第6回 コーパス日本語学ワークショップ予稿集

第6回　コーパス日本語学ワークショップ　予稿集

国立国語研究所 言語資源研究系・コーパス開発センター

第6回 コーパス日本語学ワークショップ

予稿集

2014年9月9日、9月10日

主催　国立国語研究所 言語資源研究系・コーパス開発センター

会場　国立国語研究所

第6回 コーパス日本語学ワークショップ

予稿集

2014年9月9日(火)／9月10日(水)

Program [プログラム]

9月9日（火）

13:00～13:10　**■挨　拶**　前川 喜久雄

13:10～15:10　**■口頭発表（1）**

事象の構造から見る二重デ格構文の発生
▷孟 会君

文体指標と語彙の対応分析
▷浅原 正幸、加藤 祥、立花 幸子、柏野 和佳子

平安初期歌合和歌の品詞比率
▷冨士池 優美

複数のコーパスを用いた新しい文法シラバス策定の試み
▷庵 功雄、宮部 真由美、永谷 直子

15:10～15:30　**休憩**

15:30～17:00　**■口頭発表（2）**

NINJAL-LWP の類義語比較機能
▷赤瀬川 史朗、プラシャント・パルデシ、今井 新悟

述語項構造を意識した名詞の意味構造アノテーションのための名詞意味構造の検討
▷竹内 孔一

テキストとアノテーションの汎用同時検索システム
▷狩野 芳伸、増田 勝也

17:00　**■閉　会**

9月10日（水）

10:00～11:00　**■ポスター発表　A グループ**

漢語サ変動詞の卓立性の再考 ―動詞形・構文形比率を手掛かりとして―
▷李 楓

均衡性と代表性に配慮した『太陽コーパス』の分析法試論
▷森 秀明

地方議会会議録コーパスを用いたオノマトペの分析
▷高丸 圭一、内田 ゆず、乙武 北斗、木村 泰知

現代日本語の類義表現に関するテクスト言語学的研究
―「焦点を当てる」と「焦点を置く」に着目して―
▷吉本 秋水

近代語コーパスにみる「結果」の用法
▷髙橋 圭子、東泉 裕子

BCCWJ における複合動詞後項の表記の実態
▷小椋 秀樹

多様な会話コーパスを対象とした発話連鎖ラベリングの試み
▷森 大毅、森本 郁代、大場 美和子、吉田 悦子、伝 康晴

和文体および漢文体をもつ資料の構造化 ―法華百座聞書抄の事例研究―
▷河瀬 彰宏、野田 高広

『虎明本狂言集』における「思ふ」と「存ず」―『虎明本狂言集』のコーパスデータを利用して―
▷渡辺 由貴

日本語点字資料の語種的特徴

▷中野 真樹

全文検索システム『ひまわり』を用いた既存言語資料の活用方法の検討

▷山口 昌也

ハ / ガ使用の計量的研究 ―有無・量的大小の述語の場合―

▷服部 匡

11:00～12:00 **■ポスター発表 B グループ**

コーパス検索による副詞の文中における基本生起位置の検討

▷難波 えみ、玉岡 賀津雄

BCCWJ と日英パラレル新聞コーパスに基づいた格外連体修飾形の研究

▷田邊 和子

テキストにおける多義語の意味の集中度

▷山崎 誠

拡張固有表現階層から SUMO への対応表

▷今田 水穂

明治後期における漢語の基本語化

▷田中 牧郎

「勉強する」と「rian」の対象語の分析

―BCCWJ と TNC（Thai National Corpus）を用いて―

▷木田 真理、Khommapat PRAWANG、生田 守

『バイリンガルコーパス・ナビゲーター』オンライン日伊並列コンコーダンサの構築と活用

▷ Zotti Patrizia、Apolloni Riccardo、松本 裕治

語学学習 SNS の添削ログからの母語訳付き学習者コーパスの構築に向けて

▷水本 智也

韻律情報にもとづいた機能表現の抽出

▷土屋 智行、伝 康晴、小磯 花絵

日本語作文推敲支援システム「ナツメグ」における学習者評価実験から見られる課題

▷八木 豊、ホドシチェク・ボル、阿辺川 武、仁科 喜久子

連濁に前部要素の音韻的特徴が与える影響：連濁データベースを利用した研究

▷太田 真理、太田 聡

12:00～13:00 **昼食・休憩**

13:00～14:30 **■指定討論**

▷森 大毅、山崎 誠、庵 功雄、田中 牧郎、山口 昌也、竹内 孔一

14:30～15:00 **■全体討論**

15:00 **■閉　会**

Contents [目次]

■ポスター発表 Bグループ

口頭発表（1）

9月9日（火）　13:10～15:10

事象の構造から見る二重デ格構文の発生

孟会君（北京外国語大学日本学研究センター）

Studies on the Double-de-case Construction from the Perspective of Event Structure

Meng HuiJun (Beijing Center for Japanese Studies)

要旨

格の重複現象に制約がかかっていることは多くの研究者に指摘されてきたが、本稿ではBCCWJ 中納言からの検索用例に基づき、二重デ格構文の発生可能性を事象の構造の面から考察する。意味格である「デ」はその事象成立での機能により、①「事態成立の基盤」という空間次元の「デ」、②「事態成立の媒介物」というモノ・コト次元の「デ」、③「事態成立のサマ」を規定する属性次元の「デ」という三種類に分けることができ、それらは事象の構造において異なる役割を果たしているので、お互いに共起できるが、同じ種類に属する意味役割同士は事象成立での機能も同じなので、事象の階層性や付け加えなどの場合を除いては基本的には共起できない。そのほか、〈様態〉デ格は副詞に近づく存在なので、その他のデ格との共起だけではなく、それ自身の二重共起でも容易に発生できる。

1. はじめに

日本語において、格の文中での並び方に関して、格の重複現象はよく問題にされている。格の重複現象というのは、一つの述語に対して同一の格形式が二つ以上現れるという現象のことであり、本稿では、格助詞の形態に従ってそれらを、「二重ガ格/ヲ格/ニ格/デ格」のように呼ぶことにする[1]。

格重複現象の発生は格助詞そのものの性質と大きな関係があるので、本稿では意味格であるデ格に研究対象を絞り、その二重共起の可能性について考察することを通し、意味格の格重複の発生メカニズムの究明に貢献してみたい。具体的なやり方としては、格重複に関する従来の先行研究では、主に作例による議論が進められてきたが、個人の内省による判断の違いがあり、格重複の使用実態を把握するために、本稿では、「現代日本語書き言葉均衡コーパス（BCCWJ）」を用い、そこから検出された実用例に基づき、二重デ格構文の発生可能性を考察してみたい。

2. 先行研究

従来、格の重複現象に関する研究は主に「ガ、ヲ、ニ」という三つの文法格に集中し、デ格の二重共起に関するものはほとんどが、格重複制約のそれへの適用を指摘することに止まっている。

たとえば、益岡（1987）は日本語での言語観察に基づき、「同じ意味を表す格は重複することができない」という原則を引き出し、それは例⑴のようにデ格にも適用できると指摘している。

⑴ a. * 大学で 教室で 音楽を聞く。
　 b. 　部屋で ヘッドホーンで 音楽を聞く。　（益岡 1987：22）

[1] 場合によって格助詞が三重以上に共起するケースも見られるが、極めてまれであり、本稿ではとくに格の二重共起と区別しないことにする。

「意味役割」を格重複の発生の決定要因とするという点において、益岡（1987）はFillmore(1971)の「単文異格の原則」[2]とほぼ同趣であるが、それですべての言語事象を説明できるかどうか疑うが残る。例⑴aにおける二つの〈場所〉デ格の隣接からは確かに抵抗感が感じられるが、それを⑵のようにすると、それほど許容できないわけでもないであろう。そのほか、例⑶のように、異なる意味役割同士なら必ず共起でき、同じ意味役割同士であったら必ずしも共起できないわけでもないようなので、この点についてはさらに検討する必要がある。

⑵　a.　大学では、いつもの教室で音楽を聞く。
　　b.　この大学で、芸術学部の教室では音楽を聞くことができる。
⑶　a.＊木切れでのみで仏像を彫った。　（杉本 2013：39）
　　b.　クリスマスイブに、全国で若者たちが路上でダンスパーティーを開いた。
（矢澤 2007：213）

そのほか、二重デ格の発生可能性を検討することにより、デ格の用法を本質的に捉えてみようとする研究も見られる。たとえば、矢澤（2007）はデ格の「階層依存」という性質により、二重デ格現象を説明している。それによると、デ格は基本的には「動きの仲介・媒体」を表すが、その前接名詞や動詞、階層などのあり方によって〈場所〉〈道具〉〈原料〉〈原因〉などの意味に解釈される。このように、同一の意味役割であっても、出現位置が異なれば、その文中での機能も異なってくるので、一文中に共起できるのであるという。例⑶bはまさにそのように発生してきたものであろう。

矢澤（2007）での論述は同一の意味役割のデ格の共起に対してかなりの説明力を持っているが、異なる階層に出たデ格ならすべて共起でき、同じ階層に現れたものであれば意味役割が異なっても共起できないのか、はっきりしていないので、デ格の重複現象を体系的に把握したとは言えないであろう。そのほか、それは〈場所〉〈様態〉のデ格をも「動きの仲介・媒体」に帰結するが、それもやはり少々無理があるであろう。この点に関して、杉本（2013）は異なる捉え方をしている。それはデ格に二つの「同形異機能格」があると主張し、そのうち、〈原因〉〈手段〉〈材料〉の「デ」が例⑷のように一文中に共起できないのはそれらが「同一の格助詞」であるからという。

⑷　a.＊材料不足で代替素材で製品を作った。
　　b.＊列車事故でバスで振り替え輸送を行った。　（杉本 2013：39）

確かに、そのような捉え方をすると、上述の三者の共起は Fillmore(1971)の「単文異格の原則」により経済的に排除できるが、そのようにすると、〈場所〉デ格の各用法、たとえば、〈範囲〉〈動作主〉なども同一の格助詞になるので、それらも一文中に共起できなくなるであろう。実際はどうなるのか、この点についてコーパスにより検証する必要がある。そのほか、例⑷bと同じような組み合わせになる例⑸は矢澤（2007）では認められているので、〈原因〉〈手段〉〈材料〉の「デ」の共起可能性についても検証する必要があるであろう。

⑸　鉄道ストでバスで登校した。　（矢澤 2007：245）

[2]　それによると、「一つの単文には同一の深層格を担う名詞句が共起できない」のである。

このように、二重デ格の発生を考察するには、デ格の用法をどのように捉えたらいいのかについて再検討する必要もあるし、個人の内省による判断のゆれをも避けなければならないので、本稿ではそれらを事象の構造での機能により捉えた上で、BCCWJ からの検索用例に基づき二重デ格構文の発生可能性について考察していきたい。

3. 事象の構造に基づくデ格の分類

3.1 「事象の構造」とは

本稿では、事象の構造の面から二重デ格構文の発生を検討するが、ここでいう「事象の構造」というのは「状態」、「活動」、「到達」、「達成」のようなアスペクチュアルな捉え方ではなく、言語により表出された客観世界の出来事の構造のことである。本稿では、デ格との関連で事象の構造を図 1 のように把握してみる。

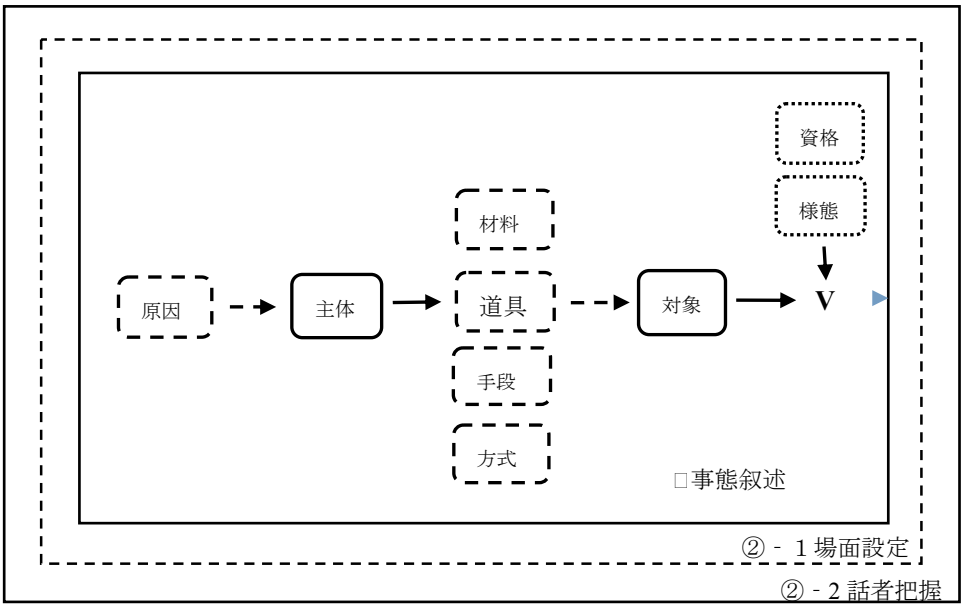


図 1　デ格との関連での「事象の構造」への把握[3]

つまり、事象の表出は、①のような動力連鎖のような構造をしている「事態叙述」とその外側を包む②－1 の「場面設定」と②－2 の「話者把握」からなり、そのうち、前者の「事態叙述」は事態成立にとっての必須的な参与者（〈主体〉〈対象〉）や、補助的な媒介物（〈道具〉〈材料〉など）により成り立ち、後者の二つは客観性において差が見られるが、いずれも事態成立の基盤として働いている。このように、事象の構造というのは実は一種の階層構造である。

実は、こういう階層性はさらに事態成立の基盤の内部にも見られる。事態成立の基盤というのは、事態の成立を空間的に位置づけるものであり、そこにも異なるレベルのものが存在している。この点について、中右（1998）は「空間認識構造の普遍的モデル」を提出しているが、本稿ではそれに倣い、そこにはない〈状況〉[4]のデ格をモデルに位置づけるこ

[3] 図 1 でのマークについて紹介してみる。②－1 の「場面設定」の破線の枠はそれが②－2 と本質的には変わらなく、ともに事態成立の基盤であるということを示す。そのほか、意味役割を囲む枠は格成分を表し、矢印は格成分と動詞との係り受け関係を表す。それらの実線と破線の区別はその文中での必須度を示す。〈様態/資格〉をその他のデ格成分と異なり、点線の枠にするのはその副詞性を表す。

[4] 〈状況〉というのは特定の場所、特定の時間帯での物事のありさまなので、本稿ではそれを主観的な

とにより、事態成立の基盤の階層性をデ格との関連で図 2 のように把握してみたい。

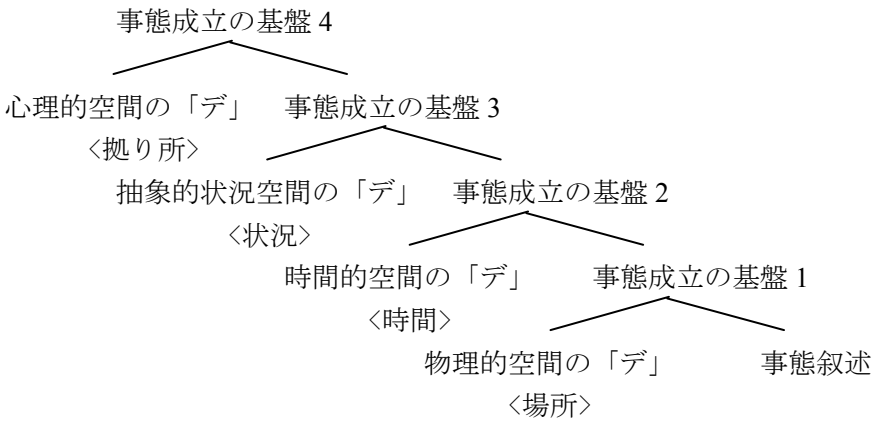


図 2　デ格との関連での事態成立の基盤の階層性

3.2　事象成立での機能によるデ格の分類

本節では、デ格の各用法をその事態成立での機能により、以下のような三種類に分ける。

A. 事態成立の基盤のデ格

つまり、空間次元のデ格のことであり、このグループに属するのは、〈場所〉、〈時間〉、〈状況〉、〈拠り所[5]〉などのデ格があり、それらは文中において、それぞれ「空間」、「時間」、「抽象的な状況」、「話者の把握」など異なる面から事態の成立を位置づける。

そのほか、〈範囲〉〈期間〉〈限定〉などのデ格用法もあるが、それらは〈場所〉〈時間〉の用法からメトニミーリンクにより派生してきたものなので[6]、本稿ではそれらを後者の二次的な用法とする。それらは機能的には文の表す事態や状態の成立の「条件的な基盤」になるので、このグループにも属すべきであろう。

さらに、〈動作主〉のデ格は事態成立の必須的な参与者のようであるが、それは〈動作主〉ガ格のように、行為者となれるすべての名詞について〈動作主〉の意味役割を果たすわけではないし、例⑹のように、文中に〈動作主〉ガ格を加えると、それは〈場所〉に降格してしまうので[7]、そこからは「場所性」が強く感じられる。ゆえに、本稿ではそれを〈場所〉の二次的な用法とし、同じグループに位置づける。

⑹　a. 警察で調べたところうそだと分かった。　（動作主）
　　b. 警察で調査員が調べたところうそだと分かった。　（場所）

B. 事態成立の媒介物のデ格

つまり、モノ・コト次元のデ格のことであり、このグループに属するのは、〈原因〉〈道具〉〈手段〉〈方式〉〈材料〉などのデ格があり、それらは文中においてともに事態成立の背景的な媒介物として働いているが、それを挟んでいる前接名詞の意味素性や述語動詞の文

「話者把握」のした、客観的な「場面設定」の上位に位置づける。

[5] 文頭における「…の調査/統計/考えでは」などのような文の表す事態や状態の情報源を表すものなので、本稿では「拠り所」と言うが、国立国語研究所（1997）ではそれを「陳述」とする。いずれも事態に対する話者の把握であるという点においては共通している。

[6] この点は森山（2002、2004）を参照のこと。

[7] こういう降格現象はすべてのデ格動作主に発生できるわけではないということは後で分析する。

法・意味的な特徴により特定の意味役割として実現する。

そのほか、〈目的〉、〈構成要素/内容物〉などのデ格用法も指摘されているが、それらはそれぞれ〈原因〉、〈材料〉と派生関係にあり、本稿ではそれらを後者の二次的な用法とし、同じグループに位置づける。

C. 事態成立のサマを規定するデ格

つまり、属性次元のデ格のことであり、主に〈様態〉デ格のことを指す。それは事態そのものの成立に直接に関係せず、ただ前接名詞の示される状態で、動作の進行に伴う動作主や対象の様態或いは、動作そのものの進行状態などを規定するだけなので、普通は副詞的な存在として扱われている。

そのほか、〈資格〉のデ格もあるが、それは実は主体や対象の身分的な〈様態〉に当たるので、本稿ではそれを〈様態〉の二次的な用法とし、同じグループに位置づける。

以上、デ格の各用法をその事態成立での機能により捉えてきた。上述の三種類は格助詞「デ」の三つの中核的な用法なので、デ格は実は事態の成立を基盤的な位置づけ、補助的な媒介物、属性規定などの面から規定する背景格であると理解していいであろう。

4. コーパス言語学の立場から見る二重デ格構文の発生

本節では前節のデ格の分類に基づき、二重デ格構文の発生可能性について考察してみたい。まずは、検証例としての二重デ格構文の検索方法について紹介する。

4.1　用例の抽出

本稿では、「現代日本語書き言葉均衡コーパス」中納言により二重デ格構文を検索した。具体的なやり方としては、キーを語彙素の「で」に、その前方共起語を品詞により「大分類―名詞」に指定する。このように一つのデ格名詞句を設定できるようになったが、その上でさらに前方共起語 2 と前方共起語 3 を同じような方法で設定する。このように検出された用例数は実は膨大になるので、筆者は二つのデ格名詞句の間の距離を 5 語以内に設定してみた。

上述の設定条件に合った用例を 40,819 件検出できたが、全て研究対象の二重デ格構文になるわけではない。そのうち、①二つの「で」が同じ文に存在しているのではない、②二つの「で」が同じ述語に掛かっているのではない、③検出された「で」が格助詞ではないなどのような二重デ格ではないものが多く、筆者はまた手作業で整理し、結局は二重デ格構文を 2,371 例検出した。本稿では、この 2,371 例の検出用例に対して定量的な分析を行うのではなく、それらを二重デ格構文の発生可能性への検証例として用いる。

4.2　二重デ格構文の発生の考察

4.2.1　事態成立の基盤のデ格と二重デ格の発生

先述した事象構造や事態成立の基盤の階層性に基いては、二重デ格構文の発生に関して、まずは以下のような仮定を立てる。つまり、

仮定 1：事態成立の基盤のデ格はその他の機能のデ格とは包み包まれの関係にあるので、それらは自由に共起できる。

仮定 2：事態成立の基盤のデ格用法の間にも階層性が存在しているので、それらはお互いに共起できる。

このようにできた二重デ格はいずれも階層性によるものなので、構造的には認められ、

許容度が高いようであるが、紙幅の関係で、ここでは、ただそのうちの〈場所〉デ格を中心に検証してみる[8]。

[1]「場所＋原因」
LBh5_00022 アメリカではオイルショックでディーゼル乗用車が売れ出した。
（沿道汚染　1993）
LBl9_00079 岬の前の海では岩盤が泥で汚れだし、海藻が根腐れを起こして枯れてゆく。
（にっぽん風景紀行　1997）
[2]「場所＋手段」
PB56_00089 東京湾では巻き網で漁獲される。　（相模湾のうまいもん 2005）
OC10_01347 人前では視線で注意を促す事はあります。　（Yahoo!知恵袋　2005）

上記の共起パターンは許容度が高いようであるが、ほとんどは「ハ」により二つのデ格のうちより上位の〈場所〉を主題化したものであり、「ハ」による主題化は実はその前接要素を後続要素から統語的に分断化することでもあるので、こういう階層マーカーの「ハ」なしでも上述の共起パターンが共起できるのか、以下のような用例を見てみよう。例文の後ろの数字はアンケート調査をした上でその許容度を点数化したものである[9]。

(7)
PB43_00053 審理は公開の法廷で、口頭弁論でなされる。　**（1.68）**
（精説不良債権処理　2004）
PB39_00161 車はジョスリン通りとトレメイン通りの交差点で赤信号で停止した。**(1.48)**
（秘密の顔を持つ女　2003）
OM61_00001 この映像のクマは、福井県名田庄村で、有害駆除目的で、おりで捕獲されました。　**（1.23）**（国会会議録　2002）

このように、事象の階層性によるデ格の共起であっても、その許容度はやはり表層的な形式に影響されているようである。上記のデータから見ると、①「ハ」による統語構造の階層化、②読点による二つのデ格名詞句の引き離し、③二つのデ格名詞句の隣接、④デ格名詞句の三つ以上の多重共起の順にその許容度が下がっていく。ゆえに、文の骨格作りに参与しない意味格のデ格であっても、その重複には表層的な制約がかかっていると結論づけていいのであろう。

次は、事態成立の基盤のデ格の各用法同士の共起例であるが、それらも予測のとおり、許容度が高いようである。

[3]「拠り所＋場所」
市民グループの調査では、生ごみ処理器を使っている家庭では、燃えるごみの排出量が

[8] 検証例として使われるのは、基本的には中納言より抽出したものであるが、当該パターンが中納言で検出できなかった場合は、補足として 1998-2000 三年間の朝日新聞（電子版）より検出することにする。
[9] 実例であっても語用などの原因で生じた場合もあり、それらは文法的に必ずしも正しいというわけではないので、本稿では補足として、周辺的な用例に対し許容度調査を行った。例文の後ろのデータはアンケート調査の結果を点数化して記すものである。調査は 2014 年 7 月 11 日に筑波大学で実施し、対象者は学生、総数 44 人である。算定法としては天野（2008）を参考に、自然・少し不自然・全く不自然の三段階の判定結果をそれぞれ 2・1・0 点に換算し平均値を示すものである。

75%も少なかった。（1999.03.14　朝日新聞）

[4]「状況＋場所」

OM11_00011 ロッキード問題をめぐる大変な疑獄、汚職がはやっている中で、福島県で知事が逮捕された。（国会会議録　1976）

[5]「時間＋場所」

関係省庁の話し合いの過程では、通産省と外務省との間で、こんなやりとりが繰り返された。（1999.08.02　朝日新聞）

以上の検証は実例の提示に止まっているが、続いては、〈動作主〉デ格とその他のデ格との共起可能性についても検討してみたい。次のような用例を見てみよう。

[6]「場所＋動作主」

新団体では 10 人ほどの報行部で運営するというが。　**(1.86)**（2000.01.19　朝日新聞）

[6]の例において、二つのデ格がともに〈場所〉と考えられるかもしれないが、「運営する」という述語動詞の情報上の完結性のために、〈動作主〉が必須項として要求されているので、二つのデ格のうちの一つは〈動作主〉として働かなければならなくなる。このように、[6]の二重デ格は実は「場所＋動作主」の組み合わせになっている。

先述したように、〈動作主〉デ格からは「場所性」が感じられるが、そこには実は度合いの差が存在している。こういう場所性の強さにより〈動作主〉デ格は二つの種類に分けられ、それらは二重デ格の発生において異なる振る舞いを見せている。

⑻　a. 警察で調べたところうそだと分かった。
　　b. 警察で調査員が調べたところうそだと分かった。
　　c. 警察では捜査部で調べたところうそだと分かった。
　　d. 中国側では、警察で調べたところうそだと分かった。

例⑻a においてデ格の前接名詞である「警察」は組織・ヒト未分化の機関名詞であり、それは両義性を持っているので、文脈によっては〈動作主〉とも、〈場所〉とも解釈できる。先述のように、そこに〈動作主〉ガ格を加えると、「警察で」は〈動作主〉から〈場所〉に降格するが、c のように、新たに加えられた〈動作主〉がたまたま「捜査部で」のようなデ格成分であったら、「場所＋動作主」という二重デ格が生じる。d は「警察」の前により広い範囲の〈場所〉を加えたもので、降格が発生しないが、「場所＋動作主」の二重デ格になるという点においては c とは変わらない。

⑼　a. 私と佐藤でその問題に取り組んだ。
　　b. わが社では、私と佐藤でその問題に取り組んだ。　**(1.95)**

一方、例⑼において、デ格の前接名詞が複数の個体であり、それは人間の集合を組織扱いしているので、場所性が多少感じられるが、究極的に〈場所〉ではないので、そこに〈動作主〉ガ格を加えても、もとの〈動作主〉が〈場所〉に降格できない。ゆえに、ここに二重デ格が発生できるのは b のような〈場所〉を新たに加える場合に限られている。

続いては、〈動作主〉デ格と〈手段〉〈様態〉のデ格との共起についても論じてみる。

[7]「動作主＋様態/手段」
同センターによると、各会場で試験開始と同時に板書で訂正したという。
（2000.01.07　朝日新聞）
OY07_02613 ギブスも取れないようだし、生活に不便があるでしょう。娘と私で交代で泊りに行くかなあ…　**(1.84)**　（Yahoo!ブログ　2008）

以下の作例もそれらの共起を証拠付けられるであろう。

⑽ a. 警察で調べたところうそだと分かった。
b. 警察でウソ発見機で調べたところうそだと分かった。　（動作主＋道具）
c. 警察でＤＮＡ鑑定で調べたところうそだと分かった。　（動作主＋方式）
d. 警察で身分秘匿で調べたところうそだと分かった。　（動作主＋様態）

以上は〈場所〉とその二次的な用法である〈動作主〉デ格を中心に、事態成立の基盤のデ格の二重デ格の発生可能性について考察してきた。そこに生じた二重デ格はほとんど階層性によるものなので、許容度が比較的高い。

4.2.2　事態成立の媒介物のデ格と二重デ格構文の発生

本節では、事態成立の媒介物というモノ・コト次元のデ格の共起可能性について検討してみたい。考察に入る前に、まずは二次的な用法とされている〈目的〉〈構成要素/内容物〉のデ格とその元用法になる〈原因〉〈材料〉のデ格との共起可能性について説明してみたい。

⑾　この飛行機は木と薄い鉄板でできている。

例⑾において、「木と薄い鉄板で」は〈材料〉とも〈構成要素/内容物〉とも解釈できるように、この三者は本質的には同一の用法である。それらの区別というと、かかっている述語動詞の意味特徴にあるようである。つまり、
①　材料名詞＋デ＋作成動詞　　○　大麦で酒をつくる　　（材料）
②　構成要素＋デ＋構成動詞　　○　うそと虚飾で成り立っている世界　（構成要素）
③　内容物＋デ＋充満動詞　○　店内はその八割ほどが客で埋まっていた。　（内容物）
このように、三者はそれぞれ特定の述語動詞にしか係らないので、それらの一文中での共起はまず有り得ないであろう。〈原因〉と〈目的〉もこういう関係にあり、それらの一文での共起も許されない。このように、本節では主に〈原因〉〈道具〉〈手段〉〈方式〉〈材料〉の間の共起可能性について考察することにする。

事象構造から見ると、「何のために誰が何を使ってどういう方法で何をやった」のように、〈原因〉〈道具〉〈方式〉などの意味役割は一文中に共起できないわけでもないようであるが、それらを同時にデ格で表していいのか、さらに検討する必要がある。

[8]「原因＋手段」
PB42_00026 旅人に生水は禁物だということでペプシコーラで代用した。　**(1.75)**
（沙漠の旅　2004）
LBo4_00044 その五日前に、突然の昏睡で救急車で入院してきた。（臨床に吹く風　2000）

例⑿は先述した異なった文法性判断が下されていたもので、ここではそれらを許容度の点数付けで挙げることにする。

⑿ a. *列車事故でバスで振り替え輸送を行った。 **(1.16)** （杉本 2013：39）[10]
b. 鉄道ストでバスで登校した。 **(0.93)** （矢澤 2007：245）

このように、そのいずれの許容度も高くないようで、〈原因〉と〈手段〉の共起はやはり許されないのであろう。ところが、[8]の PB42_00026 は高い許容度を示しているのはなぜなのであろうか、それは「…ことで…で」という形式とは関係があると考えられる。杉本（2013）は〈原因〉の「ことで」は文法的に格助詞よりは接続助詞に近い性質を持っていると指摘しており、本稿で観察された上述の現象もまさにそのためであろう。

[9]「道具＋方式」/「材料＋道具」
LBo5_00015 洋服をドラム式洗濯機でお湯と洗剤で洗い、乾燥機で素早く乾燥させる。
(1.45) （はて・なぜ・どうしてクイズ石けんと合成洗剤　2000）
OC08_01823 ブリタで作った浄水で、別の容器で麦茶を作るのが良いのではと思います。
(1.39) （Yahoo!知恵袋　2005）
？息子は色鉛筆で遠近法で絵を描いています。 **(1)** （作例）

このように、事態成立の媒介物のデ格の一文中での共起は基本的には許されないようであるが、コーパスからこのような用例を少なからず検出できたのはなぜなのであろうか。実は、そこに語用的な要素が存在しているのであろう。このタイプの検出例から付け加えの読みが強く感じられ、話し手はまさに新しい情報の付け加えの意図で、二つのデ格の間に心理的・音声的なスペースを作り、上記のような構文を出したのであろう。この点は実はこれらの意味役割の機能上の同一性の証拠にもなる。

4.2.3　事態成立のサマを規定するデ格と二重デ格構文の発生

先行研究において、こういう属性次元のデ格が〈様態〉デ句とされることがあるように、それは統語的には格の振る舞いをしているが[11]、機能的にはただ事態がどのようなさまで成立したかということに対して補足的に修飾するだけなので、そこからは副詞性が強く感じられる。まさにこういう副詞性のため、〈様態〉デ格はその他のデ格成分とは自由に共起できるようである。

[10]「場所＋様態/資格」
LBl1_00011 法律案は十三日の衆議院本会議で賛成多数で可決され、参議院に送付されました。
（改正宗教法人法　1997）
PB53_00409 赤軍軽井沢山荘事件の被告達は、国民の目の前で殺人の現行犯で逮捕された。
(1.66) （台湾を独立させよう　2005）
[11]「手段＋様態」
LBl9_00153 大名たちが、火事装束で早馬で門を走り抜けていった。

[10] 文頭につけた文法性判断のマークは著者によるものである。
[11] 李（2008）は語順の転換、とりたて詞「だけ」の挿入などのテストにより、この「状態記述 2 次述部」の「で」は格助詞であることを論述した。本稿でもこの点に賛同し、それを副詞性の強い格助詞とする。

（夜明け前の女たち　1997）

LBm7_00031 前景はやや荒っぽいタッチで強めの墨色で描いている。**（1.68）**

（墨絵独習書　1998）

[12]「様態＋様態」

LBk9_00009 中に白装束で重力のない足どりでやってくる宮廷のバラモン階級の僧侶ペダンダたちと、暖かいまなざしの老人である村…　（旅のはざま　1996）

OB5X_00230 彼女は困り切った表情で、小声で尋ねた。　（新・人間革命　1998）

このように、ほとんどは二つのデ格名詞句の隣接になるが、許容度はそれほど低くない。これもこういう属性次元のデ格の副詞性とは関係があるのであろう。つまり、〈様態〉デ格が機能的には副詞的な存在なので、それと他の機能のデ格と隣接しても格の衝突がそれほど強く感じられない。そのほか、[12]のように、修飾成分の並列や付け加えになるが、〈様態〉自身の二重共起も容易に発生できるようである。

5. 終わりに

以上はコーパスからの検索用例に基づき、二重デ格構文の発生可能性を事象の構造の面から考察してきた。以下のようにまとめられる。

[1] 事象構造において異なる機能を果たすデ格用法はお互いに共起できる。そのうち、特に階層性により支えられているものは、許容度が比較的高い。

[2] 事象構造において同じ機能を果たす意味役割同士は一文中に共起できるようであるが、事象の階層性や付け加えなどの語用の場合を除いては、同時にデ格により表せない。

[3] 属性次元のデ格成分は副詞性の強いものなので、それがその他のデ格との共起は言うまでもなく、それ自身の二重共起も容易に発生できる。

[4] 意味格のデ格であっても、その重複に表層的な制約がかかっている。

なお、紙幅の関係で、本発表では同じ意味役割の共起可能性については論じられなかったため、この点をこれからの課題として残したい。

参考文献

天野みどり（2008）「拡張他動詞文―『何を文句を言ってるの』―」『日本語文法』8 巻 1 号、pp.3-19

国立国語研究所（1997）『日本語における表層格と深層格の対応関係』三省堂

杉本　武（2013）「原因の『～で』と『～ことで』について」『文藝言語研究　言語篇』63 巻、pp.37-52、筑波大学文藝・言語学系

田中春美（1987）『格文法の原理―言語の意味と構造』　三省堂

中右　実（1998）『構文と事象構造』　研究社出版

日本語記述文法研究会（2009）『現代日本語文法 2』　くろしお出版

益岡隆志（1987）『ケーススタディ日本文法』　桜楓社

森山　新（2002）「知的観点から見たデ格の意味構造」『日本語教育　巻 115』、pp.1-10

森山　新（2004）「格助詞デの放射状カテゴリー構造と習得との関係」『日本認知言語学会論文集』4 巻、pp.66-76

李　昇祐（2008）「状態記述 2 次述部と『で』」『日本語と日本文学』47 号、pp.70-80

矢澤真人（2007）『日本語状態修飾関係の研究』　筑波大学博士学位論文

文体指標と語彙の対応分析

浅原 正幸 †　加藤 祥 †　立花 幸子 †　柏野 和佳子 ‡†
（国立国語研究所　† コーパス開発センター　‡ 言語資源研究系）

Correspondence Analysis between Writing Styles and Lexicon

Masayuki Asahara, Sachi Kato, Sachiko Tachibana, and Wakako Kashino
(National Institute for Japanese Language and Linguistics)

要旨

柏野 (2013) は文体を分類するために設計した指標として，専門度，客観度，硬度，くだけ度，語りかけ性度の 5 種の分類指標を提案し，現代日本語書き言葉均衡コーパス (BCCWJ) の図書館サブコーパス 10,551 サンプルに対して悉皆的に付与を行った。本研究では，この分類指標に対して語彙素を特徴量とした対応分析を行い，各指標と特徴的な語彙素の分布の対応を品詞ごとに定量的に評価する。対応分析によってもたらされる第 1 主成分の寄与度に基づき，語彙素の分布のみによってとらえることが可能な指標とそうでない指標を明らかにする。さらに，語彙素の分布のみによってとらえられない指標については，どのような語彙素以外に利用すべき特徴量を検討する。

1. はじめに

コーパス調査において重要な要素として，利用するサンプルの文体情報がある。柏野 (2013), 柏野・奥村 (2012b) は文体を計量する指標として，専門度，客観度，硬度，くだけ度，語りかけ性度の 5 種の分類指標を提案し，『現代日本語書き言葉均衡コーパス』(BCCWJ) の図書館サブコーパス 10,551 サンプルに対して悉皆的に付与を行った。このデータに対して，硬度・語りかけ性度を中心に，定量的・定性的な分析が進められてきた (柏野ほか (2012a), 保田ほか (2012b,a,c, 2013d,a,c,b), 加藤ほか (2015))。

本研究では統計的手法に基づいて，指標ごとに特徴的な語彙素を選定することを試みる。具体的には品詞ごとに高頻度語彙素を抽出し，指標に基づく対応分析を行い，語彙と指標との対応関係を分析する。この指標ごとに特徴的な語彙素の分布を確認し，語彙素のみにより表現可能な指標とそうでない指標を明らかにする。さらに，現在までに行われてきた定性的な分析と今回行った統計的な分析との比較を行う。

本研究の貢献は以下の通りである。各文体指標に対する人間の判断が，単純な語彙素の統計的な偏りにより表現可能な指標とそうでない指標を明らかにする。さらに語彙素の分布によってとらえられない指標については，語彙素以外に利用すべき特徴量を検討する。

2. 分析手法

2.1 文体指標

柏野 (2013) は文体指標として以下の 5 種類を規定した：

- 【専門度】: 1 専門家向き，2 やや専門的な一般向き，3 一般向き，4 中高生向き，5 小

学生・幼児向きの 5 段階指標

- 【客観度】: 1 とても客観的，2 どちらかといえば客観的，3 どちらかといえば主観的，4 とても主観的の 4 段階指標
- 【硬度】: 1 とても硬い，2 どちらかといえば硬い，3 どちらかといえば軟らかい，4 とても軟らかいの 4 段階指標
- 【くだけ度】: 1 とてもくだけている，2 どちらかといえばくだけている，3 くだけていないの 3 段階指標
- 【語りかけ性度】: 1 とても語りかけ性がある，2 どちらかといえば語りかけ性がある，3 特に語りかけ性はないの 3 段階指標

対象は BCCWJ に収録されている図書館サブコーパス 10,551 サンプル（書籍サンプル）とし，20〜50 代女性作業者延べ 9 名に可変長サンプルを呈示して文体指標付与を行った。作業において，インタビューなどのテキスト構造が文体付与に適さないものや外国語や数式などが多いサンプルなど内容や表現が文体付与に適さないものなど 1,664 サンプルを，文体指標付与対象から除外している。

2.2　対応分析

対応分析はクロス集計表の行と列の双方を並び替えることにより，行の項目と列の項目との相関関係を最大化するような処理を行う。基本的には主成分分析と同様にデータの分散を最大化する方向の軸（主成分）を逐次的に求め，説明変数を合成するという処理を行う。軸の選択は条件付極値問題として定式化でき，ラグランジュの未定乗数法によって解くと相関行列の固有値，固有ベクトルを求める問題に帰着する。値の大きい固有値に対応した軸から順に第 1 主成分，第 2 主成分と呼び，各軸は直交する。全固有値の総和で，各主成分に対応する固有値を割ったものを寄与率と呼び，各主成分によりどの程度説明ができているかの尺度となる。同様に第 1 主成分から第 α 主成分までの寄与率の和を第 α 主成分までの累積寄与率と呼び，当該主成分まででどの程度説明ができているかの尺度となる。

2.3　特徴量の設計

語彙素と品詞（細分類）の 2 つ組を 1 特徴量として設定する。先行研究で多く言及されている品詞大分類（動詞，連体詞，副詞，助動詞，助詞）ごとに頻度順に 30 語を抽出し特徴量とし，各指標に対して有効な語彙素と品詞を明らかにする。人手による指標の付与は可変長サンプルの地の文のみに対して行われた。対応分析は固定長サンプルと可変長サンプルの両方について行ったが，地の文と台詞の区別は行わず，全ての語彙について調査した。

固定長サンプルは文単位で 1000 字前後を選定するというランダムサンプリングに基づいたものであり，統計処理的にはより厳密なものである。一方，元の人手による指標の付与は文章構造上まとまった単位に行っており，実際のアノテーションにおいても文章構造に基づいた判定が行われている。

3.　結果（可変長サンプル）

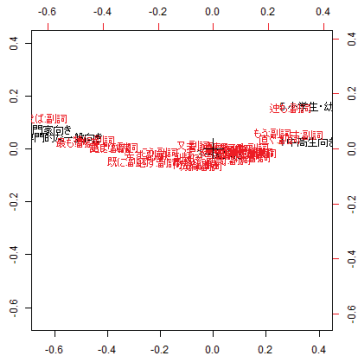
対応分析は固定長サンプルと可変長サンプルの両方について実施した。以下では可変長サンプルの分析結果について，指標ごとに特徴的な結果が得られた品詞や先行研究に多く言及され

ている品詞のみを示す。固定長サンプルの結果および可変長サンプルの他の品詞の結果については，第一著者に問い合わせていただければ提供する。

3.1　専門度

専門度については，ほぼすべての品詞において，統計的な手法によって得られた語彙と人手による判断との間に一致が見られた。以下では 5 品詞の中でもっとも第 1 主成分の寄与度が高かった副詞について言及する。

副詞



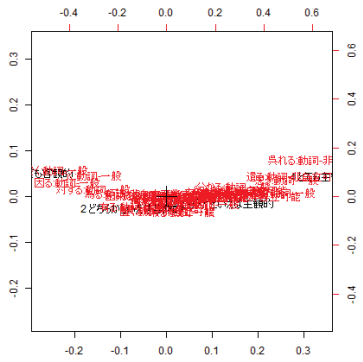
基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。『1. 専門家向き』『2. やや専門的な一般向き』間が第 1 主成分方向の差分が少なく，『4. 中高生向き』『5. 小学生・幼児向き』間は第 1 主成分方向の尺度が逆転しているが，第二成分方向に差分があり，識別可能である。第 1 主成分の寄与度は 97.1% である。第 1 主成分方向に見ると「例えば」「最も」が『1. 専門家向き』の語彙，「ずっと」「もう」が『5. 小学生・幼児向き』の語彙であることが確認できる。

佐藤・柏野 (2012) は，今回用いた専門度ではなく obi/B9 難易度 (佐藤 (2011)) を用い，4 つのテキストを難易度順に並び替える被験者実験結果と obi/B9 との比較を行っている。被験者実験においては，1000 文字程度に揃える，テキストの NDC が同じものにするなどの工夫が見られる。一方，定量的な評価に終始しており，定性的な分析があまりおこなわれていない。

3.2　客観度

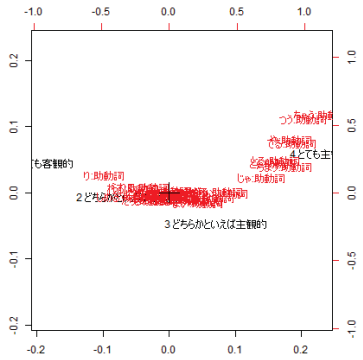
客観度については，保田ほか (2013d) において語りかけ性度との 2 軸の定性的な分析が行われている。以下では動詞・助動詞についての結果を示す。

動詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。第 1 主成分の寄与度は 97.0% である。「於く」「つく」が『1. とても客観的』の語彙，「遣る」「呉れる」「聞く」が『4. とても主観的』の語彙であることが確認できる。「遣る」「呉れる」については，保田ほか (2013d) の p.151 参考図において言及されている。

助動詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。第 1 主成分の寄与度は 91.7% である。「り」が『2. どちらかといえば客観的』の語彙，「ちゃう」「てる」「ちまう」「てく」が『4. とても主観的』の語彙であることが確認できる。一方，保田ほか (2013d) では，『4. とても主観的』群に特徴的な語として「たい」をあげているが，本分析では『3. どちらかといえば主観的』あたりに位置している。作業者は「たい」を直接主観度の高い表現とみなしているわけではなく，主観度の高い文書に頻出する「個人的な考え・感情を述べる表現」に頻出する表現としての助動詞「たい」あげていることから，「個人的な考え・感情を述べる表現」でない「たい」の影響があるのではないかと考える。

保田ほか (2013d) においては，主観度毎に出現頻度の高い表現をいくつかあげている。主観度の高い表現として，「個人的な考え・感情を述べる表現」のほかに「主体の経験を表す表現」や「根拠のない断定」をあげている。客観度の高い表現として，「裏付けがある表現」・「数値データの提示」・「伝聞の使用」をあげている。一旦，これらの表現を想定したうえで，各表現の言語の表層的な特徴を考えるという二段階の処理を経て，以下のような特徴をあげている：

主観度が高い事例の特徴

- 数値表現が少ない
- 主体の経験をあらわす受影受動文 (益岡 (1991))
- 伝聞のモダリティを表出しない複合辞「という」

客観度が高い事例の特徴

- 数値表現が多い
- 降格受動文 (益岡 (1991))
- 伝聞のモダリティを表出する複合辞「という」

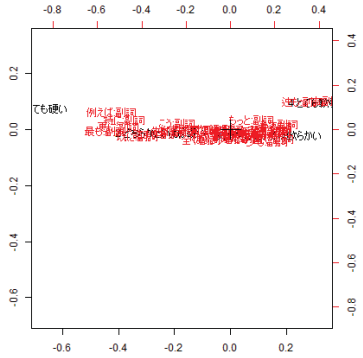
これらの特徴のうち，数値表現については固有表現抽出などの技術や数詞の割合なので判定可能であるが，残りの表現については用法の判定が必要となり何らかの人手によるアノテーションが必要になる。

しかしながらここで重要なのは，作業者のアノテーションにおいては先に述べた表現を探すことを前提として，表層的な特徴をあげている点にある。指標と表現と特徴の 3 つ組を想定してモデル化しなければ，本質的に解けない問題であると考えられる。

3.3　硬軟

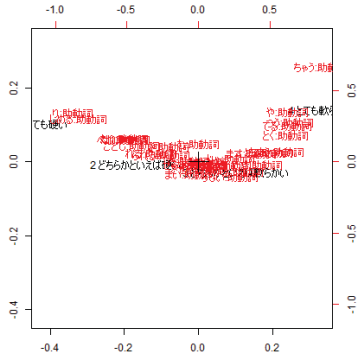
柏野ほか (2012a) は硬軟とくだけ度の 2 軸での定量・定性的な分析を行った。以下では副詞と助動詞と助詞についてとりあげる。

副詞



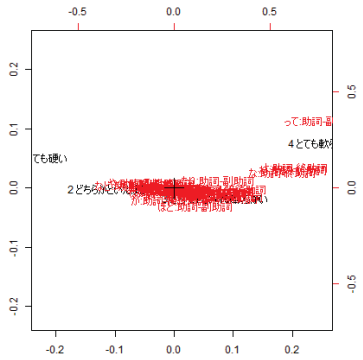
基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。第 1 主成分の寄与度は 97.7%である。第 1 主成分方向に見ると「例えば」「更に」「最も」が『2. どちらかといえば硬い』の語彙，「迚も（とても）」「ずっと」「一寸」が『4. とても軟らかい』の語彙であることが確認できる。これに対し，柏野・奥村 (2012b) の p.162 表 2 では，「いかに」「より」などの副詞が硬い文書の特徴的表現として挙げられているが，今回処理した頻度上位 30 語にこれらは入っていなかった。柏野・奥村 (2012b) ではさらに，硬い文書では副詞出現頻度が低く，軟らかい文書では副詞のバリエーションが豊富であることについて言及している。

助動詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。第 1 主成分の寄与度は 92.7%である。第 1 主成分方向に見ると「り」「しめる」が『1. とても硬い』，「ちゃう」「つう」「や」「てる」が『4. とても軟らかい』の語彙であることが確認できる。一方柏野・奥村 (2012b) で言及されている，断定（硬い表現）や「です・ます」（軟らかい表現）などの表現については特別な統計的偏りが見られなかった

助詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。第 1 主成分の寄与度は 95.0%である。「など : 助詞-副助詞」「や : 助詞-副助詞」が『2. どちらかといえば硬い』の語彙，「ね : 助詞-終助詞」「よ : 助詞-終助詞」「って : 助詞-副助詞」が『4. とても軟らかい』の語彙であることが確認できる。柏野ほか (2012c) は，硬い印象を与える特徴として終助詞がほとんど出現しないことを，軟らかい印象を与える特徴として，特徴的な副助詞（〜か，たり，や，まで）の存在をあげている。前者についてはグラフから読み取れるが，後者についてはグラフから読み取れなかった。

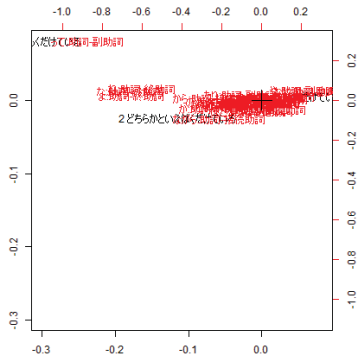
柏野ほか (2012a) は硬度判定の参考情報としていくつかの例をあげている。硬い印象を与える特徴として，「だ・である」調，断定・定義の文，抽象物が主語の受身文，硬い印象を持つ接続詞・副詞の語彙，親密度の低そうな語の頻度，難解な内容，疑問・回答が対応しているなどをあげ，軟らかい印象を与える文の特徴として「です・ます」調，語りかける文末表現，軟らかい印象を持つ接続詞・副詞，平易な語の頻度，平易な内容などをあげている。このうち，今回統計処理で扱っている語彙素のみにより表出するのは一部の文末表現と接続詞・副詞のみである。

また，柏野ほか (2012c) の p.162 表 2 では硬度判定の参考情報がまとめられている。硬い印象を与える特徴として，数詞が多い，文語助動詞などを上記のほかにあげている。

3.4　くだけ度

以下では助詞についての統計処理結果について示す。

助詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布しているが，語彙はほぼ中央に分布している。第 1 主成分の寄与度は 96.1% である。「って : 助詞-副助詞」「から : 助詞-接続助詞」「か : 助詞-終助詞」が『1. とてもくだけている』の語彙，「から : 助詞-格助詞」は『2. どちらかといえばくだけている』の語彙，「や : 助詞-副助詞」「の : 助詞-格助詞」「など : 助詞-副助詞」が『3. くだけていない』の語彙であることが確認できる。

柏野ほか (2012a) は，くだけた印象を与える特徴として以下のものをあげている：

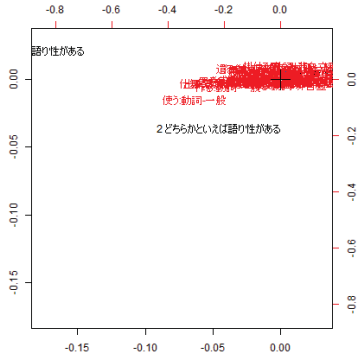
- 体言止めや述語修飾がある
- 一人称が主語の文が多い
- 平易な語に加え，俗語がある
- 卑近な内容や説明である
- オノマトペが多い
- 感覚や感情表現が多い
- 音変化（拗音化・撥音化など）の語がある
- 回答のない，いいっぱなしの疑問文がある

語彙素からは得られにくいものが多く，特に音変化は表層形からは得られるが，語彙素を用いたがために得られない特徴量になっている。

3.5　語りかけ性度

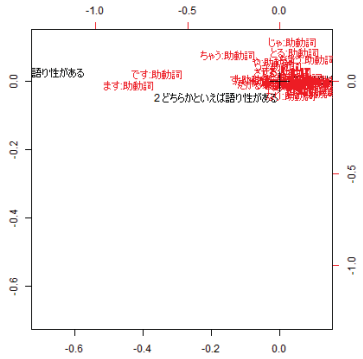
語りかけ性度については，動詞と助動詞を示す。

動詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。「使う」「作る」が『2. どちらかと言えば語りかけ性がある』付近に分布し，それ以外の語彙は「3. 語りかけ性がない」付近に分布している。第 1 主成分の寄与度は 93.7% である。

助動詞



基本的に第 1 主成分（横軸）方向に指標の尺度順に分布している。「です」「ます」が『2. どちらかと言えば語りかけ性がある』付近に分布し，それ以外の語彙は「3. 語りかけ性がない」付近に分布している。第 1 主成分の寄与度は 99.3% である。

語りかけ性度についても，単純な語彙素による対応分析においてとらえにくいことがうかがえる。語りかけ性については，加藤ほか (2015) が詳細な論考を示している。語彙については，与えられた語りかけ性度に基づいてカイ二乗値を分析し，それぞれの群について有意差のある語を定量的に分析しているほか，実作業者のコメントに基づいて頻度によらない「語りかける」という印象を表出する以下のような特徴的な表現について言及している：

- 昔話に特徴的な表現
- 教示的表現（希望・注意・禁止・勧誘・可能・評価）
- 特定の人称表現（「私たち」「われわれ」）
- 文末表現（「のである」「わけです」「からです」「ものです」，読み手を想定した表現，婉曲表現，読み手の判断を想定した表現）

これらは「語りかける」群にのみ出現率の高い表現ではないか，コーパス全体において出現数が少ないために統計処理によって表出しにくいものであると言及している。

3.6　第 1 主成分の寄与度

表 1 に固定長サンプルの対応分析における第 1 主成分の寄与度を，表 2 に可変長サンプルの対応分析における第 1 主成分の寄与度を示す。可変長サンプルが作業者が見ているサンプルと過不足なく一致するために，統計処理により表出する寄与度も一般的に可変長サンプルの方が高い（表中太字）傾向にある。得られた語彙をみると，寄与度が高い品詞空間において，特別説得力のある語彙素の分布が得られていないこともわかった。

表 1　指標と第 1 主成分の寄与度（固定長サンプル）

	専門度	客観度	硬度	くだけ度	語りかけ性度
連体詞	**91.7%**	96.2%	96.4%	**98.1%**	80.2%
副詞	95.8%	96.2%	96.5%	96.4%	89.7%
助動詞	87.4%	90.3%	92.5%	**88.5%**	99.0%
助詞	91.5%	94.6%	94.4%	**96.4%**	90.7%
動詞	95.1%	96.7%	**97.9%**	**98.8%**	**94.0%**

表 2　指標と第 1 主成分の寄与度（可変長サンプル）

	専門度	客観度	硬度	くだけ度	語りかけ性度
連体詞	**91.7%**	**98.2%**	**97.9%**	98.0%	**91.4%**
副詞	**97.1%**	**97.0%**	**97.7%**	**97.8%**	**92.7%**
助動詞	**88.1%**	**91.7%**	**92.7%**	87.9%	**99.3%**
助詞	**92.9%**	**95.7%**	**95.0%**	96.1%	**90.9%**
動詞	**95.7%**	**97.0%**	97.5%	98.7%	93.7%

4. 考察

専門度・客観度については対応分析によって得られた語彙が先行研究の定性的な分析や作業者の判断基準などと一致する傾向がみられた。一方，硬度・くだけ度については，一部の特徴的な語彙素が得られているが，必ずしも作業者の判断基準としていた参考情報と一致しない部分も見られた。さらに語りかけ性度については語彙素のみによってはとらえられず，アノテーション作業においてより高度な認知的な判断が行われていることが示唆された。

以下，語彙素以外でどのような特徴量を含めるべきかについて検討する。

語彙の印象評定：感覚・感情表現や，語彙そのものの硬軟・くだけ度などが文体判定に用いられている。これらの語彙の分類は単純な形態素解析結果のみからは情報が得られない。単語親密度 (天野・近藤 (1999)) や分類語彙表 (国立国語研究所 (2004)) など様々な語彙分類があるが，文体指標に特化した語彙表の構築が必要である。

オノマトペ・語彙の音変化：オノマトペや語彙の音変化（拗音化・撥音化）など音声言語としての印象が文体に影響を与えている。今回の語彙素による分析では表記ゆれなどを正規化して分析しており，このような音変化が消された分析になっている。発音などの情報を含めることが必要である。

文末表現：特定の文末表現が指標の判定に影響を与えることが示唆されている。文末からの N-gram などを特徴量として含めることにより，指標判定の性能向上がはかれると考える。丸山 (2012a,b) はレジスターごとの文末表現（文末からの N-gram）の傾向を分析調査している。文末表現の構造として，動動詞や終助詞が表出する文法カテゴリ（ヴォイス，アスペクト，肯否，テンス，モダリティ）に注目しているが，実際には 文字 N-gram を展開したうえで，モダリティについての対人（依頼・質問）・対時（禁止・義務・許可）についての出現率の分析を

行っている。しかしながら，BCCWJ のコアデータにはこれらの多様な文法カテゴリの人手によるアノテーションが進められており (松吉ほか (2014), 4.1 節)，これらを用いて分析することが考えられる。

統語的な用法分類：態の用法（受影受動-降格受動）や複合辞が表出するモダリティなどが指標の判定に用いられている。しかしながら，現状のコーパスアノテーションにおいては，このレベルの情報の網羅的な情報付与がされていない。

文の談話的機能：1 文の談話的機能や複数文の談話的機能の推移が文体に影響を与えている。文の談話機能としては，希望・注意・禁止・勧誘・可能・評価・断定・定義・疑問・回答などがあげられている。一部については統語レベルのアノテーション，もしくは言語処理で用いられる応用よりのアノテーションから情報を得られるが，文体そのものを評価するための談話的機能の情報が付与されていない。

内容：文章に含まれる内容そのものが文体に影響を与える。これらについては NDC や C コードなどメタデータの情報から得ることが可能であり，先行研究では NDC や C コード別の分析が多く行われている。一部の指標で NDC と C コードごとの指標の分布の偏りが観察されるが，多様な文体が混在する分類もあり，これらのメタデータが適切な内容分類の粒度かどうかについては議論の余地がある。

各特徴量については，現状の言語資源や言語処理技術により実現可能なものもあるが，何らかのアノテーションが必要なものも多くある。一方で，指標がアノテーションされている状況から，指標から特徴的な構造・用法を統計的に抽出する方法も検討する必要がある。

5. おわりに

本稿では，現代日本語書き言葉均衡コーパス (BCCWJ) の図書館サブコーパス 10,551 サンプルに対して付与された専門度，客観度，硬度，くだけ度，語りかけ性度の 5 種の分類指標に対して，品詞ごとに語彙素を特徴量とした対応分析を行い，先行研究で言及されている定量的・定性的分析との比較調査を行った。語彙素のみに基づく手法では，作業者によって認知される文体指標を部分的にしかとらえられないことがわかった。

識別的な手法を用いて，特定の指標のみに表出する低頻度の語彙素については特徴的な語彙素に対して分類を付与することは可能であろう。しかしながら，全体に同じ表現が偏在するが，用法の違いにより表出するような指標については，表層に基づく単純な手法ではとらえられない。今後，既存のアノテーションでとらえられる特徴，新たにアノテーションを行うことでとらえられる特徴などを調査する。さらに，隠れ変数を立てた統計モデルを用いることにより，いままでとらえられなかった特徴量がとらえられるのかについて分析していきたい。

謝辞

本研究の一部は国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

天野成昭・近藤公久 (1999).『NTT データベースシリーズ日本語の語彙特性』　三省堂　1 巻.

柏野和佳子・立花幸子・保田祥・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織 (2012a).「テキストの硬さと柔らかさの考察–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　第 1 回コーパス日本語学ワークショップ, pp. 131–138.

柏野和佳子・奥村学 (2012b).「書籍テキストへの分類指標人手付与の試み–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　言語処理学会第 18 回年次大会, pp. 1260–1263.
柏野和佳子・立花幸子・保田祥・飯田龍・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織・椿本弥生・沼田寛 (2012c).「書籍テキストへの文体情報付与の試み–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　第 1 回コーパス日本語学ワークショップ, pp. 155–164.
柏野和佳子 (2013).「書籍サンプルの文体を分類する」　国語研プロジェクトレビュー, 4:1, pp. 43–53.
加藤祥・柏野和佳子・立花幸子・丸山岳彦 (2015).「語りかける書きことばの表現」　国立国語研究所論集 (印刷中), 8.
国立国語研究所 (2004).『分類語彙表増補改訂版』　大日本図書.
丸山岳彦 (2012a).「『現代日本語書き言葉均衡コーパス』を用いた文末表現のバリエーション」　言語処理学会第 18 回年次大会発表論文集, pp. 591–594.
丸山岳彦 (2012b).「『現代日本語書き言葉均衡コーパス』を用いた文末表現のバリエーション (2)」　第 2 回コーパス日本語学ワークショップ, pp. 207–214.
益岡隆志 (1991).『受動表現と主観性（『日本語のヴォイスと他動性』）』　pp. 105–121.　くろしお出版.
松吉俊・浅原正幸・飯田龍・森田敏生 (2014).「拡張 CaboCha フォーマットの仕様拡張」　第 5 回コーパス日本語学ワークショップ, pp. 223–232.
佐藤理史 (2011).「均衡コーパスを規範とするテキスト難易度測定」　情報処理学会論文誌, 52:4, pp. 1777–1789.
佐藤理史・柏野和佳子 (2012).「テキストの難易度に対する人間の判断と機械の判断」　第 1 回コーパス日本語学ワークショップ, pp. 195–202.
保田祥・柏野和佳子・立花幸子・丸山岳彦 (2012a).「「語りかけ性」を有すると判断される書きことばの表現」　第 2 回コーパス日本語学ワークショップ, pp. 43–50.
保田祥・柏野和佳子・立花幸子・丸山岳彦 (2012b).「「語り性」を有する書き言葉の典型例の分析」　第 1 回コーパス日本語学ワークショップ, pp. 139–146.
保田祥・柏野和佳子・立花幸子 (2012c).「総体として印象を与える表現 :「語りかけ性」を有すると判断する根拠」　人工知能学会第 41 回ことば工学研究会.
保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013a).「「ベテランは足を保護する」が語りかけるとき」　第 4 回コーパス日本語学ワークショップ, pp. 345–354.
保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013b).「アノテーターコメントを用いた「語りかけ性」分析の試み–頻度情報から捉え難いテキスト性質の解明に向けて–」　言語処理学会年次大会発表論文集.
保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013c).「語りかけると判断される文体–大規模コーパスを用いた特徴的表現の分析–」　日本文体論学会第 104 回大会.
保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013d).「書きことばにおける「語りかけ」は何のため用いられるのか」　第 3 回コーパス日本語学ワークショップ, pp. 143–152.

平安初期歌合和歌の品詞比率

冨士池　優美（中央大学）†

Part of Speech Ratio of Japanese Poetry of *Utaawase* in the Early Heian Period

Yumi Fujiike　(Chuo University)

要旨

和歌の品詞比率については、限られた音数の中での表現が求められるため名詞の比率が高いことが知られている一方で、一つのジャンルの中にも品詞比率にばらつきがあることも指摘されている。平安初期和歌のうち歌合と勅撰集を対象とし、『日本語歴史コーパス平安時代編』『歌合コーパス』の長単位データに基づく名詞率と MVR を用いて、和歌の内容の違いの品詞比率との関係を検討した。その結果、恋歌、季節歌といった和歌の内容により、品詞比率に差があることが明らかになった。また、散文との比較から、今回調査対象とした和歌のテキストの特徴は「要約的な文章」として位置づけられ、和歌の内容による品詞比率の差は文章のジャンルを超えるものではないことが明らかになった。

１．はじめに

文章のジャンルによって品詞の割合が決定されると考えられている。その中で、短歌や俳句といった形式は、限られた音数の中での表現が求められるため名詞の比率が高いことが知られている[1]。その一方で、菅原優美（2003）では、勅撰集（八代集）、中古散文資料、私家集、歌合の和歌を対象に品詞比率を調査した結果、勅撰集の語彙が限定されており一様であるのに対し、散文資料・私家集・歌合の和歌語彙は多様であることを示した。つまり、和歌という一つのジャンルの中にも、品詞比率にばらつきがあるということになる。しかし、八代集、中古散文資料は年代の幅が広く[2]、品詞比率の違いが生じる要因が年代によるものなのか、詠まれた場や内容によるものなのか、わかりにくいところがあった。

本発表では、平安初期の主要歌合 3 作品の和歌を対象とし、同時代の勅撰和歌集である『古今和歌集』所収歌と比較することで、品詞比率からみた和歌の特徴を明らかにすることを目的とする。調査対象とするコーパスは、『日本語歴史コーパス　平安時代編』と、「中古中世歌合コーパスに基づく和歌評論の語彙論的研究」（研究課題番号：25770179）で構築中の『歌合コーパス』である。「長単位」に基づく名詞率と MVR（100×相の類の比率／用の類の比率）を用い、和歌の内容の違いと品詞比率との関係を見出す。

† fujiike@tamacc.chuo-u.ac.jp

[1] 樺島忠夫(1979)

[2] 八代集でいうと、『古今和歌集』から『新古今和歌集』まで、約 300 年の開きがある。

２．調査対象

２．１　使用するコーパス

（１）日本語歴史コーパス　平安時代編

2014 年 3 月、国立国語研究所で構築された『日本語歴史コーパス　平安時代編』が公開された。収録作品は中古和文 14 作品（竹取物語、古今和歌集、伊勢物語、土佐日記、大和物語、平中物語、落窪物語、枕草子、源氏物語、紫式部日記、和泉式部日記、更級日記、堤中納言物語、讃岐典侍日記）である。データは小学館刊行の新編日本古典文学全集の本文に基づく。『日本語歴史コーパス』では、テキストを言語単位に分割し、品詞等の情報が付与される。採用した言語単位は、言語の形態的側面に着目して規定された「短単位」、構文的側面に着目して規定された「長単位」の 2 種類である。また、『日本語歴史コーパス　平安時代編』には「本文種別」として「会話」「手紙」「歌」「詞書」といった情報が付与されている。これを利用し、『古今和歌集』のうち、和歌のみを調査対象とした。

『古今和歌集』は 905 年醍醐天皇の勅命により撰ばれ、914 年頃成立したとされる最初の勅撰和歌集であり、和歌約 1100 首を全 20 巻に収める。歌体は、長歌 5 首、旋頭歌 4 首を除き、すべてが短歌である。その所収歌は『万葉集』に次ぐ時代から『古今和歌集』撰者の時代まで、約 140 年にわたる。その歌風は、『万葉集』に次ぐ時代から 850 年頃までの「よみ人知らずの時代」、850 年頃から 890 年頃までの「六歌仙の時代」、890 年頃から成立までの「撰者の時代」の 3 期に分類される。「よみ人知らずの時代」の歌が約 4 割を占めるが、素朴でありつつも、撰者時代の美意識により撰ばれた歌と考えることができる。

（２）歌合コーパス

発表者は現在、中古から中世初期にかけて歌合を対象としたコーパス『歌合コーパス』を構築中である。歌合とは、和歌を左右に分けてつがわせ、その優劣を判定した文芸的な遊戯である。平安時代中期の遊楽の中で代表的な行事となり、平安時代末期から鎌倉時代初期にかけて、特に高度な文芸的内容を持つ行事に発展した。この『歌合コーパス』で対象とする歌合は、平安初期の「寛平御時后宮歌合（かんぴょうのおおんときさいのみやのうたあわせ）」「亭子院女郎花合（ていじいんのおみなえしあわせ）」「延喜十三年亭子院歌合（えんぎじゅうさんねんのていじいんのうたあわせ）」、後世の歌合の手本となった平安時代の代表的な歌合である「天徳四年内裏歌合」、歌合の最高峰の一つと言われる鎌倉時代の「六百番歌合」である。歌合は序文・歌・判詞・日記等、多様な要素を持つため、必要に応じて検索が可能になるように、『日本語歴史コーパス平安時代編』の本文種別と同様に、序文・歌・判詞・日記の別や、題・番・左右といった情報を付す。また、和歌だけでなく、序文・判詞・日記を含めた歌合全体に対し、形態論情報を付す[3]。

本発表では、この『歌合コーパス』のうち、平安初期歌合 3 作品「寛平御時后宮歌合」「亭子院女郎花合」「延喜十三年亭子院歌合」を対象とする。本文は、新編日本古典文学全集 11『古今和歌集』巻末の「平安初期歌合」を使用した。収録歌数は、「寛平御時后宮歌合」が

[3] 『歌合コーパス』に付した情報については、冨士池（2014a）を参照方。なお、形態論情報については、『日本語歴史コーパス平安時代編』と共通の仕様としている。

191 首[4]、「亭子院女郎花合」が 51 首、「延喜十三年亭子院歌合」が 80 首であり、歌体はすべて短歌である。成立年代は、「寛平御時后宮歌合」が 889～893 年の間、「亭子院女郎花合」が 898 年、「延喜十三年亭子院歌合」が 913 年であり、『古今和歌集』とほぼ同時代である。これらの歌合の歌の一部は『古今和歌集』に入集している。この時代は「この九世紀の終りから十世紀の初めにまたがる二十年間が、平安朝の貴族和歌の典型が完成した時代である」[5]とされ、いわゆる「たをやめぶり」と言われる優美繊細で理知的な歌風の形成期に当たる。

２．２　言語単位

ここで『日本語歴史コーパス　平安時代編』『歌合コーパス』で採用した言語単位について説明したい。『日本語歴史コーパス　平安時代編』の言語単位は、『現代日本語書き言葉均衡コーパス』で採用した単位を中古和文用に修正・拡張したものであり、『歌合コーパス』の言語単位も共通の仕様としている。採用した言語単位は、言語の形態的側面に着目して規定された「短単位」、構文的側面に着目して規定された「長単位」の 2 種類である。これまでに国立国語研究所が実施してきた語彙調査における言語単位のうち、短い単位の系列に属するものが「短単位」、長い単位の系列に属するものが「長単位」である[6]。

この 2 種類の言語単位のうち、本発表で用いるのは「長単位」である。長単位は文節を自立語と付属語に分割した言語単位である。原則として付属語を 1 長単位とする、助詞・助動詞を伴わない自立語は、主語・主題、連用修飾、連体修飾の各成分の後ろで切る等の規定に基づき、単位認定を行う[7]。長単位では合成語を認めており、結合回数の制限はないため、「木綿つけ鳥」「遊びありく」「時めき給ふ」「小野小町」といった語が 1 まとまりとなる。また、「あさましがる」「たへがたし」「うつくしげ」のような接辞を含めた形式が 1 長単位となる。『日本語歴史コーパス　平安時代編』では複合辞は認めていないが、係り受けを重視し付属語を切り出すのは不適切なものを連語として認めている。連語には「知らず読み」「我は顔」「事の他」等がある。長単位では文脈に即して品詞を付与する方針をとっており、同じ語に対して異なる品詞を与えることがある。例えば、「哀れ」の場合、「もののあはれ知りすぐし、」は名詞を、「皇子もいとあはれなる句を作りたまへるを」は形状詞を付与するといった判別を行っている。図 2-1 に長単位例を示す。調査対象となる「歌」の部分に網掛けを付した。

---

[4] 一般に 190 首とされるが、96 番歌が秋歌と冬歌とに重出しているため、191 首とする。
[5] 新編日本古典文学全集 11『古今和歌集』解説（p.513）
[6] 小椋ほか（2011）第 1 章参照。
[7] 中古和文における長単位認定の概要に関しては冨士池（2012）参照。

キー	語彙素(L)	語彙素読み(L)	品詞(L)	活用型(L)	活用形(L)	本文種別
題	題	ダイ	名詞-普通名詞-一般			詞書
しら	知る	シル	動詞-一般	文語四段-ラ行	未然形-一般	詞書
ず	ず	ズ	助動詞	文語助動詞-ズ	終止形-一般	詞書
読人	読み人	ヨミヒト	名詞-普通名詞-一般			詞書
しら	知る	シル	動詞-一般	文語四段-ラ行	未然形-一般	詞書
ず	ず	ズ	助動詞	文語助動詞-ズ	終止形-一般	詞書
心ざし	志	ココロザシ	名詞-普通名詞-一般			歌
深く	深し	フカシ	形容詞-一般	文語形容詞-ク	連用形-一般	歌
そめ	染む	ソム	動詞-一般	文語下二段-マ行	連用形-一般	歌
て	て	テ	助詞-接続助詞			歌
し	し	シ	助詞-副助詞			歌
をり	折る	オル	動詞-一般	文語四段-ラ行	連用形-一般	歌
けれ	けり	ケリ	助動詞	文語助動詞-ケリ	已然形-一般	歌
ば	ば	バ	助詞-接続助詞			歌
消えあへ	消え敢う	キエアウ	動詞-一般	文語下二段-ハ行	未然形-一般	歌
ぬ	ず	ズ	助動詞	文語助動詞-ズ	連体形-一般	歌
雪	雪	ユキ	名詞-普通名詞-一般			歌
の	の	ノ	助詞-格助詞			歌
花	花	ハナ	名詞-普通名詞-一般			歌
と	と	ト	助詞-格助詞			歌
見ゆ	見ゆ	ミユ	動詞-一般	文語下二段-ヤ行	終止形-一般	歌
らむ	らむ	ラム	助動詞	文語助動詞-ラム	連体形-一般	歌

図 2-1　長単位例

３．名詞率と MVR

本発表では、品詞比率に基づきテキストの特徴を示す指標として、名詞率と MVR を用いる。名詞の比率は文章の特質を表し、名詞の比率に応じて他の品詞もある傾向を持って変化する、つまり文章のジャンルによって品詞の割合が決定されると考えられる。ここでは延べ語数を用いて、品詞比率を求める。樺島忠夫・寿岳章子（1965）は、自立語について品詞をその機能によって体（名詞）・用（動詞）・相（形容詞・形容動詞・副詞・連体詞）・他（接続詞・感動詞）の四つに分類したとき、体の類と、用・相それぞれの類の関係を見るにあたり、MVR という「100×相の類の比率／用の類の比率」の式で表される指標を提案し、名詞率と MVR の組み合わせから見出せる文体的特徴として、名詞率が高く MVR が小さいものを「要約的な文章」、名詞率が低く MVR が大きいものを「ありさま描写的な文章」、名詞率が低く MVR も小さいものを「動き描写的な文章」と位置づけた。『日本語歴史コーパス　平安時代編』の品詞体系では、体の類に「名詞−普通名詞−一般」「名詞-固有名詞−｛一般・人名・地名｝」「名詞-数詞」「代名詞」が、用の類に「動詞−一般」が、相の類に「形容詞−一般」「形状詞−｛一般・タリ｝」「副詞」「連体詞」が分類される。

歌の内容の違いに着目するために、『古今和歌集』は巻ごとに[8]、「寛平御時后宮歌合」「延喜十三年亭子院歌合」は題ごとに、「亭子院女郎花合」は一つとして集計した。ここでは『古今和歌集』の巻、「寛平御時后宮歌合」「延喜十三年亭子院歌合」の題を「部立」と呼ぶこととする。「亭子院女郎花合」はその名のとおり花合であり、その歌の題はすべて女郎花、つまり秋の花である。「亭子院女郎花合」から『古今和歌集』に採録された歌はすべて巻第四（秋歌上）にあり、行事の歌であると同時に、秋の歌と言える。資料・部立ごとの歌数・名詞率（%）・MVR を表 3-1 に、表 3-1 に基づく資料・部立ごとの名詞率（%）・MVR の散

[8] 墨滅歌（俊成本において見せ消ちにされた歌）は調査対象外とした。

布図を図 3-1 に示す。

図 3-1 から、一口に和歌といっても品詞比率には違いがある様子が見てとれる。

まず、内容について検討していきたい。大きく分けて、恋歌は名詞率が低く MVR が大きい「ありさま描写的な文章」、季節歌は恋歌と比較すると名詞率が高く MVR が小さい「要約的な文章」と見ることができる。この傾向は、歌合の歌と『古今和歌集』所収歌との間に差はなく、資料の違いではなく内容の違いと考えられる。季節歌の中を見ると、夏歌は名詞率が低く MVR が大きくなっている一方で、春・秋・冬歌は名詞率が高く MVR が小さくなっている。ここから、品詞比率から見る限り、夏歌は季節歌の典型から離れていると言える。夏歌の大部分はほととぎすを詠んだ歌であり、縁語など相の類が増える要素は特にない。夏歌の MVR が大きくなる原因は不明である。秋歌は同じ『古今和歌集』でも上下で品詞比率に差が見られる。恋歌と季節歌以外は内容による差が大きい。離別歌は名詞率が低く MVR も小さい「動き描写的な文章」と見ることができる一方で、大歌所御歌・神遊びの歌・東歌は最も名詞率が低い「要約的な文章」と見られる。また、「亭子院女郎花合」は行事の歌であると同時に秋の歌であると考えられるが、賀歌とも秋歌とも離れた中間的な位置付けとなった。

表 3-1　資料・部立ごとの歌数・名詞率（%）・MVR

資料	部立	label	歌数	名詞率	MVR
古今和歌集	春歌　上	古_春上	68	57.76	17.79
古今和歌集	春歌　下	古_春下	66	55.19	18.05
古今和歌集	夏歌	古_夏	34	57.75	26.74
古今和歌集	秋歌　上	古_秋上	80	58.82	22.58
古今和歌集	秋歌　下	古_秋下	65	59.89	17.49
古今和歌集	冬歌	古_冬	29	56.05	17.20
古今和歌集	賀歌	古_賀	22	59.35	21.15
古今和歌集	離別歌	古_離別	41	47.73	18.38
古今和歌集	羇旅歌	古_羇旅	16	56.07	20.51
古今和歌集	物名	古_物名	47	53.60	27.94
古今和歌集	恋歌　一	古_恋1	83	54.47	30.58
古今和歌集	恋歌　二	古_恋2	64	55.76	24.61
古今和歌集	恋歌　三	古_恋3	61	49.90	32.12
古今和歌集	恋歌　四	古_恋4	70	52.96	27.36
古今和歌集	恋歌　五	古_恋5	82	52.38	28.40
古今和歌集	哀傷歌	古_哀傷	34	57.52	29.89
古今和歌集	雑歌　上	古_雑上	70	55.56	26.53
古今和歌集	雑歌　下	古_雑下	68	57.72	29.21
古今和歌集	雑躰歌	古_雑躰	68	56.31	28.88
古今和歌集	大歌所御歌・神遊びの歌・東歌	古_大・神・東	32	64.79	20.51
寛平御時后宮歌合	春歌	寛_春	40	55.56	16.42
寛平御時后宮歌合	夏歌	寛_夏	36	56.69	27.10
寛平御時后宮歌合	秋歌	寛_秋	49	58.38	21.01
寛平御時后宮歌合	冬歌	寛_冬	37	55.21	13.18
寛平御時后宮歌合	恋歌	寛_恋	38	49.68	24.60
亭子院女郎花合		女郎花	51	52.53	23.35
延喜十三年亭子院歌合	春歌	亭_春	40	53.13	17.16
延喜十三年亭子院歌合	夏歌	亭_夏	19	56.95	25.00
延喜十三年亭子院歌合	恋歌	亭_恋	21	52.43	33.33

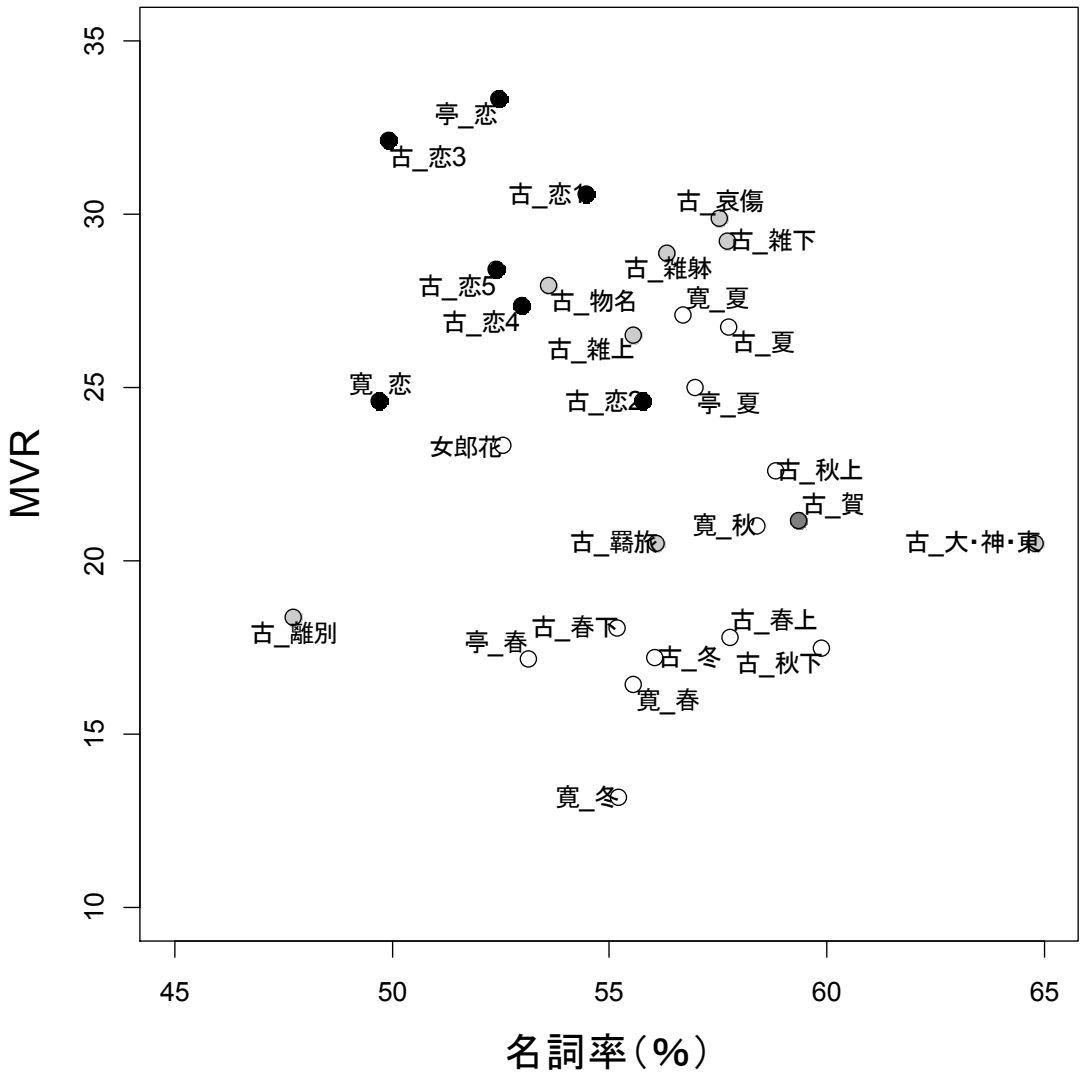


図 3-1　資料・部立ごとの名詞率（%）・MVR の散布図

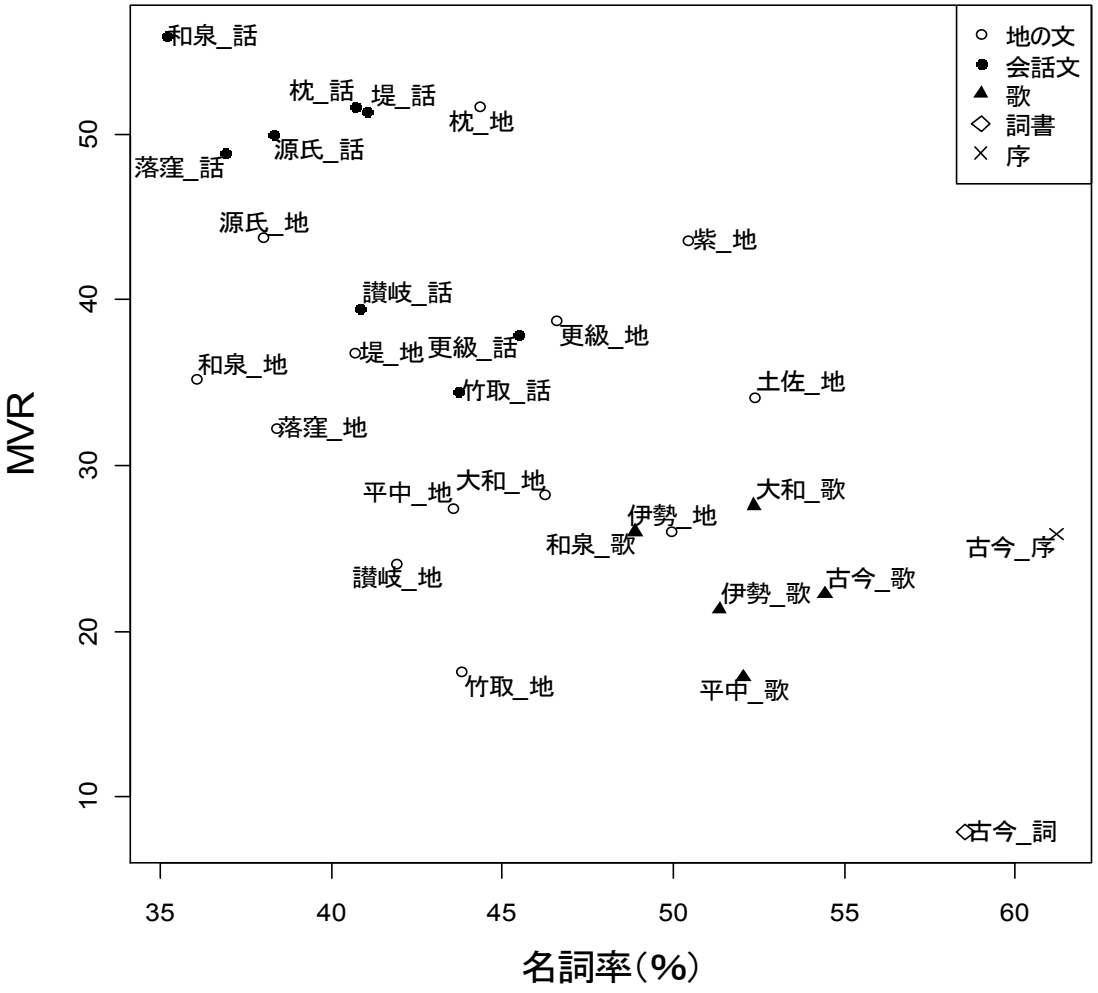


図 3-2　中古 14 作品の名詞率（%）・MVR の散布図

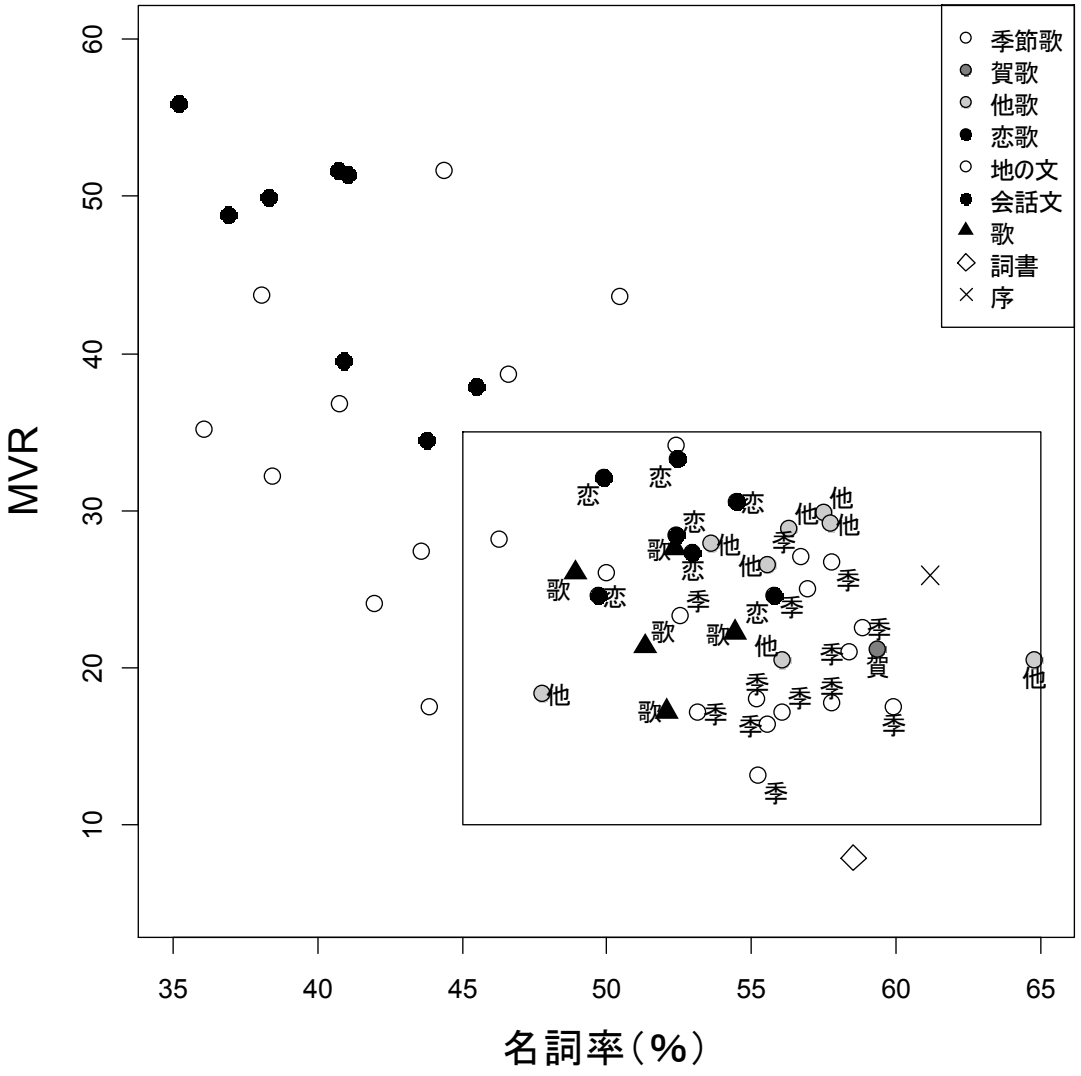


図 3-3　中古 14 作品と古今和歌集・平安初期歌合の名詞率（%）・MVR の散布図

次に、形式について検討する。限られた音数の中での表現が求められる場合には名詞の比率が高いことが指摘されているが、今回の調査ではどのようになっているのだろうか。長歌や旋頭歌といった短歌以外の形式を含む部立は雑躰歌である。図 3-1 を見ると、雑躰歌の名詞率は和歌の中では中程度である。68 首中長歌 5 首、旋頭歌 4 首を含むが、自立語の延べ語数でいうと、長歌が 38%、旋頭歌が 4%を占める。五七七五七七の形式である旋頭歌はともかく、長歌は五七調ないし七五調という音数律を持つものの、長さの制限はない。長歌のみを抽出して集計すると、名詞率 57.78%、MVR 26.95 となる。長さの制限がない長歌の名詞率は今回の調査結果の中ではやや高い方と言える。長歌 5 首と少ないサンプルではあるが、今回の調査結果からは音数の制限と名詞率の高さの関連は見られなかったということになる。

