＜全文＞言語資源活用ワークショップ2016発表論文集

著者	国立国語研究所コーパス開発センター
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	1-428
発行年	2017
URL	http://doi.org/10.15084/00001450

言語資源活用ワークショップ 2016

発表論文集

2017 年 3 月 6 日 (月)　『語彙資源活用シンポジウム』
2017 年 3 月 7・8 日 (火・水)　『言語資源活用ワークショップ 2016』

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 編

Programme

Programme:語彙資源活用シンポジウム

2017 年 3 月 6 日 (月)

【セッション 1】 (2F 講堂)

10:10-10:15　**趣旨説明**
……浅原正幸 (国立国語研究所)

10:15-10:45　**『UniDic』の拡張計画**
……岡照晃 (国立国語研究所)

10:45-11:15　**単語分かち書き用辞書『mecab-ipadic-NEologd』を公開して得た知見について**
……佐藤敏紀 (LINE)

11:15-11:45　**拡張型 NLP『JMAT』における実利用に向けた形態素解析のリソースチューニング**
……北浦雅子・紀伊馬章 (ジャストシステム)

11:45-13:00　休憩

【セッション 2】 (2F 講堂)

13:00-13:30　**『JUMAN++』の大規模語彙獲得へ向けた取り組み**
……森田一 (京都大学)

13:30-14:00　**『分類語彙表』の特徴と問題点**
……山崎誠 (国立国語研究所)

14:00-14:15　休憩

【セッション 3】 (2F 講堂)

14:15-14:45　**『日本語歴史コーパス』に出現した新規語の『UniDic』への登録について**
……鴻野知暁 (国立国語研究所)

14:45-15:15　**『日本国語大辞典』の編集方法—これまでとこれから**
……佐藤宏 (小学館)

15:15-15:45　**中型国語辞典『大辞林』編集と見出し語の収集・選定について—未知語・新語を中心に**
……山本康一 (三省堂辞書出版部)

15:45-16:00　休憩

【パネルセッション】 (2F 講堂)

16:00-17:00　パネルセッション・総合討論

Programme:言語資源活用ワークショップ 2016

2017 年 3 月 7 日 (火)

10:00-10:15　■挨　拶 (2F 講堂)　前川喜久雄

10:15-11:05　■口頭発表　A グループ (2F 講堂)

[O-A-1]

国語教科書と高校生作文の複文構造比較—従属節の構造と節形式の量的比較—

..... 松本理美 (立命館大:学生)

[O-A-2]

友人への「断り」に対する評価に関する質的考察 —日本語母語話者と中国人日本語話者の評価を通して—

..... 滕越 (東京大:学生)

11:05-11:55　■招待講演 (2F 講堂)

[I-1]

講演・講義の音声認識と字幕作成へのコーパスの活用

..... 秋田祐哉 (京都大学)

12:00-13:00　休憩

13:30-15:00　■『国語研日本語ウェブコーパス』検索系『梵天』デモ (2F セミナー室 238 室)

13:00-14:15　■ポスター発表　A グループ (2F フロア・多目的室)

[P-A-1]
もし小学生が『現代日本語書き言葉均衡コーパス』並みに漢字を使ったら
……今田水穂 (文部科学省)

[P-A-2]
コーパス構築における発話アライメントの現状
……石本祐一 (国語研)

[P-A-3]
発話文への発話者情報付与の基本設計 ―『現代日本語書き言葉均衡コーパス』収録の小説を対象に―
……宮嵜由美・柏野和佳子・山崎誠 (国語研)

[P-A-4]
夢梅本『倭玉篇』全文テキストデータベースの構築
……高橋大希・劉冠偉 (北海道大:学生)・池田証壽 (北海道大)

[P-A-5]
『日本語諸方言コーパス』の構築について
……木部暢子・佐藤久美子・中西太郎 (国語研)
中澤光平 (与那国町与那国語辞典編集業務嘱託員)

[P-A-6]
相談における談話構造　―修辞機能と脱文脈化の観点からの分析―
……田中弥生 (国語研・東京大:学生)

[P-A-7]
『UniDic』と『分類語彙表』の見出し対応表データの構築
……近藤明日子 (国語研)・田中牧郎 (明治大)

[P-A-8]
『名大会話コーパス』の比較に基づく教室談話における「中途終了型発話」の特徴
……矢田真菜 (東京学芸大:学生)

14:15-14:20　休憩 (ポスター切替)

14:20-15:35 **■ポスター発表　B グループ (2F フロア・多目的室)**

[**P-B-1**]
『多言語母語の日本語学習者横断コーパス』の母語話者データにおけるタスクと産出語彙の関連
..... 小西円 (国語研)

[**P-B-2**]
『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行
..... 加藤祥・浅原正幸・山崎誠 (国語研)

[**P-B-3**]
「大規模日常会話コーパス」プロジェクト—コーパスに基づく話し言葉の多角的研究—
..... 小磯花絵 (国語研)

[**P-B-4**]
日本語語構成情報データベースの構築
..... 淺尾仁彦 (情報通信研究機構)

[**P-B-5**]
発話文自動生成のための日本語表現文型辞書の作成
夏目和子 (名古屋大)・刀山将大 (名古屋大:学生)・佐藤理史 (名古屋大)

[**P-B-6**]
スマホで古辞書 —『篆隷万象名義』の IDS 検索を例に—
.....劉冠偉・李媛 (北海道大:学生)・池田証壽 (北海道大)

[**P-B-7**]
機械翻訳用超大規模辞書データ資源
..... 春遍雀來 (日中韓辭典研究所)

[**P-B-8**]
モンゴル語アクセント研究のためのデータベース
.....玉栄 (内モンゴル大・国語研)・西川賢哉・前川喜久雄 (国語研)

[**P-B-9**]
多重の読みを持つテキストのコーパス化
..... 小木曽智信 (国語研)

15:35-15:45 **休憩**

15:45-17:25　**■口頭発表　B グループ (2F 講堂)**

[O-B-1]

次元形容詞にみる母語話者らしい日本語形容詞の使用

. . . . . 西内沙恵 (国語研・立教大)

[O-B-2]

日本語コーパスの包括的検索環境の実現に向けて

前川喜久雄・浅原正幸・小木曽智信・小磯花絵・木部暢子・迫田久美子 (国語研)

[O-B-3]

機能語用例文データベース『はごろも』の今後の展開

. . . . . 堀恵子 (東洋大・筑波大)・内丸裕佳子 (岡山大)・加藤恵梨 (朝日大)
小西円・山崎誠 (国語研)・江田すみれ (日本女子大)
建石始 (神戸女学院大)・中俣尚己 (京都教育大)・李在鎬 (早稲田大)

[O-B-4]

日本語学習者コーパスの教育応用における留意点—『多言語母語の日本語学習者横断コーパス』に見る母語話者 L1 産出データの安定性検証を中心に—

. . . . . 石川慎一郎 (神戸大)

18:00-19:30　**■懇親会**

2017 年 3 月 8 日 (水)

10:10-11:00 **■口頭発表　C グループ (2F 講堂)**

[O-C-1]

漢語の仮名表記—実態と背景—

. . . . . 間淵洋子 (明治大:学生)

[O-C-2]

『日本語歴史コーパス』短単位アノテーション作業効率化に向けた形態素解析用辞書『UniDic』の段階的特殊化の検討—近松コーパスを例として—

. . . . . 岡照晃 (国語研)

11:00-11:50 **■招待講演 (2F 講堂)**

[I-2]

言語資源の設計・再設計と言語資源を活用した実習授業の設計

. . . . . 松吉俊 (電通大)

11:50-13:00 **休憩**

13:00-15:30 **■『国語研日本語ウェブコーパス』検索系『梵天』デモ (2F セミナー室 238 室)**

13:00-14:15　**■ポスター発表　C グループ (2F フロア・多目的室)**

[P-C-1]

全文検索システム『ひまわり』における言語分析支援機能の拡張

. . . . . 山口昌也 (国語研)

[P-C-2]

児童生徒の「手」作文に於ける経年変化の計量的分析

阿部藤子 (東京家政大)・今田水穂 (文部科学省)・宗我部義則 (お茶の水女子大付属中)
冨士原紀絵 (お茶の水女子大)・松崎史周 (日本女子体育大)・宮城信 (富山大)

[P-C-3]

『日本語日常会話コーパス』収録の進捗状況

. . . . . 田中弥生・柏野和佳子・角田ゆかり (国語研)
伝康晴 (千葉大)・小磯花絵 (国語研)

[P-C-4]

『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消

. . . . . 鈴木類 (茨城大:学生)・古宮嘉那子 (茨城大)・浅原正幸 (国語研)
佐々木稔・新納浩幸 (茨城大)

[P-C-5]

形態素解析ソフトウェア 『Web 茶まめ』の改良と Web API の試作

. . . . . 川口寛治・薦田龍輝 (東京電機大:学生)・堤智昭 (東京電機大)

[P-C-6]

『現代日本語書き言葉均衡コーパス』を用いた「～ ていく」「～ てくる」構文の意味分析

. . . . . 加藤麟太郎 (東京大:学生)・藤井聖子 (東京大)

[P-C-7]

明治初期教科書『物理階梯』のコーパス作成による語彙の考察

. . . . . 田中牧郎 (明治大)・島田むつみ・髙橋雄太 (明治大:学生)

[P-C-8]

話し言葉コーパスの転記タグ :『多言語母語の日本語学習者横断コーパス』と『日本語話し言葉コーパス』の比較

. . . . . 西川賢哉 (国語研)

[P-C-9]

『日本語日常会話コーパス』の転記基準と作業工程

. . . . . 川端良子 (国語研・千葉大:学生)・臼田泰如・西川賢哉 (国語研)
徳永弘子 (国語研・東京電機大)・小磯花絵 (国語研)

14:15-14:20　**休憩 (ポスター切替)**

14:20-15:35 **■ポスター発表　D グループ (2F フロア・多目的室)**

[P-D-1]
『現代日本語書き言葉均衡コーパス』と『分類語彙表』を利用した漢字 3 文字略熟語の抽出
……山崎誠 (国語研)

[P-D-2]
名詞項構造付与データの構築
……竹内孔一 (岡山大)

[P-D-3]
『名大会話コーパス』中納言版・ひまわり版公開データの作成
……柏野和佳子・西川賢哉・小磯花絵 (国語研)

[P-D-4]
『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション—基準策定，仕様書作成の必要性について—
……松本理美 (立命館大:学生)・浅原正幸 (国語研)・有田節子 (立命館大)

[P-D-5]
『日本語話し言葉コーパス』における発声様式の自動分類
森大毅 (宇都宮大)・藤本雅子 (国語研)・浅井拓也 (北陸先端大:学生)・前川喜久雄 (国語研)

[P-D-6]
近代文語文の通時的変化の分析　—語種率・品詞率に着目して—
……近藤明日子 (国語研)

[P-D-7]
結合の強度を測る指標としての Log-r の有用性：日・英語のバイグラムデータに基づく MI、LLR などとの比較
……藤村逸子 (名古屋大)・青木繁伸 (群馬大)

[P-D-8]
語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成
……長谷川守寿 (首都大)・西尾広美 (国語研)

[P-D-9]
固有表現抽出におけるアノテーション手法の比較
……鈴木雅也 (茨城大:学生)・古宮嘉那子 (茨城大)
岩倉友哉 (富士通研)・佐々木稔・新納浩幸 (茨城大)

15:35-15:45　休憩

15:45-16:35　**■口頭発表　D グループ (2F 講堂)**

[**O-D-1**]

『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの分析

. . . . . 宮内拓也 (国語研・東京外大:学生)・浅原正幸 (国語研)
中川奈津子 (千葉大・学振)・加藤祥 (国語研)

[**O-D-2**]

読み時間と情報構造について（ちょっとながめ）

. . . . . 浅原正幸 (国語研)

16:35-17:00　**■クロージング (2F 講堂)**

目次

発表論文

国語教科書と高校生作文の複文構造比較
－従属節の構造と接続形式の量的比較－

松本 理美（立命館大学大学院）

Comparison about Structures of Complex Sentence between Textbooks of Japanese Language and Composition written by High School Students: Quantitative Comparison about Structures of Subordinate and Connection forms

Satomi Matsumoto (Ritsumeikan University)

要旨

コーパスという言語資源を活用した文体研究は、語彙、品詞、文法などに関するものなど、数多く見られるが、複文構造や従属節の研究への活用が十分であるとは言えない。これは、現時点で、解析器による従属節への情報付与技術の発展に対し、データ分析技術の普及が追いついていないことも起因していると考える。また、複文に着目した文体研究において、高校生作文や学校教科書を対象としたものは、管見の限りない。そこで、本研究では、文章中の従属節に着目し、各種学校の国語教科書と高校生作文における文体特徴を比較することを試み、文章カテゴリーごとに従属節の出現割合を求め、副詞節については、意味別に接続形式を出現頻度でランキングした。従属節の分析からは、国語教科書と高校生作文において、名詞修飾節と副詞節の出現割合に大きな差が見られ、副詞節の意味別接続形式ランキングからも文体特徴を捉えることができた。

1. はじめに

言語に関する研究において、昨今の多種多様なコーパス開発の恩恵に与り、先達が果てしない時間と労力をかけた語や用例等の分析が、大量の言語資源を活用して瞬時に行えるようになってきている。コーパス言語学の発展とともに、研究の対象は文字、語彙から文法まで、書き言葉から話し言葉まで、多種多様な領域に渡り、調査・分析が進められている。

解析器により付与された節情報を利用して連用節の出現を定量的に分析した丸山 (2014) は、コーパスを利用した調査が「母語話者の内省では知りえない言語事実を実証的に明らかに」（丸山 2014：402）できるという利点を挙げている。

一方で、石黒（2016）は、昨今のコーパス言語学の隆盛を歓迎しながらも、敢えて小規模な言語データベースにより、接続詞に着目した文体研究を行った。ここで石黒が指摘する通り、「コーパス・ネイティブ」（石黒 2016：161）の研究者の時代にして、地道な手作業による研究により、「見落とされている言語事実」（石黒 2016：161）が新たに発見されることもまた十分にあり得ると考える。

そこで、本研究では、コーパス開発の現段階では、まだ人手による作業に頼るところもある複文の従属節、特に副詞節の接続形式に着目して、文体特徴を捉えることを試みる。

そして、その対象として、日本語研究ではもとより、言語教育の分野でも谷間的存在であり、研究対象となることが少ない日本語母語、非母語の両高校生の作文を取り上げる。これは、以下の二つの理由からである。一つは、国際化が急進する現代において、日本語学習者としてではなく、生活者として来日を余儀なくされた年少者の日本語が、日本語のバリエーションの一つとして研究対象になり得るのではないかということである。もう一つは、日本人高校生を含め、「不完全な書きことば」から日本語を捉えなおすことで、新たな発見があるのではないかと考えたからである。

2. 研究目的

本研究では、文章中の複文割合や副詞節の接続形式の出現頻度が文章カテゴリーにより異なることを明らかにし、これらにより文体特徴を捉えることができること、またこれらが文章レベルの指標となり得ることを示すことを目的とする。

3. 研究方法

高校生の作文（日本語母語話者、日本語非母語話者[1]）、小学校教科書、中学校教科書、高校教科書の文章について、手作業により、従属節に機能分類と意味分類のタグ付けを行う。

本研究では、益岡・田窪 (1992) に基づき、述語を中心とする文節のひとまとまりを「節」と定義する。ただし、単独で述語の役割と名詞を修飾する役割を持っている形容詞については、補語を伴わずに名詞を修飾している際には節とみなさないものとする。以下に例文を挙げて説明する。

(1) 「赤いマフラーを巻いている少女は、私の妹だ。」

例文(1)の下線部「赤い」は形容詞であり、節ではない。

(2) 「ふさが赤いマフラーを巻いている少女は、私の妹だ。」

例文(2)の下線部「ふさが赤い」は「ふさが」という補語を伴うため、「マフラー」に係る名詞修飾節とする。

従属節については、益岡・田窪 (1992) を元に従属節を詳細に分類した池原 (2009) に従い、意味分類体系[2]1 段目の補足節・名詞修飾節（連体節）・副詞節・並列節の 4 つに分類する。また、副詞節については、意味分類体系 2 段目の 16 種まで分類する。

16 分類した副詞節のうち、本研究で着目するのは、時を表す副詞節、原因・理由を表す副詞節、条件・譲歩を表す副詞節、付帯状況・様態を表す副詞節、逆接を表す副詞節の 5 つの副詞節[3]である。本研究では、これらの節を、順に時間節、原因・理由節、条件・譲歩節、付帯状況・様態節、逆接節と呼ぶことにする[4]。

---

[1] 日本語母語話者の高校生を日本人高校生、日本語非母語話者の高校生を外国人高校生とする。

[2] 池原 (2009) の第 8 章付録「3.主節従属節間の意味分類体系」では、従属節を意味機能により 1 段から 4 段まで順に細かく 4 段階に分類しており、1 段目は 4 種、2 段目は 27 種、3 段目は 37 種、4 段目は 154 種に分類している。例えば、「このワインはおいしいだけに値段が高い。」という文の従属節「このワインはおいしいだけに」は、1 段目は副詞節、2 段目は因果関係、3 段目は原因、4 段目は特定原因に分類される。なお、1 段目は文法的、機能的な分類であり、2 段目以降は、意味や特徴に基づいた詳細分類である。

[3] 本研究では調査対象とした文章データの全てに出現した従属節に着目する。

[4] 本研究の意味分類の基準とした池原 (2009) では、時間節を時、原因・理由節を因果関係としているが、一般的呼称を採用し、このように定義する。

以上の従属節についてタグ付けを行い、文章カテゴリーごとに計量を行う。

対象とした文章は、本研究の調査協力校より回収した高校生作文 43 編（日本人 10 編、外国人 33 編）と、小学校教科書、中学校教科書、高校教科書から各 2 編である。

4. 結果と考察

4.1 従属節について

文章中の従属節を機能による 4 分類にタグ付けし、文章カテゴリーごとに各種従属節の数を計量したものを、表 1 に示す。

表 1 文章カテゴリーごとの各種従属節数 （ ）内は割合(%)

	補足節数	名詞修飾節数	副詞節数	並列節数	従属節の総数
外国人高校生	77 (18)	45 (10)	244 (56)	71 (16)	437
日本人高校生	55 (27)	23 (11)	90 (44)	37 (18)	205
小学校教科書	59 (29)	49 (24)	72 (36)	23 (11)	203
中学校教科書	88 (33)	87 (32)	77 (28)	19 (7)	271
高校教科書	96 (29)	134 (41)	80 (25)	15 (5)	325

文章カテゴリーにより、文の数、従属節の数が異なるため、表 1 に基づき、各種従属節が従属節全体に占める割合を求め、グラフにしたものを図 1 に示す。

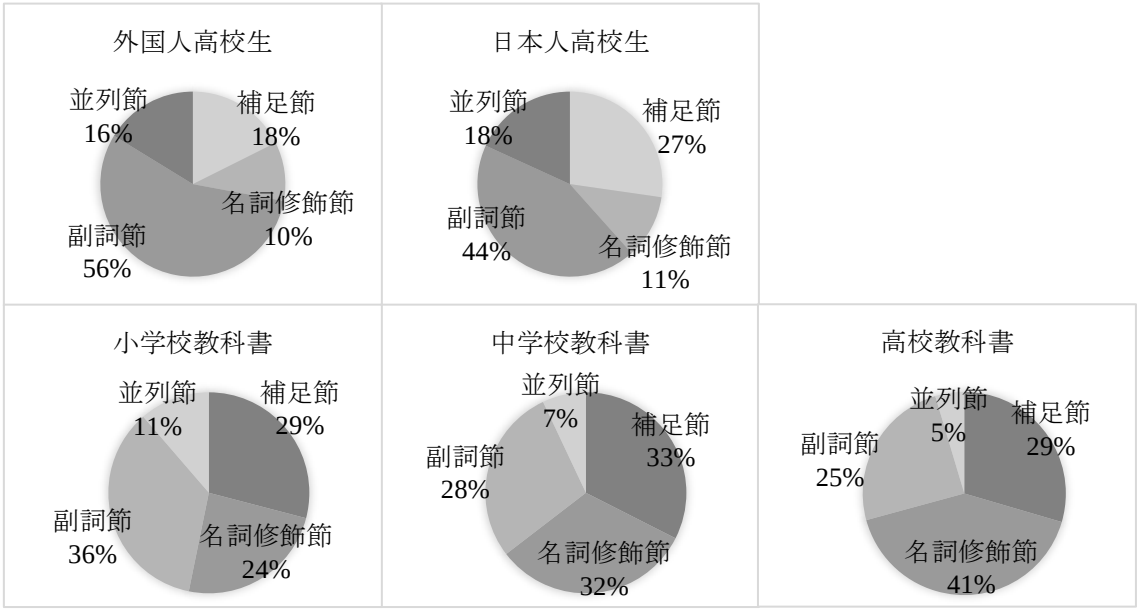


図 1 文章カテゴリーごとの各種従属節の割合

日本人高校生作文と外国人高校生作文を高校生作文群、小学校教科書、中学校教科書、高校教科書を教科書群とすると、高校生作文群と教科書群には、各種従属節の出現割合に大きな違いが認められる。特に大きな差は、補足節、名詞修飾節を併せた連体系の節と副

詞節、並列節を併せた連用系の節の間に、はっきりと現れている。高校生作文群は教科書群に比べ、連体系の節が少なく、連用系の節が多いことが明らかになった。

また、教科書群だけをみると、学年の上昇とともに名詞修飾節が増加し、副詞節、並列節が減少する傾向にあることも確認できた。

これらから、文章における従属節の割合が文章レベルを測る指標となる可能性が示唆できる。

4.2 副詞節のついて

16 種類に意味分類した副詞節のうち、本研究で分析対象とした全ての文章中に出現する副詞節に着目し、文章カテゴリーごとの意味別出現頻度を求めた。着目した副詞節は、時間節、原因理由節、条件・譲歩節、付帯状況・様態節、逆接節である。文章中に出現する頻度は以下のとおりである。

表 2 文章カテゴリーごとの副詞節の意味別出現頻度（ ）内は割合(%)

	時	原因	条件	付帯	逆接	他	合計
外国人高校生作文	54 (22)	50 (20)	12 (5)	26 (11)	38 (16)	64 (26)	244
日本人高校生作文	12 (13)	33 (37)	8 (9)	17 (19)	10 (11)	10 (11)	90
小学校教科書	9 (12)	15 (21)	13 (18)	22 (31)	2 (3)	11 (15)	72
中学校教科書	9 (12)	12 (16)	17 (22)	14 (18)	7 (9)	18 (23)	77
高校教科書	12 (15)	7 (9)	20 (28)	16 (17)	4 (5)	21 (26)	80

表 2 で示した割合を、図 2 のグラフに示す。[5]

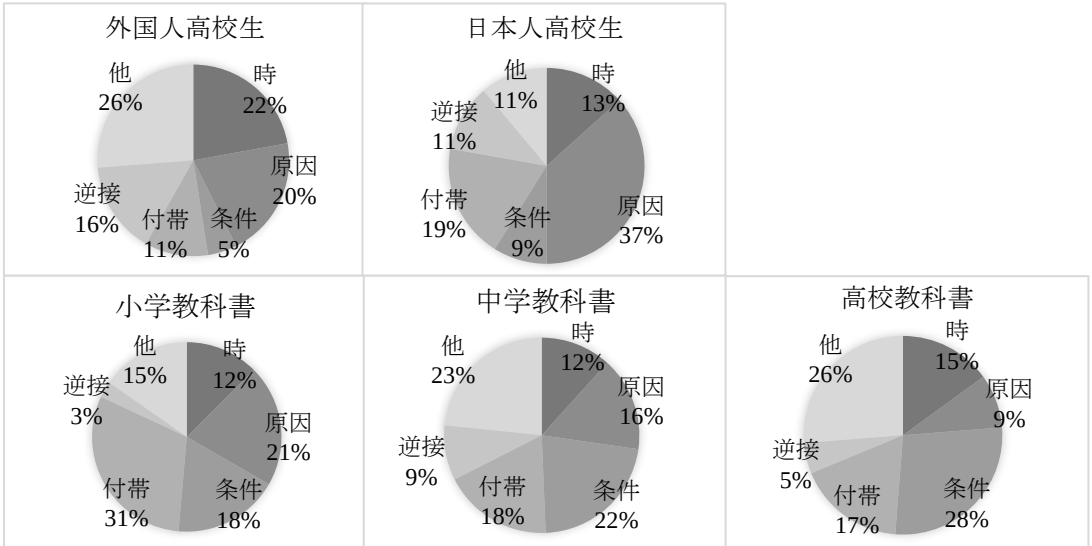


図 2 副詞節の意味別出現割合

[5] 表 2、図 2 では、時間節を「時」、原因・理由節を「原因」、条件・譲歩節を「条件」、付帯状況・様態節を「付帯」、逆接節を「逆接」と略して表記した。

ここで注目すべきは、条件節と逆接節である。高校生作文群において非常に少ない条件節が、教科書群では、学年上昇とともに増加していること、逆接節は、教科書群よりも高校生作文群に多く見られることが確認できた。

教科書群においては、学年上昇とともに、原因理由節、付帯状況・様態節の減少傾向が見られた。

このように、副詞節の意味分類によっても、文章カテゴリーの特徴が捉えられた。

4.3 副詞節の接続形式

本節では、文章カテゴリー別に、副詞節の接続形式について、意味に関係なく、節末の形式にのみ着目した分析を行う。4.3.1 では、石黒 (2016) を援用し、接続形式に見られる文体特徴を、文章カテゴリー間の共通点、相違点から探ることを意図し、文章カテゴリーごとに接続形式のランキングを行い、上位 3 位までを示す[6]。また、4.3.2 では、前項で、全文章カテゴリーにおいて頻度ランキング 1 位となった接続形式「て」に着目する。それぞれの文章カテゴリーにおいて、副詞節の接続形式「て」の意味をどのような割合で分担しているか、文章カテゴリーごとに意味分担割合を求め、その結果について考察する。

4.3.1 文章カテゴリー別接続形式の出現頻度ランキング

本項では文章カテゴリー別に接続形式の出現頻度に着目してランキングを行い、上位 3 位を示す。上位 3 位とした根拠は、データ数が多くないため、出現頻度が少ない接続形式も多く、下位になると出現頻度が「2」や「1」で同順位となる接続形式が多かったためである。なお、カッコ内には、カテゴリー中の副詞節全ての接続形式に占める割合を示す。

表 3 文章カテゴリー別 接続形式出現頻度ランキング （カッコ内は割合）

外国人高校生	
て	60 (32%)
けど	29 (15%)
から	24 (13%)

日本人高校生	
て	32 (40%)
ので	9 (11%)
けど	6 (7%)

小学教科書	
て	13 (21%)
ように	9 (15%)
から	5 (8%)
と	5 (8%)
ながら	5 (8%)

中学教科書	
て	9 (15%)
が	6 (10%)
から	6 (10%)
と	6 (10%)

高校教科書	
て	8 (13%)
ば	7 (11%)
連用中止	6 (10%)

意味に関係なく出現頻度での節続形式ランキングの上位 3 位を表 3 に示した。全ての文章カテゴリーにおいて、接続形式「て」が 1 位を占めているが、2 位との差を見ると、教科書群と高校生作文群では、その出現傾向が明らかに異なっていることがわかる。

[6] 石黒 (2016) を援用し、節形式のランキングを行うことで、文章カテゴリーの文体特徴を捉える試みを行う。石黒 (2016) のように、同じ節形式について、それぞれの文章カテゴリーでの順位を確認したいと考えたが、データ数が少な過ぎたために、今回はランキングだけを行う。

また、教科書類ではほとんど出現しない接続形式「けれど」[7]や、教科書群では学年上昇に伴い出現頻度に減少が見られる接続形式「たら」「から」「ので」[8]が、高校生作文群ではいずれも上位にランキングされていることも確認できた。

この結果から、文章レベルと接続形式に相関がある可能性が示唆された。

4.3.2 副詞節の接続形式「て」について

次に、前項の接続形式の出現割合（頻度）ランキングで、いずれの文章カテゴリーにおいても 1 位であった接続形式「て」に着目して、文章カテゴリーごとに、接続形式「て」が分担する意味について分析を行う。文章カテゴリーごとに、接続形式「て」が分担する意味の割合を求め、考察を行う。

表 3 から、学年上昇に伴い、接続形式「て」の出現が減少していることが読み取れる。図 3 は、外国人高校生作文、日本人高校生作文、小学校教科書、中学校教科書、高校教科書の中で、接続形式「て」が時間節、原因理由節、付帯状況・様態節において、どのような割合でその意味を分担しているかをグラフにしたものである。

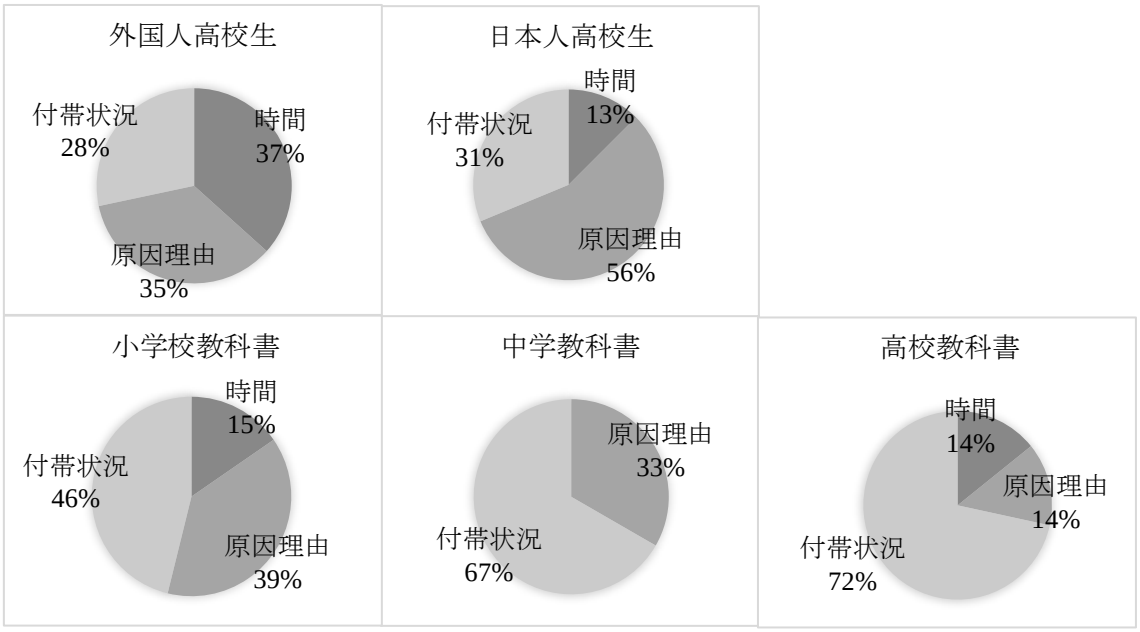


図 3 接続形式「て」が分担している意味

接続形式「て」の出現割合が高かった高校生作文群についてみると、外国人高校生の作文では、時間節、原因理由節、付帯状況節の三つの節において、ほぼ同じ割合で接続形式「て」を使用しているのに対し、日本人高校生の作文では、原因理由節での使用が多く、時間節としてはほとんど使用されていないことがわかる。

[7]「けれども」や、高校生作文群での「けど」「けども」を含む。

[8]「たら」は外国人高校生 4 位、日本人高校生 5 位、小学校教科書 9 位、中学校教科書 10 位、高校教科書出現なし、「から」は外国人高校生 3 位、日本人高校生 5 位、小学校教科書 3 位、中学校教科書 2 位、高校教科書 13 位、「ので」は外国人高校生 8 位、日本人高校生 2 位、小学校教科書 6 位、中学校教科書 7 位、高校教科書 13 位であった。

教科書の接続形式「て」を見ても、時間節での使用が極めて少なく、学年が上がるほどに付帯状況・様態節としての使用に比重が移っていくことが明らかになった。

このように、接続形式「て」が分担している意味が、文章カテゴリーによって異なることが確認され、これによっても、文体特徴がとらえられることが示された。

5. おわりに

本研究では、高校生作文群（外国人、日本人）と教科書群（小学校、中学校、高校）の複文構造や副詞節の意味、接続形式に着目し、それぞれの文体特徴を捉えることを試みた。

高校生作文群は教科書群に比べ、従属節の名詞修飾節の出現割合が少なく、副詞節、並列節の出現割合が多いことが明らかになった。教科書間の比較でも、学年上昇に伴い名詞修飾節が増加し、副詞節、並列節が減少する傾向が確認できた。

また、副詞節の意味別出現割合では、高校生作文群は教科書群と比較し、条件節の出現割合において少なく、逆接節において多いという結果が得られた。

そして、副詞節の接続形式に着目した分析において、高校生作文群で出現頻度ランキング上位の「けれど」「たら」「から」「ので」などは、教科書群での出現が少ないなど、副詞節の接続形式の出現頻度が文章カテゴリーによって異なることも明らかになった。

さらに、意義同形の副詞節接続形式が分担する意味が、高校生作文群と教科書群では異なることが確認された。特に、外国人高校生作文では、接続形式「て」が、時間節、原因理由節、付帯状況節において、ほぼ同じ割合で出現するのに対し、高校教科書では接続形式「て」の 72%が付帯状況を表す副詞節の接続形式であることが確認された。

以上の結果から、複文構造や副詞節の意味、接続形式の出現割合によって、文体特徴を捉えることが可能であり、これらには文章レベルを測る指標となる可能性があることが示唆される。

文構造の量的比較を文章レベルに関連づけた研究としては、柴崎 (2009) [9]や、石崎・伊佐原 (1988)[10] の研究などがあり、バトラー後藤 (2011) は、教科ごとの言語的な特徴を捉えることは、児童生徒への教科指導に直接役に立つばかりでなく、教科書の執筆者にも学年に応じたわかりやすい文章の執筆への配慮を促す際の、具体的な情報として有益であると述べている。

さらに、バトラーは、年少者のリテラシーや文法、とりわけ複文構造の問題については、教科書における言語表現のレベルの移行などを考えることにおいても、今後更なる研究が必要な課題であることを指摘している。

このように、多面的、多角的に文体を捉えることは、日本語学、日本語教育、国語教育など多くの分野の研究に資するものであり、本研究は、研究方法、研究対象ともに、意義のある研究であると考える。

---

[9] 石川他 (2010) は、「柴崎 (2009) は、小学校 1 年生から中学 3 年生までの国語教科書コーパスを用いて、６つの変数を手掛かりに、学年レベルの予測」を行ったことを報告し、その結果から「予測学年氏は実学年に充分近く、信頼性が高いために、小中学生を対象とした図書・国語教材の選択に役立つ」と述べている。

[10] 小学校、中学校、高等学校の教科書について、教科書の複文率が 75%になっていること、学齢が上がるほどに、用言数、埋め込み構造数、並列構造数などが増えていく傾向にあることを示している。

しかし、高校生作文、教科書の文章ともに分析したデータが少なく、筆者の経験、知識の不足からも十分な結果が得られたとはいい難い。引き続き、高校生作文データの集積を行い、教科書を含め、分析データを増進することを今後の課題とし、複文構造から文体を捉える研究を発展させたいと考える。

文　献

池原悟 (2009)『非線形言語モデルによる自然言語処理』 岩波書店

石黒圭 (2016)「社会科学専門文献の接続詞の分野別文体特性」 庵功雄他編『日本語文法研究のフロンティア』pp.161-182. くろしお出版

石崎俊・伊佐原均 (1988)「日本語文の複雑さの定性的・定量的特徴抽出」『自然言語処理』67:6, pp.1-8.

柴崎秀子 (2009)「日本語リーダビリティー測定尺度の構築とソフトウエアの実用化」『科学研究費補助金研究成果報告書』基盤研究(B), 2007~2008, 研究番号 19300277

https://kaken.nii.ac.jp/ja/file/KAKENHI-PROJECT-19300277/19300277seika.pdf（2017 年 12 月 10 日確認)

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

益岡隆志・田窪行則 (1992)『基礎日本語文法』くろしお出版

丸山岳彦 (2014)「現代日本語の連用節とモダリティ形式の分布－BCCWJ に基づく分析－」 益岡隆志他編『日本語複文構文の研究』pp.399-425. ひつじ書房

データとして使用した文献

一川誠 (2015)「時計の時間と心の時間」 『国語 6 創造』 36-41 東京：光村図書出版株式会社

今道友信 (2016)「温かいスープ」 『国語 3』 236-239 東京：光村図書出版株式会社

清水哲郎 (2014)「死と向き合う」 『精選現代文 B』 406-411 東京：筑摩書房

中村桂子 (2015)「生き物はつながりの中に」 『国語 6 創造』 226-229 東京：光村図書出版株式会社

鷲田清一 (2014)「ふわふわ」 『精選現代文 B』 134-140 東京：筑摩書房

鷲田清一 (2016)「誰かの代わりに」 『国語 3』 198-203 東京：光村図書出版株式会社

友人への「断り」に対する評価に関する質的考察 ——日本語母語話者と中国人日本語話者の評価を通して——

滕越（東京大学大学院総合文化研究科）†

A Qualitative Research on the Evaluation for the Refusals to Friends —through the Evaluation of Japanese Native Speakers and Chinese Non-native Speakers of Japanese—

Yue TENG（The University of Tokyo, Graduate School of Arts and Science）

要旨

異文化間の「断り」に関しては,中間言語語用論などの分野で,「言語や社会的規範の違いにより衝突が起きやすい」と論じられることが多い。本研究では,個人差に焦点を当て,評価の視点から研究を進めた。『BTSJ コーパス』から 5 つの「友人の依頼への断り」の音声データを選択し,日本語母語話者 3 名と中国人日本語話者 3 名に,断られる側の視点に立って,5 つの音声の好ましさをプロトコル分析とインタビューを通して評価してもらった。その結果,録音ごとに評価が比較的一致しているものとばらけているものがあり,特に評価のばらつきが大きかった 2 つの録音は,評価者の「友人への断り」における基本的態度が,「合理性・効率性重視」か,「心情・気遣い重視」かで評価が分かれていた。また,今回のデータからは,評価のばらつきと評価者の母語との関連性は見いだせなかった。

1. はじめに

異文化間における「断り」に関する研究は,Beebe et al.(1990)に始まり,中間言語語用論や対照語用論の分野で盛んにおこなわれている。これらの研究の多くは,ロールプレイや談話完成テストを通して得た「断り」の例における意味公式[1]の使用を分析し,2 つの言語間の差あるいは中間言語と目標言語の差を分析している。また,多くの研究には,異文化間の「断り」は,言語や社会文化的規範の違いによって衝突が起きやすいと論じられている。

一方,近年では母語話者が学習者の「断り」を受けてどう感じるかに焦点を当てた評価研究も現れているが,これらの研究は量的調査を通して母語話者の評価の全体的な傾向を求め,「ある種の断りは母語話者に不快感を与える恐れがあるため,学習者はそれを避けた方が望ましい」という結論が提示されている。例えば,マスデン（2011）では,大学生日本語母語話者に対し,学習者が産出した①「今夜は無理/だめ,疲れているんだ」,②「ちょっと用事があって…ごめんね」,③「私は疲れていて眠いよ。残念だけど, 今日はできない。またほかの日に誘って」の「断り」のうち, どれが最も失礼と感じるかを調査しているが,①が最も失礼と答えた母語話者は 54%,②が最も失礼であると答えた母語話者は 44%であるため,学習者は①や②の表現は避けたほうが望ましいと述べていた。

このように,同じ言語の母語話者でも,「断り」に対する評価はばらつきが大きいことがわかる。本研究は,一個人の「断り」への評価には,母語の影響はどの程度あるかという課題のパイロットスタディとして,少数の日本語母語話者と中国人日本語話者を例に,「断り」[2]の評

† yueteng0808*gmail.com（お手数ですが,*の部分をアットマークに変えてください）

[1] 「断り」の意味公式とは,「『謝罪』『言い訳』『代案』など,人がものを断るときに使,うことばを,その意味内容によって分類したものである」（生駒・志村,1993）

[2] 本研究での「断り」は,「口頭での会話において,相手の働きが始まってから,会話が収束するまでの,断り

価のばらつきの様相を調査し,その背後にある「断り」への態度の違いの解明を目指した。

2. 研究方法

本研究では,『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の「4. 女性同士の断りの電話会話」の音声データ（以下,『BTSJ コーパス』の「断り」）を利用して,3 名の日本語母語話者と 3 名の中国人日本語話者（以下,評価者）の評価の様相を調査した。調査の方法は,まず,音声データを評価者に聞かせ,好ましい順に順位付けをしながら,その過程での心的思考をできる限り口に出して録音するという,プロトコル分析[3]の方法を用いた。その後,順位付けの理由やプロトコルの内容,「断り」に対する価値観などについてのインタビューを行った。

2.1 評価対象となる音声データ

『BTSJ コーパス』の「断り」は,データ採集の協力者が,同性の先輩,後輩,同級生に対し,「近日中の国立国語研究所での言語実験の代行」の依頼を実際に行い,依頼から断りまでの一連の自然談話（電話会話）を録音したものである。今回の実験では,同級生の依頼に対する「断り」の録音のうち,タイプの異なる 5 つの録音（A～E）を筆者が選択[4]し,評価者に聞かせた。ここに,A～E の録音の,断る側の主な発話と意味公式[5]，録音の特徴を掲出する。

録音 A（1’30”）
あ, ごめん, 明日授業。（謝罪＋理由）→
うん, ごめん。（謝罪）→
行って調べるの？（働きかけ内容の詳細確認）→
代わりを立てられるの？（同情・関心）→
ほんとにごめんね, なんか。（謝罪）→
大丈夫そう？（同情・関心）
特徴：具体的な理由, 躊躇ない断り, 複数回謝罪, 関心

手の一連の発話。働きかけを聞いているときの反応や,相槌,笑いなどのパラ言語的要素も含む」とする。
[3] 理解や問題解決の過程など本来内的な認知的処理を,それらの処理に伴って起きる言語化など,観察可能な行動から分析する研究方法。「考えていることをできるだけ声に出して説明してください」などの教示によって言語化を誘導し,そこに現れた言葉づかい,表現などを分析する。発話思考法。（中島ら,1999）
[4]本研究と『BTSJ コーパス』の対応は以下に示すとおりである。A : 68-4-JBI03-JSK03, B : 69-4-JBI04-JSK04, C: 70-4-JBI05-JSK05, D : 72-4-JBI07-JSK07, E : 75-4-JBI10-JSK10。
[5] Beebe et al.(1990)の意味公式を基に,筆者が一部改正したものを用いている。

録音 B（2’31”）
空いてはいるけど行きたくない。（直接的な断り・願望/能力の否定）→
25 日でしょ？（働きかけの一部を繰り返す）→
パスポート取らせてくれ。（理由）→
どこなの？（働きかけ内容の詳細確認）→
えー, いやだ。それもいやだ。うんーーー, い, いやだ。
（直接的な断り・願望/能力の否定）→
苦手。（理由）→
えーっと。（フィラー）→
すいません。（謝罪）→
なんか電話すごく不思議なかけかたしたね。（話題転換）
ごめんね。（謝罪）
特徴：あいまいな理由, 明確な断り, 感情豊か, 最後に謝罪

録音 C（2’13”）
あー, そっかーー。（不明確であいまいな返事）→
9 時から 3 時間かーー。（働きかけの一部を繰り返す）→
沈黙→
そっかー, なんか, ちょっと…（言葉を濁す）→
ちょっときついかも。（直接的な断り・願望/能力の否定）→
申し訳ない。ごめんね, なんかね。（謝罪）→
私の知り合いみたいな人で, 聞いてみようか？（代案の提示）→
申し訳ない。（謝罪）
特徴：理由なし, 躊躇しながらの断り, 力み口調, 複数回謝罪, 代案提示

録音 D（2’07”）
うん, そうなんだ。（不明確であいまいな返事）→
あ, あたしにやらないかってこと？（働きかけ内容の詳細確認）→
えー[↑]。（直接的な断り・否定的感情の表出）→
うん, 明日テストあるんだ。（理由）→
ちょっと無理だね。（直接的な断り・願望/能力の否定）→
しかも, 明日あれだよね, あの, グロックじゃん。（理由）→
悪いね。（謝罪）→
特徴：受動的な理由, 否定的感情の表出, 最後に謝罪

録音 E（1’10”）
あ, なんかバイト？（働きかけ内容の詳細確認）→
それはなんかだめかもしれない。（直接的な断り）→
まだテスト終わってないので。（理由）
特徴：躊躇ない断り, 明確な理由, 簡潔で機械的, 謝罪なし

2.2 評価者概要

今回の研究の評価者は,日本語母語話者 3 名と中国語母語話者 3 名[6]で,いずれも 20 代である。評価者の属性の詳細を表 1 に示す。

表 1　評価者属性

日本語母語話者				
評価者	性別	職業	外国語学習歴	海外滞在歴
ア	女性	大学院研究生（言語学） 英語塾講師経験あり 日本語教育修士号取得	英語 10 年 フランス語 2 年	なし
イ	女性	大学非常勤講師（英語,3 年目）	英語 15 年 中国語 3 年	アメリカ 1 年
ウ	男性	大学生（言語学） 英語家庭教師経験あり	英語 10 年 ロシア語 3 年, イタリア語 2 年, ラテン語,サンスクリット語,ドイツ語,フランス語半年, 韓国語 1 年	なし
中国人日本語話者				
評価者	性別	職業	外国語学習歴	日本滞在歴
カ	女性	同時通訳アシスタント	日本語 13 年 英語 10 年	1 年 2 か月
キ	女性	大学院生（言語学） 中国語家庭教師の経験あり	英語 19 年 日本語 13 年 韓国語 3 か月	3 年 2 か月
ク	男性	大学院研究生（国際関係学） 日本語教師の経験あり	英語 13 年 日本語 8 年	1 年 3 か月

今回の調査では,評価対象となる音声が「女性同士」の会話であるため,より依頼する側の

[6] 現在日本滞在中,日本語能力試験 N1 レベル,8 年以上の日本語学習歴を持つ。十分な日本語能力があり,プロトコル分析やインタビューも日本語で行った。

立場に立っての評価が容易と考えられる女性の評価者に多く依頼した[7]。また,実験の前に,研究の目的,実験方法,個人情報保護等について説明を行い,承諾書に署名をいただいた。

3. プロトコル分析における順位付け

表 2　評価者の順位付けの結果

評価者	評価者属性	順位付け結果
ア	日本語母語話者,女性	C>A>D>E>B
イ	日本語母語話者,女性	A>D>B>C>E （インタビュー後,C>B>D>A>E に変更[8]）
ウ	日本語母語話者,男性	B>E>D>A>C
カ	中国人日本語話者,女性	B>A>C>E>D
キ	中国人日本語話者,女性	C>A>D>B>E
ク	中国人日本語話者,男性	C>A>D>E>B

表 2 の順位付けの結果から,直観的に 2 点の考察が得られる。

まずは,評価者の属性と順位付けの結果には,必ずしも明確な関連性はないということである。ア,ク両氏は母語も性別も異なるが,順位付けの結果は同じであった。ウ,ク両氏は性別は同じであるが順位付けの結果は正反対で,ア,ウ両氏は母語は同じであるが順位付けの結果は正反対であった。このことは,主に母語の差に注目して行われている先行研究の不足点を裏付ける結果となった。

また,それぞれの録音への評価は,比較的まとまっているもの（A, D, E）と評価者ごとにばらつきが大きいもの（B, C）があることがわかる。次節では,特に評価のばらつきが顕著であった B と C の 2 つの録音に絞って,「断り」のどのような要素で特に評価のばらつきが大きいかについて考察を進めていきたい。

4. 録音 B と録音 C への評価

4.1　録音 B への評価

録音 B は,順位付けの結果,1 位（最も好ましい）とした評価者が 2 名（ウ,カ）,3 位とした評価者が 1 名（イ）,4 位とした評価者が 1 名（キ）,5 位（最も好ましくない）とした評価者が 2 名（ア,ク）であった。録音 B において,多くの評価者が言及していたポイントについて,その評価の性質を表 3 にまとめた。

[7] ただし, 今回の調査では, 男性の評価者から, 性別の違いが原因で女性の評価者よりも依頼した側の視点からの評価が難しい, という意見は得られなかった。

[8] イ氏は,インタビューの後半で,録音を聞き直し,再考して順位付けを変更している。ただし,イ氏本人が「でも, 最初のほうが正しかったのかも。あとは, あとのほうはちょっと考えすぎちゃってるから」と述べていたため,本稿では主に変更前の順位と評価について分析する。ただし,必要な場合は変更の理由や変更後の順位にも言及する。

表 3 録音 B への評価の性質

「断り」の要素	プラス評価	マイナス評価	言及なし
「空いてるけど行きたくない」,「いやだ」	イ,ウ,カ	ア,キ,ク	
「パスポートとらせてくれ」	イ,ウ,キ	ア,カ,キ[9]	ク
笑い	イ,カ,（ウ）		ア,キ,ク

録音 B において,6 名の評価者の評価を決定づけた要素と言えるのは「空いているけど行きたくない」,「いやだ」といった発話であろう。

この発話に対して,プラスの評価をした評価者は次のように述べている：

(1) まあ印象がいいなというか,正直で。その方が助かります。なんか,相手の断りたい気持ちがちゃんとわからないと困るし,（中略）ちゃんと自分の都合を話してくれるので,なあなあにされない方がお互いにやりやすいかなって。（ウ氏）
(2) 「えー,空いてる,でも行きたくない」って,冗談,何っていえばいいのかな,ふざけてるみたいな感じで,シリアスさを感じない。（中略）結構ポップな感じだったよね。雰囲気がいい。（イ氏）
(3) さっぱり断るのもいいと思いました。「時間あるけど行きたくない」って。なんかこの人,面白くて,かわいくって,素直かなって。（カ氏）

プラスの評価の理由は,B の断り手の人間性の実直さ,そして会話の全体的な雰囲気へのプラスの影響などであった。この 3 名の評価者は,録音 B について,断り手の人間性の暖かさ,両者の関係性の良さ,そして笑いや相槌など,会話全体の雰囲気に関連する言及がほかの 3 名より多く見られた。

一方,「空いているけど行きたくない」,「いやだ」について,マイナスの評価をした評価者は次のように述べている：

(4) それちょっと素直過ぎない,って思うんですよ。たとえ行きたくないから断る場合でも,もう少し,相手に負担をかけないようなって言ったらいいんですか,こっちとしては協力したいんだけど,物理的に無理,時間的に無理っていうような風に断られるのであれば納得できるんですが。（ア氏）
(5) うーん,ちょっとなんか,言い方がはっきりしすぎますね。たぶん普段の性格とかにもよるんですけど,でももう少し,相手の気持ちを,はい,考えてほしい気持ちもあります。（キ氏）
(6) そういわれたら,まあ,親友だから納得できないっていうか,傷つくと思う。（中略）本当は行きたくないんだけど,言い訳をして,相手の気持ちを守るために,ちょっと他の言い訳で,「行けない」っていうとか。自分的にはそっちのほうがいいと思う。（ク氏）

[9] プラス評価,マイナス評価の両方に名前がある評価者は,その項目に対し,プラス面,マイナス面の両方の評価をしたことを指す。（ ）で名前を囲った評価者は,明確にその項目に言及していたわけではないが,関連の事項について言及していたことを指す。

マイナス評価の理由は,3 名とも比較的一致しており,感情の表出がはっきりしすぎており,断り手の依頼する側への配慮が不足しているということにあった。

また,録音 B では,「行きたくない」と断ってから,「パスポート取らせてくれ」という理由の後付けを行っている。この発話について,イ,ウ両氏は,用事があるということがよく分かったというように,単に理由を述べている風にとらえていたが,ア,カ両氏は,「正当な理由があるなら先に言って（ア氏)」,「ちょっと理由つけている感じ（カ氏)」とやや否定的にとらえていた。また,キ氏は,

(7)「取らせてくれ」の言い方がちょっと,自分的には好きじゃない。でも,たぶんこの二人の関係はすごくいい,からかもしれないですね。（キ氏）

と述べており,全体的に否定的な評価をしながらも,二名の発話者の関係性の良さは認めていた。

4.2 録音 C への評価

録音 C は,順位付けの結果,1 位（最も好ましい）とした評価者が 3 名（ア,キ,ク）,3 位とした評価者が 1 名（カ）,4 位とした評価者が 1 名（イ）,5 位（最も好ましくない）とした評価者が 1 名（ウ）であった。録音 C において,多くの評価者が言及していたポイントについて,その評価の性質を表 4 にまとめた。

表 4 録音 C への評価の性質

断りの要素	プラス評価	マイナス評価	言及なし
「私の知り合いみたいな人で,聞いてみようか？」	ア,イ,ウ,カ,キ,ク	イ,ウ,カ	
沈黙,力み口調[10],「うーん」	カ,キ,ク	イ,ウ,カ	ア
謝罪	ア,キ,ク	イ	ウ,カ

録音 C は録音 B とは対照的で,依頼を受けてからの沈黙の時間が長く,言いよどみや力みの口調を用いるなどの特徴があり,全体的な評価に大きく影響していた。

この特徴について,プラスの評価をした評価者は次のように述べており,ここに誠実さや依頼側への配慮を感じ取っていた。

(8) 結構悩んでる感じがしました。（中略）間に,10 秒くらいの空白の時間があって,本当に,行くかどうか結構この人は悩んでいる人だなって思いましたね。たぶん,本当に行っ

[10] 定延（2005）で,りきみ口調とは,重い荷物を運ぼうとしてなかなか持ち上がらなかったときに,つめていた息が漏れ出して出る「んんん,んん,んん」のような声,または足の小指を思いきりたんすの角にぶつけてしまったとき,涙をにじませ,小刻みに体を震わせながら漏れる,「ん,んんん,んん」のような声と描写している。また,話し手がりきむ局面として,苦しみ,関心,「いわゆる強調」の三つを挙げており,「苦しみ」の局面は心理的な苦痛による場合もあるとしており,「相手の前で恐縮してみせる」ときにもりきみ口調は良く用いられるとされている。録音 C のりきみ口調はこの類であろうと予想される。

てあげたいからこそ,そんな風に悩んでいるんだと思います。(キ氏)

(9) 真剣に考えている。行くか行かないか考えてくれて,そっちのほうが,行きたくないんだけど,そのまま断ったらこっちの気持ちを傷つけるっていうのが,それがよくないって考えてくれているみたいで,大事にしてくれているような感じがします。(ク氏)

一方,マイナスの評価には次のようなものがあった:

(10) 「うーん,んん」って。その,考えるんだったら何を考えているのか教えてほしい。(中略)何も言ってくれないのは,なんかこう,隠しているように感じるのかな。(イ氏)

(11) 行きたくなさそうなのはすごくわかるから,さっさと断ってくれても仕方ないかな。むしろそうしてくれないと,こっちのほうがちょっと悪いことしている感じがする。(ウ氏)

沈黙に対してマイナスの評価をした評価者は,何を考えているのかわからない,または断りたそうなのに断らないために,ストレスを感じるようであった。プラスマイナスの両面に対して言及のあったカ氏は,「ちゃんと考えてくれた」が「考える時間ちょっと長すぎ」と両面の評価をしていた。

また,録音 C では,自分の知り合いに聞いてみるという代案の提示を行っていた。6 名の評価者すべてが,この代案提示に対してプラスの評価をしていたが,イ,ウ,カ三氏は「自分は悪くない状況にしたいから」,「聞きたくないのに言っているようでずるいかも」などのマイナス評価も同時にあった。

さらに,録音 C の複数回の謝罪については,ア,キ,ク三氏は「申し訳ない気持ちがすごく伝わる」と高く評価していたが,イ氏は,最後の謝罪の声の調子が淡々としていて,途中の力み口調とのギャップに誠意を感じないと述べていた。

5. 「直接性」への評価からみる「断り観」

録音 B と録音 C への評価を分析すると,「断りの直接性」という要素への評価の背後に,6 名の評価者の「断り」に対する内的価値観(以下,「断り観」)の違いが垣間見える。本節では,どのような「断り観」が「直接性」の評価に影響していたのかを見ていきたい。

イ,ウ,カ三氏は,録音 B の「いやだ」や「行きたくない」に対して,「素直でわかりやすい」とプラスの評価をしており,録音 C の沈黙や言いよどみに対して「何を考えているのかわからず,ストレス」とマイナスの評価をしていた。この三名の評価者は,友人に対して断るときは,腹を割って率直に断るのを良しとしており,「断り」において効率性や合理性を重視していると考えられる。

ア,キ,ク三氏は,録音 B に対しては,「はっきりしすぎて傷つく」とマイナスの評価をしており,録音 C に対しては「自分のために悩んでくれている」とプラスの評価をしていた。この三名の評価者は,友人に対して断るときは相手への配慮が大切と考えており,「断り」において心情や気遣いを重視していると考えられる。

以上について,図 1 にまとめる。

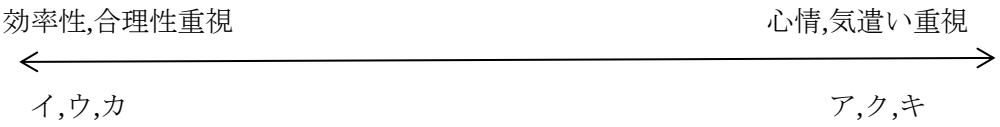


図 1　6 名の評価者の「断り観」

この図からみると,「直接性」に関する「断り観」は評価者の母語とは直接的な関連性は見受けられず,必ずしも我々が考えるような,「日本人は心情重視で,中国人は直接的」というステレオタイプには一致していなかった。

但し,「効率性重視か,気遣い重視か」は,「断り観」を構成する要素の内の一つであり,この要素だけで今回の順位付けの結果を解釈できるわけではない。また,自分が断る側に立ったときに用いる断り方と,断られる側に立ったときに好ましいと思う断り方が一致していない評価者もおり,今後さらに深い考察が必要と言えよう。

6. 結論と今後の課題

本研究では,「断り」に対する評価に焦点を当て,『BTSJ コーパス』の 5 つの断りの音声を,日本語母語話者と中国人日本語話者に聞かせ,プロトコル分析とインタビューを通して,一部の音声に対する評価の様相と,その背後にある「断り観」の一端を見出すことができた。

今回のデータに対する分析では,特に「感情を表出した直接的な断り」と「力み口調や沈黙などのパラ言語的要素」への評価に大きなばらつきがみられ,その背後には「断りにおいて重視するのは効率性や合理性か,それとも心情や気遣いか」という「断り観」が作用していることが示唆された。また,今回収集したデータが,日本語母語話者や中国人日本語話者を代表しているという保証はないが,プロトコル分析の順位付けの結果も,インタビューの分析からも,評価者の母語の違いが,「断り」への評価に影響するという根拠は示されなかった。

今回の調査では, 主に「効率性/合理性」⇔「心情/気遣い」という視点から「断り観」の違いの一端を明らかにしたが, 「断り観」には, このほかにも, 様々な構成要素があると考えられる。今後は, より大規模な量的調査を実施し, 統計的な処理を施すなどの方法を通して, 「断り観」の構成要素を解明し, 「断り」の母語差と個人差の関係についてより深く考察していきたい。

謝　辞

本論文の作成に当たって,指導教員の東京大学総合文化研究科宇佐美洋先生には,実験の方法,データ処理,論文の書き方に至るまで,様々な面から多くのアドバイスをいただいた。また,チューターとしてお世話になっている同研究科の王牧さん,ディスカッションで火花を散らし合っている宇佐美ゼミの皆さま,そして評価者として本研究に参加してくださった 6 名の評価者の皆さまからも,たくさんのアイディアをいただいた。あらためて感謝申し上げたい。

参考文献

生駒知子・志村明彦（1993）「英語から日本語へのプラグマティック・トランスファー ;『断り』という発話行為について」『日本語教育』79, pp.41-52.

宇佐美まゆみ (2007)「改訂版 : 基本的な文字化の原則 (Basic Transcription System for Japanese: BTSJ）2007 年 3 月 31 日改訂版」『談話研究と日本語教育の有機的統合のための基礎的研

究とマルチメディア教材の試作』平成 15-18 年度 科学研究費補助金 基盤研究 B(2)（研究代表者　宇佐美まゆみ）研究成果報告書
宇佐美洋（2014）『「非母語話者の日本語」はどのように評価されているか――評価プロセスの多様性をとらえることの意義』ココ出版
王源・山本裕子（2015）「親しい友人に対する断り行動の日中対照研究」『人文学部研究論集』34, pp.19-35.
近藤佐智子（2009）「中間言語語用論と英語教育」『上智短期大学紀要』29, pp.73-89.
定延利之（2005）『ささやく恋人、りきむレポーター』岩波書店
施信余（2005）「依頼に対する『断り』の言語行動について――日本人と台湾人の大学生の比較」『早稲田大学日本語教育研究』6, pp.45-61.
中島義明ら（1999）『心理学辞典』有斐閣
古村由美子（2011）『成人バイリンガルの「断り」場面における対人葛藤対処方法に関する研究――英語母語話者は日本語英語話者の対処方法をどう評価するのか――』花書院
マスデン眞理子（2011）「日本人大学生が失礼だと感じる留学生の誘い・断りの表現に関する予備調査」『熊本大学国際化推進センター紀要』2, pp.51-73.
李海燕（2013）「『断り』表現の日中対照研究」東北大学博士論文
Leslie M. Beebe. Tomoko Takahashi and Robin Uliss-Weltz. (1990). “Pragmatic Transfer in ESL Refusals”, In R.C.Scarcella, E.Anderson & S.C. Krashen (eds.),*Developing Communicative Competence in a Second language*. Boston, Heinle & Heinle Publishers, pp.55-73.
Shoshana Blum-Kulka and Elite Olshtain.（1984）. “Requests and Apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP)”. *Applied Linguistics*, 5:3, pp.196-213
Gabriele Kasper. and Shoshana Blum-Kulka.(eds.)(1993). *Interlanguage Pragmatics*. New York: Oxford University Press.

もし小学生が『現代日本語書き言葉均衡コーパス』並みに漢字を使ったら

今田 水穂 (文部科学省)

What If Elementary School Students Use the Chinese Characters As Much As BCCWJ?

Mizuho Imada (Ministry of Education, Culture, Sports, Science and Technology)

要旨

『児童・生徒作文コーパス』と『現代日本語書き言葉均衡コーパス』(BCCWJ) を用いて、児童が BCCWJ と同等の水準で漢字を使用した場合に、各漢字の頻度がどの程度になるかを推定し、その結果をワードクラウドを用いて可視化した。また、その結果を用いて、学年ごとの推定頻度の比較、BCCWJ における漢字頻度との比較、教科書コーパスについて同様に漢字頻度を推定したものとの比較を行い、推定頻度と学年の相関、児童作文に固有の高頻度漢字、小学校配当外の高頻度漢字、小学校配当の低頻度漢字を調べた。

1. はじめに

児童の使用する語彙は、大人の使用する語彙とは異なる。そこで、児童の書いた作文を調査することで、児童の書き言葉の産出における漢字の需要を評価することを考える。しかし児童は基本的に学習済みの漢字しか使わず、特に低学年の場合はほとんど仮名だけで作文を書くので、単に語彙を調べただけでは、潜在的な漢字の需要を評価することができない。そこで、それらの語が大人と同等の頻度で漢字書きされた場合に、そこに含まれる漢字がどの程度の頻度になるかを試算する。

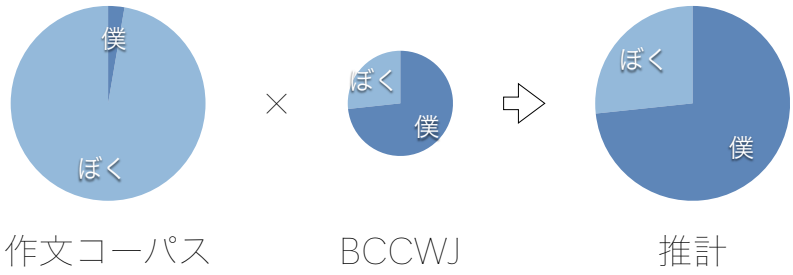


図 1　もし小学生が BCCWJ 並みに漢字を使ったら

この推計結果を用いて、児童の言語産出における漢字需要を可視化し (3 節)、学齢による漢字需要の推移 (4.1 節)、児童作文に固有の高頻度漢字 (4.2 節)、小学校配当外の高頻度漢字および配当内の低頻度漢字 (4.3 節) について考察する。

2. 方法

『児童・生徒作文コーパス』[1](作文コーパス) と、『現代日本語書き言葉均衡コーパス』[2](BCCWJ) の 2 つのコーパスを使用する。作文コーパスは小学校・中学校の児童・生徒に特定のテーマの作文課題を課し、その作文を収集・電子化したコーパスである。本調査では、2014 年度に実施した「夢」「頑張ったこと」の 2 つの作文課題について、それぞれ小学校 1〜6 年生の各 2 クラス、延べ 24 クラス分の作文資料に対して人手修正済みの形態論情報を付与したデータ[3]を使用した。

表 1　作文コーパスのサンプル数と短単位数

学年	夢		頑張ったこと		合計	
	サンプル数	短単位数	サンプル数	短単位数	サンプル数	短単位数
1	69	7196	69	10745	138	17941
2	65	11045	68	14108	133	25153
3	69	17741	69	18635	138	36376
4	78	26038	79	27481	157	53519
5	77	25265	77	29924	154	55189
6	78	26779	78	26200	156	52979
合計	436	114064	440	127093	876	241157

BCCWJ は国立国語研究所が開発した 1 億語規模の書き言葉コーパスで、13 レジスタ約 17 万サンプルの書き言葉資料によって構成される。このうち 6 レジスタ 1980 サンプル約 9 万短単位のデータがコアデータとして設定されており、この範囲のデータ全体について形態論情報の人手修正が施されている。本研究では BCCWJ コアデータのうち、他のレジスタと比べて漢字使用頻度が高い[4]新聞・白書を除いた 4 レジスタ (書籍・雑誌・ブログ・知恵袋) のデータを使用した。以下、単にコアデータというときは、この 4 レジスタを指す。

表 2　BCCWJ コアデータのサンプル数と短単位数

レジスタ	サンプル数	短単位数
書籍	83	234794
雑誌	86	241179
ブログ	471	117888
知恵袋	938	110645
合計	1578	704506

[1] 宮城・今田 (2015a)
[2] Maekawa et al. (2014)
[3] 今田水穂 (2017)
[4] 宮城・今田 (2015b)

調査は以下の手順で行った。まず、BCCWJ を使用して語別の漢字頻度表を作成した。次に作文データを使用して児童の学年別の語彙頻度表を作成した。この 2 つの数値を掛け合わせることによって、児童が BCCWJ 並みの頻度で漢字を使用した場合の漢字頻度表を作成した。この数値を以下では推定漢字頻度と呼ぶことにする。文書 a の漢字使用頻度が文書 b 並みになった時の文字 c の推定頻度を $e_{c,a,b}$ とすると、$e_{c,a,b}$ は次の式で計算できる。

$$e_{c,a,b} = \sum_{w} \frac{f_{w,a} \times g_{c,w,b}}{f_{w,b}}$$

$f_{x,y}$ は文書 y における語 x の頻度、$g_{x,y,z}$ は文書 z、語 y における文字 x の頻度である。文書 a の漢字使用頻度が文書 b 並みになった時の文字 c の 100 万字あたりの推定頻度を $\mathrm{ppm}(e_{c,a,b})$ とすると、次の式で計算できる。

$$\mathrm{ppm}(e_{c,a,b}) = \frac{10^6 \times e_{c,a,b}}{\sum_x e_{x,a,b}}$$

3. 結果

学年別の 100 万字あたり推定漢字頻度を、漢字の配当学年ごとに集計した結果を以下に示す。

表 3　100 万字あたりの推定漢字頻度

						学年
漢字分類	1 年生	2 年生	3 年生	4 年生	5 年生	6 年生
1 年配当漢字	47640	47322	47611	44966	46965	47311
2 年配当漢字	54573	57076	55670	58573	62459	63279
3 年配当漢字	38851	38364	38959	39734	42949	44269
4 年配当漢字	20867	17879	19931	22090	21998	23609
5 年配当漢字	18623	11463	11746	12535	14474	16431
6 年配当漢字	13262	13635	12116	12098	10397	13102
配当外常用漢字	22749	18668	19216	18355	18406	18103
常用外漢字	2363	1874	1687	1530	1417	1189
合計	218926	206281	206934	209882	219065	227293

全体としては 100 万字あたり 20〜23 万字が漢字であり、1 年生は例外的に漢字頻度が高いが、2〜6 年生については学年が上がるにつれて漸進的に漢字の頻度が上がることが確認できる。漢字頻度を BCCWJ 並みに調整してもこのような学年差が見られるのは、品詞や語種など語彙構成の変化を反映しているものと考える。なお、BCCWJ コアデータの漢字頻度は 100 万字あたり約 27 万字である。

個別の漢字の頻度を、ワードクラウドによって可視化したグラフを図 2 に示す。学年は低学年、中学年、高学年の 3 段階にわけ、頻度は各段階の平均を求めた。文字サイズは、頻度の平方根に比例する (従って、文字の面積と頻度が比例する)。

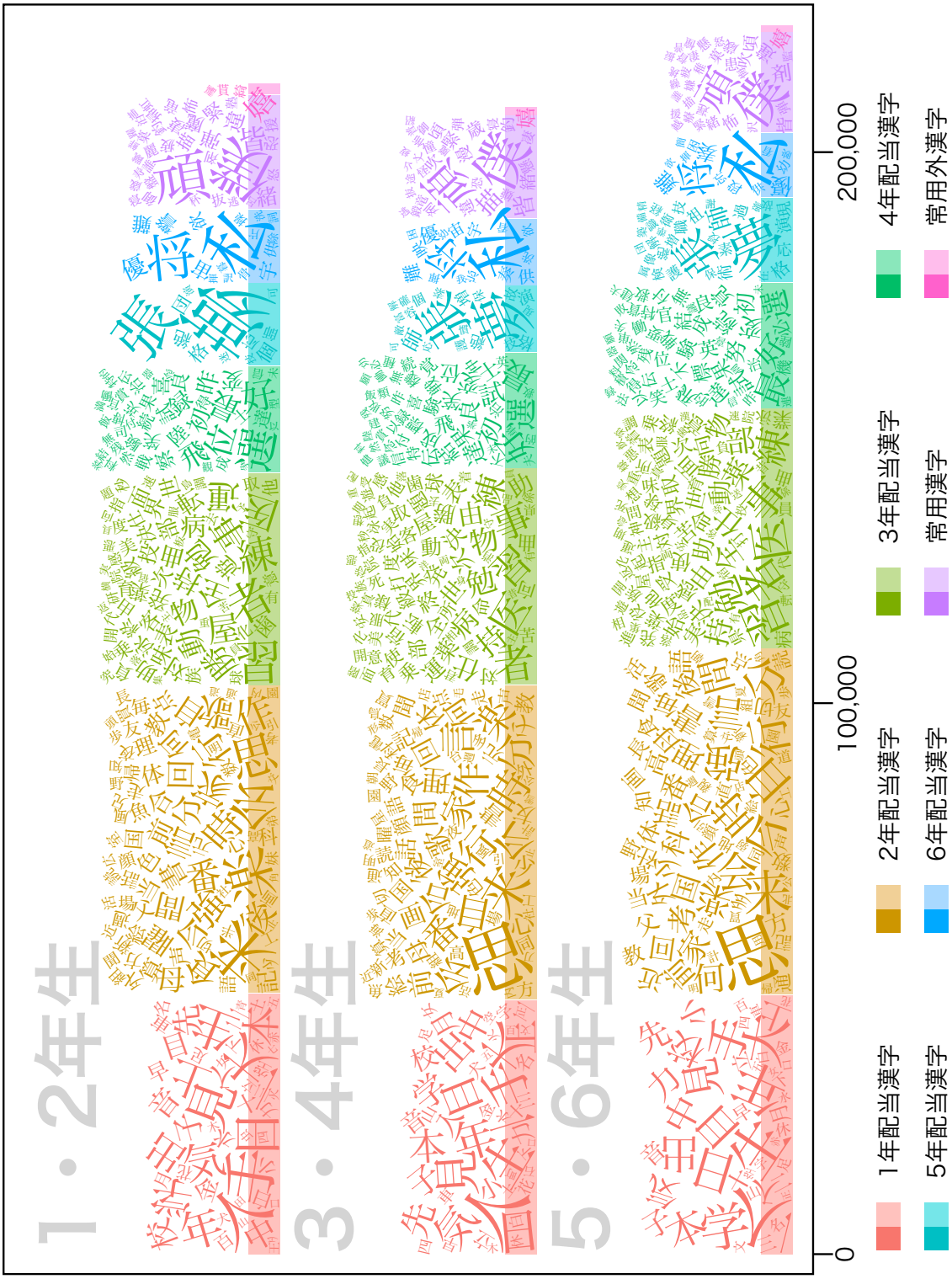

1・2年生
3・4年生
5・6年生
0
100,000
200,000
1年配当漢字
2年配当漢字
3年配当漢字
4年配当漢字
5年配当漢字
6年配当漢字
常用漢字
常用外漢字

図2　ワードクラウド

4. 考察

4.1 学年による差異

学年と推定漢字頻度の関係を調べるために、個々の漢字についてサンプルごとの推定漢字頻度を計算し、作文テーマ別に学年との相関係数を調べた。相関係数が正の値であれば学年が上がるにつれて漢字の使用頻度が上昇し、負の値であれば下降すると考えられる。図 3 は、横軸を推定漢字頻度 (全サンプル平均)、縦軸を相関係数として、各漢字を散布図で可視化したものである。頻度が 500 以上、相関係数の絶対値が 0.1 以上の漢字のみ表示する。

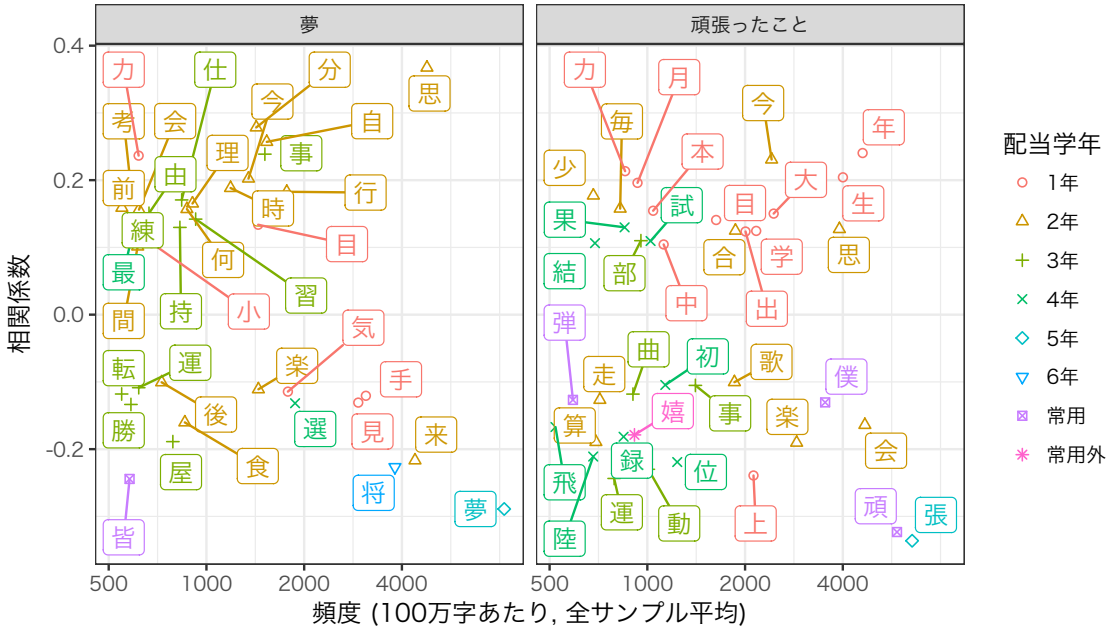


図 3　作文コーパスにおける推定漢字頻度および学年との相関係数

作文テーマによって違うが、「思、分、自、年、事、力、今、生、考」などの漢字について 0.2 以上の弱い正の相関が認められる。また、「張、頑、夢、皆、運、上、動、将、位、来」などの漢字について −0.2 以下の弱い負の相関が認められる。この結果は「思う」「考える」「自分」などの抽象的かつ一般的な語彙が学年が上がるにつれて増加するのに対して、「夢」「将来」「頑張る」「運動」など作文テーマと関連する特徴語が相対的に減少することを示唆する。減少の理由として、児童の使用語彙の変化や、1 サンプルあたりの語数の増加 (使用頻度の変化が小さい語は、相対的に単位語数あたり頻度が減少する) などが考えられる。

4.2 児童作文固有の高頻度漢字

児童作文に固有の高頻度漢字を確認するために、BCCWJ コアデータにおける 100 万語あたり漢字頻度との比較を行った。作文における推定頻度を x、コアデータにおける頻度を y とし、座標 (x, y) の原点からの距離 $\sqrt{x^2 + y^2}$ を d、x 軸からの角度 $\arctan(y/x) \times 2/\pi$ を a と

して、d を横軸、a を縦軸にプロットしたものを図 4 に示す[5]。角度が $1(=90°)$ に近いほどコアデータにおける頻度が、$0(=0°)$ に近いほど作文における頻度が高く、0.5 では同数である。

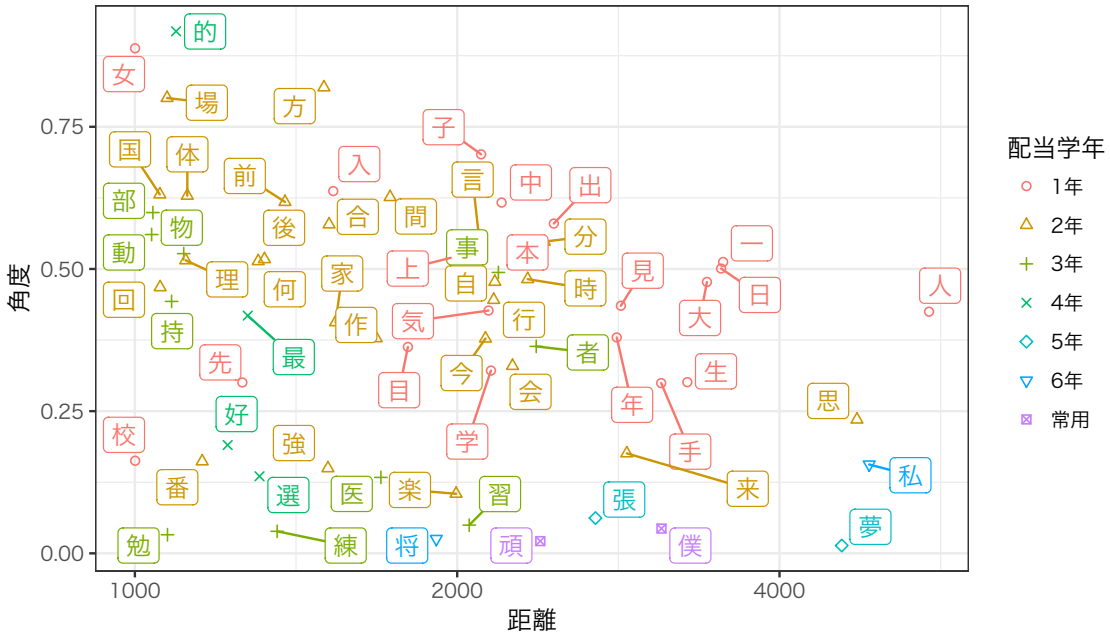


図 4　作文コーパスと BCCWJ コアデータの漢字頻度

図 4 の下方にある漢字が作文コーパスに固有の高頻度漢字と考えられるが、個々の漢字がどのような要因で児童作文において高頻度で生起するかについては、それぞれ検討を要する。「夢」「頑」「張」などは、作文テーマに固有の高頻度漢字と考えられる。「将」「来」や、医者の「医」、選手の「選」なども、作文テーマに関連した高頻度漢字である可能性がある。「私」「僕」などの 1 人称代名詞や、「思」「楽」「好」などの思考・感情語彙に含まれる漢字は、作文テーマというより生活作文などの文種に固有の高頻度漢字である可能性がある。また、「私」より「僕」の方が図の下方に位置しているのは、著者の属性 (小学生であること) の影響による可能性がある。「学」「校」「勉」「強」「練」「習」なども著者の属性に固有の高頻度漢字であろうが、このうち「勉」「強」「練」「習」などは作文テーマの影響を受けている可能性もある。これらの要因について検証するための十分な対照資料が無いため、ここでは可能性を示唆するのみに留める。

4.3　漢字の配当学年と頻度

作文は児童の言語活動の 1 つのレジスタに過ぎず、児童の漢字需要を評価するためには他のレジスタも合わせて検討する必要がある。現状、児童の言語活動を広範に調査できる均衡コーパスは存在しないが、ここでは作文コーパスの他に BCCWJ 教科書サブコーパスを使用することにする。このコーパスは BCCWJ の非コアデータに含まれるサブコーパスで、小学校か

[5] これは x を横軸、y を縦軸とする散布図について、原点を中心とする弧と両軸に囲まれた扇型の範囲を方形に変換したものに相当する。

ら高校までの検定教科書から 412 サンプル、約 93 万形態素のデータが収録されている。ここでは、小、中学校の 161 サンプル、約 36 万形態素のみを比較対象とする (以下、この範囲のコーパスを教科書コーパスと呼ぶ)。高校教科書の漢字については、中学校までに学習する漢字で対応可能であり、必ずしも小学校段階で学習する必要がないため比較対象から除外した。

小、中学校の教科書では、未履修の漢字は学習上の配慮から仮名書きに開いて表記することが多い。そのため、教科書コーパスについても作文コーパスと同様の方法で BCCWJ 並みの漢字頻度にした場合の 100 万字あたりの推定漢字頻度を計算した。作文コーパスと教科書コーパスにおける各漢字の推定頻度を用いて、小学校配当外であるが高頻度の漢字、および小学校配当であるが低頻度の漢字を調べる。

まず、高頻度の小学校配当漢字を確認する。図 5 は、作文コーパスと教科書コーパスに含まれる配当外漢字について、$d \geq 200$ のものを図 4 と同様の方法により距離と角度で表現したものである。角度が 1 に近いほど、教科書コーパスにおける頻度が高い。

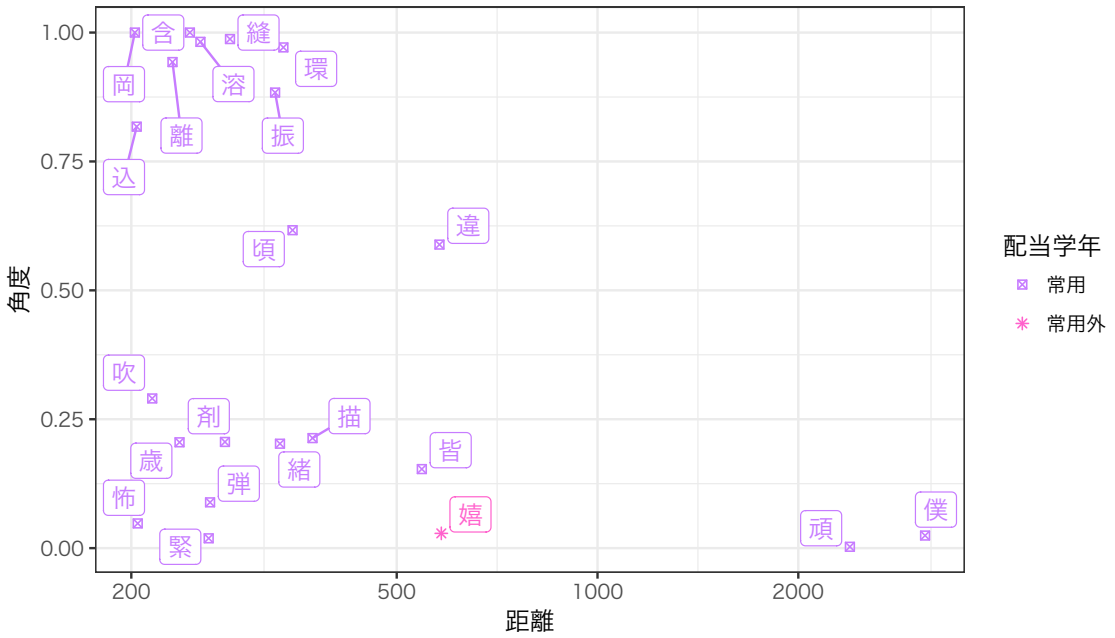


図 5　作文と教科書における高頻度の配当外漢字

$d \geq 1000$ の範囲にほとんど漢字がないことから、漢字頻度を BCCWJ 並みに調整しても、小学生の作文や小〜中学校の教科書に高頻度で生起しうる配当外漢字は少ないことが分かる。非常に頻度が高い漢字としては「僕」「頑」があるが、作文コーパスのみ高頻度で、教科書コーパスでは低頻度である。「頑」は作文テーマの影響で頻度が高くなっているものと考えるが、「僕」は本調査資料に限らず児童の書き言葉では多用される可能性があり、小学 6 年配当の「私」と合わせて学習時期を検討する余地のある漢字と言える。また、やや頻度は下がるが、作文コーパス、教科書コーパスの両方で頻度が高い「違」「頃」などについても、小学校で学習したとしても不自然ではないと考える。

次に、低頻度の小学校配当漢字を確認する。図 6 は、作文コーパスと教科書コーパスに含ま

れる小学校配当漢字について、$d < 20$ のものを表示したものである。図には含まれていないが、作文コーパス、教科書コーパスのいずれも頻度 0 だった配当漢字として、小学 5 年配当の「俵」、小学 6 年配当の「絹」「蚕」がある。

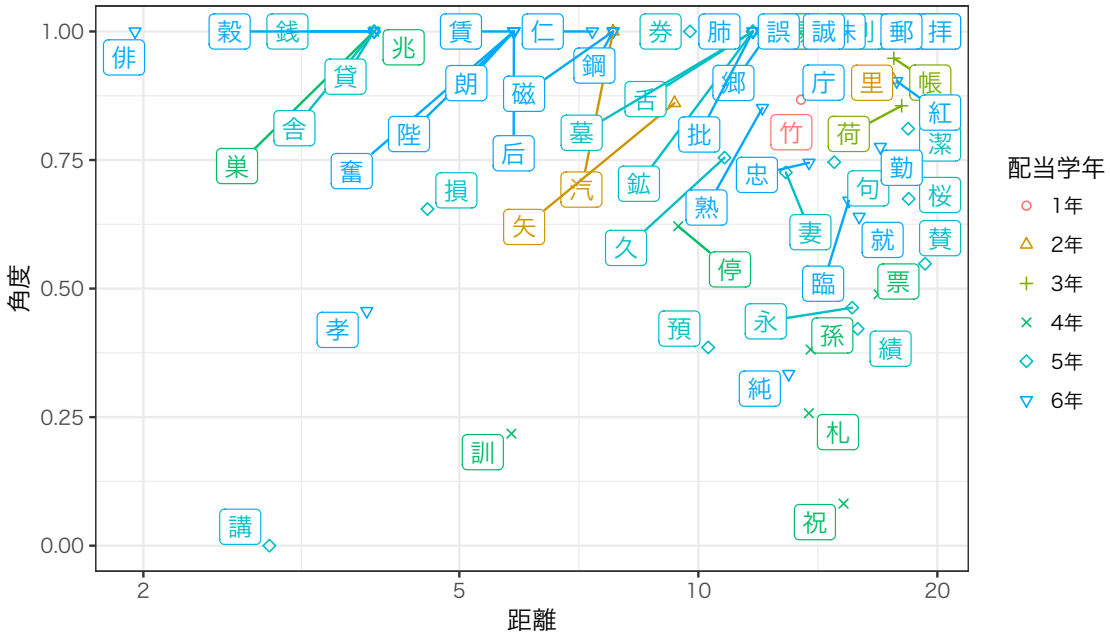


図 6　作文と教科書における低頻度の配当漢字

図中の漢字の頻度は 100 万字中 20 字以下であり、非常に頻度の低い漢字と言うことができるが、教育漢字の需要は必ずしも使用頻度のみで評価できるものではない。例えば俳句の「俳」「句」、音訓の「訓」、熟語の「熟」、批評の「批」などは国語の学習において必要になる漢字であり、頻度が低いからといって重要度が低いとは断定できない。また、小学 1 年配当の「竹」、小学 2 年配当の「矢」「里」なども頻度は低いが他の漢字の構成要素となる字であり、早い段階で教えることは一定の合理性がある。一方で、小学 2 年配当の「汽」や、前述の「俵」「絹」「蚕」などのように、必ずしもこの段階で学習する必要があるかどうか、検討の余地のある漢字も見られる。

5.　まとめ

『児童・生徒作文コーパス』と『現代日本語書き言葉均衡コーパス』の 2 つの言語資源を利用して、児童が大人と同等の使用頻度で漢字を使用した場合の推定漢字頻度を試算し、その結果の可視化と、学年差、レジスタ差、漢字の配当学年と推定頻度の関係などについて検討した。本研究で得られた知見を以下に列挙する。

- 学年差について、児童作文における推定漢字頻度は 100 万字あたり 20〜23 万字ほどで、BCCWJ コアデータにおける 27 万字よりも少なく、学年が上がるにつれて増加する傾向がある。個別の漢字を見ると、「思」「考」など学年が上がると推定頻度が増加する漢字がある一方で、「夢」「頑」「張」など作文テーマに直結する漢字は相対的に推定頻度が

低下する。

- レジスタ差について、BCCWJ と比べて児童作文に固有の高頻度漢字の中には、作文テーマ、文種、著者の属性など様々な要因の影響を受けていると考えられるものが混在している。
- 配当学年と推定頻度の関係について、「僕」「違」「頃」など配当外漢字の中にも作文や教科書において高頻度で使われうる漢字がある一方で、「汽」「俵」「絹」「蚕」など配当漢字の中にも非常に頻度の低い漢字がある。

BCCWJ を利用して教育漢字や常用漢字の分析をした研究としては、これまで棚橋 (2013)、丹保 (2014, 2016)、河内 (2015) などがある。特に丹保 (2014, 2016) は BCCWJ における高頻度漢字、低頻度漢字について配当表漢字としての妥当性を検討しており、本研究と目的、方法の重なる点が多い。

先行研究に対する本研究の新規性は、児童作文という児童の産出言語を資料として使用したこと、またその分析手法を提案したことである。資料について、児童作文は既存の他の資料にはない特徴を持つ。例えば丹保 (2016) が BCCWJ における高頻度漢字として挙げている「彼」は、本研究で使用した BCCWJ コアデータにおいても 100 万字あたり 682 字ほどで配当外漢字としては最も頻度が高いが、作文コーパスでは 20 字、教科書コーパスでは 75 字ほどと低頻度である。作文や教科書以外のレジスタも調べる必要があるが、単に大人の文章で頻出するというだけであれば小学校までに学習する必然性はなく、中学校までに学習する常用漢字に含まれていれば十分である。一方、「僕」はコアデータでは 212 字、教科書では 119 字ほどの頻度だが、作文コーパスでは 3096 字と突出して高い。大人の文章や学習教材だけを調査対象としてしまうと、このような児童の生活に固有の漢字需要を見落とす恐れがある。

また分析手法について、児童作文を対象とした漢字需要調査は、児童の漢字使用状況が既存の教育カリキュラムの影響を受ける (未履修の漢字は生起しない) という難しさがある。習得後はほぼ漢字表記されるような漢語や専門語彙であれば、語彙を調べることで漢字の需要もほぼ特定することができるが、例えば「あいつ」などの語は大人の文章でも「彼奴」と書くことは稀であり、単に全ての語彙を漢字表記に置き換えることで漢字の需要を数値化することはできない。この問題に対して、本発表は BCCWJ における漢字頻度を用いて潜在的な漢字需要を推定するという手法を提案した。この手法により、児童の漢字需要を評価するために一定の成果を示せたものと考える。

学習漢字の妥当性が頻度だけでなる様々な観点から複合的に評価すべきものであることは、先行研究の全てに共通する見解である。丹保 (2016) も、BCCWJ における頻度のみならず様々な観点から検討を行い、「彼」は高頻度漢字ではあるが用法が限られているため、配当表漢字にはふさわしくないと結論している。しかしながら、漢字の使用頻度や潜在的な需要も、その漢字の重要度を評価するための主要な指標の一つであることは疑いない。本研究で利用した作文資料は特定のテーマに沿って書かれたものであるため、児童の書き言葉の全体に対する代表性という観点からは問題の残る部分もあるが、学習漢字の評価を考える上で従来なかった新たな観点を提案するものとして、今後の研究における参考の一つとなることを期待する。

謝　辞

本研究は JSPS 科研費 JP16H00011 の助成を受けたものです。本研究で利用した言語資源のうち、『現代日本語書き言葉均衡コーパス』は国立国語研究所が開発した言語資源です。『児童・生徒作文コーパス』の本文は科研費基盤 (B)「言語研究の実践的応用に関するリサーチユニット」(代表: 矢澤真人)、形態論情報の一部は漢検研究助成「作文コーパスを資料に児童・生徒の漢字使用・選択傾向と発達の実態を明らかにする」(代表: 宮城信) による成果物です。データの利用を許諾いただいた各位に感謝します。

文　献

宮城信・今田水穂 (2015a).「『児童・生徒作文コーパス』の設計」　第 7 回コーパス日本語学ワークショップ予稿集, pp. 223–232.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). "Balanced Corpus of Contemporary Written Japanese." *Language Resources and Evaluation*, 48, pp. 345–371.

今田水穂 (2017).『『児童・生徒作文コーパス』形態論・係り受け情報データ』, (バージョン 1.3) (2017 年 2 月作成).

宮城信・今田水穂 (2015b).「『児童・生徒作文コーパス』を用いた漢字使用能力の推定」　第 8 回コーパス日本語学ワークショップ予稿集, pp. 47–56.

棚橋尚子 (2013).「学年別漢字配当表に配当された漢字と習得語彙との関係」　全国大学国語教育学会発表要旨集　125 巻, pp. 307–310.

丹保健一 (2014).「学年別漢字配当表の字種選定を巡って: 頻度下位の 10 字種を中心に」　三重大学教育学部研究紀要, 65, pp. 73–90.

丹保健一 (2016).「学年別漢字配当表の字種選定に関する基礎的研究: 使用頻度上位の非「配当表漢字」10 字種を巡って」　三重大学教育学部研究紀要, 67, pp. 33–48.

河内昭浩 (2015).「国語教育のための「常用漢字表」語例の検討」　第 7 回コーパス日本語学ワークショップ予稿集, pp. 113–122.

関連 URL

発達段階と到達目標を考慮した学齢別漢字重要度評価法の開発

https://sites.google.com/site/kaken16H00011/

作文を支援する語彙・文法的事項に関する研究プロジェクト

https://sites.google.com/site/sakubunshienpropject/

現代日本語書き言葉均衡コーパス (BCCWJ)

http://pj.ninjal.ac.jp/corpus_center/bccwj/

コーパス構築における発話アライメントの現状

石本 祐一 (国立国語研究所コーパス開発センター) *

Present Condition of Automatic Alignment of Utterance Transcription for Speech Corpus Development

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)

要旨

音声コーパスの構築にあたり、音声信号に対し発話・音韻・韻律などの各種ラベルを付与する必要がある。これらのラベルは音声分野の知識を有した作業者による目視や聴音を基に付与されることがほとんどであり、大規模コーパス構築において大きな負担となっている。特に近年研究対象となることが多い自発発話では、言い誤りや言い澱み、曖昧な発声などの現象が頻繁に生じるため、自動ラベリングを困難にしている。本稿では、転記テキストのラベリングに焦点を絞り、既存の音声認識によるシステムを応用した自動アライメントの現状について報告する。自発発話が収録されている「日本語話し言葉コーパス (CSJ)」および「日本語日常会話コーパス (CEJC)」を用いてシステムの性能評価を行い、自動アライメントの今後の課題について述べる。

1. はじめに

音声コーパスを様々な研究分野で活用することを考慮すると、音声信号から読み取れる情報が種々のラベルとして付与されていることが望ましい。例えば、言語研究では使用されている文法や語彙に着目するために単語境界や品詞などの形態論情報が求められるし、会話研究では形態統語的な情報以外に発話中のポーズや発話タイミングも重要となる。音声学的研究においてはイントネーションやアクセントなどの韻律情報が必要となるし、音声工学的研究では言語情報に加えて基本周波数やスペクトルなどの音響特徴量が用いられる。他にもパラ言語的研究では感情や態度といった発話に対する印象評価が必須となる。このように研究の目的によって音声コーパスに求められる要素が異なることから、コーパスを幅広い研究分野に供するためには付与するラベルの充実がコーパス構築における重要課題となる。

しかし、これまでに公開されている音声コーパスにそのような種々のラベルが付与されていることはほとんどない。これはラベリングに対する負担が非常に大きいためである。ラベルの多くは音声・言語分野の知識を持った作業者により人手で付与される必要があり、コンピュータによる自動解析が利用できる一部のラベルについても最終的には人手による修正が不可欠であることが多い。このラベリングの負担を軽減しコーパス構築を容易にするためには、コン

* yishi@ninjal.ac.jp

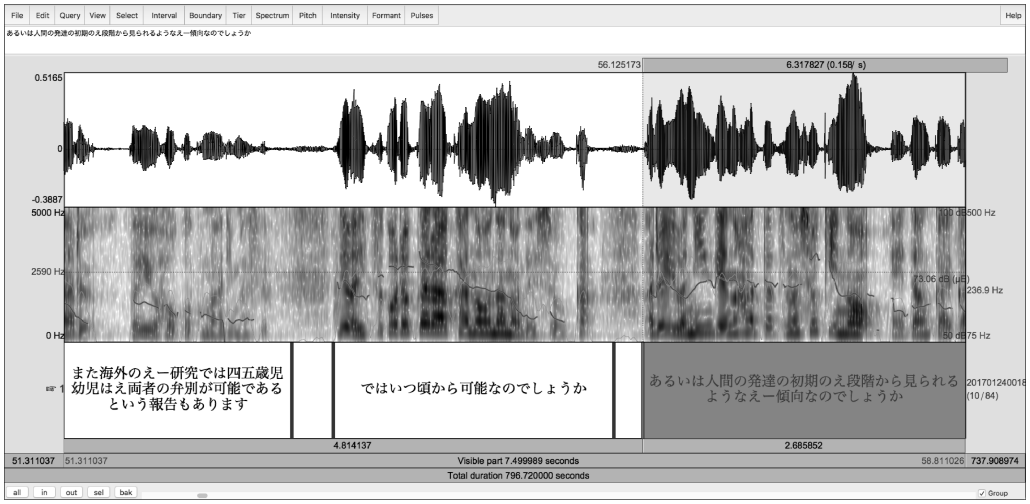


図 1　Praat による発話開始・終了時刻のアノテーション

ピュータによるラベリングの自動化が適用される範囲を広げるほかない。

本稿では、音声コーパスに付与されるラベルのうち発話を文字で書き起こしたテキスト（以下、転記テキスト）に焦点を絞り、音声データへの転記テキストの配置について、コンピュータでの自動処理における現時点での実用可能性について報告する。

2. 転記テキストのアノテーション

音声コーパスの構築においては発話に関わる様々な情報がラベルとして付与される。そのひとつである転記テキストは音声から文字への単なる書き起こしにとどまらず

- 発話単位
- 発話内の時間関係 (ポーズ)
- 発話間の時間関係 (発話の重なりや発話間の空白時間)
- 韻律・非言語情報 (強調や笑いなど)
- 非流暢性 (言い誤りやフィラーなど)

などの情報を表している。コーパスに付与される形態論情報や詳細な韻律情報といったその他のラベルはこの転記テキストを基にするため、コーパスの基盤となるものである。

しかし、転記テキストのアノテーション作業は転記基準を熟知した作業者による手作業によるところが大きく、コーパス構築における初期の問題となっている。例えば、比較的容易な発話の開始・終了時刻の認定においては、波形やスペクトログラムが表示される音声分析ソフトウェア（図 1）を用いて、実際の音声を聞き波形を見ながら数 ms 単位での調整が必要となる。つまり、発話位置を探し転記テキストを開始・終了時刻に合わせ調整（アライメント）する作業だけで発話の実時間の数倍・数十倍の時間が費やされることになり、このような作業が自動化されるだけでもコーパス構築の負担軽減が期待できる。

3. 音声認識を用いた転記テキストの自動アライメント

音声情報処理研究において、検索対象の語に適合する音声データの位置を特定する「音声ドキュメント検索」と呼ばれる問題がある (秋葉 2010)。音声ドキュメント検索は (1) 音声認識と (2) 音声と認識結果との関連づけを組み合わせた技術であり、音声ドキュメント検索が実用化されれば、その応用でコーパス構築における転記テキストの書き起こしおよびアライメント作業の自動化も可能となるであろう。しかし、実環境に存在する雑音の影響や自発発話の非流暢性などの問題から日常場面での音声認識の精度はまだ不十分である。そこで本項では、発話を書き起こしたテキストがすでに存在する状態を仮定し、テキストと音声とを関連づけることで発話位置を認定する「転記テキストのアライメント」の自動化について検討する。

3.1 自動字幕作成システム

書き起こしテキストデータから映像・音声内の位置を特定する既存システムとして、音声認識を用いた自動字幕作成システム (秋田ほか 2015, 河原ほか 2016) が公開されている。このシステムは、音声ファイルや映像ファイルを入力とし、音声認識による書き起こしをタイムスタンプ付きで出力して字幕として提示できるようにする目的で構築されており、実際に放送大学の講義の字幕付与に利用されている。また、音声認識結果をそのまま書き起こしテキストとして用いるのではなく、あらかじめ入力されたテキストに対して音声を同期させる（テキストに音声の時刻を付与する）「同期限定モード」があり、上述の転記テキストの自動アライメントを行うシステムとしての利用が期待できる。ただし、字幕作成に特化したシステムであるため、発話終了時刻は重視されていない。そこで、本稿ではアライメントについて発話開始時刻だけを取り上げることとする。

3.2 データ

すでに転記テキストが付与されているコーパスデータを用い、自動字幕作成システムによるアライメントの結果と比較することで、システムによる自動アライメントの可能性を探る。

データは、日本語話し言葉コーパス (CSJ)(Maekawa et al. 2000) と日本語日常会話コーパス (CEJC)(小磯ほか 2015) から抜粋して用いた。

CSJ からは

- 学会講演 2 名分（男女各 1 名）
- 模擬講演 2 名分（男女各 1 名）
- インタビュー対話 2 対話分（インタビュイー男女各 1 名）

を用い、学会講演発話、模擬講演発話、インタビュアーの発話、インタビュイーの発話の 4 タイプについてシステムのアライメント結果を調べた。インタビュー対話をインタビュアーとインタビュイーに分けたのは、インタビュアーの発話はフィラーや相槌が多く、インタビュイーの発話とは異なる傾向をみせると考えられたためである。システムへの入力には音声と転記テキストを用いる。CSJ では話者ごとに近接マイクを配置して音声を収録しているため、音声は雑音の非常に小さいクリアな音質となっている。テキストについては CSJ に付与されている転記テキストから転記記号を全て取り除いた上で節単位絶対境界または強境界を発話区切りと

表 1　CSJ に対する発話開始時刻の推定数

	学会講演	模擬講演	対話	
			インタビュアー	インタビュイー
正解数	185	190	317	247
推定数	185	190	309	245
検出率	100.0%	100.0%	97.5%	99.2%

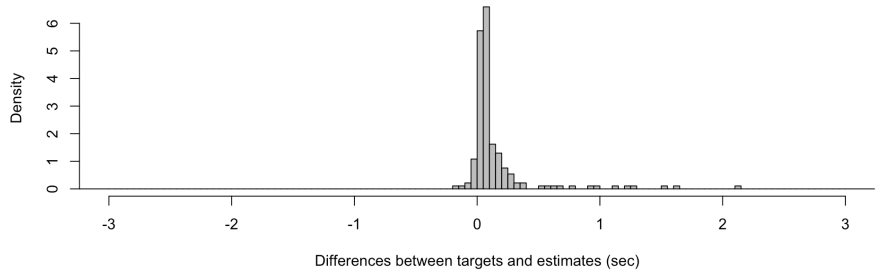


図 2　CSJ の学会講演における発話開始時刻の推定誤差

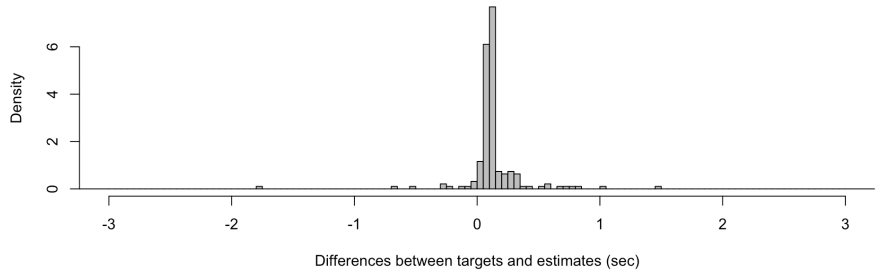


図 3　CSJ の模擬講演における発話開始時刻の推定誤差

して設定した。なお、自動字幕作成システムでは講演・スピーチ・討論の 3 つの音声認識モデルが選択できるが、講演モデルは CSJ の学会講演、スピーチモデルは CSJ の模擬講演のデータにより構築されており、CSJ データに対してそれぞれ対応する音声認識モデルを選ぶことで理想的な環境でのシステム出力とみなすことができる。

CEJC はまだ構築が済んでおらず公開されていないが、作業者による転記テキストのアライメントが完了したデータから

- 環境音の大きい飲食店内の女性 2 名の対話（以後、会話 1）
- 環境音のほとんどない室内の女性 2 名の対話（以後、会話 2）

の 2 会話を用いた。会話 1 の話者 2 名（以後、話者 A、話者 B）と会話 2 の話者 2 名（以後、話者 C、話者 D）のそれぞれについてシステムのアライメント結果を調べた。CEJC では話者ごとに IC レコーダを配置して収録しているため、システムへの入力には各話者の IC レコーダの音声を用いた。ただし、周囲の環境によって雑音や BGM、他者の音声などが入り込んでおり、話者の音声は必ずしもクリアではない。入力テキストには、書き起こしテキストを音響的

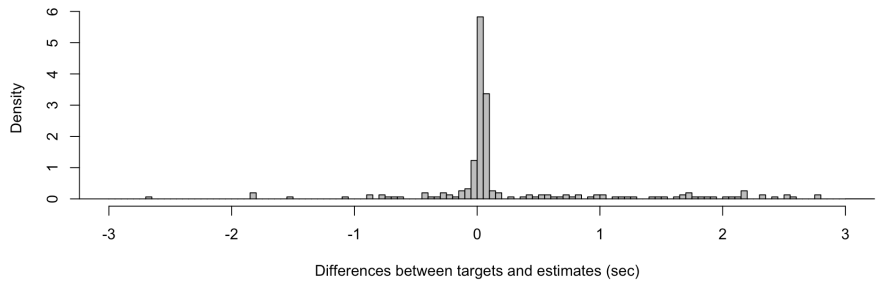


図 4　CSJ のインタビュー対話（インタビュアー）における発話開始時刻の推定誤差

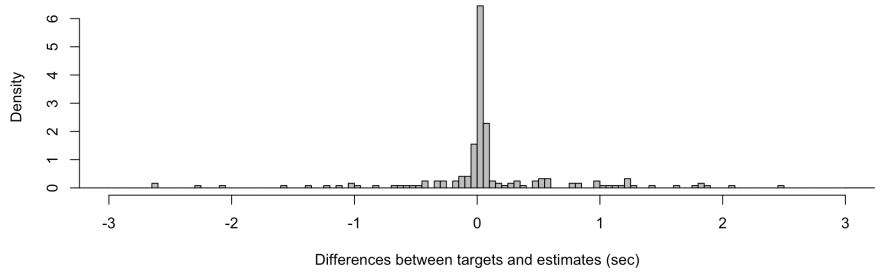


図 5　CSJ のインタビュー対話（インタビュイー）における発話開始時刻の推定誤差

な切れ目や韻律的な切れ目で区切った「短い発話単位」(Den et al. 2010) を基にした単位を用いた。そのため、CSJ よりも短い発話が多いデータとなっている。また、システムの音声認識モデルはスピーチのみを使用した。

3.3　結果

自動字幕作成システムではすべての入力テキストに対してアライメントが行われるわけではなく、発話位置の推定ができないこともある。表 1 に CSJ のデータにおいて発話開始時刻を推定できた発話数を示す。

学会講演、模擬講演ではすべての発話に対して発話開始時刻を推定できているが、インタビュー対話については少数ながらも推定できていない発話があった。推定されなかった発話は「うん」「うーん」「ええ」「はー」といった波形振幅が小さく 1 発話の長さが短い発話がほとんどであった。ただし、同様の発話であっても推定されているものもあるため、小さく短い発話がまったく推定できないわけではない。むしろ、97% 以上の発話が推定できていることから、非常に高い検出精度をシステムが有しているといえる。

次に、コーパスにあらかじめ付与されている発話開始時刻を正解値として、システムで推定された発話開始時刻との差を推定誤差として算出した。図 2–5 に CSJ のそれぞれの発話タイプにおける推定誤差のヒストグラムを示す。ヒストグラムの bin 幅は 50ms とした。±3 秒以上の誤差を生じた発話も存在したが、少数であるため図示の対象外としている。

図 2,3 からわかるように学会講演、模擬講演に対しては推定誤差が非常に小さくなっており、ほとんどが ±300ms 程度の範囲におさまっている。これは非常に高い精度で発話開始時刻を

表 2　CEJC に対する発話開始時刻の推定数

	会話 1		会話 2	
	話者 A	話者 B	話者 C	話者 D
正解数	656	798	994	1014
推定数	651	788	974	930
検出率	99.2%	98.7%	98.0%	91.7%

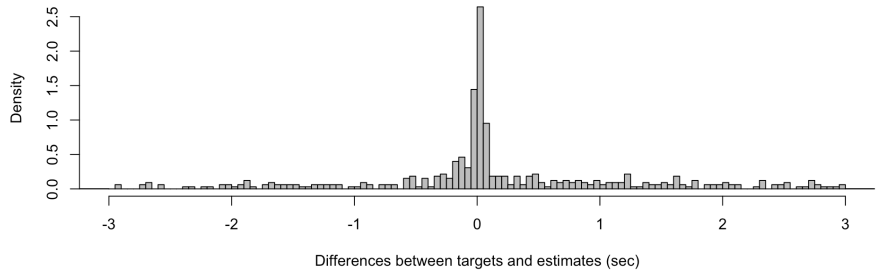


図 6　CEJC の会話 1・話者 A における発話開始時刻の推定誤差

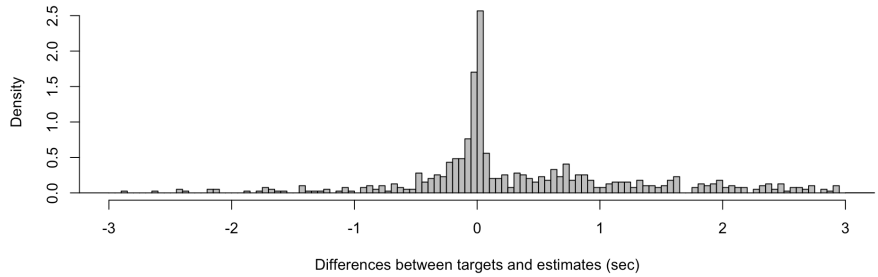


図 7　CEJC の会話 1・話者 B における発話開始時刻の推定誤差

推定できていることを示している。一方、図 4,5 に示されるインタビュー対話の推定誤差をみると、概ね ±300ms 程度におさまっているがなかには ±1,2 秒程度のズレが生じているものもあり、学会講演や模擬講演よりも精度が低下している。このような大きな誤差が生じる発話を個別にみると、ほとんどが「うん」や「うーん」といった上述の推定できなかった発話と同種のものであった。インタビュイーとインタビュアーの間で推定誤差の傾向に大きな違いは見られないが、これはインタビュアーに多いフィラーや相槌が検出不能としてある程度除かれた後の評価であるためと考えられる。

CEJC のデータにおける推定数と推定誤差についても同様に調べた。表 2 をみると、会話 1 の話者 A, B および会話 2 の話者 C に対しては 98% 以上という高い検出率を示した一方で、会話 2 の話者 D に対しては 92% 弱の検出率となった。会話 1 では環境音が大きく入り込み雑音があるにもかかわらず検出不能な発話が少ないことになる。しかし、図 6,7 で示される推定誤差からわかるように誤差が大きい発話も多く現れ、高精度の推定できているとはいえない。環境音の小さい会話 2 の結果を示す図 8,9 からも同様に推定誤差が大きく、特に検出率が低

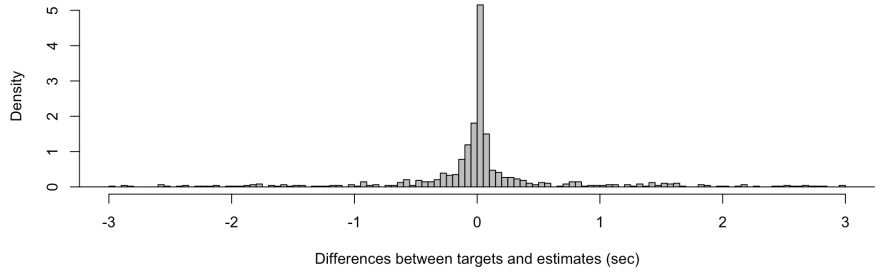


図 8　CEJC の会話 2・話者 C における発話開始時刻の推定誤差

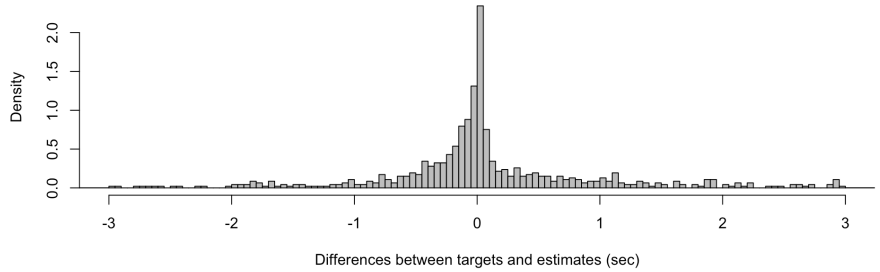


図 9　CEJC の会話 2・話者 D における発話開始時刻の推定誤差

かった話者 D に対しては誤差が大きい発話が話者 C よりも多くみられる。話者 D の検出不能の発話には「うーん」のようなフィラーだけではなく 3 秒程度のある程度の長さの発話も含まれているが、総じて大きさが小さい発話であった。また、会話 1,2 ともに推定誤差が大きい場合は全く異なる発話を指し示していることになるが、ひとつの発話の推定時刻がずれることにより後続の発話の推定時刻を誤る箇所がみられた。

3.4　考察と今後の課題

CSJ の学会講演や模擬講演において高精度で発話開始時刻を推定できているのは、音声認識の性能が大いに関係していると考えられる。すなわち、高い認識率を示す環境であれば、システムを用いた転記テキストの自動アライメントはほぼ実用的な段階に入っているといえよう。

しかし、CEJC の会話 1 のように環境音が大きい場合は検出率は高いものの推定誤差が非常に大きくなった。これはその環境音を誤って発話として認識してしまうことが原因と考えられる。また、CEJC の会話 2 のように環境音が小さい場合でも推定誤差が大きくなることがある。これはマイク位置が対象話者から離れていることにより対象話者以外の音声が入り込み、非対象話者の音声を誤って認識していることが理由のひとつとして挙げられる。以上のことから、雑音・非対象話者を含む音声に対する認識器の耐雑音性向上がシステムの適用範囲を広げるための重要な要素になっている。もっとも正しく推定できている発話も多数あることから、現段階のシステムの性能でも自動アライメントに加えて作業者による後処理を施すことを考慮すれば、コーパス構築の負担軽減には十分に役立つ状況であるといえる。

今回利用した自動字幕作成システムでは耐雑音のために振幅が小さい信号を認識対象外にす

るように構成していると考えられるが、その結果、自発会話で多く現れるフィラーや相槌への対応が難しくなっている。日常場面での収録においては収録環境の設定に制約があり理想的な収録音声を得ることが困難であることから、転記テキストの自動アライメントを推し進めるためには耐雑音性を高めるとともに小さな音も正確に認識するようなシステムの改善が必要であろう。

4. 終わりに

本稿では、音声コーパス構築における負担の軽減を目指して、音声データへの転記テキストの自動アライメントについて現時点での実用可能性について検討した。コーパス構築を目的としたものではないものの、すでに実用されている音声認識による自動字幕作成システムを応用することで、ある程度の自動アライメントが可能であることが示された。この結果を基にコーパス構築で求められる特性を考慮した自動アライメントシステムの作成を進める予定である。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」により行われたものである。また、「音声認識を用いた自動字幕作成システム」の使用を許可いただいた京都大学 河原達也教授、秋田祐哉講師に感謝いたします。

文　献

秋葉友良 (2010).「音声ドキュメント検索の現状と課題」 情報処理学会研究報告 2010-SLP-82(10), pp. 1–8.

秋田祐哉・三村正人・河原達也 (2015).「音声認識を用いた講義・講演の字幕作成・編集システム」 情報処理学会研究報告 2015-SLP-108(2), pp. 1–6.

河原達也・秋田祐哉・広瀬洋子 (2016).「自動音声認識を用いた放送大学のオンライン授業に対する字幕付与」 情報処理学会研究報告 2016-AAC-2(5), pp. 1–4.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara (2000). “Spontaneous speech corpus of Japanese.” *Proc. LREC2000*, pp. 947–952.

小磯花絵・石本祐一・菊池英明・坊農真弓・坂井田瑠衣・渡部涼子・田中弥生・伝康晴 (2015).「大規模日常会話コーパスの構築に向けた取り組みー会話収録法を中心にー」 人工知能学会研究会資料 SIG-SLUD-B5(01), pp. 37–42.

Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida (2010). “Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme.” *Proc. LREC2010*, pp. 2103–2110.

関連 URL

Praat: doing phonetics by computer

`http://www.fon.hum.uva.nl/praat/`

音声認識を用いた自動字幕作成システム

`http://caption.ist.i.kyoto-u.ac.jp/`

発話文への発話者情報付与の基本設計
-『現代日本語書き言葉均衡コーパス』収録の小説を対象に-

宮嵜由美（国立国語研究所音声言語研究領域）†
柏野和佳子（国立国語研究所音声言語研究領域）
山崎誠（国立国語研究所言語変化研究領域）

Fundamental Planning of Annotation of Speaker's Information to Utterances :Focused on Novels in "Balanced Corpus of Contemporary Written Japanese"

Yumi Miyazaki (National Institute for Japanese Language and Linguistics)
Wakako Kashino (National Institute for Japanese Language and Linguistics)
Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

現在，国立国語研究所音声言語研究領域では，『日本語日常会話コーパス』（以下，CEJC）の開発が行われている。多様な話し言葉の会話行動の収録を目指す上記プロジェクトの理念と同様，本プロジェクトの目指す，書き言葉における会話場面の「発話」への発話者情報付与も重要な“日本語の会話”の一端を担うものである。

すでに公開されている『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）の約 6 割を占める書籍のサンプルには，会話場面における大量の発話文が存在する。発話文は地の文とは言語的に異なる特徴を持つことが多いため，分析に当たっては別に扱うことが妥当であるが，現在の検索環境では難しい。

そこで，本稿では，BCCWJ 収録の小説を対象に，小説特有ともいえる発話部分特定の問題点（かぎ括弧で括られない例や非現実場面での発話など）を提示する。機械抽出のみでは同定の難しい発話箇所と発話者情報付与について，その基本設計の「発話認定箇所」基準を中心に提案する。

１．はじめに

現在，前述の CEJC や国語研究所『日本語歴史コーパス』には発話者情報が付与されているものの，BCCWJ 収録の会話文には発話者情報が付与されていない。この現代の書き言葉を収録する BCCWJ の会話文にも発話者情報が付与されれば，より深い分析や他のコーパスとの比較にも寄与できるものと考える。

そこでまず，本稿では BCCWJ 収録の小説・物語への発話者属性情報を付与するにあたり，どのように発話箇所を認定していくかを問題とする。なぜなら，実際に作業をしてみると，小説という書き言葉媒体では作者個別の文体的特徴が多くみられ，会話場面における“声

†

に出したと想定される発話”の認定にもかなりの困難が生じるためである。

例えば，発話箇所を示すことの多いカギ括弧を頼りに機械的に抽出する方法をとった場合，「釣りぼり」など看板を示す文字列も抽出され，分析対象外となる箇所も少なくない。逆に，カギ括弧で括られない場合にも，声に出したと想定される発話が多数存在し，小説の会話場面における発話の姿が十分に反映されないのが現状である。

そこで本稿では，発話箇所認定の原則としてまず，「A.発話が一重カギ括弧（以下，カギ括弧）に囲まれているかどうか」，「B.声に出したと想定される発話であるかどうか」を頼りに，以下５つの基準の提案を行う。

➢発話箇所認定の基本基準

１）カギ括弧に括られた声に出したと想定される部分

２）カギ括弧に括られた１）に準ずる部分

３）カギ括弧に括られた当該の文字列の強調などを示す部分 …<非発話>

４）カギ括弧に括られない声に出したと想定される部分

５）「場面設定」を考慮した１）に準ずる部分

上記基準に従い，具体的にどのような会話場面と，そこにどのような形式で表現される発話のバリエーションが生じているのか，発話者情報や発話状況の属性付与の概要とともに報告する。

２．作業対象

2.1 BCCWJ における「発話箇所」の収録状況

表 1 BCCWJ における発話文の割合

レジスター	サンプル数	<speech>タグを含むサンプル数	<quote>タグを含むサンプル数	発話箇所を含むサンプル数	発話箇所を含むサンプル数の割合（%）
図書館書籍（LB）	10,551	5,105	8,978	9,987	94.65
ベストセラー（OB）	1,390	917	1,080	1,321	95.04
Yahoo!知恵袋（OC）	91,445	0	0	0	0.00
法律（OL）	346	0	308	308	89.02
国会会議録（OM）	159	159	122	159	100.00
広報紙（OP）	354	244	354	354	100.00
教科書（OT）	412	0	0	0	0.00
韻文（OV）	252	0	68	68	26.98
白書（OW）	1,500	0	1,352	1,352	90.13
Yahoo!ブログ（OY）	52,680	0	0	0	0.00
出版書籍（PB）	10,117	3,479	8,646	9,250	91.43
出版雑誌（PM）	1,996	844	1,787	1,844	92.38
出版新聞（PN）	1,473	199	1,455	1,457	98.91
合計	172,675	10,947	24,150	26,100	15.12

本プロジェクトで対象とする BCCWJ には，表１，レジスター欄に示す日本語の「書き言葉」のデータが収録されている。

さらにデータには，「カギ括弧」で括られた箇所に<speech>あるいは<quote>によってタグ付けが施されている。まず，この２つのタグを暫定的な発話箇所[1]とみなし，集計したものが表１である。

この<speech>もしくは<quote>タグにより，多くの発話部分を機械的に抽出することが可能である。本プロジェクトではその出現箇所の多い，図書館書籍，出版書籍，ベストセラーを対象に，さらに NDC 番号によって分類される 913 番台「文学：日本文学：小説，物語」を作業対象の出発点とした。この，NDC913 番台の<speech>もしくは<quote>によって括られた暫定的な発話箇所はおおよそ 23 万箇所に及ぶ。

３．「発話認定箇所」と「発話者情報」

3.1 発話認定箇所と具体的データ例

前述の通り，本プロジェクトで認定する基本的な発話箇所とは，原則として「A.カギ括弧で括られた」「B．声に出したと想定される発話（以下，声に出した発話)」を指す。

ただし，対象とする小説や物語によっては，場面の流れや作家個別の文体など，例１に示す二重下線部（以下，下線部）のような，声に出した発話が必ずしもカギ括弧で括られていない場合が多数ある。

例１（ｻﾝﾌﾟﾙ ID：LBp9_00190）
<speech>[2]
<paragraph>
<superSentence><sentence>Ⅰ[3]「神林家は、わしと東吾と二人だけの兄弟である。
</sentence>
<sentence>Ⅱ東吾の同意なくば、この話は成り立たぬのだ」</sentence>
</superSentence><br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<paragraph>
<sentence> Ⅲどうじゃ、承知してくれるか、と重ねて通之進がいい、東吾は畳に手を突いて、深く頭を下げた。</sentence>

【出典】平岩弓枝（2001）「春の高瀬舟」文藝春秋

[1] <quote>タグは１発話内における<speech>の内側に括られる場合があり，必ずしも<speech>タグから独立した発話文とはならない。さらに<speech>タグ部分が，必ずしも発話箇所であるとは限らない。その詳細と具体例は「４.非発話認定箇所」に示す通り。

[2] 抽出個所の多くの前後には，例１に示したような<speech><paragraph>(<superSentence>)<sentence>のタグが付与される。本稿ではスペースの都合上<sentence>タグ以降を例として提示する。

[3] 発話と認定した箇所の冒頭に付与したローマ数字は 3.2.1 に示す「図１:属性付与の作業例」と対応するものであり，暫定的に筆者が付与したものである。

この下線部が，カギ括弧で括られていないものの，声に出した発話と認定できる根拠は，同文中に「と重ねて通之進がいい」と声に出した発話を意味する動詞が付与されている点にある。このような出現例への発話者情報付与例は 3.2.1 や 5 で詳しく述べる。

カギ括弧に拠らない声に出した発話の認定には，発話部分の認定が作業者による恣意的なものであってはならない点を十分に考慮する必要がある。しかし,「人間」が何を頼りに，どのような箇所を「発話」と認定するのかという認知過程のデータの蓄積も兼ね，機械抽出だけでは同定の難しい発話箇所の認定について以下，具体例とともに検討していく。

3.2　カギ括弧に括られた声に出したと想定される発話

3.2.1 発話認定箇所とそこに付与される属性

まず，A.カギ括弧で括られ，B.声に出した発話と判断される発話認定箇所について，必ず，話者が特定できる〈話者名〉を付与する。その他，発話と認定した箇所にどのような発話者情報が付与されているか，その内容の概略を表２に，データ入力の具体例を図１に示す。

表２　発話者情報の概略

		発話者属性	内容(概略)
①	話者ID	話者名	小説内での「発話者」の呼び名
		性別	男／女／その他／不明
		年代	若年層（~19 歳）／成年層（20 歳～59 歳）／老年層（60 歳以上）ただし，６歳以下は幼年とし，若年層を選択の上，備考欄に記載
		年代の確信レベル	書籍内に記載がない場合に「？」を付与
		非人間	ファンタジー小説，ホラー小説などに登場する人間以外の話者に「○」を付与
		会話モード	方言／外国人との会話／日本語以外での会話 通話／テレパシー／声に出した引用／独話／沈黙　など
		会話認定情報	カギ括弧がないが，声に出した発話である場合／非発話（看板，メモ，語の強調等）／心内発話　など
		備考	上記属性の補足情報
②		職業	書籍内で記載のある場合に付与
		相手	誰に対する発話かを小説内の話者名を用い付与

現在の作業段階として，まず，表２①部分の話者 ID 情報付与作業が行われており，②については，筆者が作業対象の一部のデータ（現在 100 サンプル程度）に情報付与を行っている。

例１の会話例に，表２①部分の属性を付与した作業例を図１に示す。原著では同一話者による改行が挿入されないひとつのカギ括弧内の発話であっても，作業ファイルでは，図１Ⅰ，Ⅱのように<sentence>タグを境に新たに情報付与行が設けられ，その行ごとに話者

ID を付与していく。例１の場合，具体的には〈話者名〉神林通之進，〈性別〉男，〈年代〉成年層，の話者 ID 情報がⅠ，Ⅱ，Ⅲにそれぞれ付与される。

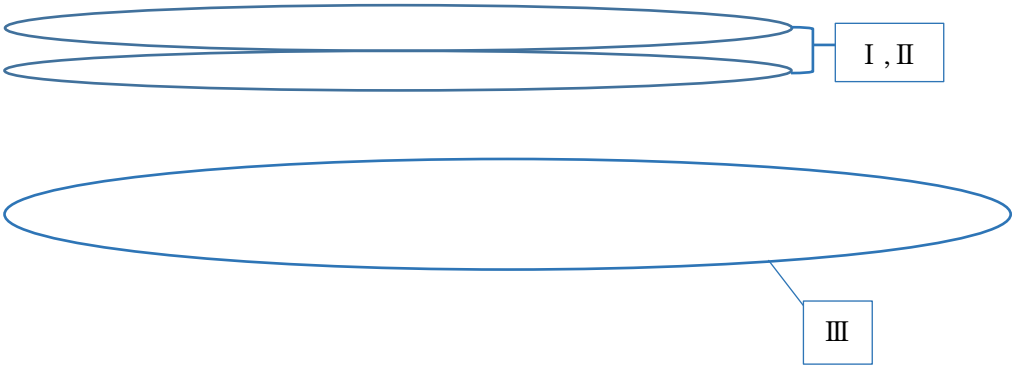


図１　属性付与作業例

ただし，Ⅲのような，カギ括弧では括られていないものの，声に出した発話として認定されるものには，〈原文〉，〈会話認定情報〉，〈備考〉欄にその旨を入力する。詳細は「５.カギ括弧が付与されていない声に出した発話」に提示する。

3.3 カギ括弧に括られたその他の発話

その他，上述 3.1 の原則に準ずるものとして，A.カギ括弧が付与され，B.声に出してはいるものの，発話者は聞き手を意識しないと想定される「独話」や，会話場面においてA.カギ括弧が付与されているが，発声の伴わない「沈黙」があげられる。また，A.カギ括弧が付与されている点を考慮し作業対象とした，「心内発話」もここで取り上げる。

3.3.1 独話

例２は，アスタシュールという男が場内のアナウンスを聞き，ひとり呻く場面である。

例２(ｻﾝﾌﾟﾙ ID：LBh9_00135)

```
<sentence>　「む…」</sentence>
<br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<paragraph>
<sentence>　すでに真昼の陽光が射す丘の上である。</sentence>
<br type="automatic_original" />
</paragraph>
```

<paragraph>
<sentence>　アスタシュールが軽い呻きをあげて立ち止まった。</sentence>
【出典】伏見健二（1993）「叛逆の獣将」中央公論社

この独話を受け，それを耳にした別の人物が「どうしたんだよ」と，カギ括弧で括られた声に出した発話が続くことから，当該下線部分もカギ括弧に括られた声に出した発話と認定した。この場合，話者 ID 情報のほか，聞き手を意識しない発話として〈会話モード〉に「独話」と記入する。

3.3.2 沈黙

「沈黙」は発言の裏返しの行為ではない。意見に対する反感や内容の吟味などその機能はさまざまである。本プロジェクトでは，沈黙は小説における会話場面内で，ある一定の意味をもつ発話の一部として抽出し，発話者情報を付与している。例３の下線部がその具体例である。

まず，「沈黙」を意味するカギ括弧で括られた三点リイダの場合であるが，①発話中のいわゆる“言いよどみ”を表す「沈黙パターン(a)」と，②カッコ内が沈黙のみで表される「沈黙パターン(b)」との２つに分類できるものと考える。

例３(ｻﾝﾌﾟﾙ ID：LBq9_00101)
<superSentence><sentence>　「…どないしたん、由香ちゃん。</sentence>
<sentence>泣いたら疲れたか？」</sentence>
</superSentence><br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<paragraph>
<sentence>　「…」</sentence> >

沈黙パターン(a)

沈黙パターン(b)

【出典】佐藤ケイ（2002）「Last kiss」メディアワークス;角川書店

「沈黙パターン(a)」の場合，発話開始の際のいわゆる言いよどみとして，開始括弧からクオーテションマーク＋終了括弧までが発話箇所と認定され，通常の発話と同様に話者 ID 情報が付与される。

「沈黙パターン(b)」の場合，(a)と同様，話者 ID 情報が付与され，さらに〈会話モード〉に「沈黙」，〈会話認定情報〉に「保留[4]」が付与される。

3.3.3 心内発話

カギ括弧が付与されているものの，声に出していない発話として，心内発話がある。例

4 「保留」とは，カギ括弧で括られてはいるものの，声に出して発話されていないものに付与される。バリエーションについては別稿にあらためたい。

４がその具体例である。これも，声に出すことで聞き手の反応を想定するものではないが，カギ括弧が付与されることを考慮し，3.1.A，B で示した原則に準ずる発話とし発話者情報付与の対象とした。

例４(ｻﾝﾌﾟﾙ ID PB29_00066)
<sentence>　<squote_A>「あれはきっと悪い夢を見たのだわ」</squote_A>と彼女は自分に言い聞かせた。</sentence>
<sentence>夢の記憶を両親に確かめるのが、なんとなく憚られた。</sentence>
【出典】森村誠一（1989）「黒魔術の女」光文社

この場合，“自分に言い聞かせた”との心内発話を意味する名詞や動詞を抽出対象の根拠とし，話者 ID 情報のほか，＜会話認定情報＞に「心内発話」と付与される。

４．非発話認定箇所

カギ括弧に括られた文字列ではあるが，当該の文字列の強調などを示すものであり，発話と認定されないものとして，例５に示す抽出例がある。

例５ （ｻﾝﾌﾟﾙ ID: LBp9_00237）
<sentence><pquote_1>「客のめし」</pquote_1>の味も、食糧の豊かな時代にあっては<pquote_2>「豚がわり」</pquote_2>に動員された屈辱を救いきれない。</sentence>
【出典】森村誠一（2001）「鍵のかかる棺　下」徳間書店

これらカギ括弧で括られた下線部は，文中における当該の語の強調を示すものであり，発話とはみなさない。このような発話として認定されないカギ括弧のデータには，話者 ID 情報を付与せず，〈会話認定情報〉を「保留」とし，発話と認定しない根拠を〈備考〉に記す。その他，声に出して読まれていない，看板，手紙やメモなどもこれにあたる。

５．カギ括弧が付与されていない声に出した発話

カギ括弧は付与されていないが，声に出した発話と判断される場合として，例１同様，例６の下線部がある。

例６ （ｻﾝﾌﾟﾙ ID: LBp9_00203）
<sentence>[それにしても…]と、蘭の方は深い安堵の吐息とともに言った。</sentence>
【出典】岩崎正吾（2001）「遥かな武田騎馬隊」角川春樹事務所

この場合，同文中にある“言った”との声に出した発話を意味する動詞を，声に出した発話と認定する根拠とする。この場合話者 ID 情報は，カギ括弧のある発話と同様に付与され，〈会話認定情報〉に「タグなし」を付与する。さらに，作業データの〈原文〉の当該発話部分を，新たにブラケット[ ]で括る。〈備考〉には，発話認定に至る根拠を示す。例えばこの場合「言った。とある」と記す。ブラケットの付与範囲は，各作業者が，地の文との境界

と判断する箇所までとする。

次にあげる例７もカギ括弧は付与されていないが，声に出した発話の例である。この例は，発話と地の文との境界を，従来の記号とその機能を頼りに機械的に抽出するには困難な例でもある。

この場合，まず，同文中にある“返事をした”との声に出した発話を意味する動詞から発話情報付与対象と判断される。また，この場合どこを地の文との境界とするかであるが，下線部ハイフンが三点リイダに相当する機能を持って付与されているものとし，その直後までをブラケットで括る。

例７（ｻﾝﾌﾟﾙID：PB59_00081）

：<sentence>　[ふうん―]と、中禅寺は感心したような馬鹿にしたような返事をした。</sentence>

<sentence>それから徐に横に視線を送って、電柱に凭れかかっていた風采の上がらない男に向けてこう云った。</sentence>

【出典】京極夏彦（2005）「百器徒然袋-雨」講談社

今日の書き言葉，打ち言葉[5]では，例えば「すみません。。」，「ありがとうございます、、」など，句読点等が，言いよどみや感情表現として使用される現象が多くみられ，本プロジェクトでも人的判断により地の文との区別を行っている。必要があれば，〈備考〉にその旨を記載する。この発話文と地の文の境界をどこに見出すか，その根拠の蓄積は，発話箇所認定に関わる新たな提案ができるものと考える。

６．場面設定による発話表示　－非現実場面での発話－

小説・物語特有の場面設定として，非現実場面があげられる。具体的には，夢の中での会話場面や，ＳＦ小説などでのロボットの会話場面，ファンタジー小説のテレパシーを使った会話場面など，まさに多種多様な場面設定がある。その多種多様性が，小説や物語という書き言葉媒体の醍醐味ともいえよう。

これらの場面設定における発話も，「声に出した発話」に準ずる「発話」と認定し，どのような場面設定での発話であったかを判別可能な状態とする情報を付与している。

6.1「夢の中」での会話例

例８は，「夢の中」という場面内での会話場面の例である。小説という書き言葉媒体の場合，場面が転換された場合や，作家固有の文体によって，声に出したと想定される発話が必ずしもカギ括弧で括られているとは言えない例のひとつとでもある。

例８作品の原著では，「夢の中」という場面での主人公の声に出した発話にはカギ括弧が，主人公以外の声に出した発話には二重カギ括弧が付与されるといった規則性がみられる。この二重カギ括弧部分は，例８の最終行にある「声」という声に出した発話を意味する名詞から，「夢の中」という場面設定における声に出した発話と認定される。

[5] 携帯メールや無料通信アプリケーション「LINE」でのやり取りでは，既に句読点は感情を表す表現として機能の拡張がみられる。

例８（ｻﾝﾌﾟﾙ ID: LBh9_00122）
<sentence>　『拓ちゃん、ごめんなさいね、駄目なのよ。</sentence>
<sentence>あたし達のせいなの』</sentence>
</superSentence><br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<paragraph>
<sentence>　あ。</sentence>
<sentence>この声は、夢ちゃんのママだ。</sentence>
【出典】新井素子（1993）「緑幻想」講談社

この場合，まず話者 ID 情報を付与し，さらに〈会話モード〉に「夢の中」が，〈会話認定情報〉に「保留」が付与される。

6.2「ファンタジー小説」－テレパシーによる会話例－

小説や物語の場面設定によっては，直接的な発声は伴わないものの，カギ括弧で会話が繰り広げられる場合がある。例９はファンタジー小説において，「意識の中」「声が響いてきた」「語りかける」など，発声は伴わないが，声に出したものと同等とされる発話を意味する名詞や動詞群から，テレパシーによる会話であることを示す例である。

例９（ｻﾝﾌﾟﾙ ID: LBd9_00046）
<sentence>　イーノゥが白い耳をピンと立てる。</sentence>
<sentence>すると、キャロルたち全員の意識の中に、マクスウェルの声が響いてきた。
</sentence>
<superSentence><quote><sentence>「聞こえるかい？</sentence>
<sentence>　私は君たちと共にいる。</sentence>
<sentence>イーノゥの目を通してすべてを見ることができるし、こうして君たちに語りかけることもできるのだ。</sentence>
<sentence>何かあった時の判断は私がしよう。</sentence>
<sentence>しかし、イーノゥの頭脳に私の意識が無理矢理入り込んでいるので、そうたくさんは話すことができない。</sentence>
<sentence>そのへんはフラッシュ、君にまかせようと思う」</sentence>
【出典】木根尚登（1989）「キャロル」CBS・ソニー出版

この場合，テレパシーの発信者に<話者名>等の話者 ID を付与する。さらに，〈会話モード〉に「テレパシー」と付与し，直接の発声を伴わない旨を〈会話認定情報〉に「保留」を記入することで，3.1 の A，B で示した原則的な「発話箇所認定」の両条件を満たすわけではないが，それに準ずるものとして検索可能となる。

７．おわりに

以上，本稿ではじめに示した発話箇所認定の基本基準１）～５）と，具体的な発話例を照らし合わせると，図２のように示すことができる。発話はＡとＢの両方の条件を備えているものが原則であるが，小説・物語という書き言葉媒体と作家の個性ともいうべき文体のあり方を考慮した上で，ＡもしくはＢの条件を備えているものを提示した。

図２　本稿で抽出した発話形態

2017 年 1 月現在，BCCWJ に収録されている小説・物語の約 2000 ファイルへの発話属性付与作業が行われている。そこには本稿で示した具体例以外にも多種多様な場面設定や話者設定があり，多種多様な発話の表現形態が存在している。作家によっては，場面転換（夢の中と現実との区別など）や発話モード（通話やテレパシーなど）の転換の演出として，カギ括弧以外の記号が使用されている例も多くみられる。

読み手である人間が，何を基準にどこまでを発話対象とするか。発話者情報属性付与作業を通し，その過程を俯瞰し整理した上で，階層的な構造を持つ日本語の書き言葉における会話行動の姿を提示していくことができると考えられる。

今後も，汎用性のある書き言葉媒体の会話場面における発話者情報コーパス構築を目指すとともに，データ整備作業を通し，人間が発話と認定する要因とその階層性についても考察していきたい。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（代表：小磯花絵），JSPS 科研費 15H03212（代表：山崎誠），16K02714（代表：宮嵜由美）の助成を受けたものです。

また，本プロジェクトの発話者属性付与作業については，国立国語研究所技術補佐員，立花幸子さん，田嶋明日香さん，平本智弥さんにご協力いただきました。ここに感謝致し

ます。

文　献

小磯 花絵・土屋 智行・渡部 涼子・横森 大輔・相澤 正夫・伝 康晴（2016）「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」国立国語研究所論集 (10), pp.85-106

小西 光・中村 壮範・田中 弥生・間淵 洋子・浅原 正幸・立花 幸子・加藤 祥・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(2015)「『現代日本語書き言葉均衡コーパス』の文境界修正」国立国語研究所論集 (9) pp. 81-100

砂川 有里子（1988.a）「引用文の構造と機能 : 引用文の 3 つの類型について」文藝言語研究. 言語篇 13,pp. 73-91

砂川 有里子（1988.b）「引用文の構造と機能(その 2) : 引用句と名詞句をめぐって」文藝言語研究. 言語篇 14, pp.75-91

村井 源（2016）「主体語彙辞書を用いた物語テキスト中の主体推定システムに向けて」人間科学とコンピュータシンポジウム発表論集 pp.209-214

山崎 誠（2007）「『現代書き言葉均衡コーパス』の設計」特定領域研究「日本語コーパス」平成 18 年度研究成果報告書『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法―現代日本語書き言葉の文字数調査― pp.20-27

関連 URL

コーパス検索アプリケーション『中納言』　https://chunagon.ninjal.ac.jp/

夢梅本『倭玉篇』全文テキストデータベースの構築

高橋 大希（北海道大学文学研究科修士課程）
劉 冠偉（北海道大学文学研究科博士課程）
池田 証壽（北海道大学文学研究科）

Construction of a Mubaibon Wagokuhen Full-text Database

Daiki Takahashi (Graduate School of Letters, Hokkaido University)
Guanwei Liu (Graduate School of Letters, Hokkaido University)
Shoju Ikeda(Graduate School of Letters, Hokkaido University)

要旨

本発表は,夢梅本『倭玉篇』の全文テキストデータベースの構築と,その利用について述べるものである。『倭玉篇』は中世に生まれ近世まで広く用いられた漢和字書である。多種の写刊本が現存し,それらについて研究が行われてきたが,その多くは部首配列や特定の部を対象にした部分的なものであった。そこで『倭玉篇』の,特に和訓に関する全体的な研究を目的として,慶長 10 年（1605)刊行の夢梅本『倭玉篇』の全文テキストデータベースを構築した。この字書は『大広益会玉篇』を中心とした中国辞書を編纂基盤としており,すでに構築されているデータを用いて効率的に入力作業を進めることができる。構築したデータベースは約 22,000 字を収録する掲出字テーブルと,約 24,000 の和訓を収録する和訓テーブルからなり,中世末期の字訓対応の資料としても価値がある。また,このデータは「平安時代漢字字書総合データベース」の和訓データの整備にも使用される予定である。

１．はじめに

『倭玉篇』は室町時代初期に成立したとされる部首分類体の字書である。鈴木（2014）によると,慶長年間までのものでも 50 種以上の写刊本が現存しており,現存最古は『延徳三年写本』（1491）である。『倭玉篇』は室町時代における漢字と訓の対応を示す資料として価値のあるものであり,いくつかの影印本や索引が出版されて活用されている。

『倭玉篇』に関する研究には,諸本の分類・系統に関する研究が多く,部首配列や少数の部をサンプルとした掲出字配列・和訓についての調査が行われてきた。一方で,それぞれの本の全巻を通した調査に基づいてその特徴を明らかにしようとした研究は少ない。この背景としては,『倭玉篇』諸本の掲出字数の多さ,文字同定の難しさから,全体的な調査を容易にするテキストデータの作成が困難であったということが考えられる。

日本古辞書のテキストデータベースとしては,すでに「平安時代漢字字書総合データベース（HDIC)」（URL : http://hdic.jp/）が一部公開されている。このデータベースは,日本字書の『篆隷万象名義』,『新撰字鏡』,図書寮本『類聚名義抄』,観智院本『類聚名義抄』と,中国字書の『玉篇』,『大広益会玉篇』,『大宋重修広韻』,『龍龕手鑑』からなる。池田（2014）が述べるように,このデータベースの構築に際しては,日本字書の解読を効率的かつ正確に行うために,関連する中国字書のデータを同時に構築し,それを参照するという方法をとっている。

以上のような状況を踏まえ,『倭玉篇』のうちの一本である夢梅本『倭玉篇』（以下,『夢梅本』と略す）の掲出字,注文のすべてをテキスト化した全文テキストデータベースを構築した。これは初の中世漢字字書のデータベースである。本データベースには,約 22,000 字の掲

出字と,約 24,000 項目の仮名書きの義注が収録されている。

『夢梅本』は慶長 10 年（1605）に刊行された,『倭玉篇』の中ではごく初期の版本である。書誌的な情報については,岡井（1933）,中田・北（1976）に詳しい。『夢梅本』の各部首内の掲出字は宋本『大広益会玉篇』を基盤としており,配列もほぼ同じであるため,HDIC でとられた方法と同様に『大広益会玉篇』のデータを用いて文字同定と入力を効率的に行うことができた。また,『大広益会玉篇』の収録字に日本漢字音と和訓を付したような『夢梅本』のデータは,他の『倭玉篇』諸本や,『大広益会玉篇』の影響を受けて成立した日本字書の電子テキスト化にも活用できる可能性がある。

本データベースは,辞書史研究のみならず,大量の漢字と訓の対応を示す言語資源として語彙史の研究にも資することを目的に構築された。『夢梅本』の国語資料としての価値は,その注文構造の特殊性にある。『倭玉篇』の多くは掲出字に対して仮名で音と和訓が付されるのみで,ほとんど漢字注を持たないのに対し,この『夢梅本』は掲出字のほとんどに漢字注が付いている。漢字注を付すことには,和訓の意味を分かりやすくすることや,当該の和訓が付されている根拠を明示するという目的があったとみられる。漢字注を持たない字書においては,漢字と和訓の対応関係は,掲出字とそれに付された和訓という構図でしかとらえることができないのに対し,ほとんどの掲出字に漢字注がある『夢梅本』では,ある和訓が掲出字自体に結びついた和訓なのか,漢字注に結びついた和訓なのかを分析することが可能となる。

本発表では,構築したデータベースの構造と入力方針を説明し,利用の一例として『大字典』和訓データベースとの比較結果を示す。また,今後のインターネット公開や,他の辞書データベースとの連携に向けた課題について述べる。

２．データベースの構造

底本には,中田・北（1976）の無窮会神習文庫蔵本の影印版を用いた。データを入力するソフトウェアには,Excel を使用した。

データベースの範囲は夢梅本の掲出字と注文のみで,目録,付録,刊記は含まない。

一般に字書データベースに求められる機能としては漢字検索と,仮名検索とがある。これに対応するため,（A）掲出字テーブルと（B）仮名注テーブルの二つを設けた。

（A）掲出字テーブル

掲出字テーブルは,掲出字一字を単位としたテーブルで,漢字検索に対応する。掲出字テーブルは①ID,②部首番号,③部首,④掲出字,⑤漢字注,⑥仮名音注,⑦仮名注,⑧備考の八つのフィールドから成る。

①掲出字 ID

掲出字の所在位置を表す ID を入力する。ID は「冊_頁_行_段」であらわす。

「冊」は,五分冊されているうちの所在を表す。数字の範囲は 1 から 5 である。

「頁」は,所在する頁が,第一冊から文字が印刷されている頁を全巻通して数えて何番目であるかを表す。この頁数は中田・北（1976）の影印本に書かれているものである。数の範囲は 001 から 594 である。

「行」は半丁七行のうち,右側から何行目にあたるかを表す。例えば,ある部首 A が一行目で終わり,一行空白があって次の行から部首 B が始まる場合,部首 B の字は三行目から始まっているとする。数の範囲は 1 から 7 である。

「段」は,ある行の中で,その字が上から何番目の掲出字であるかを表す。数の範囲は 1 から 9 で,10 字目,11 字目を表すためにそれぞれ a,b を用いる。

例：5 巻 445 ページ 4 行 11 段　→　5_445_4_b

②部首番号

当該の部首の出現順の番号を入力する。

③部首

部首名を入力する。

④掲出字

掲出字を入力する。

⑤漢字注

本文中に漢字で書かれた注と,項目の一部に付された振り仮名を入力する。「打也」「亦作赴」のように注文の意味でまとまった単位を項目と呼ぶこととする。同一のセル内に複数の項目が入る場合には「/」で区切りを示す。

⑥仮名音注

掲出字の周辺に置かれる片仮名で書かれた音注。

⑦仮名注

掲出字の下に置かれる片仮名で書かれた注文。漢字注と同様に,複数の項目の間は「/」で区切る。

⑧備考

本文と合わせて使用する上での注意点。

次の表 1 は掲出字テーブルの入力例である。上から順に,図 1～3 が対応する。引用の都合上,画像には本文の写しを使用した（以下も同様)。

表 1　掲出字テーブル入力例

①	②	③	④	⑤	⑥	⑦	⑧
1_033_6_3	4	言	討	治也	タウ	ヲサム/ウツ	
2_188_5_6	27	牛	犅	養牛羊也/今作劬	ス	ウシヒツシヲヤシナフ＊/クサ	訓点
5_548_6_2	165	血	衅	牲〈セイ〉血〈ケツ〉塗（ヌリテ）器ニ祭（マツル）也/亦作釁	キン	チヌル	訓点

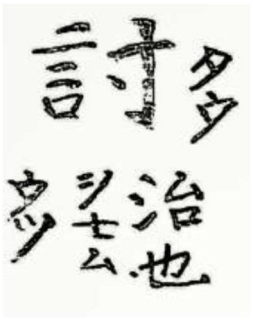


図 1　言部「討」

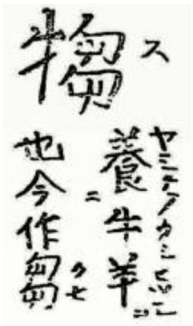


図 2　牛部「犅」

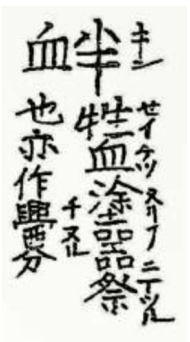


図 3　血部「衅」

（B）仮名注テーブル

仮名注テーブルは,仮名注の一項目を単位としたものであり,仮名検索に対応する。『夢梅本』では掲出字下に仮名書きされる注文には和語と漢語のどちらも含まれているため,語種を区別せずに仮名注と呼ぶこととする。

仮名注テーブルは①掲出字 ID,②仮名注 ID,③仮名注,④仮名修正,⑤掲出字,⑥漢字注,⑦部首,⑧部首番号,⑨備考の九つのフィールドから成る。

①掲出字 ID

掲出字の所在位置を表す ID を入力する。

②仮名注 ID

掲出字 ID の末尾に「出現順」を付したものを入力する。

「出現順」は仮名注を区別するために便宜的に付した値である。基本的には,掲出字下のスペースの中で,右上から右下,左上から左下に数えて何番目に出現するかを示している。

③仮名注

掲出字の下に置かれる片仮名で書かれた注文を入力する。

④仮名修正

夢梅本では,仮名遣いの乱れ,活用形の不統一,濁点の有無によって,仮名注をそのまま入力するだけでは検索に不便が生じる。そのため,仮名注の形式を修正したものを「仮名修正」に入力する。修正に際しては,影印本の索引,『日本国語大辞典』,『時代別国語辞典室町時代篇』を参考にした。また,踊り字と合字は開き,漢字は仮名に直した。和訓の検索や他の古辞書との連携にはこの列を用いる。同音異義語を区別するため,その語に対応する代表的な漢字を丸括弧に入れて末尾に付す。

⑤掲出字

掲出字を入力する。

⑥漢字注

本文中に漢字で書かれた注を入力する。

⑦部首

本文の各部首冒頭に掲げられた部首を入力する。

⑧部首番号

当該の部首の出現順の番号を入力する。

⑨備考

本文と合わせて使用する上での注意点を入力する。

次の表 2 は仮名注テーブルの入力例である。上から順に図 4～6 と対応する。

表 2　仮名注テーブル入力例

①	②	③	④	⑤	⑥	⑦	⑧	⑨
4_346_1_5	4_346_1_5_2	アニ	アニ（兄）	兄	昆也/男子先生爲一（兄）	兄	70	
5_535_6_3	5_535_6_3_1	アニ	アニ（豈）	豈	安也/焉也	喜	151	
3_290_4_1	3_290_4_1_3	サイワヒ	サイハヒ	履	皮曰一（履）/又踐也/禄也	尸	52	

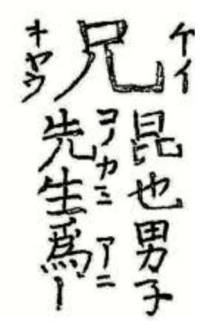


図 4　兄部「兄」

図 5　喜部「豈」

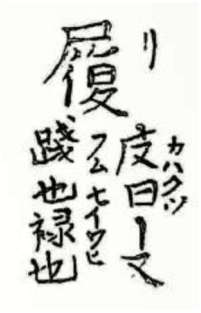


図 6　尸部「履」

３．入力方針

３．１　漢字

漢字は,Unicode の範囲内で可能な限り原本に近い字体を用いて符号化した。符号化できない字の場合,IDS(Ideographic Description Sequence)による方法で表す。たとえば,「⿰扌丁」の場合,「打」の字体を表す。IDS の部品となっている符号は「{}」で括る。IDS を使っても符号化が難しい文字に関しては対応する康熙字典体にギリシャ文字（β,γ）を付して表す。対応する康熙字典体もない場合は「■」で表す。

熟語や異体字が連続して掲出される場合は,そのそれぞれに注が付いているものとして考え,一字ずつに分ける。熟語や異体字が連続して掲出されている字には,「備考」にそれぞれ「熟字」「連字」と記す。

注文の中で掲出字を表す「一」は,その次に掲出字を丸括弧に入れて補って示す。

３．２　仮名

仮名は現在使われている字体に直す。

合字「メ（シテ）」「ㄱ（コト）」はそのまま合字を入力する。

二字連続の踊り字（くの字点）については「ヽヽ」で表す。

音読符「ー」は,漢字注のフィールドでは音読として扱うため角括弧で括って示す。仮名注のフィールドに入る場合は,括弧に入れずに示し,後ろに推定される漢字音を角括弧に入れて示す。例えば,次の図 7 の場合は部分的にしか仮名がついていないため,漢字注のフィールドになり,「ト〈ー〉メ問フヲ吉凶ヲ曰フー（敷）ト」となる。図 8 の場合は,仮名注のフィールドに入れるため,「カイコー<クワ>スル」となる。

図 7　又部「敷」

図 8　卵部「孵」

３．３　記号類

欠損などで判別不可能な字については「□」で表し,推測ができるものについては「□(ル)」のように丸括弧の中に補って示す。

明らかな誤字は,原文通りの表記の後に正しい文字を丸括弧に入れて表す。誤字がある場合,「備考」に「誤字」と記す。

注文の意味でまとまった単位を項目と呼ぶこととする。漢字注であれば「打也」「亦作赴」,仮名注であれば「アタヽカナリ」「ウツコトハナハタシ」などが項目にあたる。同一のセル内に複数の項目が入る場合には「/」で区切りを示す。

異本注記が禾部「秅」の注文に一箇所見られた。漢字注の「四百秉爲一」の中の「秉」に対して「束イ」と書かれていたため,この注記を鉤括弧に入れて「四百秉「束イ」爲一（秅）ト」のように表した。

４　符号化率

本データベースの掲出字符号化率は次の表 3 の通り 99.23%であった。符号化できなかった掲出字は 176 字で全体の 0.76%であった。この中には IDS 方式で表したものと,符号化できずに「■」を入力した字が含まれている。

表 3 掲出字符号化率

所属	字数	パーセント
CJK	13,079	57.75%
拡張 A	3,157	13.94%
拡張 B	6,186	27.32%
拡張 C	1	0.00%
互換漢字	47	0.21%
総計	22,423	99.23%

同じように,注文内の文字についての符号化率を示したのが次の表 4,5 である。符号化できなかった文字は異なりで 112 字（0.14%）,延べで 112 字（1.50%）であった。

表 4　注文異なり符号化率

所属	字数	パーセント
CJK	6,058	81.38%
拡張 A	501	6.73%
拡張 B	754	10.13%
拡張 C	1	0.01%
拡張 D	2	0.03%
互換漢字	16	0.21%
総計	7,332	98.51%

表 5　注文延べ符号化率

所属	字数	パーセント
CJK	77,863	97.72%
拡張 A	612	0.77%
拡張 B	853	1.07%
拡張 C	2	0.00%
拡張 D	2	0.00%
互換漢字	232	0.29
総計	79,564	99.86%

いずれにおいてもほとんどの文字の符号化ができており,十分に使用できる水準であると考えられる。

５.『大字典』和訓データベースとの比較

他のデータベースと連携させた利用の一例として,本データベースと同じように和訓データを持つ『大字典』和訓データベースとの比較を行った。

『大字典』（初版 1917 年）は国語学者の上田萬年によって編纂された漢和字書である。約 18,000 の掲出字を『康熙字典』と同じ部首画数順に配列している。重要視する和訓は太字で示され,品詞情報が付されている。

『大字典』和訓データベースは,『大字典』の重要和訓と品詞情報を収録するデータベースである。重要和訓が付された約 6,100 字の掲出字を収める和訓付き掲出字テキストテーブルと,約 10,500 個の重要和訓と品詞情報を収める和訓テキストテーブルからなる。
比較を行ったのは掲出字と和訓の二項目である。

まず両データベースの掲出字について共通字数を調べた。掲出字次の表 6 の「大字典」の列の値は『大字典』和訓データベースに収録されている掲出字数,「夢梅本」の列の値は夢梅本の掲出字数,「共通字数」は両本に共通して掲出されている掲出字の数を表す。共通字数は 5,388 字あり,これは『大字典』和訓データベースに収録されている掲出字のおよそ 88%にあたる。

表 6　掲出字比較

大字典	夢梅本	共通字数
6,106	22,646	5,388

不一致となったものの中には,「昂」と「昂」,「内」と「内」のようにそれぞれ微妙に異なる形で符号化しているものが含まれていた。このような符号化方針の違うデータベース間での掲出字の比較方法は今後の課題となる。

次に,『大字典』和訓データベースの和訓と,夢梅本『倭玉篇』全文テキストデータベースの仮名注で,同じ掲出字に同じ和訓がついている項目の数を調べた。夢梅本『倭玉篇』全文

テキストデータベース側で使用したのは「仮名修正」のデータである。結果が次の表 7 である。

表 7　和訓比較

大字典	夢梅本	一致数
10,519	24,446	3,111

一致する項目の数は 3,111 項目であり,これは『大字典』の和訓の約 3 割にあたる。これらの漢字と訓の対応は,比較的安定性の高いものとみることができる。

６．おわりに

本発表では,夢梅本『倭玉篇』全文テキストデータベースの構築について述べ,その利用例を示した。

最後に今後の展望と課題について述べる。

本データベースは掲出字に仮名書きの字音,和訓が付いたデータを収録しているため,同様の形式を持った他の辞書のデータ構築にも利用できる。本データベースの構築にあたってデータを利用した HDIC は,日本字書の字音・字訓が未整備の状態であるが,本データベースを用いることでデータ整備の効率化が期待できる。

また,本データベースは,広く国語資源として使用できるようインターネットでの公開を計画している。公開にあたっての使用許諾,公開形式については今後の課題としたい。

謝辞

本研究におけるデータベース構築は,JSPS 科研費 16H03422 による成果の一部である。
また成果の公表に関しては,北海道大学大学院文学研究科「共生の人文学」プロジェクトの助成を受けた。

文献

池田証壽（2014）.「平安時代漢字字書総合データベースの構築」北海道大学文学研究科紀要 142 号,pp. 79-90.
岡井慎吾（1933）.『玉篇の研究』,東洋文庫.
鈴木功眞（2014）.「字鏡集と和玉篇の境界と継承について」『国語語彙史の研究三十三』pp.147-162,国語語彙史研究会編.
中田祝夫・北恭昭（1976）.『倭玉篇夢梅本篇目次第研究並びに総合索引』,勉誠社.

