第5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5回 コーパス日本語学ワークショップ

予稿集

2014年3月6日、3月7日

主催　国立国語研究所 言語資源研究系・コーパス開発センター

会場　国立国語研究所

第5回 コーパス日本語学ワークショップ

予稿集

2014年3月6日(木)／3月7日(金)

Program [プログラム]

3月6日 (木)

13:00～13:10 **■挨　拶**　前川 喜久雄

13:10～15:10 **■口頭発表 (1)**

色彩語メタファー表現に関する研究 ―「バラ色」と「灰色」を例にして―
▷王 軒

形態素解析辞書「中古和文UniDic」を利用した古典学習教材の作成
▷須永 哲矢

闘病記からわからないこと、わかるかもしれないこと：
――闘病記コーパスを用いた患者像の一致についての考察――
▷臼田 泰如、島本 裕美子、久保 圭、荒牧 英治

動詞「教える」と共起する授受表現について
▷朱 炫姝

15:10～16:10 **■ポスター発表 (1) グループA**

コーパスコンコーダンサ『ChaKi.NET』の連続値データ型 (2) ―読み時間の表示―
▷浅原 正幸、池本 優、森田 敏生

外れ値検出手法を利用した Misleading データの検出
▷吉田 拓夢、新納 浩幸

領域間距離を利用した能動学習による語義曖昧性解消の領域適応
▷小野寺 喜行、新納 浩幸

uLSIF による重み付き学習を利用した語義曖昧性解消の領域適応
▷菊池 裕紀、新納 浩幸

トピックモデルと概念辞書による日本語の語義曖昧性解消
▷佐々木 悠人、古宮 嘉那子、小谷 善行

コーパスのメタデータを用いた日本語におけるレジスターのモデル化
▷ホドシチェク・ボル

SVM を用いたインドネシア語連体従属接続詞の判定システム
▷ Wahyu Purnomo、古宮 嘉那子、小谷 善行

合議による語義曖昧性解消の領域適応のための確信度の調整
▷古宮 嘉那子、奥村 学、小谷 善行

CRF による Web コーパスからのアニメタイトルに対する訳語の自動抽出
▷山崎 舞子、森田 一、古宮 嘉那子、小谷 善行

遡及的に構成される発話連鎖の諸特徴
▷鈴木 佳奈、山本 真理、鈴木 亮子、伝 康晴

16:10～17:10 **■ポスター発表 (1) グループB**

コーパスから取得した用例で対象物が認識可能であるのか
▷保田 祥

『現代日本語書き言葉均衡コーパス』の文境界修正作業の進捗
▷小西 光、中村 壮範、田中 弥生、浅原 正幸、今田 水穂、山口 昌也、
前川 喜久雄、小木曽 智信、山崎 誠、丸山 岳彦

自発モノローグにおける息継ぎ音自動アノテーションの試み
▷浅井 拓也、菊池 英明、前川 喜久雄

形容詞用例データベースの構築
▷山口 昌也

新しい日本語辞書定義文型の策定に向けて
▷佐藤 理史、夏目 和子

統計的現代語訳モデルを用いたセンター試験古文問題解答
▷横野 光、星野 翔

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
▷八木 豊、ホドシチェク・ボル、阿辺川 武、仁科 喜久子
BCCWJ図書館サブコーパスの文体情報検索ツールによるテキスト分析
▷柏野 和佳子、中村 壮範
独話音声と対話音声の発話末のF0変化
▷石本 祐一、小磯 花絵

3月7日（金）

10:00～12:00 **■口頭発表（2）**

Webを母集団とした超大規模コーパスの開発 ―収集と組織化―
▷浅原 正幸、今田 水穂、保田 祥、小西 光、前川 喜久雄
語義曖昧性解消を対象とした領域固有のシソーラスの構築
▷新納 浩幸、國井 慎也、佐々木 稔
複雑ネットワークの視点による二字熟語の構造について
▷山本 健
Kachakoにおけるコーパスアノテーションの重ね合わせ
▷狩野 芳伸

12:00～13:00 **昼食・休憩**

13:00～14:00 **■ポスター発表（2） グループA**

拡張CaboChaフォーマットの仕様拡張
▷松吉 俊、浅原 正幸、飯田 龍、森田 敏生
言語単位と文の長さが品詞比率に与える影響
▷山崎 誠
言語は経済に連動して変化する
▷森 秀明
形容詞による属性叙述表現の構造に基づく意味分析
▷朴 俌倧
『今昔物語集』の捨て仮名
▷冨士池 優美、岩崎 瑠莉恵
動詞「きく」のコロケーションについて
―WEBコーパスと日本語母語話者・上級日本語学習者の主観的判断から―
▷小野 正樹、朱 炫姝、デヒピティヤ・スランジ・ディルーシャ、李 国玲、
スワンナクート・パッチャラーバン
『太陽コーパス』の口語会話部分に現れる動詞の待遇表現の計量的分析
―聞き手に対する尊敬待遇表現を中心に―
▷近藤 明日子
二つの文末形式「～感じだ」と「～感じがする」について
▷中平 詩織
『書き言葉的』として指導する必要のある語の分析
―『現代日本語書き言葉均衡コーパス』を利用して―
▷木田 真理、柏野 和佳子
「Nを通じて」と「Nを通して」について
▷趙 海城

14:00～15:00 **■ポスター発表（2） グループB**

感情動詞における二格とデ格の交替について
▷張 麗
教科書コーパスにおける教科別特徴語
▷立山 智絵

『今昔物語集』における文体対立語 ―巻12と巻27の語彙比較による―
▷田中 牧郎
BCCWJのタグ情報の修正について
▷長谷川 守寿
文末表現から発話冒頭の談話標識的表現へ
▷東泉 裕子、髙橋 圭子
BCCWJにおける敬語形式の使用実態
▷小木曽 智信、神田 龍之介、近藤 明日子
『法華百座聞書抄』のテンス・アスペクト形式：『今昔物語集』との比較
▷野田 高広
『太陽コーパス』の漢字表記語に対する熟字ルビについて
▷間淵 洋子
近代雑誌資料における程度副詞類の使用状況
▷市村 太郎
最高裁判所判決の文体特徴の経年変化
▷矢野 信

15:00～16:30 **■指定討論**
▷今田 水穂、小椋 秀樹、近藤 明日子、菊池 英明、狩野 芳伸、ホドシチェク・ボル
16:30～17:00 **■全体討論**
17:00 **■閉　会**

※3月6日（木）10:30-12:00 「中納言」「ChaKi.NET」それぞれの相談タイムを設けます。

Contents [目次]

■口頭発表（1）

色彩語メタファー表現に関する研究 ―「バラ色」と「灰色」を例にして― 1
王 軒

形態素解析辞書「中古和文 UniDic」を利用した古典学習教材の作成 11
須永 哲矢

闘病記からわからないこと、わかるかもしれないこと：
――闘病記コーパスを用いた患者像の一致についての考察―― 21
臼田 泰如、島本 裕美子、久保 圭、荒牧 英治

動詞「教える」と共起する授受表現について 31
朱 炫姝

■ポスター発表（1） グループ A

コーパスコンコーダンサ『ChaKi.NET』の連続値データ型（2） ―読み時間の表示― 39
浅原 正幸、池本 優、森田 敏生

外れ値検出手法を利用した Misleading データの検出 49
吉田 拓夢、新納 浩幸

領域間距離を利用した能動学習による語義曖昧性解消の領域適応 57
小野寺 喜行、新納 浩幸

uLSIF による重み付き学習を利用した語義曖昧性解消の領域適応 63
菊池 裕紀、新納 浩幸

トピックモデルと概念辞書による日本語の語義曖昧性解消 71
佐々木 悠人、古宮 嘉那子、小谷 善行

コーパスのメタデータを用いた日本語におけるレジスターのモデル化 81
ホドシチェク・ボル

SVM を用いたインドネシア語連体従属接続詞の判定システム 91
Wahyu Purnomo、古宮 嘉那子、小谷 善行

合議による語義曖昧性解消の領域適応のための確信度の調整 97
古宮 嘉那子、奥村 学、小谷 善行

CRF による Web コーパスからのアニメタイトルに対する訳語の自動抽出 103
山崎 舞子、森田 一、古宮 嘉那子、小谷 善行

遡及的に構成される発話連鎖の諸特徴 109
鈴木 佳奈、山本 真理、鈴木 亮子、伝 康晴

■ポスター発表（1） グループ B

コーパスから取得した用例で対象物が認識可能であるのか 117
保田 祥

『現代日本語書き言葉均衡コーパス』の文境界修正作業の進捗 127
小西 光、中村 壮範、田中 弥生、浅原 正幸、今田 水穂、山口 昌也、前川 喜久雄、小木曽 智信、山崎 誠、丸山 岳彦

自発モノローグにおける息継ぎ音自動アノテーションの試み 137
浅井 拓也、菊池 英明、前川 喜久雄

形容詞用例データベースの構築 147
山口 昌也

新しい日本語辞書定義文型の策定に向けて 153
佐藤 理史、夏目 和子

統計的現代語訳モデルを用いたセンター試験古文問題解答 161
横野 光、星野 翔

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 167
八木 豊、ホドシチェク・ボル、阿辺川 武、仁科 喜久子

BCCWJ 図書館サブコーパスの文体情報検索ツールによるテキスト分析 171
柏野 和佳子、中村 壮範

独話音声と対話音声の発話末の F0 変化 181
石本 祐一、小磯 花絵

■口頭発表（2）

Web を母集団とした超大規模コーパスの開発 ―収集と組織化― 189
浅原 正幸、今田 水穂、保田 祥、小西 光、前川 喜久雄

語義曖昧性解消を対象とした領域固有のシソーラスの構築 199
新納 浩幸、國井 慎也、佐々木 稔
複雑ネットワークの視点による二字熟語の構造について 207
山本 健
Kachako におけるコーパスアノテーションの重ね合わせ 213
狩野 芳伸

■ポスター発表（2） グループA
拡張CaboChaフォーマットの仕様拡張 223
松吉 俊、浅原 正幸、飯田 龍、森田 敏生
言語単位と文の長さが品詞比率に与える影響 233
山崎 誠
言語は経済に連動して変化する 243
森 秀明
形容詞による属性叙述表現の構造に基づく意味分析 253
朴 俌傑
『今昔物語集』の捨て仮名 261
冨士池 優美、岩崎 瑠莉恵
動詞「きく」のコロケーションについて
―WEBコーパスと日本語母語話者・上級日本語学習者の主観的判断から― 271
小野 正樹、朱 炫姝、デヒピティヤ・スランジ・ディルーシャ、李 国玲、スワンナクート・パッチャラーパン
『太陽コーパス』の口語会話部分に現れる動詞の待遇表現の計量的分析
―聞き手に対する尊敬待遇表現を中心に― 279
近藤 明日子
二つの文末形式「～感じだ」と「～感じがする」について 287
中平 詩織
『書き言葉的』として指導する必要のある語の分析
―『現代日本語書き言葉均衡コーパス』を利用して― 293
木田 真理、柏野 和佳子
「Nを通じて」と「Nを通して」について 301
趙 海城

■ポスター発表（2） グループB
感情動詞における二格とデ格の交替について 311
張 麗
教科書コーパスにおける教科別特徴語 317
立山 智絵
『今昔物語集』における文体対立語 ―巻12と巻27の語彙比較による― 323
田中 牧郎
BCCWJのタグ情報の修正について 331
長谷川 守寿
文末表現から発話冒頭の談話標識的表現へ 341
東泉 裕子、髙橋 圭子
BCCWJにおける敬語形式の使用実態 347
小木曽 智信、神田 龍之介、近藤 明日子
『法華百座聞書抄』のテンス・アスペクト形式：『今昔物語集』との比較 357
野田 高広
『太陽コーパス』の漢字表記語に対する熟字ルビについて 365
間淵 洋子
近代雑誌資料における程度副詞類の使用状況 373
市村 太郎
最高裁判所判決の文体特徴の経年変化 383
矢野 信

口頭発表（1）

3月6日（木）　13:10～15:10

色彩語メタファー表現に関する研究
－「バラ色」と「灰色」を例にして－

王　軒（東北大学大学院文学研究科）†

Collocation and Metaphor: A Corpus-based Research on Color Words in Japanese

－Taking *barairo* and *haiiro* as Examples－

Xuan Wang (Graduate School of Arts and Letters, Tohoku University)

１．はじめに

日常生活においては、人々は自分の感情、考えを分かりやすく相手に伝達するとき、文字通りの発話での表現がある一方、メタファーの使用も必要不可欠である。メタファー研究の一考察として、本稿では、我々と深いかかわりを持っている「色彩語」に注目した。具体的には、色彩語「バラ色」と「灰色」を例にして、『現代日本語書き言葉均衡コーパス』(以下、BCCWJ と略す）から収集した用例を分析することによって、色彩語を用いて成立したメタファー表現の特徴及びメタファーの役割を考察することを目的とする。

２．先行研究と問題設定

坂本・古牧(2005)は、心理実験と文学作品コーパスを用いて、「赤い」「青い」「黄色い」「白い」「黒い」いわゆる基本色彩語[1]が名詞を修飾する形式のメタファー表現を対象にして考察した。心理実験の結果から、「色彩語が本来物理的色彩をもたない心的状態などの抽象物を指示する名詞を修飾している色彩語メタファーでは、色彩語単独の場合とは異なる意味が喚起され、特に否定的なイメージが喚起される傾向が見られた」と指摘した。また、色彩語が否定的なイメージをもつ抽象物を指示する名詞を修飾している例が多いことが文学作品コーパスを用いて確認された。

また、坂本・内海(2007)では、喩辞である色彩形容詞と比喩辞である名詞の意味の相互作用に基づく意味変化のパターンから、色彩形容詞の否定的意味の喚起が生じるかどうかを実験的に検討した。その結果、対象とした色彩形容詞メタファー表現について否定的意味が喚起されると指摘した。

今までの研究は主に基本色彩語を対象に、色彩語メタファーによってどのようなイメージが喚起されるのかをめぐって論じられていたが、色彩語メタファー表現の特徴及び色彩語を用いた表現の意味変化においてメタファーの役割について論じることが稀である。更に、研究対象として、基本色彩語の下位分類の色彩語と非基本色彩語をめぐって言及されていない。上述した先行研究を踏まえて、本研究は「非基本色彩語」である「バラ色」と「基本色彩語」の下位分類に属する「灰色」を対象にして、考察する。

† wangxuan777@hotmail.co.jp

1 連続した色空間を日常生活に必要な最も基本的な単位に分類したものである。最も多くの言語において共通するのは白と黒である。さらに、赤、青、黄、緑という概念も多くの言語に存在している。日本をはじめとする大部分の言語では、更に六つの中間的な色である、灰色、茶色、ピンク、オレンジ、紫を加えて、全 11 個の基本色彩語が存在する。

３．データと分析方法

コーパスデータはBCCWJのWeb版アプリケーション「中納言」を利用し、色彩語「バラ色」と「灰色」の表記法(平仮名、片仮名、漢字表記)を配慮した上で抽出した。その結果、「バラ色」と「灰色」それぞれ計230例と1160例が収集された。具体的には表1に示す。(以下説明する際、表記の形式としては「バラ色」と「灰色」に統一した。)

表1　抽出サンプルの集計

色彩語	表記法									合計
	平＋平	平＋片	平＋漢	片＋平	片＋片	片＋漢	漢＋平	漢＋片	漢＋漢	
バラ色	1	0	34	1	2	138	1	0	53	230
灰色	8	0	0	0	19	0	13	0	1120	1160

次ぎに、収集したサンプルを手作業によって、分析用例と除外例に分けてまとめた。まず、「バラ色」に関しては、以下の用例を本研究の検討の対象から除外した。(計33例)

①複合色彩形容詞：(計9例)　深いバラ色、薄バラ色、淡いバラ色など
②韻文(短歌・詩・俳句)：(計6例)　例文(略す)
③固有名詞：(計2例)「バライロガモ」など
④引用・題名：(計10例)「バラ色の雲」『ぼくのバラ色の人生』など
⑤メタ表現：(計1例)「～薔薇色は、明るい希望に満ちた世界をたとえることが多い。」
⑥明喩表現：(計2例)
「嫌煙者はまるでタバコ千円をバラ色のように考えていますね。」など

また、「バラいろ」「薔薇いろ」「ばらいろ」の用例が1件ずつ検索された。その中で、「ばらいろ」の例は引用である。「バラいろ」と「薔薇いろ」の例は色彩そのものを指しているけれども、表記法の使用頻度が高くはないため、今回の分析対象から除外した。

続いて、「灰色」に関しては、計233例の用例を本研究の検討の対象から除外した。

①複合色彩形容詞：(計69例)　薄灰色、青灰色、淡い灰色など
②韻文：短歌・詩・俳句　(計16例)例文(略す)
③固有名詞：(計48例)　灰色熊、灰色狼、灰色アルバトロス、灰色カビ病など
④引用・題名：(計32例)『灰色の潟』『灰色のなかに』「灰色の男たち」など
⑤～っぽい：(計15例)　灰色っぽい汁、灰色っぽい覇気のない会社など
⑥灰色系：(計3例)「彼女にはキスマークの灰色系の上下を進めました～」など
⑦メタ表現：(計4例)「闇夜の猫はどれも灰色」など
⑧ファンタジー人物：(計43例)「灰色の武神」「灰色の紳士」
⑨明喩表現：(計1例)「灰色のような運勢」
⑩不明：(計2例)「灰色の玉」「悲しげな灰色の風」

検索された「ハイイロ」の用例は、全部「ハイイロガン」などの固有名詞類であり、今回の分析対象から除外した。

続いて、辞書『大辞林』三省堂(1989)は色彩語「バラ色」と「灰色」の意味を次のように記述している。

・ばらいろ＜薔薇色＞
①明るい、うすくれない色。淡紅色。「バラ色の頬」
②(比喩的に)しあわせや希望に満ちている状態。輝かしい未来などを象徴する色。
「バラ色の人生」「バラ色の未来」
・はいいろ＜灰色＞
①灰のように薄黒い色。ねずみ色。グレー。「灰色の空」
②暗い気持ちで、心に喜びのない状態。さびしく陰気なこと。「灰色の人生」
③黒と白の中間の色であるところから、犯罪などの容疑が、十分には晴れていないこと。

辞書による色彩語「バラ色」と「灰色」の意味の解釈に基づいて、二つの色彩語を含む表現の意味の分布をまとめた。表 2 と表 3 が示したとおりである。

表 2　「バラ色」を含む表現の意味の分布

色彩名	バラ色（125例）				薔薇色（43例）				ばら色（29例）				計（197例）
「バラ色」の意味	字義の意味		比喩的意味		字義の意味		比喩的意味		字義の意味		比喩的意味		
表現形式	～の	以外	～の	以外	～の	以外	～の	以外	～の	以外	～の	以外	
サンプル数	37	20	39	29	18	12	3	10	11	4	6	8	
色彩統計	57				30				15				102例（52%）
メタファー表現統計			68				13				14		95例（48%）

表 3　「灰色」を含む表現の意味の分布

色彩名	灰色（911例）				灰いろ（11例）				はいいろ（5例）				計（927例）
「灰色」の意味	字義の意味		比喩的意味		字義の意味		比喩的意味		字義の意味		比喩的意味		
表現形式	～の	以外	～の	以外	～の	以外	～の	以外	～の	以外	～の	以外	
サンプル数	532	268	80	31	9	1	0	1	4	1	0	0	
色彩統計	800				10				5				815例（88%）
メタファー表現統計			111				1				0		112例（12%）

本稿では、「バラ色」を含むメタファー表現 95 例、「灰色」を含むメタファー表現 112 例を抽出した。これらの表現に対して、それぞれ「色彩語の～」という表現形式とそれ以外の表現形式に分けて考察する。色彩語メタファー表現の共起語を確認する一方、「色彩語の～」以外の表現形式をまとめることによって、その共起表現及び当該色彩語の比喩的意味の使用上の活性化を明らかにする。最後に、分析対象とする「バラ色」と「灰色」を含むメタファー表現の特徴及びメタファーの役割を考えていく。

４．結果と考察

４.１　「バラ色」を含むメタファー表現に関する考察

今回、BCCWJ を用いて、「バラ色」を含むメタファー表現 95 例を抽出した。その中に、「バラ色の～」という表現形式が 48 例であり、「バラ色の」以外の表し方は 47 例が抽出された。それぞれのまとめは下記の図 1 と表 4 のとおりである。

まず、図 1 を見ると、[バラ色の人生]が 10 例、「バラ色の未来」が 9 例、「バラ色の日々」/「バラ色の毎日」が 4 例、「バラ色の生活」が 4 例、「バラ色の夢」が 5 例である。そのほか、「バラ色の時間」「バラ色の幸福」「バラ色の期待」「バラ色の二十一世紀」「ばら色のスローガン」「バラ色の解決策」などの表現も現われた。具体例は下記のとおりである。

(1)～評価、これを繰り返すことで活気が出る、活性化する仕組みになるわけです。バラ色の人生だったら、感性が影響されますから、物事がみんな明るく見えるわけです。～

（坂元昂, 2003『感性と独創力』）

(2)～「そうじゃなく」「気にするな。ねえやが来てからは、バラ色の人生だった。そして今の俺は、お前が側にいてバラ色だ」。～

（高月まつり, 2005『家政夫様には逆らえません』）

(3)～寝太郎がいない日々を。大変だろうけれど、私を煩わすモノが一つ減った日々を。薔薇色の未来がやってくると信じるほど私はおめでたくない。～

（Yahoo!ブログ，2008）

(4)～経済法則の貫徹するところ、ついにはアヘン貿易すら合法化せざるをえないであろうと、清国貿易の前途に甘い薔薇色の夢を描きだす。～

（衞藤瀋吉, 2004『近代中国政治史研究』）

(5)～各政党は、選挙前には立派な公約を掲げ、ばら色のスローガンを叫びます。～

（Yahoo!ブログ，2008）

(6)～公務員になって**ばら色の人生**が迎えられると思って、折角の勉強を無駄にしたくないものですから。～　（Yahoo!知恵袋，2005）

(7) ～地元のヒーロー的ゴルファー、ラナルフ(マット・デーモン)と結婚して、**バラ色の生活**を送っているが、第一次大戦に召集されて戦地で～

（小野亜優, 2001『SCREEN』）

例(1)、例(2)と例(6)はともに「バラ色の人生」というメタファー表現が用いられている。例(1)の場合、我々は気分がいいので、「物事がうまくいっている」とき、「ずんずん調子がよく」「どんどん物事を進めていく」、このような「バラ色の人生」いわゆる「幸先がいい」だったら、「物事も明るく見える」ということである。例(2)はなぜ「ばら色の人生」というのか、「お前が側にいてバラ色だ」という「幸せ・幸福」に満ちているからである。似ている状態が例(7)にある「ばら色の生活」即ち「結婚生活」においても見られる。また、例(6)にある「公務員になってばら色の人生を迎える」という表現を通して、「バラ色」のその「前途が明るい状態」という意味が十分理解できる。

例(3)の場合、「煩わすモノとしての「寝太郎(旦那？)」がいないと、「薔薇色の未来」がやってくる、このように「薔薇色」と「未来」との組み合わせによって、邪魔者無しのこれからは幸先が良くなり、「幸せに満ちている状態」になると感じられる。ただ、話者はそこまで幸せになれると思わないようである。

例(4)をみると、近代中国では、「アヘン貿易の合法化」により、当時の「清国貿易の前途」も大きく開けているように思われる。このような「薔薇色」と「夢」との共起により、「輝かしい未来への憧れ」が生き生きと感じさせるようになると思われている。例(5)にある選挙前に叫んでいた「ばら色のスローガン」も、「輝かしい未来」「いい幸先」へのアピールになる。

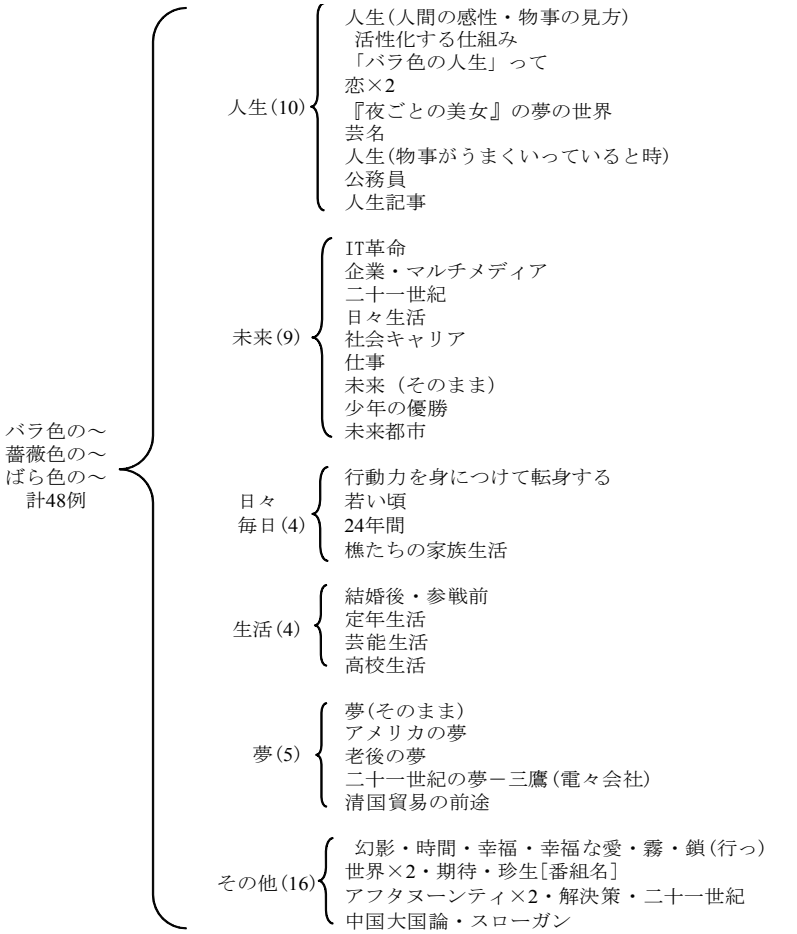


図 1　「バラ色」を含むメタファー表現① ―「バラ色」の共起語　(48 例)

図 1 が示したとおり、「バラ色の日々・毎日」「バラ色の生活」「バラ色の夢」など、それぞれいくつかの派生したメタファー表現も抽出された。「バラ色の日々・毎日」の中で、「行動力を身につけて転身する」「若い頃」「24 年間」「札掛に住んでいた樵たちとその家族の生活の日々」の形でまとめたが、これは筆者が用例の前後の文脈によってまとめたものである。また、「バラ色の生活」の中では、更に詳細になり、身の回りのことと深く関わっていることが分かった。たとえば、「定年生活」「芸能生活」「高校生活」等の表現が見られた。一方、「バラ色の夢」の場合、「夢」そのままのほか、「アメリカの夢」「老後の夢」の形で使われている。更に、「清国貿易の前途」例(4)のような指示対象を修飾する表現も現れた。

続いて、「バラ色の」以外の表し方(計 47 例)について分析を行う。表 4 に示したとおり、まず、「バラ色」の共起表現が多種多様であることが分かった。

表 4　「バラ色」を含むメタファー表現② ―「バラ色」の共起表現　(47 例)

文脈表現	共起語	用例数（計47例）
～ばかりではない/だった/です/だ/で/である/ではない/じゃない/ではなくなった	この世　毎日　未来　定年以後 結婚　恋　世界×2　企業社会 長生き　人生　政権交代　社会生活 会社勤めの毎日　超高齢化社会　心	16
～輝いて/輝かせて/に輝いている/に輝き	絶大の栄光に包まれた(未来) 頬×3 世界の縁・未来	6
～に描きすぎなかった	近代都市の自由自治	1
～とまでいわない/という話/こと/とはいう/といえば	未来 生活 定借住宅を買えば 毎日	4
～にする	人生	1
～に塗りつぶされている	情報化社会の未来	1
～に見える/見えている/見ようとした/を見た	転職 明治百年 その他	3
～になる/となる	幻想 戦略による見返り・老後・人生	4
～に語る	情報技術	1
～に包まれている	個々の人生、会社生活	1
～に染め ～に染まった	頬 結末[地獄]	2
～を期待する	二十一世紀・その他	2
～一辺倒に唱えられていた	情報社会	1
～らしい	未来	1
～に上気した	顔	1
～と思えない	未来	1
～にいろどる	北海道の開拓の前史	1

「バラ色」を含むメタファー表現の指示対象が「人生」「未来」などの他、「結婚」「会社勤めの毎日」「転職」「超高齢化社会」「長生き」「世界の縁」「政権交代」「戦略による見返り」「北海道の開拓の前史」なども抽出された。要するに、「バラ色」という色彩語の比喩的な意味は辞書による「幸せ」「希望に満ちている状態」「輝かしい未来」の他、「順風満帆の先」「恵まれた」「幸先がいい」などの感情が示されたように、われわれの日常生活に深く関わり、さまざまな場面で頻繁に使われていることが分かった。具体例を見ると、

(8)～転職先として思い描いている会社もありますが、**単に疲れがたまって転職がバラ色に見えているだけかもしれません**。みなさんなら、適職・高給・残業・休みなし、と、それなりの職・それなりの給料・残業なし・休みきっちり、ではどちらを選びますか。理由もあわせて教えていただけるとありがたいです。～

（Yahoo!知恵袋，2005）

(9)～時がたつとともに、<u>この戦略による見返りはバラ色となり</u>、長い目で見ればすべて「めでたし」「めでたし」ということになろう。～
（大前研一，1999『企業参謀―戦略的思考とはなにか』）
(10) ～しかし、残念ながら、<u>北海道の開拓の前史をばら色にいろどる逸話のみをつづりあわせて</u>、この時期をえがくことは、かえって真実をゆがめはしないかとおそれる。～
（成田龍一，1995『日本残酷物語』）

「転職の先」をめぐって、例(8)が挙げられた。「バラ色」と「～に見えているだけかもしれません」という表現との共起によって、「順風満帆の先」「恵まれた先」に予想通りになれない事実を、より躍動感のある表現として、より婉曲的に我々に伝達する効果が与えられた。

「バラ色」を含むメタファー表現は、よく「詩」や「文学作品」などで用いられるという先入観をいだいているだろう。上述した用例を見ると、「バラ色」と共起し用いられたメタファー表現はごく自然さと親切さを持ち、われわれの身の回りの様々な場面に応用し、誰でも普通に使える表現であることが分かった。

さらに、例(9)にある「戦略による見返り」と例(10)に現れた「北海道の開拓の前史」の用例を見ると、「バラ色」を含むメタファー表現の魅力、次々新しく派生してきた表現が社会と人間生活・人間活動に応じて変化しつつあり、我われの身の回りとの親近さに対して更に切実に感じられるようになった。そして、「二十一世紀」「明治百年」などのような表現から見ると、派生したメタファー表現の一特徴として挙げられるのは、一時性であると考えられる。また、「頬」「アメリカの夢」「中国大国論」などの表現は、前後の文脈や事前に一定程度の知識を持っていないと、理解しにくい傾向がある。「頬」の場合、ただの「肌の色」であるのか、恋や愛情で「幸せに満ちている頬」であるのか、断言できないゆえ、派生したメタファーが曖昧性を持っていると考えられる。

また、「バラ色の生活」の中に、「定年生活」「芸能生活」「高校生活」などの表現も現れたが、このようなメタファー表現を見ると、派生したメタファー表現のもう一つの特徴として挙げられるのは可変性であると思われている。

４.２　「灰色」を含むメタファー表現に関する考察

今回、BCCWJ を用いて、「灰色」を含むメタファー表現 112 例を抽出した。その中に、「灰色の～」という表現形式が 80 例であり、「灰色の」以外の表し方は 32 例が抽出された。それぞれのまとめは下記の図 2 と表 5 のとおりである。

辞書によると、「灰色」の比喩的な意味に関しては、「暗い気持ちで、さびしく陰気なこと」と「犯罪などの容疑が、十分には晴れていないこと」が述べられている。これらの意味をもとに、よく用いられている表現として「灰色の人生」「容疑が灰色のまま釈放される」などが挙げられる。

図 2 を見ると、「灰色高官」のメタファー表現が一番多く、計 23 例があった。似ているような表現として、「灰色議員」が挙げられる。また、「灰色決着」「灰色の判定」及び「その他②」のなかに現れた「学位」「記録」「領域」「情報」「文献」措置」「ビジネス」などの表現は、「灰色」の二番目の比喩的な意味、すなわち「疑惑のあること・十分には晴れていないこと、はっきりしないこと」を示している。具体的な用例は下記に示す。

(11) ～だが、八十六年に「労働者派遣法」が成立するまでは、中途半端な、<u>灰色のビジネス</u>と見られていた。～
（芦崎治，2004『逃げない人を、人は助ける－テンプスタッフ成長の哲学』）
(12)～二十一世紀のユーゴスラビアの未来を占う重要な闘いはしかし、蓋を開けてみれば限りなく<u>クロに近い灰色の判定</u>で幕が下りた。～
（木村元彦，2005『終わらぬ「民族浄化」セルビア・モンテネグロ』）

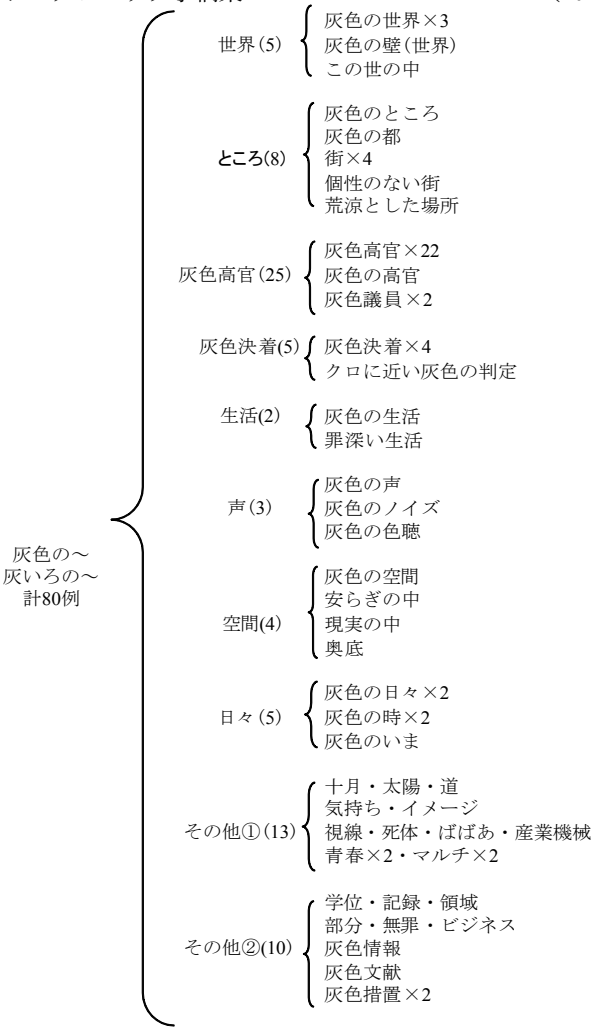


図 2　「灰色」を含むメタファー表現① ―「灰色」の共起語　(80 例)

一方、「灰色」のもう一つの比喩的な意味、いわゆる「希望がなく暗い気持ちで、活気のないこと」を用いて使われた表現として、「世界」「世の中」「ところ・場所」「街」「生活」「声」「空間」「現実の中」「日々」「気持ち」「イメージ」「青春」「灰色ばばあ(気難しいところのある方)」などが抽出された。具体例は下記のとおりである。

(13) ～漆黒の蔭、不吉な悪夢、すべての幼い男の子の脳裏に、その**灰色の奥底**に、永久に棲みつく恐怖の幻。～
(ジャッキー・ヴォルシュレガー(著)/安達まみ(訳)，1997『不思議の国をつくる』)

(14)～古代ローマ人の生活は、質素で規則正しい、**灰色の生活**であった、と書物は教えて呉れるが、それはまた、地中海や東方の質の異なる文明を一つに融合統一したローマ文明の秘密なのかもしれない。～
(森有正，1999『森有正エッセー集成』)

(15)～これから先、忍には、片足を失ったままでの長い人生が、はじまるわけである。沖の人生も、当然、暗く長い**灰色の道**となるであろう。～
(城山三郎，1976『毎日が日曜日』)

(16)～汚ない。**灰色の個性のない街**。無秩序。個々に主張しあい統一感がない。～
(松尾羊一，1989『ザ・渋谷研究―若者を吸引する渋谷現象とは何か』)

(17)～生きる意志も動く意志もない、**灰色の荒涼とした場所**にさまよい込んでいた。～
(イサドラ・ダンカン(著)/山川亜希子(訳)/山川紘矢(訳)，2004『魂の燃ゆるままに―イサドラ・ダンカン自伝』)

例(13)の場合、彼は安らぐことができない悲劇的な人物であり、「漆黒の蔭」「不吉な悪夢」のような「恐怖の幻」にずっと彼の魂はさいなまれている。「男の子の脳裏」を「灰色の奥底」で喩え、その「暗い気持ち」「寂しく陰気なこと」「憂鬱なこと」がよく表われている。また、例(14)の「灰色の生活」は「古代ローマ人の生活」のありさまであり、「質素で規則正しい」という表現を通して、「灰色」と「生活」との共起による作られたこのメタファー表現には「つまらない」という感情に満ちていることを、より切実に感じるようになった。例(15)の場合、「片足を失ったままでの長い人生」、このようなことにあったら、誰でも最初は暗い気持ちで心に喜びを持っていないはずである。「灰色の道」というメタファー表現は、婉曲的にその「暗さ」「つらさ」「悲しさ」「希望を持てず憂鬱な気持ち」を伝達した。また、例(16)と(17)に関しては、「灰色」はそれぞれ「無秩序」「統一感がない」「個性のない街」と「荒涼とした場所」と共起した。共起語前にある「個性のない」「荒涼とした」という修飾語が表わすように、「生気に欠けている」「街」と「場所」が、「灰色」との共起によって、その「さびしく陰気な気持ち・感情」は更に躍動的に感じられる。

また、図2を見ると、「灰色」を含むメタファー表現のなかでは、同じ表現の頻出度が低い。「世界」「街」「決着」「日々」「時」「青春」「措置」「マルチ」のほか、ほとんど 1 例ずつである。この点は、「バラ色」を含むメタファー表現と異なっている。

続いて、「灰色の」以外の表し方(計 32 例)について分析を行う。下記の表 5 のとおりである。

表 5　「灰色」を含むメタファー表現② ―「灰色」の共起表現　(32 例)

文脈表現	共起語	用例数
～だった/です/だ/で/ではないか/であるか	すべて×2・街の印象 灰色高官・都会 (軍事状態)・すべての理論	7
～なのだ	根拠はない	1
～が心のなかにまでしみこんできて	世界の灰色	1
～ってことで	灰色(まとめ)	1
～はなく	曖昧さもない	1
～黒/クロに近い灰色(だ)	結論・その他×2	3
～というように見られている	有罪	1
～にぬりこめられた	青春	1
～になる	人生・老老介護	2
～がかった/かって	入国ビジネスのイメージ・音色・ピアニッシモ（最弱音）	3
～灰色に灰色をかさねてえがく	哲学×4	4
～な(無血無感情な顔を)する	顔	1
～と言えるでしょう	(不倫であるかどうかの判断)	1
～に濁っていく	顔・声・心	1
～になりかねない	長生き	1
―を～にする	人生	1
その他	朝・日本の将来	2

「灰色」を含むメタファー表現の指示対象に関しては、前述した「灰色高官」「人生」のほか、「街の印象」「都会」「結論」「青春」「老老介護」「哲学」「長生き」「日本の将来」などの共起語も現れてきた。表 5 のまとめにより、「灰色」の二つの比喩的な意味の使用頻度

はほぼ同じであることが分かった。二番目の意味としての「疑惑のあること」「はっきりにしていないこと」の意味は、『国会会議録』『白書』などのような正式な場面でよく用いられているようであるが、実際には、日常生活における「不倫であるかどうかの判断」のような表現も存在していた。「灰色」を含むメタファー表現も、われわれの生活に深く関わり、どんな場面での使用も可能であると考えられる。

(18)～「**灰色なのだ。**我々もマリーダそのものを調査している。港湾工事にかこつけて世界中の港に顔を売り、麻薬を陸揚げし易くしている節もある。だが、証拠はない。少なくもチリに於いてはね」～

（仁科透，1996『黒いコンドル―泣くな！海外駐在課長』）

(19)～～ソロモンは、ここの最初のところでは、**灰色がかった音色**とゆっくりしたテンポを選んでいるのである。～

（ヨーアヒム・カイザー(著)/門馬直美(訳)，2003『ベートーヴェン 32 のソナタと演奏家たち』）

(20) ～小沢リーダーの場合、有罪と決まったわけではないのですが、限りなく**クロに近い灰色というように見られている**ということでしょう。　～

（Yahoo!ブログ，2008）

(21) ～「八十六歳と七十七歳の**老老介護**なんて、聞いただけで真っ暗になるでしょ。だから、せめて**灰色になる**くらい、明るくしなくちゃね」と。

（稲葉敬子，2000『オトコの介護力』）

例(18)は、「麻薬調査」において、「証拠はない」ため、「灰色なのだ」というメタファー表現が用いられていた。これは、「灰色」の比喩的な意味「犯罪の有無(黒白)がはっきりしないこと」をもとにしている。似ているのは例(20)、「～有罪と決まったわけではない」「クロに近い灰色というように見られている」という表現である。

また、例(19)の場合、「灰色」と「音色」との共起である。ここでの「和音」は「穏やか」「過剰な情緒と感情が必要ない」、ゆっくりとしたテンポであった。このような「無感動のように奏し出された」テンポは「灰色がかった音色」という色彩語メタファー表現を修飾することによって、その「寂しい気持ち」「穏やかな気持ち」がよりよく伝えられてきた。

最後に、例(21)を見ると、指示対象は「八十六歳と七十七歳の老老介護」になり、それと「せめて灰色になるくらい、明るくしなくちゃね」と共起し、色彩語メタファー表現として用いられていた。「老老介護」は事実上だったら、「難しい」「やり辛い」「悲しい」「憂鬱」という側面があるが、そこから「暗い気持ち」「寂しく陰気な気持ち」を含んでいるようになり、「灰色になるくらい」「明るくしなくちゃね」という表現と共起することによって、逆に、少しでも「希望に満ちている」「明るい先」を感じさせるようになった。これこそ、色彩語メタファー表現の魅力的なところであると思われる。

「バラ色」をめぐってまとめた特徴「一時性」と「可変性」については、「灰色」の中にある表現、「灰色の時」と「灰色のいま」、「灰色の生活」と「灰色の罪深い生活」からも見られる。「バラ色」と「灰色」を含むメタファー表現の中で、その共起語、あるいは指示対象は多様であり、同じ表現としての「日々」「生活」「時」などが挙げられる。一方、「高官」「文献」などの表現は「灰色」独特な修飾対象となっている。逆に「夢」という抽象的な概念に関する表現は「バラ色」を用いた表現の中に、いくつかの用例が見られたが、「灰色」を用いた表現の中に見られなかった。また、二つの色彩語が持っている「比喩的な意味」は、「色のイメージ」と深く関わっていると感じられた。「バラ色」のイメージとして、「幸福・幸せ・明るい・生命力」が挙げられ、「灰色」のイメージの場合、「薄暗い・あいまい・寂しさ・地味・無気力」が挙げられた。色のイメージは、連想からのものもあり、後天的な習慣、経験などから受けるものもある。色はただ色として認識されるだけではない。色の持つイメージをもとに、物理的な色彩を持っていない表現を潤色する。更に、修飾され

た表現はまた膨大な色彩語メタファー表現を派生しつつあり、一般的に共通認識されているものとなっている。この点から見ると、色彩語メタファー表現が「ある程度の一貫性」を持っていると考えられる。

５．まとめと課題

本研究では、コーパスベースに色彩語「バラ色」と「灰色」を含むメタファー表現の使用実態を考察した。これによって、日常生活において、色彩語メタファー表現が頻繁に使われ、次々新しく派生したメタファー表現の生産性が大きく、且つ無限大であるという結論が得られた。更に、抽出したデータを形式と意味の点から分析した結果、「バラ色」「灰色」を含む色彩語メタファー表現に「一時性」「曖昧性」「可変性」と「ある程度の一貫性」という四つの特徴を示した。

また、色彩語メタファー表現の意味変化におけるメタファーの役割について、「抽象的なものを具体化にしてものを簡潔的に説明できる」、「話し手が自分の感情を的確に表現することができ、聞き手との間に親近感を産み出し、感情を引き起しやすくなる」、「聞き手が話し手の言及した特定の行動に注意を引くようになる」三点を指摘した。

最後に本研究の考察における問題点と今後の課題について三点述べる。

本稿では、非基本色彩語「バラ色」と基本色彩語の下位類にある「灰色」を例にして、その使用実態を確認した。そのうえ、二つの色彩語を用いたメタファー表現の特徴及び色彩語表現の意味変化におけるメタファーの役割を論述した。ただ、二つの色彩語だけで表されたメタファー表現の特徴はまだ不十分であると思う。たとえば、「桃色」「黄金色」、あるいは基本色彩語「赤」を用いたメタファー表現「赤の他人」のような色彩語及び表現を取り扱う場合、どのような結果が見られるのか。この点について、更なる実証的な考察が必要である。

また、今回の考察は、日本語を中心にして行われてきた。「他文化的な要因」「社会背景」「コンテクスト」などに関しては多く触れていない。この研究をよりよく充実させるために、この大事な一環をめぐって検討することは今後の課題である。

メタファーの動機付けについて二つの要因があると考えられる。すなわち「類似性」と「経験における共起性」である。色彩語メタファー表現においては、この二つの要因がどのように表されているのか、また、メタファーの役割・機能について理論的にも実証的にも説明が必要となり、それに関しては今後引き続き研究する必要がある。

謝　辞

本稿の執筆において多くの方々に貴重なアドバイスやご指導を頂きました。特にデータの分析(メタファー表現の確認)について、指導教官から丁寧かつ熱心なご指導を賜りました。この場を借り、皆様方に深く感謝を申し上げます。

文　献

近江源太郎(2008)『色の名前に心を読む　－色名学入門－』研究社出版
楠見孝(2007)『メタファー研究の最前線』ひつじ書房
坂本真樹・内海彰(2007)「色彩形容詞と名詞の相互作用による色彩形容詞メタファーの認知効果」認知科学 14-3 pp.380-397
坂本真樹・古牧久典(2005)「心理実験とコーパスを用いた色彩語共感覚メタファーの表現効果研究」日本認知言語学会論文集 5 pp.106-116
鍋島弘治郎(2011)『日本語のメタファー』くろしお出版
松本曜(2003)『認知意味論』大修館書店
G.レイコフ, M.ジョンソン著;渡部昇一・楠瀬淳三・下谷和幸訳(1986)『レトリックと人生 』大修館書店

形態素解析辞書「中古和文 UniDic」を利用した古典学習教材の作成

須永 哲矢（昭和女子大学）†

Preparation of Teaching Materials in Classical Literature using an Electrical Dictionary for Morphological analysis "Chuko-Wabun UniDic"

Tetsuya Sunaga (Showa Women's University)

１．はじめに

国立国語研究所では形態素解析辞書「中古和文 UniDic」が公開されている。「中古和文 UniDic」は平安仮名文学作品を対象とした画期的な形態素解析辞書であり、あくまで現代語を対象とした従来の解析辞書では無力であった古典資料を高精度で解析することが可能となった。これを利用することで、日本語の歴史的研究の幅が大いに広がることが期待され、実際、国立国語研究所「日本語歴史コーパス」構築に際してもこの「中古和文 UniDic」が利用されている。今後さまざまな方面での利用が期待できる「中古和文 UniDic」であるが、その利用の在り方はなにも研究利用にとどまるものではない。教育面においても、主に高等学校での古典学習等、活躍の場面は多用に考えうる。本稿では、形態素解析辞書「中古和文 UniDic」の教育転用の一つとして、古典学習教材の作製法を紹介する。

２．高校の古典学習の在り方と、従来の一般的な学習資料

高等学校の古典学習の要点としては、助動詞の意味や識別、敬語、そして現代語訳などが定番である。文法教育に関しては、学習者にとって「文法」というもの自体が苦手意識を引き起こすせいか、このところ扱いが軽くなる傾向にあるようであり、そのため、「この語の並びでなぜこういう訳になるのかわからない」というまま学習を重ねるという学習者も多く見受けられる。これに対し現代語訳に関しては、高校で扱うような主要作品の訳なら容易に手に入る時代になっている。その結果、文法への理解がなく、自力では意味を取ることができないまま、訳文を暗記して試験に臨むことを繰り返し、学習が進んでも知識が積み重ならない、というような状態の学習者も多い。

高校生向けの訳としては、助動詞や敬語を意識した、単語ごとの逐語訳に近い形の学習用の訳文も以前から出回っているが、知識のある訳者自身が単語ごとの１対１対応を意識して訳出したにせよ、知識をこれから学んでいく学習者側は、原文と訳文を見比べて、単語間の対応がすぐにとれるわけではない。学習用逐語訳の訳文を単語ごとに切って原文と対応をとっていけば学習効果は高まるであろうし、学習用の訳の理念も本来はそこにあると考えられるが、そのような面倒な作業をする学習者は現実には少ない。学習のための逐語訳は、当初の理念・目的を理解されないまま、実際にはただ話の内容を理解するためだけに読まれ、「わかりにくい文章」との印象をもたれて終わってしまう。

単語に区切る、文法を意識する、という観点からは、いわゆる教科書ガイドと呼ばれるような、原文を全て品詞分解し、その隣に活用の種類や活用形を付した学習資料も以前からあるが、こちらは本文に上記のような品詞情報を付与することが主眼であり、結局訳は別掲であることが多く、文法知識が訳に結び付くには至らない。紙面構成からも、従来の教科書ガイドは本文の隣に品詞情報を小書で書き込んだ時点で非常に読みにくく、さらに逐語訳を付与したうえで、見やすい紙面を構成するというのは極めて困難であろう。また、原文を品詞分解してしまった以上、単語レベルではない、より大きな表現単位に対する注釈などは付与しにくくなってしまう。

古典学習においては、品詞分解、活用形や助動詞といった文法知識を前提に、単語ごと

† tetsuyasunaga@gmail.com

の対応関係を意識した逐語訳を与え、最終的な解釈としての訳文にたどりつく、という理解のしかたが理想的であると考えるが、従来の一般的な学習資料では、品詞分解は品詞分解だけ、訳文は訳文だけにならざるを得ず、各要素が結び付いた総合的な学習資料の作成は考えにくい。これに対し、「中古和文 UniDic」を用い、本文テキストを形態素解析した Excel ファイルを利用すれば、各分野の情報をを総合的に表示した学習教材が作成可能になると考える。

３．形態素解析辞書「中古和文 UniDic」

高校での学習という視点から言えば、形態素解析辞書は「機械が自動で品詞分解して、活用の種類や活用形を書き出す」という作業をしてくれる、ということになる。機械に品詞分解をさせ、その結果をもとに、必要・目的に応じて逐語訳や注釈を書き加えていくことで、必要な情報をひとつにまとめた学習教材を自由に作成することができる。

現在公開されている「中古和文 UniDic」は中古和文 UniDic ホームページより無償でダウンロードできる。ホームページでの指示に従ってダウンロード・インストールを行えば、特に機械処理に関する詳しい知識がなくとも、誰でも手軽に形態素解析を行う事ができる。

操作画面は図 1 のとおり。操作用ツール「和文茶まめ」が用意されており、ユーザはマウス操作で簡単に解析が行えるようになっている。古典本文を txt 形式で用意しておけば、あとはこの操作画面でファイルを指定してやれば、自動で品詞分解が完了する。



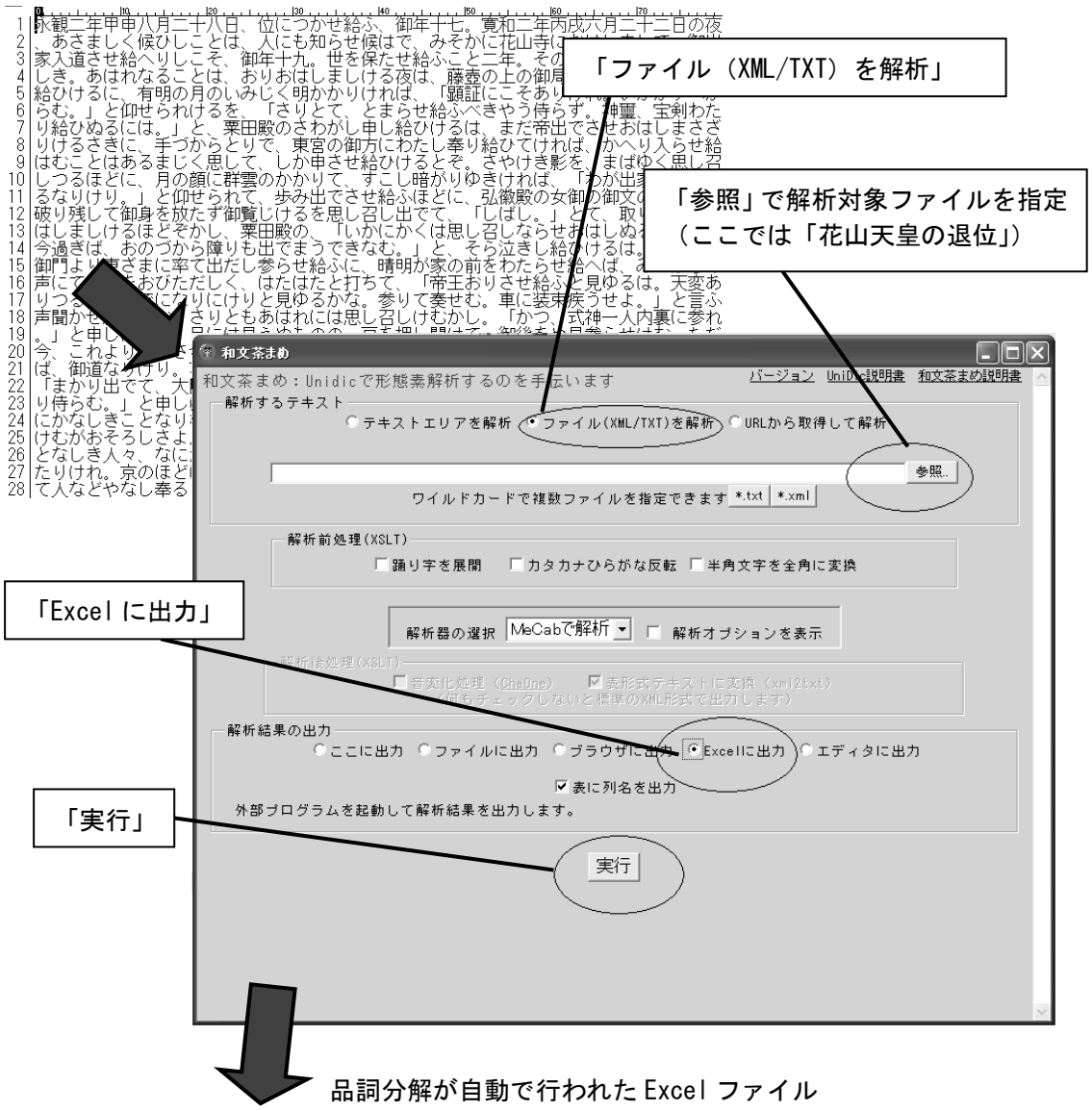
図 1　操作画面（「和文茶まめ」）

今回は txt ファイルを Excel に出力して加工する、という手順を想定している。実際の解析手順は次の通り。

・「ファイル（XML/TXT）を解析」にチェック
・「参照」をクリックし、解析対象とするファイル名を指定
・「Excel に出力」にチェック
・「実行」をクリック

以上の操作のみで図 2 に示すような、品詞分解済みの Excel ファイルが出力される。

もとの TXT ファイル（ここでは例として「花山天皇の退位」）



品詞分解が自動で行われた Excel ファイル

	A	B	C	D	E	F	G	H	I	J	K
1	出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
2	花山天皇の	B	あはれ	アワレ	アワレ	哀れ	名詞-普通	名詞-形状	詞可能	アワレ	和
3	花山天皇の	I	なる	ナル	ナリ	なり-断定	助動詞	文語助動詞	連体形-一	ナリ	和
4	花山天皇の	I	こと	コト	コト	事	名詞-普通	名詞-一般		コト	和
5	花山天皇の	I	は	ワ	ハ	は	助詞-係助詞			ハ	和
6	花山天皇の	I	、			、	補助記号-読点				記号
7	花山天皇の	I	おり	オリ	オリル	下りる	動詞-一般	文語上二段	連用形-一	オル	和
8	花山天皇の	I	おはしまし	オワシマシ	オワシマス	おわします	動詞-非自	文語四段-	連用形-一	オワシマス	和
9	花山天皇の	I	ける	ケル	ケリ	けり	助動詞	文語助動詞	連体形-一	ケリ	和
10	花山天皇の	I	夜	ヨ	ヨ	夜	名詞-普通	名詞-一般		ヨ	和
11	花山天皇の	I	は	ワ	ハ	は	助詞-係助詞			ハ	和

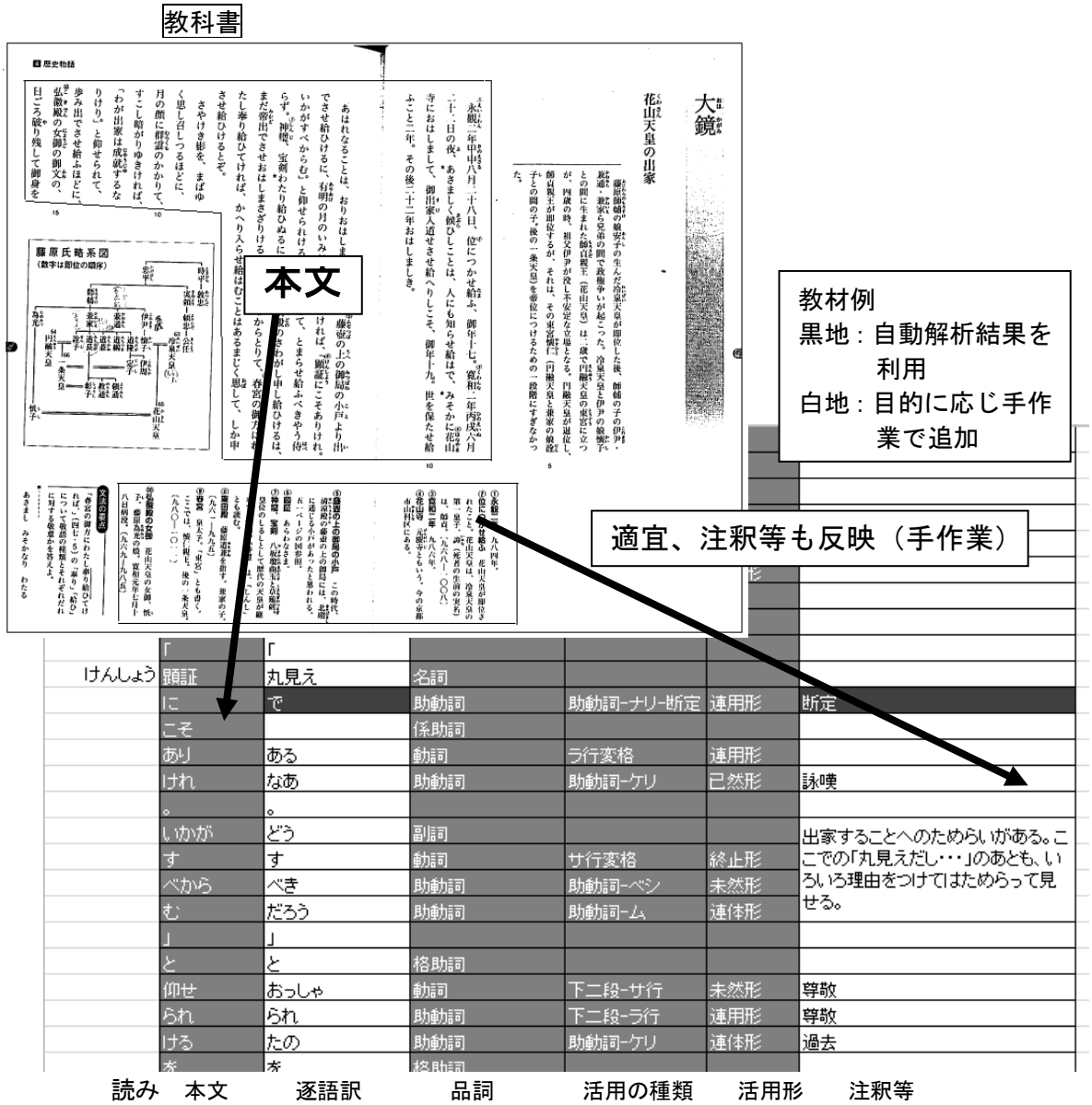
図 2　操作画面「和文茶まめ」での操作と、出力される Excel ファイル

４．形態素解析結果をベースにした学習教材の作成

４．１　教材作成の全体概要

以上の手順で出力した Excel ファイルをもとに、教材として不要な列を削除、必要な情報を追加することにより、目的に応じた教材を自由に作成することができる。

例えば一般的な教材作成としては「書字形」列（もとの本文）、「品詞」列、「活用型」列（学校教育での「活用の種類」）、「活用形」列を残し、新たに逐語訳の列や漢字の読みの列、適宜助動詞の意味や注釈を記入する列を追加する。これにより、いわゆる教科書ガイドの特色であった品詞情報と、語と語を 1 対 1 対応させた逐語訳を合わせることが可能となる（図 3）。さらに作業自体は一般的な Excel ファイルの編集作業であるため、セルの結合等を利用すれば、品詞分解を行った後でも、複数単語に渡るまとまりに注釈を与えることもでき、学習上、要注意箇所に色付けを行って目立たせることなども容易である。そして最大の特徴は、市販のもとと違い、資料の内容を自由に設定できることである。



読み	本文	逐語訳	品詞	活用の種類	活用形	注釈等
	「	「				
けんしょう	顕証	丸見え	名詞			
	に	で	助動詞	助動詞-ナリ-断定	連用形	断定
	こそ		係助詞			
	あり	ある	動詞	ラ行変格	連用形	
	けれ	なあ	助動詞	助動詞-ケリ	已然形	詠嘆
	。	。				
	いかが	どう	副詞			出家することへのためらいがある。ここでの「丸見えだし･･･」のあとも、いろいろ理由をつけてはためらって見せる。
	す	す	動詞	サ行変格	終止形	
	べから	べき	助動詞	助動詞-ベシ	未然形	
	む	だろう	助動詞	助動詞-ム	連体形	
	」	」				
	と	と	格助詞			
	仰せ	おっしゃ	動詞	下二段-サ行	未然形	尊敬
	られ	られ	助動詞	下二段-ラ行	連用形	尊敬
	ける	たの	助動詞	助動詞-ケリ	連体形	過去
	を	を	格助詞			

図 3　形態素解析結果を利用した教材のイメージ

４．２　作業手順例

形態素解析結果そのままの Excel ファイルから、学習用教材を作成する作業手順の一例を紹介する。活用の仕方と目的次第で、作成方法はいくらでもありうるはずであり、ここで示すのはあくまで一例である。

（１）不要な列の削除、必要な列の追加、列名の書き換え

図 3 に示したように、「読み／本文／訳／品詞／活用の種類／活用形／注釈等」という情報をもった教材を作成するとする。解析結果そのままのファイルでは図 4 のように「出典／文境界／書字形／発音形／語彙素読み／語彙素／品詞／活用型／活用形／語形／語種」となっているので、このうち「出典／文境界／発音形／語彙素読み／語形／語種」列（図 4 中、黒地にしてある列）を削除する（図 5）。

出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
花山天皇の	B	あはれ	アワレ	アワレ	哀れ	名詞-普通	名詞-形状詞可能		アワレ	和
花山天皇の	I	なる	ナル	ナリ	なり-断定	助動詞	文語助動詞	連体形-一	ナリ	和
花山天皇の	I	こと	コト	コト	事	名詞-普通	名詞-一般		コト	和
花山天皇の	I	は	ワ	ハ	は	助詞-係助詞			ハ	和
花山天皇の	I	、			、	補助記号-読点				記号
花山天皇の	I	おり	オリ	オリル	下りる	動詞-一般	文語上二段	連用形-一	オル	和
花山天皇の	I	おはしまし	オワシマシ	オワシマス	おわします	動詞-非自	文語四段-	連用形-一	オワシマス	和

図 4　解析結果そのままの状態

書字形	語彙素	品詞	活用型	活用形
あはれ	哀れ	名詞-普通	名詞-形状詞可能	
なる	なり-断定	助動詞	文語助動詞	連体形-一般
こと	事	名詞-普通	名詞-一般	
は	は	助詞-係助詞		
、	、	補助記号-読点		
おり	下りる	動詞-一般	文語上二段	連用形-一般
おはしまし	おわします	動詞-非自	文語四段-	連用形-一般

図 5　不要な列を削除

「書字形」列を「本文」、「活用型」を学校教育に合わせ「活用の種類」等と書き替える。また、「語彙素」列を「訳」と書き替え、逐語訳を記入する列として利用する。逐語訳は手作業で入力すべき列であるが、新規に空白の列を作って一から記入していくよりは、「語彙素」列の表示を利用して適宜書き替えたほうが作業が早い。「語彙素」とは本来辞書見出しの代表表記が表示される部分であるが、現代語形（の、活用がある場合には終止形）が表示される列であるため、名詞・助詞等はそのまま利用できる場合が多く、既存の列の中では最も逐語訳に流用しやすい。また、本文の横に漢字の読みの列を追加する（図 6）。

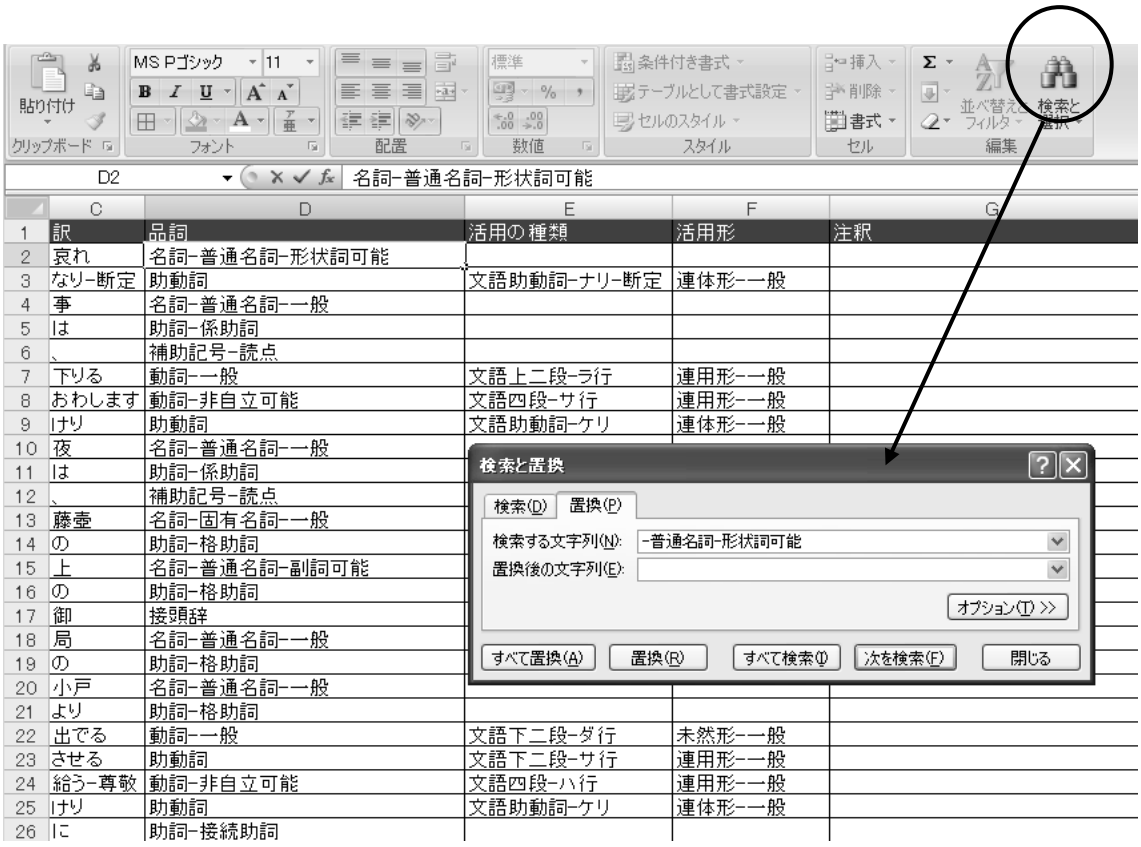
読み	本文	訳	品詞	活用の種類	活用形	注釈
	あはれ	哀れ	名詞-普通	名詞-形状詞可能		
	なる	なり-断定	助動詞	文語助動詞-ナ	連体形-一般	
	こと	事	名詞-普通	名詞-一般		
	は	は	助詞-係助詞			
	、	、	補助記号-読点			
	おり	下りる	動詞-一般	文語上二段-ラ行	連用形-一般	
	おはしまし	おわします	動詞-非自立	文語四段-サ行	連用形-一般	

図 6　列の調整完了

（２）品詞情報の書き換え

「中古和文 UniDic」での品詞情報は、例えば「名詞-普通名詞-形状詞可能」など、「（大分類）-（中分類）-（小分類）」という形で細かく出力される。これらの細かい分類は高校の古典では必要ないことが多いため、見やすさも考慮して適宜削除する。Excel の置換を

用いれば一括で変換可能である。例えば「名詞-普通名詞-形状詞可能」を「名詞」だけに書き換えたい（「-普通名詞-形状詞可能」を削りたい）なら、検索文字列を「-普通名詞-形状詞可能」、置換文字列を空欄にすればよい（図 7）。



	C	D	E	F	G
1	訳	品詞	活用の種類	活用形	注釈
2	哀れ	名詞-普通名詞-形状詞可能			
3	なり-断定	助動詞	文語助動詞-ナリ-断定	連体形-一般	
4	事	名詞-普通名詞-一般			
5	は	助詞-係助詞			
6	、	補助記号-読点			
7	下りる	動詞-一般	文語上二段-ラ行	連用形-一般	
8	おわします	動詞-非自立可能	文語四段-サ行	連用形-一般	
9	けり	助動詞	文語助動詞-ケリ	連体形-一般	
10	夜	名詞-普通名詞-一般			
11	は	助詞-係助詞			
12	、	補助記号-読点			
13	藤壺	名詞-固有名詞-一般			
14	の	助詞-格助詞			
15	上	名詞-普通名詞-副詞可能			
16	の	助詞-格助詞			
17	御	接頭辞			
18	局	名詞-普通名詞-一般			
19	の	助詞-格助詞			
20	小戸	名詞-普通名詞-一般			
21	より	助詞-格助詞			
22	出でる	動詞-一般	文語下二段-ダ行	未然形-一般	
23	させる	助動詞	文語下二段-サ行	連用形-一般	
24	給う-尊敬	動詞-非自立可能	文語四段-ハ行	連用形-一般	
25	けり	助動詞	文語助動詞-ケリ	連体形-一般	
26	に	助詞-接続助詞			

図 7　不要な品詞情報の削除

名詞・動詞等の中分類以下や、「補助記号-読点」「補助記号-読点」「補助記号-括弧開」「補助記号-括弧閉」などを削除する。また、助詞に関しては「助詞-格助詞」などと出力されるが、助詞の場合は「格助詞」等のレベルでの判別が望ましいため、「助詞-」の方を削除する。活用の種類につく「文語」、活用形につく「-一般」を削除する。以上のような工程を経て、図 7 の状態から、不要な品詞情報を削除した後の状態が図 8 である。

読み	本文	訳	品詞	活用の種類	活用形	注釈
	あはれ	哀れ	名詞			
	なる	なり-断定	助動詞	助動詞-ナリ-断定	連体形	
	こと	事	名詞			
	は	は	係助詞			
	、	、				
	おり	下りる	動詞	上二段-ラ行	連用形	
	おはしまし	おわします	動詞	四段-サ行	連用形	
	ける	けり	助動詞	助動詞-ケリ	連体形	
	夜	夜	名詞			
	は	は	係助詞			
	、	、				
	藤壺	藤壺	名詞			
	の	の	格助詞			
	上	上	名詞			
	の	の	格助詞			
	御	御	接頭辞			
	局	局	名詞			

図 8　不要な情報を削除した後の状態

以上は手作業と言っても、Excel の置換機能を使えば大した手間はかからない(4.3 参照)。あらかじめ削除すべき情報を知っている必要もなく、ファイルを上から眺めて、これは消そう、と思ったものを一括で消していく、という作業で十分対応できる。

また、「中古和文 UniDic」での品詞・活用形判定は、学校教育とやや異なるところがあるので、その点に関しては注意が必要である。学校教育に関わりそうな大きな違いは以下の２つである。１つは、学校教育での「形容動詞」は「形状詞」（語幹部分）＋断定の助動詞「なり」として出力されること、もう１つは、完了の助動詞「り」が下接する際の活用形が命令形として出力されること（学校教育では已然形扱いすることが多い）である。これらはいずれも手動で書き替えなければならないが、「形状詞」「命令形」を検索し、チェックして書き替えればよいだろう。ともに一般的なテキストであればたいした量は出ず、検索して逐一目視確認という作業は、作業量としては十分現実的である。

（３）エラー修正と、読み・訳の付与、注釈の記入

図 8 の状態にまでできれば、あとは「訳」の列を逐語訳に書き換えていき、適宜読みや注釈を自由に記入していけば、完成である。

本教材は「中古和文 UniDic」が自動で品詞分解し、品詞情報を与えた結果をもとに作成しているが、自動解析の結果はもちろん全て正しいとは限らない。機械がの自動判別には一定数の誤りもあるので、解析エラーがどこにあるかの確認と修正は必要である。ただし、「中古和文 UniDic」は平安仮名文学作品に対しては高い解析精度を実現しており、平安仮名文学作品であれば、教育上どうしても直さなければならないエラーはさほど多くは生じない。また、高校の古典では平安仮名文学作品以外も扱われるが、学校教科書での表記・仮名遣いであれば、平安期以外の資料でもある程度の精度が得られそうである (4.4 参照)。教材という性質上、エラーが無いかの確認は必須であるが、訳の付与をしながら同時に確認していけばたいていのエラーは見つけ出せる。（１）（２）に示した下準備を終えたのち、訳を付与しながら読み・注釈の適宜付与、エラーの確認と修正を同時進行で行う、というのが効率的であろう。

この他、「中古和文 UniDic」で品詞分解された言語単位（「短単位」と呼ばれる）は、学校教育の品詞分類と比べてやや短い、という点もないではない。例えば「花山寺」は「花山」／「寺」に分割して出力されるが、高校教育の感覚で品詞分解するなら「花山寺」で一語、というほうが一般的であろう。そこでこれらも修正し、もう少し長めの単位につないでいく、という工程もありうるが、このあたりの単位の切り方は高校教育の要点ではなく、解析結果のままにしておいても「間違い」というわけでもない。「これも直さなければ使えない」というのではなく、「余力があれば、直すとよりよくなる」という程度の問題だと位置づけられよう。この他、要注意箇所を色づけして目立たせたり、新たな列を追加して逐語訳ではない、最終的な「きれいな訳」を加えていく等、手の加え方は無限に拡張できるが、基本工程はここに記した（１）～（３）の作業だけで完了する（本稿末尾に作成した学習教材のサンプルを掲げる）。

以上、「中古和文 UniDic」を用いた教材作成の作業手順を紹介したが、このようにして作成が可能といっても、ここに示した作業を実際に行う場合には多大な時間がかかるというのであれば、この手法での教材作成は現実的とは言えまい。また、「中古和文 UniDic」の解析結果には、実際エラーがどのくらい出るのかも検証すべきである。エラーが余りに多く、逐一修正せねばならないようであれば、自動解析の意味がない。そこで最後に、作業時間および解析精度の検証を行った結果を簡単に紹介する。

４．３　作業コストの検証

上記作業に実際どのくらいの時間がかかるのかを検証した。サンプルは『大鏡』より教科書によく採用される「花山天皇の出家」。タイトル、リード文、系図等も含んだ教科書の版組で 4 ページ分、教科書に掲載される１話分の分量としてはよく見られる量である。解析結果による語数は 687 語であった。これを、本文を冒頭から打ち込んでテキストデー

タを作成、「中古和文 UniDic」で形態素解析後、4.1 に示した作業工程に従ってサンプルを作成した。工程（３）に関しては、注釈の付与は用途・目的・作業者の意図次第でいくらでも時間がかけられてしまうため、標準的な必須作業としてのエラー修正、読み・訳の付与までを行った。結果を表 1 に示す。

表 1　作業時間（『大鏡』より「花山天皇の退位」、687 語、教科書 4 ページ分）

テキストデータ作成	17分
形態素解析	1分
列の調整、品詞情報書き換え（工程（1）～（2））	5分
エラー確認・修正（工程（3））	16分
読み付与（工程（3））	8分
逐語訳付与（工程（3））	33分
総計	70分

テキストデータの作成（単純な打ち込み作業）に 17 分ほどかかっており、これを差し引いて、テキストデータさえあれば、逐語訳付与までの作業は 53 分で完了することが明らかになった。この後、必要に応じて注釈の記入等の作業を追加しうるし、もちろん作業に対する慣れ・不慣れによる差も、結果的な時間に反映されようが、教科書 4 ページ程度であれば、基本作業自体は 1 時間前後で完了するとみてよかろう。また、今回は検証のためエラー確認・修正と読み付与、逐語訳付与の 3 工程を別個に行ったが、実際には同時進行が可能な作業であり、作業に慣れたうえでこの工程を同時に行えば、さらなる時間短縮が見込める。

４．４　教材作成目的から見た「中古和文 UniDic」の解析精度

中古和文 UniDic は、平安仮名文学作品に対しては高い解析精度を実現しており、中古和文 UniDic　Ver0.5 の段階で、単位境界（品詞の切れ目が正しいか）で 99.3%、品詞認定で 97.8%という解析精度が報告されている（中古和文 UniDic ホームページほか)。ただしこれはあくまで平安仮名文学作品に対してのものであり、高校の古典学習においては、中世以降のテキストもさまざまに扱われるため、「中古和文」からやや離れたテキストに対しても実用に耐えるのかは、確認したいところである。そこで、サンプルデータとして作成した『大鏡』の外に、高校の教科書の定番であり、かつ、「中古和文」から外れるものとして『方丈記』『平家物語』を中古和文 UniDic を用いて解析し、解析精度を調査した。今回解析に使用したのは中古和文 UniDic　Ver1.3(2012 年 12 月公開)である。

中古和文 UniDic での解析結果には、語種や発音形など、教材作成には使用しない情報も出力されるが、今回の調査では教材作成に使用する情報のみに関し、精度調査を行った。調査項目は単位境界、語彙素、品詞、活用型（活用の種類)、活用形の５つである。これらの項目に対し、教材作成にあたって修正すべき箇所がいくつあるかを調査した。要修正箇所として数えられた「誤り」の中では、UniDic の品詞体系・規程上は誤りでないもの（「形状詞」という品詞認定や、完了の助動詞「り」の上の活用形「命令形」など）も含むため、今回の精度調査はあくまで高校の教材作成という用途から見た精度調査であり、中古和文 UniDic の、形態素解析辞書そのものの精度調査とは異なるものである点、注意されたい。

結果は表 2 のとおり。『大鏡』で精度 98~99%台であるのに対し、『方丈記』では 97~98%台、『平家物語』では 94~96%台と、「中古和文」から遠ざかるにつれ、精度がやや落ちるという予想通りの結果となった。『平家物語』あたりになると、擬態語や音便形に対応できない場合が増えてくるようである。ただそれでもなお、教材作成の素材として十分利用できる精度であることは間違いなく、『平家物語』の教材作成も実際に行ったが、中古和文と比

べるとやや手間がかかったという程度差で、作成可能であった。学習者が読みやすいよう、程よく漢字表記が用いられ、仮名遣い等も正規化されている学校教科書のテキストは、形態素解析というシステムにとって相性がよく、学校教科書というのは形態素解析辞書の活躍の場のひとつと考えてよさそうである。

表 2　各テキストにおける要修正箇所数と解析精度

			単位境界	語彙素	品詞	活用型	活用形
大鏡	総語数	要修正	10	11	10	10	3
「花山天皇の出家」	687	精度	98.5%	98.4%	98.5%	98.5%	99.6%
方丈記	総語数	要修正	14	14	13	12	17
「大火とつじ風」	1394	精度	97.7%	97.7%	97.9%	98.1%	97.2%
平家物語	総語数	要修正	50	72	58	48	46
「木曽の最期」	617	精度	96.4%	94.8%	95.8%	96.6%	96.7%

なお、表 2 の単位境界・語彙素・品詞等の要修正箇所の半数ほどは「形状詞」という品詞認定によるものであり、活用形の要修正箇所の 3 分の 1 ほどは完了の助動詞「り」の上の「命令形」である。これら UniDic と学校文法の差異による修正箇所は、規則的であり、検索・置換で簡単に修正できる。また、活用形の要修正箇所の残りの大部分は、文末の終止形・連体形の誤りである。要修正箇所の過半数は上記のタイプに尽くされるので、現実的な作業上は、①「形状詞」を検索してチェック、書き換え②「命令形」を検索し、下に助動詞「り」が続いていたら已然形に書き換え、③文末の終止形／連体形は要確認、という 3 点さえ注意すれば、他の誤りはそれこそ「見れば気付く」というような個別の誤解析ばかりであり、気付かないようなところで細かい品詞情報が誤っている、というようなケースはほとんどない。上記①②③だけ気を付け、あとは本文を流し読みしていくだけで、修正作業は完了させてよく、逐一個別に全ての項目に目を通さなくても問題ない精度である、というのが、実際に教材作成を試して得た感想である。

５．おわりに

以上、本稿では形態素解析辞書「中古和文 UniDic」の教育転用の在り方として、古典学習用の教材作成を紹介した。「中古和文 UniDic」は研究面でもさまざまな利用の可能性があるはずであり、今後ともその可能性を探っていきたい。

末尾に「中古和文 UniDic」を利用して作成した教材サンプルを掲げる。

文　献

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」『情報処理学会研究報告 人文科学とコンピュータ』Vol.2010-CH-85(No.4) pp.1-8

小木曽智信・小椋秀樹・近藤明日子・須永哲矢(2010)「形態素解析辞書「中古和文 UniDic」とその活用例」『日本語学会 2010 年度秋季大会予稿集』 pp.243-248

小椋秀樹・須永哲矢(2012)『中古和文 UniDic 短単位規程集』平成 21（2009）- 平成 23（2011）年度科学研究費補助金基礎研究(C)「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2（課題番号 21520492、代表者　小木曽智信）

関連 URL

日本語歴史コーパス「中納言」 http://maro.ninjal.ac.jp/

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

資料：教材サンプル　『大鏡』より「花山天皇の出家」（一部）

読み	本文	訳	品詞	活用の種類	活用形	注釈
	あさましく	驚きあきれ	形容詞	形容詞-シク	連用形	
	候ひ	まし	動詞	四段-ハ行	連用形	丁寧→読者
	し	た	助動詞	助動詞-キ	連体形	
	こと	事	名詞			
	は	は	係助詞			
	、	、				
	人	人	名詞			
	に	に	格助詞			
	も	も	係助詞			
	知ら	知ら	動詞	四段-ラ行	未然形	
	せ	せ	助動詞	下二段-サ行	連用形	使役
	給は	なさら	動詞	四段-ハ行	未然形	尊敬→花山
	で	ないで	接続助詞			
	、	、				
	みそかに	密かに	形容動詞	ナリ活用	連用形	
	花山寺	花山寺	固有名詞			
	に	に	格助詞			
	御座しまし	いらっしゃっ	動詞	四段-サ行	連用形	尊敬→花山
	て	て	接続助詞			
	、	、				
	御	御	接頭辞			
すけ	出家	出家	名詞			
	入道	入道	名詞			仏門に入ること
	せ	なさっ	動詞	サ行変格	未然形	
	させ		助動詞	下二段-サ行	連用形	尊敬→花山
	給へ		動詞	四段-ハ行	已然形	尊敬→花山
	り	ってしまっ	助動詞	助動詞-リ	連用形	
	し	たの	助動詞	助動詞-キ	連体形	
	こそ	こそ（驚きだ）	係助詞			結びは省略。「あさましく候ひし」など。
	。	。（当時）				
	御	御	接頭辞			
	年	年（は）	名詞			
	十九	十九（歳）	数詞			

闘病記からわからないこと、わかるかもしれないこと

——闘病記コーパスを用いた患者像の一致についての考察——

臼田 泰如（京都大学大学院 人間・環境学研究科）
島本 裕美子（京都大学 学際融合教育研究推進センター）
久保 圭（大阪大学 日本語日本文化教育センター，京都大学 学際融合教育研究推進センター）
荒牧 英治（科学技術振興機構さきがけ，京都大学 学際融合教育研究推進センター）

What We Can Know? What We May Know?

Agreement of Patient Status Estimation using Illness Narrative Corpus

Yasuyuki Usuda (Grad. School of Human and Environmental Studies, Kyoto Univ.)
Yumiko Shimamoto (Center for the Promotion of Interdisciplinary Education and Research, Kyoto Univ.)
Kay Kubo (Center for Japanese Language and Culture, Osaka Univ., Center for the Promotion of Interdisciplinary Education and Research, Kyoto Univ.)
Eiji Aramaki (JST-PRESTO, Center for the Promotion of Interdisciplinary Education and Research, Kyoto Univ.)

1　はじめに

自伝を読むと著者について分かったような気になる．同時に，著者も自らの著作を通じて，何らかのメッセージを伝えようとしていると思われる．この時，読者がえた情報はどの程度一致するのであろうか？必ずしも読者の間で理解が一致するとは限らない。こうした理解の一致／不一致に関与する要素を明らかにするのが本研究のねらいである。

しかし、テキストに書かれうる事柄、背景や著者の属性、ジャンルなど、テキストの内容や周辺状況は多様であり、内容の理解に関する検討は非常に複雑になることが予想される。適切に理解の一致度を測定するためには、著者（登場人物）の属性や内容などに一定の統制を設けられるよう、適切なテキストの集合を用意することが望ましいのではないか。

このために筆者らは闘病記というジャンルのテキストに着目する。闘病記テキストの著者はおおむね患者本人ないしはその家族など当事者であり、書かれる内容はその当事者が患者の疾患の状態や変化、あるいは疾患を取り巻く治療や環境などの状態について記述したものである。記述の重要な焦点として、患者の生活の質 (Quality of Life: QOL) も想定される。

闘病記テキストは多数の公刊された書籍の他、ブログなどインターネット上の文書類も豊富に存在する。加えて、患者の QOL を測定するための整備された尺度である SF-36v2 日本語版 (福原 and 鈴鴨, 2011) が存在する。こうしたことから筆者らは、闘病記テキストを素材として SF-36 を用い、5 人の読者がそれぞれテキストから読み取った情報がどの程度一致するかについて検討した。

2　関連研究

2.1　テキストの理解に関する研究

人がどの程度、どのようにテキストを理解するのかということについては、従来文章理解研究の範疇での研究が蓄積されてきた (大村 et al., 2001; 甲田, 2009; 針生, 2006; 秋田, 2002)。しかしながら、本研究の目的である「読者の間でどの程度内容の一致を確認できるか」という問いに答える研究は、管見の限りなされていない。

従来の文章理解研究においては、人の認知情報処理の過程を解明するという観点から、語の意味やテキストの結束性などといったテキスト側の特性と、推論などの認知能力との関わりが論じられてきた。大村 et al. (2001) は、『「何がわかったらその文がわかったことになるのか」という基準がつくりにくいことは明らか』とした上で、従来の研究を概観している。本研究の課題である「複数人で理解の一致がどの程度見られるか」に関連するものとしては、授業場面での共同問題解決のためにテキストを複数人で読む場合が類似していると考えられるが、「一致するかどうか」についての言及はみられない。

また、甲田 (2009) では、テキストの理解を促進する要因として、先行オーガナイザー（タイトルなど、あらかじめ簡略に内容を予告するものや、全体像を図式化したもの)、図の利用、読み手の知識、テキストの結束性の操作などを挙げている。特に、ややわかりにくいテキストのほうがよく理解されるとして、中程度の結束性をもつテキストにおいて推論によって補完することが理解を促進するという側面を強調している。

これらの先行研究からは、文章理解に関する研究が主にテキストの構造に関する要因（結束性）や提示方法（オーガナイザーや図)、認知的要因（知識や推論能力）に焦点を当ててきたことが示唆される。テキストの内容に関わる要因としては、読み手の知識状態（書かれている内容について知っているかどうか）への言及が見られる程度と言える。しかしながら、書かれている内容が理解に影響を及ぼす可能性もあると思われれる．本研究の試みは、内容を闘病記に限定することで，テキストに書かれている内容がとその理解の関係を調査することである．

2.2　闘病記の利用に関する研究

闘病記をとりまく社会的状況として、近年、患者が自らの医療情報を記述する試みが注目を集めている。これらの注目に伴い，Patient Reported Outcome（PRO）や患者主導レジストリという概念の提唱，闘病中の人々の日記を共有する患者 SNS（Social Networking Service）など多くの試みがなされつつある．中でも，同じ病気を抱える人、あるいはそうした状況に関心をもつ人々の間で共有することを目的とした米国の PatientsLikeMe (`http://www.patientslikeme.com/`) は大きな注目を集めている．闘病記執筆および，その利活用もこの動きのひとつといえる。闘病記をデータとしてみた場合、治療の状況などのほか、通常の医療行為を通しては知ることが難しい患者と医療者との関係、病状に伴う QOL の変化といった情報を得る手がかりとして有望である。反面、闘病記は主観的な記述が多く、確実な情報が得にくいという欠点がある。

また、闘病記の教育的応用について検討した研究として、看護学生に対する患者理解の促進の効果について検討した 太田 and 古米 (1994)、門林 et al. (2006)、岡本 et al. (2011)、岡本 and 長谷川 (2006)、現職の医療従事者が闘病記を読むことで得られた気づきについて検討した阿部 (2012) などがあり、多くは医療を学ぶ学生や医療専門職に従事する人々が闘病記を読むことで患者に対する理解を深めることができると述べている。

これらはいずれも、闘病記から読み取れる患者に関する情報を通じて、患者について理解を深めることを目的としている。この前提として、闘病記テキストは読み取りうる情報について一定の客観性が認められることがあると考えられる。しかし、その客観性はどの程度なのか。本研究ではこうした点を考察する。

3　コーパスについて

本研究で用いるコーパスは、日本語で書かれ日本国内で出版された闘病記の中から、一冊あたり 10 ページのサンプルを 2ヶ所、計 20 ページずつ無作為に取り出したものである。使用した書籍の一覧を以下に挙げる。

表 1: 使用した書籍一覧

作者	執筆者	病名	タイトル
安積政子・	闘病者母・		
藤保秀樹	BF（書簡）	白血病	〈新装版〉 永遠の愛を誓って
小西豊海	闘病者の母	急性リンパ性白血病	生きぬいて!愛娘 一 母は負けないよ
小林茂登子	闘病者本人	慢性骨髄性白血病	あたりまえの日に帰りたい 一一骨髄性白血病からの生還
大谷賀子	闘病者本人	慢性骨髄性白血病	生きてるってシアワセ!
那須田稔・			
岸川悦子	闘病者ではない著者	慢性骨髄性白血病	母さん子守歌うたって 寸越窯 ・いのちの記録
平美樹	闘病者本人	急性骨髄性白血病	病院を出ょうーネコの脱出奮闘記
相河ラズ	闘病者本人	上咽頭がん	余命半年から生きてます! 面白いほど不運な男の笑う闘病記
野実香	闘病者本人	乳癌	癌と闘わない–私の選択: 私の人生、私が選んではいけませんか?
川竹文夫	闘病者本人＋ 12	腎臓がん 甲状腺がん・	幸せはガンがくれた一心が治した 12 人の記録
柏繁男 団鬼六・	闘病者本人	前立腺がん	二度ガンが消えた　わが人生の記録
黒岩由起子	闘病者の娘	食道がん	手術は、しません一父と娘の「ガン闘病」450 日
星野史雄	闘病者本人	大腸がん	闘病記専門書店の店主が、がんになって考えたこと
横山友美佳	闘病者本人	肺がん	明日もまた生きていこう 十八歳でがん宣告を受けた私
梅本典敏・良子	闘病者の両親	悪性脳腫瘍グリオーマ	希望 ライト　麻依子の闘病記 (悪性脳腫瘍】
森下純子	闘病者の母	横紋筋肉腫（小児がん）	ママでなくてよかったよ一小児がんで逝った 8 歳 498 日間の闘い
山崎敏子	闘病者の母	ユーイング肉腫	がんばれば、幸せになれるよ一小児がんと闘った 9 歳の息子が遺した言葉
米原万里	闘病者本人	卵巣癌	打ちのめされるようなすごい本
脇英策・佳子	闘病者の妻	非ホジキンリンパ種	夕焼けの日曜日 夫婦で戦った 284 日のがん闘病記
笹塚始	闘病者本人	前立腺ガン	前立腺ガン闘病記一68 歳早期発見
安武信吾・			
千恵・はな	闘病者本人	乳がん→肺がん T細胞型急性	はなちゃんのみそ汁
穂積良洋	闘病者本人	リンパ性白血病	ワイルズの闘病記
鈴木はとこ	闘病者本人	子宮頸がん	がんよがんよとんでいけ一子宮頸がん闘病記
小実健一	闘病者本人	咽頭癌	「咽頭癌治療」と「その副作用」との闘病記
小松武幸	闘病者の夫	乳がん	ママが生きた証
片倉もとこ	闘病者本人	がん	旅だちの記
中島梓	闘病者本人	すい臓がん→肝臓がん	転移
橋照雄・真知子	闘病者の父母	急性骨髄性白血病	悠君ごめんね一父と母の白血病闘病記
藤谷ペコ	闘病者本人	乳がん	若年性乳がんになっちゃった!一ペコの闘病日記
曾山尚幸	闘病者本人	悪性リンパ腫	無差別料理研究家ｓｏｙａｍａｘの闘病記　悪性リンパ腫　１９歳の骨髄移植
尾形千秋	闘病者の娘	成人 T 細胞白血病	成人 T 細胞白血病 ATL 闘病記 (ATL シリーズ 2)
峯直法	闘病者本人	悪性リンパ腫	悪性リンパ腫なんて怖くない 直君の悪性リンパ腫格闘記
加納朋子	闘病者本人	急性骨髄性白血病	無菌病棟より愛をこめて
但馬裕子	闘病者本人	外陰がん→癌性腹膜炎	ひかりのなかへ一700 日の DIARY
横山文野・			
山口智久	闘病者本人・夫	C349	その夜、妻に最期のキスをした。
垣添忠生	闘病者の夫	肺の小細胞がん	妻を看取る日 国立がんセンター名誉総長の喪失と再生の記録
神田麻希子	闘病者本人	仙骨ユーイング肉腫	神様、私をもっと生きさせて!一小児がんと闘った青春期 4 年間の命の詩
横幕真紀	闘病者の母	急性骨髄性白血病	ずっとそばにいるよ一天使になった航平
筒井好美	闘病者本人	下咽頭の扁平上皮がん	ママの声、聞こえるよ
光武綾	闘病者の母	横紋筋肉腫	愛してるよカズ一小児ガンと闘った母親と息子の愛の記録
７２３	闘病者本人	子宮頸がん	HPV　２４歳、子宮頸がん患者デビュー
高野由美子	闘病者の母	急性リンパ性白血病	俺、マジダメかもしれない…一「急性リンパ性白血病」で逝った最愛の息子へ
佐藤由美	闘病者本人	腺様のう胞がん	余命ゼロを生きる 現役美容師、奇跡の物語
渡美佳子	闘病者の母	胚細胞腫瘍	ビビるなよ！　２５歳で逝った息子が私に放った
本田麻由美	闘病者本人	乳がん	「34 歳でがんはないよね？」一あるジャーナリストの揺れる心の軌跡
キャシー中島	闘病者の母	小細胞がん	6%の明日一もう一度あなたのお母さんになりたい
武本和枝	闘病者の妻	肝臓ガン	余命 1ヵ月! 「ガン宣告」された夫と妻の感動の記録。
橘秀夫	闘病者本人	肺がん	拝啓 肺ガンになっちゃいました
田村まり子	闘病者本人	乳がん	わたしは風にのって一ボレボレ乳がん闘病記
渥美雅子	闘病者本人	子宮がん	子宮癌のおかげです一女弁護士の全摘 57 日間の記録
仁科亜季子	闘病者本人	子宮頸がん	「子宮頸がん」一経験したからこそ伝えたい!
岸本葉子・			
中富庸介	闘病者本人・医師	虫垂がん	がんと心
中島梓	闘病者本人	下部胆管癌	ガン病棟のピーターラビット
小野允雄	闘病記本人	大腸がん	余命半年からの生還 大腸がん闘病の記録
尾関良二	闘病者本人	スキルス胃がん	僕は、慢性末期がん一「余命半年」の宣告どおりに死にたくない人におくる本
山下郁子	闘病記本人	スキルス胃がん	絶体絶命を生き抜く 一スキルス胃がん余命三カ月から三年半

以下にコーパスの抜粋を挙げる。

> 一回の化学療法で、こんなに大きな結果が出るとは思っていなかった。発汗・微熱・鼻腔のデキモノが主だった症状だったが、どれもが恢復していた。しかし、改善を確かなものとして、視覚的に確認するにはＣＴによる画像診断と血液検査、触診などしかない。それがこの病気の恐ろしいところなのだろう。つまり、胃がんや肺がんなど固形のがんは手術によって、悪性物を摘出し、転移の有無によって化学療法や放射線治療を施す。しかし、リンパ腫は全身の血液を介して悪性細胞がめぐっているのだから、どこかを外科的手術で摘出するわけにはいかないのだ。前回、施された鼻腔腫瘍摘出術もがん細胞の摘出ではなく検査が目的である。血液がんの治療法としては、全身に散らばった悪性細胞を抗がん剤の点滴によって死滅させるしかない。それが完全になされているかどうかを確実に知る方法はないようで、良い状態を維持している事で、治療に成功しているとするのだ。抗がん剤を持続して実施し続ける事はできない。なぜなら抗がん剤は良い状態の細胞も見境なしに攻撃してしまい、体力はその度に奪われていくからだ。
>
> （脇英策・佳子『夕焼けの日曜日 夫婦で戦った 284 日のがん闘病記』幻冬舎ルネッサンス）

> この抗がん剤を投与する期間中は、規則正しい生活を強いられます。お昼ごはんを食べた後は絶食、１５時３０分に“吐き気止め”を飲み、１６時にテモダールを飲みます。その後２時間は絶飲食なので、１８時まで何も食べたり飲んだりすることができません。このテモダールという抗がん剤ですが、副作用の少なさが売りのようです。少し娘の病気について説明しますと、病名は“悪性脳腫瘍”ですが、脳の細胞が腫瘍となった脳実質内発生腫瘍で“悪性神経膠腫　グリオーマ”と呼ばれるものです。レベルはⅣでした。腫瘍は脳の中心部にあり、最大時は約３センチ程度ありました。テモダールは、このような病気の治療用として開発された薬です。作用としては、薬の有効成分が体の中の血管を通って脳の組織まで到達し、腫瘍細胞がむやみに増殖するのを阻止し、さらには殺すことができる可能性があるそうです。副作用としては、嘔吐、食欲不振、便秘、疲労（倦怠感）、脱毛などがあるそうです。１６時、恐る恐るテモダールをすべて飲ませ、その後娘は、少し眠りました。１８時頃、家内と彼氏と娘で、自動車サービス店にオイル交換に行きました。薬の影響はあまりなく、娘も元気そうです。
>
> （梅本典敏・良子『希望 ライト　麻依子の闘病記 (悪性脳腫瘍)』文芸社）

これらのサンプルには、1 節で述べたように患者の健康状態、病気に対する捉え方や治療の状況、あるいは周囲の人々との関わりといった内容が綴られていることがわかる。これらのうち、どのような内容は一致しやすく、どのような内容が一致しにくいのか、ということを次節において検討する。

4　方法

本研究では認知システムに重点を置いた従来の観点に対し、テキストの内容的側面に着目することを試みる。具体的な方法として、包括的 QOL 尺度である SF-36v2 日本語版を用い、闘病記から読み取れる QOL 情報を得点化し、その得点が評定者の間でどの程度一致するかについて主に分析する。また回答の欠損についても考慮し、無回答率と得点との関連も分析する。

SF36v2 とは、認定 NPO 法人　健康医療評価研究機構 (`http://www.i-hope.jp/activities/qol/list/sf-36.html`) によって開発された、健康関連 QOL（HRQOL: Health Related Quality of Life）を測定するための尺度である。QOL とは、生命の質、生活の質などと訳

され、「身体機能」：階段を上れるか、服を着られるかなど、「心の健康」：メンタルヘルス、気分の落ち込み、不安など、「社会生活機能」：友人関係や付き合いの変化など、「日常役割機能」：仕事や家事の変化など、といったことへの主観的評価をもとに人がどの程度健康で充実した生活を送っているかを判断するための指標である (佐野, 2001)。さらに、痛み、活力、睡眠、食事、性生活なども重要な要素として含まれることがある。

SF36v2 は 36 項目の質問からなる質問紙であり、あらゆる疾患の患者の健康関連 QOL を計測することができる包括的尺度として作成されている。このため異なる疾患の患者や健康な人と患者との間で健康関連 QOL を比較することが可能となっている。質問項目の例を以下に挙げる。

> 質問 1。あなたの健康状態はいかがですか？　一番よく当てはまる番号を選んで下さい。
> 「最高に良い」場合は 1、
> 「とても良い」場合は 2、
> 「良い」場合は 3、
> 「あまり良くない」場合は 4、
> 「良くない」場合は 5。
>
> 質問 3 は 10 問あります。次の質問は、日常よく行われている活動です。あなたは健康上の理由で、次のような活動をすることがむずかしいと感じますか。むずかしいとすればどのくらいですか。それぞれの質問について一番よくあてはまる番号を選んで下さい。
>
> 質問 3 の 1、激しい活動、例えば一生けんめい走る、重いものを持ち上げる、激しいスポーツをする、など。
> 「とてもむずかしい」場合は 1、
> 「すこしむずかしい」場合は 2、
> 「ぜんぜんむずかしくない」場合は 3。
>
> 質問 6。過去 1ヶ月間に、家族、友人、近所の人、その他の仲間とのふだんのつきあいが、 身体的理由あるいは心理的な理由で、どのくらいさまたげられましたか。一番よくあてはまる番号を選んで下さい。
> 「ぜんぜんさまたげられなかった」場合は 1、
> 「わずかにさまたげられた」場合は 2、
> 「すこし、さまたげられた」場合は 3、
> 「かなり、さまたげられた」場合は 4、
> 「非常に、さまたげられた」場合は 5。
>
> （項目番号は原質問紙の通り）

36 項目の質問の得点は、それぞれ重みづけを行った上で健康関連 QOL の 9 つの下位尺度に合算される。9 つの下位尺度は表 2 の通りである (福原 and 鈴鴨, 2011)。

5 人の被験者（t1, t2, ..., t5）に 3 節のコーパスを読むよう依頼し、それぞれのサンプルごとに、患者の健康状態について SF-36v2 に回答することを求めた。臨床心理学を専攻とする大学院修了生，並びに博士後期課程・前期課程の５名（男子３名，女子２名：平均年齢 25.6 歳，SD1.82）に調査を依頼した．調査期間は 2013 年 8 月 9 日〜9 月 19 日であり，有効回答

表 2: SF36v2 日本語版の下位尺度

下位尺度	得点の解釈	
	低い	高い
身体機能 (Physical functioning: PF)	健康上の理由で、入浴または着替えなどの活動を自力で行うことが、とてもむずかしい	激しい活動を含むあらゆるタイプの活動を行うことが可能である
日常役割機能（身体） (Role physical: RP)	過去 1ヵ月間に仕事やふだんの活動をした時に身体的な理由で問題があった	過去 1ヵ月間に仕事やふだんの活動をした時に、身体的な理由で問題がなかった
体の痛み (Bodily pain: BP)	過去 1ヵ月間に非常に激しい体の痛みのためにいつもの仕事が非常にさまたげられた	過去 1ヵ月間に体の痛みはぜんぜんなく、体の痛みのためにいつもの仕事がさまたげられることはぜんぜんなかった
社会生活機能 (Social functioning: SF)	過去 1ヵ月間に家族、友人、近所の人、その他の仲間とのふだんのつきあいが、身体的あるいは心理的な理由で非常にさまたげられた	過去 1ヵ月間に家族、友人、近所の人、その他の仲間とのふだんのつきあいが、身体的あるいは心理的な理由でさまたげられることはぜんぜんなかった
全体的健康感 (General health perceptions: GH)	健康状態が良くなく、徐々に悪くなっていく	健康状態は非常に良い
活力 (Vitality: VT)	過去 1ヵ月間、いつでも疲れを感じ、疲れはてていた	過去 1ヵ月間、いつでも活力にあふれていた
日常役割機能（精神） (Role emotional: RE)	過去 1ヵ月間、仕事やふだんの活動をした時に心理的な理由で問題があった	過去 1ヵ月間、仕事やふだんの活動をした時に心理的な理由で問題がなかった
心の健康 (Mental health: MH)	過去 1ヵ月間、いつも神経質で憂鬱な気分であった	過去 1ヵ月間、おちついていて、楽しく、おだやかな気分であった

率は 100 %であった．得られた SF-36v2 の回答について、『健康関連 QOL 尺度　SF-36v2 日本語版マニュアル』(福原 and 鈴鴨, 2011) に基づき、被験者ごとに 9 つの下位尺度ごとに得点の重み和を算出した。算出された得点の重み和について、被験者ごとにケンドールの順位相関係数を算出し、平均を求めて比較した。本研究ではこの順位相関係数の値を一致度と呼ぶこととする。なお欠損値については福原 and 鈴鴨 (2011) に基づき、各下位尺度の平均点を充当して処理を行った。

ついで、各下位尺度ごとに欠損値の割合を算出し、順位相関係数との比較を行った。

5　結果

5.1　結果の概観

まず、被験者間の一致度（被験者ごとの評定の相関係数)、および下位項目ごとの無回答率を調べた。被験者間の一致度は、5 人の被験者から 2 人ずつのすべての組み合わせについてケンドールの順位相関係数を算出し、その値の平均を求めたものである。また [ ] 内は下位項目ごとの、それぞれの被験者のペアの相関係数の間の標準偏差である。算出には R ver.2.13.1 を使用し、小数点以下第 3 位で四捨五入した。

結果を概観すると、一致度に関しては、全体的健康感および身体機能については比較的高

い値が得られ、体の痛みおよび日常役割機能（精神）の値が特に低いことがわかる。また無回答率については、体の痛み、日常役割機能（精神）が高く、全体的健康感が低いことが見て取れる。

これらの結果からは、一致度の大小と無回答率に関連があることが示唆される。すなわち、回答者が回答可能であると判断した（＝無回答にしなかった）場合、一致度も高くなる傾向があると考えられる。

表 3: 無回答率 (p) と一致度の平均 (a)

下位項目	無回答率 (p)	一致度の平均 (a) [SD]
身体機能	0.11	0.469 [0.055]
日常役割機能（身体）	0.12	0.429 [0.084]
体の痛み	0.26	0.163 [0.144]
全体的健康感	0.04	0.487 [0.063]
活力	0.21	0.357 [0.089]
社会生活機能	0.15	0.344 [0.077]
日常役割機能（精神）	0.25	0.209 [0.095]
心の健康	0.19	0.375 [0.045]
健康推移	0.17	0.342 [0.151]

5.2　無回答率と一致度の関連

図 1 は、無回答の割合 p を横軸に、一致度の平均 a を縦軸にプロットしたものである。この図からは、無回答の割合 p と一致度の平均 a は弧状の関係をもち、おおむね負の比例を示していることがわかる。これは一致度の値の大小が無回答の割合から強い影響を受けている可能性を示唆する。

この中で、社会生活機能については無回答の割合が低く、かつ一致度が比較的低いという特異的な特徴を示していることが読み取れる。すなわち、社会生活機能に関わる設問項目について、各被験者はそれぞれ回答可能であると判断したにもかかわらず、被験者ごとに回答のばらつきが大きいということを意味する。

6　考察

本研究の結果より、闘病記を第三者が読むことによって著者ないしは患者についての情報を得る上で、読者の間で一致しやすい情報と一致しにくい情報があることが示唆される。

5.2 節で述べたように、「社会生活機能」は無回答率が低く、かつ一致度が低いという特徴を示している。図 1 からわかるように、多くの下位尺度においては無回答率と一致度の平均は負の相関を示し、無回答が多ければ多いほど被験者間の一致度が下がる傾向がある。このことは、被験者が判断可能であると考える場合には何らかの根拠がテキスト中に示されており、それを各被験者が利用して回答しているためであると考えられる。すなわち、「書いてある」ことについては「わかる」かつ「一致する」といえる。その中で、社会生活機能についてはそうした傾向が比較的弱く、被験者は回答可能であると判断したにもかかわらず一致

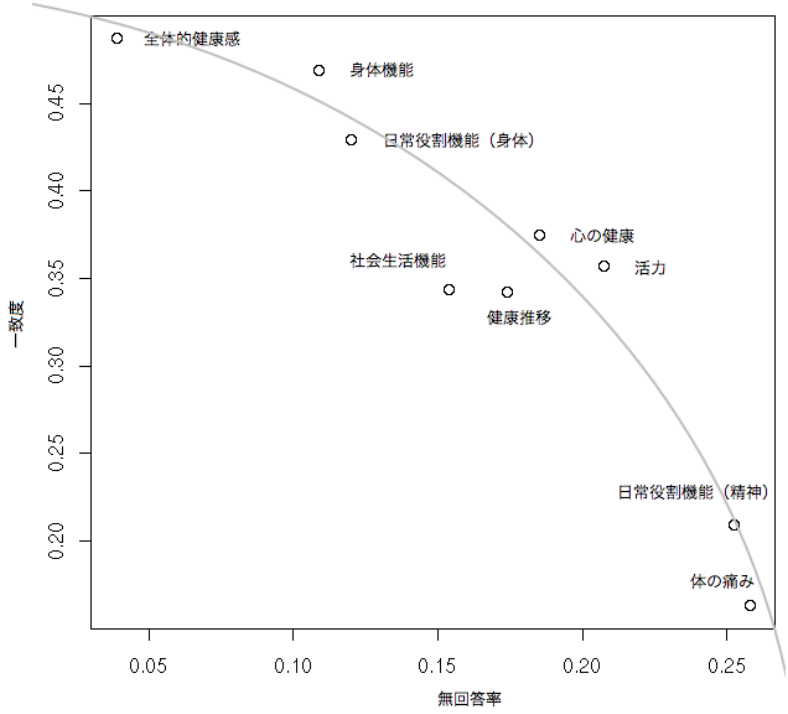


図 1: 無回答率 (p) と一致度の平均 (a) の関係

度が低い。すなわち、社会生活機能については「書いてある」から「わかる」にもかかわらず「一致しない」のである。

以上を踏まえて、闘病記読者による闘病記に書かれた人の QOL の判定に影響する要素について考察してみたい。

図 2 は、闘病記の著者が執筆する中で言及する要素の模式である。「社会生活機能」以外の尺度に関わる項目の場合、周囲の人との関わりは評定に大きな影響を持たず、もっぱら当事者の身体感覚や病状認識に関わる記述によって評定されていると考えられる。一方、「社会生活機能」に関しては、当事者の身体感覚や病状認識よりも、当事者と周囲の親しい人、あるいは会社や病院といった社会関係が大きな寄与をなすと考えられる。周囲の人との関わりは必然的に双方向的なものとなり、当事者が周囲に対して働きかける方向性と、周囲から当事者に対する働きかけの双方が影響するといえる。このため、評定者が判断の根拠とすることのできる情報も複線化すると考えられ、このことが評定者にとって「わかる」が「一致しない」という結果となったことにつながっていると考えられる。

7　まとめ

本研究では、闘病記テキストを素材とし、包括的 QOL 尺度 SF-36 を使用して執筆者の QOL を評定したデータを分析することを通じて、テキストに書かれたデータを読者がどの程度客観的に把握できるかについて検討した。分析の結果、当事者と周囲の人々や当事者を取り巻く社会との関わりに関係する指標である「社会生活機能」が、欠損値が少ないにも関わらず一致度が低いという、他の指標に比べて特異な挙動を示すことがわかった。このこと

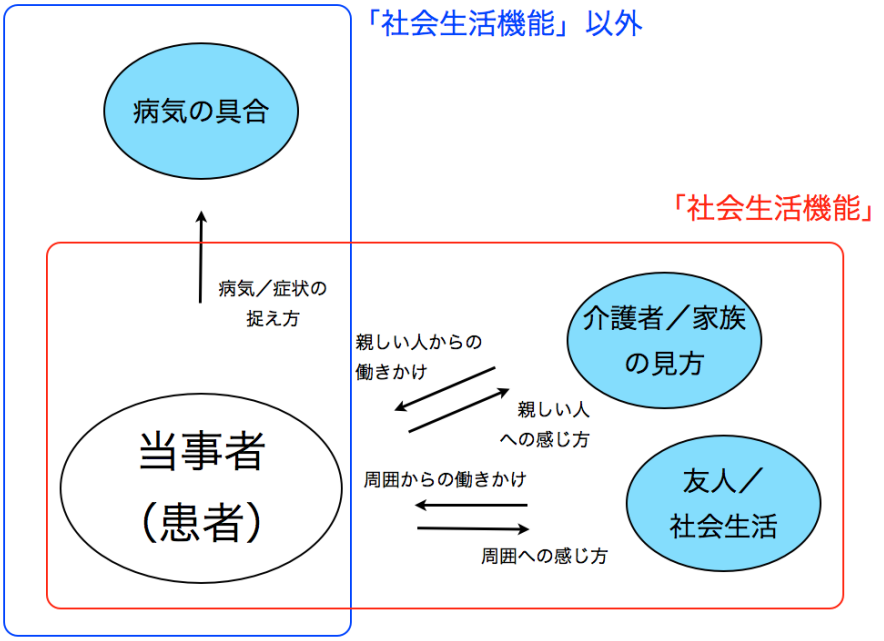


図 2: QOL 解釈のイメージ

について著者らは、「社会生活機能」には他の指標に比べて記述に関係する要素が複線化されており、そのことが評定者の判定の特異な挙動に影響しているとの考察を行った。

著者はほとんど記述されないために伝わらないと思われる「体の痛み」や「精神的な面で十分に活動できているか否か」といった内容の記述を増やすことや、記述が複雑化する「周囲の人との関係はどうか」という内容の明瞭度を上げることが有効である可能性がある。

本研究の次の段階として、株式会社メディエイドの協力を得て、同社が運営する闘病記 SNS「ライフパレット」(https://lifepalette.jp/) において闘病記を執筆するユーザーに対して SF-36 への回答を依頼し、闘病記の著者の QOL についてのデータを取得することを予定している。取得されたデータを利用して、読者の QOL 理解がどの程度著者の QOL と一致するのかを考察する。今回と同様に読者による患者の QOL 判定を行った後、その判定が著者自身による QOL 評定とどの程度一致するのかを検討することで、読者間の一致／不一致による理解のしやすさの程度だけでなく、著者の意図した内容がどの程度理解しやすいのか、理解のしやすい／しにくい情報とはどのようなものか、についての検討が可能になると考えられる。

参考文献

秋田 喜代美 (2002) 『読む心・書く心　文章の心理学入門』北大路書房.

阿部 泰之 (2012) 「医療者の闘病記読書に関する質的研究：何が得られるのか」『医学図書館』 59, pp.176–179. URL http://ci.nii.ac.jp/naid/120005196648/.

太田 にわ, 古米 照恵 (1994) 「障害者の闘病記の読書を通じての看護学生の患者理解」『岡山大学医療技術短期大学部紀要』 4, pp.89–97. URL http://ci.nii.ac.jp/naid/120002313994/.

大村 彰道, 秋田 喜代美, 久野 雅樹 (2001) 『文章理解の心理学—認知、発達、教育の広がりの中で』北大路書房.

岡本 佐智子, 長谷川 真美 (2006) 「闘病記を教材に用いた回復期看護演習のグループワークに関する検討」『埼玉県立大学紀要』 8, pp.119–124. URL http://ci.nii.ac.jp/naid/110006250563/.

岡本 寿子, 高橋 康子, 江頭 典江, 豊田 久美子 (2011) 「基礎看護技術演習に闘病記を用いる教育効果」『京都市立看護短期大学紀要』 36, pp.77–85. URL http://ci.nii.ac.jp/naid/110008609624/.

門林 道子, 真部 昌子, 小濱 優子 (2006) 「看護学生が闘病記を読む意味について：成人看護論での闘病記を用いた授業、5 年間の報告」『川崎市立看護短期大学紀要』 11, pp.13–18. URL http://ci.nii.ac.jp/naid/110004868274/.

甲田 直美 (2009) 『文章を理解するとは—認知の仕組みから読解教育への応用まで』スリーエーネットワーク.

佐野 文男 (2001) 「QOL 指標」『北海道医報』 pp.2–5 URL http://www.hokkaido.med.or.jp/iho_index.html.

針生 悦子 (2006) 『言語心理学（朝倉心理学講座）』朝倉書店.

福原 俊一, 鈴鴨 よしみ (2011) 『健康関連 QOL 尺度 SF-36v2 日本語版マニュアル 第 3 版』認定 NPO 法人 健康医療評価研究機構.

動詞「教える」と共起する授受表現について

朱 炫姝 (筑波大学大学院 人文社会科学研究科)†

A Study on the Giving-receiving Expressions that Collocate with the Japanese Verb "OSHIERU"

Hyunju Ju (University of Tsukuba, Graduate School of Humanities and Social Sciences)†

１．はじめに

本稿は『現代日本語書き言葉均衡コーパス』『日本語話し言葉コーパス』に見られる動詞「教える」と共起する授受表現を用い、ガ格名詞句と二格名詞句の省略の有無について情報構造の観点から考察したものである。事柄を叙述する文に、動作が「誰に」という方向性の表す要素が付け加えられる際、述語には方向性を示す要素が必要となる。

(1) a.　田中さんが本を買った。
　b.??田中さんが僕に本を買った。
　　+方向性　−方向性
　c.　田中さんが僕に本を買って{くれた/くださった/きた}。
　　+方向性　　+方向性　　(筆者作例[1])

例(1)のように、動詞「買う」が誰かに向けられて行われた際、ニ格名詞句を入れることだけでは不自然さを感じ(例(1b))、述語に授受表現や「てくる/いく」のような行為の方向性を示す成分を加えなければならない。一方、述語にすでに方向性が含まれていると思われる動詞[2]に関しては、以下の例(2)を用いて述べる。

(2) a.　田中さんが書き方を教えた。
　b.　田中さんが僕に書き方を教えた。
　　+方向性　　+方向性　　(筆者作例)

例(2b)が、例(1b)に比べて容認度が上がる要因について、まず動詞「買う」「教える」の方向性の有無に注目したい。情報処理振興事業協会技術センター編(1987)で「買う」「教える」の基本文型を調べると、「買う」の第 1 文型は「ガ/ヲ/V」であり、「教える」は「ガ/ニ/ヲ/V」の文型を第 1 文型として持つ[3]。例(1)を見ると「田中さんが本を買った」という構造

† murasakiju@gmail.com

[1] 例文作成は筆者によるものであるが、容認度判断については日本語母語話者 3 名(男性 1 名、女性 2 名)により判断してもらった。「?」は少し違和感を感じる文、「??」はかなり不自然さを感じ、文脈的制限がある場合のみ使用可能な文であると表示する。

[2] 方向性の意が含意されている動詞の例として、「送る」「届ける」「貸す」等があるが、本稿では「教える」に絞り、データを整理し報告する。「教える」以外の動詞の範囲まで広げて取り扱う研究は今後発表予定である。

[3] 情報処理振興事業教会技術センター編(1987)の『計算機用日本語基本動詞辞書 IPAL(Basic verbs) 辞書編』に「おしえる」と「かう」の第 1 文型の例として以下のようなものが記述されている。
・「おしえる」【教える】N1 が N2 ヲ N3 ニ(N4 デ)
　「S 先生は学生にチョムスキーを原書で教えている」
・「かう」【買う】N1 ガ N2 ニ N3 ヲ
　「彼が車を買った」

において、ニ格名詞句「僕に」を用いるためには、方向性を表す要素を含む述語との共起が必要である。その反面、例(2)の動詞「教える」は、以下のような意味を持っていることから、動詞自体が方向性を持っていることが伺える。

・行動や身を処し方などについて注意を与えて導く。いましめる。さとす。
・知っている事や自分の気持、要求などを他の人に告げ知らせる。
・知識、技芸などを身につけるようにさせる。教授する。
・おだてたりして、悪い事をするようにしむける。

(『日本国語大辞典』第２版より抜粋、下線は筆者による)

「教える」の意味から分かるように「教える」動作の対象となる人物が想定できる。つまり、動詞の意からニ格名詞句の存在が予想されるわけである。そのため、例(2b)はニ格名詞句が含まれていても通用するのである。本稿で注目したいのは、動詞「教える」にこのようにニ格名詞句が表す対象が想定できるにも関わらず、例(3)のように「てくれた」などの表現を用いることができる点である。

(3) 田中さんが僕に書き方を教えて{くれた/くださった}。
　　+方向性　　+方向性　　?

(筆者作例)

上述したように、動詞「教える」は方向性を持っているが、それに改めて行為の方向性を表す授受表現が加えられるのはなぜか。本稿では授受表現が文の情報構造においてどのような役割を果たしているかを調べる。

実際コーパスから動詞「教える」と共起する補助動詞の調査を行い、文の構造を見ると、省略された文の要素があることが分かった。次のような例がその代表的な例である。

(4) さて、私達が日本語教師になる時にまず一度は経験するのが教育実習です。学部の実習や日本語学校では皆一様に経験がなく、先生も教案の作り方から丁寧に教えてくれます。　(『日本語話し言葉コーパス』)[4]

(5) 映画っていうのは本当に人間を成長させる為に非常にいい。学校では教えてもらえないね。色んな勉強を教えてくれるものだと思うんで、是非皆さんも映画をですね、一本でも多く見ていただきたい。

(『日本語話し言葉コーパス』)

従って、本稿では、動詞「教える」と共起する授受表現の基本構造を調査し、文構造においてガ格名詞句とニ格名詞句の省略が持つ意味について考察する。第 2 章では授受表現の基本構造について考察し、第 3 章ではこの基本構造の要素の省略が表す情報構造について述べる。そして、第 4 章でコーパス・データを用い、動詞「教える」と共起する授受表現についてどのような情報構造を成しているかについて分析する。

２．授受表現の基本構造について

授受表現の基本構造について、山岡(1990:19-23)は以下のような構造で説明している。

(6)　渡辺さんが　　小林さんに　　本を　　貸し　てあげる。
　+ガ, human, Ag　+ニ, human, G　+ヲ, concrete, O　+貸ス
　+ガ, human, Ag　+ノタメニ, human, Bf　　+テアゲル

(ibid.:20)[5]

[4] コーパス・データの表記において、文字化されているデータを基とし、「、」「。」と下線は筆者による表記である。

[5] human は人間、Ag は動作主、G は目標、O は対象である。concrete は具体的で、Bf は受

動詞「貸す」は動詞「教える」と同じく動作そのものに行動の向けられる対象を含んでいるため、基本的に同じ構造を持つと言えよう。上記の例では、動詞「貸す」が向けている対象が授受表現の受益者と一致するため、「小林さんに」とニ格名詞句で表現する[6]。

このような文の構造を動詞「教える」と共起する授受表現で考えよう。

(7) 「もう少し辛くするとぐっと、うまくなるよ！」とか、言いながら　一緒に味付け直すと、な～るほど～と、なる彼女だったら、いいんですけど。。。。わたしは、ダンナにそうやって教えてもらいました。ホントにおいしくなるので、素直にきいてました。　(『現代日本語書き言葉均衡コーパス』Yahoo!知恵袋データ)

(8)　　わたしは、　　ダンナに　そうやって　　教え　　てもらいました。
　　に　　ガ　　教える
「ガ」➡「は」と表示　　ニ　　テモラウ

例(7)は恋人の手料理が美味しくないことが悩みである相手の相談に対し、話し手[7]が書いたアドバイスの文である。ここでは「教えてもらった」動作の動作主である話し手自身を主格にし、「教えた」人物をニ格で明示し、文の構造における格の明示は授受表現による格が優先されることが分かる。

次節では、動詞「教える」と共起する授受表現の情報構造を述べるための用語の整理と省略現象について述べる。

３．情報構造における省略現象

本節では情報構造における省略現象[8]について述べる。情報構造とは、Lambrecht(1994)によると、以下のように定義づけられている。

・INFORMATION STRUCTURE: That component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these structures as units of information in given discourse contexts.[9]　Lambrecht(1994:5)

また、カレン(2000)によると情報構造は「その成分の『情報価値』に基づいて分節される(ibid.:2)」と述べ、情報価値において最上位にあるのは焦点であり、その焦点を支えているものは「主題」、副詞成分による修飾は「設定部」と設定している。

益者を意味する。(山岡 1990:19-23)

[6] しかし、動詞が向けられている対象と、授受表現の受益者が一致していない場合には、以下のように表現できる。

・渡辺さんが 私のために 小林さんに 本を 貸し てあげる。(筆者作例)

上記は、ガ格名詞句の「小林さんに本を貸す」行為が「私」に向けられており、受益者は「～のために」で表している「私」である場面に限る。

[7] Yahoo!知恵袋からのデータであるため書き手となるが、書き言葉も談話の一部であると見なし、本稿では話し手とする。

[8] 本稿で取り扱う省略は、脱落とは異なる概念とする。脱落は、必要なものである要素が抜け落ちることであるが、省略は何らかの理由によって非明示化されることを示す。本稿では、授受表現におけるガ格名詞句とニ格名詞句の省略がどのような理由による現象なのかを解明する。

[9] 「文文法の一部で、事態の概念的表出としての命題が、語彙文法構造と組み合わされるもので、あるディスコースで、この情報を使用し、解釈する対話者の心的状態と一致するものである」(小野(2005:7)による訳)

先行研究の知見をまとめると、言語活動において話し手の「情報管理(Common Ground Management)」が重要となるが、適切な情報の価値をある言語形式を使って表す仕組みを情報構造であると言える。

この情報構造において、重要となる概念に「トピック」と「フォーカス」がある。Traugott(2010)によると、トピック(topic)は、文の構造の中で、話し手が現在話題にしている事柄のことを言い、フォーカス(focus)は話し手が情報を伝達するとき、聞き手の知らない新情報のことを示すとまとめられている。談話において、フォーカスを間違えると、結束性[10]を欠いたコミュニケーションとなり、円滑なコミュニケーションに失敗してしまう恐れがある。フォーカスは話し手と聞き手の間の共有知識が談話の前提となり、その管理能力で、コミュニケーションが円滑に進むかどうかが決まる。

このような情報構造において、省略とは、元々持っている文の要素の中、ある要素が省かれることである。省略の原理について「省略される要素は残される要素より情報の重要度が低いものでなければならない(中右・神尾・高見 1998:135)」という特徴がある。以下の例を見られたい。

(9) A：君は、朝 6 時に起きるんですか。
　　B :*はい、(私は) φ 起きるんです。
(中右・神尾・高見 1998:136)

上記の例の答え B が不自然となるのは、A の求めている情報が「朝 6 時に」起きるかどうかという情報であり、「起きる」ことに対しての情報の重要度は低いためである。つまり、重要度が高い情報は、省略されにくいこととなる。

授受表現の基本構造については前節で述べたが、実際に授受表現の使用を観察した結果、必ずしも基本構造の要素のすべてが明示されるわけではないことが分かった。

(10) 映画っていうのは本当に人間を成長させる為に非常にいい。学校では教えてもらえないね。色んな勉強を教えてくれるものだと思うんで、是非皆さんも映画をですね、一本でも多く見ていただきたい。
(=例(5))

省略可能な要素は、ガ格名詞句とニ格名詞句があり[11]、ガ格名詞句とニ格名詞句の両方明示される場合、ガ格名詞句のみある場合(つまり、ニ格名詞句が省略された場合)、ニ格名詞句のみある場合(つまり、ガ格名詞句が省略された場合)、ガ格名詞句とニ格名詞句の両方が省略される場合の 4 つのパターンが考えられる。

動詞「教える」と共起する授受表現の文構造における省略現象の条件について、次のような項目が設定できる。

・動詞「教える」と共起する授受表現においてガ格名詞句もしくはニ格名詞句が話し手自身または目の前にいる聞き手である場合には省略できる。
・動詞「教える」の動作主と受け手(動作の対象者)が、すでに発話で言及されており、話し手と聞き手両方が共有していると判断する重要度の低い情報は省略できる。

---

[10] 結束性とは、文と文の間における文法的・語彙的繋がりのことを示す。
(例) A:森田さんはどこ出身ですか。
　　B:東京です。
B の言葉が「私の出身は東京です」と言わなくてもいいのは、前の発話で A が森田さんの出身について聞いており、B の発話がその質問に対して繋がりがある答えを話すという結束性が期待されるためである。

[11] ここでヲ格の場合、動詞によって必須格とする場合とそうでない場合があるため、本稿では扱わないことにする。

・話し手の伝えようとする情報が、聞き手の知らない新情報であるときには省略されない。

４．各コーパスにおける動詞「教える」と共起する授受表現と省略現象に関する考察

本稿で研究対象とするのは『現代日本語書き言葉均衡コーパス』『日本語話し言葉コーパス』から収集した例の中、本動詞「教える」＋補助助詞「て」＋授受表現が使用された例である[12]。4.1 では各コーパスにおける動詞「教える」と共起する動詞について、4.2 では動詞「教える」と共起する授受表現の例を用い、ガ格名詞句とニ格名詞句の明示と省略に分け、その情報構造について述べる。

４．１　『現代日本語書き言葉均衡コーパス』『日本語話し言葉コーパス』から見られる「教える」とその補助動詞

本動詞「教える」を含む補助動詞を調べ、表 1 のような結果を得られた。

表 1 動詞「教える」と共起する補助動詞リストとその出現件数

本動詞	補助動詞	『現代日本語書き言葉均衡コーパス』の出現件数	『日本語話し言葉コーパス』の出現件数
教える	やる	266	0
	あげる	457	4
	さしあげる	14	0
	くれる	2,487	16
	くださる	16,776	4
	もらう	1,141	8
	いただく	1,393	5
	いる	830	8
	ある	10	0
	いく	74	4
	くる	44	0
	しまう	21	0
	その他	187	2
合計		23,700	51

『現代日本語書き言葉均衡コーパス』において動詞「教える」と共起する補助動詞のうち、授受表現と共起するケースが 95.1%とあり[13]、『日本語話し言葉コーパス』における動詞「教える」と共起する補助動詞のうち、72.5%が授受表現と共起していることが分かった。

次の図 1 と図 2 を通し、動詞「教える」と共起する動詞の類の中、授受表現が示す割合が高いことが分かる。授受表現以外の動詞としては、表 1 で述べた「いる・ある」「いく・くる」「しまう」などの例の出現頻度の合計した数である。

次節では、動詞「教える」と共起する授受表現における情報構造はどのようなものであ

[12] 『現代日本語書き言葉均衡コーパス』の内容は、オンライン検索システムである『NINJAL-LWP for BCCWJ』を利用し、『日本語話し言葉コーパス』は全文検索システム『ひまわり』を利用してデータを収集した。

[13] 『現代日本語書き言葉均衡コーパス』のデータで、動詞「教える」が使用された頻度は、30,502 件であるが、そのうち「教えて＋動詞」の表現形式は、23,700 件である。つまり、動詞「教える」は単独ではなく、「本動詞＋て＋補助動詞」の形式で使用される傾向があることが分かる。

るかについて述べる。

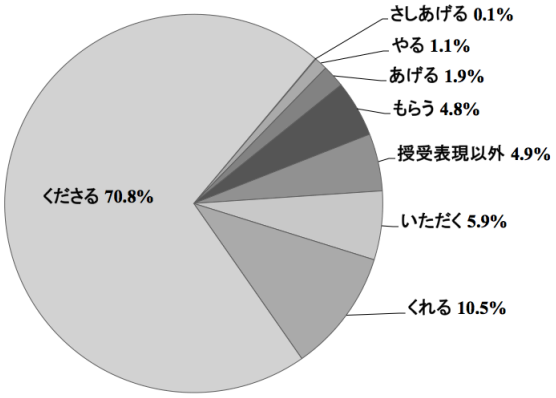


図 1 『現代日本語書き言葉均衡コーパス』における「教える」と共起する動詞リストと割合(%)

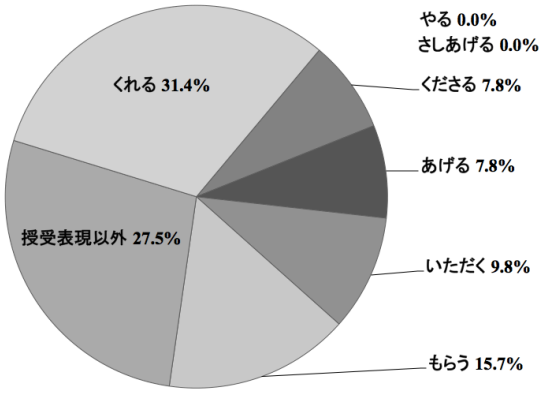


図 2 『日本語話し言葉コーパス』における「教える」と共起する動詞リストと割合(%)

４．２　動詞「教える」と共起する授受表現におけるガ格名詞句とニ格名詞句の明示と省略

前節で述べたように、動詞「教える」と共起する補助動詞には授受表現の使用が多いことが分かった。本節では、「教えて＋授受表現」の形式が使用された文を用い、文全体の情報構造において省略される情報と、明示される情報について探る。

まず、動詞「教える」と共起する授受表現の文構造について以下のように考えられる。

・A が　B に　O を　教え　てやる・あげる・さしあげる
　ガ　ニ　ヲ　教える
　ガ　ニ/ノタメニ　テヤル・アゲル・サシアゲル
・C が　A に　O を　教え　てくれる・くださる
　ガ　ニ　ヲ　教える
　ガ　ニ/ノタメニ　テクレル・クダサル
・A が　C に　O を　教え　てもらう・いただく
　ニ　ガ　ヲ　教える
　ガ　ニ　テモラウ・イタダク

上記の文構造において、授受表現で「教える」動作の方向性を表しているため、動作の主体もしくは動作の対象となるガ格名詞句とニ格名詞句を用いて方向性を表す役割が希薄となる。つまり、ガ格名詞句とニ格名詞句の明示には何らかの理由があると考えられる。

４．２．１　ガ格名詞句とニ格名詞句両方が省略される場合

コーパスの調査結果として、ガ格名詞句とニ格名詞句が両方省略された場合が顕著であったが、以下のような例がその代表的な例である。

(11) iTunes のバックアップ方法を教えてもらったのですが、バックアップしたファイルを再インストールした iTunes に入れるにはどうすれば良いのでしょうか？教えてもらえると助かります。

(『現代日本語書き言葉均衡コーパス』)

この例(11)で「方法を教えてもらった」と「教えてもらえると助かります」の両方、ガ格名詞句とニ格名詞句が省略されている。この例における省略の条件としては、話し手が読み手である聞き手への質問であり、「教えてもらう」のガ格名詞句は話し手自身となるため、省略可能である。「方法を教えてもらった対象」がニ格名詞句で明示されるはずだが、ここで省略されているのは情報構造の重要度が低い情報であるためである。つまり、ここでは「誰に」教えてもらったのかは重要な情報でないのである。

以上のように、動詞「教える」と共起する授受表現の文構造においてガ格名詞句とニ格名詞句が省略されるケースが多いが、次節ではガ格名詞句とニ格名詞句が明示されることについて述べる。

４．２．２　ガ格名詞句のみ明示される場合（ニ格名詞句が省略される場合）

次に、ガ格名詞句のみ明示された場合は、以下のような例が挙げられる。

(12) 一緒にいた観光客の人が教えてくれてその観光客の方はもう帰るって話しだったんだ。

（『日本語話し言葉コーパス』）

例(12)では、ガ格名詞句が明示された理由は、「観光客の人」が第 3 者であり、フォーカスされている。そのため、「教えてくれて」の後接部にも「観光客の人」の情報について話しが続いていることが観察できる。こうしたことをトピックの継続性の特徴を持つと言える。

４．２．３　ニ格名詞句のみ明示される場合（ガ格名詞句が省略される場合）

この節では、ガ格名詞句が省略され、ニ格名詞句のみ明示された場合について分析する。

(13) 切らないでくれ。あんたに教えてもらいたいことがある。ちょっと待て。

（『現代日本語書き言葉均衡コーパス』）

上記の例のように、電話での会話で、話し手自身が聞き手に「教えてもらいたい」ことであることは自明であるため、ガ格名詞句もニ格名詞句も省略される可能性があるが、「あんたに」のニ格名詞句が明示された理由については、他の対象からの対比の効果であり、「他の人ではなく、あなたに」という対比性を表すことができ、話し手の切実さをより感じさせるためである。

４．２．４　ガ格名詞句とニ格名詞句両方が明示される場合

ガ格名詞句とニ格名詞句の両方が明示された場面について、以下の例を見られたい。

(14) 父親の友人の家族のお姉さん達が私に泳ぎを教えてくれてんで、そうね、プールででも泳げるようになったよ。

（『日本語話し言葉コーパス』）

上記の例では、ガ格名詞句が第 3 者であり、会話の場面にいない人物であり、前の文脈においても触れたことがない新情報である。そのため、話し手は聞き手に伝える必要があることから明示され、フォーカスとしての情報構造を持つ。ニ格名詞句である「私に」は、話し手自身に向けられているため、言及しなくても分かる情報であるが、話し手自身を二格名詞句に明示することで、後節の「泳げるようになった」主体を表すトピックとして働くと言える。

５．おわりに

動詞「教える」と共起する授受表現の文構造において、ガ格名詞句とニ格名詞句は必須格でない理由は、ガ格名詞句とニ格名詞句の情報が前提化されることが無標であり、ガ格名詞句やニ格名詞句を明示することで、対比や主観性の強い有標性を持つことと考えられる。本稿の内容で取り扱った内容を表 2 のようにまとめる。

表 2 「教える」と共起する授受表現におけるガ格名詞句とニ格名詞句の明示・省略

省略の条件	ガ格名詞句とニ格名詞句が、話し手と聞き手の共通知識であるかもしくは情報の重要度が低い。
明示の条件	ガ格名詞句またはニ格名詞句が表すトピックの継続性があるかもしくは対比でトピックを強調する。

本稿では書き言葉、話し言葉に区別についての明確な区別がまだできていないことが指摘できる。この点については今後の課題とし、ウェブにおける書き込みや国会議事録の文字化データについて話し言葉的要素が強いため、書き言葉と話し言葉の中間的形式とカテゴリーを改めて設定し、分析を続けたい。

文　献

小野正樹 (2005) 『日本語態度動詞文の情報構造』、ひつじ書房.

甲斐ますみ (1995)「省略のメカニズム-談話の構造と関連性および聞き手の推論を中心に-」『岡山大学留学生センター紀要』 3、pp.1-18、岡山大学留学生センター.

カレル・フィアラ (2000) 『日本語の情報構造と統語構造』、ひつじ書房.

北原保雄 編 (2000) 『日本国語大辞典』第二版、小学館.

久野暲 (1978) 『談話の文法』、pp.5-124、大修館書店.

情報処理振興事業協会技術センター 編 (1987) 『計算機用日本語基本動詞辞書 IPAL(Basic verbs) 辞書編』、p.72、pp.86-87、情報処理振興事業協会技術センター.

坪井美樹 (2005) 「テ形接続形式と文法化」『国語と国文学』983(17:11)、pp. 13-25、東京大学国語国文学会.

中右実、神尾昭雄、高見健一 (1998) 「第Ⅱ部 第 1 章 省略」『日英語比較選書 2 談話と情報構造』、pp.114-138、研究社.

中野洋 (1991) 「『(て)いただく』文における省略」『計量国語学』18-2、pp.66-79、計量国語学会.

野元裕樹 (2013) 「文献紹介:Krifka(2008) Basic notions of Information Structure」2013 年度第 2 回研究会 東京外国語大学アジア・アフリカ言語文化研究所.

前川喜久雄 編 (2013) 『講座日本語コーパス 1.コーパス入門』、朝倉書店.

宮地裕 (1965) 「『やる・くれる・もらう』を述語とする文の構造について」『国語学』63、pp.21-33、日本語学会.

山岡政紀 (1990) 「授受補助動詞と依頼行為」『文藝言語研究 言語篇』17、pp.19-33、筑波大学文藝・言語学系.

Elizabeth Closs Traugott (2010) *Gradience, Gradualness and Grammaticalizatio*n, Amsterdam, John Benjamins. (福元広二 訳 (2011) 「第３章文法化と(間)主観化」高田博行 他 編『歴史語用論入門』、pp.59-70、大修館書店)

Knud Lambrecht (1994) *Information Structure and Sentence form : Topic, focus, and the mental representations of discourse referents,* Cambridge studies in Linguistics 71, pp.1-35, Cambridge University Press.

関連 URL

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

現代日本語書き言葉均衡コーパスのオンライン検索システム 『NINJAL-LWP for BCCWJ』
http://nlb.ninjal.ac.jp/

国立国語研究所『日本語話し言葉コーパス』 http://www.ninjal.ac.jp/corpus_center/csj/

全文検索システム『ひまわり』http://www2.ninjal.ac.jp/lrc/index.php

ポスター発表（1） グループA

3月6日（木） 15:10～16:10

コーパスコンコーダンサ『ChaKi.NET』の連続値データ型 (2)
―読み時間の表示―

浅原 正幸 (国立国語研究所コーパス開発センター) *
池本 優 (近畿大学)
森田 敏生 (総和技研)

Double Type Data on 'ChaKi.NET'
— Visualization of Reading Time —

Masayuki Asahara (Center for Corpus Development, NINJAL)
Yu Ikemoto (Kinki University)
Toshio Morita (Sowa Research Co., Ltd.)

1. はじめに

『ChaKi.NET』 (Matsumoto et al. (2005)) は、コーパスに付与されたメタデータ・形態論情報・係り受け情報を用いて文単位の検索を行ったり、形態論情報・係り受け情報に基づく頻度統計情報を取得したり、自動解析により付与された形態論情報・係り受け情報を修正したりすることができるコーパス管理システムである。開発当初は書記言語を対象にしていたためデータベース内部におけるデータ型のほとんどがテキストやアノテーションを格納する文字列型やアノテーションを抽象化した整数型が用いられてきた。前回のワークショップでは ChaKi.NET の連続値データ型 (浅原・森田 (2013)) の導入について発表し、利用法として話し言葉コーパスなどの音声コーパスの時刻情報を格納し、形態素単位の発話継続時間情報を可視化するデモを行った。

今回の発表では「現代日本語書き言葉均衡コーパス」の文の読み時間の可視化を行う。文の読み時間とは、被験者に文もしくは文章を呈示して文の一部分を読むのにかかる時間を計測したものである。読み時間の遅延を人の文処理機構を解明するための手がかりとして用いる。読み時間の測定方法として安価な機材で測定可能な「移動窓方式の自己ペース読文実験」や専用の実験装置を用いた「視線走査実験」などがある。

本稿では、移動窓方式の自己ペース読文実験や視線走査実験結果を元にどのような可視化が可能であるのか、また、ChaKi.NET 上でどのような分析が可能であるのかを議論するとともに、現在何ができないかを議論する。

2. 読み時間の測定方法

本節では読み時間の測定方法として「移動窓方式の自己ペース読文実験」や「視線走査実験」について概説する。

* masayu-a@ninjal.ac.jp

2.1　自己ペース読文実験

自己ペース読文実験は、スペースキーなどを押すごとにディスプレイ上に刺激文が単語もしくは文節単位に順次呈示し、被験者のペースでスペースバーをおしながら文を読み進め、その表示時間を記録することで読み時間を計測する実験である。Linger と呼ばれるソフトウェアが良く用いられる。自己ペース読文実験は、刺激文の呈示の方法が二種類ある。一つは中央表示方式で、もう一つは移動窓方式である。中央表示方式は、画面中央に単語または文節が表示され、スペースキーを押すごとに消えて別のものが表示する方式である。被験者から見ると同じ場所に新しい文字列が示される。移動窓方式は、最初に文字の数だけハイフンやアンダーバーなどを表示し、スペースキーを押すごとに部分的に単語または文節を表示する方式である。図 1 に移動窓方式の自己ペース読文実験の画面遷移を示す。被験者からみると見える文字列部分（窓）が移動していくように示される。調査対象の言語現象にもよるが、日本語の場合は通例文節単位に呈示し、何 msec かけて文節文字列を読むかを評価する。読文作業においては後戻りして読むことを許さないため、各文節は高々 1 回の注視時間が記録される。

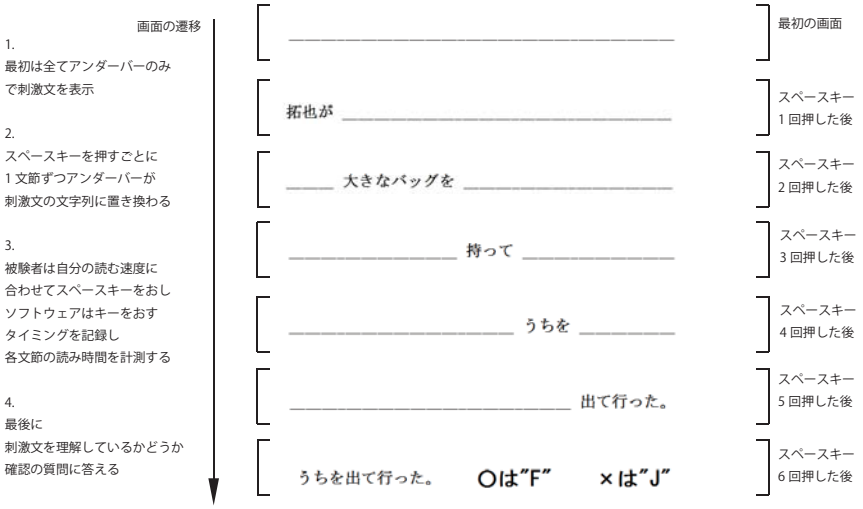


図 1　移動窓方式による画面の遷移例

現在、「現代日本語書き言葉均衡コーパス」(BCCWJ) のサンプルを用いて移動窓方式による自己ペース読文実験による読み時間評価を実施している。表 1 に BCCWJ に対する読み時間評価結果を示す。1 画面に最大 5 文呈示する。横の最大文字数は約 40 文字で、40 文字を超えた文節の次で改行して呈示する。表 1 の例は、サンプル名 PN1c_00001 の 2 画面目の一部の結果を示しており、文節 ID 4, 5, 14 の直後で改行している。文単位は基本的に BCCWJ の基準に従うが、あまりに不自然なものについては適宜修正している。

通例、呈示文節の「呈示順」「文字数」「モーラ数」などの情報で回帰を行ったうえで、（標準）残差を算出しどの部分で読み時間の遅延が発生するかを識別する。

表 1　BCCWJ に対する読み時間付与結果（移動窓方式による自己ペース読文実験）

画面 ID	文節 ID	呈示文節	停留時間 (msec)
2	0	子育ての	1428
2	1	かたわら	248
2	2	テレビ、	208
2	3	ラジオなどで	191
2	4	活躍中。	206
2	5	３３歳。	203
2	6	幼稚園から	203
2	7	大学まで	230
2	8	通った	199
2	9	青山学院では、	170
2	10	とにかく	193
2	11	活発で、	109
2	12	目立つ	192
2	13	生徒だったと	204
2	14	いう。	205
⋮	⋮	⋮	⋮



図 2　視線走査装置

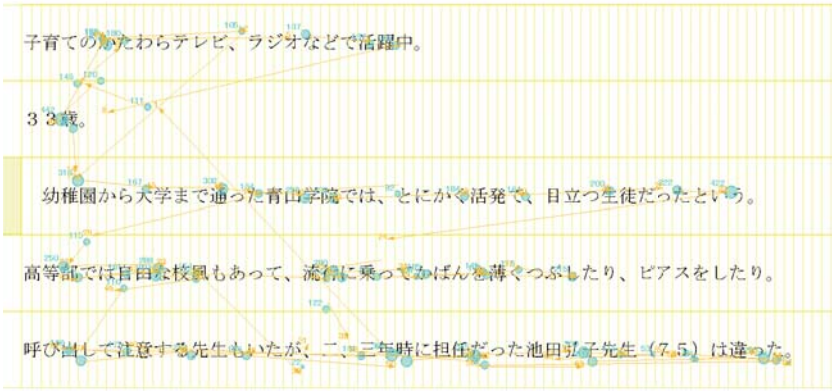


図 3　視線走査実験で得られるデータ例

2.2　視線走査実験

視線走査実験とは、視線走査装置を用いて画面上のどの文字を注視しているかを注視時間とともに計測することにより、文字列の文の読み時間を測定する方法である。心理言語実験に適した視線走査装置として SR Research 社の Eyelink CL を利用する。図 2 に機材の写真を示す。誤差を最小限にするためにハーフミラー越しに正面から瞳孔に弱い赤外線をあて、その反射をカメラでとらえるタワーマウントを用いる。あご台で顔を固定することで精度の高いデータを取得することができる。Eyelink CL は最大 1000Hz のサンプリングレートで瞳孔を走査し誤差が視野角 0.2 度未満としている。これは 50cm 離れてみる場合のモニター上の誤差が約 2mm となる。我々の実験では 21 インチの 1920x1080 ピクセル画面上の文字を 22 ポイントで呈示する。等幅フォントの文字幅は約 8mm である。

図 3 に視線走査実験で得られるデータ例を示す。表 1 と同じ PN1c_00001 の 2 画面目の結果

表 2　Dundee Eye-Tracking Corpus

WORD	TEXT	LINE	OLEN	WLEN	XPOS	WNUM	FDUR	OBLP	WDLP	LAUN	TXFR
Are	1	1	3	3	1	1	216	1	1	0	351
tourists	1	1	8	8	6	2	156	2	2	-5	3
enticed	1	1	7	7	17	3	227	4	4	-11	1
these	1	1	5	5	25	5	187	1	1	-8	73
attractions	1	1	11	11	33	6	182	3	3	-8	2
threatening	1	1	11	11	44	7	96	2	2	-11	3
Blink	1	-99	11	11	-99	-99	82	-99	-99	-99	-99
threatening	1	1	11	11	52	7	232	10	10	-99	3
very	1	1	4	4	62	9	335	2	2	-10	56
their	1	1	5	5	57	8	168	3	3	5	225
existence?	1	1	10	9	65	10	173	0	0	-8	4
existence?	1	1	10	9	71	10	188	6	6	-6	4
existence?	1	1	10	9	72	10	88	7	7	-1	4
enticed	1	1	7	7	19	3	174	6	6	53	1
these	1	1	5	5	29	5	168	5	5	-10	73
these	1	1	5	5	28	5	170	4	4	1	73
attractions	1	1	11	11	36	6	271	6	6	-8	2
attractions	1	1	11	11	34	6	88	4	4	2	2
threatening	1	1	11	11	46	7	232	4	4	-12	3
very	1	1	4	4	61	9	202	1	1	-15	56
existence?	1	1	10	9	72	10	222	7	7	-11	4
existence?	1	1	10	9	74	10	157	9	9	-2	4

例文 “Are tourists enticed by these attractions threatening their very existence?”