それでは、これらの和歌の品詞比率を、散文の品詞比率と比較するとどのような位置付けになるのだろうか。冨士池（2014b）では『日本語歴史コーパス　平安時代編』に基づく中古和文 14 作品の名詞率と MVR[9]を示した。図 3-2 に中古和文 14 作品の名詞率・MVR の散布図を、図 3-2 と今回の調査結果を重ね合わせたものを図 3-3 に示す。図 3-3 中の四角囲み部分が、図 3-1 の範囲に当たる。冨士池（2014b）では「要約的な文章」として、物語・日記所収の和歌と『古今和歌集』詞書・仮名序を挙げた。図 3-3 からも『古今和歌集』と平安初期歌合の和歌は、会話文や地の文と比較して名詞率が高く MVR が小さい「要約的な文章」であることが見てとれる。和歌の品詞比率はその内容によって異なる様子が見てとれたが、文章のジャンルを超えるものではないと言える。

４．おわりに

本発表では『日本語歴史コーパス　平安時代編』『歌合コーパス』の「長単位」データを用い、テキストの特徴を示す指標として名詞率と MVR を算出した。この指標に基づき、平安初期歌合 3 作品「寛平御時后宮歌合」「亭子院女郎花合」「延喜十三年亭子院歌合」と、同時代の勅撰集『古今和歌集』の和歌について、和歌の内容の違いと品詞比率との関係という観点から検討した。その結果、恋歌、季節歌といった和歌の内容により、品詞比率に差があることが明らかになった。その一方で、形式について検討したところ、長さの制限がない長歌の名詞率はやや高く、限られた音数の中での表現が求められる場合には名詞の比率が高いというこれまでの指摘とは異なる結果となった。また、散文との比較から、今回調査対象とした和歌のテキストの特徴は「要約的な文章」として位置づけられた。ここから和歌の内容による品詞比率の差は文章のジャンルを超えるものではないことが明らかになった。

今回は平安初期和歌に限定して調査を行った。菅原（2003）では八代集において新しいものほど名詞率が高くなる傾向が見られた。歌風の変遷とともに、品詞比率も変化すると推測される。成立年代や歌風の変遷と品詞比率の関連については、今後の課題としたい。

[9] 『古今和歌集』は歌・詞書・仮名序に、他の 13 作品は地の文・会話文・歌に分けて集計し、各作品の延べ語数の 20%以上を占める場合のみを示したもの。

付　記

本発表は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」、JSPS 科研費 「中古中世歌合コーパスに基づく和歌評論の語彙論的研究」（研究課題番号：25770179）の成果の一部である。

文　献

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）「『現代日本語書き言葉均衡コーパス』形態論情報規程集第4版」国立国語研究所内部報告書（LR-CCG-10-05-01）http://www.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-01.pdf

小沢正夫・松田成穂　校注・訳（1994）新編日本古典文学全集 11『古今和歌集』（小学館）

樺島忠夫・寿岳章子（1965）『文体の科学』（綜芸舎）

樺島忠夫（1979）『日本語のスタイルブック』（大修館書店）

菅原優美（2003）「平安時代和歌の語彙の量的構造」『国文目白』42、pp.1-9

冨士池優美（2014a）「中古中世歌合の構造化」『言語処理学会第 20 回年次大会発表論文集』、pp.205-208

冨士池優美（2014b）「品詞比率からみる中古和文テキストの特徴」『日本語学会 2014 年度春季大会予稿集』、pp.185-190

関連 URL

日本語歴史コーパス　http://www.ninjal.ac.jp/corpus_center/chj/

複数のコーパスを用いた新しい文法シラバス策定の試み

庵　功雄（一橋大学国際教育センター）
宮部真由美、永谷直子（一橋大学言語社会研究科大学院生）

Creating New Grammatical Syllabus Based on Several Corpora

Isao Iori, Mayumi Miyabe, Naoko Nagatani (Hitotsubashi University)

要旨

日本語教育を取り巻く状況の変化にともない、新しい文法シラバスが必要とされている。本稿では､そうした問題意識のもと､4技能に対応する4つのコーパスのデータにもとづき、それに日本語学、日本語教育文法の知見を加味して作成した新しい文法シラバスの試案を報告する。本稿の内容は今後の日本語教育に少なからぬインパクトを与えるものである。

１．はじめに

日本語教育を取り巻く現状の変化に合わせて、新しい文法シラバスが求められている（cf. 庵 2009, 2011a, 2014b）。こうした事情に合わせて、国立国語研究所の共同研究プロジェクトの１つとして、「学習者コーパスから見た日本語習得の難易度に基づく語彙・文法シラバスの構築」が行われ、その成果物として論文集の公刊が予定されている。本稿は、こうした問題意識のもとに、現在利用可能な 4 種類の母語話者コーパスを利用し、そこにおける頻度を重視した形で、初級から上級までを見通した文法シラバスの案を提案することを目的とする。ただし、完全に頻度のみで項目を選定するのではなく、そこに、日本語教育文法、および、「やさしい日本語」という観点を取り入れたシラバスの構築を目指すものとする[1]。

２．地域型日本語教育、学校型日本語教育と初級文法シラバス

本稿では、初級と中上級の文法シラバスについてそれぞれ（やや）異なる観点からアプローチすることにする。本節ではまず初級に関する観点について述べる。

初級の文法シラバスについては、先に論者は「やさしい日本語」という観点から Step1, 2 という文法シラバスを提案した（庵 2009, 2011a）[2]。これは、地域型日本語教育[3]の現状にそ

[1] 日本語教育文法に関する論者の考えについては庵(2011b, 2013a)を、「やさしい日本語」については庵(2013b, 2013c, 2014a)をそれぞれ参照されたい。
[2] Step1, 2 はそれぞれ初級前半、初級後半に相当するが、現行の初級シラバスに比べると大幅に項目が刈り込まれている。この点について詳しくは庵(2009, 2011a)を参照されたい。
[3] 日本語教育を「地域型」と「学校型」に分けて考える見方については尾崎(2004)などを参照。

くした「地域型初級」というものである[4]。

一方、日本国内の留学生教育をめぐる状況の変化などに対応するために、学校型日本語教育においてもこの Step1, 2 をベースにした文法シラバスが必要との観点から、論者は学校型日本語教育のための Step1, 2 を提案した（庵 2014b）。

Step1, 2 の主な特徴は以下の通りである（a〜d は地域型、学校型共通。e は学校型のみ）。

(1)a. 理解レベルと産出レベルの区別を明確にし、産出レベルを中心とする。

b. 単文については、南(1974)にもとづく階層構造を想定し、各文法カテゴリーに属する項目を取り上げることにより、森羅万象を日本語で表現できることを保障する。

c. Step1 では活用を（実質的に）廃止し[5]、Step2 でも普通形（plain form）を作るのに必要な形だけを活用形として導入する。

d. 同じ機能を担う形式が複数存在するときには「1 機能 1 形式」を原則とする。

e. 複文については、仁田(1995)にもとづく階層構造を相当し、各文法カテゴリーに属する項目を取り上げる。

これらの理念にもとづく具体的なシラバスについては 6 で提案する。

３．「バイパスとしての「やさしい日本語」」と中上級文法シラバス

上で見た初級シラバス（特に地域型）は、「やさしい日本語」が担う諸機能[6]との関係からも全ての定住外国人に対して保障されるべきものであり、その重要性は極めて大きいが、定住外国人にとって Step1, 2 だけで十分かと言うとそうは言えない。特に、定住外国人の子どもにとってはそうである。

庵(2014a, 近刊)では、定住外国人の子どもが日本語を母語とする子どもに対して負っている言語上のハンディをできるだけ早く埋め、彼（女）らが安定的に教科学習に取り組めるようにするための言語上の調整を「バイパスとしての「やさしい日本語」」と呼んでいる。そして、そうした「バイパスとしての「やさしい日本語」」のための教材は次のような要件を満たす必要があるとしている。

(2)a. 初級から上級までを見通したシラバスによって設計されている。

b. 限られた時間で学べるように、習得すべき項目が厳選されている。

c. 教材において、理解レベルと産出レベルの区別が明確で、各技能に特化した言語知識を導入できる設計になっている。

d. 教室で学ぶことを補完する形で、e-learning などの補助教材が充実している。

---

[4] これに関しては、Step1, 2 を教材化した庵監修(2010, 2011)も参照されたい。

[5] Step1 でも「書きます－書きました／書きません」などの活用はあるが、これらは「ます」を「ました」などに取り替えればいいだけで学習者の実質的な負担はないので、実質的には活用がないのと同じである。

[6] 「やさしい日本語」が担う諸機能について詳しくは庵(2013c, 2014a)を参照されたい。

これらの理念にもとづく具体的なシラバスについては 6 で提案する。

４．使用したコーパスの種類と 4 技能

本稿では、2，3 で見た問題意識にもとづいて新しい文法シラバスを提案する。その際、特に中上級においては、基本的にはコーパスにおける出現頻度を重視する（ただし、表現上の必要性などの観点から項目を入れ替えることもある）。

４.１　項目の選定基準

まず、項目の選定基準について考えるが、これは、現行の文法シラバスが「項目主義」であることを批判した野田(2005)や白川(2005)を受け、「用法主義」をとる。すなわち、重要度が高い項目については、その項目の用法を単位として項目を選定する。

次に、項目の採用基準だが、これは、現代日本語書き言葉均衡コーパス（BCCWJ）の長単位で品詞が「助詞」または「助動詞」であるもの全てとした。そして、それに連体修飾などのコーパスでは検出しにくい項目を加えた。以上の 2 つの方針にもとづいた結果、対象項目数は 246 となった。

４.２　コーパスの種類と 4 技能

ここでは、本稿で用いたコーパスと、それをどのように用いるのかについて述べる。

本稿で使用したコーパスは以下の通りである。

(3)a. 名大会話コーパス（全 129 ファイル）

b. 新聞コーパス（朝日新聞 2012 年版全ファイルのうち 24 日分をランダムサンプリングしたもの）

c. 新書コーパス（Castel/J から配付されているファイルのほぼ全て）

d. BCCWJ（書籍・コアのみ）

本稿ではこれらを「話す、聞く、書く、読む」という 4 技能に対応するものと見なす。

まず、d の BCCWJ の書籍で頻度が高いものは「読む」において必要度が高いと考えられる。一方、c の新書コーパスの内容は、「読む」ことに加え、留学生が「書く」必要があるテキストタイプであると考えられる。さらに、b の新聞コーパスの内容は、「読む」「書く」に加えて「聞く」でも必要であると考えられる。最後に、a の名大会話コーパスの内容は、「話す」を含む 4 技能全てで必要と考えられる。

以上のことから、a が最も基本的で、a→d の順で基本的でなくなる（＝高いレベルの項目になる）と考えることができる。このことを考慮に入れて、5 では各コーパスで頻度が高い項目（各コーパスの特徴語）を抽出し、6 では具体的なシラバスの試案を提示する。

５．コーパスの分析結果

ここでは各コーパスの分析結果について述べる。

５.１　コーパスサイズと総頻度による上位項目

まず、各コーパスのサイズ（形態素数）は次の通りである。

表 1　各コーパスのサイズ

名大	朝日	新書	BCCWJ
1101817	1354428	2014729	169730

次に、総頻度（100 万語単位換算）による上位 30 位までの項目は以下の通りである。

表 2　総頻度による上位項目（100 万語換算）[7]

総頻度順位	形式	品詞	名大	朝日	新書	BCCWJ	総頻度
1	の	格助詞	19225	93654	56601	57892	227372
2	だ／です／で(も)ある	助動詞	59133	29106	41369	39304	168912
3	に	格助詞	17940	56873	38415	37100	150328
4	は	とりたて助詞（係助詞）	17187	51746	37711	43104	149748
5	た	助動詞	27342	57229	23307	39433	147311
6	を	格助詞	6603	58349	29260	38496	132708
7	が	格助詞	17796	43392	25578	30372	117138
8	て	接続助詞	27273	37513	35291	16980	117057
9	と	格助詞	19242	35140	29877	21511	105770
10	で	格助詞	19646	32511	12424	13993	78574
11	も	とりたて助詞（係助詞）	18207	15954	13981	18082	66224
12	ている	助動詞	8050	10910	9900	16002	44862
13	れる／られる	助動詞	2886	12614	10832	9592	35924
14	ない	助動詞	11900	7456	7445	7553	34354
15	ね	終助詞	27062	515	180	1078	28835
16	から	格助詞	2991	9583	4084	4819	21477
17	か	とりたて助詞（副助詞）	10787	2272	3561	3971	20591
18	か	終助詞	11386	2106	2251	3682	19425
19	ます	助動詞	2945	3607	3227	9580	19359
20	ようだ	助動詞	1161	8469	4398	4725	18753
21	って	とりたて助詞（副助詞）	17280	289	136	736	18441
22	のだ	助動詞	4809	537	3530	8826	17702
23	から	接続助詞	9945	866	2475	2451	15737
24	よ	終助詞	13431	580	206	1214	15431
25	や	とりたて助詞（副助詞）	451	7440	3249	2899	14039
26	が	接続助詞	206	4619	4444	4596	13865
27	と	接続助詞	2533	3684	2798	3005	12020
28	など	とりたて助詞（副助詞）	92	6824	2375	2280	11571
29	けれど	接続助詞	8830	333	232	672	10067
30	ば	接続助詞	1475	2000	3210	2822	9507

[7] BCCWJ の検索は原則として長単位で行った。

５.２　各コーパスの特徴項目の抽出方法

次に、各コーパスの特徴項目を抽出するが、その際、次の指標を設定した。

(4) そのコーパスにおける頻度の割合が 40%を越える。

これは各コーパス間の出現頻度の割合の差からそのコーパスの特徴項目を取り出すというものである[8]。

５.３　各コーパスの特徴項目

ここでは、各コーパスの特徴項目（(4)を満たす）を取り上げる。

５.３.１　名大会話コーパスの特徴項目

表 3　名大会話コーパスの特徴項目[9]

名大比率順位	項目	品詞	名大比率	総頻度順位	名大順位	総頻度順位-名大順位
1	ようかな	終助詞	**97.2**	153	82	71
2	さ	終助詞	**96.9**	34	20	14
3	のだったら	接続助詞	**94.9**	166	91	75
4	ね	終助詞	**93.9**	15	4	11
5	って	とりたて助詞(副助詞)	**93.7**	21	11	10
6	の	終助詞	**91.5**	41	23	18
7	かな／かね	終助詞	**89.4**	66	37	29
8	けれど	接続助詞	**87.7**	29	18	11
9	もの	終助詞	**87.5**	129	56	73
10	よ	終助詞	**87.0**	24	13	11
11	なんて／なんか	とりたて助詞(副助詞)	**85.2**	55	27	28
12	な	終助詞	**82.9**	36	22	14
13	みたいだ	助動詞	**82.0**	68	40	28
14	ようか	終助詞	**72.6**	114	68	46
15	わ	終助詞	**72.2**	78	49	29
16	たら	接続助詞	**71.4**	33	24	9
17	くらい	とりたて助詞(副助詞)	**66.9**	57	35	22
18	ましょうか	終助詞	**66.0**	199	127	72
19	から	接続助詞	**63.2**	23	17	6
20	ようね／ようよ	意向形	**61.6**	169	106	63
21	たって	接続助詞	**61.3**	165	100	65
22	てある	助動詞	**60.7**	94	57	37
23	か	終助詞	**58.6**	18	15	3
24	ちゃ*	接続助詞	**58.3**	121	76	45
25	ようと{思う／考える}	意向形	**58.2**	126	80	46
26	し	接続助詞	**57.5**	56	36	20
27	からか	終助詞	**57.1**	211	159	52
28	かしら	終助詞	**54.4**	140	88	52
29	か	とりたて助詞(副助詞)	**52.4**	17	16	1
30	らしい	助動詞	**51.5**	93	60	33
31	{たら／ば／と}いい	助動詞	**49.2**	86	55	31
32	のか	終助詞	**48.9**	48	33	15
33	かもしれない	助動詞	**47.6**	82	54	28
34	からではない	助動詞	**47.2**	196	132	64
35	そうだ*	助動詞	**46.0**	89	61	28
36	てあげる	助動詞	**45.4**	151	95	56
37	てしまう／ちゃう	助動詞	**44.3**	46	34	12
38	のに*	接続助詞	**44.3**	71	52	19
39	ではない	助動詞	**44.1**	32	26	6
40	かい	終助詞	**42.9**	208	148	60
41	ておく／とく	助動詞	**41.7**	83	58	25
42	てもいい	助動詞	**40.9**	104	77	27
43	ぜ	終助詞	**40.4**	178	123	55

[8] (4)以外に、「総頻度順位と当該コーパスの順位が 40 以上であり、かつ、当該のコーパスでの比率が各コーパス間で最も高い」という指標も設定したが、これのみを満たす項目はなかった。

[9] *の項目は BCCWJ で短単位検索を行った。また、「かい」は BCCWJ と出現頻度が同じであった。

５.３.２　新聞コーパスの特徴項目

表４　新聞コーパスの特徴項目

朝日比率順位	項目	品詞	朝日比率	総頻度順位	朝日順位	総頻度順位－朝日順位
1	と見られ(てい)る	助動詞	**86.8**	139	69	70
2	をめぐる	格助詞	**84.6**	123	63	60
3	によると／によれば	格助詞	**83.0**	67	31	36
4	ために(理由)	接続助詞	**70.8**	146	55	91
5	につき	格助詞	**66.7**	209	170	39
6	など	とりたて助詞(副助詞)	**59.0**	28	18	10
7	ようと	意向形	**58.8**	113	68	45
8	際に	格助詞	**56.0**	163	123	40
9	たところ	接続助詞	**55.8**	131	81	50
10	と考えられ(てい)る	助動詞	**55.0**	135	89	46
11	てほしい	助動詞	**54.1**	100	62	38
12	や	とりたて助詞(副助詞)	**53.0**	25	17	8
13	について	格助詞	**52.3**	58	33	25
14	とともに	接続助詞	**51.3**	120	78	42
15	へ	格助詞	**50.6**	43	22	21
16	をはじめ	格助詞	**50.0**	158	120	38
17	を通じて	格助詞	**49.7**	145	103	42
18	ものの	接続助詞	**48.2**	150	110	40
19	ようだ	助動詞	**45.2**	20	15	5
20	から	格助詞	**44.6**	16	14	2
21	つつ	接続助詞	**44.4**	125	88	37
22	を	格助詞	**44.0**	6	2	4
23	上で	接続助詞	**43.7**	136	98	38
24	にわたる／にわたり／にわたって	格助詞	**43.4**	142	102	40
25	による／により／によって	格助詞	**43.1**	42	24	18
26	ずつ	とりたて助詞(副助詞)	**42.6**	130	96	34
27	たい	助動詞	**41.9**	47	25	22
28	で	格助詞	**41.4**	10	9	1
29	の	格助詞	**41.2**	1	1	0

５.３.３　新書コーパスの特徴項目

表５　新書コーパスの特徴項目

新書比率順位	項目	品詞	新書比率	総頻度順位	新書順位	総頻度順位-新書順位
1	につれて／につれ	接続助詞	**77.0**	186	140	46
2	において／における	格助詞	**67.4**	79	40	39
3	にしたがって／にしたがい	接続助詞	**66.7**	187	144	43
4	とも	接続助詞	**64.3**	147	96	51
5	に関わらず	接続助詞	**62.5**	207	184	23
6	に過ぎない	助動詞	**61.6**	149	102	47
7	までもない	助動詞	**59.6**	159	121	38
8	ているところだ	助動詞	**58.8**	198	172	26
9	に際して／に際し	格助詞	**58.5**	201	179	22
10	と思われ(てい)る	助動詞	**55.4**	137	93	44
11	にせよ	接続助詞	**55.2**	180	145	35
12	をもって	格助詞	**54.4**	174	138	36
13	にも関わらず	接続助詞	**53.0**	160	127	33
14	ようとしても／ようとも	意向形	**52.8**	191	162	29
15	まい	助動詞	**52.5**	152	118	34
16	にしても	接続助詞	**50.9**	128	87	41
17	としたら／とすれば／とすると	接続助詞	**50.2**	116	79	37
18	上に	接続助詞	**50.0**	205	188	17
19	に至るまで	格助詞	**46.3**	183	158	25
20	ことに／となる	助動詞	**45.7**	69	44	25
21	からといって	接続助詞	**45.6**	175	148	27
22	に違いない	助動詞	**45.4**	148	119	29
23	と同時に	接続助詞	**45.1**	164	137	27
24	からこそ	接続助詞	**44.9**	172	147	25
25	からだ*	助動詞	**44.8**	81	51	30
26	なければならない	助動詞	**44.7**	90	64	26
27	に限らず	とりたて助詞(副助詞)	**44.1**	192	170	22
28	たところで	接続助詞	**43.6**	194	180	14
29	だけでなく	とりたて助詞(副助詞)	**43.5**	115	84	31
30	とともに	接続助詞	**42.4**	120	91	29
31	かと思うと	接続助詞	**42.1**	210	200	10
32	ためだ*	助動詞	**41.4**	138	111	27
33	さえ	とりたて助詞(副助詞)	**40.6**	106	78	28
34	ほど(～ば～ほど)	とりたて助詞(副助詞)	**40.4**	177	156	21
35	のみ	とりたて助詞(副助詞)	**40.4**	132	104	28
36	に対する／に対して／に対し	格助詞	**40.3**	62	41	21

５.３.４　BCCWJ（書籍・コア）の特徴項目

表 6　BCCWJ（書籍・コア）の特徴項目

BCCWJ比率順位	項目	品詞	BCCWJ比率	総頻度順位	BCCWJ順位	総頻度順位－BCCWJ順位
1	AようがAまいが／AだろうがBだろうが	意向形	**88.2**	184	144	40
2	のではない	助動詞	**78.3**	95	62	33
3	おかげで*	接続助詞	**74.4**	202	183	19
4	といっても	とりたて助詞（係助詞）	**62.6**	171	145	26
5	はずがない	助動詞	**61.0**	188	167	21
6	という／といった	格助詞	**57.7**	31	17	14
7	どころ	とりたて助詞（副助詞）	**57.1**	157	124	33
8	ようとすると／ようとしたら	意向形	**57.1**	200	187	13
9	のみならず	とりたて助詞（副助詞）	**55.8**	195	182	13
10	のに（目的）*	接続助詞	**55.0**	168	151	17
11	ように（間接引用）	格助詞	**54.2**	111	91	20
12	にしろ	接続助詞	**54.0**	185	165	20
13	と考える／と考えられる	助動詞	**53.5**	107	85	22
14	やら	とりたて助詞（副助詞）	**52.7**	156	127	29
15	すら	とりたて助詞（副助詞）	**51.9**	154	123	31
16	のだ	助動詞	**49.9**	22	15	7
17	に至るまで	格助詞	**49.5**	183	164	19
18	ます	助動詞	**49.5**	19	14	5
19	のだから*	接続助詞	**49.0**	117	97	20
20	わけがない*	助動詞	**48.6**	204	194	10
21	せいで	接続助詞	**47.9**	190	179	11
22	ては*	接続助詞	**47.8**	119	99	20
23	てくださる	助動詞	**47.7**	99	84	15
24	に違いない	助動詞	**46.7**	148	121	27
25	だろうか	終助詞	**46.6**	76	55	21
26	のに（逆接）*	接続助詞	**46.6**	97	82	15
27	ようとしない	意向形	**46.1**	189	178	11
28	わけだ*	助動詞	**46.1**	70	51	19
29	ところだ	助動詞	**45.9**	133	110	23
30	ことができる	助動詞	**45.4**	77	61	16
31	はずだ	助動詞	**45.4**	103	90	13
32	のだろうか	終助詞	**45.0**	108	95	13
33	からには	接続助詞	**43.9**	203	195	8
34	てやる	助動詞	**43.2**	134	112	22
35	わりに	接続助詞	**42.9**	193	189	4
36	かい	終助詞	**42.9**	208	199	9
37	なり	とりたて助詞（副助詞）	**42.7**	182	174	8
39	たところだ／たばかりだ	助動詞	**42.4**	173	160	13
40	ようにする	助動詞	**41.6**	124	105	19
41	さえ	とりたて助詞（副助詞）	**41.4**	106	92	14
42	なら（主題）	接続助詞	**41.2**	110	98	12

６．シラバスの試案

ここでは新しい文法シラバスの試案を提案する。Step1, 2 は初級前半、初級後半に対応し、Step3 は文体を整理するレベルとする。Step4～6 はそれぞれ中級、中上級、上級に対応する。

なお、紙幅の関係上、各項目について、項目名、品詞、総頻度順位のみを記す。

このシラバスは、基本的に次の方針で策定されている。

(5)a. Step1,2 は総頻度順位の上位のものを取り入れるが、日本語文の構造上必要な項目は適宜加える。

b. Step3 は名大コーパス、Step4 は新聞コーパス、Step5 は新書コーパスにおいて頻度順で上位の項目を優先的に含める。

c. 当該の Step において表現機能上必要な項目については、頻度順とは関係なくその Step に含める。

表 7　新しい文法シラバス

Step	項目	品詞	総頻度順位
1	の	格助詞	1
1	です	助動詞	2
1	に	格助詞	3
1	は（主題）	とりたて助詞（係助詞）	4
1	た	助動詞	5
1	を	格助詞	6
1	が	格助詞（目的語）	7
1	と	格助詞	9
1	で	格助詞	10
1	も	とりたて助詞（係助詞）	11
1	ない	助動詞	14
1	から	格助詞	16
1	か	とりたて助詞（副助詞）	17
1	か	終助詞	18
1	ます	助動詞	19
1	ではない	助動詞	32
1	たい	助動詞	47
1	と思う	助動詞	52
1	より	格助詞	59
2	が	格助詞（主語）	7
2	て	接続助詞	8
2	と（引用）	格助詞	9
2	ている（進行中）	助動詞	12
2	（ら）れる（有情物主語）	助動詞	13
2	ね	終助詞	15
2	のだ（理由）	助動詞	22
2	から	接続助詞	23
3	だ／で（も）ある	助動詞	2
3	は（対比）	とりたて助詞（係助詞）	4
3	ている（結果残存、繰り返し）	助動詞	12
3	（ら）れる（使役受身）	助動詞	13
3	ようだ	助動詞	20
3	って	とりたて助詞（副助詞）	21
3	のだ（解釈）	助動詞	22
3	が	接続助詞	26
3	と（事実的）	接続助詞	27
3	ば	接続助詞	30
3	という／といった	格助詞	31
3	たら（事実的）	接続助詞	33
3	さ	終助詞	34
3	な	終助詞	36
3	だろう／であろう（推量）*	助動詞	38
3	てくる／てく（方向性）	助動詞	40
3	の	終助詞	41
3	による／により／によって	格助詞	42
3	へ	格助詞	43
3	として	格助詞	44
3	だけ	とりたて助詞（副助詞）	45
3	てしまう／ちゃう	助動詞	46
3	（さ）せる（「（さ）せてください」）	助動詞	49
3	か（おかげか／せいか／ためか／わけか／からか／のか）	終助詞	50
3	ていく／てく（方向性）	助動詞	54
3	てくれる	助動詞	63
3	みたいだ	助動詞	68
3	なら／のなら（条件）	接続助詞	84
3	てある（結果残存）	助動詞	94
3	たって	接続助詞	165
3	のだったら	接続助詞	166
3	可能形		
3	連体修飾（内の関係）（産出）		
4	ている（完了）	助動詞	12
4	（ら）れる（無情物主語）	助動詞	13
4	のだ（言い換え）	助動詞	22
4	と（条件）	接続助詞	27
4	など	とりたて助詞（副助詞）	28
4	てくる／てく（アスペクト）	助動詞	40
4	のか（説明を求める疑問文）	終助詞	48
4	（さ）せる（「（さ）せてもらう／（さ）せてくれる」）	助動詞	49
4	ず	助動詞	51
4	ていく／てく（アスペクト）	助動詞	54
4	なんて／なんか	とりたて助詞（副助詞）	55
4	し	接続助詞	56

Step	項目	品詞	総頻度順位
2	よ	終助詞	24
2	や	とりたて助詞（副助詞）	25
2	けれど	接続助詞	29
2	たら（条件）	接続助詞	33
2	よう／ましょう	意向形	37
2	でしょう（確認要求）*	助動詞	38
2	ても*	接続助詞	39
2	のか（前提のある疑問文）	終助詞	48
2	たり	とりたて助詞（副助詞）	53
2	ので*	接続助詞	60
2	ながら	接続助詞	61
2	ことが／も／はある／ない	助動詞	64
2	ために（目的）	接続助詞	72
2	てもらう	助動詞	73
2	しか	とりたて助詞（副助詞）	75
2	ことができる	助動詞	77
2	かもしれない	助動詞	82
2	ようになる	助動詞	85
2	なくてはならない	助動詞	90
2	のに（逆接）*	接続助詞	97
2	ための	格助詞	102
2	てもいい	助動詞	104
2	なら（主題）	接続助詞	110
2	ように（目的）	接続助詞	127
2	分裂文（変形）（理解）		
2	連体修飾（内の関係）（理解）		
4	くらい	とりたて助詞（副助詞）	57
4	について	格助詞	58
4	に対する／に対して／に対し	格助詞	62
4	てみる	助動詞	65
4	かな／かね	終助詞	66
4	によると／によれば	格助詞	67
4	ことに／となる	助動詞	69
4	わけだ*	助動詞	70
4	べきだ	助動詞	74
4	だろうか	終助詞	76
4	わ	終助詞	78
4	において／における	格助詞	79
4	からだ*	助動詞	81
4	ておく／とく	助動詞	83
4	[たら／ば／と]いい	助動詞	86
4	ばかり	とりたて助詞（副助詞）	87
4	にとって	格助詞	88
4	そうだ*	助動詞	89
4	ではないか（否定疑問）	終助詞	91
4	だけだ	助動詞	92
4	らしい	助動詞	93
4	てある（準備）	助動詞	94
4	のではない	助動詞	95
4	しそうだ（動詞のみ）	助動詞	98
4	てくださる	助動詞	99
4	てほしい	助動詞	100
4	わけで[は／も]ない*	助動詞	101
4	はずだ	助動詞	103
4	さえ	とりたて助詞（副助詞）	106
4	と考える／と考えられる	助動詞	107
4	のだろうか	終助詞	108
4	のではないか	終助詞	109
4	ように（間接引用）	格助詞	111
4	てはならない／てはいけない	助動詞	112
4	としたら／とすれば／とすると	接続助詞	116
4	のだから*	接続助詞	117
4	をめぐる	格助詞	123
4	もの	終助詞	129
4	ために（理由）	接続助詞	146
4	からといって	接続助詞	175
4	分裂文（情報の受け継ぎ）（理解）		
4	連体修飾（外の関係・内容補充）（産出）		
4	お～になる（尊敬語）		
4	ハーガ文（形容詞、名詞）（産出）		

Step	項目	品詞	総頻度順位
5	ている(経験・記録)	助動詞	12
5	のだ(発見)	助動詞	22
5	(さ)せる(「(さ)せる」)	助動詞	49
5	ようとする	意向形	80
5	ではないか(認識要求)	終助詞	91
5	こそ	とりたて助詞(係助詞)	96
5	に関する／に関して	格助詞	105
5	ようか	終助詞	114
5	だけでなく	とりたて助詞(副助詞)	115
5	しかない	助動詞	118
5	とともに	接続助詞	120
5	ちゃ*	接続助詞	121
5	ようにする	助動詞	124
5	つつ	接続助詞	125
5	ようと{思う／考える}	意向形	126
5	にしても	接続助詞	128
5	ずつ	とりたて助詞(副助詞)	130
5	たところ	接続助詞	131
5	ところだ	助動詞	133
5	てやる	助動詞	134
5	と考えられ(てい)る	助動詞	135
5	上で	接続助詞	136
5	と思われ(てい)る	助動詞	137
5	かしら	終助詞	140
5	にわたる／にわたり／にわたって	格助詞	142
5	ぞ	終助詞	143
5	ていただく	助動詞	144
5	に違いない	助動詞	148
5	てあげる	助動詞	151
5	ようかな	終助詞	153
5	ようね／ようよ	意向形	169
5	ぜ	終助詞	178
5	たがる	助動詞	179
5	はずがない	助動詞	188
5	せいで	接続助詞	190
5	からではない	助動詞	196
5	ましょうか	終助詞	199
5	おかげで*	接続助詞	202
5	かい	終助詞	208
5	からか	終助詞	211
5	尊敬語(不規則)		
5	お～する(謙譲語)		
5	連体修飾(外の関係・非制限節、相対補充)(産出)		
6	ている(反事実)	助動詞	12
6	まで	格助詞／とりたて助詞(副助詞)	35
6	ようと	意向形	113
6	ては*	接続助詞	119
6	としても	接続助詞	122
6	のみ	とりたて助詞(副助詞)	132
6	ためだ(理由)*	助動詞	138
6	と見られ(てい)る	助動詞	139
6	ことにする	助動詞	141

Step	項目	品詞	総頻度順位
6	を通じて	格助詞	145
6	とも	接続助詞	147
6	に過ぎない	助動詞	149
6	ものの	接続助詞	150
6	まい	助動詞	152
6	すら	とりたて助詞(副助詞)	154
6	つつある	助動詞	155
6	やら	とりたて助詞(副助詞)	156
6	どころ	とりたて助詞(副助詞)	157
6	をはじめ	格助詞	158
6	までもない	助動詞	159
6	にも関わらず	接続助詞	160
6	とはいえ	接続助詞	161
6	と見る	助動詞	162
6	際に	格助詞	163
6	と同時に	接続助詞	164
6	ざるを得ない	助動詞	167
6	のに(目的)*	接続助詞	168
6	きり	とりたて助詞(副助詞)	170
6	といっても	とりたて助詞(係助詞)	171
6	からこそ	接続助詞	172
6	たところだ／たばかりだ	助動詞	173
6	をもって	格助詞	174
6	わけにはいかない	助動詞	176
6	ほど(～ば～ほど)	とりたて助詞(副助詞)	177
6	にせよ	接続助詞	180
6	にあたって／にあたり	格助詞	181
6	なり	とりたて助詞(副助詞)	182
6	に至るまで	格助詞	183
6	AようがAまいが／AだろうがBだろうが	意向形	184
6	にしろ	接続助詞	185
6	につれて／につれ	接続助詞	186
6	にしたがって／にしたがい	接続助詞	187
6	ようとしない	意向形	189
6	ようとしても／ようとも	意向形	191
6	に限らず	とりたて助詞(副助詞)	192
6	わりに	接続助詞	193
6	たところで	接続助詞	194
6	のみならず	とりたて助詞(副助詞)	195
6	にて	格助詞	197
6	ているところだ	助動詞	198
6	ようとすると／ようとしたら	意向形	200
6	に際して／に際し	格助詞	201
6	からには	接続助詞	203
6	わけがない*	助動詞	204
6	上に	接続助詞	205
6	にしては	接続助詞	206
6	に関わらず	接続助詞	207
6	につき	格助詞	209
6	かと思うと	接続助詞	210
6	語順倒置(ハーガ)		
6	謙譲語(不規則)		

7．まとめ

本稿では、日本語教育を取り巻くさまざまな状況に対応するために必要とされる新しい文法シラバスの策定法について考えた。このシラバスは、タイプの異なる 4 つのコーパスの頻度にもとづいて、そこに日本語教育文法の知見を加味して策定されたものである。今後は、このシラバスにもとづいた教材の開発に傾注していきたい。

謝　辞

本稿は、日本学術振興会による科学研究費助成金基盤研究（A）「やさしい日本語を用いた言語的少数者に対する言語保障の枠組み策定のための総合的研究」(平成 25～28 年度　研究代表者：庵功雄)、および、国立国語研究所の共同研究プロジェクト「学習者コーパスか

ら見た日本語習得の難易度に基づく語彙・文法シラバスの構築」（プロジェクトリーダー・山内博之）の助成を受けたものである。

参考文献

庵　功雄(2009)「地域日本語教育と日本語教育文法:「やさしい日本語」という観点から」『人文・自然研究』3、pp.126-141、一橋大学

庵　功雄(2011a)「日本語教育文法からみた「やさしい日本語」の構想：初級シラバスの再検討」『語学教育研究論叢』28、pp.255-271、大東文化大学

庵　功雄(2011b)「日本語記述文法と日本語教育文法」森篤嗣・庵功雄編『日本語教育文法のための多様なアプローチ』ひつじ書房

庵　功雄(2013a)「日本語教育文法の現状と課題」『一橋日本語教育研究』創刊号、pp.1-12、ココ出版

庵　功雄(2013b)「「やさしい日本語」とは何か」庵功雄・イ・ヨンスク・森篤嗣編『「やさしい日本語は何を目指すか』、pp.3-13、ココ出版

庵　功雄(2013c)『日本語教育、日本語学の「次の一手」』くろしお出版

庵　功雄(2014a)「「やさしい日本語」研究の現状と今後の課題」『一橋日本語教育研究』2、pp.1-12、ココ出版

庵　功雄(2014b)「文法シラバスの作成を科学する」公開シンポジウム「シラバス作成を科学にする―日本語教育に役立つ多面的な文法シラバスの作成―」予稿集

庵　功雄(近刊)「言語的マイノリティに対する言語上の保障と「やさしい日本語」」『ことばと文字』2、くろしお出版

庵　功雄監修(2010, 2011)『にほんごこれだけ！１、２』ココ出版

尾崎明人(2004)「地域型日本語教育の方法論的試論」小山悟他編『言語と教育』、pp.295-310、くろしお出版

白川博之(2005)「日本語学的文法から独立した日本語教育文法」野田尚史編『コミュニケーションのための日本語教育文法』、pp.43-62、くろしお出版

仁田義雄(1995)「日本語文法概説（複文・連文編）」宮島達夫・仁田義雄編『日本語類義表現の文法（下）』、pp.383-396、くろしお出版

野田尚史(2005)「コミュニケーションのための日本語教育文法の設計図」野田尚史編『コミュニケーションのための日本語教育文法』、pp.1-20、くろしお出版

南不二男(1974)『現代日本語の構造』大修館書店

口頭発表（2）

9月9日（火）　15:30～17:00

NINJAL-LWP の類義語比較機能

赤瀬川 史朗（Lago 言語研究所）
プラシャント・パルデシ（国立国語研究所）
今井 新悟（筑波大学）

The Function to Compare Near-Synonyms in NINJAL-LWP

Shiro Akasegawa (Lago Institute of Language)
Prashant Pardeshi (National Institute for Japanese Language and Linguistics)
Shingo Imai (Tsukuba University)

要旨

NINJAL-LWP は、レキシカルプロファイリング型のコーパス検索ツールである。現在、BCCWJ 向けの NLB（NINJAL-LWP for BCCWJ）と TWC（筑波ウェブコーパス）向けの NLT（NINJAL-LWP for TWC）の 2 つのバージョンが公開されている。レキシカルプロファイリング型のコーパスツールの特長は、特定の語彙の振る舞いのおおよその全体像を把握できる点にある。なかでも、コロケーションを網羅的にしかも簡単に調査できることは、コンコーダンサにはない優れた特長と言える。今回、この NINJAL-LWP に、内容語を対象とした類義語の調査を可能にする 2 語比較機能を実装した。これまでは難しかったコロケーションレベルの類義語の研究に活用されるものと期待される。類義語のペアは、同じ品詞の内容語のほか、形容詞と連体詞（「小さい」と「小さな」）にも対応している。形容詞では、イ形容詞とナ形容詞の比較（「美しい」と「きれいな」）も可能である。また、片方の語に見られる特徴的なコロケーションだけでなく、両方の語に現れるコロケーション（「楽しい気分」と「うれしい気分」）の差異を用例レベルで簡単に比較できる機能も搭載している。本稿では、コロケーション分析におけるレキシカルプロファイリング型のコーパスツールの利点と、NINJAL-LWP の 2 語比較機能の概要を述べた上で、いくつかの比較例を提示する。

1. レキシカルプロファイリングとコロケーション分析

NINJAL-LWP は、レキシカルプロファイリング型のコーパス検索ツール（以下、レキシカルプロファイラー）である。現在、BCCWJ 向けの NLB（NINJAL-LWP for BCCWJ）と TWC（筑波ウェブコーパス）向けの NLT（NINJAL-LWP for TWC）の 2 つのバージョンが一般公開されている（関連 URL 参照）。

レキシカルプロファイリングは、あらかじめ設定された検索式に基づいて、コーパスから様々なタイプのコロケーションの情報を抽出した結果を、文法パターンごとに整理してユーザに提示するコーパス検索手法である。特定の語彙の文法的振る舞いやコロケーションをマクロ的視点から調査できる点に大きな特長がある。一方、最も一般的なコーパスツールであるコンコーダンスでは、特定の言語事象を対象にしたミクロな視点からの観察が可能である。その意味で、レキシカルプロファイリングとコンコーダンスは、コーパス検索において相互補完的な役割を果たしていると言える。

レキシカルプロファイリングがコロケーション分析に適している理由は 3 つある。一つ目は、上述の通り、検索式に基づいてコロケーションを網羅的に抽出するため、コンコーダンスに比べて、一般のユーザでも簡単にコロケーション調査ができることがある。コンコーダンサを使って、コロケーションを抽出する場合、1) コンコーダンスの出力をダウンロードする、2) 共起語を集計する、3) 集計した結果を頻度などの統計値を用いて並べ替える、という 3 つの作業が必要になる。これに対して、レキシカルプロファイラーでは、このような作業は事前に行われるため、ユーザは調査したいパターンを選ぶだけで結果が得られる。図 1 は、NLB で「冷える」のガ格名詞のコロケーションを調べた画面である。左側のパネルの「…が冷える」をクリックするだけで、右側にそのコロケーションが表示される。

図 1　NLB によるコロケーション分析

2 つ目の理由は、コンコーダンサでコロケーションを抽出する際に必要になる語数指定がレキシカルプロファイラーでは不要なことである。例えば、自動詞とガ格名詞のコロケーションを調べるときは、ガ格名詞と自動詞との間に付加詞や副詞などの修飾要素が置かれることを考慮に入れる必要がある。通常、そのような語は 5 語以内に収まることが多いと思われるが、実際の用例ではその範囲を超える可能性もある。不足なく抽出するには、抽出する範囲を広げるとよさそうに思えるが、あまり広げ過ぎると今度は、主述の関係にないガ格名詞と動詞を拾ってしまう場合が出てくる。

NINJAL-LWP では、係り受け解析の結果を利用するため、上記のような中心語と共起語の語数の問題に悩まされることはない。例えば、動詞「冷える」と共起するガ格名詞では、以下のように中心語と共起語がかなり離れた用例でも抽出される（下線は筆者）。

(1) マグマがマグマ溜りの中でゆっくり冷えるにつれて次々といろいろな鉱物ができ，マグマ溜りの下に沈んでいく。（BCCWJ 教科書サブコーパス，高等学校地学 I）

3 つ目の理由は、レキシカルプロファイラーでは、頻度以外の統計値があらかじめ用意されているため、それらの統計値を用いて並べ替えができる点である。NINJAL-LWP では、単純頻度以外に、MI と logDice という統計値が用意されている。また、NLB では、サブコ

ーパス間の使用分布を比較するために、PMW（100 万語あたりの調整頻度）が表示される。コンコーダンサの場合、頻度以外の統計値が用意されていないときは、ユーザ側で計算する必要が出てくるが、その際、共起頻度のほかに、中心語の頻度、共起語の頻度、コーパスの総語数などが分かっていなければならない。

このように、コンコーダンサを利用してコロケーションを分析するには、かなりの労力が要求される。また、3 番目に挙げた統計値の算出では、種々の頻度データを用意する必要があるが、これらは一般のユーザが簡単に準備できるものではない。レキシカルプロファイラーでは、このような一連の作業を事前に行った上で、コロケーションのリストを提示するため、ユーザは一番肝心なコロケーションの分析作業にすぐ取りかかることができる。

2. NINJAL-LWP の類義語の比較機能

まず、従来の単一の見出し語のプロファイリング画面を 2 つ開いて、類義語の比較を行うとどのようになるかをシミュレートしてみたい。図 2 は、NLB で「冷える」と「冷める」の見出し語ウィンドウをそれぞれ開いて、ガ格名詞のコロケーションのリストを比較したものである。

…が冷える　110種類

コロケーション	コーパス全体		
	頻度	MI	LD
体が冷える	44	8.63	4.98
身体が冷える	15	9.18	5.50
足が冷える	12	8.40	4.73
【人名】が冷える	8	1.43	-2.22
お腹が冷える	8	9.81	6.07
からだが冷える	7	9.77	6.02
手足が冷える	6	11.21	7.24
下半身が冷える	6	12.09	7.90
空気が冷える	5	8.58	4.87
頭が冷える	4	6.00	2.34
足元が冷える	4	9.60	5.79
腰が冷える	4	7.95	4.25
指先が冷える	4	10.50	6.55
全身が冷える	4	9.09	5.32
先が冷える	4	5.60	1.95
マグマが冷える	4	12.46	7.90
エンジンが冷える	4	8.54	4.81

Page 1 /2　100

…が冷める　77種類

コロケーション	コーパス全体		
	頻度	MI	LD
熱が冷める	31	11.86	7.59
ほとぼりが冷める	17	17.92	11.34
気持ちが冷める	17	9.03	4.79
【数字】が冷める	6	0.68	-3.54
お湯が冷める	6	11.46	7.06
酔いが冷める	5	11.95	7.44
目が冷める	4	5.32	1.09
料理が冷める	3	7.38	3.13
愛情が冷める	3	9.80	5.46
愛が冷める	3	8.13	3.87
心が冷める	3	5.62	1.39
紅茶が冷める	2	10.41	5.93
窯が冷める	2	10.97	6.39
空気が冷める	2	7.84	3.57
熱狂が冷める	2	11.76	6.96
熱意が冷める	2	11.14	6.52
怒りが冷める	2	8.43	4.14

Page 1 /1　100

図 2 単一の見出し語ウィンドウによる類義語の比較

表 1 は、図 2 をもとに、ガ格名詞を手作業でグループごとにまとめたものである。このような比較によって、2 つの動詞のガ格名詞にどのようなものがあるか、おおよその判断はつけられることが分かる。だが、共起語の数が増えてくると、手作業では、見落としや見誤りが出てきたり、客観的な基準で両者を比較することが困難になったりすることが予想される。

表 1　「冷える」と「冷める」のガ格名詞

「冷える」のガ格名詞	「冷める」のガ格名詞
体/全身/お腹/指先/手足/足/足元/腰/下半身が冷える 空気が冷える マグマが冷える エンジンが冷える	熱/熱狂/熱意/ほとぼりが冷める 酔いが冷める 愛/愛情が冷める お湯/紅茶/料理が冷める

今回実装した 2 語比較のプロファイリングの画面では、図 3 のように、2 つの動詞のコロケーションが同じパネルの左右に表示され、一方に顕著なコロケーションは段階的な濃淡でハイライト表示される。上のほうにくるコロケーションほど「冷える」に顕著なもの（「下半身が冷える」、「マグマが冷える」、「手足が冷える」など）、逆に、下のほうにくるコロケーションほど「冷める」に顕著なもの（「ほとぼりが冷める」、「熱が冷める」、「酔いが冷める」など）である。両方の動詞でともに現れているガ格名詞にどのようなものがあるかも、一目で確認できる（「体が冷える/冷める」、「気持ちが冷える/冷める」など）。

図 3　類義語の比較（2 語比較機能）

対照されるコロケーションの顕著さを測る指標として用いているのは、logDice[1]の差（画面上では「LD差」と表示）であり、Sketch EngineのSketch-Diffと同じ手法を用いている。例えば、「下半身が冷える」では、「下半身が冷める」というコロケーションが 1 件もないため、logDice差は 7.9-0=7.9 となる。また、「ほとぼりが冷める」の場合は、「ほとぼりが冷える」というコロケーションがないため、logDice差は 0-11.34=-11.34 となる。「気持ちが冷える/冷める」の場合は、logDice差は 0.7-4.79=-4.09 となり、「冷める」のコロケーションとして優勢だということになる。つまり、この例では、プラスの値の絶対値が大きくなればなるほど、「冷える」のコロケーションとして顕著であり、マイナスの値の絶対値が大きくなればなるほど、「冷める」のコロケーションとして顕著だということになる。

コロケーション比較の尺度においては、NINJAL-LWP は、Sketch-Diff の手法をそのまま採用しているが、インターフェースの面では、Sketch-Diff よりも自由度が高く、分析能力が強化されている。図 4 は、Sketch Engine の Sketch-Diff の画面である（jpTenTen11 で「冷える」の「冷める」の Sketch-Diff を all in one block モードで表示）。1 列目が共起語、2 列目が「冷める」との共起頻度、3 列目が「冷える」との共起頻度、4 列目が「冷める」の logDice、5 列目が「冷える」の logDice で、logDice の差でソートされている（logDice 差の列自体は表示されない）。初期画面では、logDice 差の大きいものがデフォルトで最大 25 まで表示されている。さらに多くの共起語の表示したい場合は、いったん設定画面に戻り、指定し直して再度実行する必要がある。Sketch-Diff の画面はスタティックな画面なので、リストを並べ替えたり、同じ画面にとどまったまま、表示する共起語の数を増減させたりすることはできない。

nounが	26,049	29,117	-1.9	-1.8
足先	0	361	--	7.6
手足	0	938	--	7.2
肝	0	583	--	6.4
マグマ	0	109	--	6.0
身体	0	1,646	--	5.9
溶岩	0	94	--	5.6
背筋	0	210	--	5.6
朝晩	0	80	--	5.6
足腰	0	89	--	5.5
足元	0	548	--	5.5
手先	0	101	--	5.4
末端	0	100	--	5.4
内臓	0	151	--	5.2
御腹	0	947	--	5.0
体	0	6,339	--	5.0
指先	0	261	--	5.0
情熱	209	0	4.8	--
熱気	156	0	5.6	--
余韻	256	0	5.7	--
ホトボリ	44	0	5.8	--
熱	4,772	75	7.4	1.4
興奮	1,243	0	6.3	--
興	158	0	6.3	--
酔い	780	0	7.8	--
ほとぼり	2,658	0	11.3	--

図 4　Sketch-Diff

これに対して、NINJAL-LWP の 2 語比較のコロケーションリストには、並べ替えやフィルターの機能が備わっているので、リストを操作しながら、さまざまな角度から分析することができる。図 5 は、コロケーションリストの簡易フィルタ機能である。頻度と logDice 差と出現位置の 3 つを自由に組み合わせて、表示する共起語の範囲を調整できる。例えば、出現位置を「両方の語」にすると、2 つの動詞で共通する共起語のみが表示される。図 6 は、「冷える」と「冷める」の両方で共通して現れるガ格名詞である。

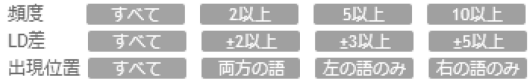


図 5　コロケーションパネルの簡易フィルタ機能

[1] logDice の算出法については、Statistics used in the Sketch Engine（http://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf）を参照してほしい。

…が冷える ― …が冷める

冷える				冷める				LD差
コロケーション	頻度	MI	LD	コロケーション	頻度	MI	LD	
体が冷える	44	8.63	4.98	体が冷める	2	4.75	0.52	4.46
コーヒーが冷える	3	8.2	4.47	コーヒーが冷める	1	7.19	2.91	1.56
空気が冷える	5	8.58	4.87	空気が冷める	2	7.84	3.57	1.3
水が冷える	2	4.78	1.13	水が冷める	1	4.36	0.13	1
肉が冷える	1	5.98	2.28	肉が冷める	1	6.56	2.3	-0.02
酒が冷える	1	5.27	1.59	酒が冷める	1	5.85	1.61	-0.02
お茶が冷える	1	7.02	3.26	お茶が冷める	1	7.59	3.29	-0.03
愛が冷える	2	6.97	3.27	愛が冷める	3	8.13	3.87	-0.6
料理が冷える	1	5.21	1.54	料理が冷める	3	7.38	3.13	-1.59
愛情が冷える	1	7.64	3.83	愛情が冷める	3	9.8	5.46	-1.63
気持ちが冷える	1	4.36	0.7	気持ちが冷める	17	9.03	4.79	-4.09

頻度　すべて　2以上　5以上　10以上
LD差　すべて　±2以上　±3以上　±5以上
出現位置　すべて　両方の語　左の語のみ　右の語のみ
page / 1

図 6　両方の語に現れるコロケーションの表示

3. 比較例

NINJAL-LWP の 2 語比較機能では、同一品詞の内容語（名詞、動詞、形容詞、副詞、連体詞）の比較が可能である。形容詞では、イ形容詞とナ形容詞の比較（「美しい」と「きれいな」）も可能である。また、形容詞と連体詞の比較（「小さい」と「小さな」）も可能である。ここでは、名詞の比較と形容詞の比較を取り上げながら、2 節で触れなかった機能を中心に紹介したい。

3.1 「図書」と「書籍」

最初に、「図書」と「書籍」の使い方の違いを文法パターンの割合に注目して NLB で調べてみることにする。図 7 は、「図書」と「書籍」の文法パターンの割合を示している。ハイライト表示になっている項目は、それぞれの語に顕著なパターンを示している。「図書」の場合、接尾辞が続くパターンや、前項または後項となって複合名詞を形成する割合が、「書籍」の場合と比べ、きわだっていることが分かる。図書／書籍＋接尾辞をクリックして、実際にどのような表現があるかを示したのが図 8 である。「図書」＋接尾辞では、「図書券」と「図書室」が全体の 85%以上を占めることが確認できる。

図書			書籍		
パターン	頻度	比率	パターン	頻度	比率
図書＋名詞・接尾	456	36.1%	書籍＋名詞・接尾	67	8.4%
図書＋名詞	422	33.4%	書籍＋名詞	117	14.7%
名詞＋図書	412	32.6%	名詞＋書籍	101	12.7%
図書＋助詞	375	29.7%	書籍＋助詞	488	61.3%
図書＋の＋名詞	110	8.7%	書籍＋の＋名詞	84	10.6%
名詞＋の＋図書	88	7.0%	名詞＋の＋書籍	81	10.2%
並立	71	5.6%	並立	110	13.8%
(未分類)	63	5.0%	(未分類)	51	6.4%
図書を…	62	4.9%	書籍を＋動詞	90	11.3%
動詞過去＋図書	40	3.2%	動詞過去＋書籍	43	5.4%
動詞基本形＋図書	40	3.2%	動詞基本形＋書籍	38	4.8%
図書に…	24	1.9%	書籍に＋動詞	25	3.1%
図書が＋動詞	24	1.9%	書籍が＋動詞	51	6.4%

すべて表示　±5%以上　±10%以上　±15%以上　±20%以上

図 7　文法パターンの割合の比較

図書＋名詞・接尾 — 書籍＋名詞・接尾

図書 コロケーション	図書 頻度	図書 MI	図書 LD	書籍 コロケーション	書籍 頻度	書籍 MI	書籍 LD	LD差
図書券	73	12.04	9.07					9.07
図書室	321	12.41	9.54	書籍室	1	6.85	1.25	8.29
図書隊	9	8.41	5.48					5.48
図書係	6	8.26	5.3					5.3
図書碑	1	7.76	4.42					4.42
図書貸	3	7.33	4.36					4.36
図書展	1	6.42	3.38					3.38
図書局	3	5.86	2.99					2.99
図書編	1	5.88	2.9					2.9
図書比	1	5.3	2.37					2.37
図書課	2	5.03	2.16					2.16
図書等	14	5.21	2.38	書籍等	5	6.49	0.9	1.48
図書学	1	4.31	1.43					1.43
図書社	1	4.13	1.25					1.25

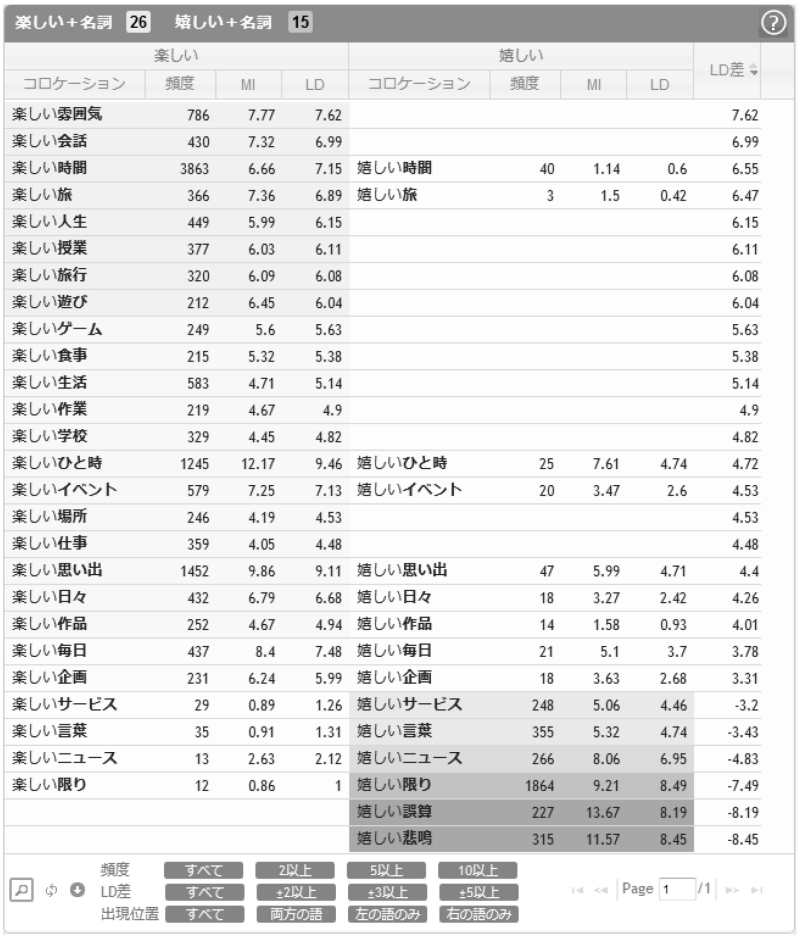
頻度 すべて 2以上 5以上 10以上
LD差 すべて ±2以上 ±3以上 ±5以上
出現位置 すべて 両方の語 左の語のみ 右の語のみ
page / 1

図 8　「図書/書籍＋接尾辞」の比較

3.2 「楽しい」と「うれしい」

次に、形容詞の「楽しい」と形容詞の「うれしい」に続く名詞を比較しながら、2 つの形容詞の違いをNLTで調べてみることにする。まず頻度 200 以上 [2]、logDice差を±3 以上に絞り込むと、図 9 のような結果になる。

2 つの形容詞に後続する名詞をまとめると、表 2 のようになる。両者の比較からそれぞれの形容詞の意味を考えると、「楽しい」は、ある時間、ある場所で活動することで満足した気分になることを表す。それに対して、「うれしい」は、外からの働きかけによって、期待したことが実現し、満足した気分になることを表す。さらに、「うれしい」では、



楽しい＋名詞 26　嬉しい＋名詞 15

楽しい コロケーション	楽しい 頻度	楽しい MI	楽しい LD	嬉しい コロケーション	嬉しい 頻度	嬉しい MI	嬉しい LD	LD差
楽しい雰囲気	786	7.77	7.62					7.62
楽しい会話	430	7.32	6.99					6.99
楽しい時間	3863	6.66	7.15	嬉しい時間	40	1.14	0.6	6.55
楽しい旅	366	7.36	6.89	嬉しい旅	3	1.5	0.42	6.47
楽しい人生	449	5.99	6.15					6.15
楽しい授業	377	6.03	6.11					6.11
楽しい旅行	320	6.09	6.08					6.08
楽しい遊び	212	6.45	6.04					6.04
楽しいゲーム	249	5.6	5.63					5.63
楽しい食事	215	5.32	5.38					5.38
楽しい生活	583	4.71	5.14					5.14
楽しい作業	219	4.67	4.9					4.9
楽しい学校	329	4.45	4.82					4.82
楽しいひと時	1245	12.17	9.46	嬉しいひと時	25	7.61	4.74	4.72
楽しいイベント	579	7.25	7.13	嬉しいイベント	20	3.47	2.6	4.53
楽しい場所	246	4.19	4.53					4.53
楽しい仕事	359	4.05	4.48					4.48
楽しい思い出	1452	9.86	9.11	嬉しい思い出	47	5.99	4.71	4.4
楽しい日々	432	6.79	6.68	嬉しい日々	18	3.27	2.42	4.26
楽しい作品	252	4.67	4.94	嬉しい作品	14	1.58	0.93	4.01
楽しい毎日	437	8.4	7.48	嬉しい毎日	21	5.1	3.7	3.78
楽しい企画	231	6.24	5.99	嬉しい企画	18	3.63	2.68	3.31
楽しいサービス	29	0.89	1.26	嬉しいサービス	248	5.06	4.46	-3.2
楽しい言葉	35	0.91	1.31	嬉しい言葉	355	5.32	4.74	-3.43
楽しいニュース	13	2.63	2.12	嬉しいニュース	266	8.06	6.95	-4.83
楽しい限り	12	0.86	1	嬉しい限り	1864	9.21	8.49	-7.49
				嬉しい誤算	227	13.67	8.19	-8.19
				嬉しい悲鳴	315	11.57	8.45	-8.45

頻度 すべて 2以上 5以上 10以上
LD差 すべて ±2以上 ±3以上 ±5以上
出現位置 すべて 両方の語 左の語のみ 右の語のみ
Page 1 /1

図 9　「楽しい」と「うれしい」の後続名詞の比較

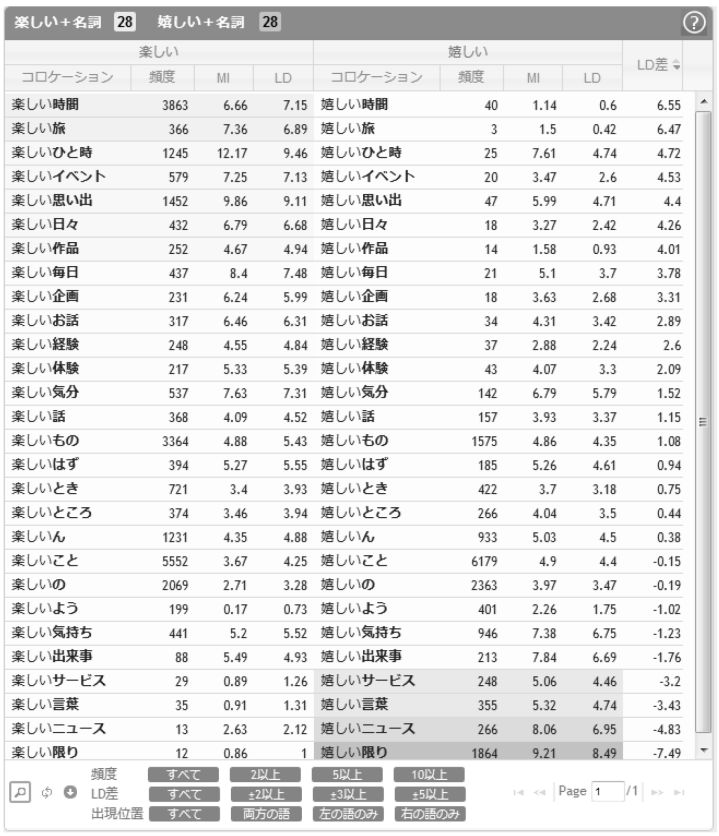
[2] フィルターダイアログを利用すると、任意の頻度、logDice 差を設定することが可能。

意外性を表す語が後続することから分かるように、期待以上の状況が生まれることも含意している。「うれしい悲鳴」という慣用表現にもそのような意外性が含まれている。

表 2　「楽しい」と「うれしい」に後続する名詞

楽しい
時間を表す語…時間、ひと時、日々、毎日 活動を表す語…旅、レッスン、授業、食事 娯楽を表す語…会話、おしゃべり、遊び、ゲーム その他…雰囲気、人生、生活、イベント
うれしい
知らせを表す語…ニュース、知らせ、報告 意外性を表す語…誤算、驚き、サプライズ その他…感想、特典、配慮、再会 慣用表現…悲鳴

では、両方の形容詞に共通して出現する後続名詞を調べてみよう。少なくとも一方の共起頻度 200 以上で、出現位置を「両方の語」にした結果が図 10 である。


楽しい＋名詞 28　嬉しい＋名詞 28

楽しい				嬉しい				LD差
コロケーション	頻度	MI	LD	コロケーション	頻度	MI	LD	
楽しい時間	3863	6.66	7.15	嬉しい時間	40	1.14	0.6	6.55
楽しい旅	366	7.36	6.89	嬉しい旅	3	1.5	0.42	6.47
楽しいひと時	1245	12.17	9.46	嬉しいひと時	25	7.61	4.74	4.72
楽しいイベント	579	7.25	7.13	嬉しいイベント	20	3.47	2.6	4.53
楽しい思い出	1452	9.86	9.11	嬉しい思い出	47	5.99	4.71	4.4
楽しい日々	432	6.79	6.68	嬉しい日々	18	3.27	2.42	4.26
楽しい作品	252	4.67	4.94	嬉しい作品	14	1.58	0.93	4.01
楽しい毎日	437	8.4	7.48	嬉しい毎日	21	5.1	3.7	3.78
楽しい企画	231	6.24	5.99	嬉しい企画	18	3.63	2.68	3.31
楽しいお話	317	6.46	6.31	嬉しいお話	34	4.31	3.42	2.89
楽しい経験	248	4.55	4.84	嬉しい経験	37	2.88	2.24	2.6
楽しい体験	217	5.33	5.39	嬉しい体験	43	4.07	3.3	2.09
楽しい気分	537	7.63	7.31	嬉しい気分	142	6.79	5.79	1.52
楽しい話	368	4.09	4.52	嬉しい話	157	3.93	3.37	1.15
楽しいもの	3364	4.88	5.43	嬉しいもの	1575	4.86	4.35	1.08
楽しいはず	394	5.27	5.55	嬉しいはず	185	5.26	4.61	0.94
楽しいとき	721	3.4	3.93	嬉しいとき	422	3.7	3.18	0.75
楽しいところ	374	3.46	3.94	嬉しいところ	266	4.04	3.5	0.44
楽しいん	1231	4.35	4.88	嬉しいん	933	5.03	4.5	0.38
楽しいこと	5552	3.67	4.25	嬉しいこと	6179	4.9	4.4	-0.15
楽しいの	2069	2.71	3.28	嬉しいの	2363	3.97	3.47	-0.19
楽しいよう	199	0.17	0.73	嬉しいよう	401	2.26	1.75	-1.02
楽しい気持ち	441	5.2	5.52	嬉しい気持ち	946	7.38	6.75	-1.23
楽しい出来事	88	5.49	4.93	嬉しい出来事	213	7.84	6.69	-1.76
楽しいサービス	29	0.89	1.26	嬉しいサービス	248	5.06	4.46	-3.2
楽しい言葉	35	0.91	1.31	嬉しい言葉	355	5.32	4.74	-3.43
楽しいニュース	13	2.63	2.12	嬉しいニュース	266	8.06	6.95	-4.83
楽しい限り	12	0.86	1	嬉しい限り	1864	9.21	8.49	-7.49

頻度　すべて　2以上　5以上　10以上
LD差　すべて　±2以上　±3以上　±5以上
出現位置　すべて　両方の語　左の語のみ　右の語のみ
Page 1 /1

図 10　「楽しい」と「うれしい」の両方に後続する名詞

このリストのなかから、比較的両方の頻度の高い「楽しい気分」と「うれしい気分」を例にとって、２つの表現にどのような意味の違いがあるかを実際の用例を見ながら調べてみたい。コロケーションリストのコロケーションの列をクリックすると、文法パターンパネルが閉じて、代わりにコロケーションリストの右に用例パネルが現れて、そのコロケーションの用例が表示される。「楽しい気分」をクリックしたときの画面が図 11 である。

楽しい コロケーション	楽しい 頻度	楽しい MI	楽しい LD	嬉しい コロケーション	嬉しい 頻度	嬉しい MI	嬉しい LD	LD差
楽しい時間	3863	6.66	7.15	嬉しい時間	40	1.14	0.6	6.55
楽しい旅	366	7.36	6.89	嬉しい旅	3	1.5	0.42	6.47
楽しいひと時	1245	12.17	9.46	嬉しいひと時	25	7.61	4.74	4.72
楽しいイベント	579	7.25	7.13	嬉しいイベント	20	3.47	2.6	4.53
楽しい思い出	1452	9.86	9.11	嬉しい思い出	47	5.99	4.71	4.4
楽しい日々	432	6.79	6.68	嬉しい日々	18	3.27	2.42	4.26
楽しい作品	252	4.67	4.94	嬉しい作品	14	1.58	0.93	4.01
楽しい毎日	437	8.4	7.48	嬉しい毎日	21	5.1	3.7	3.78
楽しい企画	231	6.24	5.99	嬉しい企画	18	3.63	2.68	3.31
楽しいお話	317	6.46	6.31	嬉しいお話	34	4.31	3.42	2.89
楽しい経験	248	4.55	4.84	嬉しい経験	37	2.88	2.24	2.6
楽しい休験	217	5.33	5.39	嬉しい休験	43	4.07	3.3	2.09
楽しい気分	537	7.63	7.31	嬉しい気分	142	6.79	5.79	1.52
楽しい話	368	4.09	4.52	嬉しい話	157	3.93	3.37	1.15
楽しいもの	3364	4.88	5.43	嬉しいもの	1575	4.86	4.35	1.08
楽しいはず	394	5.27	5.55	嬉しいはず	185	5.26	4.61	0.94
楽しいとき	721	3.4	3.93	嬉しいとき	422	3.7	3.18	0.75
楽しいところ	374	3.46	3.94	嬉しいところ	266	4.04	3.5	0.44
楽しいん	1231	4.35	4.88	嬉しいん	933	5.03	4.5	0.38
楽しいこと	5552	3.67	4.25	嬉しいこと	6179	4.9	4.4	-0.15
楽しいの	2069	2.71	3.28	嬉しいの	2363	3.97	3.47	-0.19
楽しいよう	199	0.17	0.73	嬉しいよう	401	2.26	1.75	-1.02
楽しい気持ち	441	5.2	5.52	嬉しい気持ち	946	7.38	6.75	-1.23
楽しい出来事	88	5.49	4.93	嬉しい出来事	213	7.84	6.69	-1.76
楽しいサービス	29	0.89	1.26	嬉しいサービス	248	5.06	4.46	-3.2
楽しい言葉	35	0.91	1.31	嬉しい言葉	355	5.32	4.74	-3.43
楽しいニュース	13	2.63	2.12	嬉しいニュース	266	8.06	6.95	-4.83
楽しい限り	12	0.86	1	嬉しい限り	1864	9.21	8.49	-7.49

頻度 すべて 2以上 5以上 10以上
LD差 すべて ±2以上 ±3以上 ±5以上
出現位置 すべて 両方の語 左の語のみ 右の語のみ
Page 1 /1

楽しい気分が台無しになります。
(気分屋で短気な夫（長文です） - 夫婦・家族 - 教えて！goo)
楽しい気分が台無しになります。
(気分屋で短気な夫（長文です） - 質問・相談ならMSN相談箱)
明るく、楽しい気分で過ごした。
(足立区　相談の受け方)
楽しい気分に満ちた挿絵でした。
(アンケートの答え)
院長が明るく楽しい気分になる。
(患者さんの声　::交通事故や腰痛治療のパルモ東大井整骨院::)
楽しい気分になっちゃいますよね！
(One's Wan!!　新しい家族待ってま～す。 クリスマススペシャル企画！)
リラックスして楽しい気分で乗る。
(Marinedoor.com)
だから今もとても楽しい気分だよ。
(RPG Data Library：アンジェリーク会話集２)
すっかり楽しい気分になったらしい。
(--Baltimore滞在記--)
楽しい気分にさせられてしまうのだ。
(讃岐うどん)
とっても楽しい気分になれる曲です。
(アクロ・スタント技の辞典)
楽しい気分でいることはまれである。
(障害児のきょうだい達の心の健康～きょうだい達をどう健やかに育てるか～)
と思い少し楽しい気分になりました。
(マッチスティック・メン評価)
客を楽しい気分にさせ、笑いを誘う。
(その３ユーモアトークの考え方)
楽しい気分になりたいと親子はいう。
Page 1 /6 100
537件中　1 - 100を表示

図 11　用例パネルの表示（「楽しい気分」）

同様に、「うれしい気分」をクリックして、順にその用例を確認する。以下に、「楽しい気分」と「うれしい気分」の用例を一つずつ挙げる（下線は筆者）。

(2) 賑やかな能が多く、観客は楽しい気分で家路につく。（世阿弥の生涯、http://kajipon.sakura.ne.jp/kt/haka-topic14.html）

(3) 個人懇談では先生が一生懸命褒めてくださり、ちょっと嬉しい気分です。（SkipMamaClass すきっぷママクラス PDD 子育て支援セミナー、http://www.jsnhc.org/smc/navis/navi2/syuuryouseikansou.html）

(2)では、能の鑑賞（活動）に満足していることを表している。一方の(3)は、個人面談で褒めてもらえた（外からの働きかけ）ので満足していることを表している。先ほどの「楽しい」と「うれしい」の後続名詞の分析の結果と一致していることが確認できる。

4. まとめ

本稿では、コロケーション分析におけるレキシカルプロファイリングの利点を挙げ、NINJAL-LWP に実装された 2 語比較機能の概要を述べ、実際の比較例を紹介した。2 語比較機能は、類義語の分析のみならず、反義語の分析にも応用できる。例えば、「強い」と「弱い」のガ格名詞を比較すると、片方の語にのみ共起する名詞の特徴を詳細に調べることができる（「イメージが強い」というが「イメージが弱い」とはあまり言わない、など)。今後、NINJAL-LWP がコーパス基盤の類義語の研究に広く活用されることを期待したい。

文　献

Curran, James (2004) *From Distributional to Semantic Similarity*. PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Gatto, Maristella (2014) Web As Corpus: Theory and Practice. A&C Black.

今井 新悟、赤瀬川 史朗、プラシャント パルデシ (2013)「筑波ウェブコーパス検索ツール NLT の開発」第 3 回コーパス日本語学ワークショップ予稿集、pp.199-206. (http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_web.pdf よりダウンロード可能)

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell (2004) The Sketch Engine. In *Proceeding of EURALEX 2004*, pp.105-116.

Murphy, M. Lynne (2010) Lexical Meaning. Cambridge University Press.

プラシャント・パルデシ、赤瀬川史朗 (2011)「BCCWJ を活用した基本動詞ハンドブック作成―コーパスブラウジングシステム NINJAL-LWP の特徴と機能―」『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集、pp.205-216.

プラシャント・パルデシ、赤瀬川史朗 (2012)「レキシカルプロファイリング手法を用いた BCCWJ 検索ツール NINJAL-LWP とその研究事例」日本言語学会第 144 回大会予稿集、pp.364-369.

Rychlý, Pavel (2008) A Lexicographer-Friendly Association Score. In *Proceedings of Recent Advances In Slavonic Natural Language Processing, RASLAN 2008,* pp.6-9.

関連 URL

NLB (NINJAL-LWP for BCCWJ) http://nlb.ninjal.ac.jp/

NLT (NINJAL-LWP for TWC) http://corpus.tsukuba.ac.jp/

The Sketch Engine https://the.sketchengine.co.uk/

述語項構造を意識した名詞の意味構造アノテーションのための名詞意味構造の検討

竹内 孔一 (岡山大学大学院自然科学研究科)[1]

Consideration of Semantic Structure for Nominal Noun on Predicate-Argument Thesaurus

Koichi Takeuchi (Graduate School of Natural Science and Technology, Okayama University)

要旨

動詞，形容詞，形容動詞といった述語の状態変化性を主に取り上げて述語項構造シソーラスを構築してきた．一方で，文の意味は述語だけで無く，名詞が概念を持つためこれらの意味を記述して取り込む必要がある．名詞概念の意味記述は様々な分野で提案されているが，本稿では言語学で分析されてきた西山の名詞句の分析 (西山 (2003, 2013))，ならびに影山の名詞句に関する分析 (影山 (2011)) を元にした項構造をベースの名詞意味記述を仮定してアノテーションのためのフレームワークを考察する．具体的には西山で提案されてきた非飽和名詞という考え方をベースに，必要とする要素を生成語彙論 (Generative Lexicon) の枠組みを利用して記述する影山の方法を採用する．さらに周辺の情報として，名詞述語文との類義データの構築，機能語表現と類義データの構築を一貫した項構造レベルで統一して記述し，述語項構造シソーラスの一部として拡張することを提案する．

1　はじめに

述語の持つ項構造を一様な方法で記述することで，述語間の交替関係だけでなく，状態変化と結果状態といった語の概念に特徴的な含意関係を記述してきた．例えば,「育てる」「鍛える」という語について下記の例文のように，述語の係り元 (項) に対して意味役割を付与し，述語に対しては語義概念を付与している．

- 母が [動作主] 僕を [対象 (人)] 一人前に [着点 (状態)] 育てる [成長]
- コーチが [動作主] 彼女を [対象 (人)] 一流選手に [着点 (状態)] 鍛える [成長]

ここで，名詞句における [ ] 内は意味役割を現し，述語における [ ] 内は語義概念を現す．

語義概念には語彙概念構造 (影山 (1996)) を拡張した構造を記述しており，状態変化と状態といった含意関係を記述している．現在公開してるのは動詞だけであるが[2]，内部では，形容詞，形容動詞についても付与を行っている (竹内他 (2013))．例えば，形容詞，形容動詞は下記のような例がある．

- アントシアニンは [対象] 目に [目的] 良い/効果的だ [可能]

意味役割の種類は現在，基本種類が 29 種類に対して属性との組み合わせとなり，約 90 種類程度定義されている．また述語の概念は約 1000 種類ある．各概念には述語と例文が登録されおり，約 1.1 万語に対して約 2 万例文が存在し，各例文には上記の例に示したように，意味役割と概念が付与されている．また，この体系に従って日本語書き言葉均衡コーパス (BCCWJ)(約 6000 文) の動詞に対して付与を行っている (竹内・上野 (2013))．

一方で，こうした述語項構造を言語理解の課題，例えば NTCIR-10，RITE-2 含意認識タスク[3] に利用しようと考えたとき，名詞の意味構造がほとんど課題の中心であることが分かる．

- (t1) BLT サンドイッチとは、サンドイッチの一種であり、パンに挿む食材として、ベーコン、レタス、トマト が用いられることから、それぞれの頭文字を取って名づけられた。

[1] koichi@cl.cs.okayama-u.ac.jp
[2] 動詞項構造シソーラス http://vsearch.cl.cs.okayama-u.ac.jp/.
[3] http://research.nii.ac.jp/ntcir/ntcir-10/index-ja.html.

- (t2) サンドイッチの略称として具材となるベーコン、レタス、トマトの頭文字 BLT が用いられるものがある。

含意認識タスクとは，t2 の文の内容が t1 に含まれているなら真，含まれていないなら偽と判断するタスクで，専門知識が無くてもネイティブならば読んだだけで判断できるペアが構築され公開されている．この文では,「サンドイッチ」という物のタイプを表す名詞,「一種」,「略称」,「具材」「頭文字」など名詞の中でも特別な機能を持つものが存在する．この例では「名づける」という動詞と「略称」との名前の関係,「頭文字」が何を指しているかなどが解けないと，含意認識を正しく判定することは出来ない．

こうした名詞の意味の記述にはどのような研究がなされてきたであろうか? 分野を超えて調べたところでは，言語学の他に，既に人工知能の分野で Winograd (1972) と高木・伊東 (1987) の研究が有り，人工物の概念として設計論の立場から冨山他 (1998) がある．結局のところ形式的には Minsky のフレーム理論を拡張した形 (素性構造をオブジェクトとして概念を記述) に集約できる (竹内他 (2014)) と考えられる．よって形式的な側面からは素性構造で記述することになるが，何をどう書くかが問題である．

上記の事例を整理してみると抽象的な名詞に対して 2 つの関係が存在すると考えられる．

- X の 一種/略称/頭文字 は Y

このような名詞に対する分析は日本語では西山 (2003, 2013) が非飽和名詞として「A の B」の分析から明確な分類を提案おり，さらに影山 (2011) は名詞まわりの項 (上記 X と Y) を生成語彙論 (Generative lexicon (Pustejovsky (1995))，以下 GL) の枠組みを利用して，意味構造で記述することを提案している．

そこで，本研究では影山の意味構造を参考に付与可能な名詞の意味構造について考察する．言語資源としては，既に，日本語では名詞格フレーム笹野他 (2005) として自動構築され，公開されている．名詞格フレームは大規模である一方で，人手による整備がないため，関係名や例文が整理がなされていない．英語では既に NomBank が名詞の項について，意味役割的な関係と，例文を付与して公開している (Meyers et al. (2004))．こうした先行研究を受けて，我々は，既に構築している述語項構造シソーラスの意味役割体系を利用することで一貫した意味的関係を持ちつつ，例文ベースの名詞の項構造と事例を構築する．

さらに名詞周辺の表現として，名詞を含む述語表現と動詞との関係，ならびに機能語表現の集約も行い，述語項構造辞書に統合する．これらの構造について下記に考察を行う．

2　付与可能な名詞の意味構造

まず先行研究での記述法 (日本語と英語について) を紹介してから，どのような記述が可能かについて考察する．影山 (2011) は GL の qualia structure を,【外的分類】,【内的構成】,【目的・機能】,【成り立ち】，と解釈し直して,「俳優」と「主役」の項の取り方の違いを概略，下記のような構造で記述することを提案している．これは「主役」という名詞は「そのお芝居 [w] の主役は太郎 [y] だ」

	「俳優」	「主役」
外的分類	人間 (x)	人間 (y)
目的・機能	x が芝居や映画で劇中の人物を演じる	y が芝居や映画で劇中の人物を演じる
成り立ち		y が [w] の主要人物の役を務める

のように何か主体になるもの [w] に対して成り立つ言葉で有り，さらに，主役そのものである項 [y] を取ることを意味している．さらに，外的分類で y は人間であることを示している．

一方で「俳優」の場合,「X の俳優は Y」という文では特に「俳優」が取り込む項として X が存在するわけではない．この例文での X と「俳優」との意味的関係は西山 (2003) にあるように談話から与えられる関係であり，例えば「北海道の俳優」ならば「北海道で活躍する俳優」なのか「北海道で生まれた俳優」なのかは名詞の意味構造からではなく背景知識や文脈で決まる．

この構造は「主役」などの名詞の意味を記述した構造としては機能するが一方で言語処理の観点から「X の主役は Y」の X が例えば「昨日の主役は Y」のように項ではなく，西山 (2013) が示す時間に関する表現の場合もある[4]．こうした曖昧性解消には事例が必要であり，さらにその関係を一貫した意味関係で記述できることが望ましい．これを具現化したのが NomBank である．

例えば knowledge という単語に対しては，

- students' knowledge of two-letter consonant sounds
  ARG0 = students, REL = knowledge, ARG1 = of two-letter consonant sounds

このように，knowledge の内容部分を ARG1，knowledge を所有する人を ARG0 と PropBank と同様の意味役割の定義を利用して記述する[5]．さらに，辞書内では，ARG0 は "KNOWER" で AGENT, ARG1 は "THING KNOWN OR THOUGHT" で THEME と各語に依存した詳細な意味役割が付与されている．よってこうした事例と一貫した意味役割を付与した構造は記述として参考にしたい．

これまでの議論では名詞と項との関係に着目してきたが，名詞と項の関係に大きなタイプ分けが存在して分類するという観点がある．例えば西山 (2013) では譲渡不可能名詞という「なべの蓋」や「鼻」「耳」といった必ず主体を必要とする名詞の分類を提案している．こうした全体-部分関係は，述語表現でも観測されることから述語項構造シソーラスでは，意味役割に組み込んで記述してきた．

- 彼が [動作主] 姿を [身体部分] 現す

この例の場合では,「現す」の直接的な対象は「姿」であるが，それは「彼」の一部であることを指している．

名詞の意味構造記述において，西山や影山では明示的な提案がなされていないが，言語理解の観点からは，こうした全体-部分の関係をより一般化したものとして属性 (Attribute) を分けて，素性として記述することが行われる (Winograd (1972); 高木・伊東 (1987))．例えば「車の色は赤だ」「赤い色の車」「赤い車」 はどれも「車の色が赤」であることを述べているが，こうしたものを言語理解として扱う場合,「色」という属性名詞を意味構造内で特別扱いして (属性として取り上げ)，その属性値が「赤」として扱うと，属性どうしの比較など処理が行いやすい．

上記の NomBank では名詞を 20 種類以上分けており，その中で「色」「重さ」など ATTRIBUTE という分類で記述されている．例えば color の場合

- the colors and emblems of both teams
  ATTRIBUTE, REL = color ARG1 = of both teams

と事例が記述されている．意味役割として ARG1 が THEME，ARG2 が VALUE と記述されており，事例には無いが「赤」など色がある場合も記述可能な枠組みである．こうした分類は日本語でも同様と考えられるため，言語理解を目標にすると取り込む必要がある．ただし，NomBank の事例は PennTreeBank を利用して半自動で構築されていることから，名詞にまつわる項構造の全ての要素があるわけではなく，事例として不足がある．よって例文はコーパスだけでなく作例を許し，なるべく想定される名詞の項が埋まった事例を記録していく必要がある．

さて，上記の分析を踏まえて，名詞の意味構造を下記の重要度順で構築したい．

1 名詞と名詞が取る項の例文を作成し，述語項構造シソーラスの意味役割を付与する

[4]例えば「昭和 60 年代の東京」,「あの頃の自分」など．
[5]ただしこの例はアノテーションマニュアルの事例であり，version 1.0 内には ARG0 が含まれる事例は無い．

2 名詞を属性，譲渡不可能名詞，他，カテゴリ分けする

3 BCCWJ で事例を探し，名詞の項構造を同様に付与する

4 GL の特質構造に基づく影山らが提案する構造を日本語文混じりで記述する
　　可能な限り述語項構造辞書を利用して記述する

ここで **4** については説明が必要である．上述のほとんどの議論は **1** の例文の重要性について記述したが，最初に記した影山らの GL ベースの記述では名詞の意味そのものを記述するため，名詞が動詞と明確な関係がある場合，言語理解において重要な含意関係を扱うことが出来る．

例えば,「伊坂幸太郎は『重力ピエロ』の著者である」は「伊坂幸太郎が書いた『重力ピエロ』」の意味を含意しているが，これらの関係は「著者」の特質構造における【成り立ち】内で「書く」と

	「著者」」
外的分類	人間 (x)
成り立ち	書く (動作主 [x], 対象 [w])

いった記述が成されていれば用意に結び付けることが出来る．また述語項構造シソーラス内に記述されている概念を利用すれば「記す」など「書く」の類義表現も含意関係を識別することが出来る．

3　名詞を含む述語表現の統合

例えば,「異なる」は「相違がある」といった名詞を含む述語表現と類義関係であると考えられる．述語項構造シソーラスには既に,「異なる」は意味役割とともに記述されているので，同様の形式で，こうした名詞を含む類義表現を類語辞典などから見つけ出し，登録する．

- 事実が [対象] 報道と [相互] 異なる [相違]
- 事実が [対象] 報道と [相互] 相違がある [相違]
- 報道と事実に [対象] 差がある [相違]

これらは一種の連語であるが NomBank にも Support Verb という分類で，名詞が述語の一部として振る舞う場合の項構造が記述されている．これにより述語項構造シソーラスに名詞を含む類義表現も扱うことが出来る．

4　機能語に関する類義表現の統合

機能表現辞書として「つつじ」[6]が構築されいるが，こうした機能表現との類義語は動詞で存在し，述語項構造シソーラスで既に登録されている．そのため，機能語に対して例文と意味役割を付与することが出来れば，機能語表現との言い換えを扱うことが出来る．

例えば,「予測」の意味で「〜だろう，〜かもしれない」では

- 樹齢は [対象]　千年 [補語相当] かもしれない/だろう
- 樹齢を [対象]　千年と [補語相当 (を格)] 予測する/推定する

と予測や推測といった分類の動詞と対応する．また,「放置」の意味で

- 宿題を [対象]　し [連語] ないでおく
- 宿題を [対象]　放置する/ほったらかす

「〜しないでおく」は「ほったらかす」 など関係がある．こうした機能表現を例文付きで構築することで，品詞を超えた類義表現を幅広く集約することが出来る．

[6] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/.

5　まとめ

本稿では，既に構築している述語項構造シソーラスの語義概念と意味役割の体系を利用して，非飽和名詞に対する名詞項構造事例構築のための構造を検討した．その結果，まず対象とする名詞と意味的関係がある項を含む例文の作成から着手し，意味役割の付与や名詞の分類を行うことが重要であることを明らかにした．さらに，BCCWJ への付与や Generative Lexicon ベースの意味構造が付与できれば，含意認識タスクでの扱える言い換えの範囲が広まることを示した．また名詞を含む類義語表現，ならびに機能表現の統合についても検討を加えた．

今後の課題として，意味構造を付与するための付与システムを構築し，実際に名詞の意味構造を付与しながら，付与できる範囲の意味構造を明確にする必要がある．

謝辞

本研究は，文部科学省科学研究費補助金基盤研究 (C)「言語処理及び言語分析を指向した大規模コーパスを利用した述語シソーラスの拡張」(平成 26〜28 年度，代表者: 竹内孔一) による補助を得ています．

文献

Adam Meyers, Ruth Reeves, and Catherine Macleod (2004) “NP-External Arguments: A Study of Argument Sharing in English,” in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pp. 96–103.

J. Pustejovsky (1995) *The Generative Lexicon*: MIT Press.

Terry Winograd (1972) *Understanding Natural Language*: Academic Press.

