

# 国立国語研究所学術情報リポジトリ

## <全文>言語資源ワークショップ2023発表論文集

メタデータ	言語: ja 出版者: 公開日: 2023-11-24 キーワード (Ja): キーワード (En): 作成者: 国立国語研究所 言語資源開発センター メールアドレス: 所属:
URL	<a href="https://doi.org/10.15084/0002000110">https://doi.org/10.15084/0002000110</a>

# 言語資源ワークショップ 2023

発表論文集

2023年8月28・29日(月・火)

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 言語資源開発センター 編

## 目次

Programme .....	1
発表論文集 .....	7
● 学習者コーパス研究における横断・縦断データ統合の意義—I-JAS と B-JAS をめぐって— 石川慎一郎 (神戸大学) .....	8
● 介護分野における専門用語の平易化に向けた語彙リストの構築 黄海洪 (京都大学大学院), 金丸敏幸 (京都大学) .....	26
● 日中対訳コーパスの構築と公開に向けて 宮本華瑠 (大阪大学) .....	40
● 近代中国語関係書における主体移動表現の日中対照研究 金敬玲 (國學院大學大学院) .....	52
● 説明的文章の要点把握のための読解方法の有効性評価に用いる正解文データセット (CAKeS) の作成 渡邊幸佑 (兵庫教育大学) .....	68
● 親疎関係で見る上昇下降調の使用率—『日本語日常会話コーパス』を用いて— 李海琪 (浙江大学) .....	75
● 論文指導の場面における終助詞「かな」の配慮機能—『BTSJ 日本語自然会話コーパス』の データから— 劉悦 (筑波大学) .....	83
● 程度副詞使用実態の横断的・縦断的調査—「通時話し言葉コーパス」の試み— 日暮康晴 (筑波大学大学院) .....	95
● 歩きスマホのコロケーションと意味—新聞データベースを用いた使用実態調査— 王鑫 (筑波大学) .....	109
● 国語科教科書の説明的文章における疑問-解答の段落構成—各学年テキストの比較から— 尾崎誉治 (東北大学) .....	122
● アーカイブデータを利用した言語研究とその応用可能性 鈴木成典 (国際基督教大学/日本学術振興会), 鎌野慈人 (ストーニーブルック大学), 坂本誓 (国際基督教大学), 鎌倉欧亮 (国際基督教大学), 李勝勲 (国際基督教大学/ヴェンダ大学), 閔 宇 (立命館大学), パーキンズ ジェレミー (会津大学), 五十嵐陽介 (国立国語研究所) ....	136

● 書籍の文体と修辞機能の分析のパイロットスタディ 田中弥生 (国立国語研究所), 柏野和佳子 (国立国語研究所), 加藤祥 (目白大学) .....	142
● ゲームコーパスの設計方針と構築方法 麻子軒 (関西大学) .....	151
● 読書感想文の分析と指導法—書き出しと結びに注目して— 加藤恵梨 (愛知教育大学), 角谷昌範 (愛知教育大学附属岡崎小学校) .....	159
● 中古和文資料『夜の寝覚』のコーパス構築の試み 菊池そのみ (筑波大学), 菅野倫匡 (筑波大学) .....	171
● 『日本語日常会話コーパス』にみる日常会話音声の基本周波数と談話行為の関係 石本祐一 (ものづくり大学/国立国語研究所) .....	185
● 膠着型言語における複雑さのトレードオフ 李文超 (浙江大学) .....	193
● Subjective frequency norms for 100 Japanese verb-verb compounds: The first step toward constructing a behavioral database for Japanese compound words Yijia Yao (Sophia University) .....	206
● 用言の結合価に見る, 体言のブランディング 青山文啓 (桜美林大学大学院) .....	219
● 中世期日本語比喩表現の収集の試み 菊地礼 (国立国語研究所) .....	224
● 近世・近代の日本語及び沖縄語訳聖書のパラレル・コーパスの構築 宮川創 (国立国語研究所) .....	235
● 日本語教師向け Web コンテンツ「つくば語彙チェッカー」の概要 岩崎拓也 (筑波大学), 波多野博顕 (筑波大学), 伊藤秀明 (筑波大学) .....	245
● 近現代語コーパスにおける漢語「是非」 東泉裕子 (東洋大学), 高橋圭子 (東洋大学) .....	252
● 親子の共同行為場面における振る舞いの調整—お菓子作りの事例から— 居關友里子 (国立国語研究所), 小磯花絵 (国立国語研究所) .....	263
● Building a Kansai accent dictionary using YouTube Hiroto Noguchi (Sophia University / Tokyo Medical and Dental University) .....	273
● 『分類語彙表』における多義語について 山崎誠 (国立国語研究所) .....	277

- 教科学習における抽象的思考と結びつく言語形式—数学における「とする」をケーススタディとして—  
新山聖也 (筑波大学), 竹本理美 (筑波大学), 澤田浩子 (筑波大学) ..... 284
- 『昭和・平成書き言葉コーパス』雑誌レジスターに見る順接・逆接の接続詞の通時的変化  
近藤明日子 (東京大学) ..... 297
- 実践医療用語\_語構成要素語彙試案表 Ver.3 の公開にむけて  
相良かおる (奈良先端科学技術大学院大学), 黒田航 (杏林大学), 東条佳奈 (大阪大学), 西嶋佑太郎 (京都大学), 麻子軒 (関西大学), 山崎誠 (国立国語研究所) ..... 309
- 科学技術系ライティング教育改善を目的としたシラバス分析のための spaCY-GSDLUW を利用した日本語長単位解析  
堀一成 (大阪大学) ..... 319
- 家庭での食事場面における親子会話の脱文脈度の観点からの分析  
田中弥生 (国立国語研究所), 江口典子 (国立国語研究所), 小磯花絵 (国立国語研究所) ... 329

# Programme: 言語資源ワークショップ 2023

2023年8月28日(月)

- 9:10-9:20 ■オープニング 高田智和
- 9:20-10:50 ■セッション1 口頭発表
- [o01]  
学習者コーパス研究における横断・縦断データ統合の意義—I-JAS と B-JAS をめぐって—  
..... 石川慎一郎 (神戸大学)
- [o02-s]  
介護分野における専門用語の平易化に向けた語彙リストの構築  
..... 黄海洪 (京都大学大学院), 金丸敏幸 (京都大学)
- [o03-s]  
日中対訳コーパスの構築と公開に向けて  
..... 宮本華瑠 (大阪大学)
- 10:50-11:00 休憩
- 11:00-12:00 ■セッション2 口頭発表
- [o04-s]  
近代中国語関係書における主体移動表現の日中対照研究  
..... 金敬玲 (國學院大學大学院)
- [o05-s]  
説明的文章の要点把握のための読解方法の有効性評価に用いる正解文データセット (CAKeS) の作成  
..... 渡邊幸佑 (兵庫教育大学)
- 13:00-14:00 ■招待講演1  
国立情報学研究所における言語資源共有の取り組み  
..... 大須賀智子 (国立情報学研究所)
- 14:00-14:10 休憩
- 14:10-15:10 ■セッション3 口頭発表
- [o06-s]  
親疎関係で見る上昇下降調の使用率—『日本語日常会話コーパス』を用いて—  
..... 李海琪 (浙江大学)
- [o07-s]  
論文指導の場面における終助詞「かな」の配慮機能—『BTSJ 日本語自然会話コーパス』のデータから—

- ..... 劉悦 (筑波大学)
- 15:10-15:20 休憩
- 15:20-16:20 ■セッション4 口頭発表
- [o08-s]  
程度副詞使用実態の横断的・縦断的調査―「通時話し言葉コーパス」の試み―  
..... 日暮康晴 (筑波大学大学院)
- [o09-s]  
歩きスマホのコロケーションと意味―新聞データベースを用いた使用実態調査―  
..... 王鑫 (筑波大学)
- 16:20-16:30 休憩
- 16:30-17:15 ■セッション5 ポスター発表
- [p11-s]  
国語科教科書の説明的文章における疑問-解答の段落構成―各学年テキストの比較から―  
..... 尾崎誉治 (東北大学)
- [p12-s]  
アーカイブデータを利用した言語研究とその応用可能性  
... 鈴木成典 (国際基督教大学／日本学術振興会), 鎌野慈人 (ストーニーブルック大学), 坂本誓 (国際基督教大学), 鎌倉欧亮 (国際基督教大学), 李勝勲 (国際基督教大学／ヴェンダ大学), 閔宇 (立命館大学), パーキンズ ジェレミー (会津大学), 五十嵐陽介 (国立国語研究所)
- [p13]  
コンパクトな LLM による古典日本語生成  
..... 近藤泰弘 (青山学院大学)
- [p14]  
書籍の文体と修辞機能の分析のパイロットスタディ  
田中弥生 (国立国語研究所), 柏野和佳子 (国立国語研究所), 加藤祥 (目白大学)
- [p15]  
ゲームコーパスの設計方針と構築方法  
..... 麻子軒 (関西大学)
- [p16]  
二字漢語の意味的透明性における仲介語の検討―同一サンプル内での出現位置に着目した調査―  
..... 本多由美子 (国立国語研究所)
- [p17]

テキスト生成型 AI による研究課題のクラスタリング分析

..... 中渡瀬秀一 (国立情報学研究所)



2023年8月29日(火)

- 9:35-10:20 ■セッション6 ポスター発表
- [p21]  
読書感想文の分析と指導法—書き出しと結びに注目して—  
..... 加藤恵梨 (愛知教育大学), 角谷昌範 (愛知教育大学附属岡崎小学校)
- [p22]  
中古和文資料『夜の寝覚』のコーパス構築の試み  
..... 菊池そのみ (筑波大学), 菅野倫匡 (筑波大学)
- [p23]  
日本語学習者の語りに見る語彙学習ストラテジーのプロセス  
..... 安芝恩
- [p24]  
日本語地図課題対話においてソ系指示詞はいつ使われるか?  
..... 川端良子 (国立国語研究所)
- [p25]  
『分類語彙表』への多義語の意味増補のための『計算機用日本語基本辞書IPAL』との対照  
..... 柏野和佳子 (国立国語研究所), 大阿久志緒理
- [p26]  
『子ども版日常会話コーパス』の構築に関する中間報告  
..... 小磯花絵 (国立国語研究所), 居關友里子 (国立国語研究所), 柏野和佳子 (国立国語研究所), 川端良子 (国立国語研究所), 田中弥生 (国立国語研究所), 西川賢哉 (国立国語研究所)
- [p27]  
『日本語日常会話コーパス』にみる日常会話音声の基本周波数と談話行為の関係  
..... 石本祐一 (ものづくり大学/国立国語研究所)
- 10:20-10:30 休憩
- 10:30-12:00 ■セッション7 口頭発表
- [o10]  
膠着型言語における複雑さのトレードオフ  
..... 李文超 (浙江大学)
- [o11]  
Subjective frequency norms for 100 Japanese verb-verb compounds: The first step toward constructing a behavioral database for Japanese compound words  
..... Yijia Yao (Sophia University)

	[o12]	用言の結合価に見る，体言のブランディング ..... 青山文啓 (桜美林大学大学院)
12:00-13:00		休憩
13:00-14:00	■招待講演 2	<b>Building collaborative language resources with and for language communities</b> ..... 李勝勳 (国際基督教大学)
14:00-14:10		休憩
14:10-15:40	■セッション 8 口頭発表	[o13] 中世期日本語比喩表現の収集の試み ..... 菊地礼 (国立国語研究所)
		[o14] 近世・近代の日本語及び沖縄語訳聖書のパラレル・コーパスの構築 ..... 宮川創 (国立国語研究所)
		[o15] 日本語教師向け Web コンテンツ「つくば語彙チェッカー」の概要 ..... 岩崎拓也 (筑波大学), 波多野博顕 (筑波大学), 伊藤秀明 (筑波大学)
15:40-15:50		休憩
15:50-16:35	■セッション 9 ポスター発表	[p31] 昭和女子大学近代文庫所蔵『與謝野晶子書簡』コーパスの設計 ..... 宮崎由美 (昭和女子大学), 日文コースプロジェクト 5 メンバー
		[p32] 近現代語コーパスにおける漢語「是非」 ..... 東泉裕子 (東洋大学), 高橋圭子 (東洋大学)
		[p33] 親子の共同行為場面における振る舞いの調整—お菓子作りの事例から— ..... 居關友里子 (国立国語研究所), 小磯花絵 (国立国語研究所)
		[p34] 近世読本コーパスの設計と活用—表記研究での利用を目指して— ..... 片山久留美 (国立国語研究所)
		[p35] <b>Building a Kansai accent dictionary using YouTube</b> ..... Hiroto Noguchi (Sophia University / Tokyo Medical and Dental University)
		[p36]

例外的な「基本形」—日本語会話における動詞終わりの発話—

.....大野剛 (アルバータ大学), 白田泰如 (静岡理科大学)

[p37]

『分類語彙表』における多義語について

.....山崎誠 (国立国語研究所)

16:35-16:45 休憩

16:45-17:30 ■セッション 10 ポスター発表

[p41]

教科学習における抽象的思考と結びつく言語形式—数学における「とする」  
をケーススタディとして—

.....新山聖也 (筑波大学), 竹本理美 (筑波大学), 澤田浩子 (筑波大学)

[p42]

『昭和・平成書き言葉コーパス』雑誌レジスターに見る順接・逆接の接続詞  
の通時的変化

.....近藤明日子 (東京大学)

[p43]

抄物コーパス構築に関する課題と方策

.....古田龍啓 (九州大学)

[p44]

実践医療用語\_語構成要素語彙試案表 Ver.3 の公開にむけて

.相良かおる (奈良先端科学技術大学院大学), 黒田航 (杏林大学), 東条佳奈  
(大阪大学), 西嶋佑太郎 (京都大学), 麻子軒 (関西大学), 山崎誠 (国立国語  
研究所)

[p45]

科学技術系ライティング教育改善を目的としたシラバス分析のための  
spaCY-GSDLUW を利用した日本語長単位解析

.....堀一成 (大阪大学)

[p46]

雑誌コーパス作成の試み—2015 年の資料をもとに—

.....東条佳奈 (大阪大学)

[p47]

家庭での食事場面における親子会話の脱文脈度の観点からの分析

田中弥生 (国立国語研究所), 江口典子 (国立国語研究所), 小磯花絵 (国立国  
語研究所)

17:30-18:00 ■セッション 11 クロージング 高田智和

## 発表論文集

# 学習者コーパス研究における横断・縦断データ統合の意義： I-JAS と B-JAS をめぐって

石川 慎一郎（神戸大学）<sup>†</sup>

## Significance of Integrative Analysis of Cross-sectional and Longitudinal Data in Learner Corpus Research: A Study Based on the I-JAS and B-JAS

Shin'ichiro Ishikawa (Kobe University)

### 要旨

本論文では、「多言語母語の日本語学習者横断コーパス」(I-JAS)と「北京日本語学習者縦断コーパス」(B-JAS)を用い、共通のストーリーライティング課題(SW1)における中国語母語の日本語学習者の産出データを分析した。習熟度・学習段階別に4群に分けて比較した結果、2つのコーパスは計量的指標やマクロレベルの発達傾向については同等のパターンを示すものの、各段階を特徴づける具体的な品詞および語彙項目については違いも観察された。これらの結果は、今後の学習者コーパス研究において、横断的データと縦断的データを統合的に分析する意義を示唆している。

### Abstract

This paper analyzed L2 Japanese outputs of L1 Chinese learners elicited in a common story writing task. The data was taken from a pair of learner corpora based on the common prompts: the International Corpus of Japanese as a Second Language (I-JAS) —a cross-sectional dataset— and the Beijing Corpus of Japanese as a Second Language (B-JAS) —a longitudinal dataset—. A comparison of four learner groups, who were classified according to the proficiency levels and learning stages, revealed that the two corpora showed similar patterns in terms of major quantitative indices as well as the overall developmental paths, while they showed different keywords for each of the four groups. These results suggest the significance of integrated analysis of cross-sectional and longitudinal data in future learner corpus research.

### 1. はじめに

学習者コーパスには、ある一時点において、年齢や習熟度の異なる多様な学習者のL2産出を一斉に収集する横断コーパス (cross-sectional corpus) と、1名ないし数名の学習者のL2学習過程を長期にわたってモニターし、その間、数回(線形近似や、いわゆるUカーブの検証を行うには一般に3回以上が必要とされる: Ployhart and Vandenberg, 2010, p. 97) にわたってL2産出を収集する縦断コーパス (longitudinal corpus) が存在する。

このうち、縦断コーパスは、発達の過程を直接記録し、多くの場合、学習者背景に関する詳細なメタデータが付属することから、発達コーパス (developmental corpus) や濃密コ

---

<sup>†</sup> iskwhin@gmail.com

ーパス (dense corpus) と呼ばれることもある (Meunier, 2015)。「発達を調査しようとする場合、同じ学習者集団を発達の時系列に沿って追跡していくことが理想」であり (Myles, 2015; 英文文献からの引用は拙訳による。以下同)、縦断コーパスと付属する背景データを組み合わせて解釈することで、学習者の行動や周辺環境において起こる様々な変化を加味しながら、個々の学習者の学習過程 (developmental paths) を解明することができる (Vyatkina and Cunningham, 2015)。

一方、縦断コーパスの構築はしばしば困難な取り組みとなりうる。Meunier (2015) も言うように、縦断コーパスの構築は時間がかかり、事前計画を立てにくく、研究アウトプットを出しにくく、短期プロジェクトに比べて助成金を得にくく、さらには、参加者が自然減少 (attrition) する場合も多い。また、縦断と言っても、L2 習得の始点から終点まで、つまり、L2 を習い始めた時点から、母語話者相当の L2 能力を獲得する時点までの全過程を調査することはそもそも不可能であり、多くの場合、「少数の学習者を対象とした、数か月ないし数年程度の L2 発達の記録」とならざるを得ない (Myles, 2015)。

こうした背景をふまえ、縦断コーパスに代わり、横断コーパスを用いて発達を研究する試みも広くなされている。横断コーパスの参加者を習熟度や学年といった観点で群化し、各群 (たとえば、初級・中級・上級など) を比較することで習得プロセスを間接的に観察するのである。こうした手法は、疑似縦断分析法 (pseudo-longitudinal analysis) (Johnson and Johnson, 1999) や、準縦断分析法 (quasi-longitudinal analysis) (Granger, 2002) などと呼ばれる。Myles (2015) は、各群で十分な数の学習者数が確保されており、かつ、外的基準による信頼できる習熟度情報が備わっているならば、横断データを用いて学習者の「発達の道筋」 (developmental routes) を跡付けることが可能であると示唆している。

もっとも、ここで留意すべきは、横断データを用いた疑似縦断分析で得られた結果と、縦断データを直接分析して得られた結果の一致が必ずしも検証されていないということである。こうした研究がなされてこなかったのは、相互比較可能な横断・縦断データがほとんど存在していなかったという事情による。実際、学習者の背景や、産出のトピック、また、タスクの内容が異なる横断・縦断データを比較して何らかの異同が明らかになったとしても、データ収集法の差以外の要因による可能性が高く、横断・縦断データの関係性について確定的な結論を導き出すことはできない。ゆえに、横断コーパスと縦断コーパスが学習者の L2 習得研究の資料としてどの程度の同等性を持ちうるのか、あるいは、それぞれの強み・弱みを補うため、どのように両者を統合的に利用すべきか、といった点についての共通理解はいまだ醸成されていないのが現状である。

こうした状況が長く続いてきた中、迫田久美子氏が中心となって開発した「多言語母語の日本語学習者横断コーパス」 (I-JAS) (迫田, 2020) をモデルとして、同一タスクで縦断的にデータを集めた「北京日本語学習者縦断コーパス」 (B-JAS) がリリースされたことは注目に値する。本研究は、これら 2 種のコーパスから、中国語を母語とする日本語学習者 (Chinese learners of Japanese、以下 CLJ) の L2 習得パターンを計量的に抽出し、その異同を比較することで、横断データと縦断データの関係性について検証を試みたい。

## 2. 先行研究

日本語学習者コーパスを用いた研究では、研究者の関心対象である文法項目 (テンス、アスペクト、授受表現、敬語、など) に焦点を絞り、その習得過程を質的観点から調査したものが多く、一方で、学習者の産出テキストの全体を対象として、L2 習得の総体的過程

を量的観点からモデル化しようとした研究もある。ここでは、後者のタイプの研究の例として、筆者の過去の研究の概要を示す。

まず、I-JAS を用いた研究として、Ishikawa (2017) は、中国語・韓国語・英語・トルコ語を母語とする学習者のストーリーライティング産出を調査した。学習者を習熟度別に7段階に区分して比較した結果、習熟度の上昇によって産出語数が増加すること、学習者は助詞・動詞・副詞を過小使用し、名詞・助動詞などを過剰使用すること、初級・中級では母語別の違いが出るが、上級になると母語の差は縮小していくことなどを明らかにした。

石川 (2020) では、対話タスクにおける形容詞の使用状況を分析し、中国語・韓国語・英語を母語とする中級学習者による形容詞の使用量は日本語母語話者の半分程度であること、日本語母語話者の形容詞使用は過去形とくだけた「ちっちゃい」「すごい」などによって特徴づけられることを示した。

石川 (2021a) は、世界の日本語学習者 850 人を J-CAT スコアの 10 点刻みで細かく群化し、ストーリーライティング産出のデータを用いて、習熟度の発達と動詞使用の関係を量的に概観した。その結果、(1)学習者の使用する動詞のトークン数およびタイプ数は母語話者より 22~28%程度少ないこと、また、習熟度別に見た場合、トークン数・タイプ数とも初級から中級にかけて増加し、母語話者を上回る水準に達した後、再び減少する逆 U 字型の分布を見せること、(2)母語話者・学習者・上級者・初級者を特徴づける動詞群が存在すること、(3)学習者の動詞発達過程がおおよそ4段階に分かれること、(4)少数の動詞(居る、見る、為る、作る、有る)の頻度で習熟度が予測できる可能性があること、などを示した。また、石川 (2021b) では同様の分析を韓国語母語の学習者データに適用し、石川 (2021a) で観察されたパターンが特定母語の学習者群に対してもほぼ同様にあてはまることを示した。

次に、縦断コーパスを使用した研究として、石川 (2018b) は、迫田久美子氏が中心となって構築した「中国語・韓国語母語の日本語学習者縦断発話コーパス」(C-JAS) を用い、3名の CLJ の発話データを語彙解析した。その結果、(1)トークン数とタイプ数に関しては3名の学習者ともに増加するものの、語彙多様性指標についてははっきりした変化が起こらないこと、(2)上位語は総じて安定的であるものの、初期においては機械的な応答標識が多く、中期・後期になると、次第に文法的に複雑な文を構成する上で必要となる語が多く使用されるようになること、(3)個体の差は発達段階の差よりも強く語彙使用に影響しているが、個体内では発達段階の差も見られること、(4)3名の学習者の語彙習得段階がおおよそ3期に区分でき、それぞれが異なる語群によって特徴づけられることなどを明らかにした。この研究で得られた知見の一つは、縦断データにおいては個人差が非常に大きく、3名のデータの安易な一般化は危険である、ということであった。

石川 (2018a) では、台湾の東呉大学で構築された縦断作文コーパスである LARP at SCU を用い、17名の CLJ 大学生が3年半にかけて書いた作文中のタイプ数・トークン数、句読点使用率、品詞使用率、高頻度語彙使用率などを調査した。その結果、初期においては個人差によるばらつきが大きいのが、大学入学後20か月(大学2年生後期)ごろを境界として次第に値が安定するようになり、この時期に日本語の習得が一定の定着段階に達しているのではないかという示唆を得た。

以上の筆者の研究は、いずれでも、横断データか縦断データかのいずれかを用いたもので、両者を同時に使用して結果を比較するということは試みていない。これは、一連の研究で使用した I-JAS、C-JAS、LARP at SCU という3つのコーパスが、それぞれ異なる条件でデータを集めており、相互比較が困難だと判断したためである。

これに対し、特定の文法事項を対象を絞りつつ、横断・縦断データをうまく組み合わせる主張の妥当性を高めている研究も存在する。一つの好例は、趙 (2015) であろう。同研究では、縦断型のインタビューコーパスである C-JAS と、横断型のインタビューコーパスである KY Corpus を用い、「テイナイ」の習得過程を検討している。まず、縦断データ分析より、CLJ3名と韓国語母語の学習者3名は、第1期から第8期にかけて、「テイナイ」

の各種用法に関して、ともに、「未完了」「状態」→「属性」→「全面否定」「反復」の順に使用するようになることが示された。次に、横断データ分析により、中国語・韓国語・英語の母語話者は、いずれも、「未完了」「状態」→「属性」「反復」（「全面否定」は高レベルでも少ない）の順に使用するようになることが示された。若干のずれはあるものの、全体として傾向は類似しており、著者は、これらを根拠として、「テイナイ」の用法習得順序に普遍性があることを示唆している。

今後の学習者コーパス研究においては、得られた知見の説得力を高めるため、趙（2015）のように、横断データと縦断データを併用していくことが望ましいと思われるが、その際、前提として確認すべきは、そもそも、横断・縦断データはどの程度同質であるのか、あるいは、異なるのか、という点であろう。この意味で、高度に統制された横断・縦断データを計量的に比較し、各々のデータの性質を把握しておくことは不可欠であると思われる。

### 3. リサーチデザイン

#### 3.1 目的と RQ

本研究は、I-JAS と B-JAS に含まれる CLJ のストーリーライティングデータを用い、学習者の段階によって、(1)総語数、(2)主要品詞頻度、(3)特徴語、(4)学習者群の相互関係性、がどのように変化するかを観察することで、横断データと縦断データから得られる結果の一致度を検証し、今後の学習者コーパス研究における横断データと縦断データの統合的使用の指針を得ることを主たる目的とする。本研究で使用するデータは、学習者の母語背景と産出タスクの両方がそろっているため、横断データと縦断データの性質の違いの検証に適したものとなっている。

本研究で具体的に検討しようとする研究設問（RQ）は以下のとおりである。なお、上記の観点(4)については、主要品詞頻度と、石川（2021a、2021b）でも使用した主要動詞頻度を分類の基準とする。

- RQ1 総語数に関して、横断・縦断の各データから得られる発達のパターンはどの程度一致するか？（総語数）
- RQ2 主要品詞頻度に関して、横断・縦断の各データから得られる発達のパターンはどの程度一致するか？（品詞頻度）
- RQ3 特徴語に関して、横断・縦断の各データから得られる発達のパターンはどの程度一致するか？（特徴語）
- RQ4 主要品詞頻度および主要動詞頻度に基づく段階分類に関して、横断・縦断の各データから得られる発達のパターンはどの程度一致するか？（段階分類）

#### 3.2 使用するデータ

前述のように、本研究は、I-JAS と B-JAS に含まれるストーリーライティングのデータを分析対象とする。使用したデータは、2023年7月時点における最新版である。両コーパスで実施された対面調査（インタビュー）は、はじめに、ストーリーテリング（4～5コマの漫画を見てそのストーリーを口頭で述べる；約10分）を行い、その後、対話（決められた内容・手順に基づいてインタビュワーが行う質問に答える；約30分）、ロールプレイ（謝罪・依頼という言語機能を含む状況についてロールプレイを行う；約10分）、絵描写（1枚絵を描写；約5分）を順次行い、最後に、ストーリーライティング（約20分）を行う（迫田・石川・李、2020、p. 34）。ストーリーライティングでは、冒頭のストーリーテリングで使ったものと同じ漫画イラストを見て、ストーリーをパソコン上で入力する形で実施される（同上、p. 41）。

ストーリーライティング（SW）には、SW1（ピクニックに関する5コマ漫画のストーリーを作る）と、SW2（鍵に関する4コマ漫画のストーリーを作る）の2種が存在するが、トピックの影響を可能な限り統制するため、本研究は、SW1のみを分析対象とする。



### 3.3 対象学習者

本研究は、学習者の母語背景と学習背景を統一するため、I-JAS に参加した 1,000 名の学習者中、大陸在住の CLJ (CCH および CCM) 100 名と、B-JAS の CLJ 大学生 17 名を分析対象とする。なお、両コーパスの参加者は、全員が、J-CAT (聴解・語彙・文法・読解の 4 観点別の日本語能力測定。各 100 点、400 点満点) と、SPOT (文章穴埋め式の総合的日本語能力測定。90 点満点) という 2 種のテストを受験している。

#### 3.3.1 I-JAS の CLJ

I-JAS の CLJ100 名については、疑似縦断分析を行うため、J-CAT の 4 観点スコアと SPOT スコアの計 5 種のテストスコアを主成分分析の手法で合成した上で、第 1 主成分スコアを基準として、習熟度別に 4 段階に区分する。表 1 は、5 種のスコアの量的概要である。

表 1 I-JAS の CLJ100 名の習熟度テストスコアの概要

変数	人数	平均	不偏分散	標準偏差	最小値	最大値
J-CAT 聴解	100	56.1	227.9	15.1	8.0	86.0
J-CAT 語彙	100	64.8	122.6	11.1	40.0	100.0
J-CAT 文法	100	56.9	232.3	15.2	23.0	91.0
J-CAT 読解	100	51.7	136.5	11.7	11.0	82.0
SPOT	100	73.8	46.5	6.8	57.0	90.0

J-CAT の語彙セクションを除くと、最大値が満点に達しているものではなく、これらのテストは受験者の能力値を正しく測定しているものと推定できる。

5 種のスコア内に線形結合している変数はなかった。分析で得られた第 1、第 2、第 3 主成分の寄与率はそれぞれ 61.3%、13.5%、11.4%であり、第 1 主成分で分散の 6 割以上がカバーされていた。また、第 1 主成分の負荷量の係数はすべてプラスであり (図 1)、第 1 主成分が 5 種のスコアの代表値になっていることが確認された。

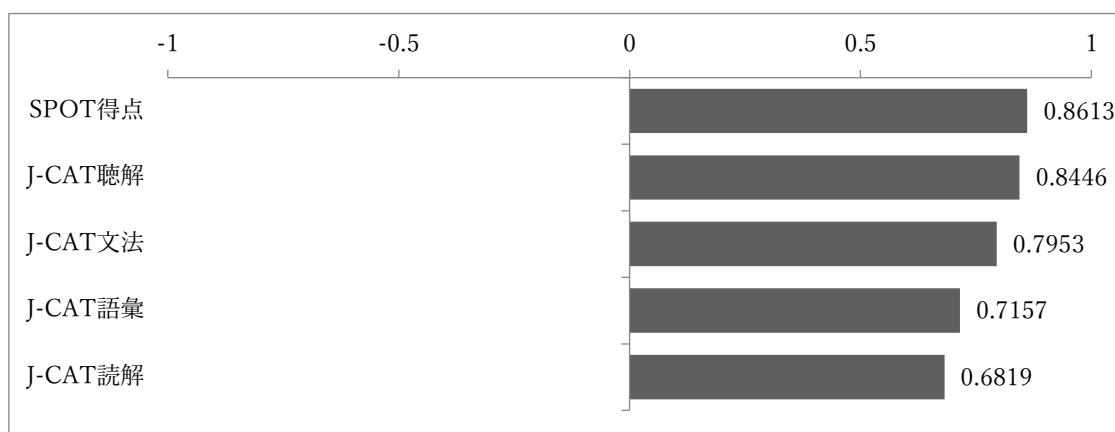


図 1 第 1 主成分の負荷量

個々の学習者ごとに第 1 主成分得点を求めたところ、主成分得点のレンジは -5.11 ~ +5.29 となった。スコアレンジをおよそ均等に分割しつつ、かつ、各群に 10 名以上の学習者が入るよう調整を行い、表 2 のような 4 段階習熟度区分を設定した。

表 2 I-JAS の CLJ100 名の 4 群化

	段階 1 (I1)	段階 2 (I2)	段階 3 (I3)	段階 4 (I4)
主成分得点	-2.0 未満	-2.0 以上	0.0 以上	2.3 以上
学習者人数	12	39	37	12
J-CAT 総合点平均値	168.3	206.2	251.2	299.2
SPOT 平均値	64.0	70.9	77.0	83.2
推定能力水準	中級	中級	中級～上級	上級

J-CATの総合点に基づいて言うと、段階1～2は「日常的な会話をこなすことができる」中級相当（中級～中級後半）、段階3～4は「学術的・専門的なコミュニケーションができる」上級の前半に相当する。また、SPOTに基づいて言うと、段階1～段階3は「自然な発話速度で日常的な場面の日本語がある程度理解できる」とされる中級相当、段階4は「自然な発話速度で幅広い場面の日本語が理解できる」上級相当ということになる（両テストのスコアの解釈については、李・小林・今井・酒井・迫田、2015に基づく）。

### 3.3.2 B-JAS の CLJ

B-JASのCLJ大学生17名については、データの収集時期を手掛かりとして、4段階に区分する。B-JASは、北京の大学で日本語を主専攻とする17名のCLJを対象として、4年間（2016年1月～2019年4月）にわたって毎年2回、計8回の縦断的データ収集を行っている。ただし、SWについては、I-JASタスクと独自タスクを交互に課しているため、今回分析対象とするSW1のデータがとられたのは表3に示す4回である（いずれもほぼ同じタイミングで2種の試験を受験している）。なお、表3中の学習時間は、2015年9月に新学期が開始したとして、各時点における通算の日本語学習時間の目安を月数で表示したものである。

表 3 B-JAS の CLJSW1/2 調査回数と学習時間

	段階 1 (B1)	段階 2 (B2)	段階 3 (B3)	段階 4 (B4)
データ収集時期	2016/1	2016/10	2017/9	2018/10
推定学習時間	5 か月	14 か月	25 か月	38 か月
学習者人数	17	17	17	17
J-CAT 総合点平均値	136.2	193.7	224.8	263.1
SPOT 平均値	60.4	71.8	75.2	80.9
推定能力水準	中級	中級	中級	上級

J-CATの総合点に基づいて言うと、段階1～3は中級相当（中級前半～中級後半）、段階4は上級の前半に相当し、SPOTに基づいて言うと、段階1～段階3は中級相当、段階4は上級相当となる。各段階の平均スコアは、J-CATではB-JASのほうがやや低いが、SPOTではほとんど差がなく、初級・中級・上級という大まかな習熟レベルでとらえれば、I-JASとB-JASの4段階の能力水準はほぼ同等であると推定される。

### 3.4 テキストデータの処理

まず、I-JASの習熟度別4群（以下、I1～I4）と、B-JASの学習時間別の4群（以下、B1～B4）の各々について、該当する学習者によるSW1の産出テキストを統合してマージファイルを作成した。その際、各行左端に記載されているコードは除去している。

なお、I-JASの公開版テキストデータには、丸括弧内に、各種のメモ（「括弧を変更」といった修正記録、「過剰使用」などの学習者の言語使用特性についての注記、「犬の名前」などのテキスト理解のための補足情報など）や、学習者の軽微な書き間違いについての修

正提案が記載されている。前者については一律で削除し、後者については修正提案を採用した。B-JAS の公開版テキストデータには丸括弧内の追記は見当たらなかったが、I-JAS 側との整合を取るため、同等レベルの誤記については筆者が修正を加えた。また、データ全体に対して、文字種レベルでの若干の統一（人名の「けん」や「まり」はカタカナに、「いぬ」は「犬」に、など）を加えた。

以上の処理後に、I1~4、B1~4 の 8 ファイルを「web 茶まめ」（現代語辞書）で解析し、結果を 1 枚のエクセルシート上に集約した。なお、書字形欄が空データ（補助記号扱いのものを含む）になっている行は削除した。また、語彙素列に含まれる注記（ハイフンで追記された外来語の原語、多義語の用法など）についても一律で削除した。また、RQ3 の特徴語分析のため、語彙素列のみを取り出して、8 種の語彙素単位のテキストファイルを作成した。

### 3.5 処理手順

RQ1（総語数）については、I1~4、B1~4 の 8 ファイルから、1 人あたりのトークン数（平均値）を調査し、I-JAS と B-JAS の間で、段階変化に伴う総語数（5 種の補助記号を除く）の変化パターンがどの程度一致しているか確認する。

RQ2（品詞頻度）については、主要品詞の中から、名詞（普通名詞：サ変可、サ変形状詞可、一般、形状詞可、助動詞可、副詞可）（「部屋」「犬」「サンドイッチ」など）、動詞（一般、非自立可）（「気付く」「飛び込む」「為る」「食べる」など）、格助詞（「の」「を」「に」など）、接続助詞（「て」など）、係助詞（「は」）の 5 品詞に注目し、1 人あたりの使用トークン数（平均値）を調査し、I-JAS と B-JAS の間で、段階変化に伴う主要品詞の頻度変化のパターンがどの程度一致しているか確認する。なお、各品詞の頻度は絶対値に大きな異なりがあるため、ここでは、すべてを自然対数に変換した上で比較を行う。

RQ3（特徴語）については、I1~4 の各語彙素単位ファイルとそれらの総体ファイル、また、B1~4 の各語彙素単位ファイルとそれらの総体ファイルにおいて、出現する語の頻度を悉皆的に調査して対数尤度比統計量（log-likelihood ratio : LLR）を計算する。今回は、データの絶対量が少なく、通常的手法では十分な数の特徴語が検出できないことから、有意傾向の判定基準となる  $\alpha = 10\%$  の下限値となる 2.71 を超えるすべての語を抽出する。上述の理由から、検定反復にもなって必要とされる多重比較補正は行わない。また、頻度の絶対的な差を示す効果量指標として、相互情報量（mutual information）を計算する。特徴語の抽出と、相互情報量の計算には、多言語対応コーパスコンコルダンスである Antconc Version 4.2.0 を使用する。その後、I-JAS と B-JAS の間で、各段階の特徴語が、どの程度、内容的に一致しているかを確認する。

RQ4（段階分類）では、上述の 5 品詞に、形容詞類（形容詞：一般、非自立可、形状詞：一般、助動詞語幹）（「悲しい」「楽しい」など）、助動詞（「れる」「ます」「た」など）、副詞（「そろそろ」「もう」「がっかり」など）、接続詞（「そして」「しかし」など）、連体詞（「この」「その」など）を加えた 10 品詞を第 1 アイテム、学習者 8 群を第 2 アイテムとする頻度表と、石川（2021a、2021b）の手法に倣い、頻度上位 30 種の動詞（語彙素）を第 1 アイテム（表 4）、学習者 8 群を第 2 アイテムとする頻度表を用意し、それぞれに対応分析を適用する。対応分析とは、頻度表の行列の相関を最大化する少数の次元を抽出し、通例、上位の 2 次元をそれぞれ横軸・縦軸とする散布図を描画し、アイテムカテゴリ間の相互関係を可視化する多変量解析手法のことである。ここでは、散布図の質的解釈を通して、I-JAS と B-JAS 間で 4 段階の位置づけにどのような差が出ているかを概観した後、各段階を特徴づける具体的な品詞・動詞項目の一致度を調べる。

表 4 RQ4 で使用する高頻度 30 動詞

高頻度動詞（上位 30 語）
為る、居る、見る、食べる、行く、入る、仕舞う、開ける、飛び出す、作る、来る、持つ、有る、着く、知る、入れる、飛ぶ、出る、成る、思う、開く、分かる、飛び込む、終わる、読む、言う、切る、驚く、つく、遊ぶ

## 4. 結果と考察

### 4.1 RQ1 総語数

横断・縦断データである I-JAS と B-JAS において、段階別に、1 人あたり総語数（トークン数）の変化を調査したところ、図 2 の結果を得た。

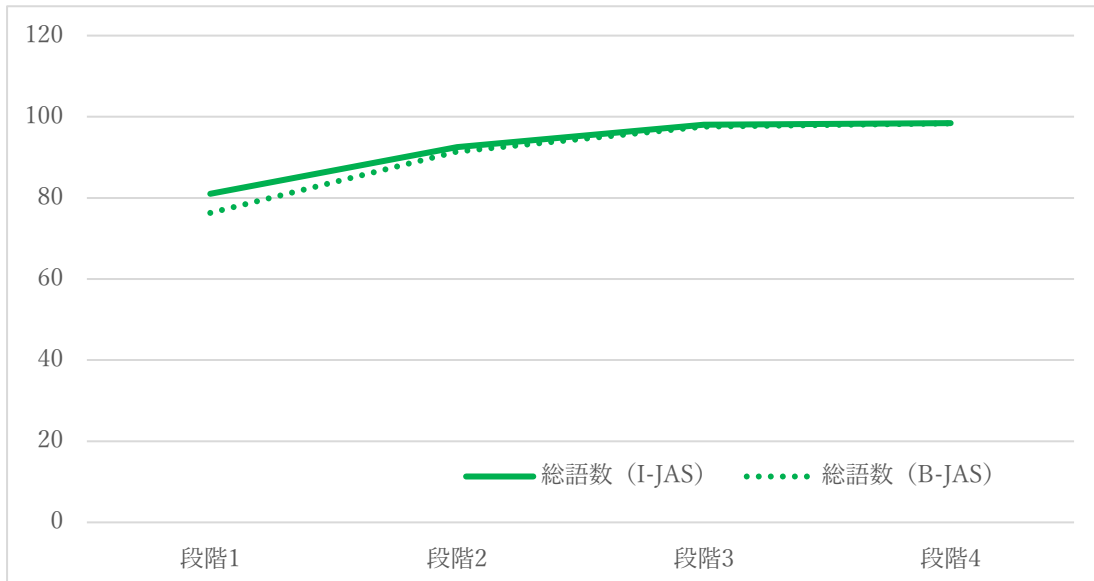


図 2 SW1 総語数 (1 人あたり) の段階別変化

上記より、I-JAS においても B-JAS においても、(1)語数はおおよそ 80~100 語の範囲に収まること、(2)段階が進むにつれて総語数が増加すること、(3)語数の増加幅は段階 1→2 の間が最も大きく、以下、段階 2→3、段階 3→4 になるにつれて小さくなること、が確認された。段階 1 において、B-JAS のほうがやや語数が少ないことを除けば、総語数の変化パターンは、I-JAS と B-JAS で高い同一性を示していると結論できる。

### 4.2 RQ2 品詞頻度

まず、I-JAS と B-JAS において、段階別に、1 人あたりの名詞・動詞頻度（トークン数を自然対数に変換）の変化を調査したところ、図 3 の結果を得た。

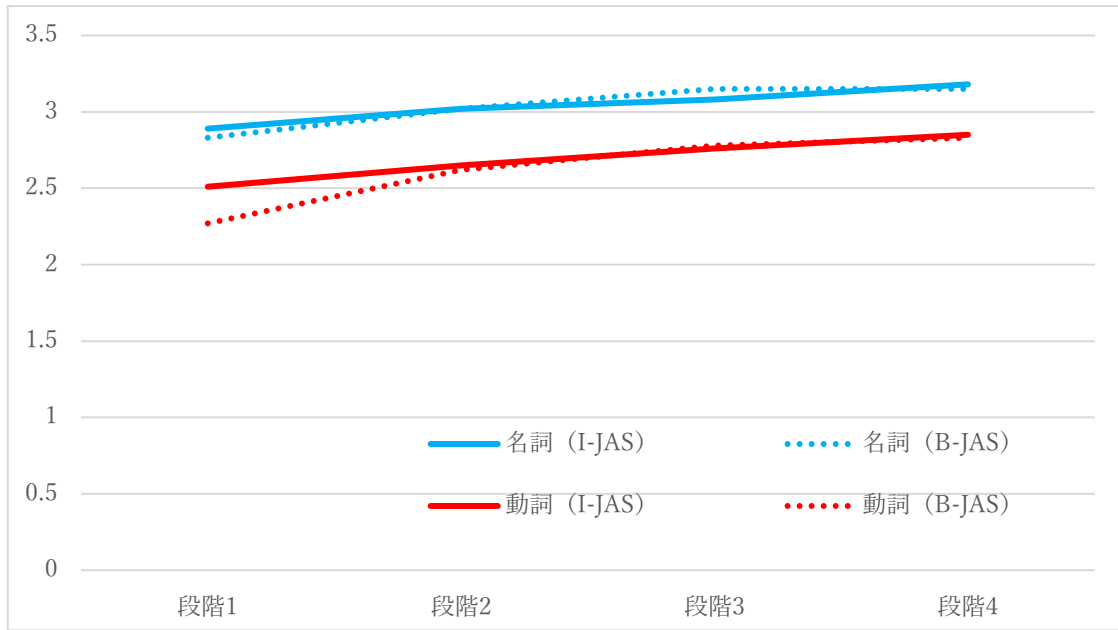


図 3 SW1 作文における名詞・動詞頻度の段階別変化

上記より、I-JAS、B-JAS とともに、(1)対数変換後の頻度で言うと、名詞はおよそ 3 前後、動詞はおよそ 2.5 前後になること、および、(2)段階が進むにつれて頻度はわずかに上昇するが、上昇幅は総じて小さいこと、が確認された。また、B-JAS の段階 1 を除くと、値も、I-JAS と B-JAS 間で非常に似通っていることが示された。

続いて、段階別に、1人あたりの3種の助詞の使用頻度（トークン数を自然対数に変換）の変化を調査したところ、図 4 の結果を得た。

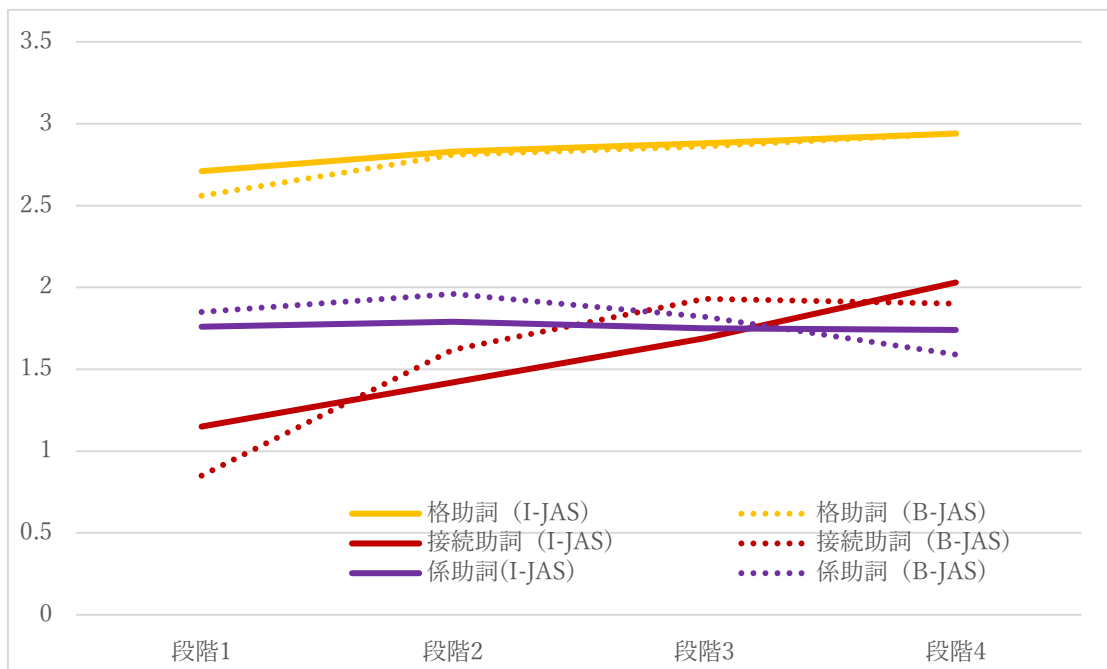


図 4 SW1 作文における助詞頻度の段階別変化

上記より、I-JAS、B-JAS とともに、(1)対数変化後の頻度で言うと、格助詞はおよそ 2.5～3、係助詞は 1.5～2、接続助詞は 1～2 となること、(2)格助詞と接続助詞は段階が進むにつれて増加すること、(3)係助詞頻度は減少すること、(4)増加幅は接続助詞が最も大きいこと、が確認された。ただ、接続助詞に関しては、I-JAS では増加幅がほぼ一定で、段階 1～4 にかけて線形的に増加するものの、B-JAS では増加幅が段階 1～2、段階 2～3、段階 3～4 になるにつれて縮小していくこともあわせて観察された。

以上、5 種類の品詞を取り上げて、その頻度変化のパターンを確認してきたが、総じて言えば、I-JAS、B-JAS 間で、高い一致度が示されたと結論できる。ただし、すべての値が同じように変化するわけではなく、接続助詞のように、増加という全体的トレンドは一致していても、増加の仕方に差が出る場合も存在することが示唆された。

#### 4.3 RQ3 特徴語

I-JAS の 4 段階別データとそれらの総体、B-JAS の 4 段階別データとそれらの総体を比較し、各段階の特徴語を調査したところ、表 5 の結果を得た。表中、語彙素に付した数字は、左側が対数尤度比統計量、右側が相互情報量である。 $\alpha=1\%$ 、 $5\%$ 、 $10\%$  (有意傾向)に相当する統計量はそれぞれ 6.63、3.84、2.71 となる。表 5 において差の有意性が確認されたものは太字で記載している。なお、ここには群内の 1 名ないし数名が著しく顕著に使っているものも含まれている。

表 5 I-JAS と B-JAS における各段階の特徴語 ( $\alpha=10\%$ )

段階	I-JAS	B-JAS
1	<b>食べ物</b> (8.4/1.1)、 <b>弁当</b> (7.6/2.2)、 <b>箱</b> (6.0/1.7)、 <b>悪い・御飯</b> (5.4/2.2)、 <b>ね</b> (4.7/1.7)、 <b>思う</b> (3.8/1.3)、 <b>郊外・驚く</b> (3.7/1.5)、 <b>彼</b> (3.0/0.6)、 <b>ながら</b> (2.9/1.3)、 <b>ます</b> (2.9/0.3)	<b>です</b> (8.2/0.7)、 <b>有る</b> (5.4/1.0)、 <b>へ</b> (4.0/0.9)、 <b>は</b> (4.0/0.3)、 <b>夫婦</b> (3.9/1.3)、 <b>ば・大変</b> (3.3/1.5)、 <b>分かる</b> (3.3/1.2)、 <b>たい</b> (3.2/1.0)、 <b>そして</b> (3.1/0.9)、 <b>ケン</b> (3.1/0.4)、 <b>マリ</b> (2.8/0.4)、 <b>楽しい</b> (2.8/0.8)
2	<b>彼</b> (5.0/0.3)、 <b>等</b> (4.7/0.3)、 <b>時</b> (3.0/0.3)、 <b>家</b> (2.8/0.5)	<b>食物</b> (3.6/1.2)、 <b>は</b> (3.2/0.2)
3	(該当なし)	<b>て</b> (4.1/0.3)
4	<b>て</b> (8.8/0.5)、 <b>二人</b> (6.4/0.6)、 <b>居る</b> (5.4/0.6)、 <b>共</b> (4.7/1.5)、 <b>たり・笑う・上・昼</b> (3.7/2.2)、 <b>内</b> (3.3/1.1)、 <b>気付く</b> (3.1/0.9)、 <b>俣</b> (2.7/1.3)、 <b>中</b> (2.7/0.6)	<b>て</b> (3.5/0.2)、 <b>来る</b> (2.9/0.9)

まず、I-JAS について言うと、段階 1 では、基本語の不適切な使用が目につく。たとえば、ピクニックのために用意するサンドイッチを「食べ物」(犬は**食べ物**を食べて (CCH25))、「弁当」(弁当の**バスケット** (CCH28))、「御飯」(ご飯を**作り**ました (CCH30)) と呼んだり、バスケットを「箱」と呼んだりする(犬が**箱**の中に入ります (CCH21))。また、サンドイッチが犬に食べられてなくなってしまったことや、それを知ってがっかりしたことを指して「悪い」と言ったり(食べ物**は悪く**になりました

(CCM07) / 気持ちは悪くになりました (CCM14))、並行的な動作の進行を含意するはずの「ながら」を文脈に合致しない形で用いたりすることもある(彼らは地図を見ながら、犬がバスケットに入った (CCH63))。このほか、丁寧体語尾の「ます」、登場人物の心理的内面を推し量る「(〜と) 思う」、話し言葉的な終助詞「ね」、の使用も多い。

- (1) ケンはマリと楽しみにしていますけど、相談しながら、近く公園に行きます。(CCH21)
- (2) ケンさんとマリさんはとても怖いとおもいました。(CCH24)
- (3) 食べ物はなかったので、残念でしたね。(CCH025)

第2段階になると、2名の登場人物を言い換える「彼等」や、場所表現としての「家」、時間表現としての「〜する時」などが使用されるようになる。

- (4) 彼らはサンドイッチを作った後、地図を見ているとき、彼らが飼う犬はバスケットのサンドイッチを食べてしまいました。(CCH31)
- (5) ある日、ケンさんとマリさんが家でサンドイッチを作っていました。(CCH52)

そして、上級相当の第4段階になると、数詞によって登場人物を代名詞的に言い換える「二人」、アスペクト形式の「ている」、列挙表現の「たり」、状態継続を含意する「まま」のほか、抽象的な場所を指す「上」「内」「中」(これらは意味が比喩的に拡張する場合もある)などが特徴的に使われるようになる。

- (6) 実は二人がどこかに行けばいいかと迷っていますので、ちょっと地図を見て場所を決めているうちに、犬はこっそりとバスケットの中には入りました。(CCM35)
- (7) ですが二人が地図を見ている隙に、家で飼っていた犬がバスケットの中に入り込みました。(CCM51)
- (8) しかし、食べ物を望んでいた犬はバスケットの中に入って、食べ物を全部食べました上に、そのままかごの中に寝てしまいました。(CCH16)
- (9) それでは、二人は歌を歌つたり、笑つたりして、手を握って一緒にピクニックに行きました。(CCM35)

以上、I-JAS の段階別特徴語を概観すると、段階 1~4 にかけて、(1)主として基本名詞に関わる誤用の減少、(2)終助詞や丁寧体を用いた話し言葉的な表現の減少、(3)人物を言い換える代名詞表現の多様化、(4)アスペクト形式(ている)の獲得、(5)抽象的场所表現の獲得、といった大まかな方向性が見て取れる。

では、これと同じ傾向が B-JAS でも見られるのであろうか。B-JAS の特徴語について言うと、段階1では、丁寧体語尾の「です」や存在動詞の「有る」(仕事がありませんでした (CCB007))、取り立ての係助詞「は」や方向を含意する格助詞「へ」(犬はバスケットの中へ跳びました (CCB001))、仮定表現「ば」(大変と言えば大変ですけど (CCB007))、意思・願望を表す助動詞「たい」(犬はサンドイッチが食べたいでした (CCB011))、接続詞「そして」(そして、ケンとマリは地図を見ました (CCB006))などが特徴的に使用される。このうち、丁寧体助動詞と過去助動詞が連結した「でした」については、各種助動詞やイ形容詞など、本来は結合できない語群が後接する事例が多く認められる(したいでした・分らないでした・したことはないでした (CCB003) / 知らないでした (CCB004) / 楽しいでした (CCB006) / 驚嘆でした・面白いでした (CCB007) / 悲しいでした (CCB013) など)。なお、この点に関して、教育出版株式会社編集局 (n.d.) は、昭和27年の国語審議会の見解も紹介しつつ、「美しいです」のような、形容詞に断定助動詞(「だ」)の丁寧体(「です」)が後節する言い方は「もはや誤用とはいえないのが実情」だとした上で、過去形については、「美しいでした」よりも「美しいかったです」と言

うほうが一般的だとしている。

段階2~3では、前述の取り立ての「は」に加え、バスケット内のサンドイッチを指して不適切に使われる「食物」(バスケットの中の食物(CCB014) / 食物は食べ終わりました(CCB005))や、接続助詞の「て」を含むテイル形(待っていました(CCB006) / 相談していました(CCB012))などが特徴的に用いられる。また、上級相当の段階4になると、引き続き接続助詞の「て」が多用され、とくに、テクル形が使用される(おなかもすいてきました(CB016) / ワンちゃんは出てきました(CCB007))。

以上を整理すると、B-JASの段階1~4においては、(1)「です」の過剰使用や「でした」の不適切使用例の減少、(2)取立て表現の増加、(3)アスペクト形式(テイル・テイク)の獲得、といった方向性が認められる。

ここで、I-JASの特徴語と比較すると、段階1については、丁寧体語尾(「ます」または「です」)の使用という点で一定の関連性が認められるが、単語のレベルで一致しているものはない。また、段階2~4についても、テイル・テイク形の構成成分としての接続助詞の「て」を除くと、個別語レベルでの重複は見られない。I-JASデータの観察から演繹された5つの発達傾向のうち、B-JASにおいて明確に再現されたものは、段階4におけるアスペクト形式の獲得だけで、そのほかについてははっきりした一致は認められなかった。RQ1およびRQ2で見たように、総語数や品詞頻度といったマクロ的なレベルではI-JASとB-JASの調査結果は高い一致度を示したわけであるが、個別語のレベルになると、両コーパスの調査結果の重なりは限定的なものとなることが示唆された。

#### 4.4 RQ4 段階分類

学習者8群と、10品詞の頻度に対して対応分析を実施したところ、第1次元の寄与率が78.7%、第2次元の寄与率が8.1%で、上位2次元で全体の8割以上が説明された。また、学習者8群と、30動詞の頻度に対して対応分析を実施したところ、第1次元の寄与率が38.6%、第2次元の寄与率が21.7%で、上位2次元で全体の6割以上が説明された。第1次元を横軸、第2次元を縦軸と散布図を作成したところ、図5-6の結果を得た。



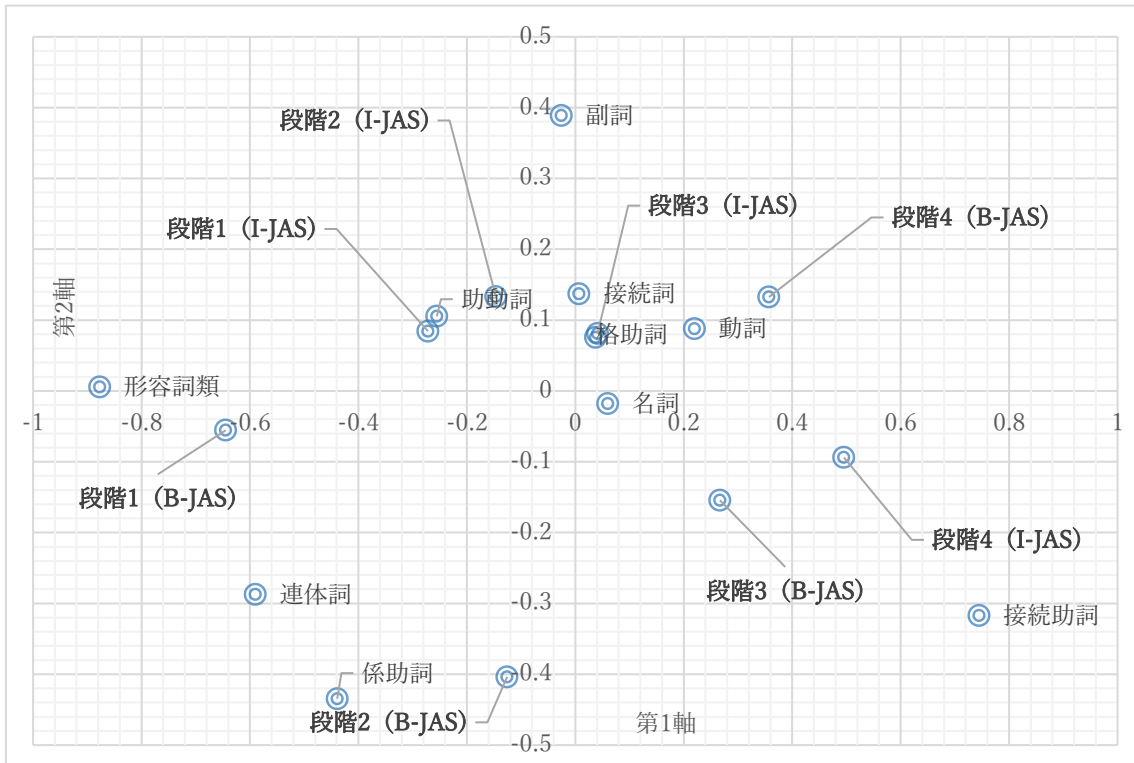


図 5 主要品詞と学習者 8 群の関係性

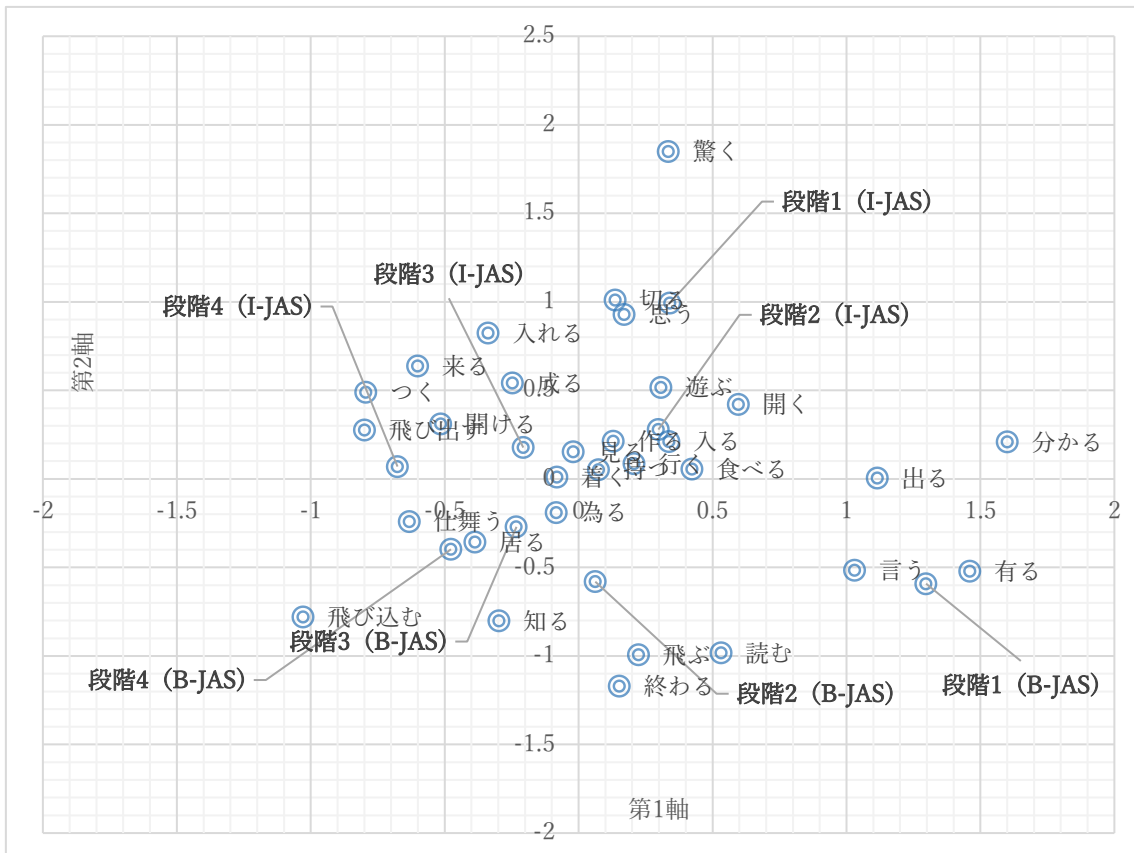


図 6 主要動詞と学習者 8 群の関係性

まず、散布図に見られるマクロ的な変化のパターンについて概観しておこう。図 5-6 に示すように、10 品詞、30 動詞のいずれを手掛かりとした場合であっても、I-JAS、B-JAS とともに、優先軸となる次元 1（横軸）上で、等しく、段階 1, 2, 3, 4 の順で布置されている。

（図 5 と図 6 で向きは逆転しているが、対応分析で得られた布置図における方向は相対的なものである）。このことは、品詞タイプ、あるいは、特定品詞に含まれる個別語セットというレベルで集約を行った場合、I-JAS であっても B-JAS であっても、発達の過程はほぼ同じように取り出せることを示す。

しかしながら、I-JAS と B-JAS がすべてにおいて同じ結果を示すわけではない。実際、次元 2（縦軸）上で、I-JAS と B-JAS は上下に分離されており、各段階を特徴づける品詞や動詞にも違いが認められる。表 6 は、I-JAS と B-JAS における各段階の近傍に布置された品詞と動詞を整理したものである。

表 6 対応分析における I-JAS、B-JAS の各段階の近傍に位置する品詞・動詞

	品詞		動詞（形態素）	
	I-JAS	B-JAS	I-JAS	B-JAS
1	形容詞	連体詞	開く、食べる、入る	有る、言う、読む
2	助動詞	係助詞	驚く、遊ぶ、行く、思う	飛ぶ、終わる
3	格助詞、接続詞、副詞	名詞	見る、着く、成る、入れる、開ける	為る、知る
4	接続助詞	動詞	来る、つく、飛び出す	居る、仕舞う、飛び込む

表 6 が示唆するのは、具体的な品詞や語彙項目といったミクロのレベルになると、I-JAS と B-JAS から演繹される変化のパターンが、一部の重複（たとえば、段階 4 におけるアスペクト形式のテクルやテイル、また、複合動詞の使用）を除き、完全には一致しないという事実である。このことは、横断データか縦断データかのいずれか一方のみを根拠として L2 習得の議論を行おうとする場合、見落とされてしまう要素がありうることを示唆する。

## 5. まとめ

### 5.1 知見の整理

本稿では、横断コーパスである I-JAS と縦断コーパスである B-JAS から取り出された CLJ の日本語習得過程の一致度を実証的に検証してきた。

まず、RQ1（総語数）については、I-JAS、B-JAS とともに、語数は 80~100 語に収まり、段階ごとに総語数は増加するものの、増加幅は段階があがるにつれて減少する、という同一のパターンを示すことが確認された。

また、RQ2（品詞頻度）については、I-JAS、B-JAS とともに、名詞・動詞・格助詞・接続助詞ともに段階が進むにつれて増加する一方、係助詞頻度は減少する、という共通のパターンを示すことが確認された。

次に、RQ3（特徴語）については、I-JAS からは、基本名詞の誤用の減少→口語的表現の減少→代名詞表現の多様化→アスペクト形式や抽象的场所表現の獲得、といった発達パ

タンが見られるのに対し、**B-JAS** からは、「です」や不適切な「でした」の減少→取立て表現の増加→アスペクト形式の獲得、といったパターンが観察され、段階 4 のアスペクト形式（テイル、テイク）の獲得を除くと、2 種類のコーパスから導かれる段階別特徴は必ずしも一致しないことが示された。

最後に、**RQ4**（段階分類）については、**I-JAS**、**B-JAS** ともに、4 段階が完全に同じ順序で布置されるものの、個々の段階を特徴づける具体的な品詞や語彙項目は一致しないことが示唆された。

以上、4 つの研究設問の検討を通して、総語数や主要品詞別頻度といったマクロレベルの変化に関しては、横断・縦断データともに、ほぼ同一の結果が得られるが、個々の品詞や語といったミクロレベルに踏み込むと、異なるパターンが析出される可能性が示された。前者は、横断コーパスを用いた疑似縦断分析が一定の妥当性を持つことを、後者は、横断コーパスと縦断コーパスの統合的使用が必要であることをそれぞれ示していると考えられる。

コーパスに限らず、第 2 言語習得研究全般における横断・縦断データの関係について、**Gass, Behney, and Plonsky (2020)** は、2 種類のデータがそれぞれ長所・短所を持ち、かつ、それらが相補的な関係であることを強調している。横断データを用いた疑似縦断分析は、多くの学習者から得られる知見を一般化しやすい一方、群を対象とした分析を行うため、個々の学習者の詳細な背景を議論しにくい。また、習熟度の比較で発達が観察できるとする仮説の妥当性も完全には証明されていない。これに対し、縦断データを用いた分析は個々の学習者背景を加味した議論が可能であるが、サンプルが少ないため、分析はケーススタディ的・記述的・質的・状況説明（ナラティブ）的になりがちで、得られた知見の一般化は行いにくい。また、産出データが当該学習者の **L2** 知識の全体を反映しているかどうか曖昧である（pp. 19-22）。**Gass** らの指摘をふまえれば、2 種のコーパスの統合的使用のメリットはいっそう明白となろう。

## 5.2 課題と展望

もっとも、本研究には、さらに検討すべき点も多い。ここでは 2 点に絞って言及する。1 点目は、コーパス比較における特殊要因の影響についてである。今回は、学習者の母語・地域とタスクの両方をそろえることで、横断・縦断というデータ収集方法以外の要因を可能な限り統制して比較を行ったわけであるが、**B-JAS** のデータは特定の 1 大学のみで集められており、当該大学における教材・教員・教授法といった特殊要因がデータに影響している可能性は残る。**B-JAS** を研究の主資料に使う場合は、データから見出される段階別の変化が自然な **L2** 発達の結果なのか、あるいは、それぞれの時期における特定の指導の結果なのか、慎重に見極める必要があるだろう。

2 点目は、分析対象とする学習者単位の妥当性についてである。今回は、**I-JAS** と **B-JAS** ともに、**CLJ** を 4 群に分けて群の単位で変化を比較したわけであるが、群に見られる傾向が群内の個人に見られる傾向を正しく代表しているかどうかははっきりしない。

そもそも、学習者コーパス研究において、分析の単位を群（例：ドイツ人日本語学習者、中級学習者、教室環境学習者など）とすべきか個人とすべきかは悩ましい判断となる。前者は議論の一般化が行いやすい一方、個人差を無視しているという批判の余地がある。後者は個人特性を丁寧に観察できる一方、知見の一般化が難しいという問題がある。伝統的に、大規模な横断コーパスでは集約的・量的な分析が、小規模な縦断コーパスでは個別的・質的な分析が好まれてきたわけであるが、この枠組みをそのまま I-JAS と B-JAS に適用することには慎重さが必要だろう。たとえば、規模の点について考えてみよう。1,000名の学習者データを包含する I-JAS が大規模で、17名のデータを包含する B-JAS が小規模であることは自明に見えるが、I-JAS で、たとえば、タイ語を母語とする学習者を習熟度で4群にわけて分析する場合、1群あたりの人数は12人前後となる。一方、B-JAS は17人からデータを8回取っているので、延べで言えば136名分とも言える。この点をふまえると、I-JAS は大規模横断データだから群単位で、B-JAS は小規模縦断データだから個人単位で調査すべきだ、とは言いきれない。実際、どちらのコーパスであっても、群分析・個人分析はともに可能であろう。

群データの傾向と、群内の個人データの傾向が一致しない可能性に関して、ここで、B-JAS の17名の学習者による4年間（Y1～Y4）のSW1の総語数の変化を見ておきたい（図7）。

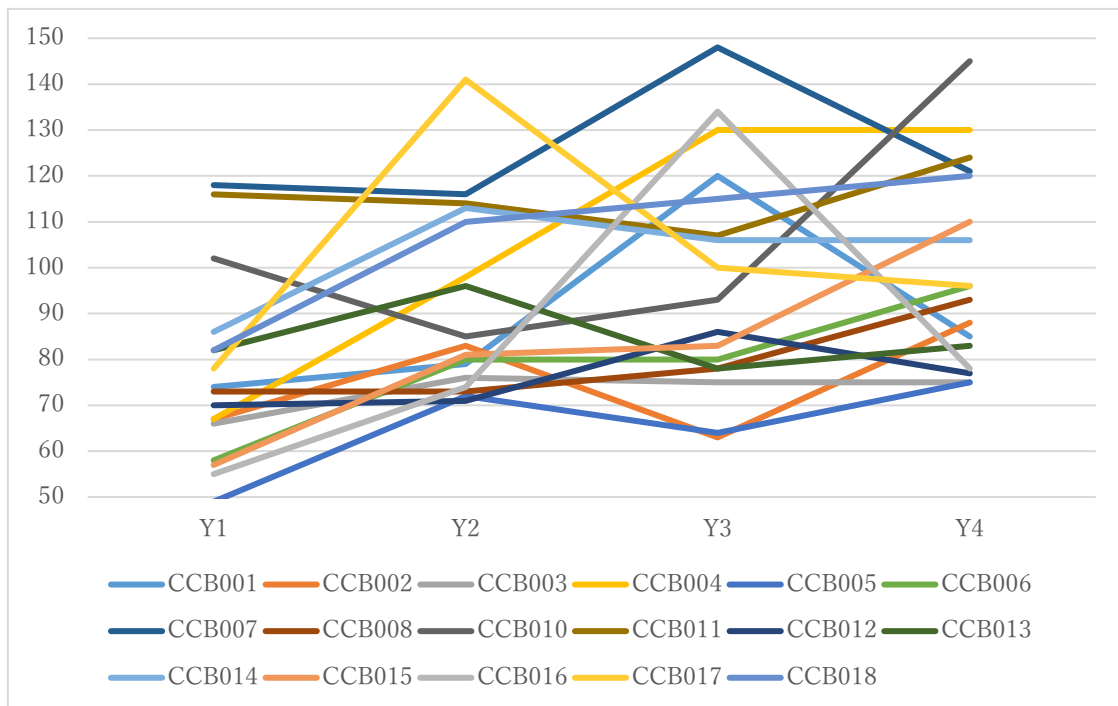


図7 B-JASの17名の参加者のSW1総語数の変化

4.1節で示したように、全体を一群と見なせば、段階1～段階4の間で増加のパターンが現れるわけであるが、総語数といったごく基本的な観点であっても個体差が非常に大きいことに気づく。これとまったく同じことは、実はI-JASにもあてはまる。

この点をふまえると、I-JAS や B-JAS を使った習得研究においても、個人単位の観察結果を組み込みながら群としての特性を議論していくという方向性が考えられるだろう。個人単位で分析を行う場合、要因（習熟度や学習段階）と結果（産出量や頻度など）の関係性を示すモデルを人数分作ることになるが、これだと結果の集約はできない。そこで、ターゲットとする要因を固定効果、個人差要因を変量効果（ランダム効果）とみなして、両者を同時に組み込んだモデルを推定するのである。これを混合効果モデル（mixed-effect model : MEM）と呼ぶ。学習者コーパス研究への MEM の適用はいまだ一般的ではないが、海外では、縦断コーパスのデータ処理に混合効果モデルを適用する実践もなされている（Paquot, Naets, and Gries, 2021）。こうした手法をうまく使っていけば、I-JAS か B-JAS か、あるいは、群調査か個人調査か、といった択一的選択を行うことなく、それらを統合した分析が可能になるだろう。

I-JAS と B-JAS のように、完全に同一のデザインで横断・縦断的にデータを集めるという試みは内外ともにほとんど先例がない。2 つのコーパスの革新性と、世界の学習者コーパス研究における意義、また、両コーパスの価値を引き出す、効果的で妥当性の高いデータ分析手法の開発の必要性を強調して本稿を閉じたい。

#### 注

本研究は、国立国語研究所のプロジェクトによる成果である「多言語母語の日本語学習者横断コーパス（I-JAS）」および「北京日本語学習者縦断コーパス（B-JAS）」を利用して行われたものである。

#### 文 献

- Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: Introductory course*. (5<sup>th</sup> ed.). Routledge.
- Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson, S. (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). John Benjamins.
- Ishikawa, S. (2017). Learners' acquisition and use of L2 Japanese vocabulary: Influence of L1 backgrounds and L2 proficiency levels: A learner corpus-based analysis. 『第二言語としての日本語の習得研究』 20, 10–27.
- Johnson, K., & Johnson, H. (Eds.). (1999). *Encyclopedic dictionary of applied linguistics*. Blackwell Publishing.
- Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 379–400). Cambridge University Press.
- Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 309–331). Cambridge University Press.

- Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, 36(1), 94–120.
- Vyatkina, N., & Cunningham, D. J. (2015). Learner corpora and pragmatics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 281–305). Cambridge University Press.
- 石川慎一郎 (2018a). 「L2 日本語語彙の習得プロセスについて—LARP at SCU コーパスを用いた台湾人学習者による日本語語彙運用の時系列分析—」『統計数理研究所共同研究リポート』 400, 1–18.
- 石川慎一郎 (2018b). 「中国語母語の日本語学習者の発話における使用語彙の変遷：発達段階の差と個体の差をめぐって」『国立国語研究所第4回学習者コーパス・ワークショップ & シンポジウム「第2言語習得における語彙の役割」予稿集』 62–76.
- 石川慎一郎 (2020). 「日本語・中国語・韓国語・英語母語話者の日本語発話における形容詞使用実態—I-JAS に基づく調査—」『国立国語研究所日本語学習者コーパス「I-JAS」完成記念シンポジウム予稿集』 27–34.
- 石川慎一郎 (2021a). 「絵描写作文課題における L2 日本語学習者の動詞使用と習熟度の関係—I-JAS の SW1 課題データの計量的概観—」『統計数理研究所共同研究リポート』 444, 1–22.
- 石川慎一郎 (2021b). 「韓国学習者の日本語動詞獲得モデル：学習者総体モデルとの比較—『多言語母語の日本語学習者横断コーパス』の絵描写作文を用いた検証—」『日本語教育研究』（韓国日語教育学会） 56, 37–54.
- 教育出版株式会社編集局 (n.d.). 「ことばのてびき Q30 : 『美しいです』『大きいです』という言い方は正しいか」 <https://www.kyoiku-shuppan.co.jp/textbook/chuu/kokugo/guidanceq030-00.html>
- Paquot, M., Naets, H., & Gries, S. Th. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 122–147). Cambridge University Press.
- 迫田久美子(2020). 「I-JAS 誕生の経緯」迫田久美子・石川慎一郎・李在鎬(編著)『日本語学習者コーパス I-JAS 入門：研究・教育にどう使うか』(pp.2–13). くろしお出版
- 迫田久美子・石川慎一郎・李在鎬 (編著). (2020). 『日本語学習者コーパス I-JAS 入門：研究・教育にどう使うか』くろしお出版.
- 趙麗雯 (2015). 「学習者コーパスに見られる『テイナイ』の使用順序—縦断的・横断的観点から」『日本語／日本語教育研究』 6, 79–96.
- 李在鎬・小林典子・今井新悟・酒井たか子・迫田久美子 (2015). 「テスト分析に基づく『SPOT』と『J-CAT』の比較」『第二言語としての日本語の習得研究』 18, 53–69.

## 介護分野における専門用語の平易化に向けた語彙リストの構築

黄 海洪（京都大学人間・環境学研究科）<sup>†</sup>

金丸 敏幸（京都大学国際高等教育院）

### The Construction of a Vocabulary List for the Simplification of Terminology in the Field of Nursing Care

Huang Haihong (Graduate School of Human and Environmental Studies, Kyoto University)

Kanamaru Toshiyuki (Institute for Liberal Arts and Sciences, Kyoto University)

#### 要旨

われわれは、介護分野において社会になじみの薄い専門用語を、平易な日本語（Plain Japanese）という考えに基づいて誰もが理解できる言葉へと言い換える語彙リストを構築した。言い換え対象となる語は、次の2段階で選定した。まず、介護福祉士国家試験を元にした介護試験コーパスを構築し、現代日本語書き言葉均衡コーパスとの比較を行って介護コーパスの特徴語を抽出する。次に、NTT 単語親密度データベースとクラウドソーシングを用いて、抽出された特徴語の単語親密度を調査する。調査によって、一般の方と専門家との間で親密度の差が1以上あった語を対象語とする。これらの結果、言い換える対象語は73語となった。その後、介護分野における日本語教育に知見のある4名の専門家の協力を得て、用途に応じた3種類の言い換えを作成した。本リストは、介護分野への理解の助けとなるほか、今後増加が見込まれる外国人介護人材への日本語教育にも活用できると考えられる。

#### 1. はじめに

専門用語は、特定の専門分野や業界でのみ使用される特別な用語が多く、特定の概念を厳密に表現するために用いられる。しかし、専門用語を理解するためには、該当専門分野の知識が必要であり、一般の人にとってはなじみの薄い存在となることが多い。

介護の仕事は、特に医療や福祉と密接に関わるため、医師や看護師、社会福祉士など、異なる専門職が連携を図りながら業務を遂行する、いわゆる多職種連携が必要となる場合も少なくない。広井（2011）によれば、介護分野には医療や福祉、心理や教育などの分野とも深く関連があり、幅広い知識が必要であると指摘している。多くの介護現場では、介護福祉士だけでなく、医師、看護師、作業療法士、管理栄養士など、様々な職種の専門家が利用者のために協力しながら支援を行っている。こうして専門家同士が連携して情報を共有したり、ケアの方針を一緒に確認し、決定したりすることが求められる。介護現場では、専門的な業務の遂行において、効率的かつ円滑なコミュニケーションを行うために、厳密な専門用語が多く用いられることが一般的である。

介護用語の一部は明治期に造語された難解な言葉が伝統の踏襲として使用し続けられている。明治時代の医学用語が看護学へと伝わり、さらに介護の分野にも広がっていたという脈絡を辿り、時代を超えて継承されてきた（遠藤, 2017）。

介護福祉士国家試験（以下：国家試験）には難しい介護用語が多用されている。EPA（経済連携協定）に基づき、来日した介護福祉士候補者は国家試験の受験が義務付けられている。

---

<sup>†</sup> [huang.haihong.46x@st.kyoto-u.ac.jp](mailto:huang.haihong.46x@st.kyoto-u.ac.jp)

メディアの報道によって、国家試験で用いられる日本語の難しさが、世間の注目を集めるようになり、社会問題として認識されるようになった。

国家試験だけでなく、遠藤（2013）は介護分野の教科書を調査し、難解な漢語を言い換えるべきだと提案している。また、遠藤（2019）は、介護用語のわかりにくさについて調査分析を行った。その結果、介護用語の多くは聞いても理解しにくい上、その語構成から意味を類推するのも難しいという問題点を明らかにした。これにより、介護分野の専門用語は介護の現場以外の人には理解が困難であることが明らかとなった。

介護分野では、財政難や離職、介護虐待といった問題が重なり合い、介護職の専門性と魅力が十分に評価されていないのが現状である。厚生労働省（2021）の推定によれば、2025年には約32万人の介護人材が不足するとされている。多くの専門用語の理解が難しいことも、介護職の選択を避ける一因と考えられる。そのため、介護分野で使用される難解な言葉を、誤解を招かないような形でわかりやすく言い換える取り組みが求められている。

専門用語の言い換えは重要であるが、あらゆる専門用語を言い換えることは時間やコストの面からも大きな負担となる。佐藤他（2004）によると、人間が持っている語彙の範囲は、個人が所属しているコミュニティによって異なるとされる。人々は日常生活のコミュニケーションに必要な基本的な語彙を共有しているが、それ以上の特殊な語彙の把握は、個人やコミュニティに依存する。

専門分野のコミュニケーションを行う上で、優先度の高い語彙を選択し、それらの語彙から言い換えを行うのが合理的である。優先度の高い語彙を選定する際には、一般的に、客観的に示すことのできる指標として出現頻度が用いられる。頻度の高い語は繰り返し出会う可能性が高いと考えられ、外国語学習などの語彙リストの作成でも使用されている。しかし、専門用語は語彙全体から見ると相対的に頻度が低い。そのため、頻度のみを指標として専門用語の中で重要性を評価する場合には、あまり有効ではない。特に、言い換えの優先度の高い専門用語は、一般になじみの薄いものが多い。そこで、単語の親密度という指標に注目する。

横川（2006）によれば、単語の親密度とは、ある単語にどれだけなじみがあるか、すなわち、人がその単語に対してどの程度よく見聞きすると感じているかを示す指標である。つまり、単語親密度は人がその語がどの程度出現すると思っているかという心的出現頻度を示すもので、単語の主観的特性の1つである。単語親密度は、頻度と同様に、またはそれ以上に単語認知の容易さに深く関わると考えられる。このため、心理言語学の研究において、単語認知の心理実験の統制要因としてよく用いられている。金澤（2020）は「親密度は頻度と似て非なる構成概念であり、頻度とは別に主観的指標としての親密度を検討することは有意義である」（p.25）と述べている。ほかにも、単語親密度の重要性は多くの先行研究で指摘されている。Richards（1970）も出現頻度の高い語は機能語や一般的な意味の語が多く、学習に役立つ具象名詞が少ないといった問題点を指摘し、語彙表を作成する際には、単語のなじみ度を示す概念として単語親密度を考慮に入れる必要性を提唱している。Connine 他（1990）は、親密度に焦点を当てることで、頻度だけを検討しては検出されない単語認知効果を確認できたと報告している。また、Lewellen 他（1993）は、親密度の違いは語彙処理の効率性の違いに影響を与えることを示している。さらに、寺田・田中（2008）は高い親密度を持つ単語が必ずしも頻繁に出現するわけではないと指摘している。

以上のことから、専門用語に対する単語親密度の調査の必要性が高いことがわかる。つまり、客観的指標の頻度だけでなく、主観的指標としての親密度も検討することが重要であ



る。このような専門用語の調査と分析を通して、介護分野の理解を妨げている難解な語彙を抽出し、それらの言い換えを行って当該分野への理解や認知を促進することで、多くの人々が介護により関心を持つ社会を構築することが求められている。

## 2. 先行研究

専門用語の平易化に関する取り組みは、行政、マスコミ、医療、法律など、国内外の様々な専門分野で行われている。

日本国内の取り組みとして、「コージェネレーション」や「セットバック」のようなわかりにくい外来語を理解しやすくするために、国立国語研究所の「外来語」委員会は2000年代の初頭に4回にわたって外来語をよりわかりやすい言葉に言い換える提案を公表した。マスコミ業界の取り組みとして、NHKが運営するウェブサイト「NHK News Web Easy」は、一般的なニュースをわかりやすい日本語に書き換えている。このサイトは、日本語学習者や日本の小中学生向けに日々情報発信をしている。書き換えの主な特徴として、難しい表現や専門用語を避け、基本的な文法と語彙を用いて書かれている。また、振り仮名を付けたり、関連する画像や音声を添えたりといったも施されている。しかし、専門用語の使用が制限されているため、提供される情報の精度と緻密さは限られることもある。

医療分野では、患者が自身の健康状態や治療方法を正確に理解するためには、難しい医療専門用語を平易な言葉にすることが求められる。この課題への対策として、「ヘルスリテラシー」という概念が導入され、重要な役割を果たしている。国立国語研究所の「病院の言葉」委員会(2009)は、病院などの医療機関で使われている言葉がわかりにくい問題を解決するために、医療分野の専門家と言語分野の専門家を集めてわかりやすくする取り組みを進めた。患者が理解しやすい説明をするために、医療従事者はどのような言葉をどのように使えば望ましいのかについて3種類に分けて提言を発表している。

法律の分野では、専門的な法律用語で書かれた複雑な法令を一般の人々が理解できるようにすることが重要である。この課題に対処するため、海外では、法律文書を平易な言葉で書くという運動がアメリカなどの国で行われている。具体的には、*legalese* と呼ばれる冗長な法律用語や専門的なジャーゴンを避け、明瞭かつシンプルな言葉を用いることを目標としている(岡部, 1983)。平易な表現を用いて伝えることで、法律の理解を向上させることを目指している。

専門分野における言葉の平易化の動きは、英語圏においては Plain English の運動まで発展している。角(2020)によれば、アメリカの多くの行政や企業は、LEP (Limited English Proficiency) を持つ人のために、文章をできるだけ Plain English の原則に従って平易化する取り組みを実施している。Plain English はメッセージを読者に容易に伝えられる様式で書かれることを目指すものである。

Cutts (2020) は Plain English のポイントとして以下の点を挙げている。

- ・テキストの内容が十分に明確であること
- ・文章の構造とデザインが優れていること
- ・子供っぽい言葉を避けること

Bailey (1990) は平易な文章は大仰な言葉 (*impressive words*) ではなく、普通の言葉 (*ordinary words*) で書くことを提唱している。藤本 (2002) は Plain English へ書き換える際には「簡潔でわかりやすい単語を使用し、文法や文のトーンを考慮すると、構成も直す必要がある」と指摘している (p.97)。

Plain Japanese は Plain English の運動から影響を受けた国際的に通用するコミュニケーション方法である。Plain Japanese と「やさしい日本語」(庵, 2016) は異なり、伝える情報の内容を削減することなく、効果的な情報伝達が目標とされる。つまり、読み手に誤解なく内容が伝わることを重視している。

Plain Japanese の主な特徴として、「情報を限定せずに伝えること」「日本語を第一言語とする人でも違和感なく使えること」の2点が挙げられる。ただし、ここで強調されるべきことは、Plain Japanese は日常的な日本語とは異なり、高度に専門化された言語であるということである。

要するに、専門用語の平易化は単に難解な言葉を簡単な言葉に置き換えるだけではない。それは情報の伝え方を見直し、最も適切な表現を見つけるという大切なプロセスである。このように、専門用語の平易化は言葉の置き換え以上の意味があり、より良い情報伝達のために行われるものだと言える。

### 3. 研究の目的

本研究は、Plain Japanese の考え方にに基づき、介護分野で使用される難解で一般の人々にはあまり知られていない専門用語を、平易で理解しやすい言葉へと言い換える語彙リストの構築を目指している。介護分野のコーパスデータを収集し、さらに介護の専門家への聞き取り調査の結果を資料として、頻度と親密度の両面からアプローチを行い、介護分野で使われる言葉をより身近に感じられる表現へ言い換えることにより、介護分野への興味や関心を高め、理解を促進することを目的とする。

また、国内生産年齢人口の減少やグローバル社会における多様な人材確保の観点から、外国人介護人材の活躍への期待が高まり、受け入れを拡大する政策が次々と打ち出されている。外国人介護人材の受け入れが拡大に伴い、介護のための日本語教育の必要性が年々高まっている。専門用語の言い換えは、今後増加が予想される外国人介護人材の教育にも有益である。本研究は、「職業目的の日本語 (JOP)」の立場から、介護の日本語を学ぶ外国人介護人材の語彙学習にも役立てられるように、介護用語の出現頻度と単語親密度の調査を通じて、根拠ある日本語教育へと繋げていくことを目指す。

### 4. 研究方法

#### 4.1 言い換え対象語の選定

##### 4.1.1 介護分野の専門用語の抽出

寺嶋 (2009) は、専門分野のコーパスと大規模コーパスである現代日本語書き言葉均衡コーパス (以下: BCCWJ) を比較することで、統計的に偏りのある特徴語を抽出できると提唱している。本研究では、BCCWJ における出現頻度と介護コーパスにおける出現頻度を比較し、介護専門用語を抽出する手法を採用する。

介護専門用語の抽出にあたっては、Dunning (1993) が提唱した統計的指標である対数尤度比 (Log Likelihood Ratio: LLR) を採用した。LLR は、コーパス言語学などで特徴語を抽出するために広く用いられている。LLR を採用した理由は、Dunning (1993) が指摘したとおり、テキストサイズが大規模であろうと小規模であろうと、妥当な値を示すという特徴があるからである。

LLR は下記の式で求められる。

$$LLR = \frac{a \log aN}{(a+b)(a+c)} + \frac{b \log bN}{(a+b)(b+d)} + \frac{c \log cN}{(a+c)(c+d)} + \frac{d \log dN}{(b+d)(c+d)}$$

ただし、

$a$  : 国家試験コーパスでの単語の出現頻度

$b$  : BCCWJ での単語の出現頻度

$c$  : 国家試験コーパスの延べ語数  $-a$

$d$  : BCCWJ の延べ語数  $-b$

である。

介護専門用語の抽出は、以下のような手順で行った。

まず、過去 10 回分 (第 24-33 回) の国家試験の内容をテキスト化し、分析対象とした。その後、形態素解析器 MeCab (Version 0.996) と形態素解析辞書 ComeJisyo (Utf8-3) を用いて、形態素解析を行った。形態素解析によって得られたリストと BCCWJ 長単位語彙リストを比較し、対数尤度比検定を行った。最後に、LLR 順の上位の語から、4 冊の用語集のうち少なくとも 3 冊に登録された語を対象に、漢語の上位 250 語とカタカナ語の上位 50 語を採用した。なお、漢語については、長単位語のうち、文字数 2 から 4 の語に限定した。

調査語彙を選定する際に参考とした用語集は以下の 4 冊である。これら 4 冊を選定した理由は、近年改訂されたり、または新たに出版されたりしたもので、新しい介護用語が反映されているからである。

- ① 中央法規出版 (2015) 『介護福祉用語辞典 七訂版』 (漢語 2,279 語, カタカナ語 423 語)
- ② 福祉教育カレッジ (2016) 『イラストでみる介護福祉用語事典 第 6 版』 (漢語 1,864 語, カタカナ語 352 語)
- ③ 佐藤富士子 (2017) 『現場で役立つ! 【早引き】介護用語辞典』 (漢語 2,011 語, カタカナ語 502 語)
- ④ 株式会社ヘルスケア総合政策研究所 (2019) 『ポケット判介護職員のための重要用語集 第 2 版』 (漢語 691 語, カタカナ語 153 語)

#### 4.1.2 介護用語の単語親密度調査

単語親密度調査の先駆的な先行研究として、天野・近藤 (1999) の調査がある。天野・近藤 (1999) は『新明解国語辞典』の見出し語約 7.7 万語を調査対象語として、「なじみがな

い」から「なじみがある」まで、7 段階の評定値を 18 歳以上 30 歳未満の男女、各 20 人から収集したものである。調査結果は、1999 年刊行の NTT データベースシリーズ『日本語の語彙特性』で調べることができる。なお、調査協力者への統制処置として、調査参加者は漢字の読み能力テスト「百羅漢」で 60 点以上を獲得した人に限定している。

浅原 (2019) は、単語親密度のアンケート調査をクラウドソーシングにより行った。クラウドソーシングは、オンラインの調査フォームを用いて、ウェブ上で評定を募るものである。調査対象の語彙は「分類語彙表」の見出し語約 10 万語である。調査協力者は 20 歳以上の男女 3,392 人であった。調査では、なじみの程度を「知っている、書いているときに出現する、読んでいるものに出現する、話すときに出現する、聞くときに出現する」の計 5 項目

で評価する。単語親密度の判定は「全く知らない」から「よく知っている」の5段階評価によって行う。収集データは、ベイジアン・リニア・ミックスモデル (Bayesian Linear Mixed Model) を用いた統計処理によって単語親密度推定が行われた。

藤田他 (2020) の調査も浅原 (2019) と同様、クラウドソーシングによる調査である。「なじみがない」から「なじみがある」まで7段階の評定値を用いて、実験協力者の18歳以上35歳未満の男女52名によって評定結果を得ている。調査結果は、2021年刊行の『令和版単語親密度データベース』に収録されている。

陳 (2014) は台湾人日本語学習者の日本語単語親密度の特徴を明らかにするために、3,000語に及ぶ日本語単語親密度データベースを構築した。学習者が親密度を低いと思う単語は、主に難易度が高い語や専門分野の専門用語であることが明らかになった。

本研究の調査方法はクラウドソーシングを用いるインターネット・パネル調査 (吉岡, 2020) である。クロス・マーケティング社所有のアンケート回答パネルに登録されている回答者へ向けて配信を行い、設問数に応じて回答者に謝礼が支給される仕組みとなっている。スクリーニングの方法について、特定の設問において1つまたは任意の選択肢を2つ選んだ回答は無効とした。

調査票の作成、調査画面の確認などを研究者側で行い、実際の調査に関してはクロス・マーケティング社が行った。調査完了後、ローデータ、単純集計表 (GT表) が納品された。

判定尺度は、先行研究と同じ1 (まったくなじみがない) から7 (とてもなじみがある) までの7段階スケールを使用した。目標の実験協力者は介護福祉士50名、一般の方50名であった。協力者の数は、事前に、有意水準: 5%, 効果量: 0.3 (中程度), 検定力: 0.8 と設定し、検定力分析のためのソフト G\*Power を用いて算出した。

#### 4.2 言い換え語の抽出

言い換え対象語彙の決定に際して、令和版 NTT 単語親密度データベース (NTTDB) を参照して取捨選択を行った。令和版 NTT DB は、時代とともに変化する可能性のある平成版 NTT データベースシリーズの第1巻と第9巻に含まれるすべての語を再調査し、さらに新しい語を追加調査して構築されたもので、総数は16万語以上に及ぶ大規模なデータベースである。

選択の基準は、介護福祉士の親密度と一般の方および NTT DB の間に親密度との差が1以上である語を対象とした。

#### 4.3 平易な日本語への言い換え語

Plain Japanese の考え方にに基づき、以下に挙げるツールとウェブサイトを参考に、平易な日本語への言い換えの検討を行った。

- ・「SNOW T15 やさしい日本語コーパス」
- ・「伝えるウェブ」
- ・「やさしにちチェッカー」
- ・「やさしい日本語でまなぶ介護専門用語集」

しかしながら、専門用語の言い換えに関して、現存の機械翻訳ツールの利用には限界があることが明らかとなった。乾 (2007) によれば、言い換えは基本的に入力文の単語を同義語に変えることで実現できるが、同義語が常に交換可能であるとは限らない。そこで、解決策として、介護の日本語教育に知見を持つ専門家4名に言い換え案の作成を依頼し、専門家に

よる言い換え案をもとにして、新しい言い換え案を作成した。

## 5. 研究結果

### 5.1 言い換え対象語の選定結果

#### 5.1.1 介護分野の専門用語

ComeJisyo による解析結果と BCCWJ の長単位語彙表 (Version1.1) を用いて、対数尤度比検定により特徴的に出現する漢語表現の抽出を試みた。表 1 に対数尤度比の上位 20 語を示す。

表 1 国家試験の特徴語 (対数尤度比による上位 20 語)

順位	特徴語	順位	特徴語
1	利用者	11	認知症対応型共同生活介護
2	認知症	12	介護支援専門員
3	訪問介護員	13	介護者
4	訪問介護	14	居宅介護
5	介護福祉士	15	事業所
6	介護老人福祉施設	16	右片麻痺
7	障害者	17	日常生活
8	通所介護	18	生活支援
9	高齢者	19	介護老人保健施設
10	要介護	20	左片麻痺

ComeJisyo による解析結果に基づくと、10 回の試験で、延べで 33,248 語、異なりで 4,782 語の漢語名詞が抽出された。そのうち、二字漢語が一番多く、約 5 割を占める。三字以上の漢語表現も少なくなく、約 4 割を占める結果となった。対数尤度比順の上位 100 語からは、「訪問介護員」「国際生活機能分類」「日常生活動作」などの長単位の専門用語が得られた。本研究では、漢語の長さは 2~4 字に限定した。最終的に、介護用語集を参照し、LLR 順に配列された特徴度の高い語から、漢語の上位 250 語とカタカナ語の上位 50 語を採用した。

#### 5.1.2 介護用語の単語親密度調査結果

本調査の調査期間は、2022 年 11 月 24 日から 2022 年 12 月 1 日である。介護福祉士 151 名、一般の方 186 名による回答が得られた。スクリーニングの結果、有効サンプル数は、介護福祉士 51 名、一般の方 64 名となった。

表 2 に介護福祉士の高親密度語の上位 20 語を示す。続いて、表 3 には介護福祉士の親密度の低い語の下位 20 語を示す。

表 2 介護福祉士を対象とした高親密度語上位 20 語

順位	高親密度語	親密度	順位	高親密度語	親密度
1	認知症	6.31	11	褥瘡	5.86

2	介護	6.14	12	要介護者	5.82
3	要介護度	6.10	13	高齢社会	5.82
4	尿失禁	6.08	14	失禁	5.82
5	血圧	6.04	15	守秘義務	5.82
6	水分補給	6.04	16	誤嚥	5.78
7	頻尿	6.00	17	便秘	5.78
8	腰痛	5.90	18	清拭	5.75
9	体位変換	5.90	19	歩行器	5.71
10	看護師	5.90	20	側臥位	5.71

表3 介護福祉士を対象とした低親密度語下位20語

順位	低親密度語	親密度	順位	低親密度語	親密度
1	ウェルビーイング	2.35	11	アドボカシー	2.78
2	昇華	2.57	12	年少人口	2.84
3	インテーク	2.59	13	反動形成	2.84
4	ゴールドプラン	2.65	14	投影	2.92
5	児童委員	2.67	15	心理検査	2.92
6	防衛機制	2.69	16	読話	3.06
7	指文字	2.71	17	同一化	3.08
8	スーパービジョン	2.71	18	味蓄	3.08
9	グリーンケア	2.76	19	破傷風	3.08
10	エコマップ	2.76	20	発達課題	3.08

以下は一般の方の調査結果である。表4に一般の方の高親密度語の上位20語を示す。続いて、表15に一般の方の親密度が低い語の下位20語を示す。

表4 一般の方を対象とした高親密度語上位20語

順位	高親密度語	親密度	順位	高親密度語	親密度
1	水分補給	5.94	11	訪問介護	5.13
2	高齢社会	5.80	12	介護保険	5.11
3	認知症	5.78	13	視覚障害	5.08
4	食物繊維	5.70	14	知的障害	5.06
5	バリアフリー	5.59	15	要介護者	5.00
6	医療保険	5.58	16	ケアマネジャー	4.95

7	感染症	5.58	17	健康寿命	4.94
8	血糖値	5.31	18	要介護度	4.89
9	デイサービス	5.23	19	ヒートショック	4.84
10	骨粗鬆症	5.20	20	訪問看護	4.83

表5 一般の方を対象とした低親密度語下位 20 語

順位	低親密度語	親密度	順位	低親密度語	親密度
1	インテーク	1.52	11	感情鈍麻	2.03
2	ゴールドプラン	1.84	12	ウェルビーイング	2.08
3	中核症状	1.84	13	反動形成	2.08
4	グリーンケア	1.89	14	構音障害	2.11
5	応益負担	1.91	15	蠕動運動	2.16
6	対	1.97	16	常同行動	2.16
7	見当識	1.98	17	エコマップ	2.27
8	日内変動	2.00	18	防衛機制	2.28
9	スーパービジョン	2.02	19	喉頭蓋	2.30
10	感情失禁	2.02	20	ユニットケア	2.30

## 5.2 言い換え語の抽出

介護福祉士の親密度と一般の方および NTT DB の間に親密度の差が 1 以上の語を言い換え対象語として選定した。結果として、75 語が抽出された。75 語のうち、意味的に重複する 2 語を除外し、最終的に 73 語を言い換え対象とした。

## 5.3 言い換え語の結果

介護分野における日本語教育に知見のある専門家 4 名が提案した言い換えは、Plain Japanese の考え方に基づくと、次の 3 つのパターンにまとめられることが明らかとなった。

パターン 1：介護分野の日本語研究者 A 氏による言い換え

パターン 2：介護専門学校の日本語主任教員 B 氏による言い換え

パターン 3：介護の日本語の研究者と日本語教師の C 氏と D 氏による言い換え

以下に、それぞれのパターンの言い換え例を 5 例ずつ取り上げ、表 6 から表 8 に示す。

表 6 介護用語の言い換え例 (パターン 1)

No.	言い換え対象語	言い換え例
1	側臥位	横向き寝
2	仰臥位	仰向け寝
3	見当識	基本的な状況の理解

4	体位変換	身体的位置や姿勢を変えること
5	感情失禁	激しい感情表現が抑えられない状態

表7 介護用語の言い換え例 (パターン2)

No.	言い換え対象語	言い換え例
1	側臥位	体を横向きにして寝ている姿勢, 体の向きのこと。ベッドに寝ているときに使うことば。体が動かせない人もいるので, 側臥位にしたり, 仰臥位にしたりして体の向きを変えてあげないと, 皮膚が病気になってしまうことがある。
2	仰臥位	仰向け。背中を下にして寝ている姿勢。顔は上を見ている。
3	見当識	今日が何月何日かわかること, 家までの帰り道がわかること, 今誰と話しているわかることなど, 生活するときに必要なことがわかること。認知症の症状で, 「見当識障害」ということばがあって, (時間) 今日が何月何日かわからない→ (場所) 帰り道がわからない→ (人) 誰と話しているかわからないなどの順番でわからなくなってくる。
4	体位変換	体の向きを変えること。自分で動けない人の場合, 体の向きを変えないと皮膚の病気になってしまうから。
5	感情失禁	小さいことが原因で急に泣いたり, 笑ったり, 怒ったりすること。脳血管疾患や脳血管性認知症などの脳の病気が原因でなることがある。

表8 介護用語の言い換え例 (パターン3)

No.	言い換え対象語	言い換え例
1	側臥位	体を横向きに寝た状態のこと。右を下にすると右側臥位, 左を下にすると左側臥位という。寝ている人の着脱介助やリネン交換をする時の体位。
2	仰臥位	あおむけ (顔や胸を上に向ける) で寝た姿勢。長く続けると褥瘡がでやすいため注意が必要。
3	見当識	時間・場所・人など基本的な状況を理解する能力。※見当識障害になると, 時間・場所・人の順に把握が難しくなる。「今日は何月何日か」という時間, 「自分は今どこにいるのか」という場所, まわりの人との関係がわからなくなる。
4	体位変換	自分で体を動かすことができない人の体の向き (体位) を2時間おきに変えること。体位を変えないと, 血行が悪くなって身体の働きが悪くなったり, 同じところが圧迫されて痛くなったり, 血流が悪くなって皮膚



		に傷ができたりする。
5	感情失禁	気持ちのコントロールが難しくなり、小さなことで泣く・笑う・怒るなどの感情が、強く出てくること。脳の疾患や精神の疾患で見られる症状。

言い換えの整理には OpenAI の ChatGPT による要約結果を採用した。ChatGPT は大規模なテキストデータを学習する能力を持ち、対話型システムとして質問に対する答えを提供することができる。さらに、文章の生成、文章の要約、翻訳といった多岐にわたる自然言語処理タスクにも利用されている。Taecharungroj (2023) によれば、ChatGPT はエッセイやプロンプト、コードの作成などを行うことができ、多くの質問への対応が可能である。そのため、言い換え結果の要約といった自然言語処理タスクにおいては高精度のパフォーマンスを期待できることから、言い換えの整理に利用することにした。

以下に、「側臥位」に関する言い換えの整理を例にして、ChatGPT に送るプロンプトを示す。

側臥位に関する下記3つの言い換えを要約してください。

1. 横向き寝。
2. 体を横向きにして寝ている姿勢、体の向きのこと。ベッドに寝ているときに使うことば。体が動かせない人もいるので、側臥位にしたり、仰臥位にしたりして体の向きを変えてあげないと、皮膚が病気になってしまうことがある。
3. 体を横向きに寝た状態のこと。右を下にすると右側臥位、左を下にすると左側臥位という。寝ている人の着脱介助やリネン交換をする時の体位。

上記のプロンプトに基づいた ChatGPT の整理結果を以下の通りである。

側臥位は、横向き寝とも呼ばれ、体を横向きにして寝る姿勢です。ベッドで寝ている際に用いられ、体が動かせない人の場合は、側臥位と仰臥位を交互に変えることが皮膚病の予防になります。側臥位には右側臥位と左側臥位があり、寝ている人の着脱介助やリネン交換をする際にも使用される体位です。

生成した言い換え結果を一語ずつ、妥当性や正確性の面から確認し、言い換え表現として問題がないことを検証した。言い換えるべき語を主に使う10の場面を以下のようにまとめた。それらは、1. 受け入れ 2. 症状 3. 食事 4. 専門家間のコミュニケーション 5. 日常生活 6. 入浴 7. 認知症 8. 排泄 9. ベッド 10. 保険サービスである。

言い換えの結果は丁寧語に統一する。介護の現場では、利用者との関係で一定の社会的距離が必要である。上野 (2007) によれば、「丁寧語は、相手とのあいだに距離を置く技法である。丁寧語を使いつづけるかぎり、「わたしはあなたとこの距離を詰めるつもりはありませんよ」というメッセージが伝えることができる。これを社会学の用語で「儀礼的距離化」という」(p.206)。介護の現場での「儀礼的距離化」を保持するために、言い換えは丁寧語に統一した。

以下に、まとめた結果の一部を表9に示す。

表9 介護用語の平易化言い換え案

介護用語	使用場面	書き換え案
側臥位	ベッド	側臥位は、横向き寝とも呼ばれ、体を横向きにして寝る姿勢です。ベッドで寝ている際に用いられ、体が動かせない人の場合は、側臥位と仰臥位を交互に変えることが皮膚病の予防になります。側臥位には右側臥位と左側臥位があり、寝ている人の着脱介助やリネン交換をする際にも使用される体位です。
仰臥位	ベッド	仰臥位は、仰向け寝とも呼ばれ、背中を下にして顔を上に向けた寝姿勢です。この姿勢で長時間過ごすとうっ滞ができやすくなるため、注意が必要です。
見当識	認知症	見当識は、基本的な状況の理解力で、日常生活に必要な時間・場所・人の認識が含まれます。認知症の症状である「見当識障害」では、時間・場所・人の順に理解が難しくなり、日常生活に影響が出ます。
体位変換	ベッド	体位変換は、身体的位置や姿勢を変えることで、特に自分で動けない人にとって重要です。定期的に体位を変えることで、血行を良くし、圧迫や皮膚の傷を防ぐことができます。
感情失禁	認知症	感情失禁は、感情表現を抑えられない状態で、小さなことで急に泣いたり笑ったり怒ったりします。主に脳の病気や精神の病気が原因となり、感情コントロールが困難になる症状です。

最終的な言い換えの結果は、独立行政法人国際交流基金日本語国際センターが開設・運営しているウェブサイト「みんなの教材サイト」にてまとめて公開する予定である。

## 6. おわりに

専門用語の習得は専門分野の学習の基盤となる。多くの介護人材は、介護施設で働きながら、国家試験の受験準備を行なっている。国家試験の資格取得は、介護職のキャリアアップの一環として、介護人材の能力向上や待遇改善をもたらすだけでなく、介護職の魅力の増加にも寄与すると考えられる。働きながら介護福祉士の資格を取得するためには、効率的な学習が不可欠である。限られた時間内で、効率よく学習目的を達成するには、学習ニーズに合わせた教育内容が求められる。

介護分野の語彙教材は、教師の主観や経験に依存していることが多く、コーパスに基づく教材の開発は限られているのが現状である。こうした現状を打開するため、本研究ではコーパス調査とアンケート調査を組み合わせて実施した。調査結果をもとに、介護分野の専門用語を頻度と親密度に基づいて抽出し、それらを Plain Japanese で言い換えた語彙リストを作成した。結果として、一般的にはなじみが薄く、難解とされる 73 語の介護用語を言い換えることができた。本研究の成果により、学習者の介護分野の日本語学習の負担が軽減されることが期待されるほか、専門的な内容をやさしく学ぶサポートにも寄与すると考えられる。

介護の専門家がよく見聞きし、なじみのある言葉は、介護現場でよく使う重要な専門用語である可能性が高い。そのため、初学者にとって学習優先度が高いと考えられる。しかし、国家試験に出る専門用語の中には、専門家にとってもなじみのない言葉が含まれている。こ

それは、そのような専門用語が実際の介護の現場であまり使用されていないことを示唆している。そのため、それらの語は現場で触れるチャンスが少なく、自然に習得するのが困難であることが考えられる。それらの語については、教師側が学習者に明示的に提示して教えることが望ましいと思われる。専門分野の語彙には、専門家であってもなじみが薄いものが存在する。それゆえ、難解な専門用語を平易な表現にすることは、専門家の育成や専門概念の普及だけでなく、その分野に関わるすべての人々にとっても有益であると言える。

### 謝 辞

本研究は全国老人福祉施設協議会の助成を受けました。この場をお借りして、深く感謝の意を表します。本稿は、研究助成を受けた研究報告書の内容を加筆修正したものです。

### 文 献

- Connine M. Cynthia, John Mullennix, Eve Shernoff, and Jennifer Yelen (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16:6, pp.1084-1096.
- Edward P. Bailey (1990). *The plain English approach to business writing*. Oxford University Press.
- Jack Croft Richards (1970) A psycholinguistic measure of vocabulary selection. *International Review of Applied Linguistics in Language Teaching*, 8:2, pp.87-102.
- Martin Cutts (2020). *Oxford guide to plain English*. Oxford University Press.
- Mary J. Lewellen, Stephen D. Goldinger, David B. Pisoni, and Beth G. Greene (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General*, 122:3, pp.316-330.
- Thad Dunning (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, pp.61-74.
- Viriya Taecharungroj (2023) “What Can ChatGPT Do ?” Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing*, 7:1, p.35.
- 浅原正幸 (2019) 「クラウドソーシングによる単語親密度の推定」『言語処理学会第 25 回 年次大会発表論文集』, pp.45-48.
- 天野成昭・近藤公久 (編) (1999) 『日本語の語彙特性』三省堂
- 庵功雄 (2016) 『やさしい日本語—多文化共生社会へ』岩波書店
- 乾健太郎 (2007) 「自然言語処理と言い換え」『日本語学』 26:10, pp.50-59.
- 上野千鶴子 (2007) 『おひとりさまの老後』法研
- 遠藤織枝 (2013) 「わかりやすい介護用語をめざして: 介護教科書の語彙調査から」『ことば』 34, pp.73-87.
- 遠藤織枝 (2017) 「明治の看護用語と現在の難解な介護用語—その継承を検証する—」『ことば』 38, pp.102-123.
- 遠藤織枝 (2019) 「わかりにくく難解な介護用語の実際」『ことば』 40, pp.196-213.
- 岡部朗一 (1983) 「アメリカの「英語簡略化法」について—消費者とのよりよいコミュニケーションを目指して—」『時事英語学研究』 22, pp.40-56.
- 金澤祐 (編) (2020) 『フォーミュラと外国語学習・教育定型表現入門』くろしお出版
- 厚生労働省 (2021) 「第 8 期介護保険事業計画に基づく介護職員の必要数について」  
[https://www.mhlw.go.jp/stf/houdou/0000207323\\_00005.html](https://www.mhlw.go.jp/stf/houdou/0000207323_00005.html) (2023 年 9 月 27 日)

- 国立国語研究所 (2009) 『「病院の言葉」を分かりやすく提案』 国立国語研究所「病院の言葉」委員会
- 佐藤浩史・笠原要・金杉友子・天野成昭 (2004) 「単語親密度に基づく基本語彙の選定」『人工知能学会論文誌』 19:6, pp.502-510.
- 角知行 (2020) 『移民大国アメリカの言語サービス—多言語とくやさしい英語>をめぐる運動と政策—』 明石書店
- 陳相州 (2014) 「台湾人日本語学習者を対象とした日本語単語親密度データベースの構築」, 『比較文化研究』 111, pp.167-179.
- 寺嶋弘道 (2009) 「日本語教育語彙を選定するための統計的指標—尤度比検定, カイ 2 乗検定, イエーツの補正公式の特徴—」, 『Polyglossia : the Asia-Pacific's voice in language and language teaching』 17, pp.71-83.
- 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.
- 広井良典 (2011) 『ケアを問いなおす:「深層の時間」と高齢化社会』 筑摩書房
- 藤田早苗・小林哲生・山田武士・菅原真悟・新井庭子・新井紀子 (2020) 「小・中・高校生の語彙数調査および単語親密度との関係分析」『言語処理学会第 26 回年次大会発表論文集』 pp.355-358.
- 藤本淳史 (2002) 「Plain English の定義とその効用」『日本実用英語学会論叢』 2002:10, pp.91-103.
- 横川博一 (編) (2006) 『日本人英語学習者の英単語親密度 文字編: 教育・研究のためのデータベース』 くらしお出版
- 吉岡洋介 (2020) 「インターネット・パネル調査の利用可能性—大学生を対象とした調査事例を通して—」『社会学評論』 pp.50-64.

