＜全文＞言語資源活用ワークショップ2018発表論文集

著者	国立国語研究所コーパス開発センター
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	1-611
発行年	2018
URL	http://doi.org/10.15084/00001631

言語資源活用ワークショップ 2018

発表論文集

2018 年 9 月 4・5 日 (火・水)　『言語資源活用ワークショップ 2017』
2018 年 9 月 6 日 (木)　『コーパスとしてのウェブテキスト活用シンポジウム』

大学共同利用機関法人 人間文化研究機構
国立国語研究所 コーパス開発センター 編

Programme:言語資源活用ワークショップ 2018

2018 年 9 月 4 日 (火)

10:00-10:10　**■挨　拶 (2F 講堂)**　前川喜久雄

10:10-12:00　**■口頭発表　A グループ (2F 講堂)**

[**O-1-01-S**]

『現代日本語書き言葉均衡コーパス』のロシア語翻訳データの構築

.....宮内 拓也 (東京外国語大学/日本学術振興会：学生)

..... Prokhorova Maria(東京外国語大学：学生)

[**O-1-02-S**]

中古語における形容詞テ形をめぐって–形容詞の意味分類との関わりから–

.....菊池そのみ (筑波大学：学生)

[**O-1-03-S**]

日本語文における連用修飾語成分に見られるパラレルについての一考察 –「赤く変わる」と「赤に変わる」とは同じか–

.....王 棟 (東京外国語大学：学生)

[**O-1-04-S**]

連体助詞の「ノ」と文体の関係

.....森 秀明 (東北大学：学生)

12:00-13:00　**休憩**

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

[**P-1-01-E**]

「日本語日常会話コーパス」への談話行為アノテーションの試み： タグ選択が困難な事例に焦点を当てて

.....居關 友里子 (国立国語研究所)

.....門田 圭祐 (早稲田大学：学生)

.....伝 康晴 (千葉大学/国立国語研究所)

[**P-1-02-E**]

「『了解』は使わないように」「了解です！」

.....髙橋 圭子 (フリーランス)・東泉 裕子 (フリーランス)

.....佐藤 万里 (フリーランス)

[**P-1-03-E-S**]

NWJC における敬語使用とレジスターとの関係

.....金 賢眞 (大阪大学：学生)

[**P-1-04-E-S**]

先天性全盲ろう児の音声言語訓練長期記録の分析状況及び保存活動
. . . . . 菊池 英明 (早稲田大学)・市川 熹 (千葉大学)
. . . . . 岡本 明 (筑波技術大学)・長嶋 祐二 (工学院大学)
. . . . . 藤本 浩志 (早稲田大学)・引田 秋生 (元山梨県立山梨盲学校)

[**P-2-09**]
『BCCWJ 図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性
. . . . . 馬場 俊臣 (北海道教育大学)

[**P-2-10**]
脚本テキストに基づくコーパス文体論の可能性 –テレビドラマ脚本に注目して–
. . . . . 松下 晶子 (専修大学大学院文学研究科：学生)
. . . . . 丸山 岳彦 (専修大学/国立国語研究所)

[**P-2-11-E**]
『UniDic』を活用した語構造情報付与の試み–『日本語歴史コーパス』を対象に–
. . . . . 村山 実和子 (国立国語研究所)

[**P-2-12**]
Twitter で使われる「深い」の意味–「強い」「すごい」と比較して–
. 加藤 恵梨 (大手前大学)・山下 紗苗 (明石工業高等専門学校：学生)
. . . . . 上 泰 (明石工業高等専門学校)

[**P-2-13**]
日本語の二重目的語構文の基本語順について
. . . . . 浅原 正幸 (国立国語研究所)・南部智史 (モナシュ大学)
. . . . . 佐野 真一郎 (慶応義塾大学)

15:35-15:45　**休憩**

15:45-17:00　**■口頭発表　B グループ (2F 講堂)**

[**O-2-01-S**]
比喩指標としての「感じる」–文法形式と比喩の関係–
. . . . . 菊地 礼 (中央大学：学生)

[**O-2-02-S**]
日本語 Wikipedia を用いた慣用句の構成性の数値化
. . . . . 岡田 優也 (関西学院大学：学生)

[**O-2-03-S**]
「ＸＸ (と)」、「ＸＸな」、「ＸＸしい」の構造・文法機能 –畳語による生産性について–
. . . . . 陳 祥 (筑波大学：学生)

2018 年 9 月 5 日 (火)

10:00-11:00 **■口頭発表　C グループ (2F 講堂)**

[O-3-01]

ニュースを対象にした手話マルチメディアコーパスの構築

..... 加藤 直人 (NHK 放送技術研究所)

..... 内田 翼 (NHK 放送技術研究所)

..... 東 真希子 (NHK 放送技術研究所)

..... 梅田 修一 (NHK 放送技術研究所)

[O-3-02]

ベイズモデルによる方言音声共通語化過程の分析

..... 前川 喜久雄 (国立国語研究所)

11:00-12:00 **■招待講演 (2F 講堂)**

[I-1-01]

言「考」不一致の言語学: コーパスはどこまで意識に迫れるか

..... 吉川 正人 (慶応義塾大学)

12:00-13:00 **休憩**

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

[P-3-01-E]

『日本語日常会話コーパス』活用環境の構築

..... 山口 昌也 (国立国語研究所)

[P-3-02-E]

「よい子」って誰？–政策ニュース映画のナレーション表現に関する研究の一環として–

..... 春木 良且 (フェリス女学院大学)・田中 弥生 (東京大学：学生)

[P-3-03-E]

敬語接頭辞異形「お ～」「ご ～」両者の用例のある語について

..... 服部 匡 (同志社女子大学)

[P-3-04-E]

撥音（の解析）は機械（UniDic）にとっても簡単ではなかったんだ！–BCCWJ を中心に–

..... 劉 志偉 (埼玉大学)

[P-3-05]

『現代日本語書き言葉均衡コーパス』書籍サンプルに対する NDC 記号拡張アノテーションと NDC 形式区分を用いた「随筆」の文体分析

..... 加藤 祥 (国立国語研究所)・櫻井 芽衣子 (日本工業大学)

..... 森山 奈々美 (津田塾大学：学生)・浅原 正幸 (国立国語研究所)

[P-3-06]

14:15-14:20 **休憩 (ポスター切替)**

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

[P-4-01-E]

日本語の非流ちょう性 –とぎれと延伸の数量調査から–

. . . . . 佐々木 藍子 (国立国語研究所／東京学芸大学：学生)

. . . . . 砂川 有里子 (筑波大学名誉教授)・浅原 正幸 (国立国語研究所)

[P-4-02-E]

『日本語日常会話コーパス』モニター公開版の概要

. . . . . 小磯 花絵 (国立国語研究所)・天谷 晴香 (国立国語研究所)

. . . . . 居關 友里子 (国立国語研究所)・臼田 泰如 (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：学生)

. . . . . 田中 弥生 (国立国語研究所：学生)

. . . . . 西川 賢哉 (国立国語研究所)

. . . . . 伝 康晴 (千葉大学／国立国語研究所)

[P-4-03-E]

日本語学習者属性別の言語行為の対話自動生成への適用に関する一考察

. . . . . 太田 博三 (放送大学：学生)

[P-4-04-E]

『現日研・職場談話コーパス』中納言版公開データの作成

. . . . . 柏野 和佳子 (国立国語研究所)・大村 舞 (国立国語研究所)

. . . . . 西川 賢哉 (国立国語研究所)・小磯 花絵 (国立国語研究所)

[P-4-05]

日本語オノマトペ共起表現レキシコン JMWEL_onomatopoeic

. . . . . 首藤 公昭 (福岡大学名誉教授)・田辺 利文 (福岡大学)

. . . . . 髙橋 雅仁 (久留米工業大学)

[P-4-06]

語彙多様性指標の可視化と単回帰分析による TTR の補正

. . . . . 今田 水穂 (文部科学省)

[P-4-07]

二字漢語を構成する漢字の造語力の変化 –『現代雑誌九十種の用語用字』データと『現代日本語書き言葉均衡コーパス』の比較を通して–

. . . . . 本多 由美子 (一橋大学：学生)

[P-4-08]

方言音声に対するテキスト自動アライメントの試み

. . . . . 石本 祐一 (国立国語研究所)

[P-4-09]

単語の分散表現を用いた領域における出現単語の特徴分析

. . . . . 佐々木 稔 (茨城大学)・古宮 嘉那子 (茨城大学)

Programme:コーパスとしてのウェブテキスト活用シンポジウム

2018 年 9 月 6 日 (木)

13:00-13:15　趣旨説明

……岡 照晃 (国立国語研究所)

【セッション 1】日本語研究に大規模ウェブテキストデータを扱うためには？ (2F 講堂)

13:15-13:55　【初級編】ウェブの検索結果を利用する

……荻野 綱男 (日本大学)

13:55-14:35　中の人が国語研日本語ウェブコーパス（**NWJC**）を使ってみた

【中級編】ウェブコーパスを “使って” みる–

……岡 照晃 (国立国語研究所)

14:35-15:15　ウェブコーパスの表と裏

……林部 祐太 (Megagon Labs)

15:15-15:30　休憩

【セッション 2】企業は大規模ウェブテキストデータをどのように活用しているか？ (2F 講堂)

15:30-16:10　利便性のあるコーパス構築へのテキストマイニング取り組み

–ビジネス分析に役立つ解析手法開発–

……三澤 賢佑 (Insight Tech)

16:10-16:50　**Wikipedia** を使った進んだ自然言語処理

……山田 育矢 (Studio Ousia)

目次

発表論文集

『現代日本語書き言葉均衡コーパス』のロシア語翻訳データの構築

宮内 拓也 (東京外国語大学大学院 / 日本学術振興会特別研究員 / 国立国語研究所共同研究員)*
プロホロワ・マリア (東京外国語大学大学院)

Construction of Russian Translation Data of "The Balanced Corpus of Contemporary Written Japanese"

Takuya Miyauchi (TUFS / JSPS / NINJAL)
Maria Prokhorova (TUFS)

要旨

『現代日本語書き言葉均衡コーパス』(の一部のデータ) には，既に英語，イタリア語，インドネシア語，中国語の翻訳データが構築されているが，新たにロシア語の翻訳データを構築した．対象となるテキストは『現代日本語書き言葉均衡コーパス』新聞 (PN) コアデータ 16 サンプル (総語数は短単位で全 16,657 語) とし，ロシア語翻訳データの総語数は 13,070 語となった．本データの構築あたっては，日本語からロシア語へ人手による翻訳を行った．また，日本語とロシア語の言語構造の違いにより，翻訳に困難を生じさせた箇所も多くあった．本稿では，翻訳データの構築方法，翻訳の際の留意点の詳細を述べる．さらに，この翻訳データの構築により，原データと並べることで疑似的な日露対訳コーパスとしての利用も可能であり，本データは日露対照研究に活用できると考えられる．本稿では，そのような活用の一例として，日本語の文末形式について，簡単にロシア語と対照させて論じる．

1. はじめに

『現代日本語書き言葉均衡コーパス』(Maekawa et al. 2014; 以下，BCCWJ) (の一部のデータ) には，既に英語，イタリア語，インドネシア語，中国語の翻訳データが構築されている．今回，新たに人手による翻訳により，BCCWJ の 16 サンプル分のロシア語の翻訳データを構築した．本稿では，翻訳データの構築方法，翻訳の際の留意点や翻訳データの活用の可能性を報告する．