影山太郎 (1996) 動詞意味論，くろしお出版．

影山太郎 (2011) 日英対照 名詞の意味と構文，大修館書店．

高木朗、伊東幸宏 (1987) 自然言語の理解，丸善出版．

笹野遼平、河原大輔、黒橋禎夫 (2005)「名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析」，自然言語処理，第 12 巻，第 3 号，pp.129–144．

西山佑司 (2003) 日本語名詞句の意味論と語用論，ひつじ書房．

西山佑司（編） (2013) 名詞句の世界，ひつじ書房．

竹内孔一、上野真幸 (2013)「日本語コーパスに対する動詞項構造シソーラスの概念と意味役割のアノテーション」，言語処理学会第 19 回年次大会，pp.162–165．

竹内孔一、竹内奈央、石原靖弘 (2013)「述語項構造のシソーラス分類と意味役割の設計について」，人工知能学会全国大会，2D4–OS–03a–1．

竹内孔一、竹内奈央、石原靖弘 (2014)「述語項構造シソーラスによる述語と名詞の構造化」，人工知能学会全国大会，2I5–OS–08b–1．

冨山哲男、桐山孝司、梅田靖、下村芳樹、吉岡真治 (1998)「第 5 章モデルに重点を置いたアプローチ」，工学知識のマネージメント，pp.180–229，朝倉書店．

付録

意味役割の種類

現段階での意味役割の種類を記述する．大きく 4 つの類にわけて，中分類で 29 種類を定義した．さらにこれらに対して属性タイプが付与される形である．まず 29 種類を以下に示す．

構文類　● 連語，外の関係，補語相当

対象類　● 経験者，被使役者，対象，基準, 相互, 起点，着点，起点・着点，通過点，経路，方向

動作主類　● 使役，原因，動作主，使役者，手段

条件周辺類　● 限界，領域，場所，時間，条件，様態，程度，目的，順接，逆接

これらに対して，属性タイプが一部付与される．例えば,「村民が悪者を懲らしめる」の場合,「懲らしめる」の対象は基本的には人であるはずである．こうした項の属性タイプが語義の概念から決まる場合にはタイプを付与している．

ここでは対象類について記述すると下記のように行列で現すことが出来る．列に意味役割，行に属性タイプを示す．下記の表における記号 “○” 印のところが，その属性タイプで限定される意味役割があることを示している．

	経験者	対象	基準	相互	起点	着点	起点・着点	通過点	経路	方向
人		○	○	○	○	○				○
操作対象	○									
程度						○				
場所					○	○				
時					○				○	
感情		○								
身体部分		○			○	○				
状態						○				
抽象						○		○	○	
材料					○					
生成物		○				○				
動作		○								
事態		○								

テキストとアノテーションの汎用同時検索システム

狩野 芳伸（科学技術振興機構 さきがけ）† 増田 勝也（東京大学 大学総合教育研究センター）

A Generic Searching System for Text and Annotations

Yoshinobu Kano (PRESTO, Japan Science and Technology Agency)　Katsuya Masuda (Center for Research and Development of Higher Education, the University of Tokyo)

テキストデータの検索は、商用・無償ともに様々なシステムが提供されており、広く利用される一般的な技術になりつつある。一方で、テキストデータに付与されたアノテーションの検索については、理論・実装いずれも整備されているとはいいがたい状況であり、標準的なツールも存在しない。しかし、複雑な構造のアノテーションから研究者が興味あるアノテーションを抜き出すことは、自然言語処理など工学的な分野だけでなく言語学など事例に基づく研究においても重要な作業である。また、異なる研究者が開発したアノテーションを同時に検索できるようにすることは、アノテーションの利用において大きな意味がある。本稿では、BCCWJ コアデータを主な対象として、テキストとアノテーションを同時に検索するためのクエリ設計と、そのクエリで検索を行う汎用検索システムの実装について述べる。

１．はじめに

本研究の目的は、将来の新規アノテーション種へも対応できるよう、検索の下位レイヤを汎用のシステムを設計し実装することにある。すなわち、ユーザビリティの向上や GUI の開発などは将来の課題とし、本研究では拡張性のある汎用システムを構築した。

前述のように、テキスト一般の検索システムについては高性能かつ安定したものが長年にわたりオープンソースで公開され、非常に多くのユーザに利用されている。この部分を新たに構築しても、性能面でも安定性でも勝るものを作成するのは現実的に無理である。そのため、テキスト検索に関わる部分はデファクトスタンダードのオープンソースライブラリである Apache Solr/Lucene[1]を内部的に利用することとした。執筆時の最新メジャーバージョンである Solr 4 では、速度向上とスケーラビリティ向上の面で大幅な機能追加がなされており、すでに多くの企業において超大規模のテキスト検索サービスを構築するために利用されている。

アノテーションの検索システムについては、知る限り標準といえるものが存在しない。本研究では内部的に Solr を用いつつ、アノテーションとテキストの検索ができるシステムを新たに構築した。アノテーション検索に最も近いと思われる技術は XML の検索であるが、整形式の XML 文書ではタグの交差が許されないため、アノテーション一般の検索をすることはできない。我々の設計では、アノテーションの検索クエリについて領域代数（Region Algebra）を基本とし、アノテーション一般の検索に対応する汎用性を担保した。

２．背景

２.１ Apache Solr

Apache Solr は、Java で実装されたオープンソースの全文検索ライブラリである。Solr のコア部分は Apache Lucene と呼ばれており、Solr はおおむね Lucene にさまざまな付加機能を追加したものになっている。Solr の機能は多岐にわたり、バージョンアップの度に新機能が追加され非常に活発に開発が続けられている。Solr 4 では SolrCloud 機能が追加

† kano@nii.ac.jp

[1] http://lucene.apache.org/solr/

され、複数台構成のサーバ群を一つの検索システムとして運用するための機能と、それを容易に管理できる各種ユーテリティが追加された。

Solr の特徴は、多くのユーザを持つことによる安定性と関連情報量の多さに加え、非常に高速な検索を実現していることが挙げられる。これは同時にスケーラビリティにも優れていることを意味する。Solr ウェブサイトのデータ [2]によると、たとえば電子図書館 HathiTrust でのサービスにおいて、500 万冊の書籍を全文検索させている [3]。100 万冊の場合のベンチマークを以下に引用する。テキストデータサイズは 666GB とあるので、100 万冊は日本語の場合 333G 文字程度、平均単語長が 3 文字程度とすると 100G（一千億）単語程度に相当すると思われる。インデックスサイズはデータの 35%程度である。このベンチマークはやや古いため、サーバは一台、サーバの搭載メモリは 4GB-8GB 程度である。ほとんどのクエリに 1 秒以内で返答している。現在においては、SSD の利用とマルチコアサーバにより、さらに数倍以上の高速化が安価に可能と考えられる。

２.２ 領域代数

領域代数(Clarke et al.,1995, Jaakkola et al. 1999, Masuda et al. 2009)とは領域集合間の関係性を記述する演算の集合である。領域は連続するテキスト列として定義され、開始・終了位置の組で表現することができる。各演算は領域集合を引数としてとり、演算結果としてまた領域集合を返す。本研究においては、タグで囲われたテキスト範囲を領域として考えることで、領域代数表現をクエリ、演算処理を検索実行処理として、領域代数をアノテーションの付与されたテキストに対する検索として利用する。領域代数には標準的な演算集合といえるものは存在しないが、我々は以下で説明する演算集合を採用する。説明をわかりやすくするため、XML/HTML 風のタグ記述を用いて説明する。また以下では領域集合 A,B に対する演算を考える。A,B はその名前を持つタグ(<A>,<B>)によるアノテーション領域の集合であり、a∈A, b∈B は単一のアノテーション領域を指す。

２.２.１ 包含演算（contain）

包含演算子は、領域 a が領域 b を包含することを表す。つまり<A>…<B>…</B>…</A>という関係である。演算結果は上記の条件を満たす領域 a の集合となる。

２.２.２ 被包含演算（contained in）

被包含演算子は、領域 b が領域 a を包含することを表す。つまり<B>…<A>…</A>…</B>という関係である。演算結果は上記条件を満たす領域 a の集合となる。包含演算子とは演算結果が包含する側の領域か包含される側の領域であるか、という点で異なる。

２.２.３ 順序演算（follows）

順序演算子は、領域 b が領域 a より後に来る領域を返す。つまり<A>…</A> … <B>…</B>という関係である。ただし領域 a と 領域 b が交差する場合は除く。領域としては、a で始まり b で終わる領域となる。

２.２.４ AND 演算（AND）

AND 演算子は、領域 a と領域 b が同時に出現する領域を返す。領域としては、a で始まり b で終わる領域又はその逆の領域を指す。主に包含演算子と組み合わせることで、”A,B の両方を含む C”のような記述が可能である。

２.２.５ OR 演算（OR）

OR 演算子は、領域 a または領域 b のいずれかの領域を返す。領域集合としては領域集合 A と領域集合 B の和集合となる。

２.２.６ 重なり演算（overlaps）

重なり演算子は、領域 a と領域 b が部分的に重なっている領域を返す。つまり<A>…<B>…</A>…</B>という関係である。領域集合としては重なっている領域、先の例での<B> … </A> の領域の集合となる。

---

[2] http://wiki.apache.org/solr/SolrPerformanceData

[3] http://www.hathitrust.org/technical_reports/Large-Scale-Search.pdf

２.３ 大規模日本語均衡コーパス BCCWJ

国立国語研究所で開発された BCCWJ(前川, 2009)は日本語において最大規模の均衡コーパスである。テキストは新聞・小説・ブログなどさまざまなジャンルのものからランダムに選択されたものが収録されている。

BCCWJ の一部データに対ついては、テキストに対しアノテーション情報が付与されている。どのアノテーションがどの部分に付与されているかはアノテーションによって異なり、重複する部分もあればしない部分もある。

BCCWJ の利用には国語研究所との間で契約が必要で、契約者にはデータを収録した DVD が配布されている。現在配布されている DVD には、テキスト情報に加え形態素レベルの情報が収録されている。形態素は短単位（SUW）と長単位（LUW）が別個に付与されている。これらの情報は、基本的に XML 形式で記述されている。本稿では BCCWJ の詳細については記述しない。

それ以外のアノテーションは、DVD には現在収録されていない。本研究では国語研が別途配布している係り受けアノテーション DepPara2[4]・奈良科学技術先端大学院大学を中心に開発された述語項構造(小町ほか, 2011)・慶応大学の開発した日本語フレーム構造(小原ほか., 2011)・岡山大学の開発した動詞項構造(竹内ほか, 2012)を用いた。

２.４ 国際標準 UIMA フレームワーク

UIMA (Unstructured Information Management Architecture)(Ferrucci et al., 2006)は様々な企業・研究機関で利用されている相互運用性のための枠組みである。UIMA はメタデータとそのための API を提供しており、実装は Apache　UIMA としてオープンソースで提供されている。最近では、テレビ番組のクイズ王に勝利した IBM の Watson システム(Ferrucci, 2012)でも用いられた。

UIMA の実行単位はコンポーネントと呼ばれ、コンポーネントを組み合わせて実行可能な UIMA ワークフローを作成する。おおむね、コンポーネントはいわゆるツールに相当し、ワークフローはアプリケーションに相当する。実行時のデータ構造は CAS と呼ばれる汎用構造に統一されている。XML のようなインライン形式と異なり、CAS ではスタンドオフ形式を前提としている。スタンドオフ形式ではテキストとアノテーションが分離され、アノテーションはテキスト内のオフセット位置を参照することでテキストに紐づけされる。CAS は概念的な構造であり、実行時は Java ヒープ内にオンメモリで格納されるが、XMI(XML Metadata Interchange)などファイル形式で保存することもできる。

UIMA は他に、データ型を階層的に定義する type system を記述する XML 形式やコンポーネントのメタデータ記述の XML 形式などを提供している。UIMA のアノテーションは type system により定義済みのデータ型で型付けされなければならない。UIMA 自体は整数や文字列といった基本的な型しか提供しないため、開発者が必要に応じてアノテーションの型階層を定義する必要がある。型階層は木構造で、型には素性と呼ぶ属性値を定義することができる。また、コンポーネントそのものも基本的に個々の開発者が作成して提供することになっており、様々な研究グループ・企業からコンポーネントが提供されている。

ウェブサービスの形式としては、ウェブサービスコンテナである Apache ActiveMQ 上で展開される UIMA-AS サービスが用意されている。

２.５ 統合自動自然言語処理システム Kachako

Kachako は UIMA 準拠のプラットフォームと互換ツールキット（UIMA コンポーネント群）からなる、統合全自動言語処理システムである(狩野, 2012)(Kano, 2012)。Kachako プラットフォームは、ツールの選択・組合せ・並列分散展開実行・視覚化・評価までを徹底的に自動化している。ツールキットでは統一 type system を定義した上で、その type system に互換な UIMA コンポーネント群を構築して、プラットフォームと互換ツール群をポータブルな形で配布し、ユーザの指定した任意の計算資源上でインストールから大規模処理ま

[4] https://sites.google.com/site/masayua/bccwjdep

で自動実行することを可能にしている。

Kachako の特徴の一つは、自動ワークフロー生成機能である。Kachako はコンポーネント群から可能なワークフローを自動計算するためにコンポーネントの入出力データ型情報を用いる。そのためデータ型の定義は自動計算が可能なように設計すると同時に、コンポーネントの入出力条件を過不足なく表現できるようにする必要がある。

我々は本研究に先行して、BCCWJ の各種アノテーションを Kachako および UIMA 互換となるよう変換するコンポーネントを実装した。本研究では入力を UIMA 形式とすることで汎用設計とし、入力データには BCCWJ データから変換した入力を用いた。

３．クエリの設計

テキストとアノテーションの統合検索を行う場合、大きく分けて三つの検索対象が考えられる。ひとつは、テキストそのものや、単一のアノテーションといったリテラル値である。二つ目は、文書内のアノテーション間の位置関係である。三つめは、アノテーション間の明示的な参照関係である。ありうる検索パターンをカバーした汎用的な検索のためには、これら三つの要素を複合的に組み合わせた検索クエリが必要である。

３.１ リテラルのクエリ

テキストのリテラル値は、二重引用符でくくった文字列で表記する。たとえば

```
"word"
```

のようになる。

アノテーションのリテラルについては、角括弧でくくった中に、一般的な XML/HTML タグと同様の表現でアノテーション名および属性名・属性値のペアを指定する。属性名と属性値は等号で接続し、属性値は二重引用符で囲う。アノテーション名と属性名は空白で区切る。たとえば

```
[annotation attr="value"]
```

のようになる。属性値を指定するかどうかは任意であり、指定しなくともよい。指定しない場合は、そのアノテーション名を持つ全てのアノテーションを表す。

３.２ 領域代数によるアノテーション間の位置関係を指定するクエリ

領域代数のクエリは、上記リテラルのクエリと領域代数の演算子を用いてポーランド記法で記述する。各演算は二項演算であり、表現としては全体が丸括弧でくくられ、最初に演算子、その後演算子のとる項が二つ並ぶ。記号と項は空白で区切る。演算子のとる項は、リテラル値または別のクエリの埋め込みである。つまり、ひとつのクエリは、複数の演算子やリテラルが複合的に組み合わさった表現であり、演算子が中間ノードでリテラルが葉ノードの二分木として表現可能である。

前述の領域代数の各演算に対する演算子として以下の記号を使用する。

包含:　`(> A B)`
被包含:　`(< A B)`
順序:　`(- A B)`
AND:　`(& A B)`
OR:　`(| A B)`
重なり:　`(@ A B)`

３.３ アノテーション間の参照を指定するクエリ

３.３.１ 変数

変数は、その変数の位置に任意の値が入ることを表現する。変数は'$'で始まる文字列で表記し、たとえば

```
[annotation attr="$1"]
```

は、annotation の attr 属性に任意の値が入ることを表現する。一クエリ中の複数箇所に同一名の変数が出現する場合は、それらの箇所には同一の値が入るとする。

３.３.２ リンクの表現

アノテーションの属性によっては、ほかのアノテーションを参照するリンクを持つことがある。このような場合は、上記の変数とアノテーション属性を用いてリンクを表現する。たとえば

```
(& [annotation attr="$1"] [annotation2 id=”$1”])
```

という表現は、annotation の attr 属性と annotation2 の id 属性が同じ値を持つことを表現する。すなわち annotation の attr 属性が annotation2 の id 属性を参照しているような領域を検索せよ、という意味になる。なお、このようなリンク表現を使用する際には、参照関係があらかじめアノテーションによって記述されている必要がある。

３.３.３ 領域を持たないアノテーションへのリンク

リンク先のアノテーションが領域を持たない（テキストの中の位置に紐づけられていない）ものに関しては、リンク先の属性名をさらにつづけて記述することで、領域を持たないアノテーションを埋め込む形で記述する。属性名はコロンで区切る。たとえば

```
[annotation attr1:attr2="val2"]
```

という場合、annotation の属性 attr1 の参照先のアノテーションの attr2 属性の値が val2 の場合、という意味になる。

４．検索システムの設計と実装

検索システムはすべて、Java 言語で実装した。外部ライブラリもすべて Java による実装のため、事実上ほとんどあらゆるサーバ、PC 上で稼働する。

４.１ リテラルの検索

リテラルの検索には、Solr を用いる。インデックス手法は標準的な転置インデックスまたは n-gram インデックスを想定する。転置インデックスはより高速な検索が必要な場合、n-gram インデックスは漏れのない検索が必要な場合に用いる。Solr のサポートするインデックス手法は多様であり、API に互換性があれば他の手法も利用可能である。

テキストのリテラル値インデックス作成は、一般のテキスト検索と変わりがない。

アノテーションのリテラル値インデックスは、非トークンを表す特殊なフィールドに保持し、インデックス自体はテキストのリテラル値と統合して検索するがトークン分割処理は行わない。アノテーションをインデックスする際は、属性値ごとにインデックスを作成する。たとえば

```
<annotation attr1="value" attr2="value2"> … </annotation>
```

というアノテーションに対しては

```
annotation:attr1="value1"
annotation:attr2="value2"
```

の二つがインデックスに追加される。これにより、インデックスの肥大を抑えつつ、インデックス検索のみで効率的な検索が可能になる。

４.２ 領域代数クエリの検索

領域代数クエリの検索は、クエリの構成する演算子二分木の各演算子ノードを一つ一つ処理することで行う。演算子の処理は、条件に合致するインデックスエントリを Solr 経由で取得することで行う。Solr の検索結果からはヒットしたアノテーションの begin/end 値が直接取得できるので、この値によるアノテーション間の位置関係についての条件の一致を行うことで高速に検索する。

この仕組みでは、演算子ごとに毎回 Solr のインデックス検索が実行され、演算子が複合的に埋め込まれている場合は取得した候補リスト中の各アノテーションについてさらにインデックス検索を実行するということが入れ子で繰り返される。

４.３ 変数の検索

変数の検索も、領域代数クエリの検索と同様であるが、まず変数に入りうる属性値を別途

検索して列挙する。そしてそれらの値を対応する変数に代入したクエリ表現を生成し、領域代数クエリの検索を行う。

５．おわりに

本研究では、BCCWJ コアデータおよびコアデータへのアノテーションを検索対象に想定して、汎用のテキスト・アノテーション統合検索システムを構築した。今後の課題としては、より大規模なデータでのスケーラビリティのテストが挙げられる。テキストのみの検索であれば Solr は十分なスケーラビリティを持つと考えらえるが、アノテーションの同時検索の場合データ量が大幅に増えるうえ検索時間が長くなるため、検証が必要である。

また、領域代数を基礎とするクエリは汎用ではあるが、文科系を含め専門外の研究者にとっては利用のハードルが高いことが想定される。より容易にクエリを入力できるようなクエリ生成支援機能や GUI の実装、検索結果表示方法の改善を検.討していきたい。

謝　辞

本研究の一部は、国立国語研究所共同研究プロジェクト「コーパスアノテーションの基礎研究」（プロジェクトリーダー：前川喜久雄）および科学技術振興機構さきがけ「情報環境と人」領域「解析過程と応用を重視した再利用が容易な言語処理の実現」（研究代表者：狩野芳伸）による補助を得ています。

文　献

C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. An algebra for structured text search and a framework for its implementation. The computer Journal, 38(1):43–56, 1995

J. Jaakkola and P. Kilpel¨ainen. Nested Text-Region algebra. Technical Report C-1999-2, University of Helsinki, 1999.

Katsuya Masuda and Jun'ichi Tsujii. Tag-Annotated Text Search Using Extended Region Algebra. The IEICE Transactions on Information and Systems, Special Section on "Natural Language Processing and its Applications," Vol.E92-D,No.12, pp.2369-2377, 2009.

前川喜久雄. (2009). 代表性を有する大規模日本語書き言葉コーパスの構築. 人工知能学会誌, *24*(5), pp.616–622.

小原京子, 加藤淳也, 斎藤博昭. (2011). 日本語フレームネットにおける BCCWJ への意味アノテーション. *日本語コーパス平成 22 年度公開ワークショップ* (pp. 513–518).

小町守, 飯田龍. (2011). BCCWJ に対する述語項構造と照応関係のアノテーション. *日本語コーパス平成 22 年度公開ワークショップ* (pp. 325–330).

竹内孔一, 竹内奈央, & 石原靖弘. (2012). 述語の分析に基づく文書解析の考察. *IPSJ SIG Notes* (Vol. 2012, pp. 1–7). Information Processing Society of Japan (IPSJ).

Ferrucci, D. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, *56*(3.4), 1:1–1:15. doi:10.1147/JRD.2012.2184356

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., et al. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report.

狩野芳伸. (2012). Kachako: 誰でも使える全自動自然言語処理プラットホーム. *2012 年度人工知能学会全国大会（第 26 回）*.

Yoshinobu Kano. (2012) Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation. *In the 1st International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012)*. Shanghai, China, November 12nd 2012.

ポスター発表　グループA

9月10日（水）　10:00～11:00

漢語サ変動詞の卓立性の再考
―動詞形・構文形比率を手掛かりとして―

李　楓（神戸大学大学院国際文化学研究科）†

Reconsideration of Lexical Status of Sino-Japanese-Verbs
- From the Perspectives of Verb Form and Sentence Structure -

Feng Li（Graduate School of Intercultural Studies, Kobe University）

要旨

漢語サ変動詞の中には、ほぼ「～する」の動詞形でのみ使うものもあれば、「～をする」の構文形と併用するものもある。同じ漢語サ変動詞であっても、これらは語としての位置づけに、質的な違いが存在すると考えられる。従来の漢語サ変動詞研究において、この点は未だ十分に解明されていないが、個々の漢語サ変動詞が構文形を持つか否かを考えることは、日本語記述の精緻化を図るうえでも、また日本語学習者に適切な情報を提供するうえでも重要であると考えられる。本研究では、李（2013）で特定した高頻度・汎用的漢語サ変動詞 93 語をサンプルとして、それらの語の動詞形、及び構文形使用状況をコーパスを用いて調査した。その結果、漢語サ変動詞は、構文形を志向するものと動詞形を志向するものに峻別されることが明らかになった。その後、重回帰分析を実行することにより、構文形併用の程度を説明するモデルの検討を試みた。

１．はじめに

いわゆる漢語サ変動詞は「～する」という動詞形を取るが、語によって、漢語部と「～する」の間に「を」を加えた「～をする」という構文形が併用される場合もあれば、構文形がほとんど使用されない場合もある。たとえば、「勉強する」についていえば、「勉強をする」という構文形も一般的である。一方、「反映する」についていうと、その構文形「?反映をする」は不自然に感じられる。このように、漢語サ変動詞は、対応する構文形を持つか持たないか、また持つとすれば、どの程度の割合で構文形が使用されるかという点において、一定の幅を持っていると言えるだろう。

ある漢語サ変動詞が対応する構文形を持ち、また構文形が相当程度用いられる場合、当該の意味は 2 種類の表現形で伝達されることとなり、意味表出手段としての漢語サ変動詞の唯一無二性、つまり卓立性は一定の制約を受けていると言える。これに対し、漢語サ変動詞が対応する構文形を持たない場合、当該の意味の伝達は漢語サ変動詞だけが唯一無二的に担っていることとなり、漢語サ変動詞はより卓立的な位置づけを持つと考えられる。先行研究もこうした現象に関心を寄せていたが、その多くは、漢語サ変動詞（動詞形）をデフォルトと見なしながら、それが構文形を取り得る際の言語的条件の解明を目指すものであった。一方、コーパスを使用し、個々の漢語サ変動詞の構文形併用率、つまりは本稿で言う漢語サ変動詞の卓立性を計量的に調査した研究は必ずしも多くない。漢語サ変動詞の卓立性の計量的調査は、言語記述の精緻化を図るうえで言語学的な意義を持つのみならず、日本語教育の観点からも重要である。

一般に、教育の現場では、ある文法形式を指導する場合に、典型的な用例を提示するこ

† lifengwang2006team@yahoo.co.jp

とが重要であると考えられている。典型性は一般にコーパス頻度で読み替えられ、様々なテキストで満遍なく高頻度に出現する語が典型例と見なされがちである。事実、李（2013）においても、そうした観点から、コーパスに含まれる多様な言語種別（ジャンル）において等しく頻出する漢語サ変動詞の特定を行った。しかしながら、仮にそうして選ばれた語が、高い構文形併用性を持ち、動詞形自体の意味表出手段としての卓立性・単独性が低いとすると、学習者に提示すべき漢語サ変動詞用例としての典型性は弱まることとなろう。つまりは、様々な日本語において、高頻度かつ汎用的に出現するだけでなく、同時に、対応する構文形を持たない（あるいは持つとしても、そうした頻度がきわめて低い）ものがより妥当な典型用例になると考えられる。

以上の点をふまえ、本研究では、すでに特定した高頻度・汎用的漢語サ変動詞をサンプルとして、その卓立性の再考を試みる。このことは、あわせて漢語サ変動詞の卓立性がどのようなテクスト内的・テクスト外的要件によって影響されているかの解明にもつながる。

２．先行研究

漢語サ変動詞については多くの研究がなされており、「～する」と「～をする」の交替の問題を論じた研究もいくつか見られる。ここでは、主要なものに限って、先行研究を概観したい。当該問題を扱った研究は、主として理論的立場に基づくものと、主として計量的立場に基づくものとに二分される。はじめに、理論的立場に立脚するものとして、田野村（1988）、影山（1993）、Uchida & Nakayama（1993）、平尾（1995）、松岡（2004）を概説する。なお、以下では、「～する」の形を動詞形、「～をする」の形を構文形と呼ぶ。

田野村（1988）は、動詞形（複合サ変動詞と称されている）と構文形（単純サ変動詞と称されている）の両方の形が表現としては成り立つが、日本語としての自然さに違いが存在する場合を検討している。氏によれば、構文形が成立するためには、（1）当該語が「する」（＝「行う」）と言うに足る動作・行為を表していること、（2）動詞概念部それ自体が「行う」ことを表していないこと（例、a.計画を実行する。b.？計画の実行をする。）、（3）動作対象に対する別の動作・行為の表現が後続する同一の文脈に存在していないこと（意味的勢力）（例、a.？シャツの洗濯をして干した。b.シャツの洗濯をして外出した。）、の 3 つの条件が必要であるという。さらに（1）については、細かく（1-1）意図的な事柄を表していること（意図性）（例、a. 警察は毒物を検出した。b.？警察は毒物の検出をした。）、（1-2）し始め、し終わるような性質の事柄を表していること（アスペクト性）（例、a.九時に受付を開始する。b.？九時に受付の開始をする。）、（1-3）純粋な心理的事柄を表していないこと（物理性）（例、a.優勝の祈願をする。b.？優勝の期待をする。）の 3 点が導かれるとされている。逆に、動詞形が成立する場合、動詞概念部が動作・行為を表すことは必須でないとされている。

影山（1993）は、構文形が成立するには、漢語名詞部分が意図的に動作を行う動作主を主語に取るような非能格名詞（例、a.家族をそろって食事をする。b.離婚をする。）であることが要求されるとしている。意図を持たず、受動的に事象に係る対象を主語にとる非対格性を持つ名詞（例、a.*持っていたピストルが暴発をした。b.*老人が会談で転倒をした。）の場合、構文形は成立せず、動詞形が使用されると論じている。

Uchida & Nakayama（1993）では、当該動詞が行為・達成動詞である場合に、構文形が成立するという。行為・達成動詞とは、漢語名詞に「～ている」が結びついた場合、動作の継続が含意されるような動詞であり、「研究する」などがその例である。一方、当該動詞

が到達・状態動詞である場合には、構文形は成り立たず、動詞形が使用されるという。到達・状態動詞とは、漢語名詞に「～ている」が加わる場合、結果の状態を含意するような動詞であり、「開始する」などがその例である。

平尾（1995）は、構文形が成立するには、「～しよう」「～したい」などと共起できる程度の意志性を持った漢語名詞が必要であるとしている（例、a.勉強しよう。→勉強をする。b.選択しよう。→選択をする。）。

松岡（2004）は、従来の研究で提唱されていた意志性や時間幅などの概念では説明がつかない場合があることを指摘しながら、動詞形と構文形の交替条件を改めて検討している。氏は構文形が成立するには、「事象への関与が高い」ことが必要であるとし、以下の 2 つの場合を説明している。一つ目は、主語が意志的に行い、その事象に持続性がある場合である（「責任をもって～をする」という表現と共起しうる。例、a.勉強をした。b.練習をする。）。二つ目は、主語が意志性や主体性を持たないが、事象の結果が主語に返ってくる（「再帰性」を持つ）ような場合である（「自分が～したのは自分の…からだ」という表現と共起しうる。例、a.怪我をした。b.失敗をした。）。

以下に、これらの主要先行研究の内容を一覧にまとめる。

表 1　主要先行研究にみる構文形の成立条件

田野村（1988）	（1）動作・行為性（a.意図性、b.アスペクト性、c.物理性）、（2）主語非動作性、（3）文脈統一性
影山（1993）	名詞非能格性（意図性・能動性を持つ動作主を主語にとる）
Uchida & Nakayama（1993）	行為・達成動詞＝動作の継続性
平尾（1995）	意志性（「しよう」、「したい」と共起しうる）
松岡（2004）	事象への強関与性（意志性は問わず）

以上の理論的研究に加え、若干ではあるが、量的研究も存在する。その一例は、田辺・中條・船戸（2012）である。同研究は、中條・木下・田辺・内山・西垣（2010）で選定された漢語動名詞上位語について、コロケーションを調査した。その結果、構文形（選挙をする）や、他の漢語名詞との結合形（選挙活動）で使用される名詞用法中心語と、主として動詞形（実施する）で使用され、構文形（? 実施をする）や漢語名詞結合形（? 実施活動）などでほとんど使用されない動詞用法中心語の 2 種類が存在することが明らかにされた。この研究は、漢語の用法分析にコーパスを使用した興味深い取り組みであるが、本研究が問題とする漢語サ変動詞（動詞形）と構文形の差異に違いを絞ったものではない。

このように、漢語サ変動詞（動詞形）と構文形の交替する条件について、先行研究では動作性・意志性など、非常に有用な知見が得られた。しかしながら、量的アプローチに基づく研究の余地はまだ多く残されていると言える。

３．リサーチデザイン

３.１　研究目的

すでに述べたように、漢語サ変動詞の卓立性を計量的に明らかにしようとした研究はきわめて限られている。ゆえに、本研究では、高頻度で汎用的な漢語サ変動詞をサンプルとして、重要な漢語サ変動詞の卓立性を検討する。これにより、学習者に提示すべきより典

型的な漢語サ変動詞を抽出し、あわせて漢語サ変動詞の卓立性がどのようなテクスト内的・テクスト外的要因と関係しているかを検討することを目的とする。

３.２　対象

漢語サ変動詞には、様々な語が含まれるが、漢語サ変動詞全体の大まかな特徴を捉えようとする場合、それらの中で、特に高頻度で汎用的な語をサンプルとして分析を進めることが妥当である。そこで、本研究では、李（2013）において特定された 93 語の漢語サ変動詞をサンプルとする。これらは「現代日本語書き言葉均衡コーパス（BCCWJ）」を構成する書籍、雑誌、新聞、ブログ、白書、知恵袋の 6 種を含む、計 15 変種（書籍は内容別 10 ジャンルをそれぞれ別個のデータとして扱う）から得られた出現頻度値を主成分分析によって合成した統計指標値を手掛かりに選ばれたものである。以下では、抽出された 93 語を提示する。抽出過程の詳細は李（2013）を参照されたい。

表 2　高頻度・汎用的漢語サ変動詞リスト

	語		語		語		語		語		語
1	存在する	17	注意する	33	要求する	49	発売する	65	実行する	81	分類する
2	利用する	18	開催する	34	完成する	50	意識する	66	変更する	82	安定する
3	紹介する	19	構成する	35	判断する	51	移動する	67	拡大する	83	観察する
4	説明する	20	評価する	36	成功する	52	指定する	68	解決する	84	保存する
5	使用する	21	設定する	37	設置する	53	強調する	69	解放する	85	心配する
6	確認する	22	変化する	38	決定する	54	発展する	70	比較する	86	重視する
7	理解する	23	提供する	39	作成する	55	増加する	71	集中する	87	考慮する
8	表示する	24	形成する	40	選択する	56	開発する	72	購入する	88	販売する
9	発見する	25	実現する	41	開始する	57	否定する	73	発揮する	89	想像する
10	発生する	26	表現する	42	採用する	58	提出する	74	一致する	90	活用する
11	注目する	27	展開する	43	登場する	59	反映する	75	無視する	91	限定する
12	指摘する	28	参加する	44	認識する	60	規定する	76	維持する	92	減少する
13	期待する	29	検討する	45	予想する	61	代表する	77	適用する	93	区別する
14	実施する	30	意味する	46	成立する	62	掲載する	78	確立する		
15	用意する	31	対応する	47	報告する	63	支配する	79	記載する		
16	発表する	32	主張する	48	導入する	64	結婚する	80	勉強する		

３.３　リサーチクエスチョン

すでに述べたように、先行研究からは多くの知見が得られたものの、未だ解明されていない点もある。たとえば、様々な漢語サ変動詞のうち、動詞形に加えて、構文形併用性の高い語と低い語にはどのようなものがあるか、言い換えれば、漢語サ変動詞の卓立性が低い語と高い語にはどのようなものがあるか、また、漢語サ変動詞の卓立性の高い語と低い語にはどのような共通特徴があるか、漢語サ変動詞の卓立性は、漢語部の頻度・意味数・語構成、漢語サ変動詞のアスペクト性・自他性、漢語サ変動詞の使用される言語変種（ジャンル）、といった様々な特性とどのように関係しているかは十分に解明されていない。

以上をふまえ、本研究で明らかにしようとするリサーチクエスチョンとして、以下の 3 点を設定した。

RQ1：高頻度・汎用的漢語サ変動詞はそれぞれどのような卓立性を持つか？

RQ2：漢語サ変動詞卓立性の高い語と低い語はどのような共通特性を持つか？
RQ3：漢語サ変動詞の卓立性を説明する統計モデルはどのようなものであるか？

３.４　研究手順

本節では、RQ ごとに研究の手順を簡潔に紹介する。

まず、RQ1（漢語サ変動詞の卓立性）については、すでに述べた BCCWJ の 15 変種を対象として、動詞形および構文形の頻度を調査する（頻度調査は変種ごとに行うのではなく、全変種を統合したデータに対して行う）。その後、動詞形と構文形の頻度の和に占める動詞形頻度の比率（動詞形使用率）を求め、これを以下の分析における卓立性指標値と見なす。なお、本調査にあたっては、漢語部に他の漢語が結合しているもの（例、自己紹介する）や、「を」と「する」の間に他の要素が付加されているもの（例、確認を繰り返しする）は対象から除外する。その後、各語の卓立性指標値を手掛かりにして、全体を降順に並べ替える。

次に、RQ2（卓立性の高いものと低いものの特性）については、卓立性指標に基づき、値が高いものおよび低いものについてそれぞれ BCCWJ からランダムに用例を抽出し、特に先行研究で示されてきた構文形の許容基準を参考にして質的な考察を行う。

また、RQ3（漢語サ変動詞卓立性の説明モデル）については、複数の説明変数で単一の目的変数を説明・予測する重回帰分析を行う。目的変数は 93 語それぞれの卓立性指標値（動詞形使用率）である。この値に影響を及ぼし得る説明変数の候補としては様々なものが考えられるが、ここでは、漢語サ変動詞の核となる漢語部に関わるパラメタと、語幹部に「する」が付加された漢語サ変動詞に関わるパラメタ、の 2 つを想定し、それぞれ以下のような説明変数を設定する。

表 3　漢語サ変動詞の卓立性に関与する要因候補

漢語部に関わる特性	漢語サ変動詞に関わる特性	
頻度・意味・語構成的特性	文法的特性	環境的特性
(1)　頻度 (2)　意味数 (3)　語構成パタン	(4) アスペクト性 (5) 自他性	(6) 非公式ジャンル出現性 (7) メディアジャンル出現性
テクスト内的		テクスト外的

上記の各要素のうち、(1) (2) (6) (7) はそれぞれ一つの説明変数しか持たないが、(3) は 5 種、(4) は 2 種、(5) は 3 種の説明変数を持つ。これらを合わせると、全体で説明変数の数は 14 種となる。以下、それぞれについて詳しく述べる。

(1) 頻度については、個々の漢語サ変動詞の漢語部の頻度データを使用する。その際、粗頻度そのものを使用すると、圧倒的に値が高くなり、重回帰モデル内で過剰な重みづけが与えられる可能性があることから、ここでは、粗頻度を自然対数に変換して使う。自然対数化により、大きな値が効率よく圧縮され、ほかの説明変数候補と比較的近い範囲に値が収束することとなる。(2) 意味数については、漢語部について、辞書の当該項目内で言及されている意味の総数を調査したデータを用いる。辞書によって意味の数は異なる場合が多いが、本研究では、一般に広く使用されている辞書の一例として、『大辞林 第三版』（三省堂、2006）を参考資料とする。たとえば、「紹介」の場合、同辞書は、①知らない人どうしを引き合わせること、なかだちをすること（「家庭教師を―する」「アルバイトの―」「自己―」）と、②未知の物事を広く知らせること（「日本文化の―」）との 2 つの意味を記載しているため、意味数は 2 となる。(3) 語構成パタンについては、漢語部を構成する 2 つの漢字間の意味的結合関係を 5 種類に区分したデータを使用する。区分は李（2013）に基づ

くもので、並立関係（選択など）、修飾関係（活用など）、客体関係（読書など）、補充関係（拡大など）、実質関係（否定など）の 5 種類である。(4) アスペクト性については、金田一（1950）に基づく継続動詞と瞬間動詞の 2 分類を使用する。金田一（1950）は日本語の動詞をアスペクトの観点から、状態動詞、継続動詞、瞬間動詞、第四種の動詞の 4 種類に分類しているが、漢語サ変動詞には、状態動詞と第四種の動詞の分類は該当しないため、継続動詞と瞬間動詞の 2 種類の可能性がある。継続動詞は動作・作用を表し、その動作の継続があるものとされ、瞬間動詞は動作作用を表すが、その動作が一瞬にして終わるものであるとされている。(5) 自他性については、李（2014）の調査に基づき、自動詞、他動詞、自他両用動詞の 3 分類を用いる。同研究は先行研究をふまえ、「～を」格を持ち、かつ、対象に対する「働きかけ」が認められるものが他動詞、それ以外は自動詞としている。(6) 非公式ジャンル出現性については、当該語がコーパス内の 15 種のジャンル中、非公式性の強い 2 ジャンル（ブログと知恵袋）における出現率データを使用する。ブログと知恵袋は、インターネット上の自由な発言の場であり、校閲のプロセスを経ていないため、言語としての公式性が低いと考えられる。(7) メディアジャンル出現性については、当該語がコーパス内の 15 ジャンル中、メディア性の強い 2 ジャンル（雑誌と新聞）における出現率データを用いる。雑誌や新聞は、圧倒的に大多数の読者を対象として、正確かつ適切な表現となるよう、慎重な校閲を行ったテキストと言える。単に非公式ジャンルに対する公式ジャンルと考えれば、ほかに政府刊行の白書類なども候補となるが、ここでは、読者層が圧倒的幅広い雑誌と新聞に限り、前述の非公式ジャンルの対称ジャンルとした。

なお、重回帰分析において、説明変数をモデルに投入する場合、「全変数導入法」と「変数増減法」の 2 つの方法が存在する。前者は手元の説明変数をすべて使って回帰モデルを作りたい場合に、後者は多数の説明変数候補の中から最適な変数の数を決めたい場合に使用される手法である（石川・前田・山崎、2010、p. 116）。ここでは、両者をともに実行し、結果を解釈することで、候補としたすべての変数の位置づけを確認すると同時に、より最適なモデルを求める。以上の処理には Seagull-Stat 2010 版を使用する。

４．結果と考察

４.１　漢語サ変動詞の卓立性

漢語サ変動詞 93 語を卓立性に基づいて並べ替えた結果、表 4 が得られた。また、漢語サ変動詞の卓立性指標値を「～90%」、「90%～92%」、「92%～94%」、「94%～96%」、「96%～98%」、「98%～」の 6 段階に区分し、それぞれの指標値レンジに含まれる語の数の比率を調べた結果、図 1 が得られた。

表 4　卓立性順漢語サ変動詞リスト（漢語部のみ）

語	比率	語	比率	語	比率	語	比率	語	比率
反映	100.00	限定	99.88	形成	99.69	検討	98.91	選択	96.22
発揮	100.00	安定	99.87	支配	99.68	変化	98.88	対応	96.09
一致	100.00	導入	99.87	作成	99.68	指摘	98.86	注意	96.07
確立	100.00	考慮	99.86	否定	99.66	無視	98.86	決定	95.75
重視	100.00	適用	99.84	構成	99.66	保存	98.85	表現	95.54
増加	99.98	減少	99.84	意識	99.63	移動	98.80	用意	95.43
存在	99.98	維持	99.83	発売	99.62	紹介	98.58	判断	95.41
完成	99.96	成功	99.80	購入	99.52	期待	98.57	心配	95.21
集中	99.96	拡大	99.79	参加	99.45	指定	98.53	設定	94.98
成立	99.95	掲載	99.77	認識	99.36	区別	98.26	説明	94.83
開始	99.95	活用	99.77	発展	99.29	比較	98.24	報告	94.42

開催	99.94		注目	99.77		分類	99.28		想像	98.19		理解	94.30
実施	99.94		代表	99.76		提供	99.24		観察	97.83		評価	93.32
登場	99.93		使用	99.76		記載	99.23		開発	97.77		発見	91.38
強調	99.93		提出	99.74		解決	99.12		主張	97.52		勉強	83.73
意味	99.93		実行	99.73		発表	99.11		要求	97.42		確認	79.54
実現	99.91		解放	99.72		予想	99.07		販売	97.16		表示	77.61
利用	99.90		規定	99.70		展開	99.04		結婚	96.98			
設置	99.88		採用	99.70		発生	98.94		変更	96.78			

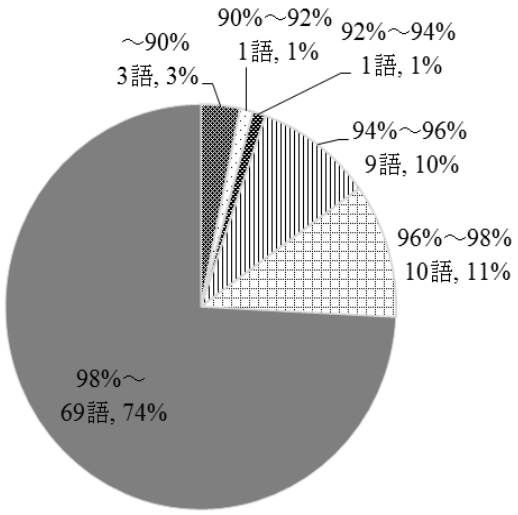


図１　卓立性指標レンジごとの漢語サ変動詞数比率

ここでは、上記よりわかることとして 2 点を示す。1 点目は、サンプルとした 93 語のうち、90 語において卓立性が 90%以上となったことである。これは、統計的手法を用いて特定された高頻度・汎用的漢語サ変動詞の大部分が動詞形・構文形の併用という観点から見ても、高い典型性を持っていることを裏付ける結果と言える。

2 点目は、ごく一部において、漢語サ変動詞の卓立性が低いものが含まれていたということである。該当するのは、指標値が 90%未満となった「勉強」「確認」「表示」の 3 語である。これらの 3 語は各種の漢語サ変動詞の中で高頻度かつ汎用的な語ではあるが、当該の意味を表出しようとする場合、動詞形に加えて、構文形も一般的に併用されることが示されている。漢語サ変動詞の「典型用例」を狭義で捉えた場合、これらの 3 語をそこに含めるかどうかについては、今後再検討の余地がある。

いずれにせよ、こういった点は従来の日本語研究においても、日本語教育においてもほとんど注目されてこなかった。漢語サ変動詞の卓立性指標に基づく本調査は、従来の研究・教育に新たな視点を示したものであると言える。

４.２　卓立性の高いものと低いものの特性

前節の分析結果をふまえ、卓立性の高い漢語サ変動詞として、指標値が 100%となった「反映」「発揮」「一致」「確立」「重視」の 5 語、及び卓立性が低いものの例として、指標値が 90%未満となった「勉強」「確認」「表示」の 3 語に注目し、それぞれの用例を質的に検討する。まず、卓立性が高い 5 語について用例を見てみよう。これらは現代日本語書き言葉均衡コーパスの 15 変種において、構文形が一例も確認されなかったものである。

（1）密着したものとなるに従い、道路交通をめぐる国民の意識、要求は、それぞれの立場を反映し、多様化してきている。（白書 OW1X_00441）
（2）ベレッタにとりつけられた新品のサイレンサーは、見事に効力を発揮した。（書籍・文学 LBl9_00257_6370）
（3）振込人名とイーバンクロ座名が一致しない場合、対象となりませんのでご注意ください。（知恵袋 OC14_02242_700）
（4）近習の地位を得たことは、生活の基盤が確立したという芽出たいことなのであるから、杯酒に沈湎して、自分は才能が発揮できないなど・・・（書籍・文学 LBg9_00012_16230）
（5）おれが仕事を重視するのが、どうしてもこの人には納得できないことだったのかもしれない、と。（書籍・文学 PB39_00045_56460）

上記の用例を用い、構文形が許容されない理由を田野村（1988）で示された構文形成立条件に基づいて考察する。前述のように、田野村（1988）によれば、構文形の成立条件としては、（A）意図性、（B）アスペクト性、（C）物理性、（D）概念部非動作性、（E）意味的勢力、の 5 点が必要であるとされている。ただし、（D）と（E）の 2 点は語に関係する条件ではなく、むしろ語が出現する文型・文脈という外的な条件であると考えられる。よって、ここでは、主として語と直接に関わる条件としての（A）（B）（C）の 3 点に注目して分析を行う。上記の用例を見てみると、（1）の「反映する」は、主語は「意識・要求」であって、まず（A）の意図性という条件は満たさない。また、「?反映し始める」や「?反映し終わる」のように、（B）アスペクト性も表しにくい。（2）の「発揮する」は無生物の「サイレンサー」が主語であるため、（A）の意図性を表していない。（3）の「一致する」と（4）の「確立する」は、単に事象の様態や状態を表しているため、（A）の意図性も（C）の物理性も備わっていない。（5）の「重視する」は、動作・行為というよりも、単に心理的活動を表しているため、（C）の物理性は認められない。これらの点をふまえると、以上の 5 語の卓立性が 100%となり、構文形が許容されないことは説明がつく。
次に、卓立性が相対的に低い 3 語について用例を見てみよう。これらは当然ながら動詞形と構文形の両方を持つが、以下に示すのは構文形の用例である。

（1）いっしょに、近づきつつある秋の中間試験のための勉強をしていたので、すっかりおそくなってしまったのである。（書籍・文学 PB29_00130_9540）
（2）また、一端選択受信を設定し、件名を読み込みに行って、メールの確認をする方法もあります。※選択受信については、下記 URL をどうぞ。（知恵袋 OC02_09229_1970）
（3）・・・また、ドアがきちんと閉まっていない場合は加熱しないでくださいという表示をしてください。（書籍・社会科学 LBm3_00019_11270）

上記について、同じく田野村（1988）の枠組みに基づいて構文形が許容される理由について確認しておきたい。まず、（1）の「勉強」について言うと、主語は人であり、「中間試験のため」という行為には明確な（A）意図性が備わっており、かつ「勉強」という行為には「勉強していた」のように、明白な（B）アスペクト性の事柄を表しており、単純な心理的活動ではないため、（C）物理性も持つと考えられる。（2）の「確認」は何らかの目的を持ち、メールをチェックする行為を表していると想像がつくため、（A）の意図性も（C）の物理性も認められる。また、この動作・行為は一定の期間をわたってのことであるため、（B）のアスペクト性も備わっている。（3）の「表示」についても、前後の文脈からも分かるように、何らかの目的を持ち、かつ動作・行為は一定の時間幅を持つものであるため、（A）意図性、（B）アスペクト性、（C）物理性の 3 つの特性をすべて備わっていると考えられる。このように、これらの例は、いずれも先行研究で提唱された構文形の成立条件にあてはまり、先行研究の議論を支持したものであると考えられる。

以上より、卓立性が高いものと低いものがそれぞれ一定の共通特徴を持ち、その特徴の大部分が先行研究で示された枠組みで説明できることが確認された。

４.３　漢語サ変動詞卓立性の説明モデル

まず 14 種の説明変数を用いて、全変数導入法で重回帰分析を行ったところ、下記の回帰式が得られた。

卓立性指標＝－941.1－1.071〔頻度〕－0.611〔意味数〕－477.9〔並立関係〕－477.0〔修飾関係〕－477.5〔客体関係〕－477.5〔補充関係〕－474.8〔実質関係〕＋0.309〔継続動詞〕＋0.111〔瞬間動詞〕＋1529.〔自動詞〕＋1526.〔他動詞〕＋1527.〔自他両用〕－0.0977〔非公式〕＋0.336〔メディア〕

上記のモデルは有意（p <.05）であり、決定係数は.23、自由度調整済相関係数は.31 となった。このことから、モデルは一定の有効性を示しているように思われるが、VIF=<10.0 となり、変数間の多重共線性が認められた。このモデルは妥当性に制約のあるものではあるが、それぞれの説明変数に付与された係数に注目すると、漢語部の頻度が低く、意味の数が少なく、並立関係・修飾関係・客体関係・補充関係・実質関係といった明確の構造を持たず、継続動詞ないしは瞬間動詞であり、かつ自動詞・他動詞・自他両用動詞のいずれかの性質を持ち、かつ非公式なジャンルであまり使用されず、むしろ、多くの読者に向けて校閲を重ねたメディアジャンルにおいて多く出現する場合、漢語サ変動詞の卓立性が上昇するという全般的な傾向が示された。この結果から、本研究で設定した説明変数の候補のうち、漢語サ変動詞の卓立性は、漢語部の語構成パタン（並立・修飾・客体・補充・実質）や、漢語サ変動詞のアスペクト性（継続動詞・瞬間動詞）や自他性（自動詞・他動詞・自他両用動詞）からははっきりした影響を受けていない可能性があり、むしろ、漢語部の頻度や意味数、あるいは漢語サ変動詞の使用される言語環境とより密接に関わることが示唆された。

ただ、すでに述べたように、全変数導入法による説明モデルは変数間の多重共線性を含み、必ずしも妥当なものではない。そこで、変数増減法により変数の取捨・整理をしたところ、以下のモデルが得られた。

卓立性指標＝108.9－1.198〔頻度〕－0.521〔意味数〕＋1.968〔自動詞〕－0.0937〔非公式〕－0.333〔メディア〕

上記のモデルは有意（p <.05）であり、決定係数は.22、自由度調整済相関係数は.42 となった。相関係数に注目すると、全変数導入法に比べ、より妥当性の高いモデルになっていることが確認される。また、VIF 値はいずれも 10.00 を超えていないため、多重共線性の問題もなく、全体としてより妥当で有効なモデルであると考えられる。

モデル中の変数に付与された係数に注目すると、漢語部の頻度が低く、意味数が少なく、漢語サ変動詞の自動詞性が強く、ブログや知恵袋のような非公式ジャンルではなく、雑誌や新聞のようなメディアジャンルで多く使用される場合に、漢語サ変動詞の卓立性が高まるという傾向が示された。大きな枠組みはすでに全変数導入法で見た場合と変わらないが、新たに自動詞性が漢語サ変動詞の卓立性に関係する指標として加わったことが重要であろう。

もちろん、この結果は 93 語という限られたサンプル語についてのみあてはまるもので、漢語サ変動詞全体の卓立性の説明モデルとしては決して十分なものではない。しかしながら、漢語サ変動詞の卓立性、あるいは構文形併用性を多様な観点から検討したことは、従来の漢語サ変動詞に関わる研究や日本語教育に新たな視点を加えたものとなりうるであろう。

５．まとめ

本研究では、李（2013）で特定された漢語サ変動詞 93 語をサンプルに、漢語サ変動詞（動詞形）と構文形の使用状況をあわせて調査することにより、各語の漢語サ変動詞の卓立性を確認した。以下、リサーチクエスチョン順に結論をまとめる。

まず RQ1 では、サンプルとした 93 語のうち、90 語において卓立性が 90%を上回り、高頻度汎用的漢語サ変動詞は、動詞形・構文形の選択という観点から見ても高い卓立性を持つことが確認された。一方で、「勉強（する）」「確認（する）」「表示（する）」のように、相対的に卓立性の低い語が存在することも示された。

次に RQ2 では、漢語サ変動詞の卓立性の高い語と低い語をそれぞれ BCCWJ から用例を抽出し、田野村（1988）で提唱された構文形成立条件（意図性、アスペクト性、物理性、概念部非動作性、意味的勢力）について確認を行った。

最後に RQ3 では、重回帰分析により、漢語部の頻度が低く、意味数が少なく、漢語サ変動詞が自動詞的に使用され、インターネット上の非公式ジャンルであまり使用されず、雑誌や新聞のようなメディアで多く使用される場合に、漢語サ変動詞の卓立性が高まることを確認できた。また、漢語部の頻度や漢語サ変動詞の使用環境などに比べると、漢語部の語構成パタンなどの影響は限定的であることも確認された。

以上のように、本研究は、漢語サ変動詞の動詞形・構文形使用に注目し、従来の研究では十分に検討されてこなかった漢語サ変動詞の卓立性について調査し、有益な知見を得た。この点をふまえれば、本研究は従来の漢語サ変動詞に関わる研究や日本語教育に新たな視点を加えたものとなりうる。ただし、得られた知見は、高頻度・汎用的漢語サ変動詞に限ったもので、漢語サ変動詞全体の特性として捉えうるかどうかを今後の課題としたい。

参考文献

石川慎一郎、前田忠彦、山崎誠（編著）(2010)『言語研究のための統計入門』、くろしお出版

Uchida. Y. and Nakayama. M. (1993) Japanese verbal noun constructions. *Linguistics.31*(4)、pp. 623-666

影山太郎(1993)『文法と語形成』、ひつじ書房

金田一春彦(1950)「国語動詞の一分類」『言語研究』（日本言語学会）15、pp.48-63、

田辺和子、中條清美、船戸はるな(2012)「新聞コーパスにおける二字漢語動名詞の動詞的・名詞的ふるまいについて」『日本女子大学紀要』61、pp.19-32

田野村忠温(1988)「『部屋を掃除する』と『部屋の掃除をする』」『日本語学』7(11)、pp.70-80、明治書院

平尾得子(1995)「VN がスルと VN スルと VN ヲスル―サ変動詞語幹と構文的制約―」宮島達夫、仁田義雄（編）『日本語類義表現の文法（上）単文編』、pp.89-98、くろしお出版

松岡知津子(2004)「漢語名詞とスルが構成する 2 種類の述語の交替」『広島大学大学院教育学研究科紀要』2(53)、pp.305-310

松村明(編)(2006)『大辞林　第三版』、三省堂

李楓(2013)「現代日本語における汎用的漢語サ変動詞の抽出とその内部構成の検討」『国立国語研究所第 4 回コーパス日本語学ワークショップ予稿集』、pp.101-110

李楓(2014)「高頻度・汎用的漢語サ変動詞の諸相」語彙研究会第 101 回例会発表資料、愛知学院大学

均衡性と代表性に配慮した『太陽コーパス』の分析法試論

森　秀明（東北大学大学院文学研究科）†

Methodological Consideration on Corpus-Balance and Representativeness of "Taiyo Corpus"

Hideaki Mori (Graduate School of Arts and Letters，Tohoku University)

要旨

『太陽コーパス』は、明治後期～大正期の総合雑誌『太陽』から５年分を抽出した全文コーパスである。近代日本語の成立期に当たるデータが集積されているため、語や文法の経年変化分析に使用されることが多い。ある形式を出版年ごとに比較する場合、それぞれのデータがある程度均質でないと正確な分析はできない。しかし、『太陽コーパス』は出版年ごとのデータに記事数・文字数・ジャンル等の偏りがあるほか、著作権問題により一部の記事が非公開となっているなど、非常に不均衡な状態にある。そこで本稿では「美術・芸術」という語の使用割合を例に、5 種類の経年変化分析の方法を検討した。その結果、これまで多用されてきた粗頻度による分析法では有効に分析できない内容語があり、テキスト平均文字数当たりの調整頻度（PTA）を使用し、ロジスティック回帰分析によってジャンルを統制する分析法（PTA・ジャンル統制法）の方が有効な分析法だと考えられた。

１．はじめに

コーパスの代表性と均衡性は同時に語られることが多いが、これらは基本的に別々の概念である。ごく単純化して定義すれば、代表性とは「あるコーパスが、推定対象の言語を正確に反映していること」であり、均衡性とは「（推定対象の言語を母集団として、そこから均衡にサンプリングした結果）コーパスが均衡な特性をもつこと」であると言えよう。

しかし『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と呼ぶ）のような、均衡に設計されたコーパスに比べ、『太陽コーパス』の代表性に疑問符が付くのは否めない。「コーパスデータの大半が特定の雑誌だけから取られていた場合、得られた結果が言語全般の特徴なのか、当該雑誌の特徴なのかは不明です。」(石川, 2012, p. 22.)という指摘や、新聞記事のような一媒体では日本語の代表としてみなしにくいことを論じた後藤(1995, 1996)の主張は、重く受け止める必要がある。それでは『太陽コーパス』に全く代表性がないのかと言えば、それも程度問題ではあるだろう。雑誌『太陽』が近代日本語を代表する資料であることは、紛れもない事実であると思われる (国立国語研究所(編), 2005 ; 田中 , 2012)。

これに比べより明確な問題点は『太陽コーパス』の不均衡性にある。『太陽コーパス』はデータに様々な偏りを持っているため、そこに何らかの代表性があったとしても、その姿は大きくかき乱されて、単純な粗頻度分析では有効な値を示さない可能性がある。そこで本稿では、どのような分析法を使用すれば、少しでも有効な分析ができるのかについて、5 種類の分析法を比較しながら検討してみる。

２．『太陽コーパス』におけるデータのばらつき

２.１　出版年別テキスト数と文字数

『太陽コーパス』のデータがばらつく大きな原因の一つに著作権問題がある。『太陽コ

† hideaki@moriharuo.com

ーパス』は 5 年分の全記事をコーパス化することを目指して作成されたが、年代が新しくなるにしたがって、著作権上の問題により削除されたデータが増加し、それが 1925 年では文字数にして約 3 割に上っている。図 1 は削除前のデータ（評価版・内部資料）と削除後のデータ（『太陽コーパス』公刊版）の文字数の比較を示したグラフ、図 2 は記事数の比較を示したグラフである（国立国語研究所(編)(2005) , P. 20.の表 5, 6 を元に作成）。

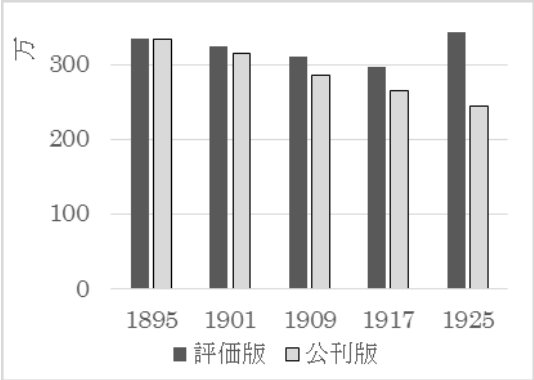


図 1 評価版と公刊版の文字数比較

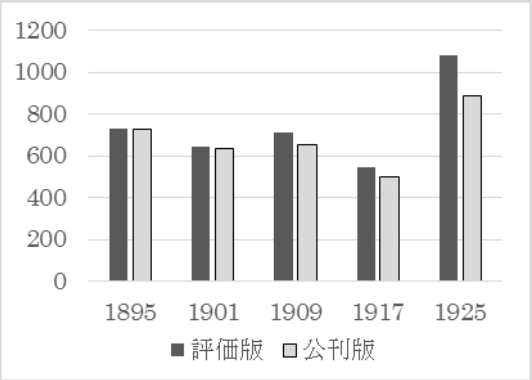


図 2 評価版と公刊版の記事数比較

図 1 で文字数の減少を追うと、1901 年から徐々に減り始め、1925 年では大きく減少しているのが分かる。著作権上の理由により一部のデータが公開できないという問題は、大規模な公開コーパスでは避けられないことであるが、一般利用者にとっては、難しい課題を抱える結果となる。例えば国立国語研究所(編)(2005)は、コーパスを制作した内部者の研究を所収しているため、これらの論文には削除前のデータが使用されている。しかし、公刊版では削除後のデータしか使用できないため、一般利用者がこれらの追試をしようとしても、初めからデータ数が食い違ってしまうのだ。このため、一般利用者は、削除後のデータを使用して、削除前のデータを使用した分析と同レベルの結果が出せるような、何らかの分析法を工夫する必要に迫られるのである。

しかし、『太陽コーパス』におけるデータのばらつきは、この著作権問題によるものだけではない。そもそも評価版の文字数自体が出版年ごとに異なっているのである。雑誌は商業的に作られた一媒体に過ぎないため、様々な要因によって各号ごとの記事数や文字数が異なる。そのような非常にばらつきのある言語資料を使用して近代日本語の平均的な姿を推定するためには、欠損の大きい 1925 年だけでなく、そもそもすべての出版年を均衡化させるような何らかの工夫が必要なのである。

図 2 は、記事数の比較である。これも年代が新しくなるにつれ徐々に公刊版の記事数が減少している。1925 年ではおよそ 2 割の記事が減少していることから、非公開となった記事には比較的長文のものが多かったことが推測される。また、出版年ごとの記事数のばらつきも激しい。特に評価版の段階から 1909 年は他の年より記事数が少なく、逆に 1925 年は非常に多い。この両年を比べると 1925 年は 1909 年のほぼ倍になっている。

２.２　１記事当たりの文字数

表 1 は、『太陽コーパス』全体の統計量である（扉・奥付等を除く）。記事数 3241 件のうち、最も短い記事の文字数は 27 字、最も長いものは 51705 字である。平均は 4442 字だが、中央値が約 3 千字であるため、『太陽コーパス』は多くの短い記事と少数の長い記事

によって構成されていることが推察される。文字数の多い記事上位 5%の文字量は全体の約 20%に及ぶ。これらごく少数の記事が分析対象の統計量を大きく左右している可能性がある。

表 1 記事の文字数統計量

記事数	3241
平均値	4442.17
中央値	2985.00
最頻値	643
標準偏差	4589.748
最小値	27
最大値	51705

表 2 出版年ごとの統計量

出版年	記事数	平均	標準偏差	最小	最大
1895	699	4752.30	4137.125	43	26643
1901	592	5304.35	4647.207	72	38152
1909	616	4624.80	4261.947	140	29343
1917	467	5669.53	6100.026	65	51705
1925	867	2812.55	3643.778	27	22482
合計	3241	4442.17	4589.748	27	51705

表 2 は出版年ごとの統計量だが、この平均の値を見ると予想通り 1917 年の平均は 5669 字、1925 年の平均は 2812 字であり、両年の記事の長さがかなり異なっている。これをさらに詳しく分析するため、記事の文字数を縦軸にして描いた箱ひげ図で検討する。図 3 は出版年ごとの全体図、図 4 は中心部を拡大した箱ひげ図、図 5 は参照のために掲げた BCCWJ「図書館書籍」サブコーパスの箱ひげ図である。

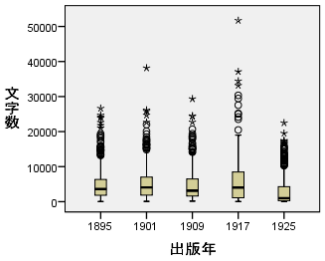


図 3 『太陽コーパス』全体

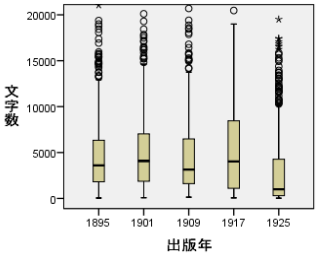


図 4 『太陽コーパス』中心部

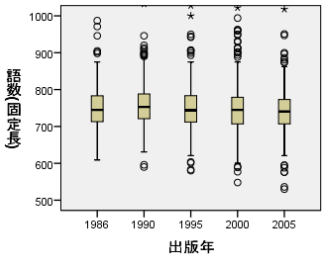


図 5 BCCWJ「図書館書籍」

図 3 を見ると各年とも多数の短い記事と、少数の長い記事によって構成されていることが確認される。特に 1917 年は長い記事が多い。図 4 で各年の箱の状態を見ると、初めの 3 年はほぼ均質だが、1917 年は箱の長さも長く、95 パーセンタイルの位置も高い。一方 1925 年は箱の長さも短く、中央値が極端に低くなっている。この年だけ極端に中央値が異なるのは、そもそもの記事数が多かったためと、著作権上の問題で長めの記事が非公開になったためと考えられる。一方、図 5 は、参考のために BCCWJ の「図書館書籍」サブコーパス（固定長）を約 5 年ごとに取り出し、その語数を描いたものである。これを見るとすべての出版年のデータが非常に均質である。BCCWJ と比較すると『太陽コーパス』がいかにばらつきを持ったコーパスであるかが良く分かる。

それではこれらのばらつきによって、具体的にどのような問題が生じるのであろうか。最も問題と思われるのが、出版年ごとに語の出現のしやすさが異なってしまう可能性である。一般にテキストの文字数が少なく制限されるほど、名詞比率は大きくなる。また 1 テキスト当たりの名詞比率が大きくなるほど、形容詞や副詞などが直線的に減少する（これを「樺島の法則」という。 樺島, 2009）。単純化していうなら、1917 年では長い記事の中で同じ語が何回も使われ、1925 年では短い記事の中で異なった語が次々と使われるため、

あるトピックに現れやすい語の頻度は、1917 年では多く、1925 年では少ない結果になりやすい。ただし Stubbs (2006) によれば助詞、助動詞などの機能語は、1 テキスト当たりの文字数にはそれほど影響されないため、これは主に内容語に生じる問題であると考えられる。

『太陽コーパス』のデータのばらつきが、単なる総文字数だけであれば、各出版年ごとの PMW などを求め、それで各年の比較を行えば問題ない。しかし、出版年ごとに 1 記事当たりの文字数に大きなばらつきがあるということは、PMW などでは平準化できない問題をはらんでいると考えられる。

２.３　ジャンル割合

『太陽』は雑誌であるため、各号のジャンルは時事的な出来事に大きく左右される。図 6 を見ると 1917 年では他の年より「社会科学」(点々)の割合が高い。この理由は 1914 年から始まった第一次世界大戦や、1917 年に勃発したロシア革命に関する記事が多く所収されたためだ。戦争のように社会的な影響力が強い出来事の場合、ジャンルの変化を言語の変化と見なす立場もあるかもしれない。しかし同年の「歴史」(横縞)などは、単に紙面上の制約によって減らされたと考える方が妥当だと思われる。

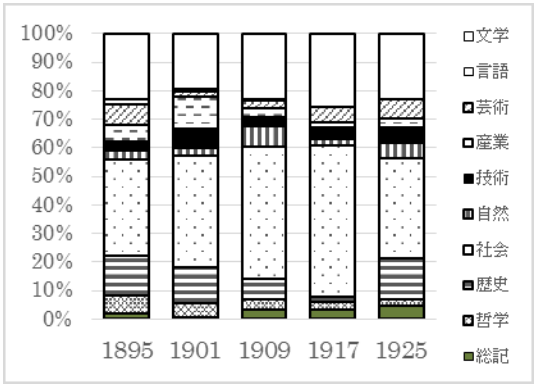


図 6 『太陽コーパス』のジャンルの割合

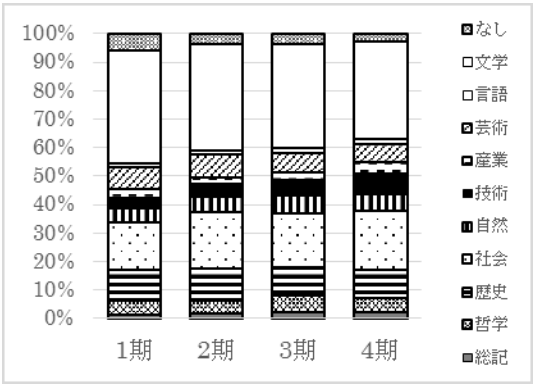


図 7 BCCWJ のジャンル割合

図 7 は、比較のため BCCWJ「図書館書籍」サブコーパスのジャンルを 5 年ごと 4 期に分けて示したものである。かなり均質な分布になっているが、古い時代ほど「技術」(黒)の割合が低く、「文学」(白)の割合が高いなどの変化も見える。これは技術的な内容の場合、技術の更新に伴って古い書籍が廃棄されやすく、文学は時代を経ても古い書籍がそのまま読み継がれるといった図書館特有の理由によるものだろう。

これに比べ、『太陽コーパス』では時代ごとに様々なジャンルのばらつきがあり、そこに一定の性質は見出せない。このようなばらつきは、その時々の編集方針によって生じたものと考えられる。ジャンルの偏りは、特定ジャンルに使用されやすい語の出現頻度に影響を与える。歴史ジャンルで使用されやすい語であれば、1917 年には歴史に関する記事そのものが極端に少ないため、その語の頻度も当然少なくなることが予想される。

３．均衡性に配慮した分析法の検討

第 2.2 節で述べたように、文字数やジャンルの影響を受けやすいのは内容語だと思われる。そこでここでは「美術・芸術」という内容語を取り上げ、データの均衡性に配慮した分析を様々に検討してみる。「美術・芸術」を取り上げるのはその語史がある程度解明されて

いるためである。なお、表計算ソフトは Excel、統計ソフトは SPSS を使用した。

３.１．「芸術・美術」の経年変化の予想

『日本国語大辞典第二版』では「芸術」の語史が以下のように記されている。

> 近世まではもっぱら「学芸・技術」の意で用いられたが、明治期に西洋文化の摂取が盛んになるに及んで、英語の art その他、美の表現・創造を共通の概念とするヨーロッパ各国語の訳語としての（中略）意味が出現した。ただし、明治初期にはむしろ同じ訳語に「美術」を用いることがより一般的であり、（中略）芸術が新しい意義で定着するのは、ほぼ明治三〇年（一八九七）前後である。（第四巻，p.1247.）

これからすると、「美術」の出現頻度は明治初期には高く、時代が新しくなるにつれ低くなる（「芸術」の頻度は逆の結果となる）ことが予想される。BCCWJ で「芸術割合」（芸術の頻度と美術の頻度の合計における芸術の頻度の割合）を調べると、「図書館書籍」サブコーパスの固定長で 46.4%、「出版書籍」サブコーパスの固定長で 53.6%、「特定目的」サブコーパスも含めた全体で 48.7%となっている。これらが最終的な使用割合であると仮定すれば、「芸術割合」はおよそ 50%前後で頭打ちになることが予想される。『太陽コーパス』を使用した分析法でこのような予想に合致する分析結果が得られるなら、その分析法はある程度有効な分析法だと考えられる。

ただし、両者の拮抗する時期が本当に 1897 年前後なのかは疑問である。この記述は同年に書かれた正岡子規の文章[1]を根拠にしているが、このような指摘は言語が変化し始めた初期に、言語変化を言葉の乱れとみなして表明されることが多い。『太陽コーパス』で正確な分析ができるなら、両者が拮抗する時期の推定もある程度正確に行えるはずである。

なお、検索ソフトは『ひまわり』を使用し、一般的な記事とは見なしにくい扉、奥付等のデータは分析から除く。また、旧漢字は新漢字に直して表記する。

３.２　粗頻度分析と出版年ごとの PMW 分析

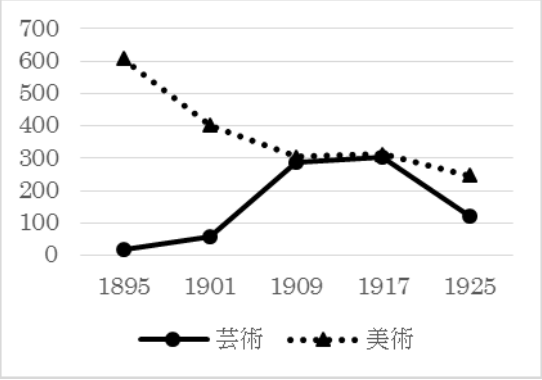


図 8 「芸術・美術」の粗頻度分析

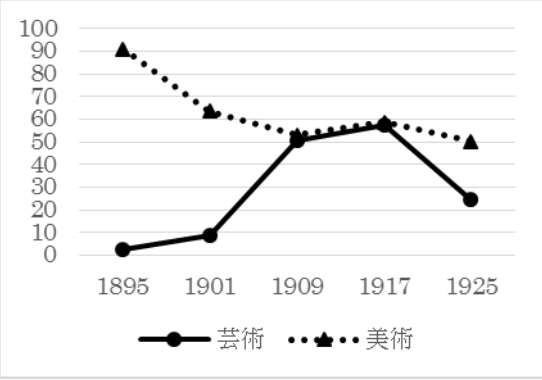


図 9 「芸術・美術」の PMW 分析

[1] 「博覧会などにて美術工芸品といふ語を用うるはやがて世人をして絵も彫刻も織物も陶器も同一美術品なりと誤想せしめ、従って美術品は工芸品なりと誤想せしむるの一大原因にやあらん」[閒人閒語]

図 8 は粗頻度（縦軸は語数）、図 9 は PMW（縦軸は PMW）のグラフである。2 つのグラフは、ほぼ似た形となっている。PMW 分析の場合、年代が新しくなるにつれて語の頻度が高くなるのは、著作権問題で削除された文字数分が補正されているからである。予想からすると「芸術」と「美術」の頻度は徐々に接近し、真ん中付近で一本の線になるはずであるから、1917 年まではある程度正確な分析になっているようにも見える。

問題は 1925 年である。この年は、もともと記事数が特別多かった上に、著作権問題のため、長めの記事が文字数で 3 割ほど非公開になっているなど、非常に偏りが大きい年であった。この年では懸念した通り両者の頻度が少なくなっている。『太陽コーパス』の場合、粗頻度分析と PMW 分析では有効に分析できない内容語があると言えるだろう。

３.３　割合による分析

前節で観察したように、粗頻度を使用した場合、有効に分析できない内容語がある。しかし、出版年ごとの割合であれば、同じばらつきの影響を受けたもの同士を割り算するため、それらの影響が相殺されて有効な分析となるようにも思われる。そこでここでは出版年別の「芸術割合」で分析してみる。

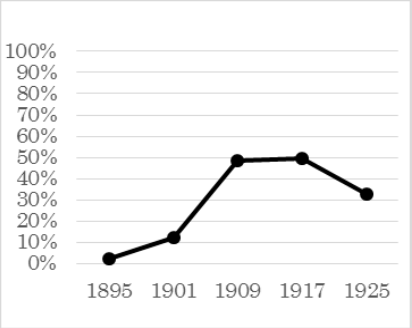


図 10 粗頻度を使用した芸術割合

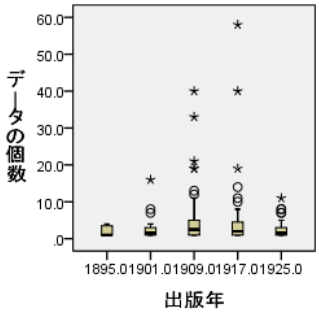


図 11 芸術の箱ひげ図

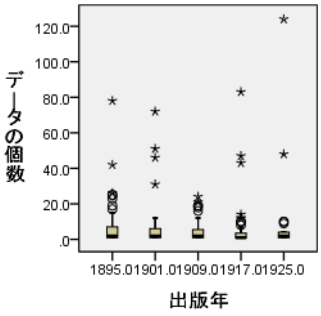


図 12 美術の箱ひげ図

図 10 は粗頻度を使用して算出した「芸術割合」である。1925 年では若干割合が高くなるが、1917 年よりまだ 20%程度も低い。割合分析であっても粗頻度を使用する限り、1925 年のばらつきが解消できない内容語があることが確認される。

この原因を探るため、1 記事当たり頻度の箱ひげ図で観察すると、図 11「芸術」の 1925 年ではすべての記事が十数頻度以下であるのに対し、図 12「美術」の 1925 年では 1 つだけ頻度 120 を超える記事がある。「美術」の 1925 年の総頻度は 73 記事合計で頻度 246 であるから、この記事だけで全頻度の約半数が出現していることになる。統計では、あまりに異質なデータは外れ値として分析に含めないことがある。この記事を外れ値として除外すれば 1925 年の「芸術割合」は 49.5%となる。これは予想に合致した値であるため、この記事は除外した方がよさそうにも思える。しかし、その判断は妥当だろうか。

『太陽コーパス』の記事は、最少文字数が 27 字、最大文字数が 51705 字というばらつきを持つ。そのためこのような現象は分析のたびに観察される可能性がある。そのたびに都合の悪いデータを除いていたのでは、データの改ざんにつながりかねない。問題の本質はこの記事だけにあるのではなく、『太陽コーパス』の多くのデータがばらついていることにあると考えるべきであろう。

３.４　テキスト平均文字数当たりの頻度（PTA）による割合分析

記事ごとの文字数のばらつきが分析を阻害する要因であるなら、記事の文字数がすべて平均文字数に近い 4500 字などであるとみなして、その分、頻度を調整した値にすればよい。例えば 450 字の記事に 1 回出現するなら、その頻度は 10（1÷450×4500）、45000 字の記事に 1 回出現するなら、その頻度は 0.1（1÷45000×4500）とカウントするのである。このようにテキスト平均文字数当たりの頻度に換算した調整頻度を PTA（per number of the text average letters）と命名する。ただし、「芸術」や「美術」は二字熟語なので、計算は（1÷検索語が出現したテキストの文字数×4500÷2）で行う。

しかし、実際にこの方法で分析しようとすると、ひとつ困難な課題に直面する。『太陽コーパス』では、記事ごとの文字数がタグ付けされていないため、PTA を求めたくとも、簡単には求められないのである。利用者各自が約 3400 記事について、文字数をタグ付けすることも考えられるが、誰しもが実施しやすい方法とは言えないだろう。

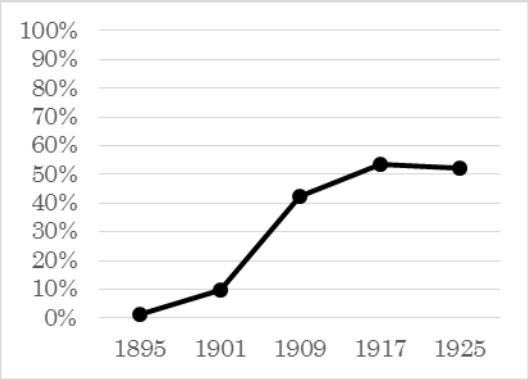


図 13 芸術 PTA 割合

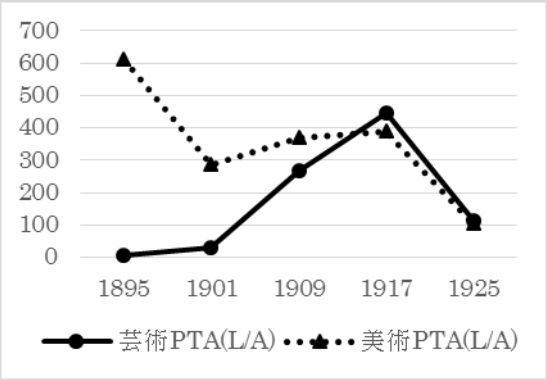


図 14 芸術と美術の PTA(L/A)の経年変化

そこで今回は検索語が出現した記事ごとの頻度リスト[2]と『太陽コーパス』全記事の文字数リスト[3]を作成し、同名の記事をできるだけ区別できるようにした上で、リストを利用した置換[4]を行って各記事の文字数を特定した。これによって「芸術 PTA 割合」を求めたのが図 13 である。図 13 の「芸術 PTA 割合」は、ほぼ予想と合致する結果である。このためテキスト平均文字数当たりの調整頻度（PTA）による割合分析は、ある程度有効であると思われる。図 14 は、各々の PTA をさらに出版年ごとの平均記事数で平準化させた調整頻度（こ

[2] 記事ごとの頻度リストの作成法：①『太陽コーパス』で検索した語の用例を Excel で開く。② ①のシートをコピーし、データ→重複の削除で、年・号・題名にチェックを入れ、重複を削除する。③元のシートで年・号・題名のデータを選択してピボットテーブルを挿入し、年・号・題名を列に、題名を値に入れて集計する。④ピボットテーブルの中身をコピー&ペーストし、ピボットテーブルは削除した上でデータに通し番号を付ける。⑤行ラベルでソートし、題名以外のデータを削除する。⑥通し番号でソートし、元の順番に直す。⑦題名とデータの個数をコピーし、②のシートの題名の横に挿入する。⑧題名が一致していることを確認したら、コピーした題名の列を削除する（同名記事で判別不能なものは平均値を使用）。

[3] 文字数リストの作成法：太陽コーパス付属のツール・プリズムを立ち上げ、①「入力 XML ファイル」ウインドウの下にある「別フォルダを指定」のボタンで、太陽コーパスの ZASSI フォルダ（Himawari_ﾊﾞｰｼﾞｮﾝ番号→Corpora→Zassi →Taiyo）を開き、corpus.xml を指定する。②「適用するスタイル」の中から csv.kiji.xsl を選び、「変換（ファイルへ出力）」を押すと、デスクトップにテキストファイルが出力される。これを Excel などで読み込む。

[4] リストを利用した置換の方法は http://stabucky.com/wp/archives/3259 で紹介されている（2014.07.22 閲覧）。

れを PTA(L/A)と記述する）でグラフ化したものである。これを見ると、どんなに文字数や記事数を平準化させても 1925 年の頻度は他の出版年と同程度にはならないことが分かる。これは 1925 年の偏りが単純な数だけの問題ではなく、「樺島の法則」に見られるような質的な問題によって偏りを持っていることを示唆している。このため、PTA(L/A)であっても頻度で分析することは難しく、出版年ごとの割合で分析するしかないと考えられる。

３.５　PTA・ジャンル統制法による分析

PTA 割合分析により、一定の成果は得られたが、『太陽コーパス』にはまだジャンルのばらつきという問題が残っている。そこで「芸術・美術」という語がどのようなジャンルで使われているのか、出版年の経過によってどのように変化していくのかを観察してみる。

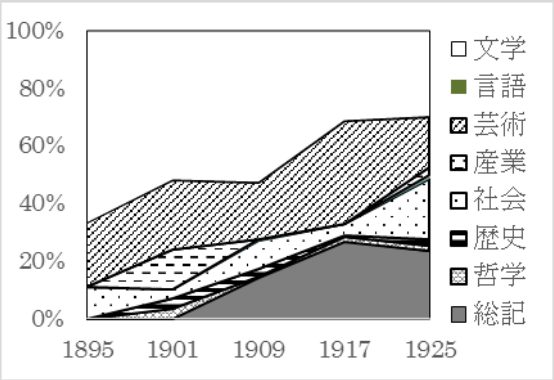


図 15 「芸術」が出現するジャンルの経年変化

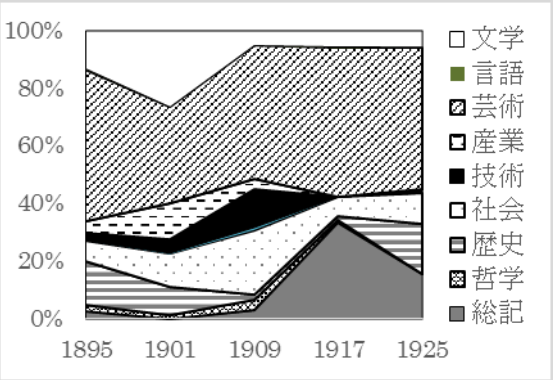


図 16 「美術」が出現するジャンルの経年変化

図 15 を見ると、「芸術」という語は、当初「文学」や「芸術」といった一部のジャンルでしか使用されていなかった。それが、年代を経るに従って多様なジャンルにも使用され、その使用割合も変容させていった様子が伺える。一方、図 16 の「美術」という語は、初めから多様なジャンルに使用され、基本的に同様の割合で使用され続けたように見える。

ただし、このグラフ自体は様々なジャンルが入り乱れている。例えば「産業」（横の鎖点）は、1901 年に多く現れるが、これは 1900 年にパリで行われた万国博覧会などの記事が所収されているからである。また、1909 年には「技術」（黒）が多く現れるが、これは黒田鵬心が 11 号と 13 号に仏教建築に関する長文の記事を書いているからである。『太陽コーパス』は時事的影響や編集方針がそのままデータに反映されるため、このような乱れが現れる。

しかし、近代日本語がこのように乱れていたわけではないだろう。近代日本語の平均な姿は、ジャンルの割合が一定か、緩やかな変化で推移していたと思われる。そこでここでは近代日本語において「美術」が使用されたジャンルの割合は一定であったと仮定し、全出版年のジャンル平均をすべての年代にあてはめた統制を行う。

分析は、ケースを PTA で重み付けし、「芸術・美術」を従属変数、出版年とジャンルを説明変数としたロジスティック回帰分析で行った[5]。変数は出版年も名義尺度とみなした。

[5] 「美術」における各偏回帰係数と標準誤差（カッコ内）は次の通り。P 値は** P<.01，* <.05 で表示。切片 –2.006** (.188)　1895 年 4.828** (.381)　1901 年 3.049** (.262)　1909 年 .503** (.160)　1917 年 .073 (.157)　（以上の参照カテゴリは 1925 年），総記 1.653** (.189)　哲学 2.026** (.437)　歴史 2.870** (.339)　社会科学 2.181 **(.203)　自然科学 1.879 * (.857)　技術・工学 5.480** (1.003)　産業 2.154** (.445)　芸術・美術 2.430** (.164)　言語 3.025 (1.780)（以上の参照カテゴリは文学）

ジャンルの効果は個々の偏回帰係数に個々の平均ジャンル割合をかけて求めた（この分析法を「PTA・ジャンル統制法」と命名する）。ロジスティック回帰分析を使用した言語分析の仕組みについては、横山・真田(2007a)を参照のこと。また、ロジスティック回帰分析を使用してデータの乱れを修正する方法は、横山・真田(2007b)に詳しい。

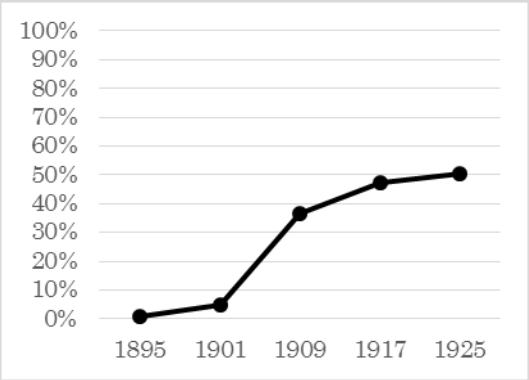


図 17　PTA・ジャンル統制法による芸術割合

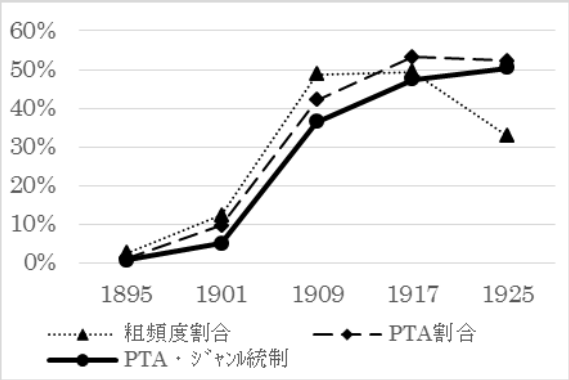


図 18 分析法による芸術割合の比較

PTA・ジャンル統制法の分析結果は、PTA 割合の分析結果より、全体的に「芸術割合」が低くなる。PTA・ジャンル統制法における 1925 年の値は 50.4%で、これはほぼ予想どおりの結果である。PTA 割合分析では 1917 年で一度高くなり、1925 年で若干落ち込んでいたが、ジャンルの統制を行うと一度も落ち込むことなく、全体がゆるやかな S 字カーブを描く。言語はその変化過程で S 字カーブを描くとされており、方言研究では井上(2000)などを初めとして多くの事例が報告されている。このような知見に合致する結果が得られたことからすると、PTA・ジャンル統制法による分析は、有効である可能性が高いだろう。

図 18 はこれまでに行った 3 種類の割合分析の結果を比較したものである。この３つのグラフの有効性を比較するため、条件を次の①～③にして「芸術・美術」を従属変数、出版年を説明変数としたロジスティック回帰分析を行った。①粗頻度を使用した場合の疑似決定係数は McFadden で.183、②PTA で重み付けした場合が.233、③PTA で重み付けし、さらにジャンルを説明変数に加えた場合が.349 となった。このことからしても、PTA・ジャンル統制法による分析がこの中では最も信頼できると考えられる。

４．代表性に配慮した分析法

これまで見てきたように、『太陽コーパス』は 1 記事文字数、出版年ごと記事数・文字数・ジャンルにばらつきがあり、さらに著作権問題によって一部のデータが非公開になっているなど、非常に不均衡なコーパスである。本稿ではこのばらつきを均衡化するため、5 種類の分析法を検討した。その中では、テキスト平均文字数当たりの調整頻度（PTA）を使用し、ロジスティック回帰分析によってジャンルを統制する分析法（PTA・ジャンル統制法）が最も有効だと考えられた。『太陽コーパス』に何らかの代表性があるとしても、その代表性の姿はデータの不均衡性によって乱されている。これを均衡にしていく分析法、すなわち均衡性に配慮した分析法がそのまま代表性に配慮した分析法にもなると考えられる。

ただし、ジャンルの統制を行うためには、近代日本語がどのようなジャンル割合になっているのかを広く調査し、それを使用するのでなければ真の代表性はないという考え方も成り立つ。田中(2012)によれば、国立国語研究所では「通時コーパス作成」のため近代語の

資料選定が行われているというが、その研究はまだ途上にあるようだ。外部資料としては『近代女性雑誌コーパス』の利用も考えられるが、データ量が少なく使用が難しい。このような状況下でも言語変化のS字カーブのように、すでに知られている知見との整合性に留意することで、一定の代表性は確保できると思われる。言語変化の分析を行った際、その結果があまりに歪んでいるなら、分析法を再検討する必要があるだろう。

もう一つ配慮が必要なのが、口語体と文語体の問題である。『太陽コーパス』では口語体と文語体が混在しているが、これらは語の選択や文法体系が異なり、基本的に異なったレジスターだと考えられる。Biber ほか(2003)では代表性の問題に関連して、「全体的一般化というものは、どのレジスターにとっても正確でないことが多く、むしろ現実には全く存在しないような言語の記述をしてしまうことになる。」(p.41.)と注意を喚起している。

本稿では詳しく触れられなかったが、「美術・芸術」の使用傾向において口語体・文語体による違いは見られなかった。しかし分析対象によっては口語体と文語体で異なる振る舞いを見せることも予想される。例えば文語体では使用率がどんどん増加するのに、口語体では徐々に使用率が減少する形式の場合、口語体と文語体を混合させて分析すると、あたかもその使用率は一定であるかのように見える。これは年代とともに文語体の記事自体が減少するためである。しかし近代日本語でそのような言語は、現実には全く存在しない。近代日本語では基本的に口語体と文語体という書き言葉しか、現実には存在していないのである。このため常に口語体と文語体による使い分けに留意し、使い分けがあった場合はこれらを分離して記述することが、もう一つの代表性に配慮した分析法だと思われる。

参考文献

石川慎一郎(2012)『ベーシック　コーパス言語学』ひつじ書房

井上史雄(2000)『東北方言の変遷』秋山書店

樺島忠雄(2009)「語彙量の実態」計量国語学(編)『計量国語学辞典』朝倉書店，pp.93-97.