#### 関連 URL

BCCWJ 長単位語彙表 (Version 1.1)	<a href="https://clrd.ninjal.ac.jp/bccwj/bcc-chu.html">https://clrd.ninjal.ac.jp/bccwj/bcc-chu.html</a>
ComeJisyo (Utf8-3)	<a href="https://ja.osdn.net/projects/comedic/releases/74810">https://ja.osdn.net/projects/comedic/releases/74810</a>
MeCab 0.996	<a href="https://taku910.github.io/mecab/">https://taku910.github.io/mecab/</a>
NHK News Web Easy	<a href="https://www3.nhk.or.jp/news/easy/">https://www3.nhk.or.jp/news/easy/</a>
SNOW T15 やさしい日本語コーパス	<a href="https://www.jnlp.org/GengoHouse/snow/t15">https://www.jnlp.org/GengoHouse/snow/t15</a>
伝えるウェブ	<a href="https://tsutaeru.cloud">https://tsutaeru.cloud</a>
やさしい日本語でまなぶ介護専門用語集	<a href="https://yasanichi-kaigo.com/en/">https://yasanichi-kaigo.com/en/</a>
やさいちチェッカー	<a href="http://www4414uj.sakura.ne.jp/Yasanichi1/nsindan/">http://www4414uj.sakura.ne.jp/Yasanichi1/nsindan/</a>
単語親密度データベース	<a href="https://www.rd.ntt/cs/team_project/icl/lirg/resources/goitokusei/">https://www.rd.ntt/cs/team_project/icl/lirg/resources/goitokusei/</a>

## 日中対訳コーパスの構築と公開に向けて

宮本 華瑠 (大阪大学)

### Toward the Construction and Publication of a Japanese-Chinese Bilingual Corpus

Haru Miyamoto (Osaka University)

#### 要旨

昨今、公開された日中対訳コーパスには、北京日本学研究中心の『中日対訳コーパス』、情報通信研究機構の『NICT 多言語対訳コーパス』、JST・NICT 共同で構築された『アジア学術論文抜粋コーパス(ASPEC)』、『GSK 通訳データベース (JNPC コーパス) 日中・日西サブコーパス』などがあげられる。しかし、『中日対訳コーパス』は現在、個人・機関問わず対訳コーパスの入手はできなくなっており、『NICT 多言語対訳コーパス』は機械翻訳の研究またはシステム開発の一環として構築されたものでデータは非公開となっている。そして、『ASPEC』コーパスと『JNPC コーパス』は、専門用語など特殊用語の偏りが多く、広く一般的に用いられている言語使用を代表できるとは言い難い。以上のことから、日中対訳研究を行う際に利用できるコーパスは極めて限定的で、言語資源が乏しい状況に陥っていることが読み取れる。発表者は個人利用を目的に2009年から対訳文の収集を続けており、その成果物を個人利用だけでなく、オープンにすべきであると考えている。収集済みデータには、雑誌『Taiwan Panorama』約45万字、『聞く中国語』2018年1月～2023年7月(67冊)のデータ約242万字、『人民網』ニュース(日中対訳文)2014年7月～2023年8月現在のデータ約272万字が含まれる。本稿では主に、日中対訳コーパスの紹介(2節)、実用に向けた活用方法の紹介(3節)、著作権問題に関する示唆(4節)、データ公開に向けての告知事項(5節)、今後の課題(6節)について述べる。

#### 1. はじめに

本稿は日中対訳データの公開と実用に向けて、発表者が個人で収集を行った日中対訳コーパスデータを紹介するものである。発表の目的は、日中対訳言語資源の存在を広めると共に、現時点で乗り越えるべき問題点を開示し情報共有を促すためである。データは今後も収集を続け、バージョンはデータが整い次第定期的に更新を行う予定である。

言語資源の著作権に関しては法律専門家への業務委託を視野に入れており、現段階の課題は「軽微な使用」の壁を如何に乗り越えるべきかである。データの試作ではなく、収集済みデータ全文を配布するには、今後著作権法の更なる展開が鍵となってくる。

#### 2. 日中対訳コーパスの紹介

対訳データは二つの捉え方が可能である。即ち、起点言語だけをデータとして切り取った場合は「単一言語資源」、起点言語と目標言語がセットになっている対訳文の場合は言語間の対照分析ができるパラレルコーパスデータとして捉えることができる。そして、目標言語を対象にすれば不自然性を隠れ持った言語資源として考えることができる。本稿でいう日中対訳コーパスとは、日本語を中国語に訳した場合と、中国語を日本語に訳した場合の両方を含む。本稿では便宜上、日本語が起点言語の場合は「日→中」、中国語が起点言語の場合は「中→日」と記す。

## 2.1 既存の日中対訳コーパス

### 2.1.1 北京日本学研究中心の『中日対訳コーパス』

「中日対訳コーパス」は北京日本学研究中心で 2003 年に公開したコーパスである。格納されたデータには、文学作品が約 1130 万字、作品は日中それぞれ 22 篇と 23 篇となり、文学以外は約 575 万字、日本原作が 14 篇、中国原作 14 篇、共同 2 篇となる。その全作品名を次に示す。

#### 【日本原作：36 作品】

明日来る人、坊っちゃん、越前竹人形、布団、雁の寺、破戒、鼻、金閣寺、こころ、高野聖、黒い雨、野火、ノルウェイの森、羅生門、青春の蹉跎、飼育、死者の奢り、砂の女、徐陽、痴人の愛、友情、雪国、日本戦後名詩百家集、百言百話、ひとりっ子の上手な育て方、激動の百年史、日本経済の飛躍的な発展、心の危機管理術、近代作家入門、マッテオ・リッチ伝、日本列島改造論、日本国憲法、サラダ記念日、タテ社会の人間関係、適応の条件、五体不満足

#### 【中国原作：39 作品】

人大報告 96, 人大報告 97, 人大報告 98, 人大報告 99, 我的父亲邓小平, 我的父亲邓小平 2, 邓小平文选第一卷, 邓小平文选第二卷, 邓小平文选第三卷, 中日飞鸿, 毛泽东选集第一卷, 毛泽东文选集第二卷, 毛泽东选集第三卷, 毛泽东选集第四卷, 毛泽东传, 中日外交两个基本文件, 插队的故事, 盖棺, 丹凤眼, 轎轡把胡同 9 号, 关于女人, 活动变人形, 红高粱, 金光大道, 家, 轮椅上的梦, 呐喊, 彷徨, 青春之歌, 倾城之恋, 棋王, 人到中年, 人啊人, 上海的早晨 (上), 霜叶红似二月花, 天云山传奇, 小鲍庄, 骆驼祥子, 钟鼓楼

作品名から読み取れるように、「坊っちゃん」の発行は 1906 年、芥川の「羅生門」「鼻」は 1915 年と 1916 年、川端の「雪国」は 1948 年であり、中国原作も同様 50 年～100 年前の作品が多く含まれている。そして、次の図 1 は中国原作の日本語訳で使われた語彙と「出典」との対応を明らかにしたものである。

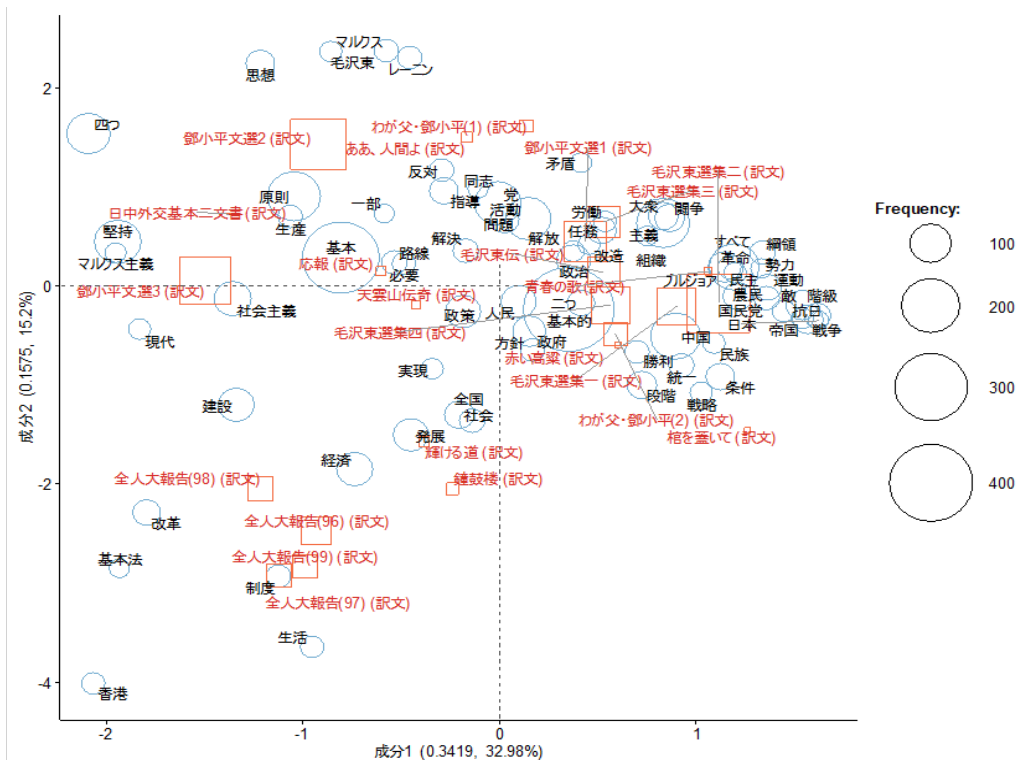


図 1 中国原作の日本語訳 語彙対応関係

上記図 1 を読み解くポイントは「原点からの距離が特徴の強さを意味する点にある。点線の交わっている場所が、横軸 0 で縦軸 0 という場所「原点」であり、原点に近い位置にある語はあまり特徴がない語になる。特徴がないというのは、外部変数の値に関係なしに、まんべんなく出現している語である(樋口他 2022:60)」とされる。また、丸いバブルの大きさは共起語の使われた回数を意味し、四角いバブルの大きさは出典の文量を意味する。

以上のことから『毛沢東選集』『毛沢東伝』『鄧小平文選』『わが父-鄧小平』『日中外交基本二文書』で用いられる特徴語として「大衆, 闘争, 綱領, 革命, 勢力, 民主, 敵, 国民党, 抗日, 帝国, 統一, マルクス, レーニン, 毛沢東, 四つ, マルクス主義, 社会主義, ブルジョア, 勝利, 制度, 戦争, 中国, 民族, 段階, 条件, 戦略, 思想, 堅持, 主義, 組織, すべて, 運動, 農民, 階級, 日本」などが確認でき、「改革, 香港, 生活, 基本法」は『全人大報告』で用いられている特徴語であることが明らかとなった。いずれにしても、中国語原文には全国人民代表大会関連が 4 作品, 鄧小平論集が 5 作品, 日中政治関連が 2 作品, 毛沢東論集が 5 作品となり、文体的偏りが眼立っている。

『中日対訳コーパス』は現在、新規の利用申し込みができない状況である。

### 2.1.2 情報通信研究機構『NICT 多言語対訳コーパス』

日中对訳コーパスには他に NICT のデータがあげられる。機械翻訳のシステム開発の一環として構築されたもので、京大コーパス(1995 年の毎日新聞の記事から抜粋された約 4 万文)を日本語原文とし、その中国語訳を作成して構成された対訳コーパスである(張他 2005; 510)。格納されているデータは全て日本語を中国語に訳した対訳文(日→中)となっている。NICT の研究成果として公開された日中对訳言語資源には「日英中基本文データ」と特許に関する専門用語のデータなどは確認できるが、「NICT 日中对訳コーパス」は公式サイトでは現在未公開となっている。

表 1. N I C T 日中对訳コーパスの詳細

	日本語	中国語
文	38, 383	
単語	947, 066	877, 859
語彙	36, 657	33, 425
一回出現の語彙	15, 036	13, 238
平均文長 (文字)	24. 7	22. 9

—張他(2008 ; 261)

### 2.1.3 JST・NICT 共同で構築された『アジア学術論文抜粋コーパス (ASPEC) 』

「アジア学術論文抜粋コーパス (ASPEC) (Asian Scientific Paper Excerpt Corpus)」に格納されているデータは全て日本語を中国語に訳した対訳文(日→中)である。並列文からなる日英論文抄録コーパス (ASPEC-JE) が 3MB, 並列文からなる日中論文抄録コーパス (ASPEC-JC) が 680KB 格納されている。ASPEC は、2006 年から 2010 年まで日本で行われた日中機械翻訳プロジェクトの成果の一つとして申し込みの申請を行うことで使用可能になる。配布用のデータには、train.txt : 672,315 対, dev.txt : 2,090 対, devtest.txt : 2,148 対, test.txt : 2,107 が含まれている。

2.1.4 GSK 通訳データベース『日中・日西サブコーパス (JNPC)』

「日中・日西サブコーパス (JNPC)」は特定非営利活動法人 言語資源協会(GSK)会員限定で公開しており、提供元は通訳コーパス作成共同研究者グループ (代表・立教大学 松下佳世) になっている。格納されたデータは、公益社団法人・日本記者クラブで行われた通訳付きの記者会見における、登壇者の原発話 (中国語・スペイン語・日本語) と通訳者の訳出を、映像、音声、文字情報を組み合わせた形でデータベース化したものである。原発話と訳出の書き起こしには、自動音声認識技術が用いられており、それぞれタイムスタンプが付与されているため、スプレッドシート等にエクスポートして分析することが可能である。

記者会見は中国語が平均約 1 時間半、スペイン語が約 1 時間 10 分、冒頭の発言ならびに質疑応答からなる。会見数は中国語 10 件 (同時通訳 8 件, 逐次通訳 2 件), スペイン語 11 件 (同時通訳 6 件, 逐次通訳 5 件) である。

2.2 自作対訳データ「日中对訳 EGA コーパス」の紹介

2.2.1 雑誌『Taiwan Panorama』

雑誌『Taiwan Panorama』は台湾で発行された日本語訳文付き (中→日) 総合誌で、中国語漢字表記は全て繁体字になっている。1976 年 1 月創刊時には『光華書報 SINORAMA』として 20 年以上発行されてきたが、2000 年 1 月号から日中对訳付初版が発売され、2006 年 2 月号からは名称を『Taiwan Panorama』に変更している。雑誌は現在 Web 版でも購読できるようになっている。

『Taiwan Panorama』の対訳文データの収集を開始した時期は 2009 年からであり、収集済みデータは 2008 年 6 月号から 2009 年 5 月号までの 12 ヶ月分に留まった状態である。データ量は、中国語原文総字数 205,251 字<sup>1</sup>、平均文長は 93.5 字; 日本語訳文総字数は 250,034 字、平均文長は 113.9 字となっており、トータル 2195 対の対訳文が含まれている。

	A	B	C	D	E
	年号	開始ページ	タイトル	中国語	日本語
2183	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	感。費心地為小吃「變身」, 例如, 將碗麵由脆的豬肉酥改成雞肉, 再改成扁薄短小的賣相, 美其名為「扁麵」; 芋頭、番薯原本黏性程度不同, 經巧手撮合為「芋薯甜糕」, 口感燒美獨夫人其實, 小吃入臺的風尚並非由阿扁起頭, 老牌台菜「欣葉」打從 32 年前台北市雙城街的 11 張桌子開始, 對台式菜席的想法就充滿價格的創見。	か出された。依頼されたビルドアップのシェフは、基準に則つて小皿料理を改良し、もち米料理の碗麵も脂っこい豚肉を鶏肉に変えて小ぶりに仕立てて、扁麵と名づけるなど工夫した。
2184	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	這份創見可追溯到 1970 年代, 當時, 百年前通化街邊官自賣出人的「江山樓」, 「蓬萊閣」早已消滅, 記憶中台式「酒家菜」(如風見大蝦、糖醋魚、肝散、四色火鍋等) 雖然在滬泉經北投找回生命, 又流入經鎮婚喪從小喜歡燒菜, 也會與人合夥開餐館的李秀英, 當年就矢志將「台菜」端上筵席檯面, 她延續了原在北投餐館做大菜的師傅陳清南與另一位資深主廚官官, 共同設計出一家筵席料理。	台湾の皿料理が宴席に並ぶようになったのは、陳前總統が最初ではない。32 年前に台北市双城街に 11 卓で開店した台湾料理の老舗欣葉は、台湾スタイルの宴席という新しいコンセプトの店である。
2186	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	可惜, 由於和當時主流的外省豪華菜不合, 欣葉度過了「一天只有 2 位客人」, 「含淚上菜」的草創期, 之後轉型確立以菜脯蛋、蛋黃肉、煎豬肝等為招牌的輕食小菜風格。邁進經濟起飛的 1980 年代, 欣葉的清粥小菜帶動飲食風潮, 平價的台菜業席也開出天地, 又陸續開創風格特殊的南餛飩、日式料理及咖啡屋, 事業不斷擴展。	小さい頃から料理好きだった李秀英は、台湾料理を宴会に志し、台湾料理のシェフと宴席料理を出すことにした。
2187	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	2006 年起, 欣葉把台菜餐飲帶入新加坡、北京與日本, 每道菜備從選料、刀工、火候、油溫控制到調味等環節, 都研究出現代化標準作業流程, 繼而李秀英的總經理李鴻鈞更帶領廚師出國談吃與觀摩, 研究精位於台北 101 大樓各種、充滿時尚感的 101 食藝軒, 那是規模驚人的最新考察成果; 道地的台式口味, 以西餐流程 (餐前酒、冷盤、前菜、主菜、甜點) 上菜, 佐以日式擺盤增添進餐情趣。	しかし、主流の中国風の好みに合わず、欣葉は一日のお客が二人という惨憺たる創業期を過ごさなければならなかった。その後は切干大根入り玉子焼きやレバー炒めなど、小皿料理を看板に路線変更した。経済発展が始まった 1980 年代、欣葉のお粥と小皿料理が人気を呼び、低価格の台湾料理の宴席もそれに連れて客がつくようになった。さらにはしゃぶしゃぶ、日本風料理店、カレー店
2188	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	以餐室中的冷盤「風華四喜拼」為例, 裡面就有海島魚子、鮮九孔、鮭魚卵與太子壽司捲, 裝盛在白色磁盤上顯得鮮艷欲滴, 精緻度媲美法國料理, 獨門的前豬肝, 食材的厚度、色澤、柔軟度已經過精挑細選, 作法有別於傳統先浸豬肝蒸製、切片, 改以大火快炒, 令醬油與糖在烹調過程中完全融化收乾, 嚼得到豬肝	2006 年、欣葉は台湾料理レストランをシンガポールと北京、日本に開店し、厳選した材料に標準化した調理プロセスを取り入れていった。家業を継いだ李鴻鈞社長は、シェフを連れて外国に視察に出かけ、より洗練されたレストランを目撃し、台北 101 の 66 階にあるフュッショナルなレストラン、101 食藝軒は、最新の視察の成果を発見している。台湾料理に食前酒から前菜、メインコース、デザートと続くフレンチのコースを取り入れ、日本風の盛り付けで興趣を添える。たとえば風華四喜と名づけられた前菜を見ると、カラスミにイクラと明太子寿司などが白い器に盛り付けられ、フレンチの皿の一角のように美しい。看板料理のレバー炒めでは、選り抜かれた食材に調理法は昔ながらの調味料に漬け込んで蒸す手順を強火で炒めるとに変え、醬油と砂糖をからめて水分を飛ばしている。レバーの口当りを残しながら、あっさりとして、赤ワインによく合う。
2190	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」		

図 2 「Taiwan Panorama」対訳文データ

2.2.2 雑誌『聞く中国語』

「聞く中国語」は日本で発行されている語学月刊誌として、ニュース、エッセイ、対談、人物紹介、日常会話集、歌詞、ドラマセリフなど幅広い題材が扱われている。使用されている中国語漢字表記は全て簡体字になっており、(対訳ではなく、日本語だけの記事もあるが) 中国語には基本的に日本語訳が付いている。各紙面には写真や絵などが満遍なく使われ、紙

<sup>1</sup> 半角、全角問わず 1 文字としてカウントしている。



面配置のバリエーションも豊富である。1冊あたり約130ページで構成されており、学習者のために漢字に発音記号のピンインを付している。データ収集を行う際には対訳文の所在を確認しながら手入力で作業を行っている。現時点の入力済みデータは2018年1月号から2023年7月号までの67冊分である。データ詳細は次の表2で示す。

表2. 雑誌「聞く中国語」入力済みデータ統計表

年	総字数	中国語字数	中国語平均文長	日本語字数	日本語平均文長
2018年	515604	212900	43.3	302704	61.6
2019年	419560	171449	41.0	248111	59.4
2020年	437739	177764	37.4	259975	54.7
2021年	395844	162202	35.3	233642	50.9
2022年	399846	162223	35.8	237623	52.5
2023年1月～7月	254923	105586	47.1	149337	66.6
統合	2423516	992124	40.0	1431392	57.6

### 2.2.3 中国日報社のニュースサイト『人民網』

『人民網』は、中国共産党中央委員会の機関紙『人民日報』で広く知られている中華人民共和国のメディア機関「人民日報社」が1997年1月1日に開設したニュースを主体とするネット情報交流プラットフォームである。『人民網』日本語版には中国語と日本語訳文付きニュースが掲載されている。

『人民網』の対訳文データは現在2014年7月から2023年8月現在まで98ヵ月分の収集が完了している（表3）。

表3. 『人民網』2014年7月～2023年8月現在 データ統計表

年度	総字数	対訳文	中国語字数	中国語平均文長	日本語字数	日本語平均文長
2014年(6ヵ月)	226,175	961	92,440	96.2	133,735	139.2
2015年	450,536	2,115	187,526	88.7	263,010	124.4
2016年	392,339	1,843	161,215	87.5	231,124	125.4
2017年	446,479	2,062	180,478	87.5	266,001	129.0
2018年	454,538	2,070	182,082	88.0	272,456	131.6
2019年	255,066	1,190	102,736	86.3	152,330	128.0
2020年	184,718	882	74,243	84.2	110,475	125.3
2021年	152,826	732	60,468	82.6	92,358	126.2
2022年	134,872	679	53,570	78.9	81,302	119.7
2023年(8ヵ月)	27,308	141	10,680	75.7	16,628	117.9
統合	2,724,857	12675	1,105,438	85.6	1,619,419	126.7

### 2.2.4 日中対訳 EGA コーパスの文体的特徴

ここでは「聞く中国語」及び「人民網」のデータを用い、語彙の文体的偏りについて確認を行う。

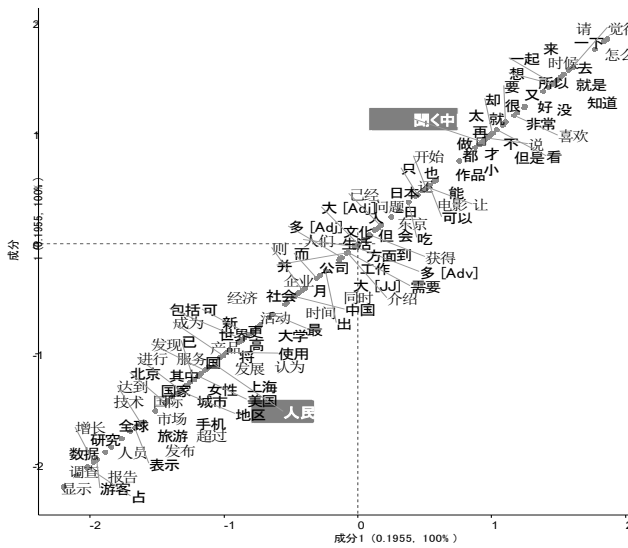


図 3. 言語資源の文体的偏りの対応分析

左の図 3 から読み取れるように、日中対訳 EGA コーパス（以降 EGA コーパス）は文体的偏りが見られず、自然言語を満遍なく収集できたデータサンプルであると考えられる。即ち、原点からの距離が特徴の強さを意味し、原点に近い位置にある語はあまり特徴がない語になる。特徴がないというのは、外部変数の値に関係なく、まんべんなく出現している語であることを現す。同様、原点から離れた語群は「聞く中国語」と「人民網」各々が持つ特徴語の集まりであることが読み取れる。

### 3. データの応用事例

本稿では主に KH Coder を用いた応用事例を紹介する。その理由の一つに KH Coder はもっとも有効なテキストマイニングソフトであると考えているからである。

KH Coder には三つの機能がある。「1 つ目は、テキストから自動的に語を取り出し、統計的な分析を行う機能。2 つ目は、分析者が注目したいコンセプトを取り出し、統計的な分析を行う機能。3 つ目は、語やコンセプトの統計分析をもとに、もとのテキストを検索・閲覧するための機能。(樋口他 2022;15)」である。KH Coder を用いた研究事例は 5,000 件を上回っている。今回、ソフトウェア KH Coder (無料版) を用いた応用事例では、雑誌「聞く中国語」2018 年 1 月～2023 年 7 月号 (約 242 万字) のデータを使用している。

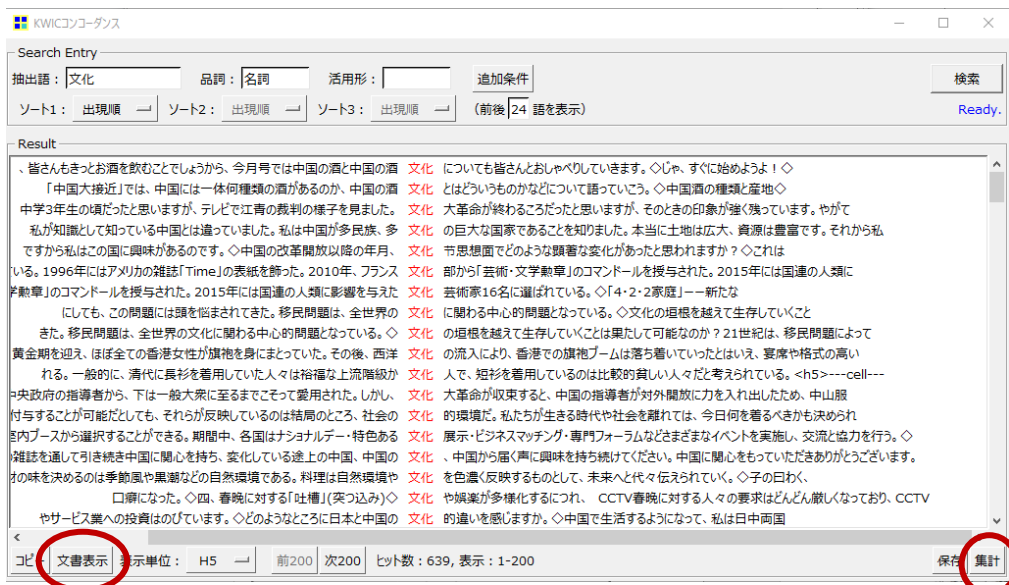


図 4. KH Coder KWIC コンコーダンスの一例

図4では抽出語を「文化」にした場合の例である。現在分析対象にしているのは日本語訳文になっているが、パネル左下の「文書表示」をクリックすることで中国語原文も同時に確認できる(図5)。ファイル内では「<<前>>」や「次>>」をクリックすることで、ヒットされた用例をスムーズに閲覧することが可能となる。更に図4右下の「集計」をクリックすると、図6のように抽出語「文化」を中心に、その前後の文脈から5グラムの範疇でもっとも共起しやすい語を確認することができる。

ここで注意すべき点は、中国語を日本語に訳した訳出文を構成する語彙は多かれ少なかれその起点言語の影響を受けていることである。

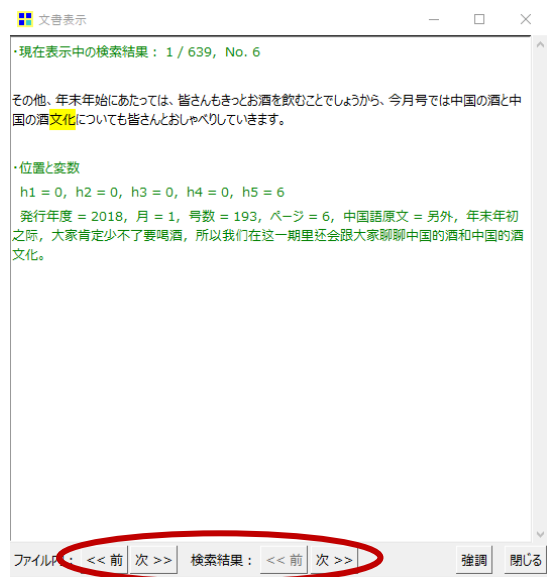


図5. 文書表示機能

図6で得られた結果はあくまでも中国語から(中→日)訳出された日本語として考え、自然言語として使用されている日本語とは区別すべきである。発表者はこの点を切口に、抽出

コロケーション統計

Node Word  
抽出語: 文化 品詞: 名詞 活用形: ヒット数: 639

Result

N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	Dice
1	日	地名	50	45	5	2	4	20	19	0	0	2	0	3	0	0.105
2	中	地名	49	44	5	2	2	3	19	18	0	0	2	0	3	0.104
3	交流	サ変名詞	42	3	39	0	2	0	1	0	27	7	1	1	3	0.101
4	違い	ナイ形容	40	1	39	0	0	0	1	0	0	26	5	7	1	0.098
5	伝統	名詞	38	36	2	2	3	0	1	30	0	1	0	1	0	0.091
6	理解	サ変名詞	38	1	37	0	1	0	0	0	1	18	6	9	3	0.088
7	日本	地名	107	95	12	5	20	17	18	35	0	5	1	2	4	0.075
8	中国	地名	122	108	14	6	13	21	19	49	0	6	3	1	4	0.067
9	遺産	名詞	21	1	20	0	1	0	0	0	17	0	2	1	0	0.062
10	芸術	名詞	25	12	13	1	1	2	4	4	8	2	0	1	2	0.059
11	両国	名詞	20	18	2	1	4	1	7	5	0	0	0	0	2	0.054
12	影響	サ変名詞	19	4	15	1	2	0	1	0	0	9	0	3	3	0.042
13	文化	名詞	26	13	13	4	7	2	0	0	0	0	2	7	4	0.041
14	歴史	名詞	18	12	6	0	2	1	6	3	0	5	0	1	0	0.041
15	漢字	名詞	13	10	3	0	1	0	0	9	0	1	0	2	0	0.034
16	無形	名詞	11	11	0	0	0	0	0	11	0	0	0	0	0	0.034
17	革命	名詞	11	0	11	0	0	0	0	0	0	11	0	0	0	0.032
18	伝える	動詞	12	1	11	0	0	1	0	0	0	8	0	2	1	0.032
19	習慣	名詞	11	3	8	1	0	0	2	0	1	4	2	0	1	0.031
20	国	名詞C	13	11	2	3	2	3	3	0	0	0	0	1	1	0.030
21	飲食	サ変名詞	9	7	2	0	0	0	0	7	0	1	1	0	0	0.026
22	最大	名詞	10	1	9	0	0	0	1	0	0	9	0	0	0	0.026
23	社会	名詞	11	5	6	1	0	1	3	0	0	5	0	0	1	0.025
24	興味	名詞	9	0	9	0	0	0	0	0	5	2	2	0	0	0.024
25	職場	名詞	8	8	0	0	2	0	1	5	0	0	0	0	0	0.024
26	深い	形容詞	10	3	7	0	0	1	2	0	0	2	3	1	1	0.023

コピー フィルタ設定 ソート: Dice 集計範囲: 左5 右5

語と共起しやすい語群の日中における違いを観察することで日中同形語「文化」の日中間のズレは何か、について考察を行っている。図6から読み取れるように、中国では「日文化、日〇文化、中文化、中〇文化、文化交流、文化～違い、伝統文化、文化～理解、日本文化、日本～文化、中国文化、中国～文化、文化遺産、芸術、両国文化、両国～文化、歴史文化、歴史～文化、漢字文化、無形文化、文化大革命」などがよく使われていることが明らかとなった。

図6. コロケーション統計の例

実際、「現代日本語書き言葉均衡コーパス(BCCWJ)」を確認したところ(図7)、「文化」は

24599例見られ、日本語高頻度コロケーションパターンとして「日本文化、日本～文化、文

化～文化，文化センター，芸術，社会，文化会館，文化活動，文化振興，伝統文化，伝統～文化，文化交流，歴史，生活，教育，地域，市民，国際，自然，文化研究，文化～研究，文化遺産，経済～文化，文化施設，文化事業，文化ホール，世界，食，スポーツ，時代，民族，文化協会，文化主義，中国，発展，文化産業，政治，都市，アメリカ，文化○革命，新しい，文化人類，影響」などがあげられる。

コロケーション統計

Node Word  
抽出語: 文化 品詞: 名詞 活用形: ヒット数: 24599

N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	Dice
1	日本	地名	1591	1333	258	149	144	278	427	335	2	54	55	83	64	0.116
2	文化	名詞	2274	1137	1137	374	472	263	26	2	2	26	263	472	374	0.092
3	センター	名詞	897	74	823	22	27	16	7	2	666	90	17	20	30	0.070
4	芸術	名詞	702	364	338	28	27	62	155	92	215	67	22	19	15	0.055
5	社会	名詞	708	458	250	45	101	158	134	20	29	87	34	51	49	0.054
6	会館	名詞	658	62	596	10	20	25	7	0	574	2	2	7	11	0.052
7	活動	サ変名詞	669	86	583	33	21	25	7	0	340	87	80	28	48	0.052
8	振興	サ変名詞	657	72	585	20	14	25	13	0	296	187	44	32	26	0.052
9	伝統	名詞	657	458	199	34	41	67	202	114	16	128	16	27	12	0.052
10	交流	サ変名詞	643	56	587	27	19	9	1	0	410	70	39	29	39	0.050
11	歴史	名詞	545	389	156	29	55	206	90	9	2	80	34	17	23	0.043
12	ない	否定助動詞	582	372	210	119	105	70	56	22	0	5	78	66	61	0.042
13	生活	サ変名詞	538	300	238	28	23	65	133	51	43	61	69	38	27	0.042
14	教育	サ変名詞	496	349	147	55	44	138	96	16	22	48	23	30	24	0.039
15	地域	名詞	486	366	120	85	68	72	104	37	3	29	14	37	37	0.038
16	市民	名詞	445	369	76	35	23	29	204	78	5	9	19	21	22	0.035
17	国際	名詞	364	224	140	25	20	44	83	52	32	36	29	22	21	0.029
18	自然	形容動詞	348	249	99	40	19	37	68	85	1	34	29	15	20	0.028
19	研究	サ変名詞	339	42	297	16	12	10	4	0	133	95	18	36	15	0.027
20	遺産	名詞	328	13	315	8	2	1	1	1	268	30	2	7	8	0.026
21	経済	名詞	335	244	91	38	47	103	51	5	10	38	14	21	8	0.026
22	施設	サ変名詞	333	60	273	18	15	19	8	0	180	28	20	22	23	0.026
23	事業	名詞	322	68	254	23	14	22	8	1	104	32	62	32	24	0.026
24	ホール	名詞	312	20	292	9	6	5	0	0	118	24	92	44	14	0.025
25	世界	名詞	320	215	105	37	30	47	65	36	2	39	22	16	26	0.025
26	食	名詞C	298	273	25	2	3	21	179	68	0	4	5	9	7	0.024
27	スポーツ	名詞	288	134	154	12	20	66	34	2	22	95	8	20	9	0.023
28	創造	サ変名詞	271	33	238	13	10	6	3	1	67	123	19	18	11	0.022
29	時代	名詞	263	153	110	39	41	44	28	1	2	25	27	32	24	0.021
30	民族	名詞	256	208	48	26	31	56	56	39	0	12	12	18	6	0.020
31	協会	名詞	235	24	211	6	7	11	0	0	178	28	0	4	1	0.019

コピー フィルタ設定 ソート: Dice 集計範囲: 左5 右5

文化遺産，経済～文化，文化施設，文化事業，文化ホール，世界，食，スポーツ，時代，民族，文化協会，文化主義，中国，発展，文化産業，政治，都市，アメリカ，文化○革命，新しい，文化人類，影響」などがあげられる。

以上のことから日中同形語「文化」は一見日中間で同形同義語として対応しているようで実際は異なった使われ方が多く存在していることが明らかである。また，日中同形語「文化」の共起ネットワークを用いた比較を行うことで，「文化」という語が含まれた文脈を構成する語群(内容語)が日中間でどのように重なり合うかの考察が可能である

(図8)．本稿では日中同形語「文化」を応用事例として取り上げることにも留めるが，日本語の「文化」は中国語と違って「会館，

センター，振興，保護，市民，活動，文化財」と結びつく傾向があることが読み取れる。

図7. 日本語「文化」の高頻度コロケーション (BCCWJ)

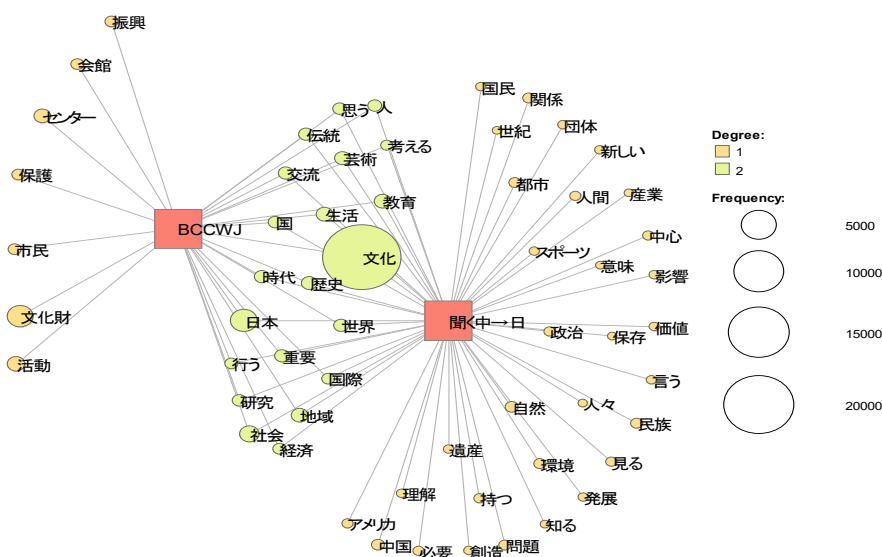


図 8. 高頻度コロケーションの比較 (上位 80)

#### 4. コーパスデータの構築及び配布に関わる著作権法

コーパスデータは個人使用の範囲では論文などで引用する際に出典を提示することで原則、著作者の承諾は不要となっているが、その成果物をオープンにしようとする様々な問題に直面する。特に用例提示が必要とされる言語研究では、著作物の第三者への全文共有は著作権問題がかかわってくる。

機械学習用に加工された学習用データセットに対して現著作者の権利を主張できるかに関しては、加工後の学習用データセットがどのような形になっているかにより結論が代わり得る。基本的に、自然言語処理分野では、データ自体は使い捨てとなるため、学習用データセットの公衆送信は著作権法 30 条の 4 が適用され、現在は原則著作者の承諾が不要となっており、機械学習等のためであれば公衆送信も認められている。ところが、翻案に関する平成 13 年の判例によれば学習用データセットが生データの本質的な特徴の同一性を維持し、生データの本質的な特徴を直接体感できるような場合には原著作者の権利を主張できるとされている。

##### 4.1 著作権法「第三十条四」とは

著作権法第 30 条 4 は次のように定義されている。

著作物は、次に掲げる場合その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合には、その必要と認められる限度において、いずれの方法によるかを問わず、利用することができる。ただし、当該著作物の種類及び用途並びに当該利用の態様に照らし著作権者の利益を不当に害することとなる場合は、この限りでない。

一 著作物の録音、録画その他の利用に係る技術の開発又は実用化のための試験の用に供する場合

二 情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、映像その他の要素に係る情報を抽出し、比較、分類その他の解析を行うことをいう。第四十七条の五第一項第二号において同じ。）

— 『著作権関係法令・条約集（令和元年版）』 pp25-26

##### 4.2 「著作権者の利益を不当に害する」とは

まず疑問になる点として「著作権者の利益を不当に害する」とはなにかである。要するに「著作権者の著作物の利用市場と衝突するか、あるいは将来における著作物の潜在的販路を阻害するかという観点から、最終的には司法の場で個別具体的に判断されることになる。」とされている（上野 2021）。即ち、データの扱いには「著作物の利用市場との衝突や潜在的販路を阻害する」要素を避けるための工夫が必要となるということである。

##### 4.3 「試験の用に供する」とは

次の疑問は「試験の用に供する」という表現であるが、平成 30 年改正前は、「電子計算機による情報解析……を行うことを目的とする場合には」と規定されていたため、情報解析を行う者が自ら著作物等の利用を行う場合が想定されていたが、同改正がこれを「情報解析……の用に供する場合」に変更したことによって、情報解析を行う者が自ら著作物等の利用を行う場合のみならず、情報解析を行う他人のために著作物等を複製したり、譲渡・公衆送信したりすることも権利制限の対象になり得るのである。例えば、情報解析を行う他人のために著作物等を収集して学習用データセットを作成することや、情報解析を行う複数事業者でこれを共有することも許容され得るのである（同上）。上記でいう「複数事業者」の範囲が明確でないこともあり、「著作物を享受する利用法を制限」にも関わることで、著作物の

共有、提供には処理内容(目的)の明確化と、契約に相当するなにかが必要となる。例えば、研究計画、情報処理の仕様書、処理の証跡のような研究上かならず必要なものがあげられる。これは万が一訴訟になった場合の証拠でもあり、安全管理上望ましいことだとされている。対訳文データの共有を著作権上問題がないものとするためには、共有先で研究計画や処理内容が準備されていることが望ましい。

#### 4.4 「情報解析」の範疇

「情報解析」の範疇に関して上野(2021)は次のように述べている。

起草者によれば、情報解析には①ウェブページや書籍等の中に含まれる特定の単語、文字列の用いられ方を分析し、多数のウェブページ、書籍等の中の異同の調査などの統計的な処理を行うウェブ情報解析や言語解析等②音声や映像、画像等に関し、それらを構成する音の波形、影像や文字列等が、どのような事物を意味するかについて、その波形の構成比、輝度・色彩、文字の構成比・出現頻度等を分析し、あらかじめ用意しておいた事物ごとの標準データパターン(特徴)のデータベースと照らし合わせて、その資料がどの事物の標準データパターンに近いのか判別(識別)を行う音声、影像、画像解析等がこれに当たるとされる[加戸 13]。また、文化庁の解説においては、「深層学習(ディープラーニング)の方法による人工知能の開発のための学習用データとして著作物をデータベースに記録するような場合も対象となるものと考えられる」と述べられている[文化庁 18]。したがって、例えば、画像認識や自動翻訳のAI開発のためにネット上の画像や文章を大量に収集することや、ディープフェイク技術開発のために特定の芸能人の音声や肖像写真を大量に収集することなども「情報解析」に当たると考えられよう。コンピュータを用いない情報解析も許容され得るのである。例えば、大量の新聞記事を情報解析するために人手でコピーをすることや、大量のテレビ番組を情報解析するために人手でハードディスクに録画する場合であっても、「情報解析」に当たり得るのである。

#### 4.5 日本初の事例

日本で著作権法第30条4を初めて適用し公開されたコーパスには「昭和・平成書き言葉コーパス(SHC)」があげられる。SHCは高度な検索システムを用いる点に関しては本稿がオープンにしようとするデータとは開示に用いるツールが大きく異なる。小木曾他(2023)は、高樹町法律事務所の小林利明弁護士にも助言を受け、著作物の利用行為が「軽微である」べき点を考慮し、SHC中納言では原文の表示される文脈長を、前後20語～30語範囲に絞ることにしている。

この「軽微である」とは何かについて発表者は、上野(2021)の著者に尋ねたことがある。その際いただいたコメントとして「『軽微』かどうかを問題にされているのは興味深い話であり、著作権法30条の4は、いわば入力段階はカバーしているが、他方で出力段階はカバーしていないとされ、その出力段階については、同法47条の5が必要になり、ここでは「軽微」であることが求められるところだということである(逆に言えば、入力段階など、30条の4だけでカバーできる範囲であれば軽微である必要はないことになる)。もっとも、この30条の4と47条の5の関係というのは非常に難問で、実は、知財学者も今さらながらこの問題に頭を悩ませているところである」と貴重な観点をご教示いただいた経緯がある(引用部分に関しては上野氏の許可を得たものである)。

本稿で紹介した自作日中対訳コーパスは全体の分量を「軽微な使用」に適した文長に調整し、更に段落の順番を全てシャッフルしたものをオープン用データとして用いる方法を視野に入れている。その引き換えに文脈情報が大きく乱れることになるが、一先ず日中対訳コーパスの試作としてEGA-Ver.1の配布を試みる。

## 5. 日中対訳コーパスの試作データ「EGA-Ver.1」の配布

EGA コーパスの試作データとして、2023年12月10日より EGA-Ver.1 の配布（無償）を開始する。データ使用には著作権法の制約があり、データの入手の際には「使用者申込」が必要となる。データの使用を希望される場合は次の「使用申込フォーム」のリンク（URL：<https://forms.gle/SFTi52Tow24shRqZA>）もしくは右側の QR コードより必要事項を記入し（誓約事項への同意を含め）送信することで申し込みは完了となる。お申込み内容に基づき当方から順次データを送付する。



同時に、2024年4月1日以降、日中対訳コーパスの全文を用いた共同研究を募集する。  
(<https://e-ga.jp/page0006.html>)

## 6. まとめと今後の課題

EGA コーパスは現在データの拡大を続けており、今後は中国語を日本語に訳した（中→日）データ及び、日本語を中国語に訳した（日→中）データの収集にも力を入れる予定である。収集済み日→中データには村上春樹の「1Q84」があり、入力済みデータ量は日本語原文約42万字、中国語訳文約28万字となる。

現段階では、データ全体を開示することは難しい状況であるが、近い将来大規模日中対訳コーパスの構築と公開の実現に向けて作業を続けていく所存である。

## 謝 辞

コーパスデータの収集及び活用手法にあたり、指導教官として終始多大なご指導を賜った、大阪大学人文学研究科基盤日本語学講座の教授石井正彦先生に深謝申し上げる。本稿は2023年3月17日に行われた自然言語処理学会ワークショップで多くの先生方からいただいた適切なお助言を賜った成果でもある。ここでコメントをいただいた先生方々に感謝の意を表す。最後に、幾度に渡り著作権法に関わる知見をご教示頂いた上野達弘先生にこれ以上なくお礼を申し上げます。

## 文 献

- 上野達弘(2021) 「情報解析と著作権——『機械学習パラダイス』としての日本」  
[https://www.jstage.jst.go.jp/article/jjsai/36/6/36\\_745/\\_pdf](https://www.jstage.jst.go.jp/article/jjsai/36/6/36_745/_pdf)
- 小木曾智信・近藤明日子・高橋雄太・間淵洋子（2023）『『昭和・平成書き言葉コーパス』の構築と公開』日本語学会 2023 年度春季大会ワークショップ予稿集 pp143-156、
- 共益社団法人著作権情報センター（2019）『著作権関係法令・条約集（令和元年版）』CRIC
- 張玉潔,馬青,内元清貴,井佐原（2005）「NICT 多言語コーパスにおける日中対訳データの構築」言語処理学会年次大会発表論文集 2005 年 3 月.pp510-513
- 張玉潔,王主龍,内元清貴,馬青,井佐原（2008）「日中対訳コーパスにおける単語・句の翻訳対応関係の付与」言語処理学会第 14 回 年次大会発表論文集 2008 年 3 月.pp261-264
- 樋口耕一・中村康則・周景龍（2022）『KH Coder OFFICIAL BOOK II 動かして学ぶ！はじめてのテキストマイニング —フリーソフトウェアを用いた自由記述の計量テキスト分析—』ナカニシヤ出版
- 樋口耕一著（2020）『社会調査のための計量テキスト分析【第2版】 内容分析の継承と発展を目指して』

- 文化庁著作権課 (2019) 「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方」(著作権法第30条の4, 第47条の4及び第47条の5関連) [https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30\\_hokaisei/pdf/r1406693\\_17.pdf](https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf)
- 北京日本学研究中心 (2013) 『日中対訳コーパス(CD-ROM)』
- 前川喜久雄 (2009) 「代表性を有する大規模日本語書き言葉コーパスの構築」『人工知能学会誌』2009年24巻5号 pp. 616-622, [https://doi.org/10.11517/jjsai.24.5\\_616](https://doi.org/10.11517/jjsai.24.5_616)
- 前川喜久雄 (2010) コーパス構築と著作権保護. 研究開発における情報利用と著作権[特集]. 人工知能学会誌 25巻5号 pp.628-632

#### 関連 URL

- コーパス検索アプリケーション『中納言』(2023年8月17日現在)  
<https://chunagon.ninjal.ac.jp/>
- 文化庁 Web サイト (2023年8月17日現在)  
[https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hosei/h20\\_05/shiryo1\\_2.html](https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hosei/h20_05/shiryo1_2.html)
- 情報処理学会 Web サイト(2023年3月現在) <https://www.ipsj.or.jp/faq/chosakuken-faq.html>
- Web サイト『人民網』(2023年8月17日現在)  
<http://j.people.com.cn/95961/index.html>



## 近代中国語関係書における主体移動表現の日中対照研究

金 敬玲 (國學院大學大学院)

### A Japanese-Chinese Contrastive Study of Fictive Motion Sentences in Modern Chinese Language Education-related Textbooks

JingLing Jin (Graduate School of Literature, Kokugakuin University)

#### 要旨

日清貿易研究所や東亜同文書院で教科書として用いられていたとされる明治 36 (1903) 年に出版された『華語跬歩』とその日本語訳である『華語跬歩総訳』(明治 37 年出版) を調査資料に「問答」という場面による会話文から抽出した中国語文 478 例、日本語文 466 例を分析対象に、同じ出来事における日中両言語の主体移動に関する表現や表現パターンを考察し、当時の主体移動表現の表現習慣として同じ事象を表現するに際し中国語の方が、移動動詞を選択する傾向が強いことと、会話文において経路関連要素である中国語の前置詞は日本語の後置詞ほど必要性が高くないことと、日本語における単独主動詞の使用は圧倒的であること及び、日中両言語ともに会話文において様態情報の必要性が非常に低いことが観察できた。

#### 1. はじめに

Talmy (2000) は、移動事象について日本語のような経路情報が主動詞で表される言語を動詞枠付け言語 (verb-framed language) と、英語のような様態が主動詞で表される言語をサテライト枠付け言語 (satellite-framed language) との 2 つの主要なタイプに分類している。Talmy (2000) はまた、中国語を英語と同様サテライト枠付け言語分類しているが、Slobin (2004) は、中国語は経路を表す動詞と様態を表す動詞が同等の地位を占める動詞連続構文を持つゆえ、等位枠付け言語 (equipollently-frame language) という 3 つ目の類型を提案している。

このように、類型論的には異なるタイプの言語ではあるが、同一の移動事象を表現するにあたり、「様態」「経路」「直示」の情報が競合することなく表現できる (小嶋 2019) という面では、日本語と中国語は同じタイプの言語と見做すことができる。

そこで、本稿では主語が自らの意志で移動する事象と無生物主語が主語となる移動事象を併せて主体移動事象とし、その事象の発生に伴う主体移動表現における日中両言語の対照研究を行う。

## 2. 先行研究

松本 (2016) は『日本語歴史コーパス』を用いて古代語の移動動詞を格助詞「より」と「を」とが「起点」と「経路」のどちらを表すかは動詞の意味的特徴によって決まるとし、移動動詞を4つのグループに分類している。

松本 (2017) は、日本語の移動表現をタイプごとに検討し、大規模コーパスに基づく数量的分析を報告している。その中で、主体移動に関しては「現代日本語書き言葉均衡コーパス・モニター公開版 (2009 年度)」で、生年が 1940 年代の著者のデータから選びだした 882 例を分析対象に、様態、経路、ダイクシスの表現頻度及び経路とダイクシスがどの位置で表現されることが多いかについて考察している。

LAMARRE (2017) は、中国語の経路を表す表現の全体像を紹介し、経路句・移動の参照点を表す場所名詞句が動詞の前と後という二つの位置に現れる要因について論じている。また、LAMARRE (2017) で用いたコーパスは5部の文学作品と1部のテレビドラマで、文学作品は老舎 (1899–1966) 著の《骆驼祥子》を除き、著者の生年は1930年代から1950年代であり、テレビドラマは2003年の作品である。

Shi&Wu (2014) は、中国語を古代中国語 (Old Chinese : 1 世紀以前)、中世中国語 (Middle Chinese : 2 世紀から 6 世紀まで)、前近代中国語 (Pre-Modern Chinese : 7 世紀から 19 世紀まで)、現代中国語 (Modern Chinese : 清末から現在まで) の4つの時代別のテキストから「物語性のあるもの、特に話し言葉のもの (we only selected narrative ones, particularly those of spoken mode)」を180例ずつ選びだし考察することを通じ、「中国語は動詞枠付け言語からサテライト枠付け言語へと典型的に移行している (Chinese has been undergoing a typological shift from a verb-framed language to a satellite-framed language)」とし、「現代中国語は、典型的な動詞枠付け言語とも、典型的なサテライト枠付け言語とも異なる多様なパターンを用いて運動事象を表現する (Modern Chinese adopts diverse patterns to encode motion events, which are different from both typical verb-framed languages and typical satellite-framed languages)」と述べている。Shi&Wu (2014) で用いた前近代中国語の資料は南宋から元末まで (1127–1368) のもので、現代中国語の資料については明示していないが、1968 年以降のものであると推測される。

## 3. 調査資料

本稿では、明治 36 (1903) 年に出版された『華語跬歩』と明治 37 (1904) 年に出版された『華語跬歩総訳』を調査資料として「問答」という場面の会話文を分析対象に主体移動を表現する文を抽出し、同じ場面における日中両言語の主体移動に関する表現や表現パターンを分析する。

御幡雅文（1859－1912）が著した『華語跬歩』<sup>1</sup>は上海にあった日清貿易研究所<sup>2</sup>や東亜同文書院<sup>3</sup>で用いられた中国語の教科書である（石田 2013）。『華語跬歩』は単語部分、短文（語句）部分、会話部分の三つの部分により構成され、本稿の対象となる部分は会話部分である。会話部分は日常会話の「家常問答」50章と立場や職業に応じた受け答えを教える「接見問答」30章からなっており、いずれも解説や訳文は付されていない。

伴直之助<sup>4</sup>（1862－1937）編『華語跬歩総訳』は『華語跬歩』の訳本で、自序に「本書の翻訳に際し（中略）京都清語講習所講師北京人任文毅君も、亦た少なからざる厚意を寄せられたり、茲に謝意を告白す」と記しており、翻訳する際、ネイティブスピーカーへの確認等があったと思われる。

『華語跬歩』から選び出した中国語文 478 例と『華語跬歩総訳』から選び出した日本語文 448 例を分析対象に近代の会話文における日中両言語の主体移動表現を考察する。

#### 4. 移動に関わる各表現について

日本語と中国語の自律移動を表す動詞として、「登る」－“登”、「入る」－“进”のような移動の経路情報を持つ「経路動詞」、「歩く」－“走”、「走る」－“跑”のような移動の様態情報を持つ「様態動詞」、「行く」－“去”、「来る」－“来”のような話者を基準点として移動を捉える「直示動詞」が存在していることは共通している。また、同一の移動事態を表現するにあたり、これらの動詞の表す情報が競合することなく表現できる同じタイプの言語である。

さらに、「部屋の中」の「中」、「城外」の“外”のような参照物との位置関係を表す位置表現が日中両言語に共に用いられている。

一方、日本語においては「から」「まで」「に」「へ」「を」「より」など経路後置詞を使用するのに対して、中国語においては介詞“从（から）”，“往（へ）”など前置詞で移動の起点や方向を表す。

#### 5. 『華語跬歩』における移動表現について

##### 5.1 移動動詞について

##### 5.1.1 “走”について

中国語の“走”は“走进教室”のように「歩く」の意でしばしば移動における様態動詞の代表例として挙げられる。“走”の「去る」「離れる」の意について論じる研究も見られており、“走”が様態動詞以外の働きもしているのは確かであるが、様態動詞としてではない動

<sup>1</sup> 石田（2013）によると、『華語跬歩』は三十年あまりの長きに渡って改訂されながら刊行されていたという。

<sup>2</sup> 1890年から1893年にかけて清国上海に存在した日本の教育機関。

<sup>3</sup> 東亜同文会が1901年上海に設立した学校で、日本の中国進出のための中堅幹部を養成する機能を担っていた。1921年専門学校、1939年大学となり、商務科（一時農工・政治科も設置）を置くが第2次大戦後廃止される。卒業生約5000名。

<sup>4</sup> 日本の実業家、政治家、衆議院議員（1期）。

詞の“走”に方向性が含まれているが否かについては議論されている。

丸尾(2005)は“走”について様態移動動詞“走<sub>1</sub>[歩く]”と方向移動動詞“走<sub>2</sub>[去る]”との二種類に分けている。それに対して、呉(2000)は“走”について移動様態動詞の他、「ある場所を離れるという移動の意味をもっているが方向性は示されていない」と述べている。

『華語跬歩』では“走”が25例みられ、下記例文のように様態を表す「歩く」の意で用いられるのが4例見られ、本稿では“走<sub>1</sub>”とする。

1) 要下来走走/降りて歩こーとすれば (接見問答第27章)

しかし、その他の3例の日本語文は移動の主体が腕時計で、無生物主語である腕時計の針の動き具合についての描写であり、人間の移動様態に注目している会話ではない。

また、以下の例文のように「この場からいなくなる」という「去る」「離れる」の意で用いられる例文が12例見られ、本稿では“走<sub>2</sub>”とする。

2) 早就该走咯/疾に往かねばならんのです (接見問答第29章)

これらの例文は丸尾(2005)で指摘している方向移動動詞に該当するものと思われる、例文中にはこの場を離れたあとの目的地が記されており、“走”の経路情報が得られ、「往く」「参る」との対応が見られる。

さらに、以下のような例文が見られる。

3) 走宅门子/屋敷廻り (接見問答第16章)

日本語訳文は「廻り」という表現を使用しているが、これは“走”の「訪問する<sup>5</sup>」を表す意で、目的地があることが示されており、本稿では経路動詞“走<sub>2</sub>”と見做す。

その他、上述の例文2)のような「経路」情報と例文1)のような「様態」情報とのいずれの情報も持たない“走”が8例見られる。

4) 共走了几天/皆なで幾日御掛りでした (接見問答第27章)

5) 不是应当走三天哪么/三日掛らねばならん所ですが (接見問答第27章)

これらの例文は移動に所要する時間に注目し、移動の「様態」情報と移動の「経路」情報を持たない中立的な移動を表しており、本稿では“走<sub>3</sub>”とする。

---

<sup>5</sup>『中日辞典第3版』(2016)は“走”を「歩く」「離れる」の外「訪問する、交際する」と解釈している。

本稿で見られる“走”を分類すると【表1】のようになる。

【表1】『華語跬歩』における“走”について

意味用法	様態動詞	経路動詞	中立移動動詞	計
用例数	4	13	8	25

### 5.1.2 “到”について

“到”について Shi&Wu (2014) は古代中国語 (Old Chinese) には見られず、中世中国語 (Middle Chinese) では「到着 (reach)」の意で経路動詞 (Path verbs) に分類されている。前近代中国語 (Pre-Modern Chinese) においても経路動詞として分類されているが「～に行く (go to)」の意で用いられ、現代中国語 (Modern Chinese) においては経路動詞として「到着 (arrive/reach)」と記述されている。

“到”は本来「到着する」「着く」を意味するが、参照物を示す名詞句を挟み直示動詞の“去/来”と組み合わせた“到…去”, “到…来”によって「～に行く」「～に来る」を意味することができる。これについて LAMARRE (2017) は「到着より前の移動部分がプロファイルされる結果、汎用的な移動動詞に転じたと考えられる」と述べている。

Shi&Wu (2014) は中世中国語においては“到”が移動動詞として単独で使用される例文のみ見られ、「到着」と記述していると思われるが、前近代中国語の移動表現のリストを見ると、“到…来”という直示動詞との組み合わせのほか、移動動詞としての単独の使用の“到”の例文が見られるにも関わらず、「～に行く (go to)」と記述していることは議論の余地があるようである。

『華語跬歩』における“到”の出現頻度は80で、以下の例文のように“到…去”, “到…来”で「～に行く」「～に来る」を表現する例文は37例見られる。

6) 没到别处去/外の處へは参りませんでした (家常問答第 20 章)

以下の例文のように移動動詞として単独で使用され「到着」を意味する“到”は32例見られる。

7) 他是昨天到的/あの方は昨日着きました (家常問答第 44 章)

その他、“到”が移動動詞として単独で使用されながら、「～に行く」を表す例文が以下のように見られる。

8) 我们先到了店里/私共最初宿屋に参りましたが (家常問答第 25 章)

9) 是到几个施主家取月例银子/はい、諸處の檀家へ月例の錢を受けに往くのです (接見問答第 21 章)

例文 8) 9) の“到”は移動を表す主動詞であり、過去の事象(例文 8)と未来の事象(例文 9)との異なる時制を表しているが、同じく方向指向の「～に行く」の意味を表している。例文 8) では“先(まず)”によって“店(店)”に行ったあとにも出来事があることを提示しており、例文 9) では“施主家(檀家)”に行ったあとの出来事である「月例の錢を受ける」を提示している。つまり、到達を前提に移動以外の出来事が後続する場合でも単独で使用される移動動詞の“到”は“去”と同義で「～に行く」を表現することができる。

丸尾(2005)は、“去图书馆借书(図書館に本を借りに行く)”と“到图书馆借书(図書館に行って本を借りる)”が形式的に同義になると指摘し、それは後続事項との関連において“去”は着点指向となるため、移動そのものではなく移動の結果に重点がある“到”と同義になると記述しているが、これは後続する出来事に重点を置いた解釈だと思われる。目的地に到着しないと後続する出来事が発生しないという目的地までの到着を重点に置く解釈となると、目的地に到着するための移動—つまり“到”の方向指向が際立つことになるという考えもあり得るだろう。

### 5.1.3 二音節動詞について

『華語跬歩』に見られる移動動詞は上述の“走”, “到”を含め単音節動詞が普通だが、「出発する」を意味する二音節移動動詞の“起身”が 8 例、“动身”が 4 例見られる。いずれも内部構造が「動詞+名詞」であるが、“起”は横になる状態あるいは座る状態から立ち上がる状態になる垂直方向の移動を表す動詞であるのに対して、“动”は「身を動かす」という中立移動動詞である。共通しているのは、前項の動詞のみの部分で移動動詞として判断することは難しく、“身”が加わることによって「出発する」という経路動詞を成す点である。また、動詞と名詞で組み合わせられるものであるため、間に“的”を挟む形式での使用もできる。

10) 他是多咱起的身/車夫は、いつ天津を登つたのら一 (家常問答第 35 章)

11) 他说他前几个早起动的身/一昨日の朝、登つたと申します (家常問答第 35 章)

## 5.2 動詞以外の経路表現について

### 5.2.1 方位詞について

『華語跬歩』に見られる方位詞を用いる例文 29 例の内訳を【表 2】に示す。

【表 2】『華語跬歩』における方位詞について

位置関係	「外」	「中」	「上」	「西」
方位詞	外	里	上	西
用例数	7	13	8	1

LAMARRE (2017) は、中国語の方位詞の統語的機能の重要性を示し、中国語の方位詞に

関して「地名など、それ自体で場所を含意する一部の名詞を除き、移動の参照点を表す名詞が前置詞と共起する場合、必ず方位詞が接続される」と述べ、“里”と“上”の使用頻度の高さを指摘している。【表2】に示すように、“里”と“上”の使用頻度は比較的高いと言えるものの、方位詞を用いた例文の割合は決して高いとは言えない。また、後節で詳しく述べるが、前置詞の使用頻度も高くなく、方位詞と共起する例文もわずかである。

以下の例文のように場所を表す名詞につき、かつ、前置詞と共起するものがある。

12) 从家里来/宅から参りました (接見問答第4章)

日本語訳文の「宅(の中)から」の「~の中」が省略できる点と異なり、中国語においては“里”は省略できない。これについてはLAMARRE(2017)でも指摘している。その他、場所を表す名詞がなく、方位詞だけで場所を表す例文が下記のように見られる。

13) 就往回里赶/直ぐに帰りましたが (家常問答第46章)

14) 他现在是在外去了/彼は只今外省へ出て居ります (家常問答第3章)

例文13)は“里”を除いた“就往回赶”でも成り立ち、むしろ“里”がない方が現代中国語として自然である。これは方位詞の“里”が「音声面でストレスを受けず接語的な側面が強い(LAMARRE2017)」ため、会話文では許容範囲内にあるとも考えられるが、その他の要因もあると思われる。その点については次節で言及する。

方位詞“外”を使用した例文は1例(“城外”)を除き、6例が例文14)のように名詞につかず、経路動詞と直示動詞の間に挟み、「外省へ行く」意を表している。同時期の中国語関係書<sup>6</sup>においては“出外去”を「旅行に出る」と翻訳しており、「遠出」を意味する固定用法ではないかなお考察の余地があると思われる。

“西”については以下の例文のように「太陽が西へ沈む」という特定用法がある。

15) 老爷儿都大平西了/御日さんが、も一入り掛つた (接見問答第15章)

## 5.2.2 前置詞について

『華語跬歩』に見られる前置詞には起点を示す“打”，“从”と方向を示す“往”が見られる。その中で、起点を示す“打”と“从”は場所を表す名詞につき、「~から」を表しているが、方向を示す“往”は場所名詞のほか、方位詞“里”，“外”との共起で二音節の経路情報を与えている。

<sup>6</sup> 『清語三十日間速成』(1904)と『支那語之勸』(1906)との2点を確認。

16) 您若是往外去，可以升什么呢/若し地方へ往かるとなれば、何役に御昇進なさることが出来ます乎ね (家常問答第9章)

前置詞“往”と方位詞の共起により二音節の経路成分を成すことについて、LAMARRE (2017) は北方語で多く使用され、着点指向の方向を表すとしている。そうすると、前節の例文 13) においても、“里”の音声面の特徴のほか、“往回”で方向を表現し、“里”で着点を表現したい話者の気持ちが含まれていると解釈することもできようが断定はしがたい。

中国語の前置詞は動詞に由来するとされている。前節 (5.1.2) で移動動詞“到”について述べたが、“到…去”構文で目的地に到着したあと後続する動作がない場合、“到”は前置詞として見なすことができる。つまり、例文 6) の“到”は方向を導く前置詞として機能しているとも解釈できる。『新著國語文法』<sup>7</sup> (1924) は“到”について「述語の前に付けることを要して、未だ到着せざる“到”は目的を表すもので“往”と通用せらる」と述べている。現代中国語においても“到”が前置詞なのか動詞なのかについて議論されている (劉 1998、LAMARRE2017)。本稿では、“到…去”構文の中で、以下の例文 17) のように“去”の前または後に動詞フレーズがあり、移動の目的や理由を述べている文における“到”を動詞とする。

17) 我现在到讲书堂送信去/手前は今まから、教会堂へ手紙を届けに参ります (家常問答第1章)

また、例文 6) のように移動に注目し、移動後の事項に注目しない例文における“到”を前置詞とする場合、“到”を 6) ’のように“往”に変えても同義になる。

6) 没到别处去/外の處へは参りませんでした (家常問答第20章)

6) ’没往别处去/外の處へは参りませんでした

以上のことを踏まえ、『華語跬步』に見られる“到”を分類すると【表3】のようになる。

【表3】『華語跬步』における“到”について

意味用法	前置詞	経路動詞		計
		到着	～に行く ～に来る	
用例数	9	32	39	80

<sup>7</sup> 黎錦熙『新著國語文法』(1924) は中国最初の本格的な現代中国語文法書であり、口語文法書である。



## 6. 『華語跬歩総訳』について

### 6.1 移動動詞について

#### 6.1.1 敬語表現の分類について

『華語跬歩総訳』における移動表現において、最も際立つ部分は敬語の使用である。「出る」「行く」「来る」の使用も見られるものの、「尊敬語」と「謙遜語」を含め敬語を使用した例文は204例(45.54%)見られる。『華語跬歩』の「接見問答」が立場や職業に応じた受け答えを教える教科書として使用されていたのは確かであるが、移動を表現するに際しては敬語の使用はほとんどない。

明治43年に出版された『華語跬歩総訳 上 増補』と『華語跬歩総訳 下 増補』は『華語跬歩』の著者である御幡雅文が翻訳者となっているが、同じく敬語が多用されており、『華語跬歩総訳』は原著者の意図に応じた翻訳だと推測される。

本稿は松本(1997)を参照に日本語の移動表現を分類するが、敬語をはじめ松本(1997)で言及されていない表現の分類を検討する。

まず、『華語跬歩総訳』で見られる主な敬語表現とそれに対応する表現及び本稿における分類を【表4】に示す。なお、「御帰りになる」のような「お～になる」が敬語標識となるものについては言及しない。

【表4】本稿における『華語跬歩総訳』の敬語表現の分類について

No.	敬語表現	対応する表現	用例数	動詞の種類
1	参る	行く・来る	77	直示
2	御出で類	出る(行く・来る)	55	直示
3	御越し類	越す(行く・来る)	5	直示
4	伺う	訪れる	7	経路
5	見える	来る	1	直示
6	いらっしゃる	行く・来る	1	直示
7	お伴致す	付き従って行く	2	付帯行為+直示
8	お暇致す	去る、いなくなる	2	経路

「参る」は「行く・来る」の謙讓語として、「いらっしゃる」は尊敬語として使用されているため直示動詞として分類し、また、「見える」が1例「来る」の尊敬語として用いられている例文が見られ直示動詞とする。

「御出で類」というのは「御出でになる」「御出で下さる」「御出でる」のようなもので「出る」ではなく「行く・来る」の尊敬語としてよく使用される敬語表現であることから、本稿では「行く・来る」の尊敬語として直示動詞と見做す。「御越し類」とは「越す」を基本形とした「御越しになる」「御越し下さる」のような表現で「行く・来る」の尊敬語として使用されているため直示動詞に分類する。

「伺う」については「訪れる」の謙讓語として経路動詞に分類し、「お伴致す」は「付き

従って行くこと」の意で付帯行為（付き従って）+直示（行く）と分類する。

「お暇致す」は「辞去の際の挨拶」の敬語表現を使用しており、「いなくなる」「去る」という意味として経路動詞に分類する。

### 6.1.2 漢語複合移動動詞について

移動に関する漢語複合動詞は16例見られ、松本（1997）で述べている「経路位置関係+基準物の包入」に相当するものが「入城する」「帰国する」「帰宅する」「上京する」「着京する」「出勤する」である。松本（1997）では「同道する」が見られていないが、「同行する」を様態+「行」と分類している。これに基づいて本稿では「同道する」を「様態+経路」に分類することにする。また、「外出する」を「方向性+経路」に分類する。

### 6.1.3 「這入る」と「出掛ける」について

「はいる」を「這入る」と表記した例が4例、「入る」と表記した例が2例見られる。

18) 私の這入つたのは哈達門です (家常問答第5章)

19) 晩には城に入ることが出来るわいと思つた (接見問答第27章)

例文18)の「哈達門」は当時の北京城に入る場合通過する門のことで、例文18)も例文19)も「城に入る」という出来事を述べており、同じ「はいる」として経路動詞に分類する。

「出掛ける」が17例見られ「出る」と同じく経路動詞に分類する。

## 6.2 後置詞

『華語跬歩総訳』は方位詞が見られない一方、後置詞を使用している例文は147例見られ、32.81%を占めている。

【表5】『華語跬歩総訳』における後置詞の現れ方

後置詞	動詞との共起の出現頻度			
	経路動詞	直示動詞	複雑述語	計
から	2	15	1	18
まで	0	1	0	1
から～まで	0	1	0	1
を	9	2	1	12
に	9	4	0	13
へ	29	66	6	101

「から」は経路動詞と直示動詞及び「帰って来る」と共起する例文が見られ「起点」を表

している。「から」を用いた日本語文と対応する中国語文を見ると、11例が起点を表す前置詞“打”，“从”と対応しており、4例がダイクシスによる場所提示例文と、3例が無生物主語の移動表現である。「から」の意味用法は様々であるが、主体移動の移動表現における「起点」を表す場合、中国語の前置詞“打”，“从”との対応関係にあると言えよう。

「まで」は一例直示動詞と共起し、「到達の限界点」を表しており、対応する中国語標識はなく、「から～まで」と対応しているのは方向を示す“往”である。

「を」と共起するのは経路動詞が多く、「城を出る」「天津を発つ」「紫竹林を立つ」のように「起点」を表すものがほとんどである。また、「通る」と共起し、「経路」を表すものが1例見られる。

その他、次のような例文が見られる。

- 21) 船で御越しでした乎、陸を御越しでした乎 (接見問答第 27 章)  
 22) 私は陸を参りました (接見問答第 27 章)

「越す」も「参る」も方向指向の移動動詞であるが、例文 21) 22) の「陸」は文脈上「起点」となっていない。そうすると、「を」の「起点」や「経路」とする解釈は上記例文には適応されない。田中 (1997) は、「Xを」の場合「動作が作用する対象としてのX」と解釈し、「を」が「経路」や「起点」を表すのはその働きの結果として「経路性」が浮かび上がったとしている。つまり、「城を出る」の「を」は「起点」としての「城」を表しているのではなく、「出る」という動作が作用する対象としての「城」を表していると認識すべきであると述べている。例文 21) と 22) の「陸」を「越す/参る」という動作が作用する対象としての「陸」とすれば「を」の解釈は成り立つと考えられる。

「に」は直示動詞と経路動詞（「入る」「戻る」「着く」「帰る」）と共起し、「着点」を表している。

「へ」は最も多く見られる後置詞で、「方向」を表す例文の他、下記 3 例のように「到着」を意味する経路動詞との共起も見られる。

- 23) 私が家へ着きましてから、間もなく、好くなりました (家常問答第 17 章)  
 24) 六時に渡し場へ着き (家常問答第 46 章)  
 25) あちらへ着して二箇月餘経つと (接見問答第 14 章)

その他、「到着」を表す動詞との共起でなくても、「香港へ参つたとき」「わたしの店へ電報が来ました」など、「帰着点」を表す「へ」が多数散見される。

『日本文法大辞典』(2001) は、「へ」の現代用法において「方向を表すのが本来の用法であるが、現代語では帰着点を表す「に」と区別して用いられることが少なく (中略) 動作・作用の帰着点を「へ」で表すことも多い」と述べている。また、山田 (1936) は江戸後期の

口語では「動作の帰着点を示す「に」が「へ」にその勢力を殆ど奪はれつくしてゐる」状態であると指摘しており、保科（1908）は「口語においてわ、その区別が殆ど消滅して、方向にも場所にも等しくエ（へ）を用いるのが普通の慣例になった」と述べている。『華語跬歩総訳』では「に」を使用した例文が13例で、「へ」を使用した例文が101例であり、同じく「へ」の支配的傾向が見られる。

## 7. 日中両言語の自律移動表現における移動動詞の使用頻度について

以上、『華語跬歩』における中国語の移動表現と『華語跬歩総訳』における日本語の移動表現をそれぞれ見てきた。方位詞に関して中国語では29例あるのに対して、日本語例文ではその使用が見られず、日本語の後置詞が147例で32.81%を占めているのに対して、中国語の前置詞は28例であり多く使用されているとは言い難い。また、主体移動の移動事象にのみ、起点を表す中国語の前置詞“打”，“从”と日本語の後置詞「から」との対応が見られる。

『華語跬歩総訳』が『華語跬歩』の翻訳版であるのにも関わらず、移動表現に関する例文の数が448例に対して478例で、日本語の方が30例少ない状況にある。このような状況に関して、まず無生物主語の例文を見ると、

26) 您的行李都来了/あなたの御荷物が、皆んな参りました (家常問答第8章)

のように、物全体の移動を日中両言語ともに主体移動表現で表しているものと、

27) 平常走的快慢怎么样/平常動き方の遅速は、どんなものです (家常問答第10章)

28) 走的也不快不慢/速くも無ければ、遅くも無く、 (家常問答第10章)

のように、物の一部の動きを中国語では移動表現で表しているのに対して、日本語では移動表現を用いない場合があることが観察できる。

次いで、中国語で「中立移動表現」に分類している“走”を使用した例文は、

29) 不是应当走三天哪么/三日掛らねばならん所ですが (接見問答第27章)

日本語の表現習慣上、移動表現とは関係がない場合が存在する。さらに、以下の例文のように、中国語の主体移動表現を日本語の客体移動表現で表すものがある。中国語文の“魚”が「海から離れる」動作の主体であるのに対して、日本語文は他動詞「出す」により動作主体が「魚」ではなくなり、中国語文の主体移動表現が日本語文の客体移動表現と対応する場が存在する。

30) 这个鱼是离海就死/此の魚は水から出すと間もなく死ぬので (接見問答第 23 章)

以上の三点の理由で中国語と日本語の移動表現に関する例文の用例数に差が生じたと言えるが、次は動詞の種類と頻度について見てみる。まず、単独で主動詞として用いられているものを【表 6】に示す。

【表 6】主体移動表現における単独用法の動詞使用頻度

動詞種類	『華語跬歩』(中国語)				『華語跬歩総訳』(日本語)		
	単純主動詞				単純主動詞		
	様態動詞	経路動詞	直示動詞	中立移動動詞	様態動詞	経路動詞	直示動詞
頻度	6 (1.26%)	108 (22.59%)	193 (40.38%)	10 (2.09%)	1 (0.22%)	161 (35.94%)	228 (50.89%)
	317(66.32%)				390 (87.05%)		

【表 6】に示すように、中国語において単純主動詞の用例数が 66.32%であるのに対して日本語は 87.05%であり非常に高い割合を占めている。これは、中国語の“回去”，“回来”のような「経路+直示」の複合的な表現を日本語では「帰る」「戻る」のような「経路動詞」という単独用法で表現できるためである。また、日本語を見ると、「直示動詞」の使用率が圧倒的に多い 50.89%に達している。これは、松本(2017)の調査結果である経路動詞(40.1%)の使用が圧倒的に多いことと異なる数値になっており、中国語文を翻訳したことに影響されているとも言えるが、本稿の分析対象が会話文であることにも影響されていると考えられる。

複合動詞及び複雑述語が使用された例文の内訳に関して、中国語文を【表 7】に、日本語文を【表 8】に示す。

【表 7】主体移動表現における複合動詞及び複雑述語の使用頻度 (中国語)

種類	複合動詞			複雑述語				
最終動詞	様態動詞	経路動詞		経路動詞	直示動詞			
組み合わせ	経路+様態	様態+経路	その他	様態・直示	様態+直示	様態+経路+直示	経路+直示	その他
頻度	1 (0.21%)	4 (8.37%)	1 (0.21%)	0	4 (8.37%)	2 (0.42%)	144 (30.12%)	5 (1.05%)
計	1(0.21%)	5(1.05%)		0	156(32.64%)			

【表 8】 主体移動表現における複合動詞及び複雑述語の使用頻度（日本語）

種類	複合動詞			複雑述語				
	最終動詞	様態動詞	経路動詞	その他	経路動詞	直示動詞		その他
組み合わせ	経路 + 様態	様態 + 経路	漢語複合動詞		付帯状況 + 経路	経路 + 直示	付帯状況 + 直示	その他
頻度	0	0	16 3.57%	3 0.67%	26 5.80%	10 2.23%		3 0.67%
計	0	0	16 3.57%	3 0.67%	36 8.04%			3 0.67%

【表 7】と【表 8】は、松本（2017）の「直示動詞は複合動詞には参加しない」と「複合動詞と複雑述語における経路表示のパターンをまとめると、直示動詞がある場合は直示動詞を、そうでなければ経路動詞を最後に置く傾向がある」という指摘に基づいて分類している。松本（2017）は日本語について述べているが、本稿では対照の便宜上、中国語文も同じ基準で表を作成している。

中国語文と日本語文ともに単独主動詞の使用が圧倒的に多いことが観察されるが、【表 7】に示すように、中国語文において「経路+直示」の使用が 144 例で 80.12%に達している。これは、“出去”，“回去”，“回来”が多用されるだと考えられる。LAMARRE（2017）ではこれらの表現を「二形態素型」とし、とくに会話において頻度が高いと指摘している。また、【表 6】～【表 8】を見ると、様態動詞の使用が非常に少ないということが観察できるが、これは調査対象が会話文であるためであろう。「様態」「経路」「直示」の情報が競合することなく表現できる言語ではあるものの、会話文では移動を報告する場面や移動を行う意思の表明など、移動様態の情報が必要と感じられる場面がほとんどない。LAMARRE（2017）においては、「テレビドラマ」と「小説」を調査し、テレビドラマで「様態」情報が含まれる移動表現は7%に過ぎないという結果となっており、本稿においても、「様態」「経路」「直示」の三つ情報が揃った例文は中国語文に 2 例のみ見られる。

- 32) 一去是坐车去的回头是骑驴回来的/往きには、車に乗って参りました、戻りには、驢馬に騎って帰りました (家常問答第 25 章)
- 33) 你滚出去罢/疾々と出て失せろ (家常問答第 40 章)

また、松本（2017）は「複雑述語ではすべてが直示動詞を最終動詞としており」と述べているが、例文 32) の日本語文のように、最終動詞が「帰る」という「経路動詞」の例文が見られるほか、「騎て出る」という最終動詞が「出る」という「経路動詞」の例文が 2 例見られ、「馬に騎る」という現代の場面では中々見られない交通手段に限定した表現と思われる。

その他、「通り越して行く」という「経路+経路+直示」の例文が 1 例見られるほか、「同道して参る」「往つて御出ででした (か)」が見られるが、これは敬語表現を使用したいという翻訳者の意図の現れではないかと推測される。

## 8. おわりに

本稿では明治期に出版された近代中国語関係書『華語跬歩』とその日本語訳である『華語跬歩総訳』の会話を対象に日中両言語の主体移動表現の対照研究を行い、当時の主体移動表現の表現習慣として同じ事象を表現するに際し中国語の方が、移動動詞を選択する傾向が強いことと、会話文において経路関連要素である中国語の前置詞は日本語の後置詞ほど必要性が高くないことと、日本語における単独主動詞の使用は圧倒的であること及び、日中両言語ともに会話文において様態情報の必要性が非常に低いことを指摘した。

### 調査資料

- 御幡雅文（1903）『華語跬歩』 文求堂  
伴直之助（1904）『華語跬歩総訳』 裕隣館

### 参考文献

- 日本語文献  
石田卓生（2013）「東亜同文書院使用以前の御幡雅文『華語跬歩』について」『同文書院記念報』第21巻 愛知大学東亜同文書院大学記念センター pp.121-132  
Christine LAMARRE（2017）「中国語の移動表現」松本曜編『移動表現の類型論』シリーズ言語対照第7巻 くろしお出版 pp.95-128  
小嶋美由紀（2019）「中国語主体移動表現の様相—ビデオクリップの口述データに基づいて—」森雄一・西村義樹・長谷川明香編『認知言語学を拓く』くろしお出版 pp.91-116  
呉念聖（2000）「中国語の移動表現」『法政大学教養部紀要．外国語学・外国文学編』111巻 pp.167-179  
商務印書館・小学館共同編集（2016）『中日辞典第3版』小学館  
田中茂範（1997）「空間表現の意味・機能」『空間と移動の表現』中右実編日英語比較選書⑥ 研究者 pp.2-119  
松本昂大（2016）「古代語の移動動詞と「起点」「経路」—今昔物語集の「より」「を」—」『日本語の研究』第12巻4号 日本語学会 pp.86-102  
松本曜（1997）「空間移動の言語表現とその拡張」『空間と移動の表現』中右実編日英語比較選書⑥ 研究者 pp.126-229  
松本曜（2017）「日本語における移動事象表現のタイプと経路の表現」松本曜編『移動表現の類型論』シリーズ言語対照第7巻 くろしお出版 pp.247-273  
保科孝一（1911）『日本口語法』 同文館  
丸尾誠（2005）『現代中国語の空間移動表現に関する研究』 白帝社  
山口明徳・秋本守英編（2001）『日本文法大辞典』 明治書院  
山田正紀（1936）『江戸言葉の研究：浮世風呂, 浮世床の語法』 普通教育研究会

英語文献

- Slobin, Dan Issac.2004. The Many Ways to Search for a Frog: Linguistic Typology and the Expression of Motion Events. In Sven Strömquist & Ludo Verhoeven (Eds.), *Relating events in narrative*, Vol. 2. *Typological and contextual perspectives* , pp. 219–257. Lawrence Erlbaum Associates Publishers.
- Talmy, Leonard.2000.A Typology of Event Integration.*Toward a Cognitive Semantics: Typology and Process in Concept Structuring*,pp.213-288. The MIT Press[高尾享幸訳(2000)「イベント統合の類型論 Leonard Talmy」坂原茂編『認知言語学の発展』ひつじ書房 pp.347-451]
- Wenlei Shi and Yicheng Wu(2014)Which way to move: The evolution of motion expressions in Chinese.*Linguistics* 52(5), pp.1237-1292

中国語文献

- 黎锦熙(黎錦熙)(1924)《新著国语文法》商务印书馆[大阪外国語学校大陸御研究所訳(1943)『黎氏支那語文法』大阪：甲文堂書店]
- 刘月华(劉月華)(1998)《趋向补语通释》北京：北京语言文化大学出版社



# 説明的文章の要点把握のための読解方法の有効性評価に用いる 正解文データセット (CAKeS) の作成

渡邊 幸佑 (兵庫教育大学) †

## Correct Answer Key Sentences Dataset for Evaluating Reading Methods for Extracting Key Sentences of Explanatory Texts

Kosuke Watanabe (Hyogo University of Teacher Education)

### 要旨

国語科の「読むこと」の指導において、説明的文章の要点を文単位で把握する読解方法(キーセンテンス把握方法)が教えられている。これまで様々な方法が提案されてきた。しかし、そのような方法によりどの程度正確に文章の要点を把握できるか定量評価されていない。他方、情報科学の自動要約の研究では、コンピュータによる要約を定量評価する方法がある。その一つに、要点として本来抽出すべき重要な文(正解文)と、コンピュータが抽出した文(抽出文)の一致度を測るものがある。この評価を実施するためには、人手で正解文を定めておく必要がある。しかし、国語教科書の説明的文章を対象に正解文を定めたデータセットはない。そこで、説明的文章を対象に研究力者3名と正解文を設定した。どの文を正解文とするかについて研究協力者間での判断は概ね一致したため、今回定めた正解文は要点として本来抽出すべき重要な文とみなしてよいと考えられる。

### 1. はじめに

『中学校学習指導要領(平成29年告示)』の国語科第一学年の「読むこと」に、説明的文章を用いた指導事項として、「文章の中心的な部分と付加的な部分、事実と意見との関係などについて叙述を基に捉え、要旨を把握すること。」(p.31)とある。この事項に関して、『【国語編】中学校学習指導要領(平成29年告示)解説』では、「キーワードやキーセンテンスなどに留意して情報を整理し、正確に要旨を捉えられるようにすることが重要である。」(p.68)とある。

これまで要点を把握するための読解方法が様々提案されてきた(大西 1981, 樺島 1983, 佐久間 1989, 岡本 1990, 石黒 2009, 川西 2014, 井上 2014, 白石 2014, 渡部 2015, 岡田 2016, 渡邊 2023a など)。しかし、従来提案されてきた読解方法は互いに異なっている。従来提案されてきた読解方法によって文章の要点をどの程度正確に把握できるか複数の文章サンプルを用いた定量評価はされていない(渡邊 2023b)。そのため、どの方法が優れており、どの方法を生徒に教えるべきか明らかでないのである。真に有効な読解方法を生徒に教えるためには、読解方法の定量的な評価が必要である。

他方、コンピュータを用いて要約を自動的に作成する自動要約の研究では、コンピュータの作成した要約を定量的に評価する手法が蓄積されてきた。たとえば、文章中から要点として重要な文を抽出させること(重要文抽出)の評価方法として、要点として本来抽出すべき重要な文(以下、正解文)と、コンピュータが抽出した文(以下、抽出文)の一致度を測るものがある(奥村・難波 2005)。正解文は人手で設定しなければならない。そこでテキスト自

---

† watten6210@gmail.com

動要約の評価型プロジェクト TSC(Text Summarization Challenge)が実施され、正解文の設定がされてきた (Fukushima and Okumura 2001, Okumura et. al. 2003, Hirao et. al. 2004)。

しかし、従来の正解文の設定は、新聞記事を対象としたものである。つまり、国語の教科書の説明的文章を対象としたものではない。国語教科書の説明的文章の要点を把握するための読解方法の有効性を評価するためには、国語教科書の説明的文章の正解文を定めたデータセットが必要である。

そこで、本稿筆者は、中学校国語科の教科書の説明的文章を対象に、研究協力者と正解文を定める作業をし、説明的文章の正解文データセットを作成した。

なお、このデータセットを「CAKeS (ケイクス)」と命名する。Correct Answer Key Sentences の下線部分をとりだしたものである。

## 2. 正解文データセット(CAKeS)の作成方法

### 2.1 対象文章

令和3年度版の中学校国語教科書における説明的文章 48 編とする。光村図書、東京書籍、三省堂、教育出版の中学校第 1 学年から第 3 学年の教科書に収められている説明的文章である。

### 2.2 研究協力者

研究協力者 3 名で作業した。国語科教育学の研究者 1 名、日本語学の研究者 2 名 (うち 1 名は小学校教員経験者) である。

### 2.3 正解文の数

1 文、並びに、文章全体の文の数の 10%、30%、及び 50%とする (以下、1 文抽出、抽出率 10%などと言及する。)

テキスト自動要約の評価型プロジェクト TSC では、抽出率 10%、30%及び 50%の文抽出を行っており、1 文抽出は行っていない。一方、国語教科書の説明的文章において重要な 1 文を抽出できることが経験的に認められるため、本稿では 1 文抽出も含め検討した。

なお、この正解文の数は目安であり、実際には研究協力者との合議により若干の増減が生じている。

設定する正解文の数に幅を持たせている。これにより、要約の制限字数の多寡に応じた最適な文抽出方法を検討できる。たとえば、最も重要な 1 文を抽出するときと、文章全体を 50%程度に縮小するときとは、異なる読み方を採用すべきと予想されるためである。

### 2.4 作業の流れ

以下のように、まず個人作業を行い、その後全体作業を行った。

- 1) 個人作業：個人で正解文を設定する。このとき他の者の作業内容は見ない。
- 2) 全体作業：その後全体で個人の結果を突合し合議により全体の正解文を定める。

### 2.5 作業の期間・場所

作業完了まで 2022 年度の 1 年間を要した。教科書会社 1 社の 1 学年分の個人作業 (文章 3~5 編) を 30 日程度の期間で行い、月末に合議を行うことを目安とした。

研究協力者は互いに遠隔地に居住していたため、合議はオンライン会議システムを利用し行った。研究協力者のうち 1 人を司会とし合議を進行した。合議の際は、個人作業の時点で 3 者とも判断が一致しているところは確認程度で済ませ、3 者間で判断が分かれた文について集中的に議論した。

### 3. 正解文データセット(CAKeS)の諸情報

以上の作業を通じ得られた正解文データセット(CAKeS)は次のものである。なお、どの文を正解文としたかについては将来的に公表したい。

#### 3.1 対象文章

対象文章のタイトル、教科書、発行者、文章全体の文の数、正解文の数は、表1のとおりである。

表1 対象文章

文章タイトル	教科書	発行者	文章全体の文の数	正解文の数(1文)	正解文の数(抽出率10%)	正解文の数(抽出率30%)	正解文の数(抽出率50%)
ダイコンは大きな根?	国語1	光村図書	34	1	3	11	17
ちょっと立ち止まって	国語1	光村図書	36	1	4	11	18
比喩で広がる言葉の世界	国語1	光村図書	29	1	4	9	14
「不便」の価値を見つめ直す	国語1	光村図書	79	1	8	24	40
「言葉」をもつ鳥、シジュウカラ	国語1	光村図書	63	1	6	19	32
モアイは語る	国語2	光村図書	79	1	8	24	40
クマゼミ増加の原因を探る	国語2	光村図書	91	1	9	27	47
君は「最後の晩餐」を知っているか	国語2	光村図書	99	1	10	31	52
人工知能との未来	国語3	光村図書	34	1	3	10	17
作られた『物語』を超えて	国語3	光村図書	71	1	7	21	36
人間と人工知能と創造性	国語3	光村図書	46	1	5	14	23
誰かの代わりに	国語3	光村図書	61	1	6	18	31
オオカミを見る目	新しい国語 1	東京書籍	55	1	6	17	28
ニュースの見方を考えよう	新しい国語 1	東京書籍	79	2	9	24	40
私のタンポポ研究	新しい国語 1	東京書籍	82	1	8	25	41
正しい言葉は信じられるか	新しい国語 2	東京書籍	72	1	7	22	36
サハラ砂漠の茶会	新しい国語 2	東京書籍	51	1	5	16	26
黄金の扇風機	新しい国語 2	東京書籍	50	1	5	15	25
ハトはなぜ首を振って歩くのか	新しい国語 2	東京書籍	99	1	10	30	50
絶滅の意味	新しい国語 3	東京書籍	85	1	9	26	43
幸福について	新しい国語 3	東京書籍	135	1	14	41	68
何のために「働く」のか	新しい国語 3	東京書籍	82	1	8	25	41
ペンギンの防寒着	現代の国語 1	三省堂	22	1	2	7	11
意味と意図	現代の国語 1	三省堂	44	1	4	13	22
クジラの飲み水	現代の国語 1	三省堂	45	1	5	14	23
玄関扉	現代の国語 1	三省堂	40	1	4	12	20
壁に残された伝言	現代の国語 2	三省堂	99	1	11	32	53
動物園でできること	現代の国語 2	三省堂	96	1	10	29	48
ジャンケン、なぜグー・チョキ・パーの三種類なのか	現代の国語 2	三省堂	35	1	4	11	18
一〇〇年後の水を守る	現代の国語 2	三省堂	86	1	9	26	43
人間は他の星に住むことができるのか	現代の国語 2	三省堂	56	1	6	17	26
俳句の世界	現代の国語 3	三省堂	46	1	5	14	23
フロン規制の物語	現代の国語 3	三省堂	88	1	9	26	44
間の文化	現代の国語 3	三省堂	42	1	4	13	21
「文殊の知恵」の時代	現代の国語 3	三省堂	41	1	4	12	21
言葉がたぐ世界遺産	中学国語 1	教育出版	65	1	7	20	33
子どもの権利	中学国語 1	教育出版	41	1	4	12	20
自分の脳を知っていますか	中学国語 1	教育出版	61	1	6	18	30
森には魔法使いがいる	中学国語 1	教育出版	89	1	9	27	45
水の山 富士山	中学国語 2	教育出版	53	1	5	16	27
紙の建築	中学国語 2	教育出版	65	1	7	20	33
日本の花火の楽しみ	中学国語 2	教育出版	49	1	5	15	25
学ぶ力	中学国語 2	教育出版	90	1	9	27	45
ガイアの知性	中学国語 2	教育出版	65	1	7	20	33
問いかける言葉	中学国語 3	教育出版	52	1	5	16	26
async——同期しないこと	中学国語 3	教育出版	62	1	6	19	31
AIは哲学ができるか	中学国語 3	教育出版	31	1	3	9	16
青春の歌——無名性の光	中学国語 3	教育出版	73	1	7	22	37

### 3.2 文の数え方

対象文章に含まれる文の数え方は次のとおりである。

#### 【原則】

句点、感嘆符、疑問符（以下、句点等）の数を数える。その数を文の数とみなす。ただし、タイトル中の句点等は数えない。

#### 【例外 1】

鍵括弧の後に「と」などの引用の格助詞がある場合、鍵括弧中の句点等はないものとみなす。

（例 1）1 文と数えるもの

「おお、賛成と反対の人が同じくらいいるんだなあ。」と思いませんか。

（池上彰「ニュースの見方を考えよう」）

（例 2）1 文と数えるもの

例えば、インタビューを受けたご本人から、「私の話したことのごとく一部しか伝えられなかった。これでは視聴者に誤解されかねない。」というような抗議が出ることも、ときにはあるのです。

（池上彰「ニュースの見方を考えよう」）

（例 3）2 文と数えるもの

「子どもは黙って大人の言うことを聞きなさい。」こんなことを言われて理不尽な思いをしたことはありませんか。

（大谷美紀子「子どもの権利」）

#### 【例外 2】

- 「比喻で広がる言葉の世界」の中の詩は文として数えない。
- 「青春の歌—無名性の光」の中の短歌とその詠み手はまとめて1文と数える。

### 3.3 一致率

複数人の読み手が1つのテキストに対して何を重要視するかについて必ずしも一致するとは限らない(浅原ら 2015)。研究協力者の個人作業の結果が互いにどの程度一致するかをまとめたものが表2である。

表2 一致率

	[3,0]	[2,1]	[1,2]	[0,3]	一致率
1文抽出	48	1	2	3008	0.941
抽出率10%	282	22	35	2720	0.832
抽出率30%	831	88	96	2044	0.819
抽出率50%	1375	156	178	1350	0.805

[3,0]：3名とも重要であると判断した文の頻度

[2,1]：2名が重要であると、1名が重要でないと判断した文の頻度

[1,2]：1名が重要であると、2名が重要でないと判断した文の頻度

[0,3]：3名とも重要でないと判断した文の頻度

一致率 =  $[3,0] \div ([3,0] + [2,1] + [1,2])$

なお、上記の式で定義する一致率の分母及び分子に[0,3]は含めていない。一致率の分母及び

分子に[0,3]を含めると正解文判断一致の実状を表せないからである。たとえば、100文からなる文章から1文を正解文として抽出する場合であって、3名とも異なる文を重要だと判断したとする。この場合、一致率の分母及び分子に[0,3]を含めて、一致率※ =  $([3,0]+[0,3]) \div ([3,0]+[2,1]+[1,2]+[0,3])$  と考えると、

$$\text{一致率※} = (0+97) \div (0+0+3+97) = 0.97$$

というように、3者の正解文判断が不一致にもかかわらず、一致率※は極めて高く、実状に合わないためである。

表2の一致率から、1文抽出の場合、研究協力者の判断は94%ほど一致している。抽出率10%の場合、研究協力者の判断はおおむね83%ほど一致している。抽出率30%の場合、研究協力者の判断は82%ほど一致している。抽出率50%の場合、研究協力者の判断は81%ほど一致している。このように、個人作業での研究協力者の判断は概ね一致している。さらにその後、合議により十分な議論を重ね、各人が納得した上で、全体での正解文を定めている。そのため、合議により定めた正解文は、要点として本来抽出すべき重要な文とみなしてよいと考えられる。

また、表2から、一致率は正解文の数が増加するにつれ減少する。ここから、制限字数の少ない要約であるほど、読み手間の判断が一致しやすいことが示唆される。

#### 4. 読解方法の有効性評価の方法

今回作成した正解文データセット (CAKeS) を用いて、説明的文章のキーセンテンスを抽出する読解方法を対象に、その有効性を評価できる。

評価尺度は、以下の再現率、精度、F値を用いる。F値が高いほど要点を捉える読解方法として有効性が高いと判断する。

$$\text{再現率} = a \div b$$

$$\text{精度} = a \div c$$

$$\text{F値} = 2 \times \text{再現率} \times \text{精度} \div (\text{再現率} + \text{精度})$$

a = ある読解方法で抽出した文に含まれる正解文の数

b = 正解文の数

c = ある読解方法で抽出した文の数

再現率は、正解文をどれだけ網羅的に拾っているか測る尺度である。精度は、正解文のみを拾っているか（不要な文を拾っていないか）測る尺度である。F値は、再現率と精度の調和平均であり、再現率と精度を総合評価するために用いる。F値が高いほど要点把握のために有効性が高い、と判断する。

#### 5. 今後の課題

今後の課題は、様々な読解方法の有効性評価、正解文データセット (CAKeS) のオープン化、正解文データセットの大規模化、正しい読みの存否の検討である。

これまで様々な読解方法が提案されてきた。「しかし」や「このように」などの特定の接

続表現を含む文を抽出する方法、内容のまとめの文や問題提起と答えを述べる文を抽出する方法、事実と意見に着目し意見を抽出する方法などである。こうした読解方法の有効性を評価し、真に有効な読解方法を見出せると期待される。なお、事実と意見に着目し意見を抽出する方法については既に有効性評価がされている(渡邊 2023c,2023d)。また、新しく提案した読解方法(渡邊 2023a)についても有効性評価を行いたい。

正解文データセット(CAKeS)を将来的にはオープン化し、あらゆる研究者の利用に供したい。その際、著作権、利用しやすさ等に留意する。なお、オープン化以前に正解文データセット(CAKeS)を利用したい方にとっては個別に相談されたい。

今回の正解文データセット(CAKeS)は、研究協力者3名で作業した。もし膨大な数の研究協力者とともに正解文データセットを作成できるならば、読みの実態がより鮮明に浮かび上がる。人々はどの文をキーセンテンスだと思うか、どのような要約を理想とするか。こうしたデータを全国・世界規模で集約することで、人間の文章理解の普遍性と独自性を明らかにしたい。

このような正解文データセットの大規模化が実現すれば、正しい読みの存否を検討できる。多くの人の間で一致する読み方を正解とみなすことができる。一方、多くの人の間で一致する読み方などないかもしれない。読みに正解がないとすれば、国語科教育はどうあるべきか。データ駆動型の国語科教育学は、従来の国語科教育の在り方を抜本的に変革する可能性を秘めている。

## 謝 辞

本稿は公益財団法人日本漢字能力検定協会、漢字・日本語教育研究助成「説明的文章の要点を把握するための読解方法の開発・評価」により実施した研究成果の一部である。

## 文 献

- 浅原正幸・杉真緒・柳野祥子.(2015)「BCCWJ-SUMM:『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス」『第7回コーパス日本語学ワークショップ予稿集』,pp.285-292.
- 石黒圭(2009).『よくわかる文章表現の技術Ⅱ 文章構成編 [新版]』明治書院.
- 井上善弘(2014).「接続語や文末からキーセンテンスを探すヒント指導」『教育科学国語教育』,No.778 pp.70-71.
- 大西忠治(1981).『説明的文章の読み方指導』明治書院.
- 岡田圭亮(2016).「中学校の説明的文章教材における要約指導の研究:初読後の文章把握を目的とした要約の方法」『上越教育大学国語研究』,30, pp42-55.
- 岡本明人(1990).「要約の力」がつく指導のヒント「要約」とは何か」『教育科学国語教育』, 32:5, pp.13-16.
- 奥村学・難波英嗣(2005).『テキスト自動要約』オーム社.
- 樺島忠夫(1983).「文章構造」『朝倉日本語新講座5 運用Ⅰ』朝倉書店,pp.118-57.
- 川西良治(2014).「段落ごとにキーセンテンスを探すヒント指導」『教育科学国語教育』, No.778, pp.68-69.
- 佐久間まゆみ(1989).「文章の統括と要約文の構造特性」『文章構造と要約文の諸相』, くらしお出版, pp.184-228.
- 白石範孝(2014).「要約指導の方法 白石流 「『問い』と『答え』と『要旨』の三点から

- の要約指導」『教育科学国語教育』,No.778, pp.24-27.
- Takahiro Fukusima and Manabu Okumura(2001) “Text Summarization Challenge: Text Summarization Evaluation at NTCIR Workshop2”, *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*,pp.45-50.
- Tsutomu Hirao, Manabu Okumura, Takahiro Fukusima and Hidetsugu Nanba(2004) “Text Summarization Challenge 3 - Text Summarization Evaluation at NTCIR Workshop 4”, *Proceedings of the Fourth NTCIR Workshop*,pp.407-411.
- Manabu Okumura, Takahiro Fukusima and Hidetsugu Nanba(2003) “Text summarization challenge 2: text summarization evaluation at NTCIR workshop 3”, *Proceedings of the HLT-NAACL 03 on Text summarization workshop* ,Vol.5, pp.49-56.
- 渡邊幸佑(2023a). 「説明的文章の要点把握のための読解方法—文脈効果に基づく文の重要度判定—」『読書科学』,64:2,pp.84-94.
- 渡邊幸佑(2023b). 「説明的文章の要点を把握するための読解方法についての批判的検討」『上越教育大学国語研究』,37,pp.27-41.
- 渡邊幸佑(2023c). 「「事実」と「意見」の再定義」『読書科学』,64:1,pp.1-14.
- 渡邊幸佑(2023d). 「事実と意見に着目した読み方により説明的文章の要点をどの程度正確に捉えられるか」『解釈』,69:5・6,pp.2-11.
- 渡部洋一郎(2015). 「授業の導入部における要約指導の効果」『教育科学国語教育』,No.792, pp. 9-13.

#### 関連 URL

NTCIR TSC <http://research.nii.ac.jp/ntcir/index-ja.html>

## 親疎関係で見る上昇下降調の使用率 — 『日本語日常会話コーパス』 を用いて —

李海琪 (浙江大学日本語科) †

### The Usage Rate of Rising-Falling Tones in Different Relationships Using Corpus of Everyday Japanese Conversation

Li Haiqi (Zhejiang University, Department of Japanese)

#### 要旨

親疎関係がイントネーションに影響することが知られているが、話者間の親疎関係と上昇下降調の使用率について定量的な分析が不十分である。本研究では『日本語日常会話コーパス』のコアデータから42会話(延べ話者数122名)を抽出し、句末音調のラベリングを話者ごとに統計し、親疎関係による上昇下降調の使用傾向を調査した。全体の傾向として、上昇下降調の使用率は、「家族親戚<友人知人<仕事関係者」であり、カイ二乗検定で有意差が見られた。親しさでは「家族親戚>友人知人>仕事関係者」だと考えられるため、上昇下降調との負の相関が示された。個人差、会話の形式と活動を考慮する場合でも、上昇下降調の使用率と話者間の親しさとの負の相関が観察されたケースが存在する。日本語母語話者の日常会話では、親しさが比較的に低く、丁寧さがより必要となる相手に対して、上昇下降調がより頻繁に使われる傾向があると言える。

#### 1. はじめに

親疎関係とイントネーションの関連性はこれまで研究されてきた。郡(2018: 111)は「話し手と聞き手の社会的・心理的關係」をイントネーションに影響する要因として挙げ、「話し手の状態や聞き手との関係がどのように表されるかは未解明の部分が多いが、高さの点では主に音域に関係するようである」と述べている。聞き手との関係と音域の関連性について、石本(2020: 374)は『日本語日常会話コーパス』の発話の平均F0を求め、「日常会話の声の高さは兄弟姉妹を除いた近親者に対しては低く、丁寧さが必要となる同僚・取引先・客相手には高くなり、友人に対してはさらに高くなる傾向がある」ことを明らかにした。

しかし、句末の声の高さの変化である句末音調が、聞き手との関係にどのような関連性があるかについて、量的に解明されていない部分がある。本研究は「ピッチが上昇したのち下降する」句末音調である上昇下降調(五十嵐ほか2006: 354)に焦点を当てる。

異なる親疎関係における上昇下降調の使用を明らかにすることは、日本語教育への貢献も考えられる。定延(2016: 217)は日本語教育で上昇下降調の指導を検討する時、「まず、その(実態と離れた)悪いイメージが日本語社会にあることを学習者に教えるべきである。さまざまな場面で、さまざまな相手に対して跳躍的上昇を発出するか否かは、学習者が判断すればいい」と述べている。しかし、どのような場面で、どのような相手に対して、上昇下降調を使えばいいかという疑問が残っている。この疑問を解答するために、まず日本語母語話者の使用傾向を明らかにする必要がある。日本語母語話者の使用傾向を目安として参考できれば、日本語学習者が判断しやすくなり、臨機応変でより適切に判断できると考えられ

---

† 22246028@zju.edu.cn



る。

親疎関係による上昇下降調の使用傾向について、量的に未解明な課題が残っている。そのため、本研究は『日本語日常会話コーパス』を利用し、実際の日常会話を調査することによって、上昇下降調の使用率と親疎関係の関連性を明らかにする。

## 2. 先行研究

上昇下降調に関する研究は主に機能と印象に集中している。上昇下降調は中止、言いよどみ、注意喚起、継続表示の手段の一つであり、説明する時やカジュアルな場面に使われることが明らかになっている（佐々木 2004、井上 2008、小磯 2014、前川 2014、郡 2020）。

上昇下降調と親疎関係の関連性を分析した論文として、郡（2016）、佐々木（2004）、金田（2007）が挙げられる。

郡（2016）は会話資料と聴取実験によって、間投助詞のイントネーション型の使い分けを分析した。親しい相手に対して、強調型上昇調に比べ上昇下降調のほうが多い傾向がある。また、疑問型上昇調の「ね」は女性が親しい相手に対して使う傾向があり、上昇下降調の「ね」より親しい相手に話しているように感じられることがわかった。しかし、間投助詞なしの文節末に現れた上昇下降調に関して、親疎関係の視点から分析しなかった。また、ほかのイントネーション型と比較したが、異なる親疎関係における上昇下降調の使用率の違いを分析しなかった。

佐々木（2004: 90）は「親密な雰囲気のかだけた場面」で上昇下降調が現れにくいとまとめたが、この結論を導き出すのに使った会話は5つしかなく、上昇下降調の使用率も計算していない。

金田（2007）は上昇下降調の時間当たり件数を使用率としたが、話速は人によって異なる。時間あたりに産出された文節が少ないと、上昇下降調が現れる機会も少なくなると考えられる。したがって、時間当たり件数を使用率として計算するのは適切であるとは言えないと思われる。

以上を踏まえ、本研究は日本語母語話者が親疎関係によって上昇下降調を使い分ける傾向を量的に明らかにすることを目的としている。

## 3. 研究方法

### 3.1 データ抽出

本研究は『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation, CEJC）（小磯ほか 2023）有償版を利用し、コアデータから42会話の延べ122名の話者のデータを抽出した。データ抽出の基準は：(1) 韻律ラベリングがある、(2) 同じ会話の話者の関係が同じ種類である。

まず、『日本語日常会話コーパス』は総計約20時間のコアデータがあり、人手で付与した韻律ラベリングがある（小磯ほか 2020）。そのため、コアデータに絞った。

次に、コアデータの52会話では、延べ話者数は169名であるが、そのうち17名は店員や方言で話した話者などであるため、韻律ラベリングの対象外である。そのため、17名を除き、残りの延べ話者数は152名である。

最後に、話者間の関係性に注目し、以下の3つの関係性に絞った。話者間の関係性は『日本語日常会話コーパス』のメタ情報に従う。

- (1) 家族親戚（「家族」と「家族・親戚」を含む）。20会話。

(2) 友人知人。15 会話。

(3) 仕事関係者（「同僚」と「仕事関係」を含む）。7 会話。

石本（2020: 373）では、発話の「向け先の種別の同定を簡便にするために、同種の関係となる参加者だけが存在する会話に限定して分析を行った」。石本（2020: 373）に従い、本研究では同じ会話に複数の話者間関係がある場合、その会話を対象外とした。「サービス場面関係」の 1 会話もあるが、備考の補足情報から、話者間は友人知人でもありと考えられるため、対象外とした。また、「先生生徒」関係は 2 会話しかないため、対象外とした。対象外としたのは総計 10 会話である。

以上の手順で、残り 42 会話（延べ話者数 122 名）を本研究の分析対象とした。会話の総計時間は 15.4 時間であり、会話形式は雑談（30 会話）、用談・相談（7 会話）、会議・会合（5 会話）である。延べ話者数 122 名（男性 51 名、女性 71 名）の中で、87 名は出身地も居住地も関東地方である。

### 3.2 ラベリングの統計

『日本語日常会話コーパス』有償版のコアデータには、簡易版 X-JToBI にもとづいた韻律ラベリングがある。簡易版 X-JToBI の Tone 層で用いる句末音調ラベルには、下降調 (L%)、上昇調 1 (H%)、上昇調 2 (LH%)、上昇下降調 (HL%)、上昇下降上昇調 (HLH%)、下降上昇下降調 (LHL%) がある（小磯ほか 2020）。本研究ではこの句末音調のラベリングを利用する。

韻律ラベリングは、Praat (Paul & David 2023) の Textgrid ファイルで提供している。Textgrid ファイルの命名は「会話 ID+話者ラベル」であるが、話者ラベルと話者 ID が異なる（例えば IC02 と C001\_001）。韻律ラベリングと話者の情報を対応するために、コーパスで提供した「話者・会話対応表」を参考し、コアデータに対して、「会話 ID・話者ラベル・話者 ID・話者情報」の対応表を作った。

42 会話に対して、上昇下降調 (HL%) の回数と非上昇下降調 (非 HL%) の回数を、話者ごとに統計した。具体的に、各 Textgrid ファイルに対して、Praat の Query-Query point tier-Count points where 機能を使った。「is equal to HL%」で算出した数を上昇下降調の回数とした。小磯ほか（2020: 36）によると、「アクセント句末最終モーラにおいて、音声は延長されかつ延長部分のピッチがほぼ一定値を保つ場合を対象に、句末境界音調の記号の後にエクステンダーの記号>を付与し、L%>や H%>のように表現する」。そのため、「ends with %>」と「ends with %」を合わせて句末音調の総数とした。非上昇下降調の回数＝句末音調の総数－上昇下降調の回数である。上昇下降調の使用率＝上昇下降調の回数/句末音調の総数である。

## 4. 結果と考察

### 4.1 全体の傾向

42 会話の話者間の関係性を説明変数として、句末音調（上昇下降調か非上昇下降調か）を目的変数として、SPSS 26 で分類木分析を行い、図 1 を得た。成長方法のアルゴリズムは CHAID を選択した。「Bonferroni 法を使用した有意確率の調整」を選択した。

上昇下降調の使用率は、家族親戚（6.4%）<友人知人（8.1%）<仕事関係者（10.2%）であり、カイ二乗検定で有意差が見られた ( $\chi^2=163.57$ ,  $df=2$ ,  $p<.001$ )。

話者間の親しさでは、一般的には「家族親戚>友人知人>仕事関係者」と考えられるのに対して、上昇下降調の使用率で並ぶと、図 2 の箱ひげ図のように「家族親戚<友人知人<仕事

関係者」という親しさと反対の傾向を示している。



図1 話者間関係別の上昇下降調（HL%）の使用率の分類木

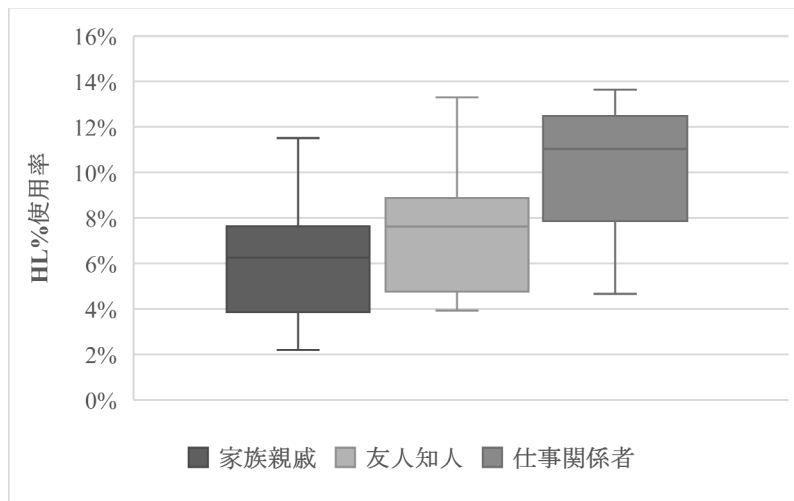


図2 話者間関係別の上昇下降調（HL%）使用率の箱ひげ図

本研究の調査結果は佐々木（2004）と金田（2007）の結論を支持したと言える。

佐々木（2004: 90）は「親密な雰囲気のできた場面」が「尻上がりイントネーション」（本研究で言う上昇下降調）が現れにくい場面の一つだとまとめ、「間投助詞が多用されることが多いからである」と述べている（佐々木 2004: 91）。

金田（2007）は大阪府・奈良県および兵庫県阪神エリア（南東部）出身の話者9名を対象に調査を行い、親しい相手は親しくない年下の人と比べた場合、句末昇降調（本研究で言う上昇下降調）の出現率が低いと結論づけた。理由について、上昇下降調は発話内容の吟味、適切な表現の考慮と関連しており、丁寧さと対人配慮の度合いが高い場合、発話内容の吟味がより必要であるため、上昇下降調を使いやすいと解釈している。

上昇下降調の「発話を考えるための時間稼ぐ」という役割は郡（2018）、小磯（2014）に言及された。本調査において丁寧さは「家族親戚<友人知人<仕事関係者」の順だと考えられ、上昇下降調の使用率の順と同じであるため、丁寧さで解釈できると思われる。

本調査の延べ話者数 122 名の中で、119 名は居住地が関東地方であり、87 名は出身地も居

住地も関東地方である。そのため、主に関東話者の使用傾向を反映したと考えられる。金田(2007)が関西話者のデータで得た結論を、本研究が主に関東話者のデータを用いて支持したと言える。

#### 4.2 同じ人の異なる関係性の会話

個人差を考慮し、同じ人が異なる関係性の人との会話において上昇下降調の使用率の傾向を考察するために、調査協力者8名に注目する。

42会話の調査協力者18名の中で、14名は家族親戚・友人知人・仕事関係者のうち二種類以上の関係性の会話がある。この14名(合計21会話)のうち、上昇下降調の使用率に関して、1名は家族親戚<友人知人<仕事関係者であり、1名は家族親戚<仕事関係者であり、6名は家族親戚<友人知人であり、合計8名は42会話全体の傾向に一致し、親しさと反対の傾向を示している。以下で上昇下降調の使用率を四捨五入で整数にし、具体的な状況を見る。

協力者T002(40代男性)は妻と雑談をした時に、上昇下降調の使用率は9%である。元同僚である友人と話す時、上昇下降調の使用率は14%である。他社の人との会話で上昇下降調の使用率は17%で最も高い。

協力者T004(60代女性)は自宅で夕食を取った時、夫と息子に対して上昇下降調の使用率は10%である。一方、地域懇談会の委員と打ち合わせした時に、仕事関係者である他の委員に対して上昇下降調の使用率は20%であり、その会話における上昇下降調の使用例を一つ挙げる。上昇下降調を使ったモーラの後に「 $\uparrow$ 」という記号をつける。Praatで表示するピッチ曲線から、句末に上昇したのち下降する山型が4箇所見られる。

会話ID:T004\_013(60代女性、仕事関係者との打ち合わせ)

だから $\uparrow$ 、ほら、わたしたちが $\uparrow$ 、地域の $\uparrow$ 、あの、未来予想図とかって思ってたって $\uparrow$ 、こう  
[permalink \(https://chunagon.ninjal.ac.jp/cejc/permalink?unit=long&position=T004\\_013,5010\)](https://chunagon.ninjal.ac.jp/cejc/permalink?unit=long&position=T004_013,5010)

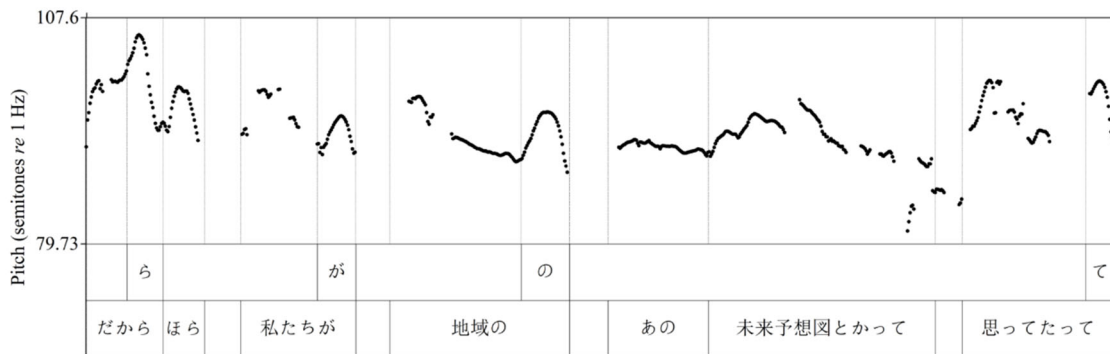


図3 Praatで見るT004の上昇下降調の使用例のピッチ曲線

協力者6名(20代女性2名、30代男性・30代女性・40代女性・50代女性各1名)の16会話において、協力者の上昇下降調の使用率は、それぞれ対家族親戚(平均4%)<対友人知人(平均8%)である。

#### 4.3 同じ人の同じ種類の関係性の会話

以上の協力者6名(上昇下降調の使用率が対家族親戚<対友人知人)から、2名の結果を取り上げる。

30代女性 T003 は家族親戚との会話が二つあり、場所と参加者が異なる。自宅で息子二人の勉強を見る時、上昇下降調の使用率は4%である。夫の実家で夫・義父母・義弟との雑談で、上昇下降調の使用率は5%である。関係がより親しいと思われる息子との会話で、上昇下降調の使用率が若干より低い。

石本 (2020) は声の高さに注目し、T003 の発話の F0 が「夫の家族が同席していても子どもや夫に対する場合とはっきりした違いが見られなかった」ことが、「T003 の夫の家族に対する心理的距離」から解釈できる可能性を指摘した。

上昇下降調の使用率に関しても、息子との会話と差があまりない理由は、T003 が夫の家族に対する心理的距離が比較的近く、夫の家族との関係が親しいことだと考えられる。

20代女性 T009 は友人知人との会話が二つあり、参加者と会話の形式が異なる。溪谷で恋人と散策する時上昇下降調の使用率は6%であり、大学祭実行委員会の幹部会議では9%である。丁寧度がより低いと思われる雑談と比べ、会議・会合のほうが上昇下降調の使用率が高く、親しさがより高いと思われる恋人と比べ、実行委員との会話のほうが上昇下降調の使用率が高いことが示された。

まとめると、同じく「家族親戚」か「友人知人」の枠にある親疎関係をより細かく見ても、より親しいと考えられる人に対し、上昇下降調の使用率がより低いケースが二つ見られた。

#### 4.4 同じ形式・活動の異なる関係性の会話

場面のフォーマル度などが上昇下降調の使用に影響する可能性が考えられるため、会話の形式と活動を統制し、親疎関係と上昇下降調の関連性を考察する。

まず、42 会話の形式では、会議・会合 (5 会話)、用談・相談 (7 会話)、雑談 (30 会話) がある。会話の形式と話者間の関係性を説明変数として、句末音調を目的変数として、分類木分析を行い、図 4 を得た。上昇下降調の使用率は、雑談 (7.4%) <用談・相談 (8.2%) <会議・会合 (9.5%) であり、場面のフォーマル度と同じ傾向を示している。

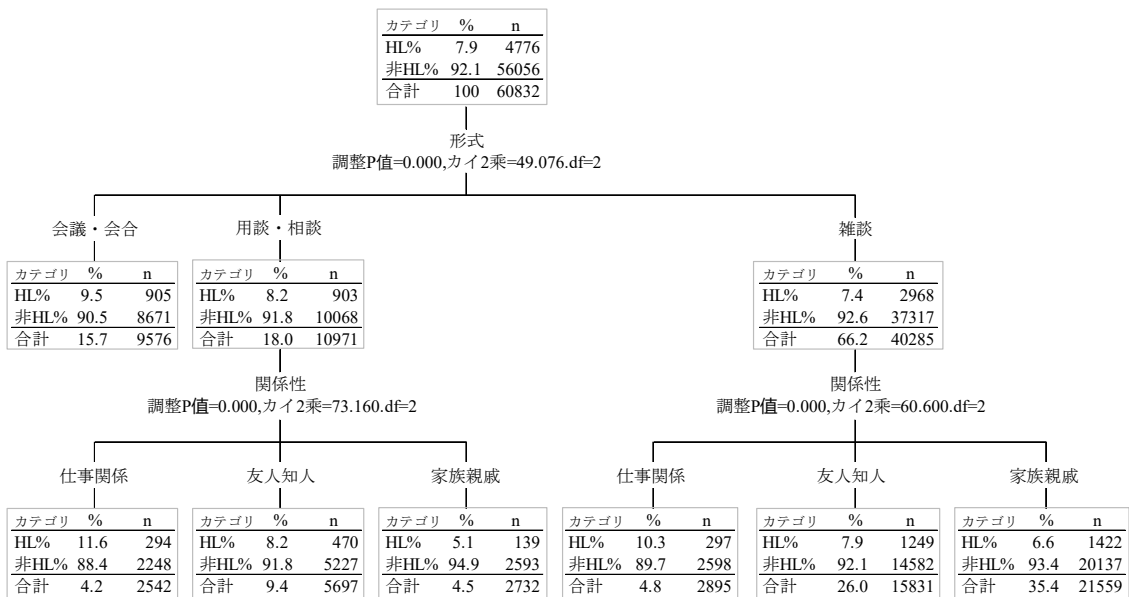


図 4 会話形式別の上昇下降調 (HL%) の使用率の分類木

用談・相談と雑談のいずれにおいても、上昇下降調の使用率は、家族親戚<友人知人<仕事関係者であり、カイ二乗検定で有意差が見られ、42 会話全体の傾向と一致する。会議・会合では、仕事関係者の 3 会話と友人知人の 2 会話がある。友人知人との 2 会話は地域活動の委員会と大学の実行委員会で行われるので、仕事関係の会話に近いと考えられる。そのため、仕事関係と友人知人の会話に有意差がないと考えられる。

次に、何をしながら会話をしていたかについて注目する。42 会話の活動で、「仕事」や「仕事・社会参加」など「仕事」を含む 8 会話を仕事類とし、「食事」や「食事・付き合い」など「食事」を含む 20 会話を食事類とする。仕事類と食事類以外の 14 会話を除外し、残り 28 会話の活動と話者間の関係性を説明変数として、句末音調を目的変数として、分類木分析を行い、図 5 を得た。

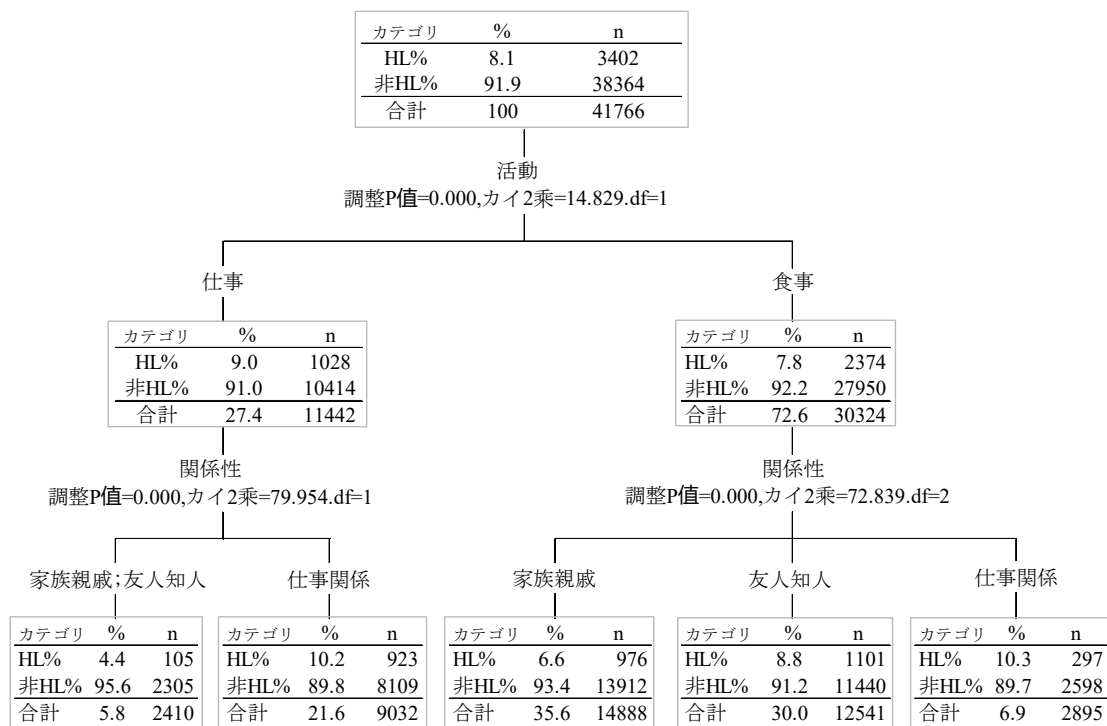


図 5 活動別の上昇下降調 (HL%) の使用率の分類木

上昇下降調の使用率は、食事類<仕事類であり、カイ二乗検定で有意差が見られた。上昇下降調の使用率は、仕事類の会話では家族親戚・友人知人<仕事関係者であり、食事類の会話では家族親戚<友人知人<仕事関係者であり、親しさとの負の相関が示された。仕事類の会話は家族との 1 会話、友人知人との 2 会話と仕事関係者との 5 会話があり、家族親戚・友人知人との 3 会話はすべて雑談である。

## 5. おわりに

本研究は『日本語日常会話コーパス』のコアデータから 42 会話（延べ話者数 122 名）を対象に、上昇下降調の使用について調査を行った。全体の傾向として、上昇下降調の使用率は、「家族親戚<友人知人<仕事関係者」のように、話者間の親しさと負の相関が見られた。同じ話者、同じ形式、同じ活動の会話においても、親しさと上昇下降調の使用率の負の相関

が示されたケースが存在する。日本語母語話者の日常会話では、比較的親しくなく、より丁寧に接する必要がある人に対して、上昇下降調がより頻繁に使われることがわかった。

## 文 献

- 五十嵐陽介・菊池英明・前川喜久雄 (2006). 「韻律情報」『日本語話し言葉コーパスの構築法』 124, pp.347-453.
- 井上史雄 (2008). 『社会方言学論考：新方言の基盤』 明治書院, pp.343-346.
- 石本祐一 (2020). 「日本語日常会話コーパスから見える会話場面と声の高さの関係性」『言語資源活用ワークショップ発表論文集』 5, pp.371-378.
- 金田純平 (2007). 「句末昇降調について一現れ方と成り立ち一」 定延利之・中川正之(編)『音声文法の対照』 くろしお出版, pp.103-128
- 小磯花絵 (2014). 「日本語自発音声における複合境界音調と統語構造との関係」『音声研究』 18:1, pp.57-69.
- 小磯花絵・菊池英明・山田高明 (2020). 「『日本語日常会話コーパス』への韻律ラベリングーラベリングの設計と日常会話の韻律の特徴一」『人工知能学会研究会資料』SIG-SLUD-B90, pp. 34-39.
- 小磯花絵・天谷晴香・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・渡邊友香 (2023). 「『日本語日常会話コーパス』設計と特徴」『国立国語研究所論集』 24, pp.153-168.
- 郡史郎 (2016). 「間投助詞のイントネーションと間投助詞的イントネーション：型の使い分けについて」『言語文化研究』 42, pp.61-84.
- 郡史郎 (2018). 「イントネーション」北原保雄(監修)上野善道(編)『朝倉日本語講座 3 音声音韻 (新装版)』 朝倉書店, pp.109-131.
- 郡史郎 (2020). 『日本語のイントネーションーしくみと音読・朗読への応用』 大修館書店.
- 前川喜久雄 (2014). 「『日本語話し言葉コーパス』の X-JToBI アノテーションから抽出される韻律上の発話スタイル」『音声研究』 18:1, pp.70-82.
- 定延利之 (2016). 「4つの発話モード」庵功雄・佐藤啄三・中俣尚己(編)『日本語文法研究のフロンティア』 くろしお出版, pp.205-223.
- 佐々木香織 (2004). 『日本語音声談話の韻律構造』 (東京外国語大学博士論文)
- Boersma, Paul and Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.10, retrieved 3 May 2023 from <http://www.praat.org/>

## 関連 URL

- |               |   |
|---------------|---|
| 『日本語日常会話コーパス』 | <a href="https://www2.ninjal.ac.jp/conversation/cejc.html">https://www2.ninjal.ac.jp/conversation/cejc.html</a> |
| Praat         | <a href="https://www.fon.hum.uva.nl/praat/">https://www.fon.hum.uva.nl/praat/</a>                               |

## 論文指導の場面における終助詞「かな」の配慮機能 — 『BTSJ 日本語自然会話コーパス』のデータから—

劉 悦 (筑波大学大学院人文社会科学研究群) †

### The Function of "Kana" in Thesis Instruction Situations: Survey Based on BTSJ Corpus

Liu Yue (Degree Programs in Humanities and Social Sciences, University of Tsukuba)