以下，2 節では翻訳対象としたデータと構築したデータの概要について，3 節では翻訳の方法と翻訳の際の留意点について，4 節では翻訳データの日露対照研究への活用の可能性についてそれぞれ述べる．5 節は本稿全体のまとめである．

* miyauchi.takuya.k0 @ tufs.ac.jp

2. 翻訳対象のデータと翻訳データの概要

翻訳の対象は BCCWJ の新聞 (PN) コアデータ 16 サンプルである．サンプルの選択は BCCWJ-ANNOTATION-ORDER[1] に基づく．対象の基本的なデータとして，表 1 に対象となるテキストの総語数 (短単位数)，文節数，文数を示す．

表 1　対象となるテキストのサイズ

サンプル名	短単位数	文節数	文数
PN1c_00001	784	236	42
PN1d_00001	783	235	34
PN1e_00001	763	219	35
PN1f_00001	797	181	38
PN2e_00001	750	214	27
PN3b_00001	975	311	43
PN3g_00001	2,640	919	142
PN4a_00001	1,244	425	51
PN4b_00001	758	246	26
PN4c_00001	737	250	31
PN4f_00001	1,047	297	40
PN4g_00001	905	296	36
PN1a_00002	1,797	611	93
PN1b_00002	1,024	277	38
PN1d_00002	734	206	28
PN1e_00002	919	272	35
計	16,657	5,195	739

構築されたロシア語翻訳データの総語数は 13,070 語であり，文数は 848 文であった．表 2 に対象とする各サンプルごとの語数，文数を示す．

既に述べたように，『現代日本語書き言葉均衡コーパス』(の一部分のデータ) には，英語，イタリア語，インドネシア語，中国語の翻訳データがある．これらの BCCWJ 外国語翻訳データと比較すると，ロシア語が今のところの規模としては最も大きい．各言語の翻訳データの総語数[2]，対象のサンプル数をまとめたものが表 3 である．

(1) BCCWJ コアデータサンプルにおけるアノテーション優先順序である．以下参照のこと．https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER

(2) ただし，中国語は字数をカウントしている．

表 2　ロシア語翻訳データのサイズ

サンプル名	語数	文数
PN1c_00001	568	40
PN1d_00001	555	54
PN1e_00001	653	34
PN1f_00001	568	56
PN2e_00001	639	33
PN3b_00001	787	39
PN3g_00001	2,227	167
PN4a_00001	1,012	62
PN4b_00001	548	23
PN4c_00001	587	33
PN4f_00001	761	38
PN4g_00001	674	36
PN1a_00002	1,409	106
PN1b_00002	763	57
PN1d_00002	564	29
PN1e_00002	755	41
計	13,070	848

3. 翻訳方法

翻訳データの構築あたっては，日本語からロシア語へ人手による翻訳を行った．翻訳者は，東京外国語大学大学院博士前期課程の，翻訳家を志望するロシア語母語話者の学生 (第 2 著者) である．

翻訳にあたり，日本語とロシア語の言語構造の違い等により，翻訳に困難を生じさせるであろう箇所が多くあることが予想されたため，一定の方針を設定した．

まず，談話レベルではロシア語としての自然さは失われてもよいとしたが，単文レベルでは自然なロシア語になるような翻訳を行った．例えば，日本語では現在形と過去形が混ざっている文章が多々あるが，ロシア語だと特定の文脈がない限りどちらかで統一するのが一般的である．今回の翻訳では，単文レベルで翻訳元の日本語の文の時制を，ロシア語文でも用いることとした．例えば，(1) のような文を見られたい．翻訳元の日本語文 (1a) では，現在形 (ル形) と過去形 (タ形) が共に用いられており，ロシア語文 (1b) でもそれに合わせて翻訳されている．

(1)　a. [...] 二、三年時に担任だった池田弘子先生（七十五）は違った。「そんな薄いかばんじゃ遊び道具も入らないよ」「体育や部活では、危ないからピアスをはずしたほうがいい」。やんわり語りかける。

表 3　ロシア語翻訳データと他の言語の翻訳データの比較

	総語数	対象のサンプル数	対象のサンプル名
ロシア語	13,070 語	16 サンプル	PN1c_00001 / PN1d_00001 / PN1e_00001 / PN1f_00001 / PN2e_00001 / PN3b_00001 / PN3g_00001 / PN4a_00001 / PN4b_00001 / PN4c_00001 / PN4f_00001 / PN4g_00001 / PN1a_00002 / PN1b_00002 / PN1d_00002 / PN1e_00002
英語	4,840 語	6 サンプル	OY04_00001 / OC01_00001 / PM25_00001 / PB12_00001 / PN1c_00001 / OW6X_00000
イタリア語	6,563 語	16 サンプル	OC01_00001 / OW6X_00000 / OY04_00001 / PB12_00001 / PM25_00001 / PN1c_00001 / OC02_00001 / OY12_00005 / OC03_00001 / OY09_00008 / OC04_00001 / OY15_00014 / OC05_00001 / OY04_00017 / OC06_00001 / OY04_00027
インドネシア語	51 語	1 サンプル	OY09_00008
中国語	7,852 字	6 サンプル	OY04_00001 / OC01_00001 / PM25_00001 / PB12_00001 / PN1c_00001 / OW6X_00000

b. Xiroko Ikéda (75 let), kotoraja byla ee klassnym
[Hiroko Ikeda]-NOM.F 75 ages who-NOM.F be-PST.F her [homeroom
rukovoditelem na vtorom i tret'em godu obučenija,
teacher]-INS on second-LOC and third-LOC year-LOC education-GEN
byla ne takoj. Ona razgovarivaet s devočkoj mjagko: ⟨⟨V
be-PST.F not such she-NOM talk-PRS.SG.3 with girl-INS softly in
takuju ploskuju sumku daže igry ne vlezut⟩⟩, ⟨⟨Na
such-ACC thin-ACC bag-ACC even toys-NOM.PL not go-PRS.PL.3 on
fizkul'ture i v sekcijax lučše snimat' serežki,
physical_education-LOC and in section-LOC better take_off-INF earrings-ACC

éto opasno⟩⟩.
it_is dangerous [3]

(読売新聞 [BCCWJ: PN1c_00001])

また，ロシア語では，ロシア国外の企業名や国外の新聞の名前などはキリル文字で表記される場合とローマ字で表記される場合がある．今回の翻訳では，企業名等は特定の場合[4]を除き，ローマ字で表記することとした．なお，このキリル文字とローマ字の表記については，単に表記の問題というわけではなく，文法面にも影響を与える問題である．一般に，ローマ字表記となる場合，その名詞は曲用しないことになるが，キリル文字表記であれば曲用する[5]．ロシア語では固有名詞でもキリル文字表記であれば曲用してしまうため，ローマ字表記にして曲用させない方がもともとの名称がわかりやすい．

固有名詞には，上位概念を示す名詞 (例えば，*kompanija*「会社」, *proizvoditel'*「メーカー」など) が同格句として前置される場合もよくある[6]．今回の翻訳では，(2a) のように上位概念を示す名詞が日本語で表示されていても，(3a) のように表示されていなくても，(2b, 3b) で示すようにロシア語ではこのような上位概念を示す名詞を加えることとした．

(2) a. 米ガートナー・グループ傘下の調査会社データクエスト

b. kompanija Dataquest, prinadležaščaja amerikanskoj Gartner
company-NOM Dataquest belong_to-PTCP.NOM American-DAT Gartner
Group
Group

(産経新聞 [BCCWJ: PN1d_00002])

(3) a. ファミリーマートは 9 日、[...] 発表した。

b. 9 čisla kompanija FamilyMart soobščila [...].
9 number company-NOM.F FamilyMart announced-PST.F

(産経新聞 [BCCWJ: PN1d_00002])

---

[3] キリル文字はローマ字に翻字する．翻字は以下の通りである：A=A，Б= B，B=V，Г=G，Д=D，E=E，Ё=E，Ж=Ž，З=Z，И=I，Й=J，К=K，Л=L，M=M，H=N，O=O，П=P，P=R，C=S，T=T，У=U，Ф=F，X=X，Ц=C，Ч=Č，Ш=Š，Щ=Šč，Ъ=”，Ы=Y，Ь=’，Э=É，Ю=Ju，Я=Ja. また，本稿で用いる文法情報の略記は以下の通りである：NOM=主格，GEN=属格 (生格)，DAT=与格，ACC=対格，INS=具格 (造格)， LOC=前置格 (処格)，M=男性，F=女性，N=中性，PRS=現在，PST=過去，PTCP=分詞 (形動詞)，INF=不定形，3=3 人称.

[4] 例外となる特定の場合とは，ロシア語で正式名称のあるものである．例えば，国外の新聞の名前で言えば，「人民日報」(*Žén'min' žibao*)，「ルモンド」(*Mond*) 「タイムズ」(*Tajms*) 「エスタド・デ・サンパウロ」(*Éštadau*) などである．

[5] 例えば，*Toyota*「トヨタ」であれば，ローマ字表記の場合は，格により形態が変化することはないが，キリル文字で表記した場合，(i) で示すように曲用する．

(i) Tojota, Tojoty, Tojote, Tojotu, Tojotoj, o Tojote
Toyota-NOM Toyota-GEN Toyota-DAT Toyota-ACC Toyota-INS about Toyota-LOC
「トヨタが，トヨタの，トヨタへ，トヨタを，トヨタによって，トヨタについて」

[6] 以降，(2-6) では，上位概念を示す名詞に下線を引く．

(3b) では，上位概念を示す名詞として *kompanija*「会社」が付加されている．もし，これが示されていない場合，*FamilyMart* を知っているロシア語母語話者は一つの店舗としての *FamilyMart* を想定してしまう可能性が高い．この場合は，(4) のように上位概念を示す名詞として *magazin*「店」を想定することになる．*magazin* が表示されていなくても，動詞は形式上これと一致し，男性形となる(7)．

(4) 9 čisla (magazin) FamilyMart soobščil [...].
9 number shop-NOM.M FamilyMart announced-PST.M

このように，上位概念を示す名詞の違いによって，内容的な齟齬をきたすだけではなく，文法面にも影響が出る．適切な上位概念を示す名詞を付加することで曖昧性を排除することができ，文がわかりやすくなるという効果がある．

地名については事情が少々複雑である．(5a) で示すように，上位概念を示す名詞があっても，ロシアの地名は普通は曲用する(8)．これと同様の方針で翻訳すれば，例えば「静岡県」は (5b) のように曲用させることになる．

(5) a. v gorode Moskve
in city-LOC Moscow-LOC
「モスクワ (という都市) で」

b. v prefekture Sidzuoke
in prefecture-LOC Shizuoka-LOC
「静岡県で」

しかしながら，この場合では，もともとの名称が分かりにくい．よって，今回の翻訳では，上位概念を示す名詞を付加し，(6) のように地名そのものは曲用させない方針とした．

(6) a. 静岡県出身。
b. Rodom iz prefektury Sidzuoka.
birth-INS from prefecture-GEN Shizuoka-NOM
(西日本新聞 [BCCWJ: PN3g_00001])