であり、●が注視点を表現し、その大きさが注視時間を表現している。注視点間を結ぶ線が、注視点の移動を表現している。画面から何度か後戻りして読むことが行われており、注視点も文節境界とは独立に分布しているように見える。

以前は Tobii や SR Research 社などの一部のベンダーが高価な視線走査装置を販売していたが、最近 Tobii 社が一般向けの視線走査装置を発売している。例えば、Tobii REX が 249 ユーロ、Tobii EyeX が 170 ユーロで発売される。精度では高価な機材に劣るが、一般に手に入る価格になり、言語研究において視線走査装置の利用が加速することが考えられる。

3.　視線走査実験データの表現手法

本節では視線走査実験データの表現手法について議論する。まず数少ない読み時間情報つきコーパスの先行研究である “Dundee Eye-tracking Corpus” Kennedy and Pynte (2005) について述べ、現在収集している日本語視線走査実験データの表現手法についても示す。

3.1　Dundee Eye-Tracking Corpus

読み時間を付与したコーパスの先行研究として Kennedy らの Dundee Eye-Tracking Corpus (以下 “Dundee Corpus”) がある。対象言語を英語とフランス語、それぞれ 10 人の母語話者を被験者とし、視線走査装置を利用して 20 ファイルの新聞社説に対する視線走査情報を記録している。各ファイルは 5 行ごとからなる 40 画面により構成されており、研究用途に一次情報が公開されている。Dundee Corpus は特定の言語現象分析を目的とせずに作成されたコーパス

に基づいており、心理言語学におけるさまざまな仮説の客観的な検証に用いられている。

表 2 に Dundee Corpus の例を示す。WORD が注視点が含まれた単語で時系列に表示される。TEXT が呈示画面番号、LINE が呈示行番号、OLEN が句読点を含めた文字列（オブジェクトと呼ぶ）長、WLEN が句読点を含めない単語文字列長、XPOS が注視点になる文字位置、WNUM が単語のインデックス、FDUR が停留時間 (単位は msec。以下時間・時刻については msec とする。)、OBLP がオブジェクト中の注視点位置、WDLP が単語中の注視点位置、LAUN が注視前のサッケード（視点を移動する眼球運動）長を相対位置で示したもの、TXFR がテキスト中の単語頻度である。尚、WORD の Blink はまたたきを表現する。

この例では 3 単語目 (1-origin) の “by” が読み飛ばされたり、 8 単語目の “their” が 9 単語目の “very” より後に読まれたり (単語を超えても LAUN が正の値) する。10 単語目の単語 “existence” は文字列長が長いために単語内で複数個所注視点が含まれる。

3.2　日本語視線走査実験データの表現手法

本節では Dundee Corpus を元に日本語視線走査実験データをいかにして表現するかについて議論する。

再度図 3 を参照されたい。分析の際には細長い長方形のグリッドを等幅フォントの半分の幅かつ文字列を中心に行間の高さで稠密に配置し、半文字単位で注視点の位置を割り当てる。視線走査装置はこのグリッド単位に注視されたかどうかを視線の停留時間とともに記録する。

分析の際にはこのグリッド上の注視点の停留時間を文字単位、形態素単位（国語研短単位 or 国語研長単位)、文節単位など用途に応じた単位に割り当てる必要がある。さらに読み戻しなどをどう扱うかを決める必要がある。

現在、データの表現方法として、最初の 1 回目の視線走査のみを表示する方法（First Pass と呼ぶ)、全ての視線走査を表示する方法の 2 種類を検討している。以下 2 種類の方法を示すが、確定したものではなく意見の募集を行っている。

3.2.1　1 回目の視線走査のみを表示する方法（First Pass）

一つ目の方法は、1 回目の視線走査のみを表示し、他の情報を捨てる方法である。以下 “First Pass” と呼ぶ。この場合、原文の文字列順序に沿って視線走査記録を示すことができる一方、読み戻しなどの一部の情報は捨てられてしまう。また読み戻しを文字・形態素・文節のどの単位で認定するかにより、捨てられる情報が変わってくる。

表 3 に文字単位の日本語の視線走査情報（First Pass）の例を示す。LINE が何行目か、CIDX が原文文字列上の文字位置、FDUR が 1 回目の注視時間、DSUM が複数回の視線停留時間の合計、FSTT が 1 回目の注視開始時刻、FEND が 1 回目の注視終了時刻、CHAR が文字、SUW が形態素、BPH が文節である。SUW は基本的に国語研短単位に準拠しているが、数値のみだけ元の書字形で記述している。実データにおいては、図 3 のように注視点は文字幅の半分単位で計測しているが、これを文字単位に集計したものである。

FSTT を若い順にたどることによって被験者がどの順で文を読んでいるかがわかる。この例では、前画面から復帰するための途中で CIDX 26 の注視点を経て CIDX 6 の注視点で読文を開始し、CIDX 5 に推移している。また CIDX 6 と CIDX 5 の関係をみると、周辺視野でとな

表 3　日本語の視線走査情報（First Pass）

LINE	CIDX	FDUR	DSUM	FSTT	FEND	CHAR	SUW	BPH
1	1	.	0	.	.	子	子育て	子育ての
1	2	.	0	.	.	育	子育て	子育ての
1	3	.	0	.	.	て	子育て	子育ての
1	4	.	0	.	.	の	の	子育ての
1	5	168	554	1055	1222	か	かたわら	かたわら
1	6	180	180	857	1036	た	かたわら	かたわら
1	7	.	0	.	.	わ	かたわら	かたわら
1	8	.	0	.	.	ら	かたわら	かたわら
1	9	.	0	.	.	テ	テレビ	テレビ、
1	10	.	0	.	.	レ	テレビ	テレビ、
1	11	.	0	.	.	ビ	テレビ	テレビ、
1	12	105	105	3120	3224	、	、	テレビ、
1	13	.	0	.	.	ラ	ラジオ	ラジオなどで
1	14	.	0	.	.	ジ	ラジオ	ラジオなどで
1	15	.	0	.	.	オ	ラジオ	ラジオなどで
1	16	187	187	1280	1466	な	など	ラジオなどで
1	17	.	0	.	.	ど	など	ラジオなどで
1	18	.	0	.	.	で	で	ラジオなどで
1	19	167	167	1490	1656	活	活躍	活躍中。
1	20	.	0	.	.	躍	活躍	活躍中。
1	21	186	186	1674	1859	中	中	活躍中。
1	22	.	0	.	.	。	。	活躍中。
2	23	.	0	.	.	３	３３	３３歳。
2	24	.	0	.	.	３	３３	３３歳。
2	25	442	442	2179	2620	歳	歳	３３歳。
2	26	146	311	662	807	。	。	３３歳。
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

りの単語や文節を確認していることがあることがわかる。このような現象をどのように扱うかが問題となる。

表 4 単語単位にデータを集計したものを示す。WIDX が単語順を表す。CIDX 6 と CIDX 5 のように連続して同じ単語を見ている場合には合算されるが、CIDX 12 のように先の方を読んでから戻ってきたと考えられるものについては排除される。

3.2.2　全ての視線走査を表示する形式（Whole Pass）

二つ目の方法は、全ての視線走査を視線走査順に表示する方法である。以下 “Whole Pass” と呼ぶ。この場合、被験者の視線走査順序に沿って視線走査記録を示すことができる一方、原文の文字列情報の可読性がなくなる。

表 5 に単語単位の日本語の視線走査情報（Whole Pass）の例を示す。FIDX が視線走査順、FCNT が注視回数である。FIDX 3 までは前画面から復帰するための注視であると考える。

4.　ChaKi.NET 上での読み時間の可視化

本節ではどのようにして読み時間を可視化するかについて示す。具体的には以下の 3 種類の可視化を検討している。

表 4　日本語の視線走査情報（First Pass）の単語単位の集計

WIDX	FDUR	DSUM	FSTT	FEND	SUW	BPH
1	.	0	.	.	子育て	子育ての
2	.	0	.	.	の	子育ての
3	348	734	857	1222	かたわら	かたわら
4	.	0	.	.	テレビ	テレビ、
5	105	105	3120	3224	、	テレビ、
6	.	0	.	.	ラジオ	ラジオなどで
7	187	187	1280	1466	など	ラジオなどで
8	.	0	.	.	で	ラジオなどで
9	167	167	1490	1656	活躍	活躍中。
10	186	186	1674	1859	中	活躍中。
11	.	0	.	.	。	活躍中。
12	.	0	.	.	３３	３３歳。
13	442	442	2179	2620	歳	３３歳。
14	146	311	662	807	。	３３歳。
⋮	⋮	⋮	⋮	⋮	⋮	⋮

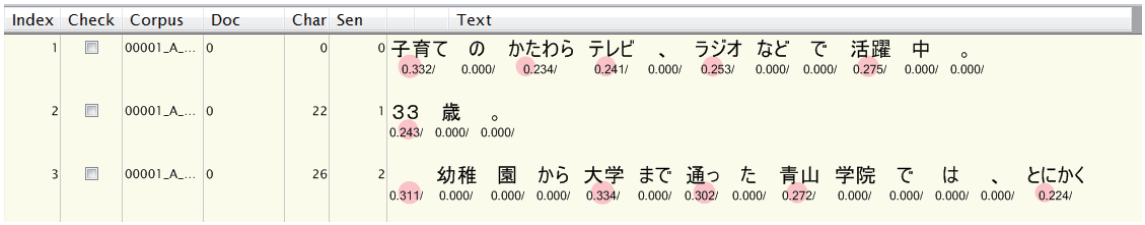
表 5　日本語の視線走査情報（Whole Pass）

LINE	FIDX	FDUR	DSUM	FCNT	FSTT	FEND	CHAR	SUW	BPH
5	1	366	366	1	7	372	年	年	二、三年時に
-99	2	111	111	1	528	638			
2	3	146	311	2	662	807	歳	歳	３３歳。
1	4	180	180	1	857	1036	た	かたわら	かたわら
1	5	168	320	2	1055	1222	か	かたわら	かたわら
1	6	187	187	1	1280	1466	な	など	ラジオなどで
1	7	167	167	1	1490	1656	活	活躍	活躍中。
1	8	186	186	1	1674	1859	中	中	活躍中。
-99	9	120	.	.	2032	2151			
2	10	442	442	1	2179	2620	歳	歳	３３歳。
1	11	234	234	1	2663	2896	か	かたわら	かたわら
1	12	152	320	2	2914	3065	か	かたわら	かたわら
1	13	105	105	1	3120	3224	、	、	テレビ、
2	14	165	311	2	3685	3849	歳	歳	３３歳。
3	15	318	318	1	3889	4206	稚	幼稚	幼稚園から
3	16	167	167	1	4234	4400	大	大学	大学まで
3	17	302	302	1	4425	4726	通	通っ	通った
3	18	259	259	1	4752	5010	学	学院	青山学院では、
3	19	239	239	1	5031	5269	院	学院	青山学院では、
3	20	184	184	1	5312	5495	く	とにかく	とにかく
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

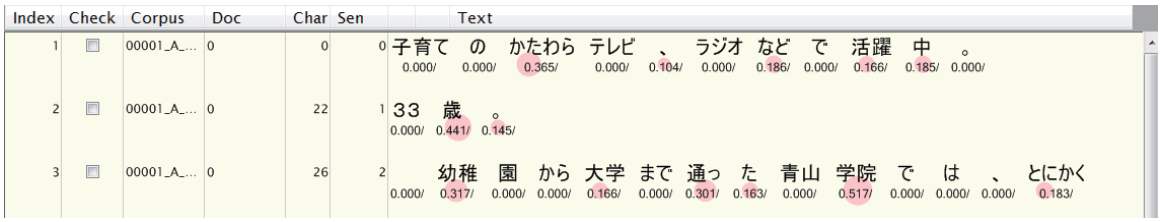
- 自己ペース読文実験で得られた読み時間の可視化
- 視線走査実験で得られた読み時間の可視化 (First Pass:コーパス出現順単語呈示)
- 視線走査実験で得られた読み時間の可視化 (Whole Pass:読み時間順単語呈示)

一つ目の自己ペース読文実験で得られた読み時間の可視化を図 4 上段に示す。自己ペース読文実験で得られた読み時間は文節単位のものである。読み時間は文節単位に割り当てられているが、ChaKi.NET において時間情報の割り当て単位は形態素単位になっている。検索の簡便

自己ペース読文実験の ChaKi.NET 上での可視化（文節頭に注視時間情報付与）



日本語視線走査実験 (First Pass) の ChaKi.NET 上での可視化（文字単位の注視時間情報を単語単位に集積）



日本語視線走査実験 (Whole Pass) の ChaKi.NET 上での可視化（文字単位の注視時間情報を単語単位に表示。集積は行わない。）

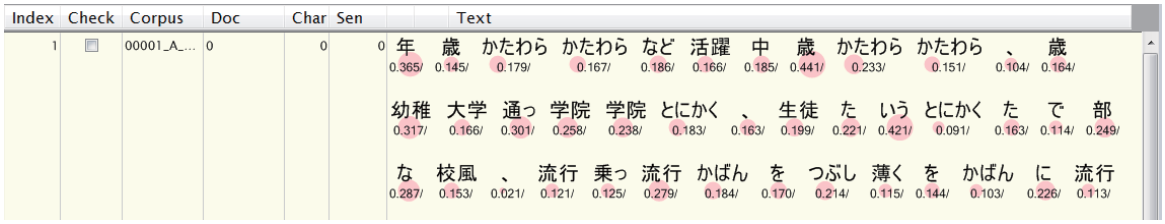


図 4　ChaKi.NET 上での文の読み時間の可視化

のために文節頭の形態素に文節全体の時間情報を付与している。読み時間の計測においては後戻りや文節飛ばしを許さないために全ての文節頭の形態素に時間情報が付与される。また視線走査装置とは異なり、注視している時間だけでなく視線を移動する時間（サッケード）も各形態素に割り当てられる。元データに付与されている形態論情報や係り受け情報との親和性がよく ChaKi.NET 上で統合して検索しやすい。

二つ目の 3.2.1 節で示したデータ表現形式 (First Pass) の可視化を図 4 中段に示す。一つ目の可視化との違いとして、全ての文節に読み時間が付与されない、文節中で真に注視された単語に読み時間が付与される、必ずしも文字列順に読まれているわけではないという点がある。この三点から時間情報に基づいて有用な検索をすることは難しい。さらに同一単語内で連続する注視点についてはその注視点間を移動する時間は読み時間として含まれているが、それ以外の注視点間を移動する時間は読み時間として含まれない点がある。また自動処理により処理されるデータでは、前の画面から復帰する途中の移動で注視されたものについては一回目の注視点 First Pass として認定されてしまうという点がある。

三つ目の 3.2.2 節で示したデータ表現形式 (Whole Pass) の可視化を図 4 下段に示す。上記二つと異なり文字列の線形順序の情報が失われており、そのままでは元の言語現象を問い合わせることが難しい一方、注視順に呈示されている。言語分析には不適かもしれないが、読み戻しや時間をかけて読んでいるところや読み飛ばしているところを重点的に観察することで要約補助用途などの利用が考えられる。

5. 課題

本節ではデータ表現形式と可視化の課題について検証する。

検証の前提として、可視化の目的について示す。我々の研究目的は心理言語研究と言語処理の二つの目的がある。一つ目の目的は例文が統制されている心理言語実験報告が、コーパス中の実際の用例でも観察できるかどうかを検証することである。コーパスで対象となる言語現象を形態論情報と係り受け構造により絞り込み検索をし、そこで観察される読み時間のふるまい俯瞰してみることが必要になる。二つ目の目的は言語処理応用である。既存の言語解析開発においてはコーパスに対して専門家がアノテーションを行い、それを機械学習に基づいて再現することで解析器を構成することが多い。一方、かな漢字変換などにおいては、系列学習のような技術が開発される一方で、予測変換といった実際の人間のふるまいを再現するようなモデルも提案されている。かな漢字変換は「書く」行為のモデル化だが、「読む」行為においても同様なモデル化が可能になると考える。今後ラップトップ PC やスマートフォンなどのカメラを利用して、大量の視線走査情報情報が得られることが考えられるが、これらを用いた自動要約・難読箇所検出・テキスト簡単化などが可能になるのではないかと考える。

このような目的のもと、現状の課題について示す。

- 周辺視野の可視化
  現在は何らかの基準で代表する形態素に読み時間情報を付与している。しかし、周辺視野で隣接している形態素や文節なども同時に見ている可能性があり、読み飛ばされているのか、周辺視野で確認しているのかを識別するような情報が必要になる。
- 複数人被験者のデータの可視化
  心理言語実験においては複数人の被験者データの分析が必要になる。現在、一つのテキストに対して一つの読み時間情報しか格納できない。
- 文字数やモーラ数による正規化
  心理言語実験においては、平均時間を評価したり、呈示順・文字列長・モーラ長に基づいて回帰分析を行いその残差を評価したりすることが行われる。読み時間そのものでは個人差などがあり何らかの正規化機構や統計論的モデル化が求められる。
- 文字列の線形順序と視線走査順序の相互関係
  自己ペース読文実験結果のように文字列の線形順序と視線走査順序が一致する場合については問題が起こらないが、視線走査実験のように文字列の線形順序と視線走査順序が必ずしも一致せず、読み戻しや読み飛ばしが起きるよう場合は、その相互作用の可視化が重要になる。
- 読み戻しの可視化
  自動要約や難読箇所検出においては読み戻しや読み飛ばしの可視化が重要になる。読み飛ばしは文字列の線形順序で First Pass の表現方法でも十分読み飛ばされていることが確認できる一方、読み戻しについては可視化できているとは言い難い。現在 Whole Pass などを用いて ChaKi.NET の WordList 機能を用いることで、仮想的に読み戻し回数の可視化が可能であるが、十分とは言えない。

このような課題に対し、ChaKi.NET のセグメント・リンク・グループの利用を検討している。ChaKi.NET では、形態論情報や係り受け構造に追加して、任意の構造としてセグメント・リンク・グループを表現することができる。例えば、周辺視野の情報をセグメントで表現したり、First Pass の表現上で視線の推移をリンクで表現したり、Whole Pass の表現上で同一形態素をグループで表現したりすることが考えられる。

6. おわりに

本稿では、コーパス管理システム ChaKi.NET の連続値データ型を用いて、読み時間を格納し可視化する手法について紹介した。現在作成中の BCCWJ に対する読み時間アノテーションは他の用途に利用されることも歓迎である。それぞれの用途に基づいて、必要な情報や利用者系は異なってくると考える。ポスター発表ではコーパスやツールに対する要望などを伺いたい。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

浅原正幸・森田敏生 (2013).「コーパスコンコーダンサ『ChaKi.NET』の連続値データ型」 第 4 回コーパス日本語学ワークショップ予稿集, pp. 249–256.

Kennedy, A., and J. Pynte (2005). “Parafoveal-on-foveal effecits in normal reading.” *Vision Research*, 45, pp. 153–168.

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

関連 URL

- 「ChaKi.NET」 Web ページ：`http://sourceforge.jp/projects/chaki/releases/`
- 「Linger」Web ページ：`http://tedlab.mit.edu/~dr/Linger/`

外れ値検出手法を利用した Misleading データの検出

吉田 拓夢 (茨城大学 工学部 情報工学科) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Detection of Misleading Data by Outlier Detection Methods

Hiromu Yoshida (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応の問題に対して，識別精度を低下させる Misleading データを検出するために，外れ値検出手法を利用する．

自然言語処理のタスクにおいて帰納学習手法を用いる際，訓練データとテストデータは同じ領域のコーパスから得ていることが通常である．ただし実際には異なる領域である場合も存在する．そこである領域（ソース領域）の訓練データから学習された分類器を，別の領域（ターゲット領域）のテストデータに合うようにチューニングすることを領域適応という (Sogaard (2013)) [3]．領域適応の問題の一つは負の転移である (Rosenstein et al. (2005))．これはソース領域のデータを使いすぎるとターゲット領域での識別精度が下がる現象である．我々は負の転移の原因を Misleading データの存在だと考えている．Misleading データとは分類器の学習に悪影響を与えるデータであり，Misleading データを検出，削除しておくことは分類器の精度向上に寄与する (Jiang and Zhai (2007))．

本論文では Misleading データはターゲット領域に対して外れ値になっていると予想し，この予想を確認する．まず訓練データ D から分類器を作成し，テストデータでその正解率 p_0 を測る．次に D の中の各データ x に対して，$D-x$ から分類器を作成し，テストデータでその正解率 p_1 を測る．$p_1 > p_0$ のとき x は Misleading データと考えられる．このようにして予め Misleading データを検出しておき，それを正解データと考えて，次に外れ値検出手法を利用して，どの程度 Misleading データを検出できるかを調べる．外れ値検出手法としては (1) 最近傍法，(2) LOF および (3) 確率密度比を試した．

実験では現代日本語書き言葉均衡コーパス (BCCWJ コーパス (Maekawa (2007))) における 3 つの領域 OC (Yahoo! 知恵袋)，PB (書籍) 及び PN (新聞) を利用する．SemEval-2 の日本語 WSD タスク (Okumura et al. (2010)) ではこれらのコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．すべての領域である程度の頻度が存在する多義語 16 単語を対象にして，WSD の領域適応の実験を行う．領域適応としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りが存在する．結果 $16 \times 6 = 96$ 通りの WSD の領域適応の問題に対して実験を行った．

実験の結果，Misleading データの存在自体は確認できたが，外れ値検出手法による Misleading データの検出精度は低かった．外れ値検出手法では本論文で設定したような Misleading データの検出は困難であるが，負の転移の有無を判定することは，ある程度可能であることが判明した．外れ値ではない Misleading データがどのような特徴を持っていたかを調べることが今後の課題である．

2　外れ値検出手法

本論文では外れ値検出手法を利用して，Misleading データの検出を試みる．利用した外れ値検出手法は (1) 最近傍法，(2) LOF および (3) 確率密度比である．以下，それぞれの手法を説明する．

[2] 10t4067l@hcs.ibaraki.ac.jp
[2] shinnou@mx.ibaraki.ac.jp
[3] 領域適応は機械学習の分野では転移学習 (神嶌敏弘 (2010)) の一種と見なされている．

2.1 最近傍法 (Erk の手法)

Erk による外れ値検出の手法 (Erk (2006)) を示す．外れ値の度合いを測るデータ点を点 x とする．この点 x に対して，対象データの中で最近傍となる点 t_n と，その点 t に対する最近傍点 t_n' を定める．これらの 3 つの点について，以下の距離を求める．

点 x と点 t_n の距離 dxt
点 t_n と点 $t_n{}'$ の距離 dtt'

この 2 つの距離を用いて，以下のように外れ値 $p_{NN}(x)$ を定める．

$$p_{NN}(x) = \frac{d_{xt}}{d_{tt'}}$$

2.2 LOF

LOF(local outlier factor) は密度をベースとした外れ値検出手法である (Breuning et al. (2000))．ある点のまわりの密度が他の点と比べて小さいほど，LOF の値は大きくなる．LOF の値を測る点を x としたとき，x の k 距離近傍集合 $N_k(x)$ を以下の様に定める．

$$N_k(x) = \{y \in D \setminus \{x\} | d(x,y) \leq kdist(x)\}$$

ここで，$kdist(x)$ は以下の条件を満たす $d(x,o)$ である．

1. 少なくとも k 個のデータ $o' \in D \setminus \{x\}$ に対して $d(x,o') \leq d(x,o)$ が成立する
2. 高々 $k-1$ 個のデータ $o' \in D \setminus \{x\}$ に対してのみ $d(x,o) < d(x,o)$ が成立する

すなわち，簡単には k 距離近傍集合 $N_k(x)$ は点 x から k 番目に近い点 o_k までの距離 $kdist(x) = dist(x,o_k)$ の範囲内にある点の集合である．LOF の算出に先立ち，まずは以下を求める．

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} rd_k(x,y)}$$

これは局所到達可能密度 (local reachability density，lrd) と呼ばれる値で，x の k 近傍内にあるデータの到達可能距離 (reachability distance，rd) の平均の逆数となっている．到達可能距離 rd は以下で定める値である．

$$rd_k(x,y) = \max\{d(x,y), kdist(y)\}$$

つまり，点 x と y の距離が y の k 距離よりも近い場合には y の k 距離に置き換えて到達可能距離 rd としている．以上をもって，LOF は次式により定められる．

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}$$

上式に示されるとおり，LOF は点 x の局所到達可能密度と点 x の k 近傍点の局所到達可能密度との平均を取っている．

2.3 確率密度比

確率密度比も外れ値検出手法として利用できる．確率密度比の算出は困難であるが，論文 (新納浩幸・佐々木稔 (2014)) では以下の簡易な手法を提案している．

対象単語 w の用例 $\boldsymbol{x}$ の素性リストを $\{f_1, f_2, \cdots, f_n\}$ とする．求めるのは領域 $R \in \{S,T\}$ 上の $\boldsymbol{x}$ の分布 $P_R(\boldsymbol{x})$ である．ここで Naive Bayes で使われるモデルを用いる．Naive Bayes のモデルでは以下を仮定する．

$$P_R(\boldsymbol{x}) = \prod_{i=1}^{n} P_R(f_i)$$

領域 R のコーパス内の w の全ての用例について素性リストを作成しておく．ここで用例の数を $N(R)$ とおく．また $N(R)$ 個の用例の中で，素性 f が現れた用例数を $n(R,f)$ とおく．MAP 推定でスムージングを行い，$P_R(f)$ を以下で定義する (高村大也 (2010))．

$$P_R(f) = \frac{n(R,f)+1}{N(R)+2}$$

以上より，ソース領域 S の用例 $\boldsymbol{x}$ に対して，確率密度比 $w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ が計算できる．

3　実験

BCCWJ コーパスの PB(書籍)，OC(Yahoo! 知恵袋) 及び PN (新聞) を異なった領域として実験を行う．SemEval-2 の日本語 WSD タスク (Okumura et al. (2010)) ではこれら領域のコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．この 3 つの領域からある程度頻度のある多義語 16 単語を WSD の対象単語とする．これら単語と辞書上での語義数及び各コーパスでの頻度と語義数を表 1 に示す．[4]領域適応の方向としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りの方向が存在する．

表 1: 対象単語

単語	辞書上の語義数	OC での頻度	OC での語義数	PB での頻度	PB での語義数	PN での頻度	PN での語義数
言う	3	666	2	1114	2	363	2
入れる	3	73	2	56	3	32	2
書く	2	99	2	62	2	27	2
聞く	3	124	2	123	2	52	2
子供	2	77	2	93	2	29	2
時間	4	53	2	74	2	59	2
自分	2	128	2	308	2	71	2
出る	3	131	3	152	3	89	3
取る	8	61	7	81	7	43	7
場合	2	126	2	137	2	73	2
入る	3	68	4	118	4	65	3
前	3	105	3	160	2	106	4
見る	6	262	5	273	6	87	3
持つ	4	62	4	153	3	59	3
やる	5	117	3	156	4	27	2
ゆく	2	219	2	133	2	27	2
平均	3.44	148.19	2.94	199.56	3.00	75.56	2.69

3.1　正解データの構築

語義曖昧性解消の領域適応において Misleading データの存在を確かめるため，以下のような実験を行った．実験のタスクにおいて 6 つの領域適応が行われるが，それぞれの領域適応のための機械学習で用いるソースデータを $\mathbf{x}^{(\mathrm{S})} = \{x_1, x_2, ..., x_n\}$ とする．このソースデータに対して，任意の i 番目のデータ x_i1 つを取り除いた新たなソースデータを $\mathbf{x}^{(\mathrm{S})}{}_i$ とする．ここで新たなソースデータ $\mathbf{x}^{(\mathrm{S})}{}_i$

[4]語義は岩波国語辞書がもとになっている．そこでの中分類までを対象にした．また「入る」は辞書上の語義が 3 つだが，OC や PB では 4 つの語義がある．これは SemEval-2 の日本語 WSD タスクでは新語義のタグも許しているからである．

で学習を行った場合に，元のソースデータ $\mathbf{x}^{(S)}$ で学習を行った場合よりも分類器の精度が向上したならば，$\mathbf{x}^{(S)}$ に含まれるデータ x_i は学習の精度を下げる Misleading データだったと考えられる．ソースデータ $\mathbf{x}^{(S)}$ に対して ${\mathbf{x}^{(S)}}_1$ から ${\mathbf{x}^{(S)}}_n$ までの新しい n 個のソースデータを作成，学習し，データ $\mathbf{x}^{(S)}$ に含まれる n 個全てのデータ点 x_i が Misleading データであるかどうかをそれぞれ 1 つずつ判別する．このようにして判別された Misleading データの集合を，その領域適応における検出すべき Misleading データの正解集合とした．結果を表 2 に示す.

表 2: 検出した Misleading データ

単語	OC		PB		PN	
	PB	PN	OC	PN	OC	PB
言う	159/666	158/666	127/1114	75/1114	82/363	35/363
	23.87	23.72	11.40	6.730	22.59	9.640
入れる	6/73	28/73	19/56	15/56	3/32	1/32
	8.220	38.36	33.93	26.79	9.380	3.130
書く	21/99	39/99	0/62	2/62	12/27	15/27
	21.21	39.40	-	3.230	44.44	55.56
聞く	26/124	21/124	26/123	0/123	4/52	27/52
	20.97	16.94	21.14	-	7.700	51.92
子供	5/77	0/77	12/93	1/93	12/29	13/29
	6.490	-	12.90	1.080	41.38	44.83
時間	1/53	8/53	0/74	0/74	0/59	5/59
	1.890	15.09	-	-	-	8.470
自分	13/128	25/128	0/308	0/308	0/71	1/71
	10.16	19.53	-	-	-	1.410
出る	14/131	10/131	39/152	32/152	22/89	10/89
	10.69	7.630	25.66	21.05	24.72	11.24
取る	6/61	5/61	10/81	18/81	12/43	22/43
	9.840	8.200	12.35	22.22	27.91	51.16
場合	0/126	0/126	7/137	13/137	14/73	9/73
	-	-	5.110	9.490	19.18	12.33
入る	36/68	11/68	38/118	27/118	27/65	42/65
	52.94	16.18	32.20	22.88	41.54	64.62
前	8/105	5/105	10/160	1/160	15/106	2/106
	7.620	4.760	6.250	0.625	14.15	1.890
見る	10/262	3/262	3/273	12/273	8/87	28/87
	38.18	1.150	1.100	4.400	9.200	32.18
持つ	8/62	0/62	2/153	11/153	1/59	1/59
	12.90	-	1.310	7.190	1.690	1.690
やる	0/117	0/117	0/156	0/156	0/27	0/27
	-	-	-	-	-	-
ゆく	17/219	0/219	15/133	1/133	3/27	3/27
	7.760	-	11.29	0.752	11.11	11.11

3.2 外れ値検出手法による検出評価

16 単語の 6 つの領域適応において，提案手法により 3 手法の外れ値の計算を行った．LOF では算出した値を正規化し，閾値 $\theta = 1.96$ より大きな値を Misleading データとみなした．Erk, 密度比の手法はそれぞれ Misleading データの検出正答率の平均が大きくなるような閾値を探し，結果，Erk の手法では閾値 $\theta = 1.9$ より大きな値を，密度比の手法では閾値 $\theta = 0.005$ より小さな値を Misleading データとした．

これらの提案手法による Misleading データの検出正答率を，6 領域適応ごとに 16 単語の平均を取った．これを表 3 に示す．検出正答率は検出した Misleading データの数でそのうちの Misleading データの正答集合に含まれる数を割ったものである．いずれの手法，領域適応においても検出正答率は著しく低い．

表 3: 提案手法による検出正答率 (%)

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	3.590	9.920	14.36	19.07	10.64	29.32	14.48
LOF	3.67	10.19	6.990	5.260	18.48	19.62	10.70
密度比	10.49	10.03	8.880	11.93	17.02	22.10	13.41

また，提案手法による misleading データを除いた場合の領域適応の正答率を表 4 に示す．LOF の手法のみが僅かに通常の領域適応の正答率を上回った．

表 4: 提案手法で misleading を除いた場合の正答率 (%)

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	70.91	68.27	69.45	75.38	69.45	73.33	70.10
LOF	70.71	67.11	70.09	75.57	68.70	73.44	70.94
密度比	69.63	67.34	68.38	76.37	59.72	66.42	67.98
NORMAL	70.77	66.96	70.29	75.56	68.49	73.26	70.89

4 考察

4.1 最近傍距離と Misleading データとの相関

外れ値検出の手法では Misleading データの検出能力が極めて低いと言える．外れ値検出はいずれの手法もデータ点の距離の差異を利用するものであるが，Misleading データと非 Misleading データの間にはその差異が認められなかったと考えられる．そこで，Misleading データと非 Misleading データについて，以下を調べる．

- ターゲットデータに対する最近傍点への距離についての差異の有無

実験で使用した 16 単語 6 種計 96 ケースの領域適応のデータについて，外れ値検出同様にそれぞれのソースデータ点のターゲットデータへの最近傍距離を算出した．これを Misleading データの正答集合により，Misleading データと非 Misleading データに分けて平均を取り，それらの相関を t 検定により判定した．なお，有意水準は 5% とした．その結果，有意差が認められたケースは 96 ケース中 12 ケースであった．また，その有意差の有無と外れ値の 3 手法による Misleading データの検出結

果を照らし合わせたところ，有意差が認められた 12 ケースの領域適応において Misleading データの検出率が特段高い訳でもなかった．Misleading データはターゲットデータへの最近傍距離において非 Misleading データとの差異があるとは言えず，そのため，距離を利用する外れ値検出の手法をもって Misleading データを判別することは難しい．

4.2　負の転移が生じない対象単語の検出

外れ値検出手法を用いても，ここで設定したような Misleading データを検出することは困難であった．本節では Misleading データと関連の深い負の転移の有無を外れ値検出手法で判定できるかを調べてみる．

まず論文 (新納浩幸・佐々木稔 (2014)) では本論文と同じデータを利用して，負の転移が生じなかった対象単語を選出している．その結果は以下の表 5 にまとめられる．表 5 でチェックが付いているものが，負の転移が生じなかった単語である．

表 5: 負の転移が生じていない領域適応

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う		✓	✓	✓	✓	
入れる		✓	✓	✓	✓	✓
書く	✓			✓	✓	
聞く	✓					
子供			✓		✓	
時間	✓		✓		✓	
自分	✓	✓				
出る				✓		✓
取る			✓		✓	✓
場合		✓	✓		✓	✓
入る		✓	✓		✓	✓
前		✓				
見る	✓					
持つ	✓	✓				✓
やる		✓		✓		✓
ゆく		✓		✓	✓	

次に本論文で行った外れ値検出手法で検出された Misleading データの割合が，全体のデータの 1 割以下である場合に，負の転移が生じないという判定を行う．これによって外れ値検出手法を利用して，負の転移が生じない対象単語の検出評価を行うことができる．検出の正解率，再現率，F 値をそれぞれ表 6，表 7，表 8 に示す．Mislead は本実験で用いた Misleading の正解データを利用した検出を示す．

表 6: 外れ値検出手法を利用した負の転移が生じない単語の検出・正解率

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	0.286	0.333	0.500	0.533	0.250	0.615	0.420
LOF	0.375	0.375	0.438	0.563	0.438	0.563	0.458
密度比	0.000	1.000	0.000	0.500	0.000	0.000	0.250
Mislead	0.125	0.333	0.375	0.583	0.286	0.429	0.355

表 7: 外れ値検出手法を利用した負の転移が生じない単語の検出・再現率

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	0.667	0.833	1.000	0.889	0.429	0.889	0.784
LOF	1.000	1.000	1.000	1.000	1.000	1.000	1.000
密度比	0.000	0.333	0.000	0.556	0.000	0.000	0.148
Mislead	0.167	0.500	0.429	0.778	0.286	0.333	0.415

表 8: 外れ値検出手法を利用した負の転移が生じない単語の検出・F 値

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	0.400	0.476	0.667	0.667	0.316	0.727	0.542
LOF	0.545	0.545	0.609	0.720	0.609	0.720	0.625
密度比	-	0.500	-	0.526	-	-	0.513
Mislead	0.143	0.400	0.400	0.667	0.286	0.375	0.378

表 8 を見ると，Misleading の正解データを用いても負の転移のない単語を検出する能力は高くない．それに比較すれば外れ値検出手法を利用した場合の検出する能力は高い．外れ値検出手法を利用して負の転移が生じない単語を判定できる可能性もあり，この点で精度改善が可能であると考える．

5　おわりに

本論文では WSD の領域適応における Misleading データの検出に外れ値検出手法を利用することを試みた．総当たり的に各データが Misleading データと見なせるかどうかを調べることで，Misleading データの存在を確認できた．また，それらを正解集合として外れ値検出を用いた検出能力も調べた．結論的には外れ値検出手法を利用しても，本論文で設定したような Misleading データの検出は困難であることがわかった．ただし Misleading データと関連の深い負の転移現象の有無を判定することには利用可能だと考えている．今後は外れ値ではない Misleading データの特徴を調査することで，新たな Misleading データの検出法を考えたい．

文献

Markus M. Breuning, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander (2000) “LOF: Identifying Density-Based Local Outliers,” in *ACM SIGMOD 2000*, pp. 93–104.

Katrin Erk (2006) “Unknown Word Sense Detection As Outlier Detection,” in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 128–135.

Jing Jiang and Chengxiang Zhai (2007) “Instance weighting for domain adaptation in NLP,” in *Proc. of ACL-2007*, pp. 264–271.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2010) “SemEval-2010 Task: Japanese WSD,” in *Proc. of the 5th International Workshop on Semantic Evaluation*, pp. 69–74.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich (2005) “To transfer or not to transfer,” in *Proc. of the NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*.

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

高村大也 (2010) 言語処理のための機械学習入門，コロナ社.

新納浩幸、佐々木稔 (2014) 「共変量シフトの問題としての語義曖昧性解消の領域適応」，自然言語処理，第 21 巻，第 1 号，(to appear).

神嶌敏弘 (2010) 「転移学習」，人工知能学会誌，第 25 巻，第 4 号，pp.572–580.

領域間距離を利用した能動学習による語義曖昧性解消の領域適応

小野寺 喜行 (茨城大学 工学部 情報工学科) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Domain Adaptation for Word Sense Disambiguation by Active Learning Using the Distance between Domains

Yoshiyuki Onodera (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応の問題に対して，データ選択に領域間距離を利用した能動学習手法を提案する．

自然言語処理のタスクにおいて帰納学習手法を用いる際，訓練データとテストデータは同じ領域のコーパスから得ていることが通常である．ただし実際には異なる領域である場合も存在する．そこである領域（ソース領域）の訓練データから学習された分類器を，別の領域（ターゲット領域）のテストデータに合うようにチューニングすることを領域適応という (Sogaard (2013))[3]．

領域適応の問題をターゲット領域のラベル付きデータの不足からくる問題として見なせば，能動学習 (Settles (2010)) や半教師あり学習 (Chapelle et al. (2006)) を利用することは有効である．ここでは WSD の領域適応の問題に対して，能動学習を利用する．一般に能動学習はラベルなしデータの集合から学習効果の高いデータを選択し，そのデータにラベルを付けて訓練データに追加することで，徐々に分類器の精度を高めてゆく．能動学習のポイントはどのようにして学習効果の高いデータを選択するかである．通常は WSD の各対象単語毎に識別の信頼度を基準にしてその選択が行われる．つまり WSD を対象とした能動学習では，各対象単語毎に能動学習を適用する形になっている．ただしこの形では WSD の総合的な評価であるマクロ平均やマイクロ平均の評価値を効率的に向上させることができる保証はない．そこで本論文では WSD の全対象単語の全テストデータを対象にして，ラベル付けするデータを選択することを試みる．また，選択する際の信頼度は，領域間の関係に依存していると考えられる．つまり領域間の距離が大きければ，算出した信頼度よりも更に信頼度は低いし，領域間の距離が小さければ，算出した信頼度は妥当と考えられる．このように領域間距離を考慮した信頼度を用いることで更なる精度向上を図る．

実験では現代日本語書き言葉均衡コーパス（BCCWJ コーパス (Maekawa (2007))）における 3 つの領域 OC (Yahoo! 知恵袋)，PB (書籍) 及び PN (新聞) を利用する．SemEval-2 の日本語 WSD タスク (Okumura et al. (2010)) ではこれらのコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．すべての領域である程度の頻度が存在する多義語 16 単語を対象にして，WSD の領域適応の実験を行う．領域適応としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りが存在する．結果 $16 \times 6 = 96$ 通りの WSD の領域適応の問題に対して能動学習を利用した WSD の領域適応の実験を行い提案手法の有効性を示す．

2　領域間距離を利用した能動学習

2.1　全対象単語に対する能動学習

通常，能動学習では図 1 に示す Schohn の手法の様に，WSD の各対象単語毎に識別の信頼度を基準にしてその選択が行われる．つまり WSD を対象とした能動学習では，各対象単語毎に能動学習を適用する形になっている．ただしこの形では WSD の総合的な評価であるマクロ平均やマイクロ平均の評価値を効率的に向上させることができる保証はない．

[2] 10t4019s@hcs.ibaraki.ac.jp
[2] shinnou@mx.ibaraki.ac.jp
[3] 領域適応は機械学習の分野では転移学習 (神嶌敏弘 (2010)) の一種と見なされている．

そこで WSD の全対象単語の全テストデータを対象にして，その中から信頼度が最も低いデータをラベル付けするデータとして選択することを試みる．

(1) ラベル付きデータから分類器を作成する．

(2) 作成した分類器によりラベルなしデータを識別する．このとき識別の信頼度も求める．

(3) 識別の信頼度が最も低いデータにラベルを付け，ラベル付きデータに追加する．

(4) (1) に戻る

図 1: 能動学習の手順 (Schohn の手法)

2.2 領域間距離を用いた能動学習

ラベル付けするデータを選択する際の信頼度は，領域間の関係に依存していると考えられる．つまり領域間の距離が大きければ，算出した信頼度よりも更に信頼度は低いし，領域間の距離が小さければ，算出した信頼度は妥当と考えられる．そこで，領域間距離を考慮した信頼度を用いることで更なる精度向上を図る．

ここでは，確率密度比から求めた重みを領域間類似度と考え領域間距離とし，能動学習においてラベル付けするデータの選択に識別の信頼度と領域間距離から導き出した評価値を用いることにする．WSD の全対象単語の全テストデータを対象にして次に示すように評価値を求め，その値が最も低いデータをラベル付けするデータとして選択する．

まず，分類器によって識別した n 個のラベルなしデータ $x_j\,(j = 1, 2, \ldots, n)$ に対する識別の信頼度の集合 $h_j\,(j = 1, 2, \ldots, n)$ の平均値 m と標本分散値 u^2 を以下の式で求める．

$$m = \frac{\sum_{j=1}^{n} h_j}{n},\ u^2 = \frac{\sum_{j=1}^{n} (m - h_j)^2}{(n-1)}$$

次に，領域間類似度の集合 $s_j\,(j = 1, 2, \ldots, n)$ から同様にして平均値 m' と標本分散値 u'^2 を求める．

$$m' = \frac{\sum_{j=1}^{n} s_j}{n},\ u'^2 = \frac{\sum_{j=1}^{n} (m - s_j)^2}{(n-1)}$$

そして各データ x_j について以下の値を求める．

$$r_j = \frac{(h_j - m)}{u},\ w_j = \frac{u \times (s_j - m')}{u'} + m$$

この r_j と w_j の値を合計したものがデータ x_j の評価値となる．

3 実験

実験では現代日本語書き言葉均衡コーパス（BCCWJ コーパス (Maekawa (2007))）における 3 つの領域 OC (Yahoo! 知恵袋)，PB (書籍) 及び PN (新聞) を利用する．SemEval-2 の日本語 WSD タスク (Okumura et al. (2010)) ではこれらのコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．すべての領域である程度の頻度が存在する多義語 16 単語 (表 1) を対象にして，WSD の領域適応の実験を行う．

表 1: 実験対象単語

単語	辞書上の語義数	OC		PB		PN	
		頻度	語義数	頻度	語義数	頻度	語義数
言う	3	666	2	1114	2	363	2
入れる	3	73	2	56	3	32	2
書く	2	99	2	62	2	27	2
聞く	3	124	2	123	2	52	2
子供	2	77	2	93	2	29	2
時間	4	53	2	74	2	59	2
自分	2	128	2	308	2	71	2
出る	3	131	3	152	3	89	3
取る	8	61	7	81	7	43	7
場合	2	126	2	137	2	73	2
入る	3	68	4	118	4	65	3
前	3	105	3	160	2	106	4
見る	6	262	5	273	6	87	3
持つ	4	62	4	153	3	59	3
やる	5	117	3	156	4	27	2
ゆく	2	219	2	133	2	27	2
平均	3.35	193.9	2.94	150.6	2.88	75.56	2.69

領域適応としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りが存在する．結果 $16 \times 6 = 96$ 通りの WSD の領域適応の問題に対して通常の能動学習と 2 つの提案手法,「全対象単語に対する能動学習」と「領域間距離を用いた能動学習」による識別の平均正解率で比較する．なお，ラベル付けするデータの数は通常の手法は各単語 5 点ずつ計 80 点，提案手法は全体で 80 点とした．加えて，ランダムにラベル付けするデータを選択する場合も試す．こちらは各単語 5 点ずつ選択する場合と全体から 80 点選択する場合をそれぞれ 5 回行いその平均を求める．分類器は SVM を用い，SVM ツールには libsvm[4] を使用した．libsvm では `-b` オプションにより識別の信頼度が求められる．

実験結果を表 2(マクロ平均) と表 3(マイクロ平均) に示す．結果として,「全対象単語に対する能動学習」は通常の手法よりも主にマクロ平均において良い正解率となり識別精度の向上に効果が有ると言える．また,「領域間距離を用いた能動学習」は低い正解率となり精度の向上に効果は見られなかった．

表 2: 実験結果 (マクロ平均 (%))

	通常の能動学習	ランダムに選択 (各単語 5 点ずつ)	ランダムに選択 (全体から 80 点)	全対象単語に対する能動学習	領域間距離を用いた能動学習
OC → PB	74.80	73.55	72.40	75.49	71.76
PB → PN	79.74	72.29	72.40	81.14	76.94
PN → OC	73.39	71.23	69.89	74.38	68.43
OC → PN	75.23	71.59	69.33	73.86	70.47
PN → PB	78.26	76.40	72.09	79.55	77.33
PB → OC	76.22	74.30	72.09	77.03	72.04

[4] `http://www.csie.ntu.edu.tw/ cjlin/libsvm/`

表 3: 実験結果 (マイクロ平均 (%))

	通常の能動学習	ランダムに選択 (各単語 5 点ずつ)	ランダムに選択 (全体から 80 点)	全対象単語に対する能動学習	領域間距離を用いた能動学習
OC → PB	78.89	77.85	78.33	78.45	77.36
PB → PN	83.37	82.47	78.33	84.62	83.13
PN → OC	74.82	73.94	74.85	74.99	72.88
OC → PN	75.60	74.80	77.49	67.00	65.26
PN → PB	82.09	81.02	74.69	81.65	83.03
PB → OC	78.70	76.35	74.69	78.20	74.57

4 考察

4.1 領域間距離の利用について

もし領域間距離を正しく測れている場合，負の転移の有無を判定できる．負の転移が生じているかどうかを調べるために以下の実験を行った．

単語 w_i についてソース領域 S からターゲット領域 T への領域適応の実験を行う．まずターゲット領域 T のラベル付きデータをランダムに 15 個取り出し，残りを評価データとする．つまり利用できる訓練データはソース領域 S のラベル付きデータとターゲット領域 T からランダムに取り出した 15 個のラベル付きデータとなる．この訓練データを用いて手法 A により分類器を作成し，先の評価データの語義識別の正解率 $P_{i,k}$ を測る．この実験を 5 回行い $P_{i,1}, P_{i,2}, \cdots, P_{i,5}$ を得る．それらの平均 P_i を「単語 w_i の S から T への領域適応における手法 A の平均正解率」とする．上記の単語 w_i を 16 種類の各対象単語 $w_1, w_2, \cdots, w_{16}$ に変えることで，16 個の平均正解率 $P_1, P_2, \cdots, P_{16}$ が得られる．それらの平均 P を「S から T への領域適応における手法 A の平均正解率」とする．

上記の手法 A としては，以下の 3 種類を試す．(1) ソース領域のラベル付きデータのみを用いる手法 (ターゲット領域の 15 個のラベル付きデータの重みを 0 とする手法)（S-Only），(2) ターゲット領域からランダムに取り出した 15 個のラベル付きデータのみを用いる手法 (ソース領域のラベル付きデータの重みを 0 とする手法)（T-Only），(3) ソース領域のラベル付きデータとターゲット領域の 15 個のラベル付きデータを用いる手法（S+T）．

S から T への領域適応における各手法の平均正解率を表 4 に示す．

表 4: 各手法の平均正解率

領域適応	S-Only	T-only	S+T
OC → PB	0.7137	0.7559	0.7511
PB → PN	0.7678	0.7206	0.7801
PN → OC	0.6926	0.7716	0.7630
OC → PN	0.6829	0.7300	0.7324
PN → PB	0.7543	0.7561	0.7863
PB → OC	0.6988	0.7766	0.7533
平均	0.7184	0.7518	0.7611

負の転移が生じているかどうかの判定には，上記の実験でより得られた `T-Only`，`S-Only` 及び `S+T` の正解率を利用する．もしも正解率で以下の関係が成立しているなら，負の転移が生じていないと考えられる．

T-Only, S-Only < S+T

結果を表 5 に示す．チェックがつけられた箇所が負の転移が生じていない領域適応の問題である．96 種類の領域適応の問題の中で 44 種類において負の転移が生じていない．

表 5: 負の転移が生じていない領域適応

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う		✓	✓	✓	✓	
入れる		✓	✓	✓	✓	✓
書く	✓			✓	✓	
聞く	✓					
子供			✓		✓	
時間	✓		✓		✓	
自分	✓	✓				
出る				✓		✓
取る			✓		✓	✓
場合		✓	✓		✓	✓
入る		✓	✓		✓	✓
前		✓				
見る	✓					
持つ	✓	✓				✓
やる		✓		✓		✓
ゆく		✓		✓	✓	

この表をもとに，負の転移が生じていない a 個の単語については従来手法と同様の手順で各単語 5 点ずつラベル付きデータを追加し，負の転移が生じている 16-a 個の単語はその中で信頼度最低のものをラベル付きデータに追加を計 (16-A)*5 回行う手法を試す．

実験結果を表 6(マクロ平均) と表 7(マイクロ平均) に示す．通常の能動学習よりも良い結果を得ることが出来た．よって，領域間距離の情報を用いてラベル付けするデータを選択する手順を制御することは効果が有ると言える．

提案手法の平均正解率が低い原因は領域間距離を利用すること自体ではなく，「確率密度比から領域間距離を測る」という方法が正しく領域間距離を測ることが出来なかった事や，領域間距離をラベル付けするデータ選択に上手く取り入れられなかった事が考えられる．したがって，今後は領域間距離を正しく測る方法と領域間距離をラベル付けするデータ選択に取り入れる方法を考えていきたい．

表 6: 実験結果 (マクロ平均 (%))

	通常の能動学習	ランダム (各単語 5 点)	ランダム (全体から)	全対象単語に対する	領域間距離を利用	負の転移を考慮
OC → PB	74.80	73.55	72.40	75.49	71.76	75.92
PB → PN	79.74	72.29	72.40	81.14	76.94	80.71
PN → OC	73.39	71.23	69.89	74.38	68.43	74.17
OC → PN	75.23	71.59	69.33	73.86	70.47	75.01
PN → PB	78.26	76.40	72.09	79.55	77.33	78.08
PB → OC	76.22	74.30	72.09	77.03	72.04	75.66

表 7: 実験結果 (マイクロ平均 (%))

	通常の能動学習	ランダム (各単語 5 点)	ランダム (全体から)	全対象単語に対する	領域間距離を利用	負の転移を考慮
OC → PB	78.89	77.85	78.33	78.45	77.36	78.86
PB → PN	83.37	82.47	78.33	84.62	83.13	83.95
PN → OC	74.82	73.94	74.85	74.99	72.88	75.66
OC → PN	75.60	74.80	77.49	67.00	65.26	76.18
PN → PB	82.09	81.02	74.69	81.65	83.03	81.99
PB → OC	78.70	76.35	74.69	78.20	74.57	78.95

5 おわりに

本論文では WSD の領域適応の問題に対して能動学習を試みた．WSD を対象とした能動学習では，各対象単語毎に能動学習を適用する形になっているが，この形では WSD の総合的な評価であるマクロ平均やマイクロ平均の評価値を効率的に向上させることができる保証はない．

そこで，WSD の全対象単語の全テストデータを対象にしてラベル付けするデータを選択する手法と，領域間距離を考慮した信頼度を用いる手法を試した．BCCWJ コーパスの OC(Yahoo!知恵袋)，PB(書籍) 及び PN(新聞) の 3 つの領域を用いた実験の結果，全対象単語から選択する手法は効果があったが，領域間距離を考慮した手法には効果は見られなかった．

そこで，仮に領域間距離を正しく測れているとして負の転移の有無を判定し，その判定を用いてデータ選択の手順を変える手法を試した結果，領域間距離を利用するという手法には効果が有ることが確認できた．

提案手法を効果的に利用するには，今後は領域間距離を正しく測る方法と領域間距離をラベル付けするデータ選択に取り入れる方法を考えることが今後の課題である．

文献

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien et al. (2006) *Semi-supervised learning*, Vol. 2: MIT press Cambridge.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2010) “SemEval-2010 Task: Japanese WSD,” in *The 5th International Workshop on Semantic Evaluation*, pp. 69–74.

Burr Settles (2010) “Active learning literature survey,” *University of Wisconsin, Madison*.

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

神嶌敏弘 (2010) 「転移学習」，人工知能学会誌，第 25 巻，第 4 号，pp.572–580.

uLSIF による重み付き学習を利用した語義曖昧性解消の領域適応

菊池 裕紀 (茨城大学大学院 情報工学専攻) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Domain Adaptation for Word Sense Disambiguation by Weighted Learning using uLSIF

Hironori Kikuchi (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では，語義曖昧性解消（Word Sense Disambiguation, WSD）をタスクとした領域適応の問題を共変量シフトの問題であるとみなし，その問題の解法である確率密度比を重みとしたパラメータ学習により WSD の領域適応の問題解決を図る．確率密度比の算出には，Naive Bayes のモデルを利用して算出したものと uLSIF によって算出したものを用いて比較実験を行う．

自然言語処理の多くのタスクにおいて帰納学習手法が利用される．そこではコーパス A からタスクに応じた訓練データを作成し，その訓練データから分類器を学習する．そしてこの分類器を利用することでタスクの解決を図っていく．このとき，実際のタスクとなるデータはコーパス A とは領域が異なるコーパス B 内のものであることがしばしば起こる．この場合，コーパス A（ソースドメイン）から学習された分類器では，コーパス B（ターゲットドメイン）のデータを精度よく解析できない問題が生じる．これが領域適応の問題であり，近年活発に研究が行われている (Sogaard (2013))．

WSD は文 $\boldsymbol{x}$ 内の多義語 m の語義 $c \in C$ を識別する問題である．事後確率最大化に基づければ，$\arg\max_{c \in C} P(c|\boldsymbol{x})$ を解く問題といえる．例えば，単語 $m =$「ボタン」には少なくとも (c_1) 服のボタン，(c_2) スイッチのボタン，(c_3) 花のボタン（牡丹），の 3 つの語義がある．そして文 $\boldsymbol{x} =$「シャツのボタンが取れた」が与えられたときに，文中の「ボタン」が $C = \{c_1, c_2, c_3\}$ 内のどれかを識別する．通常は教師付き学習手法を用いて $P(c|\boldsymbol{x})$ を推定して解くことになる．しかし，WSD の領域適応の問題は，前述したように，学習もとのソースドメインのコーパスと，分類器の適用先であるターゲットドメインのコーパスが異なる問題である．領域適応ではソースドメイン S から S 上の確率密度 $P_s(c|\boldsymbol{x})$ は学習できるという設定なので，$P_s(c|\boldsymbol{x})$ やその他の情報を利用して，ターゲットドメイン T 上の確率密度 $P_t(c|\boldsymbol{x})$ を推定できれば良い．ここで「シャツのボタンが取れた」という文中の「ボタン」の語義は，この文がどんな領域のコーパスに現れても変化するとは考えられない．つまり $P(c|\boldsymbol{x})$ は領域に依存していないと考えられる．つまり $P_s(c|\boldsymbol{x}) = P_t(c|\boldsymbol{x})$ が成立している．今 $P_s(c|\boldsymbol{x})$ は学習できるので，$P_s(c|\boldsymbol{x}) = P_t(c|\boldsymbol{x})$ が成立していれば，$P_t(c|\boldsymbol{x})$ を推定する必要はないように見える．ただしソースドメインだけを使って推定した $P_s(c|\boldsymbol{x})$ では，実際の識別精度は低い場合が多い．それは $P_s(\boldsymbol{x}) \neq P_t(\boldsymbol{x})$ から生じている．$P_s(c|\boldsymbol{x}) = P_t(c|\boldsymbol{x})$ だが $P_s(\boldsymbol{x}) \neq P_t(\boldsymbol{x})$ という仮定の下で，$P_t(c|\boldsymbol{x})$ を推定する問題は共変量シフトの問題 (Shimodaira (2000) 杉山将 (2006)) である．

訓練データを $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ とする．共変量シフトでは $P_t(y|\boldsymbol{x})$ に確率モデル $P(y|\boldsymbol{x};\theta)$ を設定する．次に確率密度比 $w(\boldsymbol{x}_i) = P_t(\boldsymbol{x}_i)/P_s(\boldsymbol{x}_i)$ を重みにした以下の対数尤度を最大にする θ を求めることで，$P_t(y|\boldsymbol{x})$ を構築する．

$$\sum_{i=1}^{N} w(\boldsymbol{x_i}) \log P(y_i|\boldsymbol{x_i};\theta)$$

実験では BCCWJ コーパス Maekawa (2007) の 3 つ領域 OC (Yahoo! 知恵袋) PB (書籍) と PN（新聞） から共に頻度が 50 以上の多義語 16 単語を対象にして，WSD の領域適応の実験を行い，前

[2]13nm705g@hcs.ibaraki.ac.jp
[2]shinnou@mx.ibaraki.ac.jp

述した共変量シフトの問題として WSD の領域適応の問題を解いた．$P_t(y|\boldsymbol{x})$ の確率モデルとしては Naive Bayes（NB） のモデルを用いた．

2　関連研究

自然言語処理に対する領域適応は，帰納学習手法を利用する全てのタスクで生じる問題である．そのため，その研究は多岐に渡っている．その中でも現在注目しているのは，WSD をタスクとした領域適応の問題が共変量シフトの問題と同値であるとするものである．この場合，WSD の領域適応の問題は，共変量シフトの解決策である確率密度比を重みとしたパラメータ学習により解消できると考えられる．実際，この解決手法は古くから扱われている (Shimodaira (2000))(杉山将 (2006))．

教師付き領域適応手法では標準手法とも言える Daumé(Daumé III, Hal (2007)) の手法も重み付けの手法である．ただしこれは事例への重みではなく，素性への重みである．そこではソースドメインの訓練データのベクトル x_s を $(x_s, x_s, 0)$ と連結した 3 倍の長さのベクトルになるよう形成し，同様にターゲットドメインの訓練データのベクトル x_t を 3 倍の長さのベクトル $(0, x_t, x_t)$ と形成する．このベクトルを用いて，通常の分類問題として解いていく．この手法は非常に簡単でありながら，効果が高い手法として知られている．この手法は，ソースドメインとターゲットドメインに共通している特徴の重みが付くことで領域適応に効果が出ていると考えられる．

また，領域適応の問題を共変量シフト下の学習として解決する研究としては，Jiang の研究 (Jiang and Zhai (2007)) と齋木の研究 (Yosuke Saiki and Okumura (2008)) が挙げられる．Jiang は密度比を手動で調整し，モデルにはロジスティック回帰を用いている．また齋木は $P(x)$ を unigram でモデル化することで密度比を推定し，モデルには最大エントロピー法のモデルを用いている．ただしどちらの研究もタスクに WSD を扱っていない．更に確率密度比 $w(\boldsymbol{x})$ の算出には結局 $P_s(\boldsymbol{x})$ と $P_t(\boldsymbol{x})$ をそれぞれ求めてその比を取る手法になっており，本論文利用した uLSIF のように直接 $w(\boldsymbol{x})$ を推定する手法は用いられていない．

3　確率密度比

確率密度比 $w(\boldsymbol{x}) = P_t(\boldsymbol{x})/P_s(\boldsymbol{x})$ の算出には大きく 2 つの方法がある．1 つは $P_t(\boldsymbol{x})$ と $P_s(\boldsymbol{x})$ を各々求め，その比をとる方法であり，もう 1 つは直接 $w(\boldsymbol{x})$ をモデル化して推定する方法である．ここでは前者として論文 (新納浩幸・佐々木稔 (2014)) で提案された簡易な手法を利用する．また後者として uLSIF を利用する．

3.1　NaiveBayes による確率密度比の推定

対象単語 m の用例 $\boldsymbol{x}$ の素性リストを $\{f_1, f_2, \cdots, f_n\}$ とする．求めるのは領域 $R \in \{S, T\}$ 上の $P_R(\boldsymbol{x})$ である．ここでは以下で仮定される NB のモデルを用いて算出する．

$$P_R(\boldsymbol{x}) = \prod_{i=1}^{n} P_R(f_i)$$

領域 R のコーパス内の m の全ての用例について素性リストを作成しておく．ここで用例の数を $N(R)$ とおく．また $N(R)$ 個の用例の中で，素性 f が現れた用例数を $n(R, f)$ とおく．MAP 推定でスムージングを行い，$P_R(f)$ を以下で定義する．高村大也 (2010)

$$P_R(f) = \frac{n(R, f) + 1}{N(R) + 2}$$

以上で求められた $P_s(\boldsymbol{x})$ と $P_t(\boldsymbol{x})$ の比を求めることで $\widehat{w}(\boldsymbol{x})$ を求める．

$$\widehat{w}(\boldsymbol{x}) = \frac{P_t(\boldsymbol{x})}{P_s(\boldsymbol{x})}$$

3.2 uLSIF による確率密度比の算出

uLSIF は Kanamori ら (Kanamori et al. (2009)) により考案された手法である．$w(\boldsymbol{x})$ を線形モデルで直接推定することができる．線型方程式を解くことによってモデルのパラメータを推定できるため高速な算出が可能となっている．

$w(\boldsymbol{x})$ を以下のモデルで定式化する．

$$\widehat{w} = \sum_{l=1}^{b} \boldsymbol{\alpha_l} \boldsymbol{\varphi_l}(\boldsymbol{x}) \tag{1}$$

ここで，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_b)^T$ はサンプルデータから学習されるパラメータを，$\varphi_l(\boldsymbol{x})$ は，以下のような基底関数を表す．

$$\varphi_l(x) >= 0 \ \forall x \in D \ \forall l = 1, 2, \cdots, b \tag{2}$$

b は $R \in \{S, T\}$ のデータ数により決定する．また，$\varphi_l(\boldsymbol{x})$ には以下のガウスカーネルを用いることとする．

$$K_\sigma(\boldsymbol{x}, \boldsymbol{x'}) = \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{x'}\|^2}{2\sigma^2}\Big) \tag{3}$$

よって，式 (1) は，以下で表すことができる．

$$\widehat{w}(x) = \sum_{l=1}^{n_T} \boldsymbol{\alpha}_l K_\sigma(x, x_l^T) \tag{4}$$

次に，パラメータ $\boldsymbol{\alpha}_l$ の求め方を定義する．パラメータ $\boldsymbol{\alpha}_l$ は，次の J_0 を最小にするような二乗誤差を求めればよい．

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2}\int (\widehat{w}(x) - w(x))^2 p_S(x) dx \\ &= \frac{1}{2}\int \widehat{w}(x)^2 P_s(\boldsymbol{x}) dx - \int \widehat{w}(x) w(x) P_s(\boldsymbol{x}) dx + \frac{1}{2}\int w(x)^2 P_s(\boldsymbol{x}) dx \\ &= \frac{1}{2}\int \widehat{w}(x)^2 P_s(\boldsymbol{x}) dx - \int \widehat{w}(x) P_t(\boldsymbol{x}) dx + \frac{1}{2}\int w(x)^2 P_s(\boldsymbol{x}) dx \end{aligned} \tag{5}$$

なお，式 (5) の最終項は一定の値をとるため無視することができる．よって最終的に以下の式になる．

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \frac{1}{2}\int \widehat{w}(x)^2 P_s(\boldsymbol{x}) dx - \int \widehat{w}(x) P_t(\boldsymbol{x}) dx \\ &= \frac{1}{2}\sum_{l,l'=1}^{b} \boldsymbol{\alpha}_l \boldsymbol{\alpha}_{l'} \Big(\int \varphi_l(\boldsymbol{x}) \varphi_{l'}(\boldsymbol{x}) P_s(\boldsymbol{x}) dx\Big) - \sum_{l=1}^{b} \boldsymbol{\alpha}_l \Big(\int \varphi_l(\boldsymbol{x}) P_t(\boldsymbol{x}) dx\Big) \\ &= \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{H} \alpha - \boldsymbol{h}^{\mathrm{T}} \alpha \end{aligned} \tag{6}$$

$\boldsymbol{H}$ は，$b * b$ の行列，

$$\boldsymbol{H}_{l,l'} = \int \varphi_l(\boldsymbol{x}) \varphi_{l'}(\boldsymbol{x}) P_s(\boldsymbol{x}) dx \tag{7}$$

$\boldsymbol{h}$ は，b 次元のベクトル，

$$\boldsymbol{h}_l = \int \varphi_l(\boldsymbol{x}) P_t(\boldsymbol{x}) dx \tag{8}$$

である．このとき，標本平均をとり J の期待値を推定する．

$$\widehat{J}(\boldsymbol{\alpha}) = \frac{1}{2n_s}\sum_{i=1}^{n_s}\widehat{w}(x_i^s)^2 - \sum_{j=1}^{n^t}\widehat{w}(x_j^t) \tag{9}$$

$$= \frac{1}{2}\sum_{l,l'=1}^{b}\boldsymbol{\alpha}_l\boldsymbol{\alpha}_{l'}\big(\frac{1}{n_s}\sum_{i=1}^{n_s}\varphi_l(x_i^s)\varphi_{l'}(x_i^s)\big) - \sum_{l=1}^{b}\boldsymbol{\alpha}_l\big(\frac{1}{n_t}\sum_{j=1}^{n_t}\varphi_l(x_j^t)\big) \tag{10}$$

$$= \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{\alpha} \tag{11}$$

$\widehat{\boldsymbol{H}}$ は $b*b$ の行列，

$$\widehat{\boldsymbol{H}}_{l,l'} = \frac{1}{n_s}\sum_{i=1}^{n_s}\varphi_l(x_i^s)\varphi_{l'}(x_i^s) \tag{12}$$

$\widehat{\boldsymbol{h}}$ は b 次元のベクトル，

$$\widehat{\boldsymbol{h}}_l = \frac{1}{n_t}\sum_{j=1}^{n_t}\varphi_l(x_j^t) \tag{13}$$

である．最後に制約条件を考えない最適化問題の形に整えると以下のような式になる．

$$\min_{\alpha\in R}\Big[\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{\alpha} + \frac{\boldsymbol{\lambda}}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}\Big] \tag{14}$$

この無制約下での凸２次計画問題を解くことで $\boldsymbol{\alpha}$ を得ることができる．

4　実験結果

本実験では，現代日本語書き言葉均衡コーパス（BCCWJ コーパス (Maekawa (2007))）の OC(Yahoo! 知恵袋)，PB(書籍)，PN(新聞) の 3 つの異なる領域を扱い実験を行った．それぞれの領域から共に頻度が 50 以上多義語 16 単語を WSD の対象単語とする．領域適応としては，OC → PB，PB → OC，PB → PN，PN → PB，OC → PN，PN → OC の 6 通りの実験を行う．

分類器は Classias (Okazaki (2009)) を用いて学習した．(OC → PB) の結果を表 1 に，(PB → OC) の結果を表 2 に，(PB → OC) の結果を表 3 に，(PN → PB) の結果を表 4 に，(OC → PN) の結果を表 5 に，(PN → OC) の結果を表 6 に示す．

また，各表の列 Normal は確率密度比の重みを”1”とした時の結果，列 NB は NaiveBayes によって求めた確率密度比を適用した結果，uLSIF は uLSIF によって求めた確率密度比を適用した結果を表すものとする．Normal，NB，uLSIF の 3 つの手法で各単語ごとに領域適応を行い，その識別精度の平均値で評価する．

結果より，uLSIF により推定した確率密度比を適用しても正解率の改善が見込めることはなかった．

表 1: 実験結果 (OC → PB)

単語	Normal	NB	uLSIF
聞く	0.6667	0.6098	0.7642
言う	0.8205	0.7971	0.5835
子供	0.2903	0.7742	0.2473
時間	0.9054	0.9054	0.9054
自分	0.9513	0.8506	0.974
入れる	0.6727	0.5636	0.6364
出る	0.5658	0.6053	0.5592
取る	0.2593	0.2716	0.2716
場合	0.8467	0.854	0.8467
入る	0.4746	0.3729	0.4237
前	0.7438	0.8063	0.6438
見る	0.8493	0.8456	0.8382
持つ	0.7712	0.6993	0.6993
やる	0.9355	0.9355	0.9355
いく	0.8872	0.8647	0.8647
書く	0.7903	0.7903	0.9032
平均	0.7144	0.7216	0.6935

表 2: 実験結果 (PB → OC)

単語	Normal	NB	uLSIF
聞く	0.6774	0.5484	0.6613
言う	0.8363	0.8498	0.5315
子供	0.1558	0.1299	0.1299
時間	0.8302	0.8302	0.8302
自分	0.875	0.875	0.875
入れる	0.726	0.7534	0.7397
出る	0.6641	0.6183	0.687
取る	0.2787	0.3607	0.0984
場合	0.9286	0.9048	0.9762
入る	0.6029	0.5588	0.6324
前	0.8942	0.8846	0.875
見る	0.5611	0.5573	0.5611
持つ	0.8033	0.8197	0.5902
やる	0.9402	0.9402	0.9402
いく	0.6895	0.6119	0.6849
書く	0.7374	0.7374	0.7374
平均	0.7	0.6863	0.6594

表 3: 実験結果 (PB → PN)

単語	Normal	NB	uLSIF
聞く	0.7115	0.6923	0.7115
言う	0.9366	0.9284	0.7218
子供	0.5517	0.5517	0.5172
時間	0.661	0.661	0.661
自分	0.9859	0.9859	0.9859
入れる	0.7188	0.7813	0.7188
出る	0.8427	0.7753	0.7191
取る	0.4186	0.3721	0.2791
場合	0.8493	0.8493	0.8493
入る	0.5692	0.5538	0.2154
前	0.875	0.8558	0.7308
見る	0.7011	0.7586	0.6897
持つ	0.9153	0.8814	0.9153
やる	0.963	0.963	0.963
いく	0.8519	0.8148	0.8519
書く	0.7778	0.7037	0.7778
平均	0.7706	0.758	0.7067

表 4: 実験結果 (PN → PB)

単語	Normal	NB	uLSIF
聞く	0.5528	0.2602	0.7642
言う	0.8402	0.7828	0.5835
子供	0.5161	0.2473	0.2473
時間	0.8649	0.9054	0.9054
自分	0.974	0.974	0.974
入れる	0.6545	0.4	0.6364
出る	0.6447	0.5592	0.5592
取る	0.3827	0.2716	0.4074
場合	0.8394	0.1533	0.8467
入る	0.5897	0.3761	0.3761
前	0.8938	0.8562	0.6438
見る	0.8519	0.8444	0.8444
持つ	0.6993	0.6993	0.6993
やる	0.9416	0.9416	0.9416
いく	0.8647	0.8647	0.8647
書く	0.8387	0.9032	0.9032
平均	0.7468	0.6275	0.6998

表 5: 実験結果 (OC → PN)

単語	Normal	NB	uLSIF
聞く	0.4231	0.4615	0.7115
言う	0.4711	0.4545	0.7218
子供	0.4483	0.5172	0.4828
時間	0.678	0.661	0.661
自分	0.9718	0.9296	0.9859
入れる	0.7188	0.7188	0.7188
出る	0.7079	0.7416	0.7191
取る	0.2558	0.2558	0.2326
場合	0.8493	0.8493	0.8493
入る	0.3231	0.4154	0.2154
前	0.7619	0.7238	0.7238
見る	0.6897	0.7241	0.6897
持つ	0.8983	0.9153	0.9153
やる	1	1	1
いく	0.8519	0.8519	0.8519
書く	0.8148	0.8148	0.7778
平均	0.679	0.6897	0.7035

表 6: 実験結果 (PN → OC)

単語	Normal	NB	uLSIF
聞く	0.5242	0.3629	0.6613
言う	0.8318	0.7553	0.5315
子供	0.5325	0.8701	0.8701
時間	0.8302	0.8302	0.8302
自分	0.875	0.875	0.875
入れる	0.7534	0.2877	0.7397
出る	0.6947	0.5878	0.687
取る	0.2623	0.0492	0.0984
場合	0.9603	0.9762	0.9762
入る	0.4328	0.3433	0.2985
前	0.8762	0.7905	0.8667
見る	0.5692	0.5654	0.5654
持つ	0.5902	0.5574	0.5902
やる	0.9565	0.9565	0.9565
いく	0.6849	0.6849	0.6849
書く	0.6869	0.7374	0.7374
平均	0.6913	0.6394	0.6856

また，各領域適応の平均正解率を表 7 に，各手法の全領域適応の平均正解率の平均を評価値として表 8 に示す.

表 7: 各領域適応の平均値

領域適応	Normal	NB	uLSIF
OC → PB	0.7144	0.7216	0.6935
PB → OC	0.7	0.6863	0.6594
PB → PN	0.7706	0.758	0.7067
PN → PB	0.7468	0.6275	0.6998
OC → PN	0.679	0.6897	0.7035
PN → OC	0.6913	0.6394	0.6856

表 8: 各手法の評価値

手法	評価値
Normal	0.717
NB	0.6871
uLSIF	0.6916

5　考察

5.1　NB と uLSIF の比較

本実験で算出した NB と uLSIF の値を比較するために，各領域適応の各対象単語に関して両者の手法で算出した確率密度比が小さいデータ上位 10% を取り出した．NB により取り出したデータ群を H とし，uLSIF により取り出したデータ群を K として，$|H \cap K|/|H \cup K|$ の値を調べた．この値が 1 に近ければ NB と uLSIF の値は概ね同じような傾向があると言える．結果を表 9 に示す.

表 5.1 をみると明らかに NB と uLSIF は異なる値を付与している．どちらの値がより真の値に近いかは判断できないが，算出される確率密度比はかなり異なる.

表 9: NB と uLSIF の密度比が低い共通用例

単語	OC → PB	PB → OC	PB → PN	PN → PB	PN → OC	OC → PN
聞く	0.6	0.333	0.143	0.667	0.667	0.263
言う	0.158	0.121	0.099	0.241	0.108	0.128
子供	0.273	0.286	0	1	1	0.077
時間	0.429	0.167	0.4	0.2	0.2	0.429
自分	0.333	0.132	0.034	0.556	0.4	0.091
入れる	0.4	0.111	0.111	1	0.5	0.077
出る	0.3	0.111	0.304	0.385	0.2	0.529
取る	0.2	0.333	0	0.333	0.333	0.091
場合	0.412	0.13	0.13	0.556	0.556	0.2
入る	0.2	0.1	0.294	0.333	0.333	0.5
前	0.333	0.333	0.185	0.333	0.333	0.25
見る	0.083	0.256	0.059	0.455	0.455	0
持つ	0.333	0.2	0.071	0.333	0.5	0.333
やる	0.375	0.154	0	0.333	0.333	0
いく	0.257	0.182	0	0	0.333	0
書く	0.053	0.333	0.091	0	1	0.053
平均	0.429	0.258	0.071	0.333	0.5	0.132

5.2　確率密度比の p 乗の利用

共変量シフト下での学習ではパラメータ推定のバラツキが大きいことが指摘されている．このパラメータのバラツキによって識別精度が安定しないと考えられる．そこで，確率密度比 $w(\boldsymbol{x})$ を $0 < p < 1$ 乗した $w(\boldsymbol{x})^p$ を重みとして適用し，バラツキを抑えることが提案されている (杉山将 (2006))．

ここで以上のバラツキを抑える手法を用いて再度実験を行ってみる．前章の実験において NB と uLSIF で，確率密度比を $w(\boldsymbol{x})^{0.5}$ として実験を行う．その結果を表 10 と表 11 に示す．

表 10: 実験結果

領域適応	Normal	NB	uLSIF	NB $(w(\boldsymbol{x})^{0.5})$	uLSIF $(w(\boldsymbol{x})^{0.5})$
OC → PB	0.7144	0.7216	0.6935	0.712	0.7038
PB → OC	0.7	0.6863	0.6594	0.6849	0.659
PB → PN	0.7706	0.758	0.7067	0.7595	0.7203
PN → PB	0.7468	0.6275	0.6998	0.6484	0.7049
OC → PN	0.679	0.6897	0.7035	0.6929	0.6829
PN → OC	0.6913	0.6394	0.6856	0.6767	0.7075

NB も uLSIF も平方根を取ることで全体の評価値は向上しているがわずかに uLSIF の方がその値は高く，確率密度比を直接推定した方がより有効であることがここからもわかる．ただしどちらにしても Normal よりは評価値は低かった．

表 11: 各手法の評価値

手法	評価値
Normal	0.717
NB	0.6871
uLSIF	0.6916
NB $(w(\boldsymbol{x})^{0.5})$	0.6957
uLSIF $(w(\boldsymbol{x})^{0.5})$	0.6964

6　おわりに

本論文では，WSD の領域適応の問題が共変量シフトの問題であることを示し，共変量シフトの解法である確率密度比を重みとしたパラメータ学習により問題の解決をを図った．共変量シフト下での学習では，訓練データ $(\boldsymbol{x}, y)$ に対して，確率密度比 $w(\boldsymbol{x}) = \frac{P_t(\boldsymbol{x})}{P_s(\boldsymbol{x})}$ の重みをつけて学習する．ここでは $P_t(\boldsymbol{x})$ と $P_s(\boldsymbol{x})$ をそれぞれ求め，その比を取る手法と，直接 $w(\boldsymbol{x})$ を推定する手法である uLSIF を試した．どちらの手法も用いても識別の精度は向上しなかった．またそれぞれの手法の値を比較すると，かなり異なる値であることも確認できた．どちらがより有効な推定方法であるかは，今後調査する必要がある．

文献

Daumé III, Hal (2007) “Frustratingly Easy Domain Adaptation,” in *ACL-2007*, pp. 256–263.

Jing Jiang and Chengxiang Zhai (2007) “Instance weighting for domain adaptation in NLP,” in *ACL-2007*, pp. 264–271.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama (2009) “A Least-squares Approach to Direct Importance Estimation,” in *Journal of Machine Learning Research 10*, pp. 1391–1445.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Naoaki Okazaki (2009) “Classias: a collection of machine-learning algorithms for classification.”

Hidetoshi Shimodaira (2000) “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, Vol. 90, No. 2, pp. 227–244.

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

Hiroya Takamura Yosuke Saiki and Manabu Okumura (2008) “Domain Adaptation in Sentiment Classification by Instance Weighting,” *IPSJ SIG Technical Report. SIG-NL Report*, Vol. 2008, No. 33, pp. 61–67.

高村大也 (2010) 言語処理のための機械学習入門，コロナ社.

新納浩幸、佐々木稔 (2014) 「共変量シフトの問題としての語義曖昧性解消の領域適応」，自然言語処理，第 21 巻，第 1 号，(to appear).

杉山将 (2006) 「共変量シフト下での教師付き学習」，日本神経回路学会誌，第 13 巻，第 3 号，pp.111–118.

トピックモデルと概念辞書による日本語の語義曖昧性解消

佐々木 悠人（東京農工大学　工学部情報工学科）
古宮 嘉那子（東京農工大学　工学研究院）
小谷 善行（東京農工大学　工学研究院）

Word Sense Disambiguation Using Topic Model and Thesaurus

Yuto Sasaki (Faculty of Engineering, Tokyo University of Agriculture and Technology))
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

1. はじめに

語義曖昧性解消（Word Sense Disambiguation，WSD）とは，複数の語義を持つ単語（多義語）が文章中に出現した際に，どの語義を表しているのかを判断するタスクである．WSD に関する研究では語義タグ付きコーパスを利用する教師有り手法が高い精度を出しているが，語義タグ付きコーパスの作成はコストの大きい作業であるため，教師データを必要としない教師無し手法も盛んに研究されている．

教師無しの WSD に対する一つのアプローチとしてトピックモデルをベースにした手法が報告されているが，そのほとんどが英語を対象としており，日本語について行われた研究や実験は少ない．そこで，すでに提案されている既存モデルを日本語に対応させ，日本語を対象とした教師無し WSD の実験を行う．

2. 関連研究

教師無し WSD に関する研究は多く，様々な手法が考えられている．中でもトピックモデル LDA（Latent Dirichlet Allocation）（Blei ら(2003)）を使用したものとしては，（Boyd-Graber ら(2007)）や（Guo ら(2011)）がある．Boyd-Graber らは，トピックが持つ単語の確率分布を概念体系 WordNet 上での単語生成過程である WORDNET-WALK に置き換えた Latent Dirichlet Allocation with WORDNET（LDAWN）というモデルを考案し，トピックモデルを教師無しの英語 WSD へ応用した．Guo らも同様にトピックモデルと WordNet の組み合わせだが，概念構造は利用せず，辞書の定義文から事前学習を行う手法で，WSD に関しては Boyd-Graber らと同程度の精度を上げている．

本稿は，LDAWN を日本語に適用し，その成果を報告する．

3. トピックモデルによる WSD システム

システムの入力から出力までの基本的な流れは以下である．

まず，入力として語義タグのついていないタグ無しコーパスの集合を受け取る．入力として与えられたコーパスに対して形態素解析を行って得られた単語列のうち，名詞と動詞の自立語を対象語とし，これらの基本形と品詞を取得する．名詞・動詞以外の品詞の単語や非自立語については，WSD の対象として扱わない．

入力として対象語が与えられると，システムは与えられたすべての対象語に対して，トピックモデルと概念辞書を用いた手法によりその語義を推定する．最終的に，すべての対象語に対して，各対象語が取りうる語義を辞書から一つ選択して割り当てた結果が出力として得られる．

3.1. トピックモデル LDAWN

LDAWN は WORDNET-WALK を利用したモデルである．WORDNET-WALK とは，WordNet のような概念同士の上位下位関係が木構造（あるいは束構造）の形で定義されている概念体系において，ルート概念から確率的に下位概念への遷移を繰り返していき，リーフ概念まで辿りついたらその概念が示す単語を出力する，という単語生成過程である．図 3.1 に WORDNET-WALK による単語の生成確率の簡単な例を示す．丸いノードは概念，三角のノードはリーフ概念（X，Y）を語義として持つ単語であり，数字は遷移確率を表しているとすると，単語 A，B，C，D の生成確率はそれぞれ 0.03，0.27，0.28，0.42 である．トピックごとの単語の確率分布をこのようなルート概念からの遷移による確率分布に置き換えることで，単語の語義ごとの確率を扱うことができる．たとえば，単語 A と単語 C が同じ単語の場合，この単語は語義 X より語義 Y を取りやすい．

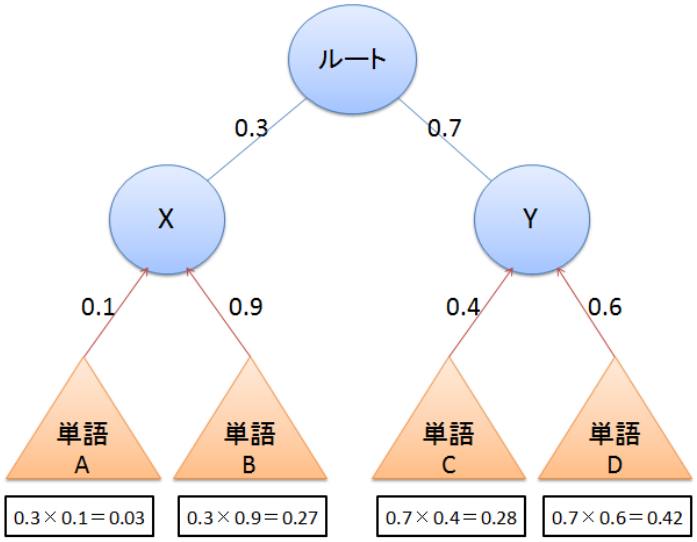


図 3.1　WORDNET-WALK による単語の生成確率の例

3.2. LDAWN における多義語の語義推定

LDAWN における多義語の語義推定方法について述べる．LDAWN では，全文書の全単語列$\boldsymbol{w}$が観測された状態での，各単語に対応する語義列$\boldsymbol{s}$の確率$P(\boldsymbol{s}|\boldsymbol{w})$に従って確率的に語義列を選択し，多義語の語義とする方法を取る．

まず，LDAWN における，全文書の生成確率は次のようになる：

$$P(\boldsymbol{z},\boldsymbol{s},\boldsymbol{w}) = \prod_{d=1}^{D}\prod_{i=1}^{N_d} P(z_{d,i})P(s_{d,i}|z_{d,i})P(w_{d,i}|z_{d,i},s_{d,i}) \tag{3.1}$$

Dは文書数，N_dは文書d中の単語数，$z_{d,i}$は文書dのi番目の単語$w_{d,i}$のトピックであり，$s_{d,i}$は概念体系上のルート概念から単語$w_{d,i}$の語義までの経路である．$P(w_{d,i}|z_{d,i},s_{d,i})$は概念（語義）から単語への遷移確率（図 3.1 の例で言うと，概念 X から単語 A への遷移確率などがこの確率に該当する）である．$P(z_{d,i})$は文書が持つトピックの確率分布からトピックを選択

する確率であり，この分布は試行回数$n=1$とした場合の多項分布に該当する．同様にWORDNET-WALK における下位概念への遷移も多項分布となる．これらの確率分布のパラメータには，事前分布として多項分布と共役な分布であるディリクレ分布を設定する．

パラメータまで考慮した全文書の生成確率は，

$$\begin{aligned}&\boldsymbol{P}(\boldsymbol{z},\boldsymbol{s},\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\beta})\\&=\prod_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{P}(\boldsymbol{\beta_k}|\boldsymbol{S\alpha_k})\prod_{\boldsymbol{d=1}}^{\boldsymbol{D}}\boldsymbol{P}(\boldsymbol{\theta_d}|\boldsymbol{\gamma_d})\prod_{\boldsymbol{i=1}}^{\boldsymbol{N_d}}\boldsymbol{P}(\boldsymbol{z_{d,i}}|\boldsymbol{\theta_d})\boldsymbol{P}(\boldsymbol{s_{d,i}}|\boldsymbol{\beta}_{\boldsymbol{z_{d,i}}})\boldsymbol{P}(\boldsymbol{w_{d,i}}|\boldsymbol{\beta}_{\boldsymbol{z_{d,i}}},\boldsymbol{s_{d,i}})\end{aligned}\tag{3.2}$$

となる．Kはトピック数，θ_dは文書dが持つトピックに関する確率分布，β_kはトピックkにおける各概念からその下位概念への遷移の確率分布であり，それぞれ多項分布のパラメータである．γ_d，$S\alpha_k$はディリクレ分布のパラメータである．これらは，パラメータのパラメータなのでハイパーパラメータと呼ばれる．βのハイパーパラメータ$S\alpha$については，下位概念への遷移に関するパラメータなので，遷移確率パラメータと呼ぶことにする．トピックkにおける概念s_iからその下位概念s_jへの遷移確率パラメータ$S\alpha_{k,s_i,s_j}$を，すべての下位概念についての和$\sum_j \alpha_{k,s_i,s_j}=1$となるように設定すると，遷移確率パラメータは定数$S$でその影響を調整できる．そのため，遷移確率パラメータは$\sum_j \alpha_{k,s_i,s_j}=1$となるように設定する．遷移確率パラメータの直観的な意味としては，事前に概念s_iからその下位概念にS回の遷移が起こり，どの下位概念に遷移したかの割合がαで決まる，というようなものである．

最終的に，文書の全単語$\boldsymbol{w}$が観測された状態での隠れ変数の条件付き確率は，

$$\boldsymbol{P}(\boldsymbol{z},\boldsymbol{s},\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{w})=\frac{\boldsymbol{P}(\boldsymbol{z},\boldsymbol{s},\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\beta})}{\boldsymbol{P}(\boldsymbol{w})}\propto\boldsymbol{P}(\boldsymbol{z},\boldsymbol{s},\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\beta})\tag{3.3}$$

となり，(3.2)式に比例する．ここから，一つの変量を除いた他のすべての変量を定数としたときの分布（全条件付き分布）を求め，ギブスサンプリングによるサンプリングを行う．(3.2)式について，s_iの経路中で通過する概念を$c_{i,1}c_{i,2}\dots c_{i,l}$とする．パラメータ$\theta$と$\beta$を積分消去し，単語$w_i$に関する変量$z_i$と$s_i$以外の変量を定数と見なすと，条件付き分布$\boldsymbol{P}(\boldsymbol{z_i},\boldsymbol{s_i}|\boldsymbol{z_{-i}},\boldsymbol{s_{-i}},\boldsymbol{w})$は，

$$\boldsymbol{P}(\boldsymbol{z_i},\boldsymbol{s_i}|\boldsymbol{z_{-i}},\boldsymbol{s_{-i}},\boldsymbol{w})\propto\frac{\boldsymbol{n}^{\boldsymbol{d}}_{\boldsymbol{-i},\boldsymbol{z_i}}+\boldsymbol{\gamma}}{\sum_{\boldsymbol{j}}\ \boldsymbol{n}^{\boldsymbol{d}}_{\boldsymbol{-i},\boldsymbol{j}}+\boldsymbol{K\gamma}}\prod_{\boldsymbol{p=1}}^{\boldsymbol{l}}\frac{\boldsymbol{T}^{\boldsymbol{-i}}_{\boldsymbol{z_i},\boldsymbol{c_{i,p}},\boldsymbol{c_{i,p+1}}}+\boldsymbol{S\alpha}_{\boldsymbol{c_{i,p}},\boldsymbol{c_{i,p+1}}}}{\sum_{\boldsymbol{c}}\ \boldsymbol{T}^{\boldsymbol{-i}}_{\boldsymbol{z_i},\boldsymbol{c_{i,p}},\boldsymbol{c}}+\boldsymbol{S}}\tag{3.4}$$

となる．数式表記の簡単化のため，w_iをc_{l+1}としている．Kはトピック数，$n^d_{-i,j}$は現在サンプリング対象としているw_iを除いた状態で，w_iを含んでいる文書d中でトピックjが割り当てられた単語の数，$T^{-i}_{z,c_p,c_{p+1}}$は同様にw_iを除いた状態で，トピックzが割り当てられた全単語が概念c_pからその下位概念c_{p+1}へのリンクを通過した回数である．さらに(3.4)式をs_iについて周辺化すると，トピックz_iの確率が求められる：

$$P(z_i|z_{-i}, s_{-i}, w) \propto \frac{n^d_{-i,z_i} + \gamma}{\sum_j n^d_{-i,j} + K\gamma} \sum_{s_i} \prod_{p=1}^{l} \frac{T^{-i}_{z_i,c_{i,p},c_{i,p+1}} + S\alpha_{c_{i,p},c_{i,p+1}}}{\sum_c T^{-i}_{z_i,c_{i,p},c} + S} \quad (3.5)$$

(3.5)式の前半部分は，文書dでトピックz_iが割り当てられる確率であり，後半部分はトピックz_iで単語w_iが出現する確率を表している．すべてのトピックz_iについて(3.5)式から確率を計算し，得られた確率に従って単語w_iにトピックz_iを割り当てる．その後，割り当てられたトピックz_iから，語義（ルート概念からの経路）s_iを選択し，単語w_iに割り当てる．この処理を全単語について行うことをイテレーション一回とし，イテレーションを何度も繰り返すことで語義の推定を行う．

3.3. ハイパーパラメータの設定

モデルで設定すべきハイパーパラメータには，各文書が持つトピックに関する確率分布θのハイパーパラメータγと，WORDNET-WALK における各概念から下位概念への遷移確率βのハイパーパラメータ$S\alpha$の二つがある．二種類のパラメータのうち，γはすべてに一様な定数を与えて，対称な事前分布とする．一方，遷移確率パラメータ$S\alpha$もそのように対称なパラメータとすると，階層が深い概念，すなわちルート概念から下位概念への遷移がより多く行われる概念の確率が低くなる．そのような性質は好ましくないため，$S\alpha$については適当な値を考える必要がある．Boyd-Graber らは実験用とは別のコーパスを用意し，各リーフ概念がコーパス中での単語の出現頻度に比例した確率となるような値を設定した．今回の実装では，単語の出現頻度を利用した値の他に，各リーフ概念への遷移確率が等しくなるような値の設定も行う．

3.3.1. 単語の出現頻度に比例する遷移確率パラメータ

コーパス中の単語の出現頻度を利用する方法では，実験用とは別のコーパスを用意してそのコーパス中に出現した単語トークンの頻度を数え，その頻度に比例した値を遷移確率パラメータとして設定する．実際に単語トークンの出現頻度から遷移確率を計算する方法としては，各概念sの確率$P(s)$を求め，概念s_iから概念s_jへの遷移確率α_{s_i,s_j}を，

$$\alpha_{s_i,s_j} = P(s_j|s_i) = \frac{P(s_i, s_j)}{P(s_i)} = \frac{P(s_j)}{P(s_i)} \quad (3.6)$$

として求める方法がある(Jiang ら(1997))．この遷移確率をそのまま遷移確率パラメータα_{s_i,s_j}の値とする．

単語の出現頻度から概念の確率$P(s)$を求める方法として，ここでは（Resnik(1995)）を採用する．Resnik の手法では，ある概念の頻度は，その概念が含む概念（その概念自身とすべての下位概念）に属する単語のコーパス中での出現頻度の総和であるとし，概念sの確率$P(s)$は，概念sの頻度を単語トークンの出現数の総和で割ることで計算する．

このようにコーパスの単語頻度を使う場合は，出現しない単語によるゼロ頻度問題を考える必要がある．この問題に対処するため，今回のシステムではリーフ概念すべての出現回数に 0.5 を加算することにした．

3.3.2. 均等な遷移確率パラメータ

コーパス中の単語の出現頻度に基づく遷移確率の他に，ルート概念から各リーフ概念への各経路の確率が均等になるような遷移確率パラメータの設定も行う．まず，3.3.1 節で述べた Resnik の手法において，各単語を語義への経路によって別のものと考えて頻度を算出する．この場合の各概念の頻度$freq(s)$を式で表すと，概念s_iから概念s_jへ到達する経路の数を$path(s_i, s_j)$，概念sが含むリーフ概念の集合を$L(s)$としたとき，

$$\boldsymbol{freq(s_i) = \sum_{s \in L(s_i)} path(s_i, s) \sum_{w \in words(s)} count(w)} \tag{3.7}$$

となる．$words(s)$は概念sとその下位概念に属する全単語であり，$count(w)$は単語wのコーパス中での出現頻度である．遷移確率は下位概念の頻度を上位概念の頻度で割ることで計算する．

この計算法ですべてのリーフ概念の頻度に 1 を割り当てると，ルート概念から各リーフ概念への各経路の確率が等しくなる．ルート概念s_{root}から任意のリーフ概念s_lへ到達する任意の経路$path_{s_l}$の遷移確率$p(path_{s_l})$は，経路$path_{s_l}$中で通過する概念を$c_1 c_2 \ldots c_n$とすると，

$$\boldsymbol{P(path_{s_l}) = \frac{freq(c_1)}{freq(s_{root})} \frac{freq(c_2)}{freq(c_1)} \cdots\cdots \frac{freq(c_n)}{freq(c_{n-1})} \frac{freq(s_l)}{freq(c_n)} = \frac{freq(s_l)}{freq(s_{root})}} \tag{3.8}$$

である．すべてのリーフ概念の頻度$freq(s_l) = 1$とすると，

$$\boldsymbol{p(path_{s_l}) = \frac{freq(s_l)}{freq(s_{root})} = \frac{1}{\sum_{s \in L(s_{root})} path(s_{root}, s)}} \tag{3.9}$$

となり，s_lによらない定数となる．よって，すべてのリーフ概念の頻度を 1 にすると，ルート概念から各リーフ概念への各経路の確率が等しくなるような遷移確率パラメータαを設定することができる．

4. 日本語 WSD の実験

4.1. 実験データ

実験には，EDR 電子化辞書[1]を利用する．EDR 電子化辞書は，概念同士の上位下位関係を木構造（束構造）として持つ，日本語に対応した概念体系である．EDR 電子化辞書は，「単語辞書」，「対訳辞書」，「概念辞書」，「共起辞書」，「専門用語辞書」，「EDR コーパス」の六つの辞書から構成されている．本実験では，単語辞書，概念辞書，EDR コーパスの三つを利用する．それぞれ日本語と英語が用意されているが，日本語のみを利用する．また，EDR 電子化辞書の最新版は平成 22 年にリリースされた Ver4.0 だが，本実験では平成 14 年の Ver2.0 を使用した．

[1] NiCT「EDR 電子化辞書」http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html

4.1.1. システムの辞書の作成

システム内部の辞書の作成には，概念辞書と単語辞書を使用する．概念辞書は，概念見出し辞書と概念体系辞書の二つの辞書から成り，見出し辞書は概念識別子（概念を識別する 16 進数の整数）と概念見出し（概念の意味内容に近い単語），概念説明を対応付けている．概念体系辞書は概念同士の関係，特に上位下位概念関係によって体系化した辞書であり，二つの概念の上位下位関係を記述したレコードから成る．単語辞書は，単語見出しや品詞などの情報と，その単語が持つ語義（概念識別子）が記述されている．

辞書に登録する単語は，すべての名詞と動詞のうち，概念体系上でルート概念から辿ることができる単語とした．また，サ変動詞については末尾の「する」を除いた名詞形も登録した．さらに，日本語では使用されない概念と未分類概念は削除した．

4.1.2. EDR 日本語コーパス

WSD の実験には EDR 日本語コーパスを使用する．EDR のコーパスは，複数の出典元文書から文章単位で情報を抜き出し，形態素，構文情報，意味情報などを付与している．コーパスの各形態素には語義タグが付与されているため，システムの出力と比較して正解判定を行うことができる．

出典元文書は「日本経済新聞」,「朝日新聞」,「アエラ」,「平凡社百科辞典」,「岩波情報科学辞典」,「雑誌」,「用例集」の七つである．それぞれに含まれる文章数と総形態素数を表 4.1 に示す．

表 4.1　出典別の文章数と総形態素数

出典	文章数	総形態素数
日本経済新聞	5018	121301
朝日新聞	91400	2272555
アエラ	49589	1183897
平凡社百科辞典	10072	284059
岩波情報科学辞典	13578	357607
雑誌	21199	528452
用例集	16946	368285

今回は実験対象として日本経済新聞の全文章を選択する．日本経済新聞の全文章中の多義語の種類数は 4822，名詞と動詞のトークン数はそれぞれ 12149，6199 だった．また，多義語の平均語義数は名詞 4.2，動詞 5.5 だった．日本経済新聞以外の六つの出典の文章は，単語の出現頻度に比例する遷移確率の計算のためのコーパスとして使用する．

トピックモデルでは文書がトピックに関する分布を持つことを仮定しているため，入力は文書単位である必要がある．EDR のコーパスは，出典元から文章単位で抽出されており基本的には文書単位ではないが，日本経済新聞，朝日新聞，アエラの三つの出典では，「日本経済新聞８７０１０１」のように各文章が得られた記事の日付に関する情報があり，その日の記事に含まれる文章を抽出することができる．そこで，一つの入力方法として，日付ごとに文書化し,「日本経済新聞８７０１０１」で一文書,「日本経済新聞８７０１０２」で一文書，というように扱う方法を採用する．しかし，これらはその日の複数の内容の記

事から文章を飛び飛びで抽出しており，複数のトピックが混在してしまっているという問題がある．もう一つの入力方法として，一文章を一文書として扱うという方法も試みる．この場合は，一つの文書に含まれる単語が少なくなり，トピックに関する情報が十分に得られないという問題がある．

4.1.3. 形態素情報の補足処理

EDR コーパスの各文章には形態素情報が付与されており，語義タグは各形態素に対して付与されている．そのため，語義タグを使用して正解判定を行うためには，EDR の形態素に合った入力を行う必要がある．しかし，EDR コーパスの形態素情報には，動詞の原形や自立語かどうかといった情報が含まれていない．システムの辞書は単語辞書から作られるため，単語は全て基本形で登録されており，動詞は原形でなければ扱うことができない．そこで，動詞の原形や自立語の情報を得るため，形態素解析器 MeCab[2]の解析結果と比較をして，対応が取れた形態素について原形化や自立語の判定を行った．

4.2. 評価方法

システムは全ての対象語（名詞と動詞の自立語で，単語辞書に載っている単語）に対して語義を割り当てるので，正解判定は対象語の中の全ての多義語に対して行う．正解かどうかの判定はコーパスの語義タグ情報を利用し，システムに割り当てられた語義が語義タグと一致した場合に正解とする．ただし，語義タグがついていない場合と，語義タグがついているが単語辞書に載っていない語義タグを指しているものは正解判定から除外した．

5. 日本語 WSD の実験結果

LDAWN を用いて, EDR コーパス中の日本経済新聞の全文章に対して WSD の実験を行った結果を示す．予備実験の結果から，トピック数は 32，トピックに関する確率分布のハイパーパラメータは$\gamma = 0.01$と設定した．

5.1. 文書の扱いに関する実験結果

一日分の文章を一文書と見なす場合と，一文章を一文書と見なす場合について比較を行った実験について述べる．表 5.1 の設定で実験を行い，正解率の平均を算出した結果を表 5.2 に示す．

表 5.1　文書の扱いに関する実験の各種設定値

パラメータ	設定値
入力文書の形態	{一日分を一文書，一文章を一文書}
遷移確率パラメータのS	{5.0，10.0}
遷移確率パラメータのα	均等な遷移確率パラメータ
イテレーション回数	1000 回（800 回分のイテレーションで割り当てられた語義から間をあけて 100 サンプル取り，最頻語義を出力）
実行回数	2 回

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

表 5.2　一日分を一文書と見る方法と一文章を一文書と見る方法の比較結果
（トピック数 32，2 回実行した平均）

条件	マイクロ平均（全体）	マイクロ平均（名詞）	マイクロ平均（動詞）	マクロ平均
一日分を一文書 $S = 5.0$	0.3230	0.3457	0.2786	0.4053
一日分を一文書 $S = 10.0$	0.3119	0.3338	0.2689	0.4059
一文章を一文書 $S = 5.0$	0.3567	**0.3785**	0.3141	0.4226
一文章を一文書 $S = 10.0$	**0.3587**	0.3755	**0.3258**	**0.4234**

この結果を見ると，一日分の文章を一文書と見なした場合の精度が全体的に低いことが分かる．また，一日分を一文書とした場合は，Sが大きくなるとマクロ平均に大きな差は出ないが，マイクロ平均が下がっていく傾向にある．一方，一文章を一文書とした場合は，マイクロ平均，マクロ平均ともに上がっていく傾向が見られた．

5.2.　遷移確率パラメータに関する実験結果

遷移確率パラメータとして，単語の出現頻度に比例する遷移確率パラメータと，ルート概念からの各リーフ概念への各経路の確率を等しくする均等な遷移確率パラメータの二つの場合について比較を行った実験について述べる．単語の出現頻度には，日本経済新聞以外の六つの出典（朝日新聞，アエラ，平凡社百科辞典，岩波情報科学辞典，雑誌，用例集）の全文章から得られた単語トークンの頻度を利用した．

表 5.3 の設定で実験を行った結果を表 5.4 に示す．なお，均等な遷移確率パラメータにおける$S = 5.0, 10.0$のときの実験結果は，5.1 節の実験結果と同じものである．表 5.4 の結果を見ると，遷移確率パラメータのαを各経路均等とした場合の方が全体的に精度が良い．

表 5.3　遷移確率パラメータに関する実験の各種設定値

パラメータ	設定値
入力文書の形態	一文章を一文書
遷移確率パラメータのS	{5.0，10.0，15.0}
遷移確率パラメータのα	{単語の出現頻度に比例する遷移確率パラメータ， 各経路の遷移確率を等しくする遷移確率パラメータ}
イテレーション回数	1000 回（800 回分のイテレーションで割り当てられた語義から間をあけて 100 サンプル取り，最頻語義を出力）
実行回数	2 回

表 5.4　単語頻度に比例する遷移確率パラメータと均等な遷移確率パラメータの比較結果（トピック数 32，2 回実行した平均）

遷移確率パラメータ	マイクロ平均（全体）	マイクロ平均（名詞）	マイクロ平均（動詞）	マクロ平均
単語頻度に比例 $S = 5.0$	0.3478	0.3772	0.2901	0.4151
単語頻度に比例 $S = 10.0$	0.3465	0.3811	0.2787	0.4191
単語頻度に比例 $S = 15.0$	0.3472	**0.3827**	0.2777	0.4180
各経路均等 $S = 5.0$	0.3567	0.3785	0.3141	0.4226
各経路均等 $S = 10.0$	**0.3587**	0.3755	**0.3258**	0.4234
各経路均等 $S = 15.0$	0.3515	0.3714	0.3126	**0.4243**

6.　日本語 WSD の実験結果の考察

6.1.　ランダムベースラインとの比較

比較用のベースラインとして，文章中のすべての多義語に対して，取りうる語義の中からランダムに選択したものを付与するランダムベースラインを設定する．ランダムベースラインの正解率を表 6.1 に示す．なお，この値は 1000 回の平均を取ったものである．

表 6.1　ランダムベースラインの実験結果（1000 回の平均）

マイクロ平均（全体）	マイクロ平均（名詞）	マイクロ平均（動詞）	マクロ平均
0.3097	0.3317	0.2666	0.3663

LDAWN による各実験結果とランダムベースラインを比較すると，マイクロ平均とマクロ平均ともにどの場合でもランダムベースラインを上回っている．LDAWN とランダムベースラインとの差が有意なものであるかを確認するため，LDAWN の結果の中で全体のマイクロ平均が最も良かったパラメータ設定の場合とランダムベースラインとの間でχ^2検定を行う．マイクロ平均が最も良かったのは「一文章を一文書とする」，「遷移確率パラメータは各経路均等」，「$S = 10.0$」の場合であり，このときのマイクロ平均は 0.3587 だった．この結果とランダムベースラインとの間で，マイクロ平均から正解数を求めてχ^2検定を行うと，有意水準 1%で有意差が認められた．この結果から，LDAWN はランダムベースラインより優れており，日本語においてもトピックという情報は有効に作用していると考えられる．

6.2.　実験結果に関する考察

5.1 節の文書の扱いに関する実験では，コーパス中の日本経済新聞を出典元とする文章について，ある日の紙面で使われた文章すべてを一つの文書として扱う「一日分を一文書」とした場合の精度が低かった．さらに，遷移確率パラメータの定数Sを大きくすると，一日

分を一文書とした場合のマイクロ平均は低下していき，$S = 10.0$で全体のマイクロ平均が0.3119 とかなり低い値となった．マクロ平均の低下は見られないため，これは出現数の多い多義語の正解率が下がっているのではないかと考えられる．パラメータSを大きくすると，イテレーション中でトピックに対して単語が割り当てられることによる確率の変化が小さくなり，トピックに関する情報が相対的に重くなると予想される．一日分を一文書として扱った場合，複数の記事の内容から文書が構成されるため複数のトピック分布が混ざってしまい，結果的にトピックという情報が文脈情報を表さないランダムな要素として働き，トピックの情報が重くなるにつれて精度が下がっていったのではないかと考えられる．

一方で，「一文章を一文書」として扱った場合，$S = 5.0$から$S = 10.0$と大きくすると精度は上がっていく傾向が見受けられた．Sを大きくするとトピックに重きが置かれるため，トピックの情報がよく効いたおかげで精度が向上したと考えられる．

5.2 節の遷移確率パラメータに関する実験では，実験用とは別に用意したコーパスから単語の出現頻度に比例する遷移確率パラメータを計算した場合の精度が低かった．単語の単純な出現頻度に基づいて遷移確率に差をつける方法はあまり有効でない可能性がある．

7. おわりに

トピックモデルをベースにした既存手法 LDAWN に対し，EDR 電子化辞書による概念体系辞書を組み込み，日本語 WSD の実験を行った．実験には EDR の日本語コーパスのうち出典が日本経済新聞となっているものを利用した．文書の入力形態を変化させた実験，遷移確率パラメータを変化させた実験の二種類の実験を行った結果，全多義語に対するマクロ平均が最大となったのは，文書の入力形態は「一文章を一文書」，遷移確率パラメータは定数$S = 15.0$でαがルート概念からの各リーフ概念への経路が均等の場合であり，マクロ平均は 0.4243 であった．全多義語に対するマイクロ平均が最大となったのは，文書の入力形態は「一文章を一文書」，遷移確率パラメータは定数$S = 10.0$でαがルート概念からの各リーフ概念への経路が均等の場合であり，マイクロ平均は 0.3587 であった．マイクロ平均が最大となった結果と，語義をランダムに割り当てるランダムベースラインの結果を比較すると，LDAWN による結果は有意水準 1%で有意に優れていた．

参考文献

David M. Blei, Andrew Y. Ng and Michael I. Jordan (2003)「Latent Dirichlet Allocation」*Journal of Machine Learning Research 3*，pp.993-1022

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu (2007)「A Topic Model for Word Sense Disambiguation」*Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*，pp.1024-1033

Weiwei Guo and Mona Diab (2011)「Semantic Topic Models: Combining Word Distributional Statistics and Dictionary Definitions」*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*，pp.552-561

Jay J. Jiang and David W. Conrath (1997)「Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy」*Proceedings of International Conference Research on Computational Linguistics*，pp.19-33

Philip Resnik (1995)「Using Information Content to Evaluate Semantic Similarity in a Taxonomy」*International Joint Conferences on Artificial Intelligence*，pp.448-453

Modeling Japanese Language Register Using Corpus Metadata

Bor Hodošček　(National Institute for Japanese Language and Linguistics) †

コーパスのメタデータを用いた日本語におけるレジスターのモデル化

ホドシチェク・ボル　（国立国語研究所）

1. Introduction

This study proposes to combine extra-linguistic and inter-linguistic features in the modeling of linguistic variation, or *register*, in Japanese. A comprehensive account of register variation in Japanese has recently become possible with the public availability of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a 100 million token corpus that contains a wide variety of written Japanese (Maekawa, 2007).

First, this study will provide necessary background on register studies and the connection between them and corpus metadata. Next, the metadata available in the BCCWJ are analyzed into several categories thought to influence register, taking into account any hierarchical properties that exist within the metadata. Finally, two pilot experiments using techniques from subgroup discovery (Wrobel, 1997; Langohr et al., 2013) and exceptional model mining (Duivesteijn, 2013) are conducted on the BCCWJ, showing some possibilities for further explorations of register variation within the BCCWJ.

2. Previous Research

2.1 Language Variation

Numerous studies on the topic of linguistic variation have been conducted using the different terminologies of style, genre, register, text type, and domain, to describe the variation observable in language (Eckert & Rickford, 2001; Biber, 1995; Irvine, 2001; Lee, 2001). A possible account of the differences between the terms is offered below (also see Lee, 2001):

- Register: A variety of language associated with the specific situation of use. Example: register of written academic Japanese; classroom conversation
- Genre: A category of language defined by a community, or associated with expected rhetorical structure and themes. Example: genre of Japanese research articles; crime novels
- Style: Variations in language associated with an individual's "unique" uses of language. Example: sensationalist style; vague written style
- Text type: A grouping of texts based purely on linguistic features. Example: informational text type
- Domain: Text devoted to a single topic or a small set of related topics, often inside one genre. Example: domain of computational linguistics

† bor.hodoscek@gmail.com

Different approaches to modeling language variation have been developed, mirroring the needs and interests of many fields, including: socio-linguistics, historical linguistics, dialect research, NLP, text classification, and authorship identification. However, for the purposes of the present work, we will mostly be concerned with register in the broader sense, covering the situation of use, community, individual unique usage of language–but all of which must be tied to actual linguistic features. The vocabulary used to describe register is furthermore limited to the metadata available within the BCCWJ, which is introduced in the following section.

2.2 Corpus Metadata

In the context of a corpus, we define metadata to be any data that describes some language-external fact about some part of a corpus. Therefore, this definition encompasses not only the common case of data describing some fact at the document level, but also includes data at the level of document groups, as well as data at the lower token, sentence, and paragraph levels (see Figure 1). A summary of the metadata annotations within the BCCWJ is provided in Maruyama (2012), while a further summary of metadata into formal and informal annotation approaches, as well as one based on granularity and dimension, is provided below through the lens of hierarchical classification systems available in the BCCWJ.

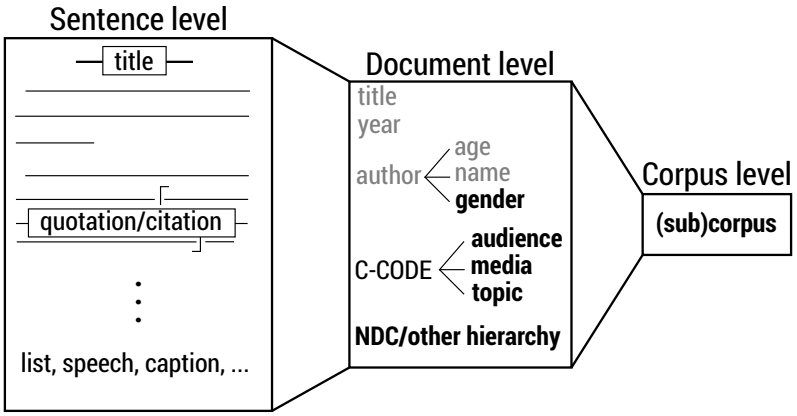


Figure 1: *Metadata at the sentence, document, and subcorpus levels. The metadata used in this study are marked in bold.*

Library classification systems such as the Dewey Decimal Classification (DDC) or the Japanese counterpart, the Nippon Decimal Classification (NDC), are classification systems which rely on expert catalogers and are arguably optimized for the efficient classification and storage of material in libraries. This guided approach to classification espouses a singular point of view and is thus susceptible to any oversights or gaps in knowledge of the catalogers. The organization of knowledge into ontologies is in many ways the most formal approach in this metadata category, and there are many cases where such a formal specification has been useful (Hirst, 2009, provides several).

In contrast, the other approach to metadata is more informal, and can involve collaborative classification of content using sets of tags in what is sometimes called a folksonomy (Vander Wal, 2007). This approach eschews the singular viewpoint (i.e., that of the compiler) of an ontology for the plurality of multiple viewpoints (i.e., that of many users). It is this property of the informal approach which

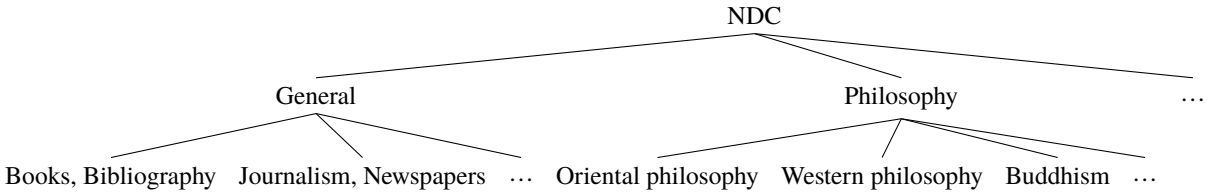


Figure 2: *A part of the hierarchical structure of the NDC.*

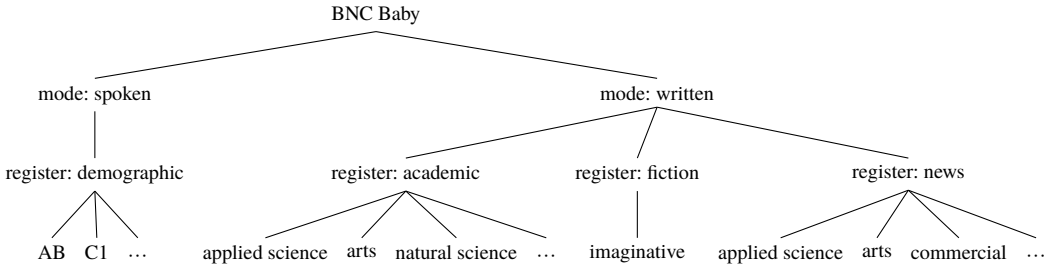


Figure 3: *A part of the hierarchical structure of the register-focused metadata within the British National Corpus Baby (as described in Gries, 2009, pp. 3–4).*

arguably makes it more flexible in describing a wider variety of metadata, given that the vocabulary of tags is unrestricted enough and that the number of taggers is sufficient to overcome incongruities in the tag data (Halpin, Robu, & Shepherd, 2007).

Finally, metadata can also be analyzed according to their granularity or dimensionality. The granularity of metadata refers to the complexity of structure contained in the metadata. For example, 1-dimensional metadata are composed of elements that are independent of one another[1]. An example of higher-dimensionality metadata is the NDC, in which 10 top-level categories further branch out into sub-categories (Figure 2).

As the annotation of corpus metadata depends on the research aims of the corpus and can thus encompass a diverse set of attributes, we limit the discussion to those corpus metadata which are involved in the identification of language variety of one kind or another. A representative example of corpus metadata that has been used for such purposes is that contained within the British National Corpus Baby (BNC Baby)[2]. For example, the text categories in the BNC Baby can be represented as a three-level hierarchy of mode, register, and sub-register, as shown in Figure 3.

Contrasting the hierarchy of the NDC with the register-focused hierarchy of the BNC Baby reveals limitations in using it as a proxy for register distinctions. A strict hierarchy such as the NDC is unable to capture all linguistic variation and situational factors. This is evident because the branching factor in the hierarchy, from parent to children, is often a combination of topic, register, and genre differences, which can interact at different levels. Instead, the realization of an ontology of linguistic variation with an associated metadata vocabulary would allow for the expression of arbitrary differences within and

[1]The Simple Dublin Core Metadata Element Set (DCMES) vocabulary is an example of this kind of linear structure (“Dublin Core Metadata Element Set, Version 1.1,” 2012).

[2]The complete list of BNC Baby categories is available from http://projects.oucs.ox.ac.uk/xaira/exercises/A1_xaira.html.

Table 1: *Number of tokens, sentences, paragraphs, and documents in the BCCWJ arranged according to media label. The last column "Hierarchy type" offers a short description of metadata at the media level.*

Code	Media label	Tokens	Sentences	Paragraphs	Documents	Hierarchy type
LB+PB+OB	Books	70,472,742	3,155,084	1,552,490	22,058	NDC, C-CODE
OY	Yahoo! Blogs	13,212,757	943,646	783,871	52,676	Yahoo! topics
OC	Yahoo! Q&A	12,088,127	780,510	624,616	91,445	Yahoo! topics
OM	Minutes of the Diet	5,600,649	139,802	45,810	159	place and type
OW	White papers	5,495,254	139,373	101,587	1,500	govt. division
PM	Magazines	5,114,752	281,765	155,260	1,996	magazine class.
OP	Local govt. newsletters	4,739,306	255,841	209,679	354	municipality
OL	Laws	1,206,481	33,289	25,364	346	subject
OT	School textbooks	1,126,214	63,370	45,952	412	subject and grade
PN	Newspapers	1,036,285	50,960	26,546	1,473	daily, evening
OV	Verse	237,685	18,974	18,974	252	haiku, etc.
	TOTAL	120,330,252	5,862,614	3,590,149	172,671	

between texts. However, the author has not yet encountered this kind of systematic encoding of the metadata inherent in corpora, and will, for the purposes of this study seek to examine only those parts of the metadata within the BCCWJ that more directly influence linguistic variation[3].

3. Materials

The BCCWJ contains a wide variety of contemporary written[4] Japanese documents with associated metadata of many forms. As can be seen in Table 1, the whole corpus, as used in this study, consists of approximately 120 million tokens[5], and is sub-divided according to general media labels. For the purposes of this study, the term "media label" is used in preference to the more established "subcorpus", as it provides a more accurate grouping of samples in the BCCWJ when one is not interested in differences between the three Books subcorpora (LB, PB, and OB).

Adding to the metadata that is unique to each media, the BCCWJ provides a variety of metadata for varying subsets of the corpus including: author name, gender, author decade of birth, publishing date, and publisher name. Most importantly for our purposes, up to 4 "genre" labels per document are provided, which allows the construction of hierarchies unique to each media label (see the "Hierarchy type" column in Table 1). It is quite evident that these "genre" labels represent different conceptualizations of genre, differing in their inclusion of topic, register, audience, and so on (see also Tanomura (2014) for another critique). The next section will detail the process of constructing hierarchies from

[3] See (Garbacz, 2006) for the beginnings of such an attempt based on the theory of document genres proposed in (Yates & Orlikowski, 1992; Orlikowski & Yates, 1994; Yates, Orlikowski, & Okamura, 1999).

[4] One notable exception is the Minutes of the Diet subcorpus, which consists of official transcriptions of dialogue.

[5] Tokens here refer to the Short Unit Words (SUWs) contained in the UniDic morphological dictionary version 2.1.2, which were extracted using the morphological analyzer MeCab version 0.966 from the C-XML variable length data of the BCCWJ, version 1.

these metadata.

4. Method

An understanding of the relationships between the metadata is a prerequisite to evaluating the discriminatory power of metadata in the modeling of language variation. Subgroup discovery, as described in the MIDOS algorithm (Wrobel, 1997), seeks to discover subgroups that have unusual distributional characteristics with respect to the entire population. Like MIDOS, we are looking for subgroups that have unusual distributional characteristics with respect to the entire population, but we are also further interested in unusual distributions with respect to the parent subgroup. First we define the language model that will be used to measure differences between different subgroups, and, in the succeeding subsection, detail the algorithm and modifications used to discover subgroups in the BCCWJ.

4.1 Language Model

The language contained in documents and subgroups is represented using a vector of word weights, calculated using a variant (wf.idf) of the well-known tf-idf (term frequency-inverse document frequency) formulation first proposed by Spärk Jones in 1972. Given a corpus containing documents, it is possible to weight words according to not only their frequency, but also to their dispersion across the number of documents they occur in. More formally:

$$\text{tf}(t, D) = \sum_{i=1}^{N} \text{number of occurrences of } t \text{ in } d_i$$

$$\text{df}(t, D) = \text{number of documents where } t \text{ occurs in}$$

$$\text{wf}(t, D) = 1 + \log_2 \text{tf}(t, D)$$

$$\text{idf}(t, D) = \log_2 \frac{N}{\text{df}(t, D)}$$

$$\text{wf.idf}(t, D) = \text{wf}(t, D) \cdot \text{idf}(t, D)$$

where t and d correspond to a term and a document, respectively. N is the total number of documents in the collection, which is 172,675 in the case of the BCCWJ. The term-frequency tf is weighted using sublinear scaling (wf), while the inverse document frequency idf is just the logarithm of the total number of documents divided by the number of documents containing t. This model, although simplistic, was chosen as a first step towards introducing more varied features that are able to capture more nuanced register variation in future work.

4.2 Subgroup Discovery and Exceptional Model Mining

Algorithm 1 gives the general outline of the SD/EMM process (a detailed explanation of the process is offered in Duivesteijn (2013), pp. 13–30). Intuitively, the algorithm starts the search from the vocabulary of features available (corresponding to the various metadata categories here). It then searches for increasingly small subsets of the data based on the most promising (according to the quality measure φ) subsets found so far. New subsets are thus comprised of conjoining the current promising subsets

(there are at most w of them) with new ones using the refinement operator η. For example, if a promising subset was (TOPIC = "politics"), the refinement operator would generate contrasting subgroups such as (TOPIC = "politics" $\wedge$ GENDER = "female") and (TOPIC = "politics" $\wedge$ GENDER != "female"). The quality score would then be computed by using the results of comparing the new subgroup and its complement with the whole dataset Ω. The search would continue until the maximum search depth (d) was reached, or until no more subgroups could be found, and the final result of which would be a list of the top-q subgroups sorted according to the chosen quality measure.

Algorithm 1: Beam search for top-q Exceptional Model Mining (reproduced from Duivesteijn, 2013, p. 19).

Input : Dataset Ω, QualityMeasure φ, RefinementOperator η, Integer w, d, q, Constraints C
Output: PriorityQueue resultSet

```
1  candidateQueue ← new Queue;
2  candidateQueue.enqueue({});
3  for Integer level ← 1; level ≤ d; level++ do
4      beam ← new PriorityQueue(w);
5      while candidateQueue ≠ ∅ do
6          seed ← candidateQueue.dequeue();
7          set ← η(seed);
8          foreach desc ∈ set do
9              quality ← φ(desc);
10             if desc.satisfiesAll(C) then
11                 resultSet.insert_with_priority(desc, quality);
12                 beam.insert_with_priority(desc, quality);
13             end
14         end
15     end
16     while beam ≠ ∅ do
17         candidateQueue.enqueue(beam.get_front_element());
18     end
19 end
20 return resultSet;
```

The following provides a detailed description of the parameters and methods used in Algorithm 1.

4.2.1 Distance function

The correlation between two wf.idf vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ is calculated using the sample correlation r:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

4.2.2 Beam-search parameters

The search width was set to 10, and the search depth was set to 8, though in practice no subgroups matching all the constraints were found beyond the fifth depth level.

4.2.3 Refinement operator

For nominal a_i with values $v_1, \ldots, v_g$ we add $\{D(a_i = v_j), D(a_i \neq v_j)\}_{j=1}^{g}$ to $\eta(D)$. The refinement of hierarchical data presents another challenge in that we must take into account the explicit constraints that hold within hierarchical data (cf. Park and Fürnkranz (2008), p. 2). Therefore, for each refinement of hierarchical metadata, all children of the current position in the hierarchy are generated.

4.2.4 Constraints

The minimum number of samples was chosen to be 20.

4.2.5 Quality measures

Following (Duivesteijn, 2013), two different quality measures were considered: φ_{scd} and φ_{ent}. The first quality measure, though statistically-oriented, revealed to be overly sensitive to the large subgroup document sizes the BCCWJ includes (which include over 170,000 documents) and measuring most subgroups' p value as zero. Thus the use of φ_{scd} was deemed to not be suitable for discovering subgroups in the present data, and other measures were considered:

- The most simple measure is simply the absolute difference between the correlation coefficient of the subgroup and its complement:

$$\varphi_{abs}(D) = |r^{G_D} - r^{G_D^C}|$$

- As the previous measure does not take into account the subgroup size or the distribution of the split of data between the subgroup and its complement, the entropy function φ_{ef} was considered:

$$\varphi_{ef}(D) = -{}^{n}/_{N}\, log_2\, {}^{n}/_{N} - {}^{n^C}/_{N}\, log_2\, {}^{n^C}/_{N}$$

- Finally, by combining φ_{abs} and φ_{ef}, we arrive at the heuristic measure φ_{ent}, which favors bigger, more balanced subgroups over those based on only the absolute difference between correlations:

$$\varphi_{ent}(D) = \varphi_{ef}(D) \cdot \varphi_{abs}(D)$$

5. Results

The results of two subgroup discovery runs conducted on the BCCWJ are presented in Table 2, both limited to the top 10 results for each run. The first run used φ_{ent} as the quality measure, while the second run used the unweighted φ_{abs} as the measure. All other parameters were kept constant during the two runs.

Table 2: *Results of top-q beam search using φ_{ent} and φ_{abs} as the quality measure (top 10 for each shown). Labels with no official English transcription translated by author.*

Quality measure ← φ_{ent} (Other parameters: q ← 40, beam width ← 10, search depth ← 8, minimum document coverage ← 20)							
Subgroup condition	φ_{ent}	φ_{abs}	φ_{ef}	r	r^C	G	G^C
CATEGORY="Yahoo! Q&A"	0.417	0.418	0.997	0.570	0.988	91,445	81,230
CATEGORY="Yahoo! Blogs"	0.293	0.330	0.887	0.647	0.977	52,680	119,995
TOPIC="literature and lit. criticism, etc." ∧ AUDIENCE="expert"	0.162	0.264	0.615	0.556	0.820	2,630	14,644
AUDIENCE="expert"	0.156	0.277	0.562	0.557	0.834	2,637	17,411
MEDIA="book" ∧ AUDIENCE="expert"	0.151	0.239	0.634	0.557	0.796	2,637	13,879
TOPIC="foreign literary novel" ∧ AUDIENCE="expert"	0.149	0.256	0.582	0.556	0.813	2,632	16,276
MEDIA="book" ∧ AUDIENCE="expert" ∧ TOPIC="medicine"	0.148	0.268	0.551	0.281	0.548	337	2,300
AUDIENCE="expert" ∧ TOPIC="medicine"	0.148	0.268	0.551	0.281	0.548	337	2,300
MEDIA="book" ∧ AUDIENCE="expert" ∧ TOPIC="elec. communication"	0.146	0.327	0.446	0.225	0.552	245	2,392
AUDIENCE="expert" ∧ TOPIC="elec. communication"	0.146	0.327	0.446	0.225	0.552	245	2,392
Quality measure ← φ_{abs} (Other parameters: q ← 40, beam width ← 10, search depth ← 8, minimum document coverage ← 20)							
Subgroup condition	φ_{abs}	φ_{ent}	φ_{ef}	r	r^C	G	G^C
CATEGORY="law"	0.833	0.017	0.021	0.166	0.999	346	172,329
TOPIC="foreign language"	0.753	0.012	0.016	0.113	0.866	29	20,019
MEDIA="comics" ∧ TOPIC="foreign language"	0.753	0.012	0.016	0.113	0.865	29	19,994
MEDIA="illustrated book" ∧ TOPIC="foreign language"	0.752	0.012	0.016	0.113	0.865	29	19,989
TOPIC="language learning"	0.735	0.013	0.017	0.130	0.866	32	20016
MEDIA="comics" ∧ TOPIC="language learning"	0.736	0.013	0.017	0.130	0.865	32	19,991
MEDIA="illustrated book" ∧ TOPIC="language learning"	0.735	0.013	0.017	0.130	0.865	32	19,986
CATEGORY="verse"	0.735	0.012	0.016	0.264	0.999	252	172,423
MEDIA="comics"	0.729	0.010	0.014	0.137	0.866	25	20,023
TOPIC="language learning" ∧ MEDIA="comics"	0.729	0.010	0.014	0.137	0.865	25	19,991

6. Discussion

Overall, the top scoring subgroup condition was CATEGORY, which is a joining of the media label with any hierarchical metadata that exist for it (see Table 1). While the Yahoo! Q&A and Blogs subcorpora displayed the largest differences in correlation to the global model when factoring in their relatively large size, the Law documents media displayed the greatest absolute difference. A perhaps more interesting subgroup condition found in these two lists is the AUDIENCE metadata extracted from the C-CODE, as it indeed describes an important element of register that is missing in otherwise more complete metadata such as the NDC hierarchy. The NDC or other media hierarchies did not feature strongly in the top results, though further examination of lower results did reveal more 3- and 4-part conjoined conditions.

7. Conclusion

Using both the metadata and the language data available within the BCCWJ, the present study conducted two pilot subgroup discovery experiments to uncover subgroups in the BCCWJ that exhibit divergent language usage based on the correlation differences between subgroups. However, the resulting subgroups from the pilot study revealed mostly those subsets of the data that were expected to exhibit specialized language, such as law documents or Internet data, with more complicated groupings of metadata not featuring prominently within the top results. Further experiments are needed to elucidate the relationships between the different metadata, both linear as well as hierarchical, within the BCCWJ.

8. Future Work

Future work should not only look to improve the subgroup discovery task or to improve feature extraction, but to also inform future annotation of corpus metadata, especially those metadata which help to uncover subgroups with divergent linguistic properties.

Finally, Wu, Markert, and Sharoff (2010) introduce a new method of quantifying genre hierarchies in terms of their visual and distributional imbalance based on tree balance entropy scores (pp. 755-757). Incorporating these measures could provide an additional quantification of subgroup distributionality alongside the entropy split function φ_{ef}, especially important in that it would permit comparisons between the different hierarchical metadata within the BCCWJ. These methods would also allow the hierarchies present in the BCCWJ to be compared to more established (from the point of view of suitability for register studies) metadata hierarchies such as those of the BNC Baby introduced in 2.2, or the Lancaster-Oslo/Bergen (LOB) Corpus[6].

Acknowledgements

This research was financially supported by JSPS Foreign Post-Doc Fellowship grant (no. P13303). A previous version of this paper was presented at the 19th Symposium of the Association for Database on the Humanities (Hodošček & Yamamoto, 2013).

References

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Dublin Core Metadata Element Set, Version 1.1. (2012, June 14). Retrieved October 30, 2013, from http://dublincore.org/documents/dces/

Duivesteijn, W. (2013). *Exceptional Model Mining* (Doctoral dissertation, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University).

Eckert, P. & Rickford, J. R. (2001). *Style and sociolinguistic variation*. Cambridge University Press.

Garbacz, P. (2006). An outline of a formal ontology of genres. In *Knowledge science, engineering and management* (pp. 151–163). Springer.

Gries, S. T. (2009). Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*. University of Liverpool.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on the World Wide Web* (pp. 211–220). WWW '07. Banff, Alberta, Canada: ACM. doi:10.1145/1242572.1242602

Hirst, G. (2009). Ontology and the lexicon. In *Handbook on ontologies* (pp. 269–292). Springer.

Hodošček, B. & Yamamoto, H. (2013). *The role of metadata in the Balanced Corpus of Contemporary Written Japanese in the analysis of register*. Presented at the 19th Symposium of the Association for Database on the Humanities, Ritsumeikan University.

[6] A list of categories used in the LOB corpus is available from http://icame.uib.no/lob/lob-dir.htm.

Irvine, J. T. (2001). “Style” as distinctiveness: the culture and ideology of linguistic differentiation. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 21–43). Cambridge University Press.

Langohr, L., Podpečan, V., Petek, M., Mozetič, I., Gruden, K., Lavrač, N., & Toivonen, H. (2013). Contrasting subgroup discovery. *The Computer Journal*, *56*(3), 289–303.

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *5*(3), 37–72. Retrieved from http://llt.msu.edu/vol5num3/pdf/lee.pdf

Maekawa, K. (2007). KOTONOHA and BCCWJ: Development of a Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the first international conference on Korean language, literature, and culture* (Vol. 2, pp. 158–177). Corpora and Language Research. Seoul.

Maruyama, T. (2012, March). The role of metadata in the analysis of large-scale corpora. In *Dai sankai kōpasu nihongo wākushoppu yokōshū [Proceedings of the 3rd Workshop on Japanese Corpus Linguistics]* (pp. 203–210). Dai sankai nihongo kōpasu wākushoppu [3rd Workshop on Japanese Corpus Linguistics]. Tokyo, Japan.

Orlikowski, W. J. & Yates, J. (1994). Genre repertoire: the structuring of communicative practices in organizations. *Administrative science quarterly*, 541–574.

Park, S.-H. & Fürnkranz, J. (2008). Multi-label classification with label constraints. In *Proceedings of the ECML PKDD 2008 Workshop on Preference Learning (PL 2008)* (pp. 157–171). Antwerp, Belgium.

Spärk Jones, K. (1972). A statistical interpretation of term importance in automatic indexing. *Journal of Documentation*, *28*(1), 11–21.

Tanomura, T. (2014). BCCWJ no shiryōteki tokusei: kōpasu rikai no jūyōsei [On the characteristics of the BCCWJ as material: The importance of understanding the corpus]. In *Kōpasu to nihongogaku [Corpus and Japanese linguistics]* (Vol. 6, Vols. 8). Kōza nihongo kōpasu [Corpus textbook series]. Asakura Publishing Co., Ltd.

Vander Wal, T. (2007). Folksonomy. Retrieved October 30, 2013, from http://vanderwal.net/folksonomy.html

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery* (pp. 78–87). Springer.

Wu, Z., Markert, K., & Sharoff, S. (2010). Fine-grained genre classification using structural learning algorithms. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 749–759). Association for Computational Linguistics.

Yates, J. & Orlikowski, W. J. (1992). Genres of organizational communication: a structurational approach to studying communication and media. *Academy of management review*, *17*(2), 299–326.

Yates, J., Orlikowski, W. J., & Okamura, K. (1999). Explicit and implicit structuring of genres in electronic communication: reinforcement and change of social interaction. *Organization science*, *10*(1), 83–103.

SVM を用いたインドネシア語連体従属接続詞の判定システム

Wahyu Purnomo（東京農工大学　工学部　情報工学科）†
古宮　嘉那子（東京農工大学　工学研究院　先端情報科学部門）
小谷　善行（東京農工大学　工学研究院　先端情報科学部門）

SVM-based System for the Determination of Adnominal Subordinating Conjunction in the Indonesian Language

Wahyu Purnomo (Department of Computer and Information Sciences, Faculty of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

インドネシア語において，いろいろな接続詞が存在する．その中で，連体従属接続詞「Yang」が最もよく使用されるものである．インドネシア語の「A Yang B」は，B が A を修飾しており，日本語の「B の A」,「BA」又は「B のほうの A」や英語の「BA」又は「A which (to be/ verb) B」に意味が似ている．A には，名詞や名詞フレーズが来る場合が多く，B には，名詞だけでなく動詞や形容詞や長いフレーズなどが来る可能性がある．

インドネシア語には「A Yang B」の他にも「AB」という書き方もある．しかし，B に来るものが長ければ「Yang」があった方が自然である．人間でも「Yang」の有無を判定するのは難しい．そのため，本稿では，文書から「Yang」の使用方法をコンピュータに学習させ，自動的に「Yang」の有無を判定させるシステムを提案する．コンピュータに手掛かりの素性を変えたりすることで，正解率が変わると見られた．

２．関連研究

（宋，浅原，古宮，小谷（2013））がインドネシア語の「Yang」に近い意味を持っている中国語の「的」の研究を報告している．（宋，浅原，古宮，小谷（2013））では，SVM を用いて中国語助詞の用法を解析した．その結果，コンピュータが判定した場合は正解率が 97.4%で，人間が判定した場合は正解率が 96.2%だった．

また，（竇，古宮，小谷（2012））では，中国語においてインターネット上でよく使用される表現を判定するシステムを提案した．（竇，古宮，小谷（2012））の結果では，100 文の語を対象にアンケートを用いて人間で判定する際に 42%の正解率が得られた．一方，同じ 100 文の語を SVM で判定する際に 92%の正解率が得られた．

３．連体従属接続詞の有無を判定するシステム

本システムは，大きく分けると文書をコンピュータに学習させる側の学習部分と学習させた上で未知の文書を判定する側の実行部分から成る．さらに，学習部分と実行部分のそれぞれは，入力文書を扱うコーパス部分，文書から手掛かりの素性を抽出する素性抽出部分，機械学習を行う SVM の部分から成る．

本システムのコーパス部分では品詞タグつきコーパスを扱う．素性抽出部分では，Python 2.7.4 で書いた素性ベクトルリストを作成し，SVM の部分では LIBSVM 3-17(Chang and Lin (2001))を利用した．

学習部分の構成は下記の図 1 のとおりである．学習部分では，コンピュータに学習させようとする品詞タグつきのインドネシア語コーパスを，素性ベクトルのリストを作成するプログラムに送り，素性の抽出を行う．このプログラムにより作成された，素性ベクトル

†50011268509@st.tuat.ac.jp

のファイルから SVM によって，モデルファイルが生成される．モデルファイルは次に述べる実行部分で利用される．

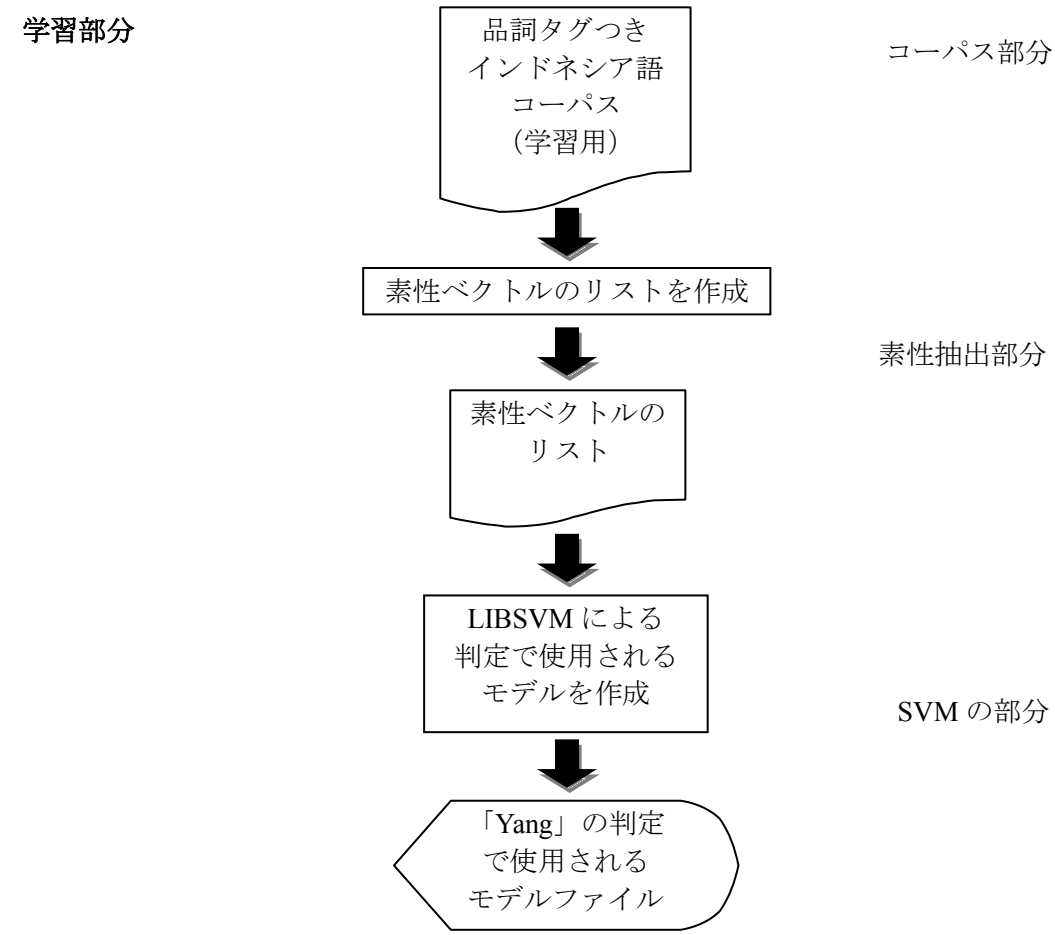


図 1　本システムの学習部分の構成

次に，実行部分では，判定のために，学習部分で使用されていない品詞タグつきコーパスを入力としている．学習の際と同様に，対象となるコーパスの素性抽出を行う．学習部分で生成されたモデルファイルと判定用の素性ベクトルリストのファイルに基づいて SVM による判定を行い，「Yang」の判定結果が得られる．本システムの実行部分は下記の図 2 のとおりである．

なお，本システムは「Yang」が名詞の直後に来ると想定しているため，コーパス中の名詞ごとに「Yang」の有無を判定した．

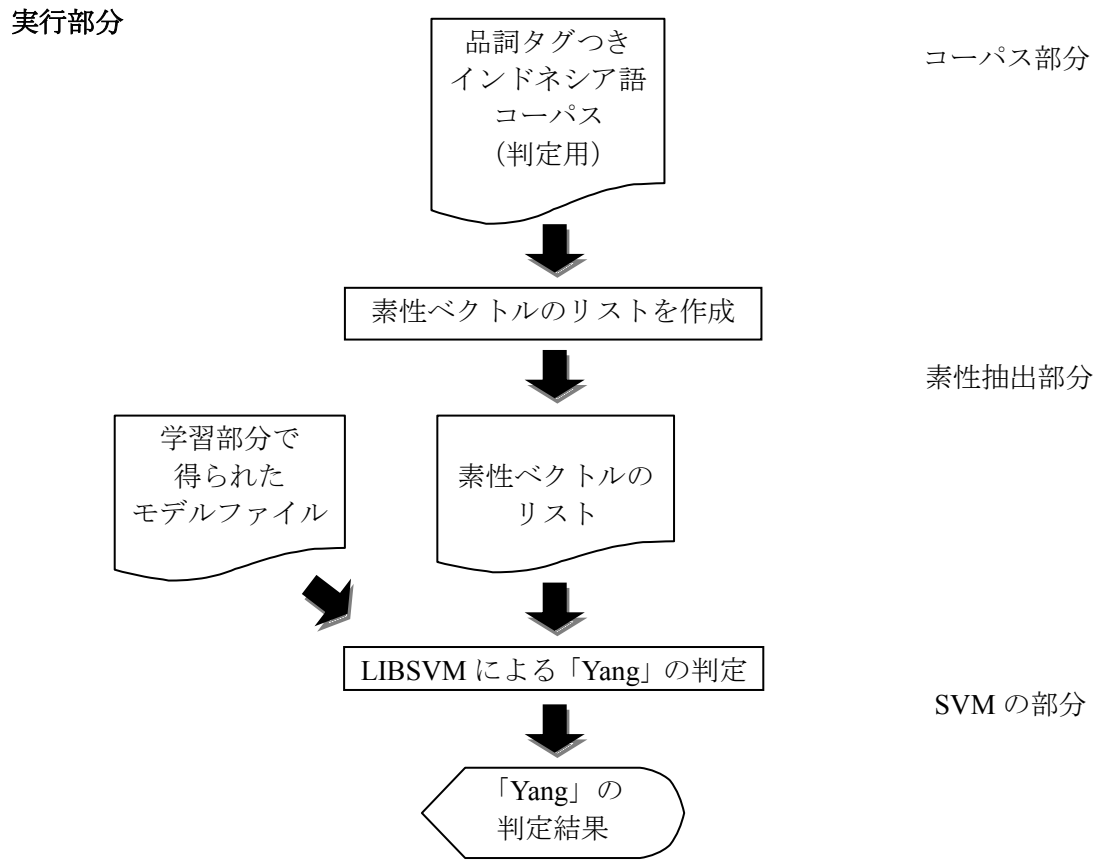


図 2　本システムの実行部分の構成

４．データ

本システムでは，One Million POS Tagged Corpus of Bahasa Indonesia[1]を利用した．このコーパスは, インドネシアの新聞から収集された 100 万単語のインドネシア語から成り，単語だけでなくその単語の品詞情報も付与されている．コーパスの例文を図 3 に示す．また，コーパス中の品詞の種類を表 1 に示す.

... pemegang/nnc saham/nnc publik/nnc ./. **Dan/nn** riil/nn **estat/nn** biasanya/jj tidak/neg diperdagangkan/nn dengan/in baik/jj dibawah/nn **kepemilikan/nn** publik/nnc ./. Salomon/nn **Brothers**/nn mengatakan/vbi ,/, Kami/prp yakin/nn **properti-properti/nn** riil/nn estat/nn akan/md ...

図 3　コーパスから取ってきた例文

[1] http://www.panl10n.net/english/OutputsIndonesia2.htm

表 1　コーパスで使用されている品詞の種類

タグ	説明	例
,	コンマ	, ;
.	文章の区切り	.
-	ダッシュ	-
SYM	記号	%
NN	普通名詞	Indeks, biaya, tenaga
PRP	人称代名詞	Kita, mereka, ia
PRN	数字代名詞	Satunya, keduanya, ketiganya
PRL	位置格代名詞	Sana, situ, sini
WRB	Wh 副詞	Apa, bagaimana, mengapa
WP	Wh 代名詞	Apa, apakah, apa-apa
VBI	自動詞	Ada, berakhir, berkata
VBT	他動詞	Membantu, menolak, menjadi
MD	助動詞	Akan, bias, telah
JJ	形容詞	Swasta, jauh, baik
CDP	基数	Satu, juta, milyar
CDO	順序	Pertama, kedua, ketiga
NEG	否定詞	Belum, bukan, tidak
IN	前置詞	Dengan, kepada, untuk
CC	等位接続詞	Atau, dan, karena
SC	従属接続詞	Bahwa, sekaligus, yang
RB	副詞	Hanya, mungkin, sebagaimana
DT	限定詞	Ini, para, tersebut
FW	洋語	Few, fiscal, for

また, 名詞には細分類がある．表 2 に細分類を示す.

表 2　名詞の細分類

タグ	説明	例
NNC	加算普通名詞	Cara, laut, tahap
NNU	不加算普通名詞	Peringatan, pikiran, system
NNG	属格普通名詞	Adanya, lainnya, misalnya
NNP	固有普通名詞	Desa, dunia, lembaga

５．　実験

本システムでは素性に単語の位置を考慮した．単語列 W：{A, B, C, D, E, F, G} とこれに対応する品詞 P：{p_A, p_B, p_C, p_D, p_E, p_F, p_G} があるとする．ここで，p_D を名詞とすると，「A B C D **Yang** E F G」という文章において，名詞の D の位置を 0，A の位置を-3，B の位置を-2，C の位置を-1，E の位置を+1，F の位置を+2，G の位置を+3 と置き, A, B, C, E, F, G の形態素と品詞を素性とした．なお,「Yang」の有無を判定するシステムであるため, 実際に E の位置に「Yang」があった場合でも素性には含まなかった．また,「Yang」が挿入される直前の名詞の形態素は素性に含めたが, 品詞は常に名詞であるため, 素性に含めなかった.

また，本システムでは関連性の低い単語を素性ベクトルに入れないために，文節（文の始まりまたは終わりやコンマの前または後）を超えた単語およびその品詞は素性に含めず,

それを超えた形態素は「NONE/none」として扱った．また，文頭と文尾は特殊な形態素として扱った．

本稿では，ふたつの実験を行った．一つ目は素性の種類を変えた実験である．具体的には，形態素だけを素性にした場合と，形態素に加えて品詞を素性にした場合の実験を行い比較した．

図 3 の例文を例に，形態素だけを利用した場合の「estat/nn」の素性ベクトルのパターンを表 3 に，形態素と品詞を利用した場合の「estat/nn」の素性ベクトルを表 4 に示す．

表 3　「estat/nn」の単語だけの素性ベクトル

「Yang」の有無	単語の位置						
	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
ない	.	Dan	riil	estat	biasanya	tidak	diperdagangkan

表 4　「estat/nn」の単語と品詞が入った素性ベクトル

「Yang」の有無	単語の位置							品詞の位置					
	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**
ない	.	Dan	riil	es tat	bia sa nya	ti dak	Diperda gangkan	bos	nn	nn	jj	neg	nn

二つ目の実験として，ウィンドウサイズを変えた実験を行った．連体従属接続詞「Yang」は，直後に長いフレーズが，また直前にはより短いフレーズが来る可能性が高い．そこで，ウィンドウサイズを変えて，本システムの正解率がどれぐらい変わるのかを確認した．この際，問題の名詞の位置より前の単語は 3 個に固定し，後ろの単語として 3 個と 5 個単語の両方を試した．

６．実験の結果

実験で使用したコーパスには，42,451 種類，計 451,339 個の名詞を含んでいた．そして，その中から「Yang」がつくものは 17,588 個であった．これは全体の 3.90%を占めているため，69.10%が最頻出ベースラインである．実験の結果は表 5 のとおりである．

表 5　実験の結果

実験		正解率 [%]
素性の種類	単語のみ	96.90
	単語と品詞	97.07
ウィンドウサイズ	前 3 後 3	97.07
	前 5 後 5	97.37

７．考察

単語と品詞の情報を素性ベクトルに入れた場合は単語のみを入れた場合より正解率が高いことが表 5 から分かる．すなわち，コンピュータに単語だけ覚えさせるのでなく，その単語の品詞も覚えさせたほうがより効果的であることが分かった．

表 5 より，ウィンドウサイズは 3 よりも 5 のときの方が正解率が高いことが見て取れる．この差は 0.30%であった．

次に，「Yang」がある際にシステムが間違って判定し 0 と出力した例を(1)と(2)に示す．

(1)「…naik pada *tingkat* **yang** jauh lebih cepat…」
「…より速いペースの段階に上がった…」
(2)「…dan *12 bulan* **yang** berakhir pada September 1988…」
「…1988 年 9 月に終わる 12 カ月や…」

上記の(1)と(2)では，「Yang」がなければ文章の構成が変わる他に，文章の意味もおかしくなるため，システムが 0 と出力する場合は，判定が間違っている．(1)における「tingkat（段階）」と(2)における「12 bulan（12 カ月）」の使用例が学習データになかったため，システムが「Yang」がないと判定したと考えられる．

コーパスを見ると，「Yang」が実際にはあるのにもかかわらず，システムはないと判定している．これは，学習データのうち，「Yang」があるデータが非常に少ないためであると考えられる．これは「oversampling」や「undersampling」によってデータの割合を変えることで改良の可能性があるが，（宋，浅原，古宮，小谷（2013））によれば，データの割合を変えない方が良い結果が出ている．

８．まとめ

本稿では，インドネシア語連体従属接続詞「Yang」の有無を判定するために，名詞の直前にある 3 個単語と名詞の直後にある「Yang」を含めず 5 個単語を素性ベクトルに入れる場合は正解率が最も高いと見られた．そして，ピリオードやコンマにより関連性の低い単語を素性ベクトルに入れないことの重要さが見られた．

文献

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

宋 東旭，浅原 正幸, 古宮 嘉那子，小谷 善行（2013），機械学習による中国語助詞の用法解析，第三回コーパス日本語学ワークショップ予稿集，pp. 111-116.

