立って、218) に相違ない、219) に増して、など。

2. 文と文の接続に関連する語句。

たとえば、6) いずれにしても、18) こういうわけで、19) こうなると、20) ここにきて、22) このために、27) これに加え、30) これに伴ない、31) これに反し、34) さもないと、37) さらにいえば、80) それがかえって、81) それがために、85) それでこそ、86) それでもって、88) それにつけても、91) それに加え、93) それに伴って、95) それに反し、97) それはさておき、101) それゆえ、など。

3. 書きことばに見られる話しことば調の語句。

たとえば、11) かしらん、48) そいでもって、79) そやけど、103) たってかまわない、104) たってしょうがない、110) だっけ、111) だってば、117) っていうか、118) ってのか、176) とやら、など。

4. 話しことばによく見られ、かつ、古語を含む語句。

たとえば、2) あげくのはて、5) いざしらず、10) かくなる上は、33) さうして、35) さもなくば、38) されども、39) ざるべからず、40) ざるを得ず、41) しかして、42) しからば、64) そのあげく、107) たるや、135) ても差し支えない、170) とはいえども、177) と思いきや、224) はおろか、228) べからず、229) べくして、230) べくもない、247) 然れど、248) 然れども、など。

1. あるいは 2. は初級、中級と段階的に学習を積み重ねてきて、上級で取り扱う内容と考えてもよい。212) の「に次いで、」や 213) の「 に従い、」は、「〜に続けて」「〜の次に」、「〜にしたがって」「〜につれて」のように初級で学んだ表現の上級のものと考えてもよい。その意味では、学習の段階としては順当であろう。

3. の「書きことばにも見られる話しことば調の語句」は、書きことばであっても、話しことばを意識しながら（実際には意識下で声には出さない発音で: subvocalization)、語句をとらえなければわかりにくいもので、マンガやシナリオ、小説などでよく使われいるが、従来の日本語の教科書の中ではあまり扱われていないようだ。

4. の「話しことばによく見られ、かつ、古語を含む語句」は、初級教科書、中級教科書で扱われなかった複合辞の最大の特徴である。日本語教育の現場では、古語を含む表現を敬遠する向きがあるのかどうかわからないが、「177) と思いきや、」「247) 然れど、」など実際の会話や文章で多々見られる。均衡コーパスを教育における言語の基準と見做すことによって、この類の語句を日本語教育で取り扱うように提案してもよいはずだ。

5.3.4 評価のまとめ

BCCWJ 複合辞辞典を初級教科書、中級教科書データベースと照合して、複合辞が段階的にどのように日本語教育で扱われるかを見た。一言でそれぞれの段階で取り上げられている複合辞の特徴

表 6　初級および中級にも見られなかった複合辞のリスト

1) あげくに、2) あげくのはて、3) あげくのはてに、4) いいかえれば、5) いざしらず、6) いずれにしても、7) いずれにせよ、8) うえは、9) かぎりは、10) かくなる上は、11) かしらん、12) かもわからない、13) からとて、14) がために、15) がゆえに、16) が早いか、17) くせして、18) こういうわけで、19) こうなると、20) ここにきて、21) ここへきて、22) このために、23) このためには、24) このためにも、25) この上は、26) これだから、27) これに加え、28) これに加えて、29) これに伴って、30) これに伴ない、31) これに反し、32) これに反して、33) さうして、34) さもないと、35) さもなくば、36) さもなければ、37) さらにいえば、38) されども、39) ざるべからず、40) ざるを得ず、41) しかして、42) しからば、43) しだいだ、44) してみれば、45) ずにいられない、46) ずにはいない、47) ずにはおかない、48) そいでもって、49) そうしたら、50) そうしてから、51) そうしないと、52) そうしながら、53) そうしながらも、54) そうでないと、55) そうでなくて、56) そうでなくては、57) そうでなくても、58) そうでなくとも、59) そうでなければ、60) そうとはいえ、61) そうはいっても、62) そこへいくと、63) そこへもってきて、64) そのあげく、65) そのあげくに、66) そのかわり、67) そのくせ、68) そのくせに、69) そのせいか、70) そのせいで、71) そのためか、72) そのとたんに、73) そのゆえに、74) その際、75) その上で、76) その前に、77) その代わりに、78) そばから、79) そやけど、80) それがかえって、81) それがために、82) それがゆえに、83) それだったら、84) それでいて、85) それでこそ、86) それでもって、87) それというのも、88) それにつけても、89) それにひきかえ、90) それにより、91) それに加え、92) それに加えて、93) それに伴って、94) それに伴ない、95) それに反し、96) それに反して、97) それはさておき、98) それはそうと、99) それはとくもかく、100) それはともかくとして、101) それゆえ、102) それゆえに、103) たってかまわない、104) たってしょうがない、105) たらいけない、106) たらない、107) たるや、108) だけのことはある、109) だけれども、110) だっけ、111) だってば、112) だとしても、113) だとするなら、114) だとするならば、115) だとすれば、116) ちなみに、117) っていうか、118) ってのか、119) ついでながら、120) つつも、121) つまりは、122) ていうか、123) ていられない、124) てかまわない、125) てさしあげる、126) てしかたがない、127) てしかたない、128) てしかるべきだ、129) てしょうがない、130) てたまらない、131) てたまるか、132) てなるものか、133) てもしかたがない、134) てもしょうがない、135) ても差し支えない、136) であればこそ、137) でいうなら、138) でいうならば、139) でなかったら、140) ではないのか、141) でもって、142) とあって、143) とあっては、144) とあれば、145) といいながら、146) というところだ、147) というものの、148) というわけだ、149) といえど、150) といえども、151) といえなくもない、152) といおうか、153) といけない、154) といったところだ、155) といったらない、156) といわず、157) ときたら、158) ときている、159) とくれば、160) ところだった、161) としたことが、162) としてみれば、163) とするなら、164) とするならば、165) となったら、166) となっては、167) となれば、168) とにかかわらず、169) とはいいながら、170) とはいえども、171) とはいっても、172) とばかりに、173) とみえる、174) ともあろうものが、175) とも限らない、176) とやら、177) と思いきや、178) どころではない、179) ないことには、180) ないことはない、181) ないこともない、182) ないといけない、183) ないとならない、184) なくしては、185) なくはない、186) なくもない、187) なければだめだ、188) にあたり、189) にあっては、190) にあらず、191) におかれましては、192) にかけても、193) にかこつけて、194) にしたところで、195) にしてからが、196) にしてみたら、197) にしてみると、198) にしてみれば、199) にたえない、200) にとってみれば、201) にとどまらず、202) には及ばない、203) にひきかえ、204) によらず、205) に及び、206) に及んで、207) に決まっている、208) に際し、209) に際して、210) に至っては、211) に至り、212) に次いで、213) に従い、214) に乗じ、215) に乗じて、216) に先立ち、217) に先立って、218) に相違ない、219) に増して、220) に代えて、221) に比し、222) に比して、223) に免じて、224) はおろか、225) はさておいて、226) はずではなかった、227) へかけて、228) べからず、229) べくして、230) べくもない、231) ましてや、232) までのことだ、233) もいいところだ、234) ものとする、235) ものなら、236) ゆえに、237) よりほかない、238) より仕方がない、239) わけにいかない、240) をめぐり、241) んことを、242) んばかりだ、243) 一方だ、244) 何となれば、245) 結局のところ、246) 最中に、247) 然れど、248) 然れども、249) 他方で、250) 他方では、

