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

著者	特定領域研究「日本語コーパス」総括班
ページ	1-245
発行年	2009-03-11
URL	http://doi.org/10.15084/00003342

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

平成21年3月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-08-03

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

2009年3月15日（日）／16日（月）

Program [プログラム]

3月15日（日）

10:00　■開　会

10:00～10:20　■領域代表者報告

「中間評価を終えて」前川 喜久雄

10:20～12:20　■計画班研究発表

「『現代日本語書き言葉均衡コーパス』における固定長サンプルと可変長サンプルの比較」山崎 誠

「オントロジーに基づく言語的アノテーション」橋田 浩一

「ジャンル別UniDic作成の試み」小木曽 智信、伝 康晴、渡部 涼子

「文法の中核と周辺 ―コーパスが観察可能にする文法の一面―」田野村 忠温

12:20～13:00　休憩・昼食

13:00～15:30　■デモ・ポスターセッション（順不同）

「『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―」
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』における電子化フォーマットとその応用」
山口 昌也、間淵 洋子、西部 みちる、小林 正行、大島 一、高田 智和

「書籍コーパス（流通実態サブコーパス）の『外字』」
高田 智和、小林 正行、間淵 洋子、西部 みちる、大島 一、山口 昌也

「著作権処理の進捗状況と著作権法改正の動きについて」
森本 祥子、前川 喜久雄、小沼 悦、新井田 貴之、大石 有香、神野 博子、竹内 ゆかり、舞木 右

「『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」
小椋 秀樹、小木曽 智信、小磯 花絵、冨士池 優美、宮内 佐夜香、渡部 涼子、竹内 ゆかり、小川 志乃、小西 光、原 裕、中村 壮範

「形態論情報データベースの構成」
小木曽 智信、小椋 秀樹、小磯 花絵、冨士池 優美、宮内 佐夜香、渡部 涼子、竹内 ゆかり、小川 志乃、小西 光、原 裕、中村 壮範

「拡張固有表現タグ付きコーパスの構築に向けて ―白書、書籍、Yahoo! 知恵袋コアデータ―」橋本 泰一

「タグ付きコーパス管理ツール『茶器』の現状と今後」松本 裕治、浅原 正幸、岩立 将和、森田 敏生

「汎用アノテーションツールSLATにおける階層構造をもつタグセットのためのインターフェース」
松井 信太朗、野口 正樹、飯田 龍、徳永 健伸

「BCCWJに見られるオノマトペの型と共起との関連」ホドシチェク・ボル、ペケシュ・アンドレイ、仁科 喜久子

「短単位を対象とした連濁の処理について」山田 篤

「Yahoo! 知恵袋にみる非規範的表現」杉本 武

「大規模コーパスの語彙統計情報の利用を支援する ―語彙情報データベースを参照するAPIの構築と活用―」
千葉 庄寿

「コーパスを用いた公共性の高い文章における表記改善への視点」斎藤 達哉

「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―」
近藤 明日子

「白書およびYahoo! 知恵袋を対象にした結合価の自動抽出 ―格助詞パターンに着目して―」荻野 孝野

「異ジャンルの種用例を用いた半教師有りクラスタリングとその語義曖昧性解消に関する効果」
杉山 一成、奥村 学

「複数の語義を積極的に取り出す動詞のクラスタリング」高橋 秀幸、竹内 孔一

「BCCWJを用いた新しい語義曖昧性解消タスク」奥村 学、白井 清昭

「フレーム意味論と『日本語コーパス』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」
小原 京子、斎藤 博昭

「日本語リーダビリティー公式の構築と測定ツールの開発」柴崎 秀子

「グラフクラスタリングを用いた語義別用例分類」佐々木 稔、新納 浩幸

「ジャンル別に見るガ格を取る名詞と共起する用言の差異」野口 慎一朗、仁科 喜久子

「規則処理のアクセント属性を導入したCRFによるアクセント結合処理」印南 圭祐、峯松 信明

「Yahoo! 知恵袋コーパスのこれから ―さらなる研究支援のために―」片山 玲文、山本 健一、岡本 真

15:30～17:30 **■計画班研究発表**

「BCCWJを利用した日本語教育語彙リスト作成の試み」橋本 直幸

「語彙政策とコーパス ―医療用語を例に―」田中 牧郎

「コーパス中の日本語の間違い」荻野 綱男

「用例クラスタと辞書の語義との対応付けによる新語義の発見」白井 清昭、中村 誠、田中 博貴

17:30 **■閉　会**

3月16日（月）

10:00 **■開　会**

10:00～12:00 **■公募班研究発表**

「所謂引用助詞『と』が標識する構文の用法再考 ―フレーム・フレーム要素・フレーム間関係の観点から―」
藤井 聖子

「エントロピーと冗長度を指標とした語彙的・統語的複合動詞の比較研究」玉岡 賀津雄

「用例間類似度測定のための属性重みの推定」新納 浩幸、佐々木 稔

「BCCWJにおける推量副詞とモダリティ形式の共起」
スルダノヴィッチ・イレーナ、ベケシュ・アンドレイ、仁科 喜久子

12:00～13:00 **休憩・昼食**

13:00～14:00 **■講　演**

「Disruptive Serviceを目指して：情報爆発プロジェクトと情報大航海プロジェクト」
喜連川 優（東京大学生産技術研究所）

14:00 **■閉　会**

Contents [目次]

領域代表者報告

3月15日（日）　10:00〜10:20

中間評価を終えて

▶前川 喜久雄

中間評価を終えて

前川喜久雄（領域代表者：国立国語研究所研究開発部門）†

After the Interim Assessment

Kikuo Maekawa (Program supervisor. National Institute for Japanese Language)

１．中間評価

特定領域研究「日本語コーパス」も、2006 年 9 月の発足以来 30 か月が経過しました。今年度は中間評価の年にあたりましたので、その話題から報告をはじめることにします。

特定領域研究課題のうち 5 年計画のプロジェクトは、3 年次に中間評価を受けることが決まっています。私どもの「日本語コーパス」も昨年 9 月に中間評価用資料を提出し、10 月 20 日に文科省が構成した評価委員会によるヒアリングを受けました。後日発表された評価の全文を以下に掲載します。

評価結果：Ａ （現行のまま推進すればよい）

（中間評価に係る意見）

本研究領域が目的に掲げた大規模日本語書き言葉コーパスの構築は、今後の日本語研究の基盤のひとつであることは言をまたず、これが特定領域研究において推進されていることのもつ意義は高い。コーパス構築は当初の計画よりもその進度は速く、さらに、構築されているコーパスは質的な面においても他国で構築されているコーパスを凌駕するものと判断する。

各計画研究の研究進展状況も順調であり、コーパス構築に必要な解析ツールを領域内の計画研究で開発する等、それぞれの特長を活かした研究成果が領域内で有機的に連結している点において、領域の目的、成果への見通しが各研究課題において共有されているものと判断する。また、これらの研究について、本邦及び海外における成果発表が積極的になされており、学術的意義は高い。

一方で、今後の課題として求められていることは、専門研究者でない方々、日本語話者以外の方々にも利用が容易になるような配慮を加え、構築したコーパスの継続的な活用のための見通しを明らかにしていくことである。国立国語研究所の組織改編によるURI（URL）の移転、混乱などにあらかじめ周到な準備を行なうなど、領域研究終了後のサービスの安定性にも注意をはらいつつ、本研究領域の研究を今後も推進していくことを望む。

「A」は最高のランクです。この評価をうけるプロジェクトがどれぐらいあるかは、興味のあるところですが、他プロジェクトの評価結果は未だ詳らかにしません。ただし、昨年 12 月には文科省から中間評価が「A」ないし「A－」のプロジェクトに追加予算を配分するので希望者は応募するようにという通達があり、実際、われわれもその恩恵に与ることができました。実質的に高い評価がえられていることの傍証だと思います。

本プロジェクトに対する学界の関心も依然として高いものがあります。国語国文学の専門商業誌『国文学解釈と鑑賞』は、74 巻 1 号（2009 年 1 月号）を特集「日本語研究とコーパス」にあてました。私も寄稿しましたが、「日本語コーパス」からは山崎誠、小椋秀樹、小木曽智信、松本裕治、小磯花絵、白井清昭、荻野綱男、田野村忠温、小林ミナ、投野由紀夫、丸山岳彦の各氏がさまざまなトピックについて寄稿しています。また 2009 年 9 月に

† kikuo@kokken.go.jp

は日本人工知能学会誌が特集として「日本語コーパス」関係者の寄稿を掲載することが決定しています。

海外の動向にもふれます。韓国では、雑誌『언어 정보와 사전 편찬』(言語情報と辞書編集) 22 巻に、延世大学の徐尚揆(ソ・サンギュ)氏の手になる KOTONOHA 計画と BCCWJ の非常に詳しい解説論文が掲載されています。また今月末に香港で開催される The 22nd International Conference on the Computer Processing of Oriental Languages (ICCPOL2009) でも招待講演のひとつが KOTONOHA 計画にあてられます。東アジア圏において KOTONOHA 計画は広く認知されはじめているようです。

最後に、常用漢字表の改定作業で BCCWJ の書籍データが利用されはじめたことも記しておこうと思います。文化審議会国語分科会漢字小委員会では、凸版印刷から提供された 5000 万字規模の印刷データを漢字の取捨選択の参考資料としていましたが、このデータがどこまで日本語を代表しているかについては、委員会の内外から問題が指摘されていました。この問題に対する客観的な解決を探る手段として BCCWJ の書籍データ(現時点においてもっとも均衡性が高い部分)の検索結果が利用され、結果として、問題の解決に寄与しえたことは、漢字小委員会だけでなく本領域にとっても非常に有意義な出来事でした。

２．コーパス構築状況

次にコーパスの構築状況を概観します。本プロジェクトは発足時からふたつの研究項目で構成されています。A01「コーパスの構築」と B01「コーパスの評価」のふたつです。B01 は A01 の成果であるコーパスを利用しなければ始まらない研究ですから、プロジェクト前半の大目標はできるだけ早く B01 が活発に活動するに足るだけのデータを提供することにあり、そのためにはデータ班に活発に活動してもらう必要がありました。本領域の提案書では、第 3 年次(つまり本年度)末の達成目標として、6300 万語程度のデータを作成することを掲げていましたが、幸いこの目標は余裕をもって達成できました。

詳しくはデータ班から別途発表がありますが、現時点でサンプリングと文字入力が終了したサンプルは 7000 万語を超えています。そのうち約 5300 万語分は昨年 7 月末に「BCCWJ 領域内公開データ(2008 年版)」として、特定領域関係者に配布しました。このデータは厳密に言えばまだ均衡したものになっていませんが、従来の大規模日本語テキストデータ(新聞、青空文庫、国会会議録など)には欠けていた現代の書籍のデータを大量(3000 万語分以上)に含んでおり、研究目的によっては、すでに実用に耐えるデータになっていると考えられます。

領域内公開データの配布に先立つ 7 月初旬には、領域内公開データのうち著作権処理が完全に終了している約 2800 万語分を「BCCWJ モニター公開データ(2008 年度版)」として、領域外の研究者にモニター公開しました。研究目的に限定した公開ですが、営利企業の研究者にも利用していただくことができ、本稿の執筆時点で 360 件以上の利用申請がおこなわれています。

３．著作権法をめぐる動向

本プロジェクトにおけるコーパス構築の最大の隘路が著作権処理であることは、これまでにも指摘してきました。昨年度の公開ワークショップでは、シンポジウム「知識資源と著作権」を設けて、利用者と権利者の双方から意見を聞く場としました。そのときは、予期していなかったのですが、その後今日までの 1 年間に著作権法改正の具体的な動きが急

速に表面化してきました。

2008 年 6 月 28 日に発表された「知的財産推進計画 2008」には著作権法改正にむけて大胆な提案が盛り込まれていました。「（２）内外リソースの積極活用のための環境を整備する」のうち「①研究開発における情報利用の円滑化に係る法的課題を解決する」の全文を以下に引用します。末尾に「(文部科学省)」とあるのは、これが文科省（具体的には文化庁著作権課）に対する要請であることを意味しています。

> **ネット等を活用して膨大な情報を収集・解析することにより高度情報化社会の基盤的技術となる画像・音声・言語・ウェブ解析技術等の研究開発が促進されること等を踏まえ、これらの科学技術によるイノベーションの創出に関連する研究開発については、権利者の利益を不当に害さない場合において、必要な範囲での著作物の複製や翻案等を行うことができるよう２００８年度中に法的措置を講ずる。（文部科学省）**

この一文の主旨はインターネットにおけるクローリングの合法化にありますが、この提案を受けて、法改正にむけての取り組みがはじまっています。7 月 25 日には文化庁の文化審議会著作権分科会法制小委員会が関係者のヒアリングを実施し、特定領域研究「情報爆発」プロジェクトの代表者である喜連川優先生と私が学術研究を代表して意見を述べました。当日の質疑の詳細な議事録がウェブ上で公開されていますが、クローリングによる情報収集に関するかぎり、委員からの明確な反対意見は表明されませんでした（http://www.bunka.go.jp/chosakuken/singikai/housei/h20_05/gijiroku.html)。

その後、10 月には「文化審議会著作権分科会法制問題小委員会平成 20 年度・中間まとめ」が公開され、パブリックコメントに付されました。その結果も現在ウェブで閲覧することができます（「電子政府の総合窓口」の「意見募集の結果一覧」参照)。現在、法律改正の準備が進められているものと思われます。

ところで、今回の著作権法改正の動きのなかでは、クローリングの問題とは別個に「日本版フェアユース規定」の導入が議論されていることが注目されます。これは「現行の著作権法は、個別具体の事例に沿って権利制限の規定を定めているため、これら規定に該当しない行為については、たとえ権利者の利益を不当に害しないものであっても形式的には違法となってしまう」（知的財産戦略本部デジタル・ネット時代における知財制度専門調査会「デジタル・ネット時代における知財制度の在り方について（報告案)」p.9）問題を解消するために、米国におけるようなフェアユース規定を日本の著作権法にも導入しようとするものです。

日本版フェアユース規定が具体的にどのような条文として記載されるかは未だ明らかでありませんが、米国と同様、利用者がこれはフェアユースであると判断すれば、権利者の許諾を待たずとも著作物を利用することができ（利用したこと自体によって罰されることがなく)、それを不当と考える権利者があれば、裁判に訴えて決着をつけることになるのだと思われます。学術的な利用を主要な目的としている我々のコーパスの場合、フェアユースを主張することには十分な合理性があると考えられますので、法律改正の動向に注目したいと思います。

４．プロジェクト後半の目標

特定領域研究「日本語コーパス」も残すところ 2 年となりました。プロジェクト前半で

はコーパスの構築に力を注ぎ、後半はコーパスを用いた研究、特に狭い意味での言語学におさまらない応用研究に力を注ぐという当初計画にしたがって、今後は研究項目 B01 に属する計画研究班の活動を活発化させていきます。もちろん、B01 の各班はすでに具体的な成果を発表してきているのですが、今後はその活動を一層活発化させていくということです。

しかし、本領域の活動目標のひとつである「コーパスによる日本語研究法の確立」を達成するためには、計画班の活動にくわえて、より広い範囲の研究者によるコーパス利用を促進する必要があります。

そのために本領域では従来から、公募研究制度を活用してきており、過去 2 年間に 4 件の公募研究班が活動していました。これらの公募研究班は今年度までで一応活動を終えますが、来春からはまた新しく 4 件の公募研究がスタートする予定です。

公募研究以外にも、本領域独自の試みとして、今年度の公開ワークショッでは初日にサテライトセッションを開催することにしました。このセッションで発表される 19 篇の論文は、すべて上述の「BCCWJ モニター公開データ」を利用した研究であり、計画班でも公募班でもない、領域外の研究者がおこなった研究です。この原稿を執筆している時点では、サテライトセッションがどのような成果を挙げるかを予見することはできませんが、願わくは、若手研究者の登竜門として、また超領域的な知見融合の場として機能してほしいと希望しています。BCCWJ のモニター公開とサテライトセッションは、平成 21 年度にも実施する予定です。

５．まとめにかえて

研究期間の 60%を終え、中間評価も無事終了したとなると、どうしても気が緩みがちになるのが人間です。しかし、プロジェクトの大成を望むならば、ここでもう一度気を引き締めなければなりません。

かつて『日本語話し言葉コーパス』の構築が最終段階をむかえた頃、『老子』に「慎終如始、則無敗事」という文字があるのを知り、古典というものの価値に深く思いをいたしたことがありました。私は、この文字をふたたび胸に刻みながら、今後の 2 年にむけて再出発しようと思います。

計画班研究発表

3月15日（日）　10:20〜12:20

『現代日本語書き言葉均衡コーパス』における固定長サンプルと可変長サンプルの比較

▶山崎 誠

オントロジーに基づく言語的アノテーション

▶橋田 浩一

ジャンル別UniDic作成の試み

▶小木曽 智信、伝 康晴、渡部 涼子

文法の中核と周辺 ―コーパスが観察可能にする文法の一面―

▶田野村 忠温

『現代日本語書き言葉均衡コーパス』における固定長サンプルと可変長サンプルの比較

山崎誠（データ班班長：国立国語研究所研究開発部門）†

Comparison between Fixed Length Sample and Variable Length Sample in BCCWJ

Makoto Yamazaki (Dept. Lang. Res., National Institute for Japanese Language)

１．はじめに

「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ と略す)には、目的に応じて固定長サンプルと可変長サンプルという 2 種類の異なるサンプルタイプがある。これらは研究目的の違いに応じて設計されたものである（山崎 2006）が、この２つのタイプはそもそも言語的に見てどのような違いがあるのか（あるいはないのか)、文字及び語彙の面から計量的に分析する。

２．固定長サンプルと可変長サンプル

固定長サンプルは、統計的に厳密な分析を目的として設計されている。1 サンプルは句読点などの記号類を除く 1,000 字から構成される。1,000 字というのは、BCCWJ で採用している短単位に換算するとおよそ 590 単位である 。

可変長サンプルは、文章の長さではなく内容的なまとまりを基準にして 1 サンプルの範囲を決定する。具体的には新聞・雑誌の 1 記事、書籍における章・節などのまとまりが 1 サンプルに該当する。ただし、無制限に長いサンプルが出来るとコーパスの分析に影響を与えるため上限を 1 万字としている。

３．使用するデータ

2008 年 7 月に配布を開始した「『現代日本語書き言葉均衡コーパス』モニター公開データ（2008 年度版)」に収録された書籍及び白書のサンプルのうち固定長サンプルと可変長サンプルの両方がそろっているサンプルを対象とした。具体的な数は次のとおりである。

書籍　　3,773 サンプル[1]
白書　　1,500 サンプル

書籍と白書の両方を対象とした理由は、両者は語彙や文体において異質の文章であることから、本稿で観察しようとしている固定長サンプルと可変長サンプルの差異の現れ方に違いが見られるかもしれないと判断したためである。分析に使用した形態素解析環境は、MeCab ver.0.97＋UniDic-1.3.9 である。なお、本稿では形態素解析の結果得られた短単位を便

† yamazaki@kokken.go.jp

[1] 候補となるサンプルは 3,795 サンプルであったが、後述の separated パターンの 22 サンプルは比較の条件を備えていないため分析の対象から除外した。

宜上「語」とみなして記述する。また、特に断らない限り分析には助詞・助動詞を含み、記号・符号を含まないこととする。品詞体系は UniDic に従う。

４．固定長サンプル、可変長サンプル全体の比較

表 1 は、対象となった各サンプル全体の延べ語数及び異なり語数と個々のサンプルごとの延べ語数及び異なり語数の平均値である。

表 1 固定長サンプルと可変長サンプルの概観

	書籍		白書	
	固定長	可変長	固定長	可変長
全体延べ語数	2,455,558	10,217,488	1,035,345	4,685,128
全体異なり語数	52,550	82,800	15,780	26,748
個別延べ語数(n)	647.2	2671.9	690.2	3213.4
個別異なり語数(k)	247.2	638.1	225.0	534.8
n/k値の平均	2.66	3.89	3.14	5.55

個々のサンプルの値で見ると、可変長サンプルは書籍で固定長サンプルの約 4.1 倍、白書で 4.6 倍の長さを持つ。延べ語数では固定長・可変長ともに書籍より白書の方が値が大きいが、異なり語数では逆に書籍の方が大きな値になっている。これは、白書の方が延べ語数の伸びに比べて異なり語数の伸びが鈍いことを意味している。1 語あたりの平均使用度数を表す n/k 値（延べ語数/異なり語数の値）も固定長・可変長ともに白書の方が大きくなっていることもそのことの表れである。固定長サンプルと可変長サンプルにおける n/k 値の分布のようすを図 1、図 2 に示した。

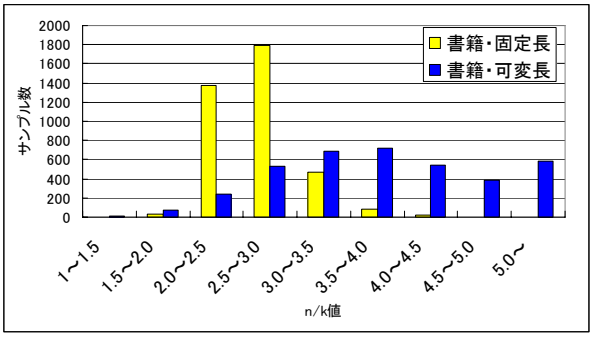


図 1　n/k 値の分布（書籍）

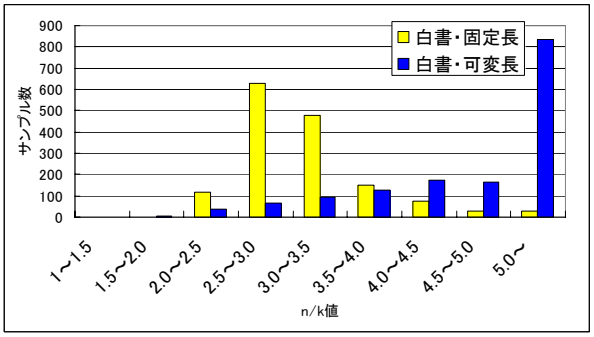


図 2　n/k 値の分布（白書）

次に文字種・語種・品詞の割合を見てみよう。表 2～4 は、それぞれ文字種、語種、品詞について、個々のサンプルにおける割合の平均値を示したものである。表 2～4 を見ると分かるように、書籍・白書ともに固定長サンプルと可変長サンプルの差異は無視できるほど小さい。これらのカテゴリの使われ方はサンプルタイプに影響を受けにくいと言える。

表 2　固定長・可変長サンプルにおける文字種の割合（延べ字数）

	書籍		白書	
	固定長	可変長	固定長	可変長
英数字	0.61	0.66	1.24	1.24
平仮名	59.19	59.21	38.66	38.87
片仮名	5.88	5.82	4.77	4.77
漢字	34.31	34.31	55.33	55.12

表 3　固定長・可変長サンプルにおける語種の割合（延べ語数）

	書籍		白書	
	固定長	可変長	固定長	可変長
和語	72.28	72.32	46.37	46.13
漢語	21.90	21.83	48.36	48.59
外来語	1.87	1.86	2.21	2.18
混種語	0.92	0.93	1.45	1.47
固有名	2.57	2.60	1.49	1.51
不明	0.16	0.16	0.06	0.06
なし	0.30	0.30	0.07	0.07

表 4　固定長・可変長サンプルにおける品詞の割合（延べ語数）

	書籍		白書	
	固定長	可変長	固定長	可変長
名詞	31.19	31.14	47.68	47.90
代名詞	1.64	1.64	0.27	0.27
動詞	14.61	14.60	10.71	10.65
形容詞	1.60	1.61	0.49	0.48
形状詞	0.98	0.97	0.60	0.60
連体詞	1.06	1.07	0.71	0.71
副詞	1.84	1.84	0.46	0.45
接続詞	0.44	0.44	0.95	0.95
感動詞	0.24	0.24	0.01	0.01
助詞	31.55	31.57	24.31	24.20
助動詞	10.05	10.07	4.26	4.25
接頭辞	0.74	0.73	1.25	1.25
接尾辞	4.07	4.07	8.29	8.29

サンプル全体を平均すると違いがないように見えるが、文字種、語種、品詞の割合の分布を個別に見るといくつか差異が見られる。図 3～図 6 は、書籍及び白書における個々のサンプルにおける文字種の割合の分布である。片仮名、平仮名、漢字の分布状況には差異はないが、英数字については書籍・白書ともに、割合が 0 すなわち英数字がひとつも現れなかったサンプルが固定長のほうに多いことが分かる。これは固定長サンプルが任意の位置からはじまる 1000 字なのに対して、可変長サンプルが章や節などの意味的なまとまりである

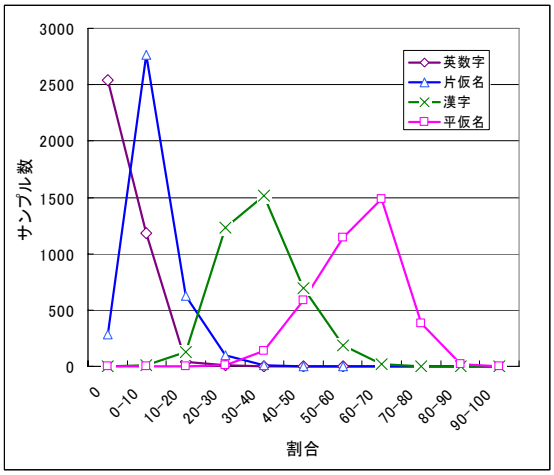


図 3　書籍固定長の文字種割合の分布

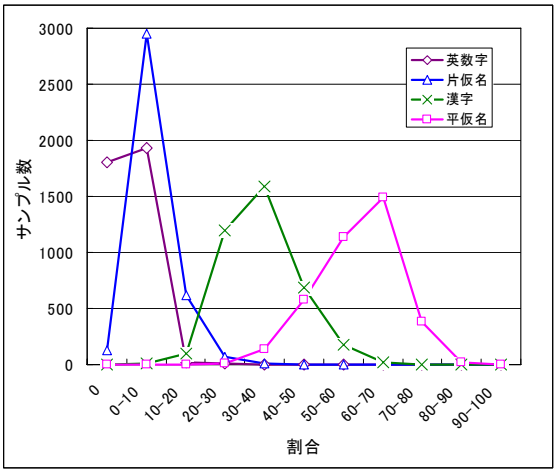


図 4　書籍可変長の文字種割合の分布

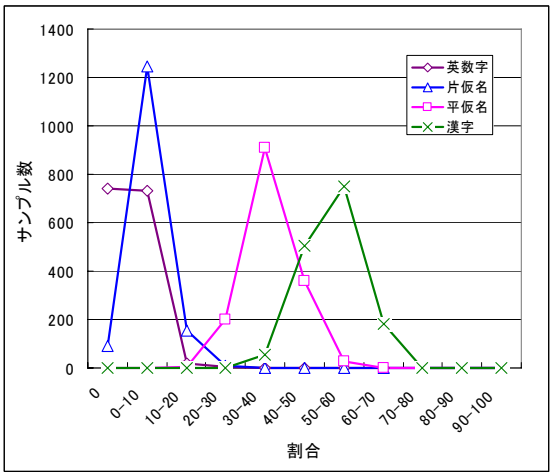


図 5　白書固定長の文字種割合の分布

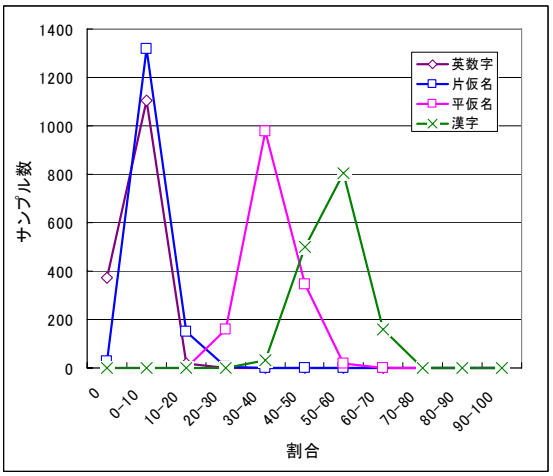


図 6　白書可変長の文字種割合の分布

ことから、章や節の見出しに関連する数字等が含まれる可能性が高いことの結果と推察される。図示はしないが、品詞では書籍において接頭辞の割合が 0 のサンプルが固定長で 257 個なのに対して可変長で 97 個、白書において形容詞の割合が 0 のサンプルが固定長で 77 個なのに対して可変長で 254 個と違いが見られた。接頭辞は見出しに使われやすいためと推察されるが形容詞については理由は不明である[2]。

表 5 に示したのは書籍の固定長・可変長のそれぞれにおける使用頻度の上位語 30 語の比較であるが、順位・使用率ともにほぼ同じである。表は掲載しないが、白書においても同様である。

表 5　上位語の比較(書籍)

	固定長			
順位	語彙素	表記	品詞	使用率
1	ノ	の	格助詞	0.05003
2	ニ	に	格助詞	0.03615
3	テ	て	接続助詞	0.03587
4	ハ	は	係助詞	0.03408
5	ダ	だ	助動詞	0.03354
6	タ	た	助動詞	0.03229
7	ヲ	を	格助詞	0.03178
8	ガ	が	格助詞	0.02441
9	スル	為る	動詞	0.02439
10	ト	と	格助詞	0.02362
11	モ	も	係助詞	0.01289
12	デ	で	格助詞	0.01235
13	イル	居る	動詞	0.01218
14	アル	有る	動詞	0.01105
15	ノ	の	準体助詞	0.01091
16	イウ	言う	動詞	0.00883
17	コト	事	名詞	0.00820
18	ナイ	ない	助動詞	0.00662
19	レル	れる	助動詞	0.00640
20	マス	ます	助動詞	0.00639
21	ナル	成る	動詞	0.00601
22	デス	です	助動詞	0.00517
23	ナイ	無い	形容詞	0.00511
24	カラ	から	格助詞	0.00475
25	ソノ	其の	連体詞	0.00443
26	ヨウ	様	形状詞	0.00409
27	ガ	が	接続助詞	0.00369
28	カ	か	副助詞	0.00339
29	ソレ	其れ	代名詞	0.00324
30	イチ	一	名詞	0.00314

	可変長		
語彙素	表記	品詞	使用率
ノ	の	格助詞	0.05028
ニ	に	格助詞	0.03620
テ	て	接続助詞	0.03597
ハ	は	係助詞	0.03446
ダ	だ	助動詞	0.03367
タ	た	助動詞	0.03279
ヲ	を	格助詞	0.03192
ガ	が	格助詞	0.02444
スル	為る	動詞	0.02433
ト	と	格助詞	0.02360
モ	も	係助詞	0.01291
イル	居る	動詞	0.01234
デ	で	格助詞	0.01219
アル	有る	動詞	0.01097
ノ	の	準体助詞	0.01096
イウ	言う	動詞	0.00884
コト	事	名詞	0.00812
ナイ	ない	助動詞	0.00679
レル	れる	助動詞	0.00635
マス	ます	助動詞	0.00607
ナル	成る	動詞	0.00595
ナイ	無い	形容詞	0.00522
デス	です	助動詞	0.00505
カラ	から	格助詞	0.00467
ソノ	其の	連体詞	0.00447
ヨウ	様	形状詞	0.00410
ガ	が	接続助詞	0.00362
カ	か	副助詞	0.00343
ソレ	其れ	代名詞	0.00330
カ	か	終助詞	0.00313

[2] 書籍、白書ともに感動詞・フィラーの割合にも差が見られたが、いくつか例にあたったところ誤解析の可能性が高いと判断したため結果には含めなかった。

固定長サンプルと可変長サンプルの範囲が異なるため、固定長サンプルにしか出現しない語あるいは可変長サンプルにしか出現しない語が存在する。表 6、表 7 はそれらを異なり語数レベル、延べ語数レベルで集計したものである。

表 6 固定長・可変長の一方にしか出現しない語(異なり)

	書籍		白書	
	語数	割合	語数	割合
固定長のみ	1,963	3.74	416	2.64
可変長のみ	32,214	38.91	11,384	42.56

表 7 固定長・可変長の一方にしか出現しない語(延べ)

	書籍		白書	
	語数	割合	語数	割合
固定長のみ	2,649	0.11	532	0.05
可変長のみ	79,170	0.77	27,682	0.59

固定長サンプルにしか出現しない語は、使用率の高い順に書籍では「滌除、*トワダ、*ナミエ、*クツナ、*兼六、*サクノスケ、スピリチュアリズム、スウェデン、邢、黄斑、*キリコ、ｔｅｌｌ」(頻度 7 以上)、白書では「スラッジ、*マレ、Ｐｕｂｌｉｃａｔｉｏｎ、主査、*セントルシア、Ｄｒｕｇｓ、サージ、*ダルフール、船溜、*アンティグア、稲叢、嗅覚、錘数、線分、*全労連、*テーラー、*フィゲーレス、領収、Ｄａｎｇｅｒｏｕｓ」(頻度 3 以上)である。*を附した語は形態素解析結果において固有名詞となっているものである。

同様に、可変長サンプルにしか出現しない語は、書籍では「*コウダユウ、蘇芳、シリウス、春雨、山車、*ジュネーブ、*足羽、*タカツネ、ストーマ、フルート、コウスケ、水虫、フランシーヌ、ダンベル、体節、景勝、思量、夕霧、*ライル」(頻度 36 以上)、白書では「燻蒸、県庁、通、氏名、着陸、*ＪＡＳ、通数、パン-pao、字幕、同感、定係、*ＪＩＳ、商船、方位、ボーナス、ミスマッチ」(頻度 30 以上)である。

両者を比較すると、可変長サンプルの方によりなじみのある語が含まれているようであるが、今後サンプル数が増えてくれば、可変長サンプルにしか出現しない語も固定長サンプルにしか出現しない語と同様の傾向を示すのではないかと思われる。

５．固定長サンプルと可変長サンプルの位置的関係

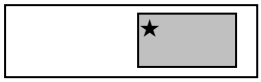
５．１．パターン

上記第 4 節で見たように、固定長サンプルと可変長サンプルは文字種、品詞、語種の割合や上位語の使用率においてほぼ同じであることが確認できたが、その最大の理由は、両者は包含関係にあるものが多いということだろう。以下その事情を説明する。

コーパス構築に当たって固定長と可変長のサンプルを別々に取得するのは作業コストがかかりすぎるため、BCCWJ では 1 回のサンプリングで当たった同一箇所から固定長と可変長の２つのサンプルを取得している(丸山・秋元 2007、同 2008)。そのため、固定長サンプルと可変長サンプルの間には包含関係を基本とする 3 種類の「パターン」が生じる。以下の図 7 に示す included, overflow, separated である。なお図 7 は左から右に文脈が進むとい

う前提で作図している。

included：固定長サンプルが可変長サンプルに包含される場合



→文脈の方向

overflow：固定長サンプルの一部が可変長サンプルからはみ出す場合

separated：固定長サンプルが可変長サンプル包含されない場合

図 7　固定長サンプルと可変長サンプルの関係

included は、固定長サンプルが可変長サンプルの中に完全に納まる場合である。図 7 の★は、ランダムに決められる「サンプル抽出基準点」を表すが、この文字を含む後続の 1,000 字が可変長サンプルの終端位置を超える場合は、overflow になる。separated は、第 2 節で述べた可変長サンプルの長さの制限(1 万字)のため、強制的に打ち切った可変長サンプルの終端が固定長サンプルの開始位置に届かなかった場合である。

今回の対象データにおけるパターンの分布は表 8 のとおりである。

表 8　サンプルのパターンの分布

	included	overflow	separated
書籍	2,319	1,454	22
白書	1,018	482	0

５．２．固定長サンプルの開始位置と言語的特徴

固定長サンプルの大多数は可変長サンプルの中にサンプルの開始位置を持つ。開始位置を決める「サンプル抽出基準点」はランダムに当てているため、固定長サンプルの開始位置は可変長サンプル内に均等に分布しているはずである。そのことを確認したのが図 8 である。図 4 は可変長サンプルをその長さにかかわらず 10 等分してその 10 個の区画のどこに固定長サンプルの開始位置が来るかを調べたものである。

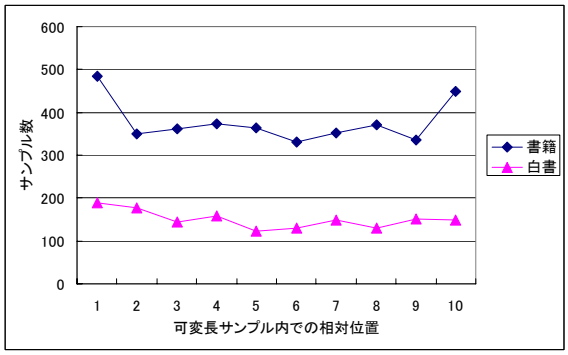


図 8　可変長サンプル内での相対位置で示した固定長サンプルの開始位置の分布

固定長サンプルが可変長サンプル内のどの位置から開始するかによって、固定長サンプルの言語的特徴に影響があると考えられる。例えば、上述の overflow パターンの場合、固定長サンプルには、一つの完結したまとまりの最後の部分とその次のまとまりの最初の部分とから構成されることになる。そのような状況で影響を受ける可能性がある指標は n/k 値である。同一の内容よりも異なる 2 つの内容の方が異なり語数を増やしやすいからである。そのことを確かめるために、可変長サンプル内での相対的位置と n/k 値との関連を調査した。

結果を図 9、図 10 に示す。書籍、白書ともに、可変長サンプル内での相対位置が後ろの方になるにつれて、n/k 値が低くなる傾向がある。白書では中間でいったん n/k 値がかなり下がるところがあるがどういう理由によるものかは分からないが、n/k 値の変動は語彙で表される同一の話題がどれだけの長さ継続するかということとも関連する。したがって、書

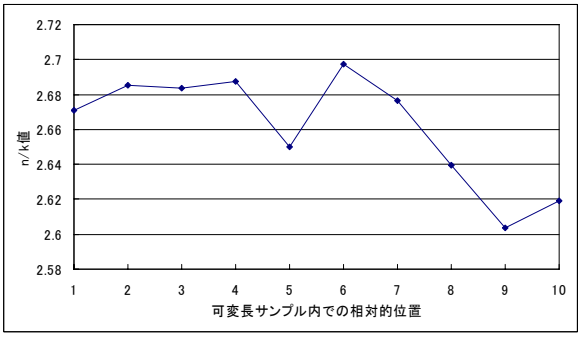


図 9　可変長サンプルにおける相対位置による固定長サンプルの n/k 値の分布（書籍）

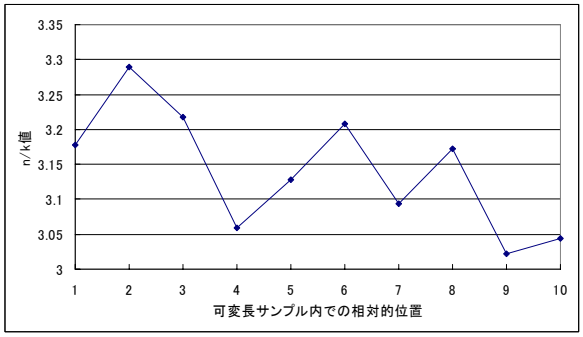


図 10　可変長サンプルにおける相対位置による固定長サンプルの n/k 値の分布（白書）

籍、白書ともに途中で値が下がっているということは、可変長サンプルの中間付近でそれ以前とは異なる語彙が多く出現しているということを示唆する。また、書籍も白書も最後の区画で n/k 値のわずかの上昇が認められるが、これはこの区画においては固定長サンプルの分量のほとんどが開始位置を持つまとまりの次のまとまりの中にあるということになり、2 つの異なる内容が同居する度合いが小さくなったためと思われる。

同様に品詞の割合を書籍、白書で調べたが、可変長サンプル内の相対的位置との関連は見いだせなかった(図 11、図 12)。

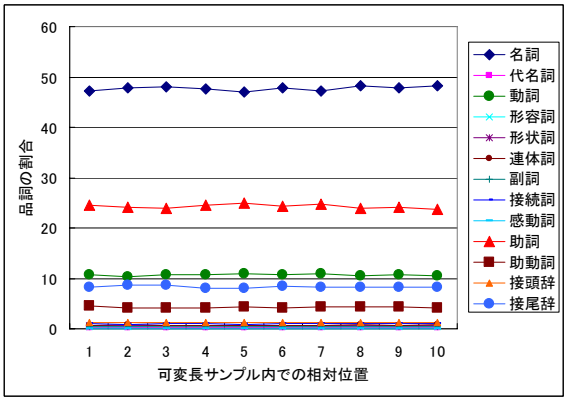


図 11　可変長サンプルにおける相対位置による固定長サンプルの品詞の割合の分布（書籍）

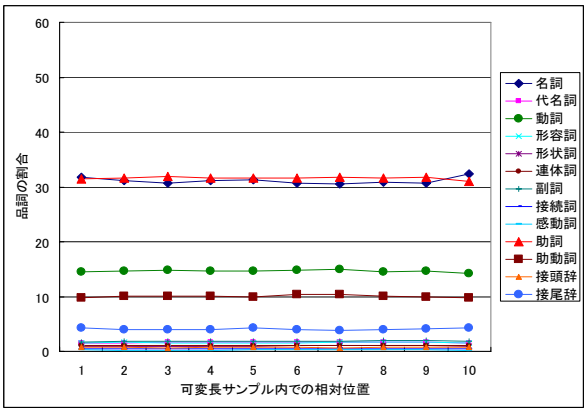


図 12　可変長サンプルにおける相対位置による固定長サンプルの品詞の割合の分布（白書）

６．まとめ

固定長サンプル、可変長サンプルは、文字種、品詞、語種等のマクロな値を見る限りではほぼ同じであり、等質なテキストであると言える。ただし、固定長サンプルの開始位置が可変長サンプルの末尾付近に当たる場合は、延べ語数と異なり語数との関係に変化が見られ、１語あたりの平均使用度数が低くなる傾向があることが分かった。BCCWJ を十分理解して使うために、今後様々な方法でサンプルの検証を行っていくことが必要である。

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）による補助を得た。

参考文献

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–』（LR-CCG-06-02）.

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－』（LR-CCG-07-01）.

山崎誠(2006)「『現代日本語書き言葉均衡コーパス』の基本設計について」、「特定領域研究「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.

オントロジーに基づく言語的アノテーション

橋田浩一（ツール班分担者：産業技術総合研究所サービス工学研究センター）†

Linguistic Annotation Based on Ontology

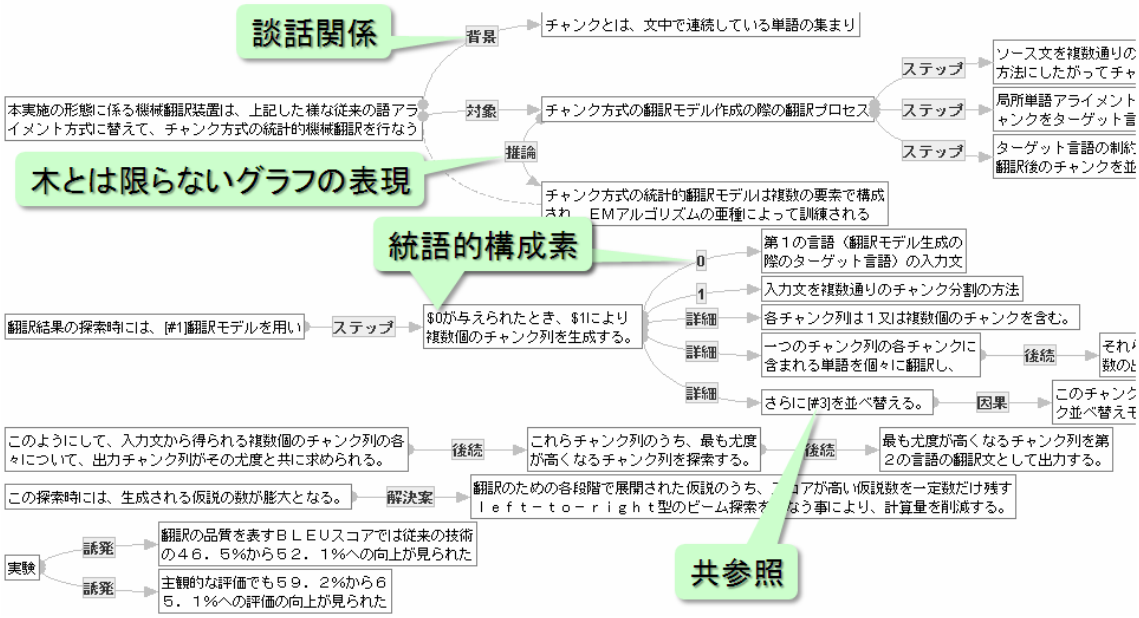
Koiti Hasida　(Center for Service Research, National Institute of Advanced Industrial Science and Technology)

1　概要

XML には標準的な意味論がないので、それに代わって RDF を用いて、つまりオントロジーに基づいて言語的な構造の記述を行う方法を研究する。

2　ネットワーク構造

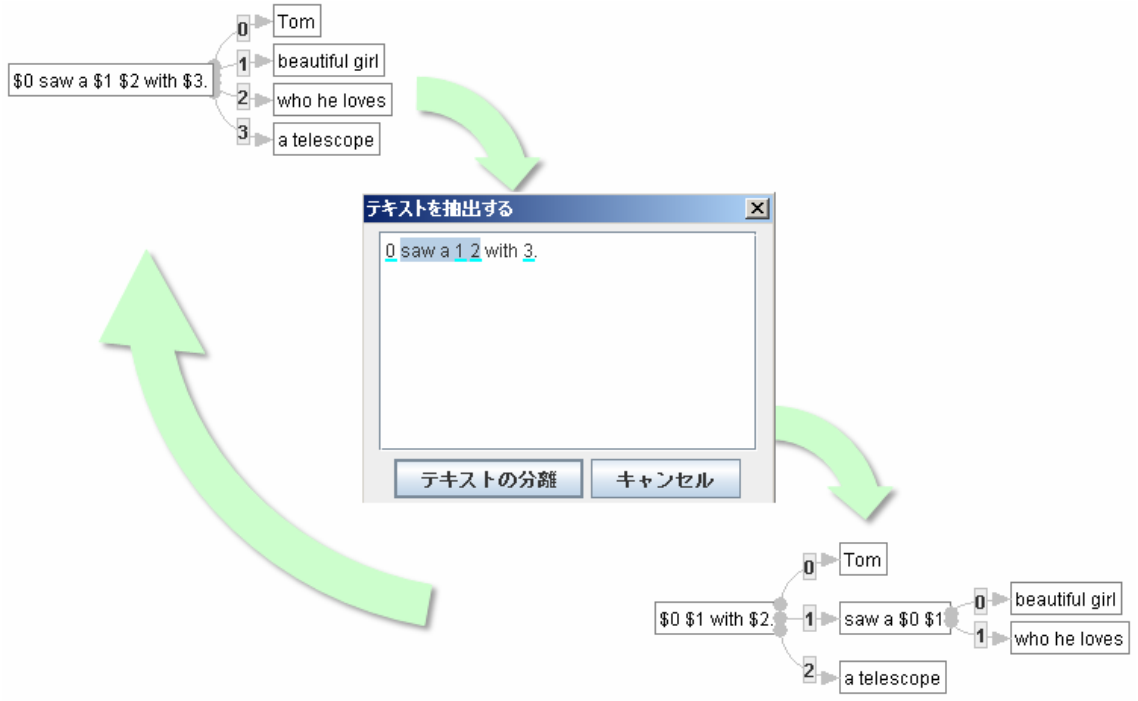
RDF のコンテンツを作成・編集するためのインフラとしてセマンティックエディタ(橋田, 2006; Hasida, 2007; 橋田・和泉, 2007)を用いる。セマンティックエディタで RDF のグラフとして文章の構造(主に談話構造)を明示した様子を下に示す。



3　言語的構造化のための諸機能の開発

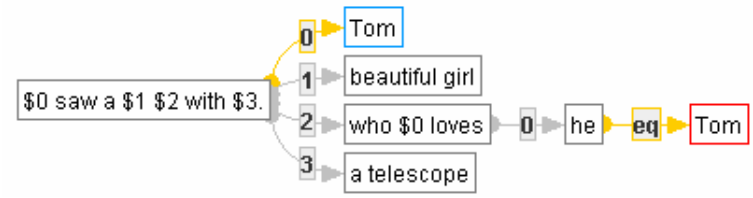
統語構造を作成・編集するためのセマンティックエディタのプラグイン機能のひとつを下図に示す。この他に、セマンティックエディタが扱う RDF のコンテンツと外部の XML データとの間で相互変換をする機能も実装した。

† hasida.k@aist.go.jp



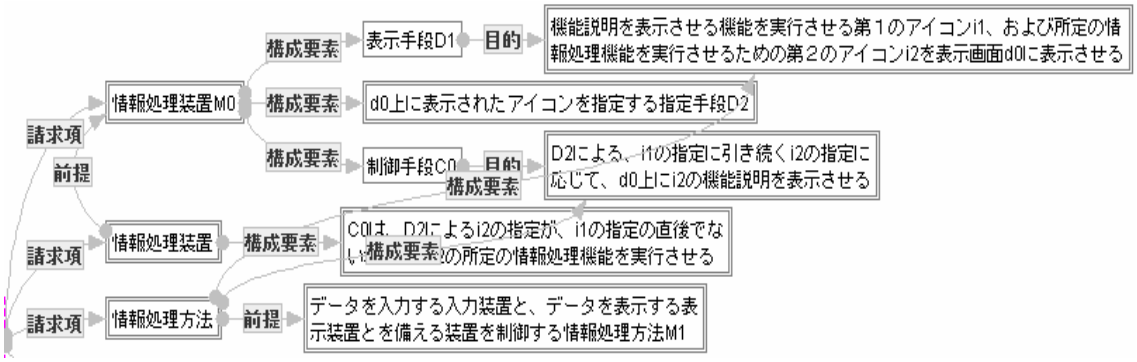
4　課題

下図に共参照の構造を明示した RDF グラフを示す。2 つの`Tom'ノードは同一のものである。このようにして任意のグラフを木構造として表現できる。同様にして任意の言語的構造を表現できるが、それは必ずしもわかりやすく扱いやすい構造とは言えない。構成素構造を表わす方法に関しても同様である。



5　展望

言語学や自然言語処理(特に後者)の研究は、言語コンテンツの分析には大いに役立っているが、創造にはあまり役立っていないように思われる。創造への支援も含めて研究成果が社会で活用され、知識の循環を活性化することが研究分野の発展には必須と考えられる。そのような観点から、本研究では一般の利用者がコンテンツ作成のために使える技術の創出も想定している。たとえばある特許の 3 つの請求項と内部構造と相互関係を明示したものを下図に示す。



第１の請求項を通常のテキストとして自動生成すると下記のようなものが得られる。

> アイコンの機能説明を表示させる機能を実行させる第１のアイコン、および所定の情報処理機能を実行させるための第２のアイコンを表示画面に表示させる表示手段と、前記表示手段の表示画面上に表示されたアイコンを指定する指定手段と、
> 前記指定手段による、第１のアイコンの指定に引き続く第２のアイコンの指定に応じて、前記表示手段の表示画面上に前記第２のアイコンの機能説明を表示させる制御手段とを有することを特徴とする情報処理装置。

このような種類の構造化と言語的なアノテーションとの相互連携を図ることを含めて今後の研究を進めて行きたい。

文献

橋田 浩一 (2006) オントロジーと制約に基づくセマンティックプラットフォーム. 人工知能学会誌, 21(6).

Hasida, K. (2007) Semantic Authoring and Semantic Computing. In Sakurai, A., Hasida, K. and Nitta, K. (*eds.*) *New Frontiers in Artificial Intelligence: Joint Proceeding of the 17th and 18th Annual Conferences of the Japanese Society for Artificial Intelligence*, 137-149, Springer.

橋田 浩一・和泉 憲明 (2007) オントロジーに基づく知識の構造化と活用. 情報処理, 48(8), 843-848.

ジャンル別 UniDic 作成の試み

小木曽 智信（電子化辞書班分担者：国立国語研究所研究開発部門）†
伝 康晴（電子化辞書班班長：千葉大学文学部）
渡部 涼子（電子化辞書班協力者：国立国語研究所研究開発部門）

An Attempt to Develop Genre-Fitted UniDics

Toshinobu Ogiso (Dept. Lang. Res., National Institute for Japanese Language)
Yasuharu Den (Faculty of Letters, Chiba University)
Ryoko Watanabe (Dept. Lang. Res., National Institute for Japanese Language)

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）は一般書籍・新聞・白書・Web データなどの多様なジャンルのテキストを含むコーパスであり、その解析を担う形態素解析辞書 UniDic にはこれら多くのジャンルに対応することが求められる。すでに UniDic には多様なジャンルのコーパスから見出し語を追加しているうえ、学習には複数ジャンルのコーパスを用いており、従来の形態素解析辞書にくらべ、多様なジャンルに対応した辞書として完成しつつあるといえる。しかし、BCCWJ に付与される形態論情報の精度を少しでも向上させるためには、このような汎用の辞書だけではなく、特定のジャンルに特化してより高精度な解析を行うことができる辞書を作成することも求められる。

本発表では、コーパスからの学習に際して特定のジャンルに適した素性のセットを用いるとともに、多ジャンルのコーパスで学習した基本辞書にジャンル別のコーパスで学習した辞書のコストを混合する方式を用いることにより、特定のジャンルに適合した形態素解析辞書を作成することを試みる。

２．利用したデータ

学習・評価に用いるジャンル別のコーパスとして、公開版 UniDic の学習コーパスのなかから表 1 に示すデータを用いた[1]。いずれのジャンルでも、評価用として 2 万語、混合率学習のためのコーパスとして 8 万語を用い、残りを基本辞書の学習コーパスとしている。Web データ（Yahoo! 知恵袋）は利用可能な人手修正済みデータが現在 10 万語分しかないため、基本辞書の学習コーパスからは外している。書籍のデータは文学作品に限らず各種の一般書籍を含んでいる。話し言葉データは CSJ のコアデータの一部を最新の短単位規定に合わせて修正したものである。

なお、語彙表には最新版の UniDic（書字形約 20 万語、活用形展開後約 53 万語）のデータを用いている。また、解析器には MeCab 0.97 を用いた。

† togiso@kokken.go.jp

[1] 公開版の UniDic では、これに加えて RWCP コーパス約 90 万語、CSJ 約 30 万語、新聞コアデータ 16 万語ほかを利用している。本発表では、ジャンルごとのコーパス量を均一にするために CSJ・新聞コアデータは約 20 万語分を文単位でサンプリングして用いている。

表 1　利用したコーパス

コーパス・ジャンル		学習用コーパス		評価用コーパス	合計
		基本辞書学習用コーパス	ヘルドアウト（混合率学習）		
CSJ	話し言葉	約 10 万語	約 8 万語	約 2 万語	約 20 万語
BCCWJ コアデータ	Web	－	約 8 万語	約 2 万語	約 10 万語
	白書	約 10 万語	約 8 万語	約 2 万語	約 20 万語
	書籍	約 10 万語	約 8 万語	約 2 万語	約 20 万語
	新聞	約 10 万語	約 8 万語	約 2 万語	約 20 万語

３．ジャンル別の素性による学習

公開版 UniDic の精度を向上させるために、さまざまな素性の組み合わせで学習した辞書を作成してジャンルごとの解析精度を調査している。その過程で、全体としては高精度でないため公開版辞書には適当ではないものの、特定のジャンルについては良い結果をもたらす素性の組み合わせがあることが分かっている。このように経験的に得られた情報を元に、各ジャンルに適した学習素性のセットを選定した。それぞれのコーパスに適した素性を用い、表 1 の学習コーパス全体を学習に利用して評価した結果を表 2・表 3 に示した。表の「汎用」素性は、公開版の UniDic で用いている学習素性を用いたものである[2]。以下の実験では、全てこのジャンル別の素性を用いる。

表 2　ジャンル別の素性を用いた解析精度（Level 3・語彙素認定）

素性＼コーパス	話し言葉	Web	白書	書籍	新聞
汎用	0.978418	0.979711	0.993486	0.984031	0.983033
話し言葉用	**0.979762**	0.979208	0.993505	0.984635	0.982832
Web 用	0.978992	**0.97973**0	0.993159	0.983689	0.983280
白書用	0.979232	0.978931	**0.993793**	0.984119	0.983500
書籍用	0.979234	0.979003	0.993312	**0.984831**	0.983612
新聞用	0.978971	0.979113	0.993428	0.984182	**0.984036**

表 3　ジャンル別の素性を用いた解析精度（Level 4・発音形認定）

素性＼コーパス	話し言葉	Web	白書	書籍	新聞
汎用	0.973534	0.975649	0.991141	0.980244	0.977645
話し言葉用	**0.974879**	0.975145	0.991160	0.980847	0.977488
Web 用	0.974109	**0.975667**	0.990815	0.979902	0.977892
白書用	0.974304	0.974868	**0.991449**	0.980289	0.978022
書籍用	0.974350	0.974941	0.990968	**0.981001**	0.978224
新聞用	0.974043	0.975004	0.991083	0.980352	**0.978559**

[2]表の Level は、1 が単位境界の認定、2 が品詞の認定、3 が語彙素の認定、4 が発音形の認定が正しく行えることを示す。語彙素の認定とは、金（キン）と金（カネ）の区別のように、境界・品詞に加え同一の語としての認定が正しく行えることを意味する。発音形の認定とは、日本（ニホン）と日本（ニッポン）の区別のように、語彙素認定に加えて語形・発音形の選択までが正しく行えることを意味する。数値は F 値。以下の精度評価でも同様。

４．混合方式によるジャンル別辞書の作成

MeCab を用いてコーパスからの学習を行う場合、図 １に示すように生起コスト(unigram コスト)、連接コスト（bigram コスト）を含むデータが配布用辞書（ソース辞書）として出力される。この辞書に手を加えることでジャンルに特化した解析用辞書を作成する。

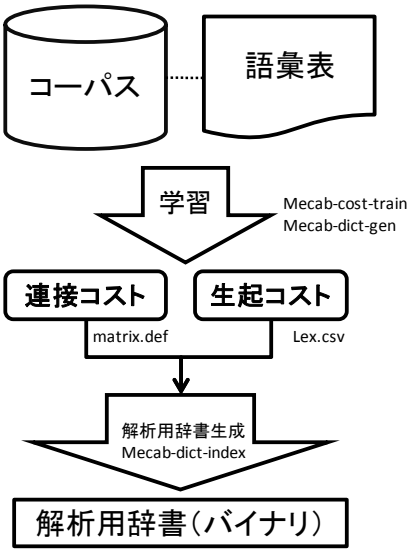


図 １　通常の MeCab 辞書作成

具体的には、多ジャンルのコーパスで学習した基本辞書の生起コスト・連接コストに、特定ジャンルのコーパスで別途学習した辞書の生起コスト・連接コストを混合する方法[3]をとる（混合方式)。ジャンル別辞書を混合する割合（混合率）は後述するヘルドアウト法による学習結果を用いる。図 ２にこの方法による解析辞書作成の流れを示した。

このようにして特定ジャンルのコーパスの性質を適切な割合で反映させることにより、ジャンルに適合した辞書を作成し、基本辞書の学習コーパスに特定ジャンルのコーパスを加えて通常の学習を行う方式（拡大方式）との精度比較を行う。

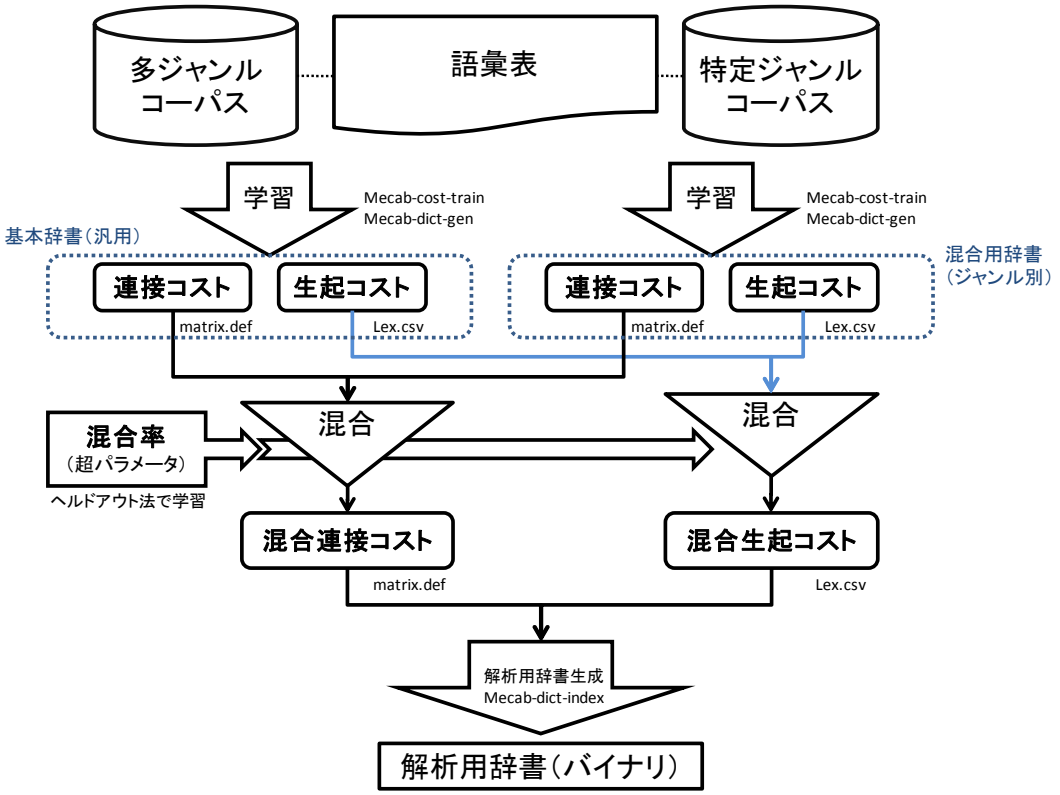


図 ２　混合方式によるジャンル別辞書作成

[3] 二つの辞書の生起コストを混合する方法は近代文語文を対象とした形態素解析辞書で試みたことがある（小木曽ほか 2009)。

５．混合率の学習

混合率の学習は図 3 に示す方法で行った。基本辞書の学習用コーパスと語彙表はそのままとし、ヘルドアウトコーパス 8 万語を学習用 7 万語、評価用 1 万語に分けて、ジャンル別辞書の学習と評価に用いた。ジャンル別の学習用コーパスと評価用コーパスは、1 万語ごとに入れ替えて交差検証を行った。混合率は 0 から 0.1 ごとに 1.0 までの 11 通りを、連接コスト・生起コストのそれぞれについて組み合わせて評価した。混合率は、たとえば 0.1 であればジャンル別辞書のコストに 0.1、基本辞書のコストに 0.9 を乗じたものを足しあわせることを意味する。評価の基準は Level 4（発音形認定）の精度とした。

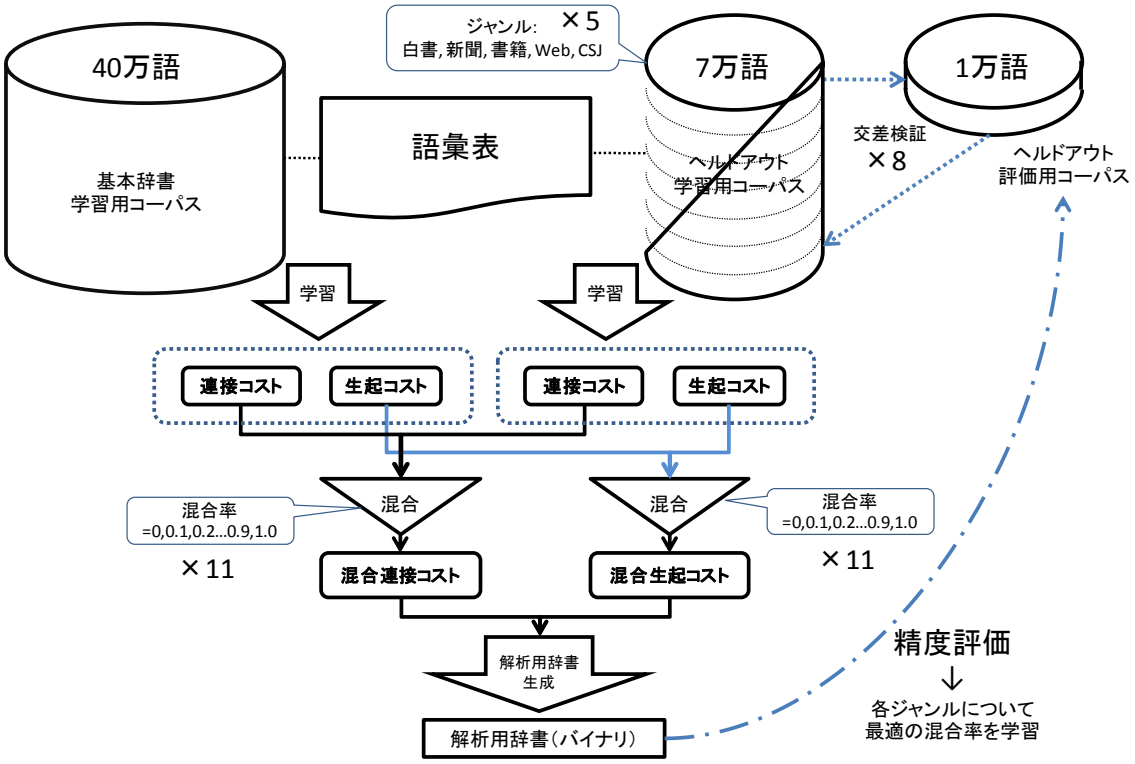


図 3　混合率の学習

この方法により学習した各ジャンルにおける最適な混合率を表 4 に示す。一般的な文章とは性格が異なる話し言葉・Web において混合率が高くなっている。特に基本辞書に当該ジャンルのデータが含まれない Web で高い。

表 4　混合率の学習結果

ジャンル	生起コスト混合率	連接コスト混合率
話し言葉	0.5	0.4
Web	0.6	0.5
白書	0.5	0.3
書籍	0.5	0.3
新聞	0.3	0.3

６．混合方式によるジャンル別辞書の精度評価

全てのジャンルについて学習した混合率を用いてジャンルに最適化した混合方式の辞書を作成した。混合方式による辞書は、対象としたジャンルに良く適合し、結果として他のジャンルでは精度が落ちている。表 5 に各ジャンル別辞書による解析精度（Level 4）を基本辞書と比較して示す。基本辞書と比較して精度が向上したものを太字で示した。ジャンル別の辞書は当該ジャンルでのみ精度が向上している。ただし、基本辞書の学習用コーパスに当該ジャンルのデータを含まない Web データでは他ジャンル用の辞書でも精度が向上する場合がある。

表 5　ジャンル別辞書の各ジャンル別精度（Level 4・発音形認定）

コーパス / 辞書	話し言葉	Web	白書	書籍	新聞
基本辞書	0.970847	0.968881	0.990870	0.979599	0.976327
話し言葉用	**0.975663**	0.961871	0.979574	0.970884	0.965253
Web 用	0.952590	**0.976022**	0.980173	0.975391	0.969444
白書用	0.954452	0.964294	**0.991985**	0.973003	0.970081
書籍用	0.960861	**0.972301**	0.986618	**0.980781**	0.972441
新聞用	0.962697	**0.970950**	0.989677	0.977937	**0.979016**

基本辞書・拡大方式・混合方式について各 Level 別の精度を表 6 にまとめた（拡大方式・混合方式については当該ジャンルの解析精度のみ)。また、Level 4（発音形認定）の解析精度を図 4 にまとめた。混合率学習の基準とした Level 4 では全てのコーパスにおいて、今回提案した混合方式がもっとも高精度となっている。単純にジャンル別コーパスを追加して学習する拡大方式に勝っており、コスト混合による方法がジャンル別辞書の作成に有効であることが示された。

表 6　各方式によるジャンル別辞書の精度評価

Level	辞書タイプ	話し言葉	Web	白書	書籍	新聞
Level 1 境界認定	基本辞書	0.994740	0.994269	0.998442	0.996039	0.994276
	拡大方式	**0.996215**	**0.995273**	0.998500	**0.996254**	0.994299
	混合方式	0.995466	0.995249	**0.998750**	0.995975	**0.994875**
Level 2 品詞認定	基本辞書	0.980527	0.977830	0.994752	0.986787	0.985279
	拡大方式	**0.983852**	0.981714	0.995041	**0.987130**	0.985794
	混合方式	0.982924	**0.983238**	**0.995674**	0.987108	**0.987659**
Level 3 語彙素認定	基本辞書	0.976083	0.973036	0.993215	0.983301	0.981448
	拡大方式	0.979759	0.977834	0.993427	**0.983902**	0.982276
	混合方式	**0.979799**	**0.979310**	**0.994176**	0.983880	**0.984139**
Level 4 発音形認定	基本辞書	0.970847	0.968881	0.990870	0.979599	0.976327
	拡大方式	0.974656	0.974227	0.991082	0.980243	0.977021
	混合方式	**0.975663**	**0.976022**	**0.991985**	**0.980781**	**0.979016**

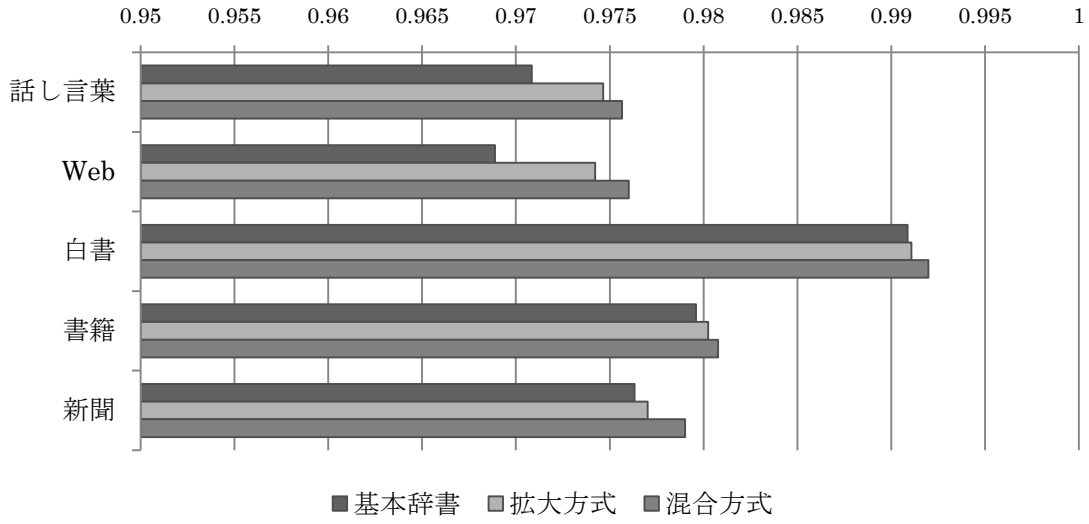


図 4　各方式によるジャンル別辞書の精度評価（Level 4・発音形認定）

7．おわりに

コスト混合方式を用いることで特定ジャンルに特化した形態素解析辞書を作成することができることを示したが、この方式は、全ての学習用コーパスが入手できない場合であっても応用可能である。すなわち、ユーザが独自に作成した特定ジャンルコーパスで学習して作った辞書を、一般公開されている汎用の辞書に混合することにより、ユーザ独自のジャンル別辞書を作成することができる。本発表で、基本辞書の学習に同種のデータを用いていない Web データにおいて良い結果が得られていることから、新しいジャンルのデータにおいても有効であると考えられる。

現在 BCCWJ の解析には汎用の UniDic を用いているが、今後は、本発表で提案した方式で作成したジャンル別 UniDic を用いることにより、BCCWJ 全体の解析精度を高めることが可能になると思われる。さらに、ジャンル別 UniDic の応用として、BCCWJ に付与されているタグを利用することで、地の文と会話文とで解析辞書を切り替えて全体の解析精度をさらに上げるような方法も考えられる。

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 号 pp.101-122.

小木曽智信・伝康晴・渡部涼子・近藤明日子（2009）「現代語コーパスの利用による近代語形態素解析の精度向上」言語処理学会第 15 回年次大会発表論文集

関連 URL

UniDic ダウンロードサイト：http://download.unidic.org /

MeCab : Yet Another Part-of-Speech and Morphological Analyzer：http://mecab.sourceforge.net/

文法の中核と周辺
―コーパスが観察可能にする文法の一面―

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 Core and Periphery of Grammar: A Corpus Linguistic View

Tadaharu Tanomura (Osaka University)

１．はじめに

文法――その用語をここでは狭く統語論（構文論）の意味に用いる――は一般に“当該言語の文法的な表現の集合を規定する規則の体系”と理解される。「文法的な表現」は「適格な表現」とも言われる。

こうした文法の概念は暗黙裡に次の２つのことを前提としている。

(1) a. 文法的な表現をそうでないものから区別できること
　　b. 文法の現象は規則の形に分析できること

しかし、実際のところ、これらの前提のいずれに関しても問題がある。まず、第１の前提に関して言えば、文法性判断の不確実性は周知の事実である。明らかに適格な表現と、明らかに不適格な表現に関しては安定した文法性判断が得られるにしても、そのことからすべての表現に関しても同様だと考えるのは根拠のない予想・期待に過ぎない。

第２の前提の妥当性が問われることは従来あまりなかったかも知れない。しかし、通念的な文法観を離れて言語の現実に向き合えば、形態論の現象に規則的な面とそうでない面があるのに似て、文法の現象にも規則化の可能な面とそうでない面があることに気付く。

文法的か否かという２値的な区別に基づく規則の形での分析・記述が可能なのは、文法の全体ではなく、その言わば中核的な部分だけだと言うべきであろう。その周辺（あるいは隙間）に位置する領域において観察される現象はしばしば曖昧で複雑な様相を呈し、言語使用の実態を単に文法上の慣習として整理・記述する以上のことを期しがたい。

この小論では、文法のそうした周辺的な領域の一端を大規模なコーパスから得られる情報に基づいて観察し考察する。取り上げる事例――規則化が困難であるばかりか、中核的な文法の規則を逸脱さえする――は一見単純な問題のようでありながら、内省による把握の遠く及ばない複雑な様相を示す。コーパスから得られる情報の統一的な解釈すら容易ではなく、以下の論述は当該の文法現象のより適切で包括的な記述のための予備考察の次元にとどまらざるを得ない。

２．引用の「と」の余剰付加

2.1　規則逸脱的な「と」付加――「そうとも言う」の無理

筆者が聞いて不自然に感じてしまう「そうとも言う」という言い回しがある。「そう言う」の「そう」に「も」を加えれば「そうも言う」になるはずであり、「と」は本来余剰的な存在である。

文法の議論の一般的な様式に則って、筆者の感覚での“自明”の結論を確認すれば次の

通りである。「×(と)」、「 (×と)」という表記は、それぞれ、「と」が必須であること、「と」が不可能であることを示す。

(2) a. 天気予報は朝夕冷え込む×(と)言った。　→　～冷え込む×(と)も言った。
　　b. 天気予報はそう(×と)言った。　→　～そう(×と)も言った。
(3) a. カボチャのことをナンキン×(と)言う。　→　～をナンキン×(と)も言う。
　　b. カボチャのことをそう(×と)言う。　→　～をそう(×と)も言う。

矢印の左の表現に「も」を加えた結果が右の表現であるが、「と」の有無による文法性の差は「も」を付加しても不変の“はず”である。つまり、(2a)を例に取れば、「～冷え込むと言った」において「と」は必須であり、「～冷え込むと」に「も」を加えてもそのことは変わらない。(2b)では、「～そう言った」に「と」を加えることはできず、「そう」に「も」を加えれば「～そうも言った」になる“はず”である。[1)]

そのような判断からすれば、「そうとも言う」における引用の「と」の付加は、文法の規則を無視した逸脱的な現象であることになる。ところが現実には、(2b)、(3b)に示した推論に従えばあり得ないはずの「そうとも言う」という言い回しが行われている。

実のところ、“「そう」への「と」の付加”という見方は事実の一面的な理解でしかないのであるが、便宜上差し当たりこの２節ではその見方に基づいて考察を進める。

2.2　「そう＋と」の広がり――「そうとも言う」の有理

「そうとも言う」という言い回しを筆者は不自然に感じるわけであるが[2)]、視野を広げて見れば、「そう」＋引用の「と」という語連続は「そうとも言う」という言い回しにおいてのみ見られるものではない。「そう」に係助詞・副助詞の類が組み合わせられるとき、「そう＋と」という語連続の出現は珍しくないのである。

拙作のWebコーパス――現在の分量約150ギガバイトのうちの約100ギガバイト分をことでは使用する――に基づく調査によれば[3)]、「そう」との共起例が相対的に多い係助詞・副助詞のうちで「と」の出現率が全般に最も高いのは、「しか」である。「そう(＋と)＋しか＋述語」における種々の述語ごとの「と」の出現率は次の表の通りで、多数の述語において90%以上の高率に上る。

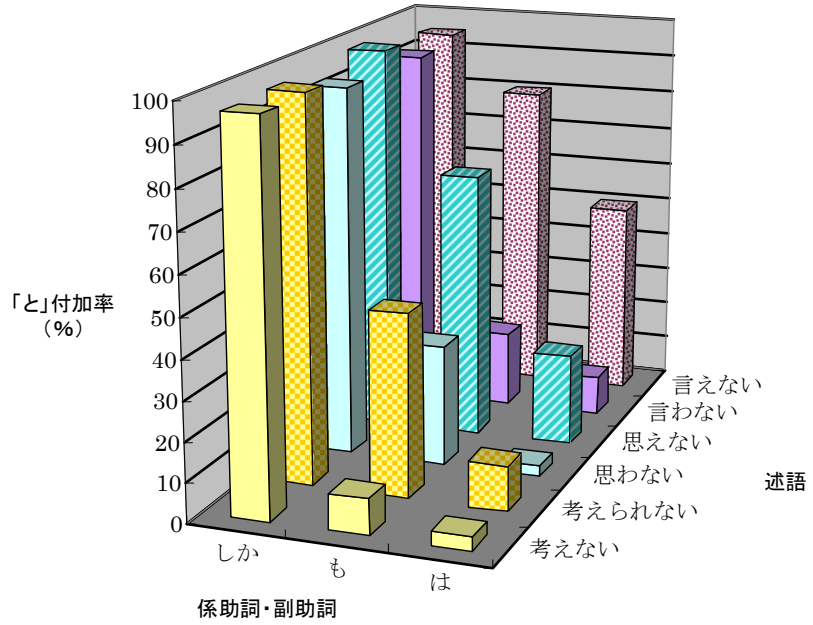
（表１）「そう(＋と)＋しか＋述語」における「と」の出現率

	そうしか	そうとしか	「と」出現率(%)
考えようがない	0	48	100
思われない	0	33	100
思えない	32	2421	98.7
考えられない	39	1275	97.0
考えない	3	89	96.7
言いようがない	28	769	96.5
言えない	32	613	95.0
聞こえない	4	73	94.8
思わない	2	30	93.8
言わない	4	53	93.0

見えない	27	352	92.9
取れない	5	48	90.6
読めない	13	77	85.6
書けない	7	15	68.2
生きられない	13	15	53.6
出来ない	101	17	14.4

（左端の欄は述語の代表形。その丁寧体や過去形なども統計に含む。）

「そう」との共起例が多く「と」の出現率が全般に高い係助詞・副助詞は、「しか」に次いで「も」、そして「は」である。「そう」と「しか」「も」「は」および６種類の述語の組合せにおける「と」の出現率の関係を（図１）に示す。



（図１）「そう（＋と）＋｛しか/も/は｝＋述語」における「と」の出現率

（図１）に見るように、「と」の出現率は場合によってまちまちで、0～100%の全範囲にわたっている。ただし、まちまちとは言っても、（図１）の限りでは次の一般化が成り立つ。

(4) (i) どの述語においても、助詞「しか」「も」「は」の順に「と」の出現率が高い。
(ii) どの動詞、どの助詞においても、動詞の単純形の否定より可能（自発）形の否定のほうが「と」の出現率が高い。

場合によって「と」の出現率に大きな開きがあるにせよ、（図１）に示す状況全体を背景として見れば、「そうとも言う」も十分に“あり得る”表現のように思えてくる。

しかしながら、筆者の感覚においてやはり、「そうとも言う」と「そうとしか言えない」のあいだには歴然とした自然度の開きがある。すなわち、「そうとしか言えない」のような言い回しは自然で、実際そうとしか言えない——つまり、「そうしか言えない」とは言いにくい——ように感じられる。それに対し、「そうとも言う」については、そのような言い回しが現実に用いられることがあるという事実をただ受動的に認識するに過ぎない。

とすれば、「そうとも言う」に感じる不自然さはそれが余剰の「と」を含むという事実だ

けでは説明できず、「そうとも言う」という表現に固有の事情があることになる。

2.3 「と」付加の普及の順序・程度――「そうとも言う」の無理再び

「しか」はもっぱら否定の述語との組合せにおいて用いられるが、「しか」以外の係助詞・副助詞にはそのような制約がない。

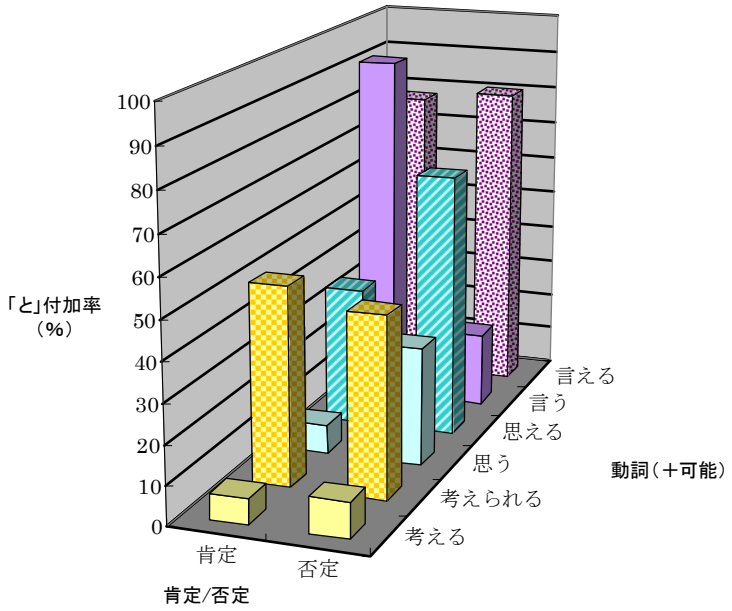
「も」の場合の、肯定・否定のさまざまな述語約 20 種類との組合せにおける「と」の出現率は次の通りである。

(表２)「そう(＋と)＋も＋述語」における「と」の出現率

	そうも	そうとも	「と」出現率(%)
知らない	0	1177	100
言う	153	1653	91.5
言いきれない	171	1340	88.7
思われない	3	13	81.3
言えない	362	1456	80.1
言える	347	1193	77.5
思えない	134	285	68.0
考えられる	106	109	50.1
見えない	39	38	49.4
考えられない	14	12	46.2
思える	103	58	36.0
聞こえる	13	7	35.0
思わない	159	70	30.6
言わない	46	11	19.3
考えない	20	2	9.1
思う	1016	82	7.5
考える	171	12	6.6

(左端の欄は述語の代表形。その丁寧体や過去形なども統計に含む。)

「言う」「思う」「考える」の各動詞については、肯定と否定、動詞の単純形と可能（自発）形という２組の対立に基づく述語の用例がすべて揃っている。そこで、「そう」、「も」、3×2×2＝12種類の述語の組合せにおける「と」出現率の関係を図示すれば（図２）のようになる。この図において、例えば最前列の２本の角柱は「考える」という動詞の肯定と否定の形――すなわち、「考える」と「考えない」――に該当し、前から２列目の２本の角柱は「考えられる」と「考えられない」に該当する。



（図２）「そう（＋と）＋も＋述語」における「と」の出現率

（図２）から次のことが分かる。

(5) (i)（図１）で、否定述語の場合に動詞の単純形より可能（自発）形のほうが「と」の出現率が高いことを見たが、肯定述語の場合にも同じことが概ねあてはまる。ただし、「言う」に関しては関係が逆転し、「言う」のほうが「言える」よりも出現率が高い。

単純　＜　可能　（全体的傾向）
（否定）考えない＜考えられない
思わない＜思えない
言わない＜言えない
（肯定）考える＜考えられる
思う＜思える
言う＞言える（逆転）

(ii) 肯定述語と否定述語とでは、否定述語のほうが肯定述語より「と」の出現率が概ね高い。ただし、６対の述語のうち「考えられる／考えられない」と「言える／言えない」の２対の場合は関係が逆転している。特に、後者は逆転の程度が大きく、例外として突出している。

肯定　＜　否定　（全体的傾向）
（単純）考える＜考えない
思う＜思わない
言う＞言わない（逆転）
（可能）考えられる＞考えられない（逆転するも僅差）
思える＜思えない
言える＜言えない

このように、「そう(と)も言う」――その丁寧体「そう(と)も言います」を含む――における引用の「と」の出現率は、「と」付加の普及拡散の順序と程度に関する傾向を外れた高率となっている。「そう＋と」という語連続を含む言い回しの中でも「そうとも言う」を特に不自然に感じる感覚はそのことと関係しているものと見られる。

もっとも、「そう(と)も言う」における「と」の付加率が関連の言い回し全体に関わる傾向に反して高くなっていることの理由は明らかではない。筆者は、「そうとも言う」は全体的な傾向に反するがゆえに新鮮な言い回しとして流行的に広まったに過ぎないのではないかと憶測する（アニメ「クレヨンしんちゃん」のテレビ放送の影響の可能性も考えられる)。それがいずれ下火になるものか、標準的な言い回しとして日本語に定着するものか将来を確実に予見することはできないが、筆者は前者の予想に傾く。

2.4　通時的な状況について

国会会議録のデータや古い文学作品などを用いてざっと調査してみたところによれば、「と」の余剰付加は実際時間の経過とともに全体的に増加しているような印象がある。現代ならば「そうとしか言えない」と言いそうなところで「と」を加えず「そうしか言えない」としている用例もそれなりの頻度で見つかる。

しかし、目下の問題の考察において通時的な事実に手がかりを求める可能性については残念ながら多くを期待できない。（表１）や（表２）に示した統計は 100 ギガバイト――小説単行本にして約 25 万冊分の量――という規模の巨大な言語資料を利用して初めて得られたものであり、過去の日本語に関してはそれに匹敵する規模の資料がない以上、「と」の付加に関わる通時的な状況を定量的に確認することは望みがたいからである。

3　関連した要因――包括的記述のための要検討課題

3.1　引用の「と」の直前におけるコピュラの潜在

以上の考察においては、「そうとも」「そうとは」といった言い回しを、本来「そうも」「そうは」などとなるべきところに余剰の「と」が付加されたものとして――その意味で、文法の規則を逸脱したものとして――扱った。

しかし、実は、「そう＋と」という語連続の現れの中には、正規の文法規則の適用によって生成されたと理解すべきものもある。それは引用の「と」の直前の位置でのコピュラの潜在の可能性に関わる。[4)] 例えば、(6a)の文に対応して、「と」の前のコピュラ「だ」が潜在している(6b)の文が可能である。

(6) a. あの人が男だとは知らなかった。
　　b. あの人が男とは知らなかった。

これと並行的に、(7a)に対応するものとして、「そうとは」を含む(7b)が可能である。

(7) a. そうだとは知らなかった。
　　b. そうとは知らなかった。

この(7b)の「そうとは」は「と」の余剰付加によって生じた逸脱的な表現ではなく、「と」の前にコピュラが潜在している――そこにあったコピュラが省略された、と言ってもよい――正統な表現である。

実際、（表２）の最上段に位置する「知らない」の行は、「そうとも知らないで」「そうとも知らず」などの言い回しの用例を意味するが、それらは「そうだとも知らないで」「そうだとも知らず」からコピュラが省かれたものと自然に解釈することができる。

そして、（表１）で見た「そうとしか思えない」「そうとしか考えようがない」などの言い回しもそのような目で見直せば、同様に「そうだとしか思えない」「そうだとしか考えよ

うがない」からコピュラが省かれたものとする解釈も可能であることが分かる。

しかしながら、注意すべき重要な点は、そのような解釈が不可能な場合においても「そうとしか」は使われるということである。例えば、2.2に書いた、

(8) 「そうとしか言えない」のような言い回しは自然で、実際そうとしか言えない（中略）ように感じられる。

という一節の加点部は「そうだとしか言えない」からコピュラが省かれたものではない。本来「そう（＝そのように）しか言えない」となるべきところに、現に余剰の「と」が付加されているのである。Webコーパス中に見出される分かりやすい同種の用例としては、次のようなものが挙げられる。

(9) a. そうしたいのではなくそうとしか出来ないのだ。
b. 悲しい現実の中ではそうとしかできなかった。
c. 間違っていると知りつつ、そうとしか生きられない。
d. そうやってここまで生きてきたし、これからもそうとしか生きられない。

可能性としては、文法規則に基づくコピュラ潜在の結果としての正統な「そう＋と」という語連続が、そのような文法操作の考えられない場合にまで類推によって影響を及ぼし、結果的に規則逸脱的な「と」の余剰付加を誘発したということが考えられる。しかし、実際のところは、それだけの単純な事情ではあり得ない。そのことには3.3であらためて触れる。

いずれにせよ、コピュラ潜在による正統な語連続としての「そう＋と」と、そうでない余剰の「と」を含む「そう＋と」は区別しがたい――すなわち、「そうとも」「そうとしか」などの個々の用例を見て、それが「と」の余剰付加によるものかコピュラの潜在によるものかを判定することは必ずしもできない――状態にあり、関連の表現をめぐる状況の分析は容易ではない。

3.2 「そう」の示す内容の種類

「そうとも言う」という言い回し、引いては、「そう＋と」という語連続を含む言い回し全般に関して考える必要のある意味上の問題がある。

「そう」が動詞「言う」にかかる言い回しにも、「そう」の表す内容のうえで性質の異なるいくつかの種類のものがある。もっとも分かりやすいのは、(2b)と(3b)に示した例のあいだの区別である。(2b)の「そう」は「朝夕冷え込むと」の代用として命題的な内容を示すのに用いられているのに対し、(3b)の「そう」は「ナンキンと」の代用としてカボチャという事物の名称、別名を示すのに用いられている。

命題的な内容を示す「そう」の用法について言えば、「天気予報は朝夕冷え込むと言った？」という問いに対する肯定的な応答としては「そう言った」と答えるのが自然であるが、「だ」を加えた「そうだと言った」という応答も柔軟に受け止めれば不可能ではないかも知れない。しかし、「カボチャをナンキンと言う？」という問いに対して「そう言う」と答える代わりに、「そうだと言う」と答えることを認めるにはいっそうの寛容が必要であろう。

10 年前に聞いて印象に残った「そうとも言う」も、今回の調査で大量に得られた「そうとも言う」の用例の圧倒的大多数も、「そう」が名称を示すタイプか、もしくは、次の例に

見るような(2)とも(3)とも少々異なるタイプである。

(10) A：仕事まだ終わらなくて。
B：どうせ１日中遊んでたんだろ。
A：そうとも言う。

(10)の「そうとも言う」の「そう」は「１日中遊んでいた」という命題を受けているという点では一見(2)に近いが、これはむしろ「そう」が名称を示す(3)に類すると考えるべきものである。その理由は次の通りである。

(2)の「そう(と)も言った」は、あることが言われたという話を受けて、それとは違う別のことも言われたということを意味する。天気予報が、例えば「あすは雨が降る」という予報に加えて、「朝夕冷え込む」ということも言ったということである。それに対し、(3)の「そう(と)も言う」は、カボチャが「ナンキン」という別名でも言及され得るということを意味する。前者のタイプでは、ある特定の時点において「朝夕冷え込む」という発話がなされた（または、なされる）ことを表すのに対し、後者では、カボチャが「ナンキン」とも呼ばれ得るという超時的な関係を表す。

(2)と(3)のそうした違いを念頭に置いて考えれば、(10)が(3)に近い性格のものであることが分かる。(10)の「そうとも言う」の「そう」は「１日中遊んでいた」を指すわけであるが、当該の文は「仕事がまだ終わらない」という先行発話に加えて何か別のことを言おうとするものではない。Aが「仕事がまだ終わらない」と表現した事実を、Bの解釈に基づいて言い換えれば「１日中遊んでいた」になるということである。Aの「そうとも言う」はその言い換えが適切であることを認めるものであり、別名の関係を認定する(3)のケースに似て、言い換えの関係が成り立つことを表す。

「そう(と)も言う」の以上の用法に便宜上名前を与えるとすれば、(2)は“追加発言タイプ”、(3)と(10)は“別名・言い換え関係タイプ”とでも呼ぶことができよう。

（表２）に示すように、「そうとも言う」の用例には非過去形のものが圧倒的に多い。これは、「と」の付加された「そうとも言う」が超時的な“別名・言い換え関係タイプ”に集中しているという事実を反映するものである。これとは対照的に、正統な「そうも言う」の場合には、非過去・過去の用例が相半ばしている。

（表２）「そう(と)も言う」における非過去の比率

	非過去	過去	非過去の比率（%）
そうも言う	70	83	45.8
そうとも言う	1629	24	98.5

（丁寧体「～言います」も統計に含む。）

3.3 包括的な記述のために

「そう」への「と」の余剰付加の原因としては、3.1 で述べたように、コピュラの潜在の結果としての「そう＋と」という語連続への類推が１つの可能性として考えられる。また、「そう」の様態を表す用法との差別化といった動機の関与もあるかも知れない。

しかし、それらが現に「と」付加の原因であるとしても、原因のすべてではあり得ない。と言うのも、2.2 の冒頭でも触れたように、「そう」への「と」付加はそもそも「そう」に

係助詞・副助詞が組み合わせられるときに限ってよく生じる現象であり、係助詞・副助詞を伴わない「そうと言う」「そうと考えられる」などの用例は非常に稀であるからである。ということは、「そうしか」「そうも」「そうは」のような「そう」＋係助詞・副助詞の語連続を避けたいという何らかの感覚が人々の心中に生じたことになるが、それを具体的に特定することはむずかしい。

いずれにせよ、この小論で述べた考察の域を超えてより適切で包括的な分析・記述を得るためには、まだ多くの表現の使用状況を調査する必要がある。まず、ここでは取り上げなかったほかの副助詞が関わる場合の状況を詳しく分析する必要がある。その中には、「そうとばかりも言えない」（「そう＋と＋ばかり＋も」）のように、「そう」に副助詞と係助詞が組み合わせられる場合の「と」付加もある。また、「と」の余剰付加と述語の種類の相関をさらに広く観察・分析する余地もある。

さらに、「そう」と同じく「と」が付加されることのある「どう」や「こう」に関わる状況も併せて検討する必要がある。「そう」の場合と並行的に、「どう」単独で述語にかかる場合は「どう言う」「どう思う」のようになり、通常「×どうと言う」「×どうと思う」とは言わないが――「どうということはない」のような言い回しはある――、係助詞・副助詞との組合せにおいては余剰の「と」が付加された「どうとでも言える」や「どうとも思わない」のような言い方が可能である。ちなみに、変化を表す述語の場合には「どうにでもなる」「どうにもならない」のように「に」が付加されるという現象もある。「こう」についても「そう」の場合と同様の「と」の付加が見られるが、「そう」に比べて付加率は全体的にかなり低い。「ああ」への「と」の付加は稀で、「ああとも言えるし、こうとも言える」のような言い回しにおいて観察される程度である。なお、「と」の余剰付加に表面上一致する表現として、「こうと決めたら後には引かない」「どうすると聞いてもああとしか言わない」のようなものがあるが、これらは心内発話ないし音声発話としての「こう」「ああ」に引用の「と」が規則通りに加わったものと見るべきであろう。