国立国語研究所(編)(2005)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社

後藤 斉(1995)「言語研究のデータとしてのコーパスの概念について　―日本語のコーパス言語学のために―」『東北大学言語科学論集』4 . pp. 71-87.

後藤 斉(1996) 「コーパスとしての新聞記事データ―終助詞「かしら」をめぐって―」『東北大学言語学論集』5. pp. 37-46.

Stubbs, M. (2002) *Words and Phrases; Corpus Studies of Lexical Semantics*. Blackwell Publishing. 南出康世・石川慎一郎(監訳)(2006)『コーパス語彙意味論―語から句へ―』研究社

田中牧郎(2012)「近代語コーパスにおける資料選定の考え方」近代語コーパス設計のための文献言語研究 成果報告書 (国立国語研究所共同研究報告 12-03)

Biber, D.,Conrad,S., Reppen,R. (1998) *Corpus Linguistic; Investigating Language Structure and Use*. Cambridge University Press.　齊藤俊雄・朝尾幸次郎・山崎俊次・新井洋一・梅咲敦子・塚本聡(訳)(2003)『コーパス言語学　―言語構造と用法の研究―』南雲堂

横山詔一・真田治子(2007a)「フィールド言語学にロジスティック回帰分析は寄与しうるか」情報処理学会研究報告. 人文科学とコンピュータ研究会報告 (9) ， pp. 9-16.

横山詔一・真田治子(2007b)「多変量S字カーブによる言語変化の解析　―仮想方言データのシミュレーション―」『計量国語学』26-3 ， pp.79-93.

地方議会会議録コーパスを用いたオノマトペの分析

高丸圭一（宇都宮共和大学シティライフ学部）†
内田ゆず（北海学園大学工学部）
乙武北斗（福岡大学工学部）
木村泰知（小樽商科大学商学部）

Analysis of Onomatopoeias in the Corpus of Regional Assembly Minutes

Keiichi Takamaru (Utsunomiya Kyowa University)
Yuzu Uchida (Hokkai-Gakuen University)
Hokuto Ototake (Fukuoka University)
Yasutomo Kimura (Otaru University of Commerce)

要旨

本研究は全国の都道府県および市区町村から収集した地方議会会議録コーパスを用いて，現代の話しことばにおけるオノマトペの出現傾向を明らかにすることを目的とする。地域差を含むコーパスからオノマトペの用例を抽出し，出現分布について具体的な分析を行う。全国403自治体から収集した2010年度の地方議会会議録コーパス約3億語を対象として，形態素解析器によってオノマトペの抽出を試みた結果，982語（186,416例）が抽出された．やや改まった公的な場での話しことばにおいてもオノマトペが豊富に使用されていることが明らかになった。特に，「しっかり」「はっきり」「どんどん」など，施策の推進や明確な言及，適切な判断を求めるために使われる表現が高い頻度で出現した。次に，正確な分析を進めるため，出現頻度が中程度のオノマトペに対象を限定し，手作業によって誤抽出を取り除いた。155語（12,512例）のオノマトペについて，地域差の検討を行った。この結果，オノマトペは西日本において有意に出現確率が高いことが明らかになった。対応分析を行い，それぞれの地方に出現する特徴的なオノマトペを分析した。この結果，方言的な語義が観察された。また，新しい用法をもつオノマトペの分布や固有表現（名称）に使用されるオノマトペの地域差が観察された。

１．はじめに

筆者らは地方議会会議録を収集し，コーパスとして学際的に利用することを目指した研究を進めている（木村，渋木，高丸，乙武，森 2012）。高丸，渋木，木村（2011）では，全国の地方自治体における会議録の公開の状況について調査し，2010年の時点で73.4%の市区町村議会の会議録がウェブに公開されていることを示した。ウェブに公開された地方議会会議録をプログラムによる自動処理によって収集・整形し，関係データベースに登録を行った（齋藤他 2011，菅原他2012）。さらに，会議録コーパスの単語n-gramデータの構築やウェブユーザインターフェイス（検索システム）の構築等を試みている（乙武，高丸，渋木，木村，森 2013）。地方議会会議録は特定の自治体に居住する者の発言が，地域別・年度別に記録されたものである。一つの地方議会の会議録を遡れば通時的な言語変化を辿ることができ，全国の地方議会会議録を横断的に調査すれば地域差を分析することができ

† takamaru@kyowa-u.ac.jp

るため，言語研究の資源として注目に値する。井上(2013)や高丸(2014)では，地方議会会議録に含まれる方言語彙や文末表現の地域差を指摘している。また，高丸(2013)では，複合名詞を対象としたテキストマイニング分析によって，地方議会における話題の遷移の可視化を試みている。

本研究では，地方議会会議録コーパスにおけるオノマトペの出現状況について分析する。日本語には豊富なオノマトペ（擬音語および擬態語）があり，音，雰囲気，程度，様子を効果的に伝えるために用いられることが知られている。近年，オノマトペの工学的な利活用を目指した取り組みが盛んに行われている（小松，中村 2012）。地方議会会議録コーパスの分析から，現代の話しことばにおけるオノマトペの使用頻度やオノマトペ使用の地域差を明らかにすることを目的とする。

２．地方議会会議録コーパス

２.１　概要

本研究では，収集した地方議会会議録コーパスのうち，全国403自治体の2010年度の会議録を用いる。収録された総単語数は293,190,430語である。地方議会会議録は議会での発言をすべて記録することを目的としている。しかし，議会を円滑に運営する目的で，議員の発言（質問）内容は事前に通告されており，読み上げ原稿が存在する発言が含まれる。また，整文の作業によって話しことばの特徴の一部が書きことば的に修正されている（高丸 2011）。この2つの点において，会議録は厳密には自由会話の書き起こし資料であるとはいえないため，議会会議録から話しことばの特徴を分析する際にはこの点に注意が必要となる。ただし，読み上げ原稿は話すことを目的に用意されたものであるため，作文の朗読とは異なると考えられる。また，地方議会会議録の整文指針（野村，鵜沼 1996）によると，オノマトペは整文によって修正される対象ではない。これらのことから，本研究における整文の影響は少ないと考える。

２.２　オノマトペ

『日本語オノマトペ辞典』（小野編 2007）に掲載された4,565語のうち，意味分類別索引に掲載された2,466語（異なり語数1,751語）を対象としてオノマトペの抽出を行う。形態素解析にはJUMAN[1]を用いる。ユーザ形態素辞書に1,751語のオノマトペをひらがなおよびカタカナの副詞として登録した上で形態素解析を行い，登録したオノマトペを含む文を用例として抜き出した。

３．形態素解析によるオノマトペの抽出

３.１　抽出結果と精度

形態素解析の結果, 約3億語の地方議会会議録コーパスから982語のオノマトペ（計186,416例）が抽出された。先行研究（高丸，内田，乙武，木村 2014）において，このうち61語4,164例について, 手作業にて誤抽出の調査を行った。誤抽出は, (i)方言に起因する解析誤り（1,527例），(ii)名称・固有名詞の一部（720例），(iii)他のオノマトペの一部（58例），(iv)言い間違い・入力ミス等（27例），(v)同音異義語（28例），(vi)その他（149例）の6つのパタンに大きく分類された。オノマトペのモーラ数別の正抽出率は表1のとおりであった。

モーラ数の短いオノマトペと一致する部分文字列が文中に多数存在するため，3モーラ以

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

下のオノマトペでは正抽出率が低い。特に，2モーラのオノマトペ14語のうち8語（例えば，「じゃん」「さっ」「すい」）は正抽出率が0%であった。4モーラのオノマトペ27語は正抽出率が72.7%と比較的高く，このうち20語（例えば「ぶつぶつ」「どっさり」「しゃんしゃん」）は正抽出率が100%であった。一方，「かったん」「かんから」の2語は正抽出率が0%であった。これらは大阪府や兵庫県内の会議録に多く，それぞれ「〜できへん**かったん**かな」「〜せなあ**かんから**」のように，方言に起因する誤抽出であった。

オノマトペの抽出においては，形態素解析器のみによって高い精度で抽出することは困難であるため，品詞や後続の形態素のパタン（木村，渋木，内田，乙武，高丸，森 2014），係り先の動詞（内田，荒木，米山 2012）などを利用した抽出手法が検討されている。本研究においては，次節では誤抽出が比較的少ないと考えられる4モーラのオノマトペを対象として，全体的な傾向を述べる。また，4章以降では分析対象のオノマトペを限定し，手作業で誤抽出を取り除いた上で，出現傾向について考察する。

表1　オノマトペ61語の誤り分析結果（モーラ数別）

モーラ数	語数	正抽出	誤抽出	合計	正抽出率
2	14	34	1006	1040	3.3%
3	20	673	1147	1820	37.0%
4	27	948	356	1304	72.7%

表2　4モーラ以上のオノマトペ（出現頻度上位15語）

順位	オノマトペ	出現頻度
1	しっかり	77,464
2	どんどん	20,680
3	はっきり	19,382
4	だんだん	5,679
5	びっくり	2,910
6	もろもろ	2,372
7	そろそろ	2,021
8	ゆっくり	1,610
9	じっくり	1,541
10	わくわく	889
11	つくづく	622
12	すくすく	542
13	ぎりぎり	513
14	がっかり	499
15	こうこう	485

３.２　出現確率上位のオノマトペ

地方議会会議録におけるオノマトペの全体的な出現傾向を確認するために，4モーラ以上のオノマトペの出現頻度上位15位までを表2に示す。「しっかり」「どんどん」「はっきり」などが多く使用された。特に「しっかり」の出現頻度が顕著に高く，抽出したオノマトペ全体の約41%を占める。出現頻度の高いオノマトペは多様な文脈で用いられるものであると考えられるが，上述の例のように地方議会会議録では，施策の推進（「しっかり」「どん

どん」等）や明確な言及や適切な判断（「はっきり」等）などを表すために高頻度で使用される。使用例を以下に示す。これらの語を手がかりに文の主題を抽出することで，発言者の主張や議論の焦点などを分析することができる可能性がある。これは今後の検討課題である。

(1)　「しっかり」の例

- ぜひ教科書の採択に沿って、やはり授業時間との関係が必ず出てくる、予測される問題でございますので、**しっかり**とした対応をお願いしておきたいと、こんなふうに思っております。（東京都荒川区）
- よっぽど気をつけて、これから早期整備、全庁的な体制、**しっかり**つくっていただくようにお願いをしておきます。（三重県津市）
- その辺は**しっかり**と地域の人と合意を、暗黙の合意ちゅうたらおかしいけど、ある程度話をしちょっていただきたいと思いますが、どうなんでしょうか。（山口県周南市）

(2)　「どんどん」の例

- ですから、私は、地域力というものを高めるために、観光の予算を**どんどん**使っていただきたいと言っているわけでありまして、ぜひ、そういったこともお願いしたいなと思っております。（北海道）
- 何も、バリアフリーなどは、うちも**どんどん**やらんとあかんという立場で物言うてるわけです。（大阪府八尾市）
- ・・・この統合庁舎は大変大きなプロジェクトでありますから、いろんな角度から検討していただき、委員会を**どんどん**開いてほしいと思います。（沖縄県うるま市）

(3)　「はっきり」の例

- やっぱり子育て支援ではだめなので、少子化対策という、そういう旗をしっかり掲げて、それがすぐできるというものではないことは**はっきり**しているんですよ。（北海道旭川市）
- ・・・地域の担い手をつくるのはいいけれども，じゃあ，行政は何をするのかというところも，**はっきり**させる必要があるということを感じました。（茨城県守谷市）
- ・・・やはり大切な税の消滅、免除ということを**はっきり**と説明するべきだと思いますが、市長の考え方をお聞きいたします。（新潟県上越市）

４．オノマトペの分析

４.１　方法

地方議会会議録にオノマトペの出現を分析するために，形態素解析から得られた抽出結果を手作業にて確認し，オノマトペではないものを除外する。982語（186,416例）をすべて確認することは困難であるため，出現数が中程度のオノマトペを用例数ベースで1割程度選び，分析対象とする。出現頻度が高いオノマトペはどの地域においても一定の頻度で出現していると考えられる。出現頻度のより低いオノマトペにおいて，用法（語義）の多様性や地域差がみられる可能性がある。一方，コーパスにおける出現頻度がきわめて低い語は出現分布を分析することに適していない。そこで，会議録コーパスにおいて，都道府県別の出現確率（オノマトペ数／総単語数）の和が5×10^{-6}以上50×10^{-6}未満の177語（18,545例）を分析対象とする。

手作業による確認作業の結果，177語（18,545例）のうち，155語（12,512例）がオノマトペであると確認された。22語はすべてが誤抽出であった。手作業でオノマトペであると確認した155語の出現頻度（全国計）を付録に示す。次節では，これらの語の地方別の出現傾向について考察する。

４.２　オノマトペ出現頻度の地域差

４.２.１　分析対象全体

オノマトペ地域差を見るために，会議録を「北海道東北」「関東」「中部」「近畿」「中国四国」「九州沖縄」の6地方区分に分け，それぞれの出現頻度および，出現確率（総単語数に占めるオノマトペ数）を求めた。分析対象155語全体の出現頻度を表3に示す。

出現確率は「近畿」（53.9×10^{-6}）が最も高く，「中国四国」（50.3×10^{-6}），「九州沖縄」（44.5×10^{-6}）がこれに続く。西日本において，オノマトペの出現頻度が高い傾向がみられる。二群の比率の差の検定（Rのprop.test関数を総当たりで実行）の結果，「北海道東北」と「中部」の間のp値が0.7156で有意差がみられなかったほかは，すべての組み合わせに有意差がみられた。

表3　分析対象155語の地域別の出現頻度と出現確率（百万分率）

	北海道東北	関東	中部	近畿	中国四国	九州沖縄
出現頻度	1,152	3,780	1,720	3,055	1,498	1,307
総単語数	31,895,757	97,129,985	48,287,270	56,699,023	29,795,513	29,382,882
出現確率（$\times10^{-6}$）	36.1	38.9	35.6	53.9	50.3	44.5

４.２.２　オノマトペごとの分布

前節でオノマトペの出現確率には地方によって差があり，西日本において出現確率が高い傾向があることを示した。本節では，どのオノマトペが出現確率の偏りに影響しているのかを確認する。このため，155語のうち全国の出現頻度が50回以上の84語を対象として，6地方区分ごとの出現頻度の対応分析を行った。第1次元と第2次元の関係を図1(a)に，第2次元と第3次元の関係を図1(b)にそれぞれ示す。なお，第1次元から第3次元の寄与率はそれぞれ，36.4%，25.5%，17.4%である。

グラフから第1次元には「九州沖縄」が，第2次元には「近畿」が，第3次元には「中国四国」が分離されていることを読み取ることができる。多くの語は中心付近に集中しており，地域差に大きく寄与していない。一方，それぞれの地方の近傍に布置している幾つかの語は，出現頻度に地域差がみられる語であると考えられる。図1(a)で「九州沖縄」，および，「近畿」の近傍に布置する語の出現頻度をそれぞれ表4，5に示す。また，図1(b)で「中国四国」の近傍に布置する語の出現頻度を表6に示す。

「九州沖縄」に頻出するオノマトペのうち「ぴしゃっ」は全112例のうち90例が九州地方で使用された。「戸口などを閉めるさま」から派生して「ある基準で完全にやめる，閉める，止めるさま」を表す用例が以下のように全国に見られる。

(4)-a　「ぴしゃっ」の例（ある基準で完全にやめる，閉める，止めるさま）

- ・・・市長はお金がないというたった一言で**ぴしゃっ**と切ってしまったんです。（和歌山県和歌山市）
- 逆にきょう呼んでしっかり話をして、それで**ぴしゃっ**とやめるとなれば、やれないことないと思うけど、呼んでちゃんと話をして・・・（長野県松本市）
- ・・・あの地域内に水が入るのを**ぴしゃっ**と入り口でとめていただいたということで、大変対応がよかったというふうなことを、まず、感謝を申し上げたいと思います。（山口県山陽小野田市）

一方，九州地方の「ぴしゃっ」には方言的な語義が存在し，「きちんと」「しっかりと」「はっきりと」に類する意味をもつ例が見られる。議会において全国的に出現頻度の高い

「しっかり」「はっきり」と言い換えられるため，九州地方の議会において多用される表現であると考えられる。

(4)-b　「ぴしゃっ」の例（九州方言の語義）

- だから、審査会の権限とか、そういうものについて**ぴしゃっ**と明らかにして示す必要があるんじゃないかというふうに思うんですが、そこら辺についてどう考えますか。（福岡県嘉麻市）
- そりゃ、職員は黒字が出ようが赤字が出ようが**ぴしゃっ**とボーナスも出ておる。（長崎県雲仙市）
- それから、そのときも出たんですけれども、資料の提出方法を、もうちょっと**ぴしゃっ**と定めた方がいいのかなというような気はいたしました。（熊本県熊本市）
- ここら辺を**ぴしゃっ**と整備しなければいけないというふうに思っておりますし、住民から信頼される自治体となるべく努力をしていかなければいけない・・・（宮崎県小林市）

「近畿」に多く出現するオノマトペ「ばくっ[2]」は，オノマトペ辞典に掲載された語義での使用は存在せず，「（棒状のものが）ばくっと折れる」（岩手県）1例のほかは，以下に示すような「漠然とした大まかなさな」の意味で用いられた。これは「漠とした」がオノマトペ化した用法であると解釈することができ，近畿を中心に使用される表現である。

(5)　「ばくっ」の例（漠然とした大まかなさま）

- ・・・まずは、何ていうんですかね、議会としての**ばくっ**とした考え方なりを聞きたいというそういった思いで委員会、特別委員会だと思っております。（愛知県尾張旭市）
- ・・・どういう課題が出てくるのかということについては**ばくっ**とは聞きましたけれども、たらればという話がいっぱいあってなかなか定かになっていない・・・（滋賀県大津市）
- 平均値がどれぐらいちゅのがもしわかれば、**ばくっ**とでも結構です。（大阪府羽曳野市）
- こうなってくると、**ばくっ**と私が考えるのは、逆にその責任体制が分散してしまって、だれがじゃあこのプロジェクトを回していくのか・・・（兵庫県豊岡市）

「中国四国」に多く出現した「きらら」は，「明るくまぶしく輝き続けているさま」をあらわすオノマトペであり，派生して雲母の別称でもある。自治体が行うイベントや取り組み，施設等の名称に多く採用されており，会議録においてもすべて固有表現（名称）として出現した。特に山口県で出現例が多く，「きらら」が使用された固有表現26種類（109例）のうち，11種類（55例）が山口県において出現している。

(6)　「きらら」の例

- ・・・業務用米としての「**きらら**397」のほか、「おぼろづき」「ふっくりんこ」といった良食味米に続きまして、「ゆめぴりか」も新たに登場するということで・・・（北海道）
- ・・・従来より山口大学医学部附属病院などとも連携を図りながら、消防防災ヘリ「**きらら**」を活用し、ドクターヘリ的な運用による救急救命活動を展開しております。（山口県宇部市）
- ・・・補助するグループホームはゆもと苑に併設されている**きらら**の里で、この施設以外は全部火災通報設備は設置されているとの説明がありました。（山口県長門市）
- 県は**きらら**博に始まり、国民文化祭、そして国民体育大会と大型イベントを行う一方で、福祉政策の費用をカットする態度は、決して許されるものではない。（山口県周南市）
- 山陽小野田市**きらら**ガラス未来館の指定管理者の指定について質疑を行います。（山口県山陽小野田市）

[2] オノマトペ辞典に掲載された語義は「勢いよく食いつく」「大きく開く」。

表4　「九州沖縄」近傍に布置する2語の出現頻度

	北海道東北	関東	中部	近畿	中国四国	九州沖縄
ぴしゃっ	0	3	4	7	8	90
ずっ	6	22	7	18	7	74

表5　「近畿」近傍に布置するオノマトペ7語の出現頻度

	北海道東北	関東	中部	近畿	中国四国	九州沖縄
ばくっ	2	3	9	52	2	1
かちっ	4	10	2	50	14	1
さんさん	2	9	7	60	2	19
めちゃくちゃ	1	9	13	42	6	4
ころっ	9	16	5	57	16	9
ぐんぐん	2	11	11	26	1	1
のびのび	5	102	13	132	12	20

表6　「中国四国」近傍に布置する2語の出現頻度

	北海道東北	関東	中部	近畿	中国四国	九州沖縄
きらら	15	4	6	5	59	20
ぴちっ	0	11	1	16	33	5

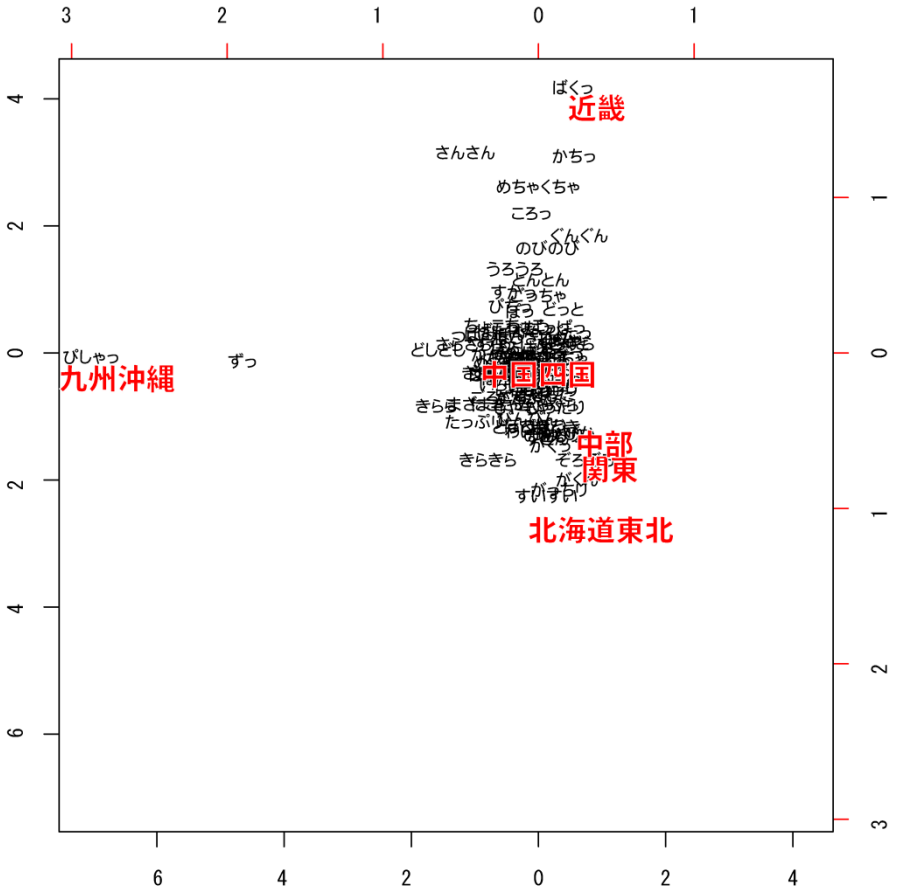


図1(a)　出現総数が50回以上のオノマトペ84語の出現地域と出現数の対応分析
（横軸：第1次元，縦軸：第2次元）

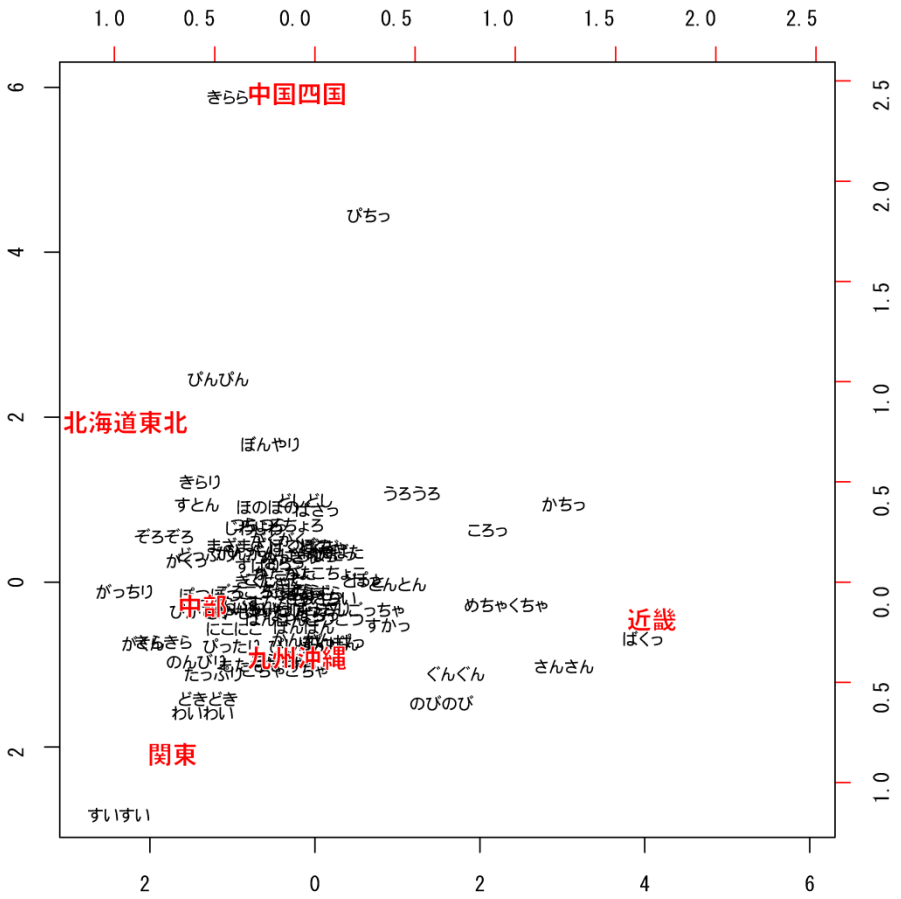


図1(b)　出現総数が50回以上のオノマトペ84語の出現地域と出現数の対応分析
（横軸：第2次元，縦軸：第3次元）

５．まとめ

本研究では，やや公的な話しことばである地方議会会議録においてオノマトペの出現に傾向ついて分析した。全国402自治体から収集した2010年度の地方議会会議録コーパス約3億語を対象として，形態素解析器によってオノマトペの抽出を試みた結果，982語（186,416例）が抽出された。オノマトペが豊富に使用されていることが明らかになった．特に施策の推進や明確な言及や適切な判断を求めるために使われる表現が顕著に高い頻度で出現した。

次に，手作業によって形態素解析における抽出誤りを除外した155語（12,512例）のオノマトペを対象に出現傾向の地域差を分析した。全体的な傾向として，西日本においてオノマトペの出現確率が有意に高いことが明らかになった。また，地方別の出現頻度を対応分析した。第1～3次元において，それぞれ「九州沖縄」「近畿」「中国四国」に出現するオノマトペがそれぞれ分離して表れることを確認した。九州沖縄地方では方言的語義の「ぴしゃっ」が多く出現することを確認した。漠然とした大まかなさまを表す「ばくっ」が近畿地方を中心として分布していることを確認した。また，中国四国地方で多く観察された「きらら」は山口県において固有表現（名称）に多用されていた。

今後は，分析対象のオノマトペ数を拡大し，オノマトペの語義の多様性や固有表現への採用を中心に検討を進めていく。語義については，構文解析器を利用して係り先動詞等と

の関係などの検討を進める予定である。商品名，店名，施設名に利用されるオノマトペについては，使用されやすいオノマトペの種類や後続語の連接などについて研究が行われている（田守 2012）。会議録に出現する施策や施設，イベント等の名称に利用されるオノマトペについてもオノマトペのイメージを利用して明るさや親しみやすさを演出するもの（例えば，「地域活動支援センターさんさん」（施設名称）（沖縄県西原町）），直感的な理解を促すもの（「ぶらぶら歩きがここちよいまち」（まちづくりの理念）（東京都昭島市）などに分けることができると考えられる。この点についても今後分析を進めたい。

謝　辞

本研究の一部は，科研費基盤研究(C)（No.26370498）「学際的応用を考慮した地方議会会議録コーパスの言語学的研究」（研究代表者：高丸圭一），および，科研費基盤研究(C)（No.25370524）「公共用語の地域差に関する社会言語学的総合研究」（研究代表者：井上史雄）による。

文　献

木村泰知，渋木英潔，高丸圭一，乙武北斗，森辰則(2012)「地方議会会議録コーパスの構築とその利用」第26回人工知能学会全国大会，3B3-NFC-4-3

高丸圭一，渋木英潔，木村泰知 (2011)「全国の市町村議会会議録のウェブ公開とデータ提供の状況」都市経済研究年報，第11号，pp.47-72

齋藤誠，大城卓，菅原晃平，永井隆広，渋木英潔，木村泰知，森辰則 (2011)「地方議会会議録の収集とコーパスの構築」言語処理学会第17回年次大会論文集，P2-21

菅原晃平，大城卓，齋藤誠，永井隆広，渋木英潔，木村泰知，森辰則 (2012)「地方議会会議録コーパスの拡充における問題点の分析と対処」言語処理学会第18回年次大会論文集，P1-15

乙武北斗，高丸圭一，渋木英潔，木村泰知，森辰則 (2013)「地方議会会議録コーパスの学際的応用を目的としたn-gramデータの構築およびウェブUIの試作」言語処理学会第19回年次大会発表論文集，pp.733-736

井上史雄(2013)「去った〇日」『ことばの散歩道』明治書院，pp.154-155.

高丸圭一(2014)「地方議会会議録コーパスにおける出現確率の相関を用いた文末表現の地域差の分析」社会言語科学会第33回研究大会，pp.174-177

高丸圭一(2013)「地方議会では何が話題になっているのか―宇都宮市議会会議録のテキストマイニング―」都市経済研究年報，13，pp.162-173

小松孝徳，中村聡史(2012)「オノマトペの利活用：オノマトペ研究の分野横断連携を目指して」，人工知能学会誌27(6)，pp.653-654

高丸圭一(2011)「規模の異なる自治体における地方議会会議録の整文の比較」社会言語科学会第27回研究大会発表論文集，pp.256-259

野村稔・鵜沼信二(1996)『地方議会実務講座 第3巻』ぎょうせい

小野正弘編(2007)『日本語オノマトペ辞典』小学館

高丸圭一，内田ゆず，乙武北斗，木村泰知(2014)「地方議会会議録におけるオノマトペの出現傾向に関する基礎的検討」言語処理学会第20回年次大会，pp.566-569

木村泰知，渋木英潔，内田ゆず，乙武北斗，高丸圭一，森辰則(2014)「地方議会会議録におけるオノマトペの自動抽出手法の提案」『第30回ファジィシステムシンポジウム（FSS2014）

講演論文集』（2014年9月発表予定）
内田ゆず，荒木健治，米山淳(2012)「ブログ記事からのオノマトペ用例文の自動抽出手法」Journal of Japan Society for Fuzzy Theory and Intelligent Informatics 24(3), pp.811-820
田守育啓(2012)「商品名および店名・施設名に利用されているオノマトペ」『人文論集』47, pp.49-70

関連URL

地方議会会議録コーパスプロジェクト　http://local-politics.jp/

付録

分析対象155語の総出現頻度

オノマトペ	出現頻度
ずばり	297
のびのび	284
ちょこちょこ	275
きらり	273
ぼん	271
ゆったり	267
すんなり	246
のんびり	244
ごちゃごちゃ	240
わいわい	240
ぽっ	234
だらだら	234
ばたばた	230
ぴったり	230
どっと	223
ばらばら	217
ずらっ	209
きらきら	197
ついつい	195
にこにこ	192
ばんばん	178
いらいら	156
さらっ	149
うろうろ	137
がくがく	137
ずっ	134
ごっちゃ	133
くどくど	132
ばっさり	132
がたがた	132
がんがん	130
ちらほら	125
じわじわ	121
がっちり	120
ごたごた	119
さらさら	118
ぎくしゃく	117
ころっ	112
ぴしゃっ	112

オノマトペ	出現頻度
きらら	109
とんとん	107
ぼんぼん	106
ぽんぽん	101
ちらちら	101
さんさん	99
ぼつぼつ	98
どしどし	97
ぴたっ	96
ふらふら	94
すとん	91
たっぷり	88
どきどき	87
かちっ	81
ほのぼの	81
ばさっ	80
こつこつ	79
ぱらぱら	78
びっしり	78
めちゃくちゃ	75
もやもや	73
かんかん	72
ぽつぽつ	72
ばくっ	69
ごちゃ	67
ぴちっ	66
ごろごろ	66
ちょろちょろ	65
めっきり	64
すかっ	63
ぼんやり	62
ぱっぱっ	61
ずらずら	61
つらつら	61
まざまざ	60
がくん	59
すいすい	59
ぶらぶら	56
ぴかぴか	53

オノマトペ	出現頻度
ぐんぐん	52
もたもた	52
ぴんぴん	51
がくっ	51
どっぷり	50
ぞろぞろ	50
どたばた	49
がっくり	49
ばりばり	48
がらり	48
ぶらり	48
こそこそ	47
ぐずぐず	47
きちん	47
ぱん	46
てきぱき	45
どかん	45
どきっ	44
じゃんじゃん	44
がばっ	43
ぼやっ	43
どろどろ	43
がちがち	41
きちっ	41
ぼこぼこ	40
とことこ	40
ころり	39
じりじり	39
ずばっ	39
ぶつぶつ	39
ぼろぼろ	38
うんざり	37
ふわっ	36
ふわふわ	36
ぴっ	35
はらはら	35
あたふた	34
どさくさ	34
がたっ	33

オノマトペ	出現頻度
はっし	33
つるつる	32
ぽっきり	32
ざわざわ	30
ぶんぶん	30
やきもき	29
ぼーん	27
ぽっかり	27
ぶくぶく	26
ちゃん	25
ぐいぐい	22
どっさり	22
ぼっ	22
ちょん	19
しゃんしゃん	17
のこのこ	12
もくもく	11
ちゃっ	10
からっ	10
とくとく	7
にやっ	7
じん	7
もやっ	7
ごん	7
くっ	5
しゅん	5
とっと	5
ぐい	4
ずきん	3
びっ	3
じっ	2
ちん	2
ごくごく	2
ぼたん	1
がしっ	1
たっ	1
ばりっ	1
ふっ	1

現代日本語の類義表現に関するテクスト言語学的研究
—「焦点を当てる」と「焦点を置く」に着目して—

吉本秋水（高知県立大学大学院人間生活学研究科）

A Text Linguistic Analysis of Synonymous Expressions in Present-Day Japanese, with Special Reference to *syooten o ateru* and *syooten o oku*

YOSHIMOTO Shuusui (University of Kochi Graduate School of Human Life Science)

要旨

現代日本語の類義表現として「A に焦点を当てる」と「A に焦点を置く」という形式が見られ、いずれも「A を議論の対象とする」という意味で使用されていると考えられるが、これらの使い分けに関する記述は国語辞典や類義語辞典等には見られない。本研究ではその使い分けを規定する要因をテクスト構造の観点から明らかにする事を目的とする。そのため、編集の強弱という点で異なる性質を持つ 2 つのコーパス『ヨミダス文書館』と『筑波 Web コーパス』を利用して用例採取を行い、文脈を構成する情報を基準とする分類を試みた。その結果ある目的が先行文脈において示されている場合、主にその目的達成の為の手段・観点に対して「焦点を当てる」が選択され、目的そのものあるいは目的達成のために取る立場には「焦点を置く」が選択されている傾向が見られることについて論じる。加えて、この使い分けはテクストのジャンルに依存していない可能性についても言及する。

1. はじめに

現代日本語の表現には同等の意味を持ちながら、使い分けの理由が不明瞭である表現が存在する。その様な表現は、「ゆれ」や「変異」という観点から、または日本語学習者に対する日本語指導時の問題点として論じられる事がある。そのような類義表現として本研究では「A に焦点を当てる」と「A に焦点を置く」を研究対象とする。「焦点を当てる」および「焦点を置く」は、いずれも「A を議論の対象にする」という意味や、次の用例(1)、(2)[1]より「ある物事に注意・関心を向ける」、「ある問題点・課題を取り上げる、主題にする」という意味で使用されていると考えられる。

(1) 三宅氏は、野田氏の「原点」の地に立ち、野田内閣を信任するかどうか、に**焦点を当てる**戦いを展開している。

（「［12 衆院選・選挙区を歩く］首相陣営「刺客」に危機感」2012.11.26）

[1] (1)、(2)は、ともに出典は『ヨミダス文書館』。本稿では新聞記事の場合は記事タイトルと日付、Web 上の文書はタイトルと URL を付す。

(2) 日本画の部で最優秀賞にあたる知事賞を受賞した伊藤明徳さんの「生きる」は、サクラの老樹のコケに覆われた幹に**焦点を置いた**大胆な構図が特徴。
（「多彩に「芸術の秋」 県展、伊藤さんら知事賞＝島根」2011.11.21）

しかし、この 2 表現の意味の違いや使い分けについては、日本語母語話者でも説明が困難であると考えられる。そして、この 2 表現の明確かつ詳細な記述は辞書には見られない。そもそも「焦点を当てる」と「焦点を置く」は辞書の見出し語「焦点」の用例として採択される事は少ない[2]が、実際には使用が認められる。

本稿では「焦点を当てる」と「焦点を置く」はテクスト言語学的の観点に立つと使い分けの確認が可能であり、出現予測が可能である事を示していく。そして「焦点を当てる」と「焦点を置く」は、特定の条件下では言葉のゆれではなく、それぞれ異なる意味を伴って使用されている事を明らかにする。

2. 先行研究

本研究が対象とする 2 表現を直接考察の対象とした先行研究は見付かっていない。そこで、類義語に関する先行研究を参考にして本研究を進めていくことにするが、それらの研究手法には本研究の立場上、以下の留意すべき点が見られる。

第一に、分析対象とする類義語が含まれた用例を自作した上で、類義語同士を置き換える[3]などの操作をした際に研究者が感じる違和感を元にした意味分析が存在する。この場合日本語使用の実態を知る事が出来ない可能性がある。

第二に、実際に記述された用例を用いた分析であっても、用例出典元の年代や使用場面が統一されていないものがある。

第三に、意味分析の際に文を最大単位としている研究が見られる事である。「話し手・聞き手の意図や主張を考察し、言語の全体像に迫るためには、文を超える談話をも対象にせざるを得ない」(児玉: 2002, p. 115)という指摘がある。

以上に挙げた留意点に対しては、コーパスの利用およびテクスト言語学の観点を導入することで本研究を進める。

3. 研究方法

3.1. テクストとテクスト言語学

テクストについて Halliday & Hasan (1976, pp. 1-2)は、“A text is best regarded as a SEMANTIC unit: a unit not of form but of meaning.”と定義している[4]。当該表現が使用されている文脈を適切に捉える為に、このテクストを意味分析の単位とする。実際に文レベルの分析では文脈の意図を捉えられない例も本稿で示す。

テクスト言語学について、庵(1999, p. 15)は以下の考えを提示している。

---

[2] 1889 年～2014 年に出版された主要な辞書 38 冊を調査した結果、「焦点を当てる」は『三省堂国語辞典』（第四版、第七版）（加えて「議論の対象にする」という説明が記述されている）と『新明解国語辞典』（第三版～第六版）、「焦点を置く」は『新明解国語辞典』（第三版）において確認された。

[3] 類義語同士を置き換える操作は学校教科書にも記述が見られる（『新しい国語２』（東京書籍）、『中学生の国語 二年』(三省堂) 等)。これは「語感を磨き語彙を豊かにする」(『中学校学習指導要領解説　国語編』pp. 59-60.）という指針に基づいたものであると考えられる。

[4] 本稿ではこの意味での‘text’の日本語表記を、引用部分を除いて「テクスト」と表記する。

テキスト言語学は，実時間内という制限された条件下で人間が行っているテキスト処理の過程の解明を主目的とする学問分野である。

加えて、庵(1999, p. 15)は「文文法は時間の制約と独立に研究してもよいが，テキスト言語学においてはそれだけでは不十分で，極めて短時間で言語処理が可能であるのはなぜかという問いに答える必要がある」という指摘もしている。本研究は「焦点を当てる」と「焦点を置く」の 2 表現は母語話者でも使い分けの説明が不明瞭である事を前提にしているが、それらの表現が使用されたテクスト、特にこれら 2 表現が同一テクスト中で使用されている例が存在する。テクスト言語学の立場は類義表現の意味分析を考える上で、テクスト作成者の言語処理の過程を考える理論的枠組みになると考える。

そして、実時間内で作成されたテクストである用例をコーパスから採取する事で、テクストの使用場面・年代の統一を図り、日本語使用の実態に即した意味分析を行う。

また、言葉のゆれに関しては土井(1963)の提示した以下の定義を採用する。

> 二つの異なる言語形式が、互に類似した意味を持つか、一方の意味や用法が変化することによって、単独でか、または他の語句と呼応して、同一共時態において、同一場面に共存する現象。(土井: 1963, p. 104)

本稿では、文脈およびテクスト構造を限定することで、2 表現に使い分けが存在する事を主張し、2 表現が言葉のゆれではない事を示す。

3.2. 使用コーパス

用例は報道文と Web 文書から採択する。用例の採択元となるコーパスは、読売新聞の提供する記事データベース『ヨミダス文書館』（以下ヨミダス）と、筑波大学が構築した『筑波ウェブコーパス』(以下 TWC) を用いる。共に Web サイト上から記事および用例を得る。用例抽出時の注意点として、①外国語文書の日本語翻訳の除去、②冗長的な表現や誤植を含むテクストも採択の対象とする（記述的立場に基づく）、の 2 点を挙げておく。特に①に関して、新聞記事の場合は日本政府と外国政府との共同宣言文の日本語訳も記事として収録されている。また TWC の検索結果で表示された用例および Web サイトのタイトルや URL だけでは、翻訳文であるかの判別が難しいテクストが存在するため、出典元の Web サイトの確認が必要となる。

なお、本稿では「焦点を当てる」と「焦点を置く」が対象とする物（「A に焦点を当てる」の A に相当する物）を「焦点対象」と呼ぶことにする。

3.3. 分析方法

分析にあたっては、先ずテクスト中の文脈を構成する情報を基準とする分類を試みた。具体的には、焦点対象が文脈中でどの様な重要度を伴って扱われているのかを分類した。テクスト中で焦点対象以外にも問題点や課題等を挙げている場合、それらと焦点対象との重要度の大小関係に着目した。更に、ある目的が文脈において示されている場合、焦点対象が目的、その目的達成の為の手段・観点、あるいは目的達成のために取る立場のいずれに該当するのかを分類した。そして、「焦点を当てる」と「焦点を置く」の使用傾向を分析した。

4. 新聞記事を用いた分析

ヨミダスでは 1986 年以降の記事を参照可能である。その内「焦点を当てる」が用いられている記事は 3602 件存在する（2013 年 7 月現在)。そのうち、2012 年 1 月 1 日から 2012 年 12 月 31 日までの、計 142 記事を分析対象として用いた。一方「焦点を置く」が用いられている記事は 65 件存在する（2013 年 7 月現在)。ただし、2012 年での使用例は見られない。また、ヨミダスに収録されている記事中での使用頻度も少ないため、ヨミダス収録分の全記事にあたる 1986 年以降の全 65 記事を対象に用例を抽出した。

次に、抽出した用例から、「焦点を当てる」と「焦点を置く」の焦点対象に傾向があるかどうかを調べた。その結果、焦点対象そのものに傾向や使い分けの規則は観察されなかった。

続いて、焦点対象の扱いに着目して分析を行った。初めに、焦点対象の重要度に関する特徴を示す。

- 焦点を当てる: 単独使用時に、焦点対象を強調して取り上げるが、他の問題点・課題と比較して、焦点対象の重要度の大小を含意しない。
- 焦点を置く: 単独使用時に、焦点対象を強調し、且つ、他の問題点・課題よりも重要度が大きい事を含意する事がある。

特にテクスト中に目標・目的、目標達成の為の手段・観点、目標達成の為の立ち位置を示す表現が出現する場合（207 記事中 58 記事）には、次の対比が観察された。

- 焦点を当てる: 手段・観点を焦点対象とする
- 焦点を置く: 目的・目標を焦点対象とする

これらのことを以下に代表的な用例を挙げて論じる。

用例(3)は党首討論に関する記事であり、目的と目的達成の手段が示されている。

(3)　野田首相が目指す消費税引き上げと、自民党の谷垣総裁が求める衆院の早期解散をともに満たす“解”はあったのかー。首相と谷垣氏の 25 日の極秘会談は、3 月以降の「消費税政局」を前に、互いに接点を見いだそうとする狙いがあったとみられる。29 日の党首討論は極秘会談でのすり合わせも反映し、首相と谷垣氏が消費増税の必要性で認識を共有するなど、議論がかみ合う場面もあった。（中略）会談で合意点があったのかどうかは明らかになっていないが、29 日の党首討論では、首相と谷垣氏の接近を印象づける場面もあった。谷垣氏の討論は、前回討論とは違い、「消費増税はマニフェスト（政権公約）違反」という理屈での解散要求を抑制。社会保障・税一体改革の中身の議論に**焦点をあてた**。

（「党首討論 消費増税 必要性は共有　首相・谷垣氏 極秘会談の影響か」2012.03.01）

このテクストでの先行文脈から読み取れる自民党の谷垣総裁の目的は、首相に対する解散要求を通す事である。その為の理屈（＝手段）であるが、記事では前回の党首討論で出された「消費増税はマニフェスト違反である」という主張から「社会保障・税一体改革の

中身の議論」に変更されていることが判る。この「中身の議論」が焦点対象となっている。そして、この焦点対象以外にテクスト中で挙げられた問題点・課題が前回討論までの理屈である「マニフェスト違反」であると考えられる。

29 日の党首討論において、谷垣総裁が考える衆院解散の理屈は「中身の議論」に変更された。これは首相と谷垣総裁の間で「消費増税の必要性」というのは共通認識として持っているのにも拘わらず、解散要求でマニフェスト違反を主張してしまうと、自民党谷垣総裁の立場をも否定することになる為である。

また、「中身の議論」が「マニフェスト違反」よりも重要度が高いということまでは判断が難しい。それは、一度はマニフェスト違反を主張した都合上、社会保障問題等での主張に舵を切ったのはやや消極的な印象をこの記事から受けるからである。

次の用例(4)はシドニー五輪開催に向けての警備体制に関する記事である。

(4) 五輪警備を統括するオリンピック警備指令センターの、ポール・マッキノン指令は、シドニー五輪期間中のテロリズム対策が、海外からのテロリストの潜入を防ぐ水際作戦に**焦点を置く**と語る。「豪州国内にテロが巣くうというより、海外からテロが輸入される可能性の方が深刻だ。我々は、米国、カナダ、英国などの諜報（ちょうほう）機関と情報交換しながら、まず、危険分子の入国を許さないことを目指す」。狭き門の空路に比べ、海からの侵入は比較的容易で狙われやすい。ＲＩＢを使ったテロ対策は、そんな諜報活動をも視野に入れる。

（「[シドニー五輪の舞台裏] (1)テロへの備え　秘密兵器で水際作戦（連載）」2000.08.15）

このテクストでの目標・目的は「五輪期間中のテロリズム対策」である。焦点対象となっている「海外からのテロリストの潜入を防ぐ水際作戦」は「テロリズム対策」を行う上での立場であると、文脈上考えられる。そして、その目的達成の為の手段が「(米国、カナダ、英国などの諜報機関と情報交換をしながら）危険分子の入国を許さないこと」である。

また、「まず」という表現も用いられており、記事上からは他のテロリズム対策の存在も想起させられるが、それらの対策よりも「テロリストの入国阻止」が重要視されていると読み取ることが可能である。

5. Web ページの文書を用いた分析

次に、TWC から抽出した用例を用いた分析結果を示す。「焦点を当てる」の頻度は 6316 で、その内 245 件を用例採択した。「焦点を置く」の頻度は 354 であり、114 件を用例採択した。用例採択にあたって、TWC の検索結果と共に表示される出典元の URL も辿り、リンク切れでない場合は出典元から用例採択を行った。

続いて、「焦点を当てる」と「焦点を置く」で焦点対象の扱いに着目して分析を行った。着眼点は、新聞の場合と同じである。分析の結果、以下の特徴が観察された。

- 焦点を当てる: 単独使用時に、焦点対象を強調するが、他の問題点・課題と比較して、焦点対象の重要度の大小を含意しない。また、「焦点を当てる」が用いられる事で、他の問題点・課題が議題から排除される事があるが、一時的な印象がある。
- 焦点を置く: 単独使用時に、焦点対象を強調し、且つ、他の問題点・課題よりも重要度が大きい事を含意する事がある。

新聞記事の分析と同様の結果が得られた。新聞記事では「焦点を置く」に、焦点対象以外の問題点・課題よりも重要である事を強調する意味が観察されたが、Web 上のテクストからは「焦点を当てる」にも同様の意味が観察された。その場合、テクスト内で焦点対象とならなかった問題点・課題に対しては、次回以降で議論の対象とするなど、一時的に焦点対象となっている例が見られた。これらの事を、以下に用例を挙げて論じる。

用例(5)はマンション管理人の、廊下上の私物に対する苦情への解決策が主題となっているテクストである。

(5) 近くにあるマンションの中で、「ここの管理は素晴らしい」と思えるところがあります。T急さんが管理するマンション(A)です。T急さんをヨイショする気がありませんが、「いいものはいい」と評価させていただきます。ただし、全部のT急さんがいいわけではなく、「だめだな、ここは」というマンションもよそにはあります。とにかく、今回は、この素晴らしさに**焦点をあてます**。今、当マンションでは、「廊下上の私物」について論争しています。管理組合がずっと放置していたために、あまりにもひどい状況になったからです。

（規則 管理人はつらいよ http://tsuraiyo.com/topic6606REIGAI.html リンク切れ）

ここでの焦点対象は、マンション A の「素晴らしさ」である。テクストの文脈上、テクスト作成者のマンションでの廊下上の私物問題を解決するための手段・方法として、マンション A の管理体制を用いるということだと考えられる。また、このテクストではマンション A の「だめだな、ここは」という難点を挙げながらも、「今回は」その難点を排除し、良い面だけを主題にしている。このテクストの主旨は、マンション A の良い面から、自分の管理するマンションを改善する手本にする事である。その主旨だけを考えると、マンション A の難点を挙げる必要はないのだが、あえて記述している事で、難点も良い面もテクスト作成者にとっては共に重要な点であるという事が窺える。従って、ここでは一時的にマンション A の難点は議論から排除されているが、良い面である「素晴らしさ」と同様に重要である事に変わりはない。

次の用例(6)からは、文レベルはなく、テクストレベルでの分析が重要である事が判る。これは部下や同僚から報告を受ける際に、それがたとえ良くない報告であっても自発的に正確な報告を受ける方法に関するテクストである。テクスト作成者は自身の子どもが通知表を貰ってきた際のやりとりを例に挙げて説明している。

(6) このような部下やメンバーがいます。これをそのまま報告したらきっと怒られる。だから、真実を少し変えて、上司に怒られないように報告をしよう。（中略）うちの子達は、必ず通知表を妻と私に堂々と見せてくれます。しかも、嬉しそうに見せてくれます。褒められるような成績なのか？というと、これがそうでもないんですね。私は学習塾の経営もしていますので、通知表には敏感でいないといけないのですが特に小学生に関しては、あまり通知用の中身には興味がありません。（中略）そうです、私は各教科の ◎ ○ △ などはあまり見ていないのです。（変な親ですよね）それより、まずは一学期きちんと学校に行ったことを認める。先生が褒めてくれている部分に**焦点を当てる**。その上で、夏休みはどうするの？と聞いて、子どもなりの

目標や計画を聞いて終わりです。以前、「お前、すごいな～、こんなに○があって！！」と言って褒めたら、「ちゃんと見てよ、○より◎のほうがいいんだよ！」と子どもに言われたこともあります。

（リーダーのための人材活用術　http://manpower.dreamstudy.jp/article/14063853.html）

このテクストで「焦点を当てる」が使用されている文とその前後の限られた部分だけを読むと、テクスト作成者は通知表の成績欄よりも、焦点対象となっている「先生が褒めてくれる部分」の方を重視しており、成績はないがしろにしていると読み取れる。「焦点を当てる」も、先に挙げた「他の問題点・課題と比較して、重要度の大小を含意しない」という用法からも外れている様に見える。しかし、この用例 (6)のテクスト全体から明らかになるのは、テクスト作成者がこの例を持ち出した本意は、敢えて成績に触れない事で子ども自らが「褒められるような成績」ではなくても、自ら積極的に成績を申告するようになるという事である。従って(6)での「焦点を当てる」は焦点対象だけでなく通知表の成績部分も同様に重要視していると考えられる。焦点対象となっている「先生が褒めてくれている部分」を褒めることが、子どもからの自己申告を促す手段である。

この(6)は、コーパスから抽出した用例に基づく分析の注意点も提起される。TWC で見出し語を検索した場合、結果として得られる用例は見出し語を含む文とその前後 1 文ずつである。仮に本研究で当該表現を含む 3 文のみを用例として分析を行うと、異なる結果を得る事になる。検索結果から表示された用例だけでなく、用例の参照元となった Web ページの本文も可能な限り参照して、テクストの全体像を把握する事が必要となる。

次の用例(7)は、大学の研究科の紹介文である。本研究では TWC から得られた用例は、新聞記事よりも比較的統制の弱いテクストであるとしているが、この用例(7)は大学の紹介という性格上、テクスト作成後に校閲がされていると考えられる。従って、Web 上の文書の中でも統制が比較的強いテクストである。

(7) 機能の創生

ものの創生の基礎である機能に**焦点を置き**、過程、構造、分子の重層した視点から、それぞれと密接な関係を持つ生産、材料、科学、評価の分野を対象とする研究分野から構成される。機能の高度な理解とそれに基づいた評価を統合し、新しい機能の創生機構を生み出すことを目的とした研究をおこなう。

（横浜国立大学大学院工学研究科｜大学院工学研究院概要｜組織
http://kenkyuin.eng.ynu.ac.jp/outline/organization/index.html）

この用例(7)は、「ものの創生の基礎である機能」の高度な理解という前提の元で、「新しい機能の創生機構を生み出すことを目的とした研究をおこなう」と、目的を明記しているテクストである。その目的達成の為の手段・観点は「過程、構造、分子の重層した視点」である。

最後に、同一テクスト内で「焦点を当てる」と「焦点を置く」の 2 表現が出現する用例を分析する。新聞記事からの用例は、テクスト内での当該表現の出現は、どちらか一方だったが、TWC では複数例得る事が出来た。また、用例(8)、(9)は、ともに目標・目的、目標達成の為の手段・観点、目標達成の為の立ち位置が示されている。

(8) 2010 年度実施計画
本年度は、次年度の本格的な現地調査に向けて、グジャラート州東部のアーメダバード県において、予備的調査を実施する。
（中略）
製作現場において、実際に製作に従事し、手工芸品の素材、道具、技術に**焦点をおく**と同時に、作業過程の中から「ものづくりの勘所」「作り手個人の創意工夫」について観察分析をおこなう予定である。また、同時に調査地への経由となるデリーやムンバイなど大都市においても、商品としての手工芸品についての流通経路や消費者の動向にも**焦点をあて**、次年度の調査へ向けての予備的聞き取り調査を実施する予定である。
（国立民族学博物館｜研究部：科学研究費補助金による研究プロジェクト
http://www.minpaku.ac.jp/research/sr/22720075.html）

(9) 魂のカウンセリングは一般的なカウンセリングとは違います。
一般的なカウンセリングでは、満たされなかった感情を追体験することや、問題の原因を知ることで、過去のトラウマ（記憶）を解消することに**焦点をおき**ますが、ここでは、それらの奥にある、本当の「私」＝魂に先に**焦点をあて**ていきます。
なぜならば「私」が育ってしまえば、過去のトラウマも、今、抱えている問題も、そこを理解し、許していかれるようになるからです。過去のトラウマや、問題、悩み、エゴ（パターン）が悪いのではなく、本当の「私」が育っていないことでそれらが自分を苦しめているということです。
（伊藤洋子 魂の道場・カウンセリング-システム・料金 http://ponta55.com/system.html）

用例(8)は次年度の本格的な現地調査に先立って行う予備的調査が大きな目的となっている。その一環として実際に工芸品の制作に従事するが、その立ち位置を「手工芸品の素材、道具、技術」（「焦点を置く」の焦点対象）に設定している。また、「焦点を当てる」の焦点対象である「流通経路」と「消費者の動向」を採り上げることは、次年度への調査に向けた予備的聞き取り調査を行う為の手段・観点であると考えられる。

用例(9)では、「一般的なカウンセリング」の目標・目的が「過去のトラウマ解消」（「焦点を置く」の焦点対象）である。一方で「魂のカウンセリング」が「一般的なカウンセリング」と異なる点は、手段・観点が「本当の「私」＝魂」（「焦点を当てる」の焦点対象）という点である。

この用例(8)と(9)は、TWC からの用例では 2 例のみであったが、同一テクスト内で「焦点を当てる」と「焦点を置く」が使用されているテクストである。これら以前の用例では、テクスト構造上「焦点を当てる」と「焦点を置く」の使い分けの要因を、それぞれ別個のテクストによる傾向から論じてきた。そこから、テクスト内容に目標・目的、目標達成の為の手段・観点、目標達成の為の立ち位置が示されている場合、「標達成の為の手段・観点に焦点は当てられる」、「目標または目標達成の為の立ち位置（前提）に焦点は置かれる」という傾向が見られた。この 2 例から予測可能な事は、当該 2 表現を両方知っているテクスト作成者も、その傾向に沿ったテクスト処理を施した結果であることが推察される。

6. 結論

本研究により、「焦点を当てる」と「焦点を置く」には、「ある物事に注意・関心を向け

る」、「ある問題点・課題を取り上げる、主題にする」という共通する意味をもつと共に、以下の特徴が明らかになった。

- 焦点を当てる: 単独使用時に、焦点対象を強調するが、他の問題点・課題と比較して、焦点対象の重要度の大小を含意しない。また、「焦点を当てる」が用いられる事で、他の問題点・課題が議題から排除される事があるが、一時的な印象がある。
- 焦点を置く: 単独使用時に、焦点対象を強調し、且つ、他の問題点・課題よりも重要度が大きい事を含意する事がある。

更に、先行文脈において目標・目的、目標達成の為の手段・観点、目標達成の為の立ち位置が示されている場合、更に同一テクスト内に当該 2 表現が出現していても、以下の使い分けが予測可能であることが示唆された。

- 焦点を当てる: 手段・観点を焦点対象とする
- 焦点を置く: 目的・目標、または目標達成の為の立ち位置（前提）を焦点対象とする

テクスト構造を限定する事で、「焦点を当てる」と「焦点を置く」に使い分けが存在し、テクスト内での出現予測が可能であると考えられる。同時に、先に挙げた土井(1963)の提示した言葉のゆれの定義の条件である「二つの異なる言語形式が、互に類似した意味を持つ」を否定する。「焦点を当てる」と「焦点を置く」は言葉のゆれではない。

また、本研究では編集の強弱という点で異なる性質を持つ 2 つのコーパス『ヨミダス文書館』と『筑波 Web コーパス』から用例採択を行った。その 2 つのコーパスから得た用例を分析した結果、「焦点を当てる」と「焦点を置く」に同様の傾向が見られたということは、編集の強弱という観点からはテクストのジャンルに依存しない傾向が可能性として窺える。

本研究より、一見すると無秩序に用いられているかに見える言語使用にも、文を超えた範疇であるテクストの持つ情報が関与し、人間の言語処理の過程を説明出来るのではないかと考える。

7. 今後の検討課題

本研究では類義語に関する先行研究の留意点を挙げながらも、検討が不十分な点がある。

第一に、用例採択において新聞記事での「焦点を当てる」の用例出典年が 2012 年なのに対して、「焦点を置く」は 1986 年から 2012 年までの 26 年間の幅がある。更に、編集の強弱という点で、Web ページと新聞記事の 2 つのコーパスを用いたが、Web ページの場合、テクスト作成の日時が不明な場合がある。TWC 自体は 2013 年に Web 上に存在した文書を元に構築されており、テクストの作成日時を 2013 年であることを保証するものではない。この為、用例の採択年代にずれが生じており、共時性の面で問題が残る。

第二に、Web 上のテクストを編集の弱い用例としたが、その中でも本稿で例示した用例(7)のように編集が比較的強いテクストも存在するため、更なるテクストの分類が必要である。また、編集の強いテクストとして新聞記事（報道文）を用いたが、一社のみの記事であり、その他の編集の強いテクスト（法令文書等）を視野に入れた分析ではない。この為、ジャンル依存性の点で問題が残る。

以上の点を今後の検討課題とする。

参考文献

庵 功雄(1999)「テキスト言語学の観点から見た談話・テキスト研究概観」『言語文化』, 36, pp. 3-19. (http://hdl.handle.net/10086/8622/よりダウンロード可能)
児玉徳美(2002)『意味論の対象と方法』.くろしお出版.
土井洋一(1963)「音韻交替についての一解釈(上)：バ・マ行音の〈ゆれ〉をめぐって」『研究年報』10, pp. 103-123.
Halliday, M. A. K. and Hasan, Ruqaiya(1976) *Cohesion in English* Longman.

教科書等

中洌正堯、長谷川孝士、尾木和英、他(2012)『中学生の国語 二年』株式会社三省堂.
三角洋一、相澤秀夫、有澤俊太郎、他(2012)『新しい国語 2』東京書籍株式会社.
文部科学省(2008)『中学校学習指導要領解説 国語編』株式会社東洋館出版社.

辞書

「焦点」.(1993). 見坊豪紀、金田一京助、金田一春彦、柴田 武、飛田良文 編、『三省堂国語辞典 第四版』(p. 546). 三省堂.
「焦点」.(2014). 見坊豪紀、市川 孝、飛田良文、山崎 誠、飯間浩明、塩田雄大 編、『三省堂国語辞典 第七版』(p. 704). 三省堂.
「焦点」.(1981). 見坊豪紀、金田一春彦、柴田 武、山田忠雄、金田一京助 編、『新明解国語辞典 第三版』(p. 555). 三省堂.
「焦点」.(1989). 金田一京助、柴田 武、山田明雄、山田忠雄 編、『新明解国語辞典 第四版』(p. 608). 三省堂.
「焦点」.(1997). 金田一京助、山田忠雄、柴田 武、酒井憲二、倉持保男、山田明雄 編、『新明解国語辞典 第五版』(p. 674). 三省堂.
「焦点」.(2005). 山田忠雄、柴田武、酒井憲二、倉持保男、山田昭雄 編、『新明解国語辞典 第六版（机上版）』(p. 715). 三省堂.

コーパス URL

読売新聞『ヨミダス文書館』https://database.yomiuri.co.jp/rekishikan/
筑波大学・国立国語研究所・Lago 言語研究所『NINJAL-LWP for TWC』
http://corpus.tsukuba.ac.jp

近代語コーパスにみる「結果」の用法

髙橋　圭子（東洋大学・非常勤講師）
東泉　裕子（東京学芸大学・非常勤講師）

Usage of *kekka* in Modern Japanese Corpora

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

要旨

現代日本語においては、「結果」などの名詞が、文頭または文中で副詞的・談話標識的に使われることがある。しかし、その成立に至る過程はまだ明らかにされていない。そこで、本発表では、「結果」の近代以前および近代語における用法を詳細に観察・記述する。そして、「結果」の用法を通時的に検討し、現代語の副詞的・談話標識的用法につながる過程について考察する。今回の調査・観察に基づき、次の３点を指摘する。(1)「結果」の語は、幕末から明治の初めごろ、現代のような意味で用いられるようになったと思われる。(2)「結果」の用法は、典型的名詞用法から名詞と副詞の中間用法を経て副詞用法に拡張したと考えられるが、そのプロセスは直線的ではなく複雑である。(3)「結果」を用いた副詞句・節の用法は、時の経過とともに、形式が簡略化され、且つ句読点の直後に用いられる例が増えてくる。この結果は、歴史語用論の知見とも軌を一にする。

１．はじめに

現代日本語では、「実際」「事実」「結果」「あげく」といった名詞に多様な用法が認められる（髙橋 2012、三枝 2013）。例えば、「結果」については『現代日本語書き言葉均衡コーパス（BCCWJ）』に次のような用法がある（東泉・髙橋 2013、髙橋・東泉 2013）[1]。

(1) 今日のヒアリングの結果ですが、・・・（OB1X_00075、特定目的・ベストセラー、山崎豊子『不毛地帯』、新潮社、1978 年）
(2) a. 今年後半のコンディションの安定ぶりには満足しているようだ。その結果として、ファーストステージの４ゴールに対して、セカンドステージでは一気に・・・（PM11_00739、出版・雑誌、『ストライカー』、学習研究社、2001 年）
b. 三十票ほど違っていることが分かり、同市選管があらためて点検した結果、ミスに気づいた。（PN5g_00006、出版・新聞、『西日本新聞』、西日本新聞社、2005 年）
(3) a.怒った音羽は仲舒に悪態の限りをつき、結果、 彼女といっしょに放り出されたのだ。(PB1n_00057、出版・書籍、伊藤遊『えんの松原』、福音館書店、2001 年）
b. この件に近い内容で、結局居住者の承諾を取らず無断で立ち入った案件がありました。結果、居住者は 300 万円相当の腕時計と指輪がなくなったと主張し警察を呼び

[1] 以下の用例の下線は筆者らによる。『BCCWJ』の用例には、順に、サンプル ID、レジスター、執筆者（書籍）、書名、出版者、出版年を記す。近代語コーパスの雑誌からの用例には、雑誌名、出版年・号、著者、題名を記す。

ました。（OC08_01946、特定目的・知恵袋、2005 年）

(1)は、格助詞やコピュラを伴う典型的な名詞としての用法である。(2)は、「結果」は名詞だがこれを含む句や節が副詞として機能している、名詞と副詞のいわば中間的な用法である。この用法は、「結果」に「として」等の複合助詞が後接するか否かにより、a.は名詞寄り、b.は副詞寄りの用法と考えられる。(3)は「結果」が単独で副詞として機能している用法で、a.は文中、b.は文頭に位置している。後者の用法は、「結果」の実質的意味が希薄化し、談話標識的用法に近づいていると見ることもできる。しかし、近代語コーパスから「結果」の用例を採取した東泉・髙橋（2013）、髙橋・東泉（2013）では、これらの用法の出現・成立に至る過程について詳しく観察することができなかった。

そこで、本稿では、まず「結果」の近代以前の用例を確認し、次いで近代以降の用例を詳細に観察・記述することにする。近代語の用例は、『明六雑誌コーパス』『近代女性雑誌コーパス』『太陽コーパス』から収集する。そして、「結果」の用法を通時的に検討し、現代語の副詞的・談話標識的用法につながる過程について考察する。

２．近代以前の「結果」

『日本国語大辞典　第二版』には、「結果」という語の初出例として(4)が挙げられている。また、同辞典には、(5)の中国元代の用例や、もとは仏教語であるとする補注[2]もある。

(4) 「人民の情と合和して、かかる結菓となりしなり」中村正直訳『自由之理』(1872)
(5) 「風餐水宿、甚日能安妥、問天天怎生結果」『琵琶記』

本節では、近代以前の「結果」の用例を確認する。まず、Japan Knowledge Lib により『新編日本古典文学全集（小学館）』の古典本文における「結果」を文字列検索したところ、計６例がヒットした。しかし、その意味はいずれも、表１に示す通り、現代語とも、また、仏教語とも異なるものであった。

表１　『新編日本古典文学全集』における「結果」の用例

ふりがな	意味	用例数	作品名	成立
かくなは	紐を結んだ形をした、油で揚げた菓子。また、その菓子の形のように縦横に切り結ぶこと。	1	「酒伝童子絵」	室町
おしかたづく	かたづける。始末をつける。	1	「しりうごと」	江戸
		4	『近世説美少年録』[3]	江戸

次に、19 世紀後半の辞典類から、「結果」に関する記述を調べ、次頁の表２にまとめた。西洋の思想や学問上の概念の訳出にあたり、中国で編纂された華英・英華辞典が幕末から明治の日本にも大きな影響を与えたことはつとに知られているが、Morrison（1815-23 刊）

[2] 【補注】「忠義水滸伝解 - 二一回」には、「結果モトハ仏語也。ソレヨリ転シテ万事物事ノシマイヲツケ片付ルコトニナルナリ」とある。

[3] 新編日本古典文学全集『近世説美少年録』（一、p.150）には、「結果　此仏語ナリ。シモフテノケルト云フガ如シ。シマイツケルコト（水滸伝字彙外集）」という注がある。

や Medhurst（1842-43 刊）の華英辞典の「結」の項には、「結果」という語は見当たらない。一方、Lobscheid の英華辞典の記述を見ると、「結果」という 1 つの語はこの頃未成立であり、「結」という動詞と「果／菓」という名詞としてそれぞれ用いられていたようである。その後、1870 年代から 80 年代にかけ、「結果」という語が成立し、現代につながったかと想像されるが、現段階では断定できない。更なる調査が必要である。

表 2　19 世紀後半の辞典における「結果」の記述

編著者	辞典名	発行年	記述
Lobscheid	『英華辞典』	1866-68	Result to result from, 故、帰結、結局、 to result in good,　結好菓、to result in evil 結悪菓 Result, consequence　果、果実 What are the results?　結何果乎
ヘボン	『和英語林集成』第三版	1886	KEKKWA　ケックヮ　結果 (*dekibai*) Effect, result, consequence.
高橋五郎	『漢英対照いろは辞典』	1888	けつくわ（名）結果。できばえ。み。なりはて。Result.
大槻文彦	『言海』	1889-91	けつ-くわ（名）　結果（和漢通用字）　事ノデキバエ。ナレノハテ。
物集高見	『日本大辞林』	1894	けつくわ　ナ（なことば＝名詞）。結果。でき。できざま。できばえ。

３．近代語コーパスにおける「結果」

３．１　「結果」の用法

本節では、19 世紀後半から 20 世紀初頭の近代語コーパスを用い、「結果」という語の様相を観察する。使用するコーパスは、国立国語研究所による『明六雑誌コーパス』『近代女性雑誌コーパス』『太陽コーパス』である。検索ツールには、「ひまわり」を用いる。

まず、それぞれのコーパスの概要と、「結果」の使用頻度を表 3 にまとめる。各コーパスの延べ語数は、『明六』は近藤（2012）、『女性雑誌』・『太陽』は近藤（2014）による。

表 3　近代語コーパスと「結果」の用例

コーパス			「結果」	
略称	収録年	延べ語数	用例数	度数/1 万語
明六	1874－75	180,605	1	0.06
女性雑誌	1894－95	586,665	102	1.74
	1909	406,889	53	1.30
	1925	272,325	30	1.10
女性雑誌　計		1,265,879	185	1.46
太陽	1895	2,031,346	622	3.06
	1901	1,929,238	899	4.66
	1909	1,725,992	948	5.49
	1917	1,619,638	812	5.01
	1925	1,456,055	580	3.98
太陽　計		8,762,269	3,861	4.41

『明六』は、(6)の一例のみである。これは、(4)とともに「結果」のごく初期の例であり、用法も典型的な名詞である。

(6)　其志向制御し易き人民を以て成立する處の國に於て苟も妄想空思行はれ愈信じて其迷を深するに至らば其結果又それ如何ぞや 人皆世事を顧みずして終に生業も廃するに至るべし（『明六雑誌』1874 年 6 号、森有礼訳「宗教」）

次に、『女性雑誌』の用例を用法別に分類し、表４にまとめた。

表４　『近代女性雑誌コーパス』における「結果」の用法[4]

発行年	用例数	用法	具体例	用例数	%
1894	41	名詞	名詞	35	85.4%
		中間的(名詞寄り)	N の結果として	2	14.6%
			その結果として	1	
			その N の結果として	1	
			V 結果として	1	
			N の結果により	1	
1895	61	名詞	名詞	53	86.9%
		中間的(名詞寄り)	N の結果として	3	6.6%
			その N の結果として	1	
		中間的(副詞寄り)	N の結果	2	4.9%
			V 結果	1	
		保留		1	
1909	53	名詞	名詞	40	75.5%
		中間的(名詞寄り)	この結果として	1	5.7%
			N の結果として	1	
			V 結果として	1	
		中間的(副詞寄り)	N の結果	2	18.9%
			その結果	1	
			V 結果	6	
			V の結果	1	
1925	30	名詞	名詞	13	43.3%
		中間的(名詞寄り)	その結果として	2	6.7%
		中間的(副詞寄り)	N の結果	3	50.0%
			その結果	4	
			V 結果	8	

[4] N＝名詞（句）、V＝動詞（句）。保留とした１例は、次のものである。「◎神戸外國婦人慈善會　昨年中の報告によれば前年の繰越金に合せ義捐金合計四百四十五弗四十七セン。（略）殘餘金二百十九弗七十七センありと。役員撰擧の結果　マクタビツシ夫人、ルーカス夫人、（略）アベル夫人等を委員とせり。」（『女学雑誌』1895 年 2 号、著者不明「片々」）下線部が見出しであれば名詞、本文の一部なら「N の結果」に分類される。同様の例は『太陽』にも数例見出せる。見出し的なものが本文化して副詞寄りの用法へ拡張していった可能性も考えられる。

表４からは、典型的名詞用法から、中間的用法（名詞寄り）、さらには中間的用法（副詞寄り）へ、用法が拡張しているさまが見てとれる。

『太陽』においても、用法拡張のさまは同様である。さらに、『太陽』には次のような例も見出せる。

(7) 彼等の刹那的な所はモダン・ガールの刹那的な所と本質的には全然の異質でありながら、結果的には交響するのではないか。（『太陽』1925 年 11 号、新居格「近代女性の社会的考察」
(8) 二十四日（日）◎公友中正合同成立　公友派は午後二時帝國ホテルに代議士會を開き合同問題を附議し結果全權を木村青木田中大木九鬼五氏に委任し五氏は合同を決して中正會に交渉し（『太陽』1917 年 2 号、著者不明「日誌」）
(9) 政本合同をやつて現内閣を倒すなどの荒藝は出來ない。結果野垂れ死ぬまでズラ～～グツタリで行くだらうと云ふのだ。（『太陽』1925 年 4 号、鬼谷庵「政界鬼語」）

(7)は、今回の調査範囲における「結果的」の初出である。(8)では、「結果」が文中で単独で副詞として用いられている。(9)は文頭の例である。なお、国立国語研究所コーパス開発センターによる『青空文庫パッケージ』からは、(10)の例も見出された。ここまでの調査の範囲では、「結果」の副詞用法の初出は(10)、その文頭の初出は(9)である。『日本国語大辞典　第二版』では(11)が初出となっているが、大きく遡るのは確かである。

(10) のでありますから、結果其物も亦三様の類型に（朝永三十郎「学究漫録」、初出「精神界　第二巻一一、一二號」1902 年 11、12 月）
(11) 女にはなぜ作曲家がいない？「そこで、女のものの考え方について非作曲家的なところを考えてみた。結果、女の考え方というのは、１＋１は２であるということだ」（藤本義一 1974～75「男の遠吠え」、『日本国語大辞典　第二版』「結果」の項）

３．２　「結果」の位置

「結果」の用法拡張により、現代語では文副詞的・談話標識的用法も見られるようになっている（髙橋 2012、東泉・髙橋 2013、髙橋・東泉 2013）。このような用法成立の端緒を探る試みの一つとして、本節では、「結果」の文中での位置を検討する。

『女性雑誌』コーパスでは、名詞用法および保留以外の「結果」は計 43 例、うち、その記事に句読点がともに用いられているのは 36 例である。その中で、句点もしくは読点のあとの用例をまとめたものが、次頁の表５である。

「結果」の用法拡張の過程において、初期の段階では「中間用法（名詞寄り）」に「N の結果において」「V 結果として」といった多様な形式が見られるが、やがて「結果」の先行部分は「その」に、後続部分は「として」に収斂されていき、さらに、先行・後続部分の脱落により単独で副詞として用いられるようになったと考えられる（髙橋・東泉 2013）。後続部分が脱落した「その結果」は、『太陽』コーパスでは 1895 年にすでに 19 例用いられており、『女性雑誌』コーパスでも 1909 年から出現する。一方、先行部分が脱落した形

式の「結果として」は両コーパスとも例がない[5]。用法拡張は直線的な過程ではなく、複雑な様相を呈しているが、表５からは、後続部分が脱落した「その結果」が、句読点とともに用いられているさまが見てとれる。

表５　『近代女性雑誌』コーパスにおける「結果」の用法と句読点

句読点との関係	X＝用法	年	用例数
句点＋X	この結果として	1909	1
	その結果	1909	1
	その結果	1925	1
句点＋接続詞＋X	そのNの結果として	1894	1
	その結果	1925	1
読点＋X	V結果として	1894	1
	Nの結果	1895	2
		1909	2
		1925	2
	V結果	1925	1
	その結果として	1925	1
	その結果	1925	2
合計			16

(12) そのまゝ解剖して見るやうなことになりました。その結果、肺病の初發病竈は、以前は肺尖だと云ふことになつて居たが（『婦人倶楽部』1925 年 6 号、宮原立太郎、「肺病並に結核病の話」）

(13) 人格ある人を選み取ることが出來ないからである。 而して其の結果、人を泣かせ又た自らも苦しむといふは、（『婦人倶楽部』1925 年 6 号、中島徳蔵、「異国人に恋されて悩む青年へ」）

(12)は「句点＋X」、(13)は「句点＋接続詞＋X」の例であるが、ともに読点が後続しており、文中での独立度は高い。現代語に見られる「結果」単独の副詞用法は、このような独立度の高い「その結果」から拡張が進行したのではないかと思われる。用例数の多い『太陽』コーパスにおける様相はより複雑であるが、基本的には同様の傾向が観察される。

３．３　現代語と異なる用法

近代語コーパスにおける「結果」には、現代語とは異なる用法がいくつか観察される。本節では、そのような用法の観察・記述を試みる。

３．３．１　「悪結果」など

近代語コーパスには、「結果」を含む漢語の複合語が観察されるが、現代語では使用されないものもある。

(14)　それと共に此の戰爭中の經濟状態は所謂變態である事も思はなければならぬ。戰爭

[5] 『太陽』には「結果において」４例、「結果的に」１例があるが、『女性雑誌』には見られない。

の直接結果として起つた各種の商工業は殊にさうである。（『太陽』1917 年 1 号、団琢磨、記者（文責）「事業界一夕話」）

(15) 元來人爲的の淘汰といへば、此の人爲的の發動の位置に立つ者は、自然淘汰の第一結果に由りて、勢力の比較に基因し、（『太陽』1895 年 3 号、千頭清臣「戦下側面的観察」）

(16) 此を誤診して腹部按摩をなし、若しくは自轉車を用ゐしめて、爲に惡結果として胃出血を起すが如きは、 大に少し。（『太陽』1901 年 9 号、吉田生（抄訳）「胃病と自転車」）

(17) それで獣の體や四肢にあんな隈取りをしたのが問題になつて居ましたが、 僕は不結果に終つたとは思ひません。（『太陽』1917 年 12 号、山村耕花「美術院日本画作家の感想」）

『女性雑誌』『太陽』コーパスを通じて「直接結果」は(14)の 1 例のみだが、「第一結果」は『太陽』に(15)を含め 2 例、「悪結果」は『女性雑誌』に 1 例、『太陽』には(16)を含め 20 例が観察される。「不結果」は『太陽』に(17)を含め 17 例、また、「大結果」も『太陽』に 3 例ある。(14)(15)は「中間的用法（名詞寄り）」に分類されるものだが、『BCCWJ』にはこのような漢語はなかった。また、(16)の「悪結果」は『BCCWJ』に 2 例あったが、いずれも名詞として使用されており、「悪結果として」という用法はなかった。(17)の「不結果」も『BCCWJ』にはなかった。一方、「好結果」は『女性雑誌』15 例、『太陽』77 例、『BCCWJ』45 例である。「好結果」「悪結果」は、Lobscheid の英華辞典の記述を想起させ、「結果」の近代語における用法として注目される。

３．３．２　「活用語連体形＋の＋結果」

活用語の連体形の主要な機能は、体言に連なりそれを修飾することであるが、近代語コーパスには、連体形と「結果」の間に「の」を介在させる用法が観察される。

(18) 佛蘭西の著者は獨逸に於ては保護を受けざるも獨逸の著者は佛蘭西に於いては保護を受くるの結果を來し、其間に公平を得ることが出來ないのである、（『太陽』1909 年 6 号、水野錬太郎「伯林に於ける著作権保護万国会議の状況」）

(19) また簡潔明瞭主義の君は、直に其本性を發露して、此間他の批評思惑等を考ふる餘裕なきの結果なるべきも、他の一方よりは確に不用意、修練不足なりとの謗りを免れぬ。（『太陽』1909 年 11 号、川尻琴湖「個人としての犬養木堂君」）

(20) また「デモクラツト」黨が「ボーア」問題によつて益英國に惡感情を抱きクルーゲルに渾身の同情を寄するの結果其感情が端なくも前例を楯として紐育市廳半旗を掲げざるの現象をなせしなり（『太陽』1901 年 5 号、森山吐虹「特別通信　英国女皇陛下の崩御と米国の態度」）

「の」を介在させるこのような用法は、(20) の「掲げざるの現象」にも見られるように「結果」以外の名詞にも見られる。また、(18)の波線部に見られるように、同一文中で「ない」「である」といった口語体の助動詞を用いながら、「活用語連体形＋の＋結果」の箇所では「受くる」といった文語体の活用形を用いている例も少なくない。更に綿密な観察が必要である。

３．３．３　「連用修飾＋動作性名詞[6]＋の＋結果」

「動作性名詞＋の＋結果」の場合にも、現代語とは異なる用法が観察される。

(21) 以前は肺尖だと云ふことになつて居たが、このＸ光線を應用の結果、主として肺門部が先に冒されると云ふことが分つて來て、（『婦人倶楽部』1925 年 6 号、宮原立太郎「肺病並に結核病の話」）

(22) 而して其の直接の衝に當るものは、 郵船會社ならざる可らず、 會社また多年之に意あり、 久しく調査の結果、其の外國航路に延長を希望する所は、 歐洲線、 濠洲線、 米國線とす、（『太陽』1895 年 11 月、著者不明「工業」）

現代語では、(21)は「このＸ光線を應用した結果」あるいは「このＸ光線の應用の結果」、(22)は「久しく調査した結果」あるいは「久しい調査の結果」とするところであろう。動詞と名詞の境界線上に位置する動作性名詞の用法は、現代語より近代語のほうが柔軟であったのかもしれない。これも、「結果」だけにとどまらない現象であると思われる。

４．まとめと課題

本稿では、漢語名詞「結果」について、まず近代以前の用法を調査し、次に近代語コーパスから収集した用例を用法別に分類した。そして、典型的な名詞用法から、中間的（名詞寄り）すなわち副詞句・節を構成する名詞としての用法へ、さらには中間的（副詞寄り）用法へと用法が拡張しているさまを観察した。また、近代語コーパスの「結果」の用法には現代語とは異なるものがあることも指摘した。

今回の調査・観察から、次のようなことが言えそうである。

- 「結果」の語は、幕末から明治の初めごろ、現代のような意味で用いられるようになったと考えられる。
- 「結果」の用法は、典型的名詞用法から名詞と副詞の中間用法を経て副詞用法に拡張したと考えられるが、そのプロセスは直線的ではなく複雑である。
- 「結果」を用いた副詞句・節の用法は、時の経過とともに、形式が簡略化され、且つ句読点の直後に用いられる例が増えてくる。このような例が、現代語の副詞的・談話標識的用法につながっていくと思われる。

以上のような仮説を立証するため、今後、歴史語用論（高田他 2011、金水他 2014、Onodera 2004、Onodera 2014 など）の理論的枠組みに基づき考察を深めるとともに、「結果」の近代から現代へかけての用法拡張の過程のより詳細な分析を行い、他の漢語名詞で「結果」と似たような用法拡張の過程を辿っていると考えられる語と比較しつつ、文頭または文中での副詞的・談話標識的用法への拡張の過程について更に検討することを目指したい。

[6] 動作性名詞とは、サ変動詞「する」を伴い、動詞として用いられる名詞。スル名詞、サ変名詞、動名詞などとも呼ばれる。

文　献

金水敏・高田博行・椎名美智編（2014）『歴史語用論の世界』ひつじ書房

近藤明日子（2012）「『明六雑誌コーパス』の語彙量」国立国語研究所共同研究報告 12-03『近代語コーパス設計のための文献言語研究成果報告書』pp.144-149.
(http://www.ninjal.ac.jp/corpus_center/cmj/doc/08kondo.pdf よりダウンロード可能)

近藤明日子（2014）「『近代女性雑誌コーパス』の小説会話部分に現れる一・二人称代名詞の計量的分析」国立国語研究所『第４回コーパス日本語学ワークショップ』、pp.135-144.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no4_papers/JCLWorkshop_No4_17.pdf よりダウンロード可能)

三枝令子（2013)「名詞から副詞、接続詞へ」『一橋大学国際教育センター紀要』4、pp.49-61.
(http://hdl.handle.net/10086/26706 よりダウンロード可能)

高田博行・椎名美智・小野寺典子編著（2011）『歴史語用論入門』大修館書店

髙橋圭子（2012）コーパスにみる名詞句の文副詞的用法」第 10 回対照言語行動学研究会
(http://www.ryu.titech.ac.jp/~nohara/taishogengokoudou/files/abst10/abst10_5takahashi.pdf)

髙橋圭子・東泉裕子（2013）「漢語名詞の副詞用法～『現代日本語書き言葉均衡コーパス』『太陽コーパス』を用いて～」国立国語研究所『第４回コーパス日本語学ワークショップ』、pp.195-202.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no4_papers/JCLWorkshop_No4_24.pdf よりダウンロード可能)

東泉裕子・髙橋圭子（2013）「『結果、こういうことが言えそうです』～コーパスにみる名詞の文副詞的用法～」国立国語研究所『第３回コーパス日本語学ワークショップ予稿集』、pp.91-96.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_12.pdf よりダウンロード可能)

Onodera, Noriko O. 2004. *Japanese Discourse Markers: Synchronic and Diachronic Discourse Analysis*. Amsterdam/Philadelphia: John Benjamins.

Onodera, Noriko O. 2014. Setting up a mental space: A function of discourse markers on the left periphery (LP). In *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*, Kate Beeching & Ulrich Detges (eds). Leiden: Brill.