を述べるのは難しいが、おおむね、初級では、平易な意味をなすひとかたまりをひとつの文法機能として教えることを意図した語句が集められている。中級では、初級で学んだ日常会話的な表現を前提に、書きことば的な要素の加わった複合辞を取り扱っている。また、初級よりも長めの文章を取り扱っていることから、文や段落をつなぐ複合辞を導入して、説明や対比の表現力強化を意図しているのがわかる。初級中級の範囲で見られなかった複合辞は、話しことばでよく見聞きする文語体の表現に特徴が見られた。

均衡コーパスを、言語のスタンダード（ここでは、尺度となりうるような一定の基準を満たしたもの、揺るがないものという意味）的な尺度として、分析に利用することによって、教育で扱われる内容の網羅性、方向性の検討、教育内容のすみやかな改善は十分可能である。今後ともコーパス研究の応用分野として研究が進められていく必要があろう。

本来、教育で扱うべき事柄は、常に根拠があってしかるべきで、教育者の内在的基準だけなく、外在的基準に照らし合わせて、決定されるべきである。均衡コーパスを元に開発した複合辞辞典によれば、なぜそれを教育で取り扱うのか、明解な根拠を与えてくれる。時代につれて変化する言語の実状を言語教育のシステムとしてとらえるためには、尺度としての均衡コーパスが必要であり、今後ともその更新が必要不可欠である。

6　研究の今後

今回は BCCWJ2008 および 2009 を用いて複合辞辞書を作成したが、BCCWJ が完成した時にはそれによって再度リストを訂正することを考えている。それはバージョン 1.1 ということになるだろう。また、今回はエクセル版、印刷版の二種類を作成したが、電子書籍としての PDF 版なども作成を予定している。iPad や SONYReader などで利用しやすい形態を考えたい。また、今回の複合辞辞書を作成するにあたっては、従来の品詞分類の限界を痛感した。単純辞機能語の場合も含め、今後、品詞分類の再考を行う必要があると考えている。複合辞リストを利用した自然言語処理的な研究も必要である。さらには、単純辞機能語を補充することで、BCCWJ 機能語辞書を作成することも必要かもしれない。これについては、他の班の作業に協力して行うことも可能だろう。

文献

[1] 近藤泰弘 (2000)『日本語記述文法の理論』ひつじ書房.

[2] 多田知子 (2010)「複合接続詞ー文の冒頭部分の階層性ー」(『国文論叢』神戸大学文学部・42 号)

[3] 多田知子 (2010)「複合接続詞の生成」(平成 21 年度研究成果報告書・辞書編集班「コーパスを利用した国語辞典編集法の研究」)

[4] 坂野収 (2010)「「言いさし表現」と文末複合辞」(平成 21 年度研究成果報告書・辞書編集班「コーパスを利用した国語辞典編集法の研究」)

複数の観点から見た用例クラスタリングに基づく新語義の発見

白井 清昭 (言語処理班分担者: 北陸先端科学技術大学院大学 情報科学研究科) [1]
中西 隆一郎 (言語処理班協力者: 北陸先端科学技術大学院大学 情報科学研究科)
中村 誠 (言語処理班分担者: 北陸先端科学技術大学院大学 情報科学研究科)