ただし，今回の翻訳の対象は PN コアデータであり，出典が新聞からのテキストであることから，地名が多く出現する．地名が出るたび，そのすべてに上位概念を示す名詞 (*gorod*「都市」，*prefektura*「県」など) を付加していくのは明らかに文章として不自然となる．よって，「東京・大阪・京都・広島」など，ロシア人にとってもなじみのある地名には (7) で示すように，基本的に日本語文に示されていない限り，このような語は付加しないこととした．

(7) ただし，このあたりのロシア語母語話者の言語感覚やロシア語の言語現実は大変複雑であるため，詳細はここでは述べない．

(8) ただし，以下 (ii) で示すように，河川の名称など，曲用しない場合もあり得る．

(ii) a. na reke Volge
at river-LOC.F Volga-LOC.F
「ヴォルガ川で」

b. na reke Enisej
at river-LOC.F Yenisei-NOM.M
「エニセイ川で」

上位概念を示す名詞 (ここでは *reka*「川」) と名称を示す名詞の性が一致している場合，曲用し (ii a)，一致しない場合は曲用しない (ii b) とされるが，詳細はここでは述べない．

(7) a. 東京のヨドバシカメラ新宿西口本店

b. v glavnom magazine Yodobashi Camera v Tokio u zapadnogo vyxoda
in main-LOC shop-LOC Yodobashi Camera in Tokyo at western-GEN exit-GEN
so stancii Sindzjuku
from station-GEN Shinjuku

(朝日新聞 [BCCWJ: PN4a_00001])

この場合, (7) の *Tokio*「東京」や *Kioto*「京都」など*-o* で終わる地名は不変化名詞 (indeclinable noun) となるため，曲用させない．しかし，*Osaka*「大阪」や *Xirosima*「広島」のように*-a* で終わる地名は (8) で示すように曲用させることになる．

(8) a. 広島、大阪各高裁長官を経て

b. zanimal post glavy Vysšego suda Xirosimy,
was_engaged_in post-ACC head-GEN high-GEN court-GEN Hiroshima-GEN
zatem glavy Vysšego suda Osaki
then head-GEN high-GEN court-GEN Osaka-GEN

(西日本新聞 [BCCWJ: PN3g_00001])

4. 日露対照研究への活用の可能性

BCCWJ のロシア語翻訳データの構築により，日本語の原データと並べることで疑似的な日露対訳コーパスとしての利用も可能であり，本データは日露対照研究へ活用できると考えられる．本稿では，その一例として，日本語の文末形式について，簡単にロシア語と対照させて論じる．

(9a-11a) の日本語の各文の文末形式と (9b-11b) のロシア語におけるその対応部分 (下線部) を見られたい．

(9) a. [...] 非常通報装置が作動。[...] 商品のビデオカメラ六十台とノートパソコン四台 [...] が盗まれていた。

b. [...] **srabotala** sistema signalizacii. [...] ukradeno 60
worked-PST.F system-NOM.F signaling-GEN stolen-PTCP 60
videokamer i 4 noutbuka iz čisla tovarov [...]
video_camera and 4 laptop from number products-GEN

(中日新聞 [BCCWJ: PN4f_00001])

(10) a. [...] 異国の食文化をどん欲に吸収。[...] パスタもレパートリーに加えた。[...] 心を奪われた。[...] 日本語学校にも通い始めた。

b. [...] on [...] žadno **vpityval** kulinarnye tradicii
he-NOM.M greedily absorb-PST.M culinary-ACC traditions-ACC
čužoj strany. On takže dobavil v svoj
foreign-GEN country-GEN he-NOM.M also added-PST.M into self's-ACC

kulinarnyj repertuar pastu [...] on byl
culinary-ACC repertory-ACC pasta-ACC he-NOM.M was-PST.M
neverojatno očarovan [...] on daže načal xodit' v
unbelievably fascinated-PTCP.M he-NOM.M even start-PST.M go-INF to
školu japonskogo jazyka.
school-ACC Japanese-GEN language-GEN

(読売新聞 [BCCWJ: PN4c_00001])

(11) a. 日本政府は [...] 無形文化遺産保護条約を締結した。[...] 佐藤禎一大使が [...] 締約受諾書を提出した。同条約は [...] 採択された。締結はアルジェリアなどに続いて**三カ国目**。

b. [...] pravitel'stvo Japonii prinjalo Konvenciju ob
government-NOM.N Japan-GEN accepted-PST.N convention-ACC about
oxrane nematerial'nogo kul'turnogo nasledija, [...]
protection-LOC intangible-GEN cultural-GEN heritage-GEN
Posol [...] Téjiti Sato pred"javil [...]
ambassador-NOM.M [Teiichi Sato]-NOM.M presented-PST.M
dokument o soglasii na prinjatie konvencii.
document-ACC about agreement-LOC on acceptance-ACC convention-GEN
Éta konvencija byla utveržden [...] Vsled za
this-NOM.F convention-NOM.F was-PST.F approved-PTCP.F following
Alžirom i t.d. Japonija **stala** tret'im gosudarstvom,
Algeria and so on Japan-NOM.F became-PST.F third-INS nation-INS
prinjavšim konvenciju.
taking-PTCP.INS convention-ACC

(西日本新聞 [BCCWJ: PN4g_00001])

(9a-11a) の日本語文の文末形式に注目すると，名詞で文を終える体言止めを用いている箇所がある．一方，日本語文で体言止めとなっている箇所の対応部分では (9b-11b) のロシア語では動詞の過去形 (*-l/-la*) で示されている(9)．

もし，日本語文で体言止めとなっている箇所を完全な文の形式とするのであれば，動名詞 (verbal noun) で体言止めにされている (9a, 10a) では「-した」を追加し，(11a) では例えば「-であった」とコピュラを追加することになろう．このようにした場合，その文末形式は，例えば (10a) では「[...] 吸収した．[...] 加えた．[...] 奪われた．[...] 始めた．」となり，過去形 (タ形) のみが続き文章が単調になってしまう．これを避けるために，適宜体言止めが用いられているものと考えられる(10)．ロシア語にも，日本語の動名詞を用いた構文のように体系的に動

(9) ボールド体で示してある．

(10) 出典のテキストが新聞からであるため，字数の制限等も関係してくる可能性がある．また，これについて結論を出すためにはより詳細な検討が必要となる．

詞を名詞化させる方法は存在するが，(9b-11b) で示すように，そのような形式はここでは用いられていない．(9a-11a) の日本語文の文末形式に対応する部分は (9b-11b) のロシア語では 1 例を除き全て動詞の過去形 (*-l/-la/-lo*) となっている(11)が，ロシア語の基本語順は SVO であり (Isačenko 1966 など)，基本的に述語動詞は文末に位置しない(12)ため，動詞の時制により文末形式が固定されることはない．そのため，文末形式を多様化するという日本語のような理由では，動詞の名詞化の構文は用いられないと考えられる．

以上のように，ロシア語では，文末に述語が位置しないことが多く，動詞の時制が文章内で同一であっても文末形式が固定されることはない．一方，日本語では述語がほぼ必ず文末に位置するために，動詞の時制が過去で統一されてしまうと，文末形式が「タ」に固定されてしまう．日本語では過去形 (タ形) の連続を避け，文末形式を多様化するために，体言止めが適宜用いられるといえる(13)．

5. おわりに: まとめと今後の課題

本稿では，BCCWJ のロシア語翻訳データの構築について述べた．

BCCWJ の新聞 (PN) コアデータ 16 サンプルを対象に日本語からロシア語へ人手による翻訳を行った結果，ロシア語翻訳データの総語数は 13,070 語となり，既に構築されていた英語，イタリア語，インドネシア語，中国語の各翻訳データと比べると最大の規模となった．

翻訳の際は，日本語文の時制をロシア語文でも用いること，企業名等はローマ字で表記すること，固有名詞にはできる限り上位概念を示す名詞を付加すること，「東京・大阪・京都・広島」などのロシア人もよく知る地名に限り上位概念を示す名詞を付加しないこと，等を方針とした．

コーパス研究が盛んである今日でも日本語・ロシア語の対訳コーパスは大変希少であり，BCCWJ の 16 サンプルであっても，その基礎となり得るデータを構築したことは日露対照研究，さらには類型論研究に対し一定の意義のある言語資料を提供できたといえる．本稿では，翻訳データの日露対照研究への活用の一例として，日本語の文末形式について，簡単にロシア語と対照させて論じた．ロシア語とは異なり，日本語の新聞には体言止めがしばしば使用されることを指摘し，その要因は文末形式を多様化するためであるとした．

Soejima (2017) は文学作品 (とその翻訳作品) を用いて日露語の対訳コーパスを構築し，不特定の動作主が関わる意図的な出来事が日本語とロシア語でどのように表現されるかについて検討している．その結果，過程を表す場合は，日本語では受動文がよく用いられる一方でロシア語では不定人称文がよく用いられるとしている．結果を表す場合は，日本語では自動詞文がよく用いられ，ロシア語では受動文や自動詞文など多様な形式が用いられるとしている．Soejima (2017) は文学作品でこの結論を導いているため，新聞という異なるレジスターでも同様の結果が得られるか今後検討したい．

さらに，今回ロシア語に翻訳した日本語のサンプルは情報構造のアノテーションもなされて

(11) (9b) の *ukradeno*「盗まれた」のみ受動分詞である．
(12) ただし，ロシア語は語順が自由であり，述語動詞を文末に位置させることも可能ではある．
(13) もちろん，これについて結論を出すためにはより詳細な検討が必要となる．

いる (Miyauchi et al. 2018)．よって，今回構築したロシア語翻訳データにも情報構造のアノテーションを施せば，日本語・ロシア語の情報構造についての対照研究が定量的に行えるようになる(14)．日本語もロシア語も共に顕在的な冠詞のない言語であるため，定性や特定性などの冠詞を持つ言語では冠詞が担う機能をどのように表現するか (またはしないか)(15)を今後詳細に検討したい．

謝　辞

本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(プロジェクトリーダー: 浅原正幸) の研究成果である．また，JSPS 科研費 (課題番号: 17J07534) の助成を受けている．

文　献

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced corpus of contemporary written Japanese.” *Language Resources and Evaluation*, 48:2, pp. 345–371.

Aleksandr V. Isačenko (1966). “O grammatičeskom porjadke slov.” *Voprosy Jazykoznanija*, 6, pp. 27–34.

Kensaku Soejima (2017). “On expressions of agent de-topicalized intentional events: A contrastive study between Japanese and Russian.” *Journal of Japanese Linguistics*, 30:1, pp. 107–128.

Takuya Miyauchi, Masayuki Asahara, Natsuko Nakagawa, and Sachi Kato (2018). “Information-Structure Annotation of the “Balanced Corpus of Contemporary Written Japanese”.” *Computational Linguistics* Vol. 781. *Communications in Computer and Information Science.*, pp. 155–165. Singapore: Springer.

Masayuki Asahara (2017). “Between Reading Time and Information Structure.” *Proceedings of The 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31.*

Catherine V. Chvany (1973). “Notes on ‘root’ and ‘structure-preserving’ in Russian.” C. Corum, T.C. Smith-Stark, and A. Weiser (Eds.), *You take the high node and I will take the low node.* Chicago, IL: Chicago Linguistic Society. pp. 252–290.