寶 梓瑜，古宮 嘉那子，小谷 善行（2012），コーパスを用いた中国語ネット語の判定システム，　第一回コーパス日本語学ワークショップ予稿集, pp.161-166.

合議による語義曖昧性解消の領域適応のための確信度の調整

古宮　嘉那子（東京農工大学　工学研究院）†
奥村　学（東京工業大学　精密工学研究所）
小谷　善行（東京農工大学　工学研究院）

Adjustment of Confidence Score for Domain Adaptation in Word Sense Disambiguation Based on the Comparison of Multiple Classifiers

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

１．はじめに

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

（Komiya and Okumura (2012)）では，語義曖昧性解消（Word Sense Disambiguation，WSD）の領域適応の適切な手法は用例によって異なると考え，supervised な領域適応において，分類器の出力する確信度の高い方の答えを採用することにより，分類の精度を向上させる手法を示した．また，（古宮，奥村，小谷 (2013)）では，同様の手法が unsupervised な領域適応においても効果を発揮することを示した．しかし，この手法は，分類器ごとの訓練事例数に大きなばらつきがある際には効果が上がらないという問題があった．そのため，本稿では，合議による語義曖昧性解消の領域適応のために分類器間の確信度を調整することを提案する．unsupervised な領域適応において，確信度に対してコーパス中に出現した語義数により調整を行った結果，正解率に有意な上昇が見られた．

２．関連研究

領域適応は，学習に使用する情報により，supervised，semi-supervised，unsupervised の三種に分けられる．本研究で扱うのは，unsupervised の領域適応，つまりラベルつきのソースデータとラベルなしのターゲットデータを利用するものである．

（古宮，奥村 (2012)）は WSD について supervised な領域適応を行った場合，最も効果的な領域適応手法はソースデータとターゲットデータの性質により異なることを示し，最も効果的な領域適応手法を，WSD の対象単語タイプ，ソースデータ，ターゲットデータの三つ組ごとに自動的に選択する手法について述べた．

また，（Komiya and Okumura (2012)）は，WSD の supervised な領域適応において，本稿でも使用する確信度という尺度を用い，用例ごとに訓練事例を自動的に選択した．（古宮，奥村，小谷 (2013)）は（Komiya and Okumura (2012)）の手法が unsupervised の領域適応に対しても有効であることを示している．本稿はこの（古宮，奥村，小谷 (2013)）についての改良を行う．

最後に，（古宮，小谷，奥村 (2013)）は，unsupervised な領域適応において，あるターゲットデータに対して複数のジャンルのソースデータが混在した場合，確信度と LOO-bound という指標を利用して，領域適応のための訓練事例の部分集合を WSD の対象単語タイプごとに自動的に選択する手法について述べた．

†kkomiya@cc.tuat.ac.jp

３．合議のための確信度の調整

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

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

ここでの分類器のスコアには確信度（Komiya and Okumura (2012)）を利用し，その値を比較することで，複数の分類器の合議を行う．ここで，確信度とは分類の確からしさを表す 0〜1 の値であり，active-learning においてラベル付けする用例を選択するのによく利用される．

しかし，（古宮，小谷，奥村　(2013)）で示されたように，この手法は，分類器ごとの訓練事例に大きなばらつきがある際には効果が上がらないという問題があった．これは，個々の分類器の訓練事例のみを基準として出力される確信度という値を，分類器間の出力結果を比較するために用いているためである．そのため本稿では，分類器間の比較に耐えるように確信度を調整する手法として，以下の三つを提案し，比較する．

・分類器が出力した語義の，訓練事例における事前確率でスコアを割る（Prior）．
・分類器の訓練事例における最頻出語義の事前確率でスコアを割る（MFS）．
・分類器の訓練事例に出現する語義数をスコアにかける(SNum).

Prior は，事前確率が高い語義を回答する分類器と，低い語義を答える分類器では，同じ確信度でも，事前確率の低い語義を答える分類器をより評価するべきだとする手法である．また MFS は同様に，最頻出語義の事前確率が低いデータで学習した分類器の確信度の方を，最頻出語義の事前確率が高いデータで学習した分類器の確信度よりも評価すべきだとする手法である．最後に SNum は，訓練事例に現れる語義が多いデータで学習した分類器の確信度の方を，訓練事例に現れる語義が少ないデータで学習した分類器よりも評価すべきだとする手法である．

４．実験

分類器としてはマルチクラス対応の SVM（libsvm）(Chang and Lin (2001))を使用した．また，libsvm の確率として出力される分類の確からしさを確信度として用いた．カーネルは予備実験の結果，線形カーネルが最も高い正解率を示したため，これを採用した．また，WSD の素性には，以下の 17 種類の素性を用いた．

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

分類語彙表の分類コードには（国立国語研究所 (1964)）を使用した.
また，実験は五分割交差検定を用いた.
また，本研究では二つのコーパスが与えられたと仮定し，それぞれのコーパスを使った場合と，さらに入手可能な二つのコーパスをすべて使った場合の三通りのうちから WSD のための訓練事例集合を選択した．また，合議の際には，最も高い確信度の分類器の結果（語義）を採用した.

５．実験データ

実験には，現代日本語書き言葉均衡コーパス（BCCWJ コーパス）(Maekawa (2008)) の Yahoo!知恵袋，白書，Yahoo！ブログ，書籍，雑誌，新聞のコアデータ[1]と，コアデータではない白書と Yahoo! 知恵袋のデータ，また RWC コーパスの毎日新聞コーパス（Hashida et al. (1998)）の合計 9 つのデータを利用した．これらのデータには岩波国語辞典（西尾ら (1994)) の語義が付与されている. 九つのコーパスのうち, ひとつをターゲットデータにし，残りのうちの二つを利用可能なソースデータとして利用して領域適応の実験を行った．この際，使用した三つのコーパス中全てに 50 トークン以上存在する単語を実験対象とした. 白書と Yahoo! 知恵袋の重複利用を避け, また実験対象となる単語が存在しない組み合わせを除くと，コーパスの選び方は全部で 171 通りとなった．また，対象単語の種類数はコーパスの選び方により変わるが，全ての論理和をとると，49 種類となった.
それぞれのデータにおける 50 用例以上の単語数, 単語ごとの用例数の平均と標準偏差を表 1 に示す.

表 1　それぞれのデータにおける 50 用例以上の単語数，単語ごとの用例数の平均と標準偏差

コーパスの種類	単語	平均	標準偏差
コア　Yahoo！知恵袋	22	157.77	153.77
コア　白書	5	79.20	21.01
コア　Yahoo！ブログ	9	245.22	431.84
コア　書籍	35	158.91	204.97
コア　雑誌	26	284.92	872.53
コア　新聞	25	92.28	78.08
非コア　白書	38	2,088.84	2,234.52
非コア　Yahoo！知恵袋	42	3,979.17	5,786.02
RWC　新聞	66	473.79	1,844.40

また，実験には岩波国語辞典の中分類の語義を採用した．語義数ごとの単語の内訳は，2 語義 :「一般」,「書く」,「考える」,「技術」,「経済」,「現在」,「子供」,「自分」,「情報」,「高い」,「作る」,「強い」,「電話」,「場合」,「他」，3 語義 :「言う」,「今」,「入れる」,「買う」,「関係」,「聴く」,「社会」,「進む」,「地方」,「出来る」,「出る」,「入る」,「開く」,「前」,

[1]コアデータは形態素解析の結果を人手で訂正しているため正確であるが，量が少ない.

「求める」，4 語義：「訴える」，「時間」，「時代」，「出す」，「乗る」，「計る」，「ひとつ」，「見える」，「持つ」，5 語義：「進める」，「やる」，「良い」，6 語義：「会う」，「見る」，「もの」，7 語義：「手」，8 語義：「する」，「取る」，9 語義：「上げる」である．

６．結果

表 2 に使用したデータの区分と適応手法別の実験結果を示す．このうちの「9 つのコーパス」が 9 つのコーパス全ての結果の平均である．また，「コアデータ」には 6 つの BCCWJ のコアデータの結果の平均を，「コアデータ以外」には BCCWJ のふたつのコアデータ以外のデータと RWC コーパスの結果の平均を示した．

表において，「Self」は，タグつきターゲットデータが手に入ったと仮定して，supervised の学習を 5 分割交差検定を用いて行った結果である．「平均的なコーパス」は，ふたつのジャンルのソースデータそれぞれをジャンルごとに分けて訓練事例とした場合の結果の平均である．入手可能なコーパスをそれぞれソースデータとして使用した場合の平均的な結果を示している．例えば，Yahoo！知恵袋のデータがターゲットデータの時のソースデータが白書と新聞だった場合，このときの「平均的なコーパス」は，白書の全データで訓練した Yahoo！知恵袋のデータの正解率と，新聞の全データで訓練した Yahoo！知恵袋のデータの正解率の平均となる．最後に，「全てのコーパス」とは，ふたつのジャンルのソースデータ全て（つまり全ソースデータ）を訓練事例とした際の結果である．例えば，上記の例において，Yahoo！知恵袋のデータがターゲットデータの時の「全てのコーパス」は，白書と新聞のコーパス全てを訓練事例として利用した際の結果である．

このとき，「Self」は upper bound であり，「平均的なコーパス」，「全てのコーパス」はベースラインである．表において，これらのうち Self 以外でコーパスごとに一番高い正解率を太字で示した．

表 2　コーパスと適応手法別の実験結果

データの種類	9 つのコーパス		コアデータ以外		コアデータ	
手法	マイクロ	マクロ	マイクロ	マクロ	マイクロ	マクロ
Self	92.35%	86.44%	91.89%	85.15%	90.07%	85.21%
平均的なコーパス	71.33%	75.47%	80.53%	79.07%	68.77%	77.53%
全てのコーパス	74.60%	77.97%	84.57%	83.31%	79.27%	80.92%
確信度	74.59%	77.92%	83.00%	81.85%	77.34%	79.85%
確信度 Prior	70.86%	75.98%	83.71%	82.06%	66.10%	77.16%
確信度 MFS	72.22%	77.33%	83.62%	81.76%	69.84%	79.26%
確信度 Snum	74.85%	78.13%	84.68%	82.51%	77.54%	80.23%

７．考察

まず，表 2 を見てみると，9 つのコーパス全ての平均では，マイクロ平均でもマクロ平均でも確信度に語義数をかけて調整する手法，「SNum」が最も良いことが分かる．コアデータ以外のデータの平均では，マイクロ平均では「SNum」が最も良いが，マクロ平均ではベースラインである「全てのコーパス」，コアデータの平均ではマクロ平均でもマイクロ平均でも「全てのコーパス」が最も良い．また，「Prior」，「MFS」の調整は効果がなかった．

一方，表 1 を見てみると，コアデータは用例数の平均がコアデータ以外のデータに比べて低いことが分かる．確信度には分類器間で訓練事例数に大きなばらつきが出た場合に効果があがりにくいという欠点があったが，これは訓練事例数が少ない際に，確信度の正確性が低くなるためである．たとえば，訓練集合の用例数が 1 件になると，他に選択肢がないため，その分類器の確信度は最高値の 1 となる．しかし 1 件の訓練事例による分類器が最も良い分類器であるとは考え難く，訓練事例数の少ない場合の確信度はあまり信用できないことが分かる．このように，確信度は訓練事例数が少ないと，多い場合よりも高くなる場合がある．これは直感に反した現象であり，少ない訓練事例数の分類器を比較対象に含める場合には，確信度は不利であると言える．このことから，コアデータは訓練事例数が少ないため，確信度の正確性が低くなり，「全てのコーパス」に及ばなかったと考えられる．

また同じく表 1 から，コアデータ以外のデータは標準偏差がコアデータに比べて非常に高いことが分かる．このことから，コアデータ以外のデータでは，コアデータよりも分類器間で訓練事例数にばらつきが出る傾向が見て取れる．また，コアデータとそれ以外のデータではもっとその傾向が強まると思われる．そのため，9 つのコーパスの平均とコアデータ以外のデータの平均では，分類器間の整合性を調整する手法が効いたと考えられる．なお，9 つのコーパスマイクロ平均での「SNum」と「全てのコーパス」の差は有意水準は 0.05 のとき，カイ二乗検定で有意であった．これらの結果から，確信度は訓練データの少ない分類器においてはその正確性が薄れるため，訓練事例数が少ない二つのコーパスに対しては全データを利用した方が WSD の正解率が高くなること，また，訓練事例数の少ないコーパスと訓練事例数が多いコーパスを比較する場合には，確信度の整合性を語義数で調整することで WSD の正解率が上昇することが分かった．

また表 2 において，先行研究（古宮，奥村，小谷　(2013)）や（Komiya and Okumura (2012)）の結果に反して，ベースラインである「全てのコーパス」が強く，調整を行わない，確信度単独による手法よりも正解率が高くなっているが，これはコアデータにおいては，使用できる訓練事例数が少なくなったためであり，またそれ以外のデータにおいては分類器間の訓練事例数の差が増したためであると考えられる．

８．まとめ

本稿では，unsupervised な領域適応において，あるターゲットデータに対して複数のジャンルのソースデータが混在した場合，用例ごとに最もよい訓練事例集合で学習された分類器を，分類器の確信度を指標として選択する手法を改良した．分類器間の確信度を，コーパス中に出現した語義数により調整を行った結果，正解率に有意な上昇が見られた．

謝　辞

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

文献

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

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

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

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

国立国語研究所（1964). “分類語彙表”. 秀英出版.

古宮嘉那子，奥村学（2012). “語義曖昧性解消のための領域適応手法の決定木学習による自動選択”，自然言語処理，Vol.19, No.3, pp.143-166.

古宮嘉那子，小谷善行，奥村学（2013). “語義曖昧性解消の領域適応のための訓練事例集合の選択”，第19回言語処理学会年次大会予稿集，pp.940-943.

古宮嘉那子，奥村学，小谷善行(2013). 分類器の確信度を用いた合議制による語義曖昧性解消の semi-supervised な領域適応. 第三回コーパス日本語学ワークショップ予稿集, pp. 1-6.

西尾実，岩淵悦太郎，水谷静夫（1994). “岩波国語辞典第五版”．岩波書店.

CRF による Web コーパスからの アニメタイトルに対する訳語の自動抽出

山崎 舞子（東京農工大学）
森田 一（東京工業大学）
古宮 嘉那子（東京農工大学）
小谷 善行（東京農工大学）

Automatic Extraction of the Translation of *Anime* Titles from Web Corpora Using CRF

Maiko Yamazaki (Tokyo University of Agriculture and Technology)
Hajime Morita (Tokyo Institute of Technology)
Kanako Komiya (Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Tokyo University of Agriculture and Technology)

１．はじめに

日本のアニメや漫画が海外でも人気となるにつれ，その関連商品も注目を集めている．しかし，海外の人が日本語サイトで売られている商品を検索するのは困難なことである．さらに，このような分野では次々に新しい用語が作り出されるため，静的な辞書で対応するのは難しい．本稿では，音訳値や距離，品詞などを素性として，Conditional Random Fields (CRF) により英語アニメタイトルに対する日本語の訳語候補を抽出する手法を提案する．この際， Web コーパスを使用し，日本語アニメタイトルが英語に翻訳される際，音訳されているものが多く含まれていることに着目した．

２．関連研究

現在，未知語に対する訳語の自動抽出ではパラレルコーパスが必要となるが，使用できる資源が限られている．そこで，Web 上のテキストを用いた訳語抽出が注目されている．Web コーパスからの未知語に対する訳語抽出に関しては，これまで多くの研究が行われてきた (Zhang (2009))．

一方，日本のサブカルチャーは世界で注目を集めており，アニメ単語の抽出が商業的に有用である．アニメ用語に着目した固有表現抽出の研究として CRF によりアニメ関連用語の固有表現抽出を行った高瀬らの研究がある (高瀬 (2013))．

本稿では，英語アニメタイトルに対する日本語の訳語候補を自動抽出する．

本研究と関わりの深い研究として，Chang(2013)の研究がある．Chang(2013)は検索エンジンによって得たスニペットをもとに訳語抽出を行い，音訳，翻訳，距離を素性とし，CRF により訳語候補を得て，出現頻度によってランク付けを行う手法を提案している．

先行研究がほとんど中国語-英語間で行われているのに対し，本研究では日本語-英語間の対訳対を対象とする．そのため，CRF でタグ付けする際，中国語では文字 (漢字) 単位でラベル付けを行うのに対し，本研究では形態素解析を行い，形態素ごとにラベル付けを行う．また，アニメタイトルを対象としているため，英語が元にあって和訳した対訳だけでなく，日本語が元にあって英訳した対訳対も多い．日本語が元にあって英訳された訳語は音訳されていること

が多いといった特徴があるため，音訳素性に着目した素性設計を行っており，カタカナやローマ字表記も考慮した．

３．システムの概要

システムは以下の手順により英語アニメタイトルに対する日本語の訳語候補の出力を得る．

1. 入力された英語アニメタイトルをクエリとして，検索エンジンによる検索結果上位最大100 件の Web ページのテキストデータ取得を行う
2. １で取得した各テキストデータに対して形態素解析を行い，各形態素に対して 11 種類の素性値を作成する．素性値については後述する
3. あらかじめ学習されたモデルによって各形態素へのラベル付けを行い，訳語候補を抽出する

４．素性の設計

本システムでは，各形態素に対する CRF の素性として以下の 11 素性を用いた．
このうち日本語読み音訳素性値と英語読み音訳素性値については４．１節，４．２節で詳しく説明する． また，音訳距離とは文字列の音としての近さをはかり，小さいほど一致度が高い値とする．

1)．表層
2)． 品詞
3)． 品詞細分類
4)． 文字種（「ひらがな」「カタカナ」「漢字」「アルファベット」「その他」）
5)． 英語タイトルと今みている形態素の位置の差（負の数も可）
6)． 翻訳辞書にのっているか否か（0 か 1）
　翻訳辞書には Wikipedia 日英京都関連文書対訳コーパス[1]を元に GIZA++[2]を用いて翻訳確率を計算後，0.1 以上の確率がついたペアが格納されている
7)．日本語読み音訳素性値
　今みている形態素をローマ字に変換した文字列と英語タイトルの各単語の編集距離を元にした音訳距離．0 から 10 の値
8)．英語読み音訳素性値
　今みている形態素をローマ字に変換した文字列と英語タイトルの各単語との動的計画法によるコストを元にした音訳距離．0 から 10 の値
9)．音訳素性値のうち小さい方
10)．括弧の中に存在するか否か（0 か 1）
11)．英語タイトルか否かのタグ

４．１　日本語読み音訳素性値

アニメタイトルには「君と僕（Kimi to Boku)」など，日本語がそのまま音訳されているも

[1] http://alaginrc.nict.go.jp/WikiCorpus/
[2] http://www.statmt.org/moses/giza/GIZA++.html

のが多く存在する．このような訳語を抽出するために，文字列の一致度をはかる日本語読み音訳素性値を導入する．

日本語読み音訳素性値は，英語タイトル中の各単語または連続する二単語と，今見ている形態素，またはその前後と連結した文字列の各組み合わせについて音訳距離を計算し，その最小値を素性値とする．

日本語読み音訳素性値の計算例として，英語タイトル「Dirty pair」として取得した文章「キディグレイド/と/ダーティペア/の/比較」(/形態素の区切りを表す)の中の「ダーティペア」の部分の素性値の計算方法を説明する．

「ダーティペア」はまず，「da-teipea」とローマ字に変換される．その後英語タイトル中の各単語と，隣り合った二つの単語の合成語，すなわち「Dirty」「Pair」「DirtyPair」の三つと「da-teipea」との編集距離がそれぞれ計算される．その中で一番小さい値である「da-teipea」と「DirtyPair」の編集距離 5 を選び，この値が「da-teipea」の文字列の長さである 9 で割られ，0.56 という値が保持される．次に今みている形態素と一つ前の形態素を足しあわせ，「とダーティペア」という形態素を「toda-teipea」とローマ字に変換し，英語タイトルの各形態素との編集距離を計算し，一番小さい「DirtyPair」と「toda-teipea」の編集距離 6 が選ばれる．これを文字列の長さ 11 で割った 0.55 という値を得る．保持しておいた値と比べると今出した値のほうが小さいので，これを 10 倍して四捨五入し，「ダーティペア」の日本語読み音訳素性値は 6 となる．

英語タイトル中の各単語だけでなく，隣り合った二つの単語の合成語を利用することで，この例のように日本語では一形態素だが英語では二単語に分かれるような場合にも対応できる．また，一つ前の形態素と今見ている形態素を足し合わせることで，月/姫（Tsukihime）のように日本語では二形態素に分割できるが，英語では一単語となるような場合にも対応した．

形態素がアルファベットだった場合は，そのままの形態素と英語タイトルに含まれる形態素それぞれとの編集距離をとることで，英語と日本語が混ざった訳語（ビッグ X など）にも対応した．

４．２　英語読み音訳素性値

「フィリックス（firikkusu)」と「Felix」の日本語読み音訳素性値を考えた場合，同じものを指しているのに編集距離が大きくなってしまう．このような英語の音をそのまま日本語にしたような訳語を抽出するために，英語読み音訳素性値を導入する．英語読み音訳値は，今みている形態素と英語タイトルの各単語との音としての近さをはかることができる．

英語読み音訳素性値は，今見ている形態素をローマ字に変換したものと，英語タイトルの各単語を動的計画法を用いて対応づけるが，この手法は動的計画法により音素とカタカナを対応付けた Tsuji（2012）の手法と関連がある．

なお，英語読み音訳素性値は注目している形態素の文字種が日本語だった場合のみ計算する．

英語タイトルに含まれる単語のうち一つを英語読み文字列，注目している形態素を日本語読み文字列とする．両者の文字列が一致した場合と英語読み変換規則による文字列の一致がおきた場合にコストは 0 となる．それ以外の場合はコストは 1 に設定した．英語読み変換規則として，英語を日本語に音訳する際の簡単な規則（igh は ai と読む，など）を作成した．

「ドロップ」と「drop」の動的計画法による対応の例を図 1 に示す．

英語読み変換規則には日本語文字列「do」と英語文字列「d」の対応，「p」と「pu」の対応が

記載されているとする．

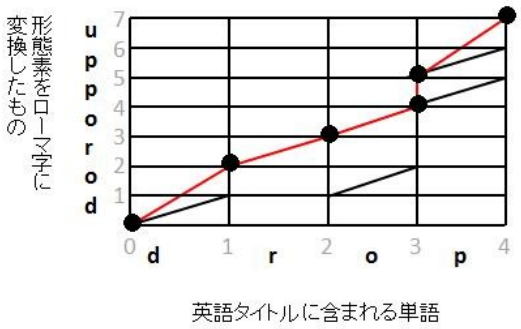


図１．動的計画法による対応付け

まず，「ドロップ」は日本語であるため，これをローマ字に直し，「doroppu」とする．英語読み文字列に「drop」を，日本語読み文字列に「doroppu」をセットする．一番左下の 0 の点から対応付けをスタートする．0 の点からは，次の 4 つの点へのパスが与えられる．座標は（英語読み文字列，日本語読み文字列）とする．

・日本語読み文字列「d」の対応がないとする(1, 0)
・英語読み文字列「d」の対応がないとする(0, 1)
・日本語読み文字列「d」と英語読み文字列「d」の一致による(1, 1)
・辞書に記載されている英語読み文字列「d」と日本語読み文字列「do」の対応による(1, 2)

0 の点から(0, 1)または(1, 0)の点に移動するにはコストが 1 かかり，(1, 1)または(1, 2)に移動するにはコストはかからない．移動する際にはそれぞれの点までのコストも記憶しておく．

次は上で求めた 4 点から移動できるパスをそれぞれ導出する．それぞれの点から移動できるパスを求め，コストを記憶していくことを繰り返し最終的に一番コストの低い経路が日本語読み文字列と英語読み文字列の対応として選ばれる．例の図 1 の場合は最終的に 0→(1, 2)→(2, 3)→(3, 4)→(3, 5)→(4, 7)の経路が選ばれ，コストは(3, 4)から(3, 5)に移動する際にかかる 1 のみとなり，これを以下の式によって計算した値である 1 が「ドロップ」と「drop」の対応を見た場合の英語読み音訳値となる．

$$\text{英語読み音訳値} = \frac{\text{コスト}}{\text{今見ている形態素の長さ}} \quad \cdot \quad 10$$

５．評価実験

５．１　実験データ

2012 年の英語版 Wikipedia ダンプデータ[3]より，日英アニメタイトルの抽出を行った．これにより 2134 対のデータを得た．

実験ではこのタイトル対をクエリとして用いて，言語指定を日本語にして検索を行い，各タイトルについて Web ページ最大 100 件を取得した．検索エンジンとしては Google[4]を利用した．

[3] http://dumps.wikimedia.org/enwiki/
[4] http://www.google.com/

５．２　実験設定

実験を行うにあたって対象とするWebページを以下のように限定した.

まず,各Webページは一つの英語タイトルと,少なくとも一つの正解の訳語を持つこととし,それ以外のページは削除した.今回の実験では，上記の条件を満たすWebページが一件でも存在し，なおかつ日英で同一のタイトルでない1282タイトルを用いた.

一つのWebページに対して，複数の英語タイトルが存在した場合，一番はじめに出てくる英語タイトル以外については英語タイトルであるというタグを消すことにより対象となる英語タイトルを一つにする操作を行った.

また,英語タイトルと距離が離れすぎている場所には訳語があらわれにくいという考えに基づき，英語タイトルから25形態素以内に限定して訳語候補の抽出を行った.

実験において形態素解析器にはMeCab[5]を，CRFの実装にはCRF++[6]を利用した.

５．３　評価方法

システム全体としては，Wikipediaから抽出した日本語タイトルを正解タイトルとし，正解タイトルとシステムが出力した訳語を比較して，完全に一致するかどうかで

システム全体の評価を行う.一つのタイトルにつき,少なくとも一つの訳語候補を出力した.すべてのタイトルのうち,訳語候補の中に一つでも正解タイトルが含まれているタイトルの割合を示したものをカバレージとし，これを評価した.

また，参考として，正解タグとシステムが出力したタグを比較して再現率，精度，F値を求め，評価した.

本実験では，すべての素性を用いたモデルと，すべての素性から「英語読み音訳値」,「日本語読み音訳値」,「音訳値のうち小さい方」の3つの音訳に関わる素性を除いたものとの比較を行う.

５．４　実験結果

システムのカバレージを表１に，タグの評価結果のマイクロ平均，マクロ平均を以下の表２に示す.

表１．カバレージ

素性	システムがとれたタイトル	カバレージ
全部	922	0.719
-音訳	806	0.629

表２．マイクロ平均・マクロ平均

	素性	再現率	精度	F値
マイクロ平均	全部	0.514	0.739	0.606
	-音訳	0.352	0.741	0.477
マクロ平均	全部	0.488	0.588	0.533
	-音訳	0.335	0.531	0.411

[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6] http://crfpp.googlecode.com/svn/trunk/doc/index.html

表１より，すべての素性から音訳を抜いた場合，取得できた訳語の数は 10 ポイント近く下がっていることから，従って，音訳素性は訳語抽出において有効であることがわかる．

また，音訳を抜いた場合では，マイクロ平均，マクロ平均ともに精度がほとんど変わらない一方，再現率が悪くなることがわかる．

６．考察

音訳素性の追加によって抽出できるようになったタイトルとして，「(Bismark，星銃士ビスマルク)」，「(Valkyria Chronicles，戦場のヴァルキュリア)」などがあげられる．これは二言語のタイトルに音的なつながりが存在するため，音訳素性が有効に働いたためと考えられる．

逆に，抽出できなくなったタイトルとしては，「(Super big，とんでぶーりん)」や「(Porco Rosso，紅の豚)」などがあげられる．これらは音的なつながりがないため，品詞のつながりや距離などから抽出しなければならないが，学習の結果音訳の重みが大きくなったため，抽出できなくなったと考えられる．

また，「・・・(アニメ) ヤマトよ永遠に　やまとよとわに・・・」というテキストから「BE FOREVER YAMATO」に対する訳語を抽出した際，正解タイトルは「ヤマトよ永遠に」であるが「ヤマトよ永遠にやまと」という誤った出力が得られた．

これは，「YAMATO」に対応する形態素が「ヤマト」と「やまと」の二つ存在し，ともに音訳距離が小さいため，同じ音をあらわす二つの形態素を含む訳語を抽出してしまったと考えられる．このような，英語タイトル中のある部分の読みに対応する日本語の形態素が複数存在する場合を許さないようにすることで，改善できる可能性がある．また，今回の実験では，抽出できたタイトルが「借りぐらしのアリエッティ」における「アリエッティ」のような一部分では不正解としたが，商品検索などの実用上は十分に目的を果たせる場合があるだろう．

７．おわりに

本稿では，英語アニメタイトルに対する日本語の訳語抽出において，音訳に着目した手法について述べた．実験により，素性として「日本語読み音訳素性値」「英語読み音訳素性値」「音訳素性値のうち小さい方」を入れることで，カバレージが 10 ポイント近く向上することを示し，訳語抽出において音訳素性が有効であることを示した．

文献

Joseph Z. Chang, Jason S. Chang, and Jyh-Shing Roger Jang (2013). Learning to find translations and transliterations on the web based on conditional random fields. *IJCLCLP*, Vol. 18, No. 1, pp. 19-45.

Rieko Tsuji, Yoshinori Nemoto, Wimvipa Luangpiensamut, Yuji Abe, Takeshi Kimura, Kanako Komiya, Koji Fujimoto, and Yoshiyuki Kotani (2012). The transliteration from alphabet queries to japanese product names. In *Proceedings of the 26th PACLIC*, pp. 456-462.

Yuejie Zhang, Yang Wang, and Xiangyang Xue. English-chinese bi-directional oov translation based on web mining and supervised learning (2009). In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 129-132.

高瀬真記、古宮嘉那子、小谷善行(2013)「CRFを用いたアニメ関連用語の固有表現抽出」第三回コーパス日本語学ワークショップ予稿集，pp. 179-182.

ポスター発表（1） グループB

3月6日（木） 16:10～17:10

遡及的に構成される発話連鎖の諸特徴

鈴木　佳奈（広島国際大学心理科学部）†
山本　真理（北星学園大学）
鈴木　亮子（慶應義塾大学経済学部）
伝　康晴（千葉大学文学部／国立国語研究所言語資源研究系）

Annotation of Retrospectively Operating Sequences in Conversation

Kana Suzuki (Hiroshima International University)
Mari Yamamoto (Hokusei Gakuen University)
Ryoko Suzuki (Keio University)
Yasuharu Den (Chiba University/National Institute for Japanese Language and Linguistics)

１．はじめに

本研究は、会話を構成する連鎖構造の一つである「遡及的連鎖（retro-sequence）」について、会話コーパスの分析を通して、その特徴を同定するものである。遡及的連鎖とは、連鎖をなす２つの発話が、通常の隣接ペア（adjacency pair）のように「予測的（prospective）」な関係ではなく、ある種の「応答」が出現することによって、その「引き金」の存在が顕在化するような関係（遡及的＝retrospective）にあるものを指す（Schelogff 2007）。これまでの話し言葉コーパスではこのような遡及的な発話関係の情報がタグとして付与されることはなく、したがって、どのような事象が遡及的連鎖に該当するのか、また、どのような基準でそのアノテーションが可能か、明らかになっていない。そこで本研究は，３種の会話コーパス（CSJ，千葉大学 3 人会話コーパス，宇都宮大学音声対話データベース）を対象に遡及的連鎖のアノテーションを試行した。本発表では，その作業の過程で見出された遡及的連鎖および関連する事象の種類と，それぞれを特徴づける言語的，発話連鎖的，対話構造上の諸要素を挙げた上で，遡及的連鎖のアノテーション手法の策定に向けて今後さらに検討すべき課題をまとめる。

２．会話の連鎖構造と遡及的連鎖

２.１　会話の連鎖構造

会話は、単にことばが無秩序に並べられているものではなく、発話の連鎖で成り立っている。さらにそれらの発話は単に連なっているのではなく、構造化されている。その最小単位は、「働きかけ（initiating action）」と「応答（responding action）」という二つの発話のペアから成る。このとき、二つの発話は異なる行為者によって産出される。通常は働きかけが行われた後に応答が行われ、その順番が逆にはならない。さらに、働きかけと応答は、同じタイプの行為が行われることによって関連づけられる。このような発話連鎖の典型が「質問―答え」、「挨拶―返答」、「依頼―受諾または拒否」といった「隣接ペア」（Schegloff & Sacks 1973）である。

実際に会話を観察すると、構造化された連鎖によって会話が組み立てられているというだけでなく、構造化されているという事実そのものが会話参与者にとって利用可能なリソースとなっていることに気づく。一人の参与者によってある種の働きかけがなされると、別の参与者が応答する義務を負う。またその際、適切な応答として行えることに制限がかけられる（例えば、「おはよう」という働きかけに対して「わかった」と応答するのは不自然だと見なされる）。このような連鎖構造があることによって、参与者は会話の「先」を予告したり予測したりすることが可能になるし、また予告／予測からの逸脱（多くの場合「応

† k-suzuky@he.hirokoku-u.ac.jp

答の不在」という形をとる）が意味を持つものとして意識化される（例えば、質問した相手から答えが返ってこなかった場合、答えるべき相手がなんらかの理由で躊躇していると理解する、など）。このようなことが可能になるのは、働きかけと応答が「予測的（prospective）」に関連づけられているためである。

２.２　遡及的連鎖

一方で、会話連鎖のなかには、予測的にではなく「遡及的（retrospective）」に関連づけられるものもあることが指摘されている（Schegloff 2007:217-219）。すなわち、先行する発話からは想定されていないような「応答」がまず出現し、それによってその「引き金」の存在が顕在化する、というものである。そのような応答には、例えば、「笑い」がある。

【事例１】（ほぼ日刊イトイ新聞　2013/01/03　「2013 年あんこの旅　第 3 回　小豆ジャム。」より。 コピーライター・糸井重里氏があんこへの熱い思いを「講演」で語っている。
糸井氏はあんこ好きであり，ジャム作り好き。 じゃあ，「小豆でジャムを作ってみたら」と思い立って実際に作ってみたが，瓶に詰めたところで，「普通のあんこ」だということに気づいた，というお話。）

観客Ｂ	これは質問ではないのですが、わたしたちの中では有名なあのお話をあらためて聞かせていただけますでしょうか。	
糸井	….あのお話、というのは？	
観客Ｂ	「小豆ジャム」のお話です。	
糸井	ああ….（苦笑）。	
観客Ｂ	記録に残しておきたいと思いますので、この場でぜひ、お話しいただけるとうれしいです。	
全観客=>	（クスクス）	**【笑い】**
糸井->	何か？　おかしいことが？	**【笑いの対象を特定する試み】**
観客Ｂ	いえ！　すみません、つい。	
糸井	どうかひとつ、真面目にお願いします。	

講演の途中、観客から笑いが漏れるが（=>の発話）、この瞬間の笑いは「想定外」のものであるとの認識が、直後の糸井氏の「何か？　おかしいことが？」の発言（->の発話）によって示される。このように、想定されていない応答（ここでは「笑い」）が出現したことで、先行発話のどこか、あるいは会話を取り巻く物理的環境のどこかに、その笑いの引き金になったもの（ここでは「笑いの対象＝laughable」）があるはずだ、という認識が共有されるのが遡及的連鎖である。

Schegloff (*ibid.*)は、遡及的連鎖を開始する「応答」として、「笑い」、「聞き返し」（本稿 4.2.1 節参照）、「気づき」（本稿 4.2.2 節参照）の３つを挙げている。また、高梨（2008, 2010）は、ある種の「評価」も遡及的連鎖を構成するとしている（事例２）。[1]

---

[1] ただし、すべての「笑い」「気づき」「評価」が遡及的連鎖を開始するわけではない。例えば、次の事例の評価発話（「高いですねー」）は、質問―応答の発話ペアの「拡張（post-expansion）」(Schegloff 2007:118-148)であり、むしろ予測的連鎖に位置づけられるものである（cf. 鈴木 2007）。

A:　身長何センチですか？
B:　170 センチです。
A:=>　高いですねー。

「笑い」「気づき」についても、直前の発話が明示的にそのような応答を「要求している」場合がある。

【事例２】

1A:–>	もうこんな時間か。	【評価対象】
2B:=>	そんなにイライラしないで。	【評価】
3A:	イライラなんかしてないさ。	

２.３　遡及的連鎖をアノテートする際の問題点

会話コーパスに対して発話連鎖の情報を付与する試みはすでに始まっているが（例えば人工知能学会「談話・対話研究におけるコーパス利用」研究グループ 2000; Dhillon, et al. 2004）、現状では「質問―返答」や「依頼―承諾」のような単純な発話ペアに限られ、様々な発話連鎖を記述する手法は存在しない。

では、遡及的連鎖のアノテーション手法を新たに開発するために、どのような問題を解決する必要があるのか。第一に、遡及的連鎖とそうでないものを区別する明確な基準の策定が必要となる。第二に、応答の言語的形式からどのように引き金が特定されうるのか、その遡及性を記述する必要がある。第三に、遡及的連鎖が開始されると同時に、同じ応答を起点とした別の予測的連鎖が開始される場合があり、そのような「あとの発話展開」も整理し記述しなければならない。そして第四に、これらの情報をコーパスに付与するための適切なタグの設計が必要となる。

なお本発表では、上記の第一から第三の問題について、現段階での見解を示す。

３．使用コーパス

本研究では、日本語話し言葉コーパス（CSJ）、千葉大学 3 人会話コーパス、宇都宮大学音声対話データベースの 3 種の会話コーパスを対象に遡及的連鎖のアノテーションを試行した。日本語話し言葉コーパスは、「独話」「対話」「朗読」のうち「対話」のみを対象とし、15 分の講演インタビューと 28 分の課題指向対話を取り上げた。千葉大学 3 人会話コーパスからは 9 分 30 秒の自由会話を、宇都宮大学音声対話データベースからは 4 分と 8 分の課題指向対話を選び、分析者 4 名で個別にアノテーションを行ったのち、協議により遡及的連鎖とそうでない事象を認定した。

４．遡及的連鎖の諸特徴

４.１　分析対象から除外された事象

まず、Schegloff (2007)および高梨（2008, 2010）で遡及的連鎖とされたもののうち、「笑い」および「評価」を最初に分析対象外とした。注１で述べたように、同じ発話形式でも遡及的連鎖のものとそうでないものがあり得、先行発話によって予測的に関連づけられているのかどうかの判定が困難であることが予想されたためである。

事例を検討する中で分析対象から除外すると判断したものに、「あ、そうかハートか」や「あなるほどね」のような相槌があった。感動詞「あ」が独立して「気づき」という認知的変化を表示しているのか、「あ、そうか」が一まとまりで「納得」を示しているのか、分析者の見解が分かれたが、これらの相槌は、何らかの情報提供が行われた発話の直後に出現することが多く、情報提供発話に対する予測的な応答であると判断した。

また、A:「四コマ中の枠内に駅前ってとりあえず書いてあって駅前で工事してんだね」B:「駅前で工事、うん」の B の発話のように、先行発話で与えられた情報を単純復唱する発話も、今回は対象外とした。

さらに、先行文脈で出されていた語句や情報を再び取り上げて、それを接ぎ穂に新しい話題を導入するような発話（例えば「でその、あのー、中国に行くきっかけなったのがチューターをやってたっていうことなんですけど…」）も、その語句・情報が最初に提示された発話そのものに連鎖的に関連づけられていないと判断し、遡及的連鎖として認定しなかった。

４.２　遡及的連鎖と認定された事象

先行研究で遡及的連鎖とされたもののうち、「聞き返し」と「気づき」は、分析を行ったデータの中にも複数見られた。また、事例数は少ないものの、本研究を通して新たに遡及的連鎖と認定されたものとして、「会話の流れの差し戻し」と「異なるポイントへの食いつき」があった。

４.２.１　聞き返し

「聞き返し」の事象については、会話分析などの研究分野で、「他者開始修復（other-initiated repair）」という名称ですでにある程度の知見が蓄積されている（Schegloff, et al. 1977; Suzuki 2010）。事例３の発話 02「私？」や発話 07「わたしのことか」のように、相手の発話の聴き取りまたは理解に問題が生じたときに、その問題の解決を目指して産出される発話である。

【事例３】（A が B にインタビューしている。直前で、B がもう６～７年ほど大学院に在籍していることが述べられている。）[2]

01A:　な［ に ］を (0.2) 研究してるんですか？ (.)［ずっと
02B:=>　［うん］　［k- (0.2) 私？
03A:　う［ん
04B:　［私？
05　(0.1)
06A:　［うん］
07B:=>　［haha］haha［ha お わた］しのことか。
08A:　［ha ha ha ha ］
09A:　a［hh ］.h.h［.h.h.h］
10B:　［ええ］　［.h.h.h］私は.h え：と：勉強自体は：
11B:　最近はあんまりしてないんです¥けど：¥とか言って

以下にその特徴をまとめる。

［言語形式の特徴］

聞き返しに用いられる言語形式には表１に示したような種類がある。

［出現する発話連鎖上の特徴］

どんな発話のあとにも出現しうる。

［引き金の特定］

原則として、その直前の発話が引き金（問題源）として理解される。「え？」「はい？」などの形式の場合、先行発話のなにが当該の問題を引き起こしているのかは特定されないが、他の形式では、用いられる形式に対応して、なにが問題なのかが特定される。

---

[2] 事例３から事例７では、発話の音声的特徴を可視化するため以下の記号を用いる。

↑↓	ピッチの上下	？	上昇イントネーション
。	下降・終了イントネーション	、	継続イントネーション
あ：	引き延ばされている音	あ-	途切れて不完全な音
（あ）	聞き取りが確定できない音		
＞ああ＜	周囲よりスピードが速い部分	＜ああ＞	周囲よりスピードが遅い部分
°ああ°	周囲より音が小さい部分	¥ああ¥	笑いながらの発話
.hhh	吸気音	h	呼気音や笑い
(.)	0.2 秒以下の沈黙	(1.0)	0.2 秒以上の沈黙（かっこ内は秒数）
［	重なっている複数の発話		

［応答に続く展開］
聞き返しを受けて、問題源発話者が、問題を解決すべく発話の修復を行う。

４.２.２　気づき

「気づき」については、他者の発話を引き金になんらかの気づきが示されるケース（事例４）と自分の発言途中で気づきが示されるケース（事例５）の２種類が見出された。また、今回分析したデータの中には見られなかったが、会話を取り巻く物理的環境が気づきの引き金になるケースも考えられる。

表１　聞き返しの形式と問題源の特定
（Suzuki 2010:152 に基づく）

聞き返しの形式	問題源の特定
「え？」「ん？」「なに？」「はい？」「あ？」「は？」「へ？」「ふん？」	先行発話に問題があることを伝えるが，どこが問題かは特定せず
疑問詞（＋上昇イントネーション）「だれ？」「いつ？」「どこ？」	先行発話の中で疑問詞と対応する要素が問題源
先行発話の一部分の繰り返し＋疑問詞「なに操業？」「おばあちゃんがなに？」	繰り返し部分が「枠」となって，問題源が特定
疑問詞＋格助詞「だれが？」「なにを？」	先行発話で省略されている要素が問題源
先行発話の一部の繰り返し＋「って」（＋疑問詞）「ばらけてるって？」「次からと言いますと，いつから：」「どこ？それ」	繰り返された部分の意味内容が問題源
「どうゆうこと？」「よくわからない」	先行発話全体の意味内容が問題源
先行発話の一部分の繰り返し名詞句：「今日？」，述部：「いらないの？」	繰り返された部分が問題源
「～じゃない？」，繰り返し＋「って」＋理解「フジキナオヒトじゃなくて？」「もうちょっとって一時間くらい？」	先行発話の問題部分が特定されるとともに，それに対する聞き手の理解も提示

【事例４】（A と B の間には仕切りがあり、お互いが見えない。４コママンガの２コマをそれぞれ手元に持っていて、言葉のみで説明し合いながら、バラバラになったコマの正しい順番を推測するという課題を行っている。）

01B:　.h でね (0.4) 人がひ- (0.2) 一人 (0.4) 主人 主人公
02　(1.1)
03A:　サラリーマン風の。
04　(1.4)
05B:　サラリーマン風ではねえけど［工事］現場の人（で／だ）しょ
06A:=>　［えっ］
07　(0.2)
08A:=>　えっ (.) ［>そうな-< そうなんだ］
09B:　［一 人 だ も ん ］だってヘルメットかぶってるよ？

【事例５】（A が B にインタビューしている。インタビューに先立つ講演の中で、B は、初めて海外旅行に行った中国・大連のトイレが衝撃的だった話をしていた。）

01A:　水洗なんですよね。
02B:　うす- (.) 一応水洗でしたね。
03B:=>あ：、そこは (.) そこは水洗じゃなかったな。

以下にその特徴をまとめる。
［言語形式の特徴］「あ」「えっ」「あれ」が単独で、または発話頭に用いられる。
［出現する発話連鎖上の特徴］ どんな発話のあとにも出現しうる。また、発言の途中で出現することもある。
［引き金の特定］　気づきを示す言語表現だけでは、なにを引き金に、どんな気づきが生まれたのかは特定できない。気づきの表現に続く発話内容によって、引き金の特定が可能

になる（しかし、引き金を明確に特定できないこともある)。
［応答に続く展開］　他者の発話が引き金になっている場合、気づきの表現のあとには、相手が言ったことへの疑問や、自分の認識との食い違いが提示される。その後、両者の間で協働して、疑問の解決や認識のすり合せが図られる。
　自己の発話の途中で気づきの表現が差し挟まれる場合は、その直後に、そこまでの発言内容についての訂正が行われる。すなわち、この場合の気づきとは、自身の「間違い」に自発的に気づき、その間違いを訂正したものと見なすことができる。

４.２.３　会話の流れの差し戻し

　分析者の協議の結果、現在の会話の流れを明示的に中断させ、その「一歩前」に戻る発話も遡及的連鎖に含めることと判断した。事例６の発話 08 のような事象である。

【事例６】（事例４とは違うペアが、同様の課題遂行対話を行っている。）

01L:	＞だか多分＜ビーが［最初で：、
02R:	［うん：うん
03L:	.hhhh (.) その：次がこう- (0.4) 眼鏡をこ、はず、
04	(0.2)
05R:	うん
06	(0.4)
07L:	外そうとしてる：、ところ：？ (.) ん？
08R:=>	えっ t-、＞待った＜エーがなん- (.) エーは：、
09L:	エーは：＞なんかね＜、.h 眼鏡を外し-

　以下にその特徴をまとめる。
［言語形式の特徴］　「えっ」「待った」のように、明確に会話の流れを止めるような表現が発話頭にある。
［出現する発話連鎖上の特徴］　直前の発話で、事例６の発話 03「その：次が」のように、現行の作業を次のステップに進めることが提案されている。このように、次のステップに進めることが明示的に示されるのは、課題遂行対話に特徴的な行動と言えるかもしれない。
［引き金の特定］　なにを引き金とみなすかについては、事例を増やしての検討が必要である。
［応答に続く展開］　「待った」に続く発話では、先に進む前に確認したいことが挙げられる。その求めに応じて、相手が情報を提供する。

４.２.４　異なるポイントへの食いつき

　「異なるポイントへの食いつき」とは、直前の発話の言わば「主旨」とは異なるところに「食いついて」反応しているような発話を指す。事例７の場合、発話 01 から 06 にかけての B の発話の主旨は「欧米人の留学生は少ない」ということであるが、インタビュアーA は、その事実に対しての反応も示しつつ（発話 07「へえ：：：：：」)、それとは違う情報（6〜7 年ぐらい大学院にいる）についての質問を次に行っている（発話 09)。この発話 09 は、分析者から見ると「そこに食いつく？」と思うような、副次的な背景情報に焦点を当て、前景化しているような印象を与える。

【事例７】（A が B にインタビューしている。B が在籍する大学院への留学生の話。）

01A:	白人はいないんですか
02B:	.h いないですね：［あのね：］.h 私大学院にもうかなりの (0.6)
03A:	［は　：　：］

04B:　ろく (.) 年 (.) な［な年ぐら］いいるんですけ［ど：］.h ひっとり：(0.2)
05A:　［う　：　ん］　［うん］
06B:　ぐらいですね：留学生で［.h あの：(0.2) 欧米の人がいたのは］
07A:　［　へ　え　：　：　：　：　］：［: .h
08B:　［うん
09A:=>　まだいらっしゃるんですね在籍してるんで［すºねº
10B:　［うん
11B:　してます .h うん

以下にその特徴をまとめる。
［言語形式の特徴］　特定の言語形式が使用されるわけではない。
［出現する発話連鎖上の特徴］　他者による自発的な、あるいは質問に答える形での情報提供がなされた発話の直後に出現する傾向がある。
［引き金の特定］　直前の発話で提供された主たる情報ではなく、副次的情報に当たる発話部分が引き金となっているように見える。
［応答に続く展開］　「異なるポイントへの食いつき」が質問という発話形式を伴った場合は、相手がその質問に応答する。

４.３　遡及的連鎖の可能性がある事象

現段階で、遡及的連鎖かどうかの判断を保留しているものに「他者の発言の言い換え」がある。例えば、「海外にどうしても行きたいっていう（0.8）気にあんまなかった」というインタビュイーに対して、インタビュアーが「海外に興味がない」と質問する場面がある。相手が言ったことを異なる表現で言い換えて質問する、というのは 4.2.1 節で紹介した「聞き返し」の例と考えることもできるが、この場面での質問は、相手の発言が理解できずに聞き返しているわけではないように見える。このような質問がどんな行為を行っているか、ということも含めて、類似の事例を集めて遡及的連鎖かどうかを検討する必要がある。

５．まとめ

アノテーションの作業をする中で、遡及的連鎖を認定するため判断基準として、以下の２点が浮かんできた。

(a) 特定の言語形式が使われているかどうか。ただし、その言語形式の有無だけでは決定しきれない。
(b) 「会話の流れ」が一時停止、ないし中断されているかどうか。会話の流れを作り出すものが発話間の予測的な関連性であるとすると、遡及的連鎖の開始は「予測からの逸脱」であり、先行発話との間にある種の「断絶感」がある。

今後、分析するデータおよび該当事例を増やし、上記の(a)と(b)について、より精密な記述をした上で、具体的なアノテーション手法の設計に移る。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト（独創・発展型）「多様な様式を網羅した会話コーパスの共有化」（平成 23 年～26 年度、代表者：伝康晴）による補助を得ています。

文　献

小磯花絵（2009）「話しことばコーパスの情報」『国文学解釈と鑑賞』,74:1, pp. 53-60

小磯花絵、伝康晴、前川喜久雄（2012）「『日本語話し言葉コーパス』RDB の構築」『第 1 回コーパス日本語学ワークショップ予稿集』、 pp. 393-400
人工知能学会「談話・対話研究におけるコーパス利用」研究グループ（2000）「様々な応用研究に向けた談話タグ付き音声対話コーパス」『人工知能学会研究会資料 SIG-SLUD-9903』、pp.19-24
鈴木佳奈（2007）「会話における数字の使用に関する一考察―『数字数量表現』と『主観的数量表現』が共在する事例をてがかりに―」『津田葵教授退官記念論文集　言語と文化の展望』、pp. 363-374、英宝社
高梨克也 (2008) 「社会的参照現象の時間的展開としての評価連鎖」『電子情報通信学会技術報告』108:187 (HCS2008-34)、pp.21-26
高梨克也 (2010) 「インタラクションにおける偶有性と接続」，木村大治・中村美知夫・高梨克也（編）『インタラクションの境界と接続―サル・人・会話研究から』、pp.39-68、昭和堂
伝康晴（2009）「隣接ペア」『多人数インタラクションの分析手法』、pp. 82-94、オーム社
伝康晴、小磯花絵、丸山岳彦、前川喜久雄、高梨克也、榎本美香、吉田奈央（2009）「対話研究にふさわしい発話単位の提案とその評価(1)～短い単位～」『人工知能学会研究会資料 SIG-SLUD-A803』、 pp. 75-80
伝康晴、小磯花絵、丸山岳彦、前川喜久雄、高梨克也、榎本美香、吉田奈央（2010）「対話研究にふさわしい発話単位の提案とその評価(2)～長い単位～」『人工知能学会研究会資料 SIG-SLUD-A903』、pp. 13-18
Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., and Yoshida, N. (2010) Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme, *Proc. LREC 2010*, pp. 2103-2110
Den, Y., Koiso, H., Takanashi, K. and Yoshida, N. (2012) Annotation of response tokens and their triggering expressions in Japanese multi-party conversations, *Proc. LREC 2012*, pp. 1332-1337
Den, Y., Yoshida, N., Takanashi, K. and Koiso, H. (2011) Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. *Proc. Oriental COCOSDA 2011*, pp. 168-173
Dhillon, R., Bhagat, S., Carvery, H., & Shriberg, E. (2004). Meeting recorder project: Dialog act labeling guide. *ICSI Technical Report TR-04-002*, International Computer Science Institute.
Heritage, J. (1984) A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, pp. 299-345.
Schegloff, E. A. (2007) *Sequence Organization in Interaction: A Primer in Conversation Analysis Volume 1*. Cambridge: Cambridge University Press.
Schegloff, E. A. and Sacks, H. (1973) Opening up closings. *Semiotica*, 8, 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*, 53:2, pp.361-382.
Suzuki, K. (2010) *Other-Initiated Repair in Japanese: Accomplishing Mutual Understanding in Conversation*. Doctoral Dissertation Submitted to Graduate School of Intercultural Studies, Kobe University.

関連 URL

ほぼ日刊イトイ新聞　「2013 年あんこの旅　第 3 回　小豆ジャム。」
http://www.1101.com/anko/2013-01-03.html

コーパスから取得した用例で対象物が認識可能であるのか

保田　祥（国立国語研究所コーパス開発センター）†

Can We Recognize Objects Using the Texts in Corpora?

Sachi Yasuda (National Institute for Japanese Language and Linguistics)

１．はじめに

テキスト情報による経験のみで，人はテキストの示す対象物がどのようなものか認識することができるのだろうか。

本研究では，対象物に関する文脈情報をコーパスから抽出し，用例からの対象物の推定実験を試みる。コーパスから取得可能な用例をまとめるとともに，対象物を同定するために有効な要素・用例について考察する。また，大規模なコーパスを用いるなど用例数を増やすことによって，取得可能な情報量もまた増加するのかという問題についても検証を行う。調査は，コーパスから取得した用例で対象物がどの程度同定可能であるのか，現代日本語書き言葉均衡コーパス（以下BCCWJ）を用いた被験者実験を行った（調査①）ほか，Google-日本語Ngram[1]（工藤・賀沢,2007）から百科事典的意味（意味的な用例）を取得することを試み，BCCWJから取得可能な用例との異同を確かめた（調査②）。これらをもとに，コーパスから取得可能な，対象物の認識を可能とするテキスト情報を考えた。

２．先行研究と本研究

たとえば辞書は，対象物の説明として適切な質・量の情報を提供していると考えられる。しかし，Fillmore & Atkins (1994) では，辞書語釈で解釈できない用例が見つかることが示されている。辞書語釈として記述された要素から対象物を同定する実験を行うと，読み手が知識を有している動物に限っても，半数程度（52%）しか同定できなかった[2]（保田ほか，2013）。

Sinclair（1991，1992）は，学習者に必要とされるのは用例であるとし，コーパスに基づく用例を重視した新たな辞書（COBUILD）作りを行っている。対象物に関する多くの用例こそが，対象物の経験を得るテキスト情報であるとも考えられる。

それでは，コーパスから取得した用例を提示することで，対象物を認識することができるのだろうか。また，コーパスから取得できる要素・用例は，対象物の同定に質・量的に十分だろうか。本研究は，コーパスから取得した用例で対象物が認識可能であるのかを調査する。

３．調査①：BCCWJ 用例から対象物を同定する

BCCWJ から取得した要素・用例で対象物がどの程度同定可能であるのか，動物を対象として被験者実験を行い，結果の分析を行う。

３.１　データ①：BCCWJ からの用例取得

BCCWJから動物（一般に知識があると考えられる単語親密度 5.000 以上[3]）に関する要素・

---

† yasuda-s@ninjal.ac.jp

[1] http://www.gsk.or.jp/catalog/gsk2007-c/

[2] 日本国内で出版されている国語辞書 10 種類を用い，5 種以上の辞書に語釈として記述されていた内容を用いた。

[3] 最大値は 7.000（天野・近藤編，1999)。

用例を抽出した。中納言[4]を用いて検索語の前後 50 文字を取得し，手作業で整理を行った。なお，文意の読み取りに文脈情報不足している場合や用例が文字数の制限によって途切れている場合などは，前後 500 文字を再取得して同様に抽出を行った。さらに，実験協力者に提示する際には，抽出した用例が句などの場合に文へ整形[5]したほか，意味的に同種と判断される用例については，以下の例のように作業者の判断でまとめた。このような複数用例をまとめた例を，本稿では意味的な用例と呼ぶ。

取得用例）

A. カモシカの被害防止対策調査　カモシカの食害発生機構の解明　カモシカの林業被害が近年，特に問題になっており～　（「環境白書」）

B. カモシカが増えたため，ヒノキの幼木を食い荒らされるという被害を受けている～　（中村幸昭「鳥羽水族館館長のジョーク箱」）

C. 最近カモシカに食われる被害が出ている『会津の伝統野菜を守る会』によって選ばれた野菜は，現在十四品目。　（丹野清志「やさい畑」）

A～C のまとめ例：意味的な用例）

この動物による林業（ヒノキの幼木）や農業（野菜）などの食害が問題とされている。

３.２　実験：BCCWJ から取得した用例を用いた対象物同定実験

10 種類の動物について，実験協力者 12 名（日本語母語話者）が，提示された用例群（対象物名は「この動物」とマスク）から対象物の同定を行った。実験協力者は，回答が可能であった場合には，どの用例が有用であったのかをあわせて記した。

実験に用いた用例については，実際に検索した結果としての件数（異表記を含むが，異義語（固有人名など動物でない）を除く）から，同サンプル内などの重複を除いたほか，同種の要素を有している用例については意味的な用例とした。以下に示す動物についての用例を提示した（異なり：提示した意味的な用例数／サンプル数／述べ：検索結果件数）。

タヌキ（41 例/372 サンプル/581 件）・カワウソ（23 例/38 サンプル/164 件）
テントウムシ（21 例/48 サンプル/69 件）・オットセイ（17 例/11 サンプル/67 件）
スズキ（12 例/36 サンプル/65 件）・カナブン（17 例/18 サンプル/36 件）
カマス（8 例/19 サンプル/26 件）・ジュウシマツ（10 例/7 サンプル/14 件）
ジャガー（5 例/9 サンプル/13 件）・ナイチンゲール（9 例/6 サンプル/8 件）

３.３　結果：BCCWJ から取得した用例を用いた対象物同定実験

10 種類の動物[6]について，BCCWJから抽出できる意味的な用例から対象物を同定した結果，動物毎の正答率のマクロ平均は 25%となった（人毎[7]の正答数のマクロ平均は 3.0 種類である）。

半数以上が正答する高正答率動物群（図１のⅠ）と，同定できないかごく稀に同定可能な低正答率動物群（図１のⅡ）に大別できる。

---

[4] https://chunagon.ninjal.ac.jp/

[5] 要素を複数含む場合など，一句を二文にした例もある。

[6] 保田ほか（2013）の実験によって，それぞれの動物に関する実際の知識率を調査している。知識率の平均は，87%であった。ナイチンゲール（20%）を除き，70%以上の実験協力者が当該塔物の知識を有していた（min：カマス（70%），max:タヌキ・テントウムシなど（100%））。

[7] 内訳は，1 種類：1 人，2 種類：5 人，3 種類：3 人，4 種類：2 人，8 種類：1 人である。

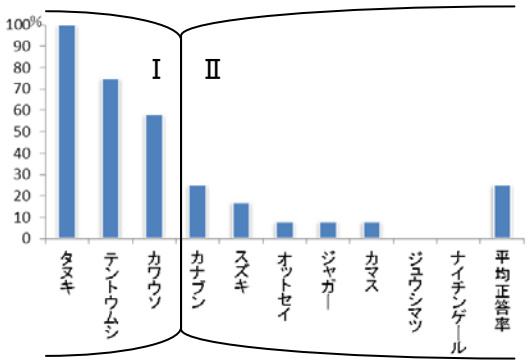


図１　BCCWJ 用例からの対象物同定正答率

３.３.１　高正答率動物群に見る判断に有用な情報

タヌキは 100%の正答が得られた。判断の際には，92%の実験協力者が「カチカチ山」(「タヌキ」全件の 2.5%と高頻度用例）を有用であったとした。同種の「ぶんぶく茶釜」は 33%にとどまるが，「カチカチ山」のように一般に知られた文化的知識レベルの情報は，タヌキを想起するトリガーとして重要であると考えられる。すなわち，用例としての出現頻度も高い人口に膾炙した物語名[8]などによって，高い正答率が得やすいことが予測される。このような文化的知識として経験した情報が，対象物の認識に最も有用である可能性があろう。

次に，75%の正答が得られたテントウムシで，判断に有用とされたのは，特徴的な形状に関する用例，とくに比喩表現でもある以下の用例（68%）であった。

例）（前略）お椀を逆さまにしてテーブルの上に伏せたような格好をしている。少し，てんとう虫にも似ている。　（垣根涼介「君たちに明日はない」）

また，一般的によく知られた情報と考えられる用例「アブラムシを捕食する。」が 42%で有用とされていた。このような用例は，テントウムシそのものの生態的な特徴説明というよりも，具体的な用途としての前提知識となっている例である。

例）毛虫は全部手で取り，アブラムシはテントウムシを連れてきて退治すると，無農薬で収穫できた。　（坂下まりあ「あたしは非定型精神病なのだよ」）

比喩表現をはじめ，前提知識として既知の情報という扱いをされている用例は，一般的な知識であると考えられている情報であろう。タヌキでも，「でっぷりした体（33%)」「寝たふりをする（42%)」のように，比喩表現や慣用的な表現についての用例が，他用例に比べて対象物の認識に有用性が高いとされていた。但し，知識や実際に経験する機会には個人差のある場合が考えられるため，テントウムシの例では，100%の正答には至らなかったようである。

また，カワウソは 58%の正答であるが，判断に用いられた用例は，正答した実験協力者のすべてが一致するということはなく，最大でも 3 人の一致であった。また，誤答は 1 人のみ（「ミンク」）で，4 人が無回答（空欄）であった。カワウソでは，複数の用例のいずれかが個人の有する個別的経験知識と合致した場合にのみ解答が可能となったと考えられる。テントウムシでも「落ち葉の裏などに集まって越冬する。」「手にとまることがある。」のような用例についても複数人（25%）が有用としていたが，経験の汎用性が下がり，個別的であるといえる。

[8]同様にオオカミならば「赤ずきん」，ウサギであれば「ウサギとカメ」のような用例が，それぞれ当該動物全件の 3%という高頻度の用例である。

このように，対象物の認識に有用とされたのは，文化的知識＞一般的経験知識＞個別的経験知識に関する用例の順であった。

なお，テントウムシの誤答は「蝶」であったが，その際「翅を広げた様子をリボンに喩えた例」が判断に有用だったとされていたことから，実験協力者によっては，印象に残る一文など，一つの経験が対象物の認識に強い影響を及ぼした可能性も残る。

３.３.２　同定の難しい動物群に見る判断に有用な情報

タヌキ・テントウムシ・カワウソ以外の動物については，正答率が 0%の動物もあり，いずれも低い正答率であった。

稀に同定が可能な場合には，「玄関や階段，バルコニーにいる」という個別的な経験に合致したと考えられる用例や，もしくは，「メソ・アメリカ文明の～」「プレ・インカ文明は～」といった専門性が高い知識と考えられる用例が判断に有用とされていた。個別的経験知識と一致する用例があった場合にのみ[9]，解答が可能であったことがわかる。

また，ジャガーの誤答は「トラ」「チーター」「ライオン」，オットセイの誤答は「アザラシ」「ビーバー」「トド」，カナブンの誤答は「タマムシ」「カメムシ」「コガネムシ」「セミ」，ナイチンゲールの誤答は「ヒバリ」「カナリア」などと，それぞれ多様であって統一性もなかった。

３.４　考察：対象物の同定に必要とされる情報がコーパスから取得可能か

知識があるはずの動物[10]であっても，その動物の情報が提示された際，その情報が実験協力者の知識と合致しなければ，対象物の同定は不可能である。

辞書（保田ほか,前掲）よりもコーパスから取得できた用例数は多く，情報量は豊富であると考えられるが，それらの記述が必ずしも対象物の同定に質的に十分とはいえない。

それでは，対象物の認識に求められる情報がコーパスから取得可能なのか。

３.４.１　経験知識を喚起する情報

経験知識を喚起する情報，「調理法」や「作品名」，「場所名（動物園・水族館など）」などは，コーパスから得られる場合が多い。

調理法例）スザンヌがよく作っていたのは，魚のクール・ブイヨン煮。ポワソニエールという魚が一匹まるまる入るような，大きな楕円形の鍋で煮た鱒や鮭，すずきなどをゆでたじゃがいもといっしょに食べた。（猪本典子「修道院のレシピ」）

作品名例）アンデルセンは，『天使』と『みにくいアヒルの子』と一緒に『ナイチンゲール』という作品を彼女のもとに送った。『ナイチンゲール』は，自然に歌う鳥の方が人工的なおもちゃよりもすばらしいということを述べた作品である。（アリソン・プリンス「ハンス・クリスチャン・アンデルセン哀しき道化」）

場所例）話が変わりますが，写真は海遊館のカワウソです。（「Yahoo!ブログ」）

実際に，コーパスから取得した要素から対象物を同定することができた場合，上記の用例のような経験知識源となり得る情報は有用性が高いと判断されている。

しかし，上記の作品名例で見たアンデルセンの作品名のほか，以下の例のような場所情

[9] なお，本実験において正答率が 0%であった「ジュウシマツ」と「ナイチンゲール」についても，ジュウシマツを飼育していた経験のある場合や，アンデルセン童話などの知識のある場合には，正答が可能であったことが別途確認された。

[10] 注 6 を参照。知識率はナイチンゲール（20%）を除き 70%以上であった。

報と関わる例などは，動物名がマスクされた場合，対象物の同定に有用とされない。

例）わが国では絶滅が心配されるカワウソを使った漁が，中国南部で行われていることが，
周達生・国立民族学博物館助教授の最近の現地調査でわかった。（花輪莞爾「猫学入門」）

「ウサギのヒゲ」や「ダチョウのクチバシ」のように，一般性が高いが比喩などに用いられるのではない（当該対象物に特徴的ではない）情報は，テキスト化されにくい（Yasuda ほか, 2012）。すなわち，一般に未知であろう情報を発信している用例が多いことから，個別的経験あるいは専門性が高い情報となりがちであり，読み手の経験知識と合致しない例も多いと考えられる。

３.４.２　想定されるカテゴリーの他メンバーとの差異情報

想定されるカテゴリーにおける他メンバーとの差異を示す情報が，コーパスから取得しにくいといえる。対象物の同定の難しい動物群においては，提示された用例では，「大型肉食獣」や「水棲動物」などと想定されたであろうカテゴリーにおいて，他メンバーとの差別化ができなかったことが推測されるためである（前掲，３.３.２）。

カテゴリーの他メンバーとの差異としては，特徴的要素のほか，大きさなどの情報が役立つ場合が考えられよう。対象物が「大きい」「小さい」のような情報はコーパスから取得しやすく，たとえばオオカミであれば，「大きい」が全用例中 2%出現する高頻度の要素として取得されるのである。しかし，用例を見ると，「大きい」はオオカミカテゴリーの中で「大きい」という情報である。対象物そのものの情報ではあるが，哺乳類や猛獣，野生の獣などのカテゴリーにおける大小の差別化情報ではないため，対象物が何であるのか認識するために有用なのではない。

例）森の主ならかれらは大きなオオカミだというだろうよ。さもなければ神聖な白いシカだと！　（栗本薫「マルガ離宮殺人事件」）

３.４.３　比喩表現

慣用化した表現や比喩として用いられていると読み取れる用例は，コーパスから取得しやすい情報であり，対象物の同定への有用性が考えられる。一般的な読み手が「当該動物に特徴的である」と想定可能な性質が，焦点とされているはずだからである。テントウムシの用例において，対象物の同定に 68%の実験協力者に有用とされたのも比喩表現であった。

動物によって異なるが，たとえば知識率が 100%と一般によく知られていると考えられるタヌキであれば，全用例中の 11%（63 件/581 件）が「みたい」「よう」などの指標を含む直喩表現，あだ名などをはじめ隠喩表現と考えられる用例である。このようにタヌキ・マムシ・オオカミなど知識率の高い動物は比喩用例が全用例の一割以上と多いため，典型的と考えられる喩えの用例が取得可能である。以下に例を示す。

例）婿には人間出身ではなく猛々しい蝮のような悪い男を待つがいい。翼を持って虚空を自在に飛行する，炎と剣をもってすべてのものを懲らしめる男を。
（高橋康雄「ギリシアの女神の物語」）

例）あんまり無防備だと，狼に食べられちゃうぞ。ま，その狼も，今は病人だからおとなしいけどさあ。でも，あんまり可愛いことばっかりしてると―わからないんだからな。
（七海花音「僕らのロビン・フッド宣言」）

但し，これらの用例は喩えられた人の特徴から当該動物の特徴が類推されるのであって，動物の特徴から対象動物が想定されるのではない。たとえば，「したたかな人や腹黒い人を

喩えて言う」という用例がタヌキの同定に有用な情報とは必ずしも判断されなかった（50%）。「体を硬直させる様子を喩えて言う」という用例が，オットセイの判断に結びつくこともなかった（0%）。必ずしも，比喩表現に用いられた特徴が，対象物の認識に役立つのではない。

しかし，「比喩表現に適切と考えられる特徴を有した動物」という ad-hoc なカテゴリーにおいては，当該動物が選定されるための他メンバーとの差異が明らかになりやすい可能性はあろう。今後，比喩表現を含む用例を詳細に分析することで，対象物を認識するためのテキスト情報が得られることが期待される。

３.５　まとめ①：コーパスから取得した用例で対象物が同定可能か

本稿の調査対象とした動物においては，コーパスから取得した用例から対象物の同定が可能な動物と難しい動物に大別できた。

対象物を読み取るための情報は，文化的知識（当該対象が登場する著名な物語名など）＞一般的経験知識（比喩表現に用いられる知識・前提的とされる知識など）＞個別的経験知識（読み手と一致する場合）の用例の順に有用とされていた。対象物を同定するために有用と考えられる知識経験を喚起する情報は，コーパスから取得しやすい。しかし，読み手と共有可能な知識経験が多い動物であれば，用例から対象物の同定が可能となるが，異文化圏の知識を要する用例や，そもそも未知であろう情報が発信されている用例などが多い動物は，対象物の同定が困難となるのである。テキスト情報は個別的あるいは専門的すぎることが多く，また，対象物そのものの情報は得られても，想定されるカテゴリーにおける他メンバーとの差別化をする情報は得にくいためである。

４．調査②：大規模 Web コーパスから用例を取得する

調査①で用いた動物は，取得できた用例数が最大のタヌキで 41 例（372 サンプル/581 件）であった。1 億語規模の BCCWJ では，取得可能な用例が少ないために，調査対象とした動物の十分な情報が得られないという可能性は考えられる。ここでは，用例数が多ければ，多くの適切な情報が取得できるのであろうかという問題について，大規模なコーパスを用いて検証する。また，個別のコーパスの特性も確認する。

４.１　データ②

Google-日本語Ngram（Web から抽出された約 200 億文（約 2,550 億語）[11]の日本語データ）を用いて，取得可能な用例を調査した。

調査①で調査した動物のうち 3 種類の動物について用例（n-gram データ（1～7 gram）・頻度 20 以上）の抽出を行った。タヌキ（異表記「狸」「たぬき」を含む：1,893,000 件），オットセイ（61,800 件），ジュウシマツ（異表記「十姉妹」を含む：33,500 件）が取得できた。

但し，本稿のような意味的情報を取得する試みにあたっては，まず，文などが最大 7gram で分割されているという問題がある。そこで，手作業により，7gram で分割されている複数語の重なる同件数の用例を合わせることで，意味の把握が可能な長さとし，以下のように最大例で 23gram を取得した。

例）老け顔アンパンマンおばさん狸顔だよね
　　コンテンツの著作権はスタジオタヌキが所有しています
　　水族館といえばイルカやオットセイなどによるショーをして

さらに，同内容と考えられる用例を，調査①の作業（データ①参照）と同様に，意味的

[11] 総単語数は 255,198,240,937，総文数は 20,036,793,177。

な用例としてまとめた。

取得用例）A．サックスのレース&可愛いオットセイ柄のブラ&ショーツ
　　B．オットセイ柄のカットソー&スパッツ

A・Bのまとめ例：意味的な用例）衣類の柄に用いられることがある

また，当該動物そのものではない固有名詞（HN・店名・キャラクター名など）と判断される用例は分けた。但し，Webベースの大規模コーパスにおいては固有名詞や固有表現が大部分を占めているといえるが，完全な分類は困難であるため，分類は作業者の調査と判断によった。

なお，この作業では，同ページから重複取得されていると考えられる用例も散見されていた。たとえばジュウシマツにおける「食事と音楽，本，ジュウシマツとラブラドール等自身のアンテナが向いたもの」という用例が 1,210 件取得されるが，この句は特定のブログ（http://suzusuzu.jugem.jp/）における説明部分に含まれると確かめられた。

このような例は，ブログやサイトのタイトル，メニューなどの説明文に検索語が含まれているために，重複カウントされている場合が多いようだが，書籍タイトル「キツネとタヌキの大研究―人間との長くてふかーいつきあい（348 件)」や，演劇タイトル「ミュージカル吾が輩は狸である（3,670 件)」，商品紹介「劇場版どうぶつの森キャラポーチ全5種 タヌキ商店 DS 小物（282 件)」などの場合もあった。また，特定の質疑や説明等が，「【オンラインゲーム】トリックスターの狸育成方法について質問します（302 件)」「(名前が分からないのですが,) よく悪代官や悪徳商人する人で顔はタヌキ顔，ちょっと太りがちで強くはない（227 件)」「東京の多摩丘陵を舞台に，そこに棲むタヌキたちが人間に反旗をひるがえすべく（346 件)」のように，別 URL から複数取得されている可能性のある場合もみられる。

これらの特定例は，「(まめ) たぬきの雑記（24,200 件)」のように「たぬき」用例全件（842,000 件）の 3%を占める高頻度のものもあるほか，文を含むレベルでも「ここをクリックすると讃岐のタヌキのランキングポイントが加算されます（1,250 件)」「ぽんぽこ狸の考え方　社会問題等様々なことについてぽんぽこ狸が，独断と偏見で説教します。(1,010 件)」のように，多数の重複用例として取得される。

大規模であるが Web ベースのコーパスを用いると，Google-日本語 Ngram の場合，同 URL から重複取得されている用例が多いことや，固有名詞や商品紹介などの重複を除いたことから，用例数は取得数の 13%程度の量となった。また，このような重複ページの多さによって，本稿で示した Google-日本語 Ngram から取得した用例の頻度情報に正確性の疑問がある。

４.２　結果と考察②

当然 Ngram であるという取得可能な文字数の制限により，取得できない情報が多いことも予測されるが，単純な検索結果としては約 190 万件の規模が取得できたタヌキ用例（異表記「狸」「たぬき」を含む）であっても，取得できる意味的な用例の種類は 28 例となり，莫大になるということはなかった。そこで，取得した意味的な用例と調査①で BCCWJ から取得した意味的な用例（データ①）との対照を行った。データ①②には，タヌキで 22 例（調査異なり計 44 例中)，オットセイで 6 例（同 25 例中)，ジュウシマツで 4 例（同 17 例）の重複（意味的に同内容）があった。差異の生じた用例に着目し，Web コーパスと BCCWJ に差が見られた情報を分類した。それぞれ得やすいと考えられる情報をまとめておく。

４.２.１　大規模 Web コーパスから得やすい情報

まず，コーパスの規模によって取得可能となる用例がある（以下 A)。単純に意味的な用例の種類が増えるというものもいくらかあるが，特に，頻度が得られることで「多い」という情報の得られることが有用となる例である。取得された用例が個別的であるのか，一

般的であるのかが，頻度情報によって分類可能となるためである。
但し，均衡コーパスと異なり，個人的経験・評価，商品情報が多く取得されることから，取得可能となる用例に見られる偏りが考えられる（以下 B)。

A.　用例件数の量に関わる情報
A-1.　コーパスの規模が大きいため取得できる（別種類の）用例による
例）剥製にされていることがあり，置物になっている。(タヌキ)
シャチなどに襲われることがある。(オットセイ)
小斑があるものもいる。(ジュウシマツ)
歌に複雑な文法があるという説があり，書籍もある。(ジュウシマツ)
A-2.　個別のタイトルや固有名詞，複合名詞などの用例の多様さによる
例）三大伝説は，證誠寺・分福茶釜・隠神刑部とされる。(タヌキ)
うどんやそばにこの動物の名がついた種類がある[12]。冷やしたものもある。
丼や握り飯など米を用いたメニューもある。(タヌキ)
A-3.　共起情報の頻度による
例）この動物に喩えるのは，特に中年以上の男性や猫が多い。(タヌキ)
イタチ，河童，ウサギ，猪などと一緒に扱われることが多い。(タヌキ)
アザラシ，ペンギン，イルカなどの海獣類と並列されやすい。(オットセイ)
子供の頃など，昔飼っていたという人が多い。(ジュウシマツ)
A-4.　同種の用例が複数得られることによる（詳細情報）
例）タマネギを炒めるなどしてこの動物の色という表現がある[13]。(タヌキ)
青いこの形に似た猫型ロボットが著名である[14]。(タヌキ)

B.　均衡コーパスとの性質的差異として得やすい情報
B-1.　一般的評価・俗説
例）可愛い。いたずらをすると思われている。(タヌキ)・愛らしい。(ジュウシマツ)
B-2.　商品情報
例）ぬいぐるみや人形に模られる。(タヌキ)
陰茎や睾丸，骨格筋から抽出したエキスが加工食品に用いられる。(オットセイ)
B-3.　一般的経験
例）同種の鳥類とあわせて，家族・ペアなどで複数飼いされる。(ジュウシマツ)
水族館でショーが見られる。(オットセイ)
鳴き声に特徴があるといわれ，鳴きまねをする人や動物がいる。(オットセイ)

４.２.２　BCCWJ から得やすい情報

本稿の調査に用いた Google-日本語 Ngram では，そもそも文単位の検索が不可能であり，文脈情報が得にくい。たとえば「タヌキみたいな猫」が何をもってタヌキに喩えられたのか，タヌキの情報を読み取るためには，前後の文脈が必要となる（以下 C)。
意味的な用例を取得するためには，文脈情報を必要とする例も多く，現実的に運用される大規模コーパスから得られる情報には制限があると考えられる。また，専門性のある情報についての用例は，Web コーパスからは得にくいものである（以下 D)。

[12] BCCWJ から取得される「タヌキ」料理は，「関西では油揚げいりのそば・うどんを示す」例のみであった。
[13] BCCWJ の「タヌキ色」の用例は，「タヌキ色のインキ（橋爪和子「ぼくはひとりぼっちじゃない」)」のみで，詳細情報はない。
[14] BCCWJ でも同種の用例は取得されたが，「猫型ロボットなのに（中略）昔の世界に行くとよく狸と間違われやすいですが（Yahoo!知恵袋)」という一例に留まっていた。

C.　文脈から得られる情報（下線部はGoogle-日本語Ngramからも高頻度で得られる情報）
C-1.　詳細・関連・補足情報
例）庭で見ることもある。餌付けされて野生に戻らなくなることがある。（タヌキ）
里山で食害の被害の原因となる。雑食性で人家の残飯をあさることもある。
荒毛の下に柔らかい上質の毛皮を持つ。長い毛1本に短い毛が約五十本もあり，保湿効果を高めている。（オットセイ）
鵺の胴がこの動物であると平家物語に描かれている。（タヌキ）
小型で11ｃｍ。（ジュウシマツ）
C-2.　文をまたぐと考えられる情報
例）高速道路などでよくはねられて死んでいる[15]。（タヌキ）
全身が硬直した様子などを，この動物になったと喩える例がある。（オットセイ）

D.　一般的に言及されにくい情報
D-1.　専門的知識[16]
例）ヒレ状の前後の脚はアシカより長く，水中の生活に適応している。（オットセイ）
プラスチックの悪影響の代表的例として，網絡まりがあげられる。（オットセイ）
平安時代は猫をこの用字で表していた。（タヌキ）
日本で作り出された。（ジュウシマツ）
D-2.　限定的知識
例）サイコロを二つ使うことをこう呼ぶゲームがある。（タヌキ）
夕食抜きの素泊まり客をこう呼ぶ。（タヌキ）
D-3.　経験が得にくい
例）アベル・タスマン国立公園では，人を恐れずじゃれついてくる。（オットセイ）
水面からひょっこり丸顔をのぞかせる。（オットセイ）
子育てが上手い。（ジュウシマツ）

４.２.３　個別のコーパスに依拠する情報

その他，本稿で使用したコーパス各々に依拠すると考えられるために，重複のない用例が見られている。サンプリングされたテキストの生産時期により生じた差異と考えられる。

E.　サンプリングされたテキストの生産時期の影響が考えられる情報
E-1.　催事など
例）世界じゅうのこの鳥の展覧会が行われる。（ジュウシマツ：Google 日本語 Ngram）
その他，演劇名・映画名などの固有名詞とそれに関する情報[17]
E-2.　流行など
例）タヌキケーキ。（タヌキ：Google 日本語 Ngram）
E-3.　その他時代的と考えられる情報
例）酒をよくのむといわれる。（タヌキ：BCCWJ）

---

[15] 「死骸」「死体」「～を見る」類は高頻度であるが，「高速道路」のような場所情報が得られておらず，文が別になっていることが考えられる。

[16] 雑学やクイズなどから専門的な用例が得られることはあり，「語源はアイヌ語といわれている。（オットセイ）」のような専門的知識が，Webコーパスからのみ得られている場合もある。

[17] 固有の作品が公開・放送などされることにより，個別的な表現をはじめとする関連情報のテキストが増加するほか，個人の感想や意見傾向がテキスト化される影響が考えられる。
例）「平成狸合戦ぽんぽこ（映画）」：1994 年 7 月 16 日公開で，テレビ放送が繰り返されている。
例）「吾輩は狸である（ミュージカルコメディ）」：2003 年 4 月 7 日初演。

５．まとめ

テキスト情報から対象物を同定する試みは，明確に対象物の認識が可能なテキスト作成のために必要な情報を整理する試みの一つである。最後に，調査①②の結果から，対象物を同定するためにコーパスから取得可能な情報について考察する。

コーパスを用いた用例からの対象物の同定は，平均正答 25%と低く，基本的には困難であるといえる。よって，テキスト情報による経験のみで人が対象物をどのようなものか認識することも当然難しいのだと考えられる。また，対象物の同定正答率は，動物によって高い群と低い群に大きく二分されたという結果でもあった。タヌキのように 100%の実験協力者が対象物を同定可能な場合には，コーパスによって提供された情報（意味的な用例）が，十分であったということである。但し，個人の有する情報と合致したかどうかという点が有用性の有無の判断ともなっていた。すなわち，対象物の同定には，文化的レベル＞一般的レベル＞個別的レベルの順に有用性があると考えられる。

これは，92%の実験協力者に有用とされた「カチカチ山」のストーリーおよび登場人物情報など，テキストに記述された情報外の情報を対象物の認識に求めるということではあろう。しかし，テキスト情報から対象物を同定するために有用とされた情報が，一般的，さらには文化的であることを求められていたという点から，大規模コーパスによって頻度情報が提供されることが望ましいともわかった。取得された用例が個別的なレベルの情報であるのか，一般的レベル以上であるのかは，ある程度，頻度情報によって分類可能となる。頻度が得られること（「多い」という情報の得られること）が有用と考えられる。

なお，コーパスからは，経験知識に関する情報は得やすいのではあるが，読み手の想定したカテゴリーの成員との差異を示す情報は得にくいこともわかった。これらの情報を取得するためには，今後，比喩表現などをはじめとする「人」の特徴と対象物との関係（人を基準とした異同情報），比較対象として他の動物と並列される場合の関係などに着目したい。比喩表現などの分析には文脈情報が必要となるが，文脈が不明の場合でも，用例数が増加することによって，並列される動物は取得できる可能性がある。対象物そのものの記述のほか，同カテゴリーの成員との異同によって情報を提供する方向性も考えられよう。

以上のように，テキストに記述された対象物の認識には，読み手が一般的に有していることが期待されるテキスト外の知識の喚起が求められている。コーパスから取得した用例から対象物を認識するためには，テキストそのものから得られる情報に留まらず，広く文脈が要されるということでもあろう。

文　献

天野成昭，近藤公久編（1999）『日本語の語彙特性』第 1 巻（単語親密度）三省堂

天野成昭，近藤公久，笠原要編（2008）『日本語の語彙特性』第 9 巻（単語親密度増補）三省堂

保田祥，浅原正幸，前川喜久雄（2013）「何が記述してあればテキストの示している対象物がわかるのか」日本認知科学会第 30 回大会大会論文集，pp.370-379.

Fillmore, Charles J. and B. T. S. Atkins (1994) “Starting where the dictionaries stop: The challenge for computational lexicography”, In Atkins, B. T. S. and A. Zampolli, Ed. *Computational Approaches to the Lexicon*, 349-393. Oxford: Oxford University Press.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1992) “Trust the text.” In M.Davies and L.Ravelli, Ed. *Advances in Systemic Linguistics* : recent theory and practice, 5-19. London: Pinter.

Yasuda, S., Okamoto, M., & Aramaki, A. (2012) “‘Mind the Gap’ between Text and Real World: A Corpus-based Study on the Prototype Effects of Animal Body Parts.” 4th UK Cognitive Linguistics Conference.

『現代日本語書き言葉均衡コーパス』の文境界修正作業の進捗

小西 光 *†　中村 壮範 †　田中 弥生 †　浅原 正幸 †　今田 水穂 †
山口 昌也 ‡†　前川 喜久雄 ‡†　小木曽 智信 ‡†　山崎 誠 ‡†　丸山 岳彦 ‡†
（国立国語研究所　† コーパス開発センター　‡ 言語資源研究系）

Revision of Sentence Boundaries in the BCCWJ DVD Edition

Hikari Konishi, Takenori Nakamura, Yayoi Tanaka, Masayuki Asahara, Mizuho Imada
Masaya Yamaguchi, Kikuo Maekawa, Toshinobu Ogiso, Makoto Yamazaki, Takehiko Maruyama
(National Institute for Japanese Language and Linguistics)

1. はじめに

本発表では現在国立国語研究所コーパス開発センターで進められている『現代日本語書き言葉均衡コーパス』(以下 BCCWJ) に対する文境界修正作業について報告する。文境界の認定には (1) 文字列レベルの情報を用いるもの，(2) 形態素列レベルの情報を用いるもの，(3) 係り受け関係を用いるものなどが考えられる。現在公開している BCCWJ DVD 第 1 版においては，(1) の文字列レベルによる処理で文境界認定が行われているが，不自然な文境界が残っていることが報告されている。人手による作業にせよ自動処理にせよ，より高レベルのアノテーションに基づくものほど高コストになる一方，より厳密な文境界の認定が可能であり，コアデータに対しては先行研究（小西ほか (2013)）において (3) の係り受け関係レベルの文境界再認定が行われた。しかしながら，非コアデータ規模になるとこのレベルの修正は非現実的であり，現在自動認定された形態論情報に基づく (2) のレベルの文境界修正作業を実施している。本稿では実作業の詳細と進捗状況を報告する。

2. 文境界認定手法についての関連研究—手がかり・CSJ における研究動向・BCCWJ の現況

本節ではまず文境界認定基準策定のために必要な手がかりについて述べ，次に BCCWJ の前に作成された『日本語話し言葉コーパス』（以下 CSJ）における文境界認定に関する関連研究を示し，最後に BCCWJ 第 1 版公開時における文境界認定とその後の研究動向について述べる。

2.1 文境界認定基準における手がかり

文境界認定基準において何らかの「手がかり」を用いて規則を人手で記述する必要がある。ある程度文境界認定作業を自動化するために何を「手がかり」に使うかが重要である。以下では「手がかり」として，(1) 文字列レベルの情報を用いるもの，(2) 形態素列レベルの情報を用いるもの，(3) 係り受け関係を用いるものの 3 種類について詳しく述べる。

(1) 文字列レベルの情報とは，句点などに基づいて文境界を用いる手法である。多くの形態素解析の前処理として句点記号「。」「.」感嘆符「！」疑問符「？」などを手がかりとして文境界認定が行われている。少し高度な情報として開き括弧や閉じ括弧を用いた規則を記述し，括弧の対応を取る手法がある。

* hkonishi@ninjal.ac.jp

(2) 形態素列レベルの情報とは，形態素解析により認定される品詞情報などを用いる手法である。句点のリストを品詞「記号-句点」などに汎化できるほか，開き括弧や閉じ括弧についても「記号-括弧開」「記号-括弧閉」と汎化して記述することができる。さらに，辞書に登録されている固有名詞や顔文字などに埋め込まれている記号などを文境界候補から除外することができる。一方，形態素解析誤りの影響をある程度見込んで処理する必要がある。

(3) 係り受けレベルの情報とは，係り受け解析により認定される係り受け関係を用いる手法である。括弧内の要素が文であるかどうかを認定するために括弧内の要素が係り受け木をなすかを判定したり，括弧の前後で係り受け関係があるかどうかで sentence 要素の入れ子を認定したりする。

この三種類以外の手がかりとして，CSJ においては形態素列レベルの情報で認定された節境界の情報や音声のポーズ長などを用いる研究が知られている。次節では CSJ における文境界認定についての様々な取り組みについて紹介する。

2.2　CSJ における文境界認定と関連技術

丸山ほか (2006) は CSJ における統語的単位について議論している。南 (1974) による従属節の分類に基づき，「絶対境界・強境界・弱境界」と呼ばれる三段階のレベルの節境界が設計・定義され，各従属節の境界にラベルが付与された。以降の研究では，「絶対境界」が CSJ における文境界として利用されてきた。

下岡ほか (2004) では CSJ の講演の書き起こしテキストの文境界認定について，話者がとるポーズ長と前後の単語情報に基づいた文境界認定手法を提案した。これに対し，田島ほか (2003) は同じデータでポーズ長が得られないことを想定し，コスト最小法の形態素解析器を用いて，句点を挿入した場合と挿入しない場合との出力コストの比較を行い文境界認定を行う手法を提案した。一種の言語モデル尤度を用いた手法とも言える。福岡・松本 (2005) は田島らの手法を拡張して言語モデル尤度を特徴量とした文境界認定手法を提案している。下岡ほか (2005) は新たに係り受け情報を用いて文境界認定する手法を提案している。この手法においては話し言葉特有の係り受け現象を扱う係り受け解析器を導入し，ポーズ長・節末表現・単語情報・文節間距離・係り受け関係などを複合的に組み合わせて文境界を認定している。西光ほか (2009) は丸山らの三つのレベルを全て認定する手法を提案している。特徴量として局所的な隣接要素間の係り受け関係のみを扱うことにより精度の向上が達成されたことを報告している。またこの論文では音声認識結果からの文境界認定についても議論している。

このように CSJ においてはさまざまなレベルの情報を利用した文境界認定手法が提案されてきた。しかしながら，CSJ 関連の文境界認定の重要な問題として，文末認定（文の最右要素）しか行われておらず，文の最右要素と最左要素の対応が取られていないという点がある。入れ子による観点に基づいた文境界を認定がなされていないために，基本的にはチャンキングなどの系列ラベリングで処理可能なレベルの文境界認定にとどまっている。

2.3　BCCWJ における文境界認定 — 本研究に至るまでの足取り

本節では，本研究に至るまでの BCCWJ における文境界認定について述べる。まず BCCWJ 第 1 版公開時における文境界認定の基準について述べ，次に係り受けアノテーション（BCCWJ-DepPara：浅原・松本 (2013)）構築時に行った文境界認定（小西ほか (2013)）について述べる。

2.3.1　BCCWJ 第 1 版における文境界認定

まず，BCCWJ 第 1 版における文境界について述べる。BCCWJ 第 1 版においては文字情報に基づく C-XML 形式と形態論情報に基づく M-XML 形式の二種類の XML 形式のファイルでデータが表現されている。この二種類の形式において認定している文境界に差異がある。

C-XML における文境界認定：

C-XML 形式においては手がかりとして文字列レベルの情報を用いた自動処理に基づく文境界認定（山口ほか (2011), pp.136–138.）が基本となっている。話し言葉や既存の書き言葉コーパスと異なり，元媒体のレイアウト情報に基づく文書構造情報（ブロック要素）が利用されている。以下 C-XML における文のスパンを表現する sentence 要素の認定規則について例（図 1）を示しながら解説する。自動認定においては句点記号「。」「.」感嘆符「！」疑問符「？」（以下 “文末記号”）やブロック要素開始位置直前を文区切り位置とみなし，直前文の末尾とみなす処理を行う（例 C-1）。論理行[*1]頭から一つ以上の sentence 要素の並びが存在する場合で行末に文末記号がない場合は sentence 要素とみなす（例 C-2）。論理行中に一つも sentence 要素がなく文末記号もない場合その論理行全体を sentence 要素とみなす（例 C-3）。これらの文末記号以外によって認定される sentence 要素は，特殊な文として属性 type=“quasi” を付与する（例 C-2，C-3；以下 “sentence@quasi 要素” と略記し，type=“quasi” がつかない sentence 要素を “正則な sentence 要素” と呼ぶ）。文字列レベルの情報として 9 種類の括弧（括弧類Ａ）[*2]の対応などを用いて，文認定時に sentence 要素の入れ子を許している。

括弧内に一つも文末記号を含まない場合，括弧内に sentence 要素を認定しない（例 C-4）。括弧内に一つ以上の文末記号が含まれる場合，括弧内に sentence 要素を認定する（例 C-5）。括弧内に一つ以上の文末記号が含まれ，且つ，閉じ括弧直前に文末記号が出現しない場合，閉じ括弧直前までの部分を特殊な文とみなし，属性 type=“quasi” を付与する（例 C-6）。

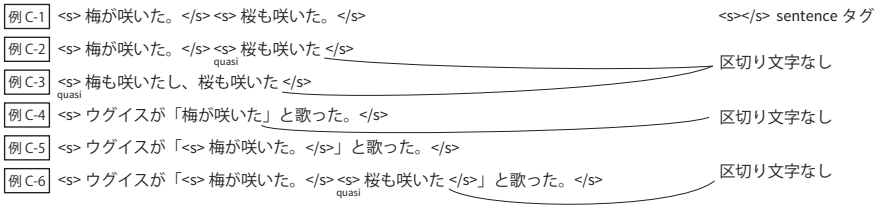


図 1　C-XML における文境界認定

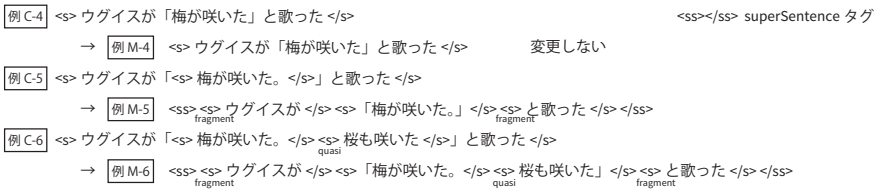


図 2　C-XML から M-XML への変換

[*1] 本稿では紙面などの物理的制約によって指示される行を「物理行」「表示行」と呼ぶのに対して、改行コードやブロック要素などにより指示される行を「論理行」と呼ぶ。

[*2] 括弧類Ａ：「補助記号-括弧開」「補助記号-括弧閉」のうち（ ）〔〕［ ］｛ ｝〈〉《》「」『』【】9 種。

M-XML における文境界認定：

M-XML 形式（国立国語研究所 (2011)，p. 94.）においては，C-XML の文境界認定を基礎としつつ，C-XML とは異なる，より単純化した文境界認定を行う方針を採用した。方針提案者は C-XML の問題点として以下の三点をあげている：sentence 要素がきわめて長くなる場合がある；形態素解析などの入力となる「文」が定めがたい；データを文番号で管理できない。

M-XML では，C-XML において sentence 要素が入れ子になっている場合に，その最も内側（下位）にあるもののみを正則の sentence 要素とし，外側（上位）にある sentence は superSentence とする。その上で，superSentence の内側にありながら正則の sentence 要素の外側に位置する部分は，新たに sentence 要素と見なすとともに type=“fragment” という属性を与えて，文断片（以下 “sentence@fragment 要素” と略記）であることを明示する。この際，括弧記号のみから成る文断片要素を作らないために，内側の sentence 要素に隣接する括弧記号を送り込む。最終的に superSentence と sentence の 2 階層からなる文境界情報が残される（図 2）。

例 C-4 においては sentence 要素に入れ子が発生していないため，C-XML 形式と M-XML 形式の sentence 要素は一致する（例 M-4）。

例 C-5 においては，括弧内の最内スパンの sentence 要素を M-XML における正則な sentence 要素と見なす（例 M-5）。例 C-5 における最外スパンを新たに superSentence 要素として認定する。正則 sentence 要素に含まれない最外スパンの連続文字列を sentence@fragment 要素として認定する。ただし正則 sentence 要素に隣接する括弧記号は sentence 要素に送り込む。

例 C-6 においては括弧内に正則な sentence 要素と sentence@quasi 要素の二つが認定されている。例 C-6 における最外スパンを新たに superSentence 要素として認定する（例 M-6）。括弧内の 2 種類の sentence 要素（正則な sentence 要素と sentence@quasi 要素）を認定し，これに含まれない前後の連続文字列を sentence@fragment 要素として認定する。ただし，内側の sentence 要素に隣接する括弧記号は内側の sentence 要素に送り込む。

しかし，例 M-5・M-6 における，「内側の sentence 要素に隣接する括弧記号は内側の sentence 要素に送り込む処理」が網羅的ではなかった。今回はこの問題を解決するために網羅的なパターンを記述し，再処理する。

2.3.2　BCCWJ-DepPara における文境界認定

前小節の状況は，C-XML の方式にしても M-XML の方式にしても係り受けアノテーションにとって好ましくない。係り受けアノテーション従事者は BCCWJ DVD データ 第 1 版 における文境界の問題点として以下の四点をあげている：基準の手がかりが文字列に基づく手法であるために，係り受けを分断するような文境界が大量に発生する；sentence@quasi 要素や sentence@fragment 要素においては，要素内に係り先が存在せず，離れた別の sentence 要素に係り先を認定するような現象が起きる；全要素を xpointer などを用いない一つの XML ファイルとして表現するために，ad hoc な後処理がなされ，文単位認定に無理が生じている；実データを見ても，必ずしも報告書通りの処理がなされていない。

そこで，小西ほか (2013) は，係り受けアノテーション向けの文境界認定基準を策定し，コアデータに対して人手による全数確認により，BCCWJ DVD 第 1 版とは異なる文境界を付与した。基本方針として，元の文書構造タグを用いず，文の内容に即して “EOS” ラベルと “Z” ラ

ベルの 2 種類の文境界を認定している。“EOS” ラベルは，係り受け関係がつながる範囲で文を連結したもので C-XML の最外スパンや M-XML の superSentence 要素に近い基準となっている。“Z” ラベルは，係り受け関係ラベルの一種（浅原 (2013)）で “EOS” ラベルで区切られる範囲内に出現する文末記号の出現に対し付与される。“Z” ラベルは文末要素にしか付与されないが，“Z” ラベルを根とする係り受け木の最大スパンを確認することで，局所的な文の文頭要素が認定できるために実質的に文の入れ子構造を認定している。括弧内の要素の扱いにおいては，コアデータに出現する括弧で括られた要素の機能を補足・発話・心内・引用・箇条書き・強調の 6 種類に分類し，要素の意味についてまで調査して，文認定を行っている。

3. 今回の文境界認定作業の概要

3.1 文境界認定の作業方針

以下に文境界認定の作業方針について述べる。BCCWJ DVD 第 1 版の文字列レベルの情報による自動処理と，BCCWJ-DepPara の係り受けレベルの情報による人手修正との中間的な処理として，**形態素列レベルの情報を用いた自動抽出結果の人手修正**をコアデータ・非コアデータ全体に対して実施する。

修正方法としては，まず C-XML 形式における文字列レベルの情報を用いた文境界認定におけるバグ相当のものを自動抽出して人手修正し，M-XML 形式に変換する際のバグ相当のものを形態素列レベルの情報を用いて自動抽出してバッチ処理および人手修正を行う。基本的に最内スパンの正則な sentence 要素を認定するとともに，その作業に伴い発生する sentence@quasi 要素，sentence@fragment 要素のような文が認定されることを許す。係り受け関係の整合性は検証しないが，括弧内の要素について最低限の確認作業（強調や補足の認定）を行う。まとめると以下のようになる：

処理 C C-XML 形式レベルで認定できる誤りの検出

BCCWJ DVD 第 1 版において，文字列レベルの情報に基づく処理により 9 種類の括弧内（括弧類 A）に，文末記号があるが文境界が設定されていない要素が約 6,000 箇所[*3]発見された。顔文字に埋め込まれた文末記号があったり，括弧が対応していない事例もあったり，全数人手で確認する。

処理 M M-XML 形式レベルで認定できる誤り検出

処理 C が完了後，形態素列レベルの情報を用いた誤り検出を行う。形態素列レベルの情報を用いた誤り検出においては，国語研コーパス開発センターに寄せられている様々な誤り報告事例や他のアノテーション作業時に問題となった事例をもとに，人手で形態素レベルの情報を用いたパターンを記述した。このパターンの認定においてはそのマッチする事例のうち修正率（真に修正すべき事例数／マッチする事例数）に基づいて 2 種類の処理を行う。

M(α) 修正率が高いパターン：マッチするほとんどの事例が真に修正すべき事例であるが，例外的に修正しなくてもよい事例が出現するパターン。これらについては，バッチ処理適用前に例外的な事例を排除するように人手で確認する。人手確認後

[*3] なお，各箇所で複数の文境界の修正が発生するために実際に修正する文境界はこの数字より大きい。

バッチ処理で修正する（自動修正箇所抽出 → 人手例外確認 → バッチ修正処理）。

M(β) 修正率が低いパターン：マッチする事例の一部のみを修正するパターン。全数確認は困難であるが，修正すべき事例が含まれうるパターンを先にバッチ処理で展開し，逐一人手で確認する（自動修正箇所抽出 → 人手修正処理）。

3.2 文境界認定基準

3.2.1 文境界認定基準の前提

文境界認定基準の前提について示す。まず現存する superSentence 要素を踏襲することを前提に sentence タグを付与する。助詞・助動詞から始まる，助詞・助動詞で終わる，助詞・助動詞のみの sentence 要素の発生を認める，括弧内に文末記号が含まれない場合には sentence タグは付与しない（例 C-4, 例 M-4 を踏襲）。以下 3.2.2 節では括弧内に文末記号が含まれる場合に対してパターンを定義し修正作業を行う（処理 M(α)）。人手で例外を確認し，必要に応じて新たなパターンを追加する。本稿に示すパターンは 2014 年 2 月時点までに設定したものであり，今後作業時に新たなパターンが増えることが考えられる。3.2.3 節ではパターンに基づく機械処理で一括処理できない事例を中心に，認定を人手で行う（処理 M(β)）。以下 sentence 要素を s 要素と略記する。

3.2.2 処理 M(α)：修正率の高いパターン・認定基準

以下修正率の高いパターンについて示す。これらは（自動修正箇所抽出 → 人手例外確認 → バッチ修正処理）の手続きで誤り修正される。

1. 句点類Ｂ[*4]のみ，もしくは，句点類Ｂの前に記号類Ｃ[*5]があり，且つ，句点類Ｂと記号類Ｃのみで構成されている s 要素は，前の s 要素の末尾に移動[*6]

例 F-1：PB26_00004 （1 行が 1s 要素，横線上が修正前・横線下が修正後）

```
<S>でも、お客様が並んでしまったら、それより早めに放送してください」</S>
<S>。</S>
----------------------
<S>でも、お客様が並んでしまったら、それより早めに放送してください」。</S>
```

2. 【原則】〔括弧開〕[*7]で終わっている s 要素は，次の s 要素の頭に〔括弧開〕を移動

例 F-2：PN1b_00009

```
<S>それより「ブラボー砦の脱出」だ、「星のない男」だ（</S>          ←注目点
<S>異議なし！</S>
<S>）。</S>
----------------------
<S>それより「ブラボー砦の脱出」だ、「星のない男」だ</S>
<S>（異議なし！）。</S>
```

(a) 【例外処理】〔括弧開〕の前がすべて空白[*8]の場合も，それらすべてを次の s 要素の頭に移動

[*4] 句点類Ｂ：「補助記号-句点」。！．？ の 4 種。

[*5] 記号類Ｃ：「補助記号-一般」（文境界を示す） ―…－・～〖〗〔〕‐‥」♪♫《》－－ー の 20 種。

[*6] 条件を規定する演算子は，打消の助動詞を否定とし、「且つ」を論理積とし、「もしくは」を論理和とした場合に，この順で優先順位が高い加法標準形で記述する。

[*7] 今回は形態論情報により括弧として定義されている「補助記号-括弧開」「補助記号-括弧閉」の全 12 種を用い、それぞれ〔括弧開〕・〔括弧閉〕と呼ぶ：〝 〟 〈 〉 「 『 【 □ （ ） < > 〔 〕 。

[*8] 例中□は全角の「空白」を表す。

例 F-3：OY14_12372

```
<S>□□□□□□□『</S>                          ←注目点
<S>今度は□一緒にファーストで行きたいね□！！</S>
<S>□』</S>
----------
<S>□□□□□□□『今度は□一緒にファーストで行きたいね□！！□』</S>
```

3. 【原則】〔括弧閉〕のみ，もしくは〔括弧閉〕で始まり，且つ，〔括弧閉〕と記号類D[*9]のみで構成されたs要素は，前のs要素の末尾に移動

例 F-4：PN1b_00009

```
<S>それより「ブラボー砦の脱出」だ、「星のない男」だ（</S>
<S>異議なし！</S>
<S>）。</S>                                ←注目点
----------
<S>それより「ブラボー砦の脱出」だ、「星のない男」だ</S>
<S>（異議なし！）。</S>
```

(a) 【例外処理】上記3.を適用した結果，〔括弧閉〕（と記号類Dのまとまり）を移動した先のs要素が，〔括弧閉〕と記号類D・E[*10]のみで構成されている場合は，前のs要素の末尾に，それらを移動

例 F-5：PN2d_00008

```
<S>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（</S>
<S>？</S>
<S>）。</S>
----------
<S>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（</S>
<S>？）。</S>                    ←注目点：ここが記号のみ
----------
<S>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（？）。</S>
```

4. 【原則】〔括弧閉〕で始まり，且つ，〔括弧閉〕に任意の短単位が後続するs要素は，前のs要素の末尾に〔括弧閉〕のみを移動

例 F-6：PN5f_00020

```
<S>（咽喉？</S>
<S>）…と其奴がね、異に蔑んだ笑い方をしたものです。</S>
----------
<S>（咽喉？）</S>
<S>…と其奴がね、異に蔑んだ笑い方をしたものです。</S>
```

(a) 【例外処理】〔括弧閉〕に記号類F[*11]が続く場合は，記号類F以外の短単位が出現するまでの範囲を前のs要素の末尾に移動

例 F-7：OC06_00325　（この例では〔括弧閉〕と読点を移動）

```
<S>峠や市街地でも、追い越し禁止道路で前を走る多少遅い車に接近して（</S>
<S>あおるつもりじゃないが。。</S>
<S>）、車が遠慮して道を譲ってくれた時、だいたい頭を下げて追い抜きます。</S>
----------
<S>峠や市街地でも、追い越し禁止道路で前を走る多少遅い車に接近して</S>
<S>（あおるつもりじゃないが。。）、</S>
<S>車が遠慮して道を譲ってくれた時、だいたい頭を下げて追い抜きます。</S>
```

(b) 【例外処理】空白で始まり，〔括弧閉〕と空白のみでs要素を構成する場合は，それ

*9 記号類D：「空白」1種，「補助記号-一般」（文境界を示す）20種，「補助記号-句点」4種，「補助記号-読点」2種。「補助記号-読点」2種は、，からなる。

*10 記号類E：「記号-一般」2,003種，「記号-文字」255種，「空白」1種，「補助記号-AA-一般」78種，「補助記号-AA-顔文字」2,405種，「補助記号-一般」（文境界を示さない）444種，「補助記号-括弧開」12種，「補助記号-括弧閉」12種。

*11 記号類F：「補助記号-一般」（文境界を示す）20種，「補助記号-読点」2種，「補助記号-括弧閉」12種。

らすべてを前の s 要素の末尾に移動

例 F-8：OY14_12372

```
<S>□□□□□□□□『</S>
<S>今度は□一緒にファーストで行きたいね□！！ </S>
<S>□』</S>　　←注目点
----------
<S>□□□□□□□□『今度は□一緒にファーストで行きたいね□！！□』</S>
```

(c) 【例外処理】上記 2.，4.，4.(a) を適用した結果，「(？)」「(！)」の文字列を s 要素に含む場合には，前後の s 要素をひとま とまりにする
(“文境界認定を打ち消して文を結合する場合” の 1. を参照)

例 F-9：PM41_00071

```
<S>この業界にしては珍しく（</S>
<S>？</S>
<S>)、可愛らしい女性編集長である。</S>
----------
<S>この業界にしては珍しく（？)、可愛らしい女性編集長である。</S>
```

5. 読点で始まっている場合は，前の s 要素の末尾に読点のみを移動

例 F-10：PB45_00024

```
<S>「ブオノ・ヴェーロ？」</S>
<S>、美味しいだろうと言ったオジサンはイタリア人で、ここに住む孫のためにナポリの店を引き払いやって来たのだという。</S>
----------
<S>「ブオノ・ヴェーロ？」、</S>
<S>美味しいだろうと言ったオジサンはイタリア人で、ここに住む孫のためにナポリの店を引き払いやって来たのだという。</S>
```

6. 【原則】発話者「～」や発話者：「～」で表記される場合は，一つの s 要素とする

例 F-11：OY13_01632

```
<S>弟子：「</S>
<S>どうすれば先生のようになれるのでしょう？」</S>
----------
<S>弟子：「どうすれば先生のようになれるのでしょう？」</S>
```

(a) 【例外処理】括弧内が二つの s 要素に分かれる場合は，括弧内一つめの s 要素のみを前の s 要素に移動

例 F-12：OY14_08173

```
<S>ゆり「</S>                    <S>ゆり「美羅！</S>
<S>美羅！</S>                    <S>真ちゃん何があったの？」</S>
<S>真ちゃん何があったの？」</S>
```

3.2.3 処理 M(β)：修正率の低いパターン・認定基準

以下の例は修正率が低いパターンで，手がかりにより候補を枚挙したうえで人手で修正すべきかどうかを判定する。大きく分けて「文境界を認定して分割する場合」と「文境界認定を打ち消して文を結合する場合」の 2 種類ある。これらは（自動修正箇所抽出 → 人手修正処理）の手続きで誤り修正される。

文境界を認定して分割する場合（特に Web 関連データ）

1. s 要素の中に顔文字を含み，且つ，その顔文字が文末だと考えられる場合

例 F-13：OC06_02963（縦線左が修正前・縦線右が修正後）

```
<S>そーですよ＾＾一番左です＾＾</S>    <S>そーですよ＾＾</S>
                                    <S>一番左です＾＾</S>
```

2. s 要素の中に（涙）等の（X）を含み，且つ，その（X）が文末表示だと考えられる場合

例 F-14 : OY14_10161

<S>イプ『</S>
<S>違う！</S>
<S>作りすぎただけだっ（照）ナマモノだから今日中に食え』</S>

<S>イプ『違う！</S>
<S>作りすぎただけだっ（照）</S> ←注目点
<S>ナマモノだから今日中に食え』</S>

3. 【特殊事例】空白で文が区切られる場合等

例 F-15 : OY14_12372

<S>□□□□□□□□□『だね、ローマが一番だったよ□日曜なのでバチカンに行ってミサを聞いた</S>
<S>□□□□□□□□□□ミケランジェロも見たよ』□うん、おいらはイタリアはしらない</S>

<S>□□□□□□□□□『だね、ローマが一番だったよ□</S>
<S>日曜なのでバチカンに行ってミサを聞いた</S>
<S>□□□□□□□□□□ミケランジェロも見たよ』□</S>
<S>うん、おいらはイタリアはしらない</S>

文境界認定を打ち消して文を結合する場合（特に雑誌・Web データ）

1.「？」「！」等に係り受け関係を結べる要素が後続し，s 要素内に含めるべきと判断される文末記号

例 F-16 : PM11_00263

<S>今が買い！</S>
<S>の中古ＭＦ一眼レフ</S>

<S>今が買い！の中古ＭＦ一眼レフ</S>

2. 補足を表す丸括弧（括弧内に句点類Ｂを含まないものに限定）

例 F-17 : OY01_00185

<S>この大会のチラシを、今夜（</S>
<S>昨夜？</S>
<S>）のハードルの練習中にわざわざセタホールまで持ってきてくださったのです！</S>

<S>この大会のチラシを、今夜（昨夜？）のハードルの練習中にわざわざセタホールまで持ってきてくださったのです！</S>

3. 【原則】係り受け関係を結べる要素が，原本レイアウト情報を反映した結果二つの s 要素に分割されていて，括弧内に文末記号が含まれない場合

例 F-18 : PB1n_00024

<S>すると、</S>　←注目点：紙面上にて改行があり，s 要素が分割されている
<S>「溶岩流が危険だから、逃げるんです」という答えが返ってきたのである。</S>

<S>すると、「溶岩流が危険だから、逃げるんです」という答えが返ってきたのである。</S>

(a) 【例外処理】括弧が強調やタイトル等の目的で用いられている場合

例 F-19 : OC01_03215

<S>ゆうべＰＭ９時から日本テレビ「</S>
<S>ものまねバトルオール新ネタ！</S>
<S>夏祭りＳＰ</S>
<S>」に出てましたよ。</S>

<S>ゆうべＰＭ９時から日本テレビ「ものまねバトルオール新ネタ！夏祭りＳＰ」に出てましたよ。</S>

4. 【特殊事例】〔括弧閉〕に丸括弧で注釈が後続する場合は変更しない

例 F-20 : PN4c_00011

<S>□だが、農業団体の韓国農業経営人中央連合会は、</S>
<S>「通貨危機で金利負担が膨らみ、農家は今も借金に苦しんでいる。</S>
<S>対策は成功していない」</S>
<S>（政策調整室）と批判的だ。</S>

4. おわりに

本発表では『現代日本語書き言葉均衡コーパス』の文境界修正作業について報告した。

以下作業進捗について示す。M-XML に基づく文境界は約 587 万箇所にのぼる。2013 年度

内に処理 C の 文字列レベルの修正作業（約 6,000 箇所）が完了する予定である。表 1 に処理 M の想定修正箇所数の現時点での集計結果を示す。処理 M(β) の非コアについては今後集計するが，母集団の規模比率からコアの 100 倍を見込んでいる。2013 年度末までに形態素列レベルの修正作業のうち処理 M(α) に着手し，2014 年度は処理 M(α) と M(β) を進める。その過程で今後新たな例外事例が認定された場合には新たなパターンを記述することになるだろう。本稿におけるパターンは先行して行われた処理において発生した不備に対処するためのものである。本作業終了後に「最初にどういう文境界基準を立てるべきであったか」という課題に取り組む必要がある。これについては機械学習などを用いて文境界認定規則を自動獲得したうえで人間にとって可読性の高い規則に汎化するということを行いたい。

表 1　処理 M の想定修正箇所数

摘要 (“s 要素” = “sentence 要素”)	コア	非コア
処理 M(α)：(3.2.2 節) 修正率の高いパターン・認定基準		
1. 句点類Bのみ，もしくは，句点類Bの前に記号類Cがあり，且つ，記号類Cのみで構成されている s 要素は，前の s 要素の末尾にそれらを移動	104	2,898
2.〔括弧開〕で終わっている s 要素は，次の s 要素の頭に〔括弧開〕を移動	222	36,086
3.〔括弧閉〕のみ，もしくは，〔括弧閉〕で始まり，且つ，〔括弧閉〕と記号類Dのみで構成された s 要素は，前の s 要素の末尾に〔括弧開〕（とそれら記号類Dのまとまり）を移動	59	4,547
4.〔括弧閉〕で始まり，且つ，〔括弧閉〕に何らかの短単位が続いていく s 要素は，前の s 要素の末尾に〔括弧閉〕のみを移動	220	35,672
5. 読点で始まっている場合は，前の s 要素の末尾に読点のみを移動	5	979
6. “発話者「～」” や “発話者：「～」” で表記される場合は，一つの s 要素とする	66	2,062
合計	676	82,244
処理 M(β)：(3.2.3 節) 修正率の高いパターン・認定基準		
文境界を認定して分割する場合（特に Web 関連データ）		
1. s 要素の中に顔文字を含み，且つ，、その顔文字が文末表示だと考えられる場合	35	未集計
2. s 要素の中に（照）（涙）等の（X）を含み，且つ，その（X）が文末表示だと考えられる場合	194	未集計
3.【特殊事例】空白で文が区切られる場合等	1,397	未集計
合計	1,626	未集計
文境界認定を打ち消して文を結合する場合（特に雑誌・Web 関連データ）		
1.「？」「！」等に係り受け関係が結べる要素が後続し，s 要素内に含めるべきと判断される文末記号	1,547	未集計
2. 補足を表す丸括弧（括弧内に句点類Bを含まないものに限定）	77	未集計
3. 原本レイアウト情報を反映した結果，係り受け関係を結べる要素が二つの s 要素に分割されていて，括弧内に文末記号が含まれていない場合	902	未集計
4.【特殊事例】〔括弧閉〕に丸括弧で注釈が後続する場合は変更しない	400	未集計
合計	2,926	未集計

謝辞

本研究の一部は国語研基幹型プロジェクト「コーパスの管理・運用」国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

浅原 (2013).「係り受け関係アノテーション基準の比較」　第 4 回コーパス日本語学ワークショップ予稿集, pp. 81–90.
浅原・松本 (2013).「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」　言語処理学会第 19 回年次大会発表論文集, pp. 66–69.
福岡・松本 (2005).「Support Vector Machines を用いた日本語書き言葉の文境界修正」　言語処理学会第 11 回年次大会発表論文集, pp. 1221–1224.
国立国語研究所 (2011).『『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版』.
小西・小山田・浅原・柏野・前川 (2013).「BCCWJ 係り受け関係アノテーション付与のための文境界再認定」　第 4 回コーパス日本語学ワークショップ予稿集, pp. 135–142.
丸山・高梨・内元 (2006).「第 5 章 節単位情報」　国立国語研究所報告書 124 日本語話し言葉コーパスの構築法, pp. 255–322.
南 (1974).『現代日本語の構造』　大修館書店.
西光・秋田・高梨・尾嶋・河原 (2009).「局所的な係り受けの情報を用いた話し言葉の節・文境界の推定」　情報処理学会論文誌, 50:2, pp. 544–552.
下岡・南條・河原 (2004).「講演の書き起こしに対する統計的手法を用いた文体の整形」　自然言語処理, 11:2, pp. 67–83.
下岡・内元・河原・井佐原 (2005).「日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化」　自然言語処理, 12:3, pp. 3–18.
田島・難波・奥村 (2003).「形態素解析器を利用した講演書き起こしの文境界検出について」　情報科学技術フォーラム (FIT 2003), pp. 155–156.
山口・高田・北村・間淵・大島・小林・西部 (2011).『特定領域研究「日本語コーパス」平成 22 年度研究成果報告書『現代日本語書き言葉均衡コーパス』における電子化フォーマット Ver.2.2』, JC-D-10-24.

自発モノローグにおける息継ぎ音自動アノテーションの試み

浅井拓也 (早稲田大学), 菊池英明 (早稲田大学), 前川喜久雄 (国立国語研究所)

Automatic Annotation of Breathing Noise in Spontaneous Monologue: A Preliminary Report

Asai Takuya(Waseda university), Hideaki Kikuchi(Waseda university), Kikuo Maekawa(NINJAL)

1. はじめに

本研究はイントネーションや談話研究への応用を念頭におき，『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese，CSJ) 前川 (2004), の音声信号中に含まれる息継ぎ音（話者の息継ぎによって生じる雑音) ラベルを付与することを目標にしている．本稿では第一報として学会講演，模擬講演から合計８名分の音声を選択して分析し，息継ぎ音自動検出器を作成した結果を報告する．

従来，話し言葉を対象にした息継ぎ音の自動検出では，Price et al. (1989) や Wightman and Ostendorf (1994) などがある．Price et al. (1989) はプロのアナウンサーが発話した音声に対して，ケプストラムを特徴量としたガウス混合分布モデル (Gaussian mixture model，GMM) ベースの検出器を使い，93% の識別率を得た．Wightman and Ostendorf (1994) は同じくケプストラムを使用したベイズ検出器を使い，オープンテストで 73% の識別率を得た．また，歌唱における息継ぎ音検出として，中野ほか (2008) は，息継ぎの自動検出に有効な音響的特徴量を解析し，検出器を作成している．この研究では，MFCC(Mel-frequency cepstral coefficients)，ΔMFCC，ΔΔMFCC，Δ パワーを特徴量とし，隠れマルコフモデル (Hidden Markov Model，HMM) を使用して，再現率 97.5%，精度 77.7% を得た．

しかし，これらの結果はいずれも，教示によって統制された音声を対象としており，また，自発音声における発話者の発話特性の変化や個人差の影響を考慮していない．そこで本研究では，自発音声の息継ぎ音に関しても上記のような音響的特徴量が息継ぎ音の検出に有効であるのかを検討し，発話者の個人性に合わせた検出器を作成することで，発話者の性質が息継ぎの自動検出の精度に与える影響を観察する．

2. 自動検出の目標

まずは，本解析における息継ぎ音自動検出器の最終的な出力について定義する．検出器の出力結果のイメージを図 1 で説明する．図 1 は今回解析の対象とした音声ファイルの一部を praat を使用して可視化したものである．最上部から音声波形，スペクトログラム，CSJ にもともと付与されている情報の一つである word 層，そして今回の自動検出結果のイメージを表示している．

息継ぎ音の検出において問題になるのは，例えば母音や摩擦音などの音響的特徴が似ている音素との分離であるが，本解析では，これを CSJ にもともと付与されている情報を利用し，分離を行った．具体的には，非発話区間 (図 1 の word 層中の#で示されている部分) のみを解析の対象としている．そのため，今回の解析では，息継ぎ音の継続時間長や，発話区間における息継ぎ音の検出は行わない．

この非発話区間に，息継ぎ音が含まれている場合には 1 を，含まれていない場合には 0 を出力すること (図 1 の最下層) が今回の検出器の目標である．

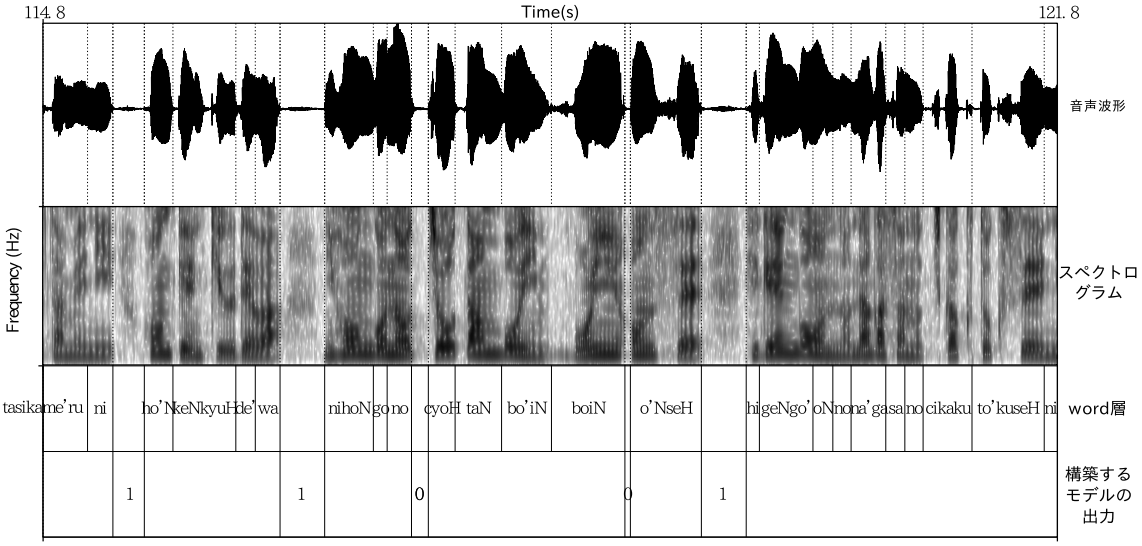


図 1　検出器の出力結果イメージ

3. データ

本章では，今回の解析で使用したデータについて説明する．今回の解析では，CSJ コーパスのコアデータから男女 4 名ずつ，合計 8 データを使用した．

息継ぎ音の特性に話速の影響がある可能性を考慮して，CSJ コーパスのコアデータの中から，特に話速の速いものと遅いものを男女それぞれ 2 名分ずつ用意した．なお，話速に関しては平均発話速度 (Mora/秒) により決定した．

これらの解析ファイルに関して，まず試験的に解析者 2 名 (エキスパート 1 名，学生 1 名) による息継ぎ音の手動アノテーションを行った．これは，そもそも自発音声における息継ぎ音は録音された音声から一貫性を持ってアノテーションを行うことが可能なのかを調査することが目的である．アノテーション範囲は上記 8 ファイルの冒頭 70sec と終端 60sec であり，非発話区間単位に息継ぎ音が含まれているか否かを評価した．

解析者の回答の一致率は κ 係数で算出した，その結果，解析者間の κ 係数は 0.92 となり，人間は息継ぎ音の検出を一貫性を持って行うことが可能であることが示された．

そのため，解析対象のデータはファイル全区間に対し学生 1 名によって手動によるアノテーションが行われた．ここでのアノテーションは，非発話区間に含まれる息継ぎ音の開始，終了時間を切り出している．この際切りだされた音声は息継ぎ音の音響的特徴量の観察と，自動検出器の教師情報及び，評価用の正解ラベルとして使用する．

本稿では中野ほか (2008) の解析を参考に，MFCC12 次元，ΔMFCC12 次元，ΔΔMFCC12 次元とパワー，Δ パワー，ΔΔ パワーを算出した．これら音響的特徴量の算出には HTK に付属している Hcopy を使用した．この際，フレームシフト幅は 10msec，フレーム幅は 25msec に設定した．対象の非発話区間の解析フレーム数は総数 171075 個であり，そのうち息継ぎ音と認定された解析フレームは 56219 個であった．

表 1 に今回の解析で使用したデータの基本情報 (発話者 ID，平均発話速度，講演時間，性別，息継ぎ音フレーム数，非息継ぎ音フレーム数) をまとめる．

表 1　解析対象データの基本情報

ID	平均発話速度	講演時間	性別	息継ぎ音フレーム数	非息継ぎ音フレーム数
A01M0015	11.57	716.24	男性	5855	10537
A03M0059	11.48	890.57	男性	6904	11746
A02M0076	8.27	1474.08	男性	7227	8494
A02M0098	7.10	1551.37	男性	10330	32441
A01F0067	10.43	942.75	女性	5996	18860
A03F0153	10.00	991.77	女性	7319	8488
A01F0122	8.17	828.17	女性	5548	15473
A06F0028	7.85	959.69	女性	7040	8817

3.1　音響的特徴量の観察

上記データに関して，音響的特徴量の観察を行った．この観察では，CSJ コアデータの音響的特徴量を，静的な特徴量 (MFCC)，動的な特徴量 (ΔMFCC，ΔΔMFCC)，パワー (パワー，Δ パワー，ΔΔ パワー) の三軸で観察する．ここでの目的は，これらの音響的特徴量のうち，息継ぎ音の自動検出器に効果的な特徴量を選択することである．

MFCC12 次元のペアプロットを図 2 に示す．この図では，それぞれ 1 次元目から 12 次元目までを上下左右に配置している．ペアプロットの下半分には，それぞれの特徴量の組み合わせの散布図を，上半分にはそれぞれの特徴量の確率密度を示す．なお，確率密度の推定にはカーネル密度推定を行い，それぞれの等高線はこの値を 10 分割して示している．最下にはそれぞれの特徴量のヒストグラムを，最右には，それぞれの次元の箱ひげ図を示した．図中，濃く示しているのが息継ぎ音，薄く示しているのが非息継ぎ音の解析フレームである．散布図や確率密度を確認すると，特に低次成分において，息継ぎ音，非息継ぎ音の分離が行えることが確認できる．これは，非発話区間における息継ぎ音には，ある種の母音的な特徴があることが原因であると思われる (図 1 を参照，息継ぎ音と認定したフレームにおいてはフォルマント成分が確認できる).

続いて ΔMFCC の観察を行う．ΔMFCC12 次元のペアプロットを図 3 に示す．分布を確認すると，ΔMFCC12 次元においては，息継ぎ音，非息継ぎ音の分布の中心点がほぼ同じ位置にあり，分離が行える次元が存在しないことが分かる．

図 4 に ΔΔMFCC12 次元のペアプロットを示す．ΔΔMFCC の分布も ΔMFCC と同様，息継ぎ音，非息継ぎ音の弁別に有効ではない可能性が強い分布となった．

これら 2 つの図の結果を考慮すると，CSJ の非発話区間から，息継ぎ音の自動検出を行う場合は，動的な特徴量よりも静的な特徴量を使用したほうが有効であることが判明した．ただし，この結果に関しては，解析フレームの幅の問題が存在する．特に動的な特徴量に関しては，解析幅の違いにより結果が異なる可能性がある．そのため，息継ぎ音の分離に適切な解析フレーム幅の検討は今後の課題である．

最後にパワーに関するペアプロットを図 5 に示す．パワーに関する分布を観察すると，特にパワーと ΔΔ パワーの組み合わせにおいて，非息継ぎ音の分離がある程度行えることが判明した．だたし，この結果に関しても，特に Δ パワー，ΔΔ パワーに関しては動的な特徴量の解析と同様，フレーム幅の問題が存在する．

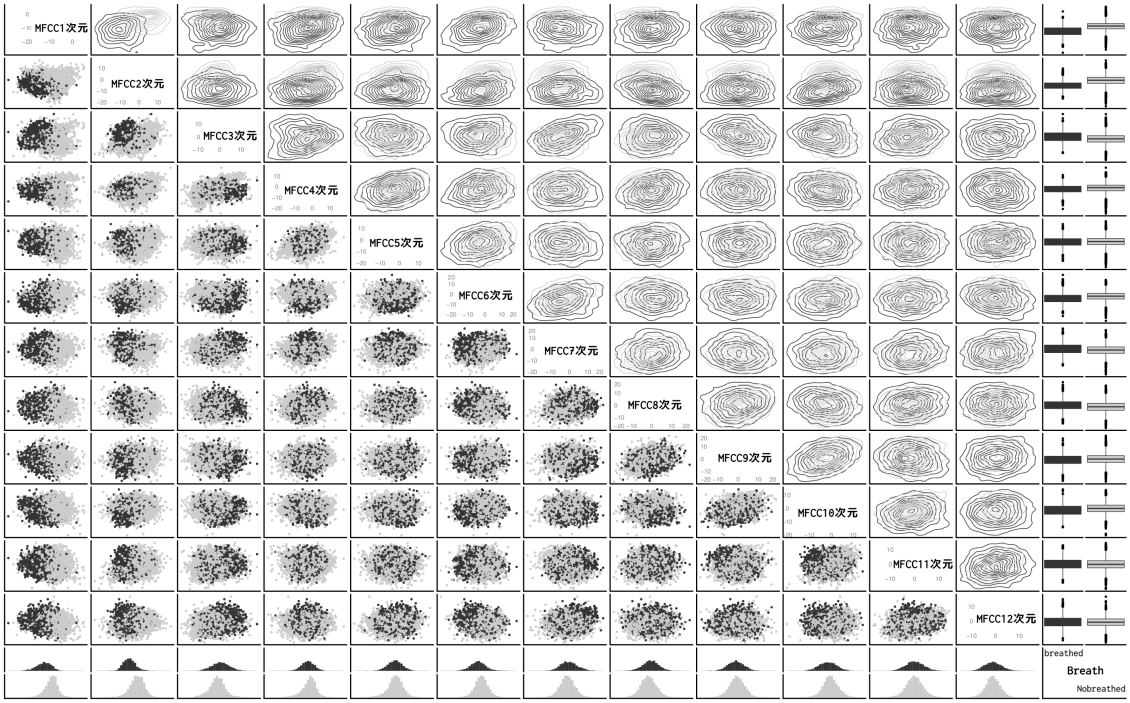


図 2　MFCC12 次元のペアプロット

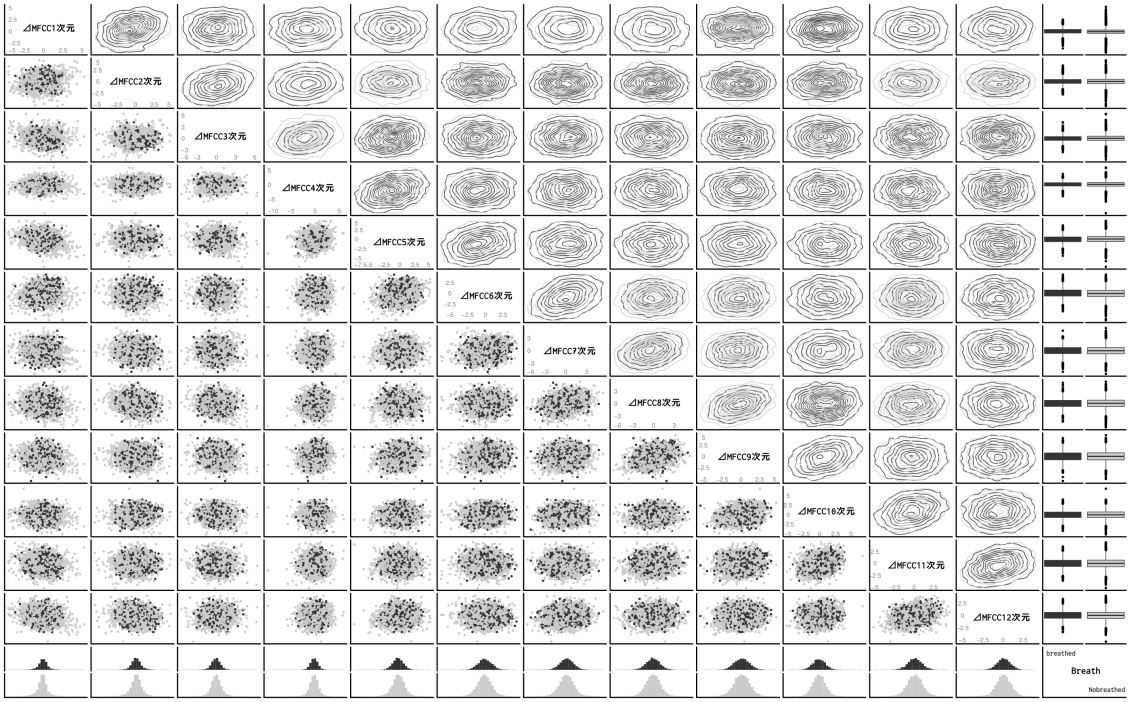


図 3　ΔMFCC12 次元のペアプロット

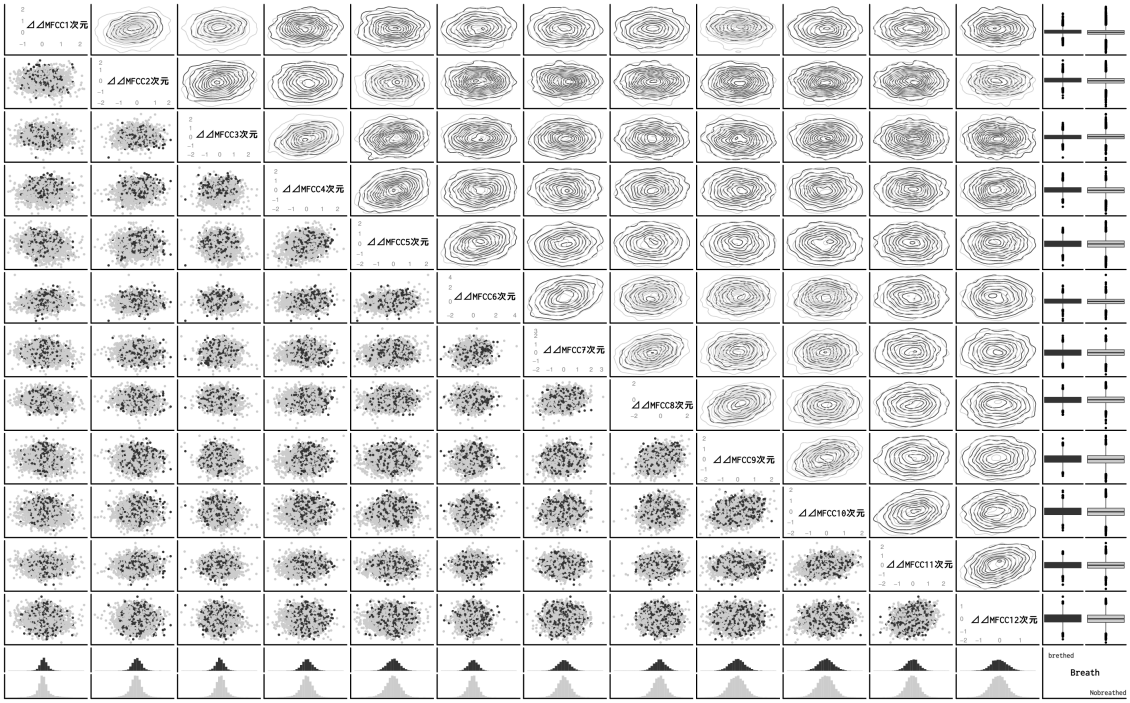


図 4　ΔΔMFCC12 次元のペアプロット

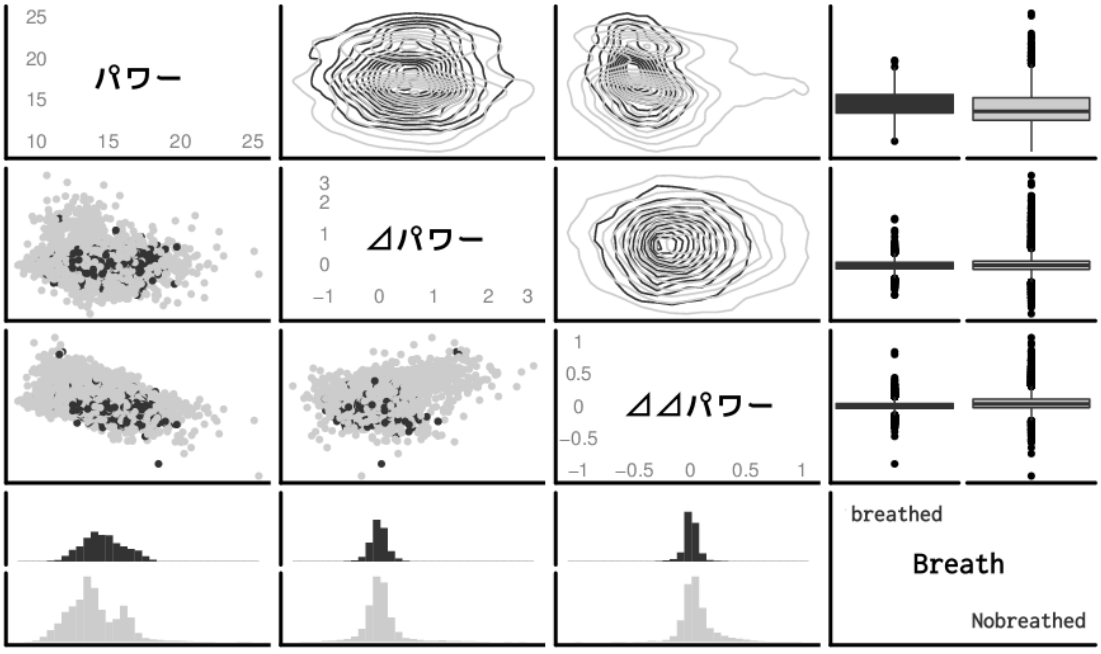


図 5　パワーのペアプロット

4. モデルの作成と評価

ここでは，上記の特徴量を利用して息継ぎ音の自動検出器の作成を行う．それぞれの特徴量を観察すると，どこか特定の次元のみで，息継ぎ音の検出を行えるような特徴量は存在せず，多次元の特徴量を使用する必要がある．また，今回の問題は，ある非発話区間に息継ぎが存在するか否かの二値の値を出力することである．このような場合において，最も効果的であるとされている分類器としてサポートベクターマシン (Support vector machine，SVM) がある．

今回の解析では，上記特徴量の観察の際に使用した解析フレームを 1 つの単位とし，それぞれのフレームが息継ぎ音であるか否かを SVM に分類させている．なお，モデルの学習時には計算コストを削減するために，すべてのデータの中から 6000 フレーム分のデータをランダムサンプリングして使用している．また，SVM のアルゴリズムは統計解析用プログラム言語 R の e1071 library を利用し，ラジアル基底関数 (RBF, radial basis function) カーネルを使用し，パラメタ C と γ はグリッド検索で最も精度の高くなる値に決定した．

以下の解析では，それぞれの条件別に，MFCC，ΔMFCC，ΔΔMFCC，それぞれのパワーを特徴量に，SVM による教師あり学習を行い，その精度の比較を行う．モデルの評価には recall，precision，F 値，accuracy，κ 係数を使用した．モデル評価時には，解析に使用したデータのうちの $\frac{2}{3}$ のデータを学習用に，$\frac{1}{3}$ をテスト用にし，3-fold cross validation 法を行った．

4.1 フルモデル vs MFCC モデル

CSJ 非発話区間における息継ぎ音の音響的特徴量を観察すると，MFCC12 次元が最も自動検出の結果に影響を与えやすいデータである．そこで，まずは，この値のみを使用しモデルの学習を行った (MFCC モデル)．モデルの学習結果を表 2 左に示す．この学習の κ 係数は 3 回の試行平均 0.63 となった．

この検討と同様に，ΔMFCC，ΔΔMFCC に関しても，それ単独に学習を行ったが，モデルの学習は収束しなかった．そのため，MFCC モデルにその他の特徴量を全て加えたモデル (フルモデル) を作成し，MFCC モデルと比較した．フルモデルの評価値を表 2 右に示す．フルモデルの 3 回試行の平均 κ 係数は 0.67 となった．わずかではあるが，MFCC モデルより分離精度の向上が確認された．本研究の目的は出来る限り精度のよい自動検出器の作成であるため，以降の検討においてはフルモデルを採用し，発話者の特性によるモデルの精度の変化を観察していく．

表 2　MFCC モデルとフルモデルの評価値比較

	MFCC モデル					フルモデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.67	0.70	0.70	0.69	0.02	0.74	0.73	0.77	0.75	0.02
recall	0.81	0.78	0.79	0.79	0.02	0.82	0.77	0.81	0.80	0.03
F	0.74	0.74	0.74	0.74	0.00	0.78	0.75	0.79	0.77	0.02
accuracy	0.83	0.84	0.85	0.84	0.01	0.85	0.85	0.87	0.86	0.01
kappa	0.61	0.63	0.64	0.63	0.01	0.67	0.64	0.70	0.67	0.03

4.2 男女差の比較

呼吸法に関しては，一般に男女で異なると言われている．そのため，息継ぎ音の自動検出においても，男性モデルと女性モデルを作成したほうが，検出力が高まる可能性がある．そこで，上記のフルモデルを利用し，学習データを男性のみにした場合と，女性のみにした場合でのモデルの精度を比較した．

男性モデルの評価結果を表 3 左に，女性モデルの学習結果を表 3 右に示す．男性モデルの 3 回の試行平均 κ 係数は 0.60，女性モデルの平均 κ 係数は 0.76 となった．

この試行の結果，男性モデルに関してはフルモデルよりモデルの精度が低くなり，女性モデルはフルモデルよりもモデルの精度が向上することが分かった．この結果は，女性の息継ぎ音は比較的安定した音響的特徴量が存在するのに対し，男性の息継ぎ音は音響的特徴量のばらつきが大きいことを示唆する結果である．

表 3　男女別モデルの評価値比較

	男性モデル					女性モデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.68	0.65	0.65	0.66	0.02	0.80	0.80	0.85	0.82	0.03
recall	0.81	0.76	0.80	0.79	0.03	0.90	0.84	0.84	0.86	0.04
F	0.74	0.70	0.72	0.72	0.02	0.85	0.82	0.84	0.84	0.02
accuracy	0.84	0.82	0.84	0.83	0.01	0.89	0.89	0.90	0.89	0.00
kappa	0.62	0.57	0.61	0.60	0.03	0.77	0.74	0.77	0.76	0.01

4.3　個人差の比較

男女差の検討に関しては，これが本当に男女差によるものなのか，それとも個人差によるものなのかを検討する必要がある．そのため，解析対象のデータに関して，各発話者一人ひとりを評価データとし，学習データはその発話者以外のデータを使用した検討を行った．この検討で，ある発話者をテストデータとした場合の評価値が低い例が見つかれば，その発話者の特殊性が判断できる．

この検討の結果を表 4 に示す．この表においては各条件の κ 係数のみを示している．Unweighted は通常の κ 係数，Weighted は重み付け κ の値である．

表 4 を確認すると男性ファイル (ID 名に M が含まれるもの) の値は女性ファイル (F が含まれるもの) に比べ評価値の値が低いことが分かる．そのため，男女差の検討での結果は個人差というよりも性別の影響を受けた結果であることが判明した．ただし，男性ファイルでは個人差間のモデルの評価値のばらつきも大きい．特に，最も精度の低かった M0098 は，発話者が高齢であり，息継ぎ音の音響的特徴量が他の発話者と性質が異なるものである可能性が強い．

表 4　各発話者別 κ 係数

	F0067	F0122	M0015	M0076	M0098	F0153	M0059	F0028
Unweighted	0.67	0.75	0.51	0.46	0.36	0.80	0.63	0.53
Weighted	0.67	0.75	0.51	0.46	0.36	0.80	0.63	0.53

4.4　話速の差の比較

最後に，話速の差に関する検討を行う．本稿では解析データの選別に関して，ファイル単位での話速の速いものと遅いものを選択的に選んでいる (表 1 参照)．そこで得られた話速別に，話速の速い発話者，遅い発話者別のモデルを作成した．話速の速い発話者モデルを表 5 左に，遅い発話者モデルを表 5 右に示す．これらのモデルにおいての κ 係数はそれぞれ，0.75，0.59 となった．そのため，この結果のみを比較すると話速の影響によりモデルの検出結果が変化するように見える．

しかし，4.2 節での検討の結果もあり，この結果が話速によるものなのか，男女差によるものなのかの検討が必要である．そこで，話速の遅い群に関しては更に，男女別にモデルを作成し，評価を行った．

表 5　話速別モデルの評価値

	話速の速い発話者モデル					話速の遅い発話者モデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.79	0.81	0.85	0.82	0.03	0.72	0.67	0.68	0.69	0.03
recall	0.89	0.82	0.84	0.85	0.03	0.77	0.72	0.73	0.74	0.03
F	0.84	0.82	0.85	0.83	0.02	0.74	0.70	0.70	0.71	0.03
accuracy	0.88	0.88	0.90	0.89	0.01	0.84	0.83	0.82	0.83	0.01
kappa	0.75	0.73	0.78	0.75	0.03	0.62	0.57	0.58	0.59	0.03

低話速度男性モデルの評価値を表 6 左に，低話速度女性モデルの評価値を表 6 右に示す．それぞれの三回の試行平均 κ 係数は 0.53，0.74 となった．つまり，同じ低話速度条件においても女性モデルの評価値は高いままである．その結果，ここでの検討結果は話速による影響ではなく，男女差の影響を反映したものであることが判明した．

表 6　低話速度男女別モデルの評価値比較

	男性モデル					女性モデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.68	0.61	0.59	0.63	0.05	0.79	0.83	0.84	0.82	0.03
recall	0.72	0.69	0.69	0.70	0.02	0.92	0.81	0.79	0.84	0.07
F	0.70	0.65	0.64	0.66	0.03	0.85	0.82	0.82	0.83	0.02
accuracy	0.83	0.80	0.80	0.81	0.02	0.89	0.89	0.87	0.88	0.01
kappa	0.58	0.51	0.50	0.53	0.05	0.76	0.74	0.72	0.74	0.02

5. まとめ

本稿では CSJ 講演音声の息継ぎ音自動検出器作成に向けて，息継ぎ音の音響的特徴量の観察を行った．その結果，CSJ コアデータの息継ぎ音自動検出には静的な特徴量である MFCC を使用するのか最も効率がよいことが分かった．続いて，SVM を使用した自動検出器を作成し，3-fold cross validation 法で評価を行った．その結果，今回提案する検出器の精度は κ 係数で 0．6 以上という大変よい精度を得た．また，発話者の性質別にモデルの学習を行い，発話者の性質とモデルの精度の関係を観察した．その結果，作成した息継ぎ音の自動検出器の精度には男女差が存在することが分かった．話速に注目した解析も行ったが，女性の発話においては検出器の精度に大きな差はなく，男性の発話においてのみ話速の差が確認された．

これらの解析は息継ぎの音響的特徴量による自動検出という課題においては，男女差を考慮する必要があること，特に男性の発話においては，話速の影響を考慮する必要のあることを示す結果である．

6. 課題

最後に今後の課題について述べる．

まず，本稿での報告の段階では冒頭で上げたモデルの出力 (各非発話区間毎の二値データ) の形式ではなく，音響解析フレーム毎に息継ぎ音か非息継ぎ音かを検出している．これを当初の目的に合わせ，非発話区間毎に変換をする必要がある．これに関しては今回の提案モデルは音響フレーム毎の評価では非

常に良い成績を出しているため，例えば，ある非発話区間に含まれる音響解析フレームの値の多い方をその区間の正解とすることで, より高い精度での検出が可能になると考えている．

また，今回の解析の結果，特徴量に MFCC12 次元を使用し，SVM による検出器を作成することで，マイクに収録されている息継ぎ音の自動検出はある程度よい精度を得られることが判明したが，最終的にこの結果をイントネーションや談話分析に応用していくことを考えると，マイクに収録されることのない，小さな息継ぎや発話区間中に発生する息継ぎ音に関しても抽出しきれることが望ましい．

実際にマイクに収録されることのなかった息継ぎ音の存在を示唆するために図 6 を示す．この図は今回解析を行った各発話者ごとの，ある息継ぎ音の発生から，次の息継ぎ音の発生までの時間をヒストグラムにしたものである．図中横軸はある息継ぎ音から次の息継ぎ音までの持続時間，縦軸はその出現回数である．

これらの分布を確認すると特に A02M0098 と A06F0028 において分布の二峰性が確認される．その他の分布に関しても分布の混合が確認される．

この図で示しているものは，それぞれの発話者の息を吐き続けている持続時間である．そして，直感に従うのなら，人間が息を吐き続けることができる時間には限界があり，おそらくは正規分布を仮定してもよいはずである．しかし，図 6 はひとつの正規分布ではなく，時間的に数カテゴリある混合分布である．つまり，この混合数分だけ，マイクに収録されることのない息継ぎが存在している可能性がある．

今後の試行として，実際に行われた息継ぎと，マイクに録音された息継ぎ音の差がどの程度のものであるのかについて厳密な実験を行う予定である．

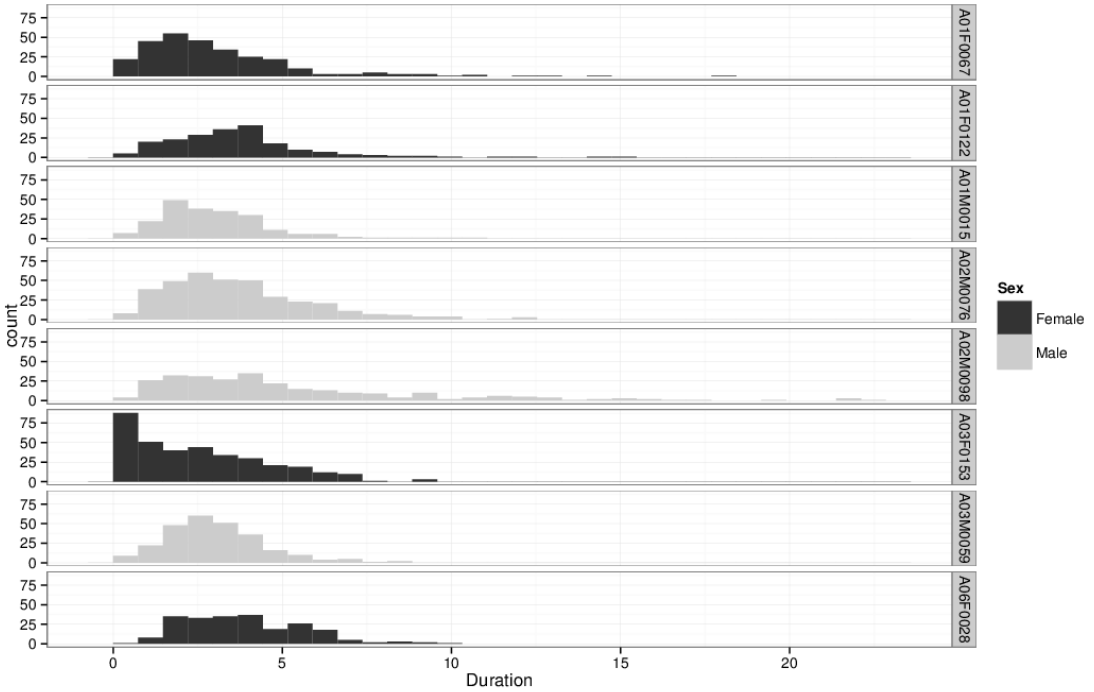


図 6　各発話者の息継ぎ音までの持続時間

参考文献

Price, P. J., M. Ostendorf, and C. W. Wightman (1989). “Prosody and parsing.” *Proc. Work. Speech Nat. Lang. - HLT ’89*, p. 5 Morristown, NJ, USA: Association for Computational Linguistics.