辞　典

日本国語大辞典編集委員会（2000-2002）『日本国語大辞典　第二版』小学館

飛田良文、松井栄一、境田稔信編（1997-1998）『明治期国語辞書大系　普及版』大空社
第二巻『漢英対照いろは辞典』・第五巻『日本辞書言海』・第八巻『日本大辞林』

ヘボン、J. C.、松村明解説（1980）『和英語林集成　復刻縮刷版』講談社学術文庫

Lobscheid, W.、那須雅之（1996）『英華辞典　復刻版』東京美華書院

Medhurst, W. H.（1994）*Chinese and English Dictionary.* 復刻版　東京美華書院

Morrison, R.、帳西平、彭仁賢、呉志良編（2008）『華英辞典　影印版』鄭州：大象出版社

コーパス

国立国語研究所『現代日本語書き言葉均衡コーパス 中納言 1.1.0』（BCCWJ）
（https://chunagon.ninjal.ac.jp/）
国立国語研究所『近代女性雑誌コーパス』
（http://www.ninjal.ac.jp/corpus_center/cmj/woman-mag/）
国立国語研究所『明六雑誌コーパス』（http://www.ninjal.ac.jp/corpus_center/cmj/meiroku/）
国立国語研究所(2005)『太陽コーパス』（国語研究所資料集 15）　博文館新社

関連 URL

国立国語研究所コーパス開発センター　http://www.ninjal.ac.jp/corpus_center/

Japan Knowledge Lib　http://japanknowledge.com/library/

BCCWJにおける複合動詞後項の表記の実態

小椋秀樹（立命館大学文学部）

Orthographic Variation of the Latter Form of Compound Verbs as Seen in the BCCWJ

Hideki Ogura (College of Letters, Ritsumeikan University)

要旨

本稿の目的は、統語的複合動詞等の表記のうち、特に後項動詞の表記に着目し、そのゆれの実態を明らかにすることにある。

具体的には、『現代日本語書き言葉均衡コーパス』の図書館サブコーパス・書籍を調査資料とし、影山(1993)、姫野(1999)で統語的複合動詞を構成する後項動詞として挙げられている動詞を中心に 30 語を取り上げた。調査では、後項動詞の度数、出版年別、著者の生年代別といった観点から、漢字表記と平仮名表記の割合を明らかにした。

１．はじめに

現代日本語における語表記のゆれについては、小椋(2012)で、『現代日本語書き言葉均衡コーパス』(以下、BCCWJ とする。)のコアデータ(1)を資料として実態調査を行った。小椋(2012)では、和語に語表記のゆれが多く見られること、和語の中でも動詞が語表記のゆれの割合の高い語群であることを明らかにした。さらに動詞を対象に、どのようなゆれの類型があるのかも調査し、漢字と平仮名の対立によるゆれ(例：「合うー あう」)が最も多いことを指摘した。漢字と平仮名の対立によるゆれは、語表記にゆれのある単純動詞の約 7 割に、同じく複合動詞の約 8 割に見られる。複合動詞では、「とりくむー取り組む」「くりかえすーくり返すー繰り返す」などのように、語全体を平仮名表記にしたものや、前項動詞又は後項動詞を平仮名表記にしたものがあり、高い割合になっているものと考えられる。

小椋(2012)は、短単位を用いた調査であるため、上に述べた複合動詞とは語彙的複合動詞を指す。短単位では、最小単位二つから成る語彙的複合動詞は「／食べ歩く／」「／走り回る／」のように 1 短単位とするが、統語的複合動詞は「／降り／出す／」「／調べ／尽くす／」のように前項動詞、後項動詞をそれぞれ 1 短単位とする(2)。そのため小椋(2012)では、統語的複合動詞を構成する前項動詞、後項動詞を単純動詞として集計している。統語的複合動詞についても、語彙的複合動詞と同様に、漢字と平仮名の対立による語表記のゆれが多く見られるのか実態を調査する必要がある。

ところで、統語的複合動詞の表記については、表記の基準の面で興味深い点がある。それは、統語的複合動詞の後項動詞について、常用漢字で表記できても、一般的な表記として平仮名表記を挙げるものが見られることである。例えば、白石、野元、高田(2010)では、

(1) BCCWJ の設計等については、前川(2008)、山崎(2011)を参照。
(2) 短単位の認定基準のうち和語の動詞については、小椋、小磯、冨士池(2011:29)を参照。

次のような例がある。

かねる　**かねる**、兼ねる　　例 申し上げかねる、母の帰りを待ちかねる

すぎる　**……すぎる**、……過ぎる　　例 考えすぎる、食べすぎる、飲みすぎる、気にしすぎる

だす　**……だす**、……出す　　例 雪が降りだす、勉強をやりだす、笑いだす

白石、野元、高田(2010:2)は、語の書き表し方について、「常用漢字・現代仮名遣い・送り仮名の付け方を原則とする、一般的と考えられる書き表し方を、ゴシック体で示し」ている。上の例では、常用漢字「兼」「過」「出」による表記ではなく、平仮名表記を「一般的と考えられる書き表し方」として示していることになる。このような基準の妥当性を検証するという意味でも、統語的複合動詞の表記の実態、特に後項動詞を漢字表記にするか、仮名表記にするかについて調査を行う必要がある。

なお近年、情報機器の普及に伴う漢字の多用化傾向ということがよく言われる。語表記のゆれに経年変化が見られるのかといったことを調査することも重要である。

以上のことから、本稿では、BCCWJ を資料として、複合動詞のうち統語的複合動詞を取り上げ、特にその後項動詞に着目して、計量的な面から表記の実態を調査することとする。なお本稿では、小椋(2012)の結果を踏まえ、漢字と平仮名の対立によるゆれを取り上げることとする。したがって、送り仮名の対立によるゆれや仮名遣いの対立によるゆれなどは、取り上げない。

以下、2 節で今回の調査対象とする統語的複合動詞の後項動詞の範囲、調査資料とするレジスター、調査方法について述べた後、3 節で調査結果を報告する。最後に 4 節で本稿をまとめる。なお本稿では、語の表記を示す際には「出す」「だす」のように鍵括弧を付け、語を示す際には《ダス》のように二重山括弧を付ける。

２．調査対象・資料・方法

２．１　調査対象

語彙的複合動詞、統語的複合動詞という区分は、影山(1993)に述べられたものである。統語的複合動詞の特徴の一つとして、後項動詞が限定されるということが挙げられる。統語的複合動詞の後項動詞として、影山(1993:96)では、《カケル》《ダス》《ハジメル》など 27 語を挙げており、姫野(1999:19)では《カカル》《ハテル》《ソコネル》の 3 語を加えている。

短単位では、原則として、統語的複合動詞の後項動詞を付属要素(接尾的要素)として扱うこととし、「／降り／出す／」「／調べ／尽くす／」のように前項動詞と結合させずに単独で 1 短単位としている(3)。

ただし例外もあり、影山(1993)、姫野(1999)で示された 30 語全てを付属要素(接尾的要素)としているわけではない。BCCWJ における出現状況から造語力が高いと判定しなかったもの、単位認定が難しいと判断したものについては、付属要素(接尾的要素)として扱わなかった。これに該当するのは、《アウ》《アキル》《カケル》《カカル》《ノコス》《アヤマル》《ナオス》の 7 語である。

一方、影山(1993)、姫野(1999)で統語的複合動詞の後項動詞として挙げられていない語

(3)　付属要素、及び短単位の認定における付属要素の扱いについては、小椋、小磯、冨士池(2011:33、(38)-(55))を参照。

であっても、BCCWJ における出現状況、単位の統一性などから付属要素(接尾的要素)に加えたものもある。例えば、《サス》《ハテル》など 6 語である。

以上のようなことから、本稿で調査対象とする語は、影山(1993)、姫野(1999)と若干異なるところがある。本稿で調査対象としたのは、以下の 30 語である。

Ａ：影山(1993)、姫野(1999)に挙げられている語

アグネル(あぐねる)	エル(得る)	オエル(終える)	オクレル(遅れる)
オワル(終わる)	カネル(兼ねる)	カワス(交わす)	キル(切る)
コナス(熟す)	スギル(過ぎる)	ソコナウ(損なう)	ソコネル(損ねる)
ソビレル(そびれる)	ソンズル(損ずる)	ダス(出す)	ツクス(尽くす)
ツケル(付ける)	ツヅケル(続ける)	トオス(通す)	ナレル(慣れる)
ヌク(抜く)	ハジメル(始める)	マクル(捲る)	ワスレル(忘れる)

Ｂ：影山(1993)、姫野(1999)に挙げられていない語

オオセル(果せる)	サス(止す)	ツヅク(続く)	ハタス(果たす)
ハテル(果てる)	ワタル(渡る)		

本稿では、影山(1993)、姫野(1999)で統語的複合動詞の後項動詞として挙げられていない語も合わせて調査対象とすることから、以下、上記 30 語を後項動詞とする複合動詞を、統語的複合動詞等と呼ぶ。

２．２　資料

本稿では、統語的複合動詞等の後項動詞の表記について、経年変化の有無も調査する。

BCCWJ で経年変化を見ることのできるレジスターは、図書館・書籍(1986-2005 年)、特定目的・ベストセラー(1971-2005 年)、特定目的・白書、特定目的・国会会議録(以上、1976-2005 年)である。このうち、特定目的・白書、特定目的・国会会議録は、公文書であり、公文書の表記の基準に基づいて書かれている。そのため、一般社会における表記とずれが見られる可能性がある。また経年変化が見られたとしても、基準の改定によるものという可能性もある。したがって、今回の調査に適したレジスターとは言えない。特定目的・ベストセラーは、多くの人々に読まれたものとして、重要なレジスターではあるが、延べ語数が約 370 万語と、データ規模の面で問題がある。

以上のようなことから、本稿では、資料として図書館・書籍を用いることとした。図書館・書籍は、全体で延べ約 3000 万語と大規模なデータである。5 年ごとに区切った場合、最も語数の少ない 1986-1990 年で延べ約 480 万語、最も語数の多い 2001-2005 年で延べ約 880 万語と、経年変化を見るのにも十分な規模のデータである。

２．３　方法

用例の収集に当たっては、短単位データ 1.0.0 を対象に、『中納言』1.1.0 により、次の検索条件式で検索した(例として、《ハテル》の検索条件式を示す。)。

キー: 語彙素 = "果てる" AND 前方共起: 品詞 LIKE "動詞%" ON 1 WORDS FROM キー IN (registerName="図書館・書籍" AND core="false") WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

本稿は、漢字と平仮名の対立による語表記のゆれの実態を把握することを目的としてい

る。その際、注意しなければならないのは、児童向けの書籍の用例である。当然のことながら、児童向けの書籍の表記には、平仮名が多用されており、児童向けの書籍に多く使われている語は、平仮名表記の割合が高くなる可能性がある。本稿では社会一般の表記の実態を把握したいと考えており、児童向けの書籍の用例は除外するのが望ましい。

そこで、児童向けの書籍の用例を除外するために、BCCWJ の DVD に格納されている書誌情報データベースを利用した(4)。書誌情報データベースでは、書籍のジャンル情報の一つとして C コード(図書分類コード)が記載されている。C コードは 4 桁の数値で、左から 1 桁目が対象読者を示す「販売対象コード」である。児童向けの書籍は、この 1 桁目が「8」となっている。この C コードの情報を使って児童向けの書籍の用例を除外した。

具体的には、まず各語の検索結果を統合した後、関係データベースにインポートした。次にサンプル ID ベース書誌情報データ(Joinud_info.txt)をインポートし、サンプル ID で検索結果と関係付けて、検索結果に C コードの情報を付与し、児童向けの書籍の用例を除外できるようにした。

３．調査結果

３．１　後項動詞の表記のゆれ

本節では、後項動詞の表記の実態について、漢字表記、平仮名表記がそれぞれどの程度用いられているのか見ていくこととする。なお、先にも述べたように図書館・書籍は、出版年で 20 年の幅を持つレジスターであるが、ここでは一つの共時態と見なして(年代幅を無視して)調査する。

30 語の後項動詞について、漢字表記、平仮名表記がそれぞれどの程度用いられているかを、表 1 にまとめた。表 1 では、漢字表記、平仮名表記の度数と、それぞれの表記が語の度数全体に占める割合とを示した(「漢字」「平仮名」の各欄)。

表 1：後項動詞の表記

	漢字		平仮名		種別		漢字		平仮名		種別
アグネル	0	0.0%	40	100.0%	固定	ソンズル	23	100.0%	0	0.0%	固定
エル	1769	37.4%	2961	62.6%	ゆれ	ダス	1660	54.1%	1410	45.9%	ゆれ
オエル	404	87.4%	58	12.6%	独占	ツクス	365	54.4%	306	45.6%	ゆれ
オオセル	0	0.0%	47	100.0%	固定	ツケル	0	0.0%	13	100.0%	固定
オクレル	96	85.0%	17	15.0%	独占	ツヅク	55	65.5%	29	34.5%	ゆれ
オワル	568	89.6%	66	10.4%	独占	ツヅケル	2925	60.6%	1899	39.4%	ゆれ
カネル	36	2.9%	1207	97.1%	独占	トオス	168	73.0%	62	27.0%	ゆれ
カワス	73	64.6%	40	35.4%	ゆれ	ナレル	297	83.7%	58	16.3%	独占
キル	778	27.4%	2065	72.6%	ゆれ	ヌク	272	69.7%	118	30.3%	ゆれ
コナス	1	0.6%	180	99.4%	独占	ハジメル	3352	48.1%	3615	51.9%	ゆれ
サス	0	0.0%	12	100.0%	固定	ハタス	76	84.4%	14	15.6%	独占
スギル	463	23.0%	1552	77.0%	ゆれ	ハテル	301	71.2%	122	28.8%	ゆれ
ソコナウ	36	36.0%	64	64.0%	ゆれ	マクル	4	1.0%	380	99.0%	独占
ソコネル	21	44.7%	26	55.3%	ゆれ	ワスレル	150	98.0%	3	2.0%	独占
ソビレル	0	0.0%	39	100.0%	固定	ワタル	272	47.1%	306	52.9%	ゆれ

(4) 書誌情報データベースについては、丸山、中村(2011)を参照。

表 1 を見ると、《アグネル》《オオセル》など語表記にゆれの見られない語が 6 語あるが、それ以外は全て漢字と平仮名の対立による語表記のゆれが見られる。

ゆれの見られる 24 語を見ると、漢字表記、平仮名表記のいずれかに集中しているものがある。例えば、《カネル》《コナス》は平仮名表記が 9 割を超えている。そこで、ゆれの程度に応じた分類を試みることとする。まず、ゆれの見られない語を「固定」、一方の表記が 8 割以上を占めている語を「独占」、それ以外を「ゆれ」と呼ぶこととし、表 1 の「種別」欄に記載した(5)。

「固定」「独占」「ゆれ」について、語の度数との関連を見ていくこととする。表 2 は、語の度数別に、「固定」「独占」「ゆれ」がどの程度、出現するかをまとめたものである。この表では、度数を 100 以下、101-200 というふうに六つに区分し、それぞれの区分における「固定」「独占」「ゆれ」の語数を示した。

表 2：「固定」「独占」「ゆれ」と語の度数

	100以下	101-200	201-300	301-400	401-500	501以上
固定	6	0	0	0	0	0
独占	1	3	0	2	1	2
ゆれ	2	2	1	1	1	8

表 2 を見ると、「固定」は、度数 100 以下の低頻度層にのみ見られることが分かる。《アグネル》《オオセル》《サス》《ソビエル》《ツケル》は平仮名表記のみが、《ソンズル》は漢字表記のみが用いられている。《アグネル》《オオセル》《サス》《ソビエル》で平仮名表記のみが用いられるのは、これらの語が常用漢字(又は表内訓)で表記できないことによると考えられる。《ソンズル》で漢字表記のみが用いられるのは、この語が漢語サ変動詞だからであろう。ただ表 1 から分かるように、「固定」に属する語は度数 50 以下であり、BCCWJ では度数が低いため、たまたま一方の表記しか出現しなかった可能性もある。

「独占」は、度数 100 以下の低頻度層、度数 501 以上の高頻度層にも見られるが、度数 101-500 の中頻度層に 6 語あり、中頻度層を中心に分布している。「ゆれ」も低頻度層から高頻度層まで出現しているが、度数 501 以上が 8 語と最も多く、高頻度層を中心に分布している。よく使われる語ほど、表記にゆれが生じるということができよう。これは「固定」が低頻度層にのみ見られることと、ちょうど逆のことといえる。なお宮島(1997:100)でも、「度数順にみると、当然度数 1 のものにはゆれがなく、度数のたかいものほど、これがおおい」述べられている。本稿は、統語的複合動詞等の後項動詞に限定した調査ではあるが、宮島(1997)と同様の傾向が確認できた。

「独占」に属する語のうち、漢字表記に集中する語は《オエル》《オクレル》《オワル》《ナレル》《ハタス》《ワスレル》の 6 語、平仮名表記に集中する語は《カネル》《コナス》《マクル》3 語で、漢字表記に集中する語の方が多い。

また「ゆれ」に属する語のうち、漢字表記の割合が高い語は《カワス》《ダス》《ツク

(5) この 3 区分は、1956 年発行の雑誌 90 種を対象に、語表記のゆれを調査した宮島(1997)を参考にしたものである。ただし宮島(1997)は、「独占」を「特定の形式が 9 割以上をしめているもの」(p.103)としており、本稿と異なる。

ス》《ツヅク》《ツヅケル》《トオス》《ヌク》《ハテル》の 8 語、平仮名表記の割合が高い語は《エル》《キル》《スギル》《ソコナウ》《ソコネル》《ハジメル》《ワタル》の 7 語である。平仮名表記の割合が高い語のうち、《エル》《キル》《スギル》《ハジメル》《ワタル》の 5 語は、度数 500 以上の高頻度語である。ゆれの見られる高頻度の語は、平仮名で表記される傾向があるといえよう。

３．２　後項動詞の表記の変化

次に、後項動詞の表記の経年変化について見ていくこととする。ここでは、漢字表記、平仮名表記の割合に変化があるのか見ていく。

改定常用漢字表(2010 年 2 月、文化審議会答申)において、「情報機器による漢字使用が一般化し、社会生活で目にする漢字の量が確実に増えていると認められる」(p.(3))と述べられているように、近年、情報機器の普及という書記環境の変化に伴って、漢字が多用される傾向にあるといわれる。図書館・書籍は、1981 年の常用漢字表(内閣告示第 1 号・同訓令第 1 号)の実施から 2010 年の改定までの 29 年間のうち、20 年をカバーするレジスターである。本稿は、統語的複合動詞の後項動詞に限定した表記の実態調査ではあるが、情報機器が一般化し、書記環境が変化していく中で、表記がどのように変化していったのか(あるいは、しなかったのか)を明らかにするための調査としても位置付けられる。

まずは、出版年代別の経年変化を見ていく。図 1 に、出版年代別に後項動詞の漢字表記、仮名表記の割合を示した。この図では、出版年が 1986-1989 年のものを 80 年代後半、1990-1994 年のものを 90 年代前半、1995-1999 年のものを 90 年代後半、2000-2005 年のものを 2000 年代と、四つの年代に区分した。語別に漢字表記、平仮名表記の割合を算出するのではなく、各年代ごとに全ての語の漢字表記、平仮名表記の度数を合計した上で、漢字表記、平仮名表記の割合を算出している。なお、その際、度数 100 以下の低頻度の語群(ゆれの見られなかった《アグネル》《オオセル》《サス》《ソビエル》《ツケル》《ソンズル》と、《ソコネル》《ハタス》)を除外して集計した。

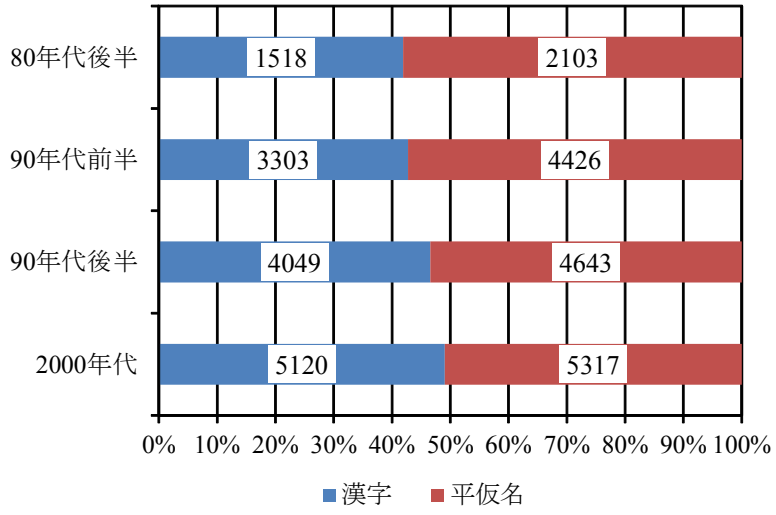


図 1：後項動詞の表記の変化(出版年代)

図 1 を見ると、年代が下るにしたがって、徐々に漢字表記の割合が高くなっていることが分かる。図 1 の用例数を基に漢字表記の割合を求めると、80 年代後半は 41.9%であったが、90 年代前半は 42.7%、90 年代後半は 46.6%、2000 年代は 49.1%となる。2000 年代には、漢字表記が平仮名表記と拮抗するまでになっているのである。

次に、著者の生年代別の経年変化を見ていく。図 2 に、著者の生年代別に後項動詞の漢字表記、仮名表記の割合を示した。図 1 と同様に、各年代ごとに全ての語の度数を合計した上で、漢字表記、平仮名表記の割合を算出しているが、その際、著者が複数のサンプルの用例は集計の対象外としている。また、語の度数の合計が 300 以上の年代を図に示した。

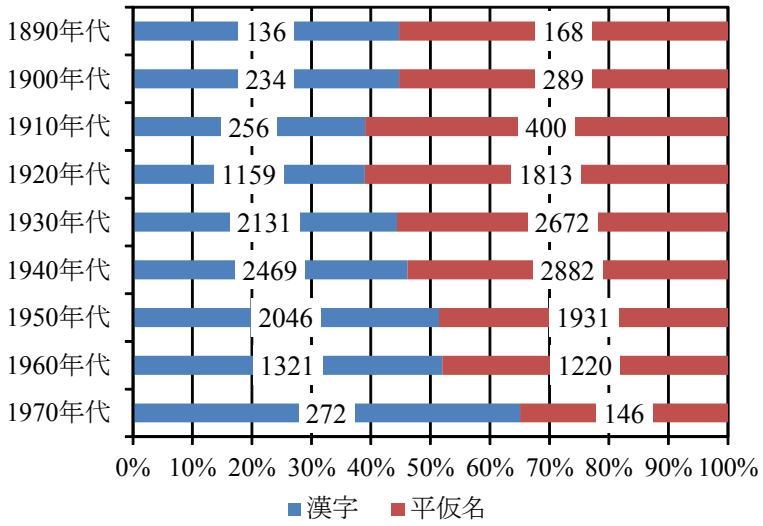


図 2：後項動詞の表記の変化(著者生年代)

図 2 を見ると、1890 年代生まれのグループから 1910 年代生まれのグループまで、漢字表記の割合が低下するが、その後は、1970 年代生まれのグループまで漢字表記の割合が増加傾向にある。1950 年代以降のグループは、1950 年代生まれのグループが 51.4%、1960 年代生まれのグループが 52.0%、1970 年代生まれのグループは 65.1%と、いずれも漢字表記が過半数を占めている。ただし 1970 年代生まれのグループは、全体の度数が 418 と少ないため、慎重に見る必要はあろう。とは言え、度数が 2000 を超える 1950 年代、1960 年代生まれのグループで漢字表記の割合が 5 割を超えており、増加傾向にあることは指摘できよう。

4．終わりに

本稿では、BCCWJ の図書館・書籍を資料として、統語的複合動詞等の後項動詞の語表記のゆれ、特に漢字と仮名の対立による語表記のゆれについて実態調査を行った。その結果、次のことが明らかとなった。

(1) ゆれの程度に応じて「固定」「独占」「ゆれ」の 3 区分で見た場合、「固定」は低頻度層に、「独占」は中頻度層を中心に、「ゆれ」は高頻度層を中心に分布している。高頻度語ほど、表記にゆれが生じるといえる。

(2)「独占」に属する語は、漢字表記に集中する語が多い(9 語中 6 語)。「ゆれ」に属する語(15 語)で、平仮名表記の割合が高い語は 7 語あるが、そのうち 5 語は度数

500 以上の高頻度語である。

(3) 出版年別の経年調査、著者の生年代別の経年調査のいずれにおいても、年代が下るにしたがって、漢字表記の割合が高くなる。

本稿で取り上げた 30 語のうち、「ゆれ」に属する語は 16 語ある。今回は、後項動詞の表記のみの調査ではあったが、統語的複合動詞は、語彙的複合動詞と同様、語表記にゆれが多いと考えられる。

また 30 語中 24 語は常用漢字で表記できる語であるが、そのうちの 23 語に漢字と平仮名の対立によるゆれが見られる。姫野(1999:28)では統語的複合動詞の後項動詞を「接尾辞的複合動詞」と呼ぶが、接尾辞的な性格から、平仮名表記が選択されやすくなっていると考えられる。

しかしながら、出版年代別、著者生年代別の調査結果を見ると、漢字表記が増加傾向にある。今後、語別に経年変化を見ていくことも必要であろう。また、漢字表記の増加傾向が情報機器の普及によるものかを検証するために、特定目的・知恵袋、特定目的・ブログを対象とした調査も必要となろう。今後の課題としたい。

なお、1 節で取り上げた《カネル》《スギル》《ダス》の表記の基準については、本稿の結果からいえば、《カネル》《スギル》は平仮名表記を一般的な表記とするのは妥当である。しかし《ダス》については、漢字表記が約 54%、平仮名表記が約 46%でゆれており、本稿の結果からは、どちらが一般的か決めることは難しい。表記の基準を考えるためにも、本稿のような実態調査を行っていくことが必要である。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト(基幹型)「コーパス日本語学の創成」(リーダー：前川喜久雄)、同「多角的アプローチによる現代日本語の動態の解明」(リーダー：相澤正夫)、JSPS 科研費「大規模コーパスに基づく現代語表記のゆれの実態解明」(代表者：小椋秀樹)による補助を得た。

参 考 文 献

小椋秀樹(2012)「コーパスに基づく現代語表記のゆれの調査 ― BCCWJ コアデータを資料として ―」、『第 1 回コーパス日本語学ワークショップ予稿集』、pp.321-328.

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

影山太郎(1993)『文法と語形成』、ひつじ書房.

白石大二編、野元菊雄新版監修、高田智和改訂新版監修(2010)『例解辞典 改訂新版』、ぎょうせい.

姫野昌子(1999)『複合動詞の構造と意味用法』、ひつじ書房.

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.

丸山岳彦、中村武範(2011)「第 8 章 書誌情報データベース」、国立国語研究所コーパス開発センター『『現代日本語書き言葉均衡コーパス』利用の手引き』第 1.0 版、pp.117-152.

宮島達夫(1997)「雑誌九十種表記表の統計」、『日本語科学』1、pp.92-103.

山崎誠(2011)「第 2 章 『現代日本語書き言葉均衡コーパス』の設計」、国立国語研究所コーパス開発センター『『現代日本語書き言葉均衡コーパス』利用の手引き』第 1.0 版、pp.113-20.

多様な会話コーパスを対象とした発話連鎖ラベリングの試み

森 大毅 （宇都宮大学大学院工学研究科）
森本 郁代 （関西学院大学法学部）
大場 美和子 （昭和女子大学人間文化学部）
吉田 悦子 （三重大学人文学部）
伝 康晴 （国立国語研究所言語資源研究系／千葉大学文学部）

On Annotating Sequence Structures for Conversation Corpora of Various Kinds

Hiroki Mori (Utsunomiya University)
Ikuyo Morimoto (Kwansei Gakuin University)
Miwako Ohba (Showa Women's University)
Etsuko Yoshida (Mie University)
Yasuharu Den (National Institute for Japanese Language and Linguistics / Chiba University)

要旨

性質が様々に異なる会話コーパスに対する基本情報の統一的な付与作業は、単一コーパスの場合にはない困難が予想される一方で、会話を形成する一連の発話が有する構造の本質に迫る有力な手段ともなり得る。本研究は、会話コーパスに付与すべき基本情報の中でも特に発話間の局所的構造に着目し、隣接ペア (Schegloff and Sacks 1973) の考え方を土台として、条件的関連性を有する連鎖の体系的記述法を確立することを目的としている。今回の試みでは、多様なコーパスを対象とした発話連鎖記述の問題点洗い出しのため、宇都宮大学パラ言語情報研究向け音声対話データベース、日本語話し言葉コーパス、千葉大 3 人会話コーパスに対し、4 名の分析者が第 1 部分／第 2 部分／第 3 部分および参照先を独立にラベリングした結果の一致性を評価するとともに、不一致の原因となる問題点を整理した。

1　はじめに

国立国語研究所共同研究プロジェクト「多様な様式を網羅した会話コーパスの共有化」(研究リーダー: 伝 康晴) は、性質が様々に異なる既存の会話コーパスを対象に共通の基本情報を付与し、相互利用を可能とすることを目的としている。インタラクションとしての会話研究が必要とする会話コーパスに付与すべき基本情報には、言語音・非言語音・非流暢性・音調など個々の発話が単独で有する特徴だけでなく、一連の会話の中での発話、すなわち談話に関する情報も含まれる。

談話に関する基本情報としては、連続した発話間の関係が最も重要なものの 1 つに挙げられる。このような情報を記述するための代表的な枠組に、隣接ペア (Schegloff and Sacks 1973) がある。隣接ペアは、Sacks らが始めた「会話分析」の流儀に基づく行為連鎖記述の最も基本的な構成要素である。会話分析の目的は、会話が動的に構成されていくために参与者が利用している社会的手続きを明らかにすることであり、分析においては徹底した観察的態度を取る。このため、雑談や多人数会話など一見秩序立っておらず体系的な談話行為の認定が難しい会話に対しても、隣接ペアの記述は談話における発話を組織化して行くための有

力な分析方法となる可能性がある。

本研究では、隣接ペアの考え方を土台として、条件的関連性を有する連鎖の体系的記述法を確立することを目的としている。この目的に迫るため、まずは多様なコーパスを対象として発話連鎖のラベリングを試み、その問題点を洗い出す。本稿では、複数の分析者による複数コーパスに対するラベリング作業の結果を比較し、不一致の原因となる問題点を整理することで、局所的連鎖の記述法の確立に向けた知見を提供する。

2　隣接ペア

隣接ペアとは、[質問]-[応答] や [依頼]-[受諾／拒否] のように、ある参与者が発話を行い、直後にその受け手が適切な応答となり得る発話を行うことで形成される、会話のやり取りの最小単位である (Schegloff and Sacks 1973)。ここで、先行する発話を隣接ペア第 1 部分 (first-pair part)、後続する発話を隣接ペア第 2 部分 (second-pair part) と呼ぶ。

隣接ペアの第 1 部分は、直後の発話に対して条件的関連性を課す。これにより、応答としての第 2 部分の生起が義務化する。例 1 は、今回の検討で対象とした日本語話し言葉コーパスのギャラタスク対話からの例である。

例 1 (CSJ D02F0015)

A1: 白黒ですか写真　┐1
B2: 白黒です　←┘2

例中、矢印で結ばれた発話対は隣接ペアを表し、矢印脇に 1, 2 と記された発話はそれぞれ第 1 部分、第 2 部分を表す (以下同様)。

A の発話は直後の B の発話に対して条件的関連性を課す。すなわち、「はい」「いいえ」あるいは「白黒です」「白黒じゃないです」場合によっては「わかりません」など、A の Yes/No 質問への返答としてふさわしい発話を期待する。B の発話が Yes/No 質問に対応しない応答 (例えば「ふーん」) だった場合には、A の質問がそもそも正しく伝わっていなかったなどの問題があったことが示唆される。また、仮に応答が後続しなかった場合はやはり聞き取り上の問題があったことや即答できないことなどが示唆される。いずれの場合も A は今コミュニケーションに問題が発生したことを認識し、恐らくその問題を解決するための例外処理 (質問を繰り返す、または聞いているかを確認するなど) を発動しようとするであろう。

第 1 部分と第 2 部分の間には、いくつかの発話が挿入される場合もある。1 つはいわゆる挿入連鎖 (Schegloff 1972, 2007) で、別の隣接ペアが入れ子になっている場合である。そのほかには、例 2 のように先行発話の一部分の繰り返しやフィラーの挿入によって第 2 部分が一時的に保留される場合がある。

例 2 (CSJ D02F0015)

A1: で上の方(0.484)トップを<H>行くのは<H>(D ン)(0.475)何でしょう　┐1
B2: 上の方　　│
B3: (F うーん)　　│
B4: トップは<H>(0.178)秋野暢子な気がするんですけど私は　←┘2

A1 に続く B の発話が、A1 によって課された条件的関連性により期待される発話ではないことによって、A は B が応答に一時的困難を覚えていることに気づくことができる。

連鎖は、隣接ペア第 2 部分で完結する場合と、さらに関連する発話によって継続する場合 (後方拡張) がある (Schegloff 2007)。後方拡張の一種に、第 2 部分に対する了解や評価によっ

て連鎖を終わらせる種類の発話があり、連鎖終結の第 3 部分 (sequence-closing third) と呼ばれる。

例 3 (CSJ D02F0015)

A1: 格好とかは何か背広とか着てますか　1
B2: 背広とか両方共背広　2
A3: 両方共背広か　3

3　ラベリング実験

3.1　対象コーパス

今回のラベリング実験で対象としたのは以下の 3 種類の対話コーパスである。

- 宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB) (Mori et al. 2011)

  4 コマまんが並べ換えタスク遂行時の 2 人対話。前半部分は主にそれぞれが持っている 2 コマの内容説明からなり、後半部分は提案に対する議論と意見の一本化の過程からなる。

- 日本語話し言葉コーパス (CSJ) (前川 2004)

  インタビュー対話とギャラタスク対話。インタビュー対話は、それに先立つ模擬講演に対してインタビュアーが質問する形式の 2 人対話。ギャラタスク対話は、与えられた芸能人のリストから、想像されるギャラの高い順に並べ換えさせるタスクを遂行時の 2 人対話。

- 千葉大学 3 人会話コーパス (Chiba3Party) (Den and Enomoto 2007)

  同性の友人 3 人による雑談。非常に自発性が高く、会話が秩序立っていない。

3.2　ラベリング作業 (1 回目)

筆者のうち 4 名がそれぞれ独立に、対象コーパスのサブセット (UUDB 3 対話、CSJ インタビュー 2 対話、CSJ ギャラタスク 1 対話、Chiba3Party 2 対話) に対し、第 1 部分・第 2 部分・第 3 部分と認められる発話にタグを付与した。また、第 2 部分・第 3 部分については参照先の発話をタグ付けした。

隣接ペアは社会的に規定された特定の行為の連鎖である。当初の計画は、対象コーパスに対し隣接ペアをラベリングするというものであった。ところが、雑談なども含む多様な会話においては、明らかに社会的に規定されているというほどではなくとも先行発話が後続発話に何らかの応答が後続することを投射すると考えられる種類の連鎖が多く観察される。実際、今回研究の対象とした多様な会話においては、規定された行為の連鎖と明確に認定できるものは多くないことがわかった。

例 4 は UUDB からの例である。A は、作中の「近所のご隠居」が「チョーさん」に話しかけても聞こえない理由を「ご隠居」が死んでいるからではないかと推定したが、B は「近所のご隠居」なのに死んでいるということが納得できない。そこで、A は「ご隠居」が自分は死んでいたのだと自分で言っていることを、台詞を再び朗読することで説明しようとしている。

例 4 (UUDB C024)

A1: そうじゃわしは死んどったんじゃ
B2: ふんふん
A3: いつまでもこうしているわけにはいかんなあ
B4: って言ってんだ ┐ 1?
A5: うん ←┘ 2?
B6: うん ×3?

B4 の発話は、A1, A3 が「ご隠居」の台詞であることを確認する意図で産出されたとも解釈できるが、言語形式上は疑問形ではない (cf.「って言ってんの?」)。[質問]-[応答] などの隣接ペアの類型に厳密に当てはめれば B4-A5 は隣接ペアではないことになる。先行発話が後続発話の型を完全に規定するというほどではないもののある程度の条件的関連性が観察される B4-A5 のようなケースは、自由な形式の会話では非常に多い。1 回目のラベリングの結果を持ち寄り、作業結果の不一致箇所について検討した結果、分析者による違いが最も顕著だったのが連鎖の記述対象の適用範囲に関する認識の違いに起因するものであった。

本研究における記述対象を厳密に隣接ペアに限定することは研究の適用範囲を非常に狭くしてしまう可能性がある。そこで、本研究の今後の発話連鎖ラベリングにおいては、条件的関連性は広く解釈すべきであることを分析者間で確認した。

このほか、発話連鎖の体系的記述法確立のために解決が必要ないくつかの問題点が指摘されたが、これらについては **4** で詳細に述べる。

3.3 ラベリング作業 (2 回目)

ラベリングの対象とする連鎖の範囲、およびいくつかの作業上の事項に関する分析者間の意識合わせを行った後、第 1 回目と同じ分析者 4 名が CSJ インタビュー 1 対話、CSJ ギャラタスク 1 対話、Chiba3Party 1 対話に対する 1 回目作業結果の見直しを行った。

4 名の 2 回目の作業結果を集計し、各発話に付与されたタグ (第 1 部分／第 2 部分／第 3 部分／それ以外) のタグの一致数を調べた。結果を図 1 に示す。4 名の一致率は、CSJ インタビュー対話およびギャラタスク対話では 60% 程度、Chiba3Party では 40%程度であった。全体的には、第 1 部分が質問のような隣接ペアの類型でなく、情報伝達や示唆、提案、希望を含むような発話の場合に一致率が低下する傾向が見られた。すなわち、条件的関連性の程度がそれほど高くない場合に、それを第 1 部分と考えるか否かが分析者により異なっていた。雑談のような自由度の高い会話では、この種の対応する義務的な応答が明確に規定されない談話行為の頻度が高いと考えられている (高梨 2012)。第 1 部分の宛先が曖昧な 3 人会話で

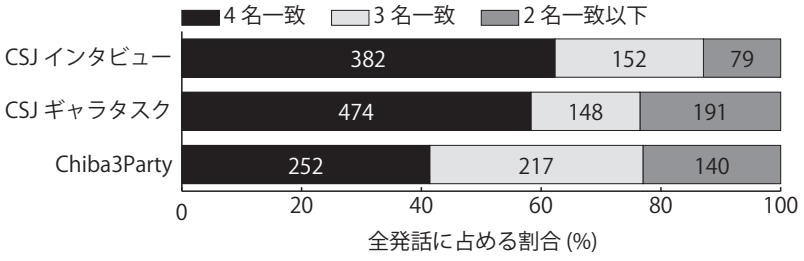


図 1: タグ (第 1 部分／第 2 部分／第 3 部分／それ以外) 一致数および一致率

は、第 1 部分の認定がさらに困難になるだけでなく、第 2 部分・第 3 部分の認定も難しくなる (伝 2009)。このことは、Chiba3Party に対する低い一致率に現れている。

4　考察

ラベリング作業を通して、今回対象としたコーパスに含まれる会話に対する発話連鎖ラベリングにおいて考慮すべき問題点がいくつか浮き彫りとなった。以下ではそれらの詳細を述べる。

4.1　相槌、同意表現

聞き手の相槌の生起が話し手の発話のイントネーションや統語構造に関連していることを示す数多くの証拠がある (Koiso et al. 1998; 岡登 1999)。その中には、話し手がある程度能動的に聞き手の相槌を招いていると解される表現が含まれる (Ward 1996)。例 5 の発話 A1 は、CSJ インタビュー対話において、最近初めて中国に行くまで海外に行ったことがないインタビュイーにその理由を尋ねるためのやり取りに先行して、多くの人は若いうちから海外に行っているものだという前提を共有するためのインタビュアーの発話である。

例 5 (CSJ D01F0023)

A1: わりとこう最近(0.192)<FV>若いうちにいっぱい
　[色んなとこ　　]行くじゃないですか=　　┐1
B2: [(F うん)(F うん)]　　　　　　　　　　┘
B3: =(F はい)(F はい)　　　　　　　　　　←2

B3 は、「じゃないですか」という反語表現に対する単純な相槌とも解釈できるが、B2 の相槌とはわざわざ表現を変えているところから、もう少し積極的に A1 に同意したものとも考えられる。いずれにしても、A1 は B の相槌ないし同意を要求する表現と言え、条件的関連性の観点から A1 と B3 の連鎖が認定できる。

例 5 の「はい」に見られるように、言語行為としての応答 (e.g. Yes/No 質問への返答) と相槌が同じ表現を持つことは、発話連鎖の認定をしばしば難しくする。例 6 は CSJ ギャラタスク対話において、2 人の男性のどちらが「蟹瀬」でどちらが「神太郎」かについて議論した結果を受けて、B が決定の提案をする場面である。

例 6 (CSJ D02F0015)

B1: じゃそれを<H>蟹瀬に(0.193)して<H>　　┐1
　(2.530)　　　　　　　　　　　　　　　　┘
A2: (F はい)　　　　　　　　　　　　　←2?
B3: でもう一人を(0.115)神太郎にしましょう　┐1
　　　　　　　　　　　　　　　　　　　　↓

B1 と A2 の間には 2 秒以上にわたる長いポーズがある。B は B1 の提案に対する A の承認を待っていたと考えられるため、B1 は第 1 部分と認定できる。一方、A は「はい」という発話によって提案を承認したとも考えられるし、単に B の発話継続を促すために相槌を打ったとも考えられる。典型的な相槌である「うん」「はい」「ええ」などの表現が肯定応答としての「うん」「はい」「ええ」などの表現と区別が付かないという問題は、日本語に特有の事情と言えそうである。この問題は、相槌の認定、つまりは発話単位の認定という最も基本的な問題と表裏一体であることを指摘しておく。

終助詞「ね」には、例 5 A1 の「か」と同様、相槌や同意を求めるニュアンスがあるが、応答が義務とまでは言えないケースが多い。例 7 は UUDB からの例で、A が自分の手持ちのコマを説明している。

例 7 (UUDB C004)

A1: でねなんか、(D ひと)、一方が=、眼鏡を外して=、なんかハンカチで拭こうとしてるのね ┐1?
B2: =うん =うん ┘
B3: うん ←2?

もし B3 の発話がなければ応答不在の印象を受けることは確かだが、B がそれまで頻繁に相槌を打っていたからそう感じられるに過ぎない可能性もある。従来の会話分析では、A1-B3 のような対が隣接ペアと認定されることはなかった。ここでは、相槌や同意表現に関わる条件的関連性は相対的であり、言語形式やイントネーションからその有無を明確に認定することが難しいことを指摘するにとどめる。

4.2 修復

例 8 は、C の発話を B が聞き返している例であり、他者開始修復 (Schegloff et al. 1977) の典型例である。

例 8 (Chiba3Party 0232)

C1: それってどこで: (0.336)(D キ)きいてんの ┐1
B2: うん [修復開始] ┐1
C3: どこできいてるの [修復実行] ←2
B4: だからマルチスレッドよ ←2

他者開始修復を含む連鎖は、基本的には例 8 に示すような挿入連鎖として記述することができる。しかし、例 8 のように第 1 部分の聞き取りに問題があったことが原因の修復の場合、言い直した本来の第 1 部分が修復時には第 2 部分となってしまうため、修復開始-修復実行の対は隣接ペアの類型とはかなり違って見える。

例 9 は A の言い直しの例であり、自己開始修復 (Schegloff et al. 1977) の典型例である。

例 9 (Chiba3Party 0232)

A1: といえば何どうですかC さん 1
A2: あ B さん<笑> [修復開始][修復実行] ┐1
B3: またかよ<笑> │
B4: またそのふりかよ │
B5: だからさっき言ったじゃん ←2

第 1 部分に対する自己開始修復の場合には、修復後の発話が新しい第 1 部分となる。その場合、第 2 部分の参照先を本来の第 1 部分にするか新しい第 1 部分にするかは分析者間で予め合意しておくべき事項である。例 9 では修復後の A2 を参照先としている。この場合、A1 は孤立した第 1 部分となる。

4.3 連鎖の単位

Schegloff and Sacks (1973) では、隣接ペアは発話の連鎖の一種として定義されている。また、隣接ペアのラベルは発話を単位とした書き起こしに付与されることが多い。しかしながら、今回のラベリング作業を通して、発話を単位とした連鎖のラベリングだけでは不十分な事例が見出された。

例 10 は、C の提案に対する B の非同意応答が「そうなのかな」という疑問形の形式を取っているため、それにさらに C が応答した例である。

例 10 (Chiba3Party 0232)

C1: こう(0.278)長い間でこう消化されるから:(0.407)腹持ちがいいとか
　　そういうのなんじゃない　┐1
B2: そうなのかな:　　←┘2　┐1
C3: いや わかんないけど　　←┘2

このやり取りは 2 つの連鎖が続けて起こったものと解釈できるが、B2 は先の連鎖の第 2 部分であり、同時に後の連鎖の第 1 部分となっている。各発話に対し第 1 部分／第 2 部分／第 3 部分／それ以外のラベルを排他的に付与するような連鎖ラベリングでは、B2 のような事例をうまく取り扱うことができない。

例 11 は、インタビュアー A の質問への答えが非常に長いケースである。

例 11 (CSJ D01F0023 途中の相槌を省略)

A1: それは実際どういう具体的にはどういうことを(0.498)色々やってあげるんですか　┐1
B2: (F あのねー)最初の方は<H>(0.34)(F あのー)(.)一緒に(D エ)(0.109)遊びに行ったりとか(.)(F あの)(0.322)浅草に行ったりとかね<笑>両国に行ったりとかねそういう(0.25)そうそうそうそうちょっと(0.238)日本ぽいところに(0.23)遊びに行くとか(0.337)そういうのもやって　←2
B3: で後は(0.382)(F あのー)その子は　←2?
B4: 日本に来て<H>　←2?
B5: (F まー)(0.265)(F あの)うち<H>私の行ってるところは日本語<H>(0.454)<FV>(D ガッ)(0.273)(D ン)日本語学っていうのをやってるところなので(0.303)(F まー)(F あのー)日本に来る前ももう日本語ぺらぺらだし<H>　←2?
B6: (F うん)日本にも来て<H>日本語学校にも行ってって感じで　←2?
B7: その(.)(D ン)(0.334)言葉に関しては(.)あんまり(0.414)トラブルが(F まー)なかったんで　←2?
B8: で後は大学院目指してるとかっていうのがあって　←2?
B9: それで(0.247)(F ん)(F あの)勉強を(.)主に(0.348)受験勉強の(0.281)家庭教師みたいな　←2?
B10: だから(0.142)週に一回とか二回とか(0.192)(F あの)勉強する時間を作って　←2?
B11: で(0.287)(F まー)二人で一緒に(0.345)受験勉強をしたと　←2?
A12: 大変ですね<H>　×3

発話を連鎖の単位としたままこのようなケースに対してラベリングするための方針としては、A1 の質問に関連がある B2 から B11 までを全て第 2 部分とする方法、A1 への直接の応答は B2 だけであると考え B2 だけを第 2 部分とする方法などが考えられる。しかしながら、B2 から B11 までは全体で 1 つの複合的単位による応答と考えた方がすっきりする。複合的単位による働きかけと共に、この問題は発話連鎖よりも大域的な構造のアノテーションと一体で考慮する必要がある。

5 おわりに

性質が様々に異なる複数の会話コーパスに対し、隣接ペアの考え方を土台として条件的関連性を有する連鎖をラベリングする試みについて述べた。4 名の分析者が独立にラベリングした結果、タグの 4 名一致率は 40%から 60%程度であった。

今回の試みを通して明確になった今後の指針の 1 つは、局所的発話連鎖以外のレベルにおけるラベリング作業と協調的・統一的に実施することである。ひとつの例は、**4.1** で述べた発話単位および相槌の認定作業である。これらのラベルが発話連鎖の認定と整合的に与えられていれば、発話連鎖のタグの一致度は間違いなく向上するし、逆に発話連鎖を意識することにより、理論的にもより強固で統一的な発話単位および相槌の認定が可能と考える。もうひとつの例は、修復 (**4.2**)、後方拡張、前方拡張、複合的単位を含む局所的連鎖の拡張 (Schegloff 2007; 伝 2009) に関するラベリング作業である。これらのうちいくつかは本研究の一部として進める必要があり、またいくつかは今後の新しい研究課題となるだろう。

今後は、これまで得られた知見を基に発話連鎖ラベリングのマニュアルを整備し、種々の会話コーパスへのラベリングを進める予定である。

参考文献

岡登洋平、加藤佳司、山本幹雄、板橋秀一 (1999) 「韻律情報を用いた相槌の挿入」情報処理学会論文誌 40:2, pp. 469–478.

高梨克也 (2012) 「発話意図のアノテーションは可能か?: 談話行為記述に学ぶ」日本音響学会 2012 年秋季研究発表会, pp. 247–250.

伝康晴 (2009) 「隣接ペア」坊農真弓 (編)、高梨克也 (編) 『多人数インタラクションの分析手法』オーム社, pp. 82–94.

前川喜久雄 (2004) 「『日本語話し言葉コーパス』の概要」日本語科学、15, pp. 111–133.

Den, Y. and Enomoto, M. (2007) “A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation” in Nishida, T. (Ed.), *Conversational Informatics: An Engineering Approach*, Hoboken, NJ: John Wiley & Sons, pp. 307–330.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998) “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs” *Language and Speech* 41:3–4, pp. 295–321.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011) “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics” *Speech Communication* 53:1, pp. 36–50.

Schegloff, E. A. (1972) “Notes on a conversational practice: Formulating place” in Sudnow, D. (Ed.), *Studies in Social Interaction*, New York: Free Press, pp. 75–119.

Schegloff, E. A. and Sacks, H. (1973) “Opening up closings” *Semiotica* 8:4, pp. 289–327.

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977) “The preference for self-correction in the organization of repair in conversation” *Language* 53, pp. 361–382.

Schegloff, E. A. (2007) *Sequence Organization in Interaction: Vol.1: A Primer in Conversation Analysis*, Cambridge University Press.

Ward, N. (1996) “In Japanese a low pitch means “Back-channel feedback please” ” 情報処理学会研究報告, 音声言語情報処理, 96:55, pp. 7–12.

和文体および漢文体をもつ資料の構造化 —法華百座聞書抄の事例研究—

河瀬 彰宏[†]（国立国語研究所コーパス開発センター）
野田 高広 （国立国語研究所コーパス開発センター）

Structuring Documents Having both Japanese and Chinese Style of Writing: Encoding *Hokke hyakuza kikigaki sho* as a Case Study

Akihiro Kawase (National Institute for Japanese Language and Linguistics)
Takahiro Noda (National Institute for Japanese Language and Linguistics)

要旨

本研究では、国立国語研究所の「通時コーパス設計」プロジェクトの一環として『法華百座聞書抄』について、和漢混淆文の文書—『今昔物語集』のテキスト—と比較しながら、文書構造化を検討し、具体的事例を示す。『法華百座聞書抄』とは、天仁 3 年（1110 年）2 月 28 日から 300 日間にわたって講じられた法華経・阿弥陀経・般若心経の説教の聞き書きである。唯一の伝本である法隆寺蔵本には、20 日分の講説（説教・説話）が片仮名主体の漢字仮名交じり体で筆録されている。本研究では、この翻刻テキストを底本として構造化を進める。このテキストには、傍書に加えて、校注者による振り仮名・振り漢字が付されている。この振り漢字は、『今昔物語集』の文書構造化には見られなかった特徴であり、構造化上の問題となる。本研究では、文書全体の構成から文字レベル（#PCDATA）に至るまでのタグセットを精緻に考慮し、構造化における問題点を整理する。

１．はじめに

国立国語研究所（以下、国語研）では、「通時コーパス設計」プロジェクトの一環として古典資料の形態素解析を実施している。形態素解析を行うためには、基礎資料となる古典テキストの電子化および構造化が必要となる。

筆者らは、これまでに平安時代の和漢混淆文の構造化（冨士池ほか 2013）、近世口語テキストの構造化（河瀬ほか 2013；Kawase et al. 2014）、狂言台本の形態素解析（河瀬ほか 2014）などを進めている。また、国語研では他にも『太陽コーパス』（田中・小木曽 2000）や『明六雑誌コーパス』（近藤・田中 2012）、BCCWJ（前川 2008、山口ほか 2009）などの様々なテキストコーパスを電子化し、公開している。

上記のテキストコーパスは、国語研が独自に考案したタグセットに基づく XML（eXtensible Markup Language）を用いて文書構造のマークアップを行っている。しかし、各々のコーパスを規定する要素には、共通のタグが使用される場合が少なからずあるものの、基本的には共通のタグセットを使用していない。そのため、同一コーパス内での文書構造の比較や文字列の抽出は可能である一方で、複数のコーパス間の構造比較や計量分析を機械的に実施することが現状では難しいという問題を抱えている。したがって、複数のコーパスの構造を高次の視点から統一的に記述することが求められている。

本研究では、この問題を解決するために、和漢混淆文の重要資料である『法華百座聞書抄』の翻刻テキスト（小林 1975）を事例に、タグセットを考案し、文書構造化を試みる。そして、過去に構造化を実施した和漢混淆文の資料である『今昔物語集』のタグセットと

[†] a_kawase@ninjal.ac.jp

の相違点を整理し、和漢混淆文の資料における構造化の問題点を明確化する。

２．『法華百座聞書抄』の特徴と電子化の意義

『法華百座聞書抄』とは、天仁 3 年（1110 年）2 月 28 日から 300 日間にわたって講じられた法華経・阿弥陀経・般若心経の講説の聞き書きである。唯一の伝本である法隆寺蔵本には、計 35 の説話を含む 20 日分の講説が片仮名主体の漢字仮名交じり体で筆録されている。『日本語学研究事典』（飛田良文ほか（編）2007）によれば、おもに次の３点の特徴をもつことから、重要な言語資料であると考えられている：（１）説法に因縁話や比喩談などの説話を加える形式をもつため、説話集の成立と説法の関係を知る手掛かりとなること。（２）典拠を仏典に求めているため、漢文訓読語の影響が著しく見受けられること。（３）中世語の萌芽と目される新語が出現していること。

一方で、片仮名主体の漢字仮名交じり体で書かれた本文は、説教・説話のどの部分に和文脈・漢文脈の要素が現れるのか、新語がどのように現れるのか、文体の指向が鎌倉時代の和漢混交文とどのような関連をもつのか、などが未解明である（飛田良文ほか（編）2007）。したがって、『法華百座聞書抄』を機械可読な形式に整備することは、これらの問題を計量的観点から分析することを実現させ、日本語史や書誌学などの人文学研究を促進する意義がある。また、同時代には、漢文脈傾向の強いテキストが数多く存在しているため、これらの文献資料をアーカイヴ化するためのフォーマットを新たに統一的観点から提供する意義がある。

３．『法華百座聞書抄』の構造

コーパス言語学の観点から『法華百座聞書抄』を分析するためには、文書の外形的情報だけでなく、文書構造および言語構造を精緻にマークアップすることが求められる。

図 1 は、『法華百座聞書抄』の翻刻テキスト（小林 1975）をワープロ印字し直したものである。

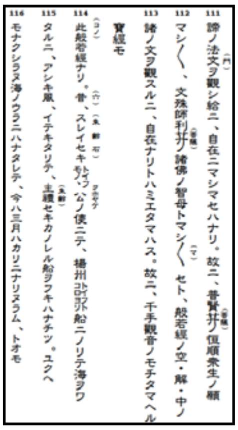

111 諦ノ法文ヲ観シ給ニ、自在ニマシマセハナリ。故ニ、普賢井ノ恒順衆生ノ願
112 マシ〳〵、文殊師利井ノ諸佛ノ智母トマシ〳〵セト、般若經ノ空・假・中ノ
113 諸ノ文ヲ観スルニ、自在ナリトハミエタマハス。故ニ、千手観音ノモチタマヘル
寶經モ
114 此般若經ナリ。昔、スレイセキ公ノ使ニテ、播州船ニノリテ海ヲワ
115 タルニ、アシキ風、イテキタリテ、主輸セキカノレル船ヲフキハナチツ。ユクヘ
116 モナクシラヌ海ノウラニハナタレテ、今ハ三月ハカリニナリヌラム、トオモ

図 1　『法華百座聞書抄』のワープロ印字（抜粋）

『法華百座聞書抄』は、（ａ）前付け部分と（ｂ）講説部分の順に構成される。（ａ）前付け部分には、内題「天仁三年二月廿八日令始修法一百座」、書写者「大安寺僧都永」、序文が記されている。（ｂ）講説部分には、計 35 の説話が収録されており、その基本構成は聞き書きされた日付、経の品名、講師名などのコンテンツ情報と、片仮名主体の漢字仮名交じり体で筆録された本文を含む（図 2）。以下では、このような構造をもつ『法華百座

聞書抄』のテキストを精緻にマークアップしていく。なお、片仮名で書かれた本文はすべて平仮名に置き換えた上で作業を実施している。

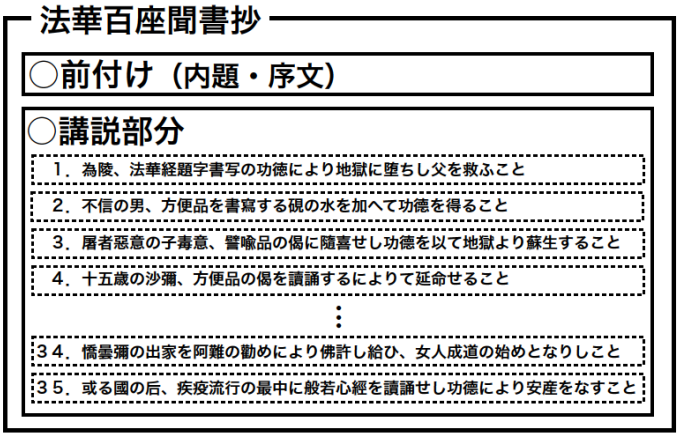


図 2　『法華百座聞書抄』の基本構成

４．文書全体の構成に関わる構造化

上述のように『法華百座聞書抄』のテキストは、（ａ）前付け部分（ｂ）講説部分をもつ。この構成は、一般的な欧文の写本と概ね一致するため、TEI P5:Guidelines（Burnard and Bauman 2007）準拠の要素を用いて構造化する場合、テキスト全体は<text>、（ａ）前付け部分は<front>、（ｂ）講説部分は<body>、をそれぞれ対応させることができる。そして、<front>内部に置かれる内題、序文、<body>内部に置かれるコンテンツ情報と講説については （text division）によって規定することができる。あるいは<article>を用いることもできる。<article>以下は基本的にコンテンツ情報と本文の塊であるパラグラフによって構成されるため、それぞれ<head>と<p>（paragraph）を対応させて明確に区別する。

以上のテキストの大局的構造を支える６つの要素を階層構造に留意して一覧表にまとめると表 1 のようになる。

表 1 <text>から<p>までの要素の一覧

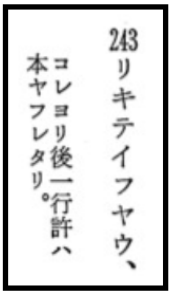
要素	説明
<text>	テキスト全体
<front>	前付け部分
<body>	講説部分
<article>	序文、説話・説教
<head>	内題、コンテンツ情報（日付・経の品名・講師名）
<p>	パラグラフ

５．パラグラフ以下の要素の構造化

次に『法華百座聞書抄』の<p>（paragraph）以下の階層について述べる。

（ａ）前付け部分<front>および（ｂ）講説部分<body>に含まれる<p>以下の内容は、講説（本文）と経典からの引用文（経文）が含まれる。本文の体裁をとるものは、欧文の電子化と同様に TEI の<s>（sentence unit）を用い、経文は<q>（quoted）と属性@type に値“経

文”を入力して表現する。また、書写者による注記（体裁注記）がある箇所については、属性に値@type=“体裁注記”を入力して区別する。図 3 は、その例である。




```
<s>
    <lb n="u243" />りきていふやう、
</s>
<q type="体裁注記">
    <s>これより後一行許は本や
        <vMark>ぶ</vMark>れたり。
    </s>
</q>
```

図 3 書写者による注記（体裁注記）の XML 表現

以上、パラグラフ以下で用いた要素を階層構造に留意してまとめた一覧を表 2 に示す。

表 2 <p>以下の要素の一覧

要素	説明
<q>	経文、書写者による注記（体裁注記）
<s>	講説（本文）、文単位

６．センテンス以下の要素の構造化

次に『法華百座聞書抄』の<s>（sentence unit）以下の階層について述べる。

（ａ）前付け部分<front>および（ｂ）講説部分<body>に含まれる<s>以下では、ルビ付き文字が頻出する。

著者らはこれまでに古典作品におけるルビ文字に対して、マークアップ上の問題点や言語構造に留意した構造化を提案してきた（e.g. 河瀬ほか 2013、Kawase et al. 2014）。ここでは先行研究—『今昔物語集』や BCCWJ のマークアップ—の方針を踏襲し、<ruby>を用いて、ルビとして振られた文字列は属性@rubyText に記述する。とくに、語（後述する短単位）を越える文字列に付与されるルビは、属性@rubyBase の値にルビが振られている文字列全体を入力して表現する（図 4）。


```
<ruby resp="校注者" rubyText="コノ" rubyBase="此の">
此
</ruby>
<corr type="hodoku">の</corr>
僧のかくのたまふ
```

図 4 @rubyBase を用いた XML 表現

通常縦書きの文書では、ルビは文字の右側に置かれる。しかし、『法華百座聞書抄』では文字の左側にも同様にルビを付与することがある。近世口語資料—洒落本—の場合、右側ルビには文字の読みや宛て語を、左側ルビには右側ルビ以外の読みや語の意味を記す傾

向があった。ここでは左側ルビについて<lRuby>を規定して表現する。

ところで、本文<s>の内容について、形態論情報（品詞・活用形・読みなど）を付与することにより、言語資源として質の高いコーパスを設計することが適う。形態論情報の付与や形態素解析の精度を向上させるためには、文字レベル（#PCDATA）の変換・修正を実施する必要がある。『法華百座聞書抄』では、具体的には（イ）本来は濁点を付けるべき文字、（ロ）踊字（ヽ・ゝ・/ \）の多用、（ハ）文字種の混在、（ニ）JISX0213 外の文字の使用といった問題がある。ここでは先行研究—『明六雑誌コーパス』と洒落本のマークアップ—の方針を踏襲し、次のようにタグを規定し、本文を整形する：（イ）<vMark>、（ロ）<odoriji>、（ハ）<char>、（ニ）<g>（gaiji）。これらの具体例を図 5 に示す。

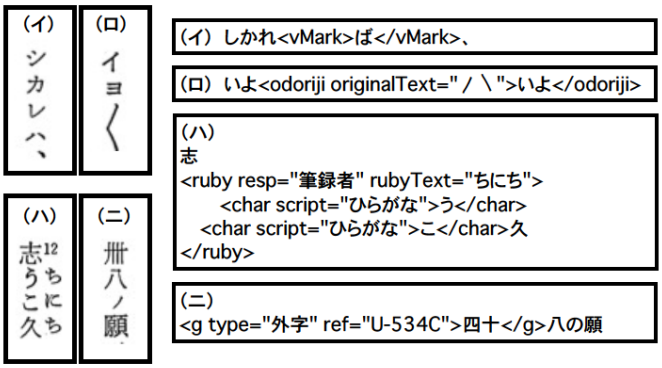


図 5 文字レベル（#PCDATA）の変換・修正
（イ）<vMark>、（ロ）<odoriji>、（ハ）<char>、（ニ）<g>

また、タグの階層構造は前後するが、<s>とは国語研が規定する短単位の集合体である。したがって、<s>直下に短単位<SUW>（Short-Unit Word）を定義し、その属性に形態論情報—語彙素、語形、書字形、品詞、活用型、活用形、発音形、語種—を付与する。この作業は<s>に該当する文の形態素解析結果に対して人手で修正しつつ付与していくものである。

以上、センテンス以下で用いた要素を階層構造に留意してまとめた一覧を表 3 に示す。

表 3 <s>以下の要素の一覧

要素	説明		要素	説明
<SUW>	短単位（語）		<ruby>	ルビ付き文字
<lRuby>	左側ルビ		<vMark>	濁点表記の文字
<corr>	本文の修正		<char>	原文表記の文字種
<odoriji>	踊字表記の文字		<g>	JISX0213 外の文字

7．本文の修正

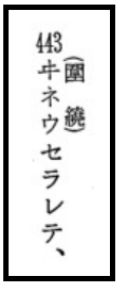
ここでは本文の修正を行う場合に用いる<corr>（correction）の用途について整理する。『法華百座聞書抄』の場合、その目的を（1）原文の誤りを正すための修正、（2）形態素解析の精度向上のための修正、に大きく分類することができる。

（1）には、原文の（a）誤字、（b）脱字の修正が存在する。（a）誤字は、属性@type に値“erratum”を、（b）脱字は、属性@type に値“omission”を入力して区別する。

そして必要に応じて属性@originalText に値として本文修正前のテキストを入力する。さらに、修正がどの段階で行われたものかを明示するために、任意に属性@resp を準備し、その値として「筆録者」、「校注者」、「作業者」を入力する。

（2）には、（c）補読（d）捨て仮名（e）振り漢字（f）返読の4種類の修正が存在する。（c）補読は、属性@type に値“hodoku”を、（d）捨て仮名は、属性@type に値“sutegana”を、（e）振り漢字は、属性@type に値“furikanji”を、（f）返読は、属性@type に値“返読前”もしくは“返読後”を入力して区別する。そして、（2）についても（1）と同様に、必要に応じて属性@originalText に値として本文修正前のテキストを入力し、修正が行われた段階を明示するために、属性@resp を準備し、その値として「筆録者」、「校注者」、「作業者」を入力する。

とくに、（f）返読については、和漢混淆文の『今昔物語集』の構造化（冨士池ほか 2013）と共通する特徴であるが、（e）振り漢字については、今回の文書構造化において新たに追加した内容である。図6に（e）振り漢字の XML 表現を示す。




```
<lb n="o443" />
<corr type="furikanji" resp="校注者" originalText="キネウ">
圍繞
</corr>
せられて、
```

図6　（e）振り漢字の XML 表現

８．位置情報および本文以外の情報

上述のタグに加え、文の改行位置には空要素の<lb/>（line break）を割り当てる。ただし、図 1（右）にも表れているように、『法華百座聞書抄』の翻刻テキストでは行番号を紙面の表裏にあわせて「オ」「ウ」とともに示している。ここでは<lb/>の属性@n に“o 行番号”や@n に“u 行番号”（o/u は表/裏に対応）を入力して表現する。

また、本文以外の情報については空要素の<info/>（information）を準備し、記述できるようにする。例えば、翻刻テキストには、書写者による傍書（注記）が本文中に一箇所だけ存在する。これを構造化する上で<info/>とその属性@text に値として傍書の記述内容を入力して表現する。

最後に、本文の欠損箇所については、<missing>を用いて表現する。とくに、文字単位の欠損と２文字以上にわたる欠損をそれぞれ図7のように表す。


```
<s>
    天仁三年二月廿八日
    令始修法一百座
    <missing>␣</missing>
    <missing>⇦␣⇨</missing>
</s>
```

図7 本文の欠損箇所の XML 表現

以上、位置情報と本文以外の情報について用いた要素を階層構造に留意してまとめた一覧表を表 4 に示す。

表 4 位置情報と本文以外の情報に関する要素の一覧

要素	説明
<missing>	欠損箇所
<lb/>	行番号（体裁注記）
<info/>	本文以外の情報（書写者による傍書）
#PCDATA	文字（character data）

９．まとめと今後の課題

本研究では、国立国語研究所の「通時コーパス設計」プロジェクトの一環として『法華百座聞書抄』について、和漢混淆文の文書―『今昔物語集』のテキスト―と比較しながら、文書構造化を検討し、具体的事例を示した。

『法華百座聞書抄』と『今昔物語集』の文書構造化における最大の違いは、とりわけ振り漢字の扱いにあった。本研究では、本文の修正処理について整理し、それらを<corr>の属性@type の値に応じて区別する方針を提案した。また同時に、修正作業の段階を明示するための工夫として<corr>の属性@resp の値として「筆録者」、「校注者」、「作業者」を割り当てる方針も提案した。これにより、原文の状態を保持しつつ、特定の修正段階における本文を抽出することが可能となった。

本研究で使用したタグセットの階層関係をダイヤグラムで表現すると図 8 のようになる。

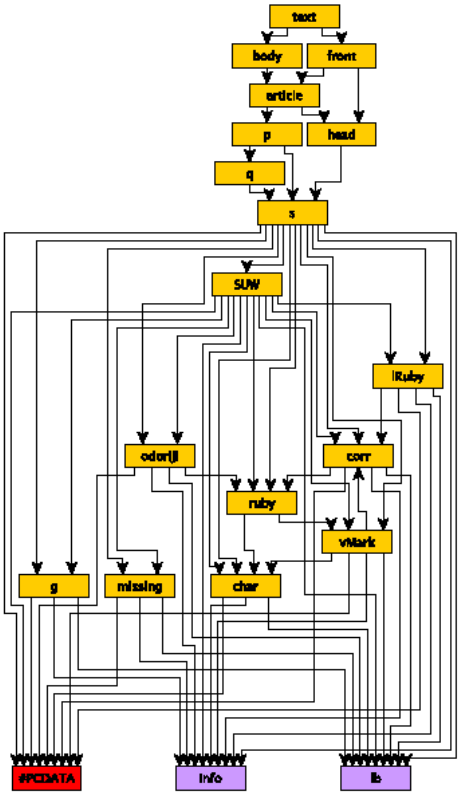


図 8 『法華百座聞書抄』のタグセットのダイヤグラム

今後の課題として、本研究において提案した文書構造化の方針を和文体および漢文体をもつ『法華百座聞書抄』以外の言語資料に適用しながら、漢文脈傾向の強いテキストを網羅的・統一的観点からアーカイヴ化するためのフォーマットを構築していく。

謝　辞

本研究は、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、代表者：田中牧郎）および国立国語研究所の共同研究プロジェクト「通時コーパスの設計」に基づく成果の一部である。

文　献

Burnard, Lou and Syd Bauman（2007）TEI P5: Guidelines for electronic text encoding and interchange, *Text Encoding Initiative.Arlington*, MA: TEI Consortium.
（http://www.tei-c.org/Guidelines/P5/）　（参照 2014-08-01）

冨士池優美・河瀬彰宏・野田高広・岩崎瑠莉恵（2013）「『今昔物語集』のテキスト整形」、第 4 回コーパス日本語学ワークショップ予稿集、pp.125-134

飛田良文ほか（編）（2007）『日本語学研究事典』、明治書院

河瀬彰宏・市村太郎・小木曽智信（2013）「TEI：P5 に基づく近世口語資料の構造化とその問題点」、じんもんこん 2013 論文集、Vol.2013、4、pp.7-12

河瀬彰宏・市村太郎・小木曽智信（2014）「『虎明本狂言集』における会話文の計量分析」、言語処理学会第 20 回年次大会発表論文集、pp.662-665

Kawase, Akihiro, Taro Ichimura, Toshinobu Ogiso（2014）Problems in encoding documents of early modern Japanese. *Proceedings of the Digital Humanities 2014*
（http://dharchive.org/paper/DH2014/Paper-934.xml）（参照 2014-08-01）

小林芳規(編)（1975）『法華百座聞書抄総索引』、武蔵野書院

近藤明日子・田中牧郎（2012）「『明六雑誌コーパス』 の仕様」、国立国語研究所共同研究報告 12-03 近代語コーパス設計のための文献言語研究成果報告書、pp.118-143

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、Vol. 4、No.1、pp.82-95

田中牧郎・小木曽智信（2000）「総合雑誌『太陽』 の本文の様態と電子化テキスト」、日本語科学、Vol. 8、pp.141-152

山岸徳平（開題）（1976）『法華修法一百座聞書抄』、勉誠社文庫 4、勉誠社

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる「『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」、Ver2.2、LR-CCG-10-04
（http://www.ni njal.ac.jp/corpus_center/bccwj/doc.html#02）（参照 2014-08-01）

関連 URL

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

『虎明本狂言集』における「思ふ」と「存ず」
―『虎明本狂言集』のコーパスデータを利用して―

渡辺由貴（国立国語研究所　コーパス開発センター）

Omou and *Zonzu* in *Toraakira-bon Kyogensyu* :Using Data from the Corpus of *Toraakira-bon Kyogensyu*

Yuki Watanabe (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所で作成中の『虎明本狂言集』のコーパスデータ（整備中、2015 年公開予定）を利用し、『虎明本狂言集』の中で多く用いられている動詞「思ふ」「存ず」の使われ方を、主に話者の属性に注目して考察する。「思ふ」および「存ず」の使用の選択にあたっては、話者の属性やその聞手、さらには名乗りや独白といった場面の問題が関与していると考えられる。

１．目的と方法

動詞「存ず」は、「思ふ」の謙譲語、あるいはあらたまり語とされる。中世軍記物語では基本的に、文末表現「と存ず」は話手が聞手と同等以下の立場の時に、「と思ふ」は話手が聞手と同等以上の立場の時に使用される（渡辺 2011）。同じく中世語の資料である『虎明本狂言集』にも「思ふ」「存ず」が多く用いられているが、両表現の使用の選択にあたり、話者の属性はどのように関与しているのであろうか。

『虎明本狂言集』における「と思ふ」類の表現を扱った研究に村上(1993)がある。定型性の強い一曲の発端部における「ばや」（「ばやと存ずる」）に注目し、狂言の構成や類の面から整理したものである。また李（2006）は、特に名乗りの場面で多用される「ウ」と「ウと思う（存ずる）」の、意味的・機能的相違に焦点を当てたものである。穐田(1975)は、狂言等における待遇表現としての「存ず」について述べている。これらの研究をふまえつつ、名乗り以外の場面も含め、総合的・数量的に「思ふ」と「存ず」両表現について検討するために、国立国語研究所で作成中の『虎明本狂言集』のコーパスデータを利用し、動詞「思ふ」「存ず」の使われ方を、主に話者の情報に注目して考察する。

『虎明本狂言集』のコーパスデータは、大塚編（2006）を底本とするもので、2015 年公開予定である。整備中のデータであり、調査結果は 2014 年 7 月 28 日現在のものである。なお、話者情報については、本文内に話者が示されていない場合にも可能な限り付与されている（小林・市村 2013，p.328）。なお、底本における両動詞の表記は「思ふ」「存ず」であるが、本コーパスの語彙素の表記にあわせ、以降、「思う」「存ずる」と現代語の表記で示す。

２．考察

２．１．「思う」「存ずる」の使用状況の概略

表 1 に、『虎明本狂言集』のコーパスデータにおいて用いられている、品詞が「動詞 - 一般」

となっている語の粗頻度の上位 10 位を示す。「存ずる」が 3 位、「思う」が 7 位となっており、『虎明本狂言集』の中でどちらの動詞も多く用いられていることがわかる。また、表 2 に、語彙素「思う」「存ずる」の用例数を[本文種別]別に示す。「思う」は 524 例中 488 例（約 93%）が、「存ずる」は 735 例中 716 例（約 97%）が「会話」の用例である。

表 1　『虎明本狂言集』において用いられる「動詞-一般」（上位 10 位）

	語彙素	粗頻度
1	言う	3305
2	取る	796
3	存ずる	735
4	然る	627
5	持つ	619
6	聞く	588
7	思う	524
8	出でる	449
9	因る	381
10	仰せる	352

表 2「思う」「存ずる」の[本文種別]別用例数

語	ト書き	引用-会話指示	引用-典拠	引用-和歌	会話	注釈	総計
思う	5	24	2	1	488	4	524
存ずる	1	16			716	2	735
総計	6	40	2	1	1204	6	1259

表 3　「と思う」「と存ずる」の[本文種別]別用例数

行ラベル	ト書き	引用-会話指示	引用-典拠	会話	注釈	総計
と思う	4	19	1	340	2	366
と存ずる	1	6		523	1	531
総計	5	25	1	863	3	897

ただし、「存ずる」の用例数には、「知る」の謙譲語としての用例が含まれているので、動詞「思う」との比較のため、ト格をとる（“語彙素「と」＋語彙素「思う」”、“語彙素「と」＋語彙素「存ずる」”）形について、表 2 と同様の調査をした（表 3）。「と思う」は 366 例中 340 例（約 93%）、「と存ずる」は 531 例中 523 例（約 98%）が「会話」の用例であった。

表 4「と思う」の話者（上位 30 名）

話者	用例数	話者	用例数
大名	35	亭主	4
主	32	次郎冠者	4
太郎冠者	31	武悪	4
夫	20	丹波	3
妻	19	見付の者	3
祖父	12	下京の女	3
出家	11	猿引	3
女	10	貸手	3
住持	9	柿主	3
果報者	8	男一	3
吉田の何某	7	継母	3
伯父	5	麻生	3
山伏	5	舅	3
教え手	5	勾当	3
おこ	4	山賊	3
鬼	4	新発意	3
すつぱ	4		

表 5「と存ずる」の話者（上位 30 名）

話者	用例数	話者	用例数
太郎冠者	58	亭主	5
主	49	果報者	5
出家	24	博打打	5
夫	22	牛博労	5
男	18	何某	5
聟	18	吉田の何某	4
大名	11	通行人	4
女	10	兄	4
甥	11	厳島の社人一	4
田舎者	8	所の者	4
妻	8	参詣人一	4
男一	8	伯蔵主	4
住持	7	猿引	4
閻魔王	7	舅	4
すつぱ	6	柑子売	4
孫一	6	見付の者	4
浅鍋売	6		

表 4・表 5 に、「と思う」「と存ず」を使用する人物上位 30 名を、[話者]情報をもとに示す。「と思う」のみを多用する「祖父」「教え手」、「と存ずる」のみを多用する「聟」「甥」「田舎者」等、その使用が偏っている人物が存在する一方、「大名」「主」「太郎冠者」「妻」「出家」等は、両表現とも多用している。次節以降、両表現の使用状況をいくつかの観点からみていく。

２．２．「と思う」のみを多用する話者・「と存ずる」のみを多用する話者について

「と思う」のみを多く使用する話者の特徴をみるため、表 4 のみにあがった話者について、「と存ずる」の使用数をあげ、「と思う」「と存ずる」の用例数の合計のうち、「と思う」の使用率が何%であるかを表 6 に示した。同様に、表 5 のみにあがった話者について、「と存ずる」の使用率を表 7 に示した。それぞれについて、２．２．１．および２．２．２．で確認する。

表 6　表 4 のみにあがった話者の「と思う」使用率

話者	と思う	と存ずる	「と思う」率（%）
祖父	12	0	100
伯父	5	1	83.3
山伏	5	3	62.5
教え手	5	0	100
おこ	4	3	57.1
鬼	4	1	80
次郎冠者	4	0	100
武悪	4	0	100
丹波	3	3	50
下京の女	3	3	50
貸手	3	1	75
柿主	3	1	75
継母	3	0	100
麻生	3	1	75
勾当	3	3	50
山賊	3	2	60
新発意	3	1	75

表 7　表 5 のみにあがった話者の「と存ずる」使用率

話者	と存ずる	と思う	「と存ずる」率（%）
聟	18	0	100
甥	11	0	100
田舎者	8	0	100
閻魔王	7	0	100
孫ー	6	1	85.7
浅鍋売	6	1	85.7
博打打	5	0	100
牛博労	5	1	83.3
何某	5	2	71.4
通行人	4	0	100
兄	4	2	66.7
厳島の社人ー	4	0	100
所の者	4	1	80
参詣人ー	4	0	100
伯蔵主	4	2	66.7
柑子売	4	0	100

＊「男」は表 5 のみにあがっていたが、表 4 の「男一」と同等に扱うこととし、表 7 にはあげなかった。

２．２．１．「と思う」のみを多用する話者

「と思う」を使用するが「と存ずる」を使用しない人物として、「祖父」「教え手」「次郎冠者」「武悪」「継母」がいる。話手が聞手と同等以上の立場にある場合は「話手≧聞手」とし、話手が聞手より立場が低い場合は「話手＜聞手」として、詳細を表 8 に示した。なお、立場の高低は、主人と使用人のような明らかな上下関係があるか否かで判断した。

表 8　「と思う」のみを使用する人物

話者	名乗り	独白	話手≧聞手	（詳細）	話手＜聞手	一人称以外	用例数
祖父			11	孫ー6、山伏（＝孫）5		1	12
教え手			5	聟4		1	5
次郎冠者			4	太郎冠者3		1	4
武悪			4	太郎冠者4			4
継母	3						3

「祖父」「教え手」については、その場面の中で高い立場にあるといえる。「祖父」の会話相手はいずれの曲においても「孫」である。同様に、「教え手」についても、聟に作法を教える人物であり、その場面において高い立場にある人物であるといえる。

「次郎冠者」および「武悪」は、立場の高い人物とはいえないが、聞手はそれぞれ同じく使

用人である「太郎冠者」であり、同等の立場の聞手に「と思う」を用いていることがわかる。

「継母」については、やや性格を異にし、「と思う」の 3 例はいずれも名乗りの場面の例である。名乗りの場面では「と存ずる」が多用されるが（村上 1993 等。２．４で後述）、「継母」は「と思ひ」「と思ひ候」を用いている。このような「と思ひ候」については２．５で述べる。

(1)わらはにもむすこが御入候へども、中／＼てうあひもなく候程に、きやうの殿をにくし／＼と思ひ候折節、寺よりくだりて候間、行人をかたらひいのりころして候、しがひをかくして候はば、ちちごの不審めされうずると思ひ、そのままおきて候、むなしくなりたるよし、ちちごに申さばやと思ひ候、いかに御入候か（ままこ　上 p.451）

２．２．２．「と存ずる」のみを多用する話者

「と存ずる」を使用するが「と思う」を使用しない人物として、「聟」「甥」「田舎者」等がいる。表 9 に、話手が聞手より立場が高い場合は「話手＞聞手」とし、話手が聞手と同等以下の立場にある場合は「話手≦聞手」として整理した。基本的には、「と存ずる」は名乗りもしくは独白、話手が聞手と同等以下の立場にある場合に用いられている。明らかに聞手より話手の立場が高い場合に「と存ずる」が使用されている例は今回の調査の範囲ではみられなかった。

表 9　「と存ずる」のみを使用する人物

話者	名乗り	独白	話手＞聞手	話手≦聞手	（詳細）	一人称以外	用例数
聟	8			10	教え手10		18
甥	4			7	伯父6、伯母1		11
田舎者	4			4	すっぱ3、目代1		8
閻魔王	5	1		1	朝比奈1		7
博打打	2			3	有徳人2、何某		5
通行人	3			1	菊一1		4
厳島の社人一	1			3	厳島の社人二3		4
参詣人一	2			2	参詣人二2		4
柑子売	2	1		1	亭主1		4

２．２．３．「と思う」のみを多用する話者・「と存ずる」のみを多用する話者のまとめ

基本的には話手が聞手と同等以上の立場の場合に「と思う」が用いられ、話手が聞手と同等以下の立場の場合に「と存ずる」が用いられる。これは中世軍記物語で見られる傾向と同様であるが（渡辺 2011）、「と思う」が特に失礼な表現であるというわけではなく、「と存ずる」というほぼ意味が同じでかつ謙譲・丁重の形式が多用されているために、高い立場の聞手に対し、敬語的ニュアンスを持たない「と思う」を使用しにくいのだと考えられる。

なお、話手と聞手の立場が同等の場合は、双方が「と思う」を使う例（太郎冠者と次郎冠者、太郎冠者と武悪）、双方が「と存ずる」を使う例（参詣人一と参詣人二）ともみられた。

２．３．「と思う」「と存ずる」両方を多用する話者について

表 10　「と思う」「と存ずる」とも 10 例以上使用する話者

話者	と思う	と存ずる	合計
太郎冠者	31	58	89
主	32	49	81
大名	35	11	46
夫	20	22	42
出家	11	24	35
女	10	10	20

次に、「と思う」「と存ずる」のどちらも 10 例以上の使用例がある、「大名」「主」「太郎冠者」「夫」「出家」「女」について検討したい。表 10 に、それぞれの人物の「と思う」「と存ずる」の用例数をあげる。また、詳細をみるため、上記の 6 話者の「と思う」「と存ずる」の使用場面を表 11・12 に示す。

表 11　話者別「と思う」の使用場面

	と思う							
話者	名乗り	独白	次第	話手≧聞手	聞手の内訳	話手＜聞手	一人称以外	合計
太郎冠者		8		8	売り手1、次郎冠者3、武悪4	3	12	31
主				23	太郎冠者22、太郎冠者・次郎冠者1		9	32
大名	1			24	太郎冠者19、女2、昆布売2、下人1		10	35
夫	1			13	妻10、告げ手2、妻・出家1		6	20
出家		5	2	3	所の者1、蛸の精1		1	11
女	2	1		3	大名1、新発意1、山賊1		4	10

「と思う」については、話手が聞手と同等以上の立場にある場合に用いられていることが多い。なお、表 11 をみると、「太郎冠者」に「話手＜聞手」の例が 3 例みられるが、いずれも鬼・小名類の例で、「主」に対し失敗の言い訳をする場面において、「と思ふ（ひ）て」の形であらわれる。高い立場の聞手に対して文末終止形の「と思う」を使用した例はみられなかったが、「と思ひて、」のように、文末以外の形であれば、高い立場の聞き手に対して「と思う」を使用しにくいという制限が弱まるのだと考えられる。

(2)私は又、かねの音をきひてこひと仰られた程に、それかと思ふてきゐて参つた、それならばそれととう仰られひで（栗やき　上 p.581）

表 12　話者別「と存ずる」の使用場面

	と存ずる							
話者	名乗り	独白	次第	話手＞聞手	話手≦聞手	聞手の内訳	一人称以外	合計
太郎冠者		9			48	主22、大名18、売り手3、次郎冠者1、仲裁人1、妻1、新座の者1、客人1	1	58
主	34	8		3	3	客人一2、仲裁人1	1	49
大名	8			1	2	女2		11
夫	12	2		1	7	出家4、妻1、仲人1、仲裁人1		22
出家	22	2						24
女	7	1			2	大名1、男1		10

一方「と存ずる」は、聞手が話手と同等以上の立場にある場合に用いられることが多い。表 12 には「話手＞聞手」の例が 5 例（「主」3 例、「大名」1 例、「夫」1 例）みられるが、いずれも「太郎冠者」に対して「と存ずる」が使われる例である。うち 4 例は(3)のように富士や鞍馬への参詣の話という共通点があり、あらたまった口調になっている背景として、神仏への畏怖の気持ちが影響していよう。穐田（1975，p.4）にも「場面的存在としての『神仏』に対する態度が、下位者である聞手に対しても向けられる」例について述べられている。

(3)「近比めでたひ、此福を某がとらふとぞんずる「某にもふくを下された「それはもろともにめでたうござる（くらままいり　上 p.527）

もう 1 例は、話手が「夫」、聞手が「太郎冠者」ではあるが、実は太郎冠者になりすました「妻」が聞手という例であり、対象は神仏ではなく妻であるが、恐れの気持ちが背景にある点で(3)と類似している。

(4)しぜん山のかみがそのふみをみたらば、なふ中／＼ただはおくまひとぞんじて（はな

ご　下 p.64）

このように、「と思う」「と存ずる」両表現とも多く使う話者であっても、特定の状況以外では、「と思う」は話手と同等以下の立場の聞手に対して用いられ、「と存ずる」は話手と同等以上の立場の聞手に対して用いられることがわかる。

２．４．名乗り・独白について

「と存ずる」については名乗りの場面での使用例が多いことも注目される。例えば、表 11・表 12 をみると、「と思う」を名乗りの場面で使用しているのは「大名」「夫」「女」の計 4 例であるのに対し、「と存ずる」を名乗りの場面で使用しているのは「太郎冠者」以外の 5 人物、計 83 例と多い。この名乗りの場面における「と存ずる」は、登場人物に対してではなく、観客への配慮として用いられていると考えられる[1]。例えば(5)の例では、新しい者を雇おうとしている件について、「大名」が名乗りとしては「と存ずる」を用い、直後に「太郎冠者」への発話で「と思う」を用いるという使い分けがみられる。

(5)新座の者をあまたおいてつかはふと存る、あるかやい　「お前に　「ねんなうはやかつた、汝がよろこぶ事がある　「いかやうな事でござるぞ　「汝一人にてはわれもめいわくにあらふず、身共もつかひたらぬ程に、新座の者をおいてつかはふと思ふがよからふか（鼻取りずまふ　上 p.192）

また、独白については、「と思う」「と存ずる」両表現ともみられる。(6)は独白の例であるが、独白であれば聞手はいないので、本来「と存ずる」のような配慮表現は不要であるが、これも名乗りと同様、観客への配慮として用いられているものであると考えられる。

(6)「太郎くわじやを留守においてござるが、何といたひているぞ　やうすを見うと存る、あらきどくや、 おくびやうなやつじやがきどくに夜まはりをするよ、(くいか人か　上 p.592）

なお、李（2006，p.62）に、名乗りや独白において、「『ウと思う』系は『ウと存ずる』あるいは『ばやと存ずる』のような敬語の形を取る。独白の時も、観客を聞き手として想定しているため、独白に「ウと思う」がその形式のまま使われることはない。」とある。実際は、「ウと思う」に限らず、聞手の存在しない場面において「と思う」が使われる際は、(7)のように「思ふたれば」「おもふが」等、文中の形式か、「候」がついた「と思ひ候」の形であり、敬語を伴わない文末終止形の「と思う」の用例は見られなかった。

(7)「言語道断の事じや、誠になくかと思ふたれば、そばに水ををひて目へぬる、扨々にくひ事じや、此よしたのふだ人に申さう（すみぬり　上 p.185）

なお、名乗りは狂言に特徴的なものであるが、中世軍記物語においても、聞手が多数いる公の場面で、行動予定を提示する「と存ずる」の形式が見られ、狂言における名乗りの形式と類似している（渡辺 2011，p.39）。

(8)進出て申けるは、「縦守殿は退給とも、維行は罷留りて、八郎殿の大矢をあたりてみんと存候。(略)」(話手：山田小三郎維行／聞手：多数　『保元物語』金刀本　p.101)

[1] 村上（1993）でも、一曲の発端部における「ばや」の用法を検討し、「にて候…ばやと候」「にて候…ばやと存る」等の形式は「観客への極めて高い配慮を示す表現」であるとしている。また、「でござる…うと存る」(「観客に対する配慮の表現として標準的なもの」)」「ぢゃ…う(「観客に対する配慮を表す表現のなかで最も低く、くだけた表現」)」等、五つの段階が認められ、虎明はこれらの段階を意図的に用いているとしている（p.570・p.576）。

２．５．「と思い候」について

ここまで「と思う」と「と存ずる」についてみてきたが、「と思う」に「候」がついた形式と、「と存ずる」には違いがあるのか考えてみたい。“語彙素「と」＋「思う」＋「候う」”で検索した結果は 13 例で、[本文種別]が「会話」の用例は、「出家」5 例、「継母」2 例、「女」「妻」「若市」「ごぜ」それぞれ 1 例であった[2]。いずれも名乗りの場面において用いられていた[3]。「と思い候」の話者は、「出家」を除いていずれも女性であることから、女性は「存ずる」よりも「思う」を使う傾向があった可能性がある。女性話者による「と思う」「と存ずる」の使用を表 13 に示す。

表 13　女性話者による「と思う」「と存ずる」

話者	と思う	と存ずる
妻	19	8
女	10	11
継母	3	
下京の女	3	3
若市	2	
尼	2	
上京の女	2	1
お寮	1	
女房	1	
伯母	1	
後家	1	1
ごぜ	1	
合計	46	24

「妻」「女」等、女性話者が「と存ずる」を使用する例も少なくないが、合計数では「と思う」の方が多い。表 3 で、会話における「と思う」が 340 例、「と存ずる」が 523 例であったこと、つまり全体では「と存ずる」の方が約 1.5 倍多くみられたことと比較すると、女性は「と存ずる」より「と思う」を使いやすいようである[4]。なお、大塚編（2006，下 p.69 注）に「虎明本ではサウラフの表記は多く『候』と漢字表記であるが、この場合のように女性のせりふ中では仮名表記にする傾向がある。」とあるように、女性の発話については意識な書き分けがなされていたようであり、「と思う」「と存ずる」の使い分けにもこのような意識があらわれているのではないだろうか。

「と思い候」は女性が使いやすい表現であることが推測されるが、待遇面ではどのようにとらえるべきであろうか。ここで、台本に書き入れられた「注釈」における「と思い候」の例についても考えてみたい。注釈において、「と思い候」の用例は(9)の例のみであったが、「とをもひ候」と「と存候」を続けて使用している。

(9)をぢさせられなと云て、つめたるがましかとをもひ候「惣じて、拍子物も、又かやうのたぐひも、うりてささやきおしへたるがましかと存候（よろい　上 p.68）

[2]他に、「注釈」「引用 - 典拠」「引用 - 会話指示」が 1 例ずつあった。
[3]村上（1993，p.577）の「『と思ひさふらふ』の話手はすべて女で、鬼の継子、石神、若市、瞽女座頭に用いられている。『と思ひ候』六例のうち五例は祐善・蝉・栄螺の出家に、もう一例は継子の継母の科白に用いられている。」という調査結果と同様である。
[4]なお、中世軍記物語においても、母娘間等、女性同士の発話では文末表現「と存ずる」の使用が確認できなかった（渡辺 2011，p.40）。

これらは台本を読む者に対する配慮表現であろうが、このように二つの表現が続けて使われている例があること、また、基本的には「と存ずる」が使われる名乗りの場面で「と思い候」の形がみられることをあわせて考えると、「候」のついた「と思い候」は丁寧な形式であり、「と存じ候」とも待遇面では大きな差はないと考えてよいかと思われる。

３．おわりに

『虎明本狂言集』における「思う」「存ずる」の使用の選択にあたっては、話者・聞手の属性が関与している。ト格をとる形式についてみてみると、基本的には話手の立場が聞手と同等以下の場合は「と存ずる」が、話手の立場が聞手と同等以上の場合は「と思う」が用いられる。また、観客への配慮表現として、名乗りや独白等の場面においても「と存ずる」が用いられることが多い。

このように目上の聞手（観客を含め）に対する発話で「と思う」があらわれにくいのは、「と存ずる」という謙譲・丁重の形式が多用されている中で、高い立場の聞手に対し、「と思う」をあえて選択し、使用することははばかられるためであると考えられる。なお、「と思い候」のように丁寧語化された形式であっても、女性による名乗りや注釈等、限られた範囲にしかあらわれない。ただし、「と思ひて」のように、文末形式以外の形であれば、「と思う」を目上の聞手に対して使用しにくいという制限は弱まるようである。なお、男性話者に比べ女性話者は「と思う」を使用する傾向があり、これも「と思う」の特徴といえる。

付記

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：田中牧郎）による成果の一部である。

参考資料・文献

大塚光信編(2006)『大蔵虎明能狂言集　翻刻　註解』上・下　清文堂出版

永積安明・島田勇雄校注(1961)『日本古典文学大系 31 保元物語･平治物語』　岩波書店

穐田定樹(1975)「『存ず』について」，大谷女子大学紀要，9，pp.1-19

小林正行・市村太郎(2013)「『虎明本狂言集』コーパスの構造化―仕様と事例の検討―」，第３回コーパス日本語学ワークショップ予稿集，pp.323-332

村上昭子(1993)「『大蔵虎明本狂言集』における終助詞『ばや』について」『小松英雄博士退官記念　日本語学論集』pp.555-578　三省堂

李淑姫(2006)「虎明本狂言集における『ウと思う』の用法―推量・意志の助動詞「ウ」との比較―」，日本學報，68，pp.55-67

渡辺由貴(2011)「中世における文末表現『と思ふ』と『と存ず』」, 早稲田日本語研究, 20, pp.34-45

日本語点字資料の語種的特徴

Lexical Characteristics of Japanese Braille

中野真樹（国学院大学）
Nakano Maki(Kokugakuin University)