関連 URL

平安時代漢字字書総合データベース（HDIC）　http://hdic.jp/

『日本語諸方言コーパス』の構築について

木部暢子（国立国語研究所言語変異研究領域）†
佐藤久美子（国立国語研究所言語変異研究領域）
中西太郎（国立国語研究所言語変異研究領域）
中澤光平（与那国町与那国語辞典編集業務嘱託員）

For building of “Corpus of Japanese Dialects”

Nobuko Kibe (National Institute for Japanese Language and Linguistics)
Kumiko Sato (National Institute for Japanese Language and Linguistics)
Taro Nakanishi (National Institute for Japanese Language and Linguistics)
Kohei Nakazawa (Education board of Yonaguni town, Okinawa Prefecture)

要旨

『日本語諸方言コーパス（Corpus of Japanese Dialects、略称：CJD）』とは、諸方言の談話資料を横断的に検索することのできるコーパスのことで、方言に関するコーパスとしては、日本で初めてのものである。資料として、1977～1985 年に実施された文化庁の「各地方言収集緊急調査」の談話データを利用し、標準語で検索してそれに対応する方言形とそれを含む談話の一節を検出する方式でデータベースを構築している。2021 年度までに最低 75 時間（３時間×25 地点）の方言データ（音声データ、転記テキスト、標準語テキスト）を公開する予定である。本発表では、CJD の概要と特徴、構築のプロセス、及び本コーパスを使った方言研究の一例を紹介し、CJD を活用することにより、方言研究にどのような研究の方向性が開けるのか、また、活用する際にどのような注意が必要なのかについて報告する。

１．はじめに

近年、大量の言語データの整備と言語コーパスの構築が世界各国で進み、それに基づく言語研究が盛んになっている。しかし、方言に関してはこれまで、地域横断的なコーパスはもちろんのこと、一地点の方言に限定したコーパスでさえ作成されていない。このような状況を踏まえ、国立国語研究所共同研究プロジェクト「消滅危機方言の調査・保存のための総合的研究」（2010～2015 年度）、「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」（2016～2021 年度）では『日本語諸方言コーパス（CJD）』の構築を行うこととし、現在、その作業を進めている。

本コーパスの特徴は、諸方言の談話を標準語で検索し、それに対応する方言形とそれを含む一定の発話単位を横断的に検索する点にある。言うまでもなく、方言と標準語は１対１で対応しない。そのため、標準語での方言検索には対応のずれの問題が生じる。しかし、各地方言の形態素辞書を作る時間と労力を考えると、すでにある日本語形態素辞書を利用して、標準語による検索を行い、並行的に方言形を検索するシステムの方がよいと判断した。また、諸方言コーパスがどのように利用されるかを考えてみると、標準語での検索システムは必須のように思われる。

資料としては、1977～1985 年に文化庁が行った「各地方言収集緊急調査」のデータを使用する。全体は、全都道府県 224 地点、１地点につき 30 時間程度の談話録音テープよりなる資料で、内容は当時 60 歳以上の地元出身者数人による自然談話である。一部は『全国方

† nkibe@ninjal.ac.jp

言談話データベース 日本のふるさとことば集成』（国書刊行会）として音声、テキスト、標準語訳が公開されているが、多くは未公開の状態である。本コーパスでは公開分、未公開分を合わせて、2021 年度までに最低 75 時間（３時間×25 地点）のデータをコーパスとして公開する。あわせて音声と方言テキスト、標準語テキストがダウンロードできるようにする予定である。

本コーパスの構築に向けて、現在、次のような手順で作業を進めている（詳細については第２節参照）。①方言音声の転記テキスト（方言テキスト）のチェック。②発話単位の認定。③方言テキストに対する時間アライメント情報の付与。④方言テキストに対応する標準語テキストのチェック。①の方言テキストと④の標準語テキストは、文化庁の事業の際に作成されたものがあり、これをもとにして、チェック作業を進めている。ただし、標準語テキストについては、全面的な見直しが必要である。前述のように、方言と標準語は１対１で対応するわけではないので、標準語テキストの付け方によっては、本来、検出されるべき方言形が検出されなかったり、検索結果が変わってきたりする可能性があるためである。標準語テキストの付け方については、現在、マニュアルの作成作業を進めており、CJD 公開の際にはマニュアルも併せて公開する予定である。②の発話単位の認定については、基本的に 0.2 秒の無音という基準で発話単位を認定している。また、話者同士の発話の重なりや相づち、フィラー、間投詞等のタグ付けも行っている。

上記の作業と並行して、本コーパスを用いた研究を試験的に行っている（詳細については第３節参照）。木部(2015)でも指摘したが、作業の中で次のような問題点が浮かび上がっている。

(a)各地の談話において、話者数、話者の属性（年齢や居住歴についてはある程度、指定があるが、男女、職業等については指定されていない）、話者同士の関係、話題等の統一が図られていない。例えば、話者同士の関係の統一が図られていないので、人称代名詞や待遇表現の地域差を単純に比較することはできない。また、話題により出現語彙に偏りがあることを十分に考慮する必要がある。

(b)標準語で検索し、方言形とそれを含む談話の一節を検索するという方法であることを念頭に置いて利用しなければならない。例えば、状態や感情には方言特有の語が使用されることが多く、秋田方言「あずましい」を標準語でどう検索するかというような問題がある。これについては、標準語テキストの問題として、コーパス構築作業の中で検討することになる。

２．日本語諸方言コーパス構築のプロセス

本節では、まず、CJD 構築のために行う作業の全体像を示した後、例を挙ながら具体的な取組みを紹介する。次に、一連の作業において特に重要となる物に関して、問題点とそれを解決するための試みを述べる。

２.１　日本語諸方言コーパス構築の流れ

本節では、CJD 構築のための一連の流れを述べる。図 1 は時間軸に沿って各作業を並べたものである。ここでは、「I. テキスト・音声の形成」にある作業に焦点を絞って詳細を述べる。

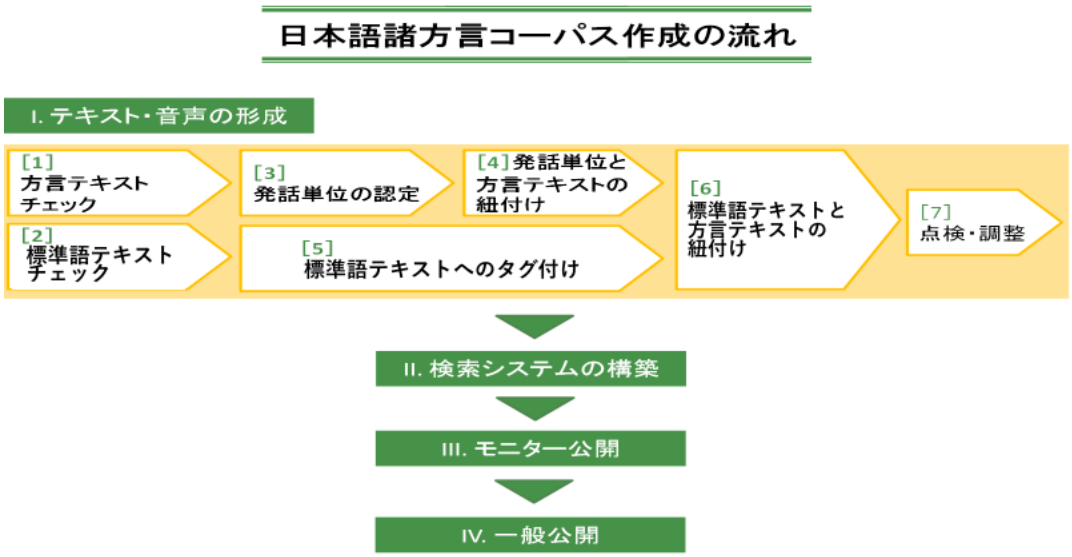


図 1. CJD 作成の流れ

［1］方言テキストチェック

1節で述べたとおり、本プロジェクトで扱う諸方言の音声データには、既に書き起こしされた「方言テキスト」と、それに対応する「標準語テキスト」がある。両テキストは複数の協力者がそれぞれの地域ごとに作成しているため、表記に揺れの問題が見られ、それらの修正が必要になる。そのために、方言間で見られる表記の揺れを統一するためのマニュアルを作成し、それに従いテキストの修正を行っている。

(1)に、方言テキストチェックの際の修正例として、東北地方を中心に見られる前鼻音の表記を示す。前鼻音を表すために地域ごとに異なる表記が用いられていたが、それらの表記を“n”に統一した。

(1)

文字化	文字化修正後
ヤンドト　シテ　カンネ	ヤｎドト　シテ　カンネ

音声の書き起こしには通常片仮名が用いられているが、必要に応じて片仮名以外の記号も用いられている。どのような記号をどのように用いるか、という点を明確にした上で、方言テキストの修正を行っている。

［2］標準語テキストチェック

標準語テキストチェックでは、標準語訳と方言が適切に対応していることを確認する。本コーパスは標準語からの検索を基本としているため、標準語テキストは、標準語としての訳の自然さではなく、方言と標準語を形態素ごとに適切に対応させることを優先している。以下に例を示す。

(2)

文字化	標準語訳
ソレオ　キカネーチャッタンダヨ	それを　きかないでしまったんだよ

標準語の訳としては自然ではないが、形態素が適切に対応するようにテキストを修正して

いる。ここまでが、「方言・標準語テキストチェック」として行う作業と、その具体例である。

［３］発話単位の認定

テキストの修正が済むと、次に、音声データの処理に進む。各地点 30 分程度の音声データを、検索上適当である単位に区切っていく。これは、表 1 に示した「発話単位認定」で行う作業である本コーパスでは方言音声が検索対象となるため、文法構造によらず、0.2 秒のポーズという音声的な基準に従って区切り目を設定している。このような基準で区切られた単位を、ここでは「発話単位」と呼ぶ。発話単位の認定作業は、プログラムと手作業で行い、この認定作業によって、音声データは発話単位に細切れにされる。

［４］発話単位と方言テキストの紐付け

続けて行うのが、この発話単位（音声）と方言テキストの紐付けである。具体的には、音声分析ソフト praat を使用し、0.2 秒のポーズで区切られた発話単位に方言テキストを貼り付けていく作業である。画面上での作業のイメージを図 2 に示す。

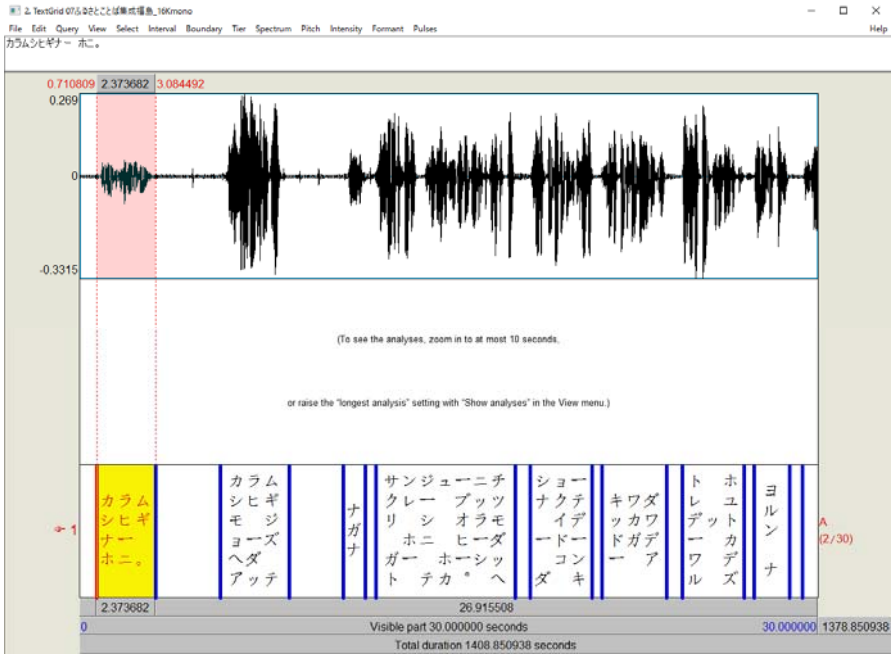


図 2. praat 上の作業イメージ図

この紐付け作業によって、コーパスの検索結果として、方言音声と方言テキストが連動して表示されることになる。紐付け作業にはいくつかの注意が必要であるが、それに関しては、２．２節で詳しく述べる。

［５］標準語テキストへのタグ付け

上記の「発話単位認定」、「音声と方言の紐付け」と平行して行う作業が、「標準語へのタグ付け」である。繰り返し述べている通り、本コーパスは標準語で方言音声・テキストを検索する仕組みとなっているため、標準語へのタグ付けはコーパス構築までの作業の中で最も重要であると言える。現時点で設定が決定しているタグ一覧を挙げる。

表 1. タグ一覧

	タグ	標準語へのタグ付けが 必要となる方言テキストの表現例
文成分の省略	［が］、［を］等	―
人称・単複	＜１・単＞　等	アッシ、オイ
終助詞	＜終助＞	ガ、ワ、ヤ
副助詞	＜副助＞	バリ、バッカイ
指小辞	＜指小辞＞	コ、メ
フィラー	＜F＞	エー、ホレ、アンタ
非言語的音	｛笑｝、｛咳｝等	―

以下では、「終助詞」と「人称・単複」のタグ付けの例を紹介する。

(3)は終助詞を表示するためのタグである。方言テキストにおいて様々形式を持つ終助詞にこのタグを付している。(4)は格標識のためのタグである。方言によっては格標識が音声的に顕在化しないことがあり、そのような場合は、方言テキストには存在しない要素をタグによって表示することになる。

(3)

文字化	標準語訳
エソエｎデ　ケサエンヤ。	急いで　ください＜終助：よ＞。

(4)

文字化	標準語訳
ヤケヒバシ　オッツケンデスヨ	焼け火箸［を］　押し付けるんです＜終助：よ＞

以上の工程で、発話単位（音声）に紐付けされた方言テキストと、タグを付された標準語テキストが揃うことになる。

［6］標準語テキストと方言テキストの紐付け

次の段階では、方言テキストと標準語テキストの紐付けが行われる。具体的に例を示すと次の通り（図 3)。

＜コーパス化作業前＞

発話No	枝番	話者	文字化テキスト	共通語訳
8	―	A	モー　ソノ　オー　オリャー　イマー　アスコラヘンニ　タキモン　カリー　イキョライ　チ　チューチ　（B　ハー　ハー）　ユー　（B　ハー）　ユーチ　ユーグライノ　コトヤッタヨ。　（B　ハー）　ナー。	もう　その　「おお　私は　今　あそこらへんに　薪［を］　刈りに　行ってるぞ」　×　と　（B　はあ　はあ）　言う　（B　はあ）　と　言うぐらいの　ことだったよ。　（B　はあ）　ねえ。

＜コーパス化作業後＞

8	000	A	モー　ソノー	もう　その
8	001	A	マー　オリャー　イマー　アスコラヘンニ　タキモン　カリー　イキョライ　チ　チューチ	「まあ　私は　今　あそこらへんに　薪［を］　刈りに　行っているぞ」　と　言って
8	002	A	ユー	言う
8	003	B	ハー	はあ
8	004	B	ハー	はあ
8	005	A	ユーチ　ユーグライノ　コト　ヤッチョローナー。	と　言って　言うぐらいの　こと　だったろうね。

図 3.『ふるさとことば集成』データのコーパス処理前後（福岡県北九州市の談話）

これによって、タグを含めた標準語を入力とし、方言音声とテキストを検索するコーパスが完成する。

［7］点検・調整

最後に、一連の作業を終えて整理された各方言における音声・方言テキスト・標準語テキストのセットの点検を行い、全体の調整を行う。様々な標準語やタグを入力として検索を行い、その結果が適切であるかどうかを確認する。地点ごとに作成されたテキストに揺れが生じている場合は、方言テキストで使用されている表記や、標準語のタグの使用に関する基準の改定や精緻化を行い、テキストの修正が必要となる。このような点検と調整を繰り返し、検索の精度を高め、方言横断的な研究に耐えうるコーパスの構築を目指す。

２.２　日本語諸方言コーパス構築作業における問題点

２.２.１　発話単位（音声）と方言テキストの紐づけ作業上の問題点

２．１節に示した通り、発話単位の認定（0.2 秒以上のポーズで区切られる有音区間を切り出す）作業は、まずプログラムによって行った。一定以上の音声の波形の振幅がある箇所とない箇所の境に自動で境界（時間情報）を入れるというプログラムである。

その結果、図 4 の矢印（➡）で示したような境界が入ったテキストグリッドが出来上がる。この境界に区切られた区間のうち、音声の波形の振幅が目立つところが有音区間であり、発話単位と認められる可能性がある区間である。

ここで「可能性がある」としたのは、後述するいくつかの理由で、この段階では正確な発話単位として切り出されていない可能性があるからである。その問題を、作業者の手で修正し（発話単位の修正）、そこに方言の文字化テキストをペースト（発話単位と方言テキストの紐づけ）していくことになる。

この作業時に問題になるのが、次のようなことである。

①背景の雑音などによる境界の過剰付与
②複数名の発話の連なり・重なりによる発話単位の結合
③単語の途中での強調された促音や、言いよどみなどによる発話単位の断絶

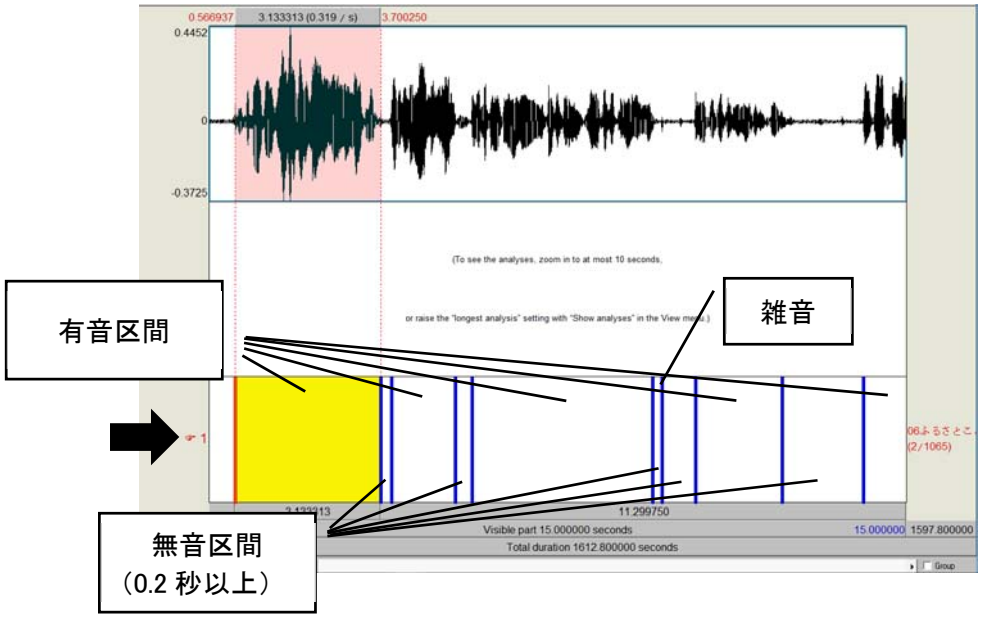


図 4．プログラムによる有音・無音区間境界付与（ふるさとことば集成：山形談話）

①については、『ふるさとことば集成』に発話とともに録音されてしまった様々なノイズ（蝉の声、電話の音、機械音など）によって、発話単位と別のところで音声波形が生じ、適切に発話単位を区切れなくなってしまうという問題である。例えば、図 4 には、前後に波形のまとまりがほぼ見られないのにも関わらず、境界が入っている部分（雑音）がある。これは録音時に入った雑音の波形が一定以上の振幅を見せたときにそれを拾ってしまった「雑音境界」である。図 4 では、一瞬の雑音であるため無視すれば問題ないが、ある程度の長さの雑音になると、そこに重なった発話が区切れなくなる。これについては、すべて作業者が音声を確認し、正確に発話単位に境界を付与する作業を行っている。

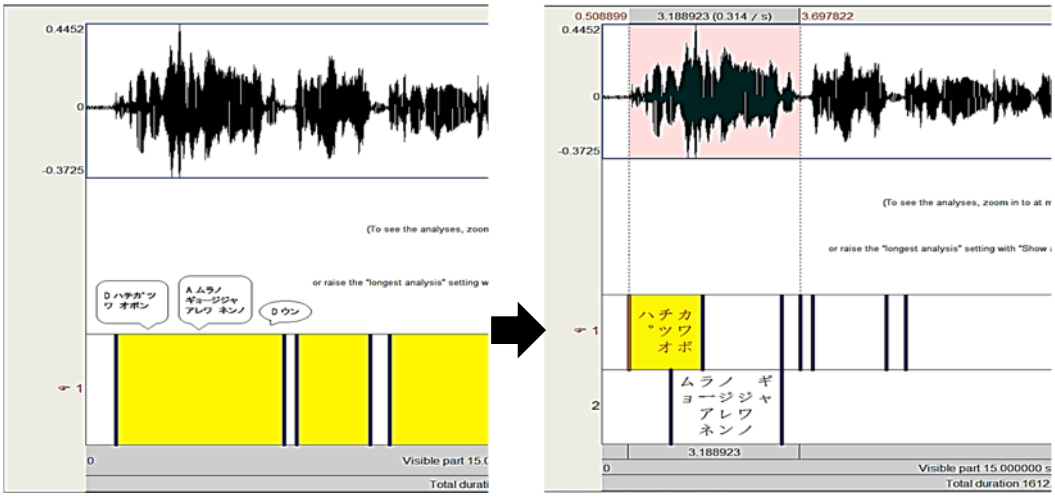


図 5a．発話の連なり・重なりによる結合　　図 5b．連なり・重なりの修正処理

②は、発話単位認定の過程で必ず処理しなければならない問題だが、複数名の発話が立て続けに（0.2 秒以内のポーズなく）なされたり、前の発話に覆いかぶさる形でなされたりすると、複数名による発話が結合された区間ができてしまうのである（図 5a）。これについては、層（Tier、図 5b の➡）を増やし、手作業で発話単位を区切ることで処理している。

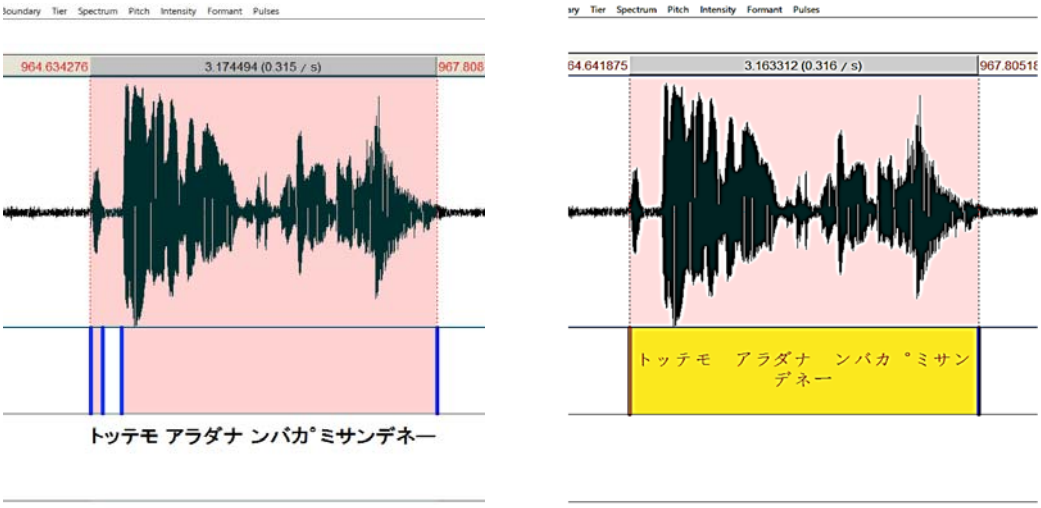


図 6a．促音による発話単位の断絶　　図 6b．促音による発話単位断絶の処理後

また、一発話単位と認めうる区間に、境界が入って断絶してしまうこともある。促音や言いよどみによる無音（無波形）区間が続く場合である。例えば、「とても」を強調して「とっても」という場合、「っ」の構えが 0.2 秒以上続くと、その間、音声波形が途切れ、境界が入って発話単位が断絶してしまうのである（図 6a）。単語の途中で言いよどみによる間ができた時も同様である。これらについて手作業で境界を消し、発話単位をつなげる処理をしている（図 6b）。ただし、この作業を行う上では、こういった途切れのうち、非語彙的な現象を、タグ付けするかどうかという点で議論がある。

２.２.２　方言テキストと標準語テキストの紐づけ作業上の問題点

CJD は、２．１節に示した通り、標準語テキストへのタグ付け作業を行う。タグの付け方次第で、コーパスは有用なものになる。例えば、西日本、特に九州では、「アンタ」「オマエ」などの形式が、発話の中の様々な箇所に頻繁に現れる。これらの形式は、標準語テキストに訳を当てる時は、直訳の方針に従って「あなた」や「おまえ」と訳すことになるが、機能としては「フィラー」として使われることが指摘されている（山本、松田 2015）。こういったものを他地域のフィラーと比較するために横断的に検索できるようにするには、「フィラー」を示すタグ「F」を用い「＜F：あなた＞」のようなタグ付けが必要になるわけである。

また、このようにタグをつけることで、各都道府県の作業担当者の違いによる標準語テキストの訳のゆれにもある程度対応できるようになる。例えば、『ふるさとことば集成』では、標準語のフィラー「ホラ」相当の形式には、全国を横断的に見渡した場合、少なくとも「ほら」という訳と、「ほれ」という訳が当てられている。これらのゆれも、＜F：ほら＞、＜F：ほれ＞のようにタグをつけることで、一括して検索することができる。

ただし、このタグの付け方に関連して、考えておくべき問題点が考えられる。

①どのような情報をタグとして採用するか
②標準語訳に対応する形式がない場合の処理をどのようにするか

①に関しては、各地の方言の特徴を踏まえて横断的に考え、先のフィラーのように、問題が生じる場合を想定し、それに対応できるタグを洗い出す作業を進めている。先に示した通り、標準語テキストの訳にゆれが生じやすいものや、直訳した標準語訳の品詞のはたらきと、運用上のはたらきにずれが生じるようなものにタグを付ける方針で設計を行っている。

また、標準語テキストに訳出するのが難しい方言形式のタグ付け処理をどのようにするかという問題もある。例えば、指小辞や副助詞、終助詞といった類いが問題となる。現時点では、対応する標準語が訳出できないものは「Z」でその存在を示し、それとともにタグ付けすることで検索を可能にする処理を施している。

(5)

文字化	標準語訳
ハヤグ**バリ**　シテ　ケサエンネ**ヤ**。	早く**＜副助：Z＞**　して　ください**＜終助：ねZ＞**。

(6)

文字化	標準語訳
ツリ**コ**サ　エク゜ダエンダゲントモ	釣り**＜指小辞：Z＞**に　行きたいんだけれども

(7)

文字化	標準語訳
オヨカ゜シェデ　クンデ　ナエ**ガワヤ**。	泳がせて　くるんで　ない**＜終助：かＺＺ＞**。

(7)のように、複合形式の場合（ガ＋ワ＋ヤ）は、対応する終助詞の単位数分を「Z」で示している。ただし、この場合も、そもそも対応する単位数を示した方がよいか、厳密に示すことができるか（融合した形式の単位認定など）という点で議論がある。

3. 日本語諸方言コーパスを用いた方言研究

本節では、CJD の検索を使って、どのような研究の方向性が開けるか、また、どのような使い方をすると不適切となるのか、具体的な検索結果を用いて、分析事例を示す。

３.１　「ホラ」相当形式に関する検索結果を用いた分析

本節では『ふるさとことば集成』を使った分析事例として、全国の「ホラ」相当形式を検索した結果を示す。なお、ここでは標準語訳の「ほら」とそのバリエーションによる検索で得られた方言形式を対象とした分析を行うため、「「ホラ」相当形式」と称した。

標準語に関する先行研究では、「ホラ」は、聞き手への注意喚起を促し、話し手と聞き手の共有知識に働きかけたり、あたかも知識などが共有可能だと表現したりするものと記述されている（大島 2001）。方言の談話展開の研究においても、情報共有喚起を促す談話マーカーとして、その使用の地域差が指摘されていたり（久木田 1990、琴 2005）、立ち上げ詞としての「ホラ」のバリエーションが記述されていたり（方言研究ゼミナール編 2006）する。だが、そもそも標準語「ホラ」の用法の全貌や、全国的な使用実態の地域差が明らかになっているとは言えない。そこで、今回『ふるさとことば集成』のデータを資料として分析を行った。

なお、『ふるさとことば集成』は、昔のことが話題に上がる会話が多いという性質から、聞き手に「ホラ」相当形式を用いて働きかける文脈が多く、用例の採取に適している。それは、話し言葉を扱った他のコーパスの検索結果との比較を見ても明らかである（表 2）。

表 2. 話し言葉のコーパスの「ホラ」検索結果

コーパスの種類	総時間（分）	「ほら」度数	出現比率
名大会話コーパス	6000	631	0.3
日常生活のことば	1058	198	0.5
ふるさとことば集成	1393	508	1.0

検索結果を抽出するにあたって、現時点ではまだデータ全体のタグ付けなどの整備が終わっていないため、検索ワードの選定の仕方が重要となる。標準語の談話研究で得られた「ホラ」のバリエーション（日本語記述文法研究会 2009、中島 2011）を参照しながら、全地点の標準語テキストのデータを形態素解析して「ホラ」と等価の形式を洗い出し、最終的に「ホラ」の音声的変異まで含めて網羅できる「ほら」「ほれ」の 2 形式で検索を行うことにした。なお、この検索ワードで検索をかけたところ、標準語テキストの訳の仕方に偏りが見られることが分かった。具体的には、静岡県の「ホラ」相当形式の検索結果のうち、40 例中 29 例が「ほら」、11 例が「ほれ」と当てられていた。そのため、検索時にこういったバリエーションへの目配りがなければ、重大な結果の相違を生んでしまう可能性がある。

今後、利用者がこういった問題に陥らないように、検索の仕方を検討する際の資料（標準語訳使用単語一覧、タグ情報一覧など）を公開し、コーパスの性質の周知を図るとともに、

検索システムの仕様の検討や研究実践を通じた利用方法の周知を行うことを検討している。

３.１.１　望ましくない分析

表 3（次頁）は、『ふるさとことば集成』の談話における「ホラ」相当形式を横断的に検索した結果である。数値は、それぞれの地域で得られた用例の度数と、何秒当たりに 1 回観察できるかという頻度を示したものである。例えば、北海道は 371 秒に 1 回「ホラ」が観察できるということを意味し、数値が小さくなるほど「ホラ」が目立つ談話ということになる。

表 3 からは、特に東北などの東日本や九州の一部の地域、さらに高知などで頻度が高く、一方、近畿を中心とした西日本にはほとんど見られないという結果が得られた。ここで、この頻度の差が「ホラ」相当形式の使用実態の地域差だと分析することについては慎重になるべきである。なぜなら、そもそもこの談話は、各地域 2 人～5 人の話者の 30 分程度の談話から導き出された結果であり、地点ごとのデータ量の少なさという問題がある上に、フィラー使用の個人差なども想定されるため、即地域差と断定するのは危険だからである。

表 3. 都道県別「ホラ」検索結果

都道府県（地域）	度数	1回／～秒	都道府県（地域）		度数	1回／～秒
北海道	6	371.0	滋賀県		0	―
青森県	45	48.6	京都府		0	―
岩手県	23	122.3	大阪府		0	―
宮城県	43	30.8	兵庫県		8	228.8
秋田県	12	129.1	奈良県		3	673.0
山形県	8	201.0	和歌山県		2	952.5
福島県	19	73.8	鳥取県		9	34.2
茨城県	4	582.5	島根県		0	―
栃木県	32	65.1	岡山県		1	1656.0
群馬県	11	215.1	広島県		0	―
埼玉県	14	162.8	山口県		0	―
千葉県	7	324.4	徳島県		10	220.5
東京都	6	348.5	香川県		0	―
神奈川県	8	260.4	愛媛県		0	―
新潟県	4	548.8	高知県		104	19.4
富山県	8	164.8	福岡県		1	1417.0
石川県	5	257.2	佐賀県		13	96.4
福井県	1	1387.0	長崎県		2	736.0
山梨県	24	66.5	熊本県		14	90.4
長野県	1	1118.0	大分県		0	―
岐阜県	0	―	宮崎県		0	―
静岡県	40	35.6	鹿児島県		19	108.9
愛知県	0	―	沖縄県A	那覇	0	―
三重県	0	―	沖縄県B	宮古島	1	720.0

３.１.２　望ましい分析

前節で「ホラ」相当形式の使用量の差について地域差とみる判断には慎重になるべきだと述べた。それならばCJDはどのような分析に有用と言えるのか。例えば、「ホラ」相当形式の分析については、次のような目的での利用が考えられる。

①各地域の「ホラ」相当形式のバリエーションの洗い出し
②日本語「ホラ」相当形式の持つ用法の幅の把握
③「ホラ」相当形式のイントネーションの比較

①に関しては、今回の検索によって、次のようなバリエーションが得られた。

(8) ホラ系（ホイ、ホエ、ホー、ホーラ、ホラ、ホラー、ホリ、ホリャ、ホレ、ホレー）、ハラ系（ハー、ハラ、ハレ）、アラ系（アラー、アリェ、アリャ、アレ、アレー、アレャ）、ソレ系（ソイ、ソラ、ソリェ、ソレ、ソレー）、オラ系（オラ、オレ、レ、レァ、レー、ロ、ロー）、オッキャ系（オッキャ、キャ、キャー）、ワヤ系（ワイ、ワヤ、ワヤー）、その他（クヤ、デ、ドー、ミナイ、メーデ）

これは、方言研究ゼミナール編（2006）を上回るバリエーション数と言える。なお、紙面の都合で割愛するが、地域ごとの使用バリエーションの差も見ることができる。

さらに用法についても、今回、従来の研究にない用法を見つけることができた。

(9) ソレン　ダンダン　アレン　ナッテクルダイナ　　アノ　ホレ　ケーケン
それが　だんだん　あれに　なってくるのだよな　あの　ほら　経験［に］

ナッテクルダイ。
なってくるのだよ。　　（『ふるさとことば集成』静岡 151、下線は筆者による）

この例は、「ソレン　ダンダン　アレン　ナッテクルダイナ」と「アレ」の適切な表現が思い出せなかったことについて、「アノ　ホレ」の部分で自分の記憶に情報喚起を促し、その結果、適切な表現としての「ケーケン」という言葉が思い浮かんだという例である。つまり、ここでの「ホレ」は、他者に注意喚起を促すようなものではなく、自分に働きかけていることを示す情報検索表示の用法と見られる。これは、内省する限り、標準語の「ホラ」でも可能で、従来の理論的な研究からは漏れていた「ホラ」の新たな用法を、今回の検索を通して見出すことができたと言える。また、このような検索で方言独自の用法が見つかれば、日本語の「ホラ」相当形式の持つ用法の広がりをつかむことにもつながる。

③に関しては分析が及ばなかったが、CJDは検索結果を通して、同じバリエーションの音声の違いを手軽に聞き比べることができることもメリットと言える。

４．おわりに

本発表では、現在、構築作業を進めている『日本語諸方言コーパス（CJD）』の概要と特徴、構築のプロセス、及び本コーパスを使った方言研究の一例を紹介し、方言コーパスの可能性と使用の際の注意点を指摘した。一般公開は2021年度の予定であるが、その前にモニター公開を行い、モニタリングの結果報告を受けてCJDをさらに改善し、一般公開へとつなげる予定である。興味のある方は、ぜひ、ご協力をお願いしたい。

謝　辞

本研究は、2010～2015 年度 国立国語研究所共同研究プロジェクト「消滅危機方言の調査・保存のための総合的研究」、2016～2021 年度 同プロジェクト「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」、2013～2015 年度 科研費基盤研究(B)「方言話し言葉コーパスの構築とコーパスを使った方言分析に関する研究」（課題番号 25284087）、2016～2020 年度 科研費基盤研究(A)「日本語諸方言コーパスの構築とコーパスを使った方言研究の開拓」（課題番号 16H01933）の支援を受けて行った。

文　献

大島弘子(2001).「「ほら」の機能について」『日本語教育』108 号, pp.34-41.

木部暢子(2015).「対格助詞ゼロの地域差－方言コーパスの可能性ー」日本方言研究会第 101 回研究発表会発表原稿集

琴鍾愛(2005).　「日本語方言における談話標識の出現傾向—東京方言、大阪方言、仙台方言の比較—」『日本語の研究』1 巻 2 号，pp.1-18.

久木田恵(1990).「東京方言の談話展開の方法」『国語学』162 号，pp.1-11.

国立国語研究所(編)(2001-2008).『全国方言談話データベース 日本のふるさとことば集成』国書刊行会.

中島悦子(編)(2011).『自然談話の文法—疑問表現・応答詞・あいづち・フィラー・無助詞—』, おうふう.

日本語記述文法研究会(編)(2009).『現代日本語文法 7 第 12 部 談話 第 13 部 待遇表現』, くろしお出版.

方言研究ゼミナール(編)(2006).『日本語方言立ち上げ詞の研究』広島大学教育学部国語教育学研究室方言研究ゼミナール.

松田美香(2015).「大分と首都圏の依頼談話」—大分方言の「アンタ」「オマエ」のフィラー的使用について—」『別府大学紀要』56 号，pp.11-22.

山本空(2015).「方言談話における対称詞の使用量の地域差」『国文学』100 号, pp.482-466.

相談における談話構造　―修辞機能と脱文脈化の観点からの分析―

田中 弥生（国立国語研究所 音声言語研究領域 / 東京大学大学院 総合文化研究科）†

Discourse Structure in Consulting: in Terms of Rhetorical Functions and the Degree of De-contextualisation

Yayoi Tanaka (National Institute for Japanese Language and Linguistics / The University of Tokyo)

要旨

本発表は，選択体系機能言語理論における談話分析手法の一つである修辞ユニット分析（Rhetorical Unit Analysis）によって，相談談話の構造を分析するものである。「修辞機能」と「脱文脈化程度」という，従来の相談談話分析にはない観点からその構造を確認する。『談話資料 日常生活のことば』（現代日本語研究会編）に収録され，「場面１」が「相談」である発話文を分析対象とする。先行研究では，ラジオ番組の医療相談や心理相談，また，インターネット上の相談コーナーともいえる Q&A サイト Yahoo!知恵袋などの談話構造の分析が行われてきたが，日常的な相談場面の分析はまだあまり行われていない。日常の生活における相談場面における談話構造を明らかにすることを検討する。

１．はじめに

相談の談話構造にかかわる先行研究に，ラジオ番組の医療相談，心理相談の談話型の解明(鈴木 2002a，2002b)や，インターネット上の Yahoo!知恵袋の情報要求モデルの提案(田中 2009，2010b)や情報構造の分析(田中 2010a)などがある。これらの談話構造分析では，佐久間(1987)の「話段」の単位によって，ザトラウスキー(1993)の発話機能が使用されることが多い。鈴木(2002a)は，ラジオの相談について，「相談内容確認」「回答」「回答確認」の 3 つの小話段が認定できるとしている。日々の暮らしにおける様々な形態の活動の様子を示す『会話行動調査』(小磯他 2016)によると，日常会話の「形式」は，「雑談」が全体の 60%強を占め，ついで「用談・相談」が 30%強であることが明らかになっている。上述のラジオ相談は，司会者がいて，専門家や知識を持つ回答者が回答し，時間が決まっている，という特徴がある。ラジオ相談やインターネットの Q&A サイトも一つのコミュニケーションの形ではあるが，より自然な会話における相談談話の解明も求められるところだろう。そこで本稿では，現在使用できる日常談話の資料として，現代日本語研究会(2016)に所収されている日常会話の相談談話を分析する。

本稿では，修辞機能と脱文脈化の観点から，相談の談話構造を明らかにすることを試みる。分析には，選択体系機能言語理論の枠組みの談話分析手法である，修辞ユニット分析（Rhetorical Unit Analysis，以下 RUA）による修辞機能と脱文脈化程度の認定を用いる[1]。以下，２.で分析方法を述べ，３.で分析結果の報告と考察を行い，４.でまとめと今後の課題について述べる。

† yayoi@ninjal.ac.jp

[1] 各種認定及び用語は原則として佐野(2010a),佐野・小磯(2011)に依った。

２．分析方法

２．１　分析対象

本稿では，『談話資料 日常生活のことば』（現代日本語研究会編）に収録され，「場面１」が「相談」，「場所」が「外出先」である発話文 1,342 件を分析対象とする。なお本資料には「場所」が「自宅」の「相談」（発話文 239 件）も含まれているが，協力者 1 名による 1 つの場面のみであるため，今回は分析の対象に含まないこととした。分析対象である 1,342 件の相談内容（場面 2）と会話人数，親疎関係と発話文数を表 1 に示す。「相談内容」（場面 2）は，現代日本語研究会(2016: 8-11)に記載されているものである。なお，以下本稿ではそれぞれの相談談話を，「美容院」「補聴器」「衣裳」「裂き織り」「アルバイト」と省略する。

表 1　分析対象―相談内容ごとの発話文数―

相談内容（場面 2）	会話人数	親疎関係と発話文数				
		本人	疎	親	不明	小計
美容院でパーマをかける前にヘアスタイルについて美容師に相談	2	99	132	0	0	231
補聴器店で家族の補聴器について店員と相談・購入	2	67	108	0	0	175
娘の通う舞の稽古場で，発表会の衣裳について先生，年配の娘の愛弟子と雑談をまじえての相談	4	159	184	0	1	344
師事する裂き織り教室の先生と住所スタンプの作成や布の販売について雑談をまじえて相談	2	247	0	209	0	456
喫茶店で，経営する居酒屋のアルバイトの女性から，新しく店を開くことについて相談を受ける	3	84	2	50	0	136
					発話文数総計	1,342

２．２　分析方法

RUA は，選択体系機能言語理論において用いられる談話分析手法のひとつで，バフチンの chronotope の概念(1981)である空間と時間の融合が言語テクストにどのように示されているかを知り，脱文脈化言語（de-contextualised language）・文脈化言語（contextualised language）の相違を捉える枠組みとして知られている(Cloran, 1994, 1999, 2010)。英語母子会話の分析の他，学校における教師と生徒の説明的な談話の様相が示され(Cloran 1999，2010)，佐野・小磯(2011)によって日本語への適用が検討され，英語と日本語の言語の違いに関わる修正が加えられている。日本語の研究では，作文指導への活用(佐野 2010)や，インターネット上の Q&A サイトの談話構造(田中 2011)やクチコミサイトの分析(田中 2013a，2013b)が行われている。テクストの意味単位を特定するための手法(佐野 2010b)だが，その過程において発話機能(speech function)，中核要素(central entity)，現象定位(event orientation)の 3 つをメッセージ単位で認定することで，修辞機能(rhetorical function)の種類を特定し，その結果として脱文脈化の程度(degree of de-contextualisation)を知ることができる。メッセージは，原則として節を最小単位として表わされるものと捉える。以下，RUA の分析方法について概説する[2]。

[2] 詳細は佐野(2010a)，佐野・小磯(2011)を参照されたい。

２．３　メッセージの認定

まず，分析対象をメッセージ単位に分割(segment)する。原則として節だが，埋め込み節はメッセージとして扱わない。メッセージの種類を図 1 に示す。主部や述部が省略されていると考えられる場合には，補足してメッセージへの分割，統合を行う。対話をデータとする場合，共話のために分割された行を統合して 1 つのメッセージと認定する場合もある。RUA では，「位置づけ」は認定対象外，「自由」と「拘束;形式的従属」を認定対象とする。「拘束;意味的従属」は従属するメッセージ（節）とともに，認定する。表 2 に「美容院」の一部の発話と，メッセージのセグメント，種類，および RUA 認定対象メッセージ数を示す。

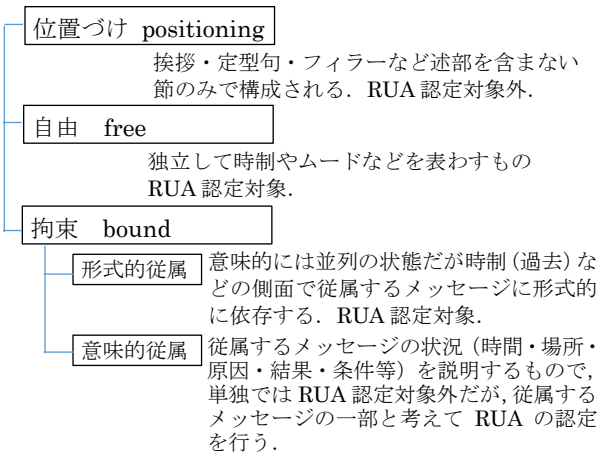


図 1　メッセージの種類

表 2 メッセージの認定例（「美容院」データの一部）

発話者	発話文番号	発話	メッセージのセグメント	メッセージの種類	認定対象数
B20f	35	＜沈黙 7 秒＞おしゃれっぽい感じにするんだったら、もう断然デジタルパーマのほうがいいんですが、エアウエーブだとー、かわいいんですよね、癖毛（げ）に近いような感じで、なって。	a. おしゃれっぽい感じにするんだったら b. もう断然デジタルパーマのほうがいいんですが、 c. エアウエーブだとー、 d. かわいいんですよね、 e. 癖毛に近いような感じで、なって。	a.「拘束；意味的従属」 b.「拘束；形式的従属」 c.「拘束；意味的従属」 d.「自由」 e.「拘束；意味的従属」	2
B20f	36	＜沈黙 10 秒＞【ヘアスタイルの雑誌を眺めながら】こういう、ちょっとこう、長い｛はい［A30f］｝感じとかだとエアウエーブでも｛うーんうんうん［A30f］｝いいかなと思うんですけどー。	a. こういう、ちょっとこう、長い感じとかだと ｛はい［A30f］｝ b. エアウエーブでもいいかなと思うんですけどー。 ｛うーんうんうん［A30f］｝	a.「拘束；意味的従属」 ｛「位置づけ」｝ b.「自由」 ｛「位置づけ」｝	1
A30f	37	＜沈黙 10 秒＞もうちょっとウエーブをかけたいんですけど。	もうちょっとウエーブをかけたいんですけど。	「自由」	1
B20f	38	こういう記事に載ってるね、大きめでもヒッカミ＃（＝意味不明）みたいな雰囲気（ふいんき）にするんだとー、デジタルパーマのほうがいいかなあと思う★んですよ。	a. こういう記事に載ってるね、大きめでもヒッカミみたいな雰囲気にするんだとー、 b. デジタルパーマのほうがいいかなあと思うんですよ。	a.「拘束；意味的従属」 b.「自由」	1
A30f	39	→あ、そうで←すか＝。	→あ、そうで←すか＝。	「位置づけ」	0
B20f	40	＝うん｛うーんうん［A30f］｝。	＝うん ｛うーんうん［A30f］｝。	「位置づけ」 ｛「位置づけ」｝	0

発話文番号 35 では，b と d の 2 つが認定対象メッセージとなる。a,c,e は，b あるいは d とともに認定を行う。なお，「～と言った」「～と思った」「～と考える」「～だと聞いた」など，話し手自身や他者による発話や考えが発話の中に引用されていると考えられる場合

は, 引用された部分 (被投射節 prefaced)を分析対象とする[3]。発話文番号 36 の b「エアウエーブでもいいかな」や，38b「デジタルパーマのほうがいいかなあ」が，該当する。また，発話文番号 39 と 40 は，「位置づけ」として分類され，この後の RUA の認定対象には含まれない。もともとデータに含まれている{　}で囲まれた他者のあいづちなども「位置づけ」である。このように，メッセージの種類を確認した結果，分析対象メッセージ数は，表 3 のとおりとなった。

表 3　相談内容ごとのメッセージ数

相談内容	メッセージ数
美容院	157
補聴器	123
衣裳	209
裂き織り	222
アルバイト	93
計	804

２．４　発話機能の認定

発話機能は，「提言 proposal」か「命題 proposition」に分類する（Halliday and Matthiessen 2004）。「提言」は表 4 の(a)の品物・行為の交換（提供あるいは命令）に関するメッセージ，「命題」は(b)の情報の交換（陳述あるいは質問）に関するメッセージが該当する。

表 4　発話機能（Halliday & Matthiessen 2004: 107）

role in exchange	commodity exchanged	
	(a)goods & service	(b)information
(i)giving	“offer” would you like this teapot?	“statement” he’s giving her the teapot
(ii)demanding	“command” give me that teapot!	“question” what is he giving her?
	提言	命題

発話機能が「提言」のメッセージは，この段階で，修辞機能は「行動」，脱文脈化指数は[1]と認定される。「＝ああー、貸してくだ★さい。」（「衣裳」発話文番号 232）「→じゃ、じゃ←、そ、そちらで発注｛はい［B20f］｝かけていただいて。」（「補聴器」発話文番号 142）など，相手に行為を要求するメッセージが「提言」である。発話機能が「命題」であるメッセージについて，この後，中核要素と現象定位の認定を行い，修辞機能と脱文脈化指数を確認する。前掲の表 2 の RUA 認定対象メッセージはすべて情報の交換で，発話機能は「命題」である。

２．５　中核要素の認定

中核要素は，コミュニケーションの当事者とメッセージの内容との空間的距離を示す。メッセージの中心となるものがコミュニケーションの場面に存在するか否かによって特定する。中核要素の分類を図 2 に示す。基本的には主語によって表現されるが，照応など前後のメッセージを用いて判断する場合もある。また，「こ

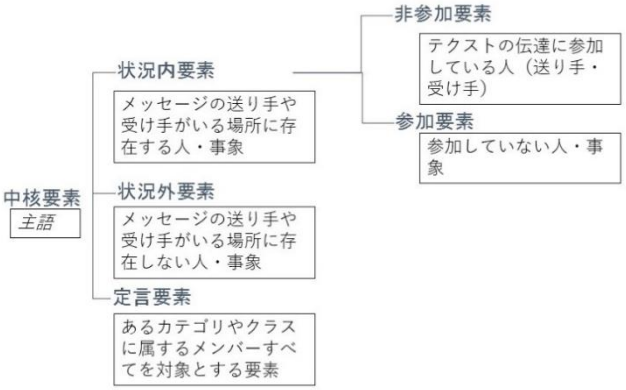


図 2　中核要素の分類（佐野・小磯 2011）

[3] 投射については佐野(2010a)を参照のこと。

のカレーは野菜がたっぷりだ」のように，述部「野菜がたっぷりだ」が「このカレー」の性質を表している場合には，「このカレー」を中核要素と認定する。

２．６　現象定位の認定

現象定位は，メッセージによって表現されている出来事がいつ起こったか，これから起こるのかなどを，メッセージが伝達されている時（Time of speaking）を基準とした時間的な位置を特定して，時間的距離を示す要素である。副詞や述部から判断する。現象定位の分類を図 3 に示す。

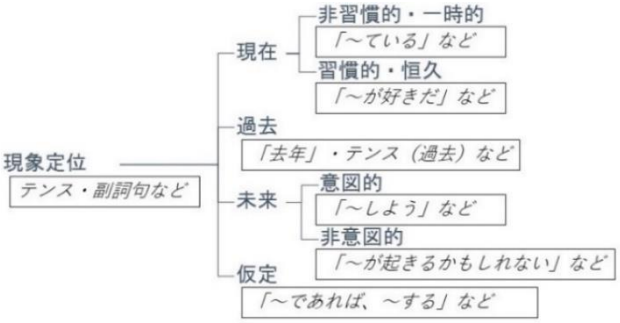


図 3　現象定位の分類（佐野・小磯 2011）

２．７　修辞機能の特定と脱文脈化指数の確認

発話機能と中核要素と現象定位の組み合わせによって，表 5 に示したように，修辞機能が特定され，脱文脈化指数を確認できる。

表 5　発話行為・中核要素・現象定位からの認定

			発話機能						
			提言	命題					
				現象定位					
				現在		過去	未来		仮定
				非習慣的 一時的	習慣的 恒久		意図	非意図	
中核要素	状況	参加	[1]行動	[2]実況	[7]自己記述	[3]状況内回想	[4]計画		[6]状況内推測
		非参加	n/a		[8]観測		[5]状況内予想		
	状況外			[9]報告	[13]説明	[10]状況外回想	[11]予測		[12]推量
	定言			n/a	[14]一般化				

「n/a」は該当なし／背景が灰色の部分が修辞機能の種類／[ ]内は脱文脈化指数

図 4　修辞機能と脱文脈化程度

脱文脈化指数とは，中核要素の here（発話地点との空間的な距離）の程度と現象定位の now（発話時点との時間的な距離）の程度によって，近いものから遠いものまで修辞機能を線上に示した際の順序の指数で，1 から 14 まである（図 4）。脱文脈化指数の数値が大きいものほど脱文脈化の程度が高く一般的・汎用的で，小さいものほど脱文脈化の程度が低く個人的・限定的であることを示す。

表 2 で示した「美容院」データに RUA 認定したものを表 7 に示す。右に「→」の後に示した脱文脈化指数と修辞機能をみると，脱文脈化指数は，13→9→2→4→13，修辞機能は「説明」→「報告」→「実況」→「計画」→「説明」と展開している。脱文脈化程度が高く会話

参加者のいる場所と空間的時間的に離れた話題から脱文脈化程度が低く会話参加者の身近な話題へ，そしてまた離れた話題へ，というように展開していることがわかる。

表 6　RUA 認定例

発話者	発話文番号	メッセージのセグメント	メッセージの種類	発話機能,中核要素,現象定位 **→脱文脈化指数. 修辞機能**
B20f	35	a. おしゃれっぽい感じにするんだったら b. もう断然デジタルパーマのほうがいいんですが、 c. エアウエーブだとー、 d. かわいいんですよね、 e. 癖毛に近いような感じで、なって。	a.「拘束；意味的従属」 b.「拘束；形式的従属」 c.「拘束；意味的従属」 d.「自由」 e.「拘束；意味的従属」	命題,状況外,現在;習慣的・恒久 **→13. 説明** 命題,状況外,現在;非習慣・一時的 **→9. 報告**
B20f	36	a. こういう、ちょっとこう、長い感じとかだと ｛はい［A30f］｝ b. エアウエーブでもいいかなと思うんですけどー。 ｛うーんうんうん［A30f］｝	a.「拘束；意味的従属」 ｛「位置づけ」｝ b.「自由」 ｛「位置づけ」｝	命題,状況内;参加,現在;非習慣一時的, **→2. 実況**
A30f	37	もうちょっとウエーブをかけたいんですけど。	「自由」	命題,状況内;参加,未来;意図的, **→4. 計画**
B20f	38	a. こういう記事に載ってるね、大きめでもヒッカミみたいな雰囲気にするんだとー、 b. デジタルパーマのほうがいいかなあと思うんですよ。	a.「拘束；意味的従属」 b.「自由」	命題,状況外,現在;習慣的・恒久 **→13. 説明**

３．　分析結果

図 5 に「美容院」談話で見られた修辞機能と脱文脈化指数の展開を示す。縦方向に談話の進行を配置し，横方向に脱文脈化程度を示した，左側ほど脱文脈化程度が低く時間的・空間的に会話当事者たちに近い「今・ここ」での修辞機能で，右側ほど脱文脈化程度が高く，会話の場面から時間的・空間的に遠い汎用的な修辞機能である。

ゆるい感じにパーマをかけたいという希望を客が述べる[4 計画] [4]ところから談話が始まる。次に美容師が，以前かけたパーマがかかっていないか[2 実況]，以前いつパーマをかけたか[3 状況内回想]，普段巻いたりするかという習慣[7 自己記述]を確認する。【A】

次いで，デジタルパーマとエアウェーブというパーマの種類についての質問[9 報告]と解説[13 説明]，これからのお手入れについての説明[8 観測][11 予測]が続く。【B】

また,過去にデジタルパーマをしたことがあるか[3 状況内回想],髪に癖があるか[8 観測][2 実況]を確認しながら，パーマをどのようにかけるとどうなるのかを解説し[13 説明][状況内予想 5]，毛質との関係からの提案をしている[2 実況][8 観測][5 状況内予想]。【C】

さらに，前に美容院に行ってからどのぐらい時間がたっているか[3 状況内回想][9 報告],具体的にどうパーマをかけるか[2 実況][11 予測][8 観測][4 計画]を相談し，決定している。【D】

映像や説明がないため定かではないが，このあたりで，会話の内容から，パーマをかける作業に入っているのではないかと推測される。パーマ作業に入っているとすれば，この後の,話題は，いわゆる雑談に分類できるものではないかと考える。

肌が弱いかどうかの確認[8 観測]や，しばらく美容院に行かなかった理由[3 状況内回想][8 観測][7 自己記述]，カラーリングをした後の一般的な話[9 報告][13 説明]，今後のカラーリ

[4] [ ]内に，脱文脈化指数と修辞機能を示す。

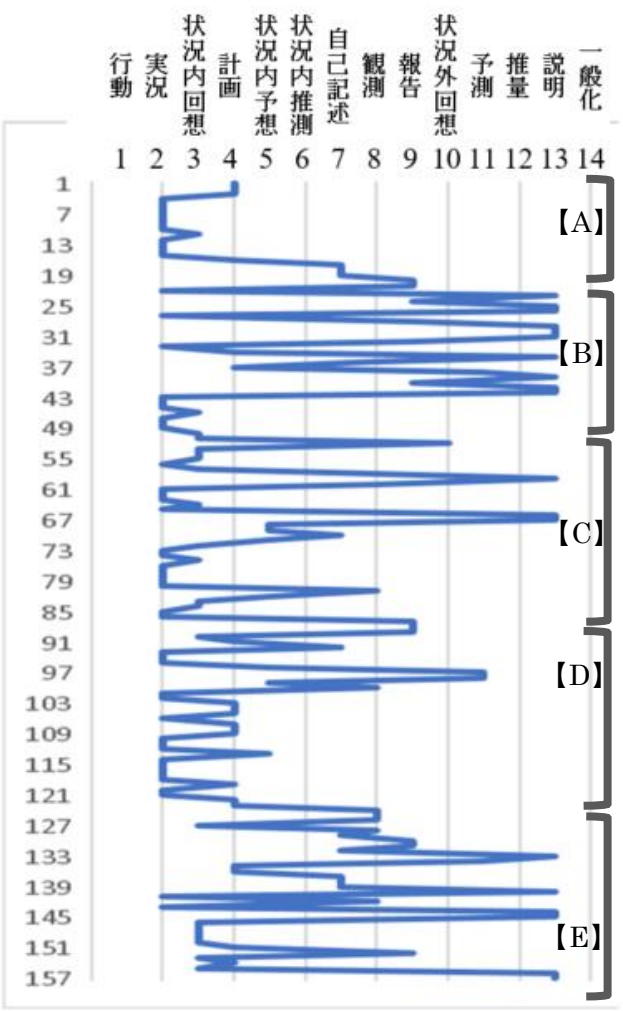


図 5　「美容院」の修辞機能・脱文脈化指数の展開

ングの予定[4 計画]，トリートメントの習慣[7 自己記述]，トリートメントの勧め[13 説明]，以前ショートカットにした失敗[3 状況内回想]，今後ショートカットにするか[4 計画]，長い方が楽だという一般論[13 説明]で終わっている。【E】

美容院における相談談話の修辞機能は，[4 計画] [2 実況]から始まり，[7 自己記述][9 報告][13 説明]と続き，全体的に，「実況」と「説明」の間で大きく揺れながら会話が進んでいることがわかる。客の要求や希望に関する問や回答と，専門家側の専門知識の提示，という大きな流れがあり，専門家と客の関係から生じる修辞機能の展開である可能性がうかがえる。

次に，「補聴器」「衣裳」「裂き織り」「アルバイト」の談話展開を図 6 に示し，確認する。

「補聴器」は，来店した本人でなく，施設等に入居中の家族が使用する機材を購入する場面である。カタログを見ながらの商品の選択[2 実況]と補聴器使用者の習慣[13 説明]の間に[9 報告][10 状況外回想]が多くみられる。状況と希望を伝える際に，その場にはいない家族（中核要素が「状況外」）について語り，器材やカタログをみながら説明をする際に，過去の機種の話（中核要素が「状況外」）をしていることが[9 報告][10 状況外回想]の出現に影響している。購入場面における相談の場合，使用者本人用か否かによって修辞機能と脱文脈化の程度は変わってくることがうかがえる。

「衣裳」は娘の発表会用の着物の柄や色について相談している。冒頭で相談者が希望[4 計画]や状況[2 実況]を述べ，その後に先生からのアドバイス[9 報告]などが提供される。途中で第三者の話題がしばらく続くため，[9 報告]の連続が見られる。

「裂き織り」は，おおまかには，前半はスタンプを作成する提案について，後半はフリーマーケットに布を出店することについてのやり取りであるが，様々な話題がところどころに入っており，他の相談とは展開が異なるようである。手元にある布（中核要素が「状況内; 非参加」）の扱いについてやりとりが続き，脱文脈化の程度の低い数値がまとまって表れている。

「アルバイト」は，お店をまかせたいといわれた相談者（アルバイト）から相談される場面である。冒頭で，ワインを注文した後に，相談があることを伝え［2 実況］，その後経緯

を伝え［9 報告］予定を話し［4 計画］たのち，相談相手からの賛成と様々なアドバイスが具体的な話［2 実況］［3 状況内回想］や一般的な励まし［9 報告］［13 説明］などが伝えられている。

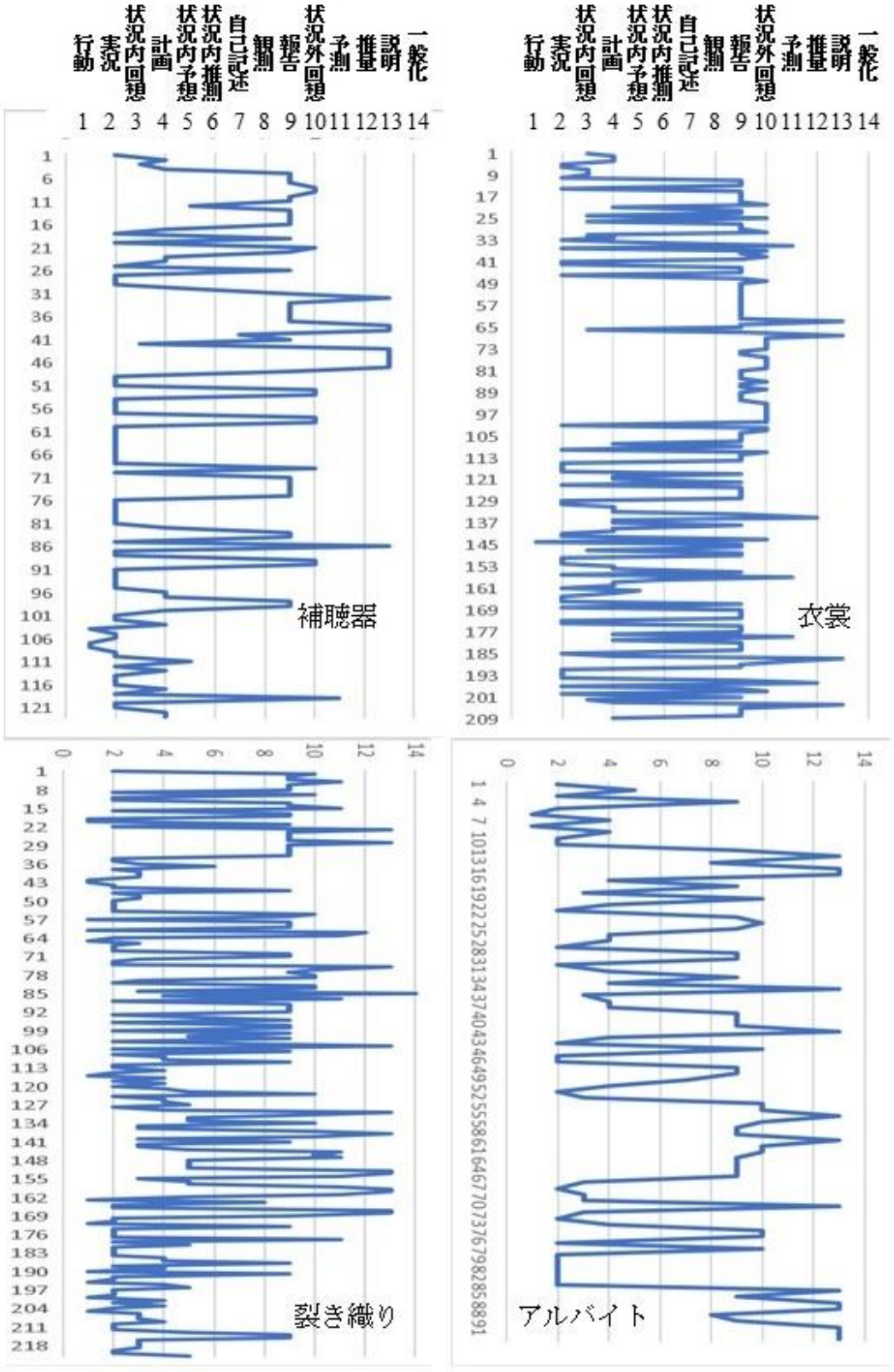


図 6　補聴器，衣裳，裂き織り，アルバイトにおける展開

相談には大きく分けて「と」相談と，「に」相談があるといえる(柏崎他 1997)。「と」相談は主に会話参加者同士が話し合って何かを決める場合が該当し，「に」相談は専門家や知識のある人の意見を聞く場合が該当する。「美容院」は美容師に，「補聴器」は店員に，「衣裳」は先生などに，「アルバイト」はアルバイト女性が本人（調査協力者）に，相談する。一方，「裂き織り」は，本人が先生と相談して決めていく「と」相談であったといえるだろう。また，鈴木(2002，2003)で分析されたラジオ番組の医療相談，心理相談は，ともに言葉（音声）で回答をもらう状況であるが，本稿で分析した相談談話のうち，「美容院」ではパーマという技術，「補聴器」では補聴器という製品にそれぞれ対価を支払うという点で，相談のゴールは言葉によるものだけではないという性質の違いがあるといえる。また，今回の分析対象の人間関係（親疎）は，「美容院」「補聴器」「衣裳」が疎，「裂き織り」「アルバイト」は親であった。人間関係が親であれば，雑談的なメッセージが多くなることは容易に想像がつく。分析の結果，技術やサービス・商品の入手を目的とした「相談」では，客側の希望・要求や背景などの情報が提供されるための脱文脈化程度の低い修辞機能と，専門家側の知識等を提供したり確認したりするための脱文脈化程度の高い修辞機能との間で揺れながら，必要に応じてその間の修辞機能が用いられる，という基本的な構造がうかがえた。

４．まとめと今後の課題

本稿では，選択体系機能言語理論における談話分析手法の一つである修辞ユニット分析(Rhetorical Unit Analysis)によって，相談の談話構造を修辞機能と脱文脈化の観点から明らかにすることを試みた。技術やサービス・商品を求める「相談」の場合，相談者の希望・要求，背景についての情報を提供する脱文脈化程度の低い修辞機能と，回答者の知識等の提供や確認のための脱文脈化程度の高い修辞機能との間で揺れながら，必要に応じてその間の修辞機能が用いられる，という基本的な構造がうかがえた。同じ「相談」という場面でも，目的（ゴール）や会話者の親疎，「と」相談か「に」相談か，などの性質によって，その構造は変異がある。その意味では今回のは性質が様々であったといえる。現在，国立国語研究所では『日本語日常会話コーパス』の構築が進められている（小磯他 2017，田中他 2017）。今後さらに「相談」の談話を分析して，本研究によって得られた基本談話構造の検証を行うとともに，その他の日常会話場面についても，談話構造の分析を行っていきたい。

謝　辞

本研究は科研費基盤(C)「「修辞機能」と「脱文脈化程度」の観点からのテキスト分析手法確立と自動化の検討」(15K02535) および国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」の助成を受けて実施した.

文　献

柏崎雅世・足立さゆり・福岡理恵子(1997)「インフォーマルな「と」相談における提案の分析」日本語教育,92, pp.60-71,日本語教育学会.

現代日本語研究会 遠藤織枝・小林美恵子・佐竹久仁子・高橋美奈子編(2016)『談話資料 日常生活のことば』ひつじ書房.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉(2017)「『日本語日常会話コーパス」の構築」『言語処理学会第 23 回年次大会(NLP2017)予稿集』

小磯花絵・土屋智行・渡部涼子・横森大輔・相澤正夫・伝康晴(2016)「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」国立国語研究所論集 10, pp.85-106.

佐久間まゆみ(1987)「文段認定の一基準（Ｉ）―提題表現の統括ー」文藝言語研究 言語篇, 11, 筑波大学 文芸・言語学系.

佐野大樹 (2010a) 「日本語における修辞ユニット分析の方法と手順 ver0.1.1―選択体系機能言語理論（システミック理論）における談話分析―（修辞機能編）」. http://researchmap.jp/systemists/資料公開/ 閲覧日 2016 年 11 月 30 日

――――― (2010b)「選択体系機能言語理論を基底とする 特定目的のための作文指導方法について ―修辞ユニットの概念から見たテクストの専門性―」専門日本語教育研究, 12, pp.19-26.

佐野大樹・小磯花絵 (2011) 「現代日本語書き言葉における修辞ユニット分析の適用性の検証ー「書き言葉らしさ・話し言葉らしさ」と脱文脈化言語・文脈化言語の関係ー」, 機能言語学研究, 6, pp.59-81, 日本機能言語学会.

鈴木香子 (2002a). 「ラジオの医療相談の談話の構造分析」早稲田大学日本語教育研究, 1, pp.117-130.

――――― (2002b). 「ラジオの心理相談の談話の構造分析」早稲田大学日本語教育研究, 3, pp.57-69.

――――― (2007). 『機能文型に基づく相談の談話の構造分析』早稲田大学日本語教育研究科博士論文 https://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/28804/3/Honbun-4605.pdf 2016 年 1 月 28 日アクセス

田中弥生 (2009). 「インターネットの知識検索サービスにおける談話構造の諸相 ―Yahoo! 知恵袋の情報要求モデルの検討―」ことばと人間, 7, pp.57-68. 「言語と人間」研究会.

――――― (2010a) 「Q&A コミュニティの談話機能と構造―「Yahoo!知恵袋」を対象に―」特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集, pp.55-62.

――――― (2010b). 「質問サイトにおける情報要求モデルと待遇コミュニケーション―「アットコスメ美容辞典」の談話機能・談話構造の分析から」待遇コミュニケーション, 7, pp.33-48. 待遇コミュニケーション学会.

――――― (2011). 「修辞ユニット分析を用いた Q&A サイトの質問と回答における修辞機能の展開の検討」社会言語科学会第 28 回大会発表論文集, pp.226-229.

――――― (2013a). 「クチコミサイトにおける修辞機能の商品評価の高低による違い ―修辞ユニット分析による検討―」機能言語学, 7, pp.59-74. 日本機能言語学会.

――――― (2013b).「評価の高低によるクチコミサイト「アットコスメ」における談話構造の特徴 ―修辞ユニット分析を用いて―」神奈川大学言語研究,35, pp.1-23, 神奈川大学言語研究センター.

田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017)「『日本語日常会話コーパス』構築における会話収録方法」『言語処理学会第 23 回年次大会（NLP2017）予稿集』

ポリー・ザトラウスキー（1993）『日本語の談話の構造分析――勧誘のストラテジーの考察』くろしお出版.

『UniDic』と『分類語彙表』の見出し対応表データの構築

近藤 明日子（国立国語研究所コーパス開発センター）†
田中 牧郎（明治大学国際日本学部）‡

Construction of a Correspondence Table between Headwords of UniDic and Headwords of "Word List by Semantic Principles"

KONDO Asuko (National Institute for Japanese Language and Linguistics)
TANAKA Makiro (Meiji University)

要旨

日本語の大規模コーパスへの網羅的・系統的な語義情報付与を目的として、各種大規模コーパスの構築に利用されている形態素解析辞書の元データである電子化辞書 UniDic の見出し（語彙素）と、大規模な現代日本語のシソーラス『分類語彙表増補改訂版データベース』の見出しとを対応づける表形式データの構築を行った（2017 年公開予定）。対応付け作業は UniDic・分類語彙表両者の見出しの読み・表記・類に基づき人手により行い、2017 年 1 月時点で、UniDic 語彙素 50,122 と分類語彙表見出し 64,045 の多対多の関連を表す対応表が構築できている。一方で、見出しの単位設計の違いにより、UniDic 語彙素と対応付けできない分類語彙表見出しの存在も明らかになった。さらに、本対応表を用いた大規模コーパスへの網羅的な語義情報付与に向けて、今後検討すべき課題の存在も明らかになった。

１．はじめに

日本語のコーパスに対する語義情報の付与は、言語研究・自然言語処理の両分野で必要度の高い課題である。意味の面から日本語の語彙全体を分析するためには、日本語の語彙を構成する語が表しうる意味の世界を系統的に分類した語義情報が付与されることが望まれる。また、語義情報を付与するコーパスは日本語の代表性を担保する大規模コーパスとし、さらにそのコーパスを構成する語すべてに網羅的に語義情報を付与することも望まれる。

そのコーパスの形態素解析に使われる形態素解析辞書の見出しデータに語義情報を付与することができれば、その解析結果であるコーパスの各語に語義情報を付与することが可能となる。そこで本研究では、複数の大規模コーパスの形態素解析に利用されている形態素解析辞書の元となる電子化辞書 UniDic の見出しデータと、大規模な現代日本語のシソーラスである国立国語研究所（2004）『分類語彙表増補改訂版データベース』（ver.1.0）[1]（以下、「分類語彙表 DB」という）において意味項目が付与された見出しデータとを対応づけた表形式データを構築した。この対応表を介して、分類語彙表 DB の意味項目を UniDic の見出しに語義情報として付与し、ひいてはそれに対応づけられたコーパスを構成する各語にも語義情報を付与することができるようになる。

２．分類語彙表 DB

まず、対応表の一方に配する分類語彙表 DB のデータについて概説する。分類語彙表 DB は、本格的な現代日本語のシソーラスの先駆である国立国語研究所（編）（1964）『分類語

† kondo@ninjal.ac.jp
‡ makiro@meiji.ac.jp

1 http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

彙表』を増補改訂した国立国語研究所（編）（2004）『分類語彙表増補改訂版』のデータベース版である。分類語彙表 DB での意味分類方式は、番号（以下、「分類番号」という）を用いてそれぞれの分類項目の体系的位置づけを示したところに特徴がある（国立国語研究所（編）2004、p.3）。分類番号は「1.3131」のような 5 桁の数字として表記され、各数字あるいはその組み合わせが「類」「部門」「中項目」「分類項目」という 4 階層の意味的範疇を示す構造となっている（図 1）。

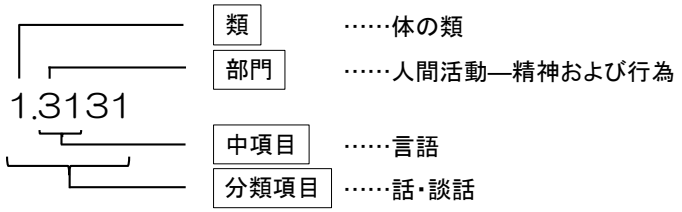


図 1　分類番号の構造

そして、この分類番号と分類項目の中をさらに分類する「段落番号」「小段落番号」、および「小段落番号」内の配列順序を表す「語番号」のもとに 98,241 の見出し[2]を配列するのが分類語彙表 DB のデータである（表 1）。

表 1　分類語彙表 DB データ例

類	部門	中項目	分類項目	分類番号	段落番号	小段落番号	語番号	見出し本体	読み
体	活動	言語	話・談話	1.3131	1	1	1	話	はなし
体	活動	言語	話・談話	1.3131	1	1	2	話	わ
体	活動	言語	話・談話	1.3131	1	1	3	トーク	とおく
体	活動	言語	話・談話	1.3131	1	2	1	談話	だんわ
体	活動	言語	話・談話	1.3131	1	2	2	談	だん
(…中略…)									
体	活動	言語	問答	1.3132	1	1	1	問答	もんどう
体	活動	言語	問答	1.3132	1	1	2	自問自答	じもんじとう
体	活動	言語	問答	1.3132	1	1	3	一問一答	いちもんいっとう
体	活動	言語	問答	1.3132	1	1	4	応酬	おうしゅう
体	活動	言語	問答	1.3132	1	2	1	禅問答	ぜんもんどう

各見出しは「分類番号」「段落番号」「小段落番号」「語番号」の 4 列により一意となる。多義語の場合は各意味の分類番号が与えられるため、1 語が複数の分類番号に配列され、それぞれ別見出しとなる。

3．UniDic

次に、対応表のもう一方に配する UniDic のデータについて概説する。UniDic とは国立国語研究所が整備している電子化辞書である。コーパスの日本語研究への応用を志向し開発された現代語に対応した辞書（伝ほか 2007）をはじめとして、「近代文語 UniDic」「中古和文 UniDic」（小木曽ほか 2013）等、各時代・文体に対応した複数の形態素解析器 MeCab[3]用

2 分類語彙表 DB 収録の全 101,070 レコードから、書籍版の分割前の見出しであることを表すレコード種別が「B」の 2,589 レコードと意味的区切り「＊」を表す 240 レコードを除いた数。
3 http://taku910.github.io/mecab/

辞書として提供されている[4]。

UniDic の特長として以下の 2 点があげられる（小椋ほか 2011、上 pp.9-10）。

(1) 見出しの単位として「短単位」を採用する。短単位とは、例えば「国立国語研究所に勤務している。」というテキストであれば、「国立｜国語｜研究｜所｜に｜勤務｜し｜て｜いる｜。」と分割する、短い語の単位である。単位の基準が分かりやすく揺れが少ないという長所がある。

(2) 表記や語形の違いにかかわらず、同じ語であれば同一の見出しを与える方針のもと、語を階層化した形で登録する。最上層に国語辞典の見出しに相当する「語彙素」、その下に語形の違いを区別する「語形」、その下に表記の違いを区別する「書字形」を設ける（表 2）。

表 2　UniDic の階層構造

語彙素	語形	書字形
ヤハリ【矢張り】	ヤハリ	矢張り
		やはり
		矢張
	ヤッパリ	やっぱり
		ヤッパリ
		やッぱり
	ヤッパシ	やっぱし
	ヤッパ	やっぱ

UniDic による形態素解析辞書で形態素解析したコーパスとして、国立国語研究所で構築された大規模コーパス『日本語話し言葉コーパス』（CSJ）[5]、『現代日本語書き言葉均衡コーパス』（BCCWJ）[6]、『日本語歴史コーパス』（CHJ）[7]がある。これらのコーパスは短単位によりテキストが区切られ、各短単位に対して UniDic のデータが形態論情報として付与されている。UniDic とコーパスの形態論情報はともに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）で管理され、UniDic とコーパスに出現する短単位が対応づけられている。よって、UniDic 見出しに語義情報が付与できれば、コーパスの各短単位に語義情報を付与することが可能となる。

４．分類語彙表 DB と UniDic の見出しの対応付け作業

ここから、本研究の主旨である分類語彙表 DB 見出しと UniDic 見出しとを対応付けた表の構築について述べる。

分類語彙表 DB の各見出しに対応づけるのは UniDic の語の複数の階層のうち語彙素とした。語彙素は、語源が同一であり、かつ意味の違いを生じていない複数の語形をまとめあげるもので（小椋ほか 2011、下 p.78）、語義情報を付与するのに適当な階層である。UniDic 語彙素は「語彙素」「語彙素読み」「語彙素細分類」「類」「語種」の 5 列により一意となる。

分類語彙表 DB の各見出しと UniDic の各語彙素との同語判別を行い、同語であれば対応付けを行った。同語判別に使う条件として以下の(A)～(C)を設けた。

4 http://unidic.ninjal.ac.jp/
5 http://pj.ninjal.ac.jp/corpus_center/csj/
6 http://pj.ninjal.ac.jp/corpus_center/bccwj/
7 http://pj.ninjal.ac.jp/corpus_center/chj/

(A) 分類語彙表 DB 見出しの「見出し本体」と UniDic 語彙素の「語彙素」が一致する[8]
(B) 分類語彙表 DB 見出しの「読み」と UniDic 語彙素の「語彙素読み」が一致する[9]
(C) 分類語彙表 DB 見出しの「類」と UniDic 語彙素の「類」との対応（表 3）が一致する

表 3　分類語彙表 DB と UniDic の「類」の対応

分類語彙表DB	UniDic
体	体 固有名 人名 姓 名 地名 国 数 接尾体
用	用 接尾用
相	相 接尾相
他	他

「類」とは品詞の上位概念に相当するもので、分類語彙表 DB では「体の類」「用の類」「相の類」「その他の類」の 4 種を設ける。UniDic 語彙素にも「類」が設けられており、その区分は分類語彙表 DB よりも細かい。そのため両者の「類」の定義や所属する見出しを参照し、表 3 の対応を設定した。

(A)～(C)の条件がすべて満たされれば同語とすることを原則とした（図 2）。

分類語彙表DB				UniDic		
見出し本体	読み	類		語彙素	語彙素読み	類
事	こと	体	—	事	コト	体

図 2　(A)～(C)による対応付け例

ただし、以下の①～③の例外ルールを設け、(A)～(C)の条件が満たされなくとも同語としたものがある。

① (A)の条件が満たされない場合でも、分類語彙表 DB の「分類項目」や UniDic 語彙素に対応づけられるコーパスの用例等を参照し、同語と判断される場合は同語とする（図 3）。

8 分類語彙表 DB の一部の「見出し本体」には「－周年」「…ている」のように UniDic 語彙素の「語彙素」との同定に不要な記号が含まれているため、この記号を除いたデータを作成し同定した。
9 分類語彙表 DB「読み」と UniDic「語彙素読み」では表記に違いがあるため、「読み」の表記を「語彙素読み」の表記にあわせて変換したデータを作成し同定した。

分類語彙表DB			
見出し本体	読み	類	分類項目
これ	これ	体	こそあど

UniDic		
語彙素	語彙素読み	類
此	コレ	体

図３　①による対応付け例

② (B)の条件が満たされない場合でも、分類語彙表 DB 見出しの「読み」と UniDic 語彙素に所属する「語形」とが一致する場合は同語とする（図 4）

分類語彙表DB		
見出し本体	読み	類
依存	いそん	体

UniDic			
語彙素	語彙素読み	類	語形
依存	イゾン	体	イソン

図４　②による対応付け例

③ (C)の条件が満たされない場合でも、UniDic 語彙素に所属する語形の「品詞」や語彙素に対応づけられるコーパスの用例等を参照し、同語と判断される場合は同語とする（図 5）。

分類語彙表DB		
見出し本体	読み	類
リアル	リアル	相
正式	せいしき	体

UniDic			
語彙素	語彙素読み	類	品詞
リアル	リアル	体	名詞−普通名詞−形状詞可能
正式	セイシキ	相	形状詞−一般

図５　③による対応付け例

以上のルールに則った同語判別作業は、専用の作業用ツールを用いて、人による判断を交え行った。分類語彙表 DB 見出しのうち UniDic に登録されていない語は、UniDic の設計上登録可能であれば新たに UniDic に登録し対応付けを行った。

５．対応づけ作業結果

2017 年 1 月現在、分類語彙表 DB の全見出しについて、ひととおり UniDic 語彙素との同語判別を終え、同語判別を保留している分類語彙表 DB 見出し約 700 を除き、分類語彙表 DB の 64,045 見出しと UniDic の 50,122 語彙素との多対多の関連を表す対応表が構築できている。

対応表に見られる対応付けの例として、分類語彙表 DB 見出しが多で UniDic 語彙素が一の対応の例を図 6 にあげる。

分類語彙表DB				
見出し本体	読み	類	分類番号	分類項目
出す	だす	用	2.3832	出版・放送
出す	だす	用	2.3770	授受
出す	だす	用	2.1531	出・出し
出す	だす	用	2.1521	移動・発着
ー出す	だす	用	2.1502	開始
出す	だす	用	2.1211	発生・復活
出す	だす	用	2.1210	出没

UniDic		
語彙素	語彙素読み	類
出す	ダス	用

図 6　分類語彙表 DB 見出しが多、UniDic 語彙素が一の対応例

このような分類語彙表 DB 見出しが多で UniDic 語彙素が一の対応は例が多く、分類語彙表 DB 見出しと対応づけられた UniDic 語彙素 50,122 の 21%にあたる 10,490 語彙素がそれぞれ複数の分類語彙表 DB 見出しと対応づけられた。一つの UniDic 語彙素に対して対応づけられる分類語彙表 DB 見出しの最大数は 13 にのぼる（表 4）。

表 4　分類語彙表 DB 見出しと対応する UniDic 語彙素数

対応する分類語彙表DB見出し数	UniDic語彙素数
1	39,632
2	8,321
3	1,421
4	501
5	113
6	72
7	28
8	15
9	8
10	6
11	2
12	2
13	1
計	50,122

逆に、分類語彙表 DB 見出しが一で UniDic 語彙素が多の対応の例を図 7 にあげる。

分類語彙表DB				
見出し本体	読み	類	分類番号	分類項目
小じゅうと	こじゅうと	体	1.2140	兄弟

UniDic		
語彙素	語彙素読み	類
小舅	コジュウト	体
小姑	コジュウト	体

図 7　分類語彙表 DB 見出しが一、UniDic 語彙素が多の対応例

このような分類語彙表 DB 見出しが一で UniDic 語彙素が多の対応は例が少なく、分類語彙表 DB の 21 見出しに限られ、対応づけられる UniDic 語彙素の最大数も 2 にとどまる。

ところで、分類語彙表 DB 全 98,241 見出しの 35%に相当する 34,822 見出しが UniDic 語彙

素と対応付けができなかったことになるが、これは見出しの単位設計の相違によるものである。UniDic は短単位の語を見出しとするのに対し、分類語彙表 DB は「有機物質」「図示する」「詭弁を弄する」といった短単位を複数つなげた合成語や連語・慣用句の類も見出しとして収録する。このような見出しは UniDic には設計上登録できないので対応付けできなかった。

６．今後の課題

構築した対応表を用いた大規模コーパスへの網羅的な語義情報付与を目指す上で、今後検討すべき課題について述べる。

第一に、コーパスへの網羅的語義情報付与のために必要な、UniDic 語彙素に対する網羅的分類番号付与についての課題がある。５．で述べたとおり分類語彙表 DB 見出しと対応づけられた UniDic 語彙素数は 50,122 であり、これは、UniDic に登録されている全 181,241 語彙素[10]（2017 年 1 月時点）の 28%に過ぎない。残る 131,119 語彙素は分類語彙表 DB に未収録の語のため、分類語彙表 DB 見出しと対応付けがとれず、分類番号が付与できない。これらの語彙素への分類番号の付与は今後の課題である。古典語であれば、宮島ほか（編）(2014)『日本古典対照分類語彙表』のデータと UniDic との対応表を別途作成し、UniDic 語彙素に分類番号を付与する方法が考えられる。それでも分類番号が付与されない語彙素については、人手による付与等の方法を検討する必要がある。

第二に、一つの UniDic 語彙素が複数の分類語彙表 DB 見出しと対応づけられる多義語についての課題がある。多義語がテキストの文脈内で用いられる場合、一般には複数の語義のうち一つが用いられる。よって、コーパスの各短単位に付与される分類番号は通常 1 種類ずつとなる。コーパスへの語義情報付与作業では、複数の分類番号が対応づけられる UniDic 語彙素の場合、その中の一つの分類番号を選択する工程が必要となる。人手による選択、語義の曖昧性解消の技術を用いた自動選択等の方法を今後検討する必要がある。

第三に、語義情報を付与する単位についての課題がある。本対応表によって実現するコーパスへの語義情報の付与は短単位に対するものである。しかし、コーパスの利用目的によっては、たとえば「勤務する」を「勤務」と「する」の短単位に分割してそれぞれに分類番号を付与するのではなく、「勤務する」全体に分類番号を付与することが要求される場合もあるだろう。この対処法として、UniDic の設計にあるもう一つの単位「長単位」に対して語義情報を付与することが考えられる。長単位は文節を自立語部分と付属語部分とに分割して得られる長い語の単位で（小椋ほか 2011、上 p.4)、たとえば「国立国語研究所に勤務している。」というテキストは「国立国語研究所｜に｜勤務し｜ている｜。」と分割される。CSJ・BCCWJ・CHJ には長単位による形態論情報も付与されており、これに語義情報を付与することは理論上可能である。分類語彙表 DB には長単位に相当する見出しも収録されており、これを利用して長単位による対応表を作成することも考えられるが、長単位の異なり語数は短単位より多くなるため、分類語彙表 DB 収録の見出しだけではコーパスへの網羅的語義情報付与には対処できない。今後別の方法の検討が必要である。

７．おわりに

以上、大規模コーパスへの網羅的・系統的な語義情報付与を目的とした、UniDic・分類語彙表見出し対応表データの構築について延べた。本対応表は 2017 年中に公開予定である。本対応表を用いた BCCWJ への語義情報付与作業は既に始まっており、６．にあげた多義語の語義選択や長単位への語義情報付与といった課題に対する検討も実作業を通じて行われつつある（加藤ほか 2017 予定)。今後、本対応表を用いて、網羅的・系統的に語義情報が

10 分類語彙表 DB に積極的に収録されない固有名詞・助詞・助動詞・記号類を除いた数。

付与されたコーパスの構築が進展することが期待される。

謝辞

本研究は、科研費特定領域研究「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」（18061008）および国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」、国立国語研究所言語変化研究領域共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の成果の一部である。

参考文献

小木曽智信・小町守・松本裕治（2013）「歴史的日本語資料を対象とした形態素解析」『自然言語処理』20(5), pp.727-748

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』国立国語研究所（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）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）（下）』（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）

加藤祥・浅原正幸・山崎誠（2017 予定）「『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行」『言語資源活用ワークショップ 2016 予稿集』

国立国語研究所（2004）『分類語彙表増補改訂版データベース』（ver.1.0）
http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

国立国語研究所（編）（1964）『分類語彙表』秀英出版

国立国語研究所（編）（2004）『分類語彙表増補改訂版』大日本図書

伝康晴・峯松信明・小木曽智信・内本清貴・小椋秀樹・小磯花絵・山田篤（2007）「コーパス日本語学のための言語資源―形態素解析用電子化辞書の開発とその応用」『日本語科学』22, pp.101-123

宮島達夫・石井久雄・安部清哉・鈴木泰（編）（2014）『日本古典対照分類語彙表』笠間書院

『名大会話コーパス』の比較に基づく 教室談話における「中途終了型発話」の特徴

矢田　真菜（東京学芸大学大学院教育学研究科・大学院生）

Characteristics of “Ellipsis of sentence endings” of classroom discourse based on the comparison of Meidai Dialogue Corpus

Mana Yada　(Graduate school of Education ,Tokyo Gakugei University)

要旨

教室談話[1]における「中途終了型発話」の特徴を明らかにすることを目的とし,『名大会話コーパス』による日常会話と,授業を書き起こした教室談話を比較した。「中途終了型発話」とは,最後まで言い切らない発話末形式のことである。分析にあたっては「ポライトネス理論」に基づき,どのようにフェイスへの配慮が行われているかに着目した。結論として,以下のことがいえた。(1)日常会話の二者間会話では発話権が均等に分布したのに対し,教室談話では教師の発話権が多く分布したことから,教師の発話権の多さが,教室における教師の権力性を表していると考えられる。(2)日常会話よりも教室談話のほうが「中途終了型発話」の生起割合が多く,「中途終了型発話」がフェイスへの配慮から生起することをふまえると,教室談話は日常会話よりもフェイスへの配慮が尊重されていると考えられる。

１．はじめに

研究の指標として、(1)話者の関係と発話数の関係,(2)発話末形式の生起割合,(3)選定語の談話機能,(4)伝達の失敗と補償行動を挙げ、これらの項目について『名大会話コーパス』による日常会話と,授業の録音データによる教室談話とを比較し,教室談話の特徴を分析した。

２．本稿における「中途終了型発話」の定義と理論の枠組み

２.１　「中途終了型発話」の定義

話し言葉の発話末形式に着目すると,最後まで言い切る形式と,言い切らない形式とがある。後者については,これまで「終助詞的な用法」,「言いさし」,「中途終了型発話」などと呼称されてきた。本稿では,先行研究を整理した上で,「中途終了型発話」と呼称することとする。これは,楠本(2015)が「分析前の段階で『言いさし』という意味解釈が生じるような言い方は避ける」としたことと,話し言葉の研究で「ポライトネス理論」や配慮行動に着目した宇佐美(1995),伊集院(2004),三牧(2015)らが「中途終了型発話」と呼称していることによる。

また,「中途終了型発話」の定義については,「発話を最後まで言い切らずに,従属節で終了しているが,意味的に完結している発話」とし,形式面と機能面の両基準を満たすものを本稿における「中途終了型発話」として認定することとした。形式面については,宇佐美(1995),三原(1995),荻原(2015)ほか多くの先行研究で用いられている「最後まで言い切らず」,「従属節で終了する」ことを基準とした。機能面については,あくまで話し手からみて,意味的に完結していたか否かを基準とした。伊集院(2004),朴(2010)は「情報の伝達が終了している」ものとしているが,それらが伝達されたかどうかは聞き手に委ねられる。本稿では,伝達がされなかった場合はその要因や補償行動を分析するため,伝達が終了したか否かに関わらず,話し手からみて意味的に完結した発話であれば,考察の対象とすることとした。

---

[1] 本稿での「教室談話」は,日本語教育の領域ではなく,学齢期の児童生徒の学校で教室において行われる談話のことを指す。さらに,「教室談話」のうち,授業場面を分析の対象とする。

２.２　理論の枠組み

本稿では,分析にあたり理論的な枠組みとして,Leech(1983)およびBrown & Levinson(1978)の「ポライトネス理論」を用いることとする。滝浦(2008)の「語用論的な発想の基本を押さえるためには,グライスの理論まで立ち返る必要がある」という指摘から,Grice(1989)の「協調の原理」,Sperber & Wilson(1995)の「関連性理論」を整理した結果,前者は聞き手の,後者は話し手の視点が欠けていることがわかった。話し手,聞き手双方の視点で人間関係に着目した理論が,Leech(1983)およびBrown & Levinson(1978)の「ポライトネス理論」である。山岡ほか(2010)は,Leech(1983)の「ポライトネスの原理」とBrown & Levinson(1978)の「ポライトネス理論」は「相補的な関係」にあると述べており,本稿でも同様に位置づけることとする。また,山岡ほか(2010)によると,「ポライトネス理論」は「対人関係をよりよいものにしたいという高度な配慮をもってなされる言語行動の原理」であることから,相互作用により成立する会話を分析する本稿においても,適当な理論であると結論づけられた。

２.３　「ポライトネス理論」概説

Brown & Levinson(1978)の「ポライトネス理論」は2つのフェイスと,Face-Threatening-Acts(以下,FTAと呼称する)が重要な概念である。まず,フェイスには,他者によく思われたい,友好的に思われたいという願望であるポジティブ・フェイスと,押しつけられたくない,自由を阻害されたくないというネガティブ・フェイスとがある。金杉(2008)はフェイスの概念について,「人間の基本的な『欲求』を土台にしていることに着目すべきである」と述べている。井上(2010)の例では,相手を家に招く場面を想定し,話し手による「さあ,入って」という発話について,聞き手が友好的に思っている場合はポジティブ・フェイスが満たされ,自由を阻害されたくないと思っている場合はネガティブ・フェイスの侵害になると説明されている。つまり,同一の話者,場面であっても,聞き手がどちらの欲求に基づいているかにより,2つのフェイスの,どちらの可能性もあるということである。

またFTAは,「フェイス脅かし行為」と邦訳され,相手のフェイスを脅かす可能性のある行為のことを指す。相手に対してなにかしらの行為をすることは,すべてFTAにつながりうることから,相手との関係に応じ,適切なFTA行動選択をすることが求められる。

滝浦(2008)は,これら「ポライトネス理論」の「対人配慮」と,情報の「伝達の効率性」の関係について,次のように整理している。滝浦(2008,p.28)は,「情報伝達の効率性を最大化すると,対人配慮は反比例的に最小化」され,「対人配慮を大きくしようとすれば,情報の伝達性を犠牲にしなければならない」と述べている。

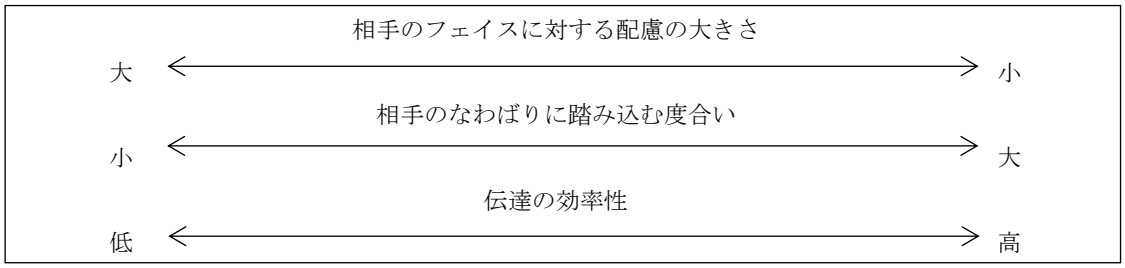


図１　滝浦(2008,p.28)による「対人配慮」と「伝達の効率性」の関係

２.４　「ポライトネス理論」による「中途終了型発話」の位置づけ

杉山(2001)は,「中途終了型発話」の生起要因として「文脈的要因」,「場面的要因」,「文

化的慣習」,「心理的要因」を挙げている。このうち,「心理的要因」に着目した宇佐美(1995),陳(2000)は,「中途終了型発話」により言明を避けることで,相手への配慮がなされると述べている。つまり,「ポライトネス理論」で「中途終了型発話」の談話機能を捉えると,相手への配慮がなされるときに生起し,相互作用の成立に寄与しているといえる。

３．日常会話と教室談話の比較

『名大会話コーパス』を日常会話の,授業の録音データを教室談話のデータとして用い,日常会話と教室談話の比較を行った。指標とする項目は,(1)話者の関係と発話数の関係,(2)発話末形式の生起割合,(3)選定語の談話機能,(4)伝達の失敗と補償行動である。データの分析にあたっては,発話の単位ごとに発話末形式を判断し,「中途終了型発話」を取り出した。本稿における発話の単位は,藤江(2000),宇佐美(2005)などが「話者交代」と「間」に基づいていることから,「話者交代」を原則とし,同一の話者による発話でも,「間」がある場合には発話の切れ目としてみなすこととした。

次に,それぞれのデータの収集方法について述べる。日常会話のデータである『名大会話コーパス』の全 129 件のデータのうち,①話者の出身が関東および中部地方である,②二者間会話である,③会話の時間が 30 分程度であるという条件のすべてを満たす 10 件を抽出した。データの詳細は表 1 の通りである。データ番号、話者の情報は『名大コーパス』に記載されていた表記に基づく。教室談話は,4 人の教師による授業を 1 コマずつ録音した全 4 コマ分の授業を,宇佐美(2011)の「基本的な文字化の原則(Basic Transcription Systems for Japanese: BTSJ)2011 年版」に従って文字化した。データの詳細は表 2 の通りである。

表 1　日常会話のデータ一覧

データ番号	話者 A	話者 B
data30	F044 (女性・90 代)	F126 (女性 60 代)
data34	F144 (女性・40 代)	F148 (女性・40 代)
data54	F074 (女性・20 代)	F087 (女性・20 代)
data65	F114 (女性・10 代)	F147 (女性・20 代)
data66	F114 (女性・10 代)	F137 (女性・20 代)
data67	F045 (女性・20 代)	F160 (女性・20 代)
data68	F119 (女性・20 代)	F160 (女性・20 代)
data71	F062 (女性・20 代)	F161 (女性・20 代)
data93	M002 (男性・20 代)	M034 (男性・20 代)
data129	F003 (女性・80 代)	F007 (女性・50 代)

表 2　教室談話のデータ一覧

データ番号	教師の情報	学校の情報	教材
A	教師 A (女性・20 代)	私立 中学3年生	「俳句十五句」 (学校図書)
B	教師 B (女性・20 代)	公立 中学1年生	メディア・ リテラシー (自作教材)
C	教師 C (女性・30 代)	私立 高校3年生	「兵隊宿」 竹西寛子 (明治書院)
D	教師 D (女性・40 代)	公立 中学3年生	「故郷」魯迅 (光村図書)

３．１　話者の関係と発話数の関係

話者の関係と発話数の関係を分析した。発話の単位に従って発話数を算出後,1 分間あたりの発話数に換算すると,日常会話は 10.2 回,教室談話は 15.8 回となった。

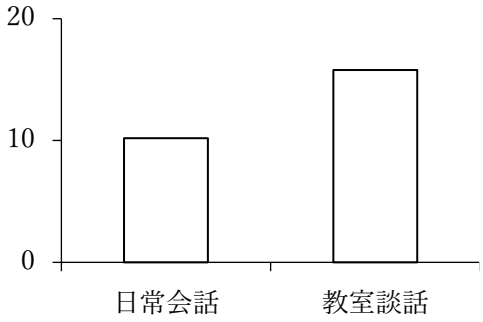


図２　1 分間当たりの発話数の平均(回)

日常会話に比べて教室談話は 1 分間あたりの発話数が多く,t 検定の結果,有意な差があった(p=.002,p<0.01)。したがって,日常会話よりも教室談話は発話数が多い傾向にあるといえる。その他,発話数に関わる要素としては,年齢差が挙げられた。日常会話において,話者間に年齢差がある data30(8.7 回),data129(8.3 回)は有意な差があり(p=.009,p<0.01),発話数が少ない傾向にあることがわかった。

次に,話者ごとに発話数を算出した,総発話数に占める発話権の分布を示す。日常会話はデータにより最大 0.6 ポイントの差はあるが,およそ 50%ずつ均等に発話権が分布した。教室談話は教師に 77.8%,生徒に 22.2%発話権が分布する結果となった。

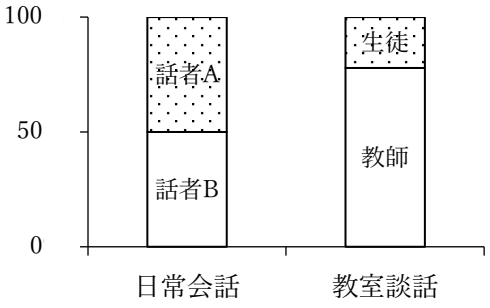


図３　話者別の発話権の分布(%)

日常会話では均等に発話権が分布しているのに対し,教室談話では教師に発話権が偏っていることがわかる。教師の発話権の多さは,教室において教師が持つ権力性に結びついており,松下(2007)はこうした権力や権限を「委譲」することで子どもの主体的な学びに繋がることを示唆している。

３．２　発話末形式の生起割合

発話の単位に従って認定したそれぞれの発話について,発話末形式の分類を行った。感動詞などを除外し,「言い切り」か「中途終了型発話」に分類した結果,日常会話は「言い切り」77.1%,「中途終了型発話」22.9%の生起割合であり,教室談話は「言い切り」50.8%,「中途終了型発話」49.1%の生起割合であった。

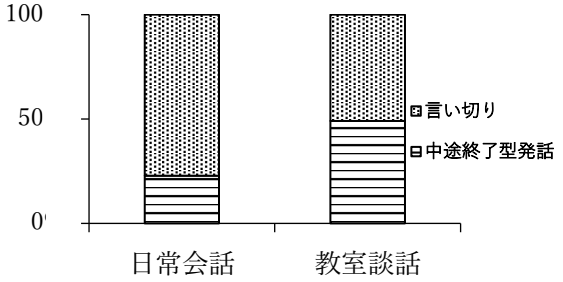


図４　発話末形式の生起割合(%)

発話末形式を比較すると,日常会話に比べて教室談話は「中途終了型発話」の生起割合が多い結果となった。このことと,２．４で述べたことを併せて考察すると,教室談話では日常

会話よりもフェイスへの配慮が尊重されているといえるのではないだろうか。図 1 では,「対人配慮」を最大化すると,「伝達の効率性」が最小化されるという関係性であったが,教室談話では発話末形式に関わらず,「伝達の効率性」は常に意識されていると考えられる。つまり,教室談話ではクラス全員に発話内容を確実に伝達するため,わかりやすい語の選択や,板書により視覚的に情報を補足するなどの工夫がなされ,常に最大限,「伝達の効率性」が尊重されている状態にあるということである。この状態を前提として,「中途終了型発話」によって,フェイスへの配慮が日常会話よりも行われているということになる。教室談話の特殊性については,藤江(2007)はじめ多くの研究で指摘されているが,このひとつの根拠として,「ポライトネス理論」の観点から日常会話との違いを指摘することができる。

３．３　選定語の談話機能

まず,日常会話,教室談話のそれぞれで生起割合が多い 10 語を,表 3 に示す。

表 3　生起割合の上位 10 語

生起順位	日常会話	教室談話
1	て　(18.2%)	名詞　(41.4%)
2	けど　(10.7%)	て　(6.3%)
3	で　(7.3%)	は　(5.6%)
4	から　(7.0%)	から　(5.3%)
5	名詞[2]　(6.7%)	けど　(5.0%)
6	って　(6.4%)	で　(4.0%)
7	は　(5.2%)	接続詞[3]　(3.7%)
8	とか　(4.6%)	ので　(3.0%)
9	が　(4.4%)	と　(3.0%)
10	し　(3.8%)	副詞[4]　(2.7%)

これら無数にある「中途終了型発話」のうち,本稿で考察する語として,先行研究で 2 件以上取り上げられているもの,かつ,日常会話と教室談話で生起割合が多かった 7 語を選定した。選定語は【けど,から,ので,て,は,が,名詞】である。

それぞれの語をカテゴリーで分類し,談話機能を分析した結果,日常会話,教室談話ともに直接的な表現を回避し,フェイスへの配慮を行う傾向があることがわかった。ただし,フェイス侵害のリスクの高いものが生起する場合は,日常会話と教室談話で異なる傾向がみられた。

この点について,教室談話において最も生起割合の多かった「名詞」を例に説明する。

３．３．１　「名詞」

以下の表 4 は,新屋(2014)による「名詞」の談話機能のカテゴリーに従い,本稿で生起した「名詞」の「中途終了型発話」を分類したものである。それぞれについては,①「話し手の知識を提供する」,②「話し手の感情・感覚や意志を表わす」,③「聞き手の行為を要求する命令,勧誘,依頼」,④「聞き手に情報の提示を要求する」ものと説明されている。

表 4　「名詞」の談話機能

番号	談話機能	日常会話(%)	教室談話(%)
①	演述型	56.2	57.6
②	情意表出型/感嘆型	20.8	19.2
③	訴え型	0.0	9.2
④	疑問型	22.6	14.0

[2] 「名詞」,「副詞」,「接続詞」については,異なり語が多いこと,重複する語がなかったことから,複数の語をまとめて品詞の単位で扱うこととした。

[3] 同上

[4] 同上

表 4 より,日常会話,教室談話双方に共通する傾向として,①が多いことが挙げられる。これは,「話し手の知識を提供する」ことのフェイス侵害のリスクの低さが要因であると考えられる。聞き手が知識をおしつけられたくない場合,ネガティブ・フェイスの侵害となることも考えられるが,話し手の知識が提供されることによって聞き手の認識改変を促すことが目的ではないため,フェイス侵害のリスクは低いと考えられる。以下,例 1 は日常会話で①に分類されたものである。

例 1　〈日常会話〉(F114 と F137 が日程の相談をしている場面)

F137　月火水がバイトだから,木曜日。

F114　木だめなんだー,あたし。

【出典】『名大会話コーパス』data66

日常会話と教室談話で異なる傾向となったのは,③と④である。これらは①よりも聞き手に対してなんらかの要求を行っていることから,フェイス侵害のリスクが高いといえる。まず,③は,日常会話では生起せず,教室談話では 9.2%生起した。「聞き手の行為を要求する命令,勧誘,依頼」が日常会話では生起しなかった要因としては,フェイス侵害のリスクのためと,このような要求は「言い切り」で発話されているのではないかと考えられる。前述の図 1 より,「対人配慮」と「伝達の効率性」は反比例する関係にあることがわかっている。このことと,教室談話のほうが日常会話よりもフェイスへの配慮が行われていることを併せて考察すると,日常会話では相手への要求をする際は「伝達の効率性」を優先させ,「言い切り」で明確に伝達しているのではないかと考えられる。「中途終了型発話」は聞き手に解釈が委ねられるため,確実に要求する内容を伝達したい場合は,最後まで言い切るという行動選択をしたほうが確実である。以下,例 2 は教室談話で③に分類されたものである。

例 2　〈教室談話〉(教師が生徒に指示を問いを解くよう指示している場面)

T　ちゃんとしたお座敷で眠らせてあげたいなあとも思っているわけ。

T　で、これを踏まえた上でー,この問いをやってみて,246 ページの,問い。

【出典】教室談話データ C

一方で④は教室談話よりも日常会話のほうが,生起割合が 8.6 ポイント高い結果となった。「聞き手に情報の提示を要求する」ことは,やはりフェイス侵害のリスクが高いものである。それにも関わらず,日常会話のほうが生起割合が多かった要因としては,聞き手に情報の提示を要求しなければ,会話の進行自体に支障をきたすことから,フェイス侵害を犯してでも要求することを優先させたためであると考えられる。これには,日常会話が 1 対 1 の二者間会話であるのに対し,教室談話は話者の役割としては教師と生徒の二者であっても,生徒が複数人であることが大きく影響していると考えられる。日常会話では情報が得られなければ伝達の失敗につながるために聞き手に情報を要求をするが,教室談話では生徒全員が情報を得られていないような場合を除き,要求を行わなくても会話が進行してしまうことが起こりうるのではないだろうか。特に,生徒が要求を行うとき,他の生徒は分かっているかもしれないという不安や,授業の進行を止めることに抵抗がある場合,自分にとっては必要な場面でも,要求を行わないことが考えられる。以下,例 3 は日常会話で④に分類されたものである。

例 3　〈日常会話〉(F007 が F003 に昔住んでいた場所を質問している場面)

F007　でー,何,東京の阿佐ヶ谷？

F003　うん,阿佐ヶ谷ね。

【出典】『名大会話コーパス』data129

以上,選定語のうち「名詞」を取り上げて分析の結果を述べてきた。日常会話,教室談話に

共通する傾向としては,フェイス侵害のリスクの高いものは避けられることが挙げられ,異なる傾向としては,フェイス侵害の高いもののうち,「中途終了型発話」によって聞き手に行動を要求する機会は教室談話のほうが多く,聞き手に情報の提示を要求する機会は日常会話のほうが多いということである。

ただし,教室談話では,原則としてフェイスへの配慮が尊重されてはいるが,教育的な配慮が優先される場合がある。例えば,生徒に注意する場面では,確実にフェイスへの侵害が生じるが,フェイス侵害を犯してでも「中途終了型発話」による言明の回避を行わずに,「言い切り」で明確に指導するということが選択される場合がある。つまり,教室談話では特殊な FTA 行動選択がされうるということである。

３．４　伝達の失敗と補償行動

「中途終了型発話」により,聞き手への情報や意図の伝達の失敗が生じた割合を算出した。失敗の判断は聞き手の反応から行った。結果,日常会話では 2.6%,教室談話では 0.7%の割合で伝達の失敗が生じた。以下,①伝達の失敗の要因,②聞き手の反応,③補償行動について考察していく。まず,①伝達の失敗の要因としては,図 1 の「伝達の効率性」ばかりが尊重された場合が挙げられた。話し手にとっての「伝達の効率性」であるため,聞き手への配慮が欠けると失敗が生じるということである。②聞き手の反応としては,情報の補完を促すものが多く生起した。教室談話では教師が聞き手の場合,情報の補完を促しつつも,生徒のフェイスへ配慮する反応がとられた。③補償行動については,日常会話は 100%の割合で生起したのに対して,教室談話は 78%の割合に留まった。補償行動の発話末形式は,日常会話では「言い切り」19.0%,「中途終了型発話」61.9%,感動詞 14.3%であり,教室談話は「言い切り」0.0%,「中途終了型発話」100%,感動詞 0.0%であったことから,「中途終了型発話」による補償行動が多いことが共通していた。ここから,話し手が伝達の失敗への補償行動として「対人配慮」を尊重したことが,「中途終了型発話」が多く生起した結果に結びついたのではないかと考えられた。つまり,伝達の失敗時には「伝達の効率性」を尊重していたのに対し,その補償行動時には「対人配慮」を尊重するようになったということである。

以下,伝達の失敗から補償行動までの一連の流れを,日常会話,教室談話それぞれの例で示す。どちらの例も,伝達の失敗が「中途終了型発話」により生じ,聞き手が話し手に対して内容の補完を促す反応を示し,「中途終了型発話」による補償行動が行われている。

	日常会話	教室談話
伝達の失敗	F045　なんか,次バイトだし。	T　半夏生。
聞き手の反応	F160　えっ？	T　はん[↑]
補償行動	F045　バイトだしね。	T　半夏生。
	【出典】『名大会話コーパス』data67	【出典】教室談話データ C

４．おわりに

本稿では,教室談話における「中途終了型発話」の特徴を明らかにすることを目的とし,『名大会話コーパス』による日常会話と,教室談話の比較を行った。結論として,以下のことがいえた。(1)日常会話の二者間会話では発話権が均等に分布したのに対し,教室談話では教師の発話権が多く分布したことから,教師の発話権の多さが,教室における教師の権力性を表していると考えられる。(2)日常会話よりも教室談話のほうが「中途終了型発話」の生起割合

が多く,「中途終了型発話」がフェイスへの配慮から生起することをふまえると,教室談話では日常会話よりもフェイスへの配慮が尊重されていると考えられる。

日常会話との比較から,教室談話の特殊性を「中途終了型発話」を指標として考察することができた。今後の課題としては,同一の話者の発話が日常会話と教室談話でどのように変化するのか,授業場面と授業場面以外での教室談話の差異があるのかなど,比較する条件を整理して,詳細に分析していくことが挙げられる。

本稿は,稿者が 2017 年東京学芸大学大学院教育学研究科国語教育専攻修士論文として提出した,「教室談話における『中途終了型発話』の特徴」を一部まとめ直したものである。

文　献

Brown,P. & Levinson,S.(1978) “Politeness:Some Universals Language Usage”,Cambridge University Press.

Grice,G.M.(1989)“Pragmatics and Natural Language Understanding. ”,Lawrence Erlbaum Associates.

Leech,G.N.(1983)“Orinciples of Pragmatics”,Longman.

Sperber, Dan & Wilson, Deirdre (1995) “Relevance: Communication and Cognition(2nd. ed.)”, Blackwell, Oxford.

井上逸兵(2001)「丁寧さ」小泉保編『入門 語用論研究―理論と応用―』研究社, pp.124-140.

伊集院郁子(2004)「母語話者による場面に応じたスピーチスタイルの使い分け:母語場面と接触場面の相違」『社会言語科学』6 巻 ,pp.12-26.

宇佐美まゆみ(1995)「談話レベルから見た敬語使用―スピーチレベルシフト生起の条件と機能」『学苑』662 巻 ,pp.27-42.

荻原稚佳子(2015)「話し手の言いさし使用の実態と聞き手の解釈―会話の目的を基にした推量を中心に―」『日本語学』34 巻 7 号 ,pp.52-64.

金杉高雄(2008)「言語ストラテジーの役割」『太成学院大学紀要』10 巻 ,pp.49-62.

楠本徹也(2015)「中途終了型発話文『～けど』『～ので』の要求・断り行為場面における待遇的談話機能」『東京外国語大学留学生日本語教育センター論集』41 巻 ,pp.47-60

白川博之(2009)『「言いさし文」の研究』,くろしお出版.

新屋映子(2014)『ひつじ研究叢書〈言語編〉第 115 巻 日本語の名詞指向性の研究』,ひつじ書房.

杉本ますよ(2001)「対話番組にみられる『中途終了型発話』―表現形式、生起理由、会話のストラテジー」『別冊論集』3 巻 ,pp.35-53.

滝浦真人(2008)『ポライトネス理論』,研究社.

陳文敏(2000)「日本語母語話者の会話に見られる『中途終了型発話』―表現形式及びその生起の理由」『言葉と文化』創刊号,pp.125-141.

藤江康彦(2000)「教室談話の成立機制に関する社会文化的研究:発話運用の柔軟性をめぐって」広島大学博士論文(未刊行).

藤江康彦(2007)「教室談話の特徴」秋田喜代美編著『改訂版 授業過程と談話分析』,pp.51-69

松下佳代(2007)「非 IRE 型の教室会話における教師の役割―エンパワメントとしての授業」『学びのための教師論』グループ・ディダクティカ編,勁草書房 pp.193-220.

山岡政紀・牧原功・小野正樹(2010)『コミュニケーションと配慮表現―日本語語用論入門―』,明治書院.

関連 URL

宇佐美まゆみ(2011)「基本的な文字化の原則(Basic Transcription Systems for Japanese: BTSJ)2011 年版, http://tufs.ac.jp/ts/personal/usamiken/btsj2011.pdf

『名大会話コーパス』,https://nknet.ninjal.ac.jp/nuc/templates/nuc.html

『多言語母語の日本語学習者横断コーパス』の母語話者データにおけるタスクと産出語彙の関連

小西 円（国立国語研究所日本語教育研究領域）†

Vocabulary Used by Native Speakers in Tasks from the International Corpus of Japanese as a Second Language

Madoka Konishi (National Institute for Japanese Language and Linguistics)

要旨

学習者コーパスを用いた研究は，学習者データと母語話者データを比較することによって行われることが多い。そのため，母語話者データの特徴を把握しておく必要がある。本研究では，『多言語母語の日本語学習者横断コーパス』（I-JAS）の母語話者データのうち，ストーリーテリング（以下、ST）2 種とロールプレイ（以下、RP）2 種を対象に，タスクの異なりが産出語彙にどのような影響を与えるか，その要因は何かについて考察した。考察にはコレスポンデンス分析の結果を用いた。その結果，タスク形態が独話か対話かによって，多くの品詞が異なる分布を示した。また，名詞や動詞は，タスク形態だけでなく，話題によっても分布が異なっていた。ST1 と ST2 は異なる話題を扱ったものとみなすことができ，名詞や動詞に分布の差があるが，RP1 と RP2 は扱う言語機能は異なるものの，話題という点からはほぼ同一のものとみなされ，名詞や動詞に分布の差があまり見られないことがわかった。一方で，感動詞や助詞はタスク形態だけでなく、機能によって分布に差が出る傾向が見られた。

１．はじめに

学習者コーパスを用いた学習者の中間言語研究は，学習者コーパスの構築が進んでいる英語教育において既にさかんに行われている（グレンジャー(編) 2008，石川 2012 他）。そのような研究は，母語話者と学習者の言語産出を比較することによって行われるものも多い。両者の比較を行う場合には，参照資料となる母語話者データの傾向を知る必要がある。そこで本研究では，『多言語母語の日本語学習者横断コーパス』[1]（以下，I-JAS と呼ぶ）の母語話者データを用いて，その言語産出の特徴について分析する。具体的には，独話形式のストーリーテリング（以下，ST と呼ぶ）2 種と，対話形式のロールプレイ 2 種を用いて，タスクの異なりが言語産出に与える影響を，語彙の観点から明らかにする。また，タスクごとに産出語彙に違いがあるとするなら，どのような品詞に違いが表れ，それらは何に影響を受けているかについて考察する。このような点が明らかになることにより，学習者データと母語話者データの比較の精度が高まると考えられる。

２．調査対象と調査方法

本稿で分析の対象とするのは，第 1 次公開データとして公開されている母語話者 15 名の，以下の 4 つのタスクである。

† komadoka@ninjal.ac.jp

[1] I-JAS の詳細は右記を参照のこと。https://ninjal-sakoda.sakura.ne.jp/lsaj/

【調査対象】　ストーリーテリング 1（以下，ST1 と呼ぶ）
ストーリーテリング 2（以下，ST2 と呼ぶ）
ロールプレイ 1（以下，RP1 と呼ぶ）
ロールプレイ 2（以下，RP2 と呼ぶ）

ST1 と ST2 は, 図 1, 図 2 に示すイラストのストーリーを話す独話形式のタスクである。ST1 は「ピクニック」というタイトルで,「朝，ケンとマリはサンドイッチを作りました。」という 1 文目が与えられ，そこに続くストーリーを述べていくタスクである。ST2 は「鍵」というタイトルで,「ケンは，うちの鍵を持っていませんでした。」という 1 文目が与えられ，そこに続くストーリーを述べていく。

RP1 と RP2 は，ロールプレイを行う対話形式のタスクである。調査者が日本料理店の店長役，調査協力者がアルバイト役になって会話を行う。RP1 の指示文(1)，RP2 の指示文(2)からわかるように, RP1 は依頼, RP2 は断りという機能をターゲットとしたタスクである。

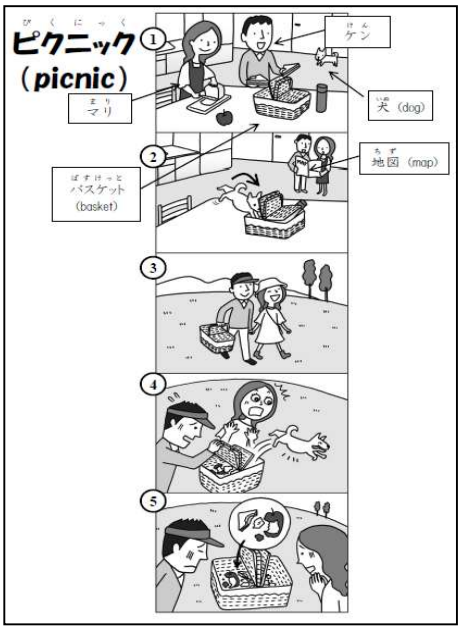


図 1　ST1 のイラスト

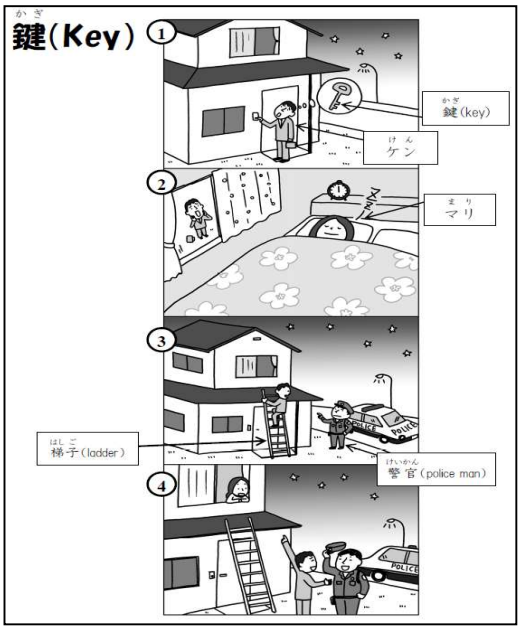


図 2　ST2 のイラスト

(1) あなたは，飲食店でアルバイトをしています。接客スタッフとして注文を取ったり，料理を運んだりしています。勤め始めてからずっと接客の仕事をしてきたので，この仕事にもすっかり慣れ，知り合いのお客さまも増えました。今は，一週間に三日アルバイトをしています。しかし，忙しくなってきたので，一週間に二日に変更したいと思っています。そこで，店長に言って三日から二日に変えてもらうように頼んでください。

(2) あなたは，飲食店でアルバイトをしています。接客スタッフとして注文を取ったり，料理を運んだりしています。店長さんから，「料理を作る人が一人やめたので，来月から料理を作る仕事を担当してほしい」と言われました。しかし, あなたは料理は苦手だし，お客さんと接する仕事がしたいので，この話を断りたいと思いました。店長に，料理の仕事の話をじょうずに断って，今の仕事を続けられるように話してください。

これら 4 つのタスクから,「空白」「記号」「補助記号」「あいづち」「解析困難箇所」「非言

語行動」を除くすべての語彙を品詞ごとに採取した。検索には，I-JAS の検索システムであるコーパス検索アプリケーション中納言[2]の短単位検索を用いた。採取した語の単位は短単位（小椋 2014）で，以下，1 短単位を 1 語と呼ぶ。分析対象となる語は表 1 の通りである。

これらの語を対象にコレスポンデンス分析を行った。コレスポンデンス分析はデータ縮約を行うための計算法である（田畑 2007）。データ表の行や列に含まれる情報を少数の成分（次元）に圧縮し，それらの関係を散布図上付置することで，視覚的なデータの俯瞰を可能にする（石川ほか(編) 2010）。本稿では，タスクを第 1 アイテム，語彙を第 2 アイテムとして分析し，両アイテムの相関を最大にするよう数量化を行う。その結果得られた 2 つの軸（第 1 主成分と第 2 主成分）で散布図を作成する。2 つの軸の解釈を行うことにより，タスクと産出語彙の対応関連や，どのような要素が語彙の分布に影響を与えているかが明らかになると考える。

表 1　分析対象となる語の数

品詞	異なり語	延べ語	品詞	異なり語	延べ語
感動詞	65	757	接尾辞	24	89
形状詞	24	63	代名詞	15	171
形容詞	32	159	動詞	172	1438
助詞	42	3028	副詞	65	379
助動詞	15	1514	名詞	361	1928
接続詞	7	35	連体詞	8	55
接頭辞	4	106	総計	834	9722

３．コレスポンデンス分析の結果

4 つのタスクと分析対象とするすべての語をプロットした図 3 を示す。横軸が第 1 主成分，縦軸が第 2 主成分を示している。横軸の上と縦軸の右の数値が 4 つのタスク（第 1 アイテム）に付与された数値，横軸の下と縦軸の左の数値が語（第 2 アイテム）に付与された数値である。図中で接近する項目は似た性質があることを示し，図中の項目を隔てる距離が大きければ大きいほど，異質性が高いことを示す（田畑 2007）。また，図 3 におけるタスクだけを取り出したものが図 4 である。

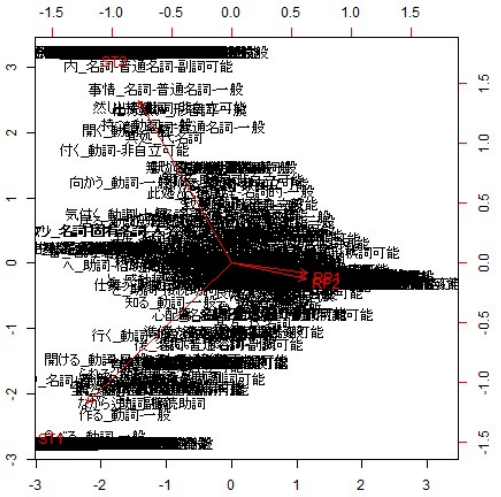


図 3　4 つのタスクとすべての語の対応関係

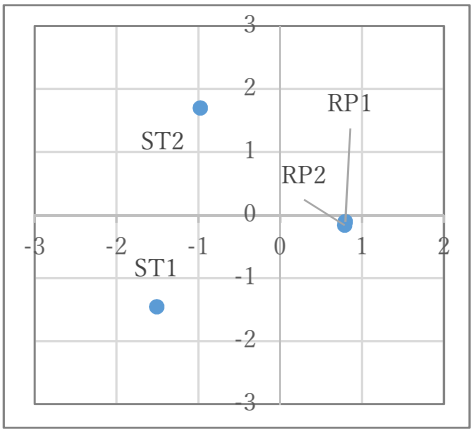


図 4　タスクの散布図

[2] 中納言はバージョン 2.2.0，データは第 1 次公開データの 20160420 版を用いた。

図 3，4 から次のことが分かる。まず，RP1 と RP2 はほぼ重なっており，第 1 主成分においても第 2 主成分においても大きな違いがない。一方，ST1 と ST2 は，第 1 主成分においてはそれほど大きな差がないものの，第 2 主成分で差がある。また，ST 群，RP 群というグループで見た場合，ST 群と RP 群は第 1 主成分において差がある。

これらの図をもとに軸の解釈を行うと，第 1 主成分は，独話か対話かというタスク形態の違いを識別していると考えられる。これは，独話と対話という形態によって，産出語彙に違いが出ることを意味する。一方，第 2 主成分は，ST1 と ST2 を識別するものの，RP1 と RP2 を識別しない。タスクの形態や内容の観点からは，ST だけを識別し，RP を識別しない要素が何であるのか判断ができない。より詳細な分析を行うため，次節では，品詞ごとにコレスポンデンス分析の結果を考察する。

４．品詞ごとの分析

４．１　３種類の分布

品詞ごとにタスクと語の対応関係を見る散布図を作成した結果，すべての語をプロットした図 3，4 とは異なる分布を示す品詞も見られた。分類の結果，タスクと語との対応関係には，以下の 3 種類の分布があることがわかった。

パターン 1：ST 群と RP 群とで分布に差があり，かつ，RP1 と RP2 にも分布の差がある（感動詞，助詞）
パターン 2：ST 群と RP 群とで分布に差があり，かつ，ST1 と ST2 にも分布に差がある（名詞，動詞，助動詞，代名詞）
パターン 3：パターン 1 にもパターン 2 にも属さない（その他の品詞）

図 3，4 と類似するのは，パターン 2 である。パターン 3 には，総出現数が少ない品詞も含まれるため，本研究ではパターン 1 とパターン 2 について分析を行う。その過程において，パターン 1 とパターン 2 の語に分布が生じる要因について考察する。

４．２　パターン１の分析

パターン 1 は，ST 群と RP 群とで分布に差がある語であり，感動詞と助詞がある。まず，感動詞と助詞についてタスクをプロットした図 5，6 を示す。ST 群，RP 群というまとまりで見た場合，両群は第 1 主成分に置いて差がある。また，RP1 と RP2 が第 2 主成分において差がある。

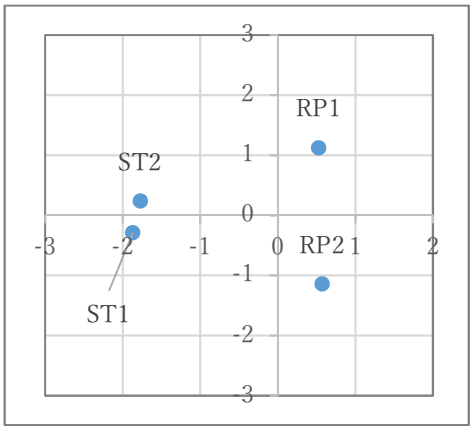


図 5　感動詞の散布図（タスク）

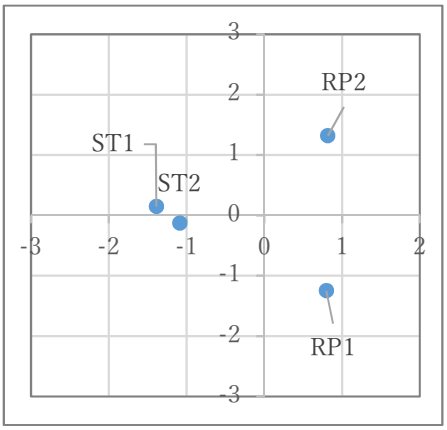


図 6　助詞の散布図（タスク）

それぞれの語を具体的に見ていく。まず，感動詞[3]をプロットした図 7 を示す。ST 群に関連して注目したい語を太字の斜体，RP 群に関連して注目したい語を太字の網掛けにした。ST 群に関連がある語は，「えーと」「えー」などのフィラーが多い（例(3)）。これは，独話形式でストーリーを語る最中に現れる。一方，RP 群では「えーっとですね」「あのですねー」など，「です」を伴って丁寧さを帯びたフィラーも用いられる。また，「はい」「いいえ」などの応答詞もある（例(4)）。これらは，対話性を帯びた感動詞であり，独話形式では通常は現れない[4]。つまり，感動詞は，タスク形態が独話か対話かによって，産出語彙に違いがあると言える。

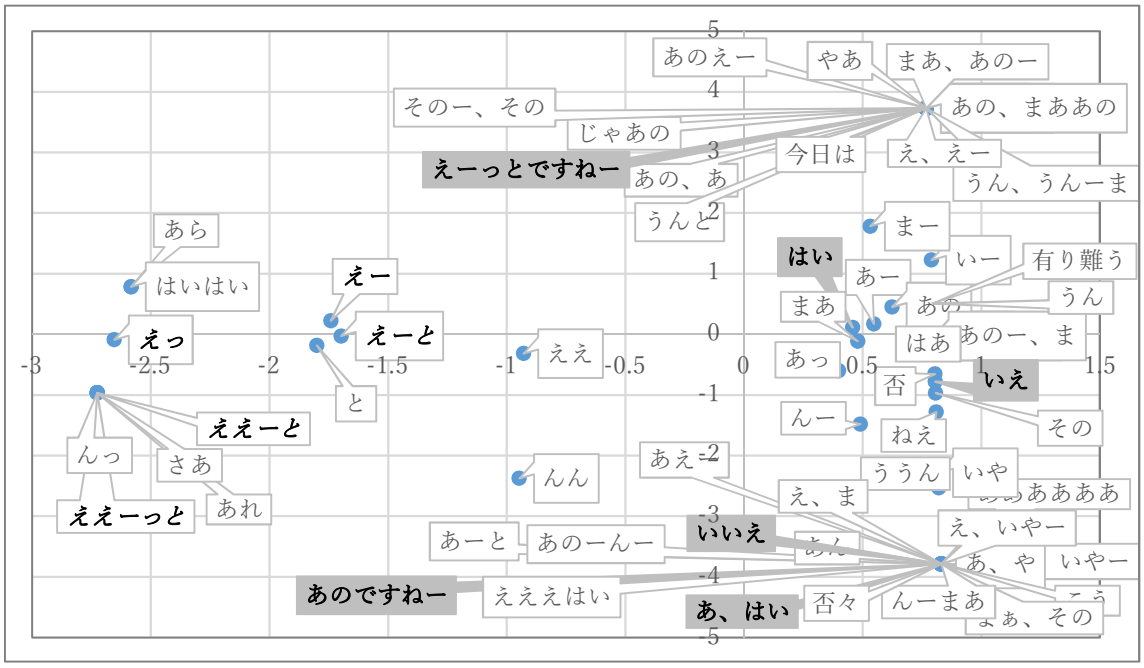


図 7　感動詞の散布図（語）

(3) K：ケンとマリはえー気づかずにバスケットを持ってピクニックに出かけました（JJJ12-ST1）[5]

(4) C：JJ さん，今ちょっと時間ありますー？
K：あ，はいいいですよー（JJJ37-RP2）

次に，助詞をプロットした図 8 を示す。格助詞を太字の斜体，終助詞を太字の網掛けにし

[3] 学習者コーパスである I-JAS には,発音の誤りや統語的に逸脱した発話が多々現れる。そのような発話に対する形態素解析の精度を高めるために,I-JAS 独自のルールに基づいてタグ付与がなされている（迫田(編)2016，迫田ほか 2016）。そのため,タグ付与の対象の 1 つである感動詞は,通常の UniDic の感動詞と異なった単位でひとまとまりの語として切り出されている場合がある（例：まあ,あのー）。

[4] ST にも対話性を帯びた感動詞である「はい」や「さあ」などが用いられることはある。「はい」は,調査冒頭で調査者が調査協力者に調査 ID を確認する箇所で用いられ,「さあ」は ST 内でセリフを述べる箇所で用いられている（例：場所に着き,「さあここでサンドイッチを食べましょう」とバスケットを開けたところ（JJJ35-ST1））。これらは ST 中で部分的に対話性が生じる箇所である。

[5] 例文末尾の「JJJ」で始まる数字は,調査協力者 ID である。また,発話先頭の「C」は調査者,「K」は調査協力者を示す。以下の例文中,〈　〉でくくられた箇所はあいづちである。#は ST における文区切りを示す。

た。ST 群がプロットされる第 1 主成分マイナスの位置には格助詞が目立つ[6]。文を構成するうえで最も基本的な格である「が」「を」や「に」は ST がプロットされる位置に隣接しており，ストーリーテリングで主題提示を行うために頻繁に用いられる係助詞「は」も，ほぼ同じ個所にある。一方，RP 群がプロットされる第 1 主成分プラスの位置には終助詞が目立ち，対話で頻繁に用いられると考えられる「よ」「ね」や，「かな」を構成する「か」「な」が位置している（例(5)(6)）。また，「んです」を構成する準体助詞「の」も，RP 群がプロットされる位置に近い。

ST 群と格助詞の相関が高い理由は，ST がケンやマリを登場人物として，第三者の視点からストーリーを語るタスクであるため，格要素を省略した文を発話しにくいからではないかと考えられる。また，RP 群と終助詞の相関が高い理由も，RP が依頼や断りという交渉を行う対話形式のタスクであることに関連していると考えられる。そのようなタスクでは，適宜助詞の省略や無助詞は起こるが，文末の終助詞は不可欠であり，また，婉曲的に話者の考えを伝える「かなと思う」などの表現も多用される。つまり，助詞の産出も，感動詞と同様に，タスク形態によって影響を受けているといえる。

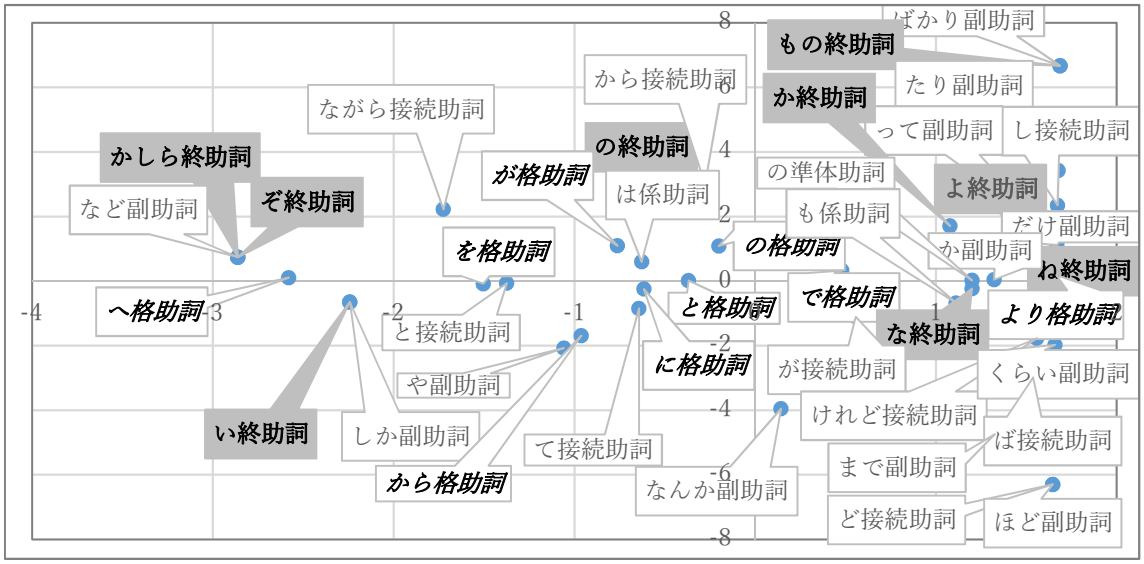


図 8　助詞の散布図（語）

(5) C：ああそうですか
　　K：そうなんですよ（JJJ14-RP2）
(6) K: ちょっとあのー，ま忙しいので，週二日に減らせないかなと思いまして（JJJ03-RP1）

パターン 1 では，RP1 と RP2 が第 2 主成分で識別されている一方で，ST1 と ST2 は差がない。第 2 主成分の軸の解釈は，パターン 2 の分析の過程で行う。

４．３　パターン 2 の分析

パターン 2 の分布は，分析対象の語すべてをプロットしたものと類似している。パターン 1 の分析により，ST 群と RP 群を識別する要素は，独話か対話かというタスク形態の違いであるといってよい。そのため，パターン 2 の分析では，第 2 主成分が識別するものついて考察する。パターン 2 は，ST1 と ST2 では分布が異なるものの，RP1 と RP2 では異ならない語であり，名詞，動詞，助動詞，代名詞がある。本節では，総語数の多い名詞と動

[6] ST 群がプロットされる付近にある終助詞「ぞ」「かしら」「い」は,ST のセリフの中で用いられており，独話形式の中で部分的に対話性を帯びる部分である。これは応答詞の「さあ」とも同じである。

詞を取り上げる。

名詞と動詞について，タスクをプロットしたものが図 9，10 である。また，名詞も動詞も総語数が 1000 を超えるため，抜粋した語をプロットしたものを図 11，12 に示す。各タスクがプロットされる位置とほぼ同じ個所にプロットされる語と，タスクとタスクの間にプロットされる語を中心に抜粋した。

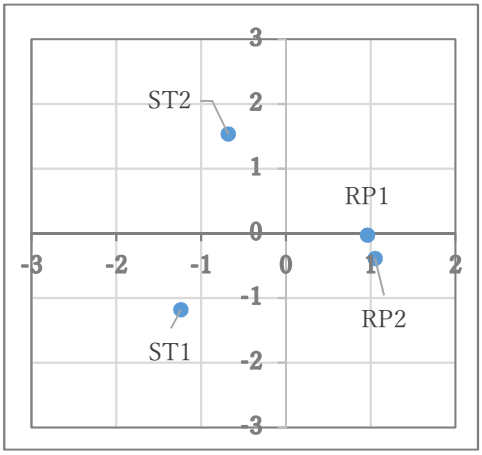


図 9　名詞の散布図（タスク）

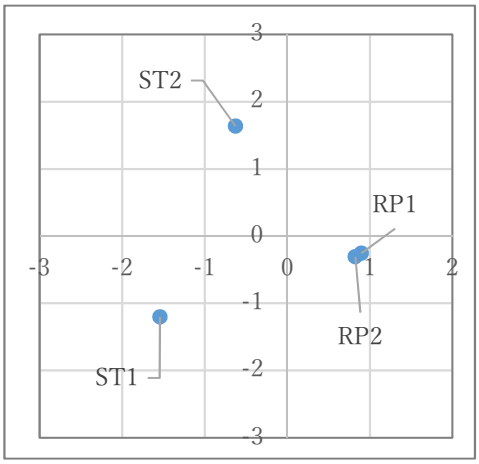


図 10　動詞の散布図（タスク）

名詞をプロットした図 11 を見ると，ST1 とほぼ同じ個所にプロットされている語は，「朝」「地図」「犬」などであり，ST1 の絵や第 1 文に現れる，ストーリーに欠かせない語である。また，ST2 の付近にプロットされる語は，「泥棒」「梯子」「警官」などであり，これも，ST2 のストーリー描写に欠かせない語である。動詞をプロットした図 12 からも同様のことがいえる，ST1 の付近にプロットされる「出掛ける」「飛び出す」「食べる」など，ST2 の付近にプロットされる「呼ぶ」「起きる」「忘れる」などは，それぞれのストーリー描写に欠かすことができない。(7)(8)に動詞の具体例を示す。

(7) その間に，バスケットに，犬が入ってしまいました# 二人はそれに気づいていません# ピクニックに出掛けました（JJJ37-ST1）
(8) ケンはうちの鍵を持っていませんでした#とケンが帰るころにマリは寝ていました（JJJ14-ST2）

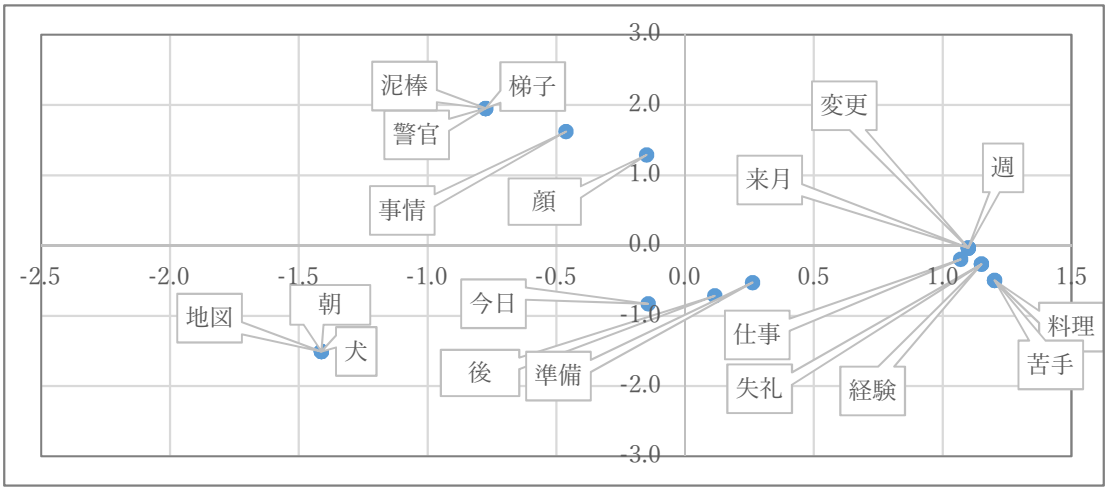


図 11　名詞の散布図（語）

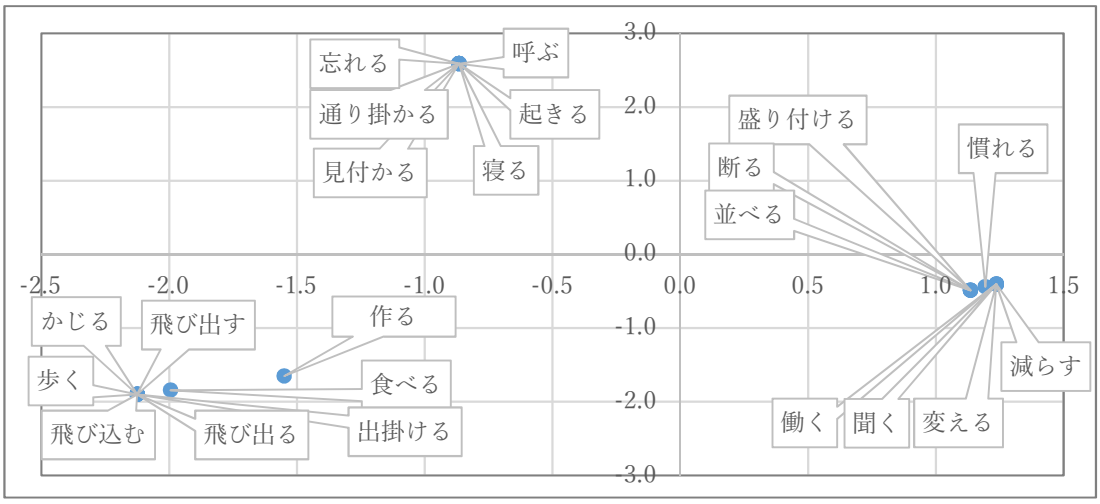


図 12　動詞の散布図（語）

ST1 と ST2 のストーリーは，登場人物のケンとマリは共通しているが，描かれている出来事の内容が異なっている。そのため，ストーリー描写に用いる名詞と動詞が重ならず，これらの産出語彙が異なる分布を示すと考えられる。つまり，ST1 と ST2 の名詞や動詞は，話題によって識別されていると考えられる。

一方，RP1 と RP2 の名詞と動詞の分布はほぼ重なっている。タスクの散布図において RP1 がプロットされる付近には，名詞では「来月」「変更」「週」などがあり，これらは働く日数を減らすという RP1 の話題に関連している（例(9)）。また，RP2 がプロットされる付近には「料理」「仕事」などがあり，これも，厨房で働くことを断る RP2 の話題に関連している（例(10)）。しかし，RP1 と RP2 がプロットされる中間あたりに「仕事」「失礼」「経験」などがあり，これらは両方のタスクで現れる（例(11)(12)）。つまり，依頼と断りという交渉の詳細は違っていても，日本料理店での仕事内容という話題の大枠が同じであるため，名詞や動詞の分布が重なったと考えらえる。つまり，RP1 と RP2 は，異なる機能を対象としているものの，異なる話題を扱っているとは言いにくい。

(9) K : あのー，ちょとあのー，来月から，あの週，二日で，お願いしたいんですが（JJJ57-RP1）
(10) K : 正直ですね，私，料理が苦手でして（JJJ30-RP2）
(11) K:今一週三日ーアルバイトで, 仕事させて頂いてるんですけれども（JJJ37-RP1）
(12) K : えー，ぜひとも，あの，このホールの接客仕事を，このまま，続けさしていただきたい，と思ってます（JJJ50-RP2）

一方で，感動詞と助詞のパターン 1 では，RP1 と RP2 が第 2 主成分において差が見られ、ST1 と ST2 にほとんど差がなかった。つまり、これらの品詞においては、第 2 主成分は依頼と断りという機能の異なりを識別していると考えられる。ST1 と ST2 は、話題は異なるものの、このタスクで行う言語行動が「第三者の視点から独話形式でストーリーを語る」という共通のものであると考えることができる。しかし、言語機能と話題とには重なりもあると思われる。パターン 1 における言語機能と、パターン 2 における話題とが、具体的な語の分布にどのような影響を与えているのかについては、さらに分析が必要である。

５．まとめ

コレスポンデンス分析の結果，I-JAS の母語話者データは，タスクの違いによって産出語

彙に違いがあることが分かった。タスクが独話か対話かによって，多くの品詞が異なる分布を示した。また，名詞や動詞は，タスク形態だけでなく，話題によっても分布が異なっていた。調査対象とした 4 タスクのうち，ST1 と ST2 は異なる話題を扱ったものとみなすことができ，名詞や動詞に分布の差があるが，RP1 と RP2 は扱う言語機能は異なるものの，話題という点からはほぼ同一のものとみなされ，名詞や動詞に分布の差があまり見られなかった。また，感動詞や助詞は，話題よりも機能によって分布が異なる傾向が見られた。

今後の課題としては，話題と機能が具体的な語の分布にどのような影響を与えるのかをさらに分析する必要がある。また、ST や RP と同じ発話データであるインタビューデータを対象とした分析や，作文課題との比較を行いたい。また，このような母語話者データの性質を理解したうえで，学習者データとの比較を行いたいと考えている。

謝　辞

本研究は国立国語研究所のプロジェクト「多文化共生社会における日本語教育研究」および科研費基盤(A)「海外連携による日本語学習者コーパスの構築－研究と構築の有機的な繋がりに基づいてー」による成果『I-JAS』を利用して行われたものである。また，コレスポンデンス分析結果の算出には，国立国語研究所コーパス開発センターの浅原正幸准教授にご支援をいただきました。ここに記して感謝します。

文　献

石川慎一郎 (2012). 『ベーシックコーパス言語学』ひつじ書房.