#### 要旨

本研究は、『BTSJ 日本語自然会話コーパス』から論文指導の場面における終助詞「かな」の配慮機能を明らかにすることを目的としている。分析は論文指導の会話データから終助詞「かな」を抽出し、個々の文脈情報を踏まえて「かな」の機能を確認し、機能カテゴリーごとにまとめるという手順で行った。その結果、学生の発話では39件の使用例が確認されたが、そのうち、31件が「意見・考えを述べる」という機能カテゴリーに分類された。また、教師の発話では75件の使用例が確認され、「意見・考えを述べる」「回答を強制しない質問をする」「助言を行う」「相手の意見を完全に受容することができない態度を間接的に示す」という機能カテゴリーに分類される「かな」の使用が確認された。本研究より、論文指導の場面において、立場の上下に関わらず教師も学生もさまざまな機能の「かな」を活用し、対人配慮を示していることが明らかになった。

#### 1. 研究背景

本研究は、論文指導の場面における終助詞「かな」の配慮機能の解明を目的とする。日本語学習者にとって日本語の終助詞は使い分けが難しく、習得が困難な学習項目とされている(白岩 2011)。しかし、これまでの終助詞の文法記述や習得に関する研究は「ね」「よ」に集中しているため、「ね」「よ」以外の終助詞についても考察を深め、日本語学習者への還元を目指した研究が必要である。

終助詞「かな」は、先行研究において語用論的機能が議論されている(鈴木 2015、中西 2015 など)ものの、自然会話データに基づく考察は限られている。そこで、本研究は配慮表現が表れやすいと想定される論文指導の場面に着目し、そこで見られる「かな」の配慮機能を明らかにする。

#### 2. 先行研究

##### 2.1 配慮表現に関する先行研究

野田ほか(2014:3)は、「聞き手や読み手に悪い感情を持たれないようにするために使う表現である」と配慮表現を定義している。一方、山岡(2015:318)は、配慮表現を「対人的コミュニケーションにおいて、相手との対人関係をなるべく良好に保つことに配慮して用いられることが、一定程度以上に慣習化された言語表現」と定義している。本研究は、配慮表現としての「かな」の機能をなるべく広範に考察するため、山岡(2015)の定義を用い、良好な対人関係の維持に使用されている言語表現を広く「配慮表現」として取り扱う。

---

† liu.yue.2022@japan.tsukuba.ac.jp



また、配慮表現は、Leech のポライトネスの原理や Brown & Levinson のポライトネス理論と深く関わっている。山岡ほか(2019)は、Leech(1983)のポライトネスの原理と配慮表現との対応関係を表 1 のように示した。表 1 では、「負担と利益にかかわる気配りの原則と寛大性の原則には、反比例的な原理である『配慮表現の原理』が存在する」(山岡ほか 2019:57)ということが強調されている。例えば、ポライトネスの原理「他者の負担を最小限にせよ」という依頼行為において、「ご面倒で恐縮ですが」など「他者の負担が大きいと述べ」ることで聞き手に配慮するという原理が存在している。

表 1 ポライトネスの原理と配慮表現の原理・機能分類との対応関係  
(山岡ほか 2019:57-65 に基づいて筆者が作成)

ポライトネスの原理	配慮表現の原理	機能分類
他者の負担を最小限にせよ	他者の負担が大きいと述べよ	負担表現／他負大
他者の利益を最大限にせよ	他者の利益が小さいと述べよ	利益表現／他利小
自己の利益を最小限にせよ	自己の利益が大きいと述べよ	利益表現／自利大
自己の負担を最大限にせよ	自己の負担が小さいと述べよ	負担表現／自負小
他者への非難を最小限にせよ		緩和表現／侵害抑制
他者への賞賛を最大限にせよ		賞賛表現
自己への賞賛を最小限にせよ		謙遜表現／自賛抑制
自己への非難を最大限にせよ		謙遜表現／自己非難
自己と他者との意見相違を最小限にせよ		緩和表現／不一致回避
自己と他者との意見一致を最大限にせよ		賛同表現
自己と他者との反感を最小限にせよ		緩和表現／不一致回避
自己と他者との共感を最大限にせよ		共感表現

Brown & Levinson(1987)は人間は誰でもフェイス(face)を持っていると想定している。そして、「フェイスを脅かす行為」、すなわち FTA(face-threatening acts)には「ポジティブ・ポライトネス」(positive politeness)と「ネガティブ・ポライトネス」(negative politeness)という二つの形の補償行為が伴う場合がある(Brown & Levinson 1987)。前者は聞き手のポジティブ・フェイス(positive face)に向けられるものであり、後者は主として聞き手のネガティブ・フェイス(negative face)を守ろうという欲求に向けられるものである。

以上の理論は、配慮表現が使用されている原因を端的にまとめている。本研究は、以上の理論の枠組みを援用し、「かな」が配慮表現として働くメカニズムを考察する。

## 2.2 「かな」に関する先行研究

終助詞に関する文法的理解においては、「かな」を終助詞「か」と「な」の意味・機能の組み合わせとして捉えるのではなく、ひとまとまりとして捉える考え方が主流であり(山下 2012、中西 2015、鈴木 2015 など)、各辞書も「かな」を一つの項目として扱っている。ここでは、各辞書における「かな」の記述を確認しておく。

『大辞林 (第四版) 』(三省堂、2019年)

(1)軽い詠嘆の気持ちを込めた疑問の意を表す。

(2)自分自身に問いかける気持ちを表す。

(3) (「ないかな」の形で) 願望の意を表す。

『広辞苑 (第七版)』(岩波書店、2018年)

(1)不確かな点を確認する意で自問し、あるいは、相手に問い掛ける語。

(2) (「ないかな」の形で) 願望の意を表す。

『精選版 日本国語大辞典』(小学館、2006年)

(1)文末にあって感動を表わす。中古以後の用法。

(2)疑問をこめた詠嘆や、判断を保留して問いかけたり、自問したりする意などを表わす。近世以後の用法。

そして、自然会話における「かな」の機能を考察した研究として、平山(2015)があげられる。平山(2015)は、「BTSJ 日本語自然会話コーパス」の10代、20代の同世代友人同士の会話データから「かな」の機能を分析し、13種類に分類した(表2)。しかし、「相手に影響を及ぼす自分の行為について述べる」「相手と異なる主張を行う際に、自分の発話を和らげる」「相手への批判を和らげる」「相手への不同意を間接的に示す」という4つの機能において、「かな」の配慮表現の側面が言及されている一方、配慮表現として機能する理由には触れられていない。また、考察対象が若い世代の友人同士の会話であったため、会話参加者の年齢や社会的地位が異なる対人関係における「かな」の使用実態はまだ明らかになっていない。

表2 「かな」の機能 (平山2015に基づいて筆者が作成)

表現類型		機能
働きかけ	依頼	相手に用件を頼む
表出	意志	未確定の自分の行為について述べる
		相手に影響を及ぼす自分の行為について述べる
	願望	自分の願望を表現する
		相手の立場で願望を述べ、共感を示す
述べ立て	疑い	記憶が曖昧なことを話す
		判断がつかないことを示す
		相手や第三者に関する、自信のない判断や認識を述べる
		相手と異なる主張を行う際に、自分の発話を和らげる
		相手への批判を和らげる
		相手への不同意を間接的に示す
問いかけ	判断の 問いかけ	相手が判断できるか不明なことについて、問いかける
		自分の判断に自信がないことについて、相手に確認を求める

### 3. 研究目的と研究課題

平山(2015)を踏まえ、本研究は論文指導の場面に着目し、配慮表現としての「かな」の機能の解明を目的とし、以下の研究課題を設定する。

【研究課題】論文指導の場面において、配慮表現「かな」はどのように使用されているか。

論文指導の場面に着目した理由は、会話参加者の教師・学生の上に上下関係が存在しており、配慮表現の多用が想定されるためである。また、使用実態の分析は質的観点から行い、個々の使用例につき、誰が何に対してどんな理由や目的で配慮を示しているかに焦点をあて、論文指導の場面における「かな」の配慮機能を明らかにする。

#### 4. 分析方法

分析データは『BTSJ 日本語自然会話コーパス 2022 年 3 月 NCRB 連動版』（以下、『BTSJ 日本語自然会話コーパス』）の論文指導という場面の会話データを使用する（表 3）。このコーパスを選んだのは、談話の流れと笑いなどのパラ言語情報が確認できるためである。データの合計会話時間は 5 時間 11 分である。

表 3 分析資料の詳細

会話番号	会話参加者	会話時間
043-03-JFT001-JFSt001	教師（女）、学生（女）	0:23:00
044-03-JMT001-JFSt002	教師（男）、学生（女）	0:42:00
045-03-JMT002-JFSt003	教師（男）、学生（女）	0:22:00
046-03-JMT002-JFSt004	教師（男）、学生（女）	0:22:00
047-03-JMT003-JFSt005	教師（男）、学生（女）	0:26:00
048-03-JMT004-JFSt006	教師（男）、学生（女）	1:04:00
049-03-JMT005-JFSt007	教師（男）、学生（女）	0:30:00
050-03-JMT001-JMSt001	教師（男）、学生（男）	0:40:00
051-03-JMT003-JMSt002	教師（男）、学生（男）	0:16:00
052-03-JMT006-JMSt003	教師（男）、学生（男）	0:26:00
合計		5:11:00

分析は、以下の 3 つの手順で行う。

(1) 使用例の抽出：

「かな」を含む発話文をエクセルの「検索」機能で抽出する。そのうち、「行かない、なかなか」など、終助詞「かな」に属さないものを目視で確認し、対象外とする。また、「かな」は「～かなと思う」などの形式で文中に使われる場合があるため、本研究は文中、及び文末に現れる「かな」全般を分析対象とする。

(2) 機能の整理：

抽出された個々の使用例につき、文脈情報を踏まえて「かな」の機能を確認する。そして、配慮表現に属するものと属さないものを分けて機能をまとめる。

(3) 配慮表現として機能する理由の考察：

ポライトネスの原理と配慮表現との関係、およびポライトネス理論を踏まえて、「かな」が配慮表現として機能する理由を分析して考察する。

#### 5. 結果と考察

論文指導場面の会話データより、「かな」の使用例は 114 件抽出された。これらを教師・学生別に機能をまとめたものが表 4、5 である。教師の発話において、75 件の使用例が

確認され、出現件数の多い順に、「意見・考えを述べる」「回答を強制しない質問をする」「記憶があいまいであったり、言い切れなかったりすることを述べる」「自分自身に問いかける」「助言を行う」「適切な表現の検索をしている状況を知らせる（フィルター）」「相手の意見を完全に受容することができない態度を間接的に示す」という7種類の機能が見られた。また、学生の発話においては、39件の使用例が確認されたが、そのうち、31件(79.5%)が「意見・考えを述べる」という機能に分類された。そのほか、「記憶があいまいであったり、言い切れなかったりすることを述べる」「困惑を示し、相手の回答を間接的に求める」「間違っていた認識を述べる」という機能も観察された。

表4 教師の「かな」の使用

言語行動	機能	「かな」の使用件数 (括弧内は比率、n=75)
述べ立て	意見・考えを述べる	27(36.0%)
	記憶があいまいであったり、言い切れなかったりすることを述べる	15(20.0%)
	自分自身に問いかける	7(9.3%)
	適切な表現の検索をしている状況を知らせる (フィルター)	2(2.7%)
	相手の意見を完全に受容することができない態度を間接的に示す	1(1.3%)
質問	回答を強制しない質問をする	16(21.3%)
助言	助言を行う	7(9.3%)
合計		75

表5 学生の「かな」の使用

言語行動	機能	「かな」の使用件数 (括弧内は比率、n=39)
述べ立て	意見・考えを述べる	31(79.5%)
	記憶があいまいであったり、言い切れなかったりすることを述べる	4(10.3%)
	間違っていた認識を述べる	1(2.6%)
質問	困惑を示し、相手の回答を間接的に求める	3(7.7%)
合計		39

以下では、具体的な会話例（断片1から断片6）を取り上げ、各機能で見られる配慮の表し方を分析する。ただし、「記憶があいまいであったり、言い切れなかったりすることを述べる」「自分自身に問いかける」「適切な表現の検索をしている状況を知らせる（フィルター）」という機能は、配慮表現としての働きが弱いいため、本稿では取り上げない。

### 5.1 意見・考えを述べる

## 断片 1 会話番号 044-03-JMT001-JFSt002 (日本人教師男性、日本人学生女性)

行番号	話者	発話内容
211	JMT001	<しかも外来語に絞る>{>}、あの、和語・漢語とか混種語は直接は扱わないって、ことですか。
212	JFSt002	えっなくはないんですが、問題は、もちろんたくさんありますが、うん、そ、あの一昔の言葉をやるほど大きくは(ふーん)ないかなっという感じですかね、うん。
213	JFSt002	あと、そのやっぱり、そうですね、うーん、そういうふうを考えれば、もともとだめなのかもしれませんが、やっぱり今の現代の外来語というのと、昔のその歌、現代の流行歌というのと、昔のそのあのしだとかそういうものっていうのを一緒に扱う、一緒についているのもどうか、とか(うん)、あと外来、今の外来語、その西洋系の外来語と漢語、その時代の漢語って言うのを、一緒に扱うって言う事の、難しさっていうか(うーん)、なんか想像がつかないっていうか、なんか恐ろしい感じがするん(なるほどね)ですよ。
214	JFSt002	<u>その、あまりにも深くて足を踏み入れると、ちょっとどうなっちゃうかなって言う感じがするんですが。</u>
215	JMT001	ちょっと僕の意図を誤解しているようなんですけどね。

断片 1 において、教師(JMT001)は 211 行目の発話で、「和語・漢語とか混種語は直接は扱わないって、ことですか」と学生(JFSt002)に質問している。213 行目の発話で、学生は「想像がつかない、恐ろしい感じがする」と、外来語と和語・漢語を一緒に扱うことへの抵抗を示している。そのため、214 行目の発話「あまりにも深くて足を踏み入れると、ちょっとどうなっちゃうかなって言う感じがするんですが」も、消極的な意見を示していると判断できる。つまり、「どうなっちゃうかな」の含意は「恐ろしいことになる」と考えられる。学生が「どうなっちゃうかな」というあいまいな言い方を選んだのはなぜであろうか。論文指導の場面においては、学問に関しても研究者としての経験に関しても、教師は学生より上位に位置づけられている。そのため、意見表明の際に、学生は常に不安を抱いており、否定されることを想定した上で話を展開していく。断片 1 の「どうなっちゃうかな」も、形式上では疑問を表す表現であるものの、文脈からみると、これは断定を避け、否定的な意見を和らげる表現である。このように、断定の回避は相手に批判されることを避け、発話者自身の人に認められたいというポジティブ・フェイスを守るための戦略として働いている。また、ポライトネスの原理の観点からみると、断定の回避は「自己と他者との意見相違を最小限にせよ」(Leech 1983:190)という原理を反映している。

小野(2006)は、「かなと思う」という表現を取り上げ、自分以外の人や意見への配慮が表れていると指摘している。鈴木(2015)も、意見表明の際に使われる「かなと思う」の自己防衛の機能を指摘しているが、両者の分析で使用されている例文はすべて「命題+かなと思う」という形式を取っている。一方、断片 1 ではもう一つの表現形式が見られ、疑問詞と「かな」が接続する場合も、文脈によっては意見表明の表現として捉えられる。さらに、意見表明の表現として、「かなという感じがする」という表現も日常会話で活用されていることがわかる。

5.2 相手の意見を完全に受容することができない態度を間接的に示す

断片2 会話番号 048-03-JMT004-JFSt006 (日本人教師男性、日本人学生女性)

行番号	話者	発話内容
318	JMT004	《少し間》んで、でこういう場合たとえば、これは一、何だって?、ど、てへ、これはリズムがどうみ、とめられるって、考えるの?。
319	JFSt006	《少し間》これは一(うん)、3年生の終わりのときに書いたレポートでは一(うん)、あのーリズムはないと言っていました。
320	JMT004	あーそうなの<か>{<}</>。
321	JFSt006	<あ、はい>{>}</>、でも、これも対照してて、鏡みたいだから面白いと思ったんです>{<}</>。
322	JMT004	<なるほど>{>}</>。
323	JFSt006	だからってどうすればいいかわかんなくて、<それで>{<}<笑いながら>,,
324	JMT004	<だからって>{>}</>どうすればいいかわかん<ない?>{<}</>。
325	JFSt006	<それで>{>}</>詩の一、その詩のこかく中に当てはまらないってということで、それで除外してた。
326	JMT004	<u>でもさ、そっかな。</u>
327	JMT004	ん、明らかに僕はリズム一、僕にはリズム、に見えるんだけどね<途中から笑いながら>。

断片2は、ある詩のリズムに関する議論である。学生(JFSt006)は、319行目で「リズムはない」と話しているが、教師(JMT004)は327行目で「ん、明らかに僕はリズム一、僕にはリズム、に見えるんだけどね」と、不同意を示している。そのため、その前の326行目の「そっかな」は、疑問を表す表現というより、不同意を表す表現として考えられる。このように、否定的態度を明示せず、疑問の形で表出するのは、相手の人に邪魔されたくないというネガティブ・フェイスへの配慮である。また、意見相違を最小限に抑えながら、不完全受容な態度を示す言語行動もポライトネスの原理「自己と他者との意見相違を最小限にせよ」(Leech 1983:190)を反映している。

5.3 回答を強制しない質問をする

断片3 会話番号 043-03-JFT001-JFSt001 (日本人教師女性、日本人学生女性)

行番号	話者	発話内容
132	JFSt001	えーっと、なにか、そういう、あの一、ちょっと古いんですけど、《少し間》まあVCDとDVDについて、どういう影響があるかっていうのをちょっと、なんか、表かなんかも混ざって書いてあった思うんですけど。
133	JFT001	<u>それはどの資料かな?。</u>
134	JFSt001	えーと、まだ、<まだ書い>{<}</>てない。

断片 3 の学生(JFSt001)は、香港映画について論文を書いている。132 行目で、学生は「VCD と DVD について」「表かなんかも混ざって書いてあった」と話している。それに対して、教師(JFT001)は 133 行目で「それはどの資料かな?。」と質問している。平山(2015:76)は、「相手が判断できるか不明なことについて、問いかける」という「かな」の機能を指摘しているものの、この質問は判断できるか不明なことではない。論文を書いた学生は論文の内容を把握しており、その質問に答えられると想定される。実際、学生も質問を受けた後、明確な答えを述べている。このような点から、教師は学生が答えを知っていると想定していて「かな」を用いて質問していると推測できる。また、学生に質問する際に、「それはどの資料?」「それはどの資料なの?」などの聞き方も考えられるが、「かな」は、独話の性質を帯びているため、相手への働きかけが弱くなり、比較的にやわらかい質問を作ることができる。このように、相手への働きかけを和らげるのは、相手の他者に踏み込まれたくない、強制されたくないというネガティブ・フェイスへの配慮として捉えられる。

#### 5.4 助言を行う

断片 4 会話番号 048-03-JMT004-JFSt006 (日本人教師男性、日本人学生女性)

行番号	話者	発話内容
482	JMT004	で、うーんと、僕のアイデアは、ここにはね、階層構造があるから、こういう階層構造があるから、その階層構造を、この図を参考にしてもいいからね(はい)、あの、階層構造ってのは、樹形図で書くことが、できるんだよね、わかりやすく。
483	JFSt006	<はい>{<}
484	JMT004	<だから>{>}、《少し間》詩学の文脈でね、うーんと、一般的に、ロシア詩っていうのは、こういう、構成要素で、こういう構造をもっているんだっていうことが書かれているけれども、それを、樹形図を用いて、わかりやすく、えーっと、概説しなすと、 <u>こういうことになるっていうやり方で、書いてみたらどうかな。</u>
485	JFSt006	そうすることで...
486	JMT004	うん。

断片 4 は、教師(JMT004)が学生(JFSt006)の論文の書き方について助言を行う断片である。482、484 行目の発話は教師の具体的な説明であるが、484 行目の終わりに、「こういうことになるっていうやり方で、書いてみたらどうかな」と、学生への働きかけが見られる。「かな」をつけずに、「書いてみたらどう?」と述べることもできるが、「かな」は独話の性質を持っているため、助言に使うと相手への働きかけが弱くなり、相手の行動への束縛も弱い。したがって、助言における「かな」の使用には相手の領域に踏み込まない態度、ひいては相手のネガティブ・フェイスへの配慮が見受けられる。また、「かな」には「判断がつかない」という意味が含まれているため、断定回避の表現としても働き、ポライトネスの原理「自己と他者との意見相違を最小限にせよ」を反映している。意見相違を最小限にすることも、対人関係を良好に保つための配慮であると考えられる。

### 5.5 間違っていた認識を述べる

断片 5 会話番号 044-03-JMT001-JFSt002 (日本人教師男性、日本人学生女性)

行番号	話者	発話内容
4	JMT001	えっとでも、あの一、あ、だいたいこういう距離には座らないですね、普通。
5	JFSt002	そう、そうなんですか？。
6	JMT001	え、そうですね、僕、はてな？、どうしてそんなふうに着てんだらう。
7	JFSt002	いや、<あの>{<},,
8	JMT001	<僕、たぶん>{>}【】。
9	JFSt002	【】カメラで、,
10	JMT001	え。
11	JFSt002	<u>こういう感じが&lt;いいのかなと思っ、考えただけなんですけど&gt;{&lt;}&lt;軽い笑い&gt;。</u>
12	JMT001	<いや、あ、そうですね、あの一でも>{>}、普通こう座らないと思いますよ。

断片 5 は面談の最初のやりとりであり、教師(JMT001)と学生(JFSt002)は二人の座り方について話している。4行目と6行目で、教師は「だいたいこういう距離には座らないですね」「どうしてそんなふうに着てんだらう」と、学生が主張していた座り方に反対意見を表明している。それを受け、学生は11行目で、「こういう感じがいいのかなと思っ、考えただけなんですけど」と説明している。「かな」を使用しない場合、「こういう感じがいいと思っ、考えただけなんですけど」という応答は、頑固で失礼な印象を相手に与え、自分の認められたいというポジティブ・フェイスを満たすこともできなくなってしまう。この発話における「かな」は、もともとの考え方をあいまいにすることで、譲歩の姿勢を見せるマーカーとなっている。同時に、ポライトネスの原理「自己と他者との意見相違を最小限にせよ」(Leech 1983:190)を反映しており、対人関係への配慮を示している。

### 5.6 困惑を示し、相手の回答を間接的に求める

断片 6 会話番号 044-03-JMT001-JFSt002 (日本人教師男性、日本人学生女性)

行番号	話者	発話内容
166	JFSt002	こうマクロ的なところから(はい)、最初アプローチしていくのか(はい)、それともその個々の現象をこうミクロからこうみていったらいいのかっていうのが、 <u>どうすればいいのかなという感じなんですけど。</u>
167	JMT001	まあ、話聞いていると、それよりもっと根本的なところがまだ定まっていないと思うんですけどね。
168	JMT001	えっとですね、修士論文まででは、あの一要するにバリエーションについてやったっていうわけなんですよ、バリエーションとは時代的なバリエーションだったんですけどね。
中略(JMT001の発話)		
175	JMT001	で、中にバリエーションを含まないものを考えるって事だと思いますよ。



断片 6 において、学生(JFSt002)は 166 行目で「どうすればいいのかなという感じなんですけど。」と自分の戸惑いを示している。この文は「なんですけど」で終わったが、疑問文の「どうすればいいのでしょうか」と等しい効力を持っている。なぜなら、論文指導の場面において、学生に助言を提供することは教師側の義務とされている。そのため、学生側は明示的に助言を求めなくても、教師からのアドバイスをもらうことが可能である。実際、教師(JMT001)は 167 行目から 175 行目にかけて、学生の疑問を解決するための指導を行っている。したがって、166 行目の「かな」は、自分自身の困惑を示し、相手のアドバイスを間接的に求めるという機能を果たしている。

なぜ学生が疑問文の使用を避けたのかというと、疑問文は相手に応答を要請する効力を持っており、相手の領域、ひいては相手のネガティブ・フェイスを脅かすからだと考えられる。ポライトネスの原理には、「他者の負担を最小限にせよ」(Leech 1983:190)という項目があるが、困惑を示して相手の回答を間接的に求めることも一つのストラテジーである。

## 6. 結論

以上が、論文指導の場面における「かな」の使用実態の考察である。教師の発話において、「意見・考えを述べる」「回答を強制しない質問をする」「助言を行う」「相手の意見を完全に受容することができない態度を間接的に示す」という 4 種類の配慮機能が見られた(表 4 再掲)。学生より上の立場にいるにもかかわらず、教師は「かな」を活用し、学生への配慮、および良好な人間関係への配慮を示している。

表 4 (再掲) 教師の「かな」の使用

言語行動	機能	「かな」の使用件数 (括弧内は比率、n=75)
述べ立て	意見・考えを述べる	27(36.0%)
	記憶があいまいであったり、言い切れなかったりすることを述べる	15(20.0%)
	自分自身に問いかける	7(9.3%)
	適切な表現の検索をしている状況を知らせる (フィラー)	2(2.7%)
	相手の意見を完全に受容することができない態度を間接的に示す	1(1.3%)
質問	回答を強制しない質問をする	16(21.3%)
助言	助言を行う	7(9.3%)
合計		75

また、学生の発話においては、「意見・考えを述べる」「困惑を示し、相手の回答を間接的に求める」「間違っていた認識を述べる」という配慮機能が見られた(表 5 再掲)。会話データより、学生も「かな」を活用して自分の意見を柔軟に述べながら、教師の意見を求めていくことが確認された。

表5 (再掲) 学生の「かな」の使用

言語行動	機能	「かな」の使用件数 (括弧内は比率、n=39)
述べ立て	意見・考えを述べる	31(79.5%)
	記憶があいまいであったり、言い切れなかったりすることを述べる	4(10.3%)
	間違っていた認識を述べる	1(2.6%)
質問	困惑を示し、相手の回答を間接的に求める	3(7.7%)
合計		39

使用件数からみると、「意見・考えを述べる」という機能の使用件数が最も多いことがわかる。これは、論文指導場面の特徴を反映していると考えられる。終助詞は、「聞き手との関係性に応じて使い分けられる」(白岩 2011:66)ため、論文指導以外の場面において、年齢差や親疎関係などの要素が「かな」の使用に与える影響についての考察を今後の課題としたい。

### 謝 辞

本研究は2023年度尚友倶楽部筑波大学日本語教育研究者育成奨学金の支援を受けたものである。

### 文 献

- 小野正樹 (2006). 「新しい文法－「かなと思う」について－」 『日本語学』 25:9, pp.46-56.
- 白岩広行 (2011). 「第二言語としての日本語の終助詞習得研究の展望」 『阪大社会言語学研究ノート』 9, pp.66-95.
- 小学館国語辞典編集部(編) (2006). 『精選版 日本国語大辞典』 小学館.
- 新村出(編) (2018). 『広辞苑 (第七版)』 岩波書店.
- 鈴木智美 (2015). 「意思表示に用いられる「かなと思う」－対立・摩擦を避け内に向かう言葉－」 『留学生日本語教育センター論集』 41, pp.61-78.
- 中西久実子 (2015). 「終助詞「かな」の語用論的特徴－「非難」「ぼかし」の用法－」 『無差』 22, pp.23-38.
- 野田尚史・高山善行・小林隆(編) (2014). 『日本語の配慮表現の多様性』 くろしお出版.
- 平山紫帆 (2015). 「自然会話における終助詞「かな」の用法」 『日本語教育実践研究』 2, pp.68-79.
- 松村明(編) (2019). 『大辞林 (第四版)』 三省堂.
- 山岡政紀 (2015). 「慣習化されたポライトネスとしての配慮表現の定義」 『日本語用論学会第17回大会発表論文集』 10, pp.315-318.
- 山岡政紀(編) (2019). 『日本語配慮表現の原理と諸相』 くろしお出版.
- 山下悠貴乃 (2012). 「依頼における文末形式「かな」「と思ったり」「と違って」の配慮表現としての機能について(第2部 発話機能論から配慮表現研究への応用と対照研究)」 『日本語コミュニケーション研究論集』 2, pp.119-127.

Brown Penelope, and Stephen Levinson(1987).*Politeness:Some universals in language usage*.Cambridge University Press.[邦訳：田中典子（監修）(2011)『ポライトネス:言語使用における、ある普遍現象』研究社.]

Leech Geoffrey(1983).*Principles of Pragmatics*.Longman.[邦訳:池上嘉彦・河上誓作(訳)(1987)『語用論』紀伊國屋書店.]

#### 調査資料

宇佐美まゆみ監修(2022) .『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2022年3月 NCRB 連動版』，国立国語研究所，機関拠点型基幹研究プロジェクト「日本語学習者のコミュニケーションの多角的解明」

## 程度副詞使用実態の横断的・縦断的調査 —「通時話し言葉コーパス」の試み—

日暮 康晴（筑波大学大学院生）<sup>†</sup>

### A Cross-sectional, Diachronic Survey of the Usage of Adverbs of Degree: Possibilities of the “Diachronic Spoken Language Corpus”

HIGURE Yasuharu (Graduate School, University of Tsukuba)

#### 要旨

本研究では『昭和話し言葉コーパス』（SSC）、『日本語話し言葉コーパス』（CSJ）、『日本語日常会話コーパス』（CEJC）を使用し、程度副詞「とても」とその類義語の使用実態についての横断的・縦断的調査を実施した。横断的調査の結果、SSC内の比較からは独話環境では「ひじょうに」の使用が特に多く、会話環境では複数の語が同程度使用されること、CSJ・CEJCの比較からは独話環境では「ひじょうに」、会話環境では「けっこう」、「すごく」の使用が特に多いことが分かった。縦断的調査からは、語によって使用傾向の変化に異なりがあり、使用頻度が高い語の中でも使用場面が狭まる語と広がる語に分かれることが分かった。また、主な使用場面が変化する語も確認された。使用場面が通時的に広がる語に注目した用例の検討からは、使用傾向の変化には類義語間の選択傾向の変化だけでなく、程度副詞の談話標識化という用法自体の変化が関わっていることが示唆された。

#### 1. はじめに

本研究は、日本語の副詞、特に「とても」や「ひじょうに」といった被修飾語の意味を高精度に限定し示す類義語群（以下、「とても」類の語と総称する）に注目し、話し言葉の場面状況による使用実態の違いとその通時的な変化に注目して調査を行う。「とても」類の語は他の副詞と同様、文体や場面状況によって使い分けがなされ、加えて、これらの語群には新語の発生・流入が激しいという特徴がある（日本語教育学会編 2005）。その近年の例には「ちょう」や「めっちゃ」がある。新語の発生は若者言葉などくだけた語の使用として捉えられることが多いが、通時的な観点からは、新語・新用法の発生は単にくだけた場での言葉遣いに限らない。例えば、「とても」の「とても面白い」のような程度副詞としての用法は2023年現在では一般的に使用されるが、これは「とても食べられない」のような否定と共起する用法から転じて1920年ごろに発生したものであり、発生当時は「はやり言葉的な用法」（播磨 1993:15）であった（播磨 1993）。

このように、「とても」類の語の使用傾向の変化は、どのような場面で使用されるのかという横断的な要素と、それが時間の経過に伴いどう変化するかという通時的（縦断的）な要素を含み得るものである。本研究では、話し言葉データによって構築された『昭和話し言葉コーパス』（SSC）、『日本語話し言葉コーパス』（CSJ）、『日本語日常会話コーパス』（CEJC）の3コーパスを使用し、日本語話し言葉における「とても」類の語の使用実態の場面による異なりとその通時的な変化の調査を試みる。

<sup>†</sup> higure.j☆gmail.com（☆を@へ変更）

## 2. 先行研究

「とても」類の語を含む日本語副詞の使用実態についての研究は、現代語を中心に近年盛んに行われている。雑談場面に注目した中俣（2016）、石川（2020）では、共通して日本語母語話者は「けっこう」、「すごく」を高頻度で使うことが明らかになっている。また、独話環境内の場面の異なりに注目した日暮（2023）では、個人的な内容についての比較的くだけたスピーチでは「すごく」、「とても」が高頻度で、学会での研究発表では「ひじょうに」が高頻度で使われるといったように、独話環境という共通条件の中でも、さらにその中で詳細な条件の違いによって「とても」類の語の使用傾向が変わり得ることが明らかになっている。ただし、これらの研究は使用頻度の高い語に注目した分析が中心になっていたり（中俣 2016、石川 2020）、限られた数の語のみを対象にしていたり（日暮 2023）と、より広範な「とても」類の語の使用実態を調査する余地は未だ残されている。

加えて、通時的な側面の先行研究には「とても」に注目した播磨（1993）・吉井（1993）や、程度副詞の程度性の発生についてまとめた鳴海（2013）などはあるものの、比較的長いスパンでの語史的なものであり、また、その調査対象も書き言葉が中心であった。前述のように程度副詞、特に本研究が注目する「とても」類の語の使用実態の変化は比較的短期間の間に起こることが予想される。これまでの研究対象が書き言葉に限られてきたことはデータ収集（録音）が不可能だったなど技術的な面によるものが大きいですが、近年では後述する複数の話し言葉コーパスの整備によって横断的・縦断的両方の視点を備えた話し言葉調査が可能となった。

さらに、語の使用頻度の通時的な変化を分析するにあたって、使用の多寡のみならず、用法そのものの変化も予想される。孫（2018）では「すごく」のうち型式[スゴイ]に、程度副詞としての具体的な意味が希薄化したフィルター的な用法が発生していることを指摘している。また、原田（2022）は「かなり」、「けっこう」に、被修飾語とされる語との間に統語的な距離があき、その用法もフィルター的なものが発生していることを指摘している。しかし、以上の研究は語の中の一型式や、類義語の一部に焦点を当てるにとどまっており、類義語内のどれくらいの範囲で同様の用法変化がみられるのかは明らかではない。語の使用傾向の変化を分析するにあたって、このような細かな用法の変化にも注目する必要がある。

本研究の研究課題は、RQ1「日本語話し言葉における程度副詞の使用傾向は発話環境によって異なりがあるか」、RQ2「日本語話し言葉における程度副詞の使用傾向の発話環境による異なりは、通時的に変化しているか」の2つである。RQ1については本稿執筆時期に近いCSJ・CEJCの結果に注目して分析を行い、RQ2はCSJ・CEJCの結果とSSCの結果との比較によって解明を目指す。

## 3. 調査手続き

### 3.1 使用コーパス情報

前述の通り本研究では『昭和話し言葉コーパス』(SSC)、『日本語話し言葉コーパス』(CSJ)、『日本語日常会話コーパス』(CEJC)の3コーパス(いずれも国立国語研究所によって公開)を使用する。これらのコーパスは日本語の話し言葉データによって構築されているという点で共通するが、データの収録対象となった言語使用場面や収録が行われた時期が異なっており、これらの比較を行うことで過去60-70年の中での日本語の横断的・縦断的調査が可能となる(丸山・小磯・西川 2022)。丸山・小磯・西川(2022)は、このような日本語話し言葉の通時的調査に向けたSSCとCSJ・CEJCといった現代のコーパスの連結を「通時話し

言葉コーパス」(丸山・小磯・西川 2022;198)と呼ぶ。ただし、各コーパスの中でも録音された場面状況は複数あり、比較分析のためには、これらの条件を確認し、可能な限り統制を行う必要がある。各コーパスにおいてそれら録音における場面状況の異なりはタグ付け(SSC、CEJCでは「形式」、CSJでは「音声タイプ」)による分類がなされており、本研究ではより精緻な比較のために、独話・会話環境それぞれにおいて類似した内容のタグ付けがなされたデータを絞り込んで比較分析を行った。具体的には、SSCからは独話環境データとして形式〔講演〕、会話環境データとして形式〔雑談〕、CSJからは音声タイプ〔学会講演〕、CEJCからは形式〔雑談〕のタグが含まれるデータを使用した。以下、それぞれの詳細をまとめる。なお、「形式」(SSC・CEJC)、「音声タイプ」(CSJ)で示される場面状況を本稿では統一して「場面」と呼ぶ。

SSC 独話環境データとして選択した形式〔講演〕は、「創立記念講演会などにおける学術的な講演」(丸山・小磯・西川 2022;213)の録音データによって構築されている。録音された講演の内容は「現代の敬語意識」(M54\_10\_LT)や「明治初期の書きことば」(M58\_12\_LA)のようにすべて言語・日本語に関するものである<sup>1</sup>。また、50名の話者のうち49名が男性、女性は1名と、話者の性差にも偏りがある(丸山・小磯・西川 2022)。SSC〔講演〕に対応するCSJの独話環境音声タイプは〔学会講演〕である(丸山・小磯・西川 2022)。CSJの〔学会講演〕は、実際に開かれた学会での研究発表の録音データで構築されている(国立国語研究所 2006)。SSC〔講演〕が言語・日本語に関するものに限られていたのに対し、CSJ〔学会講演〕学会は「理工学、人文、社会の3領域に及ぶ種々の学会」(国立国語研究所 2006;4)での研究発表である。次に、会話環境のデータとしてはSSC・CEJCの両方から型式〔雑談〕を選択した。両者は共通した基準でタグ付けがなされており、いずれも「会話の目的や話題などがあらかじめ定められていない会話」(小磯・土屋・渡部・横森・相澤・伝 2016;88)と定義付けられている(丸山・小磯・西川 2022)。

以上に挙げた、本研究で使用する3コーパスの4場面について、表1にそれぞれの収録年・語数情報を付してまとめる。

表1 使用コーパス情報

コーパス	調査対象場面	環境	収録年	総語数 <sup>2</sup> (記号入り)
SSC	講演	独話	1955-1969年	122,287
	雑談	会話	1952-1960年	306,745
CSJ	学会講演	独話	1999-2004年	284,553
CEJC	雑談	会話	2016-2020年	1,661,228

表1に確認できるように、本調査で取り扱うSSC〔講演〕は15年間、〔雑談〕は19年間に収集されたデータである。また、CSJとCEJCの収録期間はそれぞれ5年程度だが、両者の間には約12-20年の開きがある。本研究ではコーパス間の比較を中心とし、これらのコーパス内での収録年の違いには着目しない。また、講演の個別の内容(SSC・CSJ)、話者の性

<sup>1</sup> 『『独話』(50ファイル)の一覧』(SSC\_files\_M.pdf) (2023.08.15最終閲覧)より。

<sup>2</sup> 使用したコーパス語数データの出典は本稿末に掲載する。

別 (SSC・CSJ・CEJC) についても本研究では分析対象としないため、頻度の調整などによる比較は行わない。

以後、本稿において SSC [講演] で得られた結果は「SSC [独話]」、SSC [雑談] で得られた結果は「SSC [会話]」、CSJ [学会講演] で得られた結果は「CSJ」、CEJC [雑談] で得られた結果は「CEJC」と呼称する。

### 3.2 調査対象語

本研究で調査対象とする「とても」類の語は被修飾語の意味を高程度に限定し示す語と定義されるが、その中には「とても」や「ひじょうに」のように類義語が多くある。本研究では同様の意味内容を表す類義語として飛田・浅田 (2019)、中俣 (2020) を参考に 18 語を選定し、加えて、近年、特に若年代に使用が多くみられる「めっちゃ」を加えた 19 語、「おおいに」、「かなり」、「きわめて」、「けっこう」、「ごく」、「しごく」、「ずいぶん」、「すごく」、「そうとう」、「たいそう」、「だいぶ」、「たいへん」、「ちょう」、「とても」、「なかなか」、「はなはだ」、「ひじょうに」、「ひどく」、「めっちゃ」を調査対象語とした。これらの語は程度限定という点で共通するものの、詳細な語感・ニュアンスのレベルでは、評価的な視点の有無など異なりがあることが先行研究で指摘されており (渡辺 1990 など)、まったくの同レベルで入れ替え可能な語ではない。しかし、本研究ではそのようなニュアンスも使用実態に影響するという立場から、それらの差異を踏まえて比較分析・考察を行う。

### 3.3 調査の流れ

コーパス検索アプリケーション『中納言』(国立国語研究所) を使用し、各コーパスにおける各調査対象語の用例検索・結果のダウンロードを行った。実際の口頭産出の中では、同じ語でも [トテモ] / [トツテモ]、[スゴク] / [スゴイ] / [スゲー] のように様々な型式で発話される。そこで、同語と認められる中での複数型式を一括して検索できるように本研究では語彙素検索を採用した。ただし、語彙素検索で得られる結果では各用例の用法は統制できない。たとえば、語彙素「凄い」による検索を行うと、本研究で分析の対象とする「すごく面白い」といった副詞用法の他に、「すごい本」という形容詞用法、「人としてのすごさ」という名詞用法など様々な用法を含んだ結果が得られる。本研究ではダウンロードした全用例の用法を確認し、その分類を行った。その結果、程度副詞用法の他に陳述副詞用法(「とても (+否定)」、「なかなか (+否定)」)、形容詞用法(「けっこうな」、「すごい」など)、連体詞用法(「かなりの」、「ちょう」など)、動詞用法(「そうとうする」など)、名詞用法(「すごさ」、「たいへんさ」など)、メタ用法、用法不明の 7 用法が確認された。メタ用法は、例 (1) のように、その語自体に注目し、言及する中で発話されたものである。用法不明は、言いよどみや言い直し、会話参加者の割り込みなどによる発話の中断を分類した。本研究では以上の程度副詞用法以外の用法は除去し、程度副詞用法のもののみを分析対象とした。

- (1) 「どうもとてもが変な位置になってしまっているということで」

(CSJ\_A03M0016)<sup>3</sup>

<sup>3</sup> 本稿では用例の掲載に際し、調査対象として注目する「とても」類の語に下線を付し、また、引用末にコーパス名および録音データ ID (「コーパス名\_ID」) を記す。また、「とても」類の語の表記はひらがなに統一し、語の型式のゆれは各コーパス内の「キー」表記に準じる。

#### 4. 結果と考察

本章では、まず全体での副詞使用数結果を示した後に、コーパス間の比較による分析及び考察を行う。分析・考察の順番としては、まず SSC [独話] と SSC [会話] の比較、CSJ と CEJC の比較によって横断的な分析・考察を行い、次いで、SSC [独話] と CSJ、SSC [会話] と CEJC の比較を行い、縦断的な分析・考察を行う。最後に横断的・縦断的な視点の双方を合わせた総合的な考察を行う。

##### 4.1 全体における「とても」類の語の使用傾向

表2に、コーパス別の副詞使用数の集計結果をまとめる。3.3節で言及した語彙素内の異型式は合算した。各セルにはコーパス内で確認された実使用数である粗頻度と、各コーパス別に算出した調整頻度（10万語あたりの使用数）を示す。調整頻度算出には表1に示した総語数を使用した。

表2 コーパス別使用数集計結果（粗頻度・調整頻度）

	SSC [独話]	SSC [会話]	CSJ	CEJC
おいに	7 (5.72)	3 (0.98)	23 (0.69)	2 (0.12)
かなり	20 (16.35)	28 (9.13)	932 (28.05)	88 (5.30)
きわめて	13 (10.63)	2 (0.65)	101 (3.04)	0 (0.00)
けっこう	0 (0.00)	22 (7.17)	253 (7.61)	1534 (92.34)
ごく	10 (8.18)	14 (4.56)	72 (2.17)	1 (0.06)
しごく	0 (0.00)	0 (0.00)	1 (0.03)	0 (0.00)
ずいぶん	6 (4.91)	165 (53.79)	94 (2.83)	99 (5.96)
すごく	0 (0.00)	118 (38.47)	132 (3.97)	2299 (138.39)
そうとう	10 (8.18)	67 (21.84)	28 (0.84)	70 (4.21)
たいそう	0 (0.00)	0 (0.00)	2 (0.06)	1 (0.06)
だいぶ	6 (4.91)	67 (21.84)	88 (2.65)	164 (9.87)
たいへん	24 (19.63)	40 (13.04)	211 (6.35)	17 (1.02)
ちょう	0 (0.00)	0 (0.00)	1 (0.03)	238 (14.33)
とても	1 (0.82)	126 (41.08)	93 (2.80)	51 (3.07)
なかなか	13 (10.63)	55 (17.93)	125 (3.76)	91 (5.48)
はなはだ	3 (2.45)	0 (0.00)	6 (0.18)	0 (0.00)
ひじょうに	203 (166.00)	111 (36.19)	2798 (84.20)	17 (1.02)
ひどく	1 (0.82)	2 (0.65)	2 (0.06)	2 (0.12)
めっちゃ	0 (0.00)	0 (0.00)	0 (0.00)	570 (34.31)
合計	317 (259.23)	820 (267.32)	4962 (149.33)	5244 (315.67)

(カッコ内の調整頻度は10万語あたりの使用数)

表3に示した結果より、各グループにおける調査対象語全体の使用頻度はCEJC、SSC [会話]、SSC [独話]、CSJの順に高い。個々の語の使用頻度に注目すると、SSC [独話]では「ひじょうに」、SSC [会話]では「ずいぶん」、CSJでは「ひじょうに」、CEJCでは「すごく」がそれぞれのコーパス内での最頻出語であることがわかる。個別語の使用頻度については以下4.2節、4.3節で詳しく確認し、ここでは全体の傾向を述べるにとどめる。

全体の傾向として、通時的に会話環境の方が独話環境に比べて副詞が使用されやすいこと、また、SSC [独話] → CSJでは頻度が下がり、SSC [会話] → CEJCでは上がっていることから、通時的に会話環境では副詞の使用が増加し、独話環境では減少したとまとめること



ができる。ただし、SSC [独話] の内容が言語・日本語についての講演である一方で、CSJ の内容はより広い分野での学会発表である。SSC [独話] を構成する講演の中には、例 (2) のようにややリラックスした話し方も含まれる一方で、後者ではデータ量での比較など、そもそも程度副詞を使用しないような、より客観的な表現が使用されている<sup>4</sup>といったように、収録データの差異が結果に表れている可能性もある。

- (2) これはまあ効能書きを述べ立てようとするずいぶんいろいろな  
ことがあるわけではありますが (SSC\_M58\_14\_LA)

なお、全コーパス結果の合計値の中で使用率が 1%未満となる「おおいに」(35/11344 = 0.31%)、「ごく」(97/11344 = 0.86%)、「しごく」(1/11344 = 0.01%)、「たいそう」(3/11344 = 0.03%)、「はなはだ」(9/11344 = 0.08%)、「ひどく」(7/11344 = 0.06%) は以下の分析・考察の対象外とする。

## 4.2 横断的比較

### 4.2.1 SSC [独話] と SSC [会話] の比較

表 3 に示した結果より、SSC [独話] と SSC [会話] それぞれの調整頻度から作成したグラフを図 1 に示す。

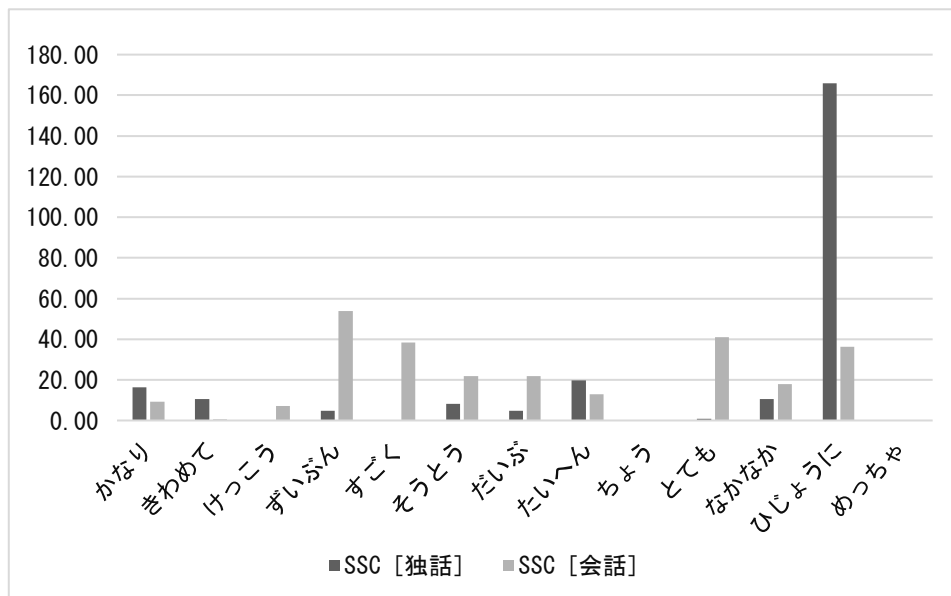


図 1 SSC [独話]、SSC [会話] における副詞使用頻度 (調整頻度より)

表 3 および図 1 の結果より、調査対象語合計の調整頻度は 259.23 (SSC [独話])、267.32 (SSC [会話]) と近似しているものの、語の分布という点では SSC [独話] では「ひじょうに」の 1 語に使用が集中しており、SSC [会話] では複数の語に使用が分散しているように、使用される語の傾向に違いがある。SSC [会話] で使用が比較的多い語には、多い順に「ずいぶん」、「とても」「すごく」、「ひじょうに」が挙げられる。特に「すごく」は

<sup>4</sup> CSJ 内の独話環境音声タイプ間の比較を行った日暮 (2023) においても、音声タイプ [学会講演] では個人のスピーチである [模擬講演] や、専門家の非専門家に対する講演である [その他講演] に比べ程度副詞そのものの使用量が少ないことが報告されている。

SSC [独話] での使用はゼロ、「とても」の使用は1回のみ、「ずいぶん」は6回と、SSC [会話] で使用が多い語は SSC [独話] での使用が少ないという傾向が認められる。一方で、SSC [独話] で使用が多い「ひじょうに」は SSC [会話] でも使用がみられる。

以下、各語の用法を、特に使用頻度の高い語に注目してまとめる。

SSC [会話] の中で最も高頻度で使用が確認された「ずいぶん」は、例 (3)、(4) のように、話し手による実感を伴った (飛田・浅田 2019) 評価的な語感を比較的強く含む語である。

(3) あの頃はずいぶん注射が流行りましたね (SSC\_C52\_10\_CT)

(4) ずいぶんあのきれいな方ですね (SSC\_C52\_12\_CT)

次いで頻度が高い「すごく」、「とても」は例 (5)、(6) のように、話し手による実感に基づく点は変わらないが、「ずいぶん」に比べ評価の印象はやや弱い。

(5) 富士山がすごくきれいなね (SSC\_C56\_04\_CT)

(6) (筆者注：病院の話) そして中はとてもきれいですものね  
(SSC\_C52\_14\_CT)

「ひじょうに」は SSC [独話] に限らず、SSC [会話] でも「すごく」の次に多い、全体では4番目に高い頻度で使用が確認された。例 (7) は SSC [独話]、例 (8)、(9) は SSC [会話] での「ひじょうに」の例である。会話環境では例 (8)、(9) のように、比較的あらたまった発話の中で使用される傾向にある。

(7) このコムは元来ヴァンダービークのこのお一百万語の調査でも  
使用率がひじょうに高いものでありますが (SSC\_M54\_12\_LT)

(8) (筆者注：身長の話) ええ今はね小学校の生徒なんかもひじょうに  
高くなりました (SSC\_C57\_23\_CT)

(9) 若い頃外国語をやるときにいつでもあのイエスノーでひじょうに  
苦勞するんですね (SSC\_C52\_20\_CT)

#### 4.2.2 CSJ と CEJC の比較

表3に示した結果より、CSJ と CEJC それぞれの調整頻度から作成したグラフを図2に示す。

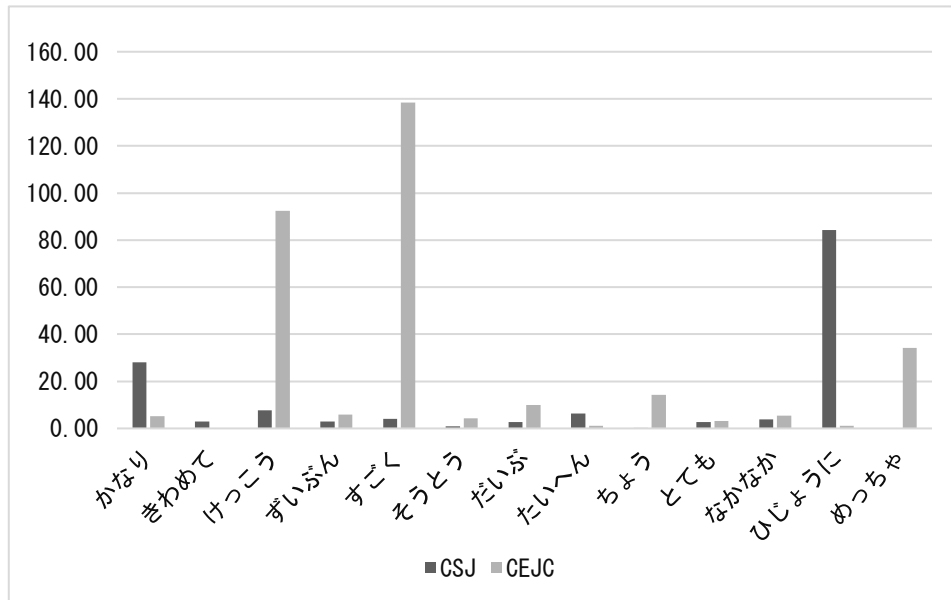


図2 CSJ、CEJCにおける副詞使用頻度 (調整頻度より)

表3 および図2の結果より、CSJとCEJCとではCEJCの方が副詞の使用頻度が高いこと、また、個別の語の使用頻度ではCSJは「ひじょうに」、「かなり」に、CEJCは「すごい」、「けっこう」、そしてやや頻度は下がるが「めっちゃ」に使用が集中している。

CSJ内で比較的高頻度の「ひじょうに」、「かなり」の例をそれぞれ例(10)、(11)に挙げる。この2語はいずれも客観的な表現という点で共通しており(飛田・浅田2019)、それゆえに研究内容の発表という場で使用が多く確認されたと考えられる。意味的な側面では、「ひじょうに」は基本的な程度限定の語、「かなり」は評価・比較の語感を含む語(飛田・浅田2019)という点で異なっており、話し手の含意によって使い分けられていると推定される。

- (10) これひじょうに曖昧な定義なんですけれども (CSJ\_A03F0108)  
 (11) かなり曖昧な基準なんです (CSJ\_A04M0883)

CEJCで最も高頻度で確認された「すごい」、「けっこう」の例を例(12)、(13)に挙げる。「けっこう」は話者が抱く予想や基準に比べて程度が高いというニュアンスを含む(飛田・浅田2019)が、実際の例でも「すごい」に比べ話者の基準による評価というニュアンスが認められる。

- (12) すごい仲のいいお母さんらしいんだよ (CEJC\_T021\_010b)  
 (13) けっこう仲よかった高校卒の子がいたから (CEJC\_T017\_016)

以上、CSJ、CEJCの結果及び用例の検討からは、CSJ・CEJCの期間においては基本的な程度限定の語として独話環境では「ひじょうに」、会話環境では「すごい」が選択され、話し手による評価・比較というニュアンスをより含ませる表現として独話環境では「かなり」、会話環境では「けっこう」が使われるといった使い分けの分布が示唆される。

また、CEJC内における「めっちゃ」は例(14)のように程度副詞としての用法自体は「けっこう」、「すごく」と変わらないものの、例(15)、(16)のようにくだけた会話での使用が比較的多く確認される。本研究では話者の年代に関して詳しい分析は行わないものの、10代、20代の若年代に使用が多くみられるという特徴も確認された。

- (14) それで入院してたらやっぱ看護婦さんとかもみんなめっちゃ忙しそう  
 だった (CEJC\_K002\_012)
- (15) (筆者注：アトラクションの待ち時間の話)  
 インディージョーンズめっちゃ長くない 死ぬくらい長い  
 (CEJC\_T018\_006b)
- (16) 妹にウ彼氏がいるんだけどめっちゃマッチョなの (CEJC\_T009\_005b)

### 4.3 縦断的比較

#### 4.3.1 SSC [独話] と CSJ の比較

表3に示した結果より、SSC [独話] と CSJ それぞれの調整頻度から作成したグラフを図3に示す。

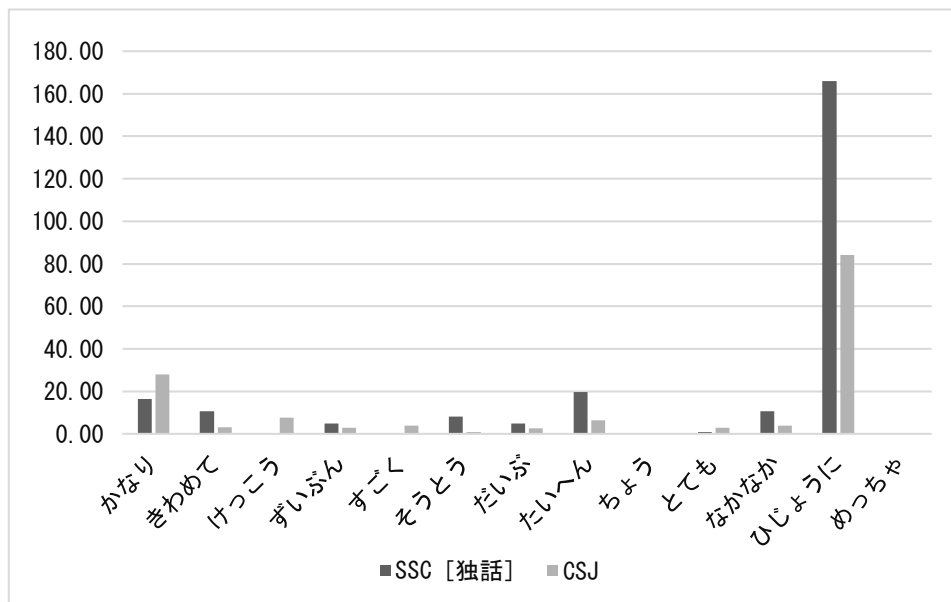


図3 SSC [独話]、CSJにおける副詞使用頻度 (調整頻度より)

SSC [独話] と CSJ とでは、「ひじょうに」が高頻度で使用されている点が共通している。ただし、CSJ での「ひじょうに」の使用頻度は SSC [独話] のおよそ半分と、「とても」類の語の使用全体が通時的に減少している。その原因には先述(4.1節)のように、講演と学会発表という録音された内容・場面の詳細な違いが要因のひとつとして挙げられる。また、全ての語が一樣に使用頻度を減少させているわけではなく、具体的には「かなり」、「けっこう」、「すごく」、「とても」は頻度上昇、「きわめて」、「ずいぶん」、「そうとう」、「だいぶ」、「たいへん」、「なかなか」、「ひじょうに」は頻度減少という傾向をみせる。使用頻度が上昇する語のうち「けっこう」、「すごく」は会話環境でも通時的に使用頻度が上昇している。これらの点については4.4節で総合的に考察を行う。

#### 4.3.2 SSC [会話] と CEJC の比較

表3に示した結果より、SSC [会話] と CEJC それぞれの調整頻度から作成したグラフを図4に示す。

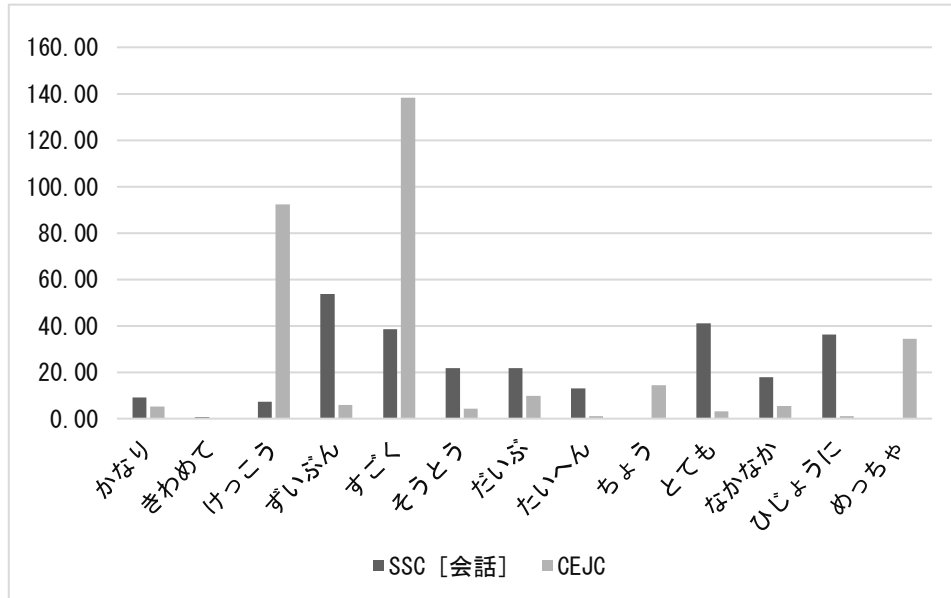


図4 SSC [会話]、CEJCにおける副詞使用頻度（調整頻度より）

SSC [会話] と CEJC の違いで顕著な点は通時的に「すごく」、「けっこう」への集中が進んでいることである。SSC [会話] では調整頻度が20を超える語には6語があるものの、最も高頻度の「ずいぶん」が使用頻度（100,000語あたり使用数）53.79回、次点で「とても」が41.08回、「すごく」が38.47回と、顕著に使用頻度が突出している語はない。それに対し、CEJCでは「すごく」が138.39回、「けっこう」が92.34回、3番目に多い「めっちゃ」が34.31回と、使用の多い語の使用が顕著に多い。CEJCにおいて使用頻度が20を超える語も、同3語のみとなっている。また、通時的に使用頻度が上昇したのも「けっこう」、「すごく」、「めっちゃ」に「ちょう」を加えた4語のみで、残りの8語は頻度が低下している。

SSC [会話] で使用頻度が最も高い「ずいぶん」は、CEJC では大きく頻度を下げる結果となった。同様の評価的なニュアンスを含む語として、CEJC では「けっこう」が高頻度で確認されたが、例(17)、(18)に挙げるように、両者を比較すると「ずいぶん」よりも「けっこう」の示す程度はやや小さく感じられる。

(17) ずいぶんいろいろあんのね (SSC\_C52\_11\_CT)

(18) (筆者注：大阪・堺に様々な施設がある話)

堺ってけっこういろいろあったな (CEJC\_T016\_004a)

ここからは、会話環境においては話し手の持つ感覚による評価は抑えられ、聞き手にとっても同意・共感しやすい程度限定表現が優先されるように変化が起こったこと、また、話し手の感覚による評価的な表現が用いられる際には、比較的程度性の低い、つまり、相手に共

感が得られにくい表現を避ける形で表現が選択されるようになったと考えることができる。

また、孫（2018）や原田（2022）に指摘されるフィラー的な用法について、SSC [会話]では被修飾語と離れて発話される例自体確認された数は少なく、その例も例（19）、（20）などのように、被修飾語との関係は例（19）は「すごい→困る」、例（20）「はずいぶん→変わる」と明らかである。

- (19) すごい陶器をさ入れるところがなくて困っちゃって (SSC\_C52\_19\_CT)  
 (20) だからはずいぶんそのジ時代のことを思うですと変わりました  
 (SSC\_C52\_18\_CT)

一方で、CEJCでは例（21）のように被修飾語との間に複数の語を挟む例の他に、例（22）のように、ひとつづきの発話で複数使用される例が見受けられた。また、例（23）は「すごい→ゆわれる（言われる）」と、程度性を含む語という、「すごく」の被修飾語となる基本的な条件から離れた例もある。また、例（24）の「めちゃ」のように、先行研究で指摘される「かなり」、「けっこう」、「すごく」以外の語でも同様に確認される。

- (21) すごいなんかドイツの洗濯機発展してるみたいな (CEJC\_K013\_011)  
 (22) かなりなんかニーどっかなイ富山とかわかんないなんかかなりだから  
 遠くで一人暮らししないといけないとかいろいろ言ってる  
 (CEJC\_T010\_013)  
 (23) でしかもさすごいさ六月の初めに締め切りだってゆわれて  
 (CEJC\_K001\_017)  
 (24) めちゃ朝日がこうふーって出てくるのが真あの真っ正面に見える  
 とこな (CEJC\_T019\_002)

孫（2018）、原田（2022）はこのような程度副詞の発展的な用法を語彙的意味が希薄化した、もしくは消失したものとして取り扱うが、これらの例を確認する限り、発話内容を強調して相手に伝えたいという話し手の意図は認めることができる。このような用法は、語に対する修飾という元の程度副詞としての文法的な要素が弱まり、発話全体に対して付される談話標識的な用法であると考えられる。

#### 4.4 総合考察

本節では、以上の結果・分析をまとめ、「とても」類の語の使用実態変化について総合的な考察を行う。

まず、独話環境では「ひじょうに」に使用が集中することがSSC [独話]、CSJに共通して確認された。ここからは、「ひじょうに」の独話環境、特に学術的な講演・発表の中で使用される語であるという特徴が認められる。さらに、「ひじょうに」はSSC [会話]では比較的あらたまった会話の中で一定程度の使用が確認されたものの、CEJCでの使用は減少していた。ここからは、「ひじょうに」の独話環境語としての性格が通時的に高まっていることが示唆される。

会話環境では使用される語に集中が起こっており、特に「すごく」、「けっこう」が高頻度で確認された。これらの語はSSC [独話] →CSJでも使用数が伸びているが、ここからは、

会話環境において高頻度で使用される中で、独話環境でも使用されるようになったという使用範囲の拡大が示唆される。この点は、独話環境語としての性格が強まった「ひじょうに」とは対称的である。第3章で確認したようにCSJとCEJCの間にも収録時期の差はあり、CEJCのデータはCSJデータの10～20年ほど後に収録が行われている。しかし、CSJとCEJCの結果の差から、日本語全体の中では会話環境での使用が独話環境に伝播したという可能性が高いと考えられる。

また、詳細な分析対象となった13語の中で、「とても」だけに、特徴的な使用傾向の変化が確認された。表3は、13語それぞれの独話・会話環境における使用頻度の通時的変化の傾向を「+」（上昇）または「-」（下降）で表したものである。「めっちゃ」は独話環境のSSC [独話]、CSJ いずれにおいても使用が確認されなかったので当該欄は空白とする。

表3 使用頻度の通時的変化傾向（独話・会話環境別）

語	独話環境	会話環境	語	独話環境	会話環境
かなり	—	—	たいへん	—	—
きわめて	—	—	ちょう	+	+
けっこう	+	+	とても	+	—
ずいぶん	+	+	なかなか	—	—
すごく	+	+	ひじょうに	—	—
そうとう	+	+	めっちゃ		—
だいぶ	+	+			

+：通時的に上昇 —：通時的に下降

以上からは、ほぼすべての語が通時的に独話・会話で同様の使用傾向を見せる、つまり、それぞれの語の全体的な使用傾向は通時的に使われなくなる、または、使われるようになる、とまとめることができる。しかし、「とても」だけは独話環境で使用頻度が上昇し、会話環境で下がるといったように、主な使用場面に変化が起きている。「とても」の使用頻度は独話ではCSJで上から5番目で、会話でもSSC [会話] で上から2番目といったように、少数の使用の中で見られた誤差の範疇に含められる差異ではなく、明確な変化である。第1章で述べたように、はじめはやはり言葉として発生した程度副詞「とても」が会話で使用される語として通時的に一般化し、さらに時を経て、今度は独話場面という改まった場面で使用される語に変化していると考えられる。今後、同様の変化が「すごく」など別の語にも予想されるが、その解明には今後数十年の継続的な調査研究が求められる。

## 5. まとめ

本研究は、『昭和話し言葉コーパス』(SSC)、『日本語話し言葉コーパス』(CSJ)、『日本語日常会話コーパス』(CEJC)の3コーパスを組み合わせた『通時話し言葉コーパス』によって、「とても」類の語19語の使用実態の横断的・縦断的調査を行った。以下に、本研究で設定した研究課題への回答をまとめる。

RQ1「日本語話し言葉における程度副詞の使用傾向は発話環境によって異なりがあるか」に対しては、「独話環境では『ひじょうに』、『かなり』の使用が、会話環境では『すごく』、『けっこう』の使用がそれぞれ比較的高頻度で確認された。意味的に『ひじょうに』は『す

ごく』と、『かなり』は『けっこう』との対応が推察される」、RQ2「日本語話し言葉における程度副詞の使用傾向の発話環境による異なりは、通時的に変化しているか」に対しては、「独話・会話の両環境で変化が確認された。独話環境の中では『ひじょうに』が最も高い頻度で使用されることは通時的に共通するが、頻度自体は通時的に下がる。他に使用頻度が上昇する語もあるが、会話環境での使用頻度上昇に伴って使用が上昇する語・会話環境では使用されなくなり独話環境で使用が増える語など、傾向は語によって異なる。会話環境では特に『けっこう』、『すごく』、『めっちゃ』が使用されるようになり、その用例の中には談話標識化という程度副詞用法の変化を示唆するものも確認された」とまとめられる。

本研究の成果は、『通時話し言葉コーパス』によって日本語の横断的な異なりとその縦断的な変化を同時に捉えることができたという点にある。今後は「とても」類の副詞や副詞以外の言語項目についての同様の調査研究、さらに、本研究で得られた結果と今後収集される未来の日本語データとの比較による継続的な調査分析もまた期待される。

### 謝 辞

本研究は、2021年度尚友倶楽部筑波大学日本語教育研究者育成奨学金およびJST 次世代研究者挑戦的研究プログラムJPMJSP2124の支援を受けたものです。

### 文 献

- 石川慎一郎（2020）「発話における副詞の使用」迫田久美子・石川慎一郎・李在鎬編『日本語学習者コーパス I-JAS 入門』くろしお出版, pp.167-184.
- 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.
- 国立国語研究所（2006）「日本語話し言葉コーパスの構築法」『国立国語研究所報告』124
- 孫琦（2018）「『すごいきれい』はほんとうに『すごい』のですか？」遠藤織枝（編）『今どきの日本語 変わることば・変わらないことば』ひつじ書房, pp.33-46.
- 中俣尚己（2016）「学習者と母語話者の使用語彙の違い—『日中 Skype 会話コーパス』を用いて—」『日本語／日本語教育研究』7, pp.21-34.
- 中俣尚己（2020）「主成分分析を用いた副詞の文体分析」『計量国語学会』32:7, pp.419-435.
- 鳴海伸一（2013）「副詞における程度的意味発生の過程の類型」『国立国語研究所論集』6, pp.93-110.
- 日本語教育学会編（2005）『新版日本語教育事典』大修館書店.
- 原田朋子（2022）「日本語の発話における副詞の意味・機能の弱まりに関する一考察：テキストマイニング手法と目視による分析を通して」『同志社大学日本語・日本文化研究』19, pp.1-28.
- 播磨桂子（1993）「『とても』『全然』などにみられる副詞の用法変遷の一類型」『語文研究』75, pp.11-22.
- 日暮康晴（2023）「日本語母語話者による程度強調副詞の使用実態—『日本語話し言葉コーパス』調査より—」『小出記念日本語教育学会論文集』31, pp.25-40.
- 飛田良文・浅田秀子（2019）『現代副詞用法辞典 新装版』東京堂出版.
- 丸山岳彦・小磯花絵・西川賢哉（2022）「『昭和話し言葉コーパス』の設計と構築」『国立国語研究所論集』22, pp.197-221.
- 吉井健（1993）「国語副詞の史的研究：『とても』の語史」『文林』27, pp.1-30.
- 渡辺実（1990）「程度副詞の体系」『上智大学国文学論集』23, pp.1-16.



URL

〈コーパス〉

中納言 – 国立国語研究所

<https://chunagon.ninjal.ac.jp/> (2023.07.04 最終閲覧)

『昭和話し言葉コーパス』(国立国語研究所)

中納言 2.7.0 データバージョン 2022.02 (2022.12.08 取得)

『日本語日常会話コーパス』(国立国語研究所)

中納言 2.7.0 データバージョン 2022.03 (2022.12.08 取得)

『日本語話し言葉コーパス』(国立国語研究所)

中納言 2.7.0 データバージョン 2018.01 (2022.12.08 取得)

〈コーパス語数データ〉

『昭和話し言葉コーパス』関連データ」(2022年7月20日取得)

(語数集計には「morph\_SUW」フォルダ内各短単位データ使用(筆者集計))

(<https://ssc-data.ninjal.ac.jp/course/view.php?id=3#section-0> よりダウンロード可能)

『日本語話し言葉コーパス』語数表(Version 201803)」(XLSX ファイル)

(2022年10月21日取得)

(<https://repository.ninjal.ac.jp/records/3278> よりダウンロード可能)

「CEJC 語数表」(7\_cejc\_suw\_wc.xlsx) (XLSX ファイル) (2023年4月21日取得)

(<https://www2.ninjal.ac.jp/conversation/cejc/cejc-wc.html> よりダウンロード可能)

〈その他コーパス情報〉

『独話』(50 ファイル) の一覧」(PDF ファイル) (2023年8月17日確認)

(<https://www2.ninjal.ac.jp/conversation/showaCorpus/> よりダウンロード可能)

# 歩きスマホのコロケーションと意味 —新聞データベースを用いた使用実態調査—

王 鑫 (筑波大学人文社会科学研究群)

## The Collocation and Meaning of “arukisumaho” A survey using newspaper database

WANG Xin (Degree Programs in Humanities and Social Sciences, University of Tsukuba)

### 要旨

「歩きスマホ」は典型的な「自動詞連用形＋モノ名詞」型複合名詞（例：空き巣）とは、以下のようなところで異なる。一つは複合名詞全体の意味がモノを表せないこと、もう一つは「歩き」と結合された「スマホ」が、本来、語用論的な意味にすぎない「スマホをすること」の読みにし解釈できないことである。本研究は、まず、新聞データベースを通して、「歩きスマホ」「寝タバコ」などの複合名詞のコロケーションと意味を調べ、その相違を明らかにする。次に、「歩きスマホ」を代表とした複合名詞は、主要部を持たない、並列の複合語としての特徴を持ち、ナガラ節に相当するような修飾関係が可能になっていると主張する。

### 1. はじめに

「動詞連用形＋モノ名詞」型複合名詞<sup>1</sup>のうち、「他動詞」とその「目的語」が結合して、行為を表せる用例が多数存在していることはよく知られている（買い物、洗い物など）。しかし、前項が自動詞の場合、(1b,c)のような行為を表す（またはそのような解釈が容認される）用例は非常に限定的である。加えて、そのほとんどが語彙的にリストされていて、項構造をもつものに比べて、規則性や生産性が見られない。そのため、先行研究を概観しても、「自VN」型複合名詞の行為性に関する研究はほとんど見られない。

「自VN」型複合名詞について、行為を表す、または行為の解釈が可能か否かという視点で観察すると、概ね以下のようなグループが確認できる。

- (1) a. 枯れ枝、流れ星、空き巣、浮き草、乾き物、曇り空、落ち葉、逃げ口、乗り場、跳び箱、寝汗、寝顔、泣き目、抜け殻、座り胼胝
- b. 寝正月、立ち湯、寄り道、寝湯、寝酒、回り道、出足、逃げ足、上がり湯、駆け足、隠れ身
- c. 歩きスマホ、歩きタバコ、寝タバコ、添い乳、迷い箸、移り箸、移り身

(1a)はモノのみを表す。その結合要素の前後関係を見ると、内項による複合と付加詞による複合に大別できる。複合名詞全体の意味は右側主名詞の表すモノの下位分類に相当し、そのモノの一つの種類または状態を指す。

<sup>1</sup> 以下は「VN」型複合名詞と略す。モノ名詞について、特に明示しない場合、具体的な物体と抽象的な概念両方を含む。行為と区別される。また、意図的に行為と区別したい場合、「モノ」の表記を採用し、そうでない場合は「もの」の表記を使用する。

(1b) は(1a)と同様に、主名詞の表すモノの下位分類に相当するような意味をもちながらも、行為としての解釈も容認される。例えば、「寝正月」は「寝て過ごすお正月」という「正月」の状態の解釈と、「寝て正月を過ごすこと」という「正月の過ごし方」の解釈、両方が可能である。「立ち湯」も同様に、「立ったまま入るお湯」というお湯の種類と「立ったままお湯に入ること」という入浴の仕方と、二つの解釈が成り立つ。また、「寄り道」のような行為の解釈が常に優勢にある用例でも、(2)のような抽象化されたモノとしての用法が見られる。

(2) その狙いを、店主は次の通り言います。「ネットは目的の本をすぐ買えるが、逆に思いがけず別の本に出会う寄り道がない。」

(OY14\_00990 9090 Yahoo!ブログ 2008)

一方で、(1c)は、モノとしての解釈は容認されない。また、構成要素のどちらが主要部なのか判断しにくい場合がある。例えば、「歩きスマホ」は「スマホ（スマートフォン）を使いながら歩くこと。特に駅の構内や屋外で歩行中にスマホを使うこと」の意味である（『デジタル大辞泉』より）。直感的にも、後の調査からもわかるように、モノとしての解釈はできない。また、語釈だけを見ると、「歩き」と「スマホ」のどちらも主要部のように見えるが、「スマホ」は「スマホを使うこと」を意味しているので、どちらも主要部ではないことを示唆している。したがって、(1c)は、「右側主要部規則」（Righthand Head Rule : Williams 1981）の例外として扱うべきであろう。

本稿は(1c)の用例を中心に、まず、コーパスと新聞データベースから、そのコロケーションと意味を確認する。次に、「自VN」型複合名詞の行為の解釈は語彙概念構造（以下、「LCS」とする）を用いて分析できると述べる。本稿の構成は以下の通りである。2節で先行研究を紹介する。3節では新聞データベースにおける使用例を確認する。4節は考察である。5節はまとめと課題である。

## 2. 先行研究

自動詞が前項にくる「VN」型複合名詞の動作性について論じたものは少ないため、本節では、まず、他動詞の用例を中心に、その動作性について体系的に論じた先行研究を整理する。

澤田（1999）は主に前項動詞が他動詞の場合、「ことの意味」を持つものと持たないものの対比から、「ことの意味」の見出しやすさを分析している。澤田(1999 : 158-160)では、「～もの」のフレームの前項に動詞がくる場合、「ことの意味」が見出しやすい動詞の特徴としては、①一回完結で短期間内に行われる動作②モノと密接に関係している、そのモノの存在に必要不可欠な動作③動作後、モノに影響が継続している動作④経験的に親密な動作であるとし、自動詞や他動詞受身のような反復的、継続的な動作は通常「こと的な意味」が見出しにくいとしている。

金（2016）は、「VN」型複合名詞は機能動詞「する」と共起できるものとできないものがあることに注目し、共起できるかどうかは「VN」型複合名詞の動作性によるものであるとしている。そして、動作性を持つ「VN」型複合名詞の特徴を「1.意味の転移が起きず、意味の結合が透明であること。2.語内部の構造が修飾関係のみならず、項関係（特に目述関係）にもあること」とまとめている（金 2016 : 126）。ここで注目したいのは、1.と2.は動作性を

持つ複合名詞の必要条件であって、十分条件ではない。したがって、1.と2.の条件を満たしても、動作性を持たない「VN」型複合名詞の反例はいくつか挙げられる。

劉 (2019) は、「VN」型複合名詞は名詞的な意味を表すのが一般的であるが、「買い物」のような名詞的な意味と動詞的な意味を併せ持つものが観察できるとした上で、動作性を持つ「VN」型複合名詞を動詞と名詞の間に格関係の有無によって二分した。そして、格関係のある「VN」型複合名詞は「1. 動詞と名詞が『他動詞—目的語』の関係であること。2. 動詞が『VN』型複合名詞の主体クオリアで、かつ『VN』には潜在的な目的クオリアが存在しないこと。3. 動詞が名詞を共有する動作性『VN』型複合名詞が既存であること。4. 『VN』型複合名詞の表す動作が想起されやすいこと。」が動作性の成立条件として挙げられている。また、格関係が存在しない「自VN」型複合名詞の動作性の成立条件について、「動詞以外に名詞と共起性の高い動詞が容易に推定できる」としている (劉 2019 : 93)。

そのほかに、「VN」型複合名詞に直接言及していないが、動詞由来複合語 (主として「名詞+動詞連用形」型複合名詞) の語形成を論じた伊藤・杉岡 (2002) の分析手法も大変参考になるため、ここで紹介しておく。伊藤・杉岡 (2002:112-125) は、内項を含む複合語は、結果産物を表すものを除くと、基本的に動作への名付け機能を有し、[-V] の品詞素性をもつ普通名詞であると述べている。また、付加詞を含む複合語の動作性は図1の示すように、LCS に含まれる異なる下位事象の基本述語が異なる意味の付加詞を選択した結果によるものであると考える。

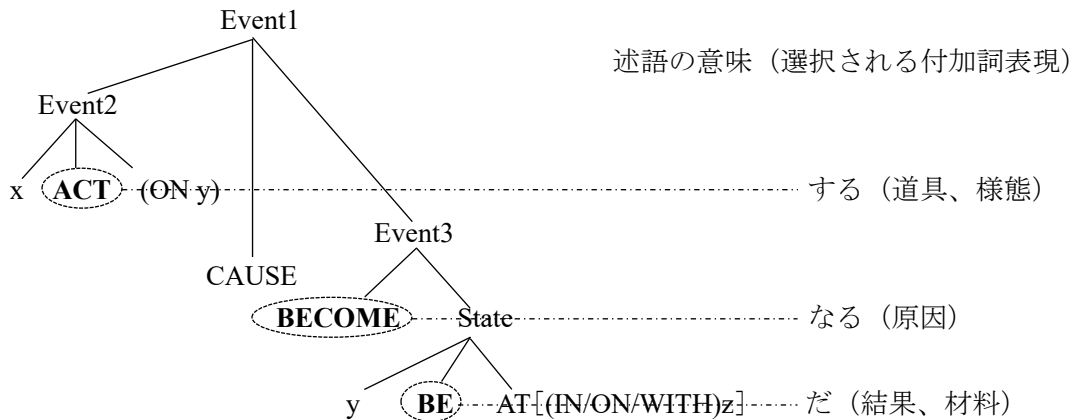


図1 〈達成動詞〉のLCS (伊藤・杉岡 2002:117 (94) による)

伊藤・杉岡 (2002:117-119) によれば、図1で示すように、Event2の基本述語は道具・様態の意味を表す付加詞を選択し、動作性をもつ(水洗い、一人歩きなど)。Event3のBECOMEを基本述語とした場合、能動性が低いので、動作よりも変化を表す(日焼け、ビール太りなど)。Event3のBEを基本述語とした場合、動作性を持たない結果状態を表すことになる(黒こげ、びしょ濡れなど)。

以上で見えてきたように、「VN」型複合名詞の動作性をめぐる研究は前項動詞が他動詞の場合に限定し、論じているものが非常に多かった。そこで取り上げられた用例は、行為とモノ両方の意味を併せ持つ場合がほとんどである。「VN」型複合名詞全体から見ると、むしろモノとしての意味が主要で、行為としての解釈は状況や文脈を考慮した副次的なものである (澤田 1999:161)。このような極めて一般化されてきた結論から、前項動詞が自動詞の「VN」

型複合名詞を見ると、行為しか表せない「歩きスマホ」のような類は、非常に異質的なものであると言えよう。このような行為に特化した「自 VN」型複合名詞が実際、どのように使われているのか、どのような特徴があるのかを明らかにする必要がある。

### 3. 新聞データベースにおける「歩きスマホ」の使用例

BCCWJ の収録年数の関係で、「スマホ」に関するデータが少ない。本研究は読売新聞データベース「ヨミダス歴史館」から「スマホ」に関連する用例を収集した。以下では「スマホ」と「歩きスマホ」がそれぞれ、新聞データベースにおいて、どのように使われているかを確認する。

#### 3.1 「スマホ」の使用例

「スマホ」の用例を 2022 年 10 月 1 日から 12 月 31 日までの 3 ヶ月の間、読売新聞全国版の新聞記事 218 篇から収集した。収集した用例の数は 243 例である。また、新聞のタイトルや見出し、および会話の直接引用や川柳などの特殊な文体における使用例を含まない。「スマホ」における前後の文脈をそれぞれ表 1 と表 2 にまとめた。

表 1 「スマホ」に前接する文脈<sup>2</sup>

前文脈	助詞	
φ (61)、文中 (読点で隔てる二つの文など) (32)、述語を含む連体修飾 (今使っている&) (12)、中古・格安 (7)、他	φ (117)	ス マ ホ
人 (対象: 子供に&を持たせる) (6)、時間 (試験中に&で撮影) (5)、場所 (会場には&の〜が並ぶ) (4)、モノ (対象: 風景に&を向ける) (2)、連用修飾 (3)	に (は) (20)	
時間 (近年は〜) (6)、人 (人相当: 配達員は〜、両社は〜) (5)、他の主題 (8)	は (19)	
モノの列挙 (パソコンや&) (16)、(後文脈に呼応し) コトの列挙 (〜の研究や&の開発をする) (1)	や (17)	
人 (所有者) (10)、手持ち (2)、〜年度 (1)、日本 (1)、連絡用 (1)、他	の (16)	
対象 (〜を&で撮影する、アプリを&にインストールする) (12)	を (14)	
人 (動作主: 人が&を操作する) (8)、モノ (メッセージが&に届いた) (2)、逆接を表す (3)	が (13)	
場所 (レジで&をかざす) (7)、道具 (日本語で&に話しかける) (2)、基準 (2)、原因 (料金の滞納で&が使えなくなる) (2)	で (13)	
	その他 (15)	

前文脈に出現する名詞は非常に豊富なため、意味の近いものをまとめて統計している。表 1 からわかるように、前文脈にはゼロ助詞 (「φ」) のものと何らかの助詞にスマホが後接するものがあり、それぞれ半分ずつを占めている。ゼロ助詞の場合、「スマホ」が文頭に立つものと、読点で隔てて文中に現れるものが最も多い。そのほか、「中古」や「格安」などの名詞に「スマホ」が後接し、複合名詞を構成するものや、「(今) 使っている」のような述語を含む連体修飾構造をなすものも少なくない。助詞を伴う場合、ほとんどの助詞が満遍なく

<sup>2</sup> 紙幅の関係で、「スマホ」を「&」で表す。〈〉内は用例の数。()内は文脈をわかりやすく示すために、意味役割や、出現環境などの具体例を明示したもの。

出現する。特に属格の「の」や、並列助詞の「や」、後文脈と合わせて「対象」として現れる「を」など（動画をスマホで撮る）から、「スマホ」は一般のモノ名詞として使われていることがわかる。

表2 「スマホ」に後接する文脈

	助詞	後文脈
ス マ ホ	を〈50〉	使う〈14〉、持つ〈11〉見る〈3〉、操作する〈2〉、手放す〈2〉、通じる〈2〉
	φ〈49〉	アプリ〈7〉、決済〈3〉、画面〈3〉、人（所有者、事業者など）〈3〉
	で〈40〉	撮影する（撮る）〈8〉、見る（視聴する）〈6〉、管理する〈3〉、利用する〈3〉
	の〈40〉	アプリ〈5〉、画面〈5〉、利用〈3〉、充電〈2〉、普及〈2〉、他
	に〈27〉	保存する〈3〉、搭載する〈3〉、届く〈2〉、向かう〈1〉、他
	や〈15〉	モノの列挙（パソコンなど）〈12〉、コトの列挙（やキャッシュレス決済の普及）〈1〉
	が〈7〉	身近にある〈2〉、使える〈2〉、手放す〈1〉
	その他〈15〉	

後文脈は、まず、助詞が「スマホ」に後接するものを確認する。助詞別で見ると、「対象」の「ヲ格」、道具の「デ格」、さまざまな意味関係を表す「ノ格」、抽象的な存在場所や動詞の行為の向かう先として現れる「ニ格」の順に並んでいる。後接する動詞も非常に豊かである。助詞を伴わない場合、名詞が直接「スマホ」に後接し、「名詞＋名詞」型複合名詞を構成し、さらに後文脈の述語の対象や道具などとして使われている。

以上のように、省略のない書き言葉的な文脈において、「スマホ」は普通のモノ名詞と同じ振る舞いを示すことがわかった。また、「スマホ」における行為の解釈が可能な環境については後述する。3.2節では、「歩きスマホ」について確認する。

### 3.2 「歩きスマホ」の使用例

「ヨミダス歴史館」から「歩きスマホ」の使用例138例<sup>3</sup>を収集し、その前後の文脈をそれぞれ表3と表4にまとめた。

表3 「歩きスマホ」に前接する文脈<sup>4</sup>

前文脈	助詞	歩 き ス マ ホ
φ〈50〉、#の定義〈18〉、文中（読点で隔てるなど）〈17〉、危険な〈1〉	φ〈86〉	
動作の場所〈16〉、道具（～の放送で#への注意喚起）〈2〉、最近〈1〉	で（は）〈19〉	
主題〈6〉、形式名詞と合わせて主題を表す（気になったのは～）〈3〉	（の）は〈9〉	
人（動作主）〈6〉、形式名詞と合わせて主語を表す（痛感したのが～）〈1〉	が〈7〉	
コトの列挙（脇見運転や#）〈5〉、動作主の列挙（盗撮犯や#をする人）〈1〉	や〈6〉	
対象（3人に#させる、謎に#を使った実験で迫る）〈3〉、場所〈2〉、時間〈1〉	に（は）〈6〉	
人（動作主）〈1〉、多数（多数の#族が出現）〈1〉	の〈2〉	

<sup>3</sup> 最初の使用例（2013年5月29日）から2023年5月31日までの読売新聞全国版の213の記事から用例を収集した。収集方針は「スマホ」と同様であり、会話の直接引用や川柳などの特殊な文体における使用例は収集しない。

<sup>4</sup> 紙幅の関係で、「歩きスマホ」を「#」で表す。

振動（から#を感知）〈1〉、人（男の人から#を咎められた）〈1〉	から〈2〉	
歩きスマホの定義（スマホを操作しながら歩くことを歩きスマホという）〈1〉	を〈1〉	

前文脈について確認すると、「歩きスマホ」が文頭に立つものや、定義の後に、鍵括弧付きで「歩きスマホ」が後接するものなど助詞を伴わないものが多い。注意すべき点としては「歩きスマホ」が被修飾語になる場合、(3)のように、「危険な」が修飾しているのは「キャラクターを探す」行為であり、その行為が「歩きスマホ」として命名されている。「危険な山道」のような「山道」の属性や特徴などを描くものではない。(4) (5)における「多数の」や「見学者の」も同様である。助詞を伴う場合、動作が行われる場所を表す「デ格」が目立つ。

- (3) 「ポケモン GO」は、スマホを持って歩くと画面にキャラクターが現れ、それを捕まえて集めるゲーム。まるで現実の場所にキャラクターが出てきたように見える。最大の問題の一つは、キャラクターをいち早く見つけようとして危険な「歩きスマホ」になることだ。  
 (2016.8.10「ポケモン GO 歩きスマホの危険 中学校長が楽しみ方「指南書」東京朝刊)
- (4) 男性が犠牲になった近鉄の事故現場もそうですが、ホームの幅が、わずか数メートルの駅はあちこちにあります。音が反響して立ち位置がわからなくなる駅、多数の「歩きスマホ族」が出現する駅……。  
 (2016.10.23「[読者と記者の日曜便]視覚障害者に声かけを」大阪朝刊)
- (5) 衆院は2014年から、国会見学者の「歩きスマホ」を禁止している。ゲーム配信後、注意喚起のための掲示板を国会内の6か所に設置した。  
 (2016.7.29「国会に「ポケモン GO」スポット 衆院が削除要請検討」東京朝刊)

表4 「歩きスマホ」に後接する文脈

	助詞	後文脈
歩 き ス マ ホ	を〈41〉	する〈21〉、禁じる〈7〉、やめる〈4〉注意する〈2〉、使う〈1〉
	の〈24〉	動作主〈12〉、危険(性)〈4〉、自粛〈2〉、怖さ〈2〉、規制〈1〉
	による〈17〉	事故〈15〉、事件〈1〉、(＃の)危険性〈1〉
	φ〈15〉	中〈3〉、だ(危険なのは「だ」)〈3〉、コトの列挙(「＃」「ながらスマホ」は事故を起こす)〈2〉、増加〈1〉、防止〈1〉、族〈1〉
	は〈10〉	危険だ〈1〉、社会問題になっている〈1〉、厳禁だ〈1〉
	に〈6〉	なる(ならない、なりがち)〈3〉、対処する〈1〉、危険を感じる〈1〉
	が〈6〉	問題となっている〈3〉、目立つ〈1〉、原因〈1〉、人の通行に影響〈1〉
	で〈5〉	(人にぶつかって)トラブル(事件)になる〈4〉、危険を感じる〈1〉
	その他〈14〉	

後文脈について、助詞が後接するものは、上位から対象を表す「ヲ格」、様々な意味関係を表す「ノ格」、原因を表す「ニヨル格」の順に並んでいる。ほかに、動作主または動作主に昇格された組織に「ガ」や「ハ」が後接して、その後に「歩きスマホ」が続くものもある。助詞を伴わない用例は、動作の最中を表す「中」や、「歩きスマホ」と並列関係にある行為

の列挙や、「増加」「防止」など変化または働きかけの対象を表すものがある。

表4から次のような点が注目に値する。まず、「スマホ」に比べて、道具の「デ格」が「歩きスマホ」に現れない。「歩きスマホ」における「デ格」は「ニヨル格」と同様に、原因を表している。次に、「ヲ格」に後接する動詞は非常にシンプルで、大きく分けて「する」と「禁じる」の2種類である。「スマホ」に後接する動詞の中でも見られた「使う」は(6)が示すように、「歩きスマホ」を一種の方法・手段として使われている。他の助詞に付く名詞も非常にシンプルで、「危険だ」「問題だ」「事故だ」のような用例が中心である。さらに、「歩きスマホ」に後接する「ノ格」は、ものの所有者という意味ではなく、動作主を表している。最後に、「歩きスマホ」に後接する「中」は、モノ名詞に付く空間的な存在を表す「中」ではなく、動作の最中を示すアスペクト的な意味である。

- (6) 人があふれるスクランブル交差点。かけ声や合図もないのに、歩行者はなぜかぶつからない。その謎に歩きスマホを使った群集実験で迫り、「パロディ版ノーベル賞」のイグ・ノーベル賞が贈られた。

(2021.10.17「顔 (Sunday) 歩行者の「阿吽の呼吸」 解明」東京朝刊)

以上の分析から、「歩きスマホ」は「スマホ」の一種、または「スマホ」の一つの状態としての解釈はできず、もっぱら行為を表すことがわかる。次の3.3節では「スマホ」と「歩きスマホ」の違いを具体的にみる。

### 3.3 「スマホ」と「歩きスマホ」の比較

本節では、先行研究でいう動名詞の判断で用いられたテストを援用し、「スマホ」と「歩きスマホ」の違いについて確認する。なお、容認差の判断は省略のない書き言葉に限る。

表5 「スマホ」と「歩きスマホ」の動作性の比較

テスト		スマホ	歩きスマホ
する		0	0
を		0	24
アスペクト的表現	～中	0	3
	～を始める	0	0
	～時(間)の～	0	1
「の」における動作主の解釈		— (属格のみ)	+

上記のテスト以外に、類別詞や表1で示した「アプリをスマホにインストールする」のような存在文の解釈が可能かどうかなどのテストもある。これらのテストと3.1節で確認した、3.2節で観察した、「スマホ」と「歩きスマホ」に前接する修飾成分の違いや後接する成分の多様性における差異などからも、書き言葉における「スマホ」はモノ、「歩きスマホ」は行為のみを表していることが裏付けられている。それでは、「スマホ」における「行為」の読みはどのような文脈において可能なのかについて考える。実際の会話の中では、発話環境、ジェスチャーなど、音声以外の要素がコミュニケーションに深く関与している。また、話し言葉においては、省略や文法の乱れが容認されるので、「スマホ」が「スマホを操作す



ること」まで意味が拡張していることは珍しいことではない。一方で、書き言葉においては、一定の条件が整っている場合にのみ、「スマホ」における行為の解釈も可能である。例えば、以下のような例である。

- (7) 運転中のスマホは大変危険ですので、おやめください。
- (8) 図書館内でのスマホはご遠慮ください。
- (9) 歩きながらのおタバコはご遠慮ください。
- (10) 火災報知器が作動しますので、トイレでのおタバコはご遠慮ください。

(7)～(10)の例は誰もが目にしたことのある「注意書き」のようなものである。このような文脈においては、行為の読みが成立する「前提」がきちんと明示されているという共通点がある。例えば、動作が行われた場所を表す「デ格」や、付帯状況を表す「ナガラ」、動作の最中を表す「～中」などが用いられることによって、「スマホ」や「タバコ」のモノとしての解釈がブロックされていて、行為の解釈のみが成立している。さらに、行為の内実まで規定される場合もある（遠慮する内容はスマホを操作することなのか、スマホを持ち込むことなのか文脈によって読み取れる場合がある）。つまり、「スマホ」におけるモノの解釈は文脈に関わらず成立するのに対して、行為としての解釈はかなり文脈（発話環境などを含む）に依存していることが言える。

同じ傾向は新聞データベースにおける使用からも確認できる。「歩きスマホ」が登場する背後には、「歩きタバコ」や「寝タバコ」のような既存語彙による類推の可能性は否定できない。「寝タバコ」が一語として最初に現れたのは1950年代<sup>5</sup>のことである。「歩きタバコ」はそれより後の1970年代<sup>6</sup>のことであって、いずれも新聞の見出しとして使われている。日常会話の中では、新聞記事より前にも、「タバコ」が「タバコを吸うこと」まで意味の拡張があったと推測できる。しかし、書き言葉においては、上述の新聞記事より前の使用例を確認すると、やはり「煙草くわえて眠って大火傷」（読売新聞 1951.3.7 東京夕刊『煙草くわえて眠って大火傷』）「歩きながらのたばこを全部やめなさいとは言いません」（読売新聞 1966.12.13『赤でんわ 愛煙家の男性方へ』）などのような文脈で使用されていた。つまり、行為であることは文脈によって明示されている。また、同時期に「歩行喫煙」「歩く喫煙」などの類義語との混用も見られ、「寝タバコ」も「歩きタバコ」も語彙として定着していないことがわかる。その後、語彙化された「寝タバコ」「歩きタバコ」が上記のような冗長な表現に取って代わり、やがて一般化された語彙として定着したと思われる。言い換えれば、文字制限のある新聞の見出しなどにおける「寝タバコ」「歩きスマホ」の登場は、話し言葉に近い場面に限られた「タバコ」「スマホ」の行為としての用法を加速させ、やがて一般化したと考える。しかし、「タバコ」「スマホ」の行為としての解釈が一般化されたとしても、モノとしての意味は依然強いので、書き言葉においては、無標のモノとしての解釈に比べて、有標の行為の解釈は文脈に強く依存していることは今でも変わらない。

<sup>5</sup> 出典は『燃えるまで知らず、寝タバコから大火 速報大切三分たてば一軒焼く』（読売新聞 1955.12.6. 朝刊）。『火災の注意 寝タバコが危い 火を扱う場所と燃料は離す』（朝日新聞 1957.1.11. 東京朝刊）である。尚、新聞記事の表記そのまま採用する。

<sup>6</sup> 出典は『歩きたばこ 老紳士ひとり新宿抵抗戦』（毎日新聞 1972.5.15 東京朝刊）である。

4. 考察

4.1 LCS を用いた分析

3節で見たように、「スマホ」が単独で使われる場合はモノのみを指し、行為を表すことはできない。一方で、「歩きスマホ」における「スマホ」はモノではなく「スマホを操作すること」を意味している。したがって、「歩きスマホ」における「スマホ」は動詞として扱ったほうが妥当である。

「歩きスマホ」における「スマホ」を動詞とみなした場合、「歩きスマホ」の意味と動作性は伊藤・杉岡（2002：117（94））のLCSを用いた分析を援用することが可能である。つまり、「歩く」ことは「スマホを操作する」ことがどのような状況のもとで行われたか、その様態に対する説明である。

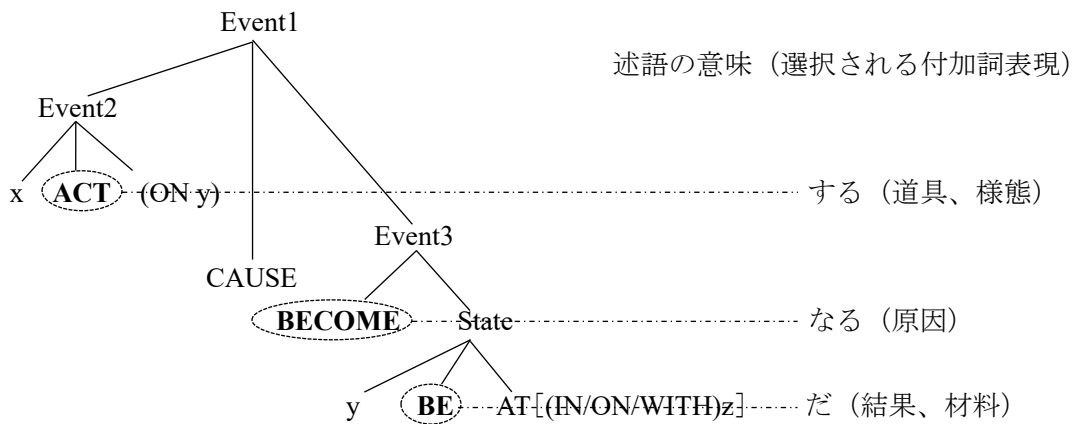


図2 〈達成動詞〉のLCS (図1再掲)

「歩きスマホ」における「スマホ」を動詞として捉えた場合、スマホに変化をもたらさない活動動詞として見るべきである。活動動詞はEvent2のみを有しているので、「歩き」は様態を表す付加詞としてACTに選択され、「歩きスマホ」全体が「する」の意味を持つ動作性名詞になる。「歩きタバコ」も「寝タバコ」も以下(11)の他の用例も同様な分析が可能である。「迷い箸」「移り箸」はそれぞれ、「食事のとき、迷ってあれこれと菜に箸を向けること」「食事の時、箸を飯、または、手許に戻さずに菜から菜へと移して食べること。一菜を食べ終わらずつぎつぎに別の菜に箸をつけること」の意味である(『日本国語大辞典』より)。「迷い」や「移り」は無意識の動作である可能性は否定できないが、それぞれ箸を使うときの様態を表し、その動作への名付けであることははっきりしている。

「添い乳」について、「乳」が「授乳すること」まで意味が拡張しているとは断定できない。しかし、歴史的な変遷を見ると、「添い乳」より以前は「添え乳」のような用例が確認できる。つまり、同じ意味の複合名詞は最初に他動詞と「乳」が複合し、後に自動詞との複合も現れ、「添い乳」として語彙化された。また、今でも「添い乳」は『日本国語大辞典』に立項しておらず、『デジタル大辞泉』などの辞書を引いても、「添え乳」に転送されることがある。このような語の成り立ちから見ると、「添い乳」は特殊な例として扱うべきであろう。いずれにせよ、(11)の用例をまとめて見るとき、後項名詞が行為として意味解釈されることによって、複合名詞全体が行為として解釈されることが可能になる。

(11) 歩きスマホ、歩きタバコ、寝タバコ、添い乳、迷い箸、移り箸、移り身

((1c) の再掲)

続いて(12)を見る。(12)に挙げた「寝正月」や「立ち湯」などの用例は、モノとしての意味が強く維持されながら、行為の解釈も完全にはブロックされていない。このような用例における行為の解釈も「歩きスマホ」と同様な分析が可能である。特別な文脈なしに、「正月」と「お湯」だけで、「お正月を過ごすこと」「お湯に入ること」を表すことは難しいであろう。しかし、「寝正月」「立ち湯」などの複合名詞を構成した場合、後項名詞は行為として解釈されなければならない。前項動詞との意味関係は、a.前項動詞が一種の付帯状況として、後項名詞の動作の様態を表す。b.前項動詞と後項名詞と一緒に、行為の連鎖を作るなどである。それぞれの複合名詞の間に、付帯状況を表す「トキ節」や「ナガラ節」または継起の「テ」を挿入することによって確認できる。

一方、「出足」「逃げ足」「移り身」は「出足が鈍い」「逃げ足が速い」「移り身が速い」のような慣用表現の用法にしか見られない。また、「足」も身体部位としての「足」ではなく、ものの動く様子を「足」に見立てて表現しているため、名詞というよりも、接尾辞である。したがって、例外として扱うべきであろう。

「寄り道」は「通行のついでに立ち寄ること。また、回り道して別の所へ立ち寄ること。また、その道」の意味である(『日本国語大辞典』より)。「寄り道」は「回り道」「迂回路」のように、目にみえる形の「道路」の意味で用いられにくい。ほとんどの場合は「寄り道になる」のような文脈で、実際に立ち寄る、または立ち寄ることを想像して、立ち寄る途中、立ち寄った結果、「寄り道」が生じたわけである。言わば一種の産物に近いものである。

(12) 寝正月、立ち湯、寄り道、寝湯、寝酒、回り道、出足、逃げ足、上がり湯、駆け足、隠れ身

((1b) の再掲)

#### 4.2 行為を表せる「自VN」型複合名詞の位置付け

行為を表せる「自VN」型複合名詞は、モノの意味が中心で、文脈が整えば、行為の解釈も容認されるものと、「歩きスマホ」のような行為に特化したものに分けられる。また、「歩きスマホ」のグループを除くと、「自VN」型複合名詞における行為の解釈が可能な用例は、モノとしての解釈は依然と強い。言い換えれば、モノとしての解釈と行為としての解釈が併存している場合が多く、排他的なものではない。「他VN」型複合名詞の行為類(買い物、洗い物)に比べて、ほとんど生産性が見られない。複数の辞書およびコーパスから「自VN」型複合名詞を網羅的に収集し、調査した王(2022)によると、「自VN」型複合名詞はもっぱらモノを表し(枯れ枝、流れ星など)、行為を表す「自VN」型複合名詞は後項名詞が動作性名詞(立ち小便、寝化粧)を除くと、ほんの数例にとどまる。

後項名詞がモノ名詞かつ行為を表せる「自VN」型複合名詞は言うまでもなく、前項動詞が非能格動詞あるいは意図的な動作として解釈可能な動詞でなければならない。また、後項名詞は人工物、あるいは身体部位を表す名詞が多い。後項名詞の意味について、クオリア構造を用いて記述することが有効であると数多くの先行研究が指摘している(Pustejovsky: 1995、小野: 2005、影山: 2011など)。また、名詞のうち、特に人工物について、その目的クオリアが最も重要である(影山 2011: 74-79)。つまり、人工物は何らかの目的で作られ、

その目的クオリアを引き立てる動作へ意味拡張しやすいのである。実際に各用例における行為の解釈もその目的クオリアに沿った行為であることが観察できる（「お湯」→お湯に入る、「タバコ」→タバコを吸うなど）。以下 (13) では、「スマホ」のクオリア構造を用いて、「スマホ」における行為の解釈は全てその目的クオリアに関連していることが確認できる。

(13) スマホ

形式クオリア：電気製品

構成クオリア：画面、金属のカバー、キーボード、内蔵マイク、カメラ、バッテリー

目的クオリア：通話、ゲーム、SNS、ビデオ・音楽の視聴

主体クオリア：設計、生産ラインにて組み立て、製造される

「歩きスマホ」のような結合パターンは劉 (2019:92) では、辞書には立項されていないことと BCCWJ における使用例が少ないことから、新しい造語法として今後定着していくと述べている。しかし、1950 年代以降に現れた「寝タバコ」「歩きタバコ」などを考えると、このような語構造をとる用例の数は確かに少ないが、造語法としては決して新しいものではないことがわかる。「歩きスマホ」のような付帯状況にある二つの行為を結合させて「自 VN」型複合名詞を構成するパターンは一般の項構造をもつ複合名詞より、意味の推測が難しいことは事実である。また、日本語においては、モノ名詞は接辞なしで品詞を変えたりすることはあまり一般的ではない（例えば「スマホ」が形を変えずに、「スマホを操作する」を表す動詞として品詞が変化することはない）。そのため、用例が非常に限定されている。しかし、ネット社会の普及に伴い、言葉が急激に変化している。一つのパターンが成立すると、新語が勢いよく生産され、伝達され、そしてその多くは消滅していくのであろう。例えば、以下のような例は Twitter で検索して出てきたものである。これらの例は明らかに「寝タバコ」や「歩きスマホ」による類推と思われるが、その行方について引き続き注視していきたい。

(14) 寝ラジオ、寝スマホ、寝テレビ、走りスマホ、座りタバコ、踊りタバコ、走りタバコ、踊りスマホ、泳ぎスマホ

5. まとめと今後の課題

本稿は新聞データベースにおける「歩きスマホ」の使用状況を通して、今まで注目されてこなかった行為の解釈が可能な「自 VN」型複合名詞の特徴を確認した。その上で、このような複合名詞は LCS を用いた分析が可能であることを見た。行為を表せる「自 VN」型複合名詞の中には、さらに行為に特化したグループとモノの解釈が優勢であるグループが存在する。モノの解釈が優勢なグループでは、行為の解釈は一つの可能性としては否定できないが、その解釈が許される文脈は限定的であり、語用論的な意味にすぎない。一方、「歩きスマホ」のような行為に特化したグループは、後項名詞の事物の一種または一つの状態としての解釈はできないため、主要部を持たない複合名詞であると言える。しかし、その深層構造から考えると、二つの行為による複合であることが言える。それに、前後の動作は複合名詞の動作性を担保する上では不可欠であって、「ナガラ節」に相当するような修飾関係が可能になっているため、並列関係にある複合名詞と見たほうが妥当であろう。

従来、LCS を用いた研究は項構造では処理できない語彙的にリストされた「NV」型複合

名詞に限って論じたものが多い。本研究は、今まで例外として扱われた行為を表せる「自VN」型複合名詞の分析においてもLCSを用いることが可能であると示した。さらに、項関係をもたない「他VN」型複合名詞の行為類（「追い腹」→「追って腹を切る」、「追い銭」→「追って（不足した）お金を払う」）にも用いられると予測できるが、これらへの検討や記述は今後の課題にしたい。

#### 参考文献

- 石井正彦（2007）『現代日本語の複合語形成論』ひつじ書房
- 伊藤たかね・杉岡洋子（2002）『語の仕組みと語形成』研究社
- 王鑫（2022）「複合名詞の語構成—「自動詞連用形＋名詞」型複合名詞を例に—」『筑波日本語研究』26,pp.20-36,筑波大学日本語学研究室
- 奥津敬一郎（1975）「複合名詞の生成文法」『国語学』101,pp.48-37,武蔵野書院
- 小野尚之（2005）『生成語彙意味論』くろしお出版
- 影山太郎（1993）『文法と語形成』ひつじ書房
- 影山太郎（1999）『形態論と意味』くろしお出版
- 影山太郎（2011）『名詞の意味と構文』大修館書店
- 村木新次郎(1985)「慣用句・機能動詞結合・自由な語結合」『日本語学』4(1),pp.15-27,明治書院
- 金恵珍（2016）「日本語『V+N』型複合名詞の動作性に関する研究」『言語文化』34,pp.111-126,韓国日本言語文化学会
- 斎藤倫明（1992）『現代日本語の語構成論的研究—語における形と意味—』ひつじ書房
- 澤田浩子(1999)「現代日本語『—もの』の複合名詞をめぐる—モノとコトの認知の世界—」『KLS19』 pp.153-163,関西言語学会
- 杉岡洋子（2020）「「動詞連用形＋名詞」複合語の多義について」『名詞をめぐる諸問題 語形成・意味・構文』 pp.2-23,開拓社
- 野田大志（2011）「「他動詞連用形＋具体名詞」型複合名詞の意味形成」『日本語の研究』7-2,pp.1-8,日本語学会
- 由本陽子（2016）「日本語複合名詞の意味解釈メカニズム」『言語文化共同研究プロジェクト』2015,pp.79-88,大阪大学大学院言語文化研究科
- 劉犀灵（2019）「動作性を持つ日本語 [V+N] 型複合名詞の成立条件」『一橋日本語教育研究』7,pp.83-94,一橋日本語教育研究会
- Pustejovsky, James.(1995)*The generative lexicon*. MIT Press
- Williams, Edwin(1981) On the Notions "Lexically Related" and "Head of a Word", *Linguistic Inquiry* 12,pp.245-274.