Finding New Word Senses Based on Clustering of Example Sentences from Multiple Viewpoints

Kiyoaki Shirai (Japan Advanced Institute of Science and Technology)
Ryûichirô Nakanishi (Japan Advanced Institute of Science and Technology)
Makoto Nakamura (Japan Advanced Institute of Science and Technology)

1 はじめに

一般に，単語の意味は日々変化し，新しい意味や用法も生まれている．単語の意味は辞書などで定義されるが，単語が辞書に定義されていない新しい意味で使われていることもある．本研究は，辞書に定義されていない単語の意味を新語義と呼び，コーパスから新語義を自動的に発見する手法を確立することを目的とする．新語義を自動的に発見することができれば，辞書編纂作業のサポートや，より完全な意味のセットを定義する辞書の整備に貢献する．辞書の整備は単語の意味を取り扱う様々な自然言語処理アプリケーションの精緻化にもつながる．

本研究が提案する新語義発見手法の概要を図 1 に示す．まず，新しい語義を発見する対象単語を w とする．w の用例をコーパスから抽出し，同じ語義を持つ用例がまとまるようにクラスタリングを行う (図 1 ①)．次に，作成された個々の用例クラスタに対し，辞書の語義との類似度を計算し，どの語義とも似ていないクラスタを新語義の用例をまとめたクラスタ (新語義クラスタ) と判定する (図 1 ②)．最後に新語義クラスタを出力する．

以下，2 節では用例のクラスタリング手法，3 節では用例クラスタに対する新語義の判定手法について述べる．4 節では提案手法の評価実験について報告する．5 節では関連研究について述べる．最後に 6 節でまとめと今後の課題について述べる．

2 用例クラスタリング

ここでは用例クラスタリングのタスクを以下のように定義する．対象単語 w を含む用例の集合 $W = \{w_i\}$ が与えられたとき，同じ語義を持つ用例のクラスタに分割し，クラスタの集合 $\mathcal{C} = \{C_k\}$ を得る．

提案手法 (中西他, 2011) の特徴は，ひとつの用例を複数の特徴ベクトルで表現し，それらを同時に利用してクラスタリングを行う点にある．一般に，語の意味の同値性あるいは類似性は様々な観点か

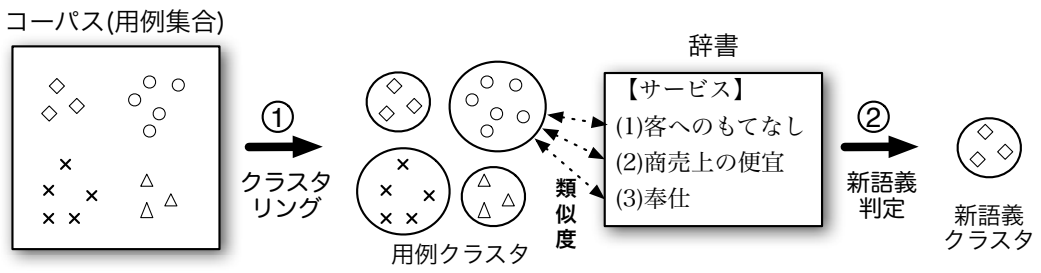


図 1: 提案手法の概要

[1] kshirai@jaist.ac.jp

(a) 時まで、あとのぶんは サービス 残業 … というわけ
その差約７００時間が サービス 残業。現在過労死が若

(b) ケーキとシャンパンを サービス されたんです。ＣＡか
とりました。飲み物を サービス したり、一緒に写真撮

(c) ファイアーウォールの サービス を開始しようとしたと
う名前でＡｐａｃｈｅ サービス をインストールするに

【サービス】(岩波国語辞典での定義)

S_1 客に対するもてなし。接待。優遇。「サービスのよい旅館」
S_2 商売で、値引きしたり客の便宜を図ったりすること。「百円サービスしておきます」
S_3 奉仕。「家庭サービス」
S_4 ↓サーブ。[2]

図 2: 「サービス」の用例ならびに語義の定義

ら認められる．例えば，図 2 に示す「サービス」の用例について考察してみよう．岩波国語辞典 (西尾他, 1994) によれば,「サービス」には図 2 右に示す 4 つの意味がある．図 2 (a) の「サービス」は，直後の単語が「残業」であることから S_3 の意味を持つと考えられる．一方，図 2 (b) は「ケーキ」「シャンパン」「飲み物」のような飲食物が周辺に出現していることから S_1 の意味を持つと考えられる．図 2 (c) の「サービス」はコンピュータに関連するテキストに出現することから，岩波国語辞典では定義されていない意味 (ネットワーク上でサーバが提供する「サービス」) であるといえる．すなわち，語の意味は，直前・直後の単語で識別できる場合，文脈に出現する単語で識別できる場合，テキストのトピックによって識別できる場合などがある．

このように，語の意味の類似性は様々な観点で測ることができる．しかし，用例クラスタリングに関する多くの先行研究では用例を 1 種類の特徴ベクトルで表現するが，これでは上記のような多様な観点を捉えることは難しい．本研究では，用例を異なる観点から見た複数の特徴ベクトルで表現し，これら複数のベクトルを同時に考慮して用例クラスタを作成することで，クラスタリングの精度を向上させることを目指す．

2.1 特徴ベクトル

本研究では用例 w_i を以下の 4 種類の特徴ベクトルで表現する (九岡他, 2008).