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』pp.245-264, くろしお出版.

小椋秀樹 (2014) 「形態論情報」山崎誠(編)『講座日本語コーパス 2 書き言葉コーパス 設計と構築』pp.68-88, 朝倉書店.

迫田久美子(編) (2016) 『海外連携による日本語学習者コーパスの構築－研究と構築の有機的なつながりに基づいて―I-JAS 構築に関する最終報告書』（平成 24-27 年度科学研究費助成事業 （基盤研究 A） 課題番号：24251010 研究代表者：迫田久美子）.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子 (2016)「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』6:3, pp.93-110.

シルヴィアン・グレンジャー(編著) 船城道雄・望月通子(監訳) (2008) 『英語学習者コーパス入門 SLA とコーパス言語学の出会い』研究社.

田畑智司 (2007) 「Mining Linguistic Variation in the Inaugural Addresses of US Presidents －米国歴代大統領の就任演説に見る言語変異：R によるテキストマイニングー」『日本行動計量学会大会発表論文抄録集』35, pp.79-82.

関連 URL

多言語母語の日本語学習者横断コーパス（I-JAS） https://ninjal-sakoda.sakura.ne.jp/lsa

コーパス検索アプリケーション『中納言』 https://chunagon.ninjal.ac.jp/

『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行

加藤 祥（国立国語研究所コーパス開発センター）†
浅原 正幸（国立国語研究所コーパス開発センター）
山崎 誠（国立国語研究所研究系言語変化研究領域）

Trial Annotation of 'Word List by Semantic Principles' information on 'Balanced Corpus of Contemporary Written Japanese'

Sachi Kato (National Institute for Japanese Language and Linguistics)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

『現代日本語書き言葉均衡コーパス（BCCWJ）』に分類語彙表番号を付与する作業を開始した。『分類語彙表増補改訂版』（2004）の分類語彙表番号を，短単位と長単位のそれぞれにアノテーションする。作業にあたり，人手で UniDic 語彙素 ID に対応させたデータ（近藤・田中，2017）を用い，該当可能性のある番号を列挙する。作業者は，該当する意味分類が選択可能であれば選択し，選択できない場合や対応のない場合には，新たに適切な番号を付与する。本発表では，番号付与作業基準と作業状況，作業結果を用いた調査例を報告する。

１．はじめに

類語の調査はもちろん，比喩をはじめとする表層的な表現と意味の差を研究する際など，意味的な情報の付与されたコーパスが有用なリソースとなる。また自然言語処理の分野では，語義曖昧性解消のタスクの学習・評価データとして様々な語義タグつきデータが整備されてきた。日本語では，古くは新聞記事を対象とした EDR コーパスや RWCP コーパスなどが国語辞典に基づく語義タグを付与していた。また，代表性を持つコーパスとして，『現代日本語書き言葉均衡コーパス』(Maekawa et al. 2014)（以下　BCCWJ）の一部に対しても，岩波国語辞典の語義が付与され，SemEval-2010 Japanese WSD Task (Okumura et al. 2011) では，日本語の語義曖昧性解消の基礎データとして用いられてきた。また，シソーラスに基づくデータとして，日本語ワードネットに基づく語義タグ付きコーパス(Bond et al. 2012) が整備されている。このコーパスは，英語データを翻訳したものであり，代表性をもつ自然な日本語コーパスに対する，シソーラスに基づく語義タグつきデータは管見の限りない。

国立国語研究所では、BCCWJ コアデータに『分類語彙表増補改訂版』（2004）の分類語彙表番号を悉皆付与する作業に着手した。現在進めているアノテーション基準・作業状況と作業結果を用いた感情表現の分析について報告する。

† yasuda-s(α)ninjal.ac.jp

２．アノテーション

２.１　概要

アノテーション作業対象として，コアデータに含まれる新聞サンプル 54 ファイル（部分集合 A：PN(A)）から，アノテーション優先順位に基づき順次作業に着手した。分類語彙表番号を人手で UniDic 語彙素 ID（小木曽・中村, 2014）に対応させたデータ（近藤・田中, 2017）により，BCCWJ の言語単位（短単位・長単位）に対応可能性のある分類語彙表番号を列挙したうえで，人手で正しい語義を選択する作業を進めている。本付与作業にあたっては，分類語彙表の 5 桁目までの番号を付与する。（例は表 1。図 1 における「3.1010 こそあど」の分類部分が該当する。）

表 1　分類番号の構造（例：この（分類番号：3.1010））

類	部門	中項目	分類項目
相（3）	関係（.1）	真偽（.10）	こそあど（.1010）

アノテーション作業は，短単位と長単位のそれぞれについて行う。列挙された分類語彙表番号の選択肢から，該当する意味分類が選択可能であれば選択し，選択できない場合や，語彙素に対応する分類語彙表番号がない場合には，新たに適切な番号を付与する。以下では，それぞれの単位のアノテーション作業基準について示す。

3 相 の 類

3.1 抽象的関係

3.10 真偽

3.1010 こそあど・他

01 この こんな こういう こうした
かかる こう かく かよう
こうこう かくかく
しかじか このまま
かくのごとく／のごとき このとおり
02 その そんな そういう そうした さる しかる
そう さ さよう
事ほどさように しかく

図 1　分類語彙表（部分例）

２.２　短単位に対する分類語彙表番号アノテーション

2.2.1　概要

機能語を除く短単位に分類語彙表番号を付与する。分類語彙表番号の付与される短単位の機能語は助動詞と助詞の一部に限られるためである。分類語彙表番号が付与される機能語の内訳を表 2 に示す。頻度は BCCWJ PN (A)サンプルのものである。

語彙素に対応して列挙された番号（曖昧性：1～11 種類）がある場合，作業者は該当番号を選択する（図 2）。

短単位	品詞	記入欄	分類語彙表番号（選択肢）		
研究	名詞-普通名詞-サ変可能		1.3065:体-活動-心	-研究・試験・調査・検査など	
費	接尾辞-名詞的-一般	1.3721			
を	助詞-格助詞				
受け入れる	動詞-一般		2.3430:用-活動-行	2.3532:用-活動-交	2.3770:用-活

図 2　番号付与作業例

新聞サンプル 54 ファイル（部分集合 A : PN (A)）における短単位数は表 3 の通りである。すなわち, 短単位 56,922 のうち, アノテーション対象となり得る自立語は 33,725 語（59.2%）あり，そのうち UniDic-分類語彙表データと語彙素番号がマッチした 28,696 語（50.4%）について，選択可能な番号が列挙されている。

表 2　分類語彙表番号が付与される機能語

頻度	語彙素番号	語彙素	品詞
261	40741	れる	助動詞
214	27905	など	助詞-副助詞
87	35891	まで	助詞-副助詞
62	39787	られる	助動詞
49	20355	せる	助動詞
47	21652	たい	助動詞
41	23122	だけ	助詞-副助詞
26	22727	たり	助詞-副助詞
8	10403	くらい	助詞-副助詞
7	34770	ほど	助詞-副助詞
6	30577	ばかり	助詞-副助詞
4	19641	ずつ	助詞-副助詞
2	29213	のみ	助詞-副助詞
2	24320	つ	助詞-副助詞
1	14185	させる	助動詞

表 3　BCCWJ PN(A)集合の短単位内訳

PN(A)短単位	のべ	56922
	機能語	23197
	自立語	33725
UniDic-分類語彙表データにマッチしたもの	全て	29513
	機能語	817
	自立語	28696

また，UniDic-分類語彙表データにマッチした自立語の，選択肢数（分類番号の曖昧性）を表 4 に示す。複数選択肢の列挙された短単位（曖昧性 2 以上）は 12,857（44.8%）ある。なお，曖昧性が 8 となる短単位数の頻出はサ変動詞「する」の頻度の影響による。短単位の番号付与にあたっては，最小限の文脈に依拠した意味とし，比喩的・慣用的な表現などは語源的な意味とする。内容に即した意味は，長単位で対応する。

「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」のような品詞の語については，体（1.で始まる分類語彙表番号）・相（3.で始まる分類語彙表番号）のどちらとも読み取れるが，BCCWJ コアに付与された人手による「名詞」「形状詞」などの用法情報に従う。

表 4　分類番号の曖昧性（BCCWJ PN(A) サンプル）

曖昧性	短単位数	曖昧性	短単位数
1	16656	6	237
2	7479	7	134
3	2621	8	1253
4	826	9	49
5	237	10	21

2.2.2　UniDic-分類語彙表対応のない場合

列挙された選択肢に，文脈上適切な番号がないと判断される場合は，新たな番号を付与する。新たに番号を付与する場合は，『分類語彙表増補改訂版』を参照し，適切な意味分類を検討する。UniDic の語彙素に対応する番号がなく，そもそも選択する番号のない場合も，分類語彙表の意味分類を確認し，適切な番号を付与する。

UniDic の語彙素に対応する番号がない例としては，未知語，固有名詞，略語などがある。未知語には「ロック」「カム」「トゥゲザー」のような外来語も多く含まれるが，それぞれ外来語の意味に相当する意味分類を選択し，分類語彙表番号として付与する。

なお，用法によっては，分類語彙表に既存の分類番号がない場合がある。その場合，分類語彙表に存在しない番号を新設して付与することもあり得る。

固有名詞

人名についてはアノテーション対象外とするが，地名や普通名詞を含む「名古屋タワープラザホール」「岡山ホテル」「阪急グランドビル」のような固有名詞については，それぞれ短単位ごとの意味分類が可能と考え，「名古屋タワープラザホール」であれば，「名古屋」「タワー」「プラザ」「ホール」のそれぞれに分類語彙表番号を付与する。

略語・掛詞等

略語についても，元の語形を考慮し，該当する分類語彙表番号を付与する。但し，「厚労」「自民」のように複数語義の組み合わせが一短単位となっている場合もある。このような場合は，「厚労」は「厚生」「労働」，「自民」は「自由」「民主」のそれぞれの短単位に相当する複数の分類語彙表番号を付与する。掛詞やダジャレなど，一短単位について複数の意味が読み取れる場合にも，複数の意味について分類語彙表番号を付与する。

その他

「一個」「一口」のように短単位での登録がある語は，その単位での分類語彙表番号候補がある場合，文脈上「一」「口」や「一」「個」が別の短単位と読むことが適切にも関わらず，「一個」「一口」を一短単位として選択肢（番号の候補）が挙がる。このような場合は，「一」「口」を一短単位と判断し，各々に分類語彙表番号を付与する。また，副詞用法の語であるが分類語彙表番号に体の類しかない場合には，対応する相の番号を新たに付与する。

２.３　長単位アノテーション

短単位と同様に，長単位についても分類語彙表番号を付与する。短単位作業時に，対応した長単位があればマークを表示し，長単位作業のあることを示している。

長単位に対応して列挙された選択可能な番号（1～11 種類）がある場合は，該当する番号を選択する。文脈上，適切な番号がないと判断される場合や，語彙素番号に対応する番号がない場合は，同様に適切な意味分類を行い，新たに番号を付与する。

「ていく」「てくる」をはじめ，「にとって」など，助動詞扱いとなるが短単位と異なる意味分類となる場合などは，機能語であっても分類語彙表番号を付与する。また，長単位より大きな単位（慣用句など）として分類語彙表番号がある場合には，メモとして番号を付与する。

３．作業状況

３.１　現在までの作業概要

作業者と担当サンプルにより，作業ペースに差が生じるが，1 時間あたり 100 語～300 語程度のアノテーションが可能である。以下の表 5 にこれまでの作業における番号付与作業量を示す。

番号付与作業量

表 5　番号付与作業量

付与対象	作業内容	作業量
短単位	番号選択（自立語）	全短単位の 47%
	新規番号追加（自立語）	全短単位の 5%
長単位（短単位の 15%程度）	番号選択	長単位の 14%
	新規番号追加	長単位の 76%

短単位へのアノテーション作業の内訳は，全短単位の 47%において番号選択，5%で新規番号追加である。両作業をあわせ，52%の短単位に番号付与を行っている。長単位は，短単位の 31%が番号付与作業対象となっている。

なお，長単位は全短単位の 15%程度に付与作業を行うこととなる。長単位におけるアノテーション作業の内訳は，14%で番号選択，76%で新規番号追加となり，新規番号追加作業の割合が高い。

３.２　作業における問題点等

作業者から質問のある点については，作業者間に揺れが生じることや，作業結果に影響のあることが予想される。現在までの作業では，作業者と QA を共有しており，作業者の記入した質問に，発表者が回答している。ここでは，作業者との QA に見られる傾向から，作業において問題となる可能性のある点について報告する。

複数の読みが可能な場合

文脈上，複数の読みが可能な場合，付与する 1 つの番号をいずれと定めるのかを作業者

個人の判断にゆだねるため，作業者によって同様の文脈でも揺れの生じる可能性が考えられる。

長単位の文法的分類

助動詞扱いになっている場合をはじめ，どの部分が主となる複合語であるのかの判断にあたり，長単位を，体・用・相のいずれに分類するのかが問題となりがちである。意味の番号は等しい場合でも，作業者によって文法的な分類が異なってくる可能性がある。また，UniDic-分類語彙表対応データの部分的な不備や不足などの影響による作業者の迷いや揺れも散見される。これらは作業の進行により，付与作業済みデータを用いた UniDic-分類語彙表対応データの補填や拡充が可能となることが期待され，今後解消され得る。

４．進捗

これまで（2016 年 10 番号付与済みデータ（2016 年 10 月～12 月）月～12 月）に付与作業の完了したデータは以下（表 6）である。

表 6　番号付与済みデータ（2016 年 10 月～12 月）

集合	短単位	付与数	選択	追加等
PN	50837	25524	24005	3074
その他	13656	6823	6505	318
総計	64493	32347	30510	3392

作業者によって追加等作業数に差があるため，現在は PN とその他に違いが見られるが，今後作業者間の揺れなどの確認を進め，整理と統一を行う予定である。

PN に付与された分類語彙表番号の類は，体の類（1）が 71%，用の類（2）が 20%，相の類（3）が 8%，その他（4）が 1%の割合となっている（図 3）。なお，PN の他の集合でも類の割合は概ね等しい結果となる。

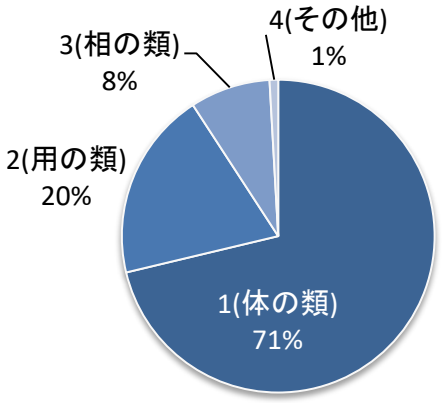


図 3 これまでに付与した分類語彙表番号の類（PN）

次に，PN に付与された上位頻度（1.0%以上）の番号を以下の表 7 に示す。アノテーションしたサンプルが新聞であることから，数記号が 1 割に及ぶ。また，地名や人間や団体の

活動や存在に関する分類番号が多い。

表 7　上位頻度付与番号（PN, 1.0%以上）

番号	頻度		例
1.1960	2665	10.4%	数記号
2.3430	1108	4.3%	行為・活動（する）
1.2590	838	3.3%	固有地名
2.1200	616	2.4%	存在（ある）
1.2000	422	1.7%	人間
1.1962	360	1.4%	助数接辞
1.1000	346	1.4%	事柄
1.2760	263	1.0%	同盟・団体
3.1010	251	1.0%	こそあど・他
付与済み PN	25510	100.0%	

５．作業結果の利活用：新聞に見られる感情表現

これまでに付与作業の完了したデータを用いた調査例を示す。調査には，2016 年 10 月から 12 月に分類語彙表番号を付与した PN の一部のデータを用いた（表 6)。以下では，意味分類のグループを検索した調査結果例として，新聞に見られる感情表現（体・用・相）の分布について報告する。

一般に客観的な記述が多いと考えられる新聞には，感情に関する表現は少ないことが予想される。それでは，新聞において感情に関する言及は，どのような場合にどのようになされるのであろうか。また，どのような感情が言及されるのか。ここでは，感情表現の出現する文脈（新聞の面情報）を見るとともに，記述された感情がポジティブ・ネガティブのどちらに多いのかを調べることで，感情に関する体・用・相の類の使い分けを見てみたい。

分類語彙表番号付与済みデータ（PN の一部; 表 6 参照）を用い，感情に関する分類の付与された短単位を調査した。「.30」は「心」の中項目であり，「.301」に分類される項目は，感情に関連している。また，体（1.301）・用（2.301）・相（3.301）のそれぞれの項目がある。以下，調査結果を類別の頻度で示す（表 8)。

新聞においては，体の類が用いられやすく相の類は用いられにくい[1]ことが予測されるが，相の類の頻度が最も高い結果となっている。それでは，新聞に用いられる感情表現はどのようなものか，ネガティブ・ポジティブの観点と出現文脈を見るため，以下で類別に用例を確認する。

[1] PN の品詞比率（token）を見ると，名詞が 46.5%（BCCWJ 全体は 35.0%，以下同様に示す），形容詞が 1.0%（1.5%），副詞が 0.8%（1.5%）などであり，新聞は体の類にあたる名詞の比率が高く，相の類にあたる形容詞や副詞の比率が低い傾向にある。

表 8　感情に関する表現

分類項目		1. 体の類	2. 用の類	3. 相の類	計
*. 3011	活動-心-快・喜び	2	13	17	32
*. 3012	活動-心-恐れ・怒り・悔しさ	5	2	5	12
*. 3013	活動-心-安心・焦燥・満足	16	4	12	32
*. 3014	活動-心-苦悩・悲哀	4	4	4	12
	計	27	23	38	88

５.１　体の類

体の類は，「エンジョイ」「不快」「怒り」「安心」などが含まれる。

ポジティブ・ネガティブどちらも含む 1.3013 が最も多く，この内訳は，「不安（7 件)」，「満足（3 件)」，「心配（2 件)」，「安心」，「楽」などであり，「不安」や「心配」というネガティブな意味の語彙が含まれている。用例は，(1）が総合面，(2）が国際面，(3）がスポーツ面の記事であるが，ニュース記事に用いられる傾向が見て取れる。新聞に現れる感情語は，ネガティブな意味の語彙である場合，漢語であることが多く，ニュース記事に体の類として現れる傾向があるといえる。

(1）開票まで不安は去らなかった。結果は，２位を八千票以上引き離し，２万票弱を獲得する圧勝。（サンプル ID：PN1b_00005，毎日新聞・総合，下線は著者による。以下同様。）

(2）失業者の増加などが原因とみられ，徴収が困難になれば保険財政の悪化につながる恐れがある。（サンプル ID：PN3g_00001，西日本新聞・総合国際）

(3）「中国の人口十三億人はいいけど，市場としては未知数。スポンサーには期待と同じくらい不安があるんだ」（サンプル ID：PN1a_00008，朝日新聞・スポーツ）

５.２　用の類

用の類は，「すっきりする」「恐れる」「ほっとする」「くよくよする」などが含まれる。具体的には，2.3011 が半数以上を占めている。内訳としても，「楽しむ（11 件)」が大半であり，このほかに「喜ぶ」などがある。以下に例示した（4）は経済面の記事だが，商品紹介部分であった。(5）は生活面の記事である。用の類は，ニュース記事ではない面において，ポジティブな感情を表す際に用いられる傾向がある。

(4）今春，摘んだ茶葉を使用し，熟成したお茶本来のコクが楽しめるという。五百ミリ・リットルペットボトル入り。（サンプル ID：PN1c_00004，読売新聞・経済）

(5）前向きな気持ちで１人のお正月を楽しめば，きっと運も向いてくるのでは？　楽しいプランのあれこれを提案したい。（サンプル ID：PN3b_00004，毎日新聞・生活）

５.３　相の類

形容詞や副詞を含む相の類は，「うれしい」「悲しい」「どきどき」「しんみり」などが含まれる。相の類に分類された表現は体・用の類よりも使用頻度が高く（表 8），新聞においても感情に関する表現として最も使われやすいといえる。

具体的には，3.3011（快・喜び）が最も多い。また，ポジティブ・ネガティブのどちらをも含む 3.3013（安心・焦燥・満足）の内訳は，「冷静（3 件）」，「気楽（2 件）」のほか，「ホッと」，「楽」，「気軽」，「伸びやか」など，概ねポジティブな意味の語彙であった。新聞で用いられる相の類は，ポジティブな感情に関する表現の現れる割合が高い傾向があると考えられる。

用例の（6）は演劇の紹介文，（7）は社説，（8）は家庭面の記事であり，用の類同様に，感情に関する表現は，ニュース記事ではない面に現れている傾向が見られた。

（6）もともとの時代劇ファンには，少し癖のある芝居が鼻につくかもしれない。だが，そういう点を割り引いても，この作品は面白い。（サンプル ID：PN5c_00002，読売新聞・エンターテインメント）

（7）入賞した作文や絵を見ていると，ロケットに乗って宇宙旅行をしたり，宇宙人と対話する近未来の夢が語られ，楽しい気分になってくる。（サンプル ID：PN2g_00004，西日本新聞・オピニオン（社説））

（8）結局，家庭訪問は受けた。内容は「元気です。おもしろい子ですね」程度。だが，学校での子どもの様子はわからないから，それだけでうれしい。（サンプル ID：PN1a_00002，朝日新聞・家庭）

５.４　新聞に見る感情表現のまとめ

新聞に見られる感情表現は，相・体・用の類の順で用いられており，一般に新聞の特徴として考えられる名詞の多さや漢語の多さに反し，相の類の使用頻度が高いという特徴がある。用例を見ると，相と用，体の類で，感情表現の出現文脈が異なる傾向にあることがわかる。相と用の類は紹介文や生活関連記事で用いられており，体の類はニュース記事に用いられているという傾向があった。

よって，新聞に用いられる感情表現は，相と用の類が主にポジティブな感情に関してどちらかというとやわらかい語彙の多い文脈で用いられ，体の類がネガティブな感情に関してどちらかというと硬い語彙の多い文脈で用いられていると考えられる。

６．まとめ

本稿では，『現代日本語書き言葉均衡コーパス（BCCWJ）』に分類語彙表番号を付与する作業について，アノテーション基準と現在までの作業状況を報告した。現在まで，月に 2 万単位（短単位）ほどのアノテーションが進行中である。

これらのデータ整備によって，BCCWJ が意味的な情報によって検索可能となり，従来用例の収集が困難であった意味上のグループに応じた分類を要する研究における新たな可能性が期待される。本稿では，品詞や特定の語彙ではなく，感情という意味上のグループを

用いて，新聞で感情に関する表現がどのように用いられているのかという調査を試みた。

この他にも，たとえば比喩研究では，隠喩のように明示された比喩指標のない用例を収集するために，結合する要素のずれを判定する必要がある。文脈と意味分類を対照することで，このような比喩の用例は格段に収集しやすくなるはずである。また，本作業によって，未知語をはじめとする分類語彙表にない語への番号付与が進んでいるほか，UniDic-分類語彙表対応データの補完となり得る番号付与はもちろん，分類語彙表にない番号の新設が要される場合もあり，既存のデータの拡充も可能となる。

今後，本データを利用した語義の曖昧性解消を自然言語処理研究者により進められることを望む。また，分類語彙表代表義データ（山崎・柏野 2017），UniDic 語彙素番号-分類語彙表番号対応表（近藤・田中 2017),『国語研日本語ウェブコーパス』に基づく word2vec モデル（浅原・岡 2017),『日本語歴史コーパス』に対する分類語彙表番号アノテーションの他,『日本語歴史コーパス』平安時代編の相の類についての分類語彙表番号アノテーション（池上 2017）や，L1 学習者作文コーパスに対する分類語彙表番号アノテーションが進められている。これらのデータに基づく，通時適応モデルの開発や作文支援システムの構築も期待される。

謝　辞

本研究の一部は国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」・言語変化研究領域共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」によるものである。

文　献

F. Bond, T. Baldwin, R. Fothergill, and K. Uchimoto. 2012. “Japanese SemCor: A Sense-tagged Corpus of Japanese” in The 6th International Conference of the Global WordNet Association (GWC-2012)

K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka and Y. Den, 2014. “Balanced corpus of contemporary written Japanese”, Language Resources and Evaluation, 48:2, 345-371.

M. Okumura, K. Shirai, K. Komiya and H. Yokono. 2011. “On SemEval-2010 Japanese WSD Task”, 『自然言語処理』 18(3), 293-307.

浅原正幸・岡照晃. 2017. 「NWJC2vec: 『国語研日本語ウェブコーパス』に基づく単語の分散表現データ」，言語処理学会第 23 回年次大会発表論文集.

池上尚. 2017. 「『日本語歴史コーパス 平安時代編』出現形容詞に対する古典分類語彙表番号アノテーション」，言語処理学会第 23 回年次大会発表論文集.

小木曽智信・中村壮範. 2014.「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」，『自然言語処理』 21(2), 301-332.

近藤明日子・田中牧郎. 2017. 「分類語彙表・UniDic 見出し対応表の構築 ―コーパスへの網羅的・系統的な語義情報付与を目指して―」，言語処理学会第 23 回年次大会発表論文集.

山崎誠・柏野和佳子. 2017.「『分類語彙表』の多義語に対する代表義情報のアノテーション」，言語処理学会第 23 回年次大会発表論文集.

国立国語研究所（編）. 2004. 『分類語彙表増補改訂版データベース』
http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

『日常会話コーパス』プロジェクト —コーパスに基づく話し言葉の多角的研究—

小磯 花絵（国立国語研究所研究系音声言語研究領域）*

Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation

Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所では，2016 年 4 月から「大規模日常会話コーパスに基づく話し言葉の多角的研究」プロジェクトを開始した。このプロジェクトでは，さまざまなタイプの日常会話 200 時間をバランス良く収録した大規模な日常会話コーパスを構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性を，レジスター・相互行為・経年変化の観点から多角的に解明することを目指す。本発表では，プロジェクト全体で推進する研究，およびそのために整備・公開する複数の言語資源の全体像について触れた上で，本プロジェクトの中核を占める『日本語日常会話コーパス』を取り上げ，コーパスの設計について報告する。

1. はじめに

日常会話は社会生活の基盤であり，日常の話し言葉の特徴や仕組み，日常生活を円滑にするための会話コミュニケーションの有様を解明することが求められている。こうした研究を支えるものとして日常会話を広く収集したコーパスの構築は急務である。

海外では，Quirk により 1959 年に開始された The Survey of English Usage 計画において，書き言葉だけでなく話し言葉が大規模に収録され，それに基づく記述文法書が作成されている。その後も，British National Corpus (BNC) や Bank of English，The Santa Barbara Corpus of Spoken American English など，会話を含む話し言葉を収録した大規模なコーパスが数多く構築され，言語学的な研究だけでなく，会話コミュニケーション研究など多様な研究が推進されている。

日本においても，1950 年代から国語研究所において日常会話を含む話し言葉の収録とそれに基づく実証的な話し言葉研究が始まり，『談話語の実態』(国立国語研究所 1955) や『話しことばの文型 (1)(2)』(国立国語研究所 1960, 1963) といった研究報告書がまとめられた。『談話語の実態』は，The Survey of English Usage 計画が始まる 4 年前に刊行されており，先進的な研究であったと言える。しかし残念ながら，収集された音声資料は公開されず，またその後も日本国内においては長らく話し言葉コーパスの構築・公開は行われてこなかった。1990 年代以降，種々の話し言葉コーパスが公開されるようになったが，特定の場面や話者層に偏った

* koiso@ninjal.ac.jp

ものが多く，BNC のように均衡性に配慮して設計されたコーパスは作られてこなかった。

このような状況を受け，国立国語研究所では，2016 年度より機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」(2016〜2021 年度）を開始した。このプロジェクトは，さまざまなタイプの日常会話 200 時間をバランス良く収録した大規模な日常会話コーパスを構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性を，「レジスター」「経年変化」「相互行為」の観点から多角的に解明することを目指すものである。

本稿では，プロジェクト全体で進める研究，およびその推進のために整備・公開する複数の言語資源について概説した上で，本プロジェクトの中核を占める『日本語日常会話コーパス』を取り上げ，コーパスの設計について報告する。

2. プロジェクトの構成

本プロジェクトでは，日常会話コーパスの構築を主導する班と，コーパスを用いて研究を推進し研究の観点からコーパスを評価する三つの研究班を組織している。

コーパス構築班　さまざまなタイプの日常会話 200 時間をバランス良く納めた大規模なコーパス『日本語日常会話コーパス』の構築・公開を主導する（班長：小磯花絵)。

レジスター班　日常会話に加え，講演などの独話や小説などの会話文をも含む多様なレジスターの話し言葉を比較し，語彙・文法・韻律などの特性を探る（班長：山崎誠)。

相互行為班　会話相互行為の中で文法が果たす役割やその特性・構造を，英語など日本語以外の会話との比較を通して総合的に分析する（班長：伝康晴)。

経年変化班　昭和期の話し言葉と現代の話し言葉を，アクセント・韻律・語彙・文法などの観点から比較し，話し言葉の経年変化過程を実証的に解明する（班長：丸山岳彦)。

3. プロジェクトで整備・公開する言語資源

国立国語研究所では，『日本語話し言葉コーパス（CSJ)』や『現代日本語書き言葉均衡コーパス（BCCWJ)』，『国語研日本語ウェブコーパス（NWJC)』など，大規模なコーパスの構築・公開を進めてきた（図 1)。特に，現代日本語の書き言葉の全体像を把握するために構築された 1 億語からなる BCCWJ やウェブを母集団とする 100 億語規模の NWJC の公開により，多様なレジスターを考慮した現代日本語書き言葉の研究をコーパス言語学の手法に基づき研究する環境が整備され，辞書編纂への活用や日本語学習者・日本語教師の利用など，基礎研究に留まらない広がりを見せている。

話し言葉については，CSJ の構築・公開により，話し言葉の言語学的・音声学的な研究や音声情報処理研究を支える基盤は整えられたと言えよう。しかし，CSJ は独話を主対象とするコーパスであり，日常生活の中で交わされる会話は含まれていない。我々は日常生活の中でどのような言葉を使い，人といかなる仕組みでコミュニケーションしているのか，また日常場面でのさまざまな活動を言葉や身体を用いていかに組織化しているのかなど，問うべき課題は多い。こうした研究を支える基盤として，実際の日常会話場面を対象とした大規模な会話コーパスの構築が求められている。本プロジェクトにおいてコーパス構築班が主導する『日本語日常

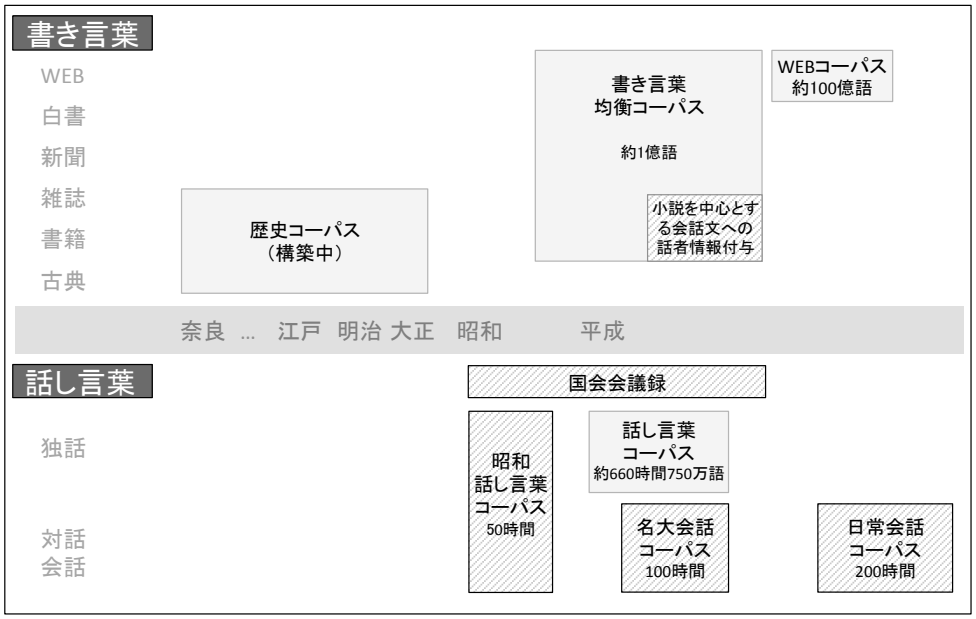


図 1　国立国語研究所で公開・構築中の主たる言語資源（斜線は本プロジェクトで構築する言語資源）

会話コーパス』は，まさにこうした状況を受けて計画したものである。

また書き言葉については，機関拠点型基幹研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」が構築を進める『日本語歴史コーパス』により，書き言葉の通時的な変化を研究する基盤も整いつつある。海外では，2006 年に The Diachronic Corpus of Present-Day Spoken English（DCPSE）という，The Survey of English Usage で 1960 年代後半から 1990 年代前半に録音されたイギリス英語の話し言葉を集めた通時コーパス（約 88 万語）が公開された(1)。こうした話し言葉の経年変化研究の基盤を整えるべく，本プロジェクトでは，経年変化班が中心となり，次の二つの言語資源の構築を計画している。

『昭和話し言葉コーパス』　先に言及した『談話語の実態』および『話しことばの文型 (1)(2)』のために 1950 年代から 1960 年代に録音された日常談話を対象にデータを整備し，『昭和話し言葉コーパス』として一般公開する (丸山 2016)。規模は会話・独話各 25 時間，計 50 時間を予定している。

『国会会議録』ひまわり検索版　国立国会図書館の許諾を得た上で，『国会会議録検索システム』に収録されている国会の会議録のうち 1947 年から 2012 年に開催された衆議院・参議院の本会議・予算委員会を対象に，全文検索システム『ひまわり』で検索できるよう整備する。話し言葉の経年変化研究を効率的に行えるよう，発言者の生年や肩書などの情報も付与し，2016 年 12 月に一般公開した。実際に経年変化研究などで利用しながら，データやシステムの改良を進める予定である。

また，日常会話を含む多様なレジスターの話し言葉を対象に，語彙や文法，韻律などの特性を分析するために，レジスター班を中心に次の三つの言語資源の整備を計画している。

(1) http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm

『名大会話コーパス』　科学研究費基盤研究 (B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」(2001〜2003 年度，研究代表者：大曽美恵子）で作成された，120 会話，計 100 時間の雑談を納めたコーパスである。収録時期は CSJ と重なる。現在は国立国語研究所に移管され，研究所のホームページから転記テキストがダウンロードできるようになっている。本プロジェクトでは，転記テキストを対象に，形態素解析用辞書 UniDic と形態素解析器 MeCab を用いて形態論情報（短単位）を自動付与し[(2)]，メタ情報として発話者の属性（性別・年齢・出生地など）と会話の情報（収録日・収録場所など）を整理した上で，オンライン検索システム『中納言』および全文検索システム『ひまわり』にて 2016 年 12 月に一般公開した (柏野ほか 2017) [(3)]。

BCCWJ **会話文発話者情報**　BCCWJ の書籍における会話文を特定し，発話者の属性情報（話者名・性別・年代）を付与する。2019 年度の公開を目指して作業を進めている (宮嵜ほか 2017) 。

『女性のことば・職場編』『男性のことば・職場編』　『女性のことば・職場編』『男性のことば・職場編』(現代日本語研究会 1998, 2002) として公開されている，職場会話の転記テキストを対象に形態論情報を付与し，全文検索システム『ひまわり』に搭載した。このデータについては，権利関係の都合で一般公開はできないが，出版社の許諾を得た上で，当該書籍を購入したプロジェクトメンバーに限定して利用している。

このように本プロジェクトでは，話し言葉の経年変化やレジスター的特性の研究を支える言語資源の開発を積極的に進め，権利関係の許す範囲で公開する。その際，UniDic に基づく短単位情報の付与や『中納言』・『ひまわり』での公開といったように，できるだけ同一の基準・同一の検索環境を整えることで，複数の言語資源にまたがる分析の利便性を高める。

4. 『日本語日常会話コーパス』の概要

本節では，プロジェクトの中核を占める『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC）を取り上げ，コーパスの設計について報告する。なお，小磯ほか (2017) でその詳細を述べたため，ここでは概要を記すに留める。

■基本方針　日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話を対象とする。幅広いレジスターをカバーするようサンプルを選ぶ。普段われわれがどのような種類の会話を行っているかを調査し (小磯ほか 2016)，その結果を参考に多様な種類の会話を納めたコーパスを構築する。

■規模　コーパスに納める会話の総時間を 200 時間に定める。これまでに収録・転記したデータから試算すると，全体で 200 万語程度になると予想される。

■収録法　日常会話をバランスよく収録するために，首都圏に在住の協力者 40〜50 人（男女 × 20 代・30 代・40 代・50 代・60 代以上 × 各 4〜5 人）に収録機材等を貸し出し，協力者自

(2) 形態論情報の一部については人手で修正を加えている。
(3) 『中納言』での公開については，国立国語研究所コーパス開発センターと共同で実施した。

表 1　プロジェクトで構築する言語資源の公開予定時期

言語資源	公開予定時期	補足
『女性のことば・男性のことば—職場編—』ひまわり版	2016 年 6 月（既公開）	プロジェクト内部限定
『国会会議録』ひまわり版	2016 年 12 月（既公開）	
『名大会話コーパス』中納言版・ひまわり版	2016 年 12 月（既公開）	
『日本語日常会話コーパス』50 時間分	2018 年度	モニター公開
『昭和話し言葉コーパス』テキストのみ	2018 年度	モニター公開
『現代日本語書き言葉均衡コーパス』話者情報	2019 年度	BCCWJ 中納言版を拡張
『昭和話し言葉コーパス』50 時間	2020 年度	本公開
『日本語日常会話コーパス』200J 時間	2021 年度	本公開

身に日常会話 15〜18 時間程度，計約 600 時間の会話を収録してもらう。この方法を「個人密着法」と呼ぶ。収録データの中から，均衡性や倫理的問題，データの質などを考慮し，コーパスに格納・公開するデータとして，各人約 4〜5 時間分の会話，計 160〜200 時間を選定する。個人密着法による会話の種類を調査し，個人密着法では収集の難しい種類の会話（例えば職場での会議や接客場面の会話など）については，調査者が主体となり収録する「特定場面法」で補う。現在は個人密着法に基づく収録を進めている。

■コーパスの構成　コーパスに格納する 200 時間の会話のうち，協力者 20 人，各 2.5 時間，計 50 時間を対象に，2018 年度にモニター公開することを予定している。またモニター公開データの中から 20 時間を選定し，「コア」データ（人手による高精度なアノテーションが付されたデータ範囲）として整備する予定である。

■研究用付加情報　会話を収録した上で，次の研究用付加情報を付与する予定である。

転記テキスト　川端ほか (2017), 臼田ほか (2017) に記した基準および手続きに基づき，映像・音声を参照しながら人手で転記テキストを作成する。

発話単位情報　「長い発話単位」(JDRI 2017) に準拠して発話単位を人手で認定する。

形態論情報（短単位情報・長単位情報）　BCCWJ の単位・品詞設計に準じて短単位情報・長単位情報を自動で付与した上で，コアについては人手で修正する。

文節間の係り受け情報　発話単位を範囲に文節間の係り受け関係の情報を自動で付与した上で，コアについては人手で修正する。

談話行為情報　国際標準化規格 ISO24617-2 に基づき日常会話用に整備した基準に基づき，コアを対象に人手で付与する。現在，基準の整備を進めている (居關ほか 2017)。

韻律情報　コアのうち，録音状態や方言の度合などを参考に選別した会話を対象に，CSJ 構築の際に整備したラベリングスキーム X-JToBI を簡略化した「簡易版 X-JToBI（仮称）」(五十嵐 2015) に準拠して人手で付与する。

5. おわりに

本プロジェクトでは，『日本語日常会話コーパス』を主軸としつつ，研究を推進する上で必要となる言語資源を各種整備する。これらの言語資源の公開予定時期を表 1 にまとめて示す。本プロジェクトの研究を進めるために，2016 年度は言語資源の整備を集中して進めた。2017 年

度からは，こうした言語資源を活用しながら本格的に研究を推進する。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」の研究成果を報告したものである。コーパスの収録にご協力・ご参加くださった皆さまに感謝します。

文　献

現代日本語研究会 (編) (1998).『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.

五十嵐陽介 (2015).「韻律情報」　小磯花絵 (編)『話し言葉コーパス　設計と構築』: 朝倉書店　pp. 81–100.

居關友里子・第十早織・伝康晴・小磯花絵 (2017).「日常会話コーパスのための談話行為タグの設計」　『言語処理学会第 23 回年次大会』.

JDRI (2017).『発話単位ラベリングマニュアル version2.1』.

柏野和佳子・西川賢哉・小磯花絵 (2017).「『名大会話コーパス』中納言版・ひまわり版公開データの作成」　『言語資源活用ワークショップ 2016』.

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵 (2017).「『日常会話コーパス』の転記基準と作業工程」　『言語資源活用ワークショップ 2016』.

小磯花絵・土屋智行・渡部涼子・横森大輔・相沢正夫・伝康晴 (2016).「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」　『国立国語研究所論集』, 10, pp. 85–106.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』の構築」　『言語処理学会第 23 回年次大会』.

国立国語研究所 (1955).『談話語の実態』国立国語研究所報告:8: 秀英出版.

国立国語研究所 (1960).『話しことばの文型 (1) –対話資料による研究–』国立国語研究所報告:18: 秀英出版.

国立国語研究所 (1963).『話しことばの文型 (2) –独話資料による研究–』国立国語研究所報告:23: 秀英出版.

丸山岳彦 (2016).「『昭和話し言葉コーパス』の計画と展望—1950 年代の話し言葉研究小史—」『専修大学人文科学研究所月報』, 282, pp. 39–55.

宮嵜由美・柏野和佳子・山崎誠 (2017).「発話文への発話者情報付与の基本設計—『現代日本語書き言葉均衡コーパス』収録の小説を対象に—」　『言語資源活用ワークショップ 2016』.

臼田泰如・川端良子・徳永弘子・西川賢哉・小磯花絵 (2017).「『日本語日常会話コーパス』の転記基準と特徴について」　『言語処理学会第 23 回年次大会』.

関連 URL

『大規模日常会話コーパスに基づく話し言葉の多角的研究』プロジェクトのウェブサイト
http://pj.ninjal.ac.jp/conversation/

日本語語構成情報データベースの構築

淺尾仁彦 (情報通信研究機構) *

Constructing a Database of Word Structures in Japanese

Yoshihiko Asao
(National Institute of Information and Communications Technology (NICT))

要旨

本研究では、形態素解析辞書『UniDic』への語構成情報の付与について紹介する。語構成情報とは、例えば名詞「招き猫」は、動詞「招く」と名詞「猫」の複合語であるといった情報を指す。日本語について語構成の情報が付与された公開データベースは、複合動詞など特定のカテゴリに限定されたものを別とすれば、管見のかぎり存在しない。このデータベースでは、『UniDic』に対して語構成情報をできるだけ網羅的に付与し、品詞・語種・アクセントなど『UniDic』に元々含まれている情報と組み合わせることにより、「名詞＋動詞の複合名詞」、「アクセントが無核の動詞の名詞化で、アクセントが有核のもの」といった複雑な条件での検索を行うことができ、語彙論・音韻論・形態論などの多様な分野で言語資源として活用可能である。合わせて、開発中の検索インタフェースの紹介を行う。

1. はじめに

近年、『UniDic』のような言語学的な観点の取り入れられた形態素解析辞書や、『日本語書き言葉均衡コーパス』(『BCCWJ』) をはじめとする形態素解析済みの大規模なコーパスが整備され、多様な分野の言語研究に活用できるようになった。

しかしながら、現在のところ、これらの言語資源からは形態論の研究で必要とされる情報を限定的にしか得ることができない。これはいわゆる「形態素解析」で解析される単位が実際には言語学的な意味での形態素（意味を担う最小単位）ではなく、それより大きい単位であるためである。一般に、形態素解析辞書、あるいはそれを用いた解析結果からは、屈折形態論に関する情報は得ることができるが、派生形態論に関する情報 (本研究では語構成情報と呼ぶ) は得ることができない。例えば、「出た」が動詞「出る」の連用形「出-」と助動詞「-た」から成るという情報は得られるが、「家出」は「家出」全体がそのまま辞書登録されており、これが名詞「家」と動詞「出る」の複合であるという情報は得られない。このため、例えば名詞と動詞が複合しているものを検索するという操作は、既存の言語資源では簡単に行うことができない。

形態素解析が言語学的な意味での形態素まで文を分割しないことには一定の合理性があると考えられる。例えば「持つ」という動詞の用法を調査する際に、「気持ち」という語の用例が全て動詞「持つ」の用例として扱われるのは通常、コーパス検索において期待される動作ではない (小椋ほか 2007)。一方で、語彙論、形態論、音韻論の研究では、しばしば語彙項目の内部構造が議論の対象となるため、既存のコーパスや辞書で語彙項目（としてその辞書で扱われているもの）の内部構造の情報に容易にアクセスできないことは、これらの分野における形態素解析やコーパスの有用性の限界となってしまう。

* asao@nict.go.jp

そこで、本研究では、形態素解析辞書である『UniDic』(伝ほか 2007) をベースとし、語構成情報を付与したデータベースを構築し、加工・再配布自由なデータとして順次公開する。また、合わせて、このデータに容易にアクセスできるよう、検索ツールを開発する。

本稿の構成は以下の通りである。2 節で、本研究で開発するデータの設計について述べる。3 節で、現在までのデータ構築状況について述べる。4 節では開発中の検索ツールについて紹介する。5 節でまとめと今後の課題について述べる。

1.1 関連研究

管見のかぎり、網羅的かつフリーで利用可能な日本語の語構成情報データベースは存在しないが、関連する言語資源として以下のものがある。『BCCWJ』の「短単位」は、ほぼ言語学的な意味での形態素に対応する「最小単位」に基づき、その組み合わせとして定義されており (小椋ほか 2011)、本研究で付与する語構成情報はこの「最小単位」のもつ情報と重なる部分がある（この最小単位自体は、公開されている形態素解析辞書では利用できない)。ただし、本研究で認定する語構成情報は、『BCCWJ』で定義されている「最小単位」と一致させることが目的ではない。また、後述のように単に形態素を認定するだけでなく、その範疇などについても、他の項目と関連づけることによって情報を付与することを意図している。

複合動詞については語構成情報を含むデータベース『複合動詞レキシコン』が公開されている (国立国語研究所 2015)。このデータベースは項構造や例文等の情報が充実する一方、収録されている項目は頻度の高いものに限定されているなど [(1)]、やや本研究とは目的が異なると考える。

英語、ドイツ語、オランダ語に関しては『CELEX2』という語彙データベースがあり (Baayen et al. 1996)、フリーではないが、各言語について網羅的な語構成情報が利用可能である。本研究はこのデータベースを 1 つのモデルとしている。

2. 設計

本研究では、語構成情報を形態素解析辞書『UniDic』をベースとして構築する。『UniDic』をベースとするメリットは、(i) ライセンス上、自由に加工・再配布を行うことができる言語資源であること、(ii) 『BCCWJ』などに付与された形態論情報と基本的に対応づけが可能なこと、(iii) 辞書への単語の収録基準が比較的明確であること、(iv) 語種やアクセントなど言語研究に有用な情報があらかじめ付与されていること、などが挙げられる。

本研究でそれぞれの語彙項目に対して付与される情報は以下の通りである。また、本研究で付与される情報のイメージ図を図 1 に示した。

- 形態素境界情報
- 語を構成する各形態素へのリンク
- 語形成に関わる付属情報（連濁、音便など）

例えば「飛び箱（とびばこ)」という項目に対しては以下のような情報が付与される。

- **飛び箱**
  - 境界情報: 飛び/箱, とび/ばこ
  - 形態素へのリンク: 「飛ぶ（とぶ)」、「箱（はこ)」へのリンク
  - 付属情報: 連濁

(1) 複合動詞レキシコンに収録されている複合動詞は 2,759 語だが、本研究では現在 7,842 語の複合動詞を認定している。

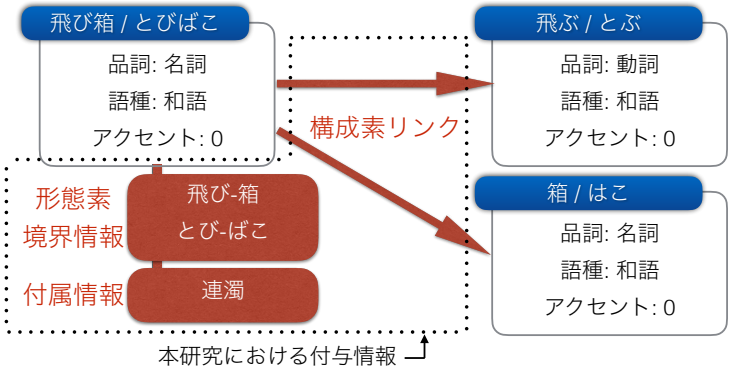


図 1　検索ツールの開発中の画面

リンク先の形態素が別途『UniDic』の見出し語として立項されている場合は、リンク先はその見出し語に紐付けられる。そのため、『UniDic』に付与されている情報を利用し、「飛び箱」の前部要素が動詞であり、アクセント型は 0 であるといった情報にアクセスすることも可能となっている。また、境界情報と形態素へのリンクを別々に付与することにより、例えばこの複合語における後部要素の形式が「ばこ」であることと、この形態素の単独での形が「はこ」であることの両方の情報にアクセス可能となっている。

合わせて、連濁など、形態音韻論に関する付属情報を付している。連濁の有無は、「ばこ」と「はこ」のような形を比較することによって機械的な判定を行うことも基本的に可能だが、検索の便宜のため直接ラベルを付与している。現状認めている付属情報には、連濁、半濁音化、音便 (促音便、撥音便)、音挿入 (促音挿入、撥音挿入、ノ挿入)、被覆形がある。

語を構成する動詞連用形および形容詞（イ形容詞）・形状詞（ナ形容詞など）語幹に関してはそれぞれ動詞・形容詞・形状詞へのリンクを付与する。

- **落ち込み**
  - 境界情報: 落ち/込み, おち/こみ
  - 形態素へのリンク: 「落ちる（おちる)」、「込む（こむ)」へのリンク
  - 付属情報: —
- **狭苦しい**
  - 境界情報: 狭/苦しい, せま/くるしい
  - 形態素へのリンク: 「狭い（せまい)」、「苦しい（くるしい)」へのリンク
  - 付属情報: —

動詞連用形や形容詞・形状詞語幹は、同じ形の名詞が立項されていても、動詞・形容詞・形状詞へのリンクを優先する（例えば形態素「落ち」は動詞の「落ちる」にリンクされ、名詞の「落ち」にはリンクされない)。そのため、名詞「落ち込み」から動詞「落ちる」、動詞「込む」へのリンクはあるが、動詞「落ち込む」や名詞「落ち」へのリンクはないことに注意が必要である。

3.　現在までの構築状況

本研究では、フリーなライセンスで提供されている『UniDic』の形態素解析用辞書 (unidic-mecab 2.1.2) に掲載されている 756,463 項目を、表記のゆれや活用の違いなどを吸収した

199,098 項目にまとめた [2]。この 199,098 項目について語構成情報を付与する。付与にあたっては、機械的な判定手法を援用しつつ、手作業によるチェックも行う。

現在までに、構成要素も『UniDic』に立項されている複合語を優先してデータの構築を行っている。原稿執筆時点では、複合動詞・複合形容詞については人手でのチェックを終えているが、複合名詞・複合形状詞については一部、人手でのチェックが残っている。現段階での暫定的な種類別の語数を、複合語を中心に表 1 にまとめた。表の数値は、今後のデータの修正によって変動する可能性がある。また、以下のようなものについては、語構成情報を整備中であり、表 1 では単純語と合わせて「その他/未処理」に含まれている。

- 派生接辞を含むもの　例：「小骨（こぼね）」「厚み（あつみ）」「羨ましい（うらやましい）」
- 漢字語根を含むもの　例：「出版（しゅっぱん）」「先手（せんて）」
- その他（3 つ以上の形態素を含むもの、複合語と考えられるが構成要素が立項されていないもの、略語、例外的な表記または音形をもつもの、語構成が不明のものなど）

なお、固有名詞、外来語、記号等に関してはその内部構造について情報を付与することは行わない予定である。表では固有名詞、外来語を名詞・形状詞には含めず、全て「その他の品詞」としている（「その他の品詞」の大部分は固有名詞と外来語である）。

表 1　現段階での暫定的な種類別の語数

語構成	名詞 (N)	動詞 (V)	形容詞 (A)	形状詞 (K)	その他の品詞
NN	7,088			28	
VN	3,279			4	
AN	1,050			43	
KN	34			5	
NV	4,489	232		29	
VV	2,225	7,842		16	
AV	340	24		5	
KV	12	1			
NA	198		153	52	
VA	26		23	2	
AA	16		28	5	
KA					
NK	15			12	
VK	3			1	
AK	3				
KK					
その他/未処理	60,108	3,179	621	1,340	106,697
合計	78,886	11,278	825	1,412	106,697

4. 検索ツール

現在、本研究で付与している情報および『UniDic』に元々付与されている情報を検索するためのウェブ UI を開発しており、現在、試験的に公開している [3]。図 2 は開発中の画面であり、動詞+動詞の複合名詞を検索した例を示している。

検索ツールを開発するのは以下のような理由による。本研究で整備するデータはそのまま

[2] 『UniDic』や『BCCWJ』で定義されている「語彙素」に近いものであるが、厳密には対応しない。
[3] http://asaokitan.net/jmorph/

テキストデータとしても公開する予定だが、直接そのデータを利用し、例えば形態素へのリンクをたどって前部要素の属性で絞り込むといった処理を行うにはある程度の知識が要求される。そのため、このような検索が簡単に行えるツールを提供することで、より広い分野の研究者にデータを利用してもらうことが可能になる。

この目的にウェブ UI を用いることには、データダウンロードなどの手間がなく、ユーザ側の環境を選ばないことや、また、データをウェブ上で公開することにライセンス上の問題がないことから、合理的であると考える。なお、ウェブページのソースコードも公開を予定している。

読み	表記	品詞	語種	前読み	前	後読み	後
サキオリ	裂き織り	名詞 普通名詞 一般 *	和	サク	裂く	オル	織る
サキワケ	咲き分け	名詞 普通名詞 一般 *	和	サク	咲く	ワケル	分ける
サグリウチ	探り撃ち	名詞 普通名詞 サ変可能 *	和	サグル	探る	ウツ	打つ
サグリツリ	探り釣り	名詞 普通名詞 一般 *	和	サグル	探る	ツル	吊る
サグリビキ	探り弾き	名詞 普通名詞 サ変可能 *	和	サグル	探る	ヒク	弾く
サゲシブリ	下げ渋り	名詞 普通名詞 一般 *	和	サゲル	下げる	シブル	渋る
サゲドマリ	下げ止まり	名詞 普通名詞 一般 *	和	サゲル	下げる	トマル	止まる
サゲフリ	下げ振り	名詞 普通名詞 一般 *	和	サゲル	下げる	フル	振る
サゲモドシ	下げ戻し	名詞 普通名詞 一般 *	和	サゲル	下げる	モドス	戻す
ササエアイ	支え合い	名詞 普通名詞 一般 *	和	ササエル	支える	アウ	合う
サシイデ	差し出で	名詞 普通名詞 一般 *	和	サス	差す-他動詞	イデル	出でる
サシオサエ	差し押さえ	名詞 普通名詞 サ変可能 *	和	サス	差す-他動詞	オサエル	押さえる

図 2　検索ツールの開発中の画面

原稿執筆時点の検索ツールでは、単語全体の表記・読み・品詞・語種、また構成素の表記・読み・品詞・語種、あるいはその組み合わせを指定して検索することができ、表記・読みについてはワイルドカードも使用可能である。ただし、アクセントや「連濁の有無」など付加情報を用いた検索、また同一の語彙素における表記や読みのバリエーションについては現段階では対応していない。音韻研究での有用性を考えると、正規表現検索、（仮名ではなく）ローマ字による検索、（字数ではなく）モーラ数を指定しての検索などが可能であればより有用なツールになると思われる。これらの点は今後の課題とする。

5.　まとめと課題

本研究では、『UniDic』への語構成情報の付加、およびその検索ツールの開発について紹介した。

本研究の主要な課題として、語構成の曖昧性が挙げられる。本研究は、語源についての情報を意図したものではなく、基本的には語構成に関する共時的な意識を反映させたものを意図している。しかしながら、語構成の意識には話者間の感覚の違いも大きく、合成語と見な

せるかどうかのグレーゾーンに位置する語も多いと考えられる。そのようなケースについての一貫した基準を現段階では持っていない。『BCCWJ』の短単位の定義の元となる最小単位の決定においては、さまざまなルールが用いられている (小椋ほか 2011)。例えば「えがく」や「まつりごと」など常用漢字表に掲げられた訓はそれ以上分割せず全体を最小単位として扱うことや、「いなずま」のように現代仮名遣いに関する内閣告示で「二語に分解しにくい」ために「ぢ」「づ」ではなく「じ」「ず」を用いると定められている語も、分割せず全体で最小単位として扱うという規定などである。これらのルールは、基準ごとに適用可能な項目が限られるうえ、(狭い意味での) 言語学的な基準とは言いにくいため、同様の基準を採用することが本研究の目的に即しているかどうかについては議論の余地がある。

本研究で構築するデータにはさまざまな拡張の可能性がある。例えば、構成要素間の関係についての情報 (項関係、付加詞関係、等位構造の区別など)、意味的情報、統語的情報 (項構造など)、頻度情報などを追加することが考えられる。現段階では、これらの情報を追加することは予定していないが、加工・再配布可能な形で公開することで、必要に応じてこれらの情報を自由に追加できるようにする。

謝　辞

本研究は JSPS 科研費 15H06258「構文形態論の形式モデルの構築に関する研究」の助成を受けたものである。

文　献

小椋秀樹・小木曽智信・小磯花絵 (2007). 「「現代日本語書き言葉均衡コーパス」の短単位解析について」　言語処理学会第 13 回年次大会発表論文集.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007). 「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」　日本語科学, 22, pp. 101–122.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011). 『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版 (下)』, 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 http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf.

国立国語研究所 (2015). 『複合動詞レキシコン』, http://vvlexicon.ninjal.ac.jp.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers (1996). *CELEX2*. Philadelphia: Linguistic Data Consortium.

発話文自動生成のための日本語表現文型辞書の作成

夏目 和子（名古屋大学大学院工学研究科）
刀山 将大（名古屋大学大学院工学研究科修士課程）
佐藤 理史（名古屋大学大学院工学研究科）

A Japanese Expression Dictionary for Automatic Generation of Conversation Sentences

Kazuko Natsume, Masahiro Tachiyama, Satoshi Sato
(Graduate School of Engineering, Nagoya University)

要旨

発話文の自動生成の実現基盤となる日本語表現文型辞書を作成した。この辞書は，依頼や勧誘といった発話の目的（発話意図）に対して，それを伝達する際に使用する複数の言語形式（表現文型）を整理したもので，現在，50 の発話意図に対して，のべ 675 件の表現文型が収録されている。たとえば，発話意図【依頼-実行】には，表現文型「V-てくださらない？」,「V-てくれんか？」,「お願い，V-て」などの 31 種類の表現文型が収録されている。この辞書の特徴は，それぞれの表現文型に，話し方の特徴を表す情報が付与されている点にある。たとえば，「V-てくださらない？」には，「女性的-2，大人っぽい-1，婉曲的-2，丁寧-1」という情報が付与されている。これらの情報を利用することにより，話者の特徴に応じた表現文型の選択が可能となる。

１．はじめに

多くの小説には，登場人物間の会話が含まれる。このような会話を構成する文（発話文）を作る際には，話し手の特徴（性別や年齢，性格など）に適した文の書き分けが必要となる。本研究では，このような発話文自動生成の実現に必要な，話し手の特徴と文型との関係を整理した辞書の編纂を行っている。

発話文生成処理の概要を図 1 に示す。我々が現在採用しているスキーマは，次のようなものである（刀山ほか 2017）。

発話内容　＋　発話意図　＋　話し方の特徴
→　発話内容　＋　表現文型
→　発話文

このスキーマでは，(1)発話内容，(2)発話意図，(3)話し方の特徴，の 3 つの入力から発話文を生成する。まず，発話意図と話し方の特徴から，その話し手に合った表現文型を決定し，これを発話内容に結合することによって，発話文を生成する。具体例を以下に示す。

「その本を取る」　＋　【依頼-実行】　＋　02010210
→　「その本を取る」　＋　「V-てくださらない？」
→　「その本を取ってくださらない？」

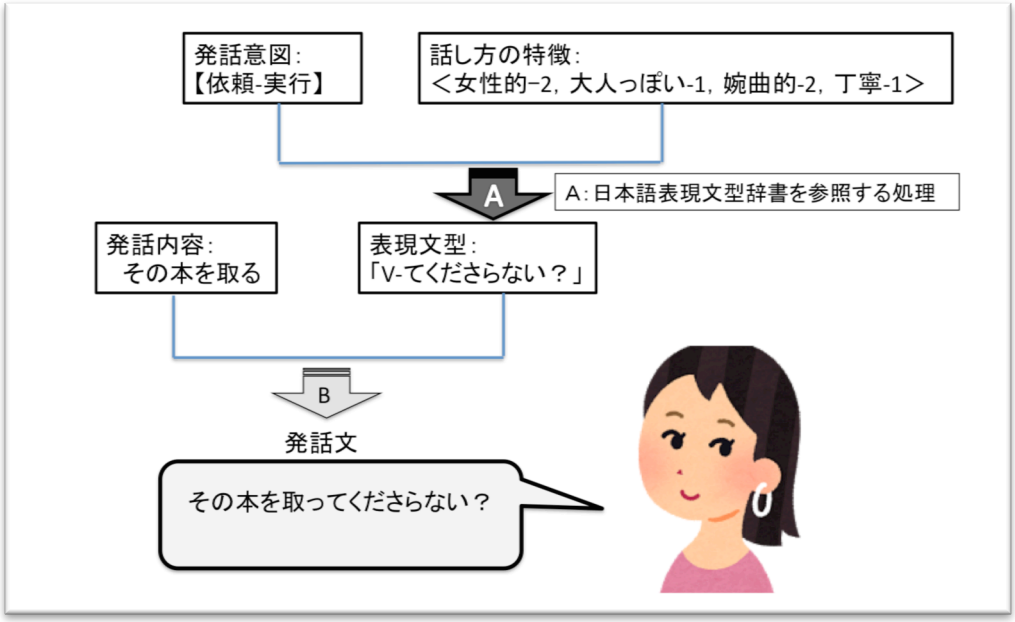


図 1　発話文生成システムの概要

この例に示すように，発話内容はその発話が内在する命題であり，発話意図はその発話の目的である。話し方の特徴は，8 次元のベクトルで表す。このベクトルは，後で説明するように「女性的-2，大人っぽい-1，婉曲-2，丁寧-1」という話者の特徴を意味する。表現文型は，特定の発話意図を伝達するための言語形式である。

本論文では，上記のスキーマの最初の処理（図 1 の処理A），すなわち，発話意図と話し方の特徴から表現文型を決定するために必要な辞書について述べる。この辞書を日本語表現文型辞書と名付ける。

２．日本語表現文型辞書の概要

日本語表現文型辞書は，ある特定の発話目的（発話意図）を伝達するために用いられる言語形式（表現文型）を整理した辞書であり，「ある話し方をする発話者が，ある目的で発話する時，この表現文型を使う」という情報を提供する。この辞書のエントリは表現文型であり，以下の情報を持つ。

1. 発話意図
2. 発話意図内番号
3. 表現文型
4. 例文
5. 話し方の特徴（8 次元のベクトル）

表 1　発話意図【依頼-実行】(E-2) を持つ 15 エントリ

					話し方の特徴							
発話意図		表現文型	例文	D	男性的	女性的	子供っぽい	大人っぽい	断定的	婉曲的	丁寧	粗雑
依頼-実行	1	V-えよ	教えろよ		2	0	0	0	0	0	0	1
依頼-実行	2	V-て	教えて		0	0	0	0	1	0	0	0
依頼-実行	3	V-てよ	教えてよ	1	0	0	0	0	0	0	0	0
依頼-実行	4	V-てくれ	教えてくれ		1	0	0	0	1	0	0	1
依頼-実行	5	V-てくれよ	教えてくれよ		1	0	0	0	0	0	0	0
依頼-実行	6	V-てくれる？	教えてくれる？		0	0	0	0	0	1	0	0
依頼-実行	7	V-てくれるか	教えてくれるか		1	0	0	0	0	1	0	0
依頼-実行	8	V-てくれない？	教えてくれない？		0	0	0	0	0	2	0	0
依頼-実行	9	V-てくれないか	教えてくれないか		1	0	0	0	0	2	0	0
依頼-実行	10	V-てくれんか	教えてくれんか		2	0	0	2	0	2	0	0
依頼-実行	11	V-てくれないかな	教えてくれないかな		0	0	0	0	0	2	0	0
依頼-実行	12	V-てくれないかしら？	教えてくれないかしら？		0	1	0	0	0	2	0	0
依頼-実行	13	V-てください	教えてください		0	0	0	0	1	0	1	0
依頼-実行	14	V-てくださる？	教えてくださる？		0	1	0	1	0	1	1	0
依頼-実行	15	V-てくださらない？	教えてくださらない？		0	2	0	1	0	2	1	0

辞書エントリの具体例を表 1 に示す。この表では，発話意図【依頼-実行】を持つ 31 エントリのうちの 15 エントリを示している。この発話意図は，「聞き手にある行為をするよう頼む」ことである。なお，この表の D 欄が 1 となっている表現文型（この場合は，「V-てよ」）は，この発話意図のデフォルトの表現文型（同一の発話意図を持つ表現文型のなかで，話し方の特徴が最もニュートラルな表現文型）であることを表す。現時点での辞書のサイズは，発話意図 50 項目，表現文型 675 エントリ（のべ）である。

３．発話意図

「発話意図」という用語は，音声対話システムにおけるユーザの発話意図推定や，コミュニケーションにおける発話意図理解などで用いられる用語である。Speech Act（言語行為／発話行為／発語内行為）という用語よりも，詳細かつ具体的なラベル付けが可能[1]と判断して，発話の目的を指し示す用語として，この用語を採用した。

３.１　発話意図の選定

辞書に収録する発話意図の候補を，下記の資料を参考にして選んだ。

a. 荒木(1999) は，日本語対話データのための 24 種類の発話単位タグを提案している。このタグは，「やりとりタグ」「発語内行為タグ」「ムードタグ（益岡 1992 より）」を統合

[1] たとえば，荒木(1999)の発話内行為タグの「情報伝達」は発話意図【説明-事情：のだ】および【説明-理由：からだ】に，「感情表出」は【感心】，【酷評】，【驚き】に，具体化した。

したものである。24 種類の発話単位タグに加えて，統合前のリストから不足するものを補い，計 40 件を収集した。

b. グループジャマシィ(1998) は，日本語の主要な文型を整理した日本語教育（非母国語話者向け）の辞書である。この辞書の意味・機能別項目索引の中から，発話内行為とみなしうる 11 件，モダリティとみなしうる 14 件の計 25 件を収集した。

c. 国立国語研究所(1960, 1963) は，話し言葉のデータを網羅的に分析した研究で，表現意図（言語主体（話し手）が文全体にこめる命令・質問・叙述・応答などの内容），および，それに対する文表現の分類が提示されている。上記の a, b との重複を除いて 16 件を収集した。

d. その他，日本語記述文法研究会(2003)からはモダリティに関する項目を，日本語記述文法研究会(2009)からは談話に関する項目を参考にし，必要に応じて候補を追加・変更した。

こうして得られた 84 件の候補から，話者の話し方の特徴が表出されやすいことを条件に，まず 36 件を選んだ。次に，この 36 件を細分化し，最終的に 68 項目の発話意図を設定した。ここでの細分化には次の 3 パターンがある。

・「-」は，階層を表す（例：【願望-行為】，【願望-物】）
・「：品詞」は，表現文型が適用できる品詞を限定する（例：【感心：A】，【感心：Na】）
・「：語句」は，表現文型で使われる語句（多くは助動詞）を限定する（例：【説明-事情：のだ】，【説明-事情：わけだ】）

末尾の付録に，68 項目の発話意図の一覧表を示す。このうち，現在までに 50 項目の表現文型の記述が完了しており，残りの 18 項目（★印）は，まだ表現文型の記述が完了していない。なお，それぞれの発話意図には，他の発話意図との違いを明確とするための説明を付与した。表の「表現文型の数」は，その発話意図を持つ表現文型の数である。

３.２　発話意図の大分類

発話意図 68 項目を，大きく 9 グループ（A から H，および K）に分類した。この大分類の設定では，主として国立国語研究所(1960, 1963)を参考にした。図 2 に表現意図およびそれに応ずる文表現(分類国立国語研究所 1963: 2)の大枠を示す。

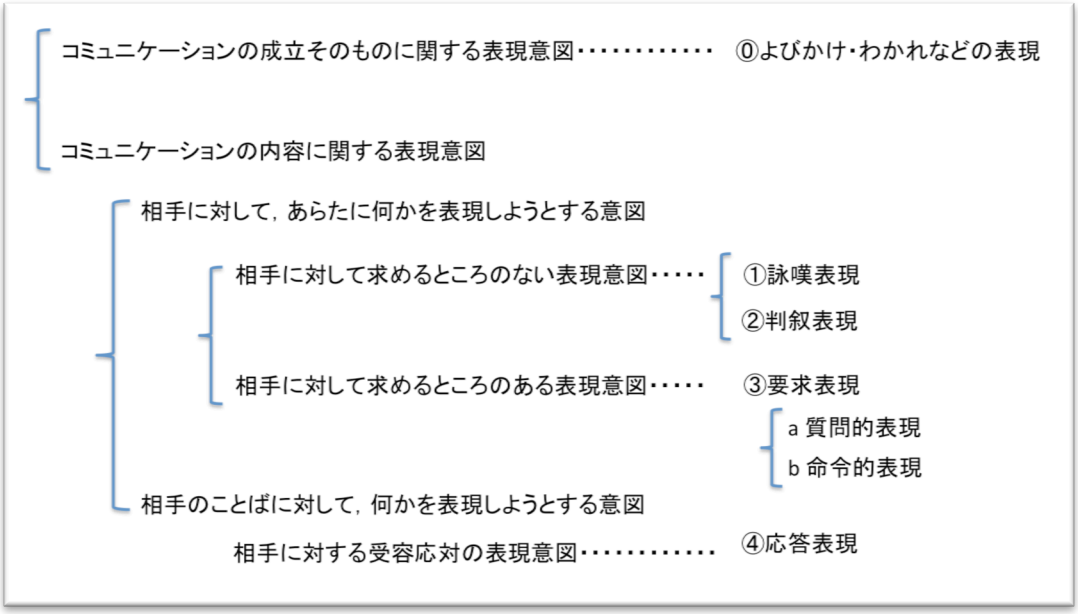


図 2　表現意図およびそれに応ずる文表現の分類（国立国語研究所 1963:2）より抜粋

図 2 の表現意図と，我々の発話意図の大分類は，おおむね以下の対応関係がある。

⓪よびかけ・わかれなどの表現		→	K（感動詞）
①詠嘆表現		→	C（感情表出）, D
②判叙表現		→	A（情報伝達）, B
③要求表現	a 質問的表現	→	F, G（質問）
	b 命令的表現	→	E（行為要求）
④応答表現		→	H（応答）

４．表現文型

日本語表現文型辞書の表現文型とは，ある発話意図を伝達するための言語形式のことで，一つの表現文型に関する情報をまとめたものが本辞書のエントリである。なお，同一の表現文型（言語形式）が複数の発話意図において使用される場合は，それぞれを別エントリとして登録する。

４.１　表現文型の収集

表現文型の収集には，主としてグループ・ジャマシィ(1998)を用いた。具体的には，この辞書の巻末の意味・機能別項目索引を利用して，該当する見出し語から＜文法情報＞を収集した。さらに，巻頭の＜文法関連の記号＞には，名詞，ナ形容詞，イ形容詞，動詞の活用を記号で表した体系表があり，これを表現文型の記述方法として利用した。日本語記述文法研究会(2003)からはモダリティに関する表現，日本語記述文法研究会(2009)からは談話に関する表現を収集した。さらに，森博嗣(1998-2001)の主要人物の会話を分析し，話者の特徴が現れる表現を参考にした。メイナード(2012)からは，ライトノベルの表現の特徴を参考にした。その他，口頭および筆記によるアンケートを基に，追加・変更した。

４.２　表現文型の記述（１）文型パターン

表現文型の典型的な記述形式は，述語文節を述語の品詞を変数として記述した文型パターンである。V は普通体の動詞，A は普通体のイ形容詞，R は動詞の連用形を表す。活用形が限定される場合は，「V(基)」「A(基)」で辞書形，「V-た」「A-かった」でタ形，「V-て」「A-くて」でテ形のように表す。N は名詞句，Na はナ形容詞の語幹を表す。P は述語一般を表す。

表 2 に発話意図【許可】を表す 15 種類の表現文型と例文を示す。【許可】は，聞き手が望んでいる行為を許して促す発話意図である。

表 2【許可】(E-8) の表現文型と例文

発話意図		表現文型	例文(V＝入る, R＝入り)	D
許可	1	お R	お入り	
許可	2	お R なさい	お入りなさい	
許可	3	お R ください	お入りください	
許可	4	どうぞ, お R ください	どうぞ, お入りください	
許可	5	V-て いいよ	入っていいよ	1
許可	6	V-て いいぞ	入っていいぞ	
許可	7	V-て いいわ	入っていいわ	
許可	8	V-て いいわよ	入っていいわよ	
許可	9	V-て いいですよ	入っていいですよ	
許可	10	V-て よろしい	入ってよろしい	
許可	11	V-て よし	入ってよし	
許可	12	V-て もいいぜ	入ってもいいぜ	
許可	13	V-て もかまわないよ	入ってもかまわないよ	
許可	14	V-て かもまわんよ	入ってもかまわんよ	
許可	15	V-て もかまいませんよ	入ってもかまいませんよ	

いくつかの発話意図では，発話内容（命題）の述語の品詞に依存して，表現文型が異なる場合がある。このような場合は，発話意図を品詞別に区分する。表 3 に，**【驚き】**の例を示す。**【驚き】**は，話し手の驚嘆した気持ちを伝える発話意図である。

述語を変数 P で一括して表す場合の例として，表 4 に発話意図【yes-no 疑問文】を表す表現文型の例を示す。【yes-no 疑問文】は，ある事柄の真偽を尋ねるという発話意図である。

表 3【驚き】(C-3) の「表現文型」と例文

発話意図		表現文型	例文	D
驚き:A/A N	1	なんて, A(基) (N な) んだ…	なんて, 可愛い(猫な)んだ…	
驚き:A/A N	2	なんて, A(基) (N な) の…	なんて, 可愛い(猫な)の…	
驚き:A/A N	3	なんて, A(基) (N な) んだろう…	なんて, 可愛い(猫な)んだろう…	1
驚き:A/A N	4	なんて, A(基) (N な) のでしょ…	なんて, 可愛い(猫な)のでしょ…	
驚き:A/A N	5	なんて, A(基) (N な) のかしら…	なんて, 可愛い(猫な)のかしら…	
驚き:Na/Na N	1	なんて, Na (な N) なんだろ…	なんて, 頑固(な父親)なんだろ…	
驚き:Na/Na N	2	なんて, Na (な N) なの…	なんて, 頑固(な父親)なの…	
驚き:Na/Na N	3	なんて, Na (な N) なんだろう…	なんて, 頑固(な父親)なんだろう…	1
驚き:Na/Na N	4	なんて, Na (な N) なんでしょ…	なんて, 頑固(な父親)なんでしょ…	
驚き:Na/Na N	5	なんて, Na (な N) なのかしら…	なんて, 頑固(な父親)なのかしら…	
驚き:N	1	なんという, N …	なんという, 美しさ…	1
驚き:N	2	なんという, N だ…	なんという, 美しさだ…	
驚き:N	3	なんという, N なんだ…	なんという, 美しさなんだ…	
驚き:N	4	なんという, N でしょ…	なんという, 美しさでしょ…	
驚き:N	5	なんという, N かしら…	なんという, 美しさかしら…	
驚き:V	1	なんと, V(基) とは…	なんと, 優勝するとは…	1
驚き:V	2	なんと, V(基) とはねえ…	なんと, 優勝するとはねえ…	
驚き:V	3	なんと, V(基) とはなあ…	なんと, 優勝するとはなあ…	
驚き:V	4	なんと, V(基) とはのう…	なんと, 優勝するとはのう…	
驚き:V	5	なんと, V(基) とは思いませんでした	なんと, 優勝すると思いませんでした	

表 4【yes-no 疑問文】(G-5-1) を表す表現文型と例文

発話意図		表現文型	例文
yes-no 疑問文	1	P ？	学校に行く？(行った？)/美味しい？(美味しかった？)/今日は残業？(残業だった？)/そこは危険？(危険だった？)
yes-no 疑問文	2	P か？	学校に行くか？(学校に行ったか？)/美味しいか？(美味しかったか？)/今日は残業か？(残業だったか？)/そこは危険か？(危険だったか？)
yes-no 疑問文	3	P かい？	学校に行くかい？(行ったかい？)/美味しいかい？(美味しかったかい？)/今日は残業かい？(残業だったかい？)/そこは危険かい？(危険だったかい？)
yes-no 疑問文	4	R-(ます/ました)か？	今日は学校に行き(ます/ました)か？
yes-no 疑問文	5	R-(ます/ました)？	今日は学校に行き(ます/ました)？
yes-no 疑問文	6	A-ですか？	それ, 美味しいですか？(おいしかったですか？)
yes-no 疑問文	7	A-です？	それ, 美味しいです？(おいしかったです？)
yes-no 疑問文	8	N/Na-(です/でした)か？	今日は残業ですか？(残業でしたか？)/そこは危険ですか？(危険でしたか？)

４.３　表現文型の記述（２）定型表現

変数を含まない表現文型を定型表現と呼ぶ。定型表現には，「感動詞およびそれに相当する句」（益岡 1992: 60）が多く，儀礼的な表現も含まれる。このような表現には話者の特徴が現れやすい。下記に発話意図【よびかけ】と【謝罪】を表す定型表現の例を示す。

【よびかけ】の「定型表現」
「ねえ」「ねえねえ」「あの」「あのね」「あのさ」「おい」「おーい」「よう」「えっと」「あ」「ちょっと」「ちょっとちょっと」「やあ」「おい，こら」「もし」「すみません」など

【謝罪】の「定型表現」
「ごめんなさい」「ごめん」「ゴメン！」「ごめんね」「ごめんな」「すみません」「すまん」「わるかったね」「申し訳ありません」「申し訳ない」など

上記以外では，発話意図 D-1【感謝】， K-2【別れ】の表現文型の記述に定型表現を用いている。

５．話し方の特徴とその利用

日本語表現文型辞書のそれぞれの表現文型には，話し方の特徴を表す 8 次元のベクトルが定義されている。

５.１　話し方の特徴を示す項目と値の基準

話し方の基本的な特徴として，以下の 4 軸 8 要素を設定した。

a. ジェンダー：男性的／女性的
b. 世代：子供っぽい／大人っぽい
c. 強弱または長短：断定的／婉曲的
d. 硬軟：丁寧／粗雑

それぞれの要素は，[0,1,2]のいずれかの値をとる。0 は，その特徴がないことを表す。1 はその特徴が弱いことを，2 は強いことを表す。4 軸を 8 要素に分けた理由，すなわち[-2,-1,0,+1,+2]を値域とする 4 次元ベクトルにしなかった理由は，マイナス表記による負のイメージを避けるため（男性 vs. 女性）と，今後の項目や値の拡張性を確保するためである。たとえば，婉曲的な表現には値 3 を付与したいものがある。また，丁寧に対する粗雑には，ぞんざいと攻撃的という異質な特徴が混在しており，将来，分離する可能性がある。

これらの 8 要素の特徴は，話者の話し方の特徴であることに注意されたい。ジェンダー(a)の 2 要素，世代(b)の 2 要素は，生物学的属性ではない。たとえば，男性的な話し方をする女性や，子供っぽい話し方をする成人もいる。

表 5 に，それぞれの要素の値と例を示す。

表 5　話し方の基本的な特徴を示す要素の値と例

	要素	値	例
1	男性的	1:やや男らしい(女性も用いる)	動詞命令形，助動詞:のだ，終助詞:よ・さ，疑問文:か・ないか
		2:とても男らしい	終助詞:ぜ，丁寧の接辞「です/ます」＋な
2	女性的	1:やや女らしい(男性も用いる)	終助詞:ね・の・わ，疑問文:かしら？，丁寧の接辞ます＋よ，意志の疑問文:しよっか，縮約:しちゃう
		2:とても女らしい	丁寧の接辞の否定「ません」＋こと？丁寧の接辞＋わ，くださない？
3	子供っぽい	1:やや子どもっぽい(若者ことば)	縮約:しちゃう，じゃん　定型表現:バイバイ　感嘆の終助詞:なあ
		2:とても子どもっぽい(幼児ことば)	でちゅ(です)
4	大人っぽい	1:やや大人っぽい(社会人ことば)	丁寧の接辞:ます・ません，語彙:
		2:とても大人っぽい(老人(高齢者)ことば)	丁寧の接辞「です/ます」＋な，感嘆の終助詞:のう
5	断定的	1:やや断定的(わかりやすい・語気が強い)	述語に，終助詞・接続助詞などが後続しない，「だ」「である」で言い切る，末尾にエクスクラメーションマークや促音がある など
		2:とても断定的(わかりやすくて語気が強い)	1 が複数該当
6	婉曲的	1:やや婉曲的(知的・弱気)	婉曲表現(疑問形式・否定形式・使役・接続助詞・「と思う」など)，比喩的表現
		2:とても婉曲的(理屈っぽい・わかりにくい)	1 が複数該当
7	丁寧	1:やや丁寧(礼儀正しい)	丁寧体，丁寧の接辞，名詞の接頭辞「お」「ご」，謙譲語，間接的表現など
		2:とても丁寧(堅苦しい)	1 が複数該当
8	粗雑	1:やや粗雑(仲間ことば・タメロ)	普通体，格助詞の省略，終助詞:よね，のさ，縮約:しちゃう
		2:とても粗雑(ぞんざい・攻撃的)	蔑語(侮蔑語・尊大語・粗雑語)　動詞性接尾辞(補助動詞):やがる，くさる

５.２　話し方の特徴を利用した表現文型の選択

前節で説明したように，表現文型のそれぞれには，話し方の特徴を表す 8 次元のベクトルが付与されている。このため，このベクトルを利用して，同一の発話意図の中から，一つの表現文型を選択することが可能となる。

具体的には，(1)発話意図，および，(2)発話者をモデル化した 8 次元ベクトル，の 2 つが与えられたとき，その発話意図を持つ表現文型のなかで，入力ベクトルと最もよく似たベクトルを持つ表現文型を選択する。この選択は，ベクトル間に距離を定義することにより可能となる。現在は，ジェンダーを表す 2 要素の重みを 2，他の要素の重みを 1 とした重み付きユークリッド距離を採用している（刀山ほか 2017)。

このように 8 次元のベクトルを直接入力する方式の他に，あらかじめ作成した話者プロファイルを利用する方式もある。話者プロファイルとは，特定の話者をモデル化したもので，具体的には，10 種類の発話意図に対して，あらかじめどの表現文型を用いるかを定義したものである。表現文型のそれぞれには，8 次元ベクトルが付与されているので，プロファイルの実体は，10 本の 8 次元ベクトルである。表 6 に，話者プロファイルの例を示す。この表の話者 A と B は，森博嗣(1998-2001)の登場人物である。これらの表現文型は，小説の中の該当する発話意図の会話を参考にして選択した。

このような話者プロファイルが用意された話者に対しては，入力として，8 次元のベクトルの代わりに，話者名を指定することが可能である。この場合，次の方法で表現文型を選択するための 8 次元ベクトルを作成する。

1．話者プロファイルに含まれる 10 本のベクトルのうち，入力として与えられた発話意図と同じ大分類を持つ発話意図に対して定義されたベクトルを取り出す。そのようなベクトルが複数ある場合は，それらの平均を計算する。
2．大分類 H と K の場合は，10 本のベクトルの平均を計算する。

表 6 で示すように，話者プロファイルは，発話意図の大分類 A から G までをカバーしている。話者プロファイルを構成する 10 本のベクトルは，必ずしも同じ値をとるわけではない。このため，似たような発話意図に対して定義されているベクトルを優先的に利用する。発話意図の大分類は，このような形で表現文型の選択の際に用いられる。

表 6　話者プロファイルの例

発話意図	話者 A	話者 B
A-1-1【説明:のだ】	今日は仕事が入っているの	今日は仕事が入っているんだよ
	01001000	10000000
B-1-1【願望-行為】	東京に行きたいわ	東京に行きたいな
	01000000	00000000
C-1-1【感心:A】	この映画は良いわね	この映画は良いと思うな
	02000000	00000100
D-4-1【非難:A】	ずるーい！	ずるいな
	01102000	10000000
E-2-1【依頼-実行】	教えてくれない？	教えてくれよ
	00000200	10000000
E-5-1【勧誘-引き込み型】	一緒に行きましょうよ	一緒に行こう
	01010010	00001000
E-8-1【許可】	入っていいよ	入っていいよ
	00000000	00000000
F-1-1【申し出】	荷物を持つわ	荷物を持つよ
	02000000	00000000
F-2-1【提案:どう】	アンケートを取るのはどうかしら	アンケートを取るのはどうかな
	01000100	00000100
G-3-1【確認-念押し】	明日は雨でしょう？	明日は雨だろう？
	01010010	10000000

６．現状のまとめと今後の目標

本論文では，話者の特徴を反映した発話文を生成するために設計・編纂した日本語表現文型辞書について述べた。この辞書は，50 項目の発話意図に対してのべ 675 の表現文型を収録しており，そのそれぞれに，話し方の特徴を表す 8 次元のベクトルが付与されている。

我々は，この辞書を，「想定した話者が，その人らしい文型で，50 項目の発話意図を表現できること」を目指して設計した。しかしながら，「その人らしさ」を表現するには，現状の，話し方の特徴を表す 8 次元ベクトルでは不十分である。なぜなら，異なる表現文型に，同一のベクトルが付与されているものが存在するからである。たとえば，発話意図【依頼-実行】を持つ表現文型は，31 エントリで，そのうちの「V-てほしいんだけど」と「V-てくれる？」のベクトルはいずれも[0000100]になっている。この 2 つの表現文型を区別するためには，前者が内向的，後者が気さく，というように，その表現から受ける印象を表す要素を加えることが必要である。

この問題を解決するために，我々は，話者の話し方の印象を示す項目を追加することを計画している。アッカーマン(2016)は，様々な性格属性を行動や態度やセリフなどで定義している辞典であり，この本の序文において，小説のキャラクタを創作する際，複数の性格的属性を組み合わせることを推奨している。この辞典から，話し方と関係がありそうな性格的属性を選んだ結果，以下のものが話し方の印象を示す要素の候補となっている。

上品・古風・気さく・外交的・内向的・知的・慎重・強引・冷淡・優しい

最終的には，話者の性格を反映した発話文を生成できるようにすることが，我々の目標である。

謝　辞

本研究では, JSPS 科学研究費挑戦的萌芽研究「発話に対するキャラクタ重畳機能の実現」(課題番号 15K12179) の助成を受けている。

文　献

刀山将大・佐藤理史・松崎拓也・夏目和子(2017).「話者属性を反映した発話文生成器の作成」言語処理学会第 23 回年次大会（発表予定）.
大沢在昌(2012).『売れる作家の全技術』KADOKAWA.
荒木雅弘・伊藤敏彦・熊谷智子・石崎雅人(1999).「発話単位タグ標準化案の作成」『人工知能学会誌』14:2, pp.53-62.
益岡隆志・田窪行則(1992).『基礎日本語文法』改訂版, くろしお出版.
グループプジャマシィ(1988).『教師と学習者のための日本語文型辞典』くろしお出版.
国立国語研究所編(1960).『話しことばの文型 1, 対話資料による研究』国立国語研究所.
国立国語研究所編(1963).『話しことばの文型 2, 独話資料による研究』国立国語研究所.
日本語記述文法研究会(2003).『現代日本語文法 4, モダリティ』くろしお出版.
日本語記述文法研究会(2009).『現代日本語文法 7, 談話; 待遇表現』くろしお出版.
森博嗣(1998-2001).『すべてが F になる』ほか S&M シリーズ全 10 巻, 講談社.
泉子 K.メイナード(2012).『ライトノベル表現論』明治書院.
A.アッカーマン・B.パグリッシ(2016).『性格類語辞典, ポジティブ編』フィルムアート社.

付　録

発話意図一覧表 68 項目（★印の 18 項目は作業中）

	記号	グループ名			
	ID	発話意図	デフォルトの表現文型	表現文型の数	説明:代表的な語句・表現
	A	情報伝達			
1	A-1-1	説明-事情:のだ	P のです	11	事情を聞き手に知らせる
2	A-1-2	説明-事情:わけだ	P わけです	10	事情を論理的に述べる
3	A-2-1★	説明-理由:からだ	P からです	0	理由を論理的に説明する
4	A-2-2★	説明-理由:もの	P だもの(もん)	0	理由を主観的に述べる
5	A-3	伝聞	P そうです	18	他者から得た情報を聞き手に伝える:だそうだ/らしい
6	A-4★	引用	P んだって	0	他者の発言などをそのまま伝える:って/とのことです
	B	話し手が自分の考えを述べる			
1	B-1	願望-行為	R-たい な	17	自分がある動作をすることを望んでいる:したい
2	B-2	願望-物	N が欲しいな	9	ある物を望んでいる:が欲しい
3	B-3	期待-事態	V-る といいな	17	ある事態が起こることを望んでいる:といい
4	B-4★	不安	V-る と困るな	0	ある事態が起こることを恐れている:と/たら(いやだ/困る)
5	B-5★	満足	V-て よかった	0	自分の行為の結果・状況を喜んでいる
6	B-6	後悔-非実行	V-ば よかったな	14	実行しなかったことを残念に思う:ば/たら(よかった)
7	B-7	後悔-実行	V-なければ よかったな	16	実行したことを残念に思う
8	B-8	決心-実行	V-よう	16	ある行為をすると決めて宣言する
9	B-9★	決心-非実行	V-る ものか	0	ある行為をしないと決めて宣言する
10	B-10	希望-他者の動作	(N に)V-て ほしいな	17	話し手の他者に対する希望を表す
11	B-11★	意見	P と思う	0	自分の考えを述べる:と思う
	C	感情表出			
1	C-1-1	感心:A	A ね	17	話し手の肯定的な評価を伝える:良い/素晴らしい/すごい
2	C-1-2	感心:Na	Na だね	16	話し手の肯定的な評価を伝える:すてき/最高
3	C-2-1	酷評:A	A ね	10	話し手の否定的な評価を伝える:ひどい
4	C-2-2	酷評:Na	Na だね	12	話し手の否定的な評価を伝える:だめ/最低
5	C-3-1	驚き:A	なんて, A(Nな)んだろう…	5	ある物事の程度が予想外であるという気持ちを伝える:疑問詞感嘆文
6	C-3-2	驚き:Na	なんて, Na(なN) なんだろう	5	ある物事の程度が予想外であるという気持ちを伝える:疑問詞感嘆文
7	C-3-3	驚き:N	なんという, N…	5	ある物事が予想外であるという気持ちを伝える:疑問詞感嘆文
8	C-3-4	驚き:V	なんと, V(基)とは…	5	だれかの行為・ある出来事が予想外であるという気持ちを伝える:疑問詞感嘆文

	D	聞き手に関することで，話し手の気持ち・態度を伝える			
1	D-1	感謝	ありがとう	12	感謝の気持ちを伝える定形表現
2	D-2	謝罪	ごめんなさい	12	謝罪の気持ちを伝える定形表現
3	D-3-1	称賛:A	A-い ね	11	聞き手の行為や性格を褒める:すごい/偉い/美しい
4	D-3-2	称賛:Na	Na だね	11	聞き手の行為や性格を褒める:すてき/立派/最高
5	D-4-1	非難:A	A-い よ	13	聞き手の行為や性格を責める:ひどい/ずるい
6	D-4-2	非難:Na	Na だよ	14	聞き手の行為や性格を責める:だめ/最低/卑怯/わがまま
7	D-5	非難-過失	A-かった/V-たね	10	聞き手の失敗・過ちを取り上げて責める
8	D-6	非難-行為-実行	どうして V の	13	聞き手の不適当な行為を責める:理由を問う疑問文
9	D-7	非難-行為-非実行	どうして V-なかった の	13	聞き手がやるべき事を実行しなかったことを責める:理由を問う疑問文
	E	行為要求			
1	E-1	命令	R-なさい	18	聞き手にある行為を要求する
2	E-2	依頼-実行	V-てよ	31	聞き手にある行為をするよう頼む。
3	E-3★	依頼-非実行	V-ないでよ	0	聞き手にある行為をしないよう頼む。
4	E-4	勧誘-グループ型	V-ようよ	19	グループとして一緒に行動するよう聞き手をその行為に誘う。
5	E-5	勧誘-引き込み型	(聞き手も/一緒に)V-ようよ	11	話し手が実行し(ようとし)ている行為に聞き手を引き込もうとする 。
6	E-6	忠告	V-た方がいいよ	16	心をこめて，過ちや欠点などを直すように言う。
7	E-7	勧告	V-べきですよ	9	ある行為をするように説きすすめる。
8	E-8	許可	V-ていいよ	15	聞き手が望んでいる行為を許して促す。
9	E-9-1	禁止:な/ない	V-ないで	15	聞き手の行為を主観的・直接的に止める:するな
10	E-9-2	禁止:ては(いけない/だめだ)	V-てはいけない	29	聞き手の行為を客観的・間接的に止めようとする:しては(いけない/だめだ)
11	E-10	勧め-行為	V-たらいいよ	23	聞き手がある行為をするよう勧める。諾否または何らかの応答を返す必要はない
12	E-11	勧め-物	N がいいよ	11	聞き手にある物を勧める。諾否または何らかの応答を返す必要はない
	F	聞き手の意向を尋ねる			
1	F-1	申し出	V-ようか	23	話し手が聞き手のためにする行為を申し出る。
2	F-2-1	提案:どう	N/V-る というのはどう?	9	話し手と聞き手に関わる行為の話し手の案に，聞き手が賛成するかどうか尋ねる。
3	F-2-2	提案:いかが	N/V-る というのはいかが?	7	話し手と聞き手に関わる行為の話し手の提案を，聞き手の考えを尋ねる。
	G	質問			
1	G-1-1	確認-肯否要求:N/Na	N/Na だったっけ?	7	話し手がはっきり記憶していないことを，聞き手に尋ねる。
2	G-1-2	確認-肯否要求:V/A	V-たんだっけ?	6	話し手がはっきり記憶していないことを，聞き手に尋ねる。
3	G-2	確認-未知情報要求	疑問詞 だったっけ	7	話し手が思い出せないことを聞き手に尋ねる。

4	G-3	確認-念押し	V-た / そう よね?	10	共通の知識についての確認で，聞き手が同意してくれるという含みがある。
5	G-4	許可要求（肯否要求）	V-て(も)いい？	10	話し手がある行為をしてもよいか，聞き手に尋ねる。
6	G-5	yes-no 疑問文	P の？	19	あることの真偽を尋ねる
7	G-6	疑問詞疑問文	P の？	19	不明なことを疑問詞で表して尋ねる
	H	応答			
1	H-1★	肯定	はい，そうです	0	yes-no 疑問文に対してその命題内容を肯定する
2	H-2★	承諾-依頼	はい，わかりました	0	依頼に対して承諾意志を示す
3	H-3★	受諾-申し出	はい，お願いします	0	申し出に対して受諾する意志を示す
4	H-4★	承諾-誘い	はい，わかりました	0	勧誘に対して承諾意志を示す
5	H-5★	否定	いいえ，ちがいます	0	yes-no 疑問文に対してその命題内容を否定する
6	H-6-1★	断り-依頼	いやです	0	依頼に対して断る意志を示す
7	H-6-2★	断り-申し出	いいえ，けっこうです	0	申し出に対して断る意志を示す
8	H-6-3★	断り-誘い	ごめんなさい，V-れません	0	勧誘に対して断る意志を示す
9	H-7-1★	不明	さあ，わかりません	0	質問に対して答えを知らないことを示す定型表現
10	H-7-2★	未定	さて，どうしましょうか	0	行為要求・申し出・提案等に対してまだ意志が決まらないことを示す定型表現
	K	感動詞			
1	K-1-1	よびかけ	あの	17	対話開始の定形表現:あの/すみません
2	K-2-1	別れ	さよなら	15	対話終了の定形表現:さよなら/じゃあ など

計 675

スマホで古辞書
－『篆隷万象名義』の IDS 検索を例に－

劉　冠偉（北海道大学文学研究科博士課程）†
李　媛（北海道大学文学研究科博士課程）
池田　証壽（北海道大学文学研究科）

Hanzi Dictionaries in Early Ages with Smartphone
A IDS Query System of Tenrei Banshō Meigi

Guanwei Liu (Graduate School of Letters, Hokkaido University)
Yuan Li (Graduate School of Letters, Hokkaido University)
Shoju Ikeda (Graduate School of Letters, Hokkaido University)

要旨

近年,スマートフォンやタブレットのようなモバイル端末が普及し,日常生活を変えつつあり,日本語教育・日本語研究にも使えるようになると予想される。

しかしながら,構築･公開が盛んである古典籍･古文書のデータベースは PC 向けが多く,PC 以外の端末で利用する際は表示サイズのずれや機能障害がしばしば発生する。そこで,モバイル端末でデータベースを利用しているユーザを想定した利便性が高い言語資源データベースの Web インターフェイスを開発したい。漢字字形の構造情報を用いて古辞書のテキスト・画像を検索することによって文字の同定に利用できる Web アプリはまだないので,篆隷万象名義の掲出字について IDS 検索と画像表示を可能にするツールを試作した。本アプリによって,漢字のパーツで篆隷万象名義に掲載している文字の画像をスマートフォンなどの携帯端末で検索でき,写本の解読・翻刻する際に役立つと期待している。

１．はじめに

スマートフォンが急速に社会で普及している。インターネット上の言語資源もスマートフォンへの対応が求められている。一方,日本の古辞書は日本語の歴史的研究に有益であり,これまでの研究と教育において利用・活用されてきた。しかし,日本の古辞書をスマートフォンで利用しようとしたときに,解決しなければならない課題は多い。

(a)　利用に制限のない,デジタル化された翻刻本文と原文画像　［対象］
(b)　パソコンで利用できる古辞書関連サイトとスマートフォン対応　［構想］
(c)　古辞書に含まれる難字・異体字を入力・表示するシステムの開発　［設計］
(d)　サーバに実装する上での問題　［実装］

上記の課題を本文の第 2 節～第 5 節にわたってその詳細を論じていく。

まず第 2 節では,モバイル端末（スマートフォン・タブレットを含む）対応古辞書検索システムを構築するには,それらのデジタル化された翻刻本文と原文画像が必要となり,利用に制限のないことが必要であることを指摘する。本研究では,我々の HDIC プロジェクトで公開している篆隷万象名義データベースの翻刻本文と,利用・公開の許諾を得ている掲出字の原文画像を利用することでこの問題を解決しようとしたことを述べる。

次に第 3 節で,古辞書を検索・表示するスマートフォン対応のサイトを構築する上での課

† toyjack@gmail.com

題を検討し,字形は明白だが,部首・画数・音訓がわかりにくい漢字は,そもそも検索のための入力が困難となるので,入力メソッドの開発が必要であることを述べる。さらに,古辞書に含まれる難字・異体字を入力・表示するシステムの開発には,IDS（詳細後述）のデータを利用するのが有効であるので,IDS データを利用する上での問題と解決策を述べる。

第 4 節では,実際の Web アプリケーションの設計について述べる。そのあと第 5 節ではサーバに実装する上での課題を述べる。

２.『篆隷万象名義』の翻刻本文と原本画像

２.１『篆隷万象名義』

『篆隷万象名義』は,9 世紀前半,唐から日本に戻った弘法大師空海が,梁・顧野王撰述の原本『玉篇』(543)を抜粋した字書である。唯一の古伝本である高山寺本『篆隷万象名義』は研究資料としての価値が高いが,誤写・誤脱が多いことも早くから言われている。一方,中国南北朝以来の字体の古い情報を残すものもあって,字体研究においても重要な資料である。

『篆隷万象名義』は,約 16,000 字の掲出字に対して,字音・字義・字体の記述を収録する。漢字字体研究において,HNG に収録された標準文献に比べて,次の二つの特徴が指摘できる。

(1) 掲出字は古辞書の説明対象としての骨組みであり,少数の重複字以外,ユニークな存在である。一方で,個々の掲出字そのもののバリエーションが僅少であるが,掲出字を網羅的に収録し,異体字も併記するため,漢字字体の多様性を備える。

(2) 異なる掲出字の間に,同一漢字部品が持つものが多く存在する。漢字部品レベルでは,字体の同一性（単一パタン）と多様性（複数パタン）が観察できる。

また,掲出字画像は,字体研究資料であると同時に,掲出字の字形の細部を確認することを可能にするもので,古写本のデータベース構築に不可欠である。さらに,データ化する過程に,Unicode テキストの不足を補う機能している。

２.２　翻刻本文

『篆隷万象名義』の全文テキスト［http://github.com/shikeda/HDIC］は TSV データで公開済みである。その詳細を李・池田（2016）では報告した。Unicode で扱える漢字の『篆隷万象名義』全掲出字に占める割合は,99.2%となる。掲出字「語」の TSV データは次の表 1 の通りである。01～10 の番号は,説明の便宜ため付けたものである。

表 1「語」の TSV データ

01	TBID	3_007_B62
02	TB_vol_radical	v9#91
03	TB_radical	言
04	Entry	語
05	Entry_type	Regular
06	Entry_diff	無
07	TB_def	魚擧反。説也,言也,喜也。
08	SYID	a082b061
09	YYID	無
10	TB_remarks	無

解　説

01　第 3 帖 7 丁裏 6 列の 2 字目（所在）
02　巻 9・部首 91 番目（巻数・部首番号）
03　言部（部首）
04　語（掲出字）
05　隷書掲出字（掲出字タイプ）
06　諸家認定に異同なし（先行研究照合）
07　魚擧反。説也,言也,喜也。
08　対応する宋本玉篇の所在は上篇 82 丁裏 6 列 1 字目（関連字書所在）
09　原本玉篇残巻に存せず（関連字書所在）
10　なし（校勘意見）

２.３　原本画像

『篆隷万象名義』掲出字の原本画像は HDIC のプロジェクトで作成したものを利用している。詳細は池田（2014）・池田他（2016）で述べた。図 1 に「語」（第 3 帖 7 丁裏）と「諒」（第 3 帖 8 丁表）の高山寺本・崇文叢書[1]の項目画像,ならびに掲出字画像を示す。

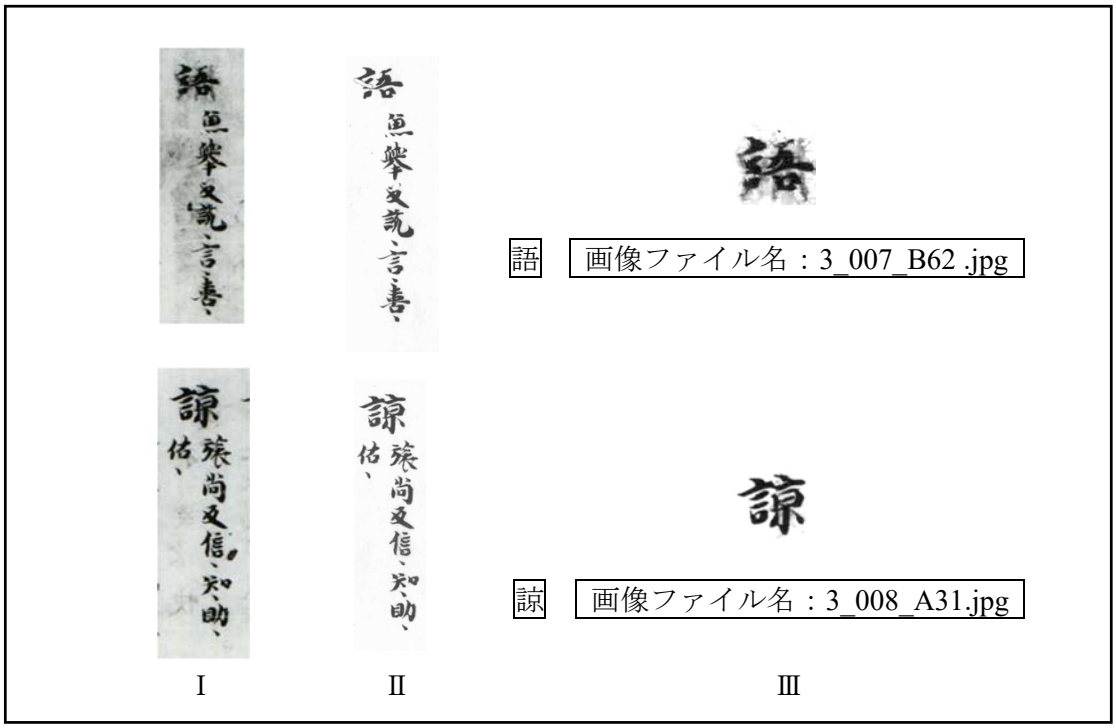


図 1 『篆隷万象名義』高山寺本（Ⅰ）と崇文叢書（Ⅱ）の原文画像・掲出字画像（Ⅲ）

掲出字のテキスト化の際に,画像データベースを構築して,掲出字のテキストの効率化をはかる。また,「諒」のように,旁の「京」の部分について,翻刻本文「京」と原本字形「亰」と相異があるが,テキスト化のとき「亰」を「京」に統一して翻字する。

[1] 図 1 に示した「語」・「諒」の崇文叢書画像は著者の個人蔵書によったが,『篆隷万象名義』崇文叢書のテキストの一部（第 1 輯の第 32 至 43）は,国立国会図書館デジタルコレクションにて公開されている。

３．IDS によっての漢字検索・入力

３.１　漢字の IDS 検索

古辞書に収録される漢字の中には,直ちには音訓がわからないような難字があり,それらの漢字を効率的に検索・入力する方法も問題である。すなわち,字形は明白だが,部首・画数・音訓がわかりにくい漢字は,そもそも検索のための入力が困難となるので,入力メソッドの開発が必要なのである。

古版本・古写本を研究するに際して,翻刻は必須な作業として研究者が多くの時間をかけている。近年,辞書・典籍の電子データベース化と公開がなされており,それらの利用によって,作業の手間が格段に軽減されているが,これらの電子データを検索・編集するために,漢字の入力が常に必要となる。その際,読み方が不明であることや,入力メソッドに未収であることが原因で,漢字を簡単に入力できないケースも少なくない。このような漢字の形しか知らずに漢字を入力したい場合は,まさに紙の字書を引く時と似ている。紙の字書のように,部首と画数を用いて漢字を検索できるデータベースでは Unihan データベースが権威的である。しかし実際に利用する際,次の二つの難点がある。

(1) 同部首同画数の字数が多い場合,欲しい漢字を探すのは難しい。

(2) 所属する部首が分からない場合,利用できない。

部首より小さい漢字構造上の要素によって検索するシステムを作ることでこの二つの問題は解決できる。そのようなシステムを実現するための漢字記述の方法として,「漢字構成記述文字列（IDS)」がある。IDS とは,漢字の構成を文字列で記述したものである。IDS は IDC[2]と漢字の部品からなる。符号化されていない漢字を表すことのできる漢字記述言語の一種である。IDS をすでに符号化した漢字に用いて,漢字の検索方法とすることもできる。このような漢字検索システムはいくつか開発されており,もっとも代表的なものは CHISE/ids-find[3]である。

CHISE は漢字符号をコード制限なしの環境で処理するためのプロジェクトである。CHISE IDS はそれのサブプロジェクトとして,漢字の IDS 情報を整備している。IDS-FIND はそれらの IDS 情報を検索するためのウェブアプリである。

図 2　CHISE/ids-find の PC 画面

図 3　CHISE/ids-find のスマートフォン画面

図 2 に示すように,CHISE の IDS-FIND 機能は PC 向けで開発されている。図 3 に示すように,PC 以外の端末でアクセスすると画面の表示が PC とほとんど変わらず,携帯端末によって操作が難しい場合が生じる。

---

[2] Ideographic Description Character 構造を表す符号であり,「⿰⿱⿲⿳⿴⿵⿶⿷⿸⿹⿺⿻」12 個からなる。

[3] http://www.chise.org/ids-find

CHISE/IDS-FIND における検索結果の数が多い場合,一回の検索結果の表示に数十秒間かかることがある。同様の問題が我々のシステムにも生じるため,解決策が必要となる。表示スピードの問題は 5.2 で検討する。

３.２　IDS データの利用

Unicode 委員会が公開している Unihan データベースは漢字データベースとして最も広く知られているものであるが,IDS に関する情報は現時点まで公開されていない[4]。

現在公開中の漢字 IDS データの中で,整備状況が一番良好なのは CHISE IDS である。CHISE IDS の ReadMe ファイルによると,「<CODEPOINT><CHARACTER><IDS>」三つのフィールドをタブで区切ったデータ構造をとっており,つまり TSV（Tab-separated Values）で示している。CODEPOINT は UCS のコードポイントである。拡張漢字 A までは「U+hhhh」のような「U+」と 4 桁の 16 進数で示す。それ以降は「U-hhhhhhhh」のような「U-」と 8 桁の 16 進数で示す。CHARACTER では CODEPOINT が対応する漢字の符号化字形を示す。IDS はその漢字の構成記述情報を示す。次の表 2 に例を示す。

表 2　CHISE IDS データの一例

CODEPOINT	CHARACTER	IDS
U+5B9A	定	⿸宀&CDP-8BCE;
U-0002A76B	儼	⿰亻⿸厂亚取
U-0002B7AA	甚	⿸甘⿱&AJ1-04307;匹

また,漢字データベースプロジェクトの「漢字構成データベース」に「字形 IDS データ」があり,現在はオープンソース共有プラットホームの GitHub［https://github.com］を用いて「CJKVI-IDS」という名称で公開している。CJKVI-IDS の構造は主に CHISE IDS と同様で,拡張漢字 B まで CHISE IDS のデータがそのまま利用されているようである。拡張漢字 C・D・E のところでは独自の IDS データを採用している。ただし,CHISE IDS と比べると,CJKVI-IDS は CDP 漢字[5]など表外漢字をそれらの画数である「①②③…」のような丸数字に変換している。Unicode の表示方法も少し異なる。IDS の問題点は川幡（2009）に詳しい。次の表 3 に例を示す。

表 3　CJKVI-IDS データの一例

CODEPOINT	CHARACTER	IDS
U+5B9A	定	⿸宀正
U+2A76B	儼	⿰亻⿸厂一④取
U+2B7AA	甚	⿸甘⑤区

今回の試作における IDS データは,高速な検索を実現するため,部品から漢字を合成するデータベースと,漢字から部品に分解するデータベースの二つに分けて作成する。合成データ

[4] http://www.unicode.org/L2/L2015/15065-ids-links.pdf

[5] Chinese Document Processing 台湾の中央研究院が開発した漢字処理システムである。2011 年の最終更新まで約 16 万 5 千の字形が収録されている。

ベースは CJKVI-IDS の ids.txt と ids-ext-cde.txt ファイルをベースとして,IDC と符号化されない部品,すなわち入力困難なものと作者の備考などを削除して,さらにシンプルなデータベースを作成する。次の表 4 に例を示す。

表 4　簡略 IDS データの一例

CODEPOINT	CHARACTER	IDS
U+5B9A	定	宀正
U+2A76B	儼	亻一取
U+2B7AA	甚	区

４．ウェブアプリケーションの設計

４.１　設計上の問題点

スマートフォンやタブレットなどのモバイル端末は PC と比べると大きな相違がある。古辞書データベースの場合では,主として次の三つの問題が生じる。

(1) 画面サイズが小さく,同時に表示できる情報が少ない。

(2) 入力メソッドが軟弱であり,難字の入力に弱い。

(3) システムフォントが不足している。また新しいフォントをインストールできない。

４.２　レスポンシブデザインの採用

(1)に対しては,レスポンシブデザインの応用によって解決できる。レスポンシブデザインとは,端末の画面サイズにより,アプリが相応な仕組みへ変えて表示するデザインである。図 4 は PC の画面,図 5 は携帯端末の画面である。同一のアプリが自由に切り替えることが可能となった。

図 4　PC での画面

図 5　携帯端末での画面

４.３　難字入力方法の開発

(2)については,Unihan データベースで公開している漢字総画数データを使用することによって,漢字の部品と残りの画数で検索できるようにした。例えば,「謐」を入力したい場合に,CHISE/ids-find などの漢字検索システムであれば,「言」と「必」との組み合わせで検索すると「謐」が出てくる。本システムは「言」と残り部分「必」の画数「10」,つまり「言 10」

の組み合わせで検索できる。画数で検索できるようにすることで,難字に対応しない入力メソッドでも負担が少なくなった。

４.４　対応するフォント

(3)については,ウェブフォント技術などを用すれば解決できるが,Unicode 漢字を全すべて含めるとフォントファイルが大きくなる,通信の制限で実現するためにさらに努力が必要と考える。ウェブフォントの圧縮・区分をさらに検討することが必要であるが,これは今後の課題とする。

５. サーバへの実装

５.１　利用したフレームワーク

ウェブアプリに用いるフロントエンドのフレームワークは多くある。データを処理する JavaScript フレームワークは jQuery［https://jquery.com/］,React［https://facebook.github.io/react/］などがよく利用され,表示の仕方を定める CSS では Bootstrap［http://getbootstrap.com/］が定番となっている。だが,今回の開発は JavaScript のフレームワーク Vue.JS［https://vuejs.org/］と CSS のフレームワーク bulma［http://bulma.io/］を用いた。

React や Bootstrap は機能性と汎用性が強く,各分野の開発によく見られるが,学習の労力を考えると,よりシンプルなフレームワークを利用して開発したいと考えた。

Vue.JS は軽量なインターフェイス利用を中心とした JavaScript フレームワークであり,学習しやすいながら強い性能を持っていることで評価されている。bulma も軽量ながら,レスポンシブデザインをサポートする CSS フレームワークである。より広範囲で使われている React と Bootstrap の代わりに,この二つのフレームワークを選択するのは,専門のプログラマーではない筆者（劉）にとって,学習が比較的容易であるというメリットがある。

５.２　検索速度の向上

検索速度を向上するため,IDS データをローカルに保存する。検索プログラムを JavaScript にして,検索の計算をクライアント側に負担させる。より効率的に IDS データベースを利用するため,オープンソース漢字検索システム刹那字引[6]［https://github.com/g0v/z0y］の部分コードを利用した。「刹那字引」は拡張漢字 E までサポートする[7]漢字検索システムである。特に検索スピードに優れている。ただし,現在では開発が止まっているようである[8]。「刹那字引」の検索がはやい理由は,検索用 IDS データベースの再構築である。「刹那字引」では,表 3 のようなデータを逆引きにして利用する。再構築したデータは次のようである。

"宀": "…亙弃宅宓宔宔宕弦宗官宙定宛宜宝实…",

"疋": "定従是琁埏媞捷疋碇蝭蝫　",

また,CHISE IDS/Find とは異なり,毎回検索で画面を更新する必要がない。ただし,画像ファイルの集合のデータ量が大きいので,サーバ側に保存する。アプリの初回利用と画像を請求する時のみサーバと通信する。

---

6 http://www.ksana.tw/kzy/

7 バージョン 1.0 までは拡張漢字 B までをサポードしており,拡張漢字 E までの検索はその以降の再開発バージョンである。コードがだいぶ変わったので,別のシステムであるともいえる。

8 最後のアップデートは 2016 年 3 月であり,筆者が提出したバッグ修復の「Pull Request」となった。その前の最後の更新は 2015 年 10 月であった。

作成したウェブアプリを KTB Image IDS Database を名付けて,https://hdic2.let.hokudai.ac.jp/ids で公開する予定である。

６．おわりに

本稿では HDIC プロジェクトの篆隷万象名義全文テキストデータベース・掲出字画像データと CJKVI-IDS データベースを利用して,篆隷万象名義の掲出字画像をスマートフォンやタブレットなどのモバイル端末での IDS および画数によっての漢字検索を実現した。

本研究は言語資源研究のツールの開発を目的に行ったものである。モバイル端末の利用法について,いろいろな意見をいただいて,さらに改良していきたい。

謝　辞

本研究は JSPS 科研費 16H03422 による成果の一部である。篆隷万象名義全文翻刻テキストと掲出字の画像公開については,高山寺当局ならびに石塚晴通教授（高山寺典籍文書総合調査団団長）のご許可・ご指導のもとに行われている。記して感謝の意を表す。

文　献

池田証壽(2014).「平安時代漢字字書総合データベースー現状と課題 2014 夏ー」『漢デジ 2014 : デジタル翻刻の未来』, 京都大学人文科学研究所附属東アジア人文情報学研究センター編.

池田証壽・李媛・申雄哲・**賈**智・斎木正直(2016).「平安時代漢字字書のリレーションシップ」『日本語の研究』12:2, pp. 68-75.

上地宏一(2005).「CHISE IDS FIND (ソフトウェア レビュー 多言語情報処理)」『漢字文献情報処理研究』2005-10:6, pp.163-165.

川幡太一(2009).「IDS による情報処理」2009 年漢字文献情報処理研究会年次大会.

守岡知彦・師茂樹(2004).「文字素性に基づく文字処理」『人文科学とコンピュータ研究会報告』2004:58(2004-CH-062), pp.53-60.

李媛(2016).「IDS データと HDIC 原本画像・翻刻テキストとを利用した古辞書の漢字字体研究について −『大広益会玉篇』を中心に−」『人文科学とコンピュータ研究会報告』,2016-CH-110:6, pp. 1-6.

李媛・池田証壽(2016). 「篆隷万象名義の全文テキストと公開システムについて」『じんもんこん 2016 論文集』, pp. 95-102.

劉冠偉・李媛・池田証壽(2015).「平安時代漢字字書総合データベースの拡張と和訓対応」『人文科学とコンピュータ研究会報告』2015-CH-106:4, pp.1-8.

The Unicode Consortium(2016). *The Unicode Standard*, Version 9.0.0, Unicode Consortium.

関連 URL

平安時代漢字字書総合データベース(HDIC)	http://hdic.jp/
KTB Image IDS Database	https://hdic2.let.hokudai.ac.jp/ids
CHISE project	http://www.chise.org
CHISE IDS	http://www.chise.org/ids/
漢字データベース	http://kanji-database.sourceforge.net/index.html
UniHan Database	http://www.unicode.org/charts/unihan.html

機械翻訳用超大規模辞書データ資源

春遍雀來（日中韓辭典研究所）

"Very Large Scale Lexical Resources for Machine Translation"

Jack HALPERN (The CJK Dictionary Institute, Inc. (CJKI))

要旨

情報交流の国際化に伴い多言語情報の充実は今や喫緊の課題である。特に固有名詞や **POI** (points of interest)は膨大な数量に加え頻繁な名称変更にも対応する必要があるため，正確で充実した多言語辞書データ資源が必須だ。そこで，機械翻訳の作業効率と精度を格段に向上させる，**超大規模辞書データ資源(Very Large Scale Lexica: VLSL)**の構築例として，固有名詞・専門用語等を含む日中韓英辞書データベースや多言語固有名詞辞書データベースを紹介する。VLSL は情報検索・形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野に応用が可能で更なる展開が期待される。

１．はじめに

近年，科学技術・学術・文化等の多方面で諸外国との相互理解・交流の重要性が再認識されている。2020 年の東京オリンピック開催に向け，多言語情報の充実は今や喫緊の課題となっている。IT 技術の発達に伴い，多言語情報は企業から一般ユーザーまで広く活用されるようになったが，そのような技術に不可欠なのが豊富な情報を包括した大規模な辞書データ資源である。

当研究所は，日中韓英を中心とする各種の辞書データベースの構築を行っており，固有名詞・専門用語の他，日本語の語彙・異表記等も含め約 2400 万項目を収録している。また，IT 関連の大手企業に広く採用されている中日・日中専門用語データベースは 20 分野に亘る専門用語を網羅した日中対訳辞書である。更に，動詞と形容詞・形容動詞を扱った日本語全活用辞典（J_FULEX）の開発もある。

これらの辞書資源は，人力による翻訳や機械翻訳の作業効率と精度を格段に向上させてきた一方，形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野で応用されている。

２．多言語固有名詞辞書データベース

辞書データ資源は翻訳のみならず，各種の言語データ処理の場面でも活用される。例えば自然言語処理に於いて特に扱いが難しい固有名詞では，多数の異表記（アラブ人名「アブドゥル・ラフマーン」には千通り以上のアルファベット表記がある）や平仮名表記の中国語訳（市町村名等）に対応しなければならない。また，一般語彙に於ける同義語（「ソフトウェア」は簡体字では「**软**件」，繁体字では「軟體」と表記）を処理する際にも辞書データベースは有用である。これらの点を踏まえ，当研究所では専門用語を含む膨大な辞書データベースの構築・拡張を続けている。

３．POI の辞書データベースと機械翻訳

地名や POI（points of interest = ホテル，公園，大学，施設等）は数が膨大である上，名称が変更される場合もある。各言語体系に基づく正しい表記が必要であるため，アルゴリズムによる全面的な自動処理での生成は不可能で，辞書データベースが必須となる。最先端のニューラル機械翻訳（NMT）ですら，POI の辞書データベースなしには学校名・道路情報での翻訳がほぼ不可能な事が Google の抜き取り調査から明らかになった。

POI の辞書データベースを含む超大規模辞書データ資源の構築は半自動的に行われ，結果に求められる精度と費用を勘案して自動翻訳と人間翻訳の割合を決定する事になる。特に固有名詞の翻訳作業では字訳・音訳・意訳・意音訳による自動変換と，人間翻訳という５通りの手法が数えられ，実際にはこれらの多様な組み合わせが可能である。つまり自動処理の割合が高く，安価で速いが精度が上がりにくいものから，人間翻訳で高価だが翻訳としての正確さを期す（定訳を選択する）ものまで様々である。

日本の地名・公共施設名

日本語	成田国際空港	京都府庁
中国語(簡体字)	成田国际机场	京都府厅
中国語(繁体字)	成田國際機場	京都府廳
韓国語	나리타국제공항	교토부청
英語	Narita International Airport	Kyoto Prefectural Office
アラビア語	مطار ناريتا الدولي	مكتب محافظة كيوطو
インドネシア語	Bandar Udara Internasional Narita	Kantor Pemerintahan Kyoto
ベトナム語	Sân bay quốc tế Narita	Tòa nhà chính quyền tỉnh Kyoto
タイ語	สนามบินนานาชาตินาริตะ	ที่ว่าการจังหวัดเกียวโต
ヒンディー語	नारिता अंतर्राष्ट्रीय हवाई अड्डा	क्योतो प्रीफ़ेक्चर मुख्यालय
ロシア語	Международный аэропорт Нарита	администрация префектуры Киото
ドイツ語	Internationaler Flughafen Narita	Präfekturverwaltung Kyoto
ポルトガル語	Aeroporto Internacional de Narita	Sede do Governo de Quioto
スペイン語	Aeropuerto Internacional de Narita	Oficina Prefectural de Kyoto
フランス語	Aéroport international de Narita	Préfecture de Kyoto
イタリア語	Aeroporto Internazionale di Narita	Sede del Governo prefettizio di Kyoto

４．日本語異表記データベース

日本語は表記の幅が広い言語であり，日本語異表記の種類には，漢字表記・平仮名表記・片仮名表記・交ぜ書き等がある。更に片仮名語の異表記（コンピュータとコンピューター，メイドとメード等）も多数出現する。また，同音異形異義語の具体例には，うまい＝美味い，上手い，巧い等，意味や表記の揺れが認められる。更に，日本語を扱う際には意味互換性の度合いや同訓異字への対応，異表記の類型（送り仮名や文字種等），詳細な属性等きめ細やかな配慮が常に求められる。

当研究所は自然言語処理で課題となるこれら異表記の問題を，データベースに全てを包括する事によって解消している。各種国語辞典・内閣告示・新聞や公用文に見られる表記・出現頻度等，様々な角度から総合的に判断した「代表表記」を定める作業が，現在も進行中である。

日本語異表記辞書データサンプル

ID	読　み	POS	SUB_ID	表　記	代表表記
F000043	あっせん	VN	a	斡旋	あっせん
			b	あっせん	
			c	あっ旋	
F000690	あかとんぼ	NC	a	赤とんぼ	赤とんぼ
			b	赤トンボ	
			c	赤蜻蛉	
			d	アカトンボ	
			e	あかとんぼ	
F000853	あきかん	NC	a	空き缶	空き缶
			b	空缶	
			c	明き罐	
			d	あき缶	
			e	あき罐	
			f	空きかん	
			g	空きカン	
			h	空き罐	
			i	空罐	
			j	空き鑵	
			k	空鑵	
F001543	あじつけ	VN	a	味つけ	味付け
			b	味付け	
			c	味付	

５．固有名詞情報と VLSL(超大規模辞書データ資源)

当研究所では，こうした異表記を網羅する日中韓英各語とアラビア語の大規模な辞書データ資源を提供しており，世界の大手企業もこれを採用している。中国語のデータベースには検証済みの正確なピンインも収録されている。先進的な計算辞書学の手法によって構築・維持された当研究所のデータ資源は，固有名詞・専門用語のほか，日本語の語彙・異表記・音韻等も含め，約2400 万項目に上る。

日中韓英固有名詞データベースの収録語数

	日英	日中	日韓
中国人名	1,000,000	1,000,000	1,000,000
中国地名	2,400	5,600	3,000
韓国人名	13,000	2,100	13,000
韓国地名	5,900	2,000	5,900
日本人名	390,000	281,000	390,000
日本人姓	150,000	91,000	150,000
日本地名	77,000	74,000	77,000
西洋人名	31,000	38,000	10,000
西洋地名	1,100	2,500	1,800
合　計	1,670,400	1,496,200	1,650,700

「日中韓英固有名詞データベース」は日中韓英語の各種固有名詞辞典を含み，総計1100 万項目に及ぶ大規模なデータベースである。その用途は機械翻訳，情報検索，形態素解析，電子辞書，入力システム，固有名認識等多岐に亘る。

日中専門用語データベース

分野	中国語	日本語
医学	肾上腺素能受体	アドレナリン受容体
生物	亲和性	親和性
生物	亲和层析法	アフィニティークロマトグラフィ
生物	琼脂扩散法	寒天拡散法
生物	琼脂糖	アガロース
生物	琼脂胶	アガロペクチン
生物	类蛋白	アルブミノイド
医学	类天花	アラストリム
医学	变应性试验	アレルギー試験
医学	变应性肉芽肿	アレルギー性肉芽腫

「中日日中専門用語データベース」は日中二ヶ国語の双方向対訳辞書である。コンピュータ科学からバイオテクノロジーに至る 20 分野に亘る幅広い専門用語を収録しており，収録語は中日・日中それぞれ約 80 万語，総計約 160 万語に及ぶ。その用途は特許翻訳を含む各種翻訳業務，用語の抽出やインデックス作成に役立つ情報検索アプリケーション，形態素解析や分節システム等，各種の自然言語処理アプリケーション，スマートフォンアプリケーションや電子辞書・CD-ROM 等多岐に亘る。

6．まとめ

POI の辞書データベースを含む VLSL（超大規模辞書データ資源）は各種の自然言語処理に向いており，とりわけ機械翻訳に有効である。コンピュータメモリーが無制限に拡大可能になった今日，自然言語処理に於いてはアルゴリズムやコーパスのみに過度に依存する必要はもはやない。VLSL や POI の辞書データベースの効果的な活用は固有名詞の翻訳精度を大幅に向上させるばかりではなく，情報検索や形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野に応用が可能であり，更なる展開が期待されるのである。

モンゴル語アクセント研究のためのデータベース

玉栄（内モンゴル大学、国立国語研究所外来研究員）†
西川賢哉（国立国語研究所コーパス開発センター）
前川喜久雄（国立国語研究所コーパス開発センター）

Database for the Research of Word Accents in Mongolian

Yurong (Inner Mongolia University, National Institute for Japanese Language and Linguistics)
Ken'ya NISHIKAWA (National Institute for Japanese Language and Linguistics)
Kikuo MAEKAWA (National Institute for Japanese Language and Linguistics)

要旨

モンゴル語のアクセントは、音韻論的には弁別的でないといわれているが、音声学的な特徴については研究者によって意見が分かれている。従来は、第一音節に固定ストレスアクセントを認める研究が主流であったが、1980 年代以降、実験音声学の影響によって、アクセントは第一音節に固定されておらず、その変異には音節構造が関係しているとの主張が広がってきた。本発表では、モンゴル語アクセントの音声学的特徴を把握するために、筆者らが設計と実装を進めている音声データベースについて報告する。このデータベースは、音節構造、母音の長短、隣接子音等に配慮した単語リストを複数の話者が発音したサンプルに、種々の音響特徴量を付与したものとなっており、モンゴル語のアクセントが種々の韻律的特徴（長さ、強さ、高さ）および分節的特徴とどのような関係にあるかを解明するために利用できる。

１．はじめに

筆者らは現在、モンゴル語の語アクセントを分析するための音声データベースを構築している。本稿ではデータベースに関する設計と実装について報告する。

モンゴル語にはいくつかの方言が認められるが、本研究では、中国内蒙古自治区を中心に使用されている内モンゴル語を対象とする。

２．モンゴル語のアクセントに関する研究概況

モンゴル語のアクセントは、音韻論的には弁別的でないという点では学者たちの意見は一致している。しかし、アクセントの性質、類型、位置、アクセントと物理的特徴の関係などは研究者によって意見が分かれている（概説として Unir and Yu 2015 を参照）。

モンゴル語アクセントの性質について伝統的な研究では、多くの学者はストレスアクセントと認める一方、第一音節にはストレスアクセント、第二音節以降はピッチアクセントと考える学者もいる(Sh.Lvbsanwandan 1986)。

アクセントの類型について、固定と自由の二つの見方があるが、多くの学者は固定アクセントと考えてきた。

アクセントの位置について、かつては第一音節にアクセントが固定されるという考えが主流であった。例えば、ロシアの学者 I.J.Schmidt (1832)の「モンゴル語の文法」には、多くの 2-3 音節語のアクセントは第一音節にあるとの記述がある。それに対し、N. N.Poppe

† umyurong[@]yahoo.co.jp

「ハルハモンゴル文法」(1951)は、語のアクセントは音節構造と関係があるとし、長母音や二重母音があれば、最初の長母音や二重母音に強勢があり、長母音や二重母音がなければ、第一音節に強勢があると指摘した（以上は、Sh.Lvbsanwandan 1986 に基づく）。また、1980 年以降の実験音声学の研究方法によって、Choijingjap (1993)、Huhe (2001)、Bayarmend (1997)は、モンゴル語のアクセントが第一音節に固定されているわけではないことを明らかにした。現在のところ、アクセントの位置には一致した意見はなく、音節構造によって決定されるという見解がある一方で、第二音節にある(Baoyuzhu 2011)との見方もある。

アクセントと物理的特徴（長さ、強さ、高さ）の関係についても、一致した意見はない。(i)強さが一番影響を与える、(ii)長さが一番影響を与える、(iii)高さが一番影響を与える、(iv)長さ、強さ、高さが合わせて影響を与える、という四つの説がある。

３．データベース

このように、モンゴル語のアクセント、特にその音声学的側面に関しては、意見の一致が見られていない。そこで筆者らは、実態を把握し、どの説が妥当なのかを実験的に検証するため、音声データベースを構築することにした。以下、このデータベースについて説明する。

３．１単語リスト

本データベースには、モンゴル語母語話者による単語の読み上げ音声、および各種研究用付加情報を収録する。単語は一音節語から四音節語の計 684 語用意した。単語の選定にあたっては、音節構造に配慮した。モンゴル語の一音節語の構造を V（母音）、C（子音）で表記すれば、基本型には V,VC,CV,CVC,VCC,CVCC の 6 種類が認められ、これに長短母音の対立が加わる。ここでは、これらの音節構造を網羅できるように単語を選定した（この際、Huhe et al. 2001 を参考にした）。さらに、先行研究でよく議論されている VCCC,CVCCC 構造を有する（と言われる）単語もリストに加えた。

３．２録音

録音は、国立国語研究所のモニター室で行なう。使用機材は、Edirol 4-Channel Portable Recorder and Wave EditorR-4,Sony Condenser MicrophoneC-357 で、44.1KHz,16bit で録音する。

話者には、単語単独で 1 回、2 種類のキャリア文に埋め込んで各 1 回発話してもらい、これを（日を置いて）2 回繰り返す。結果、一つの単語につき同じ話者の発話が 6 トークン得られることになる。単語はランダムに提示する。キャリア文は表 1 に示すものを用いる。この中から、キャリア文と当該単語の境界が「子音＋母音」あるいは「母音＋子音」

表 1. 使用するキャリア文

単語構造	キャリア文 1	キャリア文 2
/V...V/	[manɛːt __ pɔltʃɛː] （私たち（のところ）___ なりました）	[pit __ xarsan] （私たち ___ 見ました）
/V...C/	[manɛːt __ ɔltʃɛː] （私たち（のところ）___ もらいました）	[pit __ apsan] （私たち ___ 取りました）
/C...C/	[manɛː __ ɔltʃɛː] （私たち（の）___ もらいました）	[piː __ apsan] （私 ___ 取りました）
/C...V/	[manɛː __ pɔltʃɛː] （私たちの ___ なりました）	[piː __ xarsan] （私 ___ 見ました）

となるものを使用する。例えば、ターゲットとなる単語が [tʰɔːn] の場合、子音で始まり子音で終わる語なので、キャリア文としては、前文が母音で終わり、後文が母音で始まるもの、すなわち [manɛː __ ɔltʃɛː] および [piː __ apsan] を使用することになる。

現在のところ、女性話者 1 名（内モンゴル赤峰出身）の録音が終了しており、今後は女性 2 名、男性 1 名（以上、内モンゴルシリンゴル出身）を録音する予定である。

３．３アノテーション

次に、録音された音声に対するアノテーションについて説明する。ここで述べるのは現時点での仕様であり、今後変更される可能性がある点にご留意いただきたい。

アノテーションは Praat (Boersma&Weenink 2017)で行なう。Praat 用アノテーション形式である TextGrid に、ID 層、Word 層（単語層）、Seg 層（分節音層）、Comment 層を設ける[1]（図 1 参照）。

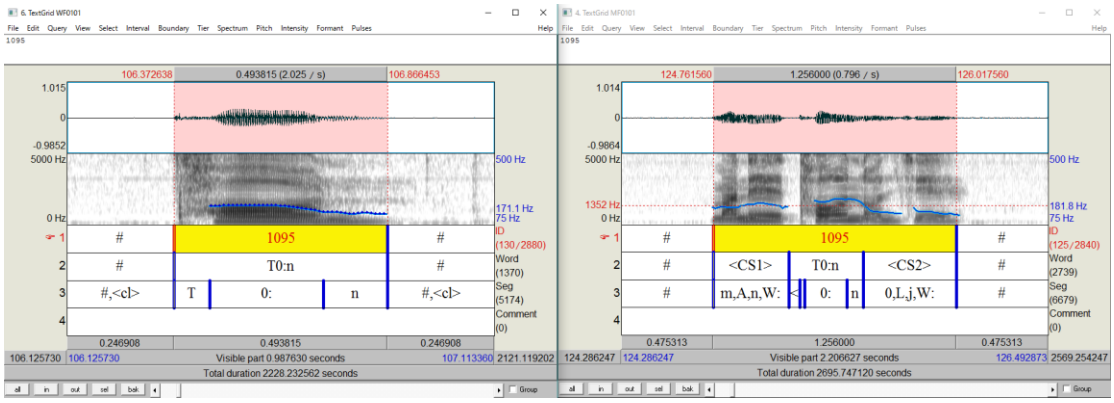


図 1. アノテーション例：
左が単語単独発話、右がキャリア文に埋め込んだ発話

ID 層では、発話（キャリア文付きで発話している場合、それを含む全体）の区間に対し、個々の単語に一意に割り当てられた 4 桁の数字(ID)を与える。

Word 層では、そこで発話されている単語をラベルとして与える。入力および検索の利便性をはかるため、ラベルには、IPA (International Phonetic Alphabet)ではなく、筆者らが独自に定義した、ASCII 文字から構成される音声表記を用いる。IPA との対応を表 2、表 3 に示す。キャリア文を発話している区間には、<CS1>, <CS2>というラベルを付与する（それぞれ、キャリア文の前文、後文を表す）。

Seg 層には当該単語を構成する分節音を与える。ここでは、表 2、表 3 に示した音声表記に加え、表 4 に示す補助ラベルを用いる。種々の理由により分節音境界を決定できない場合には、『日本語話し言葉コーパス』の分節音ラベリング（藤本・菊池・前川 2006）で考案された方式に従い、無理に境界を定めることはせず、複数の分節音をカンマで融合させたラベル（融合ラベル）を使用する。キャリア文を発話している区間（Word 層における<CS1>および<CS2>の区間）は分析の対象外であるが、キャリア文と単語の境界で分節音が融合する可能性を想定し、ひとまずその区間の分節音を融合ラベルの形で初期値として与えてある。

[1]今後 Syl 層（音節層）を追加する予定である。Syl 層の追加にあたっては、すべてを人手で行うのではなく、Seg ラベルから Syl ラベルを機械的に生成し、必要な個所について人手修正することを検討している。

Comment 層は、作業用のコメントを記述する層である。最終的には削除される。

現在、単語単独発話 1 回分およびキャリア文付き発話 1 回分の一次アノテーションが終了したところである。

表 2. 母音ラベル

IPA	ASCII	IPA	ASCII
ɐ	A	ɜ	a
ə	E	ɞ	e
i	I	ɨ	i
ɪ	1		
ɔ	0	ɵ	q
ʊ	V		
o	O	ɵ	o
u	U	ʉ	u
ɛ	W		
œ	@		
ɐː	A:		
əː	E:		
iː	I:		
ɔː	0:		
ʊː	V:		
oː	O:		
uː	U:		
eː	2:		
ɐi	AI		
ɔi	0I		
ʊi	VI		
oi	OI		
ui	UI		
ɐɛ	AW（ɐi の異音）		
ɛː	W:（ɐi の異音）		
ɔɛ	0W（ɔi の異音）		
œː	@:（ɔi の異音）		
ʊɛ	VW（ʊi の異音）		
ue	U2（ui の異音）		
oe	O2（oi の異音）		
yː	y:（ui の異音）		

表 3. 子音ラベル

IPA	ASCII
n	n
p	p
p^h	P
x	x
k	k
m	m
l	L
s	s
ʃ	sh
t	t
t^h	T
ʧ	ch
ʧ^h	CH
j	j
r	r
ŋ	N
ɸ	F（p の異音）
β	B（p の異音）
k^h	K（x の異音）

表 4. 補助ラベル

ラベル	意味
<cl>	破裂音・破擦音中の閉鎖区間
<pz>	ポーズ
<pr>	preaspiration
<ep>	挿入母音
#	非発話区間

４．おわりに

筆者らが現在構築中のモンゴル語アクセントデータベースについて報告した。今後、収録人数を増やし、音声アノテーションを進める予定である。アノテーションされた音声がある程度の量確保された段階で、種々の音響特徴量を計測する。

謝　辞

本研究にコメントをいただいた東京学芸大学の斎藤純男教授に感謝します。また来日留学している内モンゴル人発話者に感謝します。本研究は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」の成果である。また、本研究の一部は、公益財団法人 博報児童教育振興会 第11回「国際日本研究フェローシップ」の助成を受けている。

文　献

Baoyuzhu and Mengheboyan (2011)『現代モンゴル語チャハル方言の音韻研究』（中国語）民族出版社.

Bayarmend (1997)「バーリン、ホルチン方言の語アクセントについて」（モンゴル語）『モンゴル言語』.

Boersma, Paul and David Weenink (2017) *Praat: doing phonetics by computer* [Computer program]. Version 6.0.24, retrieved 23 January 2017 from http://www.praat.org/

Choijingjap (1993)「モンゴル語の語アクセントについて」（モンゴル語）『内蒙古大学学報』.

藤本雅子・菊池英明・前川喜久雄 (2006)「分節音情報」『日本語話し言葉コーパスの構築法』国立国語研究所, pp.323-346.（pj.ninjal.ac.jp/corpus_center/csj/k-report-f/06.pdf よりダウンロード可能）

Huhe (2001)「モンゴル語のプロミネンスの問題」（モンゴル語）『内蒙古大学学報』.

Huhe (2007)「モンゴル語の語アクセント問題」（中国語）『民族語文』.

Huhe (2009)『モンゴル語の語音実験研究』（中国語）遼寧民族出版社.

Huhe (2014)「モンゴル語の語アクセント問題再論」（中国語）『民族語文』.

Huhe, Chenjia you, and Zheng yu ling (2001)「モンゴル語韻律特徴声学パラメーターデーターベース」（中国語）『内蒙古大学学報』.

Jan-Olof Svantesson, Anna Tsendina, Anastasia karlsson, and Vivan Franzen (2005) *The Phonology of Mongolian*, Oxford University press.

Sh.Lvbsanwandan (1986)『現代モンゴル語の構造』（モンゴル語）内蒙古教育出版社.

Unir and Yu rong (2015)「モンゴル語の語アクセントの研究概況」（モンゴル語）『モンゴル言語』.