3.4　混同されがちな３つの概念

最後に、文法の周辺的な部分を論じる際の一般的な問題として、混同、同一視されやすい次の３つの基準、尺度の相互関係について簡単に触れておく。

(11) a. 当該の事象に関する文法性判断が容易か困難か
　　b. 当該の事象の規則化が可能か不可能か
　　c. 当該の事象が確定的か確率的か

表面的には、これらの基準、尺度は互いに一致するかのような印象がある。すなわち、単純明瞭な事象の場合には、文法性判断が容易であり、規則化が可能であり、可能な表現は１通りに確定している。また逆に、例えば（図１）や（図２）で見たような複雑な事象の場合、文法性の判断が困難であり、規則化が困難であり、表現の使い分けは用例の相対頻度という観点から見ざるを得ない。

しかし、３つの基準、尺度は実のところ互いに等価ではない。

まず、(11a)と(11b)が一致しないことは形態論の領域に例を求めれば示しやすい。例えば、動詞「行く」に「た」や「て」が続くとき、カ行五段活用動詞としては例外的に促音を含む「行った」「行って」という不規則な形になる。しかし、「行った」が適格で、規則に従う「行いた」が不適格であるという判断はゆるがない。つまり、規則化はできないが、文

法性判断は容易である。また、「見れる」「食べれる」のようないわゆるラ抜き言葉を適格な表現と見るかどうかは議論の余地のあり得る問題であるが、ラ抜き言葉を形作る規則を書くことは容易である。つまり、この例においては、規則化は容易であるが、文法性判断は容易ではない。

次に、(11c)について言えば、これは要は、あることを言うのに１通りの言い方しかないか、複数通りの言い方があるかということに過ぎない。複数通りの言い方があれば、(11a)、(11b)に関わる事情の如何を問わず、そこには必然的に使用頻度の比率の問題が生じる。したがって、(11c)は当然(11a)とも(11b)とも一致しない。

このように、文法性判断の容易性、規則化の可能性、事象の確定性という３つの基準、尺度は互いに独立しており、それらの概念に言及する際にはその点に注意が必要である。

4　おわりに

「そうとも言う」という言い回しに感じる不自然さを関心の出発点とし、日本語文法の周辺的な領域の一端を観察・考察した。

「そうとも言う」を含む一群の関連表現に関わる文法上の事実をどのように分析・記述するかは文法の立場や目的によって異なってくるが、いずれにせよ考察の基礎となる言語事実の観察の手段として内省はほとんど役に立たず、（図１）や（図２）などに示したような複雑な状況を構成するそれぞれの表現の文法性を２値的な図式に従って内省で判定することは絶望的に困難である。コーパスの時代に至って初めて、内省による把握の及ばない複雑で微妙な文法の領域を精密に観察することが可能になったと言ってよいだろう。

注

１）方言の表現を考えれば、この“はず”の推論は必ずしも成り立たない。関西方言では引用の「と」は省略可能である（「朝夕冷える(と)ゆうた」「ナンキン(と)ゆう」）が、「も」が加わる場合は「と」が必須である（「朝夕冷える×(と)もゆうた」「ナンキン×(と)もゆう」）。

２）服部匡・杉本武両氏の御教示によれば「そうとも言う」という言い回しは古くから落語家や漫才師などによって使われており、また、両氏の感覚では特に不自然な表現でもないとのことである。とすれば、「そうとも言う」の受け止め方には同世代でもかなりの個人差があることになる。なお、「そうとも言う」が古くから落語家などによって使われているという両氏の認識は、それが広く使われる一般的な表現ではなかったという認識でもあるとは言えるかも知れない。

３）コピュラの潜在については拙論(2006)をご覧いただきたい。

４）Web コーパスとは、インターネット上の日本語文書を収集・集積した言語研究資料を言う。拙作の Web コーパスについては拙論(2009a) 、同(2009b)などでその概略を述べた。

文献

田野村忠温(2006)「コピュラ再考」藤田保幸・山崎誠編『複合辞研究の現在』（和泉書院）

田野村忠温(2009a)「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(2009b) 「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号掲載予定

デモ・ポスターセッション

3月15日（日） 13:00〜15:30

『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―
▶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』における電子化フォーマットとその応用
▶山口 昌也、間淵 洋子、西部 みちる、小林 正行、大島 一、高田 智和

書籍コーパス（流通実態サブコーパス）の「外字」
▶高田 智和、小林 正行、間淵 洋子、西部 みちる、大島 一、山口 昌也

著作権処理の進捗状況と著作権法改正の動きについて
▶森本 祥子、前川 喜久雄、小沼 悦、新井田 貴之、大石 有香、神野 博子、竹内 ゆかり、舞木 右

『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
▶小椋 秀樹、小木曽 智信、小磯 花絵、冨士池 優美、宮内 佐夜香、渡部 涼子、竹内 ゆかり、小川 志乃、小西 光、原 裕、中村 壮範

形態論情報データベースの構成
▶小木曽 智信、小椋 秀樹、小磯 花絵、冨士池 優美、宮内 佐夜香、渡部 涼子、竹内 ゆかり、小川 志乃、小西 光、原 裕、中村 壮範

拡張固有表現タグ付きコーパスの構築に向けて ―白書、書籍、Yahoo! 知恵袋コアデータ―
▶橋本 泰一

タグ付きコーパス管理ツール「茶器」の現状と今後
▶松本 裕治、浅原 正幸、岩立 将和、森田 敏生

汎用アノテーションツールSLATにおける階層構造をもつタグセットのためのインターフェース
▶松井 信太朗、野口 正樹、飯田 龍、徳永 健伸

BCCWJに見られるオノマトペの型と共起との関連
▶ホドシチェク・ボル、ベケシュ・アンドレイ、仁科 喜久子

短単位を対象とした連濁の処理について
▶山田 篤

Yahoo! 知恵袋にみる非規範的表現
▶杉本 武

大規模コーパスの語彙統計情報の利用を支援する ―語彙情報データベースを参照するAPIの構築と活用―
▶千葉 庄寿

コーパスを用いた公共性の高い文章における表記改善への視点
▶斎藤 達哉

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―
▶近藤 明日子

白書およびYahoo! 知恵袋を対象にした結合価の自動抽出 ―格助詞パターンに着目して―
▶荻野 孝野

異ジャンルの種用例を用いた半教師有りクラスタリングとその語義曖昧性解消に関する効果
▶杉山 一成、奥村 学

複数の語義を積極的に取り出す動詞のクラスタリング
▶高橋 秀幸、竹内 孔一

BCCWJを用いた新しい語義曖昧性解消タスク
▶奥村 学、白井 清昭

フレーム意味論と「日本語コーパス」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
▶小原 京子、斎藤 博昭

日本語リーダビリティー公式の構築と測定ツールの開発
▶柴崎 秀子

グラフクラスタリングを用いた語義別用例分類
▶佐々木 稔、新納 浩幸

ジャンル別に見るガ格を取る名詞と共起する用言の差異
▶野口 慎一朗、仁科 喜久子

規則処理のアクセント属性を導入したCRFによるアクセント結合処理
▶印南 圭祐、峯松 信明

『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―

丸山 岳彦　（データ班分担者：国立国語研究所研究開発部門）†
山崎 誠　（データ班班長：　国立国語研究所研究開発部門）
柏野 和佳子　（データ班分担者：国立国語研究所研究開発部門）
佐野 大樹　（データ班連携研究者：国立国語研究所研究開発部門）
秋元 祐哉　（データ班協力者：国立国語研究所研究開発部門）
稲益 佐知子　（データ班協力者：国立国語研究所研究開発部門）
田中 弥生　（データ班協力者：国立国語研究所研究開発部門）
大矢内 夢子　（データ班協力者：国立国語研究所研究開発部門）

Outline of Sampling Method in the Balanced Corpus of Contemporary Written Japanese (3) : Sampling Method and Representativeness of BCCWJ

Takehiko Maruyama　(Dept. Lang. Res., National Institute for Japanese Language)
Makoto Yamazaki　(Dept. Lang. Res., National Institute for Japanese Language)
Wakako Kashino　(Dept. Lang. Res., National Institute for Japanese Language)
Motoki Sano　(Dept. Lang. Res., National Institute for Japanese Language)
Masaki Akimoto　(Dept. Lang. Res., National Institute for Japanese Language)
Sachiko Inamasu　(Dept. Lang. Res., National Institute for Japanese Language)
Yayoi Tanaka　(Dept. Lang. Res., National Institute for Japanese Language)
Yumeko Oyauchi　(Dept. Lang. Res., National Institute for Japanese Language)

1　導入

『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）』が備える最大の特徴として，それが統計的な代表性を備えた均衡コーパスであることが挙げられる。すなわち，対象となる書き言葉に含まれる総文字数を推計して母集団を定義し，その結果をメディア・ジャンル間の構成比率に比例割当することでサンプル構成比を算出するという方法でコーパスが設計されている点である。

BCCWJ の構築プロジェクトが始まり，現在 3 年目が過ぎようとしている。この間，サンプリングを担当する我々のグループ（サンプリングサブグループ）では，BCCWJ の設計段階で準備した「サンプル台帳」に基づき，さまざまなタイプの書籍・雑誌・新聞などから書き言葉を収集してきた。本稿では，BCCWJ におけるコーパスデザインについて再確認し，代表性を実現するためのサンプリングの方法論および「サンプル台帳」に基づくサンプリングの作業方法について述べる。また，現時点におけるサンプリング作業の進捗状況について報告する。

† maruyama@kokken.go.jp

2　BCCWJ のコーパスデザイン

2.1　3 つのサブコーパスと 2 種類のサンプル

まず初めに，BCCWJ のコーパスデザインについて確認しておく（詳細は，丸山・秋元 (2007,2008) を参照)。BCCWJ は,「出版サブコーパス」「図書館サブコーパス」「特定目的サブコーパス」という 3 つのサブコーパス（SC）によって構成される。BCCWJ の内部構成を，図 1 に示す。

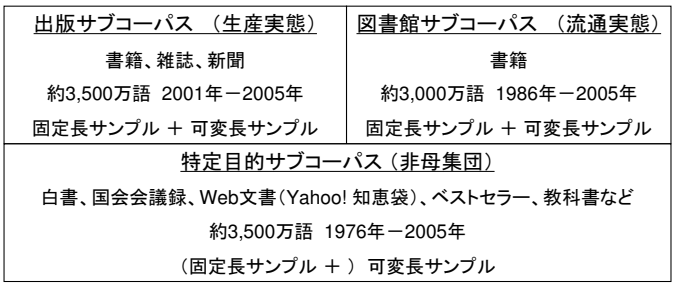


図 1: BCCWJ の内部構成

出版 SC： 出版 SC は，書き言葉の生産力という側面に着目する SC である。2001 年から 2005 年の間に国内で出版された全ての書籍・雑誌・新聞に含まれる文字の総体を母集団として，ランダムサンプリングによって得られる約 3,500 万語分のデータを収める。

図書館 SC： 図書館 SC は，書き言葉の流通・流布の実態という側面に着目する SC である。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（ただし 1986 年から 2005 年の 20 年間に発行されたもの）を対象として，ランダムサンプリングによって得られる約 3,000 万語分のデータを収める。

特定目的 SC： 特定目的 SC は，生産・流通という側面からは捉えきれない，あるいは，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。

抽出単位として,「**固定長サンプル**」「**可変長サンプル**」という 2 種類のサンプルを取得する。これは，それぞれ以下の 2 つの方針を満たすための設計である。

- 統計的に厳密な言語調査に耐え得るよう，母集団からの抽出比を重視した設計にする。
- 文体研究・テキスト研究に耐え得るよう，ある程度の文脈を確保した設計にする。

固定長サンプル： 「固定長サンプル」は，母集団に含まれる全ての文字に対して等確率を与えた上で，ある 1 文字をランダムに指定し，その文字を始点として 1,000 文字の範囲を取得するサンプルである。全ての文字に対して等確率を与えるために，母集団に含まれる文字の総数をあらかじめ推計しておく必要がある。母集団（＝推計された総文字数）からの抽出比が明確である点で，基本語彙表や漢字表の作成，語彙・文字調査など，統計的な言語研究に向く。また，母集団の層的かつ量的な構造を忠実に反映する点で，統計的な代表性を備えた均衡コーパスとしての性格を強く持つ。

可変長サンプル： 「可変長サンプル」は，固定長サンプルと同様，母集団に含まれる全ての文字に対して等確率を与えた上で，ランダムに指定した 1 文字を含む言語的な構造のまとまり（「章」や「節」など，ただし 1 万字を上限とする）を取得するサンプルである。文章・談話としてのまとまりを重視したサンプルであるため，テキストの論理構造の把握や文脈の分析，文体の調査などに向く。

可変長サンプルは，3 つの SC 全てに対して提供される。一方，固定長サンプルは，出版 SC，図書館 SC，および，特定目的 SC の一部（白書など）に対して提供される。サンプル数は，出版 SC の固定長サンプル部分を 1,000 万語分取得することとし，「1.7 文字=1 語」と換算して，17,000 文字，すなわち 17,000 サンプルと定めた。図書館 SC は，出版 SC の書籍部分と母集団がほぼ等しくなるように調整し，そこから出版 SC の書籍部分と同数のサンプル数を取得することにした。

2.2 サンプル構成比と母集団に対する代表性

出版 SC・図書館 SC では，まず母集団に含まれる文字の総数を推計し，その結果を母集団を構成する各層に比例割当することにより，各層から取得するサンプル数を算出した。実際には，母集団を「ジャンル」「発行年」の 2 側面から層別し，各層に含まれる総文字数の量的なバランスに応じて，そこから取得する固定長サンプルの文字数を算出している。このような設計により，両 SC の固定長サンプル部分は，母集団に対する**代表性（representativeness）**を備えていることが保障されている。

出版 SC・図書館 SC の母集団の総文字数の推計結果およびサンプル構成比を，表 1，2 に示す。

表 1: 出版 SC におけるサンプル構成比

層		総文字数	構成比	S 数
書籍	0. 総記	1,636,414,548	2.50%	425
	1. 哲学	2,597,610,813	3.97%	674
	2. 歴史	4,301,204,340	6.57%	1,117
	3. 社会科学	12,408,321,943	18.95%	3,222
	4. 自然科学	5,069,594,034	7.74%	1,316
	5. 技術工学	4,615,929,967	7.05%	1,199
	6. 産業	2,196,387,437	3.35%	570
	7. 芸術	3,258,432,447	4.98%	846
	8. 言語	888,800,128	1.36%	231
	9. 文学	9,341,275,486	14.27%	2,426
	n. 記録なし	2,225,954,208	3.40%	578
書籍 小計		48,539,925,351	74.14%	12,604
雑誌	1. 総合	7,421,447,806	11.34%	1,927
	2. 教育	877,875,592	1.34%	228
	3. 政治	456,459,405	0.70%	119
	4. 産業	110,640,958	0.17%	29
	5. 工業	1,468,293,360	2.24%	381
	6. 厚生	180,964,513	0.28%	47
雑誌 小計		10,515,681,634	16.06%	2,730
新聞	全国紙	2,417,622,461	3.69%	628
	ブロック紙	1,296,592,154	1.98%	337
	地方紙	2,701,855,499	4.13%	702
新聞 小計		6,416,070,114	9.80%	1,666
合計		65,471,677,099	74.14%	17,000

表 2: 図書館 SC におけるサンプル構成比

層	総文字数	構成比	S 数
0. 総記	1,003,528,880	2.01%	264
1. 哲学	2,343,849,711	4.90%	617
2. 歴史	5,010,749,621	10.47%	1,319
3. 社会科学	8,946,058,392	18.69%	2,355
4. 自然科学	3,028,276,363	6.33%	797
5. 技術工学	3,149,144,051	6.58%	829
6. 産業	1,690,150,481	3.53%	445
7. 芸術	4,057,291,256	8.47%	1,068
8. 言語	956,625,910	2.00%	252
9. 文学	15,485,091,056	32.34%	4,077
n. 記録なし	2,206,890,351	4.61%	581
合計	47,877,656,072	100%	12,604

3　代表性を実現するためのサンプリング手法

3.1　サンプリング作業の流れ

前節に示したコーパスデザインにしたがって，データ班のサンプリングサブグループでは，2006年以降，サンプリング作業を進めてきた。サンプリング作業の流れは，以下の 3 点にまとめられる。

1. 層化無作為抽出法によるサンプリングを実施するため，母集団を各層ごとにリスト化し，ランダマイズして,「サンプル台帳」を作成する。
2. サンプル台帳に従って,「サンプル抽出基準点」となる文字を指定する。
3. 一定の手続きに従って，実際の印刷紙面から固定長サンプル・可変長サンプルを取得する。

以下では，母集団に対する代表性を実現するためのサンプリング手法として，上記の 1.〜3. までの手順を我々がどのように実践しているかを示す。

3.2　手順 1：サンプル台帳の作成

一般の標本調査においては，実際の標本抽出を実施するための準備として，母集団の構成要素を記載した母集団リストを作成する。この場合の母集団は，有限個の集合として数量的に把握できるものでなければならない。例えば，住民基本台帳や選挙人名簿，事業所名簿などは，既存のリストを母集団リストとして利用できる例である。また，無作為抽出法を採用する場合，母集団リストに含まれる抽出単位に通し番号を付し，それらを乱数を用いてランダマイズする必要がある。

さて，出版 SC・図書館 SC の場合，その母集団は，表 1，2 に示した文字数によって定義されている。ここから無作為抽出を実施するためには，母集団に含まれる全ての文字に対して等確率を与え，その中からランダムに 1 文字を指定する必要がある。そして，この 1 文字を基準として，2 種類のサンプルを取得するのである。

これを概念的に述べ直すと，母集団を構成する 1 文字目から最後の文字（出版 SC では 65,471,677,099 文字目，図書館 SC では 47,877,656,072 文字目）までを 1 次元上に配置した上で，ランダムに指定された任意の 1 文字から 1,000 文字という範囲，およびその文字を含む「章」「節」などの言語的なまとまりを持つ範囲を，それぞれ固定長サンプル・可変長サンプルとして同時に取得する，ということである。このことを，出版 SC を例に概念的に図示すると，図 2 のようになる。

ここで，文字数によって定義されている母集団をどのようにリスト化してランダマイズしするかという技術的な問題がある。母集団に含まれる全ての文字に等確率を与えて 1 文字をランダムに選び出すことは，理論上は可能であるが，しかしながら，現実的には非常に困難である。

これを近似的に実現するための手段として，次のような方法を採用した。まず，母集団に含まれる全てのページをリスト化し，それらをランダマイズして優先順位を付した。さらに，優先順位の高いページとして選ばれたページの中から，1 文字を無作為に指定した。この 1 文字を，抽出単位を取り出すための基準点（「サンプル抽出基準点」）として，固定長サンプル・可変長サンプルを取得することにした。このような 2 段階の抽出（ページの無作為抽出，文字の無作為抽出）によって，母集団をリスト化し，そこから無作為に 1 文字を指定する作業に近似させることとした。

以下では，書籍の場合を例として,「サンプル台帳」の作成方法と，それをもとに「サンプル抽出基準点」を指定する方法について示す。

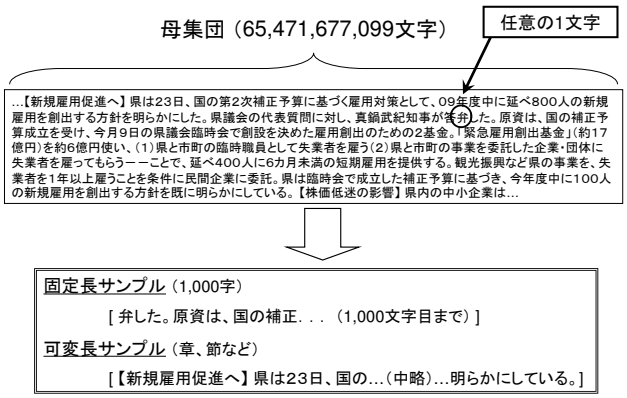


図 2: 「サンプル抽出基準点」の指定とサンプルの取得

出版 SC の「書籍」の母集団には 74,911,520 ページが，図書館 SC の母集団には 85,363,019 ページが，それぞれ含まれている。この全ページ数を， 出版 SC では 55 層（NDC の 11 分類 × 発行年の 5 年），図書館 SC では 220 層（NDC の 11 分類 × 発行年の 20 年）ごとに分類し，書誌情報つきのテーブルとしてリレーショナルデータベース上に展開した。その上で，各層に含まれる全ページをランダマイズして，各ページに対してランダムに優先順位を割り振った。

さらに，各ページに対して，ページ内の 1 点を特定する座標情報をランダムに指定した。これは，ページに 10×10 の座標枠を割り当て，特定された座標の交点に最も近い文字を「サンプル抽出基準点」として指定するためのものである。座標情報は，横軸を 0〜9，縦軸を A〜J として,「0A」から「9J」まで 100 通りの交点を指定した。

ただし，特定されたあるページが白紙であった場合,「サンプル抽出基準点」となる文字を指定することができず，次に優先順位の高いページに移らなければならない。これはほとんどの場合，その場で手に取った書籍を放棄し，次の該当ページを含む書籍を新たに探し出す必要がある。しかしながら，ランダムに指定された特定の書籍を探し出すことは実際には非常に手間のかかる作業であり，同一の書籍内から次候補を探し出す方がはるかに効率的である。そこで，作業進行上の効率を考慮して，ある書籍のうち最も優先順位が高いページから文字が指定できなかった場合は，同じ書籍の中で次に優先順位の高いページに移ってよいこととした。これを上位 20 位まで繰り返してよいこととし，20 位までのページから 1 文字も指定できなかった場合は，その書籍をサンプリングの対象から除外することとした。

また，交点の直近が図・写真であったり，交点の直近に文字がなかったりする場合,「サンプル抽出基準点」となる文字を指定することができない。そこで，1 ページあたり 10 通りの交点を準備し，それぞれに優先順位を付した。第 1 位の交点で文字が指定できなかった場合，第 2 位の交点に移ってよいこととした。第 10 位までの交点でも文字が指定できなかった場合，次に優先順位の高いページに移ることとした。

同一の書籍から優先順位の高い上位 20 位までのページ番号と，それに付随する 10 位までの座標情報を含めて，1 枚の台帳とすることにした。これに，書籍のタイトルや，サンプル管理用の ID などを表示させて，1 枚の「サンプル台帳」としてまとめた。サンプル台帳の例を，図 3 に示す。

サンプル取得数　　借り出し日

1　SampleID： PB58_00482　BibID： 20770388　タイトル: 伝え合いの言葉　2006/12/28

NDLsize　　配架情報

21cm　　国語研図書館

分類せず

優先順位	対象頁	有効	乱数1	乱数2	乱数3	乱数4	乱数5	乱数6	乱数7	乱数8	乱数9	乱数X	備考
一位	49		4E	2H	2D	0D	7D	4J	2I	7A	5I	1A	
二位	225		8D	8A	8J	6H	1C	9I	5H	4E	9D	0F	
三位	78		4J	0H	0I	7J	7C	8D	7B	2E	6E	6J	
四位	20		9J	2B	0J	0G	4I	2E	1E	7H	5J	3B	
五位	115		4D	5F	9F	8I	9E	4A	9J	2I	6H	3B	
六位	5		0H	8A	3H	1G	5I	2A	2D	6D	9E	4D	
七位	108		0G	5D	1E	0I	2I	7C	5A	2A	4H	1I	
八位	12		0C	6D	2E	3D	3F	6E	9A	0G	8J	2H	
九位	201		2J	3A	8B	0F	5E	0H	7D	4B	2B	7C	
十位	152		1E	4E	7C	1C	0F	9G	7G	4J	9D	5D	
十一位	232		2G	1D	7J	7C	0F	2H	8E	7F	5D	4C	
十二位	242		3H	2F	9J	4G	4J	6D	1G	3J	4C	2B	
十三位	51		8D	7I	4H	3E	6J	1J	1D	0J	0B	9C	
十四位	44		1I	2I	6I	7F	4H	0E	0I	0F	2E	6J	
十五位	69		5A	2E	6I	9A	7F	1J	4G	7I	4H	5C	
十六位	233		3F	1J	4I	3D	8D	5F	5C	0H	1E	1A	
十七位	153		4I	1C	2C	3G	0H	4A	4B	4G	8E	2G	
十八位	159		5E	3J	0H	3I	0G	1D	9F	8I	7H	6E	
十九位	158		7A	4B	0H	0I	2C	0C	6A	2J	8D	3I	
二十位	193		6E	6B	0F	8I	5B	8J	6G	7D	3B	2H	

図 3: サンプル台帳の例

以上のような手順によって，出版 SC では 74,911,520 ページ分の，図書館 SC では 85,363,019 ページ分のサンプル台帳を作成した。これにより，母集団に含まれる全てのページから特定の 1 ページを，さらにそのページに含まれる特定の 1 文字を，それぞれランダムに指定することができる。

3.3　手順 2：サンプル抽出基準点の指定

実際のサンプリング作業では，サンプル台帳で指定された書籍を手に取り，指定されたページ中の指定された座標に最も近い 1 文字を見つけて「サンプル抽出基準点」とする。図 3 の例では，この書籍の中で優先順位が「一位」である「49 ページ」の交点「4E」に最も近い文字を探し出すことになる。この様子を，図 4 に示す。なお，実際の作業では，座標の枠を印刷した透明のシート（「サンプル抽出基準点」指定シート）を書籍の判型ごとに用意し，印刷紙面に当てることにより，該当の 1 文字を指定している。

図 4 を見ると，図 3 のサンプル台帳で指定された第一位である「49 ページ」の交点「4E」に最も近い文字は,「た」である。そこで，この文字を「サンプル抽出基準点」として指定する。

仮に，指定された交点がページ右下のイラストに当たってしまった場合，第 2 位の優先順位を与えられた交点に移動して，その直近の文字を指定することになる。先述のように，1 ページあたり 10 通りの交点が準備されているので，少なくとも図 4 のような紙面構成であれば，10 通りのうちいずれかの交点から，1 文字を指定できることになる。また，仮に指定された 49 ページが白紙ページだったとしたら，次に高い優先順位が与えられている，同じ本の 225 ページを開き，交点「8D」に直近の文字を探すことになる。仮に 225 ページも白紙ページだったら，その次に高い優先順位のページに移動していくことになる。

なお，上位 20 位までのページから 1 文字も指定できない書籍も，まれに存在する。デザイン集やカット集，図鑑などのように，イラストや写真が主体となって書籍全体が構成されているものや，古文や外国語など非現代日本語のみで書籍全体が構成されているものなどがこれに該当する。

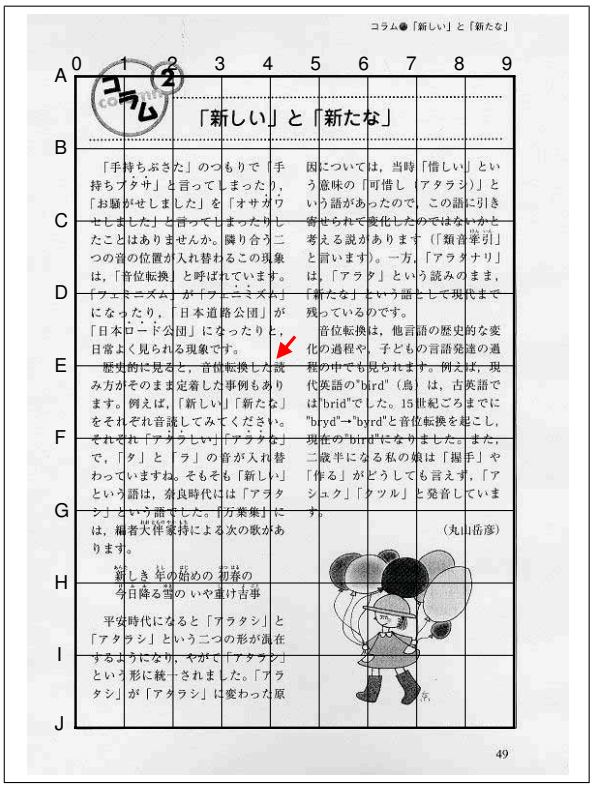


図 4: サンプル台帳で指定されたページ・座標から 1 文字を指定する例

3.4　手順 3：固定長サンプル・可変長サンプルの取得

サンプル抽出基準点が指定できたら，続いて固定長サンプル・可変長サンプルの範囲を取得していくことになる。実際の作業では，サンプル抽出基準点を起点として 1,000 文字を人手で数えて固定長サンプルの範囲を取得し，またサンプル抽出基準点を含む章や節が 1 万字を超えていないことを確認して可変長サンプルを取得している。

ここで問題となるのは，印刷紙面上に書かれている文字列のうち，どの部分をどのような順序で取得するか，という点である。一見，書き言葉として書かれている文章を取り出すのは簡単な作業のように思われるが，実際には非常に詳細な規則と判断基準が必要になり，かつ事例ごとに柔軟な対応が求められる場合が多い。実際の印刷紙面から固定長サンプル・可変長サンプルを均質的な手続きにより抽出するためには，紙面を構成する諸要素のうち，どの要素を抽出し，どの要素を抽出しないのかを前もって決めておかなければならない。

そこで以下では，書籍における書き言葉の構造を，「冊子」「印刷紙面」「文字」という 3 つの側面によって段階的に捉え，その中からコーパスに収録するサンプルとして取得する部分を絞り込んでいく基準と，それが満たすべき条件について示す。

「冊子」　書籍の冊子は通常，「本文」と呼ばれる書籍の実質的本体以外に，「口絵」「標題紙」「献辞」「前書き」「目次」「凡例」などの「前付」の要素や，「付録」「索引」「後書き」「奥付」などの「後

付」の要素などから構成されている。ここでサンプリングの対象とするのは，本文に加え，一定の文章量のある「前書き」と「後書き」とする。これ以外の要素や「広告」などは，原則，サンプリングの対象としない。

「印刷紙面」 書籍の印刷紙面は，さまざまなレイアウトを持つ要素から構成される。そのうち，「見出し」「本文」「キャプション」「注」は，サンプリングの対象とする。一方，文字を主体としないフィギュア（「イラスト」「写真」等）は，サンプリングの対象から外す。また，行列見出しを持つ「表」や，分岐型の「フローチャート」などは，文字が主体であっても一方向に読み進めることができないため，フィギュアに相当する扱いとし，サンプリングの対象外とする。さらに，実質的内容をもたない「ノンブル」「柱」も，サンプリングの対象外とする。

「文字」 印刷紙面上に印字されている文字のうち,「仮名」「漢字」「数字」「アルファベット」はサンプリングの対象とする。一方,「句読点・疑問符・感嘆符」「括弧・その他記号」などは，入力はするが，固定長サンプル 1,000 字のカウント対象とはしない。この区別は，純粋な言語表現を構成する文字種に限定して 1,000 字を取得することにより，より精密な文字調査や語彙調査を実現しようという意図による。

さらに，上記までの基準に加えて，本文部分に含まれる言語表現そのものに関する条件として,「現代日本語として書かれたもの」という条件を設ける。したがって，以下のような「非現代日本語」の表現がまとまって出現した場合，その部分はサンプリングの対象から外す。

- 非日本語（英語，フランス語，中国語等）
- 非現代語（明治元年より前に書かれた日本語）
- 非言語（数式，化学式等）

ただし,「彼は Thank you とだけ言った。」という例のように，一連の本文中に非現代日本語が混じっている場合は，その部分だけを除外することはしない。上記の各表現がサンプリングの対象外となるのは，ページや章，あるいは書籍全体が非現代日本語で構成されている場合，または，典型的には，前後に改行を伴い，主たる本文からインデントされてブロック形式で引用されている場合である。

以上のような，印刷紙面から抽出する文字列の基準を取り決めた上で，固定長サンプル・可変長サンプルを取得するのである。この作業を概念的に述べ直すと，書き言葉からサンプルを取得する作業とは，多様な構造を持つ書き言葉の実体を **1 次元の文字列（1 個以上の文字の連鎖）**として配置し，そこから一定範囲の抽出単位を取り出す作業であると考えることができる（図 2 参照）。

なお，印刷紙面から固定長サンプル・可変長サンプルを取得していくための基本方針，および技術的な問題点については，柏野ほか (2009) を参照されたい。

4　サンプリング作業の進捗状況

4.1　書籍（出版 SC・図書館 SC）の進捗状況

最後に，2008 年度末の時点におけるサンプリング作業の進捗状況について報告する。

出版 SC・図書館 SC の書籍のサンプリング作業は，2006 年度から開始し，2008 年度末の現在，図 5 に示すような達成状況となっている。横軸は，SC および NDC による層別で，括弧内は必要サンプル数である。縦軸は，必要サンプル数に対してサンプリングが完了した達成率を表す。

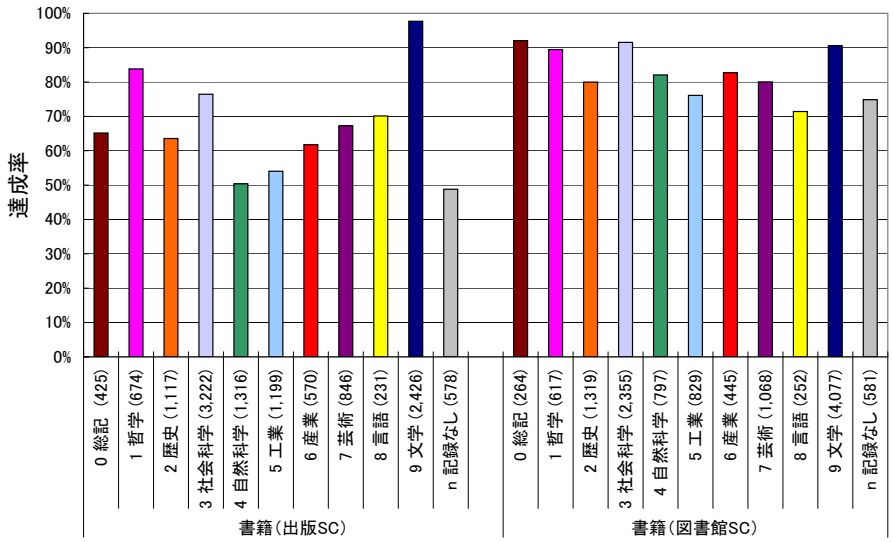


図 5: 書籍（出版 SC・図書館 SC）のサンプリングの達成状況（2008 年度末）

図書館 SC が平均 83%の達成率である一方，出版 SC は平均 67%と，低めの達成率となっている。これは，対象となる書籍の入手可能性によるところが大きい。設計当初，出版 SC で母集団を定義するのに用いたのは，2001 年から 2005 年までに出版された書籍全体のリスト（国立国会図書館の蔵書目録「J-BISC」）であった。ここからランダムに選ばれた書籍には，自費出版による書籍，地方自治体が作成した地域資料，省庁の事業報告書やハンドブック，ISBN の付いていない書籍など，一般には出回らない書籍も多く含まれる。これらの書籍は，近隣の図書館や古書店で閲覧・入手できる可能性が極めて低い[1]。一方，図書館 SC では東京都内の公共図書館に共通に所蔵されている書籍からランダムに書籍を選んでいるため，近隣の図書館や古書店で入手できる確率が高く，作業を順調に進めることができた。

4.2 雑誌・新聞（出版 SC）の進捗状況

次に，出版 SC の雑誌・新聞についての進捗状況を示す。雑誌は 2008 年度から，新聞は 2007 年度からサンプリングを開始し，2008 年度末の現在，図 6 に示すような達成状況となっている。

作業開始からまだ時間が経っていないこともあり，雑誌の進捗状況は平均 38% と低い達成率に留まっている。今後も雑誌・新聞のサンプリング作業は継続していくことになるが，特に雑誌のバックナンバーの入手が極めて困難であるという問題が生じている。特に公立図書館では雑誌のバックナンバーが廃棄されるまでの期間が短く，東京都立多摩図書館，八王子市図書館，横浜市中央図書館など，雑誌のバックナンバーを比較的多く所蔵している近隣図書館で閲覧するか，数は少ないが古書店

[1] 無論，国立国会図書館で閲覧することは可能ではある。

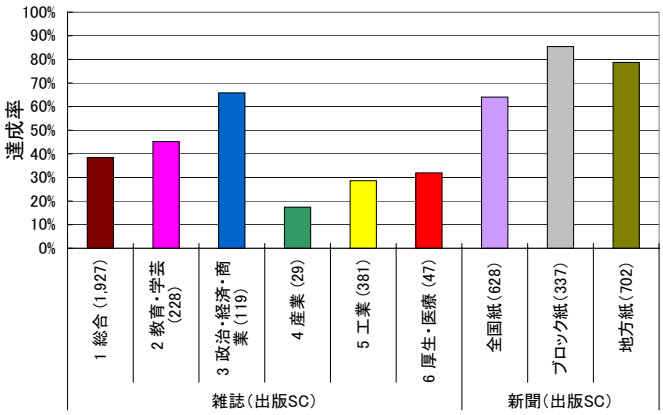


図 6: 雑誌・新聞（出版 SC）のサンプリングの達成状況（2008 年度末）

でバックナンバーを購入するかという状態が続いている。

5 まとめ

以上，本稿では，代表性を実現するためのサンプリングの方法論として,「サンプル台帳」に基づくサンプリングの作業方法について述べた。また，2008 年度末の時点におけるサンプリング作業の進捗状況について報告した。今後は，書籍・雑誌・新聞の各メディアについて，サンプリングの達成率を全体的に上昇させていくことが課題である。

十全な均衡コーパスのあり方やその実現方法については，今後試行錯誤を重ねながら議論されていくことが望まれる。その際に重要なのは，コーパスがどのような方針で設計され，どのような手順で構築されたのかという，設計方針と構築手順の 2 点であろう。BCCWJ の構築過程では，これらの点が逐一明確に提示されてきており，コーパスの設計および構築の妥当性を評価する上で，これらの情報の開示は極めて重要な意義を持つと考える。

謝辞： サンプリング作業にご協力いただいている以下の諸機関に，記してお礼申し上げます。国立国会図書館，東京都立中央図書館，東京都立日比谷図書館，東京都立多摩図書館，立川市中央図書館，八王子市中央図書館，横浜市立図書館，一橋大学附属図書館，自治大学図書室。なお，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。

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「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-01)

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「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-2)

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「日本語コーパス」平成 19 年度研究成果報告書 (JC-D-07-1)

『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とその応用

山口 昌也　(データ班分担者：国立国語研究所研究開発部門) †
間淵 洋子　(データ班分担者：国立国語研究所研究開発部門)
西部 みちる　(データ班協力者：国立国語研究所研究開発部門)
小林 正行　(データ班協力者：国立国語研究所研究開発部門)
大島 一　(データ班協力者：国立国語研究所研究開発部門)
高田 智和　(データ班分担者：国立国語研究所研究開発部門)

Applications of Text Encoding Format in the Balanced Corpus of Contemporary Written Japanese

Masaya YAMAGUCHI　(Dept. Lang. Res., National Institute for Japanese Language)
MABUCHI, Yoko　(Dept. Lang. Res., National Institute for Japanese Language)
NISHIBE, Michiru　(Dept. Lang. Res., National Institute for Japanese Language)
KOBAYASHI, Masayuki　(Dept. Lang. Res., National Institute for Japanese Language)
OSHIMA, Hajime　(Dept. Lang. Res., National Institute for Japanese Language)
TAKADA, Tomokazu　(Dept. Lang. Res., National Institute for Japanese Language)

1　はじめに

「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese, 以後，“BCCWJ” と表記) は，言語学，国語教育，日本語教育，辞書編集，自然言語処理など幅広い分野での利用を想定し，書籍，新聞，雑誌，Web データなど，さまざまな媒体のテキストが収録されている。また，個々のテキストには，書誌情報，文書構造情報，文字情報といった，さまざまな情報が付与される。

本稿では，付与される情報を利用した，研究への応用例を示す。今回利用する付与情報は，主として，章・節，発話，引用，段落，文などの文書構造情報である。これらの情報を利用して，(1) テキスト検索への応用例，(2) 文書分類への応用例を示す。

2　電子化フォーマットの概要

BCCWJ には，「可変長」「固定長」の 2 種類のサンプルが収録され，それぞれの電子化フォーマットが XML の文書型として規定される。可変長サンプルは，一つのサンプルが一つの「記事」に相当する。固定長サンプルは，1 サンプル中に 1000 文字を包含したサンプルである。この後示す例では，可変長サンプルを使用する。

XML で定義される文書要素は，大きく分けて，(1) サンプルに関するタグ (例：サンプリングポイントの情報)，(2) 文字・表記に関するタグ (例：外字，ルビなどの情報)，(3) 文書構造に関するタグの三つがある。可変長の文書型では，合計 46 種類のタグを定義している。なお，タグの詳細については，山口他 (2008)，または，電子化フォーマット仕様の Web サイト (http://www2.kokken.go.jp/densi/public/wiki/) をご覧頂きたい。

† masaya@kokken.go.jp

3　応用例

3.1　検索への応用

実際の言語研究を行う場合，コーパスから自分の研究目的にあった部分だけを抽出して，検索したり，分析に利用したい，ということがある。このような場合，抽出する際の条件としては，メディアやジャンル，著者の情報などが良く使われる。本稿では，抽出条件を詳細化するために，文書構造の情報を使うことを考える。

BCCWJ の電子化フォーマットには，文書構造を表現する 32 種類のタグが定義されているが，今回は，論理的な階層構造を表わすタグを利用してみる。具体的には，article (記事に相当)，cluster(章，節などのタイトル付きの一まとまりの文章)，list （箇条書きなどの列挙される文書要素）といったタグである。記述例を図 1 に示す。これらのタグでマークアップされる文書要素は，論理的にも視覚的にも区別しやすく，また，後述するように，文章類型との関連も深いので，抽出条件として，利用者が利用しやすいと考えられる。

上記のタグの利用例として，「平均階層数」を定義し，実際のデータと対応づけてみる。ここで言う「平均階層数」とは，章や節など入れ子構造などによる階層の度合であり，テキスト中の各文字が平均的にいくつのタグでマークアップされているかを表わす。例えば，論文などのように，深い章だてがなされている文章では，平均階層化数は大きくなり，小説などのように階層をほとんど持たない文章では，小さくなる。

図 2 は，BCCWJ 領域内公開データ 2008 年度版 (以後，BCCWJ2008) に収録されている書籍 (流通実態コーパス) の NDC5 番台 (技術・工業)，9 番台 (文学) のサンプルの平均階層数を棒グラフにしたものである。図の横軸は平均階層数，縦軸はその割合をサンプルの NDC ごとに取りまとめた値である (横軸の区間幅は 0.25)。例えば，NDC9 番台（文学）のサンプルは，約 88%のサンプルが平均階層数 1 であることを示している。

この図のとおり，NDC9 番台では平均階層数 1 にサンプルが集中し，NDC5 番台では 9 番台よりも広範囲にサンプルが分布する。この理由して，NDC9 番台には，小説やエッセーなどの階層化の度合が少ないサンプルが多く含まれていること，NDC5 番台には専門性の高く，階層化の度合が高いサンプルが含まれていることが考えられる。

平均階層数に加えて，他の付与情報を総合的に利用することにより，さらに詳細な抽出条件を指定できる。例えば，平均階層数が高く，小さい cluster 要素が多数含まれるサンプルは，辞書や商品紹介などの場合が多い，平均階層数が少なく，speech タグ (発話を表わす) を多く含む文章は，会話文を多く含む小説や対談であることが多いといった傾向が見られる。

このように，BCCWJ の付与情報は，効率的に望みのサンプルを抽出したり，逆に分析にとってノイズとなるサンプルを除外するのに有用であると考えられる。

```
<article articleID=...>
<titleBlock><title>コーパスの構築について

<cluster>
<titleBlock><title>
1章 はじめに
</title></titleBlock>
<cluster>
<titleBlock><title>
1.1 背景
</title></titleBlock>
```

図 1: XML データの例

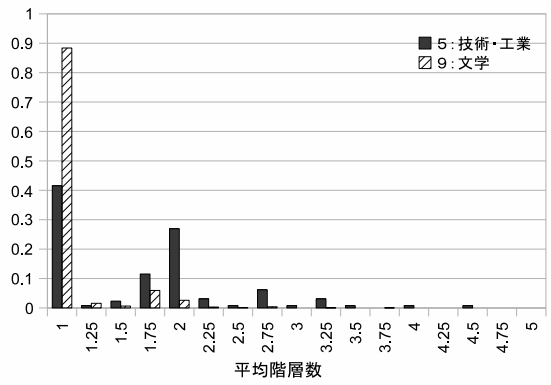


図 2: NDC 別の平均階層数

3.2　文書分類への応用

コーパスを利用する上で必要となるコンテクスト情報については，既に BCCWJ で分類情報の提供が確定している NDC 等の主題による分野分類では十分でない可能性が指摘されており，これらを補完する分類について枠組や分類手法が検討されている（佐野 2008，柏野他 2008）。

本節では，電子化フォーマットの応用例として，XML タグによる文書構造情報が文章類型の分類に利用できる可能性の一端を示したい。

文書の分類方法にはさまざまな観点に基づくものが考えられるが，ここで目指すのは，言語の位相研究に有用な分類，特に以下のような位相の差を捉えることができる分類である。

専門的-非専門的　フォーマル-インフォーマル　客観的-主観的　文章語的-口頭語的

このようなことばの位相差が，文章のどのような側面と呼応して生じるかを考えるにあたり，文章が記され，読まれる目的・動機とのかかわり[1]を取り上げる。そして，目的・機能別の文章類型に基づき文書を分類する試行として，一部の文章類型を想定し，これらの文章類型を抽出する手段として，文書構造の分析（＝電子化フォーマットによる文書の構造化）を用いる方法を探る。

3.2.1　BCCWJ2008 を用いた分析例

BCCWJ2008 に格納される書籍（生産実態，流通実態各サブコーパスおよび非母集団サブコーパスに含まれるベストセラー）可変長サンプルのうち，ある著者によってあるテーマをもとに記された文章 (以下「記事」と呼ぶ）の全体を格納し得た 1425 サンプルを対象として，以下を調査した。

A　特殊な構造要素の割合と文章類型のかかわり

B　文章の階層性，文の長さと文章類型のかかわり

A. 構造要素の割合　まず，「特殊な構造要素が，テキスト中にどの程度含まれるかは，文章類型によって異なりが見られる」という仮説を用意する。この仮説が妥当であれば，文章内において一定の役割を持つ要素の有無や割合によって，どのような目的で書かれた文章かを推定できる可能性がある。

仮説を検討するにあたって，BCCWJ データに付与されている構造化 XML タグを利用し，サンプル内の文字列要素を以下の「構造要素」に分類してそれぞれのサンプルにおける各構造要素の比率を調査した。

見出し要素類 記事見出しや下位構造要素（章や節など）の見出し，それに付随する要素など。... titleBlock, title, orphanedTitle 要素

記事情報要素類 記事の主体となる本文ではなく，記事そのものについての情報に相当するような要素。著者情報，目次，記事概要，注記など。... authorsData, contents, abstract, profile, noteBody 要素

図表関連要素類 記事の主体となる本文に対して，補足・参照の役割を担う図表や，それらに付随する図表タイトルやキャプションなど。... figure, caption 要素

発話要素 いわゆる「地の文」に相当しない発話表現。... speech 要素

引用要素 いわゆる「地の文」に相当しない引用表現。... citation 要素

[1]目的という観点からの文章分類はさまざまあり，永野 (1968)，金岡 (1968) などに整理されているが，主に森岡 (1979) の 4 分類（報告する，納得させる，印象づける，行動させる）を元に，資料の分析を通じて，分類の枠組についても検討していく予定である。

主本文要素 上記を除く文字列要素。いわゆる「地の文」にほぼ相当する。

これらの構造要素は，同じ文章中にありながら，役割が大きく異なるため，どのような目的で文章を書くかによって必要性が異なり，これらが文章中に占める割合も異なることが想定される。例えば，ある特定の構造要素の役割と，それを用いて記される文章類型との対応関係について以下のような枠組みを仮定することができる。

要素	想定される役割	対応する文章類型
発話	・発話を書き留める	記録，報告する文章
	・談話の臨場感により読み手の興味を引く	印象づける文章
図表	・視覚的に表現することで理解を助ける	納得，行動させる文章
	・根拠となるデータ等を示すことで説得力を高める	納得させる文章
引用	・引用したものについて説明をする	解説する文章
	・引用することで根拠を示す	納得させる文章

それぞれの要素の有無や文章全体に占める割合によって，類似性を認められる文書群が，特定の文章類型として認識できるものであれば，これらの要素は，文章類型を特徴付けるものとして文書分類に有用な指標であると言える。

この項では，その試みとして，発話要素と文章類型のかかわりについて観察を行なう。

構造要素の文章全体に占める割合については，文字を単位として求めた。図 3 に計測結果を示す。横軸は NDC の第 1 区分，縦軸は発話含有率（%）である。

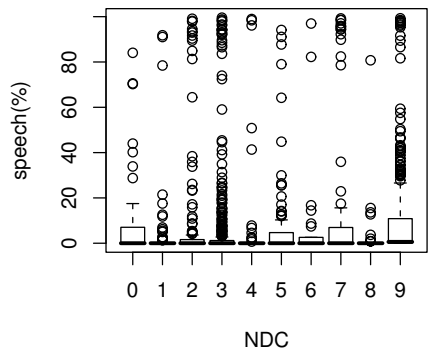


図 3: NDC 別発話含有率

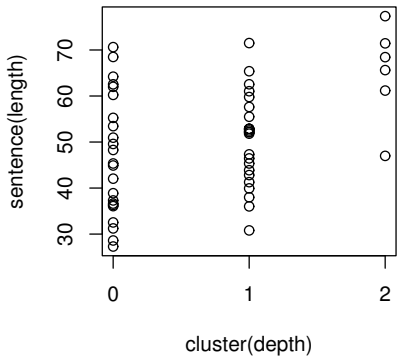


図 4: NDC210 の階層，文長分布

図 3 から，殆どのカテゴリで，発話率は低い方に集中している（発話含有率の中央値は 0 %，平均は 8.87 %）が，「NDC 9. 文学」で発話率が他のカテゴリより高く現れることが分かる。調査対象とした NDC 9 番台のサンプルは，小説がその 37 %程度を占めるため，会話文の多用などから，他のカテゴリのサンプルに比して発話率が高くなる傾向にあるのは，容易に想像できる。一方，80 %以上の部分にいくらかのまとまりが見られるが，こちらは，ある特定の NDC に特に偏って現れておらず，主題に基づかない文書の類似性を示唆している。

ここに認められる「発話主体で構成される」という性質をもつ文書群は，「発話を記録する」という目的により生成されるものと考えられ，議事録，スピーチ書き起こし，対談・インタビュー，戯曲・シナリオといった文章類型とのかかわりが想定できる。これらは，通常の書き言葉とは位相を大きく異にするものであり，他の文書と区別されるべきものであるが，NDC やタイトルからは，そのような情報を得にくい (表 1 参照)。

表 1: 発話率の高いサンプルの例

NDC	書籍タイトル
304 論文集. 評論集. 講演集	時の潮騒
491 基礎医学	養老孟司アタマとココロの正体
323 憲法	21 世紀と日本国憲法
495 婦人科学. 産科学	はじまった着床前診断
938 英米文学 作品集	マーク・トウェインコレクション
210 日本史	国のつくり方
388 伝説. 民話 [昔話]	声の神話
494 外科学	がん治療最前線

実際に発話率 70 %以上の 78 サンプルを確認したところ，対談・インタビューが 63，スピーチ・名言集が 4，議事録が 2，その他（小説，落語，エッセイ，ルポ）が 9 サンプルという結果であった。電子化フォーマットを用いて得られる発話率の高さは，それだけで特定の文章類型を特徴付ける有用な指標となりうると言える。

B. cluster 深度と sentence 長　次に，文章の階層性と文章類型とのかかわりについて観察する。その際，文長とのかかわりを同時に見ていく。

先程と同様，文章を特徴づける指標，その役割や性質，それを用いて記される文章類型との対応関係について以下に枠組みを示す。

指標		想定される役割	対応する文章類型
階層性	高	トピックを分析的に掘り下げ詳細に述べる，多くの事柄を集めて幅広く説明する	説得，報告，動作を促す文章
	低	連続性を重視して述べる，一つの事柄をじっくりと述べる	物語，記録，意見表明する文章
文長	長	事柄を詳細に解説し理解を促す	専門性の高い文章
	短	事柄を簡潔に分かりやすく表現する	一般的，実用的な文章

二つの指標により分類される文書が，想定する文章類型に当てはまるものであれば，これらの指標は，文章類型を特徴付けるものとして文書分類に有用な指標であると言える。

この項では，その試みとして，階層性・文長と文章類型との関係を観察する。階層性，文長を示す指標として，電子化フォーマットから抽出可能な情報，「最下層の cluster の深さ」「sentence の文字数（調査 A で用いた構造要素のうち，「主本文要素」の sentence を対象として調査した 1 文字あたりの文字数平均)」を用いる。

主題の影響を排除するため，特定の NDC に絞って実態を確認してみよう。例として「NDC 210 歴史」を取り上げ，分布を図 4 に示す。横軸は階層の深さ（「2」は階層 2 以上のもの)，縦軸は文の長さ（文字数。平均，中央値は共に約 50 文字）である。また，表 2 に，階層・文長と文章類型との対応例を示す。

表 2: 階層・文長と文章類型の対応例

	書籍タイトル	階層	文長	文章類型
1	日本近現代史を問う	2	68.44	論文
2	民衆史入門	1	65.38	論文
3	倭国を掘る	1	52.51	論文
4	20 世紀高度成長日本	1	39.93	歴史読み物
5	近代日本と国際社会	0	64.20	論文前書き
6	「文芸春秋」にみる昭和史	0	50.95	エッセイ
7	エッセイで楽しむ日本の歴史	0	32.53	エッセイ

実際のサンプルとの対応例として，表 2 に挙げた文書 1, 7 の本文冒頭を引用する。

文書 1 の本文冒頭 (山田敬男 (2002)『日本近代現代史を問う』学習の友社，p.66)

> 日露戦争後から一九二〇年代末までの時期は、大日本帝国憲法の体制下においては最も民主的な運動が広範に展開され、政党政治の慣習が成立するとともに、普通選挙が実現するなどの進歩がみられた時期で、昭和期のファシズムに比して、大正デモクラシーともいわれています。しかし、その一方で、この時期が、軍拡と対外膨張・権益拡大の時期であったことも確かです。

文書 7 の本文冒頭 (文芸春秋編 (1997)『エッセイで楽しむ日本の歴史』文芸春秋，p.503)

> 国定忠治が侠客か、それともギャングだったのかは、考察しても仕方がないだろう。国定忠治を一言で評するならば、暴れん坊の異端児としたほうがいい。
> それに国定忠治が歴史上の人物になり得たのは、関所破りという罪名で磔刑に処せられたからである。もし忠治が磔になっていなければ、後年ヒーローに祭り上げられることはなかっただろう。

階層化が深い階層 2 以上のサンプルは，文長が長く，いずれも論文タイプである。調査対象の NDC210 のサンプルでは，階層が深く文長が短いものは見られなかった。

階層化された階層 1 のサンプルは，文長に幅があるが，ほとんどが論文タイプである。文長の短いものに，発話主体のサンプル（対談）や，歴史読み物，論文の前付けなどが見られた。

階層化されていない階層 0 のサンプルも，同様に文長に幅があるが，こちらは歴史読み物（小説的）やエッセイが多い。文長の長いものに，引用主体の論文，解説，評論などが見られた。

以上のように，階層性・文長と文章類型についても，緩やかな関連性が見られ，他の指標と組み合わせることで，特定の文章類型を特徴付ける指標となりうることが期待できる。

4　おわりに

本稿では，BCCWJ の電子化フォーマットの概要を示すとともに，その応用例として，(1) テキスト検索への応用例，(2) 文書分類への応用例を示した。

参考文献

[1] 山口昌也，高田智和，北村雅則他 (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) 国立国語研究所.

[2] 柏野和佳子，丸山岳彦，秋元祐哉他 (2008) 『『現代日本語書き言葉均衡コーパス』における書籍サンプルの多様性』(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) 国立国語研究所.

[3] 佐野大樹 (2008)「大規模バランスとコーパスにおけるテクスト分類—システミック理論の観点から—」『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.83-90.

[4] 永野賢 (1968)「文章の分類論」『作文講座 4 文章の理論』明治書院，pp.94-141.

[5] 金岡孝 (1968)「現代における文章研究の展望と将来の課題」『作文講座 4 文章の理論』明治書院, pp.244-269.

[6] 森岡健二 (1979)「コピー研究 すぐれた表現の条件 (1)」宣伝会議, 6:7，pp.52-54.

書籍コーパス（流通実態サブコーパス）の「外字」

高田　智和（データ班分担者：国立国語研究所研究開発部門）†
小林　正行（データ班協力者：国立国語研究所研究開発部門）
間淵　洋子（データ班分担者：国立国語研究所研究開発部門）
西部みちる（データ班協力者：国立国語研究所研究開発部門）
大島　一　（データ班協力者：国立国語研究所研究開発部門）
山口　昌也（データ班分担者：国立国語研究所研究開発部門）

Un-coded Characters in the BCCWJ (the Library Subcorpus)

TAKADA, Tomokazu (Dept. Lang. Res., National Institute for Japanese Language)
KOBAYASHI, Masayuki (Dept. Lang. Res., National Institute for Japanese Language)
MABUCHI, Yoko (Dept. Lang. Res., National Institute for Japanese Language)
NISHIBE, Michiru (Dept. Lang. Res., National Institute for Japanese Language)
OSHIMA, Hajime (Dept. Lang. Res., National Institute for Japanese Language)
Masaya, YAMAGUCHI (Dept. Lang. Res., National Institute for Japanese Language)

１．はじめに

「日本語コーパス」では、JIS X 0213:2004 規格が定める文字セットに準拠して、書籍、新聞、雑誌などの文字処理を行っている。本発表では、2008 年 7 月にモニター公開されたコーパスの中から、流通実態サブコーパス（図書館の蔵書を母集団とする書籍コーパス）を対象として、文字量の計測を行い、JIS X 0213:2004 規格で表現できない文字（いわゆる「外字」）について報告する。

２．JIS X 0213:2004 文字セットにもとづく符号化

JIS X 0213:2004 規格は、第 1～第 4 水準漢字 10,050 字、非漢字（仮名・ラテン文字・記号・符号など）1,183 字、計 11,233 字の符号化文字集合である。JIS X 0213:2004 をサポートする機器が徐々に増えつつあることから、将来的に普及するであろうとの見通しのもと、「日本語コーパス」の文字処理は JIS X 0213:2004 に準拠している。しかし、情報機器の符号化表現は Unicode 方式に移行中であるので、Unicode の文字セットから JIS X 0213:2004 に対応する符号化文字を選び、その範囲内で文字処理を行う運用をしている。つまり、JIS X 0213:2004 に準拠とは、文字セットを利用するということである。

JIS X 0213:2004 規格は、199 の漢字字体の包摂規準を規定している。例えば、包摂規準連番 128（  ）は、しんにょうの点に関する規準であり、規格に則ると、しんにょうの点の違いは区別されない。コーパスの文字入力でも、JIS 包摂規準にしたがうので、サンプルに「1 点の  と 2 点の」のような事例が出てきたとしても、入力し分けることはし

† ttakada@kokken.go.jp

ない。よって、JIS X 0213:2004 をサポートする機器でコーパスを展開すると、前の例は「1 点の謎と 2 点の謎」となるであろうし、JIS X 0213:2004 をサポートしない機器では「1 点の謎と 2 点の謎」と表現されるだろう。「日本語コーパス」からは、JIS 包摂規準に記述された粒度の漢字字体の違いについて、情報を得ることはできないので留意されたい。

３．流通実態サブコーパス（書籍可変長サンプル）の文字量

モニター公開版には、流通実態サブコーパスに該当する書籍のサンプルが 1,101 収録されている。今回は、可変長サンプルを用いて文字量を計測する。まず、文字学的分類による計量結果を表 1 に示す。平仮名・片仮名・漢字を分母とする漢字含有率は 31.8%である。

表 1：流通実態サブコーパス（書籍可変長サンプル）の文字量

文字種	異なり字数	延べ字数（%）
平仮名	83	3,126,211 （56.72%）
片仮名	85	270,776 （4.91%）
漢字	4,303	1,580,973 （28.68%）
ラテン文字・ギリシア文字	60	16,391 （0.30%）
アラビア数字	10	11,929 （0.22%）
符号・記号	108	505,463 （9.17%）
計	4,649	5,511,743

次に、JIS X 0213:2004 の水準領域によって、計量結果を再分類したものを表 2 に示す。JIS X 0213:2004 文字セットの運用状況を検証するための表となる。

表 2：JIS X 0213:2004 による符号化

水準	異なり字数	延べ字数（%）
第 1 水準漢字	2,875	1,565,486 （28.40%）
第 2 水準漢字	1,311	10,167 （0.18%）
第 3 水準漢字	93	1,722 （0.03%）
第 4 水準漢字	12	13
X0208 非漢字	316	3,934,018 （71.38%）
X0213 非漢字	30	214
「外字」	12	123
計	4,649	5,511,743

JIS X 0213 は、JIS X 0208 を拡張した規格であり、第 3・第 4 水準漢字と X0213 非漢字が追加されている。最初に制定されてから 10 年余りが経過するが、環境が完全に整っていないこともあってか、JIS X 0213 を用いて文字処理を大規模に行おうとするのは、管見で

は「日本語コーパス」が初めてである。そのため、JIS X 0213 の運用状況を、実証的に確認することは、漢字情報処理にとっては意味のあることである。

第 3・第 4 水準漢字の使用は、異なり 105 字である。延べでは 1,735 字で、全体の 0.03%に過ぎず、微々たるもののように感じられる。しかし、JIS X 0213 を使うことで、「外字」が異なり 12 字（延べ 123 字）と、百分率では無視できるほど小さな値になっている。JIS X 0213 は「外字」問題の解消に一定の貢献を成し得るものと見なされる。

第 3 水準漢字の上位 10 文字は「頬（223）、嘘（204）、摑（124）、鷗（118）、噛（105）、剝（104）、呑（81）、軀（81）、痩（75）、掻（68）」であり、いずれも字典体（印刷標準字体）である。第 1 水準の対応する簡略字体の使用度数は「頬（17）、嘘（42）、掴（5）、鴎（0）、噛（14）、剥（9）、呑（88）、躯（1）、痩（5）、掻（3）」となっていて、「呑―呑」が拮抗している以外は、第 3 水準字が優勢である。今後、一般社会においては、字典体に対する需要を満たせることが、JIS X 0213 の最も大きな特長となろう。

４．JIS X 0213:2004「外字」

JIS 包摂規準を用いても符号化できない文字は、XML 形式の外字タグを用いて処理する。図 1 の例は次のようになる。

```
池中奇石<missingCharacter attribute="HanIdeograph"
unicode="U+7927" daikanwa="M24568" description="石
偏に田が三つ">＝</missingCharacter>[らい]々[らい]、
```

図 1：「外字」例

外字タグは、文字種名、Unicode、大漢和辞典番号などを属性として持つ。Unicode は、将来における Unicode への完全移行に備えるために採用した。また、「外字」には漢字が多いと予想されたので、大漢和辞典番号を記述することにした。

今回の対象サンプル中の「外字」を表 3 に示す。平仮名が 1 字（延べ 2 字）、漢字が 11 字（延べ 121 字）である。対象サンプルの複数にまたがって出現する「外字」はなく、どれも単一のサンプル中に使われた文字である。また、「外字」12 字（延べ 123 字）のうち、Unicode で表現できるものは 8 字（延べ 116 字）である。したがって、対象サンプル 4,649 字（延べ 5,511,743 字）のうち、Unicode で表現できないものは 4 字（延べ 7 字）となり、Unicode のカバー率は極めて高い。

表 3：「外字」一覧

字形	度数	Unicode	大漢和	用例区分	用例	出現サンプル
ぇ	2			感動詞	ぇーっ	LBl9_00045

	52	599E	6098	中国語	妞妞	LBt3_00046
	26	5AEA	6669	中国人名	嫽毒	LBl9_00229
	26	6BD0	16727	中国人名	嫽毒	LBl9_00229
	2	3B0B	14070	人名	羅睺	LBp9_00220
	7	6E3B	17795	中国人名	紀淆子	LBn8_00003
	1			翻刻例	「獦→獦」と	LBd9_00145
	1	7706	23184	人名	玄昉	LBb9_00052
	1	7752	23412	書名	睒子経＜せんじきょう＞	LBr9_00234
	3			神仏名	羅睺羅＜らごら＞	LBk1_00011
	1	7927	24568	漢語名詞	池中奇石礨々	LBd9_00145
	1			古代漢字	“曲想”の古い字形“”	LBm9_00253

５．むすびにかえて

記号も含めて文字の使用傾向は、文書の内容によって変動するものである。今回は書籍を対象としたが、雑誌では別な傾向が見られるだろう。今後は、外字タグを利用して、「外字」データベースへ発展させることを計画している。

参考文献

下田正弘・師茂樹（1999）「大正新脩大蔵経データベース（SAT）における外字問題」、『人文学と情報処理』25、pp.35-43

高田智和（2002）「漢字処理と『大字典』」、『訓点語と訓点資料』109、pp.99-107

田中牧郎（2005）「漢字の実態と処理の方法」、『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―（国立国語研究所報告 122）』、博文館新社、pp.271-292

富田倫生（2000）「青空文庫と外字」、『人文学と情報処理』26、pp.23-30

安永尚志（1998）『国文学研究とコンピュータ』、勉誠社

Ken Lunde（1999）*CJKV information processing*. Sebastopol : O'Reilly.

著作権処理の進捗状況と著作権法改正の動きについて

森本祥子　（データ班分担者：国立国語研究所情報資料部門）†
前川喜久雄　（総括班班長：　国立国語研究所研究開発部門）
小沼　悦　（データ班分担者：国立国語研究所研究開発部門）
新井田貴之　（データ班協力者：国立国語研究所管理部）
大石有香　（データ班協力者：国立国語研究所研究開発部門）
神野博子　（データ班協力者：国立国語研究所研究開発部門）
竹内ゆかり　（データ班協力者：国立国語研究所研究開発部門）
舞木　右　（データ班協力者：国立国語研究所研究開発部門）

Present State of Copyright Clearing Work and the Expected Revision of Copyright Law

Sachiko Morimoto　(Dept. Lang. Info. & Res., National Institute for Japanese Language)
Kikuo Maekawa　(Dept. Lang. Res., National Institute for Japanese Language)
Etsu Onuma　(Dept. Lang. Res., National Institute for Japanese Language)
Takayuki Niida　(Dept. Admin. Affairs., National Institute for Japanese Language)
Yuka Oishi　(Dept. Lang. Res., National Institute for Japanese Language)
Hiroko Kamino　(Dept. Lang. Res., National Institute for Japanese Language)
Yukari Takeuchi　(Dept. Lang. Res., National Institute for Japanese Language)
Yu Mogi　(Dept. Lang. Res., National Institute for Japanese Language)

１．　はじめに

『現代日本語書き言葉均衡コーパス』(以下，BCCWJ) 構築にあたっての著作権処理は，これまでに実質 2 年半の経験を蓄積してきた。本稿では，まずその作業進捗状況の報告を行い，次いでこの一年ほどで急速に起こった著作権に関わる動向を整理し，それをふまえた当コーパスの今後の著作権処理業務の展望を整理する。

２．BCCWJ における著作権処理の進捗状況

著作権処理は，サンプル一件ごとに著作権者を特定して許諾をとるというかたちで処理を進めているものと，著作権者から一括して使用許諾を得たもの（白書・国会会議録・ヤフー！知恵袋・新聞や雑誌の法人著作部分）がある。前者はそのほとんどが書籍であるが，その現在の処理状況は以下のとおりである。

表 1　書籍サンプル(生産サブコーパス・流通サブコーパス・ベストセラー)処理状況　（2009 年 2 月末）

必要サンプル数	25,212
許諾依頼済数	16,069
内　許諾数	9,317
拒否数	673

† morimoto@kokken.go.jp

すでに明らかになっているとおり，著作権者に連絡が取れ，回答が得られれば，ほとんどの場合に許諾が得られていることがわかる。

次に，処理作業の効率化の進展状況を確認する。

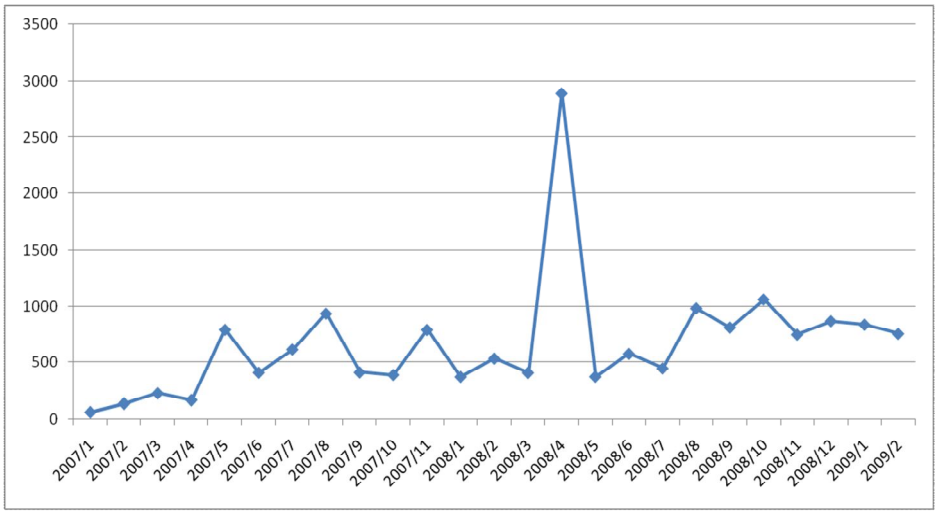


図 1　ひと月あたりの著作権処理サンプル数の推移

図 1 は，一ヶ月あたりの処理サンプル数の変化である。書籍の著作権処理業務に着手した 2007 年 1 月から表示している。折々にイレギュラーな処理業務が発生するため，毎月の処理量にはばらつきがあるが，それでも大きな流れを見たときに，2007 年中は平均すると一ヶ月に 500 件前後の処理だったものが，2008 年後半からは 1000 件弱を安定的に処理できていることがわかる。

このように処理数が増えた理由としては，担当職員が業務に熟練してきたということのほかに，以下の 2 点があげられる。

第一に，社会における著作権運用の実態を理解し，それに一定程度合わせた対応をとるようにしたこと。例えば出版業界の一部では，出版権と著作権とが混同されて運用されているが，このような場合，当初はあくまでも著作権法の原則にたって何とか出版社を除いて本来の著作権者と交渉すべく努力していた。しかし現在は出版社と執筆者との間の信頼関係を尊重し，出版社の判断のみで採録如何の判断をすることもある。これによって交渉の手間が軽減され，処理のスピードアップが図れている。

第二に，使用可否の判断を早めたこと。当初はあらゆる手段を講じて著作権者に連絡をとり「許諾率」を挙げることを目指したが，後述する著作権法改正の動きも視野に入れ，現在は努力の範囲を一定程度狭め，「連絡がとれない」という判断を早めにすることとした。

以上，著作権処理業務は当初想定した以上に，現状に合わせた柔軟な対応が求められるものとなっている。これは著作権というものの日本での定着の仕方の反映にほかならないが，著作物の多様な二次利用のニーズが高まるにつれ，既存のいわば曖昧な法理解では対応しきれなくなっていることも明らかになってきている。こうした状況を受けて，現在，

著作権法改正が相当な具体性をもって議論されている。次項でその動きについて整理する。

３．著作権法をめぐる最近の動き

３．１　著作権法改正の議論

現在，政府の知的財産戦略本部「デジタル・ネット時代における知財制度専門調査会」および文化審議会著作権分科会法制問題小委員会では，著作権法改正について議論を重ねており，2008 年秋に大きな方向性がほぼ固まったところである。

ここで議論されている主な論点は以下のとおりである。[1]

(1) デジタルコンテンツ流通促進法の整備
(2) 私的使用目的の複製の見直し
(3) リバース・エンジニアリングの法整備
(4) 研究開発における情報利用の円滑化
(5) 機器利用時・通信課程における蓄積等の合法化
(6) 包括的権利制限（日本版フェアユース）規定の可否

この中でとくに BCCWJ に関連するものは，(4)と(6)の問題である。特に(6)のフェアユースという発想は，日本の著作権法の基本的な考え方，すなわち，まずすべてを一律保護の下におき，個別の事例について権利者の権利を制限する，という方法とは異なるアプローチであるため，実現が求められつつも慎重な議論がなされている。このように制度の大きな転換を求められるものは直ちに実現することは難しいようだが，(4)については従来の枠組み内で権利制限のひとつに加えることが可能であり，じじつ次の法改正において取り入れられる予定ときく。このように，大量の著作物を利用する研究開発に伴う課題が著作権法改正の議論で取り上げられるまでには，様々な働きかけの蓄積があると思われるが，BCCWJ でも積極的に研究活動に伴う著作権処理の問題をアピールしており，その成果も少なからずあると考えている[2]。

こうした著作権法改正にむけた議論を踏まえ，BCCWJ では著作権処理の方針を一部変更することとした。具体的には，当初の「使用許諾を得られた著作物のみ利用する」という立場から，「使用を明確に拒否された著作物以外は利用する」という立場へと切り替えたのである。すなわち，最終的には一定の努力をしても著作権者と連絡がつかないものについては，公開対象に含めることにしたのである。現行の法制度下でも文化庁長官の裁定制度を利用すれば著作権者と連絡がとれないものについても利用可能だが，BCCWJ のように大量の著作権処理が必要な場合には制度の利用は現実的ではなかったため，当初は著作権者に連絡がとれないサンプルの利用は難しいと考えてきた。しかし仮に(4)の事項が実現すれば，おそらく BCCWJ では著作権処理は不要になり，少なくとも著作権者から使用を拒否されていない文章を使うことは合法化されるはずである。こうした見通しとともに，「許諾を得る」ための努力を軽減し，権利者に一通り連絡をとることを目標に据えることにした。

なお，現段階ではこうした展望を持ちつつも，原則として著作権者に理解を得てサンプルを採録したいと考えていることに変わりはなく，文化庁長官裁定で求められる水準の努

力は続けている[3]。また仮に著作権者と連絡がとれないままにサンプルを公開することになっても，公開後に著作権者から指摘があった場合には，改めて著作権処理を行う予定である。

３．２　著作権者側の主張と提案

一方，全体として著作権の権利制限につながるような法改正やその背景にある社会状況への危機感から，権利者側でも議論が活発になってきている。以前から保護期間の延長（現行の死後 50 年を 70 年に延長する）は権利者によって主張されてきているが，著作権処理にまつわる様々な困難が指摘されている現在，権利者サイドでも保護強化の主張だけでは通らないとの認識はある。そうした中で，日本文藝家協会・日本音楽著作権協会など主要な著作権者団体の連合体である「著作権問題を考える創作者団体協議会」が，著作物の利用促進のため著作権者への連絡をとりやすくするシステムを立ち上げることとなった。すでに「著作者検索データベース」として 2009 年 1 月より運用が開始されている[4]。こうした動きは，当然ながらわれわれのような著作権利用者の立場からも歓迎されることであり，当該データベースの拡充を始め，種々の手段が充実して著作権処理がしやすくなることを強く期待している。

４．おわりに

ある著作権法の専門家は，「著作権とは裁判を重ねて社会で共通認識を作り上げていくものだ」と主張するが，これは BCCWJ の著作権処理にあたって示唆に富む考え方である。われわれは当初，現行の著作権法が絶対であり，公的機関としては法を遵守して著作物を使用すべきであると考えた。そのため，コーパスで著作物を利用するというような新しい利用方法であっても，何とか現行法の枠組みに位置づけ，その中で正しい対応をしようと努力してきた。しかし，コーパス構築のための著作物の利用ということが著作権法の想定外のことであるように，著作物の利用方法はどんどん変わっていく。BCCWJ の著作権処理業務には，現行法を遵守する姿勢と同時に，新たな利用方法があることをきちんと世に訴え，それが合法的に行えるよう法の改正を求めていくという役割もあるのではないか。BCCWJ 構築プロジェクトは残すところ 2 年となったが，その間に大規模研究開発に伴う大量著作権処理の唯一の経験蓄積を持つ者として，著作権法とその制度に提言していくことも積極的に行うべきだと考えている。

---

1 「文化審議会著作権分科会法制問題小委員会　平成 20 年度・中間まとめ」（平成 20 年 10 月 1 日）を元に，(1)～(6)の課題に整理した。詳細は同「まとめ」を参照のこと。

2 例えば，前川喜久雄は「文化審議会著作権分科会法制問題小委員会の平成 20 年度中間まとめ」のパブリックコメント募集に意見を寄せるなどして積極的に発言してきた。そうした蓄積を受け，前川は文化審議会著作権分科会でのヒヤリングに招かれたりしている。

3 著作権者の連絡先が不明なため著作権処理ができないサンプルの情報を公開し，権利者によびかけるウェブサイト「KOTONOHA 著作権者検索データベース」を構築し，運用している。アクセス先は，http://www.kotonoha.gr.jp/chosakuken。

4 http://www.sousakusya.jp/　（アクセス日：2009 年 2 月 15 日）

『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

小椋秀樹　（データ班分担者：国立国語研究所研究開発部門）*
小木曽智信（電子化辞書班分担者：国立国語研究所研究開発部門）
小磯花絵　（電子化辞書班分担者：国立国語研究所研究開発部門）
冨士池優美（データ班連携研究者：国立国語研究所研究開発部門）
宮内佐夜香（データ班協力者：国立国語研究所研究開発部門）
渡部涼子　（電子化辞書班協力者：国立国語研究所研究開発部門）
竹内ゆかり（データ班協力者：国立国語研究所研究開発部門）
小川志乃　（データ班協力者：国立国語研究所研究開発部門）
小西　光　（電子化辞書班協力者：国立国語研究所研究開発部門）
原　　裕　（データ班協力者：国立国語研究所研究開発部門）
中村壮範　（データ班協力者：マンパワー・ジャパン株式会社）

Progress Report on Morphological Analysis of the Balanced Corpus of Contemporary Written Japanese

Hideki Ogura　(Dept. Lang. Res., National Institute for Japanese Language)
Toshinobu Ogiso　(Dept. Lang. Res., National Institute for Japanese Language)
Hanae Koiso　(Dept. Lang. Res., National Institute for Japanese Language)
Yumi Fujiike　(Dept. Lang. Res., National Institute for Japanese Language)
Sayaka Miyauchi　(Dept. Lang. Res., National Institute for Japanese Language)
Ryoko Watanabe　(Dept. Lang. Res., National Institute for Japanese Language)
Yukari Takeuchi　(Dept. Lang. Res., National Institute for Japanese Language)
Shino Ogawa　(Dept. Lang. Res., National Institute for Japanese Language)
Hikari Konishi　(Dept. Lang. Res., National Institute for Japanese Language)
Yutaka Hara　(Dept. Lang. Res., National Institute for Japanese Language)
Takenori Nakamura　(Manpower Japan Co., Ltd.)

１．はじめに

国立国語研究所が中心となって構築を進めている『現代日本語書き言葉均衡コーパス』（以下，BCCWJ とする。）には，様々な研究用の情報を付与する(1)。本稿で取り上げる形態論情報もその中の一つである。BCCWJ では，言語単位として長単位・短単位の 2 種類の言語単位を採用し，それぞれに代表形・代表表記・品詞・語種等の情報を付与する。

BCCWJ への形態論情報付与作業は，国立国語研究所研究開発部門言語資源グループに所属する本稿の著者 11 名（以下，形態論情報サブグループとする。）が中心となって進めている。形態論情報サブグループのメンバーは，特定領域研究「日本語コーパス」においてはデータ班・電子化辞書班のいずれかに所属しており，これによって，データ班と電子化辞書班との密接な連携に基づく作業の遂行を実現している。

＊ ogura@kokken.go.jp
（１）BCCWJ の概要等については，前川(2008)，山崎(2007)を参照。

BCCWJ で採用した長短 2 種類の言語単位の認定基準等は，既に小椋ほか(2007)・冨士池ほか(2008)で紹介したところである。これまで，これらの基準等に基づき形態論情報の付与作業を進めてきた。本稿では，その進捗状況について報告する。

２．形態論情報サブグループの任務

形態論情報サブグループの任務は，特定領域研究「日本語コーパス」で構築する書籍コーパスを含む BCCWJ 1 億語を長短 2 種類の言語単位に解析し，品詞・語種等の情報を付与することである。この形態論情報の目標精度は，長単位・短単位とも境界，品詞，代表形・代表表記の認定まで含めて 98%以上と定められている。

形態論情報サブグループは，研究計画の最終年度である 2010 年度末までに，この目標を達成するため，以下の二つを中心に作業を進めている。

(1) 形態素解析用辞書 UniDic の整備・拡充
(2) 形態素解析システム等の学習用データであるコアデータの作成

以下，第 3 節で形態素解析用辞書 UniDic の整備・拡充作業の進捗状況，第 4 節でコアデータ作成作業の進捗状況について報告する。また第 5 節で，形態論情報に関する基準の整備についても簡単に述べる。

３．UniDicの整備・拡充

３．１　未登録語の新規追加作業

BCCWJ は，1 億語から成る大規模なコーパスであるため，形態論情報の付与は自動解析システムにより行う。短単位解析には解析エンジン MeCab と解析用辞書 UniDic(2)を，長単位解析には短単位解析結果から長単位を自動構成する解析器（Uchimoto & Isahara, 2007）を使う。

長短 2 種類の言語単位のうち短単位は，長単位の基になるため，その解析精度は長単位の解析精度に大きな影響を与える。したがって，長単位・短単位とも高精度のデータとするためには，短単位の自動解析精度をできる限り向上させることが必須となる。そこで，形態論情報サブグループでは，短単位解析精度の向上に最も深くかかわる解析用辞書 UniDic の整備・拡充を最重要課題と位置付け，本研究計画開始の 2006 年度から継続して作業を行っている(3)。

整備・拡充作業の中心は，辞書未登録語の新規追加作業である。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登録している。この階層の最上位を語彙素（代表形・代表表記に相当）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。このような階層構造を持っているため，未登録語の新規追加といっても，①語彙素以下すべてを新規追加する場合，②語形以下を新規追加する場合，③書字形のみを新規追加する場合の 3 種類がある。

未登録語の新規追加作業は，2006 年度に非母集団サブコーパス（以下，サブコーパスを SC と略す。）の白書データから着手し，それ以降，生産実態 SC 新聞・書籍，流通実態 SC 書籍，非母集団 SC ベストセラーと順次対象を広げてきた。2008 年度からは非母集団 SC の Web（Yahoo!知恵袋）データからの新規追加も開始した。

（２）UniDic の概要については，伝ほか(2007)を参照。
（３）UniDic の整備・拡充作業で使用しているツール等については，小木曽ほか(2009)を参照。

未登録語の新規追加作業では，MeCab で「未知語」と解析されたものに加えて，未登録語に起因する誤解析である可能性が高い短単位連続を人手でチェックし，辞書に見出し語の追加を行っている。以下，例を挙げて説明する。

短単位は，原則として現代語において意味を持つ最小の単位（最小単位）二つの 1 回結合までを 1 短単位とする(4)。次に和語の短単位の例を挙げる。(「｜」が短単位境界，「/」が最小単位境界。)

｜雨｜　｜大/雨｜　｜食べる｜　｜食べ/歩く｜　｜白い｜　｜青/白い｜

「食べ歩く」のように，動詞（最小単位）二つから成る複合動詞は 1 短単位となるが，構成要素の「食べる」「歩く」が辞書に登録されていても，複合動詞「食べ歩く」が辞書に登録されていない場合，2 単位に分割されてしまう。このような誤解析を発見し，複合動詞の新規追加を行うために，1 最小単位から成る動詞二つの短単位連続を解析結果から抽出し，人手によるチェックを行っている。次に，実際の誤解析例を挙げる。

｜イネ｜や｜草｜を｜食い｜荒らし｜，｜
｜交互｜に｜たち｜現れる｜こと｜も｜あれ｜ば｜

「食い｜荒らし」は語彙素「食い荒らす」が登録されていなかったこと，「たち｜現れる」は語彙素「立ち現れる」の書字形に「たち現れる」が登録されていなかったことによる誤解析である。この場合，語彙素「食い荒らす」と書字形「たち現れる」を UniDic に登録することになる。このほか，1 文字の和語普通名詞の短単位連続などについても，上記と同様のチェックを行い，未登録語の新規追加を行っている。

直近 6 か月の語彙素・語形・書字形の登録数の推移を**図1**に示す。

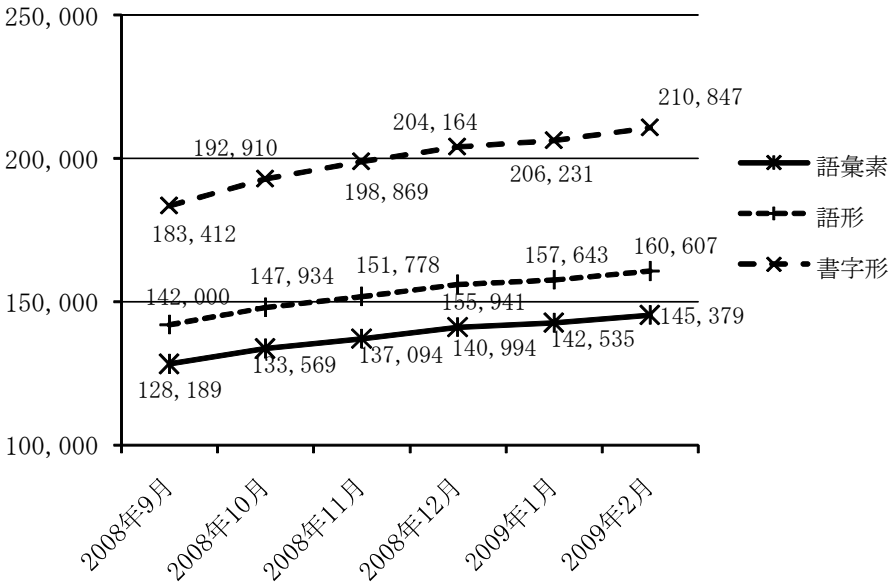


図1 語彙素・語形・書字形の登録数の推移

後に述べるコアデータの作成等，未登録語の新規追加作業以外にも並行して行っている

（4）最新の短単位の認定基準については，小椋ほか(印刷中)を参照。

作業があるため，月によって増加数に違いはあるが，毎月平均で語彙素は約 3,400，語形は約 3,700，書字形は約 5,500 増加している。2009 年 2 月 20 日現在の語彙素数は 145,379 語で，2006 年度初め（語彙素数約 46,000 語）の約 3 倍となっている。

３．２　その他の整備・拡充作業

UniDic の整備・拡充作業では，未登録語の新規追加以外に，品詞以外の情報を新たに付与したり，単語登録の方法や品詞情報の枠組み（品詞体系）等を変更したりしている。これまで行った主な変更等として，①語種情報の新規付与(5)，②人名の語彙素の統合，③活用形のうち基本形を終止形と連体形とに分割する，④「名詞−固有名詞−組織名」の廃止が挙げられる。これらのうち，以下，人名の語彙素の統合と「名詞−固有名詞−組織名」の廃止について報告する。

（１）人名の語彙素の統合

当初，人名（「名詞-固有名詞-人名-姓」「名詞-固有名詞-人名-名」「名詞-固有名詞-人名-一般」）は，同じ語形であっても表記が異なれば異なる語彙素として UniDic に登録していた。ただし「相沢」と「相澤」のように異体字の関係ととらえられる異表記は，一つの語彙素にまとめていた。この登録の方法を示したのが，**表1**の左側である。このような登録方法を取った結果，平仮名表記や片仮名表記された人名についても，それぞれ別に語彙素を立てることになり，語彙素数が非常に多くなるといった弊害が生じた。この弊害を解消するため，2007 年度に，同じ語形であれば一つの語彙素にまとめる形に登録方法を変更した。具体的には，「相澤」「相沢」「藍沢」を「アイザワ」という語彙素（語彙素は片仮名表記とする。）にまとめるというものであり，表 1 の右側のような登録方法となる。

表1 人名の登録方法の変更

語彙素	語　形	書字形
相沢	アイザワ	相沢
		相澤
藍沢	アイザワ	藍沢
アイザワ	アイザワ	アイザワ

⇒

語彙素	語　形	書字形
アイザワ	アイザワ	相沢
		相澤
		藍沢
		アイザワ

語彙素	語　形	書字形
正夫	マサオ	正夫
正雄	マサオ	正雄
雅男	マサオ	雅男
まさお	マサオ	まさお

⇒

語彙素	語　形	書字形
マサオ	マサオ	正夫
		正雄
		雅男
		まさお

（５）UniDic に付与した語種情報の概要については，小椋ほか(2008)を参照。

なお，地名（「名詞-固有名詞-地名-一般」）については，現在も，同じ語形であっても表記が異なれば，異なる語彙素として扱っている。その結果，人名と同じように語彙素数が非常に多くなるといった弊害が生じてきている。そこで，今後，人名の語彙素統合と同様の方法で「名詞-固有名詞-地名-一般」の語彙素の統合を行うことにしている。

（２）「名詞-固有名詞-組織名」の廃止

UniDic では，「ソニー」「日産」といった会社名等の品詞を「名詞-固有名詞-組織名」とし，元号やペットの名などを「名詞-固有名詞-一般」としていた。しかし 2008 年度に，この「名詞-固有名詞-組織名」を廃止して「名詞-固有名詞-一般」に統合するという品詞体系の変更を行った。その主な理由は，以下のとおりである。

1. 長単位で組織名であっても，短単位では組織名とならない場合が多い。
   例えば，「大阪大学」「明治大学」「立命館大学」は組織の名称であるが，短単位では「｜大阪｜大学｜」「｜明治｜大学｜」「｜立命｜館｜大学｜」と複数の単位に分割し，「大阪」には地名，「明治」には固有名詞（一般），「立命」「大学」には普通名詞，「館」には接尾辞という品詞を付与する。したがって，これらの例では，「名詞-固有名詞-組織名」という品詞は付与されない。
2. 組織名などの「固有表現」をテキストから抽出する技術が別途開発されている。

「名詞-固有名詞-組織名」の廃止は，現在のところ短単位の品詞体系についてのみであり，長単位の品詞体系では廃止していない。ただし，長単位の品詞体系における「名詞-固有名詞-組織名」の必要性についても慎重に検討し，最終的に必要性がないと判断されれば長単位の品詞体系でも「名詞-固有名詞-一般」に統合する予定である。

３．３　UniDicの解析精度

2008 年 7 月に UniDic-1.3.9 を公開して以降，MeCab 用の UniDic について，解析精度の向上を目指し，学習素性の再検討等を継続して行っている。現在の UniDic(1.3.11 β)の精度と UniDic-1.3.9 の精度とを**表2**に示す。

表 2 に示したのは，語彙素認定の精度（アウトサイドデータに対する F 値）である。学習用データには白書・新聞・書籍・Web（Yahoo!知恵袋）のコアデータ（第 4 節参照）のほか CSJ 等を利用している。今後更に検討を進め，解析精度の向上を図っていく。

表2 UniDicの解析精度

	白書	書籍	新聞	Web (Yahoo!知恵袋)	CSJ
UniDic-1.3.9	99.38%	98.43%	―	―	97.86%
UniDic-1.3.11 β	99.38%	98.40%	98.65%	97.89%	97.98%

※ UniDic-1.3.9 作成時点では，新聞・Web（Yahoo!知恵袋）のコアデータは完成していなかったため，これら二つの媒体に関する UniDic-1.3.9 の精度データはない。

４．コアデータの作成

４．１　コアデータの設計

2006 年度の本研究計画開始当初より，形態論情報サブグループでは形態素解析システムの学習用データを作成することとしていた。その後，ツール班でも係り受け情報等のタグ付けツールの学習用データを作成するという計画が出されたため，形態素解析システムの学習用データをツール班の学習用データとして共有することを提案し，更にツール班の要望も取り入れて**表3**のような構成の学習用データ（以下，コアデータと呼ぶ。）を作成

することとした。

なお，Web データは元々 Yahoo!知恵袋のみを採録する予定であったが，これについても，2008 年度にツール班から要望が出されたのを受け，Yahoo!知恵袋の延べ語数を 20 万語から 10 万語に変更した上で，Yahoo!ブログ 10 万語を加える形に設計を変更した。

コアデータに付与する情報のうち，形態論情報サブグループが付与する情報は，長単位・短単位の境界・代表形・代表表記・品詞・語種のほか，文節境界である。このうち文節境界については，当初，ツール班が情報付与を担当することになっていたが，作業人員の不足等の問題があり，2008 年度より形態論情報サブグループが担当することにした。以上の情報を基にツール班が係り受け情報等を付与することになっている。

コアデータは，形態素解析システム，タグ付けツールの学習用データとして使用するため，形態論情報をより高精度にすることが求められる。文節・長単位・短単位については自動解析後に人手修正を行うことによって，精度を 99%以上とすることを目標としている(6)。

表3 コアデータの構成

コーパス種別	ジャンル		延べ語数	サンプル種別
生産実態SC	新　聞		20万語	固定長
	雑　誌		20万語	可変長
	書　籍		20万語	可変長
非母集団SC	白　書		20万語	可変長
	Web	Yahoo!知恵袋	10万語	可変長
		Yahoo!ブログ	10万語	可変長

※ 新聞の可変長サンプルは固定長サンプルよりも短いものが多いため，
新聞のみ固定長サンプルを対象とすることとした。

４．２　進捗状況

（１）短単位情報

2008 年 10 月までに白書・新聞・書籍の合計延べ 60 万短単位のコアデータを作成し，特定領域内に公開している。このうち白書・書籍については，2008 年度のモニター公開で広く一般にも公開している。また，Web（Yahoo!知恵袋）延べ 10 万短単位も既にデータが完成しており，年度末には特定領域内に公開する予定である。

（２）長単位情報

2007 年度から白書を対象に長単位情報の付与作業に着手した。2008 年度はコアデータのうち，白書・新聞・書籍について，それぞれ延べ約 10 短単位分に当たるデータに長単位境界・品詞・代表形・代表表記の情報を付与している。長単位での延べ語数は，白書 81,821，書籍 120,923，新聞 76,239 となっている。これらについては，2008 年度末までに作業を終え，長単位解析ツールの学習用データとして電子化辞書班に提供する。

（３）文節情報

文節境界情報の付与は，外注により進めている。既に書籍データに対する付与作業が完了しており，サンプリングチェックによって 99%以上の精度であることを確認している。

（6）コアデータの形態論情報の人手修正作業で使用しているツール等については，小木曽ほか(2009)を参照。

2008 年度末までに更に白書・新聞への付与作業を完了し，白書・新聞・書籍の延べ約 60 万短単位分のデータをツール班・電子化辞書班に提供する予定である。Web（Yahoo!知恵袋・Yahoo!ブログ）・雑誌についても，2009 年度以降，外注により付与作業を行っていく予定である。

５．形態論情報に関する基準の整備

BCCWJ の形態論情報付与作業には，本稿の著者 11 名に加えて，アルバイター 6 名の合計 17 名が従事している。このような大人数で，効率的に高精度のデータ作成を進めるためには，単位の認定基準や品詞の判定基準等が，作業者全員に確実に理解され，共有されることが不可欠である。

そこで，UniDic の整備・拡充作業，コアデータの作成作業を行う過程で，文節・長単位・短単位の認定基準，品詞の判定基準をはじめとする種々の基準を作成するとともに，必要に応じて改定を行っている。これらの基準は作成・改定の都度，作業者に周知するとともに，作業者が参照しやすいよう，毎年度末に国立国語研究所内部報告書等として冊子にまとめている。この冊子は，申請があれば，BCCWJ の関係者以外にも頒布を行ってきたところである(7)。なお，2008 年度末にも最新版（小椋ほか印刷中）を刊行する。

６．終わりに

以上，BCCWJ に対する形態論情報付与作業のうち，形態素解析用辞書 UniDic の整備・拡充作業とコアデータの作成作業の進捗状況について報告した。第 2 節に示したように BCCWJ の形態論情報の目標精度は 98%以上（コアデータは 99%以上）と定められている。コアデータについては，全体を人手でチェックし，修正を加えていくため，目標精度の達成は可能であるという見通しを持っている。コアデータ以外については，残り 2 年間で目標を達成するため，引き続き未登録語の新規追加作業を行っていく。また，ツールによる再解析(8)やデータの一部について人手チェック及び修正等を行い，最終的に 98%以上の精度を達成したいと考えている。

参 考 文 献

小木曽智信・小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・渡部涼子・竹内ゆかり・小川志乃・小西光・原裕・中村壮範(2009)「形態論情報データベースの構成」『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』

小椋秀樹・小木曽智信・小磯花絵・冨士池優美・相馬さつき(2007)「「現代日本語書き言葉均衡コーパス」の短単位解析について」『言語処理学会第 13 回年次大会発表論文集』, pp.720-723.