Wightman, C.W., and M. Ostendorf (1994). “Automatic labeling of prosodic patterns.” *IEEE Trans. Speech Audio Process.*, 2:4, pp. 469–481.

中野倫靖・後藤 真孝・緒方 淳・平賀 譲 (2008).「無伴奏歌唱におけるブレスの音響特性とそれに基づく自動ブレス検出 (音響分析一般 (1))」　情報処理学会研究報告. [音楽情報科学], 2008:78, pp. 83–88.

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.

形容詞用例データベースの構築

山口昌也 (国立国語研究所言語資源研究系) †

Building a Database of Japanese Adjective Examples

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

1　はじめに

本稿では，日本語形容詞の用例データベースの構築方法と構築結果を報告する。本用例データベース (以後，用例 DB) の構築目的は，用法を指定した形容詞の分析である。対象とする用法は連体用法，叙述的用法，副詞的用法とし，語，および，用法によらず，分析に十分な用例が収集できるような用例 DB の構築を目指す。

以上のような用例 DB を構築する場合に問題となる事柄として，生起確率の低い語・用法の用例収集が挙げられる。特に，単一のコーパスから用例収集すると，収集可能量はコーパスの規模に依存してしまい，生起確率の低い語や用法は十分な量の用例を収集できない恐れがある。実際，約 11 億語を収録する『筑波ウェブコーパス』(今井ら 2013) を NINJAL-LWP for TWC[1] を用いて，「美しい」「すばしこい」を検索すると，用例数は次のようになる [2]。

- 「美しい」(総数：97540 例,「美しい」＋名詞：46855 例,「美しく」＋動詞：11642 例)
- 「すばしこい」(総数：85 例,「すばしこい」＋名詞：22 例,「すばしこく」＋動詞：23 例)

このように，生起確率が高い「美しい」は容易に多量の用例を収集できるが，生起確率の低い「すばしこい」では，十分な量の用例を収集しているとは言いがたい。また，2 節で示すように，語形の違いでも生起確率の差異がある。

そこで，本稿では，単一のコーパスから用例を収集するのではなく，語，および，語形ごとに専用の Web コーパスを構築し，そこから用例を収集する。この方法の利点は，大規模なコーパスを構築する必要がないこと，また，用例の不足が明らかになった時点で，再収集するのが容易なことである。筆者は複合動詞に対しても同様の試み (山口 2013) を行い，約 3900 語の複合動詞用例 DB を構築した。このうち，約 1800 語については 1000 例以上の用例が収集されているが，基本形のみでしか用例を収集していない。本稿では，複数の語形で用例収集が可能かどうかを検証する。

この後の節では，2 節で収集対象の形容詞，および，用法を定め，3 節で用例収集のための Web コーパスの構築方法を示す。次に，4 節では Web コーパスから抽出された用例を用法別に分析し，どの程度の用例を収集できるかを明らかにする。さらに，5 節で用例 DB を Web 上で利用するためのシステムについて紹介し，6 節で本稿をまとめる。

2　用例の収集対象

ここでは，用例収集の形容詞とその用法について説明する。

まず，用例対象の形容詞は，形態素解析システム JUMAN(ver.6.0) [3] 用の辞書に登録されているイ形容詞約 500 語，ナ形容詞約 1700 語 (品詞細分類が「ナノ形容詞」の形容詞も含む) とした。なお，同一の代表表記を持つ辞書項目が複数ある場合は，検索エンジン [4] の検索総数を参考にして，一つに限定している。

---

†http://www2.ninjal.ac.jp/masaya
[1]http://corpus.tsukuba.ac.jp/
[2]2014-01-30 に検索
[3]http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[4]http://www.bing.com/ を用いた。

対象とする用法は，次の三つとした。3 節で述べるように，本研究で用いるコーパス構築手法は，収集対象の形容詞に対して単一の表記を指定しなければならないが，個々の用法は複数の語形で表現しうる。例えば，連体修飾用法であれば,「白い皿」「白かった皿」で表現できる。そこで，今回はそれぞれの語形の生起確率を調査し，生起確率の高い語形の用例をコーパス構築対象とする。

連体修飾用法： 体言を修飾する用法 (例：白い皿)
副詞的用法： 用言を修飾する用法 (例：白く光る)
叙述的用法： 述語としての用法 (例：その皿の色は白い)

語形ごとの生起確率の調査には，無作為抽出文コーパス (3.2 節参照) を用いた。このコーパスは，Web から収集したページから無作為に文を抽出・集積した Web コーパスで，約 3.8 億形態素 [5] を含む。イ形容詞，ナ形容詞の出現語形の分布 (相対頻度で上位 5 語形まで) を表 1，2 に示す。なお，活用形の名称は，JUMAN に準じる。

表 1: イ形容詞の活用形ごとの分布

活用	相対度数	語例
基本形	0.586	白い
基本連用形	0.248	白く
語幹	0.078	白
タ形	0.044	白かった
タ系連用テ形	0.002	白くて
総形態素数	3855383	

表 2: ナ形容詞の活用ごとの分布

活用	相対度数	語例
語幹	0.409	きれい
ダ列基本連用形	0.186	きれいに
ダ列特殊連体形	0.175	最短の
ダ列基本連体形	0.136	きれいな
デス列基本形	0.003	きれいです
総形態素数	2205650	

表 1 を見ると，イ形容詞は基本形と基本連用形の上位 2 語形で全体の 83.4%を占める。基本形は連体修飾用法と叙述用法を，基本連用形は副詞的用法を表現しうる。一方，タ系の活用形は上記の二つの活用形よりも生起確率が低い。したがって，基本形，基本連用形を用例収集の対象とする。

一方，ナ形容詞はトップの「語幹」を除けば，上位 3 語形は語尾が「ダ」の系列であり,「デス」「デアル」系列の生起確率は低いことがわかる。また,「ダ」の系列の中で見てみると，(タ系ではなく) 基本系が上位を占めている。したがって，まず「ダ列基本連用形」「ダ列基本連体形」を収集対象の活用形と決める。これらは，それぞれ副詞的用法，連体修飾用法に対応する。これに伴い，叙述的用法も，同じ「ダ」系列の基本形 (例：きれいだ，相対頻度は 0.003) の用例を収集することにする。さらに,「ダ列特殊連体形」が 3 番目に挙がっているので，生起確率を考慮し,「ダ列特殊連体形」も連体修飾用法の語形として用例収集する。

以上をまとめると，収集対象の活用形は，次のようになる。なお，活用形の名前の煩雑さを避けるため，カッコ付きの活用形は，以後，カッコ内の表記で記す。

イ形容詞： 基本形，基本連用形 (連用形)
ナ形容詞： ダ列基本連用形 (連用形)，ダ列基本連体形 (連体形 A)，基本形，ダ列特殊連体形 (連体形 B)

3　用例 DB の構築

3.1　Web コーパスの構築と用例 DB への登録

前節で述べた形容詞と活用形に対して，個別に専用の Web コーパスを構築する。本稿で用いた構築手法は，Web 検索エンジンを用いた Web コーパス構築手法 (Baroni・Bernardini(2004)，Sharoff(2006)) を基本としている。これらの手法は，Web 検索エンジンに与えるランダムキーワードを与えることにより，収集される Web ページに特定の偏りが生じないようにしているが，本研究では対象語を検

[5]形態素解析システム JUMAN ver.7.0 による

索キーに含めることにより，対象語の用例を効率的に収集できるようにしている。なお，本稿で用いた手法は，山口 (2013) とほぼ同一である。

Web コーパスの構築と用例 DB への登録手順を次に示す。この手順は，個々の形容詞の個別の活用形に対して適用される。したがって，イ形容詞は 1 語当り 2 個，ナ形容詞は 1 語当り 4 個のコーパスを構築することになる。

(1) 検索エンジン [6] に対して，収集対象の語 (語形を指定) とランダムキーワードを与えて AND 検索し，最大 50URL を取得する。ランダムキーワードは，広範囲のジャンルで多用される語 3000 語からランダムに 1 語選択した。

(2) 検索された URL からランダムに 10URL 選択し，ダウンロードする。ダウンロードしたファイルは，文字コードを UTF-8 に統一した上で，各種クリーニング (文字の正規化，HTML タグの削除など) を行う。

(3) 規定のページ数が集まるまで，1，2 を繰り返す。今回は，各用法ごとに 500〜1000 例程度を収集することを目標として，イ形容詞の基本形は 2000 ページ，イ形容詞の連用形，および，ナ形容詞の全語形はそれぞれ 1000 ページ収集した。イ形容詞の基本形だけを 2000 ページ収集したのは，イ形容詞の基本形には二つ用法 (連体修飾用法，叙述的用法) があるからである。なお，Web ページ収集時は，重複する URL は収集しない。また，1 において，5 回連続で URL を取得できない場合は，URL の収集を終了とした。

(4) 収集した Web ページを形態素解析したのちに，収集対象の語を含む「文」を用例候補として抽出する。今回は,「文」の区切りとして，句読点，空白文字を用いた。なお，同一文は重複して用例候補とはしない。

(5) 個々の用例候補に対して，構文解析，および，格解析を行う。構文解析の結果，収録対象の語を含むと確認された場合，出典 URL とともに用例 DB に登録する。さらに，格解析結果を基に，対象語に対する共起情報 (主として，格要素の情報) を付与する。形態素解析に JUMAN(ver.7.0)，構文解析・格解析には KNP (ver.4.0)[7] を用いた。

3.2 無作為抽出文コーパスの構築

無作為抽出文コーパスは，Web から無作為に抽出した文から構成される Web コーパスで，Web における語の出現分布を把握するために作成する。本稿では，収集対象の語形の決定 (2 節参照) と構築した用例 DB の評価のために用いる。

構築した無作為抽出文コーパスの規模は約 3.8 億語，文数は約 1725 万文，出典の URL は約 402 万である。構築方法は，基本的に前節の方法と同一だが，次の点が異なる。

- 検索エンジンに与えるキーワードは，ランダムキーワード 3 語である。また，特定目的のコーパスを構築するわけではないので，対象語は指定しない。
- 合計 550 万の URL を収集し，それぞれの Web ページから，最大 5 文を無作為に抽出した。重複文の重複は許さない。
- 抽出する文は 15 形態素以上を含むものとした。これは，データやナビゲーションのリンクなどの述語を持たない「文」を排除するためである。

4 用例 DB の構築結果

4.1 収集した用例数 (語形別)

用例 DB の構築結果として，収集量を評価する。まず，語形ごとに構築した Web コーパスから収集される用例の量を調べてみる。イ形容詞，ナ形容詞の結果をそれぞれ図 1，2 に示す。なお，凡例

[6] 今回は，検索エンジンとして，Bing API(http://datamarket.azure.com/dataset/bing/searchweb) を利用した。
[7] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

部分のカッコ内の数値は，用例数の中央値である。また，これらの結果集計では用法による区別をしておらず，構築した Web コーパスに含まれるすべての当該語の用例を集計している。したがって，例えば，イ形容詞の基本形用に構築した Web コーパス中に連用形の用例が含まれていれば，それらも合算される。

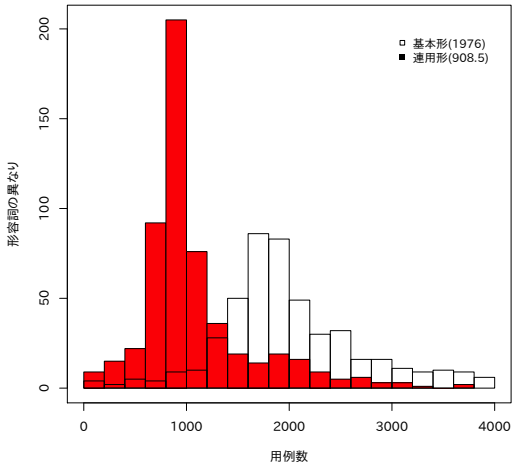


図 1: イ形容詞の用例数

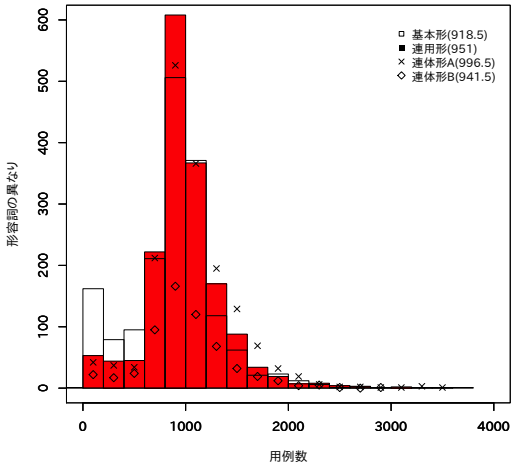


図 2: ナ形容詞の用例数

まず，イ形容詞・ナ形容詞について，用例数の中央値を見てみると，収集 Web ページ数の周辺値になっている。相対頻度の差が大きい，ナ形容詞の基本形と連用形 (それぞれ 0.003，0.186) でも，用例数の中央値の差は 32.5 であり，今回の Web ページ収集量では，生起確率による収集用例数量へ顕著な影響は現れていないと考えられる。

次に，個別の分布について見てみると，ナ形容詞の基本形の分布で，用例数の少ない形容詞 (図 2 の用例数 0 付近) が見受けられる。この原因の一つは，一般的に基本形では用いられない，「けげんだ」「並大抵だ」「むやみだ」などの存在が挙げられる。

4.2　収集した用例数 (用法別)

前節では語形ごとに，構築した Web コーパスから収集された用例数を検証した。ここでは，収集された用例の中に，目的の用法の用例がどの程度含まれているかを検証する。検証には，用例の構文解析結果を利用した。具体的には，対象の形容詞が用言に依存している場合は副詞的用法，体言に依存している場合は連体修飾用法，文末の述語となっている場合は叙述的用法に分類した。なお，この過程で，用例として適切でないもの，例えば，連用形なのにも関わらず依存先が存在しない場合 [8] や，叙述的用法なのにも関わらず格要素がない場合 (例：「寒いですね」) などは集計から外した。

まず，図 3 にイ形容詞の基本形に関連する用法の結果を示す。すでに述べたように，イ形容詞の基本形は叙述的用法と連体修飾用法で用いられうる。この二つの用法のうち，連体修飾用法の用例の割合が多く，用例数の中央値は 1287 であった。一方，叙述的用法の用例数の中央値は 241 で，連体修飾用法の用例数はその約 5.3 倍である。このような分布の差は，Web データにおける生起確率の違いが影響していると考えられる。実際に，無作為抽出文コーパスから同様の方法で二つの用法の用例を抽出すると，連体修飾用法の用例数 (中央値) は叙述的用法の約 6.1 倍であった。今回は実施しなかったが，叙述的用法の用例が不足する場合は，今回の収集量を参考にし，収集用の Web コーパスの規模を拡大する。

次に，イ形容詞の連用形と対応する，副詞的用法の結果を図 4 に示す。図 2 の結果と比較すると，中央値で 264 例減少している。これは，構築した Web コーパスの中に，他の用法も含まれていること

[8] 例：読みやすさを考慮し，文の途中で改行されている場合，文全体の抽出に失敗する

が一つの要因である。今回のデータでは，連用形の語形を指定して作った Web コーパスには，叙述的用法，連体修飾用法がそれぞれ 47，45 例 (中央値) 含まれていた。

最後に，ナ形容詞の用法別の用例数を見てみる。図 5 に連体修飾用法 (連体形 A，B の 2 種類)，図 5 に叙述的用法・副詞的用法の結果を示す。用例数の中央値は叙述的用法を除き，700 例であった。叙述的用法の用例数が少ないのは，4.1 節でも述べたように一般的に基本形では用いられない語があることと，ナ形容詞のダ系列の基本形は生起確率 0.003 と 4 種類の語形の中では最も低いことが原因だと考えられる。

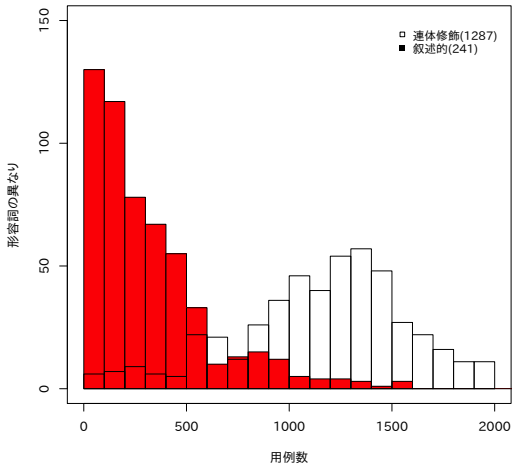


図 3: イ形容詞の連体修飾・叙述的用法の用例数

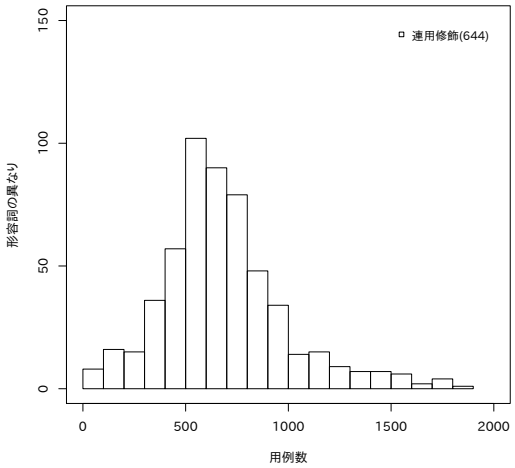


図 4: イ形容詞の連用修飾用法の用例数

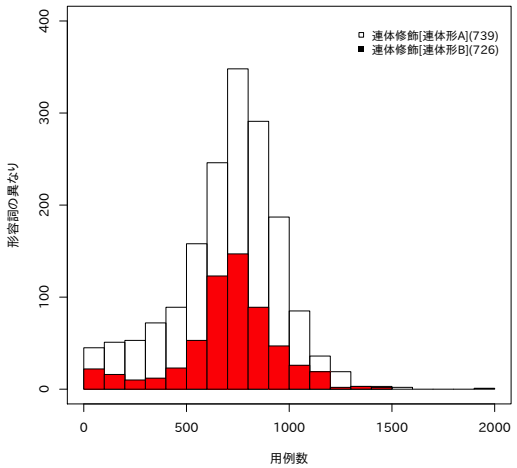


図 5: ナ形容詞の連体修飾用法の用例数

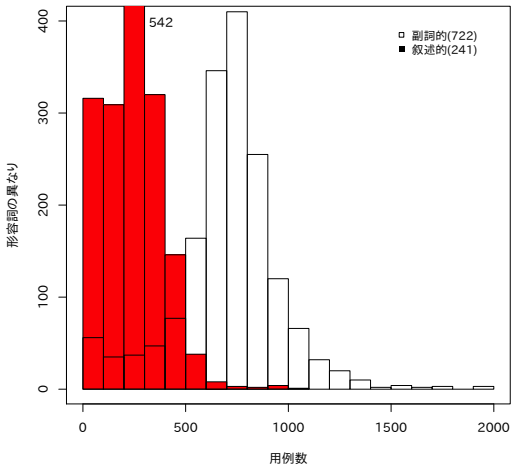


図 6: ナ形容詞の副詞的・叙述的用法の用例数

5　用例 DB の利用

5.1　想定する利用方法

用例 DB を手軽に利用できるようにするために，(a) 形容詞の検索，(b) 共起語一覧，(c) 用例一覧，の機能を実現し，Web 上で公開した [9]。想定する利用方法としては，対象の形容詞を (a) で検索し，(b) で共起語一覧を表示したのちに，選択した共起語を含む用例を (c) で閲覧するという流れである。次の節では，これらの機能のうち，(b) 共起語一覧機能の詳細について説明する。

[9] http://csd.ninjal.ac.jp/adj/

5.2　共起語一覧機能

共起語一覧機能は，対象語に依存する要素，および，対象語が依存する要素を使用頻度順に表示する機能である。共起語一覧の例として,「楽しい」の例を表 3，4 に示す (一部)。表 3 は連体修飾用法，表 4 は副詞的用法の結果である。それぞれ独立のコーパス，つまり，基本形，連用形用に構築したコーパスから集計した結果を表示している。

これらの結果において，共起語は対象語との関係ごとにまとめられる。表中の「ガ (連体)」は形容詞の連体修飾要素,「pred」は形容詞の依存先の述語を表す。共起語に併記されているカッコ内の数値は出現ページ数 (異なり) である。共起語は出現ページで降順にソートされる。それぞれの共起語をクリックすると，当該の共起語の用例が一覧表示される。

すでに述べたように，共起語の範囲は基本的に対象語に依存する要素，および，対象語が依存する要素である。ただし，副詞的用法については，依存先の用言の共起語と密接に関連している (橋本 1992) ので，共起語の範囲を対象語が依存する用言の共起語まで拡張している。例えば,「彼は楽しく数学を学んだ」であれば,「楽しく」が依存する「学んだ」だけでなく,「学んだ」と共起する「彼」「数学」も共起語とする。このように範囲拡張した場合の関係名には,「ガ̂」のように「̂」が冒頭につく。図 4 に示した「ガ̂」がその具体例である。この場合,「ガ̂」は「楽しく」が依存する用言のガ格の格要素である。

表 3:「楽しい」の共起語一覧 (連体修飾用法)

ガ (連体)	修飾	ガ
こと (45)	とても (17)	の (10)
時間 (35)	本当に (9)	人生 (7)
もの (30)	とっても (6)	みんな (4)
生活 (21)	こんなに (5)	信雄 (3)
事 (18)	もっと (5)	出社 (3)

表 4:「楽しい」の共起語一覧 (副詞的用法)

pred	̂外の関係	̂ガ
なる (78)	こと (18)	の (12)
する (48)	ため (12)	毎日 (5)
学ぶ (24)	方法 (11)	こと (5)
過ごす (21)	方 (7)	ゲーム (5)
出来る (19)	よう (5)	みんな (4)

6　終わりに

本稿では，形容詞の三つの用法を対象として用例 DB を構築した。今回用いた構築手法では，語・用法によらず，500〜1000 例の用例を収集できるよう，語，および，語形ごとに用例収集用の Web コーパスを用意した。構築の結果，叙述的用法以外の用法では目標量に達する 644〜1287 例 (中央値)，叙述的用法でも 241 例 (中央値) の用例を収集することができた。今後は，用例 DB を用いた研究として，副詞的用法における依存先の用言の多様性を決定づける要因について分析を進める予定である。

参考文献

山口昌也 (2013) 複合動詞用例データベースの構築と活用，国語研プロジェクトレビュー Vol.4 No.1，国立国語研究所

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

M. Baroni and S. Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004

Serge Sharoff (2006) Open-source corpora: Using the net to fish for linguistic data.International Journal of Corpus Linguistics, Vol. 11, pp. 435462

橋本美奈子，青山文啓（1992）形容詞の三つの用法：終止，連体，連用，計量国語学第 18 巻 5 号，pp.201-214

新しい日本語辞書定義文型の策定に向けて

佐藤 理史　(名古屋大学大学院工学研究科)
夏目 和子　(名古屋大学大学院工学研究科)

Towards Full-Sentence Definitions of Japanese Words

Satoshi Sato　(Graduate School of Engineering, Nagoya University)
Kazuko Natsume　(Graduate School of Engineering, Nagoya University)

1　はじめに

我々は、語の意味や用法を調べるために、国語辞典を引く。これは、人間による、辞書の典型的な利用法である。

一方、計算機が辞書を利用する場合もある。我々のグループでは、クロスワードパズルのソルバー [1] や大学入試センター試験『国語』現代文のソルバー [2] などを作成しているが、そこでは、しばしば、ある表現の言い換えを取得したり、2 つの言語表現の意味的等価性を判定したりすることが必要となる。国語辞書を利用してこれらを実行するプログラムを書くと、通常の利用ではなかなか気づきにくかった問題が見えてくる。語の定義に関して言えば、現存する国語辞典の多くは、次のような問題を抱えている。

- 辞書の使用者として、母語話者が想定されており、語の定義を理解するために、かなりの日本語力と常識が必要となる。(そのため、計算機は、国語辞書を十分活用できない。)
- 語の定義として、言い換え語が提示されるだけの場合があり、意味や用法に関する十分な情報が得られない。
- 語の定義 (言い換え語) が、しばしば循環する。

これらの問題を一言で要約すれば、それは、「語の定義の方法が十分に統制されていない」ということに尽きる。

これに対して、英語の学習用の辞書では、20 世紀後半から、非母語話者の辞書利用を容易にするための各種の工夫が導入されてきている。語の定義の統制に関しては、次の 2 つが導入されている。

- 語を定義するために使用する語彙の統制 (LDOCE [3] など)
- 完全な文 (full sentence) による、語の定義 (COBUILD [4])

日本語辞書に対しても、同じようなことができないであろうか。語の定義に使用する語彙の統制と文型の統制は、非母語話者にとって有益であるに留まらず、計算機が辞書を知識源として利用する際にも好都合である。

このような考えに基づき、我々は、日本語の語を定義する新しい方法の検討を開始した。今回、その一環として、約 100 語に対して実際に完全な文による定義 (full-sentence definition: FSD) を作成することを試みた。本稿では、その試みについて報告する。まず 2 節で、目指す方向性を示す。3 節では、現時点の日本語 FSD 作成ガイドライン試案を示す。4 節では、今回の試行で作成した FSD の具体例を示し、最後に 5 節で現状をまとめる。

2　目指す方向性

現在の国語辞書では、「旧態」を次のように説明する。

【旧態】　古くからの状態・ありさま。「－依然」(岩波国語辞典第 7 版)
【旧態】　以前からの状態。また、昔のままの姿。「－を脱する」「－依然 (=もとのままで少しも変化や進歩がないさま) とした経営方針」(明鏡国語辞典第 2 版)
【旧態】　昔の状態。「－を存する／－依然 (=元通りで、少しも変化・進展が無い)」(新明解国語辞典第 6 版)
【旧態】　以前 (から) のありさま。「－依然・－を脱する」(三省堂国語辞典第 7 版)

我々は、このような記述ではなく、おおよそ次のような定義が好ましいと考える。

【旧態】　《旧態依然 {の | とした | たる}》〈何か：制度・やり方〉とは、〈何か〉が昔から {変わっていない | 進歩していない} ことを［批判的に］いう。

この記述のポイントは、次の 4 点である。

1. 主に、「旧態依然」という形式で使用されることを示す。
2. 「旧態依然」が、典型的には「制度」「やり方」などを修飾することを示す。
3. 「旧態依然」と被修飾語の間には、「の」「とした」「たる」が使用できることを示す。
4. 「旧態依然」が「批判的」な意味を含意することを示す。

このように、定義の記述形式を工夫すれば、その語がどのように使用されるかについて、より豊富な情報を提供することができる。我々が設計したいのは、まさに、このような記述形式である。

3　日本語 FSD 作成ガイドライン試案

我々は、COBUILD の full-sentence difition (FSD) に倣い、語を完全な文で定義する方針を採用する。以下に、現時点でのガイドラインを示す。

3.1　対象とする語

当面、基本語以外の内容語を対象する。対象とする語の品詞は、名詞、動詞、形容詞 (イ形容詞とナ形容詞)、副詞の 4 種類とする。

3.2　見出し語

(活用する) 見出し語の語形には、伝統的な形式、すなわち、動詞とイ形容詞は終止形、ナ形容詞は語幹を採用する。

3.3　ターゲット

一つの見出し語の下に、いくつかのターゲットを設定する。ターゲットとは、定義対象とする表現形式・用法・語義のことである。次のようなものをターゲットとする。

1. 見出し語 (それ自身)　　例：【語源】　《語源》
2. 見出し語の活用形・派生形　　例：【巧妙】　《巧妙な》
3. 見出し語を用いた典型的な表現　例：【好評】　《好評を博する》

なお、本稿では、見出し語は【巧妙】のように【】で囲み、ターゲットは《巧妙な》のように《》で囲んで示す。

3.4 FSD の構成

FSD (full-sentence definition) を、前件部と後件部から構成する。以下に例を示す。

【合議】　(見出し語)
〈人々〉が〈あること〉について《合議する》とは、　(前件部)
〈あること〉について〈人々〉が互いに相談することをいう。　(後件部)

この例に示すように、「とは、」までが前件部、それ以降が後件部である。

3.5 前件部

前件部では、ターゲットの文型を示す。文型の示し方は、ターゲットの用法に依存する。以下では、主要な 4 用法 (述語用法、連体修飾用法、連用修飾用法、名詞用法) について、典型的な方法を示す。

3.5.1 述語用法の前件部

ターゲットが述語用法の場合は、その述語の補足語 (格要素と格助詞) を示すことで、その述語の文型を示す。以下に例を示す。

【心掛ける】　〈人〉が〈あること〉{を｜[をするよう] に} を《心掛ける》とは、
【向上】　〈何か〉が《向上する》とは、
【好評】　〈何か〉が [〈人々〉に]《好評だ》とは、
【好評】　〈何か〉が [〈人々〉に]《好評を得る｜博する》とは、

なお、格要素 (〈人〉、〈あること〉、〈何か〉など) の記述方法は、後述する。

3.5.2 連体修飾用法の前件部

ターゲットが連体修飾用法の場合は、連体修飾用法の形式と典型的な被修飾語 (体言) を示す。

【巧妙】　《巧妙な》〈方法・手段など〉とは、
【高尚】　《高尚 {な｜なる}》〈[人の] 精神・言動・作品・研究など〉とは、
【凝る】　《凝った》〈作品・料理など〉とは、

3.5.3 連用修飾用法の前件部

ターゲットが連用修飾用法の場合は、連用修飾用法の形式と典型的な被修飾語 (用言) を示す。

【巧妙】　[〈誰か〉が〈何か〉を]《巧妙に》〈行なう〉とは、
【公然と】　[〈人〉が〈あること〉を]《公然と》〈行なう・主張する〉とは、

被修飾語となる用言 (述語) を示すので、必然的に、その述語の補足語が必要となる。現時点では、暫定的にこれらの要素を記号 [] で囲んで示すことにする。この記号は、付加的説明要素を表す。

3.5.4 名詞用法

ターゲットが名詞用法の場合は、その用法に典型的な修飾語が存在するか否かに応じて、異なる前件部の形式を採用する。

その用法に典型的な修飾語が存在する場合は、その修飾語を示す。典型的な修飾語が必要かどうかは、後件部を具体的にどう記述するかに密接に関わる。以下の例では、「〈ことば〉」

や「〈スポーツなど〉」への言及なしには、後件部を記述できない。特に、「コーチ」のように背後に動詞が隠されている名詞は、その動詞の格要素をノ格の修飾語としてとることがある [5]。

【語源】　〈ことば〉の《語源》とは、
その〈ことば〉のもとになった［別の］ことばのことをいう。
【コーチ】　［〈誰か〉の］〈スポーツなど〉の《コーチ》とは、
［〈誰か〉に］〈スポーツなど〉を教える人のこと。

一方、典型的な修飾語を持たない場合は、前件部をターゲットのみで構成する。

【暦】　《暦》とは、
一年の月・日・曜日に行事、祝日などが順に書かれたものをいう。

3.6　クラスまたは典型例の表示

格要素、典型的な被修飾語、修飾語は、記号〈〉で囲んで表示する。おおよそ、次のような方法がある。

1. 抽象度が高いクラスを提示する。
   (a) クラス名を提示する。例：〈人〉、〈こと〉、〈もの〉、〈行ない〉
   (b) 疑問表現で提示する。例：〈誰か〉、〈何か〉、〈どこか〉
   (c) 不特定表現で提示する。例：〈ある人〉、〈あること〉、〈あるもの〉、〈ある行ない〉
   (d) 代名詞で提示する。例：〈あなた〉、〈それ〉
   (e) ラベルで表記する。例：〈A〉、〈B〉
   (f) 複数であることを明示する。例：〈人々〉
2. 具体的なクラスを提示する。例：〈ことば〉、〈スポーツ〉、〈作品・料理など〉
3. 典型例を提示する。例：〈良い・目立っている・重要だ〉

さらに、これらを組み合わせる、末尾に「など」をつける、などの選択肢もある。

現時点では、これらの記述をどのように統制するかは定めず、FSD の具体例が十分に蓄積された段階で、記述方法の指針を定める予定である。

3.7　後件部

後件部の記述は、(概念的なレベルにおいては) 次の 2 種類が存在する。

1. 言い換え的後件部：前件部と意味的に等価な言い換え表現 (置換可能な表現) を提示する
2. 説明的後件部：前件部の意味を説明する表現を提示する。

実際の後件部は、これらのどちらかに排他的に分類されるということではなく、この 2 つの両極端の中間のどこかに位置付けられる。

語の意味を基本語のみで記述できると仮定することは、その意味を基本語の意味から合成できると仮定することである。たとえば、動詞「激怒する」は、基本となる動詞「怒る」にその程度が激しいことを組み合わせた語と考えることができる。これを図式的に表すと次のようになる。

ターゲット (述語用法)	程度を表す連用修飾語	基本となる述語
激怒する	激しく	怒る

上記のように、意味の合成が比較的単純な場合は、多くの文脈で置換可能な言い換え表現を考えることができ、後件部は、言い換え的となる。しかし、意味合成が複雑になってくると、そのような言い換え表現を考えるのが難しくなり、後件部は説明的にならざるを得ない。

将来的には、典型的な意味合成のパターンを設定し、それぞれのパターンに対して、後件部の書き方を統制することが望ましい。しかしながら、現時点では、まずは、FSD の具体例を蓄積することを優先し、次に示すような最低限の原則のみを適用する。

3.7.1 原則

後件部の記述にあたっては、次のことを原則とする。

1. 前件部に現れた記号〈〉で囲まれた要素を含める。
2. 述語用法の場合は、その述語の動作性 (動作を表すか、状態を表すか)、意志性 (意志の有無) などの情報を含めるようにする。
3. 名詞用法の場合は、可能な限り、その名詞の上位語 (どのような基本語の下位に位置付けられるか) を示す。

3.7.2 言い換え的後件部

後件部が言い換え的であるということは、後件部自身が、そのまま普通に使える言語表現 (文型) であることを意味する。言い換え的である場合は、後件部の最後を「こと」で終えることを原則とする。

【小柄】　《小柄な》〈人〉とは、
体が小さい〈人〉のこと。

【巧妙】　[〈誰か〉が〈何か〉を]《巧妙に》〈行なう〉とは、
[〈誰か〉が〈何か〉を] {みごとに | うまく} 〈行なう〉こと。

【凝らす】　〈人〉が〈何か〉に《目を凝らす》とは、
〈人〉が〈何か〉をよく見ること。

【根拠】　〈判断・主張・行動など〉の《根拠》とは、
〈判断・主張・行動など〉のもとになる理由や基準のこと。

言い換え的後件部では、特に、名詞用法と述語用法において、ターゲットの上位語を示すことが比較的容易である。上記の例では、「目を凝らす」の上位語は「見る」、「根拠」の上位語は「理由」または「基準」であることを明確に示している。

3.7.3 説明的後件部

ターゲットに対して適当な言い換え表現が存在しない場合は、その表現が意味するところを、説明的に示すしか方法がない。説明性が高い場合は、後件部の末尾に「ことをいう」を用いる。

【孤独】　《孤独な》〈人〉とは、
仲間や家族がなく、ひとりで寂しい〈人〉のことをいう。

【孤独】　〈誰か〉が《孤独を〈感じる〉》とは、
〈誰か〉がひとりで寂しいと〈感じる〉ことをいう。

【後悔】　〈ある人〉が〈ある行ない〉を《後悔する》とは、
〈ある人〉が〈ある行ない〉について、あとになって残念に思うことをいう。

3.7.4　使用する語彙

後件部は、基本語のみで記述するのが理想であるが、現時点では基本語の範囲を明確には定めていないので、そのような統制は採用しない。但し、ターゲット表現より平易な表現を用いることを心掛ける。

4　FSD 作成と具体例

今回、約 100 語に対して、実際に FSD を作成する試行を行なった。

1. 100 語は、旧日本語能力試験の出題基準 [6] の語彙表の「コ」で始まる語のうち、旧 1 級に位置付けられている語から選んだ。
2. FSD の作成に当たっては、既存の辞書 (主に、岩波国語辞典 (第 7 版新版)、明鏡国語辞典 (第 2 版)、三省堂例解小学国語辞典 (第 3 版)、三省堂国語辞典（第 7 版)) と『現代日本語書き言葉均衡コーパス (BCCWJ)』の用例を参考にした。
3. FSD は、できるだけ旧 2 級以下の語彙で記述することとした。

以下に、品詞別に具体例を示す。なお、これらの具体例は、本稿執筆時点のものであり、完全に確定しているわけではない。ガイドライン試案の修正と並行して、日々修正を行なっている。

4.1　形容詞

形容詞は、一般に、述語用法、連体修飾用法、連用修飾用法をとる。それぞれの用法間で意味の変異が小さい場合は、ターゲットとして、頻度の高い用法を採用する。複数の語義や、表層形に差異がある場合は、複数のターゲットを立てる。以下に例を示す。

【固有】(1) 〈何か〉に《固有 {の｜な}》〈性質・問題・知識など〉とは、
〈何か〉に特有な〈性質・問題・知識など〉のこと。
(2) 〈何か〉に《固有の》〈財産・領土・権利など〉とは、
他から与えられたのではなく、もとからそこにあった〈財産・領土・権利など〉のことをいう。
【快い】(1) ［〈誰か〉が〈何か〉を］《快く》〈思う・感じる〉とは、
［〈誰か〉が〈何か〉を］楽しい・気持ちいいと〈思う・感じる〉こと。
(2) ［〈誰か〉が〈何か：依頼など〉を］《快く》〈行なう：引き受ける〉とは、
［〈誰か〉が〈何か〉を］気持ちよく〈行なう〉ことをいう。
(3) 〈風・音楽など〉が《快い》とは、
〈風・音楽など〉を気持ちよいと感じている様子をいう。

被連体修飾語や格要素の記述をどのように統制すべきかは、よく考える必要がある。英語の FSD では、後件部で “it” などの代名詞を使うのが自然であるが、日本語には馴染まないような気がする。〈財産・領土・権利など〉のように、典型的な例を 3 つ程度示したいが、そうすると後件部が長くなる。【快い】(2) で示したように、前件部では〈何か：依頼など〉のように〈クラス：具体例〉という形で示し、後件部では〈クラス〉で参照するという方法が、一つの候補かもしれない。

4.2　副詞

一般に、副詞の用法や意味を記述するのはかなり難しい。しかしながら、やっかいな副詞の大部分は基本語に含まれるで、現時点では対象となっていない。それでもなお、「殊(こと)に」

のような語が存在する。

【殊に】〈何か〉が《殊に》〈どうだ：良い・目立っている・重要だ〉とは
〈何か〉がとても〈どうだ〉ということ。

この例では、被連用修飾語の用言が形容詞相当の語であることを示そうとしている。

4.3 動詞

動詞では、どれだけの語義を認定し、独立したターゲットとして立てるかが問題となる(現時点では、それほど真剣に検討していない)。

【こじれる】(1)〈話し合い・関係〉が《こじれる》とは、
［人と人との意見が合わず］〈話し合い・関係〉が悪い方へ進んでしまうことをいう。
(2)〈病気・問題〉が《こじれる》とは、
〈病気〉が悪くなったり、〈問題〉が解決しなかったりすることをいう。

【漉す】〈人〉が〈［液体状の］何か〉を〈［ざるなどの］道具〉で《こす》とは、
〈人〉が〈道具〉を使い、〈液体状のもの〉からそこに含まれる［小さな］かたまりを取り除くこと。

4.4 名詞

名詞は、おおきく、普通名詞、形容詞的な名詞、サ変名詞に分けて検討している。

普通名詞は、あまり工夫のしようがない。複合名詞を作りやすい名詞の場合は、ターゲットにそれを含める。

【行為】(1)〈誰か〉の《行為》とは、
〈誰か〉の行ないのこと。
(2)［〈誰か〉の］《〈犯罪・違反・危険〉行為》とは、
［〈誰か〉による］〈犯罪・違反・危険〉を伴う行ないのことをいう。

形容詞的な名詞とは、品詞区分では名詞に分類されるが、意味的にはある種の状態や状況を表すものをさす。用法的には、やっかいなものが多い。

【交互】〈A〉と〈B〉を《交互に》〈行なう〉とは、
ＡＢＡＢのように、〈A〉と〈B〉を順に繰り返して〈行なう〉ことをいう。

【個別】(1)《個別の》〈何か〉とは、
{ひとつひとつの｜それぞれの}〈何か〉のこと。
(2)〈何か〉を《個別に》〈扱う：決める・知らせる・包むなど〉とは、
〈何か〉をそれぞれ別々に〈扱う〉こと。

サ変名詞とは、接尾辞「する」を伴って動詞化する名詞をさす。原則として述語用法をターゲットとするが、名詞用法もターゲットとすべき場合がある。

【合議】〈人々〉が〈あること〉について《合議する》とは、
〈人々〉が〈あること〉について互いに相談することをいう。

【講習】(1)〈知識・技術など〉の《講習（会)》とは、
［希望する人々を集めて］〈知識・技術など〉について教える会合のことをいう。

(2) 〈誰か〉が〈知識・技術など〉について［〈人々〉に］《講習する》とは、
［講習（会）に参加した〈人々〉に］〈誰か〉が〈知識・技術など〉について教えることをいう。
(3) 〈誰か〉が〈知識・技術など〉について《講習を受ける》とは、
〈誰か〉が講習（会）に参加して、〈知識・技術など〉について学ぶことをいう。

5　現状のまとめ

これまでの試行により、前件部の文型の示し方については、ほぼ固まってきたと考えている。主要な用法は 4 種類であり、ターゲットの用法を定めれば、示すべき文型はおおよそ決まってくる。しかしながら、たとえ動詞であっても、連体修飾用法を示すべき場合 (「凝った」) もあり、品詞だけから示すべきターゲットの形式・用法を限定することはできない。それぞれの語に対して用例を観察し、どのような形式・用法をターゲットにすべきかを決定する必要がある。

格要素や被修飾語などの〈〉の中身をどのように選定・記述すべきかは、用例を観察しても悩む場合が多い。この問題に対しては、短期的には、より簡単に用例を観察することができるようなツールを実現することで対処する予定である。しかしながら、長期的には、標準的に使用する語彙リスト (階層的シソーラス形式となっていることが望ましい) を、体言 (名詞) と用言 (動詞・形容詞) に対して定めることが必要になろう。

現時点であまり方針が固まっていないのは、後件部の文型および語彙の統制方法である。これらを統制するためには、意味合成の類型化が不可欠と考えているが、これをどのように進めるべきかについては、明確なアイディアはない。より多くの FSD を蓄積し、そこから徐々に指針が見えてくることを期待したい。

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

参考文献

[1] 内木賢吾, 佐藤理史, 駒谷和範. 日本語クロスワードを解く：性能向上の検討. 2013 年度人工知能学会全国大会 (第 27 回), 2D1-4, 2013.

[2] 佐藤理史, 加納隼人, 西村翔平, 駒谷和範. センター試験『国語』現代文の傍線部問題を解くベースライン法. 情報処理学会研究報告, Vol. 2013-NL-212, No. 5, 2013.

[3] Michael Mayer, editor. *Longman Dictionary of Contemporary English, 5th Edition*. Pearson Education Limited, 2009.

[4] John Sinclair, editor. *COBUILD Advanced Dictionary of Engish, 7th Edition*. National Geographic Learning, 2012.

[5] Sadao Kurohashi and Yasuyuki Sakai. Semantic analysis of japanese noun phrases - a new approach to dictionary-based understanding. In *Proc. of ACL-1999*, pp. 481–488, 1999.

[6] 財団法人日本国際教育協会国際交流基金. 日本語能力試験出題基準【改訂版】. 凡人社, 2002.

統計的現代語訳モデルを用いたセンター試験古文問題解答

A Similarity-based Model using Statistical Machine Translation for Solving Questions of Japanese Classics in National Center Test for University Admissions

横野 光 (国立情報学研究所)
星野 翔 (総合研究大学院大学)

1　はじめに

現在，国立情報学研究所で推進している人工頭脳プロジェクト「ロボットは東大に入れるか」は，細分化された人工知能分野を再統合することで分野全体としてどこまでできているのか，また残っている問題は何かということを明らかにすることを目的とし，そのベンチマークとして 2016 年度に大学入試センター試験を，2021 年度に東京大学入学試験を突破することを目標としている [新井，松崎 (2012)]．

センター試験は一般的に，国語，英語，数学，理科，社会の 5 科目からなっており，それぞれに対して解答に必要となる知識，手法は異なる．そのため，問題解答にはそれぞれの科目において検討する必要がある．我々は特に国語の古文問題を対象として，その解答に取り組んでいる．

本論文では，センター試験国語の古文問題ではどのような問題が出題されるのかについて分析を行い，このうち内容理解に関する問題に対して，統計的機械翻訳モデルを用いた解答手法について述べる．

2　センター試験古文問題の分析

センター試験国語古文問題について，本試験と追試験の合わせて過去 16 回分のセンター試験で出題されているか問題の分析を行った．過去に行われたセンター試験の問題は大学入試センター[1]で公開されており，自由に閲覧できる．

問われている内容に基づいた設問の分類の結果を表 1 に示す．出題数は分析に利用した 16 回分の試験問題での小問の数を表す．設問は大きく分けると，文法理解に関する問題，内容理解に関する問題，その他に分類できる．1 回分の試験での古文問題の大まかな構成としては，第 1 問目に本文傍線部の現代語訳，文法問題が 1 問程度，和歌に関する問題や文学史に関する問題は出題されない回もあるがそれぞれ 1 問程度出題される．残りは内容理解に関する問題であり，半分以上の割合を占める．そのため，内容理解の問題が解けるかどうかが高得点を得るためには重要となる．

以降，それぞれの分類について説明する．

2.1　文法問題

文法問題では主に単語の活用や助動詞の用法に関する問題，敬意表現に関する問題が出題される．単語の活用などに関する問題では，傍線部の表現がどのような形態素で構成されているか，それぞれの品詞や活用形，助動詞の場合にはその用法などが問われる．問題の形式としては傍線部の構成要素を問うものの他に，本文中の表層が同じ語のなかで，品詞が違うものを答えるという形式のものもある．

[1] http://www.dnc.ac.jp/modules/center_exam/

表 1: 問題の分類

大分類	小分類	出題数
文法理解	表現の構成要素 (品詞など) の同定	12
	主語の同定	1
	会話文の発話者推定	2
	敬意表現 (敬語の種類など)	4
内容理解	現代語訳	16
	内容理解	29
	心情説明	8
	理由説明	5
	和歌の表現技法	1
	和歌の解釈	5
	文章全体の内容理解	10
その他	古典知識	4

敬意表現に関する問題では，傍線部の表現が尊敬語であるか，謙譲語であるか，丁寧語であるか，またどの登場人物からどの登場人物への敬意であるかなどが問われる．

この他にも文中の述語における主語や発話文の話し手が明示的に現れていない文に対して，それらを同定するという問題もある．

単語の活用などに関する問題は，例えば中古和文 Unidic[小木曽ら (2010)] を用いた形態素解析器 MeCab[2]によって形態素解析を行い，その形態素情報を利用することで解答が可能であると考えられる．しかし，助動詞の用法には多義性があるため，より精度を上げるためには複数の用法を持つ助動詞に対して，その表現中で用いられている用法が何かを推定するモデルが必要となる．

敬意表現に関しては，その表現がどのタイプの敬語であるかを推定する必要がある．解答方法としてはそのような情報を形態素辞書に記述しておき，それを利用して解答するという手法が考えられるが，全ての語が一つの種類の敬意を持つわけではなく，複数の用法を持つ語もあり，そのような語に対しては文脈からタイプを推定しなければならない．また，例えば，天皇は身分関係の上位に必ず来るといったような，人物間の身分に関しての知識少なからず必要となる．

2.2 内容理解

古文の問題の大半は本文の内容理解に関する問題である．実際に問われるものには，単純な現代語訳から，本文中の記述についての理由や説明，そこから推測される登場人物の心情など様々なものがある．このような問題を解くためには，選択肢に書かれてある記述が本文と照らし合わせて正しいかどうかを推定する必要がある．

内容理解に関する問題には，登場人物の心情推定など現代文の小説問題でも出題されるようなものがある．しかし，必要となる知識は古文と現代文では異なると考えられる．例えば，小説の心情推定の問題では，その心情を表す直接的な表現が本文中に出現することはあまりなく，“父は縁側で背中を丸めて足の爪を切っていた。” といった人物の行動に関する表現から，その心情 (この場合は，父親が落ち込んでいる，あるいは悲しんでいる) を推測するというようなことが必要となる．このように本文中に直接的に書かれていないことを使って解かなければならないような問題が小説の問題では多い．

[2]https://code.google.com/p/mecab/

一方で，古文の問題における心情推定の問題では，手掛かりとなる表現が小説等に比べて直接現れていることが多い．例えば，図 1 の問題では，①が正解であるが，選択肢中の “兵衛佐の妹なら美

問 5 この文章を通して、兵衛佐の妹に対する兵部卿宮の気持ちはどのように移り変わっているか。その説明として最も適当なものを、次の①～⑤のうちから一つ選べ。解答番号は 27。

① 最初はお気に入りの兵衛佐の妹なら美しいだろうから会ってみたいと執心していたのに、何度も贈った和歌をはねつけられて落胆した。しかし、常磐に兵衛佐の邸に忍び込むように進言されて再び希望を見出した。
② 最初はほんの遊び心で臣下の兵衛佐の妹に会ってみようと思ったが、和歌のやりとりをするうちに本気で兵衛佐の妹を想うようになってきた。そしてついに、兵衛佐が不在の夜に勇んで兵衛佐の邸に忍び込むことにした。
(以下省略)

図 1: 内容理解の問題例 (2009 年度国語本試験より引用)

しいだろうから会ってみたいと執心していた” という箇所は，本文中の “女にてかれが妹ならば、いかにいつくしからん。あはれ、見ばや」と、深く御心移りて” に合致し，“落胆した” という箇所は本文中の “宮の思し召し沈み給へる” に合致している．これは論説文や小説のような現代文の問題では，本文で書かれてあることは受験者は大まかなところは理解できるとしたうえで，設問ではより深い理解が必要となる問題が出題されるのに対し，古文の問題では「古文が正しく理解できるか」が焦点となり，現代文として解釈できるかどうかに重点が置かれているからだと考えられる．従って，本文において表層に現れていることが理解できれば，古文の問題の解答は可能であると考えられる．

また，本文中に和歌が出現している場合，その和歌で使用されている表現技法についての説明を求める問題やその和歌の意味を問う問題が出題されることがある．これらの問題では縁語や枕詞といった和歌独自の修辞技法を理解する必要がある．また，古文としての表層的な解釈だけでなく，それによってその歌の詠み手が何を伝えようとしているのかといった，深い解釈が必要となる．

2.3　その他 (文学史)

文法知識，内容理解の問題に比べると数は少ないが，文学史の知識に関する問題が出題されることもある．文学史の問題では本文の内容ではなく，本文に用いられている文献の種類やその成立時期に関係した設問が出されることがある．これらの問題は本文の内容とは直接関係はなく，あらかじめ文学史に関する知識を保有しておき，その知識から選択肢の内容が含意できるかどうかを判定することで解答する．これは世界史や日本史の問題で見られるタイプの問題であり，自然言語処理における含意関係認識のタスク [田，宮尾 (2013)] として解くことができると考えられる．

3　統計的現代語訳モデルを用いた問題解答

2.2 節で述べたように古文の内容理解の問題では古文本文を現代文として解釈できるかどうかが重要であると考えられる．そこで本研究では，古文を機械翻訳モデルによって現代語へと翻訳し，それを用いて内容理解に関する問題の解答を考える[3]．

[3]いわゆる “古典” は新しいものは生成されないため，全ての古典に対してその現代語訳を付与したコーパスを構築するという方法が考えられるが現実的ではない

3.1 古文-現代語翻訳モデル

統計的な機械翻訳では翻訳元の言語のテキストと翻訳先の言語のテキストの対応が取れたコーパスが必要となる．現時点において我々が利用可能な古文-現代文コーパスとしては小学館の新編古典文学全集の電子化データ (以降，小学館コーパスと呼ぶ) がある．このコーパスは古代から近世までの古典文学作品に対して，その本文と現代語訳，注釈が対応づけられたデータである．

このコーパスの古文と現代語訳との対応は文よりも大きなセグメント単位でしかなされておらず，統計的機械翻訳に望ましい文対応付けされたデータではない．そこで対応するセグメント内の文の数に注目し，古文と現代文で文の数が等しければ，前から順に文の対応付けを行い，文の数が等しくなければ，文の数が多いセグメントを少ない方のセグメントの数に合わせて分割することで，文対応の取れた対訳コーパスを作成する [星野ら (2014)]．

学習データ作成に使用した小学館コーパスの統計を表 2 に示す．このコーパスから得られた対訳コーパスの規模は 86684 文対であった．このデータを学習コーパスとして Moses[Koehn et. al. (2007)] を用いて翻訳モデルを構築する．

表 2: 小学館コーパス統計情報

	古文	現代語	合計
単語数	2,837,101	3,720,257	6,557,358
文字数	12,763,402	17,300,081	30,063,483
セグメント数		19,102	

3.2 類似度に基づいた問題解答

内容理解に関する問題では基本的に本文の内容にあった選択肢が正答であると考えられる．そこで 3.1 節で述べた翻訳モデルによって現代語訳した本文と選択肢との類似度を計算し，その値をもとに解を決定する．

本文と選択肢の類似度はコサイン類似度で計算する．コサイン類似度は本文の特徴ベクトル T と選択肢の特徴ベクトル c_i に対して，以下の式で求める．

$$sim(T, c_i) = \frac{T \cdot c_i}{|T||c_i|}$$

本文と選択肢の特徴ベクトルは形態素の n-gram を要素とする．

内容理解に関する問題の多くは，本文中で指示されている傍線部についてのものである．このような問題では，その傍線部の近くに解答の手がかりが存在することが多い．そこで，類似度の計算において問題本文全てを利用するのではなく，傍線部を含む文の前 l 文から後ろ m 文のみを利用する．

提案モデルでは，基本的には類似度が高くなる選択肢を解として出力するが，問題には “適切でないものを選べ” と指示されたものもある．このような問題に対しては類似度が最も低いものを解として出力する．

4 実験

2009 年度，2005 年度のセンター試験国語問題から傍線部の現代語訳問題と内容理解に関する問題を人手で選別し，これらを用いて提案手法の評価を行った．提案手法では本文として傍線部を含む文の前後何文を利用するか，また，n-gram の n の値をどうするかを与える必要がある．これらに対しては別年度の問題を開発データとして決定した．

表 3: 実験結果

問題	正答率
2009 年度国語本試験	0.4(2/5)
2009 年度国語追試験	0.4(2/5)
2005 年度国語 1 本試験	0.33(2/6)
2005 年度国語 1 追試験	0.17(1/6)
2005 年度国語 1・2 本試験	0.2(2/5)
2005 年度国語 1・2 追試験	0.43(3/7)
合計	0.32(11/34)

実験結果を表 3 に示す．括弧の中は正答数と実験の対象となった設問数を表している．実験で使用した問題数が多くないため，この結果から提案手法の有効性を主張することはできないが，1 設問につき選択肢の数は 5 個程度であり，ランダムに答えた場合の正答率が約 0.2 であることを踏まえると，全く見込みがないとは言えないであろう．しかし，試験毎に正答率にばらつきがあり，安定した性能ではないというのは今後の検討課題である．

翻訳モデルの性能に関して，実験で使用した問題の本文とそれに対する翻訳結果を図 2 に示す．

問題文　ある時、中将、昼寝せさせ給ひける御夢に、いづちともなく荻薄生ひ茂りたる野原の、まことに心すごき所に、うす絹のすそ、露にうちしほれたる女房ただひとり立ち給へり。いたはしと思ひて立ち寄り見給へば、わが母にておはせり。中将を見たてまつりて、袖もしぼりあへず、仰せけるは、「都に捨ておき給ひしその嘆きに、月日の行くもおぼえはべらねども、はや六年になりぬ。この思ひゆゑ、われこの世になき身となりにき」とて、さもうらめしげなる気色にて、道もそこはかとなき野中を西へ向きて行き給ふ、とおぼして、夢うちさめぬ。

参照訳　あるとき、中将が昼寝をなさっていたその御夢に、どことはわからない萩やすすきがおい繁っている野原で、本当に人けがなくものさびしい所に、薄衣のすそが露に濡れている女性がたった一人で立っていらっしゃる。中将がお気の毒にと思って近寄ってご覧になると、自分の母上でいらっしゃった。母は中将を見申し上げると、絞りきれないほど袖を涙でぬらして、「あなたが私を都に見捨てたまま消息を絶ってしまわれたその嘆きのために、月日のたつのも忘れておりますけれども、もはや六年になりました。この悲しい思いのために、私はこの世の者ではなくなってしまいました」とおっしゃって、いかにもうらめしい様子で、道もはっきりと分からない野原の中を西方浄土の方へ向かっていらっしゃる、と思われたところで夢からさめた。

翻訳モデルによる翻訳結果　ある時、中将は、ちょうどお昼寝をなさった時の夢に、どこへともなく荻薄の生い茂っている野原のは、本当にうつ所に、うす絹の裾、露に内でうちしおれて、ただ一人で立っていられた。不憫でならないと思って、お立ち寄りご覧になると、ご自分の母でいらっしゃいました。中将のを拝見して、袖も涙でぬらし終らないうちに、おっしゃったことには、「都に捨ておきれたその嘆きに、月日の行方も思われませんが、早く六年になった。この思いから、私はこの世にいない身となってしまった」といって、さも恨めしそうな様子で、道もどこがどうということもない野中を西の方へ向って行った、とお思いになって、夢がさめた。

図 2: 翻訳例 (2005 年度センター試験国語 1・2 より引用)

敬意表現や助動詞に関してはある程度は翻訳できているが，一部の内容語に関しては翻訳できておらず，翻訳元の表現がそのまま用いられている．これは学習に使用したデータが少ないことが原因の一つであると考えられる．このような場合，一般的には対訳データを増やすことによって改善できるが，我々が取り組んでいる古文-現代文翻訳においては，利用可能な電子化された古文データはそれほど多くないため，対訳データを作成するためのコストが大きくなる．このことから，対訳辞書を

構築し，それを用いることによる翻訳精度の向上を図る．

5　おわりに

本論文では，センター試験国語古文問題の分類を行い，そのうち内容理解に関する問題に対して，統計的翻訳モデルを用いて古文を現代文に翻訳し，それと選択肢との類似度に基づいて解答するモデルについて述べた．

2 節で述べたように，古文問題には内容理解に関する問題の他に，文法に関する問題や和歌に関する問題などがある．文法に関する問題に関しては，中古和文 Unidic を用いた形態素解析を行うことで表現の品詞情報を得ることができるため，これを用いて問題に解答するという手法が考えられる．

和歌に関する問題に関しては，本論文で述べた手法と同様に現代語訳をすることで和歌の理解に関する問題に解答することは可能であると考えられるが，和歌には独特の表現技法や訳し方があるため，これをどのようにモデルに組み入れるかが重要であると考えられる．

現状では古文を対象とした言語処理ツールや言語資源は多くなく，また言語資源の規模を拡大していくというのは現代語のテキストと比べると困難である．この，小規模な言語資源しかなく，規模の拡大にコストがかかる，という状況は古文に限らず存在する．そのため，本研究では古文問題の解答モデルの構築という応用を通じて，小規模の言語資源を如何に効率的に活用するかということにも焦点を当て，その方法論の確立に取り組む予定である．

謝辞

本研究では，国語研究所通時コーパスプロジェクトから小学館新編日本文学全集データの提供を受けた．また，センター試験過去問データは独立行政法人大学入試センター，株式会社ジェイシー教育研究所から提供を受けた．

参考文献

[星野ら (2014)] 星野 翔, 宮尾 祐介, 大橋 駿介, 相澤 彰子, 横野 光 (2014),「対照コーパスを用いた古文の現代語機械翻訳」, 言語処理学会第 20 回年次大会 (to appear).

[Koehn et. al. (2007)] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst (2007)「*Moses: Open source toolkit for statistical machine translation*」In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp.177-180.

[新井，松崎 (2012)] 新井 紀子, 松崎拓也 (2012),「ロボットは東大に入れるか? —国立情報学研究所「人工頭脳」プロジェクト—」, 人工知能学会誌, Vol.27, No.5, pp.463-469.

[小木曽ら (2010)] 小木曽 智信, 小椋 秀樹, 田中 牧郎, 近藤 明日子, 伝 康晴 (2010),「中古和文を対象とした形態素解析辞書の開発」, 情報処理学会研究報告 人文科学とコンピュータ CH-85.

[田，宮尾 (2013)] 田 然，宮尾 祐介 (2013),「関係代数に基づく推論の含意関係認識への応用」, 2013 年度人工知能学会全国大会 (JSAI2013).

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

八木 豊（株式会社ピコラボ）
ホドシチェク・ボル（国立国語研究所）
阿辺川 武（国立情報学研究所）
仁科 喜久子（東京工業大学）

Evaluation of Error Detection in Japanese Composition Support System “Nutmeg”

Yutaka YAGI (Picolab Co., Ltd.)
Bor HODOŠČEK (National Institute for Japanese Language and Linguistics)
Takeshi ABEKAWA (National Institute of Informatics)
Kikuko NISHINA (Tokyo Institute of Technology)

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）、八木（2012））。

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。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、Halliday（2004）はレジスター機能として次の 3 項目を挙げている。

(1) コミュニケーションの目的と主題に関わる「フィールド」（Field of Discourse）
(2) コミュニケーションを行うための手段に関わる「モード」（Mode of Discourse）
(3) コミュニケーションパートナー同士の関係に関わる「テナー」（Tenor of Discourse）

つまり、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・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の誤り検出について示し、学習者作文コーパスに基づいてその精度を評価する。また、日本語学習者に実際にレポート形式の作文を書いてもらい、その過程で学習者がシステムを利用して推敲する評価実験の経過を報告する。

２．レジスターの誤り検出

ホドシチェク（2011）では、書籍、雑誌、新聞など様々なレジスターの書き言葉を収録している BCCWJ と、科学技術論文のコーパスを使用してレジスターの誤り判定を行う手法を

提案している。学習者が作文の目的とするレジスターを想定し、BCCWJ の各レジスターと科学技術論文の中から目的のレジスターに近いデータを準正用データ、目的のレジスターから遠いデータを準誤用データとして設定する。そのうえで、「名詞＋格助詞＋動詞」、「名詞＋格助詞＋形容詞」、「形容詞＋名詞」の 3 つのパターンに一致する共起表現の出現分布をもとにカイ二乗検定を行い、準誤用データにおける出現が有意に多い場合に、その共起表現は学習者が作文の目的とするレジスターの下ではふさわしくない表現、即ち、誤用であると判定するものである。八木（2012）では、学習者の作文の目的を論文やレポートを書くことと想定し、科学技術論文、白書など比較的硬い文章のレジスターを準正用データ、Yahoo!知恵袋、Yahoo!ブログ、国会会議録など比較的くだけた文章のレジスターを準誤用データとすることで、「ことがある」、「問題が起きる」、「一緒にする」、「いい経験」のように日本語として誤っているわけではないが論文やレポートの場合には別の表現に書き換えたほうがよい表現を、レジスターの誤りとして自動的に検出できることを述べた。

本稿では、レジスターの誤り検出対象が上記の特定の 3 つのパターンに制限されていた点を改善し、任意の句の 3-gram、2-gram、1-gram および、単独の形態素について同様の手法でレジスターの誤りを検出できるように拡張したものを使用する。作文の目的として想定するのは同様に論文やレポートで、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た。ただし、科学技術論文には言語処理学会年次大会予稿集の本文が新たに追加されている。

３．学習者作文コーパスに基づく評価

レジスターの誤り検出の精度を確認するために、学習者作文コーパス「なたね」のデータを利用して評価を行った。

「なたね」は、我々が独自に収集した学習者作文に対して日本語教師による添削を行った誤用タグ付きデータで、2014 年 2 月現在、大学院や大学あるいは語学学校に在籍する 192 人の日本語学習者による 285 作文（総文字数 205,520 字）に含まれる約 6,500 箇所の誤用に対しておよそ 9,000 件の誤用タグを付与して公開中である[1]。誤用タグは大きく「誤用の対象」、「誤用の内容」、「誤用の要因・背景」という 3 つの視点から成り、さらにそれぞれを 3 階層に細分類している。レジスターに関する誤用は「誤用の要因・背景」の一つで、483 件が登録されている。そこから 100 件を含む文を取り出してシステムによるレジスターの誤り検出を行い、日本語教師が添削で指摘したレジスターの誤用との一致度合いを人手により確認した。結果は、日本語教師が指摘した箇所の再現率が 78.0%、システムの検出結果全体の精度が 77.6%であった。内訳を表 1 に示す。

まず、日本語教師が添削で指摘した誤用 100 件のうち、78 件をシステムが自動的に検出した。この中には、以下に挙げる誤用例 1、2 のような副詞に関するものや、誤用例 3、4

表 1 日本語教師の指摘箇所とシステムの検出箇所の比較

	システムによる検出有り	システムによる検出無し
日本語教師による指摘有り	78 件	22 件
日本語教師による指摘無し	92 件 （うち 38 件は誤検出）	n/a

[1] 学習者作文コーパス「なたね」　http://hinoki.ryu.titech.ac.jp/natane

のような句に関するものも含まれている。これらは、レジスターの誤り検出の拡張前には検出対象に含まれておらず、今回、拡張した効果といえる。

【誤用例 1】なぜなら日本は今少子化により、労働不足の問題はどんどん大きくなる。
【誤用例 2】ちょっとロボットに興味をもっている私はわくわくした。
【誤用例 3】今中国ではお父さん一人の給料では家族全員を養うことが非常に難しい。
【誤用例 4】でも、外国人の先輩は大学院で研究しながら、会社に務めるそうです。
※例中の下線部は日本語教師が添削で指摘したレジスターの誤用箇所

一方、誤用例 5、6 はシステムでは検出できなかった誤用の一例である。

【誤用例 5】たぶん誇張だと思いました。
【誤用例 6】別に利益のためじゃない将来保証のためである。

データを確認したところ、準正用データとしている科学技術論文のなかに言語処理学会の会誌「自然言語処理」および年次大会予稿集の本文が登録されており、「たぶん」や「じゃない」のような表現が含まれていた。その結果、準正用、準誤用における出現頻度に有意差が出ず、誤用として検出できなかったことになる。ただし、「自然言語処理」および年次大会予稿集の当該箇所では例文の一部として使用されており、出現頻度を集計する際は、そういった例文や他から引用されている文を除外するなどの対応が必要であると考える。

また、システムは、同一文中に含まれる日本語教師が指摘しなかった 92 箇所の誤用を検出した。そのうちの 54 件は誤用として妥当なもので、残り 38 件は誤検出だった。さらに、54 件の半数以上は「です」、「ます」の使用を指摘したものである。これらについては、日本語教師が見落としたのではなく、日本語教師のほうで意図的にそこまでチェックせず、その他の添削を優先したものと思われる。

４．日本語学習者による評価実験

2014 年 1 月から、上記のレジスター誤り検出機能を備えた作文推敲支援システムの評価実験を行っている。学習者に依頼する内容の大まかな手順は以下の通りである。

1. J-CAT（Japanese Computernized Adaptive Test）を受験する
2. 背景調査のアンケートに回答する
3. 4 つの課題について、システムを使用して作文を書く
4. システムに関するアンケートに回答する

実験に先立ち、学習者の現在の日本語能力および言語的な背景を確認し、後日、分析に使用するため、J-CAT の受験と背景調査のアンケートを必須とした。実験では、あらかじめ決められた 4 つの課題について、順番に作文を書いていく。各課題は、最初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、1 回だけ、レジスターの誤り検出機能による添削を受けて書き直せるようにした。1 回だけに制限した理由は、添削前後の作文の比較を容易にするためである。また、1 つの課題を終えて次の課題に取り掛かるまでに最低 3 日間空けなければならないよう設定した。

図 1 は、レジスターの誤り検出機能による添削を受けた後、学習者に表示される添削結果画面である。画面左下に添削結果を表示しており、背景色を変えた箇所や下線の箇所が、システムが誤りの可能性を指摘している単語や表現である。学習者はそれぞれの単語や表現をクリックして指摘された内容を確認し、必要なら、辞書や「なつめ」で使い方を調べ



図　1　システムによる添削結果画面

たうえで、画面右下の入力欄で作文を書き直す。

日本語学習者による評価実験は開始したばかりであるが、システムの添削結果を受けて、以下のように適切な修正を施す学習者も散見される。

【修正例 1】私が日本人について理解できないことはギャルです→である。

【修正例 2】人をあっ→に会ったりするのは私が考えたより少なかった。

【修正例 3】性についての認識も違うだ→異なると思う。

※矢印の左側の下線部がシステムによる指摘箇所、矢印の右側が学習者による修正内容

システムによる指摘箇所と学習者による修正内容を始め、作文途中の文章（10 秒毎に記録）、作文完了までに要した時間、クリックして確認した箇所など、システム内における学習者の操作はバックエンドでデータベースに記録しており、今後は、学習者の日本語能力や言語的な背景と合わせて分析する予定である。

謝辞

本研究は、文部科学省科学研究費補助金基盤研究（C）「日本語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。

参考文献

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.

八木豊、ホドシチェク・ボル、仁科喜久子（2012）「BCCWJ と学習者作文コーパスを利用した日本語作文支援－表記と共起に関する誤用添削プロトタイプ構築－」第 1 回コーパス日本語学ワークショップ予稿集, pp.315-320.

Halliday M.A.K. and C.M.I.M. Matthiessen (2004). An Introduction to Functional Grammar. 3rd ed. London: Arnold

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.

BCCWJ 図書館サブコーパスの 文体情報検索ツールによるテキスト分析

柏野 和佳子*（国立国語研究所 言語資源研究系）
中村 壮範（国立国語研究所 コーパス開発センター）

Text Analysis of the Library Subcorpus of the BCCWJ Using a Text Search Tool

Wakako Kashino （Dept. Corpus Studies, NINJAL）
Takenori Nakamura （Center for Corpus Development, NINJAL）

１．はじめに

国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」（2009-2012 年度）において，『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）に収録されている全ての書籍サンプルを対象に，文体情報を付与した（柏野ほか 2012a，柏野ほか 2012b）。文体情報とは，テキスト構造が単純（例：章節構造）なもの（全体の約 8 割）に付与した「専門度，客観度，硬度，くだけ度，および語りかけ性度」といった評定値や，テキスト構造・紙面形式などの点で上記分類になじまないもの（約 2 割）にその特徴を表すものとして付与した「対談，Q&A 形式，図解，用語解説」等の分類情報である。今回，我々は，それらの付与した文体情報をもとに，特定の文体的特徴をもつテキストを検索し，形態論情報等を一覧できる文体情報検索ツールを構築した。本ツールを用いて概観できる，図書館サブコーパスの文体的特徴について報告する。

２．付与した文体情報

２．１　アノテーション作業の概要

文体情報のアノテーション作業の概要は次のとおりである。

●対象：BCCWJに収録されている図書館サブコーパス（10,551サンプル）の書籍サンプル[1]。

●1サンプルの範囲と長さ：サンプル全体の印象判定を行うため，1サンプル全体を範囲とする。サンプルの一部より一定量の字数を揃えて抽出するようなことはあえてしない。1サンプルの平均はおよそ3,000語。

●作業者：言語データ構築経験有のおおよそ20～50代の女性，延べ9名。

●内容：

①文体判断が可能と判断されるもの，即ち，テキスト構造が単純（例：章節構造）なものを内容・表現の文体的特徴の印象判定により細分類する。（→2.2節）

②①以外，文体判断が単純にいかないと判断されるもの，即ち，テキスト構造・紙面形式に特徴をもつものを選別，分類する。（→2.3節）

●態勢：

①典型例抽出を目的とする作業（Kashino and Okumura 2010）

・判断のゆれを検証するために同一サンプルを作業者3人で判定。最終的には3人の結果を見直

* waka@ninjal.ac.jp

[1] 1986 年から 2005 年までの 20 年間に発行された書籍のうち，東京都内の 13 自治体以上の公共図書館で共通に所蔵されていた書籍を母集団とし，そこから抽出したサンプルから成るサブコーパスである。

し判定値を一つに決定。3,324サンプルまでこの態勢。（全体の31.5%）

②全体付与を目的とする作業（柏野ほか2012b）

・1サンプル作業者1人で判定。残りの①以外を作業し，全10,551サンプルを完了。

２．２　文体的特徴を表す分類指標

BCCWJ に収録されている書籍サンプルには，NDC（日本十進分類法）によるジャンルや，C コード（日本図書コード）による販売対象[2]，また，著者情報，形態論情報などに加え，テキストの内容・表現の文体的特徴を表す指標として，次のものを設定した。そのうえで，印象判定を行い，その結果を付与した。

(a) 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

(b) 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

(c) 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

(d) 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

(e) 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

２．３　文体判断が単純にいかない特徴をもつものの分類指標

文体判断が単純にいかないと判断されるもの，即ち，テキスト構造・紙面形式に特定の特徴をもつものがある。そこで，次のような分類指標を設け，該当するサンプルの選別，分類を行った。理由は，それらは，必要に応じて，その特徴による類型で整理分類をし，別途，その文体を吟味すべきと考えたからである。なお，いずれの分類指標も該当サンプルすべてに付与したため，複数の分類指標が付与されたサンプルも存在する。

［テキスト構造・紙面形式上の特徴］

(a) 対話系（対話，対談・座談，インタビュー，往復書簡，シナリオ，その他対話形式）

(b) 引用系（Q&A 形式，投稿形式，その他引用編集形式）

(c) 視覚表現多用系（コマ割多用，図解，その他写真やイラストの多用）

(d) データベースやリスト系（用語解説，辞書形式，見本・カタログ形式，その他リスト形式）

［内容や表現上の特徴］

(e) 前書きや後書きである

(f) 明治時代より以前の古い言葉が多い

(g) 外国語が多い

(h) 数式やプログラミング言語などが多い

(i) 法律文が多い

(j) 教育現場で使いがたそうである[3]

(k) その他一定量の「本文」が認めがたい

---

[2] C コードの第 1 桁は販売対象を示す。0:一般　1:教養　2:実用　3:専門　5:婦人　6:学参 I（小中）　7:学参 II（高校）　8:児童　9:雑誌扱い

[3] 例えば，暴力的な描写や性的な描写を含むものを区別するための指標である。文体情報付与のための指標というよりは，コーパス活用のためのテキスト整理の指標として設けたものである。

２．４　アノテーション文体情報のリスト

アノテーション文体情報はリスト形式でまとめている。次の表 1 にはじめの 10 件分を例として示す[4]。このリストを用いることにより，特定の文体的特徴をもつテキスト群の，該当するテキスト数や，コーパス収録時に付与されているサンプル ID，書名，NDC（日本十進分類法），C コード（図書分類コード）といった書誌情報を得ることまでは簡単にできるようになっている。しかしながら，このようなリストだけでは，絞り込んでいったテキストの内容分析にすぐにとりかかることができない。テキストを分析したい場合は，ここで得られたサンプル ID をもとに，別途，BCCWJ の収録 DVD より該当テキストを抽出するという大きな手間が必要になる。そこで，アノテーション文体情報をもっと簡便に活用できるように，特定の文体的特徴をもつテキストを検索し，形態論情報等をすぐさま一覧できる文体情報検索ツールを構築することとした。

表 1 アノテーション文体情報のリストの例

番号	書誌情報				文体的特徴					備考	非対象とする該当理由										
	SampleID	NDC	Cコード	タイトル	専門度	客観度	硬度	くだけ度	語りかけ性度	小説の主人公あるいは語り手の人称	対話系か	引用系か	視覚表現多用系か	データベースやリスト系か	前書きや後書きか	明治時代より以前の古い言葉が多いか	外国語が多いか	数式やプログラミング言語などが多いか	法律文が多いか	教育現場で使いたそうであるか	その他，一定量の「本文」が認めがたいか
1	LBa2_00007	288	1323	光明皇后	3 一般向き	2 どちらかといえば客観的	2 どちらかといえば硬い	3 くだけていない	3 特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否
2	LBa2_00027	292		NHK大黄河	3 一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	2 どちらかといえばくだけている	2 どちらかといえば語り性がある	[私+達]	否	否	否	否	否	否	否	否	否	否	否
3	LBa5_00003	596		日本の郷土料理	3 一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	3 くだけていない	3 特に語り性はない	私	否	否	否	否	否	否	否	否	否	否	否
4	LBa7_00003	762		フルトヴェングラー	3 一般向き	3 どちらかといえば主観的	3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否
5	LBa9_00002	933		呪われたブルー・エラー	3 一般向き		3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	わたし	否	否	否	否	否	否	否	否	否	否	否
6	LBa9_00062	913	0193	汚名	3 一般向き		2 どちらかといえば硬い	2 どちらかといえばくだけている	3 特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否
7	LBa9_00068	910	3391	日本文芸史	2 やや専門的な一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	3 くだけていない	2 どちらかといえば語り性がある	わたくし	否	否	否	否	否	否	否	否	否	否	否
8	LBa9_00107	913		アローン・アゲイン	3 一般向き		3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	わたし	否	否	否	否	否	否	否	否	否	否	否
9	LBan_00008			時間と空間の冒険	5 小学生・幼児向き		3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	わし	否	否	否	否	否	否	否	否	否	否	否
10	LBan_00022		8093	ちょっと気になる転校生	5 小学生・幼児向き		4 とても軟らかい	1 とてもくだけている	3 特に語り性はない	おれ	否	否	否	否	否	否	否	否	否	否	否

３．Web アプリケーション「図書館コーパス検索ツール」の構築

アノテーション文体情報のより簡便な活用を目的に，特定の文体的特徴をもつテキストを検索し，形態論情報等を一覧できる文体情報検索ツールとして，Web アプリケーション「図書館コーパス検索ツール」を構築している。先行事例としては， 1 億語規模の BNC に対し，Lee(2001)が 70 種類（書き言葉 46 種，話し言葉 24 種）の言語使用域や，対象読者の性別や年齢層別，著者の属性等によって分類したものがある。

３．１　「図書館コーパス検索ツール」の仕様

「図書館コーパス検索ツール」は，Internet Explorer などのブラウザから利用することができる Web アプリケーションである。インターネットが利用できる環境と標準的なブラウザがあれば，特別なソフトをインストールすることなく利用することができる。ただし，現在はまだ開発段階であるため，研究所内のネットワークからのみアクセス可能である。

BCCWJ の図書館サブコーパスに付与された文体情報及び，書誌情報に基づく検索機能と，

4 表の 11 列目の「備考：小説の主人公あるいは語り手の人称」とは，アノテーション作業時にわかる範囲で，小説の場合は主人公を，それ以外は筆者自身をどの人称で表現しているかの抽出を試みたものである。

文体情報と形態論情報(短単位)を利用した簡単な集計とグラフ表示の機能を備える。

システムの概要は以下のとおりである。

【ソフトウェア】

・開発ツール	Microsoft Visual Studio 2012
・フレームワーク	asp.net
・サーバ OS	Microsoft Windows Server 2008R2 Standard
・Web サーバソフト	Microsoft Internet Information Services ver7.5
・DBMS	Microsoft SQLServer2012

【ハードウェア】

・サーバ本体	DELL PowerEdge R310
・CPU	Intel Xeon X3430(2.40GHz，8MB キャッシュ，1333MHz)
・メモリ	24GB
・HDD	300GB 15000RPM(6Gbps SAS HDD/3.5 インチ)×3 ※RAID5 構成で実質容量 600GB

【動作確認】

Google Chrome32，Internet Explorer10

３．２　「図書館コーパス検索ツール」の表示画面

次のように 3 つの表示画面から構成される。

⑴検索画面

文体情報及び，書誌情報に基づく検索が可能。

図 1 は，NDC の一桁目が「6」であり，文章の硬軟が「とても硬い」という指定を行っている画面である。左下の「検索」の BOX に，自動的に検索式を生成し，表示する。

⑵検索結果の表示画面

図 2 のように，該当サンプルのリストを表示する。該当件数が多い場合は，複数ページに分割表示する。

⑶データ表示画面

図 1 の左下にある，「データ表示」というリンクテキストをクリックすると，検索結果のテキストデータの各種情報を表示する。現在のところ，(a) 品詞（大分類），(b) 語種，(c) 活用形，(d) 平均文長，(e) 文末の文体，(f) 品詞，(g) 語彙素，以上のデータ表示が可能である。(a)～(e)までは，円グラフとリストを表示し，(f)，(g)はリストのみ表示する。画面例は次の節に示す。

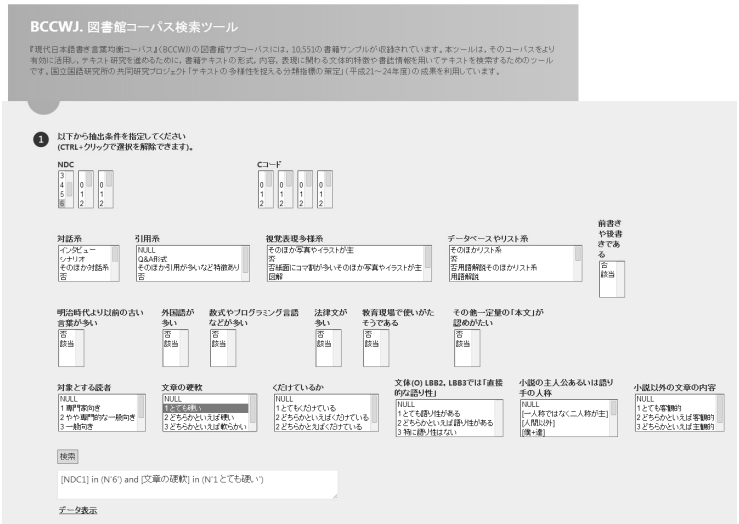


図 1 検索画面

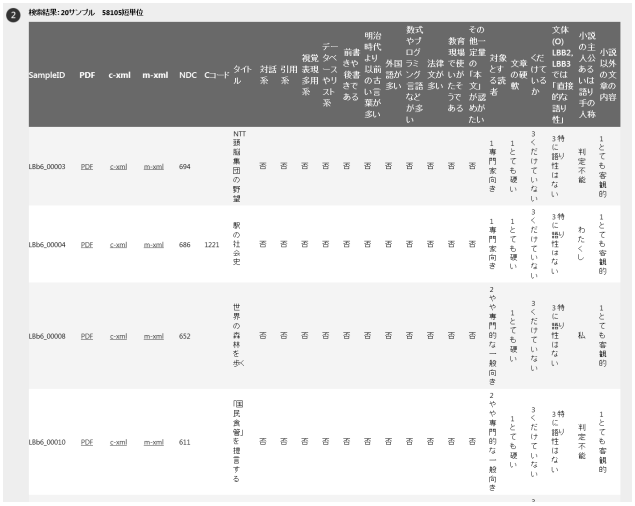


図 2 検索結果の表示画面

３．３　テキスト分析例「とても硬い」と「とても軟らかい」の場合

最後に，「図書館コーパス検索ツール」を用いたテキスト分析例として，「とても硬い」と「とても軟らかい」のテキストを比較する。図は，「データ表示画面」で得られるものである（ただし，印刷の都合上，円グラフは別途作成したものと差し替えている）。

(1) 該当テキストのサンプル数と短単位数

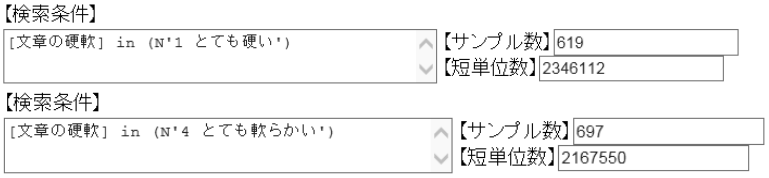
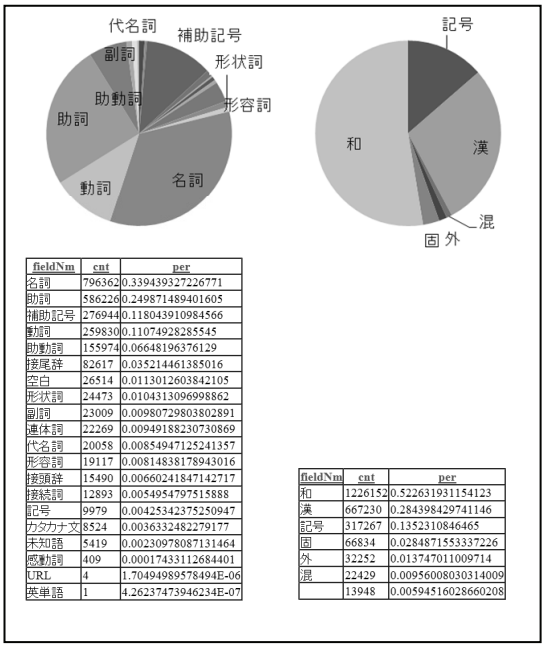


図 3 「とても硬い」「とても軟らかい」のサンプル数と短単位数

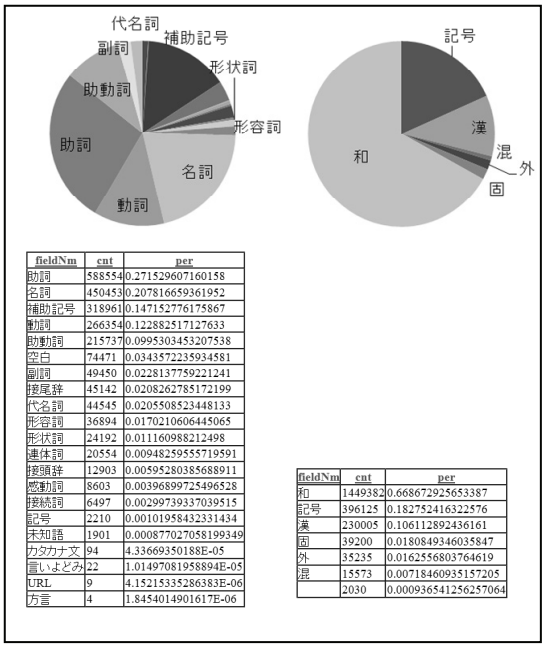
(2) 品詞（大分類），語種



fieldNm	cnt	per
名詞	796362	0.339439327226771
助詞	586226	0.249871489401605
補助記号	276944	0.118043910984566
動詞	259830	0.11074928285545
助動詞	155974	0.06648196376129
接尾辞	82617	0.035214461385016
空白	26514	0.011301260384 2105
形状詞	24473	0.0104313096998862
副詞	23009	0.00980729803802891
連体詞	22269	0.00949188230730869
代名詞	20058	0.00854947125241357
形容詞	19117	0.00814838178943016
接頭辞	15490	0.00660241847142717
接続詞	12893	0.0054954797515888
記号	9979	0.0042534237525094 7
カタカナ文	8524	0.003633248227917 7
未知語	5419	0.00230978087131464
感動詞	409	0.0001743311268440 1
URL	4	1.70494989578494E-06
英単語	1	4.26237473946234E-07

fieldNm	cnt	per
和	1226152	0.522631931154123
漢	667230	0.284398429741146
記号	317267	0.1352310846465
固	66834	0.0284871553337226
外	32252	0.013747011009714
混	22469	0.00956008030314009
	13948	0.00594516028660208

図 4 「とても硬い」品詞（大分類）と語種



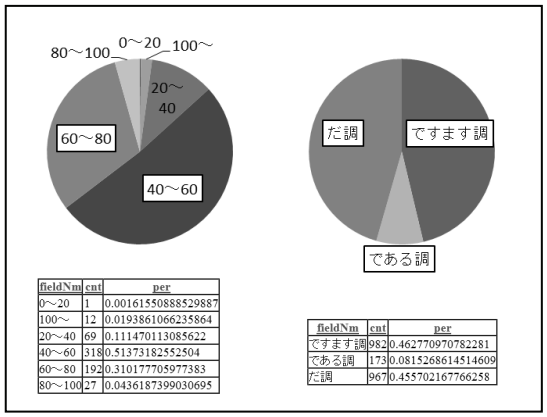
fieldNm	cnt	per
助詞	588554	0.271529607160158
名詞	450453	0.207816659361952
補助記号	318961	0.147152776175867
動詞	266354	0.122882517127633
助動詞	215737	0.0995303453207538
空白	74471	0.0343572235934581
副詞	49450	0.0228137759221241
接尾辞	45142	0.0208262785172199
代名詞	44545	0.0205508523448133
形容詞	36894	0.0170210606445065
形状詞	24192	0.011160988212498
連体詞	20554	0.00948259555719591
接頭辞	12903	0.00595280385688911
感動詞	8603	0.00396899725496528
接続詞	6497	0.00299739337039515
記号	2210	0.00101958432331434
未知語	1901	0.000877027058199349
カタカナ文	94	4.33669350188E-05
言いよどみ	22	1.01497081958894E-05
URL	9	4.15215335286383E-06
方言	4	1.8454014901617E-06

fieldNm	cnt	per
和	1449382	0.668672925653387
記号	396125	0.182752416322576
漢	230005	0.106112892436161
固	39200	0.0180849346035847
外	35235	0.0162556803764619
混	15573	0.00718460935157205
	2030	0.000936541256257064

図 5 「とても軟らかい」品詞（大分類）と語種

図 4，図 5 に示すように，このように，該当テキストに対して，BCCWJ に付与されている形態論情報を利用し，品詞（大分類）や語種の円グラフと一覧とを得ることができる。品詞（大分類）を比べると，「とても硬い」は名詞と助詞の二つが半分以上を占め，一方，「とても軟らかい」はその二つでは半分以下であり，代わりに，補助記号，助動詞の比率が高いことなどがわかる。語種を比べると，「とても硬い」での漢語比率，「とても軟らかい」での和語の比率の高さが顕著にみてとれる。

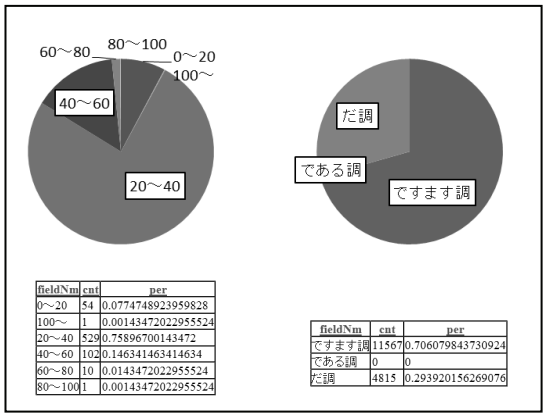
(3) 平均文長，文末の文体



fieldNm	cnt	per
0〜20	1	0.00161550888529887
100〜	12	0.0193861066235864
20〜40	69	0.111470113085622
40〜60	318	0.51373182552504
60〜80	192	0.310177705977383
80〜100	27	0.0436187399030695

fieldNm	cnt	per
ですます調	982	0.462770970782281
である調	173	0.0815268614514609
だ調	967	0.455702167766258

図 6 「とても硬い」平均文長と文末の文体



fieldNm	cnt	per
0〜20	54	0.0774748923959828
100〜	1	0.0014347202295552 4
20〜40	529	0.75896700143472
40〜60	102	0.146341463414634
60〜80	10	0.0143472022955524
80〜100	1	0.0014347202295552 4

fieldNm	cnt	per
ですます調	11567	0.706079843730924
である調	0	0
だ調	4815	0.293920156269076

図 7 「とても軟らかい」平均文長と文末の文体

平均文長は，サンプル全体の文字数を，BCCWJ に付与されている「sentence 開始タグ」の数で割って求めている。その結果，「とても硬い」の平均文長は長く，「とても軟らかい」の平均文長は短いという特徴がはっきりわかる。

BCCWJ では公開ファイル形式のひとつとして XML 形式がある。文体の文末は，会話部分に付与されている XML 形式の「quote タグ」を参照し，その部分を除外し，それ以外の文末に含まれる「です/ます。」「である。」「だ。」の数を合計数で割って求めている。その結果，「とても硬い」は「だ調」と「である調」が多く，「とても軟らかい」は「ですます調」が多い，と，はっきりとした違いがみてとれる。しかしながら，文体の文末の計数機能は現在，改良中である。「です/ます。」「である。」「だ。」の終止形以外の形で終わる場合や，常体「だ・である体」の文末として，動詞，形容詞，名詞などが文末にくる場合も数え，個別の数を出し，さらに詳細に分析できるようにする予定である。

(4) 品詞と語彙素

表 2 品詞一覧（抜粋）

	とても硬い				とても軟らかい		
	fieldNm	cnt	per		fieldNm	cnt	per
1	助詞-格助詞	389312	16.6%	1	助詞-格助詞	306269	14.2%
2	名詞-普通名詞-一般	364536	16.6%	2	名詞-普通名詞-一般	256923	11.9%
3	名詞-普通名詞-サ変可能	172170	7.3%	3	動詞	216737	10.0%
4	動詞	166974	6.6%	4	動詞-一般	142013	6.6%
5	動詞-非自立可能	146186	6.2%	5	補助記号-読点	135973	6.2%
6	補助記号-読点	122366	5.2%	6	動詞-非自立可能	124341	5.7%
7	動詞-一般	113644	4.8%	7	補助記号-句点	103699	4.8%
8	名詞-数詞	91859	3.9%	8	助動詞-接続助詞	99697	4.6%
9	助動詞-係助詞	83696	3.6%	9	助動詞-係助詞	88136	4.1%
10	助動詞-接続助詞	76109	3.3%	10	空白	74471	3.4%
11	補助記号-句点	60661	2.6%	11	名詞-普通名詞-サ変可能	54406	2.5%
12	接尾辞-名詞的-一般	56993	2.4%	12	副詞	49450	2.3%
13	名詞-普通名詞-副詞可能	47680	2.0%	13	名詞-普通名詞-副詞可能	46342	2.1%
14	補助記号-括弧開	37632	1.6%	14	代名詞	44545	2.1%
15	補助記号-括弧閉	37335	1.6%	15	接尾辞-名詞的-一般	36940	1.7%
16	名詞-普通名詞-助数詞可能	36908	1.6%	16	動詞-副詞的	33326	1.6%
17	空白	26614	1.1%	17	補助記号-括弧開	33070	1.6%
18	副詞	23009	1.0%	18	補助記号-括弧閉	33066	1.6%
19	連体詞	22269	0.9%	19	助詞-終助詞	31268	1.4%
20	動詞-副詞的	20963	0.9%	20	動詞-連体詞的	27667	1.3%
21	代名詞	20058	0.9%	21	名詞-数詞	23847	1.1%
46	助詞-終助詞	2965	0.1%	46	接尾辞-形容詞的	920	0.0%

表 2 は，「データ表示画面」から得られる品詞一覧をコピーし，編集しなおした表である。このように，ツールから得たデータを，別途編集し，利用することもできるようになっている。表 2 より，「とても硬い」では，「名詞-普通名詞-サ変可能」と「名詞-数詞」の多いことがわかる。また，「とても軟らかい」では，「副詞」「代名詞」「助詞-終助詞」の多いことがわかる。

「データ表示画面」には，頻度 1 に至るまでの全ての語彙素について，語彙素，語彙素読み，品詞，頻度，割合の一覧が示される。そこから，「名詞-普通名詞-サ変可能」と，「助詞-終助詞」を抽出したものが，次の表 3 である。両者に重なって出現する語には網掛けをして表示する。また，「助詞-終助詞」には頻度を添えて示す。表 3 をみると，「とても硬い」と「とても軟らかい」では重なりが少ないことがわかる。「とても硬い」の方に多く現れる語には，「研究，教育，存在・・・」といった学術的な語が多くあり，「とても軟らかい」の方に多く現れる語には，「料理，結婚，食事・・・」といった生活的な語が多くある。そして，終助詞の一覧をみると，「とても軟らかい」では多くの種類の終助詞が高頻度で出現していることがわかる。

表 3　語彙素一覧より得た「名詞-普通名詞-サ変可能」と「助詞-終助詞」の語

「とても硬い」名詞-普通名詞-サ変可能（上位 20 語）：
物，関係，研究，意味，教育，存在，生活，労働，活動，生産，利用，戦争，運動，規定，著作，裁判，決定，影響，支配，計画

「とても軟らかい」名詞-普通名詞-サ変可能（上位 20 語）：
物，話，仕事，心，一緒，電話，生活，関係，意味，料理，結婚，勉強，食事，説明，涙，存在，びっくり，意識，約束，旅

「とても硬い」助詞-終助詞（頻度 50 以上）：
か 2165，よ 213，な 150，ね 103，わ 63，の 58，ぞ 53

「とても軟らかい」助詞-終助詞（頻度 50 以上）：
か 7391，よ 7026，ね 5442，な 3730，の 2109，わ 1650，さ 1018，ぞ 608，い 514，ぜ 356，もの 356，かしら 323，や 120，ねん 119，け 117，じゃん 96，で 71

４．おわりに

Web アプリケーション「図書館コーパス検索ツール」の構築について報告した。本ツールを用いることで，特定の文体的特徴をもつテキストを検索し，形態論情報等を簡単に一覧することができるようになった。その一例として，「とても硬い」「とても軟らかい」と判定されたテキストが，どのようなテキストの集合となるのか，ツールから得られる，品詞，語種，平均文長，文末文体，品詞一覧，語彙素一覧を用い，比較分析を行った。そして，それらツールから得られる情報によって，両者の文体差を支える言語的特徴の違いをはっきり確認できることを示した。本ツールを活用することで，さまざまに付与した文体的特徴を支える言語的特徴が何であるかの分析を進めていくことができる。

本ツールは，インターネットが利用できる環境と標準的なブラウザがあれば，特別なソフトをインストールすることなく利用することができる。現在は，研究所内のネットワークからのみアクセス可能であるが，広く利用してもらえる方法を検討していきたい。

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」，「コーパスアノテーションの基礎研究」による成果を含みます。また，文部科学省科学研究費補助金基盤研究(C)「コーパス分析に基づく辞書の位相情報の精緻化」（課題番号:23520572）の助成を得たものです。

文　献

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

Wakako Kashino and Manabu Okumura(2010) An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese, *Proc. of PACLIC24*, pp.433-438.

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

柏野和佳子，立花幸子，保田祥，飯田龍，丸山岳彦，奥村学，佐藤理史，徳永健伸，大塚裕子，佐渡島紗織，椿本弥生，沼田寛(2012b)「書籍テキストへの文体情報付与の試み」『第 2 回コーパス日本語学ワークショップ』予稿集,pp.155-164.

附録

文体的特徴を表す分類指標の付与サンプル例

(1)専門度：1 専門家向き（LBi4_00021『がんと遺伝子』）

> Ｅ２Ｆ以外のＲＢ結合タンパク質としては，転写因子ＲＡＸ，Ｔ細胞が活性化するときに誘導されるＩＬ－２，ＧＭ－ＣＳＦ，ＨＩＶ－２などの転写を活性化する転写因子Ｅｌ Ｆ－１や先に述べた細胞周期を制御するサイクリンＤなどがある．おもしろいことに，Ｅｌ Ｆ－１やサイクリンＤのＲＢ結合ドメインにはｌａｒｇｅＴ抗原やＥ１Ａタンパク質と同じようにＬＸＣＸＥというアミノ酸配列が存在する．また，ＲＢタンパク質は骨格筋分化を支配する重要な遺伝子群ＭｙｏＤファミリー（ＭｙｏＤ，ｍｙｏｇｅｎｉｎ，ＭＲＦ４，ｍｙｆ－５）の産物とも複合体を形成し筋分化にも関与しているらしい．

(2)専門度：4 中高生向き（LBf9_00090『超魔炎獄変』）

> 白く薄い空気のヴェールが、漂うように揺らめいている。
> シャ…ァァン、シャラ…ァァン…。
> 闇を抜け、霧の中を渡る金属の響き。それは魔を覇する浄化の音。
> 響きに道を開けるかのようにすぅ…っと霧が左右に分かれた。
> それは。
> 霧の中にたたずむそれは。
> 闇。
> …いや。闇ではない。

(3)客観度：1 とても客観的（LBo3_00158『行政法要論』）

> たとえば委員会の開催が「急施を要する場合」にあたるかどうかとか、公衆浴場の施設が「公衆衛生上不適切」かどうかは通常人の経験則によって十分判断できる事柄であるから、羈束裁量であって裁判所の終局的な判断に服すべきものとする。これに対し、外国人の在留期間の更新を適当と認めるに足る相当の理由があるかどうかは、出入国管理行政の責任者である法務大臣の政治的判断に委ねらるべきであり、また、原子炉の安全性の認定は高度の科学的専門技術的知見に基づく総合的判断であるから、行政庁の便宜裁量事項であり、その当否は裁判所の審理・判断にはなじまないとする（最判昭和五三年一〇月四日民集三二巻七号一二二三頁、同平成四年一〇月二九日民集四六巻七号一一七四頁）。

(4)客観度：4 とても主観的（LBo3_00132『教師をめざす若者たち』

どんなに上手な言葉を使っても、思っていないことを発すれば、子供に伝わらない。どんなに下手な言葉でも、心から伝えたいという愛情があれば、伝わるものであるということを信じることができました。この実感は日本でも通じる「教育の原則」であると思いました。

二日目、子供たちと綿花摘みを一緒にしました。敦煌の子供たちの手は「仕事をしている手」でした。

(5)硬度：1 とても硬い（LBi3_00033『現代法社会学入門』）

取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。

(6)硬度：4 とても軟らかい（LBa4_00010『恐竜の世界をたずねて』）

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへしたへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三畳紀のはじめにいた、「テコドント」（図８６）という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

(7) くだけ度：1 とてもくだけている（LBf9_00067『男はオイ！女はハイ…』）

最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをば何となく眺めているうちに、どうしても欲しくなった商品があった。

よし、こいつひとついってやれとばかりすぐ電話にとびついた。

「ハイ、こちら―です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。

「商品番号をおっしゃって下さい」

といわれて答える。

さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

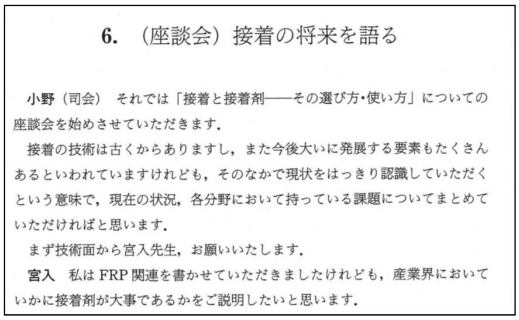
(8) 語りかけ性度：1 とても語りかけ性がある（LBt1_00013『5 分間集中力トレーニング』）

精神的に疲れていると、「ああなったら、どうしよう」「こうなったら、どうしよう」と常に不安だらけになります。

動物病院にいらっしゃる飼い主さんには、過剰な不安を抱えている人や心配性の人がとても多い。実はそれがペットの病気をさらに悪化させることになっていますが、そういう認識をお持ちの飼い主さんは、あまりいません。詳しい説明は避けますが、不安や心配性を放っておくと、動物の具合が悪くなり、当然それが自分にも返ってきます。

それでは、どうすればいいのでしょうか。

文体判断が単純にいかない特徴をもつものの分類指標の付与サンプル例



6. （座談会）接着の将来を語る

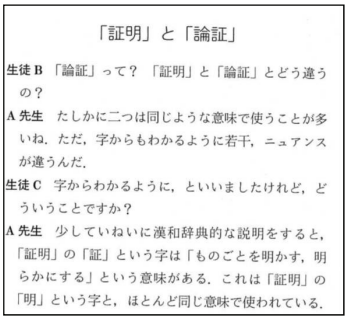
小野（司会） それでは「接着と接着剤――その選び方・使い方」についての座談会を始めさせていただきます．

接着の技術は古くからありますし，また今後大いに発展する要素もたくさんあるといわれていますけれども，そのなかで現状をはっきり認識していただくという意味で，現在の状況，各分野において持っている課題についてまとめていただければと思います．

まず技術面から宮入先生，お願いいたします．

宮入　私は FRP 関連を書かせていただきましたけれども，産業界においていかに接着剤が大事であるかをご説明したいと思います．

図 1 座談：LBd5_00012『接着と接着剤』



「証明」と「論証」

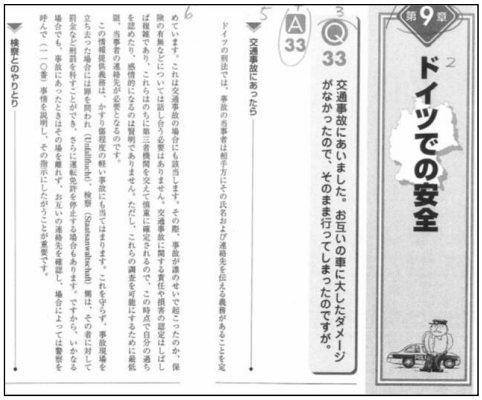
生徒 B 「論証」って？ 「証明」と「論証」とどう違うの？

A 先生　たしかに二つは同じような意味で使うことが多いね．ただ，字からもわかるように若干，ニュアンスが違うんだ．

生徒 C 字からわかるように，といいましたけれど，どういうことですか？

A 先生　少していねいに漢和辞典的な説明をすると，「証明」の「証」という字は「ものごとを明かす，明らかにする」という意味がある．これは「証明」の「明」という字と，ほとんど同じ意味で使われている．

図 2 対話：LBpn_00038『なぜ数学を学ぶのか』



第9章

ドイツでの安全

Q33 交通事故にあいました。お互いの車に大したダメージがなかったので、そのまま行ってしまったのですが

A33

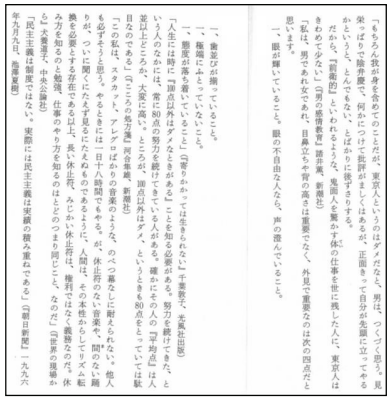
交通事故にあったら

ドイツの刑法では、事故の当事者は相手方にその氏名および連絡先を伝える義務があることを定めています。これは交通事故の場合に限ります。その際、事故が誰のせいで起こったのか、保険の有無などについて話し合う必要はありません。交通事故に関する責任や損害の認定はしばしば複雑であり、これらのために第三者機関を交えて、交通事故の認定はしばしば数日から数週間を要することもあります。ただし、これらの調査を可能にするために最低限、当事者の連絡先が必要となるのです。

この情報提供義務は、かすり傷程度の軽い事故にも当てはまります。これを守らず、事故現場を立ち去った場合には罰金が科されます。ですから、いかなる場合でも、事故にあったときはその場を離れず、お互いの連絡先を確認し、場合によっては警察を呼んで（一一〇番）事情を説明し、その指示にしたがうことが重要です。

検察とのやりとり

図 3 Q&A：LBr3_00037『ドイツ暮らしの法律 Q&A』



「もちろん我が身を含めてのことだが、東京人というのはダメだなと、男は、つくづく思う。見栄っぱりで臆病で、何かにつけて批評がましくあるが、正面きって自分が先頭に立ってやるかというと、とんでもない、とばかりに後ずさりする。

だから、「臆病的」（「男の感情表現」堺井興、新潮社）きわめて少ない」（「男の感情表現」堺井興、新潮社）

「私は、男であれ女であれ、目鼻立ちや背の高さは重要でなく、外見で重要なのは次の四点だと思います。

一、眼が輝いていること。眼の不自由な人なら、声の澄んでいること。

一、歯並びが揃っていること。

一、姿勢に気をつけていること。

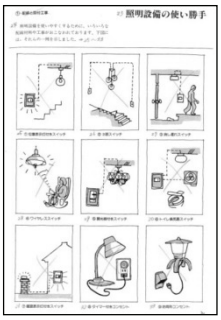
一、服装がきちんとしていること。

「人生には時に「頭以外はダメなとき」があり、努力を続けてきた、と言い切れる人がいる。努力を続けてきた人が必ずしも「平均点」は人並み以上のところが「大変に高い」ところが「頭点以外」は人並み以下のところがある。「頭点以外」は人並みに高い、というときも80点をとっていては駄目なのである」（「こころの処方箋」河合隼雄、新潮社）

「このほか、スタカット、アレグロばかりの音楽のような、のべつ間なしに耐えられない、他人を必ずうっとうしがり、やるときには一日十八時間でもやる、が、休止符のない音楽や、間のない踊りがいつまでたっても見るにたえないものであるように、人間は、その本性からしてリズム転換を必要とする存在である以上、長い休止符、みじかい休止符、機能ではなく義務なのだ、休止符の場合の力を知るのは、とどのつまり同じこと、なのだ」（「世界の現場から」大宅壮一、中央公論社）

「民主主義は制度ではない。実際には民主主義は実績の積み重ねである」（「朝日新聞」一九九六年九月九日、夕刊）

図 4　引用：LBl8_00017『試行錯誤の文章教室』



照明設備の使い勝手

図 5 イラスト：LBb5_00011『絵とき インテリアライティングの技法早わかり』



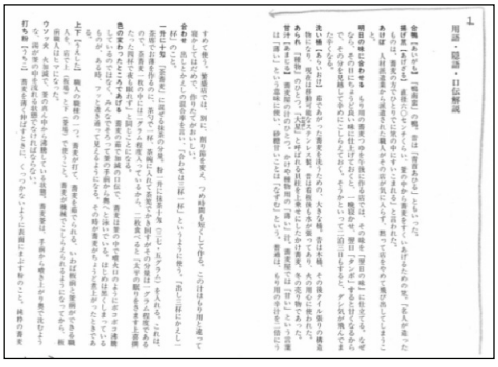
今、気になるタウン……品川

品川グランドコモンズってどんなところ？

素材で勝負の実力派レストランが集結！

アロマクラシコ

図 6　コマ割り：LBsn_00022『東京で遊ぼ』



1 用語・略語・日伝解説

図 7　辞書：LBp6_00009『蕎麦屋のしきたり』



その他の仲間の魚たち

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

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

キイロサンゴハゼ

Cryloden xanourus

分布/西部太平洋

全長 3cm
水温 24～27℃
餌 普通
水槽 40cm以上
混泳 同大の温和な魚と
入荷 普通
価格 11,000～2,000円
エサ 冷凍エサ各種、フレークフード、動植状乾燥餌、アサリのミンチ、クリル、魚肉、魚卵

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

図 8　カタログ：LBj6_00025『熱帯魚・水草カタログ』

独話音声と対話音声の発話末の F0 変化

石本 祐一（国立国語研究所 研究情報資料センター）†
小磯 花絵（国立国語研究所 理論・構造研究系）

F0 Changes at the End of Utterance in Monolog and Dialog

Yuichi Ishimoto (Center for Research Resources, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

1. はじめに

発話の韻律的特徴のひとつとして，発話末において基本周波数（F0）が急激に低下する final lowering 現象が生じることが指摘されている（Pierrehumbert and Beckman (1988)）。小磯・石本 (2012) は，『日本語話し言葉コーパス（*Corpus of Spontaneous Japanese*:以下 CSJ）』（前川 (2004)）のコアに収録されている独話音声を対象に，イントネーション句を単位として発話中の F0 の大局的な変動を調べたところ，発話冒頭から発話末にかけて F0 最大値・最小値がともに下降する傾向にあること，この下降傾向は発話中の強い統語境界でリセットされること，発話末では発話全体にわたる F0 下降とは異なる final lowering に相当するような著しい F0 下降が見られること，このような著しい下降は発話中の強い統語境界では観察されないことを明らかにした。また前川 (2013) は，同じく CSJ のコア（全体，ただしその大半は独話）を対象に，1〜5 つ連続する有核アクセント句からなる連鎖に限定して final lowering の生起領域を詳細に調べ，統語的境界にあたる最終アクセント句全体にわたって final lowering が生じることを明らかにした。

一方，石本・小磯 (2013) では，CSJ の対話音声を対象に石本・小磯 (2012) と同種の方法で発話中の F0 変動を調べ，発話末での F0 変動の程度はそれ以前の F0 変動と同程度であること，また発話末の F0 最小値は発話中の強い統語境界とほぼ同じであり独話ほど下がらないことを明らかにした。この結果は，独話音声とは異なり対話音声では，発話末において final lowering が生じない可能性のあることを示唆する。

そこで本稿では，より統制した条件において CSJ の独話と対話の発話末の F0 を比較すると同時に，CSJ の対話（インタビュー・課題指向対話）とは異なるタイプの雑談会話をも合わせて比較することによって，対話における final lowering の有無について改めて検討する（3 節）。また，仮に対話において final lowering が観察されない（されづらい）として，その背景には独話と対話間の発話の計画性や自発性の程度の違いが関わりうると考え，final lowering との関係について検討する（4 節）。

† yishi@ninjal.ac.jp

2. データ

2.1 発話単位の認定

2.1.1 CSJ

CSJ 第 3 刷に基づき作成された RDB（小磯ほか (2012)）を用い，独話音声として，「コア」と呼ばれるデータ範囲から学会における口頭発表（以下「学会講演」）約 19 時間と一般話者による主に個人的な内容に関するスピーチ（以下「模擬講演」）約 20 時間を分析対象とした。また，対話音声としてコアデータ範囲中の 18 対話から，インタビュー形式の対話と課題指向対話のうち後述の韻律情報が付されたインタビュイーの発話[*1]約 220 分を選択した。

発話単位の認定にあたっては，CSJ に付与されている節単位情報（丸山ほか (2006)）を利用した。節単位は原則「節（clause）」の境界によって得られる文法的・意味的なまとまりを持った単位であり，節境界の構造的な切れ目の大きさの観点から以下の 3 つに分類される。

絶対境界　「〜です」「〜ます」などのいわゆる文末に相当する節境界[*2]。

強境界　「〜けれど」「〜が」などの後続の節に対する切れ目の度合が強い（従属度の低い）節境界。

弱境界　「〜から」「〜で」などの後続の節に対する切れ目の度合が弱い（従属度の高い）節境界。

本研究では，絶対境界で区切られる区間を発話に相当する単位として扱う。

2.1.2 千葉大 3 人会話コーパス（千葉大コーパス）

CSJ の対話とは異なる種類のデータとして千葉大 3 人会話コーパス（Den and Enomoto (2007)）を取り上げる。これは，親近性のある 3 人が定められたテーマのもとで自由に雑談を行うという内容である。分析対象として 12 会話約 120 分を用いた。

千葉大 3 人会話コーパスには，転記テキストや発話単位，形態論情報，韻律情報などの種々のアノテーションが付与されているが，節単位情報は付与されていない。ただし，節単位情報に類似した単位として，統語・語用論の観点から認定される長い単位 (LUU: Long Utterance Unit)（Den et al. (2010)）が提供されている。そこで本研究では，LUU および形態論情報・韻律情報を用い，LUU 末の品詞において活用形が終止形であるものを絶対境界，出現形が「けど」「けれど」「が」で接続助詞であるものを強境界と暫定的に認定することとした。これは，2.1.1 節に記した CSJ の絶対境界・強境界のサブセットに相当する。

2.2 F0 特徴量

本研究では節単位境界直前のアクセント句の F0 最小値を用いる。具体的には，コーパスに付与されているラベリングスキーム `X-JToBI`（五十嵐ほか (2006)）に基づく韻律情報から節末のアクセント句の下降音調（`L%`）の F0 値を求めた[*3]（図 1）。分析において性差・個人差の影響を小さくするために，F0 値は談話・話者ごとの平均 F0・標準偏差によって Z スコアへ変換

[*1] 課題指向対話の場合はインタビュー対話でインタビュイーの役割だった話者の発話。

[*2] 明示的な文末表現が置かれるもののほか，「と文末」や「体言止め」なども含む。

[*3] 千葉大 3 人会話コーパスでは X-JToBI の簡易版に準拠して韻律情報が付与されている。

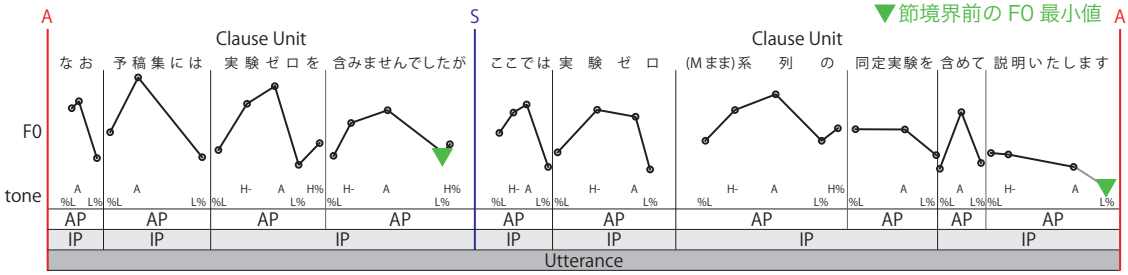


図 1　節単位末の F0 特徴量

表 1　分析 1 のデータ数

	CSJ 独話	CSJ 対話	千葉大コーパス
強境界 (S)	847	27	46
絶対境界 (A)	1369	114	295

し標準化を行っている。

3.　分析 1: 独話音声と対話音声の発話末の F0

3.1　方法

CSJ の独話と対話，および千葉大 3 人会話コーパスの雑談会話に対し，節単位絶対境界 (A) および強境界 (S) の直前の F0 の値を調べた。2.1.1 節の発話単位の定義から，絶対境界直前を発話末，強境界直前を発話中の F0 とみなす。同様の分析は石本・小磯 (2013) でも行っているが，そこでは下降調，上昇調，上昇下降調などの句末音調の違いについてあまり考慮しておらず，様々なモダリティを持つ発話が混在していたため，今回は句末音調が下降調である音声のみを分析した。また，相づちや応答などの極端に短い発話を除くため，2 つ以上のアクセント句を含む発話を対象とした。分析に用いた CSJ および千葉大 3 人会話コーパスのデータ数を表 1 に示す。

3.2　結果

CSJ の独話音声と対話音声における強境界および絶対境界直前の F0 最小値を図 2 に示す。独話に対する t 検定の結果，平均の差は有意（$t(1954.10)=21.82$, $p<0.001$）で発話末のほうがより低い F0 になっている。仮に対話においても発話末に（強境界よりも相対的に強い）final lowering が見られるならば，独話の結果と同様，強境界よりも絶対境界の方が F0 値は低くなることが予想される。しかし対話については t 検定の結果，平均の差は有意ではなかった（$t(66.30)=0.01$, $p=0.99$）。

対話音声は独話音声に比べて極端にデータ数が少なく，話者数も大きく異なるため，話者による違いが影響を与えている可能性がある。そこで，CSJ の独話と対話両方に音声が収録されている話者 1 名（Speaker ID: 514）の F0 について調べた。結果を図 3 に示す。データ数は独話音声の強境界が 15・絶対境界が 29，対話音声の強境界 10・絶対境界 31 であり，総数は少ないものの独話と対話でほぼ同じである。t 検定の結果，独話の F0 の平均は強境界と絶対境界で有

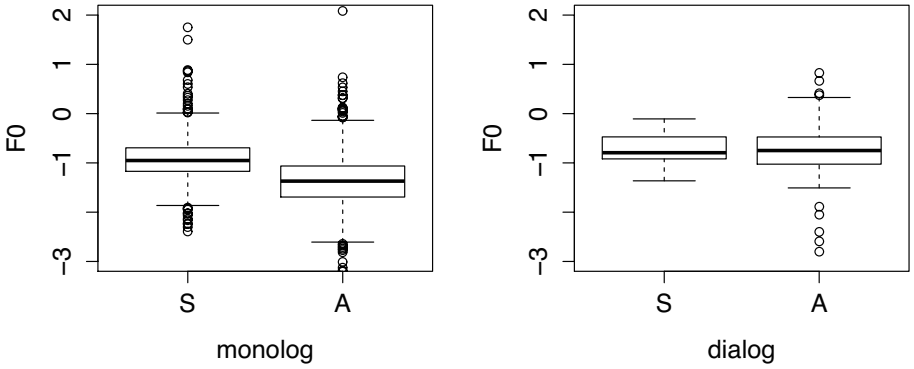


図 2　CSJ の強境界 (S) と絶対境界 (A) における F0 最小値

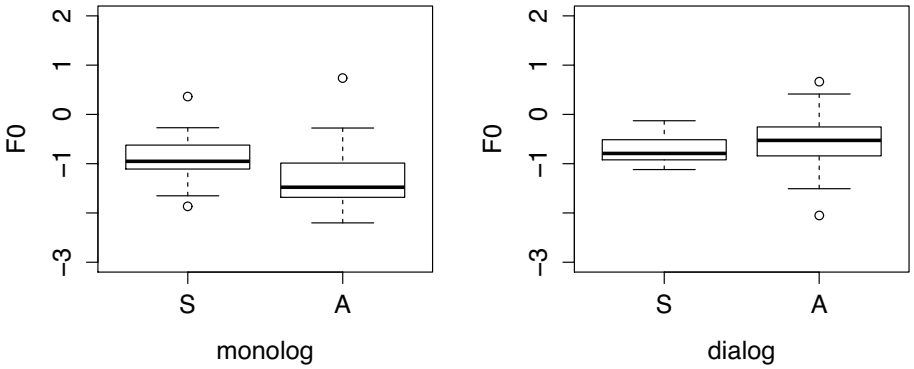


図 3　CSJ の強境界 (S) と絶対境界 (A) における F0 最小値（話者 1 名）

意差があり（t(19.12)=2.18, p<0.05），対話の F0 の平均には有意差はなかった（t(29.95)=-0.90, p=0.37）。図 3 からもわかるように同一話者においても対話音声では強境界と絶対境界で F0 に差がみられない。なお，独話と対話の両方に音声が収録されている他の 5 名の話者についてもほぼ同様の結果となった。よって，話者個人の発声様式による違いではなく，独話と対話全般における傾向であると考えられる。

次に，千葉大 3 人会話コーパスの強境界前と絶対境界前の F0 最小値を図 4 に示す。

CSJ の対話（インタビュー・課題指向対話）と同様，雑談会話においても，やはり強境界と絶対境界で F0 に差は見られないことが分かる[*4]。t 検定の結果，強境界と絶対境界の F0 平均に有意差はなかった（t(77.33)=-1.20, p=0.23）。以上の結果から，インタビューや雑談などの

[*4] 2.1.2 節に記した通り，千葉大 3 人会話コーパスの絶対境界・強境界は，定義上 CSJ のサブセットになっており，完全には一致しない。そこで，CSJ の独話・対話についても同じ基準で絶対境界・強境界のサブセットを作成して分析を行ったところ，図 2 と同様の結果を得た。

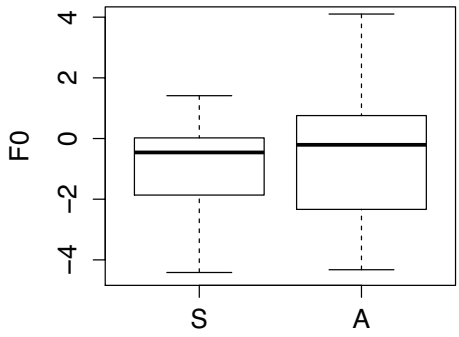


図 4　千葉 3 人会話コーパスの強境界 (S) と絶対境界 (A) における F0 最小値

対話・会話では，独話で見られるような final lowering に相当する発話末での急激な F0 下降は出現しない（出現しづらい）可能性が高い。

4. 分析 2: 発話の自発性と節単位末の F0 の関係

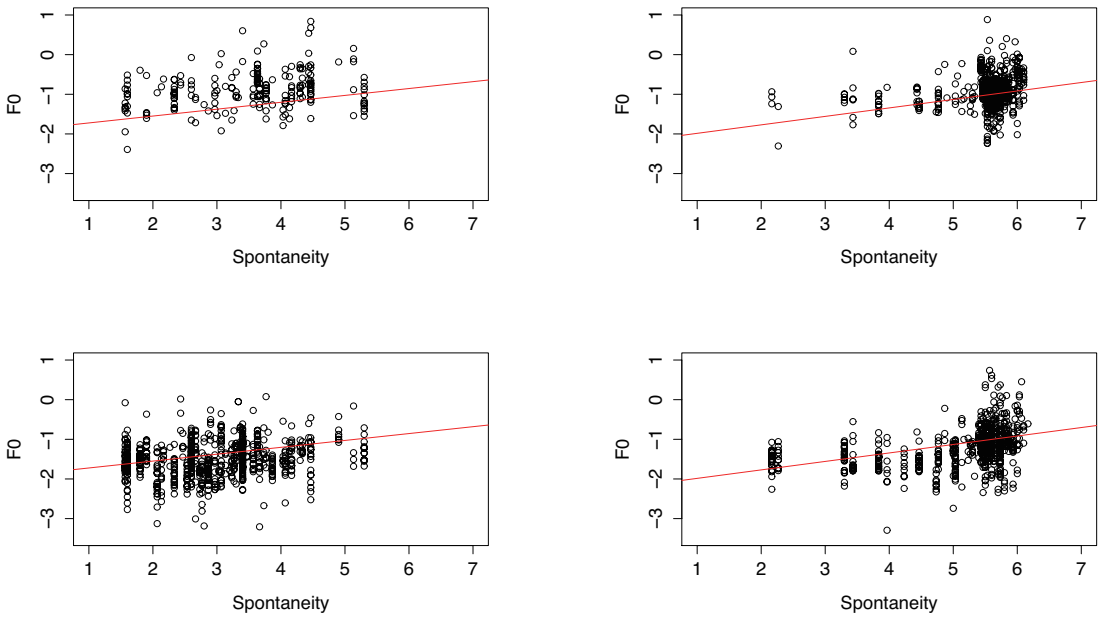
4.1 方法

対話において final lowering が観察されない（されづらい）要因として，独話と対話間の発話の計画性や自発性の程度の違いが考えられる。CSJ における独話音声（の大半）は原稿の読み上げではないが，講演という性質上，あらかじめ用意した内容に基づいて発声することから，対話と比べて相対的に発話の事前の計画性が高く自発性が低い傾向にある。こうした違いが発話末を韻律的に特徴付けるという final lowering の実現の有無に関わっている可能性が考えられる。仮にそうであるとするならば，同じ独話音声であっても，自発性の高い発話と自発性の低い発話では final lowering の出現傾向が異なることが予想される。そこで本節では，発話の自発性の程度と final lowering との関係を調べる。

CSJ コアの独話データには様々な種類の印象評定値が付与されているが，そのうち集合評定データと呼ばれる印象評定の中に「原稿を読み上げている—その場で考えて話している」という話し方の印象に関する評定項目がある（籠宮ほか (2004)）。そこで，この項目を自発性の程度を表す指標として用い，評定値が付与されている CSJ コアの独話データを対象に，自発性の度合と節単位絶対境界 (A) および強境界 (S) の直前の F0 の値との対応関係について調査した。印象評定は「原稿を読み上げている (1)」から「その場で考えて話している (7)」の 7 段階で行われており，評定値が大きいほど自発性が高いと判断されたことを意味する。また，20 名の評定者が「講演の冒頭」「講演の中盤」「講演の終盤」から 1 分前後の音声を聴いてそれぞれ評定を行っており，各講演に対し 60 個の評定値がつけられているため，講演ごとに評定値の平均値を求めて自発性の程度とした。

4.2 結果

CSJ の独話における自発性の程度と強境界および絶対境界直前の F0 最小値の対応関係を図 5 に示す。図 5(a) は学会講演，図 5(b) は模擬講演であり，それぞれ上段が強境界，下段が絶対境界のデータに対応している。図中の直線は学会講演，模擬講演全体での線形回帰直線で



(a) 学会講演　(b) 模擬講演

図 5　自発性の印象評定と F0 最小値（上: 強境界，下: 絶対境界）

表 2　一般化線形混合モデルのパラメータ（応答変数: F0 最小値）

	Estimate	Std. Error	自由度	t 値		
(Intercept)	-1.687	0.107	146.9	-15.723	p<0.001	***
自発性の程度	0.182	0.030	133.8	6.047	p<0.001	***
講演（模擬）	-0.157	0.081	135.4	-1.939	p=0.055	.
節境界（絶対）	-0.310	0.021	2159.2	-14.749	p<0.001	***

あり，上段と下段では同じ傾き・切片であることに留意されたい。学会講演と模擬講演で自発性の程度を比較すると，模擬講演ではその多くが自発性の高い講演であるのに対し，学会講演では相対的に自発性の低い講演が多くなっており，自発性の観点から学会講演と模擬講演では異なる性質にあることがわかる。また，そのような学会講演と模擬講演の性質の違いにもかかわらず，両者とも自発性が低ければ F0 が低く，自発性が高ければ F0 が高くなる傾向にある。

これらの傾向を確認するため、強境界・絶対境界直前の F0 最小値を応答変数とし，自発性の程度・節境界ラベル（強・絶対）・講演タイプ（学会・模擬）を説明変数として話者による影響を誤差項として加えた一般化線形混合モデルを構築した。なお，節境界は強境界 (0)・絶対境界 (1)，講演タイプは学会講演 (0)・模擬講演 (1) とダミー変数化している。モデルから得られた説明変数のパラメータを表 2 に示す。3 つの説明変数のうち、自発性の程度および節境界ラベル（強・絶対）が有意であった。すなわち，図 5 とあわせて考えると，自発性が高くなるほど F0 も高くなり，絶対境界の F0 は強境界よりも低くなると言える。一方，講演の種類が

F0 値に与える影響は有意ではなく，学会講演・模擬講演といった講演のタイプに限らず上記の傾向がみられることがわかる。

以上のことから，独話であっても自発性が高い発話では発話末の F0 は低くならず final lowering が観察されにくい傾向にあると言える。

5. 議論

3 節の分析から，インタビューや雑談などの対話・会話では，独話で見られるような final lowering に相当する発話末での急激な F0 下降は出現しない（出現しづらい）傾向にあることが分かった。また 4 節の分析から，同じ独話であっても，自発性の低い発話では発話末の F0 はかなり低くなるのに対し，自発性の高い発話では発話末の F0 は低くならず final lowering が観察されにくい傾向にあることが分かった。対話は基本的に事前に発話内容が決まっておらず総じて自発性が高いと考えられることから，独話・対話を通じて final lowering の生起は自発性の程度と関係していると言える。

では何故，自発性の高い独話や対話では絶対境界における F0 の低下がはっきりと現れないのだろうか。一つの可能性として，自発性の程度に伴う発話生成の認知的負荷が影響していると考えられる。自発性の低い発話では発話前に何をどのように話すかといった発話計画がある程度なされている。そのため，来たる発話末を明示するために F0 を急激に低下させるといった調整をするだけの認知的な余裕があると考えられる。一方，自発性の高い発話では，発話内容や表現を考えながら発話をするため，発話生成の認知的負荷が高く，発話末で final lowering のような韻律調整を行うだけの余裕が十分にないことが予想される。

CSJ は自発性の高い発話を収録したコーパスであるが，独話，特に学会講演ではあらかじめ発表内容が決まっていることもあり，練習を積んだ上で原稿を読み上げるように発話しているデータも含まれてる。Umeda (1982) は，F0 下降はスタイルに依存しており final lowering は朗読音声に限られる可能性を示唆しており，本研究の結果と整合的である。ただし今回の結果は，朗読か否かという二分的なスタイルに依存するようなものではなく，その背後に自発性の程度に伴う発話生成の認知的負荷が影響していることを示唆している。

また，仮に final lowering の役割が発話の終了を表示することであるならば，学会講演のように聞き手に対して談話の構造や切れ目を明確に伝えることが指向されるようなスタイルのスピーチにおいて，final lowering による発話末の表示が好まれる可能性も十分にある。つまり，談話構造を明示化するような発話の仕方を好むスタイル，またそれを可能とする自発性の低さに伴う認知的な余裕，こうしたことが合わさって学会講演では発話末の final lowering が特徴的に観察された，対話の場合はその逆の状況だったため final lowering があまり観察されなかった，という解釈である。この可能性については，今後，異なるスタイルの独話や対話を分析・比較することによって検討したい。

参 考 文 献

Den, Yasuharu, and Mika Enomoto (2007). “A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation.” Toyoaki Nishida (Ed.), *Conversational informatics: An engineering approach*. Hoboken, NJ: John Wiley & Sons. pp. 307–330.

Den, Yasuharu, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida (2010). “Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme.” *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, pp. 2103–2110. Valletta, Malta.

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

石本祐一・小磯花絵 (2012).「日本語話し言葉コーパスを用いた統語境界におけるイントネーション句変動の分析」　第 2 回コーパス日本語学ワークショップ予稿集, pp. 239–246.

石本祐一・小磯花絵 (2013).「日本語話し言葉コーパスを用いた対話音声のイントネーション句の分析」第 4 回コーパス日本語学ワークショップ予稿集, pp. 159–166.

籠宮隆之・山住賢司・槙洋一 (2004).「印象評定データの概要」　『日本語話し言葉コーパス『DVD 付属マニュアル』』.

小磯花絵・石本祐一 (2012).「日本語話し言葉コーパスを用いた「発話」の韻律的特徴の分析–イントネーション句を切り口として–」　第 1 回コーパス日本語学ワークショップ予稿集, pp. 167–176.

小磯花絵・伝康晴・前川喜久雄 (2012).「『日本語話し言葉コーパス』RDB の構築」　第 1 回コーパス日本語学ワークショップ予稿集, pp. 393–400.

前川喜久雄 (2004).「『日本語話し言葉コーパス』の概要」　日本語科学, 15, pp. 111–133.

前川喜久雄 (2013).「日本語自発音声における final lowering の生起領域」　第 27 回音声学会全国大会予稿集, pp. 47–52.

丸山岳彦・高梨克也・内元清貴 (2006).「節単位情報」　『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』　pp. 255–322.

Pierrehumbert, Janet B., and Mary E. Beckman (1988). *Japanese tone structure*. Cambridge: MIT Press.

Umeda, Noriko (1982). “F0 declination is situation dependent.” *Journal of Phonetics*, 10, pp. 279–290.