多重の読みを持つテキストのコーパス化

小木曽 智信（国立国語研究所言語変化研究領域）

Making corpus of Japanese text including multiple readings

Toshinobu Ogiso (NINJAL)

要旨　日本語のテキストには，本文漢字の通常の読みを示すのではない特殊な読みをもつ振り仮名（たとえば「強敵」と書いて「とも」とふりがなを振る類）や，掛詞（「ながめ」を「眺め」「長雨」の両用に読む類から，語形の一部から別の語を連想させる類まで），各種の洒落など，意図的に多重の読みを持たされたテキストが少なくない。従来のコーパスではこのような多重の読みは切り捨てられ，選択されたただ一つの読みを配置することが多かった。本発表では，このような多重の読みを持つテキストについて，主として『日本語歴史コーパス』の事例を整理して示すとともに，そのあるべきコーパスアノテーションの方法について論じる。

１．はじめに

テキストが多重の読みを持つと言うとき、まず想像されるのはその解釈の曖昧性かもしれない。たとえば、古典読解における文学的・文献学的なテキスト解釈の曖昧性の問題、自然言語処理における形態素解析や統語解析、先行詞の同定の曖昧性などがあげられようか。ここで言う読みは、発音形の表示というレベルから解釈のありかたまで多様である。これら多重の読みは、一種の「謎」として書き手によって残されることもありうるが、通常は唯一の解が定まっているものであって、多重の読みがあるとしても意図的に仕掛けられたものではない。これに対して、洒落や掛詞のように、意図的に多重の読みが仕込まれたテキストがある。複数の読みは形式上必ずしも明示的ではないが、複数の読みを持つこと自体に一定の価値を置くテキストである。また、ルビによって本文とはちがう別の読みが明示されている場合もある。

これまでに構築された日本語コーパスにおいては、以上のような各種の多様な読みを持つテキストであっても、原則として一つの読みだけが選択され、他の可能性は捨象されてきた。形態論情報などの言語情報アノテーションは一つの読みについてのみ行われている。しかし本来であれば、多重の読みを持つこと自体に価値があったり、はっきりと多重の読みが示されたりするテキストについては、コーパス化においても、その点への配慮が欠かせないはずである。本稿ではこのような問題意識の下、多重の読みを持つテキストとして、まずは書き手によって意図的に残されたもので、かつ、語や句を単位としたものを取り扱う。具体的には、漢字の固定的な音訓以外の読みを表示するルビと、掛詞・洒落の例である。多重の読みをどうしても扱わなければ済まないこのような場合を例として、コーパスにおいて多重の読みを取り扱う方法について検討したい。

２．自由ルビ

ルビは通常、本文のフリガナとして用いられる。フリガナは、本文の読みを一意に示すことに主眼があるのであって、本来は読みの曖昧性（＝多重の読みの可能性）を抑制するものである。それは漢字の音訓の表示に留まらず、「時雨」のような熟字訓であっても変わらな

い。しかし、この用法を逸脱して、本文の場面・文脈に即した説明的な読みを表示したり、逆にルビの読みの意味説明を親文字が行ったりするタイプのものがある。親文字とルビとの関係が一般的な慣習を離れて、ルビが自由に付与されているという観点から、本稿ではこうしたものを「自由ルビ」と呼ぶことにする。自由ルビと親文字との関係は多種多様のものがあり、混質的である。こうしたものは『現代日本語書き言葉均衡コーパス』にも次のように数多く見られる。

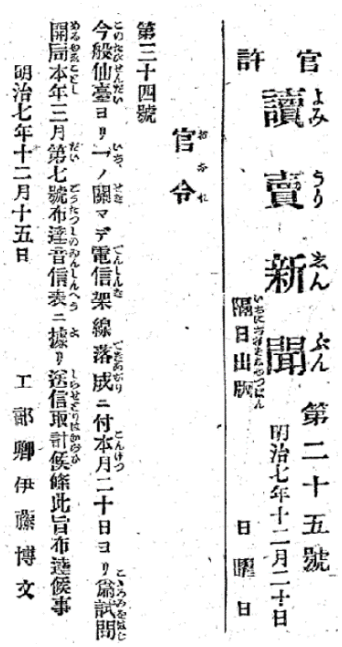
公権力　所有物　連帯責任　超魔　食べたい　ソウルフード　女主人　仮面の男
冥界の神　学際的研究　共体験　建国の父　女に磨きをかけたい症候群
オクスフォード英語辞典　お入り　海蛮人　海蛮人　海蛮人

しかしこれらは臨時的であり、頻度も必ずしも高くはない。

一方、近世・近代の資料にはシステムティックに自由ルビが使われる資料がある。一つのタイプは、図1に示す讀賣新聞のように、「隔日」に「いちにちおき」、「官令」に「おふれ」、「今般」に「このたび」、「落成」に「できあがり」とルビを振るように、難しい漢語に平易な日常語の読みを付けて理解を助けるものである。固い文語文と、日常語のルビによる平易な読みの二重のテキストとなっている。

もう一つのタイプは大部分が話し言葉の台詞からなり、ルビでその台詞の音形を示しつつ意味を本文の漢字列によって示すものである。図2に『太陽』所収の近代の作品の例を挙げる。「誤魔化す」「鳥渡」などは当時通行のフリガナともいえるが、「騙取し」「隔意」などは「騙取」「隔意」の漢語を本文にあて漢字によって意味を表示したものと思われる。このタイプのテキストは近世からあり、洒落本（市村・村山 2017）、人情本（藤本ほか 2017）のコーパス化において問題となったルビがそれぞれ例示されている。特に人情本においてはこの種のルビが多い。

二つのタイプの前者は固い書き言葉の本文に対しルビによって日常語の読みを説明的に示したものであり、後者は、平易な話し言葉のテキストをルビで表し、本文の漢字で意味を説明的に示したものであって、両者は対蹠的な位置にある。


官許 讀賣新聞 第二十五號
明治七年十二月二十日
隔日出版
日曜日

官令

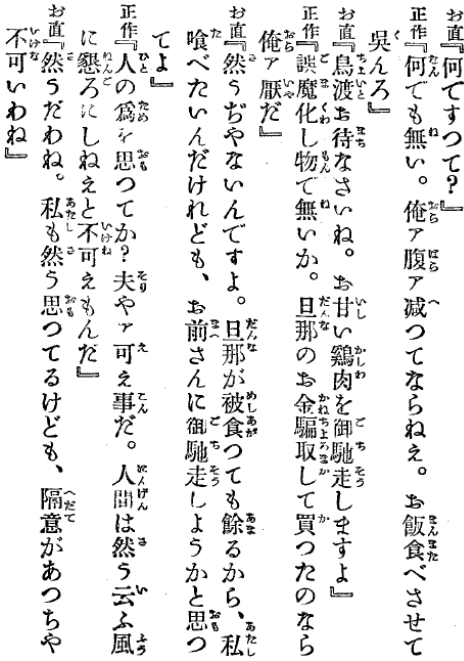
第二十四號

今般仙臺ヨリ一ノ關マデ電信架線落成ニ付本月二十日ヨリ爲試開
開局本年三月第七號布達音信表ニ據リ逓信取扱候條此旨布達候事

明治七年十二月十五日

工部卿伊藤博文

図 1　讀賣新聞 明治 7 年 12 月 20 日より


だが嘘突きだ。不可えなア』

お直『何ですつて？』

正作『何でも無い。俺ア腹ア減つてならねえ。お飯食べさせて呉んろ』

お直『鳥渡お待なさいね。お甘い鷄肉を御馳走しますよ』

正作『誤魔化し物で無いか。旦那のお金騙取して買つたのなら俺ア厭だ』

お直『然うぢやないんですよ。旦那が被食つても餘るから、私喰べたいんだけれども、お前さんに御馳走しようかと思つてよ』

正作『人の爲を思つてか？夫やア可え事だ。人間は然う云ふ風に懇ろにしねえと不可えもんだ』

お直『然うだわね。私も然う思つてるけども、隔意があつちや不可いわね』

図 2　田口掬汀「喜劇　嘘の世界」『太陽』明治 42（1909）年 12 号,p.108 より

コーパス化にあたり、讀賣新聞のような二重の本文はどちらか一方の読みを取っただけでは不足するし、人情本のようなタイプでも本文の漢語は用例として検索可能にするべきである。やはりルビと本文の二重の読みをコーパスで適切に扱う必要性が理解されよう。なお、近世・近代のテキストでは右だけでなく左側にルビが付されることがあり、その場合には三重の読みが重なる可能性もある。

3．掛詞・洒落とシンタグム

掛詞や洒落は、いずれも語形の一致や類似をもととしてもう一つの語としての読みをイメージさせるもので、その点では同じ構造を持っている。ここでは、韻文の掛詞を中心に見ていくことにする。掛詞や洒落の多重の読みは本文に内在するものといえ明示はされないが、音形の類似から多重の読みを可能にするために仮名書きされるなど、別の読みを喚起するための工夫が行われる場合もある。有名な和歌をもとに、掛詞のタイプについて確認しておきたい。次の歌は「ふる」が「古る/振る」、「ながめ」が「眺め/長雨」の二重に読まれる例である。二つの掛詞は関連するが統語関係までは持たず、個々の語が 2 重の意味を持つものとして扱える。

a. 花の色は移りにけりないたづらにわが身世に[ふる] [ながめ]せしまに（古今 113）

次は、「いなば」が「去なば/稲羽」、「まつ」が「待つ/松」 の二重に読まれる例であるが、二つの読みを介在して別の統語的なつながりが成り立っている。「立ち別れ去なば」と「稲羽の山の峰におふる松」は別の文であり、掛詞「いなば」が両者を仲介しているのである。

b. 立ち別れ[いなば]の山の峰におふる[松]とし聞かばいま帰り来む（古今 365）

掛詞は和歌に限られない。次に示すのは、現在コーパス化が進む近松の浄瑠璃（上野 2016）における例である。「なつ」が「夏」であると同時にその一部が「無」の掛詞となっている。さらに、「をりは」は「折羽(双六)」と「降り端」、「こひ目」は「乞い目」と「恋目」の二重になる。ここでも、「な」を介在にして別の文に連なっていく。

c. めぐれば。罪も[な][つ]の雲、あつくろしとて、駕籠をはや。[をりは]の[こひ目]、
（曽根崎心中 p.15）

最後は、洒落本に見られた洒落の例である。「良し」と「吉野」の二重の読みとなっている。

d. 何さ、こゝが[よし]の葛さ（傾城買四十八手 p.109）

以上の例からわかるとおり、多重の読みを持つテキストをシンタグマティックな関係として見たとき、単純な直線的な配列ではあらわせない。a.のように別の語をイメージ喚起するだけであれば、当該部分に二重の形態論情報を付与すれば済む。しかし、b. c.のような例では、イメージされたもう一つの読みを契機に別の文に乗り換える（場合によってはその語また元の文の続きに戻る）ことになるため、二つの意味の主従関係が、前文と後文で入れ替わることになる。d.の洒落は後文が続かないが、b. c.の前半と同じ構造である。

自由ルビの場合、多重化しても基本的には文の範囲は同じであるのに対し、係り結びや洒落では、このような複線的な関係が現れるため、単純に形態論情報を二重化が必要なだけでは十分でなく、文境界や統語関係のアノテーションにおいて問題を生じることになる。

４.多重の読みとコーパス化

従来の国語研究所のコーパスでは多重の読みのうちのただ一つの読みが選択されてきた。たとえば、『現代日本語書き言葉均衡コーパス』では、形態論情報が付されるのは本文文字列に対してであり、ルビはタグとしては付与されるものの形態論情報付与の対象とされていない。その反対に、『日本語歴史コーパス』の試作版として公開されている洒落本のコーパスでは自由ルビについては本文とルビを入替え、元のルビを形態論情報付与の対象とする一方で、元の本文には形態論情報が付与されていない。このような取り扱いは故あってのことではあるが、テキストが持つ重要な情報をすくい上げられていない点でやはり不十分と言わざるを得ない。

現在構築が進む『日本語歴史コーパス』では、本稿で取り上げたような近世・近代資料のコーパス化に取り組んでいるため、多重の読みを適切に扱うことが欠くことのできないこととなってきた。そこで当面の対応として、形態論情報データベース（小木曽・中村 2014）を拡張し、本文文字列に対して多重の形態論情報を付与できるようにした。文字単位で、異なる範囲にまたがる形で多重に情報を付けることが可能である。これにより、自由ルビや掛詞・洒落について、形態論情報のレベルでは対応が可能になった。今後、これを活用して和歌や近世・近代資料のコーパス構築を進める予定である。それでも掛詞の文の二重性については十分に扱えておらず、今後の課題である。

５．おわりに

本稿では、多重の読みを持つテキストの一部としてルビや掛詞・洒落を例示し、コーパス化する場合の課題について見た。その上で、限定的ながら一つの本文に多重の形態論情報を付与することでこの問題に対処した。

将来的には、冒頭に述べたような解釈の曖昧性も含めて、解が一つに定められない場合には複数の読みをそのままコーパスに格納できるようにすることも求められるだろう。コーパスでテキストの多重の読みを扱おうとする試みは緒に就いたばかりである。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「通時コーパスの構築と日本語史研究の進展開」および科研費基盤(A)「日本語歴史コーパスの多層的拡張による精密化とその活用」による成果の一部である。

文　献

小木曽智信・中村壮範（2014）．『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用，自然言語処理, 21(2), pp.301-332.

上野左絵（2016）．近松浄瑠璃本のコーパス化―「語り」のテキストをどう扱うか,人文科学とコンピュータシンポジウム論文集 2016, pp. 25-30.

小木曽智信（2016）.『日本語歴史コーパス』の現状と展望，國語と國文學, 93(5), pp.72-85.

藤本灯，北﨑勇帆，市村太郎，岡部嘉幸，小木曽智信，高田智和（2017）.「人情本コーパス」の設計と構築，国立国語研究所論集, 12, pp.1-12.

市村太郎，村山実和子（2017）. 洒落本コーパス構築の試行，国立国語研究所論集, 12, pp.29-45.

関連 URL

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

次元形容詞にみる母語話者らしい日本語形容詞の使用

西内 沙恵（国立国語研究所・立教大学）†

How native speakers of Japanese use adjectives
The case of dimensional adjective “takai”

Sae Nishiuchi (National Institute for Japanese Language and Linguistics and Rikkyo University)

要旨

日本語非母語話者は，形容詞用法の習得過程において形容動詞との活用の混同，時制の間違いなどを経るが，これらの文法規則こそが日本語形容詞使用における特性といえるか。本研究では，次元形容詞「高い」を題材にその構造と意味表出の関係を分析し，I-JAS で得られた日本語非母語話者の使用への観察から，日本語らしい使用の特性を明らかにする。

１．はじめに

日本語母語話者（以下 NS）も日本語非母語話者（以下 NNS）も，日本語の習得過程において（1）のような名詞修飾にノ格を挿入する誤用を経ることが知られている[1]。このほか，NNS の形容詞使用には（2）のような形容動詞との活用の混同や（3）のような時制の間違い，（4）のような活用の間違いが多く観察される。加えて，（5）のような意味の面での誤用も少なくない。また，これらの文法的な間違いが複合的に用いられることもある。これらはある時期を過ぎた NS にはみられなくなるが，NNS にはしばしば化石化し一定数みられ続ける誤用である。

(1) *甘いの物が好きです

【出典】 I-JAS サンプル ID：JJC14-I

(2) *父と母は忙しいだから〈はい〉誕生日のパーティは、も、していませんでした。

【出典】 I-JAS サンプル ID：JJC12-I

(3) *えー昨日はーちょっと忙しい｛笑｝、です、はい

【出典】 I-JAS サンプル ID：SES50-I

(4) *ファーストフードの中に栄養は少ないくて、・・・

【出典】 I-JAS サンプル ID：IID19-e

(5) ?古いの先生

【出典】 I-JAS サンプル ID：JJC28-I

しかし，上のような文法上の間違いがなくなれば，自然な日本語らしい発話になるだろうか。本研究では，『現代日本語書き言葉均衡コーパス（以下 BCCWJ）』で得られた実例をもとに次元形容詞の用法を考察する。さらに，I-JAS で得られた NNS と NS の発話を比べ

† snishiuchi@ninjal.ac.jp

1 ある女児の連体修飾用法の習得過程を観察した横山（1978）によれば，NS である女児 K は 628 日（1 歳 7 ヶ月）から（1）のような「[形容詞]の[名詞]」の誤用が観察され，1032 日（2 歳 8 ヶ月）に形容詞にノ格をつけた発話が観察されなくなった。

ながら構造の選択の差異を探り，日本語次元形容詞の使用における構造の特性を明らかにしたい。

２．次元形容詞「高い」の用法

形容詞を品詞として認める場合，いかなる言語でも次元性・年・評価・色彩の 4 つの意味の要素がとり出せる（Dixon 2004）。本研究では，類型論的に普遍的な意味の要素の一つである<次元性>に着目し，それを表す形容詞の一つ「高い」を扱う。

２.１　「高い」をめぐる先行研究

空間的な特性をその中心義とする形容詞の記述・分析は服部（1968），国広（1970），西尾（1972），久島（2001）など数多く，複数の次元形容詞の意味範疇が体系的にまとめられている。言語の経済性に則って，基本的な語ほど使用頻度が高く多義である（鈴木 2005）という性質から，基本的な次元形容詞が表出する意味も次元的な意味にとどまらない。

「高い」は「上方への距離が大きい（北原（編）2010: 1041）」という<次元性>のほか，温度などの<度合いの拡張性>や，評判や悪評など上方向に限らない方向性の尺度を示す<範囲の拡張性>，値段などを表す<価値の保有性>といった意味を持つ多義語である[2]。多義表出を動機づける要因は，名詞の性質と二重主語構文にみてとれる（西内 2016）。

(6)　a. ?彼女は高い。　　b. 彼女は背が高い。
(7)　a. その靴は高い。　　b. その靴はヒールが高い。

（筆者による作例）

（6a）は「彼女とデートすると，いい店に行かなければならないため高くつく」や「彼女がサービス業に従事していて，そのサービスを受けるために高額な料金を要する」など<価値の保有性>の読みは可能だ[3]が，<次元性>の読みはまずできない。同じように（7a）は「靴の値段が高い」という<価値の保有性>を意味する。彼女の「背の高さ」や靴の「ヒールの高さ」という<次元性>の表出には，「彼女」，「靴」が主題化によって示され，「背」，「ヒール」という<次元性>を受ける属性や部分がガ格で主格として指定される必要がある。では，どのような条件でこの制限が必要となるのか。

２.２　<次元性>を表出する「高い」の用法

BCCWJ で語彙素「高い」，その前方共起に書字形出現形「が」を指定し，短単位検索で抽出された 9779 件の実例を観察したところ，「高い」の<次元性>を表出する標識に被修飾名詞の場所的特性が関わっていることがみて取れた（西内 2016）。ただし固有名詞は対象への知識によって想起される意味が異なるため，観察対象は普通名詞に限定している。

２.２.１　場所的特性の観点

名詞の場所性は日本語文法において重要な問題として扱われ，名詞の場所性を文法的に検証するテストとして，寺村（1968）で（I）から（III）のーに名詞を当てはめる手法が編まれた。

(I)　ココハ–デス
(II)　–ヘイク
(III)　–デ～シタ

[2] 「低い」，「安い」などの非両立関係の反義語があることから多義語であると認定される（国広 1982）。
[3] 20 代の日本語母語話者 7 名を対象に，形容詞文の容認度と解釈を問うたインフォーマント調査による（2014 年 11 月実施）。

その後，場所性を有する名詞の意味の側面からの分析が，田窪（1984）において（A）から（E）にまとめられた。

(A) 「のところ」がつかない。
(B) 「どこ」で尋ねられる。
(C) 移動の goal，source になれる。
(D) 場所の状況語句「NP で」の NP になれる。
(E) 存在を表す文において「位置」を表す「NP に」の NP の位置に現れる。

さらに，森山（1988）で「「場所名詞」の「場所」が，何にとっての「場所」であるか（森山 1988: 176)」を観点に，場所名詞が（8）のように名詞の素性の問題として固定的に捉えられるものではなく，語用論的な意味における相対性への着目が必要であると考えられ，場所名詞の定義に下の（イ）から（ハ）の三段階を設定した。

(8) a. ペンは　デスク / デスクの上　にある。
　　b. 皿は　??デスク / デスクの上　にある。
（森山 1988：175）

(イ) 絶対的場所名詞：地名・集団・組織などの，どんな場合にも場所としてしか使われない，場所化されない名詞
(ロ) 相対的場所名詞：語用論的な相対性の観点が必要とされる「家」，「机」，「車」のような一定の空間を有する実質名詞
(ハ) 絶対的非場所名詞：一般的に場所性がない「青さ」，「正義」のような非実質的な名詞

（イ）でいう場所化とは，(9）のような「ところ」付与で「辺り一帯」の場所が取り上げられ，場所の限定が広くなる用法である。なお，これらの先行研究は対立する主張ではなく，森山（1988）の詳細な場所性の段階設定は，田窪（1984）に確立されたテストフレームに整合性があるものである。

(9) a. *猫に行く。　　b. 猫のところに行く。
（筆者による作例）

２.２.２　コーパスにみる「高い」の用法

本節では，森山（1988）の場所名詞の定義を参照しながら「高い」の実例をみてみる。例文中の下線は筆者による。分析対象の「高い」を一重線で，被修飾名詞を波線で示している。

(10) ステージが高過ぎて、センター席からは中央ステージすらよく見えないような状態。
【出典】 BCCWJ サンプル ID：OY04_03908 Yahoo!ブログ
(11) 空が高い。白い鳥も消えていた。
【出典】 BCCWJ サンプル ID：PB19_00658 川端裕人(著) 『ニコチアナ』

（10），（11）のように被修飾名詞が絶対的な場所性を有する場合，<次元性>の表出に主格による対象の明示を要していない。

(12) 鳥は垂直には飛べませんから、柵が高ければ逃げられなかったと思います。
【出典】 BCCWJ サンプル ID：OC12_00838 Yahoo!知恵袋
(13) 晴れて穏やかに見える伊勢湾も、外洋の波が高いときはそのウネリが湾内に入ってきます。

【出典】 BCCWJ サンプル ID：OY14_29732 Yahoo!ブログ

（12），（13）は絶対的な場所性ではないものの，一定の空間を有する相対的場所名詞であり，これらも<次元性>の表出に主格の補完を要さない。

(14) 眼は細いのだが、鼻が高いので、顔が引き締まって小さく見える。
【出典】 BCCWJ サンプル ID：LBj9_00218 曽野綾子(著) 『極北の光』

（14）は，主題化によって示されるはずの名詞，すなわち「眼」，「鼻」の所有者である「人」が表されていないものと考えると，主題に当たる名詞の「人」が場所性を有さないために<次元性>の表出に対象の明示が必要になっていると考えられる。

(15) 給湯床暖房とこの断熱法を併用するとより暖房効果が高くなります。
【出典】 BCCWJ サンプル ID：LBg5_00034 濱口和博(著) 『プロも見落とす家づくりの急所』)

(16) ホッキ貝（三百十五円）より赤貝が高いとはね。
【出典】 BCCWJ サンプル ID：OY03_01717 Yahoo!ブログ

（15），（16）のような場所性を帯びない名詞が被修飾名詞のとき，<次元性>以外，すなわち<価値の保有性>などの意味が想起されることがうかがえる。

(17) 諸事情があり実家には戻れない。選択肢として、１．アパートが高くても都内に住んだ方がいい。２．何があってもすぐかけつけられる
【出典】 BCCWJ サンプル ID：OC04_01022 Yahoo!知恵袋

（17）では，被修飾名詞が場所名詞であるのにも拘らず<次元性>が表出されない。これは，「アパート」が住む場所でもあり，所有の対象にもなりうる相対的場所名詞であるためだと考えられる。

以上，次元形容詞「高い」の<次元性>とそのほかの意味の表出の条件を，被修飾名詞の場所性と二重主語構文にみた。この構造を，暫定的に表１のようにまとめる。

表１　場所性を基準とした意味用法の区分

	場所的特性アリ	場所的特性ナシ
〈次元的意味〉	**直接修飾**	**ガ格の補完**
〈次元的意味〉以外	**相対的場所性なら直接修飾**	**直接修飾**

（18），（19）は，被修飾名詞が場所名詞ではないが，<次元性>を表出している用例である。これらへの分析は今後の課題としたい。

(18) ちょんまげ時代の人は、枕が高くてもちゃんと眠れたのでしょうか？
【出典】 BCCWJ サンプル ID：OC12_03936 Yahoo!知恵袋

(19) そうすると、その積み木が高くなるにつれ、不安定になって、そのうちには崩れますね。
【出典】 BCCWJ サンプル ID：OC12_06260 Yahoo!知恵袋

３．NS と NNS の形容詞使用の差異

ここまで，「高い」の実例からその用法をみてきた。では，はじめにみたようないわゆる文法規則の誤用のほかに，NS と NNS の間にはどのような使用の差異がみとめられるだろうか。

『多言語母語の日本語学習者横断コーパス（以下 I-JAS)』で語彙素「高い」を検索し得られた 373 件のうち叙述用法を観察した。対象となったのは，NNS である調査協力者（K）の使用のうち叙述用法 164 / 253 件[4]と，NS である調査協力者（K）と調査者（C）の叙述用法 8 / 17 件[5]である。

３.１　NNS の日本語形容詞の使用

NNS の「高い」の使用を観察したところ，前文で NNS 自身が用いた名詞や，調査者が発話した名詞を引き継ぎ，省略して「高い」を単独で用いる（21）から（26）のような例が 48 件と目立ってみられた[6]。なお，(20）のように助詞を省略したり，被修飾名詞と「高い」の間に調査者のあいづちを挟んだりしたものは数えていない。また，「高い」使用の直前に格の明示，とりたて助詞の使用，被修飾名詞を類推可能にする比較表現など，被修飾名詞の断定を可能にする形式が同文中にある用例も除いている。ここでも用例には，「高い」に下線を，被修飾名詞にあたる語に波線を引いている。

(20) C :「賑やかな田舎」K :「うん、と都会、近いからまあビルでも、ビルそんなに高くない、でも普通な」

【出典】 I-JAS サンプル ID : JJC28-I

（21）は，前文で用いた名詞を次に発話する文でも被修飾名詞として省略して用いている例である。「高い」は単独で用いられ，助詞句がない。

(21) K :「東京スカイツリー、ありました？」C :「あもう行ったの？」K :〈んー〉C :「えー」K :「でもー、高いすぎるた高すぎます｛笑｝」

【出典】 I-JAS サンプル ID : FFR27-I

（22），（23）は調査者の質問に現れた主格ないし述部にあたる名詞を，（24），（25）はデ格でとられた名詞を被修飾名詞として引き継ぎ，助詞句なしに「高い」で修飾している。

(22) C :「も関心が高いんあるんですか？」K :「んー」C :〈んー〉K :「そうですね」C :〈うーん〉K :「はいそうですもっと」C :〈うーん〉K :「高くてー〈うん〉もっと文化と〈うん〉文化といろいろなことを増やしたくて」

【出典】 I-JAS サンプル ID : EAU37-I

(23) C :「へー、高い山なんですか？」K :「ほんと高いですね〈ふーん〉あれ、空気がいいところ」

【出典】 I-JAS サンプル ID : JJC09-I

(24) C :「あるんですかね、お料理で」K :〈あー〉C :「何か」K :「有名なー」C :「料理？」

---

4 叙述用法以外の用法の内訳：修飾用法 81 件，連用用法 3 件，名詞用法「高さ」5 件

5 叙述用法以外の用法の内訳：修飾用法 9 件

6 同じ文脈で繰り返し使われた用例も数えている延語数である。

K：「後で有名で高い」C：「高い？」

【出典】 I-JAS サンプル ID：FFR08-I

(25) C：「よかったですね無事にタクシーで、友達に会えたんですね」K：「でもとても高かったですけど、とっても」C：「ああーそうなんですかふーん」

【出典】 I-JAS サンプル ID：HHG16-I

（26）は，調査者が発話し，また NNS 自身も前文で発話した対象を被修飾名詞として助詞句の明示なしに叙述している例である。興味深いのは，<次元性>と<価値の保有性>という異なる意味を立て続けに同一の語「高い」で表し，それに対して NS の調査者が確認している点である。<次元性>を太線で，<価値の保有性>を細線で示す。

(26) C：「スキーの有名な場所」K：「あーはいコーショベール」C：〈ふーん〉K：「はい、あーあー高いところです、」C：〈あーそうですかー〉K：「そこでとても高いです」C：「たとても高いっていう意味は、山の」K：「あーいえいえあー価格は」C：「価格が？」K：「あ高いです」C：「高いとこ？そうなんですか？どうして価格が高いんだろう」

【出典】 I-JAS サンプル ID：FFR17-I

データ種別の内訳は，発話データが 120 件，作文データが 44 件であった。データ種とタスク種別に被修飾名詞の有無を表 2 にまとめた。表 2 中の左側に実測値，行と列の割合から計算される期待値との差を右側に記す。また，期待値との差の大きさを表中にデータバーで示している。対話のタスクで現れた発話データで，被修飾名詞が明示されない使用が多かったことが読みとれる。

表２　データ種別にみる被修飾名詞の有無の期待値差

		被修飾語の明示アリ		被修飾語の明示ナシ		合計
データ種	タスク種	実測値	/期待値との差	実測値	/期待値との差	
発話データ	対話	68	−14.0	48	14.0	116
	RP1	2	0.6	0	−0.6	2
	絵描写	2	0.6	0	−0.6	2
作文データ	メール1	1	0.3	0	−0.3	1
	エッセイ	43	12.6	0	−12.6	43
総計		116		48		164

３.２　NS の日本語形容詞の使用

日本語を母語とする調査協力者の叙述用法 8 件を観察すると，作文データ，発話データのいずれでも助詞句の明示や比較表現の使用など，何らかの方法で被修飾名詞が特定される形式が用いられていた。（27）では，「私」の声の<度合いの拡張性>が表されている。

(27) C：「高音も、ちゃんと出るんですか？」K：「あー、私どっちかっていうと、その、高いほうで〈あー〉、低いのが出ないんですよ」

【出典】 I-JAS サンプル ID：JJJ15-I

４．まとめ

I-JAS で得られた NNS と NS の産出を比較したところ，NNS に文脈依存的な発話が目立つ一方で，NS の産出では被修飾名詞を特定可能にする形式が用いられていた。

日本語では，文脈で復元可能であれば結束性が維持されるため，助詞句全体の省略が可能である。にも拘らず，用例数が少ないことを差し引いても，NS に文脈依存による助詞句などの非明示が選択されていないことは興味深い。被修飾名詞及び類推を可能にする要素の明示がなくとも不自然に感じられない例を作ることは難しくなく，また例にみてきた NNS の使用は文法的におかしくないが，実際の NS の使用にはみられなかった。

使用に個人差がある「裸のハ」の出現の分布や機能の研究が発展させられている。話し言葉において，助詞句は削除されても結束性が保たれる。「裸のハ」が出現する根拠は結束性の問題ではなく，確信がない，躊躇しているといった話し手の心理が反映されている感動詞類的な振る舞いによるという見方（有田 2009，2015）や，話者間で共同して主題と解説の構造を作り上げる「コラボレーション発話行為文（三原 2016)」といった分析がなされている。NNS が対話者である調査者が発話した語を引き継ぐのは，語用論的な現象に関与しているものと思われる。一方，NS が被修飾名詞を明示するのは，形容詞の多義表出を担う文法構造によるものと考えられる。

日本語では，英語の‘high’も‘expensive’も「高い」で表すという程度の多義の認識から，その文法用法に着目されず, NNS と NS の間に使用の差異がみられたのではないだろうか。助詞句の明示が，NNS と比べて NS の使用に特徴的であることから，意味の表出に関与する「高い」の文法用法は，多義の使用での振る舞いに根ざすものであることが示唆される。

文　献

有田節子 (2009).「「裸のハ」についての覚え書き」『日本語研究センター報告』16, pp.95-107.
有田節子 (2015).「日本語疑問文の応答の冒頭に現れる「は」について：係助詞から感動詞へ」『国立国語研究所論集』9, pp.1-22.
北原保雄 (2010[2]).『明鏡国語辞典』大修館書店
久島茂 (2001).『《物》と《場所》の対立－知覚語彙の意味体系－』くろしお出版
国広哲弥 (1970).「日本語次元形容詞の体系」『言語の科学』2, pp.13-26.
国広哲弥 (1982).『意味論の方法』大修館書店
迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子 (2016).「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』6:3, pp.93-110.
鈴木智美 (2005).「多義の構造」日本語教育学会(編) 『新版日本語教育事典』pp.271-273.
田窪行則 (1984).「現代日本語の「場所」を表す名詞類について」『日本語・日本文化』12, pp.89-115.
寺村秀夫 (1968)「日本語名詞の下位分類」『日本語教育』12, pp.42-57.
西内沙恵 (2016).「現代日本語における知覚形容詞「高い」の意味基準に関する一考察－場所名詞の観点から－」『日本認知言語学会論文集』16, pp.467-473.
西尾寅弥 (1972).『形容詞の意味・用法の記述的研究』（国立国語研究所報告 44）国立国語研究所
服部四郎 (1968).「高イ，低イと high, tall; low, short」『英語基礎語彙の研究　ELEC 言語叢書』pp.119-124. 三省堂
三原健一 (2016).「コラボレーション発話行為文としての「裸のハ」構文」『日本語文法学会第 17 回予稿集』pp.89-94.
森山卓郎 (1988).『日本語動詞述語文の研究』明治書院
横山正幸 (1978).「幼児初期における連体修飾規則の習得過程」『Annual convention of the Japanese Association of Educational Psychology』20, pp.68-69.
Dixon, R. M. W. (2004) “Adjective classes in typological perspective”. In Dixon and Aikhenvald (eds.) *Adjective classes: A cross-linguistic typology*, pp.1-49. Oxford U.P.

関連 URL

コーパス検索アプリケーション『中納言』『多言語母語の日本語学習者横断コーパス』
https://chunagon.ninjal.ac.jp/

日本語コーパスの包括的検索環境の実現に向けて

前川 喜久雄（国立国語研究所音声言語研究領域）†
浅原 正幸（国立国語研究所コーパス開発センター）
小木曽 智信（国立国語研究所言語変化研究領域）
小磯 花絵（国立国語研究所音声言語研究領域）
木部 暢子（国立国語研究所言語変異研究領域）
迫田 久美子(国立国語研究所日本語教育研究領域客員教授)

Toward the Realization of a Comprehensive Searching Environment for Japanese Corpora

Kikuo Maekawa , Masayuki Asahara, Toshinobu Ogiso, Hanae Koiso, Nobuko Kibe, and Kumiko Sakoda (NINJAL)

要旨　国立国語研究所コーパス開発センターでは、従来個別に開発・提供されてきた各種日本語コーパスの検索環境を統合し、複数のコーパスを横断的に検索可能な包括的検索環境を整備する計画を進めている。既に公開済みのコーパス群だけでなく、第３期中期計画期間に種々の研究プロジェクトで開発ないし拡張を予定しているコーパス群の一部も検索対象に含める。本発表では，検索対象となる予定のコーパスを紹介した後に包括的検索環境の実現に向けてどのような問題があるかを検討し、解決の方向性を探る。

１．はじめに

国立国語研究所が当時未開拓であった日本語言語資源の整備事業に着手したのは1990年代末であった。その後、一連の事業で開発した種々の日本語コーパスは、幸い、国内外において幅広い研究領域の研究者の支持を集めることとなり、現在では言語資源整備が国立国語研究所の中核的な事業のひとつとして社会的に認知されるに至っている。

しかしながら、これまでに公開してきた各種コーパスは、それぞれ独立に検索系が開発されており、複数のコーパスを横断的に検索することができない点に運用上の制約が認められる。現在、広く利用されているコーパス検索用ウェブアプリ『中納言』も検索対象のコーパスごとに異なるバージョンを提供している。検索ロジックはほぼ同一だが検索に利用できる情報の選択肢はコーパスごとに異なっている（2節参照）。

そこで、2016 年度から 2021 年度にわたる第３期中期計画期間におけるコーパス開発センターの目標設定に際して、この問題の解消を主要な活動目標として設定することにした。この目標を達成することで、時間的、地理的変異を含む日本語コーパスが出現し、研究所がこれまでに進めてきた日本語言語資源の整備事業を一端集大成することができると考えている。

以下、２節では包括的検索環境の対象とする予定の一連のコーパスの仕様を紹介する。その後、３節で仕様にどのような問題があるかを検討した後、４節で今後どのような課題を解決する必要があるかを検討し、現時点で考えられる対応策について論じる。

† kikuo@ninjal.ac.jp

２．対象となるコーパス群

２.１　公開済みのコーパス群

最初に既に構築が終了するか、ある程度まとまった規模に達していて、国立国語研究所コーパス開発センターから公開されているコーパス群を観点に紹介する。

２.１.１　『日本語話し言葉コーパス』(略称 CSJ)

現代の標準日本語話者の自発音声コーパスである（Maekawa et al. 2000, 小磯編 2015)。規模は短単位で 752 万語。時間にして 650 時間の音声を収録している。音声認識での利用（すなわち言語モデルと音響モデルの構築）を念頭において設計されているので、内容の95%は独話である。具体的には各種学会での口頭発表と日常的な話題についての一般的なスピーチ（模擬講演）が大部分を占める。残る 5%は、独話と比較するために対話音声と朗読音声に充てられている。

アノテーションとしては、各種のタグが付与された音声の転記テキスト（発音形と基本形の 2 種類、転記単位ごとの音声信号との時間アライメント情報を含む。4.2.1 参照)、短単位と長単位による二重形態論情報、節境界ラベル等を提供している。またコアと呼ばれるサブセット（50 万語、44 時間）に対しては、X-JToBI 方式による分節音・イントネーション情報、文節係り受け構造、談話境界情報なども提供されている。コアに含まれるサンプルの形態論情報は手作業で精度を向上させている。メタ情報として、講演種別の他に、話者の属性情報（性別、年代、出身地など）を提要している。

CSJ は 2004 年の一般公開以来、DVD（第 4 刷からは USB メモリ）で頒布されている。専用の検索系は公開していないが、2011 年には、コア部分のすべてのアノテーションが RDB (SQLite)で利用可能になり、DVD 版ユーザーには無償で提供されている。また 2016 年には、コーパス全体の短単位形態論情報が『中納言』（次節参照）で検索可能になった。現在は DVD 版のユーザーのみを対象とした試験公開であるが、近日中に一般公開(無償、要登録)も開始する予定である。

２.１.２『現代日本語書き言葉均衡コーパス』（略称 BCCWJ）

現代日本語の書き言葉を対象とした均衡コーパスで、規模は短単位で 1 億語である。書籍・雑誌・新聞・広報誌・ブログ・ネット掲示板・国会会議録・法律・詩歌など多様なレジスターから抽出されたサンプルから構成されており、サンプルはすべて著作権処理済みである（Maekawa et al. 2014, 山崎編 2014)。

アノテーションとして最も重要なのは、長短両単位による形態論情報である。コア（100 万語）に含まれるサンプルの形態論情報は精度が高い。他に、文字・表記に関するタグと文書構造に関するタグも提供されている。前者にはルビ文字列、原文の誤表記、外字などの情報が、後者には「記事＞クラスター＞段落＞文」のような文書の階層構造、図表、引用、注記などの情報が含まれるが、提供されるタグの範囲はレジスターによる異動がある。メタ情報として豊富な書誌情報を提供しているのも特徴である。筆者属性のほか、原本のタイトル、巻号、出版社、出版者、ISBN、サンプル抽出位置などが提供されている。

2010 年以来、DVD 版で全データを頒布しているが、他にウェブ上で 2 種類の検索系を無償公開している。『少納言』ではユーザー登録なしに全テキストの文字列検索が可能であり、正規表現も部分的に利用できる。検索結果は書誌情報の一部とともに表示される。1 検索に対するヒット数が 500 を超える場合は、全検索結果から無作為抽出された 500 サンプルだけが画面に表示される。

『中納言』は形態論情報を検索するためのウェブインターフェースで、短単位ないし長単位の N グラム（N は 11 まで）を検索できるコンコーダンサーである。形態論情報としては、表層の文字列（書字形）の他に、語彙素(lemma)、語彙素読み、品詞（3 階層)、活用形、活用型などを指定できる。『中納言』では検索結果を上限 20 万サンプルまでダウンロードできるので、著作権保護の観点から、利用者登録をお願いしている。登録・利用は原則無償

である。
BCCWJ の公開後に作成され、公開されたアノテーション情報もある（関連 URL 参照）。文節係り受けアノテーション情報は、1 億語全体を自動解析したデータが提供されている。他に、述語項構造、述語項構造シソーラス、日本語フレームネット、時間情報・時間的順序関係、文体指標、節境界、拡張固有表現、単語係り受け構造、「れる・られる」の用法などのアノテーションが、コーパスの一部に対して提供されている。

２．１．３ 『太陽コーパス』

明治後期から大正期(1895～1925 年)の有名な総合雑誌『太陽』（博文館）から 5 年分を抽出した全文コーパスである（国立国語研究所編 2005)。2005 年の公開時には、タグ付きテキストコーパスとして頒布され、形態論情報は付与されていなかった。しかし、その後、近代語の自動形態素解析技術が実用に達したので、2016 年には短単位解析結果がウェブ上で公開された。検索系は『現代日本語書き言葉均衡コーパス』の項で紹介した『中納言』である。規模は短単位で 1100 万語（文字数で 1450 万語）である。今後は、同じく近代語を対象とした雑誌コーパス群（『近代女性雑誌コーパス』『明六雑誌コーパス』『国民之友コーパス』とともに、後述する『日本語歴史コーパス』「明治・大正編 I 雑誌」の一部を構成することになる。

２．１．４ 『日本語歴史コーパス』（略称 CHJ）

上代（奈良時代）から近代（明治・大正時代）までの日本語の歴史を通時的に研究するためのコーパスである（小木曽 2016)。2012 年より構築済みの部分から公開を開始し、現在では「平安時代編」（仮名文学 16 作品、約 86 万短単位）・鎌倉時代編 I 説話・随筆（5 作品、約 71 万短単位)、室町時代編 I 狂言（虎明本狂言集、約 24 万短単位)、明治・大正編 I 雑誌（上述の雑誌、約 1254 万短単位）が公開されている。BCCWJ と同様に短単位と長単位の二つの単位で形態論情報を付与しているが、現在のところ近世（江戸自体）以降のデータについては短単位のみである。残された貴重な資料を活用するため、「鎌倉時代編」の『今昔物語集（本朝部)』の一部と「明治・大正編」の雑誌の一部を除き、全体に人手による修正を施している。
検索インターフェースとして BCCWJ と共通の『中納言』によって公開を行っている。検索結果の各行から、外部のサービスにリンクがはられており、各作品の本文や原文の画像データなどが確認できるようになっている。たとえば、小学館の『新編日本古典文学全集』を底本とする作品はジャパンナレッジで公開されている当該ページにリンクがあり、本文・注釈・現代語訳を参照することができるほか、『今昔物語集』や近代雑誌では、原文の画像データが確認できる。

２．１．５『国語研日本語ウェブコーパス』（略称 NWJC）

BCCWJ の量的不足を補うためにウェブ上の日本語を母集団として構築された短単位 253 億語規模のウェブコーパスである。クローリング技術によって、約 1 億 URL の日本語ウェブページを繰り返し収集することで安定してアクセス可能なウェブページを決定した。公開データは、2014 年 10-12 月期に収集したデータである。
NWJC ではウェブ言語データの深刻な問題であるコピーサイトの問題を軽減するために、文単位の重複性排除を行っている。文単位の異なりを取ることによる文型パターンとしてのデータベース化を行っている。
現在提供されている形態論情報は UniDic 体系の短単位形態論情報のみである。また自動解析の結果をそのまま提供しており、CSJ や BCCWJ のように手作業で修正したサブセット（コア）は NWJC には設定されていない
NWJC の特徴として、データ全体に文節係り受け構造自動解析結果が提供されている。
NWJC は、ウェブ上の新しい検索系である『梵天』を用いて検索する。『梵天』には、文

字列検索、品詞列検索（形態論情報検索＝『中納言』の短単位検索と同等）にくわえて、係り受け検索の機能も実装されている。検索系『梵天』の実装においては 1 文中に同一語彙素が 2 回以上出現する場合、ヒットするのは最左用例だけという制約がある。ただし正確な頻度情報を必要とするユーザーには別途作成した語彙表を提供する。文字列検索機能は、前述の『少納言』と同様、ユーザー登録なしで一般公開されているが、システムに高い負荷がかかる品詞列検索ないし係り受け検索を行うユーザーには、事前に講習会を受講してもらっている。登録・利用は無償である。

２.１.６『多言語母語の日本語学習者横断コーパス』（略称 I-JAS）

I-JAS（International Corpus of Japanese as a Second Language）は日本語学習者のコーパスである。12 の異なった言語（英語、中国語、韓国語、インドネシア語、ロシア語、タイ語、フランス語、スペイン語、ドイツ語、ハンガリー語、ベトナム語、トルコ語）を母語とする日本語学習者の発話と作文のデータであり、海外の学習者と日本国内の学習者の両方が対象となっている。

これまでに公開されている日本語学習者コーパスの問題点をふまえて設計されており、多面的な角度から利用できるように、さまざまなタスクのデータと日本語母語話者のデータが含まれている。

2016 年 5 月、第一次データとして 12 言語を母語とする海外日本語学習者、国内の教室環境学習者と自然環境学習者各 15 名、日本語母語話者 15 名の計 225 名分を公開した。最終的には 2020 年春までに学習者 1000 名、日本語母語話者 50 名の計 1050 名分のデータ公開を目指している。

I-JAS の特徴としては、次の 5 点が挙げられる。(1) 既存のコーパスに比べ、規模が大きいこと。(2) データ内容が豊富であること。発話には 4 種類（ロールプレイ・ストーリーテリング・絵描写・対話）のタスク、作文は 3 種類（ストーリーライティング・メール・エッセイ）のタスクが設定されている。(3) 学習者全員が共通の日本語能力テストを受け、その評点が明示されていること。そのため、地域や機関が異なっていてもレベルの基準が統一され、比較が可能となる。(4) 多面的な利用が可能なデータ形式であること。発話データは書き起こしを行い、テキストだけでなく、検索システムの利用が可能である。また、発話の音声データも公開している。(5) 学習者の背景情報があること。全ての学習者の言語環境や学習歴などの情報が含まれている。

２.１.７『名大会話コーパス』

約 100 時間分の雑談を文字化したコーパスであり、姫路獨協大学（構築当時は名古屋大学）の大曾美恵子氏が構築されたコーパスである。一時期、国立国語研究所から『日本語自然会話書き起こしコーパス』の名称で公開されていたが、2016 年に短単位形態論情報を付与したデータを『中納言』で公開するにあたり、旧名称を復活させた。規模は短単位で 114 万語（句読点などを除く）である。

２.２　構築中のコーパス群

次に、現在構築作業が進行中のコーパス群を紹介する。前節に紹介したコーパスの拡張作業も含まれる。

２.２.１『日本語諸方言コーパス』(略称 CJD)

日本語諸方言の自然談話のコーパスである。北海道から沖縄まで全国の自然談話が横断的に検索でき、あわせて音声とテキストのダウンロードができる形で公開する。

資料としては、1977～1985 年に文化庁が行った「各地方言収集緊急調査」のデータを使用する。全体は、全都道府県 224 地点、1 地点につき 30 時間程度の談話録音テープよりなる資料で、内容は当時 60 歳以上の地元出身者数人による自然談話である。一部は 『全国方

言談話データベース 日本のふるさとことば集成』（国書刊行会）として音声、テキスト、標準語訳が公開されているが、多くは未公開の状態。本コーパスでは公開分、未公開分を合わせて、2021 年度までに最低 75 時間分のデータを公開する。

本コーパスの特徴は、諸方言の談話を標準語で検索し、それに対応する方言形とそれを含む一定の発話単位を横断的に検索する点にある。言うまでもなく、方言と標準語は 1 対 1 で対応しない。そのため、標準語での方言検索には対応のずれといった問題が生じる。しかし、方言形で検索システムを構築するには、各地方言の形態素辞書を作らなければならず、それには膨大な時間と労力が必要となる。それに対し、標準語での検索には、すでにある日本語形態素解析用辞書を利用することができ、しかも、方言間のゆるやかな横断検索が可能となる。また、諸方言コーパスがどのように利用されるかを考えてみると、標準語での検索システムは必須のように思われる。したがって、標準語による検索方法をとることとした。

本コーパスの構築に向けて、現在、次のような手順で作業を進めている。①方言音声の転記テキスト（方言テキスト）のチェック。②発話単位の認定。③方言テキストに対する時間アライメント情報の付与。④方言テキストに対応する標準語テキストのチェック。①の方言テキストと④の標準語テキストは、文化庁の事業の際にすでに作成されたものがあり、これをもとにして、チェック作業を進めているが、標準語テキストについては、全面的な見直しが必要である。前述のように、方言と標準語は 1 対 1 で対応するわけではないので、標準語テキストによっては、本来、検出されるべき方言形が検出されなかったり、検索結果が変わってきたりする可能性があるためである。標準語テキストの付け方については、作業の過程で詳細なマニュアルを作成しており、それもあわせて公開する予定である。②の発話単位の認定については、基本的に 0.2 秒の無音という基準で発話単位を認定しているが（ただし 3.2 および 4.2.1 も参照）、それに加え、話者同士の発話の重なりや相づち、フィラー、間投詞等に対し各種タグ付けをして公開する。

２.２.２『日本語日常会話コーパス』(略称 CEJC)

現代日本語の日常会話を対象とするコーパスで、規模は 200 時間（推定 200 万語相当）を目指す（小磯ほか 2017）。本コーパスは機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（平成 28～33 年度、リーダー：小磯花絵）において現在構築中のものであり、平成 33 年度末の公開を予定している。

本コーパスが対象とするのは、収録のために集められた状況での会話ではなく、日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話である。日常会話の幅広いレジスターをカバーするようサンプルを選ぶには、私たちが普段、どのような種類の会話をどの程度行っているかを把握する必要がある。そこで予備研究として約 250 人を対象とする会話行動調査を実施し、その結果を参考にしながら、多様な種類の会話をバランス良く納めたコーパスを構築する（小磯ほか 2016）。

会話データは、性別・年代の点から均衡性を考慮して選別された協力者 40～50 人に収録機材等を 2～3 ヶ月ほど貸し出し、協力者自身に日常会話 15～18 時間程度を収録してもらう方法を中心に集める。この中から、多様な種類・場面の会話となるよう、1 協力者あたり 4～5 時間を選別してコーパスに格納する。

アノテーションとしては、各種のタグが付与された音声の転記テキスト（転記単位ごとの音声信号との時間アライメント情報を含む）に加え、発話単位情報、短単位と長単位による二重形態論情報、文節係り受け情報などを付与する予定である。またコア部分（約 20 時間）に対しては、国際標準化規格 ISO24617-2 に基づき日常会話用に整備した談話行為情報や、CSJ に付与されている X-JToBI を簡略化した方式に基づくイントネーション情報を付与する予定である。コアに含まれるサンプルの形態論情報・係り受け情報は手作業で精度を向上させる。メタ情報として、話者の属性情報（性別、年代、出身地、相手との関係性など）や会話の属性情報（会話の場面、形式、人数など）を提要する。

２.２.３『日本語歴史コーパス』の拡張

『日本語歴史コーパス』は、この 3 月に「鎌倉時代編 II 日記・紀行」として、『とはずがたり』や『海道記』など 5 作品の追加公開を予定している。来年度以降は、すでに試行版を公開中の「江戸時代編」の洒落本・人情本を拡充して公開するほか、「奈良時代編 I 万葉集」「室町時代編 II キリシタン資料」の公開を予定している（いずれも 2017 年度予定)。『万葉集』やキリシタン資料では、万葉仮名やローマ字で書かれた原文と漢字仮名交じり本文とのアライメントをとり（４.２.２参照)、当該部分の原表記を確認できるようにする予定である。さらに、続日本紀宣命（奈良時代編)、近松の世話物浄瑠璃（江戸時代編)、国定読本などの教科書や近代文学作品（明治・大正編)、和歌集などのコーパス化を行い、上代から近代までの日本語を通時的に研究することのできるコーパスとする計画である。

２.２.４『多言語母語の日本語学習者横断コーパス』の拡張

2017 年春に第二次データとして、韓国語、中国語、英語、トルコ語の海外日本語学習者各 35 名、国内の環境別日本語学習者を各 25 名、日本語母語話者を 35 名、合計 225 名のデータを追加公開する予定である。これにより、日本語と言語的な類似点の多い韓国語とトルコ語、類似点の少ない中国語と英語のデータが各 50 名分となり、母語と学習者の日本語レベルの観点からの分析も容易になる。

また、韓国語、中国語、英語、フランス語については、各言語の母語話者同士での発話データの収集を計画している。I-JAS の日本語学習者に実施したタスクのうち、ロールプレイとストーリーテリング、メールに関して、母語のデータと比較することによって、母語と学習者言語（日本語）でのコミュニケーション上の問題についての研究が可能となる。

３．仕様の問題点

上に紹介したコーパス群を主にアノテーション仕様の観点から比較することで、問題の所在を明らかにすることを試みた。

３.１　形態論情報

現時点ですべての対象コーパスに付与されているという意味で、最も基本的なアノテーションは、短単位形態論情報である。日本語の膠着語的性格を考えると長単位での解析も行われていることが望ましいが、現在、両単位による二重解析が施されているのは CSJ と BCCWJ にとどまる。

同じ短単位と言っても、コーパスの開発時期によって、細部で仕様が異なることがある。CSJ と BCCWJ の間にも無視しえない差が生じていたが、現在、『中納言』でオンライン公開されている CSJ の短単位情報は、BCCWJ の規定に沿う形で統一が図られている。

また、そもそも形態素解析作業で何が解析されるかも対象コーパスによって異なる。CJD(『日本語諸方言コーパス』)の場合、形態素解析が施されるのは標準語テキストであり、方言テキストは解析対象ではない（ただし 4.3 も参照)。検索系はヒットした標準語テキストに対応する方言テキスト（と音声）を出力する。I-JAS の場合、形態素解析されるのは、いわゆる誤用を修正された日本語テキストであり、検索系は誤用を含む転記テキストを出力する（例えば、I-JAS で語彙素「経営」を検索すると、通常の「経営」以外に「けえ」「けいえ」「けいえん」「けいいん」などと転記されたサンプルが表示され、画面上にはそれらが「経営」を意図した発話であることがタグで表示される)。

３.２　発話単位

CSJ、I-JAS、CJD、名大会話コーパスなど、自発的な話し言葉をあつかうコーパスでは、発話単位をどう認定するかが問題になる。現在はコーパスごとにバラバラの状態にある。これをある程度まで企画して統一できるかどうかは緊急性の高い検討課題である。また独話と対話でも認定基準が異なってくる可能性があり、その点の検討も必要である。

CSJ のように物理的な基準（0.2 秒以上のポーズで区切るのが原則）に依拠すれば、作業基準は明確になるが、言語学的な意味づけは時に困難になる。

発話単位は、検索結果の音声再生の単位となることが予想される（4.2.1 参照）。その観点からは、極端に長いもしくは短い単位が頻出することは避けたいという要請もある。

３.３　タグ

話し言葉の転記テキストには、様々なタグが埋め込まれることが多い。対象コーパスのなかでは、CSJ と I-JAS で多数のタグセットが利用されており、CJD にも今後種々のタグが付与される予定である。

CSJ と I-JAS のタグは目的がかなり異なっており、前者では転記の正確性を高めることに主眼が置かれているのに対し、後者では、学習者のいわゆる「誤用」を含む日本語を修正して、形態素解析可能なテキストに整形することが主要な目的となっている（本ワークショップにおける西川の発表参照）。

書き言葉コーパスのテキストにもタグが付与されている。BCCWJ のタグには先に 2.1.2 節で触れた。CHJ のテキストには、本文校訂の情報の他、会話などの本文の種別、話者の情報などが付与されているが、両コーパス間でのタグの共通性は低い。CHJ ではタグの一部が『中納言』での検索対象の絞り込みに利用されているが、BCCWJ では現在そのような機能は提供されていない。今後、CHJ が万葉仮名やローマ字の文献を扱うようになれば、CHJ のタグはさらに増加するものと予想される。

３.４ その他のアノテーション

係り受け構造アノテーションは、CSJ と NWJC にだけ付与されている。両者の仕様には相違がある。韻律に関する X-JToBI アノテーションは CSJ のコアにだけ付与されている。I-JAS には上昇イントネーションを示すタグがある。CEJC にも句末イントネーションの機能に関するタグが付与される予定であり、またサブセット（コア）に対して X-JToBI 的なラベリングを施す予定がある（2.2.2 参照）。

４．議論

４.１ 包括的検索系と個別検索系

われわれは今後、上に述べたように、仕様に種々の異同をもつコーパス群を対象とした包括的検索環境を設計することになる。その際、もっとも基本的な決定のひとつは、新しい包括的検索系と従来から存在する個別検索系（『中納言』や『梵天』）の関係であり、ことに、包括的検索系が検索対象のデータをどのような方式で保持するかという問題である。

ひとつの方式は、既存対象コーパスのデータは適宜修正し、今後構築する対象コーパスのデータは当初から統一を図って、包括的検索系独自のデータを構築し、それを検索対象にするというものである。この場合、完成後に、従来の個別検索系（現在公開されている様々な『中納言』や『梵天』）をどうするかという問題が生じる。今後慎重な検討を要する問題であるが、ここでひとつの方針を述べるならば、包括的検索系において、各対象コーパスおよび対応する個別検索系の仕様の相違を十分に吸収できない場合は、少なくとも一定期間、包括的検索系と個別検索系とを併存させることにしたい。

もうひとつの方式として、包括的検索系を個別検索系に対するラッパー(wrapper)として設計する可能性も考えられる。包括的検索系は、個別検索系に対する検索リスエストを発行して、その結果を受けとり、それを適宜整形して出力するという形のシステムである。この場合、実際に検索を実施するのは個別検索系であるから、当然それらを維持しつづけることになり、結果として、検索系を再開発する困難を回避することができる。反対に、この方式の問題点としては、既存システムの出力を完全に整形して統一することが、おそらく非常に困難であろうことが挙げられる。

４.２　コーパス開発に対する技術的支援

対象コーパス群のうち、現在構築中のコーパスについては、構築作業を効率化する必要がある。コーパス開発センターでは、現在以下の二点について重点的に技術支援を行っている。

４.２.１ 音声-テキスト間アライメント

話し言葉コーパスのうち、音声ファイルを持ち、検索系を通して音声を再生しようとするコーパスでは、形態論情報等の検索対象となる転記テキストと、音声信号との時間アライメント（対応づけ）をとる必要がある。その可能性をもつのは、現状では、CSJ、I-JAS に加えて、CJD と CEJC である（名大会話コーパスは音声ファイルが公開されていない)。

既存コーパスの中でこの情報をもっとも組織的に付与できているのは CSJ であるが、CSJ の開発においては、音声・テキストアライメントは転記テキスト作成作業の一環として人手で実施した。当然、高いコストが発生した。

近年、音声認識技術の飛躍的な発展によって、この作業を全自動で実施する可能性が現実のものとなりはじめている。特に標準語音声が対象となるコーパスにおいては、自動アライメント技術を活用できる可能性が高い。一方、方言音声や学習者音声を、既存の技術でどこまで処理できるかは今後の検討を要する問題であり、基礎研究の課題である（本ワークショップにおける石本の発表参照)。

アライメントの単位は短単位などにも設定できるが、あまり短い単位を音声再生しても、知覚が困難になることも多いので、3.2 で論じた発話単位がひとつの現実的な候補であると考えられる。

４.２.２　二テキスト間アライメント

CJD では、標準語テキストを検索して、ヒットした短単位を含む標準語テキストに対応する方言テキストを出力する。そのためには、標準語テキストと方言テキストのアライメントが必要になる。方言コーパスは、その出発点となった文化庁のデータ（2.2.1 参照）が対訳形式で作成されているので、一応のアライメントはとれているのだが、今後発話単位の認定（３.２参照）などで現在のテキストを改めた場合には、アライメントの再実施が必要になる。

もうひとつ、テキストアライメント技術の応用が期待されるのが CHJ である。古典本文と現代語訳の対応がとれれば、CHJ の応用可能性が高まると期待される。また万葉仮名で書かれた文献や漢文文献の場合は、万葉仮名ないし漢文と読み下し文のアライメントが必須である。

４.３　共同研究

上記ふたつのアライメント処理技術は、音声認識や自然言語処理の研究のなかで発展してきた技術であり、現在もそこで最先端の技術が開発されつつある。その成果を効率的に取り込むために、2016 年 10 月から研究所外の研究者との共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」を実施している。

この共同研究では、アライメント処理技術以外に、意味処理技術、音声特徴自動抽出技術、教師無し言語解析技術などの共同研究を実施するか実施予定である。その成果は、各種コーパスに対する意味情報（分類語彙表番号）の付与（本ワークショップにおける加藤らの発表参照）や、音声のピッチ情報の抽出精度向上、声質(voice-quality)情報の自動付与（本ワークショップにおける森らの発表参照)、韻律情報アノテーションの部分的自動化、方言テキストの形態素解析などの形で、コーパス開発へのフィードバックを試みる予定である。

５．まとめ

本稿では、国立国語研究所コーパス開発センターで構築を進めている、複数の日本語コー

パスを包括的に検索可能なシステムについて、検索対象となる予定のコーパス群の仕様を手短に紹介するとともに、包括的検索システムの設計に関わる様々な問題を探索し、検討した。今後は、方向で指摘した個別的な問題群について技術開発の現状を報告すると同時に、全般的な開発状況についても、適宜報告していく予定である。

謝　辞

本稿は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」ですが、同時に、筆者らが過去 2 年間自主的に開催してきた SMOKKA 研究会の成果も反映されています。同研究会の参加者・発表者に深く感謝します。

文　献

Masayuki Asahara, Kazuya Kawahara, Yuya Takei, Hideto Masuoka, Yasuko Ohba, Yuki Torii, Toru Morii, Yuki Tanaka, Kikuo Maekawa, Sachi Kato and Hikari Konishi (2016). "'BonTen' Corpus Concordance System for "NINJAL Web Japanese Corpus"', Proceedings. of COLING-2016. Demo Session. (To Appear).

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato and Hikari Konishi (2014). "Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan", *Alexandria*, 25:1-2, pp.129-148.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara (2000). "Spontaneous Speech Corpus of Japanese", In Proceedings of LREC-2000 (Second International Conference on Language Resources and Evaluation), Vol. 2, pp.947-952.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Makiro Tanaka, and Yasuharu Den (2014). "Balanced Corpus of Contemporary Written Japanese", *Language Resources and Evaluation*, 48, pp.345-371.

小磯花絵 編（2015）『話し言葉コーパス:設計と構築』（講座日本語コーパス第 3 巻）朝倉書店.

小磯花絵・土屋智行・渡部涼子・横森大輔・相沢正夫・伝康晴(2016).「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」『国立国語研究所論集』Vol.10, pp.85-106.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉(2017).「『日本語日常会話コーパス』の構築」『言語処理学会第 23 回年次大会予稿集』

小木曽智信（2016).「『日本語歴史コーパス』の現状と展望」國語と國文學 93 (5), pp.72-85.

国立国語研究所編(2005)『雑誌「太陽」による確立期現代語の研究―「太陽コーパス」研究論文集―』（国立国語研究所報告 122）博文館新社刊.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016).「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』6:3, pp.93-110.

山崎誠(編)(2014)『書き言葉コーパス:設計と構築』（講座日本語コーパス第 2 巻）朝倉書店.

関連 URL

『日本語話し言葉コーパス』 http://pj.ninjal.ac.jp/corpus_center/csj/
『現代日本語書き言葉均衡コーパス』 http://pj.ninjal.ac.jp/corpus_center/bccwj/
『現代日本語書き言葉均衡コーパス』アノテーションデータ http://pj.ninjal.ac.jp/corpus_center/anno
『太陽コーパス』 http://pj.ninjal.ac.jp/corpus_center/cmj/taiyou/
『日本語歴史コーパス』 http://pj.ninjal.ac.jp/corpus_center/chj/
『国語研日本語ウェブコーパス』 http://pj.ninjal.ac.jp/corpus_center/nwjc/
『多言語母語の日本語学習者横断コーパス』https://ninjal-sakoda.sakura.ne.jp/lsaj/?cat=3
『日本語日常会話コーパス』https://www.ninjal.ac.jp/research/project-3/institute/spoken-language/
『少納言』(コーパス検索アプリケーション) http://www.kotonoha.gr.jp/shonagon/
『中納言』(コーパス検索アプリケーション) https://chunagon.ninjal.ac.jp/
『梵天』(『国語研日本語ウェブコーパス』検索系) http://bonten.ninjal.ac.jp/

機能語用例文データベース『はごろも』の今後の展開

堀恵子（東洋大学・筑波大学）
内丸裕佳子（岡山大学）
加藤恵梨（朝日大学）
小西円，山崎誠（国語研）
江田すみれ（日本女子大学）
建石始（神戸女学院大学）
中俣尚己（京都教育大学）
李在鎬（早稲田大学）

Future developments of the Japanese Grammatical Items Example Sentences Database and Searching System “HAGOROMO”

Keiko Hori(Toyo University, The University of Tsukuba)
Yukako Uchimaru(Okayama University)
Eri Kato(Asahi University)
Madoka Konishi, Makoto Yamazaki(National Institute for Japanese Language and Linguistics)
Sumire Goda(Japan Women’s University)
Hajime Tateishi(Kobe College)
Naoki Nakamata(Kyoto University of Education)
Jae-Ho Lee (Waseda University)

要旨

機能語用例文データベース『はごろも」』は，web 上で機能語の一部を検索すると，意味，難易度，作例，話し言葉と書き言葉の用例などが見られるツールで，2015 年秋から公開されている。利用者から一定の評価を得る一方，母語話者の用例だけではその項目を理解する上で難しいこともある，表現したい意味，機能に当てはまる項目群から，的確な項目を選べると産出に役立つ等の声を聞く。そこで，今後の改訂の方針として，わかりやすい用例を継続的に増やしていくことに加え，(1)見出し語を精査，(2)当該項目が文中でどの要素として働くかを明確に示す文法機能の情報をつける，(3)意味用法の記述を精査し，階層のある分類とする，(4)文法項目の前接の形式を明示する，(5)学習者作文コーパスなどから学習者レベルと，正用，誤用の文を示す，の５点を進めている。2017 年秋までに作業を終え，同年度末には改訂版を公開の予定である。

１．機能語用例文データベース『はごろも』とは

１．１　機能語用例文データベース『はごろも』の概要

機能語用例文データベース『はごろも』（以下，『はごろも』）は，web 上で機能語の一部を検索すると，項目の意味，見出し語と意味の英訳，主観判定による難易度（堀ほか 2012），旧日本語能力試験の出題基準（以下，旧『出題基準』）の級，典型例（作例），話し言葉と書き言葉のコーパスから抽出した的確な用例，参考資料のページが見られるツールで，2015 年秋から公開されている（堀ほか 2016）。また，それに先だって用例以外をダウンロード版として 2016 年 3 月に公開した。ここで「機能語」とは，日本語教育で扱われている文型，表現などを含む文法項目全般を表すものである。

文法項目は，以下の 5 種類の資料のうち，主に 2 種類の資料にある文法項目を採用した。項目数は 1,848 項目である。表 1 には，基とした資料と，そこから見出し語として採用した

項目数を示す。

(1) 国際交流基金・日本国際教育支援協会 (2002)『日本語能力試験出題基準[改訂版]』
(2) グループジャマシイ(1998)『日本語文型辞典』(以下,『文型辞典』)
(3) 国立国語研究所(1951)『現代語の助詞・助動詞』(以下,『助詞・助動詞』)
(4) 森田・松木(1989)『日本語表現文型』(以下,『表現文型』)
(5) 国立国語研究所(2001)『現代語複合辞用例集』(以下,『複合辞』)

表１　5 種の資料から取り入れた「はごろも」の項目数

旧『出題基準』	『文型辞典』	『助詞・助動詞』	『表現文型』	『複合辞』
956	1,479	413	597	346

また，用例を抽出したコーパスは，下記の通り書き言葉 4 種，話し言葉 4 種である。

書き言葉

(1) 『日英新聞記事対応付けデータ (JENAAD)』
(2) 『Kyoto University and NTT Blog コーパス(KNBC)』
(3) 『現代日本語書き言葉均衡コーパス(BCCWJ)』
(4) 『CASTEL/J　CD-ROM V1.5 』日本語教育支援システム研究会

話し言葉

(1) 『日本語会話データベース』平成 8-10 年度文部省科学研究費補助特定領域研究「人文科学とコンピュータ」公募研究（「日本語会話データベースの構築と談話分析」 研究代表者　上村隆一）の成果による
(2) 『宇都宮大学 パラ言語情報研究向け音声対話データベース (UUDB)』
(3) 『名大会話コーパス』科学研究費基盤研究(Ｂ)(２)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成 13 年度～15 年度，研究代表者：大曾美恵子）
(4) 『BTS による多言語話し言葉コーパス-日本語会話 1』宇佐美まゆみ監修(2005)東京外国語大学大学院地域文化研究科 21 世紀 COE プロジェクト「言語運用を基盤とする言語情報学拠点」

『はごろも』はスマートフォンにも対応しており，手軽にどこでも利用できる。また，Google の音声読み上げ機能を利用しているので，漢字が苦手な人も読むことができる。

１．２　『はごろも』開発の目的

『はごろも』は日本語教師の支援，特に海外の非母語話者教師支援と中上級以上の日本語学習者の支援を目的としている。用例に触れることで，文中での実際の使われ方，ニュアンス，表現形式による使用の偏りなどを理解することができると考える。2010 年に日本語能力試験が改定され，出題基準が非公開となった。そのため，どの文法項目をいつ教えたらよいかが分かりにくくなっている。そこで,『はごろも』では文法項目に難易度を付け，適切な学習段階の目安を示す（堀ほか 2012)。

また，中上級以上の学習者であれば，教師と同様に多くの用例に触れることによって項目の意味や使い方を理解する助けとなることを期待する。

１．３　利用状況

『はごろも』ダウンロード版は Excel 形式でダウンロードできるもので，日本語教師，研究者等の使用を想定している。2017 年 1 月 28 日現在のダウンロード数は 83 件である。

２．現在の問題点

『はごろも』に関して，これまで国内外での発表，デモンストレーションを行い（堀ほか 2016 など），利用者からの声を聞くことができた。それによると，文法項目の意味，難易度などの情報を手軽に知ることができる唯一のツールであり，学習者の文法学習だけでなく，読解などにも利用できるのではないかとの意見が聞かれた。しかしながら，次のような要望も聞かれた。

(1)　用例が少しで，項目についての詳細な情報がわかりにくいことがある。
(2)　意味の記述が簡単なので，もっとわかりやすくしてほしい。できれば，似ている項目との違いを明確にする解説例や，用例が欲しい。
(3)　作文などの産出活動のために，述べたい意味や表現の機能から，それを表す文法項目を探せるようになるとよい。
(4)　英語の見出し語，意味があるのはよいが，もっと多言語にしてほしい。

また，著者らが現在認識している問題点として，下記のような点がある。

(5)　活用による派生表現が異なる見出し語として上がっている。
(6)　見出し語に前接の形式を含めるか否か統一がなされていない。

３．改訂方針と作業計画・進捗

前章の問題点を解決するために，引き続きコーパスからの用例を加えることに加えて，以下の５点について改訂を行うことにした。

(1)見出し語を精査する
(2)当該項目の文中での機能を明確に示すために，文法カテゴリーの情報をつける
(3)意味の記述を精査し，階層のある分類とする
(4)文法項目の前接の形式を明示する
(5)学習者作文コーパスなどから学習者の正用，誤用の文を学習者のレベルと共に示す

以下では，その詳細について述べる。

３．１　見出し語の精査

見出し語は，基とした資料の項目立てと表記に沿ったものが多く，活用による派生表現が異なる項目となっているものがある。そこで，新たに派生語という表示項目を作成し，検索しやすい見出し語の下にまとめることとした。

例 1）見出し語「について」「についての」「につき」→見出し語「について」，派生語（についての，につき）

また，ある見出し語に「は」「も」という取り立て助詞が付く見出し語が別にあり，意味が「は」「も」を付加する以上の違いが認められない場合，それらの見出し語は削除し，派生語に含めるという扱いにした。

例 2）見出し語「については」「についても」→見出し語「について」の派生語とする

さらに，肯定形式に加えて否定形式も見出し項目にあがっている場合，肯定形式の否定という意味以上の意味が付加されない場合は削除した。

例 3）見出し語「ことがある」「ことはない」→見出し語「ことがある」のみとする

項目によっては，派生語，取り立て詞とは無関係であるが，概念としてまとめたほうが教育現場における扱い方から見てよいと判断したものは，一つの項目としたものもある。

例 4）見出し語「もっとも～が」「もっとも～けど」→見出し語「もっとも～〔逆接〕」

以上の作業によって，2016 年『はごろも』公開時に 1,848 項目であった見出し語は，2017 年 1 月末の作業の段階で 1,744 項目となった。今後も精査を続け，以下の節で述べる作業も同時に進め，2017 年度末に『はごろも』改訂版を web 上で公開する予定である。

３．２　文法カテゴリーに関する情報

これまでは見出し語が文中で果たす文法機能に関する情報は示していなかった。しかし，教師の文法への理解を深めることや教材，テスト作成などの支援のために，文法カテゴリーに関する情報は必要であると考え，付与することにした。

『はごろも』には，初級から上級までのさまざまな文法項目が含まれるため，文法項目には，語レベルだけでなく，文型，敬語，活用形など初級から上級までの多くの要素が含まれる。分類は，日本語学の立場と異なることもあるが，必ずしも日本語学に詳しくない日本語教師，日本語学習者の文法カテゴリーについての理解を促し，文の産出にも役立つ機能を示すことを優先する。そのため，文法の範囲を柔軟に考えることとした。詳細は４章で述べる。

３．３　意味記述の精査

『はごろも』の見出し語に付されている意味用法は，１．１で述べたように，基とした５つの資料の意味用法から抽出したものである。その経緯から，意味用法の記述は統一がなく，現在公開している用語数は 620 に上る。これは，執筆者の文法理論上の相違や，当該文法項目のどの側面に着目して命名するかという捉え方の相違の結果であると考えられる。現状では，利用者が作文などの産出活動を目的として，述べたい意味や機能から文法項目を探そうとした場合，意味用法に関する用語が多く，不統一なために探したい項目が見つけられないといった問題も生じうる。そこで，５名のメンバーからなる文法班では，日本語教育関係者および学習者にとって馴染みのある分類を用いた意味記述の方法について検討を行っている。主な作業は①意味大分類の検討と②意味記述の２点である。

３．１で述べたように，『はごろも』の文法項目のうち異形態や類似する表現は統合されつつあり，記述の対象としている文法項目数は，2017 年 1 月末の作業の段階で 1,744 項目である。①の意味大分類の検討では，次の作業を行っている。現在の『はごろも』の 620 の意味用法について，意味を大きく捉えるための分類枠を設けて「意味大分類」とし，1,774 項目を振り分ける。現在，『はごろも』の 639 の文法項目を 54 の意味大分類に分け，『はごろも』に記載されている意味用法と 54 の意味大分類との対応について検討を行っている。

②の意味記述では，『はごろも』の品詞分類で「活用」[1]と分類された 78 項目を除いた 1,666 項目に意味の説明を付ける作業を行っている。以下，意味大分類と意味記述における作業と課題について紹介する。

[1] ４章で述べるように，「活用」とは用言の活用形式自体を学ぶ段階で参照できるように，活用形式のみを示しており，意味記述は載せていない。

３．３．１　意味分類について

54 の意味大分類は，友松他（2010）『新装版どんなときどう使う日本語表現文型辞典』巻末の分類を参考にしている。例えば，友松他（2010）における「意志」と『はごろも』の意味用法での対応は下記の表２のようになっている。

表２　意味大分類と『はごろも』における意味用法：意志

友松他（2010）における「意志」に該当する文法項目	『はごろも』での意味用法
つもり，まい，まいとする，V ようとおもう，V ようとする，V ようにする	意志
V よう	意志・意向，勧誘・勧め，申し出
V ようとしない	否定強調

「V よう」は,『はごろも』では意志・意向の例「疲れたから, 今日は早く寝よう。(ID1663)」, 勧誘・勧めの「ちょっと休もうよ。(ID1664)」，申し出の「だれもやらないなら，ぼくがやろう。(ID1666)」の３種類ある。これらを意味大分類では「意志」としてまとめ，各項目の意味の説明でそれぞれの違いを知る設定になっている。

意味大分類において分類判断に迷う項目もあり，この点については検討が必要である。例えば，表３に挙げた文法項目は，友松他（2011）ではすべて「否定」に分類されているが，『はごろも』の意味用法では「どころではない」は「程度の強調」,「わけがない」は「判断：強い否定」,「なしに（は)」は付帯状況となっている。「断りなしに入るな」の「なしに」が「断らずに入るな」に意味が近いことを考慮すると,「ずに」は日本語教育において「付帯状況」と分類されることが多いため,「なしに（は)」も「付帯状況」に分類した方が良いだろう。このような分類の検討が今後の課題である。

表３　意味大分類と『はごろも』における意味用法：否定

「否定」の分類で一致する文法項目	一致しない文法項目
っこない，ものか（もんか）	どころではない，わけがない，なしに（は）

３．３．２　意味記述について

文法項目の意味の説明にあたっては，BCCWJ での出現傾向を調べるとともに，文法解説書の記述も参考にしている。説明の書き方は，次の３点を含めるようにしている。

① 当該文型の簡単な言いかえを示す。
② 「～の意味を持つ。／～という時に使う。」といった記述で意味の説明をする。
③ 必要に応じて話し言葉と書き言葉，プラス・マイナス評価に関する言及を加える。

表４　『はごろも』における記述

表示見出し	意味大分類	意味記述	前接形態
っぽい	傾向	「N+っぽい」は「～のような感じがする」の意味で，N の典型的な性質を持っていることを表す。「V マス+っぽい」は「すぐに～する性質である」という意味だが，前接する動詞は限られている。	N／V マス+っぽい

例えば，「っぽい」という文法項目の場合，表４のような記載となる。

３．４　文法項目の前接の形式

「活用」に含まれる項目を除いた 1,666 項目について，どのような要素とともに現れるか，前接要素の記述を進めている。BCCWJ での出現傾向，および文法解説書の記述を調査し，『はごろも』の文法項目がどのような要素と共起するかを記述している。例えば機能語の場合，次のような記載となる。

例 5）かぎりでは

前接要素の記述：V る・V た／N の＋かぎりでは

表４の「っぽい」の例では，前接形態は「N／V マス+っぽい」と書かれているが，これは名詞に接続する「っぽい」と動詞の連用形に接続する「っぽい」の２種類あることを示している。意味記述に「『V マス+っぽい』は『すぐに～する性質である』という意味だが，前接する動詞は限られている。」とあるが，これは BCCWJ の調査結果を踏まえたものである。

他にも「めく」という文法項目の場合，前接形態の記述は「N+めく」とし，意味の説明には「『～の様子が感じられる』という意味。前に来る名詞は固定的。また，『～めいた N』『～めいて』という形になることが非常に多い。」と記している。前接する名詞が固定的であることや「めく」の出現形に「～めいた N」「～めいて」が多いという言及も BCCWJ での調査によるものである。

３．５　学習者作文コーパスの用例

文法のどれが難しいのか，どれが早くから習得されるのかに関わる情報を提供することができれば，なお一層教育現場や教材作成，テスト作成に貢献できるであろう。そこで，学習者作文コーパスの用例から正用例，誤用例，非用（用いるべきところに用いていない）を載せることとした。学習者のレベルを示すことで，どのレベルでどのような正用，誤用が出現するのかが分かるようにすることが目的である。

学習者コーパスには，伊集院（2011）に基づき，正用，誤用の判断は「日本語学習者作文コーパス』に基づいた。

例 6）「〔時間〕＋に」

【正用例】たとえば、普段大学生のレポートを書くことは、二、三十年前には必ず図書館に行って、山のような本棚から需要の本をさがし出すのだ。（中級・中国語）

【誤用例】日本語を勉強していたのがもう 2 年 3 カ月に立ちました。（上級・中国語）

３．６　作業の進捗

上記 5 つの改訂作業は，2017 年度中に終了し，2017 年度末までに改訂版をアップロードする予定である。また，ダウンロード版についても，同様である。

４．文法カテゴリーについて

本章では，３．２で述べた文法カテゴリーについて，詳細に述べる。文法カテゴリーは文中の機能を示すため，最も上位のカテゴリーを「文法機能」と名付けた。その下位カテゴリーとして，「文型」「慣用表現」「活用」「敬語」「品詞」を立てる。以下では，それぞれのカテゴリーについてくわしく述べる。

４．１　「文型」

木村ほか（1989）では，「文型」を「言語単位としての文を構造の面，及び話し手の表現意図の面の両面から類型化した」ものと定義し，「N です。/N ではありません」「−は A です」「(様態)そうです」「〜ても〜」などを挙げている。これらには，文全体に関わる項目，文末表現，接続助詞に相当する項目と多様なものが含まれ，文中での機能を示すことを目的とした『はごろも』のカテゴリー分類とは必ずしも一致しない。そこで，『はごろも』では，文の類型として捉えられる「〜は〜が〜」「〔疑問詞〕＋が〜（疑問文)」などと，「複数の語が慣用的に結びつき，その結合が比較的固定化している連語」（砂川 2002）のうち，「行こうが行くまいが」の「〜ようが〜まいが」[2]のように「統語的な節の型を取り出すことのできる」(同前)ものとを文型とする。

すなわち文型とは，①2 つ以上の要素からなり，②単に 1 つの機能語としてではなく，句や文全体に関わる統語的な役割を果たし，③固定的に使用される，という 3 点を満たす。

例 7)「(のこと）を〜という」命名・定義　例文　A:国連のことを英語ではなんといいますか。B: United Nations といいます。

例 8)「かりに〜ば」　条件(仮定条件)　例文　仮にあなたの話が本当であれば、彼は嘘をついていることになる。

４．２　「慣用表現」

句レベルよりも大きく，慣用的に固定して使用されるもの。

例 9)「もういい」断り　例文　A:もう一回やってみますか。B:いや、もういいです。あきらめます。

４．３　「活用」

初級では，動詞や形容詞の活用形式自体を学ぶ段階があり，教材，テスト作成等のニーズがある。そこで，文法形式が示す表現意図とは区別して，「活用」というカテゴリーを立てた。日本語教育においては，語幹に助動詞，接続助詞を含む活用語尾がついた形を活用形として提示することが一般的であり，『はごろも』では筑波ランゲージグループ（1992）の裏扉に示されている項目を「活用」として取り上げた。

例 10）N です　N でした　N じゃありません　N じゃありませんでした　N だ　N だった

４．４　「敬語」

敬語には，「ご覧になる」のような語レベルの項目，「お〜なさる」のように複数の要素からなる項目，「ていただけませんか」のような複数の語からなる文末表現など，異なる文法カテゴリーに属するものがある。しかし，日本語教育においては，例えば「ご覧になる」と「お〜なさる」とは尊敬表現としてまとめて扱うことが一般的である。そこで，尊敬，謙譲，丁寧に関わる表現をまとめて「敬語」というカテゴリーにするほうが，教師にとっても学習者にとっても利便性が高いと考え，別のカテゴリーを作成した。

例 11)「お〜なさる」　尊敬：する　例文　来月の講演会に，館長は多くの方々をお招き

[2]砂川（2002）では，「行こうが行くまいが」から抽出される連語を「〜が〜が」としている。本稿では，『はごろも』の標記にあわせた。

なさる予定だ。

例 12)「ぞんじる＜思う＞」　謙譲：思う　例文　来月には完成すると存じます。

例 13)「させていただけないでしょうか」 謙譲:自分の行為に対する許可を求める　例文　明日休ませていただけないでしょうか。

４．５　「品詞」

品詞とは,「単語を文法的な性質によって分類したもの」(日本語記述文法研究会　2010:93)とし,「動詞，形容詞（イ形容詞・ナ形容詞)，名詞，助動詞，副詞，助詞，連体詞，接続詞，感動詞，指示詞，接辞，造語成分」を立てる。さらに下位分類は，次ページ図 1 に示すとおりである。

文中での機能は, 語のレベルで果たすこともあれば, 複合辞の場合もある。複合辞は，これまで「品詞分類」とは別に複合辞，助詞相当語などとして扱われることもあったが，文理解，産出の支援のためには文中での機能を示す同一カテゴリーに含めるほうがよいのではないかと考え,「連語」として品詞と統一的に扱うことにした。

例 14)　格助詞「〔時間〕＋に」　語レベル

格助詞「にかんして」　連語

以上に基づいて，図 1 に示すカテゴリー分類を行った。また，表 5 には各品詞に含まれる内容，例を表に示す。

５．まとめと今後の展望

本稿では，公開している『はごろも』の現状を紹介し，残された課題と利用者からの要望を踏まえ，現在進行している改訂作業について述べた。

見出し語の精査と文法機能の付与，意味記述と前接の形式の付与，学習者コーパスの用例関連づけはチームに分かれて行っているが，緊密に連絡を取っており，それぞれの関連する作業を見ながら進めている。2017 年度末には『はごろも』を改訂し，web 上で公開する予定である。

しかしながら，2017 年度末までの改訂作業の後の課題も残されている。利用者からの要望にある多言語への対応と，見出し語にまだ上がっていないが，教科書の多くで取り上げられている文法に関連する項目（「〜にくい」「〜やすい」）を取り上げるかどうかについての検討などは未定である。これらについては，今後の課題とする。

付記

本研究は，JSPS 科研費 15K02654（「日本語教師支援のための学習者コーパス文法項目データベースの構築と公開」研究代表者：堀恵子）の助成を受けたものである。

資料

伊集院郁子（2011）「日本・韓国・台湾の大学生による日本語意見文データベース」
日本語学習者作文コーパス<http://sakubun.jpn.org/>

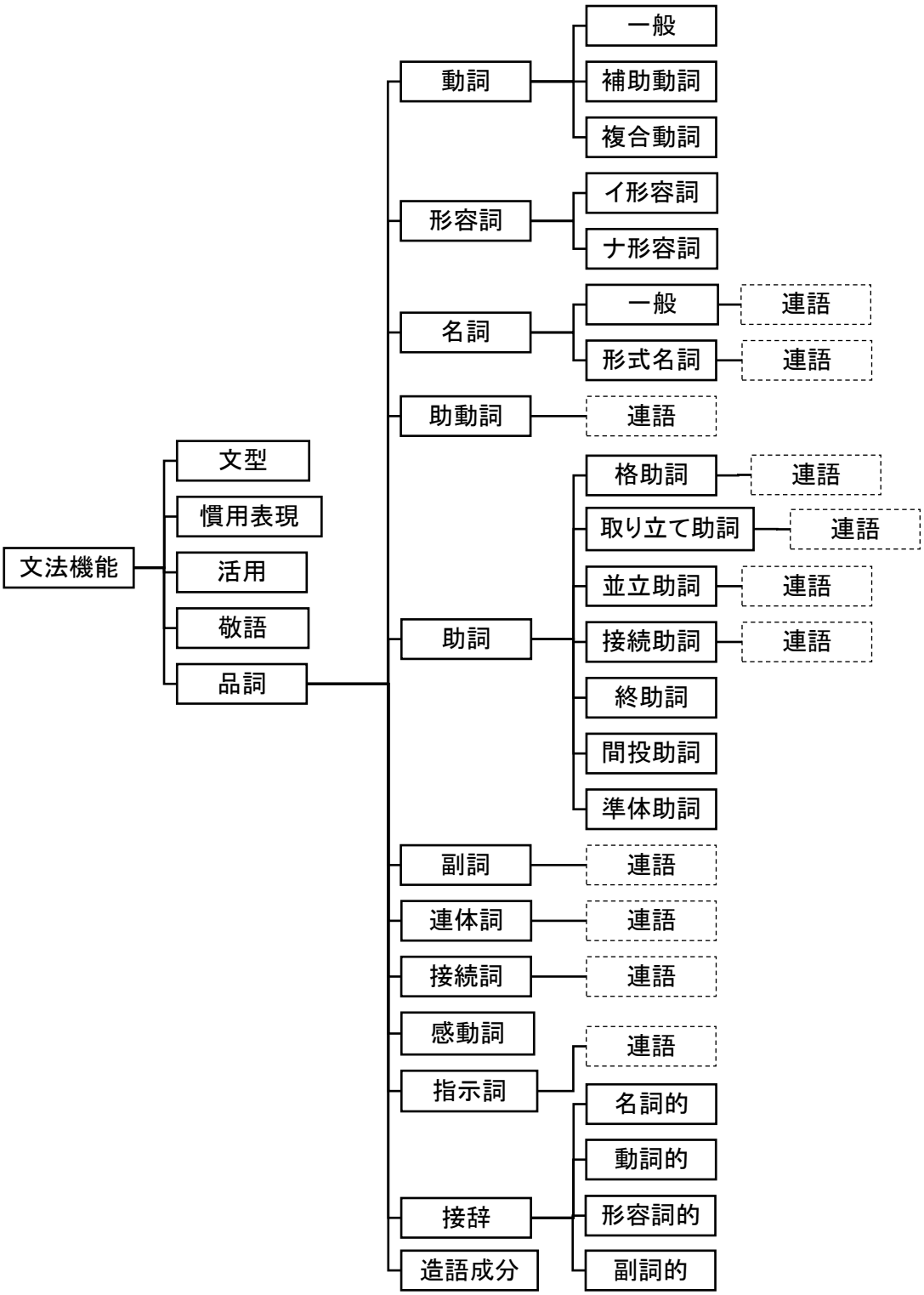


図１　『はごろも』文法機能の分類
注：連語はそれぞれの上位カテゴリーに連語的な形式を含むことを意味する。

表５　品詞の下位分類に含まれる項目，見出し語の例，注記

分類	下位分類，定義，例，*注*
動詞	一般（あげる，いたる），複合動詞（V える，V だす），補助動詞（V ていく，V ておく）
形容詞	イ形容詞（いい，イ A くて），ナ形容詞（結構，ナ A で）
名詞	一般（いたり，いつ），形式名詞（こと，の），形式名詞連語（かどうか，かなにか）　*注：疑問詞は立てず，名詞とする*
助動詞	文の述部にあって，特定の意味を加えたり，機能を果たしたりする（V よう，そうだ，わけだ），連語（V てください，かもしれない）
助詞	格助詞（〔時間〕＋に，にかんして），取り立て助詞（は，にかぎって），並立助詞（N と N，たり～たり），接続助詞（ば，とすると），終助詞(かな，もん），間投助詞（なんか，はい），準体助詞（N の，イ A の）
副詞	（もう～〔肯定〕，なにも）
連体詞	名詞を修飾する（という，わずか）
接続詞	文頭に来るもの（しかし，ことによると）
感動詞	単独で文になる。活用しない。（あ，うん）
指示詞	連体詞，副詞，名詞であるものも含む。「か」で終わる語を含まない。（これ，こう）
接辞	語基について，文中の機能をより明確に示す。名詞的（N じゅう），形容詞的（Nがたい），動詞的（Nめく），副詞的（がてら）
造語成分	語に付いて，語を形成するもの。単独で語としての用法がある語もある。（以下，以後）

注：（　）は見出し語例

参考文献

木村宗男・阪田雪子・窪田富男他編（1989）『日本語教授法』桜楓社

グループジャマシイ(1998)『日本語文型辞典』くろしお出版

国際交流基金・日本国際教育支援協会（2002）『日本語能力試験出題基準[改訂版]』凡人社

国立国語研究所（1951）『現代語の助詞・助動詞』国立国語研究所

国立国語研究所（2001）『現代語複合辞用例集』国立国語研究所

砂川有里子（2002）「国語辞書における文法的連語について−辞書と利用者に関する調査報告−」玉村文郎編『日本語学と言語学』pp.157-173 明治書院

筑波ランゲージグループ（1992）『Situational Functional Japanese』凡人社

友松悦子・和栗雅子・宮本淳（2010）『どんなときどう使う日本語表現文型辞典』アルク

日本語記述文法研究会（2010）『現代日本語文法 1』くろしお出版

堀恵子・江田すみれ・山崎誠（2016）「非母語話者日本語教師支援のために必要な品詞情報は何か」ICJLEBali<http://bali-icjle2016.com/wp-content/uploads/gravity_forms/2-ec131d5d14e56b102d22ba31c4c20b9c/2016/07/02_ICJLE2016_JP_Poster-Horiv4.pdf?TB_iframe=true>

堀恵子・李在鎬・江田すみれ（2016）「文法項目の難易度・用例文などを示す「機能語用例文データベース『はごろも』」公開」『2016 年度日本語教育学会秋季大会予稿集』pp.287-288

堀恵子・李在鎬・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ポスター発表

堀恵子・李在鎬・長谷部陽一郎（2016）「機能語用例文データベース『はごろも』について」『計量国語学』30 巻 1 号，pp.275-285，計量国語学会

森田良行・松木正恵（1989）『日本語表現文型』アルク

日本語学習者コーパスの教育応用における留意点 —『多言語母語の日本語学習者横断コーパス』に見る 母語話者 L1 産出データの安定性検証を中心に—

石川　慎一郎（神戸大学）

A Study on the Stability of L1 Production Seen in I-JAS: Japanese Learner Corpora and L2 Teaching

Shin'ichiro Ishikawa (Kobe University)

要旨

『多言語母語の日本語学習者横断コーパス』（I-JAS）を初めとする大型の日本語学習者コーパスの整備が進んだことで，母語話者と学習者の言語運用を比較し，学習者の逸脱性を客観的に明らかにした上で L2 教育の質的改善を図る可能性が拓かれつつある。しかし，こうした研究を実践する際には，母語話者データおよび学習者データの性質を十分に理解し，得られた結果を慎重に解釈する必要がある。本研究では，日本語学習者コーパスの教育応用を考える際に留意すべき問題点を概観した後，とくに母語話者による L1 産出データの安定性の問題を取り上げ，I-JAS を使った検証を行う。検証の結果，母語話者の L1 産出であっても，その正確性や言語特性については想像以上の多様性が存在することが示された。

1. はじめに

学習者コーパス研究においては，中間言語対照分析（contrastive interlanguage analysis : CIA）という分析手法が標準的に使用される（Granger, 1996 ; Granger, 1998 ; Granger et al., 2002）。中間言語対照分析では，多くの場合，母語話者による第 1 言語（L1）産出と学習者による第 2 言語（L2）産出が比較され，これにより，学習者特有の過剰使用（overuse），過小使用（underuse），誤用（misuse）などが特定される。L2 運用における学習者の逸脱の詳細が明らかになれば，それらをふまえて L2 指導の内容を改善することができる（石川，2012）。

英語においてはこうした学習者コーパスの教育応用がすでに広く試みられているが，日本語の場合，使用できるデータに制約があり，従来，同様の研究は必ずしも一般的ではなかった。しかし，近年，世界の日本語学習者と日本語母語話者による話し言葉および書き言葉の産出を大規模に収集する『多言語母語の日本語学習者横断コーパス』（International Corpus of Japanese as a Second Language：以下，I-JAS）（迫田他，2016；迫田 2016a）が開発され，今後，日本語においても，学習者コーパスを用いた中間言語対照分析と，その結果をふまえた教育内容の改善が大いに期待されるところである。

もっとも，学習者コーパスの教育応用に関しては，比較の規準となる母語話者による L1 産出の安定性，比較に使用する産出データの統制性，比較結果の教育応用の是非など，検討されるべき問題点も残されている。以下，2 節において 3 つの論点を概観した後，3 節において，とくに L1 産出の安定性の問題を取り上げ，I-JAS の日本語母語話者による L1 産出データを用いた検証を行う。なお，本研究は，2016 年 5 月 9 日にリリースされた I-JAS の第 1 次公開データ（学習者 210 名・母語話者 15 名）に基づく。

2. 日本語学習者コーパスの教育応用に関わる 3 つの問題

以下，3 つの問題について順に見ていきたい。

2.1 L1 産出の安定性

1 点目は，比較の規準となる母語話者による L1 産出が，正確性や言語特性の点で，真に安定しているかどうかということである。一般に，学習者コーパス研究では，学習者については多様性（国籍，L1，L2 習熟度，年齢，動機づけなど）を前提とした議論がなされる一方，母語話者については一枚岩的なとらえ方をすることが多い。

しかし，Leech（1998）は次のように述べて，この点に警鐘を鳴らしている。

> The conventional prescriptive view has been that the goal of foreign language learning is to approximate closer and closer to the performance of native speakers. Yet which native speakers? American, Australian, British or Caribbean? Highly educated or less so? Old or young? Such questions as these cause difficulties, although in practice teachers probably have covert answers to them. The problem becomes more noticeable when we compare learner corpora with a native-speaker 'reference corpus'.... Native-speaking students do not necessarily provide models that everyone would want to imitate. And, when we come to examine a reference corpus of native-speaker speech, the less admirable features of the native speaker's performance can show up especially clearly... （Leech，1998，p. xix）

上記でも示唆されるように，母語話者の側にも，地域・教育・年齢といった点で多様なばらつきが想定しうる。また，母語話者が常に模倣すべき「手本」となるような無謬の文を産出するとも言い切れない。

I-JAS は，母語話者について，20～50 代の年齢層にまたがる 50 名のデータを公開予定であるが（第 1 次公開データでは 15 名分のみ），これらについても，文法的正確性や基本的言語特性における安定性を実証的に検証することが重要であると思われる。

2.2 産出データの統制性

2 点目は，比較に使用する母語話者による L1 産出や学習者による L2 産出のデータが相互に比較可能な形で統制的に収集されているかどうかということである。従来の学習者コーパス研究では，多様なデータを収集することが重視され，データの内容や産出条件の統制は必ずしも十分に考慮されていなかった。たとえば，Granger et al.（2003/ 2009）によって開発された International Corpus of Learner English（ICLE）の場合，学習者による作文のトピックは 800 種を超え，辞書使用の有無，時間制限の有無といった点でも条件はまちまちである。また，比較対象となる母語話者の作文は学習者とはまったく異なる環境で収集されている。

このように，トピックや，発話条件・執筆条件が統制されていない場合，対照分析で得られた差異を合理的に解釈することは極めて困難になる。たとえば，環境問題をテーマとして，辞書なし･時間制限ありという条件で書かれた L1 作文と，夏休みの思い出をテーマとして，辞書あり・時間制限なしという条件で書かれた L2 作文があった場合，両者の対照によって得られた差を母語話者・学習者間の差とみなすのは危険である（Ishikawa, 2013；Ishikawa, 2014）。このことは，母語話者・学習者の比較だけでなく，異なる L1 を持つ学習者間の比較にもあてはまる。

I-JAS は，こうした既存の多くの学習者コーパスの制約をふまえ，母語話者・学習者の双方に共通のタスクを与えることで，相互比較を可能にする統制度の高いデータを収集している。OPI（oral proficiency interview）をアレンジした独自のプロトコルに基づき，発話については，ストーリーテリング（2 タスク），対話，ロールプレイ（2 タスク），絵描写課

題の 6 タスクが，作文については，ストーリーライティング（2 タスク）とその他の作文（4 タスク）の 6 タスクが用意されており，それぞれ，内容や条件が明示的に規定されている（迫田，2016a；迫田，2016b）。もっとも，絵描写課題とその他の作文については全員から収集しているわけではないので，分析の際には注意が必要である。

2.3 比較結果の教育応用の是非

3 点目は，比較によって得られた研究をどのような形で L2 教育に応用していくかということである。学習者コーパス研究は伝統的に L2 教育との親和性が高く，母語話者との比較によって学習者の「逸脱的」な言語使用が明らかになれば，教育的介入によって，その矯正を行うべきであるとする主張が広くなされてきた。

しかしながら，母語話者との比較で検出された差異のすべてを問題ととらえ，教育的に介入することは必ずしも適切ではない。前述のように，言語使用のすべての面において母語話者が常に無謬の手本というわけではないし，加えて，母語話者の L1 産出から乖離していたとしても，コミュニケーションを阻害するものでない限り，それらを矯正する根拠は希薄だからである。

この点に関して，Granger（2009）は図 1 のようなモデルを示し，中間言語対照分析と指導項目選定の間には，質的な取捨選択の過程が存在するべきだと強調している。

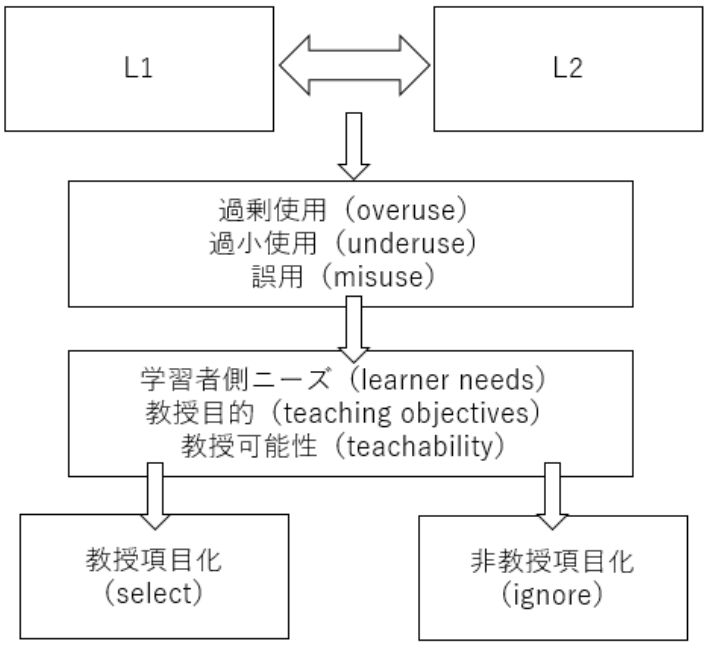


図 1　中間言語対象分析の教育応用過程（Granger，2009，Fig.4 を改変）

母語話者による L1 産出と学習者による L2 産出を対照することで，学習者の過剰使用・過小使用・誤用などの傾向が検出されるが，それらを教授項目として選定するか，問題ないとして無視するかは，学習者側ニーズ，教授目的，教授可能性といった諸点をふまえた現場の教師の判断にゆだねられるべきであるというのが Granger の主張である。つまり，学習者コーパス研究者に一義的に期待される仕事は，何を教え，何を教えないかを直接に指し示すことではなく，現場の教師が教授の内容を吟味する上で必要となる基礎資料を適切な形で提供することであると言えよう。

I-JAS について言えば，現時点では一部のデータが試行的に公開されているだけで，本格的な母語話者との対照研究や日本語教育への応用はいまだ行われていない。しかし，今後，

データの全体が公開されると，対照研究の結果を日本語教育に応用しようとする研究も増えてくるだろう。その際，安易な教育応用と一線を画する態度が研究者には求められる。

3. I-JAS に見る母語話者 L1 産出データの安定性

3.1 ねらいと RQ

前節では，日本語学習者コーパスの教育応用に関わる 3 つの問題点を概観したが，このうち，とくに重要になるのは，比較の規準となる母語話者による L1 産出の安定性の問題である。本研究は，I-JAS の母語話者 L1 産出データを用い，文法的正確性と，産出の内容に影響されることが少ないと考えられる基本的な言語特性（語数，句読点使用率，高頻度語使用状況）に着目して安定性の検証を行う。リサーチクエスチョン（RQ）は以下の 3 つである。

RQ1 母語話者による L1 産出は，文法的正確性の点でどの程度安定しているか?
RQ2 母語話者による L1 産出は，基本的な言語特性の点でどの程度安定しているか?
RQ3 母語話者による L1 産出は，高頻度語の使用状況に着目した場合，どの程度内部的に一体か?

3.2 データと手法

本研究では，I-JAS の第 1 次公開版に含まれる 15 名の母語話者によるストーリーライティング（SW1）のデータを使用する。各種のタスクの中で，とくにストーリーライティングを選んだのは，話し言葉に比べ，書き言葉のほうが安定的な産出がなされやすいことに加え，I-JAS のストーリーライティングは，すでに行ったストーリーテリングと同一課題で行われているためである。各種のタスクの中で最も高い安定性が予想されるストーリーライティングにおいて，仮に何らかの不安定性が検出されるとすれば，それは，母語話者による L1 産出の安定性の想定に対する強力な反証となろう。

SW1 では，ピクニックに関する 5 枚のイラスト（図 2）を見た後，指示された冒頭文（朝，ケンとマリはサンドイッチを作りました）に続けて，イラストに沿ってストーリーを作文する。なお，このイラストは先に行ったストーリーテリングで使用したものと同等であり，被験者はあらかじめ話の内容を十分に理解したうえで作文に臨むことができる。

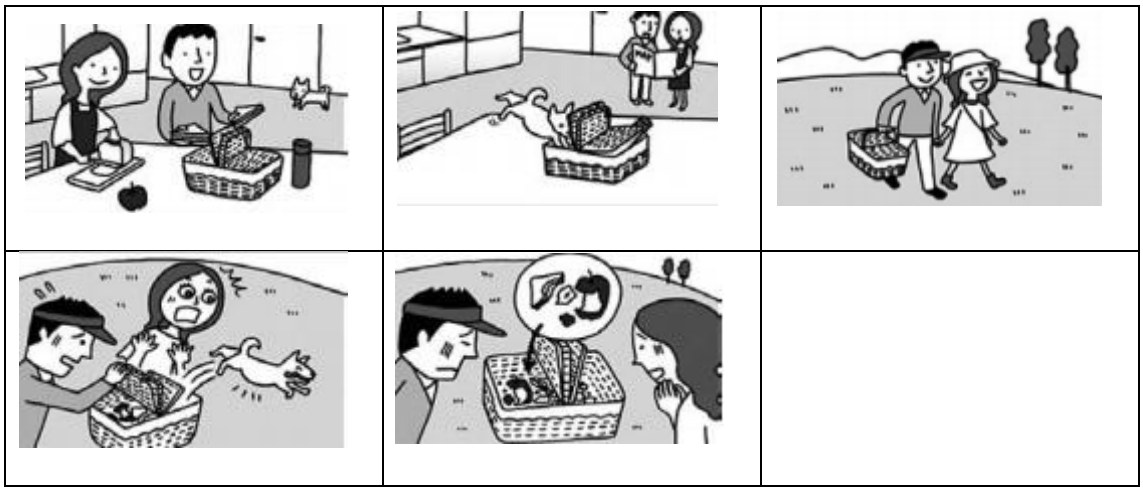


図 2 J-JAS ストーリーライティング第 1 課題（奥野・リスダ，2015 の図 1 を再構成）

SW1 のテキストは，行番号を削除した後，Chasen によって形態素解析し，以後の分析の基礎データとする。RQ1（文法的正確性）については，I-JAS に含まれている誤用修正情報のほか，全例の目視検証により，文法的に逸脱が疑われる言語使用の有無を確認する。

RQ2（言語特性）については，作文ごとに，語数 1（句読点を含まない），語数 2（句読点を含む），読点数，句点数，句読点数，読点／句点率，100 語あたりの読点数，100 語あたりの句点数，100 語あたりの句読点数を計量し，最小値と最大値の比率を確認する。また，それぞれの作文において粗頻度 5 以上となる上位形態素（句読点・記号は除く）を取り出し，15 作文中での重なりの度合いを確認する。RQ3（内部的一体性）では，15 作文中，6 種以上で共通して使用されている全 41 語（句読点・記号は除く）を資料として，ケースクラスター分析（距離は(2-2r)の平方根で定義し，合併後の距離計算は Ward 法を使用）とコレスポンデンス分析を行い，15 種の母語話者作文が内部的にどの程度一体的であるか，仮に内部でいくつかに分割される場合は，どのような語の使用がそれに影響しているのかを確認する。

3.3 結果と考察

3.3.1 RQ1 文法的正確性

中間言語対照分析において，母語話者の L1 産出からの逸脱を学習者の問題ととらえるのは，L1 産出が文法的に無謬であるという前提に基づく。しかし，今回のデータを検証したところ，I-JAS の開発者の側で誤用修正がなされた例が 2 点（JJJ03，JJJ26），他にも誤用と考えられる例が 1 点（JJJ35）見つかった。

(1) 目的地に到着し、お昼ご飯を食べるためにバスケットを開けると、犬が飛び出しして（→飛び出して）きました。（JJJ03）
(2) そんな仲睦ましい（→仲睦まじい）二人を部屋の隅から、飼い犬が見ています。（JJJ026）
(3) 行く場所を地図で確認してる［→している?］隙に愛犬がバスケットの中に入ってしまいます。（JJJ35）

誤用はいずれも軽微なものだが，母語話者作文 15 例中の 3 例において，日本語として問題となりうる言語使用が見つかったことになる。このことは，母語話者データが必ずしも言語の正確性のサンプルとなりえない場合があることを示唆している。

併せて注目すべきは，上記のようなはっきりした誤用以外にも，表記や語彙使用の点で問題を含む例が散見されたことである。下記は JJJ01 の作文の一部である。

(4) …出かける前に二人が地図を見ている間に、サンドイッチを入れたバスケットに犬が入ってしまいました…やがて突然犬がバスケットから飛び出し、二人は驚きました。

引用冒頭部の「出かける前に二人が地図を見ている間に」は，時を表す 2 種の副詞句が不適切に並列されたもので，たとえば「出かける前，二人が地図を見ている間に」や「出かけようとして二人が地図を見ている間に」などとするのが日本語としてより適切であろう。また，後半の「やがて突然犬がバスケットから飛び出し」の部分も，「やがて」と「突然」という意味の異なる副詞が同時に「飛び出し」という動詞を修飾する形になっており，「やがて，犬が突然バスケットから飛び出し」などとするのがより自然と言えよう。

上記は書き手の作文技術の不足やケアレスミスに起因するものと言えるかもしれないが，興味深いことに，それらに起因しない不自然な日本語表現も見つかる。

(5) …行く場所も決定し歩きだす二人。目的地に着いて「さあ食べよう」とバスケットを開けるとそこからは犬が・・・驚く二人。（JJJ09）（※全角省略符合は原文）
(6)…「じゃあ、ここでお昼にしましょう。」とマリが言ったので、ケンはバスケットを地面に下ろすと、まぁびっくり。…「おーい、サンドイッチも林檎も全部食べられちゃたよ。」とケン。「お昼どうしましょう。」、と困った顔のマリ。（JJJ10）

(7) …余程楽しかったのでしょう。飼い犬がその隙にサンドイッチの入ったバスケットに忍び込んだのも気づかないのですから。…さぁ歩き疲れた二人はサンドイッチを食べようと丘の上でバスケットを開けました。するとあろう事か、飼い犬が飛び出したではありませんか。…（JJJ26）

体現止め・倒置・感嘆文などを含むこうした表現は，誤りではないものの，日本語教育でモデルとするような標準的な日本語とは言い難い。こうした表現の混在は，一部の母語話者がイラストからストーリーを作るという課題を狭義の「ストーリー」，つまりは，昔話や物語文，あるいはト書きのような作文を行う課題と理解し，あえてそうしたジャンルに典型的に見られる修辞法を使って作文したことを示している。筆者の検証では，15 人中 8 人が課題を文字通りに受け取って出来事を中立的に報告しているのに対し，残りの 7 人（上記の 3 例に加えて JJJ30，35，37，50）は少なくとも部分的に物語文的な特殊な修辞法を使用している。I-JAS のストーリーライティング課題は十分に練られたものであるが，にもかかわらず，母語話者の間ですら，その受け止めに差があり，予期せぬ言語的多様性が生じていることに注意が必要である。仮に，母語話者データの中身に十分な注意を払わず，機械的に対照分析を行った場合，こうした表現を日本語の「規準」としてしまう危険性もあるだろう。

3.3.2 RQ2 基本的言語特性

まず，語数および句読点の使用状況を検証したところ，表 1 の結果が得られた。表中の語数 1 および語数 2 は句読点を除いた語数と含めた語数を，読・句・句読・読／句は読点数・句点数・句読点数・読点／句読点比率を，読%・句%・句読%は 100 語あたりの読点・句点・句読点数を示す。なお，変動係数は標準偏差を平均で割った値で，平均の異なりを補正した上でばらつきの大きさを比較するための指標である。

表 1 文長・句読点使用率

	語数 1	語数 2	読	句	句読	読／句	読%	句%	句読%
JJJ01	89	99	5	5	10	1.0	5.1	5.1	10.1
JJJ03	105	117	7	5	12	1.4	6.0	4.3	10.3
JJJ09	107	115	1	7	8	0.1	0.9	6.1	7.0
JJJ10	151	171	6	14	20	0.4	3.5	8.2	11.7
JJJ11	81	90	4	5	9	0.8	4.4	5.6	10.0
JJJ12	87	97	5	5	10	1.0	5.2	5.2	10.3
JJJ14	92	99	2	5	7	0.4	2.0	5.1	7.1
JJJ15	104	115	5	6	11	0.8	4.3	5.2	9.6
JJJ17	84	93	5	4	9	1.3	5.4	4.3	9.7
JJJ26	180	199	6	13	19	0.5	3.0	6.5	9.5
JJJ30	115	129	7	7	14	1.0	5.4	5.4	10.9
JJJ35	82	92	6	4	10	1.5	6.5	4.3	10.9
JJJ37	81	90	2	7	9	0.3	2.0	7.1	9.2
JJJ50	153	160	6	1	7	6.0	3.8	0.6	4.4
JJJ57	110	123	4	9	13	0.4	3.3	7.3	10.6
平均	108.1	119.3	4.7	6.5	11.2	1.1	4.1	5.4	9.4
標準偏差	30.4	33.1	1.8	3.4	3.9	1.4	1.6	1.8	1.9
変動係数	0.3	0.3	0.4	0.5	0.3	1.2	0.4	0.3	0.2
最大値	180.0	199.0	7.0	14.0	20.0	6.0	6.5	8.2	11.7
最小値	81.0	90.0	1.0	1.0	7.0	0.1	0.9	0.6	4.4
最大／最小比率	2.2	2.2	7.0	14.0	2.9	42.0	7.5	13.1	2.7

語数 1，語数 2 に注目すると，最小値と最大値の差が 2.2 倍に及ぶことが明らかになった。I-JAS では，イラストを与えているため，産出の量はほぼ一定になると予想されたが，実際には 2 倍を超える差が存在する。学習者の L2 産出の分析では，一般に，文章の長さは習熟度の高さを示すとされるが，習熟度スケールの終値にあるはずの母語話者間にすらこうした差が生じていることは注目に値する。

また，100 語あたりの句点，読点，句読点の数に注目すると，最小値と最大値の差は 2.7 倍～13.1 倍に及ぶことがわかった。読点／句点率ではその差は 40 倍以上に達する。これらは，内容の差の影響を受けにくいはずの句読点の使用についても，母語話者間に大きなばらつきが存在することを示唆している。なお，読点／句点率における著しい個体差は変動係数からも裏付けられる。

次に，高頻度語の使用状況に注目したい。上位形態素の一覧は表 2 の通りである。表中，サンは「サンドイッチ」，バスは「バスケット」を略記したものである。

表 2 高頻度形態素

JJJ01		JJJ03		JJJ09		JJJ10		JJJ11		JJJ12		JJJ14		JJJ15	
に	7	に	8	た	6	た	8	た	5	た	8	た	6	た	7
た	6	が	6	に	5	て	6	て	5	が	6	て	6	て	5
て	6	た	6			と	6	と	5	と	6	を	6	と	5
は	5	と	6			は	6	に	5	に	6	は	5	に	5
まし	5	を	6			が	5	まし	5	ﾊﾞｽ	5	まし	5	まし	5
		て	5			の	5			まし	5				
						を	5								
JJJ17		JJJ26		JJJ30		JJJ35		JJJ37		JJJ50		JJJ57			
て	6	た	10	に	9	を	6	て	6	て	12	た	7		
を	6	の	10	た	7	に	5	た	5	た	9	を	7		
た	5	を	9	と	7			と	5	に	7	は	6		
		は	7	を	7			は	5	を	7	まし	6		
		に	6	て	6			まし	5	が	6	て	5		
		まし	6	まし	6					まし	6	と	5		
		ｻﾝ	5	ｹﾝ	5					ｻﾝ	5	ﾊﾞｽ	5		
		と	5	は	5										
				ﾊﾞｽ	5										
				ﾏﾘ	5										

コーパス言語学では，高頻度語の使用状況は高度に安定的であると言われるが，今回のデータについて言えば，頻度 1 位の形態素が 1 語に決まる 10 名（JJJ01，03，09，10，12，15，30，35，37，50）に限ってみても，「た」が 4 名，「に」が 3 名，「て」が 2 名，「を」が 1 名とばらついており，完全な一致は確認できなかった。また，基本助詞間の頻度の多少に関して，一例として，「を」と「に」のいずれの頻度が多いかを見たところ，「を」を多用する者が 6 名（JJJ10，14，17，26，35，57），「に」を多用する者が 7 名（JJJ01，03，09，11，12，15，30），同数が 2 名（JJJ37，50）となり，やはり，完全な一致とはならなかった。

以上の結果は，語数・句読点・高頻度語といったごく基本的な言語特性に限っても，母語話者による L1 産出が想像以上に大きな多様性を持つことを例証する。こうしたばらつきの可能性を考慮せずに，母語話者と学習者を単純に比較して対照研究を行った場合，研究者の想定を超えて，極めて不適切な結果を導いてしまう危険性も否定できない。

3.3.3 RQ3 内部的一体性

母語話者のL1産出が安定的で，がある種の「一枚岩」のようなものであるとすれば，その内部にはっきりした下部構造は存在しないはずである。そこで，15人中6人以上が共通して使用している41語を資料としてクラスター分析を行ったところ，図3の結果を得た。

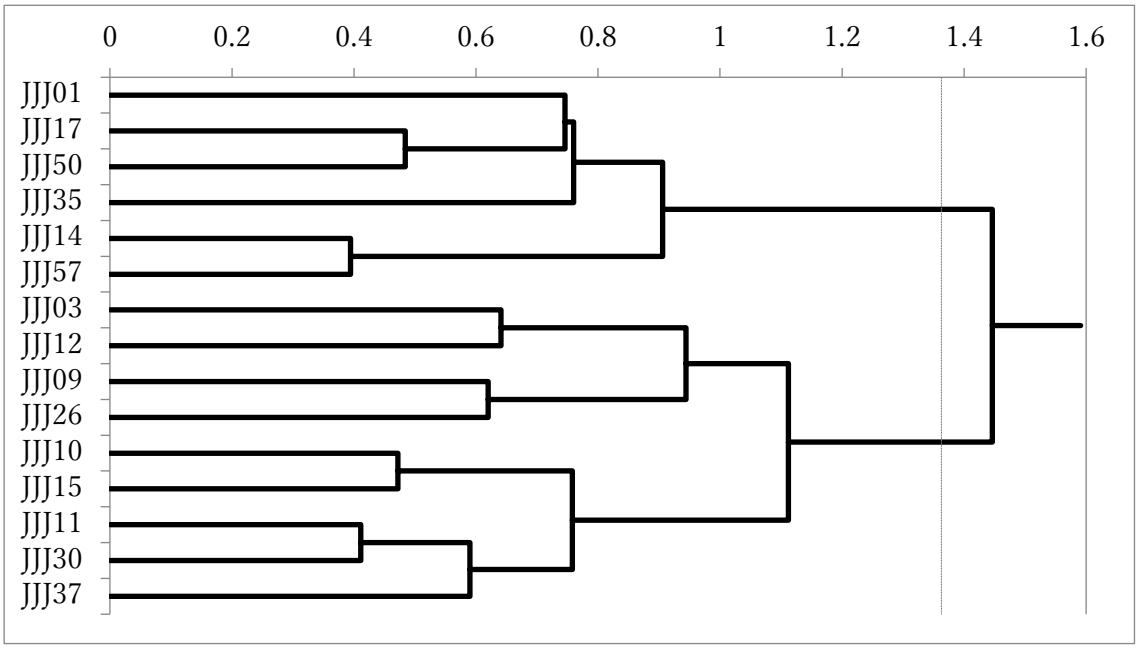


図3 クラスター分析に基づく樹形図

定常状態が最も長く継続する場所（距離1.2～1.4）にカッティングポイントを置くと，15人の母語話者は大きく2グループ（クラスターA：JJJ01, 14, 17, 35, 50, 57；クラスターB：JJJ03, 09, 10, 11, 12, 15, 26, 30, 37）に分割されることとなった。このことは，高頻度語の使用状況においても母語話者のL1産出が均質でなく，はっきりした下部構造が存在することを示している。

次に，コレスポンデンス分析の結果を概観する（図4）。散布図に見られるように，15名の母語話者は，まず，第1次元（横軸）上で，左右の2グループに分割される。この時，左側には先ほどの分析で得られたクラスターBが，右側にはクラスターAが位置する。それぞれの側に布置された形態素に注目すると，左側には，動作主体を表す名詞（「マリ」「ケン」「犬」），行為動詞（「飛び出す」「開ける」「入っ」「出かけ」「食べ」），否定辞（「ない」）などがあり，右側には，意味の希薄化した動詞（「い」「し」），丁寧性を表す文末表現（「ます」），照応指示語（「その」），意思助動詞（「う」），助詞（「の」「から」「で」「は」）などがある。母語話者のL1産出は，内容語中心型産出と機能語中心型産出に二分されると言えよう。

また，第2次元（縦軸）を合わせて分析すると，全体は第1象限（Z1＋／Z2＋），第2象限（Z1－／Z2＋），第3象限（Z1－／Z2－），第4象限（Z1＋／Z2－）に4区分され，第1象限は「その」や「し・ます」などの機能的語群，第2象限は「マリ」や「ケン」などの固有名詞，第3象限は「出かけ」「入っ」「食べ」などの動詞群，第4象限は「二・人」という人称代名詞的語群によってそれぞれ特徴づけられることがわかった。コレスポンデンス分析の結果もまた，母語話者のL1産出が必ずしも一枚岩的な均質性を持っていないことを例証している。

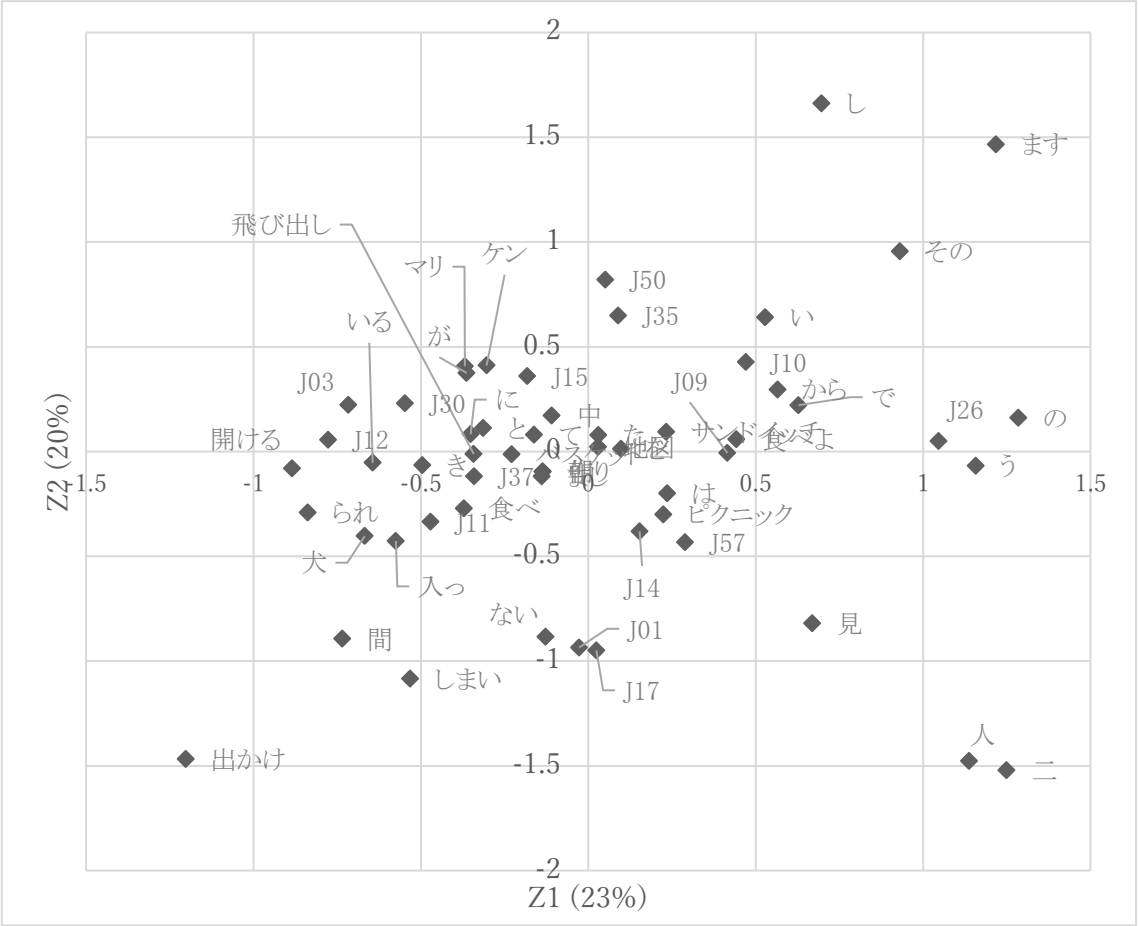


図 4 コレスポンデンス分析に基づく散布図（国籍コードを J で略記）

4. まとめ

本研究では，日本語学習者コーパスの教育応用における問題点として，比較の規準となる L1 産出の安定性，比較に使用する産出データの統制性，比較結果の教育応用の是非の 3 点を概観し，その後，I-JAS を用い，母語話者による L1 産出の安定性について実証的観点から検証を行った。その結果，3 つのリサーチクエスチョンに即して以下の事実が明らかになった。

まず，RQ1（文法的正確性）については，母語話者であっても明示的な文法的エラーがおよそ 3 分の 1 の割合で見つかり，また，明らかな誤用とまでは言えないにせよ，日本語として不自然表現も随所に散見されることが確認された。さらには，統制的なタスクを与えたにもかかわらず，書き手によってタスクの受け取り方に差があり，そのことが非標準的な日本語表現の使用につながっていることがわかった。

次に，RQ2（基本的言語特性）については，最小値と最大値の比率が，語数では 2.2 倍，句点・読点・句読点数では 2.7 倍～13.1 倍に及び，読点／句点率では 40 倍以上に達することが明らかになった。また，高頻度形態素に限っても，頻度順位上の一致度は予想以上に低いことが示された。

最後に，RQ3（内部的一体性）については，クラスター分析およびコレスポンデンス分析の結果，15 人の母語話者の L1 産出が内容語中心型と機能語中心型に大きく二分され，さらに，より細かく見れば，機能的語群，固有名詞，動詞，人称代名詞的語群に特徴づけられる 4 つのグループに区分されることが分かった。

以上の結果は，中間言語対照分析において，母語話者の L1 産出を絶対的で安定的な規準

とみなし，母語話者と学習者を比較して学習者の側の「逸脱」を論じることが，場合によってきわめて危うい行為になりかねないことを示すものである。学習者コーパス研究では，計量的なテキスト比較を行うことで異なるテキスト間の差異を容易に検出することができるわけだが，検出された差異が果たして何を意味しているのか，また，検出された差異を教育現場で扱うべきかどうかについては，慎重な判断が求められる。

謝　辞

本稿は，2016 年 12 月 3 日に開催された第 1 回学習者コーパス・ワークショップ―学習者コーパス（I-JAS）を利用するために―」（於：国立国語研究所）における招待講演「世界の英語学習者コーパス研究の潮流 ：How から Why へ」で口頭報告した内容の一部を大幅に加筆修正し，新規に原稿化したものである。同ワークショップは，石黒圭氏（国立国語研究所）がリーダーを務める「日本語学習者のコミュニケーションの多角的解明」プロジェクトおよび迫田久美子氏（国立国語研究所）による科研費（基盤研究 A）「海外連携による日本語学習者コーパスの構築および言語習得と教育への応用研究」プロジェクトに基づく。本稿執筆の契機となる講演機会をお与え下さったことに対し，迫田教授・石黒教授に御礼申し上げる。また，当日のワークショップのパネリストであり，本稿初校に対して貴重な意見をいただいた小西円氏（国立国語研究所）および奥野由紀子氏（首都大学東京）にあわせて感謝申し上げる。

文　献

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Language in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund, Sweden: Lund University Press.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). Harlow, England: Addison Wesley Longman.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). Amsterdam, The Netherlands: John Benjamins.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2003). *International corpus of learner English*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *International corpus of learner English. Version 2*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition and foreign language learning*. Amsterdam, The Netherlands: John Benjamins.

石川慎一郎 (2012).『ベーシックコーパス言語学』東京：ひつじ書房.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian Learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91-118). Kobe, Japan: Kobe University.

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2* (pp. 63-76). Kobe, Japan: Kobe University.

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). Harlow, England: Addison Wesley Longman.

奥野由紀子・リスダ=ディアンニ (2015).「『話す』課題と『書く』課題に見られる中間言語変異性：ストーリー描写課題における『食べられてしまっていた』部を対象に」『国立国語研究所論集』*9*, 121-134.

迫田久美子 (2016a).「I-JAS 使用の手引き簡易版」東京：国立国語研究所.

迫田久美子 (2016b).「学習者コーパスをどう使うか：I-JAS，C-JAS 検索法入門」『コーパスと日本語教育研究国際シンポジウム予稿集』（湖南，中国：湖南大学）38-45.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子 (2016).「多言語母語の日本語学習者横断コーパス International Corpus of Japanese as a Second Language」『国語研プロジェクトレビュー』*6*(3), 93-110.

漢語の仮名表記
—実態と背景—

間淵 洋子（明治大学国際日本学研究科・日本学術振興会）†

A Corpus Based Study of Sino-Japanese Words written in *Kana*: Actual Condition and Background

MABUCHI Yoko (Meiji University / Society for the Promotion of Science)

要旨

本発表は，本来漢字で表記されるはずの漢語が平仮名や片仮名で表記される事象を取り上げ，『現代日本語書き言葉均衡コーパス』（以下，「BCCWJ」と表記）を用いて，その実態と背景を明らかにすることを目的とする。

BCCWJ の網羅的な漢語の表記実態調査に基づき，個々の語の仮名表記率から，仮名表記が，主たる表記である語，ある程度一般的である語を特定した上で，仮名表記の定着度合いに，**字体特徴**（常用漢字表外字・音を含む語は仮名表記率が高いが，表内字でも仮名表記率の高い語がある），**語の出現状況**（語彙レベルが高い語ほど仮名表記率は低い），**音声変位形の有無**（「格好」に対する「カッコ」のような音転訛形を持つ語は仮名表記率が高い），**意味分野**（動植物や食物の分野では仮名表記率が高い），**品詞**（副詞用法を持つ語は仮名表記率が高い），**レジスター**（Web 媒体は仮名表記率が低い）等との関連性が見られることを示す。また，字体特徴にかかわらず，意味分野や品詞において特定の語彙群に同様の傾向が見られるのは，表記選択に類似性に基づく合理化作用が働くことによると主張する。

１．はじめに

日本語は，漢字，平仮名，片仮名，アルファベットといった多様な文字を持ち，最も複雑な表記体系を持つ言語だと言われる。しかし，その複雑さの中には，和語の主要な意味を担う部分や漢語は漢字で，送り仮名や文法的な役割を表す付属語は平仮名で，外来語を片仮名で，外国語をアルファベットで，といった原則的な役割分担が定まっており，これら多様な文字種は日本語文章の効率的な理解に欠かせないものとなっている。また，それぞれの文字種が持つ感情的な意味を加えたり，あえて原則を外すことで新規性や特殊性を持たせたりといった，表記戦略を取ることで，日本語表現を極めて豊かなものとしている。

このような字種の多様性と語の表記の関連性について言及したこれまでの研究の多くは，文字種と語種の対応関係に目を向け，主に外来語以外で片仮名表記される事象を取り上げ，その要因や効果，機能を述べたものであった（中山 1998，成田・榊原 2004，臼木 2008，柏野 2014 など）。これらの先行研究により，外来語以外で片仮名表記される語には，①常用漢字表外漢字，表外音訓，表外熟字訓を含む語，②動植物名を表す語，③オノマトペ，が多いこと，また，④特殊な語義の書き分け（意味の限定や専門用語的な用法など），⑤強調，⑥片仮名の持つ感情的意味の付加，といった表記戦略により，片仮名表記が選択されることが明らかにされている。一方で，特段の効果や機能に寄らない表記の慣習によるものとして，その先に踏み込むことなく残される語があり，これらの表記選択のメカニズムが明らかになっているとは言い難い。また，実際にどのような語に仮名表記の選択される

† mabuchi＠meiji.ac.jp

慣習があるのかという個々の語の実態把握・解明についても，充分に尽くされてはいない。

そこで，本発表では，特に 2 字漢語を例として，コーパスを用いて網羅的に漢語における仮名表記の実態を調査し，その背景について多角的な考察を試みる。

２．研究方法

２.１　コーパス

本研究では，漢語の仮名表記の実態をできる限り網羅的に捉えるために，国立国語研究所が 2011 年 8 月に公開した『現代日本語書き言葉均衡コーパス』（BCCWJ）を用いた調査を行う。BCCWJ は，日本語研究ばかりでなく，日本語教育，国語教育，辞書編纂，心理学・認知科学，言語政策といった多様な研究分野への応用を目指し，綿密な設計により構築された，日本初となる，また唯一の，大規模バランストコーパスである（前川 2008）。代表性を有する多種多様な書き言葉が，大量に，かつ，全てに形態論情報が付された状態で収録されており，本研究のように探索的にできるだけ多様で大量の語例を収集するためには，まさに好適なデータであると言えよう。

２.２　調査対象語の抽出

漢語の仮名表記実態把握を試みるにあたって，全ての漢語を漏れなく抽出し分析を行うことが望ましいのは明らかであるが，本研究では，敢えて 2 字漢語に限って分析対象として扱うこととした。その理由は，概ね以下の 4 点である。

1) 漢語は語構成によって，日本古来の語彙（和語）との熟合度が異なっており，特に，1 字漢語の和語との熟合度は，2 字以上の漢語のそれとは大きな隔たりがある。熟合度の違いは文法的な振る舞いや，表記にも大きく影響を及ぼしており，1 字漢語と 2 字以上の漢語とは，それぞれ別途分析の必要がある。
2) 上記熟合度の高さなども関連して 1 字漢語は，「する」などの和語と，時には連濁を伴い分かちがたく結合することがある。1 字漢語と 2 字漢語は，コーパスでの単位認定においても大きな差異があり，「愛する」「興ずる」「処する」といった 1 語の混種語として扱われる。そのため，他の漢語と同様の方法で網羅的に収集するのが難しい。
3) 一方 2 字漢語は，1 字漢語が和語と固く結束するのと同様に，漢字 2 字が離れがたく熟合し意味を持ち，これが語基となって，他の 1 字漢語や 2 字漢語と結びついて新たな語を形成する。漢語において最も基本となる形態であり，語数も多く使用も多い。よって，2 字漢語の実態を把握しておくことは，漢語全体の実態把握の基礎となる。
4) 2 字漢語は漢字 2 字が強く結合しており，それ故に語の理解・把握において漢字が担う役割が大きい。漢字そのものが重要な役割を果たしている場合には，本来漢字表記を捨てて仮名表記を選択する動機に乏しいと思われるため，それでもなお仮名表記が選択される背景を調査することは，日本語の表記・語彙体系における漢語の位置づけを探る手掛かりになる。

上記の理由から，本研究では 2 字漢語を調査対象とするが，その抽出には，Web 上のコーパス検索アプリケーション「中納言」を用いた[1]。「中納言」の検索においては，以下の条件を設けて，調査対象語として，語彙素が漢字 2 文字からなり，書字形に仮名を含む漢語を網羅的に取り出した。

[1] 現代日本語書き言葉均衡コーパス（非 NumTrans 版） 中納言 2.2.0 を用いた。

キー条件：語種「漢」（漢語），語彙素「[一-龠=][一-龠々=]」，書字形「%[ぁ-んァ-ン]%」
検索対象：「教科書（OT）」[2]を除く全てのレジスター

その結果，BCCWJ において異なり語数で約 3,500 語の漢語が抽出された。その中から，数詞や，仮名表記の粗頻度合計が 3 以下の約 1,750 語（例：「大学/だいがく」「内容/ないよう」「安全/あんぜん」「選手/せんしゅ」「会議/かいぎ」など）を分析対象外とした。これは，仮名表記の粗頻度が極めて低いものの場合，出現例が特異な例であり，仮名表記選択の背景分析に寄与しない可能性が高いためである。更に，残った半数について各語の総出現頻度（延べ語数）を計測し，それが 100 に満たない約 350 語（例：「変梃/へんてこ」「滅法/めっぽう」「鮟鱇/アンコウ」「鵞鳥/ガチョウ」など）も，分析対象外とした。これは，出現頻度が低いものの場合，後節で検討する様々な分類カテゴリやレジスターによる比較を行う際に，計量分析に耐える充分なサンプル数を確保できない可能性があるためである。

また，2 音節の 2 字漢語が平仮名 2 字で表記されている場合は，解析の誤りであることが多いため，一部表記例を確認し，誤解析が多いものについては，やはり分析対象から外した（「基地/きち」「意気/いき」「未知/ミチ」「磁気/じき」「恣意/しい」など）。以下のような例である（以下，例文においては，注目する語・表記を太字とし，下線を施す）。

(1)　携帯電話メーカーで**あゆ**がイメージキャラクターを勤める。
（解析結果：アユ,阿諛,名詞-普通名詞-サ変可能）[3]
【出典】 BCCWJ サンプル ID：PB27_00045 実著者不明・H.A シスターズ(著)『For A』2002 年

これにより，最終的な分析対象は，異なり 1，075 語，2,835 表記，延べ語数 4,022,616 語となった。

３．調査結果

３.１　仮名表記率

抽出した 1,075 語については，出現総頻度に対する仮名表記頻度（平仮名表記と片仮名表記を合わせたもの）の比率である「仮名表記率」を求め，比率により以下に層別した。

特：75%以上　高：50%以 75%未満　中：25%以上 50%未満　低：25%未満

それぞれの層に含まれる語数と，語例として出現総頻度数の高いものから上位 10 語を「仮名表記頻度 / 出現総頻度 / 仮名表記率」と共に示すと以下の通りである（表１）。

[2] 本研究では漢語の仮名表記を扱うが，学習段階の児童・生徒用の教科書においては，未習漢字を仮名で表記する，一般社会における表記とは異なる特殊な事情に基づく仮名表記が含まれているため，調査対象から外すこととした。ただし，調査対象とした書籍や雑誌においても，一部に幼児・児童・生徒向けの作品・記事が含まれているため，左記の特殊な事情に基づく仮名表記は含まれる可能性がある。

[3] 全 206 例中，正しい解析結果は，書字形「阿諛」の 23 例のみ，仮名表記例「あゆ」183 例は全て人名で誤解析であった。

表 1　仮名表記率カテゴリ別語例

カテゴリ	所属語数	所属語例
特 (75%以上)	59	勿論(19,459/20,902/93.1%), 沢山(12,918/15,543/83.1%), 所為(8,272/8,390/98.6%), 御免(3,991/4,339/92.0%), 折角(3,002/3,328/90.2%), 段々(2,861/3,301/86.7%), 林檎(2,102/2,384/88.2%), 無論(1,646/2,187/75.3%), 胡麻(1,767/2,119/83.4%), 大分(1,640/1,913/85.7%)
高 (50%以上)	53	奇麗(8,118/11,677/69.5%), 多分(4,421/7,975/55.4%), 馬鹿(4,761/7,914/60.2%), 是非(5,438/7,909/68.8%), 随分(3,256/4,657/69.9%), 大抵(2,362/3,242/72.9%), 蛋白(2,127/3,106/68.5%), 途端(2,041/3,064/66.6%), 怪我(1,451/2,855/50.8%), 人参(1,445/1,999/72.3%)
中 (25%以上)	81	本当(10,155/35,069/29.0%), 格好(3,151/6,425/49.0%), 大体(2,862/5,985/47.8%), 御飯(1,687/5,560/30.3%), 挨拶(1,407/5,308/26.5%), 普段(1,366/5,208/26.2%), 御覧(1,333/4,283/31.1%), 味噌(1,350/3,843/35.1%), 漫画(1,515/3,523/43.0%), 醤油(1,323/3,260/40.6%)
低 (25%未満)	894	自分(549/110,611/0.5%), 問題(36/70,995/0.1), 必要(50/70,110/0.1%), 時間(79/67,391/0.1), 関係(59/57,073/0.1%), 人間(146/42,308/0.3) %, 生活(9/41,372/0.02%), 会社(13/41,115/0.03%), 研究(4/39,936/0.1%), 意味(214/38,966/0.5%)

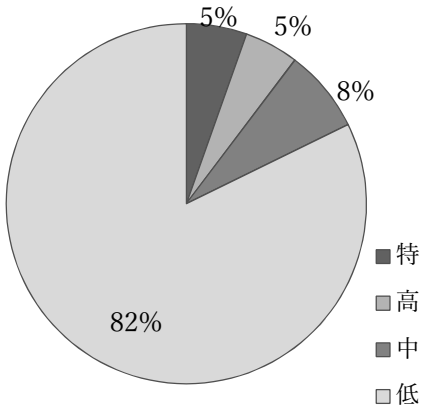


図 1 仮名表記率カテゴリの分布

表 1, 図 1 より, 仮名表記例を持つ語の約 10%は, 仮名表記の比率が 5 割以上あり，語の優勢な表記が仮名表記であることが分かった。これらは，仮名表記されることが慣用となっている語とみなすことができる。

また, 8%ほどの語で, 漢字表記が優勢であるものの, ある程度，仮名表記が一般的に用いられていることが分かった。これらは，仮名表記を選択している用例と選択していない用例を分析することで，仮名表記選択の背景についてより細かく観察することができる可能性がある。

そこで，次に，仮名表記率は何に起因するか，どのような言語的事象と相関が高いのかを確認するために，いくつかの視点から，仮名表記率の差異について検討する。

３.２　常用漢字表と仮名表記率の相関

漢語を漢字で表記するか仮名で表記するか，という選択において，最も関連が深いと思

われるのは，語を構成する漢字やその読みが常用漢字表に含まれているものか否かという点である。常用漢字は，公用文や新聞社・出版社での表記基準となっているため，漢語内に表外字が含まれる場合は，仮名に置き換えて表記されるのが表記原則だからである。

そこで，語を構成する漢字に，常用漢字表[4]に含まれない字（音）が含まれる漢語を「常用外」，常用漢字のみからなる漢語を「常用内」として層別し，各層における仮名表記率を求めたところ，表外字を含む漢語の平均は 20.9%，表内字のみからなる漢語の平均は 3.7%と大きく開きがあった。

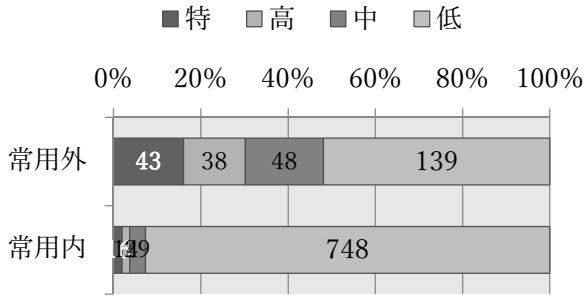


図 2　常用漢字表と仮名表記率

また，3.1 節に示した表記率による層別の分布を示した図 2 からは，表内字のみの漢語は圧倒的に仮名表記率が低いが（約 93%），表外字を含む漢語では，約半数が仮名表記率「中」以上，残りの半数が「低」であり，常用漢字かどうかは仮名表記率に大きく関わるものの，表外字を仮名表記とするかどうかには，かなり大きな幅があることが分かった。

３.３　出現頻度と仮名表記率の相関

前節において，仮名表記率は，常用漢字かどうかによって一意に決まるものではないことを示した。そこで，次に，語の使用状況と仮名表記率との関わりについて見ていきたい。

表 2　語彙レベルの基準・範囲と所属語数

レベル	基準値	頻度範囲	累積%	語数	語例（斜体は常用外，太字下線は仮名率「中」以上）
A	60%	9015-	60.01%	131	自分，問題，必要，時間，関係，人間，生活，会社，研究，意味，学校，**本当**，利用，現在
B	80%	1809-9014	80.00%	280	相当，材料，職員，移動，細胞，次第，気分，天皇，携帯，***所為***，興味，以来，距離，義務
C	90%	464-1808	90.01%	291	休憩，衣装，電波，寄付，体操，最悪，景色，*親戚*，親切，頻繁，*稽古*，感心，悪魔，***饂飩***
D	95%	157-463	95.00%	271	*氾濫*，***途轍***，堕落，白鳥，母音，催促，***火燵***，大麻，**褒美**，*傲慢*，変態，***暖簾***，伝言，繁盛
E	99%	20	99.04%	102	*漆喰*，*扶持*，面相，*躁鬱*，**野暮**，飛脚，銀山，剣幕，幻滅，*睡蓮*，*咀嚼*，*蹂躙*，母艦，*軋轢*

語の使用状況は，その語の日本語語彙における重要度や馴染み度（親密度）と相関がある(天野・近藤 2000，寺田・田中 2008 等)。高頻度の語の多くは，日本語の語彙において中心的・基本的な語であり，それらの語を構成する漢字は常用漢字に収録されている可能

[4] 調査データである BCCWJ は，2005 年までに出版された新聞・雑誌・書籍と，2008 年までに収集された Web データに基づくため，ここでは，2010 年に改訂された現行の常用漢字ではなく，1981 年内閣告示による常用漢字表を用いた。

性が高く，また表記や用法に大きな揺れはないことが推測される。一方，頻度の低い語は，馴染みが薄く，場合によっては難解・難読の語である可能性が高く，情報伝達の必要性から可読性の高い仮名表記が選択されるという背景が推測される。

そこで，実際に語の使用状況を確認するために，調査対象語を「語彙レベル」により層別し使用状況のラベルを与え，仮名表記率との相関を見る。「語彙レベル」は，語の使用頻度の累積度数による「カバー率」（対象データの延べ語数に対して累積頻度が占める比率）に基づき与えられる（田中 2013）。ここでは，田中 2013 による「語彙レベル」の求め方を参考に，独自に基準を設けてラベルを与え（表 2），仮名表記率との関連を見ていく。

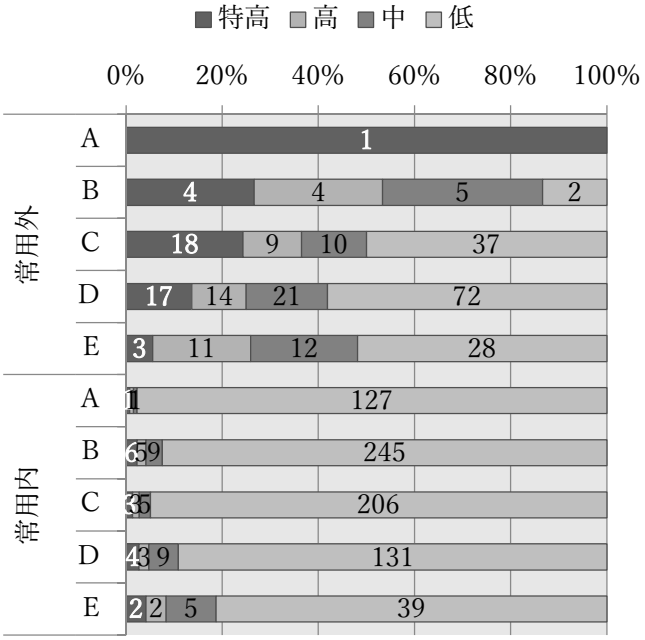


図 3　語彙レベルと仮名表記率の分布

図 2 に，常用漢字かどうかの層別で，語彙レベルごとの仮名表記率の分布を示す。表内字のみからなる漢語においては，総じて仮名表記率が低い中で，語彙レベルが高いほど仮名表記率は低く，語彙レベルが低いほど仮名表記率は高くなる傾向が見られる。

一方，表外字を含む漢語は高レベルの語が少ないものの，語彙レベルが高いほど仮名表記率は高い。この結果は，語の一般性が高いほど，ルールに基づく表記（表内字は漢字表記，表外字は仮名表記）が選択され，一般性が低いほど，それと異なる表記が許容されるという表記傾向の表れと見られる。

３.４　音転訛形と仮名表記率の相関

次に，通常の漢字音からの逸脱により，漢字と語形（音）とに乖離が見られる場合には，仮名表記される頻度が高くなると予想し，音声転訛形を持つか否かによる分析を行った。

(2) 私どもからすると、正直いって、対処法とか考えているだけ<u>面倒</u>くさい。

【出典】 BCCWJ サンプル ID : PB20_00106,青田吉弘(著)『情報化社会対話集』2002 年

(3) 毎日毎日水をとり替えて<u>めんど</u>くさいでしょ

【出典】 BCCWJ サンプル ID : PB27_00182, 渡辺襄（著)『窯焚き三昧』2002 年

音声転訛形とは，標準形で現れる例(2)に対して，例(3)に見られる語末長音短呼形（面倒：メンドウ→メンド）のようなものや，融合形（絶対：ゼッタイ→ゼッテー）などを指すが，加えて，一時的に現れる促音付加（最高：サイコウ→サイッコウ）や，意図的な連母音の長音化（携帯：ケイタイ→ケータイ）も含めた。

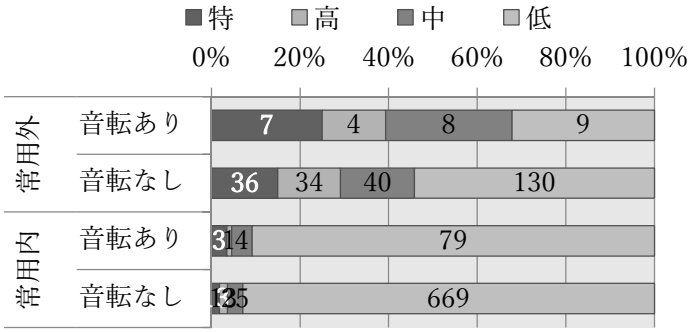


図 4　音声転訛形の有無と仮名表記率の分布

図 4 を見ると，常用漢字表外字を含む漢語では，音声転訛形のある語で，より仮名表記率の高い語が高い傾向にあることが分かる。ただし，常用漢字のみからなる漢語の場合は，そのような差が見られなかった。

3.5　語義分野と仮名表記率の相関

語の意味分野と表記の関連性については，既に先行研究によって動植物名が片仮名表記される傾向にあることが明らかにされているが，それらと常用漢字表との関連性，また，動植物名以外の仮名表記率に関与する分野の有無について検討するために，語義分野と仮名表記率の相関を見た。

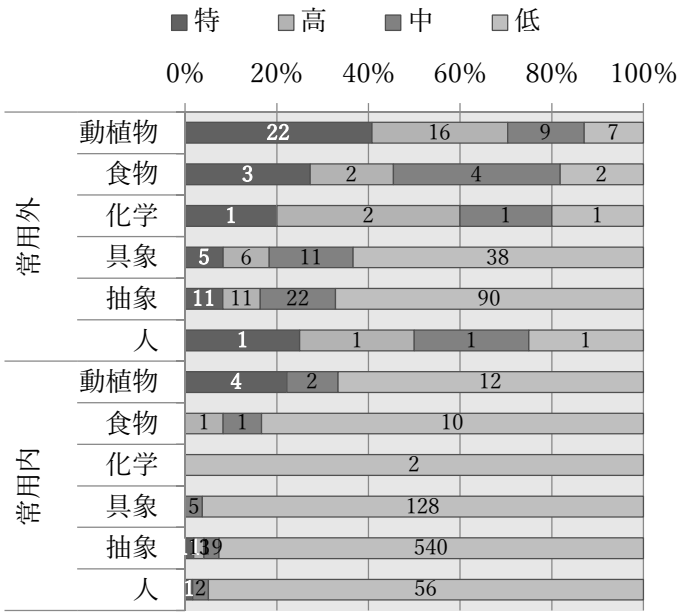


図 5　語義分野と仮名表記率の分布

まず，常用漢字のみからなる語に関しては，全体的に仮名表記率が低いものの，他分野に比して動植物名（「昆布」「隠元」など）で仮名表記率が高い「特」「高」の割合が高い。表外字でも同様に，動植物，加えて食物（「味噌」「饂飩」「煎餅」など）に関連する語彙で仮名表記率の高い語の割合が多く，抽象概念を表す語（「完璧」「贅沢」「傲慢」「躁鬱」など）で仮名表記率が低い。

通常漢字で表記されるべき，常用漢字のみからなる語で，動植物名に仮名表記率の高い語がある要因は，語義的な分野としての動植物名に表外字や外来語が多いほか，学術用語として用いられる際に片仮名で表記される用字法が浸透していることの影響とも思われる。また，動植物名は，例(4)で常用漢字のみからなる「大根」が片仮名表記されているように，他の語と列挙される場合，常用漢字か否かに依らず，統一的に仮名表記が用いられることが少なくない。このような表記の統一化・合理化が動植物語彙における仮名表記率の高さに関与しているのではないか。また同様に，動植物名は食品となるものが多く，意味分野が近接していることによって，動植物ではない食物の語彙についても，上述の統一化・合理化による仮名表記選択がなされている可能性が高い。

(4) ・<u>ダイコン</u>・コマツナ・コマツナ・ネギ・ハクサイ・ハス・ホウレンソウ・芽キャベツ・

ヤマイモ・ユリネ

【出典】 BCCWJ サンプル ID : PB15_00156, 田中元(著)『熟年世代からの元気になる「食生活」の本』2001 年

(5) 食料は、スイカ１０００円、花火１０００円、肉５１６円、たまご・たまねぎ・しょうゆ・さとう・みそ・シーチキン１３００円、きゅうり・とまと・レタス７００円、

【出典】 BCCWJ サンプル ID : LBq7_00020, 菅原道彦(著)『あそびの達人』2002 年

３.６　品詞と仮名表記率の相関

表 3　コーパスの品詞情報と品詞ラベル

コーパスの品詞情報	正規化したラベル
名詞-普通名詞-サ変可能	名詞・動詞
名詞-普通名詞-サ変形状詞可能	名詞・動詞・形状詞
名詞-普通名詞-形状詞可能	名詞・形状詞
名詞-普通名詞-副詞可能	名詞・副詞
その他「名詞」から始まる品詞	名詞
形状詞-タリ／形状詞-一般	形状詞
接尾辞-名詞的-一般	名詞
副詞	副詞

次に，それぞれの漢語に付加されている品詞情報によって漢語を層別し，仮名表記率との相関を見る。コーパスの形態論情報として付与された品詞については, 表 3 に示す正規化を行い，「結構」のように複数品詞を持つ語（形状詞-一般，副詞，名詞-普通名詞-一般）は同様に「名詞・副詞・形状詞」のように品詞を繋いだラベルを付けた。

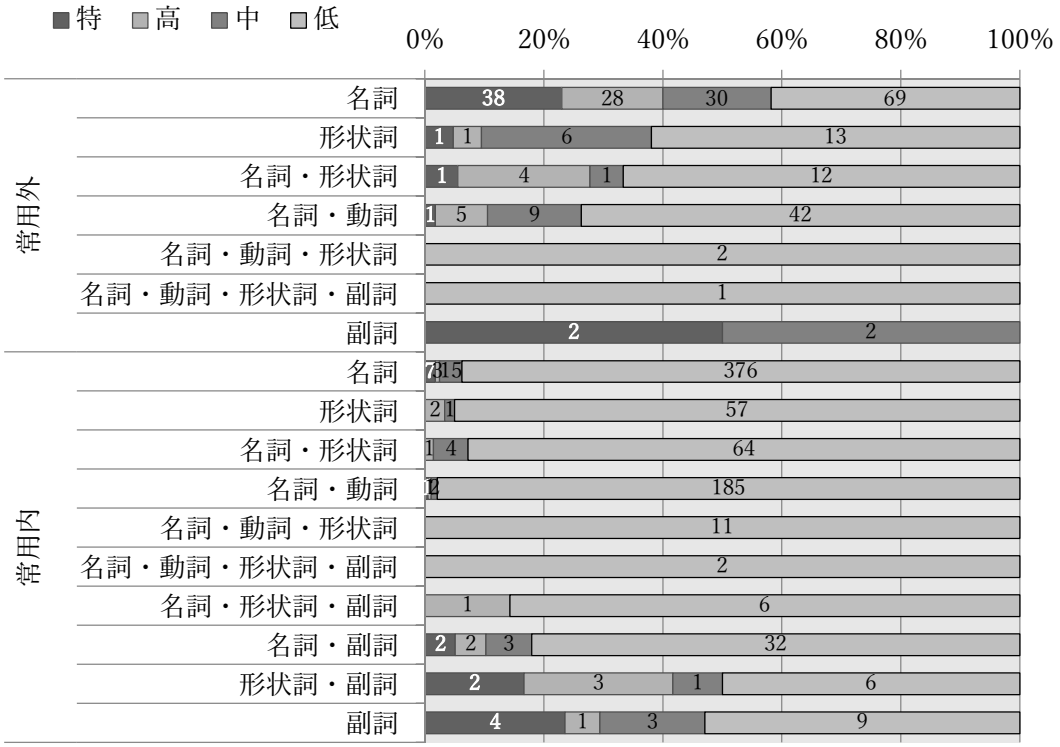


図 6　品詞と仮名表記率の分布

図 6 より，表外字を含む漢語では，動詞用法を含まない語（「勿論(副詞)」「所為(名詞)」「馬鹿(名詞・形状詞)」など）で仮名表記率の高い語の割合がより高く，常用漢字のみから

なる語では「副詞」の用法を持つ漢語（「沢山(形状詞・副詞)」「多分(名詞・形状詞・副詞)」「是非(副詞)」など）で，仮名表記率の高い語の割合が高いことが分かる。ここから，動詞は仮名表記になりにくく，副詞は仮名表記になりやすい傾向にあることが予想される[5]。

ただし，この品詞ラベルはコーパスの形態論情報に基づくもので，個々の漢語の使用実態や用法に基づくものではないため，用法と仮名表記率との相関を確認するために，いくつかの語を取り上げ，実際の用法に基づいて仮名表記率の分布を見てみよう。ここでは，「是非」（表内字）を取り上げ，例 6 のように，格助詞等に接続する，あるいは，例 7 のように連体修飾を受け，「是と非，良し悪し」の意で用いられているものを「名詞」，例 8 のように連用修飾成分になり「強く願う」意を表すものを「副詞」として，用例の分析を行った。

(6) それが不服とおっしゃるならば、天宮において**是非**を論じられるがよろしい。
【出典】 BCCWJ サンプル ID：PB19_00498, 井上祐美子(著) 『乱紅の琵琶』1950 年
(7) 今回参拝を止めた理由と参拝の**是非**。
【出典】 BCCWJ サンプル ID：OC05_01310, Yahoo!知恵袋, 2005 年
(8) 一度は**ぜひ**お目に懸からなければなるまいと思います
【出典】 BCCWJ サンプル ID：PB32_00008 杉山秀子(著)『プロメテウス』2003 年

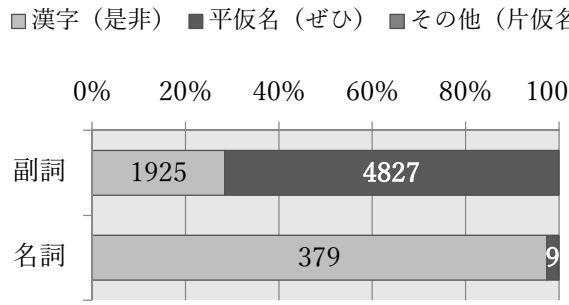


図 7 「是非」の用法と表記の相関

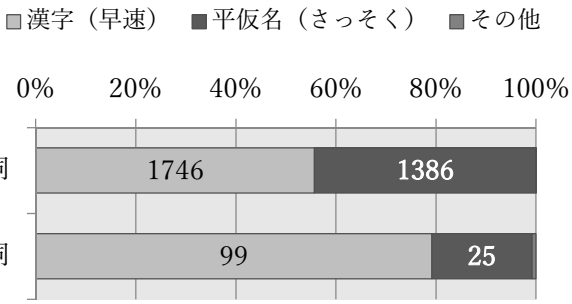


図 8 「早速」の用法と表記の相関

図 7 より，常用漢字のみからなる漢語「是非」は，名詞用法では 97%で漢字表記が選択されているのに対して，副詞では漢字表記 28%に対して，仮名表記が 71%を占め表記傾向が逆になっている。また，「是非」のように用法による意味の相違が明確なもの以外の例として，「早速」（常用内，語彙レベル B，仮名表記率中）を取り上げ，「早速やる」のような副詞用法と，「早速ですが」「早速に」「早速の回答」のような名詞用法における表記を確認すると，図 8 に示す通り，同様に副詞用法でより仮名表記が選択される傾向にあることが分かる。

このように，副詞において仮名表記が選択される傾向は，例えば漢字使用に関する指針となりうる「公用文における漢字使用及び送り仮名の付け方について」（昭和 57 年 1 月 14

[5] 漢語動詞は，構成要素となる漢字の字義によって想起される意味と語義が緊密であるため漢字表記を捨てにくいが，漢語副詞の多くは，元の語義からの意味変化により副詞用法が生まれており，構成漢字の字義と語義の結びつきが希薄であること，また，漢語動詞は実質的な意味を持つのに対して，漢語副詞は程度や時間などの意味を付加する機能的な性質が強いため，機能語や形式名詞・補助動詞等と同様，仮名表記が選択されやすいなどの背景があるものと思われる。

日例規（総）第 2 号）において「原則として、仮名で表記する副詞」として挙げられる「かなり　ふと　やはり　よほど」に含まれない語においても多く見られる傾向であり，漢字か仮名かという表記の選択が，①常用漢字か否かという知識ベースの区分に寄らず，②用いられ方（品詞）差により行われている点は，3.5 節に指摘した，意味分野による統一的な表記傾向と同様，統一性・合理性を志向した表記法の一つと意味づけることができる。

３.７　レジスターと仮名表記率の相関

前節まで，BCCWJ の全体に対する表記傾向の分析を行ったが，採取した用例がどのような媒体に掲載されたものかによって，表記の傾向が異なる可能性があるため，本節では，コーパスの付加情報「レジスター」を用い，使用媒体と表記の関係について検討する。

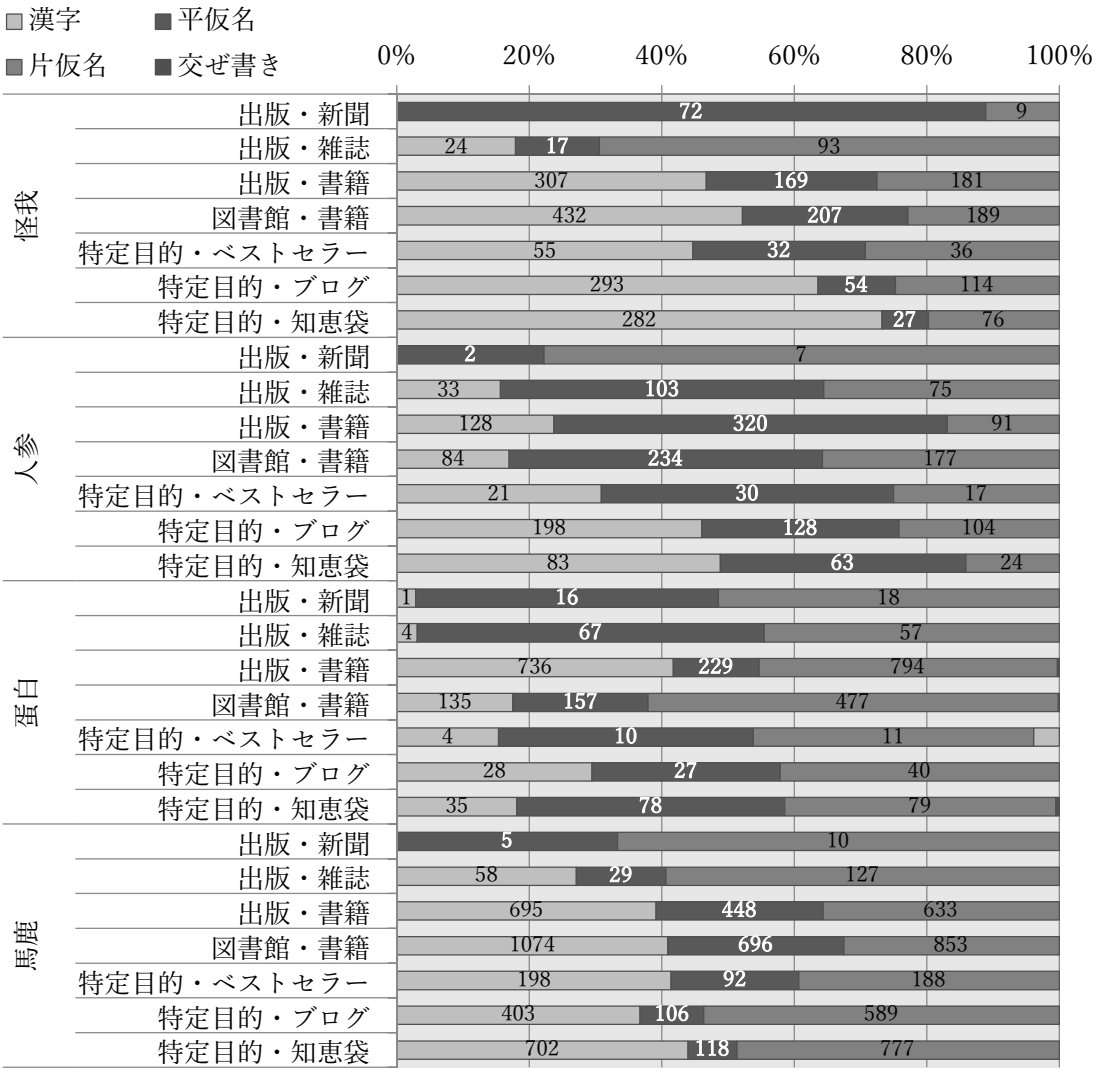


図 9　表外字を含む漢語のレジスター別表記分布

調査対象とした 1075 語から，適切な用例数が確保できる語彙レベル B（語彙レベル A は表記の揺れが少なく，また用例数が多すぎるため分析が困難であるため除く）の語で，仮名表記率が「高」の語から，表外字（音）を含む「馬鹿」「蛋白」「怪我」「人参」，常用漢

字のみからなる「多分」「是非」「随分」「大抵」を取り上げ，表記実態を調査した[6]。

表外字を含む語について，図 9 を見ると，語によるばらつきがあるものの，概観すれば，新聞ではほぼ 100%に近く仮名表記が用いられ，次いで雑誌，書籍と漢字表記率が増え Web 媒体のブログ・知恵袋では他より漢字表記が多いという傾向が見て取れる。一方，常用漢字のみの語について図 10 を見ると，語によって新聞，雑誌，書籍の分布は様々だが，総じて Web 媒体のブログと知恵袋で漢字表記が多い。

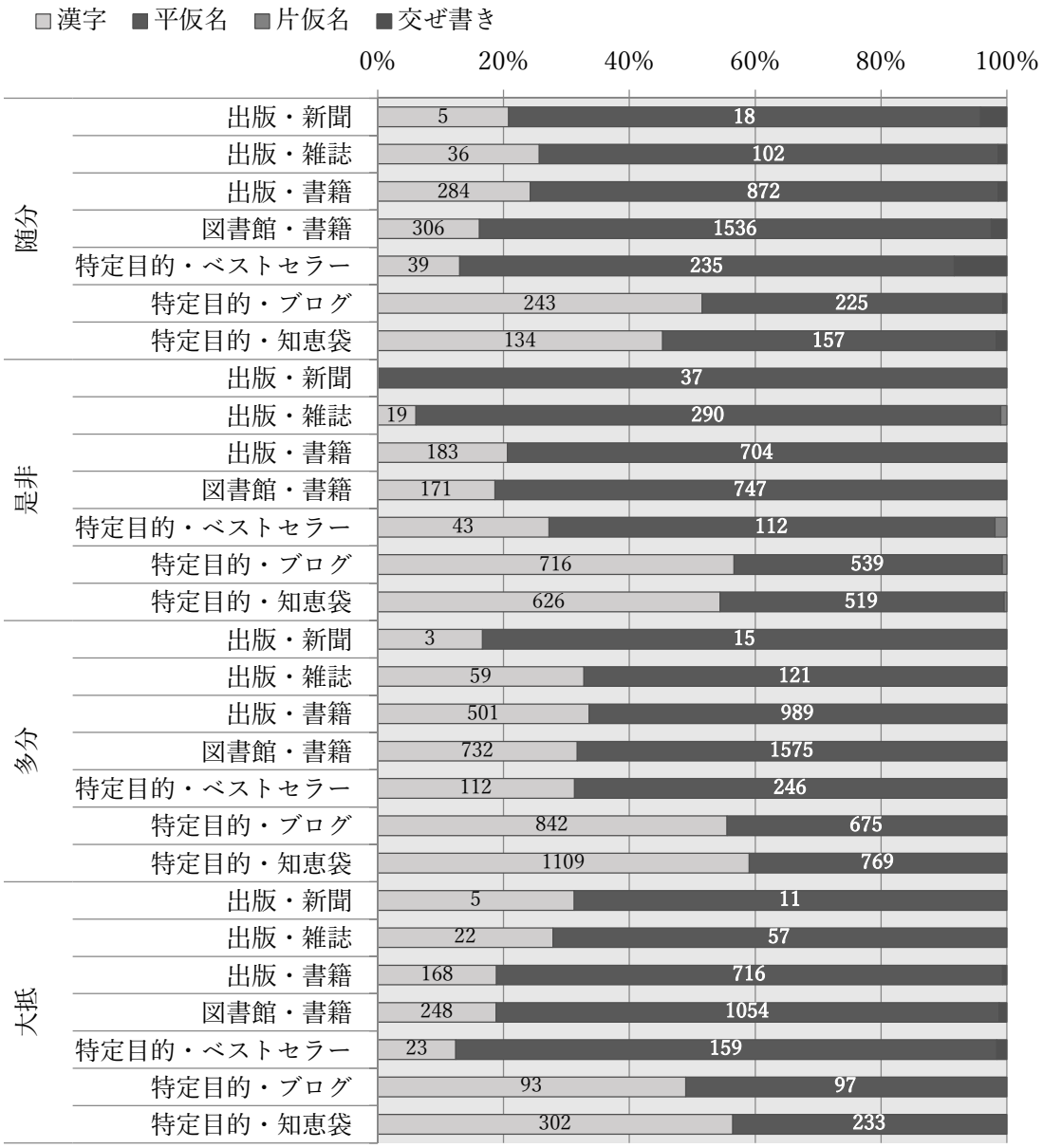


図 10　常用漢字のみからなる漢語のレジスター別表記分布

また，図 11 として，先に品詞による表記傾向の異なりを示した「是非」について，レジ

[6] 特定目的 SC のうち国会会議録，広報誌，韻文，法律，白書は，調査対象語を含まない（あるいは極めて低頻度）場合があるため，分析対象から除外した。

スターの差を含めた表記分布を示す。品詞による表記の使い分けについても，新聞・雑誌ではこれが顕著であり，書籍類でも明確な差が見られる。一方 Web 媒体のブログ・知恵袋では，用法の差は見られるものの，副詞用法における仮名表記は半数以下であり，他の媒体に見られる仮名表記の優位性が見られない。

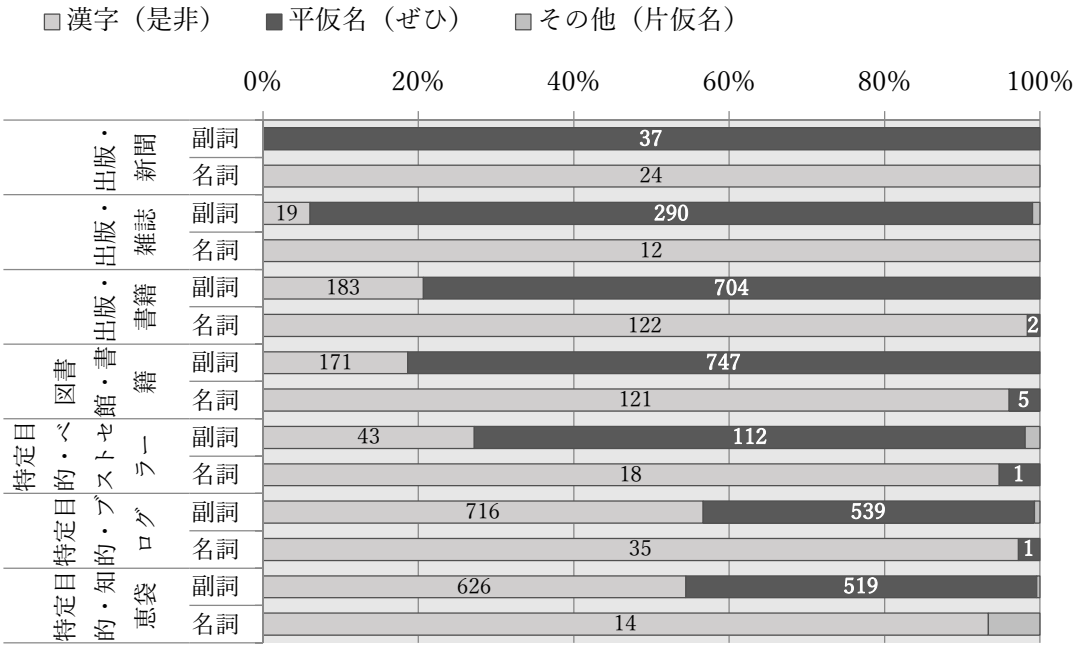


図 11　レジスター別にみた「是非」の品詞別表記分布

ここから，新聞社や出版社の記者・編集者による用字は統制が利いており，表外字や副詞用法を持つ語で平仮名表記の比率が高いが，個々の著者の用字がより反映されやすい書籍においては，表外字の漢字表記が新聞・雑誌より多く，更に標準的表記への志向性や表記統制の影響を受けにくい Web 媒体の用字においては，字体や用法による表記選択の傾向が他の媒体より低いことが分かる。Web 媒体の個人的な用字においては，仮名漢字変換の影響として，変換候補に現れる漢字表記を無意識・無意図的に選択している可能性が高い。

４．まとめ

以上，BCCWJ の網羅的な表記実態調査に基づく漢語の仮名表記率（3.1 節）から漢語の層別を行い，複数の指標との関連性について検討することで，本来漢字で表記される漢語が仮名で表記される背景として，以下の六つの要因があることを指摘した。

① 字体特徴：常用漢字表外字（音）を含む語は仮名表記率が高いが，表内字でも仮名表記率の高い語がある。…3.2 節
② 語彙レベル：出現頻度の高い漢語ほど仮名表記率は低い。…3.3 節
③ 音声変位形の有無：「面倒／メンドウ」に対する「メンド」のような音声転訛形を持つ語は仮名表記率が高い。…3.4 節
④ 意味分野：動植物やそれと近接する食物の分野では仮名表記率が高い。…3.5 節
⑤ 品詞：副詞用法を持つ語は仮名表記率が高く，複数の品詞があるものは，表記の使い分けがある場合が多い。…3.6 節

⑥ レジスター: 新聞や雑誌は標準的な表記選択が行われているが，Web 媒体のテキストではそれによらず，仮名表記率が低い傾向がある。…3.7 節

また，仮名表記選択に最も強い影響を与えると思われる①の字体特徴にかかわらず，意味分野や品詞において特定の語彙群が，同様の傾向（字体特徴から予測される標準的な表記から逸脱した表記を選択する）を見せることがあることを指摘し，これらの表記の選択や嗜好には，類似性に基づく合理化作用が働いている可能性を示した（3.5 節，3.6 節）。

５．おわりに

本発表では，漢語の仮名表記の背景について複数の要因を示したが，要因相互の関係性については検討が至らなかった。多変量解析手法などを取り入れ，漢語の仮名表記の要因について更に整理を進めたい。また，発表者は，近代から現代への漢語の変化を研究の主題としており，漢語の仮名表記についても，実態調査から近代と現代とで差異が見られることが分かっている。これについては，機を改めて報告したい。

謝　辞

本稿は，日本学術振興会特別研究員奨励費 16J08872「コーパスを利用した近現代漢語の表記・語法の多様性に関する計量的・通時的研究」（代表：間淵洋子）による成果の一部である。

文　献

天野成昭・近藤公久(2000)『日本語の語彙特性―朝日新聞の語彙・文字頻度調査〈第 7 巻〉頻度 (NTT データベースシリーズ)』三省堂.