小椋秀樹・小木曽智信・原裕・小磯花絵・冨士池優美(2008)「形態素解析用辞書 UniDic への語種情報の実装と政府刊行白書の語種比率の分析」『言語処理学会第 14 回年次大会発表論文集』, pp.935-938.

小椋秀樹・小磯花絵・冨士池優美・原裕(印刷中)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 改定版』(LR-CCG-08-03)

（７） 国立国語研究所内部報告書の頒布については，国立国語研究所 KOTONOHA 計画のホームページ http://www.kokken.go.jp/kotonoha/ex_7.html を参照。

（８） 自動形態素解析で誤解析になることの多い助詞・助動詞に対する再解析については，中村・伝(2008)を参照。

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』21, pp.101-123.
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.
冨士池優美・小椋秀樹・小木曽智信・小磯花絵・相馬さつき・中村壮範(2008)「「現代日本語書き言葉均衡コーパス」の長単位認定基準について」『言語処理学会第 14 回年次大会発表論文集』, pp.931-934.
前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」『日本語の研究』4-1, pp.82-95.
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』, pp127-136.
Uchimoto, K., & Isahara, H.（2007）. Morphological annotation of a large spontaneous speech corpus in Japanese, *Proceedings of IJCAI*, pp.1731-1737.

付記　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。

形態論情報データベースの構成

小木曽 智信（電子化辞書班分担者：国立国語研究所研究開発部門）†
小椋 秀樹（データ班分担者：国立国語研究所研究開発部門）
小磯 花絵（電子化辞書班分担者：国立国語研究所研究開発部門）
冨士池 優美（データ班連携研究者：国立国語研究所研究開発部門）
宮内 佐夜香（データ班協力者：国立国語研究所研究開発部門）
渡部 涼子（電子化辞書班協力者：国立国語研究所研究開発部門）
竹内 ゆかり（データ班協力者：国立国語研究所研究開発部門）
小川 志乃（データ班協力者：国立国語研究所研究開発部門）
小西 光（電子化辞書班協力者：国立国語研究所研究開発部門）
原 裕（データ班協力者：国立国語研究所研究開発部門）
中村 壮範（データ班協力者：マンパワージャパン株式会社）

Construction of Database for Morphologically Analyzed Corpus

Toshinobu Ogiso (Dept. Lang. Res., National Institute for Japanese Language)
Hideki Ogura (Dept. Lang. Res., National Institute for Japanese Language)
Hanae Koiso (Dept. Lang. Res., National Institute for Japanese Language)
Yumi Fujiike (Dept. Lang. Res., National Institute for Japanese Language)
Sayaka Miyauchi (Dept. Lang. Res., National Institute for Japanese Language)
Ryoko Watanabe (Dept. Lang. Res., National Institute for Japanese Language)
Yukari Takeuchi (Dept. Lang. Res., National Institute for Japanese Language)
Shino Ogawa (Dept. Lang. Res., National Institute for Japanese Language)
Hikari Konisi (Dept. Lang. Res., National Institute for Japanese Language)
Yutaka Hara (Dept. Lang. Res., National Institute for Japanese Language)
Takenori Nakamura (Manpower Japan Co., Ltd.)

はじめに

特定領域「日本語コーパス」データ班・電子化辞書班では、『現代日本語書き言葉均衡コーパス』（BCCWJ）の形態論情報付与作業を行うために、コーパスと辞書データを格納し、両者の連係をとりながら修正作業を行うためのデータベースシステム（形態論情報データベース）を構築している。本発表ではこのデータベースの全体の構成について報告する。なお、このデータベースの詳細については、小木曽・中村（印刷中）『『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』にまとめられている。

１．形態論情報データベースの概要

形態論情報データベースは、形態素解析辞書 UniDic の見出し語を格納した「辞書データベース」と、BCCWJ サンプルに形態素解析を行った結果を格納した「コーパスデータベース」からなる。両者は中間に語彙表（辞書の見出し語を活用形まで展開した表）を挟んで連係しており、コーパスデータベース中の全ての語が、原則として辞書データベースの見出し語と関連づけられており、辞書見出しの修正はコーパスに反映されるようになってい

† togiso@kokken.go.jp

る（図　1）。これにより、コーパス側からは出現した語の辞書上の全情報が、辞書側からは見出し語のコーパス中の頻度などの情報が取得できる。

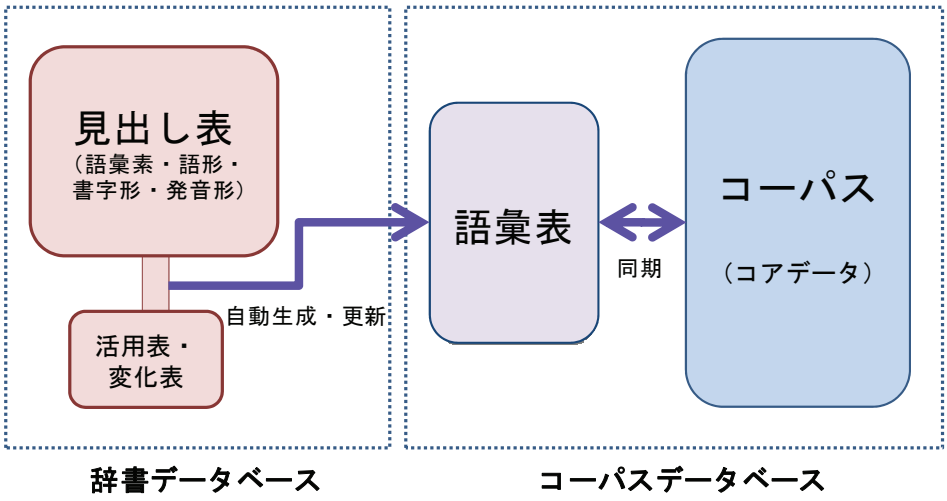


図　1　形態論情報データベース全体図

なお、データベース管理システムには Microsoft SQL Server 2005 を採用し、作業用アプリケーションは Microsoft Access で開発している。

2．辞書データベース

辞書データベースは、形態素解析辞書 UniDic の元となる見出し語のデータベースである。見出し語のテーブルのほか、活用表などの辞書作成に必要な情報からなる。

辞書データベースの基本となる見出し語表は、「短単位語彙素」、「短単位語形」「短単位書字形」「短単位発音形」の４つからなる。UniDic の見出し語階層をそのままデータベースの構造に反映させている。

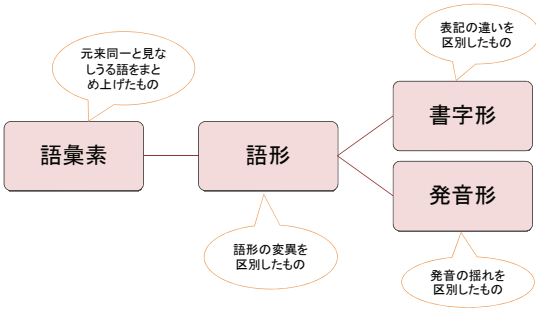


図　2　UniDic の見出し語階層

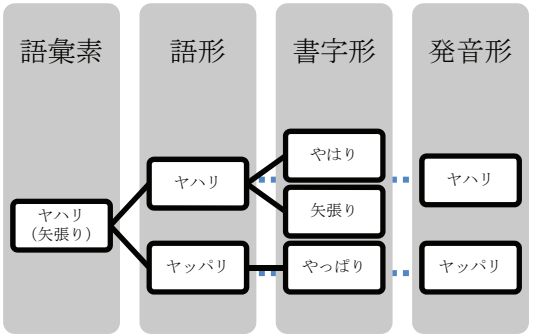


図　3　見出し語の例

現在、形態論情報データベースには、語彙素　約 14.5 万語、語形　約 16 万語、書字形　約 21 万語、発音形　約 16 万語が格納されている。

データベースでは、書字形の下にさらに書字形構成漢字表が関連づけられる。この表は、書字形が漢字を含んでいる場合に、その漢字がどのように読まれているかという情報を持つ。書字形構成漢字表とコーパスを結びつけることにより、コーパス中の漢字の音訓別頻

度表を作成することができる。また、単漢字の情報を含む漢字表と結合することにより、常用漢字や教育漢字の音訓がコーパス中の漢字の読みをどれだけ網羅しているかといった情報も得られる。

見出し語の追加・修正作業には、見出し語表の階層をそのまま表示し、修正が可能な辞書管理ツール UniDic Explorer を利用している（図 4）。

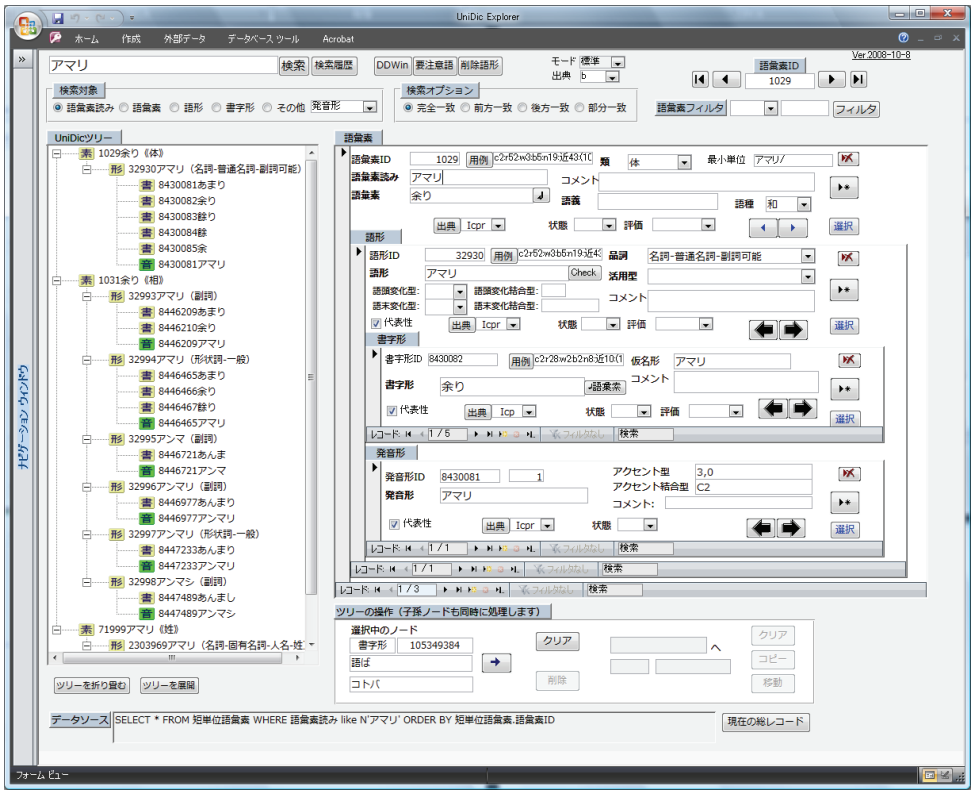


図 4　辞書管理ツール UniDic Explorer

3．語彙表

辞書データベースとコーパスデータベースをつなぐ語彙表は、4 階層の見出し語と変化表・活用表を組み合わせて生成される。この際、語形のレベルでは語頭・語末変化によって複数の変化形に展開される。語頭変化とは、連濁のような語頭で起きる語形変化、語末変化とは「一（イチ）」が「一（イッ）（本（ポン））」に変化する促音化などの語形変化を指す。さらに、活用語では活用変化によって各活用形が展開される（図 5）。

図 5　語彙表生成の流れ

たとえば、「カライ（辛い）」という語形は、濁音化を起こすことが辞書に記載されているため、基本形「カライ」と濁音形「ガライ」が展開される。さらに、「カライ」は活用語であるから表 1 に示す形容詞の各活用形が展開される。語形の下にある書字形・発音形についても全ての活用形が展開される。図 6 にこの展開の様子の一部を示した。

表 1　活用形の例（形容詞「辛(から)い」）

活用形	語形（活用後）
意志推量形	カラカロウ， カラカロ
仮定形-一般	カラケレ
仮定形-融合	カラキャ， カラケリャ
語幹-一般	カラ
終止形-一般	カライ， カレエ
終止形-促音便	カラッ
連体形-一般	カライ， カレエ
連用形-ウ音便	カロウ
連用形-一般	カラク
連用形-促音便	カラカッ

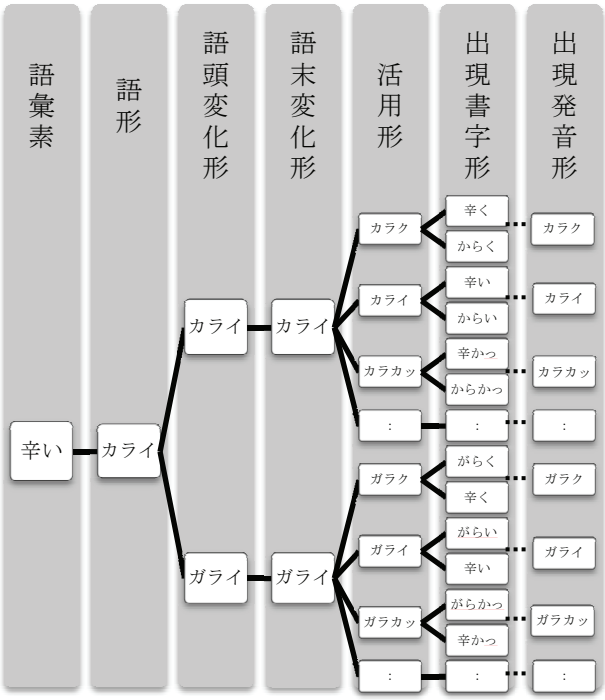


図 6　語彙表生成の例（「辛(から)い」）

4．コーパスデータベース

コーパスデータベースは、BCCWJ サンプルなどの文章データに形態素解析を施した結果を取り込んだものである。形態素解析はデータベースとは別に行い、テキスト形式の解析結果をインポートする。BCCWJ のデータは XML で記述されている。コーパスデータベースでは、この情報を関係データベースの一般的な表で表現するために、「文字表」「短単位表」「文字修正表」「数字タグ表」「ルビ表」「タグ表」の各表に分けて取り込んでいる。形態論情報の処理に直接関連するタグのみ専用テーブルに書き込み、その他のタグは一括してタグ表で保管する。いずれのテーブルもサンプル ID と原文における文字位置をキーとして関連づけられている（図 7）。

コーパスデータベース上での修正は専用ツール「大納言」によって行う。「大納言」では、文脈を参照しながら、辞書データベースと関連づけて形態素解析結果を修正できるほか、数字変換処理結果の修正、原文文字の修正等が行える。また、コーパスに対して文節や「長単位」情報[1]の付与・修正が行えるようになっている。こうした作業に対応するために高度な検索機能を備えており、短単位での検索のほかに、単位境界を越えた全文検索、サンプル単位での検索、短単位の情報を組み合わせた高度な検索が可能になっている（図 8）。

「大納言」を使って人手で修正されたデータ（コアデータ）は、もとの XML 文書に形態素タグを埋め込んだ XML 形式でエクスポートすることができる。各テーブルを SQL で結合し、データベース内部で XML 型のデータとして生成した後、ファイル出力する。

[1] 長単位、及び短単位と長単位の関係については小椋ほか（近刊）参照。

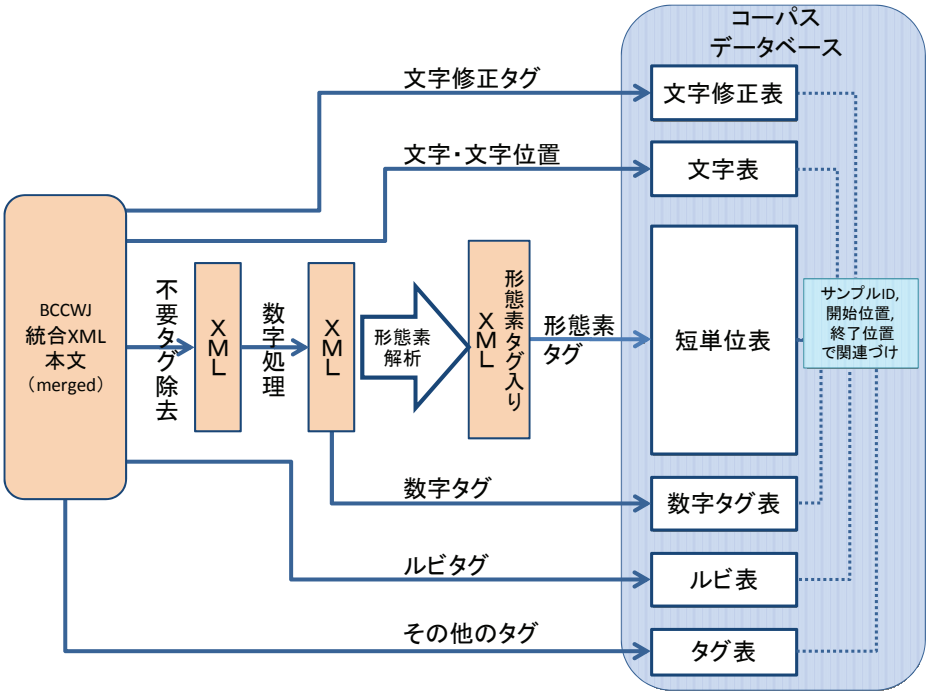


図　7　コーパスデータベースの構造と BCCWJ サンプルの解析・取り込み手順

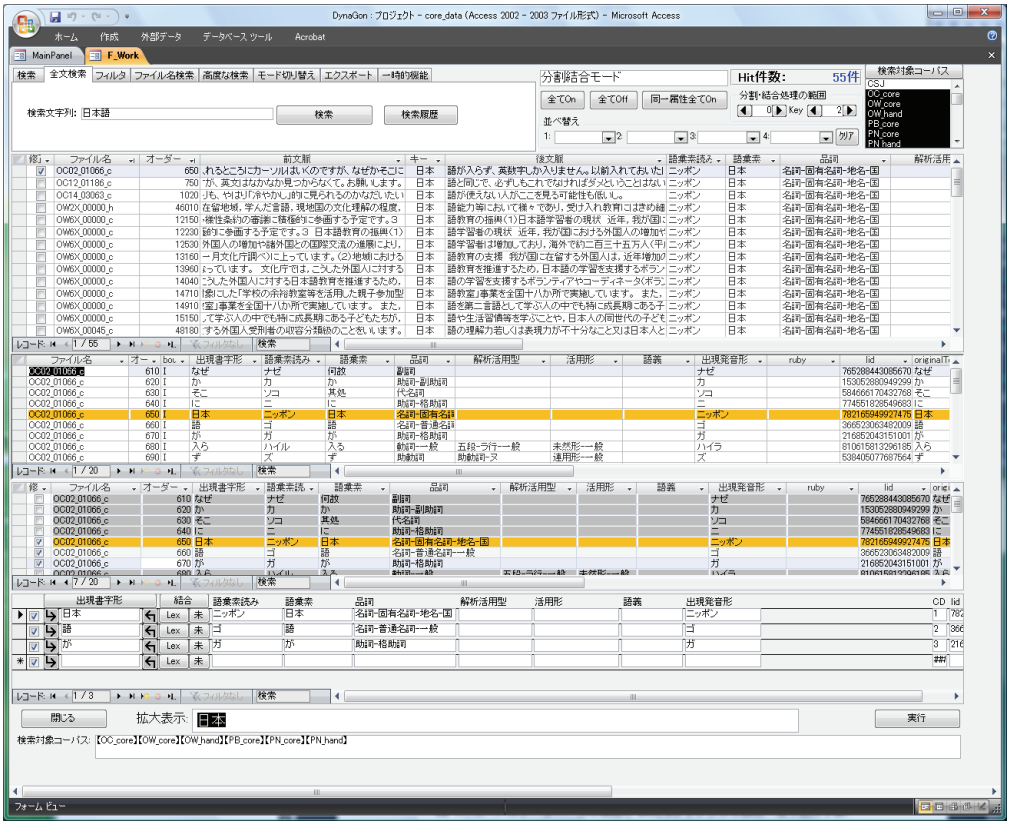


図　8　コーパス修正ツール 大納言

コーパス修正ツールの検索機能は、そのまま研究に利用することができる。とはいえ、専用のアプリケーションを必要とするうえ、データベースへの接続にも問題も生じる。そ

こで、Web 上でコーパス検索を可能にする「中納言」を開発している（図 9）。「中納言」は「大納言」の検索インターフェイスを Web 用に作り直したもので、ブラウザ上の検索や結果のエクスポートが可能となっている。公開用に別サーバで稼働しているが、「中納言」が接続するデータベースの構造は、コーパスデータベースの構造と同一である。「中納言」は下記の URL で公開予定である。

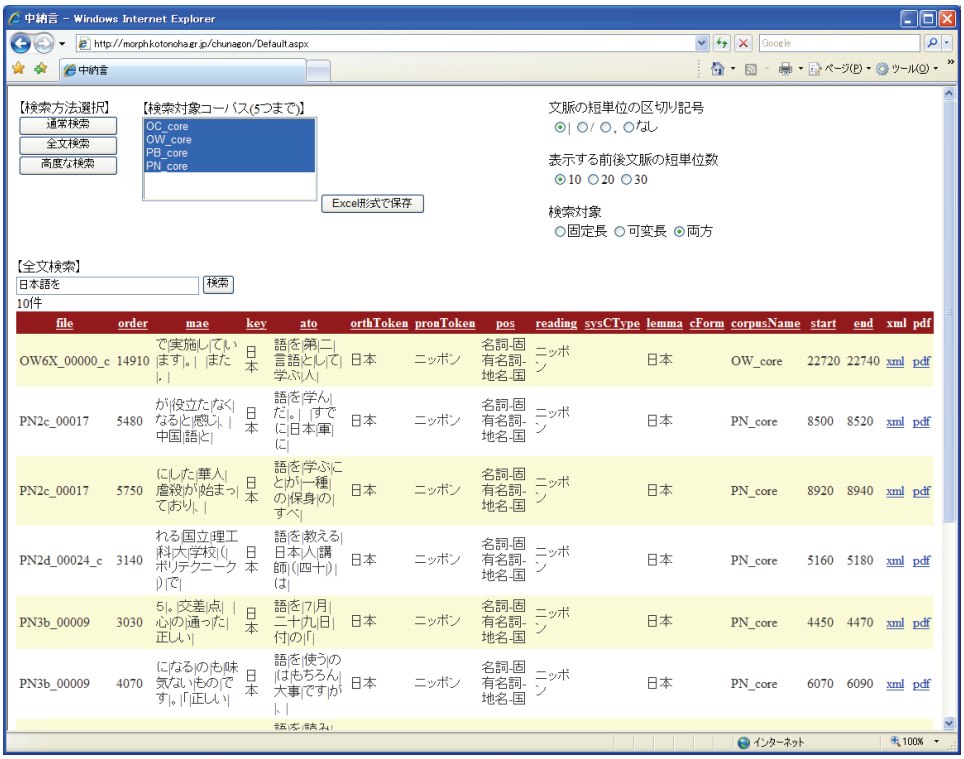


図 9　Web 版コーパス検索ツール 中納言

おわりに

以上、BCCWJ の形態論情報付与に用いているデータベース全体の構成について述べた。システムの詳細は小木曽・中村（印刷中）を参照されたい。また、UniDic の基本設計については伝（2007）を、データの言語単位に関する仕様については小椋ほか（印刷中）をそれぞれ参照されたい。

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 号 pp.101-122.

小木曽智信・中村壮範（印刷中）『『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』

小椋秀樹・小磯花絵・冨士池優美・原裕（印刷中）『『現代日本語書き言葉均衡コーパス』形態論情報規程集 改定版』

関連 URL

Web 版コーパス検索ツール：http://morph.kotonoha.gr.jp/chunagon/

拡張固有表現タグ付きコーパスの構築に向けて
- 白書，書籍，Yahoo!知恵袋コアデータ -

橋本泰一 (ツール班分担者：東京工業大学 統合研究院)†

Constructing Extended Named Entity Annotated Corpora

Taiichi Hashimoto (Integrated Research Institute, Tokyo Institute of Technology)

1．はじめに

固有表現とは，「机」「椅子」「空」「愛」といった一般的な概念を表す表現ではなく，物，イベントや考え方を表す言語表現（例：夏目漱石，東京オリンピック，日本）であり，質問応答，情報抽出，機械翻訳，テキストマイニングなどに用いられる自然言語処理における重要な基礎知識である．これまで，日本語においては，評価型ワークショップ IREX において，新聞記事に対して固有表現タグ付きコーパス（CRL 固有表現データ）が構築され，そのデータをもとに日本語における固有表現抽出に関する研究が進み，様々な抽出手法 [6–8, 10, 11] 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 種類であり，毎日新聞記事にのみタグが付与されている．しかし，このコーパスを利用して開発された固有表現抽出器を，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抽出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．さらなる高度な言語処理システムの発展に向けて，より詳細に定義された固有表現の定義のもと様々な分野のタグが付与された言語資源の作成が必要である．

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，白書，書籍，Yahoo!知恵袋コアデータへの固有表現タグの付与の結果について報告する．

2．関根の拡張固有表現

固有表現タグ付きコーパス構築に向けて，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)*1を採用した．「関根の拡張固有表現階層」は，MUC(Message Understanding Conference) プロジェクトで策定された固有表現の定義 [1]，それを基に策定された日本における IREX プロジェクトの定義 [3]，ACE(Automatic Content Extraction) プロジェクト*2の定義をもとに，関根が拡張を行った固有表現の定義 [2, 4, 5] である．関根は，質問応答システム，情報抽出，機械翻訳，情報検索，要約などの自然言語処理技術への応用を目的として，この定義の策定を行っている．

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

† hashimoto@iri.titech.ac.jp

*1 http://nlp.cs.nyu.edu/ene/

*2 http://www.nist.gov/speech/tests/ace/

表 1　拡張固有表現タグ付きコーパスの概要 (2009 年 2 月 22 日現在)

	文書数	総文字数	1 文書当りの平均文字数	表現数		1 文書当りの平均表現数
				のべ	異なり	
白書コア	62(62)	352,775	5689.9	11,819	5,277	190.6
書籍コア	49(83)	190,529	3888.3	6,963	2,797	142.1
知恵袋コア	600(939)	127,226	212.0	3,429	2,141	5.7
毎日新聞	8,584	3,643,361	424.4	252,763	63,545	29.4
白書	400	2,340,364	5850.9	74,203	23,857	185.5
CRL	1,174	593,763	505.8	19,254	7,153	16.4

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 200 種類のタグの定義を行っている．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．

３．拡張固有表現タグ付きコーパス (白書，書籍，Yahoo!知恵袋コアデータ)

平成 19 年度に，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 [9]．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．従来のコーパスに比べ，十分に大規模なコーパスを構築することができた．

しかし，これまで構築したコーパスは新聞記事と白書と 2 種類のジャンルのコーパスのみであり，研究対象をさらに広げるためにもジャンルを増やす必要がある．また，従来のコーパスには，形態素情報が付与されていないため固有表現抽出手法の比較において問題がある．従来の固有表現抽出タスクにおける従来手法のほとんどは，形態素情報が必須であるが，CRL 固有表現データには形態素情報が付与されていないため，ChaSen などの形態素解析器を用いる必要があった．しかし，形態素解析器の種類やバージョンによって解析結果が変化するため，手法の比較検討を行う際に論文に記載された性能を再現することが困難であった．固有表現抽出手法の比較検討を容易にするために共通の形態素情報付きのコーパスの作成が必要である．

平成 20 年度においては，白書，書籍，Yahoo!知恵袋各コアデータに対してタグ付けを行った．これまでの新聞記事と白書の 2 種類に加え，新たに書籍と Web の 2 種類のジャンルのデータを構築した．さらに，コアデータには短単位の形態素情報が人手により付与されているため，共通の形態素情報が利用できるようになる．

2009 年 2 月 22 日現在において，白書コアデータ (全 62 文書)62 文書のタグ付けが終了している．白書コアデータに付与されたタグは，のべ 11,819 個，異なり 5,277 個であった．書籍コアデータ (全 83 文書)49 文書のタグ付けが終了している．書籍コアデータに付与されたタグは，のべ 6,963 個，異なり 2,797 個であった．Yahoo!知恵袋コアデータ (全 939 文書)600 文書のタグ付けが終了している．Yahoo!知恵袋コアデータに付与されたタグは，のべ 3,429 個，異なり 2,141 個であった．各コーパスの比較を表 1 に示す．

４．拡張固有表現タグ付けの作業者の比較

拡張固有表現タグ付け作業における作業者間のタグ付けの一致に関して評価を行った．タグ付け作業を行っている作業者 2 名に同一の文書に対してタグ付けをしてもらい，タグ付け結果の一致数を表 2 に示す．

表 2　作業者間のタグ付け結果の一致した表現数と一致率

	文書数	一致数						一致率			
		表現のみ			表現 + タグ			表現のみ		表現 + タグ	
		A1 のみ	一致	A2 のみ	A1 のみ	一致	A2 のみ	A1	A2	A1	A2
白書コア	10	213	1,404	198	268	1,349	253	86.8	87.6	83.4	84.2
書籍コア	10	97	757	133	147	707	183	88.6	85.0	82.8	79.4
知恵袋コア	57	28	317	52	49	296	73	91.6	85.9	85.8	80.2

評価に用いた文書は，白書 10 文書，書籍 10 文書，Yahoo!知恵袋 57 文書である．作業者 1 と作業者 2 を比べた場合，作業者 1 の方がタグ付け数が少ない．作業者 1 は，表現のみの一致率が約 89%，表現とタグの一致率が約 84% であった．作業者 2 は，表現のみの一致率が約 86%，表現とタグの一致率が約 81% であった．タグ付け結果は，80% 以上は一致することがわかった．ジャンル別においてもほとんど一致率の変化はなく，文書のジャンルによってタグ付け結果が一致しなくなるということがわかった．

この結果から作業者間のタグ付けの揺れは全体の 20% ぐらいであるため，拡張固有表現タグ付け作業は，一人の作業者によってタグ付け作業を行い，もう一人の作業者によってタグ付け間違い，タグ付け忘れをチェックするという作業工程でも十分対応できると考えられる．

5．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，固有表現タグを付与した白書，書籍，Yahoo!知恵袋各コアデータへのタグ付け結果について報告した．「関根の拡張固有表現階層」の定義 (Version 7.1.0) に則って，2009 年 2 月 22 日現在白書 (62 文書)，書籍 (49 文書)，Yahoo!知恵袋 (600 文書) に対してタグ付けを行った．また，2 名の作業者間におけるタグ付け結果の比較を行い，約 80% がタグ付け結果が一致することがわかった．拡張固有表現タグ付け作業においては，1 名のタグ付け作業者と 1 名のタグ付け結果の確認の作業者により効率的にタグ付け作業を行うことが可能であると思われる．

今後は，本プロジェクトで構築するコーパスのコアデータすべてに対して，拡張固有表現タグの付与を目指す．加えて，以前構築したコーパスの見直しと修正，タグ付けツールや拡張固有表現抽出ツールの構築を行う予定である．

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

[1] Grishman, R. and Sundheim, B. Message Understanding Conference - 6: A Brief History, COLING-96 (1996).
[2] Sekine, S. Extended Named Entity Ontology with Attribute Information, In Proceedings of the 5th International Conference on Language Resources and Evaluation (2008).
[3] Sekine, S. and Isahar, H. IREX: IR and IE Evaluation project in Japanese, LREC2000 (2000).
[4] Sekine, S. and Nobata, C. Definition, Dictionary and Tagger for Extended Named Entities, In Proceedings of the Forth International Conference on Language Resources and Evaluation (2004).
[5] Sekine, S., Sudo, K. and Nobata, C. Extended Named Entity Hierarchy, LREC2002 (2002).
[6] 山田寛康 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) (2007).

[7] 山田寛康, 工藤拓, 松本裕治 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, **43**, 1 (2004), 44–53.
[8] 浅原正幸, 松本裕治 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, **45**, 5 (2004).
[9] 橋本泰一, 乾孝司, 村上浩司 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).
[10] 中野桂吾, 平井有三 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, **45**, 3 (2004).
[11] 渡辺一郎, 桝井文人, 福本淳一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).

付録 A　白書，書籍，Yahoo!知恵袋コアデータへのタグ付け結果

白書，書籍，Yahoo!知恵袋コアデータへの拡張固有表現タグの頻度分布を表 3，表 4，表 5 に示す.

表 3　拡張固有表現タグの頻度分布 1

拡張固有表現階層				白書		書籍		知恵袋	
				のべ	異なり	のべ	異なり	のべ	異なり
名前	名前 その他			0	0	27	15	1	1
	人名			138	87	863	253	236	197
	神名			0	0	11	10	0	0
	組織名	組織名 その他		55	42	30	17	23	14
		国際組織名		284	125	0	0	3	3
		公演組織名		0	0	0	0	34	29
		家系名		0	0	15	6	0	0
		民族名	民族名 その他	6	4	15	10	9	8
			国籍名	38	8	14	7	15	8
		競技組織名	競技組織名 その他	0	0	0	0	1	1
			プロ競技組織名	1	1	2	2	32	29
			競技リーグ名	1	1	0	0	3	3
		法人名	法人名 その他	139	90	23	9	4	3
			企業名	14	13	97	42	131	102
			企業グループ名	0	0	0	0	1	1
		政治的組織名	政治的組織名 その他	28	23	9	8	4	4
			政府組織名	656	145	28	18	9	5
			政党名	7	6	3	3	8	4
			内閣名	1	1	0	0	0	0
			軍隊名	23	15	3	3	4	4
	地名	地名 その他		2	2	28	22	9	8
		温泉名		3	3	0	0	0	0
		ＧＰＥ	ＧＰＥ その他	28	9	117	58	7	3
			市区町村名	76	61	112	77	18	17
			郡名	0	0	13	13	0	0
			都道府県州名	107	52	78	34	22	13
			国名	848	107	220	28	109	31
		地域名	地域名 その他	3	2	0	0	0	0
			大陸地域名	193	82	32	20	4	3
			国内地域名	40	29	39	26	13	7
		地形名	地形名 その他	0	0	6	3	0	0
			山地名	0	0	0	0	0	0
			島名	4	3	32	18	2	2
			河川名	3	3	22	11	1	1
			湖沼名	2	1	5	5	0	0
			海洋名	2	1	16	11	1	1
			湾名	4	1	8	5	0	0
		天体名	天体名 その他	0	0	0	0	0	0
			恒星名	0	0	0	0	0	0
			惑星名	2	1	11	2	1	1
			星座名	0	0	0	0	0	0
		アドレス	アドレス その他	0	0	0	0	0	0
			郵便住所	0	0	0	0	1	1
			電話番号	0	0	0	0	0	0
			電子メイル	0	0	0	0	7	7
			ＵＲＬ	8	8	2	2	14	14

表 4　拡張固有表現タグの頻度分布 2

拡張固有表現階層				白書		書籍		知恵袋	
				のべ	異なり	のべ	異なり	のべ	異なり
名前	施設名	施設名 その他		29	15	5	3	3	3
		施設部分名		4	3	29	19	4	4
		遺跡名	遺跡名 その他	0	0	5	5	0	0
			古墳名	0	0	0	0	0	0
		ＧＯＥ	ＧＯＥ その他	21	12	40	27	8	6
			公共機関名	47	33	8	6	2	2
			学校名	30	25	38	15	24	17
			研究機関名	28	11	1	1	0	0
			取引所名	0	0	0	0	0	0
			公園名	1	1	9	9	0	0
			競技施設名	4	4	0	0	3	3
			美術博物館名	5	3	0	0	0	0
			動植物園名	0	0	2	2	0	0
			遊園施設名	2	2	0	0	1	1
			劇場名	1	1	1	1	0	0
			神社寺名	0	0	21	15	0	0
			停車場名	0	0	0	0	0	0
			電車駅名	0	0	14	11	1	1
			空港名	4	4	1	1	0	0
			港名	0	0	9	6	0	0
		路線名	路線名 その他	0	0	0	0	0	0
			電車路線名	2	1	5	4	0	0
			道路名	3	3	9	4	2	2
			運河名	0	0	2	2	0	0
			航路名	0	0	0	0	0	0
			トンネル名	0	0	1	1	0	0
			橋名	1	1	4	2	0	0
	製品名	製品名 その他		517	205	177	95	354	252
		材料名		94	34	30	17	17	11
		衣服名		2	1	68	40	31	22
		貨幣名		0	0	0	0	0	0
		医薬品名		1	1	2	2	20	15
		武器名		162	33	8	7	4	3
		株名		0	0	0	0	0	0
		賞名		8	6	0	0	3	3
		勲章名		0	0	0	0	0	0
		罪名		143	50	30	17	23	13
		便名		0	0	0	0	0	0
		等級名		36	15	13	11	14	11
		キャラクター名		0	0	59	12	25	21
		識別番号		1	1	0	0	8	7
		乗り物名	乗り物名 その他	0	0	7	4	1	1
			車名	0	0	8	6	34	31
			列車名	0	0	0	0	0	0
			飛行機名	5	5	11	8	0	0
			宇宙船名	0	0	0	0	0	0
			船名	3	3	43	23	0	0
		食べ物名	食べ物名 その他	90	15	105	46	32	19
			料理名	4	4	72	37	28	21
		芸術名	芸術名 その他	0	0	0	0	1	1
			絵画名	0	0	0	0	0	0
			番組名	4	3	1	1	42	36
			映画名	0	0	4	3	27	21
			公演名	0	0	3	3	2	1
			音楽名	0	0	2	2	18	18
			文学名	2	2	49	34	18	16
		出版物名	出版物名 その他	218	112	12	8	4	4
			新聞名	0	0	6	4	1	1
			雑誌名	4	2	3	3	5	5
		主義方式名	主義方式名 その他	808	333	245	113	201	139
			文化名	7	2	3	3	0	0
			宗教名	4	3	11	6	1	1
			学問名	28	20	57	28	36	20
			競技名	2	2	44	12	21	16
			流派名	0	0	7	7	0	0
			運動名	6	6	2	2	0	0
			理論名	2	1	6	6	0	0
			政策計画名	331	189	1	1	2	2
		規則名	規則名 その他	50	29	6	5	5	5
			条約名	118	70	0	0	1	1
			法令名	243	130	4	4	2	2
		称号名	称号名 その他	15	3	332	24	117	8
			地位・職業名	1046	280	776	276	221	110
		言語名	言語名 その他	0	0	2	1	1	1
			国語名	24	10	9	5	5	2
		単位名	単位名 その他	0	0	0	0	0	0
			通貨名	1	1	0	0	0	0

表 5　拡張固有表現タグの頻度分布 3

拡張固有表現階層				白書		書籍		知恵袋	
				のべ	異なり	のべ	異なり	のべ	異なり
名前	イベント名	イベント名 その他		71	37	10	8	3	3
		催し物名	催し物名 その他	58	43	20	15	9	7
			例祭名	0	0	43	17	0	0
			競技会名	0	0	0	0	0	0
			会議名	146	109	1	1	0	0
		事故事件名	事故事件名 その他	21	14	8	7	1	1
			戦争名	7	6	6	5	1	1
		自然現象名	自然現象名 その他	42	8	13	8	1	1
			自然災害名	13	8	0	0	1	1
			地震名	24	6	0	0	0	0
	自然物名	自然物名 その他		4	3	2	2	3	1
		元素名		23	7	0	0	6	4
		化合物名		28	9	6	5	12	8
		鉱物名		5	4	25	9	7	4
		生物名	生物名 その他	3	2	5	4	5	3
			真菌類名	2	2	0	0	0	0
			軟体動物 節足動物名	0	0	0	0	0	0
			昆虫類	5	4	22	8	10	6
			魚類	1	1	75	22	1	1
			両生類	0	0	0	0	0	0
			爬虫類	0	0	0	0	0	0
			鳥類	1	1	83	12	0	0
			哺乳類	6	6	253	51	19	6
			植物名	135	27	76	35	11	11
		生物部位名	生物部位名 その他	0	0	4	3	6	4
			動物部位名	6	6	439	108	275	112
			植物部位名	15	2	41	16	1	1
	病気名	病気名 その他		0	0	0	0	0	0
		動物病気名		102	42	164	65	109	63
	色名	色名 その他		0	0	19	12	1	1
		自然色名		3	2	72	29	24	14
時間表現	時間表現 その他			1	1	0	0	0	0
	時間	時間 その他		0	0	6	1	1	1
		時刻表現		12	7	73	25	45	28
		日付表現		1688	671	167	115	82	57
		曜日表現		6	4	18	14	3	3
		時代表現		31	12	52	27	3	3
	期間	期間 その他		11	7	9	9	8	5
		時刻期間		32	21	20	15	15	12
		日数期間		0	0	0	0	0	0
		週期間		3	3	8	4	12	7
		月期間		20	15	6	6	23	15
		年期間		152	41	59	39	39	25
数値表現	数値表現 その他			22	16	9	8	31	23
	金額表現			97	86	78	42	71	48
	株指標			0	0	0	0	1	1
	ポイント			52	43	0	0	1	1
	割合表現			707	470	59	30	23	17
	倍数表現			13	10	7	7	4	4
	頻度表現			16	11	47	15	41	14
	年齢			142	77	53	36	93	54
	学齢			24	19	23	13	35	24
	序数			105	65	57	38	34	29
	順位表現			5	4	25	11	48	16
	緯度経度			0	0	0	0	0	0
	寸法表現	寸法表現 その他		2	2	7	5	30	28
		長さ		13	10	29	24	11	11
		面積		10	8	3	3	4	3
		体積		15	14	0	0	0	0
		重量		31	30	13	12	10	10
		速度		0	0	0	0	4	4
		密度		0	0	0	0	0	0
		温度		0	0	1	1	0	0
		カロリー		0	0	0	0	0	0
		震度		0	0	0	0	0	0
		マグニチュード		0	0	0	0	0	0
	個数	個数 その他		61	31	89	31	15	11
		人数		403	280	113	31	49	13
		組織数		112	100	4	3	11	10
		場所数	場所数 その他	52	44	12	11	3	2
			国数	67	36	1	1	0	0
			施設数	64	64	12	8	2	2
			製品数	130	95	50	38	32	24
			イベント数	79	65	6	5	1	1
		自然物数	自然物数 その他	0	0	29	15	0	0
			動物数	0	0	37	24	1	1
			植物数	1	1	0	0	0	0

タグ付きコーパス管理ツール「茶器」の現状と今後

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）

The Current and Future Perspective of *ChaKi*: Annotated Corpus Management 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

ツール班のタスクとして，日本語コーパスの自動タグ付け，および，タグ付け支援ツールの開発を行っている．「茶器(ChaKi)」は，本プロジェクトの当初から開発しているコーパス管理ツールであり，主として，形態素解析，文節まとめ上げ，文節間の係り受け解析の施されたコーパスを格納し，コーパス利用のための種々の機能を提供している．本稿では，茶器に関する本年度の活動とシステムの現状，および，今後の方針について紹介する．

２．タグ付きコーパス管理ツール「茶器」の現状

コーパスを有効に利用するためには，種々のタグ付けを欠かすことができない．日本語については，形態素（単語）への分かち書きが必須の処理であり，データ班および電子化辞書班により UniDic を用いた短単位に基づく形態素情報の付与が行われている．それ以上の情報も言語研究・応用には重要であり，ツール班では，文節区切り，文節係り受け，固有表現，述語や事象名詞の項構造情報，照応詞と先行詞，事象間の時間関係，文書構造など様々なレベルの情報のコーパスへの付与を想定している．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と利用環境の提供を目指したシステムである．

２．１　茶器の基本機能と現状

茶器はタグ付きコーパス管理ツールという位置づけで，次のような機能を実装してきた．以下では，典型的な機能を列挙するが，多くの機能は利用者がカスタマイズすることができる．例えば，異なる品詞体系に対応したり，表示する情報や文字の色，フォントなどを変更することが可能である．

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，文節係り受け解析を想定し，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 を用い，茶筌，あるいは，MeCab によって解析された形態素解析済みコーパス，および，南瓜によって解析された係り受け解析済みコーパスを格納するが，品詞や活用情報，係り受け関係名などは自由に定義できるので，どのような品詞体系で定義されたコーパスも格納可能である．Juman で用いられている田窪・益岡文法に基づく品詞

† matsu@is.naist.jp

体系用の品詞定義ファイルも用意されている．本年度は，茶筌，MeCab，南瓜を呼び出し，解析結果を MySQL データベースへ格納するモジュールを茶器に実装した．また，文字コード(UTF-8,，または，Shift-JIS)をコーパス毎に指定できるようにした．なお，茶器は，日本語に特化したシステムではなく，他に，中国語や英語の品詞タグ付き，および，単語係り受け解析済みコーパスを取り扱うことも可能である．

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るインタフェースを提供しており，これらの任意のパターンでの検索が可能である．複数のコーパスに対する串刺し検索も可能である．文字列検索では，簡易型の正規表現による検索を提供している．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）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．
3. 検索結果の表示：検索対象は，コーパス中の個々の文であり，検索結果は，KWIC(key word in context)形式で１行１文の形で表示される．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，形態素情報のすべてを表示する window があり，マウスを置いた形態素がもつ情報を確認することができる．KWIC のような前後文脈の表示が不要であり，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，単語リストというオプションがあり，検索要求の記述は文検索と変わらないが，検索結果を単語の一覧として新しい window に表示する機能がある．表示された単語一覧の活用情報や出現形などの表示を抑制することにより，原形だけのカウントを表示したりすることが簡単にできる．これらの検索結果は, Excel ファイルとして出力することも可能である．また，文の係り受け構造を表示する機能を備えており，別 window で，文節間の係り受け構造の表示を行う．
4. 統計情報の取得と表示：検索された文集合に対して，前後に現れる形態素の出現回数の表示を行う．また，頻度ではなく，KWIC 中心の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している．
5. タグ付け誤り修正機能：自動解析を行ったタグ付きコーパスは解析誤りを含む．誤りを発見した場合に，上記の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．形態素情報の修正のためのインタフェース TagEdit では，検索結果の中から選択した複数の文（の同じ誤り）に対して，形態素の分かち書きや品詞等の形態素情報の修正を行うことができる．分かち書き誤りに対しては，文字単位での切れ目の挿入や削除を行うことによって分かち書き誤りを修正し，切り出された各断片については，辞書中のその文字列に対応する形態素の一覧を表示し（そのために辞書を読み込ませておくことが可能．辞書が指定されない場合は，コーパスに出現した語の集合が辞書として扱われる)，その中から正しい語を選択することで形態素解析レベルの誤りを修正できる．係り受け誤り修正のためのインタフェース TreeEdit では, 係り受け木を表示するインタフェースをそのまま利用し，文節区切りの誤りや文節の係り先の誤りをマウス操作によって修正することができる．

２．２　茶器の公開に関する活動

茶器の公開，配布については，今年度は，以下の活動を行った．

- 拡大ツール班会議（８月２８日，キャンパスイノベーションセンター）：ツール班の活動と進捗を領域内メンバーに紹介し，今後の活動について意見交換を行った．茶器へのコーパスのポート，エラーメッセージ，マニュアルの整備，コーパス中の単語の一覧表示，種々の用例抽出の支援機能など，様々な要望が寄せられ，今後の開発の参考とした．
- 自然言語処理技術講習会（京都大学主催）への参加（９月８日～９日）：茶器のチュートリアルと実習を行った．講習会へ向けて，茶器の簡易インストーラを作成した．インストーラの指示に従うことにより，コーパスの解析に必要な言語解析ツールやデータベースシステムを簡単にインストールできるようになった．
- ツール講習会（１１月２８日，キャンパスイノベーションセンター）：主に領域内メンバーを対象として，ツール班で開発しているツールの講習会を行った．茶器については，データベースをネットワーク経由でアクセスできるようになったので，奈良先端大のサーバにあるコーパスを検索対象にして講習を行った．また，本年度に一般公開されたみモニター公開コーパスを茶器にインストールするための DVD の配布，および，モニター公開コーパス登録者に対して NAIST の茶器データベースにアクセスするためのユーザ名とパスワードを発行した．

３．タグ付きコーパス管理ツールの今後

プロジェクト当初に計画していたタグ付きコーパス管理機能は，茶器にほぼ実装されたが，茶器自体はかなり複雑なシステムとなってしまった．今年度途中から，Microsoft の .NET framework 上で，新しい ChaKi.NET の開発を並行して行ってきた．これにより，複数の関係データベースの利用が容易になり，MySQL だけでなく，SQLite の利用を開始した．後者は特別なデータベースシステムをインストールする必要がないので，ローカルマシンをデータベースサーバとして用いる場合，MySQL をローカルマシンに実装するという手間を省けることになった．Visual C++で開発してきた現在の茶器の開発は完了し，今後は，現在の茶器の機能を ChaKi.NET に移植していく予定である．

ツール班では，茶器とは別に，汎用のタグ付けツール SLAT を開発している（本ワークショップの松井，野口，飯田，徳永による研究報告を参照）．茶器が提供しているタグ付け機能は，形態素，文節，係り受けに特化しているが，これらは SLAT の機能を用いてタグ付けすることも可能である．一方，SLAT は茶器が提供するような検索機能や検索結果の表示機能を持たない．SLAT の機能を .NET framework に移植することにより，今後，両者の機能の統合を図る予定である．

４．モニター公開コーパスへの係り受け情報の付与

モニター公開コーパスの一部には，UniDic 品詞体系に基づく形態素解析情報が付与されているが，文節および文節係り受け情報は付与されていない．また，現存する南瓜などの係り受け解析ツールは UniDic とは異なる品詞体系のコーパスに付与された係り受け解析

情報をもったコーパスに機械学習手法を適用することによって構築されており，UniDic 品詞体系に基づくコーパスの係り受け解析を行うことができない．正確な係り受け情報が付与されたコーパスは，言語研究のためにも，南瓜のような係り受け解析システムの学習データとしても有用であり，今後は，データ班，電子化辞書班とも連携して，文節および文節係り受けの詳細な仕様を確定し，それに基づく係り受け解析情報付きコーパスを構築していく．

本年度は，機械学習に基づく係り受け解析の高精度化のため，トーナメントモデルという方法を利用し，係り先候補を直接比較することによって最適な係り先を選択する手法を提案し，有効性を確認した(Iwatate et al 2008)(岩立他 2008)．また，この手法によって得られる係り受け関係の強さが，係り受け関係の確信度として他手法より正確な値を出すことを確認し，これを用いることによって，係り受け解析が難しい文例を選択的に抽出することで，少ない学習データで性能の立ち上がりの早い係り受け解析システムの構築が可能であることを示した(岩立他 2009)．このアイデアに基づき，動的に例文を選択しながら，係り受け解析済みコーパスを構築していく予定である．

５．あとがき

タグ付きコーパス管理ツール茶器の現状と予定，および，タグ付きコーパスの構築に関する今後の予定について述べた．本ツールおよびコーパス構築に関して領域内外から忌憚のないフィードバックを期待する．

今年度の主な発表文献

渡邉陽太郎，浅原正幸，松本裕治 (2008), “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.

大熊秀治, 原一夫, 新保仁, 松本裕治 (2008), “機械学習と系列アラインメントを応用した日本語並列句解析,” 人工知能学会全国大会（第 22 回）論文集, 1H1-03, June.

Masakazu Iwatate, Masayuki Asahara and Yuji Matsumoto (2008), “Japanese dependency parsing using a tournament model,” In Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008), pp.361-368. Manchester, UK, Augusut.

Yotaro Watanabe, Masakazu Iwatate, Masayuki Asahara and Yuji Matsumoto (2008), “A Pipeline Approach for Syntactic and Semantic Dependency Parsing,” Proceedings of the 12th Conference on Natural Language Learning (CoNLL-2008), pp.228-232, , August.

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, November.

松本裕治 (2009), “統語情報の付与,” 国文学と鑑賞, Vol.74, No.1, pp.44-52, January.

岩立将和, 浅原正幸, 松本裕治 (2009), “係り受け解析器の部分解析精度評価とその利用,” 情報処理学会, 情報学基礎自然言語処理合同研究会, 2009-FI-93, 2009-NL-189, pp.41-48, January.

大熊秀治，原一夫，新保仁，松本裕治（2008），“バイパス付き編集グラフを用いた日本語並列構造解析,” 情報処理学会 自然言語処理研究会，2009-NL-190, March.

関連 URL

「茶器」配布用ページ：http://chasen.naist.jp/hiki/ChaKi/

汎用アノテーションツール SLAT における 階層構造をもつタグセットのためのインターフェース

松井信太朗　(ツール班協力者：東京工業大学大学院理工学研究科)[1]
野口正樹　(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)
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)

Annotation Interfaces for Hierarchical Structured Tagsets for SLAT

Shintaro Matsui　(Department of Computer Science, Tokyo Institute of Technology)
Masaki Noguchi　(Department of Computer Science, Tokyo Institute of Technology)
Ryu Iida　(Department of Computer Science, Tokyo Institute of Technology)
Takenobu Tokunaga　(Department of Computer Science, Tokyo Institute of Technology)

1 はじめに

近年，自然言語処理の研究分野において，対象とする課題に関する情報 (タグ) を付与した結果 (タグ付きコーパス) を用いた統計的手法が，形態素解析のような基盤処理から情報抽出のような応用処理までさまざまな分野で成果をあげている．統計手法の品質はタグ付きコーパスの品質に依存しているため，コーパスへの網羅的な揺れのないタグ付け作業を実現すること自体にも研究者の関心が高まっている **?**.

コーパスに対する情報の付与を全て人手で行うことは非常に多くの時間を費やすうえ，入力ミスなどの誤りが増える原因になる．コーパスに付与される情報の偏りや誤りは，開発する解析器の性能や解析時の評価に大きく影響を与えるため，コーパスには付与された情報の一貫性や精度が求められる．各コーパスの構築プロジェクトでは情報の付与をサポートするツールを開発し **???**，入力の簡略化や制約を加えることで情報付与のコスト削減を実現した．しかし，これらのツールは特定の仕様に特化しており，別の目的のためにコーパスを構築をする場合には直接利用することができないなど，新しいコーパスを作成する際には新しいツールを別に開発する必要があった．そこで，この問題を解決するために，様々な情報を付与することが可能な汎用アノテーションツール Segment and Link-based Annotation Tool (SLAT) を開発した **?**.

SLAT ではツール内で扱う情報をセグメントとリンクに抽象化し，タグを付与する作業をこれらの追加・削除操作と結びつけることで多様なアノテーションに対応している．現在，多様なアノテーションに対応するためのタグの設定方法や SLAT を用いたアノテーション方法などを記載したマニュアルを Web にて公開しており[2]，問い合わせや機能要望についても受け付けている．

SLAT はタグ付けにおける制約などにより効率的なアノテーションを実現したが，図 **??**に示すように一度に表示するタグの数に制限がある．タスクに応じたインターフェースを選択することでより効率的に作業を行うことができると考えられる．そこで，本稿では付与するタグを選択する問題に着目し，特に階層構造を持つタグセットから適切なタグを選択するインターフェースについて考える．

階層構造を持つ言語資源の例としては，WordNet **?**，分類語彙表 **?**，関根の拡張固有表現階層**?**などがある．たとえば，これらの言語資源で定義されている意味クラスや固有名をタグとして付与することを考えると，階層の深さや同一階層の要素数が増えるほど全体を表示することが困難になる．このため，タグ付けの作業者があらかじめ付与したいタグ名を把握していた場合でも，効率的にそのタグを選択することが難しくなる．BOEMIE[3]プロジェクト **?**で利用されているオントロジーアノテー

[1] smatsui@cl.cs.titech.ac.jp
[2] http://www.cl.cs.titech.ac.jp/SLAT/, mailto:SLAT@cl.cs.titech.ac.jp
[3] Bootstrapping Ontology Evolution with Multimedia Information Extraction

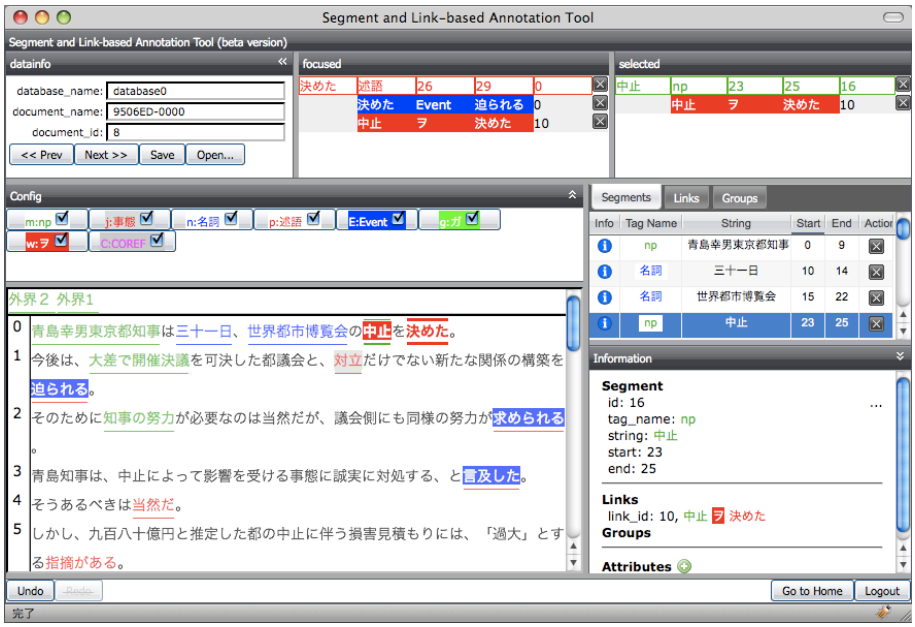


図 1: SLAT のスナップショット

ションツールや関根の拡張固有表現階層 ?のタグ付けに利用されている Fuu Tag [4]では，Windows のエクスプローラで採用されている Folder Tree 形式のインタフェースを利用している．近年，情報視覚化の研究も進められており ??，Folder Tree 形式が階層構造の情報を付与するために必ずしも最適な選択とは言えない．

そこで，本稿では階層構造のタグセットから必要なタグを選択するためのインタフェースとして Folder Tree 形式の表示と WordNet の構造を可視化するために採用されている Hyperbolic Tree 形式の表示の 2 種類を用いて，実際に階層構造を持つタグセットについてタグ付け作業を行うことで，それぞれの特性について調査した．

2 階層構造の表示形式

本稿で採用した Folder Tree 形式と Hyperbolic Tree 形式の 2 つの表示形式について，それらの特徴を以下でまとめる．いずれの表示形式においてもタグはグラフのノードとして表現される．

2.1 Folder Tree 形式

この表示形式は，フォルダの開閉により着目しているノードの子ノードの表示/非表示を切り替える機能を持つ表示形式であり，階層の深さがインデントの深さに対応するという特徴を持つ．このため，階層構造の深さが同一であるノードを把 f 握しやすいという利点を持つ．また，Windows のエクスプローラなどでも導入されており，一般作業者が操作する機会が多く，利用開始時の負荷が小さいと考えられる．階層構造を持つタグセットを用いてタグ付けを行うことができる FuuTag や BOEMIE のツールなどはこの形式を採用している．フォルダの開閉によりすべてのタグへアクセスできるが，タグの種類が多い場合は画面内にすべて表示できないため，作業に必要となるタグが頻繁に変わると選択が困難になるという欠点を持つ．

2.2 Hyperbolic Tree 形式

この表示形式では，親ノードの周りを子ノードが放射状に取り巻く形で配置され，画面の中央ほど大きく，周辺に行くほど小さく表示し，中心から一定以上離れた位置に配置されたノードは表示されない．2 つのノード間の親子関係は矢印で表現され，矢印の先が親ノード，矢印の元が子ノードを

[4]http://nlp.cs.nyu.edu/ene/

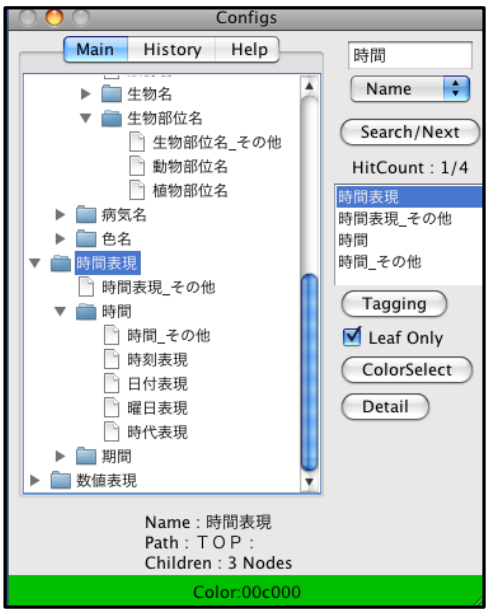


図 2: Folder Tree 表示

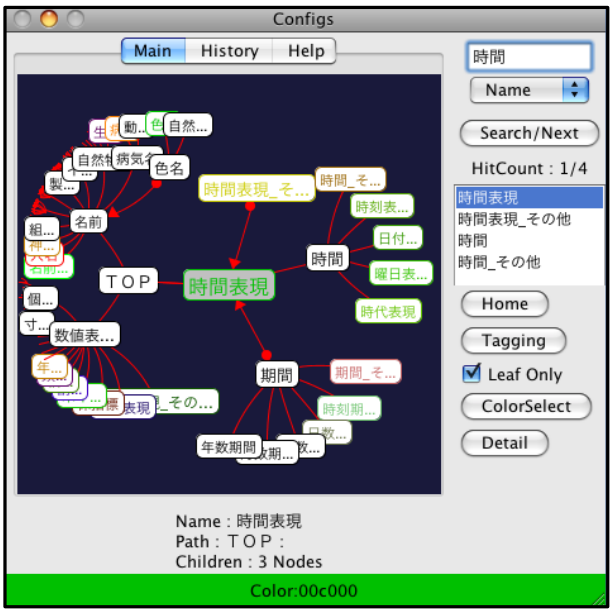


図 3: Hyperbolic Tree 表示

表す．表示されている任意のノードを選択することで，そのノードを中心に再表示されるため，ノードの選択を繰り返すことによりアクセスしたいタグへ移動できる．また，画面上の任意の点を選択しドラッグすることで，見たい方向への視点の切り替えをより簡単に行うことができる．あるノードに対して多くの子ノードが存在する場合，Folder Tree 形式では子ノードを表示するとそれ以外の階層関係がわかりにくくなるのに対し，Hyperbolic Tree 形式では階層の上下が把握しやすいという利点を持つ．ただし，Folder Tree 形式と比較して一般のユーザが操作する機会はほとんど無く，作業開始時にはタグ選択の操作に慣れる必要がある．また，着目しているタグを中心に表示するため，タグ名を表示しただけでは今着目しているタグが全体のどの階層に位置しているのかを把握しづらいという欠点がある．

3 タグ付けツールの実装

インターフェースの違いによる作業効率，精度の違いを見るため Folder Tree 形式と Hyperbolic Tree 形式でタグ選択ができるタグ付けツールを実装した．今回実装したツールでは，タグ選択のインターフェース (Folder Tree 形式か Hyperbolic Tree 形式のいずれか) を表示するウインドウ，タグ付け対象となる文書を表示するウインドウの 2 つで構成されている．タグ選択ウインドウでは，タグ選択のインターフェースに加え，**??**で後述するように，選択しているタグの補足情報の表示やタグ選択の支援となる操作を行うことができる．

3.1 タグ選択ウィンドウ

今回実装したツールのタグ選択ウィンドウのスナップショットを図 **??**と図 **??**に示す．

図 **??**の Folder Tree 形式では，葉ノードはファイルのアイコン，中間ノードはフォルダのアイコンで表現される．また，画面内に表示できないノードについては，スクロールバーで移動することによりアクセス可能となる．

図 **??**の Hyperbolic Tree 形式の実装には TREEBOLIC2 ライブラリ[1]を利用した．選択しているノードは背景を暗くすることで示され，画面右側の “Home” ボタンによりルートノードを選択できる．

また，このウィンドウでは以下のような情報を表示もしくは操作できる．

[1] http://treebolic.sourceforge.net/

- 情報表示: ウィンドウ下部に，選択しているタグに関して，葉となるタグでは親情報と使用例を，中間ノードのタグでは親情報と子ノードの個数を表示する．
- タグ付け履歴: “History” タブを選択することで，これまでに付与したタグのうち最大 20 個の履歴がリスト表示される．このリストからもタグを選択できる．
- タグ検索: 画面右上部に入力した文字列と，タグセット全体から日本語名，英語名，使用例のいずれかの部分文字列が一致するタグを検索できる．検索結果はタグの一覧リストとして表示し，このリストからもタグを選択できる．
- タグ表示色の設定: “ColorSelect” ボタンにより，タグ付け作業ウィンドウの表示色を設定できる．これによりテキストのどの位置にどのタグを付与したのかを視覚的に提示する．図 **??**に示すように，HyperbolicTree 形式の場合はノードの色にも反映される．
- タグの仕様の詳細: “Detail” ボタンにより，選択しているタグの情報 (仕様や作業例など) を表示する．また，“Help” タブを選択すると個別のタグでなく，複数のタグ，あるいはタグセット全体に関連する注意事項が提示される．

3.2 タグ付け作業ウィンドウ

このウィンドウでは，タグ付け対象となる文章とその文章に付与されたタグのリストを表示する．文章中のタグを付与する範囲をマウスで選択し，タグ選択ウィンドウでタグを選択した後に “Tagging” ボタンを押すことで，タグを付与できる．タグのリスト表示部には既に付与されたタグの一覧を表示する．このリストはタグ名，付与された範囲の文字列などでソートができるため，誤ったタグを付与していないかのチェックを行うことができる．

4 評価実験

Folder Tree 形式と Hyperbolic Tree 形式の作業による影響を調査するために比較実験を行った．熟練者と初心者での挙動の違いに付いても合わせて調べる．

4.1 共通の実験設定

関根の拡張固有表現階層 **?**の version7.1.0 のタグセットについて，熟練者あるいは初心者の作業者それぞれ 2 名が 2 つのタグ選択インターフェースを使って作業を行う．タグ付け作業中に作業者が行ったすべての動作は，その時刻と共に作業ログに記録する．

使用したタグセットは固有表現を詳細に分類したもので,「時間表現」,「人名」など 243 の固有名ノードからなる．そのうち葉ノードの数は 196 である．階層の深さは，最も深いところで 5 階層である．また，個別のタグの詳細情報やタグ付けの際の注意事項は**??**で述べた “Detail” ボタンを押すことで参照可能にし，HTML で記述されたオリジナルの仕様書の閲覧は禁止した．

4.2 熟練者によるタグ付け作業

熟練者 2 名によるタグ付け作業では，実験に使う文書として特定研究「日本語コーパス」[2]で構築中の日本語コーパスのコア・データから白書・書籍の 20 文書を選択した．これを 5 文書ずつ 4 つのグループに分け，それぞれの作業者が Folder Tree 形式を使用する場合と Hyperbolic Tree 形式を使用する場合の 4 つの組み合わせの試行で用いた．また，作業対象となる文章には，予め蓄積された過去の固有名タグ付けの履歴からパターンマッチによってある程度のタグを付与した状況で作業を開始する．

本作業前に，タグ付けツールに慣れてもらうために，両作業者は両方のツールを用いて 10 文書ずつの練習作業を行った．

この比較実験では，タグ付け作業の作業時間，グループ内の文書ごとに使用された異なりタグ数の平均，以下に示す作業のタグ選択の一致率，不一致率，タグ情報の参照回数 (“Detail” ボタンが押さ

[2]http://www.tokuteicorpus.jp/

表 1: 熟練者による実験の結果

		FT_1-FT_2	FT_1-HT_2	HT_1-FT_2	HT_1-HT_2
順序	作業者 $_1$	1	2	3	4
	作業者 $_2$	1	2	4	3
総作業時間	作業者 $_1$	3.01	4.94	7.42	3.89
[時間]	作業者 $_2$	1.83	2.55	3.19	1.39
一致率	作業者 $_1$	73.8	81.2	78.4	83.1
[%]	作業者 $_2$	76.5	78.1	84.8	90.6
タグの不一致率 [%]		10.97	5.81	5.40	3.23
平均異なりタグ数	作業者 $_1$	20.8	24.8	25.8	20.8
[種類]	作業者 $_2$	18.5	22.4	22.8	18.3
詳細参照回数	作業者 $_1$	64	43	29	56
[回]	作業者 $_2$	23	38	46	24
履歴使用回数	作業者 $_1$	1	10	25	73
[回]	作業者 $_2$	29	162	151	61

れた総数)，履歴からタグが選択された回数を比較する．タグ選択の一致率と不一致率は以下の式に基づいて算出する．

$$\text{一致率}_i = \frac{\text{両作業者が同一箇所に同一タグを付与した数}}{\text{作業者}_i\text{が付与したタグの数}}$$

$$\text{不一致率} = \frac{\text{両作業者が同一箇所に異なるタグを付与した数}}{\text{両作業者が同一箇所にタグを付与した数}}$$

4.3 熟練者による作業結果と考察

表 **??**に作業者 $_2$ 名の作業の結果をまとめる．FT_i は作業者 $_i$ が Folder Tree 形式で作業した場合を指し，同様に HT_i は作業者 $_i$ が Hyperbolic Tree 形式で作業した場合を表す．たとえば，FT_1-FT_2 は同じ 5 文書に対して両作業者とも Folder Tree 形式で作業をした試行を表わす．なお，FT_1-FT_2 と HT_1-HT_2 については，システムのバグのために両作業者の挙動が変わってしまった文書が 1 つずつあったので結果から除いた．これら 2 つのグループは残り 4 文書の結果を示している．

表 **??**にまとめた内容とログの情報から以下に示す内容がわかった．

- タグ付けの総時間は一貫して作業者 $_1$ の方が時間がかかっており，作業効率について表示形式の影響は観察できなかった．作業者 $_2$ の方が作業時間が少ない理由としては，表 **??**の履歴使用回数の差から分かるように，作業者 $_2$ の方が作業を効率的に行うために履歴を多用していることがあげられる．今回の作業では，1 文書で使用するタグの異なりは平均して 20 種類前後であり、最大 20 項目であるタグ履歴の効果はかなり大きなものである。この時間と履歴使用回数との関係は、どのような種類のタグが使用されるかが把握できればそれまでの履歴を用いてタグを選択する方が作業を効率的に進められることを示している．
- 表 **??**の一致率を見ると，作業者 $_2$ 名が Folder Tree 形式を使用した場合に比べ，共に Hyperbolic Tree 形式を使用した場合の方が高くなっており，また不一致率も減少している．これは，Hyperbolic Tree 形式で表示した場合には，例えばルートノードを表示している状況で視界に入ったタグから次に付与するタグを選択するといったバイアスがかかるのに対し，Folder Tree 形式の場合は画面に表示できるタグが隣接する兄弟ノードの個数によって制限され，より貪欲にタグを探索する必要があるために，作業者２人でタグ選択の差がでたのだと考えられる．

4.4 正解データとの比較

??. の両作業者による成果物を別の熟練者の手により統合・修正したものをここでは正解データと呼ぶことにする．この正解データとの比較結果を表**??**に示す．作業者 i の精度，再現率，F 値は以下の式から算出される．

$$\text{精度}_i = \frac{\text{作業者 } i \text{ が正解データと同一箇所に同一タグを付与した数}}{\text{作業者 } i \text{ が付与したタグの数}}$$

表 2: 正解データとの比較

		FT_1-FT_2	FT_1-HT_2	HT_1-FT_2	HT_1-HT_2
総タグ数	作業者 $_1$	374	798	871	361
[個]	作業者 $_2$	361	830	805	331
	正解	393	844	828	358
精度	作業者 $_1$	82.4	85.5	82.0	90.3
[%]	作業者 $_2$	99.4	93.3	93.8	97.9
再現率	作業者 $_1$	78.4	80.8	86.2	91.1
[%]	作業者 $_2$	91.3	91.7	94.6	90.5
F 値	作業者 $_1$	80.3	83.1	84.0	90.7
[%]	作業者 $_2$	95.2	92.5	94.2	94.0
総異なりタグ数	作業者 $_1$	54	64	76	53
[種類]	作業者 $_2$	51	59	70	46
	正解	54	62	72	52

$$\text{再現率}_i = \frac{\text{作業者 } i \text{ が正解データと同一箇所に同一タグを付与した数}}{\text{正解データに付与されているタグの数}}$$

$$F \text{ 値}_i = \frac{2 \times \text{精度}_i \times \text{再現率}_i}{\text{精度}_i + \text{再現率}_i}$$

表 **??**にまとめた内容とログの情報から以下に示す内容がわかった．

- 両作業者とも実験で用いたタグセットを使ったタグ付けの経験は約 1 年であるが，作業者 $_1$ は他のタグセットによるタグ付けの経験もある．しかし，これまでの実験での作業時間の短さや詳細参照回数の少なさ，全体的な F 値の高さを考えると，作業者 $_2$ の方がよりこのタグセットの理解度が高いと推測できる．このことから，タグ付けの経験よりも，作業を行うタグセットへの理解度がタグ付けに重要な要素であることが分かる．
- 表**??**の総タグ数について見たとき，Hyperbolic Tree 形式を用いているグループの方が Folder Tree 形式を用いる時よりもタグ数が多い．特に正解データと比較した際，作業者 $_1$ の方は Hyperbolic Tree 形式を用いるグループのみ正解データのタグ数を上回るという顕著な結果まで現れている．この結果より，Hyperbolic Tree 形式の方がタグを概観できて付与すべきタグを把握しやすいことと，その反面必要のない箇所に付与してしまいやすいことが推測される．その上，タグセットの理解度が高いと，タグを付ける必要のある箇所が分かると共に，タグを付ける必要の無い切り捨てるべき箇所も分かるようになるのだが，逆に自らの理解度に不安があると，切り捨てるべきであるという判断が困難となることも影響していると考えられる．また，HT_1-FT_2 が過剰に付与されているのは，表**??**の詳細参照回数を見ると，このグループだけタグ数の割に詳細を見返す回数が少ないことも要因に挙げられる．
- F 値を見ると，作業者 $_1$ は Hyperbolic Tree 形式の方が高く，作業者 $_2$ は Folder Tree 形式の方が高い．先のタグの理解度の考察をふまえると，作業者 $_1$ が Hyperbolic Tree 形式の方が正しく付与できるようになるのは，タグセットの理解度に不安があるためにインターフェースを確認することも多く，その際により多くのタグを纏めて概観できる方が良かったからだと考えられ，作業者 $_2$ が Folder Tree 形式の方が正しく付与できるようになるのは，これまでに用いていたタグ付けツール (Fuu Tag) が Folder Tree 形式であり，仕様書でのタグ一覧も Folder Tree 形式での配置に近い表の形であったことから，より慣れていて考えた通りのものを選べるためだと考えられる．タグセットの構造に通じていればいるほど使い慣れた Folder Tree 形式の方が精度が良くなるが，まだ理解度に不安がある場合は多くのタグを概観できる Hyperbolic Tree 形式の方が正しく付与できるのだと推測される．

表 3: 初心者による実験の結果

		FT_1-FT_2	FT_1-HT_2	HT_1-FT_2	HT_1-HT_2
順序	作業者 $_1$	1	2	4	3
	作業者 $_2$	3	2	4	1
総タグ数 [個]	正解	175	171	168	174
総作業時間	作業者 $_1$	0.87	0.69	0.98	0.92
[時間]	作業者 $_2$	0.42	0.68	0.47	1.28
正解との一致率	作業者 $_1$	93.7	98.8	89.4	93.2
[%]	作業者 $_2$	89.2	92.4	88.2	93.7
平均異なりタグ数	作業者 $_1$	16.0	15.0	21.5	23.5
[種類]	作業者 $_2$	15	15.5	21.5	24.5
	正解	16.5	18.0	20.5	23.5
タグの不一致率 [%]		11.4	7.6	13.1	10.3%
詳細参照回数	作業者 $_1$	23	17	25	23
[回]	作業者 $_2$	7	24	7	82

4.5 初心者によるタグ付け作業

タグ付け経験の無い初心者として，計算機科学を専攻する大学生 2 名を対象に**??**. と同様の要領でタグ付け作業を行わせた．**??**. の実験との違いは以下の通りである．

- タグセットの縮小: 用いるタグセットは**??**. と同様関根の拡張固有表現階層であるが，この階層構造を全く把握していない初心者が対象であることをふまえ，もともと 5 階層の深さであったものを 4 階層まで縮小したものを用いた．5 階層に位置していたタグを付与するべき箇所は，その上位の中間ノードであったタグを用いてタグ付けさせる．これにより，選択を行う異なりタグ数は 193 から 103 まで減少する．
- 文書数の減少: 使用する文書は，**??**. で用いた文書より 8 文書を抜き出して用いた．
- タグ付け位置の指定: タグ付け作業に馴染みのない初心者にとって，文書中のタグを付与すべき箇所の発見は非常に困難なタスクであり，本稿では付与すべきタグの選択を支援することに焦点を当てているため，予め文書にタグを付与すべき場所を指定することでタグの選択作業に集中させた．タグを付与すべき場所は**??**. の正解データに従う．
- タグ付けツールの修正: 5 階層に位置するタグは使用しないので表示せず，4 階層に位置する中間ノードのタグが新たに葉ノードとなるため，その下位のノードとなるタグの詳細な定義，使用例を参照できるようにした．検索時に 5 階層に位置するタグが検索された場合，その上位の中間ノードのタグが選択される．また，**??**. における履歴利用の影響と，タグセットの縮小による異なりタグ数の減少を考慮すると，さらに履歴利用の影響が強くなることが予想できる．そのため，タグ履歴の機能を外し，純粋に 2 種類のインターフェースを用いてタグの選択を行わせた．
- 比較項目: 今回は，タグを付与すべき場所が指定されているためタグの付与漏れが発生せず，一致率とタグの不一致率とが同じ意味を持つ．精度・再現率も同様である．そのため，一致率をタグの不一致率に，精度・再現率を正解との一致率にまとめた．

4.6 初心者による作業結果と考察

表 **??**に作業者 $_2$ 名の作業の比較結果をまとめる．表**??**と同様に，FT_i は作業者 i が Folder Tree 形式で作業した場合を指し，同様に HT_i は作業者 i が Hyperbolic Tree 形式で作業した場合を指す．

- 作業者はタグ付けの経験がないため，作業を行うにつれてタグ付けに習熟する要素が大きくなると考えられる．しかし，両者とも 2 つ目のタグ付けツールを用いた 3 番目，4 番目のグループがそれぞれその作業者内で 3 番目，4 番目の一致率であり，また 3 番目のグループで異なりタグ数を比べた際もう片方の作業者よりも少なくなっている．これより，最初のタグ付けツールの使

用を終えたことで集中が途切れてしまったのだと推測される．また，インターフェース，作業者に関わらずこの現象が起きたことから導入時の精度の点では両者に差はないと言える．

- 作業時間を見ると，Folder Tree 形式を用いた時の方が作業順に関わらず短い時間で終わっている．つまり，初心者にとって，作業効率の観点では Folder Tree 形式の方がよい．これは，Folder Tree 形式の方が馴染みがあり，抵抗が少なかったからだと考えられる．
- **??**. でのタグセットの理解度が低いと Hyperbolic Tree 形式の方がよいという結果と，初心者の作業能率は Folder Tree 形式の方がよいという結果は食い違っている．これは，完全に知識が 0 の状態から始めた初心者と，1 年は作業を行っている熟練者とでは差があり，熟練者の理解度は低くともどの辺りを探せば目的のタグが見つかるか程度の把握は出来ていたからだと考えられる．

5 まとめ

本稿では，SLAT の現状について報告するとともに，階層的な構造を持つタグセットを用いてタグ付けをおこなう場合に，タグセットの表現形式の違いによってタグ付けの効率や作業の一致率にどのような影響が出るかを調査した．その結果，Hyperbolic Tree 形式の方が作業者同士での一致度は高いものの，タグセットへの理解が深い熟練者は配置を覚えている Folder Tree 形式の方が精度がよかった．また，初心者でも精度にこそ差はないものの，操作に慣れている Folder Tree 形式の方が作業効率の点で有利であった．これを踏まえると，Hyperbolic Tree 形式よりも Folder Tree 形式を導入した方が広く有用であると考えられる．ただし，今回行った比較実験では，2 種類のタグ選択インターフェースに加え，履歴や文字列検索を利用したタグの選択も可能であるために厳密な比較とは言えず，またサンプルも少ないため統計的な優位とも言えず，今後さらに調査を続ける必要がある．今後は SLAT への実装も視野に入れた表示形式の改良や他の表示形式の導入についても吟味したい．

参考文献

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech and Communication)*. MIT Press, 1998.

Pavlina Fragkou, Georgios Petasis, Aris Theodorakos, Vangelis Karkaletsis, and Constantine Spyropoulos. Boemie ontology-based text annotation tool. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

GATE. GATE, a general architecture for text engineering. *http://gate.ac.uk/*.

Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, and Stephanie Strassel. Annotation tool development for large-scale corpus creation projects at the linguistic data consortium. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

Constantin Orăsan. Palinka: A highly customisable tool for discourse annotation. In *Procceedings of the 4th SIGDIAL Workshop on Discourse and Dialogue*, 2003.

NEGRA Project. @nnotate. *http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/*.

Ramana Rao, Jan O. Pedersen, Marti A. Hearst, Jock D. Mackinlay, Stuart K. Card, Larry Masinter, Per-Kristian Halvorsen, and george G. Robertson. Rich interaction in the digital library. *COMMUNICATIONS of the ACM*, Vol. 38, No. 4, pp. 29–39, 1995.

国立国語研究所（編）. 分類語彙表 (国立国語研究所資料集). 大日本印刷, 2004.

菊池司, 伊藤貴之, 岡崎章. Web ナビゲーション技術にみる情報デザイン・情報視覚化の最近の動向. 芸術科学会論文誌, Vol. 4, No. 1, pp. 1–12, 2005.

野口正樹, 三好健太, 徳永健伸, 飯田龍, 小町守, 乾健太郎. 汎用アノテーションツール SLAT. 言語処理学会第 14 回年次大会予稿集, 2008.

関根聡, 竹内康介. 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.