隣接ベクトル w_i の直前または直後に現われる単語で w_i を特徴付けるベクトル．具体的には，w_i の前後 2 語の単語の出現形ならびに品詞をベクトルの素性とする．

文脈ベクトル w_i の周辺に現われる単語で w_i を特徴付けるベクトル．また，w_i の周辺に直接現われる単語 x だけではなく，x と同一のトピックを持つ単語もベクトルの素性とすることにより，ベクトルの過疎性を緩和する．単語のトピックは LDA(Latent Dirichlet Allocation) (Blei et al., 2003) によってコーパスから自動的に推測する．

連想ベクトル 文脈ベクトルと同じく，w_i の周辺に現われる単語で w_i を特徴付けるベクトル．ただし，ベクトルの過疎性を緩和するために，事前にコーパスから作成された単語の共起行列を用いる．単語の共起行列の列を，ある単語が別の単語とどの程度共起しやすいかを表わす共起ベクトルとみなし，w_i の文脈に出現する単語の共起ベクトルの和を文脈ベクトルと定義する．

トピックベクトル PLSI (Probabilistic Latent Semantic Indexing) (Hofmann, 1999) によって推定されるトピックによって w_i を特徴付けるベクトル．具体的には，w_i を含む文書を d_i としたとき，$P(z_l|d_i)$ (z_l は PLSI の隠れ変数 (トピック)) を素性とするベクトルを作成する．

これらの特徴ベクトルは用例間の類似度を計算するために用いるが，隣接ベクトルは図 2 (a) の例のように直前・直後に出現する単語が似ているかという観点，文脈ベクトルと連想ベクトルは図 2 (b) のように周辺文脈に出現する単語が似ているかという観点，トピックベクトルは図 2 (c) のようにテキストのトピックが似ているかという観点で語義の類似性を測っている．用例をクラスタリングする際，これら 4 つの特徴ベクトルを併用することで，様々な観点から語義の類似性を捉えることを狙う．

[2]他の見出し語「サーブ」と同じ意味を持つことを表わす．

2.2 クラスタリング

2.1 項で述べた 4 つの特徴ベクトルを同時に利用するために，凝集型クラスタリングのアルゴリズムを拡張する．その概要は以下の通りである．まず，個々の用例を 1 つのクラスタとみなして初期のクラスタ集合$\mathcal{C} = \{C_1 \cdots C_n\}$を作成する．次に，全てのクラスタの組についてクラスタ間類似度$sim(C_i, C_j)$を計算し，それが最大となるC_i, C_jを求める．両者を併合したクラスタC_kを作成し，その重心ベクトルと後述するクラスタラベル$L(C_k)$を更新した後，$\mathcal{C}$を更新する[3]．この処理を停止条件を満たすまで繰り返す．

2.2.1 クラスタ間類似度

クラスタ間類似度は 2.1 項で述べた 4 つの特徴ベクトルを用いて式 (1) のように計算する．

$$sim(C_i, C_j) = \max_{v \in \{\text{ 隣接, 連想, 文脈, トピック }\}} s(v, C_i, C_j) \tag{1}$$

$s(v, C_i, C_j)$ は特徴ベクトル v によって計算されるクラスタ間の類似度である．具体的には，用例を特徴ベクトル v で表現したときのクラスタの重心ベクトル[4]のコサイン類似度と定義する．式 (1) は，クラスタ間の類似度を，隣接，文脈，連想，トピックベクトルで計算される類似度の最大値と定義している．これは，4 つの特徴ベクトルで考慮されている複数の観点のうち，どれか 1 つについてでも類似度が十分高ければ，それらは同じ語義を持つ可能性が高いという考えに基づく．

さらに，クラスタを作成する際には，同一の特徴ベクトルによる類似度が高い用例をまとめるという制約を設ける．例えば，最初に類似度が最大となるクラスタの組を併合して新しいクラスタを作成したとき，式 (1) で 4 つの特徴ベクトルのうち隣接ベクトルの類似度が最大であった場合には，以後は隣接ベクトルの類似度が十分高いときのみそのクラスタに新しい要素を併合する．作成されたクラスタは隣接，文脈，連想，トピックベクトルのいずれかによって計算される類似度が高い用例をまとめたものとなる．これにより，作成された個々のクラスタについて，それがどのような観点で似ている用例をまとめたものなのかを容易に解釈できる．

この制約はクラスタラベル $L(C_k)$ を導入することで実現する．$L(C_k)$ はクラスタ C_k がどの特徴ベクトルの観点から用例をまとめたかを示すラベルである．初期クラスタでの $L(C_k)$ は「未定」とする．また，C_i と C_j が併合されて C_k が作成されたとき，式 (1) の $s(v, C_i, C_j)$ が最大となるベクトルの種類に応じて「隣接」「文脈」「連想」「トピック」のいずれかを $L(C_k)$ とする．さらに用例間類似度 $sim(C_i, C_j)$ を式 (2) のように再定義する．

$$sim(C_i, C_j) = \begin{cases} \text{式 (1)} & \text{if } L(C_i) = L(C_j) = \text{未定} \\ s(L(C_i), C_i, C_j) & \text{if } L(C_i) = L(C_j) \text{ or } L(C_j) = \text{未定} \\ s(L(C_j), C_i, C_j) & \text{if } L(C_i) = L(C_j) \text{ or } L(C_i) = \text{未定} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