※ 本研究は JSPS 科研費 24700109, 25370505 および萌芽・発掘型共同研究「会話の韻律機能に関する実証的研究」（リーダー：小磯花絵）による成果である。

口頭発表（2）

3月7日（金）　10:00～12:00

Web を母集団とした超大規模コーパスの開発 —収集と組織化—

浅原　正幸（国立国語研究所コーパス開発センター）
今田　水穂（国立国語研究所コーパス開発センター）
保田　祥（国立国語研究所コーパス開発センター）
小西　光（国立国語研究所コーパス開発センター）
前川　喜久雄（国立国語研究所言語資源研究系・コーパス開発センター）

Early Results of Page Collection and Linguistics Annotation in Ultra Large Scale Web Corpus Construction

Masayuki ASAHARA (Center for Corpus Development, NINJAL)
Mizuho IMADA (Center for Corpus Development, NINJAL)
Sachi YASUDA (Center for Corpus Development, NINJAL)
Hikari KONISHI(Center for Corpus Development, NINJAL)
Kikuo MAEKAWA(Dept. of Corpus Studies, Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では 2006～2010 年度の期間に 1 億語規模の書き言葉コーパス『現代日本語書き言葉均衡コーパス』（以下“BCCWJ”）（前川 2007, 前川・山崎 2008）を構築し，2011 年より一般公開した。BCCWJ は種々の母集団に沿った無作為抽出を実施することによって，高度な均衡性・代表性を備えた均衡コーパスとなっている。しかし，その規模は，現代のコーパス言語学の趨勢からすれば十分とはいいがたく，生起頻度の低い言語現象の被覆に問題がある。そのためより大規模な日本語コーパスの構築が望まれている。この問題を解消するため,国立国語研究所コーパス開発センターでは 2011 年度から 6 年の期間で，Web を母集団とした 100 億語規模の超大規模コーパスを構築する計画に着手した。

現在，6 年のプロジェクトの 3 年が過ぎ，収集と組織化について進捗している。またプロジェクトの初期において収集と組織化のための環境構築に工数をかけた。現時点の環境整備・収集・組織化についてはある程度目途がつきつつある。2012 年 10 月より本収集を開始し 2013 年 9 月まで 1 年間収集したデータについて組織化が進んでおり，基礎統計量が得られている。

本稿ではプロジェクトの進捗状況を過去 1 年間に収集されたデータの基礎統計量とともに示し，本言語資源を利用した言語研究の可能性について論じる。２節では収集の進捗状況について述べる。３節では組織化の進捗状況について述べる。４節では本言語資源とグーグルによる「Web 日本語 N グラム第 1 版」（1）との頻度上位 10 語の比較を定性的に分析する。５節で言語資源の大規模化による理論的・応用的研究の可能性について論じ，６節にまとめと今後の予定について述べる。

２．収集の進捗状況

本節では収集の経過と現在までに得られている統計量について示す。

2012 年 6 月に収集のための環境整備が完了し，環境整備に関して所内のセキュリティ委員会に諮り，クローラ運用の許諾を得た。許諾後すぐにクローラの試験運用 1 か月前よりクローラに関する情報提供･問い合わせ窓口としての Web ページ/メールアドレス/電話を設置した。1 か月の周知期間を経て，2012 年 7～8 月より試験収集を開始し，Heritrix（2）の各種パラメータを調整した。9 月に最終試験収集となる 1000 万 URL 規模の収集を行い，

その際に得られた収集速度の情報からクローラ運用方針を年間 4 回収集，1 回あたりの収集量は 1 億 URL 規模で 3 か月間に設定した。URL は 1 年間通して 4 回収集し，季節に偏らないように配慮する一方，URL の更新頻度およびリンク先の情報に基づき，収集する URL を 1 年毎に変更する方針にした。2012 年 10 月より 3 カ月ごとに収集を行い，2013 年 12 月末現在, 5 回目の収集の最中で次の 1 年間に収集すべき URL サンプルを決定する状況にある。

以下収集の統計量について示す。全体については 2012 年第 4 四半期（2012-4Q）～2013 年第 3 四半期（2013-3Q）についての統計量を，各論については 2012 年第 4 四半期のみの統計量を示す。

表 1 ：2012 年第 4 四半期から 2013 年第 3 四半期の収集ページ数

	2012-4Q	**2013-1Q**	**2013-2Q**	**2013-3Q**
ページ数　（1 期）	61,668,805	58,844,092	61,479,268	57,892,917
内容の重複なしページ数	45,933,605	42,932,982	45,111,527	42,192,931
4 期通しての統計				
総異なり URL 数（4 期）	64,539,233			
(内) 内容の重複ありページ数	27,604,915			
(内) 内容の重複なしページ数	36,934,706			

表 1 に収集したページ数の統計量を示す。1 億 URL を収集しても robots.txt の順守や各種 HTTP エラーにより，ページとして収集できたものが約六割にすぎない。重複検出は URL 毎に各ページのハッシュ値を計算し同一性を認定する。各期において内容の重複なし（異なり）ページ数は 4000 万強になる。4 期通しての総異なり URL 数は約 6400 万 URL と 1 億 URL に至らない。4 期中 2 期以上収集できたページ数の内，内容の重複があるページ数は約四割の 2700 万ページ，反対に内容の重複がないページ数は 3600 万ページになる。

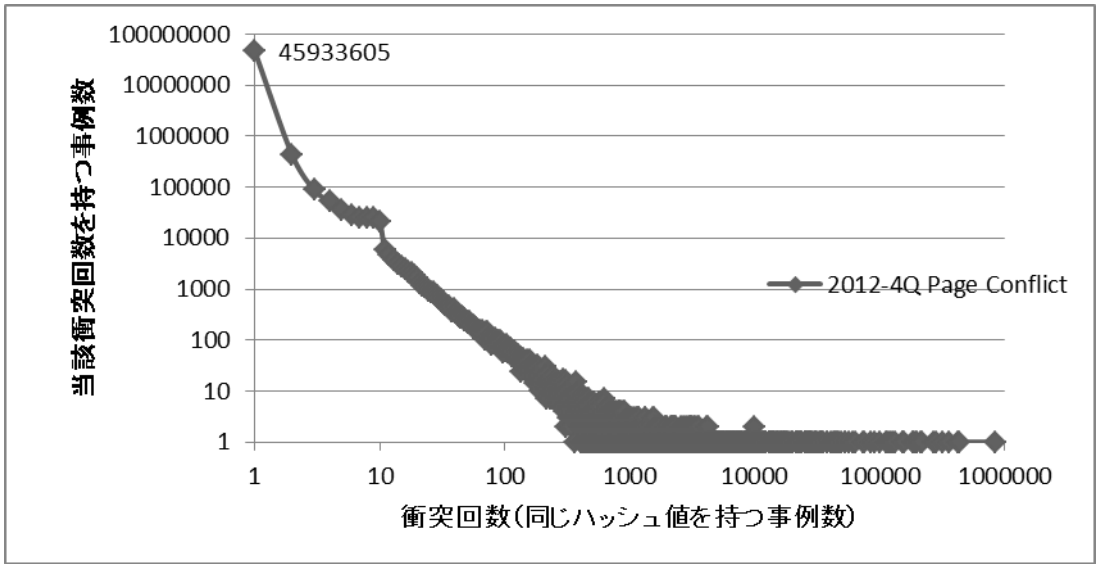


図 1 ：2012 年第 4 四半期収集ページの重複

図 1 に 2012 年第 4 四半期収集 Web ページの重複検出結果について示す。同じハッシュ値を持つ URL が複数存在することを「衝突」と呼ぶ。グラフ横軸は同じハッシュ値を持つ URL 数を示し，「衝突回数」と呼ぶ。グラフ縦軸は横軸の「衝突回数」を持つ衝突事例数（URL 数ではなくハッシュ値の異なり数で計算）を示す。グラフは両軸とも対数で表示している。グラフ中の左上の点が表の「内容の重複なしページ数」（他の URL と内容が重複しないページ数）に相当する。衝突回数 10 以下のものは同一内容の異なる URL 表示もしくはいわゆるコピーサイトであると考える。それ以上の衝突については，robots.txt や，「ソフト 404」

と呼ばれる当該 URL はサーバ上にないということを 404 HTTP ステータスコードでは返さず 200 HTTP ステータスコードで返し当該ページがないことを示すコンテンツを返すページである。

表２：2012 年第 4 四半期から 2013 年第 3 四半期の収集リンク数

	2012-4Q	**2013-1Q**	**2013-2Q**	**2013-3Q**
リンク先（のべ）	6,905,805,383	6,610,763,700	7,064,611,259	7,222,958,033
	69 億 URL	66 億 URL	70 億 URL	72 億 URL
リンク先（異なり）	892,135,930	843,166,672	865,694,816	855,684,918
	8.9 億 URL	8.4 億 URL	8.6 億 URL	8.5 億 URL
4 期通しての統計				
リンク先（異なり）	1,642,699,579			
	16 億 URL			

表２に 2012 年第 4 四半期（2012-4Q）～2013 年第 3 四半期（2013-3Q）の収集リンク数を示す。おおよそ 6000 万 URL の収集に対し，のべ 70 億前後，異なり 9 億弱の URL が収集できている。4 期を通した集計によるリンク先数が異なり 16 億 URL であることから 1 年間通して同じ URL を 4 期収集することにより 1 期のみクロールするのみ比べてリンクが約 1.8 倍（8.5 億～8.9 億→16 億 URL）に成長していることがわかる。

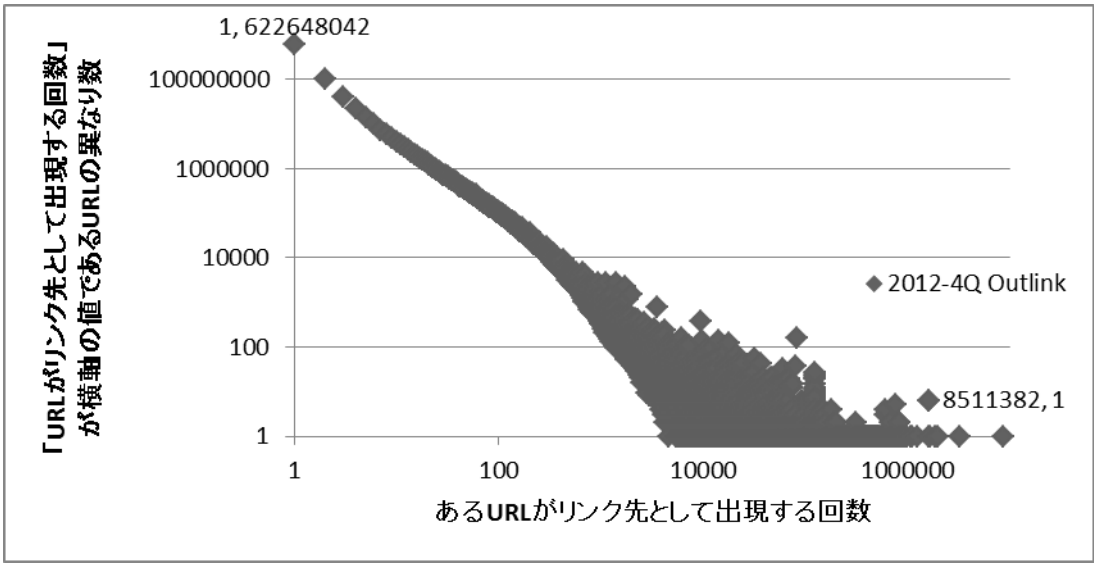


図２：2012 年第 4 四半期収集リンク先の統計

図２に 2012 年第 4 四半期の収集リンク先統計を示す。グラフ横軸はある URL がリンク先として出現する回数（被リンク数）で，グラフ縦軸が横軸相当の被リンク数を持つ URL の異なり数である。グラフの両軸は対数である。グラフ左上が 1 回しかリンクされていない URL 数で約 6 億件である。グラフ右下が最も多い被リンク数約 850 万を持つページで 1 件ある。これは有名ブログサイトのトップページであり，ブログの個々のページからリンクされている。

これらの統計情報から 2014 年以降の収集対象 URL を決定する。収集対象 URL は，4 期にクロールした同一 URL のうち内容が毎回変わっていた URL と，収集した Web ページのリンク先 URL の二種類を想定している。現在までに収集した Web ページに対するレジスタ分析は進んでいないが，レジスタ分析が進み次第，レジスタ分析結果の分散を見ながら収集対象 URL を決定したい。

３．組織化の進捗状況

本節では組織化の経過と現在までに得られている統計量について示す。

現在までのところ，正規化・形態素解析・係り受け解析までが部分的に進捗している。正規化においては nwc-toolkit による文字コード統制・文抽出を行った。利用するライブラリの関係でサーバ群ではなく安価な 8-core のワークステーション上で正規化作業を行った。8-core の CPU の並列処理により約 1 週間で 1 年分の収集データの正規化作業が完了する。正規化されたデータは MeCab/IPADIC，MeCab/UniDic，JUMAN により形態素解析を行う。32-core の計算サーバ上の並列処理により，それぞれ 1 日弱で 1 年分の収集データの形態素解析作業が完了する。さらに IPA 品詞体系に基づく CaboCha と益岡・田窪品詞体系に基づく KNP により係り受け解析を行う。32-core の計算サーバ上の並列処理により 1〜2 週間で係り受け解析作業が完了する。

表３に組織化したデータの基礎統計量を示す。Heritrix は収集 Web ページを圧縮 1GB サイズの WARC データに分割して出力する。展開すると約 3 倍程度になるため，表中の収集 WARC ファイル数に 3GB をかけた値が収集 Web ページ容量（メタデータを含む）と概算することができる。URL 数は前節の収集における URL 数である。正規化処理は nwc-toolkit による。正規化処理の際に文抽出なしに形態素解析（MeCab/IPADIC）を行うと各期のべ約 620〜647 億形態素になる。文抽出を行うと形態素数は各期約 300 億強になることから大体半分の形態素が日本語の文中の形態素ではないとして排除されている。抽出された文数はのべ数で各期 25 億文前後，文単位の同一性を認定すると文の異なり数は各期 10 億文になる。

表３：組織化したデータの基礎統計量

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
収集 WARC ファイル	814	870	910	905
URL 数	61,668,805	58,844,092	61,479,268	57,892,917
形態素数（文抽出なし）	64,714,650,129 647 億形態素	62,077,520,745 620 億形態素	63,414,252,638 634 億形態素	65,736,027,334 647 億形態素
形態素数（文抽出あり）	33,767,409,441 337 億形態素	32,651,138,004 326 億形態素	33,073,991,355 330 億形態素	30,923,912,566 309 億形態素
文数（のべ数）	2,678,315,774 26 億文	2,600,122,908 26 億文	2,659,617,620 26 億文	2,478,309,312 24 億文
文数（異なり数）	1,097,011,506 10 億文	1,048,772,913 10 億文	1,063,649,324 10 億文	1,007,771,383 10 億文

図３に 2012 年第 4 四半期の収集文の重複を示す。横軸が同一文の出現回数で縦軸が当該出現回数の文の異なり数を表す。両軸とも対数で表現する。10 億文のうち約 9 割の 8.9 億文が 1 回しか出現しない文である。以下，ページの重複も含めて同一文が異なりで 1 億文規模存在する。これらは定型的な表現やリンクの見出し語であることが多く，一番多く出現した文は 2,885,654 回出現する「職業とキャリア」（Yahoo!知恵袋のカテゴリ名）であった。ここで文抽出処理の妥当性について検証する。検証データとして BCCWJ に nwc-toolkit を適用して文数とバイト数がどの程度変化するかについて示す。

表５に nwc-toolkit のうち文抽出を行う nwc-toolkit-text-filter を BCCWJ に対して適用した場合のジャンル毎の文数とバイト数の変化を示す。表中左の文数はそれぞれの文境界基準により認定しており，OC（Yahoo!知恵袋）のように句点・感嘆符・疑問符が文境界として利用されない場合が多いジャンルについては nwc-toolkit により文数が増える場合があり，単純にどの程度削減されるかは評価できない。表中右のバイト数が純粋にデータとして削

減されたテキスト量を表す。バイト数に関して変化率が低い（多く削除されている）ジャンルは OP（広報紙）と PM（雑誌）であった。これらに対し，何が削除されているのかを確認したところ，人口などの統計情報・イベントの日時・窓口やお店の営業時間・連絡先（電話番号・住所）などが削除されていることがわかった。情報抽出の分野においては重要な情報ではあるが，言語研究においては削除しても多くの研究で影響は少ないと考える。

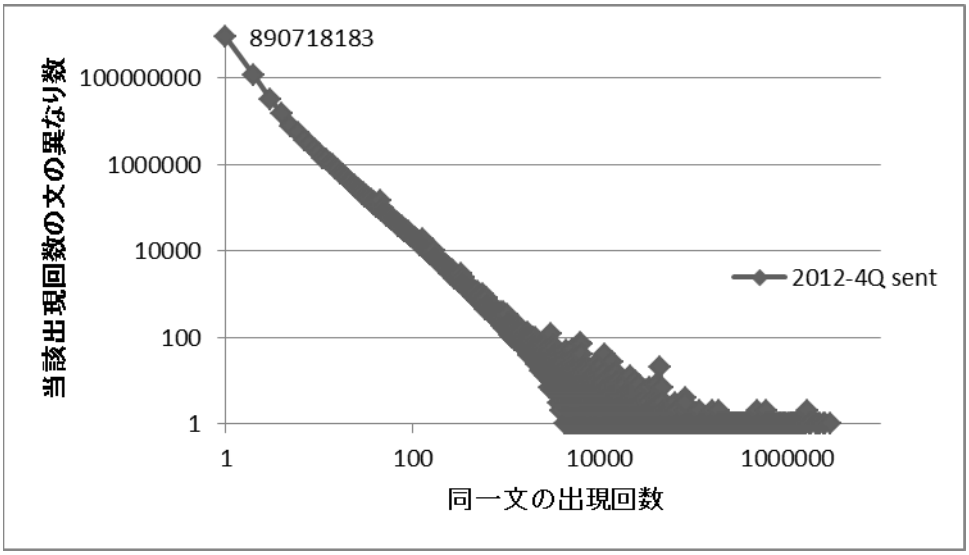


図３：2012 年第 4 四半期収集文の同一性

表４：BCCWJ の nwc-toolkit-text-filter の適用による変化

ジャンル名	文数			バイト数		
	処理前	処理後	変化率	処理前	処理後	変化率
LB	451,273	394,782	0.87	186,908	176,792	0.95
OB	222,437	203,467	0.91	23,236	22,242	0.96
OC	42,506	45,082	1.06	369,004	368,572	1.00
OL	38,827	34,761	0.90	6,004	5,864	0.98
OM	140,422	116,295	0.83	26,948	26,052	0.97
OP	257,796	142,660	0.55	22,372	14,636	0.65
OT	64,100	47,680	0.74	5,992	5,340	0.89
OV	18,977	17,807	0.94	1,656	1,632	0.99
OW	146,402	126,140	0.86	28,072	24,876	0.89
OY	117,816	76,205	0.65	224,064	213,480	0.95
PB	392,301	308,827	0.79	176,812	163,476	0.92
PM	301,399	227,089	0.75	30,248	25,812	0.85
PN	80,563	65,055	0.81	10,164	9,152	0.90
合計	2,274,819	1,805,850	0.79	1,111,480	1,057,926	0.95

現在収集しているコーパスデータについて語彙表の作成や n-gram データを進めている。unix コマンド sort, uniq, wc などに基づく単純な集計手法では sort アルゴリズムや途中で出力される中間ファイルにより時間計算量・空間計算量が想定よりも大きく時間がかかることがわかった。現在様々なソフトウェアを組み合わせて効率化をはかっており，2012 年第 4 四半期の n-gram 統計のみ得つつある。

４．本言語資源とグーグル「Web 日本語 N グラム」との比較

最後に 2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較を行う。

表５：2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較

	本言語資源(2012-4Q) MeCab/IPADIC 頻度 3 以上（文抽出有） 文単位での重複性排除有	本言語資源(2012-4Q) MeCab/IPADIC 頻度 3 以上（文抽出有） 文単位での重複性排除無	「Web 日本語 N グラム」 MeCab/IPADIC 頻度 20 以上 [工藤・賀沢 2007]
総形態素数 （のべ）	18,075,529,620 180 億形態素	33,767,409,441 337 億形態素	255,198,240,937 2550 億形態素
総文数	1,097,011,506 （異なり） 10 億文	2,678,315,774 （のべ） 26 億文	20,036,793,177 200 億文
1-gram	0.039 億	0.050 億	0.025 億
2-gram	0.47 億	0.85 億	0.80 億
3-gram	1.6 億	4.4 億	3.9 億
4-gram	2.1 億	8.7 億	7.0 億
5-gram	1.7 億	10.3 億	7.7 億
6-gram	1.2 億	9.7 億	6.8 億
7-gram	0.84 億	8.5 億	5.7 億

表 5 に 2012 年第 4 四半期収集データと 2007 年公開のグーグル「Web 日本語 N グラム」の比較結果を示す。「Web 日本語 N グラム」では頻度 20 以上の n-gram データのみを配布している。本研究ではできるだけ低頻度のデータを被覆するべく適切な頻度を検討中であるが，速報値として頻度 3 以上の概算値を報告する。総文数においては「Web 日本語 N グラム」の規模の 20 分の 1 から 10 分の 1 くらいの規模である。しかし低頻度のものも組織化することにより n-gram データの被覆では遜色ないレベルにできると考える。

表 6 ・表 7 は 2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」の頻度順位上位 10 件を比較するものである。頻度順位において，文境界メタ記号や句読点・記号類を含む n-gram は排除している。これらより各 n-gram の定性的な分析を行う。

まず本言語資源と「Web 日本語 N グラム」とで標本の量を比較すると 10 倍（文単位での重複性排除無）～20 倍（文単位での重複性排除有）くらい「Web 日本語 N グラム」の方が規模が大きい。単純に収集規模が小さい本言語資源において文単位の重複性排除無で n-gram を取ると 3-gram や 4-gram のレベルで複数のページで出現する定型的な表現が頻度順位上位にきてしまう。2-gram のレベルでも「利用規約」という定型的な表現が確認された。一方，文単位での重複性排除有の本言語資源では，規模が小さくても「Web 日本語 N グラム」と同様に機能語中心の n-gram が得られる。このように文単位での重複性排除を行うことにより Web 特有の定型的な表現の頻度の偏りをなくすことができ，定型的な表現の影響が少なくなる，より規模の大きな言語資源と同じ傾向のデータが得られていることがわかる。

5-gram 以上になると「Web 日本語 N グラム」においても定型的な表現の影響が避けられず，多くみられるようになる。一方，文単位での重複性を排除したものについては，文の一部を変えるような定型的な表現（「タグ が 付け られ た 質問」「に関する ウェブ 上 の 情報 を 探す」などは見られるが，基本的には機能表現中心のコロケーションが得られていることがわかる。

特殊な事例として一文中に同じ文字が連続している表現がある。「ああ ああ ああ ああ

ああ ああ ああ」や「え え え え え え え」などは，一文中に同じ文字が連続している表現の形態素解析結果である。例えば「え」が p 回出現するような一文に対して，p-6 回の「え え え え え え え」7-gram が枚挙される。一文中に複数の n-gram が出現する際にどのように数え上げるかは議論の余地がある。これは構造をもつデータに対する問い合わせにおいて起きる常に問題で，現在は多くのコンコーダンサにおいて正規化は行っていない。

表６：2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較
頻度順位上位 10 件（1-gram～4-gram）

	順位	1-gram	2-gram	3-gram	4-gram
本言語資源	1	の	し て	て い ます	し て い ます
MeCab/IPADIC	2	に	まし た	て い た	て い まし た
	3	て	て い	し て い	さ れ て いる
2012-4Q	4	が	て いる	し て いる	し て い た
文単位での重複性排除有	5	は	し た	と 思い ます	さ れ て い
	6	を	で は	さ れ て	た の です が
	7	た	に は	に なっ て	て き まし た
	8	で	さ れ	の です が	れ て い ます
	9	と	ませ ん	し まし た	は あり ませ ん
	10	し	い ます	さ れ た	に なり まし た
本言語資源	1	の	まし た	記事 へ の	記事 へ の トラック
MeCab/IPADIC	2	に	でしょ う	お願い し ます	専用 ページ を 表示
	3	を	行っ て	Q & A	利用 する こと が
2012-4Q	4	は	思っ て	続き を 読む	機能 を 利用 する
文単位での重複性排除無	5	て	情報 を	マーク へ 投稿	おすすめ の 知恵 ノート
	6	が	利用 規約	専用 ページ を	正確 性 の 保証
	7	た	おすすめ の	機能 を 利用	お客様 自身 の 責任
	8	で	記事 へ	済み の 質問	回答 を 指示 する
	9	と	追加 する	おすすめ の 知恵	便利 に 新規 取得
	10	し	場合 は	エンターテインメント と 趣味	はてな ブック マーク へ
「Web 日本語 N グラム」	1	の	し て	て い ます	し て い ます
MeCab/IPADIC	2	に	まし た	し て い	さ れ て いる
	3	を	て い	て い た	さ れ て い
[工藤・賀沢 2007]	4	は	て いる	し て いる	は あり ませ ん
	5	て	し た	さ れ て	れ て い ます
	6	が	ませ ん	に なっ て	て い まし た
	7	た	さ れ	し まし た	に なり まし た
	8	で	に は	さ れ た	し て おり ます
	9	と	で は	れ て いる	て き まし た
	10	し	い ます	あり ま せん	し て い た

例えば，「N1 が N2 が V1 する N3 を V2」といった文に対して「『N が』が先行する『V』」といった問い合わせを行う際に何件出現すると言えるだろうか？多くのコンコーダンサは <N1 が,V1>, <N1 が,V2>,<N2 が,V1>,<N2 が V2>の全組み合わせを出力し 4 件と出力するだろう。

このように n-gram を含めて 2 語以上の組み合わせを行う際には，重複して枚挙される現象に注意する必要がある。さらに，係り受け構造など構造を持ったデータに対する部分構造の枚挙に際しては，同様な事例が多く出現する。このような部分構造を共有する事例に

ついて，n-gram データとして正規化することは一切行わないが，利用者においては統計処理を行う場合に必要に応じて対処する必要がある。

本 n-gram データについては近いうちにデータそのものを公開するとともに，簡単な検索環境を Web API つきで対外的に公開し利用可能なようにする。2014 年度以降進めていく利用者系に関する意見などをいただければ幸いである。

表 7 ：2012 年第 4 四半期収集データとグーグル「Web 日本語 N グラム」との比較
頻度順位上位 10 件（5-gram～7-gram）

	順位	5-gram	6-gram	7-gram
本言語資源	1	さ れ て い ます	で は な い でしょ う か	の で は な い でしょ う か
MeCab/IPADIC	2	で は あり ませ ん	て い た の です が	の タグ が 付け られ た 質問
	3	と 思っ て い ます	の で は な い でしょ う	で は な い か と 思い ます
2012-4Q	4	し て い まし た	の で は な い か と	に関する ウェブ 上 の 情報 を 探す
文単位での重複性排除有	5	で は な い で しょう	に 行っ て き まし た	ああ ああ ああ ああ ああ ああ ああ
	6	の で は な い か	よう な 気 が し ます	の で は な い か と 思い
	7	は な い でしょ う か	タグ が 付け られ た 質問	し て いた の です が
	8	に なっ て い ます	の タグ が 付け られ た	思っ て い た の です が
	9	て い まし た が	さ せ て いただき まし た	え え え え え え え え
	10	て い た の です	たい と 思っ て い ます	と 思っ て い た の です
本言語資源	1	記事 へ の トラック バック	機能 を 利用 する こと が	機能 を 利用 する こと が でき
MeCab/IPADIC	2	機能 を 利用 する こと	利用 する こと が でき ませ	利用 する こと が でき ませ ん
	3	利用 する こと が でき	正確 性 を 保証 し て	正確 性 を 保証 し て おり
2012-4Q	4	正確 性 を 保証 し	お客様 自身 の 責任 と 判断	お客様 自身 の 責任 と 判断 で
文単位での重複性排除無	5	お客様 自身 の 責任 と	すべて の 機能 を 利用 する	すべて の 機能 を 利用 する こと
	6	はてな ブック マーク へ 投稿	知恵袋 の すべて の 機能 を	知恵袋 の すべて の 機能 を 利用
	7	更新 情報 が 届き ます	おすすめ の 解決 済み の 質問	ニックネーム の My 知恵袋 で 確認 でき
	8	おすすめ の 解決 済み の	記事 へ の トラック バック URL	質問 年月 や 画像 の 有無 を
	9	すべて の 機能 を 利用	ニックネーム の My 知恵袋 で 確認	質問 や 知恵 ノート は 選択 さ
	10	質問 年月 や 画像 の	する こと が でき ませ ん	以上 更新 が ない ブログ に 表示
「Web 日本語 N グラム」	1	さ れ て い ます	無料 で お 届け し ます	料 無料 で お 届け し ます
MeCab/IPADIC	2	で は あり ませ ん	料 無料 で お 届け し	配送 料 無料 で お 届け し
	3	で お 届け し ます	配送 料 無料 で お 届け	国内 配送 料 無料 で お 届け
[工藤・賀沢 2007]	4	無料 で お 届け し	国内 配送 料 無料 で お	以上 国内 配送 料 無料 で お
	5	1500 円 以上 国内 配送	円 以上 国内 配送 料 無料	円 以上 国内 配送 料 無料 で
	6	料 無料 で お 届け	以上 国内 配送 料 無料 で	1500 円 以上 国内 配送 料 無料
	7	配送 料 無料 で お	1500 円 以上 国内 配送 料	は インライン フレーム を 使用 し て
	8	国内 配送 料 無料 で	を 使用 し て い ます	フレーム を 使用 し て い ます
	9	以上 国内 配送 料 無料	インライン フレーム を 使用 し て	インライン フレーム を 使用 し て い
	10	円 以上 国内 配送 料	この 記事 へ の トラック バック	部分 は インライン フレーム を 使用 し

５．本言語資源を利用した理論的・応用的研究の可能性

本節では本言語資源を利用した理論的・応用的研究の展望について述べる。

まず語彙研究のために Web を母集団とするコーパスに対してどのような統計処理をすればよいかについて述べる。コーパスに基づく語彙研究は大きく分けて，生コーパスから得られる統計情報のみから語彙の性質を明らかにする教師なし統計学習に基づく手法と，何らかのアノテーションを付与してその情報を未知データに再現するような教師あり統計学習に基づく手法の二つが考えられる。

前者の教師なし統計学習においては，近年ベイズ統計学が注目されトピックモデルに基づく手法が盛んに利用されている。コーパスを用いてトピックモデルを構築する際に「文書」という単位と「単語」という単位が重要になる。本言語資源においては「文書」が個々の Web ページ，「単語」が形態素解析により認定された形態素が対応する単位になる。文書内の単語の共起などに基づき潜在変数を推定する。単語の与え方によっては，トピックを文書に遍在するレジスタとみなしレジスタ分析にも利用することができる。ここで本研究では「文」という単位を重要視する。Web の場合，コピーサイトや複数のページで共有する部分ページなどがあり，単純に「文書×単語関係」のみにより表現するとこれらの影響を受けてしまう。３節に示した統計量のように約 1 割の 1 億文については複数の文書に共有されている。そこで文書と単語の間に文という単位を仮定し，「文書×文関係」「文×単語関係」と階層的にモデル化することにより，コピーサイトなどの影響を前者の「文書×文関係」で吸収しながらレジスタをモデル化することが可能になると考える。さらに Web の場合には「文書×文書間関係」としてリンク―被リンク関係が規定されている。単リンクが張られる関係，もしくは同一文書にリンクを張る関係など，レジスタに影響を与える関係がいろいろ考えられる。この「文書×文書間関係」「文書×文関係」「文×単語関係」の多層の共起関係を，現実的に処理可能な計算量でどのようにモデル化するかを検討する。

後者の教師あり統計学習においては，古くは二値分類器が中心に用いられたが，最近では，順位・極性・系列などといった構造学習が自然言語処理の分野で広く用いられている。また特徴量としても木構造や有向非循環グラフなどといったものの部分構造の重複などを効率的に枚挙しながら距離（類似度）を設定して識別する学習手法が多く提案されている。本言語資源では形態論情報だけでなく，係り受け構造や述語項構造などといった情報を付与する。これにより言語研究の分野における構造の利用（特徴量としての利用と推定する対象としての利用）が促進されればと考えている。

利用者系において，３節で通常の n-gram データと文単位での重複を排除した n-gram の二つについての統計量を提示した。n-gram 検索ツールについては，この双方を提供して利用者による評価を待ちたい。また KWIC 検索においては後者の文単位での重複を排除したコーパスに対する検索ツールの開発のみを検討している。３節や４節に示すような様々な情報を提供しながら文単位での重複性の影響について議論していきたい。

形態論研究においては，Web 上の多様な語彙を観察することが可能になるだけでなく，保存する予定であるデータを通時的に観察することで経年変化を分析することが可能になる。Web の世界においては多様な形態論変化が生産され，すぐに廃れていく傾向にある。これらを適切にとらえることができるような環境を構築できればと考えている。

コーパス日本語学において，統語論研究に必要な情報が現状のコーパスから得られているとは言い難い状況にある。まず BCCWJ では品詞情報や係り受け構造に基づく問い合わせをおこなったとしても真に分析したい言語現象が数例しかないということが起きうる。例えばガ格が二つ以上出現する文（複文）の従属節境界の左端のあいまい性（pre-head attachment; Kamide and Mitchel 1999）を研究する際に，BCCWJ に様々な構造を問い合わせてみると「N ガ N ヲ VP N ガ」という語順が 1 例，「N ガ ADV N ガ」という語順が 3 例，「N ガ N ニ N ガ」という語順が 1 例しか出現しないことがわかる。規模を 100 倍にすることでこういった言語現象を発見する可能性は高くなる。また人文系の研究者側の問題として，多くの研究が語彙情報を単純なベクトル表示にする bag-of-words 的な共起以上の情報を使いこなせていないという問題がある。自動解析ながらもあらかじめ係り受け構造等を付与し，使いやすい利用者系を構築することで，語順や部分木構造などを考慮したコーパス分析手法が根付くような土壌を整備していきたい。

意味論研究においても，頻度情報によらない研究にまで踏み込むことが重要であると考える。この分野においては，共起に基づく研究は盛んに行われている一方，アノテーションを行い，さらに構造学習を導入してアノテーションを未知データに復元することで何かを明らかにするような水準に達していないと考える。ここで重要なのは，解明しようとし

ている研究課題が本質的にコーパスを用いる手法が適しているかどうかを見極める必要があるということである。コーパスを用いる手法が何も証拠を提供できない問題は，言語研究において多々あると考える。そういった問題に対しても，コーパスが仮説を立てるための手がかりを研究者に与えることは可能な場合が多いが，手がかりのみをもって証拠とすることのないように注意されたい。

６．おわりに

本論文では国立国語研究所コーパス開発センターで進めている Web を母集団とした超大規模コーパス開発プロジェクトの進捗について報告した。6 年間のプロジェクトのうち半分の 3 年が経ち，単純に 100 億語規模のコーパスを構築するだけでなく，継続的に同規模のスナップショット的なコーパスを 1 年間に複数回構築可能である環境が構築されつつある。収集においては Heritrix クローラを利用して年間 4 回のクロールを行い，組織化においては正規化・形態素解析・係り受け解析まで進捗している。組織化における残る課題は述語項構造解析とレジスタ分析と語彙表構築であるが，これについても早急に環境を整えたい。得られた基礎統計量から 3 か月単位で目標の 100 億語規模のコーパスが構築できていることがわかった。今後より一層の効率化を進めたい。

残りの期間で人文系の研究者が柔軟に利用可能な利用者系と保存環境の構築を行う。語彙調査や用例検索に留まらない，自然言語処理で培われた構造に基づく問い合わせ環境を構築したい。これにより高い被覆性を持ちながらも柔軟なコーパス調査を可能にし，統語論・意味論研究を前に進める研究環境を提供できると考える。

一方で，本言語資源に基づく調査で解明できない問題が何であるのかを示すが重要だと考える。規模を大きくすることだけでは解明できない問題についても示していきたい。

謝　辞

本研究は国語研「超大規模コーパス構築プロジェクト」によるものです。

文　献

Kamide, Yuki and D. C. Mitchell （1999） Incremental Pre-head Attachment in Japanese Parsing, *Language and Cognitive Processes,* 14(5/6), 631-662.
工藤拓・賀沢秀人（2007）『Web 日本語 N グラム第 1 版』，言語資源協会発行 (http://www.gsk.or.jp/catalog/GSK2007-C/).
前川喜久雄（2007）「コーパス日本語学の可能性–大規模均衡コーパスがもたらすもの–」日本語科学, 22, 13-28.
前川喜久雄・山崎誠（2008）「『現代日本語書き言葉均衡コーパス』」国文学解釈と鑑賞, 932（74 巻 1 号）, 15-25.
矢田晋（2010）.『日本語ウェブコーパス 2010（NWC 2010）』, (http://s-yata.jp/corpus/).

関連 URL

（1） Web 日本語 N グラム第 1 版：(http://www.gsk.or.jp/catalog/GSK2007-C/).
（2） クローラ Heritrix-3.1.1：(http://webarchive.jira.com/wiki/display/Heritrix/Heritrix).
（3） 日本語ウェブコーパス用ツールキット：(http://code.google.com/p/nwc-toolkit/).
（4） 形態素解析器 MeCab-0.996：(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html).

語義曖昧性解消を対象とした領域固有のシソーラスの構築

新納 浩幸 (茨城大学 工学部 情報工学科) [1]
國井 慎也 (茨城大学大学院 情報工学専攻) [2]
佐々木 稔 (茨城大学 工学部 情報工学科) [3]

Construction of Domain Dependent Thesaurus for Word Sense Disambiguation

Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)
Shinya Kunii (Ibaraki University, Department of Computer and Information Sciences)
Minoru Sasaki (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

多くの自然言語処理システムでは，その適用領域を限定することで，システムのパフォーマンスを改善できる．語義曖昧性解消 (Word Sense Diaambiguation, WSD) においても，適用領域を限定することで，識別精度を改善できると考えられるが，その具体的方法は明らかではない (Navigli (2009))．一方 WSD では，一般に，シソーラスを利用することで，その識別精度を改善できる．そこで本論文では領域固有のシソーラスを構築することで，適用領域を限定した WSD の精度向上を図る．

WSD は文中の多義語である対象単語に対して，その単語がどの語義で使われているかを識別するタスクである．このタスクは自然言語処理の中心課題であり，従来より多数の手法が提案されているが，近年は教師付き学習を用いる手法が中心である．そこではまず少量の対象単語の用例を用意し，その用例中の対象単語に語義 $c \in C$ を付与する．次に用例中の対象単語の周辺の文脈を特徴ベクトル $\boldsymbol{x}$ で表し，先に付与した語義とのペア $(\boldsymbol{x}, c)$ を作る．このペアの集合が訓練データとなり，この訓練データから SVM 等の分類器を学習することで WSD を解決する．

このアプローチで問題となるのは，訓練データのスパース性 (Feature Sparseness) である．例えば素性として周辺の単語表記だけを利用する場合，訓練データから得られる素性の集合は，取り得る素性全体に対してスパースとなる．そのためテストデータでは訓練データには全く出現しない単語のみで特徴ベクトルが構成される場合も多く，その場合，正しく語義識別ができない可能性が高い．そこでスパース性を解消するためにシソーラスを利用することが一般に行われている．シソーラスとしては分類語彙表 (分類語彙表 (2004)) などの手作業で作られたものが一般に利用されるが，コーパス内の単語をクラスタリングすることで自動構築したシソーラスも利用できる．WSD の適用領域を限定した場合，その領域のコーパス S から自動構築されるシソーラス T_S を利用すれば，一般的なシソーラスを利用するよりもより高い精度が実現できると考えられる．つまり上記のシソーラス T_S が領域 S の固有のシソーラスになると考えられる．ただしシソーラスは単語の語義の上位下位関係を表したものであり，シソーラスが領域に依存するとは考えづらい．そのため WSD の対象領域のコーパスに様々なコーパスを加え合わせたコーパス G からシソーラス T_G を構築した場合，シソーラス T_G はシソーラス T_S を包含していると考えられ，T_S が T_G よりも有用であるとは単純には言えない．

本論文では，基本的に T_G よりも T_S を使う方が WSD の識別精度の観点では有利であるが，G が S を拡大するような形になっている場合に，T_G を使う効果が現れると考える．そこで T_G が T_S に比較して有効な度合い α を導入し，T_G と T_S の以下で示す線形結合により領域 S に固有のシソーラス $T_{S'}$ を作成することを提案する．

$$T_{S'} = (1-\alpha)T_S + \alpha T_G \tag{1}$$

[3] shinnou@mx.ibaraki.ac.jp
[3] 13nm707s@hcs.ibaraki.ac.jp
[3] msasaki@mx.ibaraki.ac.jp

問題は α の算出方法と (式 1) の実現方法である．α の算出のために，コーパスに対する代表単語群という単語集合を定義する．ここではコーパス内の単語を頻度順に並べ，累積頻度が全体のある割合になるまでの単語群を代表単語群と定義する．今，コーパス S と G の代表単語群をそれぞれ W_S，W_G とする．これらを用いて α を以下で定義する．

$$\alpha = \frac{|W_S \cap W_G|}{|W_S|}$$

また (式 1) の実現方法としては，S と G のそれぞれから単語 w の縮約ベクトル w_s と w_g を作成し，w_s と w_g に (式 1) に対応する重みを乗じて結合させることで，単語 w の特徴ベクトル w_{sg} を構築し，$\{w_{sg}\}_{w \in G}$ をクラスタリングすることで $T_{S'}$ を構築する．

実験では現代日本語書き言葉均衡コーパス（BCCWJ (Maekawa (2007))）における 3 つの領域 OC (Yahoo! 知恵袋)，OY (Yahoo! ブログ) 及び PB (書籍) を利用する．提案手法を用いて各領域の固有のシソーラスを構築した．また東京工業大学の奥村研究室では BCCWJ に語義タグをつけたデータを構築しており，本実験の WSD ではこのデータを利用させていただいた．ある程度の頻度を持つ多義語を選び，OC では 32 単語，OY では 23 単語，PB では 46 単語の WSD を行う．構築できた領域の固有のシソーラスの他，いくつかの自動構築したシソーラスを用いて WSD を行う．5 分割交差検定による WSD の平均正解率からシソーラスの評価を行う．

2　領域固有シソーラスの作成

2.1　コーパスに対する代表単語群の非被覆

本論文では T_G が T_S に比較して有効な度合い α を導入し，α を用いた T_G と T_S の線形結合 (式 1) により領域 S に固有のシソーラス $T_{S'}$ を作成する．

α の算出法を示す．本論文では基本的に T_S は T_G よりも有効であると考える．そして T_G が T_S よりも有効になるとしたら，G が S を拡大するような形になっている場合だと考える．この拡大の度合いが α を意味する．そして本論文ではこの拡大の度合いを以下のように考える．

まず G や S には WSD で利用される重要な単語がある．それをここでは代表単語群とよぶ．そして G の代表単語群を W_G，S の代表単語群を W_S とする．ここで注意として $S \subset G$ だが $W_S \subset W_G$ とは限らない．今，そこで $W_S - W_G = W_S - W_S \cap W_G$ の意味を考えると，これは S においてのみ重要である単語と考えられ，その度合いは以下で示させる．

$$\frac{|W_S - W_G|}{|W_S|} = 1 - \frac{|W_S \cap W_G|}{|W_S|}$$

ここから α は以下となる．

$$\alpha = 1 - \frac{|W_S - W_G|}{|W_S|} = \frac{|W_S \cap W_G|}{|W_S|} \tag{2}$$

次にコーパスに対する代表単語群の定義であるが，本論文ではコーパス内の単語を頻度順に並べ，コーパス S の場合は，累積頻度が全体の 8 割になるまでの単語群，またコーパス G の場合は，累積頻度が全体の 7 割になるまでの単語群を代表単語群とした．

2.2　シソーラスの線形結合

前述した T_G が T_S に比較して有効な度合い α を利用して，領域 S に固有のシソーラス $T_{S'}$ の構築方法を述べる．

S と G のそれぞれから単語 w の縮約ベクトル w_s と w_g を作成する．具体的にはトピックモデルである Latent Dirichlet Allocation (LDA) (Blei et al. (2003)) を用いて作成する．コーパス S に対して LDA により $p(w|z_i)$ が得られる．ここで z_i は i 番目の潜在的トピックである．またここではトピックの数を 100 に設定する．そして w_s を以下とする．

$$w_s = \frac{1}{Z}(p(w|z_1), p(w|z_2), \cdots, p(w|z_{100}))$$

ここで $Z = \sum_{i=1}^{100} p(w|z_i)$ である．同様にしてコーパス G に対して LDA を適用して w_g を得る．注意として $w \notin S$ の場合，w_s は 0 ベクトルとなる．

LDA から求まった w_s や w_g はクラスタ数を 100 とした場合の，単語 w のソフトクラスタリング結果に相当する．これらのクラスタリング結果に重み α をつけて統合することで，領域 S に固有のシソーラス $T_{S'}$ を構築する．具体的には重み付きアンサンブルクラスタリングを行う．処理的には w_s に $1-\alpha$ を乗じた 100 次元のベクトルと，w_g に α を乗じた 100 次元のベクトルを結合させ，200 次元のベクトル w_{sg} を作成する (図 1 参照)．この w_{sg} を単語 w の特徴ベクトルと考えて G 内の単語をクラスタリングする．このときのクラスタ数は 1000 に設定する．またクラスタリングアルゴリズムには k-means を用いた．

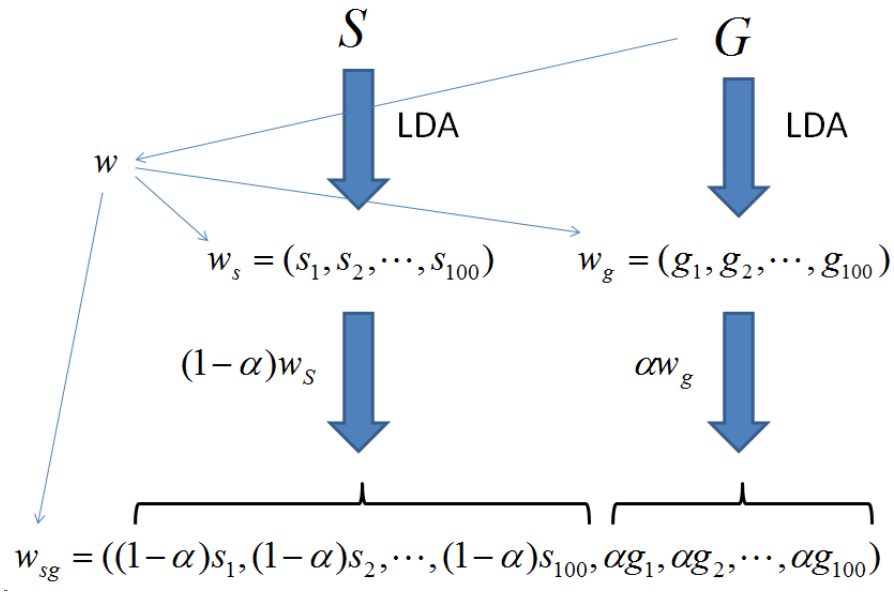


図 1: w_{sg} の作成

3　実験

シソーラスを評価するために，WSD の対象単語 w に対する訓練データ作成の際に，w の周辺の名詞に対して，シソーラスから得られるその名詞のクラスタ番号を素性として含めて識別を行う．また WSD の領域は固定する．ここでは BCCWJ の 3 つの領域 OC (Yahoo! 知恵袋)，OY (Yahoo! ブログ) 及び PB (書籍) を利用する．つまり提案手法により 3 つの領域固有のシソーラスを構築し，それぞれの領域の WSD において，領域固有のシソーラスの効果を確かめる．また対象領域を含むより大きなコーパス G としては，上記 3 つの領域のコーパスを合わせたコーパスとした．

表 1 にそれぞれの領域におけるコーパスのサイズ[4]，名詞の種類数，本実験における WSD の対象単語の数，その対象単語の平均語義数，及びそのた対象単語の平均用例数を示す．また表 2 に対象領域で用いた WSD の対象単語と語義数（括弧内の数）を示す．

[4]配布形式の（VARIABLE の）XML ファイルのサイズの合計であり，テキスト部分のサイズではない．

表 1: 利用コーパスのデータ

	OC	OY	PB	G (OC+OY+PB)
サイズ (KB)	125,347	129,384	255,528	510,260
名詞の種類数	83,130	141,950	181,356	278,629
WSD 対象単語数	32	23	46	-
WSD 平均語義数	2.81	2.91	3.06	-
WSD 平均用例数	158.4	139.5	128.4	-

表 2: 対象単語とその領域内の語義数

OC	ある (4)，いう (2)，意味付ける (3)，いれる (2)，教える (3)，買う (2)，書く (2)，聞く (2)，来る (2)，くれる (2)，子供騙し (2)，時間割り (2)，自分自身 (2)，ため (3)，できる (3)，出る (3)，とき (3)，ところ (3)，とる (7)，ない (2)，なか (3)，なる (2)，願う (2)，場合 (2)，入る (4)，前 (3)，みる (5)，持つ (4)，やる (3)，行く (2)，よい (3)，よる (3)
OY	買う (2)，来る (2)，くれる (2)，こと (3)，時 (2)，時間割り (4)，自分自身 (2)，する (8)，ため (3)，できる (3)，出る (3)，度 (3)，とき (2)，ない (2)，なか (2)，年 (3)，日 (3)，人 (3)，分 (5)，まあ (2)，前 (2)，もの (4)，やる (2)
PB	あう (6)，あげる (6)，歩く (2)，言う (2)，生きる (2)，意味付ける (3)，入れる (3)，書く (2)，考える (2)，関係付ける (3)，聞く (2)，技術屋 (2)，来る (2)，こどもだまし (2)，これ (2)，時間割り (2)，時代物 (5)，自分自身 (2)，社会問題 (3)，する (8)，それ (3)，高い (2)，出す (3)，立つ (4)，立てる (5)，強い (2)，手 (4)，出る (3)，とる (7)，人間業 (2)，場合 (2)，入る (4)，場所割り (2)，人 (2)，一つ (3)，ほとんど (2)，前 (2)，身 (2)，見える (4)，見る (6)，持つ (3)，モデルチェンジ (3)，山 (2)，やる (4)，行く (2)，呼ぶ (2)

また WSD で利用した素性は以下の 9 種類である．(e0) w の表記，(e1) w の品詞，(e2) w_{-1} の表記，(e3) w_{-1} の品詞，(e4) w_1 の表記，(e5) w_1 の品詞，(e6) w の前後 3 単語までの自立語の表記，(e7) e6 のクラスタ番号，(e8) w_{-1}，w_1，(e6) の各単語に対するクラスタ番号．なお対象単語の直前の単語を w_{-1}，直後の単語を w_1 としている．

シソーラス T を用いた各領域における WSD の実験を述べる．まず各対象単語の用例を素性リストに変換するが，その際，シソーラス T を利用して素性の一部に T におけるクラスタ番号が付与されている（上記の素性 (e8) に相当）．実験は 5 分割交差検定で行う．つまり対象単語 w の用例の素性リストの集合を 5 分割し，1 つをテストデータ，残り 4 つを訓練データとして SVM で学習し，w に対する正解率を測る．テストデータを変更することで 5 つの正解率が得られるが，それらの平均を w に対する正解率とする．領域 S における各対象単語ごとに上記の正解率を測り，それらを平均した値を領域 S におけるシソーラス T の平均正解率とする．

シソーラス T としては，シソーラスを使わないもの None とその領域 S のコーパスのみから作成したシソーラス T_S，コーパス G から作成したシソーラス T_G および提案手法に従って作成したシソーラス $T_{S'}$ を用いた．実験の結果を表 3 に示す．表 3 には $T_{S'}$ を作成する際に用いた α の値も示した．α の値は前章で説明した方法から算出した．

表 3: 利用したシソーラスに対する平均正解率 (%)

	OC	OY	PB	平均
None	**83.66**	84.58	82.58	83.51
T_S (対象コーパス)	83.54	84.92	83.18	83.88
T_G (全コーパス)	83.21	84.38	**83.21**	83.60
$T_{S'}$ (提案手法)	83.45	**85.49**	82.74	**83.89**
α	0.139	0.404	0.538	

総合的にみると，提案手法により作成したシソーラスによる正解率は，対象コーパスのみを用いて作成したシソーラスによる正解率や全体のコーパスを用いて作成したシソーラスによる正解率よりもわずかではあるが高かった．

4　考察

4.1　α の算出

本論文では T_S と T_G を以下の方法で結合している．

$$T_{S'} = (1-\alpha)T_S + \alpha T_G$$

この式において α は T_G の重みである．本論文では α の意味を T_G が T_S に比較して有効な度合い α と考えたが，それ自体は上記の式に沿っている．問題は α の算出方法であり，本論文で考案した (式 2) が妥当な値を示しているかどうかは不明である．

ここでは α の値を $0.2, 0.4, 0.6, 0.8$ と変化させた場合の平均正解率の変化を各領域に対して求めた．その結果を図 2，図 3，図 4 に示す．

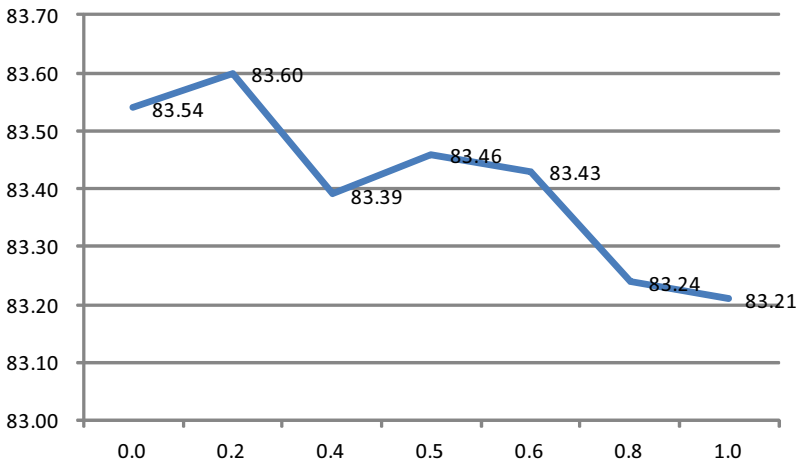


図 2: α に対する平均正解率（ＯＣ）

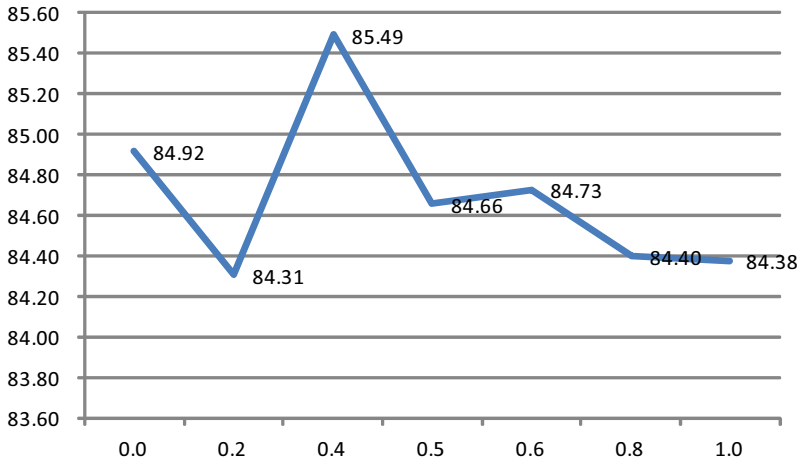


図 3: α に対する平均正解率（ＯＹ）

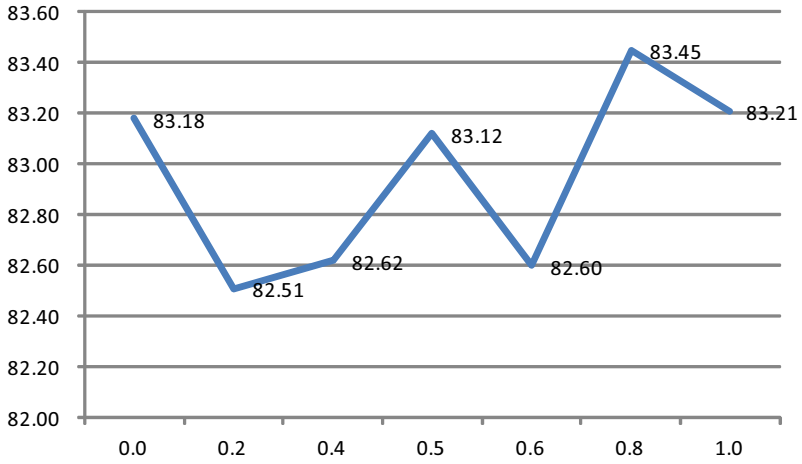


図 4: α に対する平均正解率（ＰＢ）

これらの図からわかるように適切な α の値を求めることができれば，T_S や T_G を単独で用いるよりも，それらを結合した $T_{s'}$ を用いた場合の方が WSD の精度が高まる．

しかしどの図も α と平均正解率が線形あるいは単峰性の関係になっていない．これは α の算出が困難であることを示すと共に，$T_{S'}$ を (式 1) でモデル化することが不適切であることも示唆している．対象単語毎に適切なシソーラスが異なることが原因だと予想している．

4.2　その他のシソーラスとの比較

T_S と T_G の結合方法として，本論文では LDA を用いて，それぞれのコーパスに対する単語の次元圧縮表現 (100 次元) を導き，それらを重み付きで結合して単語をベクトル化し，それをもとにクラスタリングする (クラスタ数 1000) という手順をとった．

ただし T_S と T_G の結合方法としては，コーパス S とコーパス G を結合させたコーパス $S+G$ を作り，$S+G$ からシソーラス T_{S+G} を作ることも考えられる．

またシソーラスとしては本論文で自動構築したもの以外に，既存の分類語彙表を用いることも可能である．さらに本論文で用いたコーパスとは別の一般に言語研究で用いられる新聞記事（毎日新聞 2005 年〜2008 年度版，4 年分）から作成したシソーラス T_{news} を用いることもできる．ここではこれらのシソーラスを用いた場合の正解率を求めた．表 4 に結果を示す．

表 4: その他のシソーラスによる平均正解率 (%)

	OC	OY	PB	平均
T_{S+G}	83.37	84.04	83.05	83.49
分類語彙表	83.04	84.53	**83.17**	83.58
T_{news}	**83.72**	84.45	82.59	83.59
提案手法 ($T_{S'}$)	83.45	**85.49**	82.74	**83.89**

提案手法のシソーラス $T_{S'}$ を用いる場合が最も良かった．その他のシソーラスには大きな差はなかった．

4.3　分野依存知識と一般知識の利用

本研究は WSD へ分野依存知識を利用する研究の一つである．従来，WSD へ分野依存知識を利用する研究は Domain-Driven Disambiguation (DDD) と呼ばれるアプローチの中で行われてきた (Strapparava et al. (2004))．これは基本的に対象単語の語義を内容語の領域と対象語義の領域との比較から選択する手法を取る (Gliozzo et al. (2004)，Gliozzo et al. (2005)，Magnini et al. (2002))．ただし利用している分野依存知識は語義分布である．本タスクのように領域をコーパスとして設定し，更に WSD の対象単語が一般的な単語である場合，語義分布が分野依存知識として活用できるケースは少ない．

また WSD へ分野依存知識を利用するアプローチは，WSD の領域適応の問題と関連が深い．領域適応とは学習元のデータが存在するソース領域と，実際に学習により得られた分類器を適用する先のターゲット領域が異なる問題である (Sogaard (2013))．ターゲット領域の知識をどのように取り込めるかが 1 つの課題である．我々は論文 (新納浩幸・佐々木稔 (2013)) において，WSD の領域適応にターゲット領域のトピックモデルを利用することを提案したが，そこではターゲット領域のシソーラス (本論文の T_S に相当) のみを利用している．一方，森は形態素解析の領域適応の問題を扱い，ターゲット領域の知識を含むより広範囲の領域の知識を利用することで，ターゲット領域の知識のみの利用よりも更に精度が上がったことを報告している (森信介 (2012))．しかし WSD などの他のタスクでもこの性質があるのかどうかは不明である．

ターゲット領域の知識は分野依存知識に相当し，ターゲット領域の知識を含むより大規模な知識は一般知識に相当する．WSD では分野依存知識（シソーラス T_s に相当）が概ね有効であるが，一般知識（シソーラス T_G に相当）を併用して利用することで精度向上が期待できる．論文 (Kunii and Shinnou (2013)) では T_s と T_G のそれぞれから得られる素性に重みをつけて，その素性を WSD の分類器の学習に利用している．本論文では T_s と T_G を融合した $T_{S'}$ を構築するというアプローチをとっている．どちらも分野依存知識と一般知識を融合して利用する研究である．

5　おわりに

本論文では適用領域を限定させた WSD の精度向上のために，WSD で利用されるシソーラスを，領域固有のシソーラスとして構築する方法を提案した．概略，適用領域のコーパス S と，S に別領域のコーパスを加えたコーパス G の各々からシソーラスを作成し，それらを線形結合した．線形結合する際の重みの算出には代表単語群の（非）被覆という概念を導入した．

BCCWJ の 3 つの領域 OC，OY，PB の各々で様々なシソーラスを用いて WSD を行うことで提案手法を評価した．提案手法により構築したシソーラスを用いた場合の正解率が，総合的に見れば，最も優れていたが，その差はわずかであり有意性には疑問もある．

ただし S のみから構築したシソーラス T_S や，G のみから構築したシソーラス T_G よりも何らかの形で融合させれば，より有効なシソーラスが構築できることは明らかである．今後はこの融合の方法を更に考えたい．

文献

David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003) “Latent dirichlet allocation,” *Machine Learning Reseach*, Vol. 3, pp. 993–1022.

Alfio Massimiliano Gliozzo, Bernardo Magnini, and Carlo Strapparava (2004) “Unsupervised Domain Relevance Estimation for Word Sense Disambiguation.,” in *EMNLP*, pp. 380–387.

Alfio Gliozzo, Claudio Giuliano, and Carlo Strapparava (2005) “Domain kernels for word sense disambiguation,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 403–410, Association for Computational Linguistics.

Shinya Kunii and Hiroyuki Shinnou (2013) “Combined Use of Topic Models on Unsupervised Domain Adaptation for Word Sense Disambiguation,” in *Proc. of PACLIC-2013*, pp. 415–422.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo (2002) “The role of domain information in word sense disambiguation,” *Natural Language Engineering*, Vol. 8, No. 04, pp. 359–373.

Roberto Navigli (2009) “Word sense disambiguation: A survey,” *ACM Computing Surveys (CSUR)*, Vol. 41, No. 2, p. 10.

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

Carlo Strapparava, Alfio Gliozzo, and Claudio Giuliano (2004) “Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3,” in *Proc. of SENSEVAL-3 Third International Workshop on Evaluation of Systems for the Semantic Analysis of Text*, pp. 229–234.

新納浩幸、佐々木稔 (2013)「k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応」，自然言語処理，第 20 巻，第 5 号，pp.707–726.

森信介 (2012)「自然言語処理における分野適応」，人工知能学会誌，第 27 巻，第 4 号，pp.365–372.

分類語彙表 (2004)「国立国語研究所」.

複雑ネットワークの視点による二字熟語の構造について

山本健（中央大学理工学部）

Structure of Two-letter Kanji Compounds Based on Complex Network Analysis

Ken Yamamoto (Faculty of Science and Engineering, Chuo University)

1　複雑ネットワーク

複雑ネットワークは、多数の要素の「つながり方」について調べる道具として発展してきた。人と人との知人関係、インターネットのルータの接続関係、および企業間の取引関係など様々なシステムが複雑ネットワークの視点でモデル化され分析されている。複雑ネットワークを構成する要素を**頂点**、頂点どうしのつながりを**辺**とよぶ。

数珠つなぎに知り合いをたどっていくと、少ない人数を経由するだけで誰にでも行き着けるという都市伝説がある[1]。複雑ネットワークのことばでは、ある頂点から辺をたどって別の頂点に行くための最短距離（経路長）が小さいということである。現実の多くのネットワークでも平均的な経路長が小さいことが知られている。また、「自分の知人の知人」にあたる人物が自分の知人である可能性は、世界中からでたらめに選んだ人物が自分の知人である可能性よりも格段に大きい。つまり、知人関係のネットワークには「三角形」が多く含まれるということである。ネットワーク中の「三角形」の多さを測る量が**クラスター係数**である。ネットワークが三角形を全く含まなければクラスター係数は 0 となり、ネットワークのどの 3 頂点も三角形をつくっていればクラスター係数は 1 になる。平均経路長が小さくクラスター係数が大きいとき、ネットワークは**スモールワールド**であるという。

ネットワークの統計的指標として**次数分布**がある。ちょうど k 本の辺が接続している頂点（次数が k である頂点）の割合を $P(k)$ で表し、これを次数 k に関する確率分布とみなしたとき、$P(k)$ を次数分布とよぶ。次数分布がベキ乗関数

$$P(k) \propto k^{-\gamma}$$

で表されるようなネットワークを、**スケールフリー**ネットワークとよぶ。$k \to \infty$ におけるベキ乗関数の減衰は指数関数や正規分布関数よりも緩慢であり、スケールフリーネットワークでは、次数がとても大きな頂点が存在しやすいという特徴がある。

Watts, and Strogatz (1998) によるスモールワールド性および Barabási, and Albert (1999) によるスケールフリー性の発見が、近年の複雑ネットワーク研究の潮流をつくるきっかけとなった。複雑ネットワークの諸概念をまとめた成書としては、増田、今野 (2010), 矢久保 (2013), Newman (2010) など豊富に刊行されている。以下、「複雑ネットワーク」のことを単に「ネットワーク」とよぶこととする。

本稿で取り扱うのは、二字熟語を「2 つの漢字を結ぶ辺」とみなして構成されるネットワークである。このネットワークがスモールワールドかつスケールフリーであることを示し、これらの特徴を定量的に再現するシンプルなネットワーク生成モデルを提案する。

[1] いわゆる「6 次の隔たり」

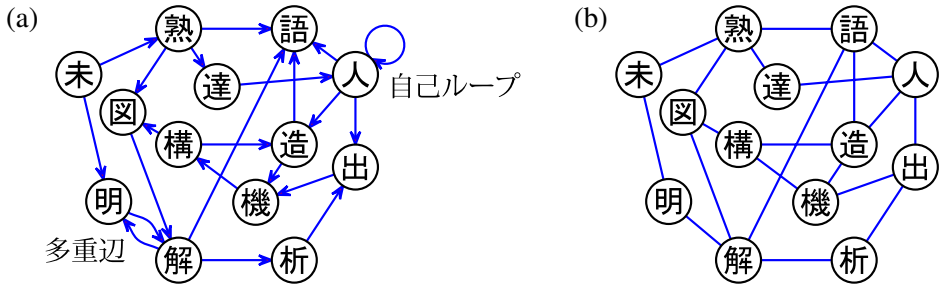


図 1: 二字熟語ネットワークの一部分。(a) 辺は向きをもち、ネットワークには自己ループおよび多重辺が存在する。(b) 本研究では辺の向きを無視して単純化したネットワークに注目する。

2　二字熟語ネットワーク

図 1 は二字熟語ネットワークの一部分である。例えば、「構造」という熟語に対応して「構」と「造」を結ぶ辺が引かれている。漢字の順序（1 字目か 2 字目か）があるので、ネットワークの辺には自然に向きがつけられる。「人人」という熟語のように、ある頂点から自分自身へ向かう辺（自己ループ）が存在する。また、「明解」と「解明」のように複数の辺をもつ頂点の対も存在する。しかし本研究では、漢字と漢字の最も単純なつながり方に関する特徴を明らかにするという点から、辺の向きは考えないこととする。自己ループを消去し、多重辺は 1 本にまとめたネットワーク（無向単純ネットワーク）について調べる（図 1 (b)）。

二字熟語のデータとして 3 冊の国語辞典『広辞苑　第四版』、『岩波国語辞典　第五版』、および『三省堂国語辞典　第 4 版』を利用した。それぞれの国語辞典の見出し語から漢字 2 文字のものをすべて抜き出し、対応するネットワークの解析をおこなった。分析結果を表 1 に示す。ネットワークの規模を特徴づける頂点数（= 漢字の種類）、辺数（≈ 熟語の個数[2]）、および平均次数（1 つの漢字が平均して何個の熟語に使用されているか）はいずれも広辞苑が他の 2 冊より大きくなっている。これは、直感的な「辞書の規模」と一致する結果である。平均経路長 l が小さく、クラスター係数 C がランダムネットワークで期待される値 C_{random} と比べてはるかに大きいことから、3 つのネットワークはスモールワールドである。これらの二字熟語ネットワークはスケールフリーでもあり、3 つの次数分布のベキ指数 γ は近い値（1.14 から 1.16）をとっている。Yamamoto, and Yamazaki (2009) にこれらの詳細が述べられている。

次節以降では、なぜ二字熟語ネットワークがスモールワールドかつスケールフリーにな

表 1: 3 冊の国語辞典から得た二字熟語ネットワークの特性値。平均経路長 l が小さく、クラスター係数 C は同じ規模のランダムネットワークでの値 C_{random} よりもはるかに大きい。次数分布 $P(k)$ のベキ指数 γ の値も記す。

辞書	頂点数	辺数	平均次数	経路長	クラスター係数		ベキ指数
			$\langle k \rangle$	l	C	C_{random}	γ
広辞苑	5458	74617	27.3	3.14	0.138	0.00501	1.14
岩国	3904	32150	16.5	3.31	0.085	0.00424	1.16
三国	3444	28358	16.5	3.32	0.086	0.00483	1.14

[2]自己ループおよび多重辺の分だけ、辺の数は熟語の数より少ない

るのかを考察する。二字熟語の形成を単純化して数理モデルを立てて、モデルが生成するネットワークと実際のネットワークを定量的に比較する。

3　二字熟語ネットワークの特徴を再現するシンプルなモデル

『広辞苑』の二字熟語ネットワークにおいて次数が大きい漢字を順に挙げると大、人、水、山、子、… となっている。これら上位の漢字は、日常で頻繁に使われる重要度の高い漢字である。一方で、ごく限られた特殊な場面でしか使われない漢字も数多くある（ただし、日常生活でそういった漢字に出会うことはほとんどない）。このように、漢字のもつ重要度には大きな格差がある[3]。漢字の重要度は熟語の「つくりやすさ」に直結すると考えるのは自然な仮定である。

そこで、以下では**重要度の高い字ほど熟語をつくりやすい**という特徴を反映した数理モデルを構築する。まずは「重要度」を数値化する必要がある。本研究では漢字の「重要度」のデータとして、文化庁がおこなった『漢字出現頻度数調査 (3)』(2007) を用いた。この調査は出版物の組版データから漢字ごとの使用頻度数を算出したもので、のべ 49072315 の漢字、8576 種類の異なる漢字からなる。「人」が最も大きな出現頻度数 610660 で、一、日、大、年、… がそれに続く。モデルの説明のため、それぞれの漢字に $i = 1, 2, \ldots, 8576$ の番号をつけ、i 番目の漢字の出現頻度数を x_i とする。「漢字 i の重要度が x_i である」と解釈する[4]。

漢字 i と j が辺で結ばれるかどうか（つまり、熟語をつくるかどうか）を確率的に決める。$p_{ij} = cx_ix_j$ という確率で i と j を辺で結ぶことにする（c は定数）。例えば、$c = 10^{-6}$ の場合、$x_i = 1000, x_j = 100$ だとすると、$p_{ij} = 0.1$ であり、頂点 i と j は確率 10% で辺を結び、90% で辺を結ばないことになる。すべての頂点の組に対して辺を結ぶかどうかを判定し、ネットワークをつくっていく。出現頻度数 x_i が大きいほど、他の頂点と辺でつながる確率が高くなっている。

定数 c の値は、モデルが実際の二字熟語ネットワークに近くなるように定める。$c = 0$ のときは辺が全く張られず、すべての頂点がばらばらのネットワークができる。c を充分に大きくすると、どの 2 頂点も辺で結ばれたネットワーク（完全グラフ）ができる（図 2）。つまり、c が大きくなると「辺の密度」も大きくなることが分かる。ネットワークの辺の密度を表すのが**平均次数**$\langle k \rangle$ である。そこで、c を変化させていき、モデルの平均次数が実際の二字熟語ネットワークと一致するような c の値を探す。ここではモデルを広辞苑のネッ

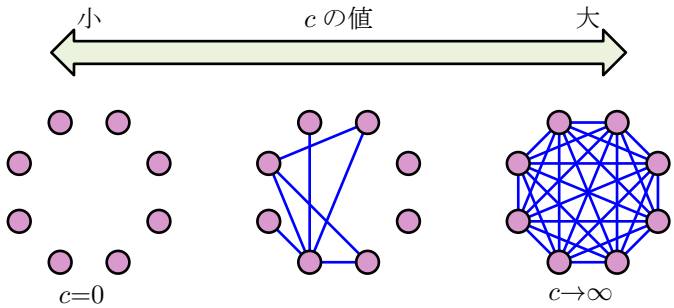


図 2: c の変化によるネットワークの構造変化のイメージ。$c = 0$ のときは辺が張られない（左）。c が充分大きいときは完全グラフとなる（右）。

[3]教育漢字、常用漢字、人名用漢字などの分類は重要度の階層を表しているといえる
[4]抽象的な「重要度の高さ」を「出現頻度数」という数値で定量化している

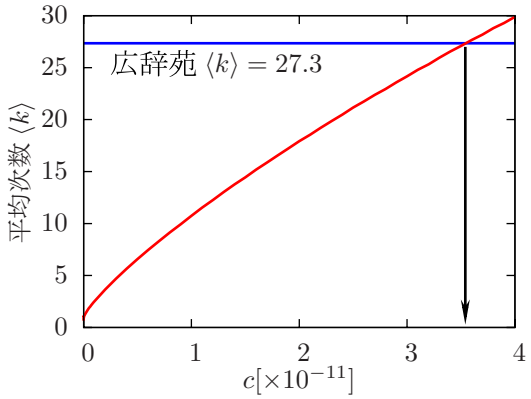


図 3: 確率 p_{ij} の比例定数 c の決定。実線はモデルの数値計算結果、破線は広辞苑のネットワークの値 $\langle k \rangle = 27.3$ である。2 つのグラフの交点から $c = 3.54 \times 10^{-11}$ が得られる。

トワークと比較することにする。c の値に対する平均次数の変化を図 3 に示す。広辞苑のネットワークの平均次数 $\langle k \rangle = 27.3$ と一致するのは $c = 3.54 \times 10^{-11}$ のときであることが分かった。

4　モデルの計算結果

前節で提案したモデルによるネットワークと実際の二字熟語ネットワーク（広辞苑）を比較したのが表 2 である。モデルには確率的な要素が入っているので、試行のたびに異なるネットワークが生成される。表に示した計算結果は 1000 個のネットワークについての平均である。『漢字出現頻度数調査 (3)』では 8576 種類の漢字が挙げられているのに対して、モデルの計算結果では頂点数が半分以上減少している。この原因は、出現頻度が小さな漢字は 1 本でも辺を獲得することが難しく、ネットワークから孤立してしまうためである（孤立した頂点は除外している）。モデルの平均次数 $\langle k \rangle$ が広辞苑と一致しているのは c の決め

表 2: モデルによるネットワークと広辞苑の二字熟語ネットワークの比較。両者の平均経路長 l, クラスター係数 C, およびベキ指数 γ はきわめて近い値である。

	頂点数	辺数	平均次数 $\langle k \rangle$	経路長 l	クラスター係数 C	ベキ指数 γ
モデル	3012	41130	27.3	2.85	0.191	1.12
広辞苑	5458	74617	27.3	3.14	0.138	1.14

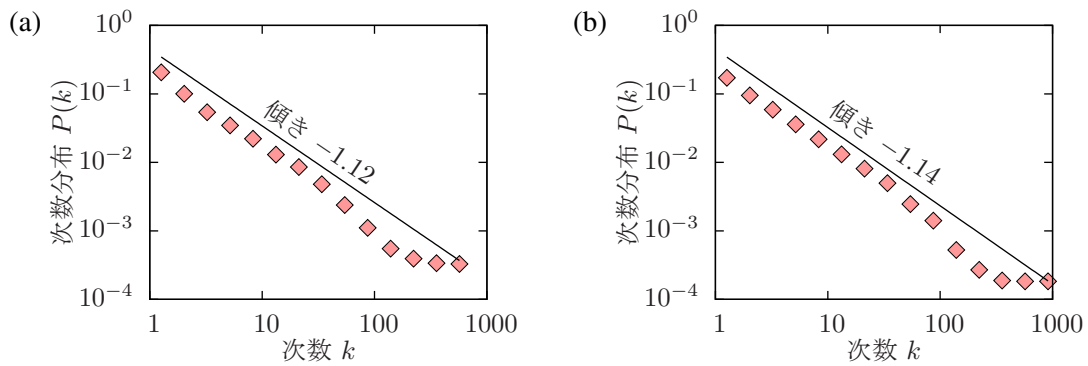


図 4: 次数分布の比較。(a) モデルが生成するネットワークの次数分布。(b) 実際の二字熟語ネットワーク（広辞苑）の次数分布。

方から当然であるが、平均経路長 l, クラスター係数 C, および次数分布のベキ指数 γ もかなり近い値となっていることは注目すべき結果である。他方、頂点数および辺数はともに“モデル”の方が“広辞苑”に比べて小さい。モデルが生成するネットワークは実際のネットワークのミニチュアであると解釈するのがよいだろう。モデルと広辞苑のネットワークの次数分布を図 4 に示す。両者はきわめて似かよっている。

5 検討

Peng, Minett, and Wang (2008) は、中国語の熟語ネットワークの構造を解析した。中国語のネットワークの経路長およびクラスター係数は日本語と近い値であるが、次数分布の指数の値は普通話および広東語でそれぞれ $\gamma = 1.40, 1.49$ となっていて日本語よりも大きい値をとる。γ が小さいほど、少数の頂点に大量の辺が集まる（次数の大きな頂点が出現しやすい）。つまり、日本語の熟語は中国語の熟語よりも重要な漢字に集中する傾向にあり、重要な漢字とそうでない漢字の格差が大きいといえる。

今回のモデルは非常に大雑把なものである。例えば、モデルでは熟語における漢字の順序を定めることができない。また、実際に二字熟語が形成する際には 2 つの漢字の意味や機能など様々な条件が考慮されるはずであるが、モデルでは熟語の形成において漢字の重要度しか考慮していない。本稿のモデルに付加的に条件を課すことで、より詳細な分析に堪えうるネットワークをつくることができるだろう。ただし、スモールワールド性およびスケールフリー性という特徴を再現するためには、本稿のような単純な設定だけで充分であることを強調する。

熟語ネットワークのスケールフリー性の起源として、Peng, Minett, and Wang (2008) は優先的接続の効果を推測している。優先的接続のモデルは、ネットワークのスケールフリー性について最初に論じた Barabási, and Albert (1999) によって提案された。成長するネットワーク（新たな頂点が次々に加わる）において、次数が大きな頂点ほど新たな頂点から辺を受け取りやすいというモデルである[5]。熟語ネットワークの形成過程で優先的接続の効果が作用したのか検証するためには、各々の熟語が「いつ形成されたのか」（各々の漢字が熟語ネットワークに「いつ加わったのか」）を知らねばならない。しかし、熟語が形成された時期の特定は困難であると考えられるので、この仮説は定性的な検証でさえ容易ではないであろう[6]。本研究のモデルではネットワークの成長を取り上げるのは避け、かわりに各漢字に内在する「重要度」に基づいて熟語のつくられ方を規定した[7]。漢字の重要度を表す数値として出版物における出現頻度の調査結果を利用し、モデルと実際のネットワークの定量的な比較をおこなうことができた。

さらに、本稿のモデルに立脚すると、日本語と中国語で γ の値が異なっているのは漢字の「重要度」の分布の仕方が両言語で異なっているためであることが導かれる。中国語における漢字の使用状況の統計を利用してこのことを確かめれば、本研究のモデルの信頼性を評価することも可能であるだろう。

---

[5]「金持ちがさらに金持ちになる (The rich gets richer.)」とたとえられる

[6]Barabási らのモデルから導かれるのは $\gamma = 3$, これを拡張した Dorogovtsev, Mendes, and Samukhin (2000) のモデルにより得られるのは $2 < \gamma < \infty$ である。中国語のネットワークのスケールフリー指数 $\gamma = 1.40$ および 1.49 がこの範囲に入らないことから、熟語ネットワークのスケールフリー性を優先的接続で説明しようとするのはそもそも的外れであると思われる

[7]本研究のモデルはフィットネスモデル (Caldarelli, Capocci, De Los Rios, and Muñoz (2002)) とよばれるモデルの一種といえる

6 まとめ

本研究では、複雑ネットワーク解析の手法を用いて二字熟語の統計的性質を調べた。漢字を頂点、二字熟語を辺とみなすことで、二字熟語は自然にネットワークの形で表現される。3 冊の国語辞典から構成したネットワークはいずれもスモールワールドかつスケールフリーであることが分かった。ネットワークの特徴を定量的に説明するために、重要度が高い漢字ほど二字熟語をつくりやすいという特徴を反映したモデルを構築した。その際、漢字の重要度のデータとして、出版物における漢字の出現頻度数の調査結果を利用した。2 つの漢字の重要度の積に比例する確率で辺を張るというシンプルな設定のモデルであるが、実際のネットワークと近い構造を再現することができた。

謝辞

漢字出現頻度数調査の冊子を提供してくださった武田康宏氏（文化庁）に感謝いたします。

文献

Duncan J. Watts, and Steven H. Strogatz (1998) “Collective dynamics of ‘small-world’ networks”, *Nature* 393, pp. 440–442.

Albert-László Barabási, and Réka Albert (1999) “Emergence of scaling in random networks”, *Science* 286, pp. 509–512.

増田直紀、今野紀雄 (2010)『複雑ネットワーク—基礎から応用まで』近代科学社.

矢久保孝介 (2013)『複雑ネットワークとその構造（連携する数学 4)』共立.

Mark E. J. Newman (2010) *Networks: An Introduction*, Oxford University Press.

Ken Yamamoto, and Yoshihiro Yamazaki (2009) “A network of two-Chinese-character compound words in the Japanese language”, *Physica A* 388:12, pp. 2555–2560.

新村出 編 (1991)『広辞苑　第四版』岩波.

西尾実、岩淵悦太郎、水谷静夫 編 (1992)『岩波国語辞典　第五版』岩波.

見坊豪紀、金田一京助、金田一晴彦、他 編 (1992)『三省堂国語辞典　第 4 版』三省堂.

文化庁 (2007)『漢字出現頻度数調査 (3)』.

Gang Peng, James W. Minett, and William S.-Y. Wang (2008) “The networks of syllables and characters in Chinese”, *J. Quant. Ling.* 15:3, pp. 243–355.

Sergey N. Dorogovtsev, José F. F. Mendes, and Alexander N. Samukhin (2000) “Structure of growing networks with preferential linking”, *Phys. Rev. Lett.* 85:21, pp. 4633–4636.

Guido Caldarelli, Andrea Capocci, Paolo De Los Rios, and Miguel A. Muñoz (2002) “Scale-free networks from varying vertex intrinsic fitness”, *Phys. Rev. Lett.* 89:25, p. 258702 [4 pages].

Kachako におけるコーパスアノテーションの重ね合わせ

狩野 芳伸（科学技術振興機構さきがけ）†

Layering BCCWJ Corpus Annotations in Kachako

Yoshinobu Kano (PRESTO, Japan Science and Technology Agency)

１．はじめに

テキストコーパスの利用にあたっては、テキスト本文の利用のみならず、そこに付与されたアノテーションの利用が重要である。日本語コーパスについては、国立国語研究所を中心に開発されてきた大規模日本語均衡コーパス BCCWJ(前川, 2009)が現在最も大規模なものである。BCCWJ には、国語研と個々の協力研究グループによって形態素・係り受け(浅原ほか, 2011)・述語項構造(小町ほか, 2011)・フレーム構造(小原ほか, 2011)・動詞項構造(竹内ほか, 2012)など様々なアノテーションが付与されてきており、アノテーションという点でも最大規模と考えられる。BCCWJ のアノテーション部分については現在も開発が進められている。本稿では執筆時点の情報に基づいて記述する。

同一テキストに対して異なる種類のアノテーションが重層的に付与されている場合、利用者としてはこれらのアノテーションを統一的に扱えれば理想的である。もっといえば、アノテーションを利用する労力が少ないほどよい。しかし実際には、アノテーションの開発は個々の担当者によって独立して行われ、フォーマットだけでなく基にするコーパステキスト文字列の差異なども存在する。そのためアノテーションの統一的利用にはかなりの労力を必要とするのが現状である。利用者からすれば、データが手元にあり内容も把握しているのに、利用者にとって本質的でない部分で時間を使わねばならないのは時間の浪費である。

本稿では、これらの問題点を解決すべく、異なる種類のアノテーションの重ね合わせを行った我々の研究について報告する。我々は、自然言語処理ツールの利用を自動化するための統合プラットフォーム Kachako を開発してきた。プラットフォーム上では、各種ツールが互換化されて統合されており、ツールの組み合わせや実行、評価が容易に行える。コーパスアノテーションを読み込むリーダーについてもいくつか統合されている。本研究では、Kachako 上で動作するコーパスリーダーの一種として BCCWJ Reader を実装した。

第２節では本研究の背景となった既存研究について述べる。第３節ではアノテーションの重ね合わせの設計とシステムについて述べ、第４節で Kachako システムを通じた応用の可能性を議論し、第５節で将来の展望や応用に触れつつ締めくくる。

２．背景

２.１ 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 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

† kano@nii.ac.jp

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[1]・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)に対応した。それぞれのアノテーションの詳細は３節で詳述する。

２.２ 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。

２.３ 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, 2012a)。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。

`Morpheme`（形態素）, `Segment`（文節等）, `Dependency`（係り受け）など本稿の対象とする基本的なデータ型はすでに Kachako で定義されており、これを利用・拡張することで Kachako の既存コンポーネントと互換にできる。なお、本稿において UIMA の型名や属性名は `TypeName` のような `Courier` フォントを用いて記す。データ型定義の詳細は３節で述べる。

---

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

３．アノテーションの重ね合わせ

本節では、BCCWJ に付与されている各種のアノテーションを重ね合わせるためのデータ型階層の設計（３．１）、テキストと文字位置の正規化（３．２）、それを用いた UIMA 準拠・Kachako 互換の BCCWJ コーパスリーダーの実装（３．３）について述べる。

３.１ データ型階層の設計

３.１.１ BCCWJ DVD 内の文書構造および形態素アノテーション

BCCWJ では、文書の構造を表すため以下のような階層構造を用いている：

<article articleID="OW6X_00000_V001" isWholeArticle="false">
　　<titleBlock>
　　<title>
日本文化発信による国際文化交流(実際にはテキストアノテーションが付与されている)
　　</title>
　　</titleBlock>
　　<cluster>
　　<titleBlock>
　　<title>
文化庁文化交流使事業(同上)
　　</title>
　　</titleBlock>
　　<paragraph>
（本文テキストおよびアノテーション）
　　:

このような文書構造タグ対応して UIMA のデータ型階層を設計した。データ型は共通して `DocumentClassAnnotation` から派生した型を使用した。`DocumentClassAnnotation` は Kachako 標準のトップレベルの文書構造アノテーション型である。また、各タイプは共通して `begin, end` の素性を持っており、それぞれ BCCWJ のデータ中のタグの表層文字列への開始位置、終了位置を格納している。`begin, end` は UIMA 全体で共有されている汎用アノテーション型の `Annotation` 型で定義された素性である。BCCWJ における文書構造タグの入れ子関係は、これらの位置情報により判断できる。`begin, end` はほとんどのアノテーションに共通のため、以下のデータ型の説明では説明を省略する。

BCCWJ における文書構造は(a)サンプル、(b)階層構造、(c)図表、(d)引用、(e)注記、(f)その他１、(g)その他２、(h)文字・表記、(i)可変長タグセットに無いタグ、の各タグにより表現されている。表 1 に各タグに対応して設計した Kachako 互換の UIMA データ型を示す。BCCWJ

表 1 文章構造タグに対応する BCCWJ のタグ名

BCCWJ の各タグ名に `org.kachako.types.japanese.document.` の接頭辞を付与したものを Kachako のデータ型名とし、対応する属性値も素性値として保持する。

sample, sampling, type, article, blockEnd, cluster, titleBlock, list, paragraph, sentence, figureBlock, figure, caption, table, quotation, citation, source, speech, speaker, noteBody, noteBodyInline
abstract, authorsData, contents, profile, rejectedBlock, verse, verseLine, ruby, correction, missingCharacter, enclosedCharacter, cursive, image, superscript, subscript, fraction, delete
br, info, rejectedSpan, substitution, denominator, div, listItem, noteMarker, mergedSample, numerator, orphanedTitle, superSentence, supplement, title, vinculum, webBr, webLine

表 2　SUW（左）および LUW（右、SUW との差分のみ）に対応するデータ型

SUW（短単位）タグ	LUW（長単位）タグ
BCCWJ 属性値 start, end, pos, cType, cForm, lForm, lemma, pron, pronBase, orthBase, orth, goshu, iType, iForm, fType, fForm Kachako データ型: org.kachako.types.japanese.syntactic.morphCeme.SUW 素性 begin --- "start"の値を 10 で割った値を入れる end --- "end"の値を 10 で割った値を入れる pos --- "pos"の値を'-'で分割した 1 番目の要素 detailedPos --- "pos"の値を'-'で分割した 2 番目の要素 posList[0] --- "pos"の値を'-'で分割した 3 番目の要素 posList[1] --- "pos"の値を'-'で分割した 4 番目の要素 conjugateType --- "cType"の内容を格納する surfaceForm --- "orthBase"の内容を格納する readings[0] --- "lForm"の内容を格納する baseForm --- "formBase"の内容を格納する conjugateForm --- "cForm"の内容を格納する pronunciations[0] --- "pron"の内容を格納する pronunciations[1] --- "pronBase"の内容を格納する semanticInformations --- その他の情報を纏めて格納する elements --- 子要素となるタグ情報を保持する	BCCWJ 属性値 B, SL, l_lemma, l_lForm, l_cType, l_cForm, l_pos, l_orthBase Kachako データ型: org.kachako.types.japanese.syntactic.morpheme.LUW 素性 begin --- documentText の開始位置 end --- documentText の終了位置 B --- "B" 'B'=文節境界、'S'=文境界のいずれか SL --- "SL" baseForm --- "l_lema" readings[0] --- "l_Form" conjugateType --- "l_cType" conjugateForm --- "l_cForm" pos --- "l_pos"の値を'-'で分割した 1 番目の要素 detailedPos --- "l_pos"の値を'-'で分割した 2 番目の要素 posList[0] --- "l_pos"の値を'-'で分割した 3 番目の要素 posList[1] --- "l_pos"の値を'-'で分割した 4 番目の要素 surfaceForm --- "l_orthBase" elements --- 子要素を保持する

の各タグ名に org.kachako.types.japanese.bccwj.document.の接頭辞を付与したものを Kachako のデータ型名とし、対応する属性値も素性として定義した。

短単位形態素情報（SUW タグ）の情報は Morpheme の子として定義した SUW に保持する（表 2 左）。Morpheme は形態素一般のデータ型として定義したため、SUW タグの属性に対応する素性の定義が無いものは、semanticInformation 素性にまとめて格納している。

長単位形態素情報（LUW タグ）の情報は、LUW 型に保持する。構造的には SUW と同様であるが、一部属性の項目が異なっているほか、文やサンプル長の属性が追加されている（表 2 右）。

３.１.２ DepPara2 係り受けアノテーション

DepPara2 は DVD が利用可能であることを前提に、DVD データ相当の拡張 CaboCha 形式データへの差分パッチという形式で追加アノテーションを配布している。DepPara2 では単に係り受けを追加するだけでなく、DVD 内の一部アノテーションに対して修正が行われる内容になっている。また、差分とすることで本文テキストのデータが含まれないため、著作権の問題を回避し DepPara2 データのみを一般に無料配布している。

DepPara2 はアノテーションツール茶器(松本ほか, 2010)等で用いられている拡張 CaboCha 形式に基づいた表記になっている。拡張 CaboCha 形式では形態素および係り受け情報を独自のフォーマットで表現する。DepPara2 では茶器にインポートできるよう、以下の a),b)を実行する Ruby スクリプトを提供している(図 1)。

a)　DVD 内の XML ファイルを拡張 CaboCha 形式に変換。係り受け部分はダミーを挿入

b)　拡張 CaboCha 形式間の差分 patch により、a)のダミー係り受け情報を実データに置換
これに対し、我々は実行効率を考慮して patch ファイルと DVD の XML ファイルから直接 Kachako 互換の UIMA 形式への変換を実装し、BCCWJ Reader 内で実行するようにした。ユーザは BCCWJ Reader のオプションに patch ファイルの位置を指定するだけでよい。

DepPara2 における文境界 Sentence は、BCCWJ DVD のものとは異なる可能性があり、どちらが正しいというものでもないため、双方を残しておくのが妥当である。そこで BCCWJ DepPara2 の処理後の文境界情報は JapaneseSentence に保持するようにした。保持する情報は DepPara2 処理後の開始位置と終了位置のみである。

BCCWJ DVD では sentence タグにより文が区切られ EOS が追加されるが、DepPara2 の処理時は DepPara2 のデータで EOS 情報を付与、削除、移動し、その情報を元に Kachako 標準の日本語文境界を表すデータ型 JapaneseSentence を作成している。

表 3　DepPara2 および拡張 CaboCha 形式に対応するデータ型

	Kachako データ型名	素性の説明
DepPara2 独自文境界	org.kachako.types.japanese.syntactic.JapaneseSentence	begin, end
DepPara2 独自文節境界	org.kachako.types.japanese.syntactic.CabochaSegment	begin, end
文節係り受け関係（ラベル付き）	org.kachako.types.japanese.syntactic.Dependency	begin --- 係り元の文節の開始位置 end --- 係り元の文節の終了位置 source --- 係り元の文節 target --- 係り先の文節 label --- ラベル
セグメント情報	org.kachako.types.japanese.syntactic.CabochaSegment	label --- ラベル begin --- 開始位置 end --- 終了位置
グループ情報	org.kachako.types.japanese.syntactic.CabochaGroup	label --- ラベル begin --- 開始位置 end --- 終了位置
リンク情報	org.kachako.types.japanese.syntactic.CabochaLink	label --- ラベル begin --- 開始位置 end --- 終了位置 source --- 開始セグメントの ID target --- 終了セグメントの ID

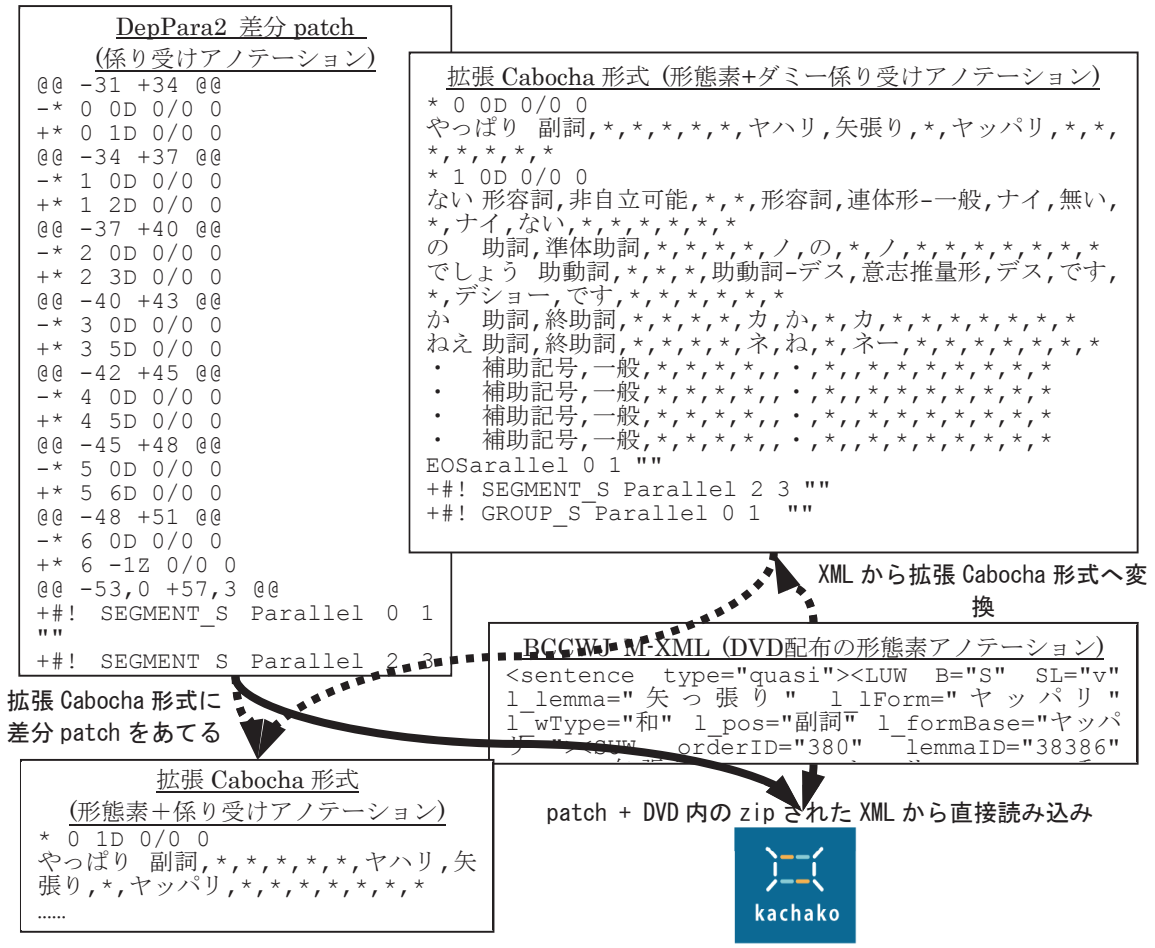


図１　DepPara2 のデータ変換の流れ。点線は DepPara2 提供のスクリプトで拡張 Cabocha 形式を生成する場合、実線は Kachako の BCCWJ Reader 内部での変換の流れを示す。

BCCWJ DepPara2 での文節情報は、CabochaSegment に保持する。BCCWJ DepPara2 では使用していないが、信頼値が入っている場合には Kachako 標準のデータ型である AnnotationMetadata に保存する。

DepPara2 の patch データが無い場合の文節境界は、BCCWJ DVD データの LUW タグの"B"の値に"S"あるいは"B"が入っている場合、そこで文節境界として決定している。

文節係り受け情報は、Kachako の汎用係り受け関係用データ型である Dependency に保持する。セグメント情報(SEGMENT_S)は org.kachako.types.japanese.syntactic.ExtraSegment に保持する。グループの情報(GROUP_S)は、CabochaGroup に保持する。またリンク情報(LINK_S)については、CabochaLink に保持する。これらのデータ型は、茶器等で用いられている（拡張）CaboCha 形式を表現するために定義された汎用データ型である。表 3 に各データ型の一覧を示す。

３.１.３ 述語項構造および並列関係アノテーション

述語項構造アノテーションは、BCCWJ 内のいくつかの文書について付与されている。述語を Pred 型、名詞句を Np 型とし、ガ格・ヲ格・ニ格の項構造のリンクを保持するようにした(図 2)。また、並列関係を Coref 型とし、並列関係のリンクも保持するようにした(図 3)。

３.１.４ 日本語フレームネットアノテーション

日本語フレームネットアノテーションは、BCCWJ 内のいくつかの文書の先頭部分にのみ付与されている。これは旧版の BCCWJ DVD に対し付与されたものを引き継いでいるた

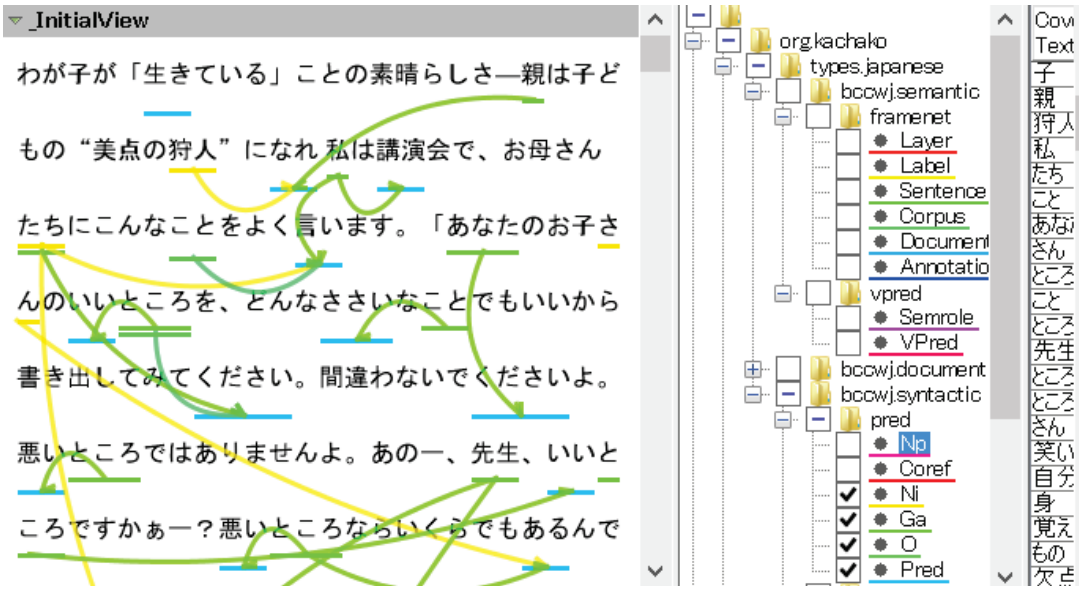


図 2　述語項構造を Kachako の概観ビューで表示した例。水色が述語、黄色がニ格、濃緑がヲ格、緑がガ格に対応している。右側には表示させたいアノテーション種を選択するリストがあり、任意の種類のアノテーションを同時に表示させることができる。

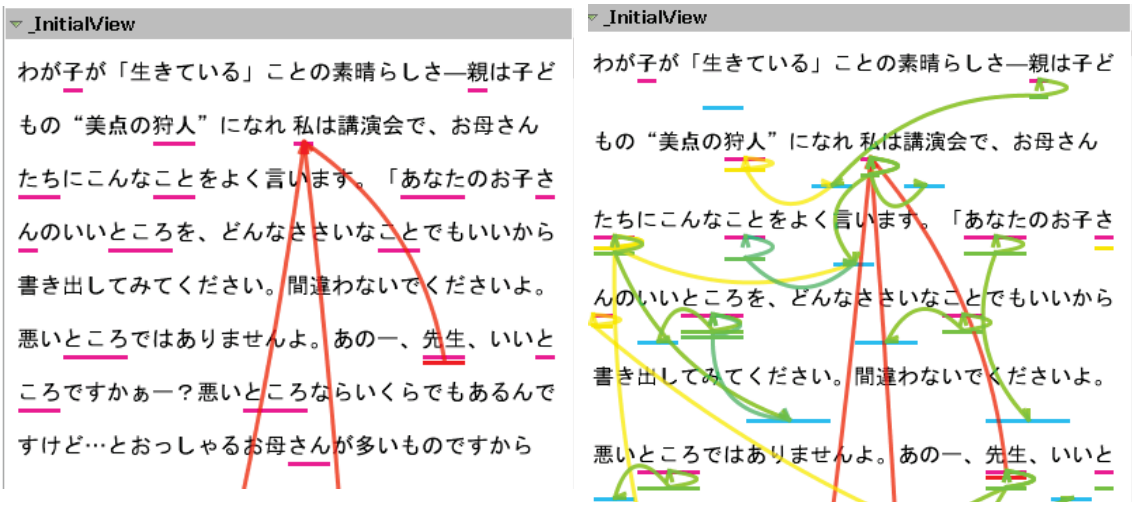


図 3　並列構造を Kachako の概観ビューで表示した例。赤が並列関係を示している。左は並列構造のみ、右は述語項構造を同時に表示した例。

め、そのままではテキスト位置が一致しない。BCCWJ Reader ではテキストが一致しなくても文字数が一致する場合はそのまま読み込んで重ねるようにし、文字数が一致しないデータは読み込まないものとした。

日本語フレームネットでは `Sentence`、`AnnotationSet`、`Layer`、`Label` の順に、親から子に入れ子になっており、親はそれぞれ複数の子を保持しうる。また、属性値としてフレーム名や役割名などが格納されている。属性値の意味については、アノテーション開発者の提供する情報を参照する必要があり、Kachako では単に文字列として扱っている。データ型には `org.kachako.types.japanese.bccwj.document.semantic.framenet` の接頭辞を付与した。

３.１.５ 動詞項構造アノテーション

動詞項構造のアノテーションは、BCCWJ 内のいくつかの文書のうち、一部の動詞について付与されている。アノテーションは改行・タブ区切り形式で保存されており、各行に中心となる動詞と動詞のとる項構造のリストが繰り返し記述され、行内には属性値がタブ区切りで格納されている。動詞を `VPred`、項構造を `Semrole` として定義し、`VPred` から対応する `Semrole` にリンクを張るようにした。これらのデータ型には接頭辞として `org.kachako.types.japanese.bccwj.document.semantic.vpred` を付与した。

３.２ テキスト位置の処理

BCCWJ の形態素単位にはいくつか種類があり、アノテーションの重ね合わせのためには基になるテキスト上でのアノテーションの開始位置、終了位置の計算を適切に行いテキストが重なるようにする必要がある。

SUW 以下のアノテーション位置については、BCCWJ のデータの開始位置・終了位置は、基本的には SUW のタグにある start/end の値を元に計算を行っているが、SUW より下位のタグについては開始位置・終了位置が無いため、XML データ中のテキストの長さから開始位置・終了位置を求めている。SUW の下位タグとしてオフセット計算を行っているものは correction, ruby, subscript, superScript の 4 つのタグである。

BCCWJ では、アラビア数字による数値の表層文字列を漢数字に置き換えたうえで、もともとの表記を NumTrans タグとして追記している。NumTrans タグの情報は `org.kachako.types.japanese.document.NumTrans` に保持する。NumTrans は元のタグ同様 `originalText` 素性を持つ。

BCCWJ の NumTrans タグには現在不具合があり、SUW より上位のタグであっても NumTrans タグのタグ以下では start/end の値に正しくない値が入っているため、タグ開始位置として使用すると不正な値になってしまう。このため NumTrans より下にある SUW は、start/end の値は使用せず、実際の XML 中のテキスト要素の文字列の長さを元に UIMA 側アノテーションの `begin/end` の値を設定している。また、NumTrans のタグより後では、誤った値が SUW の start/end に入っているため、以降のアノテーションに対応する `begin/end` には、この NumTrans タグで誤った値の差分を計算して設定している。

３.３ BCCWJ Reader コンポーネントの設定と処理

コーパスとアノテーションを読み込み、上記のデータ型階層互換かつ UIMA の形式に変換するツールは、BCCWJ Reader という名称の Kachako 互換 UIMA コンポーネントとして実装した。単にデータ型互換というだけでなく、Java 実装とパッケージングにより自動インストール・自動実行に対応したコンポーネントになっており、Java 7 の稼働するマシンであればどこでも自動実行できる。表 4 に BCCWJ データの読み込みと処理にかかる時間を計測した結果を示す。ここで行った処理には、zip ファイルの展開・XML 解析・各種重ね合わせ処理と UIMA 形式への変換が含まれる。

BCCWJ Reader の実行時オプションは以下のとおりである。

まず、`Directory` に BCCWJ DVD あるいはそのコピーの格納されている先頭ディレクトリを指定する。DVD データについては、ユーザ指定が必須のオプションはこれだけである。BCCWJ Reader は実行時に `Directory` 下の M-XML 以下のフォルダに含まれる zip 形式で圧縮されたデータをその場で展開しつつ読み込む。

`InputEncoding` はデフォルトで UTF-8 である。DVD で配布されているデータは UTF-8 形式のため通常は変更の必要がない。

DepPara2 のデータを読み込みたい場合は、以下のオプション群を適宜指定する。`AddDepPara2Data` は DepPara2 で使用する `CabochaSegment` 等の各種データをとりこむか否かを指定する。`DepPara2PatchFile` には DepPara2 サイトで配布されている patch 形式のデータを指定する。`OutputOnlyDepPara2Data` には BCCWJ DepPara2 の patch ファイルに patch 情報が存在するデータのみを対象にするか否かを指定する。この項目がチェックされていない場合、BCCWJ DVD にあるデータで DepPara2 の patch 情報が無い

ものは BCCWJ DVD 情報から生成した CabochaSegment 情報を出力する。

述語項構造・日本語フレームネット・動詞項構造などのアノテーションを読み込む場合も、同様にアノテーションを記述したファイルの指定が必要である。なおこれらのアノテーションは現時点では一般に公開されておらず、今後詳細が変更される可能性もあるため、本稿では詳細には触れない。

以下は Kachako システムで実行するコンポーネントに共通で設定可能なパラメータである。KachakoSystem_Vmargs には実行時に Java VM に渡すパラメータを指定できる。デフォルトは-Xss16m (スタックメモリの使用値を 16MB に指定)である。KachakoSystem_IvyUrls には BCCWJ Reader 実行時にコンポーネント実行に必要なファイルをダウンロードする URL が指定されている。デフォルトは Kachako の配布サイトになっており変更の必要はない。

表 4 実行時間の計測結果

データ	ファイルサイズ	処理時間(分)	平均処理速度(KB/sec)
LB.zip	1.31GB	366	357.9
OB.zip	163MB	47	346.8
OC.zip	546MB	132	413.6
OL.zip	35.9MB	13	276.1
OM.zip	198MB	463	42.7
OP.zip	159MB	109	145.8
OT.zip	39.3MB	12	327.5
OV.zip	10MB	2	500.0
OW.zip	191.1MB	58	329.4
OY.zip	525.7MB	156	336.9
PB.zip	1.22GB	357	341.7
PM.zip	202.7MB	56	361.9
PN.zip	61.8MB	11	561.8
合計	4.66GB	1782	261.5

４．Kachako システムを通じた応用の可能性

Kachako プラットフォームは現在開発を継続しており、視覚化機能以外にも多様な機能を提供する予定である。Kachako の中核的な機能は、ツールの組合せと自動実行である。自動実行では任意のマシンに対しリモート自動インストール・自動実行可能なうえ、Kachako の Hadoop-UIMA ブリッジ機構と Java・Hadoop 全自動インストール機構により大規模並列処理も自動実行できる。待ち受けウェブサービスとして自動展開することも可能である。Kachako プラットフォーム側で機能が追加されれば BCCWJ Reader をはじめとするコンポーネント側では何も変更しなくとも新たな機能が利用できるようになる。これは標準化・互換化を考慮して実装を行ったことの大きな利点といえる。

そうした対応ツール、すなわち Kachako に互換化統合済みで利用可能なコンポーネントは、現在のところ日本語・英語のツールを中心に多数実装済である。応用的なツールとしては、Javelin および Minerva の二つの日本語質問応答システムをコンポーネント分割し互換化実装した(狩野, 2012b)。ほかにも互換コンポーネント化実装が進行中・テスト中の対応ツールとして、音声認識の Sphinx-4(Walker et al., 2004)や NTCIR MedNLP の医療言語情報処理のもの(狩野, 2013)などがある。これらすべてのツールが全自動で組み合わせ・実行・結果の重ね合わせが可能であり、そこに BCCWJ アノテーションが加わったことによるメリットは大きいと考えられる。

５．おわりに

本稿では、BCCWJ に付与されたアノテーションを重ね合わせ、Kachako との標準化・互換化により統一的に扱えるようする設計と実装について述べた。重ね合わせのもっとも直接的な利用例として、Kachako の汎用視覚化機能による表示を紹介した。Kachako の他の機能や、対応ツール群との組み合わせにより、さまざまな応用が期待できる。

将来の課題として、今後各研究グループから公開される予定の、未対応の種類のアノテーションへの対応があげられる。アノテーションは随時更新される可能性があり、配布形態も今後決まるものと思われるので、適宜対応が必要である。また、テキストと重層的なアノテ

ーションを同時に検索できるような検索機能の実装を検討したいと考えている。

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

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.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., et al. (2004). *Sphinx-4: A Flexible Open Source Framework for Speech Recognition*. Mountain View, CA, USA: Sun Microsystems, Inc.

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).

松本裕治, 浅原正幸, 岩立将和, 森田敏生. (2010). 形態素・係り受け解析済みコーパス管理・検索ツール「茶器」. 情報処理学会研究報告. 自然言語処理研究会報告, *2010*(18), pp.1–6.

浅原正幸, 岩立将和, 松本裕治. (2011). BCCWJ コアデータへの係り受け・並列構造のアノテーション. *日本語コーパス平成 22 年度公開ワークショップ*. pp. 317–324.

狩野芳伸. (2012a). 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）*.

狩野芳伸. (2012b). 統合研究基盤：質問応答システムの互換コンポーネント化による再利用性向上と開発自動化支援(<特集>ロボットは東大に入れるか?). 人工知能学会誌, *27*(5), pp.492–495.

狩野芳伸. (2013). 医療言語処理ツールとコーパスの互換化: Kachako をベースとした大規模処理に向けて. *言語処理学会第 19 回年次大会* (pp. 81–82).

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).

ポスター発表（2） グループA

3月7日（金） 13:00～14:00

拡張 CaboCha フォーマットの仕様拡張

松吉 俊 (山梨大学大学院医学工学総合研究部)
浅原 正幸 (国立国語研究所コーパス開発センター)
飯田 龍 (東京工業大学大学院情報理工学研究科)
森田 敏生 (総和技研)

RFC: Requirements in Extended CaboCha Formats

Suguru Matsuyoshi (Interdisciplinary Grad. School of Med. and Eng., Univ. of Yamanashi)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Ryu Iida (Dept. of Computer Science, TITECH)
Toshio Morita (Sowa Research Co., Ltd.)

1. はじめに

各研究機関において「現代日本語書き言葉均衡コーパス」(BCCWJ) に対する様々なレベルの言語情報のアノテーションが進められている。形態論情報が人手により修正されているコアデータを中心に、共通のサンプルに複数のアノテーションが施され、これらを統合して分析・モデル化することが可能になりつつある。実際に分析・モデル化を行うためにはこれらを統合的に記述できる形式が必要になる。コーパス管理ツール『ChaKi.NET』の入出力で利用されている拡張 CaboCha フォーマット[*1]は意味的に異なるアノテーションを共通の形式で表現することを目的とし、形態論情報と係り受け情報に対する追加のアノテーションを表現するために、部分文字列を表現する SEGMENT、SEGMENT 間の有向関係を表現する LINK、SEGMENT 間の同値類を表現する GROUP を規定している。しかし、実際に重ね合わせ作業を進めてみると、アノテーション要素に属性情報を付与できない、複数のアノテーション間で名前の衝突が起きるなどの問題が発生した。そこで本研究ではこれらを解決する仕様拡張について示し、実際のアノテーションの形式表現について紹介する。

2. 現在の拡張 CaboCha フォーマットの仕様と問題点

拡張 CaboCha フォーマットは、形態論情報・文節・文節係り受けを表現する日本語係り受け解析器 CaboCha の出力フォーマット (`-f1` オプション) の出力に対して、**部分文字列の範囲を表現するセグメント**と**セグメント間の有向関係を表すリンク**と**セグメントの同値類（推移律を持つ無向リンク）を表すグループ**を表現できるようにしたものである。前提として拡張 CaboCha フォーマットはアノテーションとして形式のみを規定し、どのような情報を格納するかは自由に規定してよい。表 1 に従来の拡張 CaboCha フォーマットの概要を示す。

以下、各要素について解説する。`##` は大域的な付加情報を表現するコメント行を示す。

[*1] `http://sourceforge.jp/projects/chaki/wiki/`拡張 CaboCha フォーマットへのエクスポート を参照。

表 1　従来の拡張 CaboCha フォーマットの概要

タグ	摘要
##	コメント記号
#! DOC <id>	文書開始タグ (ID の宣言)
#! DOCID\t<id>\t<Bibinfo>	文書単位の書誌情報
#! SEGMENT_S <TagName> <StartLPos> <EndLPos> "<Comments>"	文内で閉じたセグメント
#! SEGMENT <TagName> <StartGPos> <EndGPos> "<Comments>"	文間の関係を指定するセグメント
#! LINK_S <TagName> <FromSegSNo> <EndSegSNo> "<Comments>"	文内有向リンク
#! LINK <TagName> <FromSegNo> <EndSegNo> "<Comments>"	文間有向リンク
#! GROUP_S <TagName> <SegSNo> <SegS> ... "<Comments>"	文内同値類
#! GROUP <TagName> <SegNo> <SegNo> ... "<Comments>"	文間同値類

DOC は文書単位に <ID> 相当の自然数 (1-origin) を割り当てる。DOCID は文書単位の書誌情報を格納する。書誌情報を表す<Bibinfo>は XML フラグメントで与えることとする。先に述べた 3 種類のアノテーションは識別子として <TagName> を持つ SEGMENT(_S)、LINK(_S)、GROUP(_S) により表現する。セグメント間の関係はセグメント番号に対するリファレンスにより表現し、セグメント番号は以下に述べる範囲で上から順に 0-origin で付与する。_S をつける場合は、関連するアノテーションが全て文内で閉じている場合に利用し、後で示すオフセット値 (<StartLPOS>: セグメント開始位置、<EndLPos>：セグメント終了位置) やリファレンスセグメント番号 (<FromSegSNo>：リンク元、<ToSegSNo>：リンク先、<SegSNo>：同値類）も文内で閉じたものを割り当てる。_S をつけない場合は、<DOC>行で分割される範囲で閉じたオフセット値 (<StartGPOS>: セグメント開始位置、<EndGPos>：セグメント終了位置) やリファレンスセグメント番号 (<FromSegNo>：リンク元、<ToSegNo>：リンク先、<SegNo>：同値類) を割り当てる。いずれのオフセット値も 0-origin の文字位置で表現する。以下にオフセット値の計算例を示す。

オフセット値の計算

```
0 駒 1 と 2 盤 3 は 4 も 5 っ 6 て 7 い 8 ま 9 せ 10 ん 11 。12
-----------------------------------------------------------------
#! SEGMENT_S Parallel 0 1 "駒"
#! SEGMENT_S Parallel 2 3 "盤"
```

<Comments> には局所的な付加情報を表現するコメントを記述する。本稿では便宜的に SEGMENT(_S) の<Comments>に表層文字列を記述する。

以上の現在の拡張 CaboCha フォーマットに対して今回扱いたい問題が二つある。

一つ目は属性情報である。XML などで表現されるアノテーションは XML タグに属性情報が表現できる。しかしながら現状の拡張 CaboCha 形式は属性情報を表現する手立てがない。

二つ目は名前空間である。形態論情報・文節・文節係り受けが付与されたうえに多様なアノテーションを重ね合わせる際にタグ名が衝突する事態が起きうる。その際に衝突を回避する手立てがない。

次節ではこの二つの問題を解決するための仕様拡張案について示す。

3. 仕様拡張案：属性情報と名前空間

前節で示した問題点に対処するために、新たに次の二つの機構を導入する。

一つ目は SEGMENT(_S)、LINK(_S)、GROUP(_S) に対する属性情報を表現するための ATTR の導入である。代表して以下に SEGMENT_S に対する属性付与方法を示す。

属性付与

```
#! SEGMENT_S <TagName> <StartLPos> <EndLPos> "<Comments>"
#! ATTR <Key1> "<Value1>"
#! ATTR <Key2> "<Value2>"
#! ATTR <Key3> "<Value3>"
```

ATTR は、その上に存在する一番近い要素に対する属性と解釈される。1 つの要素に対して付与できる属性の数に制限はない。拡張 CaboCha フォーマット形式のファイルを可視化するツールは、要素の属性を綺麗な形で表示できることが望ましい。

二つ目は <TagName> に対して名前空間<ns> を付与し名前衝突を回避する機構の導入である。次の記法により、要素名と属性名に対して、名前空間を記述する。

名前空間定義

```
#! SEGMENT_S <ns>:<TagName> <StartLPos> <EndLPos> "<Comments>"
#! ATTR <ns>:<Key> "<Value>"
```

名前空間を使用する目的は情報提供者を明示し、複数のアノテーションを重ね合わせた時に要素名が衝突するのを防ぐためである。例えば、pred(Pred) という名前の要素が二つ以上のアノテーションで異なる意味で用いられている場合、bccwj-pth:pred と bccwj-pas:Pred と表記することにより、これらを容易に区別することができる。

4. 仕様拡張案の使用例

表 2 に各機関で進められている BCCWJ に対するアノテーションと表現方法を示す。

表 2　BCCWJ に対するアノテーションと表現方法

アノテーション	開発者	名前空間	セグメント	リンク	グループ
並列・同格構造	奈良先端大・国語研	-	○	×	○
時間情報	国語研	bccwj-timebank	○	○	×
拡張固有表現	東工大	bccwj-ene	○	×	×
語義	東工大	bccwj-wsd	○	×	×
項構造シソーラス	岡山大	bccwj-pth	○	○	×
れる・られる	国語研	bccwj-auxv	○	×	×
日本語フレームネット	慶應大	bccwj-jfn	○	×	○
拡張モダリティ	奈良先端大・東北大	bccwj-eme	○	×	×
否定の焦点	山梨大	bccwj-wsb	○	○	×
述語項構造	奈良先端大	bccwj-pas	○	○	○

以下では様々な機関で実施されている BCCWJ に対するアノテーションを例に仕様拡張案の使用例を示す。セグメントとリンクとグループに属性情報を付与することで全てのアノテー

ションが表現可能である。また、10 種類のアノテーションの間で名前の衝突は頻繁に起こっており、pred(Pred)、event(Event)、type、class などが複数のアノテーションで利用されていた。新たに名前空間の機構を与えることにより衝突を回避することが可能になる。尚、いくつかのアノテーションについては、元データに含まれていた最終更新日時や表示時の色指定などの属性情報など、言語情報に直接関係しないものを省略して示す。

4.1 アノテーション事例

■並列構造・同格構造　並列構造・同格構造は奈良先端大・国語研において係り受けとともにアノテーションが進められている (浅原・松本 (2013))。並列構造・同格構造は文内に閉じており、同値類として三つ以上のセグメントに対しても規定できるため SEGMENT_S、GROUP_S により表現する。係り受けと一体となっているため名前空間は定義しない。

例: OC01_00001（サンプル名を表す; 以下同様）

```
* 0 1D 0/0 0
駒 名詞,普通名詞,一般,*,*,*,コマ,駒,駒
と 助詞,格助詞,*,*,*,*,ト,と,と
* 1 2D 0/0 0
盤 名詞,普通名詞,一般,*,*,*,バン,盤,盤
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 2 -1D 0/0 0
持っ 動詞,一般,*,*,五段-タ行,連用形-促音便,モツ,持つ,持っ
て 助詞,接続助詞,*,*,*,*,テ,て,て
い 動詞,非自立可能,*,*,上一段-ア行,連用形-一般,イル,居る,い
ませ 助動詞,*,*,*,助動詞-マス,未然形-一般,マス,ます,ませ
ん 助動詞,*,*,*,助動詞-ヌ,終止形-撥音便,ズ,ず,ん
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S Parallel 0 1 "駒"
#! SEGMENT_S Parallel 2 3 "盤"
#! GROUP_S Parallel 0 1 ""
```

■時間情報　時間情報のアノテーションは国語研でアノテーションが進められた (小西ほか (2013), 保田ほか (2013))。時間情報表現 (TIMEX3)・事象表現 (EVENT) を SEGMENT(_S) で切り出し、時間的順序関係 (TLINK) を LINK(_S) で表現する。名前空間として bccwj-timebank を用いる。

例: PN4b_00001 (bccwj-timebank)

```
* 0 2D 0/0 0
　空白,*,*,*,*,*,,　,
「 補助記号,括弧開,*,*,*,*,*,,「,「
十 名詞,数詞,*,*,*,*,ジュウ,十,十
一 名詞,数詞,*,*,*,*,イチ,一,一
月 名詞,普通名詞,助数詞可能,*,*,*,ガツ,月,月
に 助詞,格助詞,*,*,*,*,ニ,に,に
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 1 2D 0/0 0
ジョージ 名詞,固有名詞,人名,一般,*,*,ジョージ,ジョージ,ジョージ
・ 補助記号,一般,*,*,*,*,,・,・
ブッシュ 名詞,固有名詞,人名,一般,*,*,ブッシュ,ブッシュ,ブッシュ
を 助詞,格助詞,*,*,*,*,ヲ,を,を
* 2 -1D 0/0 0
倒す 動詞,一般,*,*,五段-サ行,終止形-一般,タオス,倒す,倒す
ぞ 助詞,終助詞,*,*,*,*,ゾ,ぞ,ぞ
」 補助記号,括弧閉,*,*,*,*,,」,」
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-timebank:TIMEX3 2 5 "１１月"
#! ATTR bccwj-timebank:tid "t4"
#! ATTR bccwj-timebank:type "DATE"
#! ATTR bccwj-timebank:definite "FALSE"
#! ATTR bccwj-timebank:valueFromSurface "XXXX-11"
#! ATTR bccwj-timebank:value "2004-11"
#! SEGMENT_S bccwj-timebank:EVENT 17 19 "倒す"
#! ATTR bccwj-timebank:class "OCCURRENCE"
#! ATTR bccwj-timebank:eid "e8"
#! ATTR bccwj-timebank:modality "UNDEF"
#! ATTR bccwj-timebank:tense "UNDEF"
#! ATTR bccwj-timebank:eventID "e8"
#! ATTR bccwj-timebank:aspect "UNDEF"
#! ATTR bccwj-timebank:cardinality "UNDEF"
#! ATTR bccwj-timebank:polarity "UNDEF"
#! ATTR bccwj-timebank:signalID "UNDEF"
#! ATTR bccwj-timebank:eiid "ei8"
#! ATTR bccwj-timebank:nf_morph "UNDEF"
#! LINK_S bccwj-timebank:TLINK 0 1 ""
#! ATTR bccwj-timebank:task "T2E"
#! ATTR bccwj-timebank:relatedToEventInstance "ei8"
#! ATTR bccwj-timebank:relTypeA "contains"
#! ATTR bccwj-timebank:relTypeB "contains"
#! ATTR bccwj-timebank:relTypeC "contains"
#! ATTR bccwj-timebank:timeID "t4"
EOS
```

■拡張固有表現　拡張固有表現は東工大においてアノテーションされた (橋本・中村 (2010))。固有表現の階層構造を属性値に level1-4 というキーを与え表現する。名前空間として bccwj-ene を用いる。

例: OC11_00001 (bccwj-ene)

```
* 0 3D 0/0 0
ラーメン 名詞,普通名詞,一般,*,*,*,ラーメン,ラーメン,ラーメン
花月 名詞,普通名詞,一般,*,*,*,カゲツ,花月,花月
の 助詞,格助詞,*,*,*,*,ノ,の,の
* 1 3D 0/0 0
ホッピー 名詞,固有名詞,一般,*,*,*,ホッピー,ホッピー,ホッピー
の 助詞,格助詞,*,*,*,*,ノ,の,の
* 2 3D 0/0 0
百 名詞,数詞,*,*,*,*,ヒャク,百,百
円 名詞,普通名詞,助数詞可能,*,*,*,エン,円,円
で 助詞,格助詞,*,*,*,*,デ,で,で
* 3 5D 0/0 0
焼酎 名詞,普通名詞,一般,*,*,*,ショウチュウ,焼酎,焼酎
追加 名詞,普通名詞,サ変可能,*,*,*,ツイカ,追加,追加
サービス 名詞,普通名詞,サ変可能,*,*,*,サービス,サービス,サービス
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 4 5D 0/0 0
何 名詞,数詞,*,*,*,*,ナン,何,何
回 名詞,普通名詞,助数詞可能,*,*,*,カイ,回,回
まで 助詞,副助詞,*,*,*,*,マデ,まで,まで
* 5 -1D 0/0 0
頼める 動詞,一般,*,*,下一段-マ行,連体形-一般,タノム,頼む,頼める
の 助詞,準体助詞,*,*,*,*,ノ,の,の
でしょう 助動詞,*,*,*,助動詞-デス,意志推量形,デス,です,でしょう
か 助詞,終助詞,*,*,*,*,カ,か,か
？ 補助記号,句点,*,*,*,*,,？,？
#! SEGMENT_S bccwj-ene:Name 0 6 "ラーメン花月"
#! ATTR bccwj-ene:level1 "名前"
#! ATTR bccwj-ene:level2 "施設名"
#! ATTR bccwj-ene:level3 "ＧＯＥ"
#! ATTR bccwj-ene:level4 "ＧＯＥ_その他"
#! ATTR bccwj-ene:ENE "GOE_Other"
#! SEGMENT_S bccwj-ene:Name 7 11 "ホッピー"
#! ATTR bccwj-ene:level1 "名前"
#! ATTR bccwj-ene:level2 "製品名"
#! ATTR bccwj-ene:level3 "食べ物名"
#! ATTR bccwj-ene:level4 "食べ物名_その他"
#! ATTR bccwj-ene:ENE "Food_Other"
#! SEGMENT_S bccwj-ene:Numex 12 14 "１００円"
#! ATTR bccwj-ene:level1 "数値表現"
#! ATTR bccwj-ene:level2 "金額表現"
#! ATTR bccwj-ene:ENE "Money"
```

■語義　語義は東工大においてアノテーションされた (奥村ほか (2013))。SENSEVAL-2 Japanese WSD タスクとして語義曖昧性解消のベンチマークデータとして利用された (奥村・白井 (2009))。セグメントとして文字列を切り出し、属性として語義を表す ID を付与する。ID は岩波国語辞典の語釈文に対応しており、これを属性値 explanation として展開する。名前空間として bccwj-wsd を用いる。

例: PN1c_00001 (bccwj-wsd)

```
* 0 1D 0/0 0
キャスター 名詞,普通名詞,一般,*,*,*,キャスター,キャスター,キャスター
　空白,*,*,*,*,*,,　,
* 1 -1D 0/0 0
蓮舫 名詞,固有名詞,人名,一般,*,*,レンホウ,レンホウ,蓮舫
さん 接尾辞,名詞的,一般,*,*,*,サン,さん,さん
#! SEGMENT_S bccwj-wsd:wsd 0 5 "キャスター"
#! ATTR bccwj-wsd:sense "11335-0-0-1-0"
#! ATTR bccwj-wsd:explanation "&lt;1&gt; ●ニュースキャスター。"
EOS
```

■項構造シソーラス　項構造シソーラスの情報は岡山大学でアノテーションされている (竹内・上野 (2013))。述語と項をセグメントとして切り出し、述語項関係をリンクで表現する。名前空間として bccwj-pth を用いる。本アノテーションは複数人によりアノテーションされており、アノテータを識別するための userid が属性値に含まれている。名前空間としてはアノテーションを識別するが、複数のアノテータを識別する機構がなく属性値を利用する。

例: PB12_00001 (bccwj-pth)

```
* 0 2D 0/0 0
単に 副詞,*,*,*,*,*,タンニ,単に,単に
* 1 2D 0/0 0
コーヒー 名詞,普通名詞,一般,*,*,*,コーヒー,コーヒー,コーヒー
と 助詞,格助詞,*,*,*,*,ト,と,と
* 2 4D 0/0 0
間違え 動詞,非自立可能,*,*,下一段-ア行,連用形-一般,マチガエル,間違える,間違え
```

```
て 助詞,接続助詞,*,*,*,*,テ,て,て
* 3 4D 0/0 0
紅茶 名詞,普通名詞,一般,*,*,*,コウチャ,紅茶,紅茶
を 助詞,格助詞,*,*,*,*,ヲ,を,を
* 4 7D 0/0 0
飲ん 動詞,一般,*,*,五段-マ行,連用形-撥音便,ノム,飲む,飲ん
だ 助動詞,*,*,*,助動詞-タ,連体形-一般,タ,た,だ
だけ 助詞,副助詞,*,*,*,*,ダケ,だけ,だけ
で 助動詞,*,*,*,助動詞-ダ,連用形-一般,ダ,だ,で
、 補助記号,読点,*,*,*,*,,、,、
* 5 6D 0/0 0
衝撃 名詞,普通名詞,一般,*,*,*,ショウゲキ,衝撃,衝撃
的 接尾辞,形状詞的,*,*,*,*,テキ,的,的
な 助動詞,*,*,*,助動詞-ダ,連体形-一般,ダ,だ,な
* 6 7D 0/0 0
味覚 名詞,普通名詞,一般,*,*,*,ミカク,味覚,味覚
体験 名詞,普通名詞,サ変可能,*,*,*,タイケン,体験,体験
に 助詞,格助詞,*,*,*,*,ニ,に,に
* 7 -1D 0/0 0
なっ 動詞,非自立可能,*,*,五段-ラ行,連用形-促音便,ナル,成る,なっ
て 助詞,接続助詞,*,*,*,*,テ,て,て
しまう 動詞,非自立可能,*,*,五段-ワア行,終止形-一般,シマウ,仕舞う,しまう
と 助詞,格助詞,*,*,*,*,ト,と,と
は 助詞,係助詞,*,*,*,*,ハ,は,は
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-pth:semrole 0 2 "単に"
#! SEGMENT_S bccwj-pth:semrole 2 11 "コーヒーと間違えて"
#! SEGMENT_S bccwj-pth:pred 7 9 "間違え"
#! SEGMENT_S bccwj-pth:semrole 11 14 "紅茶を"
#! SEGMENT_S bccwj-pth:pred 14 16 "飲ん"
#! LINK_S bccwj-pth:semrole6 4 0 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "6"
#! LINK_S bccwj-pth:semrole9 4 0 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "9"
#! LINK_S bccwj-pth:semrole6 4 1 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "6"
#! LINK_S bccwj-pth:semrole9 4 1 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "9"
#! LINK_S bccwj-pth:semrole6 4 3 ""
#! ATTR bccwj-pth:role "対象"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "6"
#! LINK_S bccwj-pth:semrole9 4 3 ""
#! ATTR bccwj-pth:role "対象"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "9"
EOS
```

■れる・られるの用法　国語研により助動詞れる・られるの用法をアノテーションした(小山田ほか (2012))。対象となる助動詞をセグメントで切り出し、属性情報として用法分類のラベルを付与する。名前空間として bccwj-auxv を用いる。

例: PN2f_00003 (bccwj-auxv)

```
* 0 2D 0/0 0
俺 代名詞,*,*,*,*,*,オレ,俺,俺
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 1 2D 0/0 0
もう 副詞,*,*,*,*,*,モウ,もう,もう
* 2 3D 0/0 0
やら 動詞,非自立可能,*,*,五段-ラ行,未然形-一般,ヤル,遣る,やら
れ 助動詞,*,*,*,助動詞-レル,連用形-一般,レル,れる,れ
た 助動詞,*,*,*,助動詞-タ,終止形-一般,タ,た,た
と 助詞,格助詞,*,*,*,*,ト,と,と
* 3 -1D 0/0 0
思っ 動詞,一般,*,*,五段-ワア行,連用形-促音便,オモウ,思う,思っ
た 助動詞,*,*,*,助動詞-タ,終止形-一般,タ,た,た
ね 助詞,終助詞,*,*,*,*,ネ,ね,ね
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-auxv:auxv 6 7 "れ"
#! ATTR bccwj-auxv:label "受身"
#! ATTR bccwj-auxv:direct "直接"
#! ATTR bccwj-auxv:sentiment "有情"
#! ATTR bccwj-auxv:agentivity "-"
#! ATTR bccwj-auxv:projection "-"
#! ATTR bccwj-auxv:existence "-"
#! ATTR bccwj-auxv:inducement "-"
EOS
```

■日本語フレームネット　日本語フレームネットは慶應義塾大学でアノテーションされている (小原 (2013))。本稿で説明するアノテーションの中で最も複雑な構造を持っている。フレームに参加する表現をセグメントで切り出し、その代表表現の属性情報 FE-feID にフレーム辞書へのリファレンスを含める。セグメント間の関係はフレームに参加するセグメント全てについてグループを定義して表現する。名前空間として bccwj-jfn を用いる。項構造シソーラスと同様に複数人によりアノテーションされており、属性情報にアノ

テータの情報が含まれている場合がある。

例: PB42_00003 (bccwj-jfn)

```
* 0 1D 0/0 0
マザー 名詞, 普通名詞, 一般,*,*,*, マザー, マザー, マザー
テレサ 名詞, 固有名詞, 人名, 一般,*,*, テレサ, テレサ, テレサ
の 助詞, 格助詞,*,*,*,*, ノ, の, の
* 1 -1D 0/0 0
施設 名詞, 普通名詞, サ変可能,*,*,*, シセツ, 施設, 施設
にて 助詞, 格助詞,*,*,*,*, ニテ, にて, にて
#! SEGMENT_S bccwj-jfn:SEG 0 7 "マザーテレサの"
#! ATTR bccwj-jfn:Lemma-name "N"
#! ATTR bccwj-jfn:FE-feID "1598"
#! ATTR bccwj-jfn:FE-name "Name"
#! SEGMENT_S bccwj-jfn:SEG 7 9 "施設"
#! ATTR bccwj-jfn:Lemma-name "N"
#! ATTR bccwj-jfn:Target-name "Target"
#! ATTR bccwj-jfn:GF-name "Target"
#! ATTR bccwj-jfn:PT-name "Target"
#! ATTR bccwj-jfn:Postpos-name "Target"
#! GROUP_S bccwj-jfn:Frame 0 1 "0"
EOS
```

■拡張モダリティ　拡張モダリティは奈良先端大・東北大でアノテーションされている (松吉ほか (2011))。対象となる事象表現をセグメントで切り出し、属性情報として用法分類のラベルなどを付与する。名前空間として bccwj-eme を用いる。

例: PB49_00005 (bccwj-eme)

```
* 0 2D 0/0 0
　空白,*,*,*,*,*,,　,
平七郎 名詞, 固有名詞, 人名, 名,*,*, ヘイシチロウ, ヘイシチロウ, 平七郎
は 助詞, 係助詞,*,*,*,*, ハ, は, は
* 1 2D 0/0 0
花川戸 名詞, 固有名詞, 地名, 一般,*,*, ハナカワド, ハナカワド, 花川戸
河岸 名詞, 普通名詞, 一般,*,*,*, カシ, 河岸, 河岸
へ 助詞, 格助詞,*,*,*,*, へ, へ, へ
と 助詞, 格助詞,*,*,*,*, ト, と, と
* 2 -1D 0/0 0
急い 動詞, 一般,*,*, 五段-ガ行, 連用形-イ音便, イソグ, 急ぐ, 急い
で 助詞, 接続助詞,*,*,*,*, テ, て, で
い 動詞, 非自立可能,*,*, 上一段-ア行, 連用形-一般, イル, 居る, い
た 助動詞,*,*,*, 助動詞-タ, 終止形-一般, タ, た, た
。 補助記号, 句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-eme:Event 12 14 "急い"
#! ATTR bccwj-eme:morphIDs "80"
#! ATTR bccwj-eme:pseudo ""
#! ATTR bccwj-eme:source "wr:筆者"
#! ATTR bccwj-eme:time "非未来"
#! ATTR bccwj-eme:conditional "0"
#! ATTR bccwj-eme:pmtype "叙述"
#! ATTR bccwj-eme:actuality "成立"
#! ATTR bccwj-eme:evaluation "0"
#! ATTR bccwj-eme:focus ""
#! ATTR bccwj-eme:degree ""
#! SEGMENT_S bccwj-eme:Event 15 16 "い"
#! ATTR bccwj-eme:morphIDs "100"
#! ATTR bccwj-eme:pseudo ""
#! ATTR bccwj-eme:source "wr:筆者"
#! ATTR bccwj-eme:time "非未来"
#! ATTR bccwj-eme:conditional "0"
#! ATTR bccwj-eme:pmtype "叙述"
#! ATTR bccwj-eme:actuality "成立"
#! ATTR bccwj-eme:evaluation "0"
#! ATTR bccwj-eme:focus ""
#! ATTR bccwj-eme:degree ""
EOS
```

■否定の焦点　否定の焦点は山梨大でアノテーションされている (松吉ほか (2013))。対象となる否定辞・焦点・手がかり句をセグメントで切り出し、属性情報として用法分類のラベルなどを付与し、否定辞-焦点間, 焦点-手がかり句間にリンクを付与する。名前空間として bccwj-wsb を用いる。

例: PN2f_00003 (bccwj-wsb)

```
* 0 1D 0/0 0
　空白,*,*,*,*,*,,　,
カ 名詞, 普通名詞, 一般,*,*,*, チカラ, カ,*, チカラ
を 助詞, 格助詞,*,*,*,*, ヲ, を,*, オ
* 1 2D 0/0 0
出し 動詞, 非自立可能,*,*, 五段-サ行, 連用形-一般, ダス, 出す,*, ダシ
切っ 動詞, 非自立可能,*,*, 五段-ラ行, 連用形-促音便, キル, 切る,*, キッ
て 助詞, 接続助詞,*,*,*,*, テ, て,*, テ
* 2 3D 0/0 0
敗れ 動詞, 一般,*,*, 下一段-ラ行, 連用形-一般, ヤブレル, 敗れる,*, ヤブレ
た 助動詞,*,*,*, 助動詞-タ, 連体形-一般, タ, た,*, タ
* 3 -1Z 0/0 0
わけ 名詞, 普通名詞, 一般,*,*,*, ワケ, 訳,*, ワケ
で 助詞, 格助詞,*,*,*,*, デ, で,*, デ
は 助詞, 係助詞,*,*,*,*, ハ, は,*, ワ
ない 形容詞, 非自立可能,*,*, 形容詞, 終止形-一般, ナイ, 無い,*, ナイ
。 補助記号, 句点,*,*,*,*,,。,*,
#! SEGMENT_S bccwj-wsb:Negation 15 17 "ない"
#! ATTR bccwj-wsb:morphID "4450"
#! ATTR bccwj-wsb:POS "否定複合辞"
#! SEGMENT_S bccwj-wsb:Focus 5 7 "切っ"
#! ATTR bccwj-wsb:morphID "4380"
#! ATTR bccwj-wsb:argTypes "テ節; 動詞句"
```

```
#! ATTR bccwj-wsb:class "付加-連用修飾"
#! ATTR bccwj-wsb:description "様態の副詞句が使用されている"
#! SEGMENT_S bccwj-wsb:Clue 11 17 "わけではない"
#! ATTR bccwj-wsb:morphIDs "4420.4430.4440.4450"
#! LINK_S bccwj-wsb:FocusOfNegation 0 1 ""
#! LINK_S bccwj-wsb:Clue 1 2 ""
EOS
```

■述語項構造・照応関係　述語項構造は奈良先端大でアノテーションされている (小町・飯田 (2011))。対象となる述語と項をセグメントで切り出し、述語-項間にリンクを付与する。また照応関係を同一の実体を指し示す同値類として捉えてグループを付与する。名前空間として bccwj-pas を用いる。

例: PB40_00003 (bccwj-pas)

```
* 0 1D 0/0 0
客先 名詞，普通名詞，一般,*,*,*，キャクサキ，客先，客先
の 助詞，格助詞,*,*,*,*,*，ノ，の，の
* 1 2D 0/0 0
安全 名詞，普通名詞，形状詞可能,*,*,*，アンゼン，安全，安全
基準 名詞，普通名詞，一般,*,*,*，キジュン，基準，基準
が 助詞，格助詞,*,*,*,*,*，ガ，が，が
* 2 -1D 0/0 0
厳しく 形容詞，一般,*,*，形容詞，連用形-一般，キビシイ，厳しい，厳しく
て 助詞，接続助詞,*,*,*,*,*，テ，て，て
* 3 4D 0/0 0
守れ 動詞，一般,*,*，下一段-ラ行，連用形-一般，マモル，守る，守れ
そう 形状詞，助動詞語幹,*,*,*,*,*，ソウ，そう，そう
に 助動詞,*,*,*,*，助動詞-ダ，連用形-ニ，ダ，だ，に
* 4 -1D 0/0 0
ない 形容詞，非自立可能,*,*，形容詞，終止形-一般，ナイ，無い，ない
EOS
    :
#! SEGMENT bccwj-pas:Np 0 2 "客先"
#! SEGMENT bccwj-pas:Np 5 7 "基準"
#! SEGMENT bccwj-pas:Pred 8 11 "厳しく"
#! ATTR bccwj-pas:type "述語"
#! SEGMENT bccwj-pas:Pred 12 14 "守れ"
#! ATTR bccwj-pas:type "述語"
    :
#! LINK bccwj-pas:Ga 2 1 ""
#! LINK bccwj-pas:O 3 1 ""
    :
#! GROUP bccwj-pas:Coref 0 36 37 ""
```

外界照応のような参照対象が文書中に出現しないようなリンクは、セグメントに対する属性表現 Exophora_Ga により表す。

```
#! SEGMENT bccwj-pas:Pred 392 395 ""
#! ATTR bccwj-pas:type "述語"
#! ATTR bccwj-pas:Exophora_Ga "その他"
```

4.2　重ね合わせ事例

アノテーションを重ね合わせた事例を稿末の付録Ａ．に示す。時間情報 (bccwj-timebank) と語義 (bccwj-wsd) と述語項構造 (bccwj-pas) と日本語フレームネット (bccwj-jfn) を重ね合わせたものである。また、ChaKi.NET による可視化事例についても稿末に示す。セグメント相当にマウスカーソルを合わせることで属性情報を見ることが可能である。

5.　おわりに

本稿では拡張 CaboCha フォーマットに対して、属性情報と名前空間を表現する機構を新たに追加することを提案し、BCCWJ に対して付与されている十種類のアノテーションについてどのように表現できるか例を示した。言語学・言語処理双方の利用にとって、このような表現が有効か否かについて広く意見を求めたい。

今後、アノテーション開発者・配布者・利用者の意見を伺いながら、仕様の最終案を策定し、各アノテーションの重ね合わせを進めたい。また、コーパス管理ツール ChaKi.NET (Matsumoto et al. (2005)) に重ね合わせたアノテーション情報に対する検索・問い合わせを行う機構を実装し、利用者系の整備も進めていきたいと考える。

謝辞

本研究の一部は科研費若手研究 (B)「否定焦点コーパス構築と焦点自動解析に関する研究」（課題番号：25870278、代表：松吉俊）、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

浅原正幸・松本裕治 (2013).「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」　言語処理学会第 19 回年次大会発表論文集.

橋本泰一・中村俊一 (2010).「拡張固有表現タグ付きコーパスの構築」　言語処理学会第 16 回年次大会発表論文集.

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」　『現代日本語書き言葉均衡コーパス』完成記念講演会.

小西光・浅原正幸・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」　自然言語処理, 20:2, pp. 201–222.

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

松吉俊・佐尾ちとせ・乾健太郎・松本裕治 (2011).「拡張モダリティタグ付与コーパスの設計と構築」　言語処理学会第 17 回年次大会発表論文集.

松吉俊・大槻諒・福本文代 (2013).「日本語における否定の焦点アノテーション」　第 3 回コーパス日本語学ワークショップ予稿集.

小原京子 (2013).「日本語フレームネット：文意理解のためのコーパスアノテーション」　言語処理学会第 19 回年次大会発表論文集.

奥村学・白井清昭 (2009).「BCCWJ を用いた新しい語義曖昧性解消タスク」　言語処理学会第 15 回年次大会発表論文集.

奥村学・白井清昭・古宮嘉那子・横野光 (2013).「BCCWJ 上の語義タグ付コーパス」　言語処理学会第 19 回年次大会（ライトニングトーク）.

小山田由紀・柏野和佳子・前川喜久雄 (2012).「助動詞レル・ラレルへの意味アノテーション作業経過報告」　第 2 回コーパス日本語学ワークショップ予稿集.

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 回年次大会発表論文集.

保田祥・小西光・浅原正幸・今田水穂・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパス』に対する時間表現・事象表現間の時間的順序関係アノテーション」　自然言語処理, 20:5, pp. 657–682.