BCCWJ に見られるオノマトペの型と共起との関連

ホドシチェク・ボル（作文支援システム班協力者：東京工業大学大学院社会理工学科）†
ベケシュ・アンドレイ（作文支援システム班協力者：リュブリャーナ大学）
仁科喜久子（作文支援システム班班長：東京工業大学大学院社会理工学科）

Onomatopoeic Patterns and their Relation to Collocations inside the BCCWJ

Bor Hodoscek (Tokyo Institute of Technology)
Andrej Bekes (University of Ljubljana)
Kikuko Nishina (Tokyo Institute of Technology)

１．はじめに

オノマトペとは、動物の声や物の様態と様子を表す語であり、もともとに存在した音から生み出された単語が言語学的時間の中で、変形をもたらしたものと考えられている。日本語では、擬態語・擬声語・擬音語などについて様々な研究が行なわれているが、本稿ではそれらを包含した意味でオノマトペという用語を用いる。日本語ではオノマトペが広く用いられ、新しい語が生成しつつあるとともに、既存のオノマトペも用法が変化するため、国語辞書や教科書に載っていないオノマトペが少なからずある。定着したオノマトペないし新しいオノマトペの法則と使い方を明確にする必要がある。日本語のオノマトペを外国人学習者に理解してもらうため、定着したオノマトペないし新しいオノマトペの法則と使い方を明確にする必要がある。

オノマトペには様々な型があり、その活用に音声的・韻律的な法則と言語的な規則が見られる。例えば「ぱしぱし」のように単位となる 2 モーラを繰り返す「ABAB」型、「ぱしっ」と単位となるモーラの後に促音を取る「AB ッ」型、「ばしん」のように単位となるモーラの後に「ん」が来る「AB ン」型というようなバリエーションが見られる。これは「ぱし」というもとの単位「AB」型があり、その活用形・異形（濁音化・半濁音化）としてみることができる（小野 2003 p. 333）。

本稿では、「現代日本語書き言葉均衡コーパス」（BCCWJ）を対象にオノマトペの抽出を行い、「核」となる型とその異形を共起の観点から構文的な構造を明らかにすることを目的とする。オノマトペは意味の要素の構造を明らかにすることも必要であり、「核」が同じオノマトペでも、意味が異なる場合が多々あるが、共起情報を用いた意味的分析が必要であるが、今回は構文に焦点を当てた。

２．オノマトペの抽出

２．１　先行研究

コーパスからオノマトペを自動抽出する方法として、奥村らと笹野らが提案したものがある（奥村 2003、笹野 2007）。奥村らは「ABAB」「A ッ B リ」など 10 個の高頻度の型を中心に、全パターンを生成し、それがオノマトペかどうかを Web コーパスで確認する。さらに係り受け関係を行い、共起情報でシソーラスに基づいて概念辞書を作成した。一方、笹野らは形態素解析の精度をあげるため、2–4 文字の繰り返しを新聞とウエブコーパスから抽出し、その中からオノマトペである語を人手で辞書に書き込み、反復型のオノマトペを精度高く認識できたという。

その他に、自動抽出を用いず、人手で作成したリストに基づいてオノマトペのオンライン辞書を作成した研究もある（浅賀 2007、2008）。浅賀らはオノマトペを副詞的用法、連

†hodoscek.b.aa@m.titech.ac.jp

体詞的用法、複合名詞的用法に分けている。オノマトペが複数の意味を持つ場合、そのオノマトペがある文を構文解析し、係り先の情報をベクター化してクラスタリングし、意味を自動分類する。

２．２　本手法

オノマトペの抽出には形態素解析器 MeCab を使用し、MeCab の辞書としては UniDic 辞書を使用した。オノマトペの可能な品詞としては、副詞、形容動詞、サ変名詞がある。UniDic では、それらに相当する品詞でオノマトペのほとんどを含めているのは、副詞と形状詞（形容動詞の語幹部分）である。他にもサ変名詞がオノマトペになることが考えられるが、BCCWJ の解析したデータの中、サ変名詞のオノマトペは「びっくり」の 1 語のみだったため、調査の対象からサ変名詞を外した。上記の品詞を制限し、形態素内のパタンマッチングによって型を得られた。パタンマッチングに用いた形態素情報は UniDic の発音形の出現形である。それは、「くた / \」「くたくた」「クタクタ」など表記の揺れの問題が避けられ、同じオノマトペとして抽出できたからである。

オノマトペの型を決めるときはいくつかの修正が必要となった。まず、「ふうっと」「フーッと」を「A一ッ」型にするなど、長音が「ー」と平仮名「う」で書いてある場合、一つの形に統一した。また、「ジーンと」「のほほん／と」のように、UniDic 辞書の中では格助詞「と」が形態素に含まれている場合と自立した形態素としてある場合がある。その際は格助詞「と」「っと」や「っ」が形態素の中にあるにもかかわらず、「AB」型の B として認められない。さらに型の「核」を定めるときに、清音と濁音・半濁音の違いを無視して、同じ核であるように設計し、共起を計算する段階でその正しさを確かめる。

抽出された語の中に過ちの最も大きい原因は形態素解析の間違いである。それは「すみ／ませ／ん／が／、／髭／のそり／跡／が／青い／感じ／で／気持ち／悪い／です／。」の「のそり」と「がん／保険／は／がんと／糖尿／病／の／因果／関係／が／立証／さ／れ／て／い／ない」の「がんと」のように、形態素解析の区切りの誤りに生じる問題である。他に、「剣／を／グル／グル／と／回し／ます」の「グルグル」のようにオノマトペが２つの形態素に分けられていることもある。

高頻度で現れた語の多くが形式上、オノマトペに似ているが、音と様態を表していない定着した用法の語、または擬態と擬声とは異なる語源の語がある。それらがオノマトペかどうかを『日本語オノマトペ辞典』と『暮らし言葉の擬態語擬音語辞典』に参照しながら、オノマトペではない語を非オノマトペリストに登録した。次の段階で、得られたオノマトペのリストに対して、ぞれぞれの抽出された語がオノマトペかどうかを日本語母語話者 3 人に人手で判断してもらった。そして、オノマトペではない語を非オノマトペリストに追加し、その一部を表 2 に示す。

表 1:オノマトペの型と出現形の関係

A	B	型	発音形	出現形
カ	タ	AB ッ	カタッ	カタッ
			ガタッ	ガタッ
		ABAB	カタカタ	かたかた
				カタカタ

表 2:オノマトペとして認めない語

ナカナカ	イロイロ	モトモト
マスマス	イヨイヨ	ワザワザ
タチマチ	タマタマ	タビタビ
ソコソコ	ツギツギ	ハナハダ
タダタダ	マタマタ	クレグレ
ハルバル	トモドモ	マチマチ

非オノマトペリストを参照しながら再度抽出を行なった。その結果、抽出できたオノマトペの異なり数は 470 語であり、延べ数は 25658 語である。また、型の数は 32 個を抽出した。得られた型は「A一、A一B リ、A一ン、A一ッ、AA、AA ッ、AB、AB一、ABAB、ABB、ABB ン、AB ン、AB ン A、AB ン AB ン、AB ッ、AB ッ A、AB リ、AB リ A、AB リ AB リ、A ン、A ン B、A ン B ン、A ン B リ、A ッ、A ッ A、A ッ B、A ッ BA ッ B、A ッ B ン、A ッ B リ、A リッ、A リリ、A」である。

３．BCCWJ にみられるオノマトペ

書籍は型の異なり数が一番多いが、それは書籍の量の大きさが原因のではなく、書籍の特徴といえる。また、500 万形態素に対して各コーパスのオノマトペの延べ数が最も多い Yahoo!知恵袋でも、書籍の 2 倍の容量でも、型の種類が少ないということがわかる。

表 3:各コーパスの特殊なオノマトペ

書籍	頻度	Yahoo!	頻度	国会	頻度
ポカン	61	ゴックン	3	バーント	2
フワリ	31	モッチリ	3	パタ	1
ジロリ	28	ウルウル	3		
ダラリ	25	ヘロヘロ	3	／	
ズングリ	25	プリット	2		
異なり数	延べ数	異なり数	延べ数	異なり数	延べ数
138	898	12	22	2	3

表 4:各コーパスにおける頻度が上位 5 個のオノマトペの出現割合

書籍	%	Yahoo!	%	国会	%	白書	%
ズット	8.42	チャント	15.86	ズット	44.75	ユッタリ	22.03
ユックリ	6.44	ズット	12.61	チャント	19.34	ユックリ	8.47
チャント	4.74	ユックリ	4.38	ソロソロ	3.54	ズット	6.78
ジット	4.66	ソロソロ	4.08	サッパリ	2.15	ノンビリ	6.78
スッカリ	4.14	スッキリ	3.42	ズバリ	1.52	ホット	6.78
異なり数（延べ数）		異なり数（延べ数）		異なり数（延べ数）		異なり数（延べ数）	
455(18631)		316(5386)		127(1582)		26(59)	

表 5:各コーパスのオノマトペの型の異なりと延べ数の割合

書籍	異なり数	延べ数	Yahoo!	異なり数	延べ数	国会	異なり数	延べ数	白書	異なり数	延べ数
ABAB	33.63	17.01	ABAB	37.66	21.05	AッBリ	26.77	10.87	AッBリ	46.15	52.54
AッBリ	19.12	31.43	AッBリ	22.15	30.63	ABAB	24.41	7.96	ABAB	23.08	13.56
ABリ	9.45	5.57	Aッ	9.18	21.05	Aッ	19.69	54.55	Aッ	15.38	20.34
Aッ	7.47	27.47	ABリ	6.01	1.97	Aン	8.66	21.62	AンBリ	7.69	8.47
ABン	5.27	1.44	AンBリ	5.06	4.29	AンBリ	5.51	1.52	ABリ	7.69	5.08
AンBリ	5.05	6.09	ABン	4.43	0.69	ABリ	4.72	2.02			
Aン	3.96	7.03	AB	3.48	0.52	ABン	3.15	0.25			
AB	2.86	0.34	Aン	3.16	16.95	AB	1.57	0.25	／		
Aーッ	1.76	1.03	Aリッ	1.58	0.33	Aリッ	1.57	0.19			
ABッ	1.54	0.13	Aーッ	1.58	0.54	ABッA	0.79	0.06			

４．共起

オノマトペの共起を考える場合、構文的な形式と意味的な関係をみなければならない。ここではまず構文的な構造に焦点を当てることとし、オノマトペと共起する語との意味的関係は、今後の課題とする。

オノマトペの品詞の面からみると、名詞、副詞、形容動詞など複数の品詞があるといえる。このことから各品詞に対応する共起関係の抽出を行なう必要がある。オノマトペは下記の 1)〜3）のような構文的な用法に分けられる。

1)「かちかち山」「コチコチ先生」「こてこての方言」など、名詞修飾になっている用法がある。これはオノマトペが形容詞的な働きをして、名詞を修飾する例である。

2)「よれよれなのが分かるコート」、「さらさらな髪」のように「よれよれ」[さらさら]というオノマトペが形容動詞的な振る舞いをしている例がある。オノマトペの語尾に「の」がつくか、「な」がつくかは、他の名詞修飾の接続と同様の問題を含んでおり、Yahoo!知恵袋などにみられる「〜な」の新しい用法といえる。

3)「外見はヨレヨレ。」のように述語形となっている用法がある。

オノマトペの付属語として可能な助詞は「の」「に」「と」「な」である。他に、付属語の動詞「する」、または、「とした」のような決まったパターンが多い。たとえば「すらすら／すらり／ずらり」のグループの中で、「すらすら」は「すらすら（と）よむ」、「ずらり」は「ずらりと並ぶ」というが、「すらり」は「すらりとした美人」というようにパターンが決まっている。これらのパターンをさらに詳細に調査することで、オノマトペの構文的な構造を明らかにできると考える。

5. まとめと今後の課題

本稿では、BCCWJ からオノマトペの抽出方法について述べ、抽出されたオノマトペをコーパス毎に分け、それぞれの特殊なオノマトペとコーパス間の差異を検討した。型とオノマトペの異なり数とそのコーパスへの関係も議論し、「核」となる型とその異形を共起の観点から考察し、オノマトペの構文的な用法の違いとして名詞修飾の用法、形容動詞的な用法、述語的な用法があることを示した。

今後の課題としては、ウエブからオノマトペのデータを取得し、型と共起の関連とともに、意味的な構造も調査する。

参考文献

浅賀千里、渡辺知恵美（2007）「Web コーパスを用いたオノマトペ用例辞典の開発」『電子情報通信学会データ工学ワークショップ（DEWS2007）』

浅賀千里、Yusuf Mukarramah、渡辺知恵美（2008）「オンラインオノマトペ用例辞典「オノマトペディア」における用例を意味により分類するための係り受け関係を考慮したクラスタリング手法」『電子情報通信学会データ工学ワークショップ（DEWS2008）』

奥村敦史、斎藤豪、奥村学（2003）「Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築」『言語処理学会第 9 回年次大会(NLP2003)発表論文集』pp. 63–70

小野正弘（編）（2003）『日本語オノマトペ辞典』講談社

笹野遼平、黒橋禎夫（2007）「形態素解析における連濁および反復形オノマトペの自動認識」『言語処理学会第 13 回年次大会(NLP2007)発表論文集』

山口仲美（編）（2007）『暮らし言葉の擬態語擬音語辞典』小学館

関連 URL

形態素解析エンジン MeCab 0.97. http://mecab.sourceforge.net/

形態素解析辞書 UniDic 1.3.9. http://www.tokuteicorpus.jp/dist/

「現代日本語書き言葉均衡コーパス」モニター公開データ（2008 年度版）. http://www.kokken.go.jp/kotonoha/ex_8.html

短単位を対象とした連濁の処理について

山田篤（電子化辞書班連携研究者：京都高度技術研究所）†

Processing “Rendaku” for Short-unit Words

Atsushi Yamada (ASTEM RI/Kyoto)

１．はじめに

連濁とは，カ，サ，タ，ハ行ではじまる語頭清音が，語の複合によって濁音化（ハ行については半濁音化を含む）する現象のことである。たとえば，「株式会社」において，「カイシャ」の語頭清音「カ」が濁って「ガイシャ」と読まれる。連濁が起こる必要条件は，語頭音がカ，サ，タ，ハ行であることであるが，語頭音がカ，サ，タ，ハ行であるからといって，すべての語が常に連濁を起こすわけではない。語によっては，全く連濁を起こさないもの，条件によって連濁を起こしたり起こさなかったりするものがある。本稿では，電子化辞書班で作成している電子化辞書 UniDic（伝 2007）における連濁の取り扱いについて述べる。

２．UniDic における連濁の取り扱い

連濁とは語頭清音の濁音化現象であるため，語境界でしか起こらない。これは，どのような単位で語を認定するかによって，処理の仕方が変わってくることを意味する。

UniDicでは，辞書に登録される語の単位として，国立国語研究所で規定された「短単位」という揺れのない斉一な単位を採用している。短単位とは，原則として，最小単位（現代語で意味を持つ最小の単位）を1回結合したものである。たとえば，先の例では「株式」「会社」がそれぞれ一短単位となり，それらの間に語境界が存在し，「株式/会社」は2つの短単位から構成される。よって，「会社」という語の語頭音が「株式」という先行語と接続することにより濁音化するという現象を連濁処理として取り扱う必要が生じる。

一方，「雨傘」は，最小単位の「雨」と「傘」の1回結合で構成されているので，これ自体は一短単位となり，間に語境界は存在しない。このとき，「カサ」が「ガサ」と濁音化しているが，辞書には短単位である「雨傘」という単位で登録されるため，語の読みないし発音形として「アマガサ」と登録しておくことで，この濁音化は処理されるので，連濁処理の対象とはならない。ところが同じ「傘」でも，「相合い/傘」は二つの短単位から成り，「傘」の前に語境界が存在するため，「傘」の語頭音が「相合い」という先行語と接続することにより濁音化するという連濁処理を行う必要がある。

このように，UniDic においては，一般に連濁と呼ばれる現象のうち，短単位内で起こるものについては辞書に登録しておくことにより正しく処理できるが，短単位の語頭の濁音化については，なんらかの方策が必要になる。これに対し，UniDic では，「会社」や「傘」のような語については，語頭変化形として「濁音形」（ないし「半濁音形」）を持たせるこ

† yamada@astem.or.jp

とにより，濁音化する可能性がある語であることを表している。

ただし，濁音形を持つ語が常に連濁するわけではなく，一定の語の境界をまたいでは起こらないことが知られている。たとえば，「万葉/仮名」の「仮名」は連濁するが，「外来/語/仮名/表記」の「仮名」は連濁しない。UniDic では階層的な単位設計を採用しており，短単位の上に中単位，長単位という単位を設けている。この例に対しては，中単位では「外来語/仮名表記」という分割となり，長単位では全体で一長単位となる。中単位は，語の内部構造に基づき設定され，中単位を超えて連濁を起こすことはない。現状では中単位の分割ツールはまだ完成していないが，UniDic ではこの中単位の存在を仮定し，中単位の範囲内で，語頭変化形を複数持つ短単位の語頭音の選択の問題として連濁を取り扱っている。現在，UniDic 1.3.9 中，語頭変化形に「濁音形」ないし「半濁音形」をもつものは 3516 個ある。このうち，語頭変化形を複数もつものは 1792 個あり，これらが中単位の先頭ではなく内部に出てきたときに，いずれの変化形であるかを選ぶことが UniDic における連濁の処理となる。

３．助数詞の連濁処理

中単位内部の短単位の語頭の連濁現象の典型的なものとして，助数詞の語頭の音変化がある。助数詞は数詞と接続して，一つの中単位を構成する。たとえば，「本」という助数詞に対する発音形には「ホン」「ボン」「ポン」があり，「イッ/ポン」「ニ/ホン」「サン/ボン」のように，先行する数詞によって規則的に変化する。これを，助数詞の語頭変化型 (iType)・語頭変化形 (iForm) と前接する数詞の語頭変化結合型(iConType)を引数とする以下のような関数を定義して，もっとも値の大きい変化形を選択するツール ChaOne（山田 2007）を作成している。

$$F_i(\text{iType}, \text{iForm}, \text{iConType})$$

たとえば，「一/本」に対しては，「一」が語頭変化結合型として”N1”，「本」が語頭変化型として”ホ混合”を持ち，F_i (ホ混合，半濁音形, N1) = 1.0 と定義されており，これが最大値となるため，結果として半濁音形である「ポン」が選択される。

４．連濁を起こす条件

助数詞の連濁については，数詞との接続に限定し，それぞれの数詞に語頭変化結合型を持たせることにより，規則的な変化を表す関数を用いて定式化することにより対応したが，これをそのまま助数詞以外の一般的な語の連濁の処理に拡張することは，先行語にどのような結合型を付与すればよいかが明らかではなく難しい。このため，助数詞以外の語の連濁現象を取り扱うにあたり，（佐藤 1989）で述べられている連濁に関する条件を UniDic の枠組みの中で検討した。

４．１　対象となる語に関する条件

以下の項目は，連濁を起こす語そのものに関する条件である。

(V-1L) 漢語の語頭音は原則として濁らない

連濁を起こさない語については，辞書中で「濁音形」を持たせないことで表現できる。

漢語において，例外とされる「砂糖」，「会社」，「菓子」，「稽古」等については，辞書中で「濁音形」を持たせればよい。

(V-2L) **外来語は連濁を起こさない**

V-1L と同様に，辞書中で「濁音形」を持たせなければよい。そもそも外来語は多くの場合，片仮名表記のため，濁音形を持つことがない。「歌留多」，「煙管」のように，漢字表記で連濁を起こすもののみ「濁音形」を持たせればよい。

(V-3L) **擬声，擬態語は連濁を起こさない**

やはり同様に，辞書中で「濁音形」を持たせなければよいが，そもそも擬声語，擬態語は一短単位になることが多いため，UniDic では問題とならない。

(V-4L) **名詞および動詞の畳語は連濁を起こす**

畳語は基本的に一短単位となるため，連濁の対象とならない。

４．２　音韻に関わる条件

以下は，連濁を起こす語の音韻に関する条件である。

(V-5L) **第二音節以降に濁音を含む語の語頭音は濁らない**

「言葉」等が該当するが，辞書中で「濁音形」を持たせないことで対応できる。

(V-9) **無声摩擦音音節が連続する語は連濁を起こさない**

同様に，辞書中で「濁音形」を持たせないことで対応できる。

４．３　先行語の音韻に関する条件

(V-6) **連濁によって，同種または類似の音が連続するときは，濁音化が避けられ易い**

先行語の最後尾の音と，当該語の先頭の音を比較する。ただし，（佐藤 1989）にあげられている例のうち，短単位境界を含むものは，「位置/付ける」，「もらい/火」，「しあげ/鉋」のみであり，この素性の有効性については検討が必要である。

(V-7) **漢字造語成分の連濁は，それに付く漢字のモーラ数や濁音の有無等によって決まる**

漢字造語成分は，UniDic では「接尾辞」に相当すると考えられ，名詞に対しては適用できない。

(1)「停留/所」，「派出/所」のように，先行語に濁音が含まれなければ連濁するが，「裁判/所」，「発電/所」のように，濁音が含まれていれば連濁しない。

(2) 「会議/所」，「刑務/所」，「事務/所」，「社務/所」，「登記/所」のように，隣接漢字が 1 モーラの場合は連濁しない。

(V-8) **直前が撥音のときは連濁を起こし易い**

先行語の最後尾の音が撥音の場合は連濁するというものであるが，（佐藤 1989）にあげられている例は，UniDic ではすべて短単位内の現象であった。有効な素性か否かについては検証が必要である。

４．４　隣接する形態素に関わる条件

(V-10) 接頭語（御，真，片，唐等）は連濁を起こしにくい

先行語がこのような接頭辞の場合は，連濁を起こさないというものである。すべての接頭辞について成り立つわけではない。

(V-11) 形容詞素性の形態素が付く場合，連濁を起こしにくい

（佐藤 1989）にあげられている例は，UniDic ではほぼすべて短単位内の現象であったため，より詳しい検討が必要である。

４．５　語構造に関わる条件

(V-12L) 意義の類似した，あるいは相反する語が並置された複合語は連濁を起こさない

（佐藤 1989）にあげられている例は，UniDic ではすべて短単位内の現象であったため，より詳しい検討が必要である。

(V-13L)（動詞＋動詞）の複合動詞は濁音化しない

（佐藤 1989）にあげられている例は，UniDic ではすべて短単位内の現象であったため，より詳しい検討が必要である。

(V-14) 三語から成る複合語において，その構造が「左枝分かれ」ならば後続語が連続的に濁音化することが可能であるが，「右枝分かれ」では第二語は濁音化しない

UniDic では短単位と中単位との組み合わせで表現され，この条件で濁音化しない場合は，中単位の語境界に相当する。

(V-15)（名詞＋動詞連用形）において，副詞的連用修飾関係では連濁を起こし易く，格関係では連濁を起こしにくい

語の用法に関わるため，取り扱いが困難である。

(V-16)（名詞＋動詞連用形）で「…する人」の意味のときは連濁を生じない

語の意味に関わるため，取り扱いが困難である。

５．連濁正解コーパスの作成

現状では，連濁現象を取り扱うにあたって，評価に用いることの出来る正解コーパスが存在しないため，連濁を起こす可能性のある語について，実例をもとに先行語とのペアで実際に連濁するか否かを記述した正解コーパスを作成した。作業は第 1 期と第 2 期に分かれ，第 1 期は，名詞を対象にして，連濁を起こす可能性のある名詞 275 個について，主にウェブから手作業で事例を収集し，6,560 事例を集めた。このうち，短単位の語境界を含んでいたものは 4,740 事例であった。

第 2 期は，Web N グラム（工藤 2007）より，UniDic で短単位の語境界に対象語がくるものを取り出して判定を行っており，現在，41 名詞，110,844 事例を収集している。

以下では，第 1 期の正解コーパスを用いた分析結果を示す。

５．１　先行語の語種の影響

基本形を選択した 1222 事例のうち,
　当該語の語種が和語：422
　　先行語種が和語：184，漢語：88，外来語：60
　当該語の語種が漢語：527
　　先行語種が和語：100，漢語：302，外来語：49
　当該語の語種が外来語：1
　　先行語種が和語：0，漢語：0，外来語：0
濁音形を選択した 3510 事例のうち,
　当該語の語種が和語：2213
　　先行語種が和語：670，漢語：900，外来語：262
　当該語の語種が漢語：1002
　　先行語種が和語：308，漢語：370，外来語：86
　当該語の語種が外来語：6
　　先行語種が和語：2，漢語：2，外来語：2
半濁音形を選択した 8 事例のうち,
　当該語の語種が和語：3
　　先行語種が和語：1，漢語：0，外来語：0
　当該語の語種が漢語：4
　先行語種が和語：0，漢語：2，外来語：0
当該語の語種が外来語：0

以上の結果から，先行語の語種の影響はほとんど見られない。

５．２　先行語が「お」「ご」「御」といった接頭辞である場合

基本形を選択した 1222 事例のうち，先行語が接頭辞のもの：89
　うち,「お」: 38,「ご」: 3,「御」: 12
濁音形を選択した 3510 事例のうち，先行語が接頭辞のもの：109
　うち,「お」: 0,「ご」: 0,「御」: 0
半濁音形を選択した 8 事例のうち，先行語が接頭辞のもの」: 2
　うち,「お」: 0,「ご」: 0,「御」: 0

以上の結果から，先行語が「お」「ご」「御」といった接頭辞の場合，連濁はしないと言える。

５．３　直前が撥音の場合

基本形を選択した 1222 事例のうち，直前が「ン」: 115
濁音形を選択した 3510 事例のうち，直前が「ン」: 384
半濁音形を選択した 8 事例のうち，直前が「ン」: 2

以上の結果から，少なくとも名詞については，直前が撥音か否かは影響しない。

６．おわりに

本稿では，UniDic を用いた解析結果に対する連濁の処理について述べた。UniDic では，当該語が連濁を起こす可能性を持つか否かは，複数の語頭変化形を持つか否かによって表現される。このうち，複数の変化形をもつもののうち，特に助数詞については，数詞との組み合わせに限定して定式化し，処理を行うツールを作成している。それ以外の語については，複数の語頭変化形の中から適切に選択を行うために，先行語および当該語のどのような素性を見ればよいかについて，正解コーパスを作成して検討した。こうして抽出した連濁に影響を与える可能性のある素性を観測素性として，条件付確率場 (Lefferty et al., 2001) を用いた実験を行い，現在，妥当性を検証している。今後，助数詞の処理と同様にツール化して配布する予定である。

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.

山田篤(2007). 『ChaOne マニュアル』

佐藤大和(1989). 「複合語におけるアクセント規則と連濁規則」講座日本語と日本語教育第 2 巻 日本語の音声・音韻(上), pp.233-265, 明治書院.

工藤拓, 賀沢秀人(2007).「Web 日本語 N グラム第 1 版」言語資源協会.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001) “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *Proceedings of the 18th International Conference on Machine Learning*, pp. 282-289.

Yahoo!知恵袋にみる非規範的表現

杉本 武 (日本語学班分担者: 筑波大学大学院人文社会科学研究科) †

Non-Standard Expressions in *Yahoo! Chiebukuro*

Takeshi Sugimoto (Graduate School of Humanities and Social Sciences, University of Tsukuba)

1 はじめに

しばしば「言葉の乱れ」(あるいは、評価的な意味合いのない「言葉のゆれ」) として取り上げられる言語現象がある。例えば、その代表的なものは、「ら抜き言葉」「さ入れ言葉」のような語形に関わるものであったり、否定と共起しない「全然」のような用法に関わるものであったりする。このような表現は、いずれ正用とみなされることになる可能性はあるとしても、現時点では規範的には正用とされない、「非規範的な表現」である。

このような非規範的な表現は、話し言葉では現れやすいが、書き言葉、特に、コーパスとしてしばしば用いられる新聞、白書のような規範性の高い資料においては、当然のことながら現れにくい。これに対して、「Yahoo!知恵袋」[1](以下「知恵袋」と呼ぶ) は、インターネット上の掲示板という性格上、書き言葉であるとはいえ、話し言葉的な要素が入りやすく、また、規範意識も働きにくいことから、非規範的な表現が現れやすくなる。

本発表は、知恵袋のような資料[2] が日本語研究 (特に文法研究) にどのように用いることができるのかということを、非規範的表現を例に模索するものである。

2 非規範的表現の資料としての知恵袋

知恵袋にみられる非規範的表現の一例として、さ入れ言葉をみると、本発表で用いた知恵袋のデータ[3]では、次のような「さ入れ言葉」が、数百例規模で現れる。

(1) 「次回、落札されることが御座いましたら、必ず差し引かさせていただきますので、暫くの間、当方で預からせていただきます。これからもどうぞよろしくお願いします。」と、メールを返信し、次回落札時に、実際に精算しました。 (知恵袋 04/06/03 [Yahoo!オークション])

(2) あなたの言う邦楽は現在では純邦楽と呼ばれて区別されています。と、以前誰かが同じような質問に答えたのをパクらさせていただきました。 (知恵袋 05/02/22 [音楽])

(3) みなさんは子供とスーパーに行ったとき、子供に「試食荒らし」をさせないようにちゃんとしつけできていますか？何度も何度も無くなるまで食べに行かさせたり、楊枝やゴミを床に捨てさせたり弁当箱持参で試食品をつめて帰らせたり … みっともないことになっていませんか？ (知恵袋 05/08/31 [幼児教育、幼稚園、保育園])

一方、「さ」の連続する、次のような用例も少なからず観察される。

---

†sugi@lingua.tsukuba.ac.jp

[1]Yahoo!知恵袋は、Yahoo! JAPAN が運営するインターネット上の掲示板で、利用者がカテゴリで分けられた掲示板上で質問をし、他の利用者が回答するという形をとるものである。以下に挙げる用例の [ ] 内は、質問のカテゴリを示す。なお、「現代日本語書き言葉均衡コーパス」のためにヤフー株式会社より提供された評価用データを使用した。

[2]「日本語書き言葉コーパス」として今後公開されるブログとともに、知恵袋は、他のサブコーパスと比べ、かなり異質な資料であると考えられる。

[3]約 15 億文字から成る。

(4) 冬眠前なんか、今にも増してすんごい食欲です。今は1日3回くらいでいいんですが、冬眠前（秋）はしっかり食べささせてあげて下さいね。私は夏も秋もしっかり食べさせてあげましたが（欲しがるだけ••• (＾＾;))、 さほど大きくなりませんでした。 (知恵袋 04/07/05 [動物、植物、ペット])

佐野 (2008b) は、国会会議録を用いた、さ入れ言葉の研究であるが、「さ」の連続を含むさ入れ言葉が一つもなかったことから、「二重「さ」制約」というものを提案している。この点からすると、上のような例は、「食べさせ」という規範形も同時に現れていることもあり、タイプミスのような書き誤まりの疑いが極めて強い。一般的に、インターネットの書き込みの場合、このような書き誤まりはつきものであり、タイプミスを誘発しやすい表現の場合、データの信頼性が問題とされるところである。

また、佐野 (2008a)、佐野 (2008b) で詳細な分析があるように、さ入れ言葉のような表現は、生年、性別、居住地の影響を受ける。一方、知恵袋の場合、特に質問、回答に記載がない限り、生年、居住地はおろか、性別すらわからないことがほとんどである。

このような点で、知恵袋のような資料には限界もある。それでは、どのような現象ならば、知恵袋を非規範的表現の分析に使うことができるのであろうか。

3 「違かった」という表現

近年みられる非規範的な表現として、「違かった」という表現がある。これは、動詞「違う」が形容詞化したものであるが、その成立については、井上 (1998) などに詳しい。井上 (1998) は、このような表現の存在理由とでも言うべきものを次のように述べている。

> 動詞「違う」にそのまま過去のタを付けると「ちがった」になる。これは、「違う」という動詞の性質上、変化の瞬間をさす。例えば化学の実験をやっていて、試薬を垂らしながら、今か今かと色の変わる瞬間を待っているときなら、「あっ違った」と言ってもいい。しかし「昔はバレーボールのルールは……」という文脈だと「違った」ではふさわしくない。「違っていた」の方がいい。テイタを付けると過去の継続的状態を示せる。ところが別の便法があったのだ。「違う」を形容詞であるかのように見なし、カッタを付けて、「昔はバレーボールのルールは違かった」と表現すると、「違っていた」にあたる過去の継続的状態を、もっと簡潔に表わせるのである。
>
> (井上 (1998:67f.))

また、井上 (1998:69) によると、この表現の成立は、「福島県あたりで数十年前に発生して、北関東を経て、東京に入ってきた」と考えられるようである。先に述べたように、知恵袋では、このような年代、地理的分布をみることはできないが、使用実態としては、「違かった」[4]というタ形の用例は、「違った」が2万例ほどあるのに対して、225例がみられる。

(5) 違う気がする。誰だかは知らないけど。どうなんでしょう？ 調べたら、ちがかった。ＭＯＤＥＬのとこにプロフィールあったよ。 (知恵袋 04/07/14 [芸能人、タレント])

(6) 異母兄弟だと思ってたら違かったらしい。結局くっつくし。 (知恵袋 04/12/22 [テレビ、ラジオ])

(7) 多分、この方だと思います！＾＾ 違かったらすみません••••••••••••＾＾; (知恵袋 05/09/04 [コスメ、美容])

[4]以下の用例数には、「ちがかった」などの仮名表記のものも含まれる。

それぞれ、タ形が文末に使われているもの、助動詞、助詞が続くものである。なお、(5) では、「ちがかった」と「違う」が共存している。

さらに、井上 (1998:70f.) は、終止形にあたると想定される「ちがい」の変化形「ちげー」という語形の存在についてふれた上で、次のように各活用形について述べている。

> これで、「ちがかろう・ちがくない・ちがかった・ちがい・ちがければ」という新しい形容詞の活用形が、ほぼそろったことになる。もっとも現代東京の口語の実際の使用状況に合わせれば、「ちがかろう」とは言わずに「ちがうだろ (う)」と言うし、「ちがければ」もあまり使われず、「ちがけりゃ」「ちがきゃ」「ちがったら」「ちがうなら」が使われる。だから「ちがかった」「ちがくない」に加えて「ちげー」が出た段階で、もう現代語形容詞としては完成の域に達していたと考えていい。
>
> (井上 (1998:71))

「ちげー」の形は、音変化を伴うことから当然であるが、知恵袋では、次のように話し言葉的な文脈でしか現れない。

(8) 引越しと死んでしまうじゃ全然ちげーよ。Σ (゜Д゜;≡;゜д゜) (知恵袋 5/07/15 [コミック])

(9) 部分分数分解？ちげーか。左辺を展開しちゃダメなんですか？ (知恵袋 5/08/13 [数学、サイエンス])

この点で、「ちげー」とその他の活用形では、文体的な位置づけが異なるものと思われる。

一方、確かに、「違かろう」の形は 4 例[5] と少ない (3 例のみを挙げる)。

(10) 自分と同じだろうが違かろうが全部大切な意見なのです。 (知恵袋 04/07/18 [Yahoo! オークション])

(11) 女の子の涙には弱いから、好きな子だろうが、違かろうが優しくしちゃうと思う。 (知恵袋 04/10/29 [恋愛相談、人間関係の悩み])

(12) でも、夫婦という家庭内での立場は年がちがかろうと「対等」じゃないと子供の教育にしても、親戚づきあいにしても何かと苦労することが多いように思います。 (知恵袋 05/06/15 [恋愛相談、人間関係の悩み])

また、「違ければ」の形は 27 例みられる。

(13) 育った環境も違ければ受けてきた教育も違う。 (知恵袋 04/11/02 [恋愛相談、人間関係の悩み])

(14) また、充電池と充電器はメーカーが違ければ使えませんよね?? (知恵袋 05/07/22 [デジタルカメラ])

(15) 全然文化も違ければ住んでる人間だって違うのにそれをどっちが良いか決めるのはどうかと … (知恵袋 05/09/26 [政治、社会問題])

ただし、このような活用形の出現率については、動詞型の「違う」の各活用形の出現率などと比較しなければならない。

次に、井上 (1998) の指摘と関連して興味深いのは、「違かった」という形の用法である。よく知られているように、動詞のタ形は、連体修飾において、過去の出来事ではなく現在の状態を表すことがある。

[5]「言葉、語学」のカテゴリなどで、このような表現について述べられている用例は除いた。

(16)　帽子をかぶった人 (帽子をかぶっている人)
(17)　曲がった棒 (曲がっている棒)

井上 (1998) の言うように、「違かった」が、「違った」が表せない過去の状態を表す「違っていた」の代わりに用いられるとすると、連体修飾において、「違った」でもって表すことのできる現在の状態を「違かった」が表すことがないのではないかと予測される。実際、「違かった」の用例の中で、連体修飾の用法のものは 8 例のみである[6]。次に、同様な例は除いて示す。

(18)　資産の桁が違かった場合には一から教えないと無理です。　(知恵袋 05/01/12 [家計、貯金])
(19)　はっきり直接言おうよ。。そんなさぐり合いなんて、違かったとき悲しいだけですよ。(知恵袋 05/07/13 [恋愛相談、人間関係の悩み])
(20)　たしかスレとレスでは意味が違かった気がします。　(知恵袋 04/10/30 [インターネット])

これらも、修飾されているのは、「場合」「とき」「気」のような形式名詞的なもので、内の関係の連体修飾ではないことが特徴的である。

一方、「違った」の場合は、次のように、状態を表す用例が数多くみられる。

(21)　麻婆豆腐の素で何か違った調理方法などありませんか？　(知恵袋 04/12/18 [レシピ、調理法])
(22)　猫以外でも、左右違った目の色をした動物を飼っていらっしゃる方いらっしゃいますか? (知恵袋 05/02/17 [動物、植物、ペット])
(23)　それにこういう場に書き込むということは、自分と違った意見の人もいるのは当たり前です。　(知恵袋 04/08/03 [恋愛相談、人間関係の悩み])

このような用例が、「違かった」ではみられないのである。

4　おわりに

前節で「違かった」について観察した現象には、ミスタイプによる誤った用例が入り込む恐れはほとんどない。また、観点も、あくまでも文法内的なものであり、話者の属性を捨象してみることも (ある程度は) 可能であろう。この点で、このような現象は、知恵袋のような資料を用いることが可能なものであろう。

ここでは、その一例として示したまでであり、知恵袋あるいはブログのような資料を日本語研究に利用する可能性の一端を示したにすぎない。この他にどのような研究に利用することができるのか、さらにはそれだけでなく適しているのかは、十分な検討が必要である。

また、このような現象は、純然たる話し言葉においても当然、観察可能なものであり、そこでも同じような結果が得られるのか、「話し言葉コーパス」との比較も必要になってくるであろう。

参考文献

井上史雄 (1998)『日本語ウォッチング』、岩波新書
佐野真一郎 (2008a)「『日本語話し言葉コーパス』に現れる「さ入れ言葉」に関する数量的分析」、『言語研究』133、pp.77-106
佐野真一郎 (2008b)「国会会議録によるさ入れ言葉の分析」、松田謙次郎 (編)『国会会議録を使った日本語研究』、pp.159-184、ひつじ書房

[6]「こと」「の」を修飾するものは除く。なお、これも、全体の用例数の中で相対的に捉えなければならないが、ここではふれないことにする。

大規模コーパスの語彙統計情報の利用を支援する
—語彙情報データベースを参照する API の構築と活用—

千葉庄寿（日本語教育班協力者：麗澤大学外国語学部）†

Supporting the Use of Statistic Information Taken from Large-corpora: A Case Study in Building APIs for Lexical-Statistical Database

Shoju Chiba (College of Foreign Studies, Reitaku University)

１.「ブラックボックス」としてのコーパス分析ツール

本稿では，従来の高機能なコーパス分析ツールの問題点である「ブラックボックス化」と「拡張性の制限」を解決するため，拡張性に優れたコーパスデータの利用環境の構築に必要な，コーパス利用ツールのコンポーネント化を提案する。

コーパス検索・分析ツールはこれまで，コーパスそのものの分析にその重点を置いてきた。しかし，日本語教育，国語教育，国語政策など，実際に応用が求められる分野においては，コーパスから得られた情報を使い，異なるコーパスのデータを相互に分析・比較・評価することが重要な作業となってくる。本特定領域が構築する『現代日本語書き言葉均衡コーパス』(以下，BCCWJ と略す) においては，他のコーパスの分析の基礎資料として活用できる「均衡性」に大きな特徴がある。また，著作権処理を通じ広く公開が予定されていることから，そのデータは幅広い分野において活用できる素地を持っている (前川 2006)。このような要請を踏まえ，今後コーパスを用いた分析環境の構築においては現行のコーパス分析の枠組みを超えたツールの開発・提供が今後強く意識されなければならない。

一方，コーパスは，その規模が大きくなり，またデータの構造が複雑になればなるほど利用者に高い情報処理能力を要求する。分析者が自分でツールを作成するのでなければ，必要なコーパス解析機能が実装されたコーパス検索・分析ツールを利用することになるが，データが大規模になり，処理が複雑になればなるほど，ツールに対する要請は増し，反面そのツールが行っている処理方法の検証や分析結果の正確さの把握が困難になる。単純な用例検索にとどまらないコーパス分析を行いたい場合，ツールの利用者はこのような「ブラックボックス化」の問題に多かれ少なかれ悩ませられる。また，このような複雑なツールの構造について大きな責任を負わなければならないツールの開発者の負担も大きい。

一つの解決方法として，分析ツールをコンポーネント化し，プログラミングにおけるソフトウエアの構成原理 (例えば Model-View-Controller, MVC) を利用して，それぞれについて明確な規格のもとツールを構築し，処理を分散させることが考えられる。**図１**は MVC の原理を図示したものである。入出力画面 (VIEW) は，入力内容を処理するコンポーネント (CONTROLLER) からは独立しており，処理自体は CONTROLLER からデータベース (MODEL) に流れ，VIEW が MODEL に直接アクセスすることはない。

† schiba@reitaku-u.ac.jp

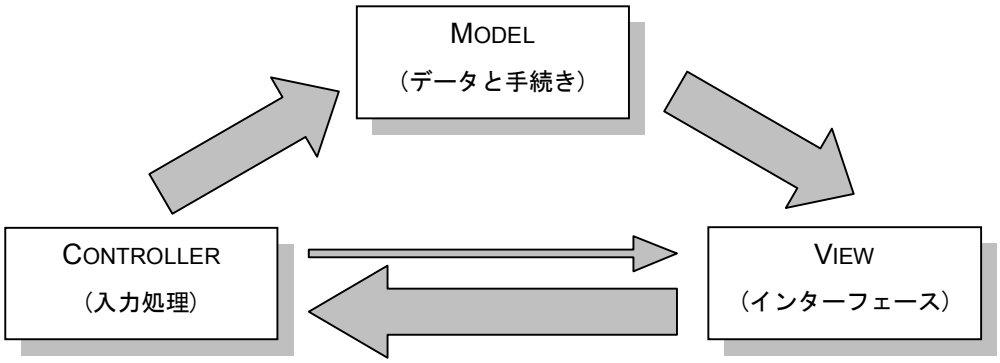


図１ ソフトウエア構築における MVC 原理

このようなコンポーネント化をコーパス利用ツールに当てはめると，MODEL はコーパス，VIEW はコーパス利用ツールのインターフェースとなろう。しかし，コーパスそれ自体には VIEW への出力をおこなう手続きは含まれておらず，コーパスを組み込み，かつ VIEW への出力の手続きをもつデータベースの仕組みが必要となる。

MVC はユーザインターフェースをもつアプリケーションの構築において有効な原理であるが，このように問題点も持っている。特に，データ蓄積のためのデータベースが入出力の仕組みを備えた独立したシステムで動作することが多い現在，コンポーネント間の関係を以下のように合理化することが望ましいと考えられる。

- CONTROLLER の再定義 (データベースやインターフェースのコンポーネントから独立して，データ入出力を一貫してコントロールできる API を構築する) [1]
- コーパス情報をデータベースへのアクセスに必要な, SQL や XQuery 等の問い合わせ言語のカプセル化 (問い合わせ言語を直接記述せずに必要なデータが取得できること)
- 言語分析・言語教育の要請に合わせた問い合わせパターンの事前設定 (データベースを最大限活用した情報を引き出すことができること)
- API の標準化 (データベースの仕様が変更されても検索ツールが維持できること)
- クロスプラットフォーム化 (OS などプラットフォームが異なっても利用できること)

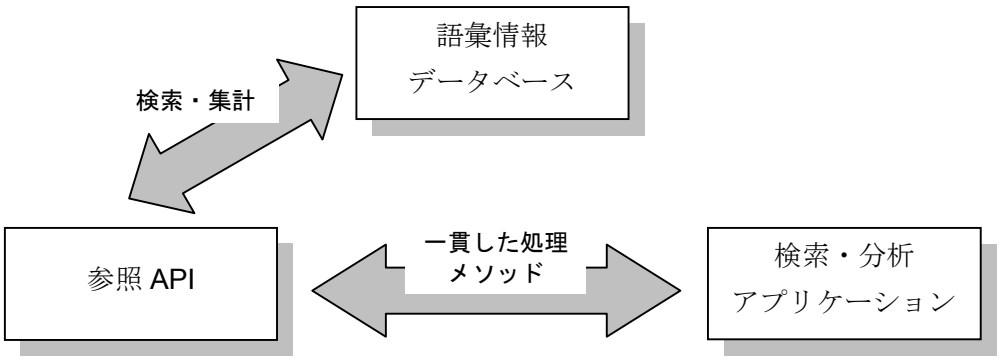


図２ コーパスの語彙情報データベースの利用モデル

[1] このような API 思想を実装したコーパス検索ツール構築の例として, フィンランド語のコーパス検索システム `Lemmie` (Grönroos & Miettinen 2002) がある。

本稿では，語彙情報を収録したデータベースとそれを参照する API を構築し，**図２**に示したようなモデルに従うことで，多様な目的に応えるコーパス利用ツールの構築が可能になることを示す。

２．語彙情報のデータベース化

本稿では，BCCWJ のデータのうち，領域内に公開されているデータ (2008 年版) について，書籍サンプル (plain text 版) のみを対象として形態素解析辞書 UniDic (version 1.3.9, ChaSen 版 [2]) にて解析した結果を用い，語彙情報データベースのプロトタイプを構築した。

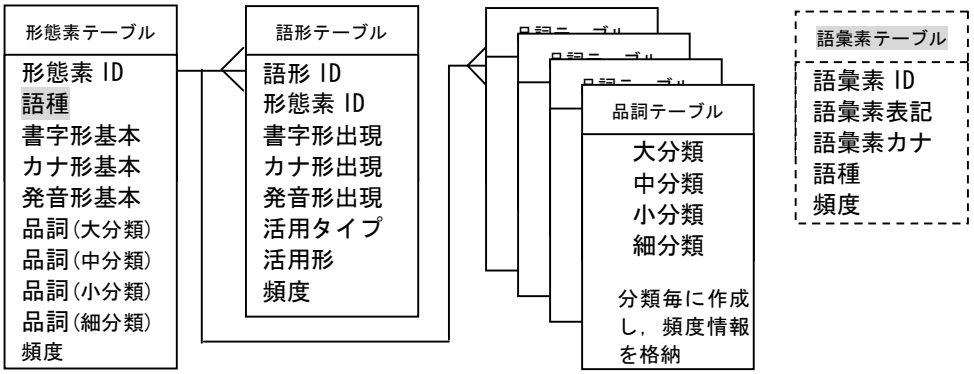


図３ 語彙データベースの構造

各テーブルはフリーの関係データベース管理システム (RDBMS) である MySQL 5.0 に格納し，語形テーブル，各品詞テーブルから形態素テーブルへの関係づけをおこなっている。

今回構築する語彙情報データベースでは，簡便を期すため「語彙素」テーブルは省略し，検索対象から外している。語彙素情報を利用することで多様な書字形をまとめて検索することが可能になる。[3] 今回は，語彙素情報の一つである「語種」情報のみ「形態素」テーブルに格納し利用している。

語彙情報データベースは，格納しているデータの精度の検証とともに，どのような情報を提供するかの要請に応じた拡張が必要になる。今回は単純な語彙統計情報のみを格納の対象としたが，コーパス自体の収録により用例自体の検索が可能になるほか，コロケーション情報などを収録することで，データベースの利用価値は大きく高まるだろう。

３．語彙情報データベースを参照する API の構築

本稿では，API のプロトタイプとして，プログラミング言語 Perl (version 5.8) のモジュール (cf. Foster-Johnson (1998), モジュール名 Bccwj::Lex) として語彙データベースへのアクセスメソッドを作成し，データベースとツールのインターフェースの連携を支援する。

---

[2] 解析作業には「茶まめ」(version 1.57，UniDic version 1.3.9 パッケージに付属) を用い，解析器として ChaSen を使いた。解析の際，以下の正規化オプションを使用している (このうち (3) は現行バージョンでは解析器として ChaSen を用いた場合のみ利用可能である)：(1) 半角英数字の全角変換 (2) NumTrans による数字処理 (3) ChaOne による音変化処理

[3] 特に数詞の音変化に関しては語彙素による正規化が効果的である。cf. 山田 (2007)

語彙情報データベースから参照できるデータとしては以下を想定し，それぞれを検索するためのメソッドを**図４**のように作成した (メソッドは対応するものを一部のみ提示する)。

- 検索キー
  - 書字形，カナ形，発音形，品詞，活用型／形，頻度などの検索キーに基づく検索 (ワイルドカード含む)
  - 対応する基本形・出現形の検索
- 検索対象
  - 形態素 (基本形・出現形) の検索 (単独，または)
  - 頻度情報の検索
  - 品詞情報・活用型・活用形の検索

```
Bccwj::Lex::Search("検索キー", "パターン", "{0/1}", "検索対象")
    ※シンプルなサーチ機能を提供する
      第３引数：0 = 基本形; 1 = 出現形
      第４引数：出力したい属性を/で区切って記述
Bccwj::Lex::SearchFreq("頻度", "{>/</>=/<=}", "検索キー", "パターン
    ", "{0/1}", "検索対象")
    ※頻度情報を含めたシンプルなサーチ機能を提供する
      第２引数： > = より大きい; < = 未満; >= = 以上; <= = 以下
      第５引数：0 = 基本形; 1 = 出現形
      第６引数：出力したい属性を/で区切って記述
Bccwj::Lex::SearchPattern("検索パターン", "{0/1}", "検索対象")
    ※複雑なサーチ機能を提供する
      第１引数：[検索キー1='パターン'●検索キー2='パターン'…] [4]
      第２引数：0 = 基本形; 1 = 出現形
      第３引数：出力したい属性を/で区切って記述
Bccwj::Lex::SearchPatternFreq("頻度", "{>/</>=/<=}", "検索パターン
    ", "{0/1}", "検索対象")
    ※複雑なサーチ機能を提供する
      第２引数： > = より大きい; < = 未満; >= = 以上; <= = 以下
      第３引数：[検索キー1='パターン'●検索キー2='パターン'…]
      第４引数：0 = 基本形; 1 = 出現形
      第５引数：出力したい属性を/で区切って記述
Bccwj::Lex::SortMorph(配列, "{0/1}", "{0/1}")
    ※検索した形態素（書字形，出現形）をカナ形でソートする
      第２引数：0 = 形態素頭から; 1 = 形態素末から
      第３引数：0 = 昇順; 1 = 降順
```

図４ 語彙情報データベースへのアクセス API の例

[4] 検索パターンの入力方式は CQL (corpus query language) の記述に準じたものを採用している。cf. Christ (1994)

語彙情報データベースを参照する API を利用することにより，コーパスからの語彙情報を利用する各種ツールを比較的簡単に構築し，利用することができる (ツールの詳細はデモにゆずる)。§２．でみたように，頻度情報をふくむデータを事前にデータベース化することにより，必要な情報を効率よく得ることができる。

- 語彙リストを BCCWJ の頻度順に並べ替える
- 例文の形態素を BCCWJ の頻度により色分け表示する
- 評価対象にあるコーパス中の形態素の出現頻度を BCCWJ のそれと対照する
- 特定の品詞について，頻度別に一覧表示する

４．終わりに：BCCWJ の応用言語学的活用にむけて

本稿で開発したツール例は，スタンドアロンの分析ツールとして単体のコンピュータ上で動作させることもできるほか (図５参照)，組織内 LAN にある共用の語彙情報データベースにアクセスする形で運用することも考えられる。

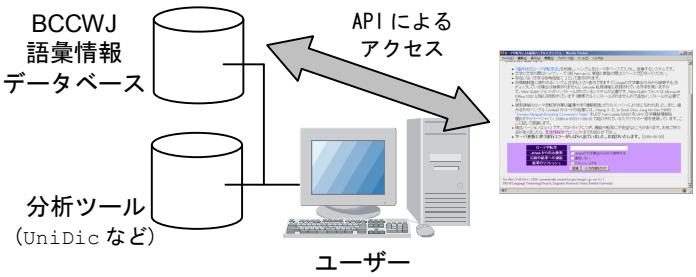


図５　スタンドアロン分析ツールと API

さらに，このような API を HTTP 経由でデータを提供する Web サービスとして実現することにより，Web ブラウザ上で利用できる様々なツールに語彙情報データベースから得られる情報を活用することが可能になる。マッシュアップと呼ばれるこの種の技術は他の分析ツールとの組み合わせを可能にし，幅広い用途にコーパスからの情報を利用することを可能にするだろう (図６)。[5]

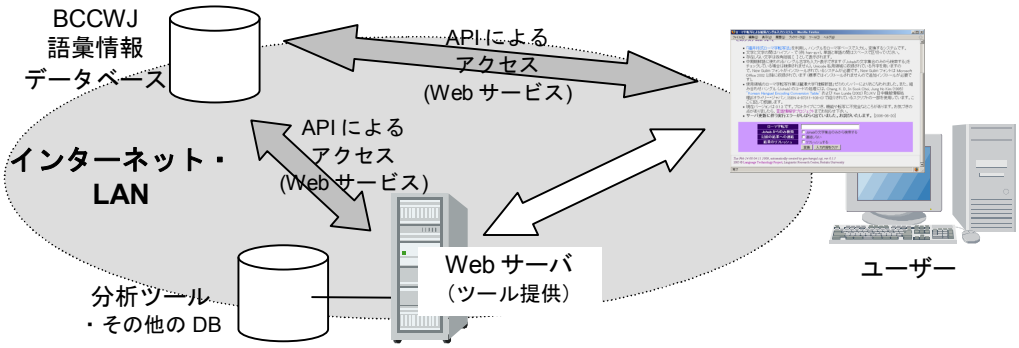


図６　Web サービスと API

[5] ただし，現時点では，オンラインの分析ツールと形態素解析辞書 UniDic との連携については，UniDic の配布条件など検討しなければならない点があることに注意。

本稿で開発しているシステムの今後の課題としては，以下のようなものが挙げられる。

- 語彙情報データベースの拡充と対応する API の開発
  1. 新聞，雑誌，書籍，インターネットなどジャンルによる分類（ないしその下位分類），出典情報などメタデータの付加
  2. 語彙素情報，長単位情報の検索
  3. *n*-gram，コロケーション情報の検索
  4. 本特定領域研究の言語政策班，日本語教育班などが構築している語彙頻度表，語彙表との連携
- `Perl` モジュール以外の形態での API の構築・提供；特に Web サービスとしてのデータ提供方法の試行
- 分析ツールとの連携：形態素解析辞書 `UniDic` 等，各種解析プログラムとのより緊密な連携と，具体的なツールの開発
- BCCWJ の用例検索との連携
- データベース及び API のパッケージ化と公開

BCCWJ の開発は，今後応用分野において，文法，辞書，教材など具体的な形での成果の公刊が期待される。しかしながら，均衡コーパスとしての真価が最も発揮されるのは，その具体的なデータが他の様々なジャンルや内容のコーパスの比較に用いられ，その評価の主たる根拠として積極的に用いられるようになることにある。このような，BCCWJ の応用言語学的利用のためには，BCCWJ の情報を効果的 (= 高速かつ網羅的に) に，かつ信頼できる形で (= よく検証された高精度なデータとして) 提供できるデータベースとその利用のための環境が構築され，安定して提供なければならない。本稿で構築しているようなシステムの試案がその一助となれば幸いである。

文献

Christ, Oliver (1994) “A modular and flexible architecture for an integrated corpus query system,” In *Proceedings of COMPLEX '94. 3rd Conference on Computational Lexicography and Text Research, Budapest, Hungary, July 7-10, 1994*, pp. 23-32.

Foster-Johnson, Eric (1998) *Perl Modules*. New York: M & T Books.

Grönroos, Mickel & Manne Miettinen (2002) “Kielipankki ja `Lemmie`-ohjelmisto,” (フィンランド語言語バンクと `Lemmie` パッケージ) *Puhe ja kieli*, 3, pp. 125-136.

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://www.tokuteicorpus.jp/result/pdf/2006_014.pdf` よりダウンロード可能)

山田篤 (2007) 「数字列への読み付与―`Numtrans` と `ChaOne`―」特定領域「日本語コーパス」平成 18 年度全体会議予稿集.

コーパスを用いた公共性の高い文章における表記改善への視点

斎藤達哉（言語政策班分担者：国立国語研究所研究開発部門）†

Perspective to Improve the Representation of Public Writing

SAITOO Tatuya　(Dept. Lang. Res., National Institute for Japanese Language)

１．はじめに―――本発表の目的

本発表は「日本語コーパス」を分析から得られる知見に基づき，公用文の表記基準の特殊な面を明らかにすることによって，公共性の高い文章における表記の改善への視点を見いだすことを目的としている。

２．公用文の表記基準と公共性の高い文章

日本の中央省庁における公用文の表記基準は，次の文書によって通知されている。

・「公用文作成の要領」

1951 年に国語審議会が建議し, 1952 年に内閣官房長官によって依命通知したもの。1981 年に「常用漢字表」の実施に伴う「一部読替え」が，1986 年に「現代仮名遣い」の実施に伴う「読替え」が行われた。

・「公用文における漢字使用等について」

1981 年に内閣官房長官から通知したもの。

また，文部省では上記の通知に基づいた「文部省用字用語例」作成し，現在も文部科学省に引き継がれている。

こうした公用文の表記基準は，中央省庁においても必ずしも徹底されているとは言えない。その原因としては，

(1) 「読替え」が行われたことによって生じた変更点が十分に理解されていないこと

(2) 副詞・接続語の中に，仮名表記と，漢字表記との 2 種が示されていること

(3) 一般的な表記と異なるために，記述の際に揺れが生じること

などが想定される。

ところで，公用文の表記基準が適用されているのは，通達，通知，依頼，回答といった事務的な文書だけではない。白書，広報など様々な立場の人が目にする文章でも，公用文の表記基準がよりどころとされることになる。こうした公共性の高い文章の中で，一般の社会生活で目にするものとは異なる表記や，表記の揺れが存在した場合，読み手が違和感や迷いを持つ原因にもなる。

３．表記方針の転換

３．１　仮名書き優先から適切な漢字使用へ

「当用漢字表」（1946 年内閣告示，1981 年廃止）は，その「まえがき」に「法令・公用文書・新聞・雑誌および一般社会で，使用する漢字の範囲」を示したものであること，「今日の国民生活の上で，漢字の制限があまり無理なく行われることをめやすとして選んだ」

† tsaito@kokken.go.jp

ことが記された。この姿勢の下，同表の「使用上の注意事項」では，

イ　この表の漢字で書きあらわせないことばは，別のことばにかえるか，または，かな書きにする

ロ　代名詞・副詞・接続詞・感動詞・助動詞・助詞は，なるべく仮名書きにする

と明記されており，漢字制限の色彩が強いものとして受け止められた。

現行の「常用漢字表」(1981)では，「法令，公用文書，新聞，雑誌，放送など，一般の社会生活において，現代の国語を書き表す場合の漢字使用の目安を示すものである」とされ，制限的色彩は弱いものとなっている。そこには，「当用漢字表」に見られたイ項，ロ項などに類する記述は見られない。なお，本稿２．に紹介した「公用文作成の要領」中での「一部読替え」(1981)も，この流れを反映したものである。

公共性の高い媒体である新聞でも，漢字制限への取り組みが積極的に行われてきた[1]。日本新聞協会の「用字用語集」[2]では，前書きに当たる部分に仮名書き優先の姿勢を示す。

・用例の漢字書きになっているものは，ひらがなまたはかたかなで書いてもよい。ただし，ひらがな書きを漢字で書いたり，かたかな書きをひらがなあるいは漢字で書いたりはしない。（1961～1964 年版）

・用語例中，漢字書きにしてあるものは，場合によって，仮名書きにしてもよい。また，平仮名書きのものは片仮名書きにしてもよいが，逆に仮名書きのものを漢字で，片仮名のものを平仮名で書かないことを原則とする。（2007 年版）

さらに，1960 年代後半からは，次に掲げるような具体的な要領が追加された。

副詞（副詞的に使う語を含む）の書き方は，次の要領によった

a　訓読みの副詞は，原則としてかな書きにするが，漢字を使った方が意味の明らかになる場合には，漢字書きにしてもよい

b　音読みの副詞は，原則として漢字書きにするが，その漢字にあまり意味がない場合には，かな書きにする　（1967，1969 年版。引用に当たって語例は省略。）

副詞等の表記については，常用漢字表が告示された年以降も，一部の表現を改めたものの，削除されることなく最新版まで踏襲されている。

一，助詞，助動詞は平仮名で書く。

二，接続詞，感動詞，補助動詞，形式名詞は，できるだけ平仮名で書く。

三，副詞は次のように書く。

(1) 訓読みの副詞のうち，漢字を用いた方が意味の明らかになる場合は漢字書きにしてもよい。

(2) 音読みの副詞は，原則として漢字で書く。ただし，その漢字にあまり意味がない場合には，平仮名書きにする。　（2007 年版。引用に当たって語例はした。）

３．２　仮名書き優先は現在も続いているのか

国立国語研究所(1983)『現代表記のゆれ』 では，1966 年 1 年間に発行された朝日，毎日，読売の 3 紙を対象にして，表記のゆれを調査している。その中には，1966 年当時の

---

[1] 新聞が，漢字制限への取り組みを積極的に行ってきたことは，須藤（2006）にも見える。

[2] 「用字用語集」は，(社)日本新聞協会発行の『新聞用語集』に掲載されている。国立国語研究所図書館が所蔵する 1961,1962,1963,1964,1967,1969,1976,1981,1996,2007 年の各版を調査した。

表外字・表外訓といった漢字制限の影響によるゆれも指摘されるが，それ以外の，日本語表記そのものが持つゆれについては，以下のような特徴を拾うことができる。

まず，漢語では，「漢字で表記できるにもかかわらず，カナで表記される語」は「副詞など形式的な用法の語」「漢語起源の語であるが，はなしことばとして日常よくもちいられ，和語的な語感をもつようになったもの」であることが指摘される(pp.40-42)。和語にも「漢字とヒラガナの対立によるゆれ」が見られ，「副詞の類」「同訓異字の語」が「カナ書きされやすい」と指摘される(p.43)。

また，これに先立って，宮島(1969)は，国立国語研究所の「雑誌九十種の用字用語」調査に基づき，当用漢字で書ける語であっても，副詞，動植物名，日常語化した漢語に仮名書きの傾向が見られることを指摘している。

［表 1］は，宮島が例示した語例について，その後の『現代雑誌 200 万字言語調査語彙表 公開版』（国立国語研究所），BCCWJ 書籍コアデータ・白書コアデータでの各表記形の頻度を付加して整理したものである。

［表 1］を見る限りでは，「今度」「丁寧」「特に」「菊」が，1994 年の現代雑誌 200 万字調査と BCCWJ モニター公開データ(2008 年度版)で，仮名表記から漢字表記へ傾斜の傾向がうかがえる。「菊」は，上記３．１に紹介した「当用漢字表」から「常用漢字表」への流れという外的要因を単純に反映していると思われる。しかし，「今度」「丁寧」「特に」の類は，国立国語研究所(1983)の指摘するような，語の運用のされ方といった内的要因も考慮に入れる必要がある。

［表1］宮島(1969)の示した語と各種調査での表記形の頻度

		雑誌90種（1956年・90誌）	現代雑誌（1994年・70誌）	BCCWJ	
				書籍コア	白書コア
今度	仮名	83	14	3	－
	漢字	139	68	18	－
散々	仮名	10	4	3	－
	漢字	4	4	－	－
随分	仮名	40	32	14	－
	漢字	12	8	3	－
	交ぜ書き	12	2	－	－
精々	仮名	8	5	6	－
	漢字	4	－	－	－
折角	仮名	29	19	6	－
	漢字	14	1	5	－
大分	仮名	17	13	2	－
	漢字	28	4	－	－
丁寧	仮名	11	10	3	－
	漢字	11	18	3	－
特に	仮名	65	88	3	－
	漢字	118	217	34	119
途端	仮名	20	16	4	－
	漢字	8	9	3	－
無論	仮名	25	17	5	－
	漢字	12	2	2	－
菊	仮名	13	1	－	－
	漢字	7	14	－	－

４．現代日本語の表記傾向の把握

４．１　比較に用いるデータ

仮名表記から漢字表記へ傾斜が，ほかの語でも生じているのかについて，以下に調査を試みたい。

［表 2］～［表 5］は，BCCWJ（現代日本語書き言葉コーパス）モニター公開データ(2008 年度版)に収められた「書籍」と「白書」の形態素解析済みコアデータ[3]を用いて，副詞，連体詞，接続詞の表記別の出現頻度を集計したものである。書籍と白書の合計頻度が 10 以上のものを対象とした。

書籍は，一般の社会生活において，目にする機会が多い文章を代表するものとして，白書は，国の機関による公共性の高い文章を代表するものとして位置づけられる。

副詞，連体詞，接続詞を対象としたのは，上記１．の（1）（2）によって混乱の生じた可能性のある品詞だからである。「公用文作成の要領」では，当初，代名詞，副詞，接続

[3] 形態素解析辞書 UniDic-1.3.9（MeCab 版）で解析されたデータで，人手による修正が施されて精度 99%以上としたコアデータ（データ班作成）である。サンプル数，語数（記号等を含む）は，BK：書籍コアデータ・80 サンプル・225,583 語，OW：白書コアデータ・62 サンプル・228,658 語。

詞などは「できるだけかな書きにする」という方針をとっていたが，1981 年の「一部読替え」によって削除されている。また，「公用文における漢字使用等について」では，副詞と接続詞（語）は，原則として漢字で書くものと，原則として仮名で書くものの 2 種が示されている。

なお，［表 2］～［表 5］では，次の項目も参考情報として付加した。

・公用文…「公用文における漢字使用等について」(1981)の 1 の（2）のイ項，オ項に例示された副詞，連体詞，接続語の表記に○印を付した。
・新聞用語（［表 2］［表 3］にだけ掲出。）…日本新聞協会(2007)所収の「用語用字集」の「表記の原則」部分に例示された副詞の表記に○印を付した。
・現代雑誌(1994 年・70 誌)…国立国語研究所(2006)での集計表での出現頻度数。母数は，短単位 26,362 語。完全に一致した語形を表示できない場合は，「－」または参考となる頻度数を括弧内に入れて示した。
・新聞表記のゆれ(1966 年・3 紙)……国立国語研究所(1983)の「付－1 新聞における語表記のゆれ一覧」に示された頻度数。母数は，短単位 34,477 語。完全に一致した語形を表示できない場合は，「－」印または参考となる頻度数を括弧に入れて示した。

４．２　BCCWJ から見た傾向

以下では，〔　〕を付したときは〔語彙素〕として，「　」または無括弧は出現形（表記形）として示す。

４．２．１　副詞

〔音読みの副詞〕

(1) 漢字で表記される傾向にあるのは，次の語である。
・漢字表記しか見られない語…直接，結局，依然，当然，単に，十分，一切，真に
・漢字表記の方が多い語…特に，一層，一番，決して，実に，突然，全然

(2) 〔折角〕は，漢字表記「折角（頻度 5）」と仮名表記「せっかく（頻度 6）」とが均衡しており，表記が揺れていると見られる。

(3) 仮名で表記される傾向にあるのは，次の語である。
・仮名表記の方が多い語
…もちろん，たぶん，いったい，ずいぶん，だいたい，ぜひ，たいてい，いったん
・仮名表記しか見られない語
…かなり，たくさん，きっと，ともかく，とにかく，ちょうど，よほど・よっぽど，だんだん，ついに

(4) 〔一層〕は，書籍と白書とで，表記の傾向が異なる傾向にある。書籍では，仮名表記「いっそう」の方が多く，白書では漢字表記「一層」の方が多い。

〔訓読みの副詞〕

(1) 多くが，仮名で表記される傾向にあるが，その中で，仮名表記の方が多いのは，次の語である。
・仮名表記の方が多い語
…ともに・とともに，よく，さらに，まだ，まず，また，まったく，なぜ，すでに，あまり，いかに，たとえ，おそらく，あらかじめ，ようやく，もはや

［表2］ 音読みの副詞

〔語彙素〕	合計	BK	OW	公用文	新聞用語	原文文字列	BK	OW	現代雑誌（1994年・70誌）	新聞表記のゆれ（1966年・3紙）
特に	156	37	119			とくに	10	0	(88)	408
				○	○	特に	27	119	(217)	93
勿論	68	64	4			もちろん	63	4	248	—
						勿論	1	0	8	—
一層	45	4	41			いっそう	3	1	14	56
						一層	1	40	13	42
						一そう	0	0	—	3
一番	40	39	1			いちばん	8	0	73	77
					○	イチバン	0	0	1	—
						一番	31	1	162	139
						一ばん	0	0	1	4
可成	39	23	16	○	○	かなり	23	16	151	—
沢山	33	32	1		○	たくさん	32	1	120	92
						沢山	0	0	9	8
直接	31	14	17			ちょくせつ	0	0	—	1
						直接	14	17	99	162
急度	28	28	0			きっと	28	0	68	—
*結局	27	25	2			けっきょく	0	0	2	5
					○	結局	25	2	80	114
依然	23	3	20			いぜん	0	0	—	29
						依然	3	20	9	38
多分	22	22	0			たぶん	14	0	18	5
					○	多分	8	0	17	17
一体	21	21	0			いったい	17	0	32	28
						一体	4	0	47	19
当然	21	20	1			とうぜん	0	0	2	—
						当然	20	1	105	—
決して	20	19	1			けっして	4	0	6	8
					○	決して	15	1	56	51
単に	20	16	4			たんに	0	0	(2)	4
						単に	16	4	(60)	36
十分	18	5	13			じゅうぶん	0	0	2	5
						充分	2	0	56	6
					○	十分	3	13	108	257
随分	16	16	0			ずいぶん	12	0	32	21
					○	随分	3	0	8	1
						ずい分	1	0	2	1
*実+に	16	16	0			じつに	4	0	(34)	(9)
				○	○	実に	12	0	(159)	(108)
突然	14	13	1			とつぜん	1	0	2	1
				○		突然	12	1	50	62
兎も角	13	13	0			ともかく	13	0	28	45
						とも角	0	0	—	1
兎に角	12	12	0			とにかく	12	0	119	73
						とに角	0	0	—	2
丁度	12	12	0			ちょうど	12	0	52	47
						ちょーど	0	0	1	—
						丁度	0	0	3	1
*大体	12	12	0			だいたい	8	0	26	15
					○	大体	4	0	9	22
是非	11	11	0		○	ぜひ	8	0	81	93
						ぜヒ	0	0	1	—
						是非	3	0	28	12
折角	11	11	0		○	せっかく	6	0	19	39
						折角	5	0	1	4
大抵	11	11	0			たいてい	9	0	16	17
					○	大抵	2	0	4	3
						大てい	0	0	—	1
余程	11	11	0			よっぽど	1	0	2	—
				○	○	よほど	10	0	11	—
段々	11	11	0			だんだん	11	0	17	—
						段々	0	0	4	—
遂に	10	10	0			ついに	10	0	(48)	89
						遂に	0	0	(7)	9
						終に	0	0	(1)	—
*一切	10	10	0			いっさい	0	0	11	29
					○	一切	10	0	43	64
一旦	10	7	3			いったん	6	2	15	39
						一旦	1	1	5	—
						一たん	0	0	—	2
*真+に	10	2	8		○	真に	2	8	(34)	—

合計が9以下の〔語彙素〕は省略した。
*印は追加分。形態素解析辞書UniDicで付加された品詞が副詞以外のもの。

［表3］ 訓読みの副詞

〔語彙素〕	合計	BK	OW	公用文	新聞用語	原文文字列	BK	OW	現代雑誌（1994年・70誌）	新聞表記のゆれ（1966年・3紙）
*何	590	479	111		○	何	324	111	870	290
						何ん	2	0	2	—
						なに	39	0	104	73
						なにっ	1	0	—	—
						なん	112	0	230	—
						なんに	1	0	—	—
						ナニ	0	0	1	1
						ナン	0	0	3	—
						なン	0	0	1	—
そう	339	320	19			せ	1	0	1	—
						そ	5	0	15	—
						そう	314	19	727	—
						さう	0	0	1	—
						ソウ	0	0	2	—
						そー	0	0	3	—
*とも+に	333	36	297		○	ともに	6	19	(299)	(235)
						共に	5	8	(112)	(37)
						とともに	24	270	—	—
						と共に	1	0	—	—
どう	236	225	11			どう	225	11	546	—
						どお	0	0	1	—
						どー	0	0	6	—
*今	226	221	5			いま	54	0	235	756
						イマ	0	0	1	—
					○	今	167	5	618	199
こう	197	136	61			こう	136	61	346	—
						こー	0	0	1	—
もう	136	135	1			もう	135	1	427	—
						も	0	0	1	—
						もぉ	0	0	1	—
						もー	0	0	1	—
良く	123	120	3			よう	3	0	—	—
						よく	115	2	115	—
						良く	2	1	1	—
最も	112	33	79			もっとも	14	2	50	143
				○	○	最も	19	77	102	160
例えば	89	64	25			たとえば	36	0	126	148
				○	○	例えば	28	25	58	26
更に	88	42	46			さらに	40	24	(349)	573
				○	○	更に	2	22	(29)	15
未だ	79	71	8			いまだ	2	3	16	7
						まだ	69	4	219	391
						マダ	0	0	—	1
						未	0	0	—	3
						未だ	0	1	7	2
少し	77	77	0			すこし	3	0	11	15
				○		少し	74	0	267	169
先ず	77	62	15			まず	61	14	247	395
						先ず	1	1	3	1
又	75	74	1			また	71	1	739	1312
						亦	1	0	1	—
						又	2	0	28	8
初めて	69	56	13			はじめて	19	0	35	167
				○		初めて	37	13	186	131
						初メテ	0	0	—	1
						始めて	0	0	2	2
矢張り	67	65	2			やっぱ	2	0	10	—
						やっぱり	10	0	118	—
				○	○	やはり	53	2	150	—
						ヤッパ	0	0	1	—
より	66	14	52			より	14	52	149	—
若し	64	64	0			もし	64	0	82	—
						若し	0	0	1	—
直ぐ	62	62	0			すぐ	62	0	238	219
						スグ	0	0	8	8
						即ぐ	0	0	1	—
						直ぐ	0	0	2	—
必ず	60	46	14			かならず	3	0	9	10
				○	○	必ず	43	14	158	113
						必らず	0	0	2	1
全く	59	57	2			まったく	33	0	133	103
						マッタク	0	0	—	1
				○	○	全く	24	2	61	120
						ったく	0	0	2	—
何故	53	52	1			なぜ	44	1	115	111
					○	何故	8	0	12	4
						ナゼ	0	0	1	—
詰まり	51	50	1			つまり	50	1	146	—
既に	49	39	10			すでに	32	2	119	303
				○	○	既に	7	8	21	11
もっと	48	48	0			もっと	48	0	184	—
一寸	48	48	0			ちと	1	0	—	—
						ちょっと	46	0	284	143
						チョット	1	0	—	—
						ちょいと	0	0	1	—
						チョッと	0	0	—	1
						一寸	0	0	1	2
唯	40	39	1			ただ	31	1	204	112
						たった	8	0	37	40
						タッタ	0	0	1	—
						タダ	0	0	6	1
						唯	0	0	5	2
						只	0	0	—	1
						たゞ	0	0	—	6
迚も	39	37	2			とっても	3	0	21	—
						とても	34	2	168	—
余り	39	26	13			あまり	21	13	187	(186)
						あんまり	4	0	21	—
				○		余り	1	0	27	(35)
						余	0	0	11	(71)

〔語彙素〕	合計	BK	OW	公用文	新聞用語	原文文字列	BK	OW	現代雑誌（1994年・70誌）	新聞表記のゆれ（1966年・3紙）
						きわめて	14	1	26	112
極めて	38	17	21			キワメテ	0	0	―	1
				○	○	極めて	3	20	10	1
						ずっと	35	0	―	―
ずっと	35	35	0			ずうっと	0	0	1	―
如何に	35	33	2			いかに	31	2	―	(137)
						如何に	2	0	―	(2)
					○	まさに	33	0	97	35
正に	33	33	0			正に	0	0	5	2
						応に	0	0	―	2
次いで	33	1	32	○	○	次いで	1	32	―	7
						ついで	0	0	―	53
						ふたたび	2	0	―	22
再び	32	23	9			再び	0	0	―	1
				○		再びび	21	9	12	163
					○	なかなか	28	2	56	―
中々	30	28	2			中々	0	0	1	―
						仲仲	0	0	1	―
暫く	29	28	1			しばらく	28	1	48	51
						暫く	0	0	1	1
元々	29	28	1			もともと	28	1	51	―
						元々	0	0	6	―
嘗て	28	25	3			かって	1	0	1	―
						かつて	24	3	73	―
はっきり	26	26	0			はっきり	26	0	―	177
						ハッキリ	0	0	―	3
				○		互いに	19	4	(54)	(59)
*互い+に	25	21	4			互に	0	0	―	(3)
						たがいに	1	0	(4)	(12)
						おたがいに	1	0	―	―
寧ろ	24	21	3			むしろ	21	3	56	―
略	24	8	16			ほぼ	8	16	57	―
仮令	23	23	0			たとえ	22	0	53	―
						例え	1	0	1	―
幾ら	20	20	0			いくら	19	0	70	(65)
						いっくら	1	0	―	―
						幾ら	0	0	2	1
僅か	20	16	4			わずか	16	4	64	―
						僅か	0	0	5	―
やや	20	6	14			やや	6	14	―	―
						稍	0	0	62	―
						すくなくとも	0	0	(1)	1
*少なく+と+も	20	15	5	○		少なくとも	15	5	(22)	33
						少くとも	0	0	(2)	1
						まぁ	1	0	―	―
まあ	18	18	0			まあ	17	0	―	29
						マア	0	0	―	1
						まア	0	0	―	1
確り	18	18	0			しっかり	18	0	133	63
						シッカリ	0	0	2	―
						確り	0	0	―	1
ちゃんと	17	17	0			ちゃんと	17	0	―	―
恐らく	17	17	0			おそらく	13	0	34	44
				○	○	恐らく	4	0	2	4
色々	16	16	0			いろいろ	16	0	151	157
						イロイロ	0	0	1	―
						色々	0	0	15	5
猶	16	14	2			なお	14	2	100	362
						尚	0	0	8	3
						猶	0	0	1	―
予め	16	11	5			あらかじめ	8	5	19	―
						予め	3	0	4	―
*常+に	16	11	5			つねに	3	0	(12)	(36)
				○		常に	8	5	(58)	(40)
丸で	15	15	0		○	まるで	15	0	72	―
						丸で	0	0	1	―
きちんと	15	15	0			きちっと	1	0	―	―
						きちんと	14	0	―	10
						キチンと	0	0	―	2
改めて	15	13	2			あらためて	7	0	11	―
						改めて	6	2	12	―
聴て	13	12	1			やがて	12	1	52	―
益々	14	10	4			ますます	10	4	47	84
						益々	0	0	4	2
すっかり	12	12	0			すっかり	12	0	―	―
行成	12	12	0			いきなり	12	0	35	―
						イキナリ	0	0	3	―
漸く	12	12	0			ようやく	11	0	49	―
						漸く	1	0	3	―
流石	11	11	0			さすが	11	0	62	―
						流石	0	0	1	―
じっと	11	11	0			じっと	11	0	―	14
						ジット	0	0	―	1
極く	11	10	1			ごく	10	1	36	―
						極	0	0	5	―
ふと	10	10	0			フッと	2	0	―	―
				○		ふと	8	0	24	13
						フト	0	0	―	1
最早	10	10	0			もはや	9	0	39	―
						最早	1	0	1	―
態々	10	10	0			わざわざ	10	0	19	―
未だ未だ	10	10	0			まだまだ	10	0	37	―
ゆっくり	10	9	1			ゆっくり	9	1	―	―
同じく	10	8	2			同じく	8	2	―	―
						同ジク	0	0	1	―
						おなじく	0	0	1	―
直ちに	10	7	3			ただちに	3	0	―	39
				○	○	直ちに	4	3	―	49
						直ニ	0	0	―	2

合計が9以下の〔語彙素〕は省略した。
*印は追加分。形態素解析辞書UniDicで付加された品詞が副詞以外のもの。

［表4］ 連体詞

〔語彙素〕	合計	BK	OW	公用文	原文文字列	BK	OW	現代雑誌（1994年・70誌）	新聞表記のゆれ（1966年・3紙）
					その	879	387	2671	3368
其の	1268	881	387		そん	0	0	1	―
					其	2	0	5	1
					其の	0	0	1	―
					この	646	325	2237	4600
此の	973	648	325		コノ	1	0	―	―
					此	1	0	―	―
					此の	0	0	2	1
					わが	15	42	64	378
我が	208	19	189	○	我が	4	147	34	9
					我	0	0	4	1
					吾が	0	0	3	1
					吾	0	0	―	1
					おなじ	3	0	6	6
同じ	133	114	19		おんなじ	0	0	1	―
					同じ	111	19	487	407
					（おおきい・な）	0	0	(1)	―
大きな	127	79	48		巨きな	0	0	(1)	1
				○	大きな	79	48	(598)	336
そんな	101	101	0		そんな	101	0	368	―
					んな	0	0	1	―
何の	84	50	34		どの	50	34	136	―
或る	82	69	13		ある	68	13	169	―
					或る	1	0	3	―
					あの	79	0	215	―
彼の	82	82	0		アノ	1	0	1	―
					あん	0	0	1	―
					かの	2	0	―	―
こんな	62	62	0		こんな	62	0	285	―
小さな	59	56	3		ちいさな	1	0	1	2
				○	小さな	55	3	111	75
どんな	55	55	0		どんな	55	0	206	―
所謂	46	20	26		いわゆる	20	26	69	―
主な	24	7	17		主な	7	17	―	―
あらゆる	19	10	9		あらゆる	10	9	74	―
単なる	14	11	3		たんなる	0	0	(2)	5
					単なる	11	3	(60)	34

合計が9以下の〔語彙素〕は省略した。
*印は，追加分。形態素解析辞書UniDicで付加された品詞が連体詞以外のもの。

［表5］ 接続詞

〔語彙素〕	合計	BK	OW	公用文	原文文字列	BK	OW	現代雑誌（1994年・70誌）	新聞表記のゆれ（1966年・3紙）
				○	また	114	487	739	1312
又	690	119	571		又	5	84	28	8
					亦	0	0	1	―
					および	14	13	51	250
及び	637	18	619	○	及び	4	606	59	41
					及	0	0	4	3
					しかし	179	56	457	―
然し	235	179	56		然かし	0	0	1	―
					然し	0	0	3	―
そして	202	196	6		そして	196	6	629	―
					そし	0	0	1	―
更に	170	42	128		さらに	41	119	349	573
					更に	1	9	29	15
且つ	117	11	106	○	かつ	11	105	35	33
					且つ	0	1	2	1
					なお	8	65	100	362
猶	73	8	65		尚	0	0	8	3
					猶	0	0	1	―
					あるいは	50	15	97	137
或いは	71	54	17		或いは	4	2	2	1
					或は	0	0	1	2
*ところ＋が	46	45	1	○	ところが	45	1	(166)	―
一方	42	16	26		いっぽう	1	0	11	4
					一方	15	26	97	372
けれど	34	34	0		けど	1	0	―	―
					けれど	33	0	―	―
					ただ	33	1	204	112
唯	34	33	1		たゞ	0	0	―	6
					タダ	0	0	6	1
					唯	0	0	5	2
					只	0	0	―	1
但し	30	11	19	○	ただし	11	18	109	50
					但し	0	1	19	8
扨	30	30	0		さて	30	0	73	52
					サテ	0	0	―	1
然も	29	29	0		しかも	29	0	200	―
*したがっ＋て	28	20	8	○	したがって	17	8	―	112
					従って	3	0	―	27
即ち	26	20	6		すなわち	19	6	29	42
					則ち	1	0	1	1
若しくは	16	8	8		もしくは	6	1	23	6
				○	若くは	2	7	―	3
*ところ＋で	11	11	0	○	ところで	11	0	(166)	―
尤も	10	10	0		もっとも	10	0	31	―

合計が9以下の〔語彙素〕は省略した。
*印は，追加分。形態素解析辞書UniDicで付加された品詞が接続詞以外のもの。

(2)〔改めて〕は，仮名表記「あらためて（頻度 7）」と漢字表記「改めて（頻度 8）」とが均衡しており，表記が揺れていると見られる。
(3) 漢字で表記される傾向にあるのは，次の語である。
・漢字表記の方が多い語
…何，今，最も，例えば，少し，初めて，必ず，再び，互いに，常に，直ちに
・漢字表記しか見られない語…次いで，少なくとも，同じく
(4)〔既に〕は，書籍と白書とで，表記が異なる傾向にある。書籍では仮名表記「すでに」が多く，白書では漢字表記「既に」が多い。
(5)〔更に〕は，白書の中で，仮名「さらに（頻度 22）」と漢字表記「更に（頻度 24）」とが均衡している。公用文の表記基準では，〔更に〕は副詞のときは漢字表記「更に」，接続詞のときは仮名表記「さらに」となっているが，BCCWJ からは，副詞のときも仮名表記に傾斜していることがわかる。

〔音読みの副詞〕の（3），〔訓読みの副詞〕の（1）の各語に共通するのは，白書での頻度が書籍に比べて少ないことである。特に前者では，不確定な表現，伝聞，疑問，推量，勧誘，希望などの表現が多く見られることから，白書での用例が少ないと思われる[4]。

また，白書での漢字表記「一層」「既に」は，公用文の基準に沿ったものであるが，社会に多く流通している表記とは，ずれが生じている。

副詞の表記について，1699 年の新聞，1994 年の現代雑誌と見比べると，以下の傾向が見られる。
・漢字表記へと傾いているもの…〔特に〕〔例えば〕〔初めて〕
・仮名表記へと傾いているもの…〔多分〕〔大体〕[5]〔全く〕

４．２．２　連体詞

(1) 仮名で書かれる傾向にあるのは，次の語である。
・この，その，どの，あの，こんな，そんな，どんな，いわゆる，あらゆる，ある
(2) 漢字で表記される傾向にあるのは次の語である。これらは，書籍での使用も多く，漢字表記が一般にも定着していると考えられる。
・単なる（音読み），　同じ，大きな，小さな，主な（訓読み）
(3)〔我が〕は，書籍と白書では，表記が異なる傾向にある。書籍では仮名表記「わが」が多く，白書では仮名表記「我が」が多い。なお，白書では，189 例すべてが「我が国」という用例であり，使用され方に偏りが見られる。

連体詞の表記について 1699 年の新聞，1994 年の現代雑誌と見比べても，〔単なる〕〔同じ〕〔大きな〕〔小さな〕〔この〕〔その〕などは，同様の傾向に見える。

〔我が〕も書籍では仮名表記「わが」が多い点で 1699 年の新聞，1994 年の現代雑誌と

---

4 〔可成り〕が白書で使われるときは，次のような「選択肢」の引用の例が含まれている。
◇「家事を半分以上している」は「すべて自分がしている」，「かなりの家事を自分でしている」，「半分程度の家事を自分でしている」と回答した人の割合の合計。（白書）

5 〔大体〕は，仮名書きの方が多いが，「大部分」の意味で使われるとき漢字表記となっている。
◇「1 つのものを，だいたい 1 分くらいで包装するつもりでやってください。…」（書籍）
◇「…ちゃんと帰ってくるの？　私はもういや，だいたいあの人は」（書籍）
◇「じきに部長が見える。大体話してあるからね」（書籍）

同様である。ただし，白書において，公用文の表記基準の影響によって，漢字表記「我が」が多くなっており，社会に多く流通している表記とは，ずれが生じている。

４．２．３　接続詞

(1) 仮名で表記される傾向があるのは，次の語である。
　・また，さらに，なお，したがって，すなわち
(2) 漢字で表記される傾向にあるのは，次の語である。
　・一方
(3) 〔及び〕〔若しくは〕は，書籍と白書では，表記が異なる傾向にある。書籍では仮名表記「および」「もしくは」が多く，白書では漢字表記「及び」「若しくは」が多い。

接続詞の表記について 1699 年の新聞，1994 年の現代雑誌と見比べても，〔又〕〔更に〕〔猶〕〔一方〕〔若しくは〕は，書籍においては，揺れの状況は変わらない。

なお，〔又〕は，白書では以下のように，「また」と「又は」との区別をもって使われる傾向にある。これは，「公用文における漢字使用等について」に則った表記である。

◇…学校，試験研究所，事業者，事業者の団体又は学識経験を有する者に対し，…
◇申請は，関係都道府県のいずれか一つの知事に対してされなければならない。また，審査会等は申請があった事件が…　（いずれも，白書）

５．まとめ―――公用文表記の改善への視点

以上，BCCWJ の書籍と白書での頻度が 10 以上の語彙素に絞って検討してきた。白書の表記のずれと揺れについて再度まとめると，以下のようになる。

(1) 書籍では仮名表記が主流であっても，白書では漢字表記のする語がある。〔一層〕〔既に〕〔我が〕〔及び〕〔若しくは〕の類である。
(2) 書籍では仮名表記なのに，白書でだけ，仮名表記漢字表記との揺れが見られる語がある。副詞の〔さらに〕がそれである。

いずれも，公用文の表記基準に準じて，漢字表記を取り入れているがために生じたものである。公用文の基準は，副詞，連体詞，接続詞の表記において漢字表記に偏っていることがある。公用書そのものを作成する場合はともかく，公共性の高い文章を作成する際には，これらを念頭に置けば，読み手に配慮した判断の助けとなるであろう。

参考文献等

国立国語研究所(1983)『現代表記のゆれ』，秀英出版
国立国語研究所(2006)『『現代雑誌の語彙調査』に基づく表記一覧』，国立国語研究所
須藤久士(2006)「新聞と常用漢字の歴史的考察　未完の日本語―「世の中」と教育のはざまで」，『朝日総研リポート AIR21』196，pp.2-34，朝日新聞社総合研究本部
日本新聞協会(2007)『新聞用語集』，日本新聞協会
宮島達夫(1969)「近代日本語における漢語の位置」，『教育 国語』16，pp.17-44，麦書房

公用文の表記基準，施策関連資料

文化庁(2001)『公用文の書き表し方の基準（資料集）』増補二版，第一法規
文化庁国語施策情報システム，http://www.bunka.go.jp/kokugo/

付記：データの集計・整理に当たっては，宮嵜由美さん（専修大学大学院）の協力を得た。

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―

近藤明日子（言語政策班連携研究者：国立国語研究所研究開発部門）†

Extraction and a Study of Subject-Specific Vocabulary in Textbooks for Japanese Junior High Schools:
In Comparison with the BCCWJ Vocabulary

KONDO Asuko (Dept. Lang. Res., National Institute for Japanese Language)

１．はじめに

本稿筆者は、学校教育での言語活動の充実を図るための基礎的な資料とすべく、中学校教科書の教科特徴語の抽出について試行してきた（近藤、2008a；近藤、2008b）。教科特徴語とは、当該教科において特徴的に高頻度に出現する語彙のことであり、当該教科の語彙指導において重要な位置を占める教科特徴語のリスト化は、言語活動の充実に欠かせないものと考える。

本稿は、これまでの試行の成果を発展させ、「教科書コーパス」（後述）と『現代日本語書き言葉均衡コーパス』（BCCWJ）とを利用して、中学校教科書の全 8 教科の特徴語の抽出を試みるものである。さらに、語彙指導において教科特徴語を利用する際に留意すべきと考えられるいくつかの観点から教科特徴語について考察する。

２．中学校教科書の教科特徴語の抽出

２．１．コーパスの用意と語彙調査

中学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。

対象コーパスは、特定領域研究「日本語コーパス」言語政策班が構築している「教科書コーパス」から抽出した。「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る（田中、2008a）。その「教科書コーパス」から中学校教科書部分を抽出し、さらに「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」の 8 種のコーパスに分割し、それぞれを各教科の対象コーパスとして用意した。また、参照コーパスとして、「BCCWJ 領域内公開データ（2008 年度版）」から書籍サンプル（plain text 版、計 13587 サンプル）を用意した。

これら 8 種の対象コーパスと 1 種の参照コーパスの形態素解析は、形態素解析器 MeCab 0.97[1]、形態素解析辞書 UniDic-1.3.9（MeCab 版）[2]を用いておこなった。ただし、短単位[3]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

† kondo@kokken.go.jp
[1] http://mecab.sourceforge.net/
[2] http://download.unidic.org/
[3] 短単位の規定の詳細は小椋・小磯・冨士池・原（2008）を参照のこと。

／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[4]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た。

形態素解析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のうち「語彙素読み」「語彙素表記」「語義」「品詞」「活用型」を用い、これらの値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た[5]。そして、UniDic の付与する品詞属性の大分類が名詞・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・接尾辞の見出し語と、短単位連続から再構成したものからなる見出し語を考察対象とした。

その結果、対象コーパスと参照コーパスの延べ語数・異なり語数は表 1 のようになった。

表 1　対象コーパス・参照コーパスの延べ語数・異なり語数

		延べ語数	異なり語数
対象コーパス	国語	96,264	14,381
	数学	47,406	2,019
	理科	60,115	5,141
	社会	88,290	13,616
	外国語	11,799	1,969
	技術家庭	57,457	7,745
	芸術	32,288	6,161
	保健体育	21,301	4,038
参照コーパス		20,060,940	774,233

２．２．教科特徴語の抽出

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」）を数値化し、その値の大きい語を教科特徴語として抽出する。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）を用いた。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）。

$$G^2=2(alna+blnb+clnc+dlnd-(a+b)ln(a+b)-(a+c)ln(a+c)-(b+d)ln(b+d)-(c+d)ln(c+d)+(a+b+c+d)ln(a+b+c+d)) \quad \cdots[1]$$ [6]

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

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 の特徴度とした。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。本稿ではこの特徴度が 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、各教科の特徴語として抽出した（高見、2003）。抽出された各教科の特徴語の異なり語数・延べ語数を表２に示す。

---

[4] 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 単位として再構成した。ただし、短単位連続の先頭に品詞属性値が「接尾辞」で始まる単位は来ないものとした。

[5] 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。また、合成名詞に近似するものとして再構成した単位の属性値には、構成前の短単位の属性値を結合したものを用いた。

[6] 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した（高見、2003）。

３．中学校教科書の教科特徴語の考察

３．１．当該教科語彙との比較

２において抽出した教科特徴語について、語彙指導での利用の際、留意すべきと考えられるいくつかの観点から分析し、各教科における教科特徴語の性質を考察する。

はじめに、各教科の特徴語が当該教科全体の語彙をどの程度カバーするのか見てゆく。各教科の総語数（表１）に対する教科特徴語の語数（表２）の割合を異なり語数・述べ語数ベースそれぞれで示したものが図１である。

表 2　教科特徴語の延べ語数・異なり語数

	延べ語数	異なり語数
国語	37,092	1,800
数学	38,097	882
理科	40,300	2,033
社会	50,795	4,009
外国語	7,423	566
技術家庭	37,666	2,768
芸術	18,120	1,959
保健体育	11,738	1,186

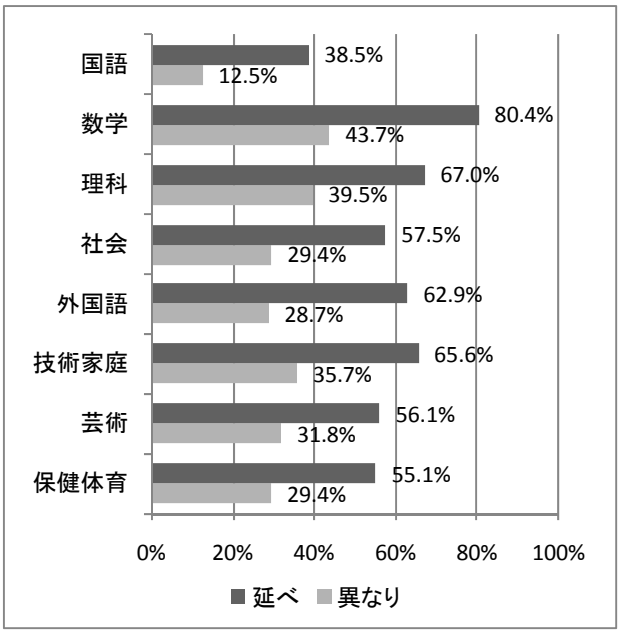


図 1　教科語彙に対する教科特徴語の割合

これを見ると、教科特徴語が教科全体の語彙をどの程度カバーするのかは、教科によって差があることがわかる。数学のように教科特徴語の占める割合が高い教科は、教科特徴語の指導が教科全体の語彙指導において大きな位置を占めるものと考えられる。一方、国語のように教科特徴語の占める割合の低い教科では、教科全体の語彙指導において、教科特徴語だけでなく非教科特徴語にも十分な目配りが必要と考えられる。なお、他教科と比べて国語で教科特徴語の割合が特に低いのは、国語教科書では小説や論説文などの教材部分が多くを占めており、その語彙のありようはBCCWJ書籍サンプルに近い性質を持つため、そこに出現する語彙が教科特徴語として抽出されにくかったものと考えられる。ゆえに国語については、教科書を教材部分と非教材部分に分けて、それぞれ特徴語を抽出することが必要であろう。今後の課題としたい。

３．２．他教科の教科特徴語との比較

次に、各教科の特徴語が他教科でも特徴語となる程度について見てゆく。他教科との関連を視野に入れた、より広がりのある語彙指導を行うためには、当該教科だけでなく他教科での語彙のありようを把握することが重要と考える。

表３は各教科の特徴語について、特徴語となる他教科数ごとに異なり語数を示したものである。図２は表３にもとづき、他教科数を0・1・2～7の3段階に分け、各段階に属する異なり語数の占める割合を示したもの

表 3　特徴語となる他教科数別の教科特徴語の異なり語数

	特徴語となる他教科数							
	0	1	2	3	4	5	6	7
国語	1,417	207	73	41	30	15	9	8
数学	617	138	46	32	21	12	8	8
理科	1,495	325	94	52	33	17	9	8
社会	3,616	234	72	36	23	12	8	8
外国語	396	98	23	19	9	8	5	8
技術家庭	2,243	301	103	57	32	15	9	8
芸術	1,699	145	50	23	16	10	8	8
保健体育	894	142	73	28	21	13	7	8

である。

ここからまず、すべての教科において他教科数 0 の語の割合が圧倒的に高いことが指摘できる。つまり、教科特徴語の多くは、指導の機会が当該教科のみに限られ、他教科では教科特徴語として指導する機会はないことになる。ただし、教科ごとに他教科数 0 の語の割合には相違があり、社会・芸術のような特に割合の高い教科から、数学・外国語のような比較的割合の低い教科まで幅がある。他教科数 0 のような、他教科での指導の機会の望めない特徴語については、当該教科での指導に重点をおく必要があり、一方、他教科数の多い特徴語は、他教科との関連性を念頭においた指導が見込まれると言える。

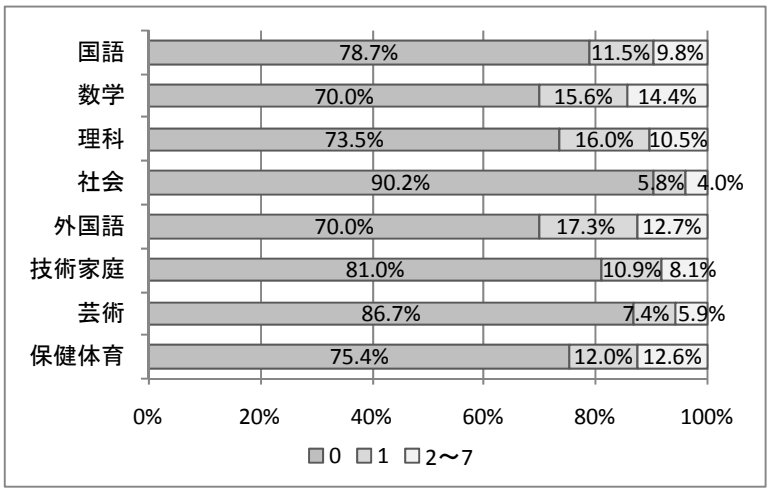


図 2　特徴語となる他教科数別の教科特徴語の割合（異なり）

３．３．BCCWJ 書籍サンプルの語彙との比較

次に、各教科の特徴語が、語彙を学習する生徒にとって馴染みのある語で構成されている程度について見てゆく。語彙指導において、対象語彙がどの程度生徒に馴染みのある語であるのか把握することは、重要なことと考える。

生徒に馴染みのある語を「生徒が接触し得る一般的な書き言葉の中で、高頻度・広範囲に用いられる語」（以下、「一般語」）と定義し、教科特徴語の中で一般語の占める割合を教科別に見てゆく。本稿では、「一般的な書き言葉」として、教科特徴語抽出の際に参照コーパスとして利用した BCCWJ 書籍サンプルを選択し、そこに出現する語の中で「①度数が 34 以上」で、かつ「②散布度が 0.3 より大きい」語を一般語として抽出する。条件①については、BCCWJ 書籍サンプルにおいて、度数 34 以上の語は異なり語数で 32,185 語、延べ語数で 18,073,564 語であり、延べ語数ベースでは書籍サンプル全体の語数（表１）の 90%を占める点に着目し、条件の基準として利用した[7]。ただし、条件①を満たす高頻度な語であっても出現するジャンルに偏りがある語は一般語とは見なしがたいと考え、条件②も設けた。条件②にいう散布度とは、コーパスを複数のジャンルに分割した場合に、当該語の出現がジャンル上に散らばっている程度を表す数値である。本稿では Juilland's D と呼ばれる統計値を散布度として用いた。コーパスに出現する単語 W の Juilland's D は次の式[2]によって求めることができる（Oaks、1998、pp.189-190；上田、1998、pp.44-46）。

$$D = 1 - \frac{SD}{M\sqrt{n-1}} \qquad \cdots[2]$$

[7] BCCWJ での度数を利用した語彙のレベル分けについては田中（2008b）を参考にした。

SD　：単語 W の各ジャンルでの使用率の標準偏差
M　：単語 W の各ジャンルでの使用率の平均
n　：コーパスのジャンル数

Juilland's D は 0 以上 1 以下の値をとり、単語 W の出現が広範囲のジャンルに散らばっている（特定のジャンルに偏らない）程度が高いほどその値は大きくなる。Juilland's D を算出するためにはコーパスにジャンル区分を設定する必要があるが、BCCWJ 書籍サンプルでは、サンプルに付された日本十進分類法（NDC）の一次区分の番号に従って 0～9 の 10 ジャンルに区分し、散布度を算出した。

その結果、条件①と条件②を満たす一般語は、異なり語数で 28,396 語となった。この一般語が各教科の教科特徴語に占める割合を異なり語数ベースで示したものが図３である。

これを見ると、外国語のように一般語の割合の比較的高い教科から、芸術のように一般語の割合の低い教科まで、教科によって差があることが分かる。一般語の割合が低い教科の特徴語は、生徒にとって馴染みのない語が多く含まれていることになり、その点に留意して語彙指導を行う必要があると考える。

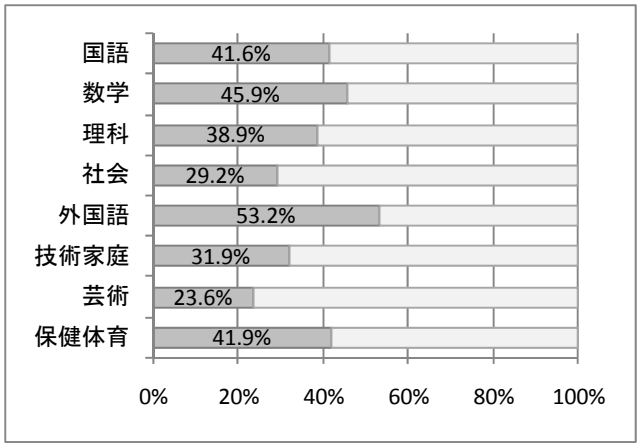


図 3　教科特徴語に対する一般語の割合（異なり）

４．おわりに

以上、「教科書コーパス」と BCCWJ を利用して、中学校教科書の教科特徴語を抽出し、その性質について考察した。各教科の語彙指導で利用しやすい教科特徴語のリスト作成に向けて、今後リストに掲載すべき情報の取捨選択を行うことになるが、その際は本稿での考察結果も反映させたいと考えている。稿末に、現時点での教科特徴語リストの試作版（抜粋）を掲載したので参照されたい。

なお、特徴語の抽出方法や一般語の定義などについては今後さらに検討が必要である。また、語彙指導上重要と考えられるその他の観点の有無についても検討も重ね、教科特徴語リストの完成を目指したい。

文献

上田博人（1998）　『パソコンによる外国語研究（Ｉ）　数値データの処理』　くろしお出版

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・原裕（2008）　『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　『現代日本語書き言葉均衡コーパス』形態論情報規定集』

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
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（2008a）　「教科書コーパスの基本設計」　『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.81-86
田中牧郎（2008b）　「語彙レベルの設定」　『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.7-12
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）
Michael Oakes（1998）　*Statistics for Corpus Linguistics*　Edinburgh University Press

付表　社会の教科特徴語リスト試作版（抜粋）

特徴語（ヨミ）	度数 使用率(‰)	一般語	特徴語となる他教科
魏（ギ）	5 (0.05663)	★	
紀伊国（キイコク）	2 (0.02265)		
生糸（キイト）	6 (0.06796)		
生糸生産（キイトセイサン）	1 (0.01133)		
生糸作り（キイトツクリ）	1 (0.01133)		
議員（ギイン）	18 (0.20387)	★	
議院（ギイン）	4 (0.04531)		
議員数（ギインスウ）	2 (0.02265)		
議員定数訴訟（ギインテイスウソショウ）	1 (0.01133)		
議院内閣制（ギインナイカクセイ）	8 (0.09061)		
気温（キオン）	10 (0.11326)	★	数学・理科・保健体育
祇園祭り（ギオンマツリ）	4 (0.04531)		
紀元前後（キゲンゼンゴ）	3 (0.03398)		
気候（キコウ）	61 (0.6909)	★	
気候学者（キコウガクシャ）	1 (0.01133)		
気候記録（キコウキロク）	1 (0.01133)		
気候区分（キコウクブン）	3 (0.03398)		
気候区分図（キコウクブンズ）	1 (0.01133)		
気候帯（キコウタイ）	4 (0.04531)		
技術（ギジュツ）	26 (0.29448)	★	技術家庭
技術型（ギジュツガタ）	1 (0.01133)		
技術協力（ギジュツキョウリョク）	3 (0.03398)		
技術力（ギジュツリョク）	4 (0.04531)	★	
基準（キジュン）	18 (0.20387)	★	数学