Alan Timberlake (1975). “Hierarchies in the Genitive of Negation.” *The Slavic and East European Journal*, 19:2, pp. 123–138.

(14) さらに，Asahara (2017) は，情報構造と読み時間の関係を考察している．ロシア語のデータに対するアノテーションを充実させれば，同様の研究が可能となり，Asahara (2017) で示された成果が日本語特有のものか通言語的なものかを調べることも可能となる．

(15) 一般にロシア語は定性を語順によって表現する傾向があるとされる (Chvany 1973 など) が，あくまで傾向に過ぎない．また，否定属格の現象も定性との関係がよく指摘される (Timberlake 1975 など) が，これもはっきりとした規則ではない．

中古語における形容詞テ形をめぐって
—形容詞の意味分類との関わりから—

菊池 そのみ（筑波大学大学院人文社会科学研究科）

Te-form adjectives in Early Middle Japanese:
From the perspective of semantic classifications of adjectives

Sonomi Kikuchi (Graduate School of Humanities and Social Sciences, University of Tsukuba)

要旨

本稿は中古和文資料を対象として中古語における形容詞テ形の出現傾向を明らかにするものである。『日本語歴史コーパス平安時代編』を使用し、形容詞テ形と形容詞ゼロ連用形の用例を抽出して両者の比較から以下の 3 点を明らかにした。まず、形容詞の連用形全体に占める形容詞テ形の割合はおよそ 1 割であり、動詞の場合にはテ形が 9 割を超えることと対照的な結果が得られた。これに加えて現代語における同形式との比較によって通時的な変化についても問題を提起した。次に形容詞テ形の出現傾向は文章のスタイルに影響を受けないことを指摘した。最後に「あり」、「をり」などの存在動詞が後続する場合についてテ形の出現傾向を形容詞の意味分類を踏まえて分析した。その結果、テ形の場合には「感情」や「評価」を表す形容詞が多い一方でゼロ連用形の場合には「状態」を表す形容詞が多いということが明らかとなった。

１．はじめに

中古語においては形容詞の連用形[1]に接続助詞「て」が後接した形式と接続助詞「て」が後接しない形式とはどちらも（1a）、（2a）のようないわゆる副詞的な修飾や（1b）、（2b）のような文と文とをつなぐ接続を担うことができる。本稿は『日本語歴史コーパス平安時代編』（以下、「CHJ」）を使用し、これらの形式の出現傾向や形容詞の意味との関係などについて用例数及びその割合から中古語における傾向を捉えるものである。

（1）a. それを見れば、三寸ばかりなる人、いとうつくしうてゐたり。　（竹取物語）[2]
　　b. いとなやましうしたまひて、物などたえてきこしめさず、日を経て青み痩せたまひ御気色も変るを、内裏にもいづくにも思ほし嘆くに、いとどもの騒がしくて、御文だにこまかには書きたまはず。　（源氏物語・浮舟）

（2）a. 三の宮三つばかりにて中にうつくしくおはするを、こなたにぞ、また、とりわきておはしまさせたまひける、走り出でたまひて（略）　（源氏物語・横笛）
　　b. よそに見やりたてまつりつるだにはづかしかりつるに、いとあさましう、さし向ひきこえたる心地うつつともおぼえず。（枕草子・宮にはじめてまゐりたるころ）

[1] 活用語尾が「ク」、「シク」の場合とウ音便化した「ウ」、「シウ」の場合とを含む。

[2] 引用する用例の本文、作品名、章段名は全て『新編日本古典文学全集』による。なお、縦書きを横書きに改め、文の一部を引用する場合には省略部分に「(略)」と記した。また、用例中の下線は全て発表者が施したものである。

２．先行研究の整理と本稿の目的

初めに『源氏物語』における形容詞の連用形について検討した進藤（1978）[3]、接続助詞「て」の用法をまとめた山口（1998）、形容詞、形容動詞に接続助詞「て」が下接する形式と下接しない形式との比較から接続助詞「て」の機能について検討した竹部（2000）を取り上げる。次に先行研究を踏まえた上で本稿の目的を述べる。

２.１．先行研究の整理

まず、進藤（1978）は中古の資料として『源氏物語』を対象に「あさし」、「あさまし」、「あたらし（惜）」、「いとほし」、「うつくし」、「うれし」、「おもし」、「かぎりなし」、「ことごとし」、「くはし」、「つよし」という 11 の形容詞について、活用形別の用例数を比較した上で「「く（しく）」「う（しう）語尾の連用形」の使用例について検討している。この検討に際して形容詞の連用形の「構文上の性質」を「中止法」、「上文の述語形容詞（単一形容詞の場合も含む）が「て」助詞を介して重文関係に立つもの」、「形容詞連用形が「あり」「侍り」「おはします」等の補語となっているもの」、「形容詞連用形が「思ふ」「おぼす」「見る」「聞く」等の補語になっているもので、「形容詞終止形＋と」と同義なもの」、「形容詞連用形が「成る」「為（す）」等の語の補語となっているもの」、「形容詞連用形がいわゆる副詞的修飾をなしているもの」、「形容詞連用形が他の形容詞に上接している」ものの 7 つに分類し、用例数を集計している。その結果から「源氏物語の形容詞の連用形は下文の述語に構文的に結合して行く性格の活用形であり、その下文との係わり方は極めておおらかな性質にある如くであり、必ずしも副詞的修飾を主たる職能とするものではないようである」と述べ、更に当該の形容詞の「程度分量を表す性格の濃厚なものは自然に文意構成上副詞修飾に近い結合になることが多く、その語義が感情を表す性格の濃厚なものは自然に中止法や、あり、おぼゆ系の動詞の補語となる結合を生ずることが多いようである」と述べている（進藤 1978:20）。

次に山口（1998）は接続助詞「て」による接続法には「形容詞・形容動詞の連用法や副詞による成分、格助詞を伴う成分」などの「語的連用成分」とのつながりが認められる場合があることに着眼し、これを「「て」連用句」[4]と呼んだ。この「「て」連用句」の前件と後件との意味関係として「内容表示」、「批評表示」、「状態表示」、「空間表示」、「時間表示」、「方法表示」、「因由表示」の 7 つがあることを示している。形容詞類については「「て」連用句」と「連用法」（接続助詞「て」が下接しないもの）との比較から「「て」連用句」は主句とは「別個の主述関係として、より自覚的な判断を担った」とし、一方で「連用法」は「主句内の連用成分としての修飾語にとどまる」ことを述べた（山口 1998:243）。

竹部（2000）は『源氏物語』の第一部（桐壺巻～藤裏葉巻）を対象に形容詞、形容動詞に「助詞テ」[5]が下接する場合と下接しない場合とを比較することによって「助詞テ」の機能を明らかにした。まず、「助詞テ」の前件と後件との意味関係に着眼し、「助詞テ」が下接する場合には「因果関係」、「並立関係」、「全体部分関係」、「情態修飾関係」といった意味関係を表す用例が見られるのに対し、「助詞テ」が下接しない場合は上記の 4 つの意味関係に加え

[3] 進藤（1978）では近代文芸文として芥川龍之介の短篇作品を対象とした調査も実施した上で『源氏物語』を対象とした調査の結果と比較しているがここでは取り上げないこととする。

[4] 山口（1998）は「「て」連用句」を形成する形容詞、形容動詞、動詞、名詞（＋「にて」）について検討しており、形容詞に限った議論ではない。また、形容詞と形容動詞とを合わせて「形容詞類」としている。

[5] 竹部（2000）は「助詞テ」や「テ」と表記しているが、本稿で扱う接続助詞「て」の範囲と異なるものではないと考える。

て「知覚内容」、「評価」を表す用例が見られることを指摘した。また、それぞれの意味関係に該当する用例数（割合）の比較を通して「テは、前件が後件に直接係るところには現れず、前件だけでのまとまりの強い場合に現れると考えることができ」、「助詞テ」が持つ機能は「テの上接部分の叙述をいったんまとめ、したがって、その叙述を後続の叙述に対して独立性の高いものとしてその部分でいったん切り、更に後続の叙述へと接続する」というものであると述べた（竹部 2000:270）。

２.２. 本稿の目的

上述の先行研究において形容詞に接続助詞「て」が下接する形式（以下、「テ形」）と下接しない形式（以下、「ゼロ連用形」）のそれぞれの用法については用例に基づいた整理、考察が進められてきた。上述の先行研究を踏まえれば接続助詞「て」の機能は前件のまとまりを形成することと捉えられるが（1a）のようにいわゆる副詞的修飾を担うテ形における接続助詞「て」の機能はこのような説明で事足りるのであろうか。また、ゼロ連用形については進藤（1978）が形容詞の意味によって副詞的な修飾をとりやすいか中止法になりやすいかという違いがあることを指摘しており、テ形の場合にも同様に検討することによってテ形とゼロ連用形との差がどのような部分に現れるのかが明らかになる。

本稿は CHJ を使用し、テ形とゼロ連用形の出現の傾向や出現する環境（資料、本文種別による差や前後の要素）やテ形の出現傾向と形容詞の意味分類との関わりについて用例数及び全体に占める割合から中古語における傾向を捉えることを目的とする。これはテ形とゼロ連用形との違いを検討する上で基礎的なデータを提示するという意義を持つものである。これまでの研究においては特に助詞や特定の活用形を対象とする場合には索引を使用した調査によって大量の用例を精確に拾うことは困難であり、限られた資料を対象に調査されてきた。しかし、本稿の調査では CHJ を使用することにより、CHJ 所収の平安和文 16 資料の全文を対象に検索することが可能となり、全体の傾向を捉え得ることが期待される。また、現代語の同形式と比較することによりテ形とゼロ連用形との通時的な変化についても検討するための下地となる。

３. 研究方法

本稿では CHJ（データバージョン 2018.3）を使用し、中古和文 16 資料からテ形の用例とゼロ連用形の用例とを抽出した[6]。検索条件式は以下（3）、（4）、（5）に示した通りである。

（3）キー: (品詞 LIKE "形容詞%" AND 活用形 LIKE "連用形%")
IN subcorpusName="平安" AND core="true"
WITH OPTIONS tglKugiri="" AND tglBunKugiri="#" AND limitToSelfSentence="1" AND tglWords="100" AND unit="1" AND encoding="UTF-16LE" AND endOfLine="CRLF"

（4）キー: (品詞 LIKE "形容詞%" AND 活用形="連用形-補助")
IN subcorpusName="平安" AND core="true"
WITH OPTIONS tglKugiri="" AND tglBunKugiri="#" AND limitToSelfSentence="1" AND tglWords="100" AND unit="1" AND encoding="UTF-16LE" AND endOfLine="CRLF"

[6] コーパス検索アプリケーション「中納言」（ver. 2.4.2）を使用し、言語単位は短単位を用いた。

（5）キー: (品詞 LIKE "形容詞%" AND 活用形 LIKE "連用形%")
AND 後方共起: (語彙素="て" AND 品詞="助詞-接続助詞") ON 1 WORDS FROM キー
IN subcorpusName="平安" AND core="true"
WITH OPTIONS tglKugiri="" AND tglBunKugiri="#" AND limitToSelfSentence="1" AND tglWords="100" AND unit="1" AND encoding="UTF-16LE" AND endOfLine="CRLF"