関連 URL

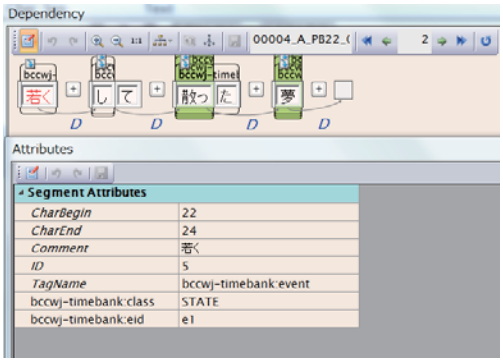
- 「現代日本語書き言葉均衡コーパス」Web ページ：`http://www.ninjal.ac.jp/corpus_center/bccwj/`

- 「CaboCha」 Web ページ：https://code.google.com/p/cabocha/
- 「ChaKi.NET」 Web ページ：http://sourceforge.jp/projects/chaki/releases/
- 「BCCWJ-DepPara」Web ページ：https://github.com/masayu-a/BCCWJ-DepPara
- 「BCCWJ-DepPara2」Web ページ：https://github.com/masayu-a/BCCWJ-DepPara2
- 「BCCWJ-TimeBank」Web ページ：https://github.com/masayu-a/BCCWJ-Timebank
- 「拡張固有表現タグ付きコーパス」Web ページ：http://www.gsk.or.jp/catalog/gsk2013-b/
- “SENSEVAL-2 Japanese WSD task” Web ページ：http://www.lr.pi.titech.ac.jp/wsd.html
- 「岡山大学動詞項構造シソーラス」 Web ページ：http://cl.cs.okayama-u.ac.jp/rsc/data/index.html
- 「BCCWJ-AUXV」 Web ページ： https://github.com/masayu-a/BCCWJ-AUXV
- 「日本語フレームネット」Web ページ：http://jfn.st.hc.keio.ac.jp/ja/
- 「拡張モダリティタグ付与コーパス」Web ページ：http://cl.cs.yamanashi.ac.jp/nldata/modality/
- 「否定の焦点アノテーション」Web ページ：http://cl.cs.yamanashi.ac.jp/nldata/negation/
- 「BCCWJ:述語項構造と照応関係のアノテーション」Web ページ：http://cl.naist.jp/nldata/bccwj/pas/

付録A. 重ね合わせの例

例: PB22_00002

```
* 0 1D 0/0 0
若く 形容詞，一般,*,*，形容詞，連用形-一般，ワカイ，若い，若く
* 1 2D 0/0 0
し 動詞，非自立可能,*,*，サ行変格，連用形-一般，スル，為る，し
て 助詞，接続助詞,*,*,*,*，テ，て，て
* 2 3D 0/0 0
散っ 動詞，一般,*,*，五段-ラ行，連用形-促音便，チル，散る，散っ
た 助動詞,*,*,*，助動詞-タ，連体形-一般，タ，た，た
* 3 -1D 0/0 0
夢 名詞，普通名詞，一般,*,*,*，ユメ，夢，夢
#! SEGMENT_S bccwj-timebank:event 0 2 "若く"
#! ATTR bccwj-timebank:eid "e1"
#! ATTR bccwj-timebank:class "STATE"
#! SEGMENT_S bccwj-timebank:event 2 3 "し"
#! ATTR bccwj-timebank:eid "e2"
#! ATTR bccwj-timebank:class ""
#! SEGMENT_S bccwj-timebank:event 4 7 "散った"
#! ATTR bccwj-timebank:eid "e3"
#! ATTR bccwj-timebank:class ""
#! SEGMENT_S bccwj-wsd:wsd 4 6 "散っ"
#! ATTR bccwj-wsd:sense "33687-0-0-4-0"
#! ATTR bccwj-wsd:explanation "&lt;4&gt; 比ゆ的に、人がいさぎよく死ぬ。多く、戦死にいう。「花と―」"
#! SEGMENT_S bccwj-wsd:wsd 7 8 "夢"
#! ATTR bccwj-wsd:sense "52859-0-0-3-0"
#! ATTR bccwj-wsd:explanation "&lt;3&gt;（現在のところ実現してはいないが）将来は実現させたい願い・理想。「彼の―は大きい」「―を描く」"
#! SEGMENT_S bccwj-pas:Pred 2 3 "し"
#! ATTR bccwj-pas:type "述語"
#! SEGMENT_S bccwj-pas:Pred 4 6 "散っ"
#! ATTR bccwj-pas:type "述語"
#! SEGMENT_S bccwj-pas:Np 7 8 "夢"
#! SEGMENT_S bccwj-jfn:JFN 4 6 "散っ"
#! ATTR bccwj-jfn:Lemma-name "V"
#! ATTR bccwj-jfn:Target-name "Target"
#! ATTR bccwj-jfn:GF-name "Target"
#! ATTR bccwj-jfn:PT-name "Target"
#! SEGMENT_S bccwj-jfn:JFN 7 8 "夢"
#! ATTR bccwj-jfn:Lemma-name "V"
#! ATTR bccwj-jfn:FE-feID "6803"
#! ATTR bccwj-jfn:FE-name "Entity"
#! LINK_S bccwj-pas:Ga 6 7 ""
#! GROUP_S bccwj-jfn:JFN 8 9 "0"
EOS
```



言語単位と文の長さが品詞比率に与える影響

山崎　誠（国立国語研究所言語資源研究系）†

Influence of Word Unit and Sentence Length on the Ratio of Parts of Speech

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

語類構成比の問題は語彙の分布問題の重要なテーマとして 1950 年代から研究が行われているが、大野(1956)、樺島(1954)(1955)などの初期の研究以降、目立った進展がないようである。水谷(1983)における語類構成比の記述と計量国語学会編(2009)の品詞構成比率の記述(pp.95-96)の間に進歩が見られないことはこの分野の研究が停滞していることを示していると言ってよいだろう。

２．品詞構成比

語類構成比について唯一研究が行われているのが品詞の比率である。テキストにおける品詞構成比を異なり語数のレベルで分析し定式化につながる傾向を見いだしたのが大野(1956)であり[1]、延べ語数のレベルで分析し定式化したのが樺島(1954)(1955)である。大野は古典、樺島は現代文という違いはあるものの、どちらも品詞の比率がテキストのジャンルによって異なり、一定の法則が見いだせるということを指摘している。

本稿はこれまで十分検討されてこなかった、文の長さが品詞比率とどう関わっているか[2]という点について調査したものである。日本語では文の長さは文字数で計測することが多いが、本稿では操作的に作成した言語単位によって計測する。具体的には『現代日本語書き言葉均衡コーパス』（以降、BCCWJ と略す）で用いられている短単位(SUW)及び長単位(LUW)である。この 2 つの単位の違いにより、品詞構成比がどのように違ってくるかということも併せて調査する。

３．データと方法

本稿では、BCCWJ の DVD 版に納められている M-XML ファイルを利用した。M-XML ファイルには<sentence>というタグが埋め込まれており、これを元に文の候補となるデータ抽出した[3]。さらに<sentence>タグの中に含まれる<suw>タグで短単位の情報を、<luw>タグで長単位の情報を抽出した。

品詞の認定は BCCWJ で用いられている UniDic の品詞体系に従った。なお、本稿では樺島(1954)(1955)(1979)に基づいて品詞を類別する[4]が、その際の基準は接尾辞を除き、UniDic の大分類を利用した。接尾辞については UniDic の中分類までの情報を利用した。品詞分類

† yamazaki@ninjal.ac.jp

1 いわゆる「大野の法則」を数学的に定式化したのは水谷(1965)(1981)である。

2 樺島(1979:218)に「名詞の比率が大きくなるにつれて、文の長さの平均値も次第に大きくなっているようである。」という指摘がある。

3 小木曽他(2011)によれば、「サンプルは sentence の集合として捉えられる」ということなので、抽出の際に漏れた部分は存在しない。

4 類別した品詞の各類の名称は樺島(1954)では、名詞、動詞、形・形動・副・連体詞、感動・接続詞、樺島(1955)では略称を用いて、N、V、Ad、I、樺島(1979)では、N、V、M、I となっている。いずれも 4 類に分類し、所属する品詞は同じである。本稿では樺島(1979)の名称を用いる。

の詳細を表 1 に示す。短単位、長単位ともこの基準を用いた。本稿の分析ではことわりのない限り、品詞の属性が「空白」および句読点などの「補助記号」を除外している。

表 1　本稿での品詞の類別

分類	UniDic の品詞
M（名詞類）	名詞
	代名詞
	接尾辞－名詞的
	記号[5]
V（動詞類）	動詞
	接尾辞－動詞的
M（形容詞・形状詞・副詞類）	形容詞
	形状詞
	副詞
	連体詞
	接頭辞
	接尾辞－形容詞的
	接尾辞－形状詞的
I（接続詞・感動詞類）	接続詞
	感動詞
P（助詞・助動詞類）	助詞
	助動詞
O（その他）	（その他）[6]

４．文の長さの分布

４．１　全体について

BCCWJ 全体における<sentence>タグで定義された「文」の数は 5,515,952 個であった。ただし、この全てがいわゆる通常の文というわけではない。間淵(2011:237)が指摘しているように、<sentence>は自動的に付与されているため、「文と認定されるべき要素であっても、sentence と認定されないものがある」または「文には相当しない sentence や，明らかに文中である不適切な位置で sentence が分断されているものがある」からである。そのことはさておき、上記の「文」について、長さの分布を見てみよう。

表 2　文の長さの分布（<sentence>タグ全体）

言語単位	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
短単位(SUW)	1.0	113.2	225.5	275.0	352.5	5527.0
長単位(LUW)	1.0	89.8	178.5	232.0	283.5	5019.0

[5] この「記号」は、「M氏」の「M」や「Ｆ１５」の「F」のような部分に付けられた品詞名である。機能的に名詞として位置付けられるとしてここに置いた。

[6] その他に該当するのは、品詞欄に以下の属性を持つものである。各属性の後ろの括弧内の数字は、(短単位/長単位)の順に該当する言語単位数を示した。「言いよどみ」(2931/318)、「web 誤脱」(173/166)、「英単語」(2931/3093)、「カタカナ文」(65450/44093)、「漢文」(551/555)、「ローマ字文」(21/0)、「新規未知語」(2/0)、「方言」(236/232)、「未知語」(194825/194756)、「URL」(18465/18232)。なお、品詞欄が空欄のものが長単位に 1 つあったが、これも（その他）に含めている。（その他）の占める割合は、短単位・延べ語数で約 0.24%、長単位・延べ語数で約 0.27%である。この数字は補助記号、空白も含めた全ての言語単位に占める割合である。

短単位、長単位とも最大値となっているのは、ジェイムズ・ジョイスの『ユリシーズ』からのサンプル（LBr9_00057、丸谷才一他訳、集英社、2003 年刊）で、この小説の特異な文体の特徴であり、かなり例外的なものであろう[7]。文の長さと文数の関係を図 1 に示した。

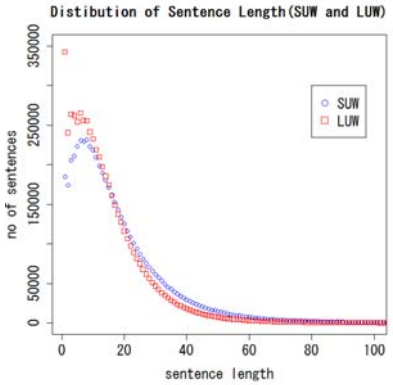


図 1　文の長さの分布

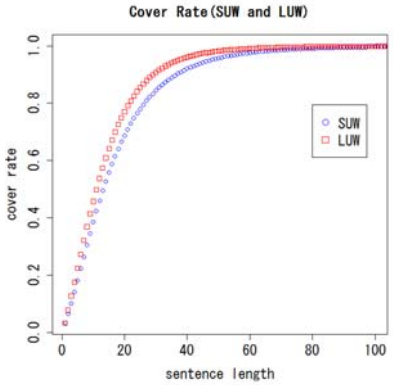


図 2　カバー率

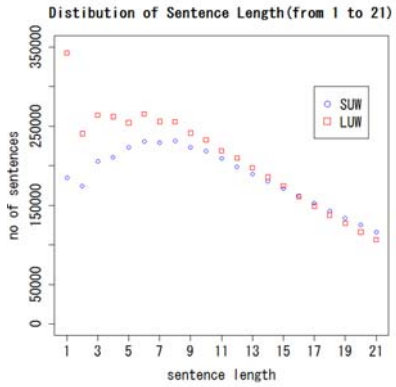


図 3　文の長さの分布（1～21）

図 1 から、文の長さが短い値を示す部分、とくに 1～10 の付近が変則的な動きになっていることが見て取れる。

図 2 のカバー率から、短単位は文の長さが 37 語で全体の 90%をカバー、長単位は 29 語で全体の 90%をカバーしていることが分かる。

図 3 は、文の長さが 20 語以下の分布を示したものである。これを見ると、短単位、長単位ともに文の長さが 1 から 2 にかけて下降、2 から上昇に転じている。短単位では、文の長さが 3～5 にかけて下降し、その後 8 まで上下を繰り返している。長単位ではやはり 8 まで上昇し、そこから下降に転じている。この不規則な動きの原因は後述のように、文の認定にある。

４．２　通常の文について

前節では<sentence>タグで認定された文全体を対象としたが、これには新聞の見出しや書籍の章のタイトル、図表のキャプションのような、通常は文とみなさないものも含まれる。このような「通常でない文」を排除するために、<sentence>タグで囲まれた文の末尾[8]の要素の語彙素が「。」「.」「！」「？」「」」で終わるものを本稿で言う「通常の文」と見なして抽出した。「」」以外は、間淵(2011:236)によると、<sentence>タグの自動認定の際の文終止マーカーである。「」」をそこに加えたのは、小説などの会話文では「」」で終わるものが多いのではないかと予想したからである。通常の文は 4,374,273 個であり、<sentence>でタグ付けされた文の約 80%である。

表 3 は「通常の文」の長さの分布を示したものである。表 2 の<sentence>タグ全体に比べると第 1 四分位数～最大値の値が低くなっていることが分かる。

---

[7] 短単位で 2 番目に長い文（文長 3299）および長単位で 2 番目に長い文（文長 2917）も『ユリシーズ』からのサンプルである。

[8] 品詞の属性が「空白」と「補助記号」を除いた末尾の要素を指す。

表 3　文の長さの分布（「通常の文」）

言語単位	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
短単位(SUW)	1.0	105.0	209.0	242.2	323.0	3279.0
長単位(LUW)	1.0	84.5	168.0	204.1	261.5	2917.0

図 4 は通常の文について、文の長さの分布を示したものである。図 1 と違って文の長さが少ないあたりのカーブがなだらかになっていて不規則性が解消されていることが分かる。図 5 は通常の文のカバー率である。カバー率が 90%に達する文の長さは図 3 と比べると若干長くなり、短単位で 40 語、長単位では 31 語である。

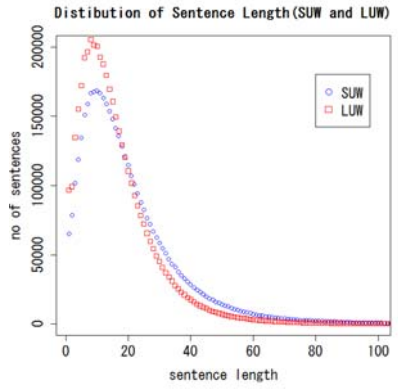


図 4　文の長さの分布（通常の文）

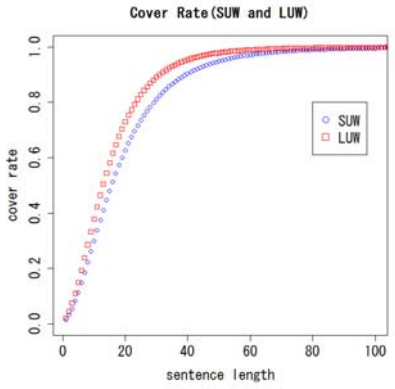


図 5　カバー率（通常の文）

４．３　通常でない文について

表 4 は、「通常でない文」すなわち、「。．！？」」で終わっていない文の長さの分布、図 6 は「通常でない文」における文の長さの分布、図 7 はそのカバー率である。ここには、見出し相当の語句などが含まれるため、文の長さが短い方に偏っていることが分かる。また、図 6 の分布の形も図 4 と異なる形になっている。90%のカバー率になるのは短単位で 18 語、長単位で 13 語であり、こちらも前記 2 つのカバー率よりも急峻であることが分かる。

表 4　文の長さの分布（「通常でない文」）

言語単位	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
短単位(SUW)	1.0	72.3	143.5	194.1	217.8	5527.0
長単位(LUW)	1.0	53.0	105.0	160.2	164.0	5019.0

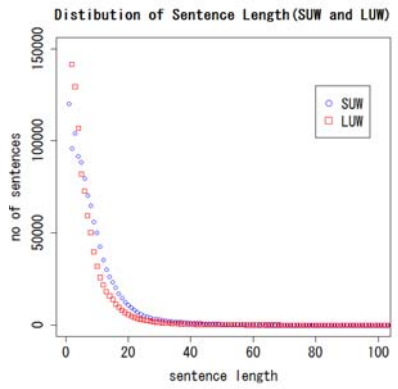


図 6　文の長さの分布（通常でない文）

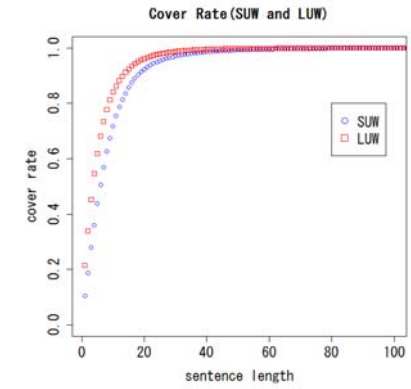


図 7　カバー率（通常でない文）

４．４　文末の記号別の分布

本稿では、「。．！？」」を「通常の文」の文末を示す記号として文の抽出を行った。その各記号別に文の長さの分布を見てみよう。表 5 は、「通常の文」に占める各文末の記号の数である。また、表 6、表 7 に分けて言語単位別の分布の様子である。最小値と最大値を除いて文の長さは短単位も長単位も＜。　」　．　？　！＞の順になっていることが分かる。

表 5　文の長さの分布

文末の記号	。	.	!	?	」
文数	3643999	62346	118680	163037	386211
割合	83.30	1.43	2.71	3.73	8.83

表 6　文末の記号別の文の長さの分布（短単位）

文末の記号	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
。	1.0	102.5	204.0	238.3	313.5	3279.0
.	1.0	45.8	90.5	105.2	136.5	461.0
!	1.0	30.8	60.5	66.1	92.3	182.0
?	1.0	33.0	65.0	73.0	97.0	433.0
」	1.0	51.0	101.0	113.7	158.0	430.0

表 7　文末の記号別の文の長さの分布（長単位）

文末の記号	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
。	1.0	82.3	163.5	200.8	252.8	2917.0
.	1.0	36.8	72.5	85.8	114.5	367.0
!	1.0	27.5	54.0	58.5	82.5	150.0
?	1.0	29.5	58.0	65.7	86.5	403.0
」	1.0	44.3	87.5	97.5	131.8	354.0

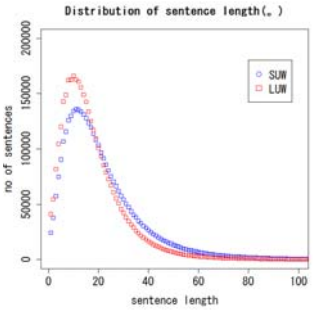


図 8 文の長さの分布（。）

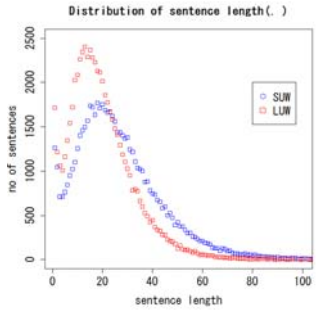


図 9 文の長さの分布（.）

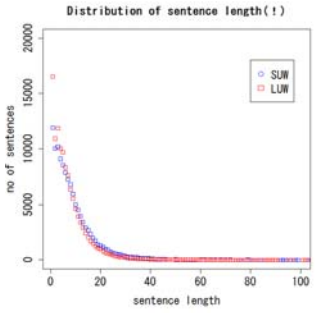


図 10 文の長さの分布（！）

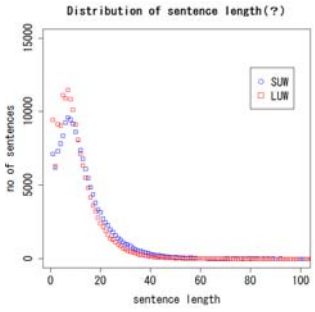


図 11 文の長さの分布（？）

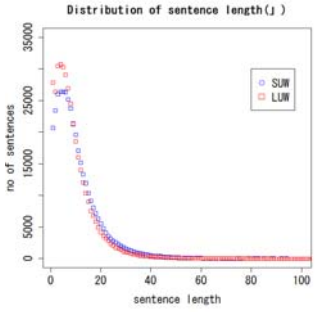


図 12 文の長さの分布（」）

図 8～図 12 は各文末記号別に文の長さの分布を観察したものである。図 10 の感嘆符の場合が他と異なる分布になっていること、また、図 9 のピリオドの場合、短単位と長単位の差が大きいが、感嘆符、疑問符、カギ括弧の場合は短単位と長単位の差が少ないことが分かった。これらは具体的にどのような文が用いられているかを見ないとこれらの違いの言語学的な解釈ができない。今回はあくまで現象の指摘にとどまる。

５．文の長さと品詞類の比率

５．１　全体の傾向

文の長さにより品詞類の比率がどのように変化するか（しないか）を観察した。品詞類は表 1 に挙げたように、N（名詞類）、V（動詞類）、M（形容詞・形状詞・副詞類）、I（接続詞・感動詞類）、P（助詞・助動詞類）の 5 つに分類した。P 以外は樺島(1954)以降の一連の分類と同じものである。樺島の分類には助詞・助動詞を含めていないが、延べ語数の水準での品詞類の構成比を見るには助詞・助動詞は言語量が多いため重要な要素であると考え、考察の対象とすることにした。表 8 は BCCWJ 全体の「通常の文」における各品詞類の割合である。短単位と長単位の違いとしては、N の比率が長単位では相対的に低く、P の比率が相対的に高くなることが挙げられる。V、M、I の比率は短単位でも長単位でもあまり変わらない。表 9、表 10 はそれぞれ短単位、長単位での文末記号別の品詞類の比率である。短単位では N と P の変動が大きいが、長単位ではどの品詞類も変動が相対的に小さくなる。また、句点で終わる文に比べて疑問符、かぎ括弧で終わる文で、N の割合が低く、P の割合が多くなっているのは話し言葉的な要因が関係している可能性がある。

表 8　品詞類の比率（通常の文）

	N	V	M	I	P
短単位	0.376	0.141	0.067	0.006	0.410
長単位	0.297	0.129	0.078	0.010	0.485

表 9　文末の記号別の品詞類の比率（短単位）

文末の記号	N	V	M	I	P
。	0.380	0.142	0.066	0.006	0.407
.	0.460	0.125	0.058	0.006	0.350
!	0.373	0.124	0.090	0.015	0.398
?	0.312	0.124	0.078	0.007	0.480
」	0.300	0.142	0.086	0.013	0.459

表 10　文末の記号別の品詞類の比率（長単位）

文末の記号	N	V	M	I	P
。	0.300	0.130	0.077	0.009	0.484
.	0.351	0.117	0.069	0.010	0.454
!	0.296	0.123	0.101	0.022	0.458
?	0.255	0.120	0.088	0.011	0.527
」	0.254	0.133	0.093	0.020	0.500

図 13、図 14 はそれぞれ短単位、長単位における文の長さによる品詞類の比率の推移を示したものである。ここでは 4．2 節で採り上げた「通常の文」を対象としている。図 13、図 14 からは、文の長さが 10 を超えたあたりから各品詞類の比率が一定化することが見て取れる。ただし、短単位では長さ 43 から N（名詞類）の比率が P（助詞・助動詞類）の比率を上回る。また、特徴的なのは文の長さ 1 における分布で、短単位では I(0.3997)、N(0.342)、

M(0.1645)、P(0.0469)、V(0.0468)、長単位では N(0.4974)、I(0.293)、M(0.1275)、V(0.0492)、P(0.033)となっている。括弧内の数字は構成比の値である。I（接続詞・感動詞類）の比率が高いのは、長さ 1 の文にはそれらだけで成り立っているものが多いということを示している。

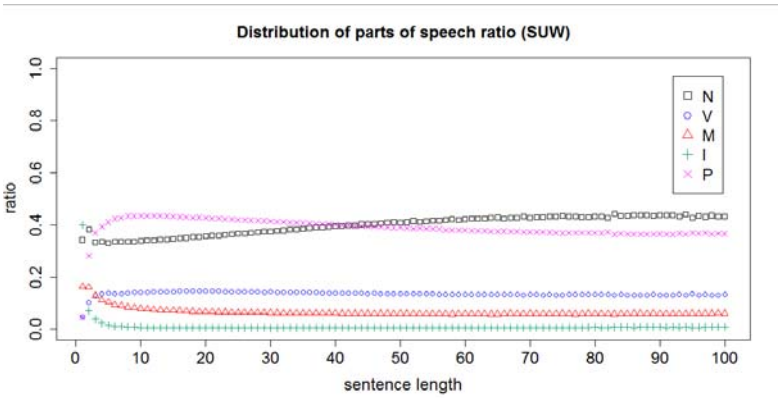


図 13　文の長さによる品詞類の比率の推移（短単位）

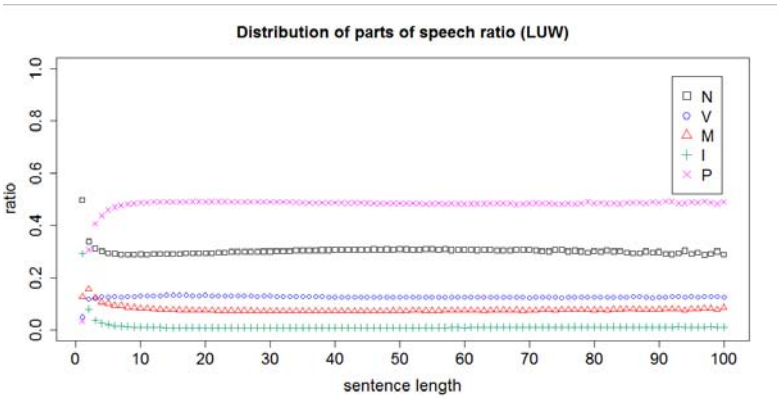


図 14　文の長さによる品詞類の比率の推移（長単位）

５．２　文の長さと品詞類の相関

樺島(1979:218)には、名詞の比率が高くなると文の長さの平均値も大きくなるという指摘があるが、今回のデータで文の長さと各品詞類および各品詞類どうしの相関がどうなっているかを調査した。文の長さが 1～100 について相関を調べた結果を表 11、表 12 に示す。相関係数（ピアソンの積率相関係数）が絶対値で 0.7 以上の部分のセルを網掛けで示したが、文の長さとの相関があったのは、短単位の N（名詞類）のみであった。各品詞類どうしの相関は、短単位で V と P、V と I、I と P の３組、長単位で I と P、N と V、V と I、N と I、N と P、V と P、M と P の 7 組であった。

表 11　文の長さと品詞類との相関行列（短単位）

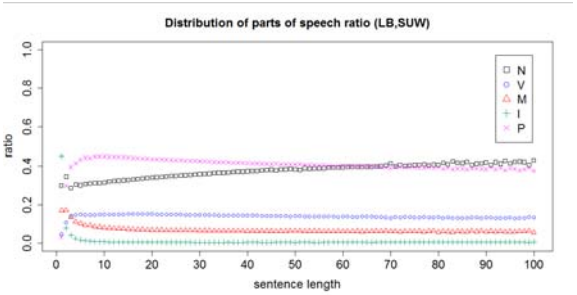
	文の長さ	N	V	M	I	P
文の長さ	1.000	0.951	-0.157	-0.575	-0.210	-0.307
N		1.000	-0.210	-0.566	-0.187	-0.358
V			1.000	-0.568	-0.885	0.975
M				1.000	0.681	-0.438
I					1.000	-0.832
P						1.000

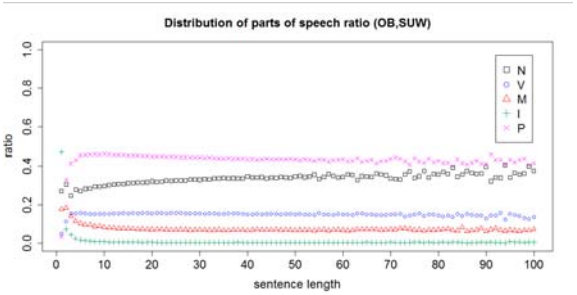
表 12　文の長さと品詞類との相関行列（長単位）

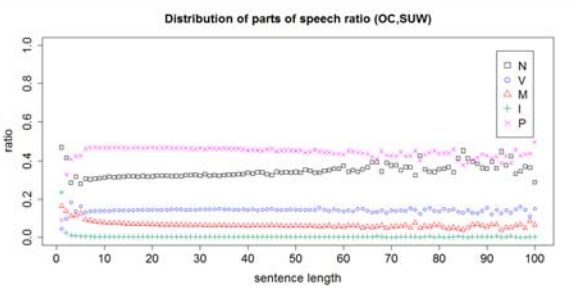
	文の長さ	N	V	M	I	P
文の長さ	1.000	-0.156	0.033	-0.306	-0.213	0.260
N		1.000	-0.965	0.454	0.940	-0.931
V			1.000	-0.493	-0.954	0.928
M				1.000	0.637	-0.726
I					1.000	-0.986
P						1.000

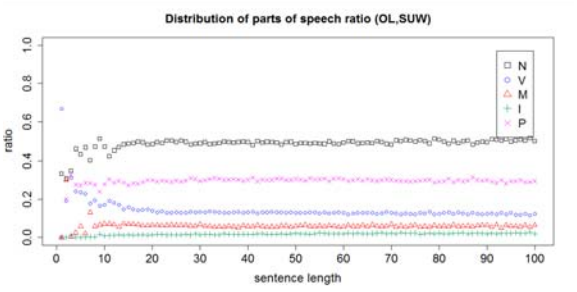
５．３　レジスターによる違い

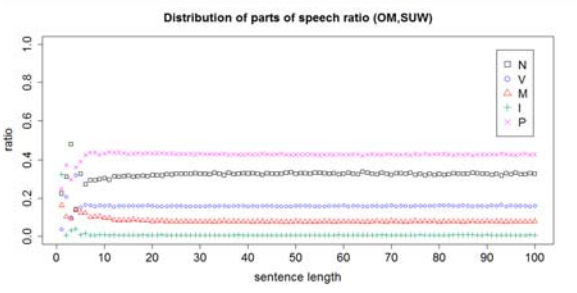
文の長さによる品詞類の比率はレジスターによって違いがあるかを見てみよう。図 15 に短単位における文の長さ 1～100 の推移を示した。

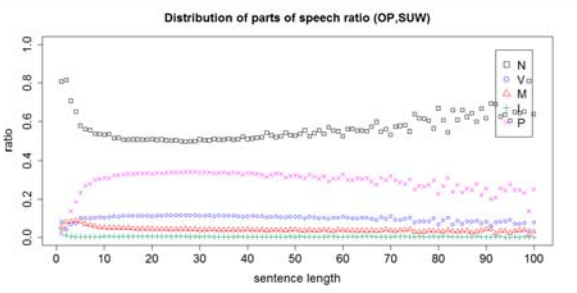


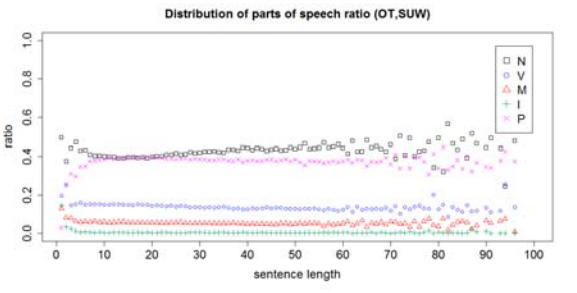


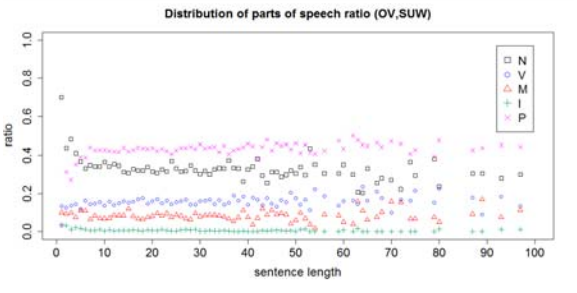












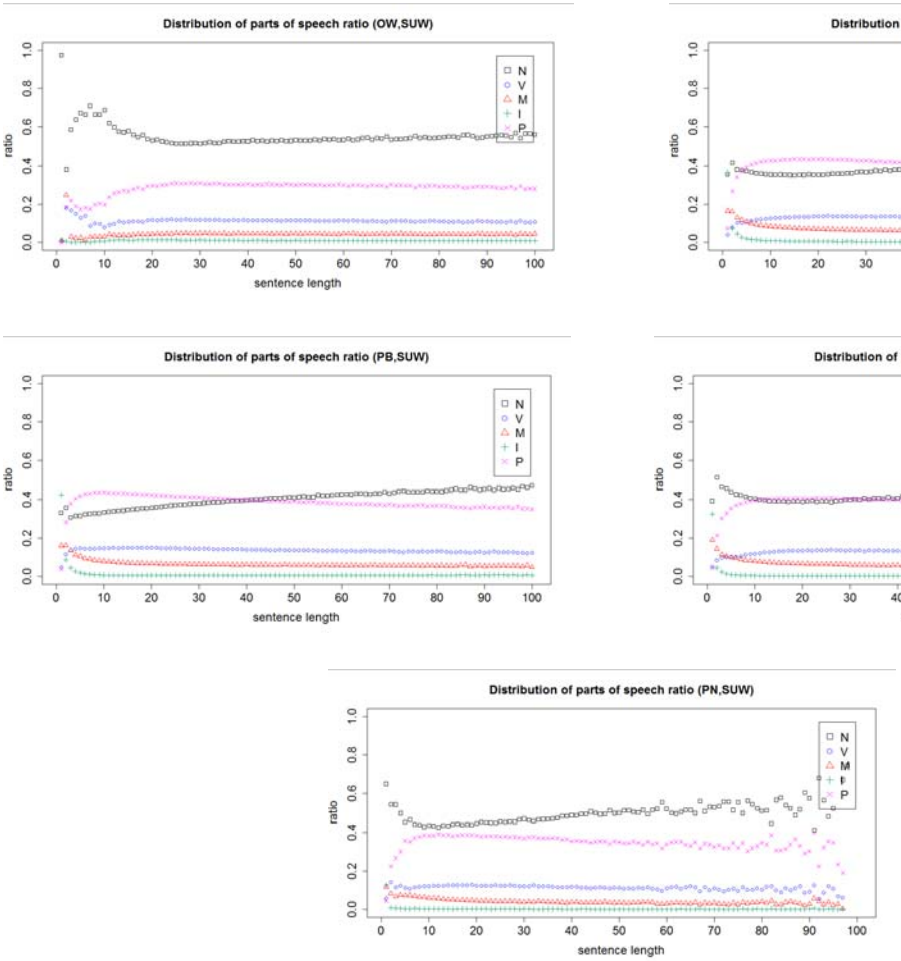


図 15　レジスターごとの文の長さによる品詞類の比率の推移（短単位）

N（名詞類）と P（助詞・助動詞類）との関係が各レジスターで違いが見られた。全体の傾向と同じような、最初 P が大きく、途中で N が上回るようになるのは、LB（図書館・書籍）、PB（出版・書籍）である。OY（Yahoo!ブログ）もその傾向に近いが、N が P を上回るところまでいかず、ほぼ同じ値に収束している。OT（教科書）、PM[9]（雑誌）は最初 N と P がほぼ同じ値であり、途中から N が P より大きくなる傾向がある。OB（ベストセラー）、OC（Yahoo!知恵袋）、OM（国会会議録）は OV(韻文)P が終始 N を上回っている。OL（法律）、OP（広報紙）、 OW（白書）、PN（新聞）は終始 N が P を上回っている。N と P の関係は長単位ではほとんどのレジスターで P が N を上回っている。ただし、OL では N と P の比率がほぼ同じであった。紙幅の関係で対照的な分布を示す LB と OL の 2 つ

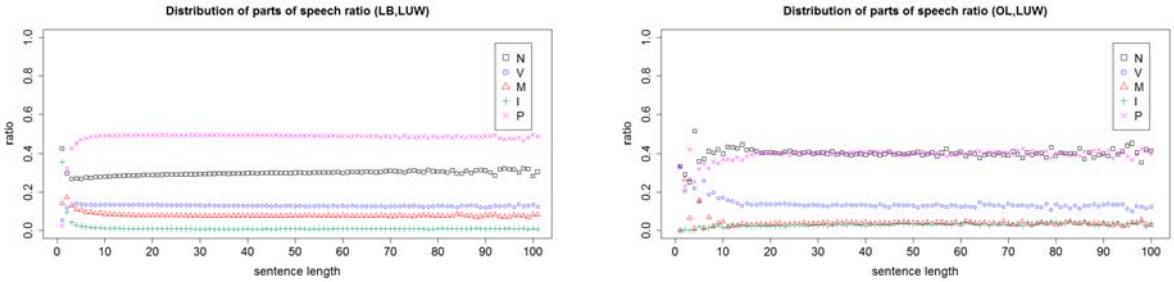


図 16　レジスターごとの文の長さによる品詞類の比率の推移（長単位）

[9] PM は OP と同じグループとも考えられる。

を図 16 に挙げた。

６．まとめと今後の課題

本稿では言語単位と文の長さによる品詞類の構成比を概観した。品詞類の分類は樺島(1954)等にしたがって、N（名詞類）、V（動詞類）、M（形容詞・形状詞・副詞類）、I（接続詞・感動詞類）を設けさらに P（助詞・助動詞類）の追加した 5 分類とした。結果としては以下の傾向が観察された。

（１）短単位と長単位とでは N と P の比率に変化が見られた。N は短単位より長単位の方が低く、逆に P は短単位より長単位のほうが高い。

（２）文末の記号による品詞類の比率に変化が見られた。N の値が「」、？」で低く、「.」で高い。「。！」はその中間であった。

（３）文の長さによる品詞類の比率の推移については、短単位、長単位ともに長さ 10 くらいから値が一定化する傾向が見られた。ただし、短単位ではそれまで P＞N だった傾向が長さ 43 から N＞P となり逆転する。

（４）文の長さとの相関が見られた品詞類は短単位の N の場合のみであった（長さ 1～100 における相関係数において係数 0.951）。

（５）レジスターによる違いについては､短単位において N と P について違いが見られた。

常に N＞P：OL　OP　OW　PN
常に N＜P：OB　OC　OM　OV
N＜P から N＞P へ変化：LB　PB
N＜P から N＝P へ変化：OY
N＝P から N＞P へ変化：OT　PM

今後は個別の品詞についての分析を進めるとともに、今回観察された現象の言語学的な解釈について考察を進めたい。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した 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

大野晋(1956),「基礎語彙に関する二三の研究―日本の古典文学作品における」, 国語学, 24, pp.296-329.
小木曽智信・間淵洋子・前川喜久雄(2011),「『現代日本語書き言葉均衡コーパス』における形態論情報付き XML フォーマット」, 言語処理学会第 17 回大会発表論文集, pp.352-355.
樺島忠夫(1954)「現代文における品詞の比率とその増減の要因について」，国語学，18，pp.15-20.
樺島忠夫(1955)「類別した品詞の比率に見られる規則性」，国語国文，24(6)，pp.55-57.
樺島忠夫(1979)『日本語のスタイルブック』大修館書店
計量国語学会(2009)『計量国語学事典』朝倉書店
間淵洋子(2011)「自動認定によって付与されるタグ」,『現代日本語書き言葉均衡コーパス』における電子化テキストの構築，国立国語研究所内部報告書(LR-CCG-10-03)
水谷静夫(1965)「大野の語彙法則について」，計量国語学，35，pp.1-13.
水谷静夫(1981)「構成比の線型回帰調整，併せて再び大野の語彙法則」計量国語学，13(2), pp.92-97.
水谷静夫(1983)『朝倉日本語講座 2 語彙』朝倉書店

言語は経済に連動して変化する

森　秀明（東北大学大学院文学研究科）

Language Change Links Economy

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

１．研究の目的

日本語の歴史において最も言語が変化した時代は室町時代であるといわれる。そしてその要因のひとつにこの時代の著しい経済発展が指摘されている。言語が社会の変動によって変化すること、そしてその社会変動の一つに経済があることについて、異論は少ないと思われる。しかし、言語が経済に連動して変化すると言ったら、首をかしげる人も多いのではないだろうか。経済言語学の嚆矢といえるクルマス(1993)では、言語には富を背景にして普及していくという経済性と、言語自身が自律的に効率化していくという経済性の二つがあり、それらは密接に結びついているという仮説が提示されている。しかし同時にこれらを結びつける知見のあまりに少ないことが指摘されている。

本研究はこの二つの経済性を結びつける可能性を秘めたいくつかの知見を述べるものである。まず 2.では「言語はなぜ変化するのか」について、先行研究で述べられた諸説を概観・整理しながら変化の理由を考える。3.では中村・米田ほか(2011)の分析を使用し、方言の共通語化における世代ごとの言語変化と経済成長率とに相関があることを述べる。4.では『現代日本語書き言葉均衡コーパス』「図書館書籍」を使用し、出版年ごとの言語変異形出現率と経済成長率に相関があることを述べる。最後に 5.ではこれらの知見がクルマスの述べる二つの経済性とどのように結びつけられるのかについて考察する。

２．言語はなぜ変化するのか

あらゆる言語は変化する。しかし言語がなぜ変化するのかという問いに答えるのは容易ではない。ここでははじめに 2.1 言語的要因と 2.2 社会的要因の観点からこれまで述べられてきた主な変化理由を整理したうえで、2.3 でこれまで言及されることの少なかった言語形成期の人間をとりまく言語環境要因について私見を述べる。

２.１　言語的要因

言語的要因は非常にシンプルである。言語の自律的な変化要因は、つまるところ生理的・認知的な効率化であると整理できるであろう。すなわち言語は、生理的な要因として発音負担を軽減させる方向へと変化する。また認知的な要因として文法面での記憶負担を軽減させる方向へと変化する。特に文法面では類推や再分析などによって基本的に整合性のとれた表現へと変化する。（真田（編）,2006、渋谷,2008、大木,2013）。

２.２　社会的要因

社会的要因には様々なものがある。分類の仕方や今後の研究によっては、次の４要因以外にも様々な要因が考えられうるだろう。

①新しいものや概念の新生・消滅・交替：パーソナルコンピューターが生み出されるまで、パソコンという言葉はなかった。一方、現代では囲炉裏のない家屋がほとんどになり、囲炉裏という言葉が使用されなくなってきた。このように社会の変化によって生み出された新しいものや概念の新生・消滅・交替によって、言語の変化が起きる（大木,2013）。

②言語接触：ある地方で話されていた方言が中央語である共通語と接触することによって次第に使われなくなったり、位相によって使い分けられたりするなど、言語接触によっ

て言語は変化する（真田（編）,2006）。

③アイデンティティー管理・威信：同じ一人称でもわたくし、わたし、あたしでは表現価値が異なる。人間は自己のアイデンティティー管理のために言語変異形の使い分けを行うため、その管理行為が言語変化を生み出す（福島,2002）。

社会的に高く位置づけられている言語形式が持つ魅力を顕在的威信、低く位置づけられている形式が持つ魅力を潜在的威信と呼ぶ。若年層は潜在的威信を志向する傾向があるため、潜在的威信を持つ変異形が若年層に取り入れられ、顕在的威信に従う保守層との世代交代が進むことで言語変化が進行する（真田（編）,2006、渋谷,2008）。

④インターアクション：伝達の相互行為を行う中で、コミュニケーションをより効率化させるために言語は変化する。生理的な労力の軽減と異なり、私立（わたくしりつ）、市立（いちりつ）などのように音韻数が多い形に変化することもある。また絶対敬語から相対敬語に変化するなど、社会的システムや文化的特徴に合わせて変化することもある（大木,2013、渋谷,2008）。

以上、ここでは言語的要因と社会的要因に分けて変化の要因を整理した。しかし言語的要因と社会的要因とは区別しにくい点も多い。例えば開かれた都市社会では複数の方言が接触した結果、複雑な形式や規則が淘汰され、比較的単純な言語システムである「コイネ」が発達することが多いという（渋谷,2008）。そして井上(1998)では、共通語のモデルとなる東京語が様々な方言を取り入れて簡略化・合理化していく様相が詳しく述べられている。

コイネが発達するにはインターアクションが不可欠である。しかし、言語的要因で生み出された様々な言語変異も社会に定着する過程でインターアクションによる淘汰を受ける。あらゆる言語変異が生き残るのではなく、比較的単純で効率的な変異だけが社会に定着すると考えれば、インターアクションによる淘汰も言語の自律的変化であると言えよう。

２.３　言語形成期の人間を取り囲む言語環境要因

図１は、言語形成期の人間を取り囲む言語環境をモデル化したものである。色が薄いほど言語変化が進み、色が濃いほど保守的な言語を使用していることを表している。言語形成期の年齢を定義することは難しいが、ここでは幼少期から十代前半を想定している[1]。

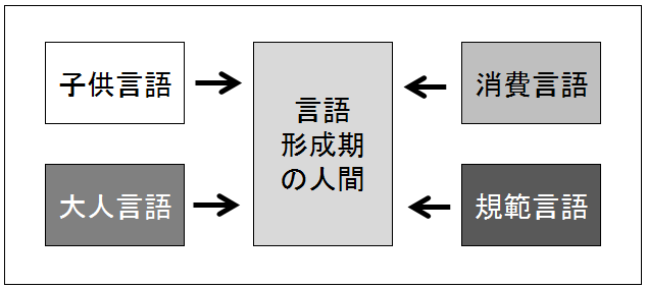


図１　言語形成期の人間を取り囲む言語環境モデル

図１の子供言語は、言語形成期にある人間が同世代を中心とした子供同士でコミュニケーションを行う際に使用する言語を意味している。乳幼児期における親との接触を別にすれば、言語形成期において最も頻度の高い言語活動が行われるのは、この子供言語であろう。子供言語は言語能力が不十分な言語形成期にある子供たち同士が使用するため、類推や再分析などによる言語変化が生じやすく、受容もされやすい。また、若年層は潜在的威信を好む傾向があることから、保守的な言語より変異形が使われやすい。

[1] 基本的な言語は幼少期から十代前半ごろに形成されると思われるが、フォーマルな言葉遣い、敬語、書き言葉などの運用能力は、それより高い年齢で形成される可能性が高い。その意味である位相の言語を習得する過程にある人間は、すべて言語形成期の人間と言えるかも知れない。ここではごく基本的な言語の習得を念頭においている。

消費言語は、社会に流通する言語の中で特に子供向けの書籍・漫画・テレビ・音楽などに使用される言語を意味している。つまり社会に流通する言語の中で、言語形成期の人間が好んで消費する言語である。子供は、ニュース番組や知的なドキュメンタリー番組より、アニメやドラマなどを好んで見ることが考えられる。そしてそれらの番組で使用されている言語には、保守形より変異形が使用されることが多い。さらに漫画やアニメなどより若年層向けに特化して作られたものでは、若年層が潜在的威信を好むことを見越して、一般には使用されていないような、より革新的な言語が使用されることもある。

大人言語は、言語形成期の人間とコミュニケーションを行う機会の多い大人たちが使用する言語を意味している。一般に大人の言語は保守形が多いと考えられやすいが、子供たちが身近に接する親や習い事の先生などは、社会全体の平均的な年代より若いと思われる。また子育て世代の母親たちは子供言語の影響を受け、言語を変化させやすいことも知られている。言語形成期の人間が接する大人言語は、社会平均より変化が進んでいるだけでなく、子供言語で生じた言語変異を受け入れやすい環境である可能性がある。

規範言語は、教育で使用される言語やフォーマルな位相で使用される言語を意味している。言語形成期の人間には規範言語の習得が強く要求されるため、規範言語が言語形成期の人間に与える影響は非常に大きいという考え方がある。しかし規範言語の習得はフォーマルな位相で使用すべき言語を習得しているだけで、日常的に使用される言語に与える影響はさほど大きくないという見方もできる。実際に言語形成期の人間が規範言語を使用して生活する場面はそれほど多くない。さらに学習者の能力によっては規範言語を十分に習得できないケースもある。このように考えれば、規範言語が言語形成期の人間に与える影響は限定的であり、生活の多くの場面で使用する言語にはさほど影響を与えないことも考えられる。

以上、言語形成期の人間を取り囲む言語環境には、子供言語、消費言語、大人言語、規範言語の 4 つが考えられる。このうち子供言語、消費言語、大人言語は、社会の平均的な世代が使用する言語より、変化の進んだ状態にあることが想定される。また規範言語も日常的に使用する言語にはそれほど大きな影響を与えていないことが考えられる。

若年層ほど言語変化が激しいのは、この世代が言語変異を生み出しやすく、潜在的威信を好むためでもあるが、若年層の言語形成に影響を与える言語環境で、すでに言語変化が進行しているからだと考えられる。言語変化の多くが世代交代によって起こるとすれば、言語形成期の人間が言語を習得する環境そのものに、言語変化を進行させる仕組みが組み込まれている。この言語環境要因も言語を変化させる社会的要因の一つと言えるであろう。

３．方言の共通語化における世代効果と経済成長率

前節では先行研究で説明されてきた言語変化の要因を整理し、社会的要因の一つに言語形成期の人間を取り囲む言語環境要因があることを述べた。言語環境要因が働く限り、特別な社会変動が起きなくても言語は常に変化していくと考えられる。そして言語環境要因に特別な変化が生じない限り、その変化の割合はほぼ一定であると思われる。

しかし、鶴岡方言をコーホート分析した中村・米田ほか(2011)によると、鶴岡方言の共通語化率は、各世代でそれぞれ異なっている。方言の共通語化は言語形成期の人間が生み出した変化ではなく、言語接触によって生じた変化である。その変化率が世代ごとに異なるということは、言語環境に刻々とした変化が生じていたことを意味するのであろうか。

ここでは初めに 3.1 で、コーホート分析の考え方を紹介する。3.2 では中村・米田ほか(2011)の研究内容を検討し、世代によって共通語化率が異なる理由の仮説を設定する。3.3 では仮説を検証するための分析の枠組みを述べ、3.4 で分析結果を述べる。

３.１　コーホート分析の考え方

コーホート分析とは、社会調査や市場の分析などに使用されている分析法で、社会や市場の変化を「時代による変化（時代効果・Period）」「加齢による変化（加齢効果・Age）」

「世代差による違い（世代効果・Cohort）」の 3 つの要因の効果に分解して分析する方法である。コーホート分析の背景には、子供時代に形成された感覚や性癖は、生涯にわたって影響し続けるという考え方がある。そのため同じ時代背景を元に成長期を過ごした世代であれば、基本的に行動パターンは一定であると考える。これに基づいて例えば団塊の世代をターゲットとしてある商品を売り出したいとき、この世代に特有な行動傾向を分析してプロモーションを行うなどの戦略に利用されている。

加齢効果とは世代や時代にかかわりなく、人間の生理的な側面やライフステージと関連して変化していく効果のことである。学生時代は若者言葉を話していたが、社会に出てからは社会人らしい言葉遣いになるなどの変化として現れる。加齢効果のみが働く場合は、個人は大きく変化しても社会の年齢構成は緩やかにしか変化しないため、社会的な言語変化は少ない。

時代効果は年齢や世代を問わず、社会全体が同じ方向に変わっていく効果のことである。パーソナルコンピューターが発明されて誰もがパソコンという言葉を使用するようになるなどの変化として現れる。時代効果が著しい場合は、社会全体が大きく変化する。なお社会環境の変化は生育環境の変化でもあり、次世代との世代差をもたらすことが考えられる。

世代効果は生まれ育った時代環境を反映した、他の世代と区別できる特徴を生み出す効果のことである。戦前に言語形成期を迎えた世代と戦後に言語形成期を迎えた世代では仮名遣いが異なるなどの変化となって現れる。世代効果が大きい場合、世代交代によって社会全体は緩やかに変化していく。

言語変化研究におけるコーホート分析では、言語変化の要因として時代・加齢・世代の 3 つを考え、それぞれに効果を仮定したうえで、研究対象となる言語変化を説明するモデルを決定する。

３.２　中村・米田ほか(2011)の検討と仮説の設定

国立国語研究所は長年にわたり世界的にも貴重で精密な言語変化の調査を行ってきた。その成果の一つが山形県鶴岡市における共通語化の実態を調査した「鶴岡調査」である。「鶴岡調査」は 1950 年の第 1 回調査から 2011 年までほぼ 20 年おきに 4 回実施され、31 種類の音声項目とアクセント 5 項目などについて約 400 名のサンプリング調査と同一人物への追跡調査を行っている。図 2～4 は第 3 回までの調査のうち、音声項目 216（番）の「スイクゥ→スイカ」の発音の変化をコーホート分析し、それぞれ時代効果、加齢効果、世代効果の別に分離してグラフ化したものである（中村・米田ほか,2011 より転載。縦軸はロジット尺度[2]、横軸は図 2 が実年代、図 3 が年齢、図 4 が生年代である）。

図 2 の時代効果は、調査年の 1950 年、1971 年、1991 年の 3 点でプロットされている。これを見ると 1950 年と比較して 1971 年では時代効果は大きく上昇したものの、1991 年では逆に下降していることが見て取れる。時代によって共通語化を促進する力に強弱があったことになる。図 3 は加齢効果である。このグラフはフラットなので、スイカの音声項目では加齢効果は認められていない。図 4 は世代効果で、調査対象者の生年を 8 年で一つの世代として区切り、それの中央値がプロットされている。実際の世代効果はもっと滑らかな推移を示すと考えられるが、ここでは 8 年幅で区分したモデルが選択されており、階段状になって示されている。

注目してほしいのは、図 4 の世代効果である。図 4 では各世代のグラフの傾きが一定でなく、それぞれ異なっている。これはなぜであろうか。先に考察したように、言語環境要因に特別な変化がなければ、その変化の割合は一定であると思われる。時代の進行とともに共通語を話す親などが少しずつ増え、そのような親に子育てされたり、周りの子供たち

[2] ロジット尺度とは、共通語化率を定数＋加齢効果＋時代効果＋世代効果で表すモデルを、log[π(i,j)/{1-π(i,j)}] = β^G + β^A(i) + β^P(j) + β^C(k)と数式化した際、β^A(i)などの値に使用される尺度。i は年齢区分、j は調査年。

の中で共通語を話す人間が多くなれば、次第に言語形成期の人間の多くが共通語を話すようになる。やがてその子たちが親の世代になれば、さらに共通語を話す人間が増えていく。しかしそのような要因による変化であれば、世代によるグラフの傾きはほぼ一定であると思われる。各世代によってグラフの傾きが異なるということは、言語環境要因が時代によって刻々と変化したことを意味しているのではないだろうか。

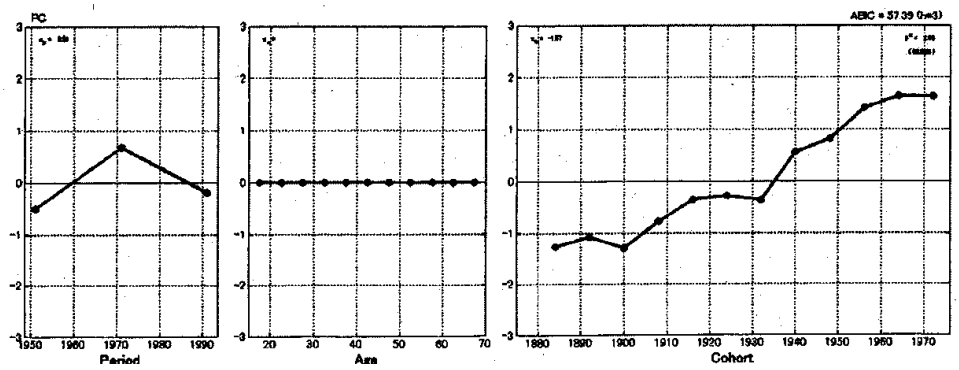


図 2　時代効果　　図 3　加齢効果　　図 4　世代効果
（中村・米田ほか(2011)より引用）

言語環境要因が変化する理由の一つとして、方言に対する社会的な言語意識を反映した教育の変化が考えられる。明治期から現代までででは、方言に対する社会的な言語意識が変化している。明治期では方言が日本の近代化を妨げるものとして撲滅運動が起こった。一方、方言が消滅の危機にさらされている現代では、方言の保護運動が行われている。このような方言に対する言語意識の変化を受け、学校教育で共通語を強く指導された世代もあれば、方言の大切さを教えられた世代もあると思われる。しかしそれが原因であれば、古い年代ほど共通語化率の傾きは急になり、現代に近くなるほど傾きが平らになると思われるが、図 4 では 1930 年～1940 年頃が最も急な傾きになっている。ただし古い年代では就学率が低く、年代が新しくなるにしたがって就学率が高くなると考えると、古い年代の変化率が新しい年代より低くなることはありうる。しかしそのように考えても 1900 年、1920～1930 年、1960~1970 年などで世代効果が低下する理由は説明できない。

この他の要因ではラジオやテレビの普及なども考えられる。これまでもテレビの普及によって共通語化が進んだことはよく指摘されてきた。ラジオ放送が始まったのは 1925 年、テレビの普及が進んだのは 1950 年代である。この頃に言語形成期を迎えた世代はおよそその 10 年前の 1915 年～1940 年生まれとなる（10 歳で考えた場合）。しかし図 4 を見ると、1915 年～1930 年頃に生まれた世代は、ほとんど共通語化率が上がっていない。1940 年~1950 年頃に生まれた世代はその前の世代より共通語化率が低くなっている。この世代はラジオに加えテレビまで普及し始めたのであるから、当然共通語化率が上昇すると思われるが、前の世代より共通語化率が低くなっている。ラジオやテレビの普及率と世代効果はなぜ連動しないのであろうか。

そこで着目したいのが図 2 の時代効果のグラフである。これについて中村・米田ほか(2011)では、上昇の要因はテレビの普及、下降の要因は今後の検討課題としている。確かに共通語化率を変化させる要因に教育やテレビの影響があるとは思われるが、連動の低さを考慮すれば、それらの影響力を左右するさらに高次の影響力が働いている可能性も検討する必要がある。1971 年に起こった社会的変動ではニクソンショックが有名である。アメリカ政府は 1971 年 7 月 15 日にブレストン・ウッズ体制の終結を宣言し、日本は固定為替制から変動相場制への転換を余儀なくされた。この年 1 ドル 360 円だった円は 1991 年には 130 円台になっている。日本が飛躍的に経済成長を遂げたのは 1954 年から 1973 年頃であるといわれる。それ以降は一貫して経済成長率が低下していく。このような経済の影響が言語の

変化に影響を与えている可能性はないであろうか。言語を使用しない経済活動は考えにくいため、経済が活発化すれば言語活動も活発化し、経済が停滞すれば言語活動も停滞すると思われる。

さらに近年、経済学と心理学を結びつけた行動経済学が盛んになり、人間の心理と経済に密接な関係があることが分かってきた。行動経済学者にとって景気変動が人間の心理に影響を与えるのは、もはや常識なのかも知れない。関沢・桑原(2012)では、「景気変動を始めとする経済現象が人間の感情に影響を及ぼすことはほぼ異論がないと思われるが,その逆の因果関係は存在するだろうか.」(p118)として、人間の感情が抑うつ的になると消費行動が消極的になる可能性を論じている。

経済が活性化することによって人間の心理がポジティブでリスク愛好的になるのであれば、言語変化も受容しやすくなることが考えられる。逆に経済が不活発になった時、人間心理がネガティブでリスク回避的になるのなら、言語変化が受容されなくなる可能性がある。方言に対する教育やテレビの影響なども経済的活性度の影響を受けていると考えればこれまでの問いがうまく説明できると思われる。そこで以下の仮説を立てる。

(1) 言語変化における時代効果と世代効果を変動させる要因の一つに経済成長がある。

３.３　分析の枠組み

経済成長の指標にはいくつか考えられるが、ここでは最も一般的と思われる経済成長率を使用する。時代効果については中村・米田ほか(2011)では観測ポイントが３箇所しかなく、分析には適さない。そこでここでは世代効果のみを分析の対象とする。また、分析に当たっては経済成長率の作用年代も変数に加え、詳しい分析データが得られる重回帰分析を行う。リサーチクエスチョンは「世代効果は、経済成長率で説明できるか」である。

ただし、現在使用されている一貫した指標で得られる経済成長率は、1956 年（内閣府　平成 2 年基準）までしか遡れない。それ以前の 1931 年までは概念も推計手法も異なる国民総支出（経済企画庁　『国民所得白書　昭和 40 年版』）から推計するしかなく、さらにそれ以前となると大川・高松・山本推計などの学術研究の成果を利用するしかない。これらの異なる指標を単純につなげただけでは、専門的には問題があるとの指摘があるものの、その調整は難しい。そのため、ここでは 1955 年以前は総務省統計局の『日本長期統計総覧　第 3 巻』の数値を、1956 年以後は内閣府平成 2 年基準の数値を利用して分析する[3]。

世代効果の数値は図 4 のグラフから読み取ったものを使用する。図 4 ではコーホート区分幅が 8 年で、グラフに記された点はその区分の中央値が掲載されている。つまり満 1 歳～満 8 歳を一つの世代とし、その代表値を満 4 歳の値としたということである。これと比較する経済成長率は満 1 歳時～満 8 歳時の経済成長率の平均とし、その中央の年を作用年代と呼ぶことにする（つまり 1881 年～1888 年なら 1884 年が作用年代）。また、時系列分析では平準化を行わないと分析に適しないため、各々の年代の差分を取って分析する。

ここで問題となるのが、生年代と経済成長率の作用年代を何年ずらせばよいかである。これは言い換えれば「スイカ」の発音が何歳で固定化されたのかという問題である。これについては、0 歳から 20 歳までの分析を行い、統計的に有意になったものを採用した。結果的にスイカは満 12 歳の値だけが有意となった。

３.４　分析結果

スイカの世代効果の説明変数には経済成長率のみが選択され、修正済み決定係数は.409 となった。ダービン・ワトソン比とは残差の中に系列相関がある可能性を示す指標で、時系列分析の際に重視される。この数値は 0～4 の値を取り、2 に近いほど良好とされるが、

[3] 経済成長率は暦年（1 月から 12 月まで）と、年度（4 月から翌年の 3 月まで）の集計があるが、分析に使用したのは 1947 年～1952 年を除いて暦年のデータである。また入手可能な経済成長率は 1886 年以後となる。

この値も 1.907 で信頼性の高い結果となった。この結果によればスイカの世代効果は、経済成長率で 4 割ほど説明できることになる。わずか 1 例の二次分析では実証性に乏しいが、ひとまず仮説を立てた二要因のうちの一方は支持されたと思われる。

表 1 スイカの世代効果と経済成長率の重回帰分析結果

	経済成長率の回帰係数	定数の回帰係数	p 値	自由度	修正済み決定係数	ダービン・ワトソン比
スイカ	0.079	0.262	0.020	10	0.409	1.907

観測値と推定値のグラフを比較すると、1952 年以前で乖離が目立った。正確な経済成長率のデータが得られるのは 1956 年以降であることを考えると、より正確な経済成長率のデータがあればさらに説明力が高まる可能性がある。多くの言語データで世代効果と経済成長率の相関が確認されるなら、やがて言語データを使用して古い時代の経済成長率が推測できるようになるかも知れない。この分析結果がどのような意味を持つかについては、次の時代効果の分析結果も含めて 5.で考察する。

４．「図書館書籍」における出版年ごとの変異形出現率と経済成長率

世代効果については、その要因の一つに経済成長率がある可能性が示唆された。しかし、時代効果については、時点数が少なく分析できなかった。時代効果を分析するためには一定期間にわたる均衡なデータが必要である。長期の話し言葉のデータとしては国会会議録が考えられるが、1 年ごとに精度の高いデータを得るには困難が予想される。そこで時代効果の分析に当たっては、『現代日本語書き言葉均衡コーパス』「図書館書籍」の固定長データを使用する。書き言葉は規範意識が働きやすく、編集作業なども行われるため、言葉の乱れと意識されやすい変異形は出現しにくい。しかし、1986 年～2005 年という 20 年間のスパンで、統計的に均衡なデータが得られるのは大きな魅力である。「図書館書籍」で抽出した変異形出現率（変異形数÷（保守形数＋変異形数））が、出版年ごとの経済成長率に連動するなら、経済成長率が時代効果の要因の一つである可能性が高い。現段階では変異形出現率のコーホート分析ができていないため、観察対象は時代効果を分離する前の生データとなるが、ここでは出版年別に分析した変異形出現率を時代効果とみなすことにする。以下、4.1 で分析の枠組みを述べ、4.2 で分析の結果を述べる。

４.１　分析の枠組み

分析には用例数が多く、比較的規範意識が働きにくいと思われるヤハリ→ヤッパリ、デナイ・デハナイ→ジャナイ、ノデス→ンデスを使用する。語彙素「矢張り」で抽出するとヤッパシ、ヤッパなどの変異形も出現するが、これらもヤッパリに含める。ジャナイはデナイ・デハナイを含めた 3 形式を抽出しジャナイの割合を求める。ノ／ンデスは語彙素「の」＋語彙素「です」で抽出し、ンデスの割合を求める。また経済成長率には平成 12 年基準の暦年を使用する。

はじめにこれらの変異形出現率が、前年の経済成長率と連動しているのかをグラフで確認する。次に、これらの保守形と変異形が経済成長率の増加とともにどのように増減するかを、各年の短単位 100 万語当たりの出現率（WPM）のグラフで確認する。最後に世代効果の分析と同様に、経済成長率と出版年を説明変数とした重回帰分析を行う。

４.２　分析結果

４.２.１　変異形出現率と前年の経済成長率

図 5 は全てを重ねて図示したものである。変異形出現率はンデス率が最も高く 45%前後、次いでヤッパリ率の 35%前後、最も低いのがジャナイ率で 25%前後となっている。一番下が経済成長率で 2%前後である。しかしこれでは比較しにくいため、各変異形の 1985 年の

割合が 10%になるように相対化したのが図 6〜8 のグラフである。最も前年の経済成長率と似ているのが図 6 のヤッパリ率で、ここまで連動していると単なる偶然とは見なしにくい。

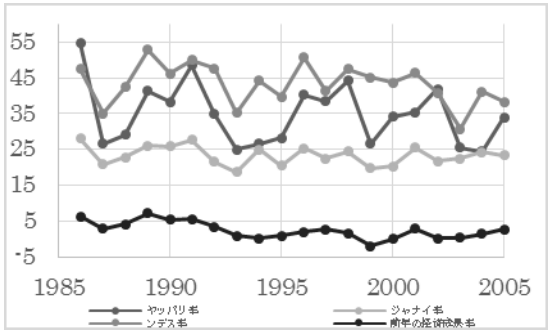


図 5　各変異形の割合と前年の経済成長率

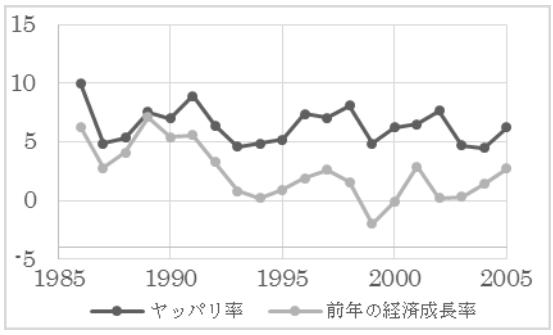


図 6　相対化ヤッパリ率と前年の経済成長率

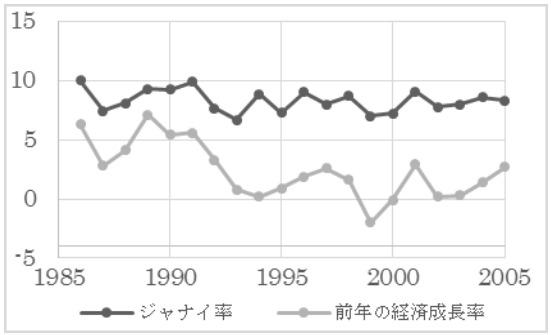


図 7　相対化ジャナイ率と前年の経済成長率

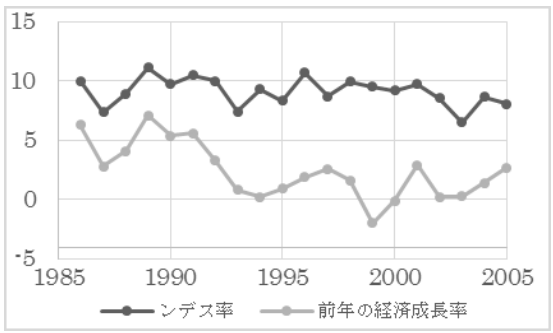


図 8　相対化ンデス率と前年の経済成長率

ここでもう一つ着目されるのが普及の S カーブとの関係である。これまでの先行研究では変異形の普及は slow-quick-quick-slow の S 字型カーブを描くとされてきた。各変異形の割合からするとジャナイ率は slow から quick に移行するあたり、ンデス率やヤッパリ率は quick が加速するあたりに位置するはずである。しかしいずれもこの 20 年間でわずかながら減少している。この事実からは自律的な普及力以外の別の力が働いていることが示唆される。

４.２.２　経済成長率と短単位 100 万語当たりの出現数

図 9〜11 は、縦軸が各出版年の短単位 100 万語当たりの出現数（PMW）、横軸が経済成長率である。デナイ・ジャナイは 2 短単位、デハナイは 3 短単位として計算した。これらのグラフで振幅が少ないものほど経済成長率とよく連動していることになる。反対に振幅が激しいものは経済成長率以外のものにも影響されている可能性が高い。

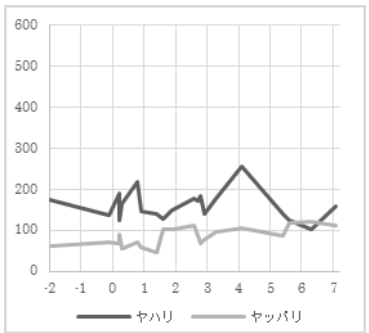


図 9　ヤハリ・ヤッパリ PMW

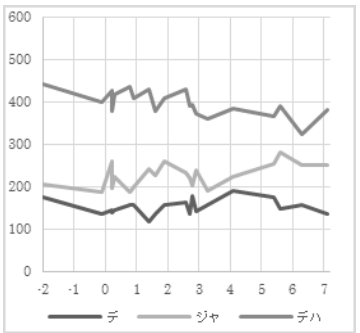


図 10 デナイ・ジャナイ・デハナイ PMW

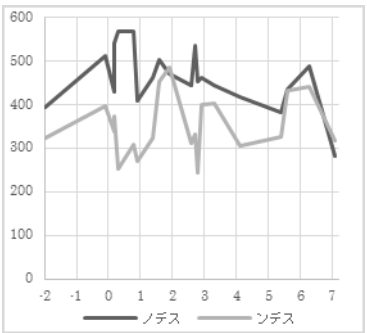


図 11 ノデス・ンデス PMW

これらの中では、ヤッパリ、ジャナイ、デハナイが比較的きれいな連動を示した。ノデス・ンデスにも経済成長率に連動した一定の傾向は見えるが、これだけ激しく振幅するということは、経済成長率以外の要因でも使い分けがなされている可能性が高い。ただしいずれの形式でも経済成長率が増加するに従って、保守形が減少すると同時に変異形が増加する傾向が観察できる。

４.２.３　変異形出現率の重回帰分析

重回帰分析の結果、いずれの形式も説明変数には経済成長率だけが選ばれ、ヤッパリ・ジャナイの説明力はスイカの世代効果と同程度の約 4 割となった。図 11 で激しい振幅が観察されたンデスの説明力はやはり低く、24%にとどまった。しかし、いずれも一定の説明力を持つことから、仮説(1)は支持されたと考えてよいだろう。ただしジャナイ・ンデスのダービン・ワトソン比がやや高いため、これらはなお詳細な検討が必要かも知れない。

表 2 変異形出現率と経済成長率の重回帰分析結果

	経済成長率の回帰係数	定数の回帰係数	p 値	自由度	修正済み決定係数	ダービン・ワトソン比
ヤッパリ	3.640	−0.392	0.002	18	0.410	2.575
ジャナイ	1.385	−0.088	0.002	18	0.406	3.130
ンデス	0.021	−0.001	0.019	18	0.240	3.163

５．考察

ここでははじめに時代効果と世代効果との関係を考える。時代効果はその時代に生きているすべての人間に及ぶ効果である。一方世代効果は、団塊の世代なら同世代の膨大な人口、ゆとり世代なら他の世代とは大きく異なる教育指導要領の方針などのようにその世代を特徴づける原因に由来すると考えられる。しかし今回の調査ではそのどちらもが経済成長率と相関していた。人間は言語形成期の経済状況にもその後の経済状況にも影響を受けながら言語を変化させていく。ただし言語形成期はその期間の経済成長率の影響力が強く固定化される。これが世代効果である。一方その後に受ける影響はその時代時代で移り変わるだけか蓄積されるとしても変化の余地がある。これが時代効果であると思われる。

スイカの発音の共通語化には教育やテレビの影響などが考えられるが、それらの変動は必ずしも世代効果と連動しなかった。これを図 1 のモデルで考えれば、規範言語や消費言語側からの働きかけは、直接的には言語形成期の人間の言語能力と連動しないということである。その一方で世代効果は経済成長率と連動していた。そして景気変動は人間心理と深い関係があった。これは子供言語の使用者である言語形成期の人間や大人言語の使用者たちの心理が経済成長率に連動して変化している可能性を示唆する。

第二言語習得研究では、言語習得において動機づけが重要な役割を果たすことが指摘されている。経済成長率の変化に連動して共通語を習得する動機づけが変動するなら、共通語の習得結果＝共通語化率が経済成長率に連動して変化する理由が説明できると思われる。

経済成長は全国一律に起きる変化ではない。まず都市部が発展し、その後地方に波及する。経済成長のさなかでは都市部と地方の格差は拡大する。この時、地方に住む子供たちが都市部にあこがれを持つことは想像に難くない。またその親をはじめとする周囲の大人たちも、将来子供は地方で暮らすより、都市部で暮らした方が幸福だと考えやすくなると思われる。都市部に対するあこがれは、共通語を習得する際の統合的な動機づけや内発的動機づけとなる。さらに都市部に出ていくためには共通語を覚える必要があると考えれば、道具的動機づけも働く。また、親や周りの大人たちもそれを望むなら外発的動機づけも与えられる。つまり言語習得に影響を与えるすべての動機づけが働くことになる。これがスイカの世代効果が経済成長率に連動する最大の理由であると思われる。共通語にはクルマ

ス(1993)の述べる富を背景にした普及力があるということであろう。

一方 2.2 で述べたように、共通語のモデルとなっている東京語は、複数の方言が接触した結果、複雑な形式や規則が淘汰され、比較的単純な言語システムであるコイネとなっている。東京語のコイネ化は経済成長が進むほど進行する。地方で経済成長率に連動して共通語化が進むということは、経済成長率に連動して効率的な言語が全国に普及することに他ならない。クルマス(1993)の述べる二つの経済性はこのようにして結びつくと考えられる。

それでは「図書館書籍」の変形出現率と経済成長率の関係はどのように考えればよいのであろうか。詳細は稿を改めて述べるが、「図書館書籍」では出版年ごとの文学ジャンルの増減と経済成長率に高い相関があった。この理由には景気が良い時代には文学がよく売れたことなどが考えられる。ここから類推すると、経済が活性化するに従って購買者層が拡大し、普段は書籍を購入しないような読者層が、馴染みやすい口語的な文体で書かれた書籍を多く購入したことなども考えられる。そしてその年によく売れたか代表的であるとみなされた書籍が、最終的に図書館に収蔵されることになったと思われる。ただし景気の変動によって読者層の拡大や縮小が起きるだけでは言語変化とは無関係である。しかし出版社が読者層の拡大を求めてより多くの人間に受け入れられる書籍を出版するとしたら、書籍に使用される言語は変化していくと考えられる。現段階では推測の域を出ないが、このような市場による淘汰と富を追及する生産者とのインターアクションが、変異形出現率と経済成長率との相関となって現れている可能性がある。これらの仕組みの解明は今後の課題であるが、経済成長率が高くなると変異形出現率が増加することは事実である。また、デハ→ジャなどの変化は言語的要因による効率化と考えられる。よって「図書館書籍」においても富の拡大によって言語の効率化が進んでいるとみなすことができるであろう。

共通語には、富を背景にした普及力があると考えられる。そして共通語が普及するということは、言語の効率化が進むことだと考えられる。「図書館書籍」でも、経済成長率に連動して変異形が増加する現象が観察された。そしてその変異は言語の効率化によって生じたものだと考えられる。このことから「図書館書籍」でも、富の拡大によって言語が効率化していく可能性が推測される。以上より、方言の共通語化における世代効果と「図書館書籍」における出版年ごとの変異形出現率に経済成長率との相関が見られるという現象は、クルマス(1993)の述べる二つの経済性を結びつける知見であると考えられる。

謝　辞

中村・米田ほか(2011)の分析に当たっては、中村隆氏より不明点をお教えいただいた。また、重回帰分析に当たっては石川慎一郎・前田忠彦・山崎誠（編）(2010)『言語研究のための統計入門』 くろしお出版 所収の Seagull-Stat を、コーパスデータの分析に当たっては田野村忠温氏開発の sortKWIC を使用した。皆様に深く感謝申し上げます。

文　献

クルマス,フロリアン(1993)『ことばの経済学』大修館書店（原著は 1992 年刊）
大木一夫(2013)『ガイドブック日本語史』ひつじ書房
真田信治（編）(2006)『社会言語学の展望』くろしお出版
福島直恭(2002)『〈あぶない ai〉が〈あぶねえ e:〉にかわる時―日本語の変化の過程と定着』笠間書院
渋谷勝己(2008) 金水敏ほか『日本語史のインターフェース』第 4,5,6 章　岩波書店
井上史雄(1998)『日本語ウォッチング』岩波書店
中村隆・米田正人・阿部貴人・前田忠彦(2011)「「鶴岡市における共通語化の調査」データのコウホート分析―ベイズ型コウホートモデル(XXII)―『日本行動計量学会大会発表論文抄録集』39　pp.121-122.
関沢洋一・桑原進(2012)「感情が消費者態度に及ぼす影響についての予備的研究」『行動経済学』第 5 巻 pp.118-136.

形容詞による属性叙述表現の構造に基づく意味分析

朴 俌傑（東北大学大学院）

A Semantic Analysis on the Property Predication of Adjectives

PARK, Bokyung (Graduate School of Arts and Letters, Tohoku University)

１．はじめに

形容詞が述語になる形容詞文は「対象表示部分＋解説（属性表示部分）」で構成される(益岡 2000)。形容詞文の構造的な形式の区別に関して、野田(1996: 135)では「比較」の表現を基準に、無題文は比較を表す傾向が強く、有題文は比較を表さない文として位置づけられている。しかし、有題文を無題文と区別するために、無題文の性格を持たないものとして捉えることは形容詞文の有題文の性格を明確にしているとは考えにくい。形容詞文の有題文に焦点を当て事物の属性の分類に基きその性格を論じている研究に益岡(2000)がある。属性の分類には事物の本質的な属性としての「内在的属性」と、一定の時間的限定のもとで成り立つ「非内在的属性」がある。事物の内在的属性に対しては(1a)の有題文では事物の属性を、(1b)の無題文では事物の一時的状態を表現すると益岡は論じている。なお、事物の非内在的属性(2)と感情形容詞(3)にも内在的属性と同様の基準で有題文と無題文が区別される。

(1) a. 雪は白い。　b. 雪が白い。　(益岡 2000: 45)
(2) a. あたりは騒がしい。　b. あたりが騒がしい。　(益岡 2000(45)を改変)
(3) a. この話は怖い。　b. この話が怖い。　(益岡 2000(46)の改変)

事物の内在的属性の場合は、属性が科学的に証明ができ一般に共有されているものであるため有題文の性格が無題文とはっきり区別できる。しかし、非内在的属性と感情形容詞による有題文の属性叙述表現は何をもって事物の属性の表現として読み取れるかが不明である。本研究では、こうしたあいまいさを事物の属性を認識する話し手（主体）に焦点を当て有題文と無題文の性格を明確にする。

以下では、まず形容詞による属性叙述表現の性格を主体の事物の属性に対する認識に焦点を当てて分析する必要性について述べる（第２節）。それを実証的に分析するために、コーパスを用いて形容詞の属性叙述文を抽出し（第３節）、形容詞の属性叙述文と共起する表現を用いて意味的特徴の分析を行う（第４節）。

２．主体の環境との相互作用としての事物の属性

形容詞による属性叙述表現に関しては、その構造が格助詞の「は」と「が」の形式を持つため、格助詞選択の問題の一部としてされている(野田 1996)。格助詞「は」と「が」に関して様々な観点が提唱されているが、異なる機能を持つ品詞においてどの観点が優先的であるかを検証することも必要であるだろう。属性叙述文の研究のもう一つの限界として、形容詞に焦点を当てることはあまりなされず、動詞による事象表現に大きな注目が集まる傾向が続いていることである(益岡 2008)。形容詞はよく名詞か動詞の下位分類として扱われるが、事物の静的な属性を詳細に表現できる形容詞の特徴を認め、独立した品詞としての形容詞に目を向ける必要がある。

形容詞の意味は対象に備わる事物(物事も含む)の性質として捉えられ、対象の事物と属性の関係に依存している。上記の益岡(2000)において属性叙述文の性格を記述する際に用いられている属性の概念は事物の性質が反映された。しかし、属性の認知に関わる主体の環境を考慮すると、属性は生き物とも無関係に存在するものではなく、人間と環境の相互作用の結果として理解されるべきである(Lakeoff 1987: cf. 深田・仲本 2008)。例えば、事物の「カバン」に対する「重い」の属性は、カバンその物の重量として述べられるというより、カバンを運ぶなどの活動に関わるときに初めて認識される属性である(仲本 2006)。このように、形容詞によって表される事物の属性の表現は、事物の性質として表されるのではなく、背後にある事物の認知者(主体)の環境において解釈される属性と言える。

事物の属性を主体の解釈として扱う際にさらに考えるべきことは、主体の捉え方である。「カバン」の「重い」という属性に対して、主体は状況によらず常に重いと受け取ることもある一方、状況によってはそう思わないものとして捉えることもある。文法的使用条件と関係せず認知者が物語の登場人物をどのように捉えるかという観点が格助詞「は」と「が」の使い分けに関与することが報告されている(ゲオルギエバ 2008)。

以上のことから、形容詞による属性叙述表現の構造、つまり有題文と無題文の異なる形式を事物の属性に対する主体の認識の相違が反映されるものとして研究していくことが求められる。そこで本研究では、抽象的な主体の認識を実証的に記述できるように、コーパスを用い、属性叙述文の意味分析を行う。それによって、それぞれの形式に主体のどのような意図や観点が反映されるかを明らかにする。

３．データと処理手順

３．１　調査対象

調査対象とするデータには『現代日本語書き言葉均衡コーパス』を用い「中納言」を介して例文を抽出した。対象は形容詞による属性叙述文で、「キー：語彙素読みで用いられる形容詞を入力し、前方共起：キーから３語、品詞は名詞の設定」で検索した。このうち、形容詞の活用形が終止形(例：風はなお激しい。)と連用形(例：この借金が思いのほか重く、)だけを扱った。そして、事物の属性に対する主体の判断が複数の要因から成り立つことを統制するために、連用形での共起表現を最小限にした[1]。対象表示部分に関しては、主体が現される事例と、「象は鼻が長い」に代表される対象表示部に連体修飾部のある事例を排除した。このほか、野田(1996)で無題文の性質として取り上げられている「～ほうが」と「一番」を排除した。また、用いられる形容詞のサンプルは主観性に着目した細川(1993)の分類に従った。細川によると、形容詞は感覚形容詞、感情形容詞、属性形容詞、評価性形容詞に分けられる。そして、各タイプでは「わたし」の心の様子を表現することが可能な語群と、そうでない語群に分けられる。この下位分類に従い均等に形容詞を選択した。

３．２　調査結果

本調査は事物の認知者である主体の観点を取りいれ、形容詞による属性叙述表現の性格を分析することが目的である。そこで、有題文と無題文において事物の属性の表現とともに共起される表現の意味分析を行う。そのために、形容詞の４つのタイプごとに均等に共

[1] 連用形で扱われる用例は、「重く、重くて、重いと思った。」の３種類である。以下のような例は排除した。例：後頭部が痛かったそうです。知恵袋がとても重く感じますが。

起表現が収集できるように200件程度の事例数を設定し、形容詞を追加しながら事例を検索した。以下の表１に検索に用いられた形容詞を、意味分類によるタイプごとに有題文(は)と無題文(が)の多い順に表示した。

表１　形容詞の意味分類における属性叙述文の分布

感覚形容詞

形容詞	は	が
痛い	17	82
冷たい	29	16
寒い	34	4
涼しい	7	1
臭い	1	2
計	88	105
総数	193	

感情形容詞

形容詞	は	が
嬉しい	23	42
楽しい	43	21
つらい	26	15
悲しい	6	5
恥ずかしい	4	5
寂しい	7	2
計	109	90
総数	199	

属性形容詞

形容詞	は	が
美しい	37	40
重い	27	30
明るい	28	16
高い	9	25
赤い	5	7
計	106	118
総数	224	

評価性形容詞

形容詞	は	が
望ましい	10	64
激しい	22	50
羨ましい	4	27
懐かしい	3	28
計	39	169
総数	208	

有題文と無題文の頻度を概観すると、評価性形容詞の属性叙述文の場合のみ有題文より著しく無題文が多い。これは用いられた評価性形容詞の全てにも同様のパターンが見られる。評価性形容詞以外のタイプでは形容詞によって形式の優位性が異なる。次節からは形容詞の意味分類とは関係せず、有題文の342例と無題文の482例を基に主体のどのような認識が形式の使い分けに関与しているかを見ていく。

４．分析

本研究では、形容詞による属性叙述文の共起表現として、形容詞が示す状態をより明確にする程度副詞や前文の文法成分を主に取り扱う。コーパスから得られた事例の分類の結果、有題文あるいは無題文と共起する表現は択一的に用いられることはなく、その頻度に差が見られた。まずどちらの形式からも高い頻度で見られる共起表現を基に主体の観点の相違を記述し、次により有題文を好む表現、無題文を好む表現を特定する。

４．１　有題文と無題文の主体の観点

有題文と無題文の両方ともに高い頻度を見せる表現として、表２のような程度の高い副詞と比較の成分がある。

表２　有題文と無題文に共通の共起表現

成分	表現	有題文	無題文
程度が高い	とても / かなり / 大変 / なかなか など	74	62
比較	最も / さらに / ますます / 特に　など	22	28

4.3節で述べるように、程度副詞のうち程度の低いことを表す場合(例：少し)は有題文と無題文で頻度に差がみられるが、高い程度を表す、特に「とても」は両形式でどの形容詞タイプでも頻度が高い。しかし、異なる形式から異なる主体の観点が見られる。有題文では事物の属性に対する主体の心理との距離が近いが、無題文では主体の心理に周辺の心情ま

でが含まれる。次の(4a)のように有題文では主体自身の楽しい感情が前文で表され、「嬉しい」の情意は主体中心の心情である。(4b)では前文の「みんな楽しそうである」の表現から、周りの感情を認識しながらの自身の心情が述べられる。（出典はサンプルID）

(4) a. その後はわいわいがやがやしながら、楽しい時間を過ごしました。ちょっと忙しい一日でしたが、楽しい一日でもありました。家に誰かが遊びに来てくれる事はとても嬉しいし、 その人達の食事の用意をする事も嫌いではない、喜んで食べて貰える顔を見るのが好きなくらい・・・ (OY14_23122)

b. まわりを見回すと、なんだかみんな楽しそうである。句会は、俳句が好きな人たちと、たくさんの俳句作品との出会いで成り立っている。俳句という、ただそれだけの縁で出会った思いがけなさ。そんな出会いがとても嬉しいと思える。いちばん大切なことは、同じ場にいるということ、同じ空気を吸うということ。 (PM42_00067)

次に、無題文の性質として論じられる「比較」の表現を用いて有題文と無題文を比較すると、２つの形式には事物の属性と主体との関係性の程度に違いが見られる。(5a)は主体の通学路における属性の判断で主体の日常における認識として主体との関係性が強い。一方、(5b)は教科書で抽出された事例で事物の属性に主体の関与は見られない。

(5) a. 宋世平はこういう辺鄙なところまでカーネル・サンダースを引っ張ってくることはできないんだよな、やっぱり適当なところですませよう、と言った。寮から校門まで歩いて十分かかる。夏の夕暮れは最も美しい。雨翔は南三高の広い道を歩きながら、夕日を見てその美しさに驚嘆していた。 (PB29_00231)

b. 望遠鏡で太陽像を白紙に投影すると、その中心部が最も明るく、周縁部にいくほどしだいに暗くなっている。 (OT23_00083)

以上、有題文と無題文の異なる観点を概観した。属性叙述文と共起する表現はまったく同じであるが、両形式に文脈からの状況と、事物と主体との関係性にはっきりとした相違が見られた。有題文では事物に対する主体の強まった情意と日常の経験と関わる事物の属性が述べられ、無題文では周りの心情を念頭に入れた主体の心情と、主体の間接的な経験の属性が述べられる。ここで注目すべきことは、上記に挙げられた４つの例とも別の形式に置き換えても許容度はそれほど変わらないことである。これは形式の使い分けに文法的使用条件より主体の認識が優先されることを示唆する。

次に、形式ごとに、好まれる共起表現を基に精密な特徴を見ていく。分析は５例以上見られた共起成分を対象にする。

４．２　有題文の共起表現

形容詞による属性叙述表現と共起する表現で、無題文より有題文と好んで共起するものがある。それが表３にまとめたものである。全般的に有題文には主体が事物の属性を持続的なものとして捉える表現が用いられ、またはその属性を主体自身の一部のように受容するような強く率直な心情が表される。

表３　「有題文」を好む共起表現

成分	表現	有題文	無題文
持続性	まだ / ひたすら / 引き続き / 現在でも など	21	6
一般性	概して / 世間一般 / 問わず / 総じて	5	1
時間の経過	もう / すでに	5	0
強い決意	決して / それにしても / かならず / だんぜん / やっぱり / でも	28	5
主体の心情	正直 / さぞ / かくも / 思いっきり	7	1

まず時間副詞を見ていくと、主に感覚形容詞と属性形容詞で「まだ」や「引き続き」などのように事物の属性が現在を基準にして以前から継続する副詞と共起する例が多い。これは、次の一般性の「世間一般」のように時間の継続性を前提とする表現の成分とも類似する。次の(6)の例は両方とも「まだ」が使われ、その形式だけ異なる。有題文(6a)の文脈では主体と事物の属性の関係はないため、無題文の形式に置き換えられる（提示された文脈の前文は無い）。無題文(6b)では渡月橋から山を見上げた時点から認知したその状態に焦点が当てられている。（無題文に見られる主体の観点に関しては4.3節で論じる）この無題文の観点に基づいて有題文の選択を推測すると、有題文の使用を用いて主体の以前の経験の流れの中で物語を始めようとする意図の可能性がうかがえる。

(6) a. 耳元で電話が鳴り続けた。仁科はソファの上で起き上って二、三度強く頭を振った。窓の外はまだ明るかった。 時計は五時ちょっと前を指していた。シャワーを浴びてそのままソファに横になったのだが、三十時間ぶりのシャワーが余程効いたようだ。 (LBd9_00131)

b. 夕暮れの渡月橋から、何気に山を見上げたら、少し時間の早い日没・・・。夕日がまだ明るくて、光り輝いていました。貴金属に例えるならプラチナのような輝き・・・。 (OY14_24682)

主体の以前の経験と関連する性質が「もう」、「すでに」のような表現からも見られる。「もう」は当該事態がその時点で突然認識された属性の表現であるより、主体が事物の属性を認識する前に推定していた事柄と関係がある(cf. 仁田 2002: 255)。

以上の時間副詞の共起表現を踏まえると、有題文には事物の属性に対する認識に主体の内的思考が含まれていると言える。そこで、次には主観性の強い表現を用いて分析する。

否定副詞の「決して」や強い判断の「それにしても」のような表現は有題文で表２の程度の高い副詞以外に最も高い頻度を見せる。「決して」は多くの先行研究によると前提を踏まえた表現であり、それにもかかわらず強く打ち消す気持ちが表される。「決して」が用いられる例として、「先行きは決して明るくはない」、「脳は決して美しいとは言えない」のように客観的な属性を表す形容詞との共起があることが興味深い。「それにしても」、「かならず」、「だんぜん」などの場合は、誰かにいくら反論されるとしても事物の属性に対する判断は曲げないという気持ちが含まれる。これらの表現に対して無題文では「やっぱり」の例しかなく、その用法は有題文のそれとは異なる。有題文の(7a)では主体であるアイドルグループのファンがツアーでいつもどおり満喫した感情がうかがえる。無題文の(7b)ではその瞬間

改めて考え直す様子が読み取れる。

(7) a. めっちゃ楽しかったです。金沢まで来てよかったよ。やっぱりトニセンは楽しいですね。オレキミに手の振りができてました。 (OY14_00565)
b. その後は、彼女らの大学の話や、元のクラスの子らの近況を聞いたりして盛り上がる。大学は大学で楽しんでるみたいやけど、やっぱ高校時代が懐かしい懐かしいって言いながら、昔話に花を咲かせ、、、 (OY08_00329)

次に、有題文と好んで共起する心情表現として「正直」のような主体の内面を表す表現が主に感情形容詞と評価性形容詞のタイプに見られる。そのうち、「さぞ」は話し手の経験上の事物の属性の認識ではなく、他者の経験の事物と関連することで有題文で見られる他の共起表現と性格が異なるとも考えられる。しかし、「さぞ」は他者の経験を対象に共感的に推量する性質を持ち、主体の率直な心情の表現である。また、「さぞ」は当該事態の程度が高い場合だけに使われる特徴を持ち、次節で述べる「少し」とは共起しない有題文の性質とも一致する (cf. 杉村 2003)。

(8) 客はそれを聞いて杉の枝を立てた普斎の小屋に入り、後入りのお茶を楽しんで帰ったということです。あえての演出ではなく、機に応じての計いとして、こうした機智に富んだ趣向はさぞ楽しいと思います。さて、『淡交』誌連載にあたっては、編集部の堤勇二氏と星原理恵氏に大変お世話になりました。 (LBp7_00014)

以上、形容詞による属性叙述文と共起する表現から特に有題文を好むものを分析した。分析した５つの項目の特徴は次の２点にまとめられる。第一に、事物の属性の認知は当該事態に限らず主体の以前の経験または想定の認識と関連する。第二に、事物の属性に関して主体の心情が膨らんでいるような認識が伴われる。このことは、4.1節で分析した有題文と無題文の共通の共起表現から明らかになった有題文に見られる主体の観点とも類似する。

４．３　無題文の共起表現

次に、無題文でより好まれる共起表現としては表４のように「限定」、「理由」などがある。

表４　「無題文」を好む共起表現

成分	表現	有題文	無題文
程度が低い	少し / 少々 / やや / ほんのり / ちょっぴり	0	22
限定性	最近 / 今の季節は / 場合には / 〜では など	4	24
条件	ば / と / たら / なら	0	21
理由	ため / ので / から / せいか / せいで	2	11
可変性	いつもではないのですが / なんとなく　など	0	6

まず、副詞に関して無題文と有題文を比較しながら分析する。程度副詞の場合、4.1節で述べたように程度が高い副詞に関しては無題文と有題文の両方とも同じ頻度で観察された。しかし、興味深いことに程度の低い副詞は無題文を好む。事物の属性の程度が低く限定さ

れていることは主体とその属性との関わりが低いか事物の一部として認めないことである。「少し」が用いられる例は、(9a)のように事物の属性を解決しようとする文脈が続く。時間副詞の場合は、有題文での「引き続き」とは逆に、無題文では持続する時間を限定する表現の頻度が高い。そして、「場合には」や「にとっては」など、状況が限定される表現の頻度が高い。(9b)では、主体は事物の恒常的な属性として判断せず限定された場合の属性として主体は認識している。

(9) a. 古い洗面台は撤去してとても綺麗に生まれ変わりました～♪神田佐久間町三丁目、○伍商会さ～ん♪ありがとうございました！工事を担当した若旦那は俳優の高島政宏にウリ似のイケ面で～す。洗面所が綺麗になったのに浴室が少々寂しい・・・。何か物足りなかったので沖縄で買って来た浴室用のポスターを貼って湯船に浸かり飲みながら沖縄旅行を思い出すのでありました。　(OY11_04724)
b. 頻回な下痢では個室管理が望ましい。　(PB24_00252)

次に 無題文で特徴的なのは、条件文の表現の頻度が高いということである。格助詞に関する研究によると条件文は格助詞「は」と共起しないため、当然の結果とも考えられる。しかし、その条件文と格助詞の関係には、従属節の条件文の中の格助詞の使い方が論じられている(野田 1996)。一方、表４の結果は、従属節の条件文をもつ主文の中の属性叙述文の形式が対象となる。そのため、従来の研究と性格が異なると言える。従属節の条件文のいくつかの例は以下のようである。

・感覚形容詞：今朝起きたら、冬になると、喋るために息をすると、触ってみると　など
・感情形容詞：話をすると
・属性形容詞：それによると、罪名別にみると
・評価性形容詞：なるべくだったら、ご心配なら、国土交通省からすれば　など

条件文との共起は主に感覚形容詞と評価性形容詞に見られる。そして、「それによると」のように事物の属性と関わる出所の例が多い。これは先の「については」や「にとっては」といった他の共起表現とも類似する成分である。もう一つ注目すべき点は、条件文の事柄が主体の事物の属性の認知を引き起こす点である。(10)の肺の痛みは吸気のときに初めて認知される属性である。

(10) ここ最近咳をする風邪にかかっています。今日朝起きると、息を吸うと、背中の右の肩甲骨の下の少し真ん中よりのところ一帯が痛くて息が吸いづらいです。肺がどうにかなったんでしょうか？　(OC09_03543)

条件文で見られる事物の属性の認知を招くような事柄は「ため」、「ので」などの目的や理由を表す表現とも類似する。ある目的、理由があって生じる事物の属性であり、その目的、理由ではないと同様の属性は発生しない。

最後に注目する相違点は、有題文では以前と同様の属性が持続することを表すものが多いのに対し、無題文では以前とは異なる事物の属性の認識が述べられているという点である。(11)では、施設内の死亡率が以前は高いと言えるほどではなかったのが、事態が変わ

ったことが認識されている。しかし、死亡率が高いという属性が事物の属性として定着していることではないこととして考えられる。

(11) 二月に入ってから、三階の一般病棟で死者が増えた。療養病棟の患者の病状が悪化すると、一般病棟に移されるのだ。いつもより施設内の死亡率が高いと病院内で噂になった。 (LBs9_00119)

以上、形容詞による属性叙述文で特に無題文と共起する表現について分析を行った。共起表現にはいくつの種類があるが、その成分には次のような類似点が見られる。一つ目は、時間と状況を区切る限定の表現が多いことから、無題文では主体が当該事態を前の状況や後の状況と関連させず、事物の属性を認知するその時点に注目を向けている。二つ目は、無題文では文脈に事物の属性を引き起こす外的の要因が存在し、事物の属性に対する主体の認識に外的思考が含まれている。これは、4.1節の(4b)で見られた主体の観点とも一致する。三つ目は、4.2節の(7b)と同様に当該事態が通常の状況とは異なるという主体の認識が読み取れる。最後は、程度の低い副詞の共起頻度が高いことから、主体が認知する属性を事物の完全なる性質としては捉えない観点を取る際には無題文を用いると考えられる。

５．おわりに

本研究では、形容詞による属性叙述文における有題文と無題文の性質を事物の属性に対する主体の観点に焦点をあて分析した。大規模日本語コーパスを用いて属性叙述文とその共起表現の調査を行った結果、評価性形容詞以外は有題文と無題文の形式に均等に分布されており、両形式の性質の相違は形容詞の意味とは相関しないと考えられる。

属性叙述文の共起表現の意味分析で最も興味深い相違点は、有題文では「決して」のように事物の属性に対する主体のゆるぎない判断が見られるが、無題文では条件文のように主体の判断に外部の状況が関与し、事物の属性の認識に主体の判断より外的（周辺）思考に焦点が当てられている。これは、形容詞の属性叙述文の形式の選択に事物の属性に対する「主体の安定感の程度」が関係するを示唆する。

参考文献

ゲオルギエバ，ベロニカ・トドロバ(2008)『「が」と「は」の使いわけとその理由：母語話者と非母語話者の実態調査を比較して』早稲田大学修士論文.

杉村泰(2003)「日本語の副詞サゾの意味分析−「共感」と「程度性」−」『名古屋大学言語文化論集』25:1 pp.67-81.

仲本康一郎(2006)「属性の意味論と活動の文脈：椅子が荷物になるとき」『日本語・日本文化』32 pp.39-61.

仁田義雄(2002)『副詞的表現の諸相』くろしお出版.

野田尚史(1996)『はとが』くろしお出版.

深田智、仲本康一郎(2008) 『概念化と意味の世界』研究社.

細川英雄(1993)「形容詞の主観性について−対象内容による形容詞の分類とその位置づけ−」『早稲田日本語研究』pp.78-65.

益岡隆志(2000)『日本語の文法の諸相』くろしお出版.

益岡隆志(2008)『叙述類型論』くろしお出版.

『今昔物語集』の捨て仮名

冨士池優美　（国立国語研究所　コーパス開発センター）†
岩崎瑠莉恵　（国立国語研究所　コーパス開発センター）

A Subsidiary Kana Character in *Konjaku-Monogatarishū*

Yumi Fujiike　(Center for Corpus Development, NINJAL)
Rurie Iwasaki　(Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所「通時コーパスの設計」プロジェクトでは、古典テキストに対して形態素解析を施す研究を進めている。その中で発表者らは『今昔物語集』コーパスの形態論情報付きコーパスの構築に向け、新編日本古典文学全集『今昔物語集』1～4（小学館、以下「新編全集」とする）に基づくテキスト整形を行った。

『今昔物語集』には、他の読み方をされる可能性のある漢字の読みの一部を補うことで漢字の読みを明確にする捨て仮名が多用されている。例えば「汝ヂ」「此カク」のようなもので、形態素解析にあたり、これが形態素の重複となり解析精度低下の原因となるため、捨て仮名部分を形態素解析対象から除外し、捨て仮名として除外された文字を XML タグに記録した。捨て仮名のテキスト整形にあたっては新編全集の頭注を参考にしたが、実際に形態素解析結果の人手修正をはじめたところ、頭注に指摘はないが似たような表記（例「薬リ」）が散見された。本発表では、現存最古の写本である鈴鹿本を底本とした巻 12 について、『今昔物語集』の捨て仮名の実際とその機能について考察する。また、他の和漢混淆文資料を対象にした捨て仮名の処理についても述べた上で、歴史的な資料を扱うコーパス構築にあたり、捨て仮名をどのように処理するとよいか検討する。

２．問題の所在

和漢混淆文に対して形態素解析を施すにあたり、漢文の要素が交じっていることに起因する問題がある。その一つが形態素の重複であり、形態素の重複があると文字と形態素との対応を上から順に取れないことが問題となる。捨て仮名はこの形態素の重複に当たる。『今昔物語集』コーパスの形態論情報付きコーパスの構築に向け、形態素の重複を解消すべく、前もって整備する必要があると考え、テキスト整形を行った[1]。

『今昔物語集』の捨て仮名には、「汝チ」の「チ」のように最後の 1 字が主に捨て仮名となるもののほか、「候フウ」のように「さぶらう」の「ぶ」の部分を表記したと思われるものもある。そこで、テキスト整形の際、頭注の記述を参考に捨て仮名を洗い出し、XML タ

† yfujiike@ninjal.ac.jp

1 『今昔物語集』全体についてのテキスト整形については、冨士池ほか（2013）に示した。

グ付けにより、捨て仮名部分を形態素解析対象から除外した[2]。XML タグには、捨て仮名として除外された文字列であることのほか、元の表記を記録した。処理対象となった捨て仮名は 1312 箇所、表記にして 240 種類超であった。以下に捨て仮名の処理例を示す。

汝**チ** → 汝	此**カク** → 此	候**フ**ウ → 候ウ
努々**メ** → 努々	今夜**ヒ** → 今夜	（**太文字**：捨て仮名）

このように『今昔物語集』コーパス構築の前処理としてのテキスト整形においては、頭注の記述を参考に捨て仮名かどうかの認定を行った。だが、コーパス構築にあたり、「捨て仮名」であるかどうか参考にできる情報があるとは限らない。むしろないことの方が多いだろう。そうなると、情報がないがテキスト整形は行いたいとなった場合、どのような表記を捨て仮名とするのかということがまず問題となる。

そこで「捨て仮名」とは何なのか、確認したい。『日本語学研究事典』では「捨て仮名」の項目はなく、「送り仮名」の項目に「副仮名・捨て仮名と称したこともある」とある。また『日本国語大辞典（Japan Knowledge）』には

> (1)漢文を訓読する時、漢字の意味を日本語にいい替えるために、助詞、助動詞、活用語尾の類を漢字の下に小さく添えて書くかなのこと。送りがな。
> (2)促音・拗音などを表わすのに用いる小さな字。「つ」「や」「ゆ」「よ」「イ」「オ」の類。
> (3)他の読み方をされるおそれのある漢字の読みの最後の一字をかたかなで、その漢字の下に小さく添えたもの。これによって、その場合の読み方をはっきりさせる。「様ン（さん）」「二人リ（ふたり）」の類。

とある。捨て仮名は送り仮名の別称であるとともに、狭義では「他の読み方をされるおそれのある漢字の読みをはっきりさせる」ために「小さく添えた」仮名ということができるだろう。ここで「最後の一字を」の部分を省略したのは、前述した『今昔物語集』中の「候フウ」「此カク」のように、添えられる仮名は必ずしも最後でもなければ、1 字とも限らないことによる。以下、本発表では、他の読み方をされるおそれのある漢字の読みをはっきりさせるために小さく添えた仮名を捨て仮名とする。

３.『今昔物語集』巻 12 の捨て仮名

３．１　調査の概要

「他の読み方をされるおそれのある漢字の読みをはっきりさせるために小さく添えた仮名」が捨て仮名であるとするならば、捨て仮名がある場合とない場合、複数の表記をもって、「読み」ないし「語」の書き分けを行っているはずである。

『今昔物語集』現存最古の写本である鈴鹿本を底本とした巻 12 について、形態素解析結果（人手修正済み）を用い、①頭注で指摘された捨て仮名の語の表記のバリエーション、

[2] 頭注の記述は一様ではないため、捨て仮名に関する頭注を洗い出し、誤写の可能性があるもの、衍字か捨て仮名かが明確でないもの等については、テキスト処理対象から除外した。

②頭注で指摘された捨て仮名の語に対して、捨て仮名がない表記でどのような語が表されているのか、③頭注で指摘された以外に仮名が添えられるかどうかで書き分けされる語があるのか、という 3 点を調査し、影印本で鈴鹿本での表記を確認した。なお、「漢字の読みをはっきりさせる」の読みが活用語尾を表す場合は、いわゆる「送り仮名」に該当すると考え、活用語は今回の調査の対象外とした。

３．２　頭注で指摘された捨て仮名の語の表記バリエーション

表 1　頭注で指摘された捨て仮名の語の表記バリエーション

	読み	種類	全訓	品詞	捨て仮名	漢字のみ	総計
1	ナンジ	汝チ		代名詞	20	24	44
2	ムカシ	昔シ		普通名詞	5	49	54
3	マタ	亦タ		副詞	4	93	97
4	サトリ	智リ		普通名詞	3	3	6
5	ヨル	夜ル		普通名詞	2	4	6
6	タテマツル	奉ツ		動詞	2	130	132
7	タビ	度ビ		普通名詞/助数詞	2	7	9
8	ヒトエニ	偏ヘ		副詞	2	13	15
9	カタチ	形チ		普通名詞	1	12	13
10	シルシ	験シ		普通名詞	1	13	14
11	モノ	者ノ		普通名詞	1	31	32
12	ノチ	後チ		普通名詞	1	104	105
13	ヒル	昼ル		普通名詞	1	4	5
14	タメ	為メ		普通名詞	1	53	54
15	アイダ	間ダ		普通名詞	1	120	121
16	ソモソモ	抑モ		普通名詞	1	1	2
17	ヒト	人ト		普通名詞	1	233	234
18	タマタマ	適マ		副詞	1	0	1
19	ユメ	夢メ		普通名詞	1	41	42
20	キミ・ギミ	君ミ		接尾辞	1	6	7
21	シルシ	注シ		普通名詞	1	0	1
22	トモ・ドモ	共モ		普通名詞	1	7	8
23	ヒジリ	聖リ		普通名詞	1	0	1
24	カズ	員ズ		普通名詞	1	6	7
25	ココロ	心ロ		普通名詞	1	81	82
26	アラズ	非ラ		動詞=助動詞	1	37	38
27	ツユ	露ユ		副詞	1	8	9
28	ヨワイ	齢ヒ		普通名詞	1	2	3
29	イウ	云イ	○	動詞	1	416	417
30	ナガイ	永ガ		形容詞	1	15	16
31	イワオ	巌ヲ		普通名詞	1	2	3
32	ナル	成ナル	○	動詞	1	51	52
33	ホラ	宝ウ螺		普通名詞	1	0	1
	総計				65	2838	2903

表 1 に巻 12 中の頭注で指摘された捨て仮名の語の表記バリエーションを示した[3]。「成ナル」のような全訓捨て仮名については、「全訓」列に○を付した。また、「汝」（「汝チ」の「チ」がない表記）のような捨て仮名のない表記については、「漢字のみ」列にその用例数を示した。巻 12 においては、頭注にて 33 種類の捨て仮名が指摘されており、延べ 65 箇所にタグ付け処理を行った。



この中で、「云イ」「成ナル」は全訓捨て仮名であり、活用語に捨て仮名を付すことにより読みをはっきりさせる目的があったものと考えられ

[3] 読み・種類欄は必要に応じて濁音表記とした。

る。その一方で、捨て仮名が添えられた表記が比較的多く出現している「ナンジ」「ムカシ」「マタ」のような語は、漢字の読みが複数あるわけではなく、仮名が付されていなくても他の読み方をされるおそれはない語である。このことから、『今昔物語集』の捨て仮名は、必ずしも、「漢字の読みをはっきりさせる」ものであるとは限らないということがわかる。

また、捨て仮名が添えられた表記と漢字のみの表記の出現頻度をそれぞれの語で比較してみると、「サトリ」のように捨て仮名が添えられた表記とない表記とで同頻度のものもあるが、「形チ」「験シ」等のように、ほとんどの捨て仮名表記が頻度 1 の捨て仮名である。つまり、臨時的に付されたことが多い様子がうかがえる。このようなことから、『今昔物語集』の捨て仮名は、決して体系的なものではないということが言える。「タマタマ」「ヒジリ」のような漢字のみの表記が出現していない語に関しては、「適マ」「聖リ」といった表記が当時の慣習的な表記であった可能性があると言えるだろう。

３．３　捨て仮名表記の語と捨て仮名がない表記の語

表 2　捨て仮名表記の語と捨て仮名がない表記の語

	読み	仮名あり			仮名なし表記														
		用例	巻12	全体	用例(読み)	数	用例(読み)	数	用例(読み)	数	用例(読み)	数	用例(読み)	数	用例(読み)	数	用例(読み)	数	
1	ナンジ	汝ヂ	20	261															
2	ムカシ	昔シ	5	33															
3	マタ	亦タ	4	5															
4	サトリ	智リ	3	45	智(シル)	1													
5	ヨル	夜ル	2	56	夜(ヨ)	314	夜(ヤ)内2例接尾辞	20											
6	タテマツル	奉ツ	2	19	奉(ウケタマワル)	11													
7	タビ	度ビ	2	6	度(ド)	78													
8	ヒトエニ	偏ヘ	2	5															
9	カタチ	形チ	1	119															
10	シルシ	験シ	1	26	験(ゲン)	10	験(ゲンズル)	6											
11	モノ	者ノ	1	22	者(シャ・ジャ)	115	者(ハ)係助詞	3											
12	ノチ	後チ	1	16	後(ウシロ)	80	後(シリエ:後方)	38	後(シリ)	7	後(アト)	3	後(ゴ)	2	後(オクレル)	2	後(カ:彼)	1	
13	ヒル	昼ル	1	16															
14	タメ	為メ	1	13	為(スル)	839	為(ナス)	9	為(ヰ)お経の一部	4	為(イル)誤写か	1							
15	アイダ	間ダ	1	12	間(マ)	36	間(ケン・ゲン)	30	間(ヒマ)	1	間(ヒラク)	1							
16	ソモソモ	抑モ	1	12	抑(オサエル)	6													
17	ヒト	人ト	1	9	人(ニン)	247	人(タリ)	84											
18	タマタマ	適マ	1	9	適(カナウ)人名	1													
19	ユメ	夢メ	1	9															
20	キミ・ギミ	君ミ	1	7	君|達(キン|ダチ)	41													
21	シルシ	注シ	1	6	注(シルス)	4													
22	トモ・ドモ	共モ	1	5															
23	ヒジリ	聖リ	1	4	聖(ショウ)	1													
24	カズ	員ズ	1	3															
25	ココロ	心ロ	1	3	心(シン・ジン)	20													
26	アラズ	非ラ	1	3	非(ナガ)	1													
27	ツユ	露ユ	1	2															
28	ヨワイ	齢ヒ	1	2															
29	イウ	云イ	1	1															
30	ナガイ	永ガ	1	1															
31	イワオ	巌ヲ	1	1	巌(イワ)	4													
32	ナル	成ナル	1	1	成(ナス)	63													
33	ホラ	宝ウ螺	1	1	宝(タカラ)	8													
	総計		65	733															

表 2 に巻 12 中の捨て仮名表記の語と捨て仮名がない表記の語を示した。「仮名なし表記」については、『今昔物語集』全体に範囲を広げて、捨て仮名がない表記でどのような語が表されているのか調査した。この調査の結果、捨て仮名によって語の書き分けがなされ、捨て仮名が漢字の読みを特定させる機能を持つ語があることが確認された。

例えば「夜」は、新編全集『今昔物語集』中で、「ヨル」のほかに「ヤ」「ヨ」といったルビが付されている。以下の（1）は「ヤ」、(2)「ヨ」、(3)（4）は「ヨル」と読む例である。新編全集のルビによると（1）「一夜」は「ヤ」、(2)「夜モ」(3)「夜深更」は「ヨ」と

読む。(3)（4）の「夜」には、「夜ル」と仮名が添えられており、「ヨル」と読むことがわかる。これらの例は、「夜」に続く語が漢字で記されており、語と語の切れ目が不明瞭であるが、(3)（4）の例のように捨て仮名が用いられている場合は、境界が明確になり、読みを特定することが可能になる。

(1) 自ラ一ノ樹ノ下ニ留テ一夜ヲ過シツ。
（巻第十七「依地蔵示従鎮西移愛宕護僧語第十四」）
(2) 物語ナドシテ夜モ漸ク深更レバ入ヌ。
（巻第十七「比叡山僧依虚空蔵助得智語第三十三」）
(3) 夜ル薪ヲ拾テ、火ヲ打テ木ヲ焼ク間、夜深更テ、…
（巻第十七「籠鞍馬寺遁羅刹鬼難僧語第四十三」）
(4) 天皇夜ル清涼殿ノ夜ルノ大臣ニ御マシケルニ、…
（巻第二十九「九条堀河住女殺夫哭語第十四」）

同様に、「アト」「ウシロ」「ノチ」のように複数の漢字の読みがある「後」についても、捨て仮名が用いられていることによって、漢字の読みが明確になっていることが確認できる。以下の（5）は「ウシロ」、(6）は「アト」、(7）(8）は「ノチ」と読む例である。新編全集のルビによると（5）「後ノ」は「ウシロ」、(6）の「後ニ」は「アト」と読む。(7）(8）は「後」に続く語が漢字で記されているが、「後チ」と仮名が添えられており、この捨て仮名によって「ノチ」と読むことが明確になっている。

(5) 牛、童ニ違テ堂ノ後ノ方ニ下リ来レリ。
（巻第二十四「関寺駈牛化迦葉仏語第二十四」）
(6) 盗人ノ出ヌル後ニ、門ニ走リ出デヽ、音ヲ挙テ、…
（巻第二十九「明法博士善澄被殺強盗語第二十」）
(7) 其ノ後チ本国ニ返テ此ノ事ヲ語ル。
（巻第十七「備中国僧阿清依地蔵助得活語第十八」）
(8) 此レヨリ後チ仏法ニ随テ、逆罪ヲ造ル事無カレ。
（巻第十二「越後国神融聖人縛雷起塔語第一」）

このように「夜ル」「後チ」の用例は、捨て仮名の有無と後続の文字種には関連があることを示唆していると言えるだろう。その一方で、(3）で「夜深更テ」とあるように、漢字に続く場合であれば必ず捨て仮名が添えられるというわけではない。ここから『今昔物語集』では漢字の読みをはっきりさせることを目的に臨時的に捨て仮名を付すことがあったと考えられる。

３．４　頭注で指摘された以外の漢字のみ表記と漢字仮名交じりの双方がある語

表 3 に頭注で指摘された以外の漢字のみ表記と漢字仮名交じりの双方がある語を示した。普通名詞には「営／営ミ」のような動詞連用形転成名詞が多く見られる。これらは活用語尾を明記したものと同様と考えられ、一種の送り仮名と考えていいだろう。送り仮名と考

えられるものを除くと、「哀／哀レ」「明カ／明ラカ」「愚／愚カ」「専／専ラ」「薬／薬リ[4]」等がある。漢字表記と漢字に仮名が添えられた表記があるわけだが、漢字の読みが複数ある語とは言えず、「漢字の読みをはっきりさせる」という点から考えると、捨て仮名とは言いがたく、頭注で捨て仮名と指摘されなかった一因と考えられる。その一方で、形態素解析という観点から考えると、例えば「明カ」「専」等は形態素の一部の不足、「薬リ」は形態素の重複と捉えられ、形態素解析辞書を対応させる必要がある[5]表記と言える。

表 3　頭注で指摘された以外の漢字のみ表記と漢字仮名交じりの双方がある語[6]

	語彙素読み	語彙素	書字形	品詞	頻度	
1	アワレ	哀レ	哀	普通名詞/形状詞可能	5	
			哀レ			5
2	イトナミ	営ミ	営	普通名詞	3	
			営ミ			1
3	オモイ	思イ	思	普通名詞	2	
			思ヒ			2
			念ヒ			1
4	チカイ	誓イ	誓	普通名詞	4	
			誓ヒ			3
5	クルシミ	苦シミ	苦	普通名詞	2	
			苦ビ			2
6	サトリ	悟リ	智	普通名詞	6	
			悟リ			3
7	チギリ	契リ	契	普通名詞	4	
			契リ			2
8	ナガレ	流レ	流	普通名詞	1	
			流レ			1
9	ネガイ	願イ	願	普通名詞	10	
			願ヒ			1
10	ハジメ	初メ	初	普通名詞	3	
			初メ			6
		始メ	始メ			2
11	メグリ	巡リ	迺	普通名詞	4	
			廻リ			1
12	クスリ	薬	薬	普通名詞	2	
			薬リ			1
13	ケムリ	煙	煙	普通名詞	2	
			煙リ			1
14	ヒカリ	光	光	普通名詞	5	
			光リ			3
15	アキラカ	明ラカ	明カ	形状詞	1	
			明ラカ			4
16	オオキ	大キ	大	形状詞	17	
			大キ			14
17	オロカ	愚カ	愚	形状詞	5	
			愚カ			1
18	タイラカ	平ラカ	平カ	形状詞		1
			平ラカ			2
19	イカデ	如何デ	何	副詞	1	
			何デ			3
20	イカニ	如何ニ	何	副詞	6	
			何ニ			9
21	シカ	然	然	副詞	1	
			然カ			2
22	シバラク	暫ク	暫	副詞	2	
			暫ク			10
23	スコシ	少シ	少	副詞	5	
			少シ			8
24	モシ	若シ	若	副詞	3	
			若シ			15
25	モッパラ	専ラ	専	副詞	9	
			専ラ			2
	総計				103	106

３．５.『今昔物語集』巻 12 における捨て仮名の機能

3.2 節から 3.4 節まで、『今昔物語集』巻 12 を対象とした捨て仮名に関する調査結果を見

4 「クスリ」に対して「クスル」という動詞もあるが、『日本国語大辞典（Japan Knowledge）』の用例は和訓栞であり、『今昔物語集』の頃に名詞からの転成動詞はなかったと考える。

5 実際には「薬リ」は『今昔物語集』の形態素解析結果人手修正時に辞書に追加することになったが、「明カ」「専」は先行して進められていた近代語コーパス整備時に辞書に追加されていた。

6 構築中のコーパスは漢字片仮名交じり文を漢字平仮名交じり文に変換しており、UniDic 中の語彙素・書字形も漢字平仮名交じり表記となっている。

てきた。『今昔物語集』巻 12 の捨て仮名の機能に関して、以下の 4 点を指摘した。

①全訓捨て仮名は活用語に見られ、「漢字の読みをはっきりさせる」ものであったと考えられる。

②『今昔物語集』の捨て仮名は必ずしも「漢字の読みをはっきりさせる」ものではない。臨時的に付されたもの、当時の慣習的な表記と考えられるものが多く存在する。

③複数の読みを持つ漢字と漢字が続く場合に捨て仮名が付されることがあり、これは「読みをはっきりさせる」目的があった可能性がある。

④新編全集の頭注で捨て仮名との指摘がなかった表記は、送り仮名に相当するものか、「読みをはっきりさせる」機能がないものである。

『今昔物語集』巻 12 の捨て仮名を見ると、①や③で示したような「漢字の読みをはっきりさせる」機能を持つものがある一方で、②で示したように、多くは、読みをはっきりさせるという読み弁別機能は見出されず、慣習的な表記と見た方がいいものであった。また、②で示したような、後続の文字種によって捨て仮名が付される例から、捨て仮名は読み弁別機能だけではなく、語の弁別機能を持っていたと考えられる。

4.『法華百座聞書抄』における捨て仮名

現在、『法華百座聞書抄』のコーパス化を進めているが、この資料にも「佛ケ」「前キ」といった捨て仮名ないし捨て仮名のような表記が散見される。形態素解析においては、形態素末尾の重複に当たるため、前処理をすることが望ましいが、捨て仮名であることを示す情報がない。そこで、総索引の情報を参考に捨て仮名らしき表記を作業者が独自に認定し、タグ付けを行った。その結果、捨て仮名らしき表記を持つ語は 56 種類が捨て仮名タグ付けの対象となった。その後、形態素解析結果の人手修正を経て、捨て仮名らしき表記を持つ語は 62 種類となった。

表 4　『法華百座聞書抄』中の捨て仮名

	読み	全訓	種類	数	今昔あり
1	ムカシ		昔シ	6	○
2	イノチ		命チ	6	
3	トキ		時キ	5	○
4	ナンジ		汝チ	4	○
5	トシゴロ		年シコロ／年シ来ロ	4	
6	タビ		度ヒ	4	○
7	ミル	○	見ミ	3	
8	ホトケ		佛ケ	3	○
9	ノチ		後チ	3	○
10	モロモロ		諸ロ	3	
11	アイダ		間ダ	3	○
12	ニョシキ	○	女職シキ	2	
13	トク	○	説ト	2	
14	ロウ	○	樓ロウ	1	
15	ラク	○	樂ラク	1	
16	カンバツ	○	旱魃ハツ	1	
17	チュウトウ	○	偸チウ盗タウ	1	
18	ラン	○	亂ラン	1	
19	ヨウモウ	○	楊ヤウマウ	1	
20	ワレ	○	我ワレ	1	
21	リヤク		利益ク	1	
22	ヨルヒル		夜ルヒル	1	
23	ヨル		夜ル	1	○
24	ヨ	○	世ヨ	1	
25	ヨシ		由シ	1	
26	ヨウゴ		擁護コ	1	
27	ミ|ナ	○	御ミ|名	1	
28	ミチシルベ		道チシルベ	1	
29	ミチ		道チ	1	
30	サキ		前キ	1	○
31	フチュウ	○	不注チウ	1	
32	ハヤシ		林シ	1	
33	ナニゴト	○	何ナニコト	1	
34	トモ	○	共トモ	1	
35	ノリ	○	説ノリ	1	
36	トシ		年シ	1	○
37	トコロ		所ロ	1	○
38	メ	○	妻メ	1	
39	チョク	○	勅チヨク	1	
40	フボ	○	父フ母ホ	1	
41	チチハハ		父チ母ハ	1	
42	タマウ	○	給タマ	1	
43	タグイ		類クヒ	1	○
44	ハベル	○	侍ハムヘリ	1	○
45	コト	○	殊コト	1	
46	コ	○	子コ	1	
47	シシ	○	完シヽ	1	
48	スガタ	○	躰スカタ	1	
49	カミ		髪ミ	1	
50	カゼ		風ゼ	1	
51	カズ		數ス	1	
52	オトコ	○	男ヲトコ	1	○
53	ナンガン	○	男ナム官クワン	1	
54	オケ	○	桶ヲケ	1	
55	イマ		今マ	1	○
56	ヒト|タビ		一ト|度ビ	1	
57	ワレ		我レ	7	
58	タトエ		喩ヒ	2	
59	イヨイヨ		彌ヨ	1	
60	ウエ		上へ	1	○
61	オノオノ		各ノ	1	○
62	カタチ		形チ	1	○
	総計			104	

表 4 に『法華百座聞書抄』中の捨て仮名らしき表記ないし捨て仮名のような表記を持つ語を示す。このうち、『今昔物語集』で捨て仮名のタグ付け対象となった表記は、表 4 の「今昔あり」に「○」としたもので、18 種類あった。『今昔物語集』は 58 万短単位超であるのに対して『法華百座聞書抄』は約 2 万短単位（ともに延べ語数）と資料規模は 30 分の 1 程度と格段に小さいものの、捨て仮名タグ付けのカバー率としては約 31%であり、3 分の 2 は『法華百座聞書抄』にしかない捨て仮名であった。ここから、資料独自の表記がある様子がうかがわれる。

『法華百座聞書抄』には全訓捨て仮名が比較的多く見られる。さらに捨て仮名表記された語が片仮名のみで表記されることもあった。読みをはっきりさせるという目的からすると、片仮名表記が最も明確であり、次いで全訓捨て仮名ということになる。全訓捨て仮名が付された語には漢語が多い。また、「利益ク」「擁護コ」といった漢語の読みの末尾を表す仮名も見られた。これらは読みをはっきりさせる機能を持つものと考えられる[7]。そのほか、「父フ母モ」（ブモ）と「父チ母ハ」（チチハハ）、「説ト」（動詞：説く）と「説ノリ」（名詞）のような表記から、同じ漢字で表記される語それぞれに対して捨て仮名を添えることによって語の書き分けをしている様子が確認された。

『法華百座聞書抄』と『今昔物語集』はほぼ同時代の資料であり、内容的にも仏教という共通点があるが、捨て仮名の使用実態をみるかぎり、表記に大きな違いがあることが明らかになった。『法華百座聞書抄』の捨て仮名には、当時の慣習的な表記のほかに、語の弁別機能を持つものがある。また、『法華百座聞書抄』は『今昔物語集』巻 12 と比較して、漢字の読みの明示を期する資料であると言えるだろう。

５．おわりに

捨て仮名（ないし捨て仮名的な表記）は平安時代の和漢混淆文だけではなく、その後、長く続く。捨て仮名の問題は、近代語まで含めた歴史的な資料を扱うコーパスの問題と言える。ここで、形態論情報付きコーパスの構築にあたり、捨て仮名（ないし捨て仮名的な表記）をどのように処理していくとよいかという問題を考えてみたい。

形態素解析という点から捨て仮名を考えたとき、「漢字の読みをはっきりさせる」という観点はあまり関係なく、形態素の一部の過不足が問題となる。そのため、前処理で可能な限り問題を取り除くことが望ましい。その一方で、形態素解析の対象となるテキスト中に、形態素の過不足がある語が存在するかどうか、またどのような表記なのか、事前に把握することは難しい。例えば『今昔物語集』で得られた捨て仮名リストがあっても、『今昔物語集』（本朝部）の 30 分の 1 程度の規模である『法華百座聞書抄』の捨て仮名は 3 分の 1 弱しかカバーされないことが明らかになった。これは、捨て仮名（ないし捨て仮名的な表記）のありようは資料によって異なり、その結果として前処理のみで網羅的に処理することは

[7] 読みの冒頭ないし末尾のみを仮名で付したものは機能としては振り仮名に近いと考えられる。しかし『法華百座聞書抄総索引』において、例えば「りやくす（利益す）」項に「利益〻セシメ給へ」「利益シタマヒケレ」が示されていることから、振り仮名（ルビ）としては処理しないこととした。

困難であることを意味している。形態素の重複を解決するという目的を考えると、捨て仮名に関するテキスト整形のタイミングは前処理となるが、これを効率的に行うためには、形態素の重複となった表記のリストを作成し、新しい資料を処理するたびにそのリストを拡充していくといった方策が考えられる。また、形態素解析結果の人手修正を通し、読みを確定させていく中で、前処理で形態素重複を解決した結果に過不足はないか確認する必要がある。

形態素の重複を解決するための前処理のほかに、コーパスに付与する情報として、機能まで勘案した上で捨て仮名であるという情報を付与することも考えられる。『今昔物語集』と『法華百座聞書抄』の処理結果から捨て仮名の機能は資料によっても異なる様子が明らかになった。つまり、同じ表記であっても捨て仮名であるかどうかは資料により異なることになる。そのため、機能まで勘案した捨て仮名情報を付与するには、対象資料ごとに付された形態論情報を利用し、「漢字の読みをはっきりさせる」機能の有無を確認するといった作業が必要になるだろう。

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」（プロジェクトリーダー：近藤泰弘）、及び、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、研究代表者：田中牧郎）の成果の一部である。

文　献

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」情報処理学会研究報告 人文科学とコンピュータ vol.2010-CH85、No.4

栗原さよ子・安部清哉（2009）「平安期の「袪」の訓タモト―『今昔物語集』の捨て仮名付きの「袪ト」をめぐって―」学習院大学大学院日本語日本文学 5、pp.34-45

小林芳則（1976）『校註法華百座聞書抄』（武蔵野書院）

小林芳則（1975）『法華百座聞書抄総索引』（武蔵野書院）

酒井憲二（1968）「表記史上の一問題　―捨てがなの場合―」語文 31、pp.35-45

佐藤麻衣子（2010）「享保期浄瑠璃本における捨て仮名 ：『出世握虎稚物語』の調査から」日本女子大学大学院文学研究科紀要 16、pp.59-71

飛田良文・遠藤好英・加藤正信・佐藤武義・蜂谷清人・前田富祺（2007）『日本語学研究事典』（明治書院）

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

馬淵和夫・国東文麿・稲垣泰一（1999-2002）『新編日本古典文学全集 今昔物語集』1〜4（小学館）
安田章（1997）『鈴鹿本　今昔物語集　影印と考証』（京都大学学術出版会）
山岸徳平（1976）『法華修法一百座聞書抄』勉誠社文庫 4（勉誠社）

動詞「きく」のコロケーションについて
-WEB コーパスと日本語母語話者・上級日本語学習者の主観的判断から-

小野正樹(筑波大学 人文社会系) †
朱炫姝††・デヒピティヤ スランジ ディルーシャ§・李国玲§§・
スワンナクート パッチャラーパン#(筑波大学大学院 人文社会科学研究科)

On the Collocation of the Japanese Verb " KIKU": Comparative Research Based on a Corpus Survey and Subjective Judgments by Native Speakers and Advanced Learners

Masaki Ono (University of Tsukuba, Faculty of Humanities and Social Sciences) †
Hyunju Ju††・Surangi Dilusha, Dehipitiya§・Guoling Li§§・
Patcharaphan Suvanakoot#(University of Tsukuba, Graduate School of Humanities and Social Sciences)

１．動詞「きく」について

本研究は動詞「きく（聞・聴・効・利）」の基本義（小野他 2009）をもとに、現実のデータ及び主観的な判断のコロケーション頻度に関する調査を行った。「聞く」「きく」「効く」「利く」を同一に扱う理由は、音韻的に同じ語彙について、日本語学習者がどれほど違いを意識しているかという発想に基づく。例として、「口を利く」という表現を日本語学習者に尋ねたところ、「口を聞く」と回答したケースが非常に多く、その際に学習者は、漢字ではなく、音声中心で理解を進めているのではないかという発想に至ったからである。

(１)先生に生意気な口を聞いて、考えてみると恥ずかしかった。（中国人日本語学習者）

漢字圏の学習者でさえも文字の区別を感じることなく、音声言語として日本語を理解しているようである。

小野他(2009)では、「きく」について、①「耳で感じる」、②「相手の言うことを受け入れる」、③「質問する」、④「能力や機能を発揮する」、⑤「利用できる、可能である」、⑥「技能がすぐれている」の６つの基本義を設定した。動詞「きく」は「音、声、言葉などを耳に感じとる」ことが原義であるが、「その内容を知り」、「人の言葉に従う」ことへの意味拡張が起き、「答えを耳に入れようとして、人に尋ねる」という意味を獲得する。さらに、『広辞苑』第六版では、「聞く」の項目に「物事をためし調べる。」（「利く」とも書く）という記述や、「効く」と「聞く」は「同源」という記述もあり、意味が連続しているという立場に立つ。以下に６つの基本義の関連と例を示す。

① 耳で感じる：例「音楽を聞く」
② 相手の言うことを受け入れる：①を踏まえて、より積極的に理解すること。例「頼みを聞く」
③ 質問する：①が可能になることを相手に依頼する。例「答えを聞く」
④ 能力や機能を発揮する：例「気が利く」
⑤ 利用できる、可能である：例「融通が利く」
⑥ 技能がすぐれている：例「腕が利く」

† ono.masaki.ga@u.tsukuba.ac.jp
†† murasakiju@gmail.com
§ sd.dehipitiya@gmail.com
§§ rikokurei@gmail.com
nongtib@hotmail.com

２．調査について

『NINJAL-LWP for TWC』を利用し、初めにインターネット上の頻度上位 100 のコロケーションを示す。次に、頻度上位 100 のコロケーションを含め、日常的に有益だと思われるコロケーションを日本語母語話者と日本語上級学習者の調査協力者に、それぞれ抽出を依頼する。

2.1. 調査１　「聞く」「きく」「効く」「利く」の頻度の高いコロケーション

「きく（聞・効・利・きく）」の５つの表記語彙について、各名詞句のコロケーション数を示す。その場合に複数の基本義を持つこともあり、述べ語数として示す。

〈表１〉動詞「聞く」のコロケーション上位 100 に見られる意味（数値は述べ語数）

名詞＋助詞情報	件数	①耳で感じる	②相手の言うことを受け入れる	③質問する
総数	330,942			
「…が聞く」	18,860	97	67	76
「…を聞く」	211,825	97	15	46
「…に聞く」	70,570	94	3	88
「…で聞く」	29,687	99	1	78

本データから分かることは、基本義①については、いずれの名詞句においても基本義①を維持していることである。次に、「利く」と「効く」についてはかなりコロケーションが異なっている。

〈表２〉動詞「効く」のコロケーション上位 100 に見られる意味（数値は述べ語数）

名詞＋助詞情報	件数	④能力や機能を発揮する	⑤利用できる、可能である	⑥技能がすぐれている
総数	31,687			
「…が効く」	10,630	79	22	1
「…を効く」	1,096	87	11	0
「…に効く」	7,467	100	0	0
「…で効く」	1,544	100	3	0

〈表３〉動詞「利く」のコロケーション上位 100 に見られる意味（数値は述べ語数）

名詞＋助詞情報	件数	④能力や機能を発揮する	⑤利用できる、可能である	⑥技能がすぐれている
総数	10,674			
「…が利く」	4,978	56	46	4
「…を利く」	1,793	49	36	1
「…に利く」	1,013	68	48	4
「…で利く」	667	27	85	1

上記データにおいては、「利く」については名詞句により数値が異なっていることが特徴である。特に、⑤の意味ならば「…で利く」が多いことは固定した表現が多いようである。

〈表４〉動詞「きく」のコロケーション上位 100 に見られる意味（数値は述べ語数）

名詞＋助詞情報	件数	①　耳で感じる	②相手の言うことを受け入れる	③質問する	④能力や機能を発揮する	⑤利用できる、可能である	⑥技能がすぐれている
総数	14,702						
「…がきく」	4,980	13	0	9	42	45	0
「…をきく」	6,163	64	10	33	14	11	0
「…にきく」	2,458	34	1	58	22	25	0
「…できく」	1,101	0	0	0	0	0	0

「きく」については、音声的に敢えて曖昧となっていた名詞句関係が、「がきく」であれば①「耳で感じる」ことであるのに対し、「にきく」ならば基本義②「質問する」、「がきく」ならば④「能力や機能を発揮する」あるいは、⑤「利用できる、可能である」のように実は音声上で規定されていることを示していよう。

2.2. コーパス上位語に対する母語話者と日本語上級学習者の判断

2.2.1　調査概要

『NINJAL-LWP for TWC』で上位 50 位に見られるコロケーションが、母語話者と日本語学習者にどれほど重要かを問うデータを示す。母語話者は、茨城県内に住む学部生および大学院生４名で、上級日本語学習者はいずれも Ｎ１レベルである。以下のデータは、上位 25、50、75、100 位のスケールに、各インフォーマントが重要だと判断したコロケーションがどれほど含まれているかを示したのである。

2.2.2　「聞く」の場合

〈表５〉動詞「が聞く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	8	3	0	4	0	10	0	7
上位 50	12	1	0	2	0	8	0	9
上位 75	13	1	0	2	0	6	0	10
上位 100	13	0	0	2	0	3	0	10
100位以下	54	0	0	1	0	6	0	15

〈表６〉動詞「を聞く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	21	18	0	3	0	0	0	0
上位 50	21	8	0	10	0	0	0	0
上位 75	17	9	0	7	0	0	0	0
上位 100	18	7	0	2	0	0	0	0
100位以下	23	12	0	24	0	0	0	0

〈表７〉動詞「に聞く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	12	10	0	6	0	1	0	6
上位 50	15	4	0	4	0	7	0	3
上位 75	14	1	0	7	0	3	0	3
上位 100	10	1	0	3	0	4	0	0
100位以下	49	0	0	8	0	29	0	0

〈表８〉動詞「で聞く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	16	8	0	5	0	5	0	0
上位 50	13	7	0	10	0	5	0	0
上位 75	15	2	0	14	0	1	0	0
上位 100	14	1	0	11	0	5	0	0
100位以下	42	1	0	10	0	15	0	0

本調査から、例えば〈表５〉の１点が多いことから、日本語上級学習者には選択の基準がそれぞれ異なることが予想される。

2.2.3　「効く」の場合

〈表９〉動詞「が効く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	21	11	0	4	0	6	0	2
上位 50	16	1	0	2	0	6	0	8
上位 75	14	0	0	2	0	4	0	7
上位 100	14	0	0	1	0	4	0	4
100位以下	35	0	0	3	0	12	0	23

〈表 10〉動詞「を効く」のコロケーション頻度上位 100 と日本語母語話者と日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	0	0	16	0	7	1	0	7
上位 50	0	0	10	0	10	0	0	0
上位 75	0	0	5	0	10	0	0	0
上位 100	0	0	3	0	7	0	0	0
100位以下	1	0	13	0	18	0	0	6

〈表 11〉動詞「に効く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	24	6	0	8	0	5	0	5
上位 50	14	4	0	8	0	4	0	7
上位 75	10	1	0	1	0	6	0	1
上位 100	12	0	0	0	0	7	0	0
100位以下	40	1	0	9	0	27	0	0

〈表 12〉動詞「で効く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	5	0	12	5	3	6	0	6
上位 50	0	0	3	1	12	3	0	6
上位 75	0	0	5	0	9	0	0	6
上位 100	3	0	1	0	6	0	0	6
100位以下	3	0	24	0	14	0	0	16

〈表 12〉の「で効く」調査から、1 点が多いことから、日本語上級学習者によって、選択の基準が異なることが予想される。

2.2.4 「利く」の場合

〈表 13〉動詞「が利く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	22	4	2	9	0	8	0	0
上位 50	17	0	5	2	0	6	0	12
上位 75	8	0	5	1	0	2	0	4
上位 100	5	0	9	0	0	1	0	2
100位以下	15	0	12	0	0	2	0	14

〈表 14〉動詞「を利く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	4	1	13	0	4	3	0	2
上位 50	1	0	7	0	10	0	0	2
上位 75	2	0	1	0	10	1	0	4
上位 100	0	0	1	0	5	0	0	0
100位以下	0	0	3	0	39	1	0	2

〈表 15〉動詞「に利く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	8	1	11	1	0	5	0	8
上位 50	3	0	9	3	0	0	0	6
上位 75	1	0	9	0	0	1	0	6
上位 100	2	0	7	0	0	0	0	4
100位以下	4	0	46	0	0	1	0	7

〈表 16〉動詞「で利く」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	0	0	11	0	5	4	0	6
上位 50	0	0	1	0	9	2	0	5
上位 75	0	0	2	0	8	0	0	3
上位 100	0	0	2	0	9	0	0	2
100位以下	1	0	13	0	39	0	0	2

2.2.5 「きく」の場合

最後に、「きく」という表記で限定されていないものについて観察する。

〈表 17〉動詞「がきく」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	20	6	0	5	0	4	0	4
上位 50	19	3	0	1	0	5	0	7
上位 75	13	1	0	3	0	3	0	6
上位 100	17	0	0	1	0	2	0	5
100位以下	31	0	0	0	0	2	0	21

〈表 18〉動詞「をきく」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	17	9	0	4	0	1	0	7
上位 50	18	6	0	5	0	3	0	4
上位 75	19	3	0	5	0	5	0	5
上位 100	13	0	0	1	0	4	0	7
100位以下	33	0	0	0	0	14	0	17

〈表 19〉動詞「にきく」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	19	2	0	7	0	7	0	6
上位 50	19	3	0	5	0	7	0	5
上位 75	14	2	0	3	0	9	0	2
上位 100	16	0	0	3	0	6	0	5
100位以下	32	0	0	0	0	11	0	17

〈表 20〉動詞「できく」のコロケーション頻度上位 100 と日本語母語話者・日本語学習者の判断の一致数

判定 / 順位	4 点		3 点		2 点		1 点	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
上位 25	15	1	1	7	0	7	0	2
上位 50	3	0	5	2	0	4	0	5
上位 75	7	0	3	0	0	0	0	4
上位 100	3	0	7	0	0	0	0	1
100位以下	27	0	29	0	0	4	0	11

本データから、音に依存しない意味を有したものこそ、母語話者は非常に厳格な判断基準をもっていることがわかる。

３．頻度と意味の関係

コロケーションの有効性は、頻度と必ずしも比例するわけではない。しかしながら、第２言語に関しては、初級語彙ほど多義で有り、それが様々な意味用法を持つことは確かであろう。使用頻度の少ない語彙を洗練語 (sophisticated words) と言う（http://www.u.tsukuba.ac.jp/~ushiro.yuji.gn/Publishing/SLAglossary.htm）が、全体の傾向として、母語話者には固定的なコロケーションがあるが、学習者には限定された基本義の語彙にこそ、重要性を感じることや、学習者により選択が異なることがわかった。

また、ガ格こそ、基本義が限定されることがわかり、周辺格のニ格やデ格になると、基本義も異なって行くことが分かった。一方で、「聞く」ほど多義ではない「利く」にはその傾向が見えなかったことは、意味の限定性があることを示していよう。

謝　辞

本研究は、文部科学省科学研究費基盤研究(A)「汎用的日本語学習者辞書開発データベース構築とその基盤形成のための研究」（平成 23～26 年度、領域代表者：砂川有里子）による補助を得ています。

文　献

小野正樹・小林典子・長谷川守寿(2009)『コロケーションで増やす表現－ほんきの日本語－』くろしお出版.

小野正樹・朱ヒョンジュ・許允瑄・山下悠貴乃・新井優子・孫思琦・伊藤秀明・デヒピティヤ スランジ ディルーシャ・グルミラ アリモヴァ(2013)「５言語の“丁寧”の意味・用法と WEB 上のコロケーション分析 : 日本語・韓国語・中国語・シンハラ語・ドイツ語の分析」『筑波大学留学生センター日本語教育論集』28、pp.1-21、筑波大学留学生センター.

新村出 編(2008)『広辞苑』第６版 岩波書店.
堀恵子・李在稿・砂川有里子・今井新悟・江田すみれ(2012)「文法項目の主観判定による ６段階レベルづけとその応用」2012 年日本語教育国際研究大会（ICJLE2012）ポスターセッション、日本語教育学会.
李在鎬・砂川有里子(2012)「コーパスを活用した日本語語彙表の構築」2012 年日本語教育国際研究大会（ICJLE2012）パネルセッション 日本語教育につながるコーパス研究－現状と今後の展望－、日本語教育学会.

関連 URL

『筑波ウェブコーパス(NINJAL-LWP for TWC)』http://corpus.tsukuba.ac.jp/

『太陽コーパス』の口語会話部分に現れる動詞の待遇表現の計量的分析 —聞き手に対する尊敬待遇表現を中心に—

近藤 明日子（国立国語研究所コーパス開発センター）†

A Quantitative Analysis of Respectful Verbs to Addressees of Conversations in *Taiyo Corpus*

KONDO Asuko (National Institute for Japanese Language and Linguistics)

１．はじめに

近代の話し言葉における動詞の待遇表現形式は現代語以上に多様性があり、先行研究において形式間の用法の差異の解明が進められてきた[1]。形式差は多くの場合位相差に基づくものと考えられ、使用年代や話し手の性別、話し手の社会的階層、話し手と聞き手との関係といったものが考察の観点として取り上げられてきた。

このように複数の形式と複数の位相との関係性の考察に重点がおかれる待遇表現研究は、多数の変数の分析に長けたコーパス言語学的手法と親和性が高いと考えられる。本稿はその一つの試みとして、代表的な近代語のコーパスである『太陽コーパス』（国立国語研究所（編）2005）中の小説・戯曲に含まれる口語体の会話部分から、主要な動詞の尊敬待遇表現の形式を抽出し、テキスト分類に用いられるランダムフォレストを用いて、形式の選択における位相の重要度を分析する。

２．調査対象テキスト

調査に利用するコーパスは、1895（明治 28）年から 1928（昭和 3）年まで刊行された総合雑誌『太陽』をコーパス化した『太陽コーパス』である。『太陽コーパス』には、1895・1901・1909・1917・1925 年の 5 カ年分、各年 12 冊、計 60 冊の全文テキストと、XML タグによる記事・引用・文字・表記等に関する情報が収録されている（田中 2005）。このコーパスに対して、近代文語文を対象とする形態素解析辞書「近代文語 UniDic」（小木曽 2009）と旧仮名遣いの口語文を対象とする形態素解析辞書（小木曽 2012）を用いて形態素解析を行い、形態論情報を付与したデータが国立国語研究所の形態論情報データベース（小木曽・中村 2011）に格納されている。本稿ではこのデータベースの 2013 年 12 月時点のデータを用いた[2]。

調査対象はコーパス中の小説・戯曲に含まれる口語体の会話部分である。これは、コーパス中で最も話し言葉的性質の強い文章であり、当時の話し言葉での実態を明らかにするのに最適なテキストと考えられる。ただし、近代の日本を舞台とする作品に限り、時代物や翻訳作品等は除外した。対象の抽出方法は次のとおりである。まず XML 記事タグのジャンル属性を利用し、属性値の NDC 番号が「912」（日本文学／戯曲）または「913」（日本文学／小説・物語）の作品を抽出し、そのなかから手作業で近代を舞台とするものを選別した。そこから、XML 引用タグの種別属性値が「会話」、記事タグまたは引用タグの文体属性が「口語」の部分を抽出し調査対象テキストとした。

† kondo@ninjal.ac.jp

[1] 近代のある程度の年代にわたる複数の資料を対象に、複数の動詞の待遇表現形式について考察を行った先行研究として、山西（1972）・辻村（1974）・近藤（2005）等がある。また、命令表現に焦点をあてたものとして陳（2005a）（2005b）（2006a）（2006a）等の一連の研究がある。

[2] データは整備途中のものであり、今後データの変更に伴い、本稿に挙げた数値も変動する場合がある。

調査対象テキストの延べ語数（助詞・助動詞を含む）と口語会話を含む作品数[3]とその著者数（異なり）を刊行年別に示したものが表 1 である。

表 1　調査対象テキストの言語量

	1895	1901	1909	1917	1925	通年
延べ語数	41704	53124	96412	76562	31678	299480
作品数	15	13	21	22	22	93
著者数	14	8	17	17	21	67

３．動詞の尊敬待遇表現形式の抽出

調査対象とする動詞の尊敬待遇表現形式の抽出手順は次のとおりである。まず、先行研究や辻村（編）(1991）にあげられた諸形式についてコーパスでの出現頻度の予備調査を行い、出現頻度の多い 13 形式を対象に選択した。ただし、尊敬の助動詞「れる・られる」による形式については「れる・られる」の意味分類が容易でないため、本稿では調査対象外とした。コーパスからの抽出は形態論情報に拠った。各形式とその検索条件を示したものが表 2 である。形式中の「～」には動詞連用形あるいはサ変動詞語幹が入り[4]、「御」は「～」に入る語によって「お」または「ご」になることを意味する。

表 2　調査対象とする動詞の尊敬待遇表現形式

形式	検索条件	
御～あそばす	語彙素「遊ばす」	語彙素「御覧」／「御免」
いらっしゃる	語彙素「いらっしゃる」	
～ていらっしゃる		
おっしゃる	語彙素「仰る」	
御～くださる	語彙素「下さる」	
～てくださる		
御～だ	語彙素「御」＋（1短単位）＋語彙素「だ」／「の」／「です」	
～たまう	動詞連用形＋語彙素「給う」	
御～なさる	語彙素「為さる」	
～なさる		
なさる		
御～になる	語彙素「御」＋（1短単位）＋出現書字形「に」＋語彙素「成る」	
御～	語彙素「御」＋動詞連用形	

この検索条件にマッチしたレコードから手作業により調査対象である形式を選別した。各形式の粗頻度と 1 万語あたりの相対頻度（括弧内）を示したものが表 3・表 4 である。命令形[5]をとる場合とそれ以外では出現する形式に偏りがあるため、表 3 に命令形以外をとる場合、表 4 に命令形をとる場合を分けて示した。

[3]複数回にわたる連載作品はまとめて 1 作品とした。

[4] 「～ていらっしゃる」「～てくださる」「～たまう」の「～」には動詞連用形のみが入り、「御～」には動詞連用形と一字漢語サ変動詞語幹「覧」「免」のみが入る。

[5] 本稿でいう「命令形」とは、①該当形式の命令形（例「～てください」「御～なさい」）、②該当形式に助動詞「ます」の命令形が付くもの（例「～てくださいませ」「御～なさいませ」）、③「御～」を指す。

表 3 動詞の尊敬待遇表現形式の頻度（命令形以外）

	1895		1901		1909		1917		1925		通年	
御～なさる	48	(11.5)	57	(10.7)	65	(6.7)	47	(6.1)	9	(2.8)	226	(7.5)
御～だ	59	(14.1)	69	(13.0)	41	(4.3)	35	(4.6)	18	(5.7)	222	(7.4)
おっしゃる	24	(5.8)	38	(7.2)	31	(3.2)	46	(6.0)	20	(6.3)	159	(5.3)
～てくださる	29	(7.0)	33	(6.2)	32	(3.3)	31	(4.0)	18	(5.7)	143	(4.8)
御～になる	22	(5.3)	14	(2.6)	25	(2.6)	41	(5.4)	35	(11.0)	137	(4.6)
～なさる	18	(4.3)	23	(4.3)	30	(3.1)	22	(2.9)	4	(1.3)	97	(3.2)
～ていらっしゃる	6	(1.4)	12	(2.3)	11	(1.1)	36	(4.7)	4	(1.3)	69	(2.3)
なさる	7	(1.7)	4	(0.8)	22	(2.3)	15	(2.0)	6	(1.9)	54	(1.8)
いらっしゃる	2	(0.5)	2	(0.4)	20	(2.1)	20	(2.6)	6	(1.9)	50	(1.7)
御～あそばす	10	(2.4)	4	(0.8)	12	(1.2)	7	(0.9)	0	(0.0)	33	(1.1)
御～くださる	2	(0.5)	4	(0.8)	15	(1.6)	6	(0.8)	5	(1.6)	32	(1.1)
～たまう	7	(1.7)	6	(1.1)	6	(0.6)	1	(0.1)	0	(0.0)	20	(0.7)
計	234	(56.1)	266	(50.1)	310	(32.2)	307	(40.1)	125	(39.5)	1242	(41.5)

表 4 動詞の尊敬待遇表現形式の頻度（命令形）

	1895		1901		1909		1917		1925		通年	
～てくださる	33	(7.9)	21	(4.0)	82	(8.5)	53	(6.9)	20	(6.3)	209	(7.0)
御～なさる	30	(7.2)	27	(5.1)	62	(6.4)	33	(4.3)	24	(7.6)	176	(5.9)
～たまう	23	(5.5)	9	(1.7)	51	(5.3)	63	(8.2)	17	(5.4)	163	(5.4)
御～	31	(7.4)	37	(7.0)	32	(3.3)	21	(2.7)	18	(5.7)	139	(4.6)
～なさる	6	(1.4)	6	(1.1)	21	(2.2)	7	(0.9)	2	(0.6)	42	(1.4)
御～くださる	7	(1.7)	1	(0.2)	18	(1.9)	5	(0.7)	7	(2.2)	38	(1.3)
いらっしゃる	0	(0.0)	0	(0.0)	5	(0.5)	7	(0.9)	3	(0.9)	15	(0.5)
～ていらっしゃる	1	(0.2)	0	(0.0)	1	(0.1)	9	(1.2)	4	(1.3)	15	(0.5)
なさる	2	(0.5)	2	(0.4)	2	(0.2)	2	(0.3)	1	(0.3)	9	(0.3)
おっしゃる	3	(0.7)	0	(0.0)	3	(0.3)	1	(0.1)	2	(0.6)	9	(0.3)
御～あそばす	5	(1.2)	0	(0.0)	2	(0.2)	1	(0.1)	0	(0.0)	8	(0.3)
計	141	(33.8)	103	(19.4)	279	(28.9)	202	(26.4)	98	(30.9)	823	(27.5)

4．形式選択における位相の重要度の分析

表 3･表 4 にあげた形式のうち､命令形以外と命令形のそれぞれ頻度上位 6 形式について、形式選択における位相の重要度を分析する。

使用する分析手法はランダムフォレスト（random forest）である。ランダムフォレストはカテゴリー化されたデータを用いて分類のルールを作成する機械学習方法の一つで、テキストマイニングの分野でも用いられる（金 2009)。ランダムフォレストは分類ルール作成の過程で、副次的に分類に用いた説明変数の重要度を計算する。ここでは、形式を目的変数、位相を説明変数としてランダムフォレストを行い、算出される説明変数の重要度を用いて形式選択に対する位相の重要度を測ってみたい。

説明変数とする位相は次の①～④の 4 種類である。

①年

該当形式が出現する雑誌の刊行年である。コーパスの XML 雑誌タグの年属性値に拠る。「1895・1901・1909・1917・1925」の 5 種。

②話者性別

該当形式が出現する会話の話者の性別である。コーパスには話者性別に関する情報は付いていないため、作品内容から手作業により判別した。「男・女・不明」の 3 種に分類し、「男・女」の 2 種に該当するレコードを分析対象とした。

③話者階層

該当形式が出現する会話の話者の社会的階層である。コーパスには話者階層に関

する情報は付いていないため、作品内容から手作業により判別した。「知識層・非知識層・不明」の 3 種に分類し、そのうち「知識層・非知識層」の 2 種に該当するレコードを分析対象とした。

④口語文体

該当形式が出現する口語会話の文体である。コーパスに付与した形態論情報に拠り、「ございます体・ですます体・その他」の 3 種に分類した。「ございます体」は該当会話中に「ござい（り）ます」が 1 例以上ある場合、「ですます体」は「ござい（り）ます」がなく「です」または「ます」が 1 例以上ある場合、「その他」は「ござい（り）ます」「です」「ます」の出現がない場合を示す。

口語文体は話し手の聞き手に対する丁寧さに関する待遇度、ひいては話し手と聞き手との関係性の一端を示すものであるとも言える。

また、本稿では聞き手を待遇対象とする形式のみを分析対象とした。聞き手以外の第三者を待遇対象とする形式を加えた場合、説明変数に「聞き手と待遇対象の関係」を加える必要が出てくるが、その分類が容易でないためである。また、東京語でない方言による会話に出現する形式も分析対象外とした。

ランダムフォレストの実行には統計分析ソフト R[6]の randomForest パッケージの randomForest()関数を用いた[7]。

４．１．命令形以外

命令形以外での頻度上位 6 形式「御～なさる」「御～だ」「おっしゃる」「～てくださる」「御～になる」「～なさる」についてランダムフォレストを行った。使用する 482 レコード中、382 レコードを学習用、100 レコードを評価用とした。

ランダムフォレストによって得られた分類ルールによる形式の分類実験の結果を表 5、説明変数（位相）の重要度を図 1 に示す。

表 5 分類実験の結果（命令形以外）

	実験結果					
	～てくださる	～なさる	おっしゃる	御～だ	御～なさる	御～になる
～てくださる	1	0	2	10	7	0
～なさる	0	5	0	2	3	0
おっしゃる	1	1	4	2	7	0
御～だ	0	3	2	9	5	0
御～なさる	2	3	3	0	16	0
御～になる	0	1	1	7	3	0

[6] http://www.r-project.org/

[7] スクリプト作成では金（2009）、石田・小林（2013）等を参照した。

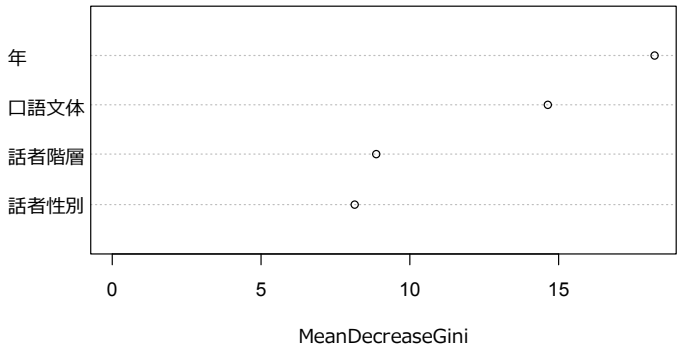


図 1　説明変数の重要度（命令形以外）

表 5 の分類実験の正解率は 35%と非常に低く、位相による形式の分類ルール作成は事実上失敗している。これは、6 形式の選択では 4 種の位相の影響力は限定的であることを意味する。ただし、こうした分類ルール作成の過程においてではあるが、図 1 から 4 種の位相のうち年と口語文体が他の位相と比べて重要度が高いという結果が得られた。重要度上位 2 位の各位相でどの形式が選択されやすいのか見るため、年別の形式比率を図 2 に、口語文体別の形式比率を図 3 に示す。

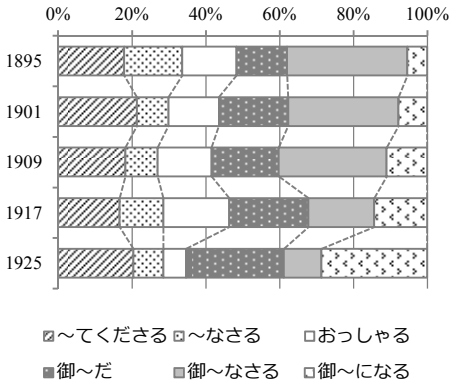


図 2　年別の形式比率（命令形以外）

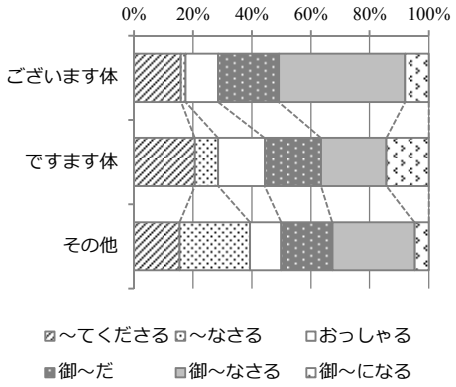


図 3　口語文体別の形式比率（命令形以外）

例えば、図 2 から 1895・1901・1909 年では「御～なさる」が、1925 年は「御～になる」「御～だ」が選択されやすく[8]、図 3 からは「ございます体」では「御～なさる」が選択されやすいことなどがわかる。

４．２．命令形

命令形での頻度上位 6 形式「～てくださる」「御～なさる」「～たまう」「御～」「～なさる」「御～くださる」についてランダムフォレストを行った。分析に使用する 575 レコード中、475 レコードを学習用、100 レコードを評価用とした。ランダムフォレストによって得られた分類ルールによる形式の分類実験の結果を表 6、説明変数（位相）の重要度を図 4 に示す。

[8] 当時の動詞の尊敬待遇表現形式の主な経年変化として、「御～なさる」の頻度の減少と「御～になる」の頻度の増加が指摘されており（辻村 1974、近藤 2005 等）、それと一致する結果と言える。

表 6 分類実験の結果（命令形）

	実験結果					
	～たまう	～てくださる	～なさる	御～	御～くださる	御～なさる
～たまう	26	0	0	0	0	0
～てくださる	7	16	0	1	0	4
～なさる	3	2	1	2	0	1
御～	4	0	0	7	0	0
御～くださる	0	6	0	0	0	0
御～なさる	5	7	0	2	1	5

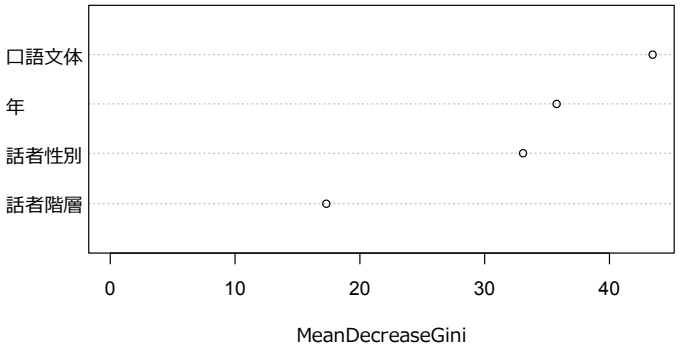


図 4 説明変数の重要度（命令形）

表 6 の分類実験の正解率は 55%で、命令形以外よりは高い値ではあるものの、分類ルールの作成に成功しているとは言い難い。命令形以外ほどではないにせよ、命令形においても、6 形式の選択に 4 種の位相はそれほど大きな影響力を持たないことがわかる。

また、図 4 から、4 種の位相のうち、重要度の高いものは第一に口語文体、次いで年と話者性別という結果が得られた。命令形以外の場合とは順位が異なることは注意される。重要度上位 3 位の各位相でどの形式が選択されやすいのか見るため、口語文体別の形式比率を図 5 に、年別の形式比率を図 6 に、話者性別別の形式比率を図 7 に示す。

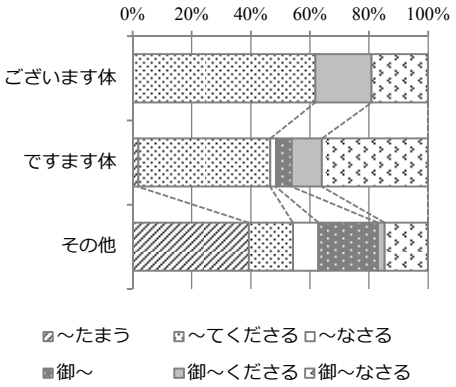


図 5 口語文体別の形式比率（命令形）

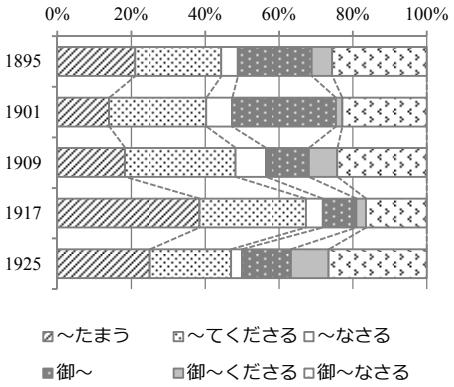


図 6 年別の形式比率（命令形）

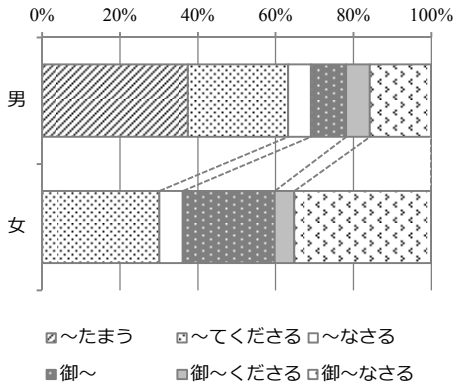


図 7　話者性別別の形式比率（命令形）

例えば、図 5 から「ございます体」では「～てくださる」が、「ですます体」では「～てくださる」「御～なさる」が、「その他」では「～たまう」が選択されやすく、図 6 からは 1917 年では「～たまう」が選択されやすく、図 7 からは「男」では「～たまう」が、「女」では「御～なさる」「～てくださる」が選択されやすいことなどがわかる。

５．おわりに

本稿ではランダムフォレストというコーパス言語学的な手法を用いて、動詞の尊敬待遇表現形式の選択における位相の重要度を分析することを試みた。その結果、特定の位相において選択されやすい傾向にある形式はあるものの、位相は形式選択に決定的な影響力を有することはないことが明らかになった。ただし、命令形においては命令形以外の場合よりも位相の影響度が高く、位相の重要度の順位も命令形とそれ以外では異なることも明らかになった。

取り上げた 4 種の位相が形式選択の決め手になり得ていないことは、これ以外に形式選択に重要な役割を果たす要素が存在する可能性があること、あるいは形式選択にはそもそも曖昧性が伴うことを示唆する。今後、より詳細な分析を行い、コーパス言語学的見地から新たな知見を得ることを目指したい。

付　記

本稿は、日本学術振興会科研費（23720242）および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」による研究成果である。

文　献

石田基広・小林雄一郎（2013）「第 10 章　方言データのテキストマイニング　「茸」のアクセントパターンを例に」『R で学ぶ日本語テキストマイニング』ひつじ書房、pp.173-190

金明哲（2009）「第 14 章　テキストの分類」『テキストデータの統計科学入門』岩波書店、pp.177-202

国立国語研究所（編）（2005）『太陽コーパス―雑誌『太陽』日本語データベース』博文館新社

近藤明日子（2005）「尊敬待遇表現―動作性の名詞や動詞連用形に付く形式について―」『国立国語研究所報告 122　雑誌『太陽』による確立期現代語の研究　『太陽コーパス』研究論文集』博文館新社、pp.227-250

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社、pp.1-48

陳慧玲（2005a）「明治前期東京語の直接命令表現の考察―位相との関連を視点として―」『文学研究論集』22、pp.1-20
陳慧玲（2005b）「明治前期東京語における士族・知識層男性の命令表現の考察」『文化継承学論集』1、pp.27-43
陳慧玲（2006a）「明治期東京語における直接命令表現の諸相」『明治大学日本文学』32、pp.77(19)-64(32)
陳慧玲（2006b）「明治期東京語における下層男性の命令表現の考察」『文学研究論集』25、pp.21-40
辻村敏樹（1974）「明治大正時代の敬語概観」『敬語講座 5　明治大正時代の敬語』明治書院、pp.7-33
辻村敏樹（編）（1991）『敬語の用法』角川書店
山西正子（1972）「「いらっしゃる」考」『国語学』88、pp.56-78

二つの文末形式「～感じだ」と「～感じがする」について

中平詩織（九州大学大学院人文科学研究科）

Study of Sentence Final-Form “kanji-da” and “kanji-ga-suru”

Shiori Nakahira (KYUSHU UNIVERSITY Graduate School of Humanities)

1. はじめに

本稿では、名詞「感じ」が文末で用いられる二つの形式について考える。

（1）　食べ過ぎておなかが張っている感じ｛だ／がする｝。

「感じ」は修飾部を必要とする名詞文（人魚構文）のひとつとして述べられる例のひとつである。その一方、動詞「スル」を伴って「感じがする」という形式も存在する。このとき、動詞「スル」は動詞としての本来の意味を失っており、名詞に付随している。

名詞「感じ」と、それにかかる修飾部までは共通しているが、文末の形式は二通り存在する。この二通りの形式には、何らかの違いが存在するのだろうか。それは例えば、文法的に承接する部分にずれはあるのか、使用される文体に異なりがあるのか、どちらのほうがより使われているか、といった疑問が生じる。

2. 調査概要

実例を調査するために、中納言を用い、調査を行った。どちらも「感じ」と共起するように設定し検索した。その後、考察の対象外となるものを手作業で除外した。動詞の活用の一部なのに異分析がなされているものや、文の中で今回の条件から外れるものである。

・「感じ＋助動詞」　7325 例

キー: (語彙素 = "感じ" AND 語彙素読み = "カンジ" AND 品詞 LIKE "名詞%") AND 後方共起: 品詞 LIKE "助動詞%" ON 1 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="50" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"

・「感じ＋（助詞）＋スル」　2729 例

キー: (語彙素 = "感じ" AND 語彙素読み = "カンジ" AND 品詞 LIKE "名詞%") AND 後方共起: (品詞 LIKE "動詞%" AND 語彙素 = "為る") WITHIN 3 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="50" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2";

「感じダ」は助動詞の中から検索したため、「ダ」以外の助動詞・文中で用いられる「感じデ」の例を除外した。「感じがスル」の場合、「感じ｛が／は／も｝する」等の助詞は同じ意味を表わすと見なすが、それ以外の「感じ｛ニ／デ｝する」等の助詞は別の表現（文末に限らず、また「感じ」の連体修飾節として見なされる）となるため、除外した。動詞「為る」は「実感する」「ビックリする」などの複合動詞も含まれてしまうため、別語を含んでしまう動詞は例外扱いとする。

例外を取り除いた結果、今回考察対象として取り扱う用例は、「感じダ」5450 例・「感じがスル」2403 例。

3. 用例の検討

3.1　「感じダ」

内容節を受け、「感じ」にコピュラがつく形式を観察する。(2)(3)のように形容詞修飾の形式が見られるが、状態を述べる形式は(4)のようなものも見受けられる。
また、「って感じ」「といった感じ」のように、発言・思考内容を受ける形式も見られた。

(2)　今日は久々にナポリタンを作ってみましたぁ☆　味付けは、塩こしょうとケチャップにチリパウダーをたっぷりかけてピリ辛にしました（＾＾）　まろやかな豆乳シチューと一緒なのでちょうどいい感じです♪　[OY03_11473]

(3)　ポリドールのところでお話したように、貴志康一はドイツ人にとっては珍しかったのでしょうが、われわれにとってはどうも陳腐な感じでした。[LBm7_00050]

(4)　「上体倒し」はおなかがプルプル震えていればＯＫ。「ワイドスクワット」は、太ももやすねの裏側の筋肉に力が入る感じです。　[PB34_00316]

(5)　ひよこ豆のカレーを食べてみました。感想は・・・まあ普通かなｗ　ちょっと濃厚な豆。でも肉の方がイイナー。尼さんになった気分だなー。って感じでした。　[OY03_08681]

(6)　お隣の岐阜県で有名な根尾谷の「薄墨桜」も少し出てきますよ。形態は絵本ですが、非常に落ち着いた色が使われていて、和の絵本といった感じです。[OP60_00004]

(7)　彼女なんですが、冬になると風邪を引きやすくなるようで、現在は咳が出て声が出にくそうな感じです。何か喉に効果があるお勧めがあったら教えて下さい。　[OC09_11705]

3.2　「感じがスル」

(8)　―まさか降雨があるのでは…？　御厨は空模様が気になる。関東大震災や広島の原爆投下の直後、大雨が降った話を思い出した。頭上の雲の色がどす黒い。厭な感じがしてきた。　[LBo9_00074]

(9)　Ｑスイーティーを食べたら、一粒だけに唇・のどにしびれがありました。原因は

何でしょうか。残留農薬等ではありませんか。Ａ　相談品の果肉部分はしびれる感じはしませんでした。 [PB23_00189]

（10）　信介がたずねると、トミちゃんはすこし首をかしげて、　「そうね、うまく言えないけれども、何かちょっと生気がないみたい。生気がないって言うより、迷っているみたいな感じがするわ。 [OB1X_00023]