#### 辞書類

- 大辞泉編集部（編），Ver.202104，『デジタル大辞泉』小学館
- 日本国語大辞典第二版編集委員会ほか（編），2000，『日本国語大辞典 第二版』小学館

コーパスと新聞データベース

国立国語研究所, 『現代日本語書き言葉均衡コーパス』(中納言 2.7.2) Ver.2021.03,

<https://chunagon.ninjal.ac.jp/bccwj-nt/search>

国立国語研究所, 『国語研日本語ウェブコーパス NWJC』(中納言 2.7.1) Ver.2021.03,

<https://chunagon.ninjal.ac.jp/nwjc/search>

筑波大学・国立国語研究所・Lago 言語研究所, 『筑波ウェブコーパス』 NLTver.1.40,

『NINJAL-LWP for TWC』, <http://nlt.tsukuba.lagoinst.info>

読売新聞データベース 『ヨミダス歴史館』

<https://database.yomiuri.co.jp/about/rekishikan/>

朝日新聞データベース 『朝日新聞クロスサーチ』

<https://xsearch.asahi.com>

毎日新聞社データベース 『毎索』

<https://mainichi.jp/contents/edu/maisaku/>

## 国語科教科書の説明的文章における疑問-解答の段落構成 —各学年テキストの比較から—

尾崎 誉治（東北大学文学研究科）

### Paragraph Structure by Question and Answer in Expository Texts in Multi-Grade Japanese Language Texts

Takaharu Ozaki (Graduate School of Arts and Letters, Tohoku University)

#### 要旨

本研究では、小学校・中学校のテキストの「問題を提示」「結論を述べる」部分を、それぞれ疑いの疑問文とその疑問内容に答える文によって捉え、疑問文、解答文とそれらが含まれる段落について、数、大きさ、疑問-解答間の距離の観点から10の変数を定めて主成分分析を行った。その結果として次の傾向が指摘できた。①第1学年では、疑問-解答を段落として卓立させつつ、ペアとして示す構成が重用される。②第2・第3学年では大きな疑問と解答のみのシンプルな型による構成が増えてくる。③第4・第5学年では、次第に大きくなる文章の中でも、疑問-解答の段落を一文段落などで卓立させながら示している。④小学校第6学年以降は疑問-解答の構成が多用されなくなり、説明的文章内の多様なジャンルにより文章構成に幅が生じる。⑤中学校のテキストでは、さらに後に解答される疑問だけでなく、提示されるだけのものや、明確な解答の見つからない疑問の使用が増える。

#### 1. はじめに

段落とは、何らかの内容的なまとまりを持ち、原則として改行1字下げによって示される文章中の構成要素である。段落の学習は、平成29年告示の小学校学習指導要領では全技能共通の「知識及び技能」、「書くこと」、「読むこと」の中に位置づけられており、第3学年及び第4学年の内容の中で「段落には問題を提示したり、具体例を示したり、理由を述べたり、結論を述べたりするなどの役割がある。」(p.83)と説明される「段落の役割」、「(これらの役割に基づく)段落相互の関係を理解することで、内容を把握したり必要な情報を的確に見付けたりすることができる。」(同)と説明される「段落相互の関係」の学習は、中学校学習指導要領では「書くこと」の中で引き継がれ、繰り返し言及されている。

児童・生徒は、学習の中で読む文章を通して、段落についての認識を培っていく。個々の教材が持つ特性は様々であると考えられるが、それらにおいて段落はどのように用いられているのだろうか。また、学年の段階に合わせつつ、どのように配列されているのだろうか。全文にわたって段落が一文で構成されているような小学校1年生の文章を始めとし、多学年の段落の性質を俯瞰的に確かめていくことは、段落に託される意味的なまとまりの段階性を考えることにつながり、段落それ自体の研究にも知見を与え得る。

そこで、本研究では、特に段落の学習に用いられる教材の一つである教科書の説明的文章において、学習指導要領解説で「段落の役割」として例示されていたもののうち、説明的文章の性質上、骨子となる要素であると考えられる「問題の提示」(疑問)や「結論を述べる」(解答)といった働きをする部分が、どのように段落を成しつつ、学年ごとに文章構造を構成しているかを分析する。

## 2. 対象と方法

小学校・中学校国語科教科書に収録されている小学校第1学年～中学校第3学年までの文章183編を対象に、疑問文および疑問内容に対応する答えを持つ解答文を頼りとして、問題の提示される段落、結論が述べられる段落を取り出し、それらの頻度、大きさ、置かれる距離などの特徴を変数として主成分分析を行う。これにより、各学年の説明的文章がどのような構造をもっているかを俯瞰するとともに、学年によってどのような性格が見られるかを分析する。以下、対象とする要素、テキストおよび分析方法について説明する。

### 2.1 対象テキスト

対象とするテキストは、小学校・中学校の国語科教科書の説明的文章としたい。大熊(2018:88)では説明的文章は次のように説明されている。

「説明的文章」とは、ある物事(知識・情報など)を、それについて知らない人に要点を整理してよく分かるように説明する文章の事である。

「説明的文章」の種類は、広義には、新聞、辞典・事典、図鑑などの「解説」、入門書、手引き、製品、器具の取扱い説明書、薬品や食品の効能書きなどと幅が広い。一方、狭義には、「説明文」「意見文」「観察記録文」「論説文」などをいう。

国語科教科書における説明的文章は広義の意味で示されるものであると思われるが、その大部分は狭義の説明的文章の下位区分として挙げられるタイプの文章のようである。狭義に捉える場合に筆頭に挙げられている「説明文」は、「基本的に問いに答えるという構造を持つ」(櫻本2009:249)などと説明されるように、読み手にとって不明な内容を、説明されるべき問題として提示したうえで解答を示していくタイプの文章である。

本研究で説明的文章を対象とするのは、国語教科書において段落の指導の中心となっている教材が説明的文章であること、文学作品などの非説明的文章と比べて論理的な構造を取り出しやすいと想像されることによる。

本研究で対象とする説明的文章は、5社分の小学校・中学校国語科教科書<sup>1</sup>に収録されている小学校第1学年～中学校第3学年までの文章183編とした。分析にあたって、これらの教科書から光学文字認識とその出力データの目視での修正といった処理を経て電子データを作成し、文単位で分析に必要な情報を施して集計した<sup>2</sup>。

各学年における資料の基本的な情報を以下の表1に示す。

<sup>1</sup> 小学校教科書は『小学校国語』光村図書、『新編新しい国語』東京書籍、『小学生の国語』三省堂、『広がる言葉小学国語』教育出版、『みんなと学ぶ小学国語』学校図書(すべて平成26年検定版)、中学校教科書は『国語』光村図書、『新しい国語』東京書籍、『現代の国語』三省堂、『伝え合う言葉中学国語』教育出版、『中学国語』学校図書(すべて平成27年検定)を用いた。対象とした説明的文章のタイトルについては、本研究末尾で資料として挙げる。

<sup>2</sup> 本研究執筆時点では、段落などの文章レベルでの書式情報が整備された電子データはまだ少ないように思われる。そのため、本研究では、今後も継続的にデータを拡充させていくことも視野に入れ、私家データを構築・使用している。



表1 各学年資料の基本情報

項目	小学校						小計	中学校			小計	合計
	1年	2年	3年	4年	5年	6年		1年	2年	3年		
資料数	22	21	20	19	19	15	116	22	22	23	67	183
総段落数	239	246	259	254	283	232	1513	286	377	372	1035	2548
	10.86	11.71	12.95	13.37	14.89	15.47	13.04	13.00	17.14	16.17	15.45	13.92
(一文段落数)	179	93	78	52	52	45	499	46	50	47	143	642
	8.14	4.43	3.90	2.74	2.74	3.00	4.30	2.09	2.27	2.04	2.13	3.51
文数	319	533	680	751	873	762	3918	1012	1515	1376	3903	7821
	14.50	25.38	34.00	39.53	45.95	50.80	33.78	46.00	68.86	59.83	58.25	42.74
文字数	8529	16792	24802	30144	36108	34336	150711	43458	59043	61031	163532	314243
	387.68	799.62	1240.10	1586.53	1900.42	2289.07	1299.23	1975.36	2683.77	2653.52	2440.78	1717.17

\*段落数、文数、文字数の下段は1資料あたりの数を表す。

## 2.2 段落の役割：「問題の提示」と「結論を述べる」

先に述べたように、説明的文章は読み手にとって不明な事柄について説明する文章であり、教科書の説明的文章に多く見られる狭義の説明文では、文章の目的上、問いと答えという構造が基本となる。さらに、樺島（1980：130-136）が、理解を高めたり問題解決を促したりする説明的文章の構成を、大きく「問題提起」「問題解決」に分けて説明しているように、説明文を中心としつつも、広く説明的文章において、同様の構造が取り出せる可能性がある。しかし、「問題の提示」「結論を述べる」部分が判定可能な形で明示されていなければ、取り出して分析することは難しい。樺島（1980：139）では、読み手の正確な理解を助けるには、問題提起、観察結果、結論などの内容をそれぞれ「～はなぜだろうか。」「次のことが分かった。」「結論は～である。」などとはっきりと言語化するとよいとしている。この点、指導上「読み手の正確な理解を助け」になる表現を使用する動機を持つ小・中学校の教科書テキストでは、これらの表現が比較的明示されていることが期待でき、段落の役割を考えやすいものと思われる<sup>3</sup>。

本研究では、具体的な表現をもとに「問題の提示」「結論を述べる」部分を判定することで、各学年の説明的文章に連続する段落構造の一端を捉えたい。それぞれの部分を判定するための考え方について次節に述べる。

### 2.2.1 「問題の提示」部分

問題の提示は、個々の文章によってさまざまな形で表示されるものと考えられる。しかし、「問題の提示」という内容を表示する最も単純な形は疑問文であると思われることから、本研究では疑問文を頼りとして「問題の提示」を担う箇所を見ていくこととする。

日本語記述文法研究会編（2003：21-22）では、疑問文のタイプとして、質問、確認要求、

<sup>3</sup> 例えば、小学校4～6学年の説明的文章における疑問文の使用を調査した尾崎（2023）では、「(疑問詞) ...だろう (か)。」といった表現が45編の資料中の68か所に用いられていた。

疑いの疑問文を挙げている。質問には、①その内容に関して不明の点があるために話し手の判断が成り立たず、②話し手は聞き手に問いかけることによってその疑問を解消しようとする、という二つの基本的性質があり、この①の基本的性質が欠けたものを確認要求、②の欠けたものを疑いの疑問文としている。本研究の対象資料に見られるこれらの疑問文の例を以下に挙げる。

【質 問】主な形式：…（ない・の）か。

(1) ところで、紙には大きく分けて二つの種類があることを知っていますか。

(小5：町田誠之「和紙の心」学校図書)

【確認要求】主な形式：…ね。／…だろう。

(2) みなさんは、学校や家庭で、リサイクルや省エネルギーに取り組んだことがあるでしょう。

(小6：石田秀輝「自然に学ぶ暮らし」光村図書)

【疑 い】主な形式：…（ない・の）だろう。／…（ない・の）だろうか。

(3) 的中率は、どうして高くなったのでしょうか。

(小5：武田康男「天気を予想する」光村図書)

説明的文章で提示される問題となる疑問には、書き手によって説明・解答が示される必要がある。そのため、読み手に問いかけることによって疑問内容を満たそうとする②の性質を持つような疑問文は、本研究で対象としたい疑問文とは性質が異なっている。この点は上に挙げた以外のタイプの疑問文であっても同様である<sup>4</sup>。そのため、「…だろう（か）」などの形式を頼りに(3)のような疑いの疑問文を取り出し、その文とその文が含まれる段落を分析対象とすることにした。

### 2.2.2 「結論を述べる」部分

前節で、問題となる部分を疑問文によって定めたため、「問題の提示」を担う疑問文が不明点としている内容に答えている箇所を、提示した問題に対する結論と考えることができる。例えば、次のような場合である。

(4) いろいろなとりのくちばしのかたちをみてみましょう。

さきがするどくどがったくちばしです。

これは、なんのくちばしでしょう。

これは、きつつきのくちばしです。

(小1：村田浩一「くちばし」光村図書)

このような箇所を、結論を述べている部分(=解答文)として取り出す。ただし、不明点に対する答えとなる内容は、一文中に認められる場合もあれば、段落やそれ以上に大きな部分にまたがって複数文に認められることもある。例えば次のような場合である。

<sup>4</sup> このほかにも、感嘆の「(どれほど) ~だろう」や情報を持ちかける「~のではない(だろう)か」のような類も、疑問の解消を目的としておらず、対応する説明・解答を持つとは考えにくい。

- (5) Q1 それぞれのじどう車は、どんなしごとをしていますか。  
 Q2 そのために、どんなつくりになっていますか。  
 A1-1 バスやじょうよう車は、人をのせてはこぶしごとをしています。  
 そのために、A2-1 ざせきのところが、ひろくつくってあります。 そののけしきがよく見えるように、A2-2 大きなまどがたくさんあります。  
 A1-2 トラックは、にもつをはこぶしごとをしています。  
 そのために、A2-3 うんてんせきのほかは、ひろいにだいになっています。 A2-4 おもいにもつをのせるトラックには、タイヤがたくさんついています。  
 A1-3 クレーン車は、おもいものをつり上げるしごとをしています。  
 そのため、A2-5 じょうぶなうでが、のびたりうごいたりするように、つくってあります。 車たいがかたむかないように、A2-6 しっかりしたあしが、ついています。  
 (小1:「じどう車くらべ」 光村図書)

このようなときでも、Q1 に対して、A1-1～A1-3 が、Q2 に対し A2-1～A2-6 が不明点に対する答えとなる内容であるため、それぞれを結論を述べている部分とする。ただし、疑問内容に対応する形で答えていない場合や、疑問内容の提出のみにとどまっている場合など、不明点に対する答えとなる内容が見当たらない場合もあり、このようなときには結論を述べている箇所がないものとして扱った。

### 2.3 分析方法

指導内容の深化や児童生徒の能力の発達に応じるため、学年の段階に伴って教科書テキストの文章の構造にも何らかの変化が見られるものと思われる。本研究では、問題の提示となる部分と結論を述べる部分に着目して構造を探りたいが、これらに着目して得られるどのような特徴が文章構造の実態や差異を掴むのに有効かは定かではない。そのため、まず疑問文および解答文（の含まれる段落）について、頻度、大きさ、両者の距離などの値を細かく調査したのち、これらの値を変数として主成分分析（Principal Component Analysis, PCA）を行う<sup>5</sup>。主成分分析は、多くの変数を持つデータについて、その変数の情報を保存しつつ、少数の成分へと圧縮する手法である。この手法を用いることで、データの情報を代表する値をとり出せるだけでなく、変数や個々のケースの関係を俯瞰することができる。そのため、本研究においても、複数の変数を総合的に分析することができるため、それぞれの変数から各学年の資料の特性を考えるにあたって有効なものであると考え、取り入れることとした。

以上を踏まえ、次節以降では、各資料における「問題の提示」「結論を述べる」部分について疑問文・解答文をもとに集計したデータを概観し、そのデータを変数として用いた主成分分析を行った結果を順次考察していく。

## 3. 分析と考察

### 3.1 本資料における「問題の提示」「結論を述べる」段落の概観

まず、対象となる疑問文とその疑問内容に対応した答えを持つ解答文について、各学年の資料中での用いられ方を集計したのが表 2 である。

<sup>5</sup> 主成分分析にはフリー解析ソフト R (ver.4.2.2) の関数 `prcomp()` を用いた。Scale は T として相関係数行列を指定した。

疑問文は対象とする疑問文数である。解答ありとしたのは、疑問内容に答える解答文をもつ疑問文であり、解答文なしとしたのは疑問内容に答える解答文を持たない疑問文である。解答ありの疑問文のうち、一文で段落となっている箇所（以降、一文段落）に用いられている場合も特に数えた。本資料では、疑問に対する解答が複数文に分散して見受けられる場合も多かったため、これを疑問文に対する解答文数で示している。疑問文・解答文を含む段落の大きさは当該段落を構成する文数である。また、対応する疑問-解答段落間が、段落にしていくつに隔てられているかについて、各資料において最小値、中央値、最大値を求めた。

表2 各学年のテキストにおける疑問文および解答文の出現状況

項目	小学校						小計	中学校			小計	合計	
	1年	2年	3年	4年	5年	6年		1年	2年	3年			
疑問文 (/文数)	32 10.0%	17 3.2%	26 3.8%	39 5.2%	37 4.2%	28 3.7%	179 4.6%	42 4.2%	84 5.5%	38 2.8%	164 4.2%	343 4.4%	
解答文あり (/疑問文)	31 96.9%	14 82.4%	21 80.8%	32 82.1%	34 91.9%	22 78.6%	154 86.0%	34 81.0%	40 47.6%	27 71.1%	101 61.6%	255 74.3%	
一文段落かつ 解答文あり (/解答文あり)	21 67.7%	4 28.6%	5 23.8%	9 28.1%	8 23.5%	5 22.7%	52 33.8%	4 11.8%	5 12.5%	5 18.5%	14 13.9%	66 25.9%	
解答文なし (/疑問文)	1 3.1%	3 17.6%	5 19.2%	7 17.9%	3 8.1%	6 21.4%	25 14.0%	8 19.0%	44 52.4%	11 28.9%	63 38.4%	88 25.7%	
解答文 (/文数)	73 22.9%	20 3.8%	48 7.1%	38 5.1%	43 4.9%	31 4.1%	253 6.5%	33 3.3%	45 3.0%	35 2.5%	113 2.9%	366 4.7%	
疑問文に対する解答文数	2.4	1.4	2.3	1.2	1.3	1.4	1.6	1.0	1.1	1.3	1.1	1.4	
疑問段落の大きさ	1.35	2.29	2.96	3.23	2.97	3.52	2.75	4.34	6.32	4.63	5.43	3.32	
解答段落の大きさ	1.79	3.30	3.10	4.03	3.91	3.68	3.10	4.66	6.29	5.03	5.43	3.63	
疑問-解答 の距離	最小値	0.95	1.24	1.35	0.89	0.37	1.27	1.01	0.32	1.77	0.70	0.93	0.98
	中央値	0.25	0.28	0.23	0.22	0.12	0.17	2.00	0.12	0.18	0.09	1.00	0.18
	最大値	2.41	2.14	2.70	3.26	2.84	2.87	2.68	3.95	3.73	2.52	3.27	2.94

### 3.2 主成分分析の結果と考察

#### 3.2.1 主成分分析の結果

主成分分析にあたって、各資料における、①疑問文・解答文の数に関する情報、②疑問文・解答文を含む段落の距離に関する情報、③疑問文・解答文を含む段落の大きさに関する情報を変数として用いる。具体的な項目は次の通りである。

① 各資料における疑問文・解答文の数に関する情報

…解答文あり疑問文数、解答文あり疑問文に占める一文段落の疑問文の割合、解答文なし疑問文数、解答文数、疑問-解答のペア数

- ② 疑問文・解答文を含む段落の距離に関する情報  
...最小距離、距離中央値、最大距離の当該文章比
- ③ 疑問文・解答文を含む段落の大きさに関する情報  
...疑問文を含む段落の大きさ、解答文を含む段落の大きさの当該文章比

また、疑問、解答の構造に着目するため、以降は183の資料のうち、疑問文が1文以上用いられている135の資料を分析対象とする。各資料から得られたこれらの項目を変数として用いて主成分分析を行った結果、10の主成分を得ることができた。それぞれの主成分の固有値、寄与率、累積寄与率は以下の通りである（表3）。

表3 固有値・寄与率

項目	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
固有値	1.73	1.56	1.20	1.09	0.87	0.72	0.55	0.48	0.26	0.23
寄与率	0.30	0.24	0.14	0.12	0.08	0.05	0.03	0.02	0.01	0.01
累積寄与率	0.30	0.54	0.69	0.81	0.88	0.93	0.97	0.99	0.99	1.00

本研究では、固有値が1以上である第1主成分（PC1）から第4主成分（PC4）までを用いる。これらの寄与率はそれぞれ0.30、0.24、0.14、0.12であり、PC4までの累積寄与率は0.81となる。PC4までで元データの81%程度の情報量を持つことになり、分析に耐えうる量の情報が反映されていると考えられる<sup>6</sup>。

次に、PC1からPC4までの主成分負荷量を表4に示す。なお、本研究では主成分負荷量の解釈にあたり、一般に中程度の相関が期待される絶対値が0.4以上の値に着目することとする。次節では、PC1からPC4の主成分負荷量とその主成分によって特徴づけられた各資料の実例をもとに、それぞれの主成分を解釈していく。

表4 主成分負荷量

変数	PC1	PC2	PC3	PC4
解答文あり疑問文	0.57	-0.70	0.14	0.00
一文段落かつ解答文あり疑問文	0.14	-0.05	0.37	0.72
解答文なし疑問文	-0.08	-0.21	0.27	-0.76
解答文	0.69	-0.45	-0.01	0.03
疑問-解答のペア数	0.58	-0.73	0.22	-0.13
最小距離	0.43	0.75	0.27	-0.13
距離中央値	0.66	0.67	0.24	-0.08
最大距離	0.80	0.38	0.19	-0.07
疑問段落の大きさ	0.33	0.11	-0.83	-0.15
解答段落の大きさ	0.69	0.02	-0.56	0.15

\*0.4以上の値に網掛けを、-0.4以下の値に下線を施した。

<sup>6</sup> 一般に、固有値1.0以上の主成分を有効とし、これに併せて累積寄与率が60~80%以上となる主成分を目安に選択される（水本2010：202-203）。

## A PC1：総合値

PC1では、ほとんどの変数において正の相関を持っており、変数を縮約した総合値と見ることができる。その中でも、疑問-解答段落間の最大距離において強い正の相関を持っている。そのため、典型的には、疑問文-解答文のペアを多く備えながらも、疑問-解答間で説明等を行うことで距離をおき、説明や主張と併せて最後に結論となる解答を示すような文章構造の場合に正の方向に位置する軸と言える。この値が大きな正の値をとる構成の典型と言えるのは、疑問に順次答えていく先の(5)や、冒頭で提示した疑問について説明を加えていき文章末尾で解答を述べる次のような構成である。

- (6) 秋になると、この木にもどんぐりがたくさんなりました。じゅくしたどんぐりは、親木のねもとにおちはじめました。これらのどんぐりは、どうなるでしょう。

(中略)

<sup>A-1</sup> 地面におちたままのどんぐりは、そのままかわいてだめになっていました。ねが地面にささらなかったのです。<sup>A-2</sup> リスやネズミのすあなふかくうめられたどんぐりは、めが土の上まで出られず、かれています。ところが、<sup>A-3</sup> 地面にあさくうめられたどんぐりの中で、食べわすられたものからは、めが出ていました。

(小2：こうやすすむ「どんぐり」学校図書)

一方で、疑問-解答のペアをほとんど持たず、明示的な解答が行われない疑問しか持たない次のような場合には、この値が大きく負に偏っている。(7)の文章では、動物すら意識や認識が誤魔化されてしまう例として挙げられた、蝶と造花のエピソードの中で、自身すら答えられない疑問内容に疑問文を使用している。

- (7) チョウは紫外線も含めた花の色に、非常に敏感であるとのことでした。それでも、四時間も偽物の花と戯れる理由があったのでしょうか。人間が作り出した偽の花には、昆虫にとって必要な栄養分はありません。本能もごまかされたとしたら、待っているのは死のみです。

(中3：毛利衛「文化としての科学技術」教育出版)

## B PC2：疑問の量と説明の長さ

PC2では、解答文あり疑問文、解答文およびそれらのペアの数と負の相関を持ち、疑問-解答段落間の最小距離、中央値と正の相関を持っている。疑問-解答の構造自体が少ないのにもかかわらず疑問-解答段落間の距離が大きくなるのは、文章全体を貫く大きな疑問に派生する疑問をあまり生じさせずに時間をかけて答えているような場合であると想像される。そのため、疑問の量と説明の長さを示す軸であると推察される。例えば、PC1で大きな値を与えられるような文章よりも、次の(8)のように、冒頭で与えられる大きな問いのみを持ち、文章末尾でそれに答えている文章の方が大きな値を与えられる傾向がある。

- (8) ありは、ものがよく見えません。それなのに、なぜ、ありの行列ができるのでしょうか。

(中略)

このように、においをたどって、えさの所へ行ったり、巣に帰ったりするので、

ありの行列ができるというわけです。

(小3：大滝哲也「ありの行列」光村図書)

#### C PC3：「問題の提示」「結論を述べる」部分が段落として独立するか

PC3では、疑問文を含む段落の大きさ、解答文を含む段落の大きさで負の相関を持っている。また、0.37と本研究の基準以下ではあるが一文で段落になる場合との正の相関も見られる。そのため、この軸は、「問題の提示」「結論を述べる」部分が段落として独立するかを示している軸であると解釈する。(9)は森林の恵みや働きについて、様々な観点から細かく疑問を立てては解答を述べることで進んでいく構成の文章である。そのため、細分化された疑問と解答が置かれる段落自体も小さくなっている。

(9) Q1 下流に下された木材は、では、どのように使われたのでしょうか。

A1 まっすぐなスギやヒノキの大木は、神社やお寺を建てるのに使われました。

(中略)

Q2 森林の土は、なぜ雨に流されてなくなってしまうのでしょうか。

わたしたちが、ベランダやコンクリートの道の上に、植木ばちの土をひっくり返したりして放っておくと、雨や風でいつのまにか土はなくなってしまう。ところが、山のしゃ面にはいつも土があります。

それもまた、森林のおかげです。A2 森林の根がしっかりと土をかかえて、しゃ面にはり付けているのです。

(小5：富山和子「森林のおくりもの」東京書籍)

#### D PC4：疑問の投げかけか疑問だけを卓立させるか

PC4では、解答文あり疑問文に占める一文段落の割合において正の相関を、解答文なし疑問文数において負の相関を持っている。解答文なしの疑問文は、明示的な解答の示されないままに終結する場合の他にも、文章の末尾などで派生する疑問文が提示される場合や問題を提出して投げかける場合などが考えられる。(10)は、この値が負の方向に大きくなっている例である。Q1～Q3は、いずれも本文末尾で示される解答されることのない疑問であるが、解答されることがないために、単なる疑問点として同じ段落の中で複数列举されている。逆に、(4)のように一文段落で疑問を卓立させ説明を加えていく場合には大きく正に傾く。そのため、この軸は、疑問の投げかけか疑問だけを卓立させるかの対立が表現された軸であると言える。

(10) Q1 では、日本の近くの海にもどってきたウミガメは、どのような生活をするのでしょうか。長崎の海でつかまえたウミガメを調べると、まだ、たまごを産んだことのないメスでした。Q2 このまま、日本の近くにとどまって、いずれたまごを産むのでしょうか。Q3 たまごを産むとしたら、それまでの間、どのような生活をするのでしょうか。

これらのぎもんをとき明かすために、このウミガメをもう一度、つかまえた海の近くで放流することになりました。

(小4：松田乾「ウミガメの命をつなぐ」教育出版)

## 3.2.2 各学年テキストにおける問い-答えの段落構成

本節では、以上の PC1~PC4 の示す傾向の解釈を踏まえ、各学年の資料において、それぞれの値の正負がどのように分布しているかをまとめた表 5 をもとに、各学年の資料全体の大まかな傾向を考える。

表 5 各学年資料の主成分得点の正負

正負のパターン				学年									合計
PC1	PC2	PC3	PC4	e1	e2	e3	e4	e5	e6	j1	j2	j3	
+	+	+	+	2	2	1	2	1	0	2	0	0	10
+	+	+	-	0	1	2	2	1	1	1	3	1	12
+	+	-	+	1	0	0	0	0	0	0	0	0	1
+	-	+	+	0	1	1	1	2	1	0	1	1	8
+	+	-	-	4	2	2	0	0	0	0	0	1	9
+	-	+	-	0	0	0	1	2	0	2	3	2	10
+	-	-	+	2	0	2	2	2	0	0	0	0	8
+	-	-	-	2	0	0	1	0	1	3	2	0	9
-	+	+	+	0	1	0	1	1	1	0	0	2	6
-	-	+	+	4	0	1	0	0	2	0	0	0	7
-	+	-	+	1	1	0	0	0	1	1	0	1	5
-	+	+	-	0	2	1	1	1	2	3	3	2	15
-	-	-	+	0	0	1	1	2	0	3	1	2	10
-	-	+	-	0	0	0	0	0	0	1	2	1	4
-	+	-	-	0	3	3	0	0	1	2	2	3	14
-	-	-	-	0	0	1	1	1	2	0	2	0	7
合計				16	13	15	13	13	12	18	19	16	135

\*+は正の値を、-は負の値を表す。網掛けは3以上の値を持つ類型に施した。

第1学年のテキストでは、PC1が正である資料が多いほか、PC3が負あるいはPC4が正のものも比較的多い。第1学年のテキストでは、疑問-解答による構成を基本としながら、ときに解答を複数文に分け、独立させながら対応付けて説明していく基本構成が窺える。一文段落が基調となることで、文章の理解が助けられているものと考えられるほか、総段落数が比較的小さい文章が多いため、PC3が正になるには段落の相対的な大きさが他の資料に比べて大きいものと推察される。また、PC4が負となるような、解答が示されないような結び方や叙述が少ないことについては、一定の型の存在や、幅広い説明的文章としてのジャンルに属する文章が少ないことなどが要因として想定される。

第2・第3学年では、PC1あるいはPC2を正とする資料が過半数を占めている。PC1の値が正の資料を中心としているのは第1学年同様であるが、PC2の値が大きい資料が多くなっているのは、疑問から解答に至るまでの間の説明が大きくなって対応できる読解能力の育成や、説明的文章の型の幅の広がりなどの要因が考えられる。



第4・第5学年では、PC1とPC3を正とする資料が半数程度を占めている。第4・第5学年では、第6学年同様、一文段落自体の割合が下級学年に比べて少なくなっている一方で、問題や結論の部分を段落として卓立させて疑問-解答の構造に組み込むことで、文章の構造や段落の役割の把握を助けているものと推察される。

第6学年では、PC1を負とする資料が過半数を占める。本資料では、第6学年の資料を境に、PC1の値の正負をとる資料数が逆転している。疑問-解答の構造をふんだんに用いた文章構成ではなく、必ずしも疑問-解答の明示的な形をとらずに説明が進み、疑問を用いても提示のみに留められたり、文単位で解答となるような箇所が見つからなかったりするような文章構成が次第に増えてくる様子が表れているものと思われる。中学生のテキストに向けて、説明的文章として扱われるジャンルがさらに広がっている可能性がある。

中学校テキストでは、さらにPC4が負となる資料が多くなっている。第6学年での流れがさらに強まり、評論・論説に類する文章などが増えることで、小学校における説明的文章の構造外の方法での疑問の使用が際立ったことが考えられる。

#### 4. まとめ

本研究では、特に段落の学習が期待される教材である説明的文章において、学習指導要領解説で「段落の役割」として例示されていたもののうち、説明的文章の性質上、骨子となる要素であると考えられる「問題の提示」(疑問)や「結論を述べる」(解答)といった働きをする部分が、どのように段落を成しつつ、学年ごとに文章構造を構成しているかを分析した。その結果は以下のようにまとめられる。

- ① 小学校・中学校のテキストにおける「問題を提示」「結論を述べる」部分を、それぞれ疑いの疑問文とその疑問内容に答える文によって捉えた。その結果、135の資料から、343の疑問文、366の解答文を収集した。
- ② 疑問文・解答文とそれらが含まれる段落について、その数、大きさ、間の距離の観点から10の変数を定めて主成分分析を行った。その結果、PC1~PC4の四つの変数が得られた。主成分負荷量から、これらはそれぞれ、PC1が総合値、PC2が疑問の量と説明の長さ、PC3が「問題の提示」「結論を述べる」部分が段落として独立するか、PC4が疑問の投げかけか疑問だけを卓立させるかを表すものと考えられた。
- ③ 各学年の資料における主成分得点の正負から次のような傾向が指摘できた。
  - a. 疑問-解答の構成を多用し、基調とする小学校第1学年~第5学年までのテキストのうち、第1学年では、疑問-解答を段落として卓立させつつ、ペアとして示す構成が重用される。
  - b. 第2・第3学年では大きな疑問と解答のみのシンプルな型が増えてくる。
  - c. 第4・第5学年では、文章自体のサイズが大きくなるものの、疑問-解答の段落は一文段落などで卓立させながら示されている。
  - d. 小学校第6学年以降は、疑問-解答の構成が多用されなくなり、説明的文章内の多様なジャンルにより文章構成にも幅が生じる。
  - e. 中学校のテキストでは、後に解答することを予定した疑問だけでなく、提示されるだけのものや、明確な解答の見つからない疑問の使用が増える。

以上により、各学年のテキストにおいて、説明的文章の骨子であると考えられる疑問と解答からなる構造がどのように用いられているかを俯瞰することができた。段落から情報を読み取る術を示すためには、段落の実態がどのようなものかをより深く知っておくことは重要であると考え。今回検討できなかった、疑問文を節として文の中に含む文や、疑問文・解答文中の形式（疑問詞、理由の「から」、文末の「のだ」など）との関わりについても、今後検討を重ねていきたい。

## 謝 辞

本研究は、JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2102 の支援を受けたものです。

## 文 献

- 大熊徹（2018）「説明的文章」中村明・佐久間まゆみ・高崎みどり・十重田裕一・半沢幹一・宗像和重（編）『日本語文章・文体・表現事典：新装版』, p.88, 朝倉書店.
- 尾崎誉治（2023）「説明的文章における段落と疑問文の関係性」『国語学研究』62, pp.116-130.
- 樺島忠夫（1980）『文章構成法』講談社.
- 櫻本明美（2009）「説明文」日本国語教育学会（編）『国語教育辞典：新装版』, p.249, 朝倉書店.
- 日本語記述文法研究会（編）（2003）『現代日本語文法4：第8部 モダリティ』くろしお出版.
- 水本篤（2010）「主成分分析：データの情報を圧縮する」石川慎一郎・前田忠彦・山崎誠（編）『言語研究のための統計入門』pp.139-161, くろしお出版.

## 資 料

本研究で用いた文章について、タイトルと学年のみ以下に挙げる。なお、分析に用いた 135 編の文章についてはタイトルの前に●を付して示した。

### 【小学校第1学年】

- 「うみのかくれんぼ」●「くちばし」「くらべてよもう」●「じどう車くらべ」●「どうぶつの赤ちゃん」「いろいろなふね」●「かんじのはなし」●「じゃんけん」●「どうやってみをまもるのかな」●「歯がぬけたらどうするの」「かんじのはじまり」●「しっぽしっぽ」「なにができるかな」●「ぼうしのはたらき」●「すずめのくらし」●「だれが、たべたのでしょうか」「はたらくじどう車」「みぶりでつたえる」●「いきもののあし」●「くらしをまもる車」●「なぜでしょうか」●「めだかのぼうけん」

### 【小学校第2学年】

- 「おにごっこ」「しかけカードの作り方」●「たんぼぼのちえ」「どうぶつ園のじゅうい」●「あなのやくわり」「たんぼぼ」「ふろしきは、どんなぬの」「ビーバーの大工事」●「たねのたび」●「つばめのすだち」「紙パックで、こまを作ろう」「きつつき」●「さけが大きくなるまで」●「すみれとあり」●「あいさつのみぶりとことば」●「たこのすみいかのすみ」「とべとべ回れ」●「どんぐり」●「ほたるの一生」●「エンペラーペンギンの子そだて」●「食べるのは、どこ」

**【小学校第3学年】**

- 「ありの行列」●「こまを楽しむ」「すがたをかえる大豆」●「言葉で遊ぼう」●「ほけんだよりを読みくらべよう」「もうどう犬の訓練」●「ヨットカーの作り方」「人をつつむ形：世界の家めぐり」●「自然のかくし絵」●「道具を使う動物たち」「農業をする魚」●「米と麦」●「身ぶりのはたらき」●「くらしと絵文字」●「どちらが生たまごでしょう」●「めだか」●「ネコのひげ」「ミラクルミルク」●「冬眠する動物たち」●「合図としるし」

**【小学校第4学年】**

- 「アップとルーズで伝える」「ウナギのなぞを追って」「大きな力を出す」「動いて、考えて、また動く」「着るロボットを作る」●「くらしの中の和と洋」●「ヤドカリとイソギンチャク」●「色さいとくらし」●「打ち上げ花火のひみつ」●「じゃんけんの仕組み」●「月のかげ絵」●「『便利』ということ」●「ウミガメの命をつなぐ」●「花を見つける手がかり」「さわっておどろく」●「アメンボはにん者か」●「手で食べる、はしで食べる」●「だまし絵で分かる脳のしくみ」●「『落ち葉』ではなく『落ちえだ』」

**【小学校第5学年】**

- 「想像力のスイッチを入れよう」●「生き物は円柱形」●「天気を予想する」●「ニュース番組作りの現場から」「見立てる」●「和の文化を受けつぐ：和菓子をさぐる」●「テレビとの付き合い方」●「動物の体と気候」●「森林のおくりもの」●「『十秒』が命を守る」●「動物の『言葉』人間の『言葉』」●「ブナの森が支える豊かな自然」●「白神山地の自然保護：緩衝地域の役割」●「まんがの方法」●「言葉と事実」「生命のかて・塩」●「メディア・リテラシー入門」●「東京スカイツリーのひみつ」●「和紙の心」

**【小学校第6学年】**

- 「時計の時間と心の時間」●「生き物はつながりの中に」「笑うから楽しい」●「自然に学ぶ暮らし」●「町の幸福論：コミュニティデザインを考える」●「未来に生かす自然のエネルギー」●「イースター島にはなぜ森林がないのか」●「『なべ』の国、日本」●「ぼくの世界、君の世界」●「森林のはたらきと健康」●「『本物の森』で未来を守る」●「自分の脳を自分で育てる」●「国境なき大陸 南極」「発明・発見は、はてなから」●「ジブシーとの約束」

**【中学校第1学年】**

- 「幻の魚は生きていた」●「ちょっと立ち止まって」●「ダイコンは大きな根？」●「スズメは本当に減っているか」●「ニュースの見方を考えよう」●「オオカミを見る目」●「クジラの飲み水」●「月を思う心」●「玄關扉」●「食感のオノマトペ」●「言葉がつなぐ世界遺産」●「笑顔という魔法」●「花の形に秘められたふしぎ」●「言葉のゆれを考える」●「電子レンジの発想」●「ものづくりに生きる」「変わる動物園」●「片言を言うまで」「生物が消えていく」●「絵巻を読む」「言葉のいのち」「言葉の向こうに」

**【中学校第2学年】**

- 「モアイは語る」●「君は『最後の晩餐』を知っているか」●「生物が記録する科学：バイオロギングの可能性」●「科学はあなたの中にある」「メッセージをどう聞くか」●「哲学的思考のすすめ」●「正しい言葉は信じられるか」●「鯉節：世界に誇る伝統食」●「壁に残された伝言」●「動物園でできること」●「人間は他の星に住むことができるのか」●「水の上富士山」●「学ぶ力」●「日本の花火の楽しみ」●「ガイアの知性」「プロセスの建築」「古典芸能に見られる古典解釈」●「孫が読む漱石：坊ちゃん」●「短歌：佐藤正午『ありのすさび』と解説」●「若者が文化を創造する」●「言葉の力」●「逃げることは、ほんとにひきょうか」

**【中学校第3学年】**

- 「『批評』の言葉をためる」「作られた『物語』を超えて」●「月の起源を探る」「誰かの代わりに」●「いつものように新聞が届いた：メディアと東日本大震災」●「絶滅の意味」●「サハラ砂漠の茶会」●「何のために『働く』のか」●「黄金の扇風機」●「『文殊の知恵』の時代」「情報社会を生きる：メディア・リテラシー」●「フロン規制の物語：〈杞憂〉と〈転ばぬ先の杖〉のはざままで」●「間の文化」●「新しい博物学の時代」●「文化としての科学技術」●「歴史は失われた過去か」「ディズニーランドという聖地」●「俳句：小林恭二『俳句とい

うゆしみ』と解説」「先人の達成と共に：本歌取りなど」●「武蔵野の風景：二次的な自然環境について」「言葉との出会い」「運動会」●「顔の見える国際協力」

### 関連 URL

文部科学省『小学校学習指導要領解説：国語編』（平成 20 年告示）

[https://www.mext.go.jp/component/a\\_menu/education/micro\\_detail/\\_icsFiles/afieldfile/2010/12/28/1231931\\_02.pdf](https://www.mext.go.jp/component/a_menu/education/micro_detail/_icsFiles/afieldfile/2010/12/28/1231931_02.pdf)

文部科学省『中学校学習指導要領解説：国語編』（平成 20 年告示）

[https://www.mext.go.jp/component/a\\_menu/education/micro\\_detail/\\_icsFiles/afieldfile/2014/04/15/1234912\\_2\\_1.pdf](https://www.mext.go.jp/component/a_menu/education/micro_detail/_icsFiles/afieldfile/2014/04/15/1234912_2_1.pdf)

文部科学省『小学校学習指導要領解説：国語編』（平成 29 年告示）

[https://www.mext.go.jp/content/20220606-mxt\\_kyoiku02-100002607\\_002.pdf](https://www.mext.go.jp/content/20220606-mxt_kyoiku02-100002607_002.pdf)

文部科学省『中学校学習指導要領解説：国語編』（平成 29 年告示）

[https://www.mext.go.jp/component/a\\_menu/education/micro\\_detail/\\_icsFiles/afieldfile/2019/03/18/1387018\\_002.pdf](https://www.mext.go.jp/component/a_menu/education/micro_detail/_icsFiles/afieldfile/2019/03/18/1387018_002.pdf)

## アーカイブデータを利用した言語研究とその応用可能性

鈴木成典（国際基督教大学大学院アーツ・サイエンス研究科、日本学術振興会）<sup>1</sup>

鎌野慈人（ストーニーブルック大学大学院）

坂本誓（国際基督教大学 RA）

鎌倉欧亮（国際基督教大学教養学部）

Seunghun Lee（国際基督教大学, ヴェンダ大学）

Yu Yan（立命館大学）

Jeremy Perkins（会津大学）

五十嵐陽介（国立国語研究所）

## Linguistic Research using Archive Data and its Application

Michinori Suzuki (International Christian University Graduate School & Japan Society for the Promotion of Science)

Shigeto Kamano (Stony Brook University Graduate School)

Chikau Sakamoto (International Christian University RA)

Osuke Kamakura (International Christian University, College of Liberal Arts)

Seunghun Lee (International Christian University & University of Venda)

Yu Yan (Ritsumeikan University)

Jeremy Perkins (University of Aizu)

Yosuke Igarashi (National Institute for Japanese Language and Linguistics)

### 要旨

本稿は、国立国語研究所の共同利用型共同研究で利用可能である豊富な音声データベースに対して、我々が実際に行っている音声データの処理方法を紹介する。使用したデータベースは大規模な録音実験に基づくものであり、1000人以上の話者の録音が存在する一方で、方言ごとに刺激リストが異なっている。そこで、初めに各方言の録音を確認し、セクション・刺激・話者が識別可能なアーカイブ ID を作成した。その際、同一方言内の話者間でも刺激の順番や繰り返しの回数にばらつきが確認されたが、Praat スクリプトや Excel を用いることで対応を可能とした。様々な方言話者を対象に産出実験を実施する際、必ずしも刺激リストの語彙や順番の通りに録音されないため、本処理方法を用いることでアーカイブデータを利用した今後の研究に役立てることができるだろう。

### 1. はじめに

本稿では、国立国語研究所の共同利用型共同研究（B）において利用可能な、過去の研究プロジェクトにて収集された豊富なアーカイブデータを取り上げ、これを用いた言語資源の整理とメタデータの設計、及び音声データの処理方法に関し、工夫した点について紹介する。共同利用型共同研究は、国立国語研究所の保有する多種多様な研究資料や言語資源等を使用して研究を行うことができる制度であり、著者らは、「国立国語研究所研究資料室収蔵資料」の「fo0245：日本語音声における韻律的特徴の実態とその教育に関する総合的研究」において収集・アーカイブされた録音音声データを用いて日本語の有声性に関する研究を

---

<sup>1</sup> michinorisuzuki19[at]gmail.com

実施している。本データベースは様々な日本語方言を対象にした大規模な録音実験に基づいて作られたものであり、1100人以上の話者の豊富な録音データを含んでいる。同時に、アーカイブされている音声データは各被験者につき録音実験全体の音声ファイルのみであるため、まず初めに各方言の録音を聴き、どのような刺激の種類でセクションが分かれているかを確認することで刺激のID作成を試みた。しかしながら、方言間で録音されているセクションの種類や、同一セクション内での刺激リストなど、複数の点においてばらつきが散見されており、同一の刺激ID作成は困難であった。方言内においても話者ごとに刺激の順番や繰り返しの回数が異なっていた。そのため、録音をもとに Section ID、Word ID、繰り返しの回数、Speaker ID が識別可能な刺激IDを効率的に作成する方法を考案した。本データベースのように、様々な方言及び年代の話者を対象とした録音実験を実施する場合、必ずしも一つの刺激リスト通りの刺激や順番で録音できない可能性がある。そのような場合でも、本稿で紹介する手順を用いることで、識別可能な刺激IDを付与することで、様々なトピックの言語研究を行うことが可能となると考えられる。

## 2. 使用しているデータベースについて

本節では、国立国語研究所の研究資料室収蔵資料「fo0245: 日本語音声における韻律的特徴の実態とその教育に関する総合的研究」の実験録音データ全体について説明を行う。

### 2.1 被験者

「日本語音声における韻律的特徴の実態とその教育に関する総合的研究」は、そのプロジェクト名の通り、日本語の韻律的特徴とその教育についての研究プロジェクトであり、日本語諸方言における様々な年代の話者の録音データが含まれている。各方言と各年代における話者の数を表1に示す。

表1 「fo0245: 日本語音声における韻律的特徴の実態とその教育に関する総合的研究」プロジェクトにおいて収集された方言の種類、ファイル数、及び録音時間

方言(方言コード)	ファイル数	録音時間(時間:分:秒)
札幌 (SPR)	127	78:44:45
弘前 (HRK)	99	85:56:21
仙台 (SED)	76	49:11:05
新潟 (NGT)	99	105:22:46
浅草 (ASK)	87	55:34:51
五日市 (ITK)	54	45:00:03
名古屋 (NGY)	83	60:41:20
富山 (TYM)	69	35:48:44
大阪 (OSK)	70	35:30:30
高知 (KOC)	77	51:49:09
広島 (HRM)	77	43:36:54
福岡 (FKO)	61	25:45:01
鹿児島 (KGS)	76	50:31:56
那覇 (NAH)	53	49:16:21
その他	120(43 都道府県)+148(琉球諸語辞典:奄美、石垣、今帰仁)	340:11:12
計: 14 方言+その他	計: 1376	計: 1113:00:58

表1が示すように、本データベースには1100時間以上の豊富なデータが含まれていることが分かる。

## 2.2 刺激

本データベースに含まれる音声ファイルには、(1)に示したような刺激の種類(セクション)が存在する。

### (1) 刺激の種類

- a. 名詞
- b. 動詞とその活用
- c. 形容詞とその活用
- d. 文章
- e. 童話(桃太郎)の朗読
- f. 天気概況
- g. 五十音・数字(1から9までと四桁の数字)
- h. 固有名詞(地名・人名)
- i. 会話

これらの刺激の種類は、オリジナルの研究プロジェクトの資料でも確認できた。しかし同時に、方言間で刺激のセクションが異なっていることや、刺激の重複や形式の一致により同一セクション内と考えられる部分でも異なる刺激が録音されていることがあると確認された。

## 2.3 実験手順

実際の録音では、実験者が被験者に渡す紙に書かれた単語や文章の読み上げや、絵を見せてその名前を言ってもらおうという方法を取っていた。また、被験者の言い間違いや周囲の騒音が入った際には再度言い直しを依頼していた。

## 3. データ整理

ここまで、使用したデータベースの説明を行なったが、本節ではデータ整理及び処理の際に困難であった点とその解決方法について述べる。

### 3.1 Speaker ID

アーカイブされていたエクセルにおいて、各音声ファイルに「通し番号」と「ファイル名(DAT ID)」が付けられており、これを参考に Speaker ID をつけていたが、1つのファイルに複数の話者が録音されていることや、複数の音声ファイルに同じファイル名がついていることがわかったため、新たに話者が識別可能な Speaker ID を作成した。整理後の各方言での話者の人数は表2の通りである。

表 2 「fo0245: 日本語音声における韻律的特徴の実態とその教育に関する総合的研究」プロジェクトにおいて収集された話者の人数 (方言ごと) <sup>2</sup>

	高齢層 <sup>3</sup>	壮年層	若年層	中学生	小学生
札幌 (SPR)	27	29	31	20	20
弘前 (HRK)	14	19	18	24	24
仙台 (SED)	13	10	11	20	21
新潟 (NGT)	28	18	13	20	18
浅草 (ASK)	31	26	2	30	0
五日市 (ITK)	32	20	2	30	0
名古屋 (NGY)	12	14	12	25	20
富山 (TYM)	10	6	13	20	20
大阪 (OSK)	10	6	14	20	20
高知 (KOC)	12	13	12	20	20
広島 (HRM)	9	11	12	23	22
福岡 (FKO)	10	10	10	17	14
鹿児島 (KGS)	10	14	12	20	20
那覇 (NAH)	12	4	7	15	15
計	230	200	169	304	234

### 3.2 Section ID

2.2 節でも述べた通り、方言間での刺激のセクションのばらつき、及び方言内での順番や繰り返しの回数のばらつきが存在していた。そのため、複数の話者に対し単一の刺激リストをもとに識別可能な **Stimuli ID** を作成し、各刺激に対しラベリングを行うというデータ処理の手順をそのまま適用することができなかった。

そこで、まず全ての方言の録音を若干名ずつ確認し、「どの方言にどのセクションが含まれているのか」を確認した。その結果、「ほぼ全ての方言データに共通するもの」「一部の方言データにしか存在しないもの」「セクション自体は他の方言と共通しているものの、実際の刺激が異なるもの」が存在していた。これらのセクションに対し、「fo0245: 日本語音声における韻律的特徴の実態とその教育に関する総合的研究」の研究資料を参考に **Section ID** を付与した。

### 3.3 Stimuli ID

方言ごとにセクションにバリエーションがあったものの、少なくとも同一方言データの同一セクションでは、刺激自体は共通であり、差があるのは繰り返しの回数のみであった。そのため、以下 (2) の手順で **Stimuli ID** の作成を行なった。

#### (2) 本データベースに対する刺激 ID 作成の手順

- a. まず一人の被験者の録音を参考に Excel に刺激のリスト (①) を書き出す
- b. 音声分析ソフトウェア Praat (Boersma & Weenink, 2023) を使用し、各話者のデータに対して Praat スクリプトを使用した刺激間の境界の配置

<sup>2</sup> 表 1 に記載されているデータに加え、43 都道府県から 120 名分の録音データや琉球諸語のデータも存在する。

<sup>3</sup> 録音をもとに、高齢層: 65 歳以上、壮年層: 40-64 歳、若年層: 19-39 歳 (録音当時) と推測される。



- c. Praat 上で各刺激の単語を Textgrid に手入力
- d. 別の Praat スクリプトを使用して入力した単語を取り出し（テキストファイル形式）、Excel（①と同じファイルの別シート）へ貼り付け
- e. Speaker ID、繰り返しの回数、貼り付けたも刺激の番号を入力
- f. VLOOKUP 関数で①を参照して word ID、ひらがな、英語訳を付与
- g. CONCATENATE 関数で「Section ID - word ID - 繰り返しの回数 - speaker ID」を組み合わせる（参考：図 1）

TokenID	Speaker	Section	Japanese	wordID	repetit	Hiragana	English
A1-W001-1-KOC012	KOC012	A1	富士	W001	1	ふじ	Fuji
A1-W001-2-KOC012	KOC012	A1	富士	W001	2	ふじ	Fuji
A1-W002-1-KOC012	KOC012	A1	藤	W002	1	ふじ	wisteria
A1-W002-2-KOC012	KOC012	A1	藤	W002	2	ふじ	wisteria
A1-W003-1-KOC012	KOC012	A1	鈴	W003	1	すず	bell
A1-W003-2-KOC012	KOC012	A1	鈴	W003	2	すず	bell
A1-W004-1-KOC012	KOC012	A1	地図	W004	1	ちず	map
A1-W004-2-KOC012	KOC012	A1	地図	W004	2	ちず	map
A1-W005-1-KOC012	KOC012	A1	巻き寿司	W005	1	まきずし	Sushi rolls
A1-W005-2-KOC012	KOC012	A1	巻き寿司	W005	2	まきずし	Sushi rolls
A1-W006-1-KOC012	KOC012	A1	三日月	W006	1	みかづき	crescent moon
A1-W006-2-KOC012	KOC012	A1	三日月	W006	2	みかづき	crescent moon
A1-W007-1-KOC012	KOC012	A1	頭巾	W007	1	ずきん	hood
A1-W007-2-KOC012	KOC012	A1	頭巾	W007	2	ずきん	hood
A1-W008-1-KOC012	KOC012	A1	缶詰	W008	1	かんづめ	canning
A1-W008-2-KOC012	KOC012	A1	缶詰	W008	2	かんづめ	canning
A1-W008-3-KOC012	KOC012	A1	缶詰	W008	3	かんづめ	canning
A1-W008-4-KOC012	KOC012	A1	缶詰	W008	4	かんづめ	canning
A1-W009-1-KOC012	KOC012	A1	火事	W009	1	かじ	fire
A1-W010-1-KOC012	KOC012	A1	舵	W010	1	かじ	rudder
A1-W010-2-KOC012	KOC012	A1	舵	W010	2	かじ	rudder

図 1：刺激 ID を作成した Excel ファイルの一例

作成した刺激 ID（及びひらがなと英語訳も）を Praat スクリプトでラベリングし、また別の Praat スクリプトを使用して個別刺激ファイルへの切り分け（chopping）を行う。その後、切り分けられた個別刺激ファイルに対しアノテーションを行った。この点に関しては、言語資源ワークショップ 2022 で発表したデータ処理方法（Suzuki, Igarashi, and Lee, 2022）を参照されたい。

#### 4. おわりに

パンデミックを機に、オリジナルの実験データだけでなく、過去の研究で収集されたアーカイブデータなどの言語資源を利用した研究が以前よりも大きく注目されることとなった。しかし、アーカイブデータは過去の一時点における研究手法や概念に基づいて収集されたものであるため、必ずしも現在ほど再現可能性が高くなるような手順で行われていない可能性がある。また、そもそも産出実験を実施する際、必ずしも刺激リストの通りに録音でき

ないことや、言い直しなどで刺激の順番が前後する可能性がある。このような一見処理しづらいアーカイブデータであっても、本稿で紹介した整理手順を応用することで再現性を高め、今後様々なトピックの言語研究へ活用が可能であると考えている。

また、本稿で紹介した手順は過去のアーカイブデータのみならず、今後の実験においても重要だと考えられる。識別可能な刺激 ID とともに個別刺激ファイルをアーカイブすることで、再現性を保ちつつ広く言語研究に応用可能なデータベースを作成することができるためである。

### 謝 辞

本稿は、国立国語研究所の共同利用型共同研究 (B) プロジェクト「東北・東京方言における有声性の対立への音響指標の影響」(研究代表者: 鈴木成典) の研究成果である。また、本研究は JSPS 科研費 23KJ1921 の助成を受けたものである。

### 文 献

- Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.14, retrieved 4 August 2023 from <http://www.praat.org/>
- 鈴木成典, 五十嵐陽介, & 李勝勲. (2023). NINJAL データベースを活用した言語研究の実施について. In 言語資源ワークショップ発表論文集= Proceedings of Language Resources Workshop (Vol. 1, pp. 79-82). 国立国語研究所.

## 書籍の文体と修辞機能の分析のパイロットスタディ

田中 弥生 (国立国語研究所 研究系) \*

柏野和佳子 (国立国語研究所 研究系)

加藤 祥 (目白大学)

## A Pilot Study of Analyzing Rhetorical Functions and Writing Styles of Books

Yayoi TANAKA (National Institute for Japanese Language and Linguistics)

Wakako KASHINO (National Institute for Japanese Language and Linguistics)

Sachi KATO (Mejiro University)

### 要旨

本発表は、人手により文体的な観点の印象評定が付与されたテキストについて、修辞機能と脱文脈度の観点によって検討する研究のパイロットスタディである。分析対象は、文体の分類指標 (柏野 2013) が付与された『現代日本語書き言葉均衡コーパス』(BCCWJ) の図書館サブコーパスのうちの4サンプルである。修辞機能分析の分類情報を付与し、修辞機能と脱文脈度を特定した。その結果、4つのサンプルに共通する修辞機能があること、また、それぞれのサンプルに特徴的な修辞機能を確認することができ、アノテーションの有効性を示すことができた。しかしながら、少数のサンプルでは文体情報と修辞機能の対応を確認するまでには至らなかった。今後は、文体指標別に選定した分析対象データを増やし、アノテーションを進める。

### 1. はじめに

本発表は、人手により文体的な観点の印象評定が付与されたテキストの文体特徴について、修辞機能と脱文脈度の観点によって検討する研究のパイロットスタディである。修辞機能と脱文脈度は、修辞機能分析の分類法によって特定される。修辞機能分析は選択体系機能言語理論の英語談話分析手法 Rhetorical Unit Analysis (Cloran 1994, 1999) を日本語に適用した修辞ユニット分析 (佐野 2010, 佐野・小磯 2011) をもとに、日本語文法の枠組みで修正を加えたものである (田中 2022)。修辞は、技巧的なものと捉えられることがあるが、本研究では「修辞機能」を「話し手書き手が発信する際に、言及する対象である事態や事物や人物等を捉え表現する様態を分類し概念化したもの」と定義する。また、文脈や脱文脈化という言葉は、研究分野によって異なる用いられ方をしているが、本研究では、脱文脈度を「発話がコミュニケーションの場「いま・ここ・わたし」にどの程度依存しているか」の程度を表す概念とする。地元のお祭りについて言及するときに、「この神社はにぎやかだね」と述べるのと、「私は先週お祭りに参加した。」と述べるのと、「地元のお祭りが先週開催された。」や「地元のお祭りは毎年9月に行われる。」「祭りは日本の伝統行事だ。」のように述べるのとでは、修辞機能と脱文

---

\* yayoi@ninja.ac.jp

脈度が異なる。これまで、修辞機能と脱文脈度の観点から、児童作文、家族の談話、職場の談話、高齢者グループの談話など(田中ほか 2021, 佐尾ほか 2023, 田中・小磯 2019, 田中 2017, 田中ほか 2022, 2023)の分析から、目的やテーマ、話題内容、状況によって、用いられる修辞機能が異なり、脱文脈度が推移することなどが明らかになっている。

柏野(2013)は文体を分類するための指標として、専門度、客観度、硬度、くだけ度、語りかけ性度の分類指標を提案し、現代日本語書き言葉均衡コーパス(BCCWJ)の図書館サブコーパスのサンプルに人手による印象評定から文体情報を付与している。では、テキストが硬い、あるいは軟らかいと感じたり、専門的、あるいは子供向けだと感じたりするのは、どのような言語的特徴によるものか。例えば浅原ほか(2014, 2015)では語彙の観点から文体との対応を確認している。本研究では、文体に見られる言語特徴の修辞機能と脱文脈度の観点との関連を調査する。

本発表では、様々な文体特徴が現れると考えられる書籍テキストに修辞機能分析の分類情報を付与して修辞機能と脱文脈度を特定し、文体の分類指標との関連を検討する。修辞機能分析では、これまで話し言葉中心に分析を行ってきた。児童作文、高齢者の小作文なども対象にしているが、それ以外の一般的な書き言葉についてはまだ分析が進んでいない。本発表は、書籍という媒体のテキストに初めて修辞機能分析のアノテーションを行う、パイロットスタディの位置付けである。本発表の目的は、文体特徴と修辞機能・脱文脈度の関連を検討することである。課題として、分析対象とした書籍テキストの修辞機能と脱文脈度の特徴を確認する。

以下、第2節で分析データと分析方法について述べ、第3節で分析結果を示し、第4節で考察を述べ、第5節でまとめを述べる。

## 2. 分析データと分析方法

### 2.1 分析データ

本発表の分析対象データは、『現代日本語書き言葉均衡コーパス』の図書館サブコーパスに収録されている書籍サンプル4件である。この4件は、初めて書籍データに修辞機能分析のアノテーションを行うパイロットスタディのため、ランダムに選定した。表1に4件の文体特徴を示す<sup>(1)</sup>。

表1 分析対象サンプルの書名と文体特徴

SampleID	NDC	書名	専門度	客観度	硬度	くだけ度	語りかけ性度
LBn3.00077	3	2010年日本実現	3 一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	3 くだけていない	3 特に語りかけ性はない
LBn3.00084	3	時代が動くとき	3 一般向き	3 どちらかといえば主観的	3 どちらかといえば軟らかい	3 くだけていない	1 とても語りかけ性がある
LBn4.00016	4	天文学者の虫眼鏡	3 一般向き	3 どちらかといえば主観的	3 どちらかといえば軟らかい	3 くだけていない	3 特に語りかけ性はない
LBn4.00041	4	遺伝子組み換え食品を検証する	3 一般向き	2 どちらかといえば客観的	2 どちらかといえば硬い	3 くだけていない	3 特に語りかけ性はない

<sup>(1)</sup> これらの文体特徴情報は柏野(2015)からダウンロードできる。<https://doi.org/10.15084/00003109>

## 2.2 分析方法

修辞機能分析は、Rhetorical Unit Analysis(Cloran 1994, 1999) を日本語に適用した修辞ユニット分析(佐野 2010, 佐野・小磯 2011) を元に、日本語文法の枠組みで修正を加えた分類法である(田中 2022)。分析手順は、次のとおりである。

1. 分析単位(メッセージ)に分割し、分析対象を特定する。
2. 分析対象のメッセージについて発話機能を分類する。
3. 発話機能が「命題」のメッセージについて、時間要素と空間要素を分類する。
4. 発話機能・時間要素・空間要素の組み合わせから、修辞機能と脱文脈度を特定する。

以下に手順の概要を示す。

### 2.2.1 分析単位の分割と分析対象の特定

分析単位であるメッセージは概ね節に相当するが、連体修飾節は独立したメッセージとして扱わない。メッセージは「定型句類」(相槌、挨拶、定型句、節の形でないものなど)、「主節」(単文、及び主節)、「並列」(従属度の低い従属節)、「従属」(従属度の高い従属節)、「引用」(“と思う”などで引用されている部分)に分類する。「主節」「並列」「引用」についてこのあとの分類を行う。

### 2.2.2 発話機能・時間要素・空間要素

メッセージの種類が「主節」「並列」「引用」に分類されたメッセージについて、発話機能・時間要素・空間要素を分類する。表2に示したように、これらの組み合わせから修辞機能と脱文脈指数が特定される。【行動】[1]がもっとも文脈に依存した表現で、【一般化】[14]がもっとも脱文脈度の高い表現である<sup>(2)</sup>。

表2 発話機能・時間要素・空間要素からの修辞機能と脱文脈指数の特定

定義	↑ 高↑空間的距離のレベル↓低						一般化 [14]
状況外	報告 [9]	状況外回想 [10]	予測 [11]		推量 [12]	説明 [13]	
状況内	参加 ↓	行動 [1]	実況 [2]	状況内回想 [3]	状況内予想 [5]	状況内推測 [6]	観測 [8]
空間要素							自己記述 [7]
時間要素	← 低 ← 時間的距離のレベル → 高						
			現在	過去	未来意志的	未来非意志的	仮定
発話機能	提言	命題					

発話機能は「提言」か「命題」に分類する。「提言」は、品物・行為の交換に関する提供・命令で、基本的には同じ時空に存在する相手に働きかけたり、会話者同士の行為にかかわる発話内容が該当し、【行動】[1]と特定される。例えば、同じ時空にいる相手への「お醤油を取って」のような行為や物を要求する場合である。「命題」は、情報を交換する陳述・質問で、「私は桜が大好き」「この桜はピンクが濃いね」「桜はバラ科の植物だ」などが該当する。発話機能が

<sup>(2)</sup> 以下、修辞機能を【】で、脱文脈指数を[]で示す。

「命題」のメッセージについて、このあと時間要素と空間要素を認定する。目の前の相手にグラスを渡して「ほら、お水を飲んで」と言うのはその場で行為を要求しているため「提言」に分類するが、「熱中症予防のためにはお水を飲んで」と伝えるのは、その場での行為を要求しているのではなく、熱中症予防のために水分補給が大切だという情報を提供していると考え「命題」とする。書籍サンプルの分析では、著者と読者の存在する時空を紙面と考え分類することができ、「お水を飲んで」のように行為を要求するメッセージは、例えば引用文や小説などの会話文以外では、現れることは少ないと考えられる。

時間要素は、話者のいる時間「いま」を基準として、メッセージで表現されている出来事がいつ起こったかを示す要素である。基本的にテンスや時間を表す副詞などによって表現され、「習慣・恒久」<sup>(3)</sup>「現在」「過去」「未来意志的」「未来非意志的」「仮定」に分類する。「太郎が花びらを拾っている」は「現在」、「去年見た青森の桜はすごかった」は「過去」、「来週、お花見に行こう」は「未来意志的」、「私は桜が大好き」は嗜好であるため「習慣・恒久」、「桜はバラ科の植物だ」は恒久的と判断し「習慣・恒久」に分類する。書籍サンプルで基準となる「話者のいる時間」については、著者が執筆した時間と考えることができる。例えば、「今、ここにりんごが一つある。」という文が書籍に書いてある場合、読者が読んでいる時間を基準にすると「過去」と考えることになるが、あくまでも著者が執筆した時間を基準とし、さらに、「今」という副詞と「ある」という現在を示す表現から、「現在」に分類する。

空間要素は、話者のいる場所「ここ・わたし」を基準として、メッセージの中心となる語句との空間的距離を示す要素で、主語、主題、述部の主体から判断し、「参加」「状況内」「状況外」「定義」に分類する。「私」「あなた」が主語であれば「参加」、「太郎が花びらを拾っている」は太郎が話者と同じ時空にいると考えられるので「状況内」<sup>(4)</sup>、「去年見た青森の桜はすごかった」「桜は日本中で楽しめる」の桜は話者のいる時空には存在していないと考えられるので「状況外」、「桜はバラ科の植物だ」は桜一般について述べているので「定義」に分類する。

### 2.2.3 修辞機能と脱文脈化度の特定

表2を参照し、発話機能、時間要素、空間要素の組み合わせから、修辞機能と脱文脈度を特定する。

## 3. 分析結果と考察

### 3.1 アノテーション例

表3、表4、表5、表6に、本発表の分析対象4サンプルそれぞれの冒頭部分とアノテーション結果を示す。

2節で述べたように、分析単位であるメッセージごとに発話機能、時間要素、空間要素を認定し、表2を参照して修辞機能と脱文脈化指数を特定する。本研究の分析対象データは書籍サンプルであり、章や節の見出しなども含まれているが、これらの見出しは節の形式でないことが多く、本研究ではメッセージの種類が「(対象外)」に分類される。表3～表6では便宜上、

(3) 「習慣・恒久」には、属性、嗜好、評価も含む。

(4) 「状況内」には、話者の身体や所有物、思想、また、その談話の中で話題になっている事柄も該当すると考える。

表3 LBN3-00077 冒頭 (SF:発話機能 時間:時間要素 空間:空間要素)

	メッセージ	SF	時間	空間	修辞機能 [指数]
a	【政策提言】	(対象外)			
b	私が描く国家像	(対象外)			
c	【政策提言】一序	(対象外)			
d	今まさに、二十世紀が終わろうとしている。	命題	現在	状況外	報告 [09]
e	西暦二千一年は新しいミレニアム(千年紀)の始まりでもある。	命題	習慣・恒久	状況外	説明 [13]
f	後の審判の後、訪れる「至福の千年紀」は、キリスト教徒にとって幻想ではない。	命題	習慣・恒久	状況外	説明 [13]
g	では、私たちにとって、現実を訪れる二十一世紀、近未来の二千年は至福の時代となるだろうか。	命題	未来非意志的	状況外	予測 [11]
h	もともと日本には世界に通用する非常に高い精神性を持った文化があった。	命題	過去	状況外	状況外回想 [03]

表4 LBN3-00084 冒頭 (SF:発話機能 時間:時間要素 空間:空間要素)

	メッセージ	SF	時間	空間	修辞機能 [指数]
a	NPO 法の特徴	(対象外)			
b	NPO 法の特徴を申し上げておきます。	命題	現在	参加	実況 [02]
c	簡単に言うと	(従属)			
d	七点あります。	命題	習慣・恒久	状況外	説明 [13]
e	(1) 議員立法	(対象外)			
f	私は、この法律は社会の根幹に関わる法律だと	命題	習慣・恒久	状況外	説明 [13]
g	思っています。				

発話機能(表内のSF)欄に(対象外)と示している。表3のdでは、「今」と「終わろうとしている」から時間要素が「現在」、「二十世紀が」から空間要素が「状況外」と分類する。eでは、「西暦二千一年は」が空間要素の対象で、執筆しているその場にはない物なので「状況外」、二千一年がミレニアムの始まりであることは変わらないことなので時間要素は「習慣・恒久」と分類する。

表4のcは、従属度の高い従属節のため、メッセージの種類が「従属」である。本表では便宜上、発話機能の列に記入している。「従属」の場合、発話機能以降の分類は不要である。また、gの「思っています」は「この法律は社会の根幹に関わる法律だ」という命題に対して断言しないために付与したものと考え、「思っています」の部分は分類しない。ただし、「思っています」や「思います」「思う」などをすべて分類しないのではなく、例えば、「私は彼の気遣いをありがたく思っています。」のように引用節を伴わない場合には時間要素として分類する。

表5のcの空間要素の対象は「二つの種族」で、その場に存在するものではないため、「状況外」に分類する。また、dでは、空間要素となる主語や主題が明示されていないが、cの主語で

表 5 LBN4-00016 冒頭 (SF:発話機能 時間:時間要素 空間:空間要素)

	メッセージ	SF	時間	空間	修辞機能 [指数]
a	二 眠れる猫、力学の法則を学ぶ	(対象外)			
b	猫好き族と犬好き族	(対象外)			
c	世に、猫好き族と犬好き族という二つの種族があるらしい。	命題	習慣・恒久	状況外	説明 [13]
d	世の移ろいを怜悧に眺めている孤高の猫が好きな種族と、飼い主一家の一員と思い込んで気楽に戯れてくれる犬が好きな種族である。	命題	習慣・恒久	状況外	説明 [13]

表 6 LBN4-00041 冒頭 (SF:発話機能 時間:時間要素 空間:空間要素)

	メッセージ	SF	時間	空間	修辞機能 [指数]
a	第九章 日本 義務表示の衝撃	(対象外)			
b	予想を超える波紋	(対象外)			
c	八月の末、カナダ、オンタリオ州で大豆畑に向かう車のなかだった。	命題	過去	状況外	状況外回想 [10]
d	案内役の三菱商事カナダの社員がつぶやいた。	命題	過去	状況外	状況外回想 [10]
e	「キリンビールが、使っているコーンスターチを遺伝子組み換えフリーのものに変えようというニュースが生産者に衝撃を与えていますよ。	命題	現在	状況外	報告 [09]
f	これだけの大企業が方向転換しようとしていることが驚きなんです」	命題	習慣恒久	状況外	説明 [13]

ある「二つの種族」が d で主題になっていると推測できるので、同様に「状況外」に分類する。

表 6 の c では、主語や主題が明示されていないが、後続する文から「案内役の社員がつぶやいたのは」のような主題が推測できるため、「状況外」に分類した。e では、読点の位置から、「キリンビールが」が主語に思われるが、述部は「与えていますよ」で、その主体は「ニュースが」であるため、空間要素の対象になるのは「ニュースが」である。

### 3.2 修辞機能の出現と考察

サンプルごとの修辞機能の出現頻度と割合を表 7 と図 1 に示す。いずれのサンプルでも、修辞機能【説明】脱文脈化指数 [13] が多く用いられている。これは、サンプルの NDC が 3 (社会科学) と 4 (自然科学) であることと関係があると考えられる。

また、サンプルと修辞機能との対応関係を調べるために、対応分析を行った<sup>(5)</sup>。分析には R の ca 関数を用いた。結果を図 2 に示す。

<sup>(5)</sup> 出現頻度が 10 件以下の修辞機能は除外した。



表 7 サンプルごとの修辞機能の出現頻度

	01 行動	02 実況	03 状況 内 回想	04 計画	05 状況 内 予想	06 状況 内 推測	07 自己 記述	08 観測	09 報告	10 状況 外 回想	11 予測	12 推量	13 説明	14 一般 化
LBn3_00077	0	4	1	17	1	1	9	3	6	16	14	14	62	1
LBn3_00084	0	0	12	12	0	0	3	1	6	30	20	5	120	0
LBn4_00016	0	1	8	2	0	1	15	13	3	42	0	8	136	4
LBn4_00041	0	7	12	6	3	1	15	15	26	72	20	2	168	5

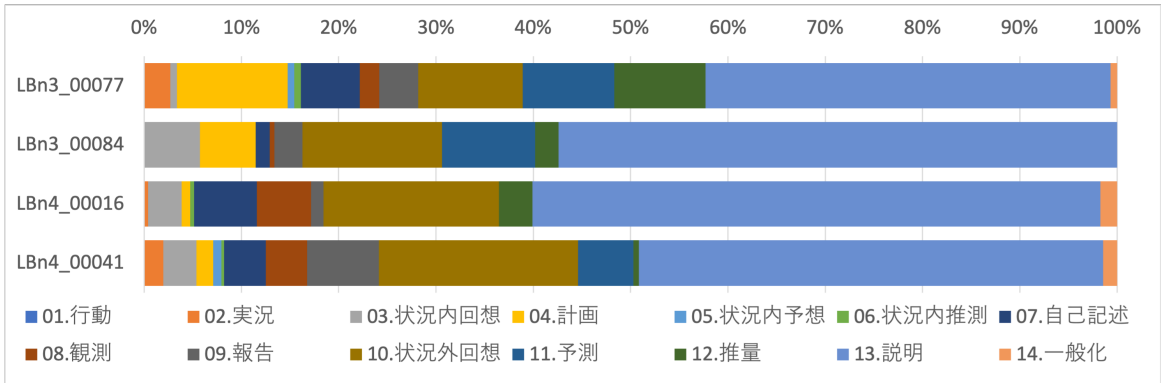


図 1 サンプルごとの修辞機能の出現割合

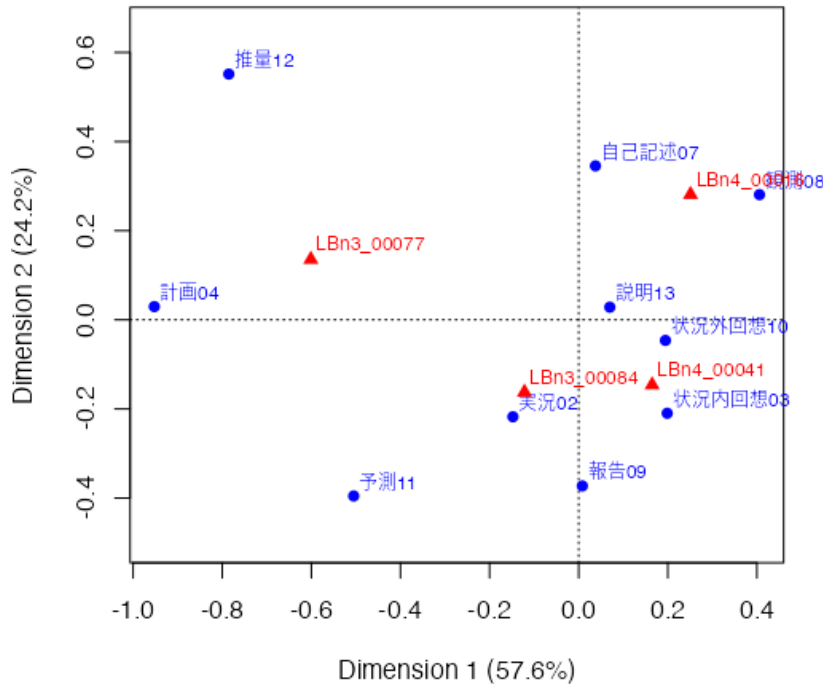


図 2 サンプルと修辞機能の対応分析結果

対応分析の結果でも【説明】[13]が4つのサンプルの中央に位置しており、いずれのサンプルでも用いられていることがわかる。また、対応分析の結果から、サンプルごとに次のような修辞機能との関連が見られた。

LBn3\_00077では【計画】[04]【推量】[12]、LBn3\_00084で【実況】[02]【報告】[09]【予測】[11]、LBn4\_00016で【観測】[08]【自己記述】[07]、LBn4\_00041で【状況内回想】[03]【状況外回想】[10]が特徴的と考えられる。LBn3\_00077のタイトルは「2010年日本実現」だが、出版年が1999年のため、未来のことを述べていることがうかがえ、そのため、【計画】[04]や【推量】[12]が特徴的な修辞機能として現れていると考えられる。LBn3\_00084は、「時代が動くとき」というタイトルの書籍である。【実況】[02]【報告】[09]ともに、時間要素が「現在」の修辞機能のため、執筆時のことについて言及している内容であることがうかがえる。LBn4\_00016のタイトルは「天文学者の虫眼鏡」である。【観測】[08]【自己記述】[07]ともに、時間要素が「習慣・恒久」の修辞機能で、空間要素が「状況内」か「参加」かの違いである。著者自身のことや、著者の感情、思考、所持品、また、著者がいる空間にあるものについて、性質や習慣などが言及されていることがうかがえる。LBn4\_00041のタイトルは「遺伝子組み換え食品を検証する」である。表4のアノテーション例にあげたメッセージで見られたように、回想場面を描写していることから、【状況内回想】[03]【状況外回想】[10]が特徴的になったことがうかがえる。

以上のことから、書籍データについても、修辞機能分析の分類法によって、修辞機能の観点からその書籍の特徴を捉えることが可能であると考えられる。

しかし、本研究の目的である、文体特徴との関わりについては、本分析では明らかにすることはできなかった。今回の4サンプルのうち、客観的なものが1つあったが、ほかの3つとその1つとでは、違いは見られなかった。また、硬度についても、語りかけ性度についても、その差異と対応するような修辞機能の差異はわからなかった。サンプル数が少ないためとも考えられるが、今回専門度が4サンプルとも同じであったことから、専門度に修辞機能の差異が見られることも考えられる。今後は、サンプル数を増やすとともに、典型的なサンプルを選択して分析することを検討する必要がある。

#### 4. おわりに

本発表は、人手による印象評定から付与された文体特徴について、修辞機能と脱文脈度の観点によって検討する研究のパイロットスタディである。文体特徴と修辞機能・脱文脈度の関連を検討することを目的とし、課題として、異なる文体特徴をもつサンプルの修辞機能と脱文脈度の特徴を確認した。『現代日本語書き言葉均衡コーパス』図書館サブコーパスから4つのサンプルに修辞機能分析の分類法でアノテーションした。分析の結果、4つのサンプルに共通する修辞機能があることがわかり、また、それぞれのサンプルに特徴的な修辞機能が確認できた。しかし、文体特徴と修辞機能の関わりについては、明らかにできなかった。今後の課題として、分析サンプルを増やすことが挙げられるが、その際に、スタイル（丁寧体か普通体か）などを揃える、文体特徴の典型的なものを対象とするなど、分析サンプルを検討する必要があると考えられる。

#### 謝 辞

本研究はJSPS 科研費 JP19K00588、JP23K00569 によるものです。

## 文 献

- C. Cloran (1994). "Rhetorical units and decontextualisation: An enquiry into some relations of context, meaning and grammar." Unpublished doctoral dissertation, University of Nottingham Nottingham.
- C. Cloran (1999). "Contexts for learning." Frances C (Ed.), *Pedagogy and the Shaping of Consciousness: Linguistic and Social Processes*. London: Continuum International Publishing. pp. 31–65.
- 浅原正幸・加藤祥・立花幸子・柏野和佳子 (2014). 「文体指標と語彙の対応分析」 第6回コーパス日本語学ワークショップ予稿集, pp. 11–20.
- 浅原正幸・加藤祥・立花幸子・柏野和佳子 (2015). 「文体指標と語彙系列の対応分析」 第7回コーパス日本語学ワークショップ予稿集, pp. 7–16.
- 柏野和佳子 (2013). 「〈共同研究プロジェクト紹介〉萌芽・発掘型：テキストの多様性を捉える分類指標の策定 書籍サンプルの文体を分類する」 国語研プロジェクトレビュー, NINJAL Project Review, 4:1, pp. 43–53.
- 柏野和佳子 (2015). 『『BCCWJ 図書館サブコーパスの文体情報』(2015年公開第1版)』.  
<https://doi.org/10.15084/00003109>
- 佐尾ちとせ・宮城信・田中弥生 (2023). 「修辞機能分析を活用した作文指導」 日本語習熟論研究:1, pp. 140–158.
- 佐野大樹 (2010). 『日本語における修辞ユニット分析の方法と手順 ver.0.1.1：選択体系機能言語理論(システミック理論)における談話分析(修辞機能編)』, <https://researchmap.jp/kotonoha/>資料公開/.
- 佐野大樹・小磯花絵 (2011). 「現代日本語書き言葉における修辞ユニット分析の適用性の検証-「書き言葉らしさ 話し言葉らしさ」と脱文脈化言語 文脈化言語の関係-」 機能言語学研究, 6, pp. 59–81.
- 田中弥生 (2017). 「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」 言語資源活用ワークショップ発表論文集, 1, pp. 69–78.
- 田中弥生 (2022). 「修辞機能と脱文脈化の観点からの日本語談話分析」 博士論文(未公刊), 東京大学大学院総合文化研究科.
- 田中弥生・小磯花絵 (2019). 「家庭での幼児の発話の修辞機能：脱文脈化の観点からの検討」 言語資源活用ワークショップ発表論文集:4, pp. 106–118.
- 田中弥生・小磯花絵・大武美保子 (2022). 「脱文脈化の観点から見た共想法に基づく高齢者談話の分析」 国立国語研究所論集:22, pp. 137–155.
- 田中弥生・小磯花絵・大武美保子 (2023). 「共想法による話し言葉・書き言葉における修辞機能の特徴-テーマとの関係に着目して-」 言語処理学会第29回年次大会発表論文集, pp. 1356–1360.
- 田中弥生・佐尾ちとせ・宮城信 (2021). 「児童作文の評価に向けた脱文脈化観点からの検討」 言語処理学会 第27回年次大会 発表論文集, pp. 750–755.

## ゲームコーパスの設計方針と構築方法

麻 子軒（関西大学）<sup>†</sup>

### Approaches to Design and Construction of a Game Corpus

Tzu-Hsuan Ma (Kansai University)

#### 要旨

ゲームコーパスを体系的に構築するにあたり、その目的を明確にした上で、それに合わせた形で発売年代やジャンルの諸観点から代表的なゲームを選定する必要がある。また、ゲームの場合はプレイヤーの操作によって表示されるテキストの内容と量が変化するため、テキストの認定基準をはじめとする構築方法は書籍をベースとしたコーパスと異なるように考えられる。本稿ではゲームにおける言語的特徴の解明と日本語教育への応用との二つの目的を意識し、現在筆者が構築しているゲームコーパスの設計方針を発売年代、ジャンル、発売本数、世界観の4つの観点で述べた。なお、収録環境の整備、作業人員の募集、プレイ動画と画像の収録、文字化作業などの構築方法、及び構築時の問題点についても言及した。

#### 1. 経緯

麻（2022）では、ゲームコーパスの構築理由とそれを用いた研究事例を提示された。しかし、同研究で挙げられた事例は単発的なもので一貫性に欠け、また、複数のゲームを体系的にコーパスに取り入れる構想にはまだ至っていない。ゲームコーパスを体系的に構築するにあたり、その目的を明確にした上で、それに合わせた形で発売年代やジャンルなどの諸観点からゲームを選定する必要がある。また、ゲームの場合はプレイヤーの操作によって表示されるテキストの内容と量が変化するため、テキストの認定基準をはじめとする構築方法は、書籍をベースとしたコーパスと異なるように考えられる。以上を踏まえ、本稿では、筆者が現在構築しているゲームコーパス（以下、本コーパス）の設計方針、及びその構築方法と問題点について述べる。

これより、2節では設計方針、具体的には収録ゲームの発売年代やジャンルなどの選定基準を述べ、3節では構築方法、具体的にはテキストの認定基準や電子化の方法などを述べ、最後に4節で構築時の問題点を挙げる。

#### 2. 設計方針

コーパスを構築するにあたり、どのような目的で、どのようなデータが必要なのかを、まず明らかにしなければならない。なぜなら、調査目的によって、選定されるサンプルやその量、ないし必要なアノテーション情報も異なるからである。現在筆者が関心を持っているのは、ゲームにおける言語的特徴の解明、及びゲームを日本語教育に応用する可能性の2点であるため、以下ではこれらの目的を意識し、発売年代、ジャンル、発売本数、世界観の4つの観点からゲームの選定基準を述べる。

---

<sup>†</sup> kenji.ma@kansai-u.ac.jp

## 2.1 発売年代

ゲームにおける言語的特徴も時代によって変化する可能性がある。ただ、ゲームの歴史は書籍と異なり、日本で広く世に知られているのは1990年代以降のことである。2023年現在でも30年ほどの歴史しかないため、本格的な通時的調査には適していない。とはいえ、技術の進歩にともない、ゲームの表現手段も進化しており、それによる言語面での変化は確実に存在すると思われる。特にここ数年、記憶媒体の容量向上により、漢字表記の多用や、キャラクターの音声付与ができるようになったため、ゲームの描写手法にも大いに影響を与えたと予想される。

本コーパスでは、据置ゲーム機の全盛期（1990～2000年、以下前期）と、直近十年間（2010～2022年、以下後期）の代表的なゲームを選定する予定である。前期は容量制限で2Dかつ音声なしのゲームがほとんどで、後期は3Dかつ音声ありのゲームが主流となっている。これにより、技術の進歩による言語的变化が観察できると思われる。発売のプラットフォームは、ひとまず据置ゲーム機に限定する。携帯ゲーム機は、後述する収録の技術的な理由により、現段階では対象としない。

## 2.2 ジャンル

また、ゲームはその遊び方によって、十数種類のジャンルが存在する。ゲームの言語的特徴を調査するには、なるべく全ジャンルを網羅的に取り入れるのが理想的だが、限定された時間でそれが困難であるため、主流的、かつストーリー性のあるジャンルを優先的に選定する。具体的に、アクションゲーム（以下ACT）、ロールプレイングゲーム（以下RPG）、シミュレーションゲーム（以下SLG）、アドベンチャーゲーム（以下AVG）の4ジャンルに限定する。それぞれのジャンルの説明と代表例は、表1に示す。

表1 本コーパスで収録するゲームのジャンル

ジャンル	説明	代表例
アクションゲーム	プレイヤーがキャラクターの行動を直接操作し、ストーリーを進めていくゲーム	スーパーマリオ、ロックマンX
ロールプレイングゲーム	プレイヤーがゲーム内のキャラクターとなり、情報を収集するなどして世界を探索し、ストーリーを進めていくゲーム	ドラゴンクエスト、ファイナルファンタジー
シミュレーションゲーム	プレイヤーが指揮者となり、戦略を考慮しながらキャラクターに行動を指示し、ストーリーを進めていくゲーム	ファイアーエムブレム、スーパーロボット大戦
アドベンチャーゲーム	プレイヤーがテキストで提示された状況に対し、キャラクターの行動を選択肢で決定し、ストーリーを進めていくゲーム	ときめきメモリアル、逆転裁判

ただし、「ゼルダの伝説」のように、ACTとRPGの性質が両方揃っており、分類が複数のジャンルにまたがるゲームもある。なお、言語研究であるため、ストーリー性が薄く、最初からテキストがあまりないと分かるジャンル、例えば格闘ゲームは除外している。

### 2.3 世界観

書籍同様、ゲーム内に現れた言語的表現は、何を描写するかによって大きく左右される。あまり良い例えではないが、書籍の場合、哲学・歴史・芸術・文学などの分類があり、それぞれに現れた言語的特徴も異なる。ゲームの場合は、書籍と同じ分類では難しいが、ゲーム内の世界観である程度分けることができる。現段階は暫定的に、中世王道風、近現代風、未来SF風の3つに分ける。本コーパスは、それぞれの世界観の作品をなるべくバランスよく収録したいのだが、ゲームの性質上、結果的に中世王道風が多めに入っていることになった。なお、「スターオーシャン」のように、宇宙の探索ができる時代に未発達惑星での冒険ができるなど、分類が難しいゲームもある。表2はゲームの世界観をまとめたものである。

表2 本コーパスで収録するゲームの世界観

世界観	説明	代表例
中世王道風	剣と魔法で作られた中世を時代背景とし、現実的な世界とかけ離れている幻想的な世界	ドラゴンクエスト、テイルズオブファンタジア
近現代風	物理的法則や施設が現実に近い世界を背景とし、実在する場所をベースとしているゲームもある	龍が如く、ペルソナ
未来SF風	発達した技術によって構築された世界を背景とし、主に機械やロボットが登場している	ロックマンX、スーパーロボット大戦

### 2.4 発売本数

本コーパスは均衡コーパスではないため、母集団のすべての性格を反映させるよりも、代表性のあるものを取り入れるのが目的である。そのため、収録するゲームも人気のあるものを想定している。また、ゲームの日本語教育への応用も筆者が関心のあることであるため、多くの人が興味を持つゲームを考察して得られた知見のほうに教育現場で役に立つと思われる。知名度を定義するには非常に困難であるが、原則的に日本国内での発売本数が10万本以上の作品に限定する。ただし、ゲームジャンルによって、どうしても作品数が揃わない場合は、例外的に発売本数が10万本未満の作品を対象とすることがある。

以上の4観点で選定する予定のゲームと現在の進捗状況は、表3に示す。

表3 選定される予定のゲームと進捗状況

	ジャンル	ゲーム名	発売年代	世界観	状態
前期	ACT	スーパーマリオワールド	1990	中世王道風	未着手
		ゼルダの伝説 神々のトライフォース	1991	中世王道風	未着手
		ロックマン X3	1995	未来SF風	済
	RPG	ドラゴンクエスト3 (リメイク)	1996	中世王道風	済
		クロノ・トリガー	1995	中世王道風	作成中
		ファイナルファンタジー7	1997	中世王道風	作成中
		テイルズオブファンタジア	1995	中世王道風	作成中
		マザー2	1994	近現代風	未着手
		スターオーシャン1	1996	未来SF風	未着手
		SLG	ファイアーエムブレム 紋章の謎	1994	中世王道風
		第4次スーパーロボット大戦	1995	未来SF風	未着手
AVG	ときめきメモリアル1	1994	近現代風	未着手	

後期	ACT	モンスターハンター：ワールド	2018	中世王道風	未着手
		ゼルダの伝説 ブレスオブワイルド	2017	中世王道風	作成中
		龍が如く7	2020	近現代風	未着手
	RPG	ドラゴンクエスト11	2017	中世王道風	未着手
		キングダムハーツ3	2020	中世王道風	未着手
		オクトパストラベラー1	2018	中世王道風	作成中
		ポケットモンスター バイオレット	2022	近現代風	未着手
		ペルソナ5	2016	近現代風	未着手
		ゼノブレイド2	2017	未来SF風	未着手
	SLG	ファイアーエムブレム 風花雪月	2019	中世王道風	未着手
		FRONT MISSION 1ST (リメイク)	2022	未来SF風	未着手
AVG	大逆転裁判2	2017	近現代風	未着手	

前期と後期のゲームがなるべく同数になるように選定した。ジャンルと世界観に関しては、これまで発売されたゲームの全体的な内訳を見て、偏りが出たのはある程度妥協しなければならないと思われる。

### 3. 構築方法

本節では、どのようにゲーム内のテキストをコーパスにするかの手順を説明する。直接ゲーム媒体の内部データにアクセスして、文字を抽出する方法が最も作業の手間が省けるが、基本的にゲームの文字コードがPCと異なり暗号化されており、一々解読するのは現実的ではない。また、許可なしに内部データにアクセスすると、法律に違反するおそれもある。そのため、代わりに地味な方法を採用するしかない。具体的には、ゲームをプレイし、テキスト情報を画像に保存しておき、後に一気に文字化する方法である。そのためには、収録環境の整備、作業人員の募集、プレイ動画収録と画像キャプチャー、文字化作業、最終確認、以上の5つのステップが必要となる。

#### 3.1 収録環境の整備

まず、必要なゲーム機器とゲームソフトを用意する必要がある。前期のゲームをプレイするための機器であるスーパーファミコンとプレイステーション2は20年以上前に発売されたもので、すでに生産終了となっているため、中古品を購入した。後期のゲームをプレイするためのニンテンドースイッチとプレイステーション4は新品で購入した。ゲームソフトに関しても同様で、基本的に前期のゲームは中古で購入する以外方法はない。

また、コーパスを作成するため、PCも必要である。文字化するだけであれば、スペックの低いPCでも問題ないが、動画を収録するために、ある程度高性能のPCが必要である。収録に使用したのはASUS Zenbook 14 OLED UX3402ZA (CPU インテル® Core™ i7-1260P、メモリ 16G、ストレージ 512GB) である。なお、録画するために、ゲームの画面をPC経由で表示させる必要がある。外付けキャプチャーボードを繋げば、PCをテレビ代わりにできるため、録画ソフトで画面の録画と画像キャプチャーが可能である。使用したキャプチャーボードは、AVerMedia Live Gamer EXTREME 2 GC550 PLUS (HDMI 端子対応) である。前期のゲーム機にはHDMI端子が付いていないため、AV端子をHDMI端子に変換するコンバーターも必要である。図1と図2は、HDMI端子に対応するゲーム機とそうでないゲーム機の収録環境をイメージしたものである。

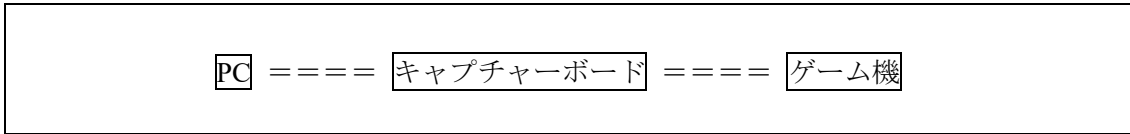


図1 ゲームプレイ動画の収録環境 (HDMI 端子対応ゲーム機の場合)

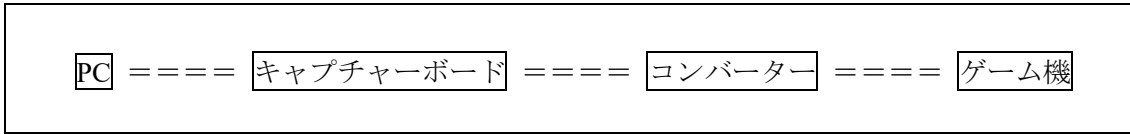


図2 ゲームプレイ動画の収録環境 (HDMI 端子非対応ゲーム機の場合)

使用した録画ソフトは **Bandicam** である。録画の画質設定は **FPS30**、画質 **640\*480** (1 時間 **600M** 程度) にした。画像の画質設定は、ゲーム機もともとの解像度に依存するが、ニンテンドースイッチとプレイステーション 4 の場合は **1920\*1080** になる (ゲームのシーンにもよるが、平均して 1 枚 **600~800K** 程度)。これで収録環境の整備は完了である。

### 3.2 作業人員の募集

本コーパスを構築するためにメインとなる作業は、ゲームのプレイ動画収録と画像キャプチャー、及び文字化 (アノテーション情報付与とデータ整形を含む) である。作業の効率性を考慮し、最も時間がかかるプレイ動画収録と文字化作業を中心に、作業人員に依頼することにした。作業者はゲーム経験者であり、かつ日本語の入力ができることが条件となる。実際募集の際に出した条件は、①ゲーム経験者 (特に **RPG**、**SLG**)、②PC での日本語入力、③EXCEL の基本操作、④国籍不問 (ただし非母語話者の場合は日本語能力試験 **N1** 必須) である。現在 2 名体制で作業を進めている。

### 3.3 動画収録と画像キャプチャー

このステップの作業は、プレイしながら画面のテキスト情報を保存することが目的である。必要な情報はテキスト情報だけであるため、それ以外の画面情報は不要である。この前提であれば、画像キャプチャーだけでもよいのだが、それでも同時に動画を収録させたのは、画像のキャプチャーに失敗した際の保険と、後に文字化する際に文脈 (場面) を確認するためである。時間を節約するため、作業者が当該ゲームの経験者であることが望ましいが、やむを得ずプレイしたことがないゲームを収録・文字化させる場合もある。

テキスト認定は、方針未定の部分があるが、現段階では作業者に下の指示を出している。

- (1) テキストが画面上に出てきたらキャプチャーする。
- (2) できればすべてのルートのすべての文字をテキストする。(攻略サイトを参照)
- (3) 同じ人物には少なくとも 2 回話しかける。(内容が変わる可能性があるため)
- (4) ストーリーが進むと、会話が変わる可能性のある人物に再度話しかける。
- (5) 「はい」「いいえ」のような選択肢は、どちらも選んで、収録する。
- (6) ミッションがあれば、失敗するバージョンも収録する。
- (7) 特定の行動を取る (特定の人物に話しかける、または特定の場所に移動する) とフラグが立ち、他の人物と話せなくなったり、会話内容が変わったりすることがあるため、話しかける順番に注意する。



- (8) 新しい町に到着したら、行動をする前に、まずセーブを別データに保存する。(やり直しに備える)

4節で詳述するが、ゲームはプレイヤーの操作によって現れるテキストも異なるため、どこまでこだわるかによってテキスト量と作業の所要時間が変わる。現在は、プレイ動画はなるべくすべての要素を網羅できるように収録してもらい、あとで筆者が文字化データを最終確認の際に必要なものを選別する方法を採用している。

### 3.4 文字化作業

文字化作業に関して、OCR という方法も試したが、ゲームのフォントは書籍と異なり独自のものを使用するものがほとんどで、認識精度が著しく悪かったため、手作業で入力する方法を採用することにした。

文字化作業に使用する EXCEL の記録フォーマットは図 3 の通りである。「シーン」欄は今後、やり取りのある会話を分析するために、筆者が後に付与する情報である。なお、「シーン」欄以外は麻 (2022) で説明されたため、そちらを参照されたい。

番号	画像	シーン	区分	スプライト	名前	場所	内容
2	2023-06-10 22-08-11-798.jpg						
3	2023-06-10 22-08-18-464.jpg						
4	2023-06-10 22-08-26-326.jpg						
5	2023-06-10 22-08-28-027.jpg						
6	2023-06-10 22-08-29-759.jpg						

図 3 文字化作業の EXCEL 記録フォーマット

具体的な手順は、キャプチャーした画像のファイル名をプログラムで「画像」の列に書き込ませた後、作業者に「区分」「スプライト」「名前」「内容」の列を入力させる。「番号」の列は、次の最終確認のステップで、自動で付与する通し番号である。実際に作業者に出示した指示は以下の通りである。

- (1) 一枚の画像を、一行のレコードに入れる。入力が必要な欄は「区分」「スプライト」「名前」「内容」の 4 か所である。
- (2) 「区分」のセルについて、キャラクターの発話は「セリフ」、システムメッセージは「メッセージ」、魔法・アイテム欄の選択候補は「メニュー」と入れる。
- (3) 区分が「セリフ」の場合、発話者は「名前」、発話内容は「内容」のセルに入力する。発話者名が表示されない場合は「\*」と入れる。
- (4) 原則、表記も含め、元テキストを忠実に再現する（会話の最初の「」は入力しない）。ただし、ひらがな・カタカナ・英数字・記号はすべて「全角」で入れる。

(5) 改行は「半角スペース」に変えて入力する。

### 3.5 最終確認

入力が終了したファイルを、筆者が最終確認する。具体的には、①入力ミスの確認、②「シーン」「場所」など必要なアノテーションの追加、③不要な行の削除、以上の3つの作業を行う。

すべての手順の所要時間に関して、最も時間がかかる RPG というジャンルは、ゲームによってクリア時間が異なるが、平均的に30~40時間で1本クリアできる。ACTの場合は攻略方法さえ分かれば3時間程度でクリアできるゲームもある。文字化作業は約プレイ時間の2~3倍時間がかかる。延べ週12時間体制で作成させているため、理論的に文字化作業も含めて2か月にRPGが1本コーパス化できる計算になっている。

## 4. 今後の課題

本節では、文字化作業を行う際に、実際に遭遇した問題点を述べる。

### 4.1 対象とするテキストの認定

書籍は、文字が物理的に紙に印刷されているため、意図的に読み飛ばさない限り、すべてのテキストを目にすることができる。一方、ゲームの場合はストーリーの分岐や任意のサブイベントがあるため、プレイヤーの操作次第で、表示されないテキストもある。ここで、ゲーム内にあるテキストをすべて収録する方法（以下、やり込み方式）と、一部のルートのみ収録する方法（以下、一周クリア方式）との2つの選択肢がある。2つの方法で同ゲームを収録するイメージを図4と図5に示す。

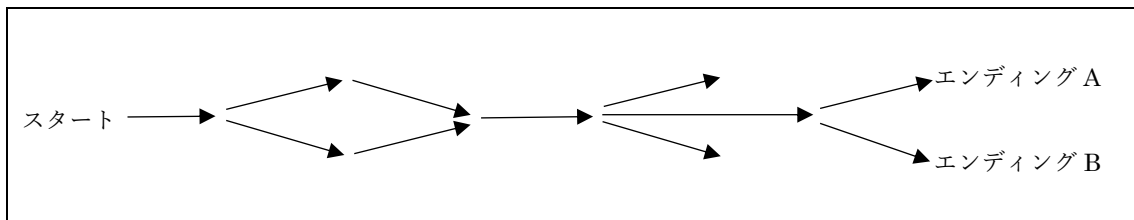


図4 やり込み方式（実線すべて収録）

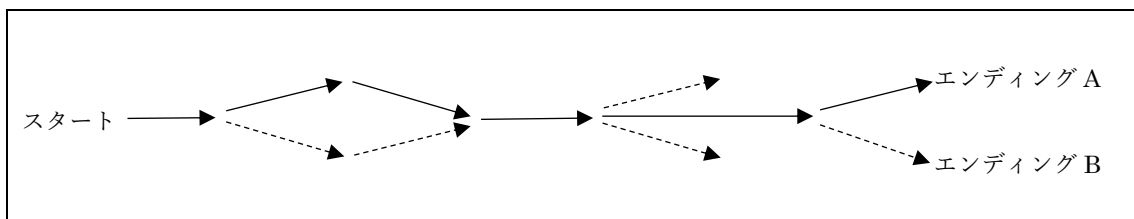


図5 一周クリア方式（実線のみ収録、点線は収録しない）

後期のRPGはやり込み方式にこだわると100時間以上かかるため、本コーパスでは基本的に一周クリア方式を採用する。ただ、前期のゲームは容易にやり込み方式を達成できるため、ゲームによってやり込み方式と一周クリア方式に分けて収録することも考えられる。ただ、一周クリア方式で収録した場合は、客観性と再現性を担保するために、選択肢の選び方やサブイベントの定義について厳密に規定する必要がある。

#### 4.2 キャラクターの特徴の記述

ゲームでは、主人公などの重要人物は名前が付いている一方、一度しか登場しないキャラクターは名前がないことが多い。コーパスでは、後に発話者を特定できるように、当該人物の特徴を記述する（図3の「スプライト」の欄）必要があるが、前期のゲームは解像度が低く、キャラクターの特徴を記述することが困難な場合がある。記述がなくてもできる研究であれば問題ないが、役割語のような社会言語学の研究では、話者の属性が重要であるため、研究に使うには限界がある。

#### 4.3 テキストの分類

現在、テキストの分類として、キャラクターによる発話の「セリフ」と、ナレーションかシステム説明の「メッセージ」と、魔法・道具欄の選択候補の「メニュー」の3種類を定義したが、これらの分類に当てはまらないテキストがある。例えば、キャラクターが心の中で思っている内容がテキストとして表示された場合は現在「セリフ」に入れているが、厳密に言うと「セリフ」ではないため、別項目を立てることを検討している。

もう一つの例は、セリフかどうか判定しにくいパターンである。例えば、キャラクターが本棚を調べたときに、書籍の内容を読み上げることがあるが、この場合、表示されたテキストは確かにそのキャラクターのセリフではあるものの、本質的には書籍の内容であるため、別扱いにする必要があるように思われる。また、特定のキャラクターが別のキャラクターに憑依した場合の発話も、これと類似したパターンである。

#### 4.4 テキスト以外の情報の記録

本コーパスの目的はテキストを収録することであるため、テキスト以外の情報、例えばフォントのサイズ、フォントの色、キャラクターの表情などは収録できていない。この点は書籍のコーパスも同様だが、ゲームではフォントの大きさや色はパラ言語的に使われることがあるため、研究する価値はあると思われる。

#### 4.5 アノテーション情報欄の設定

最後に、ゲームによって必要なアノテーション情報が異なる点が挙げられる。例えば、特に前期のRPGはスプライトの情報が重要であるが、ACTの場合、それがあまり意味を成さないため、記録フォーマットに当該欄をあえて設定する必要性が薄い。如何にすべてのゲームに適用する記録フォーマットを設計するかが重要な課題となる。

以上挙げた5つの点は、一部未解決のものもあるため、今後の課題としたい。

### 謝 辞

本研究は JSPS 科研費（若手研究）「テレビゲームの日本語教育における可能性の探索とテレビゲームコーパスの構築（課題番号：23K12220）」の助成を受けている。

### 文 献

麻子軒（2022）「テレビゲームコーパスの構築とその利活用」『言語資源ワークショップ発表論文集』2022, pp.117-126.

## 読書感想文の分析と指導法 —書き出しと結びに注目して—

加藤 恵梨 (愛知教育大学)

角谷 昌範 (愛知教育大学附属岡崎小学校)

### Analysis of Book Report and its Application in Class —Focusing on the Opening and Closing Lines—

Eri Kato (Aichi University of Education)

Masanori Sumiya (Okazaki Primary School Affiliated to Aichi University of Education)

#### 要旨

本研究は、愛知教育大学附属岡崎小学校で発行している児童文集「ひばり」(2019年～2023年の5年分)に収録されている読書感想文を調査資料とし、それらがどのような書き出しで文章がはじまり、どのような結びで文章が終結しているのかを明らかにすることを目的とする。また、低学年・中学年・高学年によって書き出しと結びに違いがあるのかについても分析する。さらに、読書感想文の書き方について指導する際、教師はどのような点に注意して指導したら良いかについて提案する。

#### 1. はじめに

鈴木(2016: 75-76)が指摘するように、『学習指導要領解説・国語編』における「書くこと」の言語活動例をみると、小・中学校を通してみても、感想文ないし読書感想文はあげられておらず、現今の「書くこと」指導において、読書感想文に関する指導は極めて少ない。また、小学校における児童の読書感想文に対する意識や本の選び方、読書感想文の書き方等に関するアンケートを作成し、3年生から6年生を対象に調査を行った天野・鈴木(2019: 24-25)は<sup>1</sup>、その結果から、①読書そのものは好きであるが、読書感想文を書くことに抵抗感を抱いている児童が多いこと、さらにその理由として②読書感想文の書き方に不安を抱えていること、という二つの課題を指摘し、そのことは、読書感想文が、通常の授業の中で扱われることが稀であり、夏休みの宿題等で課されることが多いため、書き方についての指導が不十分であることによって生じていると考えている。

先行研究の指摘から、児童にとって読書感想文を書くことは容易ではなく、児童は書き方に不安を抱えていることがわかる。太田(1965: 10)が指摘するように、特に子どもが作文を書く時に最初にぶつかる大きな壁が書き出しである。また、田近・井上・中村(編)(2018: 133)にも、「『書き出し』をどう始めればよいかわからないという子どもがいる。徒競走の時でも、スタートの時の失敗は後々までなかなか取り返しがつかないものである。」という記述があるように、書き出しは児童にとって難しいものである。しかし、書き出しに書く文のパターンがいくつか示されていれば、それらの中から自身の書きたいことにあわせて児童が選び、文章を書き出すことができると考えられる。

そこで本研究では、児童が書いた読書感想文の書き出しと結びにどのようなことが書かれているかを学齢別に調査し、それらを類型化することを目的とする。また、調査結果をもとに、教師が読書感想文の書き出しについて指導する際、どのような点に注意して指導した

<sup>1</sup> アンケート調査は、2018年10月に読書感想文を必須とする東京都内のA小学校で行い、調査対象は児童数434名、回答数426名である(天野・鈴木 2019: 24)。

ら良いかについて提案する。なお、書き出しは作文の冒頭の1段落を指すこともあれば、1文を指すこともあるが、本研究では、書き出しを作文の冒頭の1文とする。同様に、結びについても、作文の結尾の1段落ではなく、最後の1文とする。

## 2. 書き出しと結びについての先行研究

まず、書き出しについての先行研究には太田(1965)、加藤(2023a, 2023b, 2023c)がある。加藤(2023c)は、太田(1965)と加藤(2023a, 2023b)の分類をもとに、調査した児童作文の書き出しを10に大別している。しかし、加藤(2023c)があげている「動物からの発想」は本研究の調査では見られなかったため、表1のように書き出しを9つに大別する。

表1 加藤(2023c)をもとにした書き出しの類型

大分類	中分類	例
1. 人物的発想	自己中心的発想	わたしは、イタイイタイびょうしりょうかんについて、イタイイタイびょうのことをはじめてしました。
	人物中心の発想	ぼくのひいひいおばあちゃんは、イタイイタイ病の患者だった。
	共同体的な発想	私たちはとてもめぐまれていると思います。
2. 日・時的発想	特定な時から	きょう、イタイイタイびょうしりょうかんに行きました。
	不特定な時から	昔、この富山市婦中町にある神通川にカドミウムというものがたくさんふくんで流れてきました。
3. 場所的発想	場所から	富山県は、水がきれいでおいしいです。
	場面から	社会の授業で四大公害病について勉強しました。
4. テーマからの発想	テーマから	イタイイタイ病は、神通川の上流から流れてきたカドミウムが原因の病気だ。
	テーマに関連する事物から	公害病を二度と起こしたくない。
5. 事象・状態からの発想		ここに一枚の写真があります。
6. 心理状態からの発想		「今、起こったらどうなるんだろう...」
7. 抽象的な説明からの発想		真実を真実として伝えていくことがぼくたちの使命だと思う。
8. 会話からの発想		「わあ。こんなものまですてである。」
9. 行動からの発想		天気予報のニュースを見て、家族はみんな、ため息をしています。

次に、結びについての先行研究には加藤(2023a, 2023b, 2023c)がある。加藤(2023c)は加藤(2023a, 2023b)をもとに、調査した児童作文の結びを次の表2のように8つに大別している。

表2 加藤(2023c)の結びの類型

分類	例
1. 願望表示による帰結	水をムダづかいせずに、大切につかいたいと思いました。
2. 意志表示による帰結	神通川はこれからもきれいな川で、そして、イタイタイ病をわすれないようにします。
3. 意見・主張の表示による帰結	これからもきれいな水が飲めるように一人一人努力しなければいけないと思いました。
4. 感じたこと・思ったことの表示による帰結	イタイタイ病のことを知り、自然の大切さが学べてよかったです。
5. 考えの表示による帰結	二度と公害を繰り返さず美しく豊かな環境を未来に引き継ぐこと、それが私たちの使命だと思いました。
6. 事象の説明による帰結	この工事は二千十二年に終わりました。
7. 前で述べたことのまとめによる帰結	これがぼくが出した答えです。
8. 読み手への呼びかけによる帰結	みなさんも、この先イタイタイ病を起こさない、ちょっとした気づかいを心がけましょう。

上の表1と表2をもとに、4節で読書感想文の書き出しと結びの分類を行う<sup>2</sup>。

### 3. 調査資料について

本研究が調査資料とするのは、愛知教育大学附属岡崎小学校で発行している児童文集「ひばり」の70号(2019年)～74号(2023年)に収録されている読書感想文(1・2年生は「本は楽しいよ」、3～6年生は「読書感想文」に分類されているもの)である。各号における学年別の具体的な作文数は次の表3の通りである。

表3 各号における学年別の作文数

	1年	2年	3年	4年	5年	6年	合計
70号(2019)	3	4	8	3	3	6	27
71号(2020)	7	12	5	9	4	3	40
72号(2021)	5	7	10	5	4	6	37
73号(2022)	8	7	9	3	2	4	33
74号(2023)	3	8	6	6	5	5	33
合計	26	38	38	26	18	24	170

<sup>2</sup> 安藤(2023: 120)のように、読書感想文の基本の型は、次の①～③の順で書くことであると述べているものもある。

- ①その本を選んだ理由
- ②本のあらすじを、最初・途中・最後に分けて書く
- ③その本でよかったところ

また、安藤(2023: 126)は、心に残る読書感想文の型として、次の5つをあげている。

- ①本を読んで、一番心にのこった言葉や場面
- ②その言葉や場面が一番心にのこった理由
- ③自分の体験や気持ちと共通する部分
- ④自分と共通する部分について、具体的な例
- ⑤本を通して気づいたことや、読む前と後で変化した自分の気持ち

#### 4. 読書感想文の書き出しの分析

まず、書き出しの分析を行う。

##### 4.1 書き出しの類型

先ほど見た表 1 をもとに、読書感想文の書き出しを分類したものが次の表 4 である。本研究で新たに見られた、「10. 本文の引用からの発想」を加えた 10 に分類した。

表 4 読書感想文における書き出しの類型

大分類	中分類	低学年		中学年		高学年		
		1年	2年	3年	4年	5年	6年	
1. 人物的発想	自己中心的発想	15(58%)	20(53%)	23(61%)	17(65%)	7(39%)	13(54%)	95
	人物中心の発想	3(12%)	5(13%)	4(11%)	2(8%)	1(6%)	2(8%)	17
	共同体的な発想							0
2. 日・時的発想	特定な時から		1(3%)		1(4%)	1(6%)		3
	不特定な時から				1(4%)			1
3. 場所的発想	場所から		1(3%)	2(5%)		1(6%)		4
	場面から							0
4. テーマからの発想	テーマから	1(4%)	2(5%)					3
	テーマに関連する事物から							0
5. 事象・状態からの発想		1(4%)	1(3%)					2
6. 心理状態からの発想		2(8%)	5(13%)	7(18%)	2(8%)	4(22%)	7(29%)	27
7. 抽象的な説明からの発想			1(3%)	1(3%)		2(11%)		4
8. 会話からの発想			1(3%)		1(4%)			2
9. 行動からの発想		3(12%)	1(3%)		1(4%)		1(4%)	6
10. 本文の引用からの発想		1(4%)		1(3%)	1(4%)	2(11%)	1(4%)	6
合計		26	38	38	26	18	24	170

1～6年生までのすべての学年で最も多い書き出しは「1. 人物的発想」の「自己中心的発想」であり、「わたし／ぼく」から書き始められている。以下で、各学年の特徴について詳細に見ていく。

##### 4.2 学齢別の書き出しの特徴

###### 4.2.1 低学年の書き出しの特徴

1年生の書き出しで最も多いのは、「1. 人物的発想」のうちの「自己中心的発想」であり、次の(1)と(2)のように、「わたし／ぼくは「本のタイトル」をよみました」と書かれている。このように、作文のタイトルとほとんど同じような表現で文章が始まっているものが多い。

- (1) (作文のタイトル) 「しんせつなともだち」をよんで  
わたしは、「しんせつなともだち」というえほんをよみました。

(第 70 号、1 年生)

- (2) (作文のタイトル) 木のまつり  
ぼくは、『木のまつり』というはなしをよみました。 (第 73 号、1 年生)

また、2 番目に多い「9. 行動からの発想」においても、次の(3)のように、「よみました」という動詞が使われており、上の(1)や(2)と同様、作文のタイトルと書き出しが同じような表現となっている。

- (3) (作文のタイトル)『ゴリラの おとうちゃん』をよんで  
『ゴリラのおとうちゃん』という本をよみました。 (第73号、1年生)

「9. 行動からの発想」と同数で2 番目に多い「2. 人物的発想」の「人物中心の発想」については、次の(4)のように「おかあさん」が主語となっている文が多く見られる。

- (4) おかあさんが ぼくに  
「この ほんを よんでみて。」  
と みせて くれました。 (第72号、1年生)

2 年生においても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、1 年生の書き出しのように「わたし／ぼくが「本のタイトル」をよみました」といった文だけでなく、次の(5)のように本を読んで思ったことや、(6)のように本を選んだ理由について書かれている。

- (5) ぼくは、「スタンリーとちいさな火星人」という本を読んで、本当に火星人がいたらおもしろいと思いました。 (第71号、2年生)  
(6) ぼくが、なぜこの本をえらんだかというと、この本がすきで、何回も読んでいます。 (第70号、2年生)

2 番目に多いのは、「2. 人物的発想」の「人物中心の発想」であるが、2 年生では次の(7)のように、本の登場人物 ((7)ではバート・ダウじいさん) が主語となっているものが見られる。

- (7) バート・ダウじいさんがふねにいろをぬって、ふねで海にでかけてくじらをつっちゃいました。 (第71号、2年生)

また、同率で2 番目に多い「6. 心理状態からの発想」においては、本の内容だけではなく、次の(8)のように、表紙を見て思ったことを書いているものも見られる。

- (8) はじめに、ひょう紙を見て、なすにかんけいがある話だなあとと思いました。  
(第73号、2年生)

このように、1 年生の書き出しの特徴は「(わたし／ぼくは)「本のタイトル」をよみました」のように、作文のタイトルとほとんど同じような表現が用いられているものが多い点である。一方、2 年生の書き出しは、「わたし／ぼくが「本のタイトル」をよみました」といった内容で文が始まってもそれだけでは終わらず、その後本を読んで思ったことなどが短く書かれている点にある。



#### 4.2.2 中学年の書き出しの特徴

3年生においても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、その内容を見ると1、2年生よりも工夫が見られる。

- (9) (作文のタイトル)「ぞうれっしゃがやってきた」を読んで  
わたしは先月、東山動物園に行ってぞうを見ました。(第70号、3年生)

(9)は作文のタイトルが「ぞうれっしゃがやってきた」であることから、書き出しにぞうに  
関係する内容(著者が動物園に行ってぞうを見たこと)が書かれている。このように、いき  
なり本の内容について書くのではなく、本の内容と間接的に関係のあることを書くという  
工夫が見られる。

また、1・2年生のように「本を読んだ/選んだ」といった表現だけではなく、次の(10)の  
ように「本に出合った」といった表現も用いられており、どのような語を使うかに関して  
も工夫している様子が見られる。

- (10) ぼくがこの本に出合ったのは、読書感想文の本をさがして見つからないので、  
し書さんにジャンルをつたえて場所を教えてもらい、その中の一つがこの本です。  
(第72号、3年生)

さらに3年生になると、次の(11)や(12)のように、書き出しで理由が複数あることを述べ、  
その後に続く文でそれらの理由を列挙するという書き方も見られる。

- (11) わたしがこの本をえらんだ理由は、二つあります。一つ目は、ゾウとポテトチップス  
がすきだからです。二つ目は、「ゾウの森とポテトチップス」という題名をみて、ど  
んな本か気になったからです。(第73号、3年生)
- (12) この本を読もうと思った理由は、二つあります。(第74号、3年生)

4年生においても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、  
次の(13)と(14)のように、本を読もうと思った理由が詳しく書かれているのが特徴的である。

- (13) ぼくがこの本を読もうと思ったきっかけは、まだ生きることについて深く考えたこ  
とがなく、「どう生きるか」という問いに興味をもったからです。  
(第70号、4年生)

- (14) 私は、この『青いあいつがやってきた』という題名を見て、青いあいつとはどんな人  
物なんだろう。って思って本を読み始めた。(第72号、4年生)

(13)では、本を読もうと思ったきっかけが書かれており、(14)では題名に疑問を持ち、本  
を読み始めたことが書かれている。

また、次の(15)のように、本を読んで強く思ったことが書き出しに書かれているものも見  
られる。

- (15) ぼくがこの本を読んで一番思ったことは、戦争は絶対にしてはいけないということです。(第71号、4年生)

以上から、3・4年生においても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、3年生の書き出しの特徴は1・2年生のように「本を読んだ／選んだ」といった表現だけではなく、内容や表現に工夫が見られる点にある。また、「理由、二つあります」のように、理由が複数あることを述べる文が見られるのも特徴的である。一方の4年生においては、本を読もうと思った理由が詳しく書かれていたり、本を読んで強く思ったことが書き出しに書かれているという特徴が見られる。

#### 4.2.3 高学年の書き出しの特徴

5年生においても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、他の学年に比べて数が少なく、その一方で「6. 心理状態からの発想」が多く用いられている。「6. 心理状態からの発想」においては、次の(16)のように本を初めて見て思ったことが鉤かっこつきで書かれていたり、(17)のように一回目に読んだ時に感じたことが書かれているものが見られる。いずれも、他の学年と比べて書き出しがより詳細に書かれているため、一文も長くなりがちである。

- (16) 「神様ってどんな姿をしているんだろう。」  
これが、この本を初めて見たときに思ったことです。(第74号、5年生)
- (17) あれほど学級委員に選ばれたかった少年だったのに、いざライバルの榎本君に勝ったとき、不安になってしまうのが不思議だなと感じたのが一回目に読んだときのことです。(第70号、5年生)

また、次の(18)や(19)のように、筆者の心に残った一文の引用によって、文章が書き始められているものも見られる。

- (18) 「ほんとうのものをみるゆうきがあれば、かべはきえる。ぜんぶきえたあとには、きっとすばらしいせかいがあるはずだよ。」  
この本を読んで強く心に残ったのが、青い鳥のこの言葉でした。(第71号、5年生)
- (19) 「目が見えない」からって周りに差し伸べられるその手はどうしても好きになれなくて。始めに読み、さっそくドキッと不安になりました。(第72号、5年生)

6年生の書き出しにおいても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、現時点での出来事やその時に抱えている感情だけではなく、次の(20)のように過去を振り返って書いているものや、(21)のように日々抱えている思いを本のタイトルにふれながら書いているものも見られる。そのような傾向は(22)のような「9. 行動からの発想」においても同様に見られる。

- (20) 私は四年生のときに、自由研究で第二次世界大戦中にナチスドイツが多くユダヤ人などを迫害した「ホロコースト」について調べました。(第72号、6年生)
- (21) 僕はこの本の題名の通り、よく「どうせ無理」だと思ってしまいます。  
(第72号、6年生)

(22) ドラえもののアニメを見ていて、私とのび太は似ているなあと思うことがあります。  
(第70号、6年生)

以上から、5・6年生においても最も多いのは「1. 人物的発想」のうちの「自己中心的発想」であるが、5年生は他の学年に比べて「1. 人物的発想」の数が少なく、「6. 心理状態からの発想」が多いのが特徴的である。また、「6. 心理状態からの発想」では、他の学年と比べ、書き出しがより詳細に書かれているため、一文も長くなりがちである。一方の6年生の特徴は、現時点での出来事やその時に抱いている感情だけではなく、過去を振り返って書いているものや、日々抱いている思いを書いているものが見られる点にある。

## 5. 読書感想文の結びの分析

次に、結びの分析を行う。

### 5.1 結びの種類

表2をもとに、読書感想文の結びを分類したものが次の表5である。なお、結びの多くが「～と思いました／と思います」で終わっているため、そのような場合には「～と思いました／と思います」に前接する表現にもとづいて分類を行った。

表5 読書感想文における結びの種類

分類	低学年		中学年		高学年		合計
	1年	2年	3年	4年	5年	6年	
1. 願望表示による帰結	11(33%)	8(26%)	26(60%)	11(52%)	11(61%)	14(58%)	81
2. 意志表示による帰結	7(21%)	6(19%)	2(5%)	4(19%)		1(4%)	20
3. 意見・主張の表示による帰結			1(2%)			1(4%)	2
4. 感じたこと・思ったことの表示による帰結	13(39%)	12(39%)	9(21%)	4(19%)	6(33%)	1(4%)	45
5. 考えの表示による帰結	2(6%)	4(13%)	4(9%)	1(5%)	1(6%)	4(17%)	16
6. 事象の説明による帰結							0
7. 前で述べたことのまとめによる帰結			1(2%)			1(4%)	2
8. 読み手への呼びかけによる帰結		1(3%)		1(5%)		2(8%)	4
合計	33	31	43	21	18	24	170

表5を見ると、低学年で最も多いのは「4. 感じたこと・思ったことの表示による帰結」であり、中学年と高学年で最も多いのは「1. 願望表示による帰結」である。以下で、各学年の特徴について詳細に見ていく。

### 5.2 学齢別の結びの特徴

#### 5.2.1 低学年の結びの特徴

まず、1年生の結びを見ると、「4. 感じたこと・思ったことの表示による帰結」が最も多い。その内容は次の(23)のように、「この話を読んで良かった」や、(24)や(25)のような「うれしいと思いました」「たのしいです」のように本を読んで抱いた感情が述べられているものが多いのが特徴である。

- (23) このはなしをよんで、じぶんのせいかくをみつめなおすことができよかつたとおもいます。(第71号、1年生)
- (24) みんながへいわになれると、うれしいとおもいました。(第73号、1年生)
- (25) どうしてかというと、ぼくのおとうさんはゴリラのおとうちゃんみたいに、おとうちゃんのぼりやおとうちゃんブランコをやっているからです。たのしいです。(第73号、1年生)

2年生の結びにおいても「4. 感じたこと・思ったことの表示による帰結」が最も多いが、本を読んで思ったことを書く際にも1年生のように「この話を読んで良かつた」といった表現だけではなく、次の(26)のように「本を選んで良かつた」という表現が使われている。また、感情を示す際にも「うれしい」「たのしい」といった表現だけではなく、次の(27)のような「うれしくてほっとしました」や(28)のような「すてきなはなしでした」のように、1年生の時よりも多様な表現が使われているのが特徴的である。

- (26) たかいとうが三人にそっくりで、とてもすてきなはなしでした。(第71号、2年生)
- (27) 先生が門をあけてくれたときは、すごくうれしくてほっとしました。(第72号、2年生)
- (28) いつもは、おもしろいお話とかを読んでいるけれど、このお話はいつもよりハラハラドキドキのお話なので、この本をえらんでよかつたと思います。(第70号、2年生)

以上のように、1・2年生の結びは「4. 感じたこと・思ったことの表示による帰結」が最も多いが、その内容を見ると、1年生は本を読んで抱いた感情を「よかつた」「うれしい」「たのしい」といった表現を用いて表現しているのに対し、2年生では「この話を読んで良かつた」といった表現だけではなく、「本を選んで良かつた」という表現が使われていたり、感情を示す際にも1年生よりも多様な表現が使われているという特徴が見られる。

### 5.2.2 中学年の結びの特徴

3年生の結びにおいて最も多いのは「1. 願望表示による帰結」であり、全体の半数以上を占めている。3年生では読んだ本の内容に触発され、次の(29)や(31)のように「～てみたい」「～ていきたい」という表現を用いて今後したいと思っていることを表したり、(30)のように「～たくなりました」を用いて気持ちの変化を表したりしている。

- (29) いつか食べてみたいです。(第72号、3年生)
- (30) どこかに針葉じゅだけの森がないか調べたくなりました。(第73号、3年生)
- (31) 私もそういう出会いからしんせつな人になり、友だちもどんどん作っていききたいと思いました。(第70号、3年生)

一方の4年生の結びにおいても最も多いのは「1. 願望表示による帰結」であるが、「～たい」といった表現だけではなく、次の(32)のように「ぼくは願います」といった表現も使われている。また、(33)のように主人公に「近づけるようにしたい」と述べられていたり、(34)で「これから先」という表現が使われているように、将来したいと思っていることが記述されているのが特徴である。

- (32) そして、ネモ船長が再びこの世に現れたとき、その頭のうが戦争ではなく、平和のために役立てられることを期待したい。」と言うアロナクス博士の言葉をしっかりと心

- のおくそこに、残してくれることをぼくは願います。(第71号、4年生)
- (33) 私は、アイサトのような勇氣は、まだないけれど、自分でできることを考えて行動してアイサトに近づけるようにしたいです。(第72号、4年生)
- (34) これから先、世界が平和でいられるように、ぼくたちができることは何かをつねに考えながら生きていきたいです。(第70号、4年生)

以上のように、3・4年生で最も多く使われているのは「1. 願望表示による帰結」であるが、3年生では読んだ本の内容に触発され、「～てみたい」「～たくなりました」「～ていきたい」といった表現を用いて近い将来にしたいと思っていることを記述しているのに対し、4年生では「願います」といった表現も用いられていたり、将来や今後したいと思っていることが記述されているという特徴が見られる。

### 5.2.3 高学年の結びの特徴

5年生の結びにおいても「1. 願望表示による帰結」が多く用いられているが、「4. 感じたこと・思ったことの表示による帰結」も多く用いられている。「4. 感じたこと・思ったことの表示による帰結」を述べている例を見ると、低学年のように本を読んで良かったなどと表現しているだけでなく、次の(35)のように本が筆者に与えた印象を述べたり、(36)のように本の内容全体の感想や、(37)のように本を読んで抱いた感情が述べられている。

- (35) この本は、私を二つも成長させてくれたので、よい本だなと思いました。(第72号、5年生)
- (36) それに、最後まで読まないで話の結末がわからないので、そこがこの銭天堂のおもしろいところだと思います。(第70号、5年生)
- (37) 人間も動物も、今の環境や時代に合わせて変化したり、変化しきれないものもあつたりして、命って不思議だなと、この本を読んで思いました。(第72号、5年生)

一方の6年生においても最も多く使われているのは「1. 願望表示による帰結」である。それ以外にも数は少ないものの、「6. 事象の説明による帰結」以外のすべての表現が用いられているのが特徴的である。次の(38)は「8. 読み手への呼びかけによる帰結」であり、ネコに「やさしくしてあげてください」と言って文章を終えている。また、(39)のように本を読んで考えたことを主張しているもの(「3. 意見・主張の表示による帰結」)や、(40)や(41)のように本で学んだことを今後に生かそうという姿勢、自身が成長していきたいと思っていることを述べているものもある。さらに、次の(42)のように、自身だけではなく、他の人と助け合っているいろいろなことにチャレンジしていきたいと記述されているものも見られる。

- (38) ネコはあなたとの時間を大切にしています。なので、のらネコにあつたら、いじめたりしないで、やさしくしてあげてください。(第70号、6年生)
- (39) この平和がいつまでも続くように、世界中の人々が戦争とはどれほどおそろしいものなのかを自覚して、問題が起こったときに議論を重ねていく必要があると思います。(第74号、6年生)
- (40) わたしは、この本を読んだ経験を生かし、うそのうわさに流されないようにしたいです。(第74号、6年生)
- (41) これからいろいろなことに挑戦していき、颯太のように成長していきたいと思ます。(第72号、6年生)
- (42) ベルナとななえさんは互いに助け合うことで、うまくいろいろなことができたと思うので、ぼくもみんなと助け合っているいろいろなことにチャレンジしていきたいです。(第71号、6年生)

以上のように、5・6年生の結びにおいても「1. 願望表示による帰結」が多く用いられているが、5年生は「4. 感じたこと・思ったことの表示による帰結」も多く用いられているのが特徴的である。「4. 感じたこと・思ったことの表示による帰結」を述べる際には、低学年のように本を読んで良かったなどと表現するだけでなく、本が筆者に与えた印象を述べたり、本の内容全体の感想、本を読んで抱いた感情についても記述されている。一方の6年生は「6. 事象の説明による帰結」以外のすべての表現が用いられているのが特徴的である。6年生になると、さまざま表現を使って結びを記述することができるようになると考えられる。

## 6. 指導法について

読書感想文の書き方について指導する際、教師はどのような点に注意して指導したら良いかを、3つの観点から述べる。

### 6.1 読書感想文の相手意識や目的意識の指導

読書感想文で何を書けば良いのかと悩む子どもは多い。そのため、読書感想文の指導では、誰に向けて何のために書くのかという相手意識や目的意識をもたせることが大切である。例えば「自分が読んだ本で、なるほどと思ったことと、それで自分の考えがどう変わったかをクラスの仲間に伝えよう」といった具合に、具体的に示す。そうすることで、何を書いたら良いのかを明確にして書き始めることができる。また、相手意識や目的意識が定まれば、その後のテーマ決め等も、一貫性を持たせることができる。さらに、過去に書かれた作品があれば、このタイミングで示すと良い。どのようなことを書いたら良いのかと、具体的なイメージをもちながら読書感想文を執筆できるようになるからである。なお、本を選ぶためのアドバイスでは、主人公と自分に共通点があるものや、自分の考えや経験と関わらせて読むことができる内容のものが良いと伝えると効果的である。

### 6.2 テーマや内容に関する指導

テーマや内容に関する指導では、自分の考えや体験と関わりのあるテーマにすると良いことを伝える。自分事として書くことで、内容をより生き生きと書くことができるようになるとともに、テーマや内容決めに苦勞しなければ、文章の表現を工夫することに力を注げるようになるからである。そして、変容のきっかけとなった場面や言葉をできるだけ簡潔に書くように指導するとよい。できるだけ簡潔に書くように指導しておくことで、子どもは、選択する言葉をよく吟味するようになるからである。

### 6.3 形式に関する指導

形式に関する指導では、段落構成の決定、文章全体のおおよその完成、表現の工夫の順で指導を行うと効果的である。読書感想文では、いきなり文頭から書き始めるのではなく、段落の中心となりそうな文をいくつか考えるよう指導する。文が決まったら、次に段落の構成を考えるよう指導する。文を並び替えたりするなどし、最も伝わりそうな段落の構成が決まったら、段落の中心となる文を膨らませて書くよう指導する。そして、文章の全体が書けた段階で、表現の工夫の指導に入っていく。この段階になると、文章の全体を捉えながら、効果的な書き出しや印象に残る結びを書くことができるようになる。そのため、学齢別の書き出しのパターンを示すのであれば、このタイミングが効果的である。また、表現の工夫の指導と並行して、主述のねじれと誤字脱字の指導を行う。教師にとって指導しやすい内容であ

るとともに、小学生段階であれば、それらを修正するだけで、文章の仕上がりの印象が大きく変化するからである。

以上のように、読書感想文の指導には、相手意識や目的意識の指導、テーマや内容に関する指導、形式に関する指導がある。そして、教師はそれら全てを同時に行うのではなく、子どもの思考に沿いながら、順を追って指導していくとよいのである。