まず、(3）の式によって形容詞の連用形全体 16,609 例を抽出し、次に（4）の式によって形容詞の連用形のうち補助活用 1,066 例を抽出し、続いて（5）の式によってテ形 1,732 例を抽出した。最後に（3）の式によって得られた用例数から（4)、(5）の式によって得られた用例数を除くことによってゼロ連用形 13,811 例を得た。なお、補助活用の「かり」は対象から除外したため[7]、対象とする用例数の合計はテ形 1,732 例とゼロ連用形 13,811 例とを足し合わせた 15,543 例である。形容詞の語数として見た場合に延べ語数は 15,543 語であり、異なり語数は 520 語である。分析においては各形容詞についてテ形の用例数とゼロ連用形の用例数とを足し合わせた数における両者の割合を算出することによってテ形とゼロ連用形とのどちらを取りやすいのか（以下、「出現傾向」）を検討する方法を採った。

４．調査の結果とその分析

ここでは調査の結果を示すと共に結果に基づく分析を試みる。初めに全体の傾向を概観し、中古語におけるテ形の出現傾向を捉える。次にテ形の出現傾向と形容詞の意味分類との関わりについて検討する。

４.１． テ形の出現傾向の概観

まず、全体の傾向について述べる。本節では試みに動詞の連用形に接続助詞「て」が下接した形式と下接しない形式についても調査を実施し、その結果を形容詞の結果と比較する場合がある。以下では動詞の連用形に接続助詞「て」が下接した形式を「動詞テ形」、下接しない形式を「動詞ゼロ連用形」と呼ぶ。また、形容詞と動詞との混乱を防ぐために本節に限って形容詞の場合にも「形容詞テ形」、「形容詞ゼロ連用形」と呼ぶことがある。

４.１.１． 形容詞テ形の出現傾向と動詞テ形の出現傾向との比較

まず、対象とした 15,543 例のうち、テ形は 1,732 例（11.14%）であり、ゼロ連用形は 13,811 例（88.86%）であることから活用形としての連用形全体を見れば 9 割弱がゼロ連用形であるということが分かる。これに対して動詞についても同様に算出した結果、動詞テ形が 80.01%であり、動詞ゼロ連用形が 19.99%であることから形容詞の場合と比較するとテ形とゼロ連用形との関係が逆転していることが読み取れる[8]。これについては山口（1998:217）が

[7] 補助活用の連用形「かり」（CHJ における「連用形-補助」）は助動詞を後接させる「くあり」の形が変化したものである。1,060 例が「かり」に助動詞が後接する用例であり、「かりて」の用例はなかった。進藤（1978:24）が「助動詞に連なるための活用形カリ活用は、述語的構文であるので副詞的修飾であるか述語的用法であるかの調査の対象にはならない」と述べていることを踏まえ、調査の対象から除外した。

[8] 用例抽出にあたっては複合動詞の前部要素を排除するため、言語単位に長単位を用いた。また、動詞連用形 58,200 例のうち、助動詞が後続する 27,934 例を除いた結果、動詞テ形 24,216 例、動詞ゼロ連用形 6,050 例が得られた。用例の確認は行っていないため、数値は参考程度に示すこととする。

形容詞の「連用法」が担う働きの一部について「動詞の場合は、述語性の強さから連用法には立ちにくいため、「て」連用句で補完することになった」と述べていることと整合する結果である。また、いわゆる叙述を担う活用形である終止形を全ての形容詞が有している訳ではないという点も踏まえておく必要があろう[9]。特に安本（2009:119）が中古語の形容詞についての調査を踏まえて「連体修飾用法・連用修飾用法はほぼすべての形容詞が持つが、述語用法を持つ語は全体の 75%だけである」と述べていることからも中古語の形容詞は修飾用法を中心に持っていたことが窺える。

更に試みに『現代日本語書き言葉均衡コーパス』（以下、「BCCWJ」）と『日本語話し言葉コーパス』（以下、「CSJ」）とを使用し[10]、現代語の形容詞、動詞それぞれのテ形と連用形との割合を算出した。前述の CHJ を使用して算出した結果と併せて表 1 に示す。BCCWJ の「出版-新聞」、「出版-雑誌」、「出版-書籍」のコアデータを対象として算出した結果、形容詞テ形は 7.73%であり、形容詞ゼロ連用形は 92.27%であることから中古語と同様にゼロ連用形が 9 割を占めることが読み取れる。一方で動詞テ形は 54.16%であり、動詞ゼロ連用形は 45.84%であることから中古語の動詞に比べて動詞テ形と動詞ゼロ連用形との比率の差が小さくなっていることが分かる。また、CSJ 全体のコアデータを対象として算出した結果、形容詞テ形は 21.01%であり、形容詞ゼロ連用形は 78.99%であった。また、動詞テ形は 82.01%であり、動詞ゼロ連用形は 17.99%であることから CSJ を使用した調査において形容詞は CHJ、BCCWJ と概ね同様の傾向が見られたが動詞は CHJ の中古語の結果と近い傾向を示すという結果が得られた。ここで提示した結果については、更に詳細な検討が求められる。形容詞、動詞ともに接続助詞「て」の用法や活用形としての連用形の文中での機能の変化などを視野に入れる必要がある。また、本稿の調査において対象とした中古和文資料が中古語のいかなる面を反映した言語資料であるのかについても議論の余地がある[11]。

表 1　テ形とゼロ連用形との出現傾向

	中古語（CHJ・コア）				現代語（BCCWJ・出版・コア）				現代語（CSJ・コア）			
	形容詞		動詞		形容詞		動詞		形容詞		動詞	
テ形	1732	11.14%	24216	**80.01%**	196	7.73%	6414	54.16%	411	21.01%	9517	**82.01%**
ゼロ連用形	13811	**88.86%**	6050	19.99%	2338	**92.27%**	5428	45.84%	1545	**78.99%**	2087	17.99%
計	15543	100.00%	30266	100.00%	2534	100.00%	11842	100.00%	1956	100.00%	11604	100.00%

[9] これについてはまず、終止形を有する形容詞群について調査を実施した新里（1983）の研究がある。更に形容詞ごとに連用形、終止形、連体形の出現頻度に差があることを指摘した吉田（1990）や活用形の出現頻度と形容詞自体の意味との関係に着眼した吉田（1995）、安本（2009）などに代表されるような古典語に関する研究がある。安本（2009）では連用形、終止形、連体形といういわゆる活用形を更に「構文的機能」として捉え直して検討している。なお、現代語においても活用形ごとの出現頻度が形容詞によって異なっているということが橋本・青山（1992）やそれに続く宮島（1993）によって示されている。

[10] CHJ と同様にコーパス検索アプリケーション「中納言」（ver. 2.4.2）を使用した。BCCWJ における検索は形容詞の抽出には短単位を、動詞の抽出には複合動詞の前項や「ていく」、「てみる」などの形式を排除するために長単位を用いた。CSJ は中納言を使用すると長単位を用いた検索はできないため形容詞の抽出も動詞の抽出も短単位を用いた（ただし、複合動詞の前項となる連用形や助動詞が下接するものや「ている」、「てみる」といった補助動詞を除いて集計した。形容詞の抽出に際しては助動詞が下接するものを除いて集計した）。

[11] 例えば福島（2008:94）は文同士の関係表示や従属節の特徴に関する調査から平安和文資料が「口頭言語的な性格を持つ言語変種である」と述べている。

４.１.２. テ形の出現傾向と資料や本文種別との関わり

表 2 にテ形の出現傾向を資料ごとに示した。資料ごとに長さ（言語量）や成立時期、作者など考慮すべき点は多数あるが表 2 からテ形の割合とゼロ連用形の割合とに資料による差が存するように見える。物語や日記といった資料のジャンルによる差は明確でないものの各資料において使用される形容詞にも着眼して検討していく必要がある。

表 2　資料ごとのテ形の出現傾向

	テ形頻度	ゼロ連用形頻度	頻度計	テ形割合	ゼロ連用形割合
竹取物語	6	96	102	5.88%	94.12%
古今和歌集	11	140	151	7.28%	92.72%
伊勢物語	25	106	131	19.08%	80.92%
土佐日記	5	42	47	10.64%	89.36%
大和物語	34	227	261	13.03%	86.97%
平中物語	12	99	111	10.81%	89.19%
蜻蛉日記	119	557	676	17.60%	82.40%
落窪物語	159	800	959	16.58%	83.42%
枕草子	110	1185	1295	8.49%	91.51%
源氏物語	1021	8595	9616	10.62%	89.38%
紫式部日記	20	254	274	7.30%	92.70%
和泉式部日記	36	144	180	20.00%	80.00%
堤中納言物語	33	278	311	10.61%	89.39%
更級日記	33	282	315	10.48%	89.52%
大鏡	89	836	925	9.62%	90.38%
讃岐典侍日記	19	170	189	10.05%	89.95%
計	1732	13811	15543	11.14%	88.86%

表 3　「本文種別」ごとのテ形の出現傾向

	テ形頻度	ゼロ連用形頻度	頻度計	テ形割合	ゼロ連用形割合
地の文ほか	1211	8895	10106	13.61%	88.02%
会話	453	4301	4754	10.53%	90.47%
会話-発話引用	24	201	225	11.94%	89.33%
歌	34	335	369	10.15%	90.79%
古注-歌	0	2	2	0.00%	100.00%
詞書	4	27	31	14.81%	87.10%
手紙	6	50	56	12.00%	89.29%
計	1732	13811	15543	11.14%	88.86%

また、表 3 にテ形の出現傾向を CHJ における「本文種別」ごとに示した。「本文種別」は文章のスタイルに対応するものと捉えられる。表 3 から「本文種別」によるテ形の出現傾向の大きな差は認められない。この結果から中古語におけるテ形は地の文、会話文、和歌などの文章のスタイルに関わらず、おおよそ同様の出現傾向を示すということが読み取れる。

これらの観点については中世以降の資料を対象とした調査においても同様の傾向が捉えられるのかを検討することも課題のひとつである。

４.１.３． 後続する要素

表 4 にテ形、ゼロ連用形に後続する要素を品詞ごとに示した。品詞の認定は CHJ の品詞の「大分類」に従っている。ゼロ連用形は形容詞や動詞が後続する割合がテ形に比べてある程度多いことが窺える。これはゼロ連用形がいわゆる副詞的な修飾としての用法のバリエーションをテ形よりも多く持つという竹部（2000）の指摘と関係していることが予測される。

表 4　後続する要素

	テ形		ゼロ連用形	
	頻度	割合	頻度	割合
名詞	119	6.87%	884	6.40%
代名詞	4	0.23%	42	0.30%
副詞	17	0.98%	88	0.64%
形容詞	36	2.08%	1864	13.50%
形状詞	3	0.17%	254	1.84%
動詞	424	24.48%	6393	46.29%
助詞	169	9.76%	2030	14.70%
助動詞	1	0.06%	0	0.00%
感動詞	0	0.00%	2	0.01%
接頭辞	11	0.64%	330	2.39%
補助記号	946	54.62%	1912	13.84%
空白	2	0.12%	12	0.09%
計	1732	100.00%	13811	100.00%