要旨

現在、日本語文字表記システムとして、触読文字である点字と視読文字である墨字（すみじ）が平行してつかわれている。漢字かなまじり文を基本とする墨字にたいして点字は6点点字1字がほぼかな1字に相当するかな専用文でかかれる。そのため、しばしば「漢字をつかわないのでは漢語の同音異義語などの問題で文章の可読性に支障がでるのではないか」などという疑問をきくことがある。仮にそれが事実であるとすれば点字使用者を対象読者としてかかれた点字文章は、漢語をへらして和語の比率がたかくなることが予測される。そこで、現代点字新聞『点字毎日』近代点字新聞『点字大阪毎日』について語種比率を調査したところ、各時代の墨字版新聞と語種比率がほとんど差がないことがあきらかになった。これにかんしては、かな専用文であっても文脈による理解のため、漢語の同音異義語は可読性に大きな影響をあたえていないことがすでに指摘されている。

1．はじめに

日本語点字は、墨字日本語漢字かなまじり文とならんで日本語をかきあらわすための日本語文字表記システムである。明治期に考案されて以来つかわれつづけている。また表記についても、視覚により視読する墨字（すみじ）とはかなづかい等の面で一部ことなる部分があり、独立した歴史・文化をもつ文字体系である。とはいうものの、点字と墨字はまったく無関係ではなく、墨字でかかれた新聞や文学作品等がさかんに点字に翻字されている。そのための情報処理技術に関する研究もおこなわれており、特に漢字を使用しないため文節わかちがきをする日本語点字への翻字作業の補助として、形態素解析器が利用される場合もある。このように日本語点字は日本語情報処理の観点からの研究は数多くあるものの、日本語学研究はあまりなされていないといえる。

そこで、本研究ではかな専用文である日本語点字を日本語研究資料として用いることとし、日本語点字の語種について調査をおこなった。

2．日本語点字の表記上の特徴

日本語点字は、世界でもひろく普及している6点点字を使用しており、点字1字がかな1字にほぼ相当する。ただし、ひらがなとカタカナの区別はない。日本語点字による文章は基本的には漢字をもちいないかな専用文でかかれ、そのため独自の規範をもつ文節わかちがきをおこなう。また、かなづかいは「現代仮名遣い」とは一部ことなり、たとえば墨字では「は」「へ」と書字する助詞にかんしてはワ行の「わ」ア行の「え」に相当する点字かなでかく。そして和語・字音語の一部の語のウ列・オ列長音表記に長音符「ー」をもちいるなど、『現代仮名遣い』とくらべて表音性のたかいかなづかいである。このような表記の方針は日本語点字成立時からおおきな変更はなく、一貫してかな専用文・文節わかちが

きがおこなわれていることがあきらかとなっている[1]。

これにかんして、日本語点字かな専用文は、日本語墨字漢字かなまじり文とくらべて同音異義語や同訓異字などの問題から可読性や情報量の点においておとっているのではないかという仮説が想起・提唱される場合がある。またその仮説を検証もしないまま事実のこととして提示するような例もみうけられる。

たとえば日本語能力試験の点字受験に関する研究である秋元ら（2014）では、以下のようにかかれている。

> 漢字と仮名を用いて書かれた文章を、表音文字である点字に訳すことによって、損なわれる情報がある。しかし、試験においては漢字を知らなくても文章を理解できるという側面もある。また、現在 JPLT 点字冊子試験では、点訳時に損なわれる情報に対して必要に応じて注釈を付加しているが、注釈によって情報量が増えることが回答に影響する可能性もあろう。
>
> （秋元ら 2014：291）

ここで日本語能力試験に影響をおよぼすほどに「表音文字である点字に訳すことによって」「損なわれる情報がある」とのべるのであれば、実際にどのような情報がどれほど欠落しているかは明示する必要があろう。また、かな専用文でかかれた文章の可読性や情報量が、日本語能力試験の受験にどれほどの影響をおよぼすのか、検討する必要があるとおもわれる。

しかしながら、墨字漢字かなまじり文でかかれた文章が点字かな専用文に翻字される機会はおおく、「点訳者」とよばれる翻字者の育成もさかんである。翻字にさいしては「点訳注」などとよばれる語句の説明を付加することはあるが、それも最小限にとどめられている[2]。墨字漢字かなまじり文でかかれた文章を点字かな専用文になおす場合に、わずらわしさをかんじるほどに同音異義語にかんする点訳注が必要というのであれば、翻訳や翻案が必要となろう。しかし実際には墨字でかかれたかきことばをそのまま点字に翻字する方法が主流であり、さまざまな分野で墨字から点字への翻字文章が情報保障のために作成されている。

また、このような「かな専用文は漢字を使わないから同音異義語などの点で読解に困難が生じる」という説は、もっともらしくきこえるが実際には言語学的には事実とはいいきれない「漢字幻想」であり、そのような説を検証もなしに採用する言説については、言語学をはなれて「日本語表記漢字不可結論」などといったイデオロギー上のたち位置から説明するべきであるという指摘が文字学や社会言語学の研究者からはなされている。そして、実際には日本語かな専用文が日本語漢字かなまじり文とくらべて可読性や情報量において読者に深刻な不利益をもたらすほどにすくないということは証明できないという主張もある。

たとえば、Unger（2004）では、「漢字をめぐる 6 つの迷信（Six Myths about Chinese Characters）

---

1 日本語点字の表記の歴史については、金子（2007）がくわしい。

2 点訳者育成のための入門書である当山（1998）では以下のようにのべられている。「点訳者挿入符は、文の流れを中断して説明を行なうものですから、その使用は必要最小限にとどめるべきです。同音異義語であっても前後の文脈から意味が判断できる場合は用いる必要はありませんし、用いる場合もできるだけ簡潔にしなければいけません」（当山:166）」

[3]」のひとつとして、「絶対不可欠の迷信（the Indispensability Myth）」をあげている。そのなかでは、このように述べられている。

> 日本語の漢語の同音語群の中には英語の群より多くの同音異義語がふくまれているものが多い。辞書やワープロの入力作業では、「コウセイ」に少なくとも 2 ダース（固有名詞を除く）の熟語が見られる。しかし、そのうちの 10 語だけが、通常の書き物でよく出会う。しかも、それらは「構成」「攻勢」「校正」「厚生」「恒星」「更生」と、まったく共通点がない。しかも、これらの語は、辞書以外では、まれにしか並べて使われることがないし、その頻度には違いがある。（略）日本語は音節数が比較的少ないので、必然的に同音語の数が異常に増えた、と繰り返し論じられてきているが、英語の同音語の数のほうがはるかに多いのである。
>
> （アンガー・奥訳 2005：59）

次に、論文の一部をかな専用文でかくことをこころみるかどや（2012）を引用する。

> たとえば、ささいなことがだ、漢字かなまじり文でかかれているものの、漢字使用を減らしている本稿程度のものですら、「よみにくい」とかんじている読者（あなた）がいるとすれば、「自分を守ってきた鎧」にしがみつき、かたくなに変化をこばんでいる可能性がたかい。あるいわ、かりに　この　ぶんしょーが　かんぜんな　「ひょーおん　かな　わかちがき」で　かかれていたら　どーだろーか。　にほんごが　だいいちげんごで、　かつ　ひらがなお　しっているひとで　あるならば、　こーゆー　ひょーきの　にほんごお　よめない・りかいできない　はずわない。　あるのわ、　よむこと・りかいすることお　こばむとゆー　たいどだけである（よみにくさわ　ほんの　すこしの　じっせんに　よって　なれることで　たやすく　かいしょー　できる）。
>
> （かどや 2012：150）

かどや（2012）のかな専用文による引用部でも、漢語もおおくあらわれ、そのなかの一部の語についてはたとえば「じっせん（実践・実戦・実線）」、「かいしょー（解消・改称・快勝・甲斐性）」といった同音異義語が問題となりそうなものもふくまれている。しかしながら、それを「よめない・りかいできない　はずわない」とのべられている。そして、よみにくさをかんじるのだとすれば、それは「なれ」の問題であるとしている。

また、これらを計量的にうらづける論文としては、現代日本語点字資料の語種の比率について調査した羽山（2014）がある。次章で、羽山（2014）を参照しつつ、日本語点字の語種の比率について検討していく。

３．かな専用文としての日本語点字資料にあらわれる語種について

３．１　現代日本語点字の語種的特徴

羽山（2014）は、「一般に、漢字を使用しないで日本語を書こうとすると、漢字に依存せ

[3] Unger(2004:1-12)。日本語訳であるアンガー・奥村（2005）が存在するため、本予稿では日本語訳を引用する。

ず耳で聞いてもわかるような語選択が行なわれると考えられる」という仮説をあげ、点字新聞『点字毎日』[4]および墨字新聞『毎日新聞』の、2011 年の一面記事のなかから、両者でほぼ同じ内容をとりあつかっているとかんがえられる記事を「国内政治」「国外政治」「経済」「災害」「事件」という分野ごとに 1 つずつ抜き出して、そこにあらわれる語種の比較をおこなっている。

その結果を整理したものが表 1 である。点字の「わかちがき法」を基準に語を分割し、助詞助動詞をのぞいた自立語のうち、固有名詞以外について、のべ語数と異なり語数のなかの語種の割合を調査したものである。のべ語数、異なり語数それぞれについて残差分析をおこなった結果をあわせてしめした。その結果、点字か墨字かという文字種によって語種比率にはとくにおおきな差というものはみられず、どちらも新聞の語彙特徴をしめすように漢語の割合がおおくなっていることがわかった[5]。かな専用文であるからといって、点字新聞が漢字かなまじり文とくらべて漢語の使用をへらしているわけではないということが、この調査からうかがえる。

この調査は現代語における語彙調査であるが、『点字毎日』は 1922（大正 11）年 5 月に創刊された近代点字新聞『点字大阪毎日』をその前身としており、点字新聞の経年的調査も可能である[6]。また、墨字版新聞として同時期に『大阪毎日新聞』が発行されていた。

【表 1　『点字毎日』『毎日新聞』（2011 年発行）の語種比率】[7]

		のべ語数			ことなり語数		
		和語	漢語	その他	和語	漢語	その他
点字	国外政治	31.4	64.7	3.9	32.5	65.0	2.5
	災害	30.5	67.1	2.4	32.1	65.0	2.9
	事件	21.6	67.0	▲▲11.4	22.4	68.4	▲9.2
	国内政治	32.0	66.0	2.1	36.4	62.3	1.3
	経済	29.6	62.2	8.1	35.3	▼55.9	▲8.9
墨字	国外政治	26.8	68.6	4.6	27.8	69.1	3.1
	災害	34.9	63.9	1.2	32.1	66.0	1.9
	事件	▼17.9	76.4	5.6	21.3	72.5	6.3
	国内政治	19.6	▲79.4	0.9	22.7	▲76.0	1.3
	経済	30.7	63.3	5.9	32.7	62.4	1.2

＊▲　5%水準で有意に多い　▼　5%水準で有意に少ない
▲▲　1%水準で有意に多い

近代雑誌文献にみられる語種比率は、漢語の比率がたかったことが指摘されている[8]が、そ

[4] 「点字毎日」は「毎日新聞」の記事を点字に翻字したものではなく、独自の編集室を持ち点字使用者のためにかかれた新聞である。そのため、語や文体の選択にあたっては点字使用者の可読性の利便を考慮しているものとかんがえられる。
[5] 新聞で漢語の比率がたかいことは、佐竹・岸本（1998）などで指摘されている。
[6] 毎日新聞社『点字毎日』（http://www.mainichi.co.jp/corporate/tenji.html）を参照
[7] 羽山（2014）表 1 および表 2 を整理したもの。
[8] 田中（2013:19-20）によると、明治前期の「明六雑誌コーパス」では漢語の比率がきわめ

れでは『点字大阪毎日』の漢語比率はどのようなものであったのか。羽山（2014）の調査方法をもとに、『点字大阪毎日』創刊号から約 1 ヶ月間に刊行された点字新聞の語種比率について調査し、同時期に刊行された墨字版『大阪毎日新聞』との比較をおこなった。

3．2　近代日本語点字新聞の語種比率

近代日本語点字新聞『点字大阪毎日』は、1922（大正 15）年 5 月 11 日に創刊された。それ以来、週 1 回毎週木曜日に刊行されており、全 16 ページからなる。内容は、墨字版の新聞と共通するような国内外の時事ニュースと、視覚障害者が関心をもつようなトピックに特化したニュースとに大別できる。また、連載小説として菊池寛作『恩讐の彼方に』が掲載されている[9]。これらのなかから、墨字版『大阪毎日』との比較にあたって語彙種をなるべく共通点のおおいものとするため、国内外の時事ニュースのみに調査対象をしぼった。また、ページ数の関係から一つ一つの記事の分量は墨字版の新聞記事と比べると少なくなっているため、調査する記事を選定せずに 1922 年 5 月に刊行された分の記事について、すべてを調査した。

墨字版『大阪毎日』については、『点字毎日』発行日の 1 面記事のなかから『点字大阪毎日』で採取した語数とほぼ同数となるまで記事単位のランダムサンプリングをおこなった。

調査結果を表 2 にしめす。

【表 2　『点字大阪毎日』『大阪毎日』（1922 年 5 月発行）語種比率】

	のべ語数			ことなり語数		
	和語	漢語	その他	和語	漢語	その他
点字	50.9	48.0	0.9	45.0	54.5	1.4
墨字	51.5	48.0	0.4	41.0	59.5	0.7

ここから、現在の『点字毎日』および『毎日新聞』と比較すると漢語の比率がひくくなっている。また、「その他」に分類した外来語が非常にすくなくなっていた。このように、時代によって新聞の語種比率は変化していることがわかるが、一方、点字か墨字かという文字種による語種比率の差というものは、近代新聞でも現代新聞と同様にほとんどみられなかった。

4．おわりに

以上でみてきたように、現代点字新聞も、近代点字新聞も語種比率という面では、その時代の墨字新聞との間では差はほとんどみられなかった。従来「漢語の同音異義語のためにかな専用文では可読性がおとるのではないか」というようないいかたで予測されていた、点字か墨字かという文字種による語種比率の差というものは観察できず、かな専用文でかかれた点字新聞でもある時代や新聞というメディアの特徴があらわれる結果となった。今後は、範囲をひろげて各時代の点字新聞や、『毎日新聞』以外の墨字新聞についてもさらに

てたかく、明治中期から大幅に減少していることがあきらかにされておりその後徐々に減少していくものの、大正期の「太陽コーパス」においても、現在よりも漢語の比率は高くなっている。

[9] 1919（大正 8）年に発表された時代小説。

詳細に調査していく必要があろう。

また、これらの結果をうけて、「かな専用文で漢語をもちいてもなにも問題がない」という結論をみちびくわけにはいかない。漢字に依存した漢字かなまじり文を「日本語」の基準とするのであれば、点字使用者などのかな専用文を使用するひとにとっては漢語が日本語文章読解や文章発信の障害となる場合があることは、あべ（2002）でのべられている。

今後、墨字をおもに使用する研究者であっても、自分のかいた学術論文やエッセイなどが点字などのかな専用文に翻字されてよまれる場合を想定し、その場合に漢語がどのように文章読解に作用するのか検討する必要もあろう。

しかしながら、検証もしないまま「かな専用文では漢語の同音異義語を区別できないので可読性におとっている」などとかるがるしくいうことは、点字使用者などのかな専用文を使用して生活しているひとびとを不必要におびやかし、偏見を助長してしまう結果にもなりかねないことについても、とくにことばについての発言に一定の影響力をもつ日本語研究者や日本語教師、国語教師は留意するべきである。

謝辞

今回の予稿執筆にさいして情報提供および助言をしてくださった羽山慎亮氏に感謝いたします。また、調査の便宜をはかってくださった筑波大学附属視覚特別支援学校資料室に感謝いたします。

参考文献

秋元美晴、河住有希子、藤田恵（2014）「点字使用者の日本語学習に関する調査―日本語能力試験点字冊子試験受験者の日本語学習―」『恵泉女学園大学紀要』26 pp.283-294

あべ・やすし（2002）「漢字という障害」『社会言語学』2 pp.37-55

アンガー・J・マーシャル、奥村睦世/訳（2005）「漢字をめぐる 6 つの迷信」『社会言語学』5 pp.53-62

かどや・ひでのり（2012）「識字／情報のユニバーサルデザインという構想―識字・言語権・障害学」『ことばと社会』14 pp.141-159

金子昭責任編（2007）『資料に見る点字表記法の変遷―慶応から平成まで』日本点字委員会

佐竹秀雄、岸本千秋（1998）「新聞第一面の語彙　1997 年の新聞 3 紙を資料として」『武庫川女子大学言語文化研究所年報』10 pp.5-20

田中牧郎（2013）「『明六雑誌コーパス』『太陽コーパス』から見る近代語彙」『国語研プロジェクトレビュー』14-1 pp.18-27

当山拓（2002）『改訂版　点字・点訳入門』産学社

羽山慎亮（2014）「点字新聞の語彙的特徴」『社会言語学』14（未刊・掲載予定）

Unger, J. Marshall(2004) ”Ideogram Chinese Characters and the Myth of Disembodied Meaning” University of Hawai’i Press

全文検索システム『ひまわり』を用いた既存言語資料の活用方法の検討

山口昌也 (国立国語研究所言語資源研究系) †

Exploitation of Existing Language Resources by Full-Text Search System “Himawari”

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

要旨

本稿では，筆者が開発している全文検索システム『ひまわり』を対象として，既存の言語資料を活用する方法を検討した。既存の言語資料を利用するための検索ツールを考える場合，言語資料の記述形式に対応できるよう機能を実装するとともに，言語資料の再配布可能性や規模などの性質に応じた運用方法を検討する必要がある。本稿では，三つの言語資料を全文検索システム『ひまわり』にインポートし，他のユーザと共有可能な状態にする過程をとおして，(1) 多様なデータ形式に対するインポート能力を確認した，(2) 大規模なデータに対応するためにサブコーパス単位のインポート機能を実現した，(3) 利用条件に応じた配布形態に対応するために，配布資料のパッケージ化を図った。

1　はじめに

本稿では，筆者が開発している全文検索システム『ひまわり』(山口，田中 2005)[1] を用いて，既存の言語資料を活用する方法を検討する。現在，言語資料の活用を支援するためのシステムが数多く提案されており，Web ベースのコーパス検索システム (今井ら 2013，小木曽ら 2011 など)，コーパス管理機能を備えた検索システム (松本ら 2006 など)，高度な分析機能を備えた検索システム (樋口 2003 など) などが利用可能になっている。これらの検索システムに対して,『ひまわり』は，(1)XML により記述された多様な形式の言語資料の全文検索・閲覧，(2) 多様な形式の言語資料のインポート機能などの特徴を持っている。本稿では，特に，(2) の特徴を活かして，既存の言語資料を『ひまわり』で活用する方法を考える。

山口 (2013) では，既存の言語資料を『ひまわり』用のデータ形式に変換するための方法を検討・実装した。しかし，個人が作成するような小規模な言語資料を想定しており，言語資料の規模や，他の研究者と共有する際の問題については，十分考慮していなかった。また，現在のところ，実際の言語資料への適用は 1 例 (『青空文庫』XHTML 版) のみであり，特に独自形式のテキストデータをどの程度インポートできるのか，検証が十分でなかった。

そこで，本稿では，(1) 既存の言語資料を『ひまわり』にインポートし，他の研究者と共有可能な状態にするよう試みる，(2) その過程で発生する問題を明らかにする，(3) 解決するための仕組みを『ひまわり』に実装する，という手順で既存の言語資料を活用する方法を検討した。

2　既存の言語資料の活用

2.1　対象とする言語資料

今回使用した言語資料は,『日本語話し言葉コーパス』[2](以後，CSJ),『CD- 毎日新聞データ集』[3](以後,「毎日新聞」)，米国議会図書館蔵『源氏物語』[4](以後,「源氏物語」) の 3 種類である。これらを『ひまわり』にインポートし，他の研究者に配布する場合，どのような問題が発生するかを，(a) データ形式，(b) データの規模，(c) 利用条件，の三つの観点から考えてみる。上記の言語資料は，これら三つの観点の上で特徴がでるように選択した。各言語資料の内訳を表 1 に示す。

†http://www2.ninjal.ac.jp/masaya
[1]http://www2.ninjal.ac.jp/lrc/
[2]http://www.ninjal.ac.jp/corpus_center/csj/
[3]http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[4]http://textdb01.ninjal.ac.jp/LCgenji/

表 1: 対象とする言語資料の内訳

言語資料	データ形式	規模	利用条件
CSJ	XML	2.9GB	有償・再配布不可
毎日新聞	独自形式テキスト	300〜600MB(1 年分)	有償・再配布不可
源氏物語	独自形式テキスト	2.8MB	無償・再配布可能

2.2　活用する際の問題と対策

2.2.1　データ形式

言語資料は，個々の形式に従って記述されている。そのため，それらを解釈して『ひまわり』で検索するための形式 (以後,『ひまわり』形式) に変換する必要がある。

この問題に対して，山口 (2013) では，言語資料ごとに変換規則を用意することにより，多様な形式の言語資料のインポートに対応している。想定するデータ形式は，独自形式のテキスト，および，XML(HTML は内部的に XHTML へ変換) である。変換規則は，独自形式のテキストの場合は正規表現による文字列置換規則 (詳細は，後述)，XML の場合は XSLT スタイルシートにより記述する。前述のとおり，この機能はすでに公開中の『ひまわり』に実装されているため，本稿では実際の言語資料に対する検証のみを行う。

2.2.2　データの規模

一般的に，検索システムで言語資料を利用する場合，扱えるデータの規模に制限がある。これは，検索システムのハードウェア上の制限のほか，規模を大きくしすぎると実用的な検索時間で結果を得られないなど，利用上の制約が存在するからである。『ひまわり』の場合，広範な PC での動作を考慮すると，単一の『ひまわり』形式データの上限は，おおむね 150MB 程度である。

データの規模の問題に対する対策としては，サブコーパスに分割し，それぞれ独立した『ひまわり』形式データして管理することが考えられる。この方法は，CSJ のように，複数のサブコーパスからなるコーパスにとっては自然な対応であろう。また,「毎日新聞」のように，1 年分が数百 MB になるテキストデータベースの場合も，単年ごとにサブコーパスとして管理するほうが使い勝手がよいと考えられる。例えば，検索時間や検索量を試しつつ，必要に応じて，検索対象を調節するといったことが可能になる。

『ひまわり』では，複数の『ひまわり』形式データを個別に検索し，結果を統合する機能がすでに実装されている。ただし，言語資料のインポート時は，常に単一の『ひまわり』形式データとなる。そこで，インポートするファイル群のディレクトリ構造に基づいて，複数の『ひまわり』形式データを生成するように，インポート機能を拡張する。具体的には，インポート対象のファイルを収録したディレクトリをルートディレクトし，その直下のディレクトリごとに『ひまわり』形式データを構築するようにする。

2.2.3　データの利用条件

今回対象とする言語資料の利用条件は，２種類に分けられる。一つは,「源氏物語」のような「改変・再配布可能」である。この利用条件であれば，誰でも変換後の言語資料を配布することができる。一方，CSJ と「毎日新聞」のように「有償・再配布不可」の場合，言語資料の権利者でなければ，変換結果の言語資料を配布することができない。

後者のような言語資料を多くの利用者が『ひまわり』から利用できるようにする対策としては，インポート用の変換規則を配布し，利用者自身がインポートすることである。ただし，現状の『ひまわり』ではインポートはできても,『ひまわり』の設定ファイル自体は，汎用の設定ファイルが自動生成されるだけである。そのため，個々の言語資料が持つ特徴を活かした検索や閲覧がしづらい。そこ

で，インポート時に，指定された『ひまわり』用の設定ファイルが同時にインストールされるように，変換規則と設定ファイルのパッケージ化を図った。

また，これに合わせて，変換後の言語資料を配布する場合のインストールもパッケージ化した。具体的には，従来，インストールするファイルのコピーをユーザが手動で行わなければならなかったが，自動的にインストールできるようにした。

3　インポートの詳細

3.1　CSJ

CSJ は XML で記述されたコーパスであり，利用条件は有償・再配布不可である。1 講演 1 ファイルで合計 3302 ファイルで構成される。有償・再配布不可であることから，言語資料自体の配布を行うのではなく，変換規則を配布する形態を考える。

CSJ はさまざまな形式で配布されているが，本稿では CSJ の「XML 文書」形式のデータ [5] を用いた。この形式のデータには，CSJ に収録されているほとんどの付与情報を含んでいる。ただ，今回は，形態素解析済みのテキストとして利用することを目的にし，「XML 文書」から転記テキスト，および，「短単位・長単位」の情報を抽出して，『ひまわり』形式に変換するようにする。XML で記述されているので，変換は XSLT により行う。変換用のスタイルシート自体は，すでに一般公開済みのものを用いた。詳細は，公開ページ [6] を参照されたい。

CSJ の規模は全体で約 2.9GB あることから，2.2.2 節で述べたように，複数のサブコーパスへ分割する必要がある。分割の単位は，音声タイプ (例：学会講演，模擬講演など) と音声タイプの詳細情報 (例：学会の別，模擬講演テーマの別など)，人手解析・自動解析の別を基準として，合計 16 個に分割した。なお，分割時は 16 個のフォルダに 3302 個のファイルを移動することになり，ユーザの手間が大きい。そのため，ファイル振分け用のシェルスクリプトを用意することにより対処する。

3.2　「源氏物語」

「源氏物語」は，独自形式で記述されたテキストデータであり，1 冊 1 ファイルで計 54 ファイルから構成される。利用条件は無償・再配布可であることから，変換後の『ひまわり』形式のデータを配布する形態を考える。

「源氏物語」のテキストデータの例を図 1 に示す。全体的な構造としては，資料のタイトル，資料説明，本文，作成者情報，本文修正情報から構成されている。本文中の付与情報は，ページ情報，和歌の範囲の２種類である。他の２資料と比較して，特徴的なのは，ページや行の区切りの情報が改行文字で記述されていることである。このようなテキストデータを全文検索する場合，行やページをまたぐ語や表現が検索できなくなるため，『ひまわり』形式データへの変換時に対策が必要である。

『ひまわり』形式データへの変換に際しては，(a) 本文部分の全文検索が確実にできること，(b) 検索文字列が含まれる作品タイトル，ページ位置番号を取得できることを目標として，次のような文字列置換規則を作成した。規則数は 9 個である。結果の一部を図 3 に示す。なお，図中の「→」の左辺は，正規表現で記述された変換対象の文字列，右辺が変換結果の文字列である。

- 1 冊を genji 要素とした。冊のタイトルを genji 要素の title1, title2 属性 (それぞれ漢字表記，原文表記) に変換した。[規則 1]
- ページ区切り位置は page 要素 (空要素) とし，ページ番号はその no 属性に記述した。[規則 2]
- 行区切りのための改行は，改行を表す br 要素 (空要素) とした。また，page 要素前後の改行文字はすべて削除し，1 冊が 1 文字列になるように連結した。これにより，行，ページにまたがる語，表現の検索ができるようになる。[規則 3，5]

[5] http://www.ninjal.ac.jp/corpus_center/csj/manu-f/xml.pdf
[6] http://www2.ninjal.ac.jp/lrc/ 中の『ひまわり』のホームページから，「『日本語話し言葉コーパス』を『ひまわり』で利用する方法 」を参照。

```
米国議会図書館蔵『源氏物語』
桐壺
--------------------------------------------------------------------------------
記号の説明
1. くの字点は／＼で表す。
2. 和歌は「」で括る。
      :
--------------------------------------------------------------------------------

きりつほ

（１オ）
いつれの御時にか女御更衣あまたさふらひ給けるなかに
いとやむことなききはにはあらぬかすくれてときめき
給ふありけりもとより我はと思ひあかりたまへる御かた／＼
めさましき物におとしめそねみ給ふおなし程それより
けらうの更衣たちはましてやすからすあさ夕のみや
つかへにつけても人の心をのみうこかしうらみをおふつもり
にやありけんいとあつしくなりゆき物心ほそけにさとかちに
なるをいよ／＼あかすあはれなる物におほして人のそしりをも
```

図 1: 「源氏物語」のテキストデータ例

```
<genji main_title="米国議会図書館蔵『源氏物語』" title1="桐壺" title2="きりつほ" >
<comment>
--------------------------------------------------------------------------------
記号の説明
1. くの字点は／＼で表す。
2. 和歌は「」で括る。
     :
--------------------------------------------------------------------------------
</comment>
<body>
<page no="０１オ" />いつれの御時にか女御更衣あまたさふらひ給けるなかに<br />いとやむことなききはにはあらぬかすくれてときめき<br />給ふありけりもとより我はと思ひあかりたまへる御かた／＼<br />めさましき物におとしめそねみ給ふおなし程それより<br />けらうの更衣たちはましてやすからすあさ夕のみや<br />つかへにつけても人の心をのみうこかしうらみをおふつもり<br />にやありけんいとあつしくなりゆき物心ほそけにさとかちに<br />なるをいよ／＼あかすあはれなる物におほして人のそしりをも<br /><page no="０１ウ" />えは〻からせたまはす世のためしにもなりぬへき御もてなし也かん<br />たち
```

図 2: 「源氏物語」の変換例

```
# 規則１　genji 要素（開始）と冒頭の comment 要素
^(?s)(.+?)\n(.+?)\n(.*?---+\n.+?---+\n)\n*(.+?)\n
 → <genji main_title="$1" title1="$2" title2="$4" >\n<comment>\n$3</comment>\n<body>\n
# 規則２　page 要素
(?s)（(.{2,3}?)）\n(.+?)\n\n　→ <page no="$1" />$2\n</dummy>
# 規則３　page 要素内の改行を br 要素で置換する
(?s)(?<=<page [^<]{0,30}?>[^<]{0,1000}?)\n(?!</dummy>)　→ <br />
# 規則４　和歌部分を waka 要素とする
(　+「.+?」)　→　<waka>$1</waka>
# 規則５　ページ末尾を br にする
\n+</dummy>　→ <br />
# 規則６　資料末尾のコメントを comment 要素とし，gengi 要素の閉じタグを出力する
(?s)\n(---+\n[^<]+)$　→ \n</body>\n<comment>\n$1\n</comment>\n</genji>\n
```

図 3: 「源氏物語」用の文字列置換規則 (一部)

- 和歌を waka 要素とした。[規則 4]
- 本文を body 要素とし，その前後にある資料の説明，本文，作成者情報，本文修正情報などは comment 要素とした。全文検索時は通常 body 要素内のみを検索するように設定することにより，本文のみを検索対象とすることができる。[規則 1 ， 6]

なお，それぞれの規則は，入力テキストファイルに対して，上から順に一つづつ適用される。「源氏物語」の場合， 1 ファイル 1 冊なので，54 回分独立に規則が適用され，結果は一つの『ひまわり』形式データに合併される。注意すべきことは，規則の適用が 1 行ごとではなく， 1 ファイルを 1 文字列とした状態で行うことである。これにより，複数の行にまたがった文字列に対する置換ができるようになっている。

3.3　毎日新聞

「毎日新聞」は，独自形式で記述されたテキストデータであり，1 年分の記事が 1 ， 2 ファイルにまとめられている。「毎日新聞」の場合，1991 年から 2013 年までの 23 年分のデータが販売されている。利用条件は有償，再配布不可であることから，文字列置換規則を配布する形態を考える。

「毎日新聞」のテキストデータの例を図 4 に示す (日外アソシエーツの Web ページ [7] から引用)。データは，各行がデータ種別とその内容になっている構造である。例えば，「＼Ｔ１＼」で始まる行は記事見出し，「＼Ｔ２＼」で始まる行は記事全文である。記事中に付与情報は記述されていないが，「＼Ｔ１＼」のような形式で記事の情報が記述されている。

```
＼ＩＤ＼００００００１０
＼Ｃ０＼０１０１０１００１
＼ＡＤ＼０１
＼ＡＥ＼Ｎ
＼ＡＦ＼０１０１０１Ｍ０１
＼Ｔ１＼［余録］カンボジアの太陽は日本で見るより大きく見える…
＼Ｓ１＼　　’０１．１．１　朝刊　１頁　写図無　（全７３５文字）
＼Ｓ２＼　カンボジアの太陽は日本で見るより大きく見える。そのせいか１２月でも暑い。日が落ちて午後６時前に空は濃紺に染まり、やがて黒一色になる▲アンコールワット近郊のモンドルバイ村。対人地雷の犠牲になった人々の住む障害者村だ。電気はない。月がないときは真の闇だ。光を奪われた村民のために、ＮＧＯ「国際人権ネットワーク」代表、緒方由美子さんが使いかけのろうそくを集めて贈ることにした話は昨年５月、この欄でご紹介した▲「新聞で見て、雨戸をしめ、真っ暗な中でろうそくを１本つけました。何とホッとすることか。心が安らぎます。私たちも協力させて下さい」と電話をかけてきたお年寄りもいる。こうして日比谷花壇から結婚式のろうそく４０００本、記事を読んだ人、伝え聞いた人から４０００本が集まった▲昨年１２月１日、大勢の人の思いを乗せて、モンドルバイ村６００世帯にろうそくが配られた。
```

図 4: 「毎日新聞」のテキストデータ例

『ひまわり』形式データへの変換では，記事に付与されている情報を検索時に取得できるようにすることを目標にした。図 5 に変換結果を示す。付与情報は，記事を表す at 要素の属性として，発行年・月・日，面種，タイトル，朝夕刊の別，記事文字数などを取り込んだ。文字列置換規則数は，60 個である。「源氏物語」に比べて，規則が多くなったのは，コード化されている面種をデコードしたり，at 要素の数値属性値を半角文字列に統一するなどの置換を行っているためである。

前述のとおり，「毎日新聞」は 1 ファイル (半年，もしくは， 1 年分) のサイズが 300〜600MB と巨大なので， 1 ファイル 1 サブコーパスとして，複数のサブコーパスに分割する。今回は， 3 年分 (4 ファイル) をインポート対象とした。

3.4　インポート結果

表 2 にインポート結果を示す。インポートに使用した PC のスペックは，CPU: Intel Corei5 2.53GHz，メモリ:4GB，OS: MacOS 10.9.3 である。変換自体はどの言語資料も成功した。

[7] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

<at y="2001" m="01" d="01" g="1 面" t="［余録］カンボジアの太陽は日本で見るより大きく見える…" p="朝刊" c="735">
　カンボジアの太陽は日本で見るより大きく見える。そのせいか１２月でも暑い。日が落ちて午後６時前に空は濃紺に染まり、やがて黒一色になる▲アンコールワット近郊のモンドルバイ村。対人地雷の犠牲になった人々の住む障害者村だ。電気はない。月がないときは真の闇だ。光を奪われた村民のために、ＮＧＯ「国際人権ネットワーク」代表、緒方由美子さんが使いかけのろうそくを集めて贈ることにした話は昨年５月、この欄でご紹介した▲「新聞で見て、雨戸をしめ、真っ暗な中でろうそくを１本つけました。何とホッとすることか。心が安らぎます。私たちも協力させて下さい」と電話をかけてきたお年寄りもいる。こうして日比谷花壇から結婚式のろうそく４０００本、記事を読んだ人、伝え聞いた人から４０００本が集まった▲昨年１２月１日、大勢の人の思いを乗せて、モンドルバイ村６００世帯にろうそくが配られた。

図 5: 「毎日新聞」の変換例

最後に，2.1 節で示した三つの観点から結果を評価する。まず，データ形式については，懸案だった独自形式のテキストに対しても，目標に沿った変換を行うことができた。プログラミング言語を用いず，正規表現の置換規則のみで変換できるので，利用者の学習コストを抑えることができると思われる。

データの規模に対しても，サブコーパスのインポート機能を実装したことにより，大規模なデータのインポートが容易になった。ただし，サブコーパスの単位を決める際のファイルの振り分けにシェルスクリプトが必要になるなど課題もある。

利用条件の問題については，利用条件に応じた資料配布形態 (変換規則の配布，変換後の言語資料の配布) に対応するため，配布資料のパッケージ化を図った。今回の三つの資料についての動作は確認している。

表 2: インポート結果

言語資料	入力データサイズ	結果データサイズ	変換時間
CSJ	2.9GB	1.9GB	72min
毎日新聞	1.5GB	0.98GB	25min
源氏物語	2.8MB	2.2MB	0.23min

4　終わりに

本稿では，三つの言語資料を全文検索システム『ひまわり』にインポートし，他のユーザと共有可能な状態にする過程をとおして，(1) 多様なデータ形式に対するインポート能力を確認した，(2) 大規模なデータに対応するためにサブコーパス単位のインポート機能を実現した，(3) 利用条件に応じた配布形態に対応するために，配布資料のパッケージ化を図った。

参考文献

山口昌也, 田中牧郎 (2005) 「構造化された言語資料に対する全文検索システムの設計と実現」，自然言語処理 vol.12, No.4, pp.55–77

今井新悟，赤瀬川史朗，プラシャント・パルデシ (2013) 筑波ウェブコーパス検索ツール NLT の開発，第３回コーパス日本語学ワークショップ予稿集，pp.199–206

小木曽智信，中村壮範，鈴木泰山，八木豊，山崎誠，前川喜久雄 (2011) 「コーパス検索システム「中納言」デモンストレーション」，日本語コーパス完成記念講演会予稿集，pp.43–46

樋口耕一 (2003)「コンピュータ・コーディングの実践 —漱石『こころ』を用いたチュートリアル—」，年報人間科学 24，pp.193–214

松本裕治，浅原正幸，橋本喜代太，投野由紀夫，大谷朗，森田敏生 (2006) 「タグ付きコーパス管理/検索ツール『茶器』」，言語処理学会第 12 回年次大会論文集，pp.460–463

ハ/ガ使用の計量的研究 —有無・量的大小の述語の場合—

A Statistical Analysis of Uses of *wa* and *ga*: Cases with Predicates of Existence/Non-existence and Largeness/Smallness of Quantity

服部 匡（同志社女子大学表象文化学部）
Tadasu Hattori (Doshisha Women's College of Liberal Arts)

要旨

主文での「X {ハ/ガ} P」で、ハとガのどちらが用いられるかを、新聞記事データを用いて計量的に調査する。述語 P としては、「ある / ない」「多い / 少ない」「大きい / 小さい」「強い / 弱い」「高い/低い」の 5 組を取り上げ、各組の両述語での、名詞類 X 別の主題化率 (= ハの用例数 / (ハの用例数+ガの用例数)) の関係を分析する。一般的な傾向としては、大部分の X に対して、各組の前者の述語（「ある」および大値形容詞）の方が、後者の述語（「ない」および小値形容詞）に比べて主題化率が低いのであるが、「高い / 低い」に関してはかなりの例外がある。先行研究を踏まえ、上記の事実への解釈を与えた。

1. はじめに

主題を表すハの働きやそれを含む文の性質については、松下（1930）、久野（1973）をはじめ、多くの研究が過去に公刊されている。また、題目を持つ文と対比して、中心的対象がガで表示される文の性質を問題にする研究も上記のものを含めて数多い。

有無の述語「ある」「ない」の文で、対象を表す名詞句の表示について、次のような事実が知られている[1]。例えば、ある公園で発話するとして、(1a)－(2b) のうち、肯定文で名詞句を主題化した(1b)や、否定文で名詞句を主題化しない (2a) の使用状況は限定されているように感じられる。例えば、(1b) は他の木と対比する状況など、(2a) はクヌギの木があることを期待していた状況などで使用できる。

(1a) クヌギの木がある
(1b) クヌギの木はある (使用状況が限定的)

(2a) クヌギの木がない (使用状況が限定的)
(2b) クヌギの木はない

また、「多い」「少ない」を述語とする文でもある程度上と同様なことが成り立ち[2]、(3b) や (4a) の使用状況は限定されているように思われる（服部 2002)。

(3a) クヌギの木が多い。
(3b) クヌギの木は多い。 (使用状況が限定的)

---

1 存在文におけるハと肯否定の関係については堀口（1995)、丹羽（2006）を、より広いタイプの文でのガと肯否定の関係については三上（1963)、Kuroda (1965)、仁田（1986）などを参照。

2 「ある/ない」の関係と「多い/少ない」の関係が完全に平行するわけではないが、その点は別に論じる。

(4a)　クヌギの木が少ない。(使用状況が限定的)

(4b)　クヌギの木は少ない。

もっとも、この種の、内省に基づく議論は、特定の名詞句と想起しやすい状況の組合せに基づく観察からの一般化になりがちである。実際、寺村 (1988) は、(2a) のパターンに当たる「時間がありません」という文が自然に用いられることを指摘している。「時間はありません」とはあまり言わないように思われる。

そこで、有無(存否)や多寡の文での様々な名詞句に対するガとハの選択、つまり主題化の有無を統計的に調査することが望まれる。本稿では、「ある / ない」、「多い / 少ない」の他に、「大きい / 小さい」、「強い / 弱い」、「高い / 低い」の 3 組の形容詞も含めた統計調査と分析を行う。それらの 3 組は極性反義対 (polar antonym) を構成し、また、「可能性」などの多くの名詞(命題や対象の尺度的属性を表すもの) に対する量的述語となることがある点で「多い/少ない」と共通点を有している[3]。

データのサイズの大きさを重視し、新聞記事[4] (テキストとして約 8GB) をコーパスとして使用する。各新聞記事データに MeCab (0.994) と電子辞書 UniDic (1.3.12) による形態素(短単位) 解析を施し、 (5) に当たる用例を調査対象として抽出した。

(5) 名詞類 ＋ {が/は} ＋ 述語 ＋ 。　(＋の箇所に他要素は介在しない)

名詞類とは名詞・名詞性接尾辞の短単位要素 (書字出現形) [5]である。そのうち例えば「性」は実際には「可能性」「重要性」などの末尾要素である場合もある。また、「よう」「つもり」のような形式的な要素も排除しない。述語は終止形に限る。末尾の句点により当該の例は主文に限られる (ガの場合、いわゆる二重主語文の一部になっているもの、所有文と分析されることのあるものなども含む)。

分析で用いる用語を定義しておく。

(6)　主題化率　＝“X は P”の用例数　/　(“X が P”の用例数 +“X は P”の用例数)

以下、それぞれの述語に対して、名詞類別の主題化率を求め、反義述語の組により二次元にして表示する。その際は、どちらの述語に対しても (ガとハ合わせて) 10 回以上の用例のある名詞類を選ぶ。該当する名詞類が多い場合は、用例数が上位 50 の要素のみを示す。

もちろん、“X__P”でガとハのどちらが選ばれるかは状況 (を踏まえた話者の意図) によって決まるが、状況という変数を直接に調査に取り入れることは、コーパス調査では現実的にできない。しかし X によってよく用いられる状況群の相違があり、それはある程度、用例分布に反映していると仮定する。

ハには対比性の有無、ガにはいわゆる総記性の有無などの点で異なるものがあるが、そうした区別は行わない。

3　第一点に関しては「長い / 短い」なども、また両方の点に関して「濃い / 薄い」なども同じであるが、それらは用例数が少ないため扱わない。

4　毎日新聞 (1999-2005 年)、読売新聞 (1987-92 年、2000-08 年、2010-13 年)、朝日新聞 (1988-1998 年)の記事データ。各新聞社の許諾を得て研究用に利用している。

5　同一形式の意味用法による区別は行っておらず異表記のまとめも行っていない。

2. 主題化率の分布

まず、「ある」「ない」に関する主題化率の分布を図１に示す。

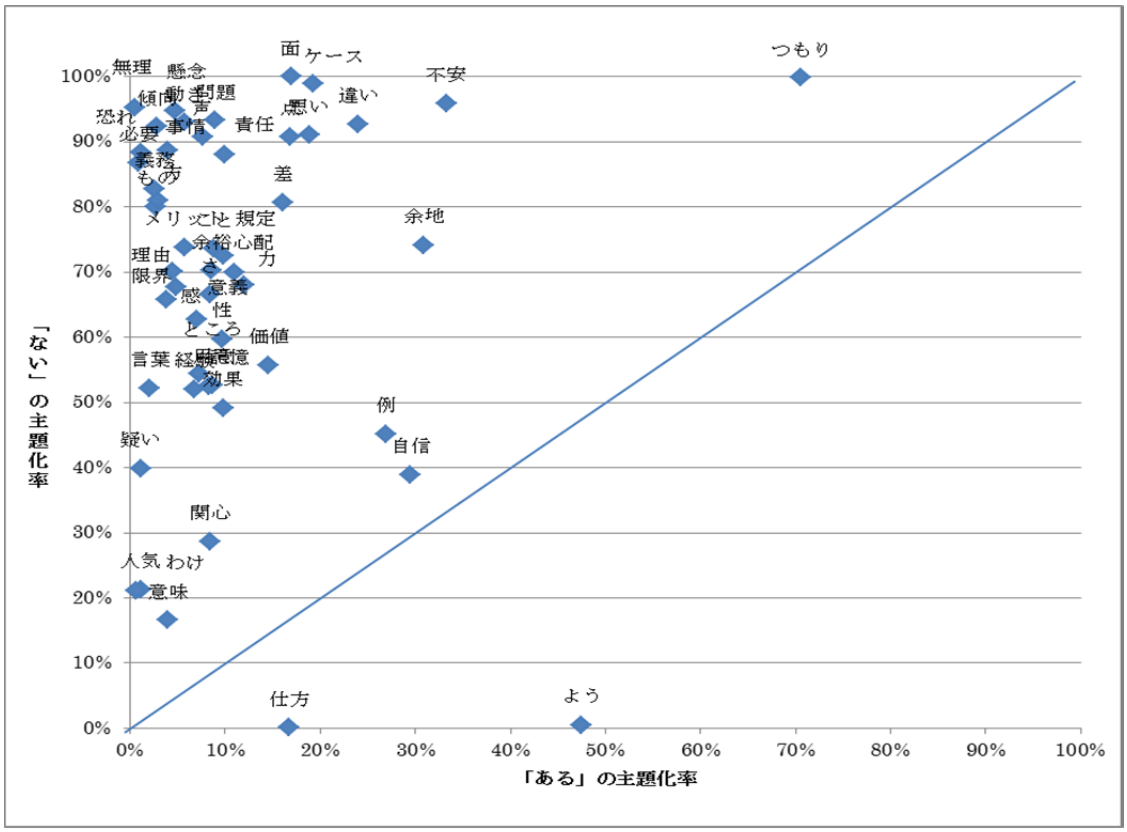


図１　「X__ある」と「X__ある」での主題化率の分布

図に登場する名詞的要素には、「問題」「ケース」など、可算的に用いられるものもあるが、「差」「自信」「関心」「恐れ」など、あまり可算的に用いられないものもある。

図上に示した対角線より上の領域に位置する要素では、「ない」での主題化率が「ある」での主題化率よりも高い。実際には、ほとんどの要素がその領域に位置している。(1)-(4)に示した感覚と矛盾しないことになる。

上記の傾向に、かなり単純化した説明を与えれば次のようになる。たとえて言えば、白紙の上に「X が存在する」絵を描くことは容易であるが「X が存在しない」絵を描くことは困難である[6]。「X が存在する」状態をいったん想定してはじめてそれと X だけの差のある絵を描くことができる。「X がない」のような文は新情報として提示されにくい。一方「X は {ある/ない}」のような文は、X があるかないかをあらかじめ問題として設定した上で「ある/ない」と認定する文であり (丹羽 2006)、不存在を表すのに用いられやすいと思われる。

全体的な傾向に反して「ある」と「ない」での主題化率が比較的近い要素に「自信・意味」などがある。これらは「X がある・ない」で、主体の性質を表している。

ちなみに、用例数が少ないため図にはあがっていないが、寺村氏が問題にした「時間」では、「ある」の主題化率が 31% (用例数 716)、「ない」の主題化率が 34% (用例数 1,018)

6　このたとえは、野矢(2002)を参考にした。

と両者同様に低く、たしかに特徴的な使用傾向を示す要素であると言える。

次に「多い」「少ない」での主題化率の分布を図２に示す。

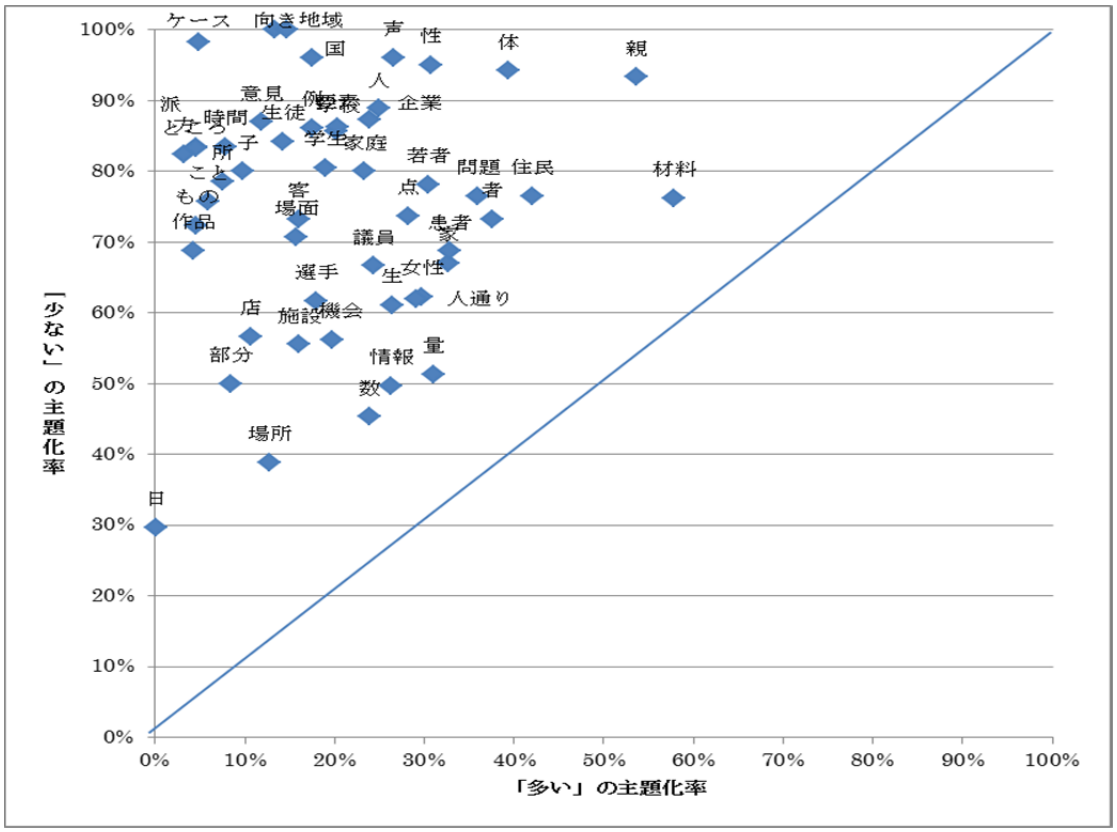


図２　「X__多い」と「X__少ない」の主題化率の関係

登場する名詞類は「ケース」「企業」など可算的な用いられ方をするものが多いが、「数」「量」「(～) 性」など、あまり可算的には用いられないものも一部ある。すべての名詞類に関して「少ない」での主題化率が「多い」での主題化率を上回ることが分かる。

上記の傾向におおまかな説明を与えると次のようになる。いわば、何も存在しない白紙の状態を心理的基盤としても「X が多く存在している」状態を把握することができるが、「X が少ない量しか存在しない」(このように否定文に言い換えられること自体注目される (服部　2002) ことは、X がある程度の量存在している状態の想定を基盤としてしか把握できない。一方「X は｛多い/少ない｝」では、X の多寡があらかじめ問題として設定される。

図 1 と異なる点は、図 1 では「ある」での主題化率は 0%から 10%の間に集中しているが、図 2 では「多い」での主題化率は、10%から 30%の間にあることが多いことである。何かの不存在と対比して存在を述べることはあまりないが、何かの多寡を問題にした上で何かの多さを述べることはそれよりはよくあるということであろうか。

例外的に「多い」と「少ない」での主題化率の近い語に「数」「量」などがあるが、この 2 語は数量そのものを表す名詞であり、「{数/量} が多い」は単なる「多い」と意味が近い。

次に、「大きい」「小さい」での主題化率の分布を図３に示す。

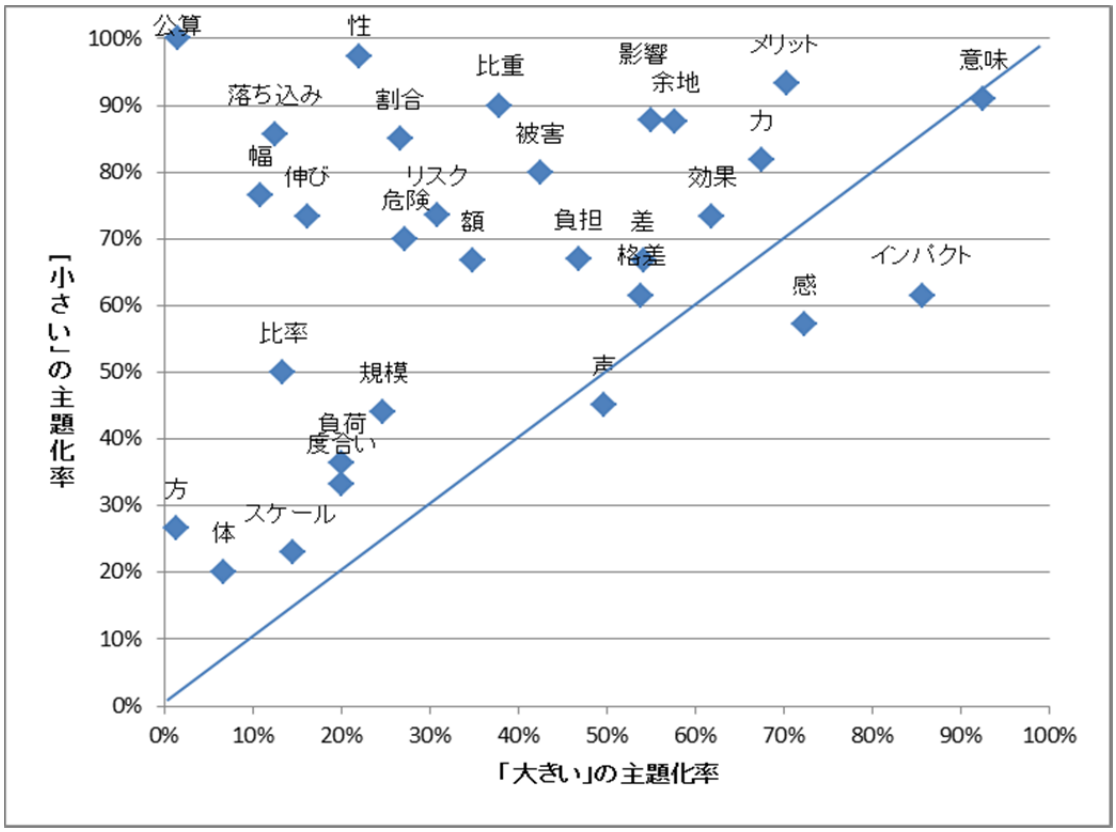


図３　「X__大きい」と「X__小さい」の主題化率の関係

登場する要素の多くは、命題や対象の持つ尺度的属性とみなしうるものを表す不飽和名詞である。例えば、「公算」はあることが実現することの公算であり、「幅」はある物体の幅である。

ここでもやはり、「小さい」での主題化率が「大きい」での主題化率を上回る要素がほとんどで、例外は「意味」「感」「インパクト」「声」の 4 要素のみである（「感」は「存在感」「抵抗感」「期待感」などを構成するものを含む）。中でも「公算」では、「大きい」の主題化率がほぼ 0%、「小さい」の主題化率がほぼ 100%と極端である。ただし、図 2 の「多い」の場合とは異なって、「大きい」での主題化率が 0%から 100%近くにまで幅広く分散していることがこの語対の特徴である。

何かの属性の値が大きいことは、小さいことよりも新情報として提示されやすいということは、ある程度言えそうである。

続いて、「強い」「弱い」、「高い」「低い」での主題化率の分布を図4・5に示す。

図４「強い」「弱い」では、「強い」での主題化率を「弱い」での主題化率が上回る要素が大部分であることは図１－３と変わらないが、その上回りの幅が概して小さい点に特徴がある。“X が弱い”は、多くの場合ネガティブな意味合い（十分に強くない）を持つため期待に反する新情報になり易いのかもしれない。「基盤」「支持」「性」「面」は両述語の主題化率がほぼ等しい。

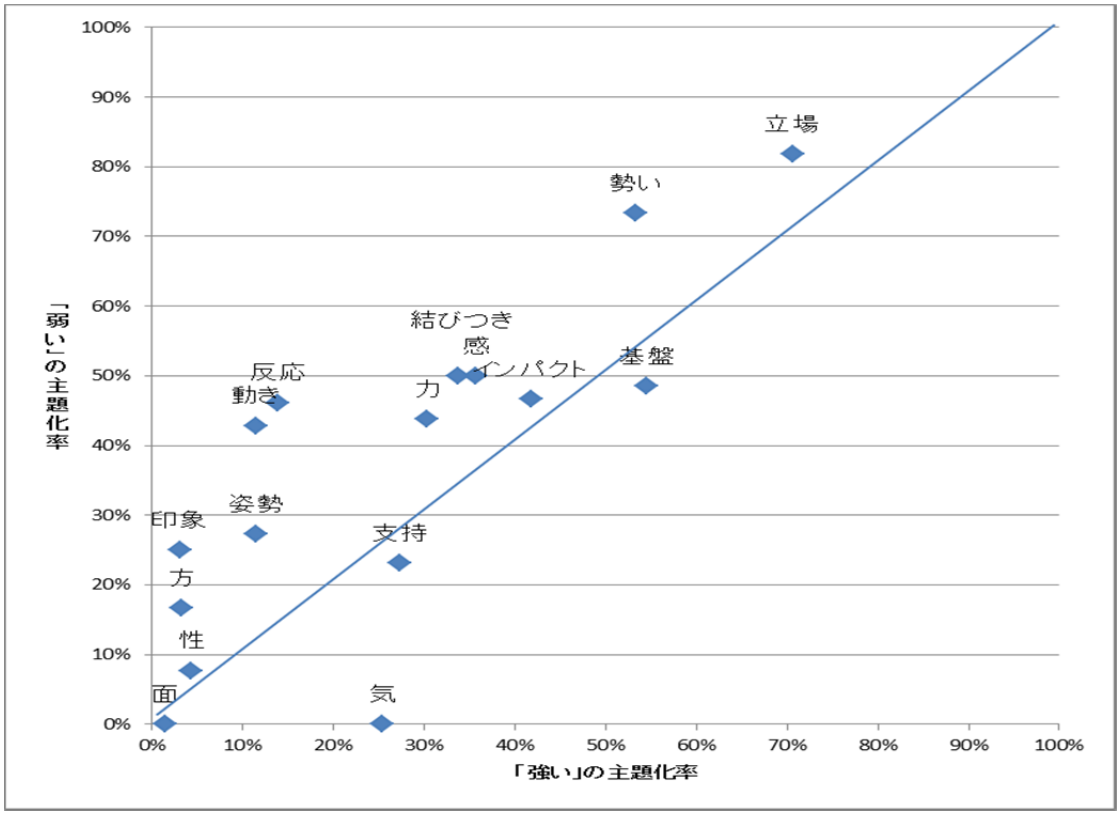


図４　「X__強い」と「X__弱い」の主題化率の関係

図５「高い」「低い」では、出現要素の多くは、やはり、対象や命題の尺度的属性と見なしうる不飽和名詞である。「確率」「比率」「割合」「濃度」「数値」など、数的な値を取るものも多い。

図５の分布は、図４に似ているものの、次の 2 点は異なる 。まず、「高い」での主題化率が「低い」での主題化率を明確に下回る要素が 8 要素あり、また、「高い」と「低い」の主題化率がほぼ等しい要素が 5 要素ある。この形容詞対は、上向き・下向きの方向性と結びつき、質や価値の大きさ/小ささを表すことがある（西尾 1972）こと、量というより段階を言うことがあることと関連する可能性がある。

特に、「性」（可能性、危険性、など）では、「高い」の主題化率の上回りが顕著である。この語は、多くの形容詞対と共起することが知られている（服部 2011）。

なお一部の名詞では、「安い」との関連を考慮すべきである。

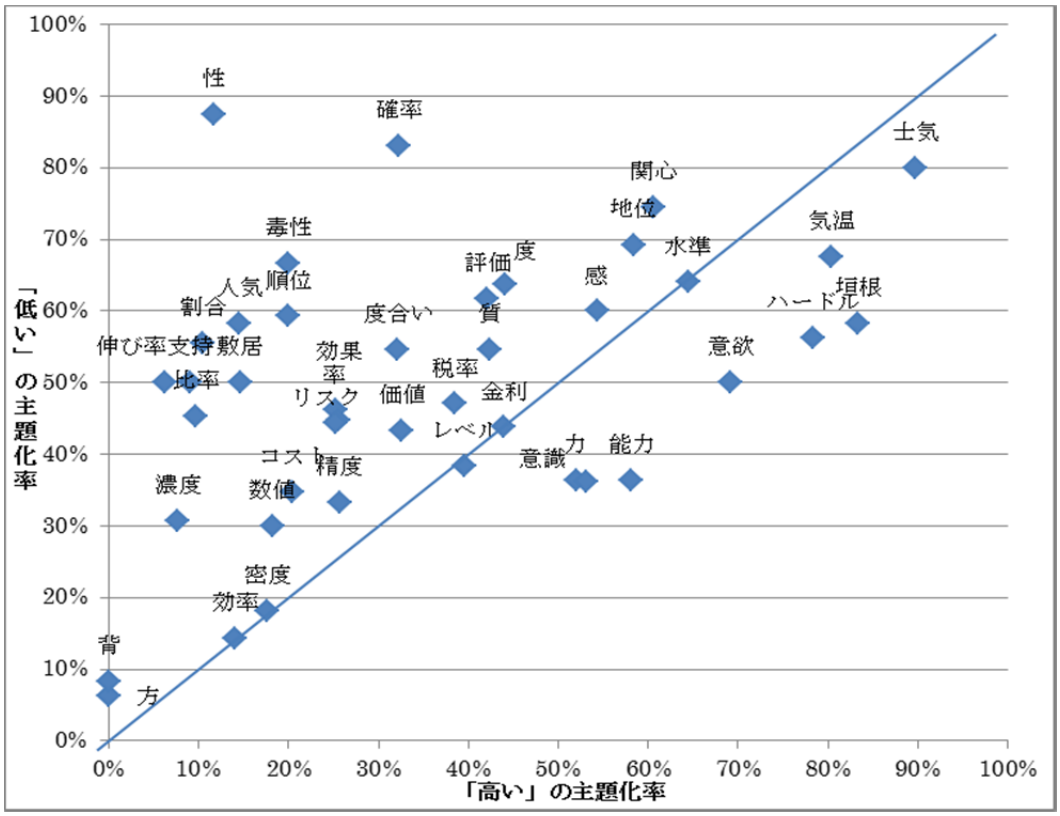


図５　「X__高い」と「X__低い」の主題化率の関係

3.　まとめ

有無の述語および、反義形容詞対を述語とする文で、ガ格が主題化される割合を名詞類別に求め、両述語での値の関係を観察した。およそ、「ある」および大値の形容詞では「ない」および小値の形容詞よりも主題化の割合が低いと言えるが、「高い」「低い」などでは例外もあり、また、「ある」や大値の形容詞での主題化率の分布には述語による相違が見られた。述語別の全体的分布には一応の解釈を与えたものの、個々の名詞類の使用傾向の相違がその名詞類のどのような性質と関連するのかは十分説明できていない。名詞類をグループ化するなどの方法を用いて解釈を試みることが今後の課題である。また、他形容詞等の場合との比較も興味ある問題である。

付　記

本研究は、学術研究助成基金助成金（基盤研究（C）「大規模通時コーパスを用いた発見的研究方法の開拓」、課題番号 26370516）および、国立国語研究所共同研究プロジェクト「コーパス日本語学の創成」による研究成果である。

文　献

久野暲 (1973)　『日本文法研究』 大修館書店.
西尾寅也 (1972)　『形容詞の意味用法の記述的研究』　秀英出版.
丹羽哲也 (2006) 『日本語の題目文』 和泉書院.
野矢茂樹(2002) 『ウィトゲンシュダイン『論理哲学論考』を読む』 哲学書房.
服部匡 (2002) 「多寡を表す述語 の特性について―肯定/否定関係との平行性を中心に」 玉

村文郎編 『日本語学と言語学』.
服部匡 (2004) 「小さな量を表わす表現の意味的性質について」 『言語研究』 124 号.
服部匡 (2011) 「程度的な側面を持つ名詞とそれを量る形容詞類との共起関係—通時的研究—」 『言語研究』 140 号.
堀口和吉 (1995) 『〜は〜のはなし』 ひつじ書房.
松下大三郎 (1930) 『標準日本口語法』中文館.
三上章 (1963) 『日本語の構文』くろしお出版.
寺村秀夫 (1988) 「文法随筆—思い出す学生たち」 『月刊日本語』1 巻 1 号. (『寺村秀夫論文集』 1993 に再録されたものによる)
仁田義雄 (1986) 「現象描写文をめぐって」 『日本語学』 5 巻 2 号.
Kuroda, S.-Y. (1965) *Generative Grammatical Studies in the Japanese Language*. Garland.

ポスター発表　グループB

9月10日（水）　11:00～12:00

コーパス検索による副詞の文中における基本生起位置の検討

難波 えみ（名古屋大学大学院国際言語文化研究科・大学院生）
玉岡 賀津雄（名古屋大学大学院国際言語文化研究科・教授）

A Corpus-Based Investigation of the Canonical Position of Adverbs in a Sentence

Emi Namba (Graduated School of Languages and Cultures, Nagoya University)
Katsuo Tamaoka (Graduated School of Languages and Cultures, Nagoya University)

要旨

日本語の副詞に生起位置があるとする主張がある(Koizumi, 1993)。小泉・玉岡(2006)は陳述、時、様態、結果の副詞を含む文処理実験により、処理負荷が低い生起位置を特定し、Koizumi(1993)の主張を支持した。さらに、副詞に基本生起位置があるとすれば，大規模コーパスでもそれが確認できるはずである。そこで、1991 年から 1999 年までの毎日新聞のコーパス(総語数 273,514,662 語)を用いて、様態の副詞 23 語と結果の副詞 17 語が、目的語を伴って他動詞と共に使われた場合の文中での生起位置を調べた。様態の副詞は、他動詞の前での生起が 3,398 回(50.0%)、与格・対格名詞句の前が 3,288 回(48.4%)、主格名詞句の前が 114 回(1.7%)であった。一方、結果の副詞は、他動詞の前が 908 回(80.7%)、与格・対格名詞句の前が 202 回(18.0%)、主格名詞句の前が 15 回(1.3％)であった。本コーパス研究は、統語理論研究の Koizumi (1993)および文処理実験研究の小泉・玉岡(2006)の主張を支持したが，より厳密に、様態の副詞は、与格・対格名詞句の前後にほぼ同じくらいの頻度で生起するのに対して、結果の副詞は、主に動詞の前に生起することを示した。

１．本研究の目的

動詞の項として要求される名詞句と違い，副詞の文中での生起位置は比較的自由であると言われている。そのため文中における生起位置が変わっても、文の意味が大きく変わることはない。このことから、副詞の文中での生起位置は自由であるとされ、議論の対象になることはあまりなかった。しかし、副詞の生起位置はほんとうに自由なのであろうか。副詞の基本生起位置は存在しないのであろうか。本研究では大規模コーパスを用いて様態と結果の副詞を検索し、生起位置の頻度を算出することで、副詞の基本生起位置を探ることにした。

２．統語理論と副詞の生起位置

2.1 副詞の統語上の生起位置をめぐる議論

三原（2008）は副詞の統語的位置を議論している。例えば、図 1 に表したように、副詞の「急いで」は「(▲）社長は（▲）本社に（▲）応援を（▲）依頼した」という文では、▲のどの位置でも生起可能である。つまり、副詞は、主格名詞句の前後、与格名詞句の後、対格名詞句の後(つまり動詞の前)にも生起することができる。副詞の生起位置が比較的自由であると言われるのはこのためである。三原は、生起位置が自由だとしながらも、「文副詞」「動詞句副詞」に大きく分けている。

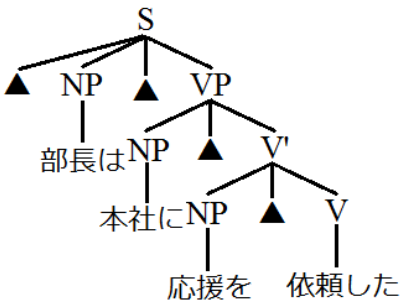


図１　副詞の位置（三原, 2008）

一方、Koizumi（1993）は、統語上の生起位置で副詞を分類し、生起位置から 3 種類の副詞群に分けた。まず、第１に、MP 副詞で、モーダル句（Modal Phrase, MP）内に生起する副詞で、陳述の副詞の多くが含まれる。第２に、IP 副詞で、屈折辞句（Inflection Phrase, IP）内に生起する副詞である。時の副詞や陳述の副詞の一部などが含まれる。第３に、VP 副詞であり、動詞句（Verb Phrase, VP）内に生起する副詞で、様態や結果の副詞が含まれる。これらの 3 種類の副詞群を統語図で示したのが、図 2 である。

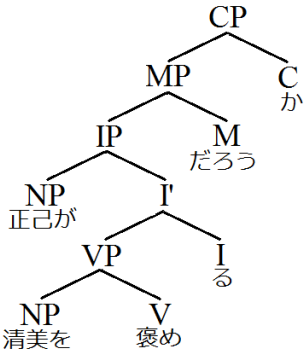


図２　日本語の統語構造（小泉・玉岡、2006 より）

2.2 文処理実験による副詞の基本生起位置の証明

文処理実験により、文中における副詞の基本生起位置の証明もなされている。小泉・玉岡(2006)では日本語の副詞類の文中での生起位置を変えた文を 24 名の被験者に示し、文正誤判断課題に要する反応時間を測定して、副詞の基本生起位置を検証した。反応時間が短いほど処理負荷が低いと想定され、それに基づいて副詞の基本生起位置を判定した。その結果は、表１の通りである。小泉・玉岡(2006)の文処理実験により判定された副詞の生起位置は、Koizumi (1993)の副詞の基本生起位置と一致し，仮説を実証した。

３.仮説

Koizumi (1993)の統語理論と小泉・玉岡(2006)の実験結果から、様態と結果の副詞の生起位置はともに，表１のように動詞句内(VP 副詞)であった。そこで、さらに本研究では様態の副詞と結果の副詞を対象に、副詞の基本生起位置をコーパス研究で検証する。様態と結果の副詞は、VP 副詞に分類され，動詞句内に基本生起位置を持つ。もしそうであれば、コ

ーパスにおける生起頻度も動詞句内に多く見られるはずである。

表 1　小泉・玉岡（2006）により証明された副詞の基本生起位置

統語的位置	副詞の種類	基本生起位置	文例
MP 副詞	陳述	Adv S O V	あいにく太郎が学校を休んだ。
IP 副詞	時	Adv S O V	今日次郎が髪を切った
		S Adv O V	太郎が昨日花瓶を壊した。
VP 副詞	様態・結果	S Adv O V	次郎がすばやく靴下を洗った。
		S O Adv V	太郎がグラスをこなごなに割った。

大規模コーパスの検索により、目的語を伴う他動詞のみを選んで副詞の生起位置別に出現頻度を計算した。他動詞句内に現れる場合は、他動詞の前[$_{VP}$ *Adv* V]と対格・与格名詞句の前[$_{VP}$ *Adv* [$_{VP}$ NP V]]の２通りの生起位置が考えられる。また、基本生起位置ではないが，主格名詞句の前[$_{S}$ *Adv* NP [$_{VP}$ NP V]]にくることも考えられるので，この 3 つの文中での位置について頻度を計算し、様態と結果の副詞が VP 副詞であるかどうかを検証した。

４.検証

4.1　コーパス

検索には、1991 年から 1999 年に発行された毎日新聞の 9 年間分を使った。総語数は、273,541,662 語である。検索には、パデゥー大学の深田淳が作成した検索エンジン「茶漉」を用いた。

4.2　検察した副詞と手順

検索の対象とした副詞は、小泉・玉岡（2006）から様態の副詞 23 語と結果の副詞 17 語である。検索した副詞を以下に示す。

様態の副詞 (*N*=23 語): ゆっくり、ちびちび、こっそり、そっと、もりもり、さっさと、テキパキ、ペラペラ、せっせと、ころころ、すばやく、ボキッと、きっぱり、こわごわ、ぼんやり、じっと、のんびり、さらりと、どんどん、難なく、うまく、のろのろ、熱心に
結果の副詞 (*N*=17 語): こなごなに、かちかちに、ペシャンコに、細かく、細く、星形に、ばらばらに、人肌に、柔らかく、かたく、パリパリに、びしょびしょに、どろどろに、カリカリに、熱く、まるく、ピカピカに

これらの副詞を検索した後、対象とした副詞を含む他動詞文を副詞の生起位置で分類した。対象となった副詞を含む他動詞の文の数は、様態の副詞が 6,800 文、結果の副詞が 1,125 文であった。

4.3 検索結果と分析

様態と結果の副詞の文中での生起位置の頻度と割合は、表２に集計した通りである。様態の副詞は、他動詞の前([$_{VP}$ *Adv* V])での生起が 3,398 回(50.0%)、与格・対格名詞句の前([$_{VP}$ *Adv* [$_{VP}$ NP V]])が 3,288 回(48.4%)、主格名詞句の前([$_{S}$ *Adv* NP [$_{VP}$ NP V]])が 114 回(1.7%)であった。一方、結果の副詞は、他動詞の前が 908 回(80.7%)、与格・対格名詞句の前が 202 回

(18.0%)、主格名詞句の前が 15 回(1.3%)であった。様態と結果の２種類の副詞と３つの文中での生起位置について、２×３のカイ二乗分布を使った独立性の検定(母比率の検定とも呼ばれ、比較的絶対頻度の影響を受けない)を行った。その結果、副詞の種類と生起位置に有意な独立した関係がみられた[$\chi^2(2)=371.12, p<.001$]。さらに、５%有意水準である 1.96 の絶対値を残差の基準として、３つの生起位置と２種類の副詞の６つのセルを比較した。その結果、様態の副詞は与格・対格名詞句の前後にほぼ同じくらいの頻度で生起するのに対して、結果の副詞は主に動詞の前に生起することが分かった。なお、主格名詞句の前にこれらの副詞が生起するのは，両副詞共にわずかに２%以内であった。主格名詞句(主語)の前にこれらの副詞句が生起することはほとんどなく、基本生起位置でないことを示した。以上のように、副詞にも、文中で最適とされる生起位置があることが証明された。

表2　他動詞と副詞の統語構造上の位置における頻度と割合

副詞の種類	[vP NP [VP Adv V]]		[vP Adv [VP NP V]]		[S' Adv [S NP [VP NP V]]]	
	頻度	割合(%)	頻度	割合(%)	頻度	割合(%)
様態(N=23)	3398	48.4	3288	50.0	114	1.7
結果(N=17)	908	80.7	202	18.0	15	1.3

５．結論

本コーパス研究は、副詞にもそれぞれに特定の基本生起位置があることを、より厳密に実証した。統語理論研究 Koizumi （1993）と三原（2008）および文処理実験研究の小泉・玉岡（2006）の主張通り、様態と結果の副詞は共に動詞句副詞であることが確認できた。さらに、動詞句内での生起位置は、様態の副詞が図３の(a)と(b)の位置である対格・与格名詞句の前後にほぼ同じくらいの割合で生起していた。一方、結果の副詞では、図３の(a)の対格・与格名詞句の後に 80%以上の割合で生起していた。様態の副詞は vP と VP の Spec，結果の副詞は VP の Spec が基本生起位置であると言えよう。先行研究（Koizumi, 1993; 小泉・玉岡, 2006; 三原, 2008）は，様態および結果の副詞が，動詞句副詞であることを示したが，本コーパス研究は，より厳密に様態と結果の副詞の動詞句内での基本生起位置を特定した。

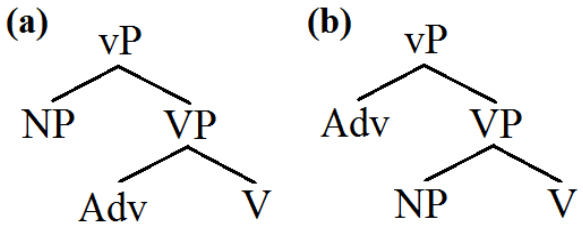


図３　様態と結果の副詞の基本生起位置

引用文献

Koizumi, Masatoshi (1993). Modal phrase and adjuncts. *Japanese/Korean Linguistics, 2*, 409-428

小泉政利・玉岡賀津雄 (2006). 文解析実験による日本語副詞類の基本語順の判定. *認知科学, 13*(*3*), 392-403.

三原健一 (2008). *構造から見る日本語文法*. 東京: 開拓社.

BCCWJ と日英パラレル新聞コーパスに基づいた 格外連体修飾形の研究

田邊　和子（日本女子大学文学部）†

Study of the Case-Outer Relative Clauses Based on the BCCWJ and Japanese-English Newspaper Parallel Corpus

Kazuko Tanabe (Japan Women's University)

要旨

本研究は、日本語の格外連体修飾形（いわゆる「外の関係」）を対象に、BCCWJ と日英パラレル新聞コーパスに基づいて使用実態を明らかにしようとするものである。日本語の格外連体修飾形は、その説明機能によって抽象的内容を叙述するのに有効に使われているにもかかわらず、質量両側面からの包括的研究は、従来、本格的になされてこなかった。

本研究では、日本語教育への応用を視野にいれて、BCCWJ と日英パラレル新聞コーパスの二つのコーパスを駆使することによって、格外連体修飾形の使用の実態を明確にしてみたい。今回の発表では、特に被修飾名詞を直前の動詞の形式に着目し、類型化を試みることにする。

１．はじめに

日本語の連体修飾節の研究は、主に、寺村（1975‐1978）による「内の関係」「外の関係」の区分に始まり、高橋（1979）、加藤（2003）、大島（2010）によって分析が進められてきた。寺村による被修飾名詞が、修飾節の中で「文に開けるか」すなわち、英語の関係代名詞のように格関係を持っているか否かという基準によって、日本語の連体修飾節を分類したことに関して、加藤は疑問を唱えたが、大島（2010：8）に至って「名詞のもつ意味的情報を鑑みることなしに日本語の連体修飾構造を考察することはできない」という見解は、これまでの一連の研究の結果として説得性のあるものである。一方、海外に目を向けると、Comrie (1996, 1998, 2010)が、格外連体修飾形を“Asian-type”(Comrie: 1996)と定め Indo-European languages における relative clauses とは基本的に異なると記述している[1]。

しかし、いずれの研究者も検証した例文は、作例だったり、他人論文からの引用や個人的な収集作業によるものであった。また、コーパス使用といっても語の検索機能で例文を引き出すといった作業によるもので、その使用例の「代表性」を認証することはできなかった。コーパス言語学では、「代表性」「均衡性」を確認できることが従来の言語学と比較して優れた点で、特に「代表性」は、使用例全体における比重をコーパス内の頻度を知ることによって、その使用例が全体の中でどれほど代表的であるか判断できる。本研究では、BCCWJ の使用により、特に「格外連体修飾形」に焦点を当てて、その使用実態を明らかにしたい。また、日英パラレル新聞コーパスをも利用して、連体修飾形が英語ではどのように訳されるか、格外連体修飾形の英訳にどのような文法的差異が見られるか考察して

† tanabeka@fc.jwu.ac.jp

1 Whitman（2011）The relative clause problem （Oxford 発表スライド）は、異なった見解を示している。

みたい。

２．格外連体修飾形に頻繁に使われる被修飾名詞の抽出

下の表は、BCCWJ コアデータから動詞もしくは助動詞(いずれも連体形)の直後に名詞がくるものを中納言により検索し、先行する動詞数を基準として用例数の多い名詞順にリストアップしたものの一部である。

表 1　中納言コアデータによる動詞・助動詞連体形接続名詞頻度順リスト

名詞	修飾動詞数	修飾助動詞数	名詞	修飾動詞数	修飾助動詞数
こと	3528	1564	うち	68	25
ため	1058	222	前	67	13
もの	564	791	意味	63	16
人	474	374	予定	63	21
わけ	246	106	点	61	38
必要	190	9	中	60	33
場合	186	220	地域	59	35
とき	177	249	方法	59	27
ところ	164	225	言葉	58	38
はず	121	61	理由	56	50
事	120	88	調査	55	20
時	112	164	方針	55	6
者	107	68	際	55	37
情報	87	82	企業	49	40
方	81	129	問題	48	57
つもり	79	14	話	46	52
ほか	75	45	声	46	24
一方	72	11	気	45	93

表 1 によると[2]、形式名詞が上位を占め、「一方」や「前」などの相対名詞が続き、普通名詞としては、「必要」「情報」「意味」が連体修飾形における被修飾名詞となりやすいことがわかる。動詞連体形か助動詞連体形のどちらが共起しやすいかという点では、動詞の比重が大きい「こと」「ため」「はず」「つもり」「一方」に対して、助動詞接続が多いのは「とき」「ところ」「時」「方」「気」が挙げられる。「〜した」の「た」も助動

2 検索式例(動詞の連体形の直後に名詞がくるもの)は,以下のようになる。

キー: (品詞 LIKE "動詞%" AND 活用形 LIKE "連体形%") AND 後方共起: 品詞 LIKE "名詞-普通名詞%" ON 1 WORDS FROM キー DISPLAY WITH KEY IN (registerName="出版・新聞" AND core="true") OR (registerName="出版・雑誌" AND core="true") OR (registerName="出版・書籍" AND core="true") OR (registerName="特定目的・白書" AND core="true") OR (registerName="特定目的・知恵袋" AND core="true") OR (registerName="特定目的・ブログ" AND core="true") WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

詞となるので、ここでは動詞の「～する（動詞の連体形）」以外の形も取ることが多いということを意味する。

表 2 は、日英パラレル新聞コーパスの日本語コーパス部分で、「する」「ている」「ていた」「される」「させる」のそれぞれ後に続く語（1 R）を検索し、用例数の頻度順にリストにしたものである。被修飾名詞によって、連体節内の文法形式に特徴があることがわかる（例：～する方針、～ていた疑い、～される見通し、～させる必要）。

表 2　日英パラレル新聞コーパスにおける被修飾名詞と文法形式のつながり

する		ている		ていた		される		させる	
用例数	1R	用例数	1R	用例数	1R	用例数	1R	用例数	1R
4764	。	19372	。	3428	。	1379	。	352	こと
4647	こと	1617	の	779	こと	403	こと	177	ため
2044	ため	1489	こと	530	が	283	の	157	。
1692	の	1358	が	438	と	152	と	91	の
1494	と	1143	と	254	の	132	よう	46	に
1058	よう	855	」	140	」	110	見通し	45	よう
1033	方針	325	ため	93	疑い	103	が	44	と
776	」	315	よう	77	ため	95	べき	44	べき
686	など	197	か	65	「	85	「	43	必要
627	「	193	「	58	もの	61	予定	41	方針
605	か	185	から	42	から	60	など	32	」
572	必要	127	、	39	として	51	可能	28	か

下記表 3～10 は、BCCWJ コアデータから着目する名詞を含む文節に係っている直前の文節の末尾に表れる動詞・助動詞(連体形)の用例数を茶器により検索したものである。それぞれに用例数の多いもの 6 語をリストアップしている(同数のものは茶器 word list での並び順による)。表 3～7 の名詞は、表 1 の中にあるものを選び出した。表 1 には含まれていないが、表 8~10 の名詞も参考として同様に調べている。表 8 の「システム」のような外来語も格外連体修飾形の対象になるので検索してみた。表 1 と同じ調査で、「システム」を修飾する動詞の用例は 35、助動詞は 13 であり、リスト中で比較的順位の高い外来語である。

表 3「必要」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		190	100		10	100
1	する	76	40	せる	3	30
2	いく	21	11.05	な	2	20
3	おく	9	4.74	れる	2	20
4	行う	4	2.11	させる	1	10
5	作る	4	2.11	た	1	10
6	図る	4	2.11	られる	1	10

表 4　「地域」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		68	100		29	100
1	する	15	22.06	た	14	48.28
2	いる	13	19.12	な	10	34.48
3	よる	8	11.76	ない	2	6.90
4	ある	6	8.82	れる	2	6.90
5	なる	3	4.41	たい	1	3.45
6	行う	3	4.41			

表 5　「理由」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		57	100		50	100
1	いる	16	28.07	た	21	42
2	いう	12	21.05	な	13	26
3	する	7	12.28	ない	11	22
4	考える	5	8.77	たる	1	2
5	こだわる	2	3.51	てる	1	2
6	ある	1	1.75	でる	1	2

表 6　「問題」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		74	100		69	100
1	いる	14	18.92	な	31	44.93
2	いう	13	17.57	た	24	34.78
3	関する	8	10.81	べき	5	7.25
4	する	4	5.41	ない	3	4.35
5	ある	3	4.05	る	3	4.35
6	かかわる	2	2.70	れる	2	2.90

表 7　「気」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		42	100		84	100
1	する	10	23.81	な	72	85.71
2	いう	6	14.29	た	8	9.52
3	いる	6	14.29	たい	1	1.19
4	かける	1	2.38	って	1	1.19
5	くる	1	2.38	てる	1	1.19
6	しまう	1	2.38	れる	1	1.19

表 8（参考）　「システム」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		63	100		25	100
1	する	25	39.68	な	14	56
2	できる	12	19.05	た	11	44
3	いう	4	6.35			
4	いく	2	3.17			
5	なる	2	3.17			
6	図る	2	3.17			

表 9（参考）　「感じ」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		27	100		78	100
1	いう	12	44.44	た	26	33.33
2	いる	4	14.81	な	21	26.92
3	言う	3	11.11	って	18	23.08
4	いく	1	3.70	てる	7	8.97
5	つき合える	1	3.70	ない	3	3.85
6	とろける	1	3.70	れる	2	2.56

表 10（参考）　「結果」に係る動詞・助動詞

	動詞	用例数	Ratio(%)	助動詞	用例数	Ratio(%)
TOTAL		33	100		51	100
1	いう	14	42.42	た	42	82.35
2	する	4	12.12	な	4	7.84
3	関する	3	9.09	だ	2	3.92
4	ある	1	3.03	る	2	3.92
5	いる	1	3.03	ない	1	1.96
6	かかる	1	3.03			

この一連の表の結果を考察すると、被修飾名詞には、大きく分けて「する」形と結びつきやすいもの（例：必要)、「～た」形と結びつきやすいもの（例：地域・結果）その他「～べき問題」「～な気」「～できるシステム」といったある特定の表現との結びつきを窺わせる名詞の 3 つのグループに大きく分けられる。さらに、「という」との接続が、高い比率を示していたり（例：感じ・結果)、否定形「ない」との接続比率が高いという特徴を持つ語もある（例：理由)。このように格外連体修飾形の機能については、被修飾名詞の内容を説明するというのが従来の一般的な見解であった。本研究でのコーパス利用によって被修飾名詞の意味に従って接続している文法形式もそれぞれ特徴があることが明確になった。

３．節単位による格外連体修飾形の分析

本項では、連体修飾節全体を視野に置き、コーパスによる使用実態の考察を試みること

にする。

Left	Center	Right
ているとか「進歩している」とか、そういった	理由	で侵略する権利を認めることは、断じてでき
の身の回りの世話をしているから」といった	理由	で同居する「非パラサイト型」の人は合わせ
親を介護する、親の家業を手伝うといった	理由	で親と同居している人も含まれている。
私が電話をかけた	理由	を私はちゃんと知っている。
登場し、今日まで連綿と生きつづけてきた	理由	はいくつか考えられますが、その一つとして
1審より減刑した	理由	について安広裁判長は、詐欺の被害者であ
意識実態調査」（二千三年）では、結婚した	理由	は何ですかという問に対し、「経済的理由か
入社を辞退した女性であること、辞退した	理由	のこと、そういうわけで結婚までに間をおく
のこだわり、つまり標準仕様として選択した	理由	は、明確でした。
SIがあらゆる産業に使われるようになった	理由	は、もちろん今までなかった新しい機能を持
ように、外敵に強いことも絶滅をまぬかれた	理由	といえます。

図 1　「～た理由」用例

上記図 1 は BCCWJ のコアデータから、表 5 で検索した、「理由」に係る助動詞の用例をエクセルに出力したものの一部である。各用例の同一のキーワードを中心に揃え左右に文を拡げた KWIC(Key Word In Context)と呼ばれる表形式である。代表用例を抽出してみた。

「～た理由」代表的用例の抽出

a. ～といった理由で、
b. （辞退した/ 結婚した/ 選択した）理由は、
c. 使われるようになった理由は、
d. 絶滅をまぬかれた理由

このように実際に使われている表現を客観的に収集できることは、コーパス利用の成果ならではのことである。

次の図 2 は日英パラレル新聞コーパスから語や句の用例を検索できるパラレルコンコーダンサーの WebParaNews[3] で「～た理由」を検索した画面の一部である。1~5 における日英対応文における理由の表現を確認してみたい。

1. 理由の叙述としては、’citing insufficient measures~’ における現在分詞 citing が有効に使われている。
2. ’argue that~’という表現の使用によって「（that 以下の内容）が理由で反対した」という訳になっている。
3. Because 節によって理由が説明されている。
4. ’for some peculiar reason, Ota reversed his position’が、「反対に転じた理由が不透明」という格外連体修飾形に対しての訳で、「不透明な理由で、反対に転じた」という逆パラフレーズ（パラフレーズの解体作業）の一例となっている。
5. ’attribute　to~’ を使って「～た理由については、」を訳している。

[3] http://www.antlab.sci.waseda.ac.jp/webparanews/

1	税制 の 是正 が 不十分 」と いっ た 理由 の ほか 、 来年 の 参院 選
2	まで 日 韓 基本 条約 を 認め なかっ た 理由 として 、 同 条約 が 、 韓国
3	制服 着用 に 踏み切っ た 理由 の 一つ は 、 生徒 たち が
4	反対 に 転じ た 理由 は 不透明 だ 。
5	率 六・二 % より 一・五 ポイント も 高かっ た 理由 について 、 厚生省 は 、 四 年
1	SDPJ Secretary General Wataru Kubo and other SDPJ members voiced opposition against handling the bills in one package, citing insufficient measures to correct inequalities in the tax system and the effect on a House of Councillors election next year.
2	The SDPJ has argued that the treaty recognizes South Korea as the only legitimate government on the Korean Peninsula.
3	Long Beach introduced uniforms because students were attacked by gang members on a number of occasions when they were wearing T-shirts or hooded sweat shirts bearing certain patterns or emblems, according to board of education officials.
4	However, for some peculiar reason, Ota reversed his position and now opposes transfer of the base's functions within the prefecture.
5	The ministry attributed the 1.5 percentage point rise in fiscal 1992 to diagnostic examination and treatment fee hikes and an increase in interferon prescriptions for hepatitis treatments, the officials said.