### 謝 辞

本研究は、令和5年度愛知教育大学学長裁量経費（連携推進分野）の助成を受けたものである。

### 文 献

- 天野未来・鈴木貴史 (2019). 「自立した読者」を育む読書感想文指導『帝京科学大学教育・教職研究』5-1、pp.23-29.
- 安藤英明 (2023). 『小学校6年生までに必要な作文力が1冊でしっかり身につく本』かんき出版
- 太田康治 (1965). 「作文指導 その反省と考察—書き出し文による発想の類型及びその傾向について—」『教育実践研究集録／新潟県立教育センター』2、pp.9-14.
- 加藤恵梨 (2023a). 「児童作文における書き出しと結びの分析」『言語資源ワークショップ2022 発表論文集・ポスター集』、pp.347-357.
- 加藤恵梨 (2023b). 「児童作文コンクール受賞作品における書き出しと結びの分析」『社会言語学会第47回大会発表論文集』、pp.135-138.
- 加藤恵梨 (2023c). 「上級中国人日本語学習者と上級韓国人日本語学習者の作文の書き出しと結びについて—日本語母語話者の作文と比較して—」『2023年度日本語教育学会春季大会予稿集』
- 櫻本明美 (1996). 「説明的表現の指導—書き出しの工夫に着目して—」『神戸親和女子大学研究論叢』30、pp.400-422.
- 鈴木恵 (2016). 「書くこと」指導への一提言—読書感想文指導を中心に—『新潟大学教育学部研究紀要人文・社会科学編』9-1、pp.75-92.
- 田近洵一・井上尚美・中村和弘（編）(2018). 『国語教育指導用語辞典〔第五版〕』教育出版

## 中古和文資料『夜の寝覚』のコーパス構築の試み

菊池 そのみ (筑波大学)

菅野 倫匡 (筑波大学)

### An Attempt to Build a Corpus of Courtly Literary Works in Early Middle Japanese Language: *Yoru no Nezame*

Kikuchi Sonomi (University of Tsukuba)

Kanno Michimasa (University of Tsukuba)

#### 要旨

本稿は中古和文資料『夜の寝覚』を対象とし、そのコーパスを構築する試みについて作業の手順と現在の進捗状況とを報告した上で今後の展望を述べるものである。まず、本稿では資料として取り上げる『夜の寝覚』の概要を説明し、この資料を対象としてコーパスを構築するための手順を提示した。次に作業を終えた約 2.3 万語のデータに基づき、コーパス全体の総語数を推計すると約 11 万語となることやそれを正解データとすると現行の「中古和文 UniDic」による解析精度が約 98%となることを報告した。また、この正解データに現れた未登録語（未知語）についても併せて一覧として示した。最後に「動詞+テ+動詞」を取り上げ、このようなコーパスを構築することによって容易になる調査の事例を紹介した。

#### 1. はじめに

国立国語研究所による『日本語歴史コーパス』(以下、CHJ)は『同平安時代編』の公開に端を発して段階的に資料や機能の拡充が図られてきた。小木曾(2022:211)によれば、それに伴う一連の構築作業を通して「形態素解析用辞書の整備や本文整備」に関する技術や知見の蓄積が進み、このことによって「研究に必要な本文のコーパスを研究者自身が作ることがすでに可能になってきた」とのことである。

このような昨今の状況を踏まえ、本稿は中古和文資料の1つである『夜の寝覚』を対象とし、そのコーパスを実際に構築する試みについて述べるものである。以下では初めに本稿において取り上げる『夜の寝覚』の概要を示し、次いでコーパスを構築するための手順を説明すると共に現在の進捗状況を報告する。また、このようなコーパスを構築することによって容易になる「動詞+テ+動詞」の調査の事例についても紹介する。

#### 2. 『夜の寝覚』の概要

ここでは『夜の寝覚』の成立、諸本、中古和文資料としての位置づけについて概要を順に示す。なお、資料名は「夜半の寝覚」、「寝覚」、「寝覚物語」などとするものもあるが(阪倉 1964b; 平林 2007)、本稿では『新編日本古典文学全集』に倣って「夜の寝覚」とする。

まず、『夜の寝覚』については11世紀初頭から12世紀初頭までの時期に成立した可能性を指摘する関根・小松(1960)の説や11世紀後半から12世紀前半までの時期に成立したとする阪倉(1964a, 1964b)の説などがあり、概ね中古後期に成立したものと見られる。その作者は『更級日記』の作者として知られる菅原孝標女であると見る説とそれを否定する説とがあり、成立時期と併せて「詳しくは不明」(平林 2007:691)とのことである。なお、成立に関する議論は鈴木(1993)に詳しい。



次に現存する諸本については鈴木(1996)の整理によると五巻本系統に属する写本(島原図書館松平文庫蔵本、天理図書館蔵竹柏園旧蔵本、国立国会図書館蔵本、東北大学狩野文庫蔵本、静嘉堂文庫蔵本、実践女子大学蔵本)と三巻本系統に属する写本(前田家尊経閣文庫蔵本)とに大別し得るとのことである<sup>1</sup>。なお、諸本間の異同は鈴木(1965)の研究や橋本(1933)、高村(1966)、関根・小林・平野・中沢・伊藤・泉(1975)などの校本に譲る。

また、鈴木(1996:583)によれば、諸本のうち「善本と目されるのは島原本と前田家本であり、ともに近世初期の書写である」とのことである。両者のうち前者を底本とするものに『日本古典文学大系』、『日本古典文学全集』、『完訳日本の古典』、『新編日本古典文学全集』があり、後者を底本とするものに関根・小松(1960, 1972)がある。なお、『夜の寝覚』は中間(五巻本における巻二と巻三との間に当たる部分)と末尾(五巻本における巻五の続きに当たる部分)とに欠落があることが知られており、『源氏物語』などに比して「不完全なもの」(永井1960:140)であるとの指摘もある。

続いて中古和文資料における『夜の寝覚』は後期の物語資料の1つに位置づけられ、特に『源氏物語』の影響の著しいもの(平林2007:691)として知られる。また、中古語の語彙や文法の研究においては『日本古典文学大系』に基づいた語彙索引である『夜の寝覚総索引』(阪倉・高村・志水1974)や文字列検索のみを備えた国文学研究資料館『日本古典文学大系本文データベース』を用いることによって研究——山崎(2013)の整理に倣えば、「定性的研究」——が進められてきた(村田2001;安部・菊池2016;辻本2016;菊池2019など)。一方、近年は『日本語歴史コーパス平安時代編』(以下、CHJ平安時代編)の公開・拡充により、それを利用した中古語の研究が増えてきているが、『夜の寝覚』はCHJ平安時代編に収められておらず、コーパスを利用せずには遂行し得ない小林・岡崎(2017)、菊池・菅野(2019)、大川(2020)などの研究——同じく山崎(2013)の整理に倣えば、「定量的研究」——において対象になっていない現状にある。更に前掲の国文学研究資料館『日本古典文学大系本文データベース』が2023年4月1日から公開停止となったことに照らせば、現時点では用例を目視で探すことや索引で引くことを除いて『夜の寝覚』を対象とした研究の実施が困難な状況にあり、定性的研究・定量的研究の双方に資するコーパスの整備が求められるところである。

### 3. コーパス構築の手順

ここではコーパスを構築する際に依拠する資料について述べた上でコーパスを構築するための手順として対象とする資料の本文の電子化と品詞や語種などの情報——小椋(2014)の言うところの「形態論情報」——の認定とを順に説明する。

まず、本稿ではコーパスを構築する際に依拠するものとして『新編日本古典文学全集』を採用することとした。前述の通り、この『新編日本古典文学全集』は鈴木(1996)において「善本」と看做されている近世初期の写本である島原図書館松平文庫蔵本を底本とするものである。なお、コーパスを構築する際に依拠した資料をも仮に「底本」(菅野・菊池2021;菊池・菅野2022)または「コーパスの底本」(小木曾2022)と呼び得るならば、『新編日本古典文学全集』はコーパスの底本、島原図書館松平文庫蔵本はコーパスの底本の底本ということになる。

<sup>1</sup> 五巻本と三巻本との対応関係については鈴木(1996)も挙げているように前者の巻一・巻二が後者の上巻に当たり、前者の巻三・巻四・巻五が後者の中巻・下巻に当たるとする永井(1960)の指摘がある。

次に対象とする資料の本文を電子化するに当たっては光学文字認識 (OCR) ソフトウェア「読取革命」(Ver.16) を用いて『新編日本古典文学全集』の頭注と現代語訳と (小) 見出しとを除き<sup>2</sup>、本文を電子化した<sup>3</sup>。また、誤りを人手で修正した。

続いて形態論情報の認定に当たっては形態素解析器「MeCab」(Ver.0.996) と解析用辞書「中古和文 UniDic」(Ver.202203) とを用いて形態素解析を実施した。また、各語に認定する形態論情報としては原文文字列、品詞大分類、品詞中分類、品詞小分類、品詞細分類、活用型、活用形、語彙素読み、語彙素表記、書字形出現形、発音形出現形、書字形基本形、発音形基本形、語種、語形、語形基本形、語彙表 ID、語彙素 ID の 18 項目であり、これに参考情報として資料名、巻情報、頁数、行数、位置 ID、ルビの 6 項目を追加し、各語について計 24 項目を認定した<sup>4</sup>。なお、形態論情報については誤りを人手で修正した。

最後にこれまでに述べた作業の手順を模式図として図 1 に示す。

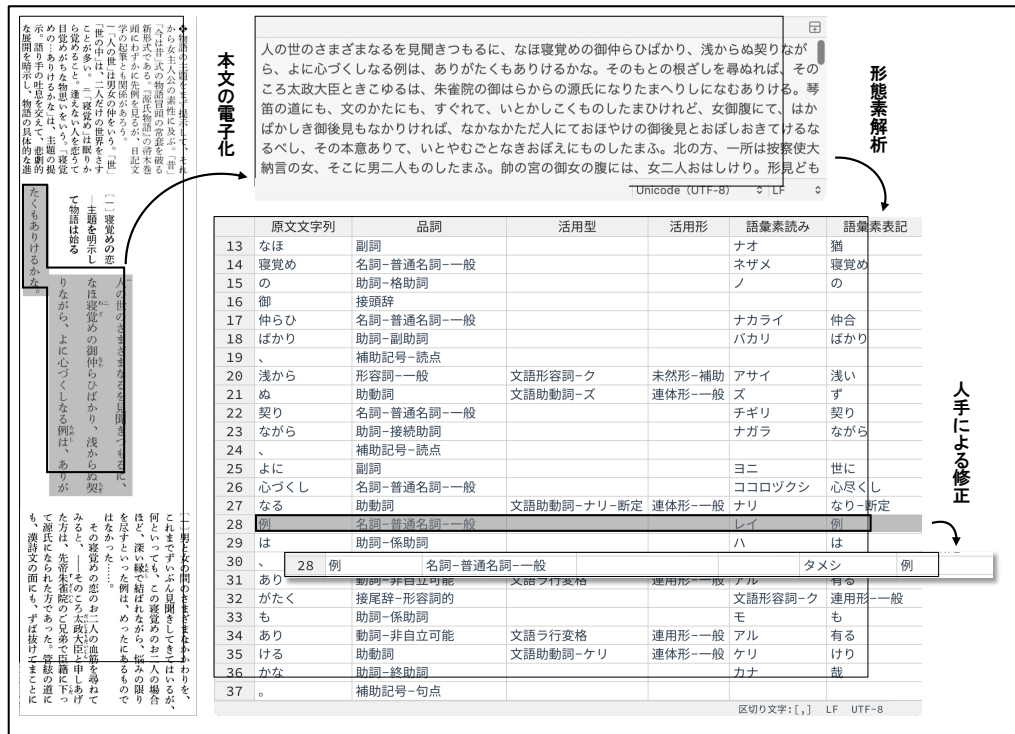


図 1 『夜の寝覚』のコーパスを構築する作業の手順

<sup>2</sup> これ以外にも原則として褐色で示されている文字 (小見出し、注番号、和歌の詠み手、図の説明など) は本文から除いた。

<sup>3</sup> 本稿では『新編日本古典文学全集』の本文を電子化するに当たって公益社団法人著作権情報センターに照会し、当該の行為が「著作物に表現された思想又は感情の享受を目的としない行為」(文化庁著作権課 2019:14) として 2018 年改正の著作権法第 30 条の 4 の権利制限の対象となる——著作権法上の例外に該当する——ことを 2023 年 7 月 26 日に確認した。また、当該のコーパスを公開するに当たっては「軽微性等の要件」(文化庁著作権課 2019:14) ——小木曾 (2023) によると具体的にはキーとなる語の前後 30 語——を満たせば、著作権法第 47 条の 5 第 1 項の準備のための行為として第 47 条の 5 第 2 項における権利制限の対象となることも同日に確認した。

<sup>4</sup> 語形については解析用辞書「中古和文 UniDic」(Ver.202203) では空欄となるが、並行して作業を進めている芥川龍之介賞受賞作品のコーパス——その構想は菅野 (2021) に述べた——と各語に認定する項目を統一して整備するために空欄のままであっても項目としては残してある。

#### 4. 進捗状況

ここでは既に作業を終えた『夜の寝覚』巻一に相当するデータに基づき、コーパス全体の推定規模（推計総語数）、現行の形態素解析の精度、巻一に見られた未登録語（未知語）の一覧を順に示し、進捗状況を報告する。

まず、『夜の寝覚』巻一に相当するデータについて語数——厳密に言えば、短単位数——は 23,185 語であり、これを CHJ 平安時代編所収の各資料の語数と比較すると『大和物語』（26,953 語）に最も近い規模であると言える<sup>5</sup>。また、コーパスの底本である『新編日本古典文学全集』に基づいて巻一のページ数と語数との関係から『夜の寝覚』巻一から巻五までに相当するコーパス全体の総語数を推計すると 113,519 語であり、これは CHJ 平安時代編所収の『大和物語』（26,953 語）と『大鏡』（84,529 語）とを足し合わせた語数に匹敵するものと言える。

次に作業を終えた巻一に相当するデータを正解データとし、現行の解析用辞書「中古和文 UniDic」(Ver.202203) による形態素解析の精度を小木曾（2013）に倣って表 1 に示す。表 1 における「境界」、「品詞」、「語彙素」は小木曾（2013:54-55）と同様に「解析結果において単語の境界が正しかったかどうか」、「境界が正しいことに加えて単語の品詞・活用型・活用形も正しく認定されていたかどうか」、「境界と品詞に加えて語彙素の認定も正しかったかどうか」という各観点からの解析精度を示すものである<sup>6</sup>。

表 1 『夜の寝覚』巻一を正解データとした場合の「中古和文 UniDic」の解析精度

	境界	品詞	語彙素
正解語数	23,185		
出力語数	23,194		
一致語数	23,041	22,856	22,653
再現率	0.99379	0.98581	0.97705
適合率	0.99340	0.98543	0.97668
F 値	0.99360	0.98562	0.97686

表 1 を見ると語彙素認定における解析精度——小木曾（2013:55）によれば、「形態素解析の精度評価では一般に F 値が用いられる」——は F 値が約 98% であり、小木曾（2013）の示す当時のそれが約 97% であることに比して高い水準と言える。また、辞書に未登録の語がデータに含まれる場合に解析精度が低下し、F 値が約 96% になるとする小木曾（2013）の指摘をも踏まえれば、未登録語（未知語）が含まれるデータにおいて高い精度で解析し得たことは特筆に値するものである。これに加えて作品別に見た解析精度について『土佐日記』の解析精度が最も高く語彙素認定で 97.9%、『古今和歌集』が最も低く 93.9% であったとする小木曾（2013:55）の報告に照らしても巻一のデータの解析精度は高い水準であるものと考えられる。

<sup>5</sup> 語数については記号・補助記号・空白を含むことを原則とするが、『夜の寝覚』巻一に相当するデータにおいて記号は 0 短単位であり、補助記号は 3,860 短単位であり、空白は本文の電子化の対象外であることを申し添える。なお、CHJ 平安時代編所収の各資料については言語資源開発センターの Web サイトにおける「語彙統計」(<https://clrd.ninjal.ac.jp/chj/chj-wc.html>) を参照した（2023 年 8 月 10 日確認）。

<sup>6</sup> 表中の「正解語数」、「出力語数」、「一致語数」、「適合率」、「再現率」、「F 値」は小木曾（2013:55）と同様に「人手による修正を行った評価データの語数」、「解析結果として出力されたデータの語数」、「出力語数のうち評価データ（正解）と一致した語数」、「一致語数／出力語数」、「一致語数／正解語数」、「再現率と適合率の調和平均」を指すものである。

このように『夜の寝覚』巻一に相当するデータを高い精度で解析し得たこととして、は継続的な更新によって解析用辞書「中古和文 UniDic」の精度の向上が図られている点に加えて次の2点が要因となっている可能性がある。まず、小木曾（2022:210）は「コーパスの底本としての本文はさまざまなレベルで異なったものが考えられる」とし、書籍と同様に「読みやすい校訂本文と、原文を尊重した翻字本文」とが役割分担の関係にあるとの前提を示した上でCHJの底本については「多くのユーザーにとって読みやすい新編全集を出発点」としたことを説明した。その上で「形態素解析の障害となる表記上の問題への対処、一般的でない表記が含まれることによる形態素解析精度の低さ」が課題になるとして「コーパスを整備する作業量の面から言えば、原文に近いほど難易度は高い」（小木曾 2022:209）ことを指摘した。この指摘を踏まえると『夜の寝覚』のコーパスの底本である『新編日本古典文学全集』は「校訂を重ねて読みやすくした本文」（小木曾 2022:209）に該当し、「形態素解析の障害となる表記上の問題」（小木曾 2022:209）などが少ないことから解析精度が高い水準になったものと考えられる。次に小木曾（2013:56）は解析用辞書「中古和文 UniDic」による形態素解析の精度について「学習用コーパスの大部分を『源氏物語』が占めるため、全体として『源氏物語』に文体（語彙・語法）に近い作品は高い精度で解析ができる傾向がある」ことを指摘した。この指摘を踏まえれば、本稿において対象としている『夜の寝覚』が中古後期の和文物語資料として『源氏物語』の影響の著しいもの（平林 2007:691）と言われることから解析精度が高い水準になったものと見える。

続いて前掲の解析精度にも大きな影響を与えるものとして『夜の寝覚』巻一に見られた語のうち解析用辞書「中古和文 UniDic」（Ver.202203）に登録のない語——小木曾（2013）によれば、「未登録語（未知語）」——を表2に一覧として示す。

表2 『夜の寝覚』巻一に見られた未登録語の一覧

新語 No.	頁-行 原文文字列	品詞/活用型/活用形	語彙素読み	語彙素表記	書字形基本形	語形基本形	語種	語彙素 ID
1011 027-3	時明	名詞-固有名詞-人名-名	トキアキラ	トキアキラ	時明	トキアキラ	固	
1012 089-14	時明	名詞-固有名詞-人名-名	トキアキラ	トキアキラ	時明	トキアキラ	固	
1021 100-9	わななかしく	形容詞-一般/文語形容詞-シク/連用形-一般	ワナナカシイ	わななかしい	わななかし	ワナナカシ	和	
2011 032-1	おもほれ	動詞-一般/文語下二段-ラ行/連用形-一般	オボホレル	溺ほれる	おもほる	オモホル	和	223745
2021 028-14	しなしなく	形容詞-一般/文語形容詞-シク/連用形-一般	シナジナシイ	品品しい	しなしなく	シナシナン	和	238050
3011 100-5	あち	代名詞	アチラ	彼方	あち	アチ	和	769
3021 077-7	あらはかす	動詞-一般/文語四段-サ行/終止形-一般	アラワカス	表わかす	あらはかす	アラワカス	和	293155
3031 099-1	御殿油	名詞-普通名詞-一般	オオトナブラ	大殿油	御殿油	オオトナブラ	和	257929
3032 103-2	御殿油	名詞-普通名詞-一般	オオトナブラ	大殿油	御殿油	オオトナブラ	和	257929
3041 040-10	垣間見	名詞-普通名詞-一般	カイマミ	垣間見	垣間見	カイバミ	和	195895
3051 058-5	公	名詞-普通名詞-一般	キミ	君	公	キン	和	120142
3061 021-3	年上	名詞-普通名詞-一般	コノカミ	兄	年上	コノカミ	和	233192
3071 042-13	好色	名詞-普通名詞-一般	スキ	数奇	好色	スキ	和	56218
3081 032-12	近勝り	名詞-普通名詞-一般	チカマサリ	近優り	近勝り	チカマサリ	和	284591
3091 019-4	早朝	名詞-普通名詞-副詞可能	ツトメテ	つとめて	早朝	ツトメテ	和	170578
3101 042-2	面杖	名詞-普通名詞-一般	ツラヅエ	頬杖	面杖	ツラヅエ	和	216783
3102 117-12	面杖	名詞-普通名詞-一般	ツラヅエ	頬杖	面杖	ツラヅエ	和	216783
3111 049-1	とほらか	形状詞-一般	トオラカ	遠らか	とほらか	トオラカ	和	309074
3121 079-8	放るかさ	動詞-一般/文語四段-サ行/未然形-一般	ハフラカス	放らかす	放るかさ	ハフラカス	和	248531
3131 051-6	ひとびとしく	形容詞-一般/文語形容詞-シク/連用形-一般	ヒトビトシイ	人人しい	ひとびとしく	ヒトビトシ	和	249452
3141 050-2	行明	名詞-固有名詞-人名-名	ユキアキラ	ユキアキラ	行明	ユキアキラ	固	340807
3151 027-2	行頼	名詞-固有名詞-人名-名	ユキヨリ	ユキヨリ	行頼	ユキヨリ	固	345242
3152 028-6	行頼	名詞-固有名詞-人名-名	ユキヨリ	ユキヨリ	行頼	ユキヨリ	固	345242
3153 029-12	行頼	名詞-固有名詞-人名-名	ユキヨリ	ユキヨリ	行頼	ユキヨリ	固	345242
3154 030-4	行頼	名詞-固有名詞-人名-名	ユキヨリ	ユキヨリ	行頼	ユキヨリ	固	345242
3161 096-11	緩し	動詞-一般/文語四段-サ行/連用形-一般	ユルス	許す	緩す	ユルス	和	38938
3171 025-2	よかん	形容詞-非自立可能/文語形容詞-ク/連体形-撥音便	ヨイ	良い	よし	ヨシ	和	36983
3181 035-2	由々し	形容詞-一般/文語形容詞-シク/終止形-一般	ヨシヨシイ	由由しい	よしよし	ヨシヨシシ	和	247943

なお、形態素解析用辞書「UniDic」は語彙素>語形>書字形>発音形の4段階の階層構造を持つことから必要な語を各階層に追加し得るとする小木曾(2013)の指摘を踏まえ、この階層に即して整理すると『夜の寝覚』巻一における未登録語は語彙素の階層に登録すべき語が異なり2語(延べ3語)、語形の階層に登録すべき語が異なり2語(延べ2語)、書字形の階層に登録すべき語が異なり19語(延べ23語)であり、最終的に合計すると異なり23語(延べ28語)となる。

以下では表2に示した『夜の寝覚』巻一における未登録語のうち「わななかしく」(新語 No.1021)、「おもほれ」(新語 No.2011)、「しなしなしく」(新語 No.2021)、「よかん」(新語 No.3171)の4語を取り上げ、各語について説明を加えることとする<sup>7</sup>。

まず、(1)の下線部は解析用辞書「中古和文 UniDic」に登録のない語と見られる。この語について『日本国語大辞典』第2版は「[形シク] (動詞「わななく(戦慄)」の形容詞化) 恐れや緊張などのため、ふるえが起こりそうになるほどである。」とし、(1)の用例のみを掲げている。このことから察するに現行の解析用辞書「中古和文 UniDic」に該当する語の登録がないのは CHJ 平安時代編所収の和文資料に用例が見られないからであるという可能性が考えられる。これを踏まえて《わななかしい》を語彙素の階層に登録すべき語であると判断した。

- (1) …御帳の外にみざり出でて、ものなど言ひ紛らはし、さらぬ顔にと思ふ心地も、いとわななかしくわびしけれど、とかく聞き言ふべきやうもなければ…  
(巻一、100頁9行目<sup>8</sup>)

次に(2a)の下線部は解析用辞書「中古和文 UniDic」に登録のない語と見られる。この語について『日本国語大辞典』第2版は「[自ラ下二] おぼれる。涙にくれる。おぼほる。」として(2a)の用例を最初に掲げており、『新編日本古典文学全集』の頭注は「「おもほれ」は「おぼほれ」の誤りか。」と説明している。このことから考えると(2a)の下線部は既に解析用辞書「中古和文 UniDic」に登録のある《涙ほれる》——実際に CHJ 平安時代編においては(2b)のように〈オボホル〉として現れる——と同じ語彙素の異なる語形である可能性がある。また、CHJ 平安時代編において(2c)の下線部と(2d)の下線部とが共に同じ語彙素《守る》であることも踏まえ、〈オモホル〉を語彙素《涙ほれる》の語形の階層に登録すべき語であると判断した。

- (2) a. …と思ひつづくるに、あたらしう、口惜しく、涙におもほれまどひながらも、思ひやりいと静かなる人にて、…  
(巻一、32頁1行目)  
b. …いとといみじきに、聞こえん方なきことどもなれば、ただ涙におぼほれたるばかりをかごとにて、はかばかしうも答へやらずなりぬ。  
(源氏物語・蜻蛉、215頁、20-源氏 1010\_00052・52270)

<sup>7</sup> 説明に当たっては特に書字形を隅つき括弧で括り、語形を山括弧で括り、語彙素を二重山括弧で括って便宜的に区別した。

<sup>8</sup> 用例の引用に当たって縦書きを横書きに改め、ルビ等を除いた。また、下線は強調の意を示し、「…」は文中の一部を省略した意を示すために引用者が施したものである。なお、出典は用例の末尾に示し、『夜の寝覚』巻一の場合には『新編日本古典文学全集』の頁番号・行番号を併記し、CHJ 所収の資料の場合には資料名と『新編日本古典文学全集』の頁番号と CHJ における位置 ID・開始位置とを併記した。

- c. …かしこにまぼりでものせむ、世の中いとはかなければ、いまはかたちをもことになしてむとてなむ… (蜻蛉日記・下, 282 頁, 20-蜻蛉 0974\_00009・44200)
- d. うち見おこせて、つくづくうちまもりて、いといみじと思ひたり。とまるはさらにもいはず。 (蜻蛉日記・上, 141 頁, 20-蜻蛉 0974\_00002・71710)

続いて (3a) の下線部は解析用辞書「中古和文 UniDic」に登録のない語と見られる。この語について『日本国語大辞典』第2版は立項しておらず、『新編日本古典文学全集』の頭注は「「しなしなし」は、いかにも上品である意。」と説明している。また、これと清濁の対立を持つように見える「しなじなし」について『日本国語大辞典』第2版は「[形シク] 素性がいい。上品である。けだかい。」と説明している。このことから察するに (3a) の下線部は既に解析用辞書「中古和文 UniDic」に登録のある《品品しい》——実際に CHJ 平安時代編においては (3b) のように〈シナジナシ〉として現れる——と同じ語彙素の異なる語形である可能性が考えられる。これを踏まえれば、〈シナシナシ〉を語彙素《品品しい》の語形の階層に登録すべき語であると判断した。

ただし、清濁の対立を持つ語形の登録については (3c) に掲げる規定があり、この規定に照らすと『日本国語大辞典』第2版における「しなじなし」の項に「古くは清音であったとの記述」のないことから〈シナシナシ〉を《品品しい》の語形の階層に登録し得ないものと解せる余地もあるという点には留意する必要がある。

- (3) a. 頭つき、様体ほそやかに、しなしなしく、きよらなるに、髪の毛とつややかにゆるゆるとかかりて、目やすき人かな、と見ゆるに、… (巻一, 28 頁 14 行目)
- b. さては、扇よりかみの額つきぞ、あやしく人のかたちを、しなじなしくも下りても、もてなすところなんめる。(紫式部日記, 156 頁, 20-紫式 1010\_00001・115850)
- c. 仮名表記の出現形に対し、現代語で濁音化するところに濁点がついていないという場合の処理に関しては、『日本国語大辞典』第2版で古くは清音であったとの記述が確認できる限り、別語形として清音形を登録し、清音で読む。

(国立国語研究所コーパス開発センター (池上尚) 2016:141, 3.1.6.清濁)

最後に (4a) の下線部は解析用辞書「中古和文 UniDic」に登録のない語と見られる。この語を含む「よかんめり」について『新編日本古典文学全集』の現代語訳は「よいようだ」としており、「よかん」に助動詞「めり」が下接したものと解し得る。このことから考えると (4a) の下線部は既に解析用辞書「中古和文 UniDic」に登録のある《良い》の連体形-撥音便——実際に CHJ 平安時代編においては (4b) のように語形〈ヨカン〉が書字形【よか】として現れる——と同じ語彙素の同一の語形の異なる書字形である可能性がある。また、CHJ 鎌倉時代編においては (4c) のように《良い》の語形〈ヨカン〉が書字形【よかん】として現れており、CHJ 平安時代編においては (4d) のように《繁い》の語形〈シゲカン〉が書字形【しげかん】として現れている。これを踏まえて【よかん】を《良い》の語形〈ヨカン〉の書字形の階層に——厳密には書字形基本形【よし】の書字形出現形【よかん】として——登録すべき語であると判断した。

- (4) a. …ただこの御身に添ひて起き臥したまふを、大臣も、「いとよかんめり。この御方にあづかりて、おぼしうしろめ」とて、… (巻一, 25 頁 2 行目)

- b. …いとをかしげにひねり縫ひたまひければ、「いとよかめり。ことなるかほかたちなき人は、物まめやかに習ひたるぞよき」とて…  
 (落窪物語・巻之一, 19 頁, 20-落窪 0986\_00001・7780)
- c. …「山里は物のさびしき事こそあるなれども、世のうきよりは住みよかんなるものを」とて、おぼしめしたたせ給ひけり。  
 (平家物語・平家灌頂巻・大原入, 506 頁, 30-平家 1250\_13002・2820)
- d. またの日も、まだしきに、「昨日は、うそぶかせたまふこと、しげかんめりしかば、えものも聞こえずなりにき。(蜻蛉日記・下, 345 頁, 20-蜻蛉 0974\_00012・10450)

## 5. 調査の事例

ここでは『夜の寝覚』のコーパスを構築することによって容易になる調査の事例について紹介する。具体的には「動詞+テ+動詞」——大木 (2022) に倣えば、「V1+テ+V2<sup>9</sup>」——について CHJ を対象とした大木 (2022) の調査結果と『夜の寝覚』巻一を対象とした調査結果とを比較する。

なお、調査に当たってはプログラミング言語「Python」(Ver.2.7.18) を用いてコーパスを検索するアプリケーションを作成し、用例を抽出した。このコーパス検索アプリケーションは図 2 のように検索条件を指定すると図 3 のように検索結果を表示する機能を備えたものである。

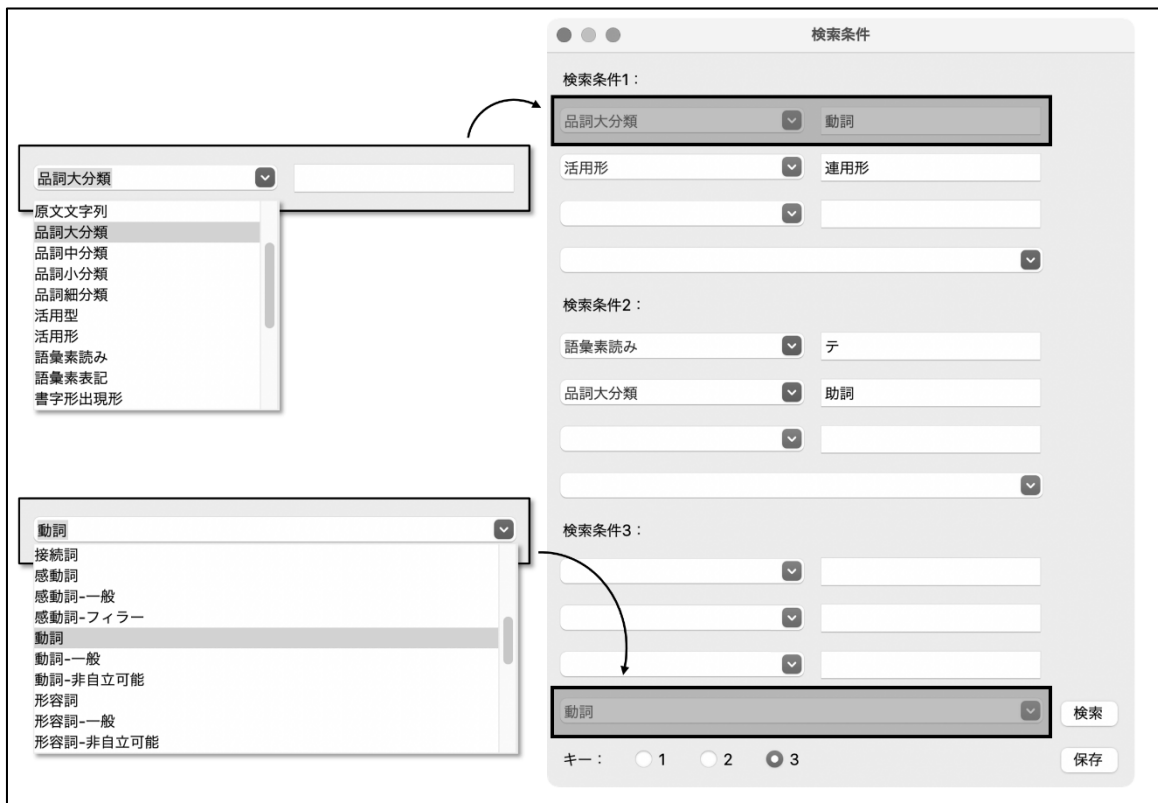


図 2 検索アプリケーションにおける検索条件の設定画面

<sup>9</sup> 大木 (2022:66) は「テ形補助動詞」の成立について述べた青木 (2020) を引用しつつ「動詞テ形に動詞が後接する場合の前項動詞を V1、後項動詞を V2 とする」と説明している。

検索結果

検索結果：108件

巻情報	頁数	行数	位置ID	前文脈	原文文字列	後文脈	品詞	活用型	活用形	語彙素読み	語彙素表記
巻1	17	6	6610	君 の いと すぐ れて	生ひ	たち たまふ に は、	動詞-一般	文語上二段-八行	連用形-一般	オイル	生いる
巻1	17	12	7630	の 年 を 経 て	弾き	しみ たる より も、	動詞-一般	文語四段-カ行	連用形-一般	ヒク	弾く
巻1	17	14	8060	に や が て 傾 き か かり て	御殿籠り	入り たる に、 小	動詞-一般	文語四段-ラ行	連用形-一般	オオントノ	御殿籠る
巻1	18	6	9610	、 琵琶 を 持 て	来	て、 「 今宵 の	動詞-非自立可能	文語力行変格	連用形-一般	クル	来る
巻1	18	10	10290	の 今宵 下 り 来 て	教へ	た て ま つ ら む 」 と て 失 せ	動詞-一般	文語下二段-八行	連用形-一般	オシエル	教える
巻1	19	2	11310	琵琶 を 取 り 寄 せ て	弾き	たまふ に、 大臣 聞 き	動詞-一般	文語四段-カ行	連用形-一般	ヒク	弾く
巻1	19	10	12890	し 月 日 を 数 へ て	待つ	に、 また の 年	動詞-一般	文語四段-タ行	連体形-一般	マツ	待つ
巻1	19	11	13160	の ある か ぎ り 出 だ し て	弾き	たまへ れ ば、 殿	動詞-一般	文語四段-カ行	連用形-一般	ヒク	弾く
巻1	20	15	16140	、 これ こ そ すぐ れて	聞こゆれ	。 昔 より と り わ き 殿	動詞-非自立可能	文語下二段-ヤ行	已然形-一般	キコエル	聞こえる
巻1	24	3	25880	の 風 に 合 は せ て	弾き	たまへ る 音 の、	動詞-一般	文語四段-カ行	連用形-一般	ヒク	弾く
巻1	25	1	28220	の 君 と き こ え て	ものし	たまふ。 こ の 殿	動詞-非自立可能	文語サ行変格	連用形-一般	モノスル	ものする
巻1	25	11	29810	の 御身 に 添 ひ て	起き	臥 した まふ を、 大臣	動詞-一般	文語上二段-カ行	連用形-一般	オキル	起きる
巻1	27	3	34090	は、 所 さ り て	忌み	たまふ べき な れ ば、	動詞-一般	文語四段-マ行	連用形-一般	イム	忌む
巻1	27	8	35000	女 な む、 渡 り て	住み	さ ぶ ら ふ な り。 月 明 き	動詞-一般	文語四段-マ行	連用形-一般	スム	住む
巻1	27	13	36160	弁 少 將 に 契 り て	かしづき	さ ぶ ら ふ 三 に あ た る は	動詞-一般	文語四段-カ行	連用形-一般	カシズク	傅く
巻1	28	1	36680	弁 少 將 に 契 り て	さぶらふ	な り。 と 申 せ ば	動詞-一般	文語四段-八行	連体形-一般	サブ라우	侍う
巻1	28	3	37060	に 歩 み 寄 り た ま ひ て	聞き	たまへ ば、 琴 の	動詞-一般	文語四段-カ行	連用形-一般	キク	聞く
巻1	28	11	38530	は、 いと すぐ れて	聞こゆ	。 こ な た も か な た も	動詞-非自立可能	文語下二段-ヤ行	終止形-一般	キコエル	聞こえる
巻1	30	5	43030	も と に 伝 ひ 寄 り て	見	たまへ ば、 池、	動詞-非自立可能	文語上一段-マ行	連用形-一般	ミル	見る
巻1	31	1	44610	て、 閉 り 入 り て	見	たまへ ば、 和 琴 の	動詞-非自立可能	文語上一段-マ行	連用形-一般	ミル	見る
巻1	32	4	47760	。 人 氣 に お ど る ぎ て	見	返 り た る ほ ど に、	動詞-非自立可能	文語上一段-マ行	連用形-一般	ミル	見る
巻1	36	13	59620	際 は な ほ 忍 び て	やみ	な む 」 と 思 ふ	動詞-一般	文語四段-マ行	連用形-一般	ヤム	止む
巻1	37	4	60690	明 け 暮 れ 出 だ し 入 れ て	見	つ べ か ら む 人 に	動詞-非自立可能	文語上一段-マ行	連用形-一般	ミル	見る
巻1	37	6	61240	、 中 宮 に 申 し て	召し	取 ら せ た て ま つ り て、	動詞-一般	文語四段-サ行	連用形-一般	メス	召す
巻1	37	10	62020	さ る 私 物 に 忍 び て	見	む。 さ て は 人	動詞-非自立可能	文語上一段-マ行	連用形-一般	ミル	見る
巻1	37	13	62450	の う ち に 忍 び て	通は	む こ そ、 い と 見 苦 し か る	動詞-一般	文語四段-八行	未然形-一般	カヨウ	通う
巻1	38	14	65300	の う ち へ 入 り て	見	たまへ ば、 げ に 物 忌	動詞-非自立可能	文語上一段-マ行	連用形-一般	ミル	見る
巻1	39	3	65910	し と、 せ め て	思ひ	く た し 思 ひ 覺 ます に も	動詞-一般	文語四段-八行	連用形-一般	オモウ	思う
巻1	39	12	67600	、 筆 に ま か せ て	書き	流 さ れ た る は、	動詞-一般	文語四段-カ行	連用形-一般	カク	書く

図3 検索アプリケーションにおける検索結果の表示画面

まず、前述した大木（2022）の調査について概要を述べる。大木（2022:67）は「テ形補助動詞」の成立について検討するための出発点として「V1+テ+V2」に「どのようなものがあるのか、また、それが概略どのように移り変わってきているのか」という点を確認するとしてCHJ所収の資料を対象に「平安時代」から「江戸時代」までの「V1+テ+V2」を抽出した上でV2に着目し、その出現頻度の上位30位までの語を一覧として示している。

これに倣って本稿においても既にコーパス構築の作業を終えている『夜の寝覚』巻一から「V1+テ+V2」の用例を抽出した。具体的には前述した検索アプリケーションにおいて図2のように検索条件1の品詞大分類を「動詞」とした上で活用形を「連用形」とし、検索条件2の語彙素読みを「テ」とした上で品詞大分類を「助詞」とし、検索条件3の品詞大分類を「動詞」とし、検索条件3をキーとして検索し、用例を抽出した<sup>10</sup>。その上でV2を出現頻度の順に並べて大木（2022:68-69）の提示する表1における「平安時代」の結果と共に表3に示す<sup>11</sup>。なお、V2を挙げる際には大木（2022）の結果と比較するための便宜を図り、大木（2022）に倣って語彙素（現代語の動詞の形）を示すこととする<sup>12</sup>。

<sup>10</sup> 図2に示すように検索条件として品詞を指定する場合は品詞大分類などを個別に指定する方法——図2の検索条件1——と一括して指定する方法——図2の検索条件3——との2通りがあるが、得られる結果は同じものである。

<sup>11</sup> 大木（2022）の表1における網掛けと丸印とを除いて示した。また、大木（2022:87）には『日本語歴史コーパス』バージョン2021.3』を利用したことや当時のCHJ平安時代編所収の資料を対象としたことなどについて記載はあるが、調査の対象とした資料の詳細や用例を抽出する際の検索条件については具体的な言及がないことを申し添える。

<sup>12</sup> 大木（2022:67）も述べているように語彙素表記の「居る」には「ゐる」「をり（る）」が含まれていることから語彙素表記では語彙素読みの異なる語を弁別し得ないという点には留意する必要がある。



表3 「V1+テ+V2」におけるV2

大木(2022)の調査結果			本稿における『夜の寝覚』巻一のデータ					
順位	語彙素表記	用例数	順位	語彙素表記	用例数	順位	語彙素表記	用例数
1	見る	426	1	見る	16	16	引く-他動詞	1
2	侍る	257	2	思う	9	16	起こす	1
3	参る	222	3	居る	6	16	隠す	1
4	来る	206	4	聞く	4	16	わななく	1
5	出でる	166	4	弾く	4	16	止む	1
6	居る	152	6	侍る	3	16	待つ	1
7	行く	145	6	思す	3	16	窺う	1
8	おわす	139	6	起きる	3	16	応える	1
9	有る	138	6	立つ	3	16	沿う	1
10	思う	114	10	問う	2	16	眺める	1
11	聞こえる	113	10	聞こえる	2	16	住む	1
12	言う	110	10	忍びる	2	16	独り言つ	1
13	奉る	109	10	有る	2	16	出だす	1
14	読む	107	10	侍う	2	16	念ずる	1
15	立つ	101	10	出でる	2	16	籠もる	1
16	返る	88	16	背く	1	16	伏す	1
17	宣う	83	16	淡める	1	16	おわす	1
18	往ぬ	80	16	来る	1	16	ものする	1
19	渡る	80	16	見える	1	16	召す	1
20	思す	74	16	通う	1	16	訪れる	1
21	おわします	73	16	為る	1	16	上る	1
22	ものする	73	16	返る	1	16	給う-尊敬	1
23	見える	72	16	御殿籠る	1	16	寝る	1
23	聞く	72	16	傳く	1	16	頼む	1
25	伏す	70	16	抗う	1	16	生いる	1
26	遣る	67	16	知る	1	16	惑う	1
27	入る	65	16	行く	1	16	然る	1
28	侍う	63	16	覚える	1	16	教える	1
29	泣く	60	16	仕る	1	16	耐える	1
30	止む	56	16	忌む	1	16	書く	1
30	書く	56					計	108

次に表3に基づいて大木(2022)におけるCHJ平安時代編所収の資料を対象とした調査結果と本稿における『夜の寝覚』巻一を対象とした調査結果との比較を試みる。両者の関係を捉えるために表3に示した動詞の出現順位についてSpearmanの順位相関係数を利用して相関分析を実施した。その結果、共通して見られる動詞(20語<sup>13</sup>)について大木(2022)の順位と巻一の順位との間に相関は認められなかった( $\rho = .386, p = .093, n.s.$ )。このように両者に相関が認められなかったことは『夜の寝覚』が『源氏物語』の影響の著しいもの(平林2007:691)と言われてきたことを踏まえれば、CHJ平安時代編所収の和文資料と比較して『夜の寝覚』が特異であることを示唆するものとは看做せず、調査の規模が小さいことの影響を受けている可能性が考えられる。

また、表3について両者の共通点を具体的にみると大木(2022)は上位30位に見られる動詞のうち「見る」、「来る」、「居る」、「行く」、「有る」、「遣る」が現代語におけるテ形補助動詞のV2として見られることを示しており、これらの動詞は『夜の寝覚』巻一にも用例が見られた。特に「見る」は大木(2022)の調査結果においても『夜の寝覚』巻一の調査結果においても最も出現頻度の高い動詞であったという点は注目し値する。更に大木(2022)は現代語と異なる点として中古和文資料においては「侍る」、「おわす」、「有る」などの「敬語

<sup>13</sup> この20語については表3において網掛けにして示したものである。

系動詞」が V2 として見られることを指摘しており、これらが『夜の寝覚』巻一においても同様に見られるものであることも確認し得た。

一方、両者の相違点についても具体的に見てみると大木（2022）の示す上位 30 位の動詞には見られないものの『夜の寝覚』巻一において上位——具体的には 4 位——に現れる動詞として「弾く」が挙げられる。これは『夜の寝覚』巻一において琵琶や琴を弾く様子が繰り返して描かれることに起因するものと見える<sup>14</sup>。また、『夜の寝覚』巻一においては見られないものの大木（2022）の示す上位 30 語の動詞——具体的には 3 位——に現れるものとしては「参る」が挙げられる。この「参る」が『夜の寝覚』巻一において V2 として現れない理由は定かでないが、物語の展開や他の動詞との関係について更に検討を深める余地のあるものと見える<sup>15</sup>。当然、「V1+テ+V2」の文法的な振る舞いを明らかにするためには複数の資料から多くの用例を収集（し、資料の個別性を捨象することによって得られる資料間の共通性を記述）することが求められるが、個別の資料に着目して当該の資料の特徴を明らかにすることも同様に重要であり、そのためには両者の相違点を見ることも必要になると言える。

このように本稿において構築を進めている『夜の寝覚』のコーパスは CHJ 平安時代編と同様に形態論情報を認定していることから索引の利用や文字列検索に比して品詞や活用形を指定した検索を容易にするものであると同時に CHJ 平安時代編を対象とした調査結果と比較することをも可能にするものであると言える。

## 6. おわりに

本稿では中古和文資料『夜の寝覚』のコーパスを構築する試みについて述べてきた。この『夜の寝覚』のコーパスが完成すれば、中古語を対象とした語彙や文法に関する定性的研究と共に宮島（1970）や小林・小木曾（2013）などに類する定量的研究も可能になり、未だに定説を見ない『夜の寝覚』の成立に関する定量的な分析を実施するための環境も整うものと考えられる<sup>16</sup>。これに関連して『夜の寝覚』の異なる底本間の比較や『夜の寝覚』と改作本と言われる『夜寝覚物語』との比較を実施することも展望として挙げられる<sup>17</sup>。

<sup>14</sup> そもそも『夜の寝覚』巻一において「弾く」は 22 例であり、巻一における動詞の合計 4,472 例のうち 0.49% を占める動詞である。一方、CHJ 平安時代編所収の和文資料において「弾く」は 133 例であり、動詞の合計 187,962 例のうち 0.07% を占める動詞である。このことから『夜の寝覚』巻一では他の中古和文資料に比して「弾く」が多く用いられている可能性が考えられる。

<sup>15</sup> ただし、試みに『夜の寝覚』巻二を見てみると V2 として現れる「参る」が数例は認められることから『夜の寝覚』に全く用例がないとは言えず、巻一に用例がないのは偶発的なものである可能性もある。

<sup>16</sup> 既に『夜の寝覚』の作者については定量的な分析を試みた北原（2015）の研究があるが、『日本古典文学大系』の本文を用いているという点において本稿の構築しているコーパスとは底本が異なることに加えて『夜の寝覚』、『浜松中納言物語』、『更級日記』、『紫式部日記』を比較することに留まっており、本来は他の資料との比較によって相対的に捉えられるはずの資料間の類似性について更なる検討の余地を残すものと考えられる。これに関連して菊池・菅野（2019）では品詞の構成比率の観点から中古和文資料の類似性について韻文と散文との比較や資料間の比較を実施しており、これに『夜の寝覚』を含めて同様に検討することによって明らかになる中古和文資料における位置づけを踏まえて『夜の寝覚』の成立に関する議論を進められる可能性がある。また、いずれも菅原孝標女によって書かれたとも言われる『夜の寝覚』、『浜松中納言物語』、『更級日記』の作者説について検討するに当たっては『浜松中納言物語』のコーパスを構築することも併せて必要になる。

<sup>17</sup> 小木曾（2022:210）も「伝本が複数ある場合には、複数の本文をコーパスとして比較、対照できるように整備することが求められる」とし、同一の資料について底本の異なるコーパスの構築を提案している。

なお、昨今では「研究者各人が自らの研究課題に即したコーパスを用意し分析に利用することも容易になりつつある」（間淵 2020:121）と言われるが、コーパスを自らの手によって用意（構築）することは困難を伴うものと言わざるを得ず、実際にはその「開発を牽引する国立国語研究所」（間淵 2020:114）の手を借りなければ、依然として実現し難い現状にあるものと見える。本稿のような試みの蓄積によって「国語研の外部の研究者であっても歴史的資料のコーパスを作ることができる環境」（小木曾 2022:211）の整備が更に進展することが望まれるところである。

### 謝 辞

本研究は JSPS 科研費 JP22K19985, JP22K19986 の成果の一部である。なお、作業の実施に当たっては黒田優月氏（筑波大学学生）の助力を得た。

### 参 考 文 献

- 青木博史（2020）「動詞連用形＋動詞」から「動詞連用形＋テ＋動詞」へ―「補助動詞」の歴史・再考― 青木博史・小柳智一・吉田永弘（編）『日本語文法史研究 5』, pp.197-226, ひつじ書房.
- 安部清哉・菊池そのみ（2016）「男が泣く日本文学の系譜を探って―男主人公が最も泣く『夜の寝覚』―」『学習院大学国語国文学會誌』 59, pp.76(57)-59(74), 学習院大学文学部国語国文学會.
- 大木一夫（2022）「テ形補助動詞成立史概略、拾遺」青木博史・岡崎友子・小木曾智信（編）『コーパスによる日本語史研究 中古・中世編』, pp.65-87, ひつじ書房.
- 大川孔明（2020）「叙述語から見た平安鎌倉時代の文学作品の文体類型」『計量国語学』 32:6, pp.331-345, 計量国語学会.
- 小木曾智信（2013）「中古仮名文学作品の形態素解析」『日本語の研究』 9:4, pp.49-62, 日本語学会.
- 小木曾智信（2022）『日本語歴史コーパス』 中古・中世のデータ構築について 青木博史・岡崎友子・小木曾智信（編）『コーパスによる日本語史研究 中古・中世編』, pp.199-214, ひつじ書房.
- 小木曾智信（2023）『昭和・平成書き言葉コーパス』の構築・公開と権利処理 日本語学会 2023 年度春季大会予稿集』, pp.145-147, 日本語学会.
- 小椋秀樹（2014）「形態論情報」前川喜久雄（監）山崎誠（編）『書き言葉コーパス―設計と構築―』, pp.68-88, 朝倉書店.
- 菅野倫匡（2021）「芥川賞作品コーパスの構築に向けて―語彙調査に関する未解決の問題との関連から―」『日本語と日本文学』 67, pp.75-88, 筑波大学日本語日本文学会.
- 菅野倫匡・菊池そのみ（2021）「和歌集における品詞の構成比率の算出方法に関する試論―計量語彙論のための標本抽出法を中心に―」『計量国語学』 33:3, pp.162-177, 計量国語学会.
- 菊池そのみ（2019）「古代語の「ての」について」『筑波日本語研究』 23, pp.113-134, 筑波大学人文社会科学研究所日本語学研究室.
- 菊池そのみ・菅野倫匡（2019）「勅撰和歌集の語彙の量的構造をめぐって―品詞の構成比率の観点から―」『国語語彙史の研究』 38, pp.360(57)-336(81), 和泉書院.

- 菊池そのみ・菅野倫匡（2022）「和歌語彙の研究に『日本語歴史コーパス』を利用する際の留意点」『筑波日本語研究』26, pp.37-62, 筑波大学人文社会科学研究所日本語学研究室.
- 北原慈子（2015）「古典文学研究における計量文献学的手法をめぐって—『更級日記』『浜松中納言物語』『夜半の寢覚』『紫式部日記』を題材として—」『Studies in Language Science』5, pp.75-103, 立命館大学大学院言語教育情報研究科.
- 国立国語研究所コーパス開発センター（池上尚）（編）（2016）『『日本語歴史コーパス平安時代編』形態論情報規程集』, 国立国語研究所.
- 小林雄一郎・岡崎友子（2017）「中古における接続表現の統計的分析—指示詞を構成要素とするものを中心に—」『国立国語研究所研究論集』13, pp.65-77, 国立国語研究所.
- 小林雄一郎・小木曾智信（2013）「中古和文における個人文体とジャンル文体—多変量解析による歴史的資料の文体研究—」『国立国語研究所論集』6, pp.29-43, 国立国語研究所.
- 阪倉篤義（1964a）「『夜半の寢覚』の文章」『國語と國文學』41:10, pp.144-156, 東京大学国語国文学会.
- 阪倉篤義（1964b）「解説」阪倉篤義（校注）『日本古典文学大系 78 夜の寢覚』, pp.3-37, 岩波書店.
- 阪倉篤義・高村元継・志水富夫（編）（1974）『夜の寢覚総索引』, 明治書院.
- 鈴木一雄（1993）「『夜の寢覚』と『更級日記』の作者」鈴木一雄『王朝女流日記論考』, pp.258-311, 至文堂.
- 鈴木一雄（1996）「解説」鈴木一雄（校注）『新編日本古典文学全集 28 夜の寢覚』, pp.561-589, 小学館.
- 鈴木弘道（1965）『寢覚物語の基礎的研究』, 塙書房.
- 関根慶子・小松登美（1960）『寢覚物語全釈』, 學燈社.
- 関根慶子・小松登美（1972）『増訂寢覚物語全釋』, 學燈社.
- 関根慶子・小林太枝子・平野由紀子・中沢礼伊子・伊藤弘子・泉民子（1975）「寢覚物語對校」関根慶子教授退官記念会（編）『関根慶子教授退官記念寢覚物語對校・平安文学論集』, pp.1-485, 風間書房.
- 高村元継（1966）『校本夜の寢覚』, 明治書院.
- 辻本桜介（2016）「主節主体の動きを表す動詞終止形に接続するトテについて—引用と異なる機能の分析—」『日本語の研究』12:2, pp.35-51, 日本語学会.
- 永井和子（1960）「『ねざめ』の構造」『平安文学研究』25, 平安文学研究会.
- 橋本佳（1933）『校本夜半の寢覚』, 大岡山書店.
- 平林文雄（2007）「夜の寢覚」飛田良文・遠藤好英・加藤正信・佐藤武義・蜂谷清人・前田富祺（編）『日本語学研究事典』, p.691, 明治書院.
- 文化庁著作権課（2019）『デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方（著作権法第30条の4, 第47条の4及び第47条の5関係）』, [https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30\\_hokaisei/pdf/r1406693\\_17.pdf](https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf)（2023年7月26日確認）, 文化庁.
- 間淵洋子（2020）「数理的研究」『日本語の研究』16:2, pp.114-121, 日本語学会.
- 宮島達夫（1970）「語いの類似度」『国語学』82, pp.42-64, 国語学会.
- 村田菜穂子（2001）「平安時代の形容動詞—〜ゲナリと〜カナリ—」『国語学』52:1, pp.16-30, 国語学会.

山崎誠（2013）「形式語研究の方法論—定性的研究と定量的研究—」藤田保幸（編）『形式語研究論集』, pp.1-18, 和泉書院.

### 資料

- 阿部秋生・秋山虔・今井源衛・鈴木日出男（校注・訳）（1994）『新編日本古典文学全集 25 源氏物語⑥』, 小学館.
- 市古貞次（校注・訳）『新編日本古典文学全集 46 平家物語②』, 小学館.
- 菊地靖彦・木村正中・伊牟田経久（校注・訳）（1995）『新編日本古典文学全集 13 土佐日記・蜻蛉日記』, 小学館.
- 阪倉篤義（校注）（1964）『日本古典文学大系 78 夜の寝覚』, 岩波書店.
- 鈴木一雄（校注・訳）（1974）『日本古典文学全集 19 夜の寝覚』, 小学館.
- 鈴木一雄（校注・訳）（1996）『新編日本古典文学全集 28 夜の寝覚』, 小学館.
- 鈴木一雄・石埜敬子（1984）『完訳日本の古典第二十五卷夜の寝覚（一）』, 小学館.
- 鈴木一雄・石埜敬子（1985）『完訳日本の古典第二十六卷夜の寝覚（二）』, 小学館.
- 日本国語大辞典第二版編集委員会・小学館国語辞典編集部（編）（2000～2002）『日本国語大辞典』第2版, 小学館.
- 藤岡忠美・中野幸一・犬養廉・石井文夫（校注・訳）（1994）『新編日本古典文学全集 26 和泉式部日記・紫式部日記・更級日記・讃岐典侍日記』, 小学館.
- 三谷栄一・三谷邦明・稲賀敬二（校注・訳）（2000）『新編日本古典文学全集 17 落窪物語・堤中納言物語』, 小学館.

### 関連 URL

- |                                |   |
|--------------------------------|---|
| 国立国語研究所『日本語歴史コーパス』             | <a href="https://clrd.ninjal.ac.jp/chj/">https://clrd.ninjal.ac.jp/chj/</a>       |
| 形態素解析器「MeCab」(Ver.0.996)       | <a href="http://taku910.github.io/mecab/">http://taku910.github.io/mecab/</a>     |
| 解析用辞書「中古和文 UniDic」(Ver.202203) | <a href="https://clrd.ninjal.ac.jp/unidic/">https://clrd.ninjal.ac.jp/unidic/</a> |
| プログラミング言語「Python」(Ver.2.7.18)  | <a href="https://www.python.org/">https://www.python.org/</a>                     |

# 『日本語日常会話コーパス』にみる 日常会話音声の基本周波数と談話行為の関係

石本 祐一 (ものづくり大学/国立国語研究所) \*

## Relationships between Dialog Acts and the Fundamental Frequencies of Everyday Speech on the Corpus of Everyday Japanese Conversation

Yuichi Ishimoto (Institute of Technologists/NINJAL)

### 要旨

パラ言語・非言語情報の影響により音声の韻律が様々に変動することが知られているが、日常生活の多様な状況を網羅するような音声データがこれまで存在せず大規模な分析を行うことが困難だった。しかし、多様な種類の日常会話をバランス良く収録した大規模コーパスである『日本語日常会話コーパス』の登場により、日常会話音声の様々な面を観察することが可能となった。そこで、『日本語日常会話コーパス』の音声の分析を通して、日常会話音声において観察される基本周波数の多様性を示す。先の研究において会話場面によって同一話者であっても発話の基本周波数が異なっていることを示したが、本稿では発話の談話機能に着目し、談話行為が基本周波数に与える影響を調べた。分析の結果、フィードバック肯定に分類される発話では基本周波数が低くなる傾向になることがわかった。また、談話行為によっては会話場面とは異なる影響を基本周波数に与える可能性が示唆された。

### 1. はじめに

2022年3月より本公開された『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC)(Koiso et al. 2022)には多様な種類の日常会話がバランス良く収録されており、これにより日常会話の言語活動を多角的に研究できるようになった。例えば、自発音声は発話意図や感情を代表とするパラ言語・非言語情報の影響により韻律が様々に変動することが知られている(森ほか 2014)が、CEJCにより日常の自発音声の多様性について大規模な定量的分析を行うことができる。本稿では、日常会話音声における韻律の多様性を示すことを目的として、声の高さを表す音響特徴量である基本周波数に着目し、CEJCに収録されている自発音声データから発話の談話行為が基本周波数の違いとしてどのように現れるのか調べた結果について報告する。

---

\* y\_ishimoto@iot.ac.jp

表 1 対象調査協力者

協力者 ID	年代	性別	職業	会話場面数	合計発話時間 (秒)
C001	40 代	女性	会社員・公務員等	9	6300.09
C002	50 代	女性	会社員・公務員等	14	6556.12
K001	30 代	女性	会社員・公務員等	12	6807.65
K002	50 代	女性	自営業・自由業	11	6201.68
K003	20 代	女性	大学生	14	6189.69
K004	40 代	女性	パート・アルバイト	12	7462.08
S001	50 代	男性	会社員・公務員等	9	2863.24
T001	30 代	男性	自営業・自由業	11	6789.28
T002	40 代	男性	自営業・自由業	13	7679.76
T003	30 代	女性	専業主婦・主夫	12	5798.67
T004	60 代	女性	専業主婦・主夫	13	7579.69
T006	20 代	男性	大学院生	10	4544.10
T009	20 代	女性	大学生	21	6320.67
T010	20 代	男性	大学生	10	6697.03
T013	60 代	男性	会社員・公務員等	17	6530.13
T014	40 代	女性	自営業・自由業	17	7817.85
T015	50 代	男性	会社員・公務員等	17	5646.89

## 2. データ

### 2.1 コーパス

CEJC は日常生活で自然に生じる会話を収録することが念頭に置かれており、研究者の介入による影響を排除するために調査協力者自身による映像の録画および音声の録音が行われている。調査協力者の家族や友人等はそれぞれの会話場面に応じて登場するが、基本的に調査協力者は必ず参加していることになることから、調査協力者の音声のみを対象とした。さらに、発話相手の属性が一つに同定できる会話場面に限定し、本稿では表 1 に示す調査協力者 17 名の分析を行う。すべての調査協力者において 9 場面以上の異なる会話場面があり、平均して調査協力者一人あたり約 100 分の発話時間となっている<sup>(1)</sup>。

また、CEJC には談話行為情報として発話単位ごとに談話行為タグ (Iseki et al. 2019) が付与されており、本稿では談話行為情報レベル 1 タグ (基本的な談話機能に関わる情報) を利用することとする。分析対象となる発話と談話行為との対応を表 2 に示す。談話行為ごとに発話数にはばらつきがあり、「注意獲得」や「注意獲得への対処」のように対応する発話が存在しない談話行為もある。

<sup>(1)</sup> CEJC では日常生活の自然な振る舞いを妨げるヘッドセットマイク等の利用は避けて各参加者が IC レコーダを胸の位置に配置して音声を録音しているため、収録状況や参加者間の位置関係によって他の話者の音声や雑音が入り込んでしまうことがあり、すべての発話時間が分析に利用できるわけではない。

表 2 談話行為情報（レベル 1 タグ）と対象発話数

タグ	談話機能	発話数	
T_Inform	タスク系	情報提供	1830
T_Question		情報要求	291
T_Answer		返答としての情報・確認提供	418
T_Request		依頼系	191
T_A-Request		依頼系への対処	167
T_Offer		申し出	19
T_A-Offer		申し出への対処	35
T_Attention		注意獲得	0
T_A-Attention		注意獲得への対処	0
T_SelfSpeech		独り言	158
S_Greeting	社会的付き合い管理系	挨拶	7
S_Apology		謝罪	20
S_A-Apology		謝罪への対処	7
S_Thanking		感謝	29
S_A-Thanking		感謝への対処	2
FB_Positive	フィードバック系	フィードバック肯定	3193
FB_Acceptance		フィードバック了承	7
FB_Repetition		フィードバック反復	126
FB_Lexical		フィードバック語彙的反応	31
O_NV	その他	非言語行動	14
O_Cue		合図	89
O_Undet		不明	67

## 2.2 基本周波数抽出

CEJC には基本周波数 ( $F_0^{(2)}$ ) の生データの情報は付与されていない。そこで、音声分析合成システム STRAIGHT の音源情報分析関数 (Kawahara et al. 2017) を用い、音声データから 1 ms 間隔で  $F_0$  推定を行った。なお、有声/無声判定は STRAIGHT から得られる非周期性情報を利用した。さらに、推定エラーの影響を低減させるために、推定された  $F_0$  のうち上位 5% と下位 5% の値を取り除いた。最終的に、性差を正規化するため各話者ごとの平均値と標準偏差を用いた z-score に変換した。

## 3. 日常会話音声の多様性

### 3.1 会話相手と基本周波数の関係

これまでに調査協力者から見た会話相手の属性と調査協力者の発話の平均  $F_0$  との関係調べており、詳細は Ishimoto and Koiso (2019) を参照されたいが図 1 に示すような関係が観察された。すなわち、発話の平均  $F_0$  は、子ども・配偶者・父母といった近親者に対して低くなる、友人に対しての発話では高くなる、同僚・取引先・客に対しては友人よりは低く近親者よ

<sup>(2)</sup> 近年ではフォルマント周波数 ( $F_1, F_2, F_3, \dots$ ) との意味的な違いを明確にするために基本周波数を  $F_0$  と表記することもある (Titze et al. 2015) が、本稿では従来から主に使用されてきた  $F_0$  という表記を用いる。



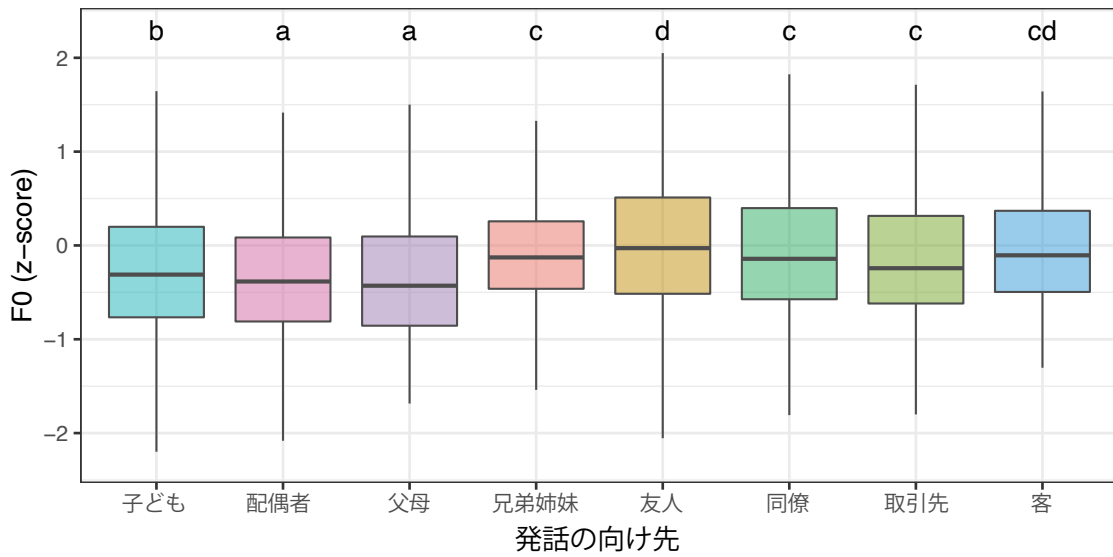


図 1 会話相手の属性に対する発話の平均 F0 (Ishimoto and Koiso (2019) より再掲)

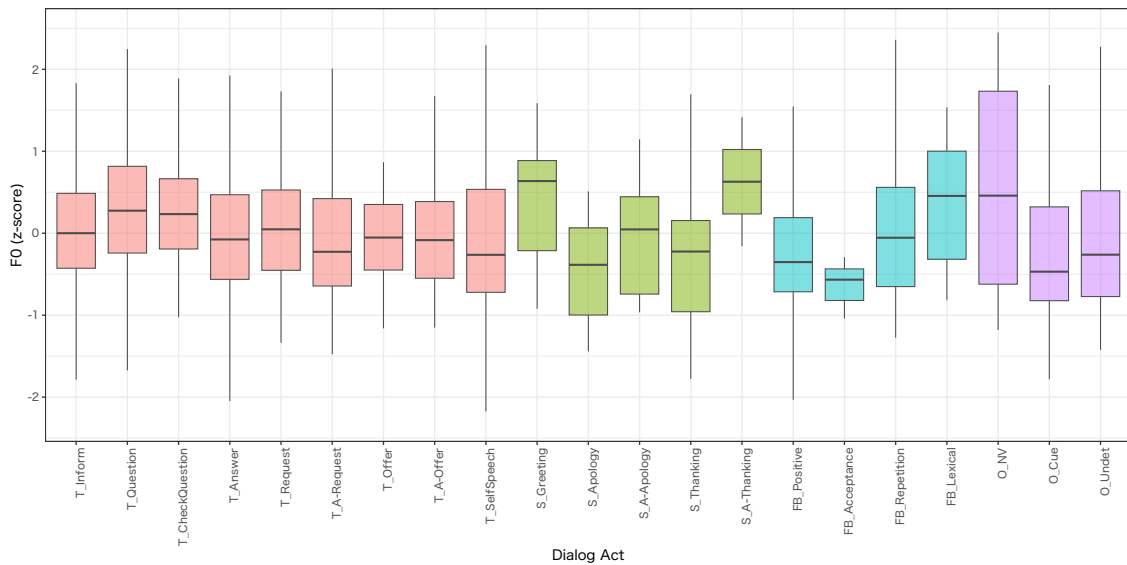


図 2 談話行為情報に対する発話の平均 F0

りは高くなる、といったように会話相手によって異なっていた。このように、会話場面や発話対象によって発話の F0 に変動が生じていることが示された。

そこで次節では、さらに発話の韻律の多様性を示すため、談話行為と F0 との関係について調べる。

### 3.2 談話行為と基本周波数の関係

発話の平均 F0 と表 2 に示した談話行為情報との関係を図示したものが図 2 である。

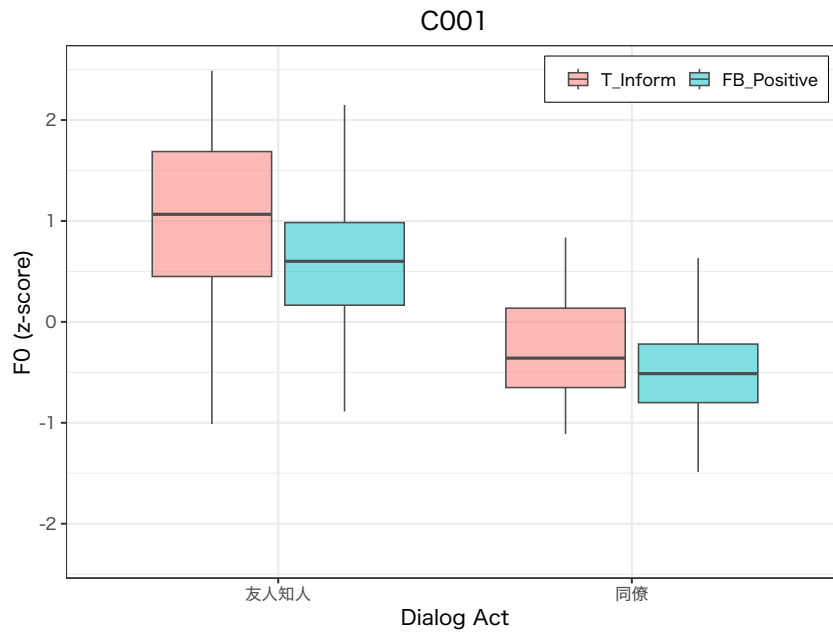


図3 C001 の談話行為情報と F0 との関係

談話行為を「タスク系」や「社会的付き合い管理系」、「フィードバック系」といった大きな括りで見たととき、F0のの違いに定性的な傾向は見られなかった。また、談話行為タグを個別に見ると、S.Greeting（挨拶）やS\_A-Thanking（感謝への対処）は他よりも高く、FB\_Acceptance（フィードバック了承）は他よりも低い、というような分布になっているが、これは対象となる発話数が少ないことが原因である可能性が大きい。

そこで、発話数が十分にある談話行為に着目し、T\_Inform（情報提供）とFB\_Positive（フィードバック肯定）に限定して、個人ごとに違いを観察する。加えて、図1にて示されたような会話相手の属性による違いも合わせて調べることにした。

調査協力者C001におけるT\_InformとFB\_Positiveの発話の平均F0を図3に示す。C001では、T\_InformよりもFB\_Positiveの方がF0が低くなる傾向が見られた。また、同じ談話行為でも会話相手の属性によって高さが変わっているが、T\_InformとFB\_Positiveの違いはそれぞれの属性で同程度であった。これは、会話場面により発話のF0のベースラインが変わる可能性を示唆していると言える。

次に、調査協力者K004におけるT\_InformとFB\_Positiveの発話の平均F0を図4に示す。K004においてもC001と同様、T\_InformよりもFB\_Positiveは低くなる傾向が見られた。しかし、「子ども」に対する発話ではT\_InformとFB\_Positiveに大きな違いは見られず、談話行為による違いが誰が相手でも生じるとは言えない結果となっている。

最後に、調査協力者T002におけるT\_InformとFB\_Positiveの発話の平均F0を図5に示す。T002では、「取引先」に対する発話では談話行為によるF0の違いが見られるが、「配偶者」に対する発話ではあまりF0は変わらない。これもK004同様、会話相手の属性により談

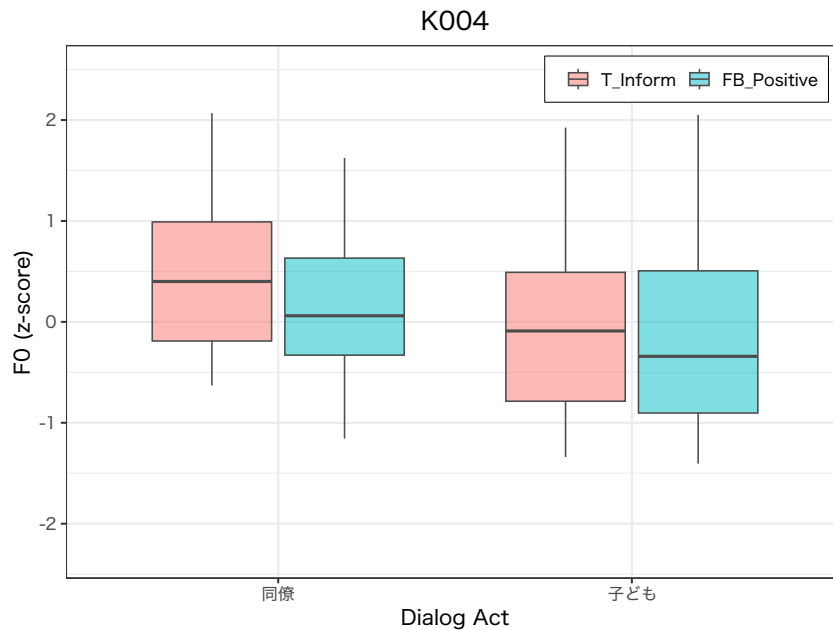


図4 K004 の談話行為情報と F0 との関係

話行為による違いが生じない場合の存在を示唆している。また、先の研究（図1）で見られた「配偶者」に対する発話の F0 が「取引先等」相手よりも低くなるという傾向が、T002 の FB\_Positive の発話では見られないという結果になっている。

### 3.3 議論

3.2 節の結果から、十分多い発話数がデータから得られた談話行為に関してのみの検証にとどまったが、FB\_Positive では発話の平均 F0 が低くなりやすいという傾向が見られた。

また、先の研究（図1）において会話相手の属性（会話場面）によって発話の F0 が異なることを示していたが、今回の調査協力者 K004 や T002 の結果からは発話の談話機能によっては会話相手の属性が発話の F0 に与える影響が小さいことが示された。これは、会話相手の属性よりも談話機能の方が発話の F0 に与える影響が強いことを示しているともいえるが、一方で調査協力者 C001 のように会話相手の属性と発話の談話機能の影響を両方ともそのまま反映した発話となっている場合もある。このような個人による違いがどのような要因によるものかはさらなる検証が必要である。加えて、T\_Inform と FB\_Positive 以外の談話行為に関しては、今回対象とした発話の数では明確な傾向を述べるのが難しく、今後さらにデータ数を増やした検証を行う予定である。

## 4. おわりに

日常会話音声における基本周波数の多様性を示すために、『日本語日常会話コーパス』(CEJC)に含まれている音声の発話の平均 F0 を、発話の談話機能と会話相手の属性の面から調べた。その結果、フィードバック肯定に該当する発話の基本周波数が低くなりやすい傾向が見られた。

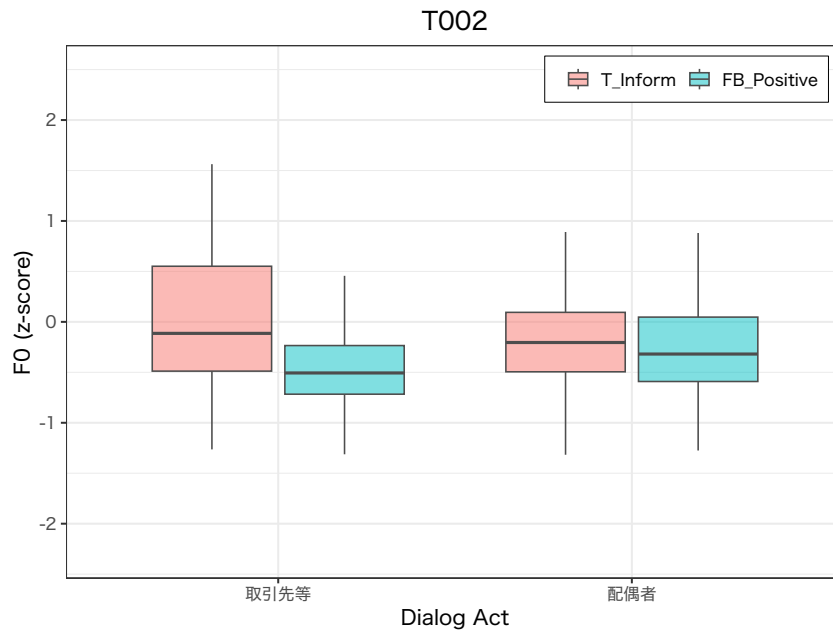


図5 T002の談話行為情報とF0との関係

また、個人によって談話機能が基本周波数に与える影響の大きさが異なり、先の研究によって示されていた会話相手の属性による影響も発話の談話機能によって現れない可能性が示唆された。しかし、情報提供やフィードバック肯定以外の多くの談話機能の影響は今回の分析においてはデータ数が少なく明確な傾向を掴むことができなかったため、今後の課題としたい。

#### 謝 辞

本研究は、国立国語研究所 共同研究プロジェクト「多世代会話コーパスに基づく話し言葉の総合的研究」、および JSPS 科研費 18K11514 の助成を受けたものです。

#### 文 献

- Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe (2022). "Design and Evaluation of the Corpus of Everyday Japanese Conversation." *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5587–5594. Marseille, France: European Language Resources Association.
- 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』音響サイエンスシリーズ/日本音響学会編:12 コロナ社.
- Yuriko Iseki, Keisuke Kadota, and Yasuharu Den (2019). "Characteristics of everyday conversation derived from the analysis of dialog act annotation." *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 1–6.

Ingo R Titze, Ronald J Baken, Kenneth W Bozeman, Svante Granqvist, Nathalie Henrich, Christian T Herbst, David M Howard, Eric J Hunter, Dean Kaelin, Raymond D Kent, Jody Kreiman, Malte Kob, Anders Löfqvist, Scott McCoy, Donald G Miller, Hubert Noé, Ronald C Scherer, John R Smith, Brad H Story, Jan G Švec, Sten Ternström, and Joe Wolfe (2015). “Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization.” *The Journal of the Acoustical Society of America*, 137:5, pp. 3005–3007.

Hideki Kawahara, Ken-Ichi Sakakibara, Masanori Morise, Hideki Banno, and Tomoki Toda (2017). “Accurate estimation of  $f_0$  and aperiodicity based on periodicity detector residuals and deviations of phase derivatives.” *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1556–1564.

Yuichi Ishimoto, and Hanae Koiso (2019). “Prosodic diversity according to relationship among participants in everyday Japanese conversation.” *Proceedings of LPSS 2019*, pp. 62–66.

#### 関連 URL

『日本語日常会話コーパス』

<https://www2.ninjal.ac.jp/conversation/cejc.html>

音声分析合成システム STRAIGHT

[https://github.com/HidekiKawahara/YANGstraight\\_source](https://github.com/HidekiKawahara/YANGstraight_source)

## 膠着型言語における複雑さのトレードオフ

李 文超 (浙江大学) †

### Complexity trade-off in agglutinative languages

Wenchao Li (Zhejiang University)

#### 要旨

本研究は 9 つの言語族にわたる 16 の膠着言語の 25 のツリーバンクを利用し、形態的豊かさと語順の柔軟性との相関関係を探った。形態的豊かさはパラダイムの移動平均形態的豊富さと移動平均サイズで測った。語順の柔軟性はコサイン類似度とエントロピーで計測した。統計分析の結果、形態素が豊富であればあるほど、語順がより柔軟になる。形態と語順の間にはかなり強い相関関係があることが確認され、複雑さのトレードオフ仮説を裏付けている。9 つの言語族の中で、オーストロネシア語、アフロアジア語、ドラヴィダ語は、アルタイ語、インド・アーリア語、ウラル語に比べて形態的にも語順的にも多様性が低いと判明した。トルコ・ウイグル・バスクおよび北サーミ語は、S、V、O の組み合わせの割合が最もバランスが取れていることが観察された。同じ言語族内に形態・語順の分離が窺えた：ウラル語族はフィン語派がフィン・ウゴル語派より柔軟であり；アルタイ語族はモンゴル語派がチュルク語派より厳格である。形態—統語の一端から言語間距離が如何に外国語習得に左右するかを考察して結果、目標言語の語順自由度の高い言語の場合、母語が形態的に豊富であれば産出した目標言語の語順が自由になると結びついた。

#### 1. はじめに

人間の言語は、複雑かつ動的で、階層的に組織されている規則的なシステムである (Fenk-Oczlon and Pilz 2021)。その規則性の 1 つは *complex trade off* である。言語内の 1 つの変数 (音韻論、形態論、構文、意味論など) が洗練されている場合、別の変数が単純化される傾向がある。言語がほぼ同じ程度の複雑さを表現できるように保つ (Menzerath's law 1954; Shosted 2006; Sinnemäki 2014; Fenk-Oczlon and Fenk 2014)。複雑さのトレードオフ仮説は Sapir (1921) 以来さまざまな言語変数間で確認されていた：音素-音節-単語 (Coloma 2017)、音韻-形態論 (Shosted 2006)、形態-構文 (Jakobson 1936; McFadden 2003; Kopleinig 2017; Yan と Liu 2021、Li, Liu, Xiong 2022)。トレードオフの考え方における重要な問題点は、トレードオフがどの程度成立するか、どの言語変数が多数の変数の中、重要な役割を果たすかという点にある。最初の問題点に関して、ほとんどの研究は、トレードオフは一部の参加者によるものであると考えられている。例えば、単語あたりの音節が多いほど、音節あたりの音素は少なくなり；音節あたりの音素が多いほど、形態的ケースは少なくなり；形態的ケースが増えるほど、語順の自由度が高まる (Fenk-Oczlon and Fenk 1999; Shosted 2006; Sinnemäki 2008, 2014; Miestamo 2009; Kopleinig et al. 2017)。2 番目の問題点に関しては、Fenk-Oczlon・Pilz (2021) は多言語の翻訳の並列データを調査した結果、音素サイズ、音節サイズ、単語の長さ、文節の長さとして人口の間の相関関係が確認できた。多変数の中、音節の複雑さが最も重要であると主張した。

---

† widelia@zju.edu.cn

本研究はトークン化、見出し語化、POS および形態的特徴のタグ付け、係り受け解析を担う自然言語処理のツールキットを使い、自然言語の形態的多様性と語順の柔軟性との相互作用を洞察する。バスク語、北サーミ語、エストニア語、フィンランド語、ハンガリー語、日本語、マラーティー語、タミル語、テルグ語、トルコ語、ウイグル語、カザフ語、ブリヤート語、ウオロフ語、インドネシア語、コプト語の 16 の膠着語を対象とし、ユニバーサル依存関係からの 25 の注釈付きコーパスを分析する。形態的豊かさは、移動平均形態的豊かさ (MAMR: Čech and Kubát 2018) とパラダイムの移動平均サイズ (MAMSP: Xanthos and Gillis 2010) の 2 つの指標によって測定される。語順の柔軟性の把握にコサイン類似度 (COS) とエントロピー (ENTR: Shannon 1948; Chen et al. 2016, Bentz et al. 2017; Yan and Liu 2021; Li et al. 2022) が使用される。スピーアマンの順位相関係数は変数の間の関連性を測定する。

## 2. 研究方法

### 2.1 データセット

以下のツリーバンクが採用される: a). 4 つのウラル言語: エストニア語 (2 つのツリーバンク)、フィンランド語 (2 つのツリーバンク)、北サーミ語 (1 つのツリーバンク)、ハンガリー語 (1 つのツリーバンク); b). 4 つのアルタイ語: ブリヤート語 (1 つのツリーバンク)、カザフ語 (1 つのツリーバンク)、トルコ語 (2 つのツリーバンク)、ウイグル語 (1 つのツリーバンク); c). 2 つのドラヴィダ語: タミル語 (2 つのツリーバンク)、テルグ語 (1 つのツリーバンク); d). インド・アーリア語のマラーティー語 (1 つのツリーバンク); e). アフロアジア系エジプト語コプト語 (1 つのツリーバンク); f). ニジェール・コンゴ大西洋言語ウオロフ語 (1 つのツリーバンク); g). 1 つのバスク語 (1 つのツリーバンク); h). オーストロネシア語 1 つ: インドネシア語 (3 つのツリーバンク); 日本語(三つのツリーバンク)。表 1 に、ツリーバンクの詳細を示します。

本研究が扱うツリーバンクはテキストサイズとテキストタイプが異なっている (ブログ、電子メール、フィクション、文法例、法律、ニュース、レビュー、ソーシャル、ウェブ、ウィキ)。そこで、コーパスのサイズが結果に影響を与えるかどうかという疑問が生じた。コーパスサイズの影響を最小限に抑えるため、以下の作業を行った。コーパスの中で最小のコーパスはマラーティー語、3847 単語である故、コーパスを 4 つのサブコーパス: 3000 トークン、9000 トークン、12000 トークン、および完全トークンにサブセットする。次に、各サブコーパスの 5 つのメトリクス値を計算し、平均値を使う。さらに、本稿は主文と、S、O、V からなる副文の両方を調査し、平均値を取ることにする。

表 1. データセット

Treebanks	Text types	Words	Sentences	Treebanks	Text types	Words	Sentences
<b>Basque-BDT</b>	News	121,443	8993	<b>Marathi-UFAL</b>	Wiki, fiction	3,847	466
<b>Buryat-BDT</b>	Fiction, grammar-examples, news	10,185	927	<b>Indonesian-PUD</b>	News, wiki	19,446	1,000
<b>Japanese-BCCWJ</b>	Fiction, news, blog, conference, nonfiction	1,253,903	57,109	<b>Indonesian-GSD</b>	Blog, news	122,019	5,598
<b>Japanese-GSDLUW</b>	blog, news	150,243	8,100	<b>Wolof-WTB</b>	bible, wiki	44,258	2,107
<b>Japanese-PUD</b>	news, wiki	28,788	1,000	<b>Uyghur_UDT</b>	Fiction	40236	3,456
<b>Tamil-TTB</b>	News	9,581	600	<b>Indonesian-CSUI</b>	News, nonfiction	28,263	1,030
<b>Tamil-MWTT</b>	News	2,584	534	<b>Wolof-WTB</b>	Bible, wiki	44,258	2,107
<b>Telugu-MTG</b>	Grammar examples	6,465	1,328	<b>Coptic-Scriptorium</b>	Bible, fiction, nonfiction	55,858	2,163
<b>Buryat-BDT</b>	Grammar examples, news, fiction	10,185	927	<b>Estonian-EDT</b>	Fiction, academic, news, nonfiction	438,245	30,968
<b>Kazakh-KTB</b>	News, fiction, wiki	10,536	1,078	<b>Finnish-TDT</b>	Fiction, legal, news, blog, grammar-examples,	202,453	15,136
<b>Turkish-Kenet</b>	News, nonfiction	183,555	16,396	<b>Finnish-TDT</b>	Poetry, medical, social, web	19,382	2,122
<b>Turkish-Boun</b>	News, nonfiction	125,212	9,761	<b>North Sami-Giella</b>	News, nonfiction	26,845	3,122
<b>Hungarian-Szeged</b>	News	42,032	1,800				



## 2.2 指標

### 2.2.1 形態的豊かさ

コーパスサイズの影響を最小限に抑えるために、テキストのサブセットのタイプトークン比 (TTR) インデックスと平均を繰り返し計算する。各言語テキストの形態学的多様性を測定する。Covington と McFall (2010) の移動平均 TTR (MATTR) を使用して、語形と補題語彙の豊かさを計算する。MATTR (W)<sub>word form</sub> は次の式で求められる。

$$\text{MATTR (W)}_{\text{word form}} = \frac{\sum_{i=1}^{N-W+1} F_i}{W (N - W + 1)}$$

N はテキストのサイズ、W はランダムに選択されたウィンドウのサイズである。F<sub>i</sub> は、特定のウィンドウサイズの拡張形式の数値である。日本語やトルコ語などの膠着語には 2 種類の語形がある。一つは語根の膠着 (さびしい→さびしかった)、もう一つは派生 (さびしい→さびしみ)。屈折言語に関して、語形には屈折 (kind → kinder) と派生 (kind → kindness) がある。本研究のウィンドウサイズは 500 ワードにする。MATTR (W)<sub>word form</sub> は共役によって得られる各ウィンドウの TTR である。各ウィンドウの TTR は、同様の式によって取得される。

$$\text{MATTR (W)}_{\text{word lemma}} = \frac{\sum_{i=1}^{N-W+1} L_i}{W (N - W + 1)}$$

N はテキストのサイズ、W はランダムに選択されたウィンドウのサイズである。L<sub>i</sub> は、一定のウィンドウ サイズの拡張形式の数値である。MATTR (W)<sub>word lemma</sub> は共役による各ウィンドウの TTR である。MAMR は、語形の多様性と補題の多様性の違いを指す。次の方法で取得される。

$$\text{MAMR (W)} = \text{MATTR (W)}_{\text{word form}} - \text{MATTR (W)}_{\text{word lemma}}$$

MAMR (W) が高いほど、言語の形態が豊かになる。語彙の多様性を捉えるためのもう 1 つの尺度は、パラダイムの平均サイズ (MSP、Xanthos and Gillis 2010) である。次の操作で取得する。まず、さまざまな語形変化の数を補題の数で割る。

$$\text{MSP} = \frac{F}{L}$$

MAMSP を次のように取得する。

$$\text{MAMSP} = \frac{\sum_{i=1}^{N-W+1} \frac{F_i}{L_i}}{W (N - W + 1)}$$

MAMP と MAMSP の値が高いほど、言語の形態が豊かになる。

### 2.2.2 語順の柔軟性

語順に関する言語構成要素は約 19 個がある (角田 2009 [1991])。本研究では、以下の理由によって、語順を主語 S、目的語 O、動詞 V の順に絞り込んだ: a). S、V、O 構成要素は自然言語の最も基本的な類型の特徴であり、大部分の文に存在している。したがって、言語間の比較に最適である; b). S、O、V 構成要素は元の注釈スキームに関係なく、すべてのツリーバンクで比較的簡単に識別できる (Kubon et al. 2016)。以上を照らし、考えられる語順パターンとしては、SVO、OVS、VSO、VOS、SOV、OSV の 6 つになる。語順の自由度が最も高い場合は、6 つのパターンが均等に分布する。即ち、各パターンの理想ベクトルは等しい: SVO:

0.166、OVS: 0.166、VSO: 0.166、VOS: 0.166、SOV: 0.166、OSV: 0.166 となる。t が期待値、s が観測値であると仮定すると、COS は次の式を使用して取得される:

$$\text{COS}(s, t) = \frac{\sum_{i=1}^n s_i t_i}{\sum_{i=1}^n s_i^2 \cdot \sum_{i=1}^n t_i^2}$$

COS(s, t) が高くなるほど、サンプルの語順の自由度が高くなる。ENTR は次の方法で取得される。

$$\text{ENTR} = -\sum_{i=1}^n t_i \times \ln t_i$$

特定の言語のエントロピーが高くなるほど、語順の自由度が高くなる。

### 3. 結果

#### 3.1 指標の信頼性

言語変数がどのように相互作用するかという研究課題に取り組む前に、本研究の指標が複数の言語的特徴を示すのに有効であるかどうかを確認しておく。形態的豊かさから始まる。ツリーバンクの MAMR、MAMSP 平均値を計算し、降順に表 2 にまとめた。

表 2. MAMR, MAMSP の値

Language family	Branch	Languages	Morphology	MAMR	MAMSP
Altaic	Turkic	Uyghur	Agglutinative	0.2418	1.4785
Altaic	Turkic	Kazakh	Agglutinative	0.1942	1.3341
Indo-European	Indo-Aryan	Marathi	Agglutinative	0.1678	1.4344
Uralic	Finnic	Finnish	Agglutinative	0.1673	1.2985
Altaic	Mongolic	Buryat	Agglutinative	0.1668	1.2700
Basque	Basque	Basque	Agglutinative	0.1569	1.2416
Uralic	Finnic	Estonian	Agglutinative	0.1406	1.2503
Indo-European	Iranian	Kurmanji	Agglutinative	0.1381	1.3164
Armenian	Armenian	Armenian	Agglutinative	0.1345	1.2518
Uralic	Finno-Ugric	North Sami	Agglutinative	0.1303	1.2280
Altaic	Turkic	Turkish	Agglutinative	0.1089	1.3600
Dravidian	Dravidian	Tamil	Agglutinative	0.1046	1.2474
Dravidian	Dravidian	Telugu	Agglutinative	0.1043	1.2466
Afroasiatic	Egyptian	Coptic	Agglutinative	0.1001	1.2001
Niger-Congo	Atlantic	Wolof	Agglutinative	0.0977	1.2545
Uralic	Finno-Ugric	Hungarian	Agglutinative	0.0707	1.1094
Austronesian	Malayo-Polynesian	Indonesian	Agglutinative	0.0509	1.0829
Japanese	Japanese	Japanese	Agglutinative	0.0253	1.0488

MAMR と MAMSP の値が高いほど、形態的複雑さがより豊かであることが示されている。形態的豊かさの最も高い値を示す上位 5 言語のうち、三つがアルタイ語族に属す: ウイグル語 (チュルク語派)、カザフ語 (チュルク語派)、ブリヤート語 (モンゴル語派)。そして、ウラル語族のフィン語派 (エストニア語とフィンランド語) は、フィン・ウゴル語派 (北

サーミ語とハンガリー語) よりも形態が多様であった。アルタイ語族では、モンゴル語派がチュルク語派よりも厳格であった。インド・アーリア語族において、マラーティー語が唯一の膠着語であり、形態的豊かさは度合いが他のインド・アーリア語、すなわち融合型のウルドゥー語とヒンディー語より高い。インドネシア語と日本語は、より低い値でランク付けされている。日本語に形態的に似ていると言われているドラヴィダ語族のタミル語は中程度の形態的多様性を見せた。

形態的豊かさを反映する MAMR と MAMSP の一貫性と信頼性をさらに調べるには、図 3 に回帰直線を使用した散布図をプロットした。

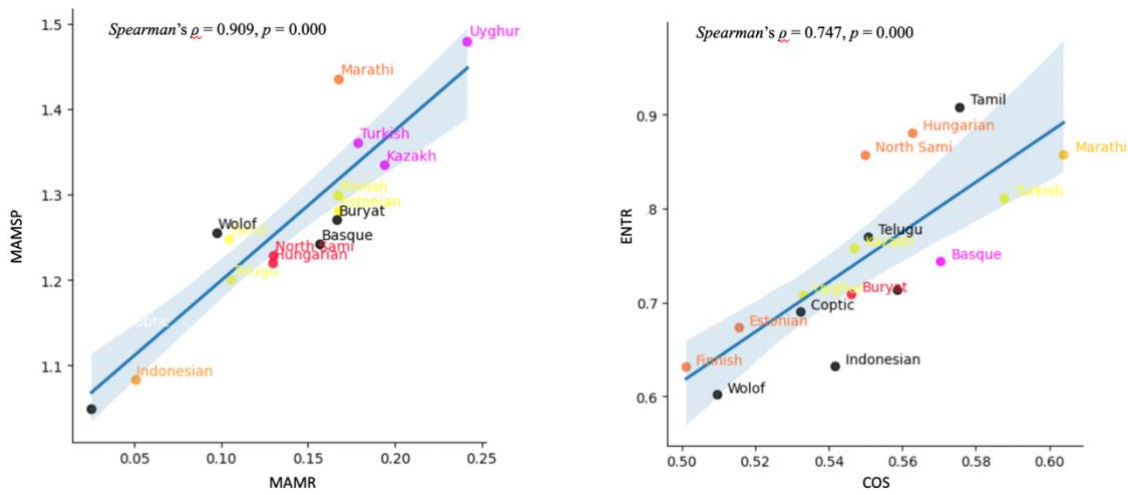


図 1. MAMR、MAMSP; COS、ENTR による散布図と回帰直線

図 1 が示すように、MAMR と MAMSP の関係は回帰直線に一致している。MAMR と MAMSP の間のスピアマンの順位相関係数は正で統計的に有意であった ( $\rho = 0.909$  および  $p = 0.000$ )。これは、2 つの指標が形態的多様性を反映できることが窺えた。

次に、語順に移る。表 3 は、ENTR と COS (語順の柔軟性を表す指標) の平均値を降順に示している。COS と ENTR の値が大きいほど、語順の自由度が高くなる。

表 3. COS と ENTR の値

Language family	Branch	Languages	COS	ENTR
Indo-European	Iranian	Kurmanji	0.7987	1.3791
Armenian	Armenian	Armenian	0.7962	1.0033
Uralic	Finno-Ugric	North Sami	0.7943	1.0021
Uralic	Finnic	Finnish	0.7788	1.3202
Niger-Congo	Atlantic	Wolof	0.7680	1.4246
Afroasiatic	Egyptian	Coptic	0.7289	0.9546
Austronesian	Malayo-Polynesian	Indonesian	0.6728	1.1602
Basque	Basque	Basque	0.6136	1.0293
Altaic	Mongolic	Buryat	0.6070	0.9685
Uralic	Finno-Ugric	Hungarian	0.5989	1.0001
Dravidian	Dravidian	Tamil	0.5963	0.8732

Dravidian	Dravidian	Telugu	0.5507	0.7992
Altaic	Turkic	Kazakh	0.5241	0.7869
Indo-European	Indo-Aryan	Marathi	0.5105	0.7813
Altaic	Turkic	Turkish	0.5081	0.6888
Altaic	Turkic	Uyghur	0.5081	0.7179
Japanese	Japanese	Japanese	0.4877	0.5400
Uralic	Finnic	Estonian	0.4773	0.5821

柔軟性の上位 5 言語は、クルマンジー語、アルメニア語、北サーミ語、フィンランド語とウォロフ語である。ドラヴィダ語、バスク語、アフリカアジア語は中級レベルで見つかった。日本語とエストニア語は語順が最も厳格と観察された。図 2 と表 4 は、各言語の順序の割合を詳しく示している。

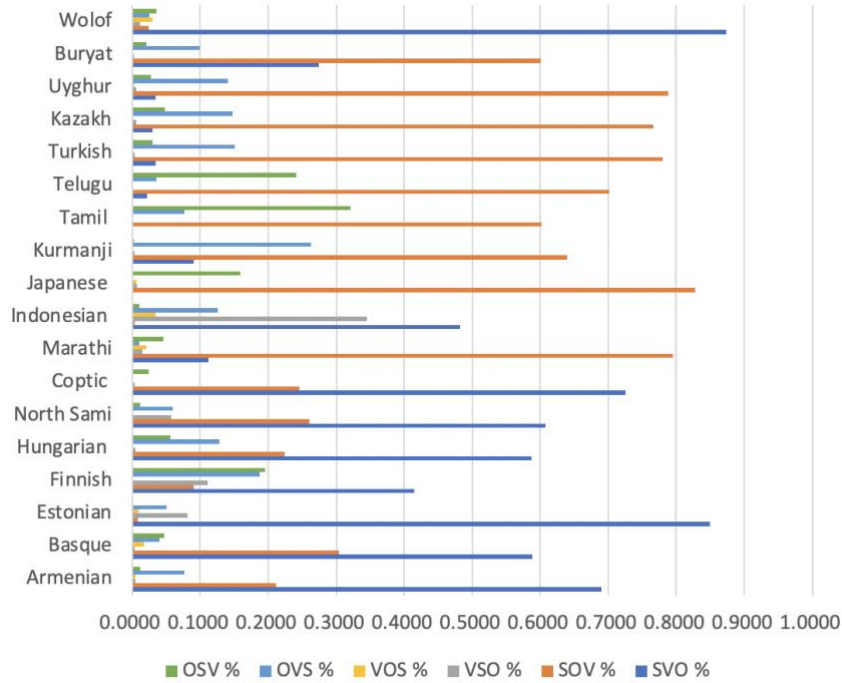


図 2. 各言語の順序の割合

表 4. 語順パターンの分布

Language family	Branch	Language	SVO %	SOV %	VSO %	VOS %	OVS %	OSV %
Indo-European	Armenian	Armenian	0.6895	0.2120	0.0044	0.0049	0.0770	0.0122
Basque	Basque	Basque	0.5884	0.3039	0.0034	0.0170	0.0397	0.0476
Uralic	Finnic	Estonian	0.8497	0.0088	0.0814	0.0091	0.0510	0.0000
Uralic	Finnic	Finnish	0.4147	0.0906	0.1116	0.0000	0.1877	0.1954
Uralic	Finno-Ugric	Hungarian	0.5869	0.2243	0.0047	0.0000	0.1280	0.0561
Uralic	Finno-Ugric	North Sami	0.6075	0.2607	0.0579	0.0020	0.0599	0.0119
Afroasiatic	Egyptian	Coptic	0.7258	0.2460	0.0040	0.0000	0.0000	0.0242
Indo-European	Indo-Aryan	Marathi	0.1123	0.7944	0.0156	0.0211	0.0111	0.0455

Austronesian	Malayo-Polynesian	Indonesian	0.4818	0.0032	0.3451	0.0343	0.1254	0.0102
Japanese	Japanese	Japanese	0.0000	0.8277	0.0076	0.0060	0.0000	0.1587
Indo-European	Iranian	Kurmanji	0.0910	0.6391	0.0033	0.0015	0.2630	0.0021
Dravidian	Dravidian	Tamil	0.0000	0.6020	0.0000	0.0000	0.0772	0.3208
Dravidian	Dravidian	Telugu	0.0220	0.7010	0.0000	0.0000	0.0360	0.2410
Altaic	Turkic	Turkish	0.0343	0.7798	0.0042	0.0010	0.1510	0.0297
Altaic	Turkic	Kazakh	0.0301	0.7661	0.0056	0.0020	0.1477	0.0485
Altaic	Turkic	Uyghur	0.0344	0.7886	0.0058	0.0021	0.1412	0.0279
Altaic	Mongolic	Buryat	0.2745	0.6009	0.0023	0.0017	0.1001	0.0205
Niger-Congo	Atlantic	Wolof	0.8734	0.0240	0.0115	0.0302	0.0255	0.0354

マラーティー語は語順に最も高い柔軟性を示した。イラン語派のクルマンジーは 2 番目の柔軟性を示し、SOV と OVS がかなりの割合が見せた。フィンランド語、ハンガリー語は、S、V、O の組み合わせの最もバランスのとれた比率を示した。ウラル族もすべての語順可能性を示したが、SVO では他の語順を大幅に上回っており、比率のバランスが悪く、値が低くなった。バスク語は本質的に凝集性の形態であり、バランスの取れた順序分布が見出されている。アルタイ語族はあらゆる順序の可能性を示すが、SOV 順序が好まれた。チュルク語派（トルコ語、カザフ語、ウイグル語）の順序の好みは、SOV (78%) > OVS (15%) > SVO (3%) > OSV になっている。(4%) > VSO (1%) > VOS (0%)。モンゴル系ブリヤート語は、トルコ系と比較して、SVO の割合が著しく高く (27.45%)、OVS の割合が低いと判明した。アフロアジア語族エジプト語派のコプト語に語順の余地が大きい。ドラヴィダ語族のテルグ語とタミル語は、一見中程度に厳格で、OV が好まれていた (SOV > OSV > OVS)。図 1 は COS と ENTR の間の相関関係を視覚化したもので、回帰直線は COS と ENTR の間の関係によく適合している。COS と ENTR の間のスピアマンの順位相関係数は正で統計的に有意であった ( $\rho = 0.747$  および  $p = 0.000$ )。これは、語順に関する 2 つの指標は信頼性があり、語順の自由度を把握するのに有効であると示した。

NLP ツールキット、ツリーバンク、計算および数学的分析に基づいて、本研究の指標は、形態的および構文的な多様性を捉える上で信頼できることが確認された。こうして、最初の研究上の疑問が解けた。

### 3.2 形態と構文の相関関係

前のセクションでは、形態的および構文的な複雑さを捉えるための 4 つの指標 (MAMR、MAMSP、COS、ENTR) の有効性を確認した。本セクションでは、2 番目の研究課題、即ち、形態的に豊かな言語は構文的により自由になる可能性が高いかどうかを探る。スピアマンの順位相関係数分析は、形態と語順の指標の間で実行し、正のかなり強い相関関係が確認されていた。MAMR 対 COS:  $\rho = 0.735$  および  $p = 0.001$ 。MAMR 対 ENTR:  $\rho = 0.656$  および  $p = 0.006$ 。MAMSP 対 COS:  $\rho = 0.624$  および  $p = 0.010$ 。MAMSP 対 ENTR:  $\rho = 0.565$  および  $p = 0.023$ 。4 つの指標の散布図行列を図 3 にプロットした。

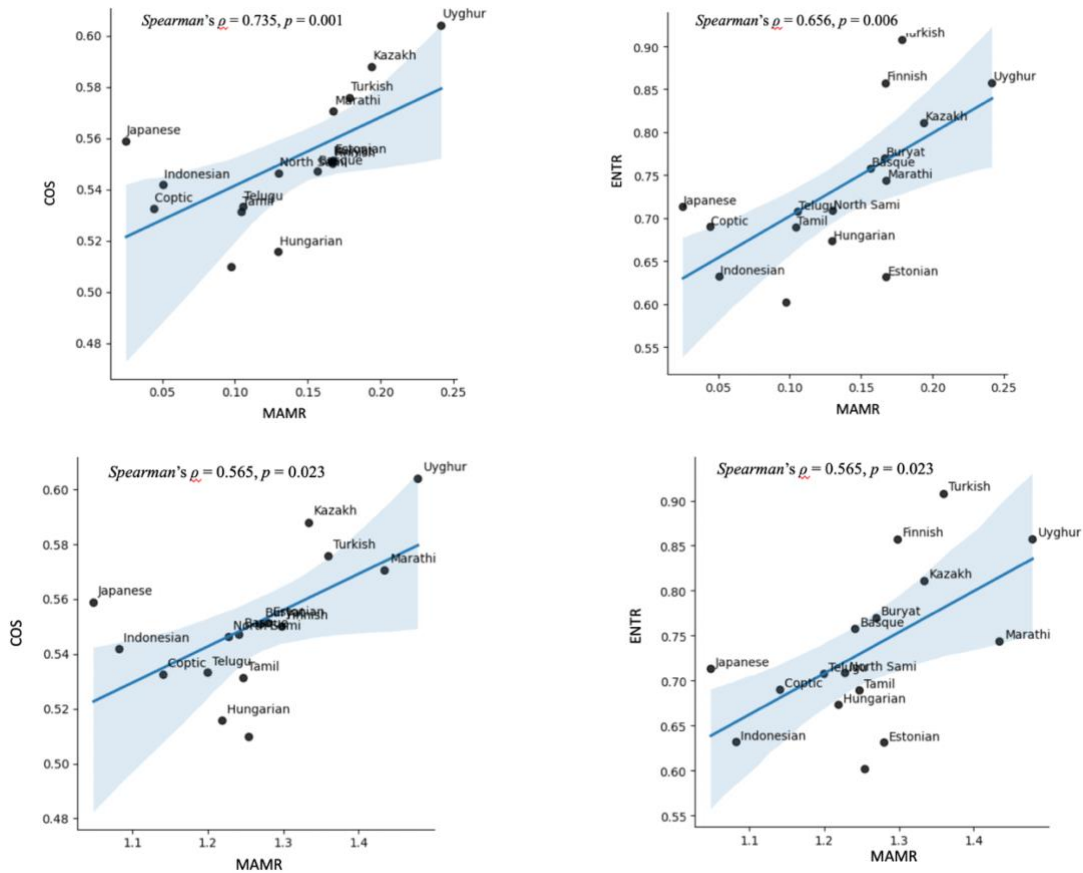


図 3. MAMR、MAMSP、COS、ENTR 間のスピアマンの順位相関係数

図 3 は、4 つの値間の正の相関関係を示している。言語の形態的豊かさが増加すると、語順の柔軟性が高まる。この発見は、スラブ言語 (Yan and Liu 2021) や日本語書き言葉 (李他 2022) と同様に、「複雑さのトレードオフ」仮説と一致している。図 4 と 5 は、形態的および統語的な多様性に基づいた 9 つの言語族の膠着言語のクラスタリングを示している。ウラル語族のフィン語派とフィン・ウゴル語派の分離が窺えた。モンゴル語派とチュルク語派の差が見られた：チュルク語派は、モンゴル語派よりも形態的に豊富で、語順に柔軟であった。日本語は他の 15 の膠着言語に比べて形態学的に豊かではないが、語順の自由度は中程度であると見出された。形態と語順に基づく言語のクラスタリングには、リトアニア語、クルマンジ語、アルメニア語、北サーミ語、フィンランド語、チェコ語、ギリシャ語とバスク語は形態的豊富さと語順の自由度に一致している。

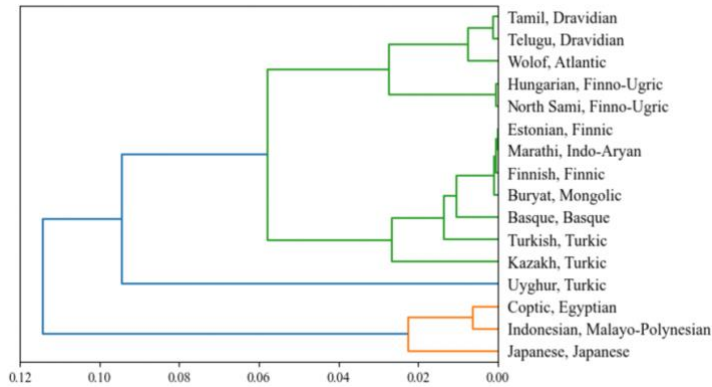


図 4. 形態的豊かさに基づく膠着言語のクラスタリング

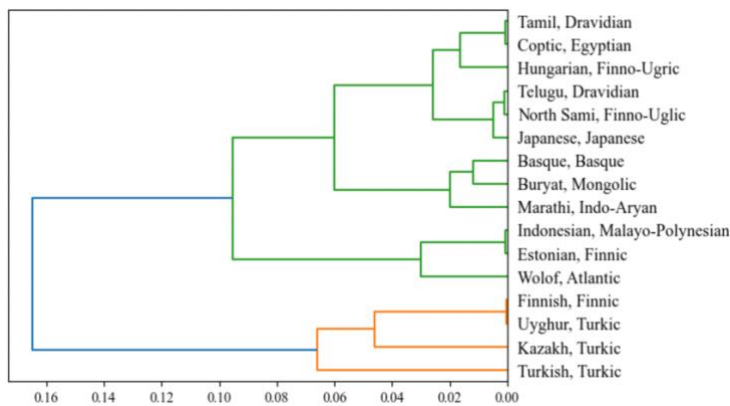


図 5. 語順の柔軟性に基づく膠着言語のクラスタリング

以上を踏まえ、形態—統語の一端から言語間距離が如何に日本語習得に左右するかを考察した。2022年10月時点で、在日外国人労働者数が182万人に達した(厚生労働省)。言語伝達は、L1、L2、Ln、学習環境、モチベーションなどの影響を伴うが(奥野 2019; 宇佐美 2006; 李・石川・砂川 2018; 迫田・細井 2020; 阿辺川他 2020)、本稿は学習初期段階に絞る。12の母国語(フランス語、ドイツ語、ドイツ語、ドイツ語、ハンガリー語、中国語、インドネシア語、韓国語、ロシア語、スペイン語、タイ語、トルコ語、ベトナム語、英語(イギリス英語、ニュージーランド英語、アメリカ英語))にわたる日本語学習者の作文を3000分横断的に分析した。その結果、孤立型・屈折型言語話者は産出した語順は膠着型言語話者よりはるかに低い。

#### 4. まとめ

本研究では、9つの言語族にわたる16の膠着言語の25のツリーバンクにより、形態的豊かさと語順の柔軟性との相関関係を調査した。指標のMAMR、MAMSP、COSとENTRは、形態的および構文的な多様性を捉えるのに有効であることが検証された。次の三点にたどり着いた。第一に、形態素が豊富であればあるほど、語順がより柔軟になる。形態と構文の間の相関関係がすべての膠着言語で確認されており、複雑さのトレードオフ仮説を裏付けている。9つの言語族の中で、オーストロネシア語、アフリカアジア語は形態的にも語順的にも固定していることが観察された。形態と統語的に多様な言語は、アルタイ語族のチュル

ク語派、ヨーロッパ語族のインド・アーリア語派、ウラル語族のフィン語派と窺えた。トルコ語、ウイグル語、バスク語および北サーミ語は、S、V、O の組み合わせの割合が最もバランスが取れていた。第二に、言語族内に差が見られた：モンゴル語派はチュルク語派より、フィン語派はフィン・ウゴル語派より語順が柔軟である。第三に、L1 の形態学的豊かさと習得言語の語順の再帰性との間に正の関係が確認された。

### 謝 辞

本研究は中国国家社会科学基金に基づいて行われたものである ((22BYY186: 李文超)。

### 文 献

- 阿辺川武, 仁科喜久子, 八木豊, ホドシチェック・ボル (2020). 日本語接続表現の計量的分析に基づく指導法の提案『計量国語学』32(7) 387-402.
- Bentz C., Alikaniotis D., Cysouw M., Ferrer-i-Cancho R (2017). The entropy of words. Learnability and expressivity across more than 1000 languages. *Entropy* 19 (6): 1–32.
- Čech, R., Kubát, M. (2018). Morphological Richness of Text. In: Fidler, M., Cvrček, V. (eds.) *Taming the Corpus. From Inflection and Lexis to Interpretation. Quantitative Methods in the Humanities and Social Sciences*. Cham: Springer, 63-77.
- Chen R., Liu H., Altmann G. (2016). Entropy in different text types. *Digital scholarship in the humanities* 32(3): 528–542.
- Coloma, G. (2015). The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics* 28: 139-159.
- Coloma, G. (2017). The Existence of Negative Correlation between Linguistic Measures across Languages. *Corpus Linguistics and Linguistic Theory* 13: 1-26.
- Coloma, G. (2020). Language complexity trade-offs revisited, *Serie Documentos de Trabajo*, No. 721, Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires.
- Fenk-Oczlon, G., Fenk, A (1999). Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology*, pp. 151-177.
- Fenk-Oczlon, G., Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznan Studies in Contemporary Linguistics*. Volume 50, Issue 2, Pages 145–155.
- Fenk-Oczlon G., Pilz J (2021). Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size. *Frontiers in Communication*. 6:626032. doi: 10.3389/fcomm.2021.626032.
- Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford UP.
- Jakobson, R. (1936). Beitrag zur allgemeinen Kasuslehre, Gesamtbedeutungen der russischen Kasus. In: *PLingCP* 6: 240-288.
- Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
- Köhler, R. (2005). Synergetic Linguistics. In: Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.). *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: Walter de Gruyter, 760-775.



- Koplenig A, Meyer P, Wolfer S, Müller-Spitzer C. (2017). The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLoS ONE* 12(3).
- Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533– 572.
- 李在鎬・石川慎一郎・砂川有里子(2018)『新・日本語教育のためのコーパス調査入門』くろしお出版.
- 李文平, 刘海涛, 熊子涵 (2022). 「日本語における語順の自由度と格標識の豊富さに関する計量的研究」. 『計量国語学』 33 (5), pp. 325-340.
- McFadden, T. (2003). On morphological case and word-order freedom. *Berkeley Linguistics Society* (BLS) 29. 295–306.
- Menzerath, P. (1954). *Die Architektur des Deutschen Wortschatzes*. Hannover; Stuttgart: Dümmler.
- Michael A., McFall, D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR), *Journal of Quantitative Linguistics*, 17:2, pp. 94-100.
- Miestamo, M. (2009). *Implicational hierarchies and grammatical complexity*. Oxford University Press.
- 奥野由紀子・呉佳穎・村田裕美子 (2019). 「日本語学習者の能動態と受動態の使用傾向にみられる母語による違いー中国語とドイツ語での語りの比較からー」『日本語研究』 39, 79-93.
- 迫田久美子・細井陽子 (2020) 異なった学習環境における日本語使用の正確さと複雑さ-日本語学習者コーパス (I-JAS) の分析に基づいて- 『計量国語学』 32 (7), 403-418.
- Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace & World Inc., 33-35.
- Shannon C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 1948, 27(4): 623–656.
- Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, 10(1), 1-40.
- Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. In Miestamo, M., K. Sinnemäki, K. and F. Karlsson, F. (eds.), *Language Complexity: Typology, Contact, Change* (Amsterdam: John Benjamins), 67–88. doi: 10.1075/slcs.94.06sin.
- Sinnemäki, K. (2014). Complexity trade-offs: A case study. In: F.J. Newmeyer and L.B. Preston (eds.), *Measuring Grammatical Complexity*, 179–201. Oxford: Oxford University Press.
- 角田太作 (1991). 世界の言語と日本語 改訂版—言語類型論から見た日本語. 東京 : くろしお出版.
- 宇佐美洋 (2006). フランス語母語話者の日本語作文における「意図不明表現」の分析—母語訳との対照から見る「分かりにくさ」の理由—『作文対訳データベースの多様な利用のために—「日本語教育のための言語資源及び学習内容に関する調査研究報告書」 81-99.
- Xanthos, A., Gillis, S. (2010). Quantifying the development of inflectional diversity. *First Language* 30, 175–198.
- Yan, J, Liu, H. (2021). Morphology and word order in Slavic languages: Insights from annotated corpora. *Voprosy Jazykoznanija* 4: 131–159.
- Zipf G. K. (1949). *Human behavior and the principle of least effort: An introduction to human*

*ecology* (Cambridge: Addison-Wesley).

# Subjective frequency norms for 100 Japanese verb-verb compounds: The first step toward constructing a behavioral database for Japanese compound words

Yijia Yao (Graduate school of Languages and Linguistics Sophia University)

日本語複合動詞 100 語の主観的出現頻度評定値 :  
日本語複合語に関する語彙データベースの構築に向けて  
姚 一佳 (上智大学)

## Abstract

Lexical frequency has been commonly used in previous studies to investigate how compound words are processed in the mental lexicon. Since the work of Balota et al., (2007), a growing number of databases have been constructed using normative and behavioral data of compound words (for English, see Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007; for French, see Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, Augustinova & Pallier, 2010; for Chinese, see Sun, Hendrix, Ma, & Baayen, 2018; for Dutch, see Keuleers, Diependaele, & Brysbaert, 2010), which make it possible to investigate and compare the mechanism of compound processing across languages. However, despite the fact that compound words are prevalently used in Japanese, to the best of our knowledge, such database have not previously been constructed for Japanese compound words. This study reported two subjective ratings and six corpus-based objective frequency estimates for 100 Japanese verb-verb compounds: familiarity, age-of-acquisition (AoA), whole-word surface frequency, whole-word lemma frequency, first-constituent surface frequency, first-constituent lemma frequency, second-constituent surface frequency, second-constituent lemma frequency. The relationship among these frequency measures was examined by using correlational and hierarchical regression analyses. Negative correlation was obtained between the familiarity and AoA ratings, indicating that more familiar compounds tended to be rated as being of earlier-acquired words by native speakers of Japanese. Moreover, the whole-word lemma frequency was the best single predictor of both familiarity and AoA ratings, revealing that when native speakers of Japanese are asked to rate a Japanese verb-verb compound on a particular variable, the verb-verb compound is processed as a whole unit. These normative data should be beneficial to researchers who are interested in selecting stimuli for psycholinguistic experiments, and it will also help us to gradually construct a normative and behavioral database for Japanese compound words.

## 1. Introduction

Lexical processing is tuned to lexical frequency. Generally, high-frequency words are processed more quickly and accurately than low-frequency words. With respect to compound words, which are composed of two or more monomorphemic words (e.g., *watermelon*), the past four decades have seen a simmering debate as to whether compound words are processed via a holistic route or via an analytic route in the mental lexicon (for review, see Libben, Gagné, & Dressler, 2020), and this debate has centered largely around the question of which frequency variable is the best predictor for compound processing. Specifically, it is assumed that if the processing efficiency of a compound word varies according to the frequency of the whole-word form but not according to the frequencies of its corresponding constituents, then the compound word might be processed as a whole unit in the mental

lexical (Full-listing models, Butterworth, 1983; Bybee, 1995). On the other hand, if the processing efficiency of a compound word is modulated by the constituent frequencies but not by the whole-word frequency, then the compound word might be processed as separate components in the mental lexicon (Full-parsing models, Libben, Derwing, & de Almeida, 1999; Taft & Forster, 1976). In addition, jointly influences of the whole-word and constituent frequencies are interpreted as results of dual-route processing (for Sub-lexical model, see Rastle, Davis, & New, 2004; for Supra-lexical model, see Giraudo & Grainger, 2000; for Parallel dual-route model, see Schreuder & Baayen, 1995).

Because of its critical importance as a diagnostic factor in investigating the mechanism of compound processing, the problem of how to select the most appropriate indicator of lexical frequency for compound words have increasingly attracted the attention of psycholinguistic researchers. Previous studies have showed that different types of frequency measures may tap into different underlying constructs (for French, see Bonin, Laroche, & Méot, 2021; for English, see Juhasz, Lai, & Woodcock, 2015; for Chinese, see Su, Li, & Li, 2023). This means that using different frequency measures may yield somewhat different experimental results, and without solid validation of the measurements that are used to gather the frequency information, comparisons across studies and languages might be misleading. Therefore, to attempt to better understand the essential frequency effects on compound processing, this paper reported normative data for 100 Japanese verb-verb compounds and statistically examined the relationships among the different frequency measures. To the best of our knowledge, this work is the first normative study on Japanese verb-verb compounds.

## 2. Previous studies

### 2.1 Corpus-based measures of objective frequency

In psycholinguistic studies, the most widely used measure to estimate lexical frequency is the corpus-based frequency measure. Generally, two types of frequency estimates can be derived from a corpus: Surface frequency and lemma frequency. Surface frequency refers to the frequency of the particular word form. In the case of compound words, such as *watermelon*, three types of surface frequency can be distinguished: Whole-word surface frequency (the surface frequency of *watermelon*), first-constituent surface frequency (the surface frequency of *water*), and second-constituent surface frequency (the surface frequency of *melon*). On the other hand, lemma frequency refers to the summed frequency of a word form and its inflectional variant. For example, for a compound verb, such as *double-check*, three types of lemma frequency can be derived: Whole-word lemma frequency (the summed frequency of *double-check*, *double-checked*, *double-checking*, *double-checks*), first-constituent lemma frequency (the summed frequency of *double*, *doubled*, *doubling*, *doubles*), and second-constituent lemma frequency (the summed frequency of *check*, *checked*, *checking*, *checks*).