石井久雄(2001)「ひらがなの文法性・語彙性」『同志社大学留学生別科紀要』1, pp.3-16

岩原昭彦・八田武志(2004)「文字言語における感情的意味情報の伝達メカニズムについて」Cognitive Studies,11:3, 271-281.

柏野和佳子(2014)「「コーパス」でさぐる和語や漢語のカタカナ表記の実態」高田智和・横山詔一編『日本語文字・表記の難しさとおもしろさ』彩流社, pp. 86-105.

児玉徳美(2013)「日本語の用字用語」『立命館文学』63, pp.410-392.

増地ひとみ(2013)「テレビ番組の文字情報における文字種の選択―番組のジャンルと語用論的要素に注目して―」『早稲田日本語研究』22, pp.24-35.

中山恵利子(1998)「非外来語のカタカナ表記」『日本語教育』(日本語教育学会)96, pp. 61-72.

成田徹男・榊原浩之(2004)「現代日本語の表記体系と表記戦略―カタカナの使い方の変化―」『人間文化研究』(名古屋市立大学)2, pp. 41-55.

則松智子・堀尾佳代子(2006)「若者雑誌における常用漢字のカタカナ表記化―意味分析の観点から―」『北九州市立大学文学部紀』72, pp.19-32

寺田博視・田中久美子(2008)「単語親密度と単語頻度の関係に関する一考察」『言語処理学会第 14 回年次大会発表論文集』, pp.713-716.

臼木智子(2008)「雑誌の片仮名表記―基準から外れる表記について―」国学院大学大学院紀要. 文学研究科 40, pp.265-280.

関連 URL

コーパス検索アプリケーション『中納言』　https://chunagon.ninjal.ac.jp/

『日本語歴史コーパス』短単位アノテーション作業効率化に向けた形態素解析用辞書『UniDic』の段階的特殊化の検討 –近松コーパスを例として–

岡 照晃 (国立国語研究所コーパス開発センター) *

An Examination of Stepwise Specialization of Morphological Analysis Dictionary "UniDic" for Efficient Word Annotation on "Corpus of Historical Japanese" –The Case of Chikamatsu Corpus–

Teruaki Oka (National Institute for Japanese Language and Linguistics)

要旨

本論文では，現在，国語研の通時コーパス構築プロジェクトで整備中の近世前期の上方資料である『近松門左衛門 世話物浄瑠璃』への短単位形態論情報アノテーションの効率化を目的に，形態素解析器 MeCab の追加学習機能を使い，既存の『洒落本』用の短単位解析用辞書から段階的に，近松専用短単位解析用辞書を作成する方法について述べる．具体的には，まず比較的時代の近い洒落本解析用辞書を，上方の洒落本コーパスのみで上方の洒落本解析用辞書にアダプテーションする．次に作成した上方の洒落本解析用辞書を，同じく上方の資料である近松コーパスで近松資料解析用辞書にさらにアダプテーションする．本手法により，従来手法よりも高い精度（語彙素認定 F1 値，地の文：86.85 → 89.60，会話文：85.07 → 88.82）で，近松資料を解析できることを確認した．また本論文で作成した短単位解析用辞書を使った近松資料のコーパス化作業が現在，進行中である．

1. はじめに

国立国語研究所では現在，日本語の通時コーパス『日本語歴史コーパス』(近藤泰弘 2012)（以下，**CHJ**）の構築を進めている．CHJ の特徴として，国語研の規定する言語単位**短単位** (伝康晴ほか 2007, 近藤泰弘 2015) での形態論情報（e.g., 品詞，活用，発音，語種…）がアノテーションされていることがある．この形態論情報の人手アノテーション効率化のため，CHJ 構築では各時代専用に形態素解析器『MeCab』(Kudo et al. 2004) 用の『解析用辞書 UniDic』(小木曽智信ほか 2013, 小木曽智信ほか 2014)（以下，単に**辞書**という場合，この**解析用辞書**を指す）のコストを学習し，MeCab による自動解析の後，解析結果を人手修正する，という作業方針を採っている．そのため自動形態素解析の結果の精度は人手作業の負担と直結し

*teruaki-oka {at} ninjal.ac.jp

ている．もし自動解析結果の精度が高ければ，人手修正が必要な箇所も少なく，形態論情報アノテーションにかかる人的・時間的負担も少ない．しかし反対に自動解析結果の精度が低いと，（人手で 0 からアノテーションするよりも格段に少労力だが）作業にかかる負担は大きい．

現状において，辞書は文献 (鴻野知暁ほか 2014) に基づき，各時代ごとに 1〜2 個（文語，口語）という粒度で作成されている．これは MeCab の内部モデルである CRF (Lafferty et al. 2001) の学習用コーパスをまとまった量で確保し，解析結果の精度を高めるためと，整備対象となる多様な資料に対して，できるだけ汎用的に使用でき，再利用可能な辞書を目指したためである．この汎用化は，外部の研究者に作成した辞書を公開することを視野に入れた考慮でもある．

これに対し本論文では，CHJ 構築現場での形態論情報アノテーション作業のさらなる効率化を第一目的に，より解析対象の資料に適合した辞書の構築について述べる．汎用的な辞書は多様な資料に使用できる反面，全体で見た時の精度は高くとも，各ドメインに特徴的な箇所の多くで，解析エラーが生じやすくなる．またドメインが一様でないコーパスを混用して学習するため，各ドメイン間でサイズが異なる場合は，その影響を受けて解析も偏りやすい[(1)]．実際，文献 (市村太郎 2014) では，近世の資料の地の文と会話文の文体・文法体系の差を指摘し，文献 (市村太郎ほか 2016) において，それ以前では学習時に混用されていた地の文と会話文を分け，それぞれに専用の辞書を学習することで解析結果の精度向上を報告している．

各資料に特化した辞書を作成する際の問題は，汎用的な辞書を作成する場合よりも学習用コーパスが絞り込まれてしまうことがある．そこで提案手法では，これまで通りの汎用辞書（のモデル）を始点として，段階的に学習用コーパスを絞り，MeCab の**追加学習**[(2)]機能によって，汎用辞書を段階的に特殊化していく方針を採る．この手法の利点として，① 整備中の資料であっても少量の人手アノテーションが終われば，逐次，辞書の学習に追加していけること，② 各時点での辞書を別の資料用辞書の学習の始点として再利用できること，③ これまで通り汎用の解析用辞書は作成していくため，国語研内部だけの特殊化した辞書と同時に，外部の研究者に有用な公開用辞書も同時に作成できることがある．

本論文では，CHJ に含まれる近世の後期の資料『洒落本』（以下，**CHJ 洒落本**）および，現在整備を進めている同じく近世の上期上方資料『近松門左衛門 世話物浄瑠璃』（以下，**CHJ 近松**）を対象に，洒落本用辞書の学習から開始し，江戸・上方各専用の辞書を作成，さらに上方専用の辞書から近松専用の辞書を作成する流れを説明する．

2. 近世期資料『洒落本』の自動形態素解析の現状と問題

CHJ では，古代語〜近代語についての代表的な資料を集め，日本語の通時的な研究に利用可能なコーパスとして順次公開を進めている．このうち『**CHJ 江戸時代編**』としてコーパス化の対象となっているのが，近世後期の洒落本と人情本，そして近世前期の上方資料であ

(1) 注） 学習時（or 前），何かしらの正規化により，このサイズのアンバランスを解消する場合は別．

(2) MeCab の公式ページ: `http://taku910.github.io/mecab/learn.html` (2017/2/4 現在) では「再学習」と記載されている機能であるが，「辞書を 0 から学習し直す」という意味と混同しやすく，また同ページに「再学習とは，現在の学習済みモデルと少量の追加学習データからモデルを再構築する仕組みです」とあることから，本論文ではこれを「追加学習」と呼称する．

表 1　本稿で使用した CHJ 洒落本の概要：全 17 作品．作品詳細は 5.1 節を参照．

地の文/会話文	地域	総文数	総短単位数	総文字数
地の文	江戸	4,255	29,983	50,348
	上方	2,384	19,525	32,451
	計	6,639	49,508	82,799
会話文	江戸	4,282	47,738	79,985
	上方	2,523	31,109	51,929
	計	6,805	78,847	131,914
	計	13,444	128,355	214,713

る近松門左衛門の世話物浄瑠璃である．この中で最も整備が進んでいるのは，『洒落本大成』(洒落本大成編集委員会 1978-88) を底本とした CHJ 洒落本であり，既に短単位アノテーション済みコーパスが 3 作品試験公開されているが(3)，国語研の内部的には，これ以上の資料についてアノテーションが進められている．また CHJ 近松に関してもアノテーションが進みつつあり，洒落本用の解析辞書 (市村太郎ほか 2016) を使った自動解析結果を順次人手で修正する作業を行なっている．

しかしながら文献 (市村太郎ほか 2016) にも取り上げられている通り，洒落本用の解析辞書を使った解析結果の精度は他の時代の辞書に比べて低い．文献 (小木曽智信ほか 2013) によると，整備が進んでいる中古和文系資料，および近代文語論説文では発音形認定のレベルで 96〜97 の F1 値を達成しているが，最新の文献 (市村太郎ほか 2016) の報告でも洒落本については，未だに語彙素認定のレベルで F1 値 90 台に止まっている(4)．文献 (小木曽智信ほか 2013) においては，語彙素認定で F1 値 95 を達成するためには最低でも 5 万短単位が必要と言われている．しかし表 1 を見ればわかる通り，CHJ 洒落本はすでにその要件を満たしている．

この問題について，文献 (市村太郎 2014) では以下の 2 つの原因を上げている．

問題 I. 表記の多様性　近世の資料の中には，一貫しない仮名遣いや送り仮名，片仮名平仮名の混ぜ書き，踊字の使用，濁点無表記 (岡照晃ほか 2013) といった，多様な表層形が出現し，辞書に網羅的に登録することが難しい(5)．

e.g., 動詞「言う」+ 接続助詞「て」で構成される表層形のバリエーションの一部：
「いつ｜て」「いゝ｜て」「言つ｜て」「いひ｜て」「いう｜て」「言ふ｜て」「云ひ｜て」「言｜て」「云｜て」「ゆう｜て」「いツて」(6)

問題 II. 地の文-会話文間の文体・文法体系の混在　CHJ の活用型が大きく文語活用と口語活用に分かれているのに対し，原文一致以前の資料である洒落本では，地の文は文語，会話文は口語で書かれている．そのため地の文と会話文とで活用型のアノテーションが別となり，地の文と会話文を辞書（=CRF のモデル）の学習・解析時に混用していると，活用型の認定でエラーが多発する．

これらに対し，問題 I については，既に文献 (市村太郎 2014) で，アノテーションレベルでの

(3) http://pj.ninjal.ac.jp/corpus_center/chj/edo.html
(4) いずれも「未知語なし」という条件の下での評価．
(5) 網羅したとしても，辞書サイズが肥大化し，かつ解析上の曖昧性も大きくなる．
(6) 「｜」は短単位の境界を表す．

対応がある程度図られており，文献 (岡照晃ほか 2014) では，その人手作業を自動化する試みが提案されている．問題 II については文献 (市村太郎ほか 2016) で，地の文と会話文とに分けて辞書を学習・解析する手法が提案されている．

しかしながら先に述べた通り，最新の文献 (市村太郎ほか 2016) でも洒落本解析用辞書の解析結果の精度は未だに他の時代の辞書に追い付いていない．問題 I が完全には解決されていないことも大きな原因ではあるが，文献 (市村太郎 2014) で言及されていないさらなる問題として，CHJ 洒落本の中に上方（関西）と江戸（関東），2 つの地域の資料が混在していることがある．

問題 III. 江戸-上方間の文体・文法体系の混在　例えば，先に上げた動詞「言う」+ 接続助詞「て」の表層形の一つである「言｜て」は，江戸では「イッ｜テ」と発音形がアノテーションされるが，上方では方言の違いから「ユー｜テ」とアノテーションされる．しかし江戸の資料と上方の資料が混在する状態で辞書の学習を行なった場合（図 1 の辞書 ii) ），「言｜て｜おく」[7]の発音形の解析結果は，上方の資料でも「イッ｜テ｜オク」となる．これは表 1 を見ればわかる通り，上方の資料よりも江戸の資料の方がデータとして多いため，解析時に曖昧な箇所が江戸側に倒れてしまうからである．もし逆に，上方の資料の方が多かった場合は，この解析結果は江戸・上方両方の資料で「ユー｜テ｜オク」となるだろう．また語彙素認定に関わる問題例として，助動詞「んす」がある．「んす」は語彙素「んす」の表層形として，洒落本解析用辞書にエントリされているだけでなく，語彙素「ます」の表層形としてもエントリされている．そして後者は主として江戸の資料にしか現れないという特徴がある．

このように，CHJ 洒落本の中に混在している文体・文法体系は地の文-会話文の別だけでなく，江戸-上方という方言の違いもある．この問題を解決するための方法として，地の文-会話文の問題を解決したときと同じように，辞書の学習用コーパスを江戸と上方で更に分割するという対応が考えられる．ただし，そもそも活用型から異なるような地の文-会話文の差に対して，江戸-上方の文体・文法体系の違いは著しく大きいわけではない．そのため学習用コーパスの縮小により，現状よりも解析結果の精度が悪くなる可能性がある．

3. 提案手法: MeCab の追加学習機能を用いた短単位解析用辞書の段階的特殊化

3.1 追加学習を用いた洒落本用解析辞書からの洒落本（江戸）・洒落本（上方）解析用辞書の作成

前節の問題 III に対し，本論文では，まず文献 (市村太郎ほか 2016) と同様に，地の文-会話文の別だけで 2 つの辞書を作る．そして次にそれぞれの辞書のモデルを江戸-上方の別で絞り込んだ学習用コーパスで MeCab の追加学習（**Regularized Adaptation**[8]）を実行する方針を採る．これにより，学習用コーパスが江戸-上方に分けて少量化される問題を回避しつつ，それぞれに適応した辞書が実現できる．

[8] 初期のモデルパラメータをできるだけ変更せずに，新しい学習データにできるだけ適応するような新しいモデルを学習する手法．MeCab の-M オプション（初期モデルの指定）で使用可能．詳細は文献 (Imamura 2013) を参照．

表 2　本稿で使用した CHJ 近松の概要.

	文数	総短単位数	総文字数
地の文	729	11,055	17,991
会話文	1,317	18,520	29,404
計	2,046	29,575	47,395

上記の手法により，次の 6 つの解析用辞書が作成される．各辞書の番号 i)〜vi) は，図 1 および，実験結果の表と対応している．

i): CHJ 洒落本（江戸 + 上方）の地の文から学習された辞書
ii): i) の辞書に CHJ 洒落本（江戸）の地の文だけ使って追加学習した辞書
iii): i) の辞書に CHJ 洒落本（上方）の地の文だけ使って追加学習した辞書
iv): CHJ 洒落本（江戸 + 上方）の会話文から学習された辞書
v): iv) の辞書に CHJ 洒落本（江戸）の会話文だけ使って追加学習した辞書
vi): iv) の辞書に CHJ 洒落本（上方）の会話文だけ使って追加学習した辞書

4. 追加学習を用いた洒落本（上方）解析用辞書からの近世前期の上方資料『近松門左衛門 世話物浄瑠璃』用解析辞書の作成

『近松門左衛門 世話物浄瑠璃』は近世前期の上方資料であり，CHJ 洒落本全体と比較した場合，まず近世前期と後期で時代に差があるほか，資料自体が上方に限定されているという特徴がある．また洒落本と異なり，地の文と会話文の文体差が小さいことも特徴の一つである．ただし CHJ 近松では，地の文と会話文の認定をそれぞれ実施し，その上で活用型に文語-口語の別でアノテーションを施している．表 2 に示す通り，CHJ 洒落本よりも少ないながら，CHJ 近松も既に形態論情報アノテーションが完了している部分が存在する．そこでこのアノテーション作業をさらに効率化するため，前節の洒落本（上方）で追加学習した辞書（辞書 iii)，vi)）に，アノテーション済みの CHJ 近松を追加学習することで CHJ 近松に特化した辞書を作成する．

よって，前節の辞書 i)〜vi) に加えて，さらに以下の 2 つの解析用辞書が作成される．各辞書の番号 vii), viii) は，図 1 および，実験結果の表と対応している．

vii): 前節 iii) の辞書に CHJ 近松の地の文だけ使って追加学習した辞書
viii): 前節 vi) の辞書に CHJ 近松の会話文だけ使って追加学習した辞書

5. 洒落本解析用辞書から近松解析用辞書までの段階的特殊化の性能評価実験

5.1 実験設定

本実験では，前節までに述べた以下の 8 つの辞書（再掲）を図 1 のように追加学習によって段階的に構築していき，各評価用コーパスでの性能を評価する．各辞書の番号 i)〜viii) は，図 1 および，実験結果の表と対応している．

i): CHJ 洒落本（江戸 + 上方）の地の文から学習された辞書
ii): i) の辞書に CHJ 洒落本（江戸）の地の文だけ使って追加学習した辞書
iii): i) の辞書に CHJ 洒落本（上方）の地の文だけ使って追加学習した辞書

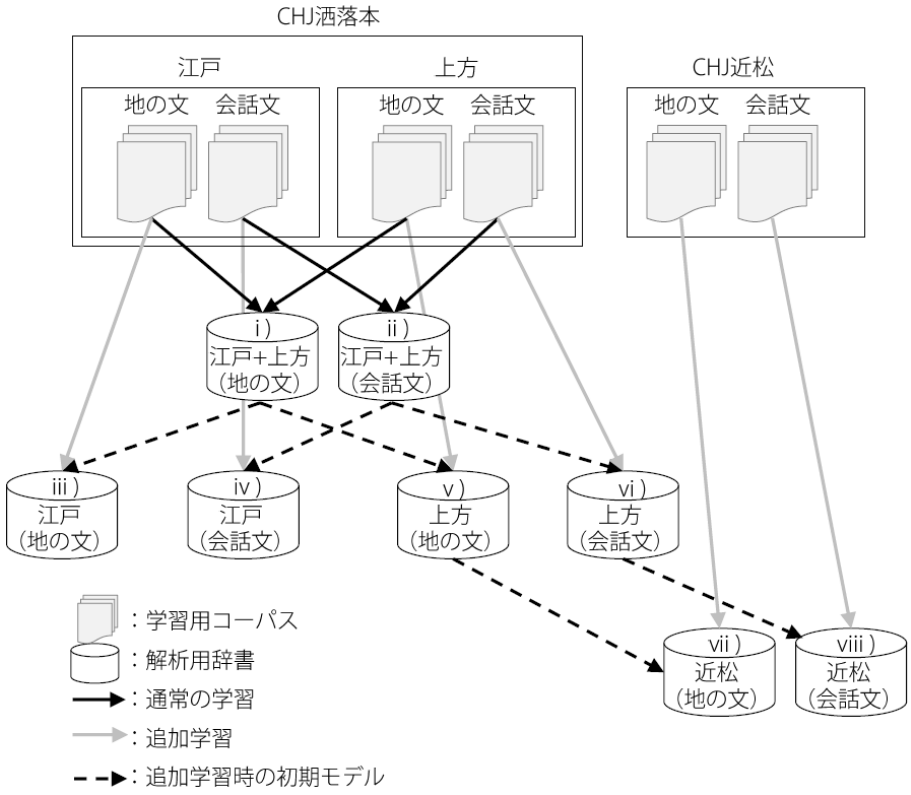


図 1　MeCab の追加学習機能を利用した解析用辞書の段階的特殊化.

iv): CHJ 洒落本（江戸 + 上方）の会話文から学習された辞書
v): iv) の辞書に CHJ 洒落本（江戸）の会話文だけ使って追加学習した辞書
vi): iv) の辞書に CHJ 洒落本（上方）の会話文だけ使って追加学習した辞書
vii): iii) の辞書に CHJ 近松の地の文だけ使って追加学習した辞書
viii): vi) の辞書に CHJ 近松の会話文だけ使って追加学習した辞書

表 3 と，表 4 にそれぞれ学習・評価用コーパスの内訳を示す．また整備中の CHJ 洒落本より使用する資料は以下の 17 作品である．

- 江戸
  - 郭中奇譚
  - 南閨雑話
  - 甲駅新話
  - 当世左様候
  - 花街鑑
  - 花街寿々女
  - 総籬
  - 仕懸文庫
- 上方
  - 原柳巷花語
  - 無論里問答

表 3　学習用コーパスの内訳.

地の文/会話文	学習用コーパス	総文数	総短単位数	総文字数
地の文	江戸（CHJ 洒落本）	3,830	27,016	45,412
	上方（CHJ 洒落本）	2,146	17,373	28,811
	江戸 + 上方（CHJ 洒落本）	5,976	44,389	74,223
	CHJ 近松	657	9,994	16,274
会話文	江戸（CHJ 洒落本）	3,854	43,312	72,501
	上方（CHJ 洒落本）	2,271	27,799	46,462
	江戸 + 上方（CHJ 洒落本）	6,125	71,111	118,963
	CHJ 近松	1,186	16,632	26,404

表 4　評価用コーパスの内訳.

地の文/会話文	評価用コーパス	総文数	総短単位数	総文字数
地の文	江戸（CHJ 洒落本）	425	2,967	4,936
	上方（CHJ 洒落本）	238	2,157	3,640
	江戸 + 上方（CHJ 洒落本）	663	5,119	8,576
	CHJ 近松	72	1,061	1,717
会話文	江戸（CHJ 洒落本）	428	4,426	7,484
	上方（CHJ 洒落本）	252	3,310	5,467
	江戸 + 上方（CHJ 洒落本）	680	7,736	12,951
	CHJ 近松	131	1,888	3,000

- 虚辞先生穴賢
- 阿闍陀鏡
- 昇平楽
- 誰か面影
- 興斗月
- 風流裸人形
- 箱まくら

解析用辞書のエントリはすべての辞書で共通とし，文献 (鴻野知暁ほか 2014) の時代情報を使った絞り込みによって UniDic データベース (小木曽智信ほか 2014) から取り出した 1,442,977 の表層形（辞書のキー）がエントリされている．このため次節に記載の実験結果は文献 (市村太郎ほか 2016) と同じく，未知語なしの条件下で実施したものとなっている．

MeCab の学習時の引数は，並列化オプションを除きすべてデフォルトのまま使用した．また CRF の正規化項のハイパーパラメータ C (=σ^2) も，全学習において共通にデフォルトの 1.0 に設定した．MeCab の学習時に使用する設定ファイルは文献 (市村太郎ほか 2016) と同じものを使用している．

評価は文献 (小木曽智信ほか 2013, 小木曽智信ほか 2014) と同じく，境界認定，品詞認定，語彙素認定，発音認定の 4 段階で F1 値を評価する．日本語の自動形態素解析における F1 値の計算方法は文献 (Kudo et al. 2004) を参照．境界認定は，文中での開始位置と終了位置が両方正しく認定できた（正解データと一致した）短単位数を評価している．品詞認定は，境界認定をパスした短単位の内，品詞大分類，中分類，小分類，細分類，活用型，活用形がすべて正しく

表 5　各評価用コーパス「地の文」における各辞書の短単位自動解析結果の精度（F1 値）.

評価用コーパス（地の文）	辞書	学習用コーパス	境界認定	品詞認定	語彙素認定	発音認定
江戸 + 上方	i)	江戸（地の文）+ 上方（地の文）	**97.19**	**91.93**	**90.11**	**89.43**
	ii)	追加学習：→江戸（地の文）	96.88	91.38	89.45	88.70
	iii)	追加学習：→上方（地の文）	97.09	91.45	89.58	88.83
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.88	89.70	87.99	87.15
	v)	江戸（会話文）+ 上方（会話文）	95.85	77.18	75.00	74.35
	vi)	追加学習：→江戸（会話文）	95.51	76.50	74.43	73.73
	vii)	追加学習：→上方（会話文）	95.91	74.99	72.39	71.73
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	95.81	74.76	73.02	72.42
江戸	i)	江戸（地の文）+ 上方（地の文）	**97.23**	**92.00**	**90.48**	**89.91**
	ii)	追加学習：→江戸（地の文）	97.10	91.94	90.42	89.75
	iii)	追加学習：→上方（地の文）	96.90	90.99	89.27	88.67
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.91	89.75	88.30	87.59
	v)	江戸（会話文）+ 上方（会話文）	95.48	77.17	74.95	74.38
	vi)	追加学習：→江戸（会話文）	95.28	76.79	74.60	73.96
	vii)	追加学習：→上方（会話文）	95.60	74.35	71.43	70.79
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	95.50	73.97	72.28	71.77
上方	i)	江戸（地の文）+ 上方（地の文）	97.14	91.83	89.60	88.76
	ii)	追加学習：→江戸（地の文）	96.58	90.62	88.10	87.26
	iii)	追加学習：→上方（地の文）	**97.35**	**92.09**	**90.00**	**89.06**
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.83	89.62	87.57	86.55
	v)	江戸（会話文）+ 上方（会話文）	96.37	77.20	75.06	74.31
	vi)	追加学習：→江戸（会話文）	95.83	76.10	74.19	73.40
	vii)	追加学習：→上方（会話文）	96.35	75.87	73.73	73.03
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.23	75.86	74.05	73.30
CHJ 近松	i)	江戸（地の文）+ 上方（地の文）	96.59	89.59	86.85	85.34
	ii)	追加学習：→江戸（地の文）	96.45	88.97	85.85	84.52
	iii)	追加学習：→上方（地の文）	96.03	89.12	86.09	84.58
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.60	**91.30**	**89.60**	**88.28**
	v)	江戸（会話文）+ 上方（会話文）	**97.25**	76.99	74.24	72.92
	vi)	追加学習：→江戸（会話文）	97.02	76.31	73.29	71.87
	vii)	追加学習：→上方（会話文）	96.60	76.86	74.03	72.52
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	97.21	80.28	78.39	77.16

認定できた短単位数を評価している．語彙素認定は，品詞認定をパスした短単位の内，語彙素読み，語彙素の両方が正しく認定できた短単位数を評価している．発音認定は，語彙素認定をパスした短単位の内，発音形出現形が正しく認定できた短単位数を評価している．評価には形態素解析器性能評価ツール『MevAL』[9]を使用した．

5.2　形態素解析性能評価実験の結果

辞書 i)〜viii) を使った各評価用コーパス地の文の解析結果の精度の比較を表 5，会話文の解析結果の精度の比較を表 6 にそれぞれ示す．表 5 を見ると，地の文の境界認定に関しては，地の文で学習した辞書と会話文で学習した辞書の間で精度に大きな差は見られない．一方で，表 6 を見ると，洒落本の会話文の境界認定では，地の文で評価した場合よりも両辞書間での精

(9) https://teru-oka-1933.github.io/meval/

表 6　各評価用コーパス「会話文」における各辞書の短単位自動解析結果の精度（F1 値）.

評価用コーパス（会話文）	辞書	学習用コーパス	境界認定	品詞認定	語彙素認定	発音認定
江戸 + 上方	i)	江戸（地の文）+ 上方（地の文）	92.65	70.90	69.35	68.78
	ii)	追加学習：→江戸（地の文）	91.72	69.65	68.05	67.45
	iii)	追加学習：→上方（地の文）	92.56	70.27	68.89	68.34
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	92.73	70.64	69.18	68.61
	v)	江戸（会話文）+ 上方（会話文）	**97.64**	**92.67**	**91.69**	**91.08**
	vi)	追加学習：→江戸（会話文）	97.33	91.93	90.93	90.30
	vii)	追加学習：→上方（会話文）	97.36	90.98	89.92	89.29
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.64	88.48	87.16	86.52
江戸	i)	江戸（地の文）+ 上方（地の文）	92.31	70.28	68.87	68.37
	ii)	追加学習：→江戸（地の文）	91.78	69.66	68.21	67.71
	iii)	追加学習：→上方（地の文）	92.15	69.37	68.10	67.60
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	92.53	70.03	68.76	68.31
	v)	江戸（会話文）+ 上方（会話文）	**97.75**	**92.95**	**92.25**	**91.73**
	vi)	追加学習：→江戸（会話文）	97.41	92.94	92.17	91.70
	vii)	追加学習：→上方（会話文）	97.34	89.95	89.07	88.53
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.51	87.63	86.57	86.09
上方	i)	江戸（地の文）+ 上方（地の文）	93.10	71.73	70.00	69.33
	ii)	追加学習：→江戸（地の文）	91.65	69.64	67.84	67.11
	iii)	追加学習：→上方（地の文）	93.10	71.48	69.96	69.32
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	92.99	71.47	69.74	69.01
	v)	江戸（会話文）+ 上方（会話文）	**97.48**	92.30	90.94	90.22
	vi)	追加学習：→江戸（会話文）	97.22	90.58	89.28	88.44
	vii)	追加学習：→上方（会話文）	97.39	**92.35**	**91.06**	**90.30**
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.82	89.61	87.95	87.08
CHJ 近松	i)	江戸（地の文）+ 上方（地の文）	95.12	76.06	72.51	71.39
	ii)	追加学習：→江戸（地の文）	94.42	75.08	71.63	70.67
	iii)	追加学習：→上方（地の文）	94.57	75.60	72.16	70.99
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	95.68	77.50	75.06	73.95
	v)	江戸（会話文）+ 上方（会話文）	95.13	87.56	85.07	83.32
	vi)	追加学習：→江戸（会話文）	95.05	86.95	84.36	82.61
	vii)	追加学習：→上方（会話文）	95.55	87.45	84.75	83.22
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	**95.81**	**89.98**	**88.82**	**88.86**

度差が大きいことが分かる．これは，地の文が文語体で比較的漢語比率が大きく，解析に曖昧性が低いのに対し，会話文は仮名文字率が高く，分割の曖昧性が高くなったためと考えられる．しかし CHJ 近松での境界認定の評価では，地の文で学習した辞書と会話文で学習した辞書の間で洒落本ほど大きな差は見られなかった．これは近松の文体が文語・口語の区別の薄い関係だと思われる．

しかしながら品詞認定になると，洒落本と CHJ 近松の両方において，真逆の文体で学習した辞書の精度が同一文体で学習した辞書よりも格段に悪くなった．これは CHJ のアノテーション方針で，地の文には文語の活用型，会話文には口語の活用型をそれぞれアノテーションするという仕様が原因である．地の文のみで学習した辞書であれば文語活用，会話文で学習した辞書であれば口語活用が優先して自動付与するようになるが，評価用辞書が逆転すれば，地の文に口語活用，会話文に文語活用を適用してしまう．地の文-会話文の区別が薄い CHJ 近松でも，

会話文を同定し，地の文と分けて活用型を付与しているため，境界認定とは異なって，洒落本と同様の結果となっている．

次に追加学習の効果を見ると，評価用コーパス:上方では，上方で追加学習した場合（辞書 iii), vii) ）に，会話文の境界認定を除いて，江戸 + 上方で学習した場合（辞書 i), v) ）よりも精度の向上が見られた．これに対し，評価用コーパス:江戸になると，江戸 + 上方で学習した場合（辞書 i), v) ）の方が，江戸を追加学習した場合（辞書 ii), vi) ）よりも精度が高くなった．これはそもそも，今回の学習用コーパスのサイズが江戸の方に偏っていたことに加え，江戸の洒落本においても上方を真似して発話している箇所がいくつも存在しているためだと考えられる．また江戸 + 上方で学習した辞書（辞書 i), v) ）と江戸を追加学習した辞書（辞書 ii), vi) ）を使って文字列「言ておく」を解析した結果，いずれも発音形は「イッ｜テ｜オク」と解析されたが，上方を追加学習した辞書（辞書 iii), vii) ）では発音形は「ユー｜テ｜オク」と解析できることを確認した．

最後に評価用コーパス:CHJ 近松の場合，上方からさらに CHJ 近松で追加学習した結果が最も高い精度となったが，地の文の境界認定の評価でのみ，会話文の江戸 + 上方で学習した場合（辞書 v) ）の精度が一番高くなった．これは前述の通り，近松では地の文-会話文の差が薄く，追加学習するよりも，量の多い会話文だけで学習したほうが優位となったためだと思われる．ただし，活用型の文語-口語アノテーションの差から，品詞認定以降の評価では，地の文で追加学習した結果の方が高い性能となった．

6. おわりに

本論文では，近世前期の上方資料である近松門左衛門の世話物浄瑠璃への短単位形態論情報アノテーションの効率化を目的に，洒落本解析用辞書から段階的な追加学習により，近松解析用辞書を構築する方法について述べた．CHJ 洒落本コーパスには，江戸と上方，2 つの地域の資料が混在しており，それらを同時に辞書の学習に使用することについての問題点を指摘し，江戸の資料よりもサイズの小さい上方資料では，実際に提案手法によって解析結果の精度向上を確認した．また上方の洒落本資料から，さらに少量の CHJ 近松を追加学習して作成した近松解析用辞書を使うことで，従来の洒落本用解析辞書よりも高い精度で CHJ 近松の解析が行えることが分かった．現在，この近松用短単位自動解析用辞書を使ったアノテーション作業が進行中である．

近松資料の特徴として，地の文と会話文の文体差が薄いことがあり，実験結果の境界認定の精度評価でもそれは確認できた．しかし CHJ 近松ではあえて，地の文と会話文を別に認定し，文語活用-口語活用の別で活用型をアノテーションしている．そのため境界認定においては，地の文で学習した辞書よりも高い精度を出した会話文用の辞書でも，品詞認定より先の評価では地の文解析用辞書より低い精度となった．このことから，近松の短単位解析用辞書の解析結果の精度をさらに向上させる方法として，地の文用の辞書を学習する際でも，会話文を，活用だけを「*」に変えて使用し，会話文用の辞書でもそれと逆の処理を実行する，という試みが今後の展開として考えられる．

謝　辞

本研究は，国立国語研究所共同研究「通時コーパスの構築と日本語史研究の新展開」の研究成果を報告したものである．

文　献

[Imamura 2013] Kenji Imamura (2013). “Case Study of Model Adaptation: Transfer Learning and Online Learning.” *Proceedings of IJCNLP-2013 (the 6th International Joint Conference on Natural Language Processing)*, pp. 1292–1298.

[Kudo et al. 2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto (2004). “Applying Conditional Random Fields to Japanese Morphological Analysis.” *Proceedings of EMNLP-2004 (the 2004 Conference on Empirical Methods in Natural Language Processing)*, pp. 230–237.

[Lafferty et al. 2001] John Lafferty, Andrew McCallum and Fernando Pereira (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pp. 282–289.

[市村太郎 2014] 市村太郎 (2014).「近世口語資料のコーパス化–狂言・洒落本のコーパス化の過程と課題–」日本語学 11 月臨時増刊号 日本語史研究と歴史コーパス 33:14, pp. 96–109.

[市村太郎ほか 2016] 市村太郎・小木曽智信 (2016).「文書構造を利用した近世期洒落本の形態素解析」言語処理学会 第 22 回年次大会 発表論文集，pp. 107–110.

[小木曽智信ほか 2013] 小木曽智信・小町守・松本裕治 (2013).「歴史的資料を対象とした形態素解析」自然言語処理，20:5, pp. 727–748.

[小木曽智信ほか 2014] 小木曽智信・中村壮範 (2014).「『現代日本語書き言葉均衡コーパス』形態論情報アノテーションシステムの設計・実装・運用」自然言語処理，21:2, pp. 301-332.

[岡照晃ほか 2013] 岡照晃・小町守・小木曽智信・松本裕治 (2013).「統計的機械学習を用いた歴史的資料への濁点付与の自動化」情報処理学会論文誌，54:4, pp. 1641–1654.

[岡照晃ほか 2014] 岡 照晃・松本裕治 (2014).「形態素解析との同時解析による歴史的資料の自動表記整理」情報処理学会第 216 回自然言語処理研究会 情報処理学会研究報告, 2014-NL216:8, pp. 1–20.

[鴻野知暁ほか 2014] 鴻野知暁・小木曽智信 (2014).「見出し語の時代情報を付与した電子化辞書の構築」言語処理学会 第 20 回 年次大会 発表論文集，pp. 209–212.

[近藤泰弘 2012] 近藤泰弘 (2012).「日本語通時コーパスの設計」NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集，pp.1–10.

[近藤泰弘 2015] 近藤泰弘 (2015).「『日本語歴史コーパス』と日本語史研究」コーパスと日本語史研究，ひつじ研究叢書＜言語編＞，第 127 巻， ひつじ書房，pp.1–16.

[洒落本大成編集委員会 1978-88] 洒落本大成編集委員会 編 (1978-1988).「洒落本大成」中央公論社.

[伝康晴ほか 2007] 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元 清貴・小磯 花絵 (2007).「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』，22 号，pp.101–123.

全文検索システム『ひまわり』における言語分析支援機能の拡張

山口昌也 (国立国語研究所音声言語領域) †

Enhancement for Supporting Language Analysis in Full-Text Search System "Himawari"

Masaya YAMAGUCHI (Spoken Language Division, NINJAL)

要旨

本稿では，筆者が開発している全文検索システム『ひまわり』の言語分析支援機能の拡張について述べる。元来，『ひまわり』は言語資料の検索と閲覧を目的に設計されたコンコーダンサであり，検索結果を分析するための機能を十分に備えていなかった。しかし，検索対象の資料の規模が大きくなると，大量の検索結果を単に表示するのではなく，集約して分析する必要性が生じる。また，検索結果の統計的な分析には，資料に含まれる文字数といった，基本的な情報を計測できなければならない。そこで，(1) 検索結果の集約機能，(2) 統計的分析のための基礎データの収集機能を『ひまわり』に実装した。拡張された機能を用いることにより，例えば『名大会話コーパス』の各会話中の発話数，文字数，単語数，特定の単語の出現数といった情報を収集できるようになる。

1　はじめに

全文検索システム『ひまわり』(山口昌也・田中牧郎, 2005)[1] は，言語研究用に設計された全文検索システムであり，XML でタグ付けされたテキストを全文検索することができる。もともと，2005 年に公開された『太陽コーパス』を検索するためのシステムとして開発された。本論文の主題である，分析支援機能としては，検索結果を閲覧するためのコンコーダンサとしての機能，および，収録された資料を指定した形式で表示する機能を持っている (図 1)。これらの機能では，利用者が検索結果や言語資料を「目で見る」ことの支援に焦点が当てられている。そのため，検索結果の集約や統計的分析については，R (統計分析用プログラミング言語)[2] や Microsoft Execl などの外部プログラムに検索結果をエクスポートして処理するという前提である。

当初の設計から 10 年以上を経て，『ひまわり』にはさまざまな改良が加えられているが，分析を支援する機能については，基本的に当初のままである。その一方で，言語資料は大規模化し，個人のレベルでも入手できるようになった。『ひまわり』でも，青空文庫，国会会議録，名大会話コーパス，Wikipedia などを手軽に検索できるようになっている [3]。

言語資料が大規模化すると，検索結果は増大するため，目で見きれなかったり，そもそも，メモリ不足などにより，検索結果を表示しきれないということも起こりうる。したがって，当初の設計のように検索結果を単純に表示するだけでなく，集約して表示する必要が生じる。また，統計的な分析を行うためには，各種の統計量を計算するための基礎的なデータ (例：総文字数，総発話数) を利用者が自由に計測できなければならないが，現状の『ひまわり』では計測自体ができなかったり，できたとしても長時間の処理が必要になっていた。

そこで，本稿では，『ひまわり』の分析支援機能の拡張として，検索結果の集約方法，および，統計的な分析の支援方法の設計を行い，実装した結果を示す。

---

†http://www2.ninjal.ac.jp/masaya

[1] http://www2.ninjal.ac.jp/lrc/ の『ひまわり』ホームページから無料でダウンロードできる。

[2] https://www.r-project.org/

[3] 『ひまわり』のホームページで配布している。簡単にインストールできるよう，パッケージ化されている。

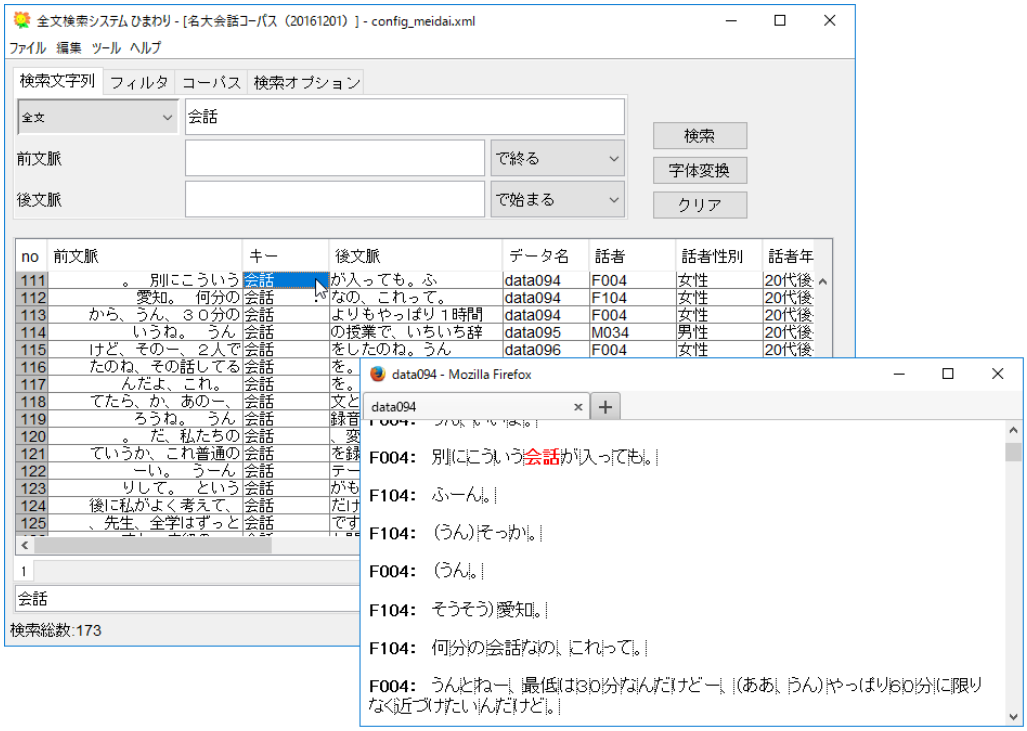


図 1: 『ひまわり』の実行例

2 既存システムと本システムの位置づけ

これまでに，言語資料の利用を支援するためのさまざまなシステムが開発されている。分析の支援という観点から見ると，支援のための機能は，大きく分けて，次の三つに分類されると考えられる。なお，例として挙げたシステムは，当該分類だけに含まれるわけではなく，複数の分類に入りうる。

(1) コンコーダンサとしての機能 (「中納言」[4](小木曽智信ほか, 2011)，「梵天」[5] など)
(2) 検索語のコロケーションを表示するなど，検索結果を集約して表示する機能 (AntConc(Anthony, 2016), NINJAL-LWP(今井新悟ほか, 2013) など)
(3) 検索結果に対する統計的分析ツールとしての機能 (KHCoder[6] (樋口耕一, 2014)) など

現状の『ひまわり』の分析支援機能は，(1) である。また，前節で述べたとおり，(2)(3) は現状の問題を解決するための解決方法となる。したがって，支援機能拡張の方向性として，『ひまわり』の特徴を活かしつつ，(2)(3) を実現することが好ましい。

他のシステムに対する『ひまわり』の特徴は，XML によりアノテーションされた多様な形式の言語資料を検索し，アノテーションされた情報を検索結果として抽出できることである。また，『ひまわり』改良の過程では，言語資料の作成を支援する機能として，テキストのインポート機能が実装されている (山口昌也, 2013)。この機能を用いると，テキストに付与された独自形式のタグや，テキスト表記上の規則 (例：「太郎：」のように，行頭の：で区切られた文字列は，後続する文字列の発話者を表す) を XML に基づいたアノテーションをインポート時に行うことができる。

以上の特徴をまとめると，言語資料の作成者が必要であると考えてアノテーションした結果を分析に活かせるということである。したがって，新しい『ひまわり』を，アノテーション結果の「分析」

[4]https://chunagon.ninjal.ac.jp/
[5]http://pj.ninjal.ac.jp/corpus_center/nwjc/bonten-overview.html
[6]http://khc.sourceforge.net/

支援機能を持ったコンコーダンサとして位置づける。アノテーションの分析機能を，上記の (2)(3) と結びつけていく方法については，次節で述べる。

3　基本的な設計方針

ここでは，アノテーション結果の分析支援という面から，検索結果の集約機能と，統計的分析機能についての設計方針を示す。

まず，『ひまわり』の検索結果を集約することは，検索結果が検索文字列とそれに付随するアノテーションから構成されるため，アノテーション結果を集約することに他ならない。現状の『ひまわり』でも，検索結果から選択した任意の列から，出現頻度付きの一覧表を作成することによって，検索結果を集約することができる。例えば，図 1 の「話者性別」列を選択して，一覧作成機能を実行すれば，検索文字列に対して，性別ごとの出現頻度を求めることができる。ただし，検索結果を求めた後でないと，一覧が作成できないという問題がある。そのため，前節で述べたように，検索結果が大量だとメモリ不足などの問題が発生する。また，後述するように，列選択による単純な一覧作成では，求める結果を得られない場合もある。そこで，本稿では検索結果の集約方法について，次の機能拡張を行う。

- 検索時の一覧作成機能 (4.3 節)
- 一覧作成機能の改善 (4.4 節)

次に，統計的分析については，分析機能自体は『ひまわり』に持たせないかわりに，統計的分析のための基礎データをアノテーション結果から得られるようにする。この理由の一つは，すでに統計的分析用の有用なツールが存在することである。もう一つの理由は，分析ツールに入力するデータ自体 (ここではアノテーション結果) を利用者が確認した上で，分析ツールを利用することが重要であると考えるからである。本稿で扱う拡張は，次の 2 点である。これらは，アノテーション結果から統計的分析用の基礎データを収集する際の問題を解決する。

- アノテーション結果の集計機能 (4.1 節)
- 外部アノテーション機能の改善 (4.2 節)

前者は，現状の『ひまわり』では，検索文字列を指定して検索しないと，アノテーション結果を抽出できない，という問題を解決する。後者は，形態素解析結果など，大量のアノテーションを行う場合に用いられる「外部アノテーション」機能の改善である。この機能はデータサイズ増大に対応するための機能であるが，現状では十分な性能が得られていない。

4　拡張機能の実現

4.1　アノテーション結果の集計機能

前述のように，『ひまわり』用の言語資料にはさまざまな研究用の情報がアノテーションされている。ここでは，付与されているアノテーションを集計する方法について考えてみる。

例えば，『ひまわり』用の「名大会話コーパス」パッケージの場合 (会話データ data016 の冒頭部分を引用)，次のように XML で記述されている [7]。

この例には，「始めまーす」「はーい」という二つの発話が含まれている。冒頭の meidai タグは，一つの会話全体に対してマークアップするタグである。u, s タグは，それぞれ発話，単語 (短単位) を表す [8]。それぞれのタグは，属性を持つことが可能である。例えば，(3 行目の)「始め」をマークアップしている s タグは，l (基本形)，p (品詞)，f (活用形)，c (活用型) の属性を持っている。

[7] 紙面の都合上，一部の属性・タグを省略している。また，見やすいように，適宜改行を入れている。
[8] 名大会話コーパスの規模は 142 万語程度なので，外部アノテーションではなく，XML でアノテーションしている。

```
<meidai name="data016" speakers="F004,F028" duration="56 分" ns="14023">
<u s="F004" i="0" sex="女性" age="20 代後半">
<s l="始める" p="動詞-非自立可能" f="連用形-一般" c="下一段-マ行">始め</s>
<s l="ます" p="助動詞" f="終止形-一般" c="助動詞-マス">まーす</s>
<s l="。" p="補助記号-句点" f="" c="">。</s>
</u>
<u s="F028" i="0" sex="女性" age="20 代後半">
<s l="はい" e="ハイ" p="感動詞-一般" f="" c="">はーい</s>
</u>
    :
</meidai>
```

名大会話コーパスには複数の会話データが含まれているが，会話データ数を計測するには meidai タグを，単語数を計測するには s タグを列挙すればよい。しかし，現状の『ひまわり』では，仕様上，言語資料の作成者が列挙内容を事前に定義しておくか，すべてのタグの情報が表示されるような検索を行う [9] 方法しか用意されておらず，いずれも，実現可能性，計算時間の点で現実的ではない。

そこで，一覧に表示するタグとその属性を利用者が対話的に指定できるようにした。図 2 は，u タグの s 属性を指定し，話者の一覧を表示した例である。一覧を作成する対象となるタグの指定を左図のダイアログで行う。ここでは，「第一階層タグ」で u タグを選択し [10]，その右のボタンを押すことにより，属性選択ダイアログ (図中央) が表示されるようになっている。この例では，タグ指定のダイアログで，「頻度」オプションをチェックしているので，単に話者の一覧を表示するだけでなく，話者 (u/s 列) ごとに発話数 (頻度列) が表示される。また，一覧の下部に表示される総数と異なりによって，総発話数，および，総発話者数を求めることができる。

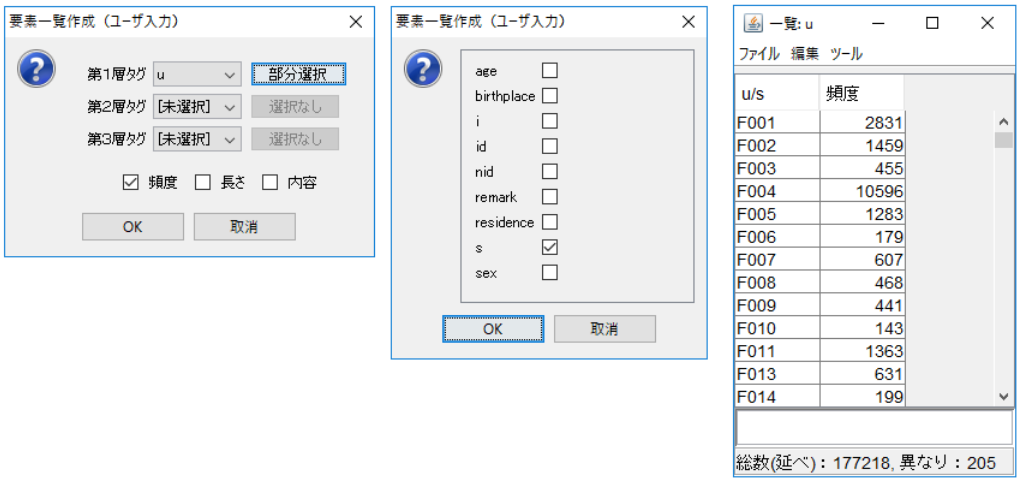


図 2: アノテーションの集計例

一覧に表示されるタグは，３階層まで指定可能である。一覧としては，最下層のタグの一覧が作成されるようになっているが，作成の過程で上位階層の属性を一覧に取り込むことができる。これらの情報は，一覧作成対象のタグに制約を加えたり，付随的な情報を集計するために用いられる。これにより，次のような一覧を作成することが可能である。

[9] 例えば，すべての単語を列挙するには，正規表現で任意の文字列 (^.+$) にマッチするような条件を指定する。
[10] 指定できるタグは，選択リストに表示される。

- 会話データごとに話者の発言数を求める
  ［第 1 階層タグ meidai (name 属性)，第 2 階層タグ u (s 属性)，「頻度」オプション］
- 会話データごとに総単語数を求める
  ［第 1 階層タグ meidai (name 属性)，第 2 階層タグ s (属性は無指定)，「頻度」オプション］

一覧表示の際のオプションには，「頻度」以外にも「長さ」「内容」がある。「長さ」は，最下層のタグでマークアップされているテキスト長 [11] を求める。例えば，これにより，会話データ (meidai タグ)，発話 (u タグ) に含まれる文字数の一覧を作成することができる。「内容」はタグでマークアップされているテキストを表示する。このオプションは，発話内容をすべて列挙する場合などに利用できるだろう。

4.2 外部アノテーション機能の改善

4.2.1 検索性能の改善

『ひまわり』では，テキストに対するアノテーションを記述する方法として，2 種類の方法が用意されている。一つは，4.1 節で示したように，XML で記述する方法である。もう一つは，外部のリレーショナル・データベースに記述する方法である。後者の場合，アノテーションする情報は，テキストの位置情報と関連付けられて，リレーショナル・データベースに格納される。

後者の方法を導入した背景には，コーパスファイルの巨大化への対応がある。特に，形態素解析システムによって得られた結果は，統計的な分析を行う上で基本的な情報であるが，書誌情報や発話情報のアノテーションなどと比較して，アノテーションの量が多くなる。また，『ひまわり』で扱える形式の XML 文書では，すべての形態素に対して，付随するすべての情報を記述しなければならず，冗長な記述とならざるを得ない。そのため，XML で直接記述すると，『ひまわり』で扱えるコーパスファイルの上限を越えてしまう場合が出てくる。

以上の背景のもと導入した方法だが，(検索はできるものの) データベース用のファイルサイズが巨大になるという問題があった。例えば，『ひまわり』用に公開している『青空文庫』パッケージの場合，データベースのサイズが約 6.2GB にもなる。

そこで，外部アノテーション検索用に用いていたリレーショナル・データベースを独自のデータベースで再実装した。検索処理は，(1) マークアップしている範囲の検索，(2) マークアップされているタグの属性検索に分けられる。(1) には山口昌也・田中牧郎 (2005) と同様に 2 分検索を使用し，(2) はアノテーション集合 (形態素解析結果の場合，辞書に相当) をメモリ上で線形検索する。(2) のタグの属性検索は，正規表現にも対応する。

表 1 に，新旧データベースのファイルサイズを示す。青空文庫は『青空文庫』パッケージ (20160401 版)，国会会議録は『国会会議録』パッケージ (20140327_rev20170201 版) を用いた。また，検索速度の参考値として，検索キーとして「あの」(基本形) を検索した時の実測値 (CPU: Xeon E5-1620 3.7GHz 4core, OS: Ubuntu 16.04, Memory: 24GB) を示す。検索総数は青空文庫で 78561 件，国会会議録で 41949 件である。使用した『ひまわり』は ver.1.5.5(旧)，ver.1.6.a20170120(新) である。

この結果のとおり，言語資料のサイズは青空文庫で約 26%に削減された。検索速度も約 2.5 倍に向上している。新しいデータベースはサブコーパスごとに分割することも可能になっているため，配布の際の問題 [12] も軽減されると考えられる。

4.2.2 外部アノテーション内容の閲覧機能

外部アノテーションの内容は，XML 文書中に直接記述されないので，実際にどのようなアノテーションがなされているのかを確認しづらい。特に，形態素解析システムなどのツールによるアノテー

[11] マークアップされているタグ，改行文字を除外した上で計測される。
[12] 一般公開する場合は，ファイルを圧縮しているが，巨大すぎて使用環境によっては展開できないなどの問題が発生する。

表 1: 言語資料のサイズと検索速度 (参考値)

言語資料	旧 (サイズ)	新 (サイズ)	旧 (検索)	新 (検索)	総語数
青空文庫	6.2GB	1.6GB	8.9sec	3.6sec	1.0 億語
国会会議録	—	4.1GB	—	2.8sec	2.7 億語

ション結果は，誤解析も含まれるため，一定の範囲で外部アノテーション結果を確認する手段を用意すべきである。

従来の『ひまわり』では，検索結果をダブルクリックすると，当該の会話データや作品全体を Web ブラウザで表示できるようになっている。今回の拡張では，外部アノテーション結果でも会話や作品全体を一覧できるようにした。

図 3 は，『ひまわり』に付属する青空文庫サンプルから芥川龍之介の「蜘蛛の糸」全体の形態素解析結果を表示した結果である。この一覧は，検索文字列「釈迦」を検索し，その検索結果の一つをダブルクリックすることにより表示される。この図のとおり，選択した検索結果 (「釈迦」) の出現位置にジャンプする。前後の形態素は上下に位置し，“_TEXT” 列がテキストでの出現形である。この一覧を使えば，作品ごとに語彙や品詞の分布を容易に求めることができる。また，『ひまわり』のソート機能を使って，ランダムにソートすれば，ランダムサンプリングが実施できる。

図 3: 外部アノテーション結果の表示例 (芥川龍之介：「蜘蛛の糸」)

4.3 検索時の一覧作成機能

検索時の一覧作成機能は，利用者が実行した検索の結果をそのまますべて表示するのではなく，集計した結果を表示するものである。この機能は，メモリ不足の問題が発生するような，大量の検索結果が得られる場合に用いることを想定している。従来版の『ひまわり』でも (検索結果をすべて表示

するのではなく) 検索件数のみを表示することはできたが，この機能では，アノテーションの集計と同様に各種の付与属性を含めて一覧表示できるようにした。

図 4 右は，出現形「が」の検索結果を集計機能を使って一覧表示した結果である。左図は，検索条件と一覧表示のための設定である。一覧表示したい属性は，左図のようにマウスで選択しておく。

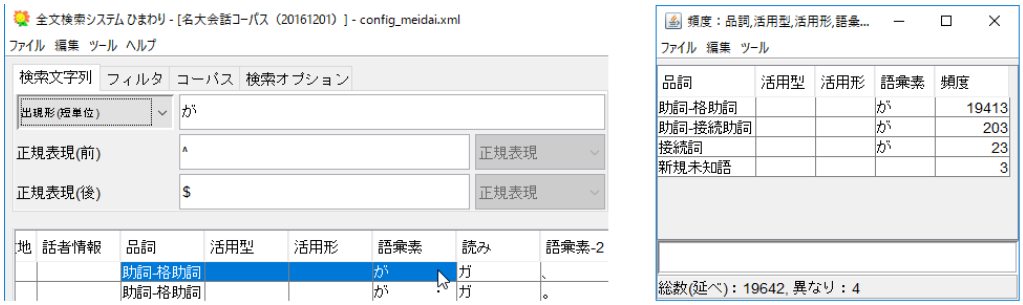


図 4: 検索時の一覧作成例

4.4 一覧作成機能の改善

4.4.1 結果の整形

検索結果の中には，そのままでは，一覧作成の際にうまく利用できないデータが含まれる場合がある。例えば，『ひまわり』用の『国会会議録』パッケージを使って，特定の検索語の経年変化を調べることを考える。このパッケージでは，検索結果として，検索文字列を含む会議の開催日が得られる。この結果に対して，年ごとに集計を行う場合，得られるのは開催日なので，年を抽出する必要がある。

このような問題を解決するために，検索結果に対する置換機能を追加した。置換は，図 5 のように，列ごとに行う。置換の条件式には，正規表現を使うことができる。図 5 では，月日の部分 (例：-03-30) は不要なので，正規表現-.*にマッチする文字列を空文字列で置換するように条件を指定する。この置換結果に対して，次節で述べる結果の再集計を行うと，年ごとの集計ができる。

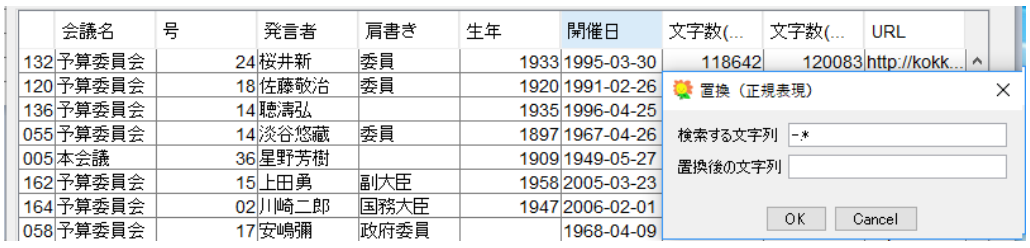


図 5: 置換機能の実行例

4.4.2 再集計

前節の置換結果のように，検索結果を再度集計しなければならない場合を考慮して，集計結果を再度集計する機能を実現する。再集計するには，検索時の一覧作成機能と同様に，再集計したい列を選択して，一覧を作成する。

図 6 左は，図 5 の「開催日」と「発言者」列を選択して，一覧を作成した結果である。このように，発言者ごとに検索文字列の出現頻度を経年変化を調べることができる。さらに，この結果から検索文字列の出現頻度の経年変化を求めるには，左図の「開催日」列を選択して，一覧を作成する (図 6 中)。この場合，左図の「頻度」の値を年ごとに合計する。それに対して，各年ごとの発言者の異な

り数を求めたい場合は，「頻度」を合計せずに「開催日」の頻度を求めればよい (右図)。元の一覧の頻度欄を考慮する・しないは，再集計するときに選択できる。

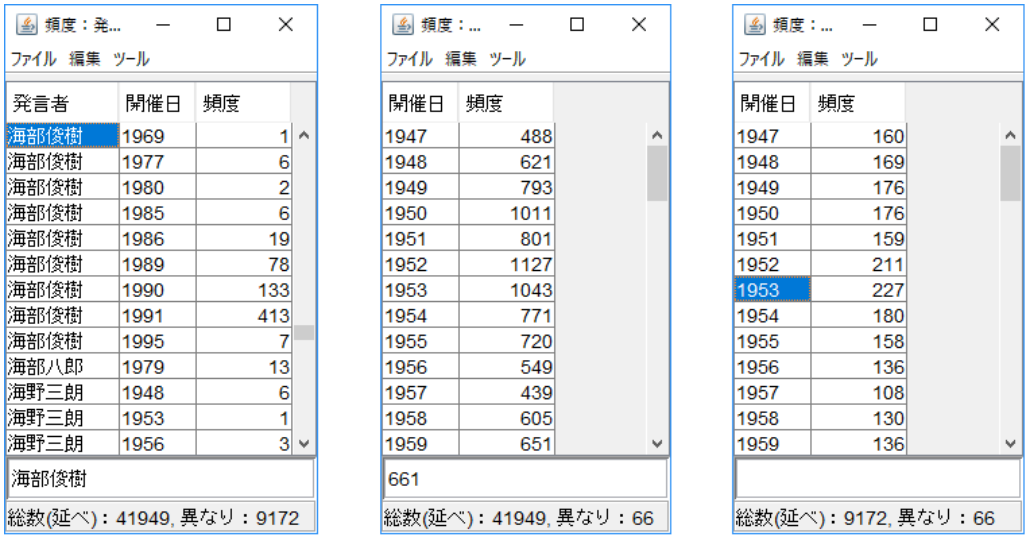


図 6: 再集計の実行例 (左：発言者ごとの経年変化，中：検索文字列，右：発言者数の異なり)

5 おわりに

本稿では，全文検索システム『ひまわり』の言語分析支援機能の拡張として，(1) 検索結果の集約を行う機能，(2) アノテーション結果から統計的分析の基礎データを収集するための機能を設計・実装した。

謝 辞

本研究は，国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」，および，科研費基盤研究 (B)『「昭和話し言葉コーパス」の構築による話し言葉の経年変化に関する実証的研究』の一環で行われたものである。

文 献

山口昌也・田中牧郎 (2005).「構造化された言語資料に対する全文検索システムの設計と実現」 自然言語処理, 12:4, pp. 55–77.

小木曽智信・中村壮範・鈴木泰山・八木豊・山崎誠・前川喜久雄 (2011).「コーパス検索システム「中納言」デモンストレーション」 日本語コーパス完成記念講演会予稿集.

Laurence Anthony (2016). *AntConc ver.3.4.4*. http://www.laurenceanthony.net/.

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.

樋口耕一 (2014).『社会調査のための計量テキスト分析――内容分析の継承と発展を目指して』 ナカニシヤ出版.

山口昌也 (2013).「個人用コーパスの作成とアノテーションを支援する環境の実現」 第３回コーパス日本語学ワークショップ予稿集, pp. 369–372.

児童生徒の「手」作文に於ける経年変化の計量的分析 —1992 年と 2016 年の作文を比較して—

阿部 藤子（東京家政大学 家政学部）
今田 水穂（文部科学省 初等中等教育局）
宗我部 義則（お茶の水女子大学附属中学校）
冨士原 紀絵（お茶の水女子大学 基幹研究院人間科学系）
松崎 史周（日本女子体育大学 体育学部）
宮城 信（富山大学 人間発達科学部）†

A Quantitative Analysis of Generation Change with Composition of 'Hands' of A Written Composition Corpus of Japanese Elementary and Junior High School Students − Comparison of 1992 and 2016 data −

Fujiko Abe （Tokyo Kasei University）
Mizuho Imada （Ministry of Education, Culture, Sports, Science and Technology）
Yoshinori Sogabe （Ochanomizu University Junior High School）
Kie Fujiwara（Ochanomizu Universtiy）
Fumichika Matsuzaki （Japan Women's College of Physical Education）
Shin Miyagi （University of Toyama）

要旨

本発表は児童生徒らの文章作成能力の経年変化を計量的分析によって明らかにすることを目的とする。その基礎資料として作文を電子化した「「手」作文コーパス」を構築した。本コーパスの資料は 1992 年及び 2016 年に児童生徒らが書いた「手」を題とする作文である（両資料は，同一の国公立大附属小中学校で同条件で作成されたものである）。両資料の調査時期にはおよそ四半世紀（24 年）の隔たりがあり，本発表の目的はその間の児童生徒らの文章作成能力の変化の有無を明らかにすることにある。予備調査を行った結果，1 サンプル当たりの文章量（総字数），語数，文節数等で両資料間に明確な差異を見いだすことはできず，文章の量的観点からは大きな経年変化は見られないことが分かった。一方で，現場の教師らから「以前に比べて子ども達が作文が書けなくなった」という指摘を聞くこともあり，使用語彙の種類や品詞の偏り，文末形式等の文体的特徴の違いを数量的差異として抽出し，2 つの資料の異動を観察する。その結果に基づき先の教師らの指摘の妥当性を検討する。

１．はじめに

近年，コーパスを利用した言語研究が盛んになってきている。国語教育学研究でも子ども達の書いた作文を資料とした文章作成能力の実態調査や指導法の開発等が行われるようになった。作文コーパスを使った研究によって，それまでベテラン教師らの経験知によってし

† miyagi@edu.u-toyama.ac.jp

か指摘できなかった児童らの書く作文の特徴が少しずつ明らかにされつつある。しかしながら，その資料となる児童・生徒の作文でコーパスとして利用可能なものは，資料の収集や公開の難しさから質量共に不足しており，十分な研究環境が整っているとは言いがたい。その課題に取り組むために，著者らは本研究に於いて小中学校の児童生徒の作文を 3 年間に亘って収集した大規模な作文コーパス『児童・生徒作文コーパス』（以下，「児童作文コーパス」と略す）の構築を進めている[1]。現在、このコーパス用いて、多角的な観点から児童生徒らの文章表現能力とその発達過程の解明を目的とした研究を進めている（関連する研究に関しては参考文献を参照のこと）。

研究を進める過程で「児童作文コーパス」で取り上げられた題「夢」や「頑張ったこと」以外の文章ではどうか，附属校と一般校とでの違いはどうかといった関連する研究課題が生まれてきた。それらの課題を解決するため、本研究全体で子ども達の文章作成能力を計り取るために様々な条件での作文調査を計画している。この計画は、本発表で報告する「「手」作文コーパス」や、同一の書き手による文体の書き分け調査を企図した中学生による文種別作文コーパス，小学生の話し言葉を記録した話し言葉コーパスなどを含む（図 1）。

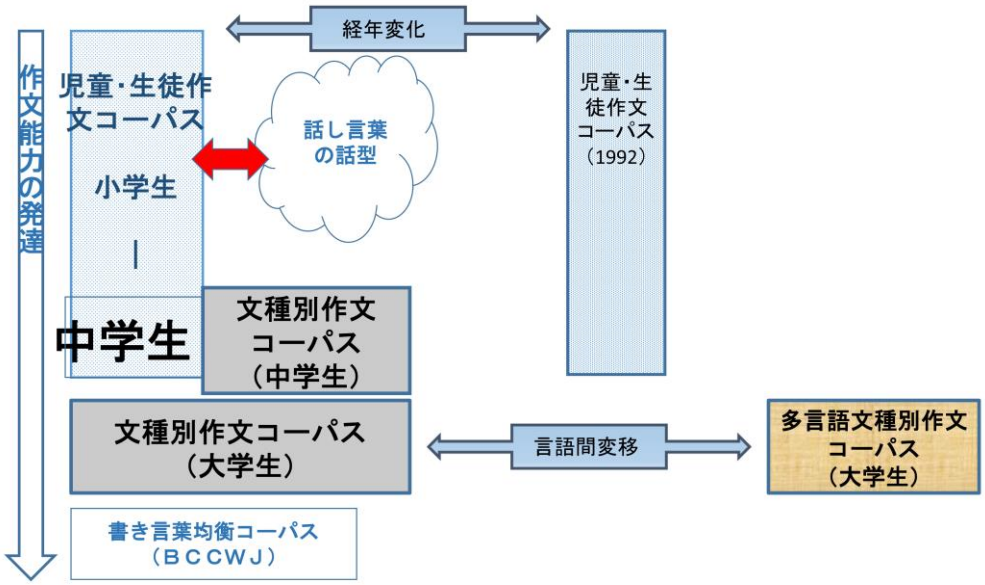


図 1　「児童・生徒作文コーパス」を中心とした関連作文コーパスの概要

本発表で報告する「「手」作文コーパス」は，児童生徒らの作文作成能力の経年変化を調査することを目的としたものである。作文コーパスの利点の一つは，相当量の資料を収集することによって，個人の書きぶりからは見いだせない学齢層に準じた文章作成能力の特徴を抽出することができることである。しかし，児童の作文能力の発達について，より正確な結果を得るためには，ある個人に着目して数年に亘って作文資料を収集し続けることが望ましい。これは現実的には難しく，またその個人が果たして代表的な作文能力を有しているのかという問題もある。それに代わる手段として，同じ条件（指示，作文の題，環境等）を整えて同レベルの小中の協力校で悉皆調査を行う方法が考えられる。この方法であれば，質

[1] 『児童・生徒作文コーパス』の規模・仕様等の詳細については，宮城・今田 2015a を参照されたい。

の高い一定程度量のデータで、かつ児童らの発達過程をかなりの精度で反映した資料を得ることができる。作文コーパスの利点としてもう一つ見逃してはならないものが，作文コーパスは調査時の児童らの実態を正確に保存しているということである（多くの場合「資料が古く現状と異なるのではないか」のようにそれが欠点として指摘されてきた）。言い換えれば，適正な資料を揃えることができれば現在なら現在の 10 年前なら 10 年前の子ども達の実態をコーパスを通じて推し量ることができると言うことになる。そこで本発表では，作成年度に隔たりのある 2 つの作文資料（1992 年と 2016 年の調査資料）を電子化した対照コーパス「「手」作文コーパス」を構築する。また，それを活用することによって，両年代間の児童生徒らの作文能力について複数の観点から調査を行い，その差異についての検討を試みる。

本発表では，この「「手」作文コーパス」がどのようなものであるのかを紹介し，実際にいくつかの文法事項について経年変化を調査してみることによって，研究資料としての妥当性を検証する。本コーパスのデータの規模は，1992 年と 2016 年の調査を合わせて 28 万形態素規模である。「児童作文コーパス」が 300 万形態素規模に達する予定なので，その 10 分の 1 ほどの規模である。調査内容によってはデータの不足が感じられるかも知れない。また，過去の資料を利用したコーパスであるため，別途意向があっても調査方法を先行する 1992 年の調査に準じざるを得ない，その性質上調査を続けて今後過去のデータを補強していくことが難しい，作文時の状況や内容について書き手の児童らにフォローアップインタビューをすることが適わないといった制限がある。様々な制約がある一方で，他に類を見ない 24 年を隔てた経年変化を反映した資料であり，今後様々な研究での活用が期待できる。

２．作文資料に基づく文章作成能力に関する経年調査の可能性

現場のベテラン中堅小中校教師との作文指導に関する議論の中で，しばしば話題にされるのが「最近の子ども達は作文が書けなくなった」という指摘である。近年の様々な電子デバイスの発達の裏側で，漢字を始めとして書く機会そのものが激減し，その結果子ども達の「書く能力」が低下していることは十分に考えられることである。一方で「書けなくなった」が具体的にどのような点に言及したものであるのかは明確ではなく，教師らに反問してみても印象以上の答えは得られなかった。子ども達の文章作成能力の経年変化を客観的に計り取るのに最も適しているのが大規模作文コーパス（できるだけ大人の手が入っていないもの）を活用した調査であろう。しかしながら，前頁の図 1 に示された中で入手が困難なものの一つがその経年変化を反映した資料であり，問題となる。もちろん過去に遡って調査することは不可能であるし，たとえ今から調査を開始してもコーパスが完成するのが 10 年以上先になってしまうからである。

調査方法に苦慮していたところ，幸運にも 1992 年に実施された小中同一題作文調査の資料を貸借することができた。さらに幸運なことに 1992 年調査の調査校が現在作文調査を進めている調査協力校の一つであったため，同校に依頼して 1992 年と同一の条件で作文調査を実施し，児童生徒の文章作成能力の経年変化を調査することを目的とした対照作文コーパスを構築することが可能となった。このコーパスも児童生徒の作文を収集したものであるが，大本の「児童・生徒作文コーパス」との混同を避けるために，調査した作文の題をとって総称で「「手」作文コーパス」と名付け，格納される年度別資料を区別する場合，調査年度に即してそれぞれ「「手」コーパス 1992」「「手」コーパス 2016」と呼び分けることにする。

３.「手」作文コーパスの設計と基本方針

３．１　「手」作文コーパスの特徴

「手」作文コーパスは，1992 年と 2016 年に小学生・中学生を対象として同一の条件で調査・収集した作文を電子化し，言語学的情報を付与した経年調査を目的としたコーパスである。以下に調査および電子化の概要を説明する。

1992 年の調査は，当時の資料によると，国立大学附属小中学校に於いて小学 1 年～中学 3 年で各学年 1 クラスずつ（1 クラス男女 20 名ずつ，計 40 サンプル）を対象に「手」という題で作文させた資料を基に構築されている。調査時の指示は「これから『手』という題で作文を書きます。どんなことを書いても自由です。原稿用紙 1 枚（400 字）で書きます」というもので，時間制限は設けないという条件で実施された。

条件を同一にそろえるため，2016 年の調査も上記の条件（指示，作文の題，字数制限）に揃えて実施した。調査協力校は国立大学附属学校 2 校（小学校 1 校，中学校 1 校）で，1992 年の調査を実施した学校と同一校である[2]。当該校で 9 学年（小学 1 年～中学 3 年）の全児童生徒に「手」（または「て」）という題で作文課題を課し，収集して電子化した。作成時間は小学校 40 分，中学校 45 分とした[3]。1992 年調査では，児童ら全員に作文させた後に学年毎に男女それぞれ 20 名ずつをランダムに抽出したようだが，2016 年の調査では調査協力校で悉皆調査を実施した後，その全てを資料として収集した（そのため総形態素数に約 4.5 倍の隔たりがある）。収集した作文は 1992 年調査が 8 クラス，2016 年調査が 32 クラスである（計 40 クラス）。各作文原本から整理番号を付し，学年，クラス，性別などの属性を区別できるようにした後，氏名を削除（用紙から切り落とした）して電子化の資料とした。なお，調査時には教師は一切の事前指導を行わず，課題作成時にも題のみを提示して内容に一切干渉していない，また質問が出ても原則返答していない[4]。

収集した作文の電子化作業は，原則として「児童作文コーパス」と同様の指針で実施した（図 1 の関連のコーパスも基本的にこの方針に従って電子化作業を進めている）。ただし，「児童作文コーパス」が判読不能な場合”＊”で置き換えを行ったのに対して，「「手」コーパス 2016」では，1992 年当時の調査方針に従って，判読不可能でも解釈可能な場合，一部推測して文章を補って記録した。以下に電子化の指針を示す。下線が「児童作文コーパス」の指針との相違である。

〇電子化の指針

・できるだけ正確に紙面を再現するよう心がける。
・段落初めの一字下げや空欄（意味不明なものも含めて）も正確に記録する。
・文字種の違いにも注意して正確に記録する。
・誤字・脱字はそのまま記録する。

[2] 本研究の関係者に 1992 年の調査・分析での参加者がいたため，かなりの程度 24 年前の調査状況を再現することができた。

[3] 1992 年の調査では時間制限を設けていないが，2016 年の調査は国語の時間を使って行われたため，作成時間の条件を変更せざる得なかった。

[4] 小学校低学年では教師が黒板に「て」と題を板書して作文させたため，「てつぼう，てさげかばん，てれび，てんとうむし，･･･」のように”「て」で始まる言葉”を列挙する児童もいたが，教師は特段の指導を行っていない。また少数であるが「手」と関わりのない内容で作文した児童もいた。

・判読不能な箇所で解釈可能な場合は推測した。（※1992 年調査の電子化方針に従った。）
・入力後に入力者以外の者が原本と照合して入力ミスを修正する。
・個人情報にかかわる部分（個人が特定される可能性のある語句や学校名，氏名・渾名等）は，当該部分を“＊”で置き換える。
・1 作文 1 ファイルで記録して整理番号を付す。（整理番号から課題・学年・クラス・性別等が判別できるようにする。）

個人情報保護の理由から，収集した作文原本は非公開とする。電子化したテキストデータは範囲を限定して利用を認める場合もある。「手」作文 2016 の本文は児童・生徒の個人情報に関する処理を施した後，学術的研究，特に学校現場への還元を目的とした研究に利用する場合での一般公開が可能になるよう現在協力校に交渉中である。「手」作文 1992 の本文は原則非公開である（文章の基本的情報や使用語彙等のデータは今後公開していく予定である）。

３．２　「手」作文コーパスの構成

本コーパスは本文テキスト，メタデータ，形態論・構文情報データで構成される。本文テキストは作文を電子化したテキストファイルである。メタデータは本文テキストには含まず，整理番号（サンプル ID）と紐付けて別に管理する。メタデータは以下の項目を含む。

作文課題の属性	サンプル ID，調査年度　（1992 | 2016）
執筆者の属性	著者 ID，学年，クラス，性別

形態論・構文情報データは，本文テキストに対して自動処理により文境界，文節境界，形態論情報，構文情報（係り受け情報）などの言語学的情報を付与したデータである。Ruby スクリプトによる文分割処理を施した後，CaboCha 0.69[5]と UniDic 2.1.2[6]で自動解析し，形態論情報と構文情報（係り受け情報）を付与した。両資料 40 ラス分の文数，形態素数，文字数（改行文字を除く）を集計したものを以下に示す。

表 1 「「手」コーパス」のデータサイズ

調査年度	サンプル	段落	文	文節	形態素	文字
1992	280	1092	2980	19159	51064	79752
2016	979	3963	12125	87079	228789	355068
総計	1259	5055	15105	106238	279853	434820

40 クラス分のコーパスの語数が約 28 万形態素で，「児童作文コーパス」が最終的に約 300 万形態素程度の規模となる予定であるので，その 10 分の 1 程度の規模のコーパスになる見込みである。なお今回貸与された 1992 年の資料では小学 5 年生の作文がすべて欠落していた。データの欠落自体は問題であるが，範囲が確定していること，学年別の違いよりも経年変化の調査に特化したコーパスであることなどから，注意して使用すれば致命的な欠陥と

[5] https://code.google.com/p/cabocha/
[6] http://sourceforge.jp/projects/unidic/

はならないと考える[7]。

３．３　作文コーパスの基本情報と基礎的な考察

自動付与した形態論・構文情報を用いて，「手」作文 40 クラス分のデータの各種情報を学年別に集計した結果を図 2 に示す。1992 年と 2016 年の調査では，調査に参加した児童生徒数が大きく異なるため，段落，文，文節，形態素，文字数をそれぞれ 1 サンプル当たりに換算して整理してある。

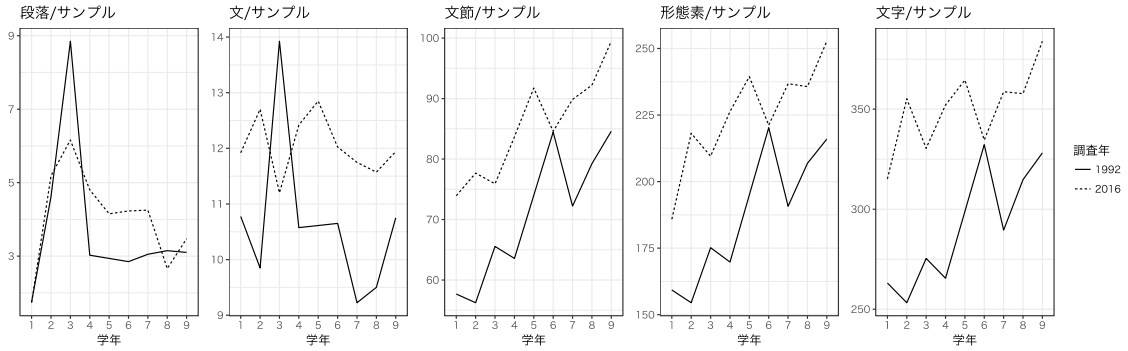


図 2　「「手」コーパス」の基本情報（学年別・1 サンプルあたり）

この図 2 に基づいて「手」コーパス 1992 と「手」コーパス 2016 のデータを比較すると，段落のように低学年から中学年にかけて大きく数値が増加した後ほぼ動きが見られないもの，文数のように揺れはあるが学年による差異がさほど顕著ではなくほぼ増減が見られないもの，文節，形態素，文字のように学年の進行によって数値が大きく増加していくものがあることが分かる。また，どの言語単位についても，ほとんどの学年で 1992 年より 2016 年の資料の方が大きな値を取っていることが分かる。

このことから，児童生徒らの 1992 年から 2016 年への経年変化を見た場合，少なくとも文章量的観点からの顕著な変化は見られず，むしろ 2016 年現在の子ども達の方が長い作文を書くことができるという 2 節の予想を覆す結果が得られた。

３．４　先行研究における作文分析の観点

先行研究に於ける作文コーパス等の言語資源を利用した代表的な計量的研究の観点を整理すると以下のようになる。

① 文章量，語数，文数等･･･成田他（1995），石井（1996），宮城・今田（2015）
② 語種，語数（異なり語数，述べ語数）･･･村上・田中（1997），成田他（同），宮城・今田（同）
③ 語彙・漢字使用 ･･･石井（同），宮城・今田（2016）

[7] これまで「児童作文コーパス」を利用した研究から，文章量変化や語彙・漢字使用の熟達等の発達過程に於いて大きな変化が見られるのが，小学 2 年と小学 3 年の間と中学 2 年と中学 3 年の間であることが確認されている（宮城・今田 2016）。その意味でも 1992 年の小学 5 年の欠陥による問題が局所的なものとなることが予想される。

④　副詞（情態・程度）･･･川口・佐々木（1997），宮城（2016）
⑤　文末形式（助動詞）･･･佐々木・川口（1994），宮城（2015）
⑥　接続詞（順接，逆接，転換等）･･･小川（1997），冨士原他（2016）
⑦　指示詞 ･･･佐々木（1997）
⑧　文法的誤り・不具合 ･･･内田・瓜生（1997），松崎（2015，2016a）
⑨　理由述べ表現 ･･･松崎（2016b）

これら先行研究の着眼点は，本発表における児童らの文章作成能力の経年変化の捉え方に示唆を与えてくれる。3.3 節で述べたように文章量的観点からの違いが見いだせないことが確認できたことから，教師らの「最近の子ども達は作文が書けなくなった」という指摘は単純に長い文章が書けなくなったということではないことが明らかになった。とすれば，次に着目すべきは，どのような語彙や形式を用いて，どのように表記しているのかといった表現レベルでの問題であろう。以下，4 節では，上記の観点を参考にして「手」コーパス 1992 と「手」コーパス 2016 を対照して文章量以外の差異を見いだすことを試みる。以下，それぞれのコーパスのデータを 1992 年の資料，2016 年の資料と呼ぶ。

４.「手」作文コーパスを用いた計量的分析

４．１　漢語使用の経年変化

村上・田中（1997）は，1992 年調査の「手」作文を資料として名詞，形容動詞，動詞の漢語使用状況を調査し，辞書量，造語性，習得時期，生活語彙と抽象語彙などの観点から詳細な考察を行っている。本発表では，両資料の差異についての基礎データを得るために個別の語彙項目の分析には立ち入らず，名詞，形容動詞，動詞における漢語の頻度について，1992 年と 2016 年の資料を比較する。UniDic 品詞体系においては，名詞と形状詞（形容動詞語幹相当）が調査対象となる。UniDic の名詞は，サ変動詞語幹として使用される名詞も含むため，名詞を調査対象とすれば漢語動詞も調べることができる。

図 3 は，1992 年と 2016 年資料における各学年の 1000 語あたりの漢語頻度（名詞，形状詞のみ）を示したものである。図 4 は，それをさらに男子と女子とに分け，1000 語あたりの漢語頻度を示したものである。

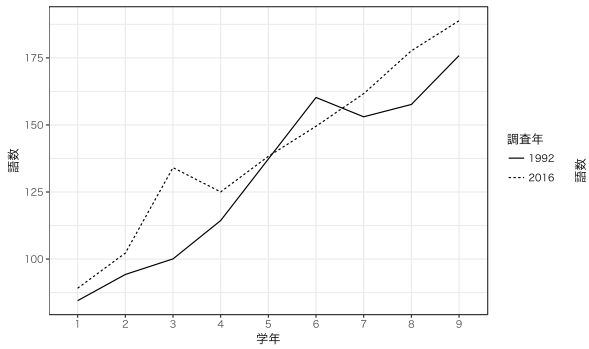


図 3　漢語表現の出現頻度

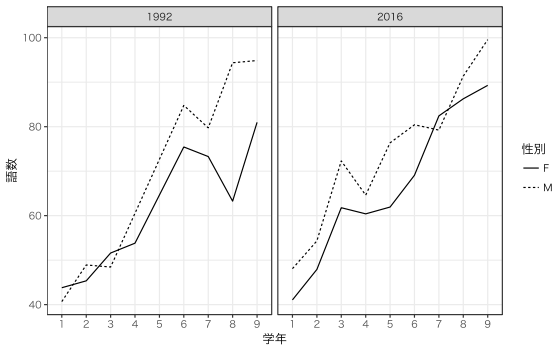


図 4　漢語表現の出現頻度（男女別）

1992 年と 2016 年の資料に共通の特徴として，図 3 から，学年が上がるほど漢語の使用頻度が増加する傾向があることが分かる。また，図 4 から，多くの学年において女子よりも男子の方が漢語の使用頻度が高いことが分かる。男女差の理由は不明だが，1992 年と 2016 年の両方で経年的に同様の特徴が観察されたことは興味深い結果である。

1992 年と 2016 年の資料の差異としては，図 4 から，多くの学年において 1992 年よりも 2016 年の資料の方が漢語の使用頻度が高いことが分かる。このことは，1992 年より現在の子ども達の方が漢語の習得時期が早いことを示唆している。また，漢語の使用頻度は学年が上がるほど増加する傾向があるが，必ずしも単調に増加するわけではなく，図 3 から，1992 年の資料は小学 6 年，2016 年の資料には小学 3 年にピークがあり，翌年度には頻度が減少している。このピークの位置も，1992 年より 2016 年の資料の方が早い学年にある。

４．２　指示表現・人称表現の経年変化

佐々木（1997）は，小学生から大学生までの作文における指示表現の使用状況を調査し，小学校の間は指示表現の頻度が増加するが，中２～高１でピークに達すること，小学校低学年ではソ系指示詞が多用されるが，次第にコ系指示詞の頻度が増加し，小学 4 年以降はコ系とソ系がほぼ 2 対 3 の割合で推移することなどを指摘している。ここでは，コ系，ソ系，ア系の指示表現の頻度について，1992 年と 2016 年の資料を比較する。調査対象とするのは，以下の指示表現である。

名詞	これ，それ，あれ，ここ，そこ，あそこ
連体詞	この，その，あの，こんな，そんな，あんな
副詞	こう，そう，ああ

図 5 は，1992 年と 2016 年の資料における各学年の 1000 語あたりの指示表現の頻度を示したものである。図 6 は，さらにコ系，ソ系，ア系に分けて，1000 語あたりの頻度を示したものである。

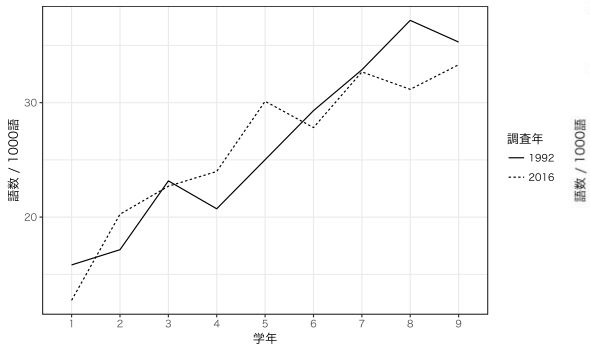


図 5　指示表現の出現頻度

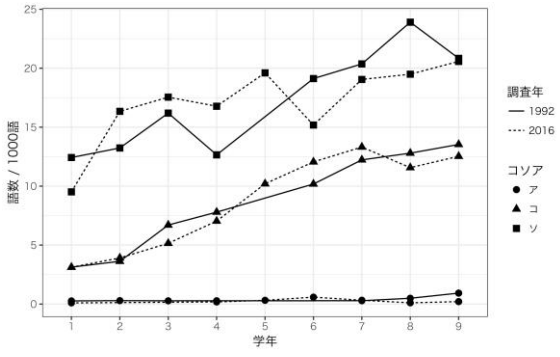


図 6　指示表現の出現頻度（コソア別）

図 5 から，1992 年と 2016 年の資料のいずれも学年が上がるに従って指示表現の頻度が増加することが分かる。低学年では 2016 年，高学年では 1992 年の資料での頻度がやや高いように見えるが，違いは明確でない。また図 6 を見ると，1992 年と 2016 年の資料いずれも，全ての学年でソ系指示詞の頻度が最も高く，次いでコ系指示詞の頻度が高く，ア系指示詞は

ほどんど使われていないことが分かる。学年によって多少の上下はあるが，全体としては1992 年と 2016 年の資料間に明確な経年変化は確認できない。

次に，先行研究では扱われていないが，人称表現の経年変化を調べる。児童作文では二人称や三人称の代名詞はあまり生起しないが，一人称代名詞は高頻度で生起する。そのほとんどは「私」「僕」のいずれかであり，男女差が顕著に観察される。ここでは「私」「僕」の 2 つの代名詞について，1992 年と 2016 年の資料における 1000 語あたり頻度を確認する。図 7 は一人称代名詞全体の頻度（「私」「僕」の合計）を男女別に示したものである。図 8 は「私」「僕」の内訳を示したもので，女子は全て「私」であったため，男子のみを示した。

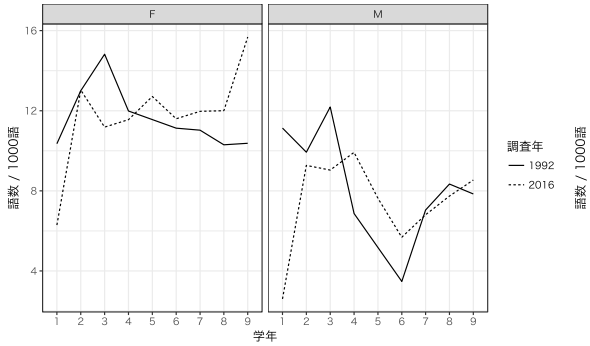


図 7　一人称表現の出現頻度（男女別）

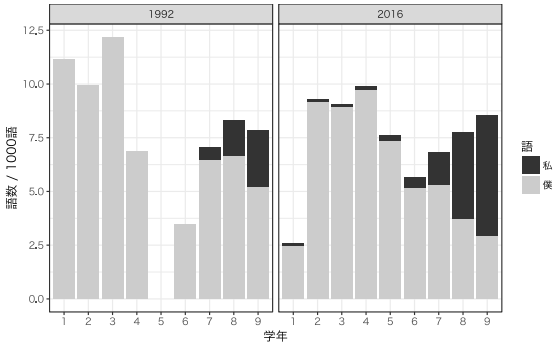


図 8　一人称表現の出現頻度（男子のみ）

図 7 から，1992 年と 2016 年の資料での共通の特徴として，全体的に男子より女子の方が代名詞の頻度が高い傾向があること，学年が上がるにつれて代名詞の頻度も単調に増加するとは言えないこと，特に男子は小学校中学年から高学年にかけて頻度が減少し，中学校以降再び上昇することなどが観察できる。1992 年と 2016 年の資料の差異としては，低学年では 1992 年，高学年では 2016 年の資料での頻度がやや高いように見えるが，全体として明確な違いは確認できない。

また図 8 からは，男子は小学校では「僕」を主に使っているが，中学校になると「私」の使用が増加することが分かる。特に 2016 年の資料では「私」の増加が顕著であり，1992 年の資料では中学 3 年でも過半数が「僕」であるが，2016 年の資料では過半数が「私」であることが確認できる。これは 1992 年と 2016 年の資料に於ける中学校段階における書き言葉文体の変化を示唆する結果と考えられる。

４．３　文末表現の経年変化

佐々木・川口（1994）では，1992 年の資料を調査して，文末表現の発達過程を明らかにしている[8]。また，宮城（2015）では「児童作文コーパス」の一部（2014 年調査）を資料として同様に文末表現の発達について言及している[9]。文末表現の発達を考える場合，個々の形

[8] 佐々木・川口 1994 では，日本語母語話者の作文では，学年を追って増加していき，「説明・真偽判断のモダリティには特に顕著な増加傾向が見られた。」（p.11）とし，「推量表現に関しては，主観性の強いものから客観性を帯びたものへと発達していく使用過程が窺えた。」（同）とも指摘している。

[9] 宮城 2015 では，佐々木他 1994 との異同を検証した上で，「全体的な傾向として，推量→蓋然性→証拠性の順に使用頻度が高くなるようである。」（pp.26-27）と指摘している。

式の語の特徴を見なければならないが，形式間の比較をする場合カテゴリ毎にまとめた方が捉えやすい（先の 2 研究も同様の立場である)。そこで本発表では,「推量」「蓋然性」「証拠性」「引用」という 4 つのカテゴリ毎に各文末形式使用の発達を調査した。各カテゴリに含まれる形式は以下の通りである。

推量	だろう
蓋然性	かもしれない，はずだ
証拠性	ようだ（伝聞)，ようだ（様態)，そうだ
引用	と考える

ここでは各カテゴリ毎に, 1992 年と 2016 年の資料における 1000 文あたり頻度を確認する。

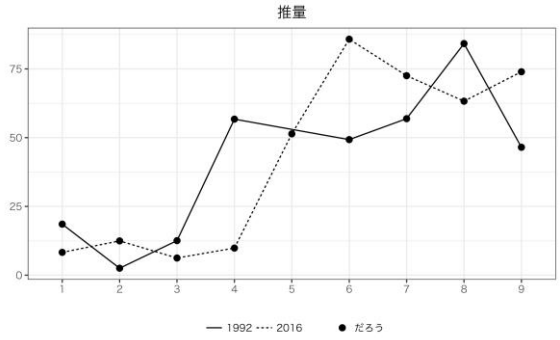


図 9　文末表現　（推量）

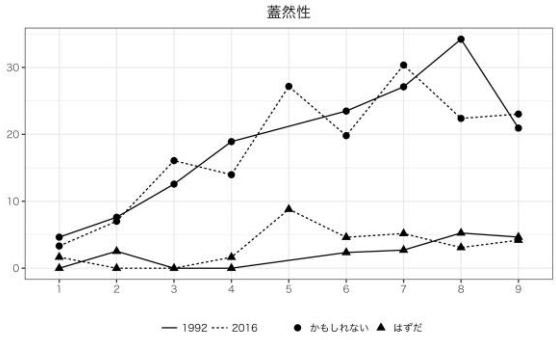


図 10　文末表現（蓋然性）

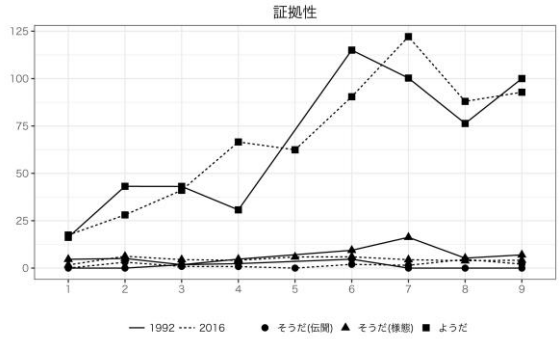


図 11　文末表現（証拠性）

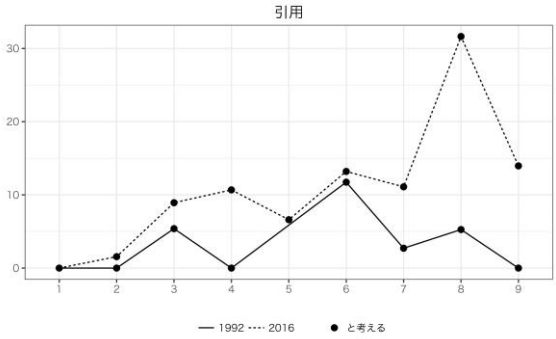


図 12　文末表現（引用）

各カテゴリでの使用頻度の変化を順に見ていこう。図 9 から，推量を表す文末形式では，1992 年と 2016 年のいずれの資料でも学年が上がるに従って同じような形で増加し，減少に転じることが分かる。ただし 1992 年の資料では小 4 から，2016 年の資料では小 5 から急激に増加している。一方で，増加のピークは，1992 年の資料で中 2，2016 年の資料で小 6 と 2 年早まっている。次に図 10 から，蓋然性を表す文末形式では，1992 年と 2016 年の間に顕著な違いは見られない。「かもしれない」は学年が上がるに従って増加し，それに対して「はずだ」は学年が上がっても頻度にさほど変化が見られない。「はずだ」は定着が遅れる（定着しない）語ということになる。続けて図 11 から，証拠性を表す文末形式では，同様に両資料間に顕著な違いはなく，「ようだ」は学年が上がるに従って同じような形で増加するの

に対して，「そうだ（伝聞・様態）」の頻度はさほど変化が見られない（定着が遅れる）。最後に図 12 から，引用を表す文末形式では両資料ともに学年が上がるに従って同じような形で増加していくが，1992 年の資料が中 1 辺りから減少に転じるのに対して，2016 年の資料は中 3 で減少に転じている。文末形式の使用を総括すると，学年別の発達はかなり似た傾向を示すが，いくつかの形式では増加と減少の学年がずれる場合があることが確認された。

４．４　程度表現の経年変化

副詞使用の発達過程を考える場合，雑多な語を含むカテゴリを一括して取り上げることはできない。形式が安定していて語数が限られており抽出しやすいこと，副詞に限らず程度表現の使用が比較的早い段階から見込まれることから，本発表では程度副詞を中心に発達過程を考察することにする。副詞の学年別使用については川口・佐々木（1996）が 1992 年の資料を調査して，副詞の学年の進行による使用傾向に言及している[10]。児童作文に於ける副詞の調査は容易ではない，学年進行による発達以外の要素が大きく，おそらく文内容に関係した使用動機等に左右されると考えられる。川口・佐々木（1996）に於いても「このような量的調査からは，学年ごとのばらつきがかなり大きいため，特徴的な発達過程を得ることはできなかった。」（p.236）と指摘される[11]。一方で「情態副詞の中で，学年による特徴が見られたオノマトペを取り上げ，～」（同）としていることから，本発表では，頻度が比較的高いと予想されるオノマトペ副詞（擬態語・擬声語）と逆に修得が遅い（頻度が低い）と予想される評価を表す副詞を合わせて取り上げることにした[12]。3 カテゴリの副詞が使用全体に占める使用頻度比（副詞の使用頻度全体に対する各カテゴリ別の使用頻度の比率）を調査した。学年間の差異については，小 2，小 4，小 6，中 2（8）を抽出して比較した。

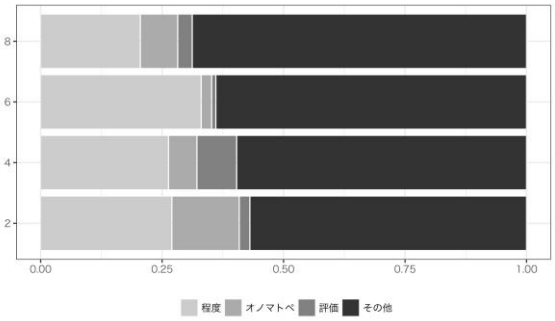


図 13　1992 年資料の資料頻度比

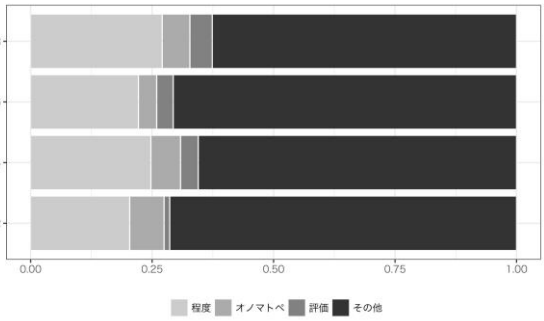


図 14　2016 年資料の資料頻度比

図 13，14 から，程度副詞の使用頻度比は，1992 年と 2016 年の資料のいずれでも他のカテゴリの語に比べて卓立して高いことが分かった。2016 年の資料では学年が上がるに従って頻度比はほぼ増加しているが，1992 年の資料では，小 6 以降で減少に転じている。次にオ

[10] 宮城 2016 では，2014 年の資料（児童作文コーパスの一部）を基に程度副詞の用法別発達を調査しているが，やはり語別の傾向記述に留まり明示的な発達過程の解明には至っていない。

[11] 実際「「手」作文コーパス」での調査に於いても副詞毎の使用頻度調査ではかなりの偏りが見られた。

[12] 本節の調査は，基本的に機械解析の語彙情報によっているので，全ての副詞を収集できているわけではない（特にオノマトペ副詞に漏れがある可能性がある）。評価を表す副詞は，上記の語彙情報を基に選出した「思いの外」「案外」「当然」のような語を含む。基準は川口・佐々木（1996）に準じている。

ノマトペ副詞の使用頻度比では，1992 年では小 6 まで減少した後，中 2 で増加に転じている。一方，2016 年の資料では各学年を通じて使用頻度比の揺れ幅は少ない。最後に評価を表す副詞では，1992 年の資料で小 4 まで増加した後減少に転じている（小 4 で特によく使用されている)。対して 2016 年の資料では，学年の進行に従って頻度比が増加した後減少しないことが分かった。ここまでの観察から，副詞の使用頻度の学年別変化については，1992 年と 2016 年の資料から差異を見いだすことができた。1992 年の資料では，副詞の使用頻度比がある段階まで増加した後減少に転じる，即ち修得後に状況に応じて使い分ける選択の段階が観察できるのに対して，2016 年の調査では，その傾向が観察されず，調査範囲では修得が進みつつある過程を示し，使い分けの段階までには至っていない可能性がある。

５．まとめ —24 年間の児童生徒らの文章作成能力の経年変化の実態—

本発表では，児童・生徒の文章作成能力の実態を映した「手」作文コーパスを利用した児童生徒の文章作成能力の学齢別発達過程と経年変化とを関連付けた立体的な計量的分析を試みた。基本的な文章量・語数・文数等の量的な違いを始めとして，漢語，指示詞（代名詞），文末形式，副詞の使用頻度の傾向を調査したが，1992 年と 2016 年の資料間での明確な差異を見い出すことはできなかった。よって，これらの観点からは「最近の子ども達は作文が書けなくなった」という指摘を否定する結果が得られたことになる。しかしながら，本発表で取り上げた観点以外にも文章の構成や表現技巧の巧拙等の文章内容に関する事項で違いが見られる可能性がある。実際に一部の文末形式，程度副詞で文体にあった選択の萌芽が確認された。これらを含め，単純に抽出できない要素をどのような形で計量化して分析することが有効であるのかが今後の課題となる。

本コーパスは規模が約 29 万形態素と中規模と呼ぶにも些か心もとないデータ量ではあるが，児童生徒らの 24 年間の文章作成能力の経年変化を反映した資料としては他を持って代えがたい価値がある（管見の限り現在このような資料は存在していない）。また，条件が統一された調査資料であること，子ども達が自身の力だけで書き上げた作文であること，「児童作文コーパス」を始め図 1 の関連するコーパス類と調査条件や解析手法，付与されるメタデータの種類などのフォーマットが統一され，本コーパスから得られた結果を容易に相対化することができる利点を勘案すると，そのデータ規模の弱点を十分に補いうると考えられる。また，調査対象の学年が義務教育課程 9 年間（小 1〜中 3）の全体をほぼカバーしていることから，経年変化と並行して学年別の発達過程を関連付けた立体的な調査を実施すること可能である。今後，本コーパスを用いた研究が積み重ねられることによって，本コーパスの真正性やどのような調査研究が可能であるのかが明らかにされていくことを期待する。それは即ち児童生徒らの文章作成能力が解明されていくこととほぼ同義である。

これまで子ども達の文章作成能力の変化は，ベテラン教師らによる経験知から語られるのみであった。その意味でも「最近の子ども達は作文が書けなくなった」という指摘の意味は大きい。一方で，教師らがそのように感じる理由が具体的に何であるのかを追求することも必須の課題である。なぜなら問題点を正確に把握していなければ，子ども達に具体的かつ効果的な文章指導をすることがほぼ不可能だからである。今後も現場教師の経験知の価値は些かも失われることはないが，一人の教師の分析能力に限界があることも十分に認識しておかなければならない。問題を明確にして解決に向かうために，国語教育学研究に於いても客観的な観点の導入が喫緊の課題である。その意味でも作文コーパス等を活用した計量的研究への期待は大きい。

本発表の内容を含む一連の研究の最終的な目標は，児童生徒らの文章作成能力の解明にある。文章作成能力は学年の進行と共に段階的に発達していくものであるので，学年毎の発達段階間の対比や本発表で示した経年変化，文章のジャンルによる差異なども十分に考慮されなくてはならない。またその成果は，教育現場における作文教育の改善と適正化を図ることにも活用されることが望まれる。「「手」作文コーパス」も含め，図 1 で示した関連作文コーパス類を現場の教師でも簡便に利用できるように「作文検索システム Kodama」も併せて調整中である[13]。

謝　辞

本発表は，博報財団第 11 回児童教育実践についての研究助成「児童・生徒の文章作成能力経年変化解明と現場と協働した指導法の開発」（2016 年度，研究代表者：宮城信，助成番号：2016053）による補助を得ています。

文　献

石井健介（1996）「児童作文の計量的分析の試み」，『学芸国語国文学』28，pp.82-90，東京学芸大学国語国文学会

内田安伊子・瓜生佳代（1997）「母語発達と文のねじれとの関係―「･･･は＋述部」の形を持つ文について―」，pp.100-108，長友和彦他（1997）

川口良・佐々木泰子（1996）「日本人と日本語学習者の作文における副詞の発達過程に関する研究」，『お茶の水女子大学人文科学紀要』49，pp.219-238，お茶の水女子大学（長友和彦他 1997 に再録）

国立国語研究所（1989）『児童の作文使用語彙（国立国語研究所報告 98）』東京書籍.（http://www.ninjal.ac.jp/s_data/drep/report_nijla/R0098.PDF よりダウンロード可能）

佐々木泰子（1997）「日本語における結束性の発達と習得―指示語と繰り返し―」，pp. 31-41，長友和彦他（1997）

佐々木泰子・川口良（1994）「日本人小学生・中学生・高校生・大学生と日本語学習者の作文における文末表現の発達過程に関する一考察」，『日本語教育』84，p.1-13，日本語教育学会（長友他1997に再録）

冨士原紀絵・宮城信・松崎史周（2016）「児童生徒作文の基礎的研究―児童生徒作文コーパスの構築と活用―」，『こども学研究紀要』4，pp.9-20，お茶の水女子大学子ども学研究会

長友和彦他（1997）『児童・生徒・学生及び日本語学習者の文章作成能力の発達過程に関する研究』，平成 8 年度文部省科学研究費補助金成果報告書

成田信子・宗我部義則・田中美也子（1995）「文章作成能力発達に関する縦断的研究 その一 ―小学生から大学生に至る同題作文の分析―」，『国語科教育』42， pp.183-192，全国大学国語教育学会（長友他 1997 に再録）

松崎史周（2014）「戦後作文・文法指導における「文法的誤り」の扱いー昭和 30 年前後を中心にー」，『目白大学人文学研究』10，pp.301-317，目白大学

松崎史周（2015）「中学生の作文に見られる「主述の不具合」の分析ー出現傾向から学習者の表現特性を探るー」，『解釈』61（5・6），pp.12-20，解釈学会

[13] 現在の「作文検索システム Kodama」は version1.3 である。詳細は宮城・今田 2015 等を参照されたい。

松崎史周（2016a）「国語教育における「だらだら文」の捉え方と扱い」,『日本女子体育大学紀要』46，pp.111-121，日本女子大学
松崎史周（2016b）「児童作文における「理由述べ」表現の分析－「将来の夢」に関する作文の場合－」,『解釈』62（5・6），pp.12-21，解釈学会
宮城信（2015）「児童・生徒作文に見る文末表現の発達―作文の表現指導との関わりから―」,『富山大学国語教育』40，pp.22-30，富山大学国語教育学会
宮城信（2016）「児童作文に見る程度修飾表現の発達」,『富山大学人間発達科学部紀要』10（2），pp.291-297，富山大学人間発達科学部
宮城信・伊集院郁子・盧妊鉉・文智暎（印刷中）「『日韓対照大学生作文コーパス』の構想」,『筑波日本語研究』21，筑波大学日本語学研究室
宮城信・今田水穂（2015）「『児童・生徒作文コーパス』の設計」,『第 7 回コーパス日本語学ワークショップ予稿集』，pp.223-232，国立国語研究所（https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no7_papers/JCLWorkshop_No.7_27.pdf よりダウンロード可能）
宮城信・今田水穂（2016）「作文コーパスを資料に児童・生徒の漢字使用・選択傾向と発達の実態を明らかにする―語彙情報付き作文コーパスの構築と学齢別語彙・漢字使用実態調査―」,『漢字・日本語教育研究』5，pp.4-20，公益財団法人 日本漢字能力検定協会
村上博之・田中美也子（1997）「同題作文における漢語表現の発達」，pp. 1-22，長友和彦他（1997）

『日本語日常会話コーパス』収録の進捗状況

田中 弥生（国立国語研究所 音声言語研究領域 ／ 東京大学大学院 総合文化研究科）†
柏野 和佳子（国立国語研究所 音声言語研究領域）
角田 ゆかり（国立国語研究所 音声言語研究領域）
伝 康晴（千葉大学文学部・国立国語研究所 音声言語研究領域）
小磯 花絵（国立国語研究所 音声言語研究領域）

Construction of "Corpus of Everyday Japanese Conversation" : Progress Report of Recording Naturally Occurring Conversations

Yayoi Tanaka (National Institute for Japanese Language and Linguistics / The University of Tokyo)
Wakako Kashino (National Institute for Japanese Language and Linguistics)
Yukari Sumida (National Institute for Japanese Language and Linguistics)
Yasuharu Den (Chiba University / National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

2016 年度から構築が始まった「大規模日常会話コーパス」プロジェクトによる『日本語日常会話コーパス』の収録手続きの概要と進捗状況について報告する。本プロジェクトでは，日常場面の中で自然に生じた会話を対象とする。そのため，性別・年代などの点からバランスを考慮して調査協力者を選別し，収録機材等を 2〜3 カ月程度貸し出して調査協力者自身に日常会話を収録してもらう方法を採用している。本発表では，こうして定めた収録方法の概要を述べるとともに，これまでに終了した 13 名の調査協力者による約 200 時間の収録について進捗状況や生じた問題などを報告する。

１．はじめに

機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）では，日常場面で自発的に生じた会話約 200 時間を収録した大規模なコーパス『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC)を構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性をレジスター・相互行為・経年変化の観点から多角的に解明することを目指している。本稿は，『日本語日常会話コーパス』構築における収録の方法を概説し，進捗状況について報告する。２節で本コーパスの概要を，３節で現在行っている収録方法を説明し，４節でこれまでの収録状況の報告を行い，５節で今後の予定について述べる。

２．コーパスの概要

日常場面での会話を収録するためには，収録場面を人工的に設定するのではなく，日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話(naturally occurring conversation)を対象としなければならない。日常の言語生活を反映したコーパスの設計に際し，我々が普段どのような場面で会話しているか実態を知るため，本コーパスの構築に先立

† yayoi@ninjal.ac.jp

って，2015 年度に「会話行動調査」を実施した。この調査結果を参考に，多様な会話をバランスよく収めたコーパスを構築する。調査の詳細は小磯他(2016)を参照されたい。

理想的には，起床してから就寝までの，自宅や職場，店舗，屋外，交通機関など，さまざまな場所で生じる会話が対象となる。このような多様な日常場面での会話を収録するために，British National Corpus(BNC)の spoken part の収録法(Burnard and Aston 1998)を参考に，以下の 2 つの収録法を採用することとした。

個人密着法　性別･年代などの点からバランスを考慮して選別された調査協力者（以下，協力者）に収録機材等を一定期間貸し出し，協力者自身に会話参加者（以下，会話者）との日常会話を収録してもらう方法。プロジェクトメンバー（以下，調査者）は原則として介在しない。

特定場面法　職場での会合や店舗での店員とのやりとりなど，個人密着法では技術的・倫理的に収録が難しいと思われる場面を特定し，調査者が主体となり収録する方法。調査者は介在するが，日常場面の中で自然に生じる会話を対象とする。

本プロジェクトではまず個人密着法に基づき収録調査を開始した。構成は以下の通りである。なお，コーパス設計の詳細は小磯他(2017)を参照されたい。

■調査期間：　2016 年 4 月～2018 年度（予定）。
■協力者の属性：首都圏（東京都，神奈川県，埼玉県，千葉県）に在住の 20 代以上の男女。出身地や生育地域の制限は設けていない。
■協力者の人数：約 40～50 名。
20 代，30 代，40 代，50 代，60 代以上の男女，それぞれ 4-5 名を予定。協力者は個人情報を取り扱うなど重い責任が生じることから，未成年者を含めないこととした[1]。
■収録時間：　協力者 1 名あたり 15～18 時間。
■コーパスへの採録時間：協力者 1 名あたり約 4～5 時間。40～50 名で合計 160～200 時間。

個人密着法での収録をある程度進めた段階で，不足する種類の会話を補うために，特定場面法を実施する。コーパス全体で 200 時間の規模を目指す。本稿では，すでに調査が進んでいる個人密着法の収録方法と，収録状況を述べる。

３．収録方法

本節では，個人密着法による収録方法について述べる。上述のとおり，この収録法では，研究者は収録場面に立ち会わず，収録に伴う一連の作業を協力者自身に担当してもらう必

[1] 協力者が集める会話の中に未成年者が含まれることはある。しかし個人密着法では，必ず協力者（つまり成人）が加わる会話のみが対象となるため，未成年者のみにより構成される会話がこの方法で収録されることはない。個人密着法で収録したデータの会話者属性の性質を調査し，仮に未成年者が少ないなどの偏りが見られる場合には，特定場面法などで補うことも検討する。

要がある。そのため，収録調査の手続きや関連資料などを入念に検討して定めた。なお，田中他(2017)でその詳細を述べたため，本節では概要を記すにとどめる。

３．１　協力者の募集方法

主に調査者の伝手により協力者を集めた。属性が偏らないよう，年代・性別の他，職業の有無も考慮した。候補となる人に，協力者募集のチラシあるいはプロジェクトのホームページにて概要を確認してもらった後，30 分～1 時間程度調査についての詳細な説明を行って，意思を確認したうえで，協力者を決定した。

３．２　協力者に依頼する作業

協力者が行う作業は，以下の通りである。

① 会話の収録（録画・録音）
② 会話者への調査内容及び公開方法の説明
③ 会話者への同意書への署名の依頼
④ 会話状況（日時，使用機材，配置など）の記録
⑤ 会話者の属性（性別，出身地など）に関するメタ情報収集のための会話者へのフェイスシート記入の依頼
⑥ 自宅等での機材や書類の管理
⑦ 定期的なデータ提出
⑧ メールや電話などでの調査者とのやりとり
⑨ 各種打合せ（収録開始時の一連の調査方法についての説明・調査終了時のフォローアップインタビュー，いずれも 3 時間程度）

一連の調査協力に対し，協力者に謝金 12 万円を支払う。謝金の額は，テスト収録に基づき作業量を試算した上で確定した。調査者が介在せずに上記の作業を行ってもらうため，マニュアルは詳細に整備し，機材については協力者の負担が最も少ない形での設定とした。また，会話者の同意書やメタ情報を適切に収集できるよう，同意書やフェイスシートの様式も改良を重ねた。同意書及びメタ情報の収集については田中他(2017)を参照されたい。協力者については，個人情報取り扱いも含めたガイドライン（複製の禁止，調査で得た個人情報を調査以外に用いない，データ保管の安全性の確保など）を作成し，調査開始時に説明の上，同意書に署名を得ている。

自然に発生する日常会話を収録するため，協力者には，収録のために人を集めるのでなく，日常の自宅での家事や食事，収録とは関係なく設定された会食や打ち合わせなどの場面に機材を持ち込み，会話者の同意を得たうえで収録するよう求めた。収録場所や場面，会話の相手などに関して，多少のバリエーションがあることが望ましいことも伝えた。

３．３　収録調査の流れ

収録調査期間（機材貸し出し期間）は基本的には 2 カ月程度，最大で 3 カ月とした。大まかな流れは以下のとおりである。一度にすべての収録を行うのではなく，4～5 回に分けて収録してもらい，随時調査者が確認し，フィードバックを行う。特に第一次収録終了後は必ずフィードバックを待ってから第二次収録を開始してもらうこととした。詳細については田中ほか（2017）を参照されたい。

① 機材等送付（調査者から協力者へ）

② 収録開始時打ち合わせ
③ 第一次収録
④ フィードバック（調査者から協力者へ）
⑤ 第二次収録
⑥ 第三次収録
⑦ 第四次収録 （必要に応じて第五次収録）
⑧ 機材等返却（協力者から調査者へ）
⑨ 調査終了時打ち合わせ

３．４ 収録機器

協力者は必ずしもカメラなどの機械類の取り扱いが得意なわけではなく，また自宅外に機材を持ち出すこともあるので，シンプルな設定，短時間での設営，簡単な操作，軽量といった制約のもとで収録方法と収録機器を検討し，次のような方法で収録を実施している。

３．４．１ 基本収録

基本的な収録の機器を表 1 に示す。また，2 名が対面する収録の基本的な配置で撮影される映像を図 1 に示す。表 1 の各機器を図 1 内に記号で表示した。

表 1 基本収録機材

	品名	設定	基本使用台数	特徴	図 1 内記号
映像	Kodak PIXPRO SP360 4K	1440×1440, 60fps	1	360 度撮影可能なカメラ。会話者たちの中央に配置。	(a)
	GoPro Hero3+	1920×1080, 60fps	2	170 度の視野角を持つカメラ。会話者を俯瞰的に記録。	(b1)(b2)
音声	IC レコーダー Sony ICD-SX734	リニア PCM, 44.1kHz,16bit	最大 6	会話者ごとにフォルダーに入れて首から下げる。	矢印
	IC レコーダー Sony ICD-SX1000	リニア PCM, 44.1kHz,16bit	1	会話全体を録音。中央に配置。	(c)

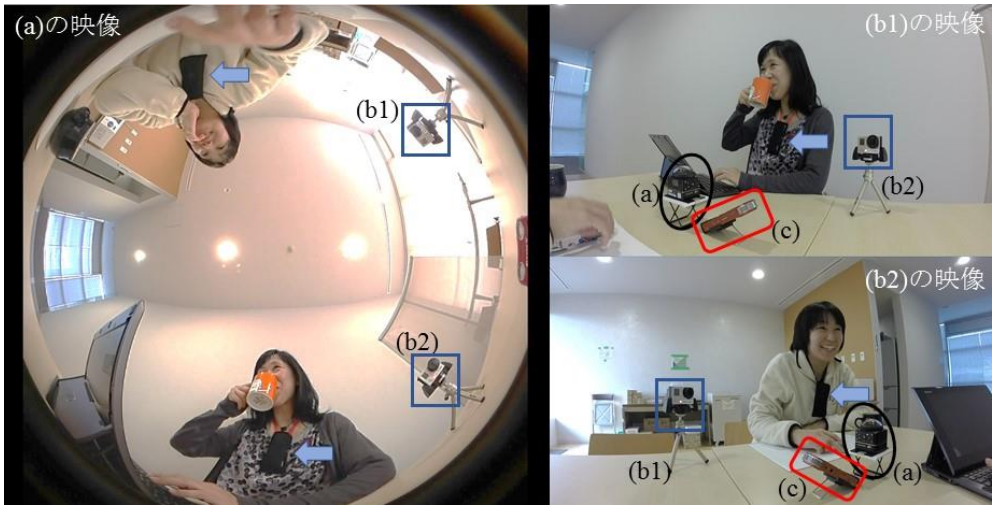


図 1 対面 2 名会話時の映像

３．４．２　基本収録以外の機材

基本収録以外の機器について，表 2 に示す。

表 2　基本収録以外の機器等

	品名	数量	特徴
移動時収録	Panasonic HX-A500	1	ウェラブルカメラ。散歩や散策，外出先への移動の際に，1 名が頭に装着。音声は基本収録の会話者ごとの IC レコーダーを使用。
車内収録	車載用アクセサリ	適宜	カメラは基本収録 GoPro を主に使用。会話者ごとの IC レコーダーも使用。
電話会話収録	小型マイクロフォン	1	スマートフォンと IC レコーダーを接続するコードをあらかじめ接続した小型マイクロフォンを使用。

３．５　マニュアル（手引き）

収録調査が問題なく進められるよう，マニュアルを用意した。調査の進め方や，機材の取り扱い方法，データの提出のタイミングや方法などを記載した『会話収録の手引き』の他，具体的な機器の操作については，別冊で『会話収録の手引き―基本収録編―』『会話収録の手引き―移動編―』『会話収録の手引き―電話編―』を作成した。

４．　収録状況の報告

上述の表 1 に示した機材を 6 セット用意し，2016 年 4 月から順次調査を開始してきた。約 10 カ月が経過し，表 3 に示したように，合計 13 名の収録調査者が調査を完了し，6 名が現在調査中である。

表 3　協力者の属性（2017 年 1 月 24 日現在）

年代	男	女	計
20 代	学生　(終了) 学生　(調査中)	学生　(終了) 学生　(終了)	4 人
30 代	自営自由業　(終了) 自営自由業　(終了)	専業主婦　(終了) 会社員等　(終了) 会社員等　(終了)	5 人
40 代	自営自由業　(終了) 会社員等　(調査中)	会社員等　(終了) 自営自由業　(調査中) 専業主婦　(調査中)	5 人
50 代	自営自由業　(調査中)	自営自由業　(終了)	2 人
60 代以上	無職　(終了) 非常勤講師　(調査中)	専業主婦　(終了)	3 人
計	9 人	10 人	19 人

本稿では，このうちすでに調査が終了した 13 名の収録状況について報告する。なお，収録されたデータのうちコーパスに格納するデータ（全体の約 3〜4 分の 1）については，小磯他(2017)を参照されたい。

４．１　収録回数と収録時間

13 名の年代別にみた収録回数と収録時間の内訳は表 4 の通りである。20 代と 40 代の一人当たり収録時間が少ないのは，それぞれ 1 名ずつ，調査期間中に転職などの理由によって，予定していた収録ができず調査期間を終えたことによる。なお，３．３で，収録調査期間（開始時打ち合わせから最終収録日まで）は基本的に 2 カ月と述べたが，最も短い協力者で 46 日，最も長い協力者で 91 日，平均すると 67 日であった。

表 4　年代別収録回数及び収録時間（調査終了分）

年代	協力者人数	収録回数	一人当たり収録回数	収録時間（時間：分）	一人当たり収録時間（時間：分）
20 代	3	51	17	42:43	14:14
30 代	5	103	20.6	86:02	17:12
40 代	2	33	16.5	25:26	12:43
50 代	1	19	19	17:02	17:02
60 代以上	2	34	17	31:27	15:43
計	13	240	18.5	202:40	15:35

４．２　収録形態

３．４で述べたように，調査には，大きく分けて，基本収録，移動時収録，さらに電話会話と車内での収録という 4 つの形態がある。図 2 に，収録形態ごとの収録回数を示す。

基本収録では，３．４．１で述べたように， SP360，GoPro（2 台），中央に配置する IC レコーダー, 会話者一人一人が首からかける IC レコーダーの，4 種類の機材を使用する。すべての機材を使用した収録（フルセット）は基本収録の約 6 割を占めている。しかし，何らかの事情により，すべての機材を使用できていない場合もある。図 3 のように，2 名が並んで座る場合，正面に 1 台の GoPro を配置することで図 4 のように 2 名が十分に撮影できるため, 1 台の

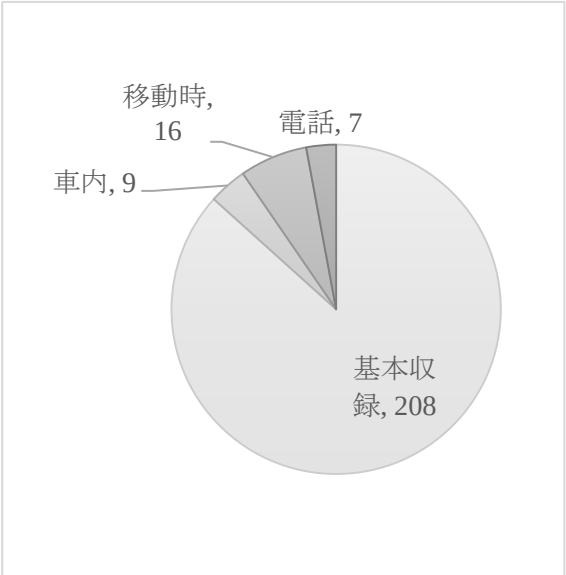


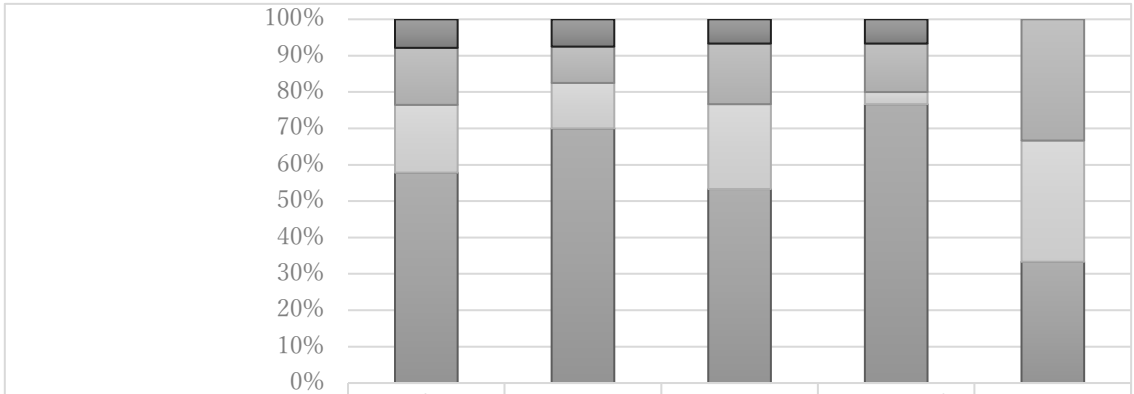
図 2　収録形態ごと収録回数

図 3　2 名並んでの収録

図 4　2 名並んでの収録時 GoPro からの映像

み使用することがある。「GoPro1 台のみフルセット」はこのような状況が含まれ，約 16%を占めている。「GoPro なしフルセット」は SP360 と IC レコーダーでの収録で，約 15%であった。その他に，収録をはじめてすぐに SP360 の電池が切れたためやむを得ず SP360 のない構成で収録を継続したケースや，貸し出している IC レコーダーよりも多い人数での収録になり，やむを得ずレコーダーを付けない人が含まれているケース，IC レコーダーの電源の入れ忘れなどがあった。

収録場所別の機材使用状況を図 5 に示す。「屋外」以外ではフルセットの使用が 5 割を超えている。ある程度落ち着いて収録準備ができることが要因と考えられる。「自宅」「その他の室内」は 7 割程度がフルセット使用である。「その他の室内」は，実家や友人宅など協力者の自宅以外に機材一式を持参して収録するケースが該当する。いずれも，時間的空間的に余裕をもって機材を配置することが可能であるためであろう。レストランや公民館などの「公共商業施設」と「職場学校」では，GoPro1 台のみフルセットと GoPro なしフルセットが合計して 3〜4 割ある。準備の時間が限られたりテーブルやスペースが小さかったりなど，時間的空間的に余裕がない場合が相対的に多いことが考えられる。

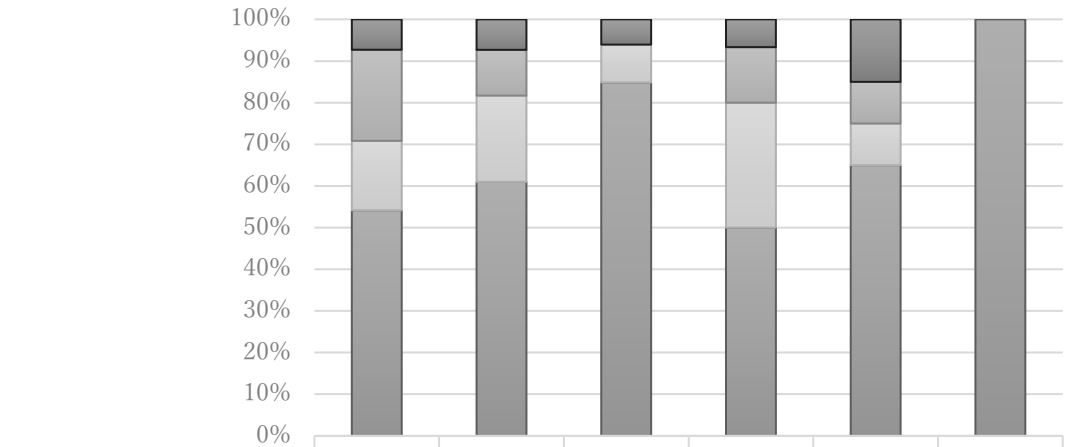


	公共商業施設	自宅	職場学校	その他の室内	屋外
その他	8	3	2	2	0
GoProなしフルセット	16	4	5	4	2
GoPro1台のみフルセット	19	5	7	1	2
フルセット	59	28	16	23	2

図 5　基本収録の収録場所別使用機材

４．３ 活動と使用された機材

次に，活動別の機材使用状況を，図 6 に示す。食事をしながらレジャー活動（付き合いを含む）を行うような忘年会や友人との会食などは「食事」「レジャー活動」の両方に分類しており，1 つの収録に複数の活動が含まれることがあるため，総数は収録合計と一致していない。なお，活動の分類には「移動」を設けているが，ウェラブルカメラ使用の収録を対象とするため，図 6 には含まれていない。



	レジャー活動	食事	休息	仕事学業	社会参加	家事雑事
その他	7	6	2	2	3	0
GoProなしフルセット	21	9	0	4	2	0
GoPro1台のみフルセット	16	17	3	9	2	0
フルセット	52	50	28	15	13	9

図 6　基本収録の活動別使用機材

活動別の機材の使用状況をみると，「家事雑事」と「休息」ではフルセットでの収録の比率が高いことがわかる。自宅など落ち着いて準備ができる状況での収録であるためと考えられる。「食事」，「レジャー活動」，「仕事学業」では，GoPro1 台や GoPro なしでの収録が増える。図 3 と図 4 で示したように GoPro1 台で会話者が十分に映る場合や，外食でスペースが狭く置き場所を確保できない場合などが考えられる。「社会参加」は，PTA や地域懇談会，祭りなどの打ち合わせで，スペースの制約で GoPro が置けないケースや，IC レコーダーより多い人数が参加して数が一致しないために「その他」に計上されている場合などがある。

４．４　収録調査において発生した問題

これまで述べたように，協力者自身が収録を行う必要があるため，収録の手続きや各種書類の書式に至るまでかなりの検討を加えてきた。その結果，これまでに大きなトラブルは発生していないが，当初想定していなかった事象も発生しているため，ここに代表的な事例と対応を報告する。こうした事例は，今後同じような収録調査を試みる研究者にとって有益な情報となるだろう。

４．４．１　機材・データについて

■カメラの電池切れ

上述のように，電池が切れてしまったために機材をフルセット使用できなかったケースが比較的多く発生した。特に SP360 は，当初，高解像度 2880×2880 で収録していたが，バッテリーの持続時間が 1 時間程度と短く，収録途中で電池切れとなることがたびたびあった。そのため，7 人目の調査から解像度の設定を 1440×1440 に下げ，バッテリーの持続時間を長くすることによって，収録途中の電池切れをできるだけ防ぐこととした。

■カメラでの静止画の撮影

カメラ（特に GoPro）のデータが動画ではなく静止画となっていたケースが見られた。これらのカメラは，電源ボタンがモード切替ボタンもかねていて，動画や静止画が切り替えられるため，誤って静止画モードで撮影してしまうことがある。動画での録画が開始されたことを，録画中に点滅するランプやカウンターの数字の上昇によって必ず確認するよう，マニュアルに記載したうえで，調査開始時の打ち合わせで強調している。また，調査期間中，数回に分けてデータを提出してもらうため，随時注意を促すことによって同様の問題はなくなっていく。

■IC レコーダー

IC レコーダーは身に着けるため，誤操作（録音レベルの変更など）が起こり得る。そこで協力者には，録音開始後，会話者に配布する前に「ホールド」（誤操作防止状態）の設定にするよう求めている。

また，中央に配置する IC レコーダーは，会話者全員の音声を収録するためのものだが，上部のマイク部分がテレビや居酒屋等での他の客の側を向いている場合，その音声が大きく録音されてしまうことがあり，向きに注意するよう喚起している。

４．４．２　書類について

３．２で述べたように，会話者への調査収録についての同意書とメタ情報収集のためのフェイスシートの記入を，協力者に依頼してもらっているが，指定欄以外への署名や記入もれなどが少なからず生じた。特にレストランなどでの収録の際は，あわただしい状況で，機材の設置を行いながら，調査についての説明と書類記入の依頼をすることになるため，十分な確認ができないものと考えられる。そこで，様式を入念に検討し，例えば，同意書の署名欄を表面に，後日撤回する場合の署名欄を 2 つ折りにした内側の面にするなど変更した結果，これらの問題はかなり解消された。

５．　まとめと今後の予定

本稿では，『日本語日常会話コーパス』構築のために現在行っている個人密着法に基づく収録の概要と現段階での収録状況について報告した。調査者が介在しない状況で，一般の人に，複数の機材を用いた収録や，同意書やフェイスシートの取得など，かなり複雑な作業をお願いしている。しかし，機器の操作が不得意な人も含め，多少の問題はあるものの，何とか無事に調査を終えている。今後，現在調査進行中の 6 名は 2 月から 4 月の間には調査が終了し，2017 年度には 16～18 名の収録調査を依頼する予定である。

謝辞

本研究は国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し

言葉の多角的研究」の研究成果を報告したものである。コーパスの収録にご協力・ご参加くださった皆さまに感謝します。

文献

Burnard, Lou, and Guy Aston (1998). The BNC Handbook. Edinburgh: Edinburgh University Press. (北村裕(監訳) (2004). 『The BNC Handbook: コーパス言語学への誘い』松柏社,)

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017)「『日本語日常会話コーパス」の構築」『言語処理学会第 23 回年次大会（NLP2017）予稿集』

小磯花絵・土屋智行・渡部涼子・横森大輔・相澤正夫・伝康晴 (2016) 「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」『国立国語研究所論集』10, pp.85–106

田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017)「『日本語日常会話コーパス』構築における会話収録方法」『言語処理学会第 23 回年次大会（NLP2017）予稿集』

『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消

鈴木 類 (茨城大学工学部情報工学科) *
古宮 嘉那子 (茨城大学工学部情報工学科)†
浅原 正幸（国立国語研究所コーパス開発センター）‡
佐々木 稔（茨城大学工学部情報工学科）§
新納　浩幸（茨城大学工学部情報工学科）¶

All-words Word Sense Disambiguation using Word Embeddings of synonym in Word List by Semantic Principles

Rui Suzuki (Department of Computer and Information Sciences, Ibaraki University)
Kanako Komiya (Department of Computer and Information Sciences, Ibaraki University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Minoru Sasaki (Department of Computer and Information Sciences, Ibaraki University)
Hiroyuki Shinnou (Department of Computer and Information Sciences, Ibaraki University)

要旨

all-words の語義曖昧性解消とは，文章中の全多義語の語義を一意に決定するタスクである．単語の語義はその周辺の文脈によって決まることから，周辺の単語同士が類似している場合その中心にある語義曖昧性解消の対象単語同士の語義も類似していると考えられる．そこで本研究では，単語の分散表現を用いて対象単語の周辺単語群と対象単語の各語義候補における類義語の周辺単語群の間の距離を測り，その距離を用いて対象単語の語義を予測した．そして，単義語の語義と多義語の予測で得た語義を基にして『分類語彙表』の概念（語義）の分散表現を作成し，“ 単語の分散表現＋概念の分散表現”を用いて周辺単語群間の距離を測りなおして再び語義を予測し，さらにこれを繰り返し行った．

『現代日本語書き言葉均衡コーパス』に『分類語彙表』のコードが付与されたコーパスを用いて実験を行ったところ，単語の分散表現のみを用いた予測では 54.2%，単語と概念の分散表現を用いた予測では最大で 59.0% の正解率となった．

* 13t4039t@vc.ibaraki.ac.jp
† kanako.komiya.nlp@vc.ibaraki.ac.jp
‡ masayu-a@ninjal.ac.jp
§ minoru.sasaki.01
¶ hiroyuki.shinnou.0828

1. はじめに

語義曖昧性解消とは，文章中の単語の語義を一意に決定するタスクである．英語のみならず，日本語を対象とした語義曖昧性解消の研究が，長年盛んに行われてきた．しかし，文中の全単語の語義を教師なし学習を用いて曖昧性解消する all-words 語義曖昧性解消の研究は，日本語においてあまり研究例がない．

本研究は，新しく作成されている途中の『現代日本語書き言葉均衡コーパス』に『分類語彙表』のコードが付与されたコー パス [3] を用いて，日本語を対象とした all-words 語義曖昧性解消を行う．

2. 関連研究

語義曖昧性解消の手法は大きく教師ありと教師なしの二つに分けることができる．一般的に，語義曖昧性解消を教師あり学習によって解決する場合，高い精度を得ることができる．しかしその反面，十分な量の教師データが必要であるためその作成にコストがかかってしまうという問題点がある．一方教師なしの場合，コストは少ないが教師あり学習を用いる場合と同等の精度を出すことは難しい．

語義曖昧性解消に『分類語彙表』を利用する手法は数多く提案されている．中でも，教師あり学習での語義曖昧性解消において，『分類語彙表』のコードや『分類語彙表』から得られる上位概念の単語などを素性として利用することは多い．教師なし学習における all-words の日本語語義曖昧性解消の関連研究には Komiya らの研究 [1] や新納らの研究 [2] がある。Komiya らの研究では、多義語の周辺に現れる語義の分布を利用する周辺語義モデルを提案している。また、新納らの研究では、単語分割をするテキスト解析のツールキットを応用し、all-words の日本語語義曖昧性解消を簡易に行えるシステムを提案している。

3. 『分類語彙表』の類義語を利用した all-words 語義曖昧性解消

本章では，本研究の手法について説明する．

3.1 単語の分散表現を利用した手法

単語の語義は周辺の単語によって決まることから，周辺の単語同士が類似している場合，その中心にある単語同士の語義も類似している，と考えることができる．本実験の手法はこの考えをもとに以下のような手順で行う．

まず，対象単語の周辺の四つの単語（前後 2 単語ずつ）のそれぞれの単語の分散表現（word2vec：以下 w2v）を求める．そしてこの四つの分散表現を連結し，一つの分散表現にしたものを「周辺単語ベクトル」とする．次に，『分類語彙表』から対象単語の語義候補ごとに類義語を列挙し，コーパス中に出現する類義語から周辺単語ベクトルを作成する．この際，周辺単語ベクトルには類義語の語義（語義候補の語義）をラベル付けしておく．最後に，対象単語と類義語の周辺単語ベクトル間の距離を測り，K 近傍法（K-NN）によって対象単語の周辺単語ベクトルと距離が近い周辺単語ベクトルのラベルを一つ求め，これを対象単語の語義と予測

する．

3.2　単語と概念の分散表現を利用した手法

本研究では，3.1 の手法での結果をもとに，さらに多義語の語義の予測を繰り返し行った．本手法の繰り返し回数が n の場合の手順を以下に示す．

まず，n-1 回目の予測で得た結果を基にコーパスを概念（分類番号）の分かち書きに変換し，word2vec で概念の分散表現（concept2vec：以下 c2v）を作成する．（繰り返し回数 0 の結果には 3.1 の手法で得た結果を利用する）次に，対象単語，対象単語の類義語の周辺単語ベクトルを作成する．この際，w2v と c2v を連結したものを単語ベクトルとし，周辺の四つの単語の単語ベクトルを連結したものを周辺単語ベクトルとする．最後に，対象単語と類義語の周辺単語ベクトルの距離を測り，3.1 の手法と同様に語義を予測する．本手法では，この操作を何度も繰り返し，精度がどこまで上昇するかを調べた．

4.　実験

4.1　『分類語彙表』の類義語

本研究の手法では『分類語彙表』から類義語を求め，利用する．ここでは本研究での類義語の定義を述べる．『分類語彙表』とは，「語を意味によって分類・整理したシソーラス (類義語集) 」である(1)．『分類語彙表』の項目は，「レコード ID 番号／見出し番号／レコード種別／類／部門／中項目／分類項目／分類番号／段落番号／小段落番号／語番号／見出し／見出し本体／読み／逆読み」となっている．『分類語彙表』では単語が分類番号によって単分類されており，この分類番号は単語の「類・部門・中項目・分類項目」を表したものである．例えば「犬」という単語は『分類語彙表』では 2 か所に存在し，分類番号はそれぞれ 1.2410, 1.5501 である（表 1）．また，『分類語彙表』には“意味的区切り”が 240 箇所に存在し，分類番号で分類された単語をさらに細かく分類している．

表 1　分類語彙表の「犬」

類	部門	中項目	分類項目	分類番号
体	主体	成員	専門的・技術的職業	1.2410
体	自然	動物	哺乳類	1.5501

本研究の 3.1 の手法では，類義語を以下のように定義した．

- 対象単語の語義（分類番号）候補と分類番号が等しく，意味的区切りがある場合は同じ区切り内の単語
- 対象単語の語義候補ごとに類義語が重複した場合，その類義語はどちらの語義の類義語からも除外する（例えば対象単語 X の語義 1 における類義語候補が A・B・C，語義 2 の類義語候補が C・D・E だった場合，どちらからも C を除外する．）

3.2 の手法において，繰り返し回数が n 回目の場合の類義語を次のように定義した．

(1) https://www.ninjal.ac.jp/publication/catalogue/goihyo/

- n-1 回目の予測において，対象単語の語義候補と等しい分類番号と予測された多義語
- 対象単語の語義候補と等しい分類番号を持つ単義語
- 対象単語の語義候補ごとに類義語が重複した場合，その類義語はどちらの語義の類義語からも除外する

4.2 実験設定

実験には『現代日本語書き言葉均衡コーパス』に『分類語彙表』のコードが付与されたコーパス [3] を用いる．このコーパスは，単語のべ数：20021, 単語異なり数：3488 からなるもので，この中に多義語は 4378 単語（のべ数）存在する．多義語の平均語義数は 3.195 であり，ランダムに語義を割り当てた場合の正解率は 31.3% である．本実験ではこの正解率を baseline とする．

単語の分散表現の作成には『国語研日本語ウェブコーパス（NWJC）』に対して word2vec というツール(2)でアルゴリズムには Continuous Bag-of-Words(C-BoW) を利用し，次元数を 200，ウィンドウ幅を 8，ネガティブサンプリングに使用する単語数を 25，反復回数を 15，として学習を行ったベクトルファイル（NWJC2vec [4]）を使用した．概念の分散表現は，コーパスを分類番号の分かち書きに変換したものを同じく word2vec で学習して作成したベクトルファイルを用いた．その際，アルゴリズムは C-BoW を利用し ，次元数を 50，ウィンドウ幅を 5，ネガティブサンプリングに使用する単語数を 5，反復回数を 3，min-count を 1，として学習を行った．

また，周辺単語ベクトルを作成する際，周辺に単語が四つない場合（対象単語が文頭や文末にある場合など）や，word2vec で学習されていない単語の分散表現などは，同じ次元の零行列を用いた．したがって，w2v のみで作成した周辺単語ベクトルは 800 次元，w2v+c2v で作成した周辺単語ベクトルは 1000 次元となる．

周辺単語ベクトルの距離を測り K-NN で分類する過程には scikit-learn(3)の KNeighborsClassifier を使用した．ここではユークリッド距離を使用し，k=1,3,5，weight=uniform, distance（uniform=重みなし, distance=重みあり）で実験を行った．

4.3 実験結果

w2v のみを用いた手法での結果を表 2 に示す．

表 2　w2v を用いた手法の結果

	K=1	K=3	K=5
重みなし	54.0	**54.2**	52.0
重みあり	54.0	52.4	51.8

次に，c2v を組み合わせ繰り返し予測した結果を表 3 に示す．

w2v を利用した予測では常に重みなしでよい精度が得られ，K=3, 重みなしでの 54.2% が最

(2) https://code.google.com/archive/p/word2vec/
(3) http://scikit-learn.org/stable/

表 3　w2v+c2v を用いた手法の繰り返し回数と正解率

K	重み	1	2	3	4	5	6
1	-	54.1	57.0	56.8	55.2	55.2	55.1
3	なし	55.3	53.0	53.6	53.3	53.2	53.2
3	あり	**59.0**	56.0	56.3	55.5	56.8	56.9
5	なし	55.1	53.0	52.4	52.5	51.7	51.9
5	あり	56.7	55.7	53.4	53.4	54.5	55.5

(一段目の数字は繰り返し回数を表す)

大となった．また，K の値，重みのあり・なしに関わらず，常に 50% 上回る，baseline を有意に超す結果となった．c2v を用いた繰り返しの予測では，繰り返し 1 回目，2 回目でさらによい精度が得られることが確認でき，K=3, 重みなり，繰り返し回数 1，の時の 59.0% が最大となった．

4.4　考察

本実験の結果により，w2v のみを利用した予測で得た結果を基に作成した w2v+c2v が有効であることが確認できた．

本手法では w2v のみを利用した予測の精度が w2v+c2v の質を左右する．最初の予測の精度を向上させる工夫の一つとして，類義語をどのように定義するか，という点が挙げられる．前述の実験での類義語の定義は，「分類番号が同じ（意味的区切りを考慮），語義候補間で重複した類義語を除外」である．語義候補間で類義語が重複した場合，まったく同じ周辺単語ベクトルに異なるラベルが付与されたものが作成されてしまい，K-NN での分類の精度が低下してしまうと考え，除外した．

本手法で距離計算に用いる類義語には，現在の条件に ① 意味的区切りを考慮しない ②意味的区切りではなく段落番号を考慮する ③『分類語彙表』における単義語のみ使用する，という条件を追加するなど，様々なバリエーションが考えられる．そこで， ① ② ③ を実際に一つずつ類義語の定義に追加して実験し，その結果から本手法の改善点などを考察する．

①を追加した結果，w2v，w2v+c2v のどちらの手法においても精度は低下した． ①を類義語の定義に追加した場合，前述の実験と比較して，対象単語の一つの語義に対してより多くの類義語が得られることになる．このため，意味が遠い単語がより多く類義語に含まれてしまい精度が低下したものと考えられる．

②を追加した場合も、どちらの手法でも精度は低下した。段落番号は意味的区切りよりも細かく単語を分類しているため， ②を類義語の定義に追加した場合対象単語により近い意味の単語を類義語として扱えることになる．しかし類義語の数は減るため，K-NN での分類の際のデータ量が減り，精度が低下したのだと考えられる． ①・ ②を類義語の定義に追加した結果から，類義語の数は多くしても少なくしても精度が低下してしまうことがわかった．また，類義語の意味の幅と類義語の量の最適なバランスはコーパスの大きさや対象単語によって左右されるものであると考えられる．

一方で ③を類義語の定義に追加した場合，どちらの手法でも精度が向上した（表 4・表 5）．

表 4　w2v を用いた手法の結果（ ③を類義語の定義に追加した場合）

	K=1	K=3	K=5
重みなし	56.3	54.3	54.8
重みあり	56.3	56.6	**56.7**

表 5　w2v+c2v を用いた手法の繰り返し回数と正解率（ ③を類義語の定義に追加した場合）

K	重み	1	2	3	4	5	6
1	-	57.0	56.3	56.9	57.1	56.9	57.1
3	なし	56.6	57.1	57.8	57.4	57.4	57.4
3	あり	58.9	59.2	**59.6**	59.3	59.3	59.3
5	なし	57.6	57.7	57.7	57.7	57.7	57.7
5	あり	59.4	59.1	59.2	59.1	59.2	59.1

前述の w2v のみを利用する実験では類義語が多義語・単義語にかかわらず類義語として利用していた． ③を類義語の定義に加えることで，類義語の周辺単語ベクトルの質が向上したものと考えられる．また，『分類語彙表』について調べた結果，掲載されている単語の 8 割近くは単義語であることがわかった．このため ③を類義語の定義に加えたとしてもそれほど類義語の数が減少せず，精度が向上したのだと推測できる．

5. おわりに

本稿では，対象単語の周辺単語ベクトルと対象単語の類義語の周辺単語ベクトルの距離から対象単語の語義を求める語義曖昧性解消の手法を提案した．また，語義を予測した結果をもとに概念の分散表現を作成し，周辺単語ベクトルに組み合わせ再び語義を求める，という操作を繰り返し行う実験も行った．実験の結果，baseline を超える精度を達成することができ，語義曖昧性解消において有効な手法であることが確認できた．また，概念の分散表現を組み合わせることでさらに精度が上がったが，何度も繰り返すことの効果はなく，精度が上がったのは繰り返し 1，2 回目のみだった．

『分類語彙表』の中のどのような条件の単語を本手法の距離計算に用いるのが最適であるかは，対象単語の語義やコーパスの大きさによって異なると考えられるが，本実験では『分類語彙表』中の単義語のみを使用することで精度を向上させることができた．

謝　辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-words WSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」 の研究成果を報告したものである。

参考文献

[1] Kanako KOMIYA, Yuto SASAKI, Hajime MORITA, Minoru SASAKI, Hiroyuki SHINNOU, and Yoshiyuki KOTANI, Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation, PACLIC 2015, pp. 35-43, (2015.10).