式 (2) の 2,3 行目は，2 つのクラスタのラベルが一致しているか，どちらか一方が「未定」のとき，「未定」でないクラスタラベルの特徴ベクトルの類似度をクラスタ間類似度とすることを表わす．また，4 行目は，C_i と C_j のクラスタラベルが異なるときは類似度を 0 とし，両者を併合しないことを表わす．

2.2.2 ベクトル間類似度の正規化

予備実験により，4 つの特徴ベクトルによって計算されるクラスタ間類似度の値には大きな差があることがわかった．式 (1) で 4 つの特徴ベクトルによるコサイン類似度を単に比較するだけでは，ベクトル間類似度が平均的に高い特徴ベクトルのみが常に選択される可能性がある．4 つの特徴ベクトルによる類似度の値を公平に比較するために，ベクトル間類似度を正規化する．

[3] $\mathcal{C}$ から C_i,C_j を削除し，C_k を追加する．
[4] クラスタ内の要素の特徴ベクトルを平均したベクトル．

まず，特徴ベクトル v によるベクトル間類似度の標本を X_v とする．X_v は，用例集合 W における全ての用例の組に対する特徴ベクトル v のコサイン類似度の値の集合とする．次に，正規化された類似度 s_R を式 (3) のように定義する．

$$s_R(v, C_i, C_j) = \frac{s(v, C_i, C_j) - min_v}{max_v - min_v} \tag{3}$$

min_v と max_v は，それぞれ標本 X_v における類似度の値の最小値，最大値である．s_R は，C_i と C_j の類似度の大きさを X_v 上で相対的に評価している．

s_R による正規化は，標本 X_v 内における類似度の分布の偏りは考慮されていない．そこで，ベクトル間類似度を正規化する別の方法として式 (4) を考える．

$$s_{SD}(v, C_i, C_j) = \frac{10(\ s(v, C_i, C_j) - \mu_v\ )}{\sigma_v} + 50 \tag{4}$$

μ_v と σ_v は，それぞれ標本 X_v における平均と標準偏差である．ただし，用例間の類似度が 0 になる場合は X_v から除く．s_{SD} は標本 X_v における $s(v, C_i, C_j)$ の偏差値である．

2.2.3 停止条件

以下の 2 つの条件を同時に満たすとき，クラスタリングを停止する．

1. クラスタの数が T_n 以下である．
2. 大きさが最大のクラスタの要素数の用例総数に対する割合が T_s $(0 < T_s < 1)$ より大きい．

2. の条件はある程度の数の用例をまとめたクラスタが作成されるまでクラスタリングを継続させるために設定した．4.1 項の実験では仮に $T_n = 10$，$T_s = 0.2$ とした．

3 新語義の判定

用例クラスタリングの後，作成された用例クラスタに対して，それが新語義の用例であるかの判定を行う．ここでは新語義判定の問題設定を以下のように定義する．クラスタ集合を $\mathcal{C} = \{C_1, \cdots, C_n\}$，辞書で定義されている語義の集合を $\mathcal{S} = \{S_1, \cdots, S_m, NS\}$ とする．NS は新語義を表わす．用例クラスタと語義とを対応付ける関数 (マッピング関数) $M : \mathcal{C} \to \mathcal{S}$ を考える．M のスコアを $score(M)$ とし，それが最大となる M を選択する．選択された M によって新語義 NS に対応付けられた用例クラスタを新語義の用例とみなす．

3.1 マッピング関数のスコア

マッピング関数のスコア $score(M)$ は式 (5) のように定義する．

$$score(M) = \alpha \frac{1}{|A|} \sum_{(i,j) \in A} sim_{cc}(C_i, C_j) + (1 - \alpha) \frac{1}{|B|} \sum_{i \in B} sim_{cs}(C_i, M(C_i)) \tag{5}$$
$$A = \{(i, j) | i \neq j \ \&\ M(C_i) = M(C_j)\}, \quad B = \{i | M(C_i) \neq NS\}$$

式 (5) における $sim_{cc}(C_i, C_j)$ は 2 つのクラスタ C_i と C_j の類似度，$sim_{cs}(C_i, S_j)$ はクラスタ C_i と語義 S_j の類似度を表わす．これらの計算方法は 3.2 項で述べる．式 (5) の第 1 項は，M によって対応付けられる語義が等しいクラスタの組に対するクラスタ間類似度の平均である．すなわち，同じ語義に対応付けられるクラスタがどれだけ似ているかを評価する．一方，第 2 項は，クラスタ C_i とそれに対応付けられる語義 $M(C_i)(= S_j)$ との類似度の平均である．ただし，新語義に対応付けられるクラスタは考慮しない．すなわち，第 2 項はクラスタと語義がどれだけ似ているかを評価する．α は両者に対する重みである．$score(M)$ は以下の考えに基づいて設計されている．各クラスタに対して，それと対応する語義を決める際，クラスタと語義との類似度を計算し，類似度が大きい語義を選択する．これは式 (5) の第 2 項で評価される．ただし，クラスタが辞書に定義されているどの語義とも似ていない場合には，新語義 NS へ対応付ける方が第 2 項の平均値が高くなり，$score(M)$ も高く