白書および Yahoo!知恵袋を対象にした結合価の自動抽出
—格助詞パターンに着目して—

荻野孝野（辞書編集班協力者：日本システムアプリケーション）†

The Valence Patterns of Japanese Verbs Extracted From *BCCWJ "Japanese Corpus"*

Takano Ogino (Japan System Application CO.,LTD)

1　はじめに

結合価のうち、ここでは格助詞パターンに着目して調査結果について述べる。荻野は格助詞パターンについて、本研究以前にも、EDR 共起辞書（日本電子化辞書研究所　1995）から格助詞パターンの抽出を行ってきた（荻野 2006）。EDR 共起辞書の元データは新聞などからなる EDR コーパスである。本報告では、これらのデータとも比較しながら、特定領域研究「日本語コーパス」で公開された白書や Yahoo!知恵袋データを対象とした格助詞パターンの調査結果について述べる。

2　結合価自動抽出の手順

2.1　対象データ

日本語均衡コーパスのうち、白書と Yahoo!知恵袋を対象として構文解析し、係り受け関係を自動的に付与したデータを利用した。これは、辞書ツール班奥村学氏が「語義タグ付けコーパスの半自動化」（奥村学他　2007）において、辞書の自動的な語義選択システム開発のために CaboCha にて構文解析して作成したデータを利用させていただいたものである。

2.2　対象データからの格助詞パターンの抽出

以下の手順にて動詞ごとに格助詞組み合わせを抽出した。

① 格助詞および格助詞相当語が後続する連用修飾語の抽出

同じ動詞にかかっている格助詞および格助詞相当語を含む連用修飾語を抽出した。以後、格助詞および格助詞相当語で導かれる連用修飾語部分を本報告では「格成分」とする。

② 格成分部分の所定位置への配置

①で認識した格成分について、あらかじめ格助詞の位置を固定した表に、該当する格成分部分を配置し、表の中にはそれらの格助詞に導かれる体言部分を記載した(表 2-1)。

ここで、固定した格助詞部分の列は、「は、が、を、に、と、へ、まで、から、より、で」の１０個の格助詞[注1)]とし、格助詞相当語などはその他の列に配置したものである。

なお、動詞部分が受身形や使役形のものは今回の作業では対象外とした。

③ 格助詞パターンの抽出

②で作成した表から、格助詞パターンを作成した。これは②の表の所定の位置の格助詞部分に体言が配置されているかどうかを判断し、その格助詞を抽出し、その組み合わせで格助詞パターンを作成したものである。

---

†ogino@jsa.co.jp

注1) 係助詞「は」を含める。

表 2-1　格成分部分への配置

動詞出現形	動詞基本形	品詞	は	が	を	に		文
向け	向ける	動詞-自立				回復	(途中列省略)	53年度中にみられた内外均衡回復に向けて
生じ	生じる	動詞-自立	動き	それぞれ				回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。

格助詞については、以下の２段階に分けて、格助詞パターンを抽出した。

(a)格助詞レベルからなるパターン：

出現した動詞ごとに、格助詞を抽出し、格助詞組み合わせ順にソートし、その異なりを格助詞パターンとしたものである。

(b)格助詞相当語を含むレベル：

本研究では、複合的な格助詞相当語「について、に関して」なども格助詞番号に置き換え、異なりパターンを作成した。なお、格助詞相当語については、表 2-2 のグループ化という列で示すように、「という、と言う、といった」のようなものは同じ格助詞相当語グループとしてパターン化を行なって番号付けを行った。表 2-2 の「グループ化」の列で同じ番号のものが、パターン化のときは同じ格助詞相当語として位置づけられる。

表 2-2　格助詞相当語のグループ化

統合番号(格助詞グループ_グループ内番号)	格助詞グループ	グループ内番号	格助詞
36_1	36	1	について
36_2	36	2	につき
36_3	36	3	につきまして
37_1	37	1	につけ
38_1	38	1	につれ
38_2	38	2	につれて
39_1	39	1	にとって
39_2	39	2	にとり

3　格助詞パターン

3.1　格助詞パターンの概要

格助詞パターンは、2.2 で述べたように、格助詞に限定してパターン化したものと、格助詞相当語を入れてパターン化したものとの２段階で行った。

格助詞パターンの数量的概要は以下のようになった。格助詞を固定した範囲での格助詞パ

ターンの異なりなどは表 3-1 に示す。

表 3-1　調査対象による格助詞パターン数

	「日本語動詞の結合価」	白書	Yahoo!知恵袋
基本的な格助詞によるパターン	173	243	213
動詞数(単語数)	9,400	4,721	8,891
１動詞あたりのパターン数	1.8	5.2	2.4
パターン延べ数	35,526	22,859	45,454
動詞数(出現数)	155,433	297,016	552,714

3.2　格助詞パターンの実態

それぞれの調査対象から抽出した格助詞パターンの異なりの状況について述べる。

3.2.1　比較の対象

○出現形と基本形

「日本語動詞の結合価」から抽出した格助詞パターン（荻野　2006）は、基本形と出現形と二つの立場での格助詞パターンを抽出した。出現形は、実際に出ている格助詞パターンでパターンの異なりを集計したものである。出現形パターンは、出現しているありのままの形で格の省略などを含む。一方で荻野（2006）で述べる基本形とは、様々な出現形から人手判断にて基本とする格助詞組み合わせパターンを設定したものである。

本研究はあくまでもコーパスからの自動抽出なので、「日本語動詞の結合価」から抽出した格助詞パターンのうち、出現形に対応する部分に相当する。

○基本的な格助詞と「その他」の格助詞相当語

3.1 でもふれたように、本研究での格助詞パターンの作成は、①格助詞範囲内のものと②格助詞相当語なども含めたものと２段階で行なった。荻野（2006）でまとめた格助詞パターンは、格助詞相当語を含まない格助詞の組み合わせパターンのみなので、格助詞パターンの比較は、①格助詞相当語を含まないレベルの比較となる。

3.2.2　出現頻度上位の格助詞パターン

「日本語動詞の結合価」（荻野他　2003）から作成した格助詞パターン、本研究で作成した格助詞パターンの出現形のうち、それぞれ頻度の多いものを**表 3-2 表 3-3 表 3-4** に掲載する。

表 3-2「日本語動詞の結合価」の上位の格助詞パターン

順位	格助詞パターン	事例頻度合計	事例の割合	概念異なり数	異なり概念の割合
1	を	32,389	20.80%	4,443	12.80%
2	が_を	24,207	15.60%	3,806	11.00%
3	が	22,240	14.30%	4,038	11.60%
4	が_に	13,536	8.70%	2,274	6.50%
5	に	9,011	5.80%	1,787	5.10%
6	を_に	8,662	5.60%	1,648	4.70%
7	を_で	8,001	5.10%	2,057	5.90%
8	が_で	5,603	3.60%	1,749	5.00%
9	無格	4,909	3.20%	2,080	6.00%
10	が_を_に	3,254	2.10%	1,004	2.90%

表 3-3 「白書」の上位の格助詞パターン（その他含まず）

順位	格助詞パターン	白書出現頻度（パターン別出現頻度総数）	割合（出現頻度）	白書出現頻度（パターン別出現語数）	割合（単語異なり）
1	を	79,693	27.41%	2,283	9.99%
2	に	45,166	15.54%	1,662	7.27%
3	無格	43,411	14.93%	2,844	12.44%
4	が	19,747	6.79%	1,394	6.10%
5	と	14,494	4.99%	429	1.88%
6	を_に	12,734	4.38%	1,243	5.44%
7	を_は	6,122	2.11%	634	2.77%
8	が_に	5,966	2.05%	801	3.50%
9	を_と	5,536	1.90%	193	0.84%
10	は	5,316	1.83%	749	3.28%

表 3-4　Yahoo!知恵袋のパターン上位（その他含まず）

順位	格助詞パターン	Yahoo!出現頻度（パターン別出現頻度総数）	割合（出現頻度）	Yahoo!出現頻度（パターン別出現語数）	割合（単語異なり）
1	無格	224,079	40.54%	7,213	15.87%
2	を	62,959	11.39%	3,741	8.23%
3	に	55,898	10.11%	3,495	7.69%
4	が	50,011	9.05%	2,990	6.58%
5	は	28,985	5.24%	2,769	6.09%
6	で	22,942	4.15%	2,965	6.52%
7	と	21,045	3.81%	1,358	2.99%
8	を_に	11,135	2.01%	1,544	3.40%
9	が_に	10,645	1.93%	1,226	2.70%
10	に_は	6,977	1.26%	1,176	2.59%

これらの表で、上位に来ている格助詞あるいは格助詞パターンから以下のような状況が把握できた。

①「が」:

出現の割合でみると、1「日本語動詞の結合価」、2「Yahoo!知恵袋」、3「白書」の順である。

「が」が単独で出現するのは、他の格も必要な時に省略されるケースと元々「が」しか取らないケースである。元々「が」しか取らないケースは「花が咲く、お腹が膨らむ」のように「現象」に該当する動詞である。3「白書」などでは、対象格などを取る動詞が多く出現し、現象に関わるような記述は少ないことが想定される。

② 無格:

Yahoo!知恵袋において、「無格」のものが割合からみても大変多い。これは、Yahoo!知恵袋が会話に近い形で記述されているものが多いことに起因するかと思われる。

③「を」:

いずれの対象においても上位を占めている。どのような表現パターンにおいても、対象格が必要なものが外される可能性は少ないと言えるかと思われる。

3.2.3　格助詞パターン別の動詞表

格助詞パターン別に実際それらの格助詞パターンがどういう動詞から抽出されたかを表3-5に示す。

表 3-5　格助詞パターン別動詞表

	「その他」を除いた異なりパターン					
格助詞パターン	白書出現頻度（パターン別出現頻度）	白書出現頻度（パターン別出現語数）	白書事例単語（最大 20 個）	YAHOO 出現頻度（パターン別出現頻度）	YAHOO 出現頻度（パターン別出現語数）	Yahoo！知恵袋 事例単語（最大 20 個）
が_を	5006	808	あおる,あきらめる,あげる,あらわす,ある,あわせる,いう,いかす,いだく,いる,うかがう, …	3262	985	あがる,あける,あげる,アピールする,ある,いう,いじめる,いじる, …
が_を_に_と	15	9	する,なる,供与する,行う,上回る,設置する,占める,与える,来す…	9	8	する,施す,取る,絶賛する,退職する,動かす,払う,誘う
が_を_に_と_から	1	1	配分する	0	0	

3.2.4　格助詞組み合わせを抽象化してみた格助詞パターンの出現状況

3.1 で述べたように、格助詞パターンの異なり数は、対象データによるが、ほぼ１７０から２００になる。どんな格助詞の組み合わせのものが多いのか、格助詞に優先順をつけて格助詞パターンをさらに単純化してその傾向をみた。

○格助詞パターンの単純化

(1)格助詞パターン内の格助詞の優先順位を決めて組み合わせ内をソートする。

「は、が、を、に、へ、から、より、まで、で、と」の順

(2)「は、が、で」など、どの動詞グループにも出現可能な係助詞、格助詞を除く。

ただし、いくつかの特別処理を含める。

①着目の格助詞組み合わせにおいて、「は」、「が」、「で」のいずれか一つ、あるいはそれらの組み合わせで出現していて、かつ「が」、「で」以外の格助詞が他に存在するなら、「は」、「が」、「で」は抽出対象外とする。これは、「は」、「が」、「で」が基本的にどの動詞グループにも出現する可能性があるので、他の格助詞があるかぎり、これらを格助詞組み合

わせの中に含めないとしたものである。

②「で」については、場所を表す「で」については、動詞に限定されずに使える任意格とみなし、格助詞パターンの構成から除く。

（3）同じ機能の格助詞の整理

「から」と「より」、「に」と「へ」は、機能的に交替可能な格助詞とし、パターンを単純にするために基本とする方の格助詞に変換する。「より⇒から」「へ⇒に」の置き換えを行なう。

これらの格助詞の単純化の状況については、**表 3-6** で参照できる。

表 3-6　格の単純化

格助詞パターン(出現形)	格の単純化	コメント
が_を_に	を_に	「が」省略
が_を_に_と	を_に_と	「が」省略
が_を_に_と_から	を_に_と_から	「が」省略
が_を_に_と_で	を_に_と	[が、で]省略
が_を_に_と_は	を_に_と	[が、は]省略
が_を_へ	を_に	[へ⇒に」変換

以上のように置き換え、その結果、同じ組み合わせになった格助詞パターンを集計して出現状況をグラフにした。ここではそのうち第１分類のグラフを図 3.1～図 3.2 に示す。

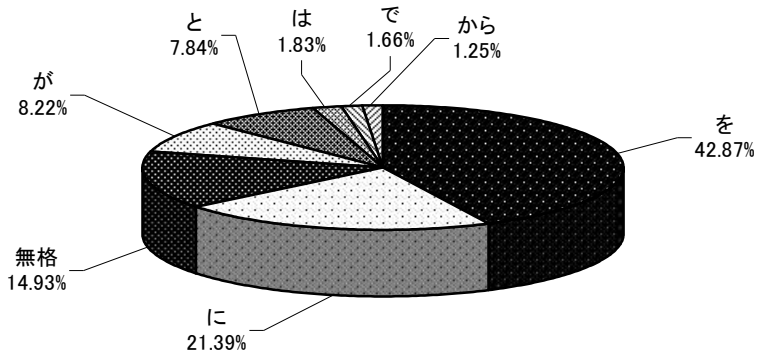


図3.1「白書」による格助詞パターンの分布（格助詞1）

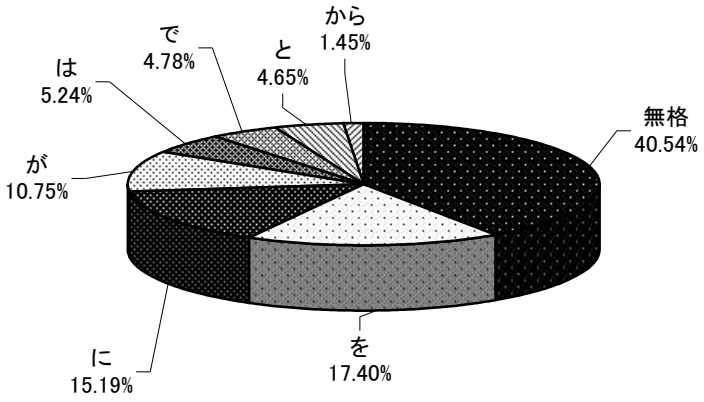


図3.2 Yahoo！知恵袋の格助詞パターン(格助詞1)

3.3　動詞ごとにみた格助詞パターン

一つの動詞あたりのパターン数は、**表 3-1** の中で１動詞あたりの平均パターン数として掲載した。これによると「日本語動詞の結合価」と「Yahoo!知恵袋」はおおまかにみて、１動詞あたり平均２パターンである。「白書」の場合の１動詞あたりのパターン数が多いのは、対象語数が他の二つのデータに比べ、少ないことが影響していることも予想される。

以上、結合価記述のうち、自動的に抽出した格助詞パターンを中心に述べた。

4　最後に

今後は、これらのパターンのうち、頻度の小さいものに着目して検討できればと思う。

文献

荻野孝野、小林正博、井佐原均（2003）．日本語動詞の結合価．三省堂

荻野孝野(2006) ．　日本語動詞の結合価の格助詞パターンと意味マーカに関する研究．神戸大学大学院自然科学研究科博士論文

荻野孝野、荻野綱男(2008 ）．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１９年度研究成果報告書『コーパスを利用した国語辞典編集法の研究』．コーパスに基づく動詞結合価の記述試作．　文部科学省科学研究費特定領域研究「日本語コーパス」辞書編集班 pp.21-64

奥村学他(2007), 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１８年度公開ワークショップ（研究成果報告会）．平成１８年度進捗状況報告：言語処理班「代表性の有るコーパスを利用した日本語意味解析」．　pp.69-78

日本電子化辞書研究所（1995）．.EDR 電子化辞書仕様説明書（第２版）EDR-TR045

異ジャンルの種用例を用いた半教師有りクラスタリングとその語義曖昧性解消に関する効果

杉山 一成 (言語処理班協力者：東京工業大学 精密工学研究所)[1]
奥村 学 (言語処理班班長：東京工業大学 精密工学研究所)

Semi-Supervised Clustering Using Seed Instances from Different Genre and Its Effect on Word Sense Disambiguation

Kazunari Sugiyama (Precision and Intelligence Laboratory, Tokyo Institute of Technology)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institute of Technology)

1 はじめに

東京工業大学の研究グループでは,「ジャンルの異なる用例に対応した半教師有りクラスタリング手法」の研究に取り組んでいる．単語の用例を語義ごとに分類するために，半教師有りクラスタリングを適用する場合には，クラスタリングの基準として，語義タグを付与した用例（以下，種用例）を導入する．しかし，このクラスタリングの際に導入する種用例は必ずしも同一ジャンルに属するものとは限らない．例えば，新聞ジャンルに属する用例を種用例として，白書ジャンルに属する用例をクラスタリングしたり，白書ジャンルに属する用例を種用例として，Yahoo 知恵袋ジャンルに属する用例をクラスタリングしたりする可能性もあると考えられる．そこで，ジャンルの異なる種用例を導入して，半教師有りクラスタリングを行った場合でも，高いクラスタリング精度を実現することが課題となる．本研究では，まず，異ジャンルの単語の用例から半教師有りクラスタリングのための種用例を選択する方法について説明する．次に，これまでに本グループで開発してきた種用例の重心の変動を抑えるクラスタリング手法 (Sugiyama, K. and Okumura, M., 2009) を適用し，種用例のジャンルと対象用例のジャンルが異なる場合に得られるクラスタリング精度について検証する．さらに，そのクラスタリング結果を利用した語義曖昧性解消の精度についても示す.

2 提案手法

2.1 システム構成

図 1 に，我々の提案する語義曖昧性解消システムを示す．本システムでは，まず，クラスタリングと語義曖昧性解消のための素性を抽出する．次に，種用例を導入して，半教師有りクラスタリングを行なう．その後，生成されたクラスタ内の単語の用例から，語義曖昧性解消のための素性を計算する．これらの素性を用いて，Support Vector Machine (SVM)，naïve Bayes (NB)，maximum entropy (ME) の 3 つの機械学習手法に基づいて分類器を構築する．

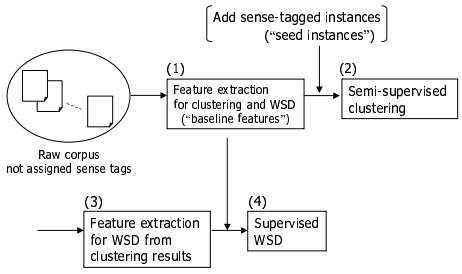


図 1: 提案する語義曖昧性解消システム

[1]sugiyama@lr.pi.titech.ac.jp

2.2 半教師有りクラスタリング

2.2.1 クラスタリングのための素性

本研究では，次の素性を用いる．

- 形態素素性
  - 対象単語および，その前後 2 語までの単語の表記
  - 対象単語および，その前後 2 語までの単語の品詞，品詞細分類
- 構文素性
  - 対象単語が名詞の場合，その名詞が係る動詞
  - 対象単語が動詞の場合，その動詞の格要素となる名詞
- 分類語彙表 (国立国語研究所, 1994) における意味カテゴリ
  - 対象語の左右の内容語に関して，分類語彙表中の 4 桁，5 桁の分類を用いる．例えば，対象語が「社会」であり，その左にある内容語が「地域」である場合に，分類語彙表における「地域」の分類は，“1.1720,4,1,3” である．この場合，4 桁の数値として “1172” を，5 桁の数値として “11720” を用いる．
- LDA (Blei, D. M., Ng A. Y., and Jordan M. I., 2003) によって推論される 5 トピック
  - “soft-tag” (Cai, J. F., Lee, W. S., and Teh, Y. W., 2007) という手法を用いて，用例の対数尤度を計算する．

なお，形態素解析器としては ChaSen[2] を，構文解析器としては CaboCha[3] を用いた．

以下において，これらの素性を「ベースライン素性」と呼ぶことにする．これらの素性は，3.4 節で述べる語義曖昧性解消システムにおいても使用する．

2.2.2 半教師有りクラスタリング

本研究では，我々がこれまでに開発してきた，種用例の重心の変動を抑える半教師有りクラスタリング手法を用いる (Sugiyama, K. and Okumura, M., 2009)．半教師有りクラスタリングのための種用例の選択法と制約の導入法は，以下に述べる通りである．

2.2.3 半教師有りクラスタリングのための種用例と制約

(1) 異ジャンルの用例を用いた種用例の選択法

異ジャンルの用例を用いた種用例の選択法を図 2 に示す．図 2 中の (a)〜(c) は，それぞれ，次の処理を行うことを示す．

(a) ソース (S) の語義の分布に基づいて種用例を選択し，対象 (T) をクラスタリングする．

(b) ソース (S) の語義の分布から，対象 (T) の語義の分布を (Chan, Y.-S. and Ng, H.-T., 2006) の手法を用いて推定し，推定した語義の分布に基づいて種用例を選択し，対象 (T) をクラスタリングする．

(c) 一度推定した語義の分布 (T) をソース (S) に加えて，(Chan, Y.-S. and Ng, H.-T., 2006) で対象の語義の分布を推定し，推定した語義の分布 (T') に基づいて種用例を選択し，対象をクラスタリングする．

なお，種用例に関して，比例代表制の政党リストにおける候補者に議席を割り当てるための手法である D'Hondt 法 (Taagepera, R. and Shugart, M. S., 1991) を用いて各語義について所定の数を選択するが，その種用例はランダムに選択する (Sugiyama, K. and Okumura, M., 2009)．

[2] http://sourceforge.net/projects/masayu-a/
[3] http://sourceforge.net/projects/cabocha/

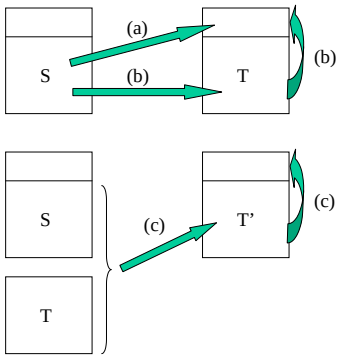


図 2: 異ジャンルの用例を用いた種用例の選択法

表 1: 語義数と対応する対象語

語義数	対象語
2	「技術」,「現在」,「自分」,「程度」,「子供」,「事業」,「情報」,「地方」,「場合」,「考える」,「図る」,「含む」,「見せる」,「生まれる」
3	「関係」,「午後」,「市民」,「意味」,「買う」,「分かる」,「作る」
4	「時代」,「一般」,「思う」,「進める」,「訴える」,「書く」,「聞く」
5	「今」,「時間」,「前」,「使う」
6	「出来る」,「見る」
7	「社会」,「開く」
8	「言う」,「乗る」
9	「進む」,「持つ」
10 以上	「入る」(11),「手」(11),「取る」(21),「かかる」(23),「出る」(25),「出す」(25)

(2) 制約の導入法

異なる語義を持つ種用例間には “cannot-link” の制約を，また，同じ語義を持つ種用例間には，“must-link” の制約を導入する．ただし，“must-link” の制約を導入する際には，外れ値を除く工夫 (Sugiyama, K. and Okumura, M., 2009) を行い，精度の良いクラスタリング結果が得られるようにしている．

3 実験

3.1 実験データ

本研究では,「SENSEVAL-2 日本語辞書タスク」で配布された RWC コーパスから新聞のデータを，また，日本語コーパスから，白書，Yahoo 知恵袋のデータを使用して実験を行った．なお，白書，YAHOO 知恵袋のデータにおいては「SENSEVAL-2 日本語辞書タスク」で使用された対象語のすべてに語義タグの付与作業が完了していないため，すでに語義タグが付与されている単語を，今回の実験対象語とした．これらの実験対象語を表 1 に示す．10 語義以上の単語の括弧内に示した数値は，各単語の語義数を表している．

3.2 語義の分布の推定精度

3.2.1 実験結果

ソースの語義の分布と対象の語義の分布を，KL divergence の値により評価した結果を表 2 に示す．また，表 2 の各組合せにおける KL divergence の値が小さい 10 語，大きい 10 語を表 3 に示す．

3.3 半教師有りクラスタリング

本実験においては，2.2.2 節で述べたように，はじめに種用例と制約を導入する．種用例は，単語の用例のデータ集合の 80%に対応する訓練用例から選択され，クラスタリングのためのテスト用例は，単語の用例のデータ集合の 20%に対応する．図 3，4 に示すクラスタリング結果は，5 分割交差検定に基づいている．

表 2: KL divergence の値

ソース	対象		KL (S-T)	KL (S-T') (T'=S+T)
新聞	白書	(a) KL 値の小さい順に 10 語の平均	3.88×10^{-2}	4.25×10^{-2}
		(b) KL 値の大きい順に 10 語の平均	8.01	8.46
	Yahoo	(c) KL 値の小さい順に 10 語の平均	2.47×10^{-3}	2.65×10^{-3}
		(d) KL 値の大きい順に 10 語の平均	4.41	5.02
白書	新聞	(e) KL 値の小さい順に 10 語の平均	6.13×10^{-2}	6.85×10^{-2}
		(f) KL 値の大きい順に 10 語の平均	5.29	5.66
	Yahoo	(g) KL 値の小さい順に 10 語の平均	1.37×10^{-4}	2.12×10^{-4}
		(h) KL 値の大きい順に 10 語の平均	4.73	4.96
Yahoo	新聞	(i) KL 値の小さい順に 10 語の平均	8.64×10^{-2}	9.02×10^{-2}
		(j) KL 値の大きい順に 10 語の平均	9.08	11.03
	白書	(k) KL 値の小さい順に 10 語の平均	7.30×10^{-2}	7.85×10^{-2}
		(l) KL 値の大きい順に 10 語の平均	8.56	8.94

表 3: 表 2 の各実験における 10 語

	各対象語 (括弧内の数値は語義数)
(a)	図る (2)、生まれる (2)、かかる (23)、時間 (5)、訴える (26)、出す (25)、取る (21)、手 (11)、午後 (3)、意味 (3)
(b)	分かる (3)、場合 (2)、書く (4)、前 (5)、地方 (2)、一般 (4)、関係 (3)、買う (3)、思う (4)、含む (2)
(c)	図る (2)、含む (2)、生まれる (2)、関係 (3)、一般 (4)、思う (4)、持つ (9)、考える (2)、時間 (5)、手 (11)
(d)	子供 (2)、使う (5)、社会 (7)、開く (7)、出す (25)、今 (5)、入る (11)、出る (25)、進む (9)、買う (3)
(e)	生まれる (2)、時間 (5)、取る (21)、現在 (2)、作る (3)、事業 (2)、見せる (2)、自分 (2)、子供 (2)、入る (11)
(f)	思う (4)、関係 (3)、前 (5)、技術 (2)、言う (8)、程度 (2)、社会 (7)、時代 (4)、見る (6)、進む (9)
(g)	かかる (23)、出る (25)、出す (25)、入る (11)、持つ (9)、情報 (2)、図る (2)、生まれる (2)、買う (3)、一般 (4)
(h)	出来る (6)、現在 (2)、市民 (3)、言う (8)、見る (6)、今 (5)、聞く (4)、午後 (3)、関係 (3)、含む (2)
(i)	図る (2)、生まれる (2)、思う (4)、持つ (9)、時間 (5)、手 (11)、取る (21)、出来る (6)、情報 (2)、程度 (2)
(j)	進む (9)、入る (11)、開く (7)、言う (8)、地方 (2)、自分 (2)、かかる (23)、時代 (4)、技術 (2)、見る (6)
(k)	かかる (23)、出す (25)、情報 (2)、一般 (4)、思う (4)、考える (2)、見せる (2)、子供 (2)、使う (5)、自分 (2)
(l)	関係 (3)、言う (8)、聞く (4)、進む (9)、程度 (2)、場合 (2)、技術 (2)、書く (4)、進める (4)、時間 (5)

3.3.1 実験結果

本研究では，“purity”，“inverse purity” の調和平均である F 値 (Hotho, A. and Nürnberger, A. and Paaß, G., 2005) に基づいて，クラスタリングの精度を評価する．KL divergence の値で最も小さい値が得られた組合せ (ソース:白書-対象:Yahoo，(g))，逆に，KL divergence の値で最も大きい値が得られた組合せ (ソース:Yahoo-対象:新聞，(j)) についての半教師有りクラスタリングの結果を，それぞれ，図 3，4 に示す．これらのグラフにおいて，各線は，各語義数に対して得られたクラスタリング精度を示す．

3.4 語義曖昧性解消

2.2.1 節で述べたベースライン素性とともに，クラスタリング結果からは，クラスタの ID，隣接する単語間の相互情報量，対象語の左右 2 語に関する情報利得を計算し，語義曖昧性解消のための素性とする (Sugiyama, K. and Okumura, M., 2009)．

3.4.1 実験結果

語義曖昧性解消の評価尺度としては，精度を用いた．この精度は，正解の語義 ID とシステムの出力する語義 ID が完全に一致していれば正解とする “fine-grained scoring” (白井清昭, 2003) に基づくものである．表 4，5 に (ソース:白書-対象:Yahoo)，(ソース:Yahoo-対象:新聞) の語義曖昧性解消結果を，それぞれ示す．

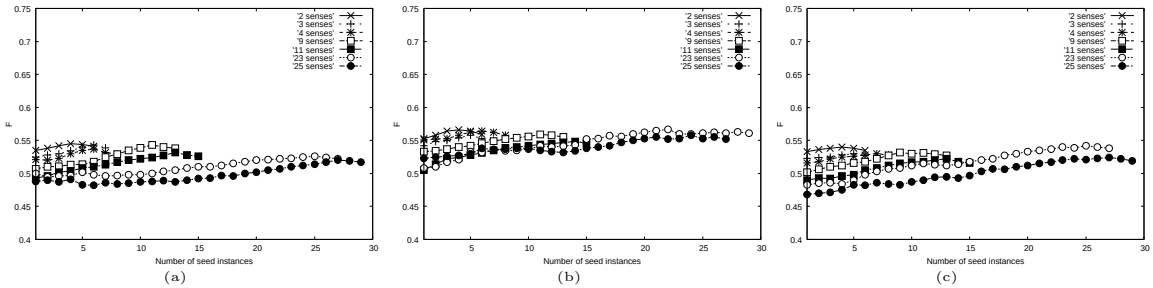


図 3: (ソース:白書-対象:Yahoo): KL 値の小さい順に 10 語のクラスタリング結果

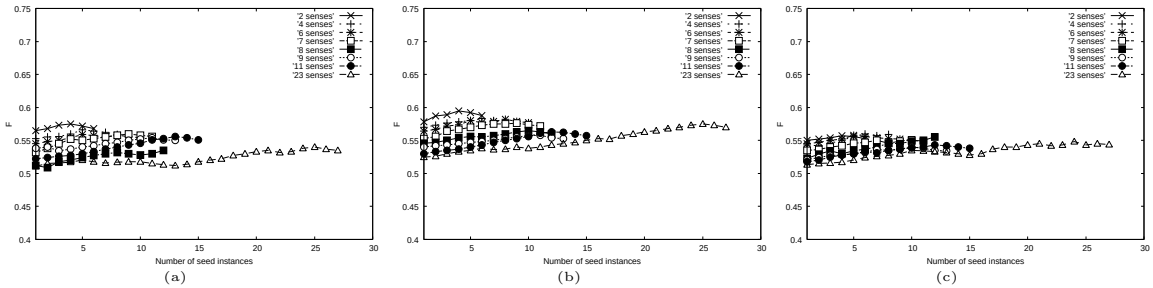


図 4: (ソース:Yahoo-対象:新聞): KL 値の大きい順に 10 語のクラスタリング結果

4 考察

まず，語義の分布の推定に関して，次のことが観察される．ソースが新聞や白書である場合，特に，(ソース:新聞-対象:Yahoo)，(ソース:白書-対象:Yahoo) の KL 値の小さい順に 10 語の平均値は，他の組合せと比べて値が小さい．新聞や白書のようなフォーマルな文書をソースとした場合，Yahoo のような口語体を中心とした文書の語義の分布は，(Chan, Y.-S. and Ng, H.-T., 2006) によって推定しやすいと考えられる．一方，ソースが Yahoo の場合，KL の小さい順に 10 語の平均値は，他の組合せと比べて値が大きい．したがって，Yahoo のような口語体を中心とした文書をソースとした場合には，新聞，白書のようなフォーマルな文書の語義の分布は，(Chan, Y.-S. and Ng, H.-T., 2006) では推定しにくいと考えられる．

次に，クラスタリング精度について，ソースの分布をそのまま適用して種用例を選択し，対象をクラスタリングする (a) と比較して，ソースの分布から対象の分布を推定し，推定された語義の分布から種用例を選択してクラスタリングを行う (b) によって，(ソース:白書-対象:Yahoo) では 2〜4%程度，(ソース:Yahoo-対象:新聞) では 1〜3%程度，精度が改善されている．また，一度，推定したものを再度ソースに加えて語義の分布を推定し，その語義の頻度分布から種用例を選択してクラスタリングを行う (c) の場合には，(a) と比較して，(ソース:白書-対象:Yahoo) では 2〜4%程度，(ソース:Yahoo-対象:新聞) では 1%程度，精度が悪くなることが観察される．

さらに，語義曖昧性解消の精度について，(ソース:白書-対象:Yahoo)，(ソース:Yahoo-対象:新聞) のいずれの場合も，クラスタリング精度の最も良い (b) の場合に，最良な語義曖昧性解消精度が得られている．また，一度推定した語義の分布を加えた (c) の場合，クラスタリング精度が，(a)，(b) と比べ，それほど高くないために語義曖昧性解消の精度も他と比べて劣っている．一度推定した語義の分布すべてを加えるのではなく，何らかの工夫をすることで，改善されることが期待される．

5 おわりに

本研究では，まず，異ジャンルの単語の用例から，半教師有りクラスタリングのための種用例を選択する方法について述べ，これまでに本グループで開発してきた種用例の重心の変動を抑えるクラスタリング手法を適用したクラスタリング精度，さらに，そのクラスタリング結果を利用した語義曖昧性解消の精度についても示した．

今後の課題として，(1) 異ジャンルの語義の分布を推定する際，2.2.3 節 (c) で述べたように，推定した語義の分布すべてを加えるのではなく，何らかの基準に基づいて選択して加える手法の開発，(2)

表 4: 語義曖昧性解消精度 (ソース (S):白書-対象 (T):Yahoo)

種用例のソース	SVM	NB	ME
(a) S	0.577	0.582	0.548
(b) (S→)T	0.592	0.607	0.559
(c) (S+T→)T'	0.565	0.573	0.541

表 5: 語義曖昧性解消精度 (ソース (S):Yahoo-対象 (T):新聞)

種用例のソース	SVM	NB	ME
(a) S	0.668	0.664	0.620
(b) (S→)T	0.687	0.682	0.643
(c) (S+T→)T'	0.649	0.657	0.616

正確に語義の分布を推定するために，個々の素性を重み付けする手法の開発，(3) 半教師有りクラスタリングを行う際，辞書に定義されているが，対象側で出現する可能性がある場合の種用例を選択する手法の開発，などが挙げられる．

文献

Blei, D. M., Ng A. Y., and Jordan M. I. (2003). "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Cai, J. F., Lee, W. S., and Teh, Y. W. (2007). "Improving Word Sense Disambiguation Using Topic Features," in *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007)*, pp. 1015–1023.

Chan, Y.-S. and Ng, H.-T. (2006). "Estimating Class Priors in Domain Adaptation for Word Sense Disambiguation," in *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (ACL 2006)*, pp. 89–96.

Hotho, A. and Nürnberger, A. and Paaß, G. (2005). "A Brief Survey of Text Mining," *GLDV-Journal for Computational Linguistics and Language Technology*, Vol. 20, No. 1, pp. 19–62.

Sugiyama, K. and Okumura, M. (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)*, pp. 266–279.

Taagepera, R. and Shugart, M. S. (1991). *Seats and Votes: The Effects and Determinants of Electoral Systems*: Yale University Press.

国立国語研究所 (1994). 分類語彙表，秀英出版.

白井清昭 (2003). 「SENSEVAL-2 日本語辞書タスク」，自然言語処理，2 巻，pp.3–24.

複数の語義を積極的に取り出す動詞のクラスタリング

高橋 秀幸 (言語処理班協力者: 岡山大学工学部)
竹内 孔一 (言語処理班分担者: 岡山大学大学院)[1]

Co-Clustering Approach for Extracting Verb Meanings from Polysemous Verb

Koichi Takeuchi (Graduate School, Okayama University)
Hideyuki Takahashi (Faculty of Engineering, Okayama University)

1 はじめに

本研究ではコーパスから動詞の類語を獲得する手法を構築することを目指している．共通する概念を持つ動詞の類義語を獲得できれば，動作 (事態) 表現の多様性を吸収することができる．例えば「富をたくえる/ためる」がほとんど同じ意味であることや「ひげをたくわえる/はやす」は「たくわえる」と言い換えできることが処理できる．こうした動詞の類義語を収集する方法として本研究ではグラフ構造を基にした同時クラスタリング法 (Aizawa (2002)) を利用した類義語抽出法を提案し一定の成果を上げた (真野・竹内 (2008))．

しかしながら動詞の多義性を積極的に取り込む機構になっていなかったため，十分に動詞のクラスタを取り出すことができていなかった．そこで本稿では初期クラスタに動詞の多義情報があると仮定して，繰り返し語義を取り出す機構を加えることで出力クラスタの精度 (purity とカバー率) を向上させることができたので報告する．さらに本特定領域研究で構築されている均衡コーパス (BCCWJ) を利用した場合，カバー率が大きく上昇し，動詞語義獲得に有効であることを示す．

2 多義語を積極的に取り出すクラスタリング

Aizawa 法による同時クラスタリングは格フレームから名詞と動詞の 2 部グラフを作成しリンクの緊密な部分を情報量の尺度を利用して抽出する．よって大規模なテキストデータに適用できる反面，山登り法であるため複数の解がある場合に取りこぼしてしまう場合がある．具体的に例を挙げて説明する．

多義性を持つ動詞がある場合の初期クラスタの様子を図 1 にイメージとして記述する．多義でかつそれぞれの語義に対する動詞集合と名詞集合が初期クラスタに存在するならば，点線で示すように重複したクラスタリングが可能であろう．Aizawa 法の場合，格要素を一定の基準で削除していき偶然残ったどちらかのクラスタが出力されて，後は捨てられていた．よって積極的に多義を取り出すには削除して得られたクラスタを元に，対応する名詞クラスタを除いてクラスタを作成すれば多義性のある場合の残りのクラスタを獲得すれば良い (Pantel and Lin (2002))．

図 1 で説明すると，初期クラスタとして動詞側に「つくる」,「にぎる」,「つかむ」,「知る」が存在し，実線で結びついている名詞側には「おにぎり」,「すし」,「情報」,「機密」,「杖」がヲ格で出現したと仮定する．まず，同時クラスタリングを適用して最初に「つくる」「にぎる」と「おにぎり」「すし」のクラスタが獲得されたとしよう (図 1 中の実線の四角内)．

そこで，これらを取り出すために，先ほど獲得したクラスタの名詞側とのリンク (「おにぎり」「すし」) だけを初期クラスタから削除して，同時クラスタリングを適用する．すると,「にぎる」,「つくる」という語義に対応する名詞へのリンクが存在しないためクラスタは作成されず,「情報」と「機密」に対応する「にぎる」,「つかむ」,「知る」というクラスタが獲得される．さらに,「情報」「機密」も初期クラスタから削除して，同時クラスタリングを適用すれば，他の語義を積極的に獲得することができる．この方法を Active Extraction Co-Clustering (AECC) と以降呼ぶ．

[1] koichi@cl.cs.okayama-u.ac.jp

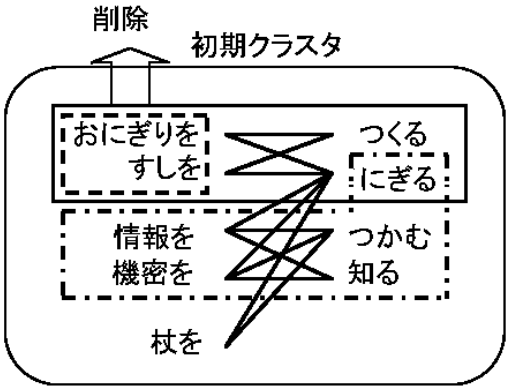


図 1: 初期のグラフ構造から繰り返し名詞リンクを削除することで他の語義を獲得

3 実験

クラスタリングによる動詞分類実験において上記で提案した AECC がどの程度従来法よりも良いか比較を行う．対象とするコーパスは新聞記事コーパス，Web 上のデータ，均衡コーパス (BCCWJ) を利用する．こうしたコーパスを利用できることから，コーパスの質の違いによる動詞クラスタ抽出の精度についても検討する．

3.1 格フレームデータ

必要とするデータは格関係付きの名詞と動詞の共起情報である．下記の 5 つのデータに対して実験を行った．

(a) Yahoo!知恵袋 45725 組の質問とベストアンサーを係り受け解析して得られた格フレーム

(b) 毎日新聞 91 年から 98 年版を係り受け解析して得られた格フレーム

(c) BCCWJ を係り受け解析して得られた格フレーム

(d) BCCWJ と同サイズの毎日新聞を係り受け解析して得られた格フレーム

(e) Web 上の 5 億文コーパスからの格フレーム

(a)(b)(c)(d) は CaboCha の係受け解析により得られた格フレームである．(e) は河原・黒橋 (2006) によって作成されたもので，構文解析器 KNP を利用して得られた格フレームである．(d) は BCCWJ と毎日新聞のコーパスの質の違いを検証するために，比較用として毎日新聞 91-92 コーパスを一部削除してバイト数を BCCWJ と同等量にしたデータである．

3.2 評価法

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 (竹内他 (2008)) と比較を行う．動詞項構造シソーラスは約 4400 語の動詞が登録されており，分類として 5 階層に細分化されている．評価は 5 階層から成る辞書のうち第 3 階層の分類と比較して行う．同じ分類に属する 2 つ以上の動詞が存在する出力クラスタを正解クラスタとし，正解クラスタ中の同じ分類に属する動詞を正解要素とする．評価の尺度として出力クラスタがどれだけ正解を含むかを平均した purity と辞書の分類をどれだけ満たしたかを示すクラスカバー率 (118 分類のうちいくつカバーしたか) とカバー率 (8588 要素のうちいくつカバーしたか) を測る (詳細は高橋・竹内 (2009))．

3.3 抽出結果

AECC を適用する前と適用した後の結果について表 1 に示す．AECC 適用前と適用後の結果を比べると，purity，クラスカバー率，カバー率の値においてほとんどのデータに対して 0.5 から 3%程度上昇した．ただし Web データの purity のみ値が下がった．これは AECC を適用することで新たに得られたクラスタに誤りが多いことを表す．原因として Web のデータには書き方の違いによる同義語 (例えば「逸れる」と「それる」) が多数存在しており，こうした要素が多く獲得される一方で，評価基準となる動詞項構造シソーラスには登録がなく評価が下がったのではないと考えられる．また，Yahoo!データの purity とカバー率は他のデータに比べて低い．これは Yahoo!データの格フレームデータ統計量が他のデータに比べて少ないことによると考えられる．

コーパスの量について見ると，毎日新聞ではコーパスの量が増加するにつれて purity とカバー率が上昇する傾向が見られるものの，最も高い精度が得られたのは毎日新聞 91 年から 95 年を利用した場合であり，単調増加とならなかった．これは信頼性の高いクラスタを得るために動詞や名詞の特定の意味に対する使われ方に関与しない語（動詞の例だと「する」や「なる」など）を排除するリンク数の制限が，コーパス量の増加に伴う各要素のリンク数の増加によって有益な要素まで制限の対象としてしまうことが原因と考えられる．この点についてはリンク数の制限とコーパスの量との関係について最適な関係を調べる必要があると考えられる．

コーパスの種類について見ると，BCCWJ は同じサイズの毎日新聞に比べて，purity とカバー率は共に高かった．これは幅広い分野からコーパスを集めていることで語義の使われ方が固定的でない BCCWJ と，語義の使われ方がある程度固定的な新聞とのコーパスの質の違いによる差であると考えられる．一方，Web と Yahoo!は毎日新聞と BCCWJ に比べ purity，カバー率共に低い精度であった．大規模 Web に対する精度はコーパス量の違いによるリンク数制限がうまく働いていないためまだ正しく手法が適用できていない可能性が大きい．さらに上記に述べたように漢字と平仮名の書き方の異なりが獲得されて評価を落とす場合が見受けられる．Web は新聞記事より表現の種類が多く，多くの語義を獲得できる期待があるが，現実には不要なデータが多く，その点均衡コーパスが有効であることが推測できる．

3.4 取り出されたクラスタに対する考察

AECC により複数のクラスタが獲得された事例を表 2 と表 3 に示す．表 2 は「充てる」という動詞について 3 つの出力クラスタが得られた例である．1 回目の出力クラスタを見ると，名詞の要素として目的語になんらかの富を指す語がニ，カラ格，ヲ格を伴って獲得されている．これらの名詞の集合と動詞の要素「流用」「充当」「充てる」「回す」という動詞のクラスタから『富を使用する』という語義が獲得できていることがわかる．次に 1 回目で得られた名詞群を削除して得られた 2 回目の出力クラスタではデ，ヲ，カラ格といった格関係は異なるが 1 回目と同様に富を表す名詞が画策されている．動詞のクラスタは「賄う」や「補てん」が新たに加わっているが，これも同様に『富を使用する』という語義の別クラスタが獲得できている．さらにこの名詞群を削除して得られた 3 回目の出力クラスタではニ，ヲ，トシテ格を伴って同様になんらかの富を指す語が獲得できている．動詞の要素には「つぎ込む」や「活用」など新たな要素が獲得できているがこれも同様に『富を使用する』という語義が獲得できている．以上から 3 つの出力クラスタの動詞集合は結局同じ語義であると考えられる．つまり，本来一つのクラスタであるべき動詞集合が 3 つの出力クラスタとして得られたことを示す．これは従来繰り返して取り出すことを行わなければ捨てられていたクラスタ要素であり，カバー率を向上させる上で重要な結果である．しかしながら本来目的としていた語義の違いまでは獲得できなかった結果であると言える．

一方，表 3 は多義性を積極的に獲得する AECC により新たな語義を得ることができるようになった出力の例である．表 3 中の「跳ね上がる」について 1 回目のクラスタでは『価格の高騰』を意味し，2 回目のクラスタでは『身体の反応』を意味する語義が得られている．また,『おもむく』について 1 回目のクラスタでは『移動』の語義が獲得され，2 回目のクラスタでは『心的変化』の語義が獲

表 1: AECC を適用しない場合と適用する場合の評価値の比較

	purity		クラスカバー率		カバー率	
	通常	AECC	通常	AECC	通常	AECC
Yahoo!	0.152	0.169	0.364	0.390	0.047	0.052
	(507/3335)	(645/3812)	(43/118)	(46/118)	(402/8588)	(448/8588)
毎日 91	0.269	0.280	0.475	0.517	0.078	0.106
	(1100/4083)	(2079/7421)	(56/118)	(61/118)	(670/8588)	(914/8588)
91-92	0.284	0.291	0.542	0.568	0.103	0.116
	(1516/5331)	(2150/7378)	(64/118)	(67/118)	(886/8588)	(994/8588)
91-93	0.289	0.298	0.551	0.593	0.113	0.126
	(1650/5708)	(2444/8194)	(65/118)	(70/118)	(968/8588)	(1081/8588)
91-94	0.282	0.297	0.602	0.627	0.116	0.131
	(1648/5854)	(2520/8471)	(71/118)	(74/118)	(995/8588)	(1128/8588)
91-95	0.315	0.325	0.636	0.653	0.130	0.148
	(1788/5685)	(2678/8241)	(75/118)	(77/118)	(1113/8588)	(1270/8588)
91-96	0.293	0.309	0.636	0.644	0.122	0.138
	(1595/5452)	(2447/7926)	(75/118)	(76/118)	(1046/8588)	(1185/8588)
91-97	0.307	0.317	0.602	0.619	0.117	0.135
	(1631/5311)	(2465/7784)	(71/118)	(73/118)	(1009/8588)	(1160/8588)
91-98	0.317	0.322	0.576	0.593	0.118	0.133
	(1574/4967)	(2268/7037)	(68/118)	(70/118)	(1010/8588)	(1146/8588)
Web	0.250	0.245	0.517	0.551	0.086	0.102
	(1004/4020)	(1522/6221)	(61/118)	(65/118)	(741/8588)	(872/8588)
BCCWJ	0.263	0.270	0.576	0.644	0.128	0.145
	(1652/6293)	(2323/8618)	(68/118)	(76/118)	(1097/8588)	(1241/8588)
毎日比較用	0.279	0.292	0.542	0.568	0.100	0.113
	(1485/5318)	(2126/7283)	(64/118)	(67/118)	(860/8588)	(968/8588)

得されている．ただし名詞側の「庁舎に」は移動の語義と結びつくべき要素でありこの場合はこの要素が誤りであると考えられる． 表中の 3 つめの事例である「たくわえる」の語義については 1 回目

表 2: 「充てる」について同義のクラスタが観測できた例

1 回目	動詞	流用 充当 あてる 充てる 回す
	名詞	費 に 代 に 財源 に 原資 に 返済 に 益 を 費 から 資金 に 経費 に 購入に
2 回目	動詞	賄う 充当 まかなう 充てる 補てん
	名詞	利子 を 財源 を 会計 から 増税 で 国債 で 会計 で
3 回目	動詞	つぎ込む 充当 活用 引き下げる 充てる
	名詞	収入 を 整備 に 輸送 に 財源 として

表 3: 多義性が観測できたクラスタ例

跳ね上がる		
1 回目	動詞	高騰 下がる 値上がり 上昇 跳ね上がる
	名詞	円 に 価格 が 物価 が 倍 に
2 回目	動詞	高鳴る 躍る 縮む 弱る 跳ね上がる
	名詞	反動 で 身体 が 心臓 が クン と
おもむく		
1 回目	動詞	おもむく 帰る 向かう 到着 行く
	名詞	都 へ 邸 へ 城 へ 屋敷 へ
2 回目	動詞	安らぐ おもむく 休まる うずく はずむ
	名詞	庁舎 に 紅屋 に 本能 の 心 の 興味 の
たくわえる		
1 回目	動詞	蓄える たくわえる 吸い上げる はやす 生やす
	名詞	髭 を 口ひげ を げ を 蜜 を 栄養を 勢力 を
2 回目	動詞	蓄える 補給 蓄積 生み出す 生む
	名詞	利子 を 財源 を 会計 から 増税 で 国債 で 会計 で

のクラスタで『ひげなどをはやす』という語義が獲得できると考えられるが，一方で名詞の要素をみると「密」や「栄養」など別の語義もクラスタ内に混在しており不要要素が少なくない．さらに 2 回目のクラスタでは『ためる』という語義とともに「増税」や「財源」と「生む」や「生み出す」との組み合わせが示すように『生成する』という異なる語義が混在している．これらは不正確なクラスタであるが,「たくわえる」に対してどのような語義があるかを人間が確認する場合には有効であると考えられる．

このように AECC による繰り返し抽出する方法では一つの語義クラスタの要素がいくつも取り出される場合と，別の語義が取り出される場合がある．どちらもどのような語義があるかを取りこぼしを減少させる (カバー率を向上させる) 上で効果的である．繰り返し削除を行ってもまだ獲得されない動詞の語義クラスタが入力データ内に存在する可能性は否定できない．これは本質的に提案手法では最適なクラスタを得ることはできないためである．しかしながら局所的なクラスタを精度良く獲得できることから多量のコーパスを用意することでカバー率を上げることが可能である．

本評価では動詞シソーラスをベースに評価しているが当然シソーラスに無い語は評価できておらず，今後人手による評価を行いたい．また，評価結果から繰り返し削除による動詞の多義クラスタを

取り出す手法はある程度成功していると考えられる．しかしながら語義が同一でクラスタの要素が少しづつ異なる例が取り出される場合との違いは明確にする必要がある．よって類似クラスタのマージを行うなどさらなるクラスタの整理を行うモジュールも検討する必要がある．

4 まとめ

本研究報告では同時共起クラスタリング手法を利用して多義性を有する動詞の語義を積極的に取り出す方法を提案した．Yahoo!知恵袋，Web コーパス，毎日新聞，BCCWJ から抽出した格フレームデータを用いて実験を行い，既存の動詞項構造シソーラスを利用して精度を比較した．実験の結果，提案手法により同時クラスタリングの出力クラスタ数が増加し，精度が改善された．また動詞の多義の出力を得ることができた．コーパスの質の違いでは，BCCWJ は毎日新聞に比べて purity を維持したままカバー率が高くなった．今後の展望として，人手によるクラスタの評価を行いどの程度の有効性があるかを数値化したい．

文献

Akiko Aizawa (2002) “A method of Cluster-Based Indexing of Textual Data,” in *Proceedings of COLING 2002*, pp. 1–7.

Patrick Pantel and Dekang Lin (2002) “Discovering word senses from text,” in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 613–619.

河原大輔、黒橋禎夫 (2006) 「高性能計算環境を用いた Web からの大規模格フレーム構築」，情報処理学会自然言語処理研究会，2006-NL，pp.67–73.

高橋秀幸、竹内孔一 (2009) 「多義性を考慮した同時共起クラスタリングによる動詞の類語抽出」，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 *NLC2008-77*，pp.37–42.

真野光平、竹内孔一 (2008) 「項関係にある名詞との共起を考慮した動詞のクラスタリング」，言語処理学会第 *14* 回年次大会，pp.1033–1036.

竹内孔一、乾健太郎、竹内奈央、藤田篤 (2008) 「意味の包含関係に基づく動詞項構造の細分類」，言語処理学会第 14 回年次大会発表論文集，pp.1037–1040.

BCCWJ を用いた新しい語義曖昧性解消タスク

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)

New Japanese WSD Task Using BCCWJ Corpus

Manabu Okumura (Tokyo Institute of Technology)
Kiyoaki Shirai (Japan Advanced Institute of Science and 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，次回の Semeval-2 は 2010 年にワークショップが開催予定である (`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，このタスクの背景, 狙う点, 課題の内容等について主に説明する．

2 BCCWJ コーパスを用いた意味解析

国立国語研究所の前川喜久雄氏を領域長として，文部科学省科学研究費補助金特定領域研究「日本語コーパス」プロジェクトが 2006 年 9 月にスタートしている (`http://www.tokuteicorpus.jp/`). このプロジェクトでは，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 Complementary 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．このコーパスは，様々なジャンルのテキストからなり代表性があるという特徴をもつと同時に，10-20 年程度の時間幅を持ったテキストからなる部分を持ち，継時性があるという特徴も合わせ持っている．これらの特徴を持つコーパスはあまりなく，これらの特徴を持つコーパスを利用し言語処理研究を行うことには大きな意味がある．そこで，我々は，このプロジェクトの一貫として，以下の 2 つのテーマを選んで，上述したコーパスの特徴を活かした日本語の意味解析に関する研究を行っている[2].

1. 意味解析では，単語の語義を自動的に同定すること (語義曖昧性解消) も 1 つのテーマであるが，この語義曖昧性解消を代表性のあるコーパスで行う場合には，どうすれば良いのだろうか．様々なジャンルのテキストでは，同じ単語についても，出現する語義の分布は異なるため，従来のように，特定のジャンルのテキスト (たとえば，新聞データ) を対象にした手法が同じようにうまく行くとは限らない．

2. 語義曖昧性解消は，あらかじめ仮定した辞書中の語義のどれであるかを決めようとするが，そもそも用例の語義が辞書中にない場合どうすれば良いのか．継時性があるコーパスでは，辞書に載っていない語義が時間経過を経て出現することはかなりあるように思われる．辞書中に載っていない，いわゆる新語義を発見できれば，辞書編集者に貢献することもできる．

†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`.

[2] 詳細は，[3] を参照していただきたい．

3　代表性のある語義タグ付コーパスの構築

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 [4], 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．

4　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．

4.1　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．

なお，語義曖昧性解消における領域適応に関する先行研究には，たとえば，[2, 1] などがある．

4.2　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．

例を上げよう．岩波国語辞典で単語「ネタ」には，(1) 新聞記事などの材料, (2) 手品の仕掛け, (3) 証拠, (4) （料理の材料としての）食物，の 4 つの語義が与えられている．このとき，以下の用例 (a) の語義は，(2) に近いが,「何かの大事な部分」を指しており，また，(b) の語義は，4 つの語義のうちにはなく,「作り話」のような意味で用いられている．

> ストーリー上、必ず使いますか？　「ネタ」ばれでも良いので教えてください。(a)
> まさにその通りですね。。。。。「ネタ」みたいに見えるけどほんとそうだよ (b)

このような用例を発見し，これらの用例の語義が辞書中には与えられていないものであると判定することが新語義発見の目標である．

4.3 新語義発見を含む語義曖昧性解消へのいくつかのアプローチ

上述したように，近年の語義曖昧性解消では, 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手法が採用されることが多い．では，新語義発見を含む語義曖昧性解消では，どのような手法を採ることができるだろうか．

新語義発見を含む語義曖昧性解消でも，単純には,「新語義」というクラスを，対象とする単語の (辞書中に列挙されている) 語義クラスの集合に追加し，その中から語義クラスを選択する分類器を学習することで，同様の手法が採れそうである (discrimination-based な手法)．ただ，この手法の場合,「新語義」クラスに対する訓練データが非常に少ない (あるいは仮定できない) という問題に対処する必要がある．

新語義発見を含む語義曖昧性解消では，これ以外にも，様々な手法が考えられそうである．これも素朴だが，コーパスに出現する単語の用例集合を，その単語が出現する文脈の類似性という観点からまずクラスタリングし，このとき，同じ意味を持つ用例集合は同じクラスタに分類されると考え，クラスタリングによって得られたクラスタが単語の語義に対応すると仮定する (induction-based な手法)．そして，得られたクラスタを辞書中の語義クラスと対応付け，対応付けできなかったクラスタを新語義に対応するクラスタと考える．

このように，新語義発見を含むように語義曖昧性解消を拡張すると，採りうる手法もより多様になり，タスクもよりおもしろさを増していると言うことができよう．

なお，新語義発見に関する先行研究はあまり多くはないが，前者のアプローチを採る研究として [6] が，後者のアプローチを採る研究として [8, 5, 7] がある．

4.4 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.

4.5 今後の日程

残念ながら，最終のワークショップの日程が確定していないため，タスクの日程も確定していないが，現時点ではおよそ以下のような日程を予定している．

1. 訓練データセット，スコアラ，入出力フォーマット等を 2009 年夏にリリース予定
2. Formal run を 2009 年度後半実施予定
3. SemEval Workshop は ACL-2010 に併設予定[3]

また，ACL に併設されるワークショップは，Semeval-2 に採択されたすべてのタスクを対象にしたワークショップであるが，本タスク独自のワークショップも別途日本で開催する予定である．

5 おわりに

多くの研究者の方々の参加を期待したい．参加を検討されている場合，オーガナイザまでご連絡下されば幸いである．タスクのメイリングリストを作成する予定である．また，タスクの詳細については，今後も参加者等との議論により，必要に応じて改訂していきたいと考えているので，タスクについての要望，意見等も是非積極的にオーガナイザまでお寄せいただければ幸いである．なお，今後タスクについての情報は随時詳細が決まり次第，上述したタスクの web ページ上で告知する予定である．

文献

[1] Eneko Agirre and Oier Lopez de Lacalle. On robustness and domain adaptation using svd for word sense disambiguation. In *Proc. of COLING'08*, 2008.

[2] Yee Seng Chang and Hwee Tou Ng. Estimating class priors in domain adaptation for wsd. In *Proc. of ACL'06*, 2006.

[3] 奥村 学 and 白井清昭. 現代日本語書き言葉均衡コーパスを用いた意味解析 – 語義の自動特定，新語義の発見 –. **言語**, 37(8):66–73, 2008.

[4] 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.

[5] 白井 清昭. コーパスにおける語の意味の自動識別. **国文学 解釈と鑑賞**, 74(1):61–69, 2009.

[6] 菊田 篤史 and 白井 清昭. 未定義語義の判別を含む語義曖昧性解消. In **言語処理学会第 *12* 回年次大会発表論文集**, pages 636–639, 2006.

[7] 田中 博貴, 中村 誠, and 白井 清昭. 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. In **第 *15* 回言語処理学会年次大会**, pages P2–31, 2009.

[8] 九岡 佑介, 白井 清昭, and 中村 誠. 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. In **第 *14* 回言語処理学会年次大会**, pages 572–575, 2008.

[3] 2010 年の ACL は 7 月 11 日から 16 日まで Sweden の Uppsala で開催される予定である．

フレーム意味論と「日本語コーパス」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

小原京子（日本語フレームネット班分担者：慶應義塾大学理工学部）†
斎藤博昭（日本語フレームネット班班長：慶應義塾大学理工学部）

Constructing a Frame-based Lexicon “Japanese FrameNet” Using BCCWJ

Kyoko Ohara (Faculty of Science and Technology, Keio University)
Hiroaki Saito (Faculty of Science and Technology, Keio University)

１．本公募班の目的

本公募班の目的は、１）フレーム意味論と、代表性を有する大規模日本語書き言葉コーパス（以下「日本語コーパス」）に基づき日本語語彙意味情報資源・日本語フレームネット（以下JFNと表記）の理論的・方法論的モデルを構築し、２）それにより「日本語コーパス」の有益な活用方法を提示することである。

今年度は特に、１）JFN の構築を進め、サンプルデータを公開する、２）JFN 構築を通じて、「日本語コーパス」の活用方法を提示する、３） JFN を他の言語資源やツールと関連付ける方法を検討する、の三項目を目標とした。

２．JFNの概要

JFNプロジェクトでは、アメリカ・バークレーで構築中の英語FrameNet（以下FNと表記）と同様に、フレーム意味論の枠組みでコーパスデータを基に語彙意味分析を行っている（Fillmore 1976, 2008）。語彙意味分析の結果をタグ付けの形で「日本語コーパス」に付与し、電子語彙体系として構築する。今年度は主に「日本語コーパス」モニター公開データ 2008 年度版を使用した。

現在JFNプロジェク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「日本語コーパス」に対するタグ付けを行っている。語彙項目アノテーションとは、語彙項目ごとに「日本語コーパス」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。全文テキストアノテーションでは、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していく。今年度は従来どおり語彙項目アノテーションを主に行いつつ、後半からは全文テキストアノテーションも行った。全文テキストアノテーションについては後の 3.2.2 で述べる。

語彙項目アノテーションによる JFN 構築プロセスは主に以下の二つの局面から成る（図１）。１）語彙項目ごとに、JFN KWIC というコンコーダンサーを用いて「日本語コーパス」を検索し、共起語や結合価パターンなどを考慮しつつ、タグ付け対象とする例文を選定・抽出する；２）抽出した文に、JFNDesktop というアノテーションツールを用いて意味フ

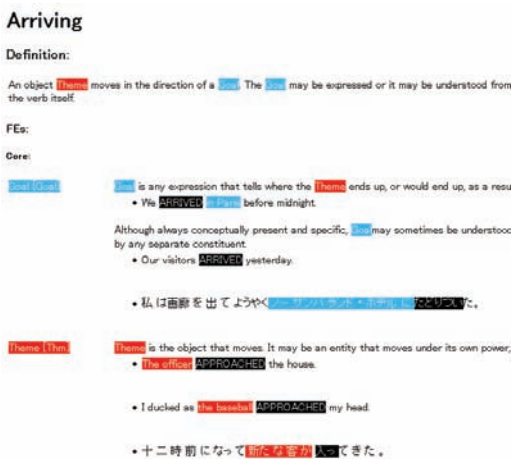
† ohara@hc.cc.keio.ac.jp

レームに基づく意味情報、統語情報、文法情報などのタグを付与する、の二つのプロセスである。

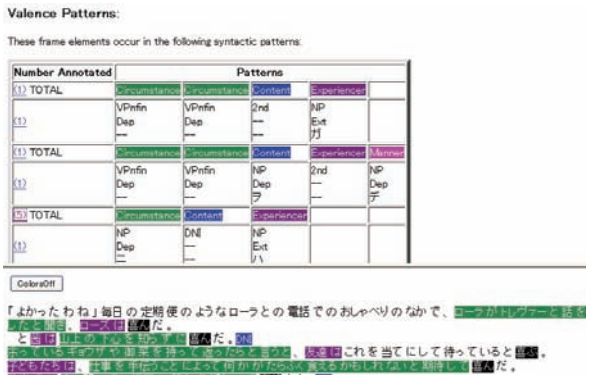
図１　JFN 構築プロセスの概要

アノテーション作業は複数の層に対して行う。フレーム要素（意味フレームに参与する意味要素）名を付与する FE（Frame Element）層、Obj などの文法機能名を付与する GF（Grammatical Function）層、NP や VP といった句タイプ名を付与する PT（Phrase Type）層、PT が NP（名詞句）の場合に格助詞名を付与する PostPos 層である。

JFNのコンテンツには、１）意味フレーム間の意味関係や、各々の意味フレームに属するフレーム要素の定義を記述した意味フレーム定義（フレームレポート）、２）語彙項目ごとのアノテーション済み例文（アノテーションレポート）、３）アノテーション済み例文から抽出した語彙項目ごとの結合価パターン情報（語彙項目レポート）がある。これらはWeb上で見ることができる（図２）。



意味フレーム定義
（例：Arriving フレーム）



結合価パターン情報とアノテーション例文
（例：語彙項目「喜ぶ」）

図２　Web上のJFNコンテンツ

３．研究の進捗状況

以下では、JFN構築の状況、JFNにおける「日本語コーパス」の活用、JFNと他の言語資源との関連付けの観点から、今年度の進捗状況について述べる。

３．１．日本語フレームネットの構築

コンコーダンサーJFN KWIC やアノテーション用ツール JFNDesktop などのツール群の整備は昨年度で一通り終わり、今年度は JFN 構築を中心に研究を進めた。昨年度は「日本語コーパス」領域内公開データ 2007 年度版を用いたが、今年度は JFN サンプルデータ公開に向けて、著作権処理が既に済んでいるモニター公開データ 2008 年度版をアノテーションに用いた。

アノテーション対象とした語彙項目の意味分野は、これまでアノテーションを行ってきた「移動」、昨年度着手した「感覚・知覚」と、今年度新たに開始した「感情」である。これまでもアノテーション対象を動詞に限定せず意味フレームごとに語彙項目を選定していたが、「移動」に関する語彙項目の大多数が動詞であったのに対し、「感情」に関する語彙項目は形容詞も多い。従って、今年度は改めて品詞別に「移動」、「感覚・知覚」、「感情」に関する語彙項目を洗い出した。また、従来は IPAL 辞書などの既存の言語資源の見出し語を参考にしつつアノテーション対象とする語彙項目を選定していたが、今年度は「日本語コーパス」デ領域内公開データにおける品詞別出現頻度順語彙表をまず作成し、その中から「移動」、「感覚・知覚」、「感情」に関する語彙項目を出現頻度の高い順に抽出した。これらを基にそれぞれの意味分野ごとに上位一定数ずつ語彙項目を選定し、それらを対象にアノテーションを行った。現在のアノテーション済み例文数は、約 2000 文である。

今年度末までに、FrameSQL というツールを使って JFN サンプルデータをアクセス制限なしで Web 上で検索表示できるようにした。FrameSQL は、専修大学の佐藤弘明氏が開発された、フレームネット形式のデータを Web 上で検索表示するためのツールである（Sato 2008）。英語、スペイン語、ドイツ語、日本語フレームネットのデータがリンクされている（図３参照）。

さらに今後は、第２節で紹介した JFN レポートシステム（図２参照）や FrameGrapher という Web 上のツールを使って JFN データを公開することも検討している。FrameGrapher は FN で開発された、フレーム間の意味関係をわかりやすく表示するためのツールで、既に JFN 用に移植済みである。第 2 節で触れたように、JFN のコンテンツにはフレーム間の意味関係の定義も含まれる。FN ならびに JFN ではフレーム間の意味関係として、継承（Inheritance）、サブフレーム（Subframe）、視点（Perspective On）、使用（Using）、使役相（Causative Of）、起動相（Inchoative）、参照（See Also）、先行（Precedes）の 8 種を定義している。例えば、`Perception_experience` フレームは `Perception` フレームと継承関係にある。図４は JFN の `Perception_experience` フレームの他フレームとの関係を FrameGrapher で表示したものである。図中の太線は継承関係を示している。

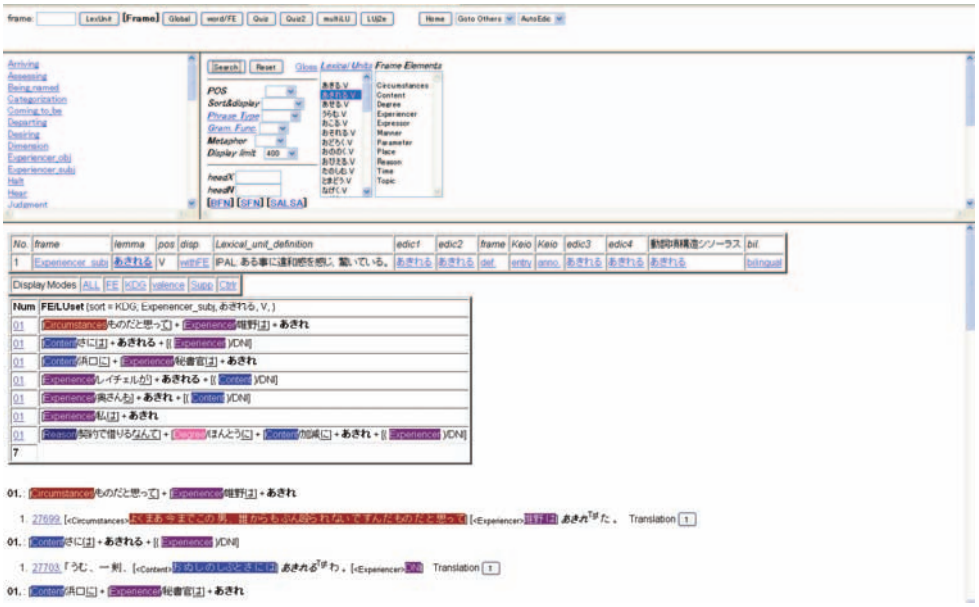


図３　FrameSQL による JFN データの表示
（例：Experiencer_subj フレームの語彙項目「あきれる」）

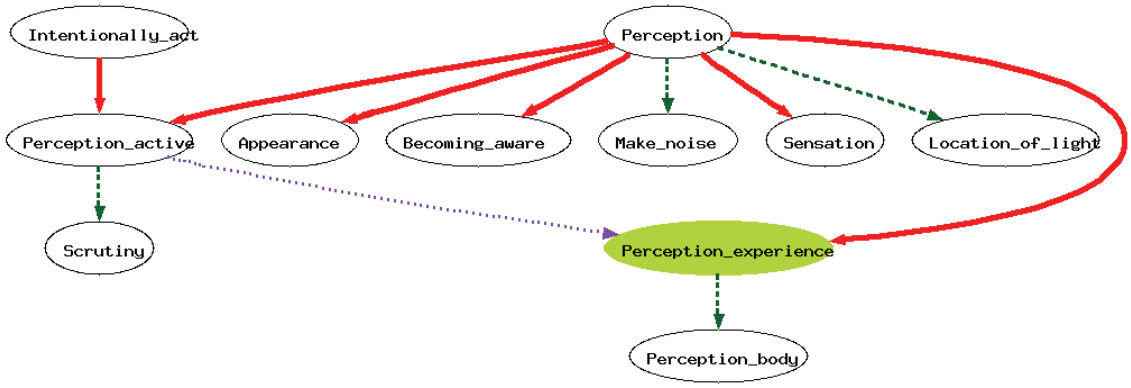


図４　FrameGrapher によるフレーム間関係の表示
(例：Perception_experience フレーム)

３．２．日本語フレームネットにおける「日本語コーパス」の活用

以下では、JFN における「日本語コーパス」活用の実態について、従来アノテーション対象としていた他のコーパスとの比較の観点からと、今年度後半に開始した全文テキストアノテーションの観点とから述べる。

３．２．１．「日本語コーパス」と他コーパスとの比較

本特定領域研究に公募班として参加する一昨年度以前は、JFN では主に新聞記事コーパスから抽出した例文に対してアノテーションを行っていた。その新聞記事コーパスと、「日本語コーパス」の書籍データ、白書データとで、それぞれの動詞の出現頻度順語彙表を作成し

たところ、動詞の意味フレームの分布に関し、ジャンル間で興味深い違いが見られた。

新聞記事、書籍データ、白書データのいずれにおいても、「いる」、「する」、「なる」の 3 動詞が上位 4 位内に入っている点では変わりがないが、それぞれのコーパスの上位 30 動詞を比べてみると、動詞の意味分類、すなわち動詞の属する意味フレームの分布に差があった。たとえば、新聞記事コーパスでは、出現頻度上位 30 動詞に「話す」、「語る」、「述べる」などの Statement フレームに属する動詞、Deciding フレームの「決める」、Request フレームの「求める」、Process_resume フレームの「始まる」が含まれるが、「日本語コーパス」の書籍データと白書データの上位 30 動詞にはこれらの動詞はいずれも含まれていない。他方、「日本語コーパス」書籍データの出現頻度上位 30 動詞の中には Becoming_aware フレームの「知る」と Perception_experience フレームの「見える」が含まれているが、新聞記事コーパスと「日本語コーパス」白書データの上位 30 動詞にはどちらも含まれない。また、「日本語コーパス」白書データでは Attempt フレームの「図る」と「努める」、Change_position_on_a_scale フレームの「達す」と「高まる」が出現頻度上位 30 位内に入っているが、書籍データや新聞記事コーパスの上位 30 位には入っていない。

また、同一語彙項目の結合価パターンについても、新聞記事コーパスと「日本語コーパス」間では違いがみられた。[1]

以上、「日本語コーパス」の均衡性・代表性について、語彙項目の意味フレーム分布や、同一語彙項目の結合価パターンのバリエーションを尺度とした評価の可能性を示唆した。書籍、白書、Yahoo!知恵袋、国会議事録、検定教科書に加え、学術論文、科学技術論文などのジャンルのテキストも加えると、さらに「日本語コーパス」の均衡性向上に寄与するのではないかと考える。

３．２．２．全文テキストアノテーション

従来の語彙項目アノテーションによる JFN 構築に加えて、今年度は全文テキストアノテーションも行った。第 2 節で述べたように、全文テキストアノテーションとは、テキスト内のすべての文の、意味フレームを喚起（evoke）するすべての語彙項目に対してアノテーションを行うことである。ただし、固有名詞などは対象とせず、あくまでも意味分析の観点から興味深いと思われる語彙項目に限定してアノテーションを行った。今年度対象としたのは「日本語コーパス」コアデータの一部サンプルである。目下コアデータは書籍と白書の二つのジャンルから構成されているが、今後は Yahoo!知恵袋、国会会議録、検定教科書などのジャンルについても全文テキストアノテーションを行いたい。

例文(1)と(2)は、「日本語コーパス」上の塩野七生著『ローマから日本が見える』のサンプルに現れる連続した文である。下線を施した合計 7 個の語彙項目がアノテーション対象（ターゲット）である。語の右下に大文字で記したのが各々のターゲットが喚起する意味フレーム名である。1)から 7)で、太字の斜字体と右肩の Target の文字で記されているのがターゲットである。また、ターゲットが喚起する意味フレームのフレーム要素に相当するものは[ ]

[1] 反対に、ジャンルや動詞の意味フレームを問わず幅広く見られる現象としては、引用の「と」がある。引用の「と」を現在JFNでどう取り扱っているかについてはOhara & Suzuki (To Appear)を参照のこと。

で囲んだ上で左下にそのフレーム要素名を記した。たとえば 2)では、動詞「出る」が Coming_to_be フレームを喚起するターゲットとして分析されている。そして、文中の「ちょうど」と「アントニウスの名前が」はそれぞれ Coming_to_be フレームのフレーム要素 TIME とフレーム要素 ENTITY に相当するものとしてアノテーションされている。

(1) ちょうど*アントニウスの* 名前BEING_NAMED 1) *が* 出COMING_TO_BE 2) *てきたので、彼についても* 解説STATEMENT 3) 願いREQUEST 4) *ましょう。*

(2) *この* 人物PERSON 5) *に対しても、あなたはずいぶん* 低いPOSITION_ON_A_SCALE 6) 評価ASSESSING 7) *を付けていますね。*

1) ・ちょうど [Entity アントニウスの] ***名前*** Target が 出てきたので、彼についても 解説願いましょう。

・ちょうど [Name アントニウス] の ***名前*** Target が 出てきたので、彼についても 解説願いましょう。

2) [Time ちょうど] [Entity アントニウスの 名前が] ***出*** Target てきたので、彼についても解説願いましょう。

3) ちょうどアントニウスの 名前が出てきたので、 [Topic 彼についても] ***解説*** Target 願いましょう。[Message CNI] [Speaker CNI] [Addressee CNI]

4) ちょうどアントニウスの 名前が 出てきたので、[Topic 彼についても] [Message解説] ***願い*** Target ましょう。

5) この [Person ***人物***] Targetに対しても、あなたはずいぶん 低い評価を付けていますね。

6) この 人物に対しても、あなたは[Degree ずいぶん] [Value ***低い***] Target [Variable 評価] を付けていますね。

7) [Phenomenon この人物に対しても]、[Assessor あなたは] ずいぶん 低い ***評価*** Target を付けていますね。

全文アノテーションで問題となったのは、語彙項目の意味情報と構文の持つ意味情報との相関関係を JFN の枠組みでどう記述するかである。これについては、FN の現行の方針にのっとり、JFN での記述の試案を Ohara (2008)にまとめたので参照されたい (Ohara 2008)。

全文テキストアノテーションを「日本語コーパス」データを対象に行うことで、フレーム意味論に基づく意味タグ付き「日本語コーパス」が作成できることになる。また、「日本語コーパス」のジャンルごと、サンプルごとに、語彙の意味フレーム（語義）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（前節参照）。さらに、「日本語コーパス」コアデータの同じサンプルに対し、たとえば本特定領域研究ツール班の述語項構造と共参照情報の枠組み (飯田他 2008) などでもアノテーションを行うことができれば、それぞれのアノテーションの枠組みを比較したり、両方のアノテーションを組み合わせることによる意味タグ付き「日本語コーパス」としての有用性を検討したり、といったことが可能となる。

３．３．他の言語資源との関連付け

JFNデータと、本特定領域研究計画班の竹内孔一氏、乾健太郎氏ら作成の動詞項構造シソーラスのデータとの対応づけを行った（竹内他 2008）。具体的には、専修大学の佐藤弘明氏のご好意により、氏の開発されたFrameSQL上で、JFNデータから対応する動詞項構造シソーラス上のデータへリンクを張っていただいた。図５にその例を示す。ここでは、JFNのExperiencer_subjフレームの語彙項目「落ち着く」と同じ見出し語「落ち着く」のエントリーとして動詞項構造シソーラス上にどのような記述があるかを表示させている。ここでは、JFNのExperiencer_subjフレームの語彙項目「落ち着く」に対応する動詞項構造シソーラス上のエントリーはID7629 である。

JFN データを動詞項構造シソーラスデータと対応づけることで、JFN と動詞項構造シソーラスの語義記述や、「日本語コーパス」データから抽出した JFN アノテーション例文と動詞項構造シソーラス上の例文とを比較対照できるようになった。

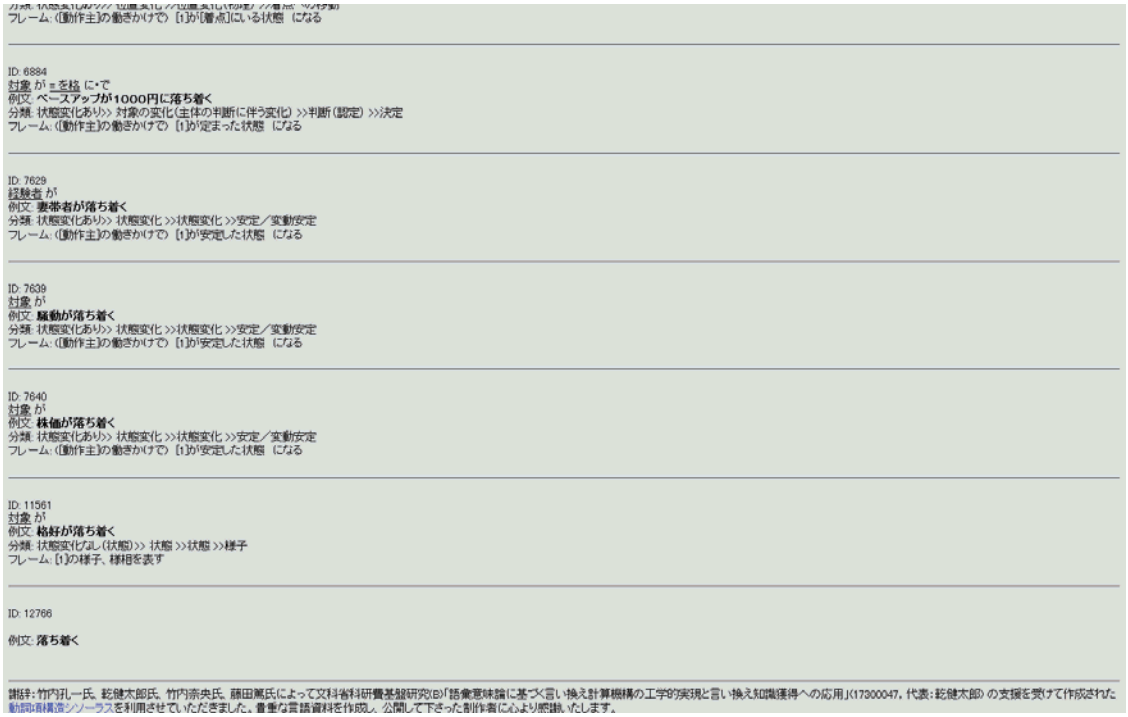


図５　FrameSQL を用いた JFN データからの動詞項構造シソーラスデータの参照
（例：「落ち着く」）

４．今後の展望

日本語フレームネット班の今年度の進捗状況について、JFN の構築、JFN における「日本語コーパス」の活用、JFNと他の言語資源とのリンクの観点から報告した。今年度は特にサンプルデータ公開に向けて語彙項目アノテーションを推進すると同時にデータの見直しを行った。また、後半には「日本語コーパス」コアデータの全文テキストアノテーションを開始した。今年度末までの研究で、文の意味を記述するには文中の自立語の意味情報を記述するのみでは不十分で、構文の持つ意味情報についても記述する必要があることがわかった。従って、今後は従来の語彙データベース（Lexicon）に加え、構文データベース

（Constructicon）をも含んだ複合資源としてのJFNの可能性を探って行きたい。

日本語フレームネット班は、班内で工学系研究者と言語学系研究者が協力して意味タグ付きコーパスを作成している点が特徴と言える。今後も「日本語コーパス」を用いて、コーパスに基づく言語分析や自然言語処理に利用できる言語コンテンツ作成に努力したい。

今年度発表論文

Saito, Hiroaki, Shunta Kuboya, Takaaki Sone, Hayato Tagami, Kyoko Ohara (2008). "The Japanese FrameNet Software Tools." *6th International Conference on Language Resources and Evaluation* (LREC2008). Marrakech, Morocco. May 28th, 2008.

Ohara, Kyoko Hirose (2008). "Lexicon, Grammar, and Multilinguality in the Japanese FrameNet." *6th International Conference on Language Resources and Evaluation* (LREC2008). Marrakech, Morocco. May 30th, 2008.

小原京子(2008).「コーパスに基づく日本語主観移動表現のフレーム意味論的分析：英語との比較から」日本認知科学会第 25 回大会発表論文集. pp. 16-17.

Ohara, Kyoko Hirose (2008). "Representing lexicon and grammar in Japanese FrameNet." *Fifth International Conference on Construction Grammar* (ICCG-5). University of Texas, Austin. September 28th, 2008.

Ohara, Kyoko Hirose (2008). 「日本語主観移動表現のコーパス分析：英語との比較から」『慶應義塾大学日吉紀要　言語・文化・コミュニケーション』第４０号. pp. 107-122.

参考文献

飯田龍，小町守，乾健太郎，松本裕治（2008）．「新聞 3 千 4 万文への述語項構造と共参照情報のアノテーション」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.15-16.

竹内孔一, 乾健太郎, 竹内奈央, 藤田篤（2008）.「意味の包含関係に基づく動詞項構造の細分類」言語処理学会第 14 回年次大会予稿集. pp.1037-1040.

Fillmore, Charles J. (1976). "Frame semantics and the nature of language." In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech.* Vol. 280: 20-32.

Fillmore, Charles J. (2008). "FrameNet meets construction grammar." In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII Euralex International Congress* (49-68). Barcelona: Institut Universitari de Lingüistica Aplicada.

Ohara, Kyoko Hirose (2008). "Representing lexicon and grammar in Japanese FrameNet." *Fifth International Conference on Construction Grammar* (ICCG-5). University of Texas, Austin. September 28th, 2008.

Ohara, Kyoko Hirose, and Ryoko Suzuki (To Appear). "A Corpus-based account of quotatives in Japanese FrameNet," To be presented in the theme session "Advances in Frame Semantics." *11th International Cognitive Linguistics Conference* (ICLC11). University of California at Berkeley.

Sato, Hiroaki (2008). "New Functions of FrameSQL for Multilingual FrameNets." *6th International Conference on Language Resources and Evaluation* (LREC2008). Marrakech, Morocco. May 28th, 2008.

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/

日本語リーダビリティー公式の構築と測定ツールの開発

柴崎秀子（リーダビリティー班班長：長岡技術科学大学工学部）†

Developing Japanese Readability Formula and Measuring Tools

Hideko Shibasaki (Nagaoka University of Technology)

１．日本語リーダビリティー公式の構築

今日までに構築されたリーダビリティー測定方法は 200 以上もあり，①公式で計算する方法，②グラフから測定する方法，③文字の出現率を使ったモデル法などがあり，測定値の示し方は，①学年レベルで示す方法と②ポイントで示す方法がある。この中で最も一般的なのは公式を使って学年レベルを計測するという方法である。本研究では，学年予測値を算出するために，小学 1 年から高校 3 年までの国語教科書 51 冊の読解教材でコーパスを作成し，変数と考えられるものを学年別に分析してみた。分析したものは，①1 文の平均文字数，②テキストの中の文字種(漢字・平仮名・片仮名・ローマ字)の割合，③語種(和語・漢語・外来語・混種語)の割合，④1 文の平均述語数，⑤係りの文節と受けの文節の距離，⑥1 文の平均文節数であるが，⑤以外の変数は，小学 1 年から中学 3 年までほぼ線形に近いことが認められたので，⑤以外の変数で線形重回帰式を使って，9 学年を予測する公式を構築した。9 学年分の 245 データの中には，一部の変数が同じ学年のテキストから大きく逸脱するものもある。例えば小学 3 年生の「じゅげむじゅげむ」は 1 文の文字数が極端に多い。そこで，まず外れ値を取り除くために，⑤を除いた①から⑥の 5 変数を予測変数，学年を従属変数として強制投入法で線形重回帰分析を行い，前予測値が±2 以上の 38 データを外れ値として除外した。次に，残った 205 データで再度，線形重回帰分析を行った（表 1，表 2）。今度は予測力のない変数を除外するため，ステップワイズ法を用いた。その結果，③が除外され，①，②，④，⑥が変数として残った。決定係数は R^2=0.858 で，この 4 変数による予測力は極めて強いことが示された。このような方法で以下の公式が構築された。

$Y=-0.148X_1+1.585X_2-0.117X_3-0.126X_4+15.581$

Y=学年

X_1＝文章中の平仮名の割合

X_2＝1 文の平均述語数

X_3＝1 文の平均文字数

X_4＝1 文の平均文節数

本研究で得られたリーダビリティー公式の信頼性について検討した。その方法はサンプルを無作為抽出し，他のサンプル及び実学年と比較する方法である。まず，この公式を使って小学１年から中学３年までの全205テキストのリーダビリティー値を求め，無作為に各学年から9つずつ選択して3つのグループに分けた。つまり1つのグループに27テキストが入ることになる。実学年(M=5.00)とランダム抽出サンプル１の公式から算出した学年(M=5.03)との相関係数(n=27)は 0.920 (p<.001)，ランダム抽出サンプル2(M=4.81)との相関係数は0.917(p<.001)，ランダム抽出

† shibalea@vos.nagaokaut.ac.jp

サンプル3(M=5.00)との相関係数は0.952（p<.001)と，非常に高い相関関係を示した。次に，実学年とランダム抽出サンプルを対応のあるサンプルの t 検定を行った。その結果，実学年とランダム抽出サンプル1の学年に有意な差はなく[t(26)=-0.127, $n.s.$], また実学年とランダム抽出サンプル2でも有意な差はなく[t(26)=0.926, $n.s.$]，さらに実学年とランダム抽出サンプル3にも有意な差はなかった[t(26)=-0.013, $n.s.$]。以上の結果から，ランダムに抽出した27種類のテキストについてのリーダビリティー測定値と実学年には有意な違いはなく，本研究の公式の信頼性は高いことを示していると言えよう。

表1 205ｻﾝﾌﾟﾙの各変数における平均と標準偏差

学年		1文の文字数		1文の文節数		平仮名の割合(%)		漢語の割合(%)		1文の述語数	
	n	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差
1	25	20.88	4.29	4.43	1.17	94.80	4.75	5.86	3.62	1.56	0.32
2	25	21.79	6.24	4.92	1.91	84.24	5.32	8.11	5.00	1.61	0.40
3	25	29.43	7.92	6.56	2.32	80.08	5.07	15.11	5.81	2.42	0.66
4	13	26.90	5.32	6.54	4.41	79.54	4.14	14.78	5.86	3.10	0.59
5	20	31.17	7.67	6.77	2.12	75.70	6.33	19.14	5.38	3.62	0.97
6	28	32.01	8.41	7.13	1.94	69.79	6.43	24.84	11.45	3.59	0.85
7	31	35.28	8.17	7.99	1.86	67.71	5.63	26.59	9.64	4.00	0.78
8	25	35.23	10.12	8.08	2.46	63.76	5.09	30.02	9.13	3.82	1.07
9	13	42.12	9.16	9.39	2.22	59.92	3.87	38.02	8.80	4.61	0.85

表2 テキストの配当学年を4変数で予測する重回帰分析の結果

変数名	β	t 値	有意確率
X_1 文章中の平仮名の割合	−0.674	−17.397	p <.001
X_2 1文の平均述語の数	0.775	12.009	p <.001
X_3 1文の平均文字数	−0.445	−5.827	p <.001
X_4 1文の平均文節数	−0.129	−2.282	p <.05
重決定係数		R^2=.858	

注: 表の重回帰分析はステップワイズ法. n=205. β は標準偏回帰係数.