[2] 新納浩幸, 古宮嘉那子, 佐々木稔, 森信介，点推定による日本語 all-words WSD システム KyWSD，情報処理学会　研究報告自然言語処理（NLP），Vol.2016-NL-227 No.2，pp.1-5，(2016,07,29).

[3] 加藤祥，浅原正幸，山崎誠，『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行，言語処理学会第 23 回年次大会発表論文集（掲載予定）

[4] 浅原正幸，岡照晃，nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ，言語処理学会第 23 回年次大会発表論文集（掲載予定）

形態素解析ソフトウェア『Web 茶まめ』の改良と Web API の試作

川口寛治(東京電機大学)
薦田龍輝（東京電機大学）
堤智昭（東京電機大学）

Improvement of “Web ChaMame” and experimental production of “Web ChaMame Web API”

Motoharu Kawaguchi (Tokyo Denki University)
Ryuki Komoda (Tokyo Denki University)
Tomoaki Tsutsumi (Tokyo Denki University)

要旨

国立国語研究所では，様々な時代の日本語資料の分析に利用可能な形態素解析辞書である UniDic を公開している．この UniDic を利用した形態素解析支援アプリケーションである『Web 茶まめ』は Web 上で公開されており，インターネットを通じて誰でも利用できる．本稿では，『Web 茶まめ』について以下の二点を報告する．一点目は，『Web 茶まめ』を公開した以降にユーザから寄せられた意見や指摘をもとに行った改良についてである．二点目は，ブラウザを用いずにインターネットを通じて Web 茶まめの機能を利用するための，WebAPI の試作についてである．

１．はじめに

国立国語研究所では，UniDic[1] [2] [3]を用いた形態素解析の支援を目的とした Web アプリケーションである Web 茶まめ[4]を公開している．Web 茶まめは Web 上で公開されており，インターネットを通じて誰でもアクセス，利用が可能である．

本稿では，Web 茶まめの解析機能の改良と，WebAPI の試作について報告する．解析機能の改良は，ユーザの使い勝手向上を目的とし，公開後にユーザから寄せられた意見を元に行った．WebAPI の試作については，Web 茶まめの機能を，ブラウザを介さず，他のシステムから利用し易い形で提供することを目的に行った．WebAPI では HTTP 通信を用いてデータを送受信することで，Web 茶まめの機能を提供する．

２．Web 茶まめ概要

Web 茶まめは，図１に示すようにサーバ内の形態素解析エンジン MeCab[5]と，９種類の UniDic と１種類の IPAdic を用いて形態素解析を行い，その結果を Web ブラウザ上に表示する形態素解析支援アプリケーションである．形態素解析処理はすべて Web 茶まめサーバ内で行われるため，特別なソフトウェアのインストールや特別な環境を用意することなく形態素解析を行う事ができる．また，Web 茶まめに新しい機能を実装したり動作の変更を行う場合には，ユーザの環境を変更する必要がなく，サーバにインストールされたソフトウェアを更新することで実現できる．

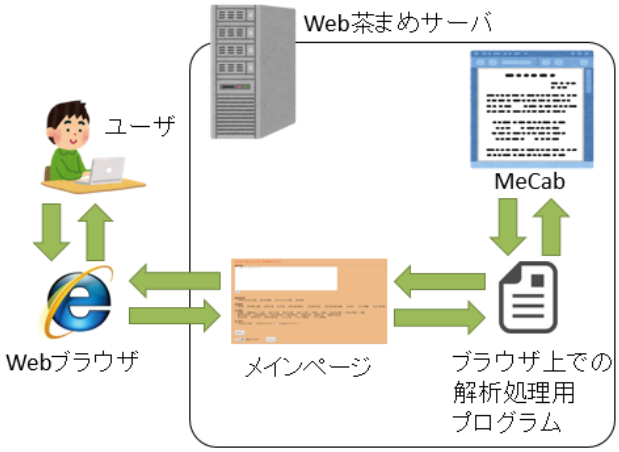


図 1　Web 茶まめ概要図

３．Web 茶まめの改良

３.１　品詞項目の細分化

従来の Web 茶まめにおいて，品詞は 1 つの出力項目として処理されていた．しかし，Web 茶まめが用いる形態素解析辞書である UniDic では，辞書内の項目として品詞の分類が，大分類，中分類，小分類，細分類の 4 つに分かれている．そこで，主によく使われる，大分類，中分類，小分類を出力項目として追加し，個別に出力するか否かを切り替えられるように変更した．

３.２　出力文字コードの選択機能の追加

Web 茶まめでは，形態素解析の結果を HTML，CSV，EXCEL の 3 形式で出力することが可能である．そのうち，CSV 出力時には，文字コードとして UTF-8 を用いていた．しかし，Windows に代表される一部の OS では ShiftJIS が使われることが多く，ユーザは出力データの文字コードを，必要に応じてその都度自前で変更する必要があった．そこで，CSV 出力を選択した場合には，出力データの文字コードを UTF-8 と ShiftJIS の 2 種類から選択可能とした．

３.３　IPAdic を用いた解析の改良

Web 茶まめでは UniDic の他に，IPAdic を用いた形態素解析が可能である．しかし，IPAdic は UniDic とは辞書内の項目数，及び項目名が異なる．そのため，UniDic と同様の項目を指定し形態素解析を行った場合，図 2 に示すように項目の違いに依存したエラーが発生してしまう．この問題を解決するためには，IPAdic の項目に準拠した出力項目の指定が必要である．

そこで，IPAdic 選択時に MeCab に送る命令文のフォーマットを IPAdic に準拠したものに変更し，IPAdic に対応した項目名と出力結果が表示されるように修正を行った．この時，IPAdic の出力項目は，IPAdic が持つ 9 つの項目を固定で出力することとした．修正後の Web 茶まめで IPAdic を用いた形態素解析を行った結果を図 3 に示す．

また，この修正に伴い，Web 茶まめの辞書・出力項目選択の GUI について，UniDic と IPAdic 利用時で出力項目が異なるために起きる，ユーザの混乱を避けるための変更を行った．具体的には図 4 に示すとおり，IPAdic を選択した場合には，出力項目チェックボックスを全て外した状態にし，チェックができないようにした．

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形
IPAdic(現代語)	I	given index is out of range						

図 2　不具合修正前の IPAdic 解析結果

辞書	文境界	書字形(＝表層形)	読み	発音	原型	品詞	品詞細分類1	品詞細分類2	品詞細分類3	活用型	活用形
IPAdic(現代語)	B	庭	ニワ	ニワ	庭	名詞	一般				
IPAdic(現代語)	I	に	ニ	ニ	に	助詞	格助詞	一般			
IPAdic(現代語)	I	は	ハ	ワ	は	助詞	係助詞				
IPAdic(現代語)	I	ニ	ニ	ニ	ニ	名詞	数				
IPAdic(現代語)	I	羽	ワ	ワ	羽	名詞	接尾	助数詞			
IPAdic(現代語)	I	鶏	ニワトリ	ニワトリ	鶏	名詞	一般				
IPAdic(現代語)	I	が	ガ	ガ	が	助詞	格助詞	一般			
IPAdic(現代語)	I	いる	イル	イル	いる	動詞	自立			一段	基本形
IPAdic(現代語)	I	。	。	。	。	記号	句点				

図 3　不具合修正後の IPAdic 解析結果

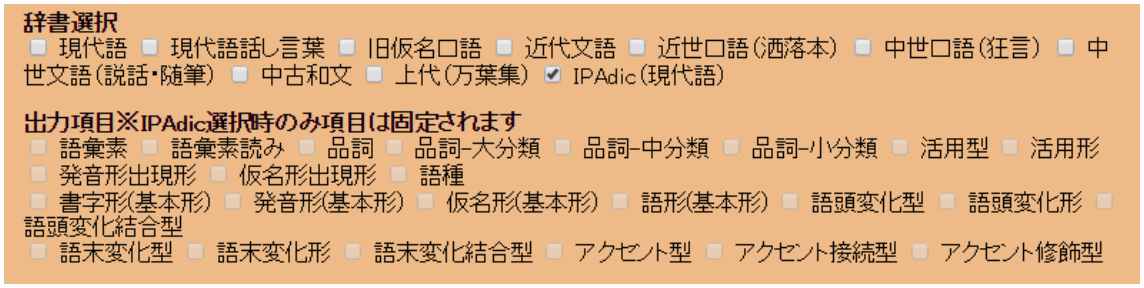


図 4　チェックボックス GUI の変更

３.４　2 辞書比較の解析の改良

Web 茶まめでは，辞書を 2 つまで選択し形態素解析を行い，解析結果を比較して見ることができる．この 2 辞書比較時の機能として 2 つの機能が存在する．1 つ目は，解析結果に差異が生じた行の文字を赤色に変更し発見しやすくする「色変更機能」である．2 つ目は，辞書毎に形態素が異なり出力結果の行がずれる場合に，行ずれを防ぐ為に空白行を挿入する「行ずれ修正機能」である．

しかし，これらの機能には，いくつかの不具合が発生していた．具体的な不具合は以下の 4 点である．実際に不具合が起こっている場合の例を図 5～7 に示す．

(1)先頭行に必ず色変更機能が実行される．
(2)差異が生じた行ではない場合でも，色変更機能が実行される場合がある．
(3)英数字に対し，行ずれ修正機能が実行されない場合がある．
(4)同一行で内容に差分があるときに色変更機能が実行されない場合がある．

Web 茶まめのプログラムを修正することにより，上記 4 点の不具合の修正を行った．その結果図 8～10 に示すように，正常に 2 辞書比較時の機能が実行されるようになった．

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現形	語種	書字形(基本形)	発音形(基本形)	仮名形(基本形)	語形(基本形)	←→	辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現形	語種	書字形(基本形)	発音形(基本形)	仮名形(基本形)	語形(基本形)
現代語	B	庭	庭	ニワ	名詞-普通名詞-一般			ニワ	ニワ	和	庭	ニワ	ニワ	ニワ	←→	現代話し言葉	B	庭	庭	ニワ	名詞-普通名詞-一般			ニワ	ニワ	和	庭	ニワ	ニワ	ニワ
現代語	I	に	に	ニ	助詞-格助詞			ニ	ニ	和	に	ニ	ニ	ニ	←→	現代話し言葉	I	に	に	ニ	助詞-格助詞			ニ	ニ	和	に	ニ	ニ	ニ

図 5　(1)の不具合例

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	←→	辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞
現代語	B	「	「		補助記号-括弧開	←→	近世口語(洒落本)	B	「	「		補助記号-括弧開
現代語	I	ワールド	ワールド-world	ワールド	名詞-普通名詞-一般	←→	近世口語(洒落本)	I	ワールドカップ			感動詞
現代語	I	カップ	カップ-cup	カップ	名詞-普通名詞-一般	←→	近世口語(洒落本)	I				
現代語	I	」	」		補助記号-括弧閉	←→	近世口語(洒落本)	I	」	」		補助記号-括弧閉

図 6　(2)の不具合例

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形現形	仮名形現形	語種	書字(基本形)	発音(基本形)	仮名(基本形)	語形(基本形)	←→	辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形現形	仮名形現形	語種	書字(基本形)	発音(基本形)	仮名(基本形)	語形(基本形)
現代語	B	2	二	ニ	名詞-数詞			ニ	ニ	漢	2	ニ	ニ	ニ	←→	近世口語(洒落本)	B	2017			名詞									
現代語	I	0	ゼロ-zero	ゼロ	名詞-数詞			ゼロ	ゼロ	外	0	ゼロ	ゼロ	ゼロ	←→	近世口語(洒落本)	I	年	年	ネン	名詞-普通名詞-数詞可能助動			ネン	ネン	漢	年	ネン	ネン	ネン
現代語	I	1	一	イチ	名詞-数詞			イチ	イチ	漢	1	イチ	イチ	イチ	←→	近世口語(洒落本)	I	カレンダー			名詞									
現代語	I	7	七	ナナ	名詞-数詞			ナナ	ナナ	和	7	ナナ	ナナ	ナナ	←→	近世口語(洒落本)	I	。	。		補助記号-句点					記号	。			
現代語	I	年	年	ネン	名詞-普通名詞-助数詞可能			ネン	ネン	漢	年	ネン	ネン	ネン	←→	近世口語(洒落本)	I	EOS												
現代語	I	カレンダー	カレンダー-calendar	カレンダー	名詞-普通名詞-一般			カレンダー	カレンダー	外	カレンダー	カレンダー	カレンダー	カレンダー	←→	近世口語(洒落本)														

図 7　(3)(4)の不具合例

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形現形	仮名形現形	語種	書字(基本形)	発音(基本形)	仮名(基本形)	語形(基本形)	←→	辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形現形	仮名形現形	語種	書字(基本形)	発音(基本形)	仮名(基本形)	語形(基本形)
現代語	B	庭	庭	ニワ	名詞-普通名詞-一般			ニワ	ニワ	和	庭	ニワ	ニワ	ニワ	←→	代話し言葉現語	B	庭	庭	ニワ	名詞-普通名詞-一般			ニワ	ニワ	和	庭	ニワ	ニワ	ニワ
現代語	I	に	に	ニ	助詞-格助詞			ニ	ニ	和	に	ニ	ニ	ニ	←→	代話し言葉現語	I	に	に	ニ	助詞-格助詞			ニ	ニ	和	に	ニ	ニ	ニ

図 8　(1)の不具合修正後の例

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	←→	辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞
現代語	B	「	「		補助記号-括弧開	←→	近世口語(洒落本)	B	「	「		補助記号-括弧開
現代語	I	ワールド	ワールド-world	ワールド	名詞-普通名詞-一般	←→	近世口語(洒落本)	I	ワールドカップ			感動詞
現代語	I	カップ	カップ-cup	カップ	名詞-普通名詞-一般	←→	近世口語(洒落本)	I				
現代語	I	」	」		補助記号-括弧閉	←→	近世口語(洒落本)	I	」	」		補助記号-括弧閉

図 9　(2) 不具合修正後の例

辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形現形	仮名形現形	語種	書字(基本形)	発音(基本形)	仮名(基本形)	語形(基本形)	←→	辞書	文境界	書字形(＝表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形現形	仮名形現形	語種	書字(基本形)	発音(基本形)	仮名(基本形)	語形(基本形)
現代語	B	2	二	ニ	名詞-数詞			ニ	ニ	漢	2	ニ	ニ	ニ	←→	近世口語(洒落本)	B	2017			名詞									
現代語	I	0	ゼロ-zero	ゼロ	名詞-数詞			ゼロ	ゼロ	外	0	ゼロ	ゼロ	ゼロ	←→	近世口語(洒落本)	I													
現代語	I	1	一	イチ	名詞-数詞			イチ	イチ	漢	1	イチ	イチ	イチ	←→	近世口語(洒落本)	I													
現代語	I	7	七	ナナ	名詞-数詞			ナナ	ナナ	和	7	ナナ	ナナ	ナナ	←→	近世口語(洒落本)	I													
現代語	I	年	年	ネン	名詞-普通名詞-助数詞可能			ネン	ネン	漢	年	ネン	ネン	ネン	←→	近世口語(洒落本)	I	年	年	ネン	名詞-普通名詞-数詞可能助動			ネン	ネン	漢	年	ネン	ネン	ネン
現代語	I	カレンダー	カレンダー-calendar	カレンダー	名詞-普通名詞-一般			カレンダー	カレンダー	外	カレンダー	カレンダー	カレンダー	カレンダー	←→	近世口語(洒落本)	I	カレンダー			感動詞									

図 10　(3)(4) 不具合修正後の例

４．WebAPI の試作

４.１　目的

Web 茶まめは，煩雑な形態素解析環境を用意せずとも利用可能であるという特徴から，形態素解析を処理として含むシステムの一部として利用されつつある [6]．しかし，Web ブラウザ上での動作を前提とした形態素解析支援アプリケーションであり，他の Web サイトやアプリケーションが Web 茶まめの機能を利用・連携して動作するのは難しい．また，利用するシステムに適した GUI に変更することはできない，といった問題がある．

そこで，Web 茶まめの機能のみを簡単に利用可能にするための WebAPI を試作した．WebAPI の利用者は，HTTP 通信を用いて，Web 茶まめサーバに解析を行うテキストデータや出力項目，使用する辞書の情報等を送信し，形態素解析結果を受信する．これにより Web サイトやアプリケーションに Web 茶まめを用いた形態素解析機能を容易に組み込むことができ，形態素解析に必要なテキストデータの入力画面や出力方法などは，WebAPI を利用するシステムが自由に組み換え可能となる．

４.２　Web 茶まめの WebAPI 概要

今回，我々が作成した WebAPI は Web 茶まめの機能の 1 つである解析前処理を行う API（以下，解析前処理 API）と形態素解析を行う API（以下，形態素解析 API）の 2 種類である．解析前処理 API は，形態素解析を行う前に実行されるため形態素解析 API とは分ける形で作成した．

作成した 2 つの WebAPI はその名の通り，Web ブラウザ版 Web 茶まめの持つ解析前処理機能と形態素解析処理を，それぞれ利用するための WebAPI である．Web ブラウザ版 Web 茶まめとの相違点としては，以下の 3 点が挙げられる．

（１） IPAdic を利用した形態素解析には非対応である．

（２） 2 辞書を同時に選択し，解析結果を比較することはできない．

（３） 出力形式は XML，及び JSON の 2 種類である．

（１）については，Web 茶まめでは UniDic を用いた解析がメインであるため，今回は IPAdic を用いた形態素解析は非対応とした．（２）については，結果の比較や表示は WebAPI を利用するシステム側が自由に設定する事項であるため，WebAPI 側では極力処理をしない方針としたため，実装を行わなかった．（３）については，HTTP 通信を用いたデータをやり取りする場合には XML と JSON が一般的によく使われ，適切であると言われるため，これらの形式を採用することとした．

これらの WebAPI は図 11 に示すように，外部のアプリケーション等から Web 茶まめサーバ内にある解析前処理 API または，形態素解析 API にそれぞれ対応した URL にアクセスすることで処理を実行することができる．実行する際には，解析対象のテキストデータや，利用する辞書の情報，実行する解析前処理の種類等を Web サーバに伝えるために，URL にクエリパラメータを付加した URL を用いる．

４.３　解析前処理 API

４.３.１　解析前処理 API の使用方法

Web 茶まめサーバの WebAPI を利用するための URL に，解析に必要な情報を付与しアクセスすることで利用可能である．試作した WebAPI では，解析前処理 API のリクエスト URL を「http://lcm.tsu-lab.sie.dendai.ac.jp/ V1/processings」とした．この URL 以降に，WebAPI の実行に必要な情報をリクエストパラメータとして付加する．解析前処理 API のパラメータは

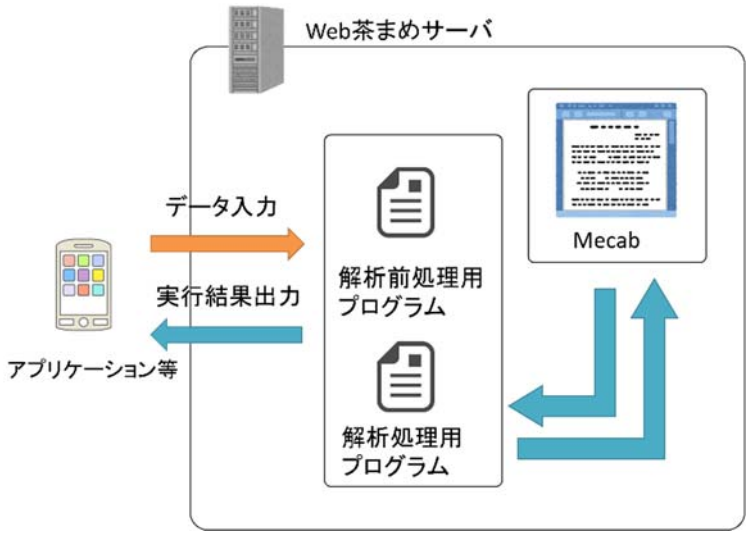


図 11　Web 茶まめの WebAPI 概要図

表 1 に示すように設定している．パラメータ名 sentence は，処理を行うテキストデータを入力する．ここでは URL で日本語を送信するためにパーセントエンコーディングを用いて入力する必要がある．パラメータ名 processings は，解析対象文章に行う前処理を選択する．前処理は複数同時に選択可能である．パラメータ名 types は出力結果を XML または JSON を入力し，出力形式を選択する．

表 1　解析前処理 API のパラメータ

パラメータ名	説明
sentence	解析前処理対象テキストを入力する．
processings	使用する出力項目を選択する．
types	出力形式を選択する．

４．３．２　解析前処理 API の出力結果

図 12 に，本 API を実行した場合の例を示す．この例では，日本語文「ｾﾞﾝｶｸ　すｽﾞ　はんてん　１２３４５」に対して，解析前処理，「半角文字を全角に変換」「 踊り字を展開 」「カタカナひらがな反転 」「数字処理」の 4 つを選択し，出力形式は XML とした．出力結果の XML タグがもつ意味は表 2 のようになっている．

```
<?xml version="1.0" standalone="true"?>
- <results>
    <zenkaku>ゼンカク すｽﾞ はんてん12345</zenkaku>
    <odorizi>ｾﾞﾝｶｸ すず はんてん12345</odorizi>
    <hanten>ぜんかく スｽﾞ ハンテン12345</hanten>
    <suuzi>ｾﾞﾝｶｸ すｽﾞ はんてん一万二千三百四十五</suuzi>
</results>
```

図 12　解析前処理 API の XML 出力結果

表 2　解析前処理 API のタグ

フィールド名	説明
results	レスポンス結果
zenkaku	半角文字を全角に変換した文
odorizi	踊り字を展開した文
hanten	ひらがなカタカナを反転した文
suuzi	数字処理をした文
error	エラー結果
message	エラーメッセージ

４.４　形態素解析 API

４.４.１　形態素解析 API の使用方法

形態素解析 API も，解析前処理 API と同様に Web 茶まめサーバの WebAPI を利用するための URL に，解析に必要な情報を付与しアクセスすることで，利用可能である．試作した WebAPI では，解析前処理 API のリクエスト URL を「http://lcm.tsu-lab.sie.dendai.ac.jp/V1/analyses」とした．

形態素解析 API のリクエストパラメータは表 3 に示すように設定している．パラメータ名 sentence は，解析前処理 API と同様にパーセントエンコーディングを用いて入力する．パラメータ名 dics は，解析で使用する辞書を 1 つ選択する．選択は半角数字 1〜9 のいずれかを入力する．それぞれに数字は，Web 茶まめが利用できる 9 つの辞書に対応しており，その順番は，Web ブラウザ版に表示されている順番である．例えば，「現代語」辞書は 1 に，「旧仮名口語」辞書は 2 に，「上代（万葉集）」辞書は 9 に対応している．パラメータ名 options_id は，出力項目を選択するためのパラメータである．選択は半角数字 1〜24 の数字を，半角カンマ（,）区切りで入力する．こちらも「語彙素」を表示したい場合は 1，「語彙素読み」を表示したい場合は 2 といったように，数字と項目は Web ブラウザ版に表示されている順番に対応している．例えば，「語彙素」「語彙素読み」「活用形」の 3 つを出力したい場合は「options_id=1,2,8」となる．パラメータ名 types は解析前処理 API と同様に XML または JSON を入力し，出力形式を選択する．

表 3　形態素解析 API のリクエストパラメータ

パラメータ名	説明
sentence	解析対象テキストを入力する．
dics	使用する辞書を 1 つ選択する．
options_id	出力項目を選択する．
types	出力形式を選択する．

４.４.２　形態素解析 API の出力結果

図 13 は，源氏物語の冒頭「いづれの御時にか，女御・更衣あまたさぶらひたまひける中に，いとやむごとなき際にはあらぬが，すぐれて時めきたまふありけり．」を解析テキストとして入力し，使用辞書に中古和文を選択，出力項目に語彙素・語彙素読み・品詞を選択，出力形式に JSON を選択した場合の出力結果である．

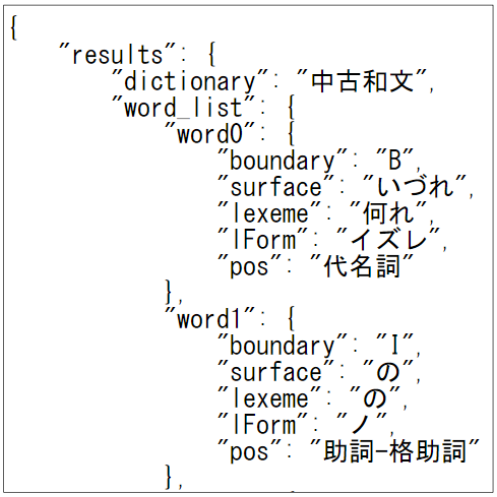
出力結果の JSON タグは，文境界と表層形，品詞-大分類，品詞-中分類，品詞-小分類，レ

スポンス結果，エラー結果とエラーメッセージ以外の出力項目名は UniDic 内部の項目名に準拠している．準拠していないタグについては，文境界は boundary，表層形は surface，品詞の大分類，中分類，小分類はそれぞれ pos1，pos2，pos3 とした．レスポンス結果，エラー結果，エラーメッセージは，解析前処理 API の XML タグと同様のタグ名である．

５．おわりに

本稿では，形態素解析ソフトウェア Web 茶まめの改良について報告を行った．ユーザから寄せられた情報を元に，不具合修正や機能の改善を行い，Web 茶まめの利便性を向上させた．また，Web 茶まめの機能を他のアプリケーションから容易に利用可能とするために，WebAPI を試作した．試作した WebAPI を用いることで，他の教育システムや形態素解析を用いた研究システムに Web 茶まめの形態素解析機能を組みこむことが可能となった．

今後の課題として，今後も寄せられるブラウザ版 Web 茶まめのユーザからの要望に答えるためにシステム改良を実施していく．また，WebAPI におけるパーセントエンコーディングを用いないデータ入力機構の実装や，WebAPI を用いた外部アプリケーションを試作し，WebAPI の機能を検証する必要があると考えている．



```
{
    "results": {
        "dictionary": "中古和文",
        "word_list": {
            "word0": {
                "boundary": "B",
                "surface": "いづれ",
                "lexeme": "何れ",
                "lForm": "イズレ",
                "pos": "代名詞"
            },
            "word1": {
                "boundary": "I",
                "surface": "の",
                "lexeme": "の",
                "lForm": "ノ",
                "pos": "助詞-格助詞"
            },
```


図 13　形態素解析 API の JSON 出力結果

参考文献

[1] 小木曽 智信，小町 守，松本 裕治（2013）「歴史的日本語資料を対象とした形態素解析」『自然言語処理』20(5)，pp. 727-748.

[2] 国立国語研究所：近代文 UniDic,(オンライン)，〈http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%B6%E1%C2%E5%CA%B8%B8%ECUniDic>(参照 2015-09-15).

[3] 国立国語研究所：中古和文 UniDic,(オンライン),〈http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%C3%E6%B8%C5%CF%C2%CA%B8UniDic〉.

[4] 堤 智昭，小木曽 智信：歴史的資料を対象とした複数の UniDic 辞書による形態素解析支援ツール『Web 茶まめ』,じんもんこん 2015 論文集,Vol.2015,NO.1,pp.179-184,(2015).

[5] 工藤 拓:MeCab：Yet Another Part-of-Speech and Morphological Analyzer, (online),入手先〈http:// mecab.sourceforge. net/〉(参照 2015-09-15)

[6] 土山 玄：絵入源氏物語のテキストデータに対する統計解析 web アプリケーションの設計,じんもんこん 2016 論文集,Vol. 2016-CH-112, NO.1, pp.1-4, (2016).

『現代日本語書き言葉均衡コーパス』を用いた「~ ていく」「~ てくる」構文の意味分析

加藤 麟太郎（東京大学教養学部 学生）†
藤井 聖子（東京大学大学院総合文化研究科）

A corpus-based semantic analysis of the *–te iku* and *-te kuru* constructions

Rintaro Kato (University of Tokyo, College of Arts and Sciences)
Seiko Fujii (University of Tokyo, Graduate School of Arts and Sciences)

要旨

『現代日本語書き言葉均衡コーパス』(以下，BCCWJ)を用いて「～ていく」「～てくる」構文の物理的移動の用法の意味分析を行った。焦点をあてた研究問題は，「～ていく」「～てくる」の前に共起する動詞の意味特性と「～ていく」「～てくる」構文の意味特性との関係は，どの程度規則的でありどの程度予測可能かという問題である。BCCWJ からの「～ていく」「～てくる」構文の無作為抽出のうち，物理的移動を表す用法をそれぞれ 497 用例，580 用例抽出し，森田(1998)に基づく仮説を立て，意味コーディングをした。コーディングに基づく定量的予備分析において，まず多くの用例がある程度予測可能な規則的傾向を示すことが明らかになったが，本稿では，左記傾向の定量的分析の報告に加えて，コーパスにみられる例外的用例の方に着目し， 例外的用例の定性的分析を示した上で，新たな意味分類を提案する。

１．はじめに

『現代日本語書き言葉均衡コーパス』(以下，BCCWJ)を用いて「～ていく」構文と「～てくる」構文の意味分析を行った。「～ていく」「～てくる」構文には，アスペクト的意味を含む様々な比喩的用法があるが，本稿では，物理的移動の用法の分析を報告する。

２．分析手法

２.１ 「～ていく」「～てくる」構文の意味分類

本稿で扱う物理的移動を表す「～ていく」「～てくる」補助動詞構文は，「いく」「くる」の意味として字義通り物理的な移動を表している。本稿で「～ていく」構文と「～てくる」構文それぞれの意味特性として着目し分類するのは，「～ていく」「～てくる」の表す物理的移動と，その「～ていく」「～てくる」の前に共起している動詞（以下，V_1）が表す事態との関係として解釈される意味である。

森田(1998)ではその関係を「動作・行為の順次性を表す」「平行して行うことを表す」「移動するときの状態を表す」「複合して一つの動作・行為・作用を表す」という特徴によって四つに分類している。以下は，森田(1998)の説明を基にそれぞれの意味に分類される基準となると考えられる特徴を示し，それぞれ名称をつけたものである。

① 順次

V_1 の表す行為が完了した後に移動が起こる。

† kato-rintaro084@g.ecc.u-tokyo.ac.jp

② 並行
V_1の表す行為と移動が同時に起こり，何かを持ってまたは伴って移動する。
③ 状態
V_1の表す行為が移動中の移動主体の状態や移動の手段を表す。
④ 一体
V_1の表す行為自体が方向性のある移動を含意するものである。

２.２ 「～ていく」「～てくる」構文の意味特性とV_1の意味との関係性に関する仮説

森田(1998)によれば， V_1の意味特性と上記①～④との関係性は以下の通りである。

(1) 他動性あり・意志性あり・アスペクト瞬間 → 順次
他動性あり・意志性あり・アスペクト継続 → 並行
他動性なし・方向性なし・アスペクト継続 → 状態
他動性なし・方向性あり・移動性あり → 一体

本分析では，上記(1)を，V_1の意味特性から「～ていく」構文・「～てくる」構文の意味特性を予測するための仮説と捉え，この予測がどの程度正確か，予測が外れる用例はどのようなものかをコーパス分析によって検証する。

２.３ V_1の意味特性，及び，「～ていく」「～てくる」構文の意味特性のアノテーション

上述の意味分析を行うにあたり，コーパスから物理的移動を表す「～ていく」「～てくる」構文が使われている用例を抽出し，それぞれの用例に，V_1の「他動性」「方向性」「意志性」「移動性」「アスペクト」に関してアノテーションを付し，さらに，人が解釈する「～ていく」構文の意味（「順次」「並行」「状態」「一体」）に関してアノテーションを付した。

２.４ コーパス母集団と抽出用例データ

BCCWJ からの検索により無作為に「～ていく」，「～てくる」の用例を抽出した上で，人の意味解釈による人手で物理的移動を表す用例（500 近似値）を抽出した。「ていく」は，検索で得られた 3500 用例のうち， 物理的移動を表す「～ていく」補助動詞構文でないものを除き 497 用例を，「～てくる」は検索で得られた 2000 用例のうち， 物理的移動を表す「～ていく」補助動詞構文でないものを除き 580 用例を抽出し，分析対象とした。

なお，「～ていく」，「～てくる」構文の無作為検索のコーパス母集団は，BCCWJ 中の以下コーパスである。書籍（1971～2005 年，22，058 件，約 6，270 万語），雑誌（2001～2005 年，1，996 件，約 440 万語），新聞（2001～2005 年，1，473 件，約 140 万語），白書（1976～2005 年，1，500 件，約 490 万語），教科書（2005～2007 年，412 件，約 90 万語），広報紙（2008 年，354 件，約 380 万語），Yahoo!知恵袋（2005 年，91，445 件，約 1，030 万語），Yahoo!ブログ（2008 年，52，680 件，約 1，020 万語），韻文（1980～2005 年，252 件，約 20 万語），法律（1976～2005 年，346 件，約 110 万語），国会会議録（1976～2005 年，159 件，約 510 万語）。

３．定量的予備分析の結果

ここでは，森田(1998)の示した V_1 の意味特性と「～ていく」「～てくる」構文の意味との関係に実際の用例はどの程度当てはまり，どの程度規則的かということを調べる。V_1 の他動性・方向性・意志性・移動性・アスペクトというそれぞれの意味特性と「～ていく」「～てくる」構文の意味特性との関係を集計し，それらについてカイ二乗検定及び残差分析を行った。

３.１　他動性

表１　V_1 の他動性と「～ていく」「～てくる」構文の意味特性との関係

	いく				くる				
	順次	並行	状態	一体	順次	並行	状態	一体	その他[1]
他動性あり	22	109	1	5	48	41		19	71
他動性なし	3		91	266	20		44	336	1

V_1 の意味特性と「～ていく」「～てくる」構文の意味特性について，森田(1998)の示した(1)の関係性が成り立つとすれば，他動性があるものは順次及び並行に，ないものは状態及び一体に分類されるはずである。「～ていく」のクロス表についてカイ二乗検定を行った結果，主効果が見られた(χ^2=454.245, df=3, p< .01)。さらに，残差分析の結果，順次及び並行で他動性ありが期待度数よりも有意に多く，状態及び一体で他動性なしが期待度数よりも有意に多いことが分かった(p< .01)。

「～てくる」のクロス表についてその他部分を除外してカイ二乗検定を行った結果，主効果が見られた(χ^2=316.240, df=3, p< .01)。さらに，残差分析の結果，順次及び並行で他動性ありが期待度数よりも有意に多く，状態及び一体で他動性なしが期待度数よりも有意に多いことが分かった(p< .01)。

３.２　方向性

表２　V_1 の方向性と「～ていく」「～てくる」構文の意味特性の関係

	いく				くる				
	順次	並行	状態	一体	順次	並行	状態	一体	その他
方向性あり	1	9		268	2	11		355	71
方向性なし	24	100	92	3	66	30	44		1

方向性については，(1)の関係性が成り立つとすれば，状態に分類されるものは方向性がなく，一体に分類されるものは方向性があるはずである。「～ていく」のクロス表についてカイ二乗検定を行った結果，主効果が見られた(χ^2=447.569, df=3, p< .01)。さらに，残差分析の結果，順次及び並行及び状態で方向性なしが期待度数よりも有意に多く，一体で方向性ありが期待度数よりも有意に多いことが分かった(p< .01)。

「～てくる」のクロス表についてその他部分を除外してカイ二乗検定を行った結果，主効果が見られた(χ^2=457.960, df=3, p< .01)。さらに，残差分析の結果，順次及び並行及び状態で方向性なしが期待度数よりも有意に多く一体では方向性ありが期待度数よりも有意に多いことが分かった(p< .01)。

[1] 「～ていく」「～てくる」構文の意味が語彙特化性を示す特殊な用例で「やってくる」71例と「経由してくる」1例である。

3.3 意志性

表 3 V_1の意志性と「～ていく」「～てくる」構文の意味特性の関係

	いく				くる				
	順次	並行	状態	一体	順次	並行	状態	一体	その他
意志性あり	25	109	82	229	66	41	25	314	72
意志性なし			10	42	2		19	41	

意志性については，(1)の関係性が成り立つとすれば，順次及び並行に分類されているものは意志性があるはずである。「～ていく」のクロス表についてカイ二乗検定を行った結果，主効果が見られた(χ^2 =23.009, df=3, p< .01)。さらに，残差分析の結果，順次で意志性ありが期待度数よりも有意に多い傾向があり(p< .1)，並行で意志性ありが期待度数よりも有意に多く(p< .01)，一体で意志性なしが期待度数よりも有意に多いことが分かった(p< .01)。状態については有意差がなかった。

「～てくる」のクロス表についてその他部分を除外してカイ二乗検定を行った結果，主効果が見られた(χ^2 =50.691, df=3, p< .01)。さらに，残差分析の結果，順次及び並行で意志性ありが期待度数よりも有意に多く(p< .05)，状態で意志性なしが期待度数よりも有意に多いことが分かった(p< .01)。一体については有意差がなかった。

3.4 移動性

表 4 V_1の移動性と「～ていく」「～てくる」構文の意味特性の関係

	いく				くる				
	順次	並行	状態	一体	順次	並行	状態	一体	その他
移動性あり		38	67	267	15	25	43	344	72
移動性なし	25	71	25	4	53	16	1	11	

移動性については，(1)の関係性が成り立つとすれば，一体に分類されているものは移動性があるはずである。「～ていく」のクロス表についてカイ二乗検定を行った結果，主効果が見られた(χ^2 =247.867, df=3, p< .01)。さらに，残差分析の結果，順次及び並行で移動性なしが期待度数よりも有意に多く，一体で移動性ありが期待度数よりも有意に多いことが分かった(p< .01)。状態については有意差がなかった。

「～てくる」のクロス表についてその他部分を除外してカイ二乗検定を行った結果，主効果が見られた(χ^2 =261.153, df=3, p< .01)。さらに，残差分析の結果，順次及び並行で移動性なしが期待度数よりも有意に多く，状態及び一体で移動性ありが期待度数よりも有意に多いことが分かった(p< .01)。

3.5 アスペクト

表 5 V_1のアスペクトと「～ていく」「～てくる」構文の意味特性の関係

	いく				くる				
	順次	並行	状態	一体	順次	並行	状態	一体	その他
継続	10	108	92	149	35	41	44	167	
瞬間	15	1		122	33			188	72

アスペクトについては，(1)の関係性が成り立つとすれば，順次に分類されているものはアスペクトが瞬間で，並行及び状態に分類されているものは，アスペクトが継続であるは

ずである。「〜ていく」のクロス表についてカイ二乗検定を行った結果，主効果が見られた(χ^2=127.706, df=3, p< .01)。さらに，残差分析の結果，順次及び一体でアスペクト瞬間が期待度数よりも有意に多く，並行及び状態でアスペクト継続が期待度数よりも有意に多いことが分かった(p< .01)。

「〜てくる」のクロス表についてその他部分を除外してカイ二乗検定を行った結果，主効果が見られた(χ^2=79.061, df=3, p< .01)。さらに，残差分析の結果，並行及び状態でアスペクト継続が期待度数よりも有意に多く，一体でアスペクト瞬間が期待度数よりも有意に多いことが分かった(p< .01)。順次については有意差がなかった。

３.６　予備分析の位置づけ

以上の結果は，「〜てくる」構文で「順次」と「アスペクト瞬間」との関係において有意差が出なかったことを除けば，(1)の関係性を支持する結果である。このように，森田(1998)で示された(1)の関係性はある程度実際の用例に当てはまり， V_1の意味特性と「〜ていく」「〜てくる」構文の意味特性との間に規則性が認められることを確かめることができた。

しかし，以上の分析(検定)において，「〜ていく」「〜てくる」構文の意味特性を人が判断する際 V_1を含める構文全体の意味解釈を含むという側面があり，その V_1の意味特性が「〜ていく」「〜てくる」構文の意味特性に影響を与えるのは当然のことと考えるべきである。（即ち，従属変数と独立変数と間の独立性が確保されているとは言えないため，カイ二乗検定自体適切とはみなせず，規則性を予備的に確認する以上の意味はない。）従って，前出の分析を，(1)の関係性に則っている用例と則っていない用例とを仕分けした上で両者の分布を確かめる予備分析として位置づけ，本稿では，4 節以降，例外的用例の分析に力点をおく。

４．例外的用例の定性的分析

４.１　予測通りの用例と例外的用例との頻度分布

以下の表 6 は，縦軸に(1)の関係性に基づく「〜ていく」「〜てくる」構文の意味特性の予測を，横軸に実際の意味特性をとった「〜ていく」「〜てくる」それぞれの出現についての表である。

表 6　「〜ていく」「〜てくる」構文の意味特性の予測と実際の意味特性の頻度分布

予測＼実	いく			
	順次	並行	状態	一体
順次	14	1		3
並行	8	108	1	1
状態	2		91	
一体				263
その他	1			4

予測＼実	くる				
	順次	並行	状態	一体	その他
順次	26			5	71
並行	21	41		13	
状態	14		44		
一体	2			336	
その他	5			1	1

これらの表を縦方向に見ると，実際の意味特性が順次であるものの中では，森田(1998)で示された(1)の関係性に基づく予測も順次となっているものが最多であり，他の三つの意味についても同様である。表を横方向に見ると，「〜ていく」「〜てくる」構文の意味特性の予測が順次となっているものの中では，予測通り実際の意味も順次であるものが最多であり，他の三つの意味についても同様である。表 6 はタイプ数ではなくトークン数で集計したものであるが，V_1の意味特性と「〜ていく」「〜てくる」構文の意味特性との関係に関する森田(1998)の仮説は，「〜ていく」では，497 用例のうち 95%以上にあたる 476 用例で正しく成り立っている。また，「〜てくる」でも，580 用例のうち約 77%にあたる 447 用例で正しく成り立っている。しかしながら，実際の意味特性が順次・一体であるものについては，

仮説に基づくと別の意味特性になると予測されていたものも少なからずある。

４.２　コーパスに見られる例外的用例

ここでは，森田(1998)に基づく予測に当てはまらなかった例外的な用例を全てみた中で，その一部を取り上げて分析した上で，実際の用例に基づく新たな意味分析を試みる。

(2) a. 私「この時間からだもん、いらないよ」ダ「帰り、店見ていくわ」私「いいって」

他動性	方向性	意志性	移動性	アスペクト	意味特性[2]	予測
あり	なし	あり	なし	継続	順次	並行

【出典】『Yahoo!ブログ』（2008）

b. さてと‥ちょっと運動してくるかな。。

他動性	方向性	意志性	移動性	アスペクト	意味特性	予測
なし	なし	あり	なし	継続	順次	状態

【出典】『Yahoo!ブログ』（2008）

c. 父が、「ちょっと、まさと出かけてくる」と母に言って、その小さな旅は始まる。

他動性	方向性	意志性	移動性	アスペクト	意味特性	予測
なし	あり	あり	あり	瞬間	順次	一体

【出典】徳光正行（2003）『せんえつですが…。』幻冬舎.

森田(1998)に基づく仮説では (2abc) における「～ていく」「～てくる」構文の意味特性がそれぞれ並行・状態・一体と予測されるが，これらの用例の実際の意味特性は順次である。これらの用例に共通することは，別の文脈では予測通りの意味特性になることがあるということである。また，四つの意味特性のうち，順次だけが移動と V_1 の表す行為とが同時に行われないという特徴がある。以上の二点から考えると，V_1 の意味特性から並行・状態・一体と予測され，実際の意味特性も予測通りになることが多い場合でも V_1 の表す行為の後に移動が起こる文脈で使われれば順次を意味することになり，順次になることを完全に予測することは不可能であるということが分かる。

また， 予測が順次でないもののうち，移動主体が話者で，かつ，到着点が話者の位置の場合は順次になりやすい。このような用例は「くる」で 27 用例あり，そのうち 21 用例の意味特性が順次である。これは，出発点（話者の位置）から「どこか」へ行って到着点（話者の位置）に戻ってくるという移動を「～てくる」構文を使って表そうとする場合というのは，上記の「どこか」で何かをする目的をもって移動する場合が多く，特に移動主体が話者の場合は「どこか」で何をするのかを聞き手に伝える目的で「～てくる」構文が用いられるからだと考えられる。

(3) a. 何を着ていくべきとは無いと思いますよ。

他動性	方向性	意志性	移動性	アスペクト	意味特性	予測
あり	なし	あり	なし	継続	並行	並行

【出典】『Yahoo!知恵袋』（2005）

b. 早速明日も学校へ傘を差していくと思うので、もしよければ教えて下さい。

他動性	方向性	意志性	移動性	アスペクト	意味特性	予測
あり	なし	あり	なし	継続	並行	並行

【出典】『Yahoo!知恵袋』（2005）

[2] 「～ていく」「～てくる」構文の意味特性を指す。以下同様。

c. 下地は自分が普段使用しているものを持ってくるように言われましたけど。

他動性	方向性	意志性	移動性	アスペクト
あり	なし	あり	なし	継続

意味特性	予測
並行	並行

【出典】『Yahoo!知恵袋』(2005)

「～ていく」「～てくる」構文の意味特性が並行と予測され，実際の意味特性も並行である用例の V_1 は，(3)のように「着る」「差す」「持つ」などである。ここであげた実際の意味特性も並行となるような V_1 は，他の動詞に比べて移動との両立がしやすいといえる。なぜなら，これらの動詞は「着る」「差す」「持つ」という動作を表すとともに，その動作を終えた後には「着ている」「差している」「持っている」という状態になっていることも表す動詞であり，「着る」「差す」「持つ」という動作をした後は移動するだけで二つの行為を並行して行っていることになるからである。しかし，移動と同時に起こっている V_1 が表す行為が「着ている」「差している」「持っている」という状態ならば，その意味特性は状態とも捉えることができ一つに定めることができない。さらに，動作が完了した後に起こる移動と動作完了に伴って発動する状態が同時に起こるということで，順次の要素も入っているため，新たな意味特性を立てる必要がある可能性もある。

(4) 集団が次々に響木とランニングシャツの男を追い抜いていく。

他動性	方向性	意志性	移動性	アスペクト
あり	あり	あり	あり	瞬間

意味特性	予測
並行	順次

【出典】斎藤純（2004）『銀輪の覇者』早川書房.

(4)は「～ていく」構文の意味特性が順次と予測されながら実際の意味特性は並行である用例だが，この用例は方向性のある移動を表しているため，一体の意味になる要件も満たしている。ここでは，追い抜かれる人を伴っている点を考慮して並行としているが，森田(1998)の意味分類では並行と一体が両立可能であることが分かった。また、この用例のように V_1 の表す瞬間動作が移動中に起こる例を考えると，以下のようなものがある。

(5) a. ボールを蹴っていく。
b. 通行人から次々と財布を奪っていく。
c. 彼はこちらに手を叩いてくる。

(5)はいずれも移動中に瞬間動作が繰り返し行われる例であるが，(5ab)はボールや財布を伴って移動していることから「～ていく」構文の意味特性が並行になる一方で，(5c)は何も伴っていないため，森田(1989)の並行の定義に則ると、状態となる。

(6) 白いストリング・ビキニの少女が，ポルシェのハンドルを握った僕に手を振っていく。

他動性	方向性	意志性	移動性	アスペクト
あり	なし	あり	なし	継続

意味特性	予測
状態	並行

【出典】大石圭（2002）『殺人勤務医』角川書店.

(6)の「振る」については，移動と「振る」行為が同時に継続的に起こっており予測通り並行となるかと思われるが，森田(1989)で示された意味分類によると並行という意味特性は「何かを持ってまたは伴っての移動」を表すものであるため，移動主体が何も持っておらず伴ってもいないこの用例は，並行とはいえない。(このことが５節で述べる検討に繋がる。)

この用例で，移動主体が手を振った状態で移動していく様子を表しているという側面を捉えると，意味特性は状態である。

５．新たな意味分類

以上のように，森田(1998)で示されている意味分類は，実際の用例と照らし合わせてみると一部ではあるが予測から外れた例外的な用例が見られ，また，識別すべき意味が捉えきれないという諸相が認められた。そこで，それらの点を考慮した上で，以下のような新しい六つの意味分類を提示する。

① 順次
V_1の表す行為が完了した後に移動が起こる。
② 並行
移動とは異なるV_1の表す行為を移動中に継続的に行っている。
③ 反復
移動中に繰り返し移動と異なるV_1の表す瞬間動作を行っている。
④ 状態
V_1の表す行為が完了した後に移動が起こり，移動中はV_1の表す行為の結果状態が継続する。
⑤ 手段
V_1の表す行為が移動の様態や手段を表す。
⑥ 一体
移動することがV_1の表す行為を行うことにもなっている。

①順次に関しては定義が変わらず分類される例も変わらない。②新並行[3]は旧並行と定義が異なる。何かを持ってまたは伴う必要がなくなり，(6)の「手を振っていく」も新並行に分類されるようになる。③反復は，主に旧並行に分類されていたもののうち，移動中に瞬間動作を繰り返し行うものを独立させた。これにより，(5abc)は全て反復に分類されるようになる。④新状態には，(3)で触れた「着る」「差す」「持つ」などのほか，旧状態に分類されていた「座る」「乗る」なども分類される。⑤手段は，旧状態を細分化したもので，「歩く」「泳ぐ」などが分類される。⑥新一体は，旧一体と定義が異なり，旧並行と旧一体が両立可能だった点を改善するため，新並行とは両立しない定義になっている。旧一体に分類されていた「入る」「落ちる」などのほか，旧並行に分類されていた「追う」「連れる」などもこの分類になる。

６．おわりに

本稿では，V_1の意味特性と「～ていく」「～てくる」構文の意味特性との関係について，森田(1998)に基づく仮説を立て， コーパスから抽出した用例への意味コーディングに基づく定量的予備分析において，まず多くの用例がある程度予測可能な規則的傾向を示すことを確認した上で，例外的用例の量的・質的分析を行った。これらの分析を通して，例外的用例の中にも傾向を見いだすことができ，新たな意味分析への示唆を得ることができた。

[3] 便宜上，森田(1998)の分類と新たな分類とで名称が同じものに関しては前者に言及するときは旧を，後者に言及するときには新を付す。

謝　辞

本分析において，横山詔一氏（国立国語研究所教授，東京大学大学院総合文化研究科客員教授）の大学院講義に，教養学部の学部生である第一著者(加藤)の聴講を特別にご許可いただいき聴講させていただいたことが大変有益であった。記して感謝申し上げる。

文　献

Fillmore, Charles. (1997). *Lectures on Deixis*. Stanford: CSLI Publications.

Kuno, Susumu. (1987). *Functional Syntax: Anaphora, Discourse and Empathy.* Chicago: University of Chicago Press.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Martin, Samuel E. (1975). *A Reference Grammar of Japanese.* New Haven: Yale University Press.

Shibatani, Masayoshi. (1990). *The Languages of Japan.* Cambridge: Cambridge University Press.

Shibatani, Masayoshi. (2003). *Directional Verbs in Japanese.* In: Shay, Erin. Seibert, Uwe. (eds.) (2003). *Motion, Direction and Location in Languages: in Honor of Zygmunt Frajzyngier*. Amsterdam: John Benjamins Publishing Company. pp.259-286.

金田一京助・山田忠雄・柴田武・酒井憲二・倉持保男・山田明雄 (1997).『新明解国語辞典』第五版，三省堂.

西尾実・岩淵悦太郎・水谷静夫 (2000).『岩波国語辞典』第６版，岩波書店.

森田良行 (1998).『基礎日本語辞典』第八版，角川書店.

山口治彦 (2009).「視点の混在と小説の語り―自由間接話法の問題をめぐって―」坪本篤朗・早瀬尚子・和田尚明(編)『「内」と「外」の言語学』pp.217-248，開拓社.

山下杉雄・村上公雄・塩谷善之・大西匡輔 (1998).『精選国語辞典』新訂版，宮地裕・甲斐睦朗監修，明治書院.

山田俊雄・築島裕・白藤禮幸・奥田勲 (2000).『新潮現代国語辞典』第二版，新潮社.

関連 URL

現代日本語書き言葉均衡コーパス（BCCWJ）　http://pj.ninjal.ac.jp/corpus_center/bccwj/

明治初期教科書『物理階梯』のコーパス作成による語彙の考察

田中 牧郎（明治大学）
島田 むつみ（明治大学大学院生）
髙橋 雄太（明治大学大学院生）

Vocabulary Analysis of Meiji Early Physics Textbook *Butsuri-Kaitei* through Corpus Construction

Makiro Tanaka (Meiji University)
Mutsumi Shimada (Meiji University Graduate School)
Yuta Takahashi (Meiji University Graduate School)

要旨

明治初期の小学校用の物理学の教科書『物理階梯』のコーパスを作成し、同時期の啓蒙雑誌『明六雑誌』のコーパスと比較することを通して、語彙の考察を行った。まず、『物理階梯』の語彙は、『明六雑誌』の語彙に比べて、異なり語数において和語の比率が高いことがわかった。また、『明六雑誌』と比較した場合の『物理階梯』の特徴語を抽出して、その性質を考察すると、物理学のテーマに関連してよく用いられる〈テーマ語〉、物理学を体系的に論じるために必要とされる〈専門語〉、物理学に限らず学術的な内容を叙述するのに適した〈学術語〉の３種に分類できた。その３種を、『増補改訂分類語彙表』の部門別に集計すると、〈テーマ語〉は「生産物および用具」に、〈専門語〉は「自然物および自然現象」に、〈学術語〉は「抽象的関係」に、それぞれ特に多いことなどがわかった。

１．はじめに

本稿は、1872（明治 5）年に刊行された、日本で最初の小学校用の物理学教科書『物理階梯』のコーパスを作成し、そのコーパスの形態論情報に基づいて、『物理階梯』の語彙について考察を行うものである。考察の方法は、主として、同時期の啓蒙雑誌『明六雑誌』の語彙と比較することによる。まず、品詞と語種の観点から語彙全体を量的に概観し、次に、『物理階梯』の特徴語を抽出し、その特徴語を〈テーマ語〉〈専門語〉〈学術語〉の３種に分類し、それぞれの性質について考察する。

２．『物理階梯』のコーパス作成

２.１　言語資料としての『物理階梯』

『物理階梯』の著者、片山淳吉（1837～1887）は、出身地舞鶴で軍事や測量を学んだ後、江戸へ出て福澤諭吉や箕作麟祥の塾で洋学を修め、明治政府の兵部省や文部省に入り、多くの教科書を執筆した人物である（岡崎 1985）。片山の書いた最初の教科書が『物理階梯』で、以後、『百科全書 植物生理学』（1874 年）、『万国地誌要略』（1879 年）、『小学物理講義』（1881 年）などを執筆している。『物理階梯』は、上中下３冊からなり、イギリス人パークル（R.G.Parker）の *First Lesson in Natural Philosophy* を抄訳する形で書かれ、1876 年に片山自身が改訂した『改訂増補物理階梯』とともに、明治前半期の長期間、全国の小学校で広く使われた。『物理階梯』の内容は、第１課「物体論」から第 40 課「日蝕、月蝕、潮汐論」まで、物体、力、音、熱、光、電気、天体など、物理学の広範囲が扱われ、豊富な図解も交えられている。文体は漢文訓読体、表記体は漢字片仮名交じり文。所々に片仮名で振り仮名が施され、おおむね、右傍のそれは読みを示し、左傍のそれは意味の説明になっている。

『物理階梯』は、科学史や教育史の重要資料として研究が重ねられてきているが、言語研究の資料としては、従来あまり使われておらず、『改訂増補物理階梯』が『日本国語大辞典第２版』（小学館）の用例採集資料として用いられていることや、杉本（1991）が『物理階梯』と『改訂増補物理階梯』の比較を行っていることが、先行研究として指摘できる程度である。しかしながら、『物理階梯』は、近代の物理学の専門語や、物理学に限らない学術的あるいは教育的な語彙や表現を把握する際に、重要な資料であると考えられ、その言語の研究の必要性は高く、コーパス化の意義も大きいことが見込まれる。

２.２　『物理階梯』の電子化と形態素解析

電子化にあたっては、翻字テキストである『日本教科書大系 近代編 22 理科(2)』（1965年、講談社）によって入力し、題簽に「官版 物理階梯」とある上中下三冊本（明治大学田中研究室蔵）によって、校訂する過程をとっている。振り仮名は、右傍・左傍を区別してタグで表示した。これに、形態素解析辞書「近代文語 UniDic」を用いて、短単位で形態素解析を施し、誤解析箇所を目視で確認し、人手で修正している。その際、右傍の振り仮名は読みとして採用したが、左傍のそれは読みには採用しなかった、振り仮名がない部分で、読みを確定する根拠が得られない場合もあるが、筆者らの判断で何らかの読みを決めている。これらの修正作業や読みの確定作業は目下進行中であり、今回報告する数値は暫定的なものにとどまる。こうして短単位ごとに形態論情報が付与された形の Excel ファイルを、コーパスとして管理している。

２.３　『物理階梯』の語彙の概要

上記の形態論情報（短単位）が付与されたデータをもとに、『物理階梯』の語彙量を集計すると、記号、補助記号、空白を除外して、延べ語数 38,120、異なり語数で 3,791 となる。これを、品詞別の構成比がわかるグラフにすると、図１のようになる。品詞の枠組は、UniDic の「大分類」をもとに、接頭辞と接尾辞は「接辞」にまとめた。その中から、助詞と助動詞を除外して、さらに固有名詞を除外した、延べ語数 25,914、異なり語数 3,657 について、語種別の構成比がわかるグラフを作ると、図２になる。

『物理階梯』の語彙を相対的に見るために、国立国語研究所によりコーパス化が行われ語彙調査の結果が示されている『明六雑誌』(『日本語歴史コーパス 明治大正編 I 雑誌』所収、『明六雑誌コーパス』としても公開）の語彙と比較する。『明六雑誌』は、1874・1875（明治 7・8）年に、洋学者の集った明六社が編集刊行した啓蒙雑誌で、人文社会科学を中心に一部自然科学も含む広範囲の分野の論説が収められており、『物理階梯』と同時期の一般性の高い言語資料として比較に適している。近藤(2012)が示す語彙統計表をもとに、『物理階梯』のデータと比較できる形に集計し直して図示すると図３・図４のようになる。

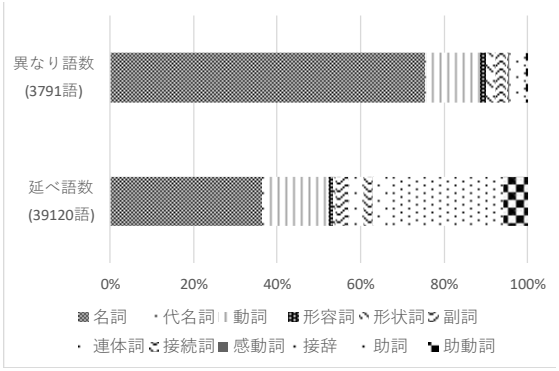


図１　『物理階梯』の品詞構成

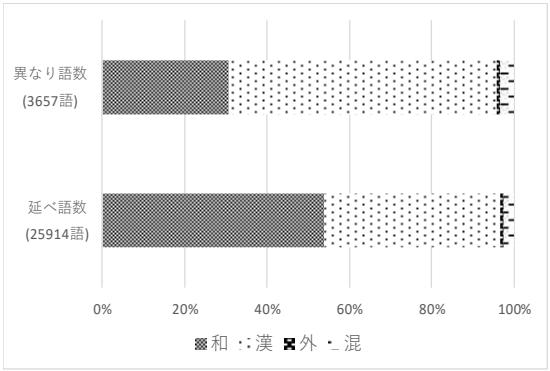


図２　『物理階梯』の語種構成

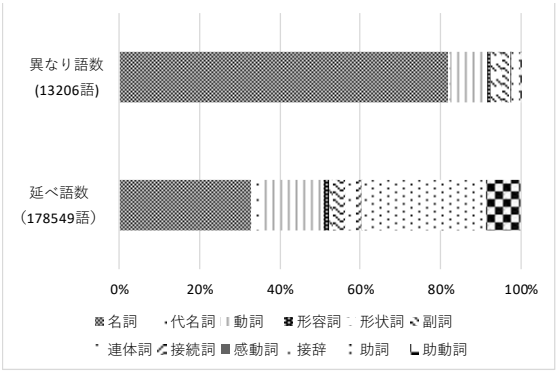


図３　『明六雑誌』の品詞構成

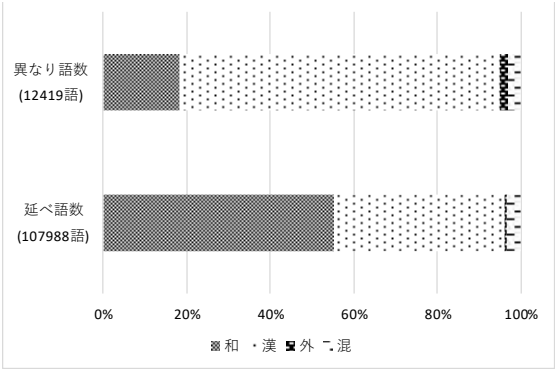


図４　『明六雑誌』の語種構成

図１～図４から、『明六雑誌』と比較したときの『物理階梯』の特徴として、次の２点が指摘できる。まず、品詞の点からは、異なり語数において名詞の比率が低く、その分動詞の比率がやや高いことである。そして、語種の観点からは、やはり異なり語数において漢語の比率が低く、和語の比率が高いことである。後者の特徴が顕著であるが、この背景には、例えば、小学校用の教科書であることで『明六雑誌』よりも平明に書かれていることや、話題が物理学に限定されることで、語彙の広がりが小さかったことなどが想定される。

３.『物理階梯』の特徴語―『明六雑誌』と比較して―

３.１　特徴語の抽出

本節では、『明六雑誌』と比較した場合の、『物理階梯』の特徴語について考察する。『物理階梯』の語彙（助詞・助動詞を含む、記号類は除く）を頻度の高い順に並べ、第１位の語から頻度を累積していき、その累積頻度が全体（延べ語数）の中で占める割合（累積使用率）を調べると、頻度４の語までで 89.98%になり、ほぼ 90%に達する。この頻度４以上の語彙を、『物理階梯』でよく使われている語（高頻度語）と扱う。同じように『明六雑誌』の累積使用率を調べると、頻度５の語までで 90.76%と 90%を超え、頻度３の語までで 94.02%と 94%を超える。そこで、頻度が２、１、０のいずれかのものは、『明六雑誌』の低頻度語または頻度ゼロの語と扱う。これら『物理階梯』の高頻度語でありながら、『明六雑誌』の低頻度または頻度ゼロの語を取り出すことで、『明六雑誌』と比較した際の『物理階梯』の特徴語を得ることができると考えた。この基準で抽出された『物理階梯』の特徴語は、全部で 393 語となった。その 393 語がどのような性質を持っているかを見ていこう。

３.２　特徴語の分類

３.２.１　テーマ語

『物理階梯』の特徴語として、第１に、物理をテーマとすることによって特徴語となっているものがある。日常の様々な場面で用いられる語であっても、物理に関することがテーマになるときに、特によく用いられるものである。ある作品によく用いられる語に関して、寿岳（1967）が〈テーマ語〉と呼ぶものに相当する。

(1)〔甲〕ノ一端ヲ密閉セシ玻璃細管ノ二尺六七寸ナルモノヲ把リ、其中ニ水銀ヲ充テ之ヲ倒マニシテ更ニ水銀ヲ盛リタル〔丙〕ノ小杯中ニ立ツレハ（中・第 18 課）
(2)光線〔丙〕ヨリ出テヽ鉛直線ニ〔乙〕ヲ射ルトキハ反射亦同線ニ復シ若シ〔甲〕ヨリ出テヽ斜メニ〔乙〕ヲ射ルトキハ其位ヲ變シテ〔丁〕ニ反射スト雖トモ其角度ノ如キハ〔甲〕〔丙〕ノ角度ト鋭鈍ヲ同フシテ〔丙〕〔丁〕ノ角度ヲ為スヘシ、（中・第 25 課）
(3)桶或ハ筒ニ、水ヲ盛リ、側面ニ、二三ノ孔ヲ穿ツトキハ（上・第 14 課）

(1)で点線を付した「倒（さかさ）ま」、(2)の「斜め」、(3)の「桶」「筒」「孔（あな）」「盛る」などは、恐らく当時の日常語でもよく使われていたと思われる。それが、文章語においては、『明六雑誌』に多い一般の論説ではあまり話題にならず、『物理階梯』のような物理をテーマにする場合はよく話題になり、使われる頻度が高くなっているのだと考えられる。

３.２.２専門語

『物理階梯』の特徴語の第２に、物理学の〈専門語〉があげられる。学術としての物理学を体系的に論じるために、明確な定義のもと、よく使われる語である。専門語は、基本的にその分野のために存在している語であり、他の分野で使われることがあっても、物理学の用語と意識されたり、物理学の知見や背景とつながった使われ方をするものである。

(4)物咸ナ重量アリテ、重ノ聚ル所、之ヲ重心ト曰ヒ（上・第８課）
(5)之ニ標スルニ、三點ヲ以テス、即チ其一ヲ、力點ト曰フ力勢ヲ加フル所ナリ、其二ヲ、重點ト曰フ、重物ニ接スル所ナリ、其三ヲ、支點ト曰フ、槓杆ヲ支撐シテ、桔槔ヲ為スノ所ナリ、（上・第９課）
(6)天文ノ學ハ天體ノ運行及ヒ其大小距離等ヲ論スル一科ニシテ日月星辰之ヲ天體ト云ヒ又其天體ヲ大別シテ四類トス即チ恒星、游星、衛星、彗星ニシテ（下・第 35 課）

(4)の「重心」、(5)の「力点」「重点」「支点」、(6)の「天体」は、いずれも「……ヲ○○ト曰（云）フ」の形式で、用語を定義している部分に使われており、(6)の「恒星」「游星」「衛星」「彗星」は、前の部分で定義された「天体」の下位概念として提示されている。このように、概念体系の一部を明確に担う専門語の多くが、特徴語として抽出されているのである。

３.２.３学術語

特徴語の第３に、物理学に限定されない学術的な叙述に必要とされる一群の語が指摘できる。一般的な論説からなる『明六雑誌』にはあまり使われないが、学術的な叙述からなる『物理階梯』にはよく使われる、〈学術語〉と呼ばれるべきものである。

(7)即チ体ヲ擲テ、直線ニ昇降セシムルトキハ、之ヲ直垂ノ擲射力ト曰ヒ、又水準ト平行シテ、擲ツトキハ、之ヲ地平ノ擲射力ト曰ヒ、其他ノ方向ニ擲ツトキハ之ヲ傾斜ノ擲射力ト曰フ、（上・第７課）
(8)夫レ音響ハ物体ヲ出テヽ四方ニ散布スト雖トモ其音ノ向フ所ハ必ス直線ヲ為シテ進行シ之ヲ響線名ク（中・第 21 課）
(9)温ノ反射スルニ角度ヲ為スヤ猶光ノ角度ニ同シク其線出ト反射トヲ驗スルニハ二個ノ凹鏡ヲ把リ相隔テヽ之ヲ左右ニ置キ其凹面ヲシテ相對セシムヘシ（中・第 23 課）

(7)で繰り返し使用されている「擲射（てきしゃ）力」は定義づけられている〈専門語〉であり、その語基をなす「擲射」も専門語と見てよいものである。一方、「擲射力」の具体例を説明する中で用いられる、下線を付した「地平」「傾斜」は、定義を伴っておらず〈専門語〉とは見なせないものである。このうち「傾斜」は、(2)で示した〈テーマ語〉の「斜め」と近い意味を持ち、学術的な文章では、「斜め」よりもふさわしい場合が多いと考えられる。同様に、「地平」も、「平ら」などと違って学術的な文章に適していたと考えられる。(8)においては、定義のある専門語「響線」に対して、「音響」にはそれがない。「音響」は、日常語でも用いられたと考えられる「音（おと）」や「響き」とは異なり、学術的な文章で使われやすい語だと見ることができよう。(9) の「験する」も、同様の性質を持っていると判断できる。

３.３　特徴語の種類と意味分野

個々の語を上記の３種に分類する際には、用例からの根拠が必ずしも十分でない語があり、『物理階梯』だけから判断するのは難しい場合も少なくない。しかしながら、種類ごとに、特徴語となる理由は異なっており、それぞれの性質の違いは明確である。表１は、『物理階梯』の特徴語 393 語について、上記の３種について、『分類語彙表 増補改訂版』（国立国語研究所）の部門ごとの語数を集計したものである。多義語の場合は、『物理階梯』において使われている語義で分類した。また、『分類語彙表 増補改訂版』に収録されていない語は、筆者らの判断でいずれかの部門に入れた。

表１　特徴語の種類と部門

番号	部門	テーマ語	専門語	学術語	計
1.1	体の類・抽象的関係	19	59	98	176
1.2	体の類・人間活動の主体	0	0	0	0
1.3	体の類・人間活動―精神および行為	3	4	13	20
1.4	体の類・生産物および用具	26	9	12	47
1.5	体の類・自然物および自然現象	22	58	33	113
2.1	用の類・抽象的関係	9	0	6	15
2.3	用の類・人間活動―精神および行為	3	1	3	7
2.5	用の類・自然物および自然現象	4	0	0	4
3.1	相の類・抽象的関係	3	1	4	8
3.3	相の類・人間活動―精神および行為	0	0	0	0
3.5	相の類・自然物および自然現象	1	0	2	3
4	その他の類	0	0	0	0
	計	90	132	171	393

表１から次のことが読み取れる。

- 全般に、1.体の類に多く、2.用の類、3.相の類には少ない。
- 1.体の類の中では、1.1 抽象的関係に最も多く、次いで 1.5 自然物および自然現象に多い。そして 1.4 生産物および用具や、1.3 人間活動―精神および行為にある程度ある。
- 1.体の類の語について、特徴語の種類と意味との関係を見ると、〈専門語〉は、1.1 抽象的関係と 1.5 自然物および自然現象に多く、他は少ない。〈学術語〉は、1.1 抽象的関係に特に多く、1.5 自然物および自然現象にも多い。

特徴語が最も多い部門である 1.1 体の類・抽象的関係のなかでは、中項目「1.15 作用」「1.17 空間」に特徴語が特に多い。ここに属する語を種類別にリスト化すると、以下の通りである。配列は、『分類語彙表 増補改訂版』の番号順である。

「1.15 作用」

〈テーマ語〉動き

〈専門語〉　反射、機力、返衝、静止、激動、動、動、回転、引衝、冪、凝集、開散、遠心、求心、粘着、衝、摩擦、圧搾、屈折、屈撓、下圧、上圧、受展

〈学術語〉　変更、流動、顫動、旋転、安置、傾斜、傾欹、経過、経路、進行、直行、通

過、導達、動植、流出、吸入、注射、膜、昇降、上騰、墜下、連合、密閉、填充
「1.17 空間」
〈テーマ語〉逆様、斜め、室内、水中、地中、物陰
〈専門語〉　支点、重点、焦点、力点、垂線、子午、鉛直、重心
〈学術語〉　中央、両端、尖頭、他端、凹面、月面、斜面、前面、側面、表面、平面、面、両面、光面、上面、体面、周囲、周辺、外辺

こうしたリストをもとに、中項目ごと、またさらに細かい分類項目ごとに、個々の語の意味・用法の分析を行い、『物理階梯』の特徴語のありようを考察することが求められる。

４．おわりに

以上、『物理階梯』の語彙について、『明六雑誌』の語彙との比較を通して、いくつかの考察を行った。和語が多い背景として、小学校用教科書という媒体に起因する側面と、話題が物理学に限定される側面とを想定したが、これについては、他の分野の教科書や、教科書以外の媒体の語彙と比較していく必要がある。また、〈テーマ語〉〈専門語〉〈学術語〉のありようについても、他の分野の教科書の調査によって教科書に特徴的な語彙について総合的な研究を進めていく必要がある。さらに、近代の〈専門語〉や〈学術語〉がどのようにして形成されたかについては、江戸時代の洋学資料をはじめ江戸以前の学術文献の調査が求められ、現代の専門語や学術語への系譜については、明治中期以後の教科書の調査が必要である。そして、本稿では十分に行えなかった、個々の語の意味・用法の考察を行うことも重要である。科学史や訳語史の分野で明らかにされている知見を参照することも課題である。

謝　辞

本研究における『物理階梯』の入力作業は、国立国語研究所共同研究プロジェクト（2009～2012 年度）「近代語コーパス設計のための文献言語研究」（プロジェクトリーダー：田中牧郎）において行われた。また、形態論情報の整備作業には、本稿の著者らのほか、小松寛子（元明治大学大学院生）が参加した。

文　献

岡崎正志（1985）「『物理階梯』の編者片山淳吉の生涯」『科学史研究第Ⅱ期』24(154), pp.84-94
国立国語研究所（1981）『専門語の諸問題』国立国語研究所報告 68，秀英出版
近藤明日子（2012）「『明六雑誌』の語彙量」『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，p.144-149，国立国語研究所
寿岳章子（1967）「源氏物語基礎語彙の構成」『計量国語学』41，pp.18-32，計量国語学会
杉本つとむ（1991）「物理学用語の翻訳とその定着―「物理階梯」から「改正増補物理階梯」へ」『国文学研究』165，pp.67-78，早稲田大学国文学会

参考 URL

日本語歴史コーパス　http://pj.ninjal.ac.jp/corpus_center/chj/
明六雑誌コーパス　http://pj.ninjal.ac.jp/corpus_center/cmj/meiroku/
近代文語 UniDic
http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%B6%E1%C2%E5%CA%B8%B8%ECUniDic

話し言葉コーパスの転記タグ：『多言語母語の日本語学習者横断コーパス』と『日本語話し言葉コーパス』の比較

西川 賢哉（国立国語研究所コーパス開発センター）†

Tags in Speech Corpora: A Comparison between `International Corpus of Japanese As a Second language' and `the Corpus of Spontaneous Japanese'

Ken'ya NISHIKAWA (National Institute for Japanese Language and Linguistics)

要旨

近年の話し言葉コーパスにおいては，発話を書き起こした転記テキストに，タグ（転記タグ）が付与されることが多い。本発表では，『多言語母語の日本語学習者横断コーパス』(I-JAS) および『日本語話し言葉コーパス』(CSJ) を対象に，そこで用いられているタグの種類・形式・目的・実際の用例を整理したうえで，両者の比較を行なう。比較の結果，(i)一方にしか存在しないタグもあるが，両コーパスでほぼ同様の機能を果たすタグも少なからず存在する（例えば，フィラー，語断片，発音誤りを表すタグ），(ii)ただし，同様の機能を果たすタグとはいえ，タグの適用範囲（転記テキストのどこからどこまでにタグを付与するか）や，タグの適用対象（タグをそもそも付与するか否か）など，細かい点では違いもある，ということが判明した。

１．はじめに

複数のコーパスを統一的な枠組みで扱うことができれば，利用者の利便性は高まる。その可能性を検討するための第一歩として，本稿では，話し言葉コーパスの転記テキストに付与されているタグ（転記タグ）の比較・検討を行なう。対象とするコーパスは次の二つである：

- I-JAS :『多言語母語の日本語学習者横断コーパス(International Corpus of Japanese As a Second language)』（cf. 迫田他 2016）
- CSJ :『日本語話し言葉コーパス(the Corpus of Spontaneous Japanese)』(cf. 前川 2006)

I-JAS の検索には，コーパス検索アプリケーション『中納言』を用いる（引用にあたり，迫田他(2016)等で用いられている形式に従ってタグを表記する)。CSJ については，現在のところ CSJ 中納言に転記タグの情報が格納されていないため，DVD 版の転記テキストを使用する（引用時，基本形のみを改行なしで表示することがある)。引用箇所中特に着目している個所を示すために，適宜下線を施す。

† nishikawa[@]ninjal.ac.jp

なお，I-JAS は現在も開発が進められており，今後タグの仕様が変更される可能性があることをあらかじめお断りしておく。

２．タグ概観

まず，I-JAS と CSJ のタグ対応表を表 1（次ページ）に示す。同種の機能を果たす，あるいは機能の一部が重複していると思われるタグは並べて表示してある。参照の便のため，両端に行番号を付す（本稿で「n 行目」とある場合，表 1 の行番号を指すことにする）。I-JAS において「タグ」とされているのは，厳密には表 1 の 1 行目から 10 行目までに挙げたものに限られ，11 行以降に「記号」として挙げたものは「タグ」とは呼ばれない。しかし，I-JAS でいう「記号」も実質的にはタグとみなしうること，また，CSJ との比較という点では「記号」と「タグ」を区別する必要はないこと，といった理由により，ここでは「記号」もタグとみなすことにする（必要な場合，「記号」も含めた I-JAS のタグ全体を「広義のタグ」，I-JAS で「タグ」とされているものを「狭義のタグ」と呼んで区別する）。

３．一方のコーパスにのみ存在するタグ

表 1 を見ると，一方のコーパスにのみ存在するタグも，両コーパスで類似の機能を果たすタグも存在することが分かる。本節ではまず，一方のコーパスにのみ存在するタグについて，両コーパスの目的・方針・使用に触れながら，簡単に見ておく。

３．１　タグ笑，泣，咳，L

24-27 行目のタグ笑，泣，咳，L は，CSJ にのみ存在し，I-JAS にはそれに対応するタグが存在しない。これは，両コーパスの目的・方針の違いによるものと思われる。I-JAS は「主として日本語学習者の文法習得，談話習得などの研究を目的として書き起こしされる」（迫田編 2016: 170）ものであり，書き起こしの基本方針の一つに「（略）発話はできるだけ発音に忠実に書き起こす。しかし，発話の重なりやポーズの長さ，声の大きさ，などの情報は付与しない。（略）」とある（迫田編 ibid；下線は引用者による）。笑，泣，咳，L に相当するタグが I-JAS に用意されていないのは，こうした基本方針を反映したものと見られる。一方，CSJ 構築の背景には，自発音声の自動認識システムの開発があった（前川 2006: 2）。この目的からすると，特殊な発声（笑いながら／泣きながら／咳をしながら／小声で…）については，タグ付けする価値のある個所だということであろう。

３．２　タグ Y

次に，9 行目のタグ Y を取り上げる。このタグは，I-JAS にのみ存在するタグで，次のように使用される：

ドアが[開いて（ひらいて）=Y]ました　【出典】I-JAS サンプル ID：JJC35-D

この例の場合，何の対応も施さなければ「開いて」の読みに曖昧性が発生するが（「あいて／ひらいて」），タグ Y を用いて読みを添えることで，その曖昧性が解消される。CSJ にこ

表 1. I-JAS, CSJ タグ対応表：迫田編(2016: 173, 176)，小磯他(2006: 80）をもとに作成

	I-JASのタグ				CSJのタグ				
	概要	内容	タグ表記	タグの由来	タグ	内容	使用例	付与対象	
1	(1)解析用の品詞を指定	フィラーを感動詞に指定	[α=F]	フィラー	(F)	フィラー，感情表出系感動詞，（応答表現）	(F あの)，(F うわ)，(F うーん)	基・発	1
2		外国語を名詞に指定	[α=N]	Noun	(O)	外国語、古語、方言、複雑な数式の読み上げ	(O ザッツファイン)	基・発	2
3		連体詞に指定	[α=R]	連体詞					3
4	(2) 解析用の語を指定	PC入力時の変換ミス	[α=K=β]	漢字・仮名					4
5		語中の長音、ポーズ	[α=T=β]	訂正	(K)	何らかの原因で漢字表記できなくなった場合	(K たち(F えー)ばな;橘)	基・－	5
6					(B)	語の読みに関する知識レベルの言い間違い	(B シブタイ;ジュータイ)	－・発	6
7		語や活用や発音の誤り	[α=G=β]	誤用	(W)	転訛や発音の怠けなど，一時的な発音エラー	(W ギーツ;ギジュツ)	－・発	7
8	(3) 解析から除外	意味不明語、語の断片	[α=X]	バツ	(D)	言い直し・言い淀み等による語断片	(D こ)これ，(D チ)チーズ	基・発	8
9	(4) 曖昧性への対応	複数の読みがある漢字語α	[α（読み）=Y]	読み					9
10		発音不明瞭（α1かα2）	[α1/α2=H]	発音	(?)	聞き取りや語の判断に自信がない場合	(? タオングー)，(? 堆積, 体積)	基・発	10
11	記号	聞き取り不能	＊				(?)	基・発	11
12		間・ポーズ	、		<P>	短単位の内部に生じる0.2 秒以上のポーズ	オ<P:00453.373-00454.013>モイ	－・発	12
13		長音	ー		<H>	非語彙的な母音の引き延ばし	ソレデ<H>，スゴ<H>イ	－・発	13
14					<Q>	非語彙的な子音の引き延ばし	カイ<Q>セキ，ス<Q>ゴイ	－・発	14
15		個人情報	【 】		(R)	話者の名前・差別語・誹謗中傷など	国語研の(R ××)です	基・発	15
16		非言語情報	{ }		<笑>	言語音と独立に生じる話者の笑い	ガクセー<笑>ノ	－・発	16
17					<咳>	言語音と独立に生じる話者の咳	ソレデ<咳>	－・発	17
18					<息>	言語音と独立に生じる話者の息	ツマリ<息>	－・発	18
19		あいづち	〈 〉						19
20		上昇イントネーション	?						20
21		直接引用	「 」						21
22		書名、映画名、ドラマ名	『 』						22
23		複数の読みがある場合のフリガナ	（ ）						23
24					(笑)	笑いながら発話している箇所	(笑ナニガ)	－・発	24
25					(泣)	泣きながら発話している箇所	(泣ドンナニ)	－・発	25
26					(咳)	咳をしながら発話している箇所	シャ(咳リン)ノ	－・発	26
27					(L)	ささやき声や独り言などの小さな声	(L アレコレナンダッケ)	－・発	27
28					(A)	アルファベット・算用数字・記号の併記	(A シーディーアール;C D － R)	基・－	28
29					(D2)	助詞・助動詞・接辞・数字の言い直し	そこ(D2 が)に，(D2 不)不自然	基・発	29
30					(M)	音や言葉に関するメタ的な引用	助詞の(M は)は(M わ)と発音	基・発	30
31					(X)	非朗読対象発話（朗読における言い間違い等）	(X 実際は)実際には，	基・発	31

の種のタグが存在しないのは，転記テキストの仕様による。CSJ では，図 1 に示すように，漢字仮名を中心に書き表される「基本形」と，実際の発音を仮名の範囲で忠実に書き表した「発音形」という 2 種の表記法が採用されている。そのため，基本形の表記において読みの曖昧性が発生するケースであっても，対応する発音形を参照することで，その曖昧性は解消される。CSJ においては，わざわざタグによって読みを与える必要はないわけである。

0285 00642.999-00645.100 L:		
十一時ぐらいに	&	ジューイチ(W イ;ジ)グライニ
うちに	&	ウチ(W ン;ニ)
帰ってきたんですけど	&	カエッテキタンデスケド
0286 00645.894-00649.273 L:		
(F その)	&	(F ソノ)
門は	&	モンワ
閉まってるんですが	&	シマッテルンデスガ
玄関が	&	ゲンカンガ
薄く	&	ウスク
開いていまして	&	アイテイマシテ

図 1. CSJ 転記テキスト例(S02M0198)：
&の左側が基本形，右側が発音形

なお，基本形と発音形という二種類の表記法が存在する CSJ においては，タグを(i)基本形に付与するのか，(ii)発音形に付与するのか，(iii)基本形と発音形の両方に付与するのか，を規定しておく必要がある。表 1 では右端の「付与対象」欄にその情報が記されている。

３．３　タグＡ

I-JAS におけるタグ（ここでは「記号」を含まない，狭義のタグ）は，もっぱら形態素解析の精度向上のために用いられる（迫田編 2016: 173ff.）。一方，CSJ においては，同様の目的で使用されるタグも多くあるが，それに限定されるわけではない（小磯他 2006: 27-28）。表 1 の 28 行目のタグ(A)は，転記テキストの可読性を高めるために導入されたタグである。このタグを用いることにより，算用数字やアルファベットを表記に添えることができる。

(A 三．ゼロ六;３．０６)　　& サンテンゼロロク　【出典】CSJ ID：A06F0075
(A エイチエムエム;HMM)は　& エイチエムエムワ　【出典】CSJ ID：A01M0099

I-JAS にはタグ A に対応するタグは用意されていない。I-JAS の表記ルールでは，数字は漢数字で，アルファベット通りの発音をしている場合はアルファベット全角で記すことにな

っている（迫田編 2016: 170-171）。

４．類似の機能を果たすタグ

本節では，両コーパスで類似の機能を果たすと思われるタグを比較する。

４．１　フィラー

フィラーおよび感情表出系感動詞には，どちらのコーパスにおいても，タグ F が付与される（表 1 の 1 行目）。

はい、[あー=F]島田店長，ちょっとよろしいですか？　【出典】I-JAS サンプル ID：JJC46-RP1
発表内容ですが(F あー)まず研究の背景として　【出典】CSJ ID：A01M0065

相違点として，フィラーが連続した場合，I-JAS では全体に一つのタグを付与する場合があるのに対し，CSJ では個々のフィラーにそれぞれタグ F を付与する。

でも実際はそんなにー、[あの、んー=F]私，　【出典】I-JAS サンプル ID：JJC32-RP1
この例に見られますように(F あの)(F んー)一方が　【出典】CSJ ID：A06F0073

I-JAS におけるこの対応は，「学習者の多様な場つなぎ的表現に関して，何を一語とするかという判断が難しいため」である（迫田編 2016: 178）。このような対応の違いにより，両コーパスをタグ F で検索した場合，フィラーの頻度や種類に差が出てくることが予想される（例えば，タグ F が付与された「あの」を I-JAS で検索しても，[あの、んー=F]はヒットしないおそれがある）[1]。

４．２　語の断片等

語の断片に対しては，I-JAS ではタグ X，CSJ ではタグ D が付与される（表 1 の 8 行目）。

今のーシフトはー[あの=F]、[し=X]週三回，　【出典】I-JAS サンプル ID：JJC32-RP1
(F えー)(D し)下の項が少し変わってきます　【出典】CSJ ID：A01M0097

I-JAS のタグ X は，語の断片の他に，不明語，学習者によるオリジナルの語に対しても付与される点で，CSJ のタグ D とは異なる。

[みやまー=X]　（迫田 2017: 183）

[1] このことは，迫田編(2016: 178)にある通り，I-JAS 開発者によって既に認識されている問題である。なお，「タグ F の内部に”、”がある場合，そこでタグ F を括り直す」という処理（例：[あの、んー=F]→[あの=F]、[んー=F]）を検討中であるとのご連絡を最近 I-JAS 開発者からいただいた。

お父さんが[ピャーピャー=X?擬音語・擬態語]怒って　（迫田 2017: 183）

また，前述のタグ F と同様の相違点だが，I-JAS ではタグ X を付与する箇所が連続して現れる場合は，まとめてタグ X を付与しているのに対し（迫田 2017: 182-183），CSJ では複数の語の断片の連続と解釈される場合には，それぞれにタグ D を付与している。

でもほんとに、[かな、悲し=X]、悲しくなった、　【出典】I-JAS サンプル ID：JJC28-I
住所とか聞きますから(D き)(D き)(D き)聞きたいんですけどって
【出典】CSJ ID：S01F0183

４．３　発音誤り

発音の誤りについては，I-JAS ではタグ G，CSJ ではタグ W が用いられる（表 1 の 7 行目）。

雰囲気が、ちょっと、[ちょと=G=ちょっと]違うと友達、
【出典】I-JAS サンプル ID：JJC28-I
ちょっと ＆(W チョト;チョット)
早めに　＆ ハヤメ(W ン;ニ)
出る　＆ デル
時に　＆ トキニ　【出典】CSJ ID：S02M0198

両コーパスの重要な相違点として，CSJ ではタグ W の範囲（スコープ）は原則として短単位なのに対し[2]，I-JAS ではそのような制約はないということが挙げられる。

記事を[かきて=G=書いて]います　【出典】I-JAS サンプル ID：GAT39-I
書いてきたんですね ＆(W カエ;カイ)テキタンデスネ　【出典】CSJ ID：S01F1522

「書いて」は短単位としては，「書い」と「て」に分割され，CSJ では「書い」の部分にだけタグ W が付与されているが，I-JAS では「書いて」全体にタグ G が付与されている。日本語学習者の日本語という観点から独自にタグ G の範囲を規定することには大きな意義があると思われるが，少なくとも，中納言のような短単位ベースの検索システムでコーパス検索をする時には，個々の短単位にこの種のタグが付与されていた方が扱いやすい，ということは言えるであろう。別の例を挙げると，

[使えな、なかた=G=使えなかった]　【出典】I-JAS サンプル ID：JJC28-I

[2] 例外として，短単位境界で音の融合などが生じそこで切り離しがたい場合，複数の短単位にまとめてタグ(W)を付与することを CSJ では認めている（小磯他 2006:104）。例：僕は ＆(W ボクワ;ボカー)

は，CSJ 方式ではおそらくタグ D とタグ W を使って，次のように書き起こされる：

使えなかった & ツカエ(D ナ)(W ナカ;ナカッ)タ

また，適用範囲とは別の問題として，そもそもこの種のタグを適用すべきか否かはっきりしないケースがある。

[助けてくれ、くれません=G=助けてくれません] 【出典】I-JAS サンプル ID：TTH27-I
入ろうと[思いですけど=G=思いますけど] 【出典】I-JAS サンプル ID：KKD20-ST2

これらの例は確かに誤用ではあるが，形態素解析に関して致命的な失敗を招くとは考えられず，その点ではタグ G は不要であるとも思える。CSJ の基準では，「助けてくれ」の部分にはいかなるタグも付与されない。また，「思いです」の類については，おそらく CSJ には明示的な規定はない―母語話者による発話のコーパスなので，そもそも想定していない―が，この部分にもタグは付与されないと思われる。

４．４　外国語

外国語については, I-JAS ではタグ N が, CSJ ではタグ O が付与される（表 1 の 2 行目）。

一番高いところは[シャンライ=N]のほう 【出典】I-JAS サンプル ID：JJC09-I
(O カナディアンレイジング)でもって 【出典】CSJ ID：A05F0039

一見同種の機能を果たすように思われるタグではあるが，実際のところ，両者は適用対象もその導入目的も大きく異なる。I-JAS のタグ N は，外国語で表現された語に加え，アニメなどの架空のカタカナ語の固有名詞にも付与される。このタグの目的は，タグが付与された要素の品詞を「名詞」とすることである（ただし，解析用辞書に登録されている語であれば，その語の情報が与えられる）。一方で，CSJ のタグ O は，外国語に加え，古語，方言，複雑な数式の読み上げに付与される。その目的は，必要に応じて CSJ の利用者が分析から除外できるようにすることである。前述のとおり，CSJ 構築の背景には，自発音声の自動認識システムの開発があった。そのためには，現代共通日本語の体系から外れた個所は除外するほうがよい。この際，タグ O の情報が利用できる。

このような目的の違いにより，CSJ ではタグ O が期待される個所で，I-JAS ではいかなるタグも付与されない，というケースが生じる。I-JAS 書き起こしマニュアル（内部資料）によると，英語の場合，2 語以上や文単位になっている場合は，タグを付与せず，アルファベットで表記することになっている。

[あー=F]、I have so many great experiences、とても、すてきな，誕生日の、経験

しかし, CSJ では下線部にタグ O が必要である。このような事例があるとすると,「外国語」という観点から，I-JAS のタグ N と CSJ のタグ O を単純にまとめて扱うことは難しい。

５． おわりに

I-JAS と CSJ で用いられているタグの比較を行なった。そもそも転記タグは，コーパスの目的や方針に基づいて設定されるものであり，一見機能的に類似しているタグであっても，本稿で見たように，細部においては違いもありうる。しかし，冒頭で述べた通り，コーパスの違いに関わらず，統一的な枠組みで扱えるのであれば，そちらのほうが利用者にとっての利便性という点では望ましい。今後は，本稿で明らかになった問題点を解消しつつ，複数のコーパスを統一的に扱う仕組みを検討する。また，他の話し言葉コーパス―例えば，現在国立国語研究所で構築中の『日本語日常会話コーパス』(川端他 2017) ―で用いられているタグについても，今後の比較の対象としたい。

謝　辞

本研究は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」の成果である。I-JAS 開発者の佐々木藍子氏および小西円氏からは，I-JAS の仕様について数々のご教示をいただいた。記して感謝する。

文　献

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵 (2017)「『日本語日常会話コーパス』の転記基準と作業工程」言語資源活用ワークショップ 2016 発表論文集.

小磯花絵・西川賢哉・間淵洋子 (2006)「第 2 章 転記テキスト」『日本語話し言葉コーパスの構築法』国立国語研究所, pp.23-132.（http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/02.pdf よりダウンロード可能）

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子 (2016)「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』6:3, pp.93-110.（https://ninjal-sakoda.sakura.ne.jp/lsaj/?page_id=364 よりダウンロード可能）

迫田久美子（編）(2016)『海外連携による日本語学習者コーパスの構築―研究と構築の有機的な繋がりに基づいて―I-JAS 構築に関する最終報告書』平成 24-27 年度科学研究費助成事業（基盤研究 A）研究成果報告書（https://ninjal-sakoda.sakura.ne.jp/lsaj/?page_id=364 よりダウンロード可能）

前川喜久雄 (2006)「第 1 章 概説」『日本語話し言葉コーパスの構築法』国立国語研究所, pp.1-21.（http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/01.pdf よりダウンロード可能）

関連 URL

コーパス検索アプリケーション『中納言』　https://chunagon.ninjal.ac.jp/

『日本語日常会話コーパス』の転記基準と作業工程

川端 良子 (国立国語研究所 音声言語研究領域 / 千葉大学) *
臼田 泰如 (国立国語研究所 音声言語研究領域)
西川 賢哉 (国立国語研究所 コーパス開発センター)
徳永 弘子 (国立国語研究所 音声言語研究領域 / 東京電機大学)
小磯 花絵 (国立国語研究所 音声言語研究領域)

Transcription Criteria and Guidelines for Processing "Corpus of Everyday Japanese Conversation"

Yoshiko Kawabata (NINJAL / Chiba University)
Yasuyuki Usuda (NINJAL)
Ken'ya Nishikawa (NINJAL)
Hiroko Tokunaga (NINJAL / Tokyo Denki University)
Hanae Koiso (NINJAL)

要旨

本稿は，平成 28 年度から構築を進めている『日本語日常会話コーパス』の転記基準と転記作業工程を紹介する。本コーパスには，日常場面で自然に生じるさまざまなタイプの会話 200 時間がバランス良く収録される予定である。日常会話には，極めてくだけた表現も頻出する。こうしたデータを多人数で書き起こしをするためには，文字化をするための基準を明確に定める必要がある。また，大量の会話を限られた期間で書き起こすために，効率的に作業をするための工夫が必要になる。本発表では，これまでに収録された会話を転記しながら策定した転記基準と効率的に作業を行うために用いている方法を紹介する。

1. はじめに

国立国語研究所では，平成 28 年度から「大規模日常会話コーパスに基づく話し言葉の多角的研究」プロジェクトを進めている。このプロジェクトでは，さまざまなタイプの日常会話をバランス良く収録した大規模なコーパス『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, 以下 CEJC と略称する) を構築し，そのコーパスの分析を通して日常会話を含む話し言葉の特性を多角的に解明することを目指している (プロジェクトの詳細に関しては本ワークショップ予稿集所収の小磯 (2017) を参照)。

話し言葉の特性を多角的に分析するためには，様々な人々の多様な言語活動の実態を記録したデータが必要である。しかし，これまでに構築されたコーパスの多くは，会話場面や参加者の属性・関係などに偏りがある。本プロジェクトでは，各世代から均等に調査協力者 (以下，協力者) を募り，収録機材を渡して，調査者が立ち合わずに，協力者自身が日常のさまざまな

* kawabata@ninjal.ac.jp

場面の会話を収録する。こうして収録されたデータから，事前に行った会話行動調査 (小磯ほか 2016) を参考に幅広いレジスターをカバーするようにサンプルを選定しコーパスを構築する設計になっている (コーパスの設計については小磯ほか (2017)，収録方法については田中ほか (2017a)，田中ほか (2017b) 参照)。このようにして，様々な場面における現実の会話データが均衡的かつ大規模に収録される点が CEJC の最大の特徴となっている。

こうしたデータを対象に研究を行うためには，収録した音声を転記したテキストが不可欠である。コーパス内で均質な転記テキストを作成するためには，文字化の基準を策定し，基準に従って転記を行った後，実際に基準に従って転記がなされているかの二重，三重のチェックが欠かせない。そのため転記作業は通常多くの時間を要する。よって，適切かつ効率的に転記を行う方法を確立することが必要であり，またその方法の記録は将来のコーパス構築に対して有効な知見になると考えられる。我々は，実際のデータで転記を行いながら，効率的に作業をするための工程を検討し，ツールの開発や転記基準の改訂を行ってきた。本稿は試行錯誤の末に定まってきた転記基準と効率的に作業を行うために用いている方法について紹介する。

2. 転記基準

本節では，転記基準について説明する。ここで述べる基準はあくまで現時点での規定である。今後転記作業が進むなかで，規定が見直される場合があることに留意されたい。

2.1 基本方針

国立国語研究所共同研究プロジェクト「均衡性を考慮した大規模日本語会話コーパス構築に向けた基盤整理」(リーダー：小磯花絵 2014 年 7 月 ～ 2015 年 8 月) での検討内容をベースに以下を転記の基本方針とした。

- 発話内容はテキストで表現できる範囲で転記し，原則として漢字仮名まじりで表記する。
- 転記テキストと音声情報の同期をとることで，転記テキストから音声情報を容易に参照できるようにする。
- 母音の延伸や発音エラーなどの会話で生じる現象は転記する対象を定め，各種タグを用いて表現する。
- 転記テキストに対して自動形態素解析を実施し，語彙素・語形・発音形等の情報を付与する。

音声情報を文字化することによって失われる情報は非常に多い。可能な限り音声情報を転記するという方針も考えられるが，そのような方針では作業にかかる時間が増大するだけでなく，転記テキストの可読性が低下する。研究者ごとに会話中に生じる現象への興味は異なるため，必要以上の情報が含まれる転記テキストはかえって使いにくいものとなってしまう。そこで CEJC では，『日本語話し言葉コーパス (CSJ)』(国立国語研究所 2006)，および『千葉大学 3 人会話コーパス』(伝・榎本 2014) の転記基準や作業手順などを参考にして，読みやすさと作業効率を重視した転記仕様を策定することにした。

CSJ の仕様と異なり，表記の統一（例：狐／きつね／キツネ）や発音を記録したテキストの作成は行わない。UniDic(1)に基づく形態素解析によって形態論情報を付与することで，転記テキスト自体に表記の揺れがあっても柔軟な検索を可能とする。また，自動解析の結果得られる発音情報を人手でチェック・修正することにより，形態論情報から正確な発音の情報を得ることができるようにする。このように，形態素解析や自作の自動変換プログラムを用いることを前提にし，作業時の転記テキストは変換時に曖昧性が残らない範囲で簡略化して効率化を図る。

2.2 転記対象

転記対象となるのは，会話の参加者 (以下，参加者) が会話中に発した言語音，言語音とは独立に生じる笑い・泣き・歌，および会話の流れに深く関わるその他の発音に類する行為（会話上意味があると考えられる舌打ちなど）である。基本的には同意書が得られた話者の発話を転記し，同意書のない話者の発話は原則書き起こさない。ただし，飲食店での収録などにおいて，店員が注文をとるなど，当り障りがないと考えられる発話に関してはその限りではない。

2.3 転記単位

転記テキストは音声との同期をとるために，以下の条件を満たす発話ごとに転記テキストを区切っている。ここで区切られた転記テキストを「転記単位」と呼ぶ。転記作業は ELAN(2)や Praat(3)などを用い，映像と音声を参照しながら人手で行い，転記単位ごとに音声にアライメントをする。

1. 知覚可能な休止がある場合
2. 異なる音種（言語音・単独の笑い・泣き・歌・その他）が続く場合
3. 発話単位の切れ目がある場合

3 の「発話単位」は，Japanese Discourse Research Initiative によって策定された「長い発話単位」(JDRI 2017) に準拠する。長い発話単位とは，話し手と聞き手が行為や情報を交換する際の基本単位に相当し，統語的・談話的・相互行為的な一まとまりに対応する単位とされる。

2.4 表記法

発話内容は原則として，現代語の表記の習慣 (現代仮名遣い) に従って，漢字仮名交じりで表記する。使用する字種は，漢字，平仮名，片仮名を中心とし，必要に応じてローマ字での表記も可とする。数字は漢数字を用いる。発話単位の境界を示すタグとして句点「。」を使用するが，読点は用いない。

■発音の扱い　語によっては表記と発音に差異があるが，それが一般的に受け入れられているものがある。たとえば，綴り字において母音が続する言葉「しいたけ (siitake)」，「通院 (tsuuiN)」，「講座 (kouza)」などは発音が長音化して「シータケ」，「ツーイン」，「コーザ」と

(1) 国立国語研究所が規定した「短単位」に対して形態論情報を付与する電子化辞書。http://pj.ninjal.ac.jp/corpus_center/unidic/
(2) http://tla.mpi.nl/tools/tla-tools/elan/
(3) http://www.praat.org/

発音される場合が多い。このように，発音が表記から予測可能である語については，一律に標準的な表記 (「しいたけ」「通院」「講座」) を用いる。

こうした予測可能な発音ではなく，強調のために「すごい」を「スゴーイ」と母音を引き延したり，「スッゴイ」のように子音を引き延して発音する場合は，後述する「:」や「%」のタグ (表 1 参照) を用いて「すご:い」「す % ごい」と表記する。

「わからない」と言おうとして「ワアンナイ」と言ってしまうような，一時的な発音のなまけやエラーについては後述するタグ (W) を用いて，「(W ワアン|分から) ない」のように，実際の発音に加え，本来言おうとした表現を補足する。一方，「こりゃすげえ (これはすごい)」のような，(1) 音の転訛を伴い，(2) くだけた場面で（意図的に）使用される表現で，(3) 一個人に限らず幅広く観察されるものという条件を満たす表現は，発音の一時的なエラーとはみなさず，口語表現としてそのままの語形を表記する。CSJ の構築の際にも，口語表現を積極的に認定したが，CEJC が対象とする日常会話では，講演を中心とする CSJ よりもこうした表現が多く見られることから，形態素解析担当班と連携して，積極的に登録する口語表現を定め，形態素解析用辞書 UniDic の拡張を行う。

感動詞類 (フィラーを含む) やオノマトペについては，基本的には語彙を定めず，聞こえた通りに表記する(4)。

フィラーの例

あー， あのー，んー，えーっとー

オノマトペの例

びゅーびゅー風が吹く， ぼろっぼろの鞄

発音が全く聞き取れなかった部分は，予想される発話の長さ（モーラ数）に応じてシャープ (＃) をタグと組み合わせて記す。

2.5 タグの設計

転記には，発音エラーや非語彙的な音（延伸，促音挿入），語の言いさしなどを体系的に示すため，『千葉大学三人会話コーパス』の転記の仕様を参考に定めたタグを使用する。タグの一覧を表 1 に示す。非語彙的な発音の変化（:，%，W）やパラ言語的情報（L，C，S，T）を記述するものや，表記に関わるもの（K，M），個人情報など仮名化や伏字化などの後処理に関わるもの（R）のほか，転記テキストを対象に行われる自動形態素解析におけるエラーをあらかじめ回避するためのもの（Y，F，A，X）などがある。形態素解析用のタグは作業上のものであり，解析後に転記テキストから削除する予定である。本節では，一部のタグについて簡単に説明する。なお，タグは発話単位末を示す句点「。」以外はいずれも半角である。

(4) ただし，応答系感動詞 (「はい」，「うん」等) は，ある程度語彙化されているため，明らかに発音のエラーと思われる場合 (例えば，「はい」と言おうとして「アイ」と言ってしまう場合) についてはタグ (W) を用いて表記する。例の場合，(W アイ|はい) とする。

表 1　転記テキストに使用されるタグの一覧

タグ	概要	使用例
:	非語彙的な母音の引き伸ばし	すご:い，デー:タ
%	非語彙的な音の詰まり	す % ごい，解 % 析
?	疑問上昇調	行きます?，コップ?
(D)	語の言いさし	(D コ) 明日から
(W)	言い誤り・発音の怠け等の一時的な発音エラー	(W コエ|これ)，(W ギーツ|技術)
(K)	タグ付与等のために漢字表記ができない箇所	(K シ:ツ|質) 問，(K リ % ツ|律)
(M)	音や言葉自体が言及の対象とされている発話	(M すごい) を (M すっごい) と発音する
(T)	小さい声で発話している箇所	(T これじゃないのか)
(L)	笑いが生じている箇所	(L)，これ (L なんですけど)
(C)	泣きが生じている箇所	(C)，(C なにが)
(S)	歌が生じている箇 所	(S)，(S ふるさと)，(S へイへイホー)
(O)	一般的でない外国語/方言が用いられる箇所	(O ボッソワー)，(O ＃＃＃)
(U)	聞き取りや語の判断に自信がない箇所	(U 外国/外交)，(U な＃＃＃)
(R)	個人情報などに関わる仮名・伏字処理候補	(R 国語研究所) の (R 佐藤) さん
。	発話単位末	食べます。，やったけど。，うん。
<>	発音に類する行為	<舌打ち>，<咳>，<口笛>
@	転記単位に対するコメント	スパ@車の愛称
(X)	語が不明な箇所	(X リョウゴ) アタック，(X ルトラ) のさ
(Y)	漢字表記の一般的な読みと発音が異なる箇所	(Y ゼツ|舌)，(Y ギョク|玉)
(F)	「その」がフィラーとして使用された場合	(F その) 研究所への行き方については
(A)	「あの」が連体詞として使用された場合	(A あの) 人が

■タグ :　語彙的には母音の引き延ばしが含まれないにもかかわらず，強調や言い淀みなどのために一時的に母音が引き延ばされた箇所に「:」(コロン) を付与する。

> 冷た:い視線で
> す:ごい腹立ったな:っていう

■タグ %　強調や言い淀みなどのために，一時的に音が詰まった箇所に「%」(パーセント記号) を付与する。

> き % ついね
> なん % かね:

■タグ ?　「?」は上昇調の句末に付与し，発話が聞き手への質問や確認などであることをを示す。上昇の音調であっても，質問や確認など聞き手への働きかけでないもの (例えば強調など) は付与対象外とする。

■タグ (D)　以下のケースで生じる「語の断片」にタグ (D) を付与する。語の断片は片仮名で表記する。なお，ここで「語」とは「短単位」(小椋 2014) を指す。

- 言いかけて語の途中で発話をやめた場合の中断した語。 言いかけた語が推測できる場合は，後述のタグ (W) と合せて用いる。推測できない場合はタグ (D) を単独で用いる。

> えー (D ダ) 例えば
> っていうだけじゃ (D ワカ) だめだよ。

- 語を言いかけたと言うよりは，発声上の問題で生じたと考えられる断片的な音声。

> その (D ン) 問題は

■タグ (W) 言い誤りや発音の怠けなどによって，一時的に非標準的な発音が生じた場合，(W 実際の発音|意図された語) の形で表記する。 実際の発音は片仮名で表記する。

> (W ワアン|わかん) ない ←「わかん (ない)」を「わあん (ない)」と発音
> (W ジュブン|自分) 一人でできるよ。 ←「じぶん」を「じゅぶん」と発音

語の断片のうち何を言いかけたか分る場合はタグ (W) を使用して言いかけた語を補い，言いさしであることを タグ (D) で示す。

> 知らない (W (D ヒ)|人) 知らない人に ←「人」と言いかけて「ひ」で中断

■タグ (M) 「あ という文字は め と非常によく似ている」のように，音や言葉自体を言及の対象としているような発話 (メタ的引用) のうち， そのままでは可読性が著しく低くなる場合や，タグ: % (W) などを用いて表記すると意図が通じなくなる場合は， その範囲にタグ (M) を付与して可読性を高める。

> (M 僕が) の (M が) は格助詞で (M 行って) の (M て) は接続助詞
> (M すごい) を (M すっごーい) のように促音を入れ強調して話す

■タグ (T) いわゆるささやき声など通常の会話時よりも明らかに小さな声で発話している箇所に付与する。声の大きさに関しては，通常の会話より音量が大きい場合と小さい場合がある。小さい場合のみタグを付与する理由は，声が小さい場合は，聞き手への働きかけではなく，いわゆる「独り言」である可能性があるからである。ただし，転記作業では独り言であるかどうかの判断を行わず，音量の小ささのみからタグの付与を判断する。

■タグ (L)(C)(S) 言語音以外として「笑い」と「泣き」および「歌」を転記対象とし，それぞれ以下のタグを付与する。

- 笑い: タグ (L)
- 泣き: タグ (C)
- 歌: タグ (S)

笑いながら，泣きながら，歌いながら発話している場合，その範囲に上記タグを付与する。非言語音が単独で出現する場合，あるいは歌詞を伴わない (聞きとれない) 歌の場合には，それぞれ (L),(C),(S) を単独で記す。

■タグ (O)　外国語など，現代標準日本語の語彙，文法体系とは異なる体系の言語のうち，日本語の日常会話では一般的に用いられない表現の箇所に付与する。発音は可能な範囲で聞きとり，片仮名で表記する。

> (O チャッチャッカンマンミヤネ)。@韓国語「待って ごめんね」か?

日本語とは異る体系の言語であっても，日常会話で一般的に使用される，あるいは理解できる表現にはタグ (O) は付与しない。

> ハロー ジャクソンとかいったら
> イエーイ
> アイムジャパニーズって言ってあげ (L れば良かっ (U た))。

■タグ (U)　聞き取りや語の判断に自信がない場合は，その範囲にタグ (U) を付与する。複数の候補がある場合は，候補を 「/」(スラッシュ) で区切り，可能性の高い順に列挙する。形態論情報はここで最も可能性が高いとされた語を解析の対象とする。

> (U 底/そこ) に付いている草や泥を取り除き
> 相手も何かきらいだ (U っていうんで)

■タグ (R)　個人情報保護などの観点から問題となる箇所については，その範囲にタグ (R) を付し，データを公開する際に仮名化・伏字化するなどの処理を施す。具体的には次のようなものが対象となる。

- 参加者を含む一般人の名前（愛称を含む，ただし著名人の名前は対象外）
- 参加者を含む一般人の所属する組織名（学校や職場の名称）など。
- 参加者を含む一般人の自宅や所属組織の住所など。
- 誹謗中傷や差別語のうち，特に問題になると判断されたもの。
- 会話者が非公開を希望した箇所。

■タグ (X)　身近な人達同士の会話では，そのコミュニティでのみ通用する語や略語が用いられることがあり，転記作業者が語を特定できないことがある。発話された表現が辞書に登録されていない場合，もしくは辞書に登録されていたとしても，その語の使用は文脈から考えて不自然である場合にタグ (X) 付与する。

> 九十 (X プチポ)。←なんらかの単位と推測できるが，そのような語が存在するか不明
> あのー (X ルトラ) のさ あのー ←文脈からブランド名か店名と推測できるが，不確定

協力者への聞きとり等によって語が判明した場合はこのタグは除かれる。最終的に語が判明しない場合は未知語とされて，タグ (U) が付与される。

■タグ (F)(A) 音の引き伸ばしや音の詰まりのある発話は，タグ「:」と「ー」(長音)，タグ「%」と「っ」(促音) のどちらで表記されているかによって，語が判定できる場合がある。例えば，「あの:」と表記された場合は，連体詞の「あの」と解析できる。「あのー」であればフィラーと判断できる。しかし，長音や促音を伴わない「あの」と「その」は，形態素解析において連体詞とフィラーを区別をすることが難しく，しかも会話中に頻出する。そのため，この 2 つの語形に対してのみ語を区別するためのタグを付与することにした。「その」がフィラーとして使用された場合は タグ (F) を，「あの」がフィラーではなく連体詞として使用された場合はタグ (A) を付与する。「その」は連体詞，「あの」はフィラーで用いられることが全般的に多かったため，作業効率の観点から，出現頻度の少ない使われ方をした場合に限定してタグを付与することにした。形態素解析後はこれらのタグを削除する。

2.6 転記テキストの例

図 1 に作業用転記テキストの例を示す。これは，ELAN で転記したものをタブ区切りテキストに変換したものである。1 行が 1 つの転記単位であり，発話の開始時間と終了時間が割り当てられている。句点「。」は発話単位の境界を示している。テキストには必要に応じて各種タグが付与されている。

図 1 転記テキストの例

発話者	開始時間	終了時間	テキスト
IC01	2502.617	2503.920	(U この前) 飲み会どこで飲んだの。
IC03	2504.661	2505.651	えっと 赤坂。
IC04	2507.718	2508.495	赤坂の
IC03	2508.791	2509.744	(L)
IC04	2509.287	2510.202	料亭。
IC03	2510.912	2511.480	(L いやいや)。
IC01	2511.432	2512.185	違う違う。
IC01	2512.749	2513.451	居酒屋。
IC03	2513.641	2514.236	(W イサカヤ|居酒屋)。
IC03	2515.464	2516.201	(X フタヘルモ)。
IC03	2516.999	2519.648	同期の (D ヒ)(D フ) 同期と二人で飲んだぐらいで。
IC05	2519.670	2521.713	芸能人もいっぱい歩いてるんじゃないですか。
IC05	2521.713	2522.074	外。
IC03	2522.237	2522.865	(W ンナ|そんな) 見る余裕。
IC03	2522.869	2526.534	もう 仕事終わったら家帰ることしか頭に (L ないです)。
IC05	2523.585	2524.039	ね:。
IC03	2526.541	2527.636	(L)
IC01	2530.214	2531.759	前ＴＢＳの地下で:
IC01	2532.456	2533.398	(R 仮名処理) さん ジュリー見た。

3. 転記作業工程

本節では，転記の作成工程について説明する。おおまかな流れを図 2 に示す。作業は大きく 5 つの工程に分けられる。以降に転記の 5 つの工程で行う作業について説明する。

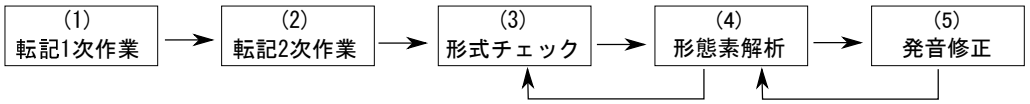


図 2　転記作業工程

■（1）転記 1 次作業　人手で会話音声の文字化と転記単位ごとの音声へのアライメント作業を行う。この作業は二つの方法で行う。一つは，上述の転記基準について知識を有する作業者が，ELAN を用いて文字化・タグ付け・音声へのアライメント作業を同時に行う方法である。もう一つは，いわゆる素起こしのレベルで文字化を外注した上で，転記基準について知識を有する作業者が Praat を用いて音声へのアライメント・タグ付け作業を行う方法である。後者は電話会話や 2 名の（比較的重複の少ない簡単な）会話を対象に行う。いずれの方法においても，次項の転記 2 次作業では，ELAN で映像音声を参照しながら修正作業をする。後者の方法を導入したのは，調査協力者へのフォローアップインタビューを行う時に文字化されたテキストが必要であり，ELAN での方法だけでは間に合わないためである。しかし実際に導入してみると，作業効率がかなり良いことから，電話や簡単な 2 名の会話についてはこの方法も積極的に採用している。現在は，文字化テキストを自動でアライメントした上で人手で修正する方法も検討中である。

■（2）転記 2 次作業　訓練を受けた作業者が，1 次作業で作成された転記テキストを対象に，ELAN 上で映像音声を参照しながら，文字化された内容や付与されたタグなどを確認・修正する。転記 1 次作業ではスピードを重視し，転記基準を完全に満たしたテキストの作成を作業者に求めていない。例えば，正しい転記テキストを作成するには，短単位の知識が必要だが，自信がない場合に詳しく調べる必要はないものとしている。発話単位の認定も，形式的・形態的に特定できる簡単なもののみの認定に留めている。また，基準ではタグはすべて半角，発音は片仮名で表記するが，作業者の入力しやすい文字（全角/平仮名）の使用も認めている。このうち，次の工程で自動変換可能なものを除き，訓練を受けた 2 次作業者が修正を行う。

■（3）形式チェック　転記テキストの形式的なチェックとして，以下の作業を行う。

- 文字種 (半角/全角，平仮名/片仮名) や典型的な転記エラーの自動修正
- タグの種類やタグの入れ子関係などの自動チェック・人手修正
- タグの範囲（短単位を範囲として付与するタグなど）の自動チェック・人手修正
- 発話単位の自動チェック・人手修正（「ケレドモ」節など形態的特徴に基づく自動チェック，発話単位長や発話単位中の無音時間などを参照したチェックなど）

修正作業は，ELAN，Praat，Excel 等のソフトウェアを用いて行う。それぞれで用いるファイル形式 (XML，TextGrid，タブ区切テキスト) を相互変換するスクリプトを整備しており，各作業ごとに最も効率の良い環境で作業できるようにしている。

■（4）形態素解析　上記 (3) の形式チェックを徹底するため，この段階で形態素解析を行う。形態素解析は，形態素解析器 MeCab(工藤ほか 2004) と形態素解析用辞書 UniDic を用いる。入力は発話単位とする。解析にあたっては以下の処理を行う。

- タグが付与されたテキストはそのまま解析できないため，タグを外して解析器に渡す。その際，タグ (D) が付与された言いよどみ要素，タグ (W) の左項（発音のなまけやエラーを含む実際の発音），タグ (U) の第 2 候補以降は解析器に渡さない。
- 短単位を範囲に付与されるタグについては，その情報を利用し，タグ付与範囲の開始・終了位置で必ず単語が分割されるようにする。
- 「(F その)」の品詞を「感動詞-フィラー」，「(A あの)」の品詞を「連体詞」にする。
- 解析器には渡さなかった要素（転記単位の開始・終了時刻の情報などを含む）を解析結果に埋め込み，転記テキストに記された情報を保持する。これにより，転記テキストが再生成できるようにする。なお，タグ (D) の範囲の品詞は「言いよどみ」とする。

以上の処理の結果を用いて，再び形式チェックを行う。

■（5）発音修正　この工程では，工程 (4) にて自動で付与された「発音形」を人手でチェック・修正する。修正対象となるのは，発音が一意に同定できない語 (例：一日「イチニチ/ツイタチ」，日本「ニホン/ニッポン」) や解析誤りによるものである。明らかな誤りや必ずしも誤りとは言えないが低頻度と思われる発音形を機械的に置換した上で，音を聴取しながら発音形の修正を行う。後者の作業は，発音形修正ツールを用いて効率化を図る。

修正した発音情報を参照することで，単位境界・付加情報も正しく解析されることがあるため，発音形修正の終了後，修正した発音に基づき，再び形態素解析を行う。

4. おわりに

本稿では，現在構築中の日本語日常会話コーパス (CEJC) の転記基準と作業工程について紹介した。現在のコーパス構築状況については小磯ほか (2017)，コーパスの特徴については，臼田ほか (2017) を参照されたい。CEJC の公開は，2021 年度末を予定している。また，2018 年度，50 時間のデータをモニター公開する予定である。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」の研究成果を報告したものである。コーパスの収録にご協力・ご参加くださった皆様に感謝します。

文　献

小磯花絵 (2017).「『日常会話コーパス』プロジェクトーコーパスに基づく話し言葉の多角的研

究－」　言語資源活用ワークショップ 2016 発表論文集.
小磯花絵・土屋智行・渡部涼子・横森大輔・相澤正夫・伝康晴 (2016).「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」　国立国語研究所論集:10, pp. 85–106.
小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』の構築」　言語処理学会年次大会発表論文集.
田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017a).「『日本語日常会話コーパス』構築における会話収録方法と進捗状況」　言語資源活用ワークショップ 2016 発表論文集.
田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017b).「『日本語日常会話コーパス』構築における会話収録方法」　言語処理学会年次大会発表論文集.
国立国語研究所 (2006).『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』　国立国語研究所.
伝康晴・榎本美香 (2014).「『千葉大学 3 人会話コーパス』使用説明書 Release 1」　. http://research.nii.ac.jp/src/files/Chiba3Party_manual.pdf
JDRI (2017).「『発話単位ラベリングマニュアル』version 2.1」　. http://www.jdri.org/open-data/ から入手可能
小椋秀樹 (2014).『書き言葉コーパス ―設計と構築―』, 第 4 章　pp. 68–86. 講座　日本語コーパス 2　朝倉書店.
工藤拓・山本薫・松本裕治 (2004).「Conditional Random Fields を用いた日本語形態素解析」情報処理学会研究報告自然言語処理（NL）, 2004:47, pp. 89–96.
臼田泰如・川端良子・徳永弘子・西川賢哉・小磯花絵 (2017).「『日本語日常会話コーパス』の転記基準と特徴について」　言語処理学会年次大会発表論文集.

関連 URL

『大規模日常会話コーパスに基づく話し言葉の多角的研究』プロジェクトのウェブサイト
http://pj.ninjal.ac.jp/conversation/

『現代日本語書き言葉均衡コーパス』と『分類語彙表』を利用した漢字 3 文字略熟語の抽出

山崎 誠（国立国語研究所研究系言語変化研究領域）†

Extraction of Clipped Compounds Comprised of Three Character Sino-Japanese Using “Balanced Corpus of Contemporary Written Japanese” and “Word List by Semantic Principles”

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

「政財界」「国内外」などの漢字 3 字で構成される「略熟語」と呼ばれる形式は，先行研究が少なく実態が明らかでない。国語辞書にも掲載されることが少ない。本発表では，現代日本語にはどのような略熟語が存在するかを『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）と『分類語彙表』を使って自動的に抽出することを試みた。具体的には，BCCWJ から，前後が非漢字という条件で漢字 3 文字連続を抜き出し，それらを構成する漢語の頻度および分類語彙表における意味番号を付与したデータを作成した。そこから，出現頻度が一定以上で，構成要素となる漢語の分類番号が一致するものとして 874 語を抽出した。内訳は「政財界」タイプ 656 語，「国内外」タイプ 297 語，重複が 79 語であった。目視で確認したところ，抽出された 3 字漢語には，略熟語でないものも多く，精度を高めるにはさらに別の条件が必要であることが分かった.