なる．また，互いに似ているクラスタは同じ語義に対応付けるようにする．これは式 (5) の第 1 項により評価される．

マッピング関数 M の数は $(|\mathcal{S}|+1)^{|\mathcal{C}|}$ である．この数は一般に非常に大きくなるため，全ての M について $score(M)$ を計算することは難しい．そこで，山登り法を用いて $score(M)$ が最大となる M を近似的に求める．まず，各クラスタについて，$sim_{cs}(C_i, S_j)$ が最大となる語義 S_j を初期の対応付けとする．次に，対応関係をランダムに変化させ，もし $score(M)$ が大きくなれば M を更新する．この操作をあらかじめ設定された回数だけ繰り返す．

3.2 類似度計算

ここではクラスタ間類似度 $sim_{cc}(C_i, C_j)$ ならびにクラスタと語義の類似度 $sim_{cs}(C_i, S_j)$ の計算方法について述べる．まず，用例クラスタ C_i を特徴ベクトル $\vec{c}_i$ に変換する．同様に，語義 S_j も特徴ベクトル $\vec{s}_i$ に変換する (田中他, 2009)．$sim_{cc}(C_i, C_j)$ は $\vec{c}_i$ と $\vec{c}_j$，$sim_{cs}(C_i, S_j)$ は $\vec{c}_i$ と $\vec{s}_j$ のコサイン類似度とする．

用例クラスタの特徴ベクトル $\vec{c}_i$ は式 (6) によって作成する．

$$\vec{c}_i = \frac{1}{N} \sum_{e_{ik} \in C_i} \sum_{t_l \in e_{ik}} \vec{o}(t_l) \tag{6}$$

e_{ik} は用例クラスタ C_i に含まれる用例，t_l は用例 e_{ik} の文脈に出現する自立語，$\vec{o}(t_l)$ は単語 t_l の共起ベクトル，N は $\vec{c}_i$ の大きさを 1 にするための正規化定数である．共起ベクトル $\vec{o}(t_l)$ は，t_l とコーパスにおける出現頻度の上位 10,000 語との共起確率を素性とするベクトルであり，コーパスから事前に獲得しておく．用例の文脈に直接出現する単語 t_l だけでなく，その共起ベクトル $\vec{o}(t_l)$ の和を特徴ベクトルとすることにより，対象語 w と間接的に共起する単語の特徴が $\vec{c}_i$ に反映される．

一方，語義の特徴ベクトル $\vec{s}_i$ は辞書の語釈文から作成する．ここでは辞書として岩波国語辞典 (西尾他, 1994) を用いる．$\vec{s}_j$ の定義式を式 (7) に示す．

$$\vec{s}_j = \frac{1}{N} \left(\sum_{t_k \in d_j} \vec{o}(t_k) + \sum_{t_l \in e_j} w_e \cdot \vec{o}(t_l) \right) \tag{7}$$

d_j は語義 S_j の定義文を，e_j は辞書に記載されている例文を表わし，それらに含まれる自立語の共起ベクトルの和を $\vec{s}_j$ とする．ここで「定義文」は単語の語義を解説した文,「例文」はその語義を用いた例文である [5]．また，ここでは用例クラスタと辞書の語義との類似度を測ることを目的としているが，意味の説明文である「定義文」よりも語義の使用例である「例文」の方が，用例クラスタの文脈に出現する単語と似ている単語が出現する傾向が強いと予想される．そのため，例文に出現する単語に対して高い重み (w_e) を与える．ここでは $w_e = 2.0$ としている．

一般に，辞書に記載されている定義文や例文は短いため，語義の特徴ベクトルがスパースになり，用例クラスタと辞書の語義との類似度を測るのに十分な情報が得られない．そこで，語義タグ付きコーパスを用意し，辞書に記載されている例文に加え，語義タグ付きコーパスから抽出した語義 S_i の例文も式 (7) の e_j として用いる．

4 評価実験

4.1 用例クラスタリング手法の評価

まず，用例クラスタリングの評価実験について述べる．評価実験には SemEval-2 日本語タスク (Okumura et al., 2010) の訓練データを利用した．同タスクの 40 語の評価単語に対し，それぞれ 40〜50 語の用例を訓練データから抽出し，用例集合 W を作成する．W をクラスタリングして得られたクラスタ集合 $\mathcal{C}$ を，用例に付与されている語義を正解ラベルとして評価する．一般に，語義識別のタスクでは，同じ語義を持つ用例をまとめてクラスタを作成することと，語義の数を推定する (語義と同じ

[5] 図 2 右にある【サービス】の例では,「サービスのよい旅館」のように括弧で囲まれている文が例文である．

数だけクラスタを作成する) ことの 2 つが要求される．しかし，本研究は，作成された用例クラスタに対し，それが辞書に定義されている語義か否かを自動判定することで，コーパスから新語義を発見することを想定している．そのため，必ずしも語義の数を推定する必要はなく，同じ語義を持つ用例をまとめたクラスタを作成することが要求される．上記の理由から，今回の実験ではクラスタの評価基準として Purity (Hotho et al., 2005) と Homogeneity (Rosenberg and Hirschberg, 2007) を採用した．これらはクラスタを構成する要素のラベルがどれだけ一致しているかを評価する指標である．

表 1: 実験結果

(A)

	Purity	Homogeneity
提案手法 (s_R)	0.771	0.357
提案手法 (s_{SD})	0.800	0.472
[九岡ら 2008]	0.751	0.294
隣接	0.811	0.487
文脈	0.750	0.282
連想	0.749	0.285
トピック	0.765	0.374
BL	0.745	0.327