４.２． テ形の出現傾向と形容詞の意味分類

次にテ形の出現傾向と形容詞の意味分類との関係について検討する。以下では算出した割合を以て比較する場合に用例数が少ない（割合が極端な値を示す可能性が高い）形容詞による影響を受けないようにテ形とゼロ連用形との用例数の合計が 15 例を超えている形容詞を考察の対象とした。用例数の合計が 15 例を超える形容詞は異なり語数では 170 語であり、延べ語数（延べ用例数に対応する）は 13,989 語である。これらが連用形全体に占める割合は異なり語数で見ると 32.69%であり、延べ語数で見ると 90.00%である[12]。

[12] 例えば『枕草子』における形容詞の活用形に関する研究であるところの吉田（1990）は当該の形容詞の総使用頻度が 10 以上のものを対象としている（全体の数は示されていないため、全体に占める割合は算出できない）。また、『源氏物語』を対象として形容詞の文中における機能と形容詞自体の意味分類との関係について調査した安本（2009）は当該の形容詞の使用頻度が 20 以上のものを対象としている。この場合に対象となる形容詞の全体に占める割合は延べ語数で見ると 87.15%であり、異なり語数で見ると 24.88%である。対象とする用例数及び全体に占める割合については引き続き検討する必要があるが先行研究の調査方法を踏まえれば傾向を捉える上で大きく不足するものではないと考える。

４.２.１. 形容詞の意味分類

まず、本稿における形容詞の意味分類を示す。本稿では基本的に安本（2009）の意味分類に従うこととし、「感情」、「評価」、「程度」、「感覚」、「状態」、「否定」の 6 項目を立てた。安本（2009）は『源氏物語』を対象に当該の形容詞の文中における機能とその形容詞自体の意味との関係について調査した研究であり、先行研究に基づいた形容詞の意味分類が示されている。ただし、安本（2009）において言及がない語については用例を概観した上で分類した[13]。また、安本（2009）は「つねなし」、「かぎりなし」のように「～なし」の形をとる形容詞の一部を「否定」に分類しているが本稿では「なし」のみを「否定」の形容詞とし、安本（2009）において「否定」とされている「～なし」の形の形容詞については「否定」以外の項目に分類した[14]。

各意味分類に該当する形容詞の一部をテ形の用例数とゼロ連用形の用例数とを足し合わせた値が大きい順に示した。各分類に該当する語の末尾の括弧内にその分類に該当する形容詞の語数を示した。各意味分類に該当する形容詞の数と全体に占める割合は表 5 の通りである。形容詞の意味分類ごとに見ると「感情」、「評価」、「状態」がそれぞれ全体の 3 割程を占めている。この値は形容詞の意味分類によって大きく結果が左右されてしまうものであり、分類の妥当性も含めて引き続き検討する必要がある。また、用例数の多い形容詞のみを対象としていることにも留意する必要がある。

感情： あやし，をかし，いとほし，浅まし，悲し，口惜し，嬉し，心細し，恥づかし，心苦し，懐かし，苦し（クルシ），心安し，心憂し，わりなし，恋し，…【56 語】
評価： めでたし，こよなし，儚し，賢し（カシコシ），良し，つれなし，悪し（アシ），やんごとなし，詳し，はしたなし，厳めし，事々し，艶めかし，…【54 語】
程度： いみじ，限り無し，著し，凄まじ，言痛し，残り無し，論無し，譬え無し【8 語】
感覚： 辛し（ツラシ），痛し，辛し（カラシ），涼し【4 語】
状態： 多し，近し，疾し，深し，久し，若し，高し，遠し，長し，麗し，同じ，白し，繁し，捗々し，早し，暗し，重し，睦まし，気近し，少なし，所狭し，…【47 語】
否定： 無し【1 語】

表 5　形容詞の分類ごとの語数と全体に占める割合

	感情	評価	程度	感覚	状態	否定	計
形容詞の語数	56	55	8	4	46	1	170
全体に占める割合	32.94%	32.35%	4.71%	2.35%	27.06%	0.59%	100%

[13] 本稿は形容詞の意味分類について検討した研究として三田村（1967）、東辻（1970）、川端（1976）、吉田（1977）、西尾（1979）、阪倉（1985）、吉田（1995）を参考としている。

[14] 国立国語研究所コーパス開発センター（池上尚）（編）（2016:94-95）には CHJ において短単位を用いて検索する場合に「体言＋「ナイ（無い）」は「『岩波国語辞典』第 6 版、『日本国語大辞典』第 2 版のいずれかで見出しになっている」場合には 1 短単位として認められていることが記されている。また、「～なし」の形の形容詞には後部要素「なし」が非存在の「無し」である場合と程度の甚だしさを表す「甚し」である場合との両系があり、岩村（1995）や村田（2005）などによって詳細に議論されているところである。これらの議論を踏まえ、一括して「否定」に分類すべきではないと判断し、このように対処した。

４.２.２. テ形の出現傾向と形容詞の意味分類

続いてゼロ連用形との合計におけるテ形の出現傾向とその出現傾向（範囲）に該当する形容詞の語数を示したのが表 6 と図 1 である。当該の形容詞のテ形とゼロ連用形との割合を 5%ごとに区切り、示している。例えばテ形 0%の場合は当該の形容詞はテ形の用例が 0 例であり、用例はすべてゼロ連用形であることを示しており、テ形が 5%の場合は当該の形容詞は連用形の用例数全体の 5%がテ形であり、残りの 95%がゼロ連用形であることを示している。前述の通り、活用形としての連用形全体を見れば 90%近くがゼロ連用形であることからも分かるようにテ形が 0%～16.00%の間に集中している。

表 6　テ形の出現傾向と形容詞の語数

テ形の出現傾向	0.00%	～6.00%	～11.00%	～16.00%	～21.00%	～26.00%	～31.00%	～36.00%	～41.00%	～46.00%	計
語数	33	45	29	24	14	13	5	5	0	2	170
全体に占める割合	19.41%	26.47%	17.06%	14.12%	8.24%	7.65%	2.94%	2.94%	0.00%	1.18%	100.00%

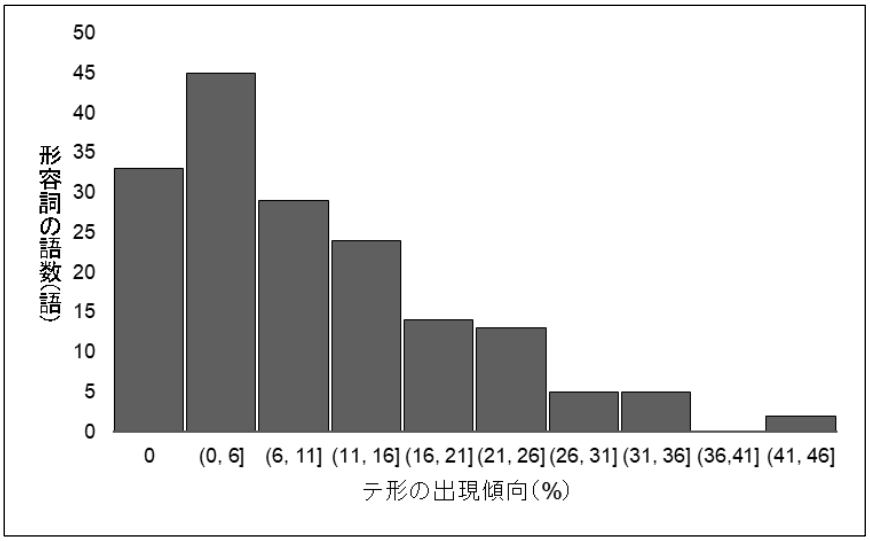


図 1　テ形の出現傾向と形容詞の語数との対応

表 7　テ形の出現傾向と形容詞の意味

		テ形の出現傾向										計
		0.00%	～6.00%	～11.00%	～16.00%	～21.00%	～26.00%	～31.00%	～36.00%	～41.00%	～46.00%	
感情	語数	6	12	9	9	8	7	2	2	0	1	56
	割合	3.53%	7.06%	5.29%	5.29%	4.71%	4.12%	1.18%	1.18%	0.00%	0.59%	
評価	語数	14	15	11	4	2	4	3	1	0	1	55
	割合	8.24%	8.82%	6.47%	2.35%	1.18%	2.35%	1.76%	0.59%	0.00%	0.59%	
感覚	語数	1	1	1	0	0	1	0	0	0	0	4
	割合	0.59%	0.59%	0.59%	0.00%	0.00%	0.59%	0.00%	0.00%	0.00%	0.00%	
程度	語数	4	4	0	0	0	0	0	0	0	0	8
	割合	2.35%	2.35%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
状態	語数	8	13	8	11	4	0	0	2	0	0	46
	割合	4.71%	7.65%	4.71%	6.47%	2.35%	0.00%	0.00%	1.18%	0.00%	0.00%	
否定	語数	0	0	0	0	0	1	0	0	0	0	1
	割合	0.00%	0.00%	0.00%	0.00%	0.00%	0.59%	0.00%	0.00%	0.00%	0.00%	
テ形の出現傾向ごとの語数の合計		33	45	29	24	14	13	5	5	0	2	170

また、表 5 と表 6 とを併せてテ形の出現傾向を形容詞の意味分類ごとに示したのが表 7 である。表 7 からはテ形の出現傾向が形容詞の意味分類によって大きく異なるといった結果は見受けられなかった。つまり、テ形の出現傾向と形容詞の意味との間には明確な対応関係が認められなかった。

４.２.３. 存在動詞が後続する場合のテ形の出現傾向と形容詞の意味分類

次に「あり」、「をり」などの存在動詞がテ形とゼロ連用形とに後続する場合について検討する[15]。ここで存在動詞を取り上げた理由は 2 点ある。1 点目は先行研究において挙げられている用例などを踏まえると存在動詞が後続する場合に中古語のテ形が（6a）のようにいわゆる副詞的な修飾の用法を持つことが多いと考えられることである。このテ形の副詞的な修飾の用法は前述の通り、先行研究の接続助詞「て」の機能に関する説明においては充分でなく、形容詞の意味に基づいた分析が求められる。2 点目はこのような副詞的な修飾の用法のテ形は現代日本語には見られないということである[16]。（6a）と同じ意味で（6b）のように「かわいくて座っている」と言うことはできず、（6c）のように「かわいらしい様子」でなどに言い換える必要があるだろう。

（6）a. それを見れば、三寸ばかりなる人、いとうつくしうてゐたり。
（竹取物語、再掲（1a））
b. #それを見ると三寸ほどの大きさの人がとてもかわいくて座っている。
c. それを見ると三寸ほどの大きさの人がとてもかわいらしい様子で座っている。

また、現代語の形容詞を対象にテ形と形容詞ゼロ連用形との用法の比較を行った津留崎（2003:25）は（7）のように「主体の主たる状態や動作が後続句節の動詞述語で表され，それと同時に存在する主体の副次的な状態を，先行句節において述べるもの」[17]を「〈副状態〉」と呼び、ゼロ連用形のみが担う用法であってテ形への交替は不可能であると指摘している。これが中古語においては（6a）のように許容されていたと考えられる。

（7）a. 花子は声もなく，立ちつくしている。（津留崎 2003:25、(93)）[18]
b. そういう会話を，私はお茶を持って行きながら寂しく聞いていた。（父）
（津留崎 2003:25、(94)）

これらのことは中古語から現代語に至る過程においてテ形、ゼロ連用形の用法にも変化が生じたことを示唆しており、存在動詞が後続する場合に着目することによって通時的な変化を捉えるための基礎的なデータを集めることができる。