１．はじめに

「政財界」「国内外」などの主に漢字 3 文字からなる複合語は，「略熟語[1]」と呼ばれ，出現頻度はさほど多くないものの興味深い特徴を持っている。しかし，先行研究が少なくその実態は明らかではない。現象としての指摘は，国立国語研究所(1958: 12)(1959: 274-280)(1962: 19)(1985: 72-73)，野村(1974: 40)，村田(1998: 3,8)，玉村(2002: 200-223)(2005)，野村(2007: 207)，石井(2007)，修(2010)にあるが，詳しい分析を加えた文献は管見の限りでは玉村(2005)のみである[2]。本稿は現代日本語の略熟語の効率的な抽出方法を提案し，言語的な分析に資することを目的とする。

２．略熟語とは

略熟語は，共通の要素を持つ 2 つ（以上）の語が複合する際に，共通の要素と，非共通の要素をつなげて作る複合語のことである。共通の要素は，非共通の要素の前に来ることもあり，後に来ることもある。ほとんどの略熟語は漢字 3 文字からなる漢語であるが，「小中学

† yamazaki (at) ninjal.ac.jp

1 この名称は国立国語研究所(1985)を執筆した玉村文郎による命名と推測される。同書以前には名称を与えている文献は管見の限りでは見当たらない。

2 「略熟語」は「日本語研究・日本語教育文献データベース」（国立国語研究所）では 1 件もヒットしない（2017 年 1 月 31 日検索）。また，CiNii の全文検索（2017 年 1 月 31 日検索）では 2 件ヒットするが 1 件は，「省略熟語」という別の概念の一部が文字列検索でヒットした例であり，もう 1 件は石井(2007)の引用であった。

校」のような漢字 4 文字や，「春夏物（はるなつもの）」のような和語も若干存在する。

略熟語は，略語と熟語の両方の性質を持っている[3]。2 つ（以上）の語が複合する際に省略が起きるという点では，「東京大学」を「東大」と略すのと似ているが，略熟語は元となる 2 つ（以上）の語に共通の要素[4]を持つという点が特徴的である。すなわち，「祖父」と「祖母」を合わせて「祖父母」としたり，「登校」を「下校」を合わせて「登下校」としたりするような例である。「祖父母」の場合は，前要素が共通，「登下校」の場合は，後要素が共通である。この共通要素の位置に着目して略熟語には 2 つのタイプがあるとされている。1 つは「祖父母」のように前要素が共通のタイプ，もう 1 つは「登下校」のように後要素が共通のタイプである。これらの語は，以下の様な過程を経て作られることが容易に見て取れる。

祖父　＋　祖母　→　祖父母
登校　＋　下校　→　登下校

国立国語研究所(1985)では，「祖父母」タイプ，「登下校」タイプをそれぞれ，XAB 型，CDY 型と呼び，玉村(2005)では，α タイプ・β タイプと呼んでいる。これらの名称はやや便宜的な名付けであり，名称から内容が推測しにくいため，本稿では「祖父母」タイプを「前要素共通型（略熟語）」，「登下校」タイプを「後要素共通型（略熟語）」と呼ぶことにする。また，元となる語から略熟語が作られることを玉村(2002)(2005)に倣い，「縮約」と呼ぶことにする。

３．略熟語の性質

これまでに先行研究で挙げられた略熟語の例を通して，その性質を見てみよう。表 1 に先行研究に出て来た略熟語の例を挙げた。

表 1：先行研究で挙げられた略熟語の例[5]

文献	前要素共通型	後要素共通型
国立国語研究所(1958:12)	絹糸布　祖父母　養父母　国内外	給排水　魚獣肉　青少年　男女児 陶磁器　市町村長
国立国語研究所(1959:277-278)[6]	実父母　古金銀　好父母　祖父母 奉迎送　輸出入　養父子	鵜鴉猟　海陸軍　可否決　華士族 金銀貨　金銀鉱　区戸長　芸娼妓 原被告　紅白旗　佐尉官　上下院 上中等　静動産　大小豆　長次官 長次男　町村会　町村費　町村割 内外勤　動植物　府県会　府県官 府県社　府県庁　府県費　陸海軍

[3] 「略熟語」という名称自体も略熟語である（略語＋熟語→略熟語）。
[4] 共通の要素のことを玉村(2005)では「軸字」と呼んでいる。
[5] 掲出順は文献に現れた順である。
[6] この文献は，明治 10(1877)年 11 月 1 日から明治 11(1878)年 10 月 31 日の郵便報知新聞を対象に行った語彙調査（サンプリング調査）で，標本延べ語数は約 10 万語である。表 1 に掲げるにあたり，字体を新字体に改めた。この調査では，三字漢語全体を分析しており，その中で「ａ×（ｂ＋ｃ）」という構成を持つものが前要素共通型，「（ａ＋ｂ）×ｃ」という構造を持つものが後要素共通型に相当する。なお，「＋」は並列を，「×」は修飾を，（　）は第一次の結合をそれぞれ表す。

国立国語研究所(1962: 19)	絹糸布	給排水
野村(1974:40)	祖父母	重軽傷　陶磁器
国立国語研究所(1985)	輸出入　転出入　移出入　流出入　国内外	許認可　乳幼児　転退職　部課長　政財界　給排水　冷暖房　校園長　本支店　送受信　預貯金　投融資　視聴覚
村田(1998: 8)[7]		許認可　鉱工業　原材料　原燃料
玉村(2002: 222)	輸出入	判検事　（出入口[8]）
玉村(2005: 39-43)	輸出入　国内外　祖父母　養親子　関東西	内外傷　前後期　中近世　自他殺　死傷者　大公使　歩車道　理美容　地家裁　判検事　統廃合　行財政　乳幼児　編著書　青少年　編著書　理美容 耳鼻咽喉科医[9]　陸海空軍 医歯薬大＋進学ガイダンス　公私立大＋付属病院　京都＋金銀糸＋平箔＋工業組合　海軍＋兵経理機関学校＋受験準備要領　定期券＋出入場＋確認システム　医歯薬理工大＋受験　全国市町村長＋会議
石井(2007)		乳幼児
修(2010: 49)	輸出入　移出入　流出入　国内外	行財政 出入り口[10] 中高生[11] 乳幼児 入退室　入退場　入退院　緑黄色　中高年　預貯金　送受信　与野党　政財界　転退職　部課長　給排水　冷暖房　校園長　本支店　投融資　視聴覚　判検事

表 1 を概観すると，おおよそ以下のようなことが見て取れる。

(1)ほとんどが漢字 3 文字である。「市町村長」や「陸海空軍」のような例もあるが，これらは 3 字漢語に比べると少ないと推測される。

(2)前要素共通型が少なく，後要素共通型が多い。表 1 における前要素共通型の異なりは 15

7 この文献は，経済学入門書に現れる漢字三字の専門用語の調査である。専門用語全体では，異なり 1,015 語（延べ 1,566 語）でそのうち漢字 3 字のものは 130 語である（村田 1998 : 1)。

8 玉村(2002: 222)では，「出入口」は，「「出口」＋「入口」の縮約なのか，「出入り」＋「口」の複合なのかの判定がむずかしい」としている。

9 この語以降は玉村(2005: 41)で「多次縮約」と位置付けられ，3 語以上が縮約したり，2 語の縮約に他の成分が結合したりする例である。「＋」記号は原文にあるとおりである。

10 修(2010: 49)に「「出入り」＋「口」の説もある」と注記あり。

11 修(2010: 49)に「稀にある 3 字熟語によるもの」と注記あり。

語であるのに対して，後要素共通型の異なりは，71 語である（玉村 2005 の多次縮約の例を除く）。

(3)語種は，略熟語及びそれらを構成する語も含めてほとんどが漢語である。「出入り口」が例外となるが，これには「出入り」＋「口」という，略熟語でない解釈も成り立つ。

(4)略熟語になるときの並び順がほぼ決まってる。すなわち，「輸出入」はあるが，「輸入出」はない。同様に「送受信」はあるが「受送信」はない。表 1 では，国立国語研究所(1959)の例として「海陸軍」と「陸海軍」が挙がっているが，これは例外と見られる。

(5)略熟語の元となる語同士は意味的に対義・類義関係にある。すなわち「冷暖房」であれば，その元となる「冷房」と「暖房」は対義語関係にあり，「預貯金」の元となる「預金」と「貯金」とは類義語関係にある。いずれにしても意味的に近い関係にある。

ほぼ同様のことが玉村(2002:222)でも指摘されている。そこでは，縮約の起きる条件として，以下のように述べられている。

> この縮約は，①縮約される原単位の個々に，共通の語彙成分が同一位置に存すること，②原単位が近似の意味分野の語であること，③同じ場面で原単位が頻用されること，の 3 条件が揃ったときに生まれると考えられる（「人事」と「炊事」から「人炊事」は生まれないし，「月曜」と「月給」から「月曜給」は造れない）。

ただし，これらの条件は必要条件であって，これらの条件が満たされても，ただちに，略熟語が出来るわけではない。玉村(2005: 41-42)では以下のように述べている。

> 「主義」と「主張」の両語はよくセットで使われ、かつ「主」という漢字を共通に有している上、語義も近接している。従って略熟語になりやすいと考えられるが、現実には「主義張」や「主張義」という語は用いられていない。「結離婚」「心肺臓」「親反日」「恒惑星」なども現在までのところ縮約形を生み出していない。必要条件を具備している 2 成分、3 成分であっても略熟語を生み出していないのは、何か十分条件を充たしていない点があるためであろう。

また，玉村(2005: 44)では，略熟語が生まれる契機として主に文章論的な観点から，書記スペースの節約，元となる語が近接して使われている，元となる語が頻用されている等の条件が指摘されている。

本稿では，主に玉村(2002:222)の①と②の条件をもとに略熟語を自動的に抽出することを試みる。

４．略熟語の抽出

４.１　データと方法

略熟語の抽出方法は大きく二通り考えられる。一つは玉村(2002)の条件①②を満たす 2 字漢語から人工的に略熟語を作り，それがコーパスに現れるかどうか（実際に使われているかどうか）を調べる方法，もう 1 つは，実際に使われた略熟語の候補を網羅的に収集し，そこから玉村(2002)の①②の条件に合うものを抜き出す方法である。前者は略熟語の可能性の最大を捉えたもので，実際に存在しないものも含まれてくる。仮に玉村(2002)の条件①②に当

てはまる語が 1,000 組あったとすると，それだけで 100 万通りの略熟語の候補ができることになる。処理の簡便さも考慮して今回は後者の方法を採ることにした。

データとして，『現代日本語書き言葉均衡コーパス[12]』（以下，BCCWJ）と『分類語彙表増補改定版データベース』（以下，分類語彙表）を利用する。BCCWJ（約 1 億語）の規模で略熟語を調査した例はないので，この規模で略熟語がどれくらい現れるのかをまず確かめるためである。分類語彙表は，意味の類似性を判定するために用いる。

略熟語を抽出する手順は以下のとおりである。

(1)BCCWJ-DVD に収録された長単位の TSV ファイルを用いて，書字形出現形を連結した文字列を作り，そこから，漢字 3 文字連続で前後が非漢字である文字列を抜き出す。その結果，延べ 2,698,888 文字列，異なり 309,652 文字列が得られた。これらが略熟語の候補[13]である。以降，便宜的に候補語と呼ぶ。

(2)309,652 の候補語に対して，略熟語であった場合に想定される略熟語を作る元となる２字漢語を 4 つ用意する（実際には重複があるので 3 つでよい）。

例えば，候補語が「教職員」であった場合，以下のようになる。

候補語	前要素共通型の場合		後要素共通型の場合	
教職員	教職	教員	教員	職員

記号化して表すと，候補語が「ＡＢＣ」であったとすると，「ＡＢ」「ＡＣ」「ＢＣ」の 3 つが必要になる。

候補語	前要素共通型の場合		後要素共通型の場合	
ＡＢＣ	ＡＢ	ＡＣ	ＡＣ	ＢＣ

その上で，候補語には，BCCWJ 長単位語彙表データ[14]からの頻度・読み・語種の情報を，元となる漢語候補のそれぞれには，BCCWJ 短単位語彙表データ[15]から頻度・読み・語種の情報と分類語彙表の分類番号[16]を付与した。

４.２　略熟語の判定

上記(2)で示した略熟語の候補語データのイメージを表 2 に示す。実際には表は横方向に展開するが，見やすさのため縦方向に表示している。

表 2 では結果のうち，いくつかのパターンを示した。「輸出入」は前要素共通型，「原材料」と「教職員」は後要素共通型であるが，略熟語と判定されるかどうかが異なる。「屋内外」は両方の型に判定されたものである。

まず，「輸出入」は想定される元の漢語が「輸出」「輸入」「出入」となるが，「輸出」と「輸入」の分類番号が 1.3760 で一致するため，前要素共通型の略熟語と正しく判定される。次に，「原材料」は，想定される元の漢語が「原材」「原料」「材料」となり，「原料」の分類番号が 1.4100，「材料」の分類番号が 1.1040/1.4100 で，共通する 1.4100 があるため，後要素共通型と正しく判定される。「教職員」は，元となる漢語が「教職」「教員」「職員」であるが，「教職」の分類番号 1.3800，「教員」の分類番号 1.2410，「職員」の分類番号が

---

12 使用したバージョンは BCCWJ-NT ver 1.1 である。

13 もちろんこれらのほとんどは略熟語でない 3 字漢語である。出現頻度順では，「可能性」(13,577 回)，「日本人」(10,667 回)，「具体的」(9,251 回)と続く。頻度順で最初に現れる略熟語は 179 位の「青少年」(1,165 回)である。

14 BCCWJ_frequencylist_luw_ver1_0.tsv （http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html）

15 BCCWJ_frequencylist_suw_ver1_0.tsv （http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html）

16 複数の分類項目に出現している場合は，それらを列挙した。

1.2400 で分類番号の一致はなく，略熟語とは判定されないケースである。ただし，分類番号の一致をやや緩和して上 4 桁までの一致でよいとするなら，この例は後要素共通型と正しく判定される。ただし，意味の一致の基準を緩めると適切でない例も混じってくるデメリットがある。「屋内外」は，想定される元の漢語が「屋内」「屋外」「内外」の 3 つである。「屋内」の分類番号が 1.1770,「屋外」の分類番号が 1.1770,「内外」の分類番号が 1.1770/1.1920 となっており，3 つ元の漢語の全てに 1.1770 が共通しているため，前要素共通型と後要素共通型のどちらにも当てはまると判定される（実際には前要素共通型である）。

このように，正しく判定されるもの，類義の条件を緩和すれば正しく判定されるもの，両方に当てはまるものがあるが，実は最も多かったのは，略熟語ではないものである。例えば，「不可能」は，元となる漢語の部類番号が，それぞれ「不可」1.1030/1.1332/1.1346，「不能」1.1346/1.3421/3.1346/3.3421，「可能」1.1346/3.1346 であり，1.1346 が共通のため，前要素共通型と判定され，かつ，1.1346 と 3.1346 が共通のため後要素共通型と判定されるが，実際にはそのどちらでもない。

表 2：略熟語の候補語データ（例）

候補語	「輸出入」	「原材料」	「教職員」	「屋内外」
出現頻度	224	482	344	8
長単位頻度	197	449	331	7
読み	ユシュツニュウ	ゲンザイリョウ	キョウショクイン	オクナイガイ
語種	漢	漢	漢	漢
AB	輸出	原材	教職	屋内
AB 頻度	6147	95	161	541
AB 読み	ユシュツ	ゲンザイ	キョウショク	オクナイ
AB 語種	漢	漢	漢	漢
AB 分類番号	1.3760	no[17]	1.3800	1.1770
AC	輸入	原料	教員	屋外
AC 頻度	8581	2053	1947	810
AC 読み	ユニュウ	ゲンリョウ	キョウイン	オクガイ
AC 語種	漢	漢	漢	漢
AC 分類番号	1.3760	1.4100	1.2410	1.1770
BC	出入	材料	職員	内外
BC 頻度	84	8892	8601	1329
BC 読み	シュツニュウ	ザイリョウ	ショクイン	ナイガイ
BC 語種	漢	漢	漢	漢
BC 分類番号	1.1530	1.1040/1.4100	1.2400	1.1770/1.1920

[17] 分類語彙表に一致する語がなかったことを意味する。

５　結果

略熟語の候補語データから，長単位頻度 5 以上，長単位の語種及び元となる漢語の語種が漢語であり，分類語彙表番号の 5 桁まで一致しているものという条件により以下の略熟語候補が抽出された。

前要素共通型　297 候補語，後要素共通型　656 候補語，重複　79 候補語

これらの全体を表 3，表 4 に示す。

表 3：前要素共通型と判定されたもの[18]（297 候補語）

悪質性	医学科	異質性	一家族	一世代	一期生	屋内外	開港場	外周部	開幕戦
加減算	加減速	家族内	学級会	下流部	感覚性	感受性	元利金	義侠心	義兄弟
騎乗馬	疑問点	給食費	九千百	九百十	給付費	給与費	教育学	競走馬	競争馬
記録帳	近視眼	近隣国	軍部隊	計算上	計算量	系統図	競馬馬	現在世	減収額
減少額	減少量	減税額	現世代	合意点	公債金	構成図	後世代	構造図	硬直化
抗日戦	高年齢	鉱物油	後方部	国内外	五十六	五千十	五千百	国家内	五百十
五百六	在位中	在京中	最終期	最初期	在職中	在世中	在籍中	在宅中	最多勝
在任中	在米中	山岳地	山間地	三千百	三百十	残余金	地獄界	室内外	実父母
市内外	資本金	社内外	謝礼金	重罪犯	重心点	執着心	終電車	重犯罪	銃砲弾
主幹事	出資金	出場校	出走馬	種目別	種類別	純利益	商工業	上申書	小中学
商店会	上流部	諸行事	食材料	食料品	食糧品	女性史	女性性	人格性	信仰心
人口数	真実性	真理性	頭蓋骨	図面上	性差別	西南部	精白米	西北部	全会員
全額国	線形状	前後半	全国紙	全国土	戦時中	全社員	全社会	全身部	前世代
全体会	全党員	全部隊	前方部	専門科	増加額	増加量	総監督	増減額	増減税
草書体	増税額	俗世界	俗世間	祖父母	対応等	耐火性	耐寒性	耐久性	大祭典
対策案	耐酸性	耐食性	耐震性	耐水性	耐熱性	単語数	男性性	地球上	地層上
地層中	地層面	地盤面	地表上	地表面	地方区	中央部	中間点	中間部	忠義心
中心核	中心間	中心軸	中心点	中心部	中枢部	忠誠心	中流部	長寿命	貯蓄金
地理学	地理上	賃貸借	定員数	停止車	適合度	敵陣地	適正度	電気灯	電磁界
電磁気	電磁場	電磁波	電話器	同一種	同一性	等価値	同業種	同時期	同質性
同種類	答申案	頭頂部	同等性	東南部	東北部	特殊性	土日曜	内服用	南西部
南東部	難破船	難問題	二十三	二千三	二千十	二千百	二百三	二百十	入院室
入国港	人間界	認定証	年度間	年度中	農作業	納入金	納付金	農牧業	排出水
配水管	配送車	配電線	八十九	発酵熱	八千百	八百十	半額分	半分量	鼻孔下
病原体	不可能	不正義	復古調	不適当	不満足	墳丘墓	分極化	分散化	文章体
分水界	文明化	閉会式	閉校式	閉講式	変異性	変化形	返還金	返済金	変動性
方位角	防衛戦	法規制	法制度	北西部	北東部	本図表	未完成	未決定	民主政
名演技	猛攻撃	目的的	文科系	門下生	役職員	有害性	輸出入	幼少時	幼少年
養殖魚	幼年時	養父母	余剰分	預託金	来場所	理科学	理科系	陸海軍	理数科

18 網掛けは，実際に前要素共通型と筆者が判断したもの。下線は，後要素共通型とも判定されたもの。

理数系　理想論　律令法　両側面　老朽化　六千十　六百十

表 4　後要素共通型と判定されたもの[19]（656 候補語）

圧痛点	暗褐色	暗紅色	暗灰色	医学科	異業種	医局長	移住民	異常時	異世界
一家族	一画面	一議員	一機種	一業種	一原因	一三塁	一時代	一車種	一集団
一宗派	一手法	一世代	一側面	一大事	一断面	一昼夜	一二塁	一品種	一分派
一兵士	一方策	一方法	一要因	胃腸炎	一流派	一箇所	異邦人	右左折	英語学
英数字	英文学	英文名	英文字	英訳文	駅職員	絵地図	演奏者	王太子	屋内外
会会員	会会長	会議場	海空軍	外国語	会社員	海水中	灰白色	外壁面	家屋内
歌曲集	各委員	各一部	学院長	学園長	各会社	各議員	各機種	各業界	各業種
各産地	学識者	各社員	各宗派	各職員	各職種	各成員	格闘技	各党派	各品種
学部長	各流派	加減算	加減速	学校外	学校区	学校長	学校内	下背部	下腹部
歌謡曲	管区長	管弦楽	漢文学	喜歌劇	奇関数	義兄弟	既裁定	騎乗馬	義祖母
客貨車	客馬車	旧街道	旧刑法	旧憲法	旧国道	旧国名	旧社名	急性病	旧町名
休廃止	給排水	旧仏教	旧民法	共依存	行財政	強大国	胸腹部	強変化	局次長
許認可	金銀貨	近現代	金鉱山	金細工	金政権	近隣国	区隊長	句読点	軍部隊
傾向性	傾斜角	劇世界	原漢文	県議会	原原種	原材料	原資料	現世代	原地図
剣闘士	高確率	高気圧	後胸部	高金利	航空路	高血圧	鉱工業	高高度	高効率
公社債	好事例	後世代	降積雪	剛速球	広帯域	高体温	皇太子	高地価	高電圧
後頭部	高得点	高能率	高倍率	後半期	後半生	好不況	好不調	公法人	公法的
公民権	高利率	後近代	国語学	国史学	国内外	国文学	御高名	古社寺	御赦免
誤使用	個人性	国家内	国境内	国公法	国公立	今学期	紺無地	再逆転	再協議
最初期	再点検	再登録	再発見	再反論	最末期	雑海草	参議院	三事業	詩歌集
市議会	資機材	資金力	四死球	市場外	死傷者	市場内	視触診	次世代	視知覚
視聴覚	実作業	室内外	実父母	治天下	市内外	詩文学	詩文集	死亡者	弱小国
社内外	受委託	終楽章	習慣性	衆議院	重軽傷	終着駅	終電車	重犯罪	銃砲声
銃砲弾	充放電	終末期	終列車	儒学者	祝祭日	主原因	手術中	出退勤	術直後
出融資	主任務	受発信	受発注	主要因	純損益	純利益	小王国	小会社	上下肢
商工業	小住宅	上前端	小太鼓	小中学	商人道	上腹部	小部隊	小文字	小論文
諸王国	諸外国	諸学説	所管外	食材料	食事後	助数詞	諸大国	初中級	助動詞
所要員	諸流派	真意義	新街道	新改訳	新幹線	新刊本	新喜劇	新機種	新契約
新建築	新憲法	新国劇	心雑音	新施設	新車種	新宗教	新首都	新制作	新品種
新仏教	心理学	新路線	垂直線	水分量	水平面	数文字	正会員	聖画像	政官界
税金額	政財界	税財源	性差別	正社員	青少年	政省令	正職員	青壮年	西南部
性非行	西北部	生没年	正礼装	聖礼典	全音域	全学校	前下部	全画面	全艦隊
全教科	前胸部	前近代	線形状	前後半	前後編	前後席	前後輪	全色盲	前時代
全重量	全盛期	前世代	船隊長	全地域	前頭部	全半壊	前半期	全部隊	全領域
全歴史	線路上	素因数	躁鬱病	総雨量	増改築	総画数	増減額	総件数	増減税
総合計	総戸数	総字数	総重量	贈収賄	送受信	総首長	草書体	総大会	総定数

[19] 網掛けは，実際に後要素共通型と筆者が判断したもの。下線は，前要素共通型とも判定されたもの。

総日数	総婦長	総容量	総連合	俗世界	俗世間	祖父母	村議会	大偉業	大異変
耐陰性	大運河	大宴会	大会戦	大街道	大学卒	大艦隊	堆厩肥	大苦戦	大激戦
大傑作	大決戦	対抗戦	大混戦	大祭典	大災厄	大作戦	大惨事	大地震	代執行
大集会	大小便	大接戦	大大名	大茶会	大帝国	大動乱	大納会	大破局	大爆笑
大迫力	大発会	大反乱	大部隊	大編隊	大暴風	多次元	他諸国	多品種	単独行
短母音	地下界	地盤面	地表面	忠義心	中高年	中上級	中心性	忠誠心	中長期
中低速	中毒性	中年期	町会長	町議会	町村会	町村長	町村民	長母音	直線上
著作者	地理学	珍回答	賃貸借	通行人	通常人	定員数	低確率	低気圧	定休日
低金利	低血圧	低山地	低湿地	低水温	低体温	低倍率	敵陣地	敵戦艦	敵本陣
適用例	鉄軌道	展延性	電磁界	電磁気	電磁力	転廃業	同一種	当会社	同機種
同基地	同業種	頭頸部	登下校	同公社	陶磁器	同時期	糖脂質	同種類	動植物
同女性	動静脈	東南端	東南部	統廃合	同比率	東北部	動名詞	投融資	同連盟
特定性	特別製	都市内	都市民	土日曜	内外装	内水面	内側面	内表面	内服用
南西端	南西部	南東端	南東部	南南東	南北朝	難問題	二三塁	二次元	入会費
入出金	入出力	乳幼児	妊産婦	熱容量	熱流量	年金額	年税額	年利率	農漁業
農漁村	農漁民	農工具	農耕地	農作業	農山村	農民兵	農用地	農林業	売買春
白砂糖	発着駅	半周期	反対論	悲喜劇	微苦笑	非合理	微震動	微振動	微積分
鼻濁音	秘伝書	非論理	不安定	部員数	風水害	不穏当	部会長	部課長	不可能
不完全	不協和	副主題	不公正	不作法	不正義	不誠実	不戦敗	部族長	不存在
部隊長	不注意	仏座像	物質量	物理性	不定休	不適当	不適法	不道徳	不必要
不分明	不満足	部門内	不愉快	不用意	文書中	米海軍	米空軍	米陸軍	変異種
編著者	法医学	法規定	法原則	法制定	法哲学	北西端	北西部	北東部	母国語
歩車道	本街道	本原則	本個体	本事業	本事件	本自体	本支店	本数字	本図表
本法案	本本堂	本問題	本訳書	本論文	未解決	未確定	未完成	未既婚	未決定
未公刊	未指定	未成熟	無香料	無際限	無彩色	無作為	無作法	無定形	無点灯
名演技	名義人	名女優	名選手	名文句	猛攻撃	猛反撃	役職員	薬理学	有彩色
養嗣子	幼少時	要相談	養祖母	陽電子	用筆法	養父母	予警報	預貯金	与野党
来訪客	理化学	理科学	利活用	陸海軍	離着陸	離転職	略系図	略地図	略礼装
榴散弾	領域内	両議院	領国内	両側面	両大国	林産業	例大祭	冷暖房	老病死
論争点	論文集	和漢文	和定食	和洋裁	和洋室				

表 3 で前要素共通型と判定された 297 語中，実際に前要素共通型であったのは，網掛けした 9 語のみであった（適合率約 3%）。一方，表 4 で後要素共通型と判定された 656 候補語中，実際に後要素共通型であったのは網掛けした 119 語であった（適合率約 18.1%）。

６．おわりに

本稿では，BCCWJ と分類語彙表を用いて漢字 3 文字の略熟語の抽出を試みた。コーパスに実際に出現した漢字 3 字文字列を候補語とし，そこから略熟語の元となる漢語を想定し，それらの意味分野の類似性を利用して略熟語の判定を行った場合，前要素共通型の略熟語は適合率約 3%，後要素共通型の略熟語は適合率約 18.1%の精度で抽出することができた。精度はかなり低いと言わざるを得ない。

今後の課題として，抽出方法の見直しがある。例えば，今回は漢字 3 文字列で前後が非漢

字という条件であったが，この場合，「駐停車禁止」のような漢字文字列に埋もれた略熟語が抽出できない。実際に「駐停車」は少納言では 20 件ヒットするが，今回の方法では出現頻度 4，BCCWJ 長単位語彙表での頻度は 2 だった。

意味分野の類似性を判断する分類語彙表も対義となる分類番号が一致していたり，1 つだけずれていたりして今回のような調査のためには統一を欠いている。類義語・対義語辞書の併用などが考えられる。

また，後要素共通型と判定された候補語の中には「各」「旧」「新」「大」「不」「本」などの接頭辞を持つものが多く含まれており，これらの多くが略熟語ではないことから，別処理として検討する必要もあろう。

謝　辞

本研究は国立国語研究所コーパス開発センターのプロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」により行われたものである。

文　献

石井正彦(2007)「略語」飛田良文他編『日本語学研究事典』171　東京：明治書院.
国立国語研究所(1958)『現代雑誌の語彙調査　総合雑誌の用語　後編』東京：秀英出版.
国立国語研究所(1959)『明治初期の新聞の用語』東京：秀英出版.
国立国語研究所(1962)『現代雑誌九十種の用語用字 第一分冊：総記および語彙表』東京：秀英出版.
国立国語研究所(1985)『語彙の研究と教育（下）』執筆：玉村文郎，東京：秀英出版.
修徳健(2010)「略熟語の中日対照」『北京日本学術センター25 周年記念シンポジウム世界日本学研究趨勢与合作論文集（予稿集）』49-50.
玉村文郎(2002)「対照語彙論」『朝倉日本語新講座 4 語彙・意味』208-235，東京：朝倉書店.
玉村文郎（2005）「日本語とドイツ語の複合語の対照」Viktoria Eschbach-Szabo, Yoko Koyama-Siebert, Martina Ebi (Hg.) “Ibunka to no deai. Sekai no naka no Nihon to Doitsu” 33-45，BUNKA - WENHUA. Tübinger Ostasiatische Forschungen Bd. 11 (volume11) Münster:Lit Verlag.
野村雅昭(1974)「三字漢語の構造」『電子計算機による国語研究VI』37-62.
野村雅昭(2007)「語彙・文字」『 国語と国文学』84(5)，202-210.
村田年(1998)「経済学専門用語三字漢語の語構成―専門分野導入期の日本語教育の方法を探る―」『日本語と日本語教育』26，1-11，慶應義塾大学日本語・日本文化教育センター.

関連 URL

『現代日本語書き言葉均衡コーパス』
http://pj.ninjal.ac.jp/corpus_center/bccwj/
『現代日本語書き言葉均衡コーパス』語彙表
http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html
分類語彙表増補改訂版データベース
http://pj.ninjal.ac.jp/corpus_center/goihyo.html

名詞項構造付与データの構築

竹内 孔一 (岡山大学大学院自然科学研究科) *

Construction of Argument Structure Data for Japanese Noun Utilizing Predicate-Argument Structure

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

要旨

含意認識タスクなど言語処理での文間の表現を取り扱う際、名詞の意味的な関係を捉える必要がある。言語学の分析から名詞の中には名詞の意味を補完する外部情報が必要なものが分かっており、生成語彙における特質構造 (クオリア構造) として記述することが提案されている。また言語資源では NomBank に代表されるように名詞の項構造を事例とともに構築されている。本研究では、先行研究で提案された特質構造を利用した名詞の項構造データを基に言語処理の観点からより形式化した構築法を提案する。具体的には名詞の項構造の例文を構築するとともに、項を同定し、述語との関係を項構造を通して結び付ける記述枠組である。述語のデータとして述語項構造シソーラスを利用し、NTCIR の RITE-2 で出現した名詞を対象に項構造の例文および対応する述語と項の関係を記述したデータを構築した。本稿では、記述枠組、および具体的に構築した名詞項構造データの事例を説明すると共に、付与での問題点や現状について記述する。

1. はじめに

含意認識タスクなど言語処理において、文同士の意味的な関係を捉えるには名詞の項構造の関係を捉える必要性がある。例えば下記の 2 つの例文は(1)文 (t2) の内容を文 (t1) が含意する例である。

(t1) シアターで上映される映像作品「もしも月がなかったら」は、米国の天体物理学者ニール・F・カミンズ教授の同名の著書をもとにした科学エンターテイメントである。

(t2) 「もしも月がなかったら」は、ニール・F・カミンズ著の書籍である。

ここでは名詞として「作品」「著書」「書籍」といった非飽和名詞が出現し、作品名と作者との関係を結び付ける役割を果たしている。例えば「X の著書は Y」という表現があった場合、「X」の部分に人に関する表現があり、「Y」に何かの名称が現れていれば、X が著者であり Y がその著作物であることを表す。こうした構造は名詞「著書」が誰の著書なのかといった情報を外部に必要とするために生まれると考えられ、名詞の構文に関する研究が行われてきた (西

* koichi 'at' cl.cs.okayama-u.ac.jp

(1) NTCIR (http://research.nii.ac.jp/ntcir/index-ja.html) の含意認識タスク (RITE2) より引用。

山 2003, Meyers et al. 2004, 庵 2007, 西山 2013, 影山 2011)。

そこで本研究では先行研究の分析を基に名詞の項構造データの構築を行う。基本的な記述枠組として、影山 (2011) が提案する生成語彙におけるクオリア構造による手法を利用する。表 1 に「作者」「主役」のクオリア構造の例を記す。

表 1　「作者」「主役」のクオリア構造

	作者	主役
外的分類 (formal)	人間 (x)	人間 (x)
目的・機能 (telic)		y が芝居や映画で劇中の人物を演じる
成り立ち (agentive)	x が w を書いた	y が [w] の主要人物の役をつとめる

表 1 では「外的分類」で名詞の分類を記述し、「目的・機能」でその名詞に役割的な機能を記述しする。「成り立ち」で名詞の意味の中である事態を引き起こした時のみその名詞の意味が成立する場合の意味を記述する。この構造の中で、特に「成り立ち」の部分は含意認識タスクで有効に利用できると考えられる。例えば「作者」の場合、「『重力ピエロ』の作者は伊坂幸太郎である」という表現は「伊坂幸太郎は『重力ピエロ』を書いた」の言い換えと考えられるが、この時「作者」という言葉と、動詞の「書く」との関係が処理できなければ、これらの含意関係を言語処理で扱うことは難しい。クオリア構造では「作者」と「書く」の関係が記述されているため、これを利用することができれば名詞に関連した言い換えを扱うことが可能となる。

しかしながら一方で、上記のクオリア構造をデータとして構築するには処理の観点から難しい点がある。例えば「外的分類」など名詞に対してどのような分類を仮定すべきかは処理タスクの応用によって異なると考えられる。さらに「目的・機能」および「成り立ち」において、述語との関係を記述する際、記号 (y や w) を埋め込んだ文による表現では、直接名詞と動詞の表現間の関係を取り出すことができないため、利用が難しい。

そこで本研究では名詞の項構造を例文として記述し、例文を基に述語との項関係を構造化する手法を提案する。こうした名詞の項構造の形式化は異なるタスクではあるが質問応答システム Watson における言語パターンとして prolog で規則化されており[(2)]既に言語処理として利用されている (Lally et al. 2012)。次節で具体的な構造化の手法について記述する。

2. クオリア構造を基にした名詞項構造データの枠組

クオリア構造は名詞の意味を記述することが目的であるため、「外的分類」や意味に合わせた記述を行うため、多くの情報を記述できる。例えば表 1 では「作者」は著者が作品を書かないと作者とは言わないため「成り立ち」では「書いた」と過去形で記述する。しかしながら本研究では言語処理の言い換えに焦点を置くため、名詞の意味構造を網羅的に扱うのではなく、

(2) authorOf(Author,Composition) という述語で構文情報や品詞情報を基に文からパターンを同定する。

名詞の項構造を述語で言い換える部分に着目して項構造を記述する。つまり名詞と動詞のように品詞をまたいだ表現を構造的に収集することとする。これらを実現するために構築手順を下記のように示し、事例および各手順での背景と問題点について記述する。

名詞項構造データの構築手順

1) RITE-2 から名詞を取り出し付与する名詞を決める。
2) 名詞 (Y) に対して「X の Y は Z だ」という例文を作成する。
3) 例文の X や Z の項に対して番号を付与 (arg0, arg1 など) して固定する。
4) 各項に対して、項どうしを表現できる述語を述語項構造シソーラスから選択する。さらに項に対して、シソーラス上で定義されている意味役割を付与する。

例えば、**1)** で RITE-2 に現れる名詞で「店員」を選択し、**2)** で例文として「そのコンビニの店員は太郎だ」を作成し、**3)** で項構造として「そのコンビニ」を arg1、「店員」を arg0 と定義する。次に **4)** で、この例文の項に対して、意味が対応する述語を述語シソーラスから探し、「働く」(「労働の語義」) を結び付ける。対応する意味役割として「そのコンビニ」を [場所][(3)] とし、「太郎」に [動作主] を付与する。この様子を図 1 に示す。

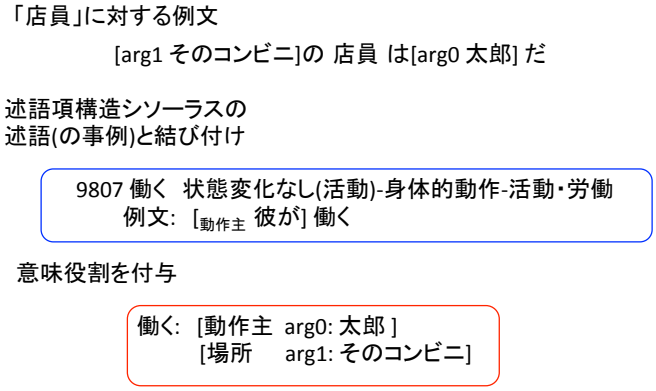


図 1 「店員」に対する名詞の項構造付与の事例

この名詞の項構造により、「そのコンビニの店員は太郎だ」と「太郎はそのコンビニで働く」の言い換え関係を取り出すことが可能になる。こうした名詞の項構造データを構築するための上記の手順の背景と問題点を下記に示す。

まず **1)** の手順を置く理由は、作成した名詞項構造データを処理として利用することを仮定しているためである。これにより含意認識タスクを例に、有効な名詞項構造データの構造を洗練できる。付与作業の問題点としては、非飽和名詞や相対名詞に作業者が気がつかず、付与を取りこぼす点である。処理の観点から付与すべき名詞を見直す枠組が必要となる。

手順 **2)** の理由は、例文があると機械学習等を利用したアプローチが期待できるためである。また後の項構造を付与して項同士の関係を結び付ける際に、単なる記号のリンクだけでなく、事例があることで、人手の付与作業の判断に貢献すると考えられる。また RITE-2 の文そのも

(3) 意味役割ラベルを [] で示す。

ので名詞の項構造を付与しなかったのは、項が出現していない場合が多いことが理由である．項構造データとして名詞の項を表す典型的な例を構築するために作成する手法を選んだ。

一方、付与作業での問題点としては項構造の例文がどのようなものか名詞の構文に関する先行研究を作業者が理解する必要がある。例えば、「X の Y は Z だ」の構文の X は Y の項でなくてはならない。例えば「あの頃の著者」など時間を表す表現が X に来た場合は Y の項では無い (西山 2003)。また Z の項は名詞でなくてはならない。しかしながら、項を持つと考えられる名詞でも Z が表現できない場合がある。例えば「性格」はの項は「彼女」など人や動物などが考えられるが Z にあたる表現は形容詞か形容動詞が普通である (「彼女の性格は穏やかだ」)。このように Z が名詞で無い場合は現段階ではあてはまる表現で記述するに留めている。

手順 **3)** では作成した例文に対して項を同定して、番号ベースの意味役割を付与する。この時、PropBank (Kingsbury et al. 2002) に従って、人にあたる意味役割を arg0、それ以外の項は直接的なものと考えられる物から順に 1、2、と番号をつけて付与する。これは名詞の項は上記の例で示すように複雑であるため、「動作主」など一貫した分類の見通しが立たないためである。一方で、例文に対して arg0、arg1 と固定することで、後のデータ化において、扱いを固定化できる。よって構文が異なっても番号ベースの項による意味役割で項を同定できるため言い換えなどの処理が可能となる。例えば「主役」に対して「[arg1 その映画] の主役は [arg0 太郎] だ」のように名詞項構造データに記録されていれば「主役」の項が arg1 で、主役そのものが arg0 であることがわかる。さらに異なる構文であっても、「[arg0 彼] が [arg1 この舞台] の主役だ」のように意味的な関係を意味役割で関係付けすることができる。

手順 **4)** では対象とする名詞の意味に関係し、例文の項をとる述語を語義の違いを考慮して述語項構造シソーラスから選択する。その際、項の役割や付加詞も含めた 72 種類の意味役割の中から選択して、対応付けする。これにより動詞と名詞を同定できる。一方、問題点としては述語シソーラスに登録されている述語は約 1.1 万語 (2.2 万事例) であり、登録されていないものは記述できない。また、名詞に対する述語を付与するのは容易ではない作業である。

次節では上記の作業手順による付与作業を行った結果と問題点について明らかにする。

3. 名詞項構造付与作業

3.1 付与作業と付与データ

付与対象のデータは含意認識タスク RITE-2(バイナリタスク) の文を MeCab[(4)] を利用して形態素解析を行い、その中から名詞を全て取り出した (3713 語)。次に項を持つ名詞に対して例文を経済学部 1 年生 2 名が作成した (1299 文)。法学部の 4 年生 1 名 (A)、言語学系研究室所属の修士の学生 2 人 (B、C) の計 3 名が意味役割と述語の同定を行った。1299 文に対して項が付与された箇所は 2554 箇所である[(5)]。項の付与件数の内訳は作業者 (A) が 297 件、(B) が 1019 件、(C) が 2017 件で合計 3333 件付与されている。付与箇所では部分的に複数の作業

(4) 辞書は IPAdic を利用した。

(5) これらの例文と項を付与した事例は先行研究の名詞項構造データ (竹内ほか 2015) とは別の新たなデータ集合である

者が付与している

図 2　CakePHP による名詞の項構造付与ツール

名詞に対する例文作成、項の同定と述語項構造シソーラスからの述語の選択、および意味役割の付与は CakePHP を利用した付与ツールで付与した。図 2 に「店員」の例文に対する付与例を示している(6)。作業者は名詞の例文を見ながら、検索の部分で関連する述語を述語シソーラスから検索し、適合する述語があれば選択できるようにしている。また、意味役割についても同様で、意味役割の一覧から各項 (arg0、arg1) に対して付与できる。図 2 中の qualia の項は、各項が生成語彙における agentive か telic のどちらに分類されるか分かる場合は任意で付与できるようにしている。ただし、図に示すように、付与が無い場合や著者と判断が異なる作業者もいるため現段階では必須のデータにしていない。意味役割付与については 3333 件の項のうち、3001 件に対して付与されている。意味役割は図にもあるように述語シソーラスで記載されている必須項だけではないため付加詞など 72 種類の中から付与が行われている。全体的には作業者にとって名詞の項構造付与は容易ではなく、付与の取りこぼしが多く見られた。また、名詞に関する述語を選択する部分が容易ではない場合があることが分かってきた。次節で問題点を挙げる。

3.2　付与作業からの考察

付与作業を通して、名詞の項構造付与の難しい点がいくつか明らかになってきた。大きく分けると、(1) 名詞項構造例文作成、(2) 項構造を有するかどうかの判定、(3) 対応する述語の選択の 3 つである。

まず、(1) では名詞の構文 (「X の Y は Z だ」) が合わないことが挙げられる。例えば「作品」という名詞では「太郎の作品はこれだ」という例文が付与された。「作品」は項を持つ名詞

(6) 作業者 koichi は著者であり、上記の付与数からは外している。

で、「の」の前に作者が来る事例で正しい。しかしながら作者と作品の関係を具体的に入れるとおかしな例文となる。「伊坂幸太郎の作品は『重力ピエロ』だ」では作品が 1 つに限定されてしまい不自然に感じるため例文として成立しない(7)。名詞項構造データとしてできれば具体的な事例が入った名詞の項構造の例文が求められるが、今回の作業で指定した構文では捉えられない名詞が存在することが明らかになってきた。これに対しては今後異なる名詞の構文を用意する必要があると考えられる。

(2) では作業者が例文に対して項構造を付与しない例が少なからず見受けられた。例えば「作品」や「小説」など項を持つと考えられる名詞の項構造が付与されていない。これはひとつには項を持つ名詞の理解が難しいこともあるが、ほとんどの場合上記 (1) のように例文が「太郎の小説」など「X の Y は Z だ」の構文にならなかった場合に作業者が項を付与しなかったことが原因である。よって (1) に加えてツールに例文をより容易に修正できる機能を加える必要があると考えられる。

(3) では名詞に対応する述語の選択が容易ではないことが作業から明らかになった。例えば「罰」の例文「[arg1 盗み食い] の罰は [arg2 反省文] だ」 に対して、動詞「罰する」を選択した。項の対応としては arg1 が [原因]、arg2 が [手段] が考えられる。つまり文で表現すると「[[原因] 盗み食いで] [[手段] 反省文で] 罰する」である。「罰」の意味を考えると「罰する」が対応すると考えられるが、一方で作業者の中には、「書く」を選択したものもいた。これは「[[原因] 盗み食いで] [[対象] 反省文を] 書く」と考えたためである。「反省文」という名詞自身が「罰」に関連した意味を持つため「反省文を書く」ことが例文の言い換えとして外れていないようにも考えられる。つまり、名詞の意味を生成語彙の枠組で述語の項として形式化した場合、名詞の概念との組合せで複数の述語が可能であることが考えられる。例えば「著者」の場合の agentive に対して「本を書く」以外に「本を出す」という表現も「書く」の意味で使われることがある(8)。よって事例をベースとした名詞との結びつきで複数の述語との対応を許す枠組が必要である。

また、(3) では述語そのものを考えるのが難しい場合がある。例えば「候補」の例文「[arg1 次の生徒会長] の候補は [arg0 田中さん] だ」に対して、「選ぶ」という述語を選択した。つまり「[[対象] 田中さんを][[補語相当 (を)] 次の生徒会長に] 選ぶ」と対応づけた。しかしながら、「候補」の意味としてはまだ「生徒会長」に選ばれたわけでは無い。その可能性があるだけである。よって「可能性がある」などモダリティに関する情報の付与も検討される。これは 2 節で述べた枠組にも関連する。名詞の意味に関連した述語表現を考えると過去形や機能表現が必要になることが考えられる。言い換え処理として利用する際、どこまでの粒度が必要か応用の観点からの見直しが必要と考えられる。

4. まとめ

本稿では述語に結び付けた名詞の項構造データの構築法について記述した。先行研究の生成語彙を利用した記述枠組を元に、RITE-2 含意認識タスクに現れた名詞を中心に例文を作成し、

(7) 倒置指定文と考えられるが「私がもっとも好きな」など文脈を与えると自然な例文になる。
(8) 例えば質問応答で著者を文から取り出す場合、「X が Y(作品) を出した」という構文は有効である。

項構造の同定と対応する述語および意味役割を述語項構造シソーラスを基に付与した。1299文に対して2554箇所の項を同定した事例を作成した。項構造データ構築作業から、名詞の項構造例文作成の難しさや、対応する述語の選択での難しさを明らかにした。今後、付与した事例データを基に名詞項構造付与の枠組を整理する予定である。

謝　辞

本研究は、基盤研究 (C) 課題番号 26370485 (研究代表者: 竹内孔一) の補助を得ている。ここに記して深く感謝する。

文　献

西山佑司 (2003).『日本語名詞句の意味論と語用論』　ひつじ書房.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman (2004). “Annotating Noun Argument Structure for NomBank.” *Proceedings of LREC2004*, pp. 803–806.

庵功雄 (2007).『日本語におけるテキストの結束性の研究』　くろしお出版.

西山佑司 (編) (2013).『名詞句の世界』　ひつじ書房.

影山太郎 (2011).『日英対照 名詞の意味と構文』　大修館書店.

A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan, P. Fodor, and J. Chu-Carroll (2012). “Question analysis: How Watson reads a clue.” *IBM Journal Research and Development*, 56:3/4, pp. 2:1–2:14.

P. Kingsbury, M. Palmer, and M. Marcus (2002). “Adding Semantic Annotation to the Penn TreeBank.” *Proceedings of the Human Language Technology Conference*.

竹内孔一・宮田周・河村一希 (2015).「述語項構造シソーラスを意識した名詞データの構築」第7回コーパス日本語学ワークショップ予稿集, pp. 143–146.

『名大会話コーパス』中納言版・ひまわり版公開データの作成

柏野 和佳子（国立国語研究所音声言語研究領域）*
西川 賢哉（国立国語研究所コーパス開発センター）
小磯 花絵（国立国語研究所音声言語研究領域）

Supplemental Arrangement for Public Data Available in the Chunagon and Himawari Versions of "Nagoya University Conversation Corpus"

Wakako Kashino (National Institute for Japanese Language and Linguistics)
Ken'ya Nishikawa (National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

『名大会話コーパス』は，科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（研究代表者：大曾美恵子，平成 13 年度～15 年度）の一環として作成された，120 会話，合計約 100 時間の日本語母語話者同士の雑談を文字化したコーパスである。国立国語研究所に移管後，文字化テキストを公開し，続けて『中納言』版，『ひまわり』版を作成し，公開している。

本稿では，『名大会話コーパス』の概要と特徴を述べる。また，『中納言』版，『ひまわり』版公開データの作成に際して行った，形態素解析結果の人手修正の内容について報告する。

１．はじめに

国立国語研究所の『日本語話し言葉コーパス』（CSJ）は，独話を主対象とするコーパスである。また，『現代日本語書き言葉均衡コーパス』（BCCWJ）及び，『国語研日本語ウェブコーパス』（NWJC）は，いずれも書き言葉のコーパスである。日常会話場面を対象とした大規模な『日本語日常会話コーパス』の構築は，国立国語研究所の機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）により，着手されたところである(小磯ほか 2017)。そのような状況の中，現時点で広く利用可能である自然会話のコーパスが『名大会話コーパス』である。

『名大会話コーパス』は，科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（研究代表者：大曾美恵子，平成 13 年度～15 年度）の一環として作成された，120 会話，合計約 100 時間の日本語母語話者同士の雑談を文字化したコーパスである。国立国語研究所に移管後，文字化テキストを公開している。さらに，「大規模日常会話コーパスに基づく話し言葉の多角的研究」のプロジェクトにおいて，文字化テキストを対象に，形態素解析用辞書『UniDic』 と形態素解析器『MeCab』を用いて形態論情報（短単位）を自動付与し，メタ情報として発話者の属性（性別・年代・出生地など）と会話の情報（収録日・収録場所など）を整理した上で，オンライン検索システム『中納言』，及び，全文検索システム『ひまわり』にて 2016 年 12 月より一般公開

* waka ＠ninjal.ac.jp

している。

本稿では，『名大会話コーパス』の概要と特徴を述べる。また，『中納言版』，『ひまわり版』公開データの作成に際して行った，形態素解析結果の人手修正の内容について報告する。

２．『名大会話コーパス』概要と特徴

２．１　『名大会話コーパス』の概要

はじめに，文字化テキストの例(data001 の冒頭)を示す。

> ＠データ１（約３５分）
> ＠収集年月日：２００１年１０月１６日
> ＠場所：ファミリーレストラン
> ＠参加者 F107：女性３０代後半、愛知県幡豆郡出身、愛知県幡豆郡在住
> ＠参加者 F023：女性４０代後半、岐阜県出身、愛知県幡豆郡在住
> ＠参加者 M023：男性２０代前半、愛知県西尾市出身、西尾市在住
> ＠参加者 F128：女性２０代前半、愛知県西尾市出身、西尾市在住
> ＠参加者の関係：英会話教室の友人
> F107：＊＊＊の町というのはちいちゃくって、城壁がこう町全体をぐるっと回ってて、それが城壁の上を歩いても１時間ぐらいですよね。
> F023：１時間かからないぐらいだね。
> ４、５０分で。
> F107：そうそう。
> ほいでさあ、ずっと歩いていたんだけど、そうすと上から、なんか町の中が見れるじゃん。
> あるよね。
> ほいでさあ、なんか途中でワンちゃんに会ったんだね。
> （ふーん）散歩をしてるワンちゃんに会ったんだ。
> F023：城壁の上をやっぱ観光客なんだけどワンちゃん連れてきてる人たち結構多くて。

上記のような会話データが全部で 129 会話(data001～data129)ある。それぞれに以下の情報が付与されている。以下，これら情報の内訳を概観する。

データ情報：収録時間，収録年月日，収録場所
参加者情報：性別，年代，出身地，居住地
参加者の関係：参加者間の関係

２．１．１　データ情報：収録時間，収録年月日，収録場所

まず，表 1 に収録時間，表 2 に収録年，表 3 に収録場所の内訳を示す。収録時間は，31～60 分のものが最も多く，平均は 47 分である。収録年月日は 2001 年 10 月 16 日～2003 年 2 月 17 日の間である。2001 年のデータが最も多い。収録場所は，テキストには例えば，「レストラン」「うどん屋」「喫茶店」「○○の実家」「○○宅」といったように示されているが，

それらを大まかに分類しなおしてみると，表 3 の通り，飲食店，自宅，大学で多く収録されている。なお，場所については，二重分類 3 件を含んでいる。

表 1　収録時間

収録時間	件数
～30分	13
31～60分	99
61～90分	16
91～分	1
合計	129

表 2　収録年

年	件数
2001年	78
2002年	48
2003年	3
合計	129

表 3　収録場所

場所	件数
飲食店	46
家	30
大学	29
大学の研究室	13
車内	8
職場	2
大学の食堂	2
学校	1
電車内	1
合計	132

２．１．２　参加者情報：性別，年代，出身地，居住地

次に，表 4 に年代別の性別，表 5 に出身地，表 6 に居住地の内訳を示す。性別は女性が多い。また，年代は 20 代が最も多い。出身地と居住地は，テキストには都道府県に加え市まで示してあるものもある。また，途中の引っ越し歴の記載があるものもある。出身地，居住地ともに中部が多いが，それ以外もある。

表 4　年代別の性別

年代	女性	男性	総計
10代	13	2	15
20代	70	18	88
30代	26	1	27
40代	16	8	24
50代	18	4	22
60代	11	4	15
70代～	6		6
不詳	1		1
合計	161	37	198

表 5　出身地

出身地	人数
北海道	11
東北	8
関東	49
中部	86
近畿	21
中国・四国	11
九州・沖縄	11
海外	1
合計	198

表 6　居住地

居住地	人数
北海道	18
東北	1
関東	49
中部	120
近畿	7
中国・四国	1
九州・沖縄	0
海外	2
合計	198

２．１．３　参加者の関係：参加者間の関係

最後に，表 7 に参加者の関係，表 8 に会話の参加者の人数の内訳を示す。参加者の関係は，テキストには例えば，「英会話教室の友人」「アルバイトの友人」「中学の同級生、F106 の母親」「F154 と F130 は友人。M004 は初対面の人。」といったように示されているが，それらを大まかに分類しなおしてみると，表 7 の通り，同級生，友人，家族，先輩，同僚の関係が多く収録されている。なお，関係については，重複分類 12 件を含んでいる。表 8

より，本データのほとんどが 2 名の対話であることがわかる。

表 7　参加者の関係

関係	件数
同級生	51
友人	31
家族	15
先輩	15
同僚	11
初対面	6
知人	5
恋人	4
親族	2
先生	1
合計	141

表 8　参加者の人数

参加者の人数	件数
2人	96
3人	28
4人	5
合計	129

２．２　『名大会話コーパス』の特徴

２．２．１　上位語

書き言葉の代表として『現代日本語書き言葉均衡コーパス』（以下，『BCCWJ』）の語彙表を用いて，『名大会話コーパス』の話し言葉としての特徴を概観する。まず，上位語の比較を表 9 に示す。

表 9　『名大会話コーパス』と『BCCWJ』の上位語の比較

順位	名大会話			BCCWJ		
1	ダ	だ	助動詞	ノ	の	助詞-格助詞
2	ウン	うん	感動詞-一般	ニ	に	助詞-格助詞
3	タ	た	助動詞	テ	て	助詞-接続助詞
4	テ	て	助詞-接続助詞	ハ	は	助詞-係助詞
5	ネ	ね	助詞-終助詞	ダ	だ	助動詞
6	ノ	の	助詞-準体助詞	ヲ	を	助詞-格助詞
7	カ	か	助詞-副助詞	タ	た	助動詞
8	ト	と	助詞-格助詞	スル	為る	動詞-非自立可能
9	デ	で	助詞-格助詞	ガ	が	助詞-格助詞
10	ノ	の	助詞-格助詞	ト	と	助詞-格助詞
11	モ	も	助詞-係助詞	デ	で	助詞-格助詞
12	ガ	が	助詞-格助詞	モ	も	助詞-係助詞
13	ニ	に	助詞-格助詞	イル	居る	動詞-非自立可能
14	ハ	は	助詞-係助詞	マス	ます	助動詞
15	ナニ	何	代名詞	ノ	の	助詞-準体助詞

表 9 において赤四角で囲んだ赤字の箇所で示した語が上位語であることが，すなわち『名大会話コーパス』の話し言葉としての特徴を表すと思われるものである。

２．２．２　品詞の分布

続いて，同じく『BCCWJ』の語彙表を用いて，『名大会話コーパス』と品詞の分布を比較する。

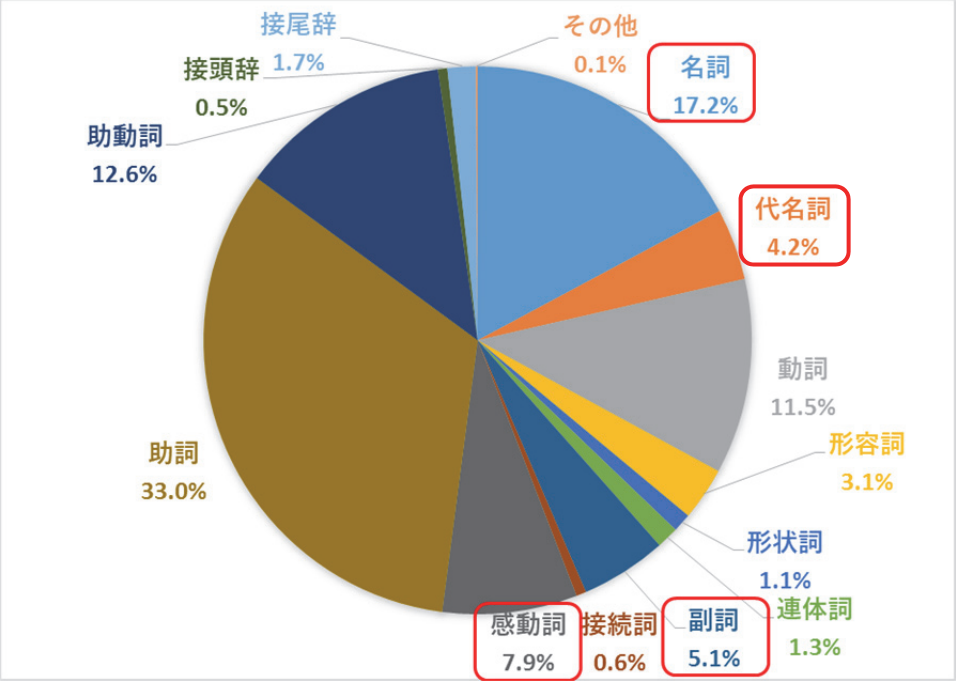


図 1　『名大会話コーパス』の品詞の分布

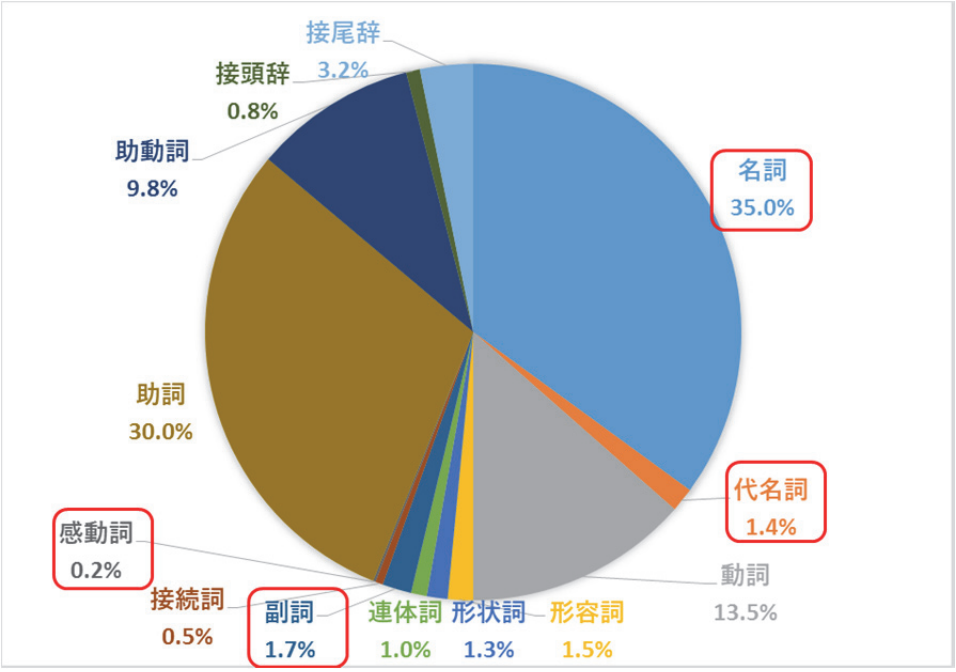


図 2　『BCCWJ』の品詞の分布

図 1 と図 2 を比べると，赤の四角形で強調している通り，『名大会話コーパス』は名詞が少なく，感動詞，副詞，代名詞が多い。ここに話し言葉としての特徴が表れていると考えらえる。

２．２．３　特徴語

Fujimura et al.(2012)では，『名大会話コーパス』にみられる話し言葉の特徴を，書き言葉の代表として『現代日本語書き言葉均衡コーパス』（モニター公開データ 2008/2009）と比較して述べている。LLR（対数尤度比）により，特徴語 10 語として「うん，だ，ね，の，か，そう，って，た，何，言う」を挙げている。表 9 で示した上位語にないものが「そう」と「言う」である。この「言う」は，次のような例（下線，太字部分）で頻出するものである。

午前中はずーっと部屋に、部屋って<u>いう</u>か玄関に入ってたんだけどー　(data019)

なお，この「っていうか」は「てか」という短縮形でも頻出する（以降にその用例を示す)。いずれも話し言葉の特徴を示す語であると言える。

山崎(2016)では，『名大会話コーパス』の特徴を『日本語話し言葉コーパス』（CSJ）の学会講演と模擬講演，『BCCWJ』の小説会話文と比較し，示している。LLR により『名大会話コーパス』の特徴語として示しているのは，例えば，10 代では「超」「うざい」などの若者語，60 代では「いらっしゃる」「おっしゃる」といった敬語や，「何しろ」「随分」などの特定の副詞である。

本稿では，いわゆる「話し言葉」であるため，『BCCWJ』では用例が得にくいが，『名大会話コーパス』で頻出することが期待される語を特徴語の例として，次の a)〜h)の 8 語を取り上げる。以下に示す方法にて，『中納言』で検索をした。その結果得られた検索結果数と用例とを示す。

a)　微妙　156 件（語彙素「微妙」で検索）※ほか，「びみょー」1 件あり
b)　やば　168 件（文字列「やば」で検索）※「やば」「やばい」同時に検索
c)　まじ　197 件（語彙素「まじ」で検索）※「まじか」は 1 例のみ
d)　無理　273 件（語彙素「無理」で検索）
e)　てか、　60 件（文字列「てか､」検索）※ほか，「､」以外にも用例あり
f)　すごい＋形容詞　344 件（書字形出現形「すごい」＋形容詞で検索）
g)　うける　37 件（語彙素「受ける」の終止形で検索）
h)　みたいな 473 件（文字列「みたいな[､。？]」で検索）※「！」はなかった。

なお，以上に示す検索結果数は，当該語の「話し言葉」ならではの用法例の正確な件数ではない。検索もれ，あるいは，別語，別用法の例が少々混じっている。より正確に検索するためには，語彙素検索と文字列検索とを併用し，さらに検索結果を絞り込むことが望ましい。

しかしながら今回の検索方法でも，下記に示す通り，当該語の「話し言葉」ならではの用例が得られることが確かめられる。

ドラえもんとかあ。ドラえもんは超うめー。ハットリ君とかあ。ハットリ君はそれは	微妙	。めっちゃ微妙。もうね、もうね、みんなね、みんながねすごい頭ん中	data005
おる？この子は、Ｅ短の子だよ。あっ、そうなんだー、	微妙	、微妙。うんそういうのばっかり。はい、ん、これはだめだな。うん	data077
何かね、ストレス感じると食べちゃうのね、すごい。うんうんだから絶対	やばい	。これ４年生の二の舞になると思って、ちょっと制限しようと思ってるん	data003
だけど、なかなか時間がないんだよね。ねーあたしもだよ。	やばー	い。あっ、ＴＯＥＩＣさ、こないだあったけど、一番初めに受けたやつさ、	data072
受かっちゃったもん、最初。５級だったしね、一番最初受けたの。	まじ	？６年のときに５級。そっか、そんなん、だって、小学校	data011
のなんなの、これ。知らないよ。大体電池がないじゃん。うそ、	マジ	？なんか電池があと２つみたいになってる。本当？うーう、携帯でも	data046
んだ、あそこで？そうそうそうそうそうそうこえーなー。私、あのカーペット系がもっと	無理	。あ、あれ、超酔う。あれ、なんかさ、なんかもう、抜けそう	data072
無理だよ。だからなんで。うーん。無理。だから、なんで。とにかく	無理	。それは、そういうことしたら、そういうことをやったっていうことを	data046
ゆくんだね。そうだねえ、ほんと。どんどんみんな大人になっちゃう、	てか、	社会人になっちゃう。うん。わせさいんときの先輩なんかもみんなもう決まって	data066
抽象画じゃないのよ。じゃないの。うんどこで見たの。	てか、	あの、日本に来たときの。昨日か、昨日かおといもテレビであった	data056
ないと、あれだね。１人でお鍋セットとかもらっても、うれしく	なくない	？そっかー。じゃ、友だち呼んで、お鍋パーティすればいい。うーん、そう	data055
そういうことは全然ないけど、それでもさ、そんなに聞きたい話じゃ	なくない	？うんうん、うんうんうん。全然。うん。でさ、何か、その、	data094
のー、なんか、中からしか選べないからとか言ってたよね。うん	すごい	(高い)んだろうねー。だってなんかゴージャスだからねー。うーんじゅうたん１つとって	data120
鼻血出そうだ。本気で出そうだ。んー、かっこいい。＊＊＊。	すごい	(かわいい)、この絵。すげーな才能のある人っていうのはすごい。うん。＊	data103
に言ったんだって。うんうんうんうんうんすごい懇願したんだって。	うける	ー。ほんで、私、ハッて思いついたんだ。うんあ、F114 ちゃんって	data066

でーとか、あ、君は日本文学専攻か、ふーん、とか言って。	受ける	ー。うーん話しかけやすい雰囲気なんじゃん。なのかね。うん困っちゃうね。	data065
いろいろ言ってたのー。うん日本語教師とかそっち系に進むには	みたいな、	でも、結構ね。うーんでも日本じゃ無理だってさ。うん、だから	data011
てー。うんいいよねー結構いろんな人にー、こういうの、どう？	みたいな。	うんだからさ、写真で生きていけるって、F141 の場合。うん。だから	data123
じゃない？うん雰囲気。今日、さむ。な、何となく寒そう、	みたいな？	うん。雪が降ったときとか、雪のお水がまだ解けきらないとき	data085

３．形態素解析結果の人手修正の内容

３．１　人手修正した範囲

『名大会話コーパス』を，オンライン検索システム『中納言』，及び，全文検索システム『ひまわり』にて公開するに際し，形態素解析用辞書『UniDic』 と形態素解析器『MeCab』を用いて形態論情報（短単位）を自動付与し，その結果を人手修正した。129 会話全ての一部分（各会話 1,500 形態素以上）を目視で修正する範囲に定め，まずはその範囲内のものを修正した。加えて，全範囲に対する一括修正も行った。『名大会話コーパス』の形態素数と人手修正した形態素数を表 10 に示す。記号・補助記号・空白を除いた形態素数に対し，人手修正で目視した作業範囲の形態素数は 26.8%にあたる。また，人手修正した形態素数は 2.5%にあたる。

表 10　『名大会話コーパス』の形態素数と人手修正した形態素数（短単位）

全形態素数	1,419,729
記号・補助記号・空白を除いた形態素数	1,131,891
人手修正で目視した作業範囲の形態素数	303,282
人手修正した形態素数	27,931

３．２　人手修正の内容

３．２．１　口語表現の誤解析の修正

以下に，口語表現に関して誤解析を修正した具体例を示す。例は左から，「対象」「テキスト（誤）」「修正後（正）」で示す。形態素の区切りは「｜」で示す。語彙素，品詞，活用形等は着目する部分のみ簡易表示する。

①「なん」

なん	そう｜なん：「何」｜だ	そう｜な：「だ」助動詞｜ん：準体助詞｜だ
なん	一緒｜なん：「など」｜じゃん。	一緒｜な：「だ」助動詞｜ん：準体助詞｜じゃん。
なん(なる)	男｜の｜人｜も｜便秘｜に｜なん：「何」｜だ｜ね。	男｜の｜人｜も｜便秘｜に｜な：「成る」連体形-省略｜ん：準体助詞｜だ｜ね。

②「そっか」「そやね」「てゆーか」

そっか	あ｜、｜そっ：副詞｜か｜そっ：「其処」代名詞｜か	あ｜、｜そっ：副詞｜か｜そっ：副詞｜か
そっか	そっ：「そう」副詞｜か：副助詞｜そっ：「そう」副詞｜かそっ：「貸す」｜か：終助詞	そっ:「そう」副詞｜か:終助詞｜そっ：「そう」副詞｜か：終助詞｜そっ：「そう」副詞｜か：終助詞
そやね	そや：「粗野」｜ね。	そ：副詞｜や：助動詞｜ね。
てゆーか	て：「で」接続詞｜ゆー：固有名詞一人名｜か：副助詞	て：副助詞｜ゆー：「言う」｜か：終助詞

③「やっとく」「やっとって」「やって」「おって」

やっとく	今日｜、｜やっと：副詞｜かん：「彼」｜と：格助詞	今日｜、｜やっ：「遣る」｜とか：「とく」助動詞｜ん：「ず」｜と：接続助詞
やっとって	３｜級｜の｜問題｜やっと:副詞｜って：副助詞	３｜級｜の｜問題｜やっ：「遣る」｜とっ：「とる」助動詞｜て：接続助詞
やって(だって)	３１｜日｜は｜休み｜やっ：「遣る」｜て。	３１｜日｜は｜休み｜やっ：助動詞｜て。
おって	同じ｜屋敷｜に｜おっ：「追う」｜て｜ね	同じ｜屋敷｜に｜おっ：「居る」｜て｜ね

④「しよー」「よ」「あーあ」「あーん」「えっと」

しよー	これ、｜どー：副詞｜し：「為る」｜よー：終助詞｜。	これ、｜どー：副詞｜しよー：「為る」｜。
よ	朝｜起き｜たら｜ちょっと｜頭痛｜いよ：感動詞	朝｜起き｜たら｜ちょっと｜頭｜痛い｜よ：終助詞
あーあ	あー：感動詞｜あっ：「有る」｜て：接続助詞｜思い｜ながら	あーあ：感動詞｜って：副助詞｜思い｜ながら
あーん	あー：感動詞｜ん：感動詞｜ラジオ｜って｜いう｜か	あーん：感動詞｜ラジオ｜って｜いう｜か
えっと	あたしはー、｜えっ：感動詞｜と：格助詞	あたしはー、｜えっと：「えーと」感動詞

⑤「あら そう」「こら」「あの」「いいやん」

あら そう[あいずち]	あらそう：「争う」｜。	あら：感動詞｜そう：副詞｜。
こら	うん｜こら：感動詞｜やばい	うん｜こら：代名詞｜やばい
あの	あの：感動詞｜人｜も？	あの：連体詞｜人｜も？
いいやん	いいや：感動詞｜ん：感動詞｜、	いい：「良い」｜やん：終助詞｜、

⑥「ねーねー」「ねー」「とか」

ねーねー	ねー：終助詞｜ねー：終助詞｜、	ねー：感動詞｜ねー：感動詞｜、
ねー	すごく｜ねー：「無い」｜。	すごく｜ねー：終助詞｜。
とか	おおーっと：感動詞｜か｜言っ｜て	おおーっ：感動詞｜と：格助詞｜か｜言っ｜て

とか	忙しかっ｜た｜ん｜じゃ｜ない｜か｜な｜っ：補助記号｜と：格助詞｜か	忙しかっ｜た｜ん｜じゃ｜ない｜か｜な｜っと：格助詞｜か

⑦「と」

と[接続助詞]	～し｜ない｜と：格助詞｜。｜	～し｜ない｜と：接続助詞｜。｜
と[格助詞]	歩い｜て｜登る｜と:接続助詞｜おっしゃる｜から	歩い｜て｜登る｜と：格助詞｜おっしゃる｜から

『日本語日常会話コーパス』でこれらの口語表現を全てそのまま表記するとは限らないが，口語表現を積極的に採用する方針であり，これらの修正履歴は今後の解析精度向上のために参考にしていく。

３．２．２　発音の誤認定の修正

発音の誤認定のタイプは主に３つに分けられる。１つ目は「数字」である。以下では一例しか示さないが，１～９すべての数字の読みについて複数ある読みの中から適切なものを自動判定することは難しいため，多く誤認定が生じている。２つ目は「清濁」である。これも多くの誤認定が生じている。このうち，以下の「座布団」のような連濁を自動判定するのには限界がある。しかしながら，その次の「橋の上」のようなものが語頭で「バシ」と濁音になることを解析時に避けることは可能であると考え，その対応を検討中である。３つ目は「音訓」である。熟語が湯桶読みや重箱読みの場合，訓が複数ある場合，熟字訓がある場合などに選択誤りがある。しかし，これらも辞書を整備することで少しでも解析精度を上げることを考えている。

以下に例を示す。着目する箇所の発音をカタカナで表記する。

①数字

一	１　イチ｜個	１　イッ｜個
四	４　ヨン｜人	４　ヨ｜人

②清濁

半濁音	５｜分　プン	５｜分　フン
濁音	座　ザ｜布団　フトン	座　ザ｜布団　ブトン
濁音	橋　バシ｜の｜上｜から	橋　ハシ｜の｜上｜から

③音訓

音訓	洗濯｜物　ブツ	洗濯｜物　モノ
音訓	大　ダイ｜掃除	大　オオ｜掃除
音訓	紅　コー｜ショウガ	紅　ベニ｜ショウガ
訓と訓	小麦｜粉　コナ	小麦｜粉　コ
訓と訓	いつ｜の｜間　アイダ｜にか	いつ｜の｜間　マ｜にか
訓と熟字訓	お｜父　チチ｜様｜と｜お｜母　ハハ｜様	お｜父　トウ｜様｜と｜お｜母　カア｜様

以上のほか，読みが複数あり，音声がないと正誤の判断がつかないものもある。例えば，

「その他（ソノタ／ソノホカ）」「毎年（マイネン／マイトシ）」「白髪（ハクハツ／シラガ）」のようなものである。なお，本データでは，「ソノタ」「マイトシ」は統一されているが，「ハクハツ／シラガ」は統一されていない。

実は，『BCCWJ』のときに整備した発音の後処理(伝ほか 2002)を今回は実施していない。そのため誤認定が多くあった。今後，『日本語日常会話コーパス』では，発音の後処理を積極的に導入する予定でいる。また，『日本語日常会話コーパス』でも漢字仮名交じりで表記するため，発音が一意に同定できないケースも生じるが，後処理として発音を確認する工程を設けることで正確な発音を保証する(川端ほか 2017)。

３．２．３　その他の修正

Unidic に登録のないオノマトペが多く出てきた。例えば，「ぴしゃーっ」「ビヨーン」「ごっしゃごしゃ」「ずきっ」などである。これらはただちに「オノマトペ」としての登録はせず，今回は「新規未知語」として今後の課題としている。

また，長音や撥音がそのまま語の途中に記述されているものが少なくなかった。例えば，「吉祥ー寺」「ひどーい」「つらーい」「なさーい」「すっげー」などである。このようなものは自動解析で語の途中で切られてしまう。そこで，人手修正した範囲内で見つけたものは一短単位に修正し，「新規未知語」としている。しかしながら，「吉祥ー寺」のようなものは同じものが複数例はないが，あとの４語は複数例あるものである。人手修正で見逃してしまったものは，語断片の誤解析のままとなっている。例えば，次の通りである。

ひどーい	ひ：「日」｜どー：「どう」副詞｜い：「いー」フィラー
	ひ：「ひい」感動詞｜どー：「どう」副詞｜い：終助詞
	ひど：「ひどい」形容詞｜ー：補助記号｜い：「いー」フィラー
つらーい	つらー：「つらー」副詞｜い：終助詞
なさーい	な：「だ」助動詞｜さー：「さ」終助詞｜い：終助詞
すっげー	す：「すっ」副詞｜げ：「気」：接尾辞｜ー：補助記号

なお，『日本語日常会話コーパス』では，強調や言い淀みのために一時的に付加された非語彙的な長音や撥音は，タグによって表現する。よって，上述の問題に『日本語日常会話コーパス』の解析時に対応することは考えていないが，ほかの話し言葉を対象とする解析時には注意が必要である。

４．おわりに

現時点で広く利用可能である自然会話のコーパスである『名大会話コーパス』の概要と特徴を述べた。書き言葉コーパスである『BCCWJ』と比べ，「うん」，終助詞の「ね」「か」，「何」が頻出し，また，感動詞，副詞，代名詞が多く，話し言葉の特徴語の用例が多く得られるコーパスであることを示した。様々な話し言葉の研究利用が期待できる。

また,『中納言』版,『ひまわり』版の公開データ作成のために, 形態素解析用辞書『UniDic』と形態素解析器『MeCab』による形態素解析を行った結果に対する人手修正の内容を報告した。「そうなんだ」「そっか」「てゆーか」などの口語表現の誤解析の修正,「１（イチ）個」「５分（ブン）」「いつの間（アイダ）にか」などの発音の誤認定の修正など，具体例を示し，今後の『日本語日常会話コーパス』などの話し言葉を対象とする形態素解析の精度

をあげるための留意事項を述べた。

謝　辞

本研究は国立国語研究所の機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）の研究成果を報告したものです。また，オリジナルの『名大会話コーパス』は，科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（研究代表者：大曾美恵子，平成 13 年度～15 年度）による研究成果です。形態素解析結果の人手修正をはじめ，本コーパスの構築，公開にご協力くださった皆さまに感謝します。

文　献

Itsuko Fujimura, Shoju Chiba, Mieko Ohso (2012) Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and Written corpora , *Proceedings of the VIIth GSCP International Conference. Speech and Corpora,* pp.393-398.

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵(2017) 「『日本語日常会話コーパス』の転記基準と作業工程」『言語資源活用ワークショップ 2016 予稿集』(収録予定).

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉(2017) 「『日本語日常会話コーパス』の構築」『言語処理学会第 23 回年次大会発表論文集』.

伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治(2002) 「話し言葉研究に適した電子化辞書の設計」『第 2 回話し言葉の科学と工学ワークショップ講演予稿集』pp. 39-46.

山崎誠(2016) 「レジスターの違いによる話し言葉の変容」『シンポジウム「日常会話コーパス」I』発表資料(http://pj.ninjal.ac.jp/conversation/pdf/sympo2016-3.pdf).

関連 URL

国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」 http://pj.ninjal.ac.jp/conversation/

『UniDic』 https://ja.osdn.net/projects/unidic/

『MeCab』 http://taku910.github.io/mecab/

『日本語自然会話書き起こしコーパス（旧名大会話コーパス）』 https://nknet.ninjal.ac.jp/nuc/templates/nuc.html

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

コーパス検索アプリケーション『中納言』 https://chunagon.ninjal.ac.jp/

『現代日本語書き言葉均衡コーパス』語彙表 http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション —基準策定、仕様書作成の必要性について—

松本 理美（立命館大学大学院言語教育情報研究科）
浅原 正幸（国立国語研究所コーパス開発センター）
有田 節子（立命館大学大学院言語教育情報研究科）

Clause Class Annotations on the 'Balanced Corpus of Contemporary Written Japanese' -- Necessity of Developing Fine-grained Criteria and Specification

Satomi Matsumoto (Graduate school of Language Education and Information Science, Ritsumeikan University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Setsuko Arita (Graduate school of Language Education and Information Science, Ritsumeikan University)

要旨

本発表では、「現代日本語書き言葉均衡コーパス」に対する節の意味分類情報アノテーションについて報告する。多様な形式を持ち、文脈の中でその意味が解釈される日本語文中の従属節の意味分類については、人手による分類が不可欠である。そこで、我々は「鳥バンク」基準互換 (池原 2009) の節の意味分類情報アノテーションを進めている。しかし、現行の作業においては、節の認定、タグ付け箇所、作業者の言語感覚に頼るところが大きい意味分類判断など、作業上の揺れも多く、改善が求められる。作業効率と信頼性の向上に繋がる基準策定と仕様書作成が必要であり、そのためには現行作業での問題点を整理することが必須であると考える。本発表では、人手による節境界アノテーション・節の意味分類タグ付け作業についての基準策定と仕様書作成が今後のコーパス開発に資することを主張し、現在の作業における問題点に焦点を当てた考察を行う。

1. はじめに

本発表では、我々が進めてきた「現代日本語書き言葉均衡コーパス」に対する節の意味分類情報アノテーションについて紹介する。

人手による節境界アノテーション・節の意味分類タグ付け作業（以下、作業とする）における問題点を整理し、作業上の基準策定や作業仕様書の必要性を示すことを本発表の目的とする。アノテーション整備のゴールが“正確な”情報付与でないことは言うまでもない。設計に関わる諸分野の専門家と多種多様な利用者による問題提起や議論、一方通行ではない情報の授受というアノテーションを介在したコミュニケーションが、言語現象に関する新たな発見とコーパス言語学の発展をもたらす可能性は大きいと考える。

そこで、節認定・タグ付け箇所・意味分類のいずれにも判断の揺れが生じることを許容したうえで、人手による作業だからこそ可能となる判断の許容範囲の探求を試みる。作業者間、あるいは解析器と人間の間に生じた齟齬の分析により技術的、理論的な問題を明らかにする。

実際のアノテーションにおける齟齬について紹介すると次のような傾向がある。補足節

における齟齬の発生は他の節と比較して少なく、特定の節に偏って齟齬が見られるということもない。作業者間で多くの齟齬が生じる名詞修飾節は他の節と認定されることは極めて少ないが、名詞修飾節内の下位分類間での齟齬が大量に発生している。同じく作業者間で多くの齟齬が生じる副詞節は、並列節と認定される齟齬が目立つ。同様に並列節は専ら副詞節との間に齟齬を生じる傾向にあり、副詞節 VS 並列節の様相を呈している。本稿では、これらの各論について節認定・タグ付け箇所・意味分類の三つの観点から議論する。

一言で節認定・タグ付け箇所・意味分類の齟齬といっても、従属節の一つ一つの個性が、その齟齬にも反映される結果となっている。本研究は、工学的な要求からコーパスを設計する工学研究者、文法的視点から節認定や意味分類を検討する日本語学研究者、節境界アノテーション・節の意味分類タグ付けを行った作業者による共同発表であるが、それぞれの観点で「どの程度の齟齬を許容すればよいか」、その“落としどころ”をどこに求めるかは、非常に難しい問題である。それでも敢えて、作業者間、作業者内での揺れを想定した作業上の基準策定や仕様書作成の必要性を主張する。それにより作業の効率化と、データの信頼性向上の可能性があると考えるからである。

次節以降、解析器への実装を前提とした作業上の問題点を論じる。

2. 作業における基準と問題点

本作業では、節を「複文を構成するところの、述語を中心とした各まとまり」(益岡・田窪 1992 : 4) と定義する。また、従属節の意味分類については、益岡・田窪 (1992)、益岡 (1997) を参考に、「実際の文型パターンに関する用例分析の結果に基づき」作成された池原 (2009) の分類体系を基準とする。

作業開始時の確認事項は、述語を含む 2 文節[1]以上からなるものを節とし、節末の接続表現の右端にタグをつけるということであった。作業においては、「節間意味分類体系」(池原 2009) のみを手掛かりに作業を行った。

以下作業の概要について示す。

最初に「鳥バンク」のパターンを UniDic 品詞体系に対応させた節自動解析器により、可能な節境界をすべて枚挙する (浅原ほか 2015: https://github.com/masayu-a/clause_pattern)。この可能な候補を見ながら、作業者 2 名 (作業者 A, 作業者 B) がお互いのアノテーションを見ずに 1 次タグ付け作業を行う。その後、作業者 A が機械の出力および 2 名分の 1 次タグ付け作業結果を見ながら 2 次タグ付け作業を行う。対象は「現代日本語書き言葉均衡コーパス」の新聞記事コアデータの一部 (PN) 54 ファイル (優先順位 00001～00054: A 集合) とした。

この作業過程において、作業者間、作業者内での判断の不一致や揺れが多数生じた。以下、特に、節認定、タグ付け箇所、意味分類で生じた作業者間の不一致や揺れについて、実例を挙げ、その問題が発生した原因について考察を行う。

3. 節の意味分類アノテーションにおける問題点と策定するべき基準

以下では、「句」ではなく「節」として認定するか否か (節認定)、「節境界」をどの位置に設定するか (タグ付け箇所)、「節」の意味分類としてどのラベルを付与するか (意味分類) についてそれぞれ示す。

[1] 文節の定義は、小椋他 (2011) の文節認定規定に従う。

3. 1　節認定について

節認定についての問題点と、それを踏まえた基準策定について述べる。

3. 1. 1　節認定の問題点

節認定の問題に関しては、2 つに分けて論じる必要があると考える。1 つは、何を節と認定するかという問題であり、2 つめは、どこからどこまでを節と認定するかという問題である。

まず、何を節と認定するかという問題について述べる。現行の作業では、「述語を含む 2 文節以上を節とする」という基準で作業をしているが、この場合、以下の例文のような節認定に問題が生じる。

なお、以下の例文について、出典を示していないものは、筆者の作例である。

(1) 彼らは一晩中、飲んだり食べたりしていた。
(2) 彼らは一晩中、酒を飲んだり、つまみを食べたりしていた。

(1) について、「彼らは一晩中、食べたりしていた。」を主節とすると、節認定には 2 文節以上を必要とする現行の作業基準では、「飲んだり」は節と認定しないが、(2) の「酒を飲んだり」は節と認定をする。(1) を節認定せず、(2) を節認定するという妥当な根拠はなく、現行の基準には問題があると考える。

また、丸山他 (2016) が指摘する通り、名詞修飾節の認定にもいくつかの問題がある。

(3) 青いビンを見つけた。
(4) ふたが青いビンを見つけた。

上記の例文では、(3) (4) ともに「青い」は「ビン」という名詞を修飾しているが、現行の作業基準では、(3) は「青い」が単独で「ビン」を修飾しているため、名詞修飾節とは認定されない。それに対し (4) の「ふたが青い」は、2 文節であることから節と認定される。このように、文節数のみを節認定の基準とすることには議論の余地があると考える。

(5) 放置された車がある。
(6) ガレージに放置された車がある。

(6) の「ガレージに放置された」を節として認定することには問題なさそうであるが、(5) の「放置された」の節認定は、どのように判断すればよいか。現行の基準で判断するならば、1 文節の「放置された」は節と認定されないことになる。しかし、「放置された」という、過去の出来事を描写している意味的な特徴を無視して、単独の文節であるという形式を根拠に節認定しないことには、問題がある。さらに「放置された」は、形態素にわけると「放置する＋れる＋た」となることからも、同じ単独文節である (3) の「青い」の節認定の問題とは性質を異にすると考えられる。

丸山他 (2016)　においては、寺村 (1981) が「主体が文脈からわかること、その述語にテンスの意識があること、という 2 点を満たす場合にのみ『節』を認める、という立場をとっている」ことに言及し、「対象の属性を規定する名詞修飾表現は (タ形をとっていても) 連

体節とは認められないことになる[2]。しかし、このような意味的な違いを表層の単語列から判定するのは極めて難しい。」(丸山他 2016 : 1116) としている。この点については、同意するところが多く、特に形式からだけでは、「テンスの意識」があるかどうかの判断は不可能である。

本作業の基準から、属性についての名詞修飾節認定を論じると、「曲がった木」の「曲がった」は単独の文節であることから節認定されず、「先が曲がった木」であれば 2 文節であることから節認定されることになる。このように、属性を表す名詞修飾表現であっても 2 文節以上で現れることはあり得ることであり、この基準にも議論の余地があると言える。

また、どこからどこまでを節と認定するかということについても、以下のような問題が指摘できる。

(7) この部屋は隅々まで掃除するのに 3 日かかった。
(8) この部屋は隅々まで掃除するのにどうしてあの部屋はしないの。

(7) では、「この部屋は」を主題とし、「隅々まで掃除する」を形式名詞「の」に係る名詞修飾節としたうえで、「隅々まで掃除するのに」を「かかった」という述語の補足節と認定するのが、厳密には正しい節認定であろう。ただ、本作業では、作業の手がかりである「節間意味分類体系」 (池原 2009) が、「のに」を補足節の「節間キーワード[3]」としているため、「隅々まで掃除するのに」を補足節と認定し、名詞修飾節の認定は行っていない。この点については、仕様書があることで作業上の揺れはなくなると思われるが、どこまでが実装の際に必要な情報であるかということには、検討の余地があると考える。

(8) については、「この部屋は」を対照主題とし、「隅々まで掃除するのに」は逆接の副詞節である。この「のに」は南の 4 分類[4]において、C 類の副詞節の節形式とされているものであり、「この部屋は」のような主題句は C 類の副詞節の構成要素になりうる。

このように、 (7) と (8) は、表層上は同じであっても、全く異なった文法機能を持った節であるといえる。

(9) 公園を散歩していた時、その事件を目撃した。

(9) では、「公園を散歩いていた時」を、時間を表す副詞節 (時間節) と認定することも、「公園を散歩していた」を「時」の名詞修飾節と認定することも可能である。

以上のように、事態性か属性かという述語の性質、意味、機能を表層上区別する手立てはなく、厳密な節認定を行う方向で議論すると、認定は益々複雑となることが予想される。そこで、文法上の議論を踏まえたうえで、実装を前提としたときの許容範囲を有した妥協点を探索する必要があると考える。その意味でも、揺れを前提とした基準と作業仕様があること

[2] 丸山他 (2016) では、寺村 (1981) のように「お茶がほしい人」を「節」として認め、「やせた人」を「節」ではなく「句」とするという立場をとると、節認定しないことになる例として、「飲むヨーグルト」「やせている男」「曲がった道」を挙げている。

[3]池原 (2009) の用語で、「従属節と主節の意味的な関係を決める接続表現部分」を指す。補足節では形式名詞「こと」「の」「ところ」などを節間キーワードとしている。

[4] 南 (1974) は、日本語の従属節について、節の構成要素を述語的部分の要素と述語的部分以外の要素に分けて論じている。そして、それぞれの要素が節内に存在することの可否を根拠に、従属節を A 類、B 類、C 類、D 類に 4 分類している。

で、齟齬が見られた節を要注意節として、実装時に役立てることができると考える。

次に、節認定における齟齬の発生割合が、従属節によって異なることについて述べる。

一方の作業者は節と認定し、他方の作業者は節と認定していない (タグ付けを行っていない) という節認定齟齬が各従属節で生じている。

作業者のいずれか一方だけが節認定をした節の総数は 631 であった。以下の表 1 に従属節の種類別にその頻度を示す。

表 1　作業者の一方だけが節認定をした従属節の頻度　(　)内は割合(%)

補足節	名詞修飾節	副詞節	並列節
102 (16)	275 (44)	207 (33)	47 (7)

表 1 の割合をグラフで表わしたものが、図 1 である。

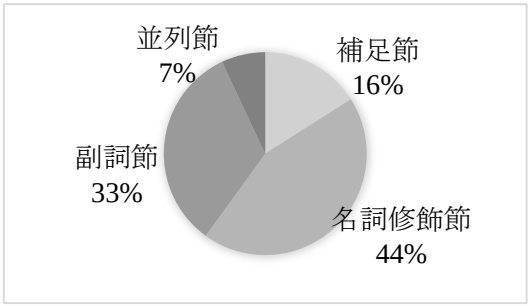


図 1　節認定に齟齬が発生する従属節の割合

作業者の一方のみが節認定した場合を見ると、齟齬の発生割合が従属節の種類によって異なることがうかがえる。作業者の一方が並列節と認定したが、他方はそれを節と認定していないという齟齬の割合が最小で、作業者の一方が名詞修飾節と認定したが、他方が節認定をしなかったという齟齬の割合が最大であった。

これを従属節の種類別にみると、作業者の一方のみが補足節と節認定した齟齬は、その 60%が引用節 : 間接引用において生じている。作業者の一方のみが節認定し、他方は節認定しなかったもので、名詞修飾節の認定における齟齬の 75%が内容節、補足語修飾節で生じ、副詞節では 50%が独立、付帯状況・様態、因果関係の原因において生じている。また、並列節における同様の節認定の齟齬のうち、70%が総記の並列で確認されている。

前述したとおり、述語の性質、意味から節を認定することは極めて難しく、節を文法上厳密に認定しようとすると文法機能上の判断も必要となるため、どの段階までの節情報を付与るすかという設計に関しても、これらの情報は有効であると考える。

3. 1. 2　節認定において策定するべき基準

文節数だけではなく、表層上の形式に着目しつつ、揺れを想定した基準を示すことが現実的であると考える。作業基準の許容範囲の幅をどの程度とするか、またどの深さまで分類する必要があるかは今後の課題である。

基準策定の際、述語になることができる品詞ごとに基準を設けることが考えられるが、本発表は作業上の問題点の整理であるため、例のみの提示とし、具体的な策定基準に関しては

次稿以降に譲る。

例） 動　詞・・・文節数や、活用の有無にかかわらず節認定を行う。

形容詞・・・補語を含まない基本形である場合のみ、節認定をしない。

文法上、節と認めることに議論の余地がある事例に関しても、広めの許容範囲を認め、節情報タグ相当の用法を検討して付与する方向での基準を策定する。

3. 2　タグ付け箇所について

タグ付け箇所についての作業上の問題点と、それを踏まえた基準策定について述べる。

3. 2. 1　タグ付け作業の問題点

タグ付け箇所の齟齬は、作業開始当初、同じ節認定をしているにもかかわらずタグ付け箇所が異なるなど多数生じていたが、作業が進むにつれ、改善、安定がみられ、節の認定上の齟齬がない場合のタグ付け箇所の齟齬は解消されている。

タグ付け箇所の齟齬は File00001 ～ 00014 (5 桁の数字は優先順位を表す https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER 参照) では 148 件確認されたが、file00015～00054 では 54 件に止まった。作業当初のタグ付け箇所の齟齬のほとんどは、名詞修飾節において発生しており、被修飾名詞の直前につけるか、被修飾名詞につけるかというものであったことを考えると、作業者の認識不足が原因であったと思われる。

3. 2. 2　タグ付け作業において策定するべき基準

開始当初と作業途中の差を考えると、仕様書に明記することにより、節認定が一致した場合は、作業者間のタグ付け箇所の齟齬は解消されると考える。

3. 3　意味分類について

意味分類についての問題点と、それを踏まえた基準策定について述べる。

3. 3. 1　意味分類の問題点

意味分類においては、節機能の齟齬[5]か、意味の齟齬かに分けて考える。

まず、節機能の齟齬について述べる。

作業者双方が節と認定をしていたものについて、作業者 A,B が付与したラベルの Level 1 における節機能の分類の結果を表 2 に示す。

表 2　作業者 A,B が付与したラベルの節機能分類 (Level 1)　数字は頻度

A \ B	補足節	名詞修飾節	副詞節	並列節
補足節	8	4	7	8
名詞修飾節	11	179	4	0
副詞節	12	6	85	125
並列節	1	0	26	2

作業者 A を基準にして考えたとき、file:00015~00054 で、作業者間で節機能分類において齟齬があったのは、補足節で 70%、名詞修飾節で 8%、副詞節で 63%、並列節で 93%であ

[5] 池原 (2009) において、副詞節とその他の 3 つの節では、Level 2 以下の分類基準が異なっており、これら全てを「意味分類」と表現することには、議論の余地があると考える。しかし、本稿では便宜上、文法的に従属節を分類した Level 1 を節機能の分類とし、Level 2 以下を意味分類とする。

った。

つまり、作業者 A と作業者 B で、認定した従属節の種類が異なるという節機能分類における齟齬は名詞修飾節で最も少なく、並列節で最も多かったということである。換言すれば、名詞修飾節は、節機能の齟齬は少ないが、名詞修飾節内での意味分類の齟齬が多く生じており、並列節では、従属節間の齟齬が大半を占めているということになる。

表 2 に示した通り、副詞節－並列節間での齟齬は多く見られるが、丸山他 (2016) が採用している分類体系では、「従属節を大きく 3 つに分けて、並列節を連用節の下位に位置づけて」(丸山他 2016 : 1114) おり、丸山他 (2016) の分類体系に基づくと、機能的なレベルでの作業者間の意味分類の齟齬はほとんど見られないことになる。

次に、作業者間において Level 1 の機能分類は一致していたが、Level 2 以降の意味分類で齟齬が生じた節について述べる。

Level 2 以下の意味分類の齟齬は、並列節、補足節ではほとんど生じておらず、齟齬の頻度が 5 を超える節はなかった。以下に、作業者間の齟齬の頻度が 10 以上であった節について、実例を挙げて詳説する。

まず、作業者双方が Level 1 で名詞修飾節と認定したが、それ以下の分類において作業者間に齟齬が生じたものを表 3 に示す。

表 3　名詞修飾節における作業者間の意味分類齟齬

頻度	作業者 A		作業者 B	
	Level 2	Level 3	Level 2	Level 3
46	補足語修飾節	限定的	補足語修飾節	非限定的
27	補足語修飾節	非限定的	補足語修飾節	限定的
23	内容節		補足語修飾節	限定的
14	縮約形修飾節		内容節	
14	縮約形修飾節		補足語修飾節	限定的
10	補足語修飾節	限定的	内容節	

表 3 に示した順に、名詞修飾節において Level 2 以下の意味分類に齟齬が生じた節から 1 例ずつ挙げる。

作業者が付与したラベルの内容を、作業者 A Level 2 (level 3) – 作業者 B Level 2 (level 3) と示し、例文を挙げて齟齬の解釈を行う。なお、例文の下線部は作業者間に意味分類の齟齬が生じた節である。

補足語修飾節 (限定的) －補足節修飾節 (非限定的)

(10) 雪舟作と伝えられる花鳥図屏風は、10 点余りが知られている。
[PN2b_00002, file:00019]

(10) は、補足語修飾節という判断で両者は一致しているが、「花鳥図屏風」が指す対象が一定しているか否かの判断において、作業者間の認識が一致しなかったため、Level 3 のタグに齟齬が生じたものである。

補足語修飾節 (非限定的) －補足節修飾節 (限定的)

(11) 警察当局が危険人物と認定した九百三十二人に対し、W杯開催の五日前までにパスポートを警察署に預ける命令 が出ているが、 [PN2c_00002, file:00020]

(11) も、(10) と同様に、補足語修飾節という判断で両者は一致しているが、「九百三十二人」が指す対象が一定しているか否かの判断において、作業者間の認識に不一致があったため、Level 3 のタグに齟齬が生じたものである。

内容節－補足語修飾節 (限定的)

(12) 再建計画に数値基準を設けた中間報告の中核的な考えに反映されている。 [PN1g_00002, file:00018]

(12) は、下線部が「中間報告」の内容を表し、修飾節と被修飾名詞が同格にあると考えた作業者 A に対し、作業者 B は、下線部は「中間報告」の指し示す対象を限定していると判断したといえる。

縮約形修飾節－内容節

(13) 診療所存続をめぐる話題が一本の柱だ。[PN1b_00003, file:00033]

(13) では、「診療所存続をめぐる」と「話題」における格関係の有無の判断により齟齬が生じた例である。作業者 A は修飾節と被修飾名詞に格関係はなく、「意味的に間接的な関係にある」 (池原 2009：293) と判断し、作業者 B は格関係を認めたものと考えられる。

しかし、益岡・田窪 (1992) では、池原 (2009) が縮約形修飾節としているものは全て内容節に含めており、格関係の有無に関係なく、「内容節とは、被修飾名詞が指し示す対象の内容を表す」(益岡・田窪 1992：203) と広く定義している。

このように、意味分類はその定義によっても判断が異なることがあり、どの深さまで分類するかは、議論の余地のあるところである。

縮約形修飾節－補足語修飾節 (限定的)

(14) 北朝鮮の核不拡散条約 (NPT) からの脱退宣言が 10 日に 90 日間の「脱退通告期間」切れとなるとされるのを受けた協議だが、（後略） [PN3a_00002, file:00024]

(14) において、作業者 A は修飾節と被修飾名詞は、格関係にも同格関係にもなく、間に「上で行われた」などが省略されていると判断し、作業者 B は修飾節が「協議」の指し示す対象を限定していると判断したものである。

補足語修飾節 (限定的) －内容節

(15) だが上品な画題とは似ても似つかぬ印象は、この屏風が型破りな作品を生涯描き続けた雪舟のまさに真筆だと明かしているようにも思う。[PN2b_00002, file:00019]

(15) で、作業者 A は、修飾節が被修飾名詞「印象」の指し示す対象を限定していると判断し、作業者 B は修飾節が被修飾名詞の内容を表すものであり、両者が同格であると判断している。

以上の齟齬の例をみると、名詞修飾節においては、修飾節と被修飾名詞の関係に判断の齟齬が見られるものと、被修飾名詞の性質についての判断に齟齬が見られるものがあることが明らかになった。

次に、作業者双方が Level 1 で副詞節と認定したが、それ以下の分類において作業者間に齟齬が生じたものを表 4 に示す。

表 4　副詞節における作業者間の意味分類齟齬

頻度	作業者 A		作業者 B	
	Level 2	Level 3	Level 2	Level 3
33	手段		因果関係	原因
12	付帯状況・様態	付帯状況	因果関係	原因

表 4 に示した順に、副詞節において Level 2 以下の意味分類に齟齬が生じた節から 1 例ずつ挙げる。

作業者が付与したラベルの内容を、作業者 A Level 2 (level 3) – 作業者 B Level 2 (level 3) と示し、例文を挙げて齟齬の解釈を行う。なお、例文の下線部は作業者間に意味分類の齟齬が生じた節である。

手段－因果関係 (原因)

(16) 他派閥からも引き抜いて三十人から五十人の新派閥をつくることができるんだ」。
[PN2e_00002, file:00021]

池原 (2009) では[6]、手段を表す副詞節は「主節の内容を行う前提」(池原 2009 : 300) を表す従属節であり、因果関係（原因）は、「従属節の内容が原因となって主節の内容が起こる」という「従属節と主節で表される事態間の因果関係を表す」(池原 2009 : 296) としている。

「主節の内容を行う前提」は当然のことながら、主節で表される事態に因果関係を持つものであることから、この齟齬は節の意味における分類の難しさ、曖昧さから生じるものであると考える。

付帯状況・様態 (付帯状況)－因果関係 (原因)

(17) 小泉内閣は、細川、橋本、小渕の各政権の積み残しを一手に引き受けて、そのすべてを処理するという重荷を背負っている。[PN1b_00004, file:00046]

これは作業者 A が、「引き受けて」を「処理する」に係る節と判断したことにより、付帯状況・様態 (付帯状況) のラベルを付与し、一方作業者 B は「引き受けて」を「背負っている」に係る節と判断したため、因果関係 (原因) のラベルを付与したと推測される。(17) は、作業者間でかかり受けの判断が異なったため、意味分類に齟齬が生じたと考える。

従属節の分類の難しさについては、南 (1974) が次のように述べていることからもうかがえる。

[6] 池原 (2009) は、益岡・田窪 (1992) および益岡 (1997) を参考にしているが、本稿では池原 (2009 : 292-303) の 8 章の付録表「3. 主節従属節間の意味分類体系」から引用している。

> (前略) 問題の従属句がどの類に属するかは、必ずしも、テとかナガラ、バといった接続助詞のいかんによってあらかじめきめられるものではないということである。それは、その句を構成しているすべての要素およびその句の文中での文法的性格による。
>
> (南 1974：130)

副詞節における意味分類は、主節と従属節の関係を文脈から判断することも多く、かかり受けや、節関係の判断に作業者の言語感覚や文法理解のレベルなどによって齟齬が生じることは必須である。

3. 3. 2 意味分類において策定するべき基準

池原 (2009) の意味分類体系においては、Level 2 以下に機能、意味、形式の混在が見られところがある。意味分類においての基準の統一について、その必要の有無も含め、どのように解決していくかという問題がある。

また、例えば、節間キーワードの「ところに」は、補足節の「トコロ型」にも、時を表す副詞節にも同義のものが示されており、作業者の混乱を招くと考えられる。ただ、これは意味分類体系に問題があるわけではなく、複数の意味に解釈される節末接続形式は存在するため、ラベル付与作業における工夫が求められる。

さらに、池原 (2009) の節間キーワードは、「約 1000 件の文型パターンを対象に、従属節と主節とを連結するキーワードとその意味に着目した用例分析を行い、分類を詳細化した」(p.259) ものである。キーワードは、形式に着目したものであるため、助詞が多いものの、助動詞も相当数見られ、動詞も混在しているなど、抽出方法に文法的一貫性がない。一般化、法則化が難しいようであれば、池原 (2009) の分類体系にパターンを追加、修正していくことが求められると考える。

意味分類において、池原 (2009) は、「彼らは会えば必ずけんかする。」の「ば必ず」を、時を表す副詞節の節間キーワードとしている。一方で、法則的条件を表す副詞節の定義として「ある事態が起こると法則的に必ず別のある事態が起こる」ことを表すとしている。ここには矛盾があり、文法的にも「会えば」を時間節とすることには議論の余地がある。他にもいくつか同様のものが見られ、若干の修正が必要ではないかと考える。

4. おわりに

本稿では、節境界アノテーション・節の意味分類タグ付けに関する現行作業の問題点を整理し、より効率的に、信頼性の高い結果を得るための議論を行った。そして、作業において、文法的に厳密で正確な判断を追及することを目的としたものではなく、文法上の議論の余地があるものについても、アノテーションとしていかに記号化していくかという観点で、許容範囲を探り、節認定や意味分類の揺れを整理することを試みた。

副詞節・並列節間の節認定の齟齬は、これらの節の分類の解消か、機能の相違点を明示し厳密に分類するかという議論をもたらす。また、従属節内の意味分類 (level 2 以下) で頻度の高い齟齬を見ると、文脈からの判断が必要なものもあり、名詞修飾節における補足語修飾節の限定か非限定かの分類や、内容節－補足語修飾節－縮約形修飾節間の齟齬を見ると、何を目的にどこまでの分類を行うか、解析器にどこまでの分類を求めるかという議論が必要

であることも示唆される。

節認定にせよ、意味分類にせよ、浅い分類であれば安定しやすいが、深くなるほど分類が細かくなり、語の性質、意味、機能などの解釈による齟齬が生じやすくなる。そこで、実装にはどこまでの情報付与が求められているかということも勘案しながら、作業の基準策定をしていくことが求められるであろう。

一旦は許容範囲を広く認め、取りこぼしなく節情報を付与することで、次の段階の意味処理において、目的に応じた分類を行うなど、利用範囲も広がると考える。

文法上の正確さと、解析器への実装という問題の最適の妥協点を探るために、現行の作業による作業者間の齟齬を中心とした問題点を述べたが、これらを基にした作業基準を策定し、作業仕様書を作成することができれば、作業の効率化も図れ、データの信頼性も向上すると考える。また、一定の基準に基づいた作業における齟齬が、節の意味機能に関する新たな発見とコーパス言語学の更なる発展に繋がることも期待できる。

謝　辞

本研究は JSPS 科研費(課題番号 : 15K12888, 研究代表者 : 浅原正幸)の助成を受けている。

文　献

浅原正幸・小西光・田中弥生・加藤祥 (2015) 「品詞列・係り受け部分木に基づくラベリングツールの設計と実装―節境界ラベリングを例に―」 第 8 回コーパス日本語学ワークショップ pp.83-92.

池原悟 (2009) 『非線形言語モデルによる自然言語処理』 岩波書店

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011) 「『現代日本語書き言葉均衡コーパス』 形態論情報規定集第 4 版（上)」 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

寺村秀夫 (1981) 『日本語の文法（下)』日本語教育指導参考書(5) 国立国語研究所

益岡隆志 (1997) 『複文』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

丸山岳彦・佐藤理史・夏目和子 (2016) 「現代日本語における節の分類体系について」言語処理学会第 22 回年次大会発表論文集 pp.1113-1116

南不二男 (1974) 『現代日本語の構造』大修館書店

日本語話し言葉コーパスにおける発声様式の自動分類

森 大毅 (宇都宮大学大学院工学研究科) *
藤本 雅子 (国立国語研究所)
浅井 拓也 (北陸先端科学技術大学院大学)
前川 喜久雄 (国立国語研究所)

Automatic classification of phonation type in the Corpus of Spontaneous Japanese

Hiroki Mori (Utsunomiya University)
Masako Fujimoto (National Institute for Japanese Language and Linguistics)
Takuya Asai (JAIST)
Kikuo Maekawa (National Institute for Japanese Language and Linguistics)

要旨

喉頭音源由来の声質の違いは，話者のパラ言語メッセージならびに心的・認知的状態を伝えるシグナルであり，自発音声コーパスに求められる重要な情報であるが，そのアノテーションは音声学の専門家でなければ難しくコストが大きい。本研究は，機械学習による声質の自動アノテーションの可能性を探ることを目的とする。本研究では，非流暢性にも関連する従来よく用いられてきた発見的な音響特徴量に加え，近年音声からの感情認識で広く用いられるようになった大規模な特徴量セットの効果を検証した結果を報告する。

1. はじめに

書き言葉と異なり，話し言葉はさまざまな音響的手がかりによって話者の意図や態度の違い、すなわちパラ言語情報を伝達している。また同時に、心的状態のように、話者の伝達の意志とは関係なく伝達される情報もある (森ほか 2014)。

話者のパラ言語メッセージや心的・認知的状態を反映する音響的手がかりには、ピッチや大きさの抑揚、テンポ、リズムなどの韻律や、母音の質に代表される声道の特徴が含まれる。中でも、喉頭における発声様式の違いに起因する声質 (Laver 1980, Ni Chasaide and Gobl 1997) は、生成面に着目した声質の基本的記述として、音声学的にも工学的にも重要である。また、自発音声に現れる声質の違いは、心的・認知的状態を無意識に伝える社会的シグナルであり、実時間コミュニケーションの動的な構成に関わるターンテイキングや非流暢性などの現象とも密接な関係がある。このため、これらの研究の基礎資料となる自発音声コーパスには、声質のアノテーションを有することが望まれる。

しかしながら、自発音声の場合には発声中の喉頭の観察は難しい。また、『日本語話し言葉コーパス』(CSJ) のように、音声以外に測定された信号を有しないコーパスも多い。この場合、

* hiroki-public@speech-lab.org

声質の記述を声帯振動の観察により行うことは不可能であり、そのかわり、専門家が聴覚的に判断する必要がある。コーパス構築において、このような声質のアノテーションに要するコストは膨大であり、このことが大規模なデータに基づいた研究を難しくしている。

本研究は，機械学習による声質の自動アノテーションの可能性を探ることを目的とする。著者らはこれまで、CSJ に含まれる長母音/aH/ /eH/を対象に、機械学習アルゴリズムによりフィラーの判別を試みた (Maekawa and Mori 2016)。音響的特徴として継続時間、強度、F0、フォルマント周波数、ジッタ、シマ、調波対雑音比、スペクトル傾斜を用い、フィラーと語彙項目のサンプルが同数の条件で交差検証を行った結果、同定精度は $F = 0.89$ であった。フィラーの多くは通常の語彙項目とは異なる声質で発声されていると考えられるため、提案した手法は声質の自動分類においても有効である可能性がある。

近年は、機械学習技術の発達に伴い、音声から話者の感情や年齢・性別などの情報を推定する問題において、これらと強く関連すると予想される少数の音響パラメータを使うのではなく、網羅的に抽出された音響パラメータ列 (LLD; 4.2 参照) から組織的に生成された非常に多数の要約統計量をそのまま使って機械学習を行うことが一般的になってきた。この種の方法は一見たいへん非効率であるが、少数精鋭のパラメータセットを用いる場合に比べ、性能が大きく改善する場合が多い。本研究では、過去の研究で用いられてきた発見的な特徴量に加え、これらの大規模な特徴量セットを用いることの有効性もあわせて検証する。

2. 対象とする声質

2.1 Creaky voice

Creaky voice(きしみ声) とは、creak または vocal fry と呼ばれる、極端に低いピッチやパルス的な音で特徴づけられる発声様式の特徴をある程度有する声であることを意味する。自発音声においては、ピッチの低下と同じように、次のような場所・場面でよく観察される。

- フィラー
- 発話末や句末
- 自信のない心理状態

Ishi et al. (2008) は、周期性と声帯パルスの類似性を利用した vocal fry の検出法を提案し、自然発話データに対して 73% の検出率と 13% の挿入誤り率を達成したと報告している。ただし、対象としたデータは、Sadanobu (2004) が「りきみ」と呼ぶ pressed かつ creaky な発声に限られており、一般の creaky voice に対する提案手法の有効性は明確ではない。

2.2 Breathy voice

Breathy (気息性) な声は、声門閉鎖の不完全により、声帯振動による周期音と同時に生じる乱流雑音によって生成される。乱流雑音の程度および様態により、whisper → whispery → breathy → modal と様々な語によって形容される。すなわち、気息性は程度の問題であり、非気息性の声との間に明確な境界は存在しない。Breathy な声は、ある種の個人性を特徴づけるほか、落胆などの心的状態を反映した低緊張の状態に関連する。音響的には、大きなスペクトル傾斜と低い調波対雑音比により特徴づけられる (Klatt and Klatt 1990)。

表 1　サンプル数

(a) フィラー

母音	全データ数	creaky	breathy
/a/	67	20	11
/e/	110	42	34
/i/	0	0	0
/o/	54	19	9
/u/	0	0	0

(b) 通常語彙項目

母音	全データ数	creaky	breathy
/a/	222	19	31
/e/	126	21	24
/i/	84	8	14
/o/	182	15	22
/u/	53	1	6

3. データ

データとして、CSJ のコア部分中の独話 (学会講演と模擬講演) を使用する。母音のみから構成されるフィラー 231 個、および通常の語彙項目中の母音 667 個をランダムにサンプリングした。

声質のラベリングは、発声様式に関する豊富な研究経験を持つ第 2 著者および第 4 著者が行った。2 名のラベラーは、全てのサンプルを 2 回ずつ聴取し、“creaky”, “breathy”, “modal” のいずれであるかを判定した。1 発声の中で、creaky から modal、または modal から creaky へのように変容が生じていると判断される場合には、その旨を記述した。

表 1 に、ラベリング結果から得た声質の分布を示す。creaky および breathy に分類されているサンプルは、ラベラー 2 名による全 4 回の判定のうち、1 回でも creaky または breathy と判定されたものである。

4. 音響特徴量

4.1 基本 15 次元

今回検討する音響特徴量のうち、この節で説明するものは、著者らが過去にフィラーの分析に使用したものである (Maekawa and Mori 2016)。

■継続時間　継続時間 (duration) は対象母音の始点から終点までの時間である。

■強度　強度 (intensity) は、単位をデシベルとして求めたものを対象母音区間において平均した。

■基本周波数　F0 は、対象母音区間から Praat の Pitch(ac) コマンドにより求め、対数を取った後に、話者ごとに平均と標準偏差を正規化した。また、ピッチの変動の指標として標準偏差 (sdPitch) を求めた。

■フォルマント周波数　フォルマント周波数 (F1, F2, F3) は、線形予測分析によって得られたものを対象母音区間において平均し、対数を取った後に、話者ごとに平均と標準偏差を正規

化した。

■ジッタ、シマ　基本周波数ゆらぎであるジッタ (jitter) としては PPQ5 (Gelzinis et al. 2008) を、振幅ゆらぎであるシマ (shimmer) としては APQ5 を用いた。

■スペクトル傾斜　スペクトル傾斜として、4 種類の特徴量を用いた。C1TL は、対象母音区間の線形予測分析から得られるスペクトル包絡の情報である LPC ケプストラムの第 1 係数から求めたスペクトル傾斜である (前川・森 2005)。H1−H2, H1−A1, H1−A2, H1−A3 は、それぞれ第 2 高調波成分、第 1 フォルマント、第 2 フォルマント、第 3 フォルマントの基本波成分に対する振幅の比である (Gordon and Ladefoged 2001)。

■調波対雑音比　調波対雑音比 (HNR) は Praat により求めた。

4.2　openSMILE

近年の音声からの感情認識研究では、数種類のパラメータを厳選するのではなく，非常に多くのパラメータ，およびそれらから二次的に導出される発話単位の最大・最小・レンジ・平均・四分位数・百分位数・標準偏差などの要約統計量を無制限に利用する方法が主流になっている。音声に関する重要な国際会議の 1 つである Interspeech では 2009 年より感情をはじめとしたパラ言語情報の認識精度を競うコンテストが開催されており， 2013 年に開催されたコンテスト Computational Paralinguistics Challenge における標準特徴は，エネルギー関連の 4 種類，スペクトル関連の 54 種類，有声音源特性に関する 6 種類からなる計 64 種類の低次特徴量 (Low Level Descriptor: LLD) を基に導出された，計 6373 の特徴から構成されている。

このような目的に特化した音響特徴量抽出ソフトウェアに openSMILE (Eyben et al. 2010) がある。今回は、openSMILE を用い、対象母音区間から Interspeech 2009 Emotion Challenge (Schuller et al. 2009) におけるベースライン特徴量を抽出した。これらは、対象母音区間の各分析フレームにおけるゼロ交差率、実効値、F0、HNR、MFCC(メル周波数ケプストラム係数) 12 次元、およびこれらの 1 次回帰係数からなる 32 種類の低次特徴量の系列に対し、区間単位での平均、標準偏差、歪度、尖度、最大値／最小値およびその位置、レンジ、線形回帰係数およびその平均 2 乗誤差からなる 12 種類の汎関数を適用して得られる 384 次元のベクトルである。

5.　声質の自動分類実験

機械学習アルゴリズムとしてサポートベクターマシン (SVM) およびランダムフォレスト (RF) を使い、声質の自動分類を行った。実験は、creaky voice を発見するタスク、および breathy voice を発見するタスクそれぞれについて行った。表 1 に示したように、声質ごとのデータ数は大きく異なる。このため、学習時には、データ数に反比例したコストを定義し、非 creaky または非 breathy に偏って学習されるのを防いだ。機械学習アルゴリズムのパラメータは F 値を基準に調整し、評価は leave-one-out 交差検証法により行った。評価尺度は F 値および ROC 曲線下面積であり、ともに 1 に近いほど検出性能が高い。

667 個の母音全てを対象にした実験の結果を表 2 に示す。表 2 から、SVM に比べランダム

表 2　自動分類結果

(a) F 値

	creaky	breathy
基本 15 次元 (SVM)	0.552	0.581
基本 15 次元 (RF)	0.562	0.624
15 + 384 次元 (SVM)	0.523	0.619
15 + 384 次元 (RF)	0.590	0.619

(b) ROC 曲線下面積

	creaky	breathy
基本 15 次元 (SVM)	0.786	0.797
基本 15 次元 (RF)	0.875	0.876
15 + 384 次元 (SVM)	0.741	0.783
15 + 384 次元 (RF)	0.872	0.903

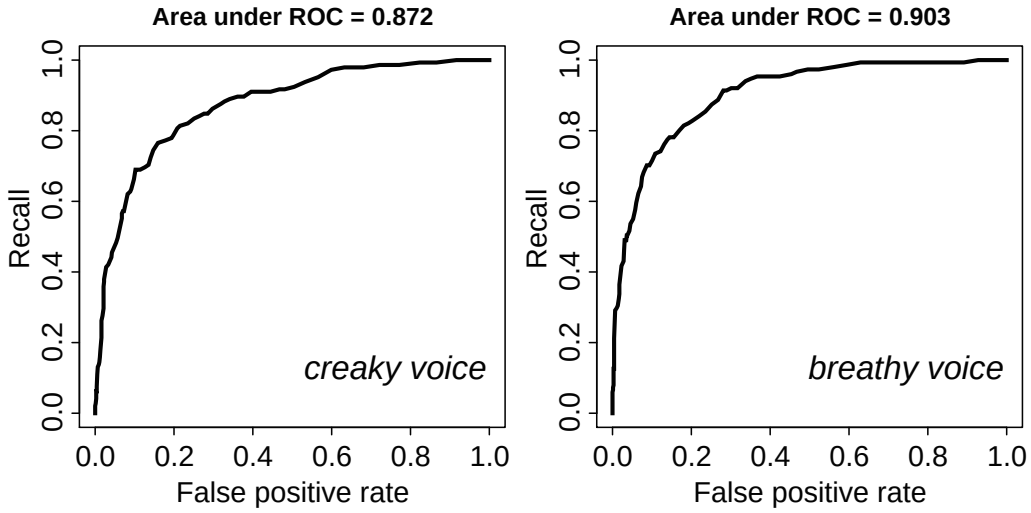


図 1　ROC 曲線

フォレストの方が性能が高いことがわかる。また、openSMILE で抽出した 384 次元を追加することで精度が向上したケースは少なく、特徴の追加の効果は限定的であったと言える。図 1 に、特徴量として基本 15 次元 + openSMILE 384 次元を用い、ランダムフォレストにより分類を行った場合の偽陽性率 (false positive rate) と再現率 (recall) の関係、いわゆる ROC 曲線を示す。この図は、例えばラベラーが creaky voice または breathy voice と判定した母音のうち 90% 以上を検出しようとすれば、ラベラーが creaky voice または breathy voice と判定しなかった母音のうちそれぞれ 39.6% および 28.1% 以上は誤って creaky voice または breathy voice と判定されてしまうことを意味する。

次に、用いた特徴量の重要度を、ランダムフォレストにおける平均不純度減少量を基準に調べた。Creaky voice 検出および breathy voice 検出における重要特徴量の上位を、表 3 (基本 15 次元) および表 4 (基本 15 次元 + openSMILE 384 次元) に示す。 基本 15 次元だけを特徴量として用いた場合、C1TL は、creaky voice 検出の重要度が高い特徴としてコンスタントに上位を占めていた。また、周期性のゆらぎである PPQ5, APQ5 も上位を占めていた。この結果は、スペクトル傾斜が小さく周期性が低い creaky voice の特徴と符合している。C1TL, PPQ5, APQ5 はまた、breathy voice 検出の重要度が高い特徴でもあった。この結果

表 3　重要特徴量 (基本 15 次元: 上位 10)

creaky voice		breathy voice	
0.44	C1TL	0.42	C1TL
0.43	PPQ5	0.41	PPQ5
0.41	APQ5	0.4	HNR
0.36	sdPitch	0.39	APQ5
0.35	H1−H2	0.38	sdPitch
0.35	HNR	0.32	H1−H2
0.33	F1	0.32	H1−A1
0.33	F2	0.31	F1
0.32	H1−A1	0.31	F3
0.31	H1−A3	0.3	H1−A3

表 4　重要特徴量 (基本 15 次元 + openSMILE 384 次元: 上位 10)

creaky voice		breathy voice	
0.58	RMSenergy range	0.67	ΔMFCC$_7$ maxPos
0.57	MFCC$_3$ min	0.66	MFCC$_2$ range
0.56	MFCC$_4$ stddev	0.62	MFCC$_1$ stddev
0.56	MFCC$_2$ minPos	0.62	MFCC$_3$ maxpos
0.54	RMSenergy stddev	0.59	ΔMFCC$_4$ minpos
0.54	MFCC$_2$ amean	0.57	MFCC$_2$ linregc1
0.54	MFCC$_1$ stddev	0.57	MFCC$_3$ minpos
0.53	F2	0.56	MFCC$_2$ minpos
0.53	MFCC$_3$ max	0.55	F1
0.53	C1TL	0.55	sdPitch

は、スペクトル傾斜が大きく周期性が低い breathy voice の特徴と符合している。基本 15 次元と openSMILE の 384 次元を併用した場合、creaky voice, breathy voice ともに、スペクトルの情報である MFCC に関連した特徴が上位を占めている。また、creaky voice についてはRMSenergy(実効値) すなわち音の強さに関連した特徴が含まれている。

6. おわりに

本研究では、『日本語話し言葉コーパス』(CSJ) に含まれる母音の発声様式 (creaky, breathy) のラベリングを行い、機械学習による声質の自動アノテーションの可能性を探った。

サポートベクターマシン (SVM) およびランダムフォレスト (RF) を用いた声質の分類実験の結果、RF の性能が SVM の性能を上回った。しかしながら、検出性能の指標である F 値は creaky voice で最大 0.59, breathy voice で最大 0.62 程度であり、人手によるアノテーション

を代替するほどの精度は得られないことがわかった。

本研究ではまた、過去の研究で用いられてきた発見的な特徴量に加え、openSMILE と呼ばれる音響特徴量抽出ソフトウェアを用いた大規模な特徴量セットを併用することの有効性も検証したが、その効果は限定的であった。

本稿で述べた手法を声質アノテーションに応用しようとする場合、専門家によるアノテーション作業の補助に使うことが考えられる。しかし、本稿で述べたように、例えば再現率 90% を目標とすると、偽陽性率は 30% から 40% 程度となり、人手による確認の負荷はかなり大きくなることが予想される。今後は、ラベラー間一致度などを基に検出した非 modal 声質の信頼度を推定する方法を検討し、自動アノテーションの実用化につなげたい。

謝辞

本研究は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および科研費（課題番号 26284062, 研究代表者 前川喜久雄）の成果である。

文　献

森大毅・前川喜久雄・粕谷英樹 (2014).『音声は何を伝えているか—感情・パラ言語情報・個人性の音声科学—』　コロナ社.

John Laver (1980). *The Phonetic Description of Voice Quality. Cambridge Studies in Linguistics.*: Cambridge University Press.

Ailbhe Ni Chasaide, and Christer Gobl (1997). “Voice source variation.” William J. Hardcastle, and John Laver (Eds.), *Handbook of Phonetic Sciences*. Oxford: Blackwell. pp. 1–11.

Kikuo Maekawa, and Hiroki Mori (2016). “Voice-Quality Difference Between the Vowels in Filled Pauses and Ordinary Lexical Items.” *Proc. Interspeech 2016*, pp. 3171–3175.

Carlos Toshinori Ishi, Ken-ichi Sakakibara, Hiroshi Ishiguro, and Norihiro Hagita (2008). “A Method for Automatic Detection of Vocal Fry.” *IEEE Transactions on Audio, Speech, and Language Processing*, 16, pp. 47–56.

Toshiyuki Sadanobu (2004). “A natural history of Japanese pressed voice.” 音声研究, 8:1, pp. 29–44.

Dennis H. Klatt, and Laura C. Klatt (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers.” *Journal of Acoustical Society of America*, 87:2, pp. 820–857.

Adas Gelzinis, Antanas Verikas, and Marija Bacauskiene (2008). “Automated speech analysis applied to laryngeal disease categorization.” *Comput. Methods Prog. Biomed.*, 91:1, pp. 36–47.

前川喜久雄・森大毅 (2005).「フィラーの声質上の特徴に関する予備的分析」　日本音響学会講演論文集, pp. 293–296.

Matthew Gordon, and Peter Ladefoged (2001). “Phonation types: a cross-linguistic overview.” *Journal of Phonetics*, pp. 383–406.

Florian Eyben, Martin Wöllmer, and Björn Schuller (2010). “openSMILE: The Munich versatile and fast open-source audio feature extractor.” *Proc. ACM Multimedia*, pp. 1459–1462.

Björn Schuller, Stefan Steidl, and Anton Batliner (2009). “The Interspeech 2009 Emotion Challenge.” *Proc. Interspeech 2009*, pp. 312–315.

近代文語文の通時的変化の分析
—語種率・品詞率に着目して—

近藤 明日子（国立国語研究所コーパス開発センター）†

Diachronic Variation in Modern Japanese Literary Text: Analysis Based on Etymological Types Ratios and Part of Speech Ratios

KONDO Asuko (National Institute for Japanese Language and Linguistics)

要旨

近代の非文芸ジャンルの文語文の通時的変化の実態を明らかにすることを目的として、『日本語歴史コーパス　明治・大正編 I 雑誌』（短単位データ 1.0）を利用した語種率・品詞率の通時的変化について分析・考察し、次の(1)～(4)の結論を得た。(1)名詞率の増加が見られ、文語体の使用の場が評論的・随筆的文章から報道的文章に移行したことが背景として考えられる。(2)男性向け雑誌では漢語率の増加が見られ、名詞率の増加の影響だけでなく、語彙自体の漢語率の増加も背景として認められる。(3)女性向け雑誌では男性向け雑誌より漢語率が低い傾向が見られ、女性と和文体との強い関係が認められる。(4)接頭辞率・接尾辞率の増加が見られ、近代語における字音接辞の発展という事象が数値として確認できる。

１．はじめに

日本近代語の文章史・文体史は言文一致運動による口語体の成立・定着によって特徴付けられる。その口語体は小説では明治 30 年代に一般化したが、一方で、実用文とも呼ばれる論説文・報道文等の非文芸の文章において口語体が定着したのは大正期に入ってからであり、明治期は依然として文語体が主に用いられた。そして、明治初期には漢文訓読体・和文体・欧文直訳体等の複数の種類の文語体が行われたものが、しだいに融合し、明治後期に普通文と呼ばれる標準的な文語体が確立・定着することが知られる。しかし、標準的ともされる普通文の実態は、漢文脈の濃厚な文体から和文脈・洋文脈の濃い文体あるいはこれらの混交した文体までさまざまあり、決して一つの統一した文体ではなかった（森岡 1991a、p.25）。

このような複雑な様相を呈する近代の文語体実用文の変遷の実態の一端を明らかにするため、本研究では明治・大正期の雑誌の大規模なコーパスである国立国語研究所（2016）『日本語歴史コーパス　明治・大正編 I 雑誌』（短単位データ 1.0）[1]（以下、「CHJ 近代雑誌」という）を資料として、基本的な文体指標である語彙の語種率・品詞率の通時的変化の分析を行う。近代の文語体実用文に関する先行研究は多くあるが、対象資料の規模や扱う期間および調査する言語項目の数に限りのあるものが多く、近代の長い期間の資料を用いて通時的変化を俯瞰するような研究はまだ多くはない[2]。本研究では、先行研究に示唆を受けつつ、語種率・品詞率という新たな観点から通時的変化を大局的に捉えることを試みる。

† kondo@ninjal.ac.jp

[1] http://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html

[2] 先行研究中、比較的長い期間の資料に基づき網羅的に言語項目を調査・分析するものとして、明治期の小学校教科書の助動詞を分析する岡本（1980）、明治・大正期の新聞の文末表現を分析する進藤（1981）、明治・大正期の小学校理科教科書や明治期の新聞の副詞・接続表現を分析する松崎（2006a・2006b）等がある。

２．利用するデータ

本研究では、CHJ 近代雑誌全 7,061 サンプルのうち、以下の①～④の条件を満たす 3,098 サンプルを抽出し、サンプル中の文語体地の文を分析対象とした[3]。

① ジャンルが「非文芸」（小説・戯曲・詩歌以外）のサンプル
② 文体が「文語」で本文種別が空値（＝地の文）の延べ短単位数が 100 以上のサンプル
③ サンプル ID の下 3 桁が「000」（雑誌本体の構造要素を扱うサンプル）のサンプルは除く
④ 著者の生年より判断し、近代より前に書かれた文章が地の文のサンプルは除く

表 1 に対象サンプルの言語量を雑誌種類別に示す。雑誌種類は雑誌タイトルと刊行年によって 10 種類に区別した。

表１　雑誌種類と言語量

雑誌種類（略称）	雑誌タイトル	刊行年	サンプル数	文語体地の文短単位数
明六雑誌（明六）	明六雑誌	1874･1875	149	159,270
国民之友（国民）	国民之友	1887･1888	1,099	907,483
太陽1895年（太陽Ⅰ）	太陽	1895	582	1,457,436
太陽1901年（太陽Ⅱ）	太陽	1901	391	1,282,299
太陽1909年（太陽Ⅲ）	太陽	1909	246	642,316
太陽1917年（太陽Ⅳ）	太陽	1917	75	235,373
太陽1925年（太陽Ⅴ）	太陽	1925	15	5,163
女学雑誌（女雑）	女学雑誌	1894･1895	463	421,777
女学世界（女世）	女学世界	1909	76	53,740
婦人倶楽部（婦人）	婦人倶楽部	1925	2	331
計			3,098	5,165,188

雑誌種類は読者層により大きく 2 つのグループに分けられる。明六雑誌・国民之友・太陽の 3 誌は男性の知識層が主な読者層であり（有山 1986、永嶺 1997）、女学雑誌・女学世界・婦人倶楽部の 3 誌は女学生や職業婦人といった知識層の女性が主な読者層である（田中 2006）。次節以降の分析結果に見られるとおり、この 2 グループ間の特に語種率のありようには大きな違いがある。

なお、両グループにおいて 1909 年以降言語量が急減するのは、当時、口語体の拡大・定着に応じて文語体が縮小・衰退していったことの反映である。1925 年の太陽・婦人倶楽部において非文芸のサンプルで文語体はほぼ用いられなくなっていたことが分かる。この言語量の少ない 1925 年の太陽と婦人倶楽部については次節以降の分析結果の扱いに注意が必要である。特に婦人倶楽部は言語量が極めて少なく、分析に堪えないと考え、以下の分析結果・考察ではとりあげない。

３．通時的変化

３．１．語種率の通時的変化

まず、語種率の通時的変化を見ていく。CHJ 近代雑誌の短単位の語種は 8 種類に分類されるが、その中で主要な 4 語種（漢語・和語・外来語・混種語）についてサンプル単位の

[3] 以下、コーパスデータの検索・集計は「CHJ 明治・大正編雑誌」のデータが収録されている国立国語研究所のデータベースを利用して行った。

比率の分布を雑誌種類ごとに箱ひげ図で示したものが図 1 である。なお、語種率の分析では、助詞・助動詞は対象外とした。

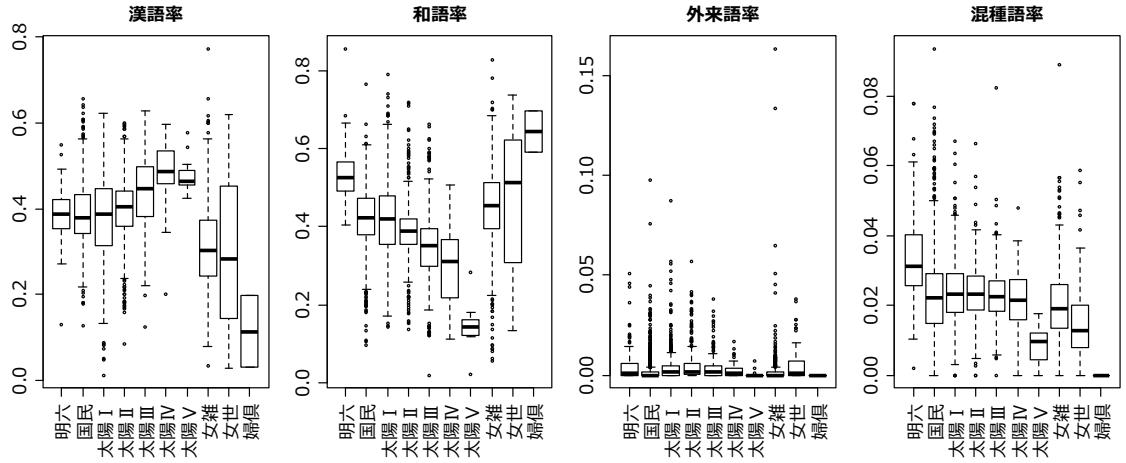


図 1　雑誌種類別に見る語種率の分布

明六雑誌・国民之友・太陽の語種率を見ると、刊行年が新しくなるにつれ漢語率が増加し和語率が減少する傾向が見られることが分かる[4]。一方、女学雑誌・女学世界は、明六雑誌・国民之友・太陽と比較して漢語率が低く和語率が高い傾向が見られる。そして、刊行年が新しくなるにつれ漢語率が減少し和語率が増加する傾向にある。このように 2 グループ間で語種率の通時的変化は全く異なる様相を見せることが分かる。

３．２．品詞率の通時的変化

次に、品詞率の通時的変化を見る。CHJ 近代雑誌の短単位の品詞は大分類で 16 種類に分類されるが、そのなかで記号類（記号・補助記号・空白）以外の 13 種類の品詞についてサンプル単位の比率の分布を雑誌種類ごとに箱ひげ図で示したものが図 2 である。

明六雑誌・国民之友・太陽の品詞率を見ると、名詞・接頭辞・接尾辞が増加、代名詞・動詞・形容詞・副詞・連体詞・接続詞・助詞が減少、形状詞・感動詞・助動詞が増減なしの傾向にある。一方、女学雑誌・女学世界は、名詞・接尾辞が増加、代名詞・動詞・形状詞・副詞・連体詞・接続詞・助動詞が減少、形容詞・感動詞・接頭辞・助詞が増減なしの傾向にある。このように品詞率は 2 グループ間で類似の傾向を見せることが分かる。

４．考察

３．で見た語種率・品詞率の通時的変化が生じた背景について、明六雑誌・国民之友・太陽と女学雑誌・女学世界に分けて考察する。

[4] 田中（2010）では同じ 3 誌の漢語率について 1909 年以降減少するとしているが、これは口語文・文語文あわせての分析であり、そこで見られる漢語率の減少は、文語文よりも漢語率の低い口語文の拡大に応じた現象と考えられる。また、森岡（1991b、pp.381-382）の明治～昭和期の新聞での調査でも和語率は増加するとしているが、これも文語体・口語体あわせての分析である。

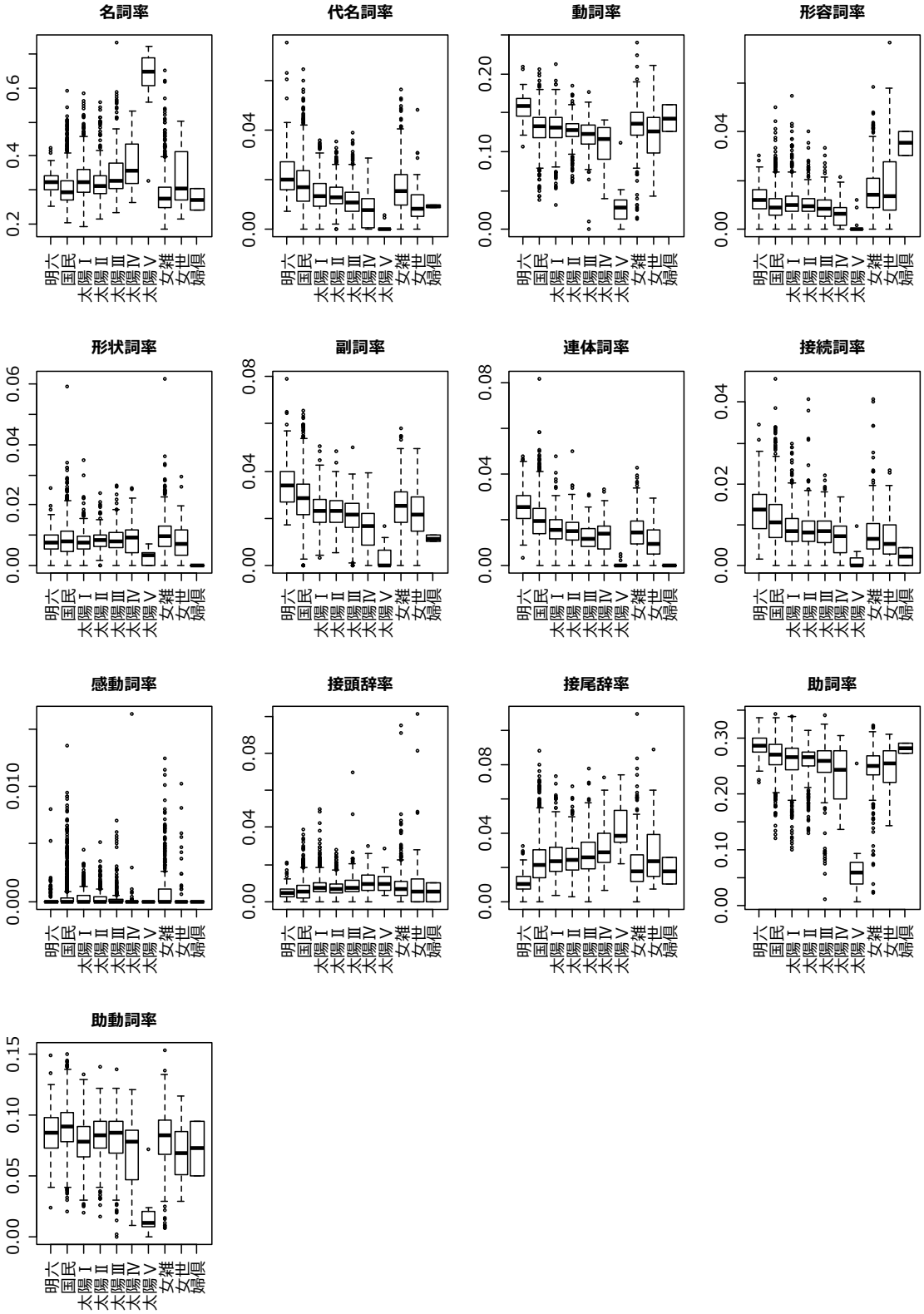


図２　雑誌種類別に見る品詞率の分布

４．１．明六雑誌・国民之友・太陽の考察

まず、明六雑誌・国民之友・太陽において名詞率が増加し、動詞・形容詞・副詞・連体

詞・接続詞の各比率は減少する点について考察する。樺島・寿岳（1965）は、自立語の品詞を名詞（N)、動詞（V)、形容詞・形容動詞・副詞連体詞（M)、感動詞・接続詞（I）の 4 組に分け、大正・昭和期の小説の調査から、N の比率が増加すれば V・M・I の比率は減少することを示し、また、少ない文字数に多くの意味を盛り込まなければならない要約的文章では名詞率が大きい傾向が見られるとした。

明六雑誌・国民之友・太陽の品詞率の通時的変化は、樺島・寿岳（1965）のいう名詞率とそれ以外の品詞率との関係に合致する。また、名詞率の高いサンプルを見ると、(1)〜(3)の例のように事実を客観的に記述した報道的文章で、樺島・寿岳（1965）のいう要約的文章に通じる性質を持つものである。

(1) 支那西洋開化之差別　有賀長雄氏著　京都　大黒屋書舗發兌　日本商業教育論　高橋義雄氏著　東京　金港堂發兌　家計簿記法例題　藤尾録郎氏編　東京　經濟雜誌社發兌　英和書尺牘書法　井上十吉氏著　東京　吉岡商店出版支那西洋開化之差別は歴史上の事實を證據として東西兩洋の文明の相異なる所以を論じたり○日本商業教育論は商人にも學問の必要なる所以を論じたるものにて町人丁稚の爲めには心得となるべき書なり　（60M 国民 1887_09018「新刊雑書」名詞率 0.543）

(2) ○政治法律◎伊藤首相歸京　久しく熱海大磯にて療養中なりし伊藤侯は客臘十日歸京せり◎臨時首相解任　上記に付伊藤侯は客冬十二月十一日徳大寺侍從長の手を經て執奏を請ひ、左の如く西園寺侯の解任手續を爲せり。樞密院議長侯爵　西園寺公望内閣總理大臣臨時代理免らる（60M 太陽 1901_01060「海内彙報」名詞率 0.558）

(3) 解答方法　一、解答は東京市赤坂區青山北町五ノ二〇金易二郎　解答方法　一、解答は東京市赤坂區青山北町五ノ二〇金易二郎氏宛に送ること　一、解答は手順を明記し、變化あるときは合せ記入すること　一、締切七月三十一日限　一、解答は七月號の本誌にて發表す　（60M 太陽 1925_09079「懸賞詰將棋新題」名詞率 0.713）

このような文章は国民之友・太陽の名詞率 0.5 以上のサンプルに集中して出現し[5]、年を経るにつれ文語体サンプル全体に占める割合が増加する傾向が見られる。図 3 は雑誌ごとのサンプル単位の名詞率の分布をヒストグラムで表したものだが、ここから名詞率 0.5 のサンプルが年を経るにつれ増加する傾向が見てとれる。そして太陽 1925 年に至り、1 サンプ

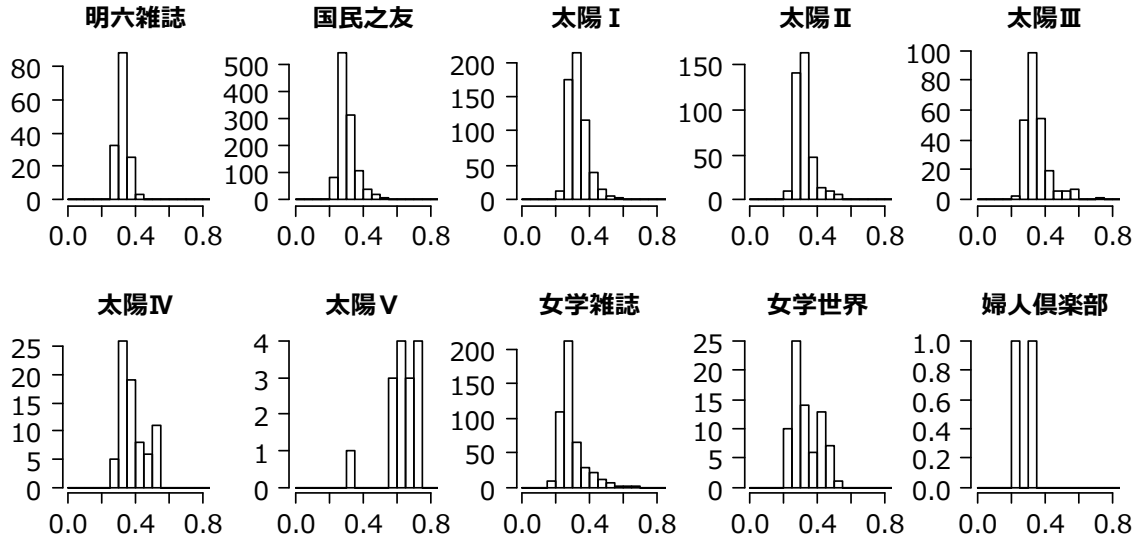


図 3　雑誌種類別に見る名詞率の分布

[5] 明六雑誌は論説文のみ掲載するため、報道的文章のサンプルは出現しない。

ルを除きすべてのサンプルが(3)と同内容の文章となるため、太陽 1925 年の名詞率は他の雑誌種類に対して突出して高くなる。

一方、名詞率の低いサンプルでは、(4)〜(6)の例のような、ある事柄について著者が主張を述べる評論的文章が多く見られる。

(4) 谷氏果然其の職を辭せり、心なき人は何とも云はば云へ、吾人は實に其の出處進退に於て、毅然たる大丈夫の動作に孤負せざるを信ず、谷氏朝を去る、然れども其の意見は朝を去らず、豈にただ朝のみならんや、顧ふに一篇の意見書は恰も噴火山の如く、万丈の光焔を吐き來りて、天下の人心を警醒作興せり、
（60M 国民 1887_07008「谷氏果然其の職を辭せり」名詞率 0.205）

(5) 太陽には社會改良意見を連載し來りしも、久しきことながら、何れも教育を以て、根本より改良するの外なしといふに歸着せざるはあらず。文部省が、新中學令に於て、音樂の一科を加へたるは、中學生をして、優美健全なる音樂の趣味を解せしめ、卑猥なる俗歌俗曲の爲めに、動もすれば導かれて、其嗜好の正當を誤まる如きの弊を一新せんとの希望に外ならず。
（60M 太陽 1901_04016「社會の腐敗救治意見」名詞率 0.226）

(6) 電車ばかり世に簡便廉價なるものはあらじ、貧しきものと、富めるものと、貴きも、賤きも、みな此便利なる文明の器械によらざるは希なり、加之近ごろは運轉手もよくなれて、怪我人も少なく、車掌も段々丁寧になりしは喜ばしき事なり。
（60M 太陽 1909_11051「牛門隨筆」名詞率 0.273）

以上のことから、名詞率の増加の背景の一つとして、文語体の用いられる文章の種類が、評論的文章から報道的文章に変化していったことがあげられると考える。

次に漢語率の増加について考察する。樺島（1963）は大正・昭和期の小説の調査で、漢語文節の比率が増すと名詞文節の比率も増す傾向が見られ、漢語の比率は名詞文節によって強く支配されるとした。延べ語数に対して最も比率の高い品詞は名詞であることから、その名詞の漢語率が全体の漢語率に最も強い影響力があるのは当然と言える。CHJ 近代雑誌の短単位では、活用語の語種は和語・混種語のいずれかであり漢語とはならないため、自立語のなかで名詞の次に比率の高い動詞の漢語率は全体の漢語率に影響力を持たず、名詞率が漢語率に与える影響はいっそう強い。よって、漢語率の増加は名詞率の増加が第一の背景となっていることは疑いない。ただし、品詞ごとにサンプル単位の漢語率の分布を箱ひげ図で表した図 4 から、名詞・形状詞・副詞・接頭辞・接尾辞で漢語率の増加が見てとれることから、全体の漢語率の増加は、名詞率の増加だけを背景とするのではなく、語彙自体の漢語率の増加という通時的変化ももう一つの背景となっていることが分かる。

最後に、接頭辞率・接尾辞率の増加について触れておく。これは、図 4 に見られる接頭辞・接尾辞の漢語率の増加傾向と合わせて、野村（1981）・松井（1987）にいう近代語における字音接辞の発展を反映したものであり、それが具体的数値として確認できたものと考えられる。

４．２．女学雑誌・女学世界の考察

まず、女学雑誌・女学世界の名詞率の増加について考察する。これは、明六雑誌・国民之友・太陽同様、文語文の用いられる文章の種類が評論的文章から報道的文章に変化していったことを背景の一つとしていると考えられる。図 3 からも女学世界のほうが女学雑誌よりも名詞率の高い区間に多くのサンプルが分布していることが分かるが、このような名詞率の高いサンプルは(7)〜(8)の例のような報道的文章で占められる。

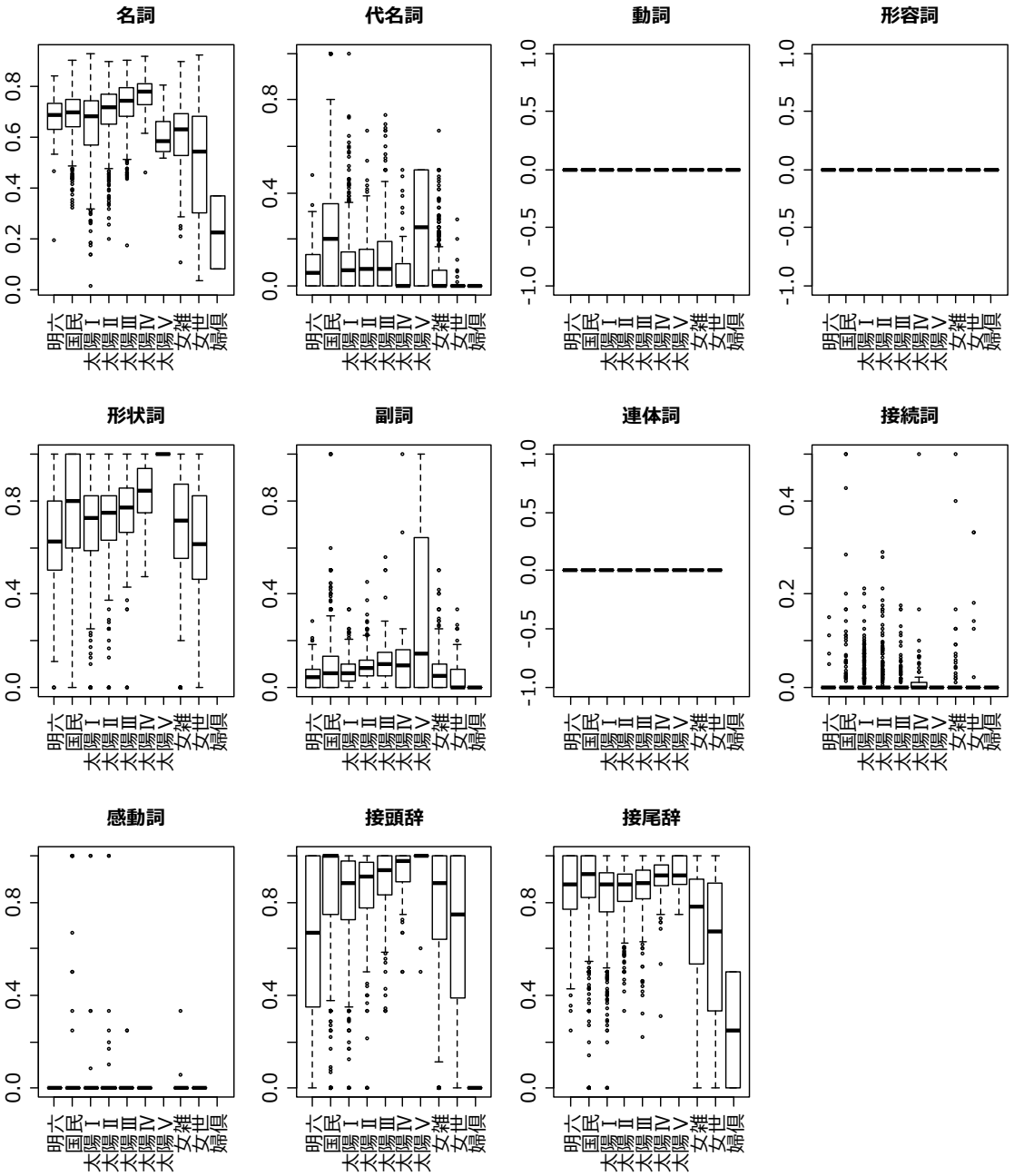


図４　品詞・雑誌種類別に見る漢語率の分布

(7) ○寄付金件　今回朝鮮事件に關し本社事業費の内金品寄付申出たるもの左の如し　一晒木綿三十段　廣島縣正社員　青盛數馬　兵庫縣正社員　三崎安二郎　同　三崎弘造　一金五拾圓　同　小谷廣吉　同　加納辰三　一金壹圓　宮城縣　鈴木志惠　一金壹圓　同　亘理きう　一梅千五樽　三斗入

（60M 女雑 1894_32016「日本赤十字社録事」名詞率 0.652）

(8) 麴町區下二番町に在りし女子音樂園は、今般豐多摩郡下澁谷に移轉新築せしにつき、五月六日午後二時より落成式を擧行せり、園主松山溢子刀自の開會の辭に次で加藤弘之男の祝詞（代讀）、坪井、三宅、兩博士及び青木文造氏の演説あり、其よりのピアノ、ヴアイオリン、合唱、箏、等生徒の演奏、數番あり、

（60M 女世 1909_08024「女子音樂園落成式」名詞率 0.503）

逆に名詞率の低いサンプルには(9)〜(10)のような評論的文章が多く見られる。

(9) 斷えず人を怨んで不平勃々たる者よ。試ろみに問はん、汝、一人にても心を許して身を托せんとするの友ありや。汝は曰はんとす、世に知己なしと。左れど、汝に問はん、抑そも汝が知り、汝が尊とみて、仕へんとするの長者はありやと。汝は曰はんとす、一人もなしと噫、あはれむべし、此種の人。
（60M 女雑 1894_34008「人相ひ評するを聽く」名詞率：0.197）

(10) 人は能く言ふ、我が理想は斯の如し、我れは斯の如くありたしと。なれど其の能く實現されたるもの世間果して幾人かある。舅姑のなき家に嫁がんと思ひしものも、實際は、之れあるのみならず、兄弟姉妹其の他親族の同居さへする家に婿を結び、又富裕なる人にと思ひ居りしものも、家もなく、衣も乏しく、食さへ豊かならざる悲運に陥り居るは、世間なべての例にあらずや。
（60M 女世 1909_13032「修養手引草」名詞率 0.245）

次に漢語率の減少について考察する。これは明六雑誌・国民之友・太陽とは逆の傾向であり、樺島（1963）のいう傾向とも合致しない。図 5 は雑誌種類ごとにサンプル単位の漢語率の分布をヒストグラムで表したものだが、女学世界では、特に漢語率の低い区間にサンプルが多く分布していることがわかる。

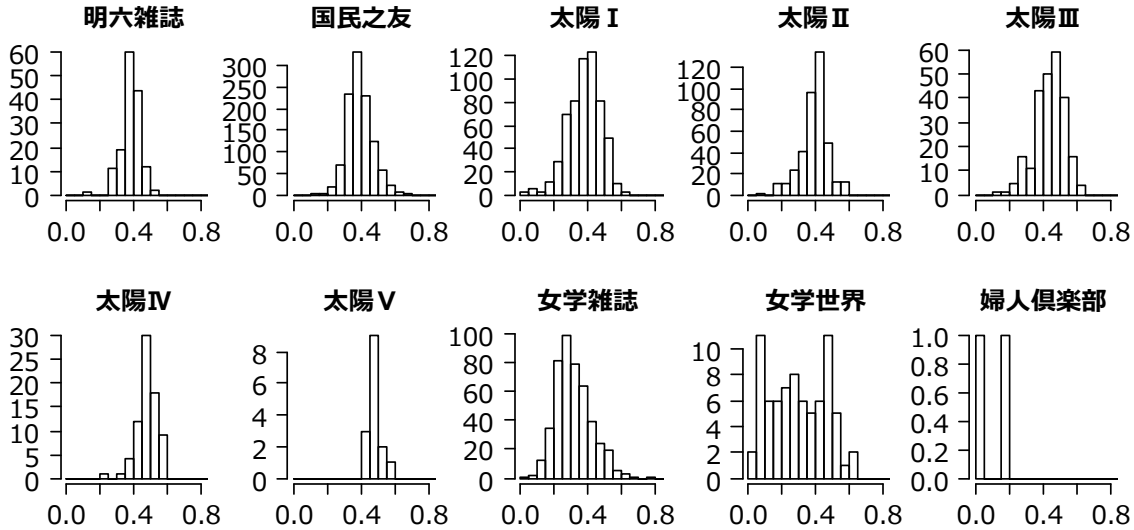


図５　雑誌種類別に見る漢語率の分布

この女学世界の漢語率の低いサンプルのほとんどが(11)のような読者投稿による随筆的文章である。

(11) 淋しさは、秋の夕ぐれならまし、たえかねて庭そぞろあるきするに、そこここより蟲の音いといとあはれに聞ゆ、花だんにたちよりて見るに、うつくしかりし昔はいづこ、花もなき小町草のうなだれて、やり水にうつせみの息つなぎ居るあはれ、げに、古しへの卒塔婆小町の、うつろひし姿、水にうつしてかへらぬ春の永き日をかこちけるにさも似たり。　（60M 女世 1909_16075「秋の夕ぐれ」漢語率：0.029）

このような読者投稿のサンプルは女学雑誌にはない。このサンプル群の存在が女学世界の漢語率の分布に大きな影響を与え、漢語率が低下しているように見えている面があると

考えられる。

最後に、女学雑誌・女学世界の漢語率が明六雑誌・国民之友・太陽よりも低いことについて考察する。特に漢語率の低いサンプルが集中して出現する女学世界のみならず、女学雑誌も図 5 に見られるように明六雑誌・国民之友・太陽よりも漢語率の低い区間にサンプルが分布する。この背景として読み手の性差を考えたい。女性は主に和文を読み書きし漢文訓読系の文章とは疎遠であった歴史があり、それが近代の文語文においても続いていたということである。それは、(9)〜(10)のような評論的文章においても(11)のような随筆的文章においても同様に見られる傾向であったことが分かる。

５．おわりに

以上、本研究では、近代の文語体実用文の通時的変化の実態を明らかにすることを目的として、CHJ 近代雑誌を利用した語種率・品詞率の通時的変化について分析・考察した。そこから明らかになったのは、名詞率の増加という通時的変化から見えてきた、文語体の使用の場が評論的文章から報道的文章に移行していく実態であった。また、漢語率の増加も見られ、その背景として名詞率の増加の影響だけでなく、語彙自体の漢語率の増加も認められた。さらに、読者の性差による文語体のありようの差異や近代語における字音接辞の発展という事象も語種率・品詞率から確認することができた。

今後、語種率・品詞率以外の観点も加え近代文語文の通時的変化を多角的に分析するとともに、口語文との比較を交えて、実態の更なる解明を進めていきたい。

謝辞

本研究は、科研費基盤研究（C）「形態論情報付きコーパスを活用した近代日本語の位相の計量的研究」（16K02750）および国立国語研究所言語変化研究領域共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の成果の一部である。

参考文献

有山輝雄（1986）「言論の商業化―明治 20 年代「国民之友」―」『コミュニケーション紀要』4、pp.1-23

岡本勲（1980）「明治文語の助動詞の位相」『中京大学文学部紀要』15:2、pp.53-98

樺島忠夫（1963）「漢語をめぐって」『計量国語学』27、pp.14-19

樺島忠夫・寿岳章子（1965）「2. 要約的表現と描写的表現」『文体の科学』pp.16-49

進藤咲子（1981）「第四章　新聞の文体」『明治時代語の研究―語彙と文章―』明治書院、pp.245-270

田中牧郎（2006）「『近代女性雑誌コーパス』の概要」『日本学術振興会科学研究費補助金研究成果報告書 基盤研究(B)「20 世紀初期総合雑誌コーパス」の構築による確立期現代語の高精度な記述』pp.55-62（http://pj.ninjal.ac.jp/corpus_center/cmj/doc/19w-mag-summary.pdf よりダウンロード可能）

田中牧郎（2010）「雑誌コーパスでとらえる明治・大正期の漢語の変動」『国際学術研究集会　漢字漢語研究の新次元　予稿集』pp.56-63

永嶺重敏（1997）「第三章　明治期『太陽』の受容構造」『雑誌の読者の近代』日本エディタースクール出版部、pp.101-132（本研究ではオンデマンド版[2004]に拠った）

野村雅昭（1981）「近代日本語と字音接辞の造語力」『文学』49:10、pp.22-34

松井利彦（1987）「漢語の近世と近代」『日本語学』6:2、pp.25-36

松崎安子（2006a）「明治期の文語文の類型―小学校理科教科書を対象として―」『文化』70:1-2、pp.92-105

松崎安子（2006b）「明治期の新聞における文語文記事の文体類型―小学校理科教科書の文体との比較から―」『文芸研究―文芸・言語・思想―』162、pp.11-22

森岡健二（1991a）『近代語の成立―文体編―』明治書院
森岡健二（1991b）『改訂近代語の成立―語彙編―』明治書院

結合の強度を測る指標としての Log-r の有用性：日・英語のバイグラムデータに基づく MI，LLR などとの比較

藤村 逸子（名古屋大学）†
青木 繁伸（群馬大学名誉教授）

Appropriateness of Log-r for calculating strength of association: Comparison with MI, LLR using Japanese and English bigram data

Itsuko Fujimura (Nagoya University)
Shigenobu Aoki (Gunma University)

要旨

２語からなるコロケーションは一般に共起頻度と２語の結合力によって特徴づけられる。本研究は，結合力の指標として Fujimura & Aoki (2016)において提案した Log-r を，同じ目的の指標として言及されることの多い MI，LLR，t-score，Dice，Jaccard と比較し，簡素な指標である Log-r の有用性を主張する。データは『現代日本語書き言葉均衡コーパス』と英語の大規模新聞コーパスから網羅的に採取した多量のバイグラムを用いる。横軸にバイグラムの共起頻度をとり，縦軸に各指標値をとった散布図を作成して各指標の特徴を視覚的に描き，散布図間の比較によって指標間の差異を明示する。

１．はじめに

大規模コーパスに基づく言語研究のひとつとしてコロケーションの研究が盛んに行われている。コロケーションは語と語の慣用的な結合と定義されるが，それには種々 のタイプのものが含まれる。それぞれのタイプを特徴づける基本的な特性として言及されることが多いのは，連語の粗頻度と，連語を構成する単語間の結合の強度の 2 つであ る（Ellis 2012; Gries 2012; Wray 2012）。粗頻度はわかりやすい特性であるが，結合の強度は名称もさまざまであり統一的に扱われてはいない。また，その指標（およびその計算式）としては MI（Mutual Information）（Church & Hanks 1990）に言及されることが多い(Ellis 2012; Evert 2009; Gries 2012; Hunston 2002)が，一方で種々の指標（および計算式）が提案され（Pecina 2010; 相澤・内山 2011），研究はいまだに途上にあると言える(Bybee 2010; Evert 2009; Gries 2013)。

日本語には「半信-半疑」，「徹頭-徹尾」，「金科-玉条」，「有象-無象」，「換骨-奪胎」，「夫唱-婦随」，「官尊-民卑」のようなコロケーションが存在する。これらがどれも 1 語性の強い連語であることは直感的に感じられる。また，本研究のコーパスの『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）においては，これらの連語の構成形態素の一方は必ず他方と共起し，その他の形態素とは共起しない（「半信」は「半疑」とのみ共起する。「半疑」も「半信」とのみ共起する。他も同様。）。しかし，結合の強度を測るはずの上記の指標によってこれらの連語を計測すると，直感や事実に反して，その値はこれらの連語間で同一とは限らない。

† fujimura@nagoya-u.jp

言うまでもなく，現象を計測するための指標の特徴が曖昧であることは望ましくないが，現段階において，それぞれの指標の特徴に関する明示的な説明はなされていないのが状況である。

我々は Fujimura & Aoki (2016) において，2 語連語（以下バイグラム）[1]の結合の強度をはかる簡素な指標として Log-r を提案し，英語とフランス語の大規模データをもとに MI と対照させて，言語現象としてのコロケーションを理解する上でのその有用性を主張した。本発表は主として日本語を扱い，Log-r を用いることによって日本語のコロケーションの記述に貢献できることを示す。また MI の他に，LLR(Log-Likelihood Ratio)，t-score, Dice, Jaccard とも比較してそれぞれの指標の特徴を明らかにし，バイグラムの結合の強度を測る指標としては Log-r が有用であることを明らかにする。

２．Log-r，対数，頻度と強度に基づく特徴づけ

本章では，Log-r を紹介する。また，Wray(2012)による連語の頻度と構成語の結合の強度に基づくコロケーションの特徴づけのモデルを示し，語彙の分布に関する研究における対数の価値を説明する。

２．１．　Log−r

2 語の結合の強さを示す指標として，2 変数（単語 x と単語 y）の属性相関を表すピアソンの積率相関係数(r)の常用対数を提案し，それを Log-r と名づける（Fujimura & Aoki 2016）。ピアソンの積率相関係数の定義式は(1)である。本研究では，ポワソン分布を仮定して，その近似式を用いる。Log-r はしたがって，(2)のように定義される。

$$r = \frac{cov_{xy}}{\sigma_x \sigma_y} \tag{1}$$

$$\text{Log-r} = log_{10} \frac{f_{xy}}{\sqrt{f_x f_y}} \tag{2}$$

(f_{xy}：バイグラム xy の頻度，f_x：x の頻度，f_y：y の頻度)

Log-r は，2 語の結合度を測る指標としてすでに提案されている z スコア，カイ二乗値 (χ^2)，phi 係数，コサインと共通した性質をもっている（cf. Pecina 2010; 相澤・内山 2011）。すなわち Log-r は全く新規の指標というわけではない。

２．２．　連語の頻度と強度による特徴づけのモデル

図１は，連語の特徴をその頻度と構成要素間の強度に基づいて特徴づける Wray (2012)によるモデルである。横軸は頻度を表し，縦軸は強度を表している。縦軸はバイグラムの構成要素の結合度の強さ，すなわちバイグラムの１語性の度合いを表す。これはバイグラムの頻度とは異なる概念である。頻度の多さと結合度の強さは独立しているはずである。第１象限には頻度大かつ強度大，第２象限には頻度小かつ強度大，第３象限には頻度小かつ強度小，第４象限には頻度大かつ強度小の連語がプロットされる。それぞれの象限の典型的なバイ

1　ここで連語とは，その共起の慣用性に関わらず単に語の連続を指す。2 語連語(バイグラム)は 2 語の連続を指す。

グラムの例としては，第 1 象限には New York，White House などの高頻度の固有名詞，第 2 象限には低頻度で結合度の強固な bovine spongiform, lingua franca などのイディオム，第 4 象限には高頻度で強度は弱い of the, I am などのレキシカルバンドル，第 3 象限は pink roses や familiar enough などのその他の平凡な 2 語の連続をあげることができる。言うまでもなく，図 1 の縦軸と横軸は連続体をなしている。本研究では，このモデルにならい，頻度を横軸にとり，強度の指標を縦軸にとって，どの指標が現実の言語現象によりよく適合するかを検討する。

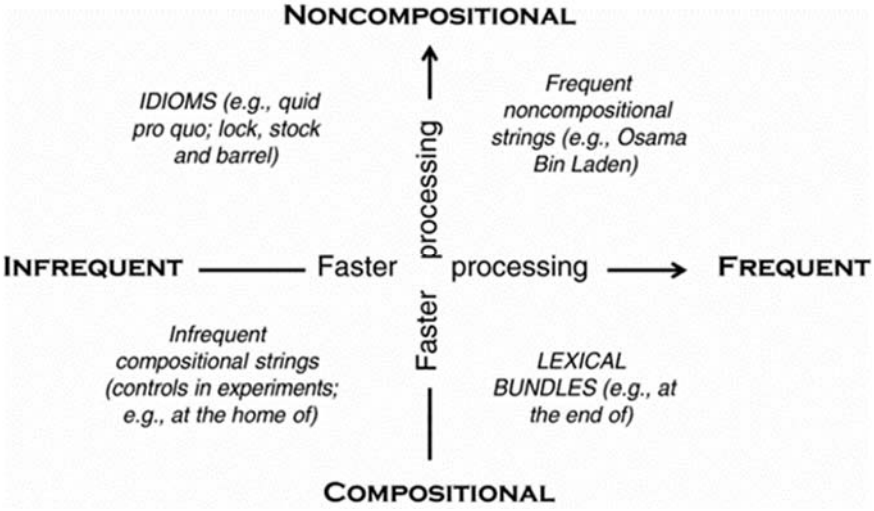


図 1　Wray(2012)によるコロケーションの特徴づけモデル

２．３．　対数

Log-r が r の常用対数をとっているのは，語の頻度分布が Zipf の法則(Zipf 1949)に従う極端に範囲の広い統計量だからである（概算では，単語のコーパス内での出現率（%）＝10/順位)。対数をとって比較すると数値の処理が容易になり，現象をグラフ化して視覚的に把握できるようになる(Baroni 2009; 青木 2009)。この点は，以下で検討する t-score，LLR，Dice，Jaccard においても同様である。MI は定義式においてすでに対数がとられているので，Log-r との比較が定義式のままで可能であるが，以上の 4 指標は通常の定義式のままでは Log-r との視覚的な比較はできない。したがって，LLR, Dice, Jaccard, t-score の検討はその常用対数値によって行う。また，横軸となる頻度それ自体もその常用対数値を基に考察する。

３．データ

データの詳細は表 1 のとおりである。

バイグラムデータの作成は日英両言語とも Unix 環境において, Unix コマンドと Perl スクリプトを使ったプログラミングによって行った。英語のデータの収集は，表 1 に挙げた新聞コーパスをもとに，形態素情報の付与は行わず，単語を単位として行った。こうしてすべての単語の頻度表とすべてのバイグラムの頻度表を取得した。日本語は，BCCWJ の短単位の TSV 形式データを用いた。バイグラムは形態素情報付きの出現形を単位として作成し，短単位形態素の頻度表とバイグラム頻度表を取得した。バイグラム頻度の低いものを削除した上で, 英語の 104 万件のバイグラムと日本語の 62 万件のバイグラムのそれぞれについて，

Microsoft Excel を用いて各指標の値を計算し，データベース化した。散布図の作成およびその他の統計処理は統計ソフトの Jmp ver.13 を用いて行った。

表１　使用コーパスとバイグラムデータ

	バイグラムの個数とコーパスの総語数・総形態素数	単位	テキストの種類	コーパスの名称と配布元
英語	・1，040，000 個（生起数 54 回以上） ・10 億語	・単語 ・形態素解析なし ・大文字・小文字は区別して扱われている	新聞	LDC ・North American News Text Corpus ・North American News Text Supplement
日本語	・615，000 個（生起数 10 回以上） ・1 億形態素	・短単位（レンマ化なし） ・UniDic による形態素解析情報付き	書籍全般，雑誌全般，新聞，白書，ブログ，ネット掲示板，教科書，法律	国立国語研究所 ・BCCWJ（DVD 版）

４．英語と日本語のバイグラムの Log-r と MI

４．１．Log-r と MI

表 2 は，英語と日本語データから取り出したバイグラムである。最上位の Log-r が 0 の場合は，バイグラムの構成要素が互いに排他的に共起する場合である。「半信半疑」においては，「半信」は常に「半疑」と結びつき，「半疑」も常に「半信」と結びつく。最下位の Log-r が-4 付近のものは，構成要素間の共起は何らかの偶然である場合である。この区間内は，結合度の強度に関する連続性が存在する。上位にあるほど２語の結合度は強く，Log-r も MI も同じ傾向を示しているように見える。MI の計算式は次の通りである。

$$\mathrm{MI} = \log_2 \frac{f_{xy}\, N}{f_x\, f_y} \qquad (3)$$

（Nはコーパスの総語数・総形態素数）

表 2 英語と日本語のバイグラム例の Log-r 値と MI 値[2]

結合の強度	英語			日本語		
	バイグラム例	Log-r	MI	バイグラム例	Log-r	MI
強	lingua franca	-0.01	22.5	半信 半疑	-0.00	19.2
⇕	apple pie	-1.01	13.8	有害 物質	-1.02	11.5
	medal winner	-2.00	8.0	環境 対策	-2.00	5.4
	earlier offer	-3.00	2.3	文化 意識	-3.02	2.4
弱	no there	-4.04	-3.4	利用 その	-4.03	-3.7

４．２．英語データ

しかし，図 2 と図 3 の散布図を見ると Log-r と MI には明確な違いがあることがわかる。

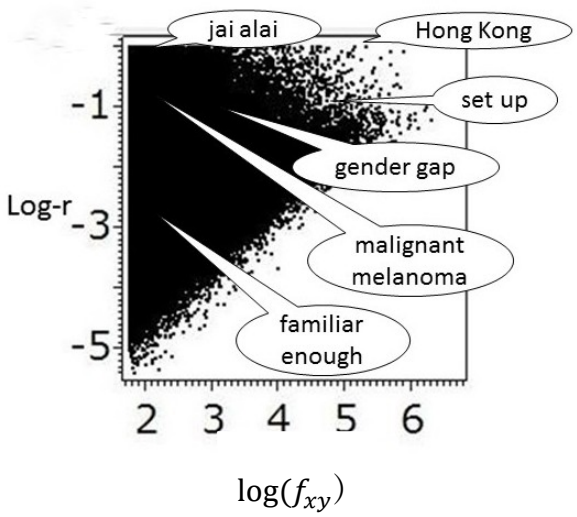


図 2　Log-r/log(f_{xy})の散布図(英語）[3]

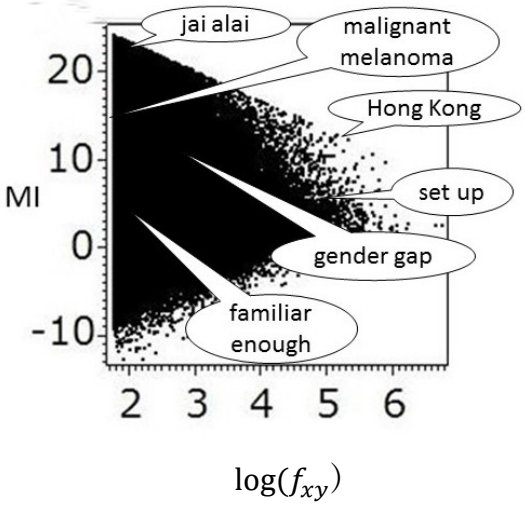


図 3　MI/log(f_{xy})の散布図(英語）

両図には約 100 万件のバイグラムがプロットされている。両図とも横軸はバイグラムの頻度（の常用対数）である。縦軸は図 2 ではバイグラムの Log-r であり，図 3 ではバイグラムの MI である。上掲の図 1 のモデルに従うと，縦軸には頻度とは独立したものとして強度がプロットされるべきである。図 2 ではそれが実現しているが，図 3 の上辺を見るとバイグラムの頻度が増加するにつれて MI 値は例外なく低下しており，２項目は独立しているとは言えない。たとえば散布図上に例として位置が示してある Hong Kong と jai alai (バスク地方のスポーツ)は，２語の結合の強度の観点においては，共通した特徴を持っている。それ

[2] MI 値は，コーパスの総語数・総形態素数 N にも左右される。N は英語データでは 10 億，日本語データでは 1 億であり，底を 2 とする N の対数値は，それぞれ 29.9 と 26.6 となる。本データの英語のバイグラムの MI 値は日本語のそれと比べてデフォルトで 3.3 大きい。この表において， MI 値を基準として日本語と英語のバイグラム（たとえば「半信半疑」と「lingua franca」）を比較するのは不可能と言ってよい。

[3] 図 2 と図 3 は，Fujimura & Aoki 2016 Fig2， Fig3 より転載。

ぞれの構成単語は他の要素とはほとんど共起せず，相互の結合は強固である[4]。すなわち，Hong の生起のうちの 97%が Kong と，Kong の生起のうちの 97%が Hong と共起している。同様に jai のうちの 91%が jai と，alai のうちの 98%が jai と共起している。図 2 においては Hong Kong と jai alai はグラフの最上部に等しくプロットされているのは理にかなっている。２つのバイグラム間で異なるのは頻度である。使用したコーパスにおいて Hong Kong は jai alai に比べて極めて高い頻度で出現している。図 2 において, Hong Kong の MI 値は malignant melanoma よりも低く gender gap のそれに近い。Hong Kong の 1 語性が強いことはデータから明らかであるので，この結果は言語使用の現実に反しているといえる。図 3 の上辺が右下がりの直線になっているのは現実の言語現象の反映ではなく, MI の計算式に由来している。強度の計測において頻度が影響を及ぼすのは，現象の正確な記述の目的には反すると考えられる[5]。なお，図 2 と図 3 において下辺が右上方向に切れ上がっているのは言語現象に対応している。バイグラムの頻度が上がるにつれて，構成語間の共起の偶然性は減じるからである。高頻度のレキシカルバンドルの結合度は必然的に高くなる。この点において，図 1 の Wray(2012)の正方形のモデルは言語現象の現実に対応してはいない。

４．３．日本語データ

次に，日本語データから作成したバイグラムを用いて，Log-r と $\log(f_{xy})$，MI と $\log(f_{xy})$の散布図を以下に示す。

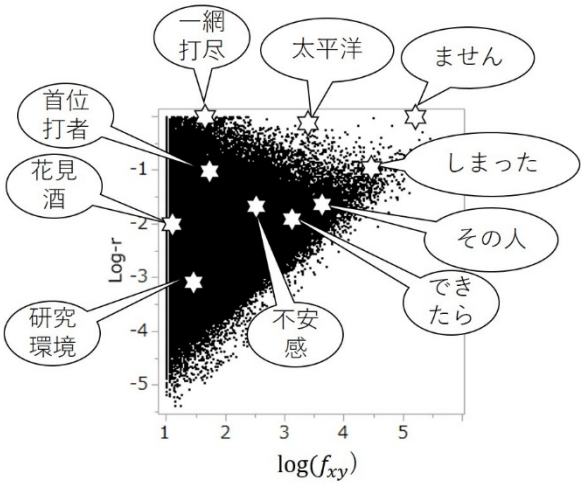


図 4　Log-r/$\log(f_{xy})$の散布図(日本語)

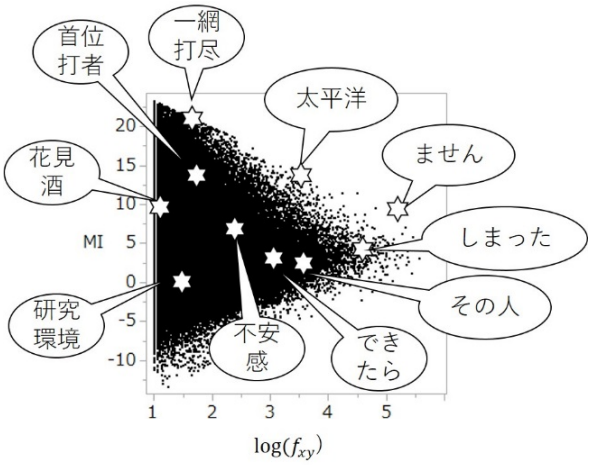


図 5　MI/$\log(f_{xy})$の散布図(日本語)

最初に言えるのは，Log-r を縦軸とする図 2 と図 4, MI を縦軸とする図 3 と図 5 の間で，散布図の形状が言語を超えて近似しているという点である。大規模なコーパス全体を網羅的に対象にする限り，言語を超えて Log-r と $\log(f_{xy})$による散布図の形状は同じであり，MI と $\log(f_{xy})$による散布図もまた同じである。この現象は，語彙の分布が言語を超えて Zipf の

[4] バイグラムの頻度とその構成要素の頻度は以下のとおりである。f(Hong Kong)：143，104，f(Hong):147，404，　f(Kong):147，852，　f(jai alai)：151，　f(jai):166，　f(alai):154.