(B)

	$\|\mathcal{C}\|$	$\|\mathcal{C}_{\geq 2}\|$	AP
提案手法 (s_R)	400	258	0.857
提案手法 (s_{SD})	396	347	0.828
隣接	400	211	0.819
文脈	400	99	0.758
連想	400	103	0.772
トピック	400	233	0.767

40 語の評価単語に対する Purity と Homogeneity の平均を表 1(A) に示す．表の 2,3 行目は提案手法で，ベクトル間類似度を正規化する方法として式 (3) と式 (4) を用いた場合を表わす．4 行目は 4 つの特徴ベクトルを単独で用いたクラスタリング結果から評価単語ごとに最良のものを自動選択する九岡らの手法 (九岡他, 2008) を表わす．5〜8 行目は隣接，文脈，連想，トピックベクトルを単独で用いたときの結果である．最後の「BL」はベースラインを表わし，凝集型クラスタリングアルゴリズムで併合する要素の組をランダムに選択する手法である．

提案手法は九岡の手法よりも Purity，Homogeneity ともに上回ることから，複数の特徴ベクトルを利用する手法として適しているといえる．また，正規化の手法としては s_{SD} の方が s_R よりも良かった．しかし，提案手法は隣接ベクトルのみを使用する手法より少し劣る．この要因を調べたところ，単独のベクトルを使用した場合には，どの要素ともマージされずに 1 つの要素だけで構成されるクラスタが多いことがわかった．このようなクラスタは明らかに有用ではない．しかし，Purity や Homogeneity はクラスタ内に同じラベルを持つ要素がどれだけまとめられるかを評価する指標なので，1 要素で構成されるクラスタが多いときには高く見積られる．

表 1(B) は提案手法を別の観点で評価した結果である．$|C|$ は評価単語 40 語の全てについて作成されたクラスタの総数を，$|C_{\geq 2}|$ はそのうち 2 つ以上の要素から構成されているクラスタの数を表わす．また，AP の定義は式 (8) であり，要素数が 2 以上のクラスタ C_i について，C_i 内で頻度が最大となる語義が占める割合 ($\mathrm{max\_prec}(C_i)$) の平均である．

$$AP = \frac{1}{|\mathcal{C}_{\geq 2}|} \sum_{C_i \in \mathcal{C}_{\geq 2}} \mathrm{max\_prec}(C_i) \tag{8}$$

提案手法は，単独のベクトルを用いる手法と比べて $|C_{\geq 2}|$ が大きいことから，他のどの用例ともマージされない用例の数が少ないという意味ではクラスタリングに成功しているといえる．また，提案手法の AP も単独のベクトルを用いる手法と比べて高い．すなわち，2 個以上の要素をまとめて作成されたクラスタについては，同じ語義を持つ用例をまとめる傾向が強い．したがって，新語義を発見するための用例クラスタリング手法として，複数の特徴ベクトルを同時に考慮する提案手法は 1 種類の特徴ベクトルのみを用いる手法よりも優れていると言える．類似度の正規化の手法 s_R と s_{SD} を比較すると，AP は s_R の方が大きいが，$|C_{\geq 2}|$ は S_{SD} の方が大きかった．

表 2: 新語義判定の評価

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Accuracy	0.308	0.571	0.706	0.731	0.745	0.747	0.743	0.727	0.704
NS(prec)	0.192	0.300	0.411	0.468	0.504	0.534	0.525	0.560	0.600
NS(recall)	0.962	0.937	0.823	0.747	0.722	0.696	0.658	0.646	0.646
NS(F)	0.321	0.454	0.549	0.577	0.594	0.604	0.584	0.600	0.621

4.2 新語義判定手法の評価

次に，3 節で述べた新語義判定手法の評価実験を行った．対象単語は SemEval-2 日本語タスクの評価単語 49 語である．同タスクのテストデータをクラスタリングの対象とする用例集合 (1 単語あたり 50 用例) とした．一方，訓練データは式 (7) で定義した語義のベクトルの作成に用いた．テストデータの中には新語義を正解の語義とする用例はあるが，その数は少ない．そこで，対象単語に新語義を正解とする用例が 1 つもない場合には，出現回数が一番小さい語義を辞書から除き，仮想的な新語義とみなした．また，ここでは新語義判定の手法だけを評価するために，用例クラスタは 2 節の手法で作成するのではなく，正解のクラスタ集合を与えた.「正解のクラスタ集合」とは，(1) クラスタの数を 10 と設定する，(2) 1 つのクラスタは同じ語義を持つ用例で構成する，(3) クラスタの大きさ (用例数) をなるべく均等にする，という 3 つの条件を満たすクラスタ集合とした．このようにして作成されたクラスタ集合に対し，提案手法により各クラスタが辞書のどの語義に対応するのか，あるいは新語義であるのかを決定した．

結果を表 2 に示す．2 行目の Accuracy はクラスタと語義の対応付けの正解率を表わす．一方，3〜5 行目はそれぞれ新語義判定の精度，再現率，F 値である．α は式 (5) における第 1 項と第 2 項に対する重みであり，これを 0.1 から 0.9 まで 0.1 きざみで変動させたときの結果を示している [6]．