２．測定ツールの開発

日本語リーダビリティー公式を応用し，測定ツール(http://readability.nagaokaut.ac.jp)を開発した。1 文の平均述語数は形態素解析で，1 文の平均文節数は CaboCha を用いて算出した。形態素解析には当初は ChaSen を使っていたが，小学校低学年の教材には平仮名が連続していることが多く，方言や口語表現も多いため解析がうまくできなかった。そこで，形態素解析の辞書を標準の IPA 辞書 から UniDic(http://www.tokuteicorpus.jp/dist/)に変更し，形態素解析器も MeCab に切り替えた。その結果，全体の解析精度は大きく向上した。アプリケーションは、Ruby を用いて，Web アプリケーションとして製作した。ツールの開発や各変数の定義は詳細については，李・長谷部・柴崎（2009）を参考にされたい。図 1 は 2008 年にベストセラーとなった『ホームレス中学生』（田村裕著）の測定結果であるが，リーダビリティー値は 5.98 と算出されたので，だいたい小学 6 年レベルだと言える。

３．他のリーダビリティー測定ツールとの比較

現在公開されている日本語リーダビリティー測定ツールとしては，名古屋大学佐藤研究室の「ことば不思議箱」と，建石・小野・山田（1998）の公式を応用したテキサス州立大学の Accessibility Institute ホームページの「Tx-Readability」がある。本研究で開発したツールを（A)，「ことば不思議箱」を（B)，「Tx-Readability」を（C）とすると，各ツールの特徴は以下のようにまとめられる。

（A）小学 1 年から中学 3 年までの 9 レベルを測定。文の長さ，文法構造，文字種の割合で学年を予測。小学 3 年から中学 1 年までの予測精度は高いが，高校レベルを予測できない。国語教科書の読解教材をデータベースとしている。

（B）小学 1 年から大学までの 13 レベルを測定。文字の出現確率によるモデル法。全科目の教科書をデータベースとしている。小学 1 年から 3 年までの予測精度は極めて高いが，中学以上の予測値には正確さを欠く。特に高校レベルは 3 学年で差別化ができない(柴崎他 2008)。

（C）0 から 100 までのポイントで難易を示す。文の長さ，文字種，同一文字種の連続性で測定する。100 が最も易しく，0 が最も難しいが，各数値は相対的な指標であり，具体的なレベルを示すものではない。小学校低学年の予測が不可能(柴崎他 2008)。データベースは不明。

この中で，（A）と（B）がどちらも学年を予測するので，精度を比較してみたい。まず，（A）の公式を構築する際に外れ値として除外した 38 テキストも含んだ 243 テキストにおける①1 から 9 までの実学年，②（B）で得られるリーダビリティー値，③（A）の公式で得られたリーダビリティー値を一元配置の分散分析を使って分析した。③でリーダビリティー値が 1 以下の小数になったサンプルについては，1 年から 9 年までを予測するという公式の性格上，整数のみを用いるべきであると判断し，1 としてカウントした。また，（B）でも，整数を算出しているので，直接比較のために整数とするのが妥当であると判断した。実学年に対して，同じテキストを（A）と（B）でリーダビリティーを測定しているので，反復測定であると考えた。分析の結果，主効果が有意であった[$F(2, 484)=7.171$, $p<.001$]。そこで，3 つの値について，単純対比で比較した。その結果，実学年(M=5.15)と（B）で算出された学年予測値(M=4.82)には有意な違いがみられた[$F(1, 242)=8.420$, $p<.01$]が，実学年と（A）柴崎で得られた学年予測値(M=5.16)には有意な違いはなかった[$F(1, 242)=0.013$, $p=.909$, *n.s.*]。つまり，本研究の公式から得られた学年予測値は実学年と有意差がないということである。以上の結果から，小学 1 年から中学 3 年までのテキストのリーダビリティー測定は，本研究で構築されたツールの方が「ことば不思議箱」よりも，より正確に実学年を予測すると言える。

具体的な作品を本研究で開発したツール（A）と「ことば不思議箱」（B）とで計測し，問題点を考えてみたい。第一には語彙の問題がある。サン・デグジュペリの「星の王子さま」は 2005 年に翻訳著作権が切れた後，多くの翻訳本が出版されている。その中で 3 つを選んだところ，倉橋由美子訳は（A）が 5.38 で（B）が 5，三田誠広訳は（A）が 4.11 で（B）が 3，内藤濯訳は（A）が 2.4 で（B）が 2 というように，ほぼ一致していた。以下に 3 つの翻訳の冒頭部分を示したが，内藤濯訳が 2 年生レベルと測定されたのは，読者の直感としては低いように思う。

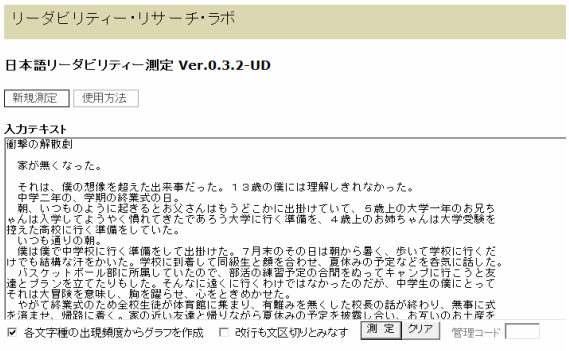


図 1　測定ツール入力画面

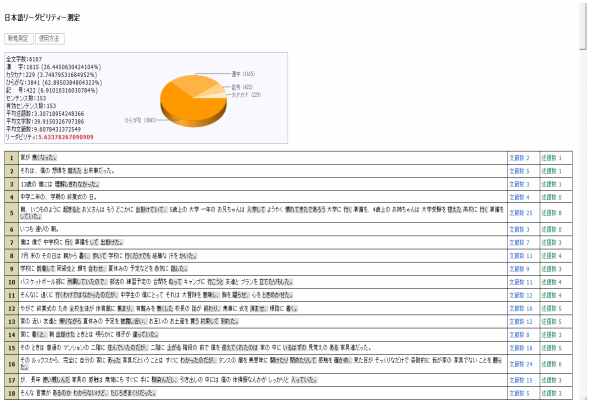


図 2　「ホームレス中学生」測定結果

＜倉橋由美子訳＞
六歳のとき、ジャングルのことを書いた『ほんとうにあった話』という本の中で、すごい絵を見たことがある。それは一匹の獣を呑みこもうとしている大蛇の絵だった。ここにその写しがある。その本には、「大蛇は獲物を噛まずに丸呑みにし、その後は動けなくなって、半年の間眠っている。その間に呑みこんだ獲物が消化される」と書いてあった。
＜三田誠広訳＞
「ヒツジの絵を描いて。」と王子さまは言った。昔のことだけどね。ぼくは六歳だった。だれも行ったことのない森の奥のことを書いた、不思議な本の中で、すごい絵を見つけたんだ。『ほんとの話』という題のその本には、ボアという巨大なヘビが、けものを、いままさにのみこもうとしている絵が出ていた。これがその絵だよ。こんなことも書いてあった。「巨大なヘビはえものをかみくだくこともなく、つるんとまるごとのみこんでしまう。そのあとは身動きできなくなって、のみこんだものが消化されるまで、半年間、ひたすら眠りつづける。」
＜内藤濯訳＞
六つのとき、原始林のことを書いた「ほんとうにあった話」という、本の中で、すばらしい絵を見たことがあります。それは、一ぴきのけものを、のみこもうとしている、ウワバミの絵でした。これが、その絵のうつしです。その本には、「ウワバミというものは、そのえじきをかまずに、まるごと、ペロリとのみこむ。すると、もう動けなくなって、半年のあいだ、ねむっているが、そのあいだに、のみこんだけものが、腹のなかでこなれるのである」と書いてありました。

内藤訳は漢字が少なく平仮名が多いために小学 2 年レベルと測定されたが，1 文の文字数は多く，「ウワバミ」「けもの」「うつし」など馴染みがない語が使われていることから，もっと学年レベルは高いように思われる。NTT 語彙データベースで単語親密度を測定すると，「獣」と「写し」という表記なら単語親密度は 5.594 という結果であるが，作品中の表記では該当語がなかった。李・柴崎（2008）でも提案されているが，今後，リーダビリティーの測定には語彙の難易を入れていく必要があると考える。

第二には，（A）は高校レベルの学年が予測できず，（B）は高校レベルの 3 学年で差別化ができない(柴崎他 2008)という問題がある。（A）は当初，高校の 3 学年レベルを国語総合，現代文Ⅰ，現代文Ⅱをデータベースとして尺度を作ることを考えていた。しかし，この分類では学年間の差異が明確にならず，線形回帰分析にも適さないことがわかった。高校は普通高校，職業高校，高等専門学校など学校によって特色があり，教科書も統一されているわけではない。また，作品によっては，現代文Ⅰと現代文Ⅱの両方に収められているものもあり（例：夏目漱石『こころ』，丸山真男『「である」ことと「する」こと』），学年でレベルを設定するのは難しい。本研究では，中学以上の教科書は学年による差異よりも，テキストの種類，すなわちが文学的文章か説明的文章かという差異のほうが大きいのではないかと考え，学年レベルでなくテキストの種類による分類を試みた。中学の国語教科書（三省堂，光村出版，東京書籍の 3 種類）と高校の国語総合，現代文Ⅰ，現代文Ⅱ（大修館書店，明治書院）に収められている読解を目的とする教材を，小説，紀行文，随筆を文学的文章，科学的な内容を説明したもの，何らかの事実や事象に対する論述文，意見文を説明的文章として分類し，平仮名の割合，1 文の平均文字数，漢語の割合を算出し，学年別に分類した場合と比較してみた（表 3，表 4）。テキストの種類で 4 レベル（中学文学的文

章，高校文学的文章，中学説明的文章，高校説明的文章）に分類すれば，レベル間の差異が明確である。この結果により，我々は今後中学以上のレベルを学年ではなく，テキストの種類で分析していく方向で考えている。

表3 中学高校用テキストの学年別変数

学年		平仮名の割合		漢語の割合		1文の文字数	
	テキスト数	平均	標準偏差	平均	標準偏差	平均	標準偏差
中学1年	34	67.71	5.63	26.59	9.64	35.28	8.17
中学2年	31	63.76	5.09	30.02	9.13	35.23	10.12
中学3年	28	59.92	3.87	38.02	8.80	42.12	9.16
国語総合	25	66.04	5.41	31.10	11.23	39.97	13.09
現代文Ⅰ	36	66.26	4.91	33.68	10.26	38.91	11.73
現代文Ⅱ	36	64.30	5.82	34.86	9.62	43.96	13.87

表4 中学高校用テキストの種類別変数

テキストの種類		平仮名の割合		漢語の割合		1文の文字数	
	テキスト数	平均	標準偏差	平均	標準偏差	平均	標準偏差
中1文学	19	72.47	3.56	20.05	8.06	27.36	6.76
中2文学	14	68.35	3.45	21.30	3.75	24.53	6.95
中3文学	13	70.51	3.58	20.06	5.61	29.86	14.02
国語総合・文学	13	67.93	3.36	24.77	7.95	34.02	10.26
現代文Ⅰ・文学	17	68.71	4.05	27.15	7.84	33.72	10.47
現代文Ⅱ・文学	14	67.13	5.16	28.06	7.87	36.88	12.38
中1説明	15	64.71	5.17	32.74	6.62	36.40	7.27
中2説明	17	64.43	6.53	32.24	8.87	36.00	4.47
中3説明	15	63.20	5.57	34.04	9.91	38.09	8.31
国語総合・論述	12	63.50	6.45	39.86	8.31	48.75	11.25
現代文Ⅰ・論述	19	63.37	4.30	41.64	7.41	44.07	10.64
現代文Ⅱ・論述	22	62.00	5.24	40.43	7.19	49.56	12.27

中学，高校レベルのテキストを分析するもう 1 つの方法として，テキストに出現した漢字をレベル別に分類してみた。漢字は 11 レベルの難易としたが，この分類は以下のように行った。まず，小学 1 年から 6 年までの教育漢字を学年配当別にレベル 1 から 6 までの 6 段階とし，中学以上は学年配当がないので，教育漢字以外の常用漢字で漢字検定 4 級相当をレベル 7，3 級相当をレベル 8，2 級相当をレベル 9 とし，さらに，常用漢字以外の JIS 第一水準漢字（漢字検定準 1 級相当）をレベル 10，JIS 第二水準漢字（漢字検定 1 級相当）をレベル 11 とした。中学，高校の国語教科書に収められた 190 の読解教材に出現した 197510 漢字（異なり字）をテキストごとに 11 段階に分類し，各段階の漢字数の全体の漢字数に対する割合を算出した。テキストを中学文学的文章，高校文学的文章，中学説明的文章，高校説明的文章で分類したところ，中学は図 3 のように，高校は図 4 のように分布が示された。点線が文学的文章，実線が説明的文章であるが，以下の点が観察される。まず，全体にレベル 2 とレベル 3 の出現割合が極めて高い。これは漢字 2 字熟語を構成する漢字に小学 2 年，3 年配当漢字が多いからである。次に，中学，高校ともに文章のタイプによって，出現した漢字のレベルが近似している。レベル 1 とレベル 2 では文学的文章における漢字出現割合が説明的文章よりも高いが，反対に，レベル 3，レベル 4，レベル 5 では説明的文章のほうが高い。レベル 6 でほぼ同じになり，レベル 7 以上の難しい漢字は文学的文章での出現率が高い。これは，文学的文章が説明的文章よりも漢字の難易に幅があることを示唆していると言えよう。最後に，毎日新聞記事（2006 年版）に出現した漢字を同じように 11 段階に分類したところ，漢字の出現割合は高校の説明的文章と極めて似ているという傾向が示された。このことにより，高校の国語教科書に使用されている漢字語彙は新聞の漢字語彙に近いことが示唆されたと言えよう。以上の漢字のレベル別分類の結果からも，中学，高校教科書については，学年間の差異よりもテキストの種類により差異でリーダビリティーの難易を決めるべきではないかと考えられる。

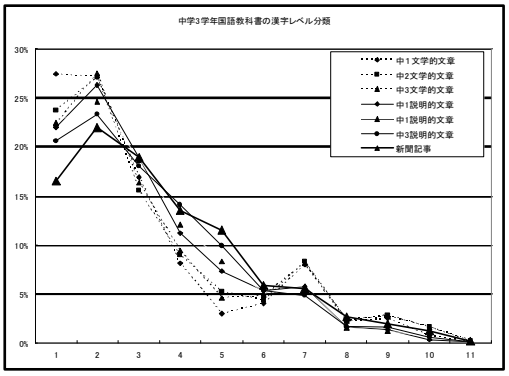


図 3　中学教科書の出現漢字レベル

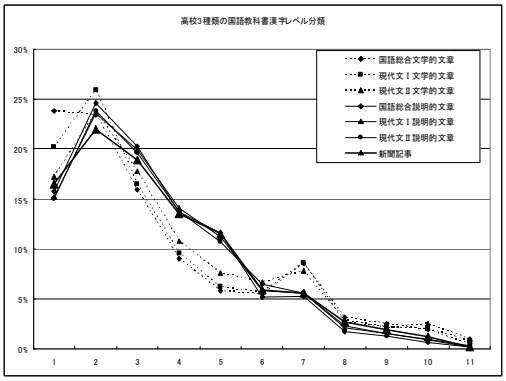


図 4　高校教科書の出現漢字レベル

４．今後の計画

これまでデータベースの作成から始めて，日本語リーダビリティーの変数を探し，公式の構築，ツールの開発まで進めてきた。現段階のツールでは小学生，中学生を対象にした図書の選択に役立てることができる。しかし，読み手が成人の場合，教育漢字が平仮名で表記された場合，文の正誤判断課題では肯定反応でも否定反応でも，漢字表記よりも反応時間がかかる（柴崎・玉岡・沢井，2008）ということが報告されている。リーダビリティー公式は１つにとどまるものではなく，成人か児童か，日本語母語話者か非母語話者か，非母語話者の場合，母語は漢字圏か非漢字圏かによって，複数の公式が必要であると考えている。

文献

建石由佳・小野芳彦・山田尚勇。(1988)「日本文の読みやすさの評価式」文書処理とヒューマンインターフェース 18-4, p1-8

李 在鎬・長谷部 陽一郎・柴崎 秀子(2009)「読解教育支援のためのリーダビリティー測定ツールについて」2009 年言語処理学会大会予稿集（印刷中）

柴崎秀子(2008)「平成 20 年度研究進捗状況報告：リーダビリティー班日本語コーパスを応用した文章の難易測定の研究」2008 年特定領域「日本語コーパス」全体会議予稿集 p125-130

李 在鎬・柴崎 秀子（2008）「日本語リーダビリティー公式構築のための国語教科書語彙の分析」計量国語学会第 52 回年次予稿集 p16-22

柴崎秀子・玉岡賀津雄・沢井康孝（2008）「漢字表記と平仮名表記が文正誤判断課題に与える影響―文字種による日本語リーダビリティー公式構築のための基礎研究－」，言語科学会年次大会予稿集 p18

グラフクラスタリングを用いた語義別用例分類

佐々木稔（クラスタリング班分担者：茨城大学工学部）†
新納浩幸（クラスタリング班班長：茨城大学工学部）

A System for Clustering Examples of Word Sense Using Graph-based Clustering

Minoru Sasaki (Faculty of Engineering, Ibaraki University)
Hiroyuki Shinnou (Faculty of Engineering, Ibaraki University)

１．はじめに

ある単語について大量の用例文を抽出し、その用例文集合を語義に基づいて自動的に分類する手法の開発を行っている。国語辞典では知りたい意味についての用例が少なく、多くの語義別用例を分類、提供することでユーザは理解を深めることができる。また、用例文集合の中には、辞典には記載されていない語義として利用された用例も存在するため、辞書の改訂作業を支援することにも応用が可能である。

用例文集合から自動的に語義を分類するために利用される方法として、教師あり学習や半教師ありクラスタリングなどが存在する。教師あり学習ではあらかじめ手作業で語義ごとに分類した教師データを利用して新しい用例を分類する。その場合、大量の教師データを語義ごとに用意することや新しい語義を見つけることへの対応は難しい。また、半教師ありクラスタリングでは、分類する用例文集合内において少量の制約条件を与えることで、大量の教師データを必要とせず、新しい語義への対応も可能な語義別の分類を行うことができる。

これらの手法の他に、最近では HyperLex などといったグラフ分割によるクラスタリングを利用して共起する単語の分類を行う手法が提案されている（Véronis 2004）。グラフを利用する場合、用例文集合に出現する単語をノード、２単語の共起頻度をエッジとして表現する。このようにして表現したグラフにおいて、高い頻度で共起する単語の組は同じ語義で使用されていると仮定して、共起頻度の高いエッジを持つノード群を求める。それらのノード群に対応する単語集合が同じ語義で共起しやすい単語の集合となる。近年、グラフクラスタリング技術は高速化や性能向上が盛んに行われ、これを使った語義分類手法も提案されている（Agirre 2007）。しかし、平均 3〜4 個の語義を持つ単語に対してクラスタリングを行ったとしても、ほとんど 1 個の語義しか得られないことが多い。そのため、多くの語義を持つ単語に対しては、性能の高いクラスタリングを行うことができない問題がある。これは単語共起のみでグラフを表現すると、頻度の高い共起のみがクラスタとして残るが、頻度は低いが特徴的な単語共起を捉えることが難しく、語義の異なりが区別できていないと考えられる。

本稿では、従来手法である出現単語をグラフにより表現する手法における問題点を改善するために、出現単語を概念に変換したグラフに対してクラスタリングを行い、その結果を基に用例文集合を語義毎に分類して表示するシステムを提案する。具体的には各出現単

† msasaki@mx.ibaraki.ac.jp

語に対して分類語彙表（国語研究所 2004）の分類番号を検索し、その分類番号をグラフ内に追加し、それにより得られたグラフ構造に対してクラスタリングを行う。そのクラスタリング結果に対して、各クラスタ内の単語を含む用例文を導き、用例文を語義毎に分類することを試みる。この用例文のクラスタリング結果を分析し、用例文集合から語義別用例文集合への分類性能を調査する。

２．システム概要

本研究における目標のひとつとして、入力単語に対して得られる用例文集合を語義別にクラスタリングして、分類結果を表示するシステムを作成することがある。本節ではそのシステムの概要について説明する。まず、用例文を分類したい単語を入力すると、その単語を含む用例文が検索される。その中に含まれる入力単語以外の共起単語に対して、グラフ構造を作成し、グラフクラスタリングを行う。その結果、単語のクラスタが得られるため、クラスタ内の単語を含む用例文を導き、最終的な出力結果とする。本稿では、このシステムを用いて用例文集合から語義別用例文集合への分類性能を調査することが目的となる。本節ではこの用例文クラスタリングシステムについて、グラフ構造作成部と名詞クラスタリング部、用例文クラスタリング部に分けて述べる。

２.１　グラフ構造作成部

キーワードの用例文集合に対して、形態素解析器 mecab[1] を用いて用例文ごとに形態素解析し、名詞だけを抽出する。抽出された名詞を、分類語彙表を用いて概念に変換する。ここで、用例文ごとに抽出された名詞は、キーワードと共起しているものと仮定する。

分類語彙表には、一つの単語に対して複数の分類番号が記載されていることがある。分類番号が一意でない場合、どの分類番号に変換するのかが問題となる。しかし、どの分類番号に変換するべきかという情報は持っていないので、全ての分類番号が出現する確率は一様分布に従うと仮定し、同じ重要度を与える。このように、一つの単語が複数の概念番号に変換されることがある。また、分類番号は 7 桁のうち上 5 桁までを使用して、同じ分類番号の単語をまとめることとする。

全ての概念の延べ出現数と、全ての概念の出現頻度を数え、一文に出現する全ての概念の組み合わせに対して相互情報量（村田 2005）を求める。これにより、各概念をノード、一文で共起する概念の組み合わせをエッジとし、エッジの重みを相互情報量とする重みつき無向グラフが作成される。

２.２　名詞クラスタリング部

作成した概念のグラフ構造に対して、グラフ分割によるクラスタリングを行う。クラスタリング手法には、マルコフクラスタリング（Véronis 2000）、または、Normalized Cut アルゴリズム（Dhillon 2004）を利用してクラスタリングを行う。これにより、グラフ内において繋がりの強い概念は同じクラスタにまとめられる。

マルコフクラスタリングは、概念の集合をいくつのクラスタに分類するかを指定しなく

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

ても分類が可能であるが、Normalized Cut アルゴリズムはクラスタ数を指定する必要がある。そこで、キーワードの語義数と概念の集合を分類する数が等しいと仮定し、あらかじめ辞書で調査したキーワードの語義数を指定する。

２.３　用例文クラスタリング部

概念のグラフ構造をクラスタリングした結果に対して、各クラスタ内の単語を含む用例文を導き、用例文を語義毎に分類する。一つの用例文に単語が複数存在し、それぞれ違ったクラスタに含まれる場合、全てのクラスタに用例文を導くということは用例文が異なる語義に重複して導かれることを意味する。そのため、一度どこかのクラスタに導かれた用例文は、それ以降どのクラスタにも導かれないようにするなど、工夫が必要である。本稿では、用例文の重複を許可する場合と禁止する場合について、それぞれ実験を行う。

３．実験方法

検索用データは、「現代日本語書き言葉均衡コーパス」（国語研究所 2008）の DVD に収められている書籍の XML データから、日本語で書かれた部分のみを抽出し、そこから一文毎に区切り、用例文を抽出する。抽出した文に対して全文検索エンジン HyperEstraier[2] を利用して検索できるように登録を行い、データベース化する。

キーワードを入力し、用例文クラスタリングシステムを実行すると、用例文がクラスタリングされ、Web ブラウザ上に語義別用例文集合が集合毎に表示される。このとき、この用例文クラスタリングシステムについて以下の (a) ～ (h) 全ての組み合わせを考慮し、それぞれに対して用例文分類結果を出し、手動で語義別に分類した正解のデータと比較する。

- グラフ構造作成部において、抽出された名詞に対して概念への変換の有無：
  - (a) 分類語彙表の分類番号を考慮しない。
  - (b) 分類語彙表の分類番号を考慮する。
- 名詞クラスタリング部での、利用するグラフクラスタリングの種類：
  - (c) マルコフクラスタリング（mcl）
  - (d) Normalized Cut アルゴリズムによるクラスタリング（nc）
- 単語クラスタから用例文集合への変換方法：
  - (e) クラスタ内の単語をすべて用例文に変換する。用例文には複数の単語があるため、異なるクラスタでも同じ用例文に変換される場合がある。このとき、用例文が複数のクラスタに重複することを許可する。
  - (f) 用例文が複数のクラスタに重複することを禁止する。複数のクラスタに重複する用例文はクラスタ内要素の少ない順に変換する。
- 使用する用例文データ数
  - (g) コーパス内に含まれる約 9 万 6 千文のデータ(NDB と表記)
  - (h) コーパス全体からなる約 23 万文のデータ(ADB と表記)

[2] http://hyperestraier.sourceforge.net/index.ja.html

４．実験結果

キーワードを「走る」として用例文を分類したときの正解データを表 1 に、分類結果を解析したデータを表 2 に示す。「走る」の語義は小学館の『デジタル大辞泉』に記載された語義をもとに、適切と思われる数を導き出した。

表 1　単語「走る」の語義数と各データにおける出現語義数と用例文数

	NDB	ADB
「走る」の語義数	12	12
出現した語義数	7	9
有効な用例文数	45	121

表 2　用例文クラスタリング実験による分類評価結果

		NDB (g)				ADB (h)			
		mcl (c)		nc (d)		mcl (c)		nc (d)	
	重複の有無	許可(e)	禁止(f)	許可	禁止	許可	禁止	許可	禁止
概念変換無 (a)	有効用例文数	39	37	176	45	142	121	378	121
	クラスタ数	32	32	12	10	86	86	12	12
	無効クラスタ数	24	25	0	1	50	53	0	2
	正解数	7	4	72	17	61	47	114	42
	正解率	0.1795	0.1081	0.4091	0.3778	0.4296	0.3884	0.3016	0.3471
	抽出された語義数	2	2	2	3	7	6	4	5
概念変換有 (b)	有効用例文数	91	45	247	45	255	121	522	121
	クラスタ数	11	11	12	9	28	28	12	11
	無効クラスタ数	1	1	0	4	7	7	0	2
	正解数	39	25	93	16	103	59	170	40
	正解率	0.4286	0.5556	0.3765	0.3556	0.4039	0.4876	0.3257	0.3306
	抽出された語義数	4	4	2	3	6	6	2	5

４.１　評価方法

用例文クラスタリングの結果について、有効用例文数、クラスタ数、用例文数が１であるクラスタ数、正解数、正解率、抽出された語義数の６つの項目について比較を行う。ここで、有効用例文数とは、ノイズを含めない用例文の数を表す。このノイズとは、本来別の単語であるが、その一部にキーワードが含まれるために抽出された用例文を指す。クラ

スタ数はクラスタリングによって得られたクラスタの数を表す。用例文数が１のクラスタは、用例文が正しく分類できていないとみなし、無効クラスタとする。クラスタ毎に分類された用例文がどの語義で使われているのかを正解データから調べ、クラスタ内で一番多く使われていた語義の用例文数を正解数として数えて正解率を求める。正解率は、分類結果内の正解用例文を a 、分類結果内の有効用例文数を b とすると a/b で与えられる。また、分類性能を評価する際には、正解率だけでなく、抽出された語義数、無効クラスタ数なども含め、総合的に評価を行う。

５．考察

用例文クラスタリング結果における以下の各組み合わせに対して、概念に変換する場合(a)と変換しない場合（b）の違いによる分類性能を比較する。

- マルコフクラスタリング（c）と固有値によるグラフ分割（d）を用いる場合
  - (c) の場合、概念を考慮する方が正解率は高く、無効クラスタ数は減少した。また、概念を考慮しない場合は、無効クラスタの数が多くなる。つまり、マルコフクラスタリングでは名詞が細かく分割されてしまうことが分かる。概念を考慮すると、名詞がある程度まとめられているので、細かく分かれすぎてしまうという事態は解消されている。
  - (d) の場合、概念を考慮すると正解率が下がるということがわかった。グラフ内のノードは概念に変換することで少なくなってしまったため、ノード間において相違が見られなくなり、固有値によって特徴を捉えることが難しくなったと考えられる。
- 用例文が複数のクラスタに重複することを許可する場合（e）と禁止する場合（f）
  - 重複を許可する（e）場合に比べ、許可しない（f）場合は有効用例文数が減るために、正解率にも影響があった。特に、マルコフクラスタリングで概念を考慮しない場合は、正解数が有効用例文数よりも大幅に減るため正解率が下がった。逆に概念を考慮する場合は、有効用例文数が半分程度になるのに対し、正解数はそこまで減らない。つまり、データ集合内で最も多く使われる語義の用例文が多くのクラスタで重複しているために、それほど正解数は減らなかったと考えられる。

次に、データ数の異なりによる（g）と（h）の違いにおいて、正解率・抽出された語義数を対象として最も良い成果を得られた（a)～(f) の組み合わせについて考察を行う。ここで、(b, c, f) は（b）と（c）と（f）の方法を組み合わせることを意味する。

- 正解率
  - (g) : (b, c, f) が最も正解率が高い。
  - (h) : (b, c, f) が最も正解率が高い。
- 抽出された語義数
  - (g) : (b, c, e)、(b, c, f) で正解率が高く、(e)、(f) に関わらず (b, c) ならば最も多くの語義が抽出された。

◆ (h) : (a, c, f) で最も多くの語義が抽出された。ただし、この方法では同じ語義を持つ用例文が異なる小さいクラスタに分類されることが多かった。同じ語義ならば、ひとつのクラスタにまとまる方が望ましい。

以上のように、実験結果から最も精度が高い組み合わせは「グラフ構造は概念を考慮して作成し、マルコフクラスタリングを利用して、用例文が複数のクラスタに重複することを禁止する (b, c, f)」であった。これは、他の組み合わせから総合的に判断して、正解率が高く、抽出された語義数が多く、クラスタ数に対する無効クラスタ数の割合が低い組み合わせを選択した結果によるものである。

６．おわりに

本稿では、概念を考慮して作成した共起単語のグラフ構造をクラスタリングし、その結果を用いて用例文を語義毎に分類するシステムの提案を行った。また、この用例文クラスタリングの結果を分析し、用例文集合から語義別用例文集合への分類性能を調査した。その結果、概念を考慮した場合は、マルコフクラスタリングを利用することで、分類性能が概念を考慮しない場合と比較して適度なクラスタ数と適切な語義数が得られ、その上で正解率の改善が見られた。しかし、固有値を用いたクラスタリングでは、今回の実験では効果がみられなかった。従って、マルコフクラスタリングを利用してクラスタリングした方が、概念を考慮することによる効果が大きいと考えられる。固有値を利用したクラスタリングによる方法は、今回提案した手法とは異なるアプローチによって分類性能が良くなる可能性があると考えられる。

参考文献

Jean Véronis (2004). “HyperLex: lexical cartography for information retrieval “, *Computer Speech & Language,* Vol.18-3*,* pp. 223-252

Eneko Agirre, and Aitor Soroa (2007). “UBC-AS: A Graph Based Unsupervised System for Induction and Classification”, Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pp. 346-349.

独立行政法人 国立国語研究所 (2004). “分類語彙表－増補改訂版”.

村田 昇 (2005). “情報理論の基礎―情報と学習の直観的理解のために”, サイエンス社.

Stijn van Dongen (2000). “Graph Clustering by Flow Simulation”, Ph.D. thesis, University of Utrecht.

I. Dhillon, Y. Guan, and B, Kulis (2004). “Weighted Graph Cuts without Eigenvectors: A Multilevel Approach”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI*)*, Vol. 29-11, pp. 1944-1957.

独立行政法人 国立国語研究所 (2008). “「現代日本語書き言葉均衡コーパス」モニター公開データ（2008年度版）”.

関連 URL

MeCab ホームページ：http://mecab.sourceforge.net/

HyperEstraier ホームページ：http://hyperestraier.sourceforge.net/index.ja.html

ジャンル別に見るガ格を取る名詞と共起する用言の差異

野口慎一朗（作文支援システム班協力者：東京工業大学大学院社会理工学科）†
仁科喜久子（作文支援システム班班長：東京工業大学大学院社会理工学科）

A Genre-Specific Look at Differences between Nouns and Co-Occurring Predicates that Take the Nominative Case “Ga”

Shinichiro Noguchi (Graduate School of Decision Science and Technology, Tokyo Institute of Technology)
Kikuko Nishina (Graduate School of Decision Science and Technology, Tokyo Institute of Technology)

１．背景と目的

日本語学習者が文章を書く上でジャンルごとのスタイルの違いを理解することは重要であると考えられる。例えば、レポートを書く場合や、メールで事務的なやり取りあるいは研究上の連絡を取る場合など、状況にあった文章表現が必要となる。

本研究では、ジャンルによる表現のスタイルの違いに着目し、その中でも格助詞の「が」と共起する名詞と用言に焦点を当て、ジャンルごとの主語となる名詞とそれと共起する用言の差異や、文章の特徴などを考察する。これはガ格を取る名詞が文を構成する重要な要素であり、ジャンルごとの違いがよく出るのではないかと考えたからである。格助詞と共起する名詞・用言に関する先行研究はかなりあるが、ジャンルによる差異に関する研究はあまり見られない。

ジャンルごとの差異を調べることにより、日本語作文支援システムを構築する上で、書きたい文章のジャンルにあわせた語彙および表現スタイルを提示できるシステム設計に役立てることを目的としている。

２．研究の概要

本研究を行う上で、BCCWJ に収録されている白書・書籍・国会議事録の３つと新聞・青空文庫の合計５つの日本語コーパスを対象に考察を行う。コーパスを解析し、格助詞「が」と共起する名詞と用言の組を抽出し、コーパスごとに名詞・用言を分類語彙表により分類する。そして、分類ごとの出現頻度を数え、出現頻度が上位の語を調べる。それにより、コーパスごとの名詞・用言の差異や傾向の分析を行い、その結果を用い、ジャンルごとの表現スタイルの差異を明らかにする。

３．共起の抽出

共起を抽出する上で本研究では、白書は 1500 件分、書籍は 4669 作品、国会議事録は 159 回分、新聞は毎日新聞 10 年分、青空文庫は青空文庫内の 4185 作品をそれぞれデータとして用いた。BCCWJ の書籍は教養書、学術書、実用書をかなり含んでいるのに対し、青空文庫は小説を主としている。

文章の構文解析には日本語係り受け解析器 CaboCha を使用し、その解析結果から格助詞と共起する名詞、用言の組を見つけ出し、抽出した。

† noguchi.s.ad@m.titech.ac.jp

３.１　名詞・用言の定義

名詞は下記の条件を満たすものと定義した。

1. CaboCha で、「名詞」と解析されたもの
2. 複合名詞の場合、最後部位にあるもの（ただし、非自立語である接尾辞・接頭辞は他の自立語、あるいは非自立語と合わせて一語と見なす）

名詞の意味は基本的に一番後ろの名詞に含まれるので 2 の条件を付けた。

次に、用言についても下記の条件を満たすものと定義した。

1. CaboCha で、「動詞」、「形容詞」と解析されたもの
2. 「名詞-形容動詞語幹」と解析され、その後ろに助動詞の「だ」が続くもの

３.２　共起の抽出

３.１の定義に従い、CaboCha による構文解析結果から、格助詞と共起する名詞・用言の組を抽出した。

この時、用言を原型に直し、受け身や使役の接尾が付く場合はそれも合わせた形とする。これは、例えば、受け身文でガ格を取る名詞は、意味的には用言の対象格となるなど、接尾により格関係が変化するためである。

表 1 はコーパスごとの文数や共起関係の数、名詞、用言数などの情報を示したものである。この中から、本研究では抽出された組の中で格助詞「が」と共起する名詞と用言について詳細を調べた。

表 1　コーパスごとの文情報

	白書	書籍	議事録	新聞	青空文庫
文数	74,035	628,465	166,793	5,514,535	1,670,294
共起関係の数	262,773	1,063,742	308,550	10,636,188	2,840,182
名詞数	17,739	70,299	18,031	172,603	94,548
「ガ格」をとる名詞数	7,284	30,897	7,720	88,954	39,671
用言数	5,946	25,946	8,104	38,447	39,774
「ガ格」と共起する用言数	3,362	14,225	4,305	23,362	20,543

数値は異なり数

４．コーパスごとの名詞・用言の出現頻度

ガ格と共起する名詞・用言について、分類語彙表を用いて分類ごとにまとめた上でそれぞれの出現頻度を調べた。

４.１　名詞の分類別出現頻度

ガ格を取る名詞について、それぞれのコーパスの出現頻度が上位の 1,000 語に関して分類語彙表を基に分類して比較する。図１は上位 1,000 語の出現頻度における分類ごとの出現頻度の割合をグラフ化したものである。

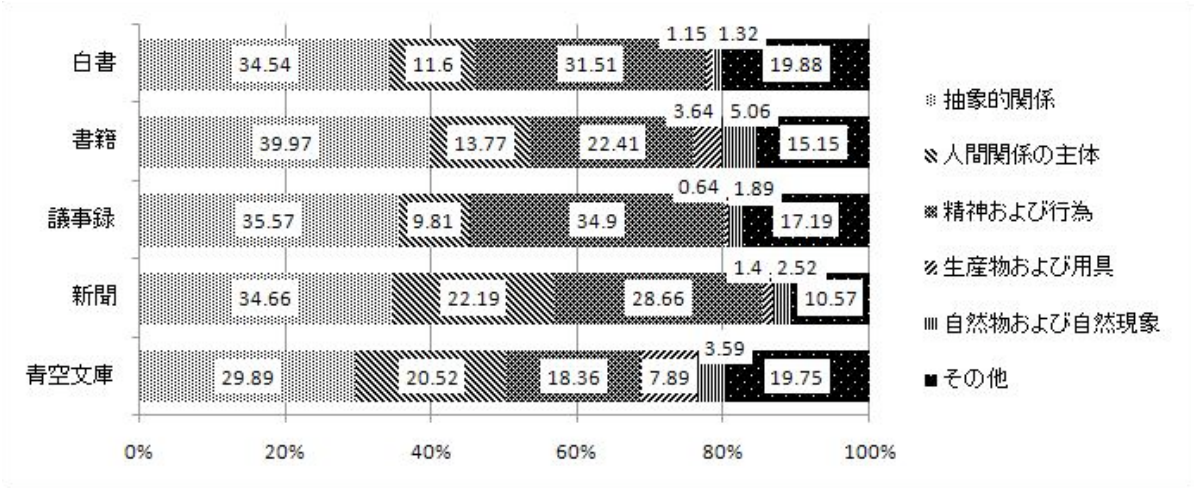


図 1　名詞の分類

図 1 を見ると「抽象的関係」、「人間活動の主体」、「精神および行為」が多くの割合を占めていることが分かる。従って、これを更に詳しく見ると、結果は表 2、表 3、表 4 の通りである。表の数値はコーパスごとの出現頻度の割合であり、括弧内の数字はコーパス内での分類の出現頻度の順位である。また、順位が 1 位のセルには網かけを、2 位のセルには斜線をかけている。

表 2　抽象的関係（名詞）

	事柄	作用	力	存在	形	時間	様相	空間	量	類	合計
白書	3.72(5)	8.77(1)	0.40(10)	1.31(8)	0.47(9)	1.84(6)	4.86(3)	1.62(7)	7.32(2)	4.23(4)	34.54
書籍	6.97(1)	1.62(8)	0.68(9)	0.43(10)	1.98(7)	6.15(3)	3.53(6)	6.08(4)	6.04(5)	6.49(2)	39.97
議事録	9.53(1)	3.16(5)	0.26(10)	0.89(8)	0.64(9)	4.34(4)	2.72(7)	4.21(3)	3.27(6)	6.55(2)	35.57
新聞	4.97(3)	3.34(5)	1.04(8)	0.61(10)	1.04(8)	2.75(6)	5.52(2)	2.15(7)	9.44(1)	3.80(4)	34.66
青空文庫	9.86(1)	0.82(8)	0.38(9)	0.25(10)	1.15(7)	2.83(4)	2.75(5)	3.21(3)	6.34(2)	2.30(6)	29.89

表 2 は「抽象的関係」について分類ごとの比率を示したものである。「抽象的関係」とは、作用、量、空間などの関係で分類したものである。この表を見ると、白書は【作用】の占める割合が他よりも高いこと、書籍と議事録は【量】が他に比べて割合が低いことが見てとれる。また、【空間】は書籍と議事録、青空文庫は他の 2 つと比べて割合が高いことも分かる。これは、書籍、議事録、青空文庫では、「ーの方がよい」、「ーの点が大事だ」という表現が多く見られ、そこに用いられる「方」や「点」など、機能語の役を担う抽象名詞の働きが大きな要因を占めているためである。

下記に記したのは、分類ごとの代表例である。

・【作用】…活動、増加、動きなど　【量】…割合、比率、格差など　【空間】…方、点、地域など

表 3　人間活動の主体（名詞）

	人物	人間	仲間	公私	家族	成員	機関	社会	合計
白書	0.87(6)	1.09(5)	0.03(8)	2.81(2)	0.28(7)	2.23(3)	3.14(1)	1.15(4)	11.6
書籍	0.57(8)	4.52(1)	0.71(7)	1.65(4)	2.01(3)	1.09(5)	0.74(6)	2.48(2)	13.77
議事録	0.85(6)	1.23(5)	0.05(8)	1.70(5)	0.08(7)	2.20(2)	2.34(1)	1.36(4)	9.81
新聞	2.42(5)	3.89(2)	0.53(8)	2.12(4)	1.74(7)	6.52(1)	3.06(3)	1.91(6)	22.19
青空文庫	1.26(4)	10.7(1)	0.92(6)	0.55(7)	3.24(2)	2.4(3)	0.24(8)	1.21(5)	20.52

表 3 は「人間活動の主体」について分類ごとの比率を示したものである。「人間活動の主体」という分類は人間や機関、社会などの関係で分類されたものです。【人間】について見ると、書籍と新聞、青空文庫で占める割合が高いことが分かる。さらに、書籍と青空文庫は【家族】が占める割合も高いが、逆に【機関】が占める割合は低いことが見てとれる。また、白書は【公私】、書籍は【社会】の割合が高いことも分かる。書籍で【社会】の割合が高いのは、教養書や学術書、実用書を含んでいるためである。

・【人間】…人称代名詞、子供など　【家族】…家、父、主人など

・【機関】…団体、機関、委員会など　【公私】…国、市町村、家庭など

・【社会】…世界、学校、店など

表 4　精神および行為（名詞）

	事業	交わり	待遇	心	生活	経済	芸術	行為	言語	合計
白書	4.35(3)	1.78(7)	1.70(8)	9.41(1)	1.92(6)	8.66(2)	0(9)	1.27(5)	2.42(4)	31.51
書籍	0.66(6)	0.42(9)	0.64(7)	11.78(1)	1.54(3)	1.25(5)	0.53(8)	1.54(3)	4.05(2)	22.41
議事録	2.20(4)	1.25(7)	2.17(5)	16.91(1)	1.00(8)	3.52(3)	0.03(9)	1.49(6)	6.33(2)	34.90
新聞	1.16(7)	1.17(6)	1.08(8)	14.16(1)	1.56(4)	2.91(3)	0.55(9)	1.30(5)	4.77(2)	28.66
青空文庫	0.02(9)	0.14(8)	0.28(7)	12.07(1)	0.92(3)	0.71(5)	0.58(6)	0.74(4)	2.90(2)	18.36

表 4 は「精神および行為」について分類ごとの比率を示したものである。これを見ると全てのコーパスで【心】の割合が高いことが見てとれる。この【心】は感覚、方法など精神的、知的活動を含むものを言う。また、白書、議事録は【事業】の割合が他の３つと比べると相対的に高く、書籍と青空文庫は【経済】の割合が他と比べて低いことが分かる。

・【事業】…企業、開発、措置など　【経済】…需要、価格、投資など

ここで、【心】について更に細かく分類して各コーパス間の違いを調べる。【心】の下位分類である【心】、【声】、【方法】、【感覚】は書籍と新聞、青空文庫では占める割合が高いが、白書、議事録では逆に低いことがわかった。また書籍と新聞は【計画・案】が高いことも明らかになった。

・【心】…心、意識、勇気など　【声】…声、歓声など　【方法】…見方、仕方、方法など

・【感覚】…気持ち、感じなど　【計画・案】…対策、計画、施策など

4.2　用言の出現頻度

用言についても名詞と同じく、コーパスごとの出現頻度上位 1,000 語を比較する。

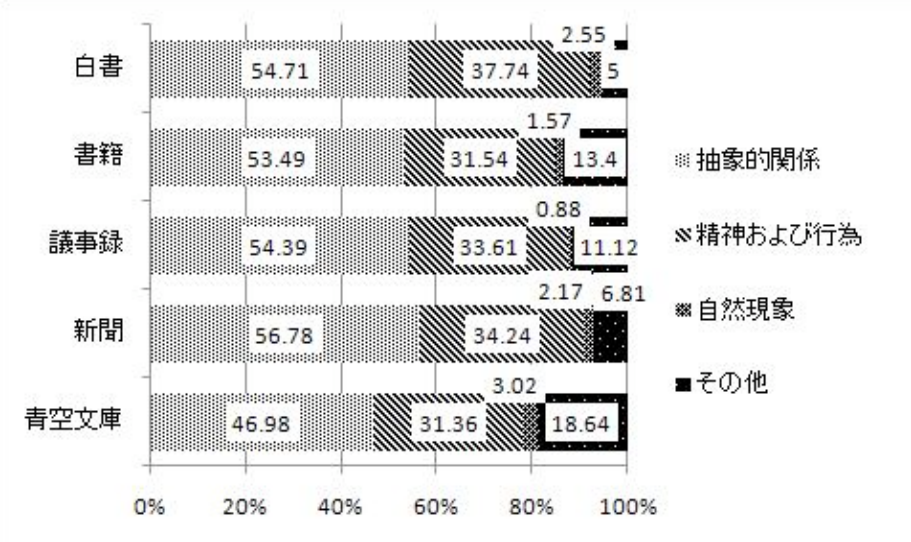


図 2　用言の分類

図 2 は用言の分類を割合で表したグラフである。図２のグラフで上位を占めている「抽象的関係」、「栄進および行為」について詳細な内訳を見る。結果は表 5, 表 6 の通りである。括弧内の数値やセルの網かけは名詞の時と同様である。

表 5　抽象的関係（用言）

	作用	力	存在	時間	様相	真偽	空間	量	類	合計
白書	24.31(1)	0.42(8)	9.18(2)	2.44(5)	1.98(6)	1.26(7)	0.18(9)	6.64(4)	8.30(3)	54.71
書籍	21.25(1)	0.26(8)	17.95(2)	1.50(4)	1.24(5)	0.11(9)	0.78(7)	0.97(6)	9.43(3)	53.49
議事録	18.01(2)	0.04(9)	21.54(1)	1.89(4)	0.51(5)	0.08(8)	0.35(6)	0.12(5)	11.85(3)	54.39
新聞	24.76(1)	1.35(7)	15.33(2)	1.69(6)	3.07(5)	0.34(8)	0.20(9)	4.22(4)	5.82(3)	56.78
青空文庫	15.04(2)	0.44(7)	20.52(1)	1.08(6)	3.13(4)	0.24(8)	0.05(9)	1.94(5)	4.52(3)	46.98

表 5 は「抽象的関係」を細かく分類したものである。「抽象的関係」とは作用、存在などの関係を表したものである。表を見ると、コーパス間で割合にはあまり差異が見られず、全体的に【作用】、【存在】、【類】の割合が高いことが分かる。

全体的に割合が高い【作用】について更に細かく分類してコーパス間の違いを調べる。【作用】の下位分類である【限定・優劣】の割合が全体的に高い。これは、「ーすることができる」という表現が多く出現するためである。また、書籍と議事録、青空文庫では【往復】、【移動・発着】の割合が高く、白書は【増減・補充】が高いことが分かった。そして、青

空文庫は【連続・反復】の割合が他と比べて低い事も見てとれた。

・【往復】…来る、帰る、行くなど　【移動・発着】…上る、着く、届くなど

・【増減・補充】…増加する、縮小する、加えるなど　【連続・反復】…受ける、続ける、重ねるなど

【存在】と【類】が全体的に高い理由は、【存在】には「ある」が、【類】には「なる」が分類され、それらの出現頻度が多いためである。「することが出来る」、「ある」は「てある」、「いる」は「ている」などとともにアスペクトを表現する機能語となることもあるので、概念をもつ内容語からは区別して考えなければならない。

表 6　精神および行為（用言）

	事業	交わり	待遇	心	生活	経済	芸術	行為	言語	合計
白書	1.29(5)	1.6(4)	0.65(7)	19.65(1)	0.64(8)	5.68(3)	0.03(9)	7.37(2)	0.83(6)	37.74
書籍	2.41(4)	1.55(6)	0.82(8)	16.72(1)	2.84(3)	1.92(5)	0.41(9)	1.24(7)	3.61(2)	31.54
議事録	0.83(7)	1.20(5)	0.71(6)	18.17(1)	1.98(4)	2.62(4)	0.05(9)	3.40(3)	4.65(2)	33.61
新聞	2.06(4)	1.75(6)	0.62(8)	21.34(1)	1.16(7)	2.51(3)	0.39(9)	1.80(5)	2.61(2)	34.24
青空文庫	2.05(2)	0.50(7)	0.24(9)	23.25(1)	1.22(4)	0.84(5)	0.41(8)	0.79(6)	2.05(2)	31.35

表 6 は「精神および行為」について分類ごとの比率を示したものである。全体的に【心】の割合が高いことが分かる。【心】とは、注意や決意など精神的活動などに関する分類を表しているものである。それ以外については、【事業】の占める割合が青空文庫で高いこと、【言語】は白書が他のコーパスに比べ低いこと、【生活】が書籍、議事録、青空文庫で高いことが見てとれる。

・【事業】…作る、建つなど　【言語】…言う、書く、話すなど　【生活】…聞く、飲む、住むなど

全体的に割合の高い【心】について、更に細かく分類して違いを調べる。【注意・認知・了解】は全て高いことが分かった。また、青空文庫は他に比べて【決心・解決・決意・迷い】の割合がかなり低く、逆に【表情・態度】、【飢渇・酔い・疲労・睡眠など】、【声】という人間の感情や状態を表す用言の割合が高い事が明らかになった。青空文庫は文学作品であり、登場人物の情緒を表現することが多いため、予想される結果と言える。

・【注意・認知・了解】…見る、聞く、認めるなど

・【決心・解決・決意・迷い】…定める、決定する、解決するなど

・【表情・態度】…笑う、泣くなど　【飢渇・酔い・疲労・睡眠など】…寝る、疲れる、覚めるなど

・【声】…叫ぶ、鳴く、歌うなど

次に、受け身を表す「れる・られる」や、「ーする」という漢語動詞についても頻度を比較し、その結果を表 7 に示す。

表 7　接尾による違い

	れる・られる	ーする
白書	27.44	30.73
書籍	3.84	5.88
議事録	7.43	10.91
新聞	8.69	13.74
青空文庫	2.66	1.67

表 7 を見ると「れる・られる」や「ーする」が使われる割合は白書が圧倒的に高く、書籍や青空文庫ではどちらもほとんど使われていないことが明らかになった。白書で「れる・られる」が多い理由は、客観的な立場をとる必要があるので人間が主語となることが少なく、「研究」や「企業」など概念的なものが主語となることが多いためであると考えられる。

５．考察

４章で得た出現頻度の結果を利用して、それぞれのコーパスの見分け方を見いだすことができる。見分け方の指標があれば、学習者がテキストを見たときに、どのジャンルの文章かを知ることができ、また、システムを利用して作文する場合もヒントを提示するためのデータになると考えられる。

まず、「人間活動の主体」の【人間】に分類される名詞の占める割合からコーパスを２つに分類する。割合が低いものが白書と議事録であり、高いものが書籍、新聞、青空文庫である。

前者は「れる・られる」や「ーする」という用言の出現頻度を調べ、両方とも出現頻度が多ければ白書、逆に少なければ議事録と判断できる。

後者については、「精神および行為」の【経済】に分類される名詞の割合から判断することができる。割合が高ければ新聞であり、低ければ書籍、青空文庫と判断できる。この二つの分類には「抽象的関係」の【量】に分類される名詞の割合を見る。高ければ青空文庫、低ければ書籍である。

また、名詞と用言それぞれについて、ピアソンの積率相関係数を用いて分類によるコーパス間の相関を求め、コーパス間での差異がどの程度あるかを見る。結果は表 8、表 9 の通りである。

表 8　名詞によるコーパス間の相関

	青空文庫	新聞	議事録	書籍
白書	0.34	0.68	0.62	0.46
書籍	0.62	0.77	0.86	
議事録	0.69	0.79		
新聞	0.75			

表 9　用言によるコーパス間の相関

	青空文庫	新聞	議事録	書籍
白書	0.81	0.94	0.83	0.88
書籍	0.93	0.96	0.98	
議事録	0.94	0.91		
新聞	0.93			

表 8 は名詞の分類によるコーパス間の相関であるが、全体的に低いことが分かる。次に、表 9 は用言の分類によるものであるが、こちらは全体的に相関が高いことが分かる。このことから、用言はコーパス間の差異が少なく判別には向かないと判断できる。

次に、名詞と用言の出現頻度の割合によりクラスタ分析を行ったところ、図 3 のような結果となった。

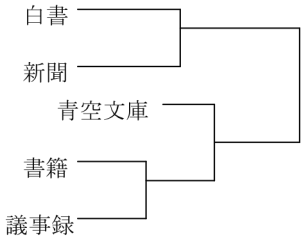


図 3　コーパスによるクラスタ分析

これにより、白書、新聞のグループと書籍、議事録、青空文庫のグループに分類されると考えられる。また、後者のグループはさらに青空文庫と書籍、議事録のグループに分類されると判断できる。

６．まとめと今後の課題

本研究ではジャンルによる表現の差異を見るために、格助詞「が」と共起する名詞、用言に焦点を当て考察を行った。コーパスごとに共起する名詞、用言に差異が明らかになり、文章内の語句の出現頻度を調べれば、ある程度コーパスを特定できることを示した。コーパスの特定ができれば、日本語学習者が文章を書くとき、文章が書きたいジャンルの表現に適しているかがわかり、適していないなら、コーパスごとの単語の出現頻度などから、どの表現がおかしいのかを判別できるシステムの研究に役立てられると考えられる。

本研究では表層的な共起関係のみを見たが、今後の課題としては、深層構造にまで踏み込んだ解析を行うことで、さらに正確な分類が出来、日本語学習者に対して、より的確な指摘を行うことで、使いやすいシステムが構築できると思われる。

参考文献

河原大輔、黒橋禎夫(2005). 「格フレーム辞書の漸次的自動構築」自然言語処理 Vol.12 No.2, pp.109-131.

小町守、飯田龍、乾健太郎、松本裕治(2006).「共起用例と出現パターンを用いた動詞性名詞の項構造解析」言語処理学会第 12 回年次大会(NLP2006)発表論文集, pp.821-824

国立国語研究所(2004),分類語彙表,大日本図書

関連 URL

日本語係り受け解析器 CaboCha, http://chasen.org/~taku/software/cabocha/

規則処理のアクセント属性を導入した CRF によるアクセント結合処理

印南 圭祐（電子化辞書班協力者：東京大学大学院新領域創成科学研究科）†
峯松 信明（電子化辞書班分担者：東京大学大学院工学系研究科）‡

Statistical Modeling of Word Accent Sandhi Based on CRF with Accent Attribute Labels Used for Rule-based Modeling

Keisuke Innami (The University of Tokyo)
Nobuaki Minematsu (The University of Tokyo)

1 はじめに

仮名漢字混じり文を入力とした日本語テキスト音声合成システムにおいて，アクセント核位置を適切に推定・付与することは，より自然な音声出力の必要条件である．日本語単語は他の語と連接して発声された場合，アクセント核位置が孤立発声時より容易に変化する (アクセント結合)．従って，アクセント結合を予測するモジュール開発が必要になる．従来，規則に基づくアクセント結合処理手法 [1] が用いられてきたが，筆者らは CR (条件付き確率場) を用いた統計的な処理手法を提案し，従来よりも高精度な処理を実現した [2][3]．また，誤り分析を通して，従来の規則処理を参考にした改善の可能性を検討した [4]．本報では，アクセント結合の規則処理で用いられるアクセント結合様式・結合アクセント価の情報を，CR 学習へ直接利用することによる性能改善について報告する．

2 CRF を用いた統計的アクセント結合処理の先行研究

筆者等はまず，JNAS[7] および ATR 音素バランス文 503 文のテキストに対し，単独ラベラにアクセント句境界，およびアクセント核の位置情報を付与させることで，アクセントデータベースを作成した [2][3]．そして利用可能な JNAS 読み上げ文 70 セット (7,280 文) を，学習用データ／評価用データとして 65 セット (6,753 文) ／ 5 セット (527 文) に分割し，CR ++を用いてアクセント核位置の推定実験を行なった．この実験は，推定対象語およびその前後に出現する語の各種語彙特性を学習素性として，文中の各形態素のアクセント核が孤立発声時の位置からどのように移動するのかを推定したものである．CR 学習に利用した素性の詳細は，[2][3] を参照して戴きたい．

実験結果は Table. 1 の通りである．精度の算出は，形態素単位とアクセント句単位で行なった．アクセント句単位の集計は，アクセント句内の全ての核位置に着目した場合と，アクセント句中の最初のアクセント核 (主核) の位置のみに着目した場合の 2 通りを行なった．また，アクセント句のうち，1 つの自立語と 1 つの付属語から構成される「単純なアクセント句」と，名詞同士の連接を伴う「名詞連続を含むアクセント句」の精度も求めた．

† innm_k@gavo.u-tokyo.ac.jp, ‡ mine@gavo.t.u-tokyo.ac.jp

表 1: 規則処理と CR　によるアクセント推定処理精度の比較 (先行研究 [2][3])

	形態素	すべての核			主核のみ		
		全ての句	単純	名詞連続	全ての句	単純	名詞連続
規則に基づく手法	—%	76.4%	94.4%	73.5%	76.8%	94.5%	74.2%
CR　による推定手法	96.5%	91.9%	97.2%	86.6%	93.5%	97.7%	87.9%

実験結果から，CR　によるアクセント推定手法を用いることで，規則処理よりも全体的に高精度なアクセント結合処理が可能であることが分かった．また，単純な句は規則・CR ともに特に誤推定が少ない構成と言える．しかし名詞連続を含む句の推定精度が低く，学習の改善が必要である．

アクセント句を構成する品詞に着目したカテゴリ分析を行なった結果 [4]，副助詞を含むアクセント句や付属語が連続で出現するアクセント句など，付属語に関連したカテゴリで誤推定が多いことが分かった．複合名詞や付属語を含む句に対するアクセント結合規則が既に整備されていることをふまえ，既存の規則の知見をより直接的に利用した学習を行なうことで改良の余地があると判断した．

3　規則処理のアクセント属性を直接利用した CRF 学習

既存の規則処理の知見を，より直接的に機械学習へ反映させるため，アクセント結合の規則処理で利用される「アクセント結合様式」および「結合アクセント価」の 2 種類のアクセント属性 (ラベル名) をデータベースに登録し，評価実験を行なった．

3.1　アクセント結合様式と結合アクセント価

匂坂らのアクセント結合規則 [1] は，アクセント結合様式と結合アクセント価の 2 種のアクセント属性を定義・利用することで，付属語／複合単語／接頭辞の各種アクセント結合の統一的な説明を可能にした．匂坂らが定義した結合規則に改良を加えたものを，Table. 2 に示す [3][5][6]．以下，規則によるアクセント結合について簡単に説明する．

3.1.1　付属語アクセント結合規則

先行する N_1 モーラ M_1 型アクセントの自立語 (名詞, 動詞, 形容詞) の後ろに N_2 モーラ $\tilde{M}_2$の結合アクセント価を持つ付属語 (助詞, 助動詞) が連接した結果，N_c モーラ M_c 型アクセントを持つ文節が構成された場合を考える．この時，後続する付属語のアクセント結合様式と，先行自立語のアクセント核の有無によってアクセント結合の変化内容が決定され，後続付属語の結合アクセント価によってアクセント核位置が決定される．(Table. 2-(a))

3.1.2　複合単語アクセント結合規則

先行する N_1 モーラ M_1 型アクセントを持つ自立語に対して，後続単語として N_2 モーラ $\tilde{M}_2$型の結合アクセント価を持つ自立語や接尾辞が連接し，N_c モーラ M_c 型アクセントを持つ複合語が構成された場合を考える．後続単語が動詞または形容詞の時は，複合語の最終モーラの直前に核がくる (M_c=N_1+N_2−1) 場合が普通である．後続単語が名詞または接尾辞

表 2: 各種アクセント結合規則
(N_1 モーラ M_1 型+N_2 モーラ$\tilde{M}_2$価 → N_c モーラ M_c 型)

(a) 付属語アクセント結合規則

結合様式	M_c	
	$M_1 = 0$	$M_1 \neq 0$
(1) 従属型	M_1	
(2@$\tilde{M}_2$) 不完全支配型	N_1+$\tilde{M}_2$	M_1
(3@$\tilde{M}_2$) 融合型	M_1	N_1+$\tilde{M}_2$
(4@$\tilde{M}_2$) 支配型	N_1+$\tilde{M}_2$	
(5) 平板化型	0	
(6@$\tilde{M}_{2a}$,$\tilde{M}_{2b}$)	N_1+$\tilde{M}_{2a}$	N_1+$\tilde{M}_{2b}$

(b) 複合名詞アクセント結合規則

結合様式	後続語の性質	$\tilde{M}_2$	M_c
(C1) 保存型	$N_2 \geqq 2$	M_2	N_1+$\tilde{M}_2$
	$M_2 \neq (N_2\text{-}1,)N_2,0$		
(C2) 生起型	$N_2 \geqq 2$	1	N_1+1
	$M_2 \neq (N_2\text{-}1,)N_2,0$		
(C3) 標準型	$N_2 \leqq 2$	0	N_1
(C4) 平板型	$N_2 \leqq 2$	*	0
(C5)	—	*	M_1
(C10)	—	*	$M_1 \mid M_2$

(c) 接頭辞アクセント結合規則

結合様式	M_c	
	$M_2 = 0,N_2$	$M_2 \neq 0,N_2$
(1) 一体化型	0	N_1+M_2
(2) 自立語結合型	N_1+1	N_1+M_2
(4) 混合型	N_1+1	N_1+M_2
	M_1	
	$M_1 \mid M_2$	
(6)	0	
(13)	M_1	
	$M_1 \mid M_2$	
(14)	M_1	N_1+M_2

の場合の処理を Table. 2-(b) に示す．複合単語アクセント結合も，付属語アクセント結合と同様に，アクセント結合様式と結合アクセント価を用いて処理体系を説明できる．ただし，処理を決定する要素の殆どを後続単語が持つ点で，付属語の処理とは異なる．

3.1.3 接頭辞アクセント結合規則

N_1 モーラ M_1 型アクセントを持つ先行する接頭辞に対して，後続単語として N_2 モーラ$\tilde{M}_2$型の結合アクセント価を持つ自立語が連接し，N_c モーラ M_c 型アクセントを持つ語が構成された場合を考える．後続単語が動詞，または形容詞の場合，前述のように単純な規則で処理可能である．後続単語が名詞の場合の処理体系を Table. 2-(c) に示す．一部のアクセント結合様式を持つ接頭辞は，構文や意味上の使い分けに対応して，複数のアクセント型を持つ場合がある．

3.2 アクセント結合様式を CRF 学習に取り入れたアクセント推定

先行研究 [2],[3],[4] の実験用データベースの作成にあたって，形態素解析辞書 UniDic[10] を利用した．UniDic では，各単語に対し，アクセント結合様式と結合アクセント価にあたる属性情報が「アクセント結合型」として登録されている．単語が自立語の場合は連続単語アクセント結合処理のためのアクセント結合様式が，接頭辞の場合は接頭辞アクセント結合処理のためのアクセント結合様式が，そして付属語の場合は付属語アクセント結合処理のた

めのアクセント結合様式が登録されている．ただし付属語の場合，各語固有の結合アクセント価が，文中発声時のアクセント核位置決定に必要な場合がある．それに該当する不完全支配型・融合型等のアクセント結合様式を取る場合，結合アクセント価が併せて登録される．また，先行自立語の品詞によって，後続付属語のアクセント結合様式，および結合アクセント価が異なる場合がある[1]．これに伴い，付属語は先行する単語が「名詞／動詞／形容詞」の場合に対応し，最大 3 種類のアクセント結合型が登録されている．これらの UniDic のアクセント結合型の登録情報を利用して，アクセント結合を規則で処理することが可能である．

従来の CR　によるアクセント推定では，各単語のアクセントに関する要素として，単独発声アクセント型しか利用していなかった．そこで，アクセント結合型の情報（即ちラベル名）を直接 CR　学習に用いた処理モデルを検討した．

なお，実験に用いたデータベースのアクセント結合型ラベルは，先行研究 [2][3] の時点における UniDic のバージョンに基づく値となっている．最新バージョンの UniDic は，アクセント結合型など規則処理に用いられる属性ラベルが更新されているが，最新のラベル情報を用いたデータベース更新・および規則処理の精度算出はまだ行なっていない，現在，その準備としてデータベースの更新作業を進めている．

3.2.1　実験手順

実験用データベースの各単語に対し，それぞれのアクセント結合型を踏まえて以下の 5 種のラベルを追加した上で，CR　によるアクセント推定の評価実験を行なった．

- 付属語アクセント結合様式 (先行語：名詞)
- 付属語アクセント結合様式 (先行語：動詞)
- 付属語アクセント結合様式 (先行語：形容詞)
- 複合単語アクセント結合様式
- 接頭辞アクセント結合様式

学習素性は，先行研究の学習素性に上記のラベルを追加したものを用いた．

先行研究 [3] では，当該形態素だけでなく，前後の形態素の語彙情報を学習素性に加えて評価実験を行ない，当該形態素とその前後の 2 形態素のラベルを学習に利用するのが最も効果的であることを確認した．これにならい，上記の 5 種類のラベルに着目する形態素の数を「当該形態素のみ／前後 1 形態素を含む 3 形態素／前後 2 形態素を含む 5 形態素」と変化させて実験を行なった．

3.2.2　実験結果

実験結果を Table. 3 に示す．前後 1・2 形態素のアクセント結合様式まで利用した場合，総合的な推定精度は僅かに向上した．前後 1 形態素を含む 3 形態素のアクセント結合様式を学習に用いることで，最も高い処理精度が得られた．また，複合名詞アクセント結合が起きる「名詞連続を含むアクセント句」に対して特に効果的な学習が行なわれていることが分かる．

[1] 匂坂らは直前の語の品詞には依存しないと述べているが [1]，アクセント属性の推定実験の結果，直前の語の品詞ごとに異なるアクセント型を持たせることの有効性が確認された [5].

表 3: アクセント結合様式・結合アクセント価を用いた CR　推定結果

		形態素	全ての核			主核のみ		
			全ての句	単純	名詞連続	全ての句	単純	名詞連続
規則手法		—%	76.4%	94.4%	73.5%	76.8%	94.5%	74.2%
従来の CR　手法		96.5%	91.9%	97.2%	86.6%	93.5%	97.7%	87.9%
CR　手法 (+結合様式)	当該のみ	96.5%	91.7%	96.8%	86.5%	93.5%	97.3%	87.9%
	前後 1 語含む	96.7%	92.2%	96.8%	87.9%	93.7%	97.3%	89.4%
	前後 2 語含む	96.5%	92.0%	97.0%	87.2%	93.6%	97.4%	88.7%
CR　手法 (+結合様式 結合アクセント価)	当該のみ	96.5%	91.9%	97.0%	86.8%	93.6%	97.4%	88.1%
	前後 1 語含む	96.6%	92.1%	96.8%	87.6%	93.7%	97.3%	89.2%
	前後 2 語含む	96.5%	92.0%	97.0%	87.2%	93.6%	97.4%	88.7%

表 4: 誤推定が多いアクセント句カテゴリの精度比較

	すべての句	副助詞	付属語連続
先行研究の素性	91.9%	84.4%	85.6%
先行研究の素性 ＋結合様式	92.2%	85.9%	85.6%
先行研究の素性 ＋結合様式組合せ	92.4%	86.7%	85.6%

前後 1 形態素までを学習に利用した場合の誤推定の減少率は，形態素単位で最大約 5.7%，アクセント句単位で最大 3.7%となった．また，前後 1 形態素まで学習に利用した場合について,「副助詞を含むアクセント句」と「付属語が連続で出現するアクセント句」のカテゴリにおける精度を求めた．その結果を Table. 4 に示す．副助詞を含む句の精度は上がったが，付属語連続を含む句の精度の向上は見られなかった．

アクセント結合様式をアクセント推定に用いることにより，従来の CR　処理モデルで特に処理精度が低いカテゴリのアクセント句に対する処理が，一部，改善できると分かった．一方，単純なアクセント句の処理精度が若干低下していることから，付属語アクセント結合の学習に関して悪影響が出ている可能性がある．

3.3　付属語の結合アクセント価を用いたアクセント推定

前述の評価実験では，付属語の結合アクセント価を機械学習に利用しなかった．そこで，アクセント結合様式に加えて結合アクセント価も利用した実験を行なった．

3.3.1　実験手順

3.2 の実験データに加えた 5 種のアクセント結合様式ラベルのうち，付属語に関連した 3 種類を，アクセント結合様式と結合アクセント価の値を組み合わせたラベルに変更した上で，3.2 の実験と同じ学習／評価データの配分，同じ学習素性を用いて評価実験を行なった．

3.3.2　実験結果

実験結果を Table. 3 に示す．アクセント結合様式のみを利用した場合の実験結果と同様，前後 1 形態素のラベルを学習素性に含めた場合に，最も効果的な学習が可能である．また，名詞連続を含むアクセント句の学習が特に改善されている．

表 5: 規則処理に関するアクセント属性の組合せ素性を用いた CR　推定結果

	形態素	すべての核			主核のみ		
		全ての句	単純	名詞連続	全ての句	単純	名詞連続
先行研究の素性	96.5%	91.9%	97.2%	86.6%	93.5%	97.7%	87.9%
従来素性＋結合様式	96.7%	92.2%	96.8%	87.9%	93.7%	97.3%	89.4%
従来素性＋結合様式の組合せ素性	96.8%	92.4%	97.1%	88.4%	94.0%	97.6%	89.8%
従来素性＋結合様式・結合価	96.6%	92.1%	96.8%	87.6%	93.7%	97.3%	89.2%
従来素性＋結合様式・結合価の組合せ素性	96.7%	92.4%	97.2%	88.7%	94.0%	97.7%	90.1%

結合アクセント価を利用した場合と利用しなかった場合では，わずかであるが利用しない場合の方が処理精度が高いという結果になった．

3.4　アクセント結合型を含めた組合せ素性の利用

前述の評価実験から，アクセント結合様式，および結合アクセント価を CR　によるアクセント推定に利用する場合，当該形態素とその前後の 1 形態素で計 3 形態素のラベルを参照するのが最も効果的だと分かった．更に規則の知見を CR　処理に反映させるため，アクセント結合型を用いた組み合わせた素性の設計を検討した．

3.4.1　アクセント結合型と単独発声アクセント型による組合せ素性

規則に基づくアクセント結合処理の結果を決定する要素として，アクセント結合型・結合アクセント型の他に，連接する語の単独発声アクセント型が挙げられる．各種アクセント結合において，連接する語の単独発声時のアクセント核位置が，アクセント結合による核位置移動の傾向に影響する．また，アクセント結合後のアクセント核位置を求めるにあたり，単独発声アクセント型への着目は重要である．そこで以下のような組合せ素性を，学習に利用することを検討した．

- { 当該形態素の単独発声アクセント型と
  直前形態素のアクセント結合様式 }
- { 当該形態素の単独発声アクセント型と
  当該形態素のアクセント結合様式 }
- { 当該形態素の単独発声アクセント型と
  直後形態素のアクセント結合様式 }

3.4.2　評価実験

上記の 3 種の組合せ素性を，3.2 でデータベースに追加した 5 種類のアクセント結合様式に関するラベルごとに作成し，先行研究 [2][3] の学習素性に計 15 の組合せ素性を追加した．そして，3.2 と同一の学習／評価データを用い，CR　によるアクセント推定の精度を求めた．

また，アクセント結合様式だけではなく，結合アクセント価も利用した場合の実験を行なった．3.2 でデータベースに追加した，付属語の結合アクセント価を含むアクセント結合

型のラベルごとに上記の組合せ素性を作成し，先行研究の学習素性に計 15 種を追加した．そして，3.2 の実験と同一の学習／評価データを用い，同様に推定精度を求めた．

3.4.3 実験結果

実験結果を Table. 5 に示す．また，副助詞を含む句，付属語が連続で出現する句の推定精度を Table. 4 に示す．実験結果から，組合せ素性の導入により全体的に推定精度が向上したことが分かった．また，組合せ素性として用いたことで，副助詞を含む句の精度は更に向上したが，付属語連続を含む句の精度の向上は見られなかった．

4 まとめ

アクセント結合の規則処理におけるアクセント属性を CR 学習に取り入れることで，アクセント推定処理の精度を向上させることができた．また，それらのアクセント属性と推定対象語の単独発声アクセント型を組み合わせることで，より効果的な学習が実現できた．その一方で，付属語連続を含む句に関しては，その精度向上を実現することが困難であることも分かった．今後は，アクセント属性ラベルの利用形式の変更や，アクセント結合様式，および結合アクセント価の組合せ素性の改良により，更なる処理精度の改善を目指す．特に，付属語が連続して出現する句に対する対策が必要である．

謝辞 本研究に全面的にご協力頂いた黒岩龍様と，度々御助言と御支援を頂いた特定領域研究「日本語コーパス」電子化辞書班の皆様に厚く感謝申し上げます．

参考文献

[1] 匂坂，佐藤，“日本語単語連鎖のアクセント規則”，電子通信学会論文誌，J66-D7，pp.849-856 (1983)

[2] 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，S 2006-174，pp.31-36 (2007)

[3] 黒岩，“日本語音声合成のためのアクセント結合規則の改善とデータベースに基づく統計的アクセント処理”，東京大学大学院情報理工学系研究科電子情報学専攻修士論文，2007.

[4] 印南，渡辺，峯松，広瀬，“CR 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)

[5] 喜多，峯松，広瀬，“日本語音声合成を目的としたアクセント結合規則の構築と改良”，電子情報通信学会音声研究会，S 2002-26，pp.13-18 (2002)

[6] 伝他，“UniDic version 1.3.9 ユーザーズマニュアル”，2008

[7] http://www.mibel.cs.tsukuba.ac.jp/jnas/

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

[9] http://hil.t.u-tokyo.ac.jp/~galatea/index-jp.html

[10] http://www.tokuteicorpus.jp/dist/

計画班研究発表

3月15日（日）　15:30〜17:30

BCCWJを利用した日本語教育語彙リスト作成の試み

▶橋本 直幸

語彙政策とコーパス ―医療用語を例に―

▶田中 牧郎

コーパス中の日本語の間違い

▶荻野 綱男

用例クラスタと辞書の語義との対応付けによる新語義の発見

▶白井 清昭、中村 誠、田中 博貴

BCCWJ を利用した日本語教育語彙リスト作成の試み

橋本直幸（日本語教育班協力者：首都大学東京オープンユニバーシティ）†

Development of a Vocabulary List for Teaching Japanese as a Foreign Language using BCCWJ

Naoyuki Hashimoto (Institute for Extended Study, Tokyo Metropolitan University)

１．はじめに

大規模コーパスの言語教育への活用法の一つとして、教育用語彙リストの作成が挙げられる。英語教育の分野では既にコーパスを利用した語彙表が多く発表され、その手法についても多くの議論がなされている。一方、日本語教育の分野では、近年、いくつかの機関・プロジェクトで独自にコーパスが作られ、日本語教育研究や第二言語習得研究の分野において活用されてきてはいるが、その中心は主に語法研究であり、教育用語彙リストの作成のように大規模コーパスを必要とする分野では、研究が立ち遅れているのが現状である。

本発表では「BCCWJ 領域内公開データ（2008 年度版）」を使用し、コーパスに基づいた日本語教育語彙リストの作成方法について考える。

２．先行研究

２．１　大規模コーパスを利用した語彙リスト（英語教育）

まず、大規模コーパスを利用した語彙リストの例として、英語教育の分野から、大学英語教育学会作成の『JACET List of 8000 Basic Words（以下、JACET 8000）』を紹介する。

『JACET 8000』は、「日本人英語学習者のための教育語彙表」として選定された 8000 語のリストで、「信頼できる大規模コーパスである BNC を基盤とし、それに日本人の教育環境を加味した「JACET 8000 サブコーパス」をからませ、全面的にコンピュータを活用しつつ、最先端の言語統計データ処理をほどこして作成したもの」（p.103）とされており、日本語教育語彙リストの作成にも大いに参考になる。作成の手順として、以下の３段階を経ている。

（1）British National Corpus（BNC）から基準データとして、「BNC 頻度順 100,000 語リスト」他を作成する。
（2）BNC での頻度と JACET 8000 サブコーパス（検定教科書、新聞・雑誌、映画、児童文学、センター試験、TOEFL・TOEIC など）での頻度を、対数尤度（log-likelihood）という視点から比較し、（1）の BNC の順位を調整し、上位 8000 語を決定する。
（3）教育的観点から、「高校教科書コーパス頻度順位表」と照合して順位を再調整する。

２．２　日本語教育における語彙リスト

現在ある日本語教育のための語彙リストのうちもっとも良く知られているものの一つとして、『日本語能力試験出題基準』（以下、『出題基準』）の語彙リストがある。これは、国際交流基金と財団法人日本国際教育協会が行っている日本語学習者のための「日本語能力

† nhashi@tmu.ac.jp

試験」の問題作成の基準となるもので、1994 年に公開、2002 年に改訂が行われている。試験は「文字・語彙」「聴解」「読解・文法」から成り、４級から１級までの４レベルに分かれている。『出題基準』では「文字」「語彙」「文法」についてリストを挙げて示しており、「語彙」については、１級レベルに必要な語彙の目安としての 10000 語のうち、8009 語を掲載している。語彙の選定にあたっては、3、4 級は既存の日本語教科書 11 種を、 1、2 級は各種語彙調査をリソースとしている。

その他、『分類語彙表』から専門家判定により教育用基本語彙 6000 語を選定した国立国語研究所『日本語教育のための基本語彙調査』(1986 年) や、専門教育出版による『品詞別・A～D レベル別 １万語語彙分類集』(1991 年、改訂 1998 年) などがある。

３．目指す語彙リストのかたち

３．１　「話題」の重要性

第２節で紹介した『出題基準』は、基本的には試験作題者のために作られたものである。レベル別にはなっているが、各レベル内は五十音順に語が配列されており、教材作成やシラバス作成には利用しにくい。では、教材作成やシラバス作成にも利用可能な語彙リストとはどのようなものだろうか。言語教育（学習）の目標が、「学習者が必要とする言語活動に必要な言語形式を教える（学習する）こと」であると考えると、各学習者が話したい内容、読みたい内容がどのような言語形式に支えられているのかが明らかになっている必要がある。そのためには、それぞれの「話題」を支える語が語彙リスト上でグルーピングされ、それぞれの話題で「まとまった話」ができるようになっていると非常に便利である。

近年、言語教育の分野において注目されている「ヨーロッパ共通参照枠（CEFR: Common European Framework of Reference for Languages)」や、全米外国語教育協会（ACTFL）の言語運用能力基準などにおいても、「話題」は能力評価の重要な指標とされており、「自分のこと」「身近な話題」から「抽象的な話題」「専門的な話題」へとレベルが上がるにつれて、扱える話題が広がっていくと考えられている。つまり、能力向上のためには、扱える「話題」を増やしていくことが重要であり、その助けとなる語彙リストが必要であると言える。

３．２　「話題別分類語彙表」の試み

橋本（2008b）では、前節の考え方に基づき、「日本語教育版分類語彙表」と題して、話題分類に沿った語彙リストの作成を試みた。具体的には、次ページ表１に挙げたように、16 の「分野」と、その下に 100 の「テーマ」を独自に設定し、『出題基準』に掲載されている約 8,000 語を筆者の主観でそれぞれのテーマに分類したものである。表１が設定した話題、表 2 が分類の例である。

「日本語教育版分類語彙表」という名前は、国立国語研究所の『分類語彙表』に倣ったものであるが、語の分類にあたって『分類語彙表』の分類枠、分類番号をそのまま採用したわけではない。『分類語彙表』(元版) の「まえがき」には次のような記述がある (pp.6-7)。

> 一つの項目に収めたのは同義類義の語の群であって、自由連想による語群ではないことである。(中略) たとえば＜ビール＞については、飲酒行動に関連して、≪酒・スタウト・ウイスキー・飲む・酔う・一杯・あわ・ジョッキ・コップ・ほろにが・ホップ・赤ら顔・ビヤホール≫等々が連想されるであろう。(中略) 連想語群をとらえるこ

表１　話題一覧（橋本（2008b））

1.文化 1.1 文化一般 1.2 食 1.3 酒 1.4 ファッション 1.5 旅行 1.6 スポーツ 1.7 建築 1.8 言葉 1.9 文芸・出版 1.10 季節・行事 2. 人生・生活 2.1 町 2.2 ふるさと 2.3 交通 2.4 日常生活 2.5 家電・機械 2.6 家事 2.7 パーティー 2.8 引越し 2.9 各種手続き 2.10 恋愛 2.11 結婚 2.12 育児 2.13 思い出 2.14 夢・目標	2.15 悩み 2.16 死 3. 人間関係 3.1 家族 3.2 友達 3.3 性格 3.4 相手への感情 3.5 容姿 3.6 人づきあい 3.7 喧嘩・トラブル 3.8 マナー・習慣 4. 教育・学問 4.1 学校(小中高) 4.2 学校(大学) 4.3 成績 4.4 習い事 4.5 試験 4.6 調査・研究 5. 芸術・趣味 5.1 音楽 5.2 絵画 5.3 工芸 5.4 写真 5.5 映画・演劇	5.6 芸道 5.7 芸術一般 5.8 趣味一般 5.9 コレクション 5.10 日曜大工 5.11 手芸 5.12 ギャンブル 5.13 遊び・ゲーム 6. 宗教・祭り 6.1 宗教 6.2 祭り 7. メディア 7.1 メディア 7.2 芸能界 8. 通信・コンピューター 8.1 通信 8.2 コンピューター 9. 経済・消費 9.1 買い物・家計 9.2 労働 9.3 就職活動 9.4 ビジネス 9.5 株	9.6 経済・財政・金融 9.7 国際経済・金融 9.8 税 10. 産業 10.1 工業一般 10.2 自動車産業 10.3 重工業 10.4 軽工業・機械工業 10.5 建設・土木 10.6 エネルギー 10.7 農林業 10.8 水産業 11. 社会 11.1 事件・事故 11.2 差別 11.3 少子高齢化 11.4 社会保障・福祉 12. 政治 12.1 政治 12.2 法律 12.3 社会運動 12.4 選挙 12.5 外交 12.6 戦争	12.7 会議 13. ヒト・生き物 13.1 人体 13.2 医療 13.3 美容・健康 13.4 動物 13.5 植物 14. 自然 14.1 気象 14.2 自然・地勢 14.3 災害 14.4 環境問題 14.5 宇宙 15. サイエンス 15.1 算数・数学 15.2 サイエンス 15.3 テクノロジー 16. 歴史 16.1 歴史 抽象的関係を表す語

表２　『日本語能力試験出題基準』をもとにした日本語教育版分類語彙表

テーマ	細目	4級	3級	2級	1級
文化一般			文化		風俗、慣習、慣行、風習
食	食事	昼御飯、朝御飯、晩御飯、夕飯		お昼、ランチ、昼食	昼飯
		お弁当			
	飲食	飲む、食べる、吸う	召し上がる、食事する、	食う、噛む、かじる、含む、しゃぶる、味わう、吐く、かぐ	なめる、飲み込む、噛み切る
	食欲			お代わり	
				食欲	
			すく	飢える、	
					ぺこぺこ、空腹
	ジャンル		西洋		洋風、和風
				和～、～風	洋～
	料理名	料理、主食、		おかず、汁、実、	
		カレー、パン、御飯	サラダ、サンドイッチ、ジャム、ステーキ、ハンバーグ	うどん、スープ、そば、刺身	ライス、粥、梅干

とも語彙論上の大切な仕事であると思われるが、ここでは、＜ビール＞をただ≪酒・ウイスキー・スタウト≫とグループをなすものとして扱い、≪飲む≫や≪ビヤホール≫との関係を断ったのである。

「話題」で分類する場合は、『分類語彙表』では扱われなかった≪酒・スタウト・ウイスキー・飲む・酔う・一杯・あわ・ジョッキ・コップ・ほろにが・ホップ・赤ら顔・ビヤホール≫を同じグループとして扱うことが望ましい。

橋本（2008b）は、8000 語を直感で分類したものであるが、次節以降では、大規模コーパスを用いた場合の話題別語彙リストの作成方法について提案する。

４．コーパスを利用した話題別語彙リスト作成の方法

４．１　話題分類のための二方法

コーパスを利用した話題別語彙リストの作成方法として、以下の二通りの方法を考えることができる。なお、ここでは作成する日本語教育語彙リストの収録総語数を仮に 20,000 語[1]として議論を行う。

Ａ：トップダウン式

大規模コーパスから出現頻度上位 20,000 語を日本語教育語彙リストの外枠とし、その１語１語について、それぞれの語がどの話題で多く使用されるかを判断し、分類する。第３節で紹介した橋本（2008b）は、この方法によるものである。

Ｂ：ボトムアップ式

コーパスのサンプル自体をまず話題別に分けておき、その話題ごとの特徴語を何らかの方法で抽出する。各話題の特徴語がすべて抽出されたところで、収録総語数 20,000 語に入らないものを削除する。また、20,000 語のうち、どの話題にも特徴語として抽出されない語は、話題別語彙とは別に収録する。これをすべての話題で行い、最後に組み合わせる。

BCCWJ の書籍サンプルには、それぞれに「NDC 分類記号（以下、NDC）」が付されているので、表１の話題と NDC を対応させれば、ボトムアップ式に話題別語彙リストを作ることができる。以下では、この方法で BCCWJ を用い、話題別語彙リスト作成を試みる。

４．２　話題特徴語の抽出

ボトムアップ式に話題別語彙リストを作成する場合、サンプルを話題で分類したのち、それぞれの話題の特徴語を出さなければいけない。本発表では、話題特徴語の抽出方法として、対数尤度比（log-likelihood ratio、LLR 値）を用いる。特徴度を出す統計指標はいくつか提案、検討されているが（内山他 2004）、対数尤度比は、テクストサイズが小さくても妥当な値を示すとされており、『JACET8000』の特徴語抽出や、コンコーダンスソフト WordSmith などのキーワード抽出に使用されている。本特定領域研究においても、近藤（2008）で教科別特徴語抽出の際に使用されている。LLR 値は以下の分割表と計算式で求められる。

	対象コーパス	参照コーパス	
単語 W	a	b	a+b
単語 W 以外	c	d	c+d
	a+c	b+d	a+b+c+d (=n)

a、b、c、d は語の出現度数

$$LLR = 2(a\log(a)+b\log(b)+c\log(c)+d\log(d) -(a+b)\log(a+b)-(a+c)\log(a+c) -(b+d)\log(b+d)-(c+d)\log(c+d) +(a+b+c+d)\log(a+b+c+d))$$

５．　語彙表の試作

５．１　話題特徴語の抽出 ―話題「旅行」を例に

ここでは、表１で提案した話題一覧の中から「旅行」という話題を例にとり、4.1 のＢ で

[1] 日本語教育用の語彙リストの収録総語数を何語にするかは別途考える必要がある。本発表で仮定する総語数 20,000 語は、本特定領域の研究成果である田中（2008）の“一般語彙リスト”の考え方に従えば、カバー率 96.60%である。日本語教育用語彙としては、一般的に上級 10,000 語とされており、それに比べると 20,000 語という数は多いが、超上級学習者までも含めた多様な学習者に対応できるよう語彙リストはできるだけ大きいものにしておく必要があると考えている。

示した方法により、話題特徴語を抽出する。手順は以下の通りである。

（1）サンプル群の選定　―話題「旅行」と NDC との対応―

「旅行」という話題に対応する NDC は【290　地理・歴史・気候】の中の【290.2 史跡・名勝. 景観】【290.6 紀行】と、【689　観光事業】であると考えられる。ただし、BCCWJ には、第 3 次区分までしか示されていない（小数点以下はない）ので、実際は、【290 地理・歴史・気候】の方は、書名を見て「旅行」に関係すると考えられるものを対象とすることになる。「旅行」に該当するサンプルは、32 サンプルであった。

（2）語彙頻度表の作成

語の対数尤度比を出すには、対象となるコーパスにおける度数と、それと比較するための参照コーパスにおける度数が必要となる。本発表では、対象コーパスを「旅行」のサンプル、参照コーパスを全書籍サンプルのうちの「旅行」以外のものとし、それぞれのコーパスで出現する度数を求め、語彙表を作成する。ここでは、UniDic-1.3.9（MeCab 版）を用い、形態素解析を行い、語彙頻度表を作成した。なお、今回の語彙リストは収録語を実質語に限定するので、語彙頻度表から機能語（助詞、助動詞）および記号類を除外した。その結果、「旅行」の延べ語数は 55088 語であった。

（3）対数尤度比の算出

（2）で作成した頻度表をもとに対数尤度比を算出する。

（4）サンプル数 1 の削除

NDC は書籍そのものにつけられた記号である。従って、BCCWJ のサンプリング箇所によっては、必ずしも NDC と内容が一致しない場合もある。よってその対処法として、たとえ LLR の値が高くても一つのサンプルにしか出てこないものは削除する。

（5）全書籍サンプルの上位 20,000 位に入らないものを削除

（4）の語彙リストのうち、全書籍サンプルにおける頻度順位が上位 20,000 位に入らないものを削除する[2]。

以上の手続きを経て「旅行」という話題における特徴語を抽出した結果が表３である（紙幅の都合で LLR 値上位 100 語のみを掲載）。上位に「陛下」「修道」「神父」「国王」など、一見すると旅行とはあまり関係ない語[3]が入ってはいるが、おおむね「旅行」に関する語が抽出されていることがわかる。

５．２　意味分類別語彙表の作成

前節で抽出した話題特徴語をさらに意味分類別に分けたものが表４である。こうすることによって、さらに教育現場で使いやすいものとなる。また、以下のようなメリットも考えられる。

（1）未収録語が明らかになる

例えば、表４の「宿泊場所」を見ると、「ホテル」「宿」が挙がっている。しかし、実際

2 全書籍サンプルの頻度順位が 20,000 位の語の度数が 63 回なので、それより少ないものを削除する。

3 これらの語はいずれも『17･18 世紀大旅行記叢書』『大航海時代叢書』において使用されているもので、削除対象にはならなかったものである。

の日本語教育現場などでは、これらに加え、「旅館」という語も重要だと考えられる。海外旅行の内容が多いBCCWJの書籍と、日本国内の旅行を内容として取り上げることが多い日本語学習者向けの教育の違いによると考えられるが、表４のように意味分類別にすることにより、「旅館」のように足りない語が何かを知ることができ、改訂への手がかりとすることができる。

(2)指導すべき重要な構文が明らかになる

表４を見ると、「移動」「宿泊」「乗降」に関する動詞やそれに対応する名詞群が多く挙がっていることに気づく。これらは「旅行」を話題とする言語活動において多用される構文であると考えられる。従って、教育現場において「旅行」という話題をタスクとして用いる際には、導入・指導すべき構文であるということが明らかになる。

６．おわりに ―今後の課題

本発表では、日本語教育における「話題」の重要性と、BCCWJを利用した話題別語彙表の作成について提案した。BCCWJの書籍サンプルにNDC分類記号が情報として付けられていることにより、従来のシソーラスのように「語彙を分類する」という発想ではなく、サンプル自体を分類し、特徴語を抽出するという新たな方法でシソーラスを作ることができる。ただし、今後、実際の現場で使える語彙表にするためには多くの課題が残っている。まず、語の難易度の問題である。表４で意味別に分類したのち、さらに難易度別に分けて提示する必要がある。もう一つは単位の問題である。本発表では短単位のままで提示しているが、実際に語彙表として発表するには、実生活で使用されている単位で語を提示する必要がある。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 ―データとしてのテクスト―』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巻3号，pp.165-197.

国際交流基金、日本国際教育支援協会（1994).『日本語能力試験 出題基準〔改訂版〕』，凡人社.

国立国語研究所（1964).『分類語彙表（国立国語研究所資料集6)』，秀英出版.

国立国語研究所（1984).『日本語教育のための基本語彙調査』，秀英出版.

近藤明日子（2008).「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（研究成果報告書）予稿集』pp.181-186.

専門教育出版（1991).『品詞別・A～Dレベル別1万語語彙分類集』（1998年改訂）.

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』.

田中牧郎（2008).「「一般語彙リスト」の設定と活用」『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）pp.37-47.

橋本直幸（2008a).「日本語教育における「話題」の扱い」『人文学報』398号， pp.58-76，首都大学東京都市教養学部人文・社会系、東京都立大学人文学部.

橋本直幸（2008b).「日本語教育版分類語彙表作成の試み」，山内博之(編)『日本語教育スタンダード試案 語彙』pp.9-91，ひつじ書房.

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巻10号，pp.50-58，明治書院.