（11）　一見、経路にはいっていない中央・篠ノ井・信越本線の名古屋市内―直江津間（三百二十五．八ｋｍ）の乗車券と、妙高高原―直江津間の急行券を買えばよさそうな感じがする。 [LBl6_00016]

接続する形式は「感じダ」に準ずるものであった。ただし、全てが同じ文脈で言い換えられるわけではないようである。

（12）a. 彼女なんですが、（中略）現在は咳が出て声が出にくそうな感じです。

b.？彼女なんですが、（中略）現在は咳が出て声が出にくそうな感じがします。

(12)は(7)の文末形式を 2 通り並べたものである。「感じダ」は自分以外の人物の意見・主観にも用いられるが、「感じがスル」ではそれが許されず、全て自分の主観を述べる形式となっている。

「感じダ」の形式は、人魚構文であったり文末名詞と言われるものである。新屋 (1989) は ”文末名詞” に関する論考の中で、本稿で扱う「感じ」についても分類しており、「感じ」は二種類の分析が与えられている。

C.主語で表されたものや一定の状況の様態を感覚的に把握して述べるもの。

D.述部が、主語（省略されることが多い）で表されたものの主観を表わすもので、さらに以下のように分けられる。

D-1. 述部が、主語で表されたものの身体的感覚を述べるもの。

D-2. 述部が、主語で表された主体の感情・心理を述べるもの。 (pp.86-85)

文末名詞は機能動詞結合との関わりが見られるとし、「～感じだ」は「～感じがする」と対応するとも述べられている。しかし、形式は対応するが意味として全く等位ではない。機能動詞結合の「感じがスル」は、一人称の主観を述べる（D）用法として用いられ、文末名詞の「感じダ」はそれよりも広く、状況や主観など広く接続することが可能な用法と考えられる。

4. レジスター

使用される文章・文体にも特徴が見られるかを、レジスターを見ることによって考える。

「感じダ」「感じがスル」は、どのような資料で用いられるか。各レジスターごとの用例数を、実例数と割合で示した。

表 1　「感じダ」「感じがスル」それぞれのレジスター使用割合

		感じダ	感じがスル	総数
図書館	書籍	833 (15.3%)	614 (25.6%)	1447 (18.4%)
出版	書籍	627 (11.5%)	477 (19.9%)	1104 (14.1%)
	雑誌	206 (03.8%)	86 (03.6%)	292 (03.7%)
	新聞	22 (00.4%)	6 (00.2%)	28 (00.4%)
特定目的	ブログ	2049 (37.6%)	564 (23.5%)	2613 (33.3%)
	知恵袋	1520 (27.9%)	449 (18.7%)	1969 (25.1%)
	ベストセラー	135 (02.5%)	92 (03.8%)	227 (02.9%)
	国会会議録	56 (01.0%)	98 (04.1%)	154 (02.0%)
	広報誌	2 (00.0%)	11 (00.5%)	13 (00.2%)
	教科書		5 (00.2%)	5 (00.1%)
	韻文		1 (00.0%)	1 (00.0%)
総計		5450	2403	7853

特定目的コーパスには他にも「法律」「白書」が存在するが、今回の調査語「感じダ」「感じがスル」の形式が用例として存在しなかったため、上記の表には現れなかった。主観を述べる表現が法律や白書に使用されないことは予想されるが、実際に用例が一例もないことによって予想は証明される。

「感じがスル」は「教科書」「韻文」レジスターでの使用が見られた。ただし、用例数が非常に少なく、例外的なものかと考えられる。

(13)　物に湿り気がないときに使ったり、川が流れる様子を表すときに使ったりする。そこにどんな共通点があるかを考えるのもおもしろい。それに、「さらさらの雪」といえば軽く乾いた感じがするが、「ざらざらした雪」というと、水分が多くて、ちょっと氷の混じったような雪を想像させる。　[OT02_00019 教科書（中学）　国語 1]

どちらもに共通する用例数が少ないレジスターは「出版・新聞」と「特定目的・広報誌」である。報道文の中に主観を述べる表現は基本的に使われない。しかしインタビュー記事等では、質問に対する意見の部分で使われる可能性があるために、数は少ないながら用例の存在が確認されるものと考えられる。

ここからは「感じダ」「感じがスル」それぞれのレジスターごとの量をみていきたい。

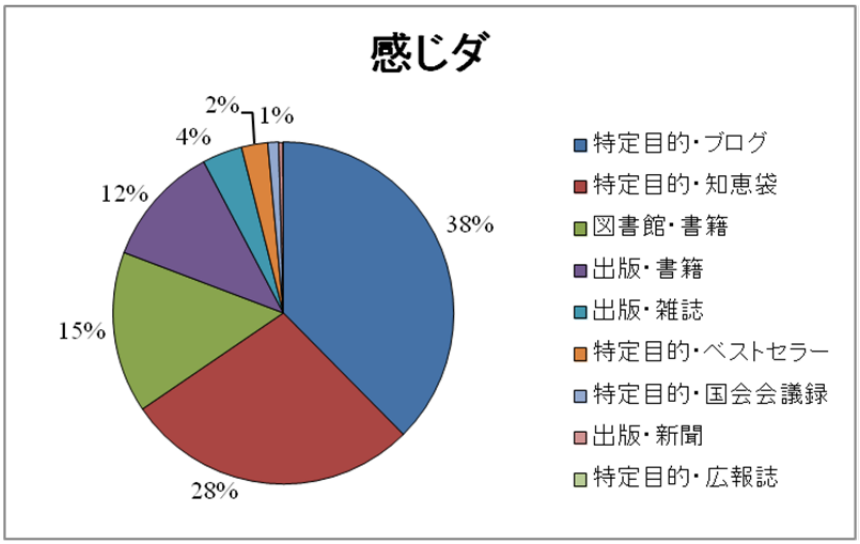


図 1　「感じダ」の使われるレジスター

「感じダ」はブログ (37.6%) ・知恵袋 (27.9%) が多い。これらのインターネット資料は「感じダ」使用例全体の 65.5%を占める。図書館・出版サブコーパスの書籍が次点となる。

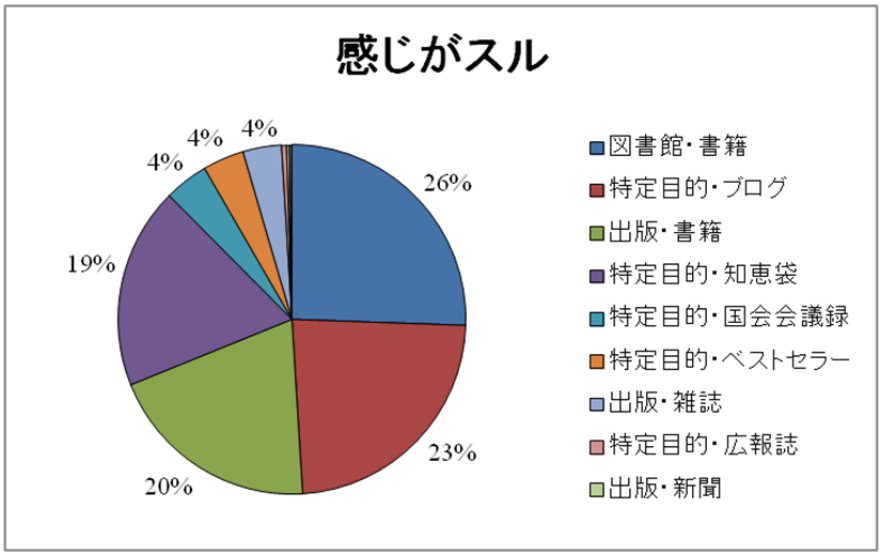


図 2　「感じがスル」の使われるレジスター

「感じがスル」は第二位にブログ (23.5%) が来るものの、図書館・出版サブコーパスの書籍の数を合わせると約半数(45.5%) に至る。この分布は「感じダ」の分布とは明らかに異

なる。

5. 表現のかたさ

それぞれの文末形式が使われるレジスターに異なりが見られたが、国会会議録に限ってみるとさらなる表現差が確認される。

国会会議録において、「感じダ」が使用されるのは1%であるのに対し、「感じがスル」の形式は 4%と僅かながら用例数が異なる。二者を比べると迂言的な表現である「感じがスル」のほうが、かたい表現として表れている。

（14）　これ以上譲歩できないという話があるわけですが、去年の暮れ以来の経過をずうっと見ていますと、政府の態度が大変対米従属といいましょうか、アメリカに対して弱腰と、こういう感じがするんですが、官房長官、その辺どうでしょうか。[OM15_00008]

6. 今後の課題

今回は共起文脈の指定により「感じダ」「感じがスル」の比較の調査であったが、文体のかたさという点では今回の調査ではカバーしきれない。おそらくコピュラをつけない「～感じゅ。」や、動詞「致す」を用いた「～感じがいたします」のような形態が存在しているが、未調査であるためである。文末に位置する名詞が文末に位置する時、どのような形式・どのような文体で用いられるかはスタイルによって自由である。

コーパスを利用した語彙研究で、複数のレジスターを縦断的に見られることには意義がある。今回のように、二つの文末形式が存在する場合にはレジスターの偏りなどが文体差に関わる。大量な例が見られる展はコーパスの利点であるが、同時に意味や用法を考える場合には、それぞれの例をよく観察する必要もあると考える。

今後は「感じ」の文末バリエーションごとの検討や、文体との対応を更に考える必要性が残される。また、「感じ」の類義語とも比較をしていきたい。

参考文献

新屋映子 (1989) 「”文末名詞”について」『国語学』159
藤田保幸 (2000) 『国語引用構文の研究』和泉書院
村木新次郎 (1991) 『日本語動詞の諸相』ひつじ書房

中納言 1.1.0、短単位データ 1.0、長単位データ 1.0

「書き言葉的」として指導する必要のある語の分析
－『現代日本語書き言葉均衡コーパス』を利用して－

木田 真理*（国際交流基金日本語国際センター）
柏野 和佳子（国立国語研究所 言語資源研究系）

Words that Need to Be Taught as "Written Language": Analysis of the BCCWJ

Mari Kida （The Japan Foundation Japanese-Language Institute,Urawa）
Wakako Kashino （Dept. Corpus Studies, NINJAL）

１．はじめに

レポートなどを書く際に、文体差を意識した適切な語彙を選択することは、上級レベルの日本語学習者にとっても難しい。そのため、学術的文章作成の指南書は、二通他（2009）、石黒（2012）など、数多く出版され、大学等では論文作成指導が行われている。しかしながら、従来「書き言葉」と「話し言葉」という文体差で整理されてきた表現も、近年、メールやブログ等に「話し言葉的」な表現が現れやすいという傾向があり、その影響から、新たな使用実態の調査が必要な語もある。学術的文章で用いるとよい語とそうでない語というような、文体差のある語の組をリスト化して示すものや，その文体差を計量的に分析するような文献や先行研究は十分にはない。

本研究では、大学院の修士課程で学ぶ留学生が、授業における課題やレポートとして書いた文章から、文体差の指導が必要なものを抽出し、文体差の修正を要する語の代表例を、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）で分析した。その分析結果を、BCCWJ の具体的な検索方法とともに報告する。

２．調査方法

２．１　調査対象データ

2012 年度日本語教育指導者養成プログラム[1]の学生 4 人が、大学院の授業における課題やレポート、研究論文などのアカデミックな文章として書いた合計 72 文書から抽出した。調査対象とした文書は、2012 年 11 月から 2013 年 3 月末までの約 5 か月間、1 年間の大学院コースの前半にあたる秋学期と冬学期に書かれたもので、1 つの文書は、A4 サイズ 2 枚～6 枚である。調査対象文書の作成者である修士課程の学生は、インドネシア、タイ、ミャンマー、チェコの大学や高校の教員で、全員日本語能力試験の 1 級または N1[2]に合格している。

２．２　調査対象の選定方法

学生が授業や発表会の課題で執筆して提出する文書に対して、週に 1 度、修正すべき箇所をチューターが添削指導しており、その修正箇所のうち、表記上、文法上、語彙選択上の明らかな誤用と思われるものを除いた修正例を抽出した。全部で 869 件の修正例が抽出された。これらの修正例には、最終的には誤用と判断されるものも含まれるが、その理由の説明が簡潔にはできない、初級レベルの明らかな誤用とは異なるものも含んでいる。そして、さらにその修正例から、筆者らの判断で、文体的特徴に関係の深い修正例を抽出すると、294 件となった。

294 件の用例について、まず全体的な傾向を見た。文体的な修正の観点には、日本語教育

* Mari_Kida@jpf.go.jp

[1] 国際交流基金日本語国際センターと政策研究大学院大学の連携大学院　現役の日本語教師を対象とした 1 年間の大学院修士課程

[2] 日本語能力試験は、2010 年に新しく改定され、1 級～4 級の 4 段階評価だったのを、N1~N5 の 5 段階評価に変更された。合格ラインは旧試験の 1 級とほぼ同じ。

の作文や論文指導の参考書である、アカデミック・ジャパニーズ研究会編（2002）、二通他（2003）、二通他（2009）、などで指導項目としてあがっている、接続表現、副詞、指示表現、文体（です、ます体、である体など）、書き言葉、なども多く見られたが、中には、従来の指導項目としてはあがっていないものも見受けられた。例えば次のような例である。

表１　修正指導が必要とされる項目例

抽出番号	修正箇所の用例 (太字、下線は修正箇所)	修正の観点	修正後 (太字、下線は修正箇所)
265	関心**を寄せている**ものの	受身表現	関心**が寄せられている**ものの
22	大学での既習者と未習者の混合**の**問題もタイの高等教育機関の課題である	の→　という	混合**という**問題
152	など**という**質問があった	という→　の	など**の**質問
24	インドネシア人学習者**には**外国語はともかく、	「は」を取る	学習者**は**外国語はともかく
224	大学入学試験の科目**に**入っておらず	「は」の挿入	科目**には**入っておらず
82	日本語**の**学習の動機になり	ＮのＮ　→　ＮＮ	**日本語学習**の動機
257	大学の**２学期試験**を受けた１年生は	ＮＮ　→　ＮのＮ	**２学期の試験**
51	一番**人気な**のは日本語だった	－な→－がある	**人気がある**
396	**様々の**経験を持ち込みながら	－の→　－な	**様々な**
100	学習者が十分に口頭コミュニケーションの練習ができ**なく**	ない→ず	**できず**
170	２回ともテーマが**絞られなくて**	なくて→ず	**絞られず**

レポートなどの文体には、受身表現のほうがふさわしい場合がある例（抽出番号 265 番）、「～という」と「の」の適切な選択（22 番と 152 番）、「は」の挿入が必要な場合とそうでない例（24 番と 224 番）、「の」の挿入に関する例（82 番と 257 番）、「－な」「－の」のいずれかの判断が必要な例（51 番と 396 番）、「ない」ではなく「ず」が適切な例（100 番と 170 番）など、指導上注目すべき項目が複数抽出された。調査対象文書の作成者は、全員日本語能力試験 1 級（N1）合格者で、日本語運用力が高く、文法や語彙選択なども、一文レベルでの使用は問題なくできている。しかし、まとまった文章を書き言葉で書く場合には、取り上げるべき指導項目がかなりあることを示している。

このような指導項目の詳細については、別稿にゆずり、本稿では、そのうち、①「N の N」か「NN」か、②「―な」か「―あるの／―の」か、③「ない」か「ず」か、の 3 点をとりあげ、BCCWJ の分析方法及び分析結果を報告する。

３．BCCWJ を用いた分析

本章では、前節でとりあげた①～③の 3 点について、実際の書き言葉における使用実態を確認する。その確認には BCCWJ を用いる（『中納言』1.1.0 を使用）。BCCWJ は一般に公開されており、書籍、新聞、雑誌や、Web 文書といった、レジスタの異なるサブコーパスが収録され、現代書き言葉の使用実態を、日本語教育の現場で誰もが検索して確認できるという利点があるため、BCCWJ を用いる。なお、本研究では、BCCWJ のサブコーパスのうち、話し言葉を書き記した国会会議録と、特殊な文体である法律文と韻文は検索対象外とする。

３．１　「N の N」か「NN」か

例えば、「日本語の教育」、「日本語教育」は、どちらも可能な表現形式である（表 2 126

番)。しかしながら、アカデミックな文章作成時においては、冗長と判断される「の」は省かれる場合が多い。今回、「の」を省くべきと修正された事例は表 2 に示すとおり、8 件あった。一方、学習者にとってはその「の」が冗長であるのか否かという判断が時に難しく、「の」を省くべきではない事例も収集されている。表 3 の 6 件である。

表 2　「N の N」から「NN」への修正例

抽出番号	修正箇所の用例 （【 】が修正箇所）	「NのN」の頻度	「NN」の頻度
82	【日本語の学習】の動機になり	13	46
94	そのイベントは【毎年の2月】に開かれ	3	524
126	【日本語の教育】の中心は	4	95
139	【大学生の生活】を送った時に	3	68
252	大使館に【就職の希望者】はN2合格者を要求する	0	33
340	2013年は【蛇の年】で	10	521
455	【研究のデータ】によると	1	23
633	【口頭の練習】としては教師からの質問に答える形が主である	0	0

表 3　「NN」から「N の N」への修正例

抽出番号	修正箇所の用例 （【 】が修正箇所）	「NのN」の頻度	「NN」の頻度
95	日本語学習者の【交流場】という目的を持っている	136	0
257	大学の【2学期試験】を受けた1年生は	0	4
268	大学院課程に指導できる【修士不足】	0	0
402	データの【収集し方】を見てみると	1	0
785	慣れない【授業雰囲気】	0	0
790	【日本事情授業】の見学体験をした	0	0

表 2 と表 3 には、「N の N」および「NN」の表現形式の BCCWJ における頻度を示している。一方の頻度が高い場合は高い方に色をつけて強調する。なお、原文通りの表現である「大学生の生活」および、「大学生生活」はいずれも頻度は 0 であったため、表 2 の 139 番は、「大学の生活」「大学生活」で検索をして得た頻度を示している。表 2 の 633 番はいずれも 0 であるが、「口頭」を「発表」に変えて検索をしてみると、「発表の練習」は頻度 1、「発表練習」は頻度 4 という結果が得られ、やや「NN」の方が多いことが確認できる。よって、表 2 に関しておおむね「NN」の頻度が高いという結果が得られる。

しかしながら、表 3 は、95 番以外は、ほとんど BCCWJ に用例がない。「NN」よりも適切と思われる「N の N」の形の用例すらとれないという点は、BCCWJ の規模では限界のある場合があることがわかる。日本語教育の現場で、大規模な N グラムデータを簡単に検索できる仕組みが整えられれば、BCCWJ での低頻度語の確認もしやすくなるかもしれない。それでも、BCCWJ の範囲内での工夫の仕方はある。例えば、268 番に関しては、「推薦人の不足」「証人の不足」で頻度 2、「職人不足」で頻度 1、402 番に関しては、「勉強の仕方（し方・しかた)」で頻度 42、「勉強仕方（し方・しかた)」で 0、785 番に関しては、「学校の雰囲気」で頻度 7、「学校雰囲気」で頻度 0、790 番に関しては、「国語の授業」で頻度 27、「国語授業」は頻度 0、との結果が得られる。つまり、似たような、頻度の得られそうな語に変えて検索をするという工夫により、低頻度語の場合でも、より優勢な形がどちらかを探ることは必ずしも不可能なことではない。

しかし、中には、検索に、コーパスを駆使する専門的知識が必要な場合がある。例えば、表 2 の 340 番の「蛇の年」などの例である。この場合、まず、「蛇の年」「蛇年」は、「巳年」

に比べ、そもそもあまり使われない（BCCWJ では、わずかに「蛇年」が頻度 1）ことをおさえ、次に十二支を「巳」に限定する必要もないことをおさえ、広く、十二支（子丑虎卯辰巳午未申酉戌）に続けて「の年」か「年」か、どちらの表現が使われているのかを検索すればよいということを思いつく必要がある。次に、例えば、一度でできる限り求める用例の頻度を得たい場合、次のように検索文字列が書けるとよい。

N の N：[^男女利父母甲乙丙丁戊己庚辛壬癸][子丑虎卯辰巳午未申酉戌亥]の年[^齢弱収頃長間金始籍少]
NN：[^男女利父母甲乙丙丁戊己庚辛壬癸][子丑虎卯辰巳午未申酉戌亥]年[^齢弱収頃長間金始籍少]

[^～]は、「～」以外の文字、[～]は、～のいずれかの文字、という指定である[3]。「子年」の「子」に連なる「男子」「女子」「利子」「父子」「母子」という用例や、「庚申の年」といった、十干（甲乙丙丁戊己庚辛壬癸）と十二支とを組み合わせる用例が、同時に検索されてしまうため、冒頭にこのように範囲外文字の指定をする。続けて、「(の) 年」の直前の[　]に、十二支の文字を入れて、いずれか 1 字に該当する場合、という指定をする。そして、「(の) 年」の直後には、「年齢」などの用例が同時に検索されないための範囲外文字の指定をする。この後、念のため目視で、不要な用例の有無を確認する。この場合は、「の年」の検索結果に、「死んだ子の年」「種子の年」というものが混じっていたため、最後にそれらを除外するという操作を行っている。

このように、検索には、分析対象語の性質によって、検索技術の難易度に差があり、検索方法に習熟を要する場合がある。しかしながら、日本語教育の現場において、ある表現が使われているかどうか、日本語母語話者でも判断に迷う場合、また、日本語学習者が正誤の判断をする際、コーパスを利用することによって使用実態を把握できるということは、非常に有益なことである。例えば、表 2 の 82 番「日本語の学習」と 94 番「毎年の 2 月」のように、「N の N」と「NN」の頻度の比較を手がかりに、誤用に近いのか、両方使用可能なのかの判断が可能となるのである。

３．２　「―な」か「―あるの／―の」か

例えば「人気」が名詞を修飾する際には、「―な」「―の」のどちらの形もあり得る。次の表現は、いずれもアカデミックな文章でも特に問題なく用いることができよう。

1)　若者に人気な店
2)　若者に人気の店

しかし、「の」を修飾する、次の表現を比べるとどうであろうか。

3)　人気なのが日本語だ。
4)　人気が（／の）あるのが日本語だ。

3)の表現よりも 4)の表現の方がアカデミックな文章にはふさわしく感じられる。今回の学習者のレポートにおいても、次の表 4 の 51 番に示すように、「人気な」から「人気がある」への修正が行われていた。

さらに、上記 1)と 2)の「人気」の例では、「―な」「―の」の形に文体差はなく感じられるが、今回の調査結果には、表 4 の 347 番から 710 番までの例に示すように、「―な」から「―の」へ修正された事例と、逆に、表 4 の 396 番に示すように、「―の」から「―な」へ修正された事例とがあった。いずれも、BCCWJ でその頻度を求めると、修正後の表現の方が高いという結果が得られている（高い頻度の方に色をつけて強調している）。それぞれの頻度を表 4 に示す。なお、中納言を用いて検索する場合には、適宜、「文字列検索」と「短単位検索」とを使い分けている。ワイルドカードを用いる場合には、「文字列検索」を、そうでなければ、表記のゆれが解消され、同音異義語の区別がされている、「短単位検索」を

[3] ワイルドカードとよばれる。中納言の文字列検索の画面にはその説明が表示されている。

表 4　「―な」「―あるの／―の」かの修正例

番号	修正箇所の用例（【　】が修正箇所）	修正後	「な」頻度	「あるの／の」頻度	「な」検索	「あるの」や「の」検索
51	一番【人気な】のは日本語だった	人気がある	28	49	文字列:人気なの[はが]	文字列:人気[のが]あるの[はが]
347	教師中心の教授法と【逆な】教育観の学習者中心の活動である	逆の	19	661	短単位:「逆」＋「な」＋名詞	短単位:「逆」＋「の」＋名詞
478	学習ストラテジーは漢字学習に【特有な】認知ストラテジーに限定した	特有の	92	988	短単位:「特有」＋「な」＋名詞	短単位:「特有」＋「の」＋名詞
702	日本語を【別な】外国語と同時に勉強する	別の	429	9537	短単位:「別」＋「な」＋名詞	短単位:「別」＋「の」＋名詞
710	【別な】ことばで言うと	別の				
396	【様々の】経験を持ち込みながら	様々な	12356	397	短単位:「様々」＋「な」＋名詞	短単位:「様々」＋「の」＋名詞

使う。表現それぞれの使用頻度をどちらの検索で調べたかをあわせて表 4 に示す。

「―な」か「―の」かは、どちらもあり得る形の場合に正誤は言い難いが、ネイティブの日本人がより文体差、あるいは、自然さを感じる違いがあるような場合には、このように検索結果の頻度差として、そのことを確認することができる。

３．３　「ない」か「ず」か

学習者のレポートの中で、「ない」が「ず」に修正された事例が、次の表 5 に示すとおり 5 件あった。「ない」も「ず」も、用例が多数あり、必ずしも、お互いに言い換え可能ではない用例も多数ある。そこで、お互いに言い換え可能であるペアになるような表現形で検索を行う。ここでは、表 5 の 100 番と 435 番より、「できなく」と「できず」の組と、170 番や 363 番のような用例をできるだけ調べるために、「なくて」（「なく、」を含む）と「ず、」の組を取り上げ、その文体差を BCCWJ のレジスタ別に分析してみることにする。

表 5　「ない」から「ず」への修正例

抽出番号	修正箇所の用例（【　】が修正箇所）	修正後
100	学習者が十分に口頭コミュニケーションの練習が【できなく】	できず
435	習った文法知識を使用し、運用【できないこと】、十分に話すことができないといった問題点があり	できず
170	2回ともテーマが【絞られなくて】	絞れず
363	焦点もはっきり【しなく】	せず
415	話し手は難しい文法や豊富な語彙は要求されて【いないで】	おらず

BCCWJ から得られる頻度は、「できなく」3,281、「できず」2,882[4]である。また、「なくて」（なく、）6,802[5]、「ず、」33,420[6]である。BCCWJ のサブコーパスを、硬[白書，教科書，広報紙，新聞]，中[書籍]，軟[雑誌，Web データ]の 3 つに区分し、それぞれの使用頻度を求める。その結果を、表 6、表 7 に示す。

表 6、表 7 では、出現率を、サブコーパス別に、使用頻度を 100 万語あたりで正規化した数値（PMW）で求めている。それから硬、中、軟別に出現率の平均を求め、硬、中、軟別に表現形の比率を求めた。表現形の比率のグラフを、次の図 1,2 に示す。

4 「できなく」も「できず」も文字列検索のあと、「キー」でソートし、人手で選別。

5 「[^かきつしむもすはっだほ][かさたなまやらわえけせてねみれ]なく[て、][^は]」で検索ののち、人手で選別。

6 単独の用例検索だと、用例が多すぎるため、「ず、」（ず：助動詞＋「、」）で検索。

表 6　「できなく」と「できず」の[硬中軟]別の出現率の比較

硬軟		サブコーパス	頻度	語数	出現率(PMW)	比率
硬	できなく	出版・新聞	26	1061729	24.5	46.0%
		特定目的・教科書	19	928448	20.5	
		特定目的・広報紙	106	3755161	28.2	
		特定目的・白書	42	4685801	9.0	
	硬の「できなく」の平均				20.5	
	できず	出版・新聞	59	1061729	55.6	54.0%
		特定目的・教科書	19	928448	20.5	
		特定目的・広報紙	39	3755161	10.4	
		特定目的・白書	47	4685801	10.0	
	硬の「できず」の平均				24.1	
中	できなく	出版・書籍	1053	28348233	37.1	52.2%
		図書館・書籍	1025	30377866	33.7	
		特定目的・ベストセラー	125	3742261	33.4	
	中の「できなく」の平均				34.8	
	できず	出版・書籍	964	28348233	34.0	47.8%
		図書館・書籍	914	30377866	30.1	
		特定目的・ベストセラー	117	3742261	31.3	
	中の「できず」の平均				31.8	
軟	できなく	出版・雑誌	139	4242224	32.8	55.4%
		特定目的・ブログ	247	10101397	24.5	
		特定目的・知恵袋	499	10162945	49.1	
	軟の「できなく」の平均				35.4	
	できず	出版・雑誌	104	4242224	24.5	44.6%
		特定目的・ブログ	307	10101397	30.4	
		特定目的・知恵袋	312	10162945	30.7	
	軟の「できず」の平均				28.5	

表 7　「なくて」と「ず、」の[硬中軟]別の出現率の比較

硬軟	語	サブコーパス	頻度	語数	出現率(PMW)	比率
硬	なくて	出版・新聞	34	1061729	32.0	7.3%
		特定目的・教科書	17	928448	18.3	
		特定目的・広報紙	25	3755161	6.7	
		特定目的・白書	10	4685801	2.1	
	硬の「なくて」の平均				14.8	
	ず、	出版・新聞	535	1061729	503.9	92.7%
		特定目的・教科書	44	928448	47.4	
		特定目的・広報紙	450	3755161	119.8	
		特定目的・白書	379	4685801	80.9	
	硬の「ず、」の平均				188.0	
中	なくて	出版・書籍	1540	28348233	54.3	12.6%
		図書館・書籍	1703	30377866	56.1	
		特定目的・ベストセラー	285	3742261	76.2	
	中の「なくて」の平均				62.2	
	ず、	出版・書籍	10644	28348233	375.5	87.4%
		図書館・書籍	13127	30377866	432.1	
		特定目的・ベストセラー	1820	3742261	486.3	
	中の「ず、」の平均				431.3	
軟	なくて	出版・雑誌	251	4242224	59.2	29.5%
		特定目的・ブログ	1044	10101397	103.4	
		特定目的・知恵袋	1893	10162945	186.3	
	軟の「なくて」の平均				116.3	
	ず、	出版・雑誌	1454	4242224	342.7	70.5%
		特定目的・ブログ	2522	10101397	249.7	
		特定目的・知恵袋	2445	10162945	240.6	
	軟の「ず、」の平均				277.7	

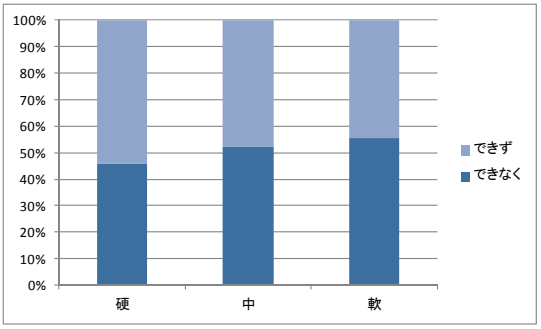


図 1　「できなく」「できず」の比率の比較

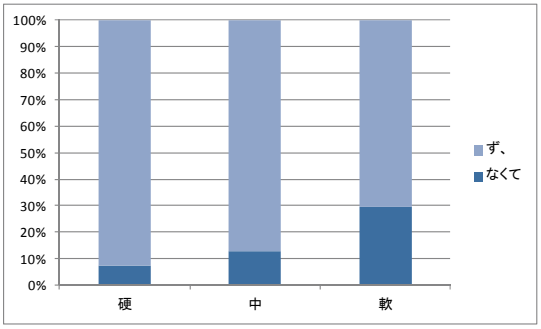


図 2　「なくて」「ず、」の比率の比較

図 1、図 2、いずれも、硬では「ず」の形（「できず」「ず、」）の比率が一番高く、軟では「ない」の形（「できなく」「なくて」）の比率が一番高く、中ではその中間となっている。つまり、文章が硬いほど「ず」が用いられ、軟らかいほど「ない」の形が用いられているといえる。アカデミックな文体は、硬い文体に位置づけられるため、アカデミックな文章作成時に、「ない」から「ず」への修正を学習者に指導する際には、こういった分析結果を学習者へ示すことで、より、文体差を理解してもらうことができるだろう。

４．おわりに

大学院の留学生が、課題やレポート、研究論文などとして書いた文書に、文体的な観点から修正が必要とされる例のうち、①「N の N」か「NN」か、②「―な」か「―あるの／―の」か、③「ない」か「ず」か、の 3 点をとりあげ、BCCWJ を用いて分析し、その検索方法及び分析結果を報告した。コーパスの分析結果が示す使用実態は、日本語教師にとっても、学習者にとっても文体差の修正が必要とされる度合いの手がかりとなる有益な情報である。日本語学習者自身にも、BCCWJ を使用して検索ができるようにする指導を行えば、学習者が必要に応じて使用実態を確認することもできるようになるであろう。本稿で調査対象データを作成した学生が在籍する日本語教育指導者養成プログラムでも、日本語学の授業で、コーパスを使った現代日本語の分析の授業を実施している。BCCWJ を用いて検索する場合、その検索方法がシンプルで容易な場合と、別な語に置き換える工夫が求められたり、複雑な検索式を書く必要がある場合などがあり、その検索技術の習熟には、時間を要する面がある。検索しやすい語とそうでない語の区別も踏まえた上で、学習者自身も、検索の技術を学んでいくことが重要である。また、本稿で分析対象とした語彙の中には、「口頭練習」「日本事情の授業」など、学習者が学んでいるプログラムの特徴を反映するものも含まれている。そして、その検索には、限界があることも確認された。今後、日本語学習者が自分で容易に検索することが可能となるためには、大きな N グラムデータにおいて簡単に検索ができる仕組みが整えられる必要があるであろう。いずれにせよ、日本語教育の現場において、BCCWJ の積極的な活用が望まれている。

今後は、文体的な修正を必要とされた例の整理、分類を継続し、BCCWJ を活用した使用実態の調査分析、そこから得られた情報を、日本語学習者向けの参考書や辞書類に記述する試みなどに取り組んでいきたい。

謝　辞

調査の補助をしてくださった国立国語研究所コーパス開発センター技術補佐員の田嶋明日香さんに感謝します。本研究は、文部科学省科学研究費補助金基盤研究(C)「コーパス分析に基づく辞書の位相情報の精緻化」（課題番号:23520572）の助成を受けたものです。

文　献

アカデミック・ジャパニーズ研究会編著（2002）『大学・大学院留学生の日本語　4論文作成編』アルク

石黒圭（2012）『この1冊できちんと書ける！論文・レポートの基本』日本実業出版社

木田真理、柏野和佳子（2012）「書き言葉均衡コーパスを活用した文体的特徴を持つ語の分析と記述の試み　―アカデミック・ライティング指導を例に―」『日本語教育国際研究大会名古屋予稿集』A268.

二通信子、佐藤不二子（2003）『改訂版　留学生のための論理的な文章の書き方』スリーエーネットワーク

二通信子、大島弥生、佐藤勢紀子、他（2009）『留学生と日本人学生のためのレポート・論文表現ハンドブック』東京大学出版会

堀一成、坂尻彰宏、石島悌（2013）「BCCWJ 教科書データより抽出した頻度情報に基づく日本語ライティング指導教材の作成」『第4回コーパス日本語学ワークショップ予稿集』pp.45-52.

「N を通じて」と「N を通して」について

趙　海城（明星大学　人文学部）

On the Japanese Compound-Postpositions: "N wo tsuuzite" vs. "N wo tooshite"

ZHAO Haicheng　(Meisei University)

１．はじめに

以下の例は泉原（2007：362）より引用したものである。a 文では、「通して」が自然で、b 文では「通じて」が自然であり、両者は置き換え不能とされている。

a. 私たちは、居ながらにして、世界中の情報を、インターネットを＋通して／×通じて、手に入れることができる。

b. ハッカーたちは、ウィルスに感染されたファイルが送りこめるインターネットを＋通じて／×通して、ユーザーのシステムを破壊していく。

a. 大自然のもとで、汗を流しながら、一緒に働いた経験を＋通して／×通じて、ふたりは生涯かわらぬ友愛をもちつづけた。

b. 戦場で、血みどろになりながら、一緒に戦った経験を＋通じて／×通して、ふたりは生涯かわらぬ友愛をもちつづけた。

a、b 両者の違いについて、泉原（2007：363）では以下のように分析している。

A＋を通して＋B：A を＋媒介／手段＋として、単純明快／公明正大＋に B が行われる場合に使われる
A＋を通じて＋B：媒介／手段＋とした A が、公にできない秘密のものであったり、不正なものであったりする場合に使われる

つまり後者は「内緒／内輪／内密／秘密裏／裏やコネを使った／おおっぴらにできない／隠しておきたい」という話し手の気分が反映される場合に多用される。逆に言えば、前者で「後ろ暗い／やましい＋ところはない」といった気持ちを表すことができる。

また、「時間的地理的一貫性」を示す以下の用例における両者の違いについて、泉原（2007：363）では「ほとんど話し手の好みの問題になり、固有日本語、つまり訓読みの和語が好きな人は、前者「～を通して」を使い、音読みの漢字語が好きな人は、後者を使うといった傾向があるだけのように思われる。」と述べている。

・このあたりは、一年／四季＋を＋通して／通じて＋花が咲き乱れる。

・彼女は、一生／生涯＋を＋通して／通じて＋伝道に身を捧げた。

果たして、「を通して」「を通じて」が使われる文には上記のような話し手の気分の違いがそれぞれ反映されているのか、また「時間的地理的一貫性」（泉原用語）を示す用例において、「を通して」を使うかそれとも「を通じて」を使うかは話し手の好みによるものなのか。このような疑問を持って、本稿では、「Nを通じて」「Nを通して」を取り上げ、前接名詞Nの違いを中心に、両者の相違点を探りたい。なお、本稿では「Nを通じて」「Nを通して」の具体的な意味・用法のズレの部分に立ち入らない。また、両者の使用に影響する

ものと思われる後続述語成分についても触れない。

２．先行研究

上記泉原（2007）以外に、以下のような研究が見られる。

森田・松木(1989)では、「格助詞の働きをするもの」の下位項目に「仕手・仲介・手段・根拠・原因を示す」を挙げている。その中で「を通して」「を通じて」[1]を取り上げて説明している(pp.20～21)。

◇人物や物事を仲立ちとして何かを行うことを表す表現で、「によって」より間接的である。“～を経由して”“～を手段として”の意を表す場合もある。

・映画を通して都会生活に憧れたこの娘の素朴な夢は……

・書物を通じて得た知識や情報を……

◇「を通して」「を通じて」はほぼ同意で相互の入れ換えが可能であるが、次例を見ると、「を通じて」の使用範囲のほうが多少狭いようである。

・学生の生態を通して＜を通じて＞日本の教育を論じている。

◇連体格の用法としては、「を通しての」「を通じての」がある。

・文通を通して＜を通じて＞の五年間の恋が実った。

また、「起点・終点・範囲を示す」項目の中で「を通じて」[2]を取り上げ、「にわって」とともに記述している(pp.34～35)。

◇場所や期間がある範囲全体に及ぶことを表す表現である。

・その基底を成す人間性とでも言うべきものは諸民族を通じてほぼ同一であるらしい事……

◇「を通じて」はさらにこの用法を一歩前進させて、物事がある期間継続して行われることを表現する用法があるとしている。

・このようなことは人類の歴史を通じていつも時代にも見られたことである。

◇「を通じて」の連体格は「を通じての」である。

・日本列島を通じての帰巣本能に駆られた民族大移動も……

グループ・ジャマシイ(1998)では、「を通して」「を通じて」を以下のように説明している。【とおして】(pp.315～316)

◇N をとおして＜仲立ち＞：人やものごと、動作を表す名詞などを受け、「それを仲立ちや手段にして」という表示を表す。それによって知識や経験などを得ることを述べるのに使う。

・私たちは友人を通して知り合いになった。

・実験を通して得られた結果しか信用できない。

◇Vーることをとおして：動詞の辞書形を受けて、上の１と同様の意味を表す。和語の動詞を受けるのが普通で、「学習する」「研究する」のような漢語の動詞の場合は、１の表現を使って「学習／研究をとおして」のように言うほうが一般的である。

---

[1] この他に「によって、により、によると、によれば、をもって、でもって、にして、につき」を取り上げている。

[2] この他に「からして、をはじめ、に至るまで、にかけて、を通じて、にわたって」を取り上げている。

・子供は、学校で他の子供と一緒に遊んだり学んだりすることを通して、社会生活のルールを学んで行く。

◇N をとおして＜期間中＞：期間を表す語を受けて、「その期間中」「期間の範囲内」といった意味を表す。期間中ずっと継続的に行為が行われたり、期間内に断続的に生じる出来事などを表す。

・5 日間を通しての会議で、様々な意見が交換された。

・この一週間を通して、外に出たのはたったの 2 度だけだ。

【をつうじて】(pp.650～651)

◇N をつうじて V：「…を経由して」という意味。何かを経由して情報を伝えたり関係ができたりするということを述べるときに使う。伝わるのは情報・話・連絡などで、交通手段は使えない。書き言葉的。「…をとおして」とも言う。

・その話は山田さんを通じて相手にもつたわっているはずです。

◇N をつうじて：時間を表す名詞に付いて、「ある一定の期間とぎれることなくずっと」という意味を表す。書き言葉的。「…をとおして」とも言う。

・その国は一年をつうじてあたたかい。

・このあたりは四季をつうじて観光客のたえることがない。

庵功雄他(2001)では、複合格助詞として「手段を表す表現」の中で、「によって、をもって」とともに、「を通じて」を取り上げている(pp.23～24)。

◇行為や情報伝達を媒介するものを表す。

・マネージャーを通じてその俳優の出演を依頼した。

・そのニュースは通信社を通じて海外に発信された。

◇変化の原因となる経験を表す用法もある。

・留学生との交流｛を通じて／によって｝彼は視野を多いに広げた。

◇節全体を手段とする場合もある。

・ボランティア活動に参加することを通じて多くのことを学んだ。

◇「を通じて」に似た形式として「を通して」がある。

・代理人を｛○を通じて／○を通して｝球団に移籍を打診した。

また、「空間的・時間的範囲を表す形式」の中で、「にかけて、にわたって」とともに、「を通じて」を取り上げている(p.34)。

◇期間や空間を表す名詞について「～の間ずっと」という意味を表す。「を通して」もほぼ同じ意味で用いられる。

・熱帯地方では 1 年を通じて寒暖の差があまり大きくない。

・北陸自動車道は雪のため全線を通じて 50 キロ規制になっている。

先行研究を概観すると、両者には「仲立ち／媒介／手段」「期間・空間」の用法に分かれて議論されていると言えよう。また、両者は書き言葉的であるとまとめられる。

３．用例採集資料と調査対象、使用傾向

資料：「現代日本語書き言葉均衡コーパス」(以下 BCCWJ と呼ぶ)を使用した。BCCWJ は語数が約 1 億 490 万語で、今回これらすべてを検索対象とした。

調査対象：「中納言」を利用して、「短単位検索」を行った。語彙素を「通す」「通ずる」に

指定し、キーを名詞に指定することにより、「N を(通し/とおし/透し/とほし）て」、「N を(通じ/つうじ）て」をすべて検索した。各レジスターの割合及びヒットした用例数は表 1 の通りである。

BCCWJ は 13 のサブコーパスからなっているが、本稿では、「出版・書籍」「図書館・書籍」「特定目的・ベストセラー」を同じ「書籍」とみなして一つにまとめた。したがって、表 1 のように、本研究では 11 のサブコーパスをレジスターとして、考察対象の出現頻度を見る。なお、本稿では前接名詞Nは複合辞の直前に位置する語とする。

「N を通して」は 4416 例[3]で、N の異なり語数は 1613 個だった。「N を通じて」は 5015 例[4]で、N の異なり語数は 1655 個だった。

N が 5 回以上出現した「N を通して」は 2333 例で、N の異なり語数は 147 個だった。同じく N が 5 回以上出現した「N を通じて」は 2933 例で、N の異なり語数は 190 個だった。本稿では N が 5 回以上出現した用例を分析対象とした。

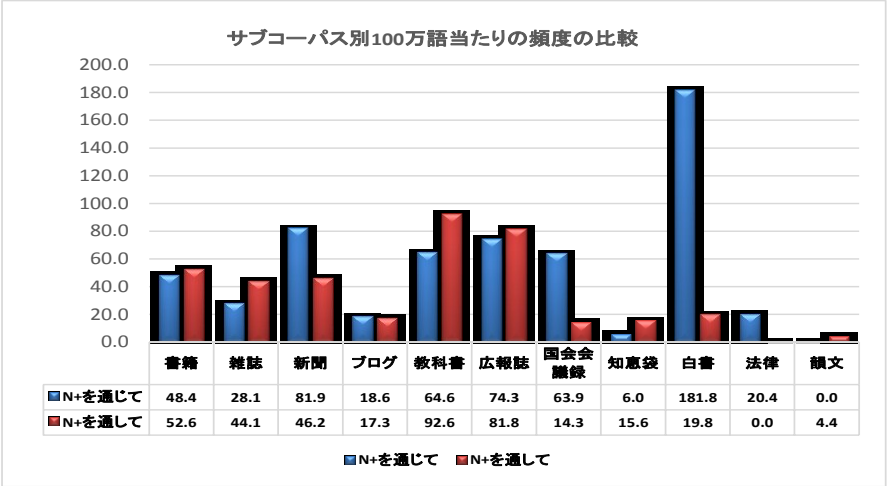
表 1　各レジスターの語数%、「通して」「通じて」の用例数、%、100 万語当たりの頻度[5]

	書籍	雑誌	新聞	ブログ	教科書	広報誌	国会会議録	知恵袋	白書	法律	韻文	合計
%	60.3	4.1	1.0	9.8	0.9	3.6	4.9	9.8	4.5	1.0	0.2	100.0
通じて	3021	119	87	188	60	279	326	61	852	22	0	5015
%	60.2	2.4	1.7	3.7	1.2	5.6	6.5	1.2	17.0	0.4	0.0	100.0
頻度	48.4	28.1	81.9	18.6	64.6	74.3	63.9	6.0	181.8	20.4	0.0	48.4
通して	3286	187	49	175	86	307	73	159	93	0	1	4416
%	74.4	4.2	1.1	4.0	0.0	1.9	7.0	1.7	3.6	0.0	0	100.0
頻度	52.6	44.1	46.2	17.3	1.1	22.9	60.2	7.2	33.9	0.0	4.4	42.6

[3] 【N＋を通して 4416 例】の検索条件式
キー: 品詞 LIKE "名詞%" AND 後方共起: (語彙素読み = "ヲ" AND 品詞 LIKE "助詞-格助詞%") ON 1 WORDS FROM キー AND 後方共起: 語彙素 = "通す" ON 2 WORDS FROM キー AND 後方共起: (語彙素読み = "テ" AND 品詞 LIKE "助詞-接続助詞%") ON 3 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

[4] 【N＋を通じて 5015 例】の検索条件式
キー: 品詞 LIKE "名詞%" AND 後方共起: (語彙素読み = "ヲ" AND 品詞 LIKE "助詞-格助詞%") ON 1 WORDS FROM キー AND 後方共起: 語彙素 = "通ずる" ON 2 WORDS FROM キー AND 後方共起: (語彙素読み = "テ" AND 品詞 LIKE "助詞-接続助詞%") ON 3 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

[5]各サブコーパスの%は中納言サイトによるものと違う。



	書籍	雑誌	新聞	ブログ	教科書	広報誌	国会会議録	知恵袋	白書	法律	韻文
N+を通じて	48.4	28.1	81.9	18.6	64.6	74.3	63.9	6.0	181.8	20.4	0.0
N+を通して	52.6	44.1	46.2	17.3	92.6	81.8	14.3	15.6	19.8	0.0	4.4

図１　サブコーパス別の 100 万語あたり「Nを通じて」「Nを通して」の出現頻度

表１に示しているように、100 万語当たりの頻度は、「Nを通じて」は 48.3 語で、「Nを通して」は 42.5 語で、「Nを通じて」のほうはやや多いが、大きな違いは見られなかった。しかし、表１、図１が示すように、サブコーパス別に見ると、出現の偏りが見られた。まず、「Nを通じて」の用例が多いのは「白書」「新聞」「広報誌」「教科書」「国会会議録」である。この５つのサブコーパスはどちらというと、より書き言葉的であると言える。法律は「Nを通して」が 0 語であるのに対し、「Nを通じて」の出現頻度が 20.4 語である。法律は書き言葉的で、漢語を多用すると思われる。以上から、「Nを通じて」が書き言葉的であることを裏付けていると言える。

一方、「Nを通して」の出現が多いのは「教科書」「広報誌」「書籍」である。雑誌は多ジャンルの内容に分かれ、一概に書き言葉的あるいは話し言葉的とは言えない。

「Nを通じて」「Nを通して」両方出現頻度が低いのは「ブログ」「知恵袋」であった。知恵袋、ブログは出版される印刷物とは異なり、一般の人が自由に書き込んでおり、「書き手と読み手という相互交渉の場という意味で「会話」の一種とみなすことができる。(秋月 2007)」という特徴を持っている。これらの書かれた発話、会話的要素の強いレジスターには両者の出現頻度が低い。

４．前接名詞の意味的分布

『分類語彙表』[6]の意味分類を使って、「N を通して」と「N を通じて」における前接するNの意味的分布を見た。『分類語彙表』にない語句はもっとも適切な意味番号を付与した。Nが多義語の場合、できるだけその文脈に合った意味分類をした。

4．1.「N を通して」だけに前接する名詞

先ほど触れたように、「N を通して」における N の出現頻度が 5 回以上のものは 147 個で、2333 例であった。この 147 個のNの意味分類は 91 種類で、そのうち、「N を通して」だけの意味分類は 35 種類で、N の異なり語数は 45 個で、用例数は 677 例であった。表 2 は「N を通して」だけに前接する名詞の意味分類及び名詞例を示している。

[6]国立国語研究所(2004)『分類語彙表―増補改訂版』大日本図書

表２　「N を通して」だけに前接する名詞の意味分類及び名詞例【大項目意味ごと】

大項目意味	中項目意味	前接名詞の意味コード	前接名詞の意味分類	前接名詞	出現数
自然物及び自然現象	自然	1.5010	光	映像(9)、	9
	身体	1.5601	頭・目鼻・顔	目(246)、眼(25)、口(18)、	289
		1.5603	手足・指	腕(5)、手(5)、	10
		1.5604	膜・筋・神経・内臓	胎盤(6)、	6
		1.5605	皮・毛髪・羽毛	皮膚(6)、	6
	物質	1.5120	空気	空気(8)、	8
		1.5151	風	風(7)、	7
		1.5161	火	火(47)、	47
人間活動の主体	家族	1.2130	子・子孫	***子ども(6)、***	***6***
	人間	1.2020	自他	***自分(5)、***	***5***
	成員	1.2410	専門的・技術的職業	***通訳(7)、***	***7***
人間活動ー精神及び行為	言語	1.3100	言語活動	**主張(6)、**	**6**
		1.3151	書き	**記録(5)、**	**5**
	行為	1.3410	身の上	**独身(8)、**	**8**
	心	1.3001	感覚	**五感(5)、感覚(11)、**	**16**
		1.3020	好悪・愛憎	**鑑賞(6)、**	**6**
		1.3070	意味・問題・趣旨など	**問題(7)、**	**7**
		1.3071	論理・証明・偽り・誤り・訂正	**筋(17)、**	**17**
	生活	1.3331	食生活	**食(7)、食事(5)、**	**12**
		1.3371	旅・行楽	**旅(6)、**	**6**
	待遇	1.3611	裁判	**裁判(7)、**	**7**
生産物及び用具	衣料	1.4200	衣料・綿・革・糸	糸(10)、	10
		1.4240	そで・襟・身ごろ・ポケットなど	袖(11)、	11
	機械	1.4610	鏡・レンズ・カメラ	カメラ(7)、レンズ(28)、眼鏡(5)、フィルター(20)、	60
	資材	1.4120	木・石・金	ガラス(14)、	14
		1.4160	コード・縄・綱など	ワイヤー(5)、紐(5)、	10
	住居	1.4440	屋根・柱・壁・窓・天井など	壁(8)、窓(17)、	25
	食料	1.4300	食料	食べ物(5)、	5
抽象的関係	空間	1.1741	上下	下(5)、	5
		1.1750	面・側・表裏	画面(6)、	6
	形	1.1800	形・型・姿・構え	姿(10)、	10
		1.1830	穴・口	穴(13)、	13
	時間	1.1635	朝晩	闇(5)、	5
	様相	1.1300	様相・情勢	状況(6)、	6

	量	1.1951	群・組・対	シリーズ(7)、	7
合計					677

「N を通して」だけに前接する名詞には和語が多いことが分かる。N の異なり語数 45 個中、2 字漢語サ変可能名詞は 6 例だけであった。また、N の意味分類を見ると、「人間活動主体」は「子供、自分、通訳」のような特定の人間を指していることが分かる。また、「人間活動―精神及び行為」は「私的な活動、感覚」を表す名詞が多いことが分かる。よって、「N を通して」は 4．2．で見る「N を通じて」と比べれば、どちらかというと私的な、具体的な事柄について述べる傾向にあると言えよう。さらに、「生産物及び用具」のグループは「糸、ワイヤー、紐」のような「移動させられるもの」、「袖、レンズ、ガラス」のような「通過場所」を表すもので、この種類の名詞が前接する「を通して」は独自の意味で、「を通じて」にない用法である。

4．2．「N を通じて」だけに前接する名詞

「N を通じて」における N の出現頻度が 5 回以上のものは 190 個、2933 例であった。190 個の名詞の意味分類は 103 種類で、そのうち、「N を通じて」だけの意味分類は 47 種類、N の異なり語数は 62 個で、用例数は 493 例であった。表 3 は「N を通じて」だけに前接する名詞の意味分類及び名詞例を示している。

表 3　「N を通じて」だけに前接する名詞の意味分類及び名詞例【大項目意味ごと】

大項目意味	中項目意味	前接名詞の意味コード	前接名詞の意味分類	前接名詞	出現数
自然物及び自然現象	自然	1.5030	音	音(7)、	7
	天地	1.5210	天体	衛星(7)、	7
		1.5250	川・湖	川(5)、	5
人間活動の主体	機関	1.2710	政府機関	***政府(5)、***	***5***
		1.2720	公共機関	***ＪＩＣＡ(6)、***	***6***
		1.2750	国際機構	***国連(6)、***	***6***
	公私	1.2540	都会・田舎	***地方(14)、***	***14***
		1.2550	政治的区画	***県(7)、***	***7***
	社会	1.2620	現場	***窓口(7)、***	***7***
		1.2630	社寺・学校	***学校(19)、大学(8)、***	***27***
	成員	1.2411	管理的・書記的職業	***大臣(5)、***	***5***
		1.2412	実業・商業	***業者(11)、***	***11***
人間活動―精神及び行為	経済	1.3710	経済・収支	**投資(8)、予算(5)、**	**13**
		1.3760	取引	**取引(6)、貿易(6)、**	**12**
		1.3770	授受	**提供(9)、**	**9**
	言語	1.3133	会議・論議	**議論(13)、会議(7)、会談(6)、協議(12)、審議(8)、話し合い(8)、**	**54**
		1.3135	批評・弁解	**評価(6)、批判(7)、**	**13**

		1.3161	目録・暦	**番組(5)、**	**5**
	交わり	1.3541	奉仕	**サービス(5)、**	**5**
		1.3542	競争	**競争(18)、試合(5)、**	**23**
		1.3551	戦争	**戦争(10)、**	**10**
	事業	1.3801	生産・産業	**開発(9)、**	**9**
		1.3833	興行	**開催(8)、**	**8**
		1.3850	技術・設備・修理	**整備(10)、**	**10**
		1.3852	扱い・操作・使用	**利用(9)、**	**9**
	心	1.3040	信念・努力・忍耐	**努力(10)、**	**10**
		1.3061	思考・意見・疑い	**検討(9)、**	**9**
		1.3081	方法	**手段(12)、**	**12**
		1.3082	制度・慣例	**制度(17)、**	**17**
	待遇	1.3600	支配・政治	**改革(8)、**	**8**
		1.3620	運営	**運営(5)、管理(5)、**	**10**
		1.3630	人事	**選挙(12)、**	**12**
		1.3650	救護・救援	**ＯＤＡ(5)、援助(7)、**	**12**
		1.3682	賞罰	**促進(7)、**	**7**
抽象的関係	作用	1.1510	動き	運転(6)、	6
		1.1523	巡回	サイクル(6)、	6
		1.1550	合体・出会い・集合など	結合(5)、	5
		1.1560	接近・接触・隔離	接触(9)、	9
		1.1581	伸縮	拡大(15)、	15
		1.1583	進歩・衰退	強化(7)、向上(5)、低下(5)、	17
	事柄	1.1040	本体・代理	他(7)、	7
	時間	1.1600	時間	時間(5)、	5
		1.1650	順序	前半(6)、	6
		1.1670	時間的前後	前後(10)、	10
	存在	1.1220	成立	形成(6)、	6
	類	1.1112	因果	影響(6)、効果(5)、	11
		1.1131	連絡・所属	連鎖(6)、	6
合計					493

「N を通じて」だけに前接する名詞には 2 字漢語が多い(63 個中 36 個が 2 字漢語サ変可能名詞)ことが分かる。和語に比べ、漢語は書き言葉的文章により多く現れるため、「N を通じて」は「N を通して」より書き言葉的であることと伺える。また、N の意味分類を見ると、「人間活動主体」は「政府、JICA、国連」のような公的機関(の擬人化)を指しており、また、「人間活動―精神及び行為」は「公的な活動」を表す名詞が多いことが分かる。よって、「N を通じて」は 4. 1. で見る「N を通して」と比べれば、どちらかというと公的な活動について述べる傾向にあると言えよう。

収集した用例全体の N の語種も確認したところ、前に来る漢語の割合は「通して」は 56.8%で、「通じて」は 71.7%であったのに対し、前に来る和語の割合は「通して」は 28.1%

で、「通じて」は 11.9%であった。全用例においても、「を通じて」は「を通して」より漢語との共起例が多句、より書き言葉的であることが分かる。

表４　前接名詞の語種

	N を通して		**N を通じて**	
行ラベル	N の語種別個数	割合	N の語種別個数	割合
外来語	535	12.1%	604	12.0%
漢語	**2510**	**56.8%**	**3595**	**71.7%**
記号	7	0.2%	30	0.6%
固有語	88	2.0%	161	3.2%
混種語	36	0.8%	29	0.6%
和語	**1240**	**28.1%**	**596**	**11.9%**
総計	4416	100.0%	5015	100.0%

4．3.「N を通じて」と「N を通して」両方に前接する名詞

N が 5 回以上出現した用例の中で、「N を通じて」と「N を通して」両者に共通するものは 79 語あった。以下の通りである。

インターネット、こと、コミュニケーション、スポーツ、テレビ、ネットワーク、ふれあい、ブログ、プロセス、マスコミ、メディア、もの、ラジオ、ルート、一生、運動、音楽、過程、会、会社、会話、絵、学習、活動、管、関係、期、期間、機会、機関、教育、局、訓練、経験、研究、言葉、交換、交渉、交流、行為、行動、作業、作品、作用、参加、仕事、四季、市場、指導、事業、事件、事例、時代、実験、実践、取り組み、取材、手、人、世界、世紀、生涯、生活、全体、全編、組織、体、体験、対話、代、調査、電話、日、年、年間、分析、本、遊び、歴史

上記共通する N の 79 語中、出現回数の開きが 5 回以上あった項目を図 2 で示す。

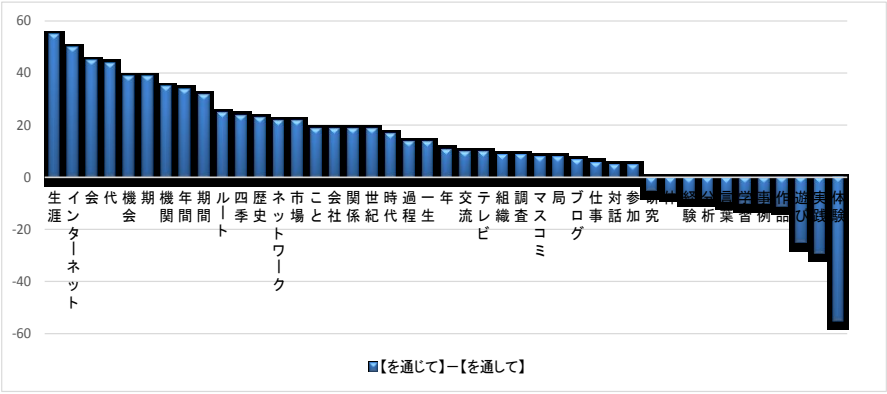


図 2「N を通じて」と「N を通して」に共通する N の比較（出現頻度の開き≧5 回）

図 2 が示すように、「N を通じて」が多い項目は「生涯～参加」の 32 語で、「N を通して」が多い項目は「体験～研究」の 11 語である。「N を通じて」の前接名詞 N は「インターネット、ネットワーク、テレビ、マスコミ、ブログ」など、情報伝達の媒体を意味するもの、

「生涯、期、年間、期間、四季、歴史、世紀、時代、一生、年」など期間を意味するものが多いことが分かる。それに対し、「N を通して」の前接名詞Nは「体験、実践、遊び、作品、事例、学習」などが多く現れ、主体の積極的な持続行為により、何かを実現し、あるいは得られることを表していることが分かる。

５．終わりに

本稿では、BCCWJ を使って「N を通じて」と「N を通して」に前接する名詞Nの異同を手掛かりに、「N を通じて」と「N を通して」の違いを探った。その結果、以下のことが明らかになった。

第一、傾向として、先行研究でも指摘されているように、「N を通じて」「N を通して」の両方が書き言葉的ではあるが､「N を通じて」がより書き言葉的であることが検証できた。第二、「N を通じて」が情報伝達の媒体としての用例、期間を表す用例が多く、それに対し、「N を通して」は主体の積極的な持続行為により、何かを実現し、あるいは得られることを表す用例が多いことが分かった。第三、「N を通じて」は公的な活動について述べる傾向にあり、「N を通して」はどちらかというと私的な、具体的な事柄について述べる傾向にあることが分かった。この第三の違いについてはコーパスのレジスターに影響されている可能性もあり、今後、別のデータで検証する必要がある。

「通す」「通じる」の本動詞の用法を削除せずに分析した。先に削除したうえで見るべきだった。また、両者の具体的な意味用法、および後続述語を検討する必要もある。今後の課題としたい。

参考文献

秋月高太郎（2007）「ブログに書かれること、書かれないこと―ブログ「会話」の含意―」、『日本語学 特集ブログのことば』、26：4、 pp.46-56.

庵功雄、高梨志乃、中西久美子、山田敏弘(2001)『中上級を教える人のための日本語文法ハンドブック』、スリーエーネットワーク.

泉原省二（2007）『日本語類義表現使い分け辞典』、研究社.

グループ・ジャマシイ(1998)『教師と学習者のための日本語文型辞典』、くろしお出版.

国立国語研究所(2004)『分類語彙表―増補改訂版』、大日本図書.

花薗悟(2004)「『N を通して』と『N を通じて』」、『留学生日本語教育センター論集』、30、pp.17-31、 東京外国語大学
（http://repository.tufs.ac.jp/bitstream/10108/20968/1/jlc030002.pdf よりダウンロード可能）

村田年(1993)「中・上級表現文型研究『…を通して/…を通じて』」、『日本語と日本語教育』、22、 pp.1-32、 慶応義塾大学

森田良行、松木正恵(1989)『日本語表現分型：用例中心・複合辞の意味と用法』、アルク.

用例出典及び検索ツール

『現代日本語書き言葉均衡コーパス』(BCCWJ) （2011） 国立国語研究所

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

ポスター発表（2） グループB

3月7日（金） 14:00～15:00

感情動詞におけるニ格とデ格の交替について

張麗（大東文化大学）

Alternation of Complements Ending in NI and DE in Emotion Verbs

Zhang Li(Daito Bunka University)

1. はじめに

格助詞の交替現象は、学習者にとっても悩みの種で、これまでも誤用例の分析などでも頻繁にとりあげられてきた。しかし、その根本的なメカニズムの解明にまでは至っていない。ニ格とデ格の交替は錯綜した現象がみられることから、統一的な説明が困難だからである。使用制限、使用基準をめぐっては、前接の名詞成分の特徴を視野に入れる必要がある。本研究では感情動詞（感性動詞、情態動詞ともいう）に前接する格助詞の異同について、交替現象に注目しながら考察を進めることにする。感情動詞に注目したのは、感情形容詞にも共通する心理的要素があり、交替現象が見えやすいという背景から、また、使用頻度も高いことから、考察の対象とした。ここで、感情動詞の定義を筆者なりに述べておくと、人間の心理、感情にかかわる動詞としてとらえ、思考動詞などは対象外とする。

2. 先行研究

2.1 ニ格・デ格についての先行研究

国語学、日本語学からニ格・デ格の分類について、国立国語研究所（1951）、言語学研究会（1983）、益岡・田窪（1987）が挙げられる。第二言語習得から、迫田（1998）などの研究がある。認知言語学から、山梨（1995）、菅井（1997）（2000）、森山（2005）が挙げられる。

2.2 ニ格、デ格の交替についての先行研究

格の交替現象が起こるデ格とニ格を中心とする先行研究は川野（2012）の壁塗り構文についての研究である。

ニ格をとる感情動詞について、先行研究では「ニ」と「デ」の言い換えができる動詞と「ニ」と「デ」の言い換えができない動詞があると指摘している。言語学研究会（1983）では、「ほれる」「あこがれる」、三原（2000）では、「呆れる」「飽きる」などが、「ニ」を「デ」に言い換えることができない動詞として挙げられている。

村上（2010）では感情動詞の補語を表す「ニ」がどのような場合に「デ」に言いかえることができるかについて、考察を行った。「驚く」「びっくりする」「おびえる」では、名詞句が「外的原因」の場合は、「デ」を使うことができる。「困る」「苦しむ」「悩む」「迷う」では、ニ格名詞句が無いことによって感情が動いたことを表す「欠乏」の場合は、「デ」を使うことができないが、「やけどのあとで悩んでいる」のように、名詞句の存在が感情を引き起こすことを表す「存在」の場合は、「デ」が使うことができることを述べた。

今までニ格、デ格についての研究は主に言語学においてニ格、デ格のそれぞれの分類を中心としている。第二言語習得の分野において場所を表すニ格とデ格の混同や、ニ格の過剰使用についての研究もなされている。ニ格とデ格の交替については、主に壁塗り構文に焦点を置いている。感情動詞を中心にしたニ格の研究は管見の限り少ない。先行研究では、

ニ格をとる感情動詞はニ格とデ格の言い換えが可能だと述べているが、実際の言語使用では、それぞれの使用率がどのぐらいなのかはっきり分からない。先行研究では七つの感情動詞について、総体的に考察したが、それぞれについての考察はまだ不十分だと考える。

3. 研究の目的

本研究では、「驚く」「おびえる」「苦しむ」「困る」「悩む」「びっくりする」「迷う」七つの代表的な感情動詞を取り上げ、以下の二つの問題を明らかにすることを目的とする。

① 七つの感情動詞それぞれニ格、デ格の使用率はどれぐらいなのか

② 感情動詞「悩む」におけるニ格とデ格の交替条件は何なのか。

４．調査方法

本研究では、『現代書き言葉均衡コーパス（中納言）』を通して、短単位検索で「驚く」「怯える」「苦しむ」「困る」「悩む」「びっくりする」「迷う」について調べる。具体的な調査方法は以下のようである。

語彙素は「驚く」、前方共起 1 は[にでを]で検索する

ほかの動詞も同じ方法で検索する。

５．調査の結果

（表 1）感情動詞におけるデ格とニ格の使用率

感情動詞	デ	ニ
驚く （使用率%）	8／500／1926 （1.6）	245／500／1926 （49.0）
怯える （使用率%）	14／469 （3.0）	330／469 （70.4）
苦しむ （使用率%）	64／500／1256 （12.8）	311／500／1256 （62.2）
困る （使用率%）	57／500／1834 （11.4）	203／500／1834 （40.6）
悩む （使用率%）	123／500／1966 （24.6）	170／500／1966 （34.0）
びっくり （使用率%）	34／778 （4.4）	182／778 （23.4）
迷う （使用率%）	67／500／825 （13.4）	273／500／825 （54.6）

「驚く」が 1926 例ヒットし、その中の 500 件が表示された。「名詞＋ニ＋驚く」は 245 例で、使用率は 49%で、「名詞＋デ＋驚く」は 8 例あり、使用率は 1.6%である。

「怯える」が 469 例ヒットした。「名詞＋ニ＋怯える」は 330 例で、使用率は 70.4%で、「名詞＋デ＋怯える」は 14 例あり、使用率は 3.0%である。

「苦しむ」が 1256 例ヒットし、その中の 500 件が表示された。「名詞＋ニ＋苦しむ」は 311 例で、使用率は 62.2%で、「名詞＋デ＋苦しむ」は 64 例あり、使用率は 12.8%である。

「困る」が 1834 例ヒットし、その中の 500 件が表示された。「名詞＋ニ＋困る」は 203 例で、使用率は 40.6%で、「名詞＋デ＋困る」は 57 例あり、使用率は 11.4%である。

「悩む」が 1966 例ヒットし、その中の 500 件が表示された。「名詞＋ニ＋悩む」は 170 例で、使用率は 34.0%で、「名詞＋デ＋悩む」は 123 例あり、使用率は 24.6%である。
「びっくり」が 778 例ヒットした。「名詞＋ニ＋びっくり」は 182 例で、使用率は 23.4%で、「名詞＋デ＋びっくり」は 34 例あり、使用率は 4.4%である。
「迷う」が 825 例ヒットし、その中の 500 件が表示された。「名詞＋ニ＋迷う」は 273 例で、使用率は 54.6%で、「名詞＋デ＋迷う」は 67 例あり、使用率は 13.4%である。

６．考察

表 1 から分かるように、七つの感情動詞の中で、デ格の使用率が一番高いのは「悩む」であり、デ格の使用率は 24.6%である。本研究では「悩む」に注目して、デ格とニ格の交替条件について考察する。前接の名詞の種類に焦点を置き、考察する。名詞には普通名詞と抽象名詞に分けている。抽象名詞には以下の四つのパターンがある。

A パターン人間の生活に関する類
B パターン変化を表す類、
C パターン対立関係・行為・事件などを表す類
D パターン感情などに関する類がある。

6.1 前接の名詞が A パターン人間の生活に関する類の場合

日常生活を表す名詞の抽象度が低く、デ格しか使えなく、ニ格とデ格の交替が難しい。

(1) Au の機種で悩んでいるんですけど…　（Yahoo!知恵袋　2005）
(2) バックの色で悩んでいます。　（Yahoo!知恵袋　2005）
(3) 十一月挙式予定なのですが引き出物で悩んでいます。　（Yahoo!知恵袋　2005）
(4) ヘアスタイルで悩んでます。　（Yahoo!知恵袋　2005）

同じ人間の生活を描く名詞が、少し抽象度が高くなるにつれ、デ格とニ格の交替ができるようになった。

（5a）友情や恋、進路などで悩み傷づく子どもたちを、そっと見守るクロ。
（『サライ』小学館 2004）
（5b）進路に悩んで、私は大学受験のため倉レの寮を出て…
（中元輝夫『雑草の記』光陽出版社 2002）
（6a）子育てで悩んでいます。　（Yahoo!知恵袋 2005）
（6b）子育てに悩む親に対して情報提供をしたり…
（衣川光正『ライブ解説職務論文文例集』公人社 2003）

前接の名詞は病気に関する類の場合、デ格とニ格の交替ができる。

（7a）こんな症状で悩んでいる人いませんか。　（Yahoo!知恵袋 2005）
（7b）戦争の悲惨さから夢がみられなくなる症状に悩む。
（福本修ほか編『精神医学の名著 50』平凡社 2003）
（8a）胃腸の丈夫な人に肩こりで悩む人はあまりいません。
（王進利『中国式目・肩・腰のツボ』大泉書店 2002）
（8b）そして、頭痛もちになったり、肩こりに悩む。
（西田達弘ほか『鼻はこれで直せる』1980）

前接の名詞は病気に関する類では、ニ格を取る用例は以下のようである。
アレルギーに、アレルギー疾患に、下痢に、高次脳機能障害に、ひざなどの痛みに、慢性病に、胃病に、便秘に、不妊症に、鬱病に、難症・難病に、大変な禅病に、更年期障害に慢性疾患に、下腹部痛に、円形脱毛症に、冷え性に、腰痛に、生理痛に、不眠に、肥満に、

薄毛に、髪の脱色に、夜泣きに、家庭内暴力に、船酔いになど。

デ格を取る用例は以下のようである。

にきびで、早期の白髪で、薄毛で、毛むくじゃら、手の爪が巻き爪で、同じ症状で、痴呆症で、対人恐怖症で、便秘で、痔で、うつ病状で、アレルギー症状で、摂食障害で、皮膚欠損で、胃下垂でなど

以上Aパターンのような、日常生活を描く抽象度の低い名詞の場合、デ格しか使えなく、デ格とニ格の交替が難しいが、人間の生活を表す名詞や病名などの固有名詞は相対的に抽象度が高くなり、デ格とニ格の交替ができるようになった。

6.2 前接の名詞が Bパターン変化を表す場合

名詞の抽象度が高いため、ニ格しか使えなく、ニ格とデ格の交替が難しい。

関係

（9）対象＝夫婦関係に悩む女性各日二十人[抽選]
（『市政だより青葉区版』宮城県仙台市青葉区 2008）

（10）上司や同僚との人間関係に悩んでいる人もいるでしょう。
（上原泰次『日々、夢に向かって邁進！』中央通信社　2003）

（11）家族がネックになったり、家族との関係に悩んだり…
（東京都豊島区立男女平等推進センター編『男が語る一家族・家庭』1994）

（12）人に接触する機会が多くて対人関係に悩む職業の方やセールスマンに多い傾向があります。
（高野正博『大腸の病気』保健同人社 2002）

不足・欠乏

（13）現在百九十 MB の RAN でリソース不足に悩んでいます。　（Yahoo!知恵袋 2005）

（14）有効需要不足に悩んでいた日本経済にとって…
（高橋乗宣『「経済白書」で読む奇跡の 50 年』1995）

（15）急速な少子化の進展で研究開発人材の不足に悩むみんなが…
（『au・au』　マガジンハウス　2003）

（16）住宅不足に悩む庶民へ住宅をという名目をかかげて、東知事の下で発足した。
（伊東光晴『現代経済の変貌』岩波書店 1997）

（17）急速な少子化の進展で研究開発人材の不足に悩むのは確実。
（高藤康浩『勝つ工場』日本経済新聞社 2005）

（18）イタリアでも税率を下げたくらいなのだが、歳入不足に悩むイギリス政府にとって
（出版太郎『朱筆』みすず書房 1990）

（19）私たちが運動を進める過程で、同じようにマンションの瑕疵・欠陥に悩んでおられる居住者の方がたくさん居られることも知りました。
（相羽宏紀『欠陥マンション改善の戦い』あけび書房 2000）

（20）曾て、塩分の欠乏に悩んで…
（井上靖ほか編『昭和文学全集』小学館 1989）

被害

（21）知的財産権侵害に悩む各種業界団体らにより設立された任意団体。
（警察庁『警察白書』行政 2003）

（22）空き巣の被害に悩む地域の解決策…　（『Breath』ソニー・マガジンズ 2003）

経済に関すること

（23）千九百七十年代には財政赤字、貿易赤字、インフレーションに悩んだ。

（風早正宏『ゼミナール経営管理入門』日本経済新聞社 2004）

（24）各県ともいま地方活性化に悩んでいるが…

（松谷みよ子『現代民話考』立風書房 1994）

（25）円高や原材料価格の高騰に悩んでいるのは、自動車業界だけではない。

（Yahoo!ブログ 2008）

過疎

（26）過疎に悩む山村においては、林業が活性化することは重要ではあるが…

（西野寿章『山村地域開発論』原書房 2004）

対人関係

（27）家庭の不和になやまなければならない自らの宿命を転換することです。

（池田大作『新・人間革命』聖教新聞社 1998）

存在

（28）怪物自身も、知性をもつがゆえに自分自身の存在に悩み、苦しむ。

（カレル・チャペック著小林恭ニほか『山椒魚戦争』1994）

矛盾、ジレンマ

（29）いつもそのジレンマに悩んで押し売りの前で首をひねるのだが…

（藤沢周平『竹光始末』新潮社 2002）

6.3 前接の名詞はCパターン対立関係、行為、事件を表す類

名詞の抽象度が高いことから、ニ格しか使えなく、ニ格とデ格の交替が難しい。

（30）もともとこの制度は、地域の迷惑行為に悩む住民が…

（山口ニ郎『ブレア時代のイギリス』岩波書店 2005）

6.4 前接の名詞はDパターン心理、感情、感覚に関する類

同じく名詞の抽象度が高いため、ニ格とデ格の交替が難しい。

（31）好きだから不安になる彼の本心に悩むあなたに 5 つのメッセージ…

（近藤裕『「運命の人」に出会う 25 のレッセン』2001）

（32）人は深い孤独に悩む。

（吉村英夫ほか『ほろっと本音キラッと青春』アールズ出版 2002）

（33）仕事全体が見えぬとき、無力感に悩み…

（松山巖『手の孤独、手の力』中央公論新社 2001）

（34）人は自分の非行を誰一人知る者がいなくても罪の意識に悩む。

（長野晃子『日本人はなぜいつも「申し訳ない」と思うのか』草思社 2003）

6.5 名詞の抽象度が高いが、ニ格とデ格と交替ができる場合

問題

（35）子供の非行、不良行為その他の問題で悩む保護者等から相談を受けて…

（『青少年白書』総務庁青少年対策本部　大蔵省印刷局）

（36）深刻なごみ問題に悩む私たちの時代に…

（コリーヌ・ブレ『山猫の愛のように』読売新聞社 1997）

ストレス

（37）思い通りにいかないというストレスで悩んでいるときは頭が空めぐりしたり、血液の循環が悪くなったりします。　（須藤和廣『めざせ！健康美容師』知道出版 2004）

（38）健康な人がストレスに悩む場合にもきわめて有効なことがわかっています。

（関谷透『ストレス・ウォーズ』現代書林 1992）

不信

（39）男性不信で悩んでいます。　（Yahoo!知恵袋　2005）

（40）不信に悩むものは傷つき狂った動物のように…

（吉田秀和『吉田秀和全集』白水社 2004）

抽象度の高い名詞が普通ニ格で表す。しかし、問題、ストレス、不信のような名詞の抽象度が高いが、ニ格だけではなく、デ格を使う用例もあった。それはニ格とデ格の交替ができるといえるだろう。

7．まとめと今後の課題

本研究では、名詞の抽象度の観点から、感情動詞におけるニ格とデ格の交替について考察した。感情動詞に接続するニ格は場合によって、デ格と交替できる。本研究は感情動詞「悩む」について、考察した結果、日常生活を描く抽象度の低い名詞と接続する場合、デ格しか使えなく、ニ格が使いにくく、ニ格とデ格の交替が難しいということが分かった。もう少し抽象度が高くなった人間の生活を表す名詞や病名を表す名詞の場合、ニ格とデ格の交替が可能だと考える。抽象度が高い名詞と接続する場合、デ格の使用が限られている。ニ格とデ格の交替が不可能という結論が得られた。しかし、一見抽象度が高い名詞でも、デ格の使用も可能の場合があるため、抽象度が高い名詞にはデ格が使えないとは言い切れないと思われる。ニ格とデ格の交替についての研究を深めたいなら、名詞の抽象度についての研究もさらに深まる必要がある。それを今後の課題とする。

参考文献

言語学研究会（1983）『日本語文法・連語論』むぎ書房

川野靖子（2012）「現代日本語の動詞「詰める」「覆う」の分析―格体制の交替の観点から―」『埼玉大学紀要（教養学部）第 48 巻第 2 号』

国立国語研究所報告 3（1951）『現代語の助詞・助動詞－用法と実例－』秀英出版　pp135－151

迫田久美子（1998）「誤用を生み出す学習者のストラテジー場所を表す格助詞「に」と「で」の使い分け―」『平成 10 年度日本語教育学会秋季大会予稿集』pp128－134

菅井三実（1997）「格助詞「で」の意味特性に関する一考察」『名古屋大学文学部研究論集 文学 43』pp23－40

菅井三実（2000）「格助詞「に」の意味特性に関する覚書」『兵庫教育大学研究紀要　第 2 分冊　言語系教育・社会系教育・芸術系教育 20』pp13－24

張善実（2008）「ニ格とヲ格の使い分けについて」日本語教育学講座研究会

益岡隆志、田窪行則（1987）『格助詞　日本語文法　セルフ・マスターシリーズ 3』くろしお出版　pp4－5

三原健一（2000）「日本語心理動詞の適切な扱いに向けて」『日本語科学』8

村上佳恵（2010）「感情動詞の補語についての一考察―「ニ」と「デ」について」『学習院大学国語国文学会誌』

森山新（2005）『認知言語学的観点を取り入れた格助詞の意味のネットワーク構造解明とその習得過程（平成 14 年度～平成 16 年度科学研究費補助金研究　基盤研究（C）（2）課題番号 14510615　研究代表者：森山新）成果報告書』

山梨正明（1995）『認知文法論』ひつじ書房

『現代日本語書き言葉均衡コーパス』

教科書コーパスにおける教科別特徴語

立山 智絵（東京女子大学大学院人間科学研究科）

Characteristic Words of Subjects in Textbook Corpus

Tomoe Tateyama (Graduate School of Human Sciences, Tokyo Woman's Christian University)

1. はじめに

日本語を母語としない児童・生徒に関する教育研究分野等では、日常会話を十分にこなせる力があっても教科学習を困難なものにする原因として、学習言語の習得が密接に関連しているということが指摘されている。本調査は、先行研究で指摘されているように、同じ語であっても教科によって使われ方が異なっていたり、使用法が特殊であったりする語についての分析を行い、教科による特徴語についての研究を精緻化させることを目的とする。

また、これまで、ある特定の教科や校種の学習言語に関する研究は盛んに行われていたが、教科や校種の枠を越えた俯瞰的な研究は未だ十分ではないと考える。これを可能にする手段として、コーパスの活用が考えられる。そこで本調査では、「現代日本語書き言葉均衡コーパス」を用いて、教科によって語の意味やコロケーションがどのように異なるのかを明らかにし、分析を行いたいと考える。

2. 調査概要

2.1 調査対象

「現代日本語書き言葉均衡コーパス」(以下、BCCWJ)に収録されている「教科書コーパス」を対象に、「中納言」を用いて調査を行う。

対象とする語については、バトラー(2010)による「学習語彙リスト」を参照する。このリストは、「教科書コーパス」の小中学校の教科書データを対象とし、一般語ではないこと、多くの教科をまたいで使用されているもの、頻度が高いものといった基準から選出された約 1500 語から成る学習語彙リストである。

リストは頻度別、重要度別に並べられており、筆者はこれまで、頻度と重要度が共に高いものから、動詞や名詞を中心に調査を行ってきた。これまでの調査から、リスト内の語をいくつかの傾向に分類できるのではないかと考えており、現在も調査を進めている。例を挙げると、使用されている意味や用法が、教科や校種によって異なっている「当てる」「分ける」などや、使用されている意味はほぼ似通っているが、コロケーションが教科や校種で異なる「与える」「通す」などに分類できると考えている。

本調査では、校種によって意味が偏っていることはほとんどないが、教科によって使用されている意味に特徴があるといえる、「伝わる」「伝える」「結ぶ」[1]を取り上げる。

---

[1] バトラー(2010)による「学習語彙リスト」では、「伝わる」「結ぶ」は「頻度最高・重要度高」の項目に分類されているが「伝える」の記載はない。しかし、自動詞「伝わる」の対

2.2　調査方法

「教科書コーパス」の全学年全校種を対象に各語を語彙素で検索し、比較的語釈が細かく分類されている『明鏡国語辞典』[2]にしたがって目視で意味の分類を行い、教科間による各語の意味やコロケーションの差異を検証する。

3.　調査結果・考察

3.1　「伝わる」「伝える」

『明鏡国語辞典』による語釈は以下の通りである[3]。

つたわ・る【伝わる】〔自五〕
①電気・熱・音や刺激・情報などがある道筋を通る。また、そのようにしてある所に達する。「階下の音が壁を―・って二階に届く」「風に乗って花の匂いが―・ってくる」「電気が手に―・ってびりっとくる」「刺激が神経細胞を通って脳に―」
②印象・感動などが情報としてあるところに達する。「屈辱の思いが胸に―・ってくる」「文面から決意のほどがひしひしと―・ってくる」
③話などが相手に通じてよく理解される。「話[こちらの意向]が先方に―」「かたことの英語[支離滅裂な発言]では話が―・らない」
④人などを仲立ちとして話が広まる。「よくないうわさが町中に―」
⑤(代々受け継がれてきて)後代・後世に残る。「わが家には先祖代々古伊万里(こいまり)の茶器が―」「写真術の創始者として後世にその名が―」「文献にあるだけで作品は現代に―・らない」
⑥遠く離れたところから文物が入ってくる。伝来する。「鉄砲はポルトガルから―・った」「大陸から梅が―・った」
⑦「伝う」のややくだけた言い方。⇒伝う「賊が屋根を―・って逃げる」「猿が木を―・って飛び移る」

なお、教科による違いを分析するため①の語釈を以下の 3 つに分類した。

①-1-1 電気・熱・音や刺激などがある道筋を通る。またそのようにしてある所に達する。
①-1-2 液体が沿って流れる。
①-2 情報などがある道筋を通る。またそのようにしてある所に達する。

つた・える【伝える】〔他下一〕
①あるものが仲立ちとなって、電気・熱・音や情報などを他方に移す。「銅線が電気を―」「空気が音を―」「電波が情報を―」「風が春の息吹を―」
②ことばなどである内容をもった事柄を知らせる。「記者が被害の状況を―」「意向[お礼のことば]を―」「伝言を―」「映像で―」「君の気持ちは私から彼女に―・えておく」「皆さんに宜(よろ)しくお―・え下さい」

応として、他動詞「伝える」もともに調査を行った。
[2] 北原保雄編(2011)『明鏡国語辞典 第二版』電子辞書版を参照する。
[3] 「語法」「注意」等については省略。

③(代々受け継いできて)後代・後世に残す。「弟子に秘伝を—」「後輩に技術を—」「伝統を未来に—」「事件を記してその詳細を後世に—」
④≪「…と—」「…と—・えられる」の形で≫それが伝承や伝説に基づく情報であることを表す。「空海が創建したと—」「この曲はモーツァルト作曲と—・えられている」
⑤遠く離れたところから文物をもってくる。もたらす。「ザビエルは日本にキリスト教を—・えた」「陶磁器[印刷術]を—」

「伝わる」は 111 例[4]、「伝える」は 217 例抽出され、教科別に意味を分類すると以下のような結果となった(表 1)(表 2)。

表1.「伝わる」教科別分類

	①-1-1	①-1-2	①-2	②	③	⑤	⑥	計
外国語					1			1
技術家庭			1					1
芸術			2	1		4		7
国語	1		7	9	2	2	4	25
社会	2		2		2	3	11	20
情報	1		3	1				5
数学	1							1
理科	41	5				5		51
計	46	5	15	11	5	14	15	111

表2.「伝える」教科別分類

	①	②	③	④	⑤	計
外国語		12	1			13
技術家庭		6	3			9
芸術	1	7	10	7	1	26
国語		39	3	3	1	46
社会		26	6	1	26	59
情報		13				13
数学		2			1	3
保健体育		5				5
理科	23	5	15			43
計	24	115	38	11	29	217

まず、表 1 の「伝わる」について意味別、教科別に見ると、理科における①-1-1 の「電気・熱・音や刺激などがある道筋を通る。またそのようにしてある所に達する。」としての使用が多く、「音」や「地震波」、「P 波」、「S 波」などが伝わる、という例が見受けられる。小学校 5 年の理科では、①-1-2 の「液は，ガラスぼうを伝わらせて入れる。」というような、液体が何かに沿って流れるという意味での使用例があり、これは理科(小学校 5 年に 2 例、高校に 3 例)にのみ出現する例である。

続いて表 2 の「伝える」では、意味別に見ると②の「ことばなどである内容をもった事柄を知らせる。」という意味での使用が 115 例(53%)と多い。教科別に見ると社会や国語、理科での使用が目立つ。この 3 教科で比較分析し残差から考えると、特に理科の①、国語の②、社会の⑤の使用が有意に多く、理科の②が有意に少ないことが分かる(χ^2=134.553, df=8,

[4] 「伝わる」では、『明鏡国語辞典』での④、⑦の使用例は見られなかった。

p<.001)。理科の①は「伝わる」と同様に電気や熱を伝えるという例が多く、理科での特徴といえるが、②では、情報や自分の気持ち、意見などを相手に伝えるという使用例が国語に限らず出現しており、コロケーションも似通っている。社会では主に、世界史や倫理の教科書で⑤の「稲作が九州北部に伝えられ…」(中 2 歴史)といったような、文物を持ってくる、もたらすという例が多く使用されている。

また、「伝わる」「伝える」に限ったことではないが、「仏教の伝播仏教はブッダガヤーから発して，北方へは，中央アジアの砂漠を越え，中国，朝鮮，日本へ伝えられた。南方へは，西インドからスリランカへ伝わり，スリランカから各地に伝わった。」(倫理/高校)のように、主語と述語が離れている例や、連体修飾節が長く文が複雑なものなど、先行研究で指摘されているような読解を困難にする可能性のある文が多々見受けられる。

3.2 「結ぶ」

続いて、「結ぶ」について述べる。『明鏡国語辞典』での語釈は以下の通りである。

むす・ぶ【結ぶ】
㊀〔他五〕
①糸・ひも・帯などをからみ合わせて離れないようにする。また、そのようにしてひもなどをある形に作る。「二本のひもを―」「靴ひも[風呂敷]を―」「草を―・んで道しるべとする」「丸帯をお太鼓に―」
②ばらばらになったり離れ離れになったりしないようにひもなどでくくりつける。結びつける。「髪をリボンで―」「荷物をひもで―」「二人は運命の赤い糸で―・ばれている」
③ひもなどを物や体の周りに巻きつける。結びつける。「神木におみくじを―」「頭の後ろで鉢巻きを―」
④「結ぶ」②ことによって、建造物などを作り出す。「草庵(そうあん)を―」「垣根を―」
⑤口・手を固く閉じる。「口を固く―」「口をへの字に―」「―・んでひらいて 手をうって―・んで〈むすんでひらいて〉」
⑥仏教で、手の指で悟りなどを象徴的に表す形を作る。「手に印(いん)を―」
⑦連絡や通信を可能にするために、複数のものをつなぐ。「○○大橋が両県を―」「ネットワークが事務所と事務所を―」「遺留品が容疑者と事件を―」
⑧「結ぶ」⑦ことによって、連絡のつく環境を作り出す。張りめぐらす。築く。「全市内に緊急連絡網を―」「鉄壁の包囲網を―」
⑨約束して強い関係を作る。「両者は一億円で売買契約を―・んだ」「隣国と通商条約を―」
⑩別々の体や心をつないで離れないようにする。結び合わせる。「愛情が二人の心をしっかりと―」「二人は強固な信頼感で―・ばれている」「めでたく―・ばれる(=愛が成就(じょうじゅ)する・結婚する)」
⑪米飯を握り飯にする。握る。「赤飯を俵形に―」
⑫形のあるもの(特に、粒状の物)を生じさせる。「空気中の水分が露を―」「ミカンが実を―」「長年の努力が実を―」「ヒマラヤ登頂に若き日の夢を―」「網膜に像を―」
⑬話・文章などをしめくくる。「感謝のことばで挨拶(あいさつ)を―」「『敬具』と記して手紙を―」
⑭係り結びで、係助詞に呼応した形で文を終止させる。「こそ」を受けて文末の活用語を已然形にする類。

㈡〔自五〕
形あるもの(特に、粒状の物)が生じる。「桑の実が―」「ハスの葉に露が―」

「結ぶ」は 150 例[5]あり、教科別に意味を分類すると以下のような結果となった(表 3) [6]。

表3.「結ぶ」教科別分類

	①	⑦	⑧	⑨	⑩	⑫	⑬	その他	計
外国語		2							2
技術家庭	2			1	2				5
芸術		2			1				3
国語	1			2	1		1	2	7
社会		36	1	59	1				97
情報			2						2
数学	1	13							14
理科		16	2			2			20
計	4	69	5	62	5	2	1	2	150

「結ぶ」では社会が 97 例あり、64.7%を占めている。意味別に見ると、⑦「連絡や通信を可能にするために、複数のものをつなぐ。」、⑨「約束して強い関係を作る。」での使用が多く、⑨の社会では、条約や同盟を結ぶという例が多数見られる。

社会では、「商工業者が株仲間を結ぶ」(中 2 歴史)のように、社会での教科特有と考えられる名詞と共起する例も見受けられる。他にも、「アテネは(中略)ついにペルシアと結んだスパルタに敗れた。」(世界史)、「大和国家は，百済や，小国が分立していた加羅（任那）地方の国々と結んで，高句麗や新羅と戦いました。」(中 2 歴史)といったような、特徴的な使用例も認められる。

「結ぶ」は特に社会で特徴的な使用が見られる語と思われるが[7]、他にも数学では、「次の四角形をつくるには，どの４つの点を結べばよいでしょうか。」(小 5 算数)のように点と点を結ぶという例や、国語では「本来結ぶべき部分が、下の接続助詞に接続させるための活用形などになって、結びが消える。」(高校国語)のように、文の結び、という教科特有の例も見受けられる。

3.3　その他の語

最後に、その他の分類について代表的な語を取り上げる。使用されている意味や用法が、教科や校種によって異なっている「分ける」と、使用されている意味はほぼ似通っているが、コロケーションが教科や校種で異なる「通す」について簡単に述べる。

「分ける」は、小学校の算数では「１人に３こずつ分けると，何人に分けられますか」(小 3 数学)のように、分配するという意味での使用が多数見受けられるが、中学・高校の数学

5 「結ぶ」では、『明鏡国語辞典』での②～⑥、⑪、⑭の使用例は見られなかった。
6 「その他」に分類した国語の 2 例は、「むすぶ　手のひらで水をすくう」(高等学校／古典)という例と、小学校 4 年生で漢字の読み(「結ケツむすぶ」)として使用されているものである。
7 今回の調査では、BCCWJ の分類にしたがって社会科の教科書は「社会」として一括りに分類しているが、倫理や政治経済など、分野によっても使用法やコロケーションが異なる可能性は十分に考えられる。この点は今後の課題である。

では「分ける」という語の使用頻度そのものが少なく、小学校算数の特徴であると言える。理科では、「じしゃくにつくものと，つかないものとに，分けることができたかな」(小 3 理科)といったように、ある事柄を区別し分類する、という意味での使用が多くを占める。

「通す」は理科や技術家庭、社会に多く見られ、理科では、ある物質が電気を通す性質を持っているか否か、といった説明で使用されていることが多い。また、阿保(2013)は、図書館・書籍のレジスターとの比較から社会科教科書における教科特徴動詞の調査を行い、社会科教科書では「通ずる」が「…を通じて」という複合辞的用法で使用されることが多いことを明らかにしているが、「通す」も同様に、「…を通して」という複合辞的用法が多数使用されている。さらに、「…を通して」という複合辞的用法は社会に限らず、他の教科でも広く使用されており、「体験」や「調査」を通して、という例が多いことが分かる。他にも、「生き物は，食べ物をとおして，かかわり合って生きている」(小 6 理科)、「それぞれの時代を生きた人々の姿を通して、古代中国の人々の考え方・生き方を読み味わってみよう」(高校国語)といったように、何らかのものを媒介・仲立ちにして、という意味でも、教科によってコロケーションに違いが見られる。

これらの語のように、教科や校種によって意味や用法、コロケーションが異なる語は、教科学習・指導の上で考慮する必要があると思われる。

4. まとめ

教科によって使用されている意味に特徴があるといえる「伝わる」「伝える」「結ぶ」について簡単にまとめ、その他特徴の見られる語を取り上げた。これらの語を概観しただけでも、日常生活では馴染みが薄いと思われる特殊な例や、ある特定の教科や校種で見られる特徴的な例が、教科書には多数使用されていることが分かる。日本語を母語としない児童・生徒はもちろん、日本語母語の児童・生徒にとっても教科学習・指導の上で留意しなければならないと考える。

今後は、このような特徴的な語について他のレジスターと比較し、より分析を深めるとともに、学習言語をどのように提示し、指導していくかについて考察していく。

文 献

阿保きみ枝(2013)「社会科教科書における教科特徴動詞の用法―教科書コーパスと図書館コーパスの比較を通して―」『第 4 回コーパス日本語学ワークショップ予稿集』国立国語研究所 言語資源研究系・コーパス開発センター, pp. 339-344

近藤明日子(2011)「中学校・高校教科書の教科特徴語リストの作成」『言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用』特定領域研究「日本語コーパス」言語政策班報告書　国立国語研究所, pp. 145-152

北原保雄編(2011)『明鏡国語辞典 第二版』大修館書店　電子辞書版

バトラー後藤裕子(2010)「小中学生のための日本語学習リスト(試案)」『母語・継承語・バイリンガル教育(MHB)研究』6, pp. 42-58

バトラー後藤裕子(2011)『学習言語とは何か　教科学習に必要な言語能力』三省堂

『今昔物語集』における文体対立語
—巻 12 と巻 27 の語彙比較による—

田中 牧郎　（国立国語研究所言語資源研究系）†

The Synonyms Belonging to Opposing Styles in *Konjaku-Monogatari-shū*: Lexical Comparison of Volume 12 and 27

TANAKA Makiro (National Institute for Japanese Language and Linguistics)

１．はじめに

平安時代後期まで（11 世紀ごろまで）は和文の文章と漢文訓読文の文章は大きく離れていて、文体[1]による語彙の違いは大きなものだった。平安時代末期（12 世紀ごろ）には和漢混淆文が成立し、それまでは一つの文章の中で使われることのなかった語彙が、一つのところに流れ込み、一緒に使われるようになる。

和漢混淆文の中で、和の語彙と漢の語彙とがどのように共存しているかを明らかにすることは、語彙史上の重要な研究課題であり、これまで、西田（1978）、佐藤（1984）、峰岸（1986）、関（1993）、藤井（2003）、舩城（2011）などで研究されてきた。ところが、その多くは、和語か漢語か、和文語か漢文訓読語かなど、語彙の出自を整理することが中心で、それらがどのような体系をなして共存しているかは明確にされていなかった。

それを明確化する方法として、和漢混淆文における和文的な文章と漢文訓読文的な文章とを比較することが考えられる。その第一の試みとして、和文寄りの文体で書かれている『宇治拾遺物語』と、漢文訓読文よりの文体で書かれている『今昔物語集』の同文説話をパラレルコーパスにして語の対応を網羅的に調べる方法を提案した（田中・山元 2014）。本発表では、その第二の試みとして、『今昔物語集』における和文の巻と漢文訓読文の巻の語彙を比較する方法を提案し、その方法で見えてくる語の対立について考える。

２．『今昔物語集』の語彙調査

『今昔物語集』（以下、『今昔』と略称）は、12 世紀前半に書かれた、日本語史上のはじめての本格的和漢混淆文の説話集である。おおまかには、前半の巻は漢文訓読文、後半の巻は和文で書かれている。国立国語研究所のプロジェクトでは、『日本語歴史コーパス』に組み込むために、『今昔』のコーパス化を進めている[2]。現在は試作段階であるが、巻 12 と

† mtanaka@ninjal.ac.jp

[1] 本稿の「文体」の定義は、「表現の媒体（音声・文字）、ジャンル、目的・意図、場面・状況などによって、言語が異なった姿を呈する現象。また、そうした観点から見た時の言語の姿」（山口 1993,p.10）により、特に、文章の種類の違いに端的に現れる言語の姿を「文体」と呼ぶ。

[2]本文は、小学館の「新編日本古典文学全集」の『今昔物語集 1～4』（馬淵和夫・国東文麿・稲垣泰一校注）により、コーパス構築のために小学館から国立国語研究所に提供された電子テキストを利用している。新編全集の『今昔』は、巻 1～10 の天竺震旦部は収録しておらず、巻 11～31 の本朝部のみを収録しており、コーパス化の対象もこの範囲になる。

巻 27 については先行してコーパス化を行っており、自動形態素解析を施したデータに目視で修正を行ってデータを整えた[3]。表 1 は、異なり語数と延べ語数、図 1 は、その語種構成比率（異なり語数）を示したものである。自立語だけを対象とし、固有名詞を除外して集計した。これらの図表には、『今昔』巻 7[4]と、平安和文の『大和物語』と『和泉式部日記』を合わせたもの[5]のデータも表示した。

表 1　各資料の語彙量

資料	異なり語数	延べ語数
今昔巻 7	1,861	12,590
今昔巻 12	2,426	17,859
今昔巻 27	1,837	14,294
大和・和泉	2,415	17,839

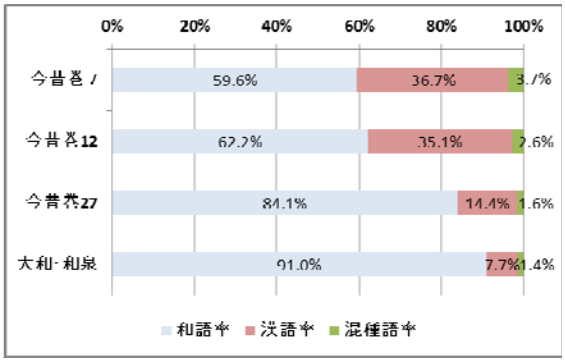


図 1　各資料の語種構成比率（異なり語数）

図 1 によると、『今昔』の巻 12 と巻 27 との間で大きな落差がある。巻 7 は純漢文を依拠資料とする漢文訓読文の巻、巻 12 は和化漢文を依拠資料とする漢文訓読文の巻であるが、そこにはほとんど差がない。巻 27 と平安和文（『大和』『和泉』）との間には差があるが、巻 12 と巻 27 の差に比べると小さい。このように、和漢混淆文である『今昔』における、巻 12（漢文訓読文）と巻 27（和文）の語彙に大きな差があることから、この 2 つの巻の語彙を比較することで、文体の違いによる語彙の違いを知ることができると予想される。

３．巻 12 と巻 27 の特徴語の抽出

『今昔』の巻 12 と巻 27 について、一方の巻によく使われ、他方の巻にあまり使われない特徴語を抽出する。抽出の手順は次の通りである。まず、巻 12、巻 27 それぞれにおいて、各語の頻度と、巻全体の延べ語数を用いて、各語の使用率を計算し、10,000 語あたりに何件使われるかを数値に表した。

使用率　＝　（頻度／延べ語数）×10,000

次に、上記で算出した巻 12 使用率と巻 27 使用率とを比較した比率を算出した。表 2 は、一定頻度以上の語彙について、50 音順に冒頭の 10 語の上記の数値を一覧にしたものである。表 2 の「1.相」、「3.会ふ」、「4.合ふ」は、巻 12 比率、巻 27 比率ともに 0.4 から 0.6 の間の数値を示しており、巻 12 や巻 27 のいずれかに偏って使用されているわけでなない。これに対して、「5.障泥」は巻 12 の比率が 0.8 を、また、「10.上がる」は巻 27 の比率が 0.92 を、それぞれ示しており、巻 12 および巻 27 それぞれに、特徴的な語と見てよいだろう。本稿では、この比率が 0.8 以上のものを、その巻の特徴語と扱うことにした。その結果、巻 12

3 語の認定基準は、小木曽・小椋・須永（2012）が、中古和文を対象として定めた、短単位規程に従うが、『今昔』に適用するにあたって一部変更したところがある。

4 『今昔』巻 7 の本文には、国文学研究資料館の日本古典文学本文データベース（底本は日本古典文学大系・岩波書店）の電子テキストを用いた。

5 『大和物語』『和泉式部日記』のデータは、『日本語歴史コーパス 平安時代編』による。この 2 作品を選んだのは、延べ語数の合計がちょうど『今昔』1 巻分に相当し、比較しやすいからである。

の特徴語が 303 語、巻 27 の特徴語が 243 語得られた。

表２　『今昔』巻 12・巻 27 の使用率の比率（50 音順冒頭 10 語）

No	読み	語彙素	品詞	巻 12 頻度	巻 27 頻度	頻度計	巻 12 使用率	巻 27 使用率	使用率計	巻 12 比率	巻 27 比率
1	アイ	相	接辞	4	3	7	2.24	2.1	4.34	0.52	0.48
2	アイダ	間	名詞	123	34	157	68.87	23.79	92.66	0.74	0.26
3	アウ	会ふ	動詞	21	19	40	11.76	13.29	25.05	0.47	0.53
4	アウ	合ふ	動詞	13	9	22	7.28	6.3	13.58	0.54	0.46
5	アオリ	障泥	名詞	5	1	6	2.8	0.7	3.5	**0.8**	0.2
6	アカイ	赤し	形容詞	2	5	7	1.12	3.5	4.62	0.24	0.76
7	アカイ	明し	形容詞	2	6	8	1.12	4.2	5.32	0.21	0.79
8	アカガネ	銅	名詞	5	3	8	2.8	2.1	4.9	0.57	0.43
9	アカツキ	暁	名詞	2	5	7	1.12	3.5	4.62	0.24	0.76
10	アガル	上がる	動詞	1	9	10	0.56	6.3	6.86	0.08	**0.92**

４．内容の違いによる特徴語

巻 12 の特徴語、巻 27 の特徴語は、それぞれ、説話の内容の違いに基づく特徴語がその多くを占める。例えば「阿闍梨」という語は、巻 12 に 44 件用いられており、巻 27 には 1 件も用いられていない。巻 12 のそれらは、6 つの説話で、飯室の阿闍梨（第 22 話）ら 6 人の人物を指して用いられているものである。「阿闍梨」とは、平安時代の日本では、儀式を執り行うなど一定の役割を担う、真言宗や天台宗の僧侶の職名として用いられることが一般的であった（『日本国語大辞典 第 2 版』小学館）。『今昔』で、「阿闍梨」が巻 12 に多く、巻 27 に皆無であるのは、それぞれの巻に収録された説話が、このような人物を多く登場させているか否かという、説話の内容の違いに基づく現象である。同じく「尼」が、巻 12 に 24 件あって、巻 27 に皆無であるのも、尼が登場する説話が一定数あるか全く無いかの違いによるものである。巻 12 と巻 27 の説話を読んで検討した結果、巻 12 特徴語 303 語中 231 語が、内容の違いに基づく特徴語だと考えられた。同じような手順で、巻 27 特徴語を一語一語検討していくと、「兄」「弟」など、243 語中 149 語が内容の違いに基づく特徴語だと考えられた。

それでは、それ以外の巻 12 特徴語 72 語、巻 27 特徴語 94 語は、どのようなものだろうか。

５．文体の違いによる特徴語

5.1　慣用句が特徴語となるもの

巻 12 にまとまった件数があり、巻 27 にはそれがほとんどない語の中には、特定の慣用句として用いられている例が多くを占めるものがある。このような場合は、その語自体が特徴的だということよりも、その慣用句が特徴的だと見るべきだろう。

例えば、「思ひ」は、巻 12 に 7 件、巻 27 に 1 件あり、巻 12 特徴語だが、そのうち、「思

ひの如く」という慣用句が、巻 12 で 4 件、巻 27 で 1 件を占め、この慣用句が巻 12 に特徴的だと見ることができる。「思ひの如く」が漢文訓読文である巻 12 に特徴的であるのは、「如思」という漢語の訓読表現だからだろう。このような特徴語は、ほかに、「首（かうべ）」（「首を挙ぐ」は「挙首」の訓読表現）、「掌」（「掌を合はす」は「合掌」の訓読表現）、「涙」（「涙を流す」は「流涙」の訓読表現）など、全部で 10 語ある。

巻 27 特徴語のうち慣用句が特徴語になるのは、「太る」（「頭の毛太る」）1 語だけである。この慣用句は話し言葉の中で使われていたものだろう。

これらは、慣用句が、漢文の訓読によってできたものか、日本語の話し言葉の中でできたものかによって、文体の違いによる特徴語になっていたものだと見ることができる。

5.2　他方の巻に対立語があるもの

文体の違いによる特徴語には、慣用句のほかに、他方の巻にほぼ同義の対立語がある場合がある。

例えば、動詞「与ふ」は、巻 12 には 20 件あるのに対して、巻 27 では 2 件にとどまり、巻 12 に特徴的である。各巻から 1 例ずつを挙げれば、次の通りである。

正則此ノ牛ヲ聖人ニ与へツ。（巻 12-24）

男懐ヨリ玉ヲ取出シテ女ニ与へツ。（巻 27-40）

第 1 例は、正則が牛を聖人に授与する例、第 2 例は、男が玉を女に授与する例であるが、

これらの例のような、誰かが何かを誰かに授与するという内容は、巻 12 にも巻 27 にもよく見られるものであり、「与ふ」が、巻 12 の特徴語となる理由を、説話が扱う内容の違いに求めることはできない。

それでは、「与ふ」は、なぜ巻 12 に多く、巻 27 には少ないのだろうか。その理由は、次のような例を見ることで推測することができる。

（遠助は）此ノ箱ヲバ女ニ取（とら）セムトテ　（巻 27-21）

この例も、誰かが何かを誰かに授与することを言っているものであるが、「与ふ」ではなく、「取らす」を用いている。巻 27 には、これと同じ意味で「取らす」を使う箇所が 12 件あり、「与ふ」よりもずっと多い。誰かが何かを誰かに授与する意味を表すのに、巻 12 では「与ふ」を用いることが一般的であり、巻 27 では「取らす」を使うことが一般的なのである。

同じ意味を表すのに、巻 12 でよく使われる語と巻 27 でよく使われる語が違うのは、漢文訓読文と和文との文体の違いによるのではないだろうか。ほぼ同じ意味の語で、他方の文体でよく使われる語のことを、「文体対立語」あるいは単に「対立語」と呼ぶことにする。

６．文体対立語

6.1　対立語のリスト

巻 12 の特徴語のうち、巻 27 に対立語を持つものについて、各巻の頻度とともに一覧にすると、表３の通りである。同じく、巻 27 の特徴語のうち、巻 12 に対立語を持つものについて一覧にすると、表４の通りである。

6.2　対立の多層性

表３で 35 対、表４で 28 対が得られているが、そのうち 13 対が双方向から得られたもの（「双方向」の列に「＊」の印があるもの）であるので、異なりで 50 対の対立が見出され

表３『今昔』巻 12 特徴語のうち巻 27 に対立語があるもの

巻 12 特徴語	語種	品詞	巻 12 頻度	巻 27 頻度	巻 27 の対立語	語種	品詞	巻 12 頻度	巻 27 頻度	双方向
疑ひ〔有らず、有らじ、無し〕	和	名詞	10	0	必ず	和	副詞	29	23	
女子（をんなご）	和	名詞	7	0	女の童	和	句	0	23	
奇異	漢	名詞	19	2	あさまし	和	形容詞・形状詞	1	31	*
悟り〔有り〕	和	名詞	10	1	賢し	和	形容詞・形状詞	1	9	*
大臣（だいじん）	漢	名詞	15	3	大臣（おとど）	和	名詞	0	18	*
為（ため）	和	名詞	54	8	とて	和	句	16	45	
端正	漢	名詞	6	0	うるはし	和	形容詞・形状詞	0	3	
並び〔に〕	和	名詞	16	0	など	和	助詞	6	95	
女人（にょにん）	漢	名詞	16	0	女（をんな）	和	名詞	17	111	
父母（ぶも）	漢	名詞	30	4	親	和	名詞	1	4	*
諸々	和	名詞	25	1	あまた	和	副詞	0	18	*
汝	和	代名詞	44	6	其（そこ）	和	代名詞	6	25	*
					己（おのれ）	和	代名詞	2	19	
与ふ	和	動詞	20	2	取らす	和	句	0	12	
哀れむ	和	動詞	10	0	哀れ	和	形容詞・形状詞	10	14	
至る	和	動詞	27	3	つく	和	動詞	19	44	
生まる	和	動詞	20	0	生む	和	動詞	1	4	*
終はる	和	動詞	26	1	果つ	和	動詞	2	10	*
悲しむ	和	動詞	35	2	悲し	和	形容詞・形状詞	9	2	
化（け）す	混	動詞	5	0	成る	和	動詞	53	74	
叫ぶ	和	動詞	11	0	声を挙ぐ	和	句	11	4	
従ふ	和	動詞	22	2	つく	和	動詞	19	44	
住す	混	動詞	14	1	住む	和	動詞	10	43	*
生ず	混	動詞	8	0	生む	和	動詞	4	1	
悩む	和	動詞	6	0	煩ふ	和	動詞	7	4	
開（ひら）く	和	動詞	22	4	開（あ）く	和	動詞	1	19	*
経（ふ）	和	動詞	26	4	過ぐ	和	動詞	12	30	
在（ましま）す	和	動詞	29	1	おはす	和	動詞	1	38	*
用ゐる	和	動詞	6	0	使ふ	和	動詞	7	5	
養ふ	和	動詞	6	1	傅（かしづ）く	和	動詞	1	2	
喜ぶ	和	動詞	32	4	嬉し	和	形容詞・形状詞	2	11	*
明らか	和	形容詞・形状詞	5	0	明（あか）し	和	形容詞・形状詞	2	6	
速やか	和	形容詞・形状詞	17	2	疾（と）し	和	形容詞・形状詞	3	11	*
具（つぶ）さ	和	形容詞・形状詞	7	1	詳し	和	形容詞・形状詞	0	3	
つゆ	和	副詞	8	0	聊か	和	副詞	3	3	
御（ご）	漢	接辞	7	0	御（み）	和	接辞	12	5	

表４『今昔』巻 27 特徴語のうち巻 12 に対立語があるもの

巻 27 特徴語	語種	品詞	巻12頻度	巻27頻度	巻 12 の対立語	語種	品詞	巻 12 頻度	巻 27 頻度	双方向
あまた	和	名詞	1	18	多く	和	名詞	14	5	＊
					諸々	和	名詞	25	1	
大臣（おとど）	和	名詞	0	18	大臣（だいじん）	漢	名詞	15	3	＊
親	和	名詞	1	4	父母（ぶも）	漢	名詞	30	4	＊
顔	和	名詞	1	15	面（おもて）	和	名詞	3	3	
侭	和	名詞	0	16	如し	和	助動詞	93	11	
					時に	和	句	75	57	
様（やう）	漢	名詞	12	62	如し	和	助動詞	93	11	
渡り	和	名詞	0	9	辺（ほとり）	和	名詞	15	16	
己（おのれ）	和	代名詞	2	19	汝	和	代名詞	44	6	＊
開（あ）く	和	動詞	1	19	開（ひら）く	和	動詞	22	4	＊
生む	和	動詞	1	4	生まる	和	動詞	20	9	＊
御（おは）す	和	動詞	1	38	在（ましま）す	和	動詞	29	1	＊
斯かり	和	動詞	1	23	然り	和	動詞	184	107	
住む	和	動詞	10	43	住す	混	動詞	14	1	＊
果つ	和	動詞	2	10	終わる	和	動詞	26	1	＊
あさまし	和	形容詞･形状詞	1	31	奇異	漢	名詞	19	2	＊
怪し	和	形容詞･形状詞	1	23	怪しむ	和	動詞	28	12	
嬉し	和	形容詞･形状詞	2	11	喜ぶ	和	動詞	32	4	＊
恐ろし	和	形容詞･形状詞	10	52	恐る	和	動詞	8	13	
賢し	和	形容詞･形状詞	1	9	悟り有り	和	句	5	1	＊
斯様（かやう）	漢	形容詞･形状詞	0	8	如此（かくのごと）し	和	句	35	1	
疾（と）し	和	形容詞･形状詞	3	11	速やか	和	形容詞･形状詞	17	2	＊
いと	和	副詞	6	33	極めて	和	副詞	31	29	
え	和	副詞	0	27	能ふ	和	動詞	4	0	
然（さ）	和	副詞	4	25	然（しか）	和	副詞	3	3	
暫し	和	副詞	1	9	暫く	和	副詞	12	10	
軈（やが）て	和	副詞	0	6	即ち	和	副詞	20	6	
あな	和	副詞	0	8	ああ	和	副詞	1	0	
人（たり）	和	接辞	1	10	人（にん）	漢	接辞	9	7	

たことになる。双方向から得られた対は、文体的対立が非常に明瞭なもの、一方向のみから得られた対は、対立がさほど明瞭でないものであり、対立のありようは多層的である。その対立と層のありようを図示すると、図２のようになる。

図のａには、表３で「双方向」に「＊」が付いている 13 語の対立が該当する。例えば「奇異」は上の「漢文訓読文に特徴的に見られる語彙の層」に位置付き、下の「和文に特徴的な語彙の層」に位置付く「あさまし」に対立している。図のａ’には、表４で「双方向」に

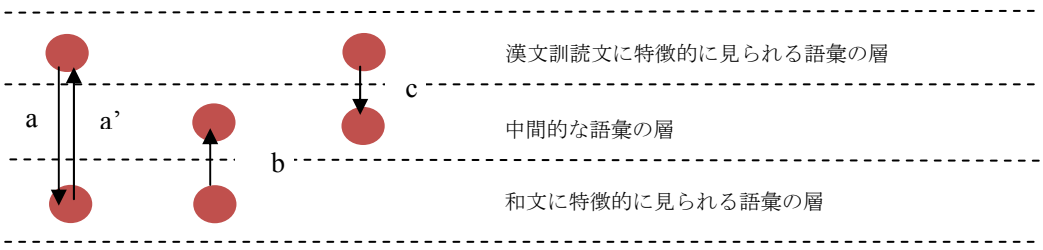


図２　語の対立と語彙の層

「＊」が付いている 13 語が該当し、例えば、下の層にある「あさまし」から上の層にある「奇異」との対立が見える。「奇異」「あさまし」の対は、上層からも下層からも対立が見える、非常に鮮明な対立を持っている。

図の b には、表４で「双方向」に「＊」が付いていない 15 対が該当する。例えば「顔」は、下層の「和文に特徴的に見られる語彙の層」に位置付き、「中間的な層」にある「面（おもて）」と対立している。「中間的な層」の語は、文体的な特徴語ではなく無色な語であるため、 a の場合ほど、対立関係が鮮明ではない。図の c には、表３で「双方向」に「＊」が付いていない 22 語が該当する。例えば、「疑い〔有らず、有らじ、無し〕」は、上層の「漢文訓読文に特徴的に見られる語彙の層」に位置付き、「中間的な語彙の層」にある「必ず」と対立している。 b の場合と同じく、文体的に無色な「必ず」からは対立が見えにくく、やはり a の場合ほどには対立が鮮明でない。

このような文体的な対立関係にある語は、従来、「漢文訓読語」と「和文語」の対立というように言われてきたが（築島 1963,1969）、そのような単純な対立にとどまるものではなく、連続的な層を背景とした対立だと見るべきであろう。

6.3　対立の類型

先に挙げた 50 対の対立を見ると、語種・品詞の観点から、次の３つの類型に分けられる。各類型における挙例では、漢の側の層（中間的な層を含む）の語／和の側の層（中間的な層を含む）の語、という形式であげる

（１）　異語種間の対立　10 対

奇異／あさまし、大臣（だいじん）／大臣（おとど）、父母（ぶも）／親、住す／住む、端正（たんじやう）／うるはし、女人／女（をんな）、化す／成る、生ず／生む、御（ご）／御（み）、人（にん）／人（たり）

漢の層と和の層の語の対立ということを考えると、漢語と和語の対立ということがまず思い浮かぶが、それは上の 10 対（20%）にとどまり、実際は多くない。

（２）　異品詞間の対立　18 対[6]

疑ひ〔有らず、有らじ、無し〕［名詞］／必ず［副詞］、女子（をんなご）［名詞］／女の童［句］、奇異［名詞］／あさまし［形容詞］、悟り〔有り〕［名詞］／賢し［形容詞］、為（ため）［名詞］／とて［句］、端正［名詞］／うるはし［形容詞］、並び〔に〕［名詞］／など［助詞］、諸々［名詞］／あまた［副詞］、与ふ［動詞］／取らす［句］、哀れむ［動詞］／哀れ［形状詞］、悲しむ［動詞］／悲し［形容詞］、叫ぶ［動詞］／声を挙ぐ［句］、喜ぶ［動詞］／嬉し［形容詞］、如し［助動詞］・時に［句］／侭［名詞］、如し［助動詞］／様〔やう〕［名詞］、怪しむ［動詞］／怪し［形容詞］、恐る［動詞］

[6] 異語種間で異品詞のものもここに計上したため、（１）と重複して分類したものが 2 対ある。

／恐ろし［形容詞］、此様〔かやう〕［形状詞］／如此〔かくのごとし〕［句］

様々な品詞間で対立があるが、漢の側の層には名詞と動詞が多く、和の側の層には形容詞が多い。特に、感情動詞と感情形容詞の対が多いことが目を引く。

（３） 同語種・同品詞の対立　24 対（48%）

諸々／あまた［名詞］、汝／其（そこ）・己（おのれ）［代名詞］、至る／着く［動詞］、生まる／生む［動詞］、終はる／果つ［動詞］、従ふ／付く［動詞］、悩む／煩ふ［動詞］、開（ひら）く／開（あ）く［動詞］、経（ふ）／過ぐ［動詞］、在（ましま）す／御（おは）す［動詞］、用る／使ふ［動詞］、養ふ／傅（かしづ）く［動詞］、明らか／明し［形容詞・形状詞］、速やか／疾（と）し［形容詞・形状詞］、具（つぶ）さ／詳し［形容詞・形状詞］、つゆ／聊か［副詞］、面（おもて）／顔［名詞］、辺（ほとり）／渡り［名詞］、極めて／いいと［副詞］、然（しか）／然（さ）［副詞］、暫く／暫し［副詞］、即ち／軈（やが）て［副詞］、ああ／あな［感動詞］

多様な品詞があるが、動詞と副詞が目立っており、この 2 つの品詞が対立関係を構成しやすいことが分かる。

（１）（２）（３）の類型ごとに、対立の実際を、個別に研究していくことが必要になる。田中（2000）は（１）の事例、田中（1998）は（２）の事例を取り上げた研究である。

７．おわりに

内容の違いによる対立語にも、慣用句があるものでも対立語があるものでもない語が、巻 12 特徴語で 27 語、巻 27 特徴語で 65 語残る。それが、どのような語であるかは、現段階では不明である。これについて研究していくことも課題である。

謝辞

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）、及び、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、研究代表者：田中牧郎）による成果の一部です。

文献

小木曽智信・小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規程集」（科研費報告書）
佐藤武義（1984）『今昔物語集の語彙と語法』（明治書院）
関一雄（1993）『平安時代和文語の研究』（笠間書院）
田中牧郎（1998）「今昔物語集の情意述語文と文体」（『国語学』194）
田中牧郎（2000）「今昔物語集における和漢の対語の意味対立―〈奇異〉〈微妙〉を例として―」（遠藤好英編『語から文章へ』）
田中牧郎・山元啓史（2014）「『今昔物語集』と『宇治拾遺物語』の同文説話における語の対応―語の文体的価値の記述―」（『日本語の研究』10-1）
築島裕（1963）『平安時代の漢文訓読語につきての研究』（東京大学出版会）
築島裕（1969）『平安時代語新論』（東京大学出版会）
西田直敏（1978）『平家物語の文体論的研究』（明治書院）
藤井俊博（2003）『今昔物語集の表現形成』（和泉書院）
舩城俊太郎（2011）『院政時代文章様式史論考』（勉誠出版）
峰岸明（1986）『平安時代古記録の国語学的研究』（東京大学出版会）
山口佳紀（1993）『古代日本文体史論考』（有精堂）

ＢＣＣＷＪのタグ情報の修正について

長谷川守寿（首都大学東京）

Correction of the Tag Information on BCCWJ

Hasegawa Morihisa(Tokyo Metropolitan University)

１．目的

本発表では、「現代日本語書き言葉均衡コーパス DVD 版」（Contemporary Written Balanced Corpus of Japanese、以後 BCCWJ と略す）の C-XML ファイルに付与されている文タグの修正箇所の検討と、その作業から見えてくる BCCWJ の特徴について述べる。

BCCWJ を使うには、検索サイト「少納言」（www.kotonoha.gr.jp/shonagon/）または「中納言」（https://chunagon.ninjal.ac.jp/）にアクセスするか、BCCWJ-DVD 版を入手し検索ツールを各自が準備して使用するか、どちらかの方法がある。

筆者は日本語学習用のコロケーション情報を抽出するため、文単位のデータが必要となった。そこで BCCWJ-DVD 版を入手し文タグ（〈sentence〉と〈/sentence〉で一つのペア）を用いて、検索用プログラムを作成していたところ、文タグに問題があり、完全には文単位に分けられたデータが得られないことが判明した。最初に発見したのが(1)であり、文タグが示す一文中（〈sentence〉と〈/sentence〉に挟まれた間）に二つ以上の文が入っている。

(1)　〈sentence〉これを「いや、よしてしと読めば、いやがっているという意味である。はたしてそれだけだろうか。句読点をすこし動かして〈quote〉「いやよ、して」〈/quote〉と読めば、（略）。〈/sentence〉　（PB10_00030）

そこで、文タグを修正してから、BCCWJ のデータを研究対象として使用する場合、修正が必要な箇所はどれくらいあるのか、修正箇所の多寡はサブコーパスや媒体により違いはあるのか、そして目的にもよるが、これらのことから文タグを修正して使用するには、どのサブコーパスや媒体が適当かを明らかにする。

２．先行研究

BCCWJ の文書構造タグ（XML）については、『第５章　文書構造タグ』（国立国語研究所 2011）に詳述されており、文タグは階層構造に関するタグ（article、cluster、paragraph など）に含まれ、これ自体は「文に相当する文書要素」を示すのに使われ、実際には〈sentence〉と〈/sentence〉によって区切られている。

文タグの不備については、田野村（印刷中）で「出版と図書館のサブコーパスだけに限っても、少なくとも文の連続の数で約 3,000 か所、文の数で約 11,000 件の文への文タグの付与が漏れている」とし、その原因として「多くの場合、複数の段落にわたる引用、または、注番号の存在のいずれか」を挙げているが、「文タグを参照する処理を自前で行うのではない限り、BCCWJ の利用への影響はほとんどない」としている。本調査のきっかけは、まさに「文タグを参照する処理を自前で行う」ものであるが、田野村（印刷中）以外に文タグの問題点を指摘した調査は見つからないのが現状である。なお国立国語研究所（2011）では「ruby タグ」「correction タグ」等と使われているが、本稿では文タグと呼ぶこととする。

３．方法

本発表の目的は、文タグの追加が必要な箇所はどのくらいあるのか、サブコーパス・媒体によって修正箇所数や出現状況に違いはあるのかを明らかにすることである。これらの考察から、BCCWJ を修正して使う場合、どのサブコーパスが適切かを考え、さらに BCCWJ の特徴にも言及したい。本研究の対象と手続きは以下の通りである。

３．１　対象

本調査の目的は BCCWJ における文タグの調査であるため、出版サブコーパス、図書館サブコーパス、特定目的サブコーパス全てを扱う。またデータ形式は、ファイル数が多く、形態素情報のタグがついていない C-XML（Charactor-base XML）のデータを使用し、文単位を対象とするため可変長（Variable）を用い、ファイルに含まれる全ての文を対象とする。

なお、BCCWJ では文字コードの符号化方式に UTF-8 が採用されているため、以後の正規表現は UTF-8 に対応した表現を用いた。検索には Ruby（1.8.7）を用いて、スクリプトを作成した。総ファイル数は、172,675 ファイルである（BCCWJ ではファイル数をサンプル数と呼んでいるが、本発表では「ファイル数」と呼ぶ）。

３．２　手順

文タグは「文に相当する文書要素」をマークするタグであり、実際には〈sentence〉と〈/sentence〉でマークされ、(2)(3)(4)のように句点“。”・感嘆符“！”・疑問符“？”の後に“〈/sentence〉”が入力されているのが適切な箇所にタグが付されている例である。

文の終端には句点などいろいろな記号が位置するが、句点で終了する文だけに調査を限定し、感嘆符や疑問符、三点リーダー“…”などで終わる場合については紙幅の都合で省略する（結果は長谷川（印刷中）を参照）。

(2)　〈sentence〉もはやまっしぐら、一直線である。〈/sentence〉　（LBa0_00002）
(3)　〈sentence〉戦前とくらべたら、なんというちがいでしょう！〈/sentence〉　（LBa9_00026）
(4)　〈sentence〉何が大丈夫か？〈/sentence〉　（LBa5_00004）

次に文タグが欠如した例について述べる。(2)(3)(4)に対し、(5)の「…んです。僕…」の部分（場所を明示するために下線を施した。以下同）には、文タグが欠如していると考える。

(5)　〈sentence〉　「カッパというのは二十歳前からのあだ名で、なぜか人々はどんな時でもぼくのことを『カッパ』としか呼ばなくなったんです。僕はそう呼ばれても平気でした。〈/sentence〉　（PB12_00071）

タグが適切に入力されていない箇所を特定する方法に先んじて、修正箇所の数え方について説明しておきたい。本調査では、(5)のような場合、“〈/sentence〉〈sentence〉”を入力すればよいので、修正箇所は１と数える。

文タグが入力されていない箇所は、(5)の「。僕」のように、句点の後に２バイト文字“僕”が続く場合と、(6)のように、句点の後に１バイト文字“〈”が続く場合がある。１バイト文字が続く場合は、後続する文に関する何らかのタグが入力されている。例えば(6)は、句点の後に、後続する文の最初の漢字「真綿」の ruby タグ（以後ルビタグと呼ぶ）が続いてお

り、他に引用情報やサンプリングの開始情報など、いくつかのタグが後続する。

本稿では、〈sentence〉で始まり〈/sentence〉タグで終わっている用例については、表記を省略する。ただし〈sentence type = "quasi"〉のように特殊なものや他のタグは、残すこととする。また、段落の存在を示す行頭の２バイト空白文字は、見やすいように“□”で示す。

(6)　□「鉛色の空から、こまかな雪があとからあとから降ってくる。〈ruby rubyText="ま"〉真〈/ruby〉〈ruby rubyText="わた"〉綿〈/ruby〉をちぎったようなという形容があるけれど、この雪は真綿ではない。（略）　（PB12_00310）

本調査では詳細に調査するために二種類の検索方法を用いた。一つは、(7)のように文字の種類や記号そのものを特定して検索する方法である（以後「文字法」と呼ぶ）。文字の種類は、ひらがな・漢字・カタカナ・アラビア数字（以後、数字）・アルファベット・ギリシャ文字・キリル文字であり、記号については文字コードが連続していないため、個別に特定して“[★☆]”のように記述する（実際には“Å”から“＝”まで）。

(7)　。[ぁ-ん]|。[一-龠々]|。[ァ-ヴー]|。[０-９]|。[Ａ-Ｚ ａ-ｚ]|。[Ａ-Ω α-ο π-ω]|。[А-я ё]|。[Å‰♯♭♪†‡¶∠⊥⌒∂∇≡≒≪≫√∽∝∵∫∬∧∨￣⇒⇔∀∃∈∋⊆⊇⊂⊃∪∩,.・:;ˇ゜ ´ ` ¨ ^ ￣＿ヽヾゝゞ〃仝々〇ー－／＼～∥｜‥‘’“”（〔［｛〈《「『【＋－±×÷＝≠＜＞≦≧∞∴♂♀°′″℃￥＄¢£％＃＆＊＠§☆★○●◎◇◆□■△▲▽▼※〒→←↑↓＝]

もう一つの方法は、特定の文字列の連続ではなく(8)のような正規表現を用いた検索である（以後「否定法」と呼ぶ）。この方法では、“。"”“。〉”“。）”“。〕”“。］”“。〉”“。》”“。」”“。』”“。】”“。□”“。。”は文末としてあり得るので、修正箇所から除外され、ほぼ文字列法と同様の検索結果が得られるが、句点の後にルビ情報が付与された漢字から始まる文などが検出できるのが異なる点である。この方法では(9)のように“。」”を含む部分は文タグが不要なので検出されない。また(10)の“。。”の部分も検出しないが、「。活」の部分は文タグが欠如した部分として検出できるようになっている。

(8)　。[^"〉〕〉」）］』》　。]|。〈r|。〈q|。〈sa|。〈no

(9)　〈sentence type="quasi"〉□「説明をしにおうかがいしたいと思いますが、お名前と住所を教えてください。」〈/sentence〉　（PB11_00006）

(10)　〈sentence type="quasi"〉ココの素材を加工できるソフトを教えてください。。。。活字が打てるといいです〈/sentence〉　（OC02_02989）

４．結果

４．１　検索結果

二つの方法を用いて修正箇所を検出した結果について、説明する。

文字法での検索では、句点にひらがな・漢字・カタカナ等が後続した場所にタグが抜けている例が検出され、(11)では「。また」「。我」「。毎」「。『」のような例が見える。

(11) 〈sentence type="quasi"〉□「―千九百十年は我々の最も得意の時代であつた。『パンの會』は毎週ひらかれた。我々はロダンの銅像の首の唇に寄せた皺の〈ruby rubyText="ねば"〉粘〈/ruby〉こさが何ういふ情を〈ruby rubyText="か"〉藏〈/ruby〉くしてゐるかゞ分るほどになつた。また〈ruby rubyText="（アラビア）"〉亜剌比亜〈/ruby〉物語や近松・三馬などに出てくる青年の心に同情を寄する程の苦勞も覚えた頃である。毎日同じ仲間と交遊して作詩し、作劇して日を暮した。…〈/sentence〉　（PB29_00042）

データは EOS まで読み込まれるが、一読み込み単位で修正箇所数が最大となったのは、48 箇所の修正が必要となる OB3X_00110 である。なお OB3X_00110 にはルビタグが多くつき、多少見にくいので、その次に 45 箇所と多かった PB15_00180 の一部を示す。

(12) □「それもあるかもしれない。それかうまくなるかですね。ゲームの中のハードであるクルマのチューニングも重要だけど、ホントのハードウエアのチューニングは大事です。(略) でも、生産技術は０。　（PB15_00180）

一方、修正すべきでない点を修正箇所として検出した例がある。以下はギリシャ文字を指定し検索した例であるが、(13)はタグが抜けている部分として正しく検出できるが、(14)ではギリシャ文字が顔文字の一部に使用されており、文タグがなくても問題ないため、検出したことは誤りである。文字種により検索精度に違いがあるので、後述する。

(13) Ｘ線のエネルギーは振動数Ｖ＝ｃ／λ）のｈ倍、つまりｈνですから、前方に散乱されるときは、エネルギーは変わりません。θが大きくなると、（略）（PB24_00213）
(14) 〈sentence type="quasi"〉□ピカチュウしか知らなくて（；σ。σ）ゞ　ゴメンネェ．．．〈/sentence〉　（OY14_49261）

４．２　異なる検索方法による違いについて

文字法と否定法を用い修正箇所を検出した結果をファイル単位で見たのが表１である。

表1　修正が必要なファイル数

検索方法	要修正	必要なし	合計
文字法	3,300(1.9%)	169,375(98.1%)	172,675(100%)
否定法	3,738(2.2%)	168,937(97.8%)	172,675(100%)

文字法の検索では、少なくとも一カ所以上の修正点をもつファイルは 3,300 あり、中でも(15)は最も多い 96 カ所の修正点が含まれていた（一部を示す）。

(15) □「多くの人がその過程に吸収されました。一九三〇年ごろ、インドは爆発的な成長をとげました。インドは四七年ごろに弾みをつけ、その後に大きく成長しましたが、今は鈍化しつづけています。一九六二年に私は大学に残るか仕事かを考えあぐね、専門家としての道、研究機関を選びました。今日、人はいっそう奮闘せねばなりませんが、これにはいい面もあるのです。　（PB29_00606）

媒体別に修正箇所のあるファイル数を調べた結果が表２であり、修正箇所数とその差について調べた結果が表３である。媒体の順序は国立国語研究所（2011,p15）に倣った。

表２　媒体別の修正ファイル数

媒体	文字法	否定法
PB	775	825
PM	102	109
PN	5	5
LB	289	306
OW	49	50
OT	22	22
OP	44	45
OB	112	114
OC	821	980
OY	1,059	1,258
OV	15	16
OL	5	5
OM	2	3
合計	3,300	3,738

表３　媒体別の修正箇所数とその差

媒体	文字法(A)	否定法(B)	(B)−(A)
PB	3,938	4,254	316
PM	434	457	23
PN	12	12	0
LB	1,004	1,053	49
OW	103	105	2
OT	86	86	0
OP	98	103	5
OB	670	737	67
OC	1,421	1,589	168
OY	2,111	2,373	262
OV	36	37	1
OL	8	8	0
OM	19	20	1
合計	9,940	10,834	894

文字法と否定法の修正箇所数を比べると、PB（出版-書籍）、OC（Yahoo!知恵袋）と OY（Yahoo!ブログ）のように 100 以上異なるものと、PM（出版-雑誌）、LB（図書館-書籍）のように差が小さいものに分かれる。否定法でのみ検出されたものを確認したところ、PB（出版-書籍）PM（出版-雑誌）PN（出版-新聞）LB（図書館-書籍）OB（ベストセラー）では、ルビタグを含むもの、引用タグ（quote）を含むもの(16)、サンプリングの開始位置を示すもの(17)、傍注を示すもの（noteBodyInline）が出現しており、そのため数値が異なった。

(16)　（略）原語の音とともにひっくるめて借用する場合が〈quote〉「音借用」〈/quote〉である。〈quote〉「借用語」〈/quote〉とか〈quote〉「外来語」〈/quote〉と呼ばれているのは〈quote〉「音借用」〈/quote〉語のことである。　（PB28_00049）

(17)　□「父の少年時代の写真を見ると、変な気持ちになります。〈sampling type="start" /〉私の孫の年齢ですから。五十年の恨のたった一つだけ、（略）　（LBi9_00146）

しかし OC・OY では、句点に“？”や“！”が続く例を検出している。これらは検索条件作成時には文の終端としては想定しておらず、否定法ではこういった用例が検出され、文が後続していない部分を誤って検出したことになる。

(18)　貼り付け方を教えてください。？　（OC02_07366）
(19)　あなたは、リカバリーＣＤを大事に保管していますか。！　（OC02_07489）

４．３　検出結果の精度について

文字種により検出されたものの精度に違いが見られたので、文字種別に検出し、その中から無作為抽出した 100 カ所（用例数が 100 以下の場合は全て）について確認する。

表4　後続文字列別の修正箇所数と検討数、その中で実際に文が後続した数

	ひらがな	漢字	カタカナ	数字	アルファベット	ギリシャ文字	キリル文字	記号	合計
箇所数	3758	3757	647	120	147	6	4	1501	9940
検討数	100	100	100	100	100	6	4	100	610
文後続	99	100	99	100	85	1	0	27	526

句点に後続する２バイト文字の文字種から、修正箇所を再集計し精度を確認したのが表４である。なお句点に後続する１バイト文字（この場合は“〈”）は除外した。以後文字種毎に検討するが、ギリシャ文字は後続する全例（６例）を確認したところ(14)のように５カ所で顔文字の一部として使われ、文が後続したのは(13)のみであったため、小節は設けない。

４．３．１　ひらがな

ひらがなが後続する例は 3,758 カ所に見られたので、無作為に抽出した 100 カ所を検討した結果、99 カ所でその後に文が続くことが確認された。文が続かないと思われる例は(20)の１例のみであり、誤入力と思われる。

(20)　〈sentence type="quasi"〉□「そう言うと兵をひきいて城外へ突撃した。張巡の兵は勇戦して賊将十四人をとらえ、八百余の首級をあげた。ん〈/sentence〉　（PB49_00170）

４．３．２　漢字

漢字が後続する例は 3,757 カ所で見られたので、100 カ所検討した。括弧“（）”に囲まれた注釈の部分は文に含まれるのか(21)、中国語の文型の部分は文に含めるのか(22)など、文の定義について再考する必要が出てくるが、この部分も文とすると、検出できている。

(21)　“地獄の天使”（オートバイの暴走族。元来はカリフォルニアの暴走族）に補助輪が必要なようにね。　（PB59_00243）
(22)　＊～且―。安〈image description="二重線のダッシュ" /〉。（抑揚）　～でさえ―である。　（OT03_00030）

４．３．３　カタカナ

647 カ所で見られたので、100 カ所を検討した。これらの文の type は「quasi」と「verse」

なので、「カラー：ＢＫ（黒）」「ヤンレ　エエ」を文の一種と考えれば、誤検出はカタカナが顔文字の一部として使われている(25)のみで、それ以外は正しく検出できている。

(23)　〈sentence type="quasi"〉ローイング、プル系トレーニングに。ナイロンパッド付きなので　手首を保護します。カラー: ＢＫ (黒) 〈/sentence〉　(OY07_00073)
(24)　〈sentence type="verse"〉お伝たちまち縄目にかかる。ヤンレ　エエ〈verseLine /〉〈/sentence〉　(LBl2_00041)
(25)　廃人ＯＸです。へ(゜∀゜へ)アヒャ！　(OY03_09472)

４．３．４　数字

120 カ所見られたので、100 カ所を検討した。　“「」”“【】”のような括弧に含まれる注釈(26)(27)や、型番(28)を文と考えれば、抽出できている。

(26)　□発端は、米国最大の商戦期“ブラックフライデー”（感謝祭翌日の金曜日。２００８年は１１月２８日）だった。　(OY14_38930)
(27)　産業振興センター（☎３６０－３１９６で。【有料講座。４時間～、５，１６５円】▸パソコン入門（略）　(OP75_00001)
(28)　〈sentence type="quasi"〉お手入れも楽々。１２３６３２　ＡＴＬＩＵＭ／アトリウムランチョンマット　アイボリー〈/sentence〉　(OY04_01548)

４．３．５　アルファベット

147 カ所で見られたので、100 カ所を検討した結果、85 カ所でその後に文が続いていることが確認された。それ以外ではアルファベットが顔文字やアスキーアートの一部に使用されている。国立国語研究所(2011、p.79)には「サンプル作成時に削除された，いわゆる「アスキーアート」」とあるが、一行単位のアスキーアートは削除されていないようである。

(29)　〈sentence type="quasi"〉（略）外さなきゃならないんですｐ（´⌒`。Ｑ）グスン〈/sentence〉　(OY14_31617)
(30)　〈sentence type="quasi"〉□（Ｔ。Ｔ）〈/sentence〉　(OY15_10298)
(31)　。ｏ○☆＊゜¨゜゜・＊：．．　(OC14_04623)
(32)　初登場第３位キタ――（゜∀゜）⌒Ｙ⌒（。Ａ。）⌒Ｙ⌒（゜∀゜）⌒Ｙ⌒（。Ａ。）⌒Ｙ⌒（゜∀゜）――！！　(OY15_00312)

４．３．６　キリル文字

４カ所で見られ、全て確認したところ、キリル文字の後に文が続く例はなく、全て誤検索である。なお、(34)は一行のデータがこのままで、対応する〈sentence〉がない。

(33)　〈sentence type="quasi"〉⊂⌒￣⊃。Д。）⊃．．．．．．．．．．．．．。〈/sentence〉　(OC14_04563)
(34)　）(。ë ë)。〈/sentence〉　(OY02_00286)

４．３．７　記号

記号は(7)の“Å”から“＝”までの部分を指すとする。1,501 カ所で見られたので、100 カ所を検討した。正しく検出したのは 27 カ所で、(35)のように文タグの欠落した場所を検索できることもあるが、(36)のように流れ星のアスキーアートの一部になっている部分も検出してしまう。このような場合、誤りを含むものが多く、また句点が文ではないものの一部で使われていることが多いことが分かる。

(35) □「民事裁判では、訴える人を『原告』、訴えられた人を『被告』と呼ぶことになっています。『被告』イコール犯罪者という意味では、（略）　（PB13_00114）
(36)　。☆．．．　（OY11_06824）

以上から文字種により検索精度が異なり、ひらがな・漢字・カタカナ・数字などが句点に続く場合は文が続くと判断して問題ないが、記号が続く場合は検出例が多く、しかも高い確率で顔文字などのアスキーアートの一部を構成していることがわかった。

そこで、ひらがな・漢字・カタカナ・数字・記号の修正箇所数の合計が 500 以上の媒体を対象に、句点に後続する文字種を調べたのが表５である。ひらがな・漢字・カタカナ・数字を、正しく修正箇所が検出できる指標、記号を修正箇所の誤検出の指標とすると、PB・LB・OB は、ひらがな・漢字・カタカナ・数字の合計の割合が多く正しく検出できているといえる。それに対し OC と OY は記号の割合が多く誤検出の可能性が高いと推測される。

なお、句点に記号が後続する場合、その句点の前が文の終端でない可能性も考えられる。それらを確認するために、「句点＋記号」が前接する文字種について調査を行う。

表５．主要媒体別の句点に後続する文字の文字種とその割合

	ひらがな	漢字	カタカナ	数字	記号	合計
PB	1821	1658	273	12	94	3858
	47.2%	43.0%	7.1%	0.3%	2.4%	100.0%
LB	482	410	63	38	4	997
	48.3%	41.1%	6.3%	3.8%	0.4%	100.0%
OB	350	259	37	1	20	667
	52.5%	38.8%	5.5%	0.1%	3.0%	100.0%
OC	347	432	75	18	514	1386
	25.0%	31.2%	5.4%	1.3%	37.1%	100.0%
OY	469	638	143	32	779	2061
	22.8%	31.0%	6.9%	1.6%	37.8%	100.0%

４．３．８　句点＋記号に前接する文字種

句点＋記号に前接する文（文字列）が何によって終わっているか、文字種別にまとめたのが表６である。例えば(35)では、句点＋記号である“。『”には“す”というひらがなが前接していると考える。表６の記号 A は、(7)の“Å”から“＝”までの記号に、“。”“）”

“…”などを追加したもので、これらは文末を構成することがあるため追加した。

表６　句点＋記号に前接する文字の内訳

	ひらがな	漢字	カタカナ	数字	アルファベット	記号A	それ以外	合計
箇所数	474	42	16	2	6	726	235	1501

表６より、句点＋記号に前接する文字の半数近くが記号 A に含まれていることが分かる。これらは(37)(38)(39)のように顔文字のようなアスキーアートを構成していることが多く、文の終端とはなっていないのである。

(37)　〈sentence type="quasi"〉（＾。＾）～〈/sentence〉　（OC01_00522）
(38)　〈sentence type="quasi"〉（。・m・）クスクス〈/sentence〉　（OC01_00542）
(39)　（。-__-。）。　（OC01_01000）

表６の「それ以外」は(40)(41)のような例で、句点に前接するのは〈sentence〉というタグのみで、文字列がないものである。そもそもこのような文字が文頭に位置し、文のタグが付されること自体がおかしいということも指摘しておきたい。

(40)　〈sentence〉。寝過ぎですよね．．．〈/sentence〉　（OY14_24922）
(41)　〈sentence〉。．〈/sentence〉　（OY03_03891 他多数）

ひらがなのように、句点＋記号が後続する例も３割強あるが、半数近くを占めるのが記号 A の例である。そこで記号 A（726 例）が、どの媒体に多いのか調べた結果（表７）、記号＋句点＋記号が続く例は、圧倒的に OY に多く、8 割強であり、OC が 2 割弱である。PB の例は、(42)のようにかなり変わった例で、実際の紙面では１バイト文字［。-゜］であろう。

表7　媒体別の記号＋句点＋記号の出現数

媒体	PB	OT	OC	OY	合計
箇所数	4	19	133	570	726
割合	0.6%	2.6%	18.3%	78.5%	100.0%

(42)　１字以上の半角カタカナをワイルドカードで検索する場合は半角文字で“［ヲ-゜］｛１，｝”（半角カタカタの記号も含める場合は“［。-゜］｛１，｝”のようにします.（略）　（PB35_00023）

このように記号を多く含む OY や OC は、顔文字やアスキーアートであることが多いため修正対象にはなりにくいであろう。修正するならば、句点に文が後続する可能性の高い PB・LB・OB を対象にするのがよいであろう。

５．結論

以上のように、修正箇所の検出とサブコーパス・媒体による修正箇所数と出現状況を見

てきた。不必要な文タグを削除したり、タグのない部分につけたりする対象としては、出版サブコーパス（PB・PM・PN）や図書館サブコーパス（LB）が適していると思われる。PM・PN は修正箇所数が少なく、PB・LB は、修正箇所数は多いが真に修正が必要な箇所である可能性が高く、文タグの追加などの修正後、データとして用いることが可能である。逆に、特定目的サブコーパスは、OM・OL・OV のように修正箇所数が少ないものも多いが、OY や OC は修正箇所数が多い割には、その場所が文の終端ではない可能性も高く、さらに修正して文として一様に扱うためには、タグの追加だけではなく、タグの削除やデータ自体の修正が必要になり、困難が予想される。また、OY では、文の終端が様々で、文末を探すのが難しいという問題があり、例えば(43)(44)では“♡”や“♥”、“♫”“♫”などが文の終端に位置し、他に“w笑”なども見られた。

(43)　〈sentence type="quasi"〉傑作ポチを戴けるととても嬉しく思います♡　御協力に感謝申し上げます（o＊。__。）oペコッ〈/sentence〉　（OY14_28649）

(44)　〈sentence type="quasi"〉σ（・・；　早く元気になってねぇ〜♥　（略）　「カシャカシャ」って、ケーキなんかも作りたい♫　デジカメ持って散歩もしたい♫　私、欲張りか？？？（略）〈/sentence〉　（OY14_35446）

さらに、OY（Yahoo!ブログ）には文自体が切れているサンプルがあり、(45)は元々のブログが検索したサイトの一部を貼り付けたような形式で、途中で文が切れていて、文単位で取り出すこと自体無理なデータも含まれている。

(45)　（略）ホースです。大事に長く乗りたい方には必需品です。■■仕様変更によりグレードアッ...　（OY14_01602）

BCCWJ は出版サブコーパス、図書館サブコーパスそれぞれが生産実態、流通実態を反映するために作成されている。特定目的サブコーパスは、上記二つのサブコーパスでは「十分な分量が集まりにくい資料を中心に収録」（国立国語研究所(2011,p.16)）されているため、DVD 版を使用する場合、用例数を増やすなど安易な目的で三つのサブコーパスを同様に扱ってはならず、研究目的に合わせ、対象とするサブコーパスを慎重に選び、さらに修正を加えていくことが重要となってくる。

参考文献

国立国語研究所(2011)「『現代日本語書き言葉均衡コーパス』利用の手引　第 1.0 版」、BCCWJ-DVD 版収録

田野村忠温（印刷中）「BCCWJ の資料的特性――コーパス理解の重要性――」『講座日本語コーパス６　コーパスと日本語学』、朝倉書店、
（http://www.tanomura.com/temporary/bccwj_tanomura_2.pdf、2013 年 2 月 24 日取得）

長谷川守寿（印刷中）「BCCWJ の文構造タグに関する一考察」、『人文学報』第 488 号、首都大学東京

文末表現から発話冒頭の談話標識的表現へ

東泉裕子（東京学芸大学）
髙橋圭子（東洋大学）

From Sentence-final Expressions to Utterance-initial Discourse-pragmatic Expressions

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

１．はじめに

現代日本語の会話では、文末表現「だろ（う）」「でしょ（う）」（以下、「ダロウ」で代表させる）が、発話冒頭に現れたり、単独で現われたりすることがある。例えば、『現代日本語書き言葉均衡コーパス』（BCCWJ）（以下、『BCCWJ』と略す）には、以下のような用例がある[1]。

(1)「うーん、証拠かあ…。あいつ、口がうまいから、またいつもの調子でいいくるめようとするだろうなあ…」卓也は眉根を寄せて考えた。「だろう、それに、卓ちゃん、一本気だから、きっとけんかになってしまうと思うんだ。・・・」　（LBnn_00031、図書館・書籍、小林しげる『いつか青空』、KTC 中央出版、1999 年）

(2)「そうなんですか。―ううん。じゃあ確かに、完全なアリバイが成立しますね」「でしょ」「ということは残りの四人の中に犯人がいるって話ですかねえ」
（PB29_00573、出版・書籍、綾辻行人『どんどん橋、落ちた』講談社、2002 年）

文末表現がなぜこのような位置に現れるのであろうか。現代語の接続詞「だから」は「指示詞＋断定辞＋接続辞」から成る「それだから」から派生したもの（「指示詞をとらないφ形＋断定辞＋接続辞」）だという（矢島 2011）。また、「だって」も「それだっても」から発達したということである（Mori 1996）。同様に、文末以外の「ダロウ」は「そうだろ（う）」「そうでしょ（う）」（以下、「ソウダロウ」）の「指示詞をとらないφ形＋断定辞」と考えることができるのではないだろうか。そこで、本稿では近代語のコーパス（『太陽コーパス』および『近代女性雑誌コーパス』）と現代語のコーパス（BCCWJ）に見られる「ソウダロウ」と文末以外の位置に現われる「ダロウ」の用例を比較することにより、「ダロウ」の文末から文末以外への用法拡張の様相を観察することにする[2]。

２．先行研究

２.１　現代日本語の「ダロウ」

現代日本語の「ダロウ」は、基本的に、1)話し手の推量を表す、2)相手に同意・確認を求めるのに用いられる文末形式とされる（『日本国語大辞典 第二版』、日本語記述文法研究会編 2003 など）。しかし、上の例(1)や例(2)にも見られるように、現代語では文末以外の位置にも生じている。このような「ダロウ」の文末以外の位置における用法に関する研究は管見ではわずかである[3]。

---

1 『BCCWJ』の用例には、サンプル ID、レジスター、出版年を付記する。さらに、書籍には執筆者、書名、出版社も記す。

2 本稿では「のだろう」「んでしょう」など「の（ん）」に後続する「ダロウ」は調査対象に含めない。

3 ただし、『日本国語大辞典 第二版』の用例に、例(2)のような単独の「ダロウ」の例がある（芥川龍之介

杉浦(2012)は、会話分析の立場から相互行為的文脈における「ダロウ」について論じている。そして、「ダロウ」は「評価−（非）同意」という行為連鎖の「（非）同意」の位置に生じ、「直前の話者の見方・判断」への応答として利用されると指摘している。さらに、杉浦(2013)は、マルチモーダル分析の立場から「ダロウ」単独ターンによって示される「同意」の強さについて詳しく論じている。

Higashiizumi and Onodera (2013)は、歴史語用論の立場から、会話に現われる「ダロウ」の生起位置の変化を調査した。現代日本語研究会編（1997, 2002）のコーパス（以下、『現代日本語』）から用例を集め、「ダロウ」が文末以外にも、例(1)のように発話冒頭に現われたり、例(2)のように単独で現われたりすることがあるということを確認した。また、次の例(3)のように、「でしょ」の前にポーズがあることを示す読点がある用例もあった[4]。

(3) [食事時の飲み物についての話]
B: ヨーロッパ行ったらね。
A: ワイン、でしょ↑
B: ヨーロッパ行ったらワインでね、（うん　Inf(女)）水をの、くれってゆうのはね、アメリカ人と日本人ぐらいだって、ヨーロッパ人はゆうんですけどー。
（現代日本語研究会編(1997) 6921–6923 行改変）

以上を踏まえ、『太陽コーパス』（以下、『太陽』）の「口語」「会話」における「ダロウ」の生起位置を調査したところ、例(1)〜(3)のような用例は『太陽』のデータには皆無であることがわかった。また、それぞれのコーパスにおける「ダロウ」全用例数に対する「ソウダロウ」の用例数の比率を比べると、現代語の会話ではわずかながら下がっていた（後述の表２参照）。

ただし、この研究は、現代語の会話の用例と近代語の雑誌の口語体会話の用例という異なるレジスターにおける比較であるため、同じレジスターでの用例の比較が必要である。そこで、本稿では、現代語と近代語における小説の会話部分に見られる「ダロウ」の用法を量的・質的に分析する。

２.２　歴史語用論における談話標識の研究[5]

ことばの使われ方が時とともにどのように変化したのかを問う分野である歴史語用論において、談話標識の歴史的発達の研究は主要なトピックのひとつとして定着しつつある（高田他 2011:30）。最近では、どの標識がどの位置に現れ、どのような機能を果たしているかという観点から議論されるようになってきている（高田他 2011:88–89、藤井 2013 など）。本研究はこれらの議論と軌を一にするものである。

1919『路上』、島崎藤村 1910〜11『家』）。

4 例(3)は、「でしょ」の前にポーズがあることを示す読点があることから、先行する文と一続きに発話される通常の文末表現「でしょう」とは異なる用法であると考えられる。

5 「だろう」は「に−て−あら−む」から変化した「で−あら−う」が変化した語で、「だろう」という形は江戸時代後期の江戸語に現れる（『日本国語大辞典 第二版』）。天明・寛政期（1781–1801年）に定着し、洒落本では聞き手への「確認要求の表現」が多いという（鶴橋2013:42–46）。一方、「だろう」の丁寧体「でしょう」は、断定の助動詞「です」の未然形に推量の助動詞「う」がついたもので、「でせう」の形で近世末期の文献で使用されている（『日本国語大辞典 第二版』）。「です・でした・でしょう」と変化する「活用のあるデス」は幕末の江戸の町人言葉から始まり、庶民向けの新聞や言文一致小説などで次第に一般していくが、1904年から使用された最初の国定教科書により全国に広まったという（田中2012: 174-178）。

３．量的調査

今回の調査では、Higashiizumi and Onodera (2013)で使用した『太陽』における「ダロウ」の用例と比較するために、『近代女性雑誌コーパス』（以下、『女性雑誌』）のデータのうち「口語」体の「会話」を検索対象とした。また、『BCCWJ』では「出版・書籍」「図書館・書籍」「特定目的・ベストセラー」[6]の「文学」の会話およびそれに準ずる部分で使用されている「ダロウ」の用例を集めた。『太陽』とともに、今回の調査に用いたコーパスをまとめると表１になる。

表１

コーパス	『太陽』	『女性雑誌』	『BCCWJ』
検索ツール	全文検索システム『ひまわり』ver.1.3.1	全文検索システム『ひまわり』ver.1.3.1	中納言 1.1.0
検索対象	文体：口語 種別：会話	文体：口語 種別：会話	出版・書籍 図書館・書籍 特定目的・ベストセラー
ジャンル	–	–	文学
期間	1895–1925	1894–1925	1971–2008

今回の調査対象である『女性雑誌』および『BCCWJ』における「ダロウ」の用法に、Higashiizumi and Onodera (2013)の調査結果を加えたものが表２である。『太陽』と同様、『近代女性雑誌』にも例(1)〜(3)のような文末以外の用例は皆無であった。

表２

	近代語（1894–1925 年）				現代語（1971–2008 年）			
	『太陽』（小説の会話）		『女性雑誌』（小説の会話）		『BCCWJ』（小説の会話）		『現代日本語』（話し言葉の会話）	
文末表現のダロウ（ソウダロウを除く）	1817	97.5%	326	95.9%	37789	97.4%	430	93.1%
ソウダロウ	47	2.5%	14	4.1%	808	2.1%	10	2.2%
文末表現以外のダロウ（例(1)-(3)）	0	0.0%	0	0.0%	192	0.5%	22	4.8%
合計	1864	100.0%	340	100.0%	38789	100.0%	462	100.0%

表２から「ダロウ」全用例数に対する「ソウダロウ」の用例数の比率は、わずかではあるが、現代語において減少傾向にあることが分かる。また、『BCCWJ』と『現代日本語』における「ソウダロウ」の用例数の比率もほぼ同程度であった。

文末以外の用法（「指示詞をとらないφ形＋断定辞」）は、『BCCWJ』のデータにも見られ、話し言葉の会話だけでなく、小説の会話でも使用されており、ある程度の広がりと

6 『BCCWJ』のデータは「コア」と「非コア」に分けられている。今回の調査ではコア・非コアを問わず、「出版・書籍」「図書館・書籍」「特定目的・ベストセラー」の３つジャンルの全データを対象にした。

定着を示していることが分かった。しかし、今回調査したコーパスでは、文末以外の用法の発生プロセスを時系列的に論じられるようなデータを得ることができなかった。今後、1920 年代から 1970 年代の間の 50 年間のデータを調査する必要があると考える。

４．質的分析

「ダロウ」の文末から文末以外の用法への拡張の様相の一端を捉えるために、ここでは「ソウダロウ」と「ソウ」を伴わない単独の「ダロウ」がどのような文脈で使われているか観察したい。

まず、「ソウダロウ」は、次の例に示すように、先行文脈の相手の発話や話者自身の発話を指している。

(4) 「わたしにしゃべられては困ったのかしら」　「たぶん、そうだろう」　「ってことは、関原さんは学君がわたしに何を話そうとしていたか、うすうす勘づいていたというの？」　「かもしれない」（LBn9_00208、図書館・書籍、新津きよみ『血を吸う夫』、角川春樹事務所、1999 年）

(5) 「だが、あなたがどれほど文句を言っても、結婚の準備は急ピッチで進んだ。浜田さんはそういう流れに逆らいはしなかった。そうでしょう？　彼は結婚をとりやめるつもりなんかなかった。　聡美さんと一緒に私に会いにきたとき、浜田さんはとても幸せそうでした」（OB6X_00071、特定目的・ベストセラー、宮部みゆき『誰か』、実業之日本社、2003 年）

次に、「ソウ」などの指示詞を伴わない単独の「ダロウ」のうち、読点に後続するものは、先行する話者自身の発話が、相手の発話の繰り返しや代弁である例がある。

(6) 「かわいいから染めちゃったらしいよ、ピンクに」　「げ」　「げ、だろ？」（PB39_00297、出版・書籍、松りんこ『誰よりいちばんっ！』、光彩書房、2003 年）

(7)「表の車の中さ。居眠りでもしてなきゃいいけどな」　リカが母親に会いに行ったりすることを考えて、一応道田に見張らせてあるのだ。　（中略）　「もし道田君が眠ってたら…」　「叩き起こせ、だろ？」　「起こさなくてもいいわよ」と、真弓は言った。 (OB3X_00137、特定目的・ベストセラー、赤川次郎『盗みに追いつく泥棒なし』、徳間書店、1987 年)

(8) 「女刑事はあんたがはじめてでね」　「困ったもんだ、でしょ？　でも世界はめまぐるしく動いてるのよ。すぐにわたしのような女も珍しくなくなるわ」　彼は少し笑い、それから顔をしかめた。（LBj9_00115、図書館・書籍、クレイグ・ホールデン(著)/近藤　純夫(訳)『リバー・ソロー』、扶桑社、1995 年）

例(6)〜(8)のような「ダロウ」は、指示詞を伴っていないが、相手の発話を繰り返したり、相手が伝えようとしている内容を話者が代弁したりして、先行文脈の発話を指していると考えると、「指示詞をとらないφ形＋断定辞」と似たような働きをしていると考えることができないだろうか。

最後に、単独や発話冒頭の「ダロウ」の中には、杉浦（2012, 2013）のいう「評価−（非）同意」という連鎖の「（非）同意」位置に生じ、「直前の話者の見方・判断」への応答としなっている、次のような例もある。

(9) 「それがね、残業なんかしてたから明細見てびっくりしちゃった。手取りで二十五

万もあったんだよ」　「それは悪くないね」　「でしょう」　「うん」　「なんだか、私すこし自信が出てきちゃった。こうやって、自分の力でこれから生きていけるんじゃないかって」（LBr9_00215、図書館・書籍、白石　一文『一瞬の光』角川書店、2003 年）

(10)「おれは今日、紫原の兄貴分をぶちのめした。（中略）そんなときに、同じ組の極道に手を出してみろ」　「それはヤバいね」　「だろう」　「だけどさ、やられっぱなしはよくないよ」（LBr9_00134、図書館・書籍、馳星周『虚の王』光文社、2000 年）

(11) 晶は唇を尖らせた。　「そんなことできるわけないじゃん」　「だろう。お前が心配すべきなのは、生活の問題じゃなくて、自分たちの音楽の変質じゃないのか。（略）」（OB5X_00248、特定目的・ベストセラー、大沢在昌、『氷舞』、光文社、1997 年）

このような「ダロウ」も「指示詞をとらないφ形」で「直前の話者の見方・判断」を指し、「指示詞をとらないφ形＋断定辞」として現れていると思われるが、詳しい分析と考察は今後の課題としたい。

５．まとめと今後の課題

文末に位置し、話し手の推量および聞き手への確認・同意要求を示すとされる「ダロウ」が、現代語の小説の会話でも、発話冒頭で使われたり、単独で現れたりしており、文末以外の位置へとその用法を拡張している様相の一端を観察した。そして、このような用法拡張のプロセスには「指示詞をとらないφ形＋断定辞」という形式が関わっている可能性があることも指摘した。これは、「だから」「だって」などの接続詞の談話標識用法の発達や、「結果」などの漢語名詞の副詞用法への拡張（東泉・髙橋 2013、髙橋・東泉 2013）などと似たような拡張のプロセスのように思われる。「共話」（水谷 2001）[7]的な文脈において、相手の発話や話者自身の発話の命題を「指示詞をとらないφ形＋断定辞」である「ダロウ」で指して自分の発話に組み込むことによって、文末以外の位置にも現われる~~表れる~~ようになったのではないだろうか。

文末表現「ダロウ」の文末以外の用法への拡張は現在進行中の現象であり、本稿では近代語と現代語のコーパスの小説の会話における用例を比較して、拡張の様相の一端を捉えようと試みた。今後は文末以外の用法の発生プロセスを時系列的に調査する必要もあると思われる。

謝　辞

本研究の一部は、第 13 回国際語用論学会（International Pragmatics Association）のパネル・セッション Cross-Linguistic Approach to Form–Function–Periphery (LP and RP) Mapping: With Special Focus on “Exchange Structure” and “Action Structure”において発表したものです。ご助言くださった方々に感謝申し上げます。

文　献

現代日本語研究会編(1997)『女性のことば・職場編』、ひつじ書房
現代日本語研究会編(2002)『男性のことば・職場編』、ひつじ書房

7 「共話」とは「聞き手があいづちを頻繁に打ちながら（中略）話し手と聞き手が対立せず一体になって話す形」のことで、「対話」と対立するものである（水谷 2001:126）。

杉浦秀行(2012) 「『でしょう』を用いた評価への応答–行為と「なわ張り」の接点–」第30回社会言語科学会研究大会ワークショップ「相互行為の中で知識を主張すること–会話分析から見た「情報のなわ張り」–」発表
杉浦秀行(2013)「『でしょう』単独ターンによって表明される同意の強さ–マルチモーダル–分析によるアプローチ–」『JCLA Conference Handbook 2013』、pp.52–55、日本認知言語学会
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・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予稿集』、pp.195-202.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no4_papers/JCLWorkshop_No4_24.pdf)
田中章夫(2012)『日本語雑記帳』、岩波書店
鶴橋俊宏(2013)『近世語推量表現の研究』、清文堂
『日本国語大辞典 第二版』、小学館、ジャパンナレッジ版
日本語記述文法研究会編(2003)『現代日本語記述文法４　第８部　モダリティ』、くろしお出版
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)
藤井聖子(2013)「条件構文の談話標識化の諸相」『第４回コーパス日本語学ワークショップ予稿集』、pp.27-34.
(http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no4_papers/JCLWorkshop_No4_04.pdf)
水谷信子(2001)『続日英比較話しことばの文法』、くろしお出版
矢島正浩(2011)「時間的・空間的比較を軸にした近世語文法史研究–ソレダカラ類の語彙化を例として–」『近代語研究のパースペクティブ』、pp.56-82、笠間書院
Higashiizumi, Yuko, and Noriko O. Onodera (2013) Form-Function-Periphery Mapping in Japanese: ‘Exchange-Structure’ and ‘Action-Structure’. *Abstracts*, pp.105–106. The 13th International Pragmatics Conference, New Delhi, India, 8-13 September 2013
Mori, Junko (1996) Historical Change of the Japanese Connective *Datte*: Its Form and Functions. In *Japanese/Korean Linguistics. 5.* pp. 201–218.