図 2　「～た理由」検索画面

1	的 自衛 権 の 行使 を 可能 に する 必要 が ある 。
2	を 抑圧 する 過剰 な 規制 を 削減 する 必要 が ある 。
3	厳格 な 検査 体制 を 速やか に 構築 する 必要 が ある 。
4	憲法 を 超える 視点 で 改革 を 議論 する 必要 は ない の か 。
5	核 拡散 を 何 として で も 阻止 する 必要 が ある 、と の 共通 の
1	Japan's use of collective self-defense should be allowed so the Japan-U.S. alliance, which has won international confidence as the stabilizer of the world, can function effectively.
2	We need to cut back excess regulation, which suppresses innovation, enterprise and creativity.
3	To reassure consumers, a rigorous inspection system should be put in place to ensure that not one cow with BSE reaches the marketplace.
4	It may be necessary to discuss how these institutions might be changed, too.
5	But the meeting ended up reaching consensus on the issue, chiefly because all members now share the view that it is essential to prevent nuclear proliferation at all costs.

図 3　「～する必要」検索画面

上記図 3 は WebParaNews で「～する必要」を検索した画面の一部である。「～する必要（がある）」の訳においては、2 ‘need’　4 ‘necessary’　5 ‘essential’ のように意味的に直接関係する単語が使用されているので、「～た理由」ほどは、多様な英語表現は使用されていない。これは、「する必要」という格外連体修飾形というよりも、「～する必要がある」という表現として使用されることが多いためで、1 と 3 における助動詞 should が「～する必要がある」の訳として使われやすいこともこのためだろう。

４．まとめ

格外連体修飾形の機能については、被修飾名詞の内容を説明するというのが従来の一般的な見解であった。本研究では、コーパスを利用することによって被修飾名詞の個別使用頻度に従って、その「代表性」を把握し、それぞれが接続している文法形式の特徴も明確にすることができた。その指標となる主なものは「する形」「た形」「～な形」である。これらは、名詞の意味によってさまざまな特徴を示していることが判明した。また、代表用例を抽出することも容易になり、使用実態を文単位でも把握できることを証明した。また、日英パラレルコーパス利用によって、名詞修飾節の日英対照比較研究を、実践的に行うことができ、特に翻訳の分野では多種多様な例を日英両方で把握することができる可能性を示唆した。

謝　辞

本研究は、文部科学省科学研究費補助金、基盤（C）課題番号 25370496 (研究代表者：田辺和子) による補助を得ています。

文　献

中條清美、アントニ・ローレンス、西垣知佳子(2012)「日英パラレルコーパス検索サイト WebParaNews の公開－開発と実践利用－」, 外国語教育メディア学会（LET）第 52 回全国研究大会, 甲南大学, 岡本キャンパス, 発表要項集, pp.94-95.

Comrie, Bernard. (1996) The unity of noun modifying clauses in Asian languages. *Pan-Asiatic Linguistics*: Proceedings of the Fourthe International Symposium on Languages and Linguistics, January 8-10, 1996, Volume 3, pp.1077-1088.

Comrie, Bernard. (1998) Rethinking the typology of relative clauses. *Language design*. pp.59-86.

Comrie, Bernard. (2010) Japanese and the other languages of the world. *NINJAL project review1*. pp.29-45.

Frellesvig, Bjarke&John Whitman. (2011) Prenominal complementizers and the derivation of complex NPs in Japanese and Korean. In William McClure(ed.) *Japanese/Korean linguistics* 18, pp.73-87. Stanford: CSLI.

加藤重広(2003)『日本語修飾構造の語用論的研究』ひつじ書房

Kawaguchi, Yuji(eds.). (2007) *Corpus-Based Perspectives in Linguistics.* John Benjamins. Amsterdam/Philadelphia.

Matsumoto, Yoshiko. (1988) Semantics and pragmatics of noun-modifying constructions in Japanese. *Berkeley Linguistics Society* 14, pp.166-175.

大島資生(2010)『日本語連体修飾節構造の研究』ひつじ書房

高橋太郎（1979）「連体動詞句と名詞のかかわりについての序説」高橋太郎（1994）むぎ書房所収

寺村秀夫(1975-1978) 「連体修飾のシンタクスと意味(1)-(4)」寺村(1992)所収

寺村秀夫(1992)『寺村秀夫論文集 I―日本語文法編―』くろしお出版

丹羽哲也(2013)「連体修飾における基本形とタ形の対立」藤田保幸編『形式語研究論集』和泉書院

テキストにおける多義語の意味の集中度

山崎　誠（国立国語研究所言語資源研究系）†

The Concentration Ratio of Polysemous Senses in Japanese Text

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

要旨

テキストにおいて多義語の意味が特定のひとつに集中して用いられる傾向があることは既に Gale et al.(1992)らによって指摘されているが、先行研究では特定の何語かを取り上げて、その出現傾向をさぐるものがほとんどであった。本研究はテキストに現れた全ての名詞について、意味の集中の度合いを測定しその結果を報告するものである。その結果、当該テキストの話題によって特定の意味に集中する傾向はあるものの、実質的な意味と形式的な意味が共存している場合は、複数の意味が生じやすいことが分かった。このことは、テキストにおける結束性は主に語彙的な部分で働き、文法的には働かないことを意味していると解釈される。

１．語彙的結束性

語彙的結束性（lexical cohesion）は、テキストを成立させる重要な条件として Halliday & Hassan(1976)によって提唱され、テキストにおける同一の語の繰り返し使用などについて計量的研究が行われてきた。例えば、多義語については、山崎(2010:30)において「テキストにおける多義語の意味実現が一つの意味に偏りやす」く、その偏りは「出現間隔が近いほど起こりやすい」ことが指摘されている。多義語の意味実現は、語彙的結束性という概念を用いず、自然言語処理の観点からの研究が早くから行われており、上掲の Gale et al.(1992)に対して、細かな意味の違いを考慮すれば、複数の意味実現が見られるという指摘もある（Krovetz(1998)）。本稿は従来のアプローチとは違い、テキスト全体における多義語の意味分布を探る試みである。そのことにより、語彙的結束性のあり方を把握すると同時に、多義の意味分布から見たテキストの特性についても考察する手がかりとする。

２．データ

本稿では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）の中の図書館書籍（LB）のデータを利用した。「BCCWJ 短単位語数」のページ[1]で公開されているファイル BCCWJ_WC_SUW_v10.xlsx を利用し、LB で空白、記号、補助記号を除外した可変長部分の短単位数の中央値が 2360 であることから、2360±50 短単位の範囲から NDC の異なるサンプルをランダムに抽出した。使用したサンプルの概要を表 1 に示した。LBd0_00003 は、クロワッサンのレシピを説明したエッセイ、LBj4_00045 はアリの生態を解説した科学書、LBn9_00088 はチェスのプレイヤーを主人公にした翻訳小説である。

† yamazaki@ninjal.ac.jp

[1] https//maro.ninjal.ac.jp/wiki/index.php?BCCWJ%2F%E7%9F%AD%E5%8D%98%E4%BD%8D%E8%AA%9E%E6%95%B0

表 1 使用したサンプル

サンプル ID	NDC	タイトル	著者	出版社	出版年	短単位数
LBd0_00003	0 総記	空飛ぶフランスパン	金子郁容(著)	筑摩書房	1989	2403
LBj4_00045	3 自然科学	アリはなぜ一列に歩くか	山岡亮平(著)	大修館書店	1995	2310
LBn9_00088	9 文学	ディフェンス	ウラジーミル・ナボコフ(著),若島正(訳)	河出書房新社	1999	2373

３．方法

３．１　形態素解析について

形態素解析は BCCWJ に付与されている短単位の情報をそのまま用いた。エラーの修正は行っていない。

３．２　多義語の認定について

該当の語が多義かどうかについては『三省堂国語辞典第七版』（以下、『三国』と略す）によった。この辞書は比較的多義を認定している可能性が高いことから、より多くの分析対象が抽出できることを期待して採用したものである。

３．３　語義の数え方

各サンプルに出現した各短単位の中から頻度 2 以上の名詞（普通名詞）[2]を抜き出し、個々の使用例ごとに『三国』のどの語義に相当するかを判断した。多義のレベルは①②などの丸付き数字のレベルで判断し、それ以下の(a)(b)などの区分は区別しない。また、[一][二]などで示された品詞の区分が異なる場合は、異なる語義とみなした。短単位と『三国』とで語のまとめ方が異なる場合、短単位を基準にした。例えば、短単位では動詞「のる」は『三国』の「乗る」と「載る」を合わせたものに相当する。したがって、「乗る」の 15 個の語義と「載る」の 3 個の語義を合わせたものを短単位の「のる」に対応させた。

以下のような事例については分析の対象から除外した。

(1)誤解析と認められるもの[3]。

(2)該当する見出し語が『三国』にないもの[4]。

(3)該当する語義が『三国』にないもの[5]。

対象となった語数を表 2 にまとめた。対象とした多義普通名詞（表 2 のいちばん下の 2 行）をサンプル全体に対する比率で見ると、LBd0_00003 は、それぞれ 13.9%（延べ）と 13.5%（異なり）、LBj4_00045 では、12.1%（延べ）と 11.8%（異なり）、LBn9_00088 では、延べ・異なりともに 7.6%である[6]。

---

[2] 名詞のサブカテゴリのうち、固有名詞と数字は多義性がほとんどないと考えられることから除外した。

[3] 例えば、LBd0_00003 において「縁（えん）」と解析された 2 語は「ふち」ないしは「へり」となるものであったため、分析から除外した。

[4] 例えば、LBj4_00045 において「大蟻」「キャピラリー」「クロマトグラフ」「巣内」「侍蟻」「他種」「吐き戻し」「山蟻」などが『三国』の見出しにないため、分析から除外した。実際上、これはら多義の可能性が少ないので分析に与える影響は少ないと考えられる。

[5] 例えば、LBj4_00045 において、多義語の例として「手」4 例が出てくるが、これらは、慣用句「手に入れる」「手を煩わす」の一部であったり、複合語「お手の物」の一部であったりするため、該当する語義がなく、分析から除外した。

[6] ちなみにこの比率の差を多重比較したところ、延べ語数、異なり語数ともに LBd0_00003 と

表 2　対象とした多義語

語数	LBd0_00003	LBj4_00045	LBn9_00088
延べ語数	2403	2310	2373
異なり語数	594	524	683
度数 2 以上の普通名詞（延べ）	511	564	249
度数 2 以上の普通名詞（異なり）	125	113	76
度数 2 以上の多義の普通名詞（延べ）	362	341	249
度数 2 以上の多義の普通名詞（異なり）	91	75	76
対象とした多義普通名詞（延べ）	335	279	181
対象とした多義普通名詞（異なり）	80	62	52

３．４　語義の集中度

語義の集中度は、ある語についてサンプル中に出現した語義数（異なり[7]）をその語の持つ、可能性としての語義数（『三国』の語義数）で割った値を 1 から引いた値とした。すなわち、語義を 3 つ持つ語がサンプル中で 1 つの語義でしか使われなかった場合、1−(1/3)で、0.667 となる。もし、3 つの語義が全部使われていれば、1−(3/3)となり、集中度は 0 となる。語義が 1 つしかない語については集中度を算出しない。

４．結果

４．１　語義数の分布

図 1 は各サンプルにおける普通名詞における、可能性としての語義数の分布である。前述のように語義数は『三国』の語義数に拠っている。3 つとも似たような L 字形をしているが、LBn9_00088 はややゆるやかなカーブになっているのが特徴的である。

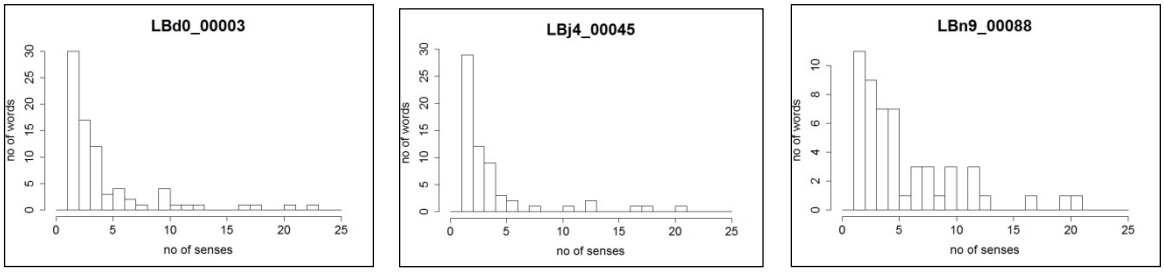


図 1　多義語の語義数の分布（『三国』の語義数の分布）

これに対して、実際に出現した語義数の分布が表 3 である。どのサンプルでも語義数 1 がいちばん多いが、LBn9_00088 は語義数 2 も割合が他より高くなっている。

表 3　出現した語義数の分布

語義数	LBd0_00003	LBj4_00045	LBn9_00088
1	69	52	38
2	10	7	12
3	1	2	-

LBj4_00045 との組み合わせは 5%水準で有意差がなかったが、そのほかの組み合わせは 5%水準で有意差が認められた。

[7] この指標は延べ語数を考慮していない。したがって、度数 10、語義数 2 の多義語があったとして、語義の分布が 9,1 の場合も 5,5 の場合は同じ値(0)になる。

4	-	1	1
5	-	-	1

４．２　集中度の分布

図 2 に各サンプルにおける多義の集中度の分布を示した。LBd0_00003 と LBj4_00045 はほぼ同じ形の分布を示しているが、LBn9_00088 は分布の形が異なっている。この差が何に由来するのかは不明であるが、LBn9_00088 が小説であることが関係している可能性を指摘しておく。

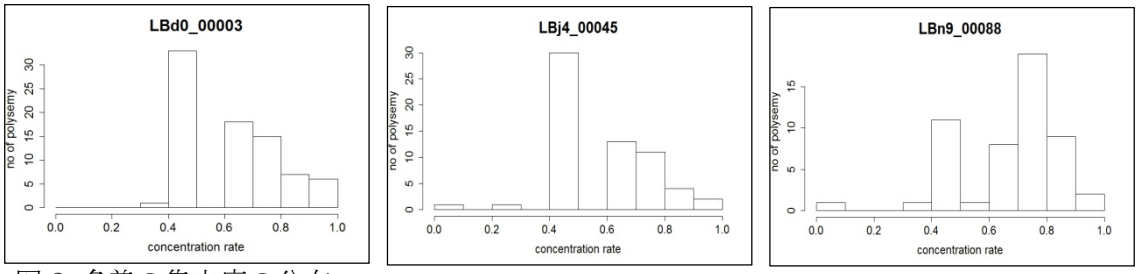


図 2 多義の集中度の分布

サンプル全体の集中度を表 4 に示した。平均値でみても、中央値でみても、LBj4_00045＜LBd0_00003＜LBn9_00088 の順になっている。事例数が少ないので憶測の域を出ないが、サンプル全体の集中度が当該テキストを特徴付ける指標になる可能性がある。

表 4　サンプル全体の集中度

代表値	LBd0_00003	LBj4_00045	LBn9_00088
平均値	0.649	0.606	0.693
中央値	0.667	0.500	0.750

表 5～7 は各サンプルにおける頻度 5 以上の語についての調査結果である。

表 5 語義の分布（LBd0_00003）頻度 5 以上

頻度	語彙素	『三国』語義数	出現語義数	内訳[8]	集中度
15	図	5	1	③	0.800
12	回	6	2	[二]①×5，[二]②×7	0.667
12	粉	2	1	①	0.500
12	方向	3	1	①	0.667
11	事	21	2	[一]①×4，[一]⑥×7	0.905
11	層	4	2	[一]①×7，二×4	0.500
9	板	5	1	①	0.800
8	ゲーム	4	1	②	0.750
8	センチ	2	1	②	0.500
8	作り	7	1	[二]①	0.857
8	度	10	2	[一]⑨×2，二×6	0.800
8	時	13	1	⑪	0.923
7	パン	2	1	①	0.500

[8] 内訳の欄の記号は『三国』の語義に対応する。×のあとの数字は出現数（token）である。ただし、出現語義数が 1 の場合、および、×1 の場合は省略した。

7	水	4	1	①	0.750
6	後	12	1	[一]②	0.917
6	固まり	2	1	②	0.500
6	時間	6	1	③	0.833
6	操作	2	1	①	0.500
6	東西	3	1	[一]①	0.667
6	中	8	2	①×3，④×3	0.750
5	上	10	2	[一]①，[一]②×4	0.800
5	三角	2	1	①	0.500
5	日	11	1	⑤	0.909
5	侭	4	1	③	0.750
5	ミリ	2	1	②	0.500

表 6 語義の分布（LBj4_00045）頻度 5 以上

頻度	語彙素	『三国』語義数	出現語義数	内訳	集中度
29	事	21	3	[一]6×24,[一]9,[一]⑪×4	0.857
17	奴隷	2	1	①	0.500
16	巣	3	1	①	0.667
13	種	3	1	①	0.667
11	炭化	2	1	①	0.500
10	相手	3	1	①	0.667
10	成分	2	1	①	0.500
10	物	17	4	[一]①×5,[一]⑥×2,[一]⑦,[一]⑬,除外 1	0.765
9	仲間	2	1	②	0.500
8	違い	3	1	一×7,除外 1	0.667
8	物質	2	1	②	0.500
8	分析	2	1	②	0.500
7	為	4	3	[一]①×5,[一]②,[二]	0.250
5	コロニー	4	1	③	0.750
5	自分	5	1	一	0.800

表 7 語義の分布（LBn9_00088）頻度 5 以上

頻度	語彙素	『三国』語義数	出現語義数	内訳	集中度
21	事	21	5	[一]①×8,[一]⑥×7,[一]⑨×4,[一]⑯,[一]⑰	0.762
19	娘	2	1	①	0.500
10	物	17	4	①×4,⑦×2,⑧×2,⑩,除外 1	0.765
8	時	13	2	⑪×6,⑬×2	0.846
6	声	5	1	①	0.800
6	二人	2	1	①	0.500
5	頭	12	2	①×2,⑤×2,除外 1	0.833
5	側	3	1	①	0.667

４．３　複数の語義で出現した語

各サンプルにおいて出現語義数が 2 以上の語は次の通りである。[] 内の数字は［可能性

としての語義数/出現語義数］である。

LBd0_00003：
回［6/2］、事（こと）［21/2］、層［4/2］、度（ど）［10/2］、中（なか）［8/2］、頃［4/2］、所（ところ）［18/2］、訳（わけ）［10/3］、最高［3/2］、地方［4/2］
LBj4_00045：
事（こと）［21/3］　物（もの）［17/4］　為（ため）［4/3］　所（ところ）［18/2］　情報［4/2］　筈（はず）［6/2］　以後［2/2］　程度［4/2］　時（とき）［13/2］　中（なか）［8/2］
LBn9_00088：
事（こと）［21/5］　物（もの）［17/4］　　時（とき）［13/2］　頭（あたま）［12/2］　後（あと）［12/2］　中（なか）［8/2］　癖（くせ）［4/2］　間（あいだ）［8/2］　顔（かお）［10/2］　最近［3/2］　姿［5/2］　全て［2/2］　展開［8/2］　訳（わけ）［10/2］

これらを見ると分かるように、複数の語義で出現した語のほとんどが抽象的な語であることが分かる。紙幅の関係で用例は省略するが、事（こと）、物（もの）、訳（わけ）、為（ため）、所（ところ）、筈（はず）など、意味が形式化して機能語に近い用法を持つ語が多い。また、「全て（名詞・副詞）」「以後（名詞、造語成分）」のように、品詞が違うことで多義として挙がっている語がある。一方、今回観察した 2400 短単位程度のテキストでは具体的な意味を持つ普通名詞で多義的に使用されている語はなかった。このことは、語彙的結束性は語彙的には強い制約としてテキストの成立条件となっているが、文法的にはその制約は弱いのではないかということが推測される。

５．まとめと今後の課題

本稿で観察したのは、普通名詞だけであったが、テキスト中では約 7 割～8 割の多義語が特定の意味でのみ使用されていることが確認された。その例外となっていたのは、ほとんどが文法的な意味での使用に関わるものであった。したがって、文法的な意味は語彙的結束性に関与する度合いが低いことを示唆した。今後はサンプル数を増やすとともに、動詞、形容詞などの他の品詞における多義の実現傾向、またテキストにおける使用頻度と出現語義数との関係も視野に入れて分析を行う予定である。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

Gale,William A. et als.(1992), “One sense per discourse”, Proceedings of the workshop on Speech and Natural Language, pp.233-237, Harriman, NY.

Krovetz, Robert.(1998), “More than One Sense per Discourse”, Proceedings of the ACL-SIGLEX Workshop (Senseval)

Halliday, M.A.K. and Hasan, R.(1976) Cohesion in English. Longman.

山崎誠(2010),テキストにおける多義語の意味実現の傾向,計量国語学会第 54 回大会予稿集,pp25-30.

拡張固有表現階層から SUMO への対応表

今田　水穂 (文部科学省初等中等教育局)

Mapping Table from ENE to SUMO

Mizuho Imada (MEXT)

要旨

複数の言語資源から取得した意味情報を上位オントロジーに統合して管理することを目的として、関根の拡張固有表現階層 7.1.0 に含まれる 243 個の固有表現名に対して、対応する SUMO クラス名を割り当てた。概要と課題について解説する。

1. はじめに

関根の拡張固有表現階層 7.1.0[1] (Sekine, Sudo, and Nobata 2002; 以下 ENE) から Suggested Upper Merged Ontology[2] (Niles and Pease 2001; 以下 SUMO) への対応表を作成した。この対応表は京都大学テキストコーパス 4.0[3] (以下京大コーパス) に含まれる名詞述語文への意味情報付与タスク [4](今田 2015) の成果物の一部である。拡張固有表現タグ付きコーパス [5]、CRL 固有表現データ [6]、日本語 WordNet[7] など複数の言語資源から得られた語義情報を SUMO に統合して管理することを目的とする。

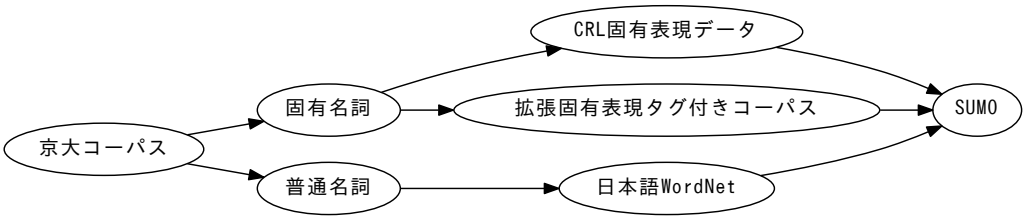


図 1　京大コーパスへの語義付与

SUMO は既存のオントロジーを統合する目的で開発された上位オントロジーである。WordNet から SUMO への変換は日本語 WordNet に収録されている既存の対応表を利用したが、拡張固有表現および CRL 固有表現から SUMO への変換は独自に対応表を作成して利用した。本発表では拡張固有表現から SUMO への対応表について説明する。

2. 方法

ENE から SUMO への対応表を構築するにあたり、次の 2 つの原則を設定した。

(1) ENE クラス名に対して割り当てる SUMO クラス名は、原則として固有表現の種類ではなく、その固有表現によって表される事物や概念のクラスとする。

(2) ENE で上位下位の関係にあるクラスは、原則として SUMO でも上位下位の関係になるように対応するクラスを設定する。

(1) は、例えば ENE には名前 (ENE) というクラスがあるが、それと対応する SUMO クラス名は Name(SUMO) ではなく Entity(SUMO) とするということである (図 2)。図中の実線は子クラス、破線は子孫クラスを表す。

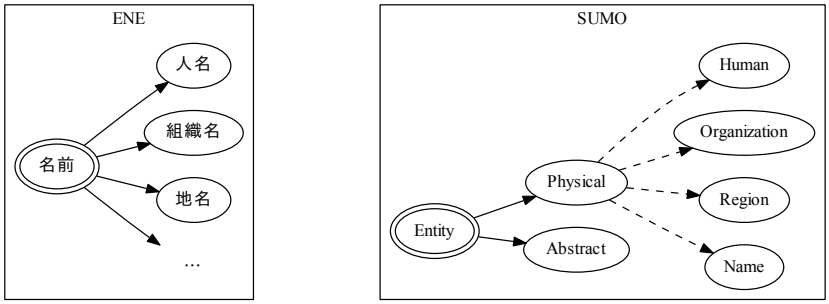


図 2　固有表現のクラスと事物や概念のクラス

(2) は、例えば施設名 (ENE) は StationaryArtifact(SUMO) と対応するので、施設名 (ENE) の下位クラスである学校名 (ENE) には StationaryArtifact(SUMO) の下位クラスである EducationalFacility(SUMO) を割り当て、Organization(SUMO) の下位クラスである School(SUMO) を割り当てないということである (図 3)。

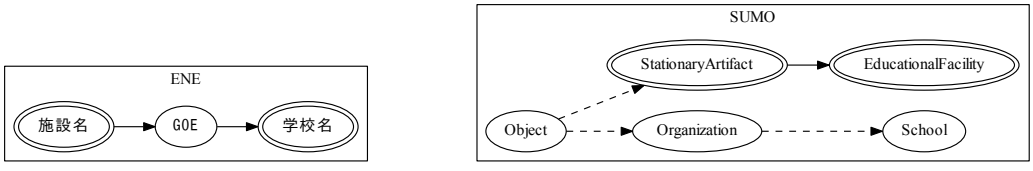


図 3　上位下位関係の保持

3. 結果

ENE クラス名 243 個に対して、対応する SUMO クラス名を割り当てた。対応表は稿末の付録に示す。また、この対応表を使用して京大コーパスに SUMO クラス名を付与するタスクを実施した。このタスクの詳細については今田 (2015) を参照されたい。

4. 考察

以下、対応表の作成および使用の過程で問題となった点について論じる。

4.1 指示対象が想定できない固有表現

(1) では ENE クラス名に対して割り当てある SUMO クラス名は固有表現の指示対象のクラスとするという原則を示したが、固有表現の中には指示対象を想定することが困難なものがある。これらについては、やむを得ず原則 (1) の例外として指示対象ではなく言語表現それ自体などに関する SUMO クラス名を割り当て、また原則 (2) のクラスの上位下位関係の保持も考慮しないものとした。具体的には以下のものである。

表 1　指示対象が想定できない固有表現

ENE	SUMO
/名前/地名/アドレス	PlaceAddress
/名前/地名/アドレス/アドレス_その他	PlaceAddress
/名前/地名/アドレス/郵便住所	PostalAddressText
/名前/地名/アドレス/電話番号	TelecomNumber
/名前/地名/アドレス/電子メイル	EmailAddress
/名前/地名/アドレス/URL	WebAddress
/名前/製品名/識別番号	Identifier
/名前/製品名/称号名/称号名_その他	Entity
/名前/製品名/単位名	UnitOfMeasure
/名前/製品名/単位名/単位名_その他	UnitOfMeasure
/名前/製品名/単位名/通貨単位名	UnitOfCurrency

アドレス (ENE) と識別番号 (ENE) は空間的実体に対して割り当てられるものであるとは限らず、具体的な指示対象を持たない場合がある。そのため、文字列それ自体を指示するものと見なして SUMO クラスを割り当てた。アドレス (ENE) の下位クラスのうち郵便番号 (ENE) は空間的実体を指示対象として想定することが可能であるが、扱いの統一のために文字列それ自体を指示するものと見なした。図 4 に関連する SUMO クラスを示す。

称号名_その他 (ENE) は主に「さま」「さん」「ちゃん」など敬称の類である。称号名_その他 (ENE)、単位名 (ENE) はそれぞれ人名や数値表現の一部を構成するものであるが、単独では指示対象を想定することができない。称号名_その他については便宜的に Entity(SUMO) を割り当てた。単位名 (ENE) については SUMO に UnitOfMeasure(SUMO) というクラスがあるた

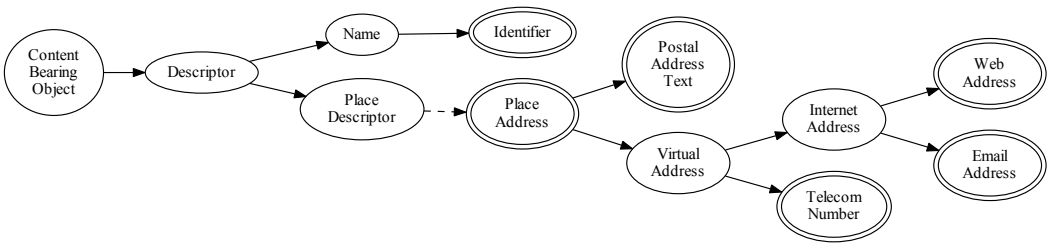


図 4　アドレス (ENE)・識別番号 (ENE) に割り当てる SUMO クラス

め、このクラスを割り当てた。人名や数値表現の一部であることから Human(SUMO) や PhysicalQuantity(SUMO) などのクラスを割り当てるか、接辞であることから ParticleWord(SUMO) を割り当てるという考え方もできるが、今後の検討課題としたい。図 5 に関連する SUMO クラスを示す。

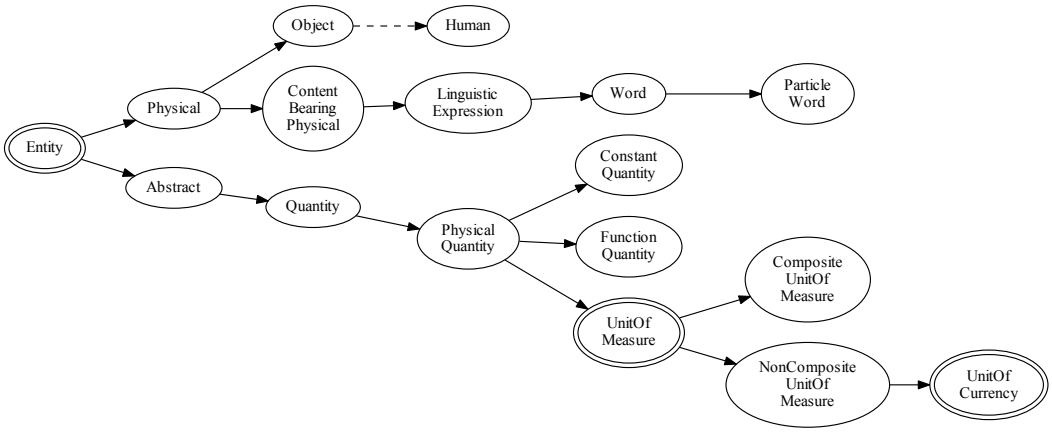


図 5　称号名 (ENE)・単位名 (ENE) に割り当てる SUMO クラス

4.2　分類粒度の不一致

ENE クラス名に対応する適当な SUMO クラス名が見つからない場合、その直近の上位クラスに相当すると考えられる SUMO クラス名を割り当てた。この結果、ENE クラス名に対して分類粒度の粗すぎる SUMO クラス名が割り当てられる場合があった。例としては、次のような事例があった。

- ENE クラス名に対応する適当な SUMO クラス名が見つからない場合
  - 食べ物名 (ENE) に対応する適当な SUMO クラス名が見つからなかったため、Object(SUMO) を割り当てた[1)]。

[1)] 食べ物名 (ENE) の子クラスである料理名 (ENE) には PreparedFood(SUMO) を割り当てた。

  - 期間 (ENE) の子クラス (時刻期間 (ENE)、日数期間 (ENE)、週数期間 (ENE) など) について、対応する適当な SUMO クラス名が見つからなかったため、全てに TimeDuration(SUMO) を割り当てた。
  - 数値表現 (ENE) の子孫クラスの多くについて、対応する適当な SUMO クラス名が見つからなかったため、RealNumber(SUMO) や PhysicalQuantity(SUMO) などを割り当てた。
- ENE クラス名に対応する適当な SUMO クラス名が見つからないが、代わりにその子クラスに相当するいくつかのクラス名が見つかった場合
  - 公演組織名 (ENE) に対応する適当な SUMO クラス名が見つからなかったが、代わりに DramaticCast(SUMO) と MusicalGroup(SUMO) が見つかったため、これら 2 つのクラス名の直近の上位クラスである GroupOfPeople(SUMO) を割り当てた。
  - 時刻表現 (ENE) に対応する適当な SUMO クラス名が見つからなかったが、代わりに Hour(SUMO)、Minute(SUMO)、Second(SUMO) などが見つかったため、これらのクラス名の直近の上位クラスである TimeInterval(SUMO) を割り当てた。日付表現 (ENE) も同様。

WordNet から SUMO への対応表には $\in$, $\subset$, $=$ などいくつか異なる種類のリンクが設定されている。ENE から SUMO への対応表にも同様の情報を付与すべきであるが、本対応表では設定していない。

4.3 意義の多面性 (facet)

ある 1 つの概念が複数の意味的側面を持つということがある。例えば「学校」は建物、場所、組織としての側面を持つ。また「本」は事物と内容の側面を持つ。こうした意義の多面性は facet (Cruse 2010) や dotted object (Pustejovsky 1998) と呼ばれる。このような事例は ENE にも SUMO にも含まれている。ENE の多面的クラスに対して SUMO にも対応する適当な多面的クラスがある場合にはそれを割り当てた。ENE の多面的クラスに対して SUMO に対応する適当な多面的クラスが無く、代わりに各側面に対応する複数のクラスがある場合には原則 (2) を考慮していずれか 1 つのクラスを割り当てた。主なものとしては以下のものがある (太字が実際に割り当てたクラス)。

表 2　意義の多面性を持つ固有表現

ENE	SUMO
/名前/地名/GPE	**GeopoliticalArea**
/名前/施設名	**StationaryArtifact**, Region
/名前/施設名/GOE	**StationaryArtifact**, Region, Organization
/名前/製品名/芸術作品名	**ContentBearingPhysical**, Proposition
/名前/製品名/出版物名	**ContentBearingObject**, Proposition

GPE(ENE) は場所と行政組織という 2 つの側面を持つ。SUMO には GeopoliticalArea(SUMO) という適当な対応クラスが存在したため、これを割り当てた。Geopo-

liticalArea(SUMO) は Region(SUMO) と Agent(SUMO) の下位クラスである (SUMO はクラスの多重継承を許す)。

施設名 (ENE) は建物と場所という 2 つの側面を持つ (ENE において「建物や場所としての属性を持つ施設の名前」と定義されている)。SUMO には StationaryArtifact(SUMO) と Region(SUMO) という 2 つのクラスがあるが、本対応表では StationaryArtifact(SUMO) を割り当てた。また、施設名 (ENE) の子クラスの GOE(ENE) は建物と場所に加えて組織としての側面を持つ (ENE において「建物や場所としての属性の他に組織名としての属性を持つ施設の名前」と定義されている)。本対応表では原則 (2) に従い、StationaryArtifact(SUMO) を割り当てた[2)]。

芸術作品名 (ENE) と出版物 (ENE) はメディア (作品、書籍など) と内容という 2 つの側面を持つ。SUMO には ContentBearingPhysical(SUMO) と Proposition(SUMO) というクラスがあるが、本対応表では ContentBearingPhysical(SUMO) かその下位クラスを割り当てた。関連する SUMO クラスを図 6 に示す。

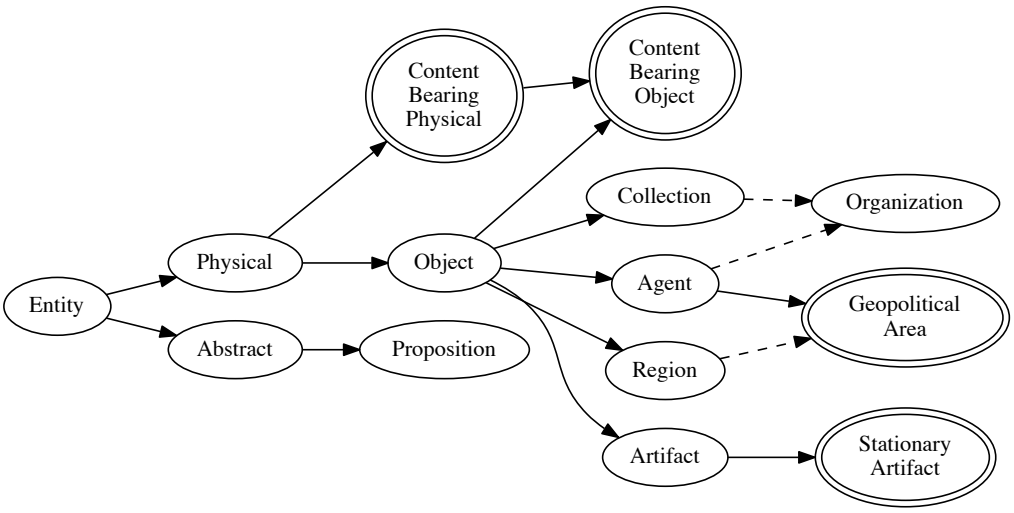


図 6　多面的概念に割り当てる SUMO クラス

以上のように、本対応表では単一の ENE クラスに対して単一の SUMO クラスのみ割り当てているため、意義の多面性に関する情報の一部が切り捨てられている場合がある。

[2)] 施設名 (ENE) の下位クラスには、対応する適当な SUMO クラスが存在するが StationaryArtifact(SUMO) ではないというものが存在する。例えば、道路名 (ENE) に対応すると考えられる Roadway(SUMO) は Region(SUMO) の下位クラスではあるが StationaryArtifact(SUMO) の下位クラスではない。この場合にも原則 (2) に従って、Roadway ではなく StationaryArtifact を割り当てた。

4.4　クラスとインスタンス

SUMO にはクラスとインスタンスの区別がある。次の図で丸囲みはクラス、四角囲みはインスタンス、矢印の実線は子クラスまたは子インスタンス関係、破線は子孫クラスまたは子孫インスタンス関係、点線はクラス-インスタンス関係を表す。

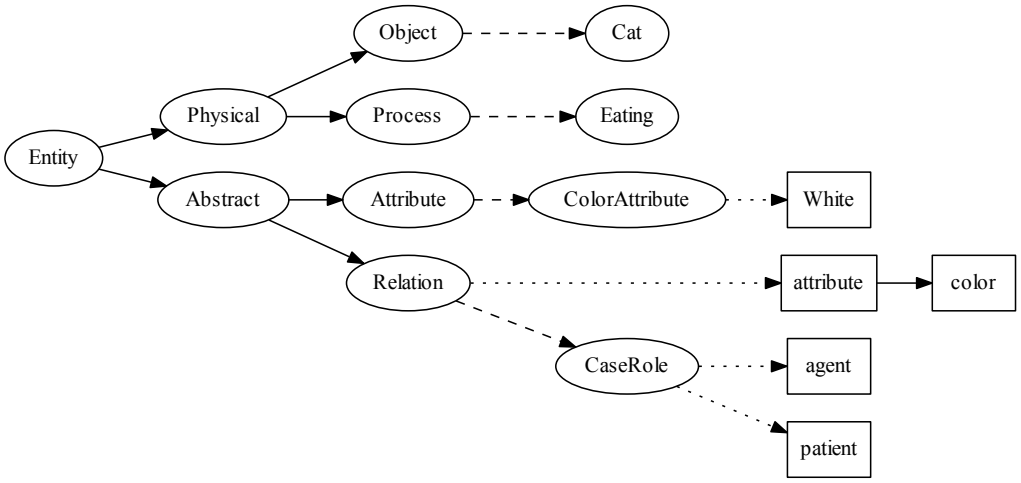


図 7　SUMO のクラスとインスタンス

標準的な述語論理の 1 項述語は SUMO では Object クラスや Process クラスなどの下位クラス (Cat, Fish, Eating など)、または Attribute クラスのインスタンス (White など) に相当し、2 項述語は Attribute クラスの下位クラス (ColorAttribute など)、または Relation クラスのインスタンス (attribute,　color など) に相当する。SUMO で述語として機能するのは Relation クラスのインスタンスのみである。標準的な述語論理の記法と SUMO の記法の対応例を以下に示す。

表 3　述語論理と SUMO 表現

標準的な述語論理	SUMO
cat(a)	instance(?A, Cat)
white(a)	color(?A, White)
∃e[eat(e) & agent(e, a) & patient(e, b)]	(exists (?E) (and (instance(?E, Eating) agent(?E, ?A) patient(?E, ?B))))

名詞述語文の意味を記述する上では、主語名詞と述語名詞がどのクラスに属するかという情報に加えて、それらがクラスかインスタンスか (あるいは述語論理における項数 (arity) がいく

つか) も重要な情報である。(3) は事物の帰属関係を表しているが、述語名詞が 1 項述語である点に特徴がある。(4) は事物の同一関係を表しているが、主語名詞と述語名詞はいずれも個体であって述語ではない。(5) は「深さ」が 2 項述語に相当し「震源」と「１０キロ」がその項に相当する。項数の情報を言語資源から取得することができれば、これらの構文の分類をある程度自動化することができる。

(3)　吾輩は猫である。

(4)　明けの明星は宵の明星だ。

(5)　震源の深さは１０キロ。

ENE はクラスとインスタンスの区別を持たない。言語学で固有名詞と呼ばれる表現の多くはインスタンスを表すが、ENE の固有表現には動物名 (ENE) などクラスを表すものも含まれる。時間表現 (ENE) や数値表現 (ENE) の下位クラスの多くはインスタンスを表すと考えられるが、「日曜日」などクラスを表すものも含まれる。また、施設部分名 (ENE) には「居間」のようなクラスを表す語も「長和殿」のようなインスタンスを表す語も含まれているなど、同一の ENE クラス内にクラスを表す語とインスタンスを表す語が混在している場合もあり、最終的には個々の語に対してクラスとインスタンスの区別を割り当てる必要がある。

本対応表では ENE クラス名に対して SUMO クラス名を割り当てることのみを行っており、その SUMO クラス自体に相当するのか、その SUMO クラスのインスタンスに相当するのかの区別までは行っていない。

5. まとめ

ENE から SUMO への対応表の概要と課題について述べた。異なるオントロジー間のリンクテーブルの構築は単に概念を 1 対 1 に対応付ければよいというものではなく、分類粒度、意義の多面性、クラスとインスタンスの区別など様々な問題について考慮する必要がある。意味論研究に利用可能な言語資源とその利用方法の開発、蓄積のために、今後さらに検討を重ねたい。

謝辞

本研究の一部は国立国語研究所共同研究プロジェクト「コーパスアノテーションの基礎研究」および国立国語研究所コーパス開発センター「超大規模コーパス構築プロジェクト」によるものである。また，本研究は JSPS 科研費 23720225「Ruby と MSXML による日本語名詞述語文の実例調査とコーパス分析ツールの構築」(研究代表者: 今田水穂) の助成を受けている。

文献

今田水穂 (2015)「日本語名詞述語文への意味情報付与」『国立国語研究所論集』8. (印刷中)

Cruse, Alan (2010) *Meaning in Language: An Introduction to Semantics and Pragmatics* . 3rd ed. OUP.

Niles, Ian, and Adam Pease (2001) “Towards a Standard Upper Ontology”. In *Proceedings of the FOIS-2001*.

Pustejovsky, James (1998) *The Generative Lexicon*. MIT Press.

Sekine, Satoshi, Kiyoshi Sudo, and Chikashi Nobata (2002) “Extended Named Entity Hierarchy” In *Proceedings of the LREC 2002*.

関連 URL

[1] 関根の拡張固有表現階層 7.1.0　https://sites.google.com/site/extendednamedentityhierarchy/

[2] Suggested Upper Merged Ontology (SUMO)　http://www.ontologyportal.org/

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

[4] Ruby と MSXML による日本語名詞述語文の実例調査とコーパス分析ツールの構築　https://sites.google.com/site/kaken23720225/

[5] 拡張固有表現タグ付きコーパス　http://www.gsk.or.jp/catalog/gsk2013-b/

[6] CRL 固有表現データ (IREX ホームページ)　http://nlp.cs.nyu.edu/irex/index-j.html

[7] 日本語 WordNet　http://nlpwww.nict.go.jp/wn-ja/

付録: 拡張固有表現階層から SUMO への対応表

テキスト

ENE	ENE 英語表記	SUMO
名前	Name	Entity
名前_その他	Name_Other	Entity
人名	Person	Human
神名	God	Deity
組織名	Organization	Group
組織名_その他	Organizaton_Other	Group
国際組織名	International_Organization	InternationalOrganization
公演組織名	Show_Organization	GroupOfPeople
家系名	Family	FamilyGroup
民族名	Ethnic_Group	EthnicGroup
民族名_その他	Ethnic_Group_Other	EthnicGroup
国籍名	Nationality	EthnicGroup
競技組織名	Sports_Organization	Group
競技組織名_その他	Sports_Organization_Other	Group
プロ競技組織名	Pro_Sports_Organization	SportsTeam
競技リーグ名	Sports_League	SportsLeague
法人名	Corporation	Organization
法人名_その他	Corporation_Other	Organization
企業名	Company	Business
企業グループ名	Company_Group	Business
政治組織名	Political_Organization	Organization
政治的組織名_その他	Political_Organization_Other	Organization
政府組織名	Government	GovernmentOrganization
政党名	Political_Party	PoliticalParty
内閣名	Cabinet	GovernmentCabinet
軍隊名	Military	MilitaryForce
地名	Location	Region
地名_その他	Location_Other	Region
温泉名	Spa	TouristSite
GPE	GPE	GeopoliticalArea
GPE_その他	GPE_Other	GeopoliticalArea
市区町村名	City	City
郡名	County	County
都道府県州名	Province	StateOrProvince
国名	Country	Nation
地域名	Region	GeographicArea
地域名_その他	Region_Other	GeographicArea
大陸地域名	Continental_Region	Continent
国内地域名	Domestic_Region	LandArea
地形名	Geological_Region	GeographicArea
地形名_その他	Geological_Region_Other	GeographicArea
山地名	Mountain	UplandArea
島名	Island	Island
河川名	River	StreamWaterArea
湖沼名	Lake	LandlockedWater
海洋名	Sea	BodyOfWater
湾名	Bay	BodyOfWater
天体名	Astral_Body	AstronomicalBody
天体名_その他	Astral_Body_Other	AstronomicalBody
恒星名	Star	Star
惑星名	Planet	Planet
星座名	Constellation	AstronomicalBody
アドレス	Address	PlaceAddress
アドレス_その他	Address_Other	PlaceAddress
郵便住所	Postal_Address	PostalAddressText
電話番号	Phone_Number	TelecomNumber
電子メイル	Email	EmailAddress
URL	URL	WebAddress
施設名	Facility	StationaryArtifact
施設名_その他	Facility_Other	StationaryArtifact
施設部分名	Facility_Part	StationaryArtifact
遺跡名	Archaeological_Place	StationaryArtifact
遺跡名_その他	Archaeological_Place_Other	StationaryArtifact
古墳名	Tumulus	StationaryArtifact
GOE	GOE	StationaryArtifact
GOE_その他	GOE_Other	StationaryArtifact
公共機関名	Public_Institution	StationaryArtifact
学校名	School	EducationalFacility
研究機関名	Research_Institute	Laboratory
取引所名	Market	PlaceOfCommerce
公園名	Park	StationaryArtifact
競技施設名	Sports_Facility	SportsFacility
美術博物館名	Museum	EducationalFacility
動植物園名	Zoo	StationaryArtifact
遊園施設名	Amusement_Park	StationaryArtifact
劇場名	Theater	Auditorium
神社寺名	Worship_Place	PlaceOfWorship
停車場名	Car_Stop	StationaryArtifact
電車駅名	Station	TrainStation
空港名	Airport	Airport
港名	Port	StationaryArtifact
路線名	Line	StationaryArtifact
路線名_その他	Line_Other	StationaryArtifact
電車路線名	Railroad	Railway
道路名	Road	StationaryArtifact
運河名	Canal	Canal
航路名	Water_Route	StationaryArtifact
トンネル名	Tunnel	Tunnel
橋名	Bridge	Bridge
製品名	Product	Entity
製品名_その他	Product_Other	Entity
材料名	Material	Substance
衣類名	Clothing	WearableItem
貨幣名	Money_Form	Currency
医薬品名	Drug	Medicine
武器名	Weapon	Weapon

ENE	ENE 英語表記	SUMO
株名	Stock	Stock
賞名	Award	UnilateralGiving
勲章名	Decoration	UnilateralGiving
罪名	Offense	CriminalAction
便名	Service	ServiceProcess
等級名	Class	Collection
キャラクター名	Character	CognitiveAgent
識別番号	ID_Number	Identifier
乗り物名	Vehicle	Vehicle
乗り物名_その他	Vehicle_Other	Vehicle
車名	Car	RoadVehicle
列車名	Train	RailVehicle
飛行機名	Aircraft	Aircraft
宇宙船名	Spaceship	Spacecraft
船名	Ship	WaterVehicle
食べ物名	Food	Object
食べ物名_その他	Food_Other	Object
料理名	Dish	PreparedFood
芸術作品名	Art	ContentBearingPhysical
芸術作品名_その他	Art_Other	ContentBearingPhysical
絵画名	Picture	PaintedPicture
番組名	Broadcast_Program	BroadCasting
映画名	Movie	MotionPicture
公演名	Show	DramaticPerformance
音楽名	Music	MusicalPerformance
文学名	Book	Text
出版物名	Printing	ContentBearingObject
出版物名_その他	Printing_Other	ContentBearingObject
新聞名	Newspaper	Newspaper
雑誌名	Magazine	Magazine
主義方式名	Doctrine_Method	Entity
主義方式名_その他	Doctrine_Method_Other	Entity
文化名	Culture	Proposition
宗教名	Religion	Proposition
学問名	Academic	FieldOfStudy
競技名	Sport	Sport
流派名	Style	Proposition
運動名	Movement	PoliticalProcess
理論名	Theory	Explanation
政策計画名	Plan	Plan
規則名	Rule	Proposition
規則名_その他	Rule_Other	Proposition
条約名	Treaty	Proposition
法令名	Law	Proposition
称号名	Title	Entity
称号名_その他	Title_Other	Entity
地位職業名	Position_Vocation	SocialRole
言語名	Language	Language
言語名_その他	Language_Other	Language
国語名	National_Language	HumanLanguage
単位名	Unit	UnitOfMeasure
単位名_その他	Unit_Other	UnitOfMeasure
通貨単位名	Currency	UnitOfCurrency
イベント名	Event	Process
イベント名_その他	Event_Other	Process
催し物名	Occasion	SocialInteraction
催し物名_その他	Occasion_Other	SocialInteraction
例祭名	Religious_Festival	SocialParty
競技会名	Game	Game
会議名	Conference	FormalMeeting
事故事件名	Incident	Process
事故事件名_その他	Incident_Other	Process
戦争名	War	War
自然現象名	Natural_Phenomenon	Motion
自然現象名_その他	Natural_Phenomenon_Other	Motion
自然災害名	Natural_Disaster	Motion
地震名	Earthquake	Earthquake
自然物名	Natural_Object	Object
自然物名_その他	Natural_Object_Other	Object
元素名	Element	ElementalSubstance
化合物名	Compound	CompoundSubstance
鉱物名	Mineral	Mineral
生物名	Living_Thing	Organism
生物名_その他	Living_Thing_Other	Organism
真菌類名	Fungus	Fungus
軟体動物_節足動物名	Mollusc_Arthropod	Invertebrate
昆虫類名	Insect	Insect
魚類名	Fish	Fish
両生類名	Amphibia	Amphibian
爬虫類名	Reptile	Reptile
鳥類名	Bird	Bird
哺乳類名	Mammal	Mammal
植物名	Flora	Plant
生物部位名	Living_Thing_Part	AnatomicalStructure
生物部位名_その他	Living_Thing_Part	AnatomicalStructure
動物部位名	Animal_Part	AnimalAnatomicalStructure
植物部位名	Flora_Part	PlantAnatomicalStructure
病気名	Disease	DiseaseOrSyndrome
病気名_その他	Disease_Other	DiseaseOrSyndrome
動物病気名	Animal_Disease	DiseaseOrSyndrome
色名	Color	ColorAttribute
色名_その他	Color_Other	ColorAttribute
自然色名	Nature_Color	PrimaryColor
時間表現	Time_Top	TimeMeasure
時間表現_その他	Time_Top_Other	TimeMeasure
時間	Timex	TimePosition
時間_その他	Timex_Other	TimePosition
時刻表現	Time	TimeInterval
日付表現	Date	TimeInterval
曜日表現	Day_Of_Week	TimeInterval
時代表現	Era	TimeInterval
期間	Periodx	TimeDuration
期間_その他	Periodx_Other	TimeDuration
時刻期間	Period_Time	TimeDuration
日数期間	Period_Day	TimeDuration
週数期間	Period_Week	TimeDuration
月数期間	Period_Month	TimeDuration
年数期間	Period_Year	TimeDuration
数値表現	Numex	Quantity
数値表現_その他	Numex_Other	Quantity
金額表現	Money	CurrencyMeasure
株指標	Stock_Index	StockIndex
ポイント	Point	PhysicalQuantity
割合表現	Percent	RealNumber
倍数表現	Multiplication	RealNumber
頻度表現	Frequency	PhysicalQuantity
年齢	Age	TimeMeasure
学齢	School_Age	TimeMeasure
序数	Ordinal_Number	RealNumber
順位表現	Rank	RealNumber
緯度経度	Latitude_Longtitude	PlaneAngleMeasure
寸法表現	Measurement	PhysicalQuantity
寸法表現_その他	Measurement_Other	PhysicalQuantity
長さ	Physical_Extent	LengthMeasure
面積	Space	AreaMeasure
体積	Volume	VolumeMeasure
重量	Weight	MassMeasure
速度	Speed	FunctionQuantity
密度	Intensity	FunctionQuantity
温度	Temperature	TemperatureMeasure
カロリー	Calorie	FunctionQuantity
震度	Seismic_Intensity	FunctionQuantity
マグニチュード	Seismic_Magnitude	FunctionQuantity
個数	Countx	PhysicalQuantity
個数_その他	Contx_Other	PhysicalQuantity
人数	N_Person	PhysicalQuantity
組織数	N_Organization	PhysicalQuantity
場所数	N_Location	PhysicalQuantity
場所数_その他	N_Location_Other	PhysicalQuantity
国数	N_Country	PhysicalQuantity
施設数	N_Facility	PhysicalQuantity
製品数	N_Product	PhysicalQuantity
イベント数	N_Event	PhysicalQuantity
自然物数	N_Natural_Object	PhysicalQuantity
自然物数_その他	N_Natural_Object_Other	PhysicalQuantity
動物数	N_Animal	PhysicalQuantity
植物数	N_Flora	PhysicalQuantity

明治後期における漢語の基本語化

田中 牧郎（明治大学・国立国語研究所）†

Inclusion of Sino-Japanese Words into Basic Vocabulary in the Late Meiji Japanese

TANAKA Makiro (Meiji University・National Institute for Japanese Language and Linguistics)

要旨

明治時代の語彙は、漢語に大きな変化がある。近代語の雑誌コーパスの年次別の語彙頻度を指標に、語彙をレベル分けして、年次によるその移行を見ることで、明治後期に基本語化した漢語を抽出することができる。その基本語化した漢語は、科学技術や社会制度を表す語、及び、抽象概念を表す語に多いが、後者の場合は、既存の和語や漢語との間で使い分けられることで、新しい語彙体系を構築していく役割がある。

１．はじめに

明治時代は、初期に非常に多かった漢語が次第に淘汰されていくが（池上 1984、今野 2014 など）、この時代に漢語がどれだけあり、どのような語が淘汰され、どのような語が定着したかということになると、資料が多種多様であることもあって、従来の研究では明らかにされてこなかった。しかし、近年、国立国語研究所で整備が進められてきた近代の総合雑誌のコーパスによって、その具体的な状況を知ることができるようになってきた。雑誌、とりわけ総合雑誌は、多彩なジャンル、幅広い執筆者、厚い読者層などの点で、当時の書き言葉をかなりの程度代表できるものと見ることができ、書籍や新聞など他の媒体に比べて、コーパスが備えるべき代表性を、一種の資料でありながら備えているからである。国立国語研究所では、総合雑誌『太陽』（1895（明治 28）-1928（昭和 3））、学術総合雑誌『明六雑誌』（1874（明治 7）-1875（明治 8））などのコーパスを開発してきた（国立国語研究所 2005、2012）。そして、まもなく、総合雑誌『国民之友』（1887（明治 20）-1898（明治 31））の、創刊時の 2 年分を対象とした『国民之友コーパス』を、公開予定である（国立国語研

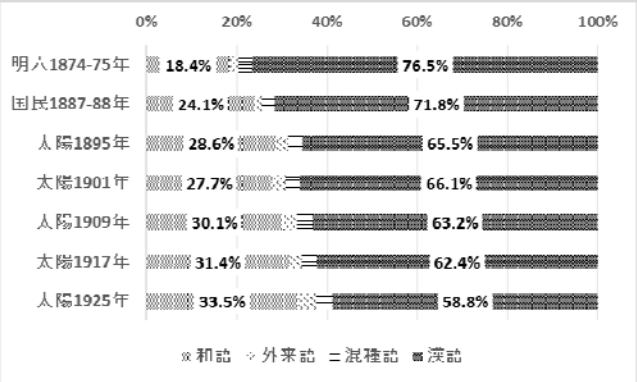


図１　語種構成比率の推移（異なり語数）

† makiro@meiji.ac.jp

究所 2014）[1]。上記の 3 つのコーパスを用いて、年次別の語種構成比率を異なり語数で算出すると、図 1 のようになる。明治初期には漢語の比率が非常に高いが、それは次第に低下していき、代わりに和語が増加していく傾向が明確に見てとれる。漢語が、明治初期に氾濫し、その後大正末期までに次第に淘汰されていく量的な様子を、具体的に知ることができる。

しかし、このような語種構成比率の調査だけでは、どのような漢語が淘汰され、どのような漢語が残存したかや、なぜこのような変化が起こったかについては、不明なままである。図 1 に見られるような大きな流れの背後にあった具体的な言語変化を研究していくには、語彙の内部に分け入って考察していく必要がある。語彙の内部に分け入る方法として、本稿では、コーパスにおける頻度情報を指標とした、基本語・周辺語の軸での語彙変化の見方を採用し、とりわけ周辺語から基本語へと変化していく「基本語化」する語彙を抽出し、それがどのようなものであるかについて考えたい。

２．基本語化のとらえ方

2.1 語彙のレベル分け

語彙を星雲に喩える樺島（1981）などの考えをもとに、田中（2012b・2013）では、語彙の中心に、安定してよく使われる「基本語」があり、その周辺に、不安定であまり使われない「周辺語」があると見立て、この基本語と周辺語の軸で近代の語彙を研究する視点を提示した。代表性を持つコーパスの場合、高頻度の語が基本語に相当し、低頻度の語が周辺語に相当すると見なすことが可能であり、『太陽コーパス』の各年次の語彙を、頻度によってレベル分けし、そのレベルを「基本レベル」「中間レベル」「周辺レベル」に分類し、語彙の変化・不変化の様子を概観した。

表１　語彙のレベル分け（『国民之友コーパス』『太陽コーパス』）

レベル	使用率の累積	国民 1887-88	太陽 1895	太陽 1901	太陽 1909	太陽 1917	太陽 1925
a	-78%	-35	-46	-51	-54	-53	-40
b	78-88%	34-12	45-17	50-19	53-19	52-19	14-39
c	88-94%	11-5	16-7	18-8	18-8	7-18	13-6
d	94-97%	4-2	6-4	7-4	7-4	6-4	5-3
e	97-100%	1	3-1	3-1	3-1	3-1	2-1
異なり語数		31, 928	49, 773	43, 049	38, 383	36, 387	38, 221

表 1 は、田中（2012b）で行った『太陽コーパス』5 年次分のレベル分けと同じ基準で、『国民之友コーパス』1887-88 年分の語彙にもレベル分けを施し、それら全体の基準を示したものである（数字は頻度）。各年次の語彙を頻度（使用率）順に配列し、使用率の累積が

[1] それぞれのコーパスの概要や、それらが持つ代表性については、田中（2005）、近藤・田中 2012）、近藤ほか（2014 予定）などを参照してほしい。

78%に達するところまでをレベルa、その後88%までをレベルbというように、a〜eの5段階に区画する。例えば、『国民之友』1987-88年では、頻度35以上の語がレベルaに属し、頻度34以下12以上の語がレベルbに属すことになる。なお、『明六雑誌コーパス』は、『国民之友コーパス』と十数年の開きがあり、他の年次が6〜8年刻みであるのに比べて離れすぎているため、頻度推移を考察するには不適切だと判断して、今回の調査資料からは外した[2]。

2.2 基本語化した語彙の抽出

個々の語において5段階に分けたレベルが、年次によってどのように変化する（変化しない）かを見ることで、語彙の中での各語の位置の変化・不変化を見ることができる。表２は、50音順の語彙表のうち、マ行の冒頭7語について、年次別のレベル情報を一覧にしたものである。例えば「間（ま)」は、すべての年次でレベルaであり、この時代基本語であり続けた語である。一方、「間合い」「マーカンタイル」は、どの年次でもレベルeまたは「-」[3]であり、ずっと周辺語だった語である。そして、すべての年次がレベルｂ・ｃ・ｄのいずれかに入っている「真（ま)」は、中間語で不変だったと言える。これらが語彙の中での位置を変えない語であるのに対して、「まあ（副詞)」は、当初レベルeだったものが、レベルｂそしてレベルaへと一定方向に変化するもので、変化した後はレベルaのままで基本語として安定する。なお、「魔」「まあ（感動詞)」は、いずれかのレベルで安定したり、一定方向に変化したりすることはなく、特徴のはっきりしない語である。

表２　年次別のレベル情報（50音順語彙表マ行冒頭7語）

語彙素読み	語彙素	品詞	語種	1887-88	1895	1901	1909	1917	1925
マ	魔	名詞	漢	−	b	d	c	b	b
マ	間	名詞	和	a	a	a	a	a	a
マ	真	接頭辞	和	d	b	c	b	c	b
マア	まあ	副詞	和	e	b	a	a	a	a
マア	まあ	感動詞	和	c	e	b	b	b	b
マアイ	間合い	名詞	和	−	−	−	−	−	e
マーカンタイル	マーカンタイル	名詞	外	−	e	−	e	−	−

このようにして作成した、レベル情報を一覧にした全語彙のリストをもとに、ここでは全年次を通じて、基本語化の方向が明確に見てとれる語を抽出することにした。具体的には、図２の●を付した範囲に収まる語を抽出し、これを基本語化した語と扱うことにした。

[2] 『明六雑誌コーパス』と『国民之友コーパス』の間の位置にある、1881年（明治14年）ごろの雑誌などのコーパス化が求められる。

[3] 「-」は、当該年次には使用例がない（頻度0）ことを示す。

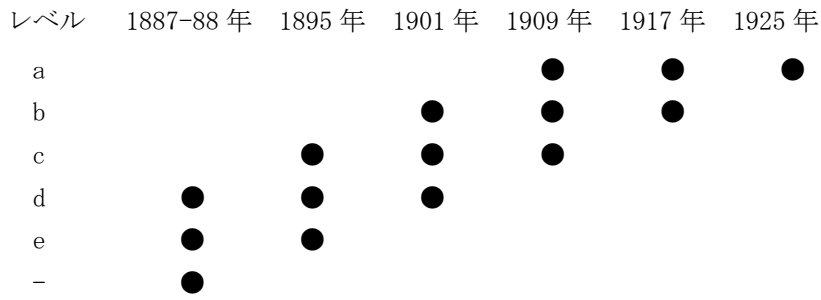


図２　基本語化の抽出基準

図２の基準で抽出した基本語化した語彙は、固有名詞を除くと、81 語となった。それを、語種別に五十音順で掲げると次の通りである。

●和語（28 語）

明るい、いらっしゃる、要る、偉い、型、決まる、心持ち、強い、直ぐ、すっかり、ずっと、立場、経つ、立つ、たらしい、小さな、ちゃんと、蔦、冷たい、何の、捕らわれる、亡くなる、何しろ、名前、太る、丸で、もっと、漸と

●漢語（50 語）

圧迫、栄養、援助、解決、階段、閣議、拡大、可能、期待、気分、共通、協定、具体、見地、講演、貢献、向上、興奮、合理、考慮、国有、色彩、支持、実現、手術、信念、推薦、生殖、節約、対抗、大変、妥協、徹底、到頭、特徴、都市、努力、内容、肉体、発展、飛行、皮肉、不安、復興、奮闘、本能、有権、有利、率、廊下

●混種語（3 語）

かなり、大した、駄目

３．基本語化した語彙の分類

3.1 口語的な語

前節で抽出した基本語化した語彙 81 語は、どのような性質を持っているだろうか。まず、第一に、口語的な語の一群がある。

例えば、和語の４番目にある「偉い」は、『国民之友コーパス』には１例のみあり、次のように、小説の会話文で使われている。

「ホ、それはおえらいな！」（『国民之友』1887 年 27 号、二葉亭四迷「あひびき」）

この語は『太陽コーパス』の初年次の 1895 年には十数例見られるが、そのほとんどは、下の例のように小説等の会話の引用文か、口語体で書かれた記事に見られ、文語部分にはほとんど使われていない。口語部分に使用範囲が限定されることから、この語は口語的な性格が強かったと考えられる。

最う奥村様はえらいお方でございますよ。（『太陽』1895 年 2 号、川上眉山「書記官」）
其えらい所は我々の師匠とするに足りるから（『太陽』1895 年 11 号、田口卯吉「歴史は科学に非ず」）

同じように、コーパスでの使用箇所を見ていくと、初期の年次では会話の引用や、口語体の記事に使用範囲が限られる傾向にある語として、次などが指摘できる。

●和語（24 語）
いらっしゃる、要る、偉い、決まる、心持ち、強い、直ぐ、すっかり、ずっと、経つ、立つ、たらしい、小さな、ちゃんと、冷たい、何の、捕らわれる、亡くなる、何しろ、名前、太る、丸で、もっと、漸と
●漢語（3 語）
気分、大変、到頭
●混種語（3 語）
かなり、大した、駄目

和語が多いが、漢語、混種語のなかにも、口語的な語はある。明治後期は、書き言葉が文語体から口語体に移行する時期であり（山本 1965、森岡 1991、野村 2013、田中 2013 など)、その変化に伴って、文章の中に口語的な語が使われることが増え、それらが書き言葉における基本語に加わっていくのだと考えられる。

3.2 科学技術や社会制度を表す語

基本語化した語彙には、第二に、科学技術や社会制度を指す一群の語彙がある。

例えば、漢語の 2 番目にある「栄養」という語は、『国民之友コーパス』には全く使われておらず、『太陽コーパス』の最初の年次には、次の 2 例だけが見られる。

身躰の榮養に充分意を注いで、少許の時間で多量の安眠を遂げる工夫を回らして、(『太陽』1895 年 12 号、石橋思案「睡眠の節減」)
吾人にありて榮養液は血液と混入し心臓の作動に因りて躰内を循環すると同じく蝶にありても亦心臓と稱するポンプ様のものありて其伸縮に依りて此營養液を全躰内に循環せしむるものなり（『太陽』1895 年 9 号、石川千代松「蝶の話」）

人体についての科学的な説明の中に用いられている。石川千代松「蝶の話」の例には、すぐ後に「営養」も見られ、この表記は、この記事に他に 3 例、『太陽』1895 年の別の記事に 1 例見られ、「栄養」と同じ意味を表している[4]。また、次のように「衛養」という表記も 1 例あり、これも「栄養」「営養」と同じ意味である。

身躰の衛養に毫も思ひ至ることなく（『太陽』1895 年 7 号、前田正名（談）「肉食の必要」）

これらの例のうち、はじめに掲げた、石橋思案「睡眠の節減」が「雑録」の欄に収められているほかは、「科学」または「農業」の欄の記事で、専門的な科学ジャンルの文章であ

[4] 『日本国語大辞典第２版』の「語誌」には、「蘭学の訳書にも「営養」「栄養」が併用された」「明治から大正にかけては「営養」の方が優勢であった」と記されている。

る。「栄養」（他の表記も含む）は、科学用語として使われ始めた語だと見られる。それが次第に一般に普及していくともに、基本語化していくと考えられる。

社会制度を指す一群の語についても同様に見ることができる。例えば「国有」は、『国民之友コーパス』には皆無で、『太陽コーパス』の初年次には、次の 1 例だけが見られる。

ペクトウヰチは前國有鐵道長官たり（『太陽』1895 年 10 号、無署名「海外彙報」）

この「国有鉄道」は、セルビアの時事を報告した記事に使われているもので、この時点では海外の制度を指す語である。ところが、次の年次の 1901 年には日本の鉄道の国有について論じる部分で多く使われるようになり、鉄道以外にも使われている。

鐵道國有の建議案が、不成立となりしは、最も滑稽なりき。（『太陽』1901 年 1 号、国府犀東「政治時評」）
我國の森林に於て國有林は一千萬町歩、御用林三百萬町歩、民有林七百三十萬町歩にして（『太陽』1901 年 5 号、高橋琢也（談）「林政論」）

以上のような、基本語化した、科学技術や社会制度を指す語はすべて漢語であり、次がそれに該当しよう。

●漢語（12 語）
栄養、閣議、協定、国有、手術、生殖、都市、肉体、飛行、本能、有権、率

3.3 抽象概念を表す語

以上の二つの理由では説明できない語が、たくさん残る。次のようなもので、いずれも抽象概念を表す語で、大部分を漢語が占める。

●和語（3 語）
明るい、型、立場
●漢語（35 語）
圧迫、援助、解決、階段、拡大、可能、期待、共通、具体、見地、講演、貢献、向上、興奮、合理、考慮、色彩、支持、実現、信念、推薦、節約、対抗、妥協、徹底、特徴、努力、内容、発展、皮肉、不安、復興、奮闘、本能、有利

これらのうち、下線を引いた語については、本稿とは別の指標を用いて、『太陽コーパス』から定着する漢語サ変動詞を抽出し[5]、その背景や事情を考察したことがある（田中 2011）。

それによれば、「向上」は、当初、文学や宗教の分野で、心を上向かせる意味を表していたが、1909 年からは、他のジャンルでも用いられるようになり、心以外の事物が上向くこ

[5] その抽出においては語の「定着」という観点を取った。そこでの「定着」は、頻度の増加傾向や新しさの度合いなどによって抽出したもので、本稿の「基本語化」の観点が語彙のレベルの変化によって抽出しているのとは、異なっている。しかし、結果的に抽出された語は、重なっているものも多い。

とも表し、同時に動詞用法を持つようにもなる。「興奮」は、当初、外的要因が精神などを高ぶらせる他動詞用法が主であったが、1909 年から自動詞用法が主になり、高ぶる要因は明示されなくなり、その頃から頻度が増加する。「考慮」は、当初、行政の分野で、自らをすり減らして苦心するような意味を持っていたが、頻度の増加と使用される分野の広がりが目立つ 1917 年ごろからは、自ら前向きに工夫するような意味へと変わっていく。さらに、「実現」は、当初は、文学のジャンルで、現実化が容易でない貴い事柄を現実のものとする意味であったが、頻度が増加しジャンルも広がる 1909 年からは、さして困難を伴わないあたり前のことを現実化する意味で使われるようになる[6]。そして、「努力」は、当初は、相当に力を込めて奮闘して頑張ることを意味したが、頻度が急増する 1909 年からは、特別な力の込め方はなくなっていく[7]。以上の基本語化する語の中には、「考慮」に対する「考える」「慮る」「顧慮」、「実現」に対する「あらわす」「あらわれる」「表現」「出現」、「努力」に対する「つとめる」のように、基本語化にあたって、類義の和語や漢語との間に使い分けの関係を形成し、新たな語彙体系を構築していくものも目立っていた。

上記のリストで下線が引かれていない語についても、基本語に際して意味変化があったのではないかと予想される。また、その意味変化の過程で類義語との使い分け関係をつくり、新しい語彙体系を構築していく動きがあったのではないかとも予想される。そのことを実証するような、個々の語の用例を分析する研究が期待される。

４．おわりに

以上本稿では、語彙頻度に基づくレベルを指標として、基本語化が明瞭にとらえられる 81 語を抽出し、それらがどのような語であるかを考察した。その結果、（１）口語的な語、（２）科学技術や社会制度を表す語、（３）抽象概念を表す語、の三つのタイプに分かれることが分かった[8]。基本語化の基準として今回採用したものは、かなり厳格なものであり、ここで抽出されたものは基本語化の傾向が極めて著しいものに限られ、そのような語に指摘できた上記三つのタイプは、この時代に基本語化する語彙の性質として重要なものと見てよいだろう。基本語化の傾向にある語は、基準を緩めることで、もっと多く抽出していくことができるだろう。

三つのタイプのうち、（１）は、言文一致運動の直後の時代である明治後期という時代に特徴的なものだと考えられる。（２）も、科学技術が著しく発展し、社会制度が大きく変革された近代化の時代だからこそ目立つものと言えるが、語が指示する対象物の発展や変革によって語が基本語化するという構図は、他の時代の語彙の変化にも適用できるものだとも考えられ、他の時代に基本語化する語と比較することが必要だろう。

（３）は、（１）（２）と違って、明治後期という時代の性質から即座にこれを説明することは難しい。しかし、語数は（１）と並んで多く、この時代の語彙変化で最も目立つ漢語に特に多いことから、語彙変化の本質に関わる現象ではないかと予想される。昭和戦後期から平成期にかけての新聞における外来語の基本語化を研究した金（2011）は、その大半を抽象概念を表す外来語が占めていることを明らかにし、「トラブル」「ケース」などを例

---

[6] 「実現」については、田中（2012a・2013）において、より詳しく考察した。

[7] 「努力」については、田中（2006）において、より詳しく考察した。

[8] このことの見通しは、田中（2013）でも述べたが、本稿は、それを実証するものという位置付けになる。

に、基本語化の背景には、既存の和語や漢語の類義語との意味的な関係があることを示している。これは、本稿で見た、明治後期に基本語化する漢語の性質や、その背景にある類義語との意味関係と、類似性が高い。明治後期の漢語の基本語化と、昭和戦後期から平成期の外来語の基本語化とを、対比的に研究し、基本語化現象を日本語史の中に位置付けていく研究が求められよう。

付記

本研究は、NINJAL 共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：田中牧郎）、及び、JSPS 科学研究費基盤研究（C）「近現代日本語彙における「基本語化」現象の記述と類型化」（26370529、研究代表者：金愛蘭）による成果の一部です。

文献

池上禎造（1984）『漢語研究の構想』（岩波書店）
樺島忠夫（1981）『日本語はどう変わるか―語彙と文字』（岩波新書）
金愛蘭（2011）「20 世紀後半の新聞語彙における外来語の基本語化」（『阪大日本語研究 別冊 3』）
国立国語研究所（2005）『太陽コーパス―雑誌『太陽』日本語データベース―』（博文館新社、CD-ROM）
国立国語研究所（2012）『明六雑誌コーパス』（国立国語研究所コーパス開発センターWeb サイト）
国立国語研究所（2014 予定）『国民之友コーパス』（国立国語研究所コーパス開発センターWeb サイト）
近藤明日子・小木曽智信・高田智和・田中牧郎（2014 予定）「『国民之友コーパス』の開発」（『日本語学会 2014 年度秋季大会予稿集』）
近藤明日子・田中牧郎（2012）「『明六雑誌コーパス』の仕様」（田中牧郎ほか（2012）『近代語コーパス設計のための文献言語研究 成果報告書』（国立国語研究所共同研究報告 12-03 所収）
今野真二（2014）『日本語の近代―はずされた漢語』（ちくま新書）
田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」（国立国語研究所『雑誌「太陽」による確立期現代語の研究』博文館新社）
田中牧郎（2006）「「努力する」の定着と「つとめる」の意味変化―『太陽コーパス』を用いて―」（倉島節尚『日本語辞書学の構築』おうふう）
田中牧郎（2011）「近代漢語の定着―『太陽コーパス』に見る―」（『文学』12-3、岩波書店）
田中牧郎（2012a）「新漢語定着の語彙的基盤―『太陽コーパス』の「実現」「表現」「出現」と「あらわす」「あらわれる」など―」（『日本語の学習と研究』160 ）
田中牧郎（2012b）「明治後期から大正期の語彙のレベルと語種」（田中牧郎ほか（2012）『近代語コーパス設計のための文献言語研究 成果報告書』（国立国語研究所共同研究報告 12-03 所収）
田中牧郎（2013）『近代書き言葉はこうしてできた そうだったんだ!日本語』（岩波書店）
野村剛史（2013）『日本語スタンダードの歴史』（岩波書店）
森岡健二（1991）『近代語の成立 文体編』（明治書院）
山本正秀（1965）『近代文体発生の史的研究』（岩波書店）

「勉強する」と「rian」の対象語の分析
－BCCWJ と TNC（Thai National Corpus）を用いて－

木田 真理*（国際交流基金日本語国際センター）
PRAWANG, Khommapat（政策研究大学院大学　日本語教育指導者養成プログラム（修士））
生田 守（国際交流基金日本語国際センター）

A Comparative Study on Object Words of 'BENKYO-suru' and 'RIAN' Using the BCCWJ and the TNC(Thai National Corpus)

Mari Kida （The Japan Foundation Japanese-Language Institute, Urawa）
PRAWANG, Khommapat（Graduate Program in Japanese Language and Culture, National Graduate Institute for Policy Studies）
Mamoru Ikuta（The Japan Foundation Japanese-Language Institute, Urawa）

要旨

タイ語の教科書や辞書などでは、日本語の動詞「勉強する」の対訳として「rian」が示されているが、タイ語母語話者の内省や学習者の誤用などから、両語は、意味や用法において異なった領域を有していると考えられる。辞書の記述からだけでは明示することができない「勉強する」と「rian」の違いを、『現代日本語書き言葉均衡コーパス』（BCCWJ）と、タイ語のコーパス『Thai National Corpus』（TNC）を用いて比較分析を試みた。本発表では、両語の対象語（「__を勉強する」、「rian__」の下線部分）に注目し、両コーパスの共通のジャンルから用例を抽出し、傾向や相違点を探っていく。

１．はじめに

第二筆者は、自らの日本語学習経験、及びタイの大学で日本語授業を行った際の学習者の誤用から、日本語の「勉強する」とタイ語の対訳の「rian」は使い方が異なっているのではないかという問題意識を持つようになった。よく見られる誤用例として、「明日テストがあるのに、本を勉強しない」、「今日寝る前に N3 の本いっしょうけんめい勉強した」「日本語学科を勉強している」（波線は誤用）などがあるが、第二筆者の日本語学習経験からも、辞書や日本語学習用教科書に記載されている情報だけでは、「勉強する」と「rian」の違いがはっきりとはわからない。学習者の誤用について、大曽・滝沢（2003）は、「誤用を訂正するには、訂正者が自らの直観に照らして訂正する方法が考えられるが、もう一つの(より妥当な)方策は、母語話者による大規模な日本語コーパスに依拠する方法である」と指摘し、「学習者の母語話者コーパスが整備されている場合には、それをもとにして、母語との比較を行うことも学習者の誤用指導の上で有益である」と述べている。

本研究では、辞書の記述や母語話者の内省だけでは明確に示すことができない「勉強する」とタイ語「rian」の違いを、日本語とタイ語のコーパスを用いて明らかにすることを目的とする。

２．研究対象と使用コーパス

辞書に記載されている「勉強する」と「rian」の語義は複数あるが、研究対象とするのは、次の表の語義とする。

* Mari_Kida@jpf.go.jp

べん・きょう【勉強】 学問や技術を学ぶこと。様々な体験を積んで学ぶこと。 『広辞苑』第 6 版	rian(動) [1] 教える人から知識を得る。理解や知識を得るためや、習熟するために、訓練を受ける。 『タイ国学士院編纂の国語辞典』

日本語のコーパスは、『現代日本語書き言葉均衡コーパス』 (以下、BCCWJ)を、タイ語のコーパスは、『Thai National Corpus』（以下、TNC）を用いる。TNC は、タイのチュラロンコーン大学により、British National Corpus の構成を基に開発されたもので、2009 年からインターネット上で一般に公開されている。TNC のホームページに記載されている情報によると、その総語数は、32,010,270 語である[2]。

分析の観点としては、「勉強する」と「rian」の対象語（「__を勉強する」、「rian__」の下線部分）に注目する。「勉強する」と「rian」の対象語を BCCWJ と TNC でそれぞれ検索・抽出し、用例中の使われ方を確認した上で、両者を比較検討し、対象語の傾向及び相違点を探る。

３．調査方法と調査データ

３．１　BCCWJ の検索方法とその結果

BCCWJ の検索には、「中納言（1.1.0)」を利用した。２つのコーパスを比較対照するためには、検索対象のメディア／ジャンルを合わせる必要がある。TNC は BCCWJ よりもメディア／ジャンルが少ないため、TNC に合わせるために、BCCWJ の検索対象を「出版・新聞」「出版・雑誌」「出版・書籍」「図書館・書籍」「特定目的・ベストセラー」と設定した。BCCWJ の「短単位検索」を用いた検索手順は、検索条件のキーを未指定とした上で、後方共起条件 1〜3 を、「を」、「勉強」、「為る（する）」と指定した。

検索の結果、「勉強する」からは 421 件の対象語が抽出された。抽出された「勉強する」の例のうち、次の(1)、(2)ような使役文の用例は、ヲ格をとっているが、「勉強する」の対象語ではないため排除した。

（1）そこから上の教養を身につけさせるには、塾に行かせるなり、私立の中学に入れるなりして子どもを勉強させるしかない。（PB43_00552）[3]

（2）「就職して十年かそこらは、会社はあんた、あんたを勉強させながらただ給料払ってるようなものなんだよ、…後略…」（OB5X_00232）

その結果、「勉強する」の対象語の用例は 419 件となった。

３．２　TNC の検索方法とその結果

TNC に収録されている用例は、66.3％が書籍、19.63％が定期刊行物からのものである。TNC の検索画面にはジャンルごとに検索結果が提示されており、今回はその中から「FICTION」「NEWSPAPER」「NON-ACADEMIC」「ACADEMIC」のジャンルで提示されている検索結果の用例を抽出した。検索方法は以下の通りである。

1）検索のキーワードを「เรียน（rian)」と入力する。

2）Collocation の検索で「*」（未指定）と入力し、「R1」と設定する。このようにすることにより、「rian」の直後に来る語を調べることができる。

3）全ての検索結果をダウンロードすることができないため、何回かに分けて検索し、全体の検索結果を全て抽出する。このように抽出した全用例を一つのエクセルファイル

[1] rian の日本語訳は第二筆者が翻訳したものである。

[2]＜http://ling.arts.chula.ac.th/TNC/contents/File/tncstat.txt＞
Counted on 8-1-2013　　2014 年 5 月 16 日参照

[3] 用例について、BCCWJ はサンプル番号を、TNC は ID 番号を付記した。

に格納した。

検索の結果、「rian」からは 20,389 件の対象語が抽出された。この 20,389 件には、本研究の対象とする「rian」の対象語以外の用例が多く含まれていた。TNC の検索機能は、文字列のみであり、検索条件に品詞を特定することができず、「rian」の後方に名詞句だけではなく形容詞なども抽出された。また「rian」の別の語義としての用例（「拝啓」：別の動詞の前に置き、より丁寧にする機能を持つ）も、「rian」が含まれる単語[4]も大量に検出された。さらに、検索機能の大きな問題点として、用例が全く同一のものが別のものだと認識されカウントされたものもあった。そこで、このような本研究の対象ではない「rian」の用例や重複している用例を、目視によって確認し排除する作業を行った。その結果、「rian」の対象語の用例は 1,370 件となった。

４．「勉強する」と「rian」の対象語

４．１　対象語の検索結果比較

BCCWJ 及び TNC から、「勉強する」と「rian」の対象語を抽出し、両語の各頻度が 1%以上のものをリストアップした結果が表 1 である。それぞれ抽出件数の多い順に、対象語とその件数、全件数（対象語の用例の件数）に占めるその対象語の割合を、対照させる形でまとめた。「勉強する」の対象語のうち頻度が 1%以上のものは、全体 419 件のうち 253 件であり、60.4%にあたる。一方「rian」の頻度が 1%以上のものは、全体 1,370 件のうち 1,238 件であり、90.4%を占めることがわかった。

表 1　「勉強する」と「rian」、頻度の割合が 1%以上の対象語、および件数と割合

「勉強する」の対象語			順位		「rian」の対象語			
割合	件数	対象語			対象語		件数	割合
					タイ語	日本語訳		
20.8%	87	～語	1	1	naŋsɯɯ	本	334	24.4%
13.6%	57	～学	2	2	phaasaa_	～語	201	14.7%
9.1%	38	～こと*1	3	3	wichaa_	～科目	138	10.1%
2.4%	10	～法	4	4	_saat	～学	101	7.4%
2.4%	10	～（し）方	4	5	ʔarai	何	77	5.6%
2.4%	10	法律		6	phɛɛt	医師	59	4.3%
2.1%	9	何	7	7	thaŋ_ **1	～方面	43	3.1%
1.4%	6	科目*2	8	8	kaan_ **2	～すること	37	2.7%
1.2%	5	理論	9	9	mɔɔ	医者	35	2.6%
1.2%	5	疑問詞+か		10	piano	ピアノ	33	2.4%
1.0%	4	～史	11	11	daan_ **3	～面（分野）	32	2.3%
1.0%	4	歴史		12	sinlapaʔ	美術	31	2.3%
1.0%	4	絵		13	ruaŋ_ **4	～こと	30	2.2%
1.0%	4	会話		14	saai_ **5	～系	26	1.9%
				15	tham_ **6	～作り	24	1.8%
				16	khanaʔ_	～学部	20	1.5%
				17	dontrii	音楽	17	1.2%
60.4%	253						1,238	90.4%
100.0%	419						1,370	100.0%

[4] rian が含まれる名詞や動詞の例は以下である。「rong rian 学校」「nak rian 学生」「puu rian 学習者」「beab rian テキスト」「tamra rian 教科書」「nangsuu rian 教科書」「rian chob 卒業する」「rian tor 進学する」「rian pised 特別な勉強：塾や家庭教師」

*1 糖尿病のこと、音楽のこと、化粧品のこと、色いろなこと、この世界のこと *2 一般教養科目、どんな科目、ニガテな科目、教養科目など、その科目、さまざまな学科目	**1 thaŋ kodmaai(法律の方面)、thaŋ daanthurakit(ビジネスの分野方面)等 **2 kaanʔaan（読むこと)、khaansɔɔn（教えること）等 **3 daandontrii（音楽の分野)、daansaŋkom(社会の分野)等 **4 ruaŋ rookpuinuaŋ（皮膚病のこと)、ruaŋ maarayaa（行儀のこと）等 **5 高校の教育での文系・理科系のこと。saai wit（理科系)、saai sin（文系）等 **6 tham khanom（お菓子作り)、thamʔaahaan（料理）等

４．２　対象語の検索結果のカテゴリー化

両語の相違を明らかにするために、表 1 に示した両語の対象語を、抽出用例の前後 200 語の文脈の中での使われ方を確認した上で、意味的に類似したものをまとめてカテゴリー化した。その結果を表 2 に示す。両対象語に共通するカテゴリーとして、「学問・知識」「方法・技術」「疑問詞」を設置した。タイ語だけにあるカテゴリーとしては、「目標である職業」「学校教育のプログラム」「趣味的な習い事」「本」等があった。「勉強する」だけにあるカテゴリーはなかった。

カテゴリー化の際、同じ対象語であっても、文脈上、異なる使われ方をしている場合は、異なるカテゴリーに分類した。「～こと」「絵」「美術」「ピアノ」「音楽」など、表 2 の（　）内に用例中の割合の数値が入っているものである。％数値表示のないものは、１つの対象語が一つのカテゴリーにしか分類されていないもので、表 1 のみに割合を示した。

表 2　「勉強する」と「rian」の対象語のカテゴリー分類、及び用例の割合

「勉強する」の対象語	カテゴリー	「rian」の対象語
～語、～学、～こと（4.1%)、法律、科目、理論、～史、歴史、絵（0.7%） 合計 46.0%	学問・知識	～語、～科目、～学、～方面、～面（分野)、美術（1.0%)、ピアノ（0.3%)、～こと、音楽（0.3%） 合計 41.4%
～（し）方、～法、会話、絵（0.3%） 合計 6.1%	方法・技術	～すること、～作り 合計 4.5%
何、疑問詞+か 合計 3.3%	疑問詞	何 合計 5.6%
	目標である職業	医師、医者 合計 6.9%
	趣味的な習い事	美術（1.3%)、ピアノ（2.1%)、音楽（0.9%）　合計 4.3%
	学校教育のプログラム	～系、～学部 合計 3.4%
	「本」	本 合計 24.4%

４．２．１　両語に共通するカテゴリー

A　学問・知識

分類した結果、対象語のうち、最も多いカテゴリーは、両語とも「学問・知識」であった。このカテゴリーには、次に該当するものを分類した。

「勉強する」の対象語：「～語」、「～学」、「法律」、「科目」、「理論」、「～史」、「歴史」

「rian」の対象語：「～語」、「～科目」、「～学」、「～方面」、「～面(分野)」

具体的な用例は次のような例である。

BCCWJ における「勉強する」の用例（下線は対象語を示す。以下同様）

（用例-1）　サーハンは大学で政治学を勉強しており、外交官になって国務省に勤め、いずれ大使になるつもりだったと話した。（OB4X_00197）

（用例-2）　ひと通りプログラミングができるようになってから，これら仕組み上のことや理論を勉強する方がよいでしょう．（PB20_00112）

（用例-3）　「たいへん具体的な話ですけれども、コロンビア大学でグリーリさんご自身が見ておられる範囲で学生さんなどが日本を勉強しようとする傾向はどんなふうに変わっていますか。」（LBo8_00003）

TNC における「rian」の用例（訳）[5]

（用例-4）　親たちが自分の子供たちにタイ語を勉強させたくない。タイ語を知る必要を感じないのだから。（ACSS011）

（用例-5）　今までしていた活動を、まだ続けている。バンコク大学でマスコミ学を勉強している今でも、その活動をするのは好きだ。（NWCOL04）

（用例-6）　「彼を応援に行かないの。今日の（授業）2 時限は、数学科目を勉強するでしょう。君は（数学が）できるからいいじゃない。」（PRNV072）

さらに、用例の文脈上の意味から考えると、学校教育の中で学問として勉強する場合をさす次のような例も「学問・知識」に分類した。

BCCWJ における「勉強する」の用例：「絵」

（用例-7）　私はそのとき、二十二歳で、絵を勉強していて、ニューヨークに憧れる、よくいるタイプの学生でした。いざ就職する段階になっても、やっぱり想いは断ち切れず、ひょいっと、ノースウエスト十八便に乗って、太平洋を越えてしまったのです。（LBj3_00109）

TNC における「rian」の用例：「美術」「ピアノ」「音楽」

（用例-8）「絵を描くのが好きだから、美術を勉強しようと思う」彼はそう言って、シンラパーコーン大学に入りたいと言った。（PRNV158）

（用例-9）　国立の大学に入学し、専攻としてピアノを勉強することを選んだ。（PRNV033）

（用例-10）このテキストで述べた分析方法は、高等教育で専攻として音楽を勉強する人のために、ある程度詳しく分析することになっています。（ACHM070）

また、両語が共通している対象語のうち「～こと」が表 1 に示したように 9.1%あったが、「～こと」の前にある名詞の意味により分類した。その結果、知識を指しているものが、4.1%にあたる 17 件あり、それらを「学問・知識」のカテゴリーに分類した。

「勉強する」の対象語：糖尿病のこと、日本のこと等。

（用例-11）　「スチュアートさんのウェールズ人のメイドも、彼は修理が苦手だといっていましたよ。そのことも思い出すべきでした。車が動かなくなったように見せかけたんですね。車が動かなくなったように見せかけたんですね。糖尿病のことを勉強するのが面倒になったのかもしれないな」（PB59_00240）

---

[5] 「rian」の対象語の用例は、全て第二筆者が翻訳した。

「rian」の対象語：皮膚病のこと、商売のこと等「rian rɯang ~」
（用例-12） その時は、皮膚病のことを勉強するなんて考えていなかった。はじめは、単にニキビのことについて考えていた。（NACMD005）
次に「学問・知識」以外のカテゴリーに分類したものの説明を、その用例とともに記す。

B 方法・技術

学問とは異なった技術的なものを、「方法・技術」というカテゴリーに分類した。
「勉強する」の対象語：「～（し）方」、「～法」、「～会話」、「絵」
（用例-13） それまで油絵をキャンバスに描いていたので、初めて紙と絵の具のつきあい方や筆の使い方を勉強することができました。（PM41_00715）
（用例-14） ロベルト沖中さんは、お父さまが七十一年前、京都で墨絵を勉強した方で、ご本人はブラジルで生まれ育ちました。（PB57_00001）
「rian」の対象語：「～すること」（「rian kaan~」）、「～作り」（「rian tham~」）
（用例-15）当時の文学部の 1 年生は、皆、文書を書くこと（作文）を勉強するために、グループ分けをした。（PRSH007）
（用例-16）妹は、今回演じること(演技)を勉強したい、歌を勉強したいと言い出した。僕は、父に妹の好きにさせてと言った。（PRNV016）

C 疑問詞

両語が共通している対象語「何」を「疑問詞」というカテゴリーにし、さらに、「勉強する」の対象語として検索された「疑問詞+か」をまとめた。
「勉強する」の対象語：「何」、「疑問詞+か」
（用例-17） したがって行政法とは一般の私人間の法律関係とどう異なるのかを勉強することと言うこともできるのです。（PB43_00696）
「rian」の対象語：「何」「rian ʔarai」
（用例-18） あの子、自分のことをあまり話さなかった。何を勉強しているか。学校の休みはどのぐらいなのか…（略）聞いても答えてくれない。（PRNV081）

４．２．２ 「rian」特有の対象語のカテゴリー

次に、「rian」特有の対象語の用例をカテゴリー別に示す。これらは、「勉強する」の対象語には抽出されなかったものである。TNC の用例の翻訳は全て第二筆者が行った。その際、「rian」直後の名詞は、「名詞＋を」で訳した。

A 目標である職業

「phɛɛt（医師）」、「mɔɔ（医者）」は、「rian」の対象語として 6.9%の用例が抽出された。タイ語の「rian mɔɔ（医者）」は、日本語の「医者になるための勉強をしている」、「医学部で勉強している」という表現に近い。ただし、「医者」の他に、頻度が 1 %未満だったため表 1 には示されていないが、「教師」、「看護士」等のような職業の用例が抽出された。
（用例-19） 医者になるための勉強をしている僕でも、注射は苦手です。（NACHM078）
（用例-20） ヴィアンさんのおうち、子供がみんな教師になるための勉強をしている。教師になったら、どこに行っても「先生、先生」と呼ばれている。（PRNV）

B 学校教育のプログラム

「～系」、「～学部」を「学校教育のプログラム」というカテゴリーにまとめた。これらは日本語では「～学部で勉強している」という表現になるが、この用例では、前置詞句（thii NP）ではなく、動詞の直後に対象語が置かれ、日本語に直訳すると、「～学部を勉強する」のようになってしまう。
（用例-21） ネースは、ファッションデザイン学部を勉強してきたから、できあがった

服にはかなり細かいところまで表現できている。（NWCOL152）

（用例-22） 伯母が工学を勉強してほしいから、理科系を勉強しなければならない。文科系に入るなんて、考えられない。（PRNV016）

C　趣味的な習い事

「rian」の対象語として抽出された「美術」、「ピアノ」、「音楽」は、文脈上、趣味としてそれらを習う、または、子供の習い事のような意味を指している。従ってこれらを「趣味的な習い事」というカテゴリーに分類した。

（用例-23） 母親は子供にいろいろなアクティビティーをさせている。例えば、サッカーを練習させたり、美術を勉強させたりする。（PRSH056）

（用例-24）…（親たちは大忙し）、朝はピアノを勉強しに子供を送り届けて、その後、子供をまた英語の勉強に連れて行かなければならない。（NWCOL065）

（用例-25）彼女は、昔、音楽を勉強したことがあるが、まだピアノを曲に弾けないうちにやめてしまった。（PRNV016）

D　「本」

「rian」の対象語として、「nangsuu（本）」は、件数 334 件抽出され、全件数に占める割合が 24.4%と最も多く見られた。しかし、「勉強する」には「本」が対象語になっている用例は抽出されなかった(表 1 タイ語の順位 1 を参照)。

「rian」の対象語として、「nangsuu（本）」は、「本」そのものを勉強するのではなく、以下の用例のように「rian nangsuu」というコロケーションで、「学校で勉強する」という意味で使われる。

（用例-26） 彼女は豊富な家庭からきて、いい教育を受けてきた。いい学校で（rian nangsuu）勉強し、麻薬とは関わっていなかった。（PRNV016）

（用例-27） 働いている人と、まだ（rian nangsuu）勉強している人とは、好きになる人のイメージが違う。（POET023）

「rian」と「nangsuu」は結びつきが強く、「nangsuu」は、「rian」の目的語というよりは補語的な役割をしていると考えられる。それゆえ、この「rian nangsuu」というタイ語の影響から、タイ語母語話者の学習者が、「*本を勉強する」という誤用文を書くことが多いと推測される。これは筆者の教授経験とも一致する。

５．「勉強する」と「rian」の対象語の類似点・相違点

「勉強する」と「rian」の対象語を分類した結果、両語には、対象語に重なりがある部分と、異なる部分があることが分かった。「rian」の対象語で、「勉強する」の対象とならないものは、「本」、「目標である職業」、「学校教育のプログラム」「趣味的な習い事」である。「本」を対象語にとる場合がタイ語ではきわめて多いが、日本語では全く見られなかった。また、「～学部」のような「学校教育のプログラム」に関しては、日本語では対象語ではなく、場所、つまり、対格ではなく場所格で表すという扱いで、「～学部を勉強する」ではなく「～学部で勉強する」と言い表す。さらに、「医者になるための勉強をしている」という場合、タイ語では「rian phɛɛt (医師)」で表すことができるが、日本語ではこのような文は考えられない。また、「趣味的な習い事」でまとめている「習い事」に対しては、日本語では「勉強する」ではなく「習う」が使用される。

６．まとめと今後の課題

これまで見てきたような「勉強する」と「rian」の対象語の違いは、タイ語母語話者の日本語学習者が書いた誤用文にもあらわれる場合がある。日本語の授業の課題として出された日記に、習い事に対して、「水泳を勉強する」、「柔道を勉強する」というような文が見られた。また、「rian」の対象となる「学校教育のプログラム」の誤用例として、「日本語学科を勉強している」という文が見られた。これまで第二筆者は、このような誤用を、単に助

詞の使い方の間違いだと理解し、助詞を訂正していたが、もしかしたら、学習者は、「日本語学科」を「場所」ではなく「対象語」と思って「を」を使用したのかもしれない。

このように、辞書には詳しい意味記述がなく、「勉強する」と「rian」の異なる部分に焦点をあてた記述まではされていないことが、学習者の誤用につながる可能性がある。そこで、タイ語母語話者の日本語学習者には、「rian」を用いて言い表すことができても、全てが「勉強する」に置き換えることができないこと、両語の対象となるものの違いなどを、何らかの方法で示す必要があると考える。この調査結果を、日本語教育の現場へ具体的に還元させるための方法、すなわち、教科書や参考書などでの説明方法や用例の提案、授業の際の用例提示案などについて考えていきたい。

謝　辞

本研究は、国立国語研究所言語資源研究系丸山岳彦准教授にアドバイスをいただきました。深く感謝致します。また、本研究は、日本語教育指導者養成プログラム（修士）（政策研究大学院大学、国際交流基金日本語国際センターの連携大学院）の「特定課題研究」の一部です。

文　献

大曽美恵子・滝沢直宏（2003）「コーパスによる日本語教育の研究ーコロケーション及びその誤用を中心にー」『日本語学』4 月臨時増刊号 22 巻 5 号、234-244.　明治書院

後藤斉（2003）「言語理論と言語資料ーコーパスとコーパス以外のデーターー」『日本語学』4 月臨時増刊号 22 巻 5 号、6-15.　明治書院

砂川有里子（2009）「コーパスを活用した日本語教育研究」『人工知能学会誌』24 巻 5 号、656-664. 社団法人人工知能学会

丸山岳彦（2011）「コーパス日本語学」『はじめて学ぶ日本語学』185-202 ミネルヴァ書房

Wirote, Aroonmanakun (2007) Creating the Thai National Corpus In *MANUSYA:Journal of Humanities*. Special issue No.13, 4-17. Chulalongkorn University.