表３　話題「旅行」の特徴語（LLR 上位 100 語）

	旅行_度数	旅行以外_度数	LLR	サンプル
ホテル	165	4136	611.56	16
観光	102	1614	464.98	13
陛下	64	938	300.85	2
修道	48	701	225.95	2
旅行	74	2880	215.00	15
バス	68	2372	210.94	10
両替	29	129	199.90	3
神父	37	403	194.38	2
ツアー	39	499	193.19	6
島々	27	195	162.38	2
客	92	6973	162.09	18
等	235	37383	161.31	21
現地	43	1206	150.42	4
国王	34	849	126.19	2
人(-じん)	197	32609	125.65	25
空港	36	1302	109.31	10
ワイン	37	1498	104.87	4
島	47	2842	100.40	5
零	411	101396	95.14	28
ジャンル	19	240	94.57	2
宿泊	24	563	91.83	6
海外	38	2149	85.50	10
予約	24	669	84.22	9
客船	13	73	84.16	3
車内	18	366	73.64	5
リゾート	17	305	73.50	5
員	61	6682	72.64	15
トイレ	28	1339	71.13	8
泊(-はく)	18	422	68.89	7
島(-とう)	31	1829	67.51	5
旅	38	2889	66.73	15
美容	17	379	66.65	3
鉄道	27	1389	65.15	6
レストラン	25	1205	63.16	11
我	38	3119	62.05	4
諸島	17	461	60.51	3
駅	41	3921	57.16	11
士	45	4798	55.37	2
パンツ	15	397	54.06	3
航海	17	574	53.72	3
場所	62	8576	53.58	17
彼	215	52120	52.75	18
社	50	6030	52.61	10
歳(-さい)	74	12333	46.53	7
乗る	48	6225	45.66	20
世界	99	19376	44.34	24
マダム	10	174	43.80	2
シティー	12	331	42.34	3
乗務	8	86	42.22	3
予算	23	1711	41.20	2

	旅行_度数	旅行以外_度数	LLR	サンプル
文化	53	7866	40.93	9
大量	25	2052	40.82	6
船	36	4140	40.32	8
コーヒー	24	1919	40.19	6
カトリック	14	568	39.62	3
土産	14	568	39.62	6
パーク	11	293	39.52	2
カフェ	12	380	39.32	3
ナンバー	15	685	39.31	2
チェック	21	1518	38.60	9
列車	20	1379	38.32	7
グッズ	9	171	37.95	2
温泉	17	981	37.66	2
聖	14	628	37.13	3
メール	19	1282	37.09	3
テーマ	22	1771	36.60	5
オペレーター	7	80	36.14	3
パソコン	20	1476	36.09	2
航空	20	1487	35.84	5
雑貨	9	197	35.59	2
到着	20	1528	34.96	10
サービス	33	3969	34.85	9
宿	19	1393	34.49	5
大陸	17	1113	34.04	5
進歩	17	1116	33.96	2
夜行	7	97	33.63	4
峰	9	244	32.04	4
販売	26	2777	31.92	8
無料	11	440	31.41	2
シャワー	12	573	30.51	4
街(-がい)	24	2513	30.17	9
我々	47	7811	29.71	11
デザイン	19	1639	29.53	3
飲み物	10	382	29.38	4
聖職	8	215	28.61	2
貨物	9	303	28.49	3
座席	12	646	28.01	5
イン	18	1551	28.00	10
入場	8	227	27.81	6
善良	7	156	27.45	3
消費	29	3778	27.40	3
船(-せん)	20	1944	27.38	3
行き	11	546	27.22	4
さん	227	65413	27.16	19
ショー	14	962	26.90	5
風景	17	1485	26.08	7
駅前	10	471	25.66	6
インテリア	7	183	25.38	2
タウン	8	279	24.81	4
得	13	903	24.74	5

（注）サンプル数1、及び固有名詞、誤解析は除く。

表４　特徴度を考慮した意味分類別語彙表　　（表３の LLR 順位 100 位以下も掲載）

項目	収録語
旅行	旅行　旅　ツアー　船旅　航海
予定	計画　行程　コース
場所	場所　現地
駅･空港など	空港　駅　港　港湾
手続き	ビザ
移動	出発する　着く　到着する 出航する　離陸する　移動する　経由する 入国する　帰国する
乗り物	鉄道　列車 電車　地下-鉄*　夜行
	バス　～車
	船　～船　客船　汽船
	航空　飛行-機*
	～行き　航路
	運行する
	定刻
乗降	乗る　乗車する　乗船する／利用する　チェック-インする*
乗り物(内)	車内　機内
手続き	予約する　チャーターする 満席　座席　空（あ）く
宿泊場所	ホテル　宿
集合	集合する
宿泊行為	宿泊する　泊まる　滞在する　チェック-インする*
	～泊
目的･イベント	観光　温泉　美容　ショッピング　ビジネス
	パーティー　ショー
行先	海外　世界　外国　国境 赤道
	温泉　リゾート　パーク　テーマパーク*　名所　聖地
	島　～島　諸島　島々　ジャングル　滝　洞穴
	街　タウン　市内　～街（がい）　エリア
	ショップ　店舗　店
お金	予算　料金　チップ　入場
	物価　外貨　通貨　両替する
	安い　無料　有料　サービス
持ち物	荷物　地図　スーツケース　パンツ　ガイド-ブック*
マナー	マナー　チップ
食事	食事　モーニング　朝食　メニュー
食べ物	チーズ　チキン
飲み物	飲み物　ドリンク
	ワイン　コーヒー　アメリカン
	飲む
施設	施設
	部屋　フロント　厨房　トイレ　シャワー　インテリア
	カフェ　レストラン　バー　厨房
	豪華　一流
土産･産物	土産　魚介　雑貨　グッズ
感想	味わう　楽しむ　謳歌する
	快適　ゆったり　快い　気軽
雰囲気	雰囲気　ムード
様子	混雑　喧騒
トラブル	トラブル
時差	時差
人	ガイド　オペレーター　（ツア）コン*　乗務-員*
	案内する
	客
風景	風景
	撮る
季節	シーズン
人気	人気　ブーム　勧め

【注】
＊は形態素解析では分けられていたが、実際には　当該形式で使われている場合が多かったので、複合語のたちで掲載したもの。

また、形態素解析において「名詞-サ変可能」となったものは、サ変動詞のかたちで掲載している。

語彙政策とコーパス　―医療用語を例に―

田中牧郎（言語政策班班長：国立国語研究所研究開発部門）†

Use of Corpus for Vocabulary Policy: A Case of Medical Words

TANAKA Makiro　(Dept. Lang. Res., National Institute for Japanese Language)

１．語彙政策に役立つデータの必要性

従来の日本の言語政策（国語政策）は，漢字政策（あるいは，仮名も含めた表記政策）が中心で，学習指導要領にも漢字について確固たる定めがある。一方，語彙については，漢字のような強力な言語政策は実施されてきていない（国語政策の歴史は，文化庁 2006 および野村 2006 を参照）。ところが近年，その状況に変化が見られるようになってきた。国語審議会や国立国語研究所が，公共的な場面で使われる分かりにくい外来語への対応を検討したことは(国語審議会 2000，国立国語研究所「外来語」委員会 2006)，語彙政策に踏み出す第一歩と位置付けられるが（野村 2006，陣内 2007，相澤 2008)，本稿では，さらに新しい二つの動きに注目したい。

第一は，専門用語の一般語化の問題である。従来は，特定の専門分野でその分野の人の間で通じ合えばよかった専門用語が，一般の人に対しても使われるようになったことで，混乱が生じている。例えば，裁判員制度によって，法廷に一般国民が加わることになり，従来は法律家だけが理解すればよかった法律用語やその意味を，一般の人にどう伝えるかが問題になっている（日本弁護士連合会 2008)。また，患者中心の医療が普及するにつれ，患者自身が医療内容を選択することが求められるようになってきているが，診療場面に出てくる医療用語が，インフォームドコンセント（説明と同意）を円滑に行う妨げになっており，改善方法が検討されている（国立国語研究所「病院の言葉」委員会 2009)。これらの問題の背景には，専門的な内容を専門家が非専門家に伝える機会が増え，また重要になってきたという社会の変化がある。この問題は，法律用語や医療用語に典型的に顕在化しているが，専門化が高度に進む一方で，国民一人一人の社会参加や自己責任が問われるようになってきた現代社会では，今後，他の分野にも問題が広がっていく可能性がある。

第二は，国民の国語力あるいはリテラシーについての議論の中で語彙力に言及されるものである。近年，児童・生徒・学生の学力低下の問題を議論する中で学力の基礎としての国語力に言及されることが増えている。また，学力低下の議論とは切り離しても，現代社会に対応した国語力の在り方を提示する試みがあったことも記憶に新しい（文化審議会 2004)。後者の審議会答申には，国語力を基礎づける「語彙力」が重視され，国語教育の文脈の中でも語彙力のあり方が議論されている。現代社会を生きる力としての語彙力という見方をすれば，社会生活に必要な各分野の情報を読み解き自分で判断していくリテラシーの基盤となる語彙力の問題は，第一の動きである専門的な情報を一般に伝えていく問題とも通じている。

これら二つの動きは，議論と試行が始まった段階であり，今後言語政策として展開されるためには，日本語の語彙の使用実態を，日本社会や社会を構成する人々と関係付けながら，正しく把握しておくことが求められよう。日本社会で生産され流通している書き言葉

† mtanaka@kokken.go.jp

を代表するものとして作られつつある「現代日本語書き言葉均衡コーパス」（BCCWJ）は，こうした要求にこたえることが期待される。

上に見たような動きに応じるためには，BCCWJ に基づいて様々なデータを用意することが望まれるが，まずは基本的なものとして，次のようなデータが求められるのではないか。

（1） 語彙のレベル分けができるデータ

一般になじみのない語彙の範囲，日常語彙の範囲，語彙の難易度などを把握するのに役立つ，語彙頻度などによるレベル分けを行ったデータ

（2） 分野特徴語が抽出できるデータ

ジャンルや媒体・教科などに特徴的な語彙の範囲を把握するのに役立つ，分野の特徴度などによって語彙を分類したデータ

これらのデータをどのような形式・内容・手順で作っていけばよいのか，作ったデータをどのように活用すれば政策展開に結びつくのかについて研究を行うことが必要である。言語政策班ではこうした作業を進めているが，本稿は，現段階での見通しを，医療用語について研究中の事例を紹介する形で述べてみたい。

２．語彙レベルの設定　―書籍コーパスの長単位語彙頻度に基づく―

田中(2008)で，コーパスの語彙頻度をもとに語彙レベルを設定する考え方と，BCCWJ2008（2008 年 7 月の領域内公開データ）のうち書籍部分の全体（以下,「書籍コーパス」と呼ぶ）を用いた短単位による実践例を示した。その考え方に従いつつ，専門用語にも対応できるように，今回実践方法を若干修正する。複合語と固有名詞が重要になる場合がある専門用語研究に使いやすいように，長単位により，固有名詞を含め，次のような手順によって，あらためて語彙レベルを設定した。まず，書籍コーパスに UniDic Ver.1.3.9(MeCab 版)を用いて，短単位による形態素解析を施し，「品詞」が，空白，記号，補助記号，未知語，付属語であるものは対象から外した。この結果をもとに，名詞や形状詞，接辞などが一定のルールで連接するものを一単位にまとめる規則を立て，この規則に合致するものは，一つの単位に認定し直した（例：起立／性／低／血圧→起立性低血圧)。

以上のような手順で作成した書籍コーパスの長単位の語彙頻度表は，異なり語数 774,233 語，延べ語数 20,060,640 語となった。この語彙頻度表をもとに，頻度の高いものから順位を付け，上位の語から順次頻度を累積していき，何位までの語で全体の語彙量のどれだけの割合を占めるか（カバー率）を計算した。そして，レベルを五段階に分けることとし，カバー率が，87%，92%，95%，97%のところで区画することにした（表１)。この区画に絶対的な根拠があるわけではないが，所属語数がピラミッド型になり，カバー率が区切りのよい数字に近い値になり，専門用語の問題が扱いやすいと考えられる線引きにした。このレベルを，国立国語研究所が行った一般国民を対象とした意識調査と対照し，評価してみたい。表２は，国民約 1400 人（語によっては約 500 人）が回答した，面接による「外来語定着度調査」（国立国語研究所 2007，http://www.kokken.go.jp/gairaigo/Yoron/index.html）を行った外来語との対照結果，表３は同じく約 1000 人が回答した，インターネットによる「非医療者に対する理解度等の調査」（国立国語研究所「病院の言葉」委員会 2009，http://www.kokken.go.jp/byoin/tyosa/rikai/）を行った医療用語との対照結果である。いず

表１　書籍コーパスによる語彙のレベル分け（長単位）

	頻度順位	頻度区間	所属語数	カバー率
レベルＡ	1-19158	68-	19,158	87.0%
レベルＢ	19159-46642	20-67	27,484	91.9%
レベルＣ	46643-97596	7-19	50,954	94.8%
レベルＤ	97597-192456	3-6	94,860	96.6%
レベルＥ	192457-774233	2-1	581,777	100.0%

れの調査でも，被験者に語を示し，その語を見聞きしたことがあるかどうか，意味が分かるかどうか（医療用語の調査では，意味を示しその意味であることを知っていたかどうか）を尋ね，見聞きがある，意味が分かる（意味を知っていた）と回答した人の比率を，認知率・理解率として算出したものである。表に示した認知率・理解率は，各レベルに属する語の平均値である。

表２ 語彙レベルと認知率・理解率（外来語）

	語数	認知率	理解率
レベルＡ	114	75.9%	58.2%
レベルＢ	82	56.9%	39.6%
レベルＣ	39	41.8%	25.6%
レベルＤ	17	25.6%	13.5%
レベルＥ	10	29.6%	16.1%
計	262	―	―

表３ 語彙レベルと認知率・理解率（医療用語）

	語数	認知率	理解率
レベルＡ	22	95.9%	79.4%
レベルＢ	24	79.1%	63.1%
レベルＣ	28	59.7%	44.9%
レベルＤ	8	59.7%	49.9%
レベルＥ	5	46.1%	34.4%
計	87	―	―

表２・表３とも，レベルＡ→Ｂ→Ｃと進むに従って，認知率・理解率いずれも低くなっていく。このことから，語彙頻度に基づくレベルは，その語がどれだけ認知され，理解されているかの度合いと相関していると考えられる（ただし個別の語については，語彙レベルと認知率・理解率とが乖離する場合もあり，その事情については考察が必要である）。レベルＤ→Ｅにおいては，そうした相関が必ずしも見られないが，これは，このレベルに属する調査対象語が少ないために，語の個別事情に左右されたのではないかと思われる。

３．医療用語の抽出　―書籍コーパスの NDC 特徴度に基づく―

語彙レベルと並ぶ基本的な情報として，分野による特徴度があげられる。近藤（2008）が述べるように，各分野における出現頻度をもとにした対数尤度比（LLR）を算出することで，語の分野特徴度を指標化することができる。この手法を医療分野の特徴度を測るのに適用し，医療分野に特徴的な語すなわち医療用語を抽出することを試みた。

書籍コーパスには，各サンプルに日本十進分類法（NDC）の分類番号が与えられている。このうち医療分野であることを示す分類は，490 医学，491 基礎医学，492 臨床医学.診断・治療，493 内科学，494 外科学，495 婦人科学.産科学，496 眼科学.耳鼻咽喉科学，497 歯科学，498 衛生学.公衆衛生.予防医学，499 薬学が相当する。書籍コーパスのすべての語（長

単位）について，これらのいずれかの番号が付与されるサンプルでの頻度と，それ以外のサンプルでの頻度を比較し，医療分野での特徴度（対数尤度比=LLR）を算出した。特徴度の高い順に並べ，上位 5669 語（LLR>14.19）を抽出し，医療用語の範囲と扱った。この基準で区切る絶対的な根拠はないが，5670 位には同じ LLR 値（LLR=14.19,医療分野の頻度２，その他の分野の頻度 0）の語が約 2000 語続くので，その直前で区切ることにした。

ところで，国立国語研究所「病院の言葉」委員会（2009）は，患者（一般国民）にとって重要でありながら分かりにくい医療用語について，57 語を例に，分かりやすく伝える工夫を提案している。この 57 語の選定母胎となった語彙リストの作成においてコーパスの語彙頻度も活用しているが，本稿で用いているコーパスや抽出方法とは別の手段によっており（田中，金，桐生，近藤 2008），最終的に 57 語を選ぶ基準は委員会の判断によっている。委員会が選んだ 57 語が，本稿で述べている書籍コーパスの NDC による特徴度によって抽出した医療用語にどの程度含まれているかを見て，本稿の抽出方法の評価をしたい。

委員会の 57 語のうち，本稿の医療用語 5669 語の中に含まれているのは 43 語（75.4%）である。含まれていない語が 14 語あるが，そのうち次の５語は，書籍コーパスでの頻度が 2 以下のものであった。医療現場ではよく用いられても，書籍には出にくい語であったり，BCCWJ が対象外とする 2006 年以後に登場した新しい語であったりするものと考えられる。

頓服（頻度２），COPD（頻度１），プライマリーケア（頻度 1），クリニカルパス（頻度 0），メタボリックシンドローム（頻度 0）

表４　医療用語として抽出されなかった語とその理由

	語	本稿の方法で抽出されなかった理由
（１）	ADL	福祉分野で高頻度　委員会は医療の周辺の語の例として選定
	グループホーム	福祉分野で高頻度　委員会は医療の周辺の語の例として選定
（２）	エビデンス	心理学分野の特定１サンプルでの頻度が極めて高い
	MRSA	文学分野の特定１サンプルでの頻度が極めて高い
（３）	ショック	医療用語とは異なる意味の日常語がある
	PET	医療用語とは同形異語の日常語がある
（４）	誤嚥	医療以外の分野でもよく使われる
	敗血症	医療以外の分野でもよく使われる
	熱中症	医療以外の分野でもよく使われる

頻度３以上で，本稿の医療用語 5669 語に入っていないのは，表４の 9 語である。なぜ本稿の方法では抽出されなかったのか，その理由も示した。(1)は，委員会があえて医療分野の外側から選んだものである。(2)(3)は，コーパスの設計やその解析方法，頻度の計算方法などの改善によって，抽出されるように改めていくことが可能であろう。そして(4)は，医療分野以外でもよく使われる医療用語があることを示しており，NDC による特徴度によって抽出する本稿の方法で，すべての医療用語を抽出することは難しいことを示唆している。しかし，全体として見れば，委員会の 57 語のうち抽出されないのは(4)の３語のみだと考えられ，本稿で試みた方法によって，重要な医療用語を含む語彙を抽出する目的は，かなりの程度まで達成できると評価してよいのではないだろうか。

次に，この方法で抽出された語彙がどんなものであるかを具体的に見るために，第１位

から 200 語刻みで 29 語を取り出して例を示すと，表５の通りである（特徴度が同じ値の場合は読み順に順位を付けた)。表のいちばん右には，２で設定した語彙レベルを示した。

表５　医療用語として抽出した語（医療特徴度上位 5669 語の例　第 1 位から 200 刻みで）

特徴度順	読み	見出し	医療頻度	医療サンプル数	医療以外頻度	医療以外サンプル数	医療特徴度（LLR）	レベル
1	カンジャ	患者	1257	158	1333	427	5408.19	A
201	カンジ	患児	40	4	1	1	274.38	B
401	ダイタイイリョウ	代替医療	34	5	25	3	162.22	B
601	レーザーチリョウ	レーザー治療	17	8	0	0	120.58	C
801	テイタイオン	低体温	15	5	2	2	94.20	C
1001	アセチルコリン	アセチルコリン	15	3	7	6	79.28	B
1201	モウハツ	毛髪	19	4	33	24	68.43	B
1401	サイボウヘキ	細胞壁	14	5	17	9	57.61	B
1601	マッサージシ	マッサージ師	10	1	5	5	52.13	C
1801	ヒロセサン	ヒロセさん	12	1	16	6	47.81	B
2001	ケイリコンピューター	経理コンピューター	6	1	0	0	42.56	D
2201	サユウ	左右	104	46	1732	1279	39.85	A
2401	インドール	インドール	5	3	0	0	35.47	D
2601	ナイフクヤク	内服薬	7	6	4	3	35.47	C
2801	オハハサン	御母さん	138	27	2749	795	31.12	A
3001	カジョウハッセイ	過剰発生	4	1	0	0	28.37	D
3201	ネンマクカソウ	粘膜下層	4	2	0	0	28.37	D
3401	ヤクサツ	扼殺	6	1	6	5	26.27	C
3601	カクビョウイン	各病院	4	3	1	1	23.43	D
3801	サイケツ	採血	9	6	34	20	21.71	B
4001	クローンハイツクリ	クローン胚作り	3	1	0	0	21.28	D
4201	セイジョウソシキ	正常組織	3	3	0	0	21.28	D
4401	ヒキツエンシャ	非喫煙者	3	3	0	0	21.28	D
4601	イデンシケンサ	遺伝子検査	4	1	2	1	20.85	D
4801	カンゲンザイ	還元剤	4	2	3	3	18.99	C
5001	ドウ	動	4	1	4	3	17.52	C
5201	ノウセイマヒシャ	脳性麻痺者	3	1	1	1	16.84	D
5401	サンバ	産婆	7	1	30	15	15.51	B
5601	レンゾクトウヨ	連続投与	3	2	2	2	14.67	D

表５の 29 語を見わたすと，医療用語と見て問題なさそうなものが多い。一方,「ヒロセ

さん」「経理コンピューター」のように，明らかに医療用語ではないものが一部に見られる。この２語はいずれも，特定の 1 サンプルだけに見られるものである。このようなものを排除するために，医療のサンプル数が１のものは除外するという細則を設けることも考えられる。ただそうすると，「過剰発生」「扼殺」「クローン胚作り」「遺伝子検査」「脳性麻痺者」「産婆」など，医療用語と考えられるものまで排除することになってしまう問題が生じる。過度な排除を避けるためには，一定頻度以上の語についてのみ適用する細則とすればよいかもしれない。また「左右」「御母さん」のように，医療のサンプル数が多いものにも医療用語とは考えにくい語があり，医療において話題になりやすい日常語も抽出されていると見られる。これらを排除したい場合は，医療分野以外の頻度が特に高いものは除外するという細則を加えることが有効だろう。この細則を加えた場合，表５では「左右」「御母さん」のほか，「患者」が排除されることになろう。

このように，特徴度によって抽出した語彙には，一部目的外のものが含まれる場合があるが，それらは別の細則を立てて適用することで，抽出されないようにすることも可能である。そうした細則の具体的設定は今後の課題であるが，その設定を行うことを前提として，今回試みた医療用語の抽出方法は実用に耐えるものだと見通すことができそうである。

４．医療用語を分かりやすく伝える工夫への語彙レベルの活用

３で抽出した 5669 語について，２で設定した語彙レベルと対照すると，レベルＡ：1417 語，レベルＢ：943 語，レベルＣ：1421 語，レベルＤ：1888 語となる（レベルＥは，極めて低頻度のため特徴度が低くなり，上位 5669 までには入らない)。

表５でレベルＡは，「患者」「左右」「御母さん」の３語であるが，これらは誰にでもなじみのある日常語である。４で述べた細則を設定して，これらを医療用語から排除することも考えられるが，レベルＡの語には，表４に示していないもの中に，「アミノ酸」「感染症」「ぜん息」「炎症」「腫れる」など，医療用語と扱いたいものも多く含まれているので，レベルＡ全体を医療用語から除外するのは適切でない。また，レベルＢの語は，「患児」「代替医療」「アセチルコリン」「毛髪」「細胞壁」「採血」「産婆」(「ヒロセさん」を除く）で，レベルＡの語に比較すると，なじみのある日常語とは言い難い語が目立っている。次にレベルＣの語を見ると，「レーザー治療」「扼殺」「還元剤」など，専門性の極めて高い語が見られるようになる。さらにレベルＤの語は，「インドール」「粘膜下層」「クローン胚作り」のような，専門的な医学用語や，「過剰発生」「非喫煙者」「連続投与」など，語形の長い複合語が目立ってくる。このように，レベルＡ→Ｂ→Ｃ→Ｄと進むにしたがって，日常語の性格が弱まって，専門語の性格が強まっていく傾向が確かめられる。

さて，「病院の言葉」委員会は，医療用語を分かりやすく伝える方法を，次ページの図１のような類型にまとめ，【分かりやすく伝える工夫】の３種５類型の各区分の代表例として 57 語を提案している。この区分と本稿の語彙レベルとを対照すると，表６のようになる。

表６から，まず，類型Ａと類型Ｂとの間で，語のレベル分布に明らかな相違が見られる。類型Ａ「日常語で」に配属される語は，レベルＢ・Ｃに分布するが，類型Ｂ「明確に説明する」の(1)(2)(3)に配属される語は，レベルＢ・Ｃだけでなく，レベルＡにも分布しているという相違である。また，類型Ｂの中で，(1)と(2)の間で，レベルＡの語の占める比率が，(1)よりも(2)で高くなっている。これらのことから，一般に知られていない語を日常語で言い換えたり，理解が不確かな語について明確な意味を伝えたりする対応方法を決める目安として，語彙レベルが役立つ面があることが示唆されよう。語彙頻度によって医療用語全

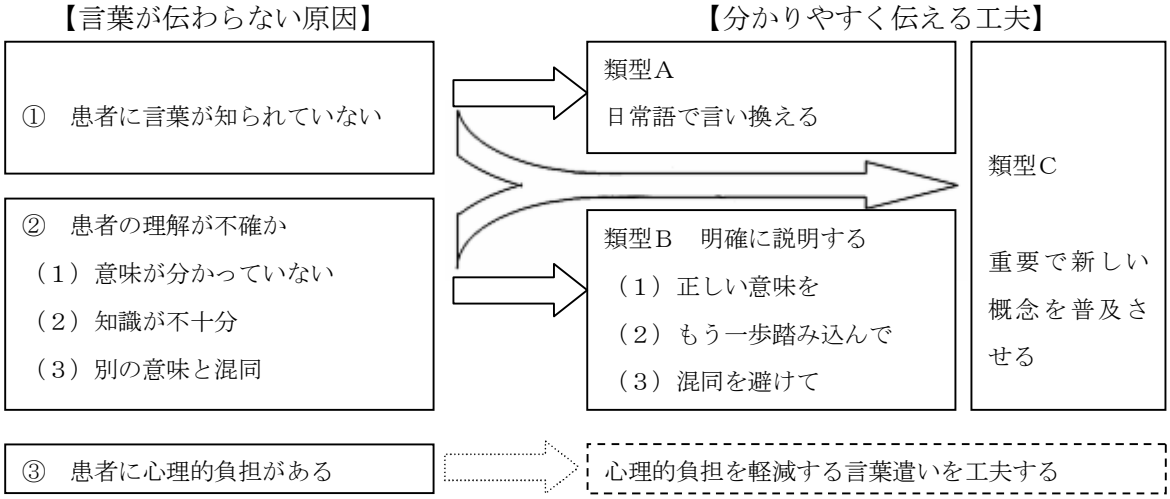


図１　「『病院の言葉』を分かりやすくする提案」における工夫の類型

表６　分かりやすく伝える工夫の類型と語彙レベルの対照

	類型 A 日常語で	類型 B(1) 正しい意味を	類型 B(2) 踏み込んで	類型 B(3) 混同を避けて	類型 C 普及を図る
レベル A	0	5　インスリン，ウイルス，炎症，潰瘍，腫瘍	8　うつ病，化学療法，抗体，ぜん息，糖尿病，動脈硬化，脳死，副作用	1　貧血	1　ガイドライン
レベル B	4　重篤，浸潤，耐性，予後	4　腫瘍マーカー，腎不全，ステロイド，対症療法	2　悪性腫瘍，肝硬変	1　合併症	4　インフォームドコンセント，セカンドオピニオン，ＱＯＬ，ＭＲＩ
レベル C	4　イレウス，寛解，生検，せん妄	2　介護老人保健施設，膠原病	5　うっ血，黄だん，尊厳死，治験，ポリープ		1　緩和ケア
レベル D	0	0	1　既往歴	0	0
抽出されない語	5　エビデンス，誤嚥，ＡＤＬ，ＣＯＰＤ，ＭＲＳＡ	4　グループホーム，頓服，敗血症，メタボリックシンドローム	1　熱中症	1　ショック	3　クリニカルパス，プライマリーケア，ＰＥＴ

＊　表中の数字は，所属する語の数

般に適切なレベル分けを行うことができれば，用語やその意味あるいは知識を伝える方法を工夫する施策にとって有意義なデータとなるだろう。言い換えたり説明したりするときに一般人向けに使うと効果的な語を選ぶ際の目安にもなろう。(金, 桐生, 近藤, 田中 2008)，「病院の言葉」委員会が提示した基本的枠組と典型例をもとに，多様な医療用語・医療場面に応用していくための参考情報として，レベル情報は有益であると予想される。 今後，

レベル設定の方法をさらに吟味するとともに，レベルを活用した工夫の実践について研究を重ねることが望まれる。

５．今後に向けて

「病院の言葉」委員会が取り組んだ医療用語を分かりやすくする検討は，難解な専門用語をどう平易に言い換えるか（類型Ａ），重要な医学知識や新しい医療理念を患者（一般国民）にどう理解してもらい（類型Ｂ），どう普及させるか（類型Ｃ），という三つの類型を導き出した。類型Ｂや類型Ｃは，冒頭に述べた語彙政策の二つの動きのうち，国民のリテラシーをどう引き上げるかという観点での語彙政策の動きに通じるものである。

言語政策班では，小中高等学校の全学年全教科の検定教科書の語彙を把握できる語彙表の作成も進めている。現代の日本社会で使用されている語彙の実態と，学校教育における語彙の実態を突き合わせる作業を行って，現代社会で必要なリテラシーという観点から語彙政策・語彙教育のあり方を議論できる基盤となるデータを整備することも，重要である。

文献

相澤正夫（2008）「『福祉言語学』事始」日本語科学 23，pp.111-123

金愛蘭，桐生りか，近藤明日子，田中牧郎（2008）「『一般向け専門用語』抽出の試み―医療用語を例に―」日本語学会 2008 年度春季大会予稿集，pp.199-206

国語審議会（2000）「国際社会に対応する日本語の在り方」国語審議会答申

国立国語研究所（2007）『公共媒体の外来語―「外来語」言い換え提案を支える調査研究―』国立国語研究所報告 126　http://www.kokken.go.jp/gairaigo/Report126/report126.html

国立国語研究所「外来語」委員会（2006）『分かりやすく伝える 外来語言い換え手引き』ぎょうせい　http://www.kokken.go.jp/gairaigo/

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　http://www.kokken.go.jp/byoin/

近藤明日子（2008）「特徴度の設定」田中，相澤，斎藤，他(2008)，pp.13-16

陣内正敬（2007）『外来語の社会言語学―日本語のグローカルな考え方―』世界思想社

田中牧郎（2008）「語彙レベルの設定」田中，相澤，斎藤，他(2008)，pp.7-12

田中牧郎，相澤正夫，斎藤達哉，他（2008）『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，金愛蘭，桐生りか，近藤明日子（2008）「コーパスによる難解語・重要語の抽出―医療用語を例に―」社会言語科学会第 21 回大会発表論文集，pp.296-299

日本弁護士連合会（2008）『裁判員時代の法廷用語―法廷用語の日常語化に関する PT 最終報告書』三省堂

野村敏夫（2006）『国語政策の戦後史』大修館書店

文化審議会（2004）「これからの時代に求められる国語力について」文化審議会答申

文化庁（2006）『国語施策百年史』ぎょうせい

コーパス中の日本語の間違い

荻野綱男（辞書編集班班長：日本大学文理学部）†

Errors in Japanese Corpora

Tsunao Ogino　(College of Humanities and Sciences, Nihon University)

１．コーパスと辞書の違い

コーパスは、実際の使用例を反映するから、ほぼ現実といってもよい。

一方、辞書は、編集者の判断が入るから、規範性も出てくるものである。単に現実を反映するだけの辞書が有用とはいえないだろう。

WWW を大規模コーパスとして利用できるようになり、日本語研究に大変有用であり、重宝している。しかし、利用していると、いくつか問題点を感じることがある。その中の一つが「間違いが多い」ということである。他にも、アダルト用語が多いという点も問題になろう。WWW には、パソコン用語やゲーム用語なども多いが、まあそれはさほど問題ではないようにも思う。しかし、「間違いが多い」ということは、特に辞書作成のプロセスを考えると、できあがった辞書の信頼性にも関わる大きな問題である。

そこで、本稿では、WWW に見られる間違いに焦点を当てて、どんなものが見られるのかを調べていきたい。

また、BCCWJ についても、2008 年 7 月にモニター版が公開され、科研費特定領域研究の関係者は、さらに大きなデータが使えるようになってきている。そこで、WWW で見られる「間違い」が BCCWJ でも見られるのかを確認していくことにする。

２．間違いの例

WWW での検索結果は数十万件もあることがあり、その全部を調べることはできない。そこで、先頭 100 件（場合によって＋α）を具体的に調べ、どんなふうに使われているかを見ていく。

検索エンジンは Yahoo! で行う。

マイナス検索を多用しているのは、参照例（使用例ではなく、その表現について議論したりコメントしたりしているもの）をのぞくためである。

"" は、フレーズ検索（Google の用語）で、指定されたものが、ちょうどその順番で並んで使われているものだけを検索することを意味している。

用例は、以下の１１通りに区分することにする。

(1)別意味　(2)参照例　(3)個人ホームページ　(4)日記、ブログ　(5)掲示板・ＢＢＳ
(6)企業・法人・団体のホームページ　(7)政府や地方公共団体などの公的なページ
(8)辞書　(9)ニュース　(10)その他　(11)表示できません

このうち、「(1)別意味」は、その文字列がまったく別の意味で使われていることを表し、本来、用例ではないものである。(3)から(10)は、どんなサイトで使われるかを区分して示したものである。「(11)表示できません」は、検索エンジンの都合で、以前は検索できたのに、サイトが削除されたりして現在はなくなっているものである。

2008-07 バージョンの BCCWJ の DVD で「ひまわり」を使って検索した結果も一緒に表示する。

† ogino@chs.nihon-u.ac.jp

２．１　漢字の読み方

２．１．１　「雰囲気」を「ふいんき」と読む

		別意味	参照例	個人ＨＰ	日記・ブログ	掲示板	企業団体ＨＰ	政府自治体ＨＰ	辞書	ニュース	その他	表示不可
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″ふんいき″-″ふいんき″」	912,000件	5	4	8	23	14	32	5	7	2	0	0
「″フンイキ″-″フインキ″」	421,000件	1	0	20	27	18	31	0	0	3	0	0
「″雰囲気″ ″ふんいき″-″ふいんき″」	156,000件	0	2	13	15	7	28	17	13	0	7	0
「″雰囲気″ ″フンイキ″-″フインキ″」	243,000件	2	1	28	28	8	28	4	0	1	0	0
「″ふいんき″-″ふんいき″」	468,000件	13	43	7	16	14	6	0	0	1	0	0
「″フインキ″-″フンイキ″」	138,000件	0	8	8	38	23	22	0	1	0	0	0
「″雰囲気″ ″ふいんき″-″ふんいき″」	219,000件	5	413	3	45	28	4	0	0	0	7	0
「″雰囲気″ ″フインキ″-″フンイキ″」	43,500件	2	56	2	17	17	5	1	0	0	0	0

「雰囲気」を「ふいんき」と読む誤用はかなり広まっていると考えていた。

検索結果は、上の４回が正しい「ふんいき」で、下の４回が間違った「ふいんき」である。

最初の４回の検索でわかるように、正しい読み方は(3)個人ホームページ、(4)日記、ブログ、(6)企業・法人・団体のホームページなどに多い。これらのヒット数の多少は文書量に比例しているのであろう。

一方、下の４回が間違った「ふいんき」の検索結果であるが、検索件数が正用とあまり変わらないくらいに多い点が気になる。誤用がそれだけ一般化しているようにも考えられる。しかし、使用状況を分類してみると、「ふいんき」は(2)参照例が非常に多いことがわかる。つまり、こういう言い方を書き込んでいる人は、誤用を誤用とわかって議論しているのであって、間違えて覚えて実際に使っているのとは状況が異なる。したがって、上の検索結果に見られるほどの「ふいんき」の多用があるわけではない。

用例を見ると、「ふいんき（←なぜか変換できない）」という一種の冗談のような使い方が多数存在する。

「"ふいんき（←なぜか変換できない）"と書いてある場合はネタか釣りである場合が多い。」と書いているサイトもある。(http://d.hatena.ne.jp/keyword/%A4%D5%A4%A4%A4%F3%A4%AD)

これは、WWWの掲示板などで流行しているフレーズのようであり、こういった日本語の誤用を「わざと」間違えて楽しむ人達もいるようだ。わざと間違えている場合と、意識せずに間違えている場合を検索結果から区別することはきわめて困難である。ここでは、こういうものも用例として扱った。

この言葉をはじめに使い始めたのはいつだかはまだわかっていないようだが(http://q.hatena.ne.jp/1124114950)、はじめに使い始めた人は「雰囲気」を「ふいんき」と読む人への皮肉として使ったといわれている。しかし現在では本当に皮肉として使っているのか曖昧な場合が多い。例えば日記のタイトルであったとすると、そのことについて説明がない場合、その人が本当に理解してその「ネタ」を使っているのかどうかがわからない。

間違った言い方の「ふいんき」は、(4)日記・ブログや(5)掲示板で多く使われ、(7)政府や地方公共団体などの公的なページ、(8)辞書、(9)ニュースではごくわずかしか使われない。(6)企業・法人・

団体のホームページでも若干使われる傾向があり、(6)は、個人サイトと公的サイトの中間的な存在であることがわかる。

このように、間違った日本語の用例は個人のページに多く、公的なページには少ない場合が多い。（そうでない場合もあるが、それについては後述する。）

BCCWJ 2008.7 には、「ふいんき」が８件見つかる。以下に示す。

1. ｙａｈｏｏ削除隊が「 ふいんき 」が変換出来ないと言う同じような質問を削除しないのは何故でしょう　Yahoo!知恵袋
2. 言葉も生きてますからね…。「 ふいんき 」が言いやすいとは別に思わないけど、そういう例は多いですよ。　Yahoo!知恵袋
3. 「雰囲気」って何て読みますか？正しくは「ふんいき」ですが、「 ふいんき 」って言う人が周りに何人かいます。　Yahoo!知恵袋
4. 言いやすいと思う人が多いなら、今後読みが「 ふいんき 」に変わっていく可能性もありますかね？言葉も生きてますからね　Yahoo!知恵袋
5. 薔薇という漢字の書き準を教えてください。「 ふいんき 」に飽きたのかな？　「書き準」は、知らないな。「書き順」なら　Yahoo!知恵袋
6. ネットを始めて初めて知ったという人も多く、その方が言い易い「 ふいんき 」を使いまくって定着させたいと思ってます？？　Yahoo!知恵袋
7. 「ふいんき」を使いまくって定着させたいと思ってます？？ いや、ふいんき言いずらいって。てか、 ふいんき ネタつきないねえ。　Yahoo!知恵袋

このように、全部参照例であり、誤用はない。

２．１．２　「缶詰」を「かんづめ」でなく「かんずめ」

		別意味	参照例	個人ＨＰ	日記・ブログ	掲示板	企業団体ＨＰ	政府自治体ＨＰ	辞書	ニュース	その他	表示不可
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″かんづめ″−″かんずめ″」	853,000 件	75	0	2	12	0	9	1	1	0	0	0
「″缶詰″ ″かんづめ″ −″かんずめ″」	65,800 件	24	0	6	16	3	42	4	5	0	0	0
「″かんずめ″ −″かんづめ″」	29,700 件	24	1	21	38	5	10	1	0	0	0	0
「″缶詰″ ″かんずめ″ −″かんづめ″」	474 件	5	6	17	42	11	16	3	0	0	0	0

「かんづめ」にはHPの名前やハンドルネームとして使用されているものが多かった。

BCCWJ では、１件の誤用が見つかる。

1. 根のすりおろしたものをかけ醤油サバ飯 - - - ご飯にサバの水煮の かんずめ（水は捨てる）を載せ、マヨネーズと醤油ゆで卵１０個　Yahoo!知恵袋

「Yahoo!知恵袋」の例であり、BCCWJ の中では、「Yahoo!知恵袋」がかなりくだけた文体を代表しており、誤用が現れやすい傾向がある。

２．２　漢字の使い方

２．２．１　「完璧」を「完壁」（※下が「玉」ではなく「土」）

		別意味	参照例	個人ＨＰ	日記・ブログ	掲示板	企業団体ＨＰ	政府自治体ＨＰ	辞書	ニュース	その他	表示不可
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″完璧″ -″完壁″」	78,100,000 件	4	1	11	4	15	45	0	4	10	3	3
「″完壁″ -″完璧″」	220,000 件	1	2	25	10	5	41	3	0	9	4	0

日本語入力システムに「かんぺき」と打ち込んで変換しても、「完壁」とは出ない上、場合によっては「『完璧』の誤用である」と指摘するシステムもあるので「完壁」を使用する場合として、基本的には本人が完全に「完壁」として思い込み入力している場合か、「かんへき」と打ち間違えて入力した場合が考えられる。

日記やブログ、掲示板に使用されている例は多くなく、結果は個人サイトや、企業・団体サイトで使用されている例が多かった。

参照例の中には『「完璧」は「完壁」と間違えやすい』というような失敗談や注意などの文章や、「完壁」という（「完璧」をもじったのであろう）商品名なども含まれている。しかし、参照例の比率が非常に低いということは、この間違いに気が付いている人が少なく、人々の話題にものぼりにくいということを意味している。実際にはたくさんの誤用があるのかもしれないが、文章を読んでいくときには正用の「完璧」と読み間違えてしまい、問題とは意識されないのであろう。

BCCWJ では、１例だけ見つかった。

1. 一つの町で何十億という農作物被害が出ておる。現在、調査がまだ 完璧 に進んでおりませんけれども、岩手県においては被害総額が百四十　　国会会議録

国会会議録の間違いというのは意外だった。

２．２．２　「年俸(ねんぽう)」を「年棒(ねんぼう)」

		別意味	参照例	個人ＨＰ	日記・ブログ	掲示板	企業団体ＨＰ	政府自治体ＨＰ	辞書	ニュース	その他	表示不可
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″年俸″ -″年棒″」	9,840,000 件	0	0	21	10	2	21	13	4	20	9	0
「″年棒″ -″年俸″」	1,030,000 件	2	0	37	25	2	27	1	0	7	1	0

誤字「年棒(ねんぼう)」の企業団体ＨＰを見ると、求人情報が多くなっている。そういう文脈で使われることばといえそうだ。

誤字の参照例がないのは、まだ人々が気が付いていない証拠である。「完璧」と同様の事情がありそうだ。

BCCWJ では、「年棒」が知恵袋 8 件/書籍 1 件ある。

1. 清原っていつから大物格闘家になったの？ 年棒 が大幅に下がって、球団事務所を出て、真っ赤な顔をして、　　Yahoo!知恵袋
2.（２例） のある選手は解雇されませんよ。他の球団が取りますよ。だいたい 年棒 が高すぎる。年棒の割りに客を呼べないのだから話にならない。　　Yahoo!知恵袋
3. 下衆な質問ですが、Ｊ１名もないレギュラーの 年棒 って平均いくらくらいなんでしょうか。Yahoo!知恵袋
4. プロ野球の審判って給料いくら貰ってんですか？ 年棒 なんですか？ 安いと言うのが定説になっていますが、　　Yahoo!知恵袋
5. 下位のほうはＪ２の中位ぐらいになります。ただ 年棒 に関しては、一流選手でも３０００万円行かないので、　　Yahoo!知恵袋
6. 確かに単純に数字だけ見ると、 年棒 の割には・・・と思うかもしれませんが、中日にとってはYahoo!知恵袋
7.サッカーの 年棒 制は、基本給＋出場給です。野球も同様のことをすればいいですね　　Yahoo!知恵袋
8. 上の説明は不要だろう。イチローの去った日本球界では最も高額な 年棒 を稼ぐ、また最も有名な選手でもあるからだ。松井 成田　好三 1950 男 『球場に数秒間の沈黙を』

「Yahoo!知恵袋」に多く現れるということは、このサイトが個人の投稿を中心に成立しているという特徴があるためであろう。誤用をたくさん含んでいることは、やや文体的に低いという解釈も可能である。

書籍で１件誤用が見つかった。単なる校正ミスと思いたいところだ。

２．２．３　「講義」を「講議」

	別意味	参照例	個人HP	日記・ブログ	掲示板	企業団体HP	政府自治体HP	辞書	ニュース	その他	表示不可
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″講義″ −″講議″」59,200,000 件	0	0	12	3	2	80	2	0	0	1	0
「″講議″ −″講義″」 570,000 件	0	0	13	7	0	79	0	0	1	0	0

これも正々堂々と間違えている。「講議」は、比率からいえばわずかだが、実数で見るとかなり大量に見つかる。大学の公式サイトや、病院、日本文学の教師のサイト、民主党の政治スクールのサイトなどでさえ誤用が見られる。ちなみに「講議　site:go.jp」での検索結果が 344 件あった。公的なところでも間違いが広がりつつあるといったところだろう。

BCCWJ では、「講議」が 4 件見つかった。書籍 3/国会会議録 1 である。

1. その講座の中での精神科医による 講議 によると、以前は精神障害者をなんとか矯正して、できるだけ健常　　平成　治世『何んだ可んだ』
2. 主幹の倉本長治先生との出会いがありました。 講議 のポイントは、戦後、悪徳商人が横行し、

商品の流通に携わる者が　　西端　春枝『経営の王道 リーダー、トップは「証券営業の神様」豊田善一に学べ！』
3. 二階のいちごちゃんの部屋（それは三畳一間だけど）で、 講議 をうけた。「では、そのまえに、ミィの最新刊で、　　上條　さなえ『スーパー・ガールいちごちゃん』
4. 非常に鮮明なテレビで画像を見たんですが、同じ大学で 講議 をしておっても、幽霊のようにゆらゆらゆらゆらと揺れたり、　　国会会議録

お堅いところでも見つけられる誤用ということになる。

２．３　複合語

２．３．１　「うろ覚え」を「うる覚え」

	別意味	参照例	個人ＨＰ	日記・ブログ	掲示板	企業団体ＨＰ	政府自治体ＨＰ	辞書	ニュース	その他	表示不可
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″うろ覚え″ -″うる覚え″」6,500,000 件	3	0	21	42	9	3	0	4	3	15	0
「″うる覚え″ -″うろ覚え″」 490,000 件	0	3	9	50	21	6	0	0	0	9	2

日記・ブログに「うる覚え」の誤用が多い。こんな内容を書きそうなのは日記・ブログの類だということで結果は納得できよう。
BCCWJ では１件だけ見つかるが、Yahoo!知恵袋 の例である。

1. ちなみにそれは７～８００ｍｌで５００円位だったと思う。 うる覚え でごめんなさいね。↓一応、参考までにレシピを。　　Yahoo!知恵袋

２．３．２　「人間ドック」を「人間ドッグ」

	別意味	参照例	個人ＨＰ	日記・ブログ	掲示板	企業団体ＨＰ	政府自治体ＨＰ	辞書	ニュース	その他	表示不可
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″人間ドック″-″人間ドッグ″」9,930,000 件	0	0	1	0	0	98	1	0	0	0	0
「″人間ドッグ″-″人間ドック″」 778,000 件	4	2	4	31	2	56	3	0	0	2	0

「人間ドッグ」は企業・団体の記事でたくさん見つかるが、その性格上、病院でたくさん使われる誤用である。ブログでもすっかり皆間違えている。
BCCWJ では、２件（知恵袋 1/書籍 1）見つかる。

1. 早期発見、早期治療に力を注ぐことではなかろうか。健康診断、 人間ドッグ の充実、禁煙対策等の強化は欠かせない。　　橋本　巌『医療費の審査 知られざるその現実』
2. やっぱり 人間ドッグ （定期検査）って必要ですか？　　Yahoo!知恵袋

この場合の例 1. の書籍は、かなり専門的な内容が扱われており、そういうところにも誤用があるというのは意外である。

２．３．３　「手持ちぶさた」を「手持ちぶたさ」

	別意味	参照例	個人HP	日記・ブログ	掲示板	企業団体HP	政府自治体HP	辞書	ニュース	その他	表示不可
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
「″手持ち無沙汰″ −″手持ちぶたさ″」629,000 件	1	2	16	46	10	8	0	2	7	4	4
「″手持ちぶさた″ −″手持ちぶたさ″」194,000 件	0	0	21	31	4	26	0	3	12	1	2
「″手持ちぶたさ″ −″手持ちぶさた″」 76,800 件	1	0	15	56	3	19	0	0	2	1	0

３つ目の指定文字列は、正しくは「"手持ちぶたさ" -"手持ちぶさた" -"手持ち無沙汰"」であった。

実際に検索されたサイトを見てみると、個人の日記やサイトのような私的なものから楽天市場（http://www.rakuten.co.jp/shopfield/460332/600892/）や禁煙用特殊タバコ通販の企業サイト（http://www.bidders.co.jp/sitem/20594202）のようなオフィシャルなものまで実に広く間違われている。また NHK 公式サイト内 NHK ボランティアネット：防災もの知りノート（http://www.nhk.or.jp/nhkvnet/bousai/vol/kokoro_set.html）でも、「手持ちぶたさや暇を気にせず〜」といった誤用が見られた。

このような誤用が広まった原因として、「誤字等の館」では「この言葉を形容詞の名詞形として使いたいという思いにあるのではないか？」という仮説を挙げている。（http://www.tt.rim.or.jp/~rudyard/kaego004.html）

この誤用についても、参照例が少なく、人々の話題にのぼらないが、誤用が浸透しつつあることを示している。

BCCWJ では、Yahoo!知恵袋で１件、書籍で１件見つかる。

1. ついついたばこをいつも以上に吸っていたのですが禁煙中なので 手持ちぶたさ です。何をしていたら良いのでしょうか？　　Yahoo!知恵袋

2. 会談に出席していた上村康広町議（その後議長）が、 手持ちぶたさ に、残っていた県職員の席に近づいて話しかけた際、　　北村　博司『原発を止めた町 三重・芦浜原発三十七年の闘い』

３．　間違いと辞書記述

以上、さまざまな例を見てきたが、現実に、誤用がたくさんあることがわかった。

単純に検索エンジンで数えただけでは、参照例などが含まれてしまうので、現実を正しく把握することはできない。今回のように先頭 100 例を人手で見ていくようなチェックが必要であろう。

使用状況を見ることで、個人が書く文章に間違いがたくさんあることもわかった。

なぜWWWには（特に個人サイトで）間違いが多いのだろうか。

第１に、校正をきちんとしていないことが考えられる。ブログなどは手軽な発信手段として発達してきた。正確な文章を書くよりは、さっさとタイムリーな話題について書くことが普通である。従来の出版であれば、編集者や校正者の目を通ってから世の中に出たのだが、WWWではそうでは

ない。

第２に、書き手に若い人が多いことが影響していることが考えられる。特にブログや掲示板では、書き手の年齢は若い傾向がある。goo のブログ検索で年代別の集計ができるが、そういう結果から見ても、利用者が若い人に傾いていることは明らかである。WWWへの親近感や使用状況において、若い人が中心になっているのが現実のようである。若い人は、各種経験が少なく、どうしても間違いが多くなりがちなのではないだろうか。

第３に、このような間違いは、実は間違いではなく、新しい変化なのかもしれない。新しい変化は若い人が受け入れやすい。

第４に、知識が不足する人でも文章を公開する傾向が強まったということもあるかもしれない。第１の指摘とも関係するが、「手軽に書ける」ことは「誰でも書ける」ことにつながり、つまりは低レベルの書き手が相対的に増えているということがあるかもしれない。

さて、辞書の作り方をどうするべきか考えてみよう。

第１に、コーパス中の何を間違いとして排除するべきかという難問がある。上でも述べたが、言語使用を見ていく上では「誤用」は存在しないという立場もある。すべては言語使用の現実である。しかし、そのようにして単に現実を受け入れるだけでは、辞書として望ましいものができあがるとは言えない。誤用と正用の境界線の引き方は現実的に非常に難しい。

第２に、コーパスから単純に辞書は作れないということがある。コーパスは言語使用の現実を反映する。辞書は規範が求められる（面がある）。この二つから出てくる当然の帰結はコーパスを単純に加工して辞書にしてはいけないということである。無論、自然言語処理において、計算機が使う辞書にはそのような単純な辞書が有効な場合も大いにある。しかし、辞書は誰が使うかを考えてみると、子供たちや学習者が手にすることが多いと考えられる。そのような実用性を考慮すると、「コーパス≒現実」と「辞書≒規範」の距離は大きい。

第３に、辞書記述にあたって、丁寧に用例を見ていくと結局コストがかかるという問題がある。「丁寧に見る」ことは、すなわち時間を通じて人件費がかかるということである。多数の項目の記述を考えると、どこまで予算がかけられるかが不透明である。このようなことをするべきかどうかを考えると、判断がむずかしい。

第３の問題は、いい辞書は、丁寧に時間をかけなければできあがらないが、そのような体制が可能かという問題につながる。これを考慮すると、どのレベルの記述をする場合、どれくらいのコストがかかるのかを明らかにするようなことも、辞書作成の一側面ではあるまいか。

BCCWJ についても、考えておくべき点がある。BCCWJ のすでに領域内公開された部分でも、間違いがたくさんあるということである。BCCWJ もコーパスである以上、間違いを含むのは当然である。作成過程の間違いではなく、原著作物がはじめから含んでいる間違いである。とすると、その利用を考える場合も、間違いがあることを前提にしなければならない。今回、BCCWJ を調査してみて、国会会議録などでも間違いがあることを発見した。どんなに丁寧に作ろうとも間違いが入り込むのだということを実感した。ならば、それを前提とした利用法が考えられるべきであろう。

注　本稿は、辞書編集班報告書 JC-C-08-01 (2009.2)中の「コーパスに含まれる「間違い」をめぐって」の簡略版である。

用例クラスタと辞書の語義との対応付けによる新語義の発見

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)
田中 博貴 (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)

Detection of New Senses by Finding Dictionary Defined Senses for Clusters of Example Sentences

Kiyoaki Shirai (Japan Advanced Institute of Science and Technology)
Makoto Nakamura (Japan Advanced Institute of Science and Technology)
Hiroki Tanaka (Japan Advanced Institute of Science and Technology)

1 はじめに

言語処理班・北陸先端大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いる．一般に，単語の意味は辞書などで定義される．しかし，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ため，辞書による単語の意味の定義は必ずしも完全ではない．そのため，あらかじめ語義を定義せず，単語のインスタンス (用例) をクラスタリングすることで単語の意味を自動的に弁別する研究が行われている (Bordag, 2006; Fukumoto and Tsujii, 1994; 九岡他, 2008; Schütze, 1998)．しかしながら，これらの先行研究では，同じ意味を持つ単語の用例をまとめたクラスタを作成することはできるが，作成された用例クラスタがどのような意味を持つのかといった意味の解釈は行われていない．

そこで，本研究では用例クラスタの意味の解釈を行う．具体的には，自動作成された用例クラスタに対し，その単語の意味が既存の辞書の語義に対応するか，あるいはどの意味にも対応しない新語義なのかを判定する手法を提案する (田中他, 2009; 田中, 2009)．

2 既存語義への対応付け

本節では，自動作成された用例クラスタに含まれる単語の意味が辞書のどの語義に対応するのかを判定する手法について述べる．なお，ここでは用例クラスタは新語義に対応するのではなく，既存の辞書の語義のいずれかに該当すると仮定する．

2.1 概要

対象単語を w とする．コーパスから抽出された w の複数の用例をクラスタリングし，作成された用例クラスタを C_i とする．ここで C_i は，同じ意味を持つとみなせる単語 w の用例の集合である．一方，S_j は，辞書によって定義された w の語義とする．このとき，用例クラスタ C_i に対応する語義 S_j を以下の手法により決定する．

まず，用例クラスタ C_i を特徴ベクトル $\vec{c}_i$ で表現する．同様に，語義 S_j を特徴ベクトル $\vec{s}_j$ で表現する．C_i に対し，ベクトル間の類似度が最大となる語義 S_j を選択する (式 (1))．

$$S_{selected}(C_i) = \arg\max_{S_j} sim(\vec{c}_i, \vec{s}_j) \quad (1)$$

ここではベクトル間の類似度はコサイン類似度とする．以下，特徴ベクトル $\vec{c}_i$ と $\vec{s}_j$ の構成方法について述べる．

2.2 用例クラスタの特徴ベクトル

用例クラスタの特徴ベクトル $\vec{c}_i$ は，クラスタ C_i に含まれる用例において，対象語 w の周辺に出現する単語を基に作成する．ただし，クラスタリングによって作成された用例クラスタの中には用例数が少ないものもある．実際に我々が用例のクラスタリングを試みたところ，1 個または 2 個の用例

[1] kshirai@jaist.ac.jp

のみで 1 つのクラスタが作成されることもあった．そこで，ベクトルがスパースになるのを避けるため，間接的な単語間の共起も考慮する．

まず，図 1 のような単語の共起行列 A を作成する．

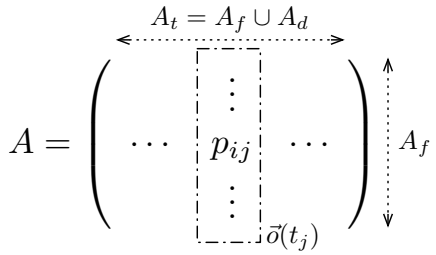


図 1: 単語の共起行列

A_f は，A の行に対応し，クラスタの特徴ベクトル $\vec{c}_i$ の素性となる単語から構成される単語集合である．ここでは，BCCWJ [2]の Yahoo!知恵袋コーパスにおいて，出現頻度上位 10,000 の自立語の集合を A_f とした．ただし，式 (2) に定義する $df(t)$ が 0.01 以上の単語 t は一般的すぎるとみなして除外した．

$$df(t) = \frac{\text{単語 } t \text{ が出現する文書数}}{\text{コーパスにおける全文書数}} \tag{2}$$

一方，A の列に対応する単語集合 A_t は，用例クラスタにおいて対象単語の周辺に出現すると仮定される単語の集合である．基本的には，A_t は A_f と同じ単語集合とする．ただし，実際には A_t は A_f と A_d の和集合としている．この理由ならびに A_d の定義については 2.3 項で述べる．最後に，行列の要素 p_{ij} は単語 t_i と t_j の文書内共起確率である．正確には，p_{ij} は，j 列目の単語 t_j が出現する文書があったとき，同じ文書内に i 行目の単語 t_i が出現する確率 $P(t_i|t_j)$ である．p_{ij} は Yahoo!知恵袋コーパスから学習する．また，j 列目のベクトルを単語 t_j の共起ベクトル $\vec{o}(t_j)$ とする．

クラスタの特徴ベクトル $\vec{c}_i$ を式 (3) と定義する．

$$\vec{c}_i = \frac{1}{N} \sum_{e_{ik} \in C_i} \sum_{t_l \in e_{ik}} \vec{o}(t_l) \tag{3}$$

e_{ik} は用例クラスタ C_i に含まれる用例を，t_l は用例 e_{ik} の文脈に出現する自立語を表わす．また，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．用例の文脈に直接出現する単語 t_l ではなく，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．

2.3 語義の特徴ベクトル

語義の特徴ベクトル $\vec{s}_j$ は辞書の語釈文から作成する．本研究では語義の定義に用いる辞書として岩波国語辞典 (西尾他, 1994) を用いた．ただし，岩波国語辞典の語釈文は，その全てが単語の意味を説明した定義文ではない．そこで，岩波国語辞典の語釈文を以下の 4 つのタイプに分類した．

①定義文 ： 単語の意味を説明した文．

②例文 ： その語義の典型的な用例．

③参照見出し ： 別の見出し語または語義への参照．

④その他 ： 上記 3 つに当てはまらない文．見出しの英語表記，注釈などが該当する．

[2] http://www.tokuteicorpus.jp/

S_1 ①型。模型。 ③▽↓もけい。
S_2 ①手本。模範。 ②「これをモデルにしてやれば間違いない」
S_3 ①美術製作の対象となるもの・人。文学作品の人物の素材となる人。
S_4 ①「ファッション　モデル」の略。流行の服装をして、客に見せたり写真に撮らせたりするのが職業の（女の）人。④▽ｍｏｄｅｌ

図 2: 「モデル」の語釈文

図 2 は岩波国語辞典における「モデル」の 4 つの語義の語釈文と，そのタイプ分けの例である．これら 4 つのうち,「参照見出し」と「その他」は単語の意味を表わしていると言い難いため，特徴ベクトル $\vec{s_j}$ の作成には用いない方がよいと考えられる．我々は，語釈文を上記 4 つのタイプに分類するため，簡単なパタンマッチによるプログラムを実装した．

次に,「定義文」に分類される語釈文のみを用いて $\vec{s_j}$ を式 (4) のように作成する．

$$\vec{s_j} = \frac{1}{N} \sum_{t_k \in d_j} \vec{o}(t_k) \tag{4}$$

d_j は語義 S_j の語釈文における定義文を，t_k は d_j に出現する自立語を表わす．また，N は正規化定数である．$\vec{s_j}$ は，用例クラスタの特徴ベクトル $\vec{c_i}$ と同様に，t_k の共起ベクトル $\vec{o}(t_k)$ の和を正規化することによって得られる．なお，図 1 の共起行列における単語集合 A_d は，辞書全体における語釈文のうち「定義文」に出現する自立語の集合である．A_f と A_d の和集合を A_t としたのは，コーパスにおける出現頻度が低いために A_f には含まれないが辞書定義文には出現する自立語に対して，共起ベクトル $\vec{o}(t_k)$ を得るためである．

2.3.1 辞書の例文の利用

辞書の語釈文のうち,「定義文」だけでなく「例文」もまた単語の意味を識別する有力な手がかりになると考えられる．そこで，定義文と例文の両方を用いて，$\vec{s_j}$ を式 (5) のように作成する．

$$\vec{s_j} = \frac{1}{N} \left(\sum_{t_k \in d_j} \vec{o}(t_k) + \sum_{t_l \in e_j} w_e \cdot \vec{o}(t_l) \right) \tag{5}$$

e_j は語義 S_j の語釈文中の例文を，t_l は e_j に出現する自立語を表わす．また，w_e は例文に出現する自立語の共起ベクトルに対して与えられる重みである．ここでは，コーパスから作成された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．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，コーパスから作成された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，例文に出現する自立語の共起ベクトルに高い重みを与えるようにした．なお，w_e の値は実験的に決定する．

2.3.2 特徴ベクトルの補正

一般に，辞書における定義文や例文の長さは短いため，作成された語義の特徴ベクトル $\vec{s_j}$ はスパースになりやすい．特徴ベクトルがスパースであるとは，ここでは多くの素性 (あるいはベクトルの次元) に対する特徴ベクトルの値が 0 となる状態を指す．例えば，図 2 の語義 S_1 においては，定義文に出現する自立語の数は 2 つしかなく，$\vec{s_j}$ におけるほとんどの素性の値が 0 になる．

用例クラスタと辞書の語義との対応付けを試みた結果，定義文が比較的長くスパースではない特徴ベクトルを持つ語義の方が，定義文が短くスパースな特徴ベクトルを持つ語義と比べて，一貫して用例クラスタの特徴ベクトルとの類似度が高くなる傾向が強いことがわかった．その結果，用例クラスタは常に同じ語義に対応付けられてしまう．この問題を解決するためには語義の特徴ベクトルのスパースネスを緩和する必要がある．

そこで，語義の特徴ベクトルを補正し，値が 0 となる素性の数を減らすことを試みた．具体的には，以下の式 (6) によって補正後の特徴ベクトル $\vec{s}'_j$ を作成した．

$$\vec{s}'_j = \frac{1}{N_1} (\vec{s}_j + w_c \cdot \vec{m}) \tag{6}$$

$$\vec{m} = \frac{1}{N_2} \sum_{t_m \in A_f} \mathrm{s_j}(t_m) \cdot \vec{o}(t_m) \tag{7}$$

式 (6) において，$\vec{s}_j$ は式 (4) や (5) で作成した語義 S_j の元の特徴ベクトル，$\vec{m}$ は補正ベクトルである．すなわち，補正後の特徴ベクトル $\vec{s}'_j$ は両者の重み付き和である．w_c は補正ベクトルに対する重み，N_1 はベクトルの大きさを 1 にするための正規化定数を表わす．次に，補正ベクトル $\vec{m}$ を式 (7) で定義する．A_f は $\vec{s}_j$ の素性となる単語集合，$\mathrm{s_j}(t_m)$ は $\vec{s}_j$ における素性 t_m に対するベクトルの値，N_2 は正規化定数である．すなわち，元の特徴ベクトルの素性 t_m に対してそれらの共起ベクトル $\vec{o}(t_m)$ の重み付き和をとることで補正ベクトルを作成し，これを元の特徴ベクトルに足すことによってベクトルのスパースネスを緩和している．

3 新語義の判定

本節では，用例クラスタが，辞書における既存の語義のいずれにも該当しない新語義の用例を集めたものであるかを判定する手法について述べる．ある対象単語に対し n 個の用例クラスタ C_i が作成されたとする．各 C_i に対し，既存語義近接度 K_i を求める．K_i は，C_i の持つ意味が既存の辞書の意味にどれだけ近いかを表わす指標で，式 (8) のように既存の語義 S_j との類似度の最大値と定義する．

$$K_i \ = \ \max_j \ sim(\vec{c}_i, \vec{s}_j) \tag{8}$$

直観的には，既存語義近接度が小さければ小さいほど，その用例クラスタは新語義である可能性が高い．そこで，用例クラスタを K_i の降順にソートする．以下，用例クラスタ $C_1 \sim C_n$ は，K_i の大きい順に並んでいるものとする．

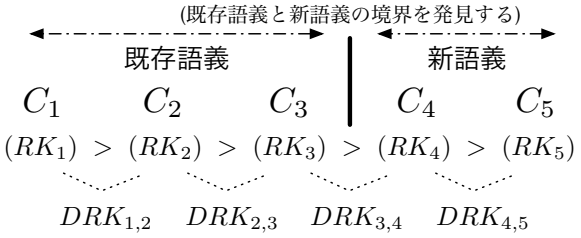


図 3: 新語義の判定

次に，上記のように並べられた C_i に対し，既存語義と新語義とを分ける境界を発見する (図 3)．まず，相対既存語義近接度 RK_i を，最も大きい既存語義近接度 K_1 に対する K_i の相対値 $(= K_i/K_1)$ とする．さらに，隣接する用例クラスタ C_i と C_{i+1} の相対既存語義近接度の差を $DRK_{i,i+1}$ とする (式 (9))．

$$DRK_{i,i+1} \ = \ RK_i - RK_{i+1} \tag{9}$$

ここでは新語義の判定を行う 2 つの手法を提案する．

- 手法 1

  $1 \leq i \leq N-1$ のうち，$DRK_{i,i+1}$ が最も大きい i を見つけ，$i+1$ 番目以降の用例クラスタ $C_{i+1} \cdots C_N$ は新語義であると判定する．これは，既存語義近接度の差が大きいところが既存語義と新語義の境界になっているという仮定に基づく．

- 手法 2

  手法 1 の条件に加え，最大の $DRK_{i,i+1}$ の値が閾値 T_k よりも大きいときに限り，用例クラスタ $C_{i+1} \cdots C_N$ を新語義と判定する．T_k 以下の場合は新語義に相当する用例クラスタは存在しないものとする．すなわち，既存語義近接度の差が十分大きくなければ既存語義と新語義との境界とはみなさない．

手法 2 で閾値 T_k を設定する際，用例クラスタと辞書の語義との類似度の大きさは対象単語によってばらつきがあるため，既存語義近接度 K_i に対して閾値を設定することは困難である．そのため，相対化した既存語義近接度 RK_i に対して閾値 T_k を設定した．

4 予備実験

4.1 実験データ

提案手法を評価する予備実験を行った．まず，対象単語として以下の 23 個の単語を用いた．

> モデル，ネタ，カバー，ウイルス，ソース，肉，サービス，地方，アルバム，コード，自分，場合，時間，意味，電話，一緒，目，以前，代，顔，系，郵便，反応

各対象単語に対し，Yahoo!知恵袋コーパスの中から 100 個の用例をランダムに選択した．

本研究では，用例集合に対してクラスタリングを行い，同じ語義を持つ用例を集めたクラスタを自動作成し，各クラスタに対応する辞書の語義を選択することを目的としている．しかし，自動的に作成されたクラスタは，違う語義を持つ用例が 1 つのクラスタにまとめられるという誤りを含む．ここでは，用例クラスタと語義との対応付け，ならびに新語義の判定手法を評価するために，人手で作成した完全に正しい用例クラスタを実験に用いた．具体的には，コーパスから抽出された用例に対して人手で語義を付与し，同じ語義を持つ用例をまとめてクラスタを作成した．用例に付与する語義は岩波国語辞典の中分類の語義とした．これは比較的荒い意味分類である．また，岩波国語辞典に該当する語義がない場合は新しい語義を定義した．

4.2 実験結果

人手で作成された用例クラスタに対し，2 節で提案した手法によって対応する辞書の語義を選択した．ここでは，語義の特徴ベクトル $\vec{s_j}$ の作成方法の違いに応じて，以下の 4 つの手法を比較した．

M_d ：式 (4) によって $\vec{s_j}$ を作成する．すなわち，辞書の語釈文のうち「定義文」に含まれる単語のみを用いる手法．

$M_{d,e}$ ：式 (5) によって $\vec{s_j}$ を作成する．すなわち，辞書の語釈文のうち「定義文」と「例文」に含まれる単語を用いる手法．

M_d^c ：式 (4) によって $\vec{s_j}$ を作成した後，式 (6) によって特徴ベクトルの補正を行う手法．

$M_{d,e}^c$ ：式 (5) によって $\vec{s_j}$ を作成した後，式 (6) によって特徴ベクトルの補正を行う手法．

実験結果を表 1 に示す．

表 1: 語義の対応付けの実験結果

	M_d	$M_{d,e}$	M_d^c	$M_{d,e}^c$
クラスタ数	63	63	63	63
正解クラスタ数	26	31	31	39
正解率	0.41	0.49	0.49	0.62

表 1 において,「クラスタ数」は対象単語 23 語に対して人手で作成した用例クラスタの総数 (新語義に対応する用例クラスタは除く),「正解クラスタ数」は正しい語義に対応付けられた用例クラスタの数，正解率はその割合を表わす．なお，式 (5) における w_e，式 (6) における w_c はいくつかの値を試し，正解率の一番大きい値を定めた．その結果，$w_e = 2$，$w_c = 5$ となった．ただし，これらのパラメタの調整はテストデータを用いて行っているという点で，今回の実験はクローズドテストである．

表 1 の結果から，辞書の定義文だけでなく例文を用いて語義の特徴ベクトルを作成した方が，また値が 0 となるベクトルの素性の数を減らすための補正を行う方が，辞書の語義との対応付けの正解率が向上することがわかった．ただし，手法 $M^c_{d,e}$ における正解率は 0.62 であり，十分高いとはいえず，改善の必要がある．

次に，4.1 項で作成した全ての用例クラスタに対し，3 節で提案した方法を用いて，用例クラスタが新語義であるか否かを判定する実験を行った．判定の精度，再現率，F 値を表 2 に示す．

表 2: 新語義判定の実験結果

手法	N_1	N_2	N_2	N_2
T_k	–	0.03	0.025	0.02
精度	0.43	0.62	0.63	0.57
再現率	0.65	0.50	0.60	0.65
F 値	0.52	0.55	0.62	0.60

表 2 において，N_1，N_2 はそれぞれ 3 節で述べた「手法 1」「手法 2」を表わす．N_2 については閾値 T_k を変えて実験を行った．新語義判定の F 値は，$T_k = 0.025$ のときの手法 2 が最大で 0.62 であった．

5 おわりに

本論文では，コーパスから作成された用例クラスタに対する意味の解釈，すなわち用例クラスタに対応する辞書の語義を選択する手法と，辞書のどの語義にも対応しない新語義であるかを判定する手法について述べた．今後の課題としては，提案手法の改良・洗練，自動作成された用例クラスタに対する実験，各手法におけるパラメタの最適化方法の検討などがあり，これらに順次取り組む予定である．

文献

Stefan Bordag (2006) “Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation,” in *Proceedings of the EACL*, pp. 137–144.

Fumiyo Fukumoto and Jun’ichi Tsujii (1994) “Automatic Recognition of Verbal Polysemy,” in *Proceedings of the COLING*, pp. 762–768.

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.

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.

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．P2-31.

田中博貴 (2009)「用例のクラスタリングに基づく単語の新語義の発見」，修士論文，北陸先端科学技術大学院大学.

公募班研究発表

3月16日（月）　10:00〜12:00

所謂引用助詞「と」が標識する構文の用法再考
―フレーム・フレーム要素・フレーム間関係の観点から―

▶藤井 聖子

エントロピーと冗長度を指標とした語彙的・統語的複合動詞の比較研究

▶玉岡 賀津雄

用例間類似度測定のための属性重みの推定

▶新納 浩幸、佐々木 稔

BCCWJにおける推量副詞とモダリティ形式の共起

▶スルダノヴィッチ・イレーナ、ベケシュ・アンドレイ、仁科 喜久子

所謂引用助詞「と」が標識する構文の用法再考
−フレーム・フレーム要素・フレーム間関係の観点から−

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）†

Uses of So-called Quotative TO Constructions Revisited:
Considering Frames, Frame Elements, and Frame-to-frame Relations

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

１．本研究の目的・焦点と背景

本研究では、所謂引用助詞「と」の用法に関して、日本語学において行なわれてきた研究からの知見や仮説に基づき、『BCCWJ 2008 領域内公開版』コーパスを用いて分析しつつ、フレームネットの理論的・記述的枠組みで「と」の用法の問題がどのように捉えられ、どのように分析できるか・どのように記述されるかを考察し提案する。

日本語の引用の研究は、日本語学において多くの優れた研究が活発に行なわれ展開してきた（その代表的研究として, 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てきた。国立国語研究所においても、様々なプロジェクトにおいてその記述が行なわれ（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、「と」などが標識する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、多くの研究によって指摘され、特に話し言葉の研究で分析されてきた（「と」に関しては, Okamoto 1996, 加藤 1998, 2008,等;「って」に関しては, 山崎 1996, S. Suzuki 1996, R. Suzuki 1999, 加藤 2008, 等;「みたいな」標識の引用に関しては、S. Suzuki 1995, 加藤 2008, Fujii 2002, 2006、等）。

本研究では、述部を伴って使用される所謂引用「と」助詞の用法に焦点を絞り、ト標識の節(または句)を受ける述部の意味的・形式的特徴にも注目し、ト標識節(句)の用法を、ト節(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。また、ト標識節内の末尾表現(文末モダリティ表現等)にも着目しつつ、そのト標識節と述部との関係を捉える。これらの意味で、「構文」の分析・記述を目指す立場をとる。

２節以降に詳述・例示するが、フレームネットの枠組みでは、ト標識の節(句)を受ける述部の述語を**フレーム喚起語**として捉え、どのフレームを喚起するフレーム喚起語であるかを考察する。また、「ト標識節と述部との関係を捉える」という点は、ト標識節(句)が喚起フレームのどのような**フレーム要素**として位置づけられるかという観点から考察する。

このような観点での本研究が依拠するフレーム意味論・フレームネットの枠組みを、まず２節で要点のみまとめておく。３節で、所謂引用助詞「と」の用法に関して、日本語学での先行研究を鑑みて、本分析での用法分類の要点のみ手短にまとめる。それらの分類と仮説に基づき、『BCCWJ 2008 領域内公開版』コーパスを質的・量的に分析した結果の概要を４節、５節で報告する。ト標識節(句)構文の出現傾向を述部のフレーム喚起語を中心に示しつつ、（本分析の第一義的目的ではないが）白書コーパスと書籍コーパスとの比較も示す。６節では、引用のト標識節(句)構文を、述部の喚起するフレーム、および、そのフレーム間関係の観点で、ト標識節(句)構文の関与フレーム・意味領域の統合を試みる。

† sfujii@boz.c.u-toko.ac.jp

２．フレーム意味論に基づく FrameNet の理論的・方法論的枠組

２.１　FrameNet と日本語フレームネット

フレームネット(FrameNet: FN, PI: Charles J. Fillmore)は、1997 年よりカリフォルニア大学バークレー校およびInternational Computer Science Instituteにおいて開発･構築されてきた(現在も拡張継続中）。フレーム意味論（Fillmore 1982, Petruck 1996, etc.）に基づき、コーパスデータを参照しつつ、英語の語彙の意味および参与する構文の分析を行い、その意味・形式の記述・情報を電子語彙体系として構築・資源化するプロジェクトである(Ruppenhofer 他 2006; http://www.icsi.berkeley.edu/~framenet/)。本特定領域の研究班で行なっている「日本語フレームネット」(http://jfn.st.hc.keio.ac.jp/ja/index.html; 小原 2006, 小原他 2005a, 2005b, Ohara *et al.* 2003, 2004, 斎藤他 2007,等を参照）は、理論的・方法論的に FrameNet に依拠し、FrameNet との密接な連携のもとに、小原京子氏を中心に進めてきた。

FrameNet では、語彙項目（Lexical Unit: LU）各々の意味的・構文的特質や語彙項目間の関係を、フレームやフレーム要素（Frame Element: FE）という概念を用いて記述し、コーパスから代表的用例を選別し、各々の語彙項目の結合価パターンごとに用例へのフレーム要素の意味タグ付けを行う。さらに、フレーム間の関係も記述することを目標とする。

２.２　フレーム、フレーム要素、フレーム間関係：引用ト標識節構文の分析のために

フレーム意味論および FN でのフレームとは、言語コミュニティにおいて言語的相互作用（その意味理解）の基盤となる概念構造や信念・慣習・制度的パターンのスキーマ化された表象である。テキストや発話における語の使用によって、人の知識構造のうち一つの要素が使われ活性化されると、他の概念集合とその有機的繋がりが喚起され利用可能になる。そのような認知操作の母体となる知識構造が、フレームである。語彙の中には、フレーム喚起要素になる語とそうでない語とがある。動詞は概ね文の中の主なフレーム喚起要素となることが多いが、動詞のみでなく、名詞や形容詞も重要なフレーム喚起要素である。支援動詞構文では、日本語の軽動詞「する」等を伴う事態性名詞（主張, 判断, 等）が主なフレーム喚起語となり、「する」や他の支援動詞は主な喚起要素とならない(藤井,上垣 2008)。

各々のフレームは、その知識構造に参与し有機的に繋がりをもつ幾つかの参与要素から成り立っている。これをフレーム要素と呼ぶ。テキストや発話においては、各々のフレームにおいて結びつくフレーム要素が、文(節)を構成する要素として組み合わされる。

このような FrameNet の枠組みでは、ト標識の節(句)を係り受ける述部の述語を、フレーム喚起語として捉える。従って、まず、どのフレームを喚起するフレーム喚起語であるかという観点からその特徴付けを行う。各々の用法で意義が異なれば、異なるフレームを喚起すると捉える。(従って、その喚起フレームの同定には述部のみでなく構文を考慮する。)

「ト標識節と述部との関係を捉える」ために、ト標識の節(句)が、フレーム喚起語(述部)の喚起するフレームにどのようなフレーム要素として参与しているかを考察し記述する。

図１に、BCCWJ2008 の本分析において同定できたト標識節(句)を受ける述部のフレーム喚起語が喚起するフレームのうち、引用部が言語表現の内容で述部が言動・思考・心的活動･判断･感情に関するフレームを喚起する用法のフレーム (紙幅制限でその一部)を示す。

FrameNet では、個々のフレームを明示するのに加え、フレーム間の関係も考察し構築している。図１に示した、ト標識節(句)構文が喚起するフレームの広がりとフレーム間関係に関しては、６節で図１に立ち戻って考察する。

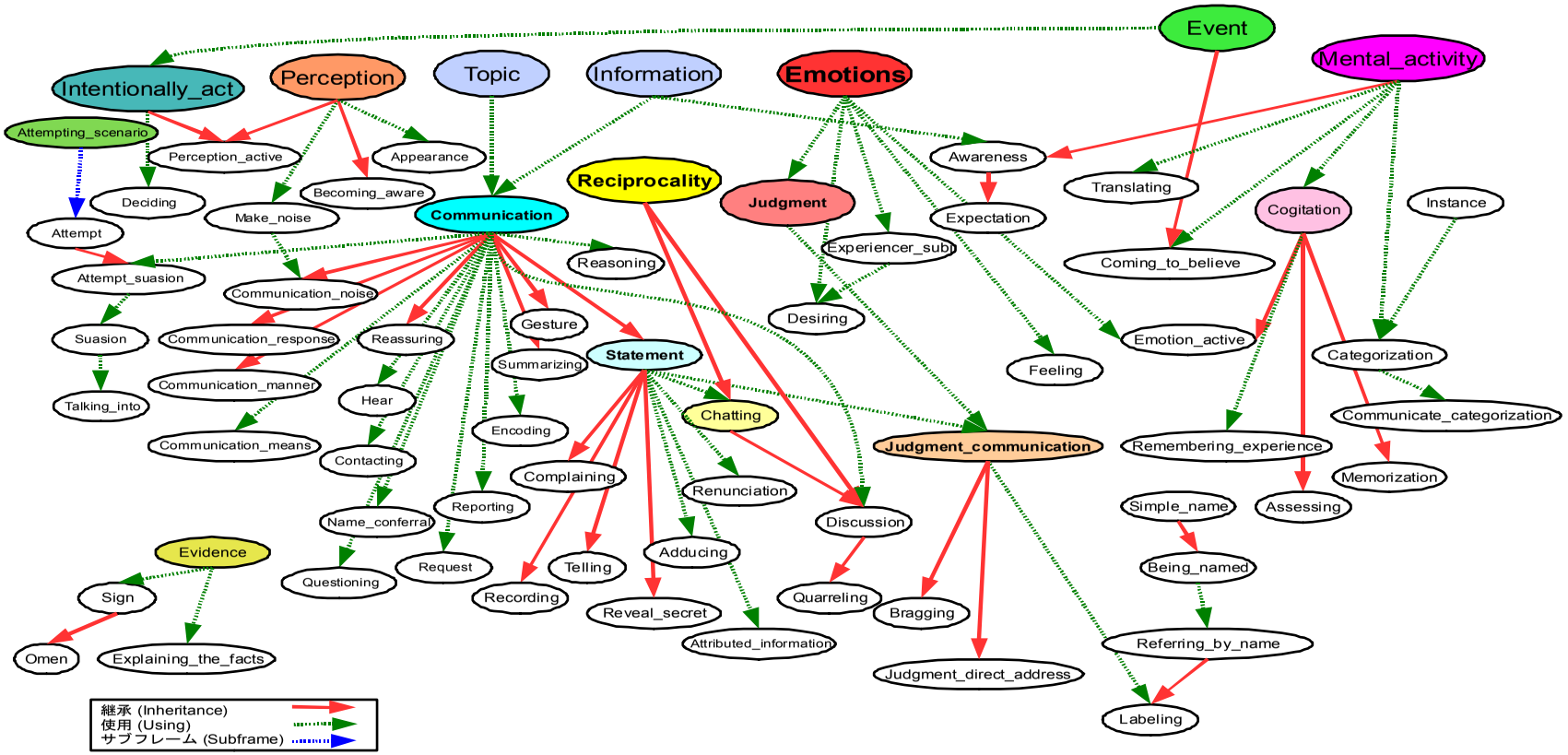


図１　ト標識節(句)構文が喚起するフレーム(BCCWJ2008 の本分析で認められたもの) の広がりとフレーム間関係[1]：
引用部が言語表現の内容であり、述部が言動・思考・心的活動・判断・感情に関わるフレームを喚起する用法のみに関して

[1] FrameNet では、個々のフレームを明示するのに加え、フレーム間の関係も考察し構築している。フレームとフレームとは様々な関係で結ばれている。現行のＦＮでは、継承 Inheritance, サブフレーム Subframe, 視点 Perspective_on, 先行 Precedes, 使用 Using, 使役相 Causative_of, 起動相 Inchoative_of, 参照 See also の関係がある。最も基本的な関係は Inheritance で、下位のフレームが上位のフレームのコア要素をすべて引き継ぐ関係である。
このフレーム間関係は、FrameGrapher というシステムで図式化できる(http://framenet.icsi.berkeley.edu/FrameGrapher/)。ただ、FrameGrapher で自動的に描いた図式は横幅が広い図になり、（上位・下位のレベル数や下位フレームの表示数は指定できるが）表示必要な関連フレームのみを取捨選択することができないため、上の図は FrameNet データベースのデジタル情報に基づいて手動で描いたものである。（従って、図作成上のエラーは著者の責任である。）
フレーム間関係は現在も更新されており、特に上記領域は改変の議論中である。図 1 は 2008 年 1 月~09 年 2 月の FN データベースに基づいて作成した。

３．所謂引用の助詞「と」の用法の大分類、着目した現象

本稿で報告する『BCCWJ 2008 領域内公開版』コーパスを用いた分析では、国立国語研究所(1951), 藤田(1986), 特に山崎 (1993)の洞察・提案を鑑みつつ、所謂引用の助詞「と」の用法の大別を以下の大分類とした。この作業分類に基づき手作業でコーディングをした。

表１．所謂引用の助詞「と」が標識する構文の用法に関する、本分析における作業分類

引用助詞「と」の用法大別	引用の類型	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	言動思考	
				言動思考:言動	いう、話す　主張、回答、説明
				言動思考: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 様々 （立つ、手渡す、等）
	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、他
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	（手作業で削除）
	頻繁に「引用助詞」と誤解析された形態素			誤:接続助詞	（手作業で削除）

題目等で「所謂引用…」としたのは、品詞分類上(言語処理での”形態素”分類で)「引用助詞」と認められる「と」助詞の用法の中には、意味的に狭義の引用でない用法も含まれるからである。この点では、表１のカテゴリーQ(上半分)とカテゴリーP(下)「狭義の引用ではない用法」との間に、重要な第一分岐があるとみなす（この点,山崎 1993 と同じ立場をとる)。

ただ一方で、(表１において明示していない指標であるが、本研究での指標の一つである)「ト標識引用部と述部との関係」という指標においては、カテゴリーQ の内部で既に別の重要な分岐があり、カテゴリーQ とカテゴリーP とをまたぐ別の重要な第一分岐となる。この分岐による特徴を本研究でまず大掴みに、「**内**の関係」vs.「**外付け**の関係」とする。

「内の関係」は、カテゴリーQ でもカテゴリーP でも「ト標識引用部と述部との関係」として統語的にも意味的にも認められる。一般的な統語的観点で言えば、ト標識引用部が述部の主幹述語の項構造の項となっているかどうかである（但し実際は、一つの動詞の項構造の項かどうかという単純な問題に留まらない)。これを、引用用法のカテゴリーQ の中で言うと、ト標識引用部が述部(の主幹述語)の補語(補文)になっているかいないか、という指標である。補語(補文)になっている用法が、Q1「言語表現補語引用型」と名付けた、典型的な引用文である。発話行為動詞、思考動詞、感情動詞等の補語(補文)としてト標識引用部が述部の内容を補完する（殆どが引用助詞トは補文標識とみなしてよいケース)。統語的に必須項であるかどうかは別であるが、述部の述語の特性としてト標識引用部を位置づけることができる。

(1)　Q1.　有難うといった。有難うと礼状を書いた。有難うと感激した。
　　Q2-1　有難うと頭を下げた。　有難うと涙を流した。
　　Q2　有難うと飛び出て行った。　有難うとバナナを手渡した。

他方、「**外付け**の関係」は、Q2「言語表現外付け引用事態型」と名付けた、述部の述語の特性としてト標識引用部を位置づけることができない用法である。この用法に関しては５節で考察する。Q1 と Q2 との中間に位置づけられる用法もある。

表２．BCCWJ2008(領域内版) 白書・書籍コーパスにおける引用助詞トと共起する**動詞**頻度順

白書					書籍:文学				書籍:言語				書籍:技術工学				書籍:自然科学				書籍:哲学				書籍:歴史				書籍:総記			
	17513			17513		49111	49111	49111						11,733				13684				19623				27846				6515			
1	する	2,501	4627	結果	1	思う	7474	9075	言動思考	1	いう	695	803	言動思考	1	いう	1,739	2007	言動思考	1	いう	1,739	2007	言動思考	1	いう	2382	2770	言動思考	1	思う	936	1055	言動思考
2	なる	1,897	1972	結果	2	いう	5420	6090	言動思考	2	思う	661	728	言動思考	2	思う	1,504	1626	言動思考	2	思う	1,504	1626	言動思考	2	思う	2269	2497	言動思考	2	いう	923	1066	言動思考
3	考える	1,589	1,614	言動思考	3	言う	3592	4734	言動思考	3	言う	465	543	言動思考	3	する	891	1956	言動思考	3	する	891	1956	言動思考	3	する	1658	3333	言動思考	3	言う	488	589	言動思考
5	いう	823	990	言動思考	4	する	3307	6364	結果	4	する	431	1013	言動思考	4	言う	651	774	言動思考	4	言う	651	774	言動思考	4	言う	1464	1764	言動思考	4	する	449	997	言動思考
8	みる	472	535	言動思考	8	考える	796	1109	言動思考	6	考える	287	322	言動思考	6	考える	455	524	言動思考	6	考える	455	524	言動思考	6	考える	882	998	言動思考	6	考える	182	205	言動思考
9	思う	471	498	言動思考	9	呼ぶ	569	626	言動思考	7	呼ぶ	172	179	言動思考	7	呼ぶ	374	417	言動思考	7	呼ぶ	374	417	言動思考	9	呼ぶ	744	822	言動思考	9	呼ぶ	150	157	言動思考
11	いえる	407	439	言動思考	10	いえる	533	905	言動思考	12	なる	86	162	言動思考	11	いえる	147	190	言動思考	11	いえる	147	190	言動思考	13	いえる	196	253	言動思考	11	ある	334	652	言動思考	11	いえる	68	87
12	認める	177	277	言動思考	13	聞く	362	462	言動思考	13	言える	71	115	言動思考	13	なる	108	262	言動思考	13	なる	108	262	言動思考	14	感じる	166	197	言動思考	12	書く	301	401	言動思考	12	書く	61	77
12	認める	88		他	14	書く	280	360	言動思考	14	書く	68	84	言動思考	14	聞く	84	101	言動思考	14	聞く	84	101	言動思考	15	なる	154	431	言動思考	13	いえる	283	398	言動思考	14	なる	54	141
14	答える	230	232	言動思考	16	なる	249	835	結果	15	ある	67	151	言動思考	16	感じる	75	80	言動思考	16	感じる	75	80	言動思考	16	述べる	140	164	言動思考	14	なる	233	558	言動思考	16	聞く	51	64
18	呼ぶ	138	145	言動思考	17	答える	244	377	言動思考	16	見る	63	72	言動思考	17	書く	68	78	言動思考	17	書く	68	78	言動思考	17	ある	127	322	言動思考	16	みる	214	238	言動思考	18	見る	41	52
21	言える	130	176	言動思考	19	言える	233	391	言動思考	18	聞く	41	47	言動思考	18	認める	57	68	言動思考	18	認める	57	68	言動思考	18	見る	109	140	言動思考	18	聞く	172	206	言動思考	19	述べる	36	40
22	見込む	61	129	言動思考	21	見る	199	295	言動思考	20	いえる	40	53	言動思考	20	思える	47	70	言動思考	20	思える	47	70	言動思考	19	答える	98	118	言動思考	19	述べる	167	193	言動思考	20	感じる	35	42
22	見込む	64		他	22	伸ず	193	217	言動思考	21	答える	36	45	言動思考	21	ある	44	172	言動思考	21	ある	44	172	言動思考	20	信じる	94	110	言動思考	21	ある	156	189	言動思考	21	見る	24	111
25	言う	111	144	言動思考	24	わかる	173	234	言動思考	22	感じる	35	39	言動思考	22	よぶ	42	45	言動思考	22	よぶ	42	45	言動思考	21	書く	94	108	言動思考	22	称ず	154	169	言動思考	22	言える	24	35
32	見る	57	110	言動思考	25	おもう	170	209	言動思考	23	述べる	35	36	言動思考	23	比べる	39	40	言動思考	25	言える	37	54	言動思考	22	聞く	93	118	言動思考	23	記す	142	171	言動思考	23	おもう	21	22
32	見る	31		他	27	感じる	166	207	言動思考	24	みなす	29	30	言動思考	25	言える	37	54	言動思考	26	述べる	37	43	言動思考	24	言える	79	116	言動思考	24	よぶ	133	142	言動思考	29	信じる	17	21
34	続く	61	80	列挙	28	ある	160	516	言動思考	27	読む	20	25	言動思考	26	述べる	37	43	言動思考	27	みる	35	40	言動思考	25	みなす	77	80	言動思考	25	言える	131	206	言動思考	30	答える	16	20
35	感じる	60	64	言動思考	30	信じる	143	190	言動思考	28	よぶ	19	19	言動思考	27	みる	35	40	言動思考	30	答える	33	38	言動思考	28	語る	69	101	言動思考	26	伝える	121	153	言動思考	31	みる	15	20
47	定める	38	43	結果:他	31	云う	138	175	言動思考	29	みる	18	20	言動思考	30	答える	33	38	言動思考	32	信じる	30	35	言動思考	29	勝う	63	70	言動思考	27	答える	103	127	言動思考	32	話す	15	20
48	みなす	37	38	結果:他	33	知る	136	249	言動思考	33	信じる	16	27	言動思考	32	信じる	30	35	言動思考	33	みなす	29	30	言動思考	30	記す	62	76	言動思考	28	感じる	89	105	言動思考	33	わかる	14	23
68	称す	5	25	言動思考	35	訊く	129	185	言動思考	34	記す	16	18	言動思考	33	みなす	29	30	言動思考	34	名づける	29	29	言動思考	31	伝える	61	86	言動思考	29	信じる	88	120	言動思考	34	叫ぶ	14	15
68	称す	18		他	36	叫ぶ	119	172	言動思考	41	称す	13	14	言動思考	34	名づける	29	29	言動思考	35	見る	28	35	言動思考	32	訊く	61	69	言動思考	31	みなす	84	86	言動思考	35	よぶ	13	13
69	題する	23	23	言動思考	37	おっしゃる	116	134	言動思考	44	使う	12	30	言動思考	35	見る	28	35	言動思考	38	称す	24	28	言動思考	33	訊ず	56	68	言動思考	33	語る	70	102	言動思考	36	称す	13	13
81	位置づける	17	18	結果:他	38	話す	111	145	言動思考	47	認める	12	17	言動思考	38	称す	24	28	言動思考	39	話す	22	28	言動思考	34	名づける	53	55	言動思考	34	思える	68	145	言動思考	37	認める	13	15
82	位置づける	17	17	結果:他	39	化す	101	101	結果	48	配す	11	11	言動思考	39	話す	22	28	言動思考	41	語る	21	28	言動思考	35	思える	51	55	言動思考	36	認める	65	72	言動思考	38	おっしゃる	11	11
84	称する	15	15	言動思考	40	見える	98	129	言動思考	52	伸ず	10	13	言動思考	40	合わせる	21	26	言動思考	42	わかる	20	27	言動思考	36	見なす	49	50	言動思考	38	名づける	58	59	言動思考	39	みなす	11	12
86	伸ばす	13	32	尺度関連	42	告げる	86	108	言動思考	53	知る	10	13	言動思考	41	語る	21	28	言動思考	44	記す	19	22	言動思考	37	よぶ	48	52	言動思考	40	見なす	56	59	言動思考	42	思える	10	12
	伸ばす	1		移動(へと)	45	笑う	78	149	行為:感情	54	訳す	10	13	言動思考	42	わかる	20	27	言動思考	47	決める	18	22	言動思考	38	称す	46	52	言動思考	44	叫ぶ	49	62	言動思考	44	笑う	10	11
93	述べる	13	14	言動思考	48	称す	76	83	言動思考	55	願う	10	12	言動思考	46	記す	19	22	言動思考	49	おっしゃる	17	20	言動思考	40	認める	45	52	言動思考	45	話す	46	63	言動思考	45	題す	10	10
89	とらえる	10	11	言動思考	49	頼む	72	101	言動思考	58	いふ	8	9	言動思考	47	決める	18	22	言動思考	51	伝える	17	21	言動思考	41	わかる	42	56	言動思考	46	化す	45	45	言動思考	46	云う	9	9
89	とらえる	1	10	他	51	みえる	71	75	言動思考	60	やく	8	8	言動思考	49	おっしゃる	17	20	言動思考	62	称する	14	14	言動思考	42	みる	40	48	言動思考	48	いふ	43	48	言動思考	47	伸ばす	9	9
108	伸び悩む	10	10	尺度関連	52	みる	71	108	言動思考	61	わかる	8	19	言動思考	51	伝える	17	21	言動思考	63	題す	14	15	言動思考	44	問う	39	49	言動思考	49	称する	43	44	言動思考	50	決める	9	11

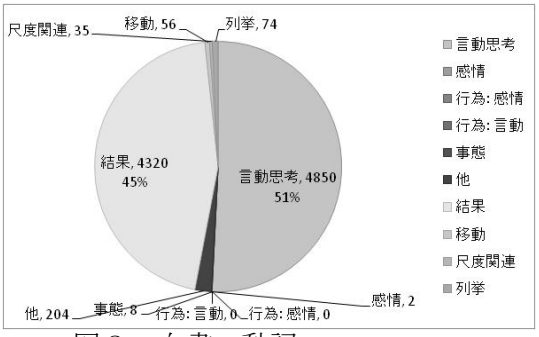


図２　白書：動詞

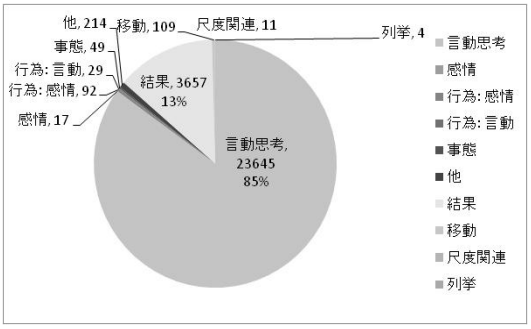


図３：書籍文学：動詞

表３．BCCWJ2008(領域内版) 白書・書籍における引用助詞トと共起する**事態性名詞**頻度順

白書					書籍:文学				書籍:言語				書籍:技術工学				書籍:自然科学				書籍:哲学				書籍:歴史				書籍:総記			
	17513			17513		49111		49111		6244	6,244			11,733				13684				19623				27846				6515			
16	回答	169	212	言動思考	44	判断	81	92	言動思考	26	説明	21	29	28	判断	35	45	23	判断	47	54	27	主張	75	91	32	主張	84	103	26	説明	20	27
24	判断	112	121	言動思考	76	主張	53	62	言動思考	32	主張	16	17	31	入力	31	34	29	主張	41	42	58	判断	30	33	35	推定	66	70	38	入力	12	12
29	増加	106	333	尺度関連	77	心配	53	98	感情	39	心配	13	17	37	主張	24	38	37	心配	23	27	60	理解	28	39	37	判断	63	69	43	指摘	10	15
30	予想	100	106	言動思考	96	説明	42	64	言動思考	40	発音	13	20	43	予想	20	22	38	期待	23	27	61	表現	28	38	43	推測	51	57	48	心配	9	11
33	減少	71	165	尺度関連	98	確信	41	53	言動思考	45	判断	12	14	44	説明	20	28	41	報告	22	38	62	努力	27	40	47	指摘	44	54	49	感心	9	13
36	推計	60	62	言動思考	107	決心	37	46	言動思考	46	表現	12	16	48	表示	18	27	43	指摘	21	29	91	想像	16	22	59	説明	36	51	51	確信	9	10
40	推定	49	49	言動思考	108	想像	36	66	言動思考	51	推測	10	11	57	確信	16	20	45	理解	20	24	75	確信	20	21	70	想像	31	45	53	認識	9	11
42	仮定	48	54	言動思考	122	約束	30	47	言動思考	57	指摘	9	12	59	仮定	15	15	49	説明	20	27	91	想像	16	22	75	報告	29	45	60	説明	8	11
43	期待	48	52	言動思考	138	話	25	92	言動思考	64	想像	8	8	66	予測	12	15	50	仮定	19	20	108	宣言	13	17	76	宣言	29	32	68	表示	7	16
46	上昇	40	77	尺度関連	140	電話	25	69	言動思考	65	推定	8	9	67	指摘	12	17	51	推定	19	21	110	評価	13	16	78	理解	29	42	77	判断	6	12
51	予測	34	38	言動思考	143	努力	23	28	言動思考	76	理解	7	10	68	発表	12	15	59	推測	16	18	111	説明	13	25	81	約束	28	30	78	努力	6	6
54	急増	32	43	尺度関連	146	期待	23	39	事態	77	表記	7	7	69	推定	11	11	60	診断	16	25	114	命名	12	13	87	発表	26	34	83	想像	6	8
58	低下	30	67	尺度関連	154	息	20	39	事態	79	解釈	7	10	70	期待	11	14	64	想像	15	17	117	指摘	12	18	89	表現	26	32	84	推定	6	6
61	繁栄	26	26	尺度関連	157	解釈	20	28	言動思考	87	報告	6	6	81	推測	9	11	67	予想	14	14	118	推定	12	12	91	心配	25	31	86	規定	6	7
62	認識	26	87	言動思考	158	証言	20	24	言動思考	88	変化	6	9	86	表現	9	10	70	解釈	14	17	119	決心	12	16	95	非難	24	29	87	解釈	6	7
63	拡大	25	54	尺度関連	167	推測	19	29	言動思考	98	記載	6	9	87	質問	9	12	75	確信	13	14	120	決意	12	12	103	評価	21	30	101	決意	5	5
64	指摘	25	62	言動思考	176	噂	17	30	言動思考	100	質問	6	9	94	努力	8	10	78	予測	12	14	121	約束	12	17	106	批判	20	33	103	納得	5	7
65	規定	25	33	言動思考	177	推定	17	18	言動思考	102	動違い	5	5	99	変化	8	12	79	努力	12	15	127	期待	11	18	107	確信	20	34	105	評価	5	7
67	評価	24	33	言動思考	178	断定	17	25	言動思考	109	感心	5	10	100	想像	8	8	81	定義	12	15	128	非難	11	15	110	期待	19	26	112	仮定	4	4
72	推測	22	22	言動思考	194	感心	16	30	感情	110	改称	5	5	101	理解	8	9	82	発表	12	14	131	定義	10	13	112	発展	19	23	115	報道	4	10
73	移行	22	23	移動	205	報告	15	24	言動思考	112	規定	5	6	108	報告	7	14	88	回答	11	16	132	心配	10	14	114	規定	19	22	126	表現	4	9
80	主張	17	25	言動思考	207	宣言	15	18	言動思考	113	訓読	5	10	109	心配	7	12	96	表記	10	10	136	断言	10	14	117	断言	18	20	127	計画	4	4
88	判定	13	14	言動思考	210	決意	15	19	言動思考	120	使役	4	4	113	発言	7	11	102	病気	9	18	137	発展	10	12	121	決心	17	18	128	証明	4	5
90	発展	13	14	移動	237	予想	13	26	言動思考	126	定義	4	4	117	規定	7	12	103	表現	9	22	138	認識	10	18	131	発音	15	20	129	損解	4	6
92	認定	13	24	言動思考	239	仮定	13	13	言動思考	130	発言	4	5	118	記憶	7	9	105	評価	9	13	158	想定	8	9	135	推察	14	14	147	判明	3	3
98	質問	12	22	言動思考	245	指示	13	23	言動思考	132	納得	4	4	130	安心	6	12	108	進行	9	9	159	推察	8	9	137	発言	14	18	150	反論	3	4
102	定義	11	11	言動思考	247	断言	13	28	言動思考	134	記憶	4	5	131	実感	6	8	112	命名	8	8	166	規定	8	10	142	想定	13	19	156	回答	3	4
107	増大	10	22	尺度関連	251	記憶	13	22	言動思考	156	一緒	3	3	133	意味	6	11	116	決意	8	11	169	質問	8	18	145	覚悟	13	14	158	報告	3	6
110	想定	10	17	言動思考	260	動違い	12	15	言動思考	159	予測	3	3	135	操作	6	8	118	記載	8	10	174	仮定	7	8	150	努力	12	20	160	定義	3	3
111	記載	10	11	言動思考	264	後悔	12	19	感情	165	努力	3	8	137	断言	6	6	125	判定	7	9	175	判示	7	7	153	強調	12	18	161	宣言	3	3