Accuracy は $\alpha = 0.6$ のときに最大で 0.747，新語義判定の F 値 (NS(F)) は $\alpha = 0.9$ のときに最大で 0.621 となった．また，α を大きくすると F 値の改善が見込まれることから，α を 0.9 から 1 まで 0.01 きざみで変動させたところ，$\alpha = 0.98$ で F 値が最大となり，その値は 0.633 であった．一方，用例クラスタと語義の類似度 $sim_{cs}(C_i, S_j)$ を計算し，それが閾値 [7]以下の場合に新語義と判定する手法を試したところ，新語義判定の F 値は 0.538 となった．この結果から，提案手法は単純に用例クラスタと語義の類似度を測って新語義か否かを判定する手法よりも優れていることがわかった．これは，マッピング関数 M を決定する際に，似ているクラスタは同じ語義に対応させるという選好 (式 (5) の第 1 項) が有効に働いたことを示唆する．

5 関連研究

本研究は，辞書を使わずに語義を自動的に推定する語義推定 (Word Sense Induction) もしくは語義識別 (Word Sense Descrimination) と呼ばれるタスクと関連が深い．語義識別に関する研究の多くは，用例を特徴ベクトルで表現し，ベクトル間の類似度を基に用例をクラスタリングする．Schütze は，コーパスから単語の共起行列を学習し，それを基に対象語と他の語との二次共起 (間接共起) の情報を反映した特徴ベクトルを作成し，Buckshot と呼ばれるアルゴリズムでクラスタリングを行う手法を提案している (Schütze, 1998)．また，意味解析に関する評価型ワークショップ SemEval では，過去 2 回にわたって英語を対象とした語義識別のタスクが実施され，用例クラスタリングに関するシステムが報告されている (Agirre and Soroa, 2007; Manandhar et al., 2010)．これらの先行研究では用例を 1 種類の特徴ベクトルで表現しクラスタリングを行っているのに対し，提案手法では用例を様々な観点から見た複数の特徴ベクトルで表現し，これらを同時にクラスタリングに用いる点に特徴

[6] α の最適化は今後の課題である．
[7] 閾値は実験的に決定した．

がある．4.1 項の評価実験では，1 種類のベクトルのみを用いるよりも複数の特徴ベクトルを同時に用いる方が，新語義判定に適した用例クラスタを作成できることが確認された．

一方，語義識別に関する先行研究は，同じ語義を持つ用例をまとめてクラスタを作成するが，クラスタとしてまとめられた用例が辞書のいずれかの意味を持つことを仮定しており，それが辞書にない新しい語義であるかまでは判定していない．これに対し，提案手法は，用例クラスタが辞書のどの語義にも対応しない新語義なのかを判定する点に新規性がある．

6 おわりに

本論文では，コーパスから新語義を自動的に発見する手法について述べた．コーパスから新語義を発見することはチャレンジングな課題であり，現状では実用レベルの精度で新語義の発見ができるとは言い難い．用例クラスタリング，新語義判定の両方について，手法の更なる洗練が求められる．用例クラスタリングについては，例えば構文的係り受けを考慮した特徴ベクトルなど，特徴ベクトルの種類を増やすことが考えられる．一方，新語義判定については，式 (5) におけるパラメタ α の最適化などが今後の課題として残されている．これらの課題について順次取り組んでいきたい．

文献

Eneko Agirre and Aitor Soroa (2007) “SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems,” in *Proceedings of SemEval-2007*, pp. 7–12.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003) “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Thomas Hofmann (1999) “Probabilistic Latent Semantinc Indexing,” in *Proceedings of the SIGIR*, pp. 50–57.

Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß(2005) “A Brief Survey of Text Mining,” *GLDV-Journal for Computational Linguistics and Language Technology*, Vol. 20, No. 1, pp. 19–62.

九岡佑介、白井清昭、中村誠 (2008) 「複数の特徴ベクトルのクラスタリングに基づく単語の意味の弁別」，第 14 回言語処理学会年次大会，pp.572–575.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan (2010) “SemEval-2010 Task 14: Word Sense Induction & Disambiguation,” in *Proceedings of SemEval-2010*, pp. 63–68, July.

中西隆一郎、白井清昭、中村誠 (2011) 「複数の観点から定義された用例間類似度に基づく語義識別」，言語処理学会第 15 回年次大会発表論文集.

西尾実、岩淵悦太郎、水谷静夫 (1994) 岩波国語辞典 第五版，岩波書店.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2010) “SemEval-2010 Task: Japanese WSD,” in *Proceedings of SemEval-2010*, pp. 69–74.

Andrew Rosenberg and Julia Hirschberg (2007) “V-measure: A Conditional Entropy-based External Cluster Evaluation Measure,” in *Proceedings of the 2007 EMNLP-CoNLL Joint Conference*, pp. 410–420.

Hinrich Schütze (1998) “Automatic Word Sense Discrimination,” *Computational Linguistics*, Vol. 24, No. 1, pp. 97–123.

田中博貴、中村誠、白井清昭 (2009) 「新語義発見のための用例クラスタと辞書定義文の対応付け」，言語処理学会第 15 回年次大会発表論文集，pp.590–593.

書　名　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
発行日　平成23年3月10日
発行者　文部科学省科学研究費特定領域研究「日本語コーパス」総括班
http://www.tokuteicorpus.jp/
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）
文書管理番号　JC-G-10-02