関連 URL

BCCWJ（中納言）　https://chunagon.ninjal.ac.jp/search

TNC（TNC Ⅱ）　http://www.arts.chula.ac.th/~ling/TNCII/

「バイリンガルコーパス・ナビゲーター」オンライン日伊並列コンコーダンサの構築と活用

Zotti Patrizia（奈良先端科学技術大学院大学、国立国語研究所）
Apolloni Riccardo，松本裕治（奈良先端科学技術大学院大学）

Compilation and Use of the Bilingual Corpus Navigator (BCN): A Japanese-Italian Online Concordancer

Patrizia Zotti (NAIST, NINJAL)
Riccardo Apolloni, Yuji Matsumoto (Nara Institute of Science and Technology)

要旨

日本語を学ぶイタリア学生の数は２０年の間にほとんど４倍に増加した（１９９３年の１９７８人から、２０１２年の７４２０人)。２０１２年には、日本語が全国２１の大学で教えられていた。それにも関わらず、オンラインリソースはほとんど存在しない。
近年では、言語処理で基本的なリソースと見なされてきた並列コーパスは、言語教育、辞書編集、翻訳の研究など多様な分野で重要であることが認められ始めている。しかし、より広く、より効果的な利用を確保するためには、簡単で直感的な方法で並列コーパスに含まれるすべての情報へのアクセスを可能にするプラットフォームを開発する必要がある。JAICOという１０，０００ペアー日伊並列コーパスと「バイリンガル・コーパス・ナビゲーター」という日伊並列オンライン・コンコーダンサを開発した。コーパスデータの検索結果をKWIC形式で、対応する対訳文と共に表示するツールである。

１．はじめに

近年では、言語処理で基本的なリソースと見なされてきた並列コーパスは、言語教育、辞書編集、翻訳の研究など多様な分野で重要であることが認められ始めている。データ駆動型学習（Data Driven Learning – DDL）では、外国語教育への実践利用には，コーパスの言語分析結果を教材やシラバスに応用する間接的利用と，コーパス検索から得られた用例を見て学習者自身が言語の規則性を発見する直接的利用が考えられている。しかし、利用の容易さを確保するためには、簡単で直感的な方法で並列コーパスを利用することができる検索ツールが必要である。

本稿では, JAICOという１０，０００ペアー日伊並列コーパスと「バイリンガル・コーパス・ナビゲーター」という日伊並列オンライン・コンコーダンサを紹介する。２節では、関連研究を紹介し、３節ではBCN (BilingualCorpusNavigator) 「バイリンガル・コーパス・ナビゲーター」の機能と使い方を紹介し、４節では，JAICO日伊コーパスを紹介し、５節でまとめを行う。

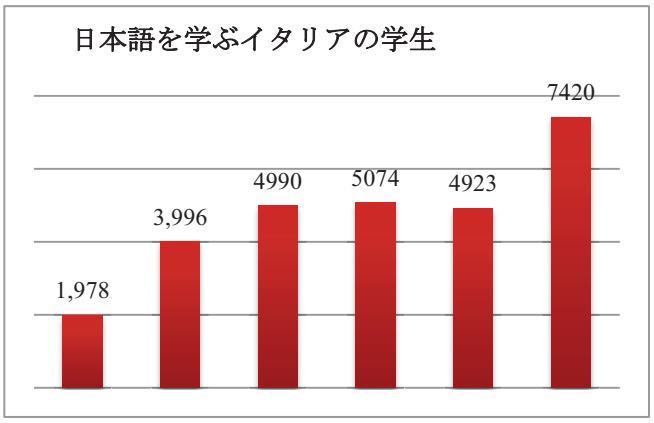


図１－日本語を学ぶイタリアの学生

出典: Japan Foundation “Survey on Japanese-Language Education Abroad” (1993-2012).

２．関連研究

伊日オンラインリソースを調べてみると、次の３つのリソースしか見つけることができない。それぞれを以下に簡単に説明する。

１）ITADICT（http://virgo.unive.it/itadict/）は、オンライン日伊辞書の計画である。現在は４万の見出し語を含んでいる (Mariotti, Mantelli 2012)。

２）OPUS Corpus (http://opus.lingfil.uu.se/)は、ウェブから取られた翻訳した文書のオープンソースデータベースである。多言語の中で、日伊パラレルデータにアクセスできるが、ジャンルは字幕だけであること、ソフトウェアのマニュアルが不十分であること、また、出力結果の日本語文が同じ語の繰り返しが含まれるなど完全な文として出力されないなどの問題がある。

３）Lang-8 （http://lang-8.com/）は言語交換ソーシャルネットワーキングサイトで、外国語で書いた文章をネイティブの話者が添消をしてくれる語学学習プラットフォームである。

３．ＢＣＮ「バイリンガル・コーパス・ナビゲーター」とは

BCN「バイリンガル・コーパス・ナビゲーター」は、イタリア語と日本語の二言語ＪＡＩＣＯコーパスの一部を搭載している（４節参照)。

BCNはhttp://cl.naist.jp/~zottip/tools.htmlにアクセスするか, あるいは「BilingualCorpusNavigator」のインターネット検索で最初に得られる検索結果をダブルクリックすると，図１のBCN の初期画面が現れ，検索作業が可能となる。

イタリア語、または日本語の検索したい語句を,「string」という文字列のボックスに入力し，ジャンルを「genre」のボックスに５つから（すべて、ニュース、小説、議事録、白書、雑録）選択し、検索方法を選択して、最後に「submit」ボタンを押すと，図２に示したようなKWIC（Key Words in Context) 検索結果の画面が得られる。

「Finding」と「Matching」の２つの検索方法がある。「Matching」の場合には、単一のトークンか、トークンの組み合わせ、または完全なテキストのセグメントを入力して、結果は優先度順に並べられる。図３示すように、すべての検索したトークンを含む文が最初に表

示され、その後検索したそれぞれトークンの１つを含む文が表示される。
図３は,「vita libertà sicurezza」の３つの伊トークンの「Matching」検索結果である。「vita libertà sicurezza」を含む伊文の検索結果が、画面の右側に１文表示され、「vita sicurezza」を含む伊文の検索結果が１文表示され、「vita libertà」を含む伊文の検索結果が１文表示され、「vita」、「sicurezza」、「libertà」のうち１つのトークンを含む伊文の検索結果が表示されている。または、それらの伊文に対応する日文が、画面の左に表示される。
「Finding」の検索方法の場合には、一つのトークンか、日本語の場合には、スペースがない文字列の検索しかできない。
ＢＣＮでは，図３と図４に示すように，検索語を含む日本語文とそれに対応するイタリア語文が一画面に表示されるので，日本語とイタリア語の文例を対照させながら学習することができる。
研究の現段階では、二つのインターフェスがある。一番目では、文のレベルで対応付けた７０００文の検索が可能で、イタリア語の検索の場合には「Matching」と「Finding」の検索が行われる。日本語の場合には「Finding」検索しか行われない。二番目では、単語レベルアラインされた白書の１００文の検索ができ、両言語でもMatchingとfindingの検索が行われる。すべてのデータに対して、単語をもちいた検索を実現することを目標にしているが、単語のアラインメントを全データに対して行うにはまだ時間を要する。現在の開発の実験段階では、データに関する検索は両言語とも文字レベルが対象であり、見出し語や品詞を用いた検索は、まだ行うことができない。

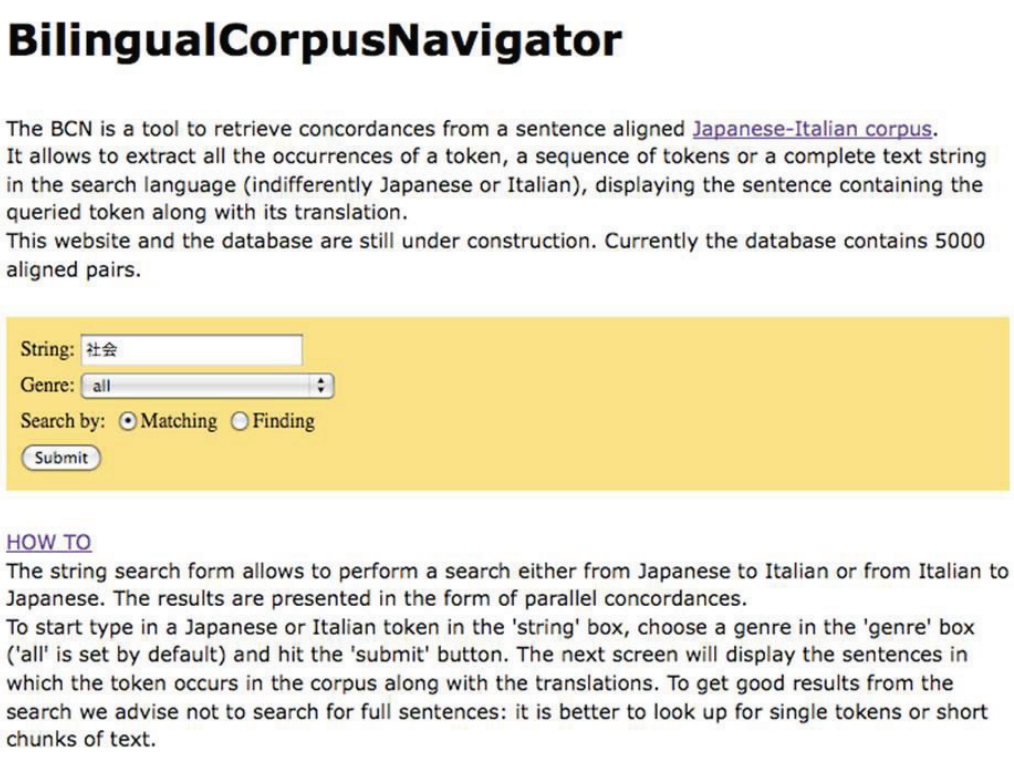


図　２　初期画面

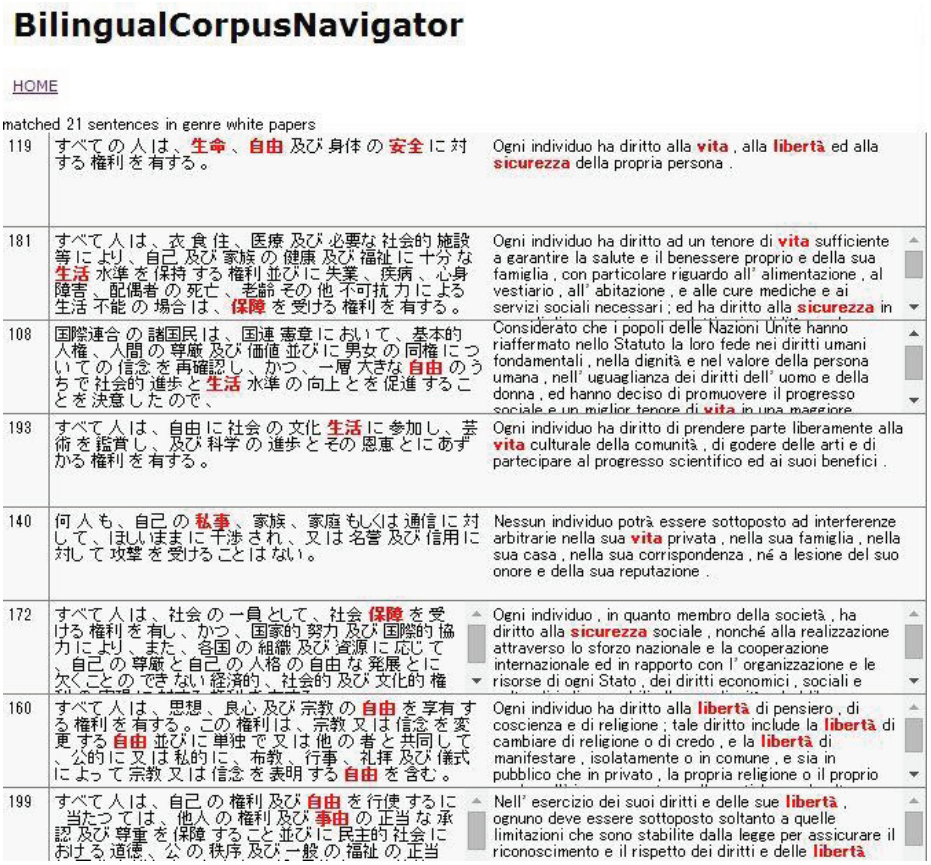


図　3　「vita libertà sicurezza」の「Matching」検索を示すス画面

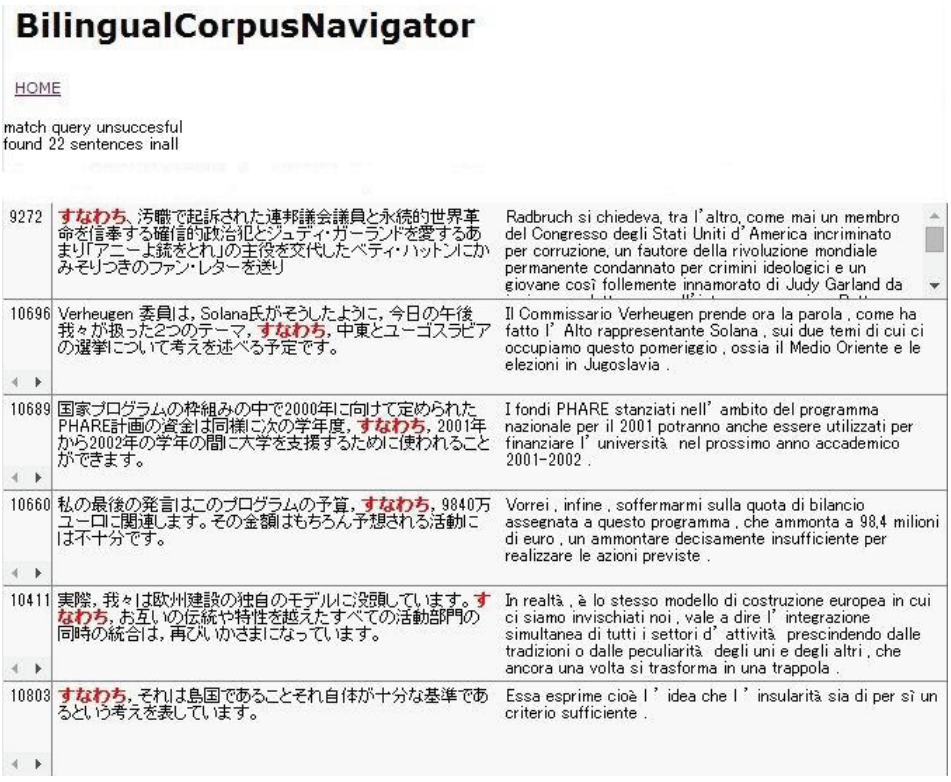


図　4　「すなわち」の「Finding」検索を示すス画面

４．JAICO 日伊パラレルコーパスとは

JAICO 「JApaneseItalianCorpus」という日伊並列コーパスのデータは，イタリア語・日本語各１万文の文対応付けコーパスである（Zotti 2013）。最初の６０００文は半自動的に対応付けされ、次の４０００文はChurch and Galeアルゴリズム（Gale, Church 1993）に基づく実装で自動的に対応付けされた（Zotti, Apolloni, Matsumoto 2010；JISA – JapaneseItalianSentence Aligner, http://cl.naist.jp/~zottip/tools.html）。

ＢＣＮで使用しているJAICOのデータは、７１００伊日並列文である（詳細については、図５を参照のこと）。最近、単語と単語の対応の作業を始めたが、まだオンラインBCNのデータベースに１００しかアップロードしていない。

			Japanese		Italian	
Domain	Year/s	Sentence Pairs	Tokens	Types	Tokens	Types
N	1989-2001	2000	54,849	5,873	41,238	6,452
PP	2000	2000	74,740	5,539	54,267	7,357
LW	1965-2004	3000	62,849	8,890	48,763	8,893
WP		100				
		7,100	**192,438**	**20,302**	**144,178**	**22,702**

N= News (Yomiuri Shinbun translated into Italian)- Utiyama, Isahara 2003; Nichols et al. 2010
PP= Parliamentary Proceedings (Europarl Shared Task ACL 2007 dataset translated into Japanese)
LW= Literary Works (Excerpts from Japanese novels and their translation)
WP= White Papers (Universal Declaration of Human Rights)

図 5　BCNで使用しているJAICOのデータ

５．まとめ

日伊並列コーパスのデータを検索するフリーウェアオンライン・コンコーダンサを開発した。現在の開発の実験段階では、7000対応付け文と100単語と単語対応付け文の検索が可能である。

これから、日本語とイタリア語について無料のツールを提供するために、コーパスやオンラインデータベースを増やす予定である。

参考文献

Barlow, M. (2008) Parallel Texts and Corpus-Based Contrastive Analysis, in Gómez González, M., Mackenzie, L. and González Alvarez, E. (eds.), *Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives*, Benjamins, 101-121.

Gale, W.A., Church, K.W. (1993) A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19/1, 75-102.

Koehn P. (2005) Europarl: a Parallel Corpus for Statistical Machine Translation, in *Conference Proceedings: the Tenth Machine Translation Summit, 79-86. Phuket, Thailand.*

Japan Foundation (1993, 1998, 2003, 2006, 2009, 2012) *Survey on Japanese-Language Education*

Abroad. Planning and Coordination Section, Japanese Language Dept., Japanese-Language Group. Tokyo.

Johansson, S. (1998) On the Role of Corpora in Cross-linguistic Research, in Johansson, S., S. Oksefjell (eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam and Atlanta, GA, Rodopi, 3-24.

Mariotti, M.M., Mantelli A. (2012) ITADICT Project and Japanese Language Learning. *Acta Linguistica Asiatica* 2/2, 65-82.

Tiedemann, J. (2012) Parallel Data, Tools and Interfaces in OPUS, in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*

Utiyama M., Isahara H. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences, in *Proceedings of the 41st Annual Meeting of the ACL - Association for Computational Linguistics*, 72-79.

Zotti, P. (2013) Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e Applicazioni, in M. Casari, P. Scrolavezza (eds), *Giappone, storie plurali*, 351-363. I libri di Emil-Odoya Edizioni. Bologna.

Zotti, P., Apolloni, R., Matsumoto, Y. (2014) Sentence Alignment of a Japanese-Italian Parallel Corpus. Towards a web-based interface. 言語処理学会第 20 回年次大会発表論文集, 23-26, 18 March 2014.

関連 URL

http://cl.naist.jp/ zottip ~ /

語学学習 SNS の添削ログからの母語訳付き学習者コーパスの構築に向けて

水本 智也 (奈良先端科学技術大学院大学)†

Toward the Construction of a Learner Corpus with Native Language Translation: Using the Data of Language Learning SNS

Tomoya Mizumoto (Nara Institute of Science and Technology)

要旨

学習者の誤用発生の理由の分析や自動誤り訂正には、学習者コーパスが使用される。学習者の意図を考慮して誤用の理由を分析する、もしくは、学習者の意図を考慮して自動誤り訂正するためには、母語訳のついた学習者コーパスが有効であると考える。しかしながら、母語訳付き学習者コーパスの構築には多大な労力を要する。現在公開されている母語訳付き学習者コーパスには、国立国語研究所によって提供されている「作文対訳 DB」があるが、その作文数は限られている。そこで、本研究では語学学習 SNS に注目する。語学学習 SNS では、学習者の書いた作文とその作文に対して添削が行なわれている。この語学学習 SNS のエッセイ中には、学習者自身が母語によって訳を書いているものも存在する。そこで母語で訳が書かれているものを抽出し、学習者の作文、その添削、母語による対訳が付いたコーパスを自動で構築する。本稿では、語学学習 SNS から作られた Lang-8 Learner Corpora の簡単な概要と現在進行中の母語訳付き学習者コーパスの構築について述べる。

1　はじめに

自分の母語以外を学習する第二言語学習者は増加傾向にある。また、第二言語学習を支援するサービスも増加しており、第二言語学習支援に関する研究も盛んに行なわれている。第二言語学習を支援するサービスとしては、多言語対応日本語読解支援システム「あすなろ」[*1]や「語学学習 SNS Lang-8」[*2]がある。第二言語学習支援に関する研究として最も盛んに行なわれているのは、自動誤り訂正である。英語の文法誤り訂正は、共通のデータセットで訂正性能を競うコンペティションである Shared Task が 2011 年から 4 年連続で行なわれている [5, 4, 12, 11]。

† tomoya-m@is.naist.jp

*1 http://hinoki.ryu.titech.ac.jp/asunaro/main.php?lang=jp

*2 http://lang-8.com

また、中国語のスペルチェックの訂正のコンペティション [15] も行なわれており、自動文法誤り訂正が盛んであることがわかる。

自動誤り訂正や学習者の誤用発生の理由の分析には、学習者コーパスを使用する。学習者の意図を考慮して誤用の理由を分析する、もしくは、学習者の意図を考慮して自動誤り訂正するためには、母語訳のついた学習者コーパスが有効であると考える。学習者コーパスの開発が盛んに行なわれており、母語訳の付いていないコーパスは多く公開されている。一方、母語訳の付いた学習者コーパスの開発はほとんど行なわれていない。その理由の 1 つは母語訳付き学習者コーパスの構築には多大な労力を要するためである。現在公開されている母語訳付き学習者コーパスに、国立国語研究所によって提供されている「作文対訳 DB」があるが、その作文数は限られている。

そこで本研究では、母語訳付き学習者コーパスの構築を行なう。学習者コーパスを開発するにあたり、一から、学習者を募り、実際に作文とその対訳を書いてもらうことは非常に大変な作業である。そこで本研究では、Lang-8 Learner Corpora [9][*3]を用いて、そこから母語訳付き学習者コーパスの構築を試みる。Lang-8 Learner Corpora は、自動誤り訂正 [9, 8, 14]、学習者の書いた作文の母語推定 [1, 2]、問題自動生成 [13] に用いられており、自然言語処理による学習者支援に関する研究で効果が実証されている。これまでの自然言語処理による学習者支援の研究では、学習者が学習言語で書いた文とその添削文のみが用いられてきた。本研究では、学習言語の文とその添削文に加えて、母語訳が付いた 3 つ組で構成される母語訳付き学習者コーパスを自動で構築することを目標とする。

2　関連研究

現在、多くの学習者コーパスが存在している。英語の学習者コーパスは、Cambridge Learner Corpus (CLC) [*4]、NUS Corpus of Leaner English (NUCLE) [3]、Konan-JIEM Corpus (KJ) [10]、International Corpus of English (ICLE) [6]、NICT Japanese Learner English (NICT JLE) [7] など数多くある。誤りの訂正、タイプ付与が行なわれているものはあるが、これらのコーパスには母語訳が付いていない。

日本語のコーパスとしては、寺村誤用データ[*5]、大曽による日本語学習者の作文コーパス[*6]、東京外国語大学の日本語学習者言語コーパス[*7]国立国語研究所の作文対訳 DB[*8]などがある。この中で母語訳が付いているコーパスは作文対訳 DB のみである。しかしながら、その数は 1,754 作文と限られており、さらに添削がついているものはおよそ 250 作文だけである。

語学学習 SNS から作られた大規模な学習者コーパスとして、Lang-8 Learner Corpora がある。自然言語処理による学習者支援の研究で用いられているが、これまで使用されたのは学習者の文とその添削文のみであった。

[*3] http://cl.naist.jp/nldata/lang-8/
[*4] http://ilexir.co.uk/applications/clc-fce-dataset/
[*5] http://teramuradb.ninjal.ac.jp
[*6] http://kaken.nii.ac.jp/d/p/08558020.ja.html
[*7] http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=ja
[*8] http://jpforlife.jp/taiyakudb

表 1　Lang-8 に含まれる学習言語ごとのエッセイ数

学習言語	エッセイ数	学習言語	エッセイ数
English	237,843	French	12,392
Japanese	185,991	German	11,111
Mandarin	28,154	Russian	4,069
Korean	21,779	Traditional Chinese	4,052
Spanish	12,606	Italian	3,339

3　Lang-8 Learner Corpora

Lang-8 Learner Corpora は語学学習 SNS Lang-8 から作られた学習者コーパスであり、現在、奈良先端科学技術大学院大学自然言語処理学研究室 (NAIST) で公開されている。Lang-8 は学習者が学習している言語で作文を書くと、その学習言語を母語とするユーザが添削してくれる。また反対に添削された学習者自身も、自分の母語で書かれた他のユーザの作文を添削できる。Lang-8 では、2011 年 10 月時点で 80 言語をサポートしており、317,307 人のユーザが登録している。

NAIST で公開している Lang-8 Learner Corpora は、2011 年までの作文データが収録されている[*9]。Lang-8 Learner Corpora は、580,549 エッセイからなり、様々な言語から構成されている。表 1 に Lang-8 Learner Corpora のページで挙げられているエッセイ数の多いトップ 10 の言語とそのエッセイ数を示す。1 番エッセイ数が多い言語は英語であり、2 番目が日本語、3 番目が中国語となっている。

現在、公開されている Lang-8 Learner Corpora は、JSON 形式で保存されている。図 1 に Lang-8 Learner Corpora の保存形式の例を示す。破線より上がデータの構造を示しており、破線より下が具体例を示している。保存されている情報は、学習者の作文とその添削に加えて、エッセイ ID、ユーザ ID、学習言語、母語である。本研究で構築する母語訳付き学習者コーパスで必要となる、学習者の文（図中の青字下線部分）、その添削文（図中の赤字破線部分）はこの構造から簡単に抽出することができる。一方、母語訳がどの部分であるかは Lang-8 Learner Corpora の JSON 形式では明示的に示されていない。母語訳が書かれているエッセイもあるが、その場合は学習者の書いた文（図中の青字下線部分）に母語訳が書かれている。そのため、母語訳付き学習者コーパスを作成するためには、学習者の書いた文から学習言語の文と母語訳の文を判別して抽出する必要がある。

4　母語訳付き学習者コーパスの構築

本節では、Lang-8 Learner Corpora から母語訳付き学習者コーパスを構築する方法について述べる。母語訳付き学習者コーパスを構築するための処理は、大きく分けると以下の 2 つに分類される。

[*9] 2012 年以降のデータを使いたい場合は、Lang-8 から買うことで使用可能である

["エッセイID", "ユーザID", "学習言語", "母語",
["学習者文1", "学習者文2",...],
[["学習者文1に対する添削文1", "学習者文1に対する添削文2",...],
["学習者文2に対する添削文1", "学習者文2に対する添削文2",...], ...],]

["772869", "227504", "English", "Spanish",
["My prefer color", "Hello people,", "Today I didn't know to tell us.",
"My prefer color is red.", "Because is funny and diferent."],
[[],[],["Today I didn't know how to say it this:"],["My favourite color is
red."],["Because it is funny and different."]]]]

図 1　Lang-8 Learner Corpora の JSON 形式で保存されている情報の例

表 2　対訳候補として抽出されたエッセイ数。「—」は言語を限定せず、全ての言語を表す。

学習言語	母語	エッセイ数
Japanese	—	28,978
Japanese	English	19,885
Japanese	Mandarin	5,586
English	—	33,533
English	Japanese	28,753
—	—	81,560

1. Lang-8 Learner Corpora から、学習言語と母語訳が含まれているエッセイを対訳候補エッセイとして抽出する
2. (1) で抽出したエッセイから学習者の文と母語訳が対訳になっているものを抽出する

現在、作業が済んでいるのは上記の (1) までであり、(2) は現在も進行中である。そのため本稿では、(1) についてのみ述べる。

Lang-8 Learner Corpora から、学習言語と母語訳が含まれているエッセイを抽出する手順は以下の通りである。

1. JSON 形式のファイルから、各エッセイごとに学習者の文とその添削文を取り出す
2. エッセイから取り出された学習者の文に対して、言語判定を行なう
3. (2) で判定された言語と、各エッセイに含まれている学習言語情報、母語情報を比べて同じであればそれぞれ数を数える
4. (3) で得た学習言語で書かれた文と、母語で書かれた文が一定の割合以上のものを対訳候補エッセイとして抽出する

以下、実際の作業について述べる。(2) の言語判定には、language-detection[*10] ツールを使用した。このツールは 53 言語の判定をすることができる。今回は (4) の学習言語と母語の割合

[*10] https://code.google.com/p/language-detection/

表 3　抽出してきた対訳候補の例（対訳になっている例）

Japanese	いま、だいがっくとてもいそがしです。
English	Right now, School is very busy.
Japanese	たくさんテストーがあります。
English	We have many tests.

表 4　抽出してきた対訳候補の例（対訳になっていない例）

English	I have my final Japanese oral exam in a few days.
English	I hope everything goes well on the exam!
Japanese	十一年間ぐらいバイオリンをひいているから、...
Japanese	そこで、夢をかなえるために来年大学で音楽を ...

が 10:3 以上となっているものを対訳候補エッセイとして抽出した。

表 2 に抽出してきた対訳候補エッセイの数を示す。対訳候補エッセイの総数は、81,560 であった。学習言語が日本語である対訳候補エッセイ数は 28,978 で、学習言語が英語の対訳候補エッセイ数は 33,533 であった。表 1 で示したように日本語で書かれたエッセイは 185,991 であるため、およそ 15.6% のエッセイが対訳候補として抽出されている。同様に英語の方も約 14.0% のエッセイが対訳候補として抽出されている。また、英語が母語で学習言語が日本語であるエッセイは 19,885 であった。

表 3 と表 4 に対訳候補として抽出してきたエッセイの一部を例として示す。表 3 は学習言語（日本語）で書かれた文と母語（英語）で書かれた文が対訳になっているような例である。一方、表 4 は学習言語で書かれた文と母語で書かれた文が対訳になっていない例である。今後は、表 4 のような対訳になっていないエッセイを取り除き、対訳になっているエッセイを取り出し、文同士の対応を自動で取る作業を行なう予定である。

5　おわりに

現在進行中である語学学習 SNS からの母語訳付き学習者コーパス構築について述べた。Lang-8 Learner Corpora の中には、学習者が母語訳を書いているエッセイがある。本稿では、学習者の書いた文に対して言語判定を自動で行ない、学習言語で書かれた文と母語で書かれている文が含まれているエッセイの抽出を行なった。その結果、学習言語が日本語であるエッセイでは、約 15.6% のエッセイが対訳候補エッセイとして抽出された。その中には、対訳となっていないエッセイも含まれているため今後は、そのようなエッセイを取り除いていく予定である。

謝辞

Lang-8 のデータ使用に関して，快諾してくださった喜洋洋さんに感謝いたします．本研究は JSPS 特別研究員奨励費の助成を受けたものです。

参考文献

[1] Brooke, J. and Hirst, G.: Native Language Detection with 'Cheap' Learner Corpora, *Proceedings of LCR 2011* (2011).

[2] Brooke, J. and Hirst, G.: Robust, Lexicalized Native Language Identification, *Proceedings of COLING 2012*, pp. 391–408 (2012).

[3] Dahlmeier, D., Ng, H. T. and Wu, S. M.: Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31 (2013).

[4] Dale, R., Anisimoff, I. and Narroway, G.: HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task, *Proceedings of BEA*, pp. 54–62 (2012).

[5] Dale, R. and Kilgarriff, A.: Helping Our Own: The HOO 2011 Pilot Shared Task, *Proceedings of ENLG*, pp. 242–249 (2011).

[6] Granger, S., Dagneaux, E., Meunier, F. and Paquot, M.: *International Corpus of Learner English v2*, Presses universitaires de Louvain (2009).

[7] Izumi, E., Uchimoto, K. and Isahara, H.: Error Annotation for Corpus of Japanese Learner English, *Proceedings of LINC-05*, pp. 71–80 (2005).

[8] Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M. and Matsumoto, Y.: The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings, *Proceedings of COLING*, pp. 863–872 (2012).

[9] Mizumoto, T., Komachi, M., Nagata, M. and Matsumoto, Y.: Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, *Proceedings of IJCNLP*, pp. 147–155 (2011).

[10] Nagata, R., Whittaker, E. and Sheinman, V.: Creating a Manually Error-tagged and Shallow-parsed Learner Corpus, *Proceedings of ACL-HLT*, pp. 1210–1219 (2011).

[11] Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H. and Bryant, C.: The CoNLL-2014 Shared Task on Grammatical Error Correction, *Proceedings of CoNLL Shared Task*, pp. 1–14 (2014).

[12] Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C. and Tetreault, J.: The CoNLL-2013 Shared Task on Grammatical Error Correction, *Proceedings of CoNLL Shared Task*, pp. 1–12 (2013).

[13] Sakaguchi, K., Arase, Y. and Komachi, M.: Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners, *Proceedings of ACL*, pp. 238–242 (2013).

[14] Sawai, Y., Komachi, M. and Matsumoto, Y.: A Learner Corpus-based Approach to Verb Suggestion for ESL, *Proceedings of ACL*, pp. 708–713 (2013).

[15] Wu, S.-H., Liu, C.-L. and Lee, L.-H.: Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013, *Proceedings of SIGHAN Workshop on Chinese Language Processing*, pp. 35–42 (2013).

韻律情報にもとづいた機能表現の抽出

土屋 智行（国立国語研究所言語資源研究系）†
伝 康晴（千葉大学文学部/国立国語研究所言語資源研究系）
小磯 花絵（国立国語研究所理論・構造研究系）

Extraction of Functional Expressions through Prosodic Information

Tomoyuki Tsuchiya (Dept. Corpus Studies, NINJAL)
Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

要旨

機能表現には，複数の語がひとつのまとまりをなしているものや，文節の境界をまたぐものが多く存在する．したがって，語や文節という基準だけでは，機能表現を規定する形式的特徴を十分に抽出することができない．特に，文節の境界をまたぐ機能表現の場合，文の係り受け構造にも影響をあたえるため，結果的に言語使用者による文の解釈との乖離が生じてしまう．そこで，本発表では，言語使用者の解釈を反映する情報として母語話者が発話する際のポーズと韻律句という 2 つの韻律情報を用いて，文節をまたぎながらも，話者がひとつのまとまりとして発話している機能表現の抽出を試みる．具体的には，日本語話し言葉コーパス（CSJ）から隣接する 2 文節の係り受け関係やポーズの有無，韻律情報などの文節間の情報を抽出し，文節内における同様の情報との比較をとおして，文節間でひとつのまとまりをなしている表現を抽出し，その機能的な意味を考察する．

1. はじめに

文の構造は，係り受けによる文節の関係によって記述される．しかし，係り受けを構成する文節および文節同士の関係性は，自立語と付属語という形態論的な規定や，統語的な規定に基いているため，母語話者の直感や，意味的な関係性を直接反映しているものとはいえない．機能表現あるいは複合辞と呼ばれる言語表現は，複数の語が結合することでひとつの機能的な意味を有し，文の意味的な記述に重要な役割を持つものであるが，この機能表現と文節は，しばしばその境界に不一致がみられる．土屋ほか (2007) は，「2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現」を機能表現として取り上げ，機能表現を反映させた係り受け解析を試みている．その中で，土屋ほか (2007) は，機能表現の特定にあたって形態素と係り受けという 2 つのレベルでの調整の必要性を述べている．

上記のとおり，この機能表現の抽出や収集にあたって問題となるのは，複数の形態素が結合しているという特徴と，文節というひとつの単位に拠らないという特徴である．これまでの研

† ttsuchiya@ninjal.ac.jp

究 (国立国語研究所 2001, 土屋ほか 2007) では，一定の意味的な特徴に基づいて機能表現を収集しているが，形態的な基準が明確ではないために，特定の形態素の結合を機能表現と定めることが難しいことが指摘されている．実際に，機能表現の用例集やデータベースの構築はなされている (注連ほか 2007) ものの，機能表現やそれに近いカテゴリーの形態的な基準や範囲は，国立国語研究所 (2001) においても明確に定められているものではない．したがって，文節に拠らない結合表現を機能表現として判断し，抽出するためには，その形態的な特徴を分析していく必要がある．

では，どのような形態的特徴から 形態素の結合を分析していくべきだろうか．本論文では，実際の話者の発話状況を観察することで，形態素の結合の度合いを分析する必要があると考える．複数の構成要素の結合によって形成されるいわば「定型的」な言語表現によって，話者の流暢性が実現されるということは，理論的にも主張されている (Fillmore 1979, Pawley and Syder 1983)．また，言語使用者の感覚との乖離を埋めるためにも，話者の実際の発話状況から抽出する必要がある．

土屋ほか (2014) は，この流暢性に注目し，文節間のポーズ率に焦点を当てて，複雑な係り受け構造をもたらす機能表現の抽出を試みた．その結果，表 1 のような表現を機能表現の例として抽出した．

表 1: 土屋ほか (2014) が抽出した機能表現

係り元文節末語彙素	係り先文節先頭語彙素
{ と/って } いう	事，風，感じ，形，話
{ と/って }	言う，為る，思う，成る，考える
ような	事，形，感じ，気，結果
ように	成る，為る，読める

土屋ほか (2014) で挙げられた機能表現は，形式的な語彙（「事」「形」「為る」「成る」など）に加え，思考や伝達にかかわる語彙（「言う」「思う」「感じ」「話」など）が多用されている例が中心であった．しかし，発話における流暢性には，語間のポーズのみならず，韻律的なまとまりもかかわる．話者がひとつのまとまりとして発話している表現を抽出するには，韻律にも焦点を当てた分析が必要である．

本論文では，係り受け関係にある隣接した文節にたいし，文節という区切りを越えて結合している言語表現の抽出を試みる．その抽出の指標として，音声的な流暢性を取り上げる．ただし，その流暢性の基準には，具体的には，土屋ほか (2014) で使用されたポーズの出現の度合いに加え，韻律に基づいたまとまりとしてアクセント句を流暢性の指標として用い，分析していく．

2. 分析

2.1 方法

分析データは『日本語話し言葉コーパス（CSJ）』（第 3 刷）の RDB 版 (小磯ほか 2012) を使用した．まず，2 つの文節が係り受け関係にあり，かつ隣接している箇所を収集した．次に，収集した箇所から，文節間の流暢性を測る指標として，

- 文節境界およびその前後の長単位間のポーズ情報
- 文節中に出現するアクセント句境界情報

の 2 点を収集した．

文節境界およびその前後の長単位間のポーズ情報としては，係り元文節の末尾から 3 つの長単位（L3, L2, L1）と係り先文節の先頭から 3 つの長単位（R1, R2, R3）を抽出した．また，文節境界およびその前後の長単位間のポーズ情報として，L3〜R3 のうち隣接した長単位同士のポーズの有無（L3L2, L2L1, L1R1, R1R2, R2R3）を抽出した．なお，抽出される文節には，長単位の数が 3 未満のものもある．その場合，存在していない長単位の情報は欠損値となる．たとえば，係り元文節が 2 つの長単位のみで構成されている場合，長単位 L3 およびポーズ情報 L3L2 は欠損値となる．

アクセント句境界の情報としては，文節をまたぎ 1 つのアクセント句が継続するか否か，およびアクセント句境界の位置情報を抽出した．アクセント句境界の位置の分類としては，長単位 L3〜R3 の内部または末尾，長単位 L3 の先頭，（R1〜R3 末尾以外の）文節末尾，その他とした．

3. 結果

3.1 文節の長単位構成

まず，CSJ から収集された文節の情報について示していく．

係り受け関係にある隣接した文節を CSJ 内で検索した結果，全体で 79,005 箇所あった． 係り元の文節のタイプ頻度は全体で 34,658 例，係り先の文節のタイプ頻度は 39,850 例，両者の文節の組み合わせのタイプ頻度は 69,104 例であった．

次に，係り受け関係にある隣接した文節の長単位の構成をみていく．文節を構成する長単位の数は，各文節によって異なるため，構成要素である長単位の数が 3 未満の文節も含まれる．それぞれの長単位の出現数は表 2 のとおりである．

表 2 にあるように，7 割以上の係り元文節が，2 つ以上の長単位から形成されていることが確認できる．係り元の文節ののうち 3 つ以上の長単位から構成されるものは 15% であるので，実質的に 2 つの長単位から構成されるかかり元文節は 6 割程度となる．係り先の文節は，9 割が 2 つ以上の長単位から形成されているが，3 つ以上の長単位で形成されている文節は 4 割弱であった．したがって，2 つの長単位で構成される文節は 5 割程度となる．

表 2: 文節を構成する長単位の出現数とその割合

長単位	出現数 (%)
L3 以上	12170 (15.40%)
L2	61050 (77.27%)
L1	79005 (100.00%)
R1	79005 (100.00%)
R2	70913 (89.76%)
R3 以上	30052 (38.04%)

3.2 長単位間ポーズの観点から

次に，流暢性の基準の 1 つである文節間・文節内ポーズの観点からの分析をおこなっていく．まず，各長単位間に出現する 0.1 秒以上のポーズの出現数とその割合を表 3 に示す．

表 3: L3〜R3 におけるポーズ（0.1 秒以上）の出現数

pause	L3L2	L2L1	L1R1	R1R2	R2R3
ポーズ無	11791	59059	70918	69218	29289
ポーズ有	379	1991	8087	1695	763
ポーズ率	3.11%	3.26%	10.24%	2.39%	2.54%

この表からわかるように，L3 から R3 の間に出現するポーズは，文節間（L1R1）では約 10%，文節内では約 2〜3% であり，一般的に文節をまたぐと流暢性が低くなることが確認できる．その一方で，文節間のポーズ率が相対的に低い場合，係り元文節と係り先文節の繋ぎ目は一息で発話されやすい，すなわち文節間の流暢性が高いと考えられる．そこで，文節間のポーズ率が文節内平均ポーズ率以下（約 2〜3%）となるような L1R1 の組み合わせを抽出した．収集した 79,005 箇所のうち，L1R1 の組み合わせを語彙素のレベルでみると，27,783 例のタイプがあった．その中で，頻度が 10 例以上であり，かつ文節間のポーズ率が 3% 以下となるような例を収集した．その結果，全体で 38 例の表現を収集することができた（表 4）．

表 4 で確認できるように，流暢性が高く，かつ頻度が 100 以上の語彙素の組み合わせとして，「という事」「の方」「の中」「という風」「って言う」「た時」「言う事」「た場合」「の場合」「の時」「言う風」がある．このうち，「という事」「という風」「って言う」「言う事」「言う風」は，「{ と/って } 言う」形式が用いられており，「機能的な意味を担うような言語表現」として土屋ほか (2014) で挙げたものと共通している．頻度が 100 未満の表現でも，共通した表現として「ている事」「という意味」が挙げられる．他にも，「の研究」「の表」「此の図」「た結果」「の影響」など，学術的な領域で用いられるような表現が見られた．

それ以外に一定の機能性を有するものとしては，「の中」「た時」「た場合」「の場合」「の内」のように命題の条件を指定する表現，「の時」「の方」「の頃」「の日」のように特定の時空間

的領域を指す表現，「ている事」「てる事」「為る事」のように特定の事態や行為を動名詞化する表現が挙げられる．

表 4: 流暢性の高い文節境界

L1R1 語彙素	生起頻度	ポーズ頻度	ポーズ率	L1R1 語彙素	生起頻度	ポーズ頻度	ポーズ率
という/事	1680	47	2.80	の/研究	46	1	2.17
の/方	811	14	1.73	に/近い	45	1	2.22
の/中	406	6	1.48	を/含む	40	1	2.50
という/風	405	4	0.99	てる/事	40	1	2.50
って/言う	401	8	2.00	という/意味	40	1	2.50
た/時	311	8	2.57	此の/図	40	1	2.50
言う/事	168	1	0.60	の/表	38	1	2.63
た/場合	163	2	1.23	其の/人	38	1	2.63
の/場合	146	2	1.37	た/結果	38	1	2.63
の/時	142	4	2.82	から/見る	38	1	2.63
言う/風	127	1	0.79	其の/子	37	1	2.70
もう/一つ	94	1	1.06	も/言う	37	1	2.70
の/内	77	1	1.30	為る/事	37	1	2.70
ている/事	74	2	2.70	を/買う	36	1	2.78
の/頃	72	1	1.39	の/数	36	1	2.78
を/受ける	70	1	1.43	の/影響	36	1	2.78
もう/少し	58	1	1.72	を/与える	35	1	2.86
を/掛ける	55	1	1.82	だ/動作継続	34	1	2.94
の/日	47	1	2.13	に/使う	34	1	2.94

3.3　アクセント句境界の観点から

次に，流暢性のもう 1 つの基準として設けたアクセント句境界情報の観点から分析をおこなっていく．先ほど収集した文節の中で，アクセント句境界の出現箇所を検索したところ，全体で 159,671 箇所あった．これらのアクセント句境界から，言いよどみにともなう語断片やフィラー表現，母音不確定の表現など非流暢性にかかわるアクセント句を除外した結果，154,451 箇所をアクセント句境界として抽出できた．それぞれの出現位置の分布は，表 5 のとおりである．

表 5: アクセント句境界とその句末境界音調の分布

出現位置	L3 先頭	L3 内部	L3 末尾	L2 内部	L2 末尾	L1 内部	L1 末尾	
頻度	340	1118	1666	1943	1341	2177	57966	
出現率 (%)	0.22	0.72	1.08	1.26	0.87	1.41	37.53	
出現位置	R1 内部	R1 末尾	R2 内部	R2 末尾	R3 内部	R3 末尾	文節末尾	その他
頻度	1910	9070	4616	39290	1380	15632	13552	2450
出現率 (%)	1.24	5.87	2.99	25.44	0.89	10.12	8.77	1.59

表 5 を見ると，全体の 4 割近くが L1 末尾，すなわち係り元文節末尾でアクセント句が終了

していることが分かる．その次に出現の割合が高いのは，R2 末尾で 25% となっており，係り先文節では 2 番目の長単位の末尾にアクセント句境界が存在していることが確認できる．

次に，収集したアクセント句境界情報から，係り受け関係にある 2 つの隣接した文節をまたいでいるアクセント句を抽出したところ， 全部で 249 表現を収集することができた．非流暢性にかかわるアクセント句を取り除いたものは，合計で 205 例であった．この 205 例のアクセント句の L1R1 の語彙素のパターンを集計し，頻度 2 以上のものを抽出した結果，表 6 のとおりとなった．

表 6: 文節をまたぐアクセント句における L1R1 の頻度（頻度 2 以上）

表現	頻度	表現	頻度
という/事	69	を/見る	3
と/思う	6	って/言う	2
という/風	5	という/感じ	2
に/成る	5	という/状況	2
何々為る/事	5	に/為る	2
が/有る	3	の/時	2
た/場合	3	の/中	2
ない/成る	3	の/方	2
は/無い	3	考慮為る/事	2

表 6 を見ると，最も生起頻度が高いのは「という事」で 69 例，次に「と思う」「という風」であった．これは，土屋ほか (2014) で挙げられた例だけでなく，文節間のポーズの観点から流暢性が高い表現と一致している．この他にも，2 回しか生起していないが，「の時」「の中」「の方」「という感じ」「という状況」など，前節の分析で挙げた表現と共通した意味的特徴をもつ表現が確認できた．具体的には，命題の条件を指定する表現（「た場合」），特定の時空間的領域を指す表現（「の時」「の方」），特定の事態や行為を動名詞化する表現（「何々為る事」「考慮為る事」）が挙げられる．

次に，文節をまたいだ 205 例のアクセント句境界の位置を見ていく．表 7 にあるように，文節をまたいだアクセント句の 4 割近くが R1 末尾までに境界が存在し，9 割近くが R2 末尾までにアクセント句が終了している．これら 205 例のアクセント句が存在する係り先文節のうち，長単位が 3 つ以上存在するものの数を確認したところ，全体で 140 例あった．これは 205 例の文節の 68.29% を占める．したがって，アクセント句境界の位置は，R3 に出現しにくく，全体（表 5 参照）と比較すると，R1 末尾に終了する傾向が見られる．

表 7: 文節をまたぐアクセント句のアクセント句境界の出現位置の分布

	R1 内部	R1 末尾	R2 内部	R2 末尾	R3 内部	R3 末尾	その他	合計
頻度	22	60	68	30	11	9	5	205
出現率	10.73	29.27	33.17	14.63	5.37	4.39	2.44	100.00

4. 考察

以上の分析の結果，文節をまたぎながらも，一定のまとまりをなしていると考えられる表現として，「{ と/って } いう事」をはじめとする形式的な語彙を用いた表現に加え，「の場合」などの命題の条件を示す表現，「の時」「の方」など特定の時空間的領域を示す表現，「為る事」など事態や行為を動名詞化する表現が確認できた．これらの表現は，2 つ以上の語が文節の境界を越えて韻律的に結合していることから，文節をつなぐ機能的な意味を担っている可能性が考えられる．それに対して，土屋ほか (2014) で挙げられた「ように」「ような」を用いた表現や，思考や伝達にかかわる語彙は，いくつか観察されたものの（「という感じ」）全体的にはあまり観察されなかった．

文節をまたいだアクセント句の句末境界は，R1 末尾と R2 末尾に終了し，R3 で終了しにくい傾向が確認できた．この点からも，文節をまたいで結合している表現は，文節境界に依存的な存在で，文節同士の関係性をあらわす機能的な意味を持つ可能性が考えられる．しかし，今回抽出した表現が実質的にそのような機能を持つか否かを知るためには，それぞれの表現の意味を，全体の文脈を踏まえてより精査する必要がある．これについては今後の課題としたい．

本研究は，これまでに対象としてきた言語表現を大幅に広げての分析となったが，結果的に土屋ほか (2014) での分析結果を再確認するようなかたちとなった．流暢性の観点から文節をまたいで結合している表現を抽出することはできたものの，意味的な機能の分析が十分ではないため，注連ほか (2007) や国立国語研究所 (2001) で挙げられている機能表現と一致していない点も多い．今後は，意味的な分析を進めつつ，先行研究で挙げられている機能表現の発話状況を流暢性の観点から確認していくことで，機能表現の形態的・意味的特徴を探っていくことが課題となる．

謝辞　本研究は国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴）による成果である。

参考文献

Fillmore, Charles J (1979). “On fluency.” Daniel Kempler, and William S. Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior*. New York: Academic Press. pp. 85–101.

小磯花絵・伝康晴・前川喜久雄 (2012).「『日本語話し言葉コーパス』RDB の構築」　『第 1 回コーパス日本語学ワークショップ予稿集』　pp. 355–364.

国立国語研究所 (2001).『現代語複合辞用例集』.

Pawley, Andrew, and Frances Hodgetts Syder (1983). “Two puzzles for linguistic theory: Nativelike selection and nativelike fluency.” *Language and communication*, 191, p. 225.

注連隆夫・土屋雅稔・松吉俊・宇津呂武仁・佐藤理史 (2007).「日本語機能表現の自動検出と統計的係り受け解析への応用」　『自然言語処理』, 14:5, pp. 167–197.

土屋智行・伝康晴・小磯花絵 (2014).「発話の流暢性を踏まえた機能表現の抽出と分析」　『言語処理学会第 20 回年次大会論文集』　pp. 19–22.

土屋雅稔・注連隆夫・松吉俊・宇津呂武仁・佐藤理史・中川聖一 (2007).「機能表現を考慮した日本語係り受け解析器学習のためのコーパス作成」　『言語処理学会第 13 回年次大会論文集』　pp. 510–513.

関連 URL

「会話コーパス」ホームページ：`http://www.jdri.org/kaiwa/`

日本語作文推敲支援システム「ナツメグ」における学習者評価実験から見られる課題

八木 豊（株式会社ピコラボ）
ホドシチェク・ボル（明治大学）
阿辺川 武（国立情報学研究所）
仁科 喜久子（東京工業大学）

Problems Found in a Learner Evaluation Experiment Using Japanese Composition Supporting System "Nutmeg"

Yutaka YAGI (Picolab Co., Ltd.)
Bor HODOŠČEK (Meiji University)
Takeshi ABEKAWA (National Institute of Informatics)
Kikuko NISHINA (Tokyo Institute of Technology)

要旨

我々は BCCWJ に科学技術論文を加えたコーパスを使用してレジスター誤り検出を行う日本語作文推敲支援システム「ナツメグ」を開発した。このシステムを使用した、日本語学習者によるレポート形式作文の推敲支援を評価する実験と分析結果について報告する。

実験は 4 つの課題ごとにテーマとプロンプトを示し、学習者が作文する際に一度だけレジスター誤り検出機能による指摘を受けて書き直せる形で実施し、36 名の学習者から 144 作文を収集した。

レポートに近い比較的硬い文章から学習した言語モデルによる評価では、いずれの課題においてもシステムによる指摘後にパープレキシティが減少しており、全体としてレポートに適した文章に修正できたことが確認された。一方、個別のレジスター誤りのうち、特に副詞については、指摘を受けると削除してしまう学習者が多く見られた。これらの事例に対しては修正候補および例文の提示法による改善を検討、提案する。

1　はじめに

国立国語研究所による「現代日本語書き言葉均衡コーパス」（以後 BCCWJ）に加えて、近年、日本語学習者の作文に含まれている誤用に対してタグ付けを施した学習者作文コーパスが公開されてきた。それに伴い、日本語教育の分野では、それらのコーパスおよび自然言語処理の技術を利用して作文に含まれる誤用を自動的に検出し、学習者の作文を支援するための研究が行われている（今村（2012）、水本（2013））。

我々も、レジスターの誤り検出を中心に、日本語作文推敲支援システム「ナツメグ」の開発を進めてきた（八木（2014））。レジスターとは、「社会的な拘束力をもつ言語学上の規範」における言語使用域の変異のことであり、書き手と読み手がどのような関係で、どのようなコンテキストのもとで言語表現を使用するかによって、異なる語彙や文法項目で記述されることを示すものである。学習者作文においては、論文や授業で提出するレポートの中で話し言葉を使用しているなど、場にそぐわない表現がレジスターの誤りに該当する。

本稿では、日本語作文推敲支援システム「ナツメグ」におけるレジスター誤り検出機能について概説し、このシステムを使用した、日本語学習者によるレポート形式作文の推敲

支援を評価する実験と分析結果について報告する。

２　「ナツメグ」におけるレジスター誤り検出機能

ホドシチェク（2011）は、学習者が作文の目的とするレジスターを想定し、目的のレジスターに近いコーパスを準正用データ、目的のレジスターから遠いコーパスを準誤用データと設定したうえで、準正用・準誤用それぞれのデータに含まれる形態素および共起表現の頻度について統計処理を行い、準誤用データの頻度が有意に多い場合に、その表現は目的のレジスターの下ではふさわしくない、即ち、誤用であると判定する手法を提案した。

日本語作文推敲支援システム「ナツメグ」では、この手法を使用してレジスター誤り検出を行う。作文の目的とするレジスターには、アカデミック・ライティングの中でも論文・レポート、申請書などの硬い文書を想定し、科学技術論文、白書、法律を準正用データ、Yahoo!知恵袋、Yahoo!ブログ、国会会議録を準誤用データとして利用する。

レジスター誤り検出の精度を確認するため、日本語教師が学習者作文に対してレジスター誤りなど様々な誤用タグを付与した学習者作文コーパス「なたね」[1]のデータの一部を利用して予備調査を実施した。その結果、日本語教師が指摘したレジスター誤り箇所の再現率が 78.0%、システムの検出結果全体の精度が 77.6%であった（八木（2014））。

３　学習者評価実験

３.１ 実験手順

本稿における評価実験は以下の手順で実施した。

(1) 母語や日本語の学習時間など、背景調査のアンケートに回答
(2) J-CAT（Japanese Computerized Adaptive Test）を受験
(3) システムを利用して作文を入力
(4) システムに対するアンケートに回答
(5) 日本語教師によるコメントを送付

まず、学習者の言語的な背景および実験開始時点の日本語能力を確認するため、背景調査のアンケートへの回答と J-CAT の受験を必須とした。次に、システムを利用した作文入力では、課題ごとにテーマとプロンプトを示し、学習者がそれに関する作文を一通り書き終えてから、一回だけレジスター誤り検出機能による指摘を受けて作文を書き直せるようにした。一回に制限した理由は、レジスター誤り検出の精度が 100%でないシステムに対して、誤用の指摘がなくなるまで修正を繰り返し試行し、過度にシステムに依存した表現になってしまうことを避けるためである。学習者には以下に挙げる 4 つの課題を一つずつ順番に提示し、それぞれ 400 字以上、各課題に取り組む間隔は最低 3 日間となるよう設定した。

課題１：日本人について理解できないこと
課題２：原子力発電の可否
課題３：日本のアニメやゲームソフトはなぜ人気があるか
課題４：インターネット社会の功罪

最後に、システムに対するアンケートを回収し、至らぬシステムのアフターケアとして、学習者が書いた全ての作文に対する日本語教師によるコメントを学習者に向けて送付した。

[1] 学習者作文コーパス「なたね」　http://hinoki-project.org/natane

表１ 学習者の習熟度

習熟度（J-CAT）	人数
母語話者相当	1
上級	7
上級前半	17
中級後半	7
中級	4
計	**36**

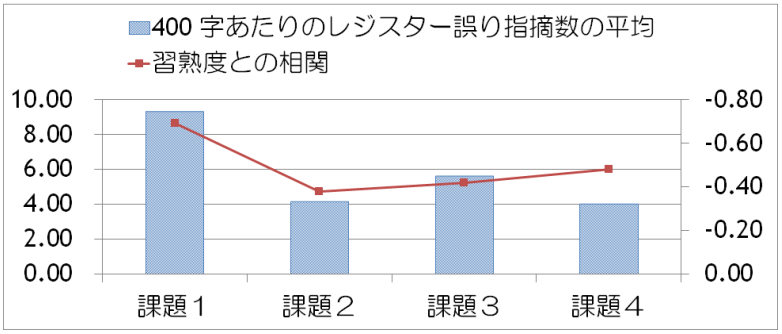


図１ 400 字あたりのレジスター誤り指摘数

３.２ 実験結果

学習者評価実験は 5 つの調査地で実施し、36 名の学習者から 144 作文を収集した。J-CAT の得点に基づいた学習者の習熟度の分布を表１に示す。

図１は、学習者が書いた作文（400 字あたり）に対するシステムのレジスター誤り指摘数の平均および、学習者の習熟度との相関を課題に示したものである。課題１と課題２を比較すると、システムによるレジスター誤り指摘数が大幅に減少していることがわかる。これは、課題１の作文時に、デスマス体による記述や「ちょっと」、「ちゃんと」のような副詞がレジスター誤りに該当するという指摘を受け、課題２の作文時には、そういった表現を使用する学習者が減少したことによるところが大きい。一方、テーマを少し柔らかいものに設定した課題３では、課題２と比較してレジスター誤り指摘数が増加しており、課題の内容によりレジスター誤り指摘数に増減があることがわかった。

また、課題１のレジスター誤り指摘数は学習者の習熟度とやや強い負の相関（-0.69）があり、最初の課題では、やはり習熟度の高い学習者のほうがレジスター誤りを犯しにくいことがわかる。ただし、課題２以降は、レジスター誤り指摘数と学習者の習熟度の相関は弱くなっており、中級以上の学習者は、システムが指摘することでレジスター誤りを学習し、自身で訂正できることを示している。

準正用データとして利用したコーパスから言語モデルを作成し、学習者の作文の訂正前後のパープレキシティの変化を確認したところ、微量ではあるが、いずれの課題においてもシステムによる指摘後にパープレキシティが減少しており、全体としてレポートに適した文章に修正できたことが確認された。パープレキシティが大きく変化しなかった要因の一つとして、今回の実験では、レジスター誤りの可能性を指摘するのみで、代替候補など訂正方針を示すことをしないため、訂正後の表現もレジスター誤りに該当してしまったことが挙げられる。

表２は、今回の実験においてシステムが指摘した全てのレジスター誤りについて、その指摘が妥当なものか否か日本語教師による判定を実施した結果である。全体の精度は 83.05%と、予備調査よりも良い結果を得られた。品詞別では、動詞の精度が著しく悪いが、これは、「驚く」、「言える」、「教える」、「探す」のような一般的な動詞にも関わらず、準正用データに出現しないものが多く見られたことによるものである。

表３は、システムが指摘したレジスター誤りに対して学習者がどのような対応をしたか品詞ごとにまとめたものである。何の対応もせずそのまま残した場合を「未対応」、削除以外の何らかの変更をした場合を「変更」、該当する表現を削除してしまった場合を「削除」としてカウントした。助動詞の削除割合が高いのは、「です」や「ます」を削除してデスマ

表２ システムによる指摘の妥当性

品詞	適切	不適切	計	精度
助動詞	342	0	342	100.00%
副詞	158	29	187	84.49%
動詞	86	97	183	46.99%
ナ形容詞	88	11	99	88.89%
イ形容詞	83	12	95	87.37%
助詞	55	21	76	72.37%
名詞	42	4	46	91.30%
補助動詞	10	0	10	100.00%
連体詞	10	0	10	100.00%
感動詞	3	5	8	37.50%
計	**877**	**179**	**1056**	**83.05%**

表３ 指摘箇所に対する学習者の対応

品詞	未対応	変更	削除	削除割合
助動詞	258	33	51	14.91%
副詞	107	51	29	15.51%
動詞	138	39	6	3.28%
ナ形容詞	69	23	7	7.07%
イ形容詞	62	31	2	2.11%
助詞	51	21	4	5.26%
名詞	33	11	2	4.35%
補助動詞	4	1	5	50.00%
連体詞	7	3	0	0.00%
感動詞	6	0	2	25.00%
計	**735**	**213**	**108**	**10.23%**

ス体による記述を修正するという正しい対応によるものである。一方、副詞の削除割合が高いのは、学習者の語彙が少なく、代替表現が思いつかないために削除してしまっているように見受けられた。また、1,056 件の指摘箇所のうち 735 件が未対応のまま残されているが、このうちのおよそ 8 割は、前述の日本語教師による妥当性判定で妥当であると判断されたものである。こういった箇所については、学習者自身が代替表現を考えたり、調べたりして、訂正を試みてほしいところではあるが、実験後に回収したシステムに対するアンケートでは、「誤り指摘箇所に対してどのように訂正したらよいかがわからない」、「訂正のヒントとなるような情報をもっと提示できないか」という意見が多く寄せられており、今後は、誤用コーパスに頻出する典型的な誤用の場合はその修正候補を、その他に検出した誤用に対しては訂正の参考となるような例文の提示を目指していきたい。

謝辞

本研究は、文部科学省科学研究費補助金基盤研究（C）「日本語作文支援システムで考慮すべき学習者属性情報と提示項目の分析研究」(研究代表者：阿辺川武、研究期間：2012 年 4 月～2015 年 3 月）による助成を得て実施しています。

文献

今村賢治、齋藤邦子、貞光九月、西川仁（2012）「小規模誤りデータからの日本語学習者作文の助詞誤り訂正」自然言語処理, vol.19, no.5, pp.381-400.

水本智也、小町守、永田昌明、松本裕治（2013）「日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得」人工知能学会論文誌, vol.28, no.5, pp.420-432.

ホドシチェク・ボル、仁科喜久子（2011）「作文支援システムにおけるレジスターの扱い」世界日本語教育研究大会 異文化コミュニケーションのための日本語教育 2, pp.522-523.

八木豊、ホドシチェク・ボル、阿辺川武、仁科喜久子（2014）「日本語作文推敲支援システム「ナツメグ」における誤用検出手法の評価」第 5 回コーパス日本語学ワークショップ予稿集, pp.167-170.

連濁に前部要素の音韻的特徴が与える影響：連濁データベースを利用した研究

太田　真理　（東京大学大学院総合文化研究科）
太田　聡　（山口大学人文学部）

Effects of First-Element Phonological Features on Rendaku: A Study Using the Rendaku Database

Shinri Ohta (Graduate School of Arts and Sciences, The University of Tokyo)
Satoshi Ohta (Faculty of Humanities, Yamaguchi University)

要旨

本研究では、連濁データベースに含まれる複合名詞のうち、先行研究で提案された音韻・統語・意味的要因からは連濁を予測できない 16,211 語に対して、前部要素と後部要素が持つ子音、母音、アクセント、モーラ数などの音韻的特徴から、連濁の生起が予測可能かどうかを検証した。これらの例外とされる複合語においても、後部要素の音韻的特徴に加えて前部要素の音韻的特徴を考慮することで、連濁の生起が予測可能である、という仮説を立てた。サポートベクターマシン (SVM) を用いた解析の結果、後部要素のみに基づくモデルの正答率は、チャンスレベルより有意に高く、さらに前部要素の音韻的特徴を加えると 90%以上の複合語で正しく連濁の生起を予測可能であった。以上の結果から、連濁現象の例外とされてきた複合語でも、後部要素の音韻情報に加えて、前部要素の音韻情報を考慮に入れることで、連濁の生起が予測可能であることが示された。

1. 日本語の連濁現象

1.2. 本居・ライマンの法則

日本語では、複合語の後部要素が、有声性に関して対立を持つ無声阻害音（清音：/k/, /s/, /t/, /h/）で始まる場合に、対応する有声阻害音（濁音：/g/, /z/, /d/, /b/）に変化する、連濁現象が知られている。例えば、「ごみ＋はこ」は「ごみばこ」となり（下線部は連濁を示す）、後部要素「はこ」が清音/h/で始まるために、対応する濁音/b/に変化する。一方で、「ごみ＋かご」が「ごみがご」にならないことから明らかなように、連濁は常に生じるわけではない。連濁を阻害する制約として、「複合語の後部要素が濁音を含む場合は連濁が生じない」という「本居・ライマンの法則」が知られている (Lyman, 1894)。この法則から、「ごみ＋はこ」は「はこ」が濁音を含まないために連濁が生じて「ごみばこ」となるが、「ごみ＋かご」では「かご」が濁音/g/を含むために連濁が妨げられて「ごみかご」となることが正しく予測される。

1.2. 連濁に影響する形態・統語・意味的要因

本居・ライマンの法則に加えて、単語の形態論的要因や統語的要因、意味的要因が連濁の生起に影響を与えることが示唆されてきた。例えば、形態論的要因として、「ごみ＋ケース」が「ごみゲース」とならないように、連濁は主に和語で観察され、漢語や外来語では観察されないという語種の効果が知られている（「会社」、「キセル」などの一部の和語化した単語では例外的に連濁が生じる）(Ito and Mester, 2003)。また、統語的要因として、

「ぬりばしいれ（塗り箸専用の入れ物）」と「ぬりはしいれ（漆塗りの箸入れ）」の対比から示されるように、「複合語の木構造中で右枝に来る要素のみで連濁が生じる」という、「右枝条件」が知られている (Otsu, 1980)。

意味的要因として、「やま＋かわ」は「やまかわ」となるが、「たに＋かわ」は「たにがわ」となることから分かるように、前部要素と後部要素が意味的に並列される複合語（並列複合語）では、連濁が妨げられることが知られている (Ito and Mester, 2003)。一方、「ひと＋ひと」が「ひとびと」となるように、同一の単語の繰り返しからなる複合語（畳語）では、連濁が生じる。さらに、「やま＋さき」が「やまざき」にも「やまさき」にもなりうるように、人名や地名などの固有名詞では、連濁の生起に関して曖昧性が生じやすいことも知られている。

1.3. 前部要素が連濁に与える影響

ここまで挙げた要因は、後部要素に含まれる要因（例えば本居・ライマンの法則や語種の効果）、あるいは、前部要素と後部要素の関係によって決まる要因（右枝条件や並列複合語）であった。これに対して、「なかじま」と「ながしま」の対立から示唆されるように、「複合語の前部要素が濁音を含む場合も連濁が生じない」という「強いライマンの法則」が提案されている (Vance, 2005)。この強いライマンの法則は上代日本語では機能していたと考えられるが、現代日本語でも機能しているかどうかについてははっきりしていない (Zamma, 2005; Kawahara and Sano, 2014)。連濁の生起に関してさまざまな要因が影響することを概観したが、いずれの要因にも例外があり、完全に連濁の生起を説明できるわけではない。このように例外が多い連濁現象から一般則を導くためには、多次元の情報に基づいて、統計的に結果を予測する機械学習を用いることが有効であると考えられる。

1.4. 研究の目的

本研究では、連濁データベース (Miyashita and Irwin, 2014) に含まれる日本語複合名詞の中で、音韻・形態・統語・意味的な要因では連濁の生起が説明できない語を対象に、前部要素と後部要素の音韻的特徴（子音・母音・モーラ数・アクセント）の組み合わせによって、連濁の生起が予測できるかどうか検証することを目的とした。我々は、これらの例外として扱われてきた複合語においても、後部要素の音韻的特徴に加えて、前部要素の音韻的特徴も考慮に入れることで、連濁の生起が予測可能である、という仮説を立てた。この仮説を検証するために、統計的機械学習の分野の標準的な手法であるサポートベクターマシン (SVM) を用いて、音韻的特徴から連濁を予測した際の正答率を調べた。

2. 研究方法

2.1. 連濁データベース

本研究で使用した連濁データベースは、国立国語研究所共同研究プロジェクト「日本語レキシコン―連濁事典の編纂」の一環として構築が進められており、本研究ではその最新バージョンである v2.3 を使用した (Miyashita and Irwin, 2014)。連濁データベースには、広辞苑または新和英大辞典に含まれる 32,241 個の複合語が収録されており、これら複合語は以下のいずれかに該当する。

(1) 後部要素が和語で、本居・ライマンの法則によって連濁が阻害されない複合語
(2) 後部要素が漢語または外来語で、例外的に連濁が生じる複合語
(3) 「はしご」のように、ライマンの法則に反して連濁が生じる複合語

このデータベースには、複合語ごとの連濁の有無に加えて、前部要素と後部要素の語種、モーラ数、品詞、さらに後部要素の使用頻度、アクセントなどの情報も記載されている。

本研究では、連濁データベースに含まれる複合語のうち、後部要素が和語の名詞であり、連濁の生起に曖昧性が存在しない複合語を解析の対象とした。また、先行研究から示唆された連濁に影響する音韻・統語・意味の要素を排除するために、以下の基準に該当する複合語は解析対象から除外した。

a. 人名・地名でのみ使用される単語
b. 並列複合語で連濁が生じない単語
c. 畳語で連濁が生じる単語
d. 省略語
e. /b/に由来する/m/を含み、連濁が生じない単語

16,211 個の複合語を解析の対象とし、このうち 13,115 個で連濁が生起し、連濁の生起率は 80.9%であった。

2.2. サポートベクターマシンを利用した連濁生起の予測

SVM とは、機械学習で使われる手法の一つであり、特にパターン認識に関して優秀な学習モデルであることが知られている。図 1 では、○と●を分類する場合を例に SVM の説明を行う。○と●を分ける分離超平面の引き方は無数に存在するが、SVM ではマージン（分離超平面とデータとの距離）が最大となるように、分離超平面を決定するする。この時の分離超平面と最も近いデータのことをサポートベクトルと呼ぶ。

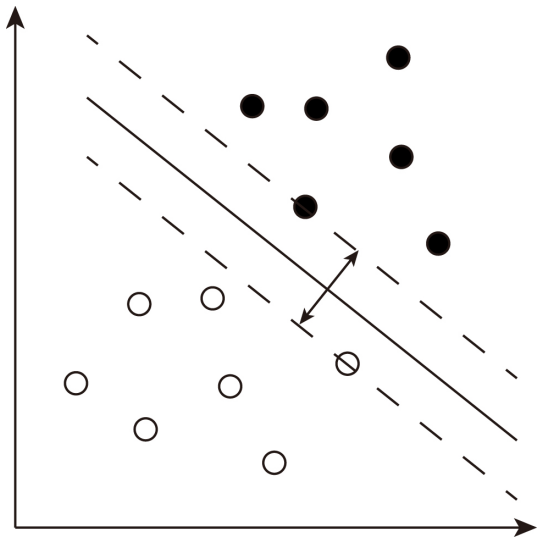


図 1. サポートベクターマシンの概略
実線は分離超平面を示し、矢印はマージンを示す。SVM はマージンが最大となる分離超平面を学習する（破線上のデータはサポートベクトル）。

音韻的特徴から連濁の生起が予測可能かどうかを SVM により検証するために、以下のように単語の音韻情報を 2 値的にコード化して、モデルに取り入れた。まず、前部要素の語末の音節と、後部要素の語頭の音節を取り出した。これらの音節をそれぞれ子音と母音に分け、個々の音素を音韻的特徴として抽出した。例えば「ごみ＋はこ」という複合語において、前部要素の最後のモーラは/m/, /i/という音素を含み、後部要素は/h/, /a/という音素を含むため、これらの音素を 1 とし、それ以外の音素は 0 とする。同様に、前部要素全体に含まれる音素についてもモデルに組み込んだ。例えば前部要素が「やま」の場合、/y/, /m/, /a/という音素を持つ。ここで、/a/は単語中に 2 個含まれるため、/y/, /m/, /a/はそれぞれ 1, 1, 2 とする。

さらに、前部要素のモーラ数、後部要素のモーラ数、複合語全体のモーラ数、後部要素のアクセント（現代語でのアクセントと古語でのアクセント）についての情報も学習に利用した。以上の結果、前部要素に含まれる子音・母音に関する特徴 26 種類、前部要素の最後の音節に関する特徴 22 種類、後部要素の最初の音節に関する特徴 10 種類、モーラ数に関する特徴 3 種類、アクセントに関する特徴 2 種類、使用頻度に関する特徴 1 種類の全 64 種類の情報を利用した。SVM の学習ではデータのスケールをそろえる必要があるため、個々の音韻的特徴 x に対して、

$$\{x - \min(x)\}/\{\max(x) - \min(x)\}$$　（$\max(x)$ は x の最大値、$\min(x)$ は x の最小値）

を適用して、x が 0 から 1 の間の値を持つように調整した (Hsu et al., 2003)。

SVM を使ったデータ解析には、統計言語 R の e1071 パッケージに含まれる svm 関数を使用した (Meyer et al., 2014)。また、学習データに対するオーバーフィッティングを避けるために、50 分割交差確認を行った。50 分割交差確認とは、データを 50 個に分割し、そのうちの 49 個を使用して学習を行い、使用していないデータで SVM の検証を行う、というプロセスを 50 個のデータそれぞれに対して行う手法である。SVM による予測がチャンスレベルよりも高いかを t 検定により調べた。解析対象にした複合語では、81%で連濁が生じていたため、81%に比べて有意に正答率が高い場合、SVM による学習は成功したと考えられる。まず、後部要素の音韻的特徴のみに基づいて連濁を予測するモデルが、連濁の生起をどの程度予測できるかを調べ、さらにこのモデルに前部要素の音韻的情報を加えた場合にどの程度予測精度が向上するかを検証した。最後に、前部要素の音韻的特徴のみに基づいて連濁を予測するモデルも構築し、モデル間の正答率を分散分析により比較した。

3. 結果

3.1. 前部要素と後部要素の音韻的特徴を組み合わせることで説明可能な連濁

後部要素の音韻的特徴のみに基づくモデルの正答率は 83 ± 1.8 %であり、チャンスレベル（81%）よりも有意に高い正答率であった ($p < 0.0001$, $t(49) = 7.5$)。これに対して、後部要素の音韻的特徴に前部要素の音韻的特徴を加えたモデルは、90 ± 1.4 %のデータに対して正しく連濁の生起を予測した（図 2）。このモデルの正答率も、チャンスレベルよりも有意に高かった ($p < 0.0001$, $t(49) = 46$)。これに対して、前部要素の音韻的特徴のみに基づくモデルの正答率は 81 ± 2.2 %であった ($p = 0.58$, $t(49) = 0.56$)。

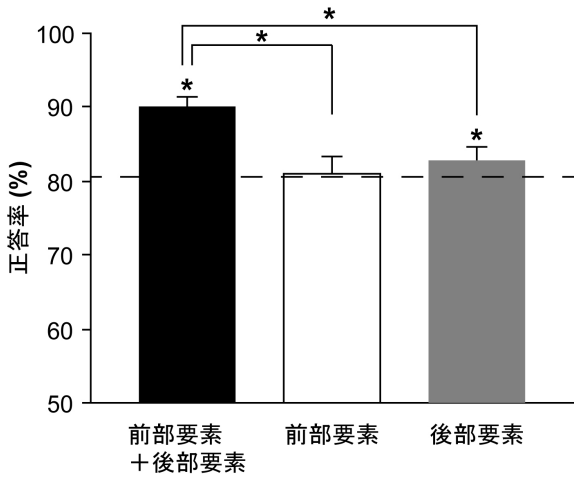


図 2. 前部要素と後部要素の音韻的特徴に基づく連濁の予測

各モデルの正答率を示す（平均±標準偏差）。破線はチャンスレベルを示す。前部要素＋後部要素：前部要素と後部要素の音韻的特徴を組み合わせたモデル。前部要素：前部要素の音韻的特徴のみに基づくモデル。後部要素：後部要素の音韻的特徴のみに基づくモデル。
*：corrected $p < 0.05$。

3.2. モデル同士の正答率の比較

これらの 3 種類のモデルの正答率に対して、繰り返しのない分散分析を行った結果、有意なモデルの主効果が観察された ($p < 0.0001$, $F(2,147) = 329$)。対応のない t 検定により、さらにモデル同士の正答率を比較したところ、前部要素と後部要素の音韻的特徴を組み合わせたモデルは、他のモデルよりも有意に正答率が高かった ($p < 0.0001$, $t(98) > 22$)。また、後部要素の音韻的特徴のみに基づくモデルは、前部要素の音韻的特徴のみに基づくモデルよりも有意に正答率が高かった ($p < 0.0001$, $t(98) = 4.3$)。以上の結果から、後部要素の音韻的特徴から、ある程度連濁の生起が予測可能であり、さらに前部要素の音韻的特徴を加えることで、非常に高い精度で連濁の生起が予測可能であることが明らかとなった。

4. 考察

本研究では、前部要素と後部要素の音韻的特徴から連濁の生起が正しく予測されるかどうかを、機械学習の手法である SVM により検証した。連濁データベースに含まれる日本語複合名詞のうち、先行研究で提案された音韻・統語・意味的要因からは、連濁が予測できない 16,211 個の複合語を対象とした。後部要素の音韻的特徴のみに基づいて連濁を予測するモデル、後部要素の音韻的特徴に加えて前部要素の音韻的特徴も考慮して連濁を予測するモデル、前部要素の音韻的特徴のみに基づいて連濁を予測するモデル、という 3 種類のモデルの比較検討を行った。後部要素の音韻的特徴のみに基づくモデルでは、83%の複合語に対して、正しく連濁の生起を予測可能であった。しかしながら、今回使用したデータのうち、81%の複合語では連濁が生じていたため、このモデルでは十分に予測精度が改善したとは言い難い。これに対して、前部要素と後部要素の音韻的特徴に基づくモデルでは、90%以上の高い精度で連濁の生起を予測可能であった。これらの結果は、連濁現象の例外とされてきた複合語では、後部要素の音韻的特徴に加えて、前部要素の音韻的特徴も考慮することが、正しく連濁の生起を予測するために必須であることを示唆する。一方

で、前部要素の音韻的情報のみに基づくモデルでは、チャンスレベルと比べて有意に予測精度が向上しなかったことから、現代日本語では「強いライマンの法則」は機能していないことが示唆される。また本研究は、このように多くの要因が関係する言語現象にとって、統計的機械学習の手法が極めて有効であることを示すものである。

本研究で用いたSVMは、与えた学習データに対して1つの分離超平面を学習する手法であり、学習データ中のどの情報が、予測力の向上に重要であったのかは明らかでない。今回考慮した音韻的特徴には、子音・母音・アクセント・モーラ数という性質の異なる特徴が混在しており、それぞれ連濁の予測に対する寄与率が異なることが予想される。今後の研究では、判別分析やロジスティック回帰分析のように、個々の説明変数に対して予測への寄与率が計算される統計手法を用いることで、重要度の高い音韻的特徴の絞り込みを行う予定である。また、今回は考慮に入れなかった統語・意味的要因や、前部要素のアクセント等をモデルに組み込むことで、さらなる予測精度の向上を目指す予定である。

謝　辞

本研究で使用した「連濁データベース」をご提供くださった山形大学のアーウィン先生並びにモンタナ大学の宮下先生に感謝いたします。

文　献

Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin (2003) “A Practical Guide to Support Vector Classification,” http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Ito, Junko and Armin Mester (2003) *Japanese Morphophonemics: Markedness and Word Structure*, Linguistic Inquiry Monograph 41, Cambridge, MA: The MIT Press.

Kawahara, Shigeto and Shin-ichiro Sano (2014) “Testing Rosen’s Rule and Strong Lyman’s Law,” *NINJAL Research Papers*, 7, pp. 111–120.

Lyman, Benjamin S. (1894) “Change from surd to sonant in Japanese compounds,” *Oriental Studies of the Oriental Club of Philadelphia*, pp. 1–17.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, and Chih-Chen Lin (2014) “Package ‘e1071,’” http://cran.r-project.org/web/packages/e1071/e1071.pdf.

Miyashita, Mizuki and Mark Irmin (2014) The Rendaku Database v2.3 (http://www-h.yamagata-u.ac.jp/~irwin/site/Rendaku_Database.html よりダウンロード可能)

Otsu, Yukio (1980) “Some Aspects of Rendaku in Japanese and Related Problems,” *MIT Working Papers in Linguistics*, 2, pp.207–227.

Vance, Timothy J. (2005) “Sequential Voicing and Lyman’s Law in Old Japanese.” In Salikoko S. Mufwene, Elaine J. Francis & Rebecca S. Wheeler (eds.), *Polymorphous linguistics: Jim McCawley’s legacy*, pp. 27–43. Cambridge: The MIT Press.

Zamma, Hideki (2005) “Correlation between Accentuation and Rendaku in Japanese Surnames: With Particular Attention to Morphemes.” In Jeroen van der Weijer, Kensuke Nanjo & Tetsuo Nishihara (eds.), *Voicing in Japanese*, pp. 157–176. Berlin: Mouton De Gruyter.

書　名　第6回 コーパス日本語学ワークショップ予稿集
発行日　平成26年9月2日
発行者　国立国語研究所 言語資源研究系・コーパス開発センター
http://www.ninjal.ac.jp/organization/chart/03/
http://www.ninjal.ac.jp/organization/chart/06/
連絡先　〒190-8561　東京都立川市緑町10-2
大学共同利用機関法人 人間文化研究機構 国立国語研究所コーパス開発センター内
電　話　042-540-4300（代表）