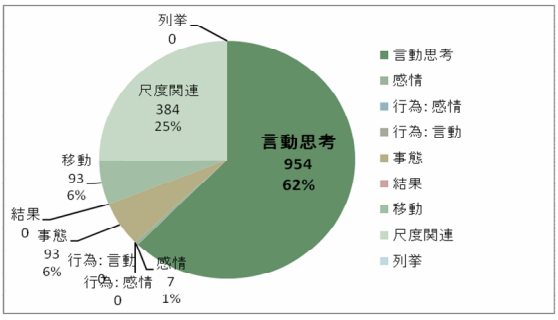


図４　白書：事態性名詞

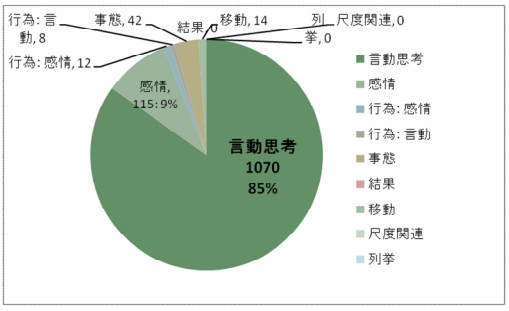


図５　書籍文学：事態性名詞

４．現代日本語書き言葉均衡コーパスを用いたト標識節(句)構文の用法の分析

『BCCWJ 2008 領域内公開版』の白書コーパスと書籍コーパスとを用いて、３節の表１で要約した作業分類に基づき、ト標識節(句)構文の引用部と述部を分析した。さらに、それとは別に、１節２節で概説した、フレーム喚起語の同定、喚起されるフレームの考察、フレーム要素の考察という観点でも分析を行った。それぞれ(関連するが)異なるステップの分析ではあるものの、紙幅の制限のため、後者も適宜言及しつつ、報告する。

対象としたコーパスは、以下８つのコーパスである：白書コーパス全体；書籍文学コーパス(8,775,301 語の内、3,683,460 語、文学全体の訳 42%)；書籍言語コーパス全体(398,497 語)；書籍技術工学コーパス全体(1,115,821 語)；書籍自然科学コーパス全体(1,074,332 語)；書籍哲学コーパス全体(1,403,199 語)；書籍歴史コーパス全体(2,141,841 語)；書籍総記コーパス全体(521,436 語)。『BCCWJ 2008 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、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。

４．１ BCCWJ を用いたト標識節(句)構文の分析（１）：内の関係

ト標識節(句)を係り受ける述部の動詞は、当然、引用助詞トの直後にくるとは限らない。が、条件を統制しかつ極力正確な方法で大量データを分析して、ト標識節(句)構文の用法の使用傾向が把握できるよう、まずは、フレーム喚起要素がト標識の直後に使われている構文を分析した。即ち、引用助詞トの直後に共起する動詞および事態性名詞（及び５語以内に共起する動詞および事態性名詞）の頻度表を作成し、手作業での分析を進めた。表２は、ChaKi による共起出力データに対して、人手で点検と用例意味解釈とコーディングを加え、各々のコーパスでの上位語から共格や接続助詞などの誤解析を取り除いた 100 語リストの一部(紙幅限度のため)を示した表である。（中には、同じ動詞が、多義に使われるだけでなく、誤解析の共格を含む場合があった。その場合は、参考まで共格カウントを残している。）

この方法で対象となる用法は（少なくとも複数出現し頻度上位にあがる語の場合は）殆どが「内の関係」の用法である。「外付けの関係」の用法も、述部のフレーム喚起語がトの直後にくる用例も観察収集しているが、それ以外の用例が重要なので、後述のとおり、「外付けの関係」は別方法で収集・分析を行った（５節)。

４．２ 白書コーパスと書籍文学コーパスにおける、ト標識節(句)構文の使用傾向

図２ー５に、白書コーパスと書籍文学コーパスでの、引用助詞トの直後にフレーム喚起語が共起する構文に関して、フレーム喚起語上位 100 語での用法の分布を示している。円グラフ２と３が動詞に関して、円グラフ４と５が事態性名詞に関してである。

白書では、P1「結果」用法が書籍文学に比べて多く、P1「移動」P2「尺度関連」P3「列挙」用法も相対的に多い。事態性名詞に関しても、白書で同様に、P1「結果」、P1「移動」、特に P2「尺度関連」が多いのに対して、書籍文学では、Q1「言動思考」用法に加え、Q1「感情」用法も相対的に多い。書籍に関しては、文学コーパスの分析のみ示しているが、他のサブコーパスを勘案しても、白書コーパスの特殊性を示唆していると思われる。

５．BCCWJ を用いたト標識節(句)構文の分析（２）：外付けの関係

４節までの分析手順で、主に「内の関係」のト標識節(句)構文の用法を対象とした。即ち、3 節の表 1 の類型における Q2「言語表現外付け引用事態型」以外の用法を考察した。しかし、(1)の Q2 で例示したように、日本語では、引用を表出するト標識節(句)が、言語表現を内容としない事態を表す述語とともに、頻繁に使用されることが、1980 年代から日本語学で指摘されてきた(寺村による指摘、藤田 1986, 山崎 1993, 藤井 2002)。

⑵ 「自分が少しでも橋渡しできれば」と取り組んできた。
⑶ 熱が出たので、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。
⑷ 千代は92年に……行方不明になってしまったのです。XXさんは「知人の責任問題になってはいけない」と千代のことは話題にしなかったそうです。

⑵-⑷の類の用例を観察すると、その多くが、「ーと思って」又は「ーと言いながら」と

補完して解釈が可能な意味関係を呈している。だが、単に「省略」と捉えるのでは現象が捉えきれない、補文としての名詞節的な用法をもつ引用ト標識節が、ある構文環境の中で、副詞節的な用法に拡張している、というのが著者の主張であった。同時に、述部の述語の主語が、ト標識引用部の内容を「そう思って」「そう言いながら」述部の事態・行為を行う、という解釈が可能である事実は重要である。従って、(i) 引用ト標識節(引用部)の主観者(主体者)が、主節の述部の主語と一致しており、(ii) 引用部は述部の行為・事態の主語(主体者)の動機・心情を表出している、というのが仮説であった(藤井 2002)。

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い。述部の述語（2.取り組む、3.行く/検査する、4.話題にしない）の喚起する意味フレームにおいて、ト標識引用節が述語によって要請されたフレーム構成要素ではない。この点で Q1 の「言語表現補語引用型」と異なる。従って、「ト標識引用部と述部との関係」を捉えるために、３節後半で提示した「内の関係」と「外付けの関係」との区別を、フレーム喚起語が喚起するフレームにおけるフレーム要素という観点で捉えると、「内の関係」の引用節は、フレーム喚起語が喚起するフレームの主要なフレーム要素[例えば Message, Content 等]として位置づけられるのに対して、Q2 の用法（「外付けの関係」）では、FrameNet において多くの副詞節(句)の位置づけに用いられている extra-thematic Frame Elements (FE) (Ruppenhofer et. al. 2006: 135-136) として、ト標識引用節を位置づけることができる。extra-thematic FE は、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描写するフレーム要素である。

しかし、３節で添えた通り、実際用例を並べてみると、extra-thematic FE とみなしてよいものと、中間的なもの（Q1-2）「言語表現引用同一事態型」に多々遭遇する。(中間的な Q1-2 等は、extra-thematic FE とするより non-core FE と捉えた方がよいと思える例が多い。)従って、用例一つ一つを綿密に意味解釈しつつ吟味することにより、extra-thematic FE とみなせるト標識引用節構文とそうでないものとの全体像を明確にすることができる。

上記仮説の検証も含め、このような問題の実証的分析を可能にしたのが、本特定領域の有難き成果 BCCWJ である。今回対象にした BCCWJ から、Q2「言語表現外付け引用事態型」の用例を、書籍文学 124 例、書籍言語 30 例；書籍技術工学 46 例;書籍自然科学 53 例;書籍哲学 119 例;書籍歴史 57 例； 書籍総記 24 例を(手作業で)採取した。白書コーパスからは、３回の読解では一例も見つからなかったが、見直し作業において、名詞修飾節内に埋め込まれた短い微妙な用例が一例のみみつかった。（この点においても、白書コーパスの特殊性が明確になった。）これまでのこれらの事例の分析では、上記仮説 (i) (ii) が支持されているが、この中でも述部との意味関係におけるグラデーションも明確になってきている。

６．BCCWJ を用いたト標識節(句)構文の分析（３）：フレームの広がりとフレーム間関係

２節表１の分類による分析に加え、述部のフレーム喚起語が喚起するフレームに関しても、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、動詞・事態性名詞それぞれ頻度上位 100 語（延べ 800 語）の用法で喚起されるフレームをコーディングした。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る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ボトムアップで、最上位フレームを同定した。図１は Q1 の用法に関してどの上位フレームに集約されるかを示した図である。他の用法に関しては図 1 に含まれない。例えば P2「尺度関連」は多くの場合 Change_position_on_a_scale (例: 高まる)というフレームを喚起する(図１から離れている)。

７．おわりに

引用のト標識節(句)が参与する構文を、BCCWJ を用いて、日本語学からの知見を鑑み照応しつつ、フレーム・フレーム要素・フレーム間関係の観点から分析した。日本語学でかねてから大きな問題として研究されてきた引用助詞トの用法をより良く理解したい、というのが最大の目的であった(ある)。と同時に、日本語学での研究手法や研究成果の蓄積とフレームネット(FN, JFN) の試みとの架け橋となる考察を深めるのが最大の願いであった。

さらに、どのタグ付きコーパス構築においても同様であるが、一貫した意味タグ付けをする際、水面下でどのような綿密な言語分析が必要であるか、また、一貫したカテゴリーを定めたタグ付けの生産物に漏れこぼれた言語現象にどのような捨て難い現象の観察があるかについて、共有・共感できるならこの上ないことである。

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、鈴木氏との議論が大変有益であった（特に鈴木 2005 で報告された「ほめる」「しかる」のフレーム要素に関する考察)。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)、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 代表 斎藤博昭氏、JFN 代表 小原京子氏、鈴木亮子氏、他）に記して感謝の意を表する。『BCCWJ2008 領域内版』のデータ処理・整理作業最終段階においては、(「日本語コーパス」科学研究補助金の支援をいただき）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、順に博士課程・修士課程)と平山仁美氏(同教養学部１年)に御協力・補助をいただいた。深謝する。また、内田諭氏と共同で行なってきていた FrameNet における副詞節記述に関する研究（内田&藤井 2007; 進行中）での考察に繋がる問題(extra-thematic frame elements の分析)も含まれており、特にフレーム間関係の考察に関する内田氏との議論に感謝する。

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 回年次大会発表論文集』言語処理学会, pp. 845-848.
内田諭、藤井聖子（2007).「FrameNet の枠組みを応用した接続語の意味記述-while の場合-」『言語処理学会第 13 回年次大会発表論文集』言語処理学会 pp. 851-854.
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)』, pp.613-616.
加藤陽子（1998). 「話し言葉における引用の「ト」の機能」『世界の日本語教育』8 国際交流基金日本語国際センター, pp.243-256.
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.
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」『日本認知言語学会論文集第 5 巻(JCLA 5)』, pp. 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(2002).「ト引用助詞節の副詞節的外付け用法に関する覚え書」『科学研究費補助金萌芽的研究(2)「対照意味論・対照語用論的研究の方法と理論の構築-日英語の談話における語用標識の研究-」研究成果報告書』.
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」『日本認知言語学会論文集第 5 巻 (JCLA 5)』, pp. 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.
Fontenelle, Thierry. (Ed.). (2003). Special Issue: FrameNet and Frame Semantics. International Journal of Lexicography. Vol.16, Special Issue 3, Oxford, Oxford University Press.
Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Chris R. Johnson, and Jan Scheffczyk. (2006). *FrameNet II: Extended Theory and Practice*.(http://framenet.icsi.berkeley.edu/book/book.pdf)
Suzuki, Ryoko. (1999). Multifunctionality: the developmental path of the quotative *tte* in Japanese. In B. A. Fox, D. Jurafsky, & L. A. Michaelis (eds.), *Cognition and Function in Language*. Stanford: CSLI Publications. 51-64.
Suzuki, Satoko. (1996). The discourse function of the quotation marker *tte* in conversational Japanese. *BLS* , 22, 387-393.

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/

エントロピーと冗長度を指標とした語彙的･統語的複合動詞の比較研究

玉岡賀津雄（リーダビリティー班分担者: 麗澤大学外国語学部）†

A Comparative Study on Lexical and Syntactic Compound Words Using Two Informatics Indexes of Entropy and Redundancy

Katsuo Tamaoka (Department of Foreign Studies, Reitaku University)

１．共起頻度の分析のための指標

クロード・シャノン(Claude Shannon)は，『通信の数学理論』(1948)という論文において，「エントロピー(entropy)」と「冗長度(redundancy)」という二つの概念を発表した。これらは，情報量の尺度である。エントロピーは，あいまいさや乱雑さの増減を示す指標である。言語研究においては，表現の種類とその使用頻度に基づいて一つの値を算出し，不規則性を示すことができる。また，冗長度は，エントロピーとエントロピー最大値(最も不規則な状態)を利用して得られる無駄の程度(規則性とも考えられる)を示す指標である。エントロピーと冗長度の尺度を組み合わせることで，ある表現の多様性と規則性とを簡単な数値で表すことができるのである。

近年，大規模コーパスが整備され，「共起頻度」が容易に算出できるようになってきた。これに伴い，複数の基準で選択された表現の「異なり頻度」や「延べ頻度」を解析する手法が求められるようになってきている。このような，それ自体はノンパラメトリック・データである共起頻度に対して，通信(あるいは情報)の数学理論で扱われるエントロピーと冗長度を適用させパラメトリック・データの指標に変換することによって，様々な多変量解析を行うことができるようになる。また，これら二つの指標をもとに各表現を二次元上でプロッティングし，個々の共起表現パターンを記述的に考察することも可能である。さらに，新聞，小説，雑誌など，異なるタイプとサイズのコーパスから得られた共起頻度のパターンについて，エントロピーと冗長度の指標に基づいて，各コーパスの特徴を比較検討することもできる。本章では，エントロピーと冗長度の算出方法とそれらの指標を用いた多変量解析の例を紹介する。

２．語彙的・統語的複合動詞の統語構造

日本語では，動詞を二つ組み合わせることで複合動詞を作ることができる。影山(1993, 1999)は，日本語の複合動詞を２種類に分けている。一つは，語彙的複合動詞で，初めにくる動詞(V_1)と次にくる動詞(V_2)の組み合わせに「語彙的な習慣化が見られる」（影山, 1999, p. 189）としている。語彙的複合動詞には，二つの動詞の間に「語の形態的親密性」を示すサエやモを挿入することができず，一つの語となっていると指摘している。もう一つは統語的複合動詞である。これについて影山（1999）は，その名の示す通り「統語的な構造」に

† ktamaoka@gc4.so-net.ne.jp

由来し，補文構造という形で捉えられるとしている(pp. 189-190)。たとえば，「彼は昼食を食べ始めた。」と「子供は手紙を投函し忘れた。」における補文構造は，「彼は[昼食を食べ]始めた。」と「子供は[手紙を投函し]忘れた。」という「V1することを(が)V2」の部分を指す。この構造を元にして，「食べ」を「始める」に，「投函し」を「忘れる」に結合させると，表面的には，「食べ始める」と「投函し終わる」という複合動詞になると説明している。つまり，統語的複合動詞では，二つの動詞は基本的に別々の語であるという考え方である。

両者を分かりやすく区別するために，統語構造から違いを説明する。図 1 の(i)に示すように，語彙的複合動詞は，先にくる動詞(V_1)である「飲む」が，次にくる動詞(V_2)の「比べる」と結合して「飲み比べる」という一つの動詞($V_1 + V_2 = V$)を作ると考える。この複合動詞に対して，「コーヒーを」という対格の名詞句(NP, noun phrase)が結びついて動詞句(VP, verb phrase)を作るという構造を持つ。それに対して統語的複合動詞は，図 1 の(ii)に示すように，「コーヒーを」という名詞句(NP)が直接に「飲む」の動詞(V_1)と結合して動詞句(VP_1)を作る。そして，この動詞句が後にくる「終える」の動詞(V_2)に結びつき，さらなる動詞句(VP_2)を作るという構造と考える。このように，両者の複合動詞は，統語的に異なる構造を持つと考えることができる。語彙的複合動詞の場合は，「コーヒー」というものを複合動詞が受けて「飲み比べる」ものと解釈され，「飲む」と「比べる」の二つの動詞を切り離すことができない。これに対して，統語的複合動詞の場合は，「コーヒーを飲む」という動詞句で示す行為を「終える」という動詞で受けることになる。

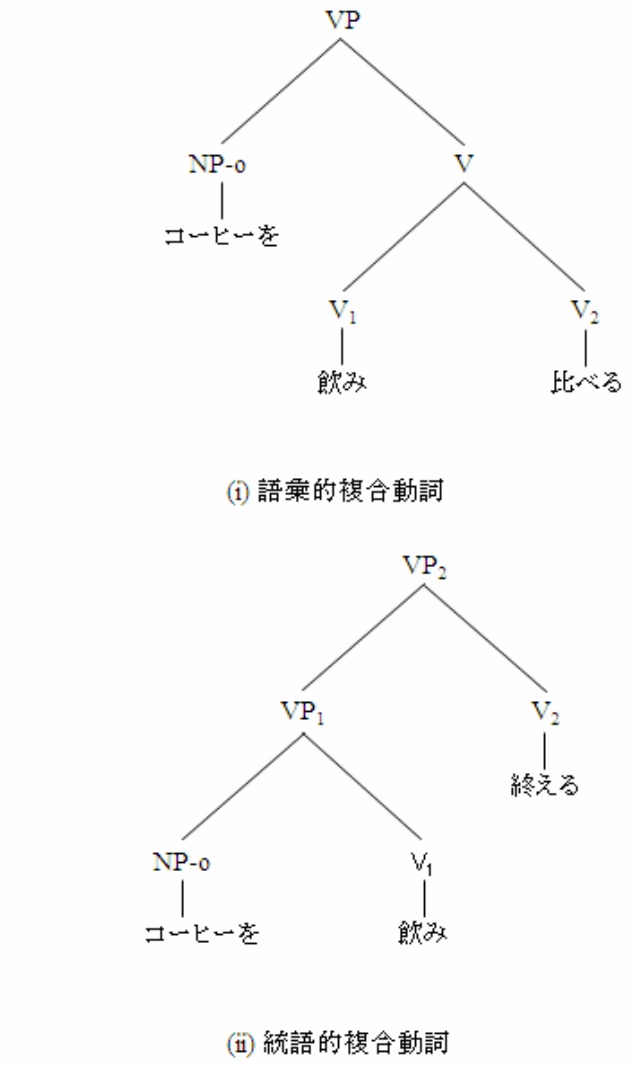


図1 語彙的・統語的複合動詞の統語的構造

注: NP-oは、対格の名詞句を示し、Vは動詞、VPは動詞句を示す.

つまり，コーヒーを飲むことを「終える」のであり，統語的複合動詞においては，二つの動詞が別々に機能している。

３．語彙的・統語的複合動詞に関する二つの仮説と個々の記述

語彙的/統語的複合動詞に関する二つの仮説について，新聞と小説のコーパスから得られる共起頻度を用いて実証し，さらに個々複合動詞の特徴を記述的に考察する．

まず，複合動詞の種類に関する仮説である。語彙的複合動詞は，「飲み比べる」というように二つの動詞が一つの語彙的な単位を成しており，慣用句のように強く結合していると考えられる。したがって，これら二つの動詞の組み合わせは限定的であり，それほど多様な動詞が組み合わされることはないであろうと予想される。一方，統語的複合動詞は，図１の例の「終える」(V_2)から分かるように，どのような行為も「終える」ことができるものであれば何にでも結合できるので，多様な第１動詞(V_1)を受けることができると考えられる。これらの構造の異なる２種類の複合動詞について，通信の数学理論であるエントロピーと冗長度の指標の違いから予測してみる。まず，語彙的複合動詞は，ある特定の二つの動詞の共起パターンは偏っていると考えられるので，エントロピーは低く，冗長度が高いであろう。それに対して，統語的複合動詞は，多様な動詞が結合して共起すると考えられるので，エントロピーが高く，冗長度は低いと予想される。これが第１の仮説である。

次に，コーパスの種類に関する仮説である。新聞は，複数の新聞記者が一般大衆に情報を伝達するために，簡潔で分かりやすい一定の表現スタイルが採られる。一方，小説は，特定の作家の個性に応じた多様な表現が駆使されている。両者は，書き手の意図と目的において大きく異なっている。仮に，語彙的/統語的複合動詞が一般的な統語構造を持っているといえるなら，語彙的・統語的複合動詞は，新聞か小説かに関係なく類似した共起パターンを示すことが予想される。これが第２の仮説である。新聞のコーパスとして，1991年から1994年までの４年間に刊行された毎日新聞を利用した。その延べ総語数は88,454,573語である。一方，小説のコーパスとしては，青空文庫コーパスに収録されている文学作品のうち，吉行エイスケ著の『地図に出てくる男女』，新美南吉著の『ごん狐』など現代語で書かれているものを選んでコーパス化したものを利用した。延べ総語数は8,370,720語である。小説のコーパスとのサイズのバランスをとって両コーパスを比較しやすくするために，４年間分に限定し，小説のコーパスの約10倍の大きさにした。

最後に，複合動詞の個別性とコーパスの種類に関係する記述を試みた。たとえば，「込む」(V_2)は，「流れ込む」，「射し込む」，「吹き込む」など多様な用いられ方をする。これらは，水が「流れ込む」，光が「射し込む」，風が「吹き込む」というように，意味的に二つの動詞が一つの動詞として一塊のものとなっていると解釈されるので，語彙的複合動詞である。「込む」(V_2)は，ちょうど英語の前置詞の‘into’のような意味が付加され多様な表現と結びつきやすいので，複合動詞としても多様な動詞(V_1)と結合されやすいと考えられる。また，「得る」(V_2)は，「動かし得る」，「保ち得る」，「考え得る」など多様な動詞(V_1)と結合して用いられる。これらの複合動詞は，「岩を動かし得る」であれば，「岩を動

かす」(VP_1)ことを「得る」(VP_2)という補文構造を成す。同様に，「平静を保ち得る」という表現であれば，「平静を保つ」ことが「できる」のであり，やはり「得る」(V_2)は単独でその前の動詞句に結びついている。つまり,これらは統語的複合動詞である。ちょうど英語の‘can’や‘be possible to’のような解釈ができ，「岩を動かす」ことが「できる」または「可能である」という意味になる。このように，第２動詞を基準とする複合動詞群について，新聞または小説それぞれに特徴的に見られるものが幾つか存在するであろうと思われる。

４．エントロピーと冗長度の公式と計算法

シャノンによると，エントロピーは以下の式で計算される。

$$H=-\sum_{j=1}^{j} p_j \log_2 p_j$$

複合動詞については，まず第２番目にくる動詞(V_2)を基準にしてエントロピーを計算する。たとえば，「歩く（V_2）」を含んだ統語的複合動詞を構成する第１動詞(V_1)を考えてみると，毎日新聞の 1991 年から 1994 年までの４年間の記事においては, 18 種類の第１動詞(V_1)と結合する。これが異なり頻度である。

これら 18 種類の動詞と「歩く」とが結合する総頻度は 44 回であるが，これが延べ頻度である。最も多いのが「売り歩く」で７回，次に「尋ね歩く」で６回であった。上記の公式の p_j に相当するのが，「売り歩く」など個別の複合動詞の頻度が「歩く」(V_2)を基準につくられる全複合動詞の総頻度に占める割合である。具体的には，「売り歩く」の頻度の７を，総頻度の 44 で割った 0.159 が p_j の値となる。$\log_2 p_j$ は，「売り歩く」の場合は $\log_2 0.159$ であり，-2.652 の値が得られる。次に，$p_j \log_2 p_j$ として，$0.159\times(-2.652)=-0.422$ と計算される。「歩く」(V_2)と結合してつくられる複合動詞は 18 種類あるので，同様の計算を個々の複合動詞について 18 回行い(∑の j の部分)，それらをすべて積算して，-1 を掛けると，「歩く」(V_2)を含む複合動詞について 3.780 というエントロピー値が算出される。

エントロピーとともにシャノンが提示した有名な概念は，冗長度である。冗長度とは，無駄の程度を表す指標である。ただし，Tamaoka et al. (2004)の研究の場合は，これを，二つの動詞から成る複合動詞の組み合わせと共起頻度の偏りに相当するものであり，同じような二つの動詞の組み合わせが繰り返し使われる度合いを示していると捉えている。シャノンによると，冗長度は，以下の公式で得られる。

$$R=(1-H/H_{\max})\times 100\ (\%)$$

H はエントロピーであり, $Hmax$ はエントロピー最大値を意味する。エントロピー最大とは，すべてが等しい確率で生起する場合である。つまり，いずれが起こっても不思議ではない混沌としたまったくの無秩序の状態を意味する。ある第２動詞を基準とした複合動詞の種類，すなわち異なり頻度を J とすると，以下の式でエントロピーの最大値が得られる。

$$H_{max}=\log_2 J$$

表1 毎日新聞の語彙的/統語的複合動詞(V_1+V_2)の頻度、エントロピー、冗長度

#	V_2の種類	V_2の延べ頻度	V_1の延べ頻度	V_1の異なり頻度	V_1とV_2の延べ頻度	エントロピー	冗長度 (%)
(1) 語彙的複合動詞							
1	込む	295	1,098,690	81	278	5.76	9.10
2	あげる	2,914	45,880	57	174	5.30	9.20
3	切れる	543	64,292	44	119	4.66	14.73
4	取る	5,947	53,493	33	94	4.39	13.04
5	回る	1,021	17,989	27	61	4.27	10.12
6	つく	2,354	8,906	19	45	3.81	10.34
7	歩く	1,554	30,414	18	44	3.78	9.35
8	上がる	1,808	40,283	31	229	3.69	25.56
9	継ぐ	355	20,382	15	33	3.68	5.88
10	死ぬ	1,376	16,929	13	14	3.66	0.97
11	たてる	632	434,024	16	55	3.66	8.46
12	かかる	4,764	61,171	14	25	3.62	4.83
13	替える	135	14,613	15	40	3.58	8.44
14	いれる	2,114	13,410	13	19	3.58	3.35
15	刺す	333	21,434	11	12	3.42	1.19
16	返す	609	9,679	23	45	3.36	25.72
17	出る	7,153	8,513	18	56	3.21	22.94
18	こめる	110	27,164	12	23	3.13	12.63
19	落ちる	755	12,213	11	33	3.07	11.26
20	落とす	599	4,939	11	20	3.05	11.94
21	おろす	253	7,527	10	58	2.88	13.29
22	きる	1,529	136,575	60	496	2.77	53.04
23	入る	6,425	19,260	10	25	2.76	17.06
24	飛ばす	244	1,056	7	13	2.57	8.62
25	つける	460	20,528	9	34	2.51	20.80
26	倒す	140	805	7	14	2.41	14.02
27	殺す	444	1,614	6	14	2.35	8.98
28	起こす	1,392	3,506	6	19	2.07	19.76
29	渡る	573	4,639	7	41	2.00	28.81
30	おりる	431	10,443	7	27	1.68	40.21
31	のぼる	3,417	13,187	4	11	1.68	16.16
32	返る	91	1,487	5	23	1.61	30.68
33	広げる	856	7,812	4	12	1.42	29.09
34	渡す	491	48,068	3	10	1.36	14.13
35	くだる	124	14,816	3	15	1.27	19.69
36	知る	1,993	2,830	5	24	1.14	50.96
37	合わせる	1,109	38,142	15	61	0.88	77.38
(2) 統語的複合動詞							
1	続ける	5,519	539,169	261	1425	6.73	16.21
2	始める	2,983	1,379,861	207	657	6.50	15.55
3	あう	2,302	295,787	170	873	6.16	16.87
4	過ぎる	3,777	368,408	130	515	5.71	18.74
5	まくる	86	708,256	32	66	4.56	8.91
6	終わる	1,884	51,545	31	56	4.50	9.18
7	終える	503	850,402	24	37	4.31	5.90
8	尽くす	687	843,270	26	89	3.72	20.86
9	ぬく	575	724,584	23	131	3.11	31.33
10	かねる	328	1,062,433	18	108	2.82	32.27
11	得る	4,478	362,068	100	1601	0.08	98.77

注1: 88種類の統語的複合動詞と21種類の語彙的複合動詞から延べ頻度が10回以上の48種類を選んだ.
注2: 1991年から1994年までの毎日新聞の記事は、88,454,573語の延べ頻度からなる.

表2 青空文庫の語彙的/統語的複合動詞(V_1+V_2)の頻度、エントロピー、冗長度

#	V_2の種類	V_2の延べ頻度	V_1の延べ頻度	V_1の異なり頻度	V_1とV_2の延べ頻度	エントロピー	冗長度 (%)
(1) 語彙的複合動詞							
1	あげる	572	25,037	48	92	5.13	8.17
2	かかる	603	28,586	46	90	5.08	8.01
3	つく	658	35,298	41	107	4.78	10.75
4	たてる	218	7,413	28	57	4.48	6.77
5	出る	1,980	6,355	31	61	4.46	9.95
6	取る	823	4,436	27	45	4.43	6.92
7	きる	311	109,842	36	69	4.42	14.52
8	のぼる	460	33,216	37	142	4.39	15.69
9	回る	162	24,004	29	79	4.31	11.24
10	歩く	368	13,012	23	34	4.26	5.91
11	刺す	150	21,216	20	25	4.21	2.51
12	上がる	182	45,880	62	174	4.19	6.10
13	返す	193	9,719	20	39	4.16	5.28
14	入る	678	20,161	23	55	4.15	8.28
15	殺す	358	48,145	19	36	4.04	4.95
16	落ちる	210	6,852	15	31	3.55	9.24
17	つける	86	20,961	15	30	3.51	10.24
18	合わせる	50	30,796	12	15	3.46	3.58
19	倒す	34	1,603	12	16	3.45	3.69
20	いれる	287	2,702	12	17	3.29	8.31
21	狂う	63	2,581	11	20	3.22	6.87
22	おろす	67	14,421	10	20	3.18	4.15
23	破る	74	3,688	10	14	3.18	4.21
24	くだす	102	21,056	9	10	3.12	1.51
25	渡る	157	8,438	12	30	3.11	13.39
26	返る	67	7,444	9	42	2.62	17.50
27	込む	281	110,771	81	220	2.13	66.48
28	起こす	53	5,335	5	15	1.93	16.87
29	消す	47	706	3	13	1.55	2.30
(2) 統語的複合動詞							
1	始める	294	148,945	100	178	6.07	8.67
2	あう	268	133,453	85	173	5.72	10.71
3	過ぎる	617	67,033	66	132	5.44	10.08
4	続ける	146	48,805	50	85	5.28	6.51
5	得る	551	159,594	98	297	5.12	22.56
6	かねる	84	121,923	44	79	5.02	7.99
7	尽くす	20	80,821	10	11	3.28	1.33
8	終わる	64	17,183	12	24	2.98	16.89

注1: 88種類の統語的複合動詞と21種類の語彙的複合動詞から延べ頻度が10回以上の37種類を選んだ.
注2: 本研究で使用した青空文庫コーパスは、8,370,720語の延べ頻度からなる.

例えば，「歩く」(V_2)が作る複合動詞は，総頻度が44回で18種類の複合動詞を作るので，異なり頻度のJは18である。エントロピーの最大値は，H=$\log_2$18で4.170となる。この数値は，18種類の複合動詞がすべて等しく生起する場合のエントロピーである。つまり，どの「歩く」(V_2)から作られる複合動詞も同数回だけ出現するので，もっとも規則性の無い状態であ

るというわけである。さて，冗長度は，得られたエントロピーをそのエントロピー最大値で割り，その数値を１から引いて100倍して，パーセントで示したものである。「歩く」(V_2)についていえば，エントロピー値が3.780であるので，これをエントロピー最大値の4.170で割って，１から引いた値である0.09348に100を掛けると，9.348%となる。これが冗長度である。

表１は，毎日新聞４年分のコーパスにおける複合動詞をつくる第２動詞(V_2)のエントロピーと冗長度を示したものである。統語的複合動詞をつくる第２動詞の88種類と語彙的複合動詞をつくる第２動詞の21種類(合計109種類)のうち，複合動詞の総頻度(延べ頻度)が10回以上である48種類を選んで計算したものである。これは，延べ頻度が10回以下では，エントロピーと冗長度の指標が信頼しうる値とはいえないために設けた基準である。表２は，同様にして算出した青空文庫の小説のコーパスにおいて複合動詞の延べ頻度が10回以上であった37種類の結果である。

一見すると，冗長度は個々のエントロピーがエントロピー最大値に占める割合を１から引いた値であるので，エントロピーが高くなればその分冗長度は低くなるという両変数の逆相関を考えてしまいそうである。表１の毎日新聞４年分の新聞のコーパスから得られた第２動詞(V_2)からつくられる48種類の複合動詞については，エントロピーと冗長度の間のピアソンの相関係数は-0.549(p<.001)であり，表２の青空文庫の小説のコーパスの37種類では，-0.217(p=0.197, *n.s.*)でしかない。このように，両指標の相関係数が-1となり完全な逆相関を示すことはなく，両変数が独立した指標であることが分かる。

５．仮説１の証明—語彙的/統語的複合動詞のエントロピーと冗長度による比較

仮説１は，語彙的複合動詞と統語的複合動詞の特徴の違いである。本研究でエントロピーと冗長度の指標を用いる最大の魅力は，第２動詞を基準として見られる複合動詞の特徴を，語彙的複合動詞と統語的複合動詞との間で直接に比較できることである。語彙的/統語的複合動詞の２つのグループについての一元配置の分散分析を，新聞と小説のコーパスとで個別に行った。その結果，新聞のコーパスについては，統語的複合動詞のエントロピー(M=4.138, SD=1.949)の方が，語彙的複合動詞のエントロピー(M=2.974, SD=1.160)よりも有意に高いことが示された[F(1,46)=8.946, p<.01]。しかし，冗長度については，統語的複合動詞(M=24.963%, SD=25.880%)と語彙的複合動詞(M=18.425%, SD=15.721%)の間に有意な違いはなかった[F(1,46)=1.069, p=.307, *n.s.*]。このエントロピーの分析結果は，統語的複合動詞のV_2が，語彙的複合動詞にV_2に比べて多様なV_1と結合して複合動詞を作っていることを示している。冗長度については，両複合動詞を作る２つの動詞の結合頻度のパターンの規則性に違いがないことが示された。小説のコーパスについても同じ一元配置の分散分析を行った。その結果，統語的複合動詞のエントロピー(M=4.864, SD=1.124)の方が語彙的複合動詞のエントロピー(M=3.717, SD=0.901)よりも有意に高いことが示された[F(1,35)=9.144, p<.01]。しかし，冗長度については，統語的複合動詞(M=10.593%, SD=6.504%)と語彙的複合動詞(M=10.117%, SD=11.679%)の間に有意な違いはなかった[F(1,35)=0.012, p=.913, *n.s.*]。新聞の

コーパスと同様に小説のコーパスでも，エントロピーの分析結果は，統語的複合動詞のV_2が，語彙的複合動詞にV_2に比べて多様なV_1と結合して複合動詞をつくっていることを示した。エントロピーの指標は，仮説１の結果を支持している。

６．仮説２の証明—語彙的/統語的複合動詞の共起パターンについての新聞と小説との違い

第２の仮説は，新聞のコーパスと小説のコーパスとで，V_2を基にしてV_1と結びつく複合動詞の特性に違いがあるかどうかである。まず，毎日新聞の４年間のコーパスは延べ総語数が88,454,573語で，青空文庫コーパスは延べ総語数が8,370,720語である。両コーパスのサイズには10.57倍の違いがある。しかし，表１と表２で示したエントロピーと冗長度は，V_2を基準とした多様なV_1との結合による共起頻度のパターンを示したデータサイズに左右されない指標であるので，コーパスのサイズに関係なく，二つのグループの複合動詞のパターンを比較することができる。新聞と小説のコーパスに共通して出現する共起頻度の述べ頻度が10回以上の複合動詞は，34種類である。エントロピーは正規分布しているので t 検定を行うこともできるが，仮説１との一貫性を考えて，これら34種類の複合動詞について反復測定による分散分析を行う。反復測定である理由は，同じ複合動詞が，新聞のコーパスと小説のコーパスで算出されていると考えるからである。

エントロピーについて新聞と小説のコーパスの違いを比較してみると，新聞(M=3.422, SD=1.581)の方が，小説(M=4.080, SD=0.991)よりも有意に低かった[$F(1,33)=6.898$, $p<.05$]。これは，新聞よりも小説のコーパスの方が，V_2を基準としてV_1と結合して作られる複合動詞のパターンが，多様性に富んでいることを示している。冗長度につても同じ分析をした結果，新聞(M=20.958%, SD=19.864%)の方が，小説(M=10.727%, SD=11.006%)よりも有意に高かった[$F(1,33)=7.358$, $p<.05$]。これは，新聞の方が小説よりも複合動詞の２つの動詞の結合関係が規則的であったことを示している。以上のように，エントロピーと冗長度の両方において，新聞と小説のコーパスの複合動詞の共起頻度パターンに有意な違いが見られ，新聞のコーパスの複合動詞は，小説のコーパスほどの多様性は無く，より規則的なパターンを示すことが分かった。当然ながら，新聞は，一般大衆に情報を伝達するために，簡潔で分かりやすい一定の表現スタイルを採っているので，小説よりもエントロピーが小さく，冗長度が大きくなったのであろう。一方，小説は，作家の個性に応じた多様な表現が現れるために，新聞よりもエントロピーが大きく，冗長度が小さくなったのであろう。

エントロピーと冗長度は，両者のコーパスの書き手の意図と目的を反映した結果を反映しており，両指標がコーパスの違いを比較するのに有効な方法であることが分かる。以上のように，エントロピーと冗長度の指標は，仮説２を支持していた。

７．個々の複合動詞の記述—新聞と小説それぞれのコーパスに特徴的な語彙的/統語的複合動詞

仮説２で，新聞と小説で複合動詞の共起頻度のパターンに違いがあることが分かった。そこで，具体的にどの複合動詞が両コーパスで顕著な違いを呈しているかを考察する。こ

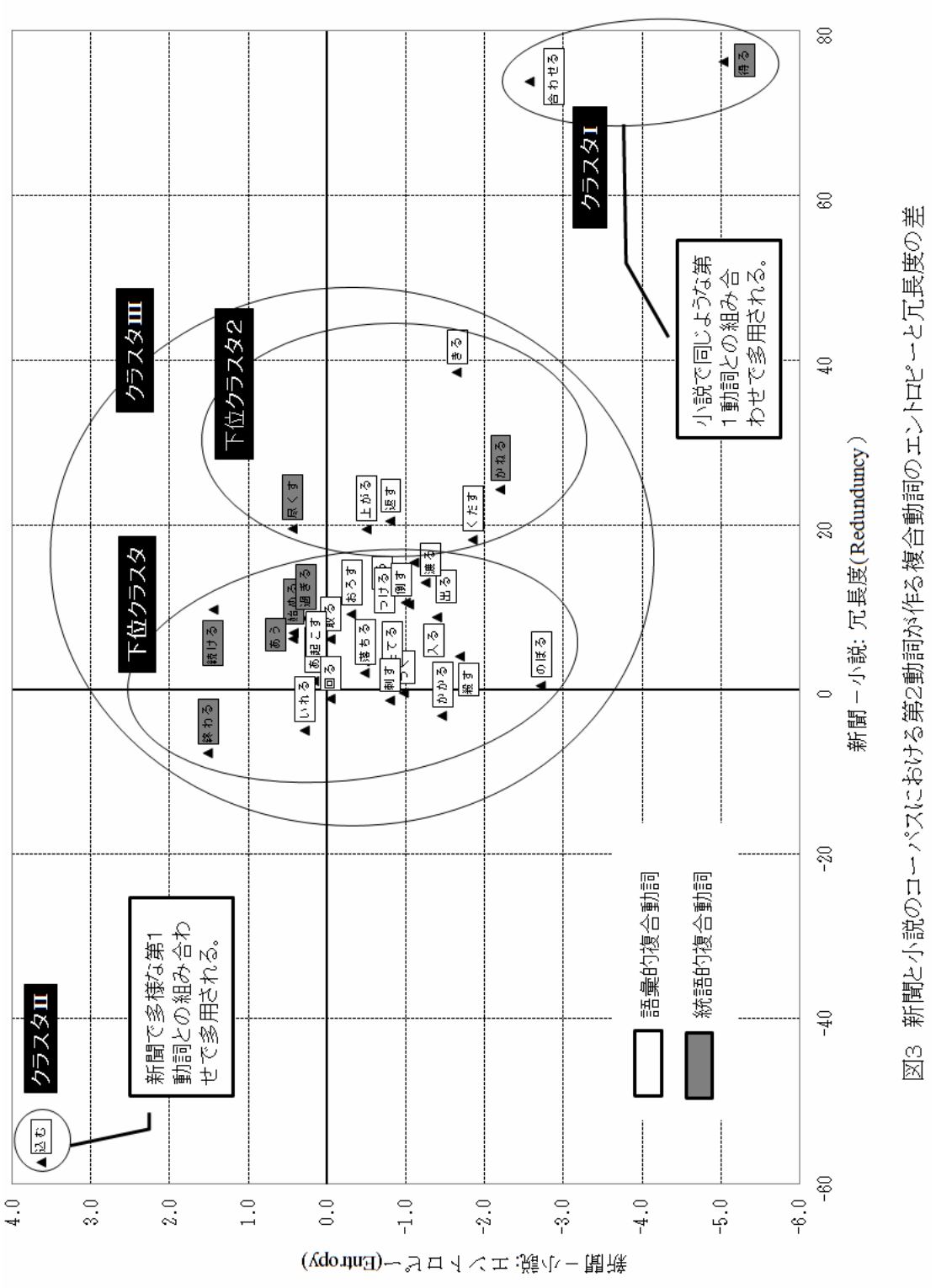


図3　新聞と小説のコーパスにおける第2動詞が作る複合動詞のエントロピーと冗長度の差

こでは，個々の複合動詞について，エントロピーと冗長度を用いて記述的にそれらの特徴

を描いてみる。個々の複合動詞のエントロピーと冗長度の両コーパス間の差を算出する（いずれか一方のエントロピーから他方のエントロピーを差し引く）と，絶対値が大きい複合動詞ほど新聞と小説での用いられ方に顕著な違いがあると判断することができる。そこで，両コーパスに共通する34種類の複合動詞のエントロピーと冗長度について，新聞のコーパスで得られたものから小説のコーパスで得られたものを差し引いて，語彙/統語的複合動詞の第二動詞V_2を二次元上にプロッティングした（図３）。さらに，34種類の複合動詞について共起パターンの観点から分類するために，階層クラスタ分析を行った。原データ間の距離は平方ユークリッド，クラスタ間の距離はウォード法をとった(分析には，SPSS15.0のBase Systemを使用した)。この方法では，複合動詞を区別するのに25ポイントが最大となる。13ポイントを分類の基準とすると，まず25ポイントでクラスタIに分類されたのが，語彙的複合動詞の「合わせる」(エントロピーの差が-2.573，冗長度の差が73.802%)と統語的複合動詞の「得る」(エントロピーの差が-5.040，冗長度の差が76.208%)であった。図３では，第４象限の右下に位置する。クラスタIIに分類されたのは，語彙的複合動詞の「込む」(エントロピーの差が3.638，冗長度の差が-57.375%)のみであった。第２象限の左上に位置している。残りは，クラスタIIIと分類された。クラスタIIIについて，さらに６ポイントで２つのクラスタに分類するなら，図３のように下位クラスタ１と下位クラスタ２を想定することもできる。しかし，一般にクラスタ化の妥当な基準は10ポイント以上であるので，これらを独立したクラスタとみなすよりは，クラスタIIIの下位クラスタと扱うほうがよいであろう。

図３のプロッティングしたグラフにクラスタ分析の結果を重ね合わせると，新聞と小説の違いが分かり易い。「合わせる」「得る」「込む」の３つのV_2がつくる複合動詞は，両コーパス間で顕著な違いを示すことが分かる。「得る」の新聞でのエントロピーは0.08と小さく，冗長度は98.77%と大きい。一方，小説では反対に，エントロピーが5.12と大きく，冗長度は22.56%と小さかった。つまり，「得る」は，新聞では，「考え得る」や「あり得る」などの特定の少数の組み合わせが頻繁に出現するといった規則的な共起パターンをみせるが，小説では，より多様なV_1動詞と結合してさまざまな複合動詞をつくることが分かる。「合わせる」も同様のパターンである。一方，「込む」は，小説より新聞のほうが，エントロピーが大きく，冗長度が小さい。したがって，新聞のほうが小説よりも多様なV_1と結合されることが分かる。V_2が作る３つの複合動詞において，両コーパスの共起頻度パターンに顕著な違いが示された。

引用文献

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423 (Part I) and 623-656 (Part II).

用例間類似度測定のための属性重みの推定

新納 浩幸 (クラスタリング班班長: 茨城大学 工学部) [1]
佐々木 稔 (クラスタリング班分担者: 茨城大学 工学部)

Estimation of Feature Weight for Measuring Similarity between Example Sentences

Hiroyuki Shinnou (Ibaraki University, Faculty of Engineering)
Minoru Sasaki (Ibaraki University, Faculty of Engineering)

1 はじめに

ある単語の用例を集め、それら用例をその単語の語義に基づいて分類するタスクに取り組んでいる。本論文では、このタスクの本質的課題となる用例間類似度の測定について論じる。

語義別の用例は本格的な意味解析にとって有用である。例えば、語義別の用例を訓練データとして利用することで語義の曖昧性を解消する分類器を学習することできる (Masaki Murata and Masao Utiyama and Kiyotaka Uchimoto and Qing Ma and Hitoshi Isahara, 2001)。また動詞の格フレームの自動構築 (Philip, 1992) は、動詞の語義別の用例があれば容易に行える。またシソーラスの自動構築 (Hindle, 1990) においては、通常、名詞の多義性は無視されるが、語義別の用例があれば、語義を考慮したシソーラスの作成も容易である。辞書の編纂、語学学習においても、語義別の用例は有用であろう。

ある単語 w の語義別の用例を収集するには、単語 w を含む用例をコーパスから抽出し、その用例中の単語 w の語義を識別できればよい。つまり語義別用例の収集は、語義の曖昧性解消のタスクとして扱える。そのために（半）教師有り学習を用いるアプローチによって解決できそうであるが、以下に示す２つの問題から、そのアプローチによる解決は困難である。第１の問題は、訓練データの作成コストである。教師有り学習では大量の訓練データを必要とし、対象となる単語が多い場合、その作成コストが大きすぎる。第２の問題は、語義の設定である。（半）教師有り学習で本タスクに挑む場合、単語 w の語義を予め設定しておかなければならない。単語 w が与えられたときに、w の語義を内省により列挙することは困難である。例えば、語義の粒度を均一に保つことは難しいし、マイナーな語義を見落としてしまう危険性もある。そのためにこのタスクに対しては用例を語義に基づいてクラスタリングするアプローチが妥当である。

ただし用例のクラスタリグを行うためには用例間の類似度（あるいは距離）を測る必要がある。用例間の類似度の設定方法として確立された手法は存在せず、これまでアドホックに対処されていた。そのためこのタスクの精度は、実質、クラスタリング手法ではなく、用例間の類似度の適切さに大きく依存している。

本論文では用例間類似度を線型モデルで表し、そのパラメータ（属性重み）を推定することを試みる。この際、用例間類似度の値は、手作業であっても設定することができないため、推定のための訓練データを構築できないという問題がある。この問題に対して、ここでは以下のような対策をとった。まず線型モデルのパラメータを経験的な値で与え、仮のモデルを作成する。次に語義識別タスクに対する訓練データを用いて、用例対が同じクラスに属する場合は、仮のモデルから類似度を与え、同じクラスに属さない場合は、類似度を 0 とすることで訓練データを作成する。この訓練データをもとにパラメータの推定を行う。

実験では SENSEVAL2 の日本語辞書タスク (白井清昭, 2003) で用いられた名詞 50 単語を対象とした。SENSEVAL2 で提供されたそれら単語に対する訓練データを用例のセットとし、(1) 単純な線型モデルによる用例間類似度、(2) 経験的なパラメータ値を用いた用例間類似度、(3) 本手法により得られたパラメータ値を用いた用例間類似度の 3 つを用いてクラスタリングを行った。エントロピー

[1] shinnou@mx.ibaraki.ac.jp

の平均値によるクラスタリングの評価を行ったところ、(3)、(2)、(1) の順でよい結果を得ることができた。

2 用例に対する素性リスト

クラスタリグを行う場合、対象のデータが実数値ベクトルで表現されている必要はない。データ間の類似度（あるいは距離）が設定されていれば十分である。ここでは用例間の類似度を定義するために、用例を素性リストで表現する。素性としてここでは以下のものを利用する。

ee1	直前の２単語（表記）
ee2	直後の２単語（表記）
e1	直前の単語
e2	直後の単語
e3	前方と後方の内容語それぞれ２つまで
e4	e3 の分類語彙表の番号

例を示す。対象の単語を「記録」として、以下の用例を考える（形態素解析され各単語は原型に戻されているとする）。

過去/最高/を/記録/する/た/。

この場合、「記録」の直前、直後の２単語の表記は「最高を」と「した」なので、‘ee1=最高を’，‘ee2=した’となる。また直前、直後の単語は「を」と「する」なので、‘e1=を’，‘e2=する’となる。次に、「記録」の前方の内容語は「過去」、「最高」なので、ここから「記録」に近い順に 2 つとり、‘e3=過去’、‘e3=最高’が作られる。また「記録」の後方の内容語は「する」だけであり、‘e3=する’が作られる。次に「最高」の分類語彙表の番号を調べると、 3.1920_4 である。ここでは分類語彙表の 4 桁目と 5 桁目までの数値をとることにした。つまり‘e3=最高’に対しては、‘e4=3192’ と ‘e4=31920’ が作られる。同様に「過去」の分類語彙表の番号 1.1642_1 から‘e4=1164’ と ‘e4=11642’ が作られる。次は「する」の分類語彙表を調べるはずだが、ここでは平仮名だけで構成される単語の場合、分類語彙表の番号を調べないことにした。これは平仮名だけで構成される単語は多義性が高く、無意味な素性が増えるので，その問題を避けたためである。もしも分類語彙表上で多義になっていた場合には、それぞれの番号に対して並列にすべての素性を作成する。

結果として、上記の用例に対して以下の 11 個の素性を要素とする素性リストが得られる。

```
( ee1=最高を，ee2=した，e1=を，e2=する，e3=最高， e3=過去，e3=する，
  e4=3192， e4=31920，e4=1164， e4=11642 )
```

3 用例間の類似度

3.1 類似度のモデル

用例 s_1 と s_2 の類似度 sim を以下で定義した。

$$sim(s_1, s_2) = a \cdot M(ee1) + b \cdot M(ee2) + c \cdot M(e1) + d \cdot M(e2) + e \cdot M(e3) + f \cdot M(e4)$$

ここで $M(x)$ は s_1 と s_2 の素性リスト中の素性 x の一致数を表す。上記の式は類似度測定の一種のモデルであり、これをここでは**線形モデル**と呼ぶことにする。線形モデルでは素性同士がどれくらい一致しているかを調べ、素性毎に重みをつけている形をしている。重みがすべて 1 の場合は、素性リストの全要素を次元で表現し、用例を高次元実数値ベクトルで表現し、それらの余弦尺度によって類似度を測ることに対応している。

問題は線形モデルのパラメータ a, b, c, d, e, f（つまり素性毎の重み）をどのように設定するかである。

3.2 訓練データの作成

線形モデルのパラメータを求めるために、重回帰分析を利用することにする。重回帰分析では線型モデルのパラメータを最小自乗法で求める。しかしここで問題がある。重回帰分析では訓練データとして観測値が必要である。この場合は、訓練データとして用例間の実際の類似度が必要であるが、そのようなものは手作業であっても与えることはできない。

この対処として、ここでは、経験的なパラメータ値から用例間の類似度を与えることにした。用例中の対象単語の語義の類似度は手作業であっても与えることはできないが、異なる語義であることは判定することができる。つまり、異なる語義である場合は類似度を 0 とし、同じ語義である場合は、経験的なパラメータ値を用いて類似度を与える。

これによって訓練データを作成することができる。具体的にここではこの経験的なパラメータ値として、以下を用いた。

$$a = b = 10,\ c = d = 5,\ e = f = 1$$

3.3 最小自乗法によるパラメータ推定

n 個のデータがそれぞれ m 次元の実数値ベクトルで表現されているとする。i 番目のデータ $x^{(i)}$ を以下で表現する。

$$x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_m^{(i)})$$

$x^{(i)}$ と $x^{(i)}$ の観測値 $y^{(i)}$ に線型モデルを当てはめる。

$$y^{(i)} = \sum_{j=1}^{m} a_j x_j^{(i)} + e_i$$

ここで $e_i \sim N(0, \sigma^2)$ である。ここから以下の残差平方和を最小にするようにパラメータを推定する。

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y^{(i)} - \sum_{j=1}^{m} a_j x_j^{(i)} \right)^2$$

以下、各パラメータで偏微分を行い、極値問題を解けばパラメータが求まる。これは最小自乗法によりパラメータ推定を行っていることと同じである。

結論だけ述べれば、パラメータは以下で与えられる。

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mm} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{my} \end{bmatrix}$$

ただし

$$S_{ij} = \sum_{k=1}^{n} (x_i^{(k)} - \bar{x_i})(x_j^{(k)} - \bar{x_j})$$

$$S_{iy} = \sum_{k=1}^{n} (x_i^{(k)} - \bar{x_i})(y_j^{(k)} - \bar{y})$$

$$\bar{x_i} = 1/n \sum_{k=1}^{n} x_i^{(k)}, \quad \bar{y_i} = 1/n \sum_{k=1}^{n} y^{(k)}$$

4 実験

まず訓練データの作成であるが、ここではSENSEVAL2 の日本語辞書タスク (白井清昭, 2003) で用いられた名詞 50 単語の訓練データを利用することにした。このデータは語義曖昧性解消のタスクに対する訓練データであり、本質的に各用例に対象単語の語義の id が付与されていると考えて良い。このデータから各用例対が同じ語義がどうかを自動で判定できる。同じ語義の場合は、線形モデルの経験的なパラメータ値から類似度を与え、異なる語義の場合は 0 を与えた。最終的に得られた線形モデルに対する訓練データ数は 846,045 個であった。

ここから最小自乗法により線形モデルのパラメータを推定した。得られたパラメータは以下であった。

$$a = 8.1987, \quad b = 8.0044, \quad c = 3.3696, \quad d = 3.8949, \quad e = 1.0512, \quad f = 0.4125$$

次に以下の類似度の定義に従って用例のクラスタリングを行った。

類似度 (1) 単純な線型モデルによる用例間類似度 (各パラメータ値は 1)

類似度 (2) 経験的なパラメータ値を用いた用例間類似度 (各パラメータ値は前述)

類似度 (3) 本手法により得られたパラメータ値を用いた用例間類似度 (各パラメータ値は上記)

クラスタリングにはクラスタリングツールの CLUTO[2]を利用した[3]。

クラスタリングの対象も SENSEVAL2 の日本語辞書タスクで用いられた名詞 50 単語の訓練データとした。これはクラスタリングの正解が付与されている形になっているので、クラスタリングの評価をエントロピーで行うことができる (新納浩幸, 2007)。エントロピーは値が小さいほどよいクラスタリングであることを意味する。結果を表 1 に示す。

表 1 より単純な類似度 (1) よりも経験的なパラメータ値を用いた類似度 (2) の方が良い結果が得られている。さらに経験的なパラメータ値を用いた類似度 (2) よりも本手法の類似度 (3) がさらに良い結果を出している。

5 考察

前述した実験において、平均のエントロピーは本手法が最良値を出したが、その差は非常に小さい。また最良値を出した単語数でみると類似度 (1) が 25 単語、類似度 (2) が 15 単語、類似度 (3) が 19 単語 であり[4]、類似度 (1) の最も単純なモデルが優れているという結果になる。これは類似度 (2) の経験的なパラメータ値が類似度 (1) よりも適切でなかったためである。類似度 (3) は類似度 (2) のパラメータを学習により改善したものと捉えることができる。そのため類似度 (3) は類似度 (2) よりも、平均のエントロピーと最良値を出した単語の数の両方において良い値となっている。つまり経験的なパラメータ値をうまく設定できれば、ここで示した学習手法により最良のパラメータを得ることができるであろう。

パラメータを推定するために、ここでは重回帰分析の手法を用いたが、判別分析の手法を用いて線形判別関数を求めることでも同様の推定が可能である。この場合、訓練データの観測値は同じ語義かどうかだけで済むので、より適切な推定が行えるようにも見える。しかし本タスクの場合、訓練データ（用例対）の大部分は異なる語義のクラスに属し、かつ原点となっている。このようなアンバランスな訓練データに対しては適切な推定を行うことが困難である。

用例間の類似度を測る場合、類似度のモデル以上に重要になるのは単語間の類似度である。用例中にある対象単語の周辺の単語は数も少なく、まったく一致しない場合も多い。その際に用例間の類似

[2] http://glaros.dtc.umn.edu/gkhome/views/cluto

[3] 起動するプログラムはデータ間の類似度からクラスタリングを行う scluster である。そこで使われているアルゴリズムはトップダウンにデータを 2 分割してゆく処理を、目的のクラスタ数が得られるまで再帰的に行う k-way clustering と呼ばれる手法である。

[4] 合計が 50 単語にならないのは、同点のものを重複して数えているかである。

単語	データ数	クラスタ数	類似度 (1)	類似度 (2)	類似度 (3)
間	266	9	0.368	0.384	0.383
頭	169	6	0.524	0.505	0.511
一般	267	5	0.635	0.652	0.649
一方	274	4	0.247	0.253	0.253
今	333	5	0.509	0.517	0.512
意味	173	3	0.964	0.920	0.930
疑い	201	2	0.080	0.080	0.080
男	213	4	0.198	0.196	0.192
開発	209	3	0.646	0.652	0.700
核	255	5	0.298	0.292	0.279
関係	414	3	0.650	0.656	0.652
気持ち	256	5	0.470	0.475	0.474
記録	236	3	0.564	0.547	0.547
技術	198	2	0.367	0.365	0.302
現在	341	2	0.362	0.404	0.404
交渉	242	2	0.143	0.145	0.145
国内	277	2	0.886	0.891	0.891
言葉	263	4	0.669	0.651	0.642
子供	354	2	0.987	0.973	0.973
午後	396	3	0.717	0.724	0.715
市場	254	4	0.485	0.483	0.449
市民	207	2	0.953	0.955	0.965
社会	340	6	0.239	0.241	0.232
少年	190	2	0.232	0.237	0.237
時間	283	4	0.537	0.584	0.567
事業	253	2	0.888	0.873	0.870
時代	360	4	0.480	0.483	0.485
自分	362	2	0.302	0.303	0.306
情報	285	3	0.523	0.507	0.504
姿	201	4	0.655	0.567	0.567
精神	157	2	0.677	0.711	0.692
対象	236	2	0.301	0.288	0.288
代表	466	3	0.354	0.343	0.329
近く	238	3	0.681	0.586	0.586
地方	271	2	0.945	0.950	0.947
中心	255	2	0.231	0.252	0.252
手	230	12	0.467	0.420	0.441
程度	202	2	0.137	0.137	0.137
電話	270	3	0.517	0.462	0.465
同日	234	2	0.799	0.796	0.798
花	175	3	0.114	0.111	0.113
反対	241	2	0.197	0.205	0.206
場合	292	2	0.751	0.789	0.789
前	426	4	0.264	0.262	0.267
民間	174	2	0.157	0.157	0.158
娘	203	3	0.362	0.359	0.357
胸	156	5	0.535	0.537	0.549
目	229	9	0.389	0.412	0.414
もの	757	14	0.548	0.522	0.516
問題	636	4	0.114	0.115	0.117
平均値	278.4	3.76	0.4824	0.4786	0.4767

表 1: 実験結果

度を測る手がかりは単語間の類似度が最有力である。通常、単語間の類似度は既存のシソーラスを用いて測るが、既存のシソーラスがその規模や構造の点から本タスクに適しているかどうかは疑問である。今後は本タスクに適した単語間の類似度を大規模に作成していく必要があるだろう。

クラスタリング一般で考えてもデータ間の類似度の設定方法が本質的であり、近年はデータ間の類似度を学習により設定するという研究が盛んであるので（(Liu Yang, 2007) など）、それらの研究成果も取り込みたい。

また用例を対象単語の語義に基づいてクラスタリングするタスクではクラスタの数（つまり語義の数）の推定も重要である (Hiroyuki Shinnou and Minoru Sasaki, 2008)。SemEval-2007 の Task-02 ではここでのタスクと本質的に同等のタスクを扱っているが、中心の問題はクラスタ数の推定とその評価方法であった (Eneko Agirre and Aitor Soroa, 2007)。

最後に、ここでは対象単語が名詞であったが、動詞の場合、用例間の類似度の測定は用例による翻訳で生じる動詞の格フレームの選択問題と同型であることも指摘しておく。用例による翻訳では用例集から入力文と類似の文を検索するが、基本的には文の中心となる動詞を対象単語として、用例間の類似度を測っている。標準的には格の種類と格に入る名詞の類似度から算出される。つまり使われているモデルは、ここで示した線形モデルであるため、本手法が応用できる。

6 おわりに

用例をその単語の語義に基づいてクラスタリングするために、本論文では用例間の類似度を測る手法について述べた。用例間類似度を線型モデルで表し、次にパラメータを最小自乗法により推定する。訓練データの構築については経験的なパラメータ値を用いた仮のモデルを使うことを提案した。実験では、本手法の類似度の定義は、単純な定義や経験的な定義よりも、よいパフォーマンスを示すことができた。今後は、本タスクに適した単語間の類似度を大規模に構築し、クラスタリングの精度を更に改善したい。

文献

Eneko Agirre and Aitor Soroa (2007) “SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems,” in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 7–12.

Donald Hindle (1990) “Noun classification from predicate argument structures,” in *Proceedings of the 28th annual meeting on Association for Computational Linguistics (ACL-90)*, pp. 268–275.

Hiroyuki Shinnou and Minoru Sasaki (2008) “Division of Example Sentences Based on the Meaning of a TargetWord Using Semi-supervised Clustering,” in *Proceedings of LREC-2008*.

Liu Yang (2007) “An Overview of Distance Metric Learning,” in *http://www.cs.cmu.edu/ liuy/dist_overview.pdf*.

Masaki Murata and Masao Utiyama and Kiyotaka Uchimoto and Qing Ma and Hitoshi Isahara (2001) “Japanese word sense disambiguation using the simple Bayes and support vector machine methods,” in *Proceedings of the SENSEVAL-2*, pp. 135–138.

Resnik Philip (1992) “WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery,” in *Proceedings of AAAI-92 Workshop on Statistically-Based NLP Techniques*, pp. 48–56.

新納浩幸 (2007) R で学ぶクラスタ解析，オーム社.

白井清昭 (2003) 「SENSEVAL-2 日本語辞書タスク」，自然言語処理，第 10 巻，第 3 号，pp.3–24.

BCCWJ における推量副詞とモダリティ形式の共起

スルダノヴィッチ・イレーナ（作文支援システム班協力者：東京工業大学）[1]
ベケシュ・アンドレイ（作文支援システム班協力者：リュブリャーナ大学）
仁科喜久子（作文支援システム班班長：東京工業大学）

Suppositional Adverbs and Modality Forms Collocations in BCCWJ

Irena Srdanović (Tokyo Institute of Technology, University of Ljubljana)
Andrej Bekeš (University of Ljubljana)
Kikuko Nishina (Tokyo Institute of Technology)

１．はじめに

日本語における推量副詞と文末モダリティ形式には強い共起関係がある。この二つの部分は文章の中で離れている位置に出現する場合が多く、「遠隔共起関係」と考えることができる。このような関係は南（1974）の入れ子構造の研究、工藤（2000）の副詞と陳述副詞の呼応の研究において既に注目されている。Bekeš(2006)では、このような共起関係が会話において発話の理解を促進させるという機能について論じられている。Srdanović ら（2008a）では、さまざまなコーパスを分析した結果、それぞれの副詞の分布がコーパスの種類またはジャンルによって異なり、副詞と文末モダリティ形式との共起傾向にも差異が見られることが明らかにされた。その上 Srdanović ら（2008b）では、現代日本語書き言葉均衡コーパス（BCCWJ）(Maekawa 2006; 山崎 2006)を利用し、他のコーパスとの比較を行なった結果、BCCWJ の書籍と大規模なウェブコーパスの JpWaC（Erjavec ら 2007; Srdanović ら 2008c）は偏りが無いデータの傾向を示していることが明らかになった。本稿では、これらのコーパスにおける推量副詞と文末モダリティの共起関係を検討する。この得られた共起関係に対して日本語の辞書に現れる推量副詞と文末モダリティの共起関係を比較する。BBCWJ など日本語のコーパスは辞典の編集への応用にも重要な役割を果たすものと期待できる。

２．副詞の分布による偏りがないコーパス

Srdanović ら(2008b)では、13 種のコーパスにおける推量副詞の分布を比較し、それぞれのコーパスの特徴との関係について述べた。利用したコーパスは表 1 に並べた順番で次の通りである。(1) BCCWJ の政府系白書コーパス（KokkenOW）、(2) 自然言語処理論文（NLP）、(3) 16 冊分の大学理系基礎科目教科書のコーパス（16K）、(4) インフォーマルな談話を含むコーパス（NUJCC）、(5) フォーマルな会話のコーパス（Oikawa）、(6) Yahoo!知恵袋のウェブコーパス（KokkenOC）、(7) 大規模なウェブコーパス JpWaC、(8) BCCWJ の書籍コーパス（KokkenBK）、(9) 一年間の毎日新聞データ（Mai2002）、(10) 小学生の国語教科書（KokugoK）、(11) BCCWJ の一部である中学生の教科書（KokkenK）、(12) KokkenK に含まれる国語教科書（KKK）、(13) 新聞、50 年前までの小説などの統一していないデータ（Kudo）である。

[1] srdanovic.i.ab@m.titech.ac.jp

表 1 はコーパスにおける推量副詞の分布の相対頻度をパーセントで示している。特定の副詞が最も高い比率で現れる場合は最も偏りのあるデータとして認められる。それらは、BCCWJ の政府系白書コーパス（KokkenOW）、自然言語処理論文（NLP）、インフォーマルな談話を含む NUJCC コーパス、理系教科書の 16K の 4 種のコーパスである。一方、バランスのとれた副詞の分布を示しているコーパスとしては BCCWJ の書籍コーパス（KokkenBK）とウェブコーパス（JpWaC)があり、偏りが無いデータの傾向を示しているといえる。[2]

表 1 複数コーパスにおける推量副詞の分布[3]

副詞/コーパス	KokkenOW	NLP	16K	NUJCC	Oikawa	KokkenOC	JpWaC	KokkenBK	Mai2002	KokugoK	KokkenK	KKK	Kudo
かならず	5%	23%	42%	7%	14%	4%	8%	15%	25%	12%	28%	16%	4%
ぜったい	2%			52%		14%	9%	6%	11%	3%	2%	4%	5%
ぜったいに	2%		4%			11%	6%	8%	12%	3%	9%	2%	
かならずしも	84%	66%	39%	1%	5%	2%	6%	6%	8%	0%	10%	6%	
よほど	0%				1%	2%	2%	3%	2%	3%	1%	2%	4%
よっぽど				2%		2%	1%	1%	1%	1%			
たいがい				2%	8%	1%	1%	1%	0%				1%
たいてい	1%	6%	4%	1%		5%	4%	2%	3%	4%	6%	12%	1%
きっと		3%		15%	8%	15%	12%	14%	10%	38%	26%	26%	28%
おおかた					1%	0%	0%	1%	0%	1%	1%	2%	3%
おそらく	1%	3%	7%	1%	8%	1%	13%	12%	9%	2%	5%	10%	19%
さぞ						0%	0%	1%	1%	4%	1%	2%	5%
たぶん	2%		3%	3%	39%	26%	16%	11%	6%	3%	4%	8%	10%
どうも	0%		1%	6%	7%	6%	8%	7%	5%	15%	2%	4%	5%
どうやら				2%		3%	5%	5%	3%	3%			5%
あんがい	0%		1%	3%	1%	0%	2%	1%	1%	1%	1%		2%
ひょっとしたら				1%	1%	1%	1%	1%	1%	1%	1%	2%	3%
ひょっとすると							0%						
ことによれば						0%	0%	0%	0%				
ことによると	3%		1%				0%	0%	1%				1%
もしかしたら				5%	8%	5%	3%	3%	1%	2%	2%	2%	5%
もしかすれば						1%	0%						
もしかすると							1%	1%	0%	1%	1%	2%	
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

さらに推量副詞の出現分布の偏りについて、分布の偏りを測るエントロピーを計算すると、同様の結果が示される（表 2 参照)。エントロピーは、各単語の出現分布が均等であるほど値は高く、特定の単語が際だって出現しているコーパスほど数値は低く分布に偏りがあるといえる。

---

[2] 全体のコーパスは偏りのないコーパスなのか、副詞以外のデータを分析して、確認することができる。両方のコーパス、BCCWJ の書籍コーパスとウェブの JpWaC は偏りのないデータを目的として作成されたコーパスである。Srdanović ら（2008c）では、新聞データと比較した結果、JpWaC のほうが偏りのデータとして評価された。その応用について Srdanović・仁科（2008）を参照されたい。

[3]「よほど」「よっぽど」は Kudo と Oikawa コーパスでは同一語として分析している。「ぜったい」ぜったいに」は Kudo と NUJCC コーパスでは同一語として分析している。「もしかしたら」「もしかすると」は NUJCC コーパスでは同一語として分析している。JpWaC のサンプル分析の結果では「どうも」37%、「おおかた」44%、「ぜったい」17%となったが、その中から副詞と認められないものを排除した結果を表 1 に示した。

ここで各推量副詞の出現率を $P(A_j)$とすると、各コーパスにおけるエントロピーI は $I=-\sum_j P(A_j)\log(1/P(A_j))$となり、ひとつの推量副詞のみが出現しているコーパスにおいて、エントロピーの値は 0 に近くなる。この計算結果を表 2 で示す(本研究では底を 2 として計算した)、BCCWJ の書籍コーパス（KokkenBK）の値が最も大きく、ウェブコーパス（JpWaC)がそれに次いでいて、白書が最も低い。情報エントロピーの観点からみると BCCWJ の書籍、Web コーパスのエントロピーが高く、偏りが無いデータであり、白書は偏りがあるデータといえる。

表 2　エントロピー形によるコーパスにおける副詞の分布

コーパス	エントロピー	
KokkenBK	3.7140703	偏りがない
JpWaC	3.7009878	
Mai2002	3.4980456	
KokkenOC	3.3438106	
KKK	3.3242558	
Kudo	3.3228428	
KokugoK	3.1359857	
KokkenK	3.0703589	偏りがある
Oikawa	2.8167623	
NUJCC	2.4855166	
16K	1.9616011	
NLP	1.4137994	
KokkenOW	1.0723104	

本稿では、書籍と JpWaC を利用して、偏りのない傾向を示しているデータにおいて推量副詞と文末モダリティ形式の共起関係はどの程度共通点と差異があるかについて検討する。

３．書籍と JpWaC における副詞と文末モダリティ形式の共起関係

本章では、BCCWJ の書籍と JpWaC における推量副詞と文末モダリティ形式の共起関係を検討する。ここで用いたデータは JpWaC 全４億語からランダムに取り出した 2 千万語の部分的なコーパスである。

表３、４は、それぞれ書籍と JpWaC における推量副詞の分布とそれぞれの副詞ともっとも高頻度およびもっとも顕著性計指数が高い値[4]で共起する５つのモダリティ形式を示している。まず、副詞の分布から見ると、両方のコーパスの結果はかなり類似しているが、高頻度の副詞において「かならず」と「たぶん」の出現に最も差異がある。書籍ではもっとも高頻度の「かならず」は、新聞のコーパス（Mai2002）や教科書、論文で高頻度であることを勘案すれば（表１参照)、フォーマルな書き言葉コーパスの特徴を表しているといえる。JpWaC でもっとも高頻度の「たぶん」は、フォーマルな会話コーパス Oikawa と知恵袋のデータ（KokkenOC）でも高頻度なので、JpWaC の方が書籍よりフォーマルな会話の性格も多少示しているといえる。

[4] JpWaC の場合には、顕著性計指数が高い項目の順番で５つのモダリティ形式が並んでいる。顕著性（salience）はコーパスにおける共起の統計的な重要性を表している。ここでは共起頻度を中心語頻度と共起語頻度の和で割って、2 倍したダイス係数の統計値が利用されている。

表 3　BBCWJ の書籍におけるもっとも高頻度の
副詞と文末モダリティ形式の共起関係

タイプ	副詞	頻度	上位の5つのモダリティ形式（共起頻度において）									
確信・推測	かならず	4548	のだ	334	はずだ	163	だろう	151	と思う	76	に違いない	44
推測	おそらく	4216	だろう	961	のだろう	625	と思う	225	に違いない	157	のだ	153
推測・確信	きっと	3547	だろう	417	のだろう	332	に違いない	251	と思う	224	はず	157
推測	たぶん	3241	だろう	487	のだろう	412	と思う	290	のだ	111	のではないか	109
推定	どうも	2320	のだ	156	らしい	149	ようだ	126	気がする	65	と思う	43
確信	ぜったいに	2114	のだ	141	はずだ	61	だろう	59	と思う	51	べきだ	25
不確定	もしかしたら	1824	かもしれない	401	のかもしれない	314	のではないか	130	のか	95	のだ	70
確信	かならずしも	1591	わけではない	127	のだ	97	ものではない	69	とはいえない	78	とは限らない	72
確信・推測	ぜったい	1581	のだ	111	と思う	58	だろう	38	はずだ	26	のか	25
推定	どうやら	1512	らしい	504	ようだ	319	のようだ	87	のだ	42	だろう	12
確信・推測	たいてい	1217	のだ	98	だろう	29	ことではない	18	ようだ	15	と思う	15
推測	よほど	1047	のだろう	95	のか	76	のだ	60	だろう	43	らしい	42
不確定	ひょっとした	967	かもしれない	218	のかもしれない	148	のではないか	111	のか	36	と思う	36
推測	さぞ	427	だろう	128	ことだろう	76	のだろう	29	に違いない	26	と思う	23
不確定	あんがい	423	のだ	53	のかもしれない	35	かもしれない	35	のではないか	8	ではないか	7
推測	おおかた	281	のだろう	31	のだ	21	だろう	16	と思う	10	ことだろう	6
確信・推測	たいがい	172	のだ	11	だろう	2	だろうか	2	はずだ	2	のか	2
不確定	ことによると	62	のかもしれない	10	かもしれない	8	のではないか	4	のか	3	ものであろう	2

表 4　JpWaC におけるもっとも高頻度の副詞と文末モダリティ形式の共起関係

タイプ	副詞	頻度	上位の5つのモダリティ形式（共起頻度、共起顕著性において）									
推測	たぶん	1527	のだろう	124 36.5	と思う	221 35.87	だろう	175 35.23	のだと思う	32 30.22	のではないかと	28 24.82
確信・推測	かならず	1448	はず	31 28.09	のだ	95 27.68	と思う	41 21.96	だろう	20 16.29	わけだ	6 11.57□
推測	おそらく	1341	だろう	269 41.2	のだろう	128 36.7	ことだろう	37 29.06	と思う	131 28.38	に違いない	21 26.0
推測・確信	きっと	1340	のだろう	160 41.04	だろう	159 34.54	に違いない	34 33.09	ことだろう	44 32.13	はず	52 29.05
確信・推測	ぜったい	1294	のだ	87 25.35	だろう	40 22.56	と思う	41 20.92	はず	17 20.12	べき	13 16.66
推定	どうも	1022	ようだ	59 34.4	らしい	44 33.65	気がする	28 27.32	のだ	63 22.16	のようだ	12 21.52
確信	ぜったいに	816	てはならない	21 34.94	のだ	79 26.43	べき	15 19.49	と思う	29 19.08	はず	12 17.88
推定	どうやら	548	らしい	121 48.54	ようだ	95 40.6	のようだ	27 31.21	みたいだ	16 26.2	そうだ	21 21.46
確信・推測	たいてい	497	のだ	45 23.71	と思う	17 16.79	だろう	14 16.36	はず	6 13.51	ように思う	3 9.99
推測	よほど	409	のだろう	34 28.12	だろう	22 18.87	と思う	25 18.55	のだ	31 17.26	のか	17 16.45
確信	かならずしも	382	とは限らない	53 51.05	わけではない	35 34.77	ものではない	11 21.89	ことではない	9 21.09	といえない	8 20.93
不確定	もしかしたら	316	かもしれない	102 43.36	のかもしれない	59 39.67	のかな	10 19.06	のではないか	8 15.87	のではないかと	5 12.04
不確定	あんがい	187	のかもしれない	11 22.14	気がする	9 19.42	のだ	18 16.13	かもしれない	8 16.1	だろうかと思う	2 15.53
不確定	ひょっとしたら	81	かもしれない	25 30.48	のかもしれない	17 28.67	と考えるのかもしれない	1 10.25	のではないかと	2 8.16	かもしれないのだ	1 7.87
確信・推測	たいがい	72	のだ	6 11.16	と思う	4 9.93	わけだ	2 8.8	のかなと思う	1 8.71	だろう	3 8.69
推測	さぞ	31	ことだろう	6 20.5	だろう	8 16.39	のだろう	5 14.41	だろうと思う	2 11.59	だろうと考える	1 9.16 □
推測	おおかた	24	ものと思う	1 8.46	ものだろう	1 7.77	と思う	2 7.42	のかな	1 6.72	と考える	1 6.52
(推定)	ことによると	2	らしい	1 7.46	べき□	1 7.02						

最も高頻度の副詞とモダリティ形式の共起関係を見ると、両方のコーパスでは類似した結果が見られる。ここから、それぞれの副詞と典型的に共起するモダリティ形式の傾向が同様だと分かる。興味深いのは、「のだろう」、「だろう」、「と思う」などの「推測」を表すモダリティ形式および「はず」、「のだ」、「に違いない」などの「確信」を表すモダリティ形式がもっとも高頻度の共起関係として出現することである。このことは人間のコミュニケーションにおいて、工藤（2000）が提唱している提案した推量を表す四つのモダリティタイプの中で「推測」と「確信」の方が「不確定」と「推定」より機能的に優先することを暗示していると考えられる。

両方のコーパスにおいて、一つの副詞と共起する「推測」のモダリティ形式の傾向があ

れば、共起の中では、「確信」のモダリティ形式も見られる。たとえば、「きっと」と「おそらく」の「に違いない」などである。逆に、副詞が「確信」のモダリティ形式と共起する場合、その副詞は「推測」のモダリティ形式とも共起する。たとえば、「かならず」、「ぜったい」と、「だろう」「と思う」という「推測」の文末形式が共起する場合である。

４．日本語の辞典との比較

本章では、3 章で得られた書籍と JpWaC における副詞と文末モダリティ形式の共起関係の結果を、三冊の日本語学習辞典と比較する。学習者によく利用されている和英辞典 2 冊（『ジーニアス和英辞典』、『ニューセンチュリー和英辞典』）と、中級レベル以上の日本語学習者にとって問題となる文型を網羅的に集めた辞典（『日本語文型辞典』）を対象にした（表 1 参照）。

副詞の分布から見ると、2 冊の和英辞典には対象となるすべての推量副詞が現われている。「文型辞典」には、現われない副詞もある。「あんがい」、「おおかた」、「ことによる」のような最も低頻度の副詞以外に、コーパス中にかなり出現している「ぜったい」と「ぜったいに」が現われていない。表 1 によると、この二つの副詞はインフォーマルの会話コーパスでは非常に高頻度が、書籍と JpWaC でも、新聞コーパス、知恵袋のコーパス、フォーマルな会話コーパスでもかなり出現するので、中級以上の日本語学習の辞典にも含んだ方が良いと考えられる。逆に、「さぞ」と「たいがい」は教科書系のコーパスでは頻度が非常に高いが、JpWaC や書籍などのコーパスにはあまり現れない。特に「さぞ」はやや古い表現だと考えられる。

推量副詞と共起するモダリティ形式についてはというと、二冊の和英辞典には、例文中にモダリティ形式が多く現れている。しかし、それぞれの辞典で例文に現れるモダリティ形式は異なっており、特定の副詞と共起する代表的なモダリティ形式が足りない場合もかなりある。たとえば、「たぶん」には、「かもしれない」と「ものだ」の代わりに、「のだろう」、「と思う」を入れた方がよい。また、「おそらく」には、「と思う」、「に違いない」も追加したほうが良いと考えられる。

「文型辞典」では、網羅されている推量副詞のすべての項目にモダリティ形式が現れているのは評価される。しかし、この辞典でも、推量副詞の分布に偏りのないコーパスに高い頻度および高い顕著性で現れるモダリティ形式のデータは応用できる。たとえば、「おそらく」という副詞に関し、「ものと思われる」のかわりに、「のだろう」、「ことだろう」、「と思う」のモダリティ形式を載せたほうがよい。「かならず」の項目では推量のモダリティ形式が現われていないので、「はずだ」、「のだ」、「だろう」、「に違いない」を追加した方が良い。また、「たいがい」と「たいてい」の出現は異なるので、辞典では同じ例文と説明で扱うのは適切ではない。

以上の指摘は、偏りがない傾向を示しているデータに基づいて行った、一般的な日本語習得を目的とした辞典の作成のためのものである。しかし、ジャンルによって代表的なモダリティ形式は異なるので（Srdanović ら 2008a,b）、学習辞典にもジャンル別のデータを反映する必要がある。

表 5　日本語辞書における副詞と文末モダリティ形式との共起

副詞	ジーニアス	ニューセンチュリー	文型辞典
あるいは	▲かもしれない	▲かもしれない	▲かもしれない　●のだろう，と思われる
あんがい	∆	▲のかもしれない	/
おそらく	▲でしょう,のだろう,だろう	∆	▲だろう,ものと思われる,にちがいない
おおかた	∆	∆	/
かならず	▲でしょう,なさい,だろう	▲でしょう,てください,必要がある,ことにしている	▲てください,しなければならない,ようにしよう
かならずしも	▲とは限らない,というわけではない,ものではない	▲とは限らない,とはいえない	▲とは限らない,ではないと私は思っている，しなければならぬものだと思っているわけではない　●わけではない，とはかぎらない
きっと	▲のです,にちがいない,んだ,だろう	▲だろう,に違いない,でしょう	▲でしょう，だろう,にちがいない,てください
ことによると	∆	▲かもしれない	/
さぞ（かし）	▲でしょう	▲でしょう,ことでしょう	▲ことだろう,ことでございましょう
たいがい	▲だろう	∆	（ことにしている,ことになっている）
たいてい	▲だろう	▲であろう	「たいがい」と同じ例文・説明
たぶん	▲だろう、かもしれない、でしょう	▲でしょう,だろう,かもしれない,ものだ	▲でしょう,ておこう,でしょうか,と思う
どうも	▲ようだ、ように思う	▲そうだ,ようだ,んだ,気がしない	▲そうだ,ようだ,らしい
もしかしたら	▲かもしれない	▲かもしれない	▲かもしれない,のではないだろうか
ひょっとしたら	▲だろう,かもしれない,でしょうか	▲のかもしれない,かと思う	/
よほど	∆	▲ようだ	▲に違いない,らしい,んだろう,んだ,んだろうと思う
ぜったい	∆	∆	/
ぜったいに	▲してはいけない,するな	不可能だ,ことはない	/

∆副詞は辞書にあるが、共起するモダリティ形式は現れない　▲モダリティ形式が例文にある
●モダリティ形式が他の表現として言及されている

６．まとめと今後の課題

本稿では､偏りが無いデータとしての BBCWJ の書籍コーパスとウェブコーパスの JpWaC において推量副詞と文末モダリティ形式の共起関係の共通点と差異を調べた。両方のコーパスでは推量副詞と文末モダリティ形式の共起関係に関する共通点が多く、類似したモダリティ形式との共起傾向が認められる。さらに、その結果を日本語の辞典と比較して、どの程度辞典には推量副詞と文末モダリティ形式の共起関係についての情報がカバーされているかについて検討した。調査した辞書では推量副詞と文末モダリティ形式の共起関係が

かなり網羅されているが、幾つかの問題点が明らかになった。コーパスから得られた結果に基づいて、日本語の辞典の改善への示唆を与えた。

謝辞

本稿を執筆するにあたり、東京工業大学大学院修士課程１年（仁科研究室）のホドシチェク・ボルさんにデータ抽出のためのプログラム作成で支援を得た。統計手法については東京工業大学グローバルエッジ研究院寺井あすか助教のご指導をいただいた。日本語語法についてはリュブリャーナ大学文学部アジア・アフリカ研究学科日本研究講座、井田尚美講師、内容については東京大学大学院教育研究科阿辺川学術研究支援員のご助力を得た。その他、多数の方々の支援を得たことをここに感謝する。

参考文献

工藤 浩 (2000)「副詞と文の陳述のタイプ」『日本語の文法３モダリティ』（森山卓郎，仁田義雄，工藤浩）岩波書店. 161–234.

南不二男（1974）『現代日本語の構造』大修館書店

山崎誠 (2006)「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.

Bekeš, A. (2006) *Japanese suppositional adverbs in speaker-hearer interaction*. Proceedings of the third conference on Japanese language and Japanese language teaching, Rome 2005. Venezia: Libreria editrice cafoscarina. 34–48.

Erjavec,T., Kilgarriff, A. Srdanović E. I. (2007) *A large public-access Japanese corpus and its query tool*. Inaugural Workshop on Computational Japanese Studies.

Maekawa, K. (2006) *Kotonoha, the Corpus Development Project of the National Institute for Japanese Language*. Language Corpora: Their Compilation and Application (Proceedings of the 13th NIJL International Symposium), Tokyo, 55-62.

Srdanović, E.I., Bekeš, A., Nishina, K. (2008a) *Distant Collocations between Suppositional Adverbs and Clause-Final Modality Forms in Japanese Language Corpora*. Large-scale knowledge resources: construction and application; Third International Conference on Large-Scale Knowledge Resources, LKR2008. Lecture Notes in Computer Science 4938 (Eds.T. Tokunaga and A. Ortega). Springer-Verlag Berlin Heidelberg, 252–266.

Srdanović, I, Bekeš, A., 仁科喜久子(2008b)「複数のコーパスに見られる副詞と文末モダリティの遠隔共起関係」特定領域研究、「日本語コーパス」平成 19 年度公開ワークショップ（研究成果発表会）予稿集, 223-230.

Srdanović, E.I., Erjavec,T., Kilgarriff, A.(2008c) *A web corpus and word-sketches for Japanese*. Journal of Natural Language Processing 15/2, 137-159.

Srdanović, E.I., 仁科喜久子(2008)「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 59–80.

参考学習辞典

（1）小西友七・南出康世『ジーニアス和英辞典』大修館書店 2003-2004（「ジーニアス」と省略）
（2）小西友七『ニューセンチュリー和英辞典』第 2 版、三省堂出版 2006（「ニューセンチュリー」と省略）
（3）砂川有里子/駒田聡/下田美津子/鈴木睦/筒井佐代/蓮沼昭子/ベケシュ・アンドレイ/森本順子, 『日本語文型辞典』, くろしお出版, 1998.（「文型辞典」と省略）

関連 URL

Sketch Engine コーパス検索ツールホームページ：http://www.sketchengine.com
KOTONOHA ホームページ：http://www2.kokken.go.jp/kotonoha/

講　演

3月16日（月）　13:00〜14:00

Disruptive Serviceを目指して：情報爆発プロジェクトと情報大航海プロジェクト

▶喜連川 優（東京大学生産技術研究所）

Disruptive Service を目指して:情報爆発プロジェクトと情報大航海プロジェクト

喜連川　優（東京大学生産技術研究所）

2008 年の CACM に Google 社の MapReduce 処理は一日に２０ＰＢと発表され、また、AT&T のネットワークを一日に流れる情報量は近日１０ＰＢを超えるとされている。膨大な情報の流れと、膨大な情報の処理は、今、まさに情報爆発時代の真っただ中にいることを示していると言えよう。特定領域「情報爆発ＩＴ基盤」の申請を行った 2004 年の秋、情報生成量はエクサバイトの単位で計量していた。2007 年及び 2008 年 3 月に IDC より出されたレポートによると 2010 年にはおおむね人類の創出する情報量はゼッタバイトに達すると予測されている。本プロジェクトの目指すところを一言で表現するならば、この情報爆発という『現象』に着目し、当該現象を情報分野の研究者の新たな課題と捉え、(1)情報爆発時代に生まれる種々の問題を同定しその解決を目指し先駆的に取り組むとともに、(2)情報爆発を問題と見做すのではなくむしろ積極的に見つめ、爆発する情報から新たな価値創出を目指す今までにない研究に挑戦しようというものである。

情報爆発特定は、安西慶応義塾長が率いられた情報学特定(01-05)の後継として、2005 年に発足した。2005 年は総括班のみの稼働となり、2006 年より計画班、公募班からなるプロジェクト全体が始動した。本プロジェクトは A01 情報爆発時代における情報管理・融合・活用基盤、A02 情報爆発時代における安全・安心 IT システム基盤、A03 情報爆発時代におけるヒューマンコミュニケーション基盤管理、B01 情報爆発時代における知識社会形成ガバナンスの大きく 4 研究グループ（13 計画研究班および 87 公募研究班）から構成される。即ち、爆発する情報からのサーチを始めとする価値創出を A01 が、爆発する情報と人間との接点を A03 が、A01、A03 を構築するシステム技術を A02 が担当、さらに社会との接点こそがこれからの IT にとって極めて重要であるとの認識から、B01 がこれを担当する。また、情報爆発時代においては爆発する情報量を処理するプラットフォーム（情報融合炉）に関する技術が肝となることを認識し、2004 年、特定領域研究に新たに導入された支援班というスキームを利用し、研究者全員で共に構築し、運用する共創プラットフォームを提案した。2006 年より InTrigger と名付けた独自の分散計算環境を一歩一歩構築し、現在、約 1000ＣＰＵコアに達している。さらに、大学の有する大規模な計算資源、TSUBAME(東工大)や T2K（東大、筑波、京大）とも連動可能とするように工夫し、大規模処理環境が次第に動作しつつある。

共創プラットフォーム上には検索エンジン『Tsubaki』が稼働し、ページの収集や大量ページに対する基本的な文法処理など低位の作業を一元化することにより研究者の負担を大幅に軽減し、存分に上位の多様な検索技術の開発に集中することを可能とする環境を提供している。また、『IMADE』では多視点の映像・音声記録システム、モーションキャプチャシステムなどを総合した実験工房を実現し、実世界インタラクションの記録・分析・支援に関する実験環境を提供している。

本特定領域では、共創プラットフォームを核として構築・活用することにより多様な情報研究者が互いに刺激しあう新しい研究のスタイルを推進し、「情報爆発」という多くの人間社会・文化・経済活動に影響を与える現象に、広い視野と高度な戦略に基づき挑んで行きたいと考えている。

講演者は　経済産業省「情報大航海プロジェクト」ステアリングコミッティ委員長を務めている。当該プロジェクトは、一言で言うならば情報爆発からの価値創出・サービス創出を目指すものであり、人々の行動履歴マイニングに基づく「便利なライフアシストサービス」、運動情報マイニングによる生活習慣病ソリューション、ひやりはっと情報のテキストマイニングによるプロアクティブな安全システムの構築など、社会に役立つ「非 google」なサービスの創出にむけて実用化を強く意識したプロジェクトを推進しており、その現況について紹介する。

書　名　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
発行日　平成21年3月11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　独立行政法人国立国語研究所研究開発部門内
電　話　042-540-4300（代表）
文書管理番号　JC-G-08-03