ここでは存在動詞として「あり」、「ゐ」、「をり」、「はべり」、「おはす」、「おはします」、「さぶらふ」、「ものす」が後続する用例を対象とした。なお、「ものす」が後続する場合には

[15] 存在動詞が後接するもののみを対象としており、助詞などの介入やテキストに読点が施されている場合は含めていない。

[16] 現代語においても「大変お若くていらっしゃいますね」（作例）など一部、許容される場合があるように見受けられる。しかし、場面や共起する語が限られるため、ここでは措くこととする。

[17] 津留崎（2003:25）の原文の該当箇所は斜体である。

[18] (7) の引用箇所の下線や四角囲みは原文のままである。

動作を表す用例は対象から外した。存在動詞が後続する形容詞についてテ形とゼロ連用形のそれぞれにおいて頻度順位の上位 12 語を示したものが表 8 である。表 8 の結果からテ形は「感情」や「評価」の形容詞が多いのに対してゼロ連用形は「状態」の形容詞が多く、上位を占めていることが分かる。以下では用例を概観し、分析を試みる。

表 8　存在動詞が後続する形容詞

テ形					ゼロ連用形				
頻度順位	形容詞	意味分類	用例数	割合	頻度順位	形容詞	意味分類	用例数	割合
1	無し	否定	26	30.95%	1	多し	状態	59	8.15%
2	をかし	感情	5	5.95%	2	近し	状態	45	6.22%
2	うつくし	感情	5	5.95%	3	無し	否定	35	4.83%
4	やむごとなし	評価	4	4.76%	4	良し	評価	34	4.70%
4	心細し	感情	4	4.76%	5	めでたし	評価	32	4.42%
6	つれなし	評価	3	3.57%	6	幼し	状態	24	3.31%
6	多し	状態	3	3.57%	6	若し	状態	24	3.31%
8	若し	状態	2	2.38%	8	いみじ	程度	23	3.18%
8	もの儚し	評価	2	2.38%	8	久し	状態	23	3.18%
8	心安し	感情	2	2.38%	9	深し	状態	19	2.62%
8	麗し	評価	2	2.38%	10	賢し(カシコシ)	評価	12	1.66%
8	あやし	感情	2	2.38%	11	苦し(クルシ)	感情	11	1.52%
…(略)…					…(略)…				
合計			84	100%	合計			724	100%

まず、表 8 からテ形においては「なし」が全体の 3 割程を占めていることが分かる。(8)、(9) のようにいわゆる副詞的修飾の用法である。

(8) 朔日の装束はとらざりければ、さりげもなくてあれどはだか姿は忘られず、おそろしきものから、をかしうともいはず。（紫式部日記）

(9) たはぶれにも御気色のものしきをば、いとわびしと思ひてはんべめるを、いと大きなることなくてはべらむには、御気色など見せたまふな。（蜻蛉日記・中巻）

次に「感情」の形容詞である「をかし」、「うつくし」のテ形について検討する。(10) ～ (13) のように「をかし」や「うつくし」は後続の動詞が表す主体の状態を修飾している。これは山口 (1998) の「状態表示」、竹部 (2000) の「情態修飾関係」にそれぞれ該当する。

(10) 聞こえさせつることの残りもまだいと多かり。艶にをかしうてはべりし。まめやかに聞こえさせはべらむ。（落窪物語・巻之一）

(11) 実方の兵衛佐、長命侍従など、家の子にて、いますこし出で入りなれたり。まだ童なる君など、いとをかしくておはす。（枕草子・小白川といふ所は）

(12) 尚侍の君つと抱きもちてうつくしみたまふに、とう参りたまふべきよしのみあれば、五十日のほどに参りたまひぬ。女一の宮一ところおはしますに、いとめづらし

くうつくしうておはすれば、いといみじう思したり。　　（源氏物語・竹河）

（13）いと若くきよらにて、かく御賀などいふことは、ひが数へにやとおぼゆるさまの、なまめかしく人の親げなくおはしますを、めづらしくて、年月隔てて見たてまつりたまふは、いと恥づかしけれど、なほけざやかなる隔てもなくて、御物語聞こえかはしたまふ。幼き君もいとうつくしくてものしたまふ。　　（源氏物語・若菜上）

一方で同じく「感情」の形容詞である「心細し」は状態の描写ではなく、（14）、（15）のように心情の描写であると言える。

（14）父はただ、われをおとなにし据ゑて、われは世にも出で交らはず、かげにかくれたらむやうにてゐたるを見るも、頼もしげなく心ぼそくおぼゆるに、きこしめすゆかりある所に、「なにとなくつれづれに心ぼそくてあらむよりは」と召すを、古代の親は、宮仕人はいと憂きことなりと思ひて過ぐさするを、「今の世の人は、さのみこそは出でたて。（略）　　（更級日記・宮仕えの記）

（15）ただ、常に候ふ侍従、弁などいふ若き人々のみ候へば、年に添へて人目まれにのみなり行く古里に、いと心細くておはせしに、右大将の御子の少将、知るよしありて、いとせちに聞こえわたりたまひしかど、かやうの筋は、かけても思し寄らぬことにて、（略）　　（堤中納言物語・思はぬ方に泊りする少将）

また、「状態」の形容詞の場合も（16）～（19）のようにいわゆる副詞的な修飾の用法が多い。しかし、中には（20）のように「人気多し」が「あらぬさまなり」（ここでは「これまでとは違う別世界である」といった意）と判断させる原因であるものもある。

（16）皆ののしりて、さざとして出でたまふすなはち、あこぎ告げに走らせやりたれば、少将、心地たがひて、例乗りたまふ車にはあらぬに、朽葉の下簾かけて、男ども多くておはしぬ。帯刀馬にてさきだちておこせたまへり。　　（落窪物語・巻之二）

（17）返りごとものして、いととげにあめれど、よにもあらじ、いまは人知れぬさまになりゆくものをと思ひ過ぐして、あさましううちとけたること多くてあるところに、午時ばかりに、「おはしますおはします」とののしる。　　（蜻蛉日記・下巻）

（18）かくいふほどに、年もかへりにけり。君の御方に若くて候ふ男、このましきにやあらむ、定めたるところもなくて、この童に言ふ。（堤中納言物語・ほどほどの懸想）

（19）このおとどは、基経のおとどの太郎なり。御母、四品弾正尹人康親王の御女なり。醍醐の帝の御時、このおとど、左大臣の位にて年いと若くておはします。菅原のおとど、右大臣の位にておはします。その折、帝御年いと若くおはします。
（大鏡・天　左大臣時平）

（20）浦島の子が心地なん。おはしまし着きたれば、殿の内悲しげもなく、人気多くてあ

らぬさまなり。御車寄せておりたまふを、さらに古里とおぼえず疎ましううたて思さるれば、とみにもおりたまはず。（源氏物語・夕霧）

続いてゼロ連用形の用例を概観する。頻度の順位が高い「多し」と「近し」とはいずれも「状態」の形容詞であり、(21)～(24)のようにいわゆる副詞的修飾の用法である。山口(1998)は「多し」を「状態表示」、「近し」を「空間表示」としている。また、(25)、(26)のように「なし」も同様に副詞的修飾と見られる。

(21) いとうれしくて、「かしこき御影に別れたてまつりにしこなた、さまざま悲しきことのみ多くはべれば、今はこの渚に身をや棄てはべりなまし」と聞こえたまへば、「いとあるまじきこと。これはただいささかなる物の報いなり。(源氏物語・明石)

(22) 御方々、君達、上人など、御前に人のいとおほく候へば、廂の柱に寄りかかりて女房と物語などしてゐたるに、物を投げ給はせたる、あけて見たれば、「思ふべしやいなや。人、第一ならずはいかに」と書かせたまへり。
（枕草子・御方々、君達、上人など、御前に）

(23) むかし、なま心ある女ありけり。男近うありけり。女、歌よむ人なりければ、心みむとて、菊の花のうつろへるを折りて、男のもとへやる。(伊勢物語・十八　白菊)

(24) 持ちしらぬ物設けて、ついたてて、入り臥し入り臥しすることよ」とのたまへば、おとどは、「近くおはしてのたまへ」とのたまへば、いらへ遠くなりぬれば、はての言葉は聞こえず。（落窪物語・巻之一）

(25) 内裏には御物の怪のまぎれにて、とみに気色なうおはしましけるやうにぞ奏しけむかし。見る人もさのみ思ひけり。（源氏物語・若紫）

(26) この家つくりはべること二年なり。そのほどまでは、音なくはべりて、かく妨げさせたまへば、いと安からずなむ嘆き申したまふ」と申せば、(略)
（落窪物語・巻之三）

５．おわりに

最後に本稿の結論と今後の課題とを述べる。

５.１．本稿の結論

本稿では大きく以下の 3 点を明らかにした。

まず、CHJ を使用した中古和文 16 資料を対象とした調査の結果によれば中古語の形容詞の連用形のうち、テ形は 11.14%であり、ゼロ連用形は 88.86%であることを示した。同様に中古語の動詞の場合には動詞テ形が 80.01%であり、動詞ゼロ連用形が 19.99%であることから形容詞の場合の両者の関係が逆転しているということを指摘した。これはそもそも中古語の形容詞が動詞に比べて「述語性」が低く、修飾用法を中心に持つという先行研究の記述によって説明できる。次にテ形の出現傾向は資料によって違いがあるが「本文種別」による

違いはないことを明らかにした。最後にテ形の出現傾向と形容詞の意味分類との関わりについて存在動詞が後続する場合を取り上げて検討を行った。その結果、テ形は「感情」の形容詞が多いのに対し、ゼロ連用形は「状態」の形容詞が多いことが明らかとなった。「感情」の形容詞のうちでも（いわゆる評価の意味が強い）状態の描写の場合と感情の描写の場合とがあることを概観した。存在動詞が後続する場合にはテ形の一部の例を除いてテ形、ゼロ連用形ともにいわゆる副詞的な修飾の用法を担っていることが分かる。テ形とゼロ連用形との違いを検討する上では形容詞自体の意味分類と後続する動詞のタイプとに留意する必要があることが改めて示された。

５.２． 今後の課題

本稿ではテ形の出現傾向と形容詞の意味との関わりについて後続する動詞が存在動詞であるもののみを取り上げたが進藤（1978）がゼロ連用形の用法として指摘した「形容詞連用形が「思ふ」「おぼす」「見る」「聞く」等の補語になっているもの」や「形容詞連用形が「成る」「為（す）」等の語の補語となっているもの」についても同様に検討してみる必要がある。

また、いわゆる副詞的な修飾や文と文とをつなぐ接続といった用法と形容詞の意味分類との関わりについても検討する必要がある。これについては「うつくしう」のようにウ音便化した連用形としていない連用形との比較によっても検討することが課題のひとつとして挙げられる。

調査資料

国立国語研究所（2016）『日本語歴史コーパス平安時代編』
http://pj.ninjal.ac.jp/corpus_center/chj/heian.html（2018/07/16 閲覧）
国立国語研究所（2017）『現代日本語書き言葉均衡コーパス』
http://pj.ninjal.ac.jp/corpus_center/bccwj/（2018/07/19 閲覧）
国立国語研究所（2018）『日本語話し言葉コーパス』
http://pj.ninjal.ac.jp/corpus_center/csj/（2018/07/20 閲覧）