These corpus-based estimates serve as important indicators of compound processing in psycholinguistic studies. For example, most studies showed that the visual recognition of compound words involves access to both whole-word and constituent representations, as evidenced by an interactive effect of whole-word and constituent frequencies on visual lexical decision times (for English, see Schmidtke, Gagné, Kuperman, Spalding, & Tucker, 2018; for German, see Günther, Marelli, & Bolte, 2020; for Spanish, see Güemes, Gattei, & Wainseboim, 2019). However, this is not always the case when data from Asian languages is used. It was shown that direct access to the whole-word representation of compound words is more preferable than morpheme-based access in Asian languages, such as Chinese (Myers, Huang, & Wang, 2006; Tse, Yap, Chan, Sze, Shaoul, & Lin, 2017), Japanese (姚 2020), and Vietnamese (Pham & Baayen, 2015), as evidenced by the predominant effect of whole-word frequency. In other words, there seems to exist a typological distinction in compound

processing. However, before we delve into this linguistic typological puzzle, the first thing we want to know is whether surface and lemma frequency measures orient to the same aspects of mental representations of compound words. This is important because differences in measurements might make previous results incomparable with each other.

Although consensus has not been reached yet, previous studies dealing with this ongoing problem reported two important findings: Firstly, the predictive power of surface and lemma frequency measures on lexical decision times is different. Surface frequency is demonstrated to be less powerful to predict the whole-word frequency on lexical decision times, as compared to lemma frequency (Ji, Gagné, & Spalding, 2011; Juhasz, Lai, & Woodcock, 2015; Keuleers, Brysbaert, & New, 2010). For example, by collecting both constituent-based surface and lemma frequency information on the same stimuli, Ji, Gagné, & Spalding (2011) found that whereas lemma frequency is the most important variable to predict lexical decision times, no influence of surface frequency on lexical decision times was observed. In addition, Keuleers, Brysbaert, & New (2010) reported that adding compound-based lemma frequency to the stepwise multiple regression model increases the explained variance in lexical decision times by 10% points for compound words. Secondly, there is a nonlinear relationship between objective frequency and lexical decision times. The predictive power of objective frequency for words in the high (i.e., above 100 per million) and low (i.e., below 1 per million) frequency bands were weaker than those in the medium frequency band (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Kuperman & Van Dyke, 2013). Moreover, objective frequency tends to be less reliable in predicting proficient readers' (i.e., readers have more reading experience or larger vocabulary size) task performance (Falkauskas & Kuperman, 2015; Kuperman & Van Dyke, 2013). For example, Kuperman & Van Dyke (2013) reported that proficient readers are equally faster in processing both high- and low-frequency words. These results indicate that although both surface and lemma frequencies are well-accepted diagnostic factors in previous studies, they may represent some related but distinct constructs, and they might not be reliable indicators of individual differences in language experience.

## 2.2 Rating-based measures of subjective frequency

In order to better understanding the effects of experience-driven individual differences on compound processing, two rating-based frequency measures are increasingly used: Familiarity and AoA. Familiarity refers to the self-reported frequency of occurrence of a word in the raters' daily lives. This estimate is typically obtained by asking participants to evaluate how often they encounter or produce a given word in daily communication, by using a Likert scale. For example, Tagalakis & Keane (2006) asked their participants to judge whether they had encountered the compound word before and how frequently using a 7-point Likert scale (1 = very unfamiliar, 7 = very familiar). Momenian, Cham, Amini, Radman, & Weekes (2021) asked their participants to rate the compound word and its corresponding constituents by using a 7-point Likert scale familiarity questionnaire (1 = very unfamiliar, 7 = very familiar). AoA refers to the age at which people learn a particular word and are often obtained by asking subjects to indicate their own age at which they learned the words. For instance, Juhasz, Lai, & Woodcock (2015) used a 7-point scale with 2-year age bands to collect the age at which compound words were acquired (7 = the word was learned at age 13 or order). Although it is known that rated AoA is less reliable for the words which are learned before the age of 4 and after the age 15 (Brysbaert, 2017), compound words that were judged to be earlier-acquired were shown to be recognized and produced faster and more accurately than those judged to be later-acquired (Juhasz, 2018; Song & Li, 2021).

By collecting these two rating-based frequency estimates on the same compound word dataset, previous studies reported several interesting findings regarding their relationship: Firstly, familiarity and AoA are associated with each other, as evidenced by a negative correlation between familiarity

and AoA rating scores (Bonin, Laroche, & Méot, 2021; Juhasz, Lai, & Woodcock, 2015; Su, Li, & Li, 2023). Familiar compound words are more likely to be rated as earlier-acquired compounds and vice versa. Secondly, both of them are reliable predictors of lexical decision times (Bonin, Laroche, & Méot, 2021; Juhasz, Lai, & Woodcock, 2015). Thirdly, adding familiarity and AoA into the statistical model produced a significant increase in the amount of variance explained compared to the baseline model which only includes corpus-based frequency estimates (Juhasz, Lai, & Woodcock, 2015). This means that rating-based subjective frequency measures were able to capture some important and unique features of compound words, which were not reflected by corpus-based objective frequency measures.

However, despite the increasingly widespread use of these two rating-based frequency measures in psycholinguistic studies, there are two major mysteries still remain unsolved. Firstly, whether familiarity and AoA reflect the same underlying constructs? With respect to this question, inspection of the literature reveals that, compared to AoA, familiarity is more strongly related to orthographic properties of compound words, such as number of strokes (Su, Li, & Li, 2023). On the other hand, AoA is more strongly associated with semantic properties of compound words, such as imageability and sensory experience (Bonin, Laroche, & Méot, 2021; Juhasz, Lai, & Woodcock, 2015). Secondly, how they are related to corpus-based measures? Although scarce, previous studies showed that familiarity were strongly associated with whole-word frequency rather than constituent frequencies, whereas AoA were weakly associated with both whole-word and constituent frequencies (Bonin, Laroche, & Méot, 2021; Juhasz, Lai, & Woodcock, 2015; Su, Li, & Li, 2023). These results indicate that although both familiarity and AoA have proven to be effective measures of personal linguistic experience, they may tap in some different aspects of word features.

### 2.3 Why Japanese verb-verb compounds?

Japanese is rich in verb-verb compounds. According to Compound Verb Lexicon (National Institute for Japanese Language and Linguistics, 2015), which is an online database for Japanese verb-verb compounds, at least 2700 verb-verb compounds are commonly used in contemporary Japanese. Besides this quantitative feature, Japanese verb-verb compounds also possess some peculiar structural features which might play special roles in processing.

Firstly, Japanese verb-verb compounds are all concatenated words. Specifically, no orthographical segmentation cues, such as spaces or hyphens, can be inserted between the two constituents of Japanese verb-verb compounds. This is distinctly different from English compound words, which can be presented either with or without a space or hyphen. In addition, unlike Greek and Polish, in which compound words are generally formed by inserting a linking vowel between their corresponding constituents, Japanese verb-verb compounds are usually formed by melting two monomorphemic words together (i.e., by eliminating the verbal ending of their corresponding first constituents). For example, to form a Japanese verb-verb compound 食べ始める “start eating”, we need to transform its first constituent 食べる “eat” into the continuous form 食べ by eliminating the verbal ending る. In other words, compound verbs are orthographically far more like monomorphemic words in Japanese than in other languages.

Secondly, although Japanese verb-verb compounds are similar in orthographic structure, there are two distinct association patterns between the first and second constituents. One pattern, the so-called syntactic verb-verb compound, is highly productive because principally it allows flexible combination of all monomorphemic verbs, the other pattern, the so-called lexical verb-verb compound, is highly restricted because it is based on the transitivity of the constituents (影山 1993). In case of lexical verb-verb compounds, for example, a transitive verb 拭く “wipe” can be combined with another transitive verb 消す “to make something disappear”, but can never be combined with an intransitive verb, such as 消える “disappear”. In contrast, with respect to syntactic verb-verb compounds, a transitive verb 拭く “wipe” can be combined with either a transitive verb 始める “start” or an intransitive verb 始

まる “started”. In other words, regular word-formation rules and exceptions co-exist in Japanese verb-verb compounds.

Thirdly, in Japanese verb-verb compounds, the orthographic-semantic associations are not always systematic. Most of the Japanese verb-verb compounds are polysemous words. For example, a verb-verb compound 仰ぎ見る may refer either to “look up” or “look up to someone as”. In addition, many Japanese verb-verb compounds are composed of homonymic constituent verbs. For example, a monomorphemic verb つく may refer either to “attached” (i.e., in 追いつく “catch up”) or “to imagine something” (i.e., in 思いつく “come up with”). In other words, the transparency of the semantic relationship between the Japanese verb-verb compound and its corresponding constituents is relatively low.

To sum up, the aforementioned quantitative and structural features make Japanese verb-verb compounds to be an ideal testing ground for exploring the relationship between different frequency measures, since the morphological productivity of the compound words (Frost, Kugler, Deutsch, & Forster, 2005; Velan & Frost, 2011), the orthographic segmentation cues (Rastle, Davis, & New, 2004), and the homonymy of the constituents (Bertram, Laine, Baayen, Schreuder, & Hyönä, 2000; Bertram, Schreuder & Baayen, 2000) are assumed to be influential factors of compound processing. However, to the best of our knowledge, this relationship has not yet been investigated in Japanese verb-verb compounds and no normative data are available for Japanese verb-verb compounds. Thus, the current study aims to fill these gaps by providing normative data for 100 Japanese verb-verb compounds and investigates how different frequency measures are related to each other.

### 3. Method

#### 3.1 Participants

A total of forty-five native speakers of Japanese participated in our study (14 males and 31 females; mean age, 24.27 years; range, 18-54 years). All participants had normal or corrected-to-normal vision. They received ¥4,500 as compensation for their participation. Participants performed familiarity and AoA ratings individually. This study involving human participants was reviewed and approved by the Sophia University Ethics Committee for Research on Human Subjects.

#### 3.2 Materials

A total of 100 Japanese verb-verb compounds were selected for familiarity and AoA rating tasks from the database BCCWJ (Balanced Corpus of Contemporary Written Japanese; National Institute for Japanese Language and Linguistics, 2012). The selection was performed by the author.

To be included in the stimulus list, a word had to have a “monomorphemic verb + monomorphemic verb” structure. The verb-verb compounds ranged in word length from four to six characters (mean word length = 4.46,  $SD = 0.54$ ) and syllable length from four to seven (mean syllable length = 5.26,  $SD = 0.8$ ). The word lengths of first constituents ranged from two to four characters (mean word length = 2.34,  $SD = 0.5$ ), and the word lengths of second constituents ranged from two to five characters (mean word length = 2.78,  $SD = 0.79$ ). The syllable lengths of first constituents ranged from two to five (mean syllable length = 2.78,  $SD = 0.79$ ), and the syllable lengths of second constituents ranged from two to four (mean syllable length = 2.8,  $SD = 0.64$ ).

Two types of corpus-based objective frequency estimates were derived from the BCCWJ corpus by using *Chunagon* interface (about 104 million words). Surface frequency was operationalized as the frequency of the particular word-form as it was presented in the experiment. Surface frequency ranged from 2 to 1 per million for verb-verb compounds (mean surface frequency = 0.37,  $SD = 0.43$ ), ranged from 9 to 1 per million for first constituent (mean surface frequency = 1.89,  $SD = 1.71$ ), and ranged from 501 to 1 per million for second constituents (mean surface frequency = 37.49,  $SD = 62.93$ ).

Lemma frequency was operationalized as the summed frequency of a word and its inflectional

variants. For instance, lemma frequency of a verb-verb compound 食べ始める “start eating” was the summed frequency of its inflectional (e.g., 食べ始める “start eating”, 食べ始めた “started eating”, 食べ始めよう “let’s start eating”, 食べ始めたら “once (I) started eating”) and spelling variants (e.g., 食べはじめる, たべはじめる, たべ始める, 食べはじめた). Lemma frequency ranged from 17 to 1 per million for verb-verb compounds (mean lemma frequency = 2.07,  $SD = 2.58$ ), ranged from 98 to 1 per million for first constituents (mean lemma frequency = 10.93,  $SD = 13.48$ ), and ranged from 820 to 1 per million for second constituents (mean lemma frequency = 166,  $SD = 229.28$ ).

Two types of rating-based subjective frequency estimates of verb-verb compounds were collected by using questionnaires. Familiarity was operationalized in terms of ratings of the familiarity of the compound word in daily communication. Familiarity of verb-verb compounds was collected by using a 7-point Likert scale. Response scales ranged from *very unfamiliar with the verb-verb compound* (1) to *very familiar with the verb-verb compound* (7). AoA was operationalized as the age at which the compound verb was acquired. AoA of verb-verb compounds was collected by using a 12-point scale with 0.5-year age bands. Response scales ranged from *word acquired at 1-year-old* (1) to *word acquired after 7-year-old* (7). For each rating task, a form was created in which the order of word presentation was randomized across participants.

### 3.3 Procedure

Participants first gave their consent. They were instructed to rate their level of familiarity or age of acquisition to each verb-verb compound intuitively. The familiarity rating task was provided first for each item. At the top of the form, participants were provided with an instruction consisting of the main question and a description of the Likert scale categories. In the header of each page, a brief statement reiterating the scale was also provided. Target words were presented on the left-hand side of the page and participants made their ratings by marking the point with a circle. Secondly, the AoA rating task was provided and the procedure was identical to the familiarity rating task.

### 3.4 Statistical analysis

All statistical analyses were conducted using SPSS version 24.0 (SPSS Inc., Chiago, IL, USA). The Pearson correlation analysis was used for assessing the correlations between different frequency estimates. The item-level hierarchical multiple regression analyses and multiple regression analyses were used for evaluating the predictive power of objective frequency estimates on subjective rating scores. For all analyses, a value of  $p < .05$  was considered significant. For each corpus-based objective frequency estimate, we added 1 to the raw frequency value, then calculated the frequency per million, and finally calculated the  $\log_{10}$  of the frequency of the per million value.

## 4. Results

The mean ratings and standard derivations, and ranges for each of the rating-based subjective frequency estimates are presented in Table 1.

Table 1 Descriptive statistics for the two rating-based subjective frequency estimates

Estimate	Minimum	Maximum	Mean	SD
Familiarity	6.87	2.22	5.58	1.16
AoA	6.96	4.48	6.24	0.62

SD = standard deviation. AoA = age of acquisition. All ratings are on a 1 to 7 scale.

### 4.1 Correlation analyses



Table 2 presents the correlations between the two rating-based subjective and the six corpus-based objective frequency estimates. Familiarity was positively correlated with whole-word surface frequency [ $r(100) = .491, p = .000$ ] and whole-word lemma frequency [ $r(100) = .664, p = .000$ ], and weakly correlated with first-constituent surface frequency [ $r(100) = .252, p = .011$ ] as well as first-constituent lemma frequency [ $r(100) = .267, p = .007$ ]. However, it was not found to be significantly correlated either with second-constituent surface frequency [ $r(100) = .057, p = .576$ ] or with second-constituent-lemma frequency [ $r(100) = .043, p = .669$ ]. These results suggest that verb-verb compounds that with higher whole-word frequency or with more frequent first-constituents tended to be rated as being more familiar by native speakers of Japanese.

On the other hand, AoA was marginally negatively correlated with whole-word surface frequency [ $r(100) = -.258, p = .01$ ] and whole-word lemma frequency [ $r(100) = -.412, p = .000$ ]. A significant but weak negative correlation was observed between AoA and first-constituent surface frequency [ $r(100) = -.212, p = .034$ ], as well as between AoA and first-constituent lemma frequency [ $r(100) = -.269, p = .007$ ]. However, no significant correlation was found either between AoA with second-constituent surface frequency [ $r(100) = -.043, p = .674$ ], or between AoA and second-constituent lemma frequency [ $r(100) = -.043, p = .673$ ]. These results suggest that verb-verb compounds that with higher whole-word frequency or with more frequent first-constituents tended to be rated as being of earlier-acquired words by native speakers of Japanese.

Finally, AoA were strongly negatively correlated with familiarity [ $r(100) = -.599, p = .000$ ], indicating that verb-verb compounds learned early in life tended to be rated as more familiar by native speakers of Japanese.

Table 2 Correlation between rating-based and corpus-based ( $\log_{10}$ ) frequency estimates

	Fam	AoA	W- surface	W- lemma	1st surface	1st lemma	2nd surface	2nd lemma
Fam	1	-.559***	.491***	.664***	.252*	.267**	.057	.043
AoA		1	-.258**	-.412***	-.212*	-.269**	-.043	-.043
W-surface			1	.795***	.280**	.178	.390***	.357***
W-lemma				1	.352***	.325***	.172	.155
1st-surface					1	.794***	.134	.159
1st-lemma						1	.070	.094
2nd-surface							1	.975***
2nd-lemma								1

Fam = whole-word familiarity. AoA = whole-word age of acquisition. W-surface = whole-word surface frequency. W-lemma = whole-word lemma frequency. 1st-surface = first-constituent surface frequency. 1st-lemma = first-constituent lemma frequency. 2nd-surface = second-constituent surface frequency. 2nd-lemma = second-constituent lemma frequency. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

#### 4.2 Regression analyses

To further examine whether the characteristics of the verb-verb compounds and their corresponding constituents influence the familiarity and AoA rating scores, firstly we conduct three-step hierarchical item-level multiple regression analyses for both familiarity and AoA ratings, with rating scores as dependent measures. Length variables (Whole-word length, First-constituent word length, Second-constituent word length, Whole-word syllable length, First-constituent syllable length, Second-constituent syllable length) were entered in Step 1, Constituent-based frequency variables (First-constituent surface frequency, First-constituent lemma frequency, Second-constituent surface frequency, Second-constituent lemma frequency) were entered in Step 2, and Compound-based

frequency variables (Whole-word surface frequency, Whole-word lemma frequency) were entered in Step 3. The results of these hierarchical regression analyses are shown in Table 3.

Table 3 Standardized regression coefficients ( $\beta$ ) and significance from Step 1 to 3 of the regression analyses for familiarity and AoA ratings (Hierarchical regression)

Predictors	Familiarity ratings	AoA ratings
Step 1: Length variables		
R-square	.123	-.009
Step 2: Constituent-based frequency variables		
R-square	.203	.045
	$\Delta R^2 = .081$	$\Delta R^2 = .089$
Step 3: Compound-based frequency variables		
R-square	.427***	.156**
	$\Delta R^2 = .293$	$\Delta R^2 = .117$

Length variables = Whole-word length, First-constituent word length, Second-constituent word length, Whole-word syllable length, First-constituent syllable length, Second-constituent syllable length. Constituent-based frequency variables = First-constituent surface frequency, First-constituent lemma frequency, Second-constituent surface frequency, Second-constituent lemma frequency. Compound-based frequency variables = Whole-word surface frequency, Whole-word lemma frequency. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

As shown in Table 3, Length variables did not account for significant variance across the rating scores in both familiarity and AoA ratings. Likewise, Frequency variables at the constituent level also did not account for significant variance in both familiarity and AoA rating scores. However, Frequency variables at the whole-word level accounted for significant additional unique variance in the two rating-based estimates (29.3% and 11.7% for rating scores in familiarity and AoA rating respectively). These results indicate that compound-based frequency variables play determinant and exclusive role in subjective ratings.

For further investigate the influence of corpus-based objective frequencies on subjective ratings, we conduct multiple regression analyses for both familiarity ratings and AoA ratings, with rating scores as dependent measures, and including the following predictors: Whole-word surface frequency, Whole-word lemma frequency, First-constituent surface frequency, First-constituent lemma frequency, Second-constituent surface frequency, Second-constituent lemma frequency. The results of these multiple regression analyses are shown in Table 4.

Table 4 Standardized regression coefficients ( $\beta$ ) and significance from models predicting familiarity and AoA ratings (Multiple regression)

Predictors	Familiarity ratings	AoA ratings
<i>Adjusted R<sup>2</sup></i>	.414	.150
Whole-word surface frequency	-.057	.125
Whole-word lemma frequency	.701***	-.498**
First-constituent surface frequency	-.037	.076
First-constituent lemma frequency	.085	-.194
Second-constituent surface frequency	.073	-.003
Second-constituent lemma frequency	-.118	-.012

AoA = age of acquisition. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

As shown in Table 4, with approximately 70% of explained variance, familiarity rating scores were predicted to a large extent by the whole-word lemma frequency, with verb-verb compounds rated as more familiar as the whole-word lemma frequency becomes higher. However, familiarity rating scores were not significantly predicted by either whole-word surface frequency or constituent-based surface/lemma frequencies. Likewise, whole-word lemma frequency was also a significant predictor of AoA rating scores. The higher whole-word lemma frequency the verb-verb compounds with, the more they were rated as early acquired. Similar to what occurred for familiarity, AoA rating scores were not significantly predicted by either whole-word surface frequency or constituent-based surface/lemma frequencies. In other words, the whole-word lemma frequency is the single best determinant for both familiarity and AoA rating scores.

## 5. Discussion

As familiarity and AoA become widely used factors in psycholinguistic studies, there has been an intense surge of interest in the investigation of their construct validity (Bonin, Laroche, & Méot, 2022; Juhasz, Lai, & Woodcock, 2015; Song & Li, 2021). However, due to data limitations, the relationship between familiarity and AoA was not yet validated in Japanese verb-verb compounds. In this article, we provided the first normative dataset of Japanese verb-verb compounds. This dataset contains two types of subjective frequency measures (familiarity, AoA) for 100 Japanese verb-verb compounds as well as two types of objective frequency measures (surface frequency, lemma frequency) for 100 Japanese verb-verb compounds and their corresponding constituents. Familiarity ratings were made on a scale ranging from 1 = *very familiar* to 7 = *very unfamiliar* (7-point Likert scale) and AoA ratings were made on a scale ranging from 1 = *1-year-old* to 7 = *after 6-year-old* (12-point scale with 0.5-year age bands) by 45 native speakers of Japanese. Surface and lemma frequencies were calculated on the basis of a corpus of approximately 104 million words gathered from modern Japanese texts. To find out whether these two frequency measures related to one another we further performed correlation and regression analyses on the aforementioned dataset. The major findings were as follows.

With respect to correlation analyses, first of all, we found that the two rating-based subjective frequency measures were strongly correlated with each other. Compound words that were judged to be familiar in daily communication were estimated to be acquired earlier in life. Secondly, we further found that the two rating-based subjective frequency measures had similar but not identical patterns of intercorrelations with corpus-based objective frequency measures. For familiarity, rating scores were strongly associated with both whole-word surface and whole-word lemma frequencies, and weakly associated with first-constituent surface and first-constituent lemma frequencies. Likewise, for AoA, rating scores were weakly related to whole-word surface frequency and first-constituent frequencies, and more closely associated with whole-word lemma frequency. However, neither of them was associated with second-constituent frequencies. On the other hand, the major findings of regression analyses shed light on whether and to what extent objective frequency variables influence subjective frequency ratings: We found that whole-word lemma frequency was the best single predictor of familiarity and AoA rating scores. This result indicates that the total probability of encountering a verb-verb compound rather than the probability of encountering the particular word-form determines native speakers' impression of the compound word.

The predominant effect of whole-word lemma frequency on subjective frequency ratings in the present study challenges the adequacy of existing Full-parsing models (e.g., Libben, Derwing, & de Almeida, 1999; Taft & Forster, 1976) of compound processing, which assume obligatory morphological decomposition during compound processing. On the basis of the coefficient patterns induced in familiarity and AoA ratings, we would claim that native speakers of Japanese tended to process verb-verb compounds as a whole unit when they were asked to rate the lexical properties of those words explicitly in un-speeded task settings. This is consistent with the predictions of Full-listing

accounts which hypothesize that compound words are activated directly via their whole-word representations during processing (e.g., Butterworth, 1983; Bybee, 1995).

The lack of effects of constituent frequencies on subjective ratings could be attributed to the orthographic and semantic features of Japanese verb-verb compounds. As mentioned in the review section, Japanese verb-verb compounds are words with low orthographic decomposability and low transparency of orthographic-semantic association. Both of these two features were assumed to influence readers' perception of the constituents of the compound words (Gagné, Spalding, Spicer, Wong, Rubio, & Cruz, 2020). For example, spaced compound words are more likely to be processed as separate units than closed compound words (Ji, Gagné & Spalding, 2011), and semantically opaque compounds were more likely to be processed as a whole unit compared to semantically transparent compounds (Marelli & Luzzatti, 2012). In other words, it might be the degree of salience of the constituents affects whether the compound word will be processed as a whole unit or as separate units, and for compound words with less salient constituents, such as Japanese verb-verb compounds, holistic processing route rather than analytic processing route became the preferable route to processing.

Another interesting result of the present study is the absence of effect of whole-word surface frequency on the subjective ratings. This result leads to the conclusion that Japanese verb-verb compounds possess their own lemma in the mental lexicon and was inconsistent with the previous results reported in English and French compound nouns, in which the subjective rating scores of compound words were co-determined by whole-word and constituent surface frequencies (Bonin, Laroche, & Méot, 2022; Juhasz, Lai, & Woodcock, 2015). This discrepancy may be due to the fact that unlike nouns, verbs are frequently used in their inflected forms rather than their original forms in daily communications. Thus, for compound verbs, whole-word surface frequency might not be powerful enough to predict or explain lexical behaviors, and instead, whole-word lemma frequency might be a better indicator of lexical processing.

## 6. Conclusion

Taken as a whole, over the past two decades, there has been an explosion of interest in using databases of behavioral performance measures to investigate questions concerning mechanisms of compound processing (for English, see Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007; for French, see Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, Augustinova & Pallier, 2010; for Chinese, see Sun, Hendrix, Ma, & Baayen, 2018; for Dutch, see Keuleers, Diependaele, & Brysbaert, 2010). Those databases allow a more complete picture of how compound words are processed in the mental lexicon and serve as a useful tool for cross-language comparisons.

However, the majority of available behavioral data are established for Indo-European languages, and there is a lack of normative and behavioral data for Japanese. In this study, we established two psycholinguistic norms for 100 Japanese verb-verb compounds derived from native speakers of Japanese. By examining whether characteristics of the verb-verb compounds and their corresponding constituents influence ratings of familiarity and AoA of verb-verb compounds, we found that whole-word lemma frequency is the best single predictor of the rating scores, and there was only a weak association between first-constituent frequencies and subjective rating scores. These findings support the holistic accounts of compound processing (e.g., Butterworth, 1983; Bybee, 1995). Our interpretation of these findings is that the low degree of salience of the constituents inhibits the morphological decomposition and facilitates the lexical storage.

As a first step toward constructing a new database of behavioral performance measures for Japanese compound words, this study develops our understanding of compound processing and the underlying organization of mental lexicon. However, further work is required to evaluate whether familiarity and

AoA are powerful predictors of behavioral performance data (i.e., lexical decision times, word naming times) of Japanese verb-verb compounds, and to improve and expand the current dataset to incorporate other lexical (i.e., morphological family size, morphological family frequency) and psycholinguistic variables (i.e., semantic transparency, concreteness, imageability).

### Acknowledgements

This research was supported by JSPS KAKENHI Grant number 22K20032.

### References

- Barbara J. Juhasz (2018). “Experience with compound words influences their processing: An eye movement investigation with English compound words”, *Quarterly Journal of Experimental Psychology*, 71:1, pp.103-112.
- Barbara J. Juhasz, Yun-Hsuan Lai, and Michelle L. Woodcock (2015). “A database of 629 English compound words: Ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience”, *Behavior Research Methods*, 47, pp.1004-1019.
- Brain Butterworth (1983). “Lexical representation”, In Brain Butterworth (Ed.), *Language Production vol.2: Development, writing and other language processes* (pp.257-294), London: Academic Press.
- Ching Chu Sun, Peter Hendrix, Jianqiang Ma, and Rolf H. Baayen (2018). “Chinese lexicon database (CLD): A large-scale lexical database for simplified Mandarin Chinese”, *Behavior Research Methods*, 50, pp.2606-2629.
- Chi-Shing Tse, Melvin J. Yap, Yuen-Lai Chan, Wei Ping Sze, Cyrus Shaoul, and Dan Lin (2017). “The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words”, *Behavior Research Methods*, 49, pp.1503-1519.
- Christina L. Gagné, Thomas L. Spalding, Patricia Spicer, Dixie Wong, Beatriz Rubio, and Karen Perez Cruz (2020). “Is buttercup a kind of cup? Hyponymy and semantic transparency in compound words”, *Journal of Memory and Language*, 113, 104110.
- Dangui Song, and Degao Li (2021). “Psycholinguistic norms for 3,783 two-character words in simplified Chinese”, *SAGE Open*, 11, 215824402110544.  
<https://doi.org/10.1177/21582440211054495>.
- Daniel Schmidtke, Christina L. Gagné, Victor Kuperman, Thomas L. Spalding, and Benjamin V. Tucker (2018). “Conceptual relations compete during auditory and visual compound word recognition”, *Language, Cognition and Neuroscience*, 33:7, pp.923-942.
- David A. Balota, Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman (2007). “The English Lexicon Project”, *Behavior Research Methods*, 39:3, pp.445-459.
- David A. Balota, Michael J. Cortese, Susan D. Sergent-Marshall, Daniel H. Spieler, and Melvin J. Yap (2004). “Visual word recognition of single-syllable words”, *Journal of Experimental Psychology: General*, 133:2, pp.283-316.
- Emmanuel Keuleers, Kevin Diependaele, and Marc Brysbaert (2010). “Practice effects in large-scale

- visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords”, *Frontiers in Psychology*, 1, 174. doi:10.3389/fpsyg.2010.00174.
- Emmanuel Keuleers, Marc Brysbaert, and Boris New (2010). “SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles”, *Behavior Research Methods*, 42, pp.643-650.
- Fritz Günther, Marco Marelli, and Jens Bölte (2020). “Semantic transparency effects in German compounds: A large dataset and multiple-task investigation”, *Behavior Research Methods*, 52, pp.1208-1224.
- Gary Libben, Bruce L. Derwing, and Roberto G. de Almeida (1999). “Ambiguous novel compounds and models of morphological parsing”, *Brain and Language*, 68:1-2, pp.378-386.
- Gary Libben, Christina L. Gagné, and Wolfgang U. Dressler (2020). “The representation and processing of compounds words”, In Vito Pirrelli, Ingo Plag, and Wolfgang U. Dressler (Eds.), *Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon* (pp.336-352), Berlin, Boston: De Gruyter Mouton.
- Georgios Tagalakis, and Mark T. Keane (2006). “Familiarity and relational preference in the understanding of noun-noun compound”, *Memory & Cognition*, 34:6, pp.1285-1297.
- Hélène Giraudo, and Jonathan Grainger (2000). “Priming complex words: Evidence for supralelexical representation of morphology”, *Psychonomic Bulletin & Review*, 8:1, pp.127-131.
- Joan Bybee (1995). “Regular morphology and the lexicon”, *Language and Cognitive Processes*, 10:5, pp.425-455.
- Hadas Velan, and Ram Frost (2011). “Words with and without internal structure: What determines the nature of orthographic and morphological processing?”, *Cognition*, 118:2, pp.141-156.
- Hien Pham, and Harald Baayen (2015). “Vietnamese compounds show an anti-frequency effect in visual lexical decision”, *Language, Cognition and Neuroscience*, 30:9, pp.1077-1095.
- Hongbo Ji, Christina L. Gagné, and Thomas L. Spalding (2011). “Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds”, *Journal of Memory and Language*, 65, pp.406-430.
- James Myers, Yu-chi Huang, and Wenling Wang (2006). “Frequency effects in the processing of Chinese inflection”, *Journal of Memory and Language*, 54:3, pp.300-323.
- Kaitlin Falkauskas, and Victor, Kuperman (2015). “When experience meets language statistics: Individual variability in processing English compound words”, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41:6, pp.1607-1627.
- Kathleen Rastle, Matthew H. Davis, and Boris New (2004). “The broth in my brother’s brothel: Morpho-orthographic segmentation in visual word recognition”, *Psychonomic Bulletin & Review*, 11:6, pp.1090-1098.
- Ludovic Ferrand, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova, and Christophe Pallier (2010). “The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords”, *Behavior Research Methods*, 42:2, pp.488-496.
- Marc Brysbaert (2017). “Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings ‘corrected’ for frequency”, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 70:7, pp.1129-1139.
- Marco Marelli, and Claudio Luzzatti (2012). “Frequency effects in the processing of Italian nominal

- compounds: Modulation of headedness and semantic transparency”, *Journal of Memory and Language*, 66:4, pp.644-664.
- Marcus Taft, and Kenneth I. Forster (1976). “Lexical storage and retrieval of polymorphemic and polysyllabic words”, *Journal of Verbal Learning and Verbal Behavior*, 15, pp.607-620.
- Mercedes Güemes, Carolina Gattei, and Alejandro Wainseboim (2019). “Processing verb-noun compound words in Spanish: Evidence from event-related potentials”, *Cognitive Neuropsychology*, 36:5-6, pp.265-281.
- Mohammad Momenian, Shuk K. Cham, Jafar Mohammad Amini, Narges Radman, and Brendan Weekes (2021). “Capturing the effects of semantic transparency in word recognition: A cross-linguistic study on Cantonese and Persian”, *Language, Cognition and Neuroscience*, 36:5, pp.612-624.
- Patrick Bonin, Betty Laroche, and Alain Méot (2021). “Psycholinguistic norms for a set of 506 French compound words”, *Behavior Research Methods*, 54, pp.393-413.
- Ram Frost, Tamar Kugler, Avital Deutsch, and Kenneth I. Forster (2005). “Orthographic structure versus morphological structure: Principles of lexical organization in a given language”, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31:6, pp.1293-1326.
- Raymond Bertram, Matti Laine, Harald R. Baayen, Robert Schreuder, and Jukka Hyönä (2000). “Affixal homonymy triggers full-form storage, even with inflected words, even in a morphologically rich language”, *Cognition*, 74, B13-B25.
- Raymond Bertram, Robert Schreuder, and Harald R. Baayen (2000). “The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity”, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:2, pp.489-511.
- Robert Schreuder, and Harald R. Baayen (1995). “Modeling morphological processing”, In Laurie B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp.131-154). Hillsdale, New Jersey: Lawrence Erlbaum.
- Victor Kuperman, and Julie A. Van Dyke (2013). “Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers”, *Journal of Experimental Psychology: Human Perception and Performance*, 39:3, pp.802-823.
- Yongqiang Su, Yixuan Li, and Hong Li (2023). “Familiarity ratings for 24,325 simplified Chinese words”, *Behavior Research Methods*, 55, pp.1496-1509.
- 影山太郎 (1993). 『文法と語形成』 ひつじ書房
- 姚一佳 (2020). 「語彙判断課題における日本語母語話者と学習者の複合動詞アクセス」『ことばの科学研究』 21, pp.81-100.

#### 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/>  
 『複合動詞レキシコン』 <https://vvlexicon.ninjal.ac.jp>  
 『現代日本語書き言葉均衡コーパス』 <https://clrd.ninjal.ac.jp/bccwj/>

## 用言の結合価に見る、体言のブランディング

青山 文啓 (桜美林大学大学院)

### Some thoughts on valency and nominals

Fumihiko Aoyama (Graduate Division, J. F. Oberlin University)

#### 要旨

《あそこに結婚式場がある》に対して《あした結婚式がある》という。「結婚式場」も「結婚式」も見出し語〈結婚〉の複合語として辞典には載せられる。こうして《あそこに学校がある》と《あした学校がある》から、見出し語〈学校〉にも二つの用法が認められる。「結婚式場」のように場所《あそこ》と結ばれる建物としての〈学校〉と、「結婚式」のように日時《あした》と結ばれるイベントとしての〈学校〉である。こうした視点から〈学校〉に二つの用法を認める例は市販の辞典にはめずらしい。しかし、一つの用言が体言どうし（あるいは副用言と）を結びつける結合価（文型）に、単文の基礎を置かないかぎり、どのような用法記述（ブランディング）にも体系性見いだせなくなるだろう。助詞の下位区分、二重主語構文、アスペクト、動詞の自他ペア、などから例をあげ、結合価の有用性について論じる。

#### 1. はじめに

ここでは一つの単語をいくつかの用法に分けることを「ブランディング」と仮称する。宮島(1996)が「カテゴリー的多義性」と呼ぶ問題は、ここでいう「ブランディング」に含めて考えたい。宮島(1996)は《大人と子ども》《親と子ども》の二つを比較し、前者は年齢を、後者は親族関係を軸に、並列された二つの体言(nominal)が呼応する例としてあげている。体言を並列させればそこにはゲシュタルト効果が生まれる。つまり、両方の並列表現に現れる〈子ども〉には、ブランチとして分けられるべきべつの側面があることになる。

おなじ問題は以下の(1)(2)の違いにはっきりと現れる。用言(verbal)イルは、体言「夫婦」「公園」を〈子ども〉とそれぞれ組み合わせる（本文中では用言と助詞はカタカナで表記し、見出し語レベルで見た体言を〈 〉に、文節、句、文などその使用例を《 》に示す）。

- (1) 夫婦には子どもが二人いる。
- (2) 公園には子どもが二人いる。

「夫婦」と組み合わせられた(1)の「子ども」は親族としての性格が表面に現れ、「夫婦の実子」という解釈に優先権が与えられる。しかし、(2)のように「公園」などの場所と組み合わせられれば、「子ども」の持つそのような性格は水面下に沈んでしまう。

さらに類例を二対ほどあげておこう。〈結婚〉という単純語は「結婚式場」や「結婚式」という複合語を造る。辞典では単純語を優先して見出し語にするため、用例として提示される可能性しか複合語にはない。用言アルは、建物としての「結婚式場」を「あそこ」などの場所と、イベントとしての「結婚式」を「あした」などの日時と、それぞれ(3)(4)のように結びつける。重要なことは二つある。一つは、用言アルのもとで、建物は場所としか結び合われないが、イベントは日時とも結ばれることである(Lyons 1968, Ch. 8.1)。

- (3) あそこに結婚式場がある。
- (4) あした結婚式がある。

つまり《あしたあそこで結婚式がある》のように、用言アルのもとで三つの体言が結合され、場所の表示にはデ格が、二格の代わりに使われる。こうしてアルには少なくとも二つ



の結合価が区別されるが、イベントの体言が用言アルの結合価に変更を加えると考えてよいだろう（さきほどは同一の体言〈子ども〉を較べたが今回はべつべつの体言である）。用言アルの持つ二つの結合価には、もう一つ見逃せない問題があるが、この点については節をあらためて取りあげたい。

ところで、「結婚式場」と「結婚式」の区別は単一の見出し語〈学校〉には共存する。(5)(6)は二つのべつの結合価だが、おなじく〈学校〉が現れるからである（「あした学校がある」は予定だが、「あしたあそこに学校がある」はバカげた予言としか受け取られない）。

(5) あそこに学校がある。

(6) あした学校がある。

しかし市販の辞典が〈学校〉に二つの用法を区別することはない。辞書業界では「二行項目」という用語があるほど体言の大半は収録語数をかせぎ、その表記を示すためだけにある。アクセントは無視され、《な[がしの さら【皿】》と《な[がしの タ]クシー》とが、単一の見出し語〈ながし【流し】〉のもとに記述される例はおおい(青山 2004)。

宮島(1996)は25年以上前の論考で、私の関わった仕事も紹介されているが、いまでも教えられることばかりである。その眼目は一般の「多義性」から「カテゴリー的多義性」を区別することに注がれる。ここでは「多義性」と「カテゴリー的多義性」とを区別せず、辞典におけるブランディング（用法記述）一般の問題としてあつかい、一つの体言に見られる多義性（用法）を識別するための目安として、用言の結合価に着目する。

## 2. 体言の結合価と、助詞の下位区分

以下、用言を中心に例をあげるが、体言が述部に現れる場合に、結合価の問題が明瞭に意識される例を先に見ておきたい。日本語の助詞が分布によって下位区分されることは暗黙の諒解のようで、ほとんど正面から論じられることはない。(7)(8)(9)に現れる助詞トは格助詞か並列助詞のいずれかである。(7)ではどちらともいえないが、この場合の文節《シャベルと》は(8)のように構成要素として移動させることができる。つまり(7)(8)のトを格助詞に区分することの妥当性は、「同義語」という体言の持つ結合価による。

(7) シャベルとハナスは同義語だ。

(8) ハナスはシャベルと同義語だ。

(9) シャベルとハナスは五段活用だ。

一方、(9)《[シャベルとハナス]は》のように、単一の構成要素内からべつの構成要素として移動できないトは並列助詞と呼ぶしかない。《[シャベルと][ハナスは]》のように、二つの文節をそれぞれ自立させられるかどうかは、述部体言の持つ結合価による。ということは、結合価を想定しなければ助詞の下位区分さえおぼつかないことになる。

## 3. 二重主語構文の拡がり、体言の性格

形容詞などの状態用言を述部に持つ、日本語の二重主語構文はひじょうに幅がひろい。係り助詞ハと助詞ガの組み合わせが表現するのは格役割(case role)ではなく、従属関係（係り）の深浅でしかない（久野 1973 が「総記のガ」と呼ぶのは係りの浅いガである）。このような結合価に流し込まれた「ゾウ」と「ハナ」の指示関係は《ゾウのハナ》であり、(10)と(11)が同義的であることはこの例文とおなじくらい有名である。二重主語構文で気をつけるべきことは二つある。

一つはその際限のなさである。(10)(11)を(12)(13)と較べればそれは明らかだ。つまり(12)(13)の「老人」と「朝」のように指示関係が見いだしにくいものまで、二重主語構文は受け入れるからである。

(10) ズウはハナが長い。

(11) ズウのハナは長い。

(12) 老人は朝が早い。

(13) 老人の朝は早い。

もう一つはこの融通無碍がどこから来るかである。《老人は朝早く起きる》のような動詞述語文が援用され、融通無碍な解釈が行なわれるとすれば、これまで以上に二重主語構文の周辺部を探索してみる余地はありそうである(青山 1998)。

数の少ない形容詞は多義の温床だが、まれに助詞の組み合わせが用法の識別に役立つ。

(14) 彼は料理がうまい。[⇒じょうずだ；とくだ...]

(15) 彼の料理はうまい。[⇒おいしい...]

「料理」は(14)では【能力】だが(15)では【生産物】である。同義語が[ ]内に示した形容名詞(形容動詞語幹)であることは、形容詞との関係を象徴している。宮島(1996)はこれを「カテゴリー的多義性」と呼び、一般の「多義性」から区別する。また、前者が換喩的であるのに対し、後者は隠喩的だという重要な指摘をしている(言い換えれば、単語の用法記述は素朴な素性論にではなく、レトリックに学ぶべきだということだろう：佐藤 1978)。

#### 4. アスペクトと、体言のブランディング

《トマトソースをかける》と《トマトソースをつくる》を較べれば自明だが、「トマトソース」の存在を前提とするか否かは、用言に決定権がある。ということは、アスペクトが用言の性格を換えれば、その結合価のなかに現れる体言の性格も換わらざるをえない。たとえ自他ペアがおなじ体言〈火〉を共有しても、他動詞(16)が行為を宣言するのに対し、自動詞が記述するのは後続する限界点(17)とその状態(18)でしかない。

(16) 火をつけます 《意志》

(17) 火がつけました 《完了》

(18) 火はついてます 《結果》

(19) 火をけします 《意志》

(20) 火がきえました 《完了》

(21) 火はきえています 《結果》

試しに、その反義的な表現を(19)(20)(21)にあげた。これら六つの表現のうち「火」の存在が前提とされるのは(17)(18)(19)だが、その存在／非在はすべての体言に関わるため、ブランディングには分けられない。ツクルの自他ペアは結合価から見ればデキルだが、先ほどの(17)(18)の順序は入れ替わって、以下のような順序になる：《[16]ソースをつくります>[18]ソースをつくっています>[17]ソースができました》。これがおそらく、進行相を表現できる動詞ツクルと、結果相しか表現しない動詞デキルの違いである。

「多義性」から遠く離れてしまった。話題を本筋にもどすため、交通手段の例と、所持物の例をあげることにしたい。三つの状態用言(22)イル、(23)アル、(24)ナイに、共通して現れるのは〈電車〉だろう。

(22) まだホームに電車がいます。 【車輛】

(23) まだ電車はある。 【手段】

(24) もう電車はない。 【手段】

「梅雨前線」「台風」などはイルを除けば、状態用言にはなじまない。

この節を閉じるにあたって自他ペアに現れる体言について考えたい。〈サイフ〉が「私」と組み合わせられた(25)が象徴的な例である。

(25) 私はサイフを落とした。 【所持物／モノ】

(26) テーブルからサイフが落ちた。 【モノ】

このような場合に【所持物】としての解釈が前面に出てくるが、その自動詞では【モノ】としての解釈しか出てこない。前者の同義語は「紛失する」「落下させる」だが、後者では「落下する」だけであり、体言のブランチとすべき問題ではないことが分かる。自他ペアがおなじ体言を共有したとしても同義的であることを保証しないが、〈子ども〉ではブランチを分けるのに〈サイフ〉で分けないとすれば反芻の余地はある。

## 5. さらにアスペクトと体言に関連して

これまで述べてきたのとは逆に、体言が用言のもとでその性格に影響をおよぼす可能性についても考えないわけにはいかない。このことに関連して、『日本文法研究』（久野 1973, Ch. 9; Kuno 1973, Ch. 10）に提示される以下のような仮説について考えよう：

[+状态的]動形詞は、現在時の状態を差し、

[-状态的]動形詞は、未来時の状態を指すか、現在時の習慣的動作、あるいは普遍的動作を指す。

著者自身のことばによれば、この本はもともと米国で出版された Kuno (1973)を、日本語で著述しなおした日本語版である（「はしがき」には英語版の出版年が“1972”と書かれている）。べつの翻訳者が日本語に訳していたらと想像してみたいくなるのは、あまり目にしない用語が日本語版で充てられているからである。「動形詞」は久野が“verbal”に充てた訳語だが、小論では「用言」を充てている。

用語の詮索はべつの機会に廻し、本題に入ることにしよう。久野は“[+状态的]動形詞は現在時の状態を指す”という。先にあげた例文(3)(4)に具体的な固有名を入れて以下の(27)(28)としよう。どちらの例文にも状態用言アルが、つまり久野の[+状态的]動形詞が現れる。しかし、(27)は現在時を指すとも考えられる（無難な解釈は《総称アスペクト》だろう）が、(28)は普遍的・習慣的（つまり総称的）でもなく、未来時を指す以外にない。

(27) 目黒に結婚式場がある。

(28) あした目黒雅叙園で結婚式がある。

ということは、イベントの体言「結婚式」が、状態用言アルを[-状态的]に変更したと考えるしかない。金田一(1950)は-テイルが後続しない動詞を状態動詞とし、それ以外の動詞とに二分した。久野は、金田一(1950)のように明確な基準は示していないが、状態動詞としてあげるのは以下の七つである：〈解ル、出来ル、聞(コ)エル、見エル、要ル、有ル “*aru/ to exist*”, 在ル “*aru/ to have*”)。漢字で書き分け〈有ル “*aru/ to exist*”)と〈在ル “*aru/ to have*”)二種類のアルがあげられる理由は示されていない（Kuno1973 から、必要なところに限りローマ字表記とその英訳を抜粋した）。一方、〈要ル〉はあっても〈居ル〉がないのはなぜか。さらに不思議なのはこうした論考が五十年間放置されてきたことである。しかし、この問題をあらためて取りあげるのは、本章の功績が動詞、形容詞、補助用言（補助動詞）を越えて状態性を認めること、補助用言スギルを論じる後半がこの章の白眉であることをここに記しておきたいからである。

## 6. まとめ

助詞の下位区分、二重主語構文、アスペクト、動詞の自他ペアなどから例をあげ結合価の有用性について論じてきた。「オシエル／オソワル」などの他動詞ペア、「勉強スル」など

の補助用言スル，体言あるいは用言の側の問題とすべきかなど，多義性の話題はつきない。すべてべつの機会を待つことにしたい。

### 謝 辞

小論をまとめるにあたりいろいろなかたにお世話になったが，お名前をあげるのは論述を整理してからにしたい。

### 文 献

- 青山文啓(1998) 二重主語構文と辞書，『言語』27(3)  
青山文啓(2000) 統語論—単語の二重分節を中心として，『一橋論叢』124(4).  
青山文啓(2004) ことばの研究と辞書に記載される情報，『桜美林論叢』31.  
青山文啓(2013) 二つの四階層モデル—『文型』と『構造』のための読書ノート，『基本文型の研究』林四郎[著]，明治図書1960，復刊：ひつじ書房2013  
金田一春彦(1950) 国語動詞の一分類，『言語研究』15，再収：金田一[編](1976)  
金田一春彦[編](1976) 『日本語動詞のアスペクト』むぎ書房  
久野暲(1973) 『日本文法研究』大修館書店  
情報処理振興事業協会(1990) 『計算機用日本語基本形容詞辞書 IPAL(Basic Adjectives)』同技術センター  
佐藤信夫(1978) 『レトリック感覚』講談社[文庫1986/1992]  
宮島達夫(1996) カテゴリー的多義性，『日本語文法の諸問題—高橋太郎先生古希記念論文集』鈴木泰／角田太作[編]，ひつじ書房  
ライオンズ，ジョン(1973) 『理論言語学』國廣哲弥ほか[訳]，大修館書店
- Kuno, S. (1973) *The Structure of the Japanese Language*. Cambridge, Mass.: MIT Press.  
Lyons, J (1968) *Introduction to Theoretical Linguistics*. (上掲翻訳書：『理論言語学』1973)

## 中世期日本語比喩表現の収集の試み

菊地 礼 (国立国語研究所) †

### An Attempt to Collect Japanese Figurative Expressions in the Middle Ages

Rei Kikuchi(National Institute for Japanese Language and Linguistics)

#### 要旨

本稿は、日本の中世期（鎌倉～室町時代）のテキストから比喩を収集する試みについて報告する。現在、日本語の比喩研究は、現代語を中心としてデータベース化が進められ、実証的な研究の機運が高まっている。古語は内省の効かない研究対象であるため、実例ベースの研究が求められる。しかし、日本の古語の比喩を実証的に研究するための研究資源が整備されていない。そのような現状を鑑み、『日本語歴史コーパス』に『分類語彙表増補改訂版』『日本古典対照分類語彙表』の分類語彙表番号を付与したメタデータ「CHJ-WLSP」を用いて、日本の古語、特に中世期のテキストから比喩を収集し、分析に必要な情報のアノテーションを行う。本項はデータベース構築の概要と現状の作業済みデータの分析例を報告する。特に、『方丈記』『虎明本狂言集』から収集した比喩について、比喩の種別、比喩を形成する構文と文法形式、比喩を構成する意味分野について記述する。

#### 1. はじめに

##### 1.1 研究対象と目的

本稿は古代日本語の比喩の実態解明を目的としたデータベースの構築の試みについて述べる。現在、日本語の比喩研究は現代語を中心として進められている。一方で、古語の比喩の研究は進展していない。古語は内省がきかないため、実例をベースにして研究を進める必要がある。しかし、研究資源となる比喩の実例データが整備されていないため、個々の研究者の用例収集に依存している。そこで、『日本語歴史コーパス』（以下、CHJ）に『分類語彙表』及び『日本古典対照分類語彙表』の分類語彙表番号を付与したメタデータである CHJ-WLSP（浅原ほか 2022）を用いて比喩を収集し、関連データをアノテーションした比喩のデータベースを構築する。その構築方法と現状で収集した比喩の分析事例を示す。

##### 1.2 研究背景と問題

先にも述べたように、日本語の比喩研究は現代語を中心的な対象として進められている。概念メタファー理論(Conceptual metaphor theory)を背景とした理論的研究(鍋島 2011、2016)が主である。しかし、近年では加藤ほか(2021)及び Kato, et al (2022)が『現代日本語書き言葉均衡コーパス』(BCCWJ)を用いた比喩の収集によるデータベース構築を行い、加藤(2023)は『比喩表現の理論と分類』(国立国語研究所報告 57)を電子化し、データベースとして活用できるようにしている。また、Komatsubara (2021)・小松原(2023)は「青空文庫」所収の日本近代文学作品から比喩をはじめとしたレトリックを収集し、意味・文法・表現効果に関する情報をアノテーションした「J-fig 日本語レトリックコーパス」を構築して

† re-kikuchi@ninja.ac.jp (○→@)

いる。実証的な比喩研究の機運が高まっている。

一方で、日本語の古語の比喩研究は進展が見られない。これには古語という対象の性質と用例収集の難しさが原因にある。半沢（1979）、半沢（1981）などによる研究もあるが、特定の作品や比喩の種別に偏り、比喩の使用実態の全体像や意味・文法的メカニズムは未解明である。古語は内省のきかない研究対象であり、実例をベースにして研究を進めることが求められる。多門（2006）は中古・中世・近世の散文資料から幅広く用例を収集し、丁寧な記述を施す。古語の比喩の用例集としての使用も意図されているが、電子データではないため、活用の幅は限定的である。つまり、実証的な研究を進めていくうえで必須となるオープンアクセスの形で公開される日本の古語の比喩のデータは存在しない。その状況を鑑みて、古語の比喩のデータベース構築を行う。特に、中世期（鎌倉～室町）の資料を対象とする。これは当該年代が近代語へと変化する過渡期であり、比喩の収集しやすさや比喩の変化を見ることができるといった利点があるためである。

## 2. データベース構築の概要

### 2.1 対象テキスト

CHJ 所収の鎌倉・室町期のテキストを対象とする。対象テキストは、説話（『今昔物語集（本朝部）（1120年以降成立）』『宇治拾遺物語（1213～1221年ごろ成立）』『十訓抄（1252年成立）』）、随筆（『方丈記（1212年成立）』『徒然草（1330～1331年ごろ成立）』）、狂言（『虎明本狂言集（1642年成立）』）とする。これらはジャンル・文体（和漢混交文、擬古文、口語）・時代が異なるため、実態を偏りなく見るために適する。

対象となるテキストには『分類語彙表』及び『日本古典対照分類語彙表』の分類語彙表番号が付与された Excel 形式のメタデータ（CHJ-WLSP）がある。比喩の収集作業には本データを用いた。

1	meta	boundary	orthToken	lemma	wlsp	pSampleID	pStart	類	類ラベル	部門	部門ラベル	中項目	中項目ラベル	分類項目	分類項目ラベル
2	0900竹取	B				20-竹取0900_00001	10								
3	0900竹取		いま	今	1.1641	20-竹取0900_00001	20	1	体	1	関係	16	時間	1641	現在
4	0900竹取		は	は		20-竹取0900_00001	40								
5	0900竹取		むかし	昔	1.1642	20-竹取0900_00001	50	1	体	1	関係	16	時間	1642	過去
6	0900竹取		、	、		20-竹取0900_00001	80								
7	0900竹取		たけとり	竹取		20-竹取0900_00001	90								
8	0900竹取		の	の		20-竹取0900_00001	130								
9	0900竹取		弱	弱	1.2050	20-竹取0900_00001	140	1	体	2	主体	20	人間	2050	老少
10	0900竹取		と	と		20-竹取0900_00001	150								
11	0900竹取		いふ	言う	2.3102	20-竹取0900_00001	160	2	用	3	活動	31	言語	3102	名
12	0900竹取		もの	者	1.2000	20-竹取0900_00001	180	1	体	2	主体	20	人間	2000	人間
13	0900竹取		あり	有る	2.1200	20-竹取0900_00001	200	2	用	1	関係	12	存在	1200	存在
14	0900竹取		けり	けり		20-竹取0900_00001	220								
15	0900竹取		。	。		20-竹取0900_00001	240								

図 1 : CHJ-WLSP

短単位ごとに行が分けられ、それぞれに分類語彙表番号（wlsp）が付与されている。これにより、比喩を構成する短単位の意味分野を確認することもできる。『宇治拾遺物語』『十訓抄』『方丈記』『徒然草』は全文に対して意味分類番号が付与されている<sup>1</sup>。『虎明本狂言集』は一部（「あさいな」「かみなり」「ゑさし」「ばくろう」「せいらい」）に付与され、『今昔物語集（本朝部）』はコアデータのみが付与されている<sup>2</sup>。

<sup>1</sup> 付属語などは『分類語彙表』に該当する番号が無いいため、CHJ-WLSP 内では付与されていない。

<sup>2</sup> コアデータとして用いられているのは、鈴鹿本を底本とした 12 巻・17 巻・20 巻・27 巻・29 巻であ

## 2.2 比喩の収集手順

上記の CHJ-WLSP を上から逐次テキストを確認し、比喩に該当する短単位をマークすることで比喩の収集を行っている。同作業にあたっては、日本古典文学を専攻する作業員一名と稿者のダブルチェックを行っている。

比喩の同定は個人の主観的判断に頼るところが大きく、作業の妥当性を担保することが難しい。そのため、比喩の判定手順である MIP (Metaphor Identification Procedure) (Pragglejaz Group 2007) 及びそれを拡張した MIP-VU (Metaphor Identification Procedure VU University Amsterdam) (Steen et al. 2010) を採用した<sup>3</sup>。本手法は Kato, et al (2022) でも使用しており、日本語の比喩の収集における妥当性は確認されている。これにより修辞性の高い比喩から概念メタファーのような一般に比喩とは気づくことの難しい比喩まで幅広く収集することができる。

(1) くさむらの蛍は、遠く檜のかがり火にまがひ、暁の雨はおのづから木の葉吹く嵐に似たり。

(30-方丈 1212\_00011,4590<sup>4</sup>)

(2) いはむや、深く思ひ深く知らむ人のためには、これにしも限るべからず。

(30-方丈 1212\_00011,6020)

(1) は暁方に降る雨の音を風に散る木の葉の音に喩える表現である。類似した表現は現状のデータ内では確認されず、比較的独自性の高い比喩と判断できる。一方で、(2) は「思ふ」や「知る」といった思考・知覚を「深く」と表現する。これは抽象的な事柄を深さを持つ容器のように表現するものである。これは概念メタファー (Lakoff & Johnson 1980) の一種の「容器のメタファー」と考えられ、人間の認知を支える慣用的な比喩と見られる。

---

る。約 15 万短単位であり、本朝部全体の約 30%の分量である (池上尚 『今昔物語集 (本朝部)』のデータについて)。

<sup>3</sup> 比喩の判定手順は以下の通りとなる。

1. Read the entire text–discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text–discourse
3. (a) For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.  
 (b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be  
 —More concrete  
 —Related to bodily action.  
 —More precise (as opposed to vague)  
 —Historically older.  
 Basic meanings are not necessarily the most frequent meanings of the lexical unit.  
 (c) If the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes, mark the lexical unit as metaphorical.

<sup>4</sup> 用例の出典は (サンプル ID,開始位置) で記す。

## 2.3 アノテーション

収集した比喩に研究を進める上で有用と思われる情報をアノテーションする。アノテーション項目は現在のところ、「比喩種別」「比喩指標」「結合情報」である。

### 2.3.1 比喩種別

比喩の種別をアノテーションする。比喩の種別は国立国語研究所(1977)における比喩の形式的な区別である「指標比喩」「結合比喩」「文脈比喩」に加え、「換喩」「提喩」を採用する<sup>5</sup>。

- (3) 朝に死に夕に生るるならひ、ただ水の泡にぞ似たりける。  
(30-方丈 1212\_00001,2340)
- (4) はてには、朱雀門、大極殿、大学寮、民部省などまで移りて、一夜のうちに塵灰となりにき。  
(30-方丈 1212\_00002,1100)
- (5) その主と栖と無常を争ふさま、いはばあさがほの露に異ならず。或は露落ちて、花残れり。  
(30-方丈 1212\_00001,3570)
- (6) 南、竹の簀子を敷き、その西に閼伽棚をつくり、北によせて障子をへだてて阿弥陀の絵像を安置し、そばに普賢をかき、前に法花経を置けり。  
(30-方丈 1212\_00010,900)
- (7) 二年があひだ、世の中飢渴して、あさましき事侍りき。  
(30-方丈 1212\_00005,330)

(3) は指標比喩と呼ばれる比喩の一種である。指標比喩はおおむね直喩と重なる。「やうなり」「ごとし」「と見ゆ」のような比喩の形式的な指標を有する表現である。ここでは「ただ～ぞ似たりける」という構文を形式的な指標としている。

(4) は結合比喩と呼ばれる比喩の一種である。結合比喩はおおむね隠喩と重なる。「やうなり」などの形式的な指標を有さず、異なる意味的カテゴリーに属する語の結合による表現である。「一夜のうち」の時間表現「一夜」と空間的な内外を表す「うち」が結合され、時間を空間的に表わす表現となっている。

(5) は文脈比喩と呼ばれる比喩の一種である。文脈比喩はおおむね諷喩と重なる。「やうなり」などの形式的な指標を有さず、かつ句・節・文を作り上げている各要素の意味的な結合にも一見逸脱が存在しない表現である。「露落ちて、花残れり」は一文だけを見ると、露のこぼれる花の描写である。しかし、前文の「その主と栖と無常を争ふさま」によって、人事に関する表現であることが理解される。前後文脈との関係で初めて、比喩と判断することができる。

(6) は「普賢菩薩の絵」を「普賢」と省略的に記述する。これは「漱石を読む」(漱石の小説作品を読む)のように、隣接関係をもとにした転義が生じており、換喩と判断される。

(7) は平安京が飢餓に苦しむことを記述したものである。「世」という広いカテゴリーによって、「平安京」というメンバーを指示しており、カテゴリー：メンバー関係の転義が生じ

<sup>5</sup> 換喩と提喩の区別については佐藤(1978)、瀬戸(1995)を参照した。



ている。提喩として判断される。

### 2.3.2 比喩指標

「指標比喩」と判定した例に対して、その形式的な指標に関する情報を記入する。指標比喩は、比喩の形式的な指標を有した表現である。その形式的な指標となる構文全体を「比喩指標」と呼び、構文を構成する要素を「比喩指標要素」と呼ぶ（国立国語研究所 1977）。指標比喩を構成している比喩指標を記述した。

いは	言う				指標比喩	いわば～異ならない	「いはばXに異ならず」
ば	ば						
あさがほ	朝顔						
の	の						
露	露						
に	に						
異	異						
なら	なり						
ず	ず						

図 2：比喩指標の記述

これにより、どのような構文が比喩を形成し、その構文を形成するためにどのような文法カテゴリーに属するアイテムを用いていたかなどを見ることができる。つまり、中世期の比喩の文法形式的な側面を確認することが可能となる。現代日本語の比喩の形式的な指標は国立国語研究所（1977）や鍋島（2016）、小松原（2023）などにより明らかとなっている。どのような文法的な要素が古語の比喩表現の形式的な指標となるかは、現状では確定していない。そのため、本データベース構築により、探索的な収集が有効となる。

### 2.3.3 結合情報

CHJ-WLSP は、『分類語彙表』及び『日本古典対照分類語彙表』の分類語彙表番号を付与したメタデータである。そのため、マークした比喩を構成している語の意味番号も同時に確認することができる。特に、「結合比喩」は、その比喩を構成している語の意味分野の結合を抜き出して記述した。

我	我が						3.1040
おもは	思う						2.3061
ず	ず						
も	も						
無常	無常			結合比喩	3.1200 / 1.5151		31200
の	の						
風	風						1.5151
に	に						
さそは	誘う						2.3520
れ	れる						
、	、						
只今	只今						3.1641
めいど	冥土						1.2600
に	に						

図 3：意味の結合情報の記述

これにより、どのような意味分野の語が結合することで逸脱的な表現を形成し、比喩を創り上げていたかを見ることができる。つまり、中世期の比喩の意味的な側面を確認することが可能となる。図 2 では「無常の風」という結合比喩に対して「3.1200（相-抽象的關係-存在）

ノ 1.5151 (体-自然物及び自然現象-風)」と記述し、意味の結合が確認しやすいようにしている。

### 3. データの分析例

本節では、以上の手順で構築作業を進めているデータベースを用いた分析事例を紹介する。現在、対象テキストのうち『方丈記』『虎明本狂言集』への比喩マーカの付与を終えている。この二つのテキストにおける「比喩の種別」「指標比喩」「結合比喩」を概観する。

#### 3.1 比喩の種別

『方丈記』と『虎明本狂言集』に出現した比喩の種別が表1である。出現数(粗頻度)と1000語(短単位)当たりの出現頻度(調整頻度<sup>6</sup>)を挙げる。調整頻度の値の降順で並べている。

表1: 比喩の種別と頻度

『方丈記』			『虎明本狂言集』		
比喩種別	粗頻度	調整頻度	比喩種別	粗頻度	調整頻度
文脈比喩	58	12.6	文脈比喩	59	12.8
結合比喩	55	11.9	結合比喩	41	8.9
換喩	35	7.6	指標比喩	11	2.4
指標比喩	25	5.4	提喩	11	2.4
提喩	21	4.6	換喩	8	1.7
？ <sup>7</sup>	9	2.0	？	5	1.1
その他 <sup>8</sup>	1	0.2	その他	0	0
合計	204	44.3	合計	135	29.3

『方丈記』から確認する。調整頻度の第一位は文脈比喩(12.6)であり、第二位は結合比喩(11.9)となる。次に、換喩(7.6)が続き、指標比喩(5.4)、提喩(4.6)と続く。

『虎明本狂言集』を確認する。調整頻度の第一位は文脈比喩(12.8)であり、第二位は結合比喩(8.9)である。次に、指標比喩(2.4)と提喩(2.4)が続き、換喩(1.7)と続く。

両者の結果から確認できることを述べる。まず、両テキストとも文脈比喩の出現頻度が最多であり、次点として結合比喩が続く。特に、『虎明本狂言集』は第三位以下の種別の出現頻度が2.5を下回っており、文脈比喩の12.8、結合比喩の8.9を大きく下回っている。比喩として文脈比喩・結合比喩を主に用いる様子が確認できる。

Kato, et al (2022) が『現代日本語書き言葉均衡コーパス』のコアデータからの比喩の収集

<sup>6</sup> 調整頻度の算出にはテキストを構成する短単位の全数を用いた。ただし、記号類(カッコ、句読点など)は除く。『方丈記』の全短単位数は4,607であり、『虎明本狂言集』の全短単位数は4,591である。

<sup>7</sup> 「？」は比喩種別の判定が現状で確定できていない例を示す。

<sup>8</sup> 「たましきの都のうちに棟を並べ、薨を争へる高き賤しき人の住ひは〔後略〕」(方丈記-1212\_00001,850)における枕詞「たましき」を「その他」として収集した。

調査によると、指標比喩が 337 例、結合比喩が 9,983 例<sup>9</sup>、文脈比喩が 1,368 例、換喩が 2,336 例、提喩が 370 例となる。現代日本語では結合比喩が非常に多く出現し、次点で換喩が出現しやすい。文脈比喩が最多で、次点が結合比喩となる『方丈記』『虎明本狂言集』の結果とは異なる比喩の出現分布であることが分かる。現代日本語と古代日本語で用いる比喩の種類が異なるのだと予測される。

### 3.2 指標比喩

本節では『方丈記』『虎明本狂言集』に出現した指標比喩を確認する。特に、指標比喩を構成する比喩指標と比喩指標要素の一覧を見る。比喩指標及び比喩指標要素の認定は国立国語研究所（1977）を参考にした。まずは比喩指標から確認する（表 2、表 3）。

表 2：『方丈記』の比喩指標の一覧

<p>「X、またかくのごとし」「ただ X にぞ似たりける」「いはば X に異ならず」「X がごとく Y になりぬ」「X がごとくして」「X、Y がごとし」「X を Y のごとく Z」「X に異ならず」「X、Y がごとくすれども」「X のたとへにかなへり」「X、Y のごとし」「X、Y に異ならず」「X と見えしかど」「X、たとへば Y がごとし」「X に似たり」「いはば X がごとし」「X のごとくして」「X、Y にたとへつべし」「X はおのづから Y に似たり」「X なり」「X もまた同じ」「X 驚く程の Y」「X に Y を類える」</p>
--

表 3：『虎明本狂言集』の比喩指標の一覧

<p>「X は Y のごとく」「X 程に」「X も Y やうなり」「たださながら X のごとくなり」「たださながら X のごとく Y」「X がごとくなり」「X 心地して」「X やうに Y」「X のごとく」「X は Y のごとく」「X は Y のごとくにして」</p>
---

『方丈記』から確認する。23 種の比喩指標が用いられている。『方丈記』の指標比喩の粗頻度は 25 であったため、同じ比喩指標を二度以上用いる例の少ないことが分かる。

『虎明本狂言集』を確認する。11 種の比喩指標が用いられている。『虎明本狂言集』の指標比喩の粗頻度は 11 であったため、延べ数と異なり数が同じとなる。同じ比喩指標を使い回さない様が確認できる。

両テキストともに、比喩指標の異なり数が多く、様々なヴァリエーションを用いていることが分かる。サンプル数が少ないため即断はできないが、現代語では助動詞「ようだ」を用いた直喩、その中でも連用形「ように」を用いた「X (の) ように Y」が指標比喩（≒直喩）の中でも出現数が突出して多い（小松原 2023）のと対照的な結果である。

次に比喩指標を構成する文法形式である比喩指標要素の一覧を確認する（表 4、表 5）。

<sup>9</sup> ただし、Kato, et al (2022) では比喩の形式的な指標 (MFlags) と結合 (Combinations) の収集であり、この数そのまま指標比喩と結合比喩の数と見なせるかは考慮が必要である。

表 4：『方丈記』の比喩指標要素一覧

品詞	比喩指標要素
助詞	「の」「に」「が」「して」「ども」「しかど」「を」「ぞ」
助動詞	「ごとし」「たり」「けり」「ず」「ぬ」「つ」「き」「べし」「なり」
動詞	「似る」「異なる」「する」「なる」「かなふ」「見ゆ」「たとふ」「驚く」「類ふ」
副詞	「また」「かく」「いはば」「たとへば」「おのづから」
名詞	「たとへ」「程」
形容詞	「同じ」

表 5：『虎明本狂言集』の比喩指標要素の一覧（下線部は表 5 と重なる例）

品詞	比喩指標要素
助詞	「は」「 <u>の</u> 」「 <u>に</u> 」「も」「が」「 <u>して</u> 」
助動詞	「 <u>ごとし</u> 」「やうなり」「 <u>なり</u> 」
動詞	無し
副詞	「ただ」「さながら」
名詞	「 <u>程</u> 」「心地」
形容詞	無し

『方丈記』から確認する。比喩指標要素として出現する品詞には、助詞・助動詞といった付属語から動詞・副詞・名詞・形容詞といった自立語まで確認できる。助詞には「の」「に」「が」「を」といった格助詞から係助詞「ぞ」、接続助詞「して」「ども」「ど」まで確認できる。助動詞には比況の「ごとし」、アスペクトの「たり」「つ」「ぬ」、テンスの「き」「けり」、みとめ方の「ず」、モダリティ「べし」「なり」の出現が認められる。比況助動詞だけが比喩を構成するのではないことが分かる。動詞には、「似る」「異なる」「類ふ」といった類似性に関わる語、「なる」のような変化に関わる語、「かなふ」「見ゆ」「驚く」といった発話者の知覚・感情に関わる語、「たとふ」のようにメタ的に比喩であることに言及する語が確認できる。副詞には前文と後文が同様であることを表す「また」、指示副詞「かく」、自発の「おのづから」、メタ的に比喩であることに言及する「いはば」「たとへば」が確認できる。名詞にはメタ的に比喩であることに言及する「たとへ」や程度の「程」が確認できる。形容詞には同一性を表す「同じ」が確認できる。多様な意味的カテゴリーが比喩を形成していることが分かる。

『虎明本狂言集』を確認する。比喩指標要素として出現する品詞には、助詞・助動詞といった付属語から副詞・名詞といった自立語が認められる。動詞・形容詞は認められない。助詞には「は」「も」といった取り立ての係助詞、「の」「に」「が」といった格助詞、「して」のような接続助詞が確認できる。助動詞には比況の「ごとし」「やうなり」、モダリティの「なり」が認められる。副詞には限定の「ただ」、比況の副詞「さながら」が認められる。名詞には程度の「程」や発話者の感情に関わる「心地」が認められる。

両者の違いを述べる。まず、『方丈記』では「やうなり」の出現が認められないが、『虎明本狂言集』では出現している。「やうなり」は現代語の比喩の代表的な助動詞「ようだ」の

前身である。比喩を表す「やうなり」の用例自体は平安時代から確認できる（永野 1969）。また、『虎明本狂言集』では取り立ての係助詞「は」「も」の出現が認められる。『方丈記』では「X、またかくのごとし」のように無助詞の構文が用いられたのに対し、『虎明本狂言集』では「XはYのごとく」のような主語・主題を助詞でマークしながら叙述する構文が用いられるようになったのである。13世紀初頭の『方丈記』から17世紀前半の『虎明本狂言集』の400年の間に生じた助詞・助動詞の歴史的な変化が比喩の構成にも影響を及ぼす様子が確認できる。

また、比喩指標要素に用いられる助動詞として、『方丈記』ではアスペクト・みとめ方・テンス・モダリティといった多様な文法カテゴリーに関わる文法形式が認められたが、『虎明本狂言集』では比況助動詞とモダリティの「なり」のみである。さらに、『方丈記』では豊富に認められた動詞の比喩指標要素が『虎明本狂言集』では出現しない。多様な意味・文法的カテゴリーを駆使して比喩を形成する『方丈記』と比況助動詞や比況副詞を用いて比喩を形成する『虎明本狂言集』という対比が見える。

### 3.3 結合比喩

最後に、『方丈記』『虎明本狂言集』の結合比喩を確認する。結合比喩は先述したように、結合する要素の意味的な逸脱によって比喩性を獲得する。どのような意味的な結合が見られるかをここでは確認する。ここで「XのY」「XはY」の前項Xを「結合要素①」とし、後項Yを「結合要素②」と呼ぶ。まずは結合の延べ数と異なり数を確認する（表6）。

表 6：結合の延べ数と異なり数

	延べ	異なり
方丈記	55	48
虎明本狂言集	41	23

『方丈記』は異なり数が大きく、様々な種類の結合のパターンが存在することが分かる。一方で、『虎明本狂言集』は『方丈記』と比較すると異なり数が小さく、同じ結合のパターンが複数回出現する様子が分かる。次に、用例数（粗頻度）が2以上の例を表7、表8に掲げる。なお、確認の便宜のため、意味番号は中項目までを示す。

表 7：『方丈記』の結合比喩の意味番号（頻度2以上）

結合要素①	結合要素②	用例数
1.26（体-人間活動の主体-社会）	1.17（体-抽象的關係-空間）	3
3.19（相-抽象的關係-量）	2.30（用-人間活動-心）	3
1.30（体-人間活動-心）	1.56（体-自然物および自然現象-身体）	2
1.30（体-人間活動-心）	3.19（相-抽象的關係-量）	2
1.19（体-抽象的關係-量）	1.16（体-抽象的關係-時間）	2

表 8：『虎明本狂言集』の結合比喩の意味番号（頻度 2 以上）

結合要素①	結合要素②	用例数
1.26（体-人間活動の主体-社会）	1.47（体-生産物および用具-土地利用）	11
3.12（相-抽象的關係-存在）	1.51（体-自然物および自然現象-物質）	4
1.34（体-人間活動-行為）	2.30（用-人間活動-心）	4
1.19（体-抽象的關係-量）	1.16（体-抽象的關係-時間）	2
1.26（体-人間活動の主体-社会）	1.17（体-抽象的關係-空間）	2

『方丈記』と『虎明本狂言集』で頻度 2 以上の結合に重複が無い。一方で、要素ごとには重なりが見られる。結合要素①に用いられる意味分野としては、「1.26（体-人間活動の主体-社会）」「1.19（体-抽象的關係-量）」が重なる。人間社会の事柄や量的な事柄が比喩に用いられやすいことが分かる。結合要素②に用いられる意味分野としては、「2.30（用-人間活動-心）」が重なる。人間の心理に関わる表現の多いことが分かる。

同時代の他のテキストの傾向は現時点で未調査のため即断できないが、『方丈記』は同じ結合を使い回さない傾向が見て取れる。一方で、『虎明本狂言集』では「1.26（体-人間活動の主体-社会）+1.47（体-生産物および用具-土地利用）」の結合が 10 回以上繰り返される。これは「六道の辻」の例であり、地獄を舞台とする場合に出現する固定的な言い回しである。演じることを前提とした狂言作品であり、口頭表現がテキストを構成していることによつて、このような固定的な言い回しが多く出現すると推測される。

#### 4. おわりに

本稿では、古語の比喩の実態解明を目的としたデータベース構築の概要と現時点で収集されたデータを用いた分析事例を示した。サンプル数は少ないが、比喩を構成する文法形式の違いや文体による比喩的な言い回しの出現の違いなどを明らかにした。

本データベースを構築することにより、Kato, et al (2022) や小松原 (2023) が構築している近代語、現代語の比喩のデータベースに、古語の比喩のデータベースが加わる。各時代の比喩を概観することが可能となり、比喩の通史的な研究を可能とする。また、作品ごと、ジャンルごとでの比較も可能となるため、各作品の特徴やジャンルと比喩の運用の関係などを明らかにすることもできる。特に、Kato, et al (2022) は「現代日本語書き言葉均衡コーパス」(BCCWJ) に『分類語彙表』の意味番号を付与した BCCWJ-WLSP (加藤ほか 2019) を用いており、本稿が紹介したデータと構築手法やアノテーション項目などの共通点が多い。これにより、現代と古代で比喩を形成する意味分野の違いなどを確認することもできる。比喩の実証的な通時的な研究を遂行していくことにつながる。

#### 謝辞

本研究は科学研究費助成事業若手研究「中世期日本語比喩表現の収集とその分析」(23K12198) の助成を受けたものである。

#### 参考文献

浅原正幸・池上尚・鈴木泰・市村太郎・近藤明日子・加藤祥・山崎誠 (2022) 「分類語彙

- 表番号を付与した『日本語歴史コーパス』データ」日本語学会 2022 年度春季大会発表予稿集.
- 加藤祥・浅原正幸・山崎誠 (2019) 「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ」『日本語の研究』15 巻 2 号、pp.134-141 : 日本語学会.
- 加藤祥・菊地礼・浅原正幸 (2020) 「『現代日本語書き言葉均衡コーパス』に基づく指標比喩データベース」『自然言語処理』27 巻 4 号、pp.853-887 : 言語処理学会.
- Kato Sachi, Kikuchi Rei, Asahara Masayuki (2022) "Figurative Expression Information Database on 'Balanced Corpus of Contemporary Written Japanese'", 15th Researching and Applying Metaphor Conference, abstract.
- 加藤祥・浅原正幸 (2023) 「『比喩表現の理論と分類』データの電子化および情報付与」『国立国語研究所論集』25 号、pp.1-19 : 国立国語研究所.
- 国立国語研究所 (1977) 『比喩表現の理論と分類』国立国語研究所報告 57 : 秀英社.
- 国立国語研究所編 (2004) 『分類語彙表増補改訂版』大日本図書.
- Komatsubara Tetsuta, 2021, "The Corpus of Japanese Figurative Language : Toward a comprehensive framework for describing figurative language", 『国際文化研究 : 神戸大学大学院国際文化学研究科紀要』55 号、pp.107-134 : 神戸大学大学院国際文化学研究科.
- 小松原哲太 (2023) 「『現代語の助詞・助動詞』の電子化とその応用 : 直喩へのアノテーションの事例」『国立国語研究所論集』24 号、pp.45-58 : 国立国語研究所.
- 佐藤信夫 (1978) 『レトリック感覚』講談社.
- 瀬戸賢一 (1995) 『メタファー思考』講談社現代新書.
- 多門靖容 (2006) 『比喩表現論』風間書房.
- 永野賢 (1969) 「ようだ - 比況 < 現代語 >」(松村明編『古典語現代語助詞助動詞詳説』学燈社、pp.313-318).
- 鍋島弘治朗 (2011) 『日本語のメタファー』くろしお出版.
- 鍋島弘治朗 (2016) 『メタファーと身体性』ひつじ書房.
- 半沢幹一 (1979) 「上代の比喩表現について - 「共通性」と素材との関連から -」『国語学研究』19 号、pp.36-47 : 東北大学文学部国語学研究室内「国語学研究」刊行会.
- 半沢幹一 (1981) 「万葉比喩論序論 - 直喩の認定と表現形式」『共立女子大学文芸学部紀要』27 号、pp.319-347 : 共立女子大学.
- 宮島達夫・石井久雄・安部清哉・鈴木泰 (2014) 『日本古典対照分類語彙表』笠間書院.
- Lakoff, G. and M. Johnson. (1980) *Metaphors We Live By*. Chicago: University of Chicago Press. (渡部昇一・楠瀬淳三・下谷和幸(訳)(1986)『レトリックと人生』東京:大修館書店).
- Pragglejaz Group (2007) "MIP: A Method for Identifying Metaphorically Used Words in Discourse." *Metaphor and Symbol* 22 (1), pp.53-62.
- Steen, G. J., Dorst, A. G., Herrman, J. B., Kaal, A., Krennmayr, T., and Pasma, T. (2010) . *A Method for Linguistic Metaphor Identification*. John Benjamins Publishing

#### 関連 URL

池上尚「『今昔物語集 (本朝部)』のデータについて」(<https://clrd.ninjal.ac.jp/chj/doc/abstract-kamakura-2016.pdf>)

## 近世・近代の日本語及び沖縄語訳聖書の平行・コーパスの構築

宮川 創 (国立国語研究所研究系) †

### Construction of a Parallel Corpus of Early Modern and Modern Japanese and Okinawan Translations of the Bible

So Miyagawa (National Institute for Japanese Language and Linguistics)

#### 要旨

現存の聖書の日本語訳のなかで最も早いものはキリシタン資料における聖書の抄訳である。しかし、まとまった形での現存する最古の聖書和訳は、19世紀前半にギュツラフがシンガポールにて刊行した『約翰福音之伝』(「ヨハネによる福音書」と『約翰上中下書』(ヨハネ書簡3通)である。この後、ベッテルハイムの沖縄語訳と漢和对訳を経て、江戸末期・明治初期にかけて、ヘボンやブラウンなどのプロテスタント訳、明治元訳、ニコライの正教会訳、ラゲのカトリック訳などが出版された。発表者はこれらの翻訳に含まれている「ヨハネによる福音書」の平行・コーパス(日本語7翻訳、沖縄語1翻訳、アイヌ語1翻訳、翻訳元の可能性が高いギリシア語批判校訂版、英訳、ラテン語訳、ドイツ語訳、中国語訳など)を作成し、JSON-LDとしてAPIを通じて外部にデータを提供できるシステムを構築した。最後に、本平行・コーパスを用いて翻訳間の影響関係を分析・視覚化するため、Rのstyloパッケージを用いて諸翻訳間のクラスター分析を行なった。

#### 1. はじめに

本稿は、江戸期と明治期を中心とした日本語および琉球語訳の聖書翻訳の平行・コーパス(並列テキスト集)を構築し、その翻訳間の影響関係を分析するものである。これまでも日本の聖書翻訳は多く研究されているが、それらを一元的にまとめたデジタル・平行・コーパスは存在していない。この研究はその初の試みとなる。このコーパスは、キリシタン資料における聖書抜粋、「ギュツラフ訳」、「ベッテルハイム沖縄語訳(琉球語訳/琉訳)」、「ヘボン/ブラウン訳」、「明治元訳」、「ニコライ/中井訳」、「ラゲ訳」、「大正改訳」など、近世・近代の多くの既存翻訳を元としている。また新たなデータとしてアイヌ語の「バチェラー訳」や「ニコライ/中井訳」の底本の一つとなった教会スラヴ語訳、「ラゲ訳」の底本となったヴルガタ訳ラテン語聖書、そのほか、プロテスタントの聖書翻訳に大きな影響を与えた、「モリソン訳」や「ブリッジマン/カルバートソン訳」の古典中国語訳などを追加している。本稿では、まず、構築中のこのコーパスの内容を述べ、次に設計、そして、最後にパイロット・スタディとして、「ヨハネによる福音書」の諸翻訳のスタイロメトリー(stylometry; 計量文体学/文体統計学)によるクラスター分析をした結果を示す。

---

† so-miyagawa@ninjal.ac.jp

1 本稿は、「言語資源ワークショップ2023」(LRW2023)で発表した内容の縮小版の予稿であり、この予稿に大幅な追記・修正を加えた論文は、『計量国語学』34巻4号(2024年3月20日発行予定)に掲載される予定である。



## 2. コーパスの内容

日本語における聖書翻訳の歴史は、近世のキリシタンの活動に端を発する。ジョン・セーリス (John Saris) による、キリシタン版日本語訳新約聖書の報告<sup>2</sup>はあるものの、その現存は確認されていない。しかし、様々な現存するキリシタン版の日本語資料には、聖書からの抜粋が収録されている。キリスト教修養書である『コンテムツス・ムンヂ』 (*Contemptus Mundi*) のキリシタン版日本語訳<sup>3</sup>は、聖書からの抜粋の日本語訳が豊富に含まれている。このテキストは複数の版が現存しているが、それらの内の一つでローマ字版 (1596 年刊) のものは、ドイツのヴォルフエンビュッテルにあるヘルツォーク・アウグスト図書館に所蔵されており<sup>4</sup>、ラテン語、ポルトガル語、そして日本語の要素が混在している。具体的には、ラテン語のウルガタ (*Vulgata*) 訳聖書からの引用があり、その後日本語訳が示されている。筆者は、ボーフム大学の研究チームと協力して、手書きテキスト自動翻刻ソフトウェア *Transkribus*<sup>5</sup>を用いて、この古典テキストのデジタル翻刻と TEI (*Text Encoding Initiative*) *Guidelines* に準拠した XML (*Extensive Markup Language*) 化を行った<sup>6</sup>。これは「日本語史研究用テキストデータ集」サイト上で公開され、CC BY 4.0 International ライセンスに基づいている<sup>7</sup>。これらのキリシタン資料における聖書翻訳の抜粋は、コーパスに追加する計画である。

一方、日本語への本格的な聖書翻訳は、19 世紀のプロテスタント宣教師カール・フリードリヒ・アウグスト・ギュツラフ (*Karl Friedrich August Gützlaff*) によって行われた。ギュツラフは特に「ヨハネによる福音書」を翻訳し、これが後の翻訳に影響を与えた。『コンテムツス・ムンヂ』や他のキリシタン版資料から抜粋された「ヨハネによる福音書」の部分は、パラレル・コーパスに追加する予定である。この「ギュツラフ訳」によって、日本語聖書翻訳の歴史と言語学的特性をさらに詳細に解析することが可能になる。本研究では、Gallica で公開されている「ギュツラフ訳」のフランス国立図書館本の「ヨハネによる福音書」を、伊波 (1997, 1998a, 1998b, 1999a, 1999b) の翻刻や岩崎 (1984) の翻刻、日本聖書教会 (2006) の翻刻と影印、ギュツラフ (2001) の影印を参考にしながら、デジタル翻刻したものである。

ジャン・バーナード・ベッテルハイム (*Jean Bernard Bettelheim*) はハンガリー出身のユダヤ教徒で、後に英国国教会に改宗し、日本で宣教活動を目指した。しかし、彼は日本本土ではなく、琉球王国 (現在の沖縄県) に到着し、1846 年から 8 年間、ほぼ軟禁状態で那覇の護国寺で過ごした。この期間中に彼は琉球語派・北琉球語群に属する沖縄語を学び、新約聖書のいくつかの部分を沖縄語に翻訳した。この聖書翻訳は、通常「琉球語訳聖書」または「琉訳聖書」と呼ばれ、国文学研究資料館の「国書データベース」で公開されている<sup>8</sup>。本稿では、「ベッテルハイム沖縄語訳」と呼称する。本パラレル・コーパスにおける「ベッテルハイム沖縄語訳」のテキストは、この「国書データベース」にて公開されている画像を、伊波 (1997, 1998a, 1998b, 1999a, 1999b) の翻刻を参考

<sup>2</sup> Saris (1900: 141) 参照。

<sup>3</sup> 現在では、『キリストに倣いて』 (*Imitatio Christi*) というタイトルで知られる。この書は、14-15 世紀の、アウグスチノ会士であったトマス・ア・ケンピス (*Thomas à Kempis*) の著作である。

<sup>4</sup> 岸本・白井 (2019) 参照。

<sup>5</sup> 宮川 (2022) 参照。

<sup>6</sup> ノイツラ・宮川 (2022) 参照。

<sup>7</sup> 「コンテムツス・ムンヂ (HAB 所蔵本)」日本語史研究用テキストデータ集 ([https://www2.ninjal.ac.jp/textdb\\_dataset/cmhab/](https://www2.ninjal.ac.jp/textdb_dataset/cmhab/), 2023 年 8 月 17 日確認)。

<sup>8</sup> <https://kokusho.nijl.ac.jp/biblio/100244876/1?ln=ja> (2023 年 8 月 17 日確認)。

にしながら、デジタル翻刻したものである。

沖縄語での翻訳は、沖縄語の動詞語尾や格助詞など、その特徴を反映しており、琉球歴史的仮名遣いが用いられている。ベッテルハイムは欧州に帰国後、沖縄語と日本語が異なることに気付き、四福音書を漢和对訳で翻訳し、ウィーンで出版した。ギュツラフとベッテルハイムという二人の宣教師による聖書翻訳は、日本及び琉球におけるキリスト教宣教と言語研究に貴重な資料となっている。特にベッテルハイムの沖縄語訳は、沖縄語や琉球文化、そして19世紀の宣教活動に対する深い理解を可能にする。

19世紀後半には、別の宣教師、ジョナサン・ゴープル (Jonathan Goble) が日本に来て、「マタイによる福音書」の日本語訳を横浜で出版した。「ゴープル訳」はデジタル化されていないが、「明治学院大学図書館デジタルアーカイブス」でその画像が公開されている<sup>9</sup>。

ジェームズ・カーティス・ヘボン (James Curtis Hepburn) は19世紀後半に活動した米国長老派教会の医療伝道宣教師で、和英語辞典『和英語林集成』の著者であり、ヘボン式ローマ字の開発者でもある。彼はシンガポールでギュツラフ訳の『約翰福音之傳』を手に入れ、その後日本へ向かい、サミュエル・ロビンス・ブラウン (Samuel Robbins Brown) と共に四福音書の和訳を行った。これらの訳は「ヘボン/ブラウン訳」と呼ばれ、当時の日本での印刷が明治政府によって禁止されたため、上海で印刷された後、日本に輸入された。この「ヘボン/ブラウン訳」は Wikisource から取得され<sup>10</sup>、ルビの形式が正規表現を用いて HTML 形式に変換されている。このテキストは概して漢字表記が少ない。

1875年には、長老派の宣教師クリストファー・カロザース (Christopher Carrothers) 名義で、日本人信徒の加藤九郎が翻訳し、編集した『畧解新約聖書』も出版されたが、この中には「ルカによる福音書」および「ヨハネによる福音書」は含まれていない<sup>11</sup>。

ヘボンやブラウンが属していたのは、カルヴァン派の系譜を継ぐ教派で、横浜を中心に活躍していた。一方、バプテスト派のネイサン・ブラウンは、洗礼 (彼の用語では「浸礼 (しづめ)」) に対する神学的違いから、ヘボンやサミュエル・R・ブラウン (Samuel R Brown) と対立した関係にあった。

このように、19世紀後半の日本では、プロテスタントの宣教師たちが聖書の翻訳において多大な貢献を果たしていた。ヘボンを中心としたプロテスタント宣教師グループは、1872年に標準的な日本語訳を作成するための翻訳委員会を設立した。ヘボン、サミュエル・R・ブラウン、アメリカン・ボードのダニエル・クロスビー・グリーン (Daniel Crosby Greene) や英国聖公会のジョン・パイパー (John Piper) やウィリアム・ボール・ライト (William Ball Wright)、バプテスト派のネイサン・ブラウンなども参加した。しかし、翻訳の際に使用する用語について意見が分かれ、特に「洗礼」か「浸礼 (しづめ)」かでバプテスト派のネイサン・ブラウンが脱会した。この委員会による新約聖書の日本語訳は、1875年から順次出版され、これは「明治元訳」と呼ばれている<sup>12</sup>。

<sup>9</sup> [https://mgda.meijigakuin.ac.jp/bible/book\\_image/1871matgbl2](https://mgda.meijigakuin.ac.jp/bible/book_image/1871matgbl2) (2023年8月17日確認)。

<sup>10</sup> <https://ja.wikisource.org/wiki/%E6%96%B0%E7%B4%84%E8%81%96%E6%9B%B8%EF%BC%88%E3%83%98%E3%83%9C%E3%83%B3%E8%A8%B3%EF%BC%89> (2023年9月30日閲覧)。

<sup>11</sup> 『畧解新約聖書 第一冊馬太』: [https://mgda.meijigakuin.ac.jp/bible/book\\_image/1875matcar](https://mgda.meijigakuin.ac.jp/bible/book_image/1875matcar) (2023年8月17日確認)。『畧解新約聖書 第二冊馬可』: [https://mgda.meijigakuin.ac.jp/bible/book\\_image/1876marcar](https://mgda.meijigakuin.ac.jp/bible/book_image/1876marcar) (2023年8月17日確認)。

<sup>12</sup> <https://ja.wikisource.org/wiki/%E6%98%8E%E6%B2%BB%E5%85%83%E8%A8%B3%E6%>

正教会では、ロシア正教会の宣教師ニコライ（Николай；出生名 Касаткин カサートキン）が 1861 年に日本での宣教を開始し、漢学者で正教徒のパウエル中井木菟麻呂と共に新約聖書を翻訳した。この訳は 1901 年に公刊され、「ニコライ／中井訳」と呼ばれている。このニコライ／中井訳のデータは Wikisource から取得され<sup>13</sup>、2023 年 8 月 17 日時点で一部テキストが不足しているため、それを補完する作業を進めている。

ローマ・カトリック教会からは、1897 年には高橋五郎が司祭のミシェル・スタイシェン（Michael Steichen）の口述から和訳した「スタイシェン／高橋五郎訳」が出ており、これは「ヨハネによる福音書」を含む四福音書の日本語文語訳である。また、1910 年にエミール・ラゲ（Émile Ragué）による翻訳、通称「ラゲ訳」が出版されており、この翻訳のデジタル・テキストも Wikisource で公開されている<sup>14</sup>。

このように、明治時代の日本においては多くの宣教師や学者が関わり、多角的な視点から聖書の日本語訳が行われた。それぞれの教派が独自の方法で翻訳を行い、日本にキリスト教を広めるとともに、日本語としても重要な文献をした。

大正時代以降で重要な文語訳聖書として、1910 年に設立された改訳委員会による 1917 年の『改訳 新約聖書』（「大正改訳」）がある。この「大正改訳」は、翻訳の誤りなど様々な問題があった明治元訳の新約聖書の改訳である。大正改訳では、外国人宣教師だけでなく、多数の日本人キリスト者も参加した。大正改訳は Wikisource でルビなし版として公開されているが、ルビ付き版は「日本語の聖書」というホームページで公開されていることが明らかにされている<sup>15</sup>。

次に、1928 年に永井直治によって発行された『新契約聖書』がある。永井版は、他の翻訳を参照せずに、Stephanus 版の Textus Receptus を底本とした日本語文語訳であり、内村鑑三の序言がついており、国立国会図書館デジタルコレクションからデジタル翻刻されている<sup>16</sup>。全体として、明治以降の聖書の文語体日本語翻訳は、多様な試みと改訂が行われ、デジタル化されたテキストも存在するが<sup>17</sup>、ルビや漢字の読みに一貫性や明確性がない場合もあるという点が強調される。

以下は、今回取り上げなかった聖書翻訳を含めた、1945 年までの言語別の聖書翻訳のリストである。聖書が書別で出版されたり、分冊で出版されたりしている場合は、「ヨ

---

96%B0%E7%B4%84%E8%81%96%E6%9B%B8\_(%E6%98%8E%E6%B2%BB37%E5%B9%B4)  
(2023 年 8 月 17 日確認)。

<sup>13</sup> <https://ja.wikisource.org/wiki/%E6%88%91%E4%B8%BB%E3%82%A4%E3%82%A4%E3%82%B9%E3%82%B9%E3%83%8F%E3%83%AA%E3%82%B9%E3%83%88%E3%82%B9%E3%81%AE%E6%96%B0%E7%B4%84> (2023 年 8 月 17 日確認)。

<sup>14</sup> <https://ja.wikisource.org/wiki/%E6%88%91%E4%B8%BB%E3%82%A4%E3%82%A8%E3%82%BA%E3%82%B9%E3%82%AD%E3%83%AA%E3%82%B9%E3%83%88%E3%81%AE%E6%96%B0%E7%B4%84%E8%81%96%E6%9B%B8> (2023 年 9 月 30 日確認)。

<sup>15</sup> ルビなし版は、[https://ja.wikisource.org/wiki/%E5%A4%A7%E6%AD%A3%E6%94%B9%E8%A8%B3%E6%96%B0%E7%B4%84%E8%81%96%E6%9B%B8\\_\(%E3%83%AB%E3%83%93%E4%BB%98\)](https://ja.wikisource.org/wiki/%E5%A4%A7%E6%AD%A3%E6%94%B9%E8%A8%B3%E6%96%B0%E7%B4%84%E8%81%96%E6%9B%B8_(%E3%83%AB%E3%83%93%E4%BB%98)) (2023 年 8 月 17 日閲覧)。ルビあり版は、<http://bible.salterrae.net/taisho/xml/john.xml> (2023 年 8 月 17 日閲覧)。

<sup>16</sup> <https://dl.ndl.go.jp/pid/1109881> (2023 年 8 月 17 日閲覧)。

<sup>17</sup> <https://ja.wikisource.org/wiki/%E6%96%B0%E5%A5%91%E7%B4%84%E8%81%96%E6%9B%B8> (2023 年 9 月 30 日閲覧)。

ハネによる福音書」の出版年で代表させてある。「ヨハネによる福音書」が含まれない場合は、福音書を中心とした他の書の出版年で代表させてある。

#### 日本語

- 16/7 世紀：キリシタン訳（未発見，ジョン・セーリスによる報告<sup>18</sup>）
- 16/7 世紀：キリシタン資料中の聖書引用
- 1837 年：ギュツラフ訳（「ヨハネによる福音書」及び「ヨハネの手紙一・二・三」）
- 1850 年：ウィリアムズ訳（「マタイによる福音書」など）
- 1871 年：ゴープル訳（「マタイによる福音書」）
- 1872 年：ヘボン／ブラウン訳（「マタイによる福音書」，「マルコによる福音書」，「ヨハネによる福音書」）
- 1873 年：ベッテルハイム漢和对訳（「マタイによる福音書」，「マルコによる福音書」，「ルカによる福音書」，「ヨハネによる福音書」）
- 1875 年：加藤九郎／カロザース訳（「マタイによる福音書」，「マルコによる福音書」）
- 1879 年：N・ブラウン訳（新約聖書）
- 1880 年：明治元訳（新約聖書；後に旧約聖書）
- 1881 年：井深梶之助訳（「マルコによる福音書」）
- 1897 年：スタイシェン／高橋五郎訳（「マタイによる福音書」，「マルコによる福音書」，「ルカによる福音書」，「ヨハネによる福音書」）
- 1899 年：中川藤四郎等訳（詩篇）
- 1901 年：ニコライ／中井訳（新約聖書，「詩篇」など）
- 1910 年：ラゲ訳（新約聖書）
- 1905/11/14 年：左近義弼訳（「マタイによる福音書」など）
- 1917 年：大正改訳（新約聖書）
- 1928 年：永井直治訳（新約聖書）
- 1944 年：塚本虎二訳（新約聖書）

#### 沖縄語

- 1855 年：ベッテルハイム沖縄語訳（四福音書など）

---

<sup>18</sup> Saris (1900: 141) 参照.

## アイヌ語

- 1897年：バチェラー訳（新約聖書、「詩篇」など）

これらの聖書のうち、「ウィリアムズ訳」（1850年）は主に「マタイによる福音書」と「創世記」からなり、「ヨハネによる福音書」は5章9節まで試訳として現存している。「ゴープル訳」（1871年）は「マタイによる福音書」のみである。「ベッテルハイム漢和対訳」（1873年）の「ヨハネによる福音書」は見つかっていない。他の三福音書および使徒行伝は既に発見されている。「カロザース／加藤訳」（1875年）は、「マタイによる福音書」と「マルコによる福音書」のみである。井深梶之助訳（1881年）は、「マルコ福音書」のみである。

### 3. コーパスの設計

初期段階にある本コーパス・プロジェクトは、新約聖書のうち、まず『ヨハネによる福音書』の書翻訳のデジタル翻刻のコーパス化に取り組んだ。

このコーパスの設計には Omeka S というデジタル・アーカイブのコンテンツ・マネジメント・システムを使用した（図1）。Omeka S は文化遺産のデジタル・アーカイブに特化しており、リンクト・オープン・データ・モジュールを実装することで、データを、API (application programming interface) を通して JSON-LD 形式で出力できるように設計されている。

#### 新日本語・琉球語ヘクサブラ ヨハネによる福音書1章1節

タイトル	新日本語・琉球語ヘクサブラ ヨハネによる福音書1章1節
口語訳聖書	初めに言があった。言は神と共にあった。言は神であった。
ギュツラフ訳	ハジマリニ カシコイモノ ゴザル。コノ カシコイモノ ゴクラクトモノ ゴザル。コノ カシコイモノ フゴクラク。
ベッテルハイム訳	ハジマリニ カシコイモノ ヲテ、コノカシコイモノヤ シヤウテイト トモニヲタン カノ カシコイモノヤ シヤウテイド。
ヘボン・ブラウン訳	元始(はじめ) 言(ことば) あり 言(ことば) は 神(かみ) とともにあり 言(ことば) は 神(かみ) なり
明治元訳	太初(はじめ) に 道(ことば) あり 道(ことば) は 神(かみ) と 偕(とも) にあり 道(ことば) は 即(すなはち) 神(かみ) なり
ニコライ・中井訳	太初(はじめ) に 言(ことば) 有(あ)り、言(ことば) は 神(かみ) と 共(とも) に 在(あ)り、言(ことば) は 即(すなはち) 神(かみ) なり。
ラゲ訳	元始(はじめ) に 御言(みことば) あり、御言(みことば) 神(かみ) の 御許(おんもと) に 在(あ)り、御言(みことば) は 神(かみ) にてありたり。
永井直治訳	初に言ありき、また言は神と偕にありき、また言は神なりき。
大正改訳聖書	太初に言あり、言は神と偕にあり、言は神なりき。
Textus Receptus	εν αρχη ην ο λογος και ο λογος ην προς τον θεον και θεος ην ο λογος
King James Version	In the beginning was the Word, and the Word was with God, and the Word was God.
書名	ヨハネによる福音書
章番号	1
節番号、節の番号	1



図1 Omeka S による本パラレル・コーパスのオンラインでの表示

聖書翻訳の多様性を考慮する上で、パラレルコーパスの作成は言語学、神学、文化研

究など多くの学術分野で重要な貢献を果たす可能性がある。まず、メタデータの構造化は研究の基盤を形成する。この目的のために、JSON-LD や RDF といった形式が使用される。これらはメタデータを機械可読かつ相互運用可能な形で整理し、各翻訳の文脈や特性、関連するリソースに対するリンクをする。続いて、Linked Open Data (LOD) の導入により、聖書の各翻訳が存在する広がりの中で他の文献や歴史的事件、地理的場所といった外部リソースとどのように関連しているかを明示する。

次に、デジタルアーカイブ専用で作られたコンテンツ管理システム (Content Management System: CMS) の Omeka S が一元管理と公開のプラットフォームとして機能する<sup>19</sup>。これは大規模なデータベースを整理し、多様な視点からのアクセスを可能とするものである。その上で、API を通じてこのパラレル・コーパスは外部の研究者やデベロッパーにも開かれ、新たな分析手法やアプリケーションの開発が促進される。

さらに、研究成果の信頼性と拡散も重要な側面である。この点で、Omeka S プラットフォーム内で各リソースに明示される引用情報と、SNS シェア機能によって、研究成果はより広範な公衆に届けられる。最後に、柔軟な検索機能が大量の翻訳データを効率的に探索するための手段をする。

以上の手法は聖書翻訳のパラレル・コーパス作成において多層的な貢献を果たす。これにより、言語学、神学、文化研究など多様な研究分野で、聖書翻訳に対するより高度で包括的な理解が可能となる。

#### 4. スタイロメトリーによるクラスター分析

以上のように作成したパラレル・コーパスは、言語学的・文学的・宗教学的・歴史学的など様々な分野で使用できる。この技術によって、コーパスのデータは広く共有されやすくなる。本節では、パイロット・スタディとして簡易な計量文体学 (スタイロメトリー) の分析方法の 1 つであるクラスター分析を行う。日本語と琉球語で翻訳された『ヨハネによる福音書』の計量的類似度分析について報告する。すなわち、スタイロメトリーの手法を用いて、聖書翻訳の成り立ちが文体統計のクラスター分析にどのように反映されるかを検討する。この研究で使用した主要なツールは、R 言語用のパッケージである *stylo* (Eder et al. 2016) である。スタイロメトリーは、テキスト間の文体や類似度を統計解析する分野であり、著者推定や年代推定などに用いられる。翻訳間の影響関係の分析では、スタイロメトリー (文体学の統計解析) を用いている。この手法によって、各聖書翻訳の文体の類似度が試験的に計測され、それを基に翻訳の影響関係を計量的に解析する。このような分析は、翻訳文献の影響力や変遷を理解する上で非常に有用なものとなる。全体として、この研究はデジタル・ヒューマニティーズの手法を活用して聖書翻訳の歴史と影響関係を新たな角度から研究する試みであり、今後の研究に対する重要な基礎となると言える。

*stylo* パッケージは主成分分析、クラスター分析、ネットワーク分析など多くの統計手法をしている。本研究では、クラスター分析を主要な統計手法として用いた。R 言語の実行環境としては *Rstudio* を使用した。入力データはプレーン・テキストであり、言語設定は CJK (中国語・日本語・韓国語) に設定した。分析単位は文字で、*n-gram* は 2 に設定した。

初回の分析では、テキストは主に文字表記 (カタカナか漢字仮名交じり表記か) によってクラスター化されてしまった。この結果は文体の類似度ではなく、文字表記に依存していることを示している。この問題を解消するために、すべてのテキストをカタカナに統一し、再分析を行った。

以下の図 2 は、このカタカナ化を行った後の、*stylo* によるクラスター分析の結果で

<sup>19</sup> 宮川 (2023) 参照。

ある。

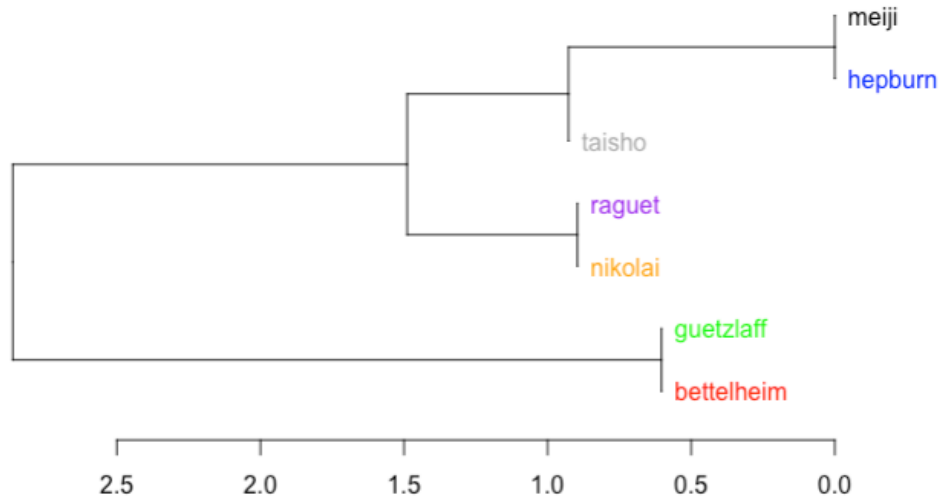


図2 「ヨハネによる福音書」の諸翻訳のクラスター分析

「明治元訳」(meiji)と「ヘボン／ブラウン訳」(hepburn)が最も近いことが示されているが、これは、翻訳者が同じであるため、説明がつく。次に「ギュツラフ訳」(guetzlaff)と「ベッテルハイム沖縄誤訳」(bettelheim)が近いことが示されている。「ギュツラフ訳」は知多半島の方言の影響を受けた言語変種で書かれている一方、「ベッテルハイム沖縄語訳」は、当時の琉球王国の沖縄語で書かれている。言語の違いによる助詞や活用語尾の違いは大きいものの、「ベッテルハイム沖縄語訳」自身が「ギュツラフ訳」を参考にして書かれたため<sup>20</sup>、その類似度を説明できる。次に「ラゲ訳」(raguet)と「ニコライ／中井訳」(nikolai)の近さが示されている。これらは、異なる教派による翻訳であり（「ラゲ訳」はローマ・カトリック教会、「ニコライ／中井訳」はロシア正教会）、かつ、異なる底本の聖書（「ラゲ訳」はラテン語訳聖書、「ニコライ／中井訳」はギリシア語聖書および教会スラブ語訳聖書）から訳されている。しかし、「ニコライ／中井訳」は1901年、「ラゲ訳」は1910年と、比較的近い時期の、漢文訓読文の影響を受けた日本語文語で書かれているため、それらの類似度はある程度説明がつく。次に、類似度が高いのは、「明治元訳」(meiji)と「ヘボン／ブラウン訳」(hepburn)のクラスターと「大正改訳」(taisho)である。「大正改訳」は、「明治元訳」の改訳であるため、容易にその類似度は説明がつく。最後に、「明治元訳」・「ヘボン／ブラウン訳」・「大正改訳」・「ラゲ訳」・「ニコライ／中井訳」がクラスターをなしているが、これらは、日本語文語で書かれているという点で、残りの、標準的な日本語文語が用いられていない「ギュツラフ訳」・「ベッテルハイム沖縄語訳」のクラスターと大きく異なっていることの説明がつく。

このように「ヨハネによる福音書」の諸翻訳において、文体統計のクラスター分析は、翻訳の成り立ちによるグループ分けを明確に示すことができた。すなわち、スタイロメトリーのクラスター分析は日本語と琉球語の「ヨハネによる福音書」の計量的な類似度を分析する有用なツールであることが確認された。ただし、初回の分析では文字表記が大きな影響を与えたため、このような外部要因を制御することが重要であると示された。今後は、この手法をさらに発展させ、多言語・多文化のテキスト分析に応用する可能性が考えられる。

<sup>20</sup> 伊波 (1997) 参照。

以上のように、*stylo* パッケージは多様な分析手法をしており、特に文字表記などの外部要因を制御した上で、高度なテキスト分析が可能であることが示された。この研究は、日本語と琉球語、さらには他の CJK 言語に対するスタイロメトリーの応用可能性を拓く一例である。

## 5. 結論と展望

このパイロット・スタディは、スタイロメトリーを用いて聖書翻訳の文体と成り立ちを解析するための基礎を築いた。今後は、現在構築中の聖書翻訳の追加を完了させるとともに、ラテン語訳聖書や教会スラブ語訳や、日本語聖書翻訳に影響を与えたと先行研究が考えているブリッジマン／カルバートソン訳やモリソン訳などの漢訳聖書を追加する。また、ある程度構築できている「ヨハネによる福音書」のみならず、他の新約聖書の諸書、さらには旧約聖書の諸書にも広げていく。そして、また、スタイロメトリー分析も、クラスター分析だけでなく、主成分分析や多変量解析など *stylo* に装備されている他の分析方法の活用法を模索したり、変数を変更しながら結果の比較を考察したりする予定である。

## 謝 辞

本研究は、人間文化研究機構共同研究プロジェクト（2022 年度～）共創先導プロジェクト共創促進研究「学術知デジタルライブラリの構築」国語研拠点（代表：高田智和）、国立国語研究所共同研究プロジェクト（2022 年度～）広領域連携型基幹研究「異分野融合による総合書物学の拡張的研究」国語研ユニット「古辞書類に基づく語彙資源の拡張と語彙・表記の史的変遷」（代表：高田智和）、国立国語研究所共同研究プロジェクト（2022 年度～）機関拠点型基幹研究「開かれた言語資源による日本語の実証的・応用的研究」基幹型「多様な語彙資源を統合した研究活用基盤の共創」統括班（代表：小木曾智信）、日本学術振興会科学研究費助成事業基盤研究（A）23H00007「日本語諸方言の形態素解析用辞書の構築と活用」（代表：小木曾智信）、日本学術振興会科学研究費助成事業基盤研究（B）19H01265「多言語による日本語学用語辞典および日琉諸語の用例に対するグロス規範の作成」（代表：マシュー・ジスク）の支援を受けている。

## 文 献

- Maciej Eder, Jan Rybicki, and Mike Kestemont (2016). “Stylometry with R: a package for computational text analysis”, *R Journal*, 8:1. pp. 107–21. Available at <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html> (accessed August 17, 2023).
- John Saris (1900). *The voyage of Captain John Saris to Japan, 1613*, Hakluyt Society. Available at <https://archive.org/details/captainjvoyageof00saririch> (accessed August 17, 2023).
- 伊波和正 (1997). 「ベッテルハイム著『琉訳聖書』の分析:『ヨハネ福音書』(『欽定訳聖書』・ギュツラフ日本語訳・ベッテルハイム琉球語訳)比較対照:資料編 I(I-5 章)」『沖縄国際大学外国語研究』 2 :1, pp. 155–206.
- (1998a). 「ベッテルハイム著『琉訳聖書』の分析:『ヨハネ福音書』(『欽定訳聖書』・ギュツラフ日本語訳・ベッテルハイム琉球語訳)比較対照:資料編 II(6-8 章)」『沖縄国際大学外国語研究』 2:2, pp. 235–278.
- (1998b). 「ベッテルハイム著『英琉聖書』の分析:『ヨハネ福音書』(『欽定訳聖書』・ギュツラフ日本語訳・ベッテルハイム琉球語訳)比較対象:資料篇 iii(9-12 章)」『沖縄国際大学外国語研究』 3:1, pp. 255–305.
- (1999a). 「ベッテルハイム著『琉訳聖書』の分析:『ヨハネ副音書』(『欽定訳聖書』・ギュツラフ日本語訳・ベッテルハイム琉球語訳)比較対象:資料編 IV(13-17 章)」



- 『沖縄国際大学外国語研究』 3:2, pp. 179–190.
- (1999b). 「ベッテルハイム著『琉訳聖書』の分析: 『ヨハネ福音書』 (『欽定訳聖書』・ギュツラフ日本語訳・ベッテルハイム琉球語訳)比較対照 : 資料編 V(18-21 章)」 『沖縄国際大学外国語研究』 4:1, pp. 129–168.
- 岩崎摂子 (1984). 『善徳纂「約翰福音之伝」: 本文ならびに総索引』 桜楓社.
- 岸本恵実・白井純 (2019). 「新出本・ヘルツォーク・アウグスト図書館蔵ローマ字本『コンテムツスムンヂ』 (1596 年天草刊) について」 『大阪大学大学院文学研究科紀要』 59, pp. 37–53.
- カール・フリードリッヒ・アウグスト・ギュツラフ (2001). 『約翰福音之伝・約翰上中下書』 覆刻版, 新教出版社.
- 日本聖書協会 (2006). 『ギュツラフ訳 ヨハネによる福音書 - 現代版, 語句の解説つき 抜粋朗読 CD つき』 日本聖書協会.
- ノイツラ・ゾフィー・宮川創 (2022). 「HTR プログラム Transkribus による日本語キリタン版『コンテムツス・ムンヂ』のデジタルアーカイブ化」 『デジタルアーカイブ学会誌』 6:s3, pp. s123–s126.
- 宮川創 (2022). 「Transkribus による手書きテキスト資料の自動翻刻」 一般財団法人人文情報学研究所 (監修)・石田友梨・大向一輝・小風綾乃・永崎研宣・宮川創・渡邊要一郎 (編) 『人文学のためのテキストデータ構築入門 TEI ガイドラインに準拠した取り組みにむけて』 文学通信, pp. 56-109.
- 宮川創 (2023). 「Omeka S を用いた言語資源デジタルアーカイブの構築」 宮川創 『日琉諸語の記述・保存研究』 1, pp. 43-59.