[5] MI は，頻度は低いが，結合が強固なバイグラムの発見のための実用的ツールとしては役に立つと思われる。MI と Log-r の散布図を比べると両者は同一のグラフの変形であることがわかる。

法則に従うという問題に通じる[6]。換言すると，各指標の特徴を評価する際のサンプルとしては，コーパス全体を対象にする必要がある。

次に図 4 と図 5 とを比較すると，上述の図 2 と図 3 との比較と同様のことが観察できる。BCCWJ において「一網打尽」は「一網_名詞-普通名詞-一般」と「打尽_名詞-普通名詞-一般」のバイグラムと分析され，「ません」は「ませ-助動詞」と「ん-助動詞」のバイグラムと分析されているが，下に掲げる表 3 に記したように，「一網」はその 95%が「打尽」に後続され，「打尽」はその 98%が「一網」に前置されている。また，「ませ」はその 98%が「ん」に後続され，「ん」はその 90%が「ませ」に前置されている。「ません」は「一網打尽」に比べてわずかに結合の強度が弱いが，いずれにせよどちらも 1 語性の極めて強いバイグラムである。「一網打尽」と「ません」の極めて大きな差異はその出現頻度である。Log-r による散布図（図 4）はこの現実をよく表していると言える。

英語に関して述べたように，MI には頻度の低いバイグラムを高く評価し，頻度の高いバイグラムを低く評価するという数式上の特徴がある。「ません」の MI 値は「一網打尽」の MI 値と比べて大変低く，散布図上では「花見酒」や「首位打者」よりも下にある。「花見」が「酒」に後続されるのはその 2%に過ぎず，「酒」が「花見」に前置されるのはその 0.5%に過ぎない(表 3 参照)。共起の強度の測定である限り，90%以上の共起率の「ません」が 2 %以下の共起率の「花見酒」よりも下位にプロットされるのはあり得ないので，MI が測っているのは共起の強度ではないということになる。

直感的に，「一網打尽」や「花見酒」や「首位打者」は内容語的なバイグラムであるのに対して，「ません」や「しまった」は機能語的なバイグラムであると感じられる。しかし，機能語的か内容語的かの差異は結合の強度とは別の問題である。

本節では英語と日本語を対象に，Log-r と MI の比較を行った。ここで言えるのは，Log-r は結合の強度を頻度とは独立に測っているのに対して，MI は結合の強度と頻度を融合させた指標であり，図 1 における縦軸を構成するには適してはいないということである。

５．日本語バイグラムに基づく Log-r と t-score，LLR，Dice，Jaccard との比較

５．１．各指標とバイグラム例の値

本節では，日本語のバイグラムデータを用いて，Log-r とバイグラムの結合度を計測する指標として言及されることの多い他の指標とを比較する。取り上げるのは，t-score，LLR（Log-Likelihood Ratio），Dice，Jaccard の４つの指標である。それぞれの指標の計算は以下の式による(Petina 2010)。　なお，N はコーパスの総語数・総形態素数である。

- t-score

$$\left(f_{(xy)} - \frac{f_{(x)} \times f_{(y)}}{N}\right) \div \sqrt{f_{(xy)}} \tag{4}$$

- LLR (Log-Likelihood Ratio)

$$-2 \sum_{ij} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}} \tag{5}$$

[6] Fujimura & Aoki (2016)では，同様の比較を英語とフランス語間で行った。二次元的散布図の形状の差異は，英語と日本語間より，英語とフランス語の間の方が小さい。英仏語と日本語のデータを比べると，次の３つの顕著な差異がある。１）言語の特徴：英仏語は日英語より近い，２）処理の単位：英仏語では単語であり，日本語では短単位形態素である，３）テキストジャンル：英仏語は新聞のみであるが，日本語は種々のものを含む。

- Dice

$$\frac{2 \times f_{(xy)}}{f_{(x)}+f_{(y)}} \tag{6}$$

- Jaccard

$$\frac{f_{(xy)}}{f_{(xy)}+f_{(x)}+f_{(y)}} \tag{7}$$

上で述べたとおり，Log-r との視覚的な比較の目的のために，これらの 4 指標は常用対数に変換して用いる。

表 3 は, すでに図 4 と図 5 において例に挙げた 10 個のバイグラムのデータを掲げている。すなわち，表 3 には，それぞれのバイグラム頻度，構成形態素の頻度，対数化されたバイグラム頻度，Log-r 値，MI 値，対数化された t-score 値，対数化された LLR 値，対数化された Dice 値，対数化された Jaccard 値が示されている。

表 3　バイグラム例，それぞれの頻度と各構成形態素の頻度，および各指標値

	$f_{(xy)}$	$f_{(x)}$	$f_{(y)}$	$\log(f_{xy})$	Log-r	MI	t-score	LLR	Dice	Jaccard
一網-打尽	42	44	43	1.6	-0.02	21.1	0.81	3.11	-0.02	-0.49
ませ-ん	142,609	144,908	158,770	5.2	-0.03	9.3	2.58	6.31	-0.03	-0.50
太平-洋	2,635	2,961	3,934	3.4	-0.11	14.5	1.71	4.73	-0.12	-0.56
首位-打者	50	501	612	1.7	-1.04	14.0	0.85	2.94	-1.05	-1.37
不安-感	321	10361	18540	2.5	-1.64	7.4	1.25	3.43	-1.65	-1.96
その-人	5,734	390,373	150,791	3.8	-1.63	3.3	1.83	4.21	-1.67	-1.98
しまっ-た	28,516	37,113	2,692,208	4.5	-1.04	4.8	2.21	5.22	-1.68	-1.99
でき-たら	1,180	97,682	110,418	3.1	-1.94	3.5	1.49	3.55	-1.95	-2.25
花見-酒	12	626	2474	1.1	-2.02	9.6	0.54	2.13	-2.11	-2.41
研究-環境	34	39847	30,652	1.5	-3.01	1.5	0.57	1.42	-3.02	-3.32

以下では，4 節と同じく，60 万件のバイグラムによる散布図（各指標／$\log(f_{xy})$）と，その散布図上にプロットされた表 3 のバイグラムをもとに，各指標の特徴を明らかにする。

５．２．　t-score

t-score は MI とともに古くからコロケーションのための指標として言及されている。しかし，図 6 からわかるように，t-score はバイグラムの頻度にほぼ相関した値をとる。図 1 の Wray(2012)のモデルを想定して, t-score を結合の強度を測る指標として用いることは全く不適切であると言える。頻度によって値が変わるので，第 1 節に挙げた熟語の場合，BCCWJ において「半信-半疑」は「夫唱-婦随」より高頻度であるため t-score の値も高いが、別のコーパスにおいてもしも「夫唱-婦随」が「半信半疑」より頻度が高い場合には，両者の t-score の順序は逆転することになる。

５．３． LLR

G スコアとも呼ばれる LLR(Log-Likelihood Ratio)は，Dunning(1993) によりコロケーションの指標として導入された。ライプツィヒ大学の Wortschatz[7]にコロケーションの指標として実装されているなど，実際によく使用されている。図 7 からわかるようにこの指標も頻度との相関性が高い。従って，Wray(2012)のモデルの共起の強度の軸とするには不適切である。

t-score と LLR の値は log(f_{xy})が高いほど増加する。この点，MI とは正反対の指標である。いずれにせよ，これらの 3 指標は頻度から独立しては計測されないので，共起の強度を特徴付ける目的には適合しないと考えられる。

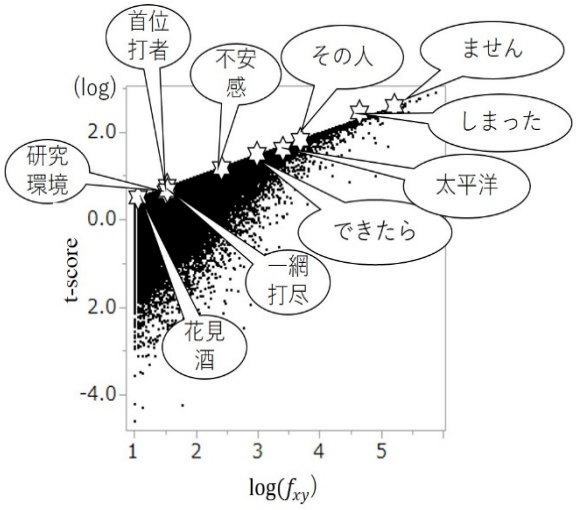


図 6　t-score/log(f_{xy})の散布図[8]

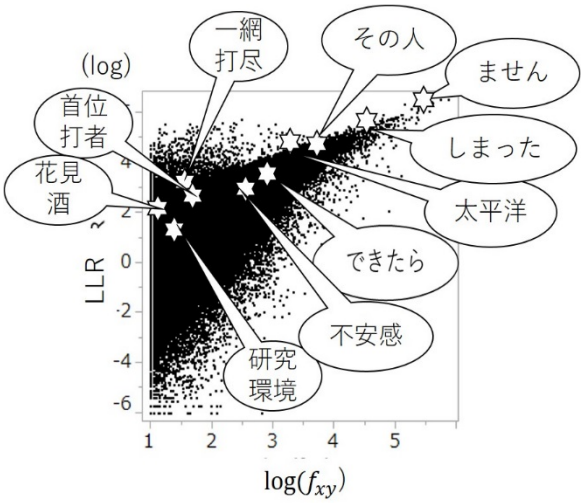


図 7　LLR/log(f_{xy})の散布図

５．４． Dice と Jaccard

最後に Dice と Jaccard を検討する。図 8 と図 9 からわかるように，Dice と Jaccard は大変よく似ている。例に挙げたバイグラムの中では「不安感」の位置が若干異なる以外は，順位に変わりはない。また，図 4 との比較からわかるように，この２つの指標は Log-r ともよく似ている。

平面的な散布図では，Dice，Jaccard, Log-r は見分けがつかないほどよく似ている。Dice と Jaccard よりも Log-r が結語の強度を測る指標として有用であることを主張するためには，これらの 3 指標の差異を別の角度から検討する必要がある。

7 http://corpora2.informatik.uni-leipzig.de/?dict=fra_mixed_2012

8 t-score の散布図において，高密度部分は上辺近辺に偏っている。LLR ではその偏りが緩和されている。

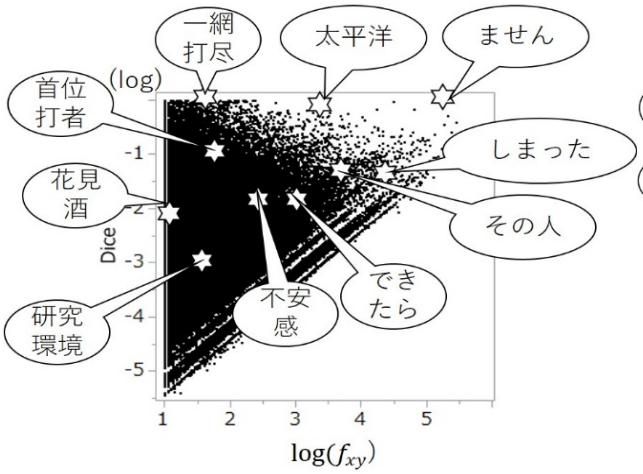


図 8　Dice /log(f_{xy})の散布図

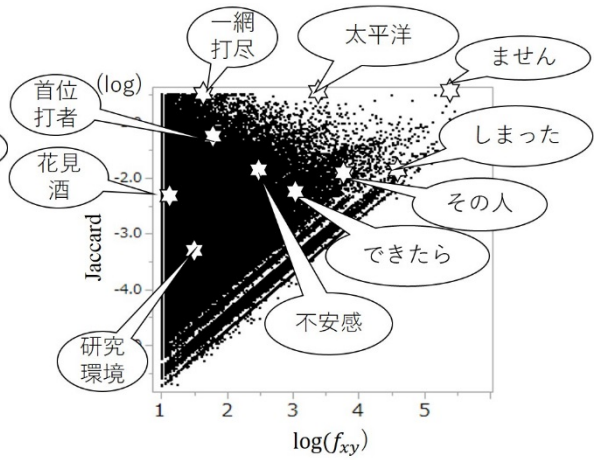


図 9　Jaccard /log(f_{xy})の散布図

６．三次元の散布図による Log-r，Dice，Jaccard の比較

Dice，Jaccard，Log-r の違いを明らかにするために，図 10，図 11，図 12 では，それぞれ図 8，図 9，図 4 と同一の散布図上に，JMP のノンパラメトリック密度推定を用いて，バイグラムの度数の密度に応じた 5%刻みの等高線を引いた。また，log(f_{xy})と各指標の頻度分布を散布図の上部と右にヒストグラムで示した。等高線は色別されており，赤色が最も密度が高く，寒色になるにつれて密度が低い。最後の等高線の外側は全体の 5%にあたるバイグラムが薄く分布する。図 10 の Dice と図 11 の Jaccard はほぼ同一の分布であるが，図 12 の Log-r は全く異なる。図 10 と図 11 では，最も密度が高いのは左端の最下部である。この位置はバイグラムの頻度と強度が最も弱いと想定される場合に当たり，典型的な出現は意味のない（多くの場合は何らかの誤りに基づく）偶然の共起である。このような共起が最も多いということは常識的に考えにくく，Dice や Jaccard は言語使用の現実を反映する指標ではないと考えられる。一方，Log-r による散布図（図 12）において密度が最も高いのは，左端最下部から少し上の位置である。この位置は図 1 の第 3 象限に当たり，Wray(2012)の言う infrequent compositional strings のための定位置である。このような特徴のバイグラムがコーパスにおいて多数存在することは Zipf の法則によって想定できる。図 12 を見ると，密度の分布はなだらかな連続を構成していて，これも Zipf の法則に適合している。

Fujimura & Aoki (2016)では，log(f_{xy})と Log-r による散布図を描くと，共起の頻度と強度に加えて，バイグラム構成単語の親密度（familiarity）[9]も測ることが可能となり，３つの観点からバイグラムの分類ができると主張した。Dice や Jaccard に基づくバイグラムの分布は Log-r による分布に比べて均整がとれておらず，語の親密度を加えたモデルを想定することは困難である。語の親密度を計算するとバイグラムの特徴がより精密に記述できるので，この観点からも Dice や Jaccard に比して Log-r の有用性は高いと言える。

[9] Log-r と log(f_{xy})の散布図を，左上を頂点とする二等辺三角形と見なした場合，最も語の親密度が低いのは左上の頂点である。最も親密度が高いのは下辺であり，機能語はこの場所に位置する。

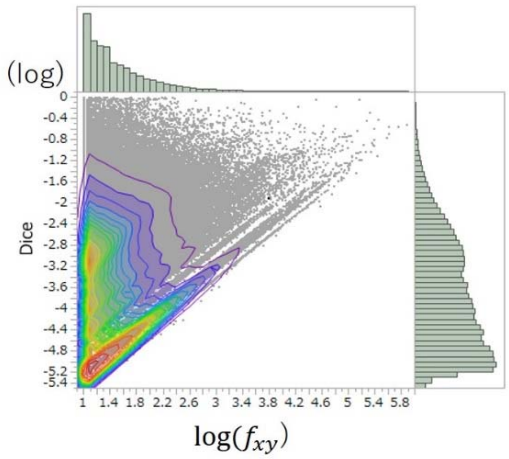


図 10　Dice /log(f_{xy})の三次元散布図

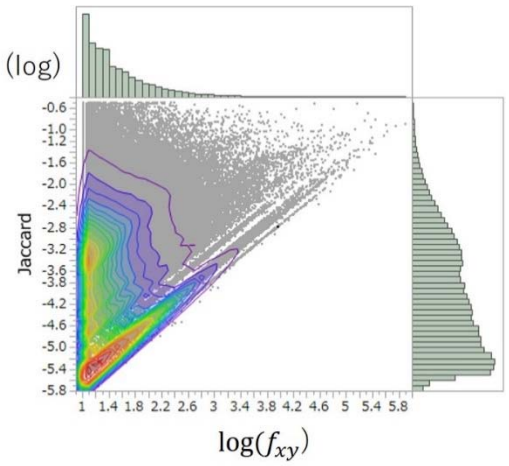


図 11　Jaccard /log(f_{xy})の三次元散布図

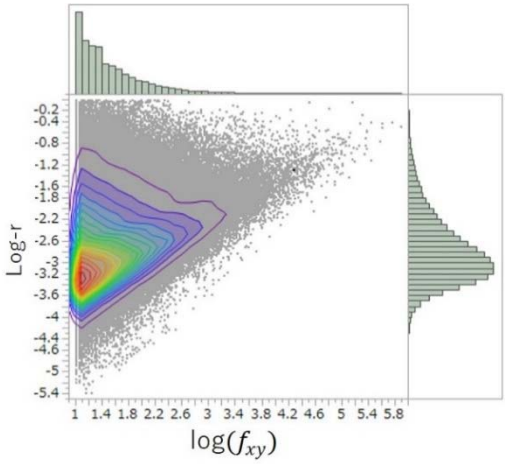


図 12　Log-r /log(f_{xy})の三次元散布図

7．おわりに

本稿では，日本語と英語の多量のデータを用いて，共起の強度を測る指標としての有用性の観点から筆者らが提案する Log-r と，コロケーションの指標として言及されることの多い MI, t-score, LLR, Dice, Jaccard とを比較した。その結果次のことがわかった。MI, t-score, LLR は頻度との相関が強く，強度の指標としては使えない。Dice と Jaccard は，一見 Log-r と近似してはいるが，現実の分布を表してはいない。Log-r は共起の強度のみを測る簡素な指標として最適である。Log-r を強度の指標として用い，頻度や密度などのその他の特徴を組み合わせることによって，バイグラムを多角的かつ正確に特徴付けることが可能になる。Log-r 以外の指標を用いる際には，その目的を明確に定める必要がある。

指標の特徴を評価するためには，本研究で行ったようにコーパス全体をサンプルとすることが必要である。バイグラムの特徴は一様ではないので，恣意的に選択したサンプルによる比較は避けるべきである。

Log-r を強度の指標として認定すると，頻度(log(f_{xy}))，強度(Log-r)，密度に加え，バイグラムの構成要素の親密度（familiarity）も散布図上にプロットでき，バイグラムの詳細な特徴付けに貢献できる (cf. Fujimura & Aoki (2016))。この作業は各形態素のレンマ形をもとに行う必要があるが，BCCWJ にはこのデータが備わっている。次の課題としたい。

謝　辞

本研究は科研費基盤(C)「大規模コーパスに基づく名詞と形容詞の使用パターンと構造化に関する日仏語対照研究」の助成による。英語バイグラムは，滝沢直宏教授（データ取得時は名古屋大学，現在は立命館大学）の提供によります。記して感謝いたします。

文　献

Baroni, M., (2009) Distributions in text, Lüdeling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, Berlin, pp. 803-822.

Bybee, J., (2010) *Language, usage, and cognition*. Cambridge University Press.

Church, K., & Hanks, P., (1990)　Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1), pp. 22-29.

Dunning, T., E., (1993)　Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19(1), pp.61-74.

Ellis, N.C., (2012) Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear.　*Annual Review of Applied Linguistics*, 32, pp.17-44.

Evert, S., (2009) Corpora and collocations, Lüdeling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, pp.1212-1248.

Fujimura, I., & Aoki, S., (2016) A New Score to Characterise Collocations: Log-r in Comparison to Mutual Information, in *Europhras2015 Computerised and Corpus-Based Approaches to Phraseology: Monolingual and Multilingual Perspectives* pp. 271-282.

Gries, S. Th., (2012) Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 36(3), pp.477-510.

Gries, S. Th., (2013) 50-something years of work on collocations What is or should be next..., *International Journal of Corpus Linguistics*, 18:1, pp.137-165.

Hunston, S., (2002) *Corpora in Applied Linguistics*, Cambridge University Press.

Pecina, P., (2010) Lexical association measures and collocation extraction, *Lang Resources & Evaluation*, 44, pp.137-158.

Wray, A., (2012) What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play, *Annual Review of Applied Linguistics*, 32, pp.231-254.

Zipf, G. K., (1949) *Human Behavior and the Principle of Least Effort*, Addison-Wesley.

相澤彰子・内山清子（2011)「語の共起と類似性」松本裕治(編)『言語と情報科学』朝倉書店 pp.58-76.

青木繁伸(2009)『統計数字を読み解くセンス―当確はなぜすぐにわかるのか?』(DOJIN 選書 27) 化学同人.

語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成

長谷川 守寿（首都大学東京）†
西尾 広美（国立国語研究所）

Building a 'Corpus of Documents Distributed in Kindergarten' for the Investigation of Vocabularies and Sentence Structures

Hasegawa Morihisa (Tokyo Metropolitan University)
Hiromi Nishio (National Institute for Japanese Language and Linguistics)

要旨

現在、多くの幼稚園では日本語を母語としない保護者（NonNativeSpeaker 保護者、以下 NNS 保護者）が見られるが、日本語学習の機会が少なく日本語が十分に理解できない場合、幼稚園の配布文書が正しく理解されず、情報伝達がうまくいかずに保育活動に支障をきたすこともある。そのため、将来的に教師と NNS 保護者を結ぶ「保護者に伝わるやさしい日本語」のテキスト化をめざし、『幼稚園の配布文書コーパス』を作成している。

コーパスの作成では、より精度の高い語彙・文型調査が行えるよう、OCR ソフトの認識誤りを人手だけで修正するのではなく、形態素解析システム（unidic-mecab2.1.2）も活用して誤りを発見して修正し、さらに正確に語に区切れない場合は表記の変更・記号の追加を行っている。本発表では、そのコーパス作成法について報告する。

１．はじめに

本稿は、『幼稚園の配布文書コーパス』の作成の手順を詳細に記述し、今後のデータ追加作成のための手順書となることを目指したものである。

現在、幼稚園児の保護者には日本語を母語としない人が見られるようになったが、中には日本語学習の機会が少なく、日本語が理解できないケースも出ている。そのような場合、幼稚園からの配布文書が正しく理解されず、情報伝達や意思疎通がうまくいかずに保育活動に支障をきたす、という問題も出てきている(西尾 2013)。

そこで地域や運営団体の異なる幼稚園で配布された文書を元に『幼稚園の配布文書コーパス』を作成し、語彙・文型調査を行い、将来的に教師と NNS 保護者を結ぶ「保護者に伝わるやさしい日本語」のテキスト化や、NNS 保護者が文書を理解する際に役立つ語彙表の作成などを予定している。本稿では、調査を行う前段階として、配布文書をどのようにテキストデータ化したのかを報告し、今後のコーパスの規模拡大へ向けた手順書とする。

本稿のコーパスの利用目的は語彙・文型調査が主であるため、作成の過程では語が正しく認定できることを優先している。より精度の高い語彙調査ができるようにするために、どのような作業をしているのか明らかにする。

２．『幼稚園の配布文書コーパス』の必要性

汎用のコーパスとしては、2011 年より国立国語研究所が『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）の DVD 版の配布を開始し、さらに少納言・中納言という検索サイトの公開を開始した。また特定目的のコーパスとしては、『日中 Skype 会話コーパス』(中俣 2015)や『児童・生徒作文コーパス』(宮城・今田 2015)、『学校お便りコーパス』(李 2016)

† hasegawa-morihisa@tmu.ac.jp

などのように、特定目的のコーパスも多数作成・公開されている。

しかし現在までのところ、我々の関心の対象である幼稚園の配布文書を収集したデータは存在しない。幼児教育の面からその分野で使用されている用語集などにあたるという方法も考えられるが、そうした用語が実際に配布文書で使われているのかというデータの真正性が保証されないため採用できない。そこで実際の配布文書を元に、語彙や文型調査に向けた『幼稚園の配布文書コーパス』を作成している。本稿ではその手順について述べる。

３.『幼稚園の配布文書コーパス』の構成と基本方針

３.１　コーパスの構成

本稿で説明する文書が実際に配布されたのは都内にある公立Ｓ幼稚園で、3 歳児クラスが 1 クラス、4・5 歳児クラスが 2 クラスずつで、合計 5 クラスからなる。

対象とする文書は、Ｓ幼稚園で平成 19 年度（4 月 9 日から翌年 3 月 13 日）に、園児の保護者に向けて配布された文書 93 種類である（幼稚園内部の文書は対象外)。ページ数は A4 用紙相当で 228 枚である（A3 用紙 1 枚は、A4 用紙 2 枚に換算)。なお、この期間に配布されたと考えられる資料『土と緑のＳ幼稚園』・『要覧』は、この幼稚園への入園を考えている幼児の保護者に向けた文書であり、入園に向けて準備する物の説明なども含まれ、非常に重要な配布文書と考えられるため、厳密にはその当時だけの園児の保護者向けではないが、対象とする。また李(2016)の『学校お便りコーパス』に含まれるような、いわゆるお便りだけではなく、保護者会などの資料も含めている。これは、保護者は幼稚園で配布される全ての文書を理解することが求められるからである。

３.２　コーパスの基本方針

紙の文書をテキスト化する際、レイアウトは無視し、文は文の形式で、箇条書きは箇条書きというように、そのまま入力することを基本とする。イラスト等は入力しない。表がある場合も語句のみ入力し、表形式では入力しない。表記の多様性を調べる目的ではないため、フォント情報等も考慮しない。

個人情報に関わる部分（個人が特定される可能性のある語句や氏名、呼び名など）は、全て“山田太郎”“太郎”で置き換え、幼稚園に関わる語句は全て“南大沢”に置き換える。これは、“○○”などの記号で置き換えた場合、正しく形態素解析できなくなる可能性があるため、正しく人名・地名と解析されるように“山田太郎”“南大沢”としたものである。

また本目的を遂行するために、配布文書をそのままテキスト化したのでは形態素解析で正しく語の境界を認定できない場合には、正しく認定できるようにテキストに修正を加える。これ以後、紙の状態のものを“プリント”、電子化されたものを“テキスト”と呼ぶ。

４.『幼稚園の配布文書コーパス』の作成法

コーパスの作成法は以下のとおりである。図 1 の流れに沿って作成方法について述べる。

４.１　配布文書の入手

幼稚園の配布文書の資料は、当時園長であった人から、平成 19 年度に園が保護者に配布した一年分の資料全てを提供してもらったものである（提供者の希望により名前は伏せる)。園長が年間の記録として保管していたということからも、真正性・網羅性の観点で問題がないと考える（研究使用の許諾も得ている)。配布文書は全てコピーし、実物は所有者に返却した。これ以降、配布文書と言及する際は全てコピーしたものを指す。

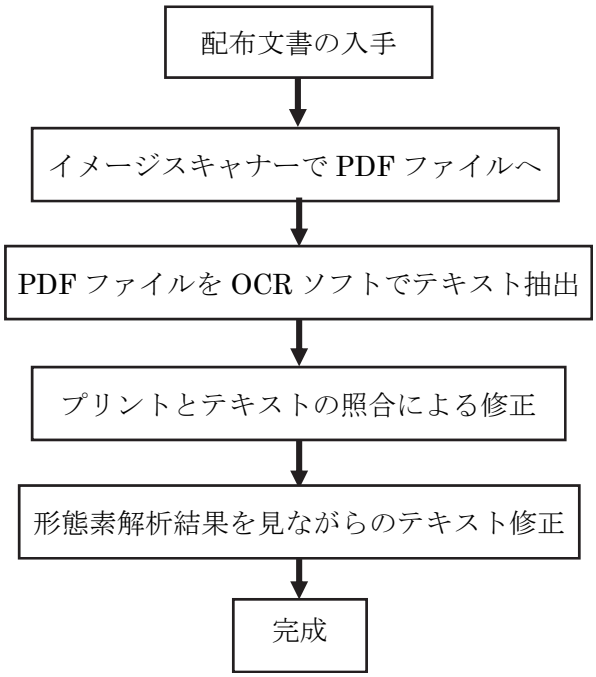


図 1. 配布文書コーパス作成の手順

４.２　イメージスキャナーで PDF ファイルへ

イメージスキャナーを用いて配布文書のイメージを取り込み、PDF ファイルにして保存する。本研究ではフォントの色等は研究の対象外となるので、白黒でスキャンする。

４.３　PDF ファイルを OCR ソフトでテキスト抽出

OCR ソフト（『読取革命 ver15』）で PDF ファイルとして保存された画像を文字化する。ファイル形式はテキストファイルにする。ファイル名は現在、イメージスキャナーの設定のままである。文字コードは S-JIS を採用する。手書きされたお知らせは OCR ソフトを用いても正しく文字認識ができないので、キーボード入力を行う。

４.３.１　プリントとテキストの照合による修正

OCR ソフトにより作成されたテキストの修正を行う。プリントとテキストを照合することで、OCR ソフトの誤りを発見する。OCR ソフトを用いたコーパスデータの作成法に関しては三井(2011)が詳しい。三井(2011)では、様々な原稿ごとにどのような誤りが見られたか述べられているが、ここでは、幼稚園の配布文書の場合に見られた誤りを挙げておきたい 。誤りは、便宜的に以下のＡからＣのように分けることができる。(1)から(3)に示すような誤り（矢印の左側）が見られたので、正しい形（矢印の右側）に変更した。

Ａ． 清音・濁音・半濁音などを含む語の認識が正しくないもの
Ｂ． 小書き文字を含む語の認識が正しくないもの
Ｃ． 字体の似ているもの（Ａ・Ｂ以外）

Ａに含まれるものには、(1)のように「キッズ」が「キッス」、「たんぽぽ」が「たんぽぼ」と認識されていたので、右側の正しい形に修正した。Ｂは(2)の「よ」と「ょ」のような小書き文字を含む誤りである。Ｃには、(3)の「せ」と「さ」、「一」と「ー」、「間」と「問」のように、三井(2011)などで指摘されているものも含まれるが、「於」と「１を」のように字体が似ていると考えにくいものまで含まれ、確認の作業では注意が必要である。

(1)　キッス＝＞キッズ、たんぽぽ＝＞たんぽぽ、ひろぱ＝＞ひろば、
　　なるぺく＝＞なるべく、楽しんたり＝＞楽しんだり
(2)　でしよう＝＞でしょう、ペインテイング＝＞ペインティング
(3)　年長せん＝＞年長さん、み一つけた＝＞みーつけた、時間＝＞時間、
　　終巣式＝＞終業式、1 を＝＞於、

このように、プリントとテキストを照合していく作業をしていく中で、OCR ソフトは正確に文字が抽出されているのであるが、中には元々プリント自体に誤字・脱字があったり、文法的な誤りがあるものがあることが分かった。

コーパス作成の方針としては、なるべくプリントに使われる語を正確に反映し、そのまま文字化することを優先するが、我々の目的が語彙表・文型リストの作成であるため、使われている意図を反映した語の抽出が必要となる。そこで前後の文脈から明らかな誤りと筆者らが判断したものは、テキストを正しい語・表記に修正することとした。

例えば、誤字脱字の例には、仮名漢字変換ではよく見られるもの(4)から、前後の文脈を確認しなければ誤りと気づけないような例(5)も見られた。また活用の問題(6)だけでなく、(7)のように判定が難しいものも含まれたが、著者らで相談し、前後の文脈から誤りと考えられるもののみ修正することとした。

また、対象とする文書は複数の教諭によって書かれていることから表記のゆれが生じる。例えば、「チーム」と「ティーム」、「貸し出し」と「貸し出」のようにゆれが見られた。そこで誤解析を防ぐために事前に「チーム」「貸し出し」に統一した。

(4)　スッテプ＝＞ステップ、として＝＞とおして、ゲー＝＞ゲーム
　　うっこつけい＝＞うこっけい、園庭解放＝＞園庭開放、機械＝＞機会
　　年長時＝＞年長児、あったたかくって＝＞あたたかくって
　　ジョカー＝＞ジョーカー、ちょとんと＝＞ちょこんと、めいいっぱい＝＞目一杯
(5)　テレビ付け＝＞テレビ漬け、持ち返す＝＞持ち返らす
(6)　ふれたりかかわりして＝＞ふれたりかかわったりして
　　楽しくようです＝＞楽しいようです
(7)　家族に一員として＝＞家族の一員として
　　子供に育てたいことを明確に捉え＝＞子供に育てたいことを明確に伝え
(8)　チーム／ティーム＝＞チーム、貸し出し／貸し出＝＞貸し出し

なお、この段階での修正には、「ドッチボール」「少しづつ」のように厳密には誤った形であるが、多く使われているため誤りと考えられないものも含まれている。我々は正しい形の方が NNS 保護者は辞書などで調べやすいであろうと推測し、「ドッジボール」「少しずつ」に変更した。なお、これらは形態素解析の際にも誤りとなるため、修正されることとなる。

４.３.２　形態素解析結果を見ながらのテキストの修正

前述のようなプリントとテキストのチェックを二度行ったが、OCR ソフトで文字認識を行い作成したテキストから、全ての誤りを目視で発見することは困難である。そこで、形態素解析にかけて、その結果を検証する中で、テキストの入力精度を上げ、表記を修正し、語彙調査に適したコーパスに変えていくこととする。

また長谷川・西尾(2016)の調査では、幼稚園のお知らせは、子供向けの文章ではないのだが、通常の表記に比べて、漢字よりもひらがなで書かれることが多いなど、いくつかの特徴を指摘した。このテキストを形態素解析にかけると、誤った結果が出てしまい、正確な語彙調査ができなくなってしまう。そのため正確な語彙調査を行うには、入力や表記の修正を行い、正しく解析できるように変更する前処理が必要となる。

そこで形態素解析器に MeCab(mecab-0.996.exe) 、形態素解析用辞書に UniDic-mecab (ver2.1.1)を使用し、入力したテキストの形態素解析を行う。その結果を基に、正確に語に区切ることができるか確認する。例えば「ハサミでシッポを切る」のような文を形態素解析すると表 1 のような結果になる（必要な部分のみ表示する)。

表 1. 形態素解析の実際（正しい解析の例）

書字形	語彙素読み	語彙素	品詞
ハサミ	ハサミ	鋏	名詞-普通名詞-一般
で	デ	で	助詞-格助詞
シッポ	シッポ	尻尾	名詞-普通名詞-一般
を	ヲ	を	助詞-格助詞
切る	キル	切る	動詞-非自立可能

形態素解析を行い、語の境界・品詞・見出し語（UniDic では語彙素）といった結果を確認する段階で、修正を行うことが必要となる箇所が多数見られる（ここでは、正しく解析できているとは、少なくても語の境界・語の品詞・見出し語の認定が正しいことを意味し、見出し語の読みについては問わない)。そこで後述の「トマトを食べる」や「ボタンかけを練習する」の解析結果（表 2）を用いて、正しく解析できていない場合について説明する。

表 2. 形態素解析の実際（誤解析の例）

書字形	語彙素読み	語彙素	品詞
トマ	トマ	トマ-Thomas	名詞-固有名詞-人名-一般
ト	ウラナイ	占い	名詞-普通名詞-一般
を	ヲ	を	助詞-格助詞
		＜略＞	
ボタン	ボタン	ボタン -button	名詞-普通名詞-一般
かけ	カケ	欠け	名詞-普通名詞-一般
を	ヲ	を	助詞-格助詞
		＜略＞	

表 2 で「トマト」は、見出し語で「トマ-Thomas」「占い」となっており、正しく語に区切られていないことから、ここに何らかの問題があることが分かる。また「ボタンかけ」は見出し語では「ボタン-button」「欠け」となっており、語彙素は「掛ける」となるべきであり、正しく語彙素が確定できていない（網掛けは問題のある箇所)。このような方法で誤りを発見し、その箇所を修正する作業を現在までに三回行った。その結果どのような誤りがあり修正したかを、OCR ソフトの問題と表記等に由来する問題に分けて述べる。

４.３.３　OCR ソフトによる誤認定の修正

まずは OCR ソフトの認識の問題である。人手による確認作業では見逃してしまったが、形態素解析の結果を確認して明らかになったものには、(9)(10)のような例がある。例えば(9)の「トマト」と「トマト」は非常に字体が似ている。しかし“トマ”はカタカナで、“ト”は漢字である。また“り”はひらがなで、“リ”はカタカナである。“口”はくにがまえ、“ロ”はクチである。目視ではフォントの微妙の違いしかない誤りは確認できなかったが、形態素解析にかけ、結果を確認し、品詞や見出し語が想定しているものと異なっているものを発見することで、(10)のような目視では見逃してしまった誤りにも気づき、修正を行うことができた。この作業によりテキストの精度向上が期待できる。

(9)　　トマト＝＞トマト、べりー＝＞ベリー、口＝＞ロ
(10)　　教青＝＞教育、雲悌＝＞雲梯

この作業は、使用した『読取革命』の設定を変更したり、使用する OCR ソフトを変更することで正しく認識できる可能性もあるが、入力の質を高める作業は必須となるであろう。

４.３.４　表記を変更したもの

形態素解析を行い、表 1 表 2 のように、語の境界・品詞・見出し語といった結果を確認することで、修正が必要となる箇所が数多く見られた。

そこで形態素解析の結果を検討し、実際に別表記で茶まめに入力して確認しながら、テキストを修正した。ここでは修正を便宜的に以下の 3 タイプに分け、説明する。

Ｍ１：文字種を変更したもの
　１．ひらがな表記だったものを、漢字表記に変えたもの
　２．ひらがな表記だったものを、カタカナ表記に変えたもの
　３．カタカナ表記だったものを、ひらがな表記に変えたもの
　４．カタカナ表記だったものを、漢字表記に変えたもの
　５．漢字表記を、別の漢字表記に変えたもの
Ｍ２：音引きを修正したもの
Ｍ３：語のつながりを修正したもの

Ｍ１：文字種を変更したもの

まず、Ｍ１－１に該当するひらがな表記を漢字に変更したものは、(11)(12)のように多数存在する。左側の括弧内には、そのまま解析した場合にどのように解析されるかを示した。

(11)　　おかずはいりません（現状では「入りません」と解析）＝＞おかずは要りません
　　テーブルふき（「吹き」）＝＞テーブル拭き
　　コップについで（「継いで」）＝＞コップに注いで
　　〇〇だより（「だ／より」）＝＞〇〇便り
　　友だちとかかわり（「とか／かわり」）＝＞友だちと関わり
　　くらいよみち（「くらい[助詞-副助詞]／よ[助詞-終助詞／道]」＝＞暗い夜道
　　うこっけい（う[感動詞]／滑稽）＝＞烏骨鶏
　　しっぽとり（しっ[感動詞]／ぽとり[副詞]）＝＞しっぽ取り
　　〇〇ぐみ（〇〇「グミ」）＝＞〇〇組
(12)　　おわん＝＞お椀

なお、(12)の「おわん」は文頭では「お（感動詞）／わん（副詞）」と解析されるが、「左手でおわんをもつ」のように文の形では「御」「椀」と正しく解析できる。(12)はイベントを説明する文書において、持ち物リストの中にあり、文頭に出現しているため、修正した。

次に、「Ｍ１－２　ひらがな表記だったものをカタカナ表記に変えたもの」について説明する。これには(13)が該当する。例えば、“どろけい”は、ひらがなのままでは（泥＋ケイ［人名］）と解析されるが、カタカナ表記に変えると「泥警」［泥棒と警察］と正しく解析できる。“すだしい”は、スダシイとカタカナ表記にすることによって「すだ椎」と一語に解析できる。ただし文中での「なす」は正しく解析できるが、括弧の後の「なす」は動詞と解析されるため、その出現位置に現れる“なす”のみ変更した。

(13)　　どろけい＝＞ドロケイ、すだじい＝＞スダジイ、なす＝＞ナス

同様に、「Ｍ１－３　カタカナ表記だったものをひらがな表記に変えたもの」(14)、「Ｍ１－４カタカナ表記だったものを漢字表記に変えたもの」(15)、「Ｍ１－５　漢字表記を別の漢字表記に変えたもの」(16)を挙げることができる。(16)は、文脈では「たくさん休んで下さい」という意味で用いており、「十分」でも「じゅうぶん」という読み方は存在するが、形態素解析では数字としてしか解析されないため、「充分」に変更した。

(14)　シトシト降る＝＞しとしと降る
(15)　ヨウシュヤマゴボウ＝＞洋種山牛蒡、ダンボール＝＞段ボール
(16)　十分休む＝＞充分休む

この他に、テキストの修正において文字コードの問題が 1 件発生した。上記のように正しく語が認定できるよう修正する作業を行ってきたが、修正できない問題も生じた。それは「鼻をかむ」という表現で、語彙素のレベルでは「鼻」「を」「噛む」と解析される点である。正しく「擤む」と判定されるには、テキストを「擤む」に変えなければならないが、S-JIS では保存できず、UTF-8 などで保存するしかない。UTF-8 を採用していれば正しく語彙素を判定できるのであるが、現在は S-JIS を採用しているため、修正できていない。

Ｍ２．音引きの修正

一部の語の音引きについては誤った解析結果になるので、(17)のような修正を加えた。ただし UniDic には「だーいすき」「できなーい」「ずーっと」「はーい」「よーい」など音引き形が辞書の書字形に登録されている語もあり、正しく解析できる語はそのままである。

(17)　いただきまーす＝＞いただきます、おいしーい＝＞おいしい
　　はいりたーい＝＞はいりたい、てんきにな～れ＝＞てんきになれ
　　楽し～い＝＞楽しい、すご～い＝＞すごい

Ｍ３．出現環境の修正

UniDic は単語（短単位）に区切られたものの組み合わせの中からコストが最小の組み合わせを正解として出力するという特徴がある（小木曽 2014）。例えば「①節分」は、「①／節分」よりも、コストが小さい「①／節／分」が解析結果として選ばれる。2 月の豆まきの予定で出てきた表現なので、このように語を認定されるのは正しくなく、「節分」で正しく語に区切ることができるように工夫する。この場合は“①”と“節分”の間に読点“、”を入れると、「①／節分」と正しく語に区切れることが確認できたので、読点を挿入し、正しく語に区切ることができるようにしておく。読点“、”や空白“□ ”ならば、記号として解析され、数える際に外せばよいので、正しく解析でき、語数に影響しないからである。

具体的にどのような修正を加えたか、例を挙げて説明する。(18)は、そのままでは「違い（名詞-普通名詞-一般）」と解析されるが、読点を入れることによって「違う（動詞-一般）」と正しく解析することができる。また(19)は、語彙素レベルでは「水揚げ／良い／ね」と解析されるが、読点を挿入することによって「水／上げる／ね」と正しく解析することができる。なお(20)のような例の場合、読点の挿入では「学びや」は「学舎（まなびや）」と解析され、結果は変わらない。このような場合は、空白“□”を挿入した。この方法により専門性が高く独特な語でかつ臨時一語的な語も、語の境界を正しく認定できることになる。例えば(21)はそのままでは「新年／中」となるが、正解は新しい年中児という意味なので、「新／年中」となるのが正しい。そのため「新□年中」とした。

(18)　製作とは違い大掛かりな作業です＝＞製作とは違い、大掛かりな作業です
(19)　水あげようね＝＞水、あげようね
(20)　その後の学びや創造性が＝＞その後の学び□や創造性が

(21)　新年中＝＞新□年中、弁当時＝＞弁当□時、再任用主事＝＞再□任用□主事

UniDic の開発がさらに進んで、より精度の高い解析結果が出せるようになったとき、上記のような前処理は必要なくなるかもしれないが、当面 UniDic を用いて調査を行うには、必須となる処理であろう。特に幼稚園の配布文書のような、かなり特殊なテキスト化したものを対象とする際には、辞書の書字形に含まれている表記も考慮する必要が出てくる。

５．今後の課題

本稿では、正しさについて語の境界と品詞のみとし、読みについては不問とした。今後語彙調査を実施するにあたり、語彙素読みの修正を行う必要がある。例えば、「年中」について、文書の中では(22)のように、“ねんちゅう”と読ませるもののみであり、“ねんじゅう”と読ませる例は一例もない。しかし形態素解析の結果、語彙素読みは“ねんじゅう”であるため修正が必要となる。同様の例が「お母様」(23)であり、読みは“おははさま”である。これは UniDic の問題でもあるが、修正が必要となる。また(24)の“お家”は(25)のように「おうち」と読ませる意図と思われるが、解析結果は“おいえ”である。語彙素の読みに誤りを含むものは他にも多数存在するため、修正が必要となると考える。

(22)　１学期は年長組を中心に行い、徐々に年中組にも広げていく予定です。
(23)　先日は、お母様方の協力のもと、楽しい夕涼み会ができました。
(24)　お家でも遊んでいるようなおもちゃを用意して安心して過ごせるようにしている。
(25)　自信をもって取り組んだ姿におうちでも（略）誉めてあげてください。

今後は読みも含めた短単位の語彙表を完成させ、長単位での語彙表を作成したいと考えている。これは語彙表作成を見すえた場合、単語表は長単位を基に作成した方が望ましいためである。これには、短単位を元に長単位を認定する解析器 Comainu を用いる予定であるが、元となる短単位が正確に区切られていなければ、正確な長単位も認定できないため、短単位の認定の精度を上げる必要があると考える。

文　献

小木曽智信(2014)「第 5 章　形態素解析」前川喜久雄監修・山崎誠編『講座日本語コーパス　２　書き言葉コーパス―設計と構築―』、朝倉書店、pp.89-115

中俣尚己(2015)「『日中 Skype 会話コーパス』について」、（http://nakamata.info/about_skype_corpus.pdf、最終確認 2017 年 2 月 3 日）

西尾広美（2013）「幼稚園における『やさしい日本語』使用の必要性―教師と非母語話者の保護者のコミュニケーションの現状調査から―」『日本語研究』33、首都大学東京・都立大学・日本語・日本語教育研究会、pp.99-102

長谷川守寿・西尾広美(2016)「『幼稚園の配布文書コーパス』の作成と試行調査」『言語処理学会 第 22 回年次大会 発表論文集』、言語処理学会、pp.246-249

三井正孝(2011)「第 1 章 コーパスデータの作成―OCR ソフトを利用して―」荻野綱男・田野村忠温編『講座 IT と日本語研究５ コーパスの作成と活用』、明治書院、pp.7-45

宮城信・今田水穂(2015)「『児童・生徒作文コーパス』の設計」『第 7 回コーパス日本語学ワークショップ予稿集』、国立国語研究所、pp.223-228

李暁燕(2016)「『学校お便りコーパス』について」(http://lixiaoyan.jp/database/、最終確認 2017 年 2 月 3 日）

固有表現抽出におけるアノテーション手法の比較

鈴木雅也 (茨城大学)*
古宮嘉那子 (茨城大学)†
岩倉友哉 (富士通研究所)‡
佐々木稔 (茨城大学)§
新納浩幸 (茨城大学)¶

Comparison of Annotating Methods in Named Entity Extraction

Masaya Suzuki (Ibaraki University)
Kanako Komiya (Ibaraki University)
Tomoya Iwakura (Fujitsu Laboratories Ltd.)
Minoru Sasaki (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要 旨

本稿では, 非専門家による固有表現抽出のタスクとしてのアノテーションを題材に, ふたつの手法について比較を行った. ひとつは既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行う手法であり, もうひとつは人手で一からアノテーションを行う手法である. 実験には現代日本語書き言葉均衡コーパス (BCCWJ) を利用し, 手法ごとに 1 テキストに対し 2 人の非専門家を割り当てて, アノテーションを行った. 評価には, アノテーションにかかる時間, 一致率, Gold Standard との比較による正解率, それぞれの手法で作成されたコーパスを訓練事例とした場合の正解率を利用し, ジャンルごと, 及び, 全ジャンルのマイクロ平均とマクロ平均を算出した. 本実験の結果から, 全ジャンルのマイクロ平均とマクロ平均で比較した場合には既存のアノテーション結果を用いた手法の方が良い結果となるが, 既存の固有表現抽出器の訓練事例から離れたジャンルで同様に比較した場合には人手でアノテーションを行う手法の方が良い結果となることが明らかになった.

1. はじめに

非専門家をアノテータとする, クラウドソーシングによるコーパスへのアノテーションは, 安価で速く仕上がることが Snow ら (Snow et al. 2008) によって明らかとなっている. しかし, アノテーション手法に起因したアノテーションの品質の違いについては, これまで言及さ

* 13t4038a@vc.ibaraki.ac.jp
† kanako.komiya.nlp@vc.ibaraki.ac.jp
‡ iwakura.tomoya@jp.fujitsu.com
§ minoru.sasaki.01@vc.ibaraki.ac.jp
¶ hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

れてこなかった. 固有表現抽出におけるアノテーションはルールが多く複雑なため, 非専門家にとってタグの付け間違いが発生しやすいタスクとなっており, この観点での議論が必要なタスクのひとつであると考えられる. そこで, 本稿では, 固有表現抽出におけるアノテーションを題材として, 非専門家の手で高品質なコーパスを作成するための手法についての考察を行った. なお, 本稿は (Komiya et al. 2016) を元にしている.

固有表現抽出におけるアノテーションでのタグの付け間違いを減らすための手法として, 既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行う手法が考えられる. しかし, 訓練事例として特定ジャンルのコーパスのみを用いている固有表現抽出器の場合, 特にそのジャンル以外のコーパスのアノテーションにおいて, タグの付け間違いが発生することがある. そこで, 本研究では, 前述の手法と既存の固有表現抽出器を使用せず, 人手でアノテーションを行う手法のふたつの手法について, アノテーションにかかる時間, タグの一致率, Gold Standard との比較による正解率の各観点から比較することで考察を行った. この際, テキストのジャンルに起因したアノテーションの品質の違いについても考察を行っている.

2. 関連研究

アノテーションに関する先行研究としては, 次のようなものが挙げられる. Snow ら (Snow et al. 2008) は, 非専門家によるコーパスへのアノテーションに関して, アノテーションにかかる時間, アノテーションの品質, コストの観点から, 専門家が行った場合と比較することで考察を行った. Alex ら (Alex et al. 2010) は, 反復的で agile なアノテーション手法を提案し, 既存の線形によるアノテーション手法との比較を行った. van der Plas ら (Plas et al. 2010) は, 英語のテンプレートを用いたフランス語のコーパスへの意味情報の付与を題材に, 言語横断的なアノテーションの信頼性について考察を行った. Marcus ら (Marcus et al. 1993) は, 品詞アノテーションや bracketing といったタスクのための Penn TreeBank を開発するため, 既存のアノテーション結果を用いる手法と人手のみで行う手法について比較を行った. しかし, 我々が知る限り, 非専門家の手で高品質なコーパスを開発するために, 既存のアノテーション結果を用いる手法と人手のみで行う手法を比較したという論文は存在しない.

本稿では, 固有表現抽出におけるアノテーションを題材にしている. 固有表現抽出とは, 固有名詞に時間や数値といった表現を加えた概念である固有表現を文章中から抽出するタスクであり, 昔から研究が行われてきた. このタスクに関する先行研究としては, 次のようなものが挙げられる. 橋本ら (橋本泰一ほか 2008, 橋本泰一・中村俊一 2010) は CD-毎日新聞'95 データ集(1)や現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa 2008)(2) を元に, 拡張固有表現タグ付きコーパス(3)を作成した. 徳永ら (徳永健伸ほか 2015) は, 固有表現抽出におけるアノテータの視線分析を行った. Sasada ら (Sasada et al. 2015) は, 部分的なタグ付きテキストを用いて訓練可能な固有表現抽出器を提案した. Sekine ら (Sekine and Isahara 2000)

(1) http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
(2) http://pj.ninjal.ac.jp/corpus_center/bccwj/
(3) http://www.gsk.or.jp/catalog/gsk2014-a/

は Message Understanding Conference-6 (MUC-6)[4] での定義 (Grishman and Sundheim 1996) を元に, Information Retrieval and Extraction Exercise (IREX)[5] で固有表現抽出の共通タスクを行うため, 8 種類の固有表現タグ (組織名, 人名, 地名, 固有物名, 日付表現, 時間表現, 金額表現, 割合表現), 及び, それらと同等に扱われるオプショナルタグからなる 9 種類のタグを定義した. しかし, IREX で用いられたのは, 新聞コーパスのみであった.

2014 年, 6 領域から構成される Project Next NLP (岩倉友哉 2015, 平田亜衣・小町守 2015, Ichihara et al. 2015)[6] において, 前述の拡張固有表現タグ付きコーパスを用いた固有表現抽出のエラー分析が行われた. Ichihara ら (Ichihara et al. 2015) は, 既存の固有表現抽出器の性能について調べ, 固有表現抽出器の訓練事例から離れたジャンルのテキストにおいて, タグの付け間違いが増加することを示した. 本稿では, 訓練事例から離れたジャンルのコーパスにおいて, 既存のアノテーション結果を用いた手法ではタグの付け間違いが発生する場合があるということを示す.

本研究では, 非専門家の手で高品質なコーパスを作成するため, 固有表現抽出のタスクについて, 既存のアノテーション結果を用いた手法と人手のみでアノテーションを行う手法のふたつの手法による, アノテーションにかかる時間, タグの一致率, Gold Standard との比較による正解率の評価を行った.

3. アノテーション手法の比較

本稿では, 次のふたつのアノテーション手法について比較を行った.

- KNP+Manual
  既存の固有表現抽出器 KNP (Sasano and Kurohashi 2008)[7] によるアノテーション結果に対し, 人手で修正を行う.
- Manual
  人手で一から固有表現のアノテーションを行う.

また, 比較を行うにあたり, それぞれのテキストに対するアノテーションにかかる時間, タグの見かけの一致率とカッパ係数, Gold Standard との比較による適合率 (精度), 再現率, F 値を指標として設定した.

ふたつの手法間で一致したタグの個数が表 1 で示されるとき, 見かけの一致率とカッパ係数はそれぞれ式 (1) と 式 (2) で与えられる.

$$d = \frac{\sum_{i=1}^{n} a_{ii}}{a_{00}} \tag{1}$$

[4] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
[5] http://nlp.cs.nyu.edu/irex/index-j.html
[6] https://sites.google.com/site/projectnextnlp/
[7] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

表 1　ふたつの手法間で一致したタグの個数

		手法 X				
		タグ 1	タグ 2	...	タグ n	合計
手法 Y	タグ 1	a_{11}	a_{21}	...	a_{n1}	a_{01}
	タグ 2	a_{12}	a_{22}	...	a_{n2}	a_{02}
	...	...	...	...	...	...
	タグ n	a_{1n}	a_{2n}	...	a_{nn}	a_{0n}
	合計	a_{10}	a_{20}	...	a_{n0}	a_{00}

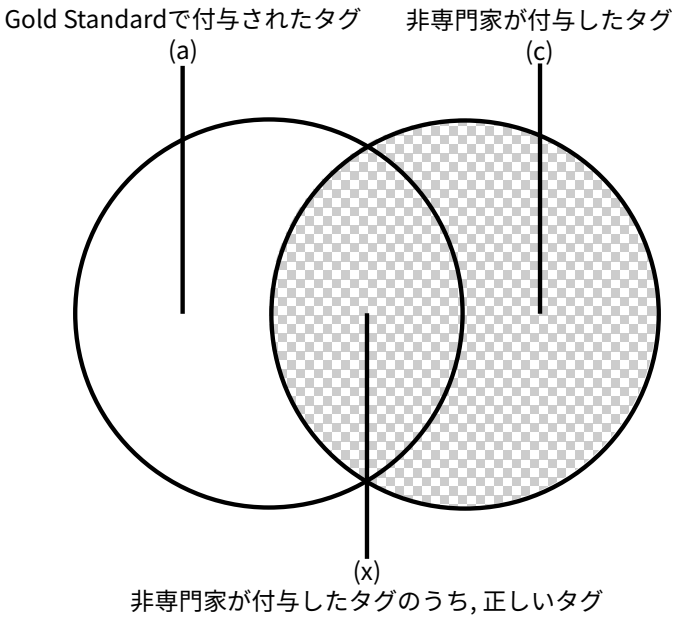


図 1　タグ集合

$$\kappa = \frac{a_{00}\sum_{i=1}^{n} a_{ii} - \sum_{i=1}^{n} a_{i0}a_{0i}}{(a_{00})^2 - \sum_{i=1}^{n} a_{i0}a_{0i}} \tag{2}$$

また, タグ集合が図 1 のように示されるとき, 適合率, 再現率, F 値はそれぞれ 式 (3), 式 (4), 式 (5) のように与えられる.

$$p = \frac{n(x)}{n(c)} \tag{3}$$

$$r = \frac{n(x)}{n(a)} \tag{4}$$

$$f = \frac{2pr}{p+r} \tag{5}$$

表 2　ジャンルごとのテキストとそこに含まれるタグの数

ジャンル	テキスト	タグ									
		Artifact	Date	Location	Money	Organization	Percent	Person	Time	Optional	合計
OC	74	44	18	65	9	18	0	6	0	8	168
OW	8	86	143	147	9	136	33	15	0	26	595
OY	34	23	61	59	7	64	10	79	3	17	323
PB	5	32	49	100	0	19	5	174	9	20	408
PM	2	9	24	36	5	18	1	216	3	1	313
PN	13	24	166	192	60	123	37	78	22	20	722
合計	136	218	461	599	90	378	86	568	37	92	2, 529

4. 実験

本実験では, ClassA-1[(8)] に分類される 136 テキストを BCCWJ より抽出して用いた. ClassA-1 に分類される BCCWJ のテキストは, Yahoo! 知恵袋 (OC), 白書 (OW), Yahoo! ブログ (OY), 書籍 (PB), 雑誌 (PM), 新聞 (PN) の 6 ジャンルで構成されている. それぞれのジャンルにおけるテキストとそこに含まれるタグの数は表 2 の通りである. なお, 本実験では固有表現抽出器として KNP Ver.4.16 (Linux 版) と JUMAN Ver.7.01 (Linux 版)[(9)] を用いており, 前者は訓練事例として新聞コーパスを用いている (Sasano and Kurohashi 2008)[(10)].

被験者は非専門家 16 人であり, IREX によるアノテーションのルール (Inf 1999) を読み合わせた後, これに従って 9 種類のタグによるアノテーションを行った. この際, 全ての被験者のアノテーション結果を集めたときに, それぞれの手法について, 2 セットのコーパスを構成できるよう, 被験者は割り当てられた 34 テキストに対し, それぞれの手法を半分ずつ適用した. また, 習熟によるバイアスがかかりにくくするため, 被験者をふたつのグループに分け, 最初に適用する手法をグループごとに変えた. なお, アノテーションの際には, テキストごとのアノテーションにかかる時間の記録も行っており, それを元に手法ごとのアノテーションにかかる平均時間を算出した. また, 本実験では Gold standard として BCCWJ NE コーパス (2016 年 2 月 1 日版) (Iwakura et al. 2016)[(11)] を用いている.

Gold standard は IREX によるアノテーションのルールに基づき作成した. Gold standard にオプショナルタグが付与されているときはその範囲を超えてタグが付与されていない場合を, それ以外のときはタグとその範囲が Gold standard と一致している場合を正解としている.

本実験では, 手法ごとに 1 テキストに対し 2 人の非専門家を割り当てて, アノテーションを行ったという条件下で, 2 人のアノテータの平均正解率と, どちらか一方でも正解のタグを付与しているならば正解とみなした際の正解率を算出した. 後者は, 実際にコーパスを作成する際, 2 人のアノテータによるアノテーション結果を統合して作成することが想定されるため, 算出を行った.

これらに加え, 機械学習における訓練事例としての品質を確かめるため, それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いてアノテーションを行った. この際に用い

(8) http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list

(9) http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

(10) 厳密には Web 上の記事も訓練事例として用いられているが, 本稿では訓練事例としてのウエイトが大きい新聞コーパスを KNP の訓練事例として扱っている.

(11) https://sites.google.com/site/projectnextnlpne/

表 3　一致率のマイクロ平均 (全体)

手法	見かけの一致率	カッパ係数
KNP+Manual	**0.79**	**0.75**
Manual	0.57	0.50
Both	0.64	0.58

表 4　一致率のマクロ平均 (全体)

手法	見かけの一致率	カッパ係数
KNP+Manual	**0.66**	**0.48**
Manual	0.52	0.29
Both	0.52	0.31

た素性は, 形態素, 文字種, 品詞タグ, 分類[(12)], キャッシュ素性, 統合素性, 格フレーム素性であり, これはオリジナルの KNP と同様である (Sasano and Kurohashi 2008). なお, それぞれの手法における 2 人分のアノテーション結果を結合したものをその手法の訓練事例としており, また, できる限り多くのジャンルのテキストを含むような形で 5 分割交差検定を行っている.

5. 結果

表 3, 表 4 はそれぞれの手法の見かけの一致率とカッパ係数のマイクロ平均とマクロ平均を示しており, 表 5, 表 6 はそれらをジャンルごとに示したものである. これらにおける **Both** はふたつの手法を用いた計 4 人のアノテータによるアノテーション結果の全てのペアを取ったときの一致率の平均を示している.

表 7, 表 8 はそれぞれの手法の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示しており, 表 9, 表 10 はそれらをジャンルごとに示したものである. これらにおける **KNP** はオリジナルの KNP によるアノテーション結果の正解率を, **Average** は **KNP+Manual** と **Manual** の平均を示している.

なお, ふたつの手法の中でより高い水準を記録した見かけの一致率, カッパ係数, 適合率, 再現率, F 値については太字で示している. また, 表 11 は, それぞれの手法における 1 テキストあたりのアノテーションにかかる平均時間を示している.

次に, 2 人のアノテータのうち, どちらか一方でも正解のタグを付与しているならば正解とみなしたとき (2 人のアノテータによるアノテーション結果を統合したとき) の性能について調べた. 表 12, 表 13 はそれぞれの手法の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示しており, 表 14, 表 15 はそれらをジャンルごとに示したものである. **KNP** と **Average** に関しては, 表 7〜表 10 と同様である.

これらに加え, それぞれの手法で作成されたコーパスを訓練事例とした固有表現抽出器の

(12) データとして存在する場合のみ.

表 5　一致率のマイクロ平均 (ジャンルごと)

ジャンル	手法	見かけの一致率	カッパ係数
OC	KNP+Manual	**0.62**	**0.54**
OC	Manual	0.47	0.34
OC	Both	0.52	0.41
OW	KNP+Manual	**0.78**	**0.73**
OW	Manual	0.41	0.28
OW	Both	0.55	0.46
OY	KNP+Manual	**0.69**	**0.63**
OY	Manual	0.58	0.50
OY	Both	0.57	0.49
PB	KNP+Manual	**0.76**	**0.68**
PB	Manual	0.67	0.56
PB	Both	0.71	0.61
PM	KNP+Manual	**0.87**	**0.84**
PM	Manual	0.61	0.55
PM	Both	0.69	0.64
PN	KNP+Manual	**0.86**	**0.75**
PN	Manual	0.81	0.65
PN	Both	0.80	0.65

性能を調べた. 表 16, 表 17 はそれぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示しており, 表 18, 表 19 はそれらをジャンルごとに示したものである.

まず, マイクロ平均について比較する. 表 7, 表 16 における適合率と再現率, 表 14 における適合率について, 有意水準 0.05 のカイ二乗検定で検定を行った場合, **KNP** と **KNP+Manual**, **KNP** と **Manual**, **Manual** と **KNP+Manual** は統計的に有意である. また, 正解率におけるジャンルごとのマイクロ平均 (表 9, 表 14, 表 18) のうち, アスタリスクが付与されている箇所においては, 適合率, または, 再現率について同様に検定を行った場合, **Manual** と **KNP+Manual** は統計的に有意である. しかし, 表 12 において, 再現率について同様に検定を行った場合, **KNP** と **KNP+Manual**, **KNP** と **Manual** は統計的に有意であるが, **Manual** と **KNP+Manual** は有意ではない. また, 正解率のマクロ平均について同様に検定を行った場合, 標本数が少ないという理由から, 統計的に有意ではない.

表 6　一致率のマクロ平均 (ジャンルごと)

ジャンル	手法	見かけの一致率	カッパ係数
OC	KNP+Manual	**0.58**	**0.27**
OC	Manual	0.50	0.15
OC	Both	0.47	0.14
OW	KNP+Manual	**0.80**	**0.73**
OW	Manual	0.45	0.36
OW	Both	0.59	0.50
OY	KNP+Manual	**0.63**	**0.47**
OY	Manual	0.50	0.29
OY	Both	0.47	0.30
PB	KNP+Manual	**0.63**	**0.54**
PB	Manual	0.60	0.43
PB	Both	0.62	0.48
PM	KNP+Manual	**0.87**	**0.83**
PM	Manual	0.62	0.55
PM	Both	0.69	0.63
PN	KNP+Manual	**0.88**	**0.74**
PN	Manual	0.74	0.56
PN	Both	0.77	0.59

表 7　2 人のアノテータの平均正解率のマイクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.78	0.68	0.73
KNP+Manual	**0.84**	**0.81**	**0.83**
Manual	0.75	0.73	0.74
Average	0.80	0.77	0.78

6. 考察

6.1 一致率とアノテーションにかかる時間

表 3, 表 4 より, **KNP+Manual** の一致率は, **Manual** の一致率よりもマイクロ平均, マクロ平均ともに高い数値となっていることがわかる．また, 表 5, 表 6 より, 全てのジャンルについて同様の傾向が見られることがわかる．これは, **KNP+Manual** の人手により修正される前のコーパスが共に同じ固有表現抽出器によってアノテーションされたものであることが影

表 8　2 人のアノテータの平均正解率のマクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.47	0.40	0.43
KNP+Manual	**0.55**	**0.55**	**0.55**
Manual	0.53	0.51	0.52
Average	0.54	0.53	0.53

響していると考えられる．さらに, 表 11 より, **KNP+Manual** における 1 テキストあたりのアノテーションにかかる時間は, **Manual** よりも平均約 2 分程度短いということがわかる．これは有意水準 0.01 の F 検定で検定を行った場合, 統計的に有意である．これらのことから, **KNP+Manual** は **Manual** よりもアノテーションにかかる時間が短く, 一致率が高いということがいえる．

また, 表 5, 表 6 より, **Both** の一致率は多くの場合, **Manual** と同等以上の数値となっているが, OC における一致率のマクロ平均は, **Both** が 0.01 ポイント以上 **Manual** を下回っていることがわかる．このことから, OC には新聞コーパスから生成したルールだけでは抽出できないような固有表現が多く含まれているということがわかる．

さらに, 表 3, 表 4 より, **Manual** のカッパ係数に関して, マイクロ平均では適度な値だったのに対し, マクロ平均では低い値となっていることがわかる．マイクロ平均は固有表現ごとの平均であり, マクロ平均はテキストごとの平均であるということから, **Manual** ではテキストごとに見たときに, 一致率の偏りが大きいということがいえる．

6.2　正解率

表 7, 表 8 より, **KNP+Manual** の正解率は, **Manual** の正解率よりもマイクロ平均, マクロ平均ともに高い数値となっていることがわかる．しかし, 表 9 より, OC における再現率と PM における適合率のマイクロ平均についてはこの傾向が見られず, また, **KNP** におけるこれらの指標は, 他のジャンルよりもかなり低い値となっていることがわかる．このことから, **KNP+Manual** の正解率は **KNP** の正解率に依存しているということがいえる．

また, 表 10 より, **KNP+Manual** の正解率は, OY, OW, PN については **Manual** の正解率よりも高い値となっている一方, OC, PB, PM については, PM の再現率を除き **Manual** の正解率よりも低い値となっていることがわかる．さらに, KNP の訓練事例である新聞コーパスに近く, **KNP** による正解率が高くなることが示されている (Ichihara et al. 2015) OW と PN において, **KNP** の正解率は **Manual** の正解率よりも高い値となっている．これらのことから, 非専門家がアノテーションを行う場合, KNP の訓練事例に近いジャンルのテキストについては **KNP+Manual** の方が良い結果を得られ, KNP の訓練事例から離れたジャンルのテキストについては **Manual** の方が良い結果を得られるということがいえる．

表 9　2 人のアノテータの平均正解率のマイクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.72	0.48	0.57
OC	KNP+Manual	***0.78**	0.75	**0.77**
OC	Manual	0.67	**0.80**	0.73
OC	Average	0.72	0.78	0.75
OW	KNP	0.79	0.79	0.79
OW	KNP+Manual	***0.82**	***0.85**	**0.83**
OW	Manual	0.65	0.67	0.66
OW	Average	0.73	0.76	0.74
OY	KNP	0.73	0.57	0.64
OY	KNP+Manual	***0.85**	***0.75**	**0.80**
OY	Manual	0.80	0.68	0.74
OY	Average	0.83	0.72	0.77
PB	KNP	0.75	0.60	0.66
PB	KNP+Manual	**0.79**	**0.74**	**0.76**
PB	Manual	0.78	0.73	0.75
PB	Average	0.78	0.73	0.76
PM	KNP	0.61	0.58	0.59
PM	KNP+Manual	0.89	**0.86**	**0.87**
PM	Manual	**0.90**	0.85	0.87
PM	Average	0.89	0.86	0.87
PN	KNP	0.88	0.78	0.83
PN	KNP+Manual	***0.88**	***0.85**	**0.86**
PN	Manual	0.77	0.72	0.75
PN	Average	0.83	0.79	0.81

6.3　2 人のアノテータによるアノテーション結果を統合したときの正解率

表 12, 表 13 より, どちらか一方でも正解のタグを付与しているならば正解とみなした場合, **KNP+Manual** の正解率は **Manual** の正解率よりも高い値となっているが, 2 人のアノテータの平均正解率 (表 7, 表 8) に比べると, その差はかなり小さいということがわかる. これは, 2 人のアノテータのうち, 少なくともどちらか一方は正しいタグを付与していることが多いためであると考えられる.

また, 表 7, 表 8 は 2 人のアノテータの正解率の平均であることから 1 人のアノテータの正解率, 表 12, 表 13 は 2 人のアノテータの正解率とみなすことができる. すると, 表 7, 表 8, 表

表 10　2 人のアノテータの平均正解率のマクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.31	0.26	0.28
OC	KNP+Manual	0.39	0.41	0.40
OC	Manual	**0.42**	**0.44**	**0.43**
OC	Average	0.40	0.42	0.41
OW	KNP	0.77	0.80	0.79
OW	KNP+Manual	**0.83**	**0.86**	**0.84**
OW	Manual	0.70	0.73	0.71
OW	Average	0.76	0.79	0.78
OY	KNP	0.58	0.44	0.50
OY	KNP+Manual	**0.68**	**0.63**	**0.66**
OY	Manual	0.56	0.49	0.52
OY	Average	0.62	0.56	0.59
PB	KNP	0.66	0.46	0.54
PB	KNP+Manual	0.71	0.65	0.68
PB	Manual	**0.81**	**0.67**	**0.74**
PB	Average	0.76	0.66	0.71
PM	KNP	0.60	0.66	0.63
PM	KNP+Manual	0.82	**0.87**	0.85
PM	Manual	**0.86**	0.84	**0.85**
PM	Average	0.84	0.85	0.85
PN	KNP	0.88	0.78	0.82
PN	KNP+Manual	**0.88**	**0.85**	**0.86**
PN	Manual	0.78	0.72	0.75
PN	Average	0.83	0.78	0.81

12, 表 13 より, 2 人のアノテータによる正解率は常に 1 人のアノテータによる正解率よりも高い値となっていることがわかる. さらに, 2 人のアノテータによる **Manual** の正解率は, 常に 1 人のアノテータによる **KNP+Manual** の正解率よりも高い値となっていることがわかる. このことから, 非専門家をアノテータとする場合, 既存の固有表現抽出器を使用すること以上に, アノテータの人数を増やすことが良い結果を得る上で重要であるといえる.

6.4　訓練事例としてのアノテーション結果

表 16, 表 17 より, それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合, **KNP+Manual** を訓練事例とした場合の正解率は **Manual** を訓練事例とした場合の

表 11　アノテーションにかかる平均時間 (手法ごと)

手法	時間
KNP+Manual	3:19
Manual	5:23

表 12　2 人のアノテータによるアノテーション結果を統合したときの正解率のマイクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.78	0.68	0.73
KNP+Manual	**0.91**	**0.89**	**0.90**
Manual	0.87	0.88	0.88
Average	0.89	0.89	0.89

表 13　2 人のアノテータによるアノテーション結果を統合したときの正解率のマクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.47	0.40	0.43
KNP+Manual	**0.63**	**0.62**	**0.63**
Manual	0.62	0.62	0.62
Average	0.63	0.62	0.63

正解率よりも高い値となっていることがわかる．しかし, 表 18, 表 19 より, PB と PN における適合率のマイクロ平均とマクロ平均, 及び, PB における F 値のマクロ平均についてはこの傾向が見られないことがわかる．このことから, **KNP+Manual** よりも **Manual** を訓練事例とした方が良いアノテーション結果となる場合もあるということがわかる．

また, 表 16, 表 17 より, オリジナルの **KNP** の正解率は **KNP+Manual** や **Manual** を訓練事例とした場合の正解率よりも高い値となっていることがわかる．これは, **KNP+Manual** や **Manual** で作成されたコーパスが, オリジナルの **KNP** の訓練事例に比べ, とても少ないためであると考えられる．一方で, 表 16, 表 17 より, **KNP+Manual** や **Manual** を訓練事例とした場合, 及び, オリジナルの **KNP** において, 適合率は再現率に比べて大きな差がなく, また, 表 18 より, OC と OY の適合率のマイクロ平均において, オリジナルの **KNP** よりも **KNP+Manual** や **Manual** を訓練事例とした場合の方が高い値となっていることがわかる．このことから, 訓練事例が少ないとしても, 適合率はオリジナルの KNP と同等以上になるということがいえる．

表 14　2 人のアノテータによるアノテーション結果を統合したときの正解率のマイクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.72	0.48	0.57
OC	KNP+Manual	**0.87**	0.86	0.87
OC	Manual	0.86	**0.91**	**0.88**
OC	Average	0.87	0.89	0.88
OW	KNP	0.79	0.79	0.79
OW	KNP+Manual	***0.91**	**0.91**	**0.91**
OW	Manual	0.76	0.89	0.82
OW	Average	0.84	0.90	0.87
OY	KNP	0.73	0.57	0.64
OY	KNP+Manual	**0.94**	**0.87**	**0.90**
OY	Manual	0.93	0.86	0.89
OY	Average	0.94	0.87	0.90
PB	KNP	0.75	0.60	0.66
PB	KNP+Manual	0.87	0.82	0.84
PB	Manual	**0.90**	**0.86**	**0.88**
PB	Average	0.89	0.84	0.86
PM	KNP	0.61	58	0.59
PM	KNP+Manual	0.93	**0.94**	0.93
PM	Manual	***0.97**	0.93	**0.95**
PM	Average	0.95	0.94	0.94
PN	KNP	0.88	0.78	0.83
PN	KNP+Manual	***0.93**	**0.90**	**0.92**
PN	Manual	0.89	0.87	0.88
PN	Average	0.91	0.89	0.90

7. まとめ

本稿では, 非専門家の手で高品質なコーパスを作成する手法について調べるため, 固有表現抽出におけるアノテーションを題材として, ふたつの手法について比較を行った. ひとつは既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行う手法 (**KNP+Manual**) であり, もうひとつは人手で一からアノテーションを行う手法 (**Manual**) である. この際, アノテーションにかかる時間, タグの一致率, Gold Standard との比較による正解率の各観点から比較を行っている. また, これに加え, 機械学習における訓練事例としての品質を確かめる

表 15　2 人のアノテータによるアノテーション結果を統合したときの正解率のマクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.31	0.26	0.28
OC	KNP+Manual	0.46	0.47	0.47
OC	Manual	**0.49**	**0.51**	**0.50**
OC	Average	0.48	0.49	0.49
OW	KNP	0.77	0.80	0.79
OW	KNP+Manual	**0.91**	0.91	**0.91**
OW	Manual	0.83	**0.91**	0.87
OW	Average	0.87	0.91	0.89
OY	KNP	0.58	0.44	0.50
OY	KNP+Manual	**0.79**	**0.74**	**0.76**
OY	Manual	0.68	0.65	0.67
OY	Average	0.74	0.70	0.72
PB	KNP	0.66	0.46	0.54
PB	KNP+Manual	0.84	0.78	0.81
PB	Manual	**0.94**	**0.86**	**0.90**
PB	Average	0.89	0.82	0.86
PM	KNP	0.60	0.66	0.63
PM	KNP+Manual	0.86	0.93	0.89
PM	Manual	**0.98**	**0.93**	**0.95**
PM	Average	0.92	0.80	0.92
PN	KNP	0.88	0.78	0.82
PN	KNP+Manual	**0.93**	**0.90**	**0.92**
PN	Manual	0.89	0.86	0.88
PN	Average	0.91	0.88	0.90

ため, それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いたアノテーションも行った. これらの実験の結果から, 全ジャンルのマイクロ平均とマクロ平均で比較した場合, **KNP+Manual** は **Manual** よりもアノテーションにかかる時間が少なく, 一致率や正解率についても高い値になることが明らかになった. 一方で, 新聞から離れたジャンルで同様に比較した場合, **Manual** の方が良い結果となることが明らかになった. これらのことから, 新聞に近いジャンルのテキストについては **KNP+Manual** を, そうでないテキストについては **Manual** を採用するのが良いといえる.

表 16　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマイクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.78	0.68	0.73
KNP+Manual	**0.74**	**0.38**	**0.50**
Manual	0.67	0.29	0.40
Average	0.71	0.33	0.45

表 17　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.47	0.40	0.43
KNP+Manual	**0.40**	**0.24**	**0.30**
Manual	0.31	0.16	0.21
Average	0.36	0.20	0.26

表 18　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマイクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.72	0.48	0.57
OC	KNP+Manual	**0.88**	**0.29**	**0.43**
OC	Manual	0.84	0.20	0.32
OC	Average	0.87	0.24	0.38
OW	KNP	0.79	0.79	0.79
OW	KNP+Manual	***0.74**	***0.53**	**0.62**
OW	Manual	0.55	0.36	0.43
OW	Average	0.65	0.45	0.53
OY	KNP	0.73	0.57	0.64
OY	KNP+Manual	**0.84**	***0.32**	**0.46**
OY	Manual	0.80	0.18	0.30
OY	Average	0.82	0.25	0.38
PB	KNP	0.75	0.60	0.66
PB	KNP+Manual	0.70	**0.31**	**0.43**
PB	Manual	**0.73**	0.28	0.40
PB	Average	0.72	0.29	0.41
PM	KNP	0.61	0.58	0.59
PM	KNP+Manual	**0.55**	**0.19**	**0.29**
PM	Manual	0.52	0.14	0.22
PM	Average	0.54	0.17	0.25
PN	KNP	0.88	0.78	0.83
PN	KNP+Manual	0.76	***0.43**	**0.55**
PN	Manual	**0.78**	0.36	0.49
PN	Average	0.77	0.40	0.52

表 19　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.31	0.26	0.28
OC	KNP+Manual	**0.24**	**0.16**	**0.19**
OC	Manual	0.17	0.12	0.14
OC	Average	0.21	0.14	0.17
OW	KNP	0.77	0.80	0.79
OW	KNP+Manual	**0.72**	**0.57**	**0.63**
OW	Manual	0.63	0.43	0.51
OW	Average	0.67	0.50	0.57
OY	KNP	0.58	0.44	0.50
OY	KNP+Manual	**0.52**	**0.24**	**0.33**
OY	Manual	0.31	0.09	0.14
OY	Average	0.42	0.17	0.24
PB	KNP	0.66	0.46	0.54
PB	KNP+Manual	0.51	**0.24**	0.32
PB	Manual	**0.65**	0.22	**0.32**
PB	Average	0.58	0.23	0.33
PM	KNP	0.60	0.66	0.63
PM	KNP+Manual	**0.55**	**0.29**	**0.38**
PM	Manual	0.53	0.25	0.34
PM	Average	0.54	0.27	0.36
PN	KNP	0.88	0.78	0.82
PN	KNP+Manual	0.75	**0.44**	**0.55**
PN	Manual	**0.78**	0.37	0.50
PN	Average	0.77	0.40	0.53

謝 辞

本研究は文部科学省科学研究費補助金 [若手 B (No.15K16046)] と富士通研究所の助成により行われました. ここに謹んで御礼申し上げます.

また, KNP についての有益な情報を提供してくださった東京工業大学の笹野遼平先生に, この場を借りて御礼申し上げます.

文 献

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng (2008). "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks." *Proceedings of the conference on empirical methods in natural language processing*, pp. 254–263., Association for Computational Linguistics.

Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, and Hiroyuki Shinnou (2016). "Comparison of Annotating Methods for Named Entity Corpora." *LAW X*, pp. 59–67.

Bea Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov (2010). "Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs." *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 29–37., Association for Computational Linguistics.

Lonneke van der Plas, Tanja Samardžić, and Paola Merlo (2010). "Cross-lingual validity of PropBank in the manual annotation of French." *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 113–117., Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a large annotated corpus of English: The Penn Treebank." *Computational linguistics*, 19:2, pp. 313–330.

橋本泰一・乾孝司・村上浩司 (2008).「拡張固有表現タグ付きコーパスの構築」 情報処理学会研究報告, 自然言語処理研究会報告 (NL-188-17), pp. 113–120.

橋本泰一・中村俊一 (2010).「拡張固有表現タグ付きコーパスの構築-白書, 書籍, Yahoo! 知恵袋コアデータ」 言語処理学会第 16 回年次大会発表論文集, 2010, pp. 916–919.

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

徳永健伸・西川仁・岩倉友哉・湯上伸弘 (2015).「固有表現認識課題におけるアノテータの視線分析」 情報処理学会研究報告自然言語処理, 2015:8, pp. 1–8.

Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata (2015). "Named Entity Recognizer Trainable from Partially Annotated Data." *Proceedings of the PACLING 2015*, pp. 10–17.

Satoshi Sekine, and Hitoshi Isahara (2000). "IREX: IR and IE Evaluation project in Japanese." *Proceedings of the 2nd International Conference on Language Resources &*

Evaluation.

Ralph Grishman, and Beth Sundheim (1996). “Message Understanding Conference-6: A Brief History..” *COLING* Vol. 96., pp. 466–471.

岩倉友哉 (2015).「固有表現抽出におけるエラー分析」 言語処理学会第 21 回年次大会 (NLP2015) ワークショップ：自然言語処理におけるエラー分析

平田亜衣・小町守 (2015).「様々なジャンルのテキストに対する固有表現認識の分析」 言語処理学会第 21 回年次大会 (NLP2015) ワークショップ：自然言語処理におけるエラー分析

Masaaki Ichihara, Kanako Komiya, Tomoya Iwakura, and Maiko Yamazaki (2015). “Error Analysis of Named Entity Recognition in BCCWJ.” 言語処理学会第 21 回年次大会 (NLP2015) ワークショップ：自然言語処理におけるエラー分析

Ryohei Sasano, and Sadao Kurohashi (2008). “Japanese Named Entity Recognition Using Structural Natural Language Processing..” *IJCNLP*, pp. 607–612.

Information Retrieval and Extraction Exercise`http://nlp.cs.nyu.edu/irex/NE/df990214.txt` (1999).『ルール、定義』.

Tomoya Iwakura, Ryuichi Tachibana, and Kanako Komiya (2016). “Constructing a Japanese Basic Named Entity Corpus of Various Genres.” *ACL 2016*, pp. 41–46.

『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの分析

宮内 拓也 (国立国語研究所 コーパス開発センター / 東京外国語大学大学院) *
浅原 正幸 (国立国語研究所 コーパス開発センター)
中川 奈津子 (日本学術振興会 特別研究員 PD / 千葉大学大学院)
加藤 祥 (国立国語研究所 コーパス開発センター)

Analysis of Annotation of Information Structure on "The Balanced Corpus of Contemporary Written Japanese"

Takuya Miyauchi (National Institute for Japanese Language and Linguistics / Tokyo University of Foreign Studies)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Natsuko Nakagawa (Japan Society for the Promotion of Science / Chiba University)
Sachi Kato (National Institute for Japanese Language and Linguistics)

要旨

日本語は冠詞のない言語である．ゆえに，日本語から冠詞を持つ言語への翻訳の際には，人によるものでも，機械によるものでも冠詞選択の問題を引き起こすことになる．冠詞選択には，ソースとなる言語における名詞句の情報構造 (定性，特定性など) が影響を与える．本稿では，翻訳における冠詞選択の問題を軽減させるため，『現代日本語書き言葉均衡コーパス』のテキスト (新聞 (PN) コアデータ 16 サンプル) 内の名詞句に対し，情報構造に関係する文法情報のタグをアノテーションした結果を報告する．特に，本稿ではそのアノテーションの概要と基礎統計について述べる．

1. はじめに

冠詞がない言語を母語とする者にとって，冠詞がある言語を習得する際の冠詞選択は難しいものである (Ionin et al. 2004, Tanaka 2013 など)．冠詞選択は，一般に定性 (definiteness) や特定性 (specificity) などの情報構造が大きな影響を与えると考えられる．英語のように定性により定冠詞と不定冠詞を使い分ける言語もあれば，サモア語のように特定性により冠詞を使い分ける言語もある (Mosel and Hovdhaugen 1992)．さらには，コヴェ語(1)のように定性と特定性が共に冠詞選択に影響を与える言語もある (Sato 2013)．

言語処理の分野では英語母語話者が産出した大量のテキストから，英語学習者の冠詞の誤り

* t-miyauchi @ ninjal.ac.jp

(1) コヴェ語は，パプアニューギニアのニューブリテン島で話されるオーストロネシア語族の言語のひとつである．

を検出する手法が提案されている (Nagata et al. 2005)．しかし，日本語母語話者が産出する他言語の冠詞選択を検討する場合，日本語における名詞句の情報構造を考慮する必要がある．さらに機械翻訳において日本語文を冠詞のある言語に翻訳する際にも，日本語の情報構造が問題となってくる．

本稿では，機械翻訳での冠詞選択の問題に関する基礎研究として，『現代日本語書き言葉均衡コーパス』(Maekawa et al. 2014; 以下，BCCWJ) のテキスト内の名詞句に対して情報構造に関わる文法情報のアノテーションを行った結果を報告する．

日本語の情報構造に関する過去のアノテーションは，主としてテキスト中に出現する情報が談話中に既出であること (情報状態 (information status)) を共参照情報として付与するものであった．本研究は，より汎用性を求めるため，情報状態，定性，特定性のみならず，情報構造関係の様々な項目で名詞句にタグを付与した(2)．

2. 関連研究

情報構造のアノテーションには，当該の言語形式の情報構造をどのように決定するかという点で，二つのタイプがある．

まず初めに，当該の言語形式に基づいて情報構造を決める研究がある．例えば，Calhoun et al. (2005) は，Vallduví and Vilkuna (1998) や Steedman (2000) に言及し，韻律を採用した．L+H*LH%(3) の韻律を持つ形式はテーマ (theme)(4)となり，H*L，H*LL%(5) の韻律を持つものはレーマ (rheme)(6)となる．そして，当該の NP が以前に言及されたか否か，またそれが以前述べられた個体から言及可能か否かという点をもとに情報構造をアノテーションした．Hajičová et al. (2000) は語順を用いた情報構造のアノテーションを提案した．この研究は情報構造についてプラハ学派の伝統に触発されたものであり，それゆえに動詞より左にある言語形式をトピックとする．これらのアノテーション基準は言語依存であり，日本語に適用可能なものではない(7)．

二つ目のタイプの研究では，言語学的なテストを採用する．Götze et al. (2007) は言語に依存せず，かつ特定の言語理論にもよらず，情報状態(8) とトピック(9)，フォーカス(10) をアノ

(2) ラベルの詳細については 3. で述べる．
(3) L+H*LH% の韻律では，低めに上昇した後，発話境界の上昇調が現れる (Steedman 2000:645-655)．
(4) 概してトピック (topic) に対応する．
(5) H*L の韻律では，急速な高いピッチから始まりそして下がる．H*LL% の韻律は，H*L に発話末の低い音調が伴ったものである (Steedman 2000:645-655)．
(6) 概してフォーカス (focus) に対応する．
(7) なお，Calhoun et al. (2005) は英語を，Hajičová et al. (2000) はチェコ語を対象にしている．
(8) 情報状態は，旧情報 (given)/補完可能 (accessible)/新情報 (new) の三つが区別される．ある表現が，それ以前の談話で明示的に言及されてる先行詞があれば旧情報であり，関連のあるものが言及されていれば補完可能であり，言及されていなければ新情報である (Götze et al. 2007:§3.2)．
(9) トピックについては，アバウトネストピック (aboutness topic)/フレームセッティングトピック (frame setting topic) の二つが区別される．アバウトネストピックは (1) で説明する．フレームセッティングトピックとは，そのトピックを含む文の述語が解釈されるべきフレームであり，典型的には時間や場所の表現がこれにあたる (Götze et al. 2007:§4.2.3)．
(10) フォーカスに関しては，新情報フォーカス (new-information focus)/対比フォーカス (contrastive focus) の二つが区別される．新情報フォーカスとは，談話を進めるために欠けている新しい情報を提供する文の要素であ

テーションするための基準を策定した．例えば，アバウトネストピックは以下 (1) の手続きで決定される．

(1) An NP X is the aboutness topic of a sentence S containing X if

「ある名詞句 X が，これを含む文 S のアバウトネストピックであるのは以下のときである:」

a. S would be a natural continuation to the announcement *Let me tell you something about X*

「『X について話させて』という予告の後に，S が自然に続き得るとき.」

b. S would be a good answer to the question *What about X*?

「S が，『X についてはどう？』という質問にふさわしい答えであるとき.」

c. S could be naturally transformed into the sentence *Concerning X, S′*, where S′ differs from S only insofar as X has been replaced by a suitable pronoun.

「S が『X に関しては，S′』に自然に変形できるとき．ただし、S′ は S における X が適切な代名詞に置き換わっている点のみで異なる.」

(Götze et al. 2007:165)

本研究は Götze et al. (2007) に沿うものであるが，いくつかの点で彼らの研究とは大きく異なっている．まず，本研究ではトピックとフォーカスを直接アノテーションしない．これはトピックやフォーカスがそれぞれに多次元であるためである (Nakagawa 2016)．実際に Götze et al. (2007:163) では，例えば，指示的 (referential) な NP，特定 (specific) 解釈や総称 (generic) 解釈を持つ不定 (indefinite) の NP など様々な種類のアバウトネストピックが区別されている．定性や特定性のような要因はトピックとは独立であると考え，そのようにアノテーションする方がより単純である．第 2 にトピックやフォーカスと相関すると知られている要因については，定性や特定性以外にも例えば，有生性 (animacy) や動作主性 (agentivity)(Givón 1976, Keenan 1976) など，より多くある．そのため，本研究では情報構造アノテーションの一環としてこれらについてもアノテーションすることとする．

3. タグとアノテーション基準

BCCWJ では，長単位と短単位という二つの単位が採用されているが，本研究では，短単位の名詞をアノテーション対象とする．ただし，複合語については，前部要素には指示性 (referentiality) がないということなどを考慮して，前部要素まで含めて一つの名詞と捉える(11)．

以下の (2) で示す項目についてラベルを設定した．

る．対比フォーカスとは，他の発話に対して対比を呼び起こす文の要素である (Götze et al. 2007:§5.2).

(11) つまり，短単位では 2 語以上の扱いを受けるものに関しても，1 語扱いになっている場合があることになる．これは BCCWJ への共参照アノテーションと同様の方針である．

(2) a. 情報状態 (information status)
b. 定性 (definiteness)
c. 特定性 (specificity)
d. 有生性 (animacy)
e. 有情性 (sentience)
f. 動作主性 (agentivity)
g. 共有性 (commonness)

以下，実例と共にそれぞれのタグのアノテーション基準を示す．

3.1 情報状態

(2a) の情報状態とは，いわゆる旧情報と新情報の区別である．ある談話において，新たな情報は「新情報 (discourse-new)」となり，聞き手が知っている情報は「旧情報 (discourse-old)」となる(12)．一つのテクスト全体を一つの談話と見なし(13)，アノテーションを行った．

(3) a. 担任だった池田弘子先生は違った。
b. スクールカウンセラーでもあった先生の授業は
(読売新聞 [BCCWJ: PN1c_00001])

(3a) の下線部の名詞「池田弘子」はこのテクストで初出の名詞であるため，新情報タグが付与される．一方，(3b) の下線部の名詞「先生」は (3a) の「池田弘子」を指示しているため旧情報タグが付与される．これらの名詞は共参照関係にある．

3.2 定性，特定性

(2b) の定性とは，指示対象を聞き手が同定できるか否かを示すカテゴリーである(14)．指示対象を聞き手が同定できると話し手が想定していれば「定 (definite)」であり，同定できないと想定している場合は「不定 (indefinite)」である．本研究では，スコープとして前後 3 文を見ることとする．

(4) 「そんな薄い$_{a.}$ かばんじゃ$_{b.}$ 遊び道具も入らないよ」
(読売新聞 [BCCWJ: PN1c_00001])

(4) の下線部 *a.* の名詞「かばん」はスコープである (4) の前 3 文以内に既出の名詞であり，ここでは具体的に聞き手の持ち物のかばんを指示している．話し手は当然この「かばん」は聞き手により同定しうると考えていると考えられるため，定のタグが与えられる．(4) の下線部 *b.* の名詞「遊び道具」は特に具体的な何か遊び道具を指示しているわけではないため，不定のタグが付与される．

(12) 情報構造自体についての詳細は Kruijff-Korbayová and Steedman (2003) や Hinterwimmer (2011) などを参照のこと．
(13) 本研究では，BCCWJ の新聞 (PN) のデータを用いたため，一つの記事が一つの談話であると見なしている．
(14) 定性そのものについては，Lyons (1999), Heim (2011) などを参照のこと．

(2c) の特定性は，定性と少々似た概念であるが，話し手が特定の事物を想定しているか否かを示す意味論的カテゴリーである(15)．話し手が特定の事物を想定しているならば「特定 (specific)」となり，想定していなければ「不特定 (unspecific)」となる．定性と同様，特定性に関してもスコープとして前後 3 文を見ることとする．

(5) 行き場を失った$_{a.}$廃タイヤがあぜ道や$_{b.}$納屋の横に放置されてきた。

(北海道新聞 [BCCWJ: PN2e_00001])

(5) の下線部 *a.* の名詞「廃タイヤ」は，北海道鷹栖町に放置された約 30,000 本のタイヤを具体的に指しており，これは (5) の前後 3 文から読み取ることが可能であるため特定のタグが付与される．(5) の下線部 *b.* の名詞「納屋」は特定の納屋が想定されているわけではなく，不特定のタグが与えられる．

3.3 有生性，有情性

(2d) の有生性とは，生きているか否かを示すカテゴリーである．生物（人間，動物など）は「有生 (animate)」であり，無生物（植物を含む）は「無生 (inanimate)」である．有生性は名詞句レベルのみで判断し，付与されるものとする．有生性と似た概念として (2e) の有情性がある．これは，情意があるか否かを示すパラメーターである．自由意志による移動が可能な場合は「有情 (sentient)」となり，自由意志による移動はないなら「非情 (insentient)」となる．日本語については，有生/無性の区別よりも有情/無情の区別が重要であるとする先行研究もあり (三上 1953, 山口 1985 など)，また，有生性と有情性の値が異なる場合もあり得ることから，このパラメーターの設定が必要となる(16)．情意の有無は名詞句単体では判定できない場合があるため，有情性は述語-項レベルまで見た上で判断し，付与されるものとする．

(6) オオクチバスなどの$_{a.}$ブラックバス類が、少なくとも四十三都道府県の七百六十一のため池や$_{b.}$湖沼に侵入し、

(読売新聞 [BCCWJ: PN4c_00001])

(6) の下線部 *a.* の名詞「ブラックバス」は生物であるため，有生のタグが付与される．また，ブラックバスに情意があるか否かは判断が難しいが，その述語は「侵入する」となっており，これは意志的な動作，行為を表しているため，ここでの「ブラックバス」は有情のタグが付与されることになる．(6) の下線部 *b.* の名詞「湖沼」は無生物であり，情意もないと判断されるため，それぞれ，無生，無情のタグが与えられる．

3.4 動作主性

(2f) の動作主性は，事態に関わる人がその事態で果たしている役割を示す．行為を意図的に実現するものは「動作主 (agent)」とし，行為によって変化を被るものを「被動作主 (patient / theme)」とする．このパラメーターについては節レベルまで見て判断し，タグを付与すること

(15) 特定性そのものに関しては，Heusinger (2011) などを参照のこと．

(16) 例えば，「ゾンビ」と「幽霊」は有生性と有情性の値の差により区別できる．ゾンビは腐敗した人体が自発的意思なく徘徊するため，有生，無情である．一方，幽霊とは死者の魂が未練や遺恨により現れたものであるため，無生，有情となる．

とする．その際，主節と従属節の両方を考慮する．また，「どちらでもよい」「どちらでもない」を許す．

(7) a. 編み笠をかぶった人なつっこい笑顔を見るだけで、
b. もみじの木にとまって仲良く寄り添う二羽のキジバト。
c. 独特な雰囲気の写真になりました。
(産経新聞 [BCCWJ: PN1d_00001])

(7a) の下線部の名詞「笑顔」は，主節では被動作主であり従属節では動作主である．このような場合に「どちらでもよい」というタグを付与する．(7b) の下線部の名詞「キジバト」は，それを含む文がこの名詞で終わる体言止めの文であるため，主節では動作主性の判断ができないが，従属節では動作主であるため，「動作主」というタグを付与する．(7c) の下線部の名詞「写真」は動作主でも被動作主でもないため，「どちらでもない」となる．

3.5 共有性

(2g) の共有性は，情報を聞き手が既に知っていると話し手が想定しているか否かを示すパラメターである．聞き手が既に知っていると話し手が想定している情報は「共有 (hearer-old)」であり，知らないと想定している情報は「非共有 (hearer-new)」である．なお，この判断の際はアノテータの世界知識 (world knowledge) を使ってもよいこととし，「想定可能」というラベルも許す．このラベルは，ブリッジング (bridging) を起こしている際に付与される．

(8) $_{a.}$ キャンティ街道を抜け、$_{b.}$ オリーブ畑に囲まれた田園地帯の$_{c.}$ レストランで、
(読売新聞 [BCCWJ: PN4c_00001])

(8) の下線部 *a.* の名詞「キャンティ街道」は，世界遺産にも登録されている，ワインで有名な街道であり，アノテータは既にこの街道について知っていたため，共有のタグが付与された．(8) の下線部 *b.* の名詞「オリーブ畑」は当該の記事からどんなオリーブ畑であるのか判断できないため，非共有のタグが与えられる．(8) の下線部 *c.* の名詞「レストラン」はキャンティ街道のレストランを指しており，ある種のブリッジングを起こしているため，想定可能のタグが付与される．

3.6 その他

固有名詞については，アノテーションの際，有名の度合いを考慮してよいこととし，アノテータの持つ世界知識を参照してもよいとする．(9a) の形式名詞や (9b) のような慣用表現は対象から外し，それぞれ「形式名詞」「慣用表現」タグを付与する．なお，慣用表現であるか否かについてはアノテータによる揺れを許すこととする．

(9) a. 様々な人がいるということが
b. 聞く耳を持たせてくれるんです。
(読売新聞 [BCCWJ: PN1c_00001])

4. 基礎統計

対象は BCCWJ の新聞 (PN) コアデータ 16 サンプルに出現する名詞 2023 件とした．サンプルの選択は BCCWJ-ANNOTATION-ORDER[17] に基づく．作業者は BCCWJ-DepParaPAS (植田ほか 2015, 浅原・大村 2016) に付与された共参照情報を確認しながら作業を行う．定性，特定性，有生性，有情性，動作主性[18]については，与えられた文脈から判断できない場合に「どちらでもよい」というタグを認めた．特定性，動作主性，共有性については，その概念が認めがたい場合に「どちらでもない」というタグを認めた．

表 1 にタグの基礎統計を示す．情報状態のラベルは以前アノテーションされた共参照情報に基づいているが，情報状態の分布と他のラベルの分布は異なっている．よって，この差異より，他のラベルは日本語からの翻訳の際の冠詞選択に影響を与えると考えられる．

表 1　タグの基礎統計

情報状態	新情報	旧情報	-	-
	1345	678	-	-
定性	定	不定	どちらでもよい	-
	1122	899	2	-
特定性	特定	不特定	どちらでもよい	どちらでもない
	1157	749	116	1
有生性	有生	無生	どちらでもよい	-
	342	1680	1	-
有情性	有情	無情	どちらでもよい	-
	337	1678	8	-
動作主性	動作主	被動作主	どちらでもよい	どちらでもない
	192	338	2	1491
共有性	共有	非共有	想定可能	どちらでもない
	1036	494	489	4

表 2 に情報状態と定性の分割表を示す．不定のラベルは新情報のラベルと共に現れることが多いが，興味深いことに定のラベルは新情報，旧情報のどちらのラベルとも現れうるという傾向がある．これは共参照情報の冠詞選択への貢献が限定的であることを示しているといえる．表 3 は情報状態と特定性の分割表である．これについても情報状態と定性のものとほぼ同様の分布を示している．

(17) BCCWJ コアデータサンプルにおけるアノテーション優先順序である．以下参照のこと．https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER

(18) ただし，先に見たように，動作主性については，「どちらでもよい」のタグは主節から見た場合と従属節から見た場合で動作主性の値が異なる場合に付与される．

表 2　情報状態と定性

	新情報	旧情報
定	497	625
不定	846	53
どちらでもよい	2	0

表 3　情報状態と特定性

	新情報	旧情報
特定	531	626
不特定	705	44
どちらでもよい	108	8
どちらでもない	1	0

次に名詞句が含まれる文節の付属語主辞[(19)]と各タグの分布について確認する．付属語主辞は係り受け解析器 CaboCha (Kudo and Matsumoto 2002, 工藤・松本 2002) に含まれる UniDic 主辞規則に基づくものである．付属語主辞と情報状態の分布を表 4 に示す．付属語主辞が「、」や「。」となっているものは，格表示されない連体止めの表現を表している．付属語主語と情報状態の関係を見ると，「が」「を」「に」は新情報が多く，「は」は旧情報が多い傾向にあることがわかる．

表 4　付属語主辞と情報状態

付属語主辞	新情報	旧情報
、	284	168
。	130	82
が	99	47
は	59	62
も	29	10
を	222	53
に	84	28
の	108	86
で	40	13
と	13	12

付属語主辞と定性の分布を表 5 に，付属語主辞と特定性の分布表 6 に示す．表 5，表 6 から，「は」「の」に定性・特定性のものが多い一方，「を」に不定性・不特定性のものが多い傾向にあることがわかる．

5.　おわりに: まとめと今後の課題

本稿では，BCCWJ に対する情報構造のアノテーションデータについて紹介した．本研究では日本語の名詞句に対し，七つの情報構造に関する概念を導入した．これらのアノテーションラベルの分布は，日本語母語話者の冠詞選択や冠詞誤りの修正に対し共参照情報だけでは十分でないことを示しており，新たに導入されたラベルは翻訳における冠詞選択に役立つと思わ

(19)　付属語主辞とは，付属語のうち主要な要素を意味する．

表 5　付属語主辞と定性

付属語主辞	定	不定
、	268	184
。	122	90
が	74	72
は	79	42
も	10	29
を	89	186
に	43	69
の	122	72
で	35	19
と	13	12

表 6　付属語主辞と特定性

付属語主辞	特定	不特定
、	270	171
。	126	77
が	74	55
は	82	33
も	10	25
を	96	155
に	46	54
の	123	61
で	36	12
と	14	8

れる．

今後の課題は以下に示すとおりである．

まず，情報構造をアノテーションする被験者実験を行う．本稿で示したラベルは言語学者によってアノテーションされたものである．被験者実験のためには，言語学の専門知識を持たない人でも回答が可能なように，わかりやすい質問に落とし込む必要がある．非言語学者により情報構造のラベルを判定する質問を作成し，非言語学者にもわかるような言語学的なテストを設計する．それにより，アノテーションの数だけでなく，標的サンプルをも増やすことができる．

第 2 に，ラベルの再検討も行っていくことを考える．名詞の飽和/非飽和 (西山 2003) や Löbner (1985, 2011) の名詞の 4 分類 (Sortal タイプ $\langle e,t\rangle$ / Individual タイプ e / Relational タイプ $\langle e,\langle e,t\rangle\rangle$ / Functional タイプ $\langle e,e\rangle$) を付与することで，名詞の指示性と他のラベルの関係性を解明する手掛かりになると考えられる．さらに，動作主性を動作主/被動作主のみでなく，もう少し詳細に意味役割として，受益者 (beneficiary)，経験者 (experiencer)，着点 (goal)，道具 (instrument)，場所 (location) などを付与することで，意味役割と他のラベルの関係性を突き止めることも考えたい．

第 3 に，機械学習に基づきシソーラスを用いて情報構造の評価モデルを開発する．それにより，日本語テクストに基づき，機械翻訳による冠詞選択の評価を行う．

最後に，本稿で示した情報構造アノテーションを視線計測のデータ (Asahara et al. 2016) と重ね合わせる．これにより，情報構造に対する読み時間のふるまいについて調査を行い (浅原 2017)，情報構造と視線の関係をより深く解明していくことを目指す．

謝　辞

本研究は JSPS 科研費 (課題番号：25284083，研究代表者：浅原正幸) の助成を受けている．

文　献

Tania Ionin, Heejeong Ko, and Kenneth Wexler (2004). “Article semantics in L2 acquisition: The role of specificity.” *Language Acquisition*, 12:1, pp. 3–69.

Junko Tanaka (2013). “A Multivariate Analysis of L2 English Article Use by Articleless L1 Learners.” *Selected Proceedings of the 2011 Second Language Research Forum*, pp. 139–147. Somerville, MA: Cascadilla Press.

Ulike Mosel, and Even Hovdhaugen (1992). *Samoan Reference Grammar*. Oslo: Scandinavian University Press.

Hiroko Sato (2013). “Definiteness and specificity in Kove.” *International workshop on information structure of Austronesian languages*, pp. 37–45. Tokyo: The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Ryo Nagata, Tatsuya Iguchi, Fumito Masui, Atsuo Kawai, and Isu Naoki (2005). “A Statistical Model based on the Three Head Words for Detecting Article Errors.” *IEICE TRANSACTINOS on Information ans Systems*, E88-D:7, pp. 1700–1706.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced corpus of contemporary written Japanese.” *Language Resources and Evaluation*, 48:2, pp. 345–371.

Sasha Calhoun, Malvina Nissim, Mark Steedman, and Jason Brenier (2005). “A framework for annotating information structure in discourse.” *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pp. 45–52. Ann Arbor: The Association for Computational Linguistics.

Enric Vallduví, and M Vilkuna (1998). “On rheme and kontrast.” P. W. Culicover, and L. McNally (Eds.), *The Limits of Syntax*. San Diego: Academic Press. pp. 79–108.

Mark Steedman (2000). “Information structure and the syntax-phonology interface.” *Linguistic Inquiry*, 34, pp. 649–689.

Eva Hajičová, Jarmila Panevová, and Petr Sgall (2000). “A manual for tectogrammatical tagging of the Prague Dependency Treebank.”Technical report, ÚFAL/CKL. (TR-2000-09)

Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel (2007). “Information structure.” Stefanie Dipper, Michael Götze, and Stavros Skopeteas (Eds.), *Information structure in cross-linguistic corpora: annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Vol. 7.: Universitätsverlag Potsdam. pp. 147–187.

Natsuko Nakagawa (2016). “Information structure in spoken Japanese: Particles, word

order, and intonation." Unpublished doctoral dissertation, Kyoto University.

Talmy Givón (1976). "Topic, pronoun, and grammatical agreement." Charles N. Li (Ed.), *Subject and Topic*. New York: Academic Press. pp. 149–187.

Edward L. Keenan (1976). "Towards a universal definition of "subject"." Charles N. Li (Ed.), *Subject and Topic*. New York: Academic Press. pp. 303–334.

Ivana Kruijff-Korbayová, and Mark Steedman (2003). "Discourse and Information Structure." *Journal of Logic, Language and Information*, 12:3, pp. 249–259.

Stefan Hinterwimmer (2011). "Information structure and truth-conditional semantics." Klaus von Heusinger, Claudia Maienborn, and Paul Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*. Vol. 2.: Mouton de Gruyter. pp. 1875–1908.

Christopher Lyons (1999). *Definiteness*. Cambridge: Cambridge University Press.

Irene Heim (2011). "Definiteness and indefiniteness." Klaus von Heusinger, Claudia Maienborn, and Paul Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*. Vol. 2.: Mouton de Gruyter. pp. 996–1025.

Klaus von Heusinger (2011). "Specificity." Klaus von Heusinger, Claudia Maienborn, and Paul Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*. Vol. 2.: Mouton de Gruyter. pp. 1058–1087.

三上章 (1953).『現代語法序説』 刀江書院, 東京.

山口光 (1985).「存在文と所有文」 金田一春彦・林大・柴田武 (編)『日本語大辞典』 大修館書店, 東京 pp. 198–200.

植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸 (2015).「『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション」 第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214.

浅原正幸・大村舞 (2016).「BCCWJ-DepParaPAS:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化」 言語処理学会第 22 回年次大会発表論文集, pp. 489–492.

Taku Kudo, and Yuji Matsumoto (2002). "Japanese Dependency Analysis using Cascaded Chunking." *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69.

工藤拓・松本裕治 (2002).「チャンキングの段階適用による日本語係り受け解析」 情報処理学会論文誌, 43:6, pp. 1834–1842.

西山佑司 (2003).『日本語名詞句の意味論と語用論: 指示的名詞句と非指示的名詞句』 ひつじ書房, 東京.

Sebastian Löbner (1985). "Definites." *Journal of Semantics*, 4, pp. 279–326.

Sebastian Löbner (2011). "Concept types and determination." *Journal of Semantics*, 28, pp. 279–333.

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto (2016). "Reading-Time An-

notations for Balanced Corpus of Contemporary Written Japanese." *Proceedings of COLING-2016*, pp. 684–694.

浅原正幸 (2017).「読み時間と情報構造について（ちょっとながめ）」　言語資源活用ワークショップ 2016 発表論文集．M.s.

読み時間と情報構造について（ちょっとながめ）

浅原 正幸 (国立国語研究所コーパス開発センター) *

Between Reading Time and Information Structure (longer version)

Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

本研究では『現代日本語書き言葉均衡コーパス』に対して付与された，文の読み時間データ『BCCWJ-EyeTrack』と，名詞句の定性などの情報構造アノテーションデータの対照分析を行った。日本語母語話者 24 人分のデータを線形混合モデルにより分析した結果，特定性 (specificity)・有情性 (sentience)・共有性 (commonness) が文の読み時間に影響を与え，それぞれ異なったパターンの読み時間の遅延を引き起こすことがわかった。特に共有性においては新情報 (hearer-new)・想定可能 (bridging) が識別可能なレベルで異なった。このことは，ある名詞句が言語受容者にとって新情報なのか想定可能なのかを読み時間データから推定することができる可能性を示唆しており，文書要約のユーザ適応などの応用に利用することが期待できる。

1. はじめに

情報構造は和文翻訳時の冠詞選択や文書要約において重要であるが，言語処理の分野ではあまり研究されてこなかった。

情報構造のうち定性 (definiteness) や特定性 (specificity) は，言語によっていずれかもしくは両方が名詞句の冠詞の選択に影響を与える。しかし，日本語をはじめとする冠詞のない言語の母語話者が産出する名詞句に対しては，それがどのような情報構造なのかを明らかにする必要がある。さらに，冠詞のない言語を，人間もしくは機械翻訳により，冠詞をもつ言語に翻訳する場合には，原言語の名詞句の情報構造を考慮する必要がある。

情報構造のほかの観点として，情報状態 (information status) と共有性 (commonnness) がある。情報状態は共参照関係に関連する概念で，先行文脈に出現しているか (`discourse-old`) 否か (`discourse-new`) を表すものである。一方，共有性は言語生産者側が想定する「言語受容者側がその情報を知っているか (`hearer-old`)，推定可能か (`bridging`)，そうではないか (`hearer-new`)」といった観点である。このうち，テキスト中の名詞句に対する推定可能性 (`bridging`) について検証する Bridging reference (Asher and Lascarides 1998) は共参照関係解析の分野で難しいとされている。書かれたテキストの表層情報のみで難しく，(Hou et al. 2013) の研究においては，タグ付けコーパスと世界知識を用いて bridging reference を解こう

* masayu-a@ninjal.ac.jp

としている。このように，完全な世界知識を用いて完璧な文書理解を行うことも重要ではあるが，言語受容者側が彼らの知識のみでは推定不可能である新情報であることを検出することも，利用者志向 (user-oriented) の情報抽出や文書要約には必要であると考える。

一般に，統語・意味情報処理が完全に行われたとしても，テキストの表層のみから機械学習に基づき情報構造の推定を行うことは困難であろう。一方，人間の文処理機構において，情報構造が文処理速度の促進・阻害の双方に影響を与えられることが考えられる。本稿では，**人間の読み時間に基づく情報構造の推定手法を確立するため**に，情報構造が人間の文処理時間のどのように影響を与えるかについて分析する。

日本人母語話者の読み時間データとして『現代日本語書き言葉均衡コーパス』(以下 BCCWJ)(Maekawa et al. 2014) に対して読み時間を付与した BCCWJ-EyeTrack (Asahara et al. 2016) を用いる。同データに対して，情報構造アノテーション (宮内ほか 2017) を重ね合わせたうえで，線形混合モデルに基づく統計分析を行う。

結果，いくつかの情報構造について，読み時間の測定値に差が出ることを発見したので報告する。

2. 関連研究

最初に読み時間付与コーパス関連の先行研究について示す。*Dundee Eyetracking Corpus* (Kennedy and Pynte 2005) は英語とフランス語の新聞社説をそれぞれ 10 人の母語話者の読み時間を視線走査装置を用いて記録したコーパスである。*Dundee Eyetracking Corpus* は特定の言語現象を仮定したものではなく，テキストに出現する自然な言語現象の中から探索的に研究することを目的として構築されている。例えば，Demberg and Keller (2008) は Gibson (2008) の Dependency Locality Theory や Hale (2001) の Surprisal Theory を検証している。このようなコーパスの存在は，被験者実験と統計的検証の分業にしたほか，統計的検証の再現可能性を可能にしており，Roland et al. (2012) は先行研究 (Demberg and Keller 2007) の結果がいくつかの外れ値により歪められたものであることを指摘している。

次に，情報構造関連の先行研究について示す。Götze et al. (2007) は言語理論と独立した情報構造アノテーション基準として，情報状態 (information status: given/accessible/new)・話題 (topic: aboutness topic/frame setting topic) ・焦点 (focus new-information focus/contrastive focus) に対するタグ集合を定義している。また，Prasad et al. (2015) は Discourse Treebank (Miltsakaki et al. 2004) や PropBank(Palmer et al. 2005) に対するブリッジングのアノテーション基準について議論している。

最後に視線情報や読み時間に基づく言語分析や解析モデルについて示す。Barrett et al. (2016) は視線のパターンに基づいた品詞タグ付け手法について提案している。Klerke et al. (2015) は機械処理したテキストの文法性判断を視線情報に基づきモデル化する手法を提案している。Iida et al. (2013) は述語項構造アノテーションの作業者の視線走査情報を分析している。

我々の研究はこれらの先行研究と異なる立場をとる。均衡コーパスに対して，自然な設定での視線情報と，言語の専門知識に基づくアノテーションを重ね合わせることで，心理言語学的

表 1　Data format

列名	データ型	摘要
surface	factor	出現書字形
time	int	読み時間
logtime	num	読み時間 (常用対数)
measure	factor	読み時間の種類
sample	factor	サンプル名
article	factor	記事情報
metadata_orig	factor	文書構造タグ
metadata	factor	メタデータ
length	int	文字数
space	factor	文節境界空白の有無
subj	factor	実験協力者 ID
setorder	factor	文節境界空白の呈示順
dependent	int	係り受け関係
sessionN	int	セッション順
articleN	int	記事呈示順
screenN	int	画面呈示順
lineN	int	行呈示順
segmentN	int	文節呈示順
is_first	factor	最左要素
is_last	factor	最右要素
is_second_last	factor	右から 2 つ目の要素
infostatus	factor	情報状態
definite	factor	定性
specificity	factor	特定性
animacy	factor	有生性
sentience	factor	有情性
agentivity	factor	動作主性
commonness	factor	共有

な統計的分析を行い，言語受容者の視線情報を含む読み時間と情報構造との対照分析を行う。

3. 分析手法

分析データとして BCCWJ-EyeTrack (Asahara et al. 2016) に情報構造アノテーション (宮内ほか 2017) を重ね合わせたもの (表 1) を用いる。以下，データの詳細について説明する。

3.1 読み時間データ

自己ペース読文法は，他の文節をマスクしたうえで 1 文節単位に逐次的に呈示する読み時間測定手法である。読み戻しができないために，文節単位の読み時間がそのままデータとなる (SELF)。視線走査法で取得したオリジナルのデータから文字の半角単位に Start Fixation Time （注視開始時刻）と End Fixation Time （注視終了時刻）と Fixation Time （注視時間）を得る。このデータを国語研文節単位でグループ化しなおした注視順データを集計して，テキスト生起順データに加工する。テキスト生起順データは以下の 5 種類からなる。

- First Fixation Time (FFT)
- First-Pass Time (FPT)

表 2　Parameters of the linear mixed model for the self paced reading time (SELF) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.893	0.062	46.51
length	0.102	0.002	**42.31**
space=T	0.003	0.004	0.86
dependent	-0.005	0.003	-1.61
sessionN	-0.021	0.022	-0.94
articleN	-0.023	0.007	**-3.23**
screenN	-0.032	0.002	**-11.19**
lineN	-0.014	0.002	**-6.10**
segmentN	-0.005	0.001	**-4.83**
is_first=T	0.047	0.006	**7.19**
is_last=T	0.040	0.008	**4.71**
is_second_last=T	-0.011	0.005	**-2.11**
space=T:sessionN	-0.019	0.044	-0.43
is=discourse-old	-0.005	0.005	-0.98
def=indefinite	0.004	0.015	0.30
spec=specific	0.044	0.016	**2.78**
spec=unspecific	0.001	0.010	0.16
ani=inanimate	-0.000	0.050	-0.02
sent=insentient	-0.105	0.067	-1.56
sent=sentient	-0.098	0.050	-1.94
ag=both	-0.058	0.049	-1.18
ag=neither	-0.004	0.007	-0.69
ag=patient	-0.013	0.008	-1.63
com=hearer-new	0.025	0.007	**3.59**
com=hearer-old	-0.020	0.009	**-2.11**
com=neither	0.000	0.025	0.01

45 data points (0.69%) were excluded in the 3-SD trimming.

- Regression Path Time (RPT)
- Second-Pass Time (SPT)
- Total Time (TOTAL)

これらの読み時間情報 (time, logtime) に対して，出現書字形 (surface)・記事情報 (sample, article)・文書構造 (metadata_orig, metadata) のほか，出現書字形文字数 (length)，文節単位の空白の有無 (space)，実験協力者 ID (subj)，係る文節数 (dependent)，実験協力者ごとの呈示順序 (sessionN, setorder, articleN, screenN, lineN, segmentN)，画面水平方向の位置 (is_first,is_last,is_second_first) を付与したデータを分析に用いる。係る文節数は BCCWJ-DepPara (Asahara and Matsumoto 2016) のものを用いる。

本研究では日本語母語話者 24 人分のデータを統計分析に用いる。データの詳細については (Asahara et al. 2016) を参照されたい。

3.2　情報構造アノテーション

情報構造は，BCCWJ の短単位について以下の情報を付与したものを用いる。基準の詳細については (宮内ほか 2017) を参照されたい。

表 3　Parameters of the linear mixed model for the first fixation time (FFT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.151	0.078	27.32
length	-0.011	0.003	**-3.55**
space=T	-0.003	0.006	-0.61
dependent	-0.002	0.004	-0.56
sessionN	-0.020	0.016	-1.26
articleN	-0.009	0.004	-1.94
screenN	-0.008	0.003	**-2.14**
lineN	-0.014	0.003	**-4.61**
segmentN	0.002	0.001	1.65
is_first=T	-0.000	0.009	-0.03
is_last=T	-0.002	0.012	-0.17
is_second_last=T	-0.010	0.008	-1.25
space=T:sessionN	0.040	0.032	1.25
is=discourse-old	-0.003	0.008	-0.42
def=indefinite	0.021	0.020	1.01
spec=specific	0.039	0.022	1.79
spec=unspecific	0.017	0.014	1.22
ani=inanimate	0.119	0.083	1.43
sent=insentient	-0.005	0.103	-0.05
sent=sentient	0.092	0.068	1.35
ag=both	0.060	0.069	0.87
ag=neither	0.009	0.009	0.93
ag=patient	-0.001	0.012	-0.14
com=hearer-new	0.003	0.009	0.38
com=hearer-old	0.016	0.013	1.25
com=neither	0.019	0.034	0.58

2 data points (0.03%) were excluded in the 3-SD trimming.

- 情報状態 (information status:speaker-new)
  談話中に同一指示名詞句が出現した (discourse-old) か否 (discourse-new) か。既存の共参照情報ラベルを見ながら判定する。想定可能 (bridging) は言語受容者側の判断として，共有性に委ねる。
- 定性 (definiteness:hearer-identify)
  言語受容者が外延の示す実体を認識できる (definite) か否 (indefinite) か。
- 特定性 (specificity:speaker-identify)
  言語生産者が外延の示す実体を認識できる (specific) か否 (inspecific) か。
- 有生性 (animacy)
  名詞句が指示しているものが生きているか (animate) か否 (inanimate) かを述語を見ないで判定する。
- 有情性 (sentience)
  名詞句が指示しているものが自由意志を持つ (sentient) か否 (insentient) かを述語-項の対を見て判定する。

表 4　Parameters of the linear mixed model for the first pass time (FPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.303	0.102	22.53
length	0.144	0.004	**33.61**
space=T	-0.032	0.007	**-4.23**
dependent	-0.005	0.006	-0.89
sessionN	-0.041	0.028	-1.46
articleN	-0.001	0.009	-0.19
screenN	-0.023	0.005	**-4.76**
lineN	-0.018	0.004	**-4.64**
segmentN	-0.008	0.002	**-4.07**
is_first=T	0.068	0.011	**5.94**
is_last=T	0.021	0.015	1.40
is_second_last=T	0.028	0.010	**2.84**
space=T:sessionN	0.062	0.056	1.11
is=discourse-old	0.005	0.010	0.50
def=indefinite	0.024	0.026	0.90
spec=specific	0.064	0.028	**2.26**
spec=unspecific	0.031	0.018	1.70
ani=inanimate	0.210	0.104	**2.01**
sent=insentient	-0.001	0.129	-0.01
sent=sentient	0.194	0.086	**2.25**
ag=both	-0.050	0.087	-0.57
ag=neither	0.014	0.012	1.19
ag=patient	-0.006	0.015	-0.43
com=hearer-new	0.024	0.012	1.95
com=hearer-old	0.000	0.017	-0.03
com=neither	0.002	0.043	0.05

13 data points (0.24%) were excluded in the 3-SD trimming.

- 動作主性 (agentivity)
  節レベルで動作主 (agent)・被動作主 (patient) になるかを判定する。従属節側と主節側の両方で検討するため，どちらも可 (both) も許す。
- 共有性 (commonness:hearer-new)
  共有性は，言語需要者側が既知であると，言語生産者が想定している (hearer-old) か否 (hearer-new) かを判定する。談話上，世界知識を利用して想定可能 (bridging) である場合を許す。

読み時間の分析が文節単位であるために，文節内の最右要素の情報構造を文節を代表する情報構造として用いることとした。

3.3　統計処理手法

まず，対象は情報構造が付与されている名詞句のみとする。データの前処理として，metadata が {authorsData, caption, listItem, profile, titleBlock} のものを除外した。さらに視線走査実験結果の 0 (fixation がない対象) のデータポイントを除外した。この時点でのデータポイント数は SELF が 6444 件，FFT・FPT・RPT・TOTAL が 5268 件，SPT が 2081 件である。

表 5　Parameters of the linear mixed model for the second pass time (SPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.318	0.164	14.099
length	0.015	0.007	**2.101**
space=T	-0.043	0.013	**-3.099**
dependent	0.001	0.011	0.175
sessionN	-0.030	0.019	-1.556
articleN	-0.005	0.008	-0.635
screenN	-0.020	0.008	**-2.497**
lineN	-0.013	0.007	-1.821
segmentN	-0.009	0.003	**-2.564**
is_first=T	-0.019	0.019	-1.013
is_last=T	-0.061	0.029	**-2.103**
is_second_last=T	0.015	0.017	0.864
space=T:sessionN	0.052	0.037	1.380
is=discourse-old	0.023	0.018	1.279
def=indefinite	0.022	0.046	0.471
spec=specific	0.000	0.049	0.010
spec=unspecific	-0.026	0.032	-0.800
ani=inanimate	-0.077	0.217	-0.355
sent=insentient	0.243	0.252	0.965
sent=sentient	0.126	0.144	0.880
ag=both	-0.118	0.156	-0.757
ag=neither	-0.023	0.021	-1.083
ag=patient	-0.045	0.027	-1.655
com=hearer-new	0.037	0.023	1.616
com=hearer-old	-0.013	0.030	-0.455
com=neither	-0.050	0.066	-0.765

1 data point (0.04%) was excluded in the 3-SD trimming.

分析は常用対数時間に対して線形混合モデルに基づいて行い，最初に一度モデル化したうえで，標準偏差 ± 3.0 を超えるデータポイントを除外した。subj と article をランダム切片として，次のような式に基づき分析を行った。なお，ランダム切片に対する係数の組み合わせによるモデル選択は行っていない。

```
logtime ~ space * sessionN + length + dependent
  + is_first + is_last + is_second_last
  + articleN + screenN + lineN + segmentN
  + infostatus + definite + specificity + animacy
  + sentience + agentivity + commonness
  + (1 | subj) + (1 | article)
```

4. 結果

自己ペース読文法 (SELF) と視線走査法 (FFT,FPT,SPT,RPT,TOTAL) の結果を表 2,3,4,5,6,7 に示す。

ここでは情報構造以外の傾向について確認しておく。文字数 (length) が多くなれば読み時間が長くなる傾向，および，実験が進むにつれて読み時間が短くなる傾向 (articleN, screenN, lineN, segmentN) がある。さらにレイアウト上，最左文節 (is_first)・最右文

表 6　Parameters of the linear mixed model for the regression path time (RPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.188	0.118	18.48
length	0.120	0.004	**24.79**
space=T	-0.021	0.008	**-2.47**
dependent	0.001	0.006	0.18
sessionN	-0.048	0.028	-1.67
articleN	0.001	0.007	0.14
screenN	-0.014	0.005	**-2.50**
lineN	-0.012	0.004	**-2.69**
segmentN	-0.014	0.002	**-5.91**
is_first=T	0.026	0.013	**2.00**
is_last=T	0.063	0.017	**3.59**
is_second_last=T	0.030	0.011	**2.65**
space=T:sessionN	0.065	0.057	1.13
is=discourse-old	-0.003	0.011	-0.30
def=indefinite	0.041	0.030	1.34
spec=specific	0.095	0.032	**2.95**
spec=unspecific	0.038	0.020	1.82
ani=inanimate	0.112	0.119	0.94
sent=insentient	0.248	0.150	1.65
sent=sentient	0.345	0.102	**3.37**
ag=both	-0.054	0.100	-0.54
ag=neither	0.013	0.014	0.91
ag=patient	-0.000	0.017	-0.01
com=hearer-new	0.001	0.014	0.09
com=hearer-old	-0.018	0.019	-0.94
com=neither	0.042	0.049	0.86

43 data points (0.81%) were excluded in the 3-SD trimming.

節 (is_last)・右から 2 番目の文節 (is_second_first) で読み時間が長くなる傾向がある。視線走査法においては，文節単位に半角空白を置いたほうが読み時間が短くなる傾向がある。一方，係る文節の数 (dependent) の効果は，対象を名詞句のみとした結果，有意差がなかった。

5. 考察

表 8 に結果のまとめを示す。0 は読み時間に有意差がなかったものである。+ は読み時間が増える傾向にあり，- は読み時間が減る傾向にある固定要因である（いずれも有意差あり: $1.96 < |\text{t value}|$）。

定性 (definite) は読み時間に影響を与えることがない一方，特定性 (specificity) が，自己ペース読文法 (SELF) と視線走査法 (FPT, RPT, TOTAL) の読み時間を長くする効果があることがわかった。

また，有生性 (animacy) は視線走査法 (FPT) の読み時間を長くし，有情性 (sentience) は，視線走査法 (FPT, RPT) の読み時間を長く効果があることがわかった。日本語の文処理においては，他言語で言及される有生性よりも，日本語学で言及される有情性のほうが読み時間に影響を与えやすいことがデータから読み取れる。

表 7　Parameters of the linear mixed model for the total time (TOTAL) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.500	0.105	23.69
length	0.135	0.004	**30.44**
space=T	-0.043	0.007	**-5.51**
dependent	-0.001	0.006	-0.30
sessionN	-0.050	0.027	-1.82
articleN	-0.000	0.009	-0.09
screenN	-0.037	0.005	**-7.17**
lineN	-0.020	0.004	**-4.84**
segmentN	-0.015	0.002	**-7.27**
is_first=T	0.061	0.011	**5.16**
is_last=T	-0.007	0.016	-0.49
is_second_last=T	0.027	0.010	**2.62**
space=T:sessionN	0.062	0.054	1.14
is=discourse-old	0.009	0.010	0.89
def=indefinite	0.036	0.027	1.32
spec=specific	0.070	0.029	**2.39**
spec=unspecific	0.016	0.019	0.88
ani=inanimate	0.177	0.108	1.63
sent=insentient	-0.027	0.133	-0.20
sent=sentient	0.130	0.089	1.46
ag=both	-0.025	0.091	-0.28
ag=neither	0.006	0.013	0.50
ag=patient	-0.011	0.015	-0.70
com=hearer-new	0.030	0.013	**2.34**
com=hearer-old	-0.000	0.017	-0.02
com=neither	0.033	0.045	0.74

5 data points (0.09%) were excluded in the 3-SD trimming.

さらに，共参照情報から得られる情報状態 (infostatus) は読み時間に影響を与えることがない一方，共有性 (commonness) については，言語受容者にとっての新情報 (hearer-new) が想定可能な要素 (bridging) に対して，有意に読み時間を長くする効果が自己ペース読文法と視線走査法 (TOTAL) に見られた一方，言語受容者にとっての旧情報 (hearer-new) が想定可能な要素 (bridging) に対して，有意に読み時間を短くする効果が自己ペース読文法に見られた。

最後に動作主性の差異は読み時間に影響を与えなかった。

このことから読み時間の差異により，特定性・(有生性・) 有情性・共有性について推定できる可能性があることがわかった。さらに，それぞれ読み時間の差異の出現傾向に違いがあることから，読み時間の測定値の組み合わせにより，各情報構造が推定できる可能性がある。

6. おわりに

本研究では，読み時間と情報構造の対照分析を行い，名詞句の特定性・有生性・有情性・共有性について，読み時間の差異が出現することを確認した。

本研究では『現代日本語書き言葉均衡コーパス』に対して付与された，文の読み時間データ『BCCWJ-EyeTrack』と，名詞句の定性などの情報構造アノテーションデータの対照分析

表 8　Summary: reading time and information structures

Fixed Effect		SELF	FFT	FPT	SPT	RPT	TOTAL
infostatus=discourse-old	(vs. discourse-new)	0	0	0	0	0	0
definite=indefinite	(vs. definite)	0	0	0	0	0	0
specificity=specific	(vs. either)	+	0	+	0	+	+
specificity=unspecific	(vs. either)	0	0	0	0	0	0
animacy=inanimate	(vs. animate)	0	0	+	0	0	0
sentience=insentient	(vs. either)	0	0	0	0	0	0
sentience=sentient	(vs. either)	0	0	+	0	+	0
agentivity=both	(vs. agent)	0	0	0	0	0	0
agentivity=neither	(vs. agent)	0	0	0	0	0	0
agentivity=patient	(vs. agent)	0	0	0	0	0	0
commonness=hearer-new	(vs. bridging)	+	0	0	0	0	+
commonness=hearer-old	(vs. bridging)	-	0	0	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0	0	0

を行った。日本語母語話者 24 人分のデータを線形混合モデルにより分析した結果，特定性 (specificity)・有情性 (sentience)・共有性 (commonness) が文の読み時間に影響を与え，それぞれ異なったパターンの読み時間の短縮・延長を引き起こすことがわかった。特に共有性においては新情報 (`hearer-new`)・想定可能 (`bridging`) が識別可能なレベルで異なった。このことは，ある名詞句が言語受容者にとって新情報なのか想定可能なのかを読み時間データから推定することができる可能性を示唆しており，文書要約のユーザ適応などの応用に利用することが期待できる。

謝　辞

本研究は JSPS 科研費 基盤 (B) 25284083 「言語コーパスに対する読文時間付与とその利用」の助成を受けました。

文　献

N. Asher, and A. Lascarides (1998). “Bridging.” *Journal of Semantics*, 15:1, pp. 83–113.

Yufang Hou, Katja Markert, and Michael Strube (2013). “Global Inference for Bridging Anaphora Resolution.” *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 907–917. Atlanta, Georgia: Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto (2016). “Reading-Time Annotations for ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tech-*

nical Papers, pp. 684–694.

宮内拓也・浅原正幸・中川奈津子・加藤祥 (2017).「『現代日本語書き言葉均衡コーパス』に対する情報構造アノテーションの構築」　言語処理学会第 23 回年次大会発表論文集.

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

V. Demberg, and F. Keller (2008). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity." *Cognition*, 109:2, pp. 193–210.

E. Gibson (2008). "Linguistic complexity: Locality of syntactic dependencies." *Cognition*, 68, pp. 1–76.

J. Hale (2001). "A probabilistic earley parser as a psycholinguistic model." *Proceedings of the second conference of the North American chapter of the association for computational linguistics* Vol. 2., pp. 159–166.

D. Roland, G. Mauner, C. O'Meara, and H. Yun (2012). "Discourse expectations and relative clause processing." *Journal of Memory and Language*, 66:3, pp. 479–508.

V. Demberg, and F. Keller (2007). "Eye-tracking evidence for integration cost effects in corpus data." *Proceedings of the 29th meeting of the cognitive science society (CogSci-07)*, pp. 947–952.

Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel (2007). "Information structure." Stefanie Dipper, Michael Götze, and Stavros Skopeteas (Eds.), *Information structure in cross-linguistic corpora: annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Vol. 7.: Universitätsverlag Potsdam. pp. 147–187.

Rashmi Prasad, Bonnie Webber, Alan Lee, Sameer Pradhan, and Aravind Joshi (2015). "Bridging Sentential and Discourse-level Semantics through Clausal Adjuncts." *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pp. 64–69.

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber (2004). "The Penn Discourse Treebank.." *LREC*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles." *Conputational Linguistics*, 31:1, pp. 71–105.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard (2016). "Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 579–584.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard (2015). "Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences." *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*,

pp. 97–105.

Ryu Iida, Koh Mitsuda, and Takenobu Tokunaga (2013). “Investigation of annotator’s behaviour using eye-tracking data.” *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 214–222.

Masayuki Asahara, and Yuji Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annotation Treebank on the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58.

編集後記

国立国語研究所に着任して5年が経ちました。着任時に依頼された仕事「超大規模コーパスプロジェクト」（2011年度～2015年度）はウェブコーパスを構築せよというものでした。100億語という数値目標が設定されていましたが、今回のワークショップでお披露目することができました。大きなトラブルもなくプロジェクトを進めることができたのは、多くの方のお力添えのおかげです。国立国語研究所コーパス開発センターでは並行して「コーパス日本語学の創成」と「コーパスアノテーションの基礎研究」のプロジェクトが進められ、その枠組の中で「コーパス日本語学ワークショップ」（年2回全8回）が開催されました。

現在、国立国語研究所コーパス開発センターでは「包括的高度検索環境の整備」（2016年度～2021年度）と題し、『中納言』『梵天』などの検索系をより使いやすくするプロジェクトを進めております。前プロジェクトの「コーパス日本語学ワークショップ」をより発展させた形の「言語資源活用ワークショップ」を今回より年1回全6回開催する予定です。2016年度のみ先に人事を進めたために年度末の3月に開催いたしましたが、2017年度～2020年度は9月に、2021年度は8月に実施する予定です。

前ワークショップはコーパス日本語学を新たに作るということを目標にしておりましたが、本ワークショップはより高度なコーパス利用方法を目標にしております。現代日本語（独話・対話・書き言葉）のみならず、方言・古典語・L1学習者の日本語・L2学習者の日本語などが対象になります。コーパスで扱う上でさまざまなことを学ばなくてはなりません。コーパス整形・統計処理などについても議論する場にしていきたいと考えております。

実際のワークショップでは、2件の招待講演と45件の一般発表がありました。招待講演をご快諾くださりました京都大学の秋田祐哉先生・電気通信大学の松吉俊先生、広報不足にもかかわらず一般発表に申込をしてくださった方々に感謝いたします。また「語彙資源活用シンポジウム」をワークショップに併設して企画いたしました。形態素解析用辞書・シソーラスなどの電子化辞書編纂者と紙媒体で出版されてきた辞書編纂者とが熱い議論をとりかわしました。次回以降もテーマを決めてシンポジウムを併催していきたいと思います。ワークショップ・シンポジウム期間中、朝倉書店・くろしお出版・ひつじ書房・勉誠出版に会場にて出店をお願いしましたところご快諾くださりました。どうもありがとうございました。

今回よりワークショップに寄せられた予稿を発表論文と呼び方を変え、国立国語研究所学術情報リポジトリ https://repository.ninjal.ac.jp/ にてDOI（Digital Object Identifier）付きで公開することになりました。ご利用いただければ幸いです。

最初のウェブコーパスの話に戻しますと、検索系『梵天』の2016年9月の試験公開・2017年3月の一般公開以降、一般公開版は24万件の利用が、高機能版は 37000件の利用がありました。『梵天』の講習会の実施も35回に及び、内2回については Youtube Live により実施しました。ウェブコーパス NWJC2vec の頒布数も70件を超えて、利用が増えていることをうれしく思います。今後ウェブコーパスを用いた発表が増えていくことを楽しみにしております。

国立国語研究所
コーパス開発センター
浅原　正幸