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』（BCCWJ）
国立国語研究所『近代女性雑誌コーパス』
国立国語研究所(2005)『太陽コーパス』（国語研究所資料集 15）　博文館新社

BCCWJ における敬語形式の使用実態

小木曽智信（国立国語研究所言語資源研究系）†
神田龍之介
近藤明日子（国立国語研究所コーパス開発センター）

Survey on the Usage of Honorific Forms in the BCCWJ

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Ryunosuke Kanda
Asuko Kondo (National Institute for Japanese Language and Linguistics)

１．はじめに

現代日本語には敬語形式のバリエーションが多く、尊敬語では「－なさる」「御[1]－なさる」「御－になる」「御－だ」「られる」、謙譲語では「御－する」「御－いたす」「御－申し上げる」等の形式が存在する。しかし、これらの敬語形式が、実際にどのような頻度で用いられているのか、どのような要因によって選択されているのかは十分に明らかでない。コーパスを用いた大規模な用例調査も行われていない。

本発表では、『現代日本語書き言葉均衡コーパス』（BCCWJ）を用いて、尊敬語・謙譲語Ｉの形式を中心に、(1) 使用域（レジスター），(2) 命令形か否か，(3) 元となる動詞語彙の違い，の３つの観点から使用実態の全体像を調査する。また、やりもらい・恩恵を意味に含む「－いただく」「－くださる」については、「－いただく」と「－くださる」の使い分け，「検討いただく」「応募ください」のような「御」の不使用，などの問題があるため、追加して、使用実態の調査・記述を行う。

以上の基礎調査を通して、敬語形式の選択要因を探り、今後の研究の足がかりとする。

２．現代日本語の敬語形式と調査対象

現代日本語における主要な敬語形式としては「－なさる」「御－になる」、「御－する」「御－いたす」などの文法的な形式や、助動詞「れる」「られる」の尊敬用法、「召し上がる」「拝読する」のような語彙的な敬語が挙げられる。

本発表ではこのうち、文法的な（非語彙的な）形式に注目し、表 1 に挙げた諸形式を調査対象とした。なお、助動詞「れる」「られる」も文法的な形式の一つであるが、受身・可能などの他の用法を持ち､その区別が BCCWJ ではアノテーションされていないため対象外とした[2]。

表 1　調査対象の敬語形式

タイプ	形式
尊敬語	－なさる
	御－なさる
	御－になる
	御－だ
	御－くださる
	－てくださる
謙譲語	御－する
	御－いたす
	御－申し上げる
	御－いただく
	－ていただく

† toigso@ninjal.ac.jp

1 「御－なさる」は「－」部分の語の語種に応じて「お－なさる」と「ご－なさる」の両形式を含む。他の「御」を含む形式も同様である。

2 BCCWJ における「れる」「られる」の用法分類については、小山田・柏野・前川（2012）によるコアデータ（及び国会会議録の一部）の調査がある。この調査結果に基づき３．４節で他形式との比較を行う。

３．敬語形式の使用頻度

３．１　調査方法

表 1 に挙げた形式の用例を BCCWJ から取得して、各形式の使用頻度を集計した。用例の調査は、

1)　BCCWJ の短単位を格納したデータベースから調査対象の用例を取得

2)　ダウンロードしたデータを表計算ソフトにインポート後、手作業でクリーニング

という手順で行った。以下、それぞれの方法の詳細を述べる。

1) データベースからの用例の取得

BCCWJ のデータが格納された「形態論情報データベース」（小木曽・中村 2011）の短単位テーブルに対し SQL 文を発行してデータを取得した。この短単位データは BCCWJ DVD 版の短単位データに相当する。SQL 文の一例として「御－なさる」の例を挙げる[3]。

```
SELECT register, coreFlag, sampleID,
dbo.fnGetContextPre(c.orderID, c.sampleID, 30) AS 前文脈, c.orthToken AS キー,
dbo.fnGetContextPost(c.orderID, c.sampleID , 30) AS 後文脈,
c.lemma, c.lForm, c.formBase, c.pos, c.goshu, c.cForm, c1.cForm AS 形式活用形
FROM corpus AS c
LEFT JOIN corpus as c_1 ON c_1.sampleID = c.sampleID AND c_1.orderID = c.orderID-10
LEFT JOIN corpus as c1 ON c1.sampleID = c.sampleID AND c1.orderID = c.orderID+10
WHERE c_1.lemma = '御' AND c1.lemma = '為さる'
```

このように、レジスターやサンプル ID などの出典情報と、出現箇所の KWIC、元となる動詞（サ変動詞の場合は語幹）部分の形態論情報、さらに形式全体の活用形を取得している。

2) データのクリーニング

SQL 文で取得したデータを表計算ソフトにインポート後、次のような調査対象外の例を手作業で除外した。

・調査対象以外の表現が外形上、偶然一致したもの
・「－」部分が動詞ではないと考えられるもの
　・「－」部分がサ変動詞語幹にならない語であるもの
　・「－」部分が連体修飾を受けるもの
・「御－」が丁寧語や美化語であると考えられるもの，固定的な言い回しとなっているもの
　　例：「お辞儀する」「お掃除する」「お稽古する」「お酌する」「おめかしする」「おみそれする」

このほか、次のような用例も集計から除外した。

・形態素解析の誤解析にもとづくもの
・敬語法に言及しているメタ的な用例

なお、現時点では、クリーニングに不十分な点があり、今後数値が変動する可能性がある。

３．２　各形式のレジスター別粗頻度

上記の方法によって確認した各形式のレジスター別の用例数（粗頻度）を表 2 に示す。総計約 14 万例、尊敬語が約 8.6 万例、謙譲語が約 5.4 万例 であった。レジスター別には OL（法律）に全く現れないことが目を引くが、敬語形式が法律文になじまないことは自明であろう。

[3] dbo.fnGetContextPre 及び dbo.fnGetContextPost は文脈を取得するためのユーザ定義関数である。なお、わかりやすくするために列名の一部を変更した。

表 2　各形式のレジスター別用例数（粗頻度）

		LB	OB	OC	OL	OM	OP	OT	OV	OW	OY	PB	PM	PN	総計
尊敬語	てくださる	6116	1081	29199	0	774	3785	49	13	12	4938	6320	775	91	53153
	なさる	2985	629	1420	0	371	13	538	7	1	1030	2461	181	22	9658
	御－くださる	1044	208	1572	0	181	5328	7	3	9	1785	827	221	36	11221
	御－だ・です	697	90	1481	0	437	35	4	1	0	464	623	177	7	4016
	御－なさる	601	107	72	0	81	2	1	3	0	243	294	17	0	1421
	御－になる	1895	448	418	0	1916	68	11	0	1	221	1101	97	7	6183
謙譲語	ていただく	2079	401	2288	0	4549	1085	12	0	5	1902	1974	340	56	14691
	御－いたす	336	69	1534	0	1906	37	1	0	0	973	223	27	0	5106
	御－いただく	456	90	289	0	1705	410	2	0	0	567	437	78	6	4040
	御－する	3718	685	9929	0	4323	1371	20	4	15	4381	2988	682	73	28189
	御－申し上げる	288	54	49	0	1350	56	1	0	0	186	166	14	1	2165
	申し上げる	54	7	2	0	79	14	2	0	0	29	34	2	0	223
総計		20269	3869	48253	0	17672	12204	648	31	43	16719	17448	2611	299	140066

３．３　命令形の除外

尊敬語の「－なさる」「－くださる」は命令形の用例が極めて多い。「－なさい」（「しなさい」）が直接的な命令形（「しろ」「せよ」）の代用として用いられたり，依頼表現として「－ください」が用いられたりするためである。これらの用法は、形式上は尊敬語の形と一致するものであるが、すでに尊敬語とは異質なものになっていると考えられる。

形式別に見た命令形の分布は表 3 の通りである。表 2 の用例数をもとに命令形の割合を計算すると、「てくださる」の 90.79%、「なさる」の 86.04%、「御－てくださる」の 91.77%、「御－なさる」の 64.60%が命令形である。尊敬語のこれら 4 形式全体の 89.83%が命令形となっている。

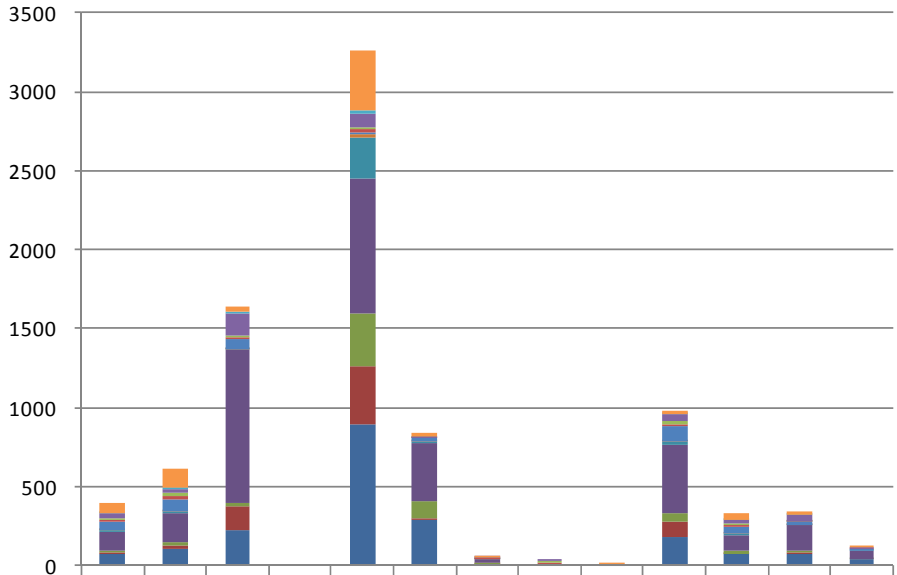
表 3　各形式の命令形の用例数

		LB	OB	OC	OL	OM	OP	OT	OV	OW	OY	PB	PM	PN	総計
尊敬語	てくださる	4619	811	28623	0	683	3733	44	13	12	3913	5032	691	81	48255
	なさる	2434	533	1217	0	251	12	536	6	1	985	2159	157	19	8310
	御－くださる	747	142	1496	0	123	5305	7	1	9	1592	623	217	36	10298
	御－なさる	359	77	48	0	4	2	1	3	0	219	193	12	0	918
謙譲語	御－いたす	1	2	0	0	0	0	0	0	0	0	2	0	0	5
	御－する	11	3	0	0	1	0	0	1	0	2	3	1	1	23
	御－申し上げる	1	1	0	0	0	0	0	0	0	0	0	0	0	2
総計		8172	1569	31384	0	1062	9052	588	24	22	6711	8012	1078	137	67811

このように、尊敬語形式の命令形は極めて数が多く、本来の敬語としての機能を失ったものであると考えられることから、敬語形式の分布を調査する上では、命令形を一括することには問題がある。そこで、以下の議論では、命令形の敬語形式は一律に除外することにする。これにより、用例数は総計約 7.2 万例、尊敬語が約 1.8 万例、謙譲語が約 5.4 万例 となる。

３．４　各形式のレジスター別調整頻度

表 1 の敬語形式の非命令形の用例数（表 2 の用例数から表 3 の用例数を減じたもの）を、各レジスターの総語数（記号等を除く）で除し、調整頻度（100 万語あたりの用例数）にしたものが図 2 である。



	LB	OB	OC	OL	OM	OP	OT	OV	OW	OY	PB	PM	PN
尊敬語 御ーになる	62	120	41	0	376	18	12	0	0	22	39	22	5
尊敬語 御ーなさる	8	8	2	0	15	0	0	0	0	2	4	1	0
尊敬語 御ーだ・です	23	24	144	0	86	9	4	4	0	46	22	40	5
尊敬語 御ーくださる	10	18	7	0	11	6	0	9	0	19	7	1	0
尊敬語 なさる	18	26	20	0	24	0	2	4	0	4	11	5	2
尊敬語 てくださる	49	72	56	0	18	14	5	0	0	101	45	19	7
謙譲語 申し上げる	2	2	0	0	15	4	2	0	0	3	1	0	0
謙譲語 御ー申し上げる	9	14	5	0	265	15	1	0	0	18	6	3	1
謙譲語 御ーする	122	182	968	0	847	365	22	13	3	430	105	153	53
謙譲語 御ーいただく	15	24	28	0	334	109	2	0	0	56	15	18	4
謙譲語 御ーいたす	11	18	150	0	374	10	1	0	0	95	8	6	0
謙譲語 ていただく	68	107	223	0	892	289	13	0	1	187	69	76	41

図 1　各形式のレジスター別調整頻度（100 万語あたりの用例数）

敬語形式の出現頻度はレジスター差が極めて大きい。敬語形式が現れやすいものから順に、OM（国会会議録）、OC（Yahoo!知恵袋）、OY（Yahoo!ブログ）、OP（広報誌）、OB,LB,PB（書籍）となっている。敬語は対者的な表現であるから、OM のような討論や、OC のような問答で現れやすいことは自然である。

全体の頻度だけでなく、使用される形式にもレジスター差が存在する。謙譲語では、OC では「御ーする」が多く、「ていただく」が少ない。尊敬語では、OM で「御ーになる」が多く「てくださる」が少ない、PM（雑誌）や OC で「御ーだ・です」が多い、OY では「御ーになる」が少ない。

ラレルの尊敬用法

本発表では、助動詞「れる」「られる」の尊敬用法は調査対象外としたが、小山田・浅原・前川（2012）は、BCCWJ のコアデータ（および国会会議録と「日本語話し言葉コーパス」CSJ の一部）における「れる」「られる」のレジスタごとの用法別用例数を調査している。この結果（「表 6 レジスター別の意味判定結果」）をもとに，助動詞「れる」「られる」の尊敬用法の調整頻度（100 万語あたり頻度）を計算すると表 4 のようになる。

この数字は、今回調査した敬語形式と比較して非常に使用頻度が高く、調査対象形式の全体を上回る。今回の調査範囲では「れる」「られる」を扱わなかったことで尊敬語が大幅

に欠落していることを確認しなければならない。

表 4　助動詞「れる」「られる」尊敬用法の調整頻度（100 万語あたりの用例数）

※小山田・浅原・前川（2012）にもとづく試算

	OC	OW	OY	PB	PM	PN	コア全体
「れる」「られる」尊敬用法	6345	25	2674	1034	806	136	140

３．５　尊敬語・謙譲語とレジスター

図 1 の数値を元に、レジスターごとに謙譲語・謙譲語がどのような割合を占めるかをまとめた結果が図 2 である。

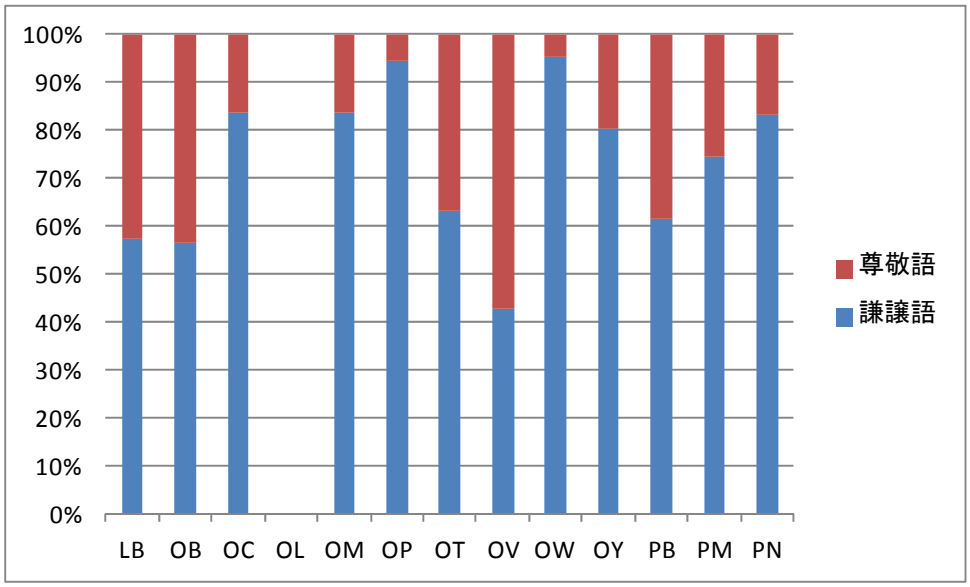


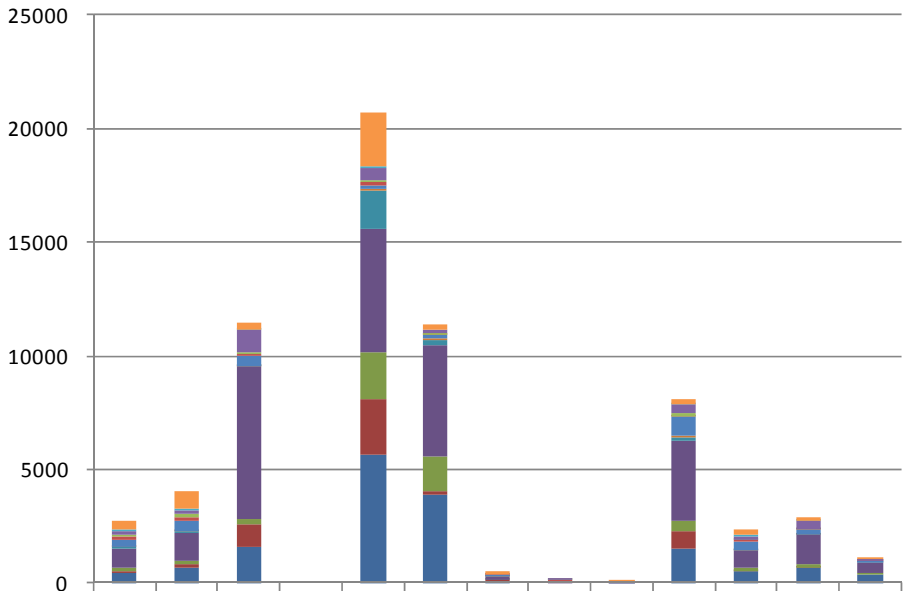
図 2 レジスター別の尊敬語・謙譲語の割合

敬語の使用が多い OM（国会会議録）や OC（Yahoo!知恵袋）、OY（Yahoo!ブログ）では、謙譲語が 8 割程度となっている。書籍は OB（ベストセラー），PB（出版），LB（図書館）ともほぼ同じで謙譲語が 6 割前後となっている。

３．６　動詞数あたりの頻度（敬語形式化率）

３．４節ではコーパスの総語数（短単位数）を用いて調整頻度を計算した（図 1）。しかし、今回調査対象としている敬語形式は、動詞を元として成り立つものであるから、「全ての動詞」の用例数を基準にして調整頻度を計算することで、いわば「敬語形式化率」を測ることができると考えられる（ただし語彙的な敬語や「れる」「られる」が敬語の用例として含まれていない）。レジスターによって総語数に占める動詞の割合は大きく違う。動詞以外の語が多く含まれているレジスターでは、敬語形式化率は総語数による調整頻度よりも小さくなるはずである。

このような考えから、各形式の動詞数 100 万語あたりの頻度をまとめたものが図 3 である。図 1 と比較したとき特に OP（広報誌）の数値の伸びが顕著である。図 1 では、OM > OC > OY > OP > OB の順に頻度が高かったが、図 3 では、OM > OC > OP > OY > OB と順も入れ替わっている。



	LB	OB	OC	OL	OM	OP	OT	OV	OW	OY	PB	PM	PN
尊敬語 御ーになる	434	787	284	0	2392	246	89	0	2	178	279	184	47
尊敬語 御ーなさる	55	53	16	0	96	0	0	0	0	19	26	9	0
尊敬語 御ーだ･です	159	158	1007	0	545	126	33	26	0	373	158	335	47
尊敬語 御ーくださる	68	116	52	0	72	83	0	52	0	155	52	8	0
尊敬語 なさる	126	169	138	0	150	4	16	26	0	36	76	45	20
尊敬語 てくださる	343	474	392	0	114	188	41	0	0	825	326	159	68
謙譲語 申し上げる	12	12	1	0	99	51	16	0	0	23	9	4	0
謙譲語 御ー申し上げる	66	93	33	0	1685	202	8	0	0	150	42	26	7
謙譲語 御ーする	848	1198	6750	0	5395	4954	163	79	29	3524	755	1288	488
謙譲語 御ーいただく	104	158	196	0	2128	1482	16	0	0	456	111	148	41
謙譲語 御ーいたす	77	118	1043	0	2379	134	8	0	0	783	56	51	0
謙譲語 ていただく	476	704	1556	0	5678	3921	98	0	10	1531	499	643	380

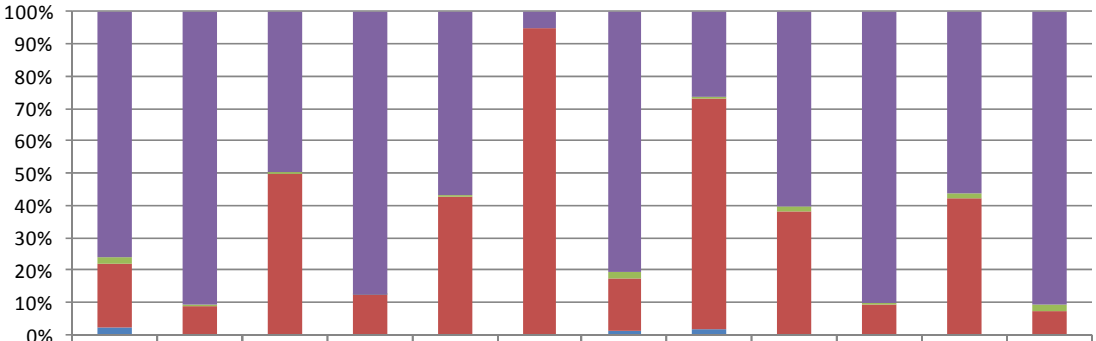
図 3　各形式のレジスター別調整頻度（動詞 100 万語あたりの用例数）

４．　元となる動詞の語彙と敬語形式

４．１　敬語形式と元の動詞の語種

各敬語形式がどのような動詞をもとにして成り立っているのかを語種の観点から調査した。例えば「御ーなさる」であれば、「お待ちなさる」のように「ー」部分が和語動詞の場合は和語とし、「御理解なさる」のようにサ変動詞の場合には、語幹（BCCWJ では名詞）の語種によって分類した。結果を図 5 に示す。グラフは語種割合を表示している。

外来語のように、「御ー」が着く場合に利用できない形式もあるが、和語・漢語についてはいずれの形式も可能である。しかし、実際の用例数は形式ごとに大きな違いがある。特に和語の割合が高いものは、謙譲語「ーていただく」「御ーいたす」「御ーする」、尊敬語「ーてくださる」「御ーだ・です」「御ーになる」である。逆に、謙譲語「ー申し上げる」、尊敬語「ーなさる」は漢語の割合が高い。



	ていただく	御ーいたす	御ーいただく	御ーする	御ー申し上げる	申し上げる	てくださる	なさる	御ーくださる	御ーだ・です	御ーなさる	御ーになる
	謙譲語						尊敬語					
和	11187	4616	2012	24726	1224	11	3943	355	557	3627	283	5609
混	272	20	6	1	14		89	10	13	13	8	115
漢	2915	465	2022	3439	925	212	803	961	353	376	212	459
外	317						63	22				

図 4　敬語形式と元の動詞の語種

4．2　敬語形式と元の動詞語彙

続いて、各敬語形式がどのような動詞をもとにして成り立っているのかを、動詞語彙（見出し語）ごとに集計した。動詞異なり語数は極めて多いため、形式ごとに頻度上位 30 語を掲げる（表 5，表 6）。表中の名詞はサ変動詞語幹として使われていたものである。

表 5　尊敬語の元の動詞（頻度上位 30 語）

順位	てくださる		なさる		御ーくださる		御ーだ・です		御ーなさる		御ーになる	
1	来る	371	居る	44	許す	53	勧める	1846	出でる	68	考え	465
2	教える	323	心配	42	出でる	36	持つ	389	休む	32	出でる	331
3	する	309	来る	35	待つ	34	分かる	318	心配	24	分かる	249
4	言う	175	御免	30	了承	28	出でる	211	苦労	22	見える	246
5	居る	113	結婚	26	越す	26	考える	209	返る	19	聞く	206
6	与える	85	言う	25	付き合う	18	有る	206	結婚	17	指摘	196
7	見る	78	する	24	教える	16	苦労	112	話す	14	遣る	169
8	送る	70	苦労	17	助ける	16	過ごす	106	利用	9	亡くなる	162
9	成る	68	為る	16	守る	15	承知	90	検討	9	帰り	153
10	読む	62	行く	15	賛同	15	呼ぶ	43	相談	9	書く	133
11	掛ける	60	御覧	15	読む	13	希望	30	指摘	7	持つ	122
12	応援	60	相談	14	伝える	13	思う	29	有る	5	思い	111
13	答える	58	反対	11	指導	12	探す	28	待つ	5	読み	106
14	聞く	58	約束	10	与える	12	忘れる	24	入る	5	話す	101
15	仰る	58	判断	10	可決	12	使う	16	答弁	5	使う	98
16	書く	52	答弁	10	参加	11	記憶	15	控える	5	会う	91
17	遣る	51	対処	10	回答	10	聞く	14	聞く	4	答え	88
18	話す	48	勉強	10	協力	10	聞き及ぶ	14	安心	4	出し	76
19	行く	46	見る	9	容赦	10	執心	14	遠慮	4	出掛ける	71
20	回答	43	利用	9	利用	9	待つ	13	勉強	4	尋ねる	64
21	付ける	40	質問	8	確認	9	利用	13	話	4	止める	62

22	紹介	38	出品	8	貸す	9	泊まる	12	遣る	3	触れ	62
23	喜ぶ	37	要求	8	話す	9	困る	12	出席	3	決める	59
24	入れる	36	遠慮	8	楽しむ	9	成る	11	許す	3	知る	52
25	為る	35	研究	8	尽力	9	立腹	10	会う	3	作る	50
26	説明	31	入院	8	送る	9	気付く	9	約束	3	調べ	48
27	取る	30	再婚	8	過ごす	9	住まう	8	渡す	3	戻る	45
28	見せる	30	出席	7	任せる	9	決まる	8	判断	3	認める	45
29	案内	30	我慢	7	教示	9	所望	8	仕事	3	休み	45
30	届ける	30	発表	7	出席	8	済む	7	調べる	3	有る	43

表 6　謙譲語の元の動詞（頻度上位 30 語）

順位	ていただく		御－いたす		御－いただく		御－する		御－申し上げる		申し上げる[4]	
1	教える	1643	願う	2668	答える	297	願う	11301	願う	455	感謝	101
2	する	1570	答える	587	理解	269	勧める	1914	答える	454	答弁	13
3	見る	445	伺う	299	分かる	219	聞く	1820	説明	307	説明	10
4	来る	352	尋ねる	168	利用	154	伺う	1543	礼	159	祈念	7
5	遣る	337	諮る	168	協力	108	話す	957	報告	101	質問	6
6	考える	254	説明	113	考える	102	紹介	806	答弁	99	要望	5
7	クリック	186	聞く	92	答弁	100	待つ	745	詫びる	76	失礼	5
8	行く	166	報告	88	越す	93	会う	733	祈る	69	報告	4
9	読む	164	勧める	75	説明	87	尋ねる	705	話す	42	提出	3
10	見せる	157	届ける	63	話す	81	答える	531	待つ	31	注文	3
11	理解	139	紹介	57	教える	75	邪魔	363	挨拶	26	拝察	3
12	分かる	128	掛ける	51	検討	66	送る	360	質問	25	提案	2
13	取る	128	案内	48	読む	63	届ける	342	提案	22	期待	2
14	言う	128	断る	41	指摘	62	伝える	341	案内	19	提供	2
15	置く	128	送る	39	述べる	58	掛ける	271	断る	14	尊敬	2
16	知る	127	話す	32	出でる	58	馳走	229	遠慮	13	要請	2
17	出す	125	伝える	25	審議	55	説明	214	察する	13	歓迎	2
18	作る	123	待つ	21	連絡	53	返す	194	紹介	12	発表	2
19	聞く	121	連絡	21	楽しむ	52	案内	177	尋ねる	11	推挙	2
20	収める	116	受ける	20	示す	49	渡す	177	祈念	9	反論	2
21	成る	114	質問	20	許す	49	任せる	165	伺う	8		
22	参加	98	教える	19	紹介	43	見せる	150	連絡	8		
23	送る	92	察する	19	付き合う	42	受ける	137	伝える	7		
24	掛ける	92	返す	17	待つ	41	断る	128	返事	7		
25	検討	91	持つ	16	支払う	40	出掛ける	126	要望	7		
26	持つ	89	会う	16	参加	40	迎える	121	相談	6		
27	付ける	89	用意	16	議論	37	教える	119	披露	6		
28	答える	88	見受ける	15	聞く	36	預かる	113	注意	6		
29	思う	88	預かる	14	持つ	36	相談	104	期待	6		
30	為る	86	邪魔	13	集まる	36	連絡	88	指摘	4		

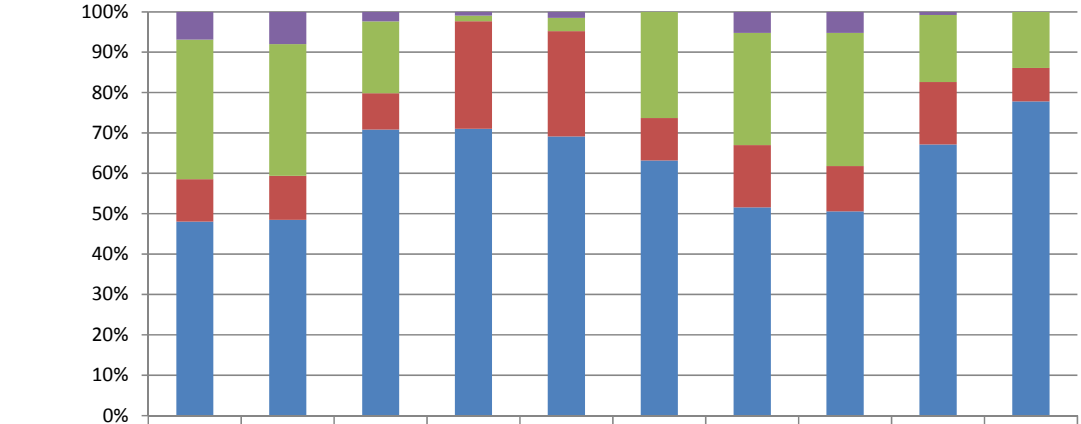
[4] 頻度 1 となる 21 位以下は省略した。

現時点では用例のクリーニングが完全でないため、表 5・表 6 には、「おすすめです」のような美化語と考えられる用例や、「お出でなさる」「ご覧なさる」などの語彙的な敬語形式、「おでかけする」「お礼申し上げる」などの固定的な表現の一部が含まれている。

５．「（御）－（て）くださる」と「（御）－（て）いただく」

５．１　「くださる」と「いただく」の用例数の非対称性

尊敬語「－てくださる」と謙譲語「－ていただく」は、いずれも恩恵のやりとりを伴う表現で、ちょうど視点を入れ替えた関係となるため、通常、相互に置き換えが可能となる。一般に「A が－してくださる」は「A に－していただく」と言い換えられる。しかし、実際の用例数には偏りがある。これを確認するため、「御－くださる」「てくださる」と「御－いただく」と「ていただく」の 4 形式に限って集計し直したものが図 5 である。



	LB	OB	OC	OM	OP	OT	OY	PB	PM	PN
御－くださる	297	66	76	58	23		193	204	4	
てくださる	1497	270	576	91	52	5	1025	1288	84	10
御－いただく	456	90	289	1705	410	2	567	437	78	6
ていただく	2079	401	2288	4549	1085	12	1902	1974	340	56

図 5　「－くださる」と「－いただく」の用例数

文化庁文化審議会（2007）『敬語の指針』はこうした用法について，「どちらの言い方も適切に敬語が用いられているもの」とした上で、「「御利用いただく」の使い方には，問題があると感じている人たちもいる」とする。しかしコーパスの用例数の上ではどのレジスターにおいても謙譲語の「いただく」系が多く、尊敬語「くださる」は少ない。中でも OM（国会会議録）と OP（広報誌）では「いただく」系が多く 95%を超えている。書籍（LB, OB, PB）は、3 つのレジスターとも同じような割合であり、「いただく」系は約 60%で、比較的「くださる」系が多い。

５．２　「御」の不使用

表 1 の調査対象には含めなかったが、漢語サ変動詞から作られる「御－くださる」「御－いただく」に近いものとしてとして、それぞれ「御」が付かない「〈漢語〉くださる」「〈漢語〉いただく」という形式が用いられる場合がある。菊地（1997）はこの用法について「誤りと言えるかどうかはともかく、あまり一般的な言い方ではない」（p.210）としている。

この形式について、BCCWJ における用例数を調査した。粗頻度と調整頻度を表 7 に挙げる。

表 7　「〈漢語〉くださる」「〈漢語〉いただく」の用例数と調整頻度（100 万語あたり）

		LB	OB	OC	OM	OP	OY	PB	PM	PN
〈漢語〉いただく	用例数	22	3	129	73	125	105	38	11	4
	調整頻度	0.72	0.80	12.58	14.31	33.29	10.30	1.33	2.47	2.92
〈漢語〉くださる	用例数	16	2	180	1	142	66	33	3	0
	調整頻度	0.53	0.53	17.55	0.20	37.81	6.47	1.16	0.67	0.00

この形式の用例数が多かった語は次の通りである（括弧内は用例数）。

「ーいただく」
回答(40), 連絡(28), 落札(26), 参加(24), 協力(17), 指摘(11), 確認(11), 答弁(11), 検討(10), 理解(10), 応募(9), 納付(9), 応援(9), 賛同(9), 利用(7), 評価(7), 説明(7)

「ーくださる」
参照(79), 連絡(68), 回答(28), 参加(27), 注意(25), 利用(24), 確認(21), 持参(14), 相談(10), 電話((9), 意見(8), 応募(7), 送付(5), 紹介(5)

６．まとめ

以上、BCCWJ を利用して現代語の敬語形式の用法について調査結果を見てきた。現代語の敬語形式について、大規模なコーパスに基づいて数量的な面から検討した研究はこれまでに行われておらず、敬語形式の使用頻度や使用域、元となる動詞の語彙の違いは、今回はじめて計量的に記述された。現代語の敬語形式の用法記述にとっても、こうした計量的な面からの研究が資するところがあると考える。現時点では粗い調査に過ぎず十分な分析も行えなかったが、今回得られたデータを基礎として、今後、詳細な研究を進めていきたい。

文　献

小山田由紀・柏野和佳子・前川喜久雄（2012）「助動詞レル・ラレルへの意味アノテーション作業経過報告」第 2 回コーパス日本語学ワークショップ予稿集, pp.59-68

小木曽智信・中村壮範（2011）「『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版」JC-U-10-01

菊地康人（1997）『敬語』講談社

文化庁文化審議会（2007）『敬語の指針』
http://www.bunka.go.jp/bunkashingikai/soukai/pdf/keigo_tousin.pdf

『法華百座聞書抄』のテンス・アスペクト形式：『今昔物語集』との比較

野田 高広（国立国語研究所非常勤研究員）†

Tense/Aspect Markers in *Hokke Hyakuza Kikigaki Shō* and *Konjaku Monogatari Shū*

Takahiro Noda (National Institute for Japanese Language and Linguistics)

１．はじめに

『法華百座聞書抄』（1110 年成立）と『今昔物語集』（1120 年以降成立）のテンス・アスペクト形式の比較考察を試みる。方法として、文体的にも近いと考えられる両資料について、アスペクト形式「たり」「り」「つ」「ぬ」にテンス形式「けり」「き」が後接する形式を抽出してその傾向の差に着目する。調査対象には「橘をひいたぞ、あやまちすな。」（平家物語・巻４）のような明確な「た」の例は現れないが、中古資料なども取り上げながら、衰退の途をたどる「けり」からテンス的意味を獲得しつつある「たり」への動態の一コマを、複合形式の傾向の違いとして記述する。いずれも国語研究所内で整備を進めているコーパスを利用して、『今昔物語集』は最古本である鈴鹿本を底本とする一部の巻（12; 27）を利用する。

２．『法華百座聞書抄』について

２.１ 言語資料としての『法華百座聞書抄』

『法華百座聞書抄』は、ある内親王の発願により 1110 年 2 月 28 日より 300 日間講じられた法華経・阿弥陀経・般若心経の説教の聞き書きであり、院政後期の抄写本（法隆寺蔵本）が唯一の伝本として現存する。法隆寺蔵本には 300 日間のうち、20 日分の説教が筆録されており、計 35 の説話を含む。原文は片仮名を主体とした漢字片仮名交じり文であり、典拠の影響もあって漢文訓読語の影響が大きいとされる（アルイハ・カルガユエニ等）。また、下二段動詞に「り」が接続する例（アタヘリ）や推量助動詞「う」（マウサウニシタカイテ）など中世語への過渡的な段階を示す例も少なからず見られ（cf. 小林 1975）、文体的な位置付けなど問題は多いながらも、日本語史資料として重要な文献であることに違いはない。

２.２ 法華百座聞書抄コーパス

調査では、『法華百座聞書抄総索引』の本文編をテキスト化したものを XML 形式でマークアップし、さらに形態素解析を施したものを使用した。また、『今昔物語集』は通時コーパスプロジェクトの一環として現在開発中のコーパス（テキストは小学館新編日本古典文

† maypole@gmail.com

学全集）を用いた。今回は比較的整備が進んでいる本朝仏法部の巻 12 と本朝世俗部の巻 27 を対象とした（いずれも鈴鹿本を底本とする）。表 1 に示すのはそれぞれの総語数（総短単位数）である。

表 1　調査資料総語数

『法華百座聞書抄』		20053 語
『今昔物語集』	巻 12	45038 語
	巻 27	31164 語

３．複合形式の一語化

通時的に助動詞「たり」は完了・存続を表すアスペクト形式だったのがテンス的な意味を獲得し単純過去を表すようになった（cf. 此島 1973: 183）。本節ではこの契機と考えられる助動詞の複合現象をあつかう。

３.１ 山口尭二（2003）

山口（2003）は、アスペクト形式「つ」「ぬ」「たり」やテンス形式「き」「けり」から構成される古代語のテンス・アスペクト体系が「た」を中心とする中世以降の体系に推移する要因を、「ぬ」と「つ」の混同、および「てけり」などの複合形式の一般化に求める[1]。

(1)　「ふたたびは慈悲をもつて汝にあたへぬ。さのみは身の力なし。かなふまじ」（撰集抄・3・7）
(2)　宇佐の宮に参り、千部の経を読まむと、大願を立、七百余部はかきよみぬ。（幸若・百合若大臣）
(3)　我も／＼とはせかさなり／＼、ほどなく一千余騎になりてげり。（金刀比羅本保元・上）
(4)　かの福田神の所為とさとりて、犬をおいのけつ。そのゝち気色なをりてけり。（古今著聞集・256）

(1)(2)は中世以前なら「つ」の選択が予想されるような例で、「与ふ」「書き読む」という人為的な動作を表す動詞に「ぬ」が選択されている。さらに、中世には(3)(4)のように、複合形の「てけり」には、本来なら「ぬ」が選択されそうな自然的な変化を表す動詞の前接例が目立つという。

(5)　良ノ申ウケタマウニヨツテ、高祖ヲモ翼蔽シテ、タスケテアツタシ。（漢書抄・項籍伝・三 23 オ）
(6)　只乗興而言タケルヤ（杜詩続翠抄・八 31 ウ）

さらに、(5)のような、中世以前の「たり」には見られない存在動詞への後接例などをあ

[1] フランス語の単純未来形 *je chanterai* や *ne* V *pas* の通時的変化に見られる fusion（融合；Hopper & Traugott 2003: 52-55, 65-66)も同様の現象だと考えられる（cf. 野田 2012: 92-97）。

げて、これらの「たし」や「たける」などの一体化した形式の出現を、「た」を中心とした近代語的体系への統合過程として捉えている。

３.２ アスペクト形式の「けり」後接傾向

山口（2003）に説かれる「た」への統合過程を確認するべく、『法華百座聞書抄』『今昔物語集』12,17 巻のアスペクト形式「たり」「り」「つ」「ぬ」にテンス形式の「けり」「き」が後接する比率を調査した。参考として先に中古作品の状況を確認する。

３.２.１ 中古作品の傾向

中納言日本語歴史コーパスを用いて、中古 10 作品の状況を調査した結果を示す。まず、表 2 は「たり」「り」「つ」「ぬ」のトークン頻度、および、それぞれの形式の使用率である。

表 2　中古 10 作品のアスペクト形式のトークン頻度と使用率（トークン頻度／総語数＊100）

	総語数	たり		り		つ		ぬ	
01竹取物語	10144	102	1.01%	59	0.58%	48	0.47%	108	1.06%
02古今和歌集	31017	76	0.25%	432	1.39%	75	0.24%	347	1.12%
03伊勢物語	13657	104	0.76%	110	0.81%	29	0.21%	180	1.32%
04土佐日記	6606	43	0.65%	107	1.62%	21	0.32%	69	1.04%
05大和物語	22878	258	1.13%	110	0.48%	74	0.32%	227	0.99%
06落窪物語	53952	653	1.21%	329	0.61%	340	0.63%	499	0.92%
07枕草子	64996	1544	2.38%	183	0.28%	292	0.45%	379	0.58%
08源氏物語	435893	4333	0.99%	3447	0.79%	1545	0.35%	3264	0.75%
09紫式部日記	17094	327	1.91%	86	0.50%	35	0.20%	76	0.44%
10和泉式部日記	10759	147	1.37%	21	0.20%	58	0.54%	133	1.24%
総計	666996	7587	1.14%	4884	0.73%	2517	0.38%	5282	0.79%

成立年代順に並べたが、取り立てて目立った傾向はない。以下の表 3、表 4 に示すのはそれぞれのアスペクト形式の総数に対する「けり」「き」の後接数および後接率である。

表 3　中古 10 作品のアスペクト形式の「けり」後接率

	たりけり		りけり		てけり		にけり	
01竹取物語	6	5.88%	1	1.69%	2	4.17%	15	13.89%
02古今和歌集	41	53.95%	61	14.12%	3	4.00%	93	26.80%
03伊勢物語	54	51.92%	42	38.18%	9	31.03%	88	48.89%
04土佐日記	0	0.00%	10	9.35%	0	0.00%	12	17.39%
05大和物語	149	57.75%	61	55.45%	17	22.97%	105	46.26%
06落窪物語	40	6.13%	16	4.86%	19	5.59%	51	10.22%
07枕草子	24	1.55%	8	4.37%	8	2.74%	49	12.93%
08源氏物語	116	2.68%	85	2.47%	80	5.18%	452	13.85%
09紫式部日記	6	1.83%	0	0.00%	1	2.86%	14	18.42%
10和泉式部日記	4	2.72%	0	0.00%	1	1.72%	18	13.53%
総計	440	5.80%	284	5.81%	140	5.56%	897	16.98%

表 4　中古 10 作品のアスペクト形式の「き」後接率

	たりき		りき		てき		にき	
01竹取物語	3	2.94%	0	0.00%	4	8.33%	3	2.78%
02古今和歌集	0	0.00%	0	0.00%	10	13.33%	30	8.65%
03伊勢物語	1	0.96%	2	1.82%	1	3.45%	7	3.89%
04土佐日記	3	6.98%	2	1.87%	0	0.00%	1	1.45%
05大和物語	5	1.94%	0	0.00%	7	9.46%	12	5.29%
06落窪物語	24	3.68%	7	2.13%	6	1.76%	27	5.41%
07枕草子	50	3.24%	11	6.01%	3	1.03%	29	7.65%
08源氏物語	168	3.88%	124	3.60%	51	3.30%	298	9.13%
09紫式部日記	12	3.67%	1	1.16%	0	0.00%	5	6.58%
10和泉式部日記	3	2.04%	0	0.00%	2	3.45%	8	6.02%
総計	269	3.55%	147	3.01%	84	3.34%	420	7.95%

この２つの表の中で特に目を引くのが『古今和歌集』『伊勢物語』『大和物語』の「けり」後接率の高さである。３作品の特徴としては韻文的要素が強いということが挙げられそうだが、以下の表 5 のように、「けり」後接例は和歌や会話文よりもむしろ地の文に多く出現する。とりわけこの３作品の突出した数値については、文法的に説明できる問題なのか、個人的な文体差の問題なのかよく分からず断定は控えたいが、表 6 に明らかなように『伊勢物語』『大和物語』の「けり」の使用率は他の作品に比して極めて高く、両者は『源氏物語』の 5 倍以上の比率を示す。このことが大きく関わっていることは確かであろう。

概して、中古和文作品では「たり」「り」「つ」への「けり」後接率が 10%を切る作品が大半であり、３作品を除き、それほど強い後接傾向は見られないことが確認できる。また、「ぬ」の「けり」後接例は「つ」に比して若干多いようだが、せいぜい 10%台にとどまるものが大半となっている。

表 5 「けり」後接：本文種別集計

作品名	歌			歌	会話		会話	詞書				詞書
	たり	つ	ぬ	集計	たり	ぬ	集計	たり	つ	ぬ	り	集計
02古今和歌集		1	49	50				36	2	31	58	127
03伊勢物語			14	14		1	1					
05大和物語	2		9	11	1	1	2					
総計	2	1	72	75	1	2	3	36	2	31	58	127

作品名	手紙		手紙	地				地	総計
	たり	ぬ	集計	たり	つ	ぬ	り	集計	
02古今和歌集				5		13	3	21	198
03伊勢物語		1	1	54	9	72	42	177	193
05大和物語	1		1	145	17	95	61	318	332
総計	1	1	2	204	26	180	106	516	723

表 6　中古 10 作品「けり」使用率（「けり」総数／総語数＊100）

作品名	総語数	「けり」総数	使用率
01竹取物語	10144	111	1.09%
02古今和歌集	31017	764	2.46%
03伊勢物語	13657	668	4.89%
04土佐日記	6606	81	1.23%
05大和物語	22878	1288	5.63%
06落窪物語	53952	511	0.95%
07枕草子	64996	346	0.53%
08源氏物語	435893	3635	0.83%
09紫式部日記	17094	80	0.47%
10和泉式部日記	10759	68	0.63%
合計	666996	7552	1.13%

３.２.２　『法華百座聞書抄』と『今昔物語集』の使用状況

表 7 に『法華百座聞書抄』の「たり」「り」「つ」「ぬ」のトークン頻度および使用率を示す。

表 7　アスペクト形式のトークン頻度と使用率（トークン頻度／総語数＊100）

	総語数	たり		り		つ		ぬ	
法華百座	20053	106	0.53%	135	0.67%	60	0.30%	82	0.41%
今昔12	45038	233	0.52%	201	0.45%	84	0.19%	202	0.45%
今昔27	31164	478	1.53%	38	0.12%	153	0.49%	291	0.93%

先の中古 10 作品と比較して顕著な傾向は見られないが、『今昔物語集』巻 27 での「り」の数値が低い点が指摘できる。これは和文体的性質が色濃いとされる本朝世俗部の巻であることも無関係ではないだろうが、中古和文作品では「給ふ」の前接例の数が際立っており、たとえば『源氏物語』では「り」3447 例中の 2621 例を「給ふ」前接形が占める。一方、『今昔物語集』の巻 12 では 201 例中の 31 例、巻 27 では 38 例中での 3 例と、「給へり」という複合形が減少傾向にある。このように、中古和文作品での「り」の頻度の高さは「給ふ」前接形の数値の影響を受けており、表 7 から『法華百座聞書抄』や『今昔物語集』において「り」が減少したとはいうことはできない。

以下、表 8・表 9 に示すのは「けり」「き」の後接率である。

表 8　アスペクト形式の「けり」後接率

	たりけり		りけり		てけり		にけり	
法華百座	11	10.38%	2	1.48%	2	3.33%	11	13.41%
今昔12	6	2.58%	3	1.49%	5	5.95%	18	8.91%
今昔27	126	26.36%	8	21.05%	28	18.30%	120	41.24%

表 9　アスペクト形式の「き」後接率

	たりき		りき		てき		にき	
法華百座	4	3.77%	6	4.44%	1	1.67%	0	0.00%
今昔12	7	3.00%	5	2.49%	0	0.00%	6	2.97%
今昔27	6	1.26%	1	2.63%	1	0.65%	8	2.75%

表 8 の『今昔物語集』巻 27 の数値が際立っていることが分かると思う。巻 27 はすべての形式において中古和文作品全体の数値を大きく上回っており、『法華百座聞書抄』や巻 12 とは傾向を異にしている。先述の通り巻 27 は和文的特性が強いとされる本朝世俗編に属する巻であり、「けり」後接形の比率の増加は、山口（2003）の「一体化」が進行するさまを表していると考えてよいだろう。一方で、数値の少なさはともかく、『法華百座聞書抄』の示す数値は『今昔物語集』巻 12 や一部を除く中古和文に近い傾向を示している。

表 9 の「き」後接例は用例数も少なく、あくまで参考にとどめざるを得ないが、「たりき」「りき」の例は先掲の(5)のような、室町期に現れる「たし」へと継承されていく形式だと考えられる。以下、参考に「たりし」と「たりけり」の例を示す。

(7)　「…をもひかけぬ國の亂いできたりしに、まどはしてその行方をしらず」（法華・オ 498）

(8)　「…母の胎より生れ出ける時に、左の手に捲て生れたりけるを、母の此のごとく申て得しめたりし也。…」（今昔・12-34）

(9)　身の才微妙く、心賢く御ければ、世の人、賢人の右の大臣とぞ名付たりし。（今昔・27-19）

(10) おもひもあへず、「阿彌陀佛」といはれたりけるをききて、この魚のわがかたにかへりてきければ…（法華・オ 271）

(11) 其の鯖荷たりける杖、今に御堂の東の方の庭に有り。（今昔・12-7）

(12) 其の時に寢殿の辰巳の母屋の柱に、木の節の穴開たりけり。（今昔・27-3）

３.３ 前接動詞の傾向

「つ」「ぬ」は中古作品では動作の意図性に対応した強い選択傾向があったのが、時代を下るに従い例外的なものが現れるようになる。３.１で述べたように、山口（2003: 230-231）は「てけり」などの複合形において両者の混同例が顕著である点を指摘しており、本節では山口（2003）の挙げる例（(1),(2)など）より若干時代をさかのぼる『法華百座聞書抄』および『今昔物語集』巻 12, 27 の実態を確認したい。

３.３.１ 「て-けり／き」

『法華百座聞書抄』の「てけり」は山口（2003）で示されたような混同例は存在しないようであり、以下のように「てけり」は意図的な動作を表す動詞句に後接する。

(13) …あかつきにのみかくいひゐたらむ、はづかしきことなりと思て、高きいはほのうへにいたりて身をなげてけり。地獄に率ていかされて、獄卒、此の僧をかなへにうちいれつ。（法華・オ 183）

(14) 弟子、五躰を地になげてをがみたてまつりて、それよりなむ深く信じてさとりをひらきてける。はふせん大師は、わが山王院の大師にまします。（法華・オ 357）
(15) 此の佛の功徳、大概さきざき申てしかば、何ごとをか申候べからむ。（法華・ウ 112）

一方の『今昔物語集』では(18)のような、「命を失ふ」という非意図的な事態の「てけり」後接例が存在するが、このような例は他には見られないようであり、大勢としては中古作品と近い傾向を示す。

(16) 「前に別れ去にし僧の、此の山に入行ひける間、生死を猒て身を投てける也」（今昔・12-31）
(17) 遠助、「哀れ、『見まじ』と云てし物を。不便なる態かな」と云て…（今昔・27-21）
(18) 嫉妬の故に遠助思懸ず、非分に命をなむ失ひてけり。（今昔・27-21）

３.３.２ 「に-けり／き」

『法華百座聞書抄』に「にき」は見られなかった。「にけり」も中古和文と似通った傾向を示しており、非意図的な動詞への後接例が大半を占める。(22)の「参る」は『源氏物語』などの平安中期の作品では「つ」「ぬ」のいずれも後接する動詞であり、やはり例外的なものは見られない。

(19) 此は法花寺のあかつきのかねの聲と思て、我懈怠しにけりとおどろきて…（法華・オ 186）
(20) わがころもをとらへさせて、ほどもなくゆきつきにけり。（法華・オ 238）
(21) …ひじりをおがみてなむさり給にける。（法華・オ 261）
(22) 魚とらむために三年ばかりがほどよるひる念佛をしけるは、みな極樂へまいりにけり。（法華・オ 276）
(23) しまのうちにひとりものこらず往生しにければ往生のしまとなむいひ…（法華・オ 277）
(24) 「かずもしらずも、ひさしくもなりなりにけるかな」とかなしくをぼえて…（法華・オ 290）
(25) 嚴をみるに、とひたてまつらむとをもひつることもわすれにけり。（法華・オ 394）
(26) 師子、酒をほしきままにのみて、ゑひてねにけるをみて…（法華・ウ 175）
(27) 給へりしに、佛になり給ひて、すでに十劫はすぎ給ひにければ…（法華・ウ 379）

『今昔物語集』の「にけり／き」にも取り立てて例外的なものは見られず、「知る」「失す」などの非意図的な動詞句に後接するもの、もしくは移動動詞などのような「つ」「ぬ」両例にまたがる動詞句の例しか見られないようである。

(28) 「隠れ得たりと思ふ穴をも知にけり」と書生思ふ程に…（今昔・12-28）
(29) 其の中大夫、従者一人を具して、我れは胡録掻負て出にけり。（今昔・27-37）
(30) 「…其の御前にして維摩経を読誦せしかば、即ち癒にき。」（今昔・12-3）
(31) …病付て有しを、繚ふ人も無くて、此の夏失にしを…（今昔・27-24）

４．結論

以上、『法華百座聞書抄』のテンス・アスペクト形式の複合形式「てけり」「たりけり」などに焦点を当てて考察をしてきた。３.２.２の表 8 に示したように『今昔物語集』の巻 27 では「たりけり」「てけり」などのような「けり」の強い後接傾向が確認できたが、『法華百座聞書抄』は『今昔物語集』巻 12 と似通った傾向を示しており、さらに、中古和文作品で見られた傾向と大きな違いは確認できなかった。また、山口（2003）で指摘された、「つ」「ぬ」の前接動詞の混同は調査対象にはほとんど見られないことを確認した。

中世資料的な性質を備える『法華百座聞書抄』はテンス・アスペクト形式の複合の様相をみる限りは平安中期の性質を温存すると考えられ、それは和漢両文脈の性質を兼ね備えるとされる『今昔物語集』巻 12 の状態に数値的には近い。一方、和文脈的な傾向が強いとされる『今昔物語集』の巻 27 は明らかにアスペクト形式への「けり」の後接率が高いのであり、「たりけり」のような複合形の増加を、「た」などを中心とした新しい体系への萌芽とみるのは山口（2003）に説かれたとおりである。

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、研究代表者：田中牧郎）による成果の一部です。

参考文献

此島正年（1973）『国語助動詞の研究』桜楓社.

小林芳規(編)（1975）『法華百座聞書抄総索引』武蔵野書院.

小林芳規（1975）「国語史研究資料としての『法華百座聞書抄』」小林芳規(編)『法華百座聞書抄総索引』, 557-558. 武蔵野書院.

坂詰力治（1975）「法華百座聞書抄における助動詞について」小林芳規(編)『法華百座聞書抄総索引』, 611-628. 武蔵野書院.

築島裕（1963）『平安時代の漢文訓読語につきての研究』東京大学出版会.

野田高広（2012）「アスペクト形式「ている」の成立について」『東京大学言語学論集』32, 85-107.

山口堯二（2003）「完了辞・過去辞の通時的統合：「たり＞た」への収斂」『助動詞史を探る』, 225-242. 和泉書院. （初出：川端善明・仁田義雄(編)（1997）『日本語文法：体系と方法』, pp. 211-227. ひつじ書房.）

山下和弘（2013）「『今昔物語集』に見られるテアリとタリの対比：「テアリ対タリ」「テアリケリ対タリケリ」」『国語国文学研究』48, 37-51.

Hopper, Paul J., & Traugott, E. C. (2003). *Grammaticalization* (2nd ed.). Cambridge: Cambridge University Press.

『太陽コーパス』の漢字表記語に対する熟字ルビについて

間淵 洋子（国立国語研究所 コーパス開発センター）†

On the Ruby Elements in Taiyo Corpus: The Case of Group-ruby

MABUCHI, Yoko (Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所コーパス開発センターでは，現在「通時コーパス」プロジェクトの一環として，形態論情報付きの近代語コーパスを構築している。これまでに，2012 年『明六雑誌コーパス』が公開されているほか，2014 年度には『国民之友コーパス』の公開も予定されており，今後も資料を拡充していく計画である。その一つが雑誌『太陽』であり，2005 年に公開された『太陽コーパス』を増補改訂し，新たに形態論情報付きコーパスとして構築し直す準備を進めている。

形態論情報を付与するにあたっては，言語の表層情報，すなわち表記を元に，語を割り当ててゆくという方法を取ることになるが，その際問題になるのが，異なる複数の語を同一の表記で表す「同表記異音語」の存在である（伝他 2008）。本発表では同一の表記で表される異なる語を「同表記異語」と呼ぶことにする。

例えば，「上方」という表記は，京阪地域を表す「かみがた」という語（A）と上の方を表す「じょうほう」という語（B）の二つの異なる語のいずれをも表し得る形である。

(1)　【上方】落語には「ぜんざい公社」という演目がある。（BCCWJ:PN2b_00005；毎日新聞 2002/9/27 朝刊）
(2)　またこれまでの推移から、主力商品、売れ筋商品の目標数を【上方】修正する。（BCCWJ:PB46_00066；宮崎文明 2004『商売の「計画・実施・検証」；実例による販売計画と商品管理の原則』）

「上方」の場合は，上例のように接続する語あるいは文脈によって，いずれの語かを判断することは容易で，例(1)は語(A)に，例(2)は語(B)にそれぞれ対応することが分かる。しかし，このような同表記異語の中には，それぞれの語同士が同義または極めて近接した意味を持つために，文脈を手掛かりに語を対応付けるのが困難な場合もある。特に，いわゆる「熟字訓」と呼ばれる表記と語の対応は，そもそも和語とそれに極めて近接的な意味を持つ漢語を結びつけることによって成り立っているために，文脈に基づく語の判定が容易でないことが多く，語種や文体といった手掛かりですら無力な場合も少なくない。

そのような場合に決定的な手掛かりとして用いることができるのが，表記に付された振り仮名（ルビ）である。ルビによって，読みを確定することができれば，その表記がどの語を表すための表層形かを知ることができる。

(3)　光代、【明日（あす）】は夙く發たうぞ。（太陽 1895 年 2 号 川上眉山「書記官」）
(4)　今では局長が引受て、萬事表面上商會の世話をして居る仲であつて見れば、すでに

† mabuchi@ninjal.ac.jp

【明日（みょうにち）】か二三日中に願書が出来て、（太陽 1895 年 2 号 川上眉山「書記官」）

『太陽コーパス』は，本文行の文字列に加えてそれに対応付けた形で詳細なルビ情報を付与しており，これを観察することによって，表記と語の結び付きの実態を把握できる可能性がある。

本発表では，同表記異語における表記と語の結び付きの実態把握を目指し，その第一段階として，『太陽コーパス』におけるルビ情報のうち，熟字訓を含む，複数文字列に対応するルビ情報（以下「熟字ルビ」と呼ぶ）を調査し，その一部について出現の実態を報告する。その上で，先に挙げたコーパスへの形態論情報付与における問題点に対する対処法を示す。

２．調査概要

２．１　コーパス

調査には，2005 年に公開された『太陽コーパス』CD-ROM を用いた。

『太陽コーパス』は，言文一致を経て口語体による書き言葉が安定し普及する時期（明治時代後期～大正時代）の書き言葉を代表できるコーパスとして作られたものであり，月刊総合雑誌『太陽』（博文館）の明治 28（1895）年，明治 34（1901）年，明治 42（1909）年，大正 6（1917）年，大正 14（1925）年について，著作権処理ができなかった記事を除くほぼ全文を対象にしたものである。

雑誌『太陽』は，分量の多さ，ジャンルの広さ，執筆陣の多彩さ，読者層の厚さなどの点で，当時の文献資料として格別の価値を持っていることが指摘されており（田中 2005），それを裏付けるように，2005 年の『太陽コーパス』公開以来，口語化，濁点や仮名遣い等の表記法の確立・統一化，新漢語の定着など，当時の語用の実態や現代語に連なる変化の過程を明らかにした論考が多く発表された。

『太陽コーパス』の収録記事が発行された明治～大正期は，ルビが活発に使用されていた（田中 2005）ことに加え，現在のように表記の規範意識が高くなく，語と表記の関係が多様であったことが知られている(今野 2012)。本研究で取り上げる漢字表記語に対して与えられた様々な「熟字ルビ」も，そのような明治～大正期の表記の多様性を映す現象の一つであり，『太陽コーパス』はこの現象を観察するのに有用な調査対象資料である。

２．２　調査対象表現

本研究で対象とする「熟字ルビ」の例を抽出するために，XML タグを利用した。

『太陽コーパス』は，XML 形式により様々な情報が付加されているが，その中に，本文行の文字に与えられた振り仮名（ルビ）の文字列を格納する要素「r」がある。

雑誌『太陽』における振り仮名の様相については，以下の報告がある（田中 2005）。

- 1910 年頃までは，記事の属する欄によって (1)総ルビ（ほとんどすべての漢字に振り仮名があるもの），(2)パラルビ（一部の漢字に振り仮名があるもの），(3)無ルビ（特別な場合を除き振り仮名がないもの）かが決まっている。
- 1910 年代からは，ほとんどすべての記事で総ルビになっている。
- 振り仮名を付けるか否かは著者の判断だけでなく編集部の判断も大きかったのではないかと推測されるが，必ずしも読みにくさや誤読の可能性などによるとは言いがたい。

・　全巻を通して常に総ルビであるのは，小説・戯曲ジャンルの記事である。

これを踏まえて，『太陽コーパス』においては，誌面に見られるルビについて，

・　小説・戯曲・詩歌の類の記事[1]については，原文にある振り仮名を入力する。

・　上記以外の記事については，原文に振り仮名があっても入力しない。

という方針に基づき，「r」タグに格納している。この結果，『太陽コーパス』全 3409 記事のうち，1 割程度に該当する 364 記事に対してルビが付与されている。

なお，「r」タグは，原則として漢字 1 字 1 字にルビを対応させる形（モノルビ）で付与されるが，熟字訓など複数文字連続にしか対応させられないルビについては，語の範囲で「r」タグを付与している（グループルビ）。今回は，この複数文字に対応付けられたルビを「熟字ルビ」と呼び，これらを調査対象表現とした。このようなルールで付された熟字ルビは，単なる字と音の結び付きではなく，語と語の結びつきを示す表現であるはずであり，同一表記を介した異語の関係性や，表記と語の結びつきの強さなどを観察するのに好対象と言える。

２．３　調査方法

XML 文書から XPath を用いて調査対象表現を抽出した。XPath は，XML 文書内で特定のノードの位置を指定することを目的とした構文である。コンピュータにおけるファイル・システムパスと同様に，スラッシュ区切りで階層を明示した式（ロケーション・パス）を使ってノードを参照することができ，比較的容易に XML 文書内の特定の要素や属性を抽出することができる。今回は，スクリプト言語 perl を用い，XPath 式「//r」を元に 2 文字以上の「r」要素（本文行文字列）と「rt」属性値（ルビ文字列）の対を抽出した[2]。

これによって抽出されるのは，『太陽コーパス』において，2 文字以上の本文文字列を範囲として r タグが付されているものすべてであり，延べ 26,165 件の表記とルビの対（ルビは語形を示していると考えられるので，以下「表記-語対」と呼ぶ）が抽出された。これらは，記事の異なりで 12,800 程度に，更に表記揺れ等を統一し，異なり 4,900 程度の表記-語対に集約した。

この表記-語対には，「茶-間／ちゃの-ま」のように，1 文字ずつ分割して r タグを付すことが可能なものや，「中合／なからい」のように音変化を伴うために分割はしがたいけれども，単漢字とそれに対応する読みが照合できるようなものも含まれる。今回の分析においては，語と語の結びつきを観察するという調査目的から外れるため対象から外した。同様に，表記事情が特殊で扱いに注意が必要となる固有名や動植物名なども，今回の試験的分析には不適切と判断し，対象から外した。その結果，分析対象として抽出されたのは，約 4,700 個の表記-語対である。

３．調査結果と分析

３．１　語種とルビのタイプ

調査対象として得られた表記-語対について，本文行文字列の語種，ルビ文字列の語種，ルビの表現タイプを分類した。分類結果を表 1 に示す。なお，語種の判断については，山田俊雄他編『新潮国語辞典　第二版』（新潮社）を参照した。

---

[1] 「記事」タグ内「ジャンル」属性に格納された NDC 番号のみを規準として対象記事を定めている。

[2] 作業環境：Windows8，Cygwin(1.7.17)，perl(5.14.2)＋XPath モジュール

語種：漢語（漢文由来表現を含む），和語，和語・漢語（対応する音読み・訓読みが両方あるもの），外来語（外国語含む），混種語，宛字（対応する漢語がないもの）
ルビの表現タイプ：表音的ルビ，表意的ルビ

表１　本文行文字列およびルビ文字列の語種別表記対数

		ルビ語種				総計
		漢語	和語	混種語	外国語	
表意ルビ		170	4200	7	259	4636
本文語種	漢語	160	3800	6	205	4171
	和語	3	290	1	26	320
	和語・漢語	2	42		3	47
	混種語		10		1	11
	宛字	5	58		24	87
表音ルビ			39		14	53
本文語種	和語		1			1
	宛字		38		14	52
総計		170	4239	7	273	4689

熟字ルビの大半は表意的なルビである。中でも，漢語の本文文字列と和語のルビ文字列を対応付ける表記対が圧倒的に多く，いわゆる熟字訓もここに含まれる。ついで，漢語に対して外国語，漢語に対して別の漢語，和語に対して別の和語を対応付ける表記対が多く見られる。それぞれ，以下に例を示す。『太陽コーパス』の熟字ルビには，語種の異なる語の結びつきのみならず，同一語種の別語を結びつける例が多いことが注目される。

漢語×和語：今日／きょう，先刻／さっき，住居／すまい
漢語×外語：硝子／ガラス，煙管／キセル，卓子／テーブル
漢語×漢語：真実／ほんとう，平常／ふだん，容貌／きりょう
和語×和語：左手／ゆんで，右手／めて，最早／もう

３．２　表記対の出現記事数

次に，出現記事数を計測した。度数分布を表２に示す。

表２　出現記事数と表記の分布

出現記事数	表記-語対数	累積%
1	3397	72%
2-3	742	87%
4-5	211	92%
6-7	117	94%
8-9	54	96%
10-141	192	100%

7 割程度が 1 記事のみに出現する表記-語対である。以下に例を示す。

出現記事数１の例：高利貸／アイスクリーム，争論／いさかい，同行／いっしょ

これらは，一時的な表記と語（読み）の結びつきである可能性もあるが，ここで示したのは単純な出現数であるため，当該の表記・語自体の出現可能性の差による部分が大きく，一概に定着していない・通用的でないとは言いがたい。中には辞書に掲載される表記-語対も含まれる。

一方，複数の記事に現れる表記-語対の場合は，この時代においてはある程度定着した通用的な表記-語対と言える。表３に出現記事数の多い順に表記-語対を挙げる（参考として，ルビ付与対象記事 364 記事に対する出現率を合わせて示す）。

「今日／きょう」のようないわゆる「熟字訓」のほか，「先刻／さっき」「四辺／あたり」といった現代では一般的と思われない表記-語対が上位に見られる点は注目される。

表３　出現記事数　上位 30 語

表記-語対	本文行語種	ルビ語種	ルビタイプ	出現記事数	出現率
一人/ひとり	漢語	和語	表意	141	39%
如何/いかん	漢語	和語	表意	134	37%
今日/きょう	漢語	和語	表意	120	33%
二人/ふたり	漢語	和語	表意	120	33%
何處/どこ	和語	和語	表意	115	32%
此方/こちら	和語	和語	表意	106	29%
何時/いつ	和語	和語	表意	92	25%
此處/ここ	和語	和語	表意	89	24%
其處/そこ	和語	和語	表意	85	23%
一寸/ちょっと	漢語	和語	表意	83	23%
可愛/かわいい	宛字	和語	表音	73	20%
何故/なぜ	和語	和語	表意	68	19%
昨日/きのう	漢語	和語	表意	64	18%
田舎/いなか	漢語	和語	表意	62	17%
明日/あす	漢語	和語	表意	57	16%
今朝/けさ	漢語	和語	表意	57	16%
這入/はいる	和語	和語	表意	54	15%
不可/いけない	漢語	和語	表意	51	14%
可笑/おかしい	宛字	和語	表意	51	14%
心地/ここち	漢語	和語	表意	50	14%
屹底/きっと	宛字	和語	表音	49	13%
先刻/さっき	漢語	和語	表意	47	13%
身體/からだ	漢語	和語	表意	46	13%
煙草/タバコ	漢語	外国語	表意	46	13%
四邊/あたり	漢語	和語	表意	44	12%
今年/ことし	漢語	和語	表意	44	12%
眞面目/まじめ	漢語	和語	表意	44	12%
眞似/まね	宛字	和語	表意	43	12%
甲斐/かい	宛字	和語	表意	42	12%
住居/すまい	和語・漢語	和語	表意	41	11%

３．３　文体別に見た表記に対する読みの出現割合

次に，　出現記事数の多い表記-語対について，ルビ付与対象記事における本文文字列と対応付けられる語（読み）の出現実態を追加調査した。調査は，本文文字列の表記をターゲットとして，XML 文書内を文字列検索し，そこに付加されたルビ情報を元に読みを確定させるという方法で行った。分析に際しては，『太陽コーパス』の記事に付与された文体情報（口語・文語）を用いた。以下に調査結果の一部を示す。

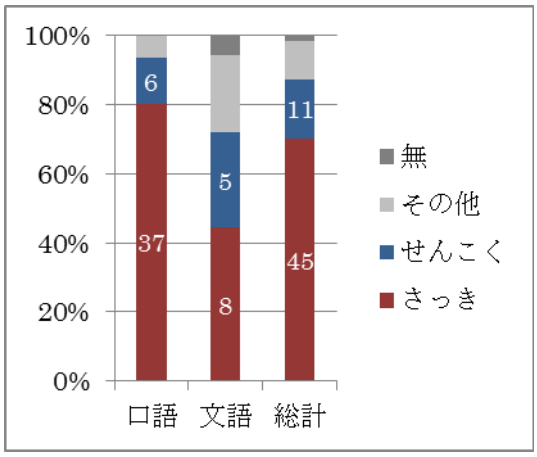


図１　本文表記「先刻」の読み

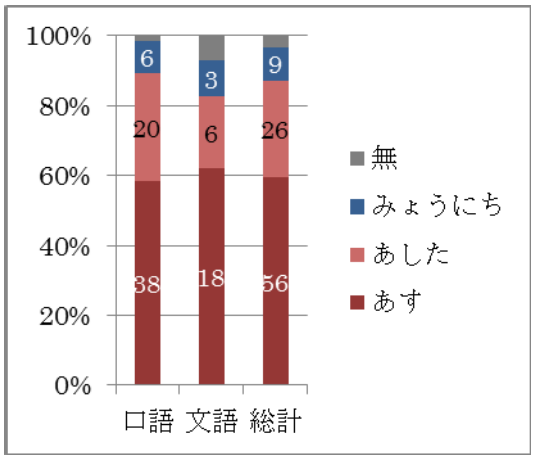


図２　本文表記「明日」の読み

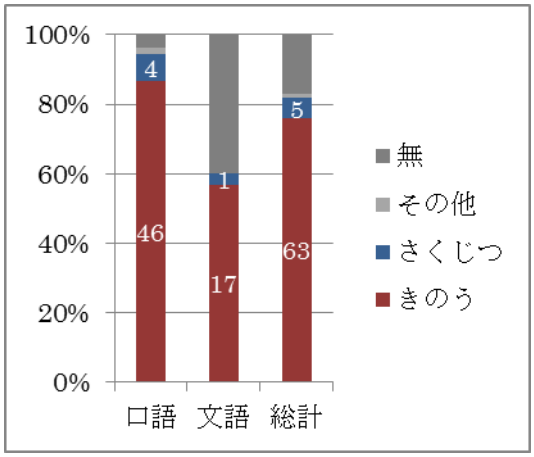


図３　本文表記「昨日」の読み

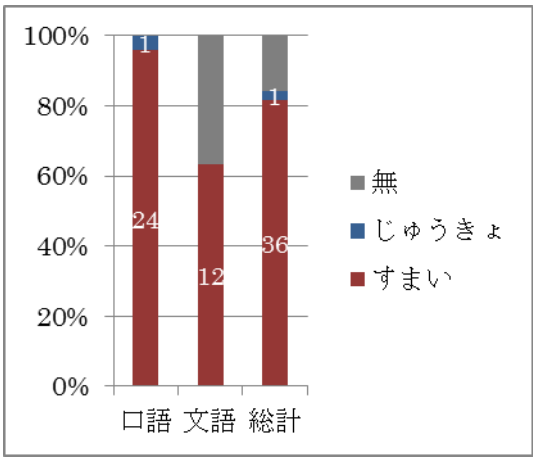


図４　本文表記「住居」の読み

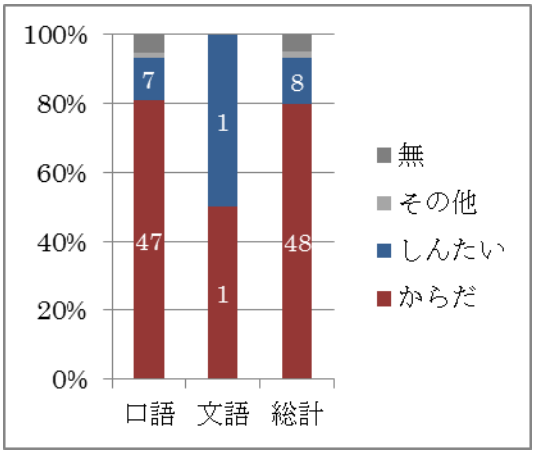


図５　本文表記「身体」の読み

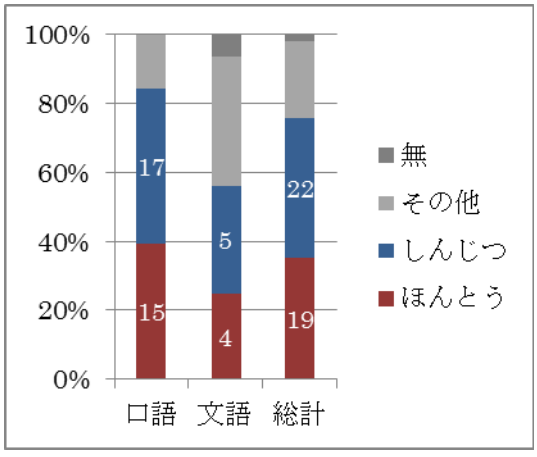


図６　本文表記「真実」の読み

図１の「先刻」を見ると，口語記事においては，圧倒的に和語「さっき」の語の表層形として用いられることの多いことが見て取れる。文語記事においても，「さっき」の割合が多いが，口語記事に比して字面通りに読む漢語「せんこく」の割合が多くなっている。

一方，図２の「明日」では，口語記事と文語記事でほぼ分布差異はなく，割合は，和語「あす」＞和語「あした」＞漢語「みょうにち」となっている。

図３，４の「昨日」「住居」は，いずれも口語記事で，ほぼ和語「きのう」「すまい」の表層形として用いられているが，文語記事では和語形に加えてルビのない形が多く見られる点に特徴がある。ルビのない形が漢語形なのか和語形なのかは不明である。

図５「身体」は，文語記事の数が少なく分析に用いることはできないが，口語記事において「からだ」と強力に結びついていることは疑いがない。

上５例については，文体差の見られる語もあるものの，元来の漢語形の読みを和語形の読みが大きく凌いでいるが，この中で現代語として一般的に「熟字訓」として認識されているのは，「明日／あす・あした」「昨日／きのう」であり，「先刻／さっき」「住居／すまい」「身体／からだ」といった表記と語（読み）の結びつきは強くない[3]。

図６「真実」は，同語種の別語との結びつきの例だが，口語記事で「ほんとう」「しんじつ」が同数ずつ，文語記事では「ほんとう」が「しんじつ」の倍以上を占め，「真実」と「ほんとう」の結びつきの強さが見て取れる。これも，現代においては希薄な結びつきの例である。

４．表記形にどの語を割り当てるべきか―形態論情報付与における指針

既存の形態論情報付き近代語コーパスである『明六雑誌コーパス』では，その形態論情報付与に関して，読みを与えるルールを以下のように定める（須永・近藤 2012）。

原則１　音読み・特に漢音を優先する。

原則４　原文にルビがあっても、その読みに従わない場合がある。

[1] 漢語に対し、外来語のルビがついている場合

[2] 原文のルビの読みが、熟字訓として定着しているとは言い難いもの

しかし，３節に示した調査結果より，少なくとも『太陽』の文芸記事においては，漢語に対して記事の文体や現代語の実態・内省による表記と語の結び付きを手掛かりに語を割当てることが不適切であることが分かった。

そこで，今後，『太陽コーパス』における形態論情報付与作業においては，今回の調査を元に，通用的と判断できる表記-語対については，以下のような方針で語の認定を行うことを検討したい。

(a) ルビに基づき語（読み）を認定する

(b) 表記に対する読みの分布に偏りがあるものについては，その情報に基づいて，ルビのない表記に熟字ルビに相当する読みを与える

５．まとめ

本発表では，『太陽コーパス』の XML タグを用いて，複数文字に対応するルビの抽出を試み，以下の調査報告を行なった。

[3] 例えば，BCCWJ コアでは「先刻-さっき」「住居-すまい」の表記-語対はなし，「身体-からだ」については書字形「身体」440 例中「からだ（体）」の語彙素認定は 71 例（16%）に過ぎない。

・ 本文表記およびルビの語種分類とそのクロス分布
・ 出現記事数による本文表記-ルビ対の定着度
・ 本文表記を元にした同表記異語の出現分布実態

また，調査結果を用いての，「時代的な表記と語の結びつき」を考慮した形態論情報付与を提案した。今回の手法あるいは時間的な制約によって残された問題点のいくつかを以下に示す。今後の課題としたい。

・ 文芸以外の記事についての実態把握

『太陽コーパス』の仕様上，文芸作品以外の記事にはルビが付与されておらず，それらの記事については実態の把握ができなかった。ルビの様態は多分にジャンルに依存するものであり，特に文学作品ではそれが顕著であることが予想される。更に，新聞等公共性の高いテキストにおけるルビの実態等とも合わせて考えるべきものである。

・ 著者別，年別の詳細な調査

テキストの表記は，著者の個人的な志向の強く現れる部分であるため，本来は，今回行った記事単位ではなく著者単位で集計を行うべきものである。また，今回指摘した表記と語（読み）の結びつきの定着度が現代語と『太陽』とで異なる点については，『太陽』の内部でも経年的な変化が見られる可能性もある。現代語との比較も合わせて今後取り組むべき課題と位置づける。

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：近藤泰弘）による成果の一部です。

コーパス

『太陽コーパス』: 国立国語研究所(2005a)『太陽コーパス―雑誌『太陽』日本語データベース―』（CD-ROM）博文館新社

『BCCWJ』: コーパス検索アプリケーション中納言（https://chunagon.ninjal.ac.jp/）

参考文献

今野真二(2012a)『百年前の日本語―書きことばが揺れた時代 (岩波新書)』岩波書店

今野真二(2012b)『ボール表紙本と明治の日本語』港の人

京極興一(1998)『近代日本後の研究―表記と表現―』東宛社

国立国語研究所(2005b)『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

須永哲矢・近藤明日子(2012)「近代語コーパスのための形態論情報付与規程の整備」田中牧郎・岡島昭浩・小木曽智信・小野正弘・小島聡子・島田泰子・朱京偉・高田智和・張元哉・陳力衛・近藤明日子・須永哲矢『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）』国立国語研究所、pp.93-117

武部良明(1981)『日本語表記法の課題』三省堂

田島優(1998)『近代漢字表記語の研究』和泉書院

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(2005b)、pp.1-48

伝康晴・中村純平・小木曽智信・小椋秀樹(2008)「語種情報を用いた同表記異音語の解消」言語処理学会年次大会発表論文集 14、pp.69-72

近代雑誌資料における程度副詞類の使用状況

市村太郎（国立国語研究所コーパス開発センター）†

The Usage of Degree Adverbs in the Magazine Material of Modern Japanese

Taro Ichimura (National Institute for Japanese Language and Linguistics)

１．はじめに

本研究では、国立国語研究所によって構築された『明六雑誌コーパス』(2012)・『太陽コーパス』(2005)を対象とし、主に程度副詞類の数量的な状況について分析する。

従来近代語における程度副詞に関する研究は、語史研究においても、各時代の計量的な調査においても、基本的には口語における使用状況を対象として進められてきた。また、中尾(2003)の『太陽』を対象とした研究などのように、文語・口語については区別されずに扱われたものもある。

一方で、当然文語文においても程度副詞は使用されるのであるが、こちらの研究は口語文での使用状況に比べてほとんどなされておらず、口語文での使用とどのように関係するのかはもちろん、何が文語的な副詞であるのかですらも、明確にはされてこなかった。近年副詞の語史研究が盛んに進められており、個々の研究においては、「口語的」という言及がなされることがしばしばあるが、基本的には「会話文で多く用いられる」などの尺度から、総合的に判断されているものとみられる。

そこで本研究では、口語文とともに文語文を多く含む近代雑誌類を資料として、文語記事・口語記事それぞれにおける程度副詞類の使用状況を、数量的な面から明らかにしたい。また、当該資料の範囲での注目すべき語の用法等についても検討する。

２．本研究における調査対象と方法

調査対象となる『明六雑誌コーパス』・『太陽コーパス』は、記事ごとにＸＭＬタグによって口語・文語の区別がなされており、本研究にではこの情報に基づいて調査を行った。

近代雑誌コーパスを利用した特定の語彙の文語・口語別の計量分析には、注目すべきものとして近藤(2013)がある。本研究においても、数値の指標等おおむねこれに倣い、それぞれの語の単純な出現数粗頻度、出現記事数を集計し、分析対象とした。なお本来ならば、言語の使用は著者・話者ごとというのが基本的な単位であるが、『太陽コーパス』においては著者不明の記事も多く、そのような記事は排除せざるを得なくなり、データ収集のうえで偏りを生じる原因となりうる。その点「記事数」ならば、同一著者の別記事を別々にカウントすることになるデメリットはあるが、データの漏れはなくなる。そのため、本稿においても著者ベースではなく、記事数ベースでの分析に重きを置いた。

次に対象とする語については、本稿では、程度副詞類の中でも最も種類の多い、後続する状態の「程度の甚だしさ」をあらわすものに限ることとし、松井(1977)などの過去の調査を参考にし、さらに実際『太陽コーパス』等での出現状況を調査し、全期を通じて粗頻度 100 を超える分量が得られるもの 14 語を対象とした。なお一定量得られるものとしては、「極僅か」の「ごく」や、古くからの「いと」も見られたが、文字列検索で得られたものを選り分けるのは難しく、また、解析済みデータにおいてもエラーが多いことが見込まれるため、本研究においては対象外とした。

また、用法が多種多様にわたるもの、当該期間において意味・機能の大きく変化しているものは除くこととした。

たとえば、真実性を基に事柄を強調する「まことに」は、程度副詞のように後続の状態

†i.taro@suou.waseda.jp

を強調するだけではなく、「実際に」を意味することもあれば、より構文的に広い範囲を修飾し、肯定的、また応答詞的に用いることもある。たとえば次のようなものは「程度副詞」と同等には扱い難いものであろう。このようなものをいわゆる程度副詞と選り分けることは困難であり、本研究においては除外した。

(1)東京九段阪上靖國神社の前に高く雲表を摩して樹てられたる青銅華表は諸大名より献納したる舊式大砲を鑄て造りたる記念物誠に千古に傳ふべきなり（『太陽』1「美術」1895 文語）

また、副詞「とても」は、松井(1977)等では集計対象とされているが、中尾(2005)によれば『太陽』の時期は程度副詞用法を派生した段階であり、また用例も多くは得られない。

その他、否定と呼応することのある「全然」や「あまり」等も含め、「後続の状態の程度の甚だしさを強調する」という単純な程度副詞としての機能以外でも多く使用される語については、用例の分別が時に困難であり、調査として平等を期すことができないため、今回は除外した。

このように、「程度の甚だしさ」を表す語に限り、さらにそこから一部の語を除外することにより、以下の語を調査対象とすることとした。

(2)「いたく」「いちじるしく」「おおいに」「きわめて」「しごく」「ずいぶん」「すこぶる」「たいそう」「だいぶ」「たいへん」「ひじょうに」「はなはだ」「ひどく」「よほど」

なお用例の収集に際しては、「ひまわり ver.1.3.1」を用い、形態論情報が付与されている『明六雑誌』については、形態論情報を利用して語を同定し収集した。また、『太陽コーパス』については、原則文字列検索を用い、それぞれの語について考えうる表記パターンにより網羅的に収集し、目視によって対象外の用例を選り分けた。その際、原則として、「連用修飾しており、後続する状態の程度の甚だしさを表しているもの」のみを集計対象とし、たとえば形容動詞用法の連体形や、名詞述語となっているような場合などは排除した。形容動詞連用形に相当する「○○に」や「いちじるしく」は取ったが、「甚だ」の場合は、古くから副詞として「はなはだ」が定着しており、形容詞とは大きく乖離していると考えられるため、「甚だし」に類する例は除外した。

(3)森中には高調凄音群猿の叫ぶを聞く、 俯して水源未知の利根を見れば、 水流混々河幅猶廣く水量甚多し、 或は岩に觸れて澎湃白沫を飛ばし、 或は瀾となり沈靜深緑を現はす、 沼田を發して今日に至り河幅水量共に甚しく減縮せるを覺えず、（『太陽』1「利根水源探検紀行」1895 文語）

また「原則文字列検索を用いた」と述べたが、「甚」が漢字１字・振り仮名なしで「はなはだ」を表す例が多くあるが、これを「甚」のみで抽出し、目視で選り分けるのは困難であり、またエラーが発生する可能性が高まるため、漢字１字・振り仮名なしの「甚」については、「近代文語 Unidic」によって形態論情報の付与された内部作業用データ（2014 年１月 31 日時点）を、誤解析箇所を除いたうえで利用した。

なお引用に関しては、「会話」「心話」については集計対象としたが、その他の文献等からの引用箇所については、付与された情報に基づき、除外することとした。

３.『明六雑誌』『太陽』における程度副詞類の使用数

２で上げた方針に基づき、表１に各語について、文語・口語および、口語記事中の会話文における使用回数と、出現記事数を、各資料・年代ごとに集計した。

表１『明六雑誌』『太陽』各期における文体別程度副詞の使用記事数と粗頻度（総頻度順）

			明六 (1874/75)	太陽１ (1895)	太陽２ (1901)	太陽３ (1909)	太陽４ (1917)	太陽５ (1925)	計
おおいに	文語	記事数	51	262	143	73	19	2	550
		粗頻度	83	914	841	263	42	2	2145
	口語	記事数	0	15	69	163	139	140	526
		粗頻度	0	42	204	388	483	300	1417
	口語会話	記事数	0	2	11	14	11	12	50
		粗頻度	0	2	18	23	12	23	78
はなはだ	文語	記事数	51	207	136	79	23	0	496
		粗頻度	73	600	665	320	65	0	1723
	口語	記事数	1	17	75	127	114	103	437
		粗頻度	3	36	190	282	264	196	971
	口語会話	記事数	0	3	12	9	2	12	38
		粗頻度	0	4	23	12	2	17	58
すこぶる	文語	記事数	29	188	123	61	20	1	422
		粗頻度	40	583	564	192	44	1	1424
	口語	記事数	0	11	49	126	119	100	405
		粗頻度	0	22	108	291	267	217	905
	口語会話	記事数	0	1	4	7	1	3	16
		粗頻度	0	3	7	11	1	3	25
きわめて	文語	記事数	21	136	104	52	19	2	334
		粗頻度	30	282	394	133	35	3	877
	口語	記事数	0	8	36	131	114	133	422
		粗頻度	0	10	69	351	311	290	1031
	口語会話	記事数	0	0	0	9	1	6	16
		粗頻度	0	0	0	21	1	12	34
ひじょうに	文語	記事数	0	39	40	13	5	0	97
		粗頻度	0	59	104	27	7	0	197
	口語	記事数	0	9	45	122	117	153	446
		粗頻度	0	11	107	319	327	409	1173
	口語会話	記事数	0	0	8	8	7	9	32
		粗頻度	0	0	23	11	13	21	68
ずいぶん	文語	記事数	1	37	16	11	2	0	67
		粗頻度	1	53	25	27	2	0	108
	口語	記事数	0	21	67	108	78	107	381
		粗頻度	0	45	126	207	180	176	734
	口語会話	記事数	0	2	13	19	13	22	69
		粗頻度	0	3	27	39	41	33	143
よほど	文語	記事数	1	16	14	6	2	0	39
		粗頻度	1	20	19	9	6	0	55
	口語	記事数	1	13	56	98	73	86	327
		粗頻度	1	28	108	205	132	145	619
	口語会話	記事数	0	2	14	14	4	9	43
		粗頻度	0	2	19	22	12	11	66

いちじるしく	文語	記事数	0	27	32	9	6	0	74
		粗頻度	0	45	72	17	12	0	146
	口語	記事数	0	3	3	29	41	42	118
		粗頻度	0	3	3	45	90	59	200
	口語会話	記事数	0	0	0	0	0	0	0
		粗頻度	0	0	0	0	0	0	0
だいぶ	文語	記事数	0	11	1	6	1	0	19
		粗頻度	0	12	1	8	1	0	22
	口語	記事数	0	2	19	49	38	61	169
		粗頻度	0	2	30	80	67	85	264
	口語会話	記事数	0	1	7	10	8	13	39
		粗頻度	0	1	10	20	10	17	58
たいへん	文語	記事数	0	3	0	1	0	0	4
		粗頻度	0	4	0	2	0	0	6
	口語	記事数	0	4	23	19	24	33	103
		粗頻度	0	10	39	29	45	77	200
	口語会話	記事数	0	0	10	7	7	11	35
		粗頻度	0	0	12	10	16	46	84
しごく	文語	記事数	0	10	6	6	2	0	24
		粗頻度	0	13	14	7	3	0	37
	口語	記事数	0	4	21	29	13	26	93
		粗頻度	0	6	27	34	13	37	117
	口語会話	記事数	0	2	2	2	1	3	10
		粗頻度	0	3	2	2	1	3	11
いたく	文語	記事数	1	53	23	4	2	1	84
		粗頻度	1	100	47	4	2	1	155
	口語	記事数	0	1	9	5	4	8	26
		粗頻度	0	1	14	5	4	8	30
	口語会話	記事数	0	0	1	0	0	0	1
		粗頻度	0	0	1	0	0	0	1
ひどく	文語	記事数	0	4	2	1	0	0	7
		粗頻度	0	5	5	2	0	0	12
	口語	記事数	0	3	6	17	22	31	79
		粗頻度	0	4	10	25	41	46	126
	口語会話	記事数	0	0	3	5	6	7	21
		粗頻度	0	0	5	11	11	7	34
たいそう	文語	記事数	0	9	1	1	0	0	11
		粗頻度	0	14	1	2	0	0	17
	口語	記事数	0	8	10	14	9	18	59
		粗頻度	0	13	14	21	10	26	84
	口語会話	記事数	0	2	4	4	4	8	22
		粗頻度	0	2	7	9	4	11	33

※「おおいに」は「大に」表記分類困難のため、「おおきに」「おおいに」を合算した。

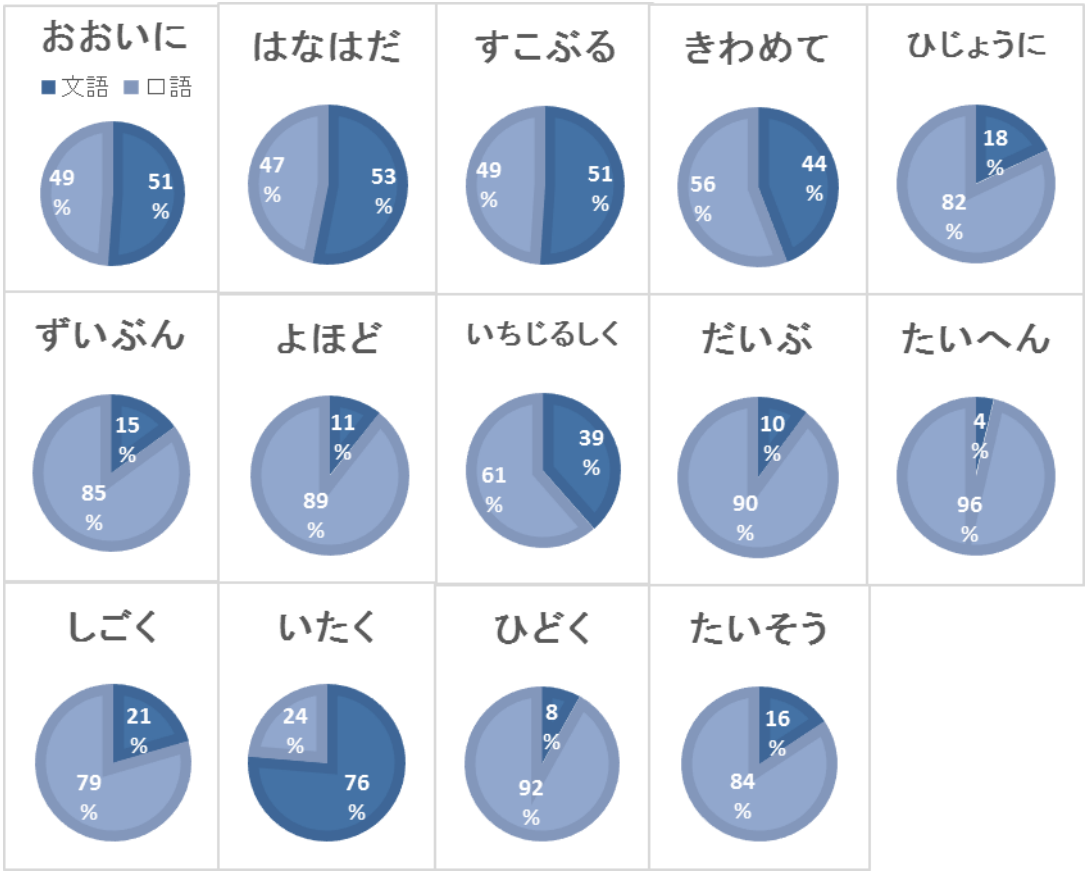
『明六雑誌』・『太陽』では、記事数ベースでは、「おおいに」(1076)が最も多く用いられ、以下「はなはだ」(933)「すこぶる」(827)「きわめて」(756)「ひじょうに」(543)「ずいぶん」(448)「よほど」(366)「いちじるしく」(192)「だいぶ」(188)「たいへん」(107)「しごく」(117)「いたく」(110)「ひどく」(86)「たいそう」(70)の順となった。

この結果は、松井(1977)や趙・祁(2010)による口語文・小説会話文を対象とした調査の結果とは大きく異なり、やはり文語文を含むという文体的な使用状況の差が影響しているとみられる。そこで表１の数値に基づき、次の表２には文語・口語別に、それぞれの出現記事率（口語・文語各出現記事数／口語・文語各全記事数）、表３には各語に就いての出現記事における文語・口語の比率を示す。

表 2『明六雑誌』『太陽』所載の文語・口語全記事中の文体別の出現記事率（%）
※総記事数は文語 1765 記事・口語 1781 記事

おおいに	文語	31.2	ずいぶん	文語	3.8	しごく	文語	1.36
	口語	29.5		口語	21.4		口語	5.22
はなはだ	文語	28.1	よほど	文語	2.2	いたく	文語	4.8
	口語	24.5		口語	18.4		口語	1.6
すこぶる	文語	23.9	いちじるしく	文語	4.2	ひどく	文語	0.4
	口語	22.7		口語	6.6		口語	4.4
きわめて	文語	18.9	だいぶ	文語	1.1	たいそう	文語	0.62
	口語	23.7		口語	9.5		口語	3.3
ひじょうに	文語	5.5	たいへん	文語	0.23			
	口語	25		口語	5.8			

表 3 各語の用例出現記事の文語・口語の比率



まず表２によると、文語記事において多く用いられる程度副詞は、「おおいに」・「はなはだ」・「きわめて」・「すこぶる」であり、口語記事において多く用いられる語は、「おおきに」・「ひじょうに」・「はなはだ」・「きわめて」・「すこぶる」・「ずいぶん」・「よほど」であった。「おおいに」・「はなはだ」・「きわめて」・「すこぶる」は文語・口語両方に用いられ、口語にはさらに「ひじょうに」「ずいぶん」「よほど」が加わるという状況であった。

また、出現記事の文体面での偏りを見る（表３）と、「いたく」「はなはだ」「おおいに」「すこぶる」「きわめて」「いちじるしく」は比較的文語記事で使われるケースが多い。このことから、「はなはだ」「おおいに」「すこぶる」「きわめて」が文語記事において頻繁に用いられ、「いたく」「いちじるしく」も、全体的な頻度としてはさほど多くはないが、語としては文語記事で使われる割合が高い語と言える。

４．文語での使用が多い副詞について

４．１　各語の使用状況と近世語との関係

本調査の結果を松井(1977)や趙・祁(2010)の調査に照らし合わせると、「おおいに」・「はなはだ」等、文語で多くみられる語であっても、口語文や小説会話文で多く用いられているものがあることがわかる。表２の結果ともあわせ、程度副詞類においては、文語で多用されるからといって、口語文で使われないというような制約はないと言える。「おおいに」「はなはだ」は、文語・口語共通の副詞として、また比較的会話文には分量が少ない「すこぶる」「きわめて」などは、文語・口語共通の書き言葉として広く用いられたと考えられる。

さて、前項において、「はなはだ」「おおいに」「すこぶる」「きわめて」「いたく」「いちじるしく」が、比較的「文語的色合いを帯びる語」と言えることを確認した。これらの語の大部分は、近代以前の資料においても文語的な文章での用例がみられるものである。

(4)又、唐書をよみて見るに、郭元振すがたはなはだうつくし。(『ひとりね』)
(5)これは決戦に比し候に大に劣り候やと被存候。(『白石先生手翰』)
(6)山ノ左右ノスソ長ク、形チ頗ル富士ニ似タリ。(『孔雀樓筆記』)
(7)此朝臣は（…中略…）父にも祖にも超て老中の職ともなりたり。されど、きはめて意地わろき根性ありて、(『折たく柴の記』)
(8)「蘭東事始」と題して、即函丈に進呈す。師、是を披閲し給ひ、「翁が素情、是にて償ひぬ」と、いたく喜ばせ給ひたり。(『蘭東事始』)

「はなはだ」「おおきに」「きわめて」「いたく」については、原(1970)において『今昔物語集』の程度副詞として挙げられたものでもあり、古くから見られるものである。また、「すこぶる」については、築島(1963)において漢文訓読語として挙げられており、途中意味的変遷を経たものの、(6)のように近世においては程度副詞として用いられている。

このように、これらの語は近世文語における程度副詞類の使用を受け継いだものと言える。また「いたく」以外は『明六雑誌』にも例が多く見え、その一方で「よほど」「たいそう」など近世口語文で用いられることの多かった他の語は殆どみられないことから、程度副詞に限れば、『明六雑誌』の文章は近世文語文を踏襲しているといえるであろう。

ただし「いたく」のみは『明六雑誌』ではわずか 1 例のみであった。太陽初期の文語文においては比較的多く用いられていたが、文語記事内での出現率は 1901 から 1909 では 4.8% から 1.5%と急落し、また口語文でも定着せず、衰退傾向にあると考えられる。

これら旧来の語に対して、「いちじるしく」は、本調査では近世期の文献には程度副詞化した例を得られておらず、また『和英語林集成』においても、程度副詞用法の記載はなく、『明六雑誌』にも使用は見られない。また、会話文には用例がみられず、松井(1977)の調査でも 32 位と下位であった。明治以前の具体的な使用状況は現時点では不明と言わざるを得ないが、この時期の書き言葉において発達しつつある述べ方と考えられる。

４．２　各語の使い分けと「いちじるしく」の意味

では各語はどのように使い分けられ、また「いちじるしく」はどのような特徴があった

のだろうか。

まず文語・口語両用の副詞として多用されている「大いに」については、古い語形「おおきに」に関する、『今昔物語集』等を対象とした原(1971)によれば「形容詞・形容動詞には係らず専ら同時を修飾する」特徴があるという。

この傾向については、(12)のような例外はあるが、10 例に満たない程度であり、近代雑誌においても動詞や動きを表す語に係るというのは同傾向であった。

(9)昔希臘の初めて興るや其民強健剛毅にして國を愛するの心深し（…中略…）然るに民風此時より破壊し驕奢淫佚輕薄狡猾の風大に行はれ國初の強健剛毅の氣悉く消磨し遂に羅馬の民に其國を奪はる（『明六雑誌』３「陳言一則」1874 文語）

(10)桂公も大いに意を強うし、翌日早速澁澤氏以下當代の實業家を招待して、同じく諒解を求められたのだといふ。（『太陽』9「時局に面して安田善次郎翁を想ふ」1925 口語）

(11)『僕は其氣で大いに勉強しなければならん（『太陽』11 喜劇「無能病」1909 口語会話）

(12)大變お手和らかで、大いに受けが好かつた。（『太陽』1「漱石先生と門下」1917 口語）

では、多くが形容詞にかかる「すこぶる」と「はなはだ」は何が異なるだろうか。

「すこぶる」は漢文訓読語由来の語で、プラス評価であろうがマイナス評価であろうが、肯定的・積極的に言い立てるような強調の意を表すとみられる。現代ではそして「*頗る少ない」では用いないとされ、「普通の程度をはるかに超えている」（『類語例解辞典』）とされる。近代雑誌資料においては、「少ない」ことを修飾する例はあるのだが、わずかであり、その一方で中尾(2003)に指摘されるように、「多い」「盛ん」に係る例が目立つ。

(13)我邦開港以來金銀の外出頗る多し近年最も甚だし大抵其源六なり(『明六雑誌』23 正金外出歎息録（貨幣四録の二）1874 文語）

(14)概して其運賃の高率なるのみならず、 列車の運轉度數亦頗る僅少であつて面白くない。（『太陽』13「清国鉄道の現況」1909 口語）

一方「はなはだ」は、否定的であっても後続事態を淡々と強める文脈でも用いうる。また現代語と異なり、「はなはだよし」などプラスの文脈でも用いられる。ただ、「微」や「少」の類で多く用いられる点や、(16）のような消極的評価となる否定文、また(17)のように「あまり」のような意でも用いられる点は、「すこぶる」とは傾向が異なる。

(15)然らずんば全國幾多の商業學校の勃興するも、其與ふる所の教育の効果、盖し甚だ微々たるものあらん。（『太陽』14「商業世界」1901 文語）

(16)作家が他の作家の畫に妄評を下したり、人身攻撃を加へたりするのは甚だよくないけれど、自己の主張を正々堂々と述べるに何の憚る所があらう。（『太陽』13「四季花鳥」1917 口語）

(17)若し國宗に異なる宗教を信ずる人民其數甚だ多からず然も政府之を國内に許し置くを不可と爲す時は其異宗人民各其所有の土地を賣却し家族と共に其國を去り又總て其家産を攜去するの通義を有すべし（『明六雑誌』6「ワツテル」1874 森有礼　文語）

中尾(2003)は「少ない」「稀」等に修飾する点に着目し、「消極的な評価」(p.136)を表すとした。このような特徴が長じて現代語では「マイナス寄り」（飛田・浅田、1994）「自分にとってマイナス」（『類語例解辞典』）というニュアンスにつながるのだろう。

また、会話文に見られる「はなはだ失礼」については「すこぶる」には例が見られない。

(18)いや、これはお待またせ致いたしまして、甚失禮を致いたしました。（『太陽』喜劇「まぜっかへし」1909 口語会話）

ではこれらの語の中で、「いちじるしく」は何を特徴として存在しえたのであろうか。目立つ点としては次のように比較や変化を表す構文に用いられることが挙げられる。

(19)實際發病の初期に於て貯藏物質が未だ發芽せざるものに比して著しく少なきこと、實驗の證明する所なり。（『太陽』13「農業世界」1901　文語）

(20)但だ江左に至りて、其の法の著しく發達せしや疑を要せざるべきか。（『太陽』3「字音の標準」1895　文語）

また、表現としては比較や変化の構文でなくても、次のように文脈的に比較や変化を背

景としている文で用いられている。

(21)近來青年文人の間に世を厭ふの思想著しく萌え出で爲に自殺を遂ぐるものすらあるは洵に慨はしき事共なり（『太陽』7「文学」1895　文語）

(22)次に事業經營上の成績の點から見ると遺憾ながら日本はヨーロツパに比較して非常に劣つて居る。 製品も粗惡であるし、原價が著しく高い。（『太陽』7「事業経営上より比較したる黄白両人種の優劣」1925　口語）

一方、「*いちじるしく嬉しくなった」などのように、自らの感情等は表さない。

つまり書き言葉において、何らかの比較を背景として客観的・観察的に（渡辺、2002 における「ひとごと」として）、後接する状態について「程度が甚だしい」「顕著である」ことを述べるのが、近代における程度副詞「いちじるしく」の特徴であると見られる。

５．口語的な副詞の使用状況

５．１口語会話文での使用状況

３では「ひじょうに」「ずいぶん」「よほど」「だいぶ」「たいへん」「しごく」「ひどく」「たいそう」が比較的口語的な語であることが分かったが、口語的性格をより詳細に検討するため、各語の全用例中、口語会話文での用例がどの程度占めるかを次の表４に示した。

表 4 各語の全用例における口語会話出現率（%）

たいへん	32.7	よほど	11.7	きわめて	2.1
たいそう	31.4	しごく	8.5	すこぶる	1.9
ひどく	24.4	ひじょうに	5.9	いたく	0.9
だいぶ	20.7	おおいに	4.6	いちじるしく	0
ずいぶん	15.4	はなはだ	4.1		

「たいへん」「たいそう」「ひどく」「だいぶ」の 4 語は 20%を超え、とりわけ「たいそう」と「たいへん」は、文語記事における用例もすべて会話文であり、特に話し言葉としての性質が強いものと言えよう。

一方、「ひじょうに」や「しごく」については、口語会話文での用例は少ない。特に「ひじょうに」は、表２を見ると『明六雑誌』では用いられないが、『太陽』1895 年・1901 年では文語文で多く用いられていることが目を引く。その後口語文でも、会話文では大きな動きが見られないものの、地の文において大きく数を伸ばしている。

このように、口語文で使用される語の中でも、「たいへん」「たいそう」など話し言葉寄りのものと、「ひじょうに」など主に書き言葉で使用されたものがあったことがわかる。

５．２「ひじょうに」と「たいへん」をめぐって

表２を経年的に見ると、近代雑誌資料において口語的な程度副詞の中で特に使用が拡大しているとみられるものに「ひじょうに」と「たいへん」がある。「ひじょうに」は『明六雑誌』では例が見られないものの、文語文・口語文両面で用いられ、『太陽』においては 1909 年以降書き言葉として使用が拡大している。一方「たいへん」は、口語記事の特に会話文において 1925 年の『太陽』で粗頻度第 1 位となっている。

この２語が伸長した契機は何だったのだろうか。まず両語の意味・機能面を確認する。

「ひじょうに」については、中尾(2003)が、下記の点を特色として挙げている。

①「変化動詞にかかることが多く、客観性や評価性のある表現となる」「客観性を表現する語で二人称、三人称の感情の程度を表現する語」②「形容詞・形容動詞の分析によれば『高い』や『大』などにみるように対象とする事柄が、『非常に』は具象的な事柄が多い」③「より客観性のある、変化を表す語や評価のよいものに関する語が多くみられた」（p.137　箇条書きは筆者）

しかしながら、実際の用例を観察すると、これらの分析には疑問が生じる。

まず「二人称、三人称の感情の程度を表現する」とされるが、たとえば「ひじょうに」の「喜ぶ」の例は口語文での例であり、文語文では「はなはだ」や「いたく」が用いられていて、「非常に」は用いられていない。文語・口語で分担している可能性があり、中尾の挙げた性質が「非常に」の他の副詞類と区別すべき本質的なものと断言するのは難しい。

また、「二人称・三人称」「客観性」という点についても次の例を見ると、疑問である。

(23)私は此事を曾て或る報告書で見た事もあるが、近頃ノースロツプ氏より直接概畧の話を聞て非常に感じた事である、(『太陽』7「樹栽日に就て」1895 口語)

(24)大層親切に世話をして呉れますので、 私は非常に喜んで居るんです(『太陽』11 喜劇「無能病」1909 口語会話)

(25)之を考へると私は非常に心細い。(『太陽』13「心頭雑草」1917 口語)

上の例ではいずれも「私は」が明示され、自らの感情や感覚について述べている。「二人称・三人称の」「客観的」なものではなく、「一人称」の「主観的」な文脈でも用いうるということになる。この差は、渡辺(2002)の「わがこと・ひとごと」の別にあたるもので、程度副詞類を分類する上で重要な観点であって、無視できない。

また「変化」構文にも表れることを特色としているが、この点については、むしろ動詞修飾を主とする「大いに」や比較を背景とする「いちじるしく」、話し言葉では「だいぶ」「ずいぶん」に顕著にみられる傾向である。中尾(2003)自身が「非常に」のかかり先として「高い」「大きい」を挙げているように、形容詞等の静的な状態性述語に修飾するのが基本であって、「非常に」のみに目立つ特徴とまではいえないように思われる。

また中尾（2003）は「積極的な評価を表現する」(p.136)とする。しかしながら、確かに「はなはだ」と比較すればその傾向はあるだろうが、「少」「低」などは「非常に」においても少なからずみられ、「はなはだ」以外の語と一線を画するほどの特徴とは言えない。

結局のところ、「非常に」は、形容詞・動詞を修飾し、「わがこと」にも使い得る、つまり程度副詞として、意味・機能的な面で比較的制限の少ない語であったとするのが、妥当と考える。だからこそ多くの類語がある中で伸長する契機があったと考えられるのである。

また「たいへん」についても、市村(2012)で挙げたように、「わがこと」についても表し得、意味・用法上制限の少ない語であったと考えられる。(26)のように「わがこと」を表す例もあれば、(27)のように「そうだ」と共起する客観的・観察的な例もあり、また(28)のように変化を表し、「少」を修飾する例もある。

(26)忌味は言ひツこなし、ね、私は今大變困つてるんですから、 虐待るのはよして頂戴よ、(『太陽』4「銅山王」1909 口語会話)

(27)急いで病院へ駈け着けて見ると、 病人の枕頭には、 前夜の通りに叔母と父親とがしよんぼり番をして居た。 朝の間は大變心持が好さ相に見えたが、 午後から何うも容態が好くない。(『太陽』1「反魂香」1917 口語)

(28)印象は大體いつもと同様だ。たゞ例年より左傾的な繪が大變少くなつたやうな氣がする。(『太陽』12「二科印象批評」1925 口語)

このように「ひじょうに」も「たいへん」も程度副詞としては制限の少ない語であったと言え、その点が近代～現代における伸長につながったと考えられる。

さて、この 2 語は、『日本国語大辞典』第二版等を見ると、両語とも明治期に程度副詞用法を派生したとみられるが、では同時期に発生した両語の「書き言葉的性質」「話し言葉性質」に明確に分かれることになった背景にはどのような事情があったのだろうか。

この点に関しては、程度副詞用法を派生する以前の状況が関わっているとみられる。

漢語「ひじょう」は、近世期には「たいへんだ」ほど口語的な用例は見られず、文語的な文の中で用いられている例が見られる。また明六雑誌においても、「非常の」が見られる。

(29)此全澤、博學の人なりしが、天性奇人にて、萬づ其好む所常人と異なりしにより、其良澤を教育せし所も亦非常なりしとなり。(『蘭東事始』)

(30) かくの如き非常の人は豈王法の刑罰に由てその道の行なはるを防ぎ遏むる㕝を得べけ

んや（『明六雑誌』37 賞罰毀誉論 1875 文語）

一方「たいへん」は、程度副詞用法を派生する前の形容動詞の用法がすでに口語的な語であり、その前身となる「大きな変事」を意味する名詞用法以外では、より話し言葉への偏りが大きく、江戸の戯作等話し言葉を反映する資料において用いられている。(32)は程度副詞の萌芽的例と言えるが、これも人情本の会話文の例である。

(31)北八「こいつは大變〳〵（『東海道中膝栗毛』）

(32)丹「ヲヤ〳〵大変につもつたぜ、こりやアおゐねへ（『春色辰巳園』）

「たいへん」と「非常に」は、程度副詞用法を派生する前の段階において既に文体的に差があるとみられ、そのような状況が、『太陽』での「大変」が話し言葉的で「非常に」が書き言葉的であるという差に反映しているのであると考える。

飛田・浅田(1994)によれば、現代における「非常に」は「やや硬い文章語」であり、「たいへん」は「慨嘆・驚き・感動・丁重さなどさまざまな暗示を伴う」(p.443)という。

近代雑誌における両語の伸長過程の違いは、このような現代語における両語のもつニュアンス違いに通じるのである。

文　献

市村太郎(2012)「副詞「たいそう」の変遷―近代語を中心に―」『国文学研究』167

近藤明日子(2013)「近代雑誌総合記事に出現する一人称代名詞の分析―単語情報付『太陽コーパス』を用いて」近代語学会編『近代語研究』17、pp.133-154、武蔵野書院

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」国立国語研究所編『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』、pp.1-48、博文館新社

趙宏・祁福鼎 (2010)「近代語の会話文における程度副詞の使用状況をめぐって―『高い程度を表す』副詞を中心にして―」『明治大学日本文学』35

築島裕(1963)『平安時代の漢文訓読語につきての研究』東京大学出版会

中尾比早子(2003)「明治・大正期における程度副詞「非常に」について」田島毓堂・丹羽一彌編『名古屋・ことばのつどい　言語科学論集』、pp.127-138、名古屋大学大学院文学研究科

――――(2005)「副詞「とても」について―陳述副詞から程度副詞への変遷―」国立国語研究所編『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』、pp.213-226、博文館新社

原栄一(1970)「程度副詞おほきに小考」『金沢大学教養部論集人文科学篇』8

飛田良文・浅田秀子(1994)『現代副詞用法辞典』東京堂出版

松井栄一 (1977)「近代口語文における程度副詞の消長―程度の甚だしさを表す場合―」松村明教授還暦記念会編『松村明教授還暦記念　国語学と国語史』、pp737-758、明治書院

渡辺　実 (2002)『国語意味論』塙書房

※辞書・資料類（資料類の引用に際しては振り仮名を省略するなど表記を一部改めた。）

『日本国語大辞典』第二版 小学館 2001

『使い方のわかる類語例解辞典』小学館 1994

中村幸彦他『近世随想集』日本古典文学大系 岩波書店 1965

松村明他校注『新井白石』日本思想大系　岩波書店 1975

小高敏郎他校注『戴恩記　折たく柴の記　蘭東事始』日本古典文学大系　岩波書店 1964

麻生磯次校注『東海道中膝栗毛』日本古典文学大系　岩波書店 1958

中村幸彦校注『春色梅兒誉美』日本古典文学大系　岩波書店 1962

関連 URL

国立国語研究所 近代語のコーパス http://www.ninjal.ac.jp/corpus_center/cmj/

最高裁判所判決の文体特徴の経年変化

矢野 信（株式会社法学館法教育研究所）†

Temporal Change of Stylistic Features in the Judgment Documents of the Supreme Court of Japan

Makoto Yano (Japan Research Institute of Law Related Education, HOUGAKUKAN CO.,LTD.)

１．はじめに

最高裁判所の判決・決定は，法律の分野における書き言葉の中においてある種の代表的な位置付けを有していると考えられる。そのような性格を有する最高裁判所の判決文（最高裁判決[1]）のうち，比較的最近のもの（特に平成年代のもの）は，接続詞の用法，言い回しのバリエーションなどの点において，ばらつきがかなり少ない，統制のとれた文体で書き表されているように感じられた（これは筆者の経験からくる実感である。）。

本発表では，そのような現在の最高裁判決の文体特徴や，それがどの時期に成立したのかを明らかにすることを目的として，最高裁判決のデータを分析する。

２．戦後の「判決文」の文体についての概観

２.１　最高裁判所の発足

明治憲法下での裁判所の判決文は，文語体であり，句読点なし，濁点のないカタカタ表記であった。これに対して，日本国憲法（新憲法）が施行された1947（昭和22）年に発足した最高裁判所の判決では，当初から，口語体・ひらがな・句読点あり・濁点ありの表記が用いられた。新憲法自体が口語体・ひらがなで表記されており，新憲法下での法令や公文書もすべてそれにならった。判決書も同じであった。

なお，判決における理由の判示では当事者の主張（準備書面・上告理由など）を一部引用してから述べることがあるが，そのようにして引用される当事者の主張の部分には，最高裁発足後もしばらくの間は文語体・カタカナ表記が見られる。また，明治憲法下から存続している法令は，文語体・カタカナ表記のままであった[2]。そのような法令が判決中に引用された場合にも，文語体・カタカナ表記が部分的に見られることになる。

２.２　第一審判決の簡易化の動き

1990（平成2）年に，東京高裁・地裁，大阪高裁・地裁の民事判決書改善委員会が共同で，「民事判決書の新しい様式について」と題する提言を行った(東京高裁・地裁民事判決書改善委員会，他1990)。それまで判決書の作成が裁判官の負担となり充実した審理から逆行していたことへの反省[3]と，当事者が真に知りたいことに簡明に答える判決書を目指し，判決書様の抜本的な改革がなされた。これは，現在ではほぼ完全に普及している[4]。

---

† klagegrund@gmail.com

[1] 以下，特に断りのない限り，「最高裁判決」には「決定」も含むものとする。それらの関係については，矢野(2013)を参照。

[2] 現時点（2014 年 1 月）においても，有効なものとして残っている文語体・カタカナ表記の法律が複数ある（手形法など）。

[3] 審理よりも判決書の作成に心理的なウエイトが置かれ，不十分な審理のままで口頭弁論が終結し，判決書の起案の段階で初めて事案の吟味に入るという悪循環が生じていたとされる(三宅 1994)。

[4] 同提言以前の様式（在来様式）による判決書が現在でもしばしば見られるが，それは，事件・法律問題の種類によっては在来様式による判決書の方が分かりやすい，という判断に基づくものと考えられる。

また，1991（平成3）年には，東京地裁・大阪地裁刑事判決書検討グループが刑事裁判の判決書の見直しについての提言を行い，刑事事件の判決についても，簡易化を目指す改革が行われた(東京地裁・大阪地裁刑事判決書検討グループ1991)。そこでは，古めかしい表現や長文を避けるなどの提案がなされている。

２.３　裁判員制度の施行

2009（平成21）年5月に施行された裁判員制度は，刑事裁判に国民の健全な社会常識を反映することを目的としたものである(司法制度改革審議会2001)。制度の施行に先立って，裁判員裁判における判決書の在り方について様々な検討が行われた。そのうち，司法研修所編(2009)では，①重要な争点にポイントを絞った平易・簡潔なものとする，②端的に検察官の主張する事実が立証されているかの理由・結論を示す，③結論を導いた実質的な理由が示されていることを要する，との3点を挙げた上で，従来の判決書に比べて圧倒的に短い判決書のサンプルを掲げている（A4で2～3ページ等）。

２.４　最高裁判所判決の文体との関係

前2項目はいずれも地方裁判所の判決文の文体についての変動要因であって，最高裁判決の文体に直接触れるものではないが，無関係ではないであろう（むしろ，実は先導的に変化していったという可能性も考えられる。）。それらの点の解明につながる要素も含めて，最高裁判決の文体特徴を明らかにしていきたい。

３．分析対象としたデータ

今回，分析に用いたデータは，最高裁判所ホームページの「最高裁判所判例集」[5]に収録されている最高裁判所判決・決定のPDFデータをテキストデータ化したものである。

年代の範囲は，1947（昭和22）～2012（平成24）年であり，全体で27,904件，33,947,691字，24,109,775語[6]である。このデータは，年によって収録件数・文字数にかなり大きなばらつきがあるが（詳細は矢野(2013)），今回は特に調整を行わず，全文を検索対象とした。

また，検索対象のデータは，すべて，法廷意見の部分と個別意見の部分に区分した。最高裁判所の判決では，各裁判官が個別に意見を述べることができるが（裁判所法11条），そのようにして各裁判官が個別に述べた部分を個別意見[7]，それ以外の，裁判官全員として述べた部分を法廷意見という（多数意見ともいう。）。多数意見部分は最高裁全体としての文章であるのに対し，個別意見部分は各裁判官が自由に述べることができるものであることから，前者に比べて後者は，文体特徴の統制が弱いことが想定される。このことから，あらかじめ両者を区分しておいた。なお，法廷意見部分が21,536,876語，個別意見部分が2,572,899語であった。

４．分析１：接続詞のバリエーションについて

４.１　意義・目的

文体特徴とは，文字・単語から文・段落・文章に至る文章の各要素に関する書き手の独

[5] 紙媒体で刊行されている「最高裁判所民事判例集」,「最高裁判所刑事判例集」,「最高裁判所裁判集民事」,「最高裁判所裁判集刑事」に収録されている最高裁判所判決・決定の本文（個別意見を含む）部分が PDF で検索可能なデータベースである（矢野 2013）。http://www.courts.go.jp/

[6] 形態素解析器 MeCab, 形態素解析辞書 UniDic を用いて形態素解析（短単位）した結果の数から, 感動詞, 記号，その他，フィラーを除いた数

[7] 個別意見には，法廷意見（多数意見）と反対の結論とする「反対意見」，法廷意見を補足する「補足意見」，法廷意見と結論を同じだが理由が異なる「意見」の 3 種類がある。今回はその区別は行っていない。

特の構成パターンを指すとされる(金2009)。その要素には様々なものがあるが，今回は，接続詞のパターンに着目した。接続詞とは，文頭において，先行する文とのつながりを示す役割を果たすものである(益岡・田窪1992)。これは，先行する文との論理的関係を表すために用いることができる。判決における文章は，結論（判決の主文）に至る理由を，事件の事実関係を法律の条文（その解釈を含む。）に照らしたときの法的判断として論理的に説明していくものであることから，接続詞の用法が重要な意味を持ってくる。

法的判断の際に用いられる論理的関係のパターンは，おそらく一定の数に限られている。このことから，判決文に登場する接続詞は，一般に接続詞といわれるもののうちの一部に限られることが予想される（観点1）。また，同じ論理的関係を表すのに一般には複数の接続詞を使うことができる場合であっても，法的判断・司法判断の場面であるという要請から，どれか一通りの接続詞が専ら好んで用いられる（用いられるようになっていく），ということがありそうである（観点2）。最後に，表記（漢字，送り仮名）の上で複数のバリエーションがある接続詞であっても，いずれか一通りの表記がなされるということがあり得る（観点3）。このうち，観点2，3は，文体の統制がとれていることと密接に関係すると考えられる（現段階では，統制が人為的なものであるか否かは区別しない。）。

以上のような観点で，最高裁判決に登場する接続詞のバリエーションを見ていく。

４.２　検索方法

前記３のデータを，形態素解析器MeCab，形態素解析辞書UniDic を用いて形態素解析した結果で，品詞「接続詞」となっている単語を単純に集計した。つまり，短単位の1語のレベルだけを対象としたため，接続詞相当句といわれるものは含まれていない。

４.３　結果１　接続詞の種類数の変化（全体）

5年ごとの区間（例えば1955年区間は1951～1955年を表す。ただし，1950年区間は1947（昭和22）年から，2015年区間は2012（平成24）年まで）に分けて，出現する接続詞の種類数の変化を示したのが図1 である（法廷意見・個別意見に区分）。それぞれを，書字形レベルと原形レベルに分けて示した。

MeCabの辞書（初期状態）には接続詞として171語（書字形レベル）が登録されていることと比べると，全体的に，判決文で用いられる接続詞はかなり限られているといえそうである（観点1，2）。

また，法廷意見においては，書字形レベルでの数が明らかに減少しているが，これは，表記の統一がとられていった経緯を想像させる（観点3に関係）。法廷意見では，原形レベルでの数も微かに減少傾向が見られる（観点2に関係）。

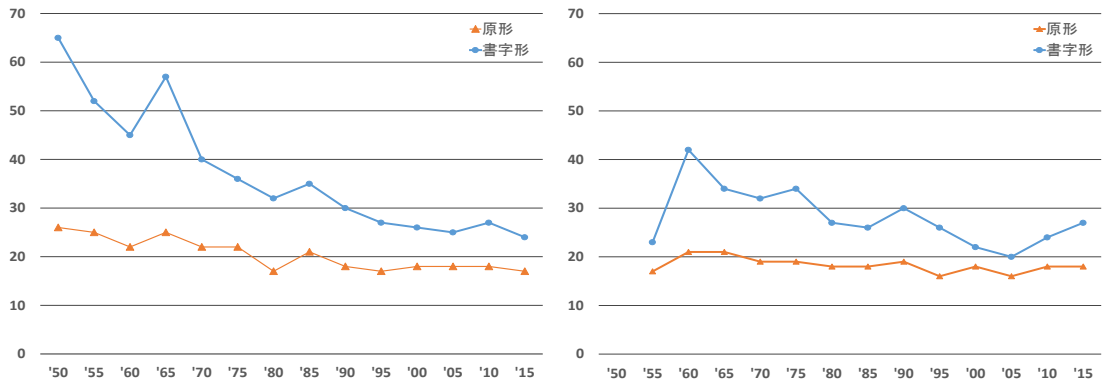


図１　使われている接続詞の種類数の変化（左：法廷意見，右：個別意見）

一方，個別意見においては，減少傾向といえるものは見られなかったが，原形レベルで，用いられている接続詞の種類数は，法廷意見のそれとほぼ同じであるといえる（観点1）。

４.４　結果２　接続詞の比率

全区間を通じて，法廷意見ベースで，頻度の高い10種類を挙げたのが，表1 である。なお，「種類」欄は，原形を同じくするものを包括し，その中で書字形の頻度が高いものをもってラベルとした。

この上位10種類で，使用される接続詞全個数の92.7%を占めており，最高裁判決で用いられる接続詞が，極めて限られていることが分かる（観点1，2）。また，法令では頻繁に用いられる「若しくは」がここに含まれていない（同様に「並びに」も含まれないが，「並びに」は短単位2語に解析されていたため，ここでは検出されなかった。）。

表1 最高裁判決で使用される接続詞の比率（%）

種類	又	及び	しかし	そして	かつ	ないし	なお	すなわち	されば	更に
比率	28.65	25.01	8.37	7.11	6.38	5.21	4.93	2.53	2.49	2.03

４.５　結果３　表記の移り変わり

代表的な接続詞について，表記の移り変わり（観点3）を示したのが，図3（次頁）である。これは，法廷意見部分のみを対象として，単純頻度をグラフ化したものである。なお，「又は」についてのみ，形態素解析結果の 接続詞：又 ＋ 書字形：は／ハ の個数を数えたものである。

これを見ると，「更に／さらに」以外は，1980年区間頃にはほぼ，一通りの表記に収束していることが分かる。使われなくなった表記が比較的最近にもごく少数現れているのは，古い法令の引用である可能性がある（特に文語表現の場合）。これに対して，「更に／さらに」は，どちらが優位であるということもなく，また，歴史的に，多数派が入れ替わっているという意味で，極めて特殊な分布である。

「更に／さらに」を除けば，公文書における表記の統一が行われたのに合わせて，表記の統一が事実上行われていったという可能性が考えられる。

５．分析２：「のだ文」の使用頻度の推移

５.１　意義・目的

文体特徴の分析のもう一つとして文末表現としての「のだ／のである」を取り上げる。いわゆる，「のだ文」とは，文末形式が「のだ」「のである」「んだ」「んです」などで終わる文の総称であり，この「のだ」とは，準体助詞「の」に判断の助動詞「だ」が付いたものであるとされる。そして，「のだ」は，一語化され，文末にくると，その文は「説明」「判断」「命令」などとして働く，という見方が一般的である(益岡・田窪1989)。

ここにいう，“説明”“判断”という語（語感）からすると，「のだ文」は，判決文に多く登場してもよさそうなものであるが，実際には（特に最近の最高裁判決には）あまり登場しない。その理由の考察につながる第一歩として，最高裁判決における「のだ文」の頻度，使われ方を見てみる。

５.２　検索方法

前記３のデータの法廷意見部分を，直接，全文検索した。具体的には，直前の一文字が「も」以外となる「のである。」を検索した（「ものである」は，別の表現になるため）。

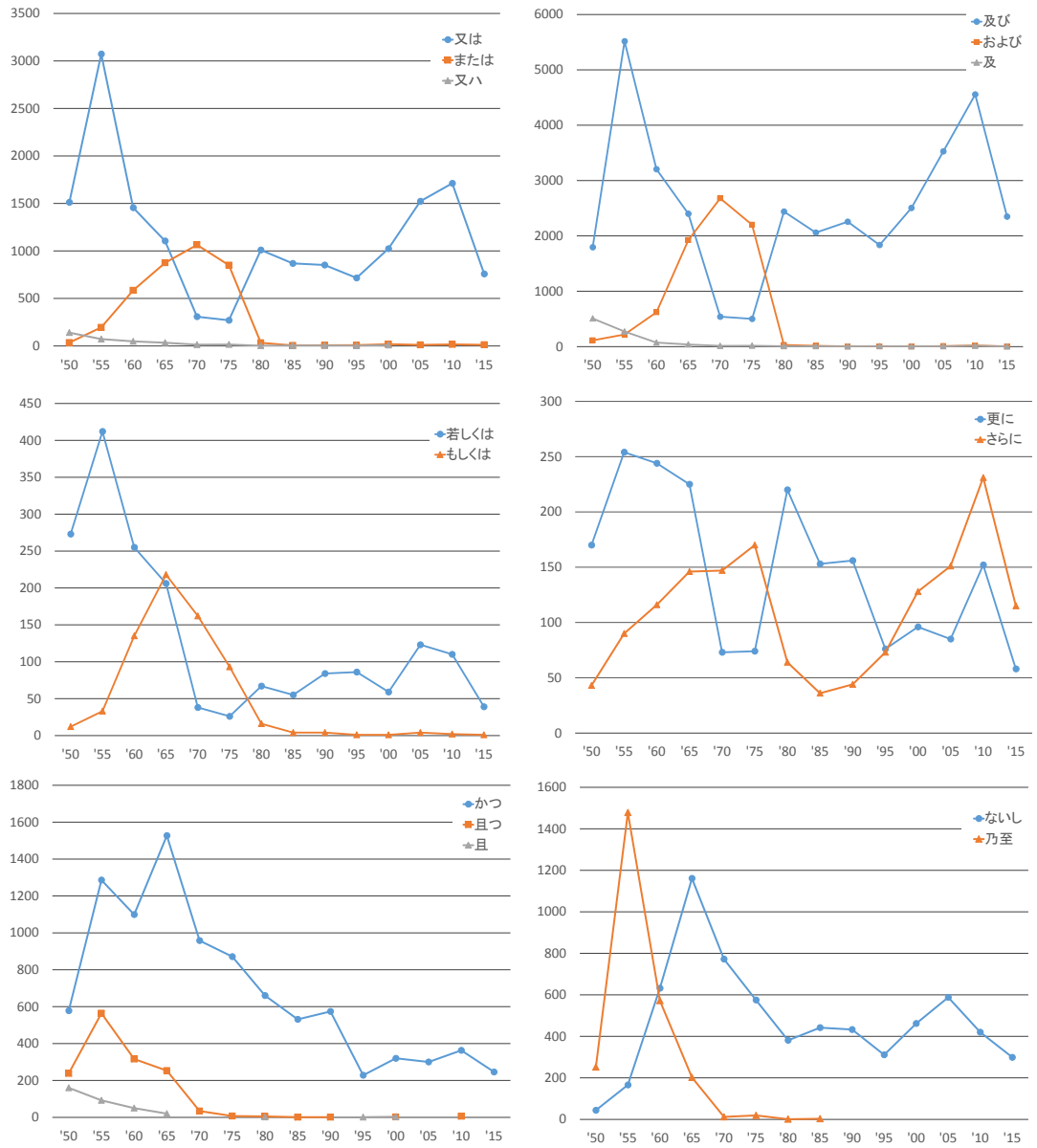


図 2　各接続詞の表記の移り変わり（縦軸は頻度（個数））

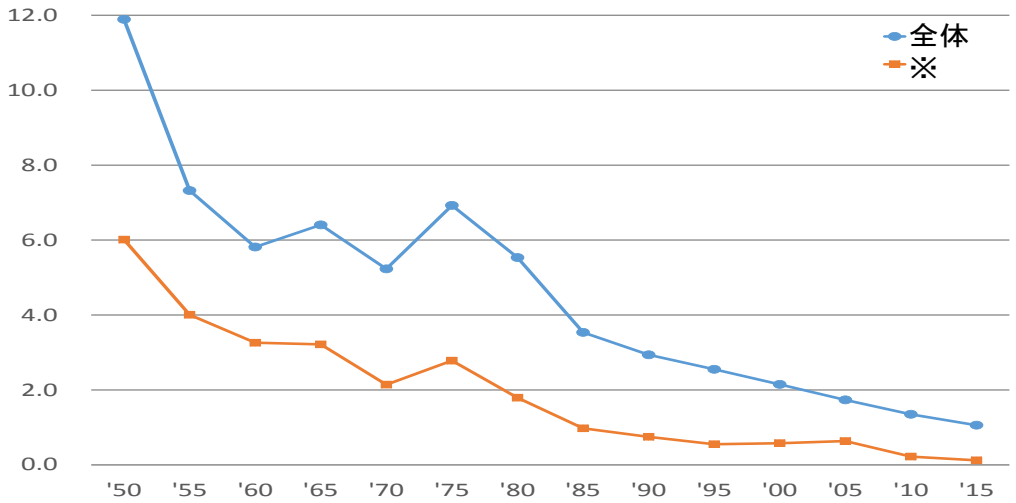


図 3　「のである。」の 10,000 語当たりの頻度（※は「というのである。」を除いたもの）

なお，判決文の文末はほとんどが「である。」形であり，「だ。」形が使われることは稀である（実際，対象データで「のだ。」を全文検索したが，ヒットしたのは8件のみであり，その全てがカギ括弧引用の内部（セリフ等）であった。）。

５.３　結果

法廷意見部分のテキストを対象として，5年ごとの区間に分けて，10,000語当たりの「のである。」の数を示したのが図2 （●印）である。これによれば，戦後すぐの頃はそれなりの数が使われていたものが，ほぼ一貫して減少していることが分かる。また，その中のかなりの割合は，「というのである。」という形であった。

最高裁判決で使われる「というのである。」は，最高裁自身の判断部分に先立って，高裁判決に対する弁護人の上告理由等の内容をまとめて「……というのである。」として提示する，という使われ方がほとんどであった。これは，最高裁自身による“説明”“判断”を述べているわけではないといえる。

そこで，図3 には，出現形「というのである。」を除外したものを■印で示した。これによれば，1985年区間以降は1.0を下回っているが，これは，年間に100件近くある判決・決定（最高裁判例集に収録される件数）の中で数〜十数か所といったオーダーである。

このことから，「のだ文」は最高裁判決ではほぼ“絶滅”したといってよいと思われるが，これは，「のだ文」の持つ特殊な（主観的な）ニュアンスを嫌っていったと考えることができそうである（逆に戦後すぐの頃には多く見られた要因も併せて検討できれば興味深い。）。

６．おわりに

今回は，最高裁判所の文体特徴の一つとして，接続詞のバリエーションと「のだ文」の頻度を見ていった。前者においては，一定の論理的関係を表すのに特定の表現・表記を用いるように統一化されていったという傾向が，後者においては，特殊なニュアンスを持つ「のだ文」が嫌われていったという傾向が見られた。今後，今回の結果をさらに深く分析するとともに，他の文体特徴についても調べていきたい。

文　献

金明哲(2009)「計量文献学」『計量国語学事典』朝倉書店，pp.238-248

司法研修所編(2009)『裁判員裁判における第一審の判決書及び控訴審の在り方』法曹会

司法制度改革審議会(2001)「司法制度改革審議会意見書―21世紀の日本を支える司法制度」pp.102-108（http://www.kantei.go.jp/jp/sihouseido/report-dex.html よりダウンロード可能）

東京高裁・地裁民事判決書改善委員会，他(1990)「民事判決書の新しい様式について」判例タイムズ，5，pp4-35

東京地裁・大阪地裁刑事判決書検討グループ(1991)「刑事判決書の見直しについて（提言）」判例タイムズ，15，pp10-23

益岡隆志・田窪行測(1989)『日本語の文法の研究』くろしお出版

益岡隆志・田窪行則(1992)『基礎日本語文法』くろしお出版

三宅弘人(1994)「判決文の簡易化について」日本語学，13，pp.71-79

矢野信(2013)「言語資料としての『判決文』の分析にまつわる問題点」，第4回コーパス日本語学ワークショップ予稿集，pp.291-298
（http://www.ninjal.ac.jp/event/specialists/project-meeting/m-2013/jclws04/ よりダウンロード可能）

書　名　第5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3月3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）

(C) 2014 Department of Corpus Studies/Center for Corpus Development, NINJAL