## 日本語教師向け Web コンテンツ「つくば語彙チェッカー」の概要

岩崎 拓也 (筑波大学 人文社会系)  
波多野 博顕 (筑波大学 人文社会系)  
伊藤 秀明 (筑波大学 人文社会系)

### Overview of the Web Content "Tsukuba Vocabulary Checker" for Japanese language teachers

Takuya Iwasaki (Faculty of Humanities and Social Sciences, University of Tsukuba)  
Hiroaki Hatano (Faculty of Humanities and Social Sciences, University of Tsukuba)  
Hideaki Ito (Faculty of Humanities and Social Sciences, University of Tsukuba)

#### 要旨

本発表では、現在構築中の「つくば語彙チェッカー」についての紹介を行う。「つくば語彙チェッカー」は、web上で形態素解析を実行し、「リーディング・チュウ太」「日本語教育語彙表 Ver1.0」「日本語文法項目用例文データベース『はごろも』ver.3」「EDR 電子化辞書」といったデータベースをもとに、入力したテキストの頻度や語彙レベルを一括して確認することができる日本語教師向けのコンテンツである。また、この「つくば語彙チェッカー」は、テキスト中の任意の箇所に空欄補充問題などを作成することが可能な機能も搭載している。本発表では、これらの紹介に加えて、「つくば語彙チェッカー」の使いやすさ向上のためにどのような改修を行っているか、今後どのような機能を実装予定なのかについても紹介する。

#### 1. はじめに

筑波大学グローバルコミュニケーション日本語・日本事情遠隔教育拠点（以下、日本語・日本事情遠隔教育拠点）は、文部科学省による教育関係共同利用拠点の認定拠点として、2010年より日本語教育に関するコンテンツ及びリソースの開発・運用・情報発信を行っている。本発表で紹介する「つくば語彙チェッカー」は、日本語・日本事情遠隔教育拠点が2021年3月より「リーディング・チュウ太」（元東京国際大学 川村よし子氏を中心に開発）の運用を引き継いだことを契機として開発が始まった。

「リーディング・チュウ太」は、1990年代末の公開から25年ほどが経つコンテンツである。入力テキスト中の単語と漢字を旧日本語能力試験の出題基準に合わせてレベル判定を行う機能や、単語の意味を多言語で提示する辞書機能など、その有益さから現在では「日本語教育のデファクトスタンダードの1つ」（北村2013）として、日本語教育機関のみならず地方行政団体など多くの機関で利用されている。また、現在では「日本語教育語彙表」や「機能語用例データベース『はごろも』」など、日本語教育用に構築された他のアノテーション・リソースに基づくwebコンテンツも公開されており、多面的なアプローチで電子テキストの解析が可能な状態になっている。

一方、上記の各コンテンツは、開発・公開時の経緯等から、当然ながらそれぞれが独立したものとして公開されている。そのため、現状では同じテキストを複数のアノテーション・リソースにあたって一度に解析することができない。また、操作方法や出力結果の表示とい

ったユーザビリティの面でも異なっているため、ユーザーはそれぞれのコンテンツの特徴を理解した上で利用する必要がある。また、伊藤・山田（2021）が日本語教師および日本語学習者を対象に「リーディング・チュウ太」のユーザビリティ調査を行った結果、シンプルな作りであってもどのように使うのかということの理解や、コンテンツ独特の用語（宝箱、道具箱など）の使い方がユーザーの負担を高めることを示している。このことから、現代のユーザーを意識したユーザビリティの改善も重要である。

以上の背景から、日本語・日本事情遠隔教育拠点では、川村よし子氏から「リーディング・チュウ太」、日本語学習辞書支援グループ（2015）から「日本語教育語彙表 ver1.0」、「はごろも」プロジェクト（2015）から「機能語用例データベース『はごろも』ver.3」の各データを提供いただくことで、様々なデータベースを一括して検索できるコンテンツ「つくば語彙チェッカー」の開発を進めている。多様なアノテーション・リソースを横断的にかつ簡便に利用できるツールの開発は、日本語教育関係だけでなく多方面の機関で有益なコンテンツとなるだろう。

## 2. 「つくば語彙チェッカー」の特徴

図1に「つくば語彙チェッカー」の現在のトップページのスクリーンショットを示す（なお、今回使用した図は全て現在開発中のコンテンツから引用したものである）。「つくば語彙チェッカー」では、左側にさまざまなタブがあり、そこから各機能に移動することができる。メインの機能としては、形態素解析をはじめとする語彙チェック機能と、問題・テスト作成機能の二つがある。以下では、これらの機能についての紹介を行う。

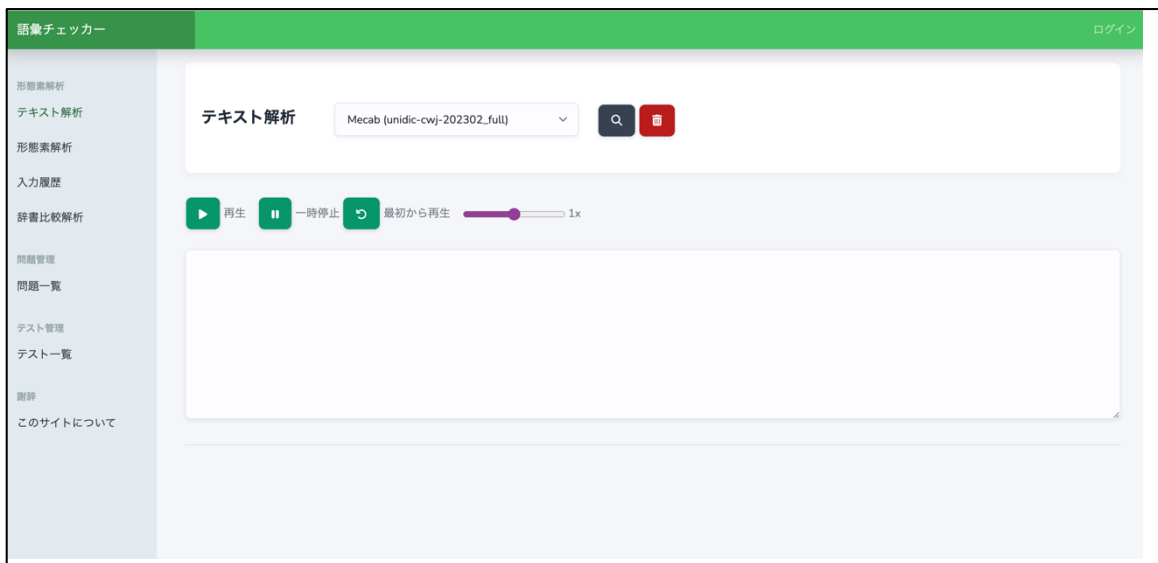


図1 「つくば語彙チェッカー」のトップ画面

### 2.1 語彙チェック機能

「形態素解析」という枠内には、「テキスト解析」「形態素解析」「入力履歴」「辞書比較解析」という四つのタブが存在する。まず、テキスト解析機能について説明する。

「テキスト解析」は、入力したテキストに形態素解析を施した上で「リーディング・チュウ太」「日本語教育語彙表 Ver1.0」「日本語文法項目用例文データベース『はごろも』ver.3」「EDR 電子化辞書」の各データベースに基づき、レベル判定や意味の参照を一括で行う機能である。この際、現在はレベル判定リストとの照合が形態素ベースになるため、派生語や

複合語の一部でオリジナル版と異なる結果となる。この不整合は今後の開発で改善していく予定である。

「テキスト解析」では、「MeCab+unidic-cwj-2023.02」と「Chasen+ipadic-2.7.0」という形態素解析器と形態素解析用辞書の組み合わせを選択して解析ができる。「Chasen+ipadic-2.7.0」を選択可能にしている理由は、もともと「リーディング・チュウ太」が「Chasen+ipadic-2.7.0」によって動作するものであったため、これに合わせた解析結果を出力できるようにするためである。そのため、通常は「MeCab+unidic-cwj-2023.02」による解析で問題ないと思われる。

テキスト解析を行った後の解析結果画面を図2に示す。この解析結果のページでは、入力したテキストの「リーディング・チュウ太」と「日本語教育語彙表 Ver1.0」における語彙レベルのクロス集計表が画面中央に表示される。また、その上には、各データベース名のタブがあり、これをクリックすることによって、画面右側に各レベルや頻度に属する形態素情報が表示される。必要に応じてチェックボックスをクリックして選択することで、まとめてCSV形式で出力結果をダウンロードすることも可能である。

それ以外にも、一括解析の結果を参照したうえで、テキスト中の語を適切なものに修正し、再度解析することが何度でも可能である。テキストから合成音声を作成する機能も備えているため、例えば、想定したレベルに合ったテキストが作成できた場合、それを音声素材として学習者に聞かせることも可能である。



図2 テキスト解析機能の画面

画面左側の機能一覧から移動できる形態素解析機能では、指定した解析器と辞書をもとに形態素解析を実行し、CSV形式で結果をダウンロードできるようになっている(図3)。解析結果の詳細をすぐに確認したい場合や、テキストから属性情報(品詞・語種など)別でリストを作りたい場合などに活用できる。

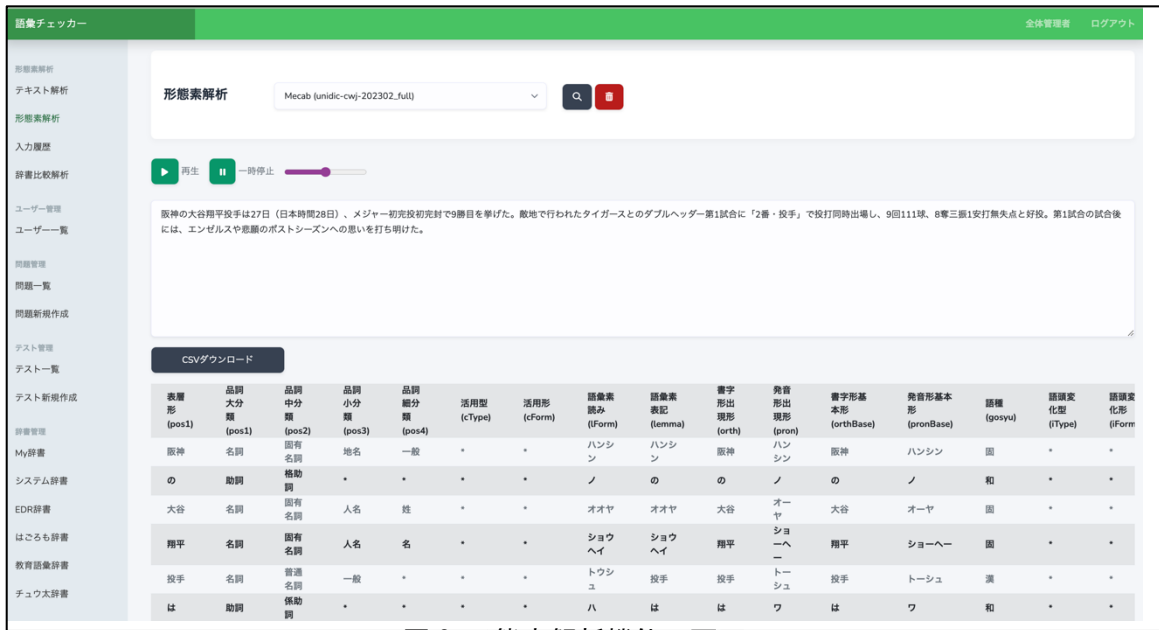


図3 形態素解析機能の画面

そのほかの機能として、「辞書比較解析機能」も搭載されている。この機能は、解析された形態素との照合結果を各リソース（「リーディング・チュウ太」「日本語教育語彙表 Ver1.0」「日本語文法項目用例文データベース『はごろも』ver.3」「EDR 電子化辞書」）に基づいて表示するとともに、その品詞と頻度も併せて一つのリストとして表示するものである（図4）。各形態素がテキストに出現した順で上から表示され、品詞や語彙のレベルを横断的に可視化することに特化している。

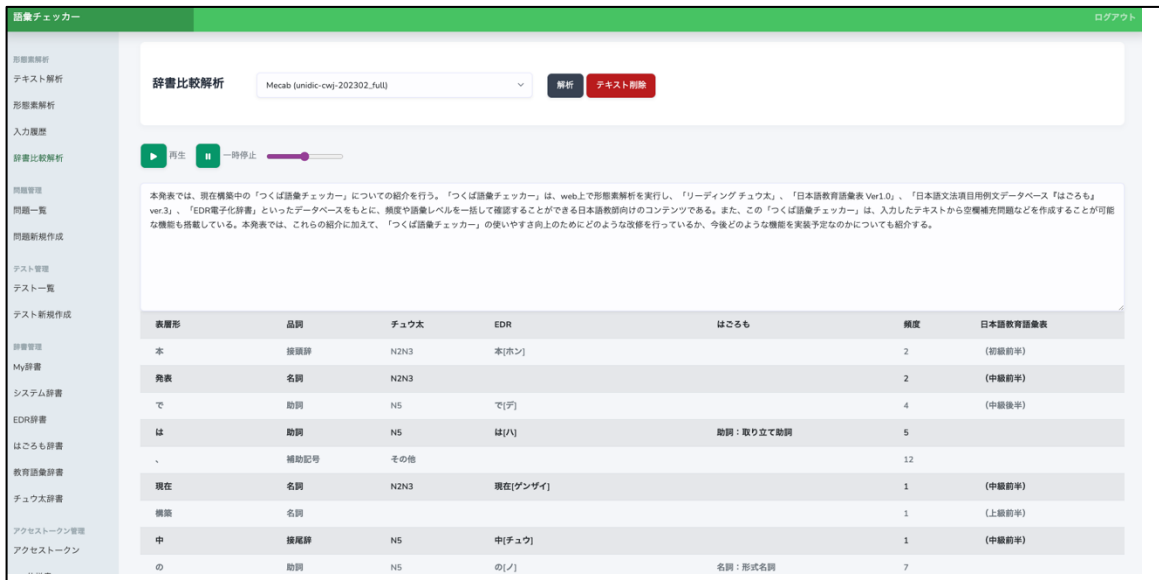


図4 辞書比較解析機能の画面

## 2.2 穴埋め問題作成機能

この「つくば語彙チェッカー」の特徴的な機能の一つに、穴埋め問題作成機能が搭載されている点が挙げられる。任意のテキストを、テキスト解析機能で形態素解析をかけ、その結果を踏まえて穴埋め問題を作成することができる。テキスト解析の結果の画面（図3）の下には、図5のように穴埋め問題作成とテスト名の一覧のスペースが存在する。穴埋め問題

を作成するためには、穴埋めにしたいテキスト部分をマウスオーバーするだけで作成することができる。また、マウスオーバーの順番がどの順番であろうとも、テキストのはじめから（1）（2）のように自動で順番を揃えてくれるようになっている（図6）。



図5 穴埋め問題作成機能の画面



図6 穴埋め問題作成機能による問題の作成画面

こうして作成した穴埋め問題は、あらかじめ作成しておいたテストに組み込むことが可能となっている（このテスト作成機能については現在開発中である）。この方法以外にも、左側にある【問題作成】ボタンをクリックし、【問題詳細】ページに移動して、穴埋め問題の作成、検索用タグの付与、作成した問題／解答の印刷を行うこともできるように計画中である。なお、この機能については現在開発中であるため、今後の改修によって仕様が変更する可能性がある点に注意されたい。

### 3. 現在行っている「つくば語彙チェッカー」の改修について

このように、「つくば語彙チェッカー」では、主に日本語教師のテキスト解析補助のために平易な形態素解析機能をもととした日本語教育関連のデータベースの一括検索を可能とする機能を搭載し、穴埋め問題を中心としたテスト作成機能を有する web コンテンツである。当初は、図 1 のように簡易的なボタンとレイアウトによる UI デザインであり、開発者にしか取り扱いができない、癖のあるデザインとなっていた。そこで、想定されるユーザーである日本語教師にとって使いやすくするために、2名の現役日本語教師と日本語教育を専攻とする大学院生に協力してもらい、使用感を確かめてもらったうえで、感想を提出してもらい、改修を試みている。

現在までには、「リーディング・チュウ太」の解析結果と同じように、解析結果のクロス集計表をデフォルトで表示されるように改修を行った（日本語教育語彙表の解析結果も併せて表示するように変更）。また、検索結果からデータベースごとの結果へのアクセスは、当初は右上に配置されていたが、分かりにくいとのことだったため、画面中央へ配置変更した（図 7）。今回の改修によって追加された機能は、すでに紹介したとおり、一度入力したテキストを再度利用できるように入力履歴のタブの作成、また、入力履歴を検索できるウィンドウの追加である。また、すでに紹介した辞書比較解析機能も新たに追加した機能の一つである。今後は、解析結果を一括して CSV でダウンロードする機能などの機能の追加に加え、テスト作成機能など新機能の搭載も予定されている。



図 7 「つくば語彙チェッカー」のトップページ（改修前後の比較）

また、先述の通り、現状では形態素解析の結果をそのままレベル判定用のリソースと照合させている。そのため、オリジナル版と一部で結果の不整合が起きている。例えば、派生語の「高さ」は、解析結果のままでは「高」と「さ」に分けられるため、各形態素について判定リストと照合してしまう（細かく切りすぎてしまう問題）。派生語・複合語が判定リストにあった場合、それらを一語として処理しレベル判定を行う必要があるため、今後はこのような接続に係る処理を進めていく予定である。

### 4. まとめ：「つくば語彙チェッカー」公開に向けて

以上、本発表では、現在構築中の「つくば語彙チェッカー」の特徴を紹介し、現在行っている改修についても報告した。この「つくば語彙チェッカー」の本公開は 2023 年度末を予定している。なお、今後は「リーディング・チュウ太」が寺・北村・（1996）の読解支援システム「dictlinker」を発展させて開発する際に、機能の充実だけではなく、「使ってもらうための雰囲気作りも重要であることを学んだ」（北村 2013）と述べているように、今回の発表に間に合わなかったテスト機能の追加などの機能の改修を行いつつ、UI 改修についても予定をしている。これらの機能が完成し、日本語教師をはじめとした日本語教育関係者に使用してもらうことで、これまでに日本語教育界に培われてきた言語資源リソースを発展させていきたい。

### 謝 辞

本研究は川村よし子氏、日本語学習辞書支援グループ、「はごろも」プロジェクトの長年の研究およびデータのご提供により成り立っております。この場を借りて深く御礼申し上げます。

### 文 献

- 伊藤秀明・山田智久 (2021). 「日本語学習デジタルリソースのコストパフォーマンスについての検討」『筑波大学シンポジウム 日本語語彙辞書を利用した新たな研究』2021年11月13日 口頭発表.
- 北村達也 (2013). 「日本語読解学習支援システム「リーディング・チュウ太」」『甲南大学紀要 知能情報学編』6:2, pp.243-253.([https://www.konan-u.ac.jp/hp/kitlab/pub/konan\\_kiyou2013.pdf](https://www.konan-u.ac.jp/hp/kitlab/pub/konan_kiyou2013.pdf) よりダウンロード可能).
- 寺朱美・北村達也・浩一郎 (1996). 「WWW ブラウザを利用した日本語読解支援システム」『日本科学教育学会 年会論文集』20, pp.103-104.([https://doi.org/10.14935/jssep.20.0\\_103](https://doi.org/10.14935/jssep.20.0_103) よりダウンロード可能).
- 日本語学習辞書支援グループ (2015). 『日本語教育語彙表 ver.1』,<http://jhlee.sakura.ne.jp/JEV/> (2023年7月27日確認).
- 「はごろも」プロジェクト (2015). 『機能語用例データベース「はごろも」 ver.3』,<http://hgrm.jpn.org> (2023年7月27日確認).



## 近現代語コーパスにおける漢語「是非」

東泉 裕子 (東洋大学)

高橋 圭子 (東洋大学)

### The Sino-Japanese *zehi* ‘right and wrong, by all means’ in Corpora of Modern and Present-day Japanese

Yuko Higashiizumi (Toyo University)

Keiko Takahashi (Toyo University)

#### 要旨

現代日本語において「ゼヒ」は副詞用法が中心であるが、感動詞的用法への拡張も見られる。「ゼヒ」は歴史的に見ると「是」と「非」という漢語の文字通りの意味の名詞から副詞へと意味・用法が拡張し、近代以降に副詞用法の使用頻度が高くなったという。本研究では、書き言葉のコーパスの会話部分ならびに話し言葉のコーパスの会話を利用して、近現代語の会話における「ゼヒ」の用例を調査した。調査の結果、(i)「ニ」を伴わない「ゼヒ」の形式の副詞用法が最も頻度が高く、名詞用法はわずかであること、(ii) 副詞用法の使用比率は20世紀のほうが19世紀後半より高いこと、(iii) 現代語の会話のコーパスでは「ゼヒ」や「ゼヒゼヒ」などの形式で相手の発話への応答としても用いられ、感動詞的用法への拡張が見られることがわかった。

#### 1. はじめに

現代日本語の「ゼヒ」には、(1)のような名詞用法、(2)・(3)のような副詞用法がある<sup>1</sup>。また、(4)・(5)のように相手の発話への応答として用いられることもある。このような用法は感動詞的用法と言える。副詞用法および感動詞的用法の場合、(3)・(5)の「ゼヒゼヒ」のように畳語の形式も観察される。なお、(1)～(5)は「現代日本語書き言葉均衡コーパス(BCCWJ)」および「日本語日常会話コーパス(CEJC)」からの用例である<sup>2</sup>。

- (1) 「あ、いえ、なんでもないのよ。まあね、代理母そのものの是非はともかくとして、自分こそ三田村明の代理母だという女が二人も現れたんじゃあねえ、ご本人の三田村氏も、妻の利香さんもそりゃびっくりするわよねえ」  
BCCW, LBf9\_00174, 35680, 1991年
- (2) 「ゼヒ、あの仔犬がほしいですね。」 BCCW, LBc9\_00163, 12950, 1988年
- (3) IC04\_広瀬 そうですね  
IC02\_義母 うん  
IC04\_広瀬 ゼヒゼヒそんな時はねうーんゆっくりしてください  
CEJC, T001\_009, 41750, 2016年

<sup>1</sup> 以下、本文中では「是非」「ゼヒ」などの複数の表記の総称として「ゼヒ」を用いる。

<sup>2</sup> コーパス検索アプリケーション『中納言』で検索した用例には、順に、コーパス略称、サンプルID/資料ID/会話ID、開始位置、出版年/収録年などを記す。なお、CEJCの発話境界を示す区切りはスペースに置き換える。以下、用例の下線は稿者らによる。

- (4) IC01\_佐竹 あ そうらしいね でもねあたしボルドーのスカートユニクロで買ったけど まだ試してないの  
 IC06\_弓絵 え  
 IC01\_佐竹 あ  
 IC06\_弓絵 ぜひ  
 IC05\_咲乃 ぜひ  
 IC06\_弓絵 だって  
 IC01\_佐竹 じゃあしたその格好で行こうかな  
 CEJC, T011\_007, 1760/1780, 2016 年
- (5) IC02\_牧島 フィリピン行ってみるかな ポピュラーポテンシャル  
 IC01\_根本 ぜひぜひ  
 IC02\_牧島 いい機会  
 CEJC, T018\_017, 2550, 2017 年

以上のような「ぜひ」の多様な用法とそこに至る歴史的過程を明らかにするために、本研究ではぜひの用法を、各種コーパスに基づき調査する。先行研究によれば、副詞用法は名詞用法から生じ、近代以降に使用頻度が高くなるという。そこで、本研究では近現代語の「ぜひ」の用法を調査することにした。

## 2. 先行研究

### 2.1 漢語「是非」の語史

漢語「ぜひ」の通時的変遷については、玉村(1991, 1993, 2018)の詳しい研究がある。玉村の一連の研究によれば、中古までは「ぜひ」は「是」と「非」という文字通りの意味の名詞用法が中心であったが、副詞用法「ぜひ(ニ)」は中世前期から、「ぜひトモ(ニ)」は中世後期から見られるようになる<sup>3</sup>。近世になると、「ぜひ(ニ)」「ぜひトモ(ニ)」は意志・希望や依頼・懇願などの表現の使用が増加し、豊語的な「ぜひ(ニ)」も見られる<sup>4</sup>。また、「ぜひ(ニ)」「ぜひトモ(ニ)」ともに時代が下るにつれて、「ニ」を伴わない形式が多く見られるようになるという。そして、玉村(1993: 67)では、(6)のような豊語的な例について、「感動詞に一步近づいているのかもしれない」と述べている。

- (6) 「真野さん、なにか話を聞かせてよ。面白い話がない？」  
 (中略)  
 「怪談でございます。小菅さん、だいじょうぶ？」  
 「ぜひ、ぜひ」  
 『道化の花』1935年 (玉村 1993: 67)

<sup>3</sup> 「ぜひ」の副詞用法は日本で発生したものであり、中国語には見られない (玉村 1993, 方 2009)。韓国語においても「是非」は個々の漢字の原義のままであるという (Seongha Rhee 氏(マヒドン大学・韓国外語大学)による)。なお、「ぜひ」の副詞用法の成立には、「ぜひニツケテ」(玉村 1991, 1993)、「ぜひナク」(方 2009) という慣用句的な副詞句が関与することが指摘されている。

<sup>4</sup> 豊語的な「ぜひ(ニ)」には、例えば、次のような用例がある。

- (i) 「是非是非シヤリムリニ求ルナリ」(『学談雑論』1716 頃)(玉村 2018: 68)  
 (ii) 「是非に[踊り字]に是非なくも下地は好也御意はよし」(『仮名手本忠臣蔵』1748 年初演)(玉村 2018: 68)

## 2.2 現代語における「ゼヒ」の用法

現代語における(2)のような副詞「ゼヒ」は、グループ・ジャマシイ(2023)によれば、(7)のような意味・用法をもつ<sup>5</sup>。

- (7) a. 「どうしても・かならず」という意味。  
 b. 依頼の表現「てください」、希望の表現「てほしい」などと共に使い、強い願望を表す。ふつう否定の希望表現と共に使わない。単に意志の表現を強めるためには「かならず」などを使う。  
 c. 人間のかかわることにしか使えない。 (グループ・ジャマシイ 2023: 143)

また、山岡他(2018)は、(8)のような例を挙げ、副詞「ゼヒ」は「相手の要求を受諾することが自分にとって負担ではなく、むしろ利益であるとして、相手の心理的負担を軽減しようとする配慮」(山岡他2018: 175)を示す配慮表現としても使用されると述べている。(8a)は相手の許可要求に対する受諾、(8b)は相手の勧誘に対する受諾である。同様の指摘は、森本(1994: 167-168)、福島(2002)にもある。

- (8) a. A 「来週のテニスの親善大会に、私も参加させていただいてよろしいでしょうか。」  
 B 「ええ、(○ゼヒ/△φ)参加してください。」  
 b. A 「来週のテニスの親善大会があるんですが、B さんもいかがですか。」  
 B 「いいですね。(○ゼヒ/△φ)参加させてください。」

(山岡他 2018: 175)

(4)・(5)のような相手の発話への応答として用いられる「ゼヒ」の感動詞的用法については、先行研究において詳しい記述や調査はなされていないようである。ただし、森本(1994: 167-168)は受諾の用法として(9)を挙げている。(9)は相手 P の発話への応答として用いられており、感動詞的用法と言える<sup>6</sup>。

- (9) P: うちへ遊びにいらっしやいませんか。  
 Q: ええ、ゼヒ。(森本 1994: 167)

(3)・(5)のような畳語の形式「ゼヒゼヒ」については、「ゼヒ」とは別に立項する辞書もある(『日本国語大辞典 第二版』、『大辞林 第四版』、『デジタル大辞泉』、グループ・ジャマシイ(2023)など)。

<sup>5</sup> 森田(1977: 259)、飛田・浅田(2018: 216)などにも同様の説明がある。

<sup>6</sup> (4)・(5)・(6)・(9)のような「ゼヒ」は、「相手の発話に応じ、何らかの反応を返す表現」(柏野 2019: 368, 2020: 331)である。柏野(2019, 2020)はこのような表現を「応答表現」と呼んでいる。そして、「うん」「はい」「ああ」などの狭義の感動詞だけでなく、「ですよ」「だろうね」などの文末表現に由来するものや、「いいね」「さすが」「すごい」「なるほど」「確かに」「了解」「もちろん」「だいじょぶ」「だめ」なども「応答表現」とする。また、鈴木(2016)は「反応表現」という語を用いている。

### 3. 調査データ

現代語の「ゼヒ」の多様な用法のうち、感動詞的用法は話者間の相互行為の中に見られる。そこで、本研究では、話し言葉のコーパスは独話以外の会話・対話を、書き言葉のコーパスは会話部分を、調査の対象とした。表1は今回の調査に用いたコーパスを「調査対象期間」の古いものから順に並べたものである。現代日本語研究会(編)(2016)(以下、「現日研(2016)」)を除き、いずれのコーパスも国立国語研究所コーパス開発センターによるものである。

国立国語研究所コーパス開発センターによるコーパスは、『青空文庫パッケージ(AO)』以外は、コーパス検索アプリケーション『中納言』を用いて検索した。短単位モードを使い、キーに語彙素読み「ゼヒ」と語彙素「是非」を指定して検索した。各コーパスについては、以下のような手順で分析対象を決めた。

『日本語歴史コーパス(CHJ)』については、「時代名」は「明治」「大正」を選択して検索を行い、「本文種別」が「会話」で「話者」に情報が入っている用例のみを抽出した。

『現代日本語書き言葉均衡コーパス(BCCWJ)』は、「話者名」に情報が入っている用例のみを抽出した。結果として「図書館・書籍」のデータの会話部分に使用されている「ゼヒ」の用例を分析対象とすることになった。

『昭和話し言葉コーパス(SSC)』は「会話」、『日本語話し言葉コーパス(CSJ)』は「対話」から用例を抽出した。CSJからは「ゼヒ」の用例が8件得られたが、話者交代の位置が不明であるため、今回は分析対象とはしないことにする。SSCの「会話」、『現日研・職場談話コーパス(CWPC)』、『日本語日常会話コーパス(CEJC)』、『名大会話コーパス(NUCC)』については話者交代の位置を確認することができるので、分析対象とした。

以上のような手順により得られた用例は目視で確認し、前後の文脈が不明のものは除去した。検索は、2023年7月から8月に行った。

『青空文庫パッケージ(AO)』については、全文検索システム『ひまわり』ver.1.7.2を利用した。文字列検索で「是非」「ぜひ」をそれぞれ検索し、ランダムに並べ替え、会話部分に現れる「ゼヒ」の用例を集めるために、前後の文脈にカギ括弧のあるもののみを120例ずつ抽出した。

現日研(2016)のデータはエクセルで読み込み、エクセルの検索の機能を使って「ぜひ」を検索し、8例を得た。「是非」は皆無であった。

表1 調査対象コーパスと本研究での分析対象用例数

コーパス名(略称)	調査対象期間	データバージョン	中納言バージョン	分析対象用例数
日本語歴史コーパス(CHJ)	1871-1947	2023.03	2.7.2	385
『青空文庫』パッケージ(AO)	1926-1962	2023-04-06	—	240
現代日本語書き言葉均衡コーパス(BCCWJ)	1976-2007	2021.03	2.7.1	234

昭和話し言葉コーパス(SSC)「会話」 (丸山他 2022)	1952-1969	2022.03	2.7.2	14
現日研・職場談話コーパス(CWPC) (現日研 2011)	1993	2018.03	2.7.2	6
日本語話し言葉コーパス(CSJ) 「対話」	1999-2001	2019.01	2.7.2	
名大会話コーパス(NUCC) (藤村他 2011)	2001-2002	2020.10	2.7.2	40
現代日本語研究会(編) (2016) (現日研 (2016))	2011-2014			8
日本語日常会話コーパス(CEJC) (小磯他 2020)	2016-2018	2023.03	2.7.2	89

#### 4. 調査結果

##### 4.1 近現代語の書き言葉のコーパスの会話における「ゼヒ」

表2は本研究で調査対象とした近現代語の書き言葉のコーパスの会話における「ゼヒ」の用例数と使用比率をまとめたものである。CHJは便宜的に19世紀と20世紀とに二分した。

表2 近現代語の書き言葉のコーパスの会話における「ゼヒ」

用法	形式	CHJ		AO	BCCWJ
		1871-1896	1901-1947	1926-1962	1986-2005
感動詞的	ゼヒ(トモ/ニモ)		1 0.4%	2 0.8%	7 3.0%
		ゼヒ <sup>7</sup>	39 35.8%	216 78.3%	149 62.1%
副詞	ゼヒニ	4 3.7%	6 2.2%	5 2.1%	1 0.4%
	ゼヒニモ			2 0.8%	1 0.4%
	ゼヒトモ <sup>8</sup>	14 12.8%	35 12.7%	38 15.8%	44 19.2%
	ゼヒトモニ	1 0.9%			
	ゼヒゼヒ	3 2.8%	1 0.4%	1 0.4	1 0.4%

<sup>7</sup> CHJ(1871-1896)の副詞用法「ゼヒ」全39例は「ぜつび」1例を含む。

<sup>8</sup> 畳語のような「ゼヒトモ」の用例がAOとBCCWJにあるが、1例と数えた。

(i) 「是非とも、是非ともそうしなくてははいけません」(AO, 山本周五郎『新潮記』1943年)

(ii) 「おお ぜひとも ぜひとも ぜひとも ぜひとも 『庵』へ おいで なせえ」(BCCWJ, Lbk9\_00045, 20440, 1996年)

	ゼヒ(ニ)+引用 <sup>9</sup>	2 1.8%	2 0.7%	6 2.5%	9 3.8%
副詞(計)		63 57.8%	260 94.2%	201 83.8%	219 93.6%
名詞	ゼヒ+助詞	9 8.3%	5 1.8%	5 2.1%	3 1.3%
	ゼヒ+無助詞				1 0.4%
	ゼヒ(ガ/モ)ナイ	25 22.9%	8 2.9%	30 12.5%	4 1.7%
	ゼヒニ及バズ	10 9.2%	2 0.7%		
	その他	2 1.8%		2 1.6%	
名詞(計)		46 42.2%	15 5.4%	37 15.4%	8 3.4%
合計		109	276	240	234

表 2 から、20 世紀以降は「ゼヒ」という形式による副詞用法が中心であることが確認できる。「ゼヒ」という形式による副詞用法の使用比率は、CHJ (1871–1896)は 35.8 %であるのに対し、CHJ (1901–1947)は 78.3%、AO は 62.1%、BCCWJ は 69.7%である。一方、19 世紀、すなわち CHJ (1871–1896) には、名詞用法も 42.2%と少なくないが、その多くは「ゼヒ(ガ/モ)ナイ」・「ゼヒニ及バズ」といった定型表現である(46 例中 35 例、76.1%)。

「ゼヒゼヒ」という疊語の形式の副詞用法は、CHJ (1871–1896)に 3 例、CHJ (1901–1947)に 1 例、AO に 1 例、BCCWJ に 1 例見出された。(10)は本調査における疊語の形式の副詞用法の初出例である。

- (10) 駐春亭で頭取であひとの一座だから金がたりないでひよつとはちをかくといけな  
いからぜひぜひたのむ とつかひをよこされたけれど

CHJ, 60C 口語 1871\_02203, 20430, 1871 年

さらに、玉村(2018: 72)の指摘のとおり、「ゼヒ(ニ)」と「ゼヒトモ(ニ)」については、今回の分析対象からも「ニ」を伴わない形式のほうが「ニ」を伴う形式よりも使用比率が高いことがわかる。各コーパスにおいて、「ゼヒ」は 35.8%~78.3%であるのに対し、「ゼヒニ」は 0.4%~3.7%である。また、「ゼヒトモ」は 12.7%~19.2%であるのに対し、「ゼヒトモニ」は CHJ (1871–1896)に 1 例(0.9%)しか見出せなかった。

感動詞的用法は、少数かつ使用比率が低いが、CHJ (1901–1947)に 1 例(0.4%)、AO に 2 例(0.8%)、BCCWJ に 7 例(3.0%)見出された。(11a)は「ゼヒトモ」、(11b)は「ゼヒニモ」、(11c)は「ゼヒ」の感動詞的用法である。表 2 の数値は、暫定的に、(11a)のような「そりゃあ」や

<sup>9</sup> 玉村(2018: 68)によれば、「さきには是非と仰られけれ共、夫の心うかがはしくてふりきり立かへりし」(『一休咄』1688 年)などのように、「ゼヒ」に続くべき述語が省略され、引用の助詞「と」が下接する例が中世後期から見られるようになり、近代以降に多用されるという。そこで、表 2・表 3 では「ゼヒ(ニ)ト」「ゼヒ(ニ)ッテ」などの用例を暫定的に「ゼヒ(ニ)+引用」に分類した。

(11c)のような「どうぞ」などが前後に現れている場合も感動詞的用法として数えたものである。(11a)の「ゼヒトモ」、(11c)の「ゼヒ」は本調査における感動詞的用法の初出例である。

- (11) a. 『(略) つまり、恐らく貴方はマルトノマー號に冷蔵庫の設備がお要りでせうがな?』『そりやあ、是非とも。』『さう思つてましたよ、あの娘を連れて戻るのに冷蔵庫が要るだらうとね。(略)』 CHJ, 60M 太陽 1925\_05048, 25730, 1925 年
- b. 「いずれ、大坂へも来いよ」秀吉は、まるで彼を、旧友あつかいにし、それ以上、かれの恥じ入るのをみながらなかった。「ぜひにも」と、成政は、礼をのべて、やっと退出した。 AO, 吉川英治『新書太閤記』11, 1939 年~1945 年
- c. 「ええ、しよっちゅう、貴女のことを云って、会いたがっていますよ。それに、路子も、たいへん貴女に、すまながっています。今度、何か機会を作りますから、子供をご覧になりませんか。」 「ぜひ、どうぞ。」 AO, 菊池寛『貞操問答』, 1934 年

#### 4.2 現代語の話し言葉のコーパスの会話における「ゼヒ」

表3は本研究で調査対象とした現代語の話し言葉のコーパスの会話における「ゼヒ」の用例数とその使用比率である。

表3 現代語の話し言葉のコーパスの会話における「ゼヒ」

用法	形式	SSC	CWPC	NUCC	現日研 (2016)	CEJC
		1952-1969	1993	2001-2002	2011-2014	2016-2019
感動詞的	ゼヒ(ゼヒ)			2 5.0%	1 12.5%	22 24.7%
副詞	ゼヒ	14 100.0%	6 100.0%	31 77.5%	7 87.5%	54 60.7%
	ゼヒトモ			2 5.0%		2 2.2%
	ゼヒゼヒ			2 5.0%		7 7.9%
	ゼヒ+引用			2 5.0%		4 4.5%
副詞(計)		14 100.0%	6 100.0%	37 92.5%	7 87.5%	67 75.3%
名詞	ゼヒ+助詞			1 2.5%		
合計		14	6	40	8	89

表3から、現代語の話し言葉のコーパスの会話では、ほとんどが「ゼヒ」という形式で副詞として使用されていることがわかる。名詞用法は1例のみであった。また、「ニ」を伴う形式の「ゼヒニ」と「ゼヒトモニ」の用例は皆無であり、会話においては「ゼヒ」と「ゼヒトモ」が定着していることがうかがえる。(12)は「ゼヒ」、(13)は「ゼヒトモ」の例である。





IC02\_高畑 はい CEJC, T015\_010, 33290, 2017年

- c. F064 F141 も来ない? 絶対おもしろいよー。 発表見物だよ。  
(略)  
F141 ふーん あ、あたしも、でも見たいかも。  
F064 ぜひ  
F141 うん。 おもしろいね。(略) NUCC, data123, 141270, 2001年

(5)のような畳語の形式の「ぜひぜひ」の感動詞的用法は、NUCCに1例、CEJCに7例あった。さらに、「ぜひぜひぜひ」と3回繰り返すものもCEJCに2例あった。(16a)は「ぜひぜひ」、(16b)は「ぜひぜひぜひ」の用例である。

- (16) a. F130 あー 見たかったわ。  
F109 うん、今度ね。  
F130 うん、ぜひぜひ。  
F109 うん。それは結構。 NUCC, data112, 30840, 2002年
- b. IC03\_堀江 これ修理修理させてくれないかなってちょっと思っ  
IC02\_園部 あ ぜひぜひぜひ  
IC01\_健 あー ネットのね あの堀江さんに仕事を提供してあげてくだ  
さい CEJC, S001\_018, 92830, 2017年

## 5. まとめと考察

ここまでの調査結果から、近現代語の「ぜひ」について次のようなことがわかった。

- (i) 20世紀以降は「ニ」を伴わない「ぜひ」による副詞用法が中心であり、(1)のような名詞用法の使用はごくわずかである。
- (ii) 副詞用法の使用比率は20世紀のほうが19世紀後半より高い。
- (iii) 現代語の会話のコーパスでは、「ぜひ」や「ぜひぜひ」などが相手の発話への応答としても用いられており、感動詞的用法への拡張が見られる。

ここで、漢語名詞の副詞化の観点から若干の考察を加えたい。漢語は基本的に名詞として日本語に受容された後に、元の語とは異なる意味・用法をもつようになるものがある(前田1983、趙2013、鳴海2014、2015など)。「ぜひ」についても、名詞から副詞へと意味・用法が拡張し、20世紀以降には「ぜひ」の形式の副詞用法が大半を占めるようになった。そして、現代語の会話では、相手の発話に対する応答としても使用されるようになっている。つまり、副詞用法から感動詞的用法へという意味・用法の拡張が見られる。

「ぜひ」の他にも現代語では感動詞的用法が観察される漢語がある。例えば、「勿論」(高橋・東泉2020)、「無理」(高橋・東泉2021)、「了解」(高橋他2018)などは、それぞれの複雑な過程をたどりつつ、名詞から副詞へ、そして感動詞的用法へという拡張が見られる。

また、「ぜひぜひ」などの畳語の感動詞的用法の類例として、「あるある」「いるいる」(鈴木2016、Ono and Suzuki 2018)、「無理無理」(高橋・東泉2021)、「すごいすごい」「いいいい

い」(Suzuki 2023)などが挙げられる。こうした畳語は、強調や感情の共有などがあると考えられる。

## 6 おわりに

本研究では、近現代語の書き言葉のコーパスの会話部分と話し言葉のコーパスの会話を用いて、近現代語の会話における「ゼヒ」の用例を調査した。そして、20世紀以降は「ニ」を伴わない「ゼヒ」の形式の副詞用法が中心であること、現代語の会話のコーパスでは「ゼヒ」や「ゼヒゼヒ」などの形式で相手の発話への応答としても用いられ、感動詞的用法への拡張が見られることがわかった。「ゼヒ」と同じように、名詞用法から副詞用法へ、さらに感動詞的用法へという拡張が見られる漢語は少なくないが、拡張の時期やプロセスは一樣ではない。このような漢語の調査は今後の課題としたい。

## 謝 辞

LRW2023 において有益なコメントをお寄せくださった皆様に感謝申し上げます。本研究は、日本学術振興会科学研究費補助金による基盤研究(C)「漢字文化圏における漢語の語用論的標識化」(研究代表：高橋圭子、課題番号: 20K00650)、同「漢字文化圏における漢語の語用論的標識の発達」(研究代表：東泉裕子、課題番号: 23K00560)の研究成果の一部です。

## 文 献

- 柏野和佳子(2019)『日本語日常会話コーパス』モニター公開版に見られる応答表現」国立国語研究所『言語資源活用ワークショップ 2019 発表論文集』pp. 368–380.  
<https://doi.org/10.15084/00002589> (2023年8月1日確認)
- 柏野和佳子(2020)『日本語日常会話コーパス』モニター公開版に見られる感動詞以外の応答表現」国立国語研究所『言語資源活用ワークショップ 2020 発表論文集』pp. 331–347.  
<https://doi.org/10.15084/00003173> (2023年8月1日確認)
- 現代日本語研究会(編)(2011)『合本 女性のことば・男性のことば(職場編)』ひつじ書房
- 現代日本語研究会(編)(2016)『談話資料 日常生活のことば』ひつじ書房
- 小磯花絵・天谷晴香・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・渡邊友香(2023)『日本語日常会話コーパス』設計と特徴」『国立国語研究所論集』24, pp. 153–168. 国立国語研究所. <http://doi.org/10.15084/00003692> (2023年8月1日確認)
- 鈴木亮子(2016)「会話における動詞由来の反応表現—「ある」と「いる」を中心に—」井出祥子・藤井洋子(監修)藤井洋子・高梨博子(編)『コミュニケーションのダイナミズム』pp. 63–83. ひつじ書房
- 高橋圭子・東泉裕子(2020)「語用論的標識としての『勿論』の歴史」『東洋大学人間科学総合研究所紀要』22号, pp. 197–208. 東洋大学人間科学総合研究所.  
<https://doi.org/10.34428/00012023> (2023年8月1日確認)
- 高橋圭子・東泉裕子(2021)「語用論的標識としての漢語『無理』の歴史」『東洋大学人間科学総合研究所紀要』23号, pp. 53–74. 東洋大学人間科学総合研究所.  
<https://doi.org/10.34428/00012355> (2023年8月1日確認)
- 高橋圭子・東泉裕子・佐藤万里(2018)『「了解」は使わないように』『了解です!』『言語資源活用ワークショップ 2018 発表論文集』pp. 57–67. 国立国語研究所.  
<https://doi.org/10.15084/00001638> (2023年8月1日確認)
- 玉村禎郎(1991)『是非』の語史：副詞用法の発生まで』『語文』56, pp. 20–38. 大阪大学国語国文学会. <https://hdl.handle.net/11094/68827> (2023年8月1日確認)

- 玉村禎郎(1993)『『是非』』『日本語学』12(7), pp. 66–72. 明治書院.
- 玉村禎郎(2018)「近世における「是非(に／とも／ともに)―副詞用法を中心に―」近代語学会編『近代語研究』第20集, pp. 61–74. 武蔵野書院.
- 趙英姫(2013)「近現代の漢語副詞の成立」野村雅昭(編)『現代日本漢語の探究』pp. 214–233. 東京堂出版
- 鳴海伸一(2014)「漢語形容動詞・副詞の品詞性と用法変化：通時的観点からみた近現代の特徴」『国立国語研究所共同研究報告 13–03 近現代日本語における新語・新用法の研究』pp. 56–75. 国立国語研究所. <http://doi.org/10.15084/00002748> (2023年8月1日確認)
- 鳴海伸一(2015)『日本語における漢語の変容の研究：副詞化を中心として』ひつじ書房
- 方香蘭(2009)「漢語副詞『是非』の成立について」広島女学院大学大学院言語文化研究科編『広島女学院大学大学院言語文化論叢』12号, pp. 102–79 (119–142).
- 福島泰正(2002)「『ぜひ』の機能と使用条件について―聞き手に何かさせることを意図した場合―」『日本語教育』113号, pp 24–33.
- 藤村逸子・大曾美枝子・大島デイヴィッド義和(2011)「会話コーパスの構築によるコミュニケーション研究」藤村逸子・滝沢直宏(編)『言語研究の技法：データの収集と分析』pp. 43–71. ひつじ書房
- 前田富祺(1983)「漢語副詞の変遷」国語語彙史研究会(編)『国語語彙史の研究 四』pp. 189–231. 和泉書院
- 丸山岳彦・小磯花絵・西川賢哉(2022)「『昭和話し言葉コーパス』の設計と構築」『国立国語研究所論集』22号, pp. 197–221. 国立国語研究所. <http://doi.org/10.15084/00003522> (2023年8月1日確認)
- 森本順子(1994)『話し手の主観を表す副詞について』くろしお出版
- 山岡政紀・牧野功・小野正樹(2018)『新版 日本語語用論入門―コミュニケーション理論から見た日本語―』明治書院
- Ono, Tsuyoshi and Ryoko Suzuki (2018) “The use of frequent verbs as reactive tokens in Japanese every day talk: Formulaicity, florescence, and grammaticization”, *Journal of Pragmatics* 123, pp. 209–219.
- Suzuki, Ryoko (2023) “Repeated adjectives as (a)typical clauses in Japanese conversation”, Panel: Pragmatics of the ‘(a)typical clause’ across languages, Organized by Ritva Laury and Tsuyoshi Ono. 第18回国際語用論会議. 2023年7月11日.

### 辞書・辞典

- グループ・ジャマシイ(編)(2023)『日本語文型辞典 改訂版』くろしお出版
- 飛田良文・浅田秀子(2018)『現代副詞用法辞典 新装版』東京堂出版
- 松村明(編)(2019)『大辞林 第四版』三省堂
- 森田良行(1977)『基礎日本語1』角川書店

### 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/>
- 国立国語研究所言語資源開発センター <https://clrd.ninjal.ac.jp>
- コトバンク『デジタル大辞泉』 <https://kotobank.jp>
- ジャパンナレッジ Lib『日本国語大辞典 第二版』 <https://japanknowledge.com/library/>
- 全文検索システム『ひまわり』 <https://csd.ninjal.ac.jp/lrc/index.php?himawari>

## 親子の共同行為場面における振る舞いの調整 —お菓子作りの事例から—

居關 友里子 (国立国語研究所) †

小磯 花絵 (国立国語研究所)

### Coordination of Behavior in Joint Action between Mother and Child: From the Case Study of Cookie Making

Yuriko Iseki (National Institute for Japanese Language and Linguistics)

Hanae Koiso (National Institute for Japanese Language and Linguistics)

#### 要旨

本研究は、子どもが日常で経験している共同行為場面の一つとして、親子でクッキーのかた抜きを行う際のやり取りの記述を行った。活動に参加したのは4歳1ヶ月の男児とその母親である。作業は単純な繰り返し作業ではなく、各回で異なる困難が生じていた。ここで、子どもと母親の両者が、状況と互いの振る舞いに注意をはらい、問題を共有、そして対処の方法を探りながら作業を進めることを通して、一つ一つの型抜きを達成している様子が観察された。これらのやり取りを通して子どもは、作業の進行や問題などに関わる情報を得るために、いつ何に注意をはらうかなどといった、クッキーの型抜きに限られない、より汎用的で基礎的な能力を育むことに繋がる経験を行っていることが示唆される。

#### 1. はじめに

子どもやそれを取り巻く人々の言語使用や身体的振る舞いなど、様々なコミュニケーションの実際について分析するための資源として、現在、国立国語研究所では『子ども版日本語日常会話コーパス (CEJC-Child)』の構築に取り組んでいる (小磯ほか, 2022)。これを構成する主要なデータが、家庭での日常生活の様子を収めた音声・映像資料である。家庭は、子どもたちが日常で経験している中心的なコミュニケーションの場の一つであると言えるだろう。子どもたちはここで、食事、遊び、着替え、片付けや親の手伝いなど、様々な日常を構成する重要な活動について経験し、家族の補助を受けながらできることを増やしていく。

このような家庭で経験される活動には、一つの目標に向けて他者と共に行う活動も含まれる。これらの活動は、他者の意図を理解し、それに応じて振る舞いを調整したり、また自身の欲求を抑えたりといった、社会生活で必要となる重要な能力が関わっている。本研究では、この共同行為がなされる状況の一つとして、お菓子作り場面を扱う。この中で、活動に参加する子どもと養育者が、どのようにお菓子づくりに関わる振る舞いを調整し、作業を進行させているのかについて記述を行い、この機会を通して子どもが得ている経験とは具体的にどのようなものなのかについて考察を行う。

#### 2. 先行研究

##### 2.1 調理・お菓子作りという活動

魚住・池沢 (2016) は、家庭や地域の中で子どもがものを作る経験が今日減少していると

---

† iseki@ninjal.ac.jp

し、その一方で、自ら課題を持ち主体的に取り組むことができるもの作りや体験活動などの重要性について指摘している。このような状況を補う試みと位置づけられるものとして、地域の子ども（主に幼児・児童）とその保護者向けの調理体験のイベントなどが行われており（木村ほか, 2014、加藤, 2022 など）、多くは食育という教育的観点に加え、親子間のコミュニケーション、そして調理そのものを楽しむことを主な目的や成果として報告している。また、自宅での手軽な調理体験に利用されるクッキングトイを扱った研究では、クッキングトイ制作企業側のコンセプトおよび実際に使用した感想の両者において、親子が共に楽しんで作業する（できた）ことに意義や効果を見出している（小口ほか, 2015）。これらのことから、親子の調理に関わる活動は、作業を楽しむことそれ自体や親子の交流を期待して設けられていることが示唆される。一方で、調理・お菓子作り活動で作られるものは、制作後にそれを食べるという行為が控えている。そのため、交流や制作を楽しみながらも、成果物はある程度の水準を満たしている必要がある。ここで保護者は、子どもに適当である作業をその時々判断して体験を促し、必要に応じて援助する必要があるといえる。そして子どもも、保護者の意図する作業を都度読み取り、自身の欲求などと折り合いをつけつつ、作業を進行する必要がある。この共同行為の様相について本研究は焦点を当てる。

## 2.2 子どもの共同行為

本研究では、親子が共にお菓子を作るという共同行為がどのようになされているのかについて見ていく。共同行為 (joint action) とは、2人以上の人が振る舞いを調整することで、環境に変化を生み出す社会的相互行為 (Sebanz et al., 2006) と広く定義される。共同行為には様々な能力が関わっているとされ (Brownell, 2011 など)、子どもの発達研究においてしばしば注目されている。子どもの共同行為に関して、分析の軸となる観点の一つは子どもの年齢である。共同行為における子どもの振る舞いが、自身のゴールに向けた反応によるものなのか、あるいは他者と共有されたゴールを共に達成しようとしてなされたものなのかといった異なりをどのように捉えるかによって、子どもが共同行為を行い始める年齢には開きがある。これらの研究における議論を整理した古畑・板倉 (2016) は、共有されたゴールに向けた振る舞い (後者) を共同行為として扱う場合、これは生後3歳以降に可能になると考えることができると述べている (Hamann et al., 2012 など)。

共同行為と年齢との関係については様々な知見が実験的手法から明らかになってきた。ではこれらの共同行為は、具体的な生活の文脈の中では、子どもたち、またそれらを取りまく人々によってどのように経験されているのだろうか。本研究は、CEJC-Child の強みである、日常にあるありのままの相互行為を捉えたという特徴を利用し、お菓子を作り場面を例に取り、ここでの共同行為における親子の振る舞いの調整がどのように行われているのかについて、やり取りを記述する。

## 3. データ

本研究で扱うのは、『子ども版日本語日常会話コーパス (CEJC-Child)』に格納予定のデータである。収録は、対象となる幼児の保護者 (以下「協力者」) に、毎月1時間程度を目安に1年から4年程度の期間、継続的に行ってもらった。ここに調査者は立ち会わず、協力者に一定期間収録機材を貸し出し、機材の設置・操作に加え、収録をいつ行うのかに関する調整や、収録に参加する人々の同意書の取得、データの提出といった一連の手続きを担ってもらう。使用機材等の詳細については小磯ほか (2022) を参照されたい。

以下で見ていくのは、4歳1ヶ月の男児であるカイと母親によるクッキー作りの場面である(会話ID: Y010\_012)。このうち、カイと母親の二人が積極的に作業に関わる「型抜き」の場面を取り扱う。以下で見る場面では、それ以前に作られたクッキー生地を扱い、(1)生地を伸ばす、(2)型を生地に押し当てて抜く、(3)型に抜かれた生地を鉄板にのせる、という作業が繰り返される。

#### 4. 分析

##### 4.1 初回の作業における調整

はじめに、一回目の型抜きの様子について見て行きたい。親子はどのようなやり取りを行っているのだろうか。

##### [断片 1]初回の型抜き

((母は伸ばした生地にクッキーが何枚分取れそうかを確認し、一枚分しかとれないかもしれないことをカイに説明))

1. 母: どうする? (1.0) 一個ではじめる?
2. カイ: うん. ((ラップで挟まれた生地の上においてあった型に手を添える))
3. 母: >ああここではできないよね?
4. (1.0)
5. 母: はい. 持ってて? ((生地の上のラップに手を掛けながら型をカイに差し出す))
6. (1.0) ((カイ、型を手に持ち、リズムを付けて机を叩く))  
((母、伸ばした生地のラップを外す(L7 まで)))
7. 母: ん:うん. この向きでいいかな.
8. (1.1) ((カイ、生地に型を配置))
9. カイ: これ? ((母を見上げる))
10. 母: うん. ((カイに視線をやり領く))
11. (0.6) ((カイ、型の上に両手を置き、母を見上げる))
12. カイ: じゃあ行くよ:?=
13. 母: =うん.
14. (1.2) ((カイ、型を生地に押し付ける))
15. カイ: お.
16. 母: お.
17. カイ: ° お:.. °
18. (1.3)
19. カイ: お:::. ((型を生地から外す))

結果のみを見ると、親子は一枚目のクッキー生地の型抜きに成功しているが(19行目)、細かな点を見ると、トラブルやそれに対する対処がなされていることがわかる。この点も含め、まずは型抜き作業の流れを追って見ていく。断片以前では、母親が型抜きに合わせて生地を伸ばしながら生地を伸ばしていた。この時カイの手は机の下に置かれており、母の作業を眺めている。そこでなされた母親の質問発話「どうする?(1.0)一個ではじめる?(1行目)」の直後、カイは同意の返答とともに生地に置かれた型に手を添える。このカイの手を

出すタイミングから、カイは母親の 1 行目の発話を一つ目の型抜きを今行うことの提案として聞き、それに応じようとしているものと観察可能である。しかしながら、机の上で伸ばされていた生地はまだラップがかかったままの状態であり、母は 1 行目の発話を、型抜きを今行うことの提案ではなく、生地を伸ばすサイズを決定するための質問として行っており、カイは型抜きの開始を早まったものとして母に理解されている。母が慌てて産出した問題点の指摘「>ああここではできない<よね? (3 行目)」によってカイは手を型から離し、母もその手を支えるようにして型から遠ざける。ここではカイと母親の間に生じた、発話の理解に関わる食い違いにより、型抜きの進行が差し止められている。

そして母親は、今型抜きができない理由となっているラップを取り除く。この時、母親はカイに型を手渡し、持っているように依頼する (4 行目)。カイはそれを受け入れ、リズムをつけて型で机を叩く (6 行目)。ここでのやり取りを通して、母は型を抜くのがカイであることを明示的に予告すること、そしてカイもそれを理解し、出番を待っていることを表示することを行っているといえる。

母が 7 行目の発話とともにラップを剥がすやいなや、カイは素早く手に持った型を生地の上に配置する。ここでは、先にカイの作業が差し止められた理由が何であったかに対するカイの理解が示されており、3 行目の母親の発話が正しく理解されていることが分かる。型の配置を終えたここでカイは、すぐに型を押して生地を抜くのではなく、「これ?」という発話とともに母親に視線をやり、型の配置に問題がないか確認を求める。母親が確認を与えたこと (10 行目) を受け、カイは手のひらを型にかざす形に変更する。これは型を生地に押し込む直前の手型である。そして再び母親を見上げ、「じゃあ行くよ:?!」と型を押す動作の実行を宣言する (12 行目)。これも同様に母親に受け入れられている。これらの、型の配置のやり直しから型を押すまでの一連のやり取りでは、カイは先に生じた型抜きの進行の差し止めを踏まえた上で、型の配置、そして型押しの実行を母親と共同で組み立てているといえる。

最後に、カイは 15 行目「お」で注目に値する変化があったことを発話で表示し、母親もその様子を発話の繰り返しによって受け入れる。さらにカイは「お:...」という発話とともに型を生地から外すことで (19 行目)、自身が今行った型を押す作業が肯定的な成果をあげたことを表示し、母親と共有する機会を生み出そうとしている。

以上が初回の型抜きにおけるカイと母親の振る舞いである。この事例の中では、問題の発生によって、一度型抜きの進行が差し止められるやり取りが生じる。この問題を受け、子どもはその後の進行の中で問題への適切な理解を実演し、また型の配置とそれを押す作業のそれぞれにおいて、確認要求、宣言を行うことにより、母親を共同での型抜きに巻き込み、さらにその成果を共有しようと振る舞っていることが観察された。

#### 4.2 問題への対処に関わる調整

二回目以降の型抜きは、初回の型抜きの問題を踏まえ、同様の流れで進められるかというところではなく、各回ごとに異なる作業の進行や問題の発生、対処といったものが生じていた。この理由の一つは、型を抜かれるクッキー生地の高さが、それ以前の型抜きによって縮小すること、つまり型抜きの難易度や、生地を伸ばし直す手続きの必要性が変化することにある。次に取り上げるのは事例 1 の直後、二回目の型抜きにおける親子のやり取りである。一回目と異なっているのは、型を抜く対象である生地の高さであり、一枚目の型を抜いたことによって、二枚目の型を生地に収めることが難しくなっている。ここではこのような

問題の解決に向けた方策が、双方の身体と言語を用いて交渉されている。

**【断片 2】二回目の型抜き（新たな問題の発生と対処）**

(型抜きした一つ目の生地を鉄板に置き終えた直後)

1. カイ: [(あ)と一個.] (型を生地に乗せる))
2. 母: [一個できた]けど>↑もう一個<できる?
3. カイ: [\*ん.]
4. 母: [あ+>はみだ-はみ-見て見て\*見て.<はみ出て\*るよ.+  
母 +-----生地の型からはみ出した箇所を指差し-----+  
カイ \*型に手をかざす-----\*手をどける-----\*
5. \* (1.5)  
カイ \*型の位置を動かす-->
6. 母: パズルの形に>なってないもん.+\*ほら.\*<  
母 +型の一部を指差す手型で型をつまみ位置をずらす->  
カイ ----->\*手を型にかざしすぐよける\*
7. (0.5)
8. 母: \*ここも足り-足りて\*ない.+  
母 ----->+  
カイ \*母の指先箇所を覗き込む\*
9. 母: もう一回やろ?
10. カイ: う\*ん.=  
カイ \*手のひらを型にかざし母を見上げる
11. 母: =もう一回+こ\*ねこねしてからやろ?+  
母 +型に手を近づける-----+型をつまむ-->>  
カイ \*手をよける

断片冒頭で、カイは直前に行った一回目の型抜きと同様の流れで、二枚目のクッキーを型で抜くことに取り掛かる。一回目で用いたものと同じ、パズルの形をしたクッキー型を生地に配置しながら「(あ)と一個. (1行目)」という発話によって自身が行わんとしている作業を宣言するカイに対し、母は問い返しの発話を返す(2行目)。これは一旦作業の進行を容認する発話として聞かれる。しかしながら、直後にカイが型を押そうとしたところで、母親が問題に気づきカイに対処を促す。ここで生じている問題は、生地の残りの面積が小さくなったことによって、型の全体が生地の中に収まらないというものである。母は4行目「はみ出している」という表現の産出を途中でせき止め(「はみだ-はみ-」)、かつこれを早口で産出してカイに緊急性を伝え、型抜き動作の実行が行われる前にこれを取りやめさせようとしている。そして、指差しと「見て見て見て」という発話によって、問題箇所にカイの注意が向くよう誘導し、その上で問題点の指摘「はみ出てるよ」を産出し直す(4行目)。ここでは、指摘した箇所を両者が見ることによって問題を共有し、カイが自ら作業を修正することが促されている。



カイが型を押す動作を取りやめ（4行目）、型の位置を慎重にずらすという適切な振る舞いによって（5行目）、ここで生じている問題が「生地に対する型の位置」についてであったことが、問題なく共有されたことが母親に理解可能になっている。しかしながら、6行目まで続いた配置の修正によっても型は生地に収まらず、結局問題は解決されなかった。この試行の中には、カイと母親の両者に、問題解決への志向が示されている。例えば、母親は問題の指摘に用いる表現を「(型が生地から) はみ出てる」から「(抜かれるクッキーの形が) パズルの形になってない(6行目)」「(抜かれるクッキーの形をパズル型にするには生地が) 足りてない(8行目)」という、抜き終えたクッキー生地の形を想像させ、そこに問題が生じてしまうという表現に変更し、カイがより問題を具体的に想像できるよう工夫している。一方のカイも、かざした手をよけることで（6行目）、問題が解消されていない可能性への志向を示す。また、母親の指差し箇所を覗き込むように見ており（8行目）、指摘されている問題が視覚で確認されるべき類のものであり、よく見ることが問題解決と結びついているという理解がここに表示されている。

問題の共有と対処が十分に試みられ、その上でこの問題が解決困難であることがさらに共有された後、母親は型の移動ではない対処として「もう一回やろ?」と、やり直しを提案する。やり直す作業が不明瞭であったこの発話は、カイによって型を押すことと理解される（10行目）。母親はこの発話の「もう一回」「やろ」の部分はそのままに、「こねこねしてから」を挿入する形で修復し（Schegloff et al., 1977）、誤解の解消を試みている。そして、母親が生地をこね直し、カイはその様子を見ながら自身の出番を待つやり取りが続く。

断片2では、一枚目の型抜きの際には生じなかった新たな問題について、母親がその共有と解決を子どもに促し、その困難さを共有、新たな方策に移行するやり取りが行われていた。ここでは子どもと母親の両者に、問題解決に向けた調整が観察された。母親は都度、問題箇所に対する注意を引きながら、問題の指摘に用いる表現を変更することで問題の理解と解決を促し、子どもは作業を止める、やり直すといった調整を、母の発話と身体動作の観察を通して実践していた。

#### 4.3 作業の分担に関わる調整

断片2の事例では、生じた問題に対する対処として、作業のやり直し（型の再配置）および作業の手順を調整する（型抜きの前に生地をこね直す）方策が取られていた。本節ではこれらとは異なる対処として、母親とカイが作業の分担を変更するという方策が観察されたため、このやり取りについて見ていきたい。ここでは、作業の分担がどのように共有され、それに沿った作業がどのように実施されるのかという点に注目する。

以下で扱う断片3の直前では、それまで用いていた生地を再度こね、型抜きの準備がなされている。この時、今支度している生地も二枚分のクッキー型をとるには微妙なサイズであるものの、断片2で行うことになったようなこね直しをすることなく二枚の型を抜きたいことを親子で共有し合っていた。そのため断片3では、限られた生地の面積の中に二枚分の型を収める必要があり、型の配置がこれまでの型抜きよりさらに難しい状況が生じているといえる。

##### [断片3] 分担の調整1（新しい分担の適用）

（型の向きを母が指示したが、カイは異なる向きで抜こうとして止められた）

1. 母: こっち\*向き↑ね.+

- 母 >>型を生地に置く----+
- カイ \*上げていた手を型に伸ばし手のひらをかざす-->
2. (0.5) \*
- カイ ----->\*
3. 母: +\*>まって-まだ待って.+<二つ入るかちよつと確+\*認してい:い?
- 母 +カイの手を軽く払う-----+一つ目の位置に型を置く-----+二つ目の位置に型を置く-->
- カイ \*手を大きく上げて保持する-----\*手をかざそうとせずよける
4. (2.1) +
- 母 ----->+
5. 母: この辺かな?((すばやく両手を引いて腹部の前で揃える))
6. (1.4) ((カイ両手を型にのせる))
7. カイ: おっと.
8. (2.2) ((カイ型を押し込む))
9. カイ: お:.

断片3の直前で、カイは型を生地に配置するものの、これは二枚分のクッキー型を抜くことが困難な向きであった。母親はこれに気づき型を取り、1行目で型の向きを変えて生地に置く。すかさずカイはこれを押そうと型に手をかざす。ここでは二人の間に理解の齟齬が生じている。母親は1行目で、型の向きを提示するために生地に型を置いたのに対し、カイはこれを一枚目の型を抜くべき位置の指定として理解していると考えられる。

母親は3行目で発話と自身の手によってカイの型抜き動作を差し止める。これを受けてカイは手を挙手するように大きく上げて保持し、自身が待つべき局面であるという理解を示している。母親は、型を生地にあてがって、二枚のクッキー型が配置可能かどうかを試し始める。ここから、母親は生地の大さき・形が型に対して余裕のない状況であることに気づき、そのため今回の型の配置の難易度を高いものとして捉え直していると考えられる。

ここで母親は、型の配置作業を自身の分担に、型を押す作業をカイの作業として分担し直す。母親は型の配置を確認する振る舞いに続けて、現在置かれた型の位置が問題ないということ「この辺かな?」と発話しながら、型を生地に置いたまま両手を引いて腹部の前で揃える(5行目)。この手を引く素早い動作と手を揃える動作は、母親がこの後の作業に手を出さないことを強調し、この後の型を押す作業がカイの分担であることを明示する、参与者間の振る舞いの調整を手助けする動作であるといえる (coordination smoother; Vesper et al., 2010)。カイはこれを受けて、型に再び手をかざし生地に押し込み、無事に型抜きを成功させている(9行目)。

型抜きを達成する以前に、カイは二度型抜きの機会を見送っている。一度目は3行目の母の明示的な指摘による差し止めを受けた際、そして二度目は、母親が型の位置を確認する最中(3行目「確認」という発話がなされている箇所)である。後者の際、カイは母親の指先に視線を固定しており、カイは自身で型抜きの機会を見出し、そしてそれをすぐに撤回することを、母親の振る舞いの観察を通して行っていることがわかる。このように、これまでカイが行っていた型の配置を母が行い、それをカイが押す、という新たな作業の分担が、分担に関わる明示的な発話を介さずに共有されている様子が観察された。そして、ここで共有さ

れた作業の分担は、この次のクッキー型を抜く作業でも利用される。以下の断片4は、断片3で型抜きした生地を鉄板に置き終えた後のやり取りである。

〔断片4〕 分担の調整2 (同じパタンの利用)

1. 母: +たらもう:もう一個ね.  
母 +型を生地の上に配置--->
2. カイ: [ もう 一 枚. ]
3. 母: [\*も+うい-\*はい.]+\*  
カイ \*型に手をかざす-----\*型に手の平をくつつける---->  
母 ----->+手を引く-----+
4. (0.6)\* (2.2)  
カイ ----->\*型を押し込む-->
5. カイ: うん.\*  
カイ ----->\*

これは、1行目の「たらもう:もう一個ね。」の発話が、先の事例と同じパターンを想起させるものであり、かつ母親が開始している型の配置動作も先ほどのパターンに沿ったものであることが、作業分担のキューになっていると考えられる。カイは母の型の配置作業中、自身の手を型を生地に押し込む手型のまま宙に浮かせて待っており、自身の担当する作業が型を押し込むことであること、そしてこれが型の配置後すぐにやってくることへの理解が表示されている。型の配置後、母は先程のような大げさな手の引き方と姿勢をとっていないが、カイはすかさず型押しを実行、問題なくこれを達成した。

5. 考察

以上では、親子が具体的にどのようにしてクッキーの型抜き作業を進めているのかについて、4つの事例を取り上げ記述を加えてきた。子どもが行う共同行為の能力に関する先行研究は、例えば、3歳頃から他者の行為のタイミングに合わせて自身の行為を調整できるようになることを見出している (Meyer et al., 2010)。用いられた実験課題は、他者 (大人) と交互にボタンを押すという簡単な操作を求めるもので、この子どもの操作における精度が測られた。本研究が見てきた子どもは4歳1ヶ月児であり、この操作を正確に行うことができる段階に達していることが予想される。そして、この操作の内容が、親子でのクッキーの型抜き活動に含まれる諸作業になるとどうなるかについては、前節までで見てきたとおりである。何の滞りもなく型抜きを終えている事例も存在したが、より多くの場合では、何らかの問題とそれへの対処が都度求められていた。そして、親子でクッキーを一緒に作ることをそのものを楽しみ、出来上がったものを食するという活動の性質上、ここで生じている問題の多くは、大人が一方的に対処してしまうことが期待されるようなものではなく、子どもがそこに可能な限り関わり、問題を共有し、そして対処の方法を探りながら作業を進め、状況に応じてそれを修正することによって乗り越えることが期待されるものであった。このような文脈で母親は、問題が生じた際、それをカイに共有し、カイ自身が対応を行う機会を設けたり、あるいは自身が一度計画していた作業を微修正し、可能な作業をカイに都度共有することで作業を進行していた。一方のカイも、都度生じる曖昧な状況や問題に対し、指

示を受け身で待っているのではなく、母親の発話や手の動きの観察を通して自らの担う作業が何なのか、そしてそれはいつ行うべきであるのかについて判断し、試行していた。また、一度経験し成功した振る舞いを、それ以降の作業の中で援用しようと試みる様子も観察された。

実生活の中にある共同行為とは、以上のような、状況に応じた対処を子どもと大人の両者が共に探り出す、そういった機会を生むようなものが多く存在していると考えられる。そして、本研究で扱ったクッキー作りのような状況で子どもが得ている経験とは、クッキーの型抜きに限られない、より汎用的で基礎的な能力を育む経験であると考えられる。何をどのような順序で誰がやり、やらないのかを理解し試すことや、問題を共有し、必要な修正を試みること、また、進行や問題などに関わる情報を得るために、いつ何に注意をはらうかを自らの振る舞いによって表示することなど、人と関わりながら何かを成し遂げる際に必要となる様々な行為の調整が、このような具体的な実践の中で経験され、学ばれていくのだと考えられる。

## 6. おわりに

本研究では、子どもが経験している日常の共同行為場面として、親子でお菓子作りを行う際のやり取りについて記述を行った。ここでは、子どもと母親が互いに振る舞いを調整し、都度生じる問題に対処しながら作業を進行させている様子が観察された。お菓子作りは食品を扱う作業であり、成果物にある程度の水準が求められる。このことは作業に関わる二人、特に母親にとっては重要な関心事の一つであり、作業の調整もこの点を踏まえたものになっていたと考えられる。これが工作物であったら、やり取りにはどのような違いが生じるだろうか。あるいは作業に関わる参加者が異なる場合はどうだろうか。日常に生じる様々な共同作業では、それぞれの文脈に応じた調整のなされ方が採用されており、子どもたちはその文脈とともに様々な行為の調整を経験していると考えられる。CEJC-Child には今回扱ったデータ以外にも様々な場面や、年齢、協力者を収めたデータが格納される予定である。今後、これらのバリエーションを活用し、子どもたちが日常生活で経験している多用な相互行為について、記述し、比較を行っていきたい。

## 謝 辞

本研究は国立国語研究所のプロジェクト「多世代会話コーパスに基づく話し言葉の総合的研究」および科研費 20H01264, 20K20695, 22K13109, 23H00630 の研究成果を報告したものである。

## 文 献

- Brownell, C. A. (2011). Early Developments in Joint Action. *Review of Philosophy and Psychology*, 2(2), 193–211.
- 古畑尚樹・板倉昭二 (2016). 「乳幼児における We-mode の可能性.」『心理学評論』59(3), 236–252.
- Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In Lerner, G. H. (Ed.), *Conversation analysis: Studies from the first generation*, 13–31.
- Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitments to joint goals. *Child Development*, 83(1), 137–145.

- 加藤康子 (2022). 「親子クッキングからの食育活動 — 「おやつをつくろう」ワークショップ —」『八戸学院大学短期大学部研究紀要』 55, 1-10.
- 木村亜希子 (2014). 「幼児期の親子クッキング活動に及ぼす調理環境の影響: あおもり食育サポーター事務局の活動事例から.」『青森中央短期大学研究紀要』 27, 1-7.
- 小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・藤越・西川賢哉 (2023). 「『子ども版日本語日常会話コーパス』の構築.」『国立国語研究所言語資源ワークショップ発表論文集』 1, 103-108.
- Meyer, M., Bekkering, H., Paulus, M., & Hunnius, S. (2010). Joint Action Coordination in 2½- and 3-Year-Old Children. *Frontiers in Human Neuroscience*, 4, 1-7.
- Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2), 195-226.
- 小口悦子・小野かお里・小池美穂 (2015). 「クッキング玩具使用後の評価.」『東京家政学院大学紀要』 55, 75-82.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, 53(2), 361-382.
- Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2), 70-76.
- 魚住明生・池沢広美 (2016). 「子どもの発達段階に即したのものづくり教育の提案.」『三重大学教育学部研究紀要. 自然科学・人文科学・社会科学・教育科学・教育実践』, 67, 461-472.
- Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, 23(8), 998-1003.

## 付 録

書き起こしに用いた主要な記号とその意味は以下の通りである。発話は Jefferson (2004)、身体的振る舞いについては Mondada (2007) の記号法を参考に一部変更を加えた。

[ ]	括弧内の箇所が同時に発話されている	うん	相対的に大きく・強く発音されている
(あ)	聞き取り不確実	° うん °	相対的に小さく・弱く発音されている
(1.0)	間合い (秒)	>うん<	相対的に速度が速く発話されている
(.)	0.2 秒未満の間合い	<うん>	相対的に速度が遅く発話されている
:	音の引き伸ばし	(( ))	その他注記 (非言語情報、文脈情報など)
.	下降調イントネーションで発話されている	+, *	振る舞いの区切れ目
?	上昇調イントネーションで発話されている	---	振る舞いが継続する
↑	直後に記す音が顕著に高く発音されている	*-> / -->*	振る舞いが後続行の同一記号まで継続する
↓	直後に記す音が顕著に低く発音されている	*->>	振る舞いが抜粋以降も継続する
さ-	発音が不完全に途切れている	>>-*	振る舞いが抜粋以前から継続している

## Building a Kansai accent dictionary using YouTube

Hiroto Noguchi (Sophia University/Tokyo Medical and Dental University)

### YouTube を利用した関西方言アクセント辞書の作成

野口大斗（上智大学／東京医科歯科大学）<sup>†</sup>

#### Abstract

This paper presents an attempt to create an accent dictionary of the Osaka dialect using user-generated content on YouTube. Speech data were extracted from videos, transcribed, and force-aligned based on the transcribed speech. The pitch of each segment was measured, and the pitch patterns were automatically detected. This paper discusses the preliminary application of the process for a single video.

#### 1. Introduction

Language processing has undergone significant advancements, particularly in text-to-speech systems and language models. However, for languages spoken by smaller populations, linguistic resources remain underdeveloped. This disparity is evident within dialects in the Japanese language. While substantial linguistic resources exist for standard Japanese, regional dialects are often overlooked. Regarding accents, well-established accent dictionaries exist for the Tokyo dialect, such as those by NHK (1998) and open-source contributions by Tachibana and Katayama (2020).

Conversely, although accent dictionaries exist for the Osaka dialect, such as Sugito (1995), comprehensive open-source versions do not exist. Open-source accent dictionaries are valuable for linguistics and language processing. However, unless researchers have a rich network of informants and generous funding, generating extensive accent data, including thousands of words read by multiple speakers, can be a tremendous task. This paper presents a unique method for automatically generating accent dictionaries using user-generated content on YouTube.

#### 2. Previous Studies

Sugito (1995) conducted in-depth interviews with three informants from two generations to determine accent patterns. As presented in Example (1), the pitch of each word is marked with an “L” for low and “H” for high. The Osaka dialect is characterized by the fact that each word begins with a lexical high or low.

(1)		
a. kodomo	‘child’	HHH
b. i’noti	‘life’	HLL
c. kimi’ra	‘you’	HHL
d. suzume	‘sparrow’	LLL
e. hata’ke	‘field’	LHL

The development of accent dictionaries has traditionally been grounded in rigorous fieldwork, employing written documentation, auditory analysis, and direct recording. Although these approaches

---

<sup>†</sup> noguchih425@gmail.com

enable efficient acquisition of word accents essential for headwords, they have limitations. Many of these dictionaries primarily rely on lists that do not reflect natural speech. Furthermore, these methods are time-consuming, posing challenges to comprehensive and swift data collection.

### 3. Methods

#### 3.1 Data

Data from YouTube were used to eliminate the burden on informants and capture authentic speech patterns. YouTubers were chosen based on the following conditions: the YouTuber (1) must be from the Kansai region, (2) must be speaking alone, and (3) must not play background music. For this study, a video from one such qualified YouTuber was selected.

#### 3.2 Procedures

The audio was separated from the video, and the utterances were transcribed using Whisper’s (Radford et al., 2023) “large” model. For each utterance, a text file with the same name (except for the file extension) was prepared with the utterance content written in hiragana. The transcribed text was converted into hiragana using mecab-ipadic-NEologd (Toshinori, 2015) because the original version of MeCab does not correctly recognize new words like “YouTuber.” Afterward, forced alignment was performed using Julius (Kawahara, 2015). The average pitch for each segment was measured in hertz and assigned an H or L based on whether it was higher or lower than the average pitch for each word. Figure 1 presents an example of the segmentation results.

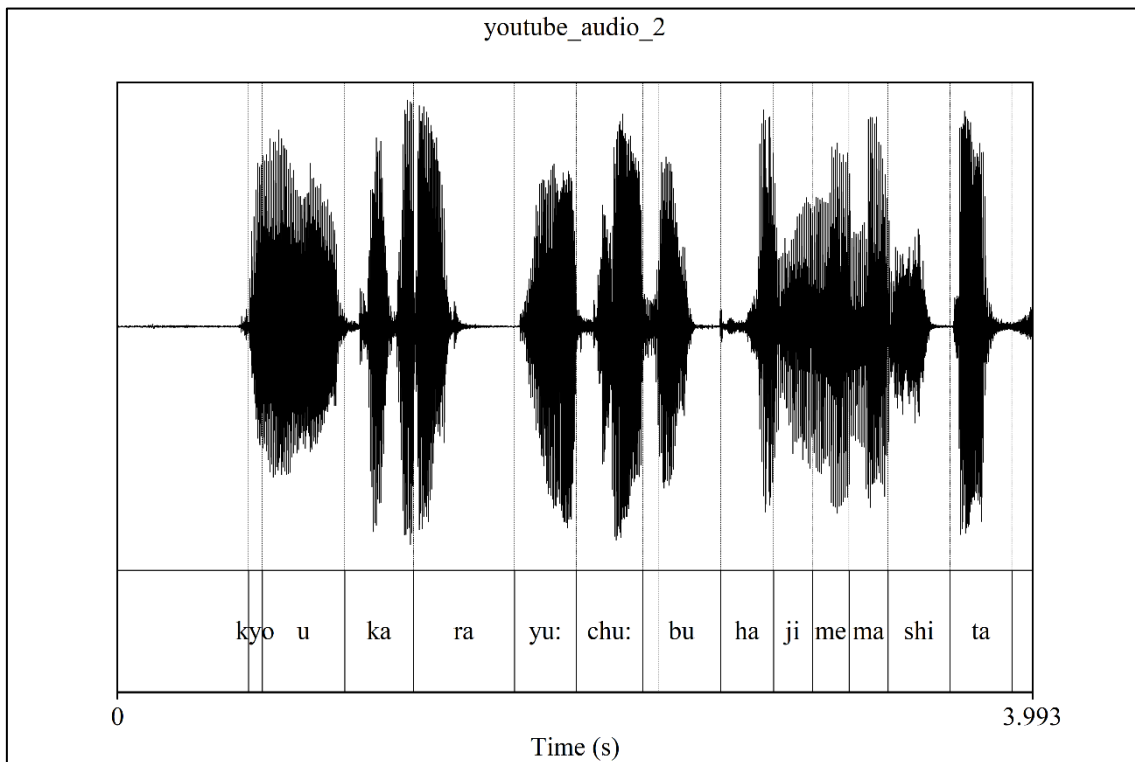


Figure 1: Segmentation example.

### 4. Results

The pitch was measured for each segment in Example (2). The average pitch for each word was compared to the segments to avoid down-step effects. Although some differences occur from the auditory impression, the pitch curve in Figure 2 can be transcribed to the symbols in Example (3). Due

to the specifications in Julius, long vowels are not divided into mora units, and pitch transcription within long vowels does not reflect the actual pitch contour.

(2) ['kyo,' 'u,' 'ka,' 'ra,' 'yu:,' 'chu:,' 'bu,' 'ha,' 'ji,' 'me,' 'ma,' 'shi,' 'ta']  
 (I started my YouTube channel today.)

(3)

a. ['kyo,' 'u']	['L,' 'H']	“today”
b. ['ka,' 'ra']	['H,' 'L']	“from”
c. ['yu:,' 'chu:,' 'bu']	['L,' 'H,' 'H']	“YouTube”
d. ['ha,' 'ji,' 'me']	['H,' 'L,' 'L']	“start”
e. ['ma,' 'shi']	['H,' 'L']	(POLITE)
f. ['ta']	['L']	(PAST)

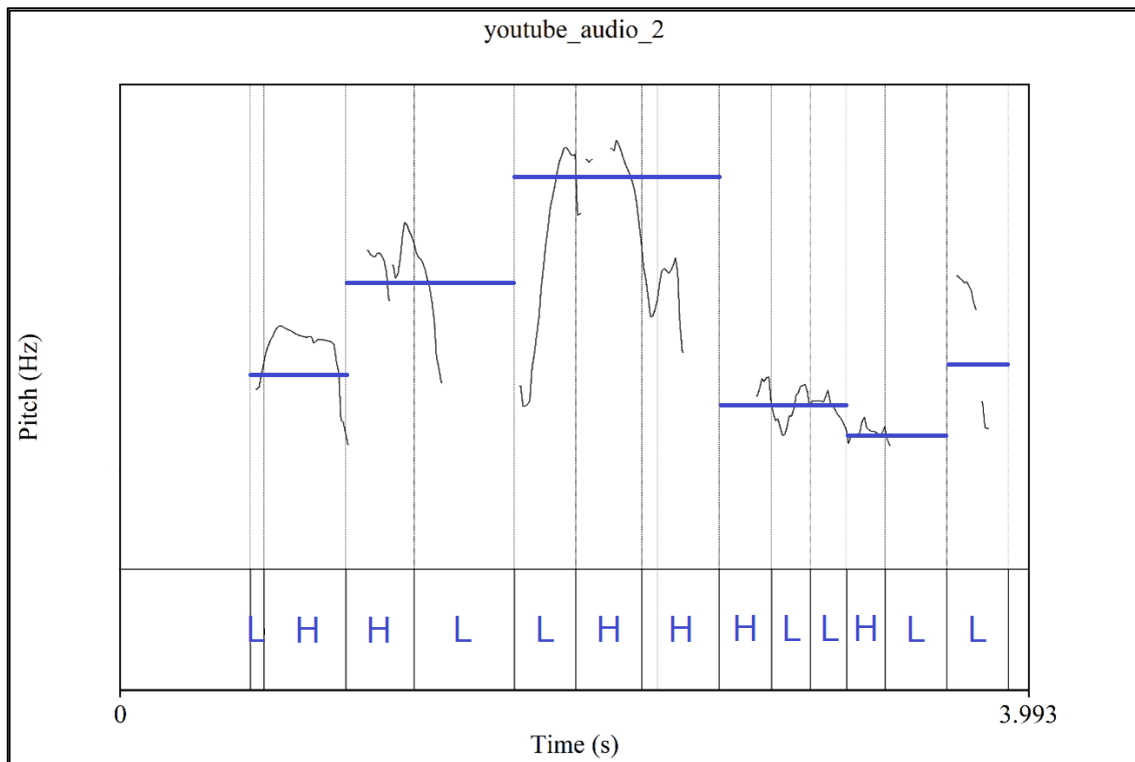


Figure 2: Pitch contour for Example (2).

## 5. Discussion

The pitch transcription method may not initially seem appropriate. However, some words appear more than once. For example, Example (4) indicates the accent pattern for the word “YouTube” was identified as in the video. The most frequently pronounced pattern, Example (4)a, which was pronounced 10 times, is also consistent with the auditory impression. Increasing the number of videos is expected to obtain more reliable accent patterns for a more extensive vocabulary. In this 16-min video alone, 474 types and 918 tokens appeared in terms of word count. Further investigation is needed in the future.



- (4)
- a. ['yu:,'chu:,'bu'] ['H,'H,'L'] \* 10
  - b. ['yu:,'chu:,'bu'] ['L,'H,'L'] \* 2
  - c. ['yu:,'chu:,'bu'] ['H,'L,'L'] \* 1
  - d. ['yu:,'chu:,'bu'] ['L,'H,'H'] \* 1

### Acknowledgments

I would like to thank the YouTube Research Program for permission to use the data in this paper.

### References

- Kawahara, T. (2005). Open-source speech recognition software Julius. *JSAL*, 20(1), 41-49.
- Kenkyūjo, N. H. B. (1998). *NHK nihongo hatsuon akusento jiten [NHK Accent Dictionary of the Japanese Language]*. Tokyo: Nihon Hōsō Shuppan Kyōkai.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518.  
<https://proceedings.mlr.press/v202/radford23a.html>
- Sugito M. (1995). *CD-ROM accent dictionary of spoken Osaka and Tokyo Japanese*. Tokyo: Maruzen.
- Tachibana, H., & Katayama, Y. (2020). Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 8059-8063.  
<https://doi.org/10.1109/ICASSP40776.2020.9054081>
- Toshinori, S. (2015). Neologism dictionary based on the language resources on the Web for Mecab.  
<https://github.com/neologd/mecab-ipadic-neologd>

## 『分類語彙表』における多義語について

山崎 誠 (国立国語研究所 (客員教授))<sup>†</sup>

### Polysemes in the "Word List by Semantic Principles"

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

#### 要旨

2004年に刊行された『分類語彙表増補改訂版』(以下、分類語彙表)はその「まえがき」によると、初版とくらべて多義語の処理を改良して、「同じ単語を意味に応じて何箇所にも出すようにした」と記述されている(P.6)。しかし、現代の小型国語辞書に掲載されている多義語と比べると、『分類語彙表』の多義語は掲出されている分類項目が少ないものがある。例えば、「切る」は『三省堂国語辞典』(第八版)では動詞の意味が27個、造語成分としての意味が3個あるが、これら30個の意味を『分類語彙表』と対照させると、単独の見出しがあるものが3個、「スイッチを切る」のように連語として見出しがあるものが6個で、計9個しか対応していなかった。残りの21個は、単独の見出しで掲出できそうなもの15個、連語として掲出できそうなもの6個であった。本発表の目的は、使用頻度の高い多義語を取り上げ、『分類語彙表』に収録されていない意味を拾い上げ、増補の候補とすることである。

#### 1. はじめに

分類語彙表には2004年の改訂の際に、多義語を構成するそれぞれの意味を増補したことが記述されている。国立国語研究所(2004: 6)に以下のように書かれている。

3) 基本語の多くは多義的である。その意味のそれぞれには、当然別の分類番号が与えられなければならない。元版においても「あがる」は、移動の意味では<2.1540 上がり下がり>に、終わる意味では<2.1502 開始・終了>に、食べる意味では<2.333 生活・衣食住>に分類されている。しかし、全体としては元版における多義語の処理は、不十分だった。今回の増補改訂版ではこの点を改良して、同じ単語を意味に応じて何箇所にも出すようにした。

しかし、多義語の持つ個々の意味に対応する見出しが見当たらないものがある。トランプを混ぜる意味の「切る」は、分類語彙表には見いだせない<sup>1</sup>。同様に、しばしば「斬る」と表記される、批判する意味の「切る」も該当する見出しが存在しない。これは上記の記述とは一致しないように思われる。そこで、分類語彙表において、多義語の持つ個々の意味がどのように収録されているかを調べてみようと思う。具体的には、小型国語辞典に掲載されている多義語の個々の意味が分類語彙表に掲載されているかどうかを確認する。とくに和語の動詞に焦点を当てる。

<sup>†</sup> yamazaki@ninjal.ac.jp

<sup>1</sup> 第17回NINJALセミナー「語彙資源の構築と活用」(2023年2月18日)で筆者が発表した際に参加者からこの指摘があった。

## 2. 分類語彙表における多義の認定方法

分類語彙表において多義語をどのように認定するかについて述べる。多義語は、「意味的に関連の認められる異なった二つ以上の意味をもつ語。」(『日本大百科全書(ニッポニカ)』の「多義語」の項目、執筆・国広哲弥)と定義される。分類語彙表では、便宜的に各分類項目が多義語の条件である「意味的に関連の認められる異なった二つ以上の意味」に対応しているとみなすことができるだろう。

次に、多義を持つ「語」をどのように特定するかであるが、こちらも便宜的に分類語彙表において、表記と読みの組み合わせで決めるのが妥当であろう。すなわち、任意の2語について、表記と読みの両方が一致していれば同じ語と見なし、表記あるは読みのいずれか、あるいは、その両方が異なれば別の語と見なすわけである。もちろん、この方法で語を特定することはやや正確さを欠く<sup>2</sup>。便宜的ではあるが、上記の方法で「分類語彙表増補改訂版データベース」<sup>3</sup>における各語の意味数を調べたものが表1である<sup>4</sup>。

表1 分類語彙表における語の意味数と語数(分類項目による集計)

意味数	語数	割合
1	67598	83.20
2	11228	13.82
3	1839	2.26
4	439	0.54
5	82	0.10
6	36	0.04
7	18	0.02
8	5	0.01
10	1	0.00
11	1	0.00
合計	81247	100.00

表1は、分類項目を意味の単位としたものであるが、例えば、1.3374(スポーツ)という分類項目に、「シュート」という見出しが2回現れる。これらは、サッカーのシュートと野球の球種であるシュートに対応する。これらは分類項目の直下の階層である段落でみると、それぞれ別の段落に現れる。そこで、これらを別の意味領域に属すると見なすことができるように、分類項目と段落番号を使って各語の意味数を集計したものが表2である。

以降、表2に基づいて話を進める。表2には、意味1の語が全体の83.16%であることが分かる。意味1の語は単義語と見なすことができるので、残りの16.74%が多義語ということになる。ちなみに、意味数が12の語は「立てる」、意味数11は「手」であった。

<sup>2</sup> 例えば、花の「リラ」(lilas)と通貨単位の「リラ」(lira)が同じ語になってしまう。また、「立てる」「建てる」のように異なる漢字をあてている場合は別語となる(ただし、このケースを同音異義語とするか、多義語をするかは議論の余地がある)。

<sup>3</sup> <https://github.com/masayu-a/WLSP/blob/master/bunruidb.txt>

<sup>4</sup> 上記データベースから、以下の2種のレコードを除外している。(1)区切りを表すレコード(見出しが「\*」)。(2)レコード種別が「B」となっているもの。これは、見出しが「瑣事・些事」のように複数の語を便宜的に合併して示したもの、あるいは、「くんだり(件)」のように、括弧で漢字を添えた項目等である。これらは、別に表記と読みの組み合わせのレコードが用意されているので、「B」は集計外とした。

表2 分類語彙表における語の意味数と語数（分類項目＋段落番号による集計）

意味数	語数	割合
1	67568	83.16
2	11230	13.82
3	1853	2.28
4	448	0.55
5	85	0.10
6	38	0.05
7	17	0.02
8	4	0.00
9	2	0.00
11	1	0.00
12	1	0.00
合計	81247	100.00

### 3. 多義語「切る」の掲載状況

この節では、多義語の例として『三省堂国語辞典』（第八版）の「切る」を取り上げ、その各意味が分類語彙表にどのように掲載されているか、あるいは、ないかを調査した。結果を表3に示す。「切る」には、品詞の違いで、動詞と造語成分（接尾辞）とがあるが、両方を対象とした。表の「分類語彙表での有無」は、以下のとおり。

- ：対応する意味が掲載されているもの
- △：対応する意味が連語として掲載されているもの（「縁を切る」「スイッチを切る」等）
- ×：対応する意味が掲載されていないもの

「分類番号と項目名」は、対応する分類項目がある場合は、その情報を、ない場合は、もし掲載されるとしたらその分類番号になるはずの情報を記した。

表3 『三省堂国語辞典』（第八版）の「切る」の意味と分類語彙表における位置

「切る」の意味（語釈と例文）	分類語彙表での有無	分類番号と項目名
①刃物(はもの)などで、一つのものをはなればなれにする。断(た)つ。「枝を一・糸を一」	○	2.1571 切断
②刃物などで きずつける。「指を一」	○	2.1571 切断
③切ってあける。「封(ふう)を一」	×	2.1553 開閉・封
④金属に刻みつける。「銘(めい)を一」	×	2.1570 成形・変形
⑤切って内がわに作る。「ポケットを一・炉(ろ)を一」	×	2.1570 成形・変形
⑥〔切りはなして〕作る。「伝票を一・手形を一」	×	2.3770 授受
⑦配布する印刷物を作る。「レジユメを一」	×	2.3832 出版・放送
⑧〔空間に〕まっすぐに線を引く。「空(くう)を一・十字を一」	×	2.1510 動き
⑨〔よけいな水分などを〕取り去る。「水を一・揚(あ)げものの油を一」	×	2.1251 除去
⑩〔関係を〕きっぱりなくす。「縁(えん)を一」	△	2.3500 交わり [縁を切る]

⑪切り捨てる。「下請(う)けを一・苦手な授業を一 [=もう出席しない]」	×	2.1251 除去
⑫とちゅうでやめる。「話を一」	×	2.1503 終了・中止・停止
⑬〔スイッチを動かして〕電流をとめる。	△	2.3850 技術・設備・修理
⑭〔カメラのシャッターボタンを〕おす。	×	2.3850 技術・設備・修理
⑮顔を動かして、その方角に向く。また、左右に回す。「正面を一・ハンドルを一」	×	2.1730 方向・方角
⑯ある かまえや態度をとる。「みえを一・鯉口(こいぐち)を一 [⇒「鯉口」の用例]・しらを一」	△	2.3030 表情・態度
⑰〔斬る〕刀で、きる。殺す。「寄らば一ぞ・泣いて馬謖(ばしょく)を一 [⇒「泣く」の【句】]」	×	2.3440 犯罪・罪
⑱〔斬る〕えんりよなく、批判する。「現代の世相を一」	×	2.3135 批評・弁解
⑲そのときまでで、それ以後は受けつけないことにする。「百番で一・期限を一」	×	2.1721 境・間
⑳ある数より下になる。「残り二〇秒を一」	×	2.1584 限定・優劣
㉑(中を分けて)勢いよく〈進む／進み出る〉。「先頭を一・スタートを一」	△	2.1525 連れ・導き・追い・逃げなど
㉒思い切って手放す。「飛車を一・札ビラを一」	△	2.3700 取得
㉓ふだをさし出す。「カードを一 [=切りふだを出す]・名刺(めいし)を一」	×(2.3770) △(2.3710)	2.3770 授受 2.3710 経済・収支
㉔そろえたままの ふだをまぜあわせる。「カルタを一 [=よくまぜ(て、くば)る]・カードを一」	×	2.1550 合体・出会い・集合など
㉕水面に当たって はねかえる。「投げた石が水を切つてとぶ」	×	2.1525 連れ・導き・追い・逃げなど
㉖<<すもう>>まわしにかかった相手の手をはなさせる。	×	2.3374 スポーツ

以下、造語成分として意味

①終わりまで…する。そこで終わりにする。「泳ぎ一・逃(に)げ一・使い一・言い一」	○	2.1250 消滅
②これ以上ないほど…する。「澄(す)み一・困り一・にがり一・弱り一・勝ち一 [=完勝する]」	×	2.1721 境・間
③ずっと…する。「かかり一・詰(つ)め一」	×	2.1504 連続・反復

「切る」は動詞の意味が 27 個<sup>5</sup>、造語成分（接尾辞）としての意味が 3 個、合計 30 個の意味があるが、これらのうち、単独の見出しがあるものは 3 個、連語として見出しがあるものが 6 個であった。「切る」の多義語充足率<sup>6</sup>は 30.0%であり、3 割という低い状況である。残りの 21 個は分類語彙表には掲載されていなかった。その 21 個のうち、単独の見出しで掲出できそうなもの 15 個、連語として見出しが妥当であるものが 6 個認められた。

<sup>5</sup> ㉓に 2 つの意味があるので、ブランチ数の合計とは合わない。

<sup>6</sup> ○と△の合計を全体数（○×△の合計）で割った値。㉓が 2 つの分類項目に対応するため、母数をブランチ数とはしなかった。

4. 使用した辞書との関係

前節では多義語「切る」について分類語彙表と『三省堂国語辞典』（第八版）との対照を試みたが、実際に分類語彙表を増補した際に参照した辞書は『三省堂国語辞典』ではない。分類語彙表における多義語の増補については中野ほか（1989: 450）に言及がある。以下、該当部分を引用する<sup>7</sup>。

(5) 一語に複数の分類番号をあてる方法

この方法は、次の手順で行う予定である。

- ①『新明解国語辞典』（MT版）の意味記述文から各意味毎にレコードを作る。
- ②①から多義語を抜き出す。
- ③②に、その意味記述によって、新しい分類番号を付ける。

しかし、現段階での増補については、積極的にはこの方法をとってはいない。機械的な方法で付けられた複数の分類番号を、しいて1つにするということはない程度である。

この記述によれば、多義語の増補はある程度まで行われたということが推察される。また、上記引用中の『新明解国語辞典』は中野（1989: 449）に約70,000語とあることから、第二版であると思われる。そこで本節では『新明解国語辞典』（第二版）との比較を試みた。『新明解国語辞典』では、意味の区分が大きくまとめられ、その中に用例として個別の意味が記述されている形式になっているため、各用例と分類語彙表との対応関係を調べた。

表4 『新明解国語辞典』（第二版）の「切る」の意味と分類語彙表における位置

「切る」の意味（語釈と例文）	分類語彙表での有無 <sup>8</sup>	分類番号と項目名
一①（他五）㊦鋭い刃物などで、勢いよく二つの部分に離すようにする。「人を一〔＝(A)傷つける。(B)殺す〕・たまを一〔＝庭球・卓球などで、たまに逆回転を与える打ち方をする〕・口を一〔＝(A)罐や封筒などの端を切って、出し入れ口を作る(B)沈黙を破って、物を言い出す〕・行列を一〔＝横切る〕・伝票を一〔＝発行する〕・切符を一〔＝切符にパンチを入れたりして、正式に乗客・入場者と認める〕」	○(2.1571) ×(2.3374) ×(2.1570) ×(2.3100) ×(2.1525) ×(2.3770) ○(2.1571)	2.1571 切断 2.3374 スポーツ 2.1570 成形・変形 2.3100 言語活動 2.1525 連れ・導き・追い・逃げなど 2.3770 授受 2.1571 切断
㊦手やからだからだなどで空気を「切る」 ㊦のような動作をする。「空気を切って〔＝空中を突き抜けるようにして〕飛んでくる・十字を一〔＝十字の形を手ではっきりと描く〕・肩で風を一〔＝勢いよく進む〕・先頭を一〔＝先頭に立って進む〕」	○(2.1525) ×(2.1520) ○(2.3030) ○(2.1525)	2.1525 連れ・導き・追い・逃げなど 2.1510 動き 2.3030 表情・態度 2.1525 連れ・導き・追い・逃げなど

<sup>7</sup> 同一の記述が中野（1989: 4）にも見られる。

<sup>8</sup> (A)(B)が同じ分類項目に対応する場合は区別していないが、異なる分類項目に対応する場合は、それぞれを記した。

		ど
㊦局面が転換するような、思い切った事を何かする。「たんかを―・しらを―みえを―・切り札を―〔＝出す〕・札びらを―〔＝見せびらかすように、人前に出・す(して使う)〕・ハンドルを―〔＝鋭く動かして、方向を変える〕・カーブを―」	○(2.3100) △(2.3030) △(2.3100) △(2.3770) ×(2.1730) ×(2.1730)	2.3100 言語活動 2.3030 表情・態度 2.3100 言語活動 2.3770 授受 2.1730 方向・方角 2.1730 方向・方角
㊧一続きのものの連絡を絶つ。「言葉を―〔＝そこで話をやめる〕・スイッチを―〔＝操作して、回路を絶つ〕・先着順十名で―〔＝申込者や入場者を限定する〕・日限を―〔＝一定の範囲に限る〕・水を―〔＝水分が無くなるようにする〕・かるたを―〔＝相互に連絡が無いように、うまく交ぜる〕」	×(2.1503) △(2.3850) ×(2.1721) ×(2.1721) ×(2.1251) ×(2.1550)	2.1503 終了・中止・停止 2.3850 技術・設備・修理 2.1721 境・間 2.1721 境・間 2.1251 除去 2.1550 合体・出会い・集合など
㊨そこまで到達しない範囲内にとどまる。割る。「百メートル競走に十秒を―・元を切って売る」	×(2.1584) ×(2.1584)	2.1584 限定・優劣 2.1584 限定・優劣
二〔接尾語的に〕㊩余す所無く何かをする。「読み―(0)・言い―」	○(2.1250) ○(2.1250)	2.1250 消滅 2.1250 消滅
㊪限界に達して、それ以上堪えられない状態になる。とことんまで…する。「弱り―(4)(2)・困り―」	×(2.1721) ×(2.1721)	2.1721 境・間 2.1721 境・間

表 4 から、分語彙表への掲載状況としては、○が 8 個、△が 4 個、×が 17 個となった。多義語充足率は 41.4%と、表 3 よりは大きい値となった。しかし、大きな区分で見ると、動詞の㊨と接尾語の㊩を除いては掲載されていたことになる。増補の際、各意味のレコードがどのような構造になっていたか分からないが、表 4 のような大きな区分であった場合、その中にどれを選んで分類番号を付けたのかが分からないので、作業的にはあまり好ましくない方法だったのかもしれない。

## 5. おわりに

本稿では 2004 年に刊行された分類語彙表増補改訂版の多義語の増補の状況を「切る」を例として観察した。その結果、多くの意味が掲載されていないことが分かった。これは作業上どのようなデータを使用したかと関わっているのではないかと推察された。他にも使用頻度が高く多義である語の中には同じように増補から漏れてしまった意味があるかもしれない。今後分類語彙表の改訂の際の参考としたい。

## 謝 辞

本研究は JSPS 科研費 JP19K00655 の助成を受けている。

## 文 献

国立国語研究所(編)(2004).『分類語彙表増補改訂版』大日本図書

中野洋(1989)『『分類語彙表』形式による語彙分類表』文部省科学研究費補助金特定研究(1)  
「言語データの収集と処理の研究」(研究代表者：野村雅昭、研究分担者：中野洋)  
中野洋、宮島達夫、石井久雄、藤田正春、鶴<sup>9</sup>岡昭夫、森由紀(1988)「日英語彙データの収  
集・比較と機械辞書の作成」

#### 関連 URL

日本大百科全書 (ニッポニカ) , JapanKnowledge, <https://japanknowledge.com> , (2023 年 8 月  
15 日参照)

---

<sup>9</sup> 原文では「鶴」。



# 教科学習における抽象的思考と結びつく言語形式 — 数学における「とする」をケーススタディとして —

新山 聖也 (筑波大学)

竹本 理美 (筑波大学)

澤田 浩子 (筑波大学)

## Linguistic Forms Linked to Abstract Thinking in Subject Learning : Case study of "To Suru" in Mathematics.

Seiya Niiyama (University of Tsukuba)

Satomi Takemoto (University of Tsukuba)

Hiroko Sawada (University of Tsukuba)

### 要旨

近年、日本語指導を必要とする外国人児童生徒が増加しており、教科学習に必要な学習言語能力の支援が問題となっている。本稿では、教科学習の中で求められる抽象的思考と結びつく言語形式を分析することを目的とし、中学校教科書のテキストを対象として形態素解析を行った。

まず、数学教科書と理科教科書の比較から、共通して出現しやすい表現と特定の教科に出現しやすい表現が存在することを指摘し、ケーススタディとして数学に特徴的な文型として「AをBとする」に注目して分析を行う。「とする」の前後文脈、出現しやすい単元に関して分析を行い、数学では「とする」が具体的事象における要素と数式に出現する要素を同定し、立式の際に思考の枠組みを設定する用法で用いられることを指摘する。

本稿の分析は、語彙だけでなく、特定の文型が教科学習で求められる抽象的な思考と結びつくことを示す事例として位置付けられる。

### 1. 研究の背景

#### 1.1 教科学習と抽象的思考

日本の公立学校において日本語指導が必要な児童生徒の数は日本国籍、外国籍を問わず増加しており、2010年度は34,007人だったのに対し、2021年度は58,307人となっており、約10年で1.7倍程度増加している(文部科学省 2022)<sup>1</sup>。このような、文化的・言語的に多様な背景を持つ子ども(Culturally Linguistically Diverse Children、以下CLD児)は、日常会話において高い日本語能力を持っているように見えても、教科学習に関わる日本語能力に課題を持っていることがある(cf.西川・青木 2020)。この言語能力の違いは、Cummins (1979)においてBICS(Basic Interpersonal Communication Skills)とCALP(Cognitive / Academic Language Proficiency)の違いとして区別されるものと対応している。

教科学習に関わる日本語能力を対象とする分析の一例として、竹本・新山・澤田(2023)は、CLD児の受動文理解を対象として研究を行っている。竹本・新山・澤田(2023)は、CLD児を対象に文理解テストを実施し、CLD児が受動文の理解を苦手とすること、その中

---

<sup>1</sup> 内訳は、2010年度は外国籍28,511人と日本国籍5,496人であり、2021年度は外国籍47,619人と日本国籍10,688人である。

でも特に物理的接触を伴わない動詞（例：「探す」「囲む」）を述語とする受動文の理解に困難を抱えていることを報告している。さらに、現代日本語書き言葉均衡コーパス(BCCWJ)の「教科書サブコーパス」による小・中学校の教科書分析から、学年があがるにつれて「含まれる」や「囲まれる」のような二者間の抽象的な関係を表す受動文が用いられやすい傾向にあることを指摘している。このように、教科学習においては、一種の抽象性に関わる日本語表現を理解することが求められ、CLD 児にとって困難となり得ると考えられる。

本稿では、さらに教科書のテキスト分析を進めることで、教科学習で求められる抽象的思考と結びつく言語形式の抽出を行う。抽象的思考と結びつく言語形式は、特定の環境で用いられやすいものと考えられ、日常会話を通じて学習することがより困難であると考えられるためである。

## 1.2 教科学習と文型の関係

教科学習と日本語の関係を取り扱った研究として、宮部 (2019, 2021) が挙げられる。宮部は、教科書に出現する言語形式の分析から、日本語学習者が教科学習を行う際に困難になり得る言語形式をまとめている。中でも、宮部 (2019) では、数学教科書に出現する助詞トに着目し、トが多様な用法を持つことやトと動詞の組み合わせが派生的な意味や機能的な働きを持つことなどを指摘している。

興味深い点として、「と言う」「とみる」「とする」「となる」のようなトと組み合わせられる動詞が、日常会話や初期の日本語学習における「言う」「見る」「する」「なる」とは異なる意味を持っているという点が挙げられる<sup>2</sup>。つまり、日常会話において高い日本語能力を持っているように見える CLD 児にとっても、「と言う」「とみる」「とする」「となる」のようなトと動詞の組み合わせは、教科学習でしか出会わない表現である可能性がある。この事実は、教科学習と日本語の関係について分析する際には、語彙レベルの言語形式に着目するだけでなく、文型レベルの言語形式に着目する必要があることを示唆している。

宮部 (2019) の問題点としては、分析対象が数学教科書に限定されているため、言語形式と教科学習の対応関係がわからない点が挙げられる。例えば、「と言う」は「意味を持つ最小の単位を形態素と言う」のように用語や概念を導入する際に普遍的に用いられる。このような「と言う」は、「ものごとを教える」という場面において普遍的に用いられ、「急に降る雨は、にわか雨って言うんだよ」のような日常会話の日本語とも連続的な側面を持つことが予想される。よって、本稿では、教科書に頻出する言語形式の中でも、特定の教科に頻出する言語形式に着目する。

## 1.3 本稿の目的

以上の背景に基づいて、本稿は、抽象的思考と結びつく言語形式を抽出し、その言語形式と教科学習の結びつきを明らかにすることを目的とする。この目的を達成するために、本稿では、以下の2点について調査を行う。

- (1) 抽象的思考と結びつく言語形式を取り出すために、教科書に頻出する言語形式を調査する。中でも、特定の教科に頻出する形式に着目する。

<sup>2</sup> 「と言う」は教科書の中ではひらがなで「という」と表記されるが、「数学という教科」のような場合に用いられる「という」と区別するため、本稿では「と言う」と表記する。

- (2) 教科書に頻出する言語形式の中でも、特定の教科に頻出する言語形式について、言語形式と教科の関係について分析を行う。本稿では、ケーススタディとして「AをBとする」という文型に着目し、「AをBとする」という文型と数学の関係について分析する。

まず、(1) は、抽象的思考と結びつく言語形式を取り出すために行う調査である。前述の通り、教科に偏りなく頻出する形式は、「ものごとを教える」という場面において普遍的に用いられ、日常会話の日本語とも連続的な側面を持つと予想される。一方、特定の教科に頻出する形式はその教科の学習で求められる抽象的思考と結びついていることが予想される。よって、本稿では、特定の教科に頻出する形式に着目することで、教科学習で求められる抽象的な思考と結びつく言語形式を取り出すことを図る。

続いて、(2) は、言語形式と教科学習の結びつきを明らかにするために行う分析である。「AをBとする」という文型が、数学という教科とどのように結びついているのかを明らかにすることで、教科学習における抽象的思考と結びつく言語形式についてのケーススタディを提示する。

## 2. 研究方法

### 2.1 調査対象

本稿では、教科学習の中でも、中学校の理数系科目である数学と理科を対象として調査を行った。中学校の数学は「抽象的な概念や関係を理解し、思考し、また、それを説明するという言語的要素もともなう教科 (宮部 2019: 118)」であり、教科学習における抽象的思考と言語形式の関係を分析する上で重要な調査対象となる。また、教科間の比較をする上で、本稿では理数系教科書に絞って調査を行った。具体的な調査対象は、(3) と (4) に記載する6冊の教科書である<sup>3</sup>。

- (3) 中学校数学教科書：『新しい数学 1』『新しい数学 2』『新しい数学 3』（東京書籍）  
(4) 中学校理科教科書：『新しい科学 1』『新しい科学 2』『新しい科学 3』（東京書籍）

本稿では、教科間の比較を行うために、同一の出版社から出版された教科書に統一して調査を行った。

### 2.2 調査方法

本稿では、上記 (3) と (4) の教科書について文字情報をすべてテキストデータ化し、キストデータについて中・長単位解析器 Comainu (小澤ほか 2014) を用いて長単位解析を行った<sup>4</sup>。長単位解析の結果、本稿で調査を行った数学教科書は合計で 121,398 語、理科教科書は合計で 280,523 語の長単位に分割された<sup>5</sup>。

<sup>3</sup> 調査対象となる教科書の詳細な書誌情報は、調査資料として論文の末尾に掲載している。

<sup>4</sup> 本稿で長単位解析のデータを用いた理由としては、「とする」と複合助詞「として」や「と言う」と脚注 2 で言及した「という」を区別して取り扱うためである。

<sup>5</sup> ここでは、長単位を数える単位として語を採用している。本稿では、長単位解析における長単位について、言語学における形態素（意味を持つ最小の単位）というより語に近いものと見なし、語という表現を用いている。

また、解析によって得られた長単位データについて、N-gram 分析を行った。N-gram とは、N 個の単位の連続のことを指しており、本稿で言えば、N 個の長単位の連続を指す。N-gram 分析では、2gram であれば「次/の」、3gram であれば「次/の/式」、4gram であれば「次/の/式/を」のような言語単位の連続を対象として、単位同士の組み合わせのパターンやその出現回数について分析を行うことが可能となる。

以上の通り、本稿では、理数系教科書を対象とし、長単位解析と N-gram 分析によってデータの分析を行った。このデータに基づいて、3 節で理数系教科書の比較を行い、4 節で数学教科書における「とする」の分析を行う。

### 3. 理数系教科書の比較

3 節では、抽象的思考と結びつく言語形式を取り出すために、教科書に頻出する言語形式を調査する。1 節で述べた通り、教科書に頻出する言語形式は、教科一般に頻出する形式と特定の教科に頻出する形式に分類できると考えられ、本稿では、後者の言語形式に着目する。これは、特定の教科に頻出する形式は、その教科の学習で求められる抽象的思考と結びついていることが予想されるためである。

本稿では、前述の通り、N-gram 分析によって、長単位の組み合わせの出現回数を調査した。これは、1.2 節で見た通り、語彙レベルの言語形式だけでなく、文型レベルの言語形式についても調査を行うためである。この調査の結果を表 1 に示す。表 1 は、中学校の理数系教科書に長単位 2-gram として出現する形式のデータを、20 位まで示している。この際、句点や読点を含むものは除外している。

表 1 では、数学教科書と理科教科書について、それぞれ 4 列で情報を示している。1 列目と 2 列目には、2-gram として出現する語彙素の連続のデータとその出現回数を記載し、3 列目と 4 列目には、語彙素が実際にどのような形であられるか、出現回数が 2 桁以上の回数になる書字形の例を、2 種類まで掲載している。語彙素「様だ」であれば、実際には書字形「ように」や書字形「ような」として出現することが確認できる<sup>6</sup>。

表 1 を見ると、数学では、「値」「数」「長さ」「式」「図」のような数学に関わる名詞語彙だけでなく、「のような」「次の」「右の」のような指示表現、「でしょうか」「求めなさい」や「考えてみましょう」のような読み手に働きかける表現が頻繁に出現していることがわかる。理科では、「調べる」「作る」「使う」のような理科の観察・実験などに関わる動词语彙の他にも、「のような」のような指示表現、「どのような」「だろうか」のような疑問表現が頻繁に出現していることがわかる。

また、理科教科書では、場所や状況に関わる「では」「には」が上位となっている。これは「がある」「れている」のように存在や状態を述べる表現、「の中」のような位置を示す表現の出現回数が多いこととも対応した結果と思われる。

注目すべき点は、数学における「とする」（12 位）と理科における「と言う」（9 位）である。「とする」と「と言う」はいずれも、宮部（2019）において数学教科書に出現し、派生的な意味や機能的な働きを持つとされていたトと動詞の組み合わせであるが、「と言う」は理科教科書においても出現頻度順位が高いことがわかる。

<sup>6</sup> 「では」や「には」のように、語彙素よりも書字形の出現回数が多い例もある。これは、語彙素の判定としては別の語彙素として判定されたが、書字形としては「には」や「では」として出現している例が存在することを示しており、数値が誤っているわけではない。

表1 数学教科書と理科教科書における 2-gram の組み合わせの出現回数

	数学教科書				理科教科書			
	語彙素	出現回数	書字形内訳	出現回数	語彙素	出現回数	書字形内訳	出現回数
1	様だ	934	ように	678	様だ	2,059	ように	1,277
			ような	256			ような	780
2	次の	713	次の	713	には	958	には	966
3	てみます	551	てみましょう	546	何の様	795	どのよう	795
4	ですか	523	でしょうか	386	が有る	746	がある	564
			ですか	137			があり	137
5	の図	474	の図	474	では	662	では	699
6	の様	452	のよう	452	の様	606	のよう	606
7	考えてみる	406	考えてみよう	332	に成る	517	になる	401
			考えてみ	74			になっ	260
8	の値	384	の値	384	れている	472	れている	302
							れてい	170
9	の数	378	の数	378	と言う	471	という	394
							といい	61
10	に成る	350	になる	322	の中	455	の中	364
			になり	53			のなか	91
11	右の	346	右の	346	だか	452	だろうか	452
12	とする	307	とする	229	の大きさ	384	の大きさ	384
			とし	78				
13	を使う	299	を使っ	256	た時	375	たとき	362
			を使う	22				
14	の長さ	298	の長さ	298	を調べる	347	を調べる	218
							を調べ	93
15	事を	289	ことを	289	を作る	346	をつくる	210
							をつくっ	55
16	を求め為さる	278	を求めなさい	278	時の	345	ときの	345
17	では	267	では	279	を使う	344	を使っ	236
							を使う	71
18	には	248	には	249	事を	327	ことを	327
19	数の	246	数の	246	ています	323	ています	292
							ていまし	18
20	の式	239	の式	239	自分の	320	自分の	320

そこで、本稿では、トと動詞の組み合わせを対象を絞り、数学教科書と理科教科書における出現回数を比較した。表2に、数学教科書と理科教科書におけるトと動詞の組み合わせの出現回数を示す。この際、実際の出現回数だけでなく、カッコ内に10万語あたりの出現回数も示した。2.2節で見た通り、理科教科書と数学教科書はデータサイズが異なるので、10万語あたりの出現回数を見ることで、教科書の分量の差を考慮に入れた上で比較を行うことができる<sup>7</sup>。表2からは、「と言う」については理科教科書でも数学教科書でも出現回数が多いのに対し、「とする」については一定の差があることが窺える。

<sup>7</sup> ここで10万語としている数は、長単位の数である。長単位を数える単位として語を使っていることについては、脚注5で説明している。

表2 数学教科書と理科教科書における「ト+動詞」

数学教科書		理科教科書	
ト+動詞	出現回数 (10万語あたりの回数)	ト+動詞	出現回数 (10万語あたりの回数)
とする	307 (252.9)	と言う	471 (168.0)
と言う	213 (175.5)	と比べる	237 (84.5)
となる	152 (125.2)	と考える	200 (71.3)
と考える	57 (47.0)	となる	183 (65.2)
と表わす	43 (35.4)	とする	173 (61.7)
と言える	40 (32.9)	と呼ぶ	93 (33.2)
と書く	38 (31.3)	と言える	54 (19.2)
と見る	28 (23.1)	と結び付く	35 (12.5)
と思う	20 (16.5)	と思う	34 (12.1)
と比べる	17 (14.0)	と異なる	14 (5.0)

このような対立は、教科書に出現する「と言う」の実例からも伺える。(5) はそれぞれ数学と理科の教科書における「と言う」の実例であるが、いずれも「原点」や「受粉」という用語を導入するために用いられている。

- (5) a. 数直線上で0が対応している点を原点という。  
(『新しい数学1』p.23 下線は筆者による)
- b. めしべに花粉がつくことを受粉という。  
(『新しい科学1』p.12 下線は筆者による)

このような「と言う」による用語の導入は、特に数学という教科や理科という教科内容に対応しているわけではなく、「ものごとを教える」という場面において普遍的に用いられることが予想される。実際、日常会話を収録したコーパスである『日本語日常会話コーパス』(CEJC)にも、(6)のような例がみられる。特に、(6b)は、身近ではない単語を教えるという意味で、教科書における用語の導入に近い用法だと考えられる。

- (6) a. (さつま揚げの呼び方について) 香川では天ぷらってゆうんだよねこれ  
(CEJC カッコ内の補足と下線は筆者による)
- b. 今はだから塾歴社会ってゆうんだって (CEJC 下線は筆者による)

つまり、「と言う」は、(5)のような例において確かに本動詞「言う」と異なる意味を持っているものの、それは日常会話でも使われる用法と連続的であり、教科学習と強く結びつく表現ではないものと考えられる。

このような、特定の教科と結びつかない「と言う」に対して、表2の結果を踏まえると、「とする」は数学という教科と結びつく言語形式と見なし得る可能性がある。これに関して、本稿では、数学教科書と理科教科書における「とする」の前接形式について調査を行うことで、「とする」がどのような文型で用いられているかについて確認した。この結果を表3に

示す<sup>8</sup>。表3では、表2と同様にカッコ内に10万語あたりの出現回数を示している。

表3 理数系教科書における「とする」の前文脈<sup>9</sup>

前接形式のパターン	数学教科書	理科教科書
	出現回数 (10万語あたりの回数)	出現回数 (10万語あたりの回数)
N とする	276 (227.4)	118 (42.1)
た とする	9 ( 7.4)	11 ( 3.9)
V とする	8 ( 6.6)	3 ( 1.1)
V よう とする	6 ( 4.9)	28 (10.0)
A とする	5 ( 4.1)	3 ( 1.1)
である (だ) とする	3 ( 2.5)	4 ( 1.4)
[オノマトペ] とする	0 ( 0.0)	6 ( 2.1)

表3は、数学に出現する「とする」が、「N(名詞)とする」という一定の形で頻繁に用いられること示している。10万語あたりの出現回数で考えると、数学における「Nとする」は、理科における「Nとする」の5倍程度出現していることになる。「Nとする」は、主に(7)の「AをBとする」、あるいは(8)の「A=Bとする」のような文型で用いられる。

- (7) 3つの続いた整数のうち、もっとも小さい整数をnとすると、3つの続いた整数はn, n+1, n+2と表される。

(『新しい数学2』p.22 下線は筆者による)

- (8) 円Oの周上の点をA, B, Cとし、 $\angle ACB = \angle a$ とする。

(『新しい数学3』p.174 下線は筆者による)

一方、理科においては「流そうとする」のような「Vようとする」の用例も一定数出現する。これは日常会話にも頻出するタイプの「とする」であると考えられ、日常会話に頻出しない「AをBとする」とは異なった性質を持っている<sup>10</sup>。このように、数学においては「AをBとする」という文型が頻出するのに対し、理科においては必ずしもそうではないことが確認できる。

3節では、抽象的思考と結びつく言語形式を取り出すために、教科書に頻出する言語形式の調査を行った。表1のデータは、様々な言語形式と教科の関係を示唆しているが、本稿では「とする」と「と言う」の対立に着目し、数学と理科に共通して頻出する「と言う」に対して、数学のみに頻出する「AをBとする」という文型の存在を示した。

<sup>8</sup> 表4におけるN, V, Aはそれぞれ名詞、動詞、形容詞を指している。

<sup>9</sup> 10万語あたりの出現回数を算出する際には、小数点第2位以下を四捨五入する形で数値を算出している。この処理は、表5においても同様である。

<sup>10</sup> 日常会話コーパス(CEJC)における「とする」の用例525例のうち、233例が「Vようとする」の例である。また、123例が「とろっとする」のような「オノマトペとする」の用例であり、「AをBとする」は28例程度である。「AをBとする」の内訳としては、「必要とする(8例)」や「よしとする(4例)」のような定型的な表現が多くみられる。

#### 4. 数学における「とする」の分析

##### 4.1 思考の枠組みを提供する「とすると」

3節では、特定の教科に頻出する形式は、その教科の学習で求められる抽象的思考と結びついているという予想に基づいて、特定の教科に頻出する言語形式の調査を行った。一方で、3節の調査では、「AをBとする」という文型が数学という教科とどのように結びついているかが明らかになっていない。よって、4節では、「AをBとする」という文型と数学という教科の関係について分析を行う。

まず、数学教科書において「AをBとする」という文型が、どのような文脈で用いられているのかを確認する。表4は、数学教科書における「とする」と「と言う」の後文脈を比較したものである。

表4 数学教科書における「とする」「と言う」の後文脈

「とする」	出現回数	「と言う」	出現回数
とすると	79	と言う。	180
とします	50	と言い、	15
とする。	49	であると言う	11
とするとき	33	と言います	6
とする点	22	距離と言う	5

数学教科書において、「とする」は接続助詞「と」や「とき」を後接し、従属節に出現するパターンが多いのに対し、「と言う」は、句点や「ます」を後接し、文末に出現するパターンが多いことが見て取れる。

表4から、「AをBとする」は接続助詞トを後接する形で用いられやすいと考えられるが、「とすると」を扱った研究として、中俣(2017)が挙げられる。中俣(2017)では(9)のような例文を取り上げ、「とすると」の機能について、(10)のように説明している。

(9) 1.9%で借り入れたとすると、毎月の返済は5万2,285円です。

(BCCWJ) (中俣2017:93)

(10) すなわち、「PとするとQ」はPという前提の元ではQが成り立つことを意味し、Qを述べるための思考の枠組みを提供する働きがあると言える。

(中俣2017:93)

中学校教科書における(11)の事例も、中俣(2017)が説明する通り、思考の枠組みを提供するような用法で用いられていると考えられる。また、(12)の事例では「とする」が文末で用いられているが、「とする」が出現する文をP、「このとき」から続く文をQと捉えれば、「Pという前提の下ではQが成り立つ」という思考の枠組みを提供する機能を持っている。なお、「提供する」という言葉は、コミュニケーションにおいて聞き手や読み手に対して教える文脈が想起されるので、本稿では「思考の枠組みを設定する」用法としておく。

(11) 紙パックの枚数をx枚とすると、次の等式ができる。

(『新しい数学1』p.92 下線は筆者による)



(12) 縦が 4cm, 横が xcm の長方形の面積を ycm<sup>2</sup> とする。

このとき, y を x の式で表すと  $y = 4x$  となり,  $y = ax$  の形で表されるから, y は x に比例する。

(『新しい数学 1』 p.120 下線は筆者による)

このように、数学における「A を B とする」は主に「とすると」や「とするとき」のような条件に関する接続助詞と共起し、思考の枠組みを設定する用法で用いられていると考えられる。ただし、数学における「とすると」は、(9) のような「たとする」の形で用いられる用例は少なく、主に「A を B とする」という文型で用いられている。この点に着目し、4.2 節では、「A を B とする」という文型が数学という教科とどのように関わっているかを分析する。

#### 4.2 具体的事象における要素と数式に出現する要素を同定する「A を B とする」

4.1 節では、数学における「とすると」が主に条件に関する接続助詞と共起し、思考の枠組みを設定する用法を持つことを確認した。一方、この説明では、数学において「A を B とする」という文型が頻出する要因について説明することができない。よって、4.2 節では「A を B とする」という文型と数学という教科について分析を行う。

結論を先取りすると、4.2 節では、「A を B とする」が数学における立式という抽象的操作と関わる言語形式であることを主張する。立式とは、清水 (2022) において「中学校で学習する方程式の立式では、文章問題から数量を読み取ってそれらの等しい関係を見つけ、等式で表すことが要求される」と述べられているように、文章題などの具体的事象から数量を読み取って数式であらわす抽象的操作のことを言う。文部科学省 (2017) による学習指導要領の解説においても「中学校数学科では、具体的な事象の中から二つの数量を取り出し、それらの変化や対応を調べることを通して、関数関係を見いだし考察し表現する力を 3 年間にわたって徐々に高めていくことが大切である。(p.50)」と述べられており、立式は中学校数学において重要な位置付けを占めている。

まず、「A を B とする」という文型について確認する。金 (2014) によると、「A を B とする」は、「 $A=B$ 」の意味をあらわすとされている。また、「A を B とする」は、(13) のように「だ」を挿入できず、トの前に節を埋め込むような引用のトとは異なるものとされており、「とすると」の前に節を埋め込む「たとする」と「A を B とする」とは、文法的に異なった構造を持つと考えられる。数学における「A を B とする」についても、(14) のように、「だとする」は不自然であり、「 $A=B$ 」という同定の意味をあらわすものと考えられる。これは、(8) のような「 $A=B$  とする」の用例からも支持される。よって、数学における「A を B とする」は、2つの要素が同一であることを示す、同定を行う形式であると考えられる。

(13) a. 父は兄を後継者とした。

b. \*父は兄を後継者だとした。

(金 2014 : 137)

(14) 一辺の長さを x (\*だ) とする。

これを踏まえた上で、「A を B とする」が数学という教科とどのように結びついている

のかを調査する。ここでは、数学教科書の中で、「とする」がどのような単元で頻出するのかを調査した。一口に数学といっても、全ての単元において等質な内容を扱っているわけではない。よって、単元と出現回数との関係は、文型と教科の結びつきを考える上で、重要な証拠となる。表5に、数学教科書の各章において、「とする」がどの程度出現しているかをまとめた表を示す。表5では、出現回数に加え、カッコ内に1万語あたりの出現回数を示している。

表5 数学教科書における「とする」の出現回数と単元の関係

1年生教科書 (『新しい数学1』)		2年生教科書 (『新しい数学2』)		3年生教科書 (『新しい数学3』)	
章タイトル	出現回数 (1万語あたりの回数)	章タイトル	出現回数 (1万語あたりの回数)	章タイトル	出現回数 (1万語あたりの回数)
正負の数	2 (2.5)	式の計算	7 (14.7)	多項式	8 (14.4)
文字式	2 (3.8)	連立方程式	8 (19.0)	平方根	17 (35.7)
方程式	9 (18.1)	1次関数	23 (28.4)	2次方程式	12 (25.2)
比例と反比例	25 (30.7)	平行と合同	13 (20.4)	関数 $y=ax^2$	20 (28.5)
平面図形	20 (31.8)	三角形と四角形	26 (36.0)	相似な図形	28 (37.7)
空間図形	7 (11.0)	確率	7 (16.8)	円	30 (76.1)
データの分析と活用	3 (6.6)	データの比較	1 (4.5)	三平方の定理	38 (84.5)
				標本調査	1 (3.7)

表5の興味深い点として、1年生であれば「文字式」や「方程式」、2年生であれば「式の計算」や「連立方程式」のように、主に計算を扱う章では「とする」の出現回数がそれほど多くないことがわかる。一方で、比例や関数、図形のように、計算だけでなくグラフや図形を扱う章においては、「とする」の出現回数が多くなる。また、3年生では全体的に「とする」の出現回数が多くなっている。

この事実は、「AをBとする」が数学における立式という抽象的操作と関わる言語形式であることを示しているものと考えられる。比例や関数、図形のような章では、グラフや図形にあらわれる要素を式としてあらわす必要がある。よって、他の章と比べ、具体的事象から数量を読み取って数式であらわす、立式という抽象的操作を多く求められることが予想される。更に、立式においては、具体的事象における要素と数式における要素を関連付ける必要があり、「AをBとする」がAとBの同定を行う文型であることも整合的である。なお、計算を主に扱う章においても立式を要するような文章題は一定数登場するが、1年生と2年生では立式を扱う分量よりも計算に割く分量が多く、「とする」の出現回数が少ないものと思われる。

表5から得られた「AをBとする」と立式との関係は、实例を見ても確かめることができる。まず、(15)において、「AをBとする」の「A」は「紙パックの枚数」という具体的事象における個数を示しており、「B」は数式の中で個数をあらわす“x”という文字を示している。そして、「AをBとする」という文型によって、具体的事象における数量を式にあ

らわすための枠組みが示されている。また、(16)において、「AをBとする」の「A」は「2辺の長さ」や「斜辺の長さ」という図形の中の空間的な長さを示しており、「B」は長さを数式の中であらわす“a”や“b”という文字を示している。同じく(16)においても、「AをBとする」という文型によって、具体的事象における数量を式にあらわすための枠組みが示されている。このように、(15)と(16)における「AをBとする」は、具体的事象に出現する要素であるAと数式に出現する要素であるBを同定し、関連付ける機能を持っている。

(15) 紙パックの枚数を  $x$  枚とすると、次の等式ができる。

$$10x + 1000 = 21000 \quad \dots\dots\textcircled{1}$$

(『新しい数学1』p.92)

(16) 直角三角形の直角をはさむ2辺の長さを  $a$ ,  $b$ , 斜辺の長さを  $c$  とすると、次の関係が成り立つ。

$$a^2 + b^2 = c^2$$

(『新しい数学3』p.189)

このように、「AをBとする」という文型は、思考の枠組みを設定するというだけでなく、数学の教科学習においては、具体的事象から数量を読み取って数式であらわす立式という抽象的操作と深くかかわる言語形式であると考えられる。よって、本稿では、数学における「AをBとする」という文型が数学における立式という抽象的操作と関連する言語形式であると捉えるために、4.1節と4.2節の内容を踏まえ、(17)のような教科と文型と機能との対応関係を提案する。

(17) 教科：数学

文型：AをBとする

機能：具体的事象から数量を読み取って数式であらわす立式という操作において、具体的事象における要素を数式に出現する要素と同定し、思考の枠組みを設定する。

4.2節では数学という教科と「AをBとする」という文型の関係について考察した結果、立式を行うという数学における操作と「A=B」という同定を行う「AをBとする」が対応関係を持っていることを示した。

## 5. まとめ

本稿では、大きく分けて以下の2点の分析を提示した。

(18) 理数系教科書の比較を行い、数学の教科学習に結びつく「AをBとする」という文型の存在を明らかにした。

(19) 「とする」の分析から、数学において「AをBとする」という文型が具体的事象における要素と数式に出現する要素を同定する機能を持っていることを明らかにし、数学における立式という抽象的な操作と「AをBとする」という文型が結びついていくことを明らかにした。

最後に、教科学習における日本語の理解と日本語の使用について述べる。前述の通り、文

部科学省 (2017) では、「中学校数学科では、具体的な事象の中から二つの数量を取り出し、それらの変化や対応を調べることを通して、関数関係を見だし考察し表現する力を3年間にわたって徐々に高めていくことが大切である。(p.50)」と述べられている。ここで「関係を見だし考察し表現する」とあるように、数学においては、理解するだけではなく、それを表現する場面も存在する。その一例として、証明問題では、日本語を使って数学的な思考について説明する必要がある。このような背景を踏まえると、数学という教科と日本語の関係は、単なる理解の問題だけでなく、日本語を使って考察し、説明を行うという使用の問題とも関連するものと考えられる。よって、教科書の調査だけでなく、教科学習でどのような形で日本語を使用する必要があるかという点についても研究が必要だと思われるが、これについては今後の課題としたい。

### 謝 辞

本研究の成果は、以下の共同研究および助成事業の成果に基づくものである。  
筑波大学・凸版印刷株式会社共同研究「ICT 教材の活用による多文化的背景を持つ生徒支援施策の研究」(研究代表者：澤田浩子)  
茨城県教育委員会，グローバル・サポート事業委託研究「オンライン学習による日本語初期指導カリキュラム開発・検証に関する研究」(研究代表者：澤田浩子)  
日本学術振興会，科学研究費補助金・基盤研究(B)「学びの場における CLD 生徒の言語使用の分析とデータベースの構築」(研究代表者：澤田浩子，22H00666)  
日本学術振興会，科学研究費補助金・基盤研究(B)「インクルーシブ教育システムにおけるインテンシブ・ニーズ支援モデルの構築」(研究代表者：米田宏樹，22H01031)

### 参考文献

- 金賢娥 (2014). 『現代日本語における助詞「ト」の研究：引用の周辺にある「ト」を中心に』筑波大学博士論文.
- 小磯花絵・天谷晴香・居關友里子・白田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 (2023). 「『日本語日常会話コーパス』設計と特徴」『国立国語研究所論集』24, pp.153-168.
- 小澤俊介・内元清貴・伝康晴 (2014). 「BCCWJ に基づく長単位解析ツール Comainu」『言語処理学会第20回年次大会論文集』pp.582-585.
- 竹本理美・新山聖也・澤田浩子 (2023). 「外国人生徒による受動文の文構造理解：教科学習におけるつまづきを探る」『全国大学国語教育学会国語科教育研究：大会研究発表要旨集』144, pp.71-74.
- 清水宏幸 (2022). 「学校数学における文字式に関する研究：式をひとまとまりと見ることに焦点を当てて」『日本数学教育学会誌』103:R117, pp.5-14.
- 西川朋美・青木由香 (2020) 「日本生まれ・育ちのJSLの子どもの格助詞の産出：記述式テストで見られたモノリンガルとの違い」『日本語教育』177, pp.47-61.
- 宮部真由美 (2019). 「トの分析からみた中学校数学科教科書の日本語の難しさ：日本語学習者の教科学習における日本語の困難点とは」『日本語／日本語教育研究』10, pp.117-131.
- 宮部真由美 (2021). 「中学校数学科教科書の内容理解における日本語の困難点：日本語を母語としない中学生の教科学習支援を目指して」『人文・自然研究』15, pp.127-139.
- 文部科学省 (2017). 「【数学編】中学校学習指導要領(平成29年告示)解説」

[https://www.mext.go.jp/component/a\\_menu/education/micro\\_detail/\\_icsFiles/afieldfile/2019/03/18/1387018\\_004.pdf](https://www.mext.go.jp/component/a_menu/education/micro_detail/_icsFiles/afieldfile/2019/03/18/1387018_004.pdf)

文部科学省 (2022). 「日本語指導が必要な児童生徒の受入状況等に関する調査結果について

[https://www.mext.go.jp/content/20221017-mxt\\_kyokoku-000025305\\_02.pdf](https://www.mext.go.jp/content/20221017-mxt_kyokoku-000025305_02.pdf)

Cummins, J. (1979). Cognitive/Academic Language Proficiency, Linguistic Interdependence, the Optimum Age Question and Some Other Matters. *Working Papers on Bilingualism*, 19, pp.197-205.

#### 調査資料

梶田隆章・真行寺千佳子・永原裕子・西原寛 ほか 131 名 (2022). 『新しい科学 1』 東京書籍株式会社.

梶田隆章・真行寺千佳子・永原裕子・西原寛 ほか 131 名 (2022). 『新しい科学 2』 東京書籍株式会社.

梶田隆章・真行寺千佳子・永原裕子・西原寛 ほか 131 名 (2022). 『新しい科学 3』 東京書籍株式会社.

藤井斉亮・真島秀行 ほか 94 名 (2022). 『新しい数学 1』 東京書籍株式会社.

藤井斉亮・真島秀行 ほか 94 名 (2022). 『新しい数学 2』 東京書籍株式会社.

藤井斉亮・真島秀行 ほか 94 名 (2022). 『新しい数学 3』 東京書籍株式会社.

『現代日本語書き言葉均衡コーパス (BCCWJ)』 <https://clrd.ninjal.ac.jp/bccwj/>

『日本語日常会話コーパス (CEJC)』 <https://www2.ninjal.ac.jp/conversation/cejc.html>

## 『昭和・平成書き言葉コーパス』雑誌レジスターに見る 順接・逆接の接続詞の通時的変化

近藤 明日子（東京大学）<sup>†</sup>

### Diachronic Variation of Illative and Adversative Conjunctions: An Analysis with the Magazine Register of the Showa-Heisei Corpus of Written Japanese

KONDO Asuko (The University of Tokyo)

#### 要旨

『昭和・平成書き言葉コーパス』雑誌レジスターを用いて、昭和・平成期の非文芸ジャンルの書き言葉における順接・逆接の接続詞について考察した。まず、コーパスから作成した短単位 n-gram を用いて接続詞語形を網羅的に抽出したところ、順接の接続詞は 31 語形、逆接の接続詞は 17 語形が確認できた。これらの語形について口語文体（常体多 / 敬体多）割合を分析すると、語形と口語文体割合には対応関係が見られ、それは各語形の書き言葉的・話し言葉的性質の強弱と対応していることが明らかになった。次に、『日本語歴史コーパス 明治・大正編 I 雑誌』のデータと合わせて、刊行年別に接続詞の各語形の使用サンプル率を算出し、明治期から平成期までの量的な通時的変化を概観した。話し言葉的性質の強い語形が増加すること、また特に常体多の口語文体では語形の整理が進み、少数の語形が使用されるようになることが明らかになった。

#### 1. はじめに

明治・大正期は、書き言葉の文体の標準が文語体から口語体へと大きく変化した時期である。そして、その中で使用される語彙も大きく変化し、接続詞もまた例外ではない。その様相は、当時の総合雑誌を収録した国立国語研究所 (2019) 『日本語歴史コーパス 明治・大正編 I 雑誌』（以下、「CHJ 雑誌」と呼ぶ）から抽出した順接・逆接の接続詞の語形群の通時的変化と文体変遷との対応関係を分析・考察した近藤 (2021a) (2021b) に詳しい。当時の総合雑誌において文語体から口語体への文体転換は大正末期に完了したが、口語体の内部の変化はそれ以降も続き、昭和・平成期を経て現在の書き言葉につながっている。そこで、昭和・平成期の接続詞の変遷の実態を辿るため、本稿では小木曾ほか (2023) 『昭和・平成書き言葉コーパス』の雑誌レジスター（以下、「SHC 雑誌」と呼ぶ）を利用し、近藤 (2021a) (2021b) の手法に倣い、順接・逆接の接続詞を網羅的に抽出し、文体（常体/敬体）との対応関係を分析する。そして、CHJ 雑誌のデータと合わせて、明治期から平成期までの接続詞の量的な通時的変化を概観する。

#### 2. 先行研究

近現代語の資料からの接続詞の網羅的抽出とその量的な通時的変化の考察を行った先行研究として、まず京極・松井 (1973) がある。これは古代から近代までの各時代の接続詞の語形とその頻度を掲出し考察するもので、そのなかで近現代に相当する接続詞の考察は明

<sup>†</sup> akondo [アットマーク] l.u-tokyo.ac.jp

治 17 年から昭和 31 年成立の文芸作品 16 作品を対象として行っている。その他、石田 (2009) は、1906～2006 年の雑誌『中央公論』から 10 年間隔で各年 1 万語をサンプリングした計 11 万語のテキストから、国立国語研究所 (編) (2004) 『分類語彙表 増補改訂版』に掲出の接続詞を抽出し、その量的な通時的変化を考察するものである。本稿は、総合雑誌という多種の分野の文章を掲載する資料における非文芸ジャンルの書き言葉を対象とする点で、文芸作品を対象とする京極・松井 (1973) とは異なる。また、石田 (2009) とは対象とする資料が重複するが、本稿は SHC 雑誌だけで 2300 万語という大規模なテキストを対象とし、また特定のリストに拠らず、できるかぎり多くの語形の抽出を試みる点で、近現代の接続詞の通時的変化をより精緻に明らかにすることができると思われる。

### 3. 使用するコーパスとテキスト

本稿で研究資料とする SHC 雑誌は、昭和・平成期の代表的な月刊総合雑誌『中央公論』『文芸春秋』の 1933 年～2015 年の間の 8 年おき 11 か年分を収録刊年とし、各年 1 誌分の通常号 12 冊、計 132 冊の全文を収録したコーパスであり、CHJ 雑誌と接続して利用することを想定した設計となっている (近藤 2023)。

SHC 雑誌は 1 記事=1 サンプルとして、テキストをサンプル単位に分割して収録し、各サンプルにはジャンル情報を付与する。本稿では、文芸ジャンル (小説・戯曲・詩歌) を除いた非文芸ジャンルのサンプルを研究対象テキストとした。ただし、非文芸ジャンルのサンプルのうちサンプル ID の末尾が「000」のサンプル (雑誌 1 冊から記事部分を除いた残りのテキストを集めたサンプル) は除いた。

また、接続詞の語形と文体との対応関係を分析するにあたり、文末辞「だ」「である」を主に使用する「常体多」のサンプルと「です」「ます」を主に使用する「敬体多」のサンプルに分類した。近藤 (2021a) (2021b) での方法に倣い、敬体の文末辞「です」「ます」「でございます」「であります」と常体の文末辞「である」「だ (意志推量形・終止形のみ)」の頻度をサンプルごとに集計し、文末辞の頻度合計に対する敬体の文末辞の頻度合計の割合が 0.5 以上のサンプルを「敬体多」、0.5 未満のサンプルを「常体多」とした。以下、「常体多」「敬体多」で分類する文体を「口語文体」と呼ぶ。

次の表 1 に、研究対象テキストのサンプル数と延べ語数 (記号類・未知語類は除く) を口語文体・刊行年ごとに示す。

表 1 研究対象テキストのサンプル数と延べ語数

刊行年	常体多		敬体多		合計	
	サンプル数	延べ語数	サンプル数	延べ語数	サンプル数	延べ語数
1933	480	2,213,394	84	233,566	564	2,446,960
1941	402	1,843,996	24	185,410	426	2,029,406
1949	167	673,923	28	172,353	195	846,276
1957	599	2,248,305	114	426,799	713	2,675,104
1965	935	1,467,400	200	217,236	1,135	1,684,636
1973	593	1,714,941	164	295,979	757	2,010,920
1981	587	1,942,241	253	365,484	840	2,307,725
1989	853	1,733,782	388	761,363	1,241	2,495,145
1997	734	1,652,672	288	650,734	1,022	2,303,406
2005	895	1,419,814	380	848,046	1,275	2,267,860
2013	750	1,128,847	530	1,307,663	1,280	2,436,510
計	6,995	18,039,315	2,453	5,464,633	9,448	23,503,948

表 1 の口語文体の延べ語数の割合の通時的変化を示したものが図 1 である。

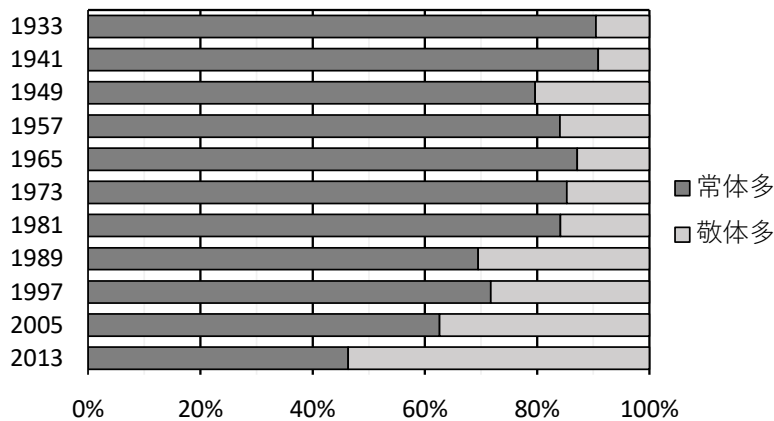


図 1 口語文体の延べ語数の割合の通時的変化

図 1 によれば、1981 年までは常体多が約 80~90%を占めていたものが、以降その割合を減らし、2013 年は敬体多のほうが過半を占めるに至る。このように SHC 雑誌の非文芸ジャンルのサンプルでは常体多の減少と敬体多の増加という口語文体の変化があることが明らかになった。

#### 4. 短単位 n-gram による接続詞の抽出

次に、3. で定めた研究対象テキストから接続詞を抽出する方法とその抽出結果について述べる。SHC 雑誌の語の単位は「短単位」と呼ばれる比較的短い単位である。接続詞は複数短単位の連続で構成される場合も多いため、品詞が「接続詞」の 1 短単位のみを抽出しても網羅的な抽出ができない。そこで、近藤 (2021a)(2021b) に倣い、対象テキストから 1~8gram の短単位の n-gram を作成し、そこから接続詞と考えられるものを独自の判断で抽出する方法をとった。接続詞の可能性の高いものを効率的に抽出するため、抽出の対象とする n-gram



は文頭<sup>1</sup>のもの、かつ出現するサンプル数が 95 以上（対象テキストの全サンプル数 9448 [表 1 参照] の 1%以上）のものに限った。そして、1gram からは品詞が「接続詞」のものを接続詞として抽出した。2～8gram からは『日本国語大辞典 第 2 版』（小学館）等の辞典類や青木 (1973) 等を参照しつつ、そこには載っていない語形も含め、独自の判断で順接・逆接の接続詞を抽出した。接続詞の用法分類は市川 (1976) に拠り、「二つの事柄を論理的に結びつけて述べるのに用いる」接続詞のうち、「前の内容を条件として、それから生じる結果を導く」ものを順接、「前の内容に反する内容を導く」ものを逆接と認定した<sup>2</sup>。

以上の方法により抽出した順接の接続詞の全語形を示したものが表 2、逆接の接続詞の全語形を示したものが表 3 である。表 2・表 3 とも、語形は五十音順に掲出し、口語文体別に粗頻度 (a) ・使用サンプル数 (b) ・使用サンプル率 (c) を示した。使用サンプル率とは、各口語文体の全サンプル数（常体多は 6995、敬体多は 2453、表 1 参照）に対する該当語形の使用サンプル数の割合を示す。

表 2・表 3 にあるように、調査対象テキストから順接の接続詞は 31 語形、逆接の接続詞は 19 語形が抽出された。

---

<sup>1</sup> 文頭に記号類がある場合は、それを除いて短単位 n-gram を作成した。

<sup>2</sup> 語形単位で接続詞を認定したため、個々の用例の中には接続詞ではないもの、接続詞であっても別用法のものが含まれる場合もある。ただし、以下の (1) (2) の場合については、該当語形からあらかじめ該当用例を除外する処理を行った。

- (1) 接続詞として認定した別の短単位 n-gram が含まれている場合。具体的には「これによって」から「これによってみれば」「これによってみると」「これによってみるに」「これによってみても」を除外、「それなら」から「それならば」を除外、「それゆえ」から「それゆえに」を除外、「しかし」から「しかしながら」を除外した。
- (2) 後接語から明らかに接続詞ではないと判断できる場合。具体的には「そのけっか」から「そのけっかは」「そのけっかと」は除外した。

表 2 順接の接続詞の粗頻度(a)・使用サンプル数(b)・使用サンプル率(c)

語形	常体多			敬体多		
	(a)	(b)	(c)	(a)	(b)	(c)
かくして	422	272	3.9%	24	17	0.7%
こうなると	166	144	2.1%	21	20	0.8%
これによって	133	122	1.7%	24	22	0.9%
されば	170	103	1.5%	11	8	0.3%
しからば	260	171	2.4%	6	6	0.2%
したがって	2,868	1,218	17.4%	342	192	7.8%
じゃあ	68	54	0.8%	58	48	2.0%
すると	993	619	8.8%	471	302	12.3%
そうしたら	76	65	0.9%	133	91	3.7%
そうすると	177	137	2.0%	273	154	6.3%
そうすれば	163	153	2.2%	88	75	3.1%
そうなると	188	168	2.4%	116	88	3.6%
そうなれば	82	75	1.1%	44	38	1.5%
そこで	3,515	1,752	25.0%	1,244	586	23.9%
そしたら	62	47	0.7%	74	61	2.5%
そのけっか	532	437	6.2%	188	132	5.4%
そのため	973	713	10.2%	374	261	10.6%
そのために	503	406	5.8%	229	182	7.4%
それで	2,382	1,374	19.6%	1,605	623	25.4%
それなら	178	157	2.2%	63	54	2.2%
それによって	94	79	1.1%	35	33	1.3%
それによると	105	92	1.3%	22	21	0.9%
それゆえ	200	137	2.0%	24	21	0.9%
それゆえに	241	139	2.0%	5	5	0.2%
だから	3,825	1,736	24.8%	2,325	675	27.5%
で	350	261	3.7%	260	147	6.0%
ですから	51	45	0.6%	766	353	14.4%
では	751	499	7.1%	282	189	7.7%
とすれば	127	110	1.6%	13	11	0.4%
ならば	92	83	1.2%	35	30	1.2%
ゆえに	369	192	2.7%	32	27	1.1%

表 3 逆接の接続詞の粗頻度(a)・使用サンプル数(b)・使用サンプル率(c)

語形	常体多			敬体多		
	(a)	(b)	(c)	(a)	(b)	(c)
が	1,673	776	11.1%	228	125	5.1%
けれど	863	433	6.2%	244	149	6.1%
けれども	739	364	5.2%	170	103	4.2%
しかし	18,128	4,072	58.2%	5,022	1,041	42.4%
しかしながら	714	381	5.4%	88	58	2.4%
しかるに	1,020	460	6.6%	39	25	1.0%
それでも	853	659	9.4%	369	284	11.6%
それなのに	211	185	2.6%	94	82	3.3%
それにもかかわらず	225	188	2.7%	18	18	0.7%
だが	6,222	1,970	28.2%	505	204	8.3%
だけど	127	82	1.2%	314	145	5.9%
だって	108	92	1.3%	118	84	3.4%
でも	837	416	5.9%	1,696	530	21.6%
といっても	148	131	1.9%	44	43	1.8%
ところが	4,316	2,047	29.3%	1,920	715	29.1%
とはいえ	203	179	2.6%	86	71	2.9%
にもかかわらず	294	253	3.6%	64	59	2.4%

### 5. 接続詞の語形と口語文体との対応関係

次に4.で抽出した接続詞の語形と口語文体との対応関係を分析する。口語文体の使用サンプル率の割合を示したものが図2・図3である。図2・図3とも常体多の割合の高い順に語形を掲載した。

図2・図3によれば、順接・逆接の接続詞ともに、常体多の割合の高い語形から敬体多の割合の高い語形まで様々である。そのなかで、各口語文体の割合が高い語形は該当口語文体との対応関係が強いものである。敬体多のサンプルは、インタビューや座談会を記録した談話体の文章も多く、話し言葉的性質の強い文章と言える。対して常体多のサンプルは相対的に書き言葉的性質の強い文章と言える。よって、常体多の割合の高い語形は書き言葉的性質が強く、敬体多の割合の高い語形は話し言葉的性質が強いと見做することができる。例えば、「しからば」「それゆえに」「かくして」「されば」「しかるに」等は書き言葉的性質の特に強い語形、「ですから」「そうしたら」「だけど」等は話し言葉的性質の特に強い語形となる。このように、接続詞の語形と口語文体との間には対応関係が認められ、それは書き言葉的あるいは話し言葉的性質の強弱との対応関係と見做される。

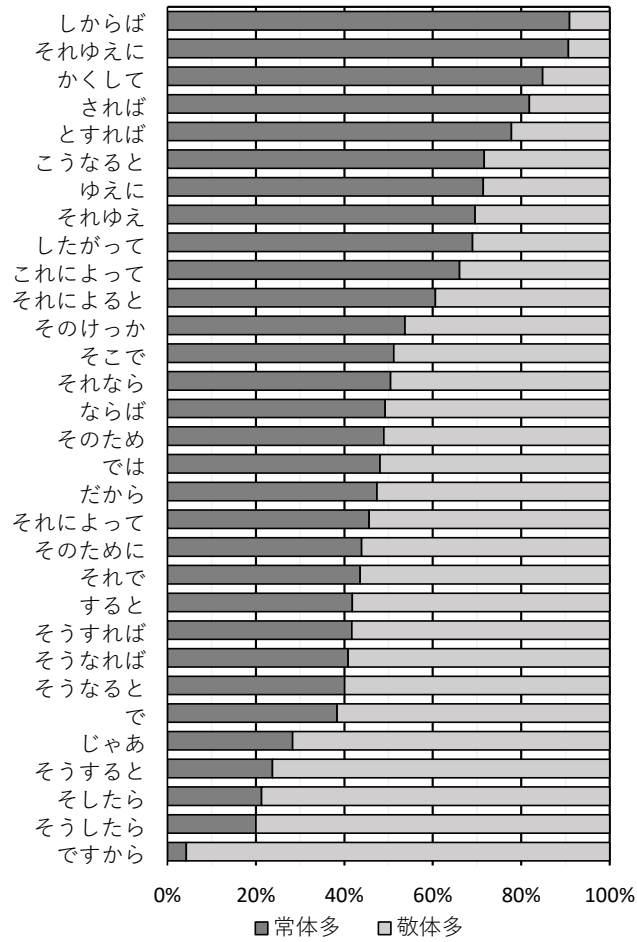


図 2 順接の接続詞の使用サンプル率の口語文体割合

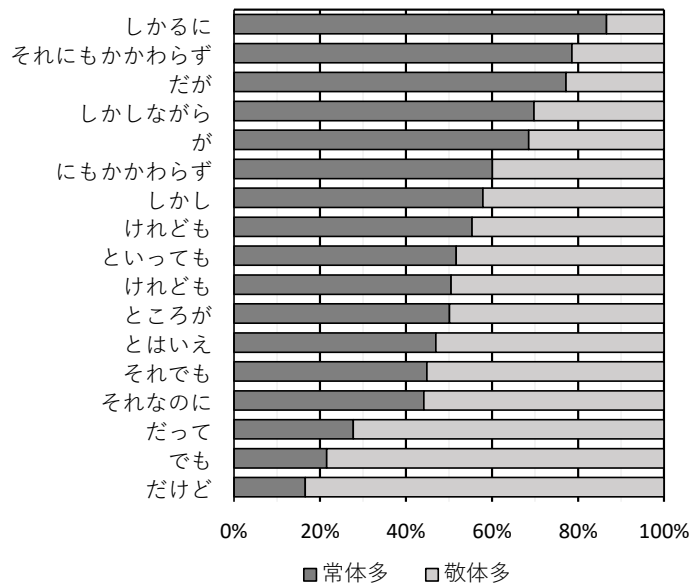


図 3 逆接の接続詞の使用サンプル率の口語文体割合

## 6. 接続詞の量的な通時的変化の分析

次に、抽出した接続詞の語形の量的な通時的変化について分析する。分析は使用サンプル率に基づくこととするが、分析の前提として、4. で抽出した接続詞が使用されるサンプルの全サンプル数に対する割合の通時的変化を確認する。表2・表3にあげた接続詞の語形のいずれかが1例以上使用されるサンプル数の全サンプルに占める割合を接続詞用法・口語文体別に示したものが図4である。

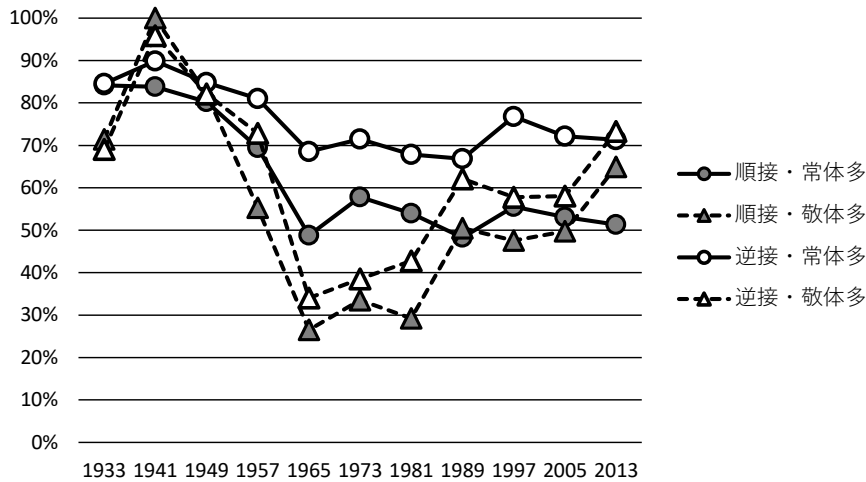


図4 各刊行年の接続詞使用サンプル率

図4によれば、接続詞の使用サンプル率は刊行年によってまちまちである。例えば、1941・1949年のようにどの用法・口語文体でも使用サンプル率の高い年もあれば、1965・1973・1981年のように敬体多の文体で使用サンプル率の低い年もある。このような高低が生じる要因として、雑誌の編集方針によって掲載される記事の種類が異なっていることが考えられる。1941・1949年は硬い内容の論説文のサンプルが多い、1965・1973・1981年は敬体多のサンプルは語数の少ない短いものが多い、といったものである。接続詞自体の消長を考察する場合、編集方針のような外的要因の影響を取り除く必要がある。そこで、以下では、順接・逆接の接続詞を使用する全サンプル数に対する該当語形を使用するサンプル数の割合を「使用サンプル率」として分析を行い、接続詞内部での各語形の勢力関係の通時的変化を見ていくこととする。

次の表4～表7は、1895年～2013年の接続詞用法・口語文体ごとに接続詞の各語形の使用サンプル率の通時的変化を示したものである。1895年～1925年のデータはCHJ雑誌の中の雑誌『太陽』の口語体のサンプル<sup>3</sup>から算出したもので、近藤(2021a)(2021b)で示したものと同一である。掲出した語形はいずれかの刊行年で使用サンプル率が20%以上のものに限定している。これらは近現代の非文芸ジャンルの書き言葉において主要な語形であると言える。表中空欄であるのは、CHJ雑誌・SHC雑誌のそれぞれで使用サンプル数が少なく、抽出対象外となった語形である。また、各表の最終行には接続詞を使用するサンプルの総数を示した。各表の使用サンプル率はこのサンプル総数を100%としたときの値である。表のセルは使用サンプル率によって色分けし、50%以上は■、20%以上は■、10%以上は■

<sup>3</sup> 研究対象のテキストの詳細は近藤(2021a)(2021b)を参照のこと。

、5%以上は で示した。

表4 常体多における主な順接の接続詞の使用サンプル率の通時的変化

語形	1895	1901	1909	1917	1925	1933	1941	1949	1957	1965	1973	1981	1989	1997	2005	2013
かくして	0.0%	5.4%	4.7%	10.8%	8.1%	15.8%	22.6%	14.2%	4.6%	3.5%	3.8%	7.6%	4.1%	2.2%	2.5%	0.8%
ここにおいてか	0.0%	51.7%	2.5%	5.2%	1.7%											
されば	28.6%	23.0%	19.5%	29.1%	7.6%	11.4%	9.8%	4.5%	0.5%	1.1%	1.2%	0.6%	0.7%	0.5%	0.0%	0.0%
しからば	28.6%	12.2%	19.5%	23.1%	18.1%	19.1%	16.6%	8.2%	1.7%	1.1%	0.9%	2.2%	0.5%	0.2%	0.4%	0.0%
したがって	0.0%	10.8%	39.0%	47.0%	48.0%	52.5%	59.6%	45.5%	43.8%	27.0%	30.3%	32.5%	21.1%	16.7%	12.2%	4.9%
すると	28.6%	14.9%	5.8%	10.8%	15.8%	14.6%	10.4%	13.4%	12.3%	16.4%	22.7%	15.8%	11.9%	19.1%	15.2%	14.0%
そこで	71.4%	55.4%	31.0%	35.1%	46.1%	53.7%	38.9%	37.3%	54.8%	42.3%	53.6%	45.7%	40.8%	39.7%	31.4%	32.5%
そのため	0.0%	0.0%	0.7%	1.6%	3.6%	9.9%	15.4%	15.7%	19.7%	15.1%	21.0%	24.9%	19.7%	19.1%	15.4%	17.1%
それで	42.9%	29.7%	13.0%	17.5%	21.0%	36.4%	22.0%	37.3%	32.5%	30.5%	35.6%	40.7%	33.0%	39.0%	33.5%	32.2%
それゆえに	28.6%	8.1%	3.2%	9.6%	2.4%	12.4%	10.7%	8.2%	2.2%	0.7%	0.6%	1.3%	0.0%	2.0%	2.3%	1.3%
だから	0.0%	6.8%	18.8%	5.6%	12.6%	48.0%	26.7%	39.6%	51.9%	45.8%	47.5%	46.4%	45.6%	40.4%	38.1%	33.8%
で	0.0%	21.6%	20.6%	11.2%	16.9%	15.3%	5.0%	11.2%	4.3%	2.9%	4.4%	7.9%	9.0%	7.1%	3.8%	3.1%
ゆえに	57.1%	37.8%	41.2%	39.0%	22.9%	16.8%	14.2%	6.7%	1.2%	0.7%	3.5%	3.8%	1.2%	1.7%	2.5%	2.9%
順接の接続詞 使用サンプル総数	7	74	276	250	419	404	337	134	416	456	343	317	412	408	475	385

表5 敬体多における主な順接の接続詞の使用サンプル率の通時的変化

語形	1895	1901	1909	1917	1925	1933	1941	1949	1957	1965	1973	1981	1989	1997	2005	2013
されば	33.3%	33.3%	11.1%	7.7%	4.8%	5.0%	8.3%	8.7%	0.0%	0.0%	1.8%	0.0%	0.0%	0.0%	0.0%	0.0%
しからば	16.7%	20.0%	0.0%	0.0%	9.5%	5.0%	4.2%	4.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%
したがって	11.1%	11.1%	22.2%	38.5%	28.6%	35.0%	33.3%	56.5%	46.0%	35.8%	25.5%	13.5%	11.2%	10.9%	9.0%	7.0%
すると	0.0%	6.7%	22.2%	7.7%	4.8%	31.7%	33.3%	13.0%	17.5%	11.3%	18.2%	18.9%	18.4%	25.5%	29.1%	30.5%
そうすると	11.1%	2.2%	11.1%	0.0%	4.8%	11.7%	37.5%	26.1%	27.0%	18.9%	21.8%	18.9%	11.7%	14.6%	7.4%	6.4%
そうすれば	5.6%	2.2%	0.0%	7.7%	0.0%	5.0%	4.2%	21.7%	12.7%	5.7%	10.9%	5.4%	4.6%	6.6%	9.0%	2.9%
そうなる						3.3%	12.5%	21.7%	14.3%	13.2%	10.9%	6.8%	7.1%	3.6%	5.3%	6.4%
そこで	44.4%	66.7%	33.3%	23.1%	47.6%	38.3%	62.5%	78.3%	65.1%	54.7%	54.5%	37.8%	42.3%	45.3%	48.1%	48.3%
そのけっか	16.7%	6.7%	0.0%	0.0%	4.8%	3.3%	16.7%	13.0%	7.9%	24.5%	10.9%	9.5%	9.7%	10.2%	11.6%	10.8%
そのため	0.0%	0.0%	0.0%	7.7%	4.8%	8.3%	20.8%	21.7%	27.0%	34.0%	18.2%	18.9%	18.4%	14.6%	30.7%	21.2%
そのために	0.0%	0.0%	0.0%	0.0%	0.0%	3.3%	20.8%	21.7%	20.6%	24.5%	12.7%	14.9%	12.2%	11.7%	21.2%	13.4%
それで	55.6%	33.3%	66.7%	53.8%	19.0%	41.7%	75.0%	65.2%	55.6%	37.7%	47.3%	47.3%	51.5%	55.5%	55.0%	48.8%
それゆえ	5.6%	24.4%	0.0%	15.4%	4.8%	6.7%	0.0%	13.0%	1.6%	1.9%	0.0%	2.7%	0.5%	2.9%	1.6%	0.6%
それゆえに	38.9%	24.4%	0.0%	0.0%	9.5%	1.7%	0.0%	4.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.1%	0.3%
だから	0.0%	6.7%	0.0%	15.4%	9.5%	26.7%	50.0%	60.9%	76.2%	41.5%	56.4%	58.1%	47.4%	63.5%	59.8%	57.0%
で	11.1%	13.3%	22.2%	0.0%	28.6%	25.0%	20.8%	4.3%	11.1%	1.9%	12.7%	28.4%	14.3%	19.7%	8.5%	5.5%
ですから						13.3%	29.2%	39.1%	47.6%	41.5%	29.1%	35.1%	35.2%	29.9%	25.4%	22.4%
では						13.3%	0.0%	8.7%	6.3%	17.0%	16.4%	16.2%	9.7%	13.1%	20.6%	20.1%
はたしてしからば	0.0%	26.7%	0.0%	0.0%	0.0%											
ゆえに	55.6%	57.8%	11.1%	0.0%	23.8%	8.3%	8.3%	0.0%	1.6%	0.0%	10.9%	0.0%	2.6%	1.5%	1.6%	0.9%
よって	27.8%	8.9%	0.0%	0.0%	9.5%											
順接の接続詞 使用サンプル総数	18	45	9	13	21	58	23	23	83	68	64	109	241	167	224	395

表6 常体多における主な逆接の接続詞の使用サンプル率の通時的変化

語形	1895	1901	1909	1917	1925	1933	1941	1949	1957	1965	1973	1981	1989	1997	2005	2013
が	36.4%	10.0%	15.9%	16.0%	30.1%	29.8%	17.0%	17.9%	16.6%	11.4%	20.9%	16.0%	17.1%	12.9%	6.8%	6.5%
けれど	9.1%	0.0%	4.7%	2.3%	3.6%	16.3%	20.8%	17.2%	12.7%	7.1%	5.8%	4.2%	5.2%	5.1%	5.2%	3.5%
けれども	18.2%	20.0%	32.9%	13.0%	23.0%	15.4%	19.7%	13.1%	12.1%	6.0%	3.9%	3.5%	5.0%	2.6%	3.7%	1.9%
しかし	72.7%	65.7%	70.8%	71.8%	64.0%	85.3%	86.8%	91.0%	86.1%	74.7%	76.8%	80.7%	76.1%	73.8%	72.3%	67.4%
しかしながら	9.1%	15.7%	21.3%	36.6%	20.9%	21.9%	32.9%	15.9%	8.1%	1.7%	3.0%	3.0%	2.8%	4.0%	3.2%	1.7%
しかるに	18.2%	57.1%	50.5%	64.9%	32.6%	40.9%	42.9%	22.1%	6.7%	3.5%	3.2%	2.7%	0.7%	0.9%	1.4%	0.0%
だが	9.1%	0.0%	4.0%	1.1%	11.5%	41.3%	27.8%	36.6%	41.2%	32.4%	42.5%	42.7%	29.9%	37.4%	40.5%	41.3%
ところが	45.5%	27.1%	24.6%	20.6%	32.0%	46.4%	31.3%	44.8%	47.1%	43.9%	45.2%	44.0%	37.5%	39.0%	30.9%	26.3%
逆接の接続詞 使用サンプル総数	11	70	301	262	469	416	371	145	495	649	431	405	578	572	650	537

表7 敬体多における主な逆接の接続詞の使用サンプル率の通時的変化

語形	1895	1901	1909	1917	1925	1933	1941	1949	1957	1965	1973	1981	1989	1997	2005	2013
けれども	16.7%	15.9%	33.3%	16.7%	50.0%	13.8%	47.8%	34.8%	18.1%	2.9%	6.3%	3.7%	3.7%	10.2%	5.4%	3.3%
しかし	61.1%	52.3%	66.7%	88.9%	75.0%	82.8%	82.6%	95.7%	92.8%	80.9%	70.3%	52.3%	73.4%	71.3%	69.2%	67.6%
しかしながら	44.4%	22.7%	33.3%	5.6%	18.8%	12.1%	30.4%	21.7%	3.6%	1.5%	1.6%	2.8%	4.6%	3.0%	1.8%	2.8%
しかるに	66.7%	54.5%	11.1%	27.8%	37.5%	19.0%	17.4%	8.7%	4.8%	1.5%	0.0%	0.0%	0.0%	1.2%	0.0%	0.3%
しかれども	5.6%	27.3%	0.0%	0.0%	0.0%											
それでも	0.0%	4.5%	0.0%	5.6%	12.5%	8.6%	8.7%	13.0%	4.8%	7.4%	18.8%	17.4%	20.3%	19.8%	22.3%	25.8%
だが	0.0%	6.8%	0.0%	0.0%	0.0%	6.9%	34.8%	30.4%	19.3%	11.8%	7.8%	19.3%	10.8%	13.2%	17.0%	12.4%
でも	0.0%	2.3%	0.0%	0.0%	0.0%	8.6%	30.4%	0.0%	9.6%	22.1%	28.1%	37.6%	33.6%	45.5%	44.6%	45.3%
ところが	22.2%	13.6%	33.3%	22.2%	25.0%	43.1%	73.9%	65.2%	61.4%	41.2%	51.6%	37.6%	50.2%	49.7%	46.0%	50.1%
逆接の接続詞 使用サンプル総数	18	44	9	18	16	60	24	23	63	53	55	74	196	137	189	344

表4～表7から、明らかな量的な増加・減少の傾向が見られる語形を取りだしまとめたものが、次の表8である。

表8 明らかな増加・減少の傾向が見られる接続詞

		増加	減少
順接	常体多	そのため	かくして、ここにおいてか、 されば、しからば、したがって、 それゆえに、で、ゆえに
	敬体多	すると、そうすると、そのため、 そのために、ですから、では	されば、しからば、それゆえ、 それゆえに、はたしてしからば、 ゆえに、よって
逆接	常体多	だが	けれども、しかしながら、しかるに
	敬体多	それでも、だが、でも	けれども、しかしながら、しかるに、 しかれども

表8から以下のことが言える。まず、増加傾向の語形と減少傾向の語形を比較する。増加傾向の語形の中には、近藤 (2021a)(2021b) によると CHJ 雑誌の小説・戯曲の会話文で主に使用される「すると」「そうすると」「そのために」「それでも」「だが」「でも」があり、減少傾向の語形の中には CHJ 雑誌の非文芸ジャンルの文語体で主に使用される「ここにおいてか」「されば」「しからば」「はたしてしからば」「ゆえに」「よって」「しかれども」がある。つまり、増加傾向の語形は話し言葉的性質の強いのに対し、減少傾向の語形は書き言葉的性質の強いことが分かる。近藤 (2021a)(2021b) は、明治・大正期の口語体の発生・定着の過程において、敬体多の口語文体ではより話し言葉的性質の強い語形の使用が増加し、一方常体多の口語文体ではそのような傾向は見られず、文語体で多用される書き言葉的性質の非常に強い語形も引き続き使用されたとする。それが、昭和・平成期に入り、敬体多では明治・大正期から引き続いて話し言葉的性質の強い語形の増加が進んだことになる。ま

た常体多でも、書き言葉的性質の強い語形が減少し、相対的に話し言葉的性質の強い語形の使用が増加したことになる。このように、口語体で使用される接続詞の語形の通時的変化は、大局的には書き言葉の話し言葉的性質の強化という変化の中に位置づけられると言える。

次に、常体多と敬体多の間で比較すると、常体多は増加傾向の語形よりも減少傾向の語形の方が数が多く、結果的に主要な語形数が少数に固定されていく。それに対し、敬体多は減少傾向の語形もある一方で、増加傾向の語形も少なくなく、結果的に主要な語形は常体多より種類に富む状態が維持されている。常体多においては、明治・大正期に見られた多様な語形が昭和・平成期で整理され、主に使用される語形が限定される時期に入っていると見られる。敬体多にはそのような傾向が見られないのは、敬体多のサンプルには、初めから書き言葉として書かれた文章だけでなく、インタビュー・座談会等の話し言葉の記録という性質の異なる文章も多く含まれており、その多様性が語形の種類の多さにつながっているためと考えられる。

## 7. おわりに

以上、『昭和・平成書き言葉コーパス』雑誌レジスターを資料として、昭和・平成期の非文芸ジャンルの書き言葉における順接・逆接の接続詞の口語文体との対応関係や量的な通時的変化を分析した。まず、短単位 **n-gram** を利用して接続詞の網羅的な抽出を行い、順接の接続詞は 31 語形、逆接の接続詞は 17 語形を得た。次にこれらの語形の口語文体割合を見ると、語形と口語文体の間には対応関係が存在し、それは書き言葉的・話し言葉的性質の強弱に対応していることが明らかになった。最後に、『日本語歴史コーパス 明治・大正編 I 雑誌』のデータと合わせて、明治期から平成期にかけての量的な通時的変化を概観した。話し言葉的性質の強い語形が増加すること、また特に常体多の口語文体では語形の整理が進み、少数の語形が使用されるようになることが明らかになった。

今後の課題としては、通時的変化についてより詳細な考察を加えることが必要と考える。変化の傾向の分類については、統計的手法を取り入れ、より多くの語形について分析対象としたい。また、変化の要因についても、語形間の用法の差異にも目を配りつつ考察を深めたい。

## 謝辞

本稿は国立国語研究所共同研究プロジェクト「開かれた共同構築環境による通時コーパスの拡張」による成果の一部である。

## 参考文献

- 青木伶子 (1973) 「接続詞および接続詞的語彙一覧」 鈴木一彦・林巨樹 (編) 『品詞別日本語文法講座 6 接続詞・感動詞』 明治書院, pp.210-253.
- 石田裕子 (2009) 「中央公論にみる接続表現の変遷」 『同大語彙研究』 11, pp.3-16.
- 市川孝 (1976) 「副用語」 『岩波講座日本語 6 文法 I』 岩波書店, pp.219-258.
- 小木曾智信・近藤明日子・高橋雄太・田中牧郎・間淵洋子編 (2023) 『昭和・平成書き言葉コーパス』 バージョン 2023.5, <https://clrd.ninjal.ac.jp/shc/>
- 京極興一・松井栄一 (1973) 「接続詞の変遷」 鈴木一彦・林巨樹 (編) 『品詞別日本語文法講座 6 接続詞・感動詞』 明治書院, pp.89-136.
- 国立国語研究所 (編) (2004) 『分類語彙表 増補改訂版』 大日本図書



- 国立国語研究所 (編) (2019) 『日本語歴史コーパス 明治・大正編 I 雑誌』 (短単位データ 1.2) [https://clrd.ninjal.ac.jp/chj/meiji\\_taisho.html#zasshi](https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#zasshi)
- 近藤明日子 (2021a) 「明治・大正期の書き言葉における文体と語彙の通時的変化—逆接の接続詞を例に—」 近代語学会 (編) 『近代語研究 第二十二集』, pp.47-67.
- 近藤明日子 (2021b) 「明治・大正期の書き言葉における文体と語彙の通時的変化—順接の接続詞を例に—」 田中牧郎・橋本行洋・小木曾智信 (編) 『コーパスによる日本語研究 近代編』 pp. 115-136.
- 近藤明日子 (2023) 「『昭和・平成書き言葉コーパス』 (バージョン 2023.5) 雑誌レジスター 概説書」 <https://clrd.ninjal.ac.jp/shc/doc/abstract-shc-zasshi-202305.pdf>
- 日本国語大辞典第二版編集委員会・小学館国語辞典編集部 (編) (2000-2002) 『日本国語大辞典 第二版』 小学館 (ネットアドバンス社提供サービス「ジャパンナレッジ」コンテンツ <https://japanknowledge.com/contents/nikkoku/> による)

## 実践医療用語\_語構成要素語彙試案表 Ver.3 の公開にむけて

相良かおる（奈良先端科学技術大学院大学）

黒田航（杏林大学）

東条佳奈（大阪大学）

西嶋佑太郎（京都大学）

麻子軒（関西大学）

山崎誠（国立国語研究所）

### Word Component Database for “Hands-On Medical Terms” (Version 3) Specification

Kaoru Sagara (Nara Institute of Science and Technology)

Kow Kuroda (Kyorin University)

Kana Tojo (Osaka University)

Yutaro Nishijima (Kyoto University)

Tzu-Hsuan Ma (Kansai University)

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

#### 要旨

医療用語の合成語の語構造および語構成要素とその意味を明らかにすることを目的に、合成語 7,087 語を分析し『実践医療用語\_語構成要素語彙試案表 Ver.2』を作成する過程で、(1)医療用語の選定方法、(2)分割単位の曖昧性、(3)語構造の記述方法、(4)意味ラベルの命名と付与方法に課題が見つかった。そこでこれらの課題を検討し、改良版の試案表 Ver.3 の作成に着手した。具体的には、(1)合成語の追加は今後の課題とし試案表 Ver.3 では行わない。(2)語分割が複数考えられる場合は、一つに絞るか分割せずに一つの語構成要素とする。(3)並列および不連続な語構造も記述できる複層化形態素解析 (Multi-Layered Morphological analysis : MLMA) を提案し、試案表 Ver.3 では、不連続な構造は原則として記述しないが、一部の並列構造を不連続な構造として記述できる場合は記述する。(4) 語構成要素の意味ラベルの集合と合成語の意味との関係がわかるように意味ラベルを合成語にも付与する。また意味ラベルが「#未定」の語構成要素の見直しを行うこととした。

#### 1. はじめに

医療記録データには、複数の語が連結された医療用語の合成語（以下、「医療合成語」という）が多く存在する。一方、施設内で蓄積される少ないデータで自然言語処理を行うためには、医療合成語の語構成やそれらの構成要素の意味の利用は有用である。しかし、医療記録は個人情報が含まれるため非公開であり、言語学的な調査があまり行われてこなかった。

そこで、我々は、(1)医療合成語の語構造を明らかにすること、(2)医療合成語を構成する語構成要素を抽出し、意味ラベルを付与すること、(3)医療合成語の日本語学または言語学的知見を得ること、(4)得られた知見を含む成果物を医療実践、医療教育の領域で、出来得れば言語学研究の領域においても利用可能な形で公開することを目的とし、『実践医療用語辞書 ComeJisyoSjis-1』より抽出した医療合成語 7,192 語を対象に 2020 年に『実践医療用語\_語構成要素語彙試案表 Ver.1』（以下、「試案表 Ver.1」という）を作成した。次いで 2021 年には、見直しの必要があった意味ラベルの妥当性（相良 2021）を踏まえて、医療合成語

や意味ラベルを見直し、『実践医療用語\_\_語構成要素語彙試案表 Ver.2』<sup>1</sup>（以下、「試案表 ver.2」）を公開した（東条ら 2022）。

医療合成語の語構造の解析の過程で、医療合成語には①同じ見出し語であっても数詞の表記法にゆれがあること、算用数字は、順序尺度が多く名義尺度が少ないこと、ローマ数字は名義尺度が多いこと（東条ら 2021）②「～性」を含む病名が多く、「先天性」「多発性」「急性」などが高頻度に用いられ、「急性細菌性髄膜炎」のように他の「～性」との共起も見られること（東条ら 2020）、③「急性」と「～性」の共起については、「急性」は語頭に多く出現し、医療現場では「急性」無しの同義の病名が多く使われていること、「急性」とその他の「～性」の語順を変えた同義語が存在すること（相良ら 2022）などの知見を得ている。また、語構造と意味に関しては、④《部位》を表す意味ラベル（例：身体部位）は「での」、《原因》を表す意味ラベル（例：病原体）は「による」、《状態》を表す意味ラベル（例：経過、症状）は「である」、《主体》を表す意味ラベル（例：患者属性）は「に生じた」を当該語構成要素の直後に加えることによって、機械的な言い換えが可能であることが明らかになっている（麻ら 2020）。

しかしながら、これらの研究結果は、日本語学的な成果というよりは、言語処理をする上で有益な成果といえる。また、医療現場では「急性」無しの病名が使われている（相良ら 2022）ということから、対象としている医療合成語は、「実践医療用語」と言えるのかという疑問も浮上している。

そこで本稿では、試案表 Ver.3 を日本語学的な分析にも利用可能な言語資源とするための問題点と解決方法について述べる。具体的には、先ず本研究で対象とする医療合成語と各試案表の概要を述べ、次いで、日本語学的知見を得る上での問題点と、試案表 Ver.3 に反映できる解決方法について述べる。

## 2. 医療合成語の概要

### 2.1 選定方法

以下に本医療合成語の選定方法を示す（相良ら 2019 再掲）。

#### 手順 1. 医療合成語候補 31,162 語の抽出

『実践医療用語辞書 ComeJisyoSjis-1』の見出し語の抽出に用いた医療記録は、それぞれ 1 施設分の医師経過記録、看護経過記録、多職種共有経過記録であり代表性はない。そこで、対象合成語のバイアスを減らすために Web 上で公開されている辞書など研究用に収集した医療用語データと本辞書の見出し語を照合し一致した 31,162 語を医療合成語候補とした。なお、これらには、英数字のみからなる語も含まれている。

#### 手順 2. 医療合成語に使われる一般的な用語 231 語の抽出

医療の知識がなくても語の境界の判定が容易になるように、一般的な日本語を含む医療合成語を対象にすることとし、医療合成語候補 31,162 語を MeCab 0.996<sup>2</sup> と Unidic-cwj-2.2.0<sup>3</sup>を用いて自動形態素解析を施し、以下の(a)～(d)に沿って語を抽出した。

- (a) 「名詞」または「形状詞」となる単一語
- (b) 二語以上の語で「名詞」+「接尾辞」、「形状詞+接尾辞」、「接頭辞」+「名詞」、

<sup>1</sup> <https://www.gsk.or.jp/catalog/gsk2020-g/>

<sup>2</sup> MeCab : <https://taku910.github.io/mecab/>

<sup>3</sup> 現代書き言葉 UniDic : [https://clrd.ninjal.ac.jp/unidic/back\\_number.html#unidic\\_bccwj](https://clrd.ninjal.ac.jp/unidic/back_number.html#unidic_bccwj)

漢字2字の「名詞」+「名詞」、「記号」のみの組み合わせであるもの

(c) 上記以外で文字数3文字以下のもの

(d) 自動形態素解析の結果、「感動詞」と誤解析された語を、目視で確認し妥当と認められたもの

次に(a)~(d)の規則に従ったものから「英数字」、「ひらがな」、「カタカナ」のみで表記された語を除いた3,728語について、2018年11月公開のComeJisyoUtf8-1の登録語75,831語における出現頻度を求め、頻度の降順上位25%の768語を抽出した。

医療合成語の語構成要素に意味ラベルを付与する上で、『分類語彙表一増補改訂版一』（以下、『分類語彙表』<sup>4</sup>）の分類項目の利用可能性を考え、768語と『分類語彙表』の見出し語を照合し、一致した231語を語構成要素候補とした。なお、ComeJisyoUtf8-2の見出し語114,957語に含まれる語を『分類語彙表』の分類番号で集計（意味分類）した結果については、山崎（2020）に詳述されている。

手順3. 一般的な語を含む医療合成語：7,192語

医療合成語候補31,162語の内、語構成要素に一般的な語231語を含む7,192語を試案表Ver.1の医療合成語とした。

## 2.2 出現頻度

表1と表2は、医療合成語候補31,162語から英数字のみの語および人名などの固有名詞を除いた30,881語（以下、「全医療合成語」という）および一般語を含む試案表Ver.1の医療合成語7,192語についてComeJisyoUtf8-2r1（見出し語数114,957語）<sup>5</sup>に記載の出処に関する情報をまとめたものである。表1は、5種類の文書における頻度を示したものであり、試案表Ver.1の医療合成語7,192語において文書頻度が0の語が57.36%占めている。

表1 文書頻度 (0~5) <sup>6</sup>

	計	5	4	3	2	1	0	未登録語
全医療合成語	30,881	583	1,401	5,897	3,553	4,320	15,104	23
	100%	1.89%	4.54%	19.10%	11.51%	13.99%	<b>48.91%</b>	0.07%
試案表 Ver.1	7,192	143	251	1,025	688	958	4,125	2
	100%	1.99%	3.49%	14.25%	9.57%	13.32%	<b>57.36%</b>	0.03%

※未登録語：ComeJisyoUtf8-2r1の見出し語にない語

表2は、電子カルテシステム導入の際に推奨されるMEDIS標準マスター<sup>7</sup>の内、病名マスター、手術・処置マスター、症状所見マスター、看護実践用語マスターの4分野と照合した

<sup>4</sup> 国立国語研究所：分類語彙表 増補改訂版，大日本図書，2004.

<sup>5</sup> <http://comedic.osdn.jp/> ※現在OSDN側の都合でダウンロードが出来ない場合がある。

<sup>6</sup> 実際の医療記録データ3種類と看護師教育で使われる看護記録データ、そして言語資源協会より公開されている模擬診療録テキスト・データの5種類における文書頻度(0~5)

<sup>7</sup> 一般財団法人医療情報システム開発センター(以下、MEDIS-DC)が厚生労働省から委託を受け、診療情報の「用語・コード」の標準化として作成したもので、以下の9分野の標準マスターがある。①病名マスター(ICD10対応標準病名マスター)、②手術・処置マスター、③臨床検査マスター、④医薬品HOTコードマスター、⑤医療機器マスター、⑥看護実践用語標準マスター、⑦症状所見マスター、⑧歯科分野マスター、⑨画像検査。なお、本稿では①②⑥⑦の4分野を扱っている。

結果の頻度を示したものである。試案表 Ver.1 の医療合成語の 85.55%が病名マスターに登録されている病名となっている。なお、試案表 Ver.2 の医療合成語 7,087 語における病名マスターの割合は 86.24%、文書頻度が 0 の割合は 57.99%である。

表 2 MEDIS 標準マスターの登録語数

	計	病名	手術・処置	症状所見	看護実践用語	マスター登録無	未登録語
全医療合成語	30,881	20,499	995	38	185	9,141	23
	100%	66.38%	3.22%	0.12%	0.60%	29.60%	0.07%
試案表 Ver.1	7,192	6,153	425	7	41	564	2
	100%	85.55%	5.91%	0.10%	0.57%	7.84%	0.03%

※未登録語：ComeJisyoUtf8-2r1 の見出し語にない語

表 3 は、全医療合成語の文書頻度と MEDIS 標準マスターの内、病名マスター、手術・処置マスター、症状所見マスター、看護実践用語マスターの 4 分野と照合した結果のクロス集計表である。全体の 66.43% (20,499 語) が病名マスターに登録の病名であり、その内 52.75% (10,814 語) が文書頻度 0 (全体に占める割合は 35.04%) である。

表 3 文書頻度と MEDIS 標準マスターのクロス集計

	計	病名	手術・処置	症状所見	看護実践用語	マスター登録無	
5	583	236	19	6	28	294	
4	1,401	653	41	6	52	649	
3	5,897	3,595	141	13	64	2,084	
2	3,553	2,255	115	5	17	1,161	
1	4,320	2,946	160	5	11	1,198	
0	15,104	10,814	519	3	13	3,755	
計	30,858	20,499	995	38	185	9,141	

### 3. 試案表の概要

表 4 と表 5 は、各試案表の概要をまとめたものである。試案表 Ver.1 と Ver.2 の変更点は、① 不要な対象医療合成語の削除、② 合成語の語構造情報の付与、③ 意味ラベルの見直し・統合である (東条ら 2022)。①については、異体字による重複と誤字および現在では使われていない語 105 語を削除した。②の語構造情報は、医療合成語に語の境界位置を [ ] で示したものである。

医療合成語：感染後甲状腺機能低下症

語構造：感染後 [甲状腺 [機能低下症]]

語構成要素：感染後、甲状腺、機能低下症、甲状腺機能低下症

語構成要素の認定は、長単位や短単位といった言語学や日本語学の観点の形態認定の単位とするのではなく、「医療の観点からみた有意味性を反映する単位 (以下、「医療の観点からみた単位」という)」とし、その判断は、医師、看護師と議論を交えて行っている。

以下に医療の観点からみた有意味性を反映する単位の例を示す。なお、医療従事者が分割



マスター登録の病名となっており、実際の医療記録に出現する合成語の殆どが病名であるとは考えにくく、偏りがあること、合成語の選定方法に問題があることは否めない。

#### 4.2 語の分割の曖昧性

言語学的な分析では、品詞などを考慮した機能的な単位で複合語を分割するのが一般的である。しかし、意味の理解なくして医療合成語を品詞単位で分割することは困難である。そして意味の解釈には曖昧性がある。同じ医学でも専門領域によって意味の解釈がことなる例として、「右上葉肺扁平上皮癌」を「概念が失われない最小の単位」に分割する際、臨床医学の領域では、「右 | 上 | 葉 | 肺 | 扁平上皮癌」となるが、解剖学の領域では「右 | 上 | 葉 | 肺 | 扁平上皮 | 癌」となることが指摘されている（劉 2000）。本研究でも「乳汁分泌抑制」の分割において、使用頻度の高い助産師は「乳汁分泌 | 抑制」と分割し、医師と看護師は「乳汁 | 分泌抑制」と分割した。なお、試案表では、医学大辞典の部分一致検索の結果<sup>8</sup>を根拠に、「乳汁 | 分泌抑制」としている。このように、語の意味の解釈には曖昧性があることから、意味による語分割において斉一性を保証することは困難である。

一方、意味が捨象される統計的自然言語処理や機械学習による語分割では、大量のデータが必要であり、秘匿性の高い医療記録データを大量に集めることは困難である。また、語分割された結果が有意味か否かの判断は人に委ねられるため、意味の解釈の曖昧性の問題に直面することになる。

本研究では、医療合成語を医療の観点から有意味性を反映する単位に分割し、語構成要素を抽出している。

医療合成語：一過性甲状腺機能低下症

語 構 造：一過性 [甲状腺 [機能低下症]]

語構成要素：一過性、甲状腺、甲状腺機能低下症、機能低下症

しかし「機能低下症」は、「機能低下」に有意味性があると判断すると「機能低下 [症]」となり、「低下症」に有意味性があると判断すると「機能 [低下症]」となる。このように医療の分野により判断が分かれ、語末が「症」や「障害」など前要素に意味を添える接尾語的な機能を持つ語の場合は、分割せずに一つの語構成要素としている。

その結果、病気を表す語末語「病」および「症」の前要素の相違または特徴を調べる際には、試案表 Ver.2 をそのまま利用することはできず、語構造の見直しが必要となる。

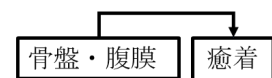
#### 4.3 語構造の記述方法

試案表 Ver.2 では、「・」で区切って列挙した語構成要素の後要素との関係は、以下に示すように2種類あり、医療の知識がないと、どちらのタイプかを判断するのは困難である。

医療合成語：骨盤腹膜癒着

語 構 造：[骨盤・腹膜] 癒着

意 味：骨盤と腹膜が癒着している状態



医療合成語：前後十字靭帯損傷

語 構 造：[[前・後] 十字靭帯] 損傷

意 味：前十字靭帯と後十字靭帯が損傷している状態



<sup>8</sup> 南山堂医学大辞典第20版の検索結果では、「分泌抑制」の「乳汁」以外の前要素として「成長ホルモン」「胃酸」「酸」がある。

#### 4.4 意味ラベル

試案表 Ver.1 では、一般的な日本語に意味ラベルを付与し、その数は 80 種類となった。しかし、「～性」となる語構成要素をはじめ、患者の状態全てに意味ラベル《状態》を付与した結果、意味ラベル《状態》が全語構成要素に占める割合は、22.24%となった。そこで Ver.2 では、医療の観点による意味ラベルの付与に注力し、意味ラベル《状態》を細分化し、一方で国語辞典<sup>9</sup>に見出し語のある語構成要素については《#未定》の意味ラベルを付与した。その結果、「サルコイドーシス性」は《病態》で、「マイコプラズマ性」は《病因》かつ《病原体》というように「～性」の前要素に関する知識がなくとも意味判断ができるようになった（東条ら 2022）。一方、一般的な日本語に《#未定》の意味ラベルを付与したことで、《#未定》が付与された語構成要素 141 語の全体に占める割合は、2.13%となっている。

複数の意味を持つ語構成要素には、複数の意味ラベルを無秩序に「,」で区切って列挙している。表 6 は、語構成要素に付与した意味ラベル数をまとめたものである。Ver.2 では、意味ラベル《状態》を細分化したことにより、複数の意味ラベルを付与した語構成要素が多くなっている。

表 6 語構成要素に付与した意味ラベル数

	1	2	3	4	計
Ver.1	5,603	745	32	0	6,380
Ver.2	4,289	2,171	138	35	6,633

多義の語構成要素の中には、連結する語構成要素により意味が一意に決まるものがある。

語構成要素：ウイルス  
意味ラベル：病因,病原体

語構成要素：ウイルス性  
意味ラベル：状態,病態,病因,病原体

「ウイルス」に付与した《病原体》と《病因》に一般的な日本語の意味での関係はない。そして後ろに「肝炎」がつく「ウイルス肝炎」での「ウイルス」の意味は《病因》となり、「肝炎」が前につく「肝炎ウイルス」での「ウイルス」は《病原体》となる。なお、「ウイルス肝炎」は「ウイルス性肝炎」の「性」が省略されたものと考えられる。

語構成要素：離開  
意味ラベル：状態

語構成要素：嚥下困難  
意味ラベル：状態,症状

語構成要素：嚥下性  
意味ラベル：状態,病態

語構成要素：呼吸困難  
意味ラベル：症状,状態,病態

一方、《症状》《病態》《状態》には、意味的な重なりがある。ものごとのありさまには《状

<sup>9</sup> 「岩波国語辞典第五版タグ付きコーパス 2004」の見出し語と照合



態》を付与し、「脱水」など病気により患者に現れた状態には《症状》を、「栄養不良」などの身体機能の状態には《病態》を付与している。

語構成要素に付与した意味ラベルは、前後の語構成要素により変わる場合がある。今後、多様な医療合成語を分析する中で、語構成要素に付与される意味ラベルの種類は増える可能性があり、語構成要素に付与した意味ラベルの組み合わせが変わる場合がある。

現在は複数の意味を持つ場合、無秩序に「,」で区切って意味ラベルを列挙しており、見かけ上意味ラベル間に包含関係があるか否の判断はできない。

#### 4.5 医療合成語の多様性

試案表の医療合成語の8割以上が病名となっている(表2)。しかし医療合成語には、病名だけでなく「残存機能」「皮膚脆弱」「局所圧迫」など、複数の一般的な日本語からなる合成語も多く、また、「冠動脈バイパス術」のような専門用語からなる合成語も多い。従って代表性を高めるために、医療分野での多様な文書に出現する医療合成語を追加する必要がある。

現在、非医療従事者が日本語学的な研究・調査を目的に医療施設で蓄積される実際の医療記録データを入手するのは困難である。しかしながら、公開されている標準マスター、例えば診療行為マスター<sup>10</sup>や診療ガイドライン<sup>11</sup>、厚生労働省が公開している国家試験問題文などは、非医療従事者であっても入手可能である。

### 5. 対処方法と試案表 Ver.3

本章では、(1)実践医療用語の選定方法、(2)語の分割単位の曖昧性、(3)語構造の記述、(4)意味分類についての対処方法を述べる。

#### 5.1 実践医療用語の選定方法

本研究に着手した当初は、全医療合成語 30,881 語の内、一般的な日本語を含む約 7 千の医療合成語の語構造分析と語構成要素の意味分類を終え最終版の試案表を公開した後に、残りの医療合成語の語構造分析と語構成要素の意味分類に着手する予定であった。

しかし今回の結果から、試案表 Ver.3 を公開した後に、新たに入手可能なデータから分析対象とする医療合成語を選定し分析を行う予定である。

#### 5.2 分割単位の曖昧性

試案表 Ver.3 において、「乳汁分泌 | 抑制」と「乳汁 | 分泌抑制」や「機能 | 低下症」「機能低下 | 症」のように分割方法が複数あった場合、医学事典や用語集などの情報を基に「乳汁 | 分泌抑制」にするなど、どちらか一つにするか、「機能低下症」のように語末が「症」や「障害」などの場合は分割せずに一つの語構成要素とする。

#### 5.3 語構造の記述

分割単位の変更は、当事者が行う必要があるが、語構成要素の抽出方法については、**並列分散形態素解析** (parallel Distributed Morphological Analysis : PDMA) を提案し、加えて PDMA より抽出精度は低いものの、記述が簡便な**複層化形態素解析** (Multi-Layered Morphological analysis : MLMA) を提案している。また、MLMA で語構造を記述したデータから語構成要素を自動抽出する Perl スクリプトを公開している (黒田ら 2023)。

語構造の記述において、並列構造、不連続な構造の記述は、MLMA の記法を用いること

<sup>10</sup> <https://www.ssk.or.jp/seikyushiharai/tensuhyo/kihonmasta/index.html>

<sup>11</sup> <https://minds.jcqh.or.jp/>

で記述することができ、また MLMA の記述から自動的に語構成要素を抽出することができる（黒田ら 2023）。

試案表 Ver.3 では MLMA を用い、不連続な構造は原則として記述しないが、並列構造を不連続な構造として記述できる場合は記述する。

#### 5.4 意味ラベル

現在、試案表 Ver.3 の作成に向けて試案表 Ver.2 の語構造を見直し、抽出した 12,336 の語構成要素に意味ラベルを付与している。その際、「#未定」としている一般的な意味を持つ語構成要素については、「行為」「状態」「物」などの大まかな分類方法を決め、該当する語構成要素に意味ラベルを付与する予定である。意味ラベルの列挙方法については、今回は定めない。

### 6. おわりに

我々は、医療合成語の語構造と意味を分析し、その結果を『実践医療用語—語構成要素語彙試案表』として公開している。本稿では、試案表を日本語学または言語学での調査研究で利用する上での問題を明らかにした。試案表 Ver.2 を改善した試案表 Ver.3 においても、対象医療合成語の代表性の問題、語分割の単位の曖昧性、意味ラベルが確定されていないなどの問題は解決できておらず、日本語学または言語学で試案表を変更することなく利用するのは難しい。

しかし、意味が捨象される機械学習の結果を解釈するのは人であることから、医療従事者が実践医療用語をどのように解釈し利用しているかを知る上で、「医療の観点からみた有意味性」を重視した医療合成語の分析は有用である。

以下は、医療合成語の語構造の分析過程で興味を持ったこと、日本語学的な問いをまとめたものである。これらの問いが、何の役に立つのかは分からない。しかし、以下の問いに面白さを感じ、試案表を改変し、または新たな言語資源の作成に着手し、医療用語の語彙研究に取り組む人が増えることを期待している。

- ① 語構成要素の語順を換えた場合に意味がことなる医療合成語（例：「ウイルス肝炎」と「肝炎ウイルス」、「腎嚢胞」と「嚢胞腎」）の語構成要素の特徴
- ② 語構成要素の語順を換えても意味が同じである医療合成語（例：「境界型糖尿病」と「糖尿病境界型」）の語形成の特徴。なぜ語順の交換が生まれるのか
- ③ 語構成要素の並列構造の意味の分析（例：「[前・後] 十字靭帯」、「[脾・腎] 静脈吻合術」、「[絞扼性・癒着性] 腸閉塞症」、「[歯状核・赤核・淡蒼球・ルイ体] 萎縮症」）
- ④ 語の重複のある医療合成語の分析（例：[乳汁<分泌] 抑制>、先天性 [中枢<神経] 異常>）
- ⑤ 医療の観点からみた一つの語構成要素を更に有意味な短い単位に分割した要素の意味素性の集まりと語構成要素の意味との関係（例：「[気管支肺胞洗浄] 異常」：「気管支肺胞洗浄」は検査法で、気管支から肺胞までを洗浄した排泄液に異常があるという意味）

### 謝 辞

本研究は JSPS 科研費 JP18H03499 ならびに JP21H03777 の助成を受けている。

## 文 献

- 相良かおる（2021）「実践医療用語における語構成要素の意味ラベルについて」『言語処理学会第27回年次大会発表論文集』 pp.559-562.
- 東条佳奈、黒田航、相良かおる、高崎智子、西嶋佑太郎、麻子軒、山崎誠（2022）「実践医療用語\_語構成要素語彙試案表 Ver.2 の構築」『言語資源ワークショップ 2022 発表論文集』 pp.2-6.
- 東条佳奈、相良かおる、西嶋佑太郎、麻子軒、山崎誠（2021）「医療用語に含まれる序数詞について」『じんもんこん 2021 論文集』 pp.194-199.
- 東条佳奈、麻子軒、相良かおる、高崎智子、山崎誠（2020）「病名における「～性」の分析：一般書籍との比較から」『言語資源活用ワークショップ 2020 発表論文集』 pp.357-364.
- 相良かおる、高崎智子、東条佳奈、西嶋佑太郎、山崎誠（2022）「『急性』を含む病名の語構成」『言語資源ワークショップ 2022 発表論文集』 pp. 43-51.
- 麻子軒、相良かおる、高崎智子、東条佳奈、山崎誠（2020）「意味ラベルを用いた『-性』を含む病名の言い換え」『じんもんこん 2020 論文集』 pp.283-288.
- 相良かおる、山崎誠、麻子軒、東条佳奈、小野正子、内山清子（2019）「実践医療用語の語構成要素意味を基準とした分割」『じんもんこん 2019 論文集』 pp.57-64.
- 山崎誠（2020）「実践医療用語を構成する語の計量的分析」『言語資源活用ワークショップ 2020 発表論文集』 pp. 164-173.
- 劉亜斌、里村洋一、佐々木哲明、木村通男、廣瀬康行、山崎俊司（2000）「構造化臨床医学用語集の構築に関する研究」『医療情報学』 20 巻 6 号 pp.513-522.
- 黒田航、相良かおる、東条佳奈、麻子軒、西嶋佑太郎、山崎誠（2023）「要素の重複と不連続性を扱える抽出型の語構成要素解析：並列分散型形態素解析の提案」『言語処理学会第29回年次大会発表論文集』 pp. 772-777.

# 科学技術系ライティング教育改善を目的とした シラバス分析のための spaCy-GSDLUW を利用した日本語長単位解析

堀 一成, (大阪大学)\*

## Long-unit-word morphological analysis on Japanese university syllabus data using spaCy-GSDLUW for providing the learning contents for science academic writing

HORI Kazunari, (Osaka University)

### 要旨

日本語アカデミック・ライティング科目のシラバスデータを新しい形態素解析ソフトウェア spaCy\_GSDLUW で長単位形態素解析することの有用性が明らかになった。発表者らは、大学学部初年次生や高校生を対象とする科学技術系日本語アカデミック・ライティング指導を改善するため、日本語アカデミック・ライティング科目のシラバスデータを収集し、内容の分析を試みている。その際、シラバスデータの言語特徴を良く抽出するため、長単位形態素解析を行うことが有効ではないかと考えている。2022年に公表された長単位解析ソフトウェア spaCy-GSDLUW は、導入作業も容易で、分析の精度もこれまでのシステムより高いものである。発表者が開講する少数の科目シラバスのみを対象にした試行結果であるが、教育関係情報の言語処理を進めるにあたり、まず長単位形態素解析ことが有用であろうとの目処がついた。試行の方法や得られた結果に対する考察を紹介する。

### 1. 発表の背景

発表者らは、大学学部初年次生や探究学習に取り組む高校生を対象に、日本語アカデミック・ライティング指導を継続している。また、その指導内容が順次改善されるよう、様々な試みを行ってきた。

大学学部初年次生を対象とした独自の日本語アカデミック・ライティング指導教材を作成し、(所属する大阪大学だけでなく) 広く学部初年次生が自由に利用できるようデータ公開している(堀・坂尻(2020), 堀・坂尻(2015))。また科学実験レポート作成に特化した理系学部初年次対象のガイドブック(堀ほか(2022))なども上梓した。

このような指導を行うにあたり、指導内容の根拠となる情報がコーパス分析の結果から得られないかと考え、科学技術分野の学術文・技術文の解析を行った(堀ほか(2016))。その際、学

---

\* hori.kazunari.celas[アットマーク]osaka-u.ac.jp

術文に特有な言語表現を抽出するためには、長単位に基づく形態素解析を行うべきだと判断し、実行し一定の成果を得た。

富士池ら(富士池ほか(2008))によると、長単位とは、文節の内部を自立語部分と付属語部分に分解することで認定される区切りである。長単位は資料の特徴語を取り出せることが利点であるとしている。対して、短単位は基準がわかりやすく、ゆれが少ないが、合成語を構成要素に分割してしまう問題点があるとしている。

一方、2016年より、発表者の所属する研究グループは、大学・大学院に在籍する日本語学習者による読解・ライティングの学習方法や文章観とその背景を、探索的調査から明らかにすることに取り組んでいる。その取り組みの一つとして(日本語非母語話者に対するものを含む)多くの大学で開講されている日本語アカデミック・ライティング指導科目のシラバスデータを収集分析し、ライティング教育の現況を把握する作業を進めている。グループの先行して研究を進めていた研究者は、導入が容易で操作も簡便な KH-Coder を主な分析ツールとして利用していた。KH-Coder は、形態素解析を MeCab か Chasen で実行する設定であり、その後の様々な解析は短単位形態素に基づくものになっている。発表者としては、シラバスに含有されているライティング教育の特徴をよく表す語彙は、長単位形態素に基づくものであろうと予想した。しかし、長単位形態素解析のソフトウェアは普段 Linux など UNIX 系システムを使い慣れないものには扱いが難しいところがあり、積極的な導入進言に逡巡する状況であった。

2022年に、後述する新しい新しい長単位形態素解析ソフトウェア spaCy\_GSDLUW が発表され、これを活用することでシラバスデータ分析の高度化が図れるのではないかとの発想に至った。

本発表は、取り組んでいるシラバスデータ分析といった教育関係情報の言語分析をする際(特に科学技術系ライティング教育の改善に資するような情報が得たい)に、長単位形態素解析システム spaCy\_GSDLUW を利用することが有用であるか、目処をつけるための試行の結果を報告するものである。

### 1.1 長単位解析ソフトウェア spaCy-GSDLUW

国立国語研究所と株式会社リクルート Megagon Labs が共同開発し、2022年に一般に公開された新しい長単位形態素解析ソフトウェア spaCy\_GSDLUW(松田ほか(2022))は、Universal Dependencies に基づく依存構造解析モデルに基づくものであり、従来手法を上回る長単位品詞推定精度を誇るものである。

また、Python (バージョン 3.10 が推奨されている) と pip システムが動作できるようであれば、pip コマンドを利用し、一行コマンドを入力するだけで GitHub のリポジトリからダウンロード・インストールが完了するため、導入が比較的容易であるところも特徴である。

(注意：本稿作成時点の 2023 年 8 月において、spaCy\_GSDLUW は Python3.11 以降のバージョンでは正常動作しない)

## 2. 頻度リストの作成方法

以下にシラバス分析長単位形態素の頻度リストを作成した手順を説明する。作業は Mac Mini (2020)(Intel Core i7-8700B 3.2GHz) 上で行った。OS は macOS Ventura 13.5 である。

### 1. 試行対象シラバス平文データの準備

試行分析対象として、発表者（堀）が大阪大学で 2023 年度に開講しているセミナー型アカデミック・ライティング指導科目 2 科目の日本語シラバスを用いた。一つは学部初年次生対象の「学術的文章の作法」であり、もう一つは大学院生対象の「学術的文章の作法とその指導」である。大阪大学の教学システム KOAN 上で提供している両科目のシラバスデータをダウンロードし、結合して改めて UTF-8 テキストデータとした。このデータを Microsoft 365 Word の文字カウント機能でカウントした、スペースを含めない文字数は、5,254 字である。

本来、形態素解析がより有効に行えるよう、記号・数字・本文内容に関係ない文字列などを取り除く前処理をすべきであるが、今回は作業省力化のため、このような処理を行っていない。

### 2. spaCy\_GSDLUW と MeCab を用いた形態素解析作業

上記作業により得た UTF-8 テキストデータを Python 3.10.10 上で spaCy\_GSDLUW を利用し長単位形態素リストを得る処理をした。入力したテキストファイルを、1 行ずつ spaCY\_GSDLUW ライブラリに渡し、長単位形態素解析したリスト情報を作成し、標準出力に出力する Python プログラムを作成し、実行した。同様の処理を MeCab(Unidic 辞書) を用いて短単位形態素リスト情報を得る Python プログラムを作成し、実行した。

### 3. Python による頻度情報抽出作業と Excel によるデータ整理

得られた spaCy\_GSDLUW の出力である長単位形態素情報付与済データから、各行の形態素原形情報が出力される第 3 列のデータのみを抜き出し、その頻度を計算する Python プログラムを作成し、実行することにより（試行対象であるシラバスデータの）長単位形態素頻度データを得た。

まったく同様の操作を MeCab の処理対象についても行い、短単位形態素頻度データを得た。

最後に、作業（堀）が目視判断により、出力形態素を一般名詞と一般動詞に分け、さらに短単位では分解され単独形態素として抽出できないであろう「漢語+する」動詞を分離し、別表データとする作業を行った。これらのデータ整理作業は Microsoft 365 Excel 上で手作業で行った。

このようにして得られたシラバス試行分析形態素頻度を、表 1 から表 5 に示す。

## 3. 得られた頻度リストに対する考察

得られた頻度リストに対して考察する。

印象的説明になるが、表 1 と表 2 に示した動詞の長単位・短単位を見比べると、長単位解析

により、「抽象名詞（多くの場合漢語）＋する」動詞を含む、より言語教育に関連した？ 表現が抽出できているといえる。特に短単位で「為る」の頻度が非常に高いことは、上記「～する」動詞が分解されてしまって、一単語として抽出できていないことを示唆している。特に短単位では頻度リストに出てこず、長単位では出てきた「漢語＋する」動詞を別表である表5に提示した。

同様に、表3と表4に示した名詞の長単位・短単位を見比べると、長単位解析により、「担当教員」「教室対面」「5パラグラフ論証文」など、よりシラバスらしい？ アカデミック・ライティング教育らしい？ 表現が抽出できているといえる。短単位名詞頻度表の最後に「オン」とあるが、長単位名詞頻度表と比べると、これは「オンデマンド」が分解されてしまったものであることがわかる。

このことから、シラバスのような教育関連情報を言語分析するにあたり、その後どのような統計処理をする場合であっても、第一処理段階の形態素解析の処理を長単位に基づいたものとするのは、（短単位処理したデータを基に処理進行することと比べ）おそらく有用な選択であろうということを示唆するものと、発表者は考える。

#### 4. 今後の展開

本発表は、シラバスデータの言語分析の第一処理に、長単位に基づく形態素解析が（短単位に基づくものに比べ）有用であるか、目処をつけるための試行報告であった。

##### ◎ 解析対象日本語シラバスデータの継続収集とコーパス化

本報告の形態素解析対象データは、大阪大学に所属する発表者（堀）個人の開講科目シラバスデータに限ったものであった。本来の目的である、多様なアカデミック・ライティング教育の現状をシラバス分析により把握し、改善の参考データとするためには、できる限り広範なシラバスデータを収集し、分析を進める必要がある。

##### ◎ シラバスデータからの有用な教育改善情報抽出方法の研究

今回は、少数のシラバスデータを長単位形態素解析して頻度情報を得るという、簡易な作業を行っただけである。今後アカデミック・ライティング教育の改善に資する情報を得るためには（長単位形態素解析したデータを出発点とし）どのような統計処理・AI利用分析などをすればよいか、有効な分析法の研究を進める必要がある。

2023年8月の発表時点において、発表者らが所属する研究チームにおいて、このようなデータ収集法、有効な分析法について議論を継続している。

#### 5. おわりに

以上のように、日本語アカデミック・ライティング科目のシラバスデータを spaCy\_GSD-LUW で長単位形態素解析することにより、有用な情報を得ることができるよう目処がついたといえる。2022年に公表された新しい長単位解析システム spaCy-GSDLUW は、導入作業も容易で、分析の精度も従来の長単位解析システムより高いものである。

本発表の事例を参考にされ、多くの研究で長単位形態素解析が利用されることを望むものである。

## 謝 辞

本発表の研究は、科学研究費基盤研究 (C) 課題番号: 20K03251 「『ダメな科学ライティング』をさせないための高大接続による探究学習教育法の研究」(研究代表者: 堀一成)、および基盤研究 (B) 課題番号: 19H01269 「日本語読解・ライティングの方法に影響する母語・母文化の教育的背景要因に関する研究」(研究代表者: 村岡貴子) による補助を受け推進しているものである。

本発表は、松田 寛氏を代表とする株式会社リクルート Megagon Labs の研究グループと国立国語研究所言語資源開発センターの共同成果である spaCy\_GSDLUW に依存したものである。有用なソフトウェアの開発と公開に対して深く謝意を表したい。

また、日本語非母語話者も対象に含む日本語アカデミック・ライティング教育について、シラバス分析の方針などについて、有用な議論を継続していただいている北九州市立大学 基盤教育センターの池田 隆介先生、大阪大学 国際教育交流センターの村岡 貴子先生にも感謝を表明する。

## 文 献

- 堀一成・坂尻彰宏 (2020). 『阪大生のための アカデミック・ライティング入門 第4版』 大阪大学 全学教育推進機構 <http://hdl.handle.net/11094/71454> から自由に PDF ファイルをダウンロードできる
- 堀一成・坂尻彰宏 (2015). 「大阪大学におけるアカデミック・ライティング教育の実践と教材作成」 大阪大学高等教育研究 Vol.3, pp. 27-32.
- 堀一成・北沢美帆・山下英里華 (2022). 『ダメ例から学ぶ 実験レポートをうまくはやく書けるガイドブック』 羊土社.
- 堀一成・坂尻彰宏・石島悌 (2016). 「ライティング教材作成を目指した日本語学術文長単位解析の試行」 言語処理学会 第22回年次大会発表論文集, pp. 685-688.
- 富士池優美・小椋秀樹・小木曾智信・小磯花絵・内元清貴・相馬さつき・中村壮範 (2008). 「現代日本語書き言葉均衡コーパス」の長単位認定基準について」 言語処理学会 第14回年次大会発表論文集, pp. 931-934.
- 松田寛・大村舞・浅原正幸 (2022). 「UD Japanese に基づく国語研長単位解析系の構築」 言語処理学会 第28回年次大会発表論文集, pp. 725-730.

## 関連 URL

長単位形態素解析ソフトウェア 『spaCy \_ GSDLUW』  
[https://github.com/megagonlabs/UD\\_Japanese-GSD/releases/tag/r2.9-NE/](https://github.com/megagonlabs/UD_Japanese-GSD/releases/tag/r2.9-NE/)



表1 堀シラバスデータから spaCy\_GSDLUW を用いて抽出した長単位動詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

長単位動詞	長単位頻度
成る	14
為る	13
行う	13
作成する	7
書く	7
学習する	7
ディスカッションする	6
求める	5
有る	4
学ぶ	4
ことができる	4
作る	4
応じる	3
得る	3
欠席する	3
助ける	3
言う	2
纏まる	2
出す	2
利用する	2
注意する	2
心掛ける	2
止む	2
決まる	2
相談する	2
取り止める	2
富む	2
習得する	2
繋がる	2
出来る	1

表2 堀シラバスデータから MeCab を用いて抽出した短単位動詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

短単位動詞	短単位頻度
為る	108
成る	16
行う	13
居る	10
有る	9
出来る	9
行く	7
書く	7
頂く	6
求める	5
学ぶ	4
貰う	4
作る	4
答える	4
付ける	3
言う	3
応ずる	3
得る	3
助ける	3
因る	3
見る	2
纏まる	2
出す	2
心掛ける	2
決まる	2
取り止める	2
繋がる	2
探す	1
読む	1
立てる	1

表3 堀シラバスデータから spaCy\_GSDLUW を用いて抽出した長単位名詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

長単位名詞	長単位頻度
題目	27
授業	23
課題	20
担当教員	15
教室対面	11
必要	10
受講者	9
受講生	9
場合	8
オンデマンド	8
曜日	8
明確	6
レポート	6
書き手	6
実際	5
説明	5
私語り	5
本授業	5
文章	5
説得力	4
論文・レポート	4
方法	4
根拠	4
身	4
技術	4
論理的	4
実習	4
5パラグラフ論証文	4
プレゼンテーション	4
分量	4

表4 堀シラバスデータから MeCab を用いて抽出した短単位名詞頻度表 (頻度上位 30 語まで) (堀一成 作成)

短単位名詞	短単位頻度
授業	50
事	35
課題	32
題目	30
受講	23
教員	22
的	20
レポート	18
作成	18
様	17
提出	16
曜日	16
担当	16
対面	15
文章	14
教室	13
ディスカッション	13
者	12
パラグラフ	12
必要	11
実習	11
プレゼンテーション	10
学習	10
大学	9
学術	9
論文	9
情報	9
評価	9
元	9
オン	9

表5 堀シラバスデータから spaCy\_GSDLUW を用いて抽出した長単位「漢語+する」動詞表 (全18語)(堀一成 作成)

作成する
学習する
欠席する
利用する
注意する
相談する
習得する
入手する
収集する
読解する
説明する
経験する
用意する
出席する
修得する
就職する
修正する
実現する

## 家庭での食事場面における親子会話の脱文脈度の観点からの分析

田中 弥生 (国立国語研究所 研究系) \*

江口 典子 (国立国語研究所 研究系)

小磯 花絵 (国立国語研究所 研究系)

### Analysis of Parent-Child Conversation During a Meal Situation at Home from the Perspective of Decontextualization

Yayoi TANAKA (National Institute for Japanese Language and Linguistics)

Noriko EGUCHI (National Institute for Japanese Language and Linguistics)

Hanae KOISO (National Institute for Japanese Language and Linguistics)

#### 要旨

本研究は、修辞機能分析の分類法による日常会話分析の一環として、家庭での親子の談話について修辞機能を確認し、脱文脈度の観点から検討するものである。本研究では、「修辞機能」を「話し手書き手が発信する際に、言及する対象である事態や事物や人物等を捉え表現する様態を分類し概念化したもの」と定義し、脱文脈度は「発話がコミュニケーションの場「いま・ここ・わたし」にどの程度依存しているか」の程度を表す概念とする。本発表では、現在構築中の『子ども版日本語日常会話コーパス』の内部公開データから、家庭での食事場面の親子の談話を対象に、食事についての話題と食事以外の話題での修辞機能の出現、および、食事以外の話題内容の特徴と脱文脈度の関連を確認した。また、子どもの年齢と脱文脈度の関連が確認された。

#### 1. はじめに

本研究は、脱文脈度の観点からのコミュニケーション分析の一貫である。子供の言語コミュニケーションに関して、語彙の理解や、発話構造、語用論的知識の発達など、さまざまな観点から研究が行われている。また、児童の発達や教室談話に関して、「脱文脈化」は議論されている(岩田ほか 1995, 橋本 2009)。教育心理学の分野では、「脱文脈化」を「思い出に相当するエピソード記憶 (episodic memory) を、知識に相当する意味記憶 (semantic memory) に転換する過程」(漁田・漁田 1999:205)としている。また、話し言葉と書き言葉を対比させ、成長とともに論じられることも多い。岩田ほか (1995) は、話し言葉は「相手の表情や身振り、発話のイントネーション、発話文脈といった手がかりに依存しながら日常のなかで自然と学ばれていく」のに対して、読み書き言葉は「対面的な文脈に依存することなく、言語の純粋に形式的な側面のみによっている」ため、「脱文脈的」であるとしている。

---

\* yayoi@ninja.ac.jp

この脱文脈化の程度を可視化できる手法が、「修辞機能分析」の分類法である。本発表では、「修辞機能分析」の分類法によって、家庭での食事場面での会話の脱文脈化の様相を可視化し、親子間のコミュニケーションにおける特徴を脱文脈度の観点から検討する。

本研究において修辞機能とは、「話し手書き手が発信する際に、言及する対象である事態や事物、人物等を捉え表現する様態を分類し概念化したもの」と定義する。また、脱文脈度は「発話がコミュニケーションの場「いま・ここ・わたし」にどの程度依存しているか」の程度を表す概念とする。

修辞機能分析の分類法は、選択体系機能言語理論の英語談話分析手法 Rhetorical Unit Analysis(Cloran 1994, 1999) を日本語に適用した修辞ユニット分析(佐野 2010, 佐野・小磯 2011) を元に、日本語文法の枠組みで修正を加えたものである(田中 2022)。テキストの分析単位(概ね、日本語文法の節に相当)ごとに、述部の時制と、主語・主題の話者からの距離の分類から修辞機能が特定され、あわせて脱文脈化指数が特定される。これによって、一般的な内容か個人的な内容か、発話の時空に依存しない内容か依存する内容か、抽象的なことか具体的なことか、などを示すことができる。例えば、子供はまず目の前の「いま・ここ・わたし」に近いことについて話せるようになり、成長とともに過去や、明日のこと、その場にはいないおばあちゃんのことなど、時空を離れた発話ができるようになる。「それ、ちょうだい」と同じ時空にいる親に言うのは文脈化しており、「はやぶさは北海道新幹線だよ」と話すのは同じ時空にいらなくても伝えられるため脱文脈度が高い。

これまで、児童作文、家族の談話、職場の談話、高齢者グループの談話など(田中ほか 2021, 佐尾ほか 2023, 田中・小磯 2019, 田中 2017, 田中ほか 2022, 2023)の分析から、目的や話題内容、状況によって用いられる修辞機能が異なり、脱文脈度は推移することなどが明らかになっている。田中・小磯(2019)では、修辞機能と脱文脈度の観点からの幼児の談話分析のケーススタディとして、家庭での食事場面における幼児と両親との会話の分析を行なっている。その結果、食事場面においては基本的に脱文脈度の低い発話が交わされ、0～4歳の女児の発話の脱文脈度が高くなるのは、両親の発話をきっかけとするものであることが明らかになっている。本発表の目的は、親子の会話における話題内容及び年齢と、脱文脈度の関連を明らかにすることである。現在国立国語研究所にて構築中の『子ども版日本語日常会話コーパス』に収録される予定の、家庭での親子の談話を分析対象とし、言語表現から修辞機能と脱文脈化指数を特定し、話題内容によってどのように修辞機能が用いられ、脱文脈度が表れているかを検討する。以下、第2節で分析データと分析方法について説明し、第3節で分析結果、第4節で考察を述べ、第5節でまとめと今後の課題について述べる。

## 2. 分析データと分析方法

### 2.1 分析データ

本研究の分析対象として、現在国立国語研究所にて構築中の『子ども版日本語日常会話コーパス』(CEJC-Child) プロジェクト内部公開データの、夕食時の親子の会話(Y006\_\_010a 17分24秒、参加者5名)の書き起こしデータを用いる。子供は3人で、年齢は長女11歳、長男

7歳3ヶ月、次女2歳3ヶ月である。書き起こしデータの個人名はすべて仮名である。

## 2.2 分析方法

修辞機能分析の分類法の手順は次のとおりである。

1. 分析単位（メッセージ）に分割し、分析対象を特定する。
2. 分析対象のメッセージについて発話機能を分類する。
3. 発話機能が「命題」のメッセージについて、時間要素と空間要素を分類する。
4. 発話機能・時間要素・空間要素の組み合わせから、修辞機能と脱文脈度を特定する。

以下に手順の概要を示す。

### 2.2.1 分析単位の分割と分析対象の特定

分析単位であるメッセージは概ね節に相当するが、連体修飾節は独立したメッセージとして扱わない。メッセージは「定型句類」（相槌、挨拶、定型句、節の形でないものなど）「主節」（単文、及び主節）「並列」（従属度の低い従属節）「従属」（従属度の高い従属節）「引用」（“と思う”などで引用されている部分）に分類する。「主節」「並列」「引用」についてこのあとの分類を行う。

### 2.2.2 発話機能・時間要素・空間要素

メッセージの種類が「主節」「並列」「引用」に分類されたメッセージについて、発話機能・時間要素・空間要素を分類する。表1に示したように、これらの組み合わせから修辞機能と脱文脈化指数が特定される。【行動】[1]がもっとも文脈に依存した表現で、【一般化】[14]がもっとも脱文脈度の高い表現である<sup>(1)</sup>。

表1 発話機能・時間要素・空間要素からの修辞機能と脱文脈化指数の特定

定義	↑ 高↑空間的距離のレベル ↓ 低						一般化 [14]
状況外	報告 [9]	状況外回想 [10]	予測 [11]		推量 [12]	説明 [13]	
状況内	行動 [1]	実況 [2]	状況内回想 [3]	状況内予想 [5]		観測 [8]	
参加				計画 [4]	状況内推測 [6]		自己記述 [7]
空間要素	← 低 ← 時間的距離のレベル → 高						
時間要素	現在	過去	未来意志的	未来非意志的	仮定	習慣・恒久	
発話機能	提言	命題					

発話機能は「提言」か「命題」に分類する。「提言」は、品物・行為の交換に関する提供・命令で、基本的には同じ時空に存在する相手に働きかけたり、会話者同士の行為にかかわる発話内容が該当し、【行動】[01]と特定される。例えば、同じ時空にいる相手への「この花見て!」「お醤油を取って」のような行為や物を要求する場合である。「命題」は、情報を交換する陳述・質問で、「私はトマトが大好き」「このトマトは真っ赤だね」「トマトはナス科の植物だ」などが該当する。発話機能が「命題」のメッセージについて、このあと時間要素と空間要素を認定する。

<sup>(1)</sup> 以下、修辞機能を【】で、脱文脈化指数を[]で示す



時間要素は、話者のいる時間「いま」を基準として、メッセージで表現されている出来事がいつ起こったかを示す要素である。基本的にテンスや時間を表す副詞などによって表現され、「習慣・恒久」<sup>(2)</sup>「現在」「過去」「未来意志的」「未来非意志的」「仮定」に分類する。

空間要素は、話者のいる場所「ここ・わたし」を基準として、メッセージの中心となる語句との空間的距離を示す要素で、主語、主題、述部の主体から判断し、「参加」「状況内」「状況外」「定義」に分類する。

### 2.2.3 修辞機能と脱文脈化度の特定

表1を参照し、発話機能、時間要素、空間要素の組み合わせから、修辞機能と脱文脈度を特定する。

## 3. 分析結果

### 3.1 修辞機能の出現

話者ごとの修辞機能の出現の様子を表2と図1に示す。「定形句類」は、修辞機能ではないが、次女の喃語や不明瞭な発話を定形句類に分類しているため、比較のためにここに示す。ただし、以降の分析には、「定形句類」は含まない。

表2 話者ごとの修辞機能の出現頻度

	定形句類	01 行動	02 実況	03 状況内回想	04 計画	05 状況内予想	06 状況内推測	07 自己記述	08 観測	09 報告	10 状況外回想	11 予測	12 推量	13 説明	14 一般化
父	94	20	97	8	0	3	0	6	15	6	2	3	0	26	1
母	4	4	5	0	0	0	0	0	0	0	0	0	0	0	0
長女	39	4	29	3	0	0	0	8	13	2	0	0	0	11	0
長男	44	1	25	2	0	0	0	5	6	4	1	0	0	23	1
次女	128	10	11	0	0	0	0	0	0	0	0	0	0	0	0

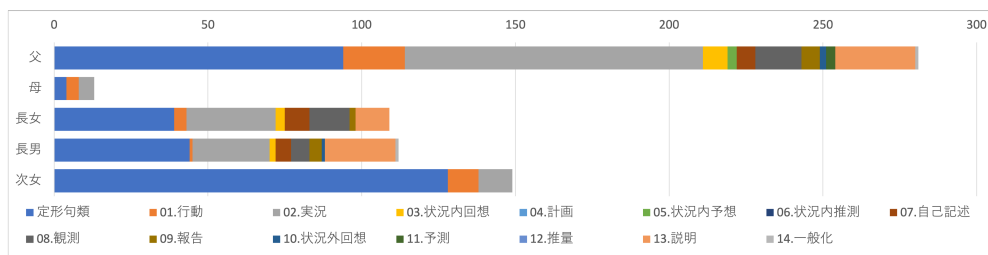


図1 話者ごとの修辞機能の出現割合

表2、図1から、次女の発話は定形句類に分類されるものがほとんどで、分析可能な発話は少なく、また脱文脈度が低いこと、別室にいる母の発話は少なく、脱文脈度が低いこと、父の発話は多く、定形句類と【実況】[02]が多いこと、脱文脈度の高い発話は、父と長女、長男によるものであることなどがわかる。

<sup>(2)</sup> 「習慣・恒久」には、属性、嗜好、評価も含む。

### 3.2 話題内容による特徴

分析対象の談話は、その話題内容から、大まかに食事に関する発話と食事以外の発話に分けることができる。表3と図2に食事と食事以外の発話における修辞機能の出現頻度と割合を示す。

表3 食事・食事以外の話題の発話における修辞機能の出現頻度

	01 行動	02 実況	03 状況内 回想	04 計画	05 状況内 予想	06 状況内 推測	07 自己 記述	08 観測	09 報告	10 状況外 回想	11 予測	12 推量	13 説明	14 一般化
食事	32	108	4	0	3	0	0	13	0	0	0	0	0	0
食事以外	7	59	9	0	0	0	19	21	12	3	3	0	60	2

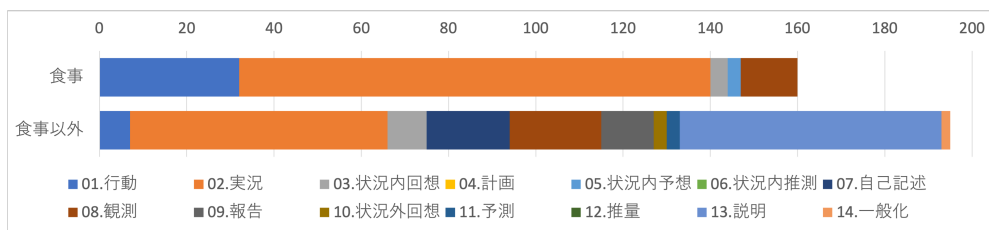


図2 食事・食事以外の話題の発話における修辞機能の出現割合

全体では【実況】[02]が多く、特に食事については脱文脈化指数[08]より低い脱文脈度のみが用いられている。脱文脈化指数[08]以下は、表1に示したように、空間要素が「参加」あるいは「状況内」であるため、食事に関する話題では、その場に関わることのみが話されていることがわかる。

食事以外の話題の発話でも【実況】[02]が多いが、ついで、【報告】[09]【説明】[13]という脱文脈度の高い発話も多い。表1に示したように、【報告】[09]、【説明】[13]は、空間要素が「状況外」で、「いま・ここ・わたし」から離れているものである。

表4と図3に、食事以外の話題の発話の修辞機能の出現頻度と割合を話題別に示した。

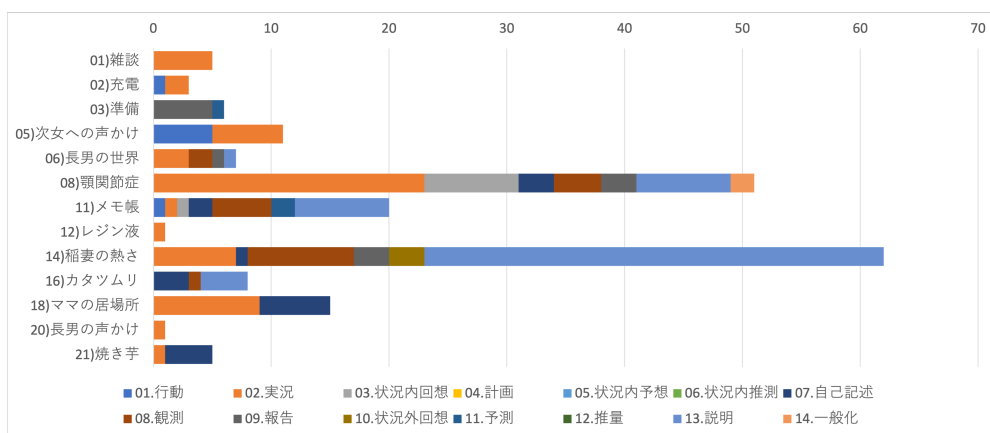


図3 食事以外の話題別の修辞機能の出現割合

表4 食事以外の話題別の修辞機能の出現頻度

	01 行動	02 実況	03 状況内 回想	04 計画	05 状況内 予想	06 状況内 推測	07 自己 記述	08 観測	09 報告	10 状況外 回想	11 予測	12 推量	13 説明	14 一般化
01) 雑談	0	5	0	0	0	0	0	0	0	0	0	0	0	0
02) 充電	1	2	0	0	0	0	0	0	0	0	0	0	0	0
03) 準備	0	0	0	0	0	0	0	0	5	0	1	0	0	0
05) 次女への声かけ	5	6	0	0	0	0	0	0	0	0	0	0	0	0
06) 長男の世界	0	3	0	0	0	0	0	2	1	0	0	0	1	0
08) 顎関節症	0	23	8	0	0	0	3	4	3	0	0	0	8	2
11) メモ帳	1	1	1	0	0	0	2	5	0	0	2	0	8	0
12) レジン液	0	1	0	0	0	0	0	0	0	0	0	0	0	0
14) 稲妻の熱さ	0	7	0	0	0	0	1	9	3	3	0	0	39	0
16) カタツムリ	0	0	0	0	0	0	3	1	0	0	0	0	4	0
18) ママの居場所	0	9	0	0	0	0	6	0	0	0	0	0	0	0
20) 長男の声かけ	0	1	0	0	0	0	0	0	0	0	0	0	0	0
21) 焼き芋	0	1	0	0	0	0	4	0	0	0	0	0	0	0

食事以外の話題でも脱文脈度の低い発話が多いことから、家庭の食事場面では脱文脈度の低い発話が基本であることがうかがえる。食事以外の話題には、長男が話しかけたが家族の誰も応答せず続かなかった話題なども含まれている。ここでは、まずそれぞれの話題の冒頭の発話を確認し、次に、まとまった発話が見られ、脱文脈度の高い発話を含む「顎関節症」の話題と、「稲妻の熱さ」の話題の2つについて、確認していく。

### 3.2.1 各話題の冒頭の発話

食事以外の各話題の冒頭の話題が変わる時の1～2発話を表5に示す。

表5 食事以外の話題の冒頭の発話と修辞機能・脱文脈化指数

01) 雑談	長女	うっ。	
01) 雑談	次女	ママは？	【実況】 [02]
02) 充電	長男	百パーセント。	【実況】 [02]
03) 準備	長女	ななちゃん。	
03) 準備	父	きょう何年何月何日だ。	【報告】 [09]
05) 次女への声かけ	母	ななちゃん 威張らないで:。	【行動】 [01]
06) 長男の世界	長男	うー へ。 カタツムリに見える。	【実況】 [02]
08) 顎関節症	父	パパ顎痛いな。	【実況】 [02]
11) メモ帳	長女	パパ。あの。ねーね。最初パパのだった黄色いねメモ帳。	
12) レジン液	長女	あ。そうだ。パパ。もうレジン液がなくなっちゃった。	【実況】 [02]
14) 稲妻の熱さ	長男	パパ。稲妻は太陽より何倍熱いと思う?。	【説明】 [13]
16) カタツムリ	長男	パパってカタツムリ食べたことあるの?。	【自己記述】 [07]
18) ママの居場所	次女	ママ。 ママは?。	【実況】 [02]
20) 長男の声かけ	長男	ジャイアンと:。	
21) 焼き芋	長女	パパってスイートポテトと焼き芋どっちが好き?。	【自己記述】 [07]

話題が変わる時の発話には、【行動】 [01] 【実況】 [02] など脱文脈度が低いもののほか、相手の好みや経験を尋ねる【自己記述】 [07] や一般的な話題の【報告】 [09] 【説明】 [13] が用いられている。長男や長女が話題をもちかけても、必ずしも返答が得られないこともあり、応答が得られても、その話題が長く続くものばかりではなかったが、以下に示す、08) 顎関節症と 14) 稲妻の熱さは、主に父と長女、長男の間のやり取りが続き、脱文脈度の低いものから高いものまで様々な修辞機能が用いられていた。

### 3.2.2 顎関節症の話題の発話と修辞機能・脱文脈化指数

図4に、顎関節症の話題における修辞機能の推移を示す。縦方向に脱文脈化指数を示し、左から右に時間の経過を示している。

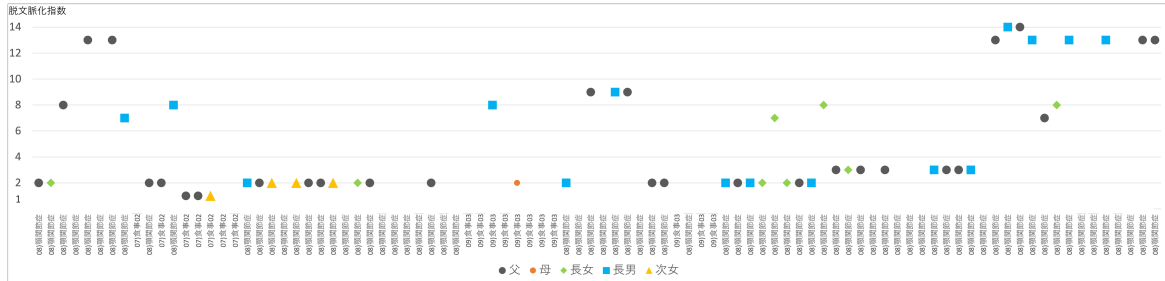


図4 顎関節症の話題場面の修辞機能の推移

食卓に並んでいる料理の確認を終えて、父も着席し、沈黙の後、父の「パパ顎痛いな、今日。」という発話がきっかけで子供3人と顎関節症の話が始まる。冒頭のやりとりを表6に示す。ここでは、顎関節症がどういうものを説明する父の発話が【説明】[13]で脱文脈度が高いが、それ以外は基本的には、子供達が父に痛いかどうかを問い、父が答える、というやりとりで、脱文脈度は低い。次女も不明瞭な発話ながら、「痛いのー?」「顎がー?」と確認し、会話に参加している。

表6 顎関節症の話題の冒頭の発話と修辞機能・脱文脈化指数

父	パパ顎痛いな。	実況 [02]
父	きょう。	実況 [02]
長女	なんで?。	実況 [02]
父	顎関節症つつうの。	
父	これ。	観測 [08]
長男	うん?。	
父	顎が痛くなる。	説明 [13]
長男	ふーん。	
父	噛めない。	説明 [13]
長男	どうして?。	自己記述 [07]
父	痛い。	実況 [02]
長男	どうして痛くなっちゃうんだろう。 (次女がお水を要求)	観測 [08]
長男	どうして噛めないんだろう。	実況 [02]
父	痛いから	実況 [02]
父	歯が。	
次女	(不明瞭) アイト:?(痛い:?)	実況 [02]
父	うん?	
次女	ここ?	実況 [02]
父	痛い痛い	実況 [02]
父	こう顎が痛い痛いって。	実況 [02]
次女	(不明瞭) クンダラ:(顎が:?)	実況 [02]
父	うん	
長女	顎?	実況 [02]
父	顎 (次女との食事についてのやりとり)	実況 [02]

この後、中盤や終盤では、表7に示すように、子供達から痛さについての質問があり、比較のために「おたふく風邪」という言葉がでて、過去におたふく風邪にかかったか、どれだけ顔が腫れたかという【状況内回想】[03]の話題や、おたふく風邪とはなにか、という【一般化】[14]や【説明】[13]のやりとりが交わされる。

表 7 顎関節症の話題の中盤・終盤の発話と修辞機能・脱文脈化指数

中盤			終盤		
長男	めちゃくちゃ痛い?。	【実況】 [02]	長男	おたふく風邪って?。	【一般化】 [14]
父	痛い。	【実況】 [02]	父	ほっぺが腫れる風邪。	【一般化】 [14]
長男	めちゃくちゃ痛い?。	【実況】 [02]	長男	悪い風邪?。	【説明】 [13]
長女	どっかが痛いつて。	【実況】 [02]	父	しゅんすけやってないか。	【自己記述】 [07]
父	これぐらい。	【実況】 [02]	長女	しゅんすけやってない。	【観測】 [08]
長男	痛い 痛い。	【実況】 [02]	長男	悪い風邪?。	【説明】 [13]
長女	おたふく風邪の痛いやつみたいなの?。	【観測】 [08]	長男	パパ 悪い風邪?。	【説明】 [13]
父	おたふくになった?。	【状況内回想】 [03]	父	うん。悪くもない。	【説明】 [13]
長女	ネーネなったよ。	【状況内回想】 [03]			
父	痛かった?。	【状況内回想】 [03]			
長女	うん。				
父	キャンプん時こんなってたよね。	【状況内回想】 [03]			

### 3.2.3 稲妻の熱さの話題

長男の「パパ、稲妻は太陽より何倍熱いと思う?。」という発話がきっかけで、父、長女、長男とのやりとりがある。図 5 に、稲妻の熱さの話題における修辞機能の推移を示す。

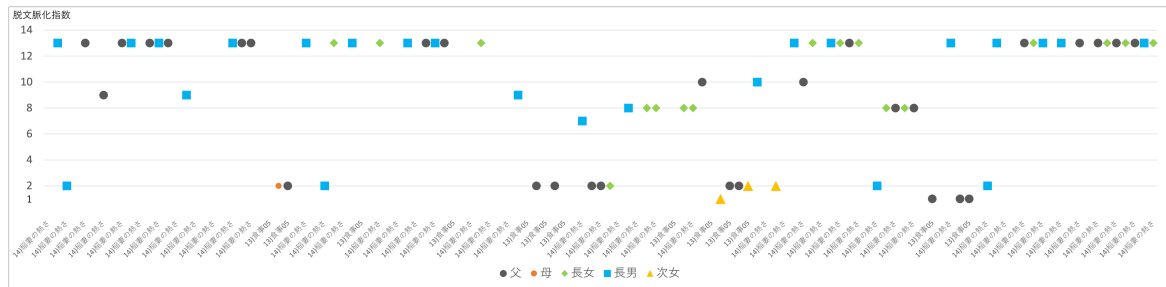


図 5 稲妻の熱さの話題場面の修辞機能の推移

表 8 に冒頭と中盤の発話を示す。長男の【説明】 [13] から始まり、稲妻の熱さから身を守るものとしてのボディガードについての話題に移り、表 9 に示すようにドラえもんにも言及する。長男の主張に対して父が質問したり、長男の発話内容の誤りと思われるところに長女がコメントをしたり補足をして、長男と長女の発話に脱文脈度の高い修辞機能が見られる。この間、次女もテーブルにいて、食事についての発話【実況】 [02] や動作をするが、稲妻やドラえもんなどの会話には加わらず、たまに不明瞭な発話をする。父と長女は、次女の対応【行動】 [01] をしながら、長男との会話に参加している。

表 8 稲妻の熱さの話題の冒頭・中盤の発話と修辞機能・脱文脈化指数

冒頭			中盤		
長男	パパ。稲妻は太陽より何倍熱い	【説明】 [13]	長男	でも:。たぶんね。	
長男	と思う?	【実況】 [02]	長男	ボディガードの があればいいよ。	【説明】 [13]
父	え?。太陽より熱いの?。	【説明】 [13]	長男	ボディガード。	【説明】 [13]
長男	ふん。		父	え。	
父	本当?。		長男	ボディガード。	【実況】 [02]
長男	ふん。		長女	ボディじゃ ボディじゃないよ。	【説明】 [13]
父	太陽のほうが熱いんじゃないの?。	【説明】 [13]	長女	しゅんすけ。	
長男	違うよ。	【説明】 [13]	長男	じゃあ何?。	

表9 稲妻の熱さの話題の終盤の発話と修辞機能・脱文脈化指数

父	見えないガードマンが何した?。	【状況外回想】 [10]
長男	ドラえもんの中で ガードマンが。	
次女	お肉:。	【行動】 [01]
父	ななちゃん これ これある これある。	【実況】 [02]
次女	嫌だ:。	【実況】 [02]
長男	しずかちゃん。しずかちゃんのことを守ってたの。	【状況外回想】 [10]
父	うーん?。	
次女	嫌だもん。	【実況】 [02]
長男	雷で。	
父	うん。	
長男	雷持てんのかな。	【説明】 [13]
父	カミが 雷落ちたの?。	【状況外回想】 [10]
長女	落ちてない。	【説明】 [13]
長男	ううん。	
長男	雷って持てんのかな。	【説明】 [13]
長女	持てるわけないじゃん。	【説明】 [13]
父	持てるってどういうこと?。	【説明】 [13]
長女	電気なんだから。	【説明】 [13]
次女	ハーカッテ。	
長男	ボディガード。	【実況】 [02]
長女	だからボディじゃないってば。	【観測】 [08]
父	何ガード?。	【観測】 [08]
長女	ガードマン。普通の。	【観測】 [08]
父	ガードマンね。	【観測】 [08]
次女	とうとうーとうーるーとうーるーとうーるーとうーるーとうーるー。	
父	ななちゃん座って。	【行動】 [01]
長女	なーちゃん。	
長男	見えない壁のやつあんじゃん。	【説明】 [13]
父	ななちゃん座って。	【行動】 [01]
父	ちゃんと座って。	【行動】 [01]

#### 4. 考察

テーマを設定されて作文を書いたり談話を行う場合、その中で用いられやすい修辞機能はテーマとの関係が見られることがわかっている (田中ほか 2021, 2022)。本稿の分析対象は設定されたテーマではなく、話者たち自身とその場でとりあげたものではあるが、話題内容によって、用いられる修辞機能が異なることがうかがえた。顎関節症はその時の父親の症状であるため、どのような病気かという説明をするために脱文脈度は高くなるが、基本的には父親の痛みに焦点があてられ、その場の話者たちの場に近い修辞機能が用いられたと考えられる。一方、稲妻の熱さの話題については、長男の「パパ。稲妻は太陽より何倍熱いと思う?。」という問いかけから始まり、話題内容自体がその場と全く関係のないものであるため、脱文脈度が高くなると考えられる。

7歳の長男は、3.2.1節で見たように、話題の始まりとなる発話をしているが、始まり以外にも、顎関節症やおたふく風邪についての父への質問やコメント、児童文学のキャラクターの話、メモ帳の使い方への意見など、脱文脈度の高い発話を行っていた。「顎関節症の話題」のように脱文脈度の高い発話も含まれるものの基本的には脱文脈度の低い発話で構成された話題では、2歳の次女は、不明瞭ではあるが会話に参加する場面が見られた。しかし、基本的に脱文脈度が高い「稲妻の話題」では次女は会話に加わらず、その合間に生じている次女の発話は稲妻関係の内容ではなく食事に関するものであった。田中・小磯 (2019) では、食事場面において、両親の脱文脈度の高い発話をきっかけに、幼児も脱文脈度の高い発話を行うことがわかっているが、本稿の分析対象データからは、2歳ではそのような脱文脈度の高い発

話を行うことはないことがうかがえた。田中・小磯 (2019) の幼児より年齢が低いこと、及び、田中・小磯 (2019) では両親と3人の会話で子供が1人であったのに対して、本データでは、両親の他に年上の姉兄が2人いるという状況の違いの影響も考えられるだろう。

## 5. おわりに

本発表の目的は、親子の会話における話題内容及び年齢と、脱文脈度の関連を明らかにすることであった。家庭での食事場面の親子による談話データを対象として、それぞれの発話について言語表現から修辞機能と脱文脈度を特定し、談話の話題内容によってどのような修辞機能が用いられているか、年齢による違いが見られるかを確認した。分析の結果、まず、食事についての話題では脱文脈度が低く、空間的距離の近い修辞機能が用いられていること、食事以外の話題では、内容に応じて脱文脈度は変わることが確認された。また、子どもの会話への参加は、2歳児は脱文脈度の低い会話には参加できるが脱文脈度の高い会話には参加しておらず、7歳児は脱文脈度の高い発話を自ら発信しており、年齢と使用できる脱文脈度とに関わりがあることが確認された。今後の課題として、同家庭の一年前や一年後のデータにより、同一兄弟の異なる時期の談話の分析を試みることで、また、他家族での食事場面の様子の分析によって、同様の傾向が見られるか確認し、親子会話に共通する特徴を明らかにしたいと考える。

## 謝 辞

本研究は国立国語研究所のプロジェクト「多世代会話コーパスに基づく話し言葉の総合的研究」、JSPS 科研費 JP19K00588、JP20H01264、JP23H00630 によるものです。

## 文 献

- C. Cloran (1994). "Rhetorical units and decontextualisation: An enquiry into some relations of context, meaning and grammar." Unpublished doctoral dissertation, University of Nottingham Nottingham.
- C. Cloran (1999). "Contexts for learning." Frances C (Ed.), *Pedagogy and the Shaping of Consciousness: Linguistic and Social Processes*. London: Continuum International Publishing. pp. 31–65.
- 漁田武雄・漁田俊子 (1999). 「授業と自宅学習の間で生じる文脈依存記憶」 静岡県立大学短期大学部研究紀要:13-2, pp. 205–211.
- 岩田純一・石田勢津子・落合正人 (1995). 『児童の心理学』 有斐閣, 東京.
- 佐尾ちとせ・宮城信・田中弥生 (2023). 「修辞機能分析を活用した作文指導」 日本語習熟論研究:1, pp. 140–158.
- 佐野大樹 (2010). 『日本語における修辞ユニット分析の方法と手順 ver.0.1.1: 選択体系機能言語理論 (システム理論) における談話分析 (修辞機能編)』, <https://researchmap.jp/kotonoha/>資料公開/.
- 佐野大樹・小磯花絵 (2011). 「現代日本語書き言葉における修辞ユニット分析の適用性の検証 - 「書き言葉らしさ 話し言葉らしさ」と脱文脈化言語 文脈化言語の関係 -」 機能言語

- 学研究, 6, pp. 59–81.
- 田中弥生 (2017). 「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」 言語資源活用ワークショップ発表論文集, 1, pp. 69–78.
- 田中弥生 (2022). 「修辞機能と脱文脈化の観点からの日本語談話分析」 博士論文（未公刊）, 東京大学大学院総合文化研究科.
- 田中弥生・小磯花絵 (2019). 「家庭での幼児の発話の修辞機能：脱文脈化の観点からの検討」 言語資源活用ワークショップ発表論文集:4, pp. 106–118.
- 田中弥生・小磯花絵・大武美保子 (2023). 「共想法による話し言葉・書き言葉における修辞機能の特徴－テーマとの関係に着目して－」 言語処理学会第 29 回年次大会発表論文集, pp. 1356–1360.
- 田中弥生・小磯花絵・大武美保子 (2022). 「脱文脈化の観点から見た共想法に基づく高齢者談話の分析」 国立国語研究所論集:22, pp. 137–155.
- 田中弥生・佐尾ちとせ・宮城信 (2021). 「児童作文の評価に向けた脱文脈化観点からの検討」 言語処理学会 第 27 回年次大会 発表論文集, pp. 750–755.
- 橋本憲尚 (2009). 「学校教育と知能観の再考：”状況に埋め込まれた学習”の視点から」 佛教大学総合研究所紀要:16, pp. 1–18.