参考文献

岩村恵美子（1995）「ナシ（甚）型形容詞―否定性接尾語を有する形容詞の考察―」『語文』，64，pp.12-25，大阪大学国文学研究室.
川端善明（1976）「用言」『岩波講座日本語 6 文法Ⅰ』，pp.169-217，岩波書店.
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』，国立国語研究所コーパス開発センター.
阪倉篤義（1985）「歌ことばの一面」『文学・語学』，105，pp.51-63，全国大学国語国文学会.
新里博樹（1983）「終止形を有する形容詞群の考察」『国語研究』，46，pp.39-50，国学院大学国語研究会.
進藤義治（1978）「源氏物語の形容詞の連用形の機能について」『南山国文論集』，3，pp.7-24，南山大学.
竹部歩美（2000）「中古のテについて―形容詞・形容動詞に下接する場合に着目して―」『国學院大學大學院紀要文学研究科』，32，pp.255-273，国学院大学大学院.
津留崎由紀子（2003）「形容詞の中止形を用いた複文における先行句節と後続句節の関係」『日本語科学』13，pp.7-32，国立国語研究所.
西尾光雄（1979）「源氏物語の形容詞について」『東京女子大学日本文学』，51，pp.1-16，東

京女子大学.
橋本三奈子・青山文啓（1992）「形容詞の三つの用法―終止、連体、連用」『計量国語学』, 18:5，pp.201-214，計量国語学会.
東辻保和（1970）「古典語感情形容詞研究の一視点」『文学・語学』，56，pp.80-91，全国大学国語国文学会.
福島直恭 (2008)『書記言語としての「日本語」の誕生―その存在を問い直す―』, 笠間書院.
三田村紀子（1967）「形容詞の意味分類」『研究年報』，10，pp.14-25，奈良女子大学文学会.
宮島達夫（1993）「形容詞の語形と意味」『計量国語学』，19:2，pp.94-104，計量国語学会.
村田菜穂子（2005）『形容詞・形容動詞の語彙論的研究』，和泉書院.
安本真弓（2009）「構文的機能から見た中古形容詞の特徴―意味との関わりから―」『国語学研究』, 48, pp.119(28)-105(42)， 東北大学大学院文学研究科国語学研究室内「国語学研究」刊行会.
山口堯二（1998）「中古語「て」連用句とその周辺」佐藤喜代治（編）『国語論究第 7 集中古語の研究』，pp.211-247，明治書院.
吉田金彦（1977）「古代語形容詞の語構成」『吉田金彦著作選 8 動詞・形容詞』，pp.341-377，明治書院.
吉田光浩（1990）「主成分分析法による形容詞の活用分析―『枕草子』を資料として―」『大妻国文』，21，pp.1-16，大妻女子大学国文学会.
吉田光浩（1995）「平安期形容詞の意味と終止用法について―『枕草子』『源氏物語』『栄花物語』を資料として―」宮地裕・敦子先生古稀記念論集刊行会（編）『日本語の研究宮地裕・敦子先生古稀記念論集』， pp.112-145，明治書院.

日本語文における連用修飾語成分に見られるパラレルについての一考察
―「赤く変わる」と「赤に変わる」は同じか―

王棟（東京外国語大学大学院生）

要旨

本発表はこれまで注目されてこなかった日本語の連用修飾語に見られるパラレル「形容詞ク動詞」(以下 A ク V)と「名詞ニ動詞」(以下 N ニ V)の棲み分けに注目した論考である。本稿は大規模コーパス、現代日本語書き言葉均衡コーパス(BCCWJ)の用例に基づいて、「赤ク／赤ニ動詞」を「A ク／N ニ V」の例として考察した。その結果、「A ク V」と「N ニ V」の棲み分けについて以下のことを指摘する。

① 通常、動詞の補語と修飾語の位置に現れるのは形容詞「赤ク」であり、「赤ニ」の使用は稀である
② 「赤ニ」は色によって表される意味概念の対立のある文脈に現れやすい。
③ 「色の多様性に言及する場合」と「色変化の始まりがある場合」と「色が詳細な説明を受ける場合」において、「赤ニ」の使用は文法的な面において義務的と言ってよいが、事象を説明する機能的な面において「赤ニ」の補足を行っている。

1. はじめに

動詞述語文において、動詞にかかって事象の情態を表す連用修飾語がある。その多くは副詞と形容詞・形容動詞の連用形である。例えば、以下のような例が挙げられる。

(1) 牛肉を きれいに 切る。
(2) マイクを しっかりと 握る。
(3) 体が 小さく 見える。

このような連用修飾語成分には、「形容詞ク動詞」と「名詞ニ動詞」とのパラレルを成す特殊な例が存在している。日本語にはこのような「形容詞・名詞」の対を成す語彙がそれほど多くはないが、そのほとんどは色彩を表す語である。例えば

(4) 壁を 赤く／赤に 染める。
(5) 星が 赤く／赤に 光る。
(6) 光が 赤く／赤に 見える。

日本語の連用修飾語成分に見られる「形容詞ク動詞」と「名詞ニ動詞」とのようなパラレルはどのような棲み分けがあるのだろうか。本発表は日本語の連用修飾語成分にみられる「形容詞ク動詞」と「名詞ニ動詞」のパラレルに注目して、現代日本語書き言葉均衡コーパスのデータに見られる「赤ク動詞」(以下「赤ク V」)と「赤ニ動詞」(以下「赤ニ V」)の棲み分けを観察し、「赤ニ V」の使用条件を明らかにすることを目的とする。

2. 「文成分」と「共起動詞の種類」から見た「赤ク」と「赤ニ」

2.1 「文成分」から見た「赤ク」と「赤ニ」

日本語の学校文法では、文の成分として、主語・述語・連体修飾語・連用修飾語・独立語の五つが挙げられる。そのうち、「連用修飾語」の分析と再構築は国立国語研究所(1963)・鈴木(1972)・早津(2010)などが成されてきた。本稿は早津(2010)の分類と定義を援用する。

早津(2010)は連用修飾語を構文的な機能の異なる成分、「補語」・「修飾語」・「状況語」・「陳述語」・「接続語」の五つに分けている。「赤ニ V」・「赤ク V」において、「赤ニ」と「赤ク」

は「補語」と「修飾語」として働く場合がある。

早津(2010)では「補語」を「述語動詞の表す動きの成立に直接的・間接的に参加している事物であって述語動詞の意味によって一定の補語が要求されるのであり、補語と述語は切り離しがたい関係で対立しつつ補語が述語を補っている」と定義している。「赤ク V」と「赤ニ V」の例としては「赤クスル／ナル」と「赤ニスル／ナル」がそれに当たる。以下の例のように、補語としての「赤ク」・「赤ニ」と述語とは切り離しがたい関係にある。

(7)　髪を赤く／赤にする。
　　＊髪をする。

(8)　果皮が赤く／赤になる。
　　＊果皮がなる。

それに対して、「修飾語」は、「述語の表す動きや状態について、その様子・程度・量などを詳しく説明して述語をかざる成分」を指す。「赤ニ」・「赤ク」が「修飾語」となる場合、専ら述語動詞の表す動きや状態についてその様子を詳しく説明している。例としては「壁を赤く／赤に染める」などがある。

本節で述べたように、文成分に注目すれば、動詞にかかる「赤ク」と「赤ニ」は「補語」として働く場合があれば、「修飾語」として働く場合があるということがわかる。

2.2 本発表が注目する「共起動詞の種類」

前節で述べた通り、「補語」は「述語動詞の表す動き」に関わり、そして「修飾語」が詳しく説明しているのは「述語の表す動きや状態」に関わるという点に注目すると、述語となる動詞の性格を考慮に入れる必要があると考えられる。

日本語の動詞は「＜動作＞か＜変化＞か」と「＜主体＞か＜客体＞か」という観点を組み合わせてみると、次のように三分類することができる(奥田 1977)。

・主体動作・客体変化動詞：主体の動作を表すと同時に、客体の変化を捉えている動詞である。テイル形を取る場合、能動は動作継続を表し、受動は結果継続を表す。すべては他動詞である。
　例：染める、塗る、腫らす

・主体動作動詞：動作のみを捉えている動詞、自動詞も他動詞もある。それのテイル形は能動の場合でも受動の場合で動作の継続を表す。
　例：光る　燃える　輝く

・主体変化動詞：主体の変化を捉え、テイル形は結果継続を表す。
　例：染まる　腫れる　濁る

「赤ク」・「赤ニ」と三種類の動詞と共起する場合、以下のようになる。

「赤ク」・「赤ニ」が主体動作・客体変化動詞と共起する場合、変化する「客体」を色彩の面から説明している。

(9)　太郎が壁を赤く／赤に染める。

また、「赤ク」・「赤ニ」が主体動作動詞と共起する場合、動作をする「主体」を色彩の面から説明している。

(10)　灯りが赤く／赤に光る。

そして、「赤ク」・「赤ニ」が主体変化動詞と共起する場合、変化する「主体」を色彩の面から詳しく説明している。

(11)　海は赤く／赤に染まる。

以上の分析に基づいて、本発表は「文成分」と「共起動詞の種類」を考慮に入れ、実例調査を通して「赤ク」と「赤ニ」のふるまいを見ていく。

3. データの収集について

データの入手は現代日本語書き言葉均衡コーパス(データバージョン 1.1)と検索アプリケーション中納言(ver.2.4)を利用する。「赤ク」もしくは「赤ニ」が修飾語となる場合は、「赤く壁を染める」のように、「A ク N ヲ V」のように動詞と直接結びつかない場合がある。したがって、検索条件は「赤ク／赤ニ」の後十語以内に動詞が含まれるものに設定した。

この検索条件で、「赤ク V」の用例を 2460 件、「赤ニ V」の用例を 912 件入手した。手作業で「赤ク」・「赤ニ」が動詞とは係り受けの関係ではない場合(赤くも腫れてもない)と慣用表現の場合(朱に交われば赤くなる)などを除外する。以上の作業を通して、「赤ク動詞」を 1405 例、「赤ニ V」を 94 例入手した

4. 「赤ク V」と「赤ニ V」の分布 ―文の成分と動詞の種類―

4.1 全体的な傾向

まず、「赤ク V」と「赤ニ V」の分布の全体的な傾向は下の通りである。

表 1 データの内訳

		赤ク V	赤ニ V
補語	なる	590	26
	する	134	4
修飾語	主体動作・客体変化動詞	146	12
	主体動作動詞	101	7
	主体変化動詞	405	45
合計		**1405**	**94**

この分布では、通常、動詞の補語と修飾語の位置に現れるのは形容詞「赤ク」であり、「赤ニ」の使用は稀であるということがわかる。

次に「補語」となる場合と「修飾語」となる場合の分布を示す。

4.2 「補語」の場合

「補語」となる場合、「赤ク V」と「赤ニ V」は「なる」と「する」との共起が見られる。

表 2「補語」となる場合

	赤ク V　合計：29 例	赤ニ V　合計：724 例
	なる 26　する 4	なる 590　する 134

4.3 「修飾語」の場合

4. 3. 1 「主体動作・客体変化動詞」と共起する場合

表 3「主体動作・客体変化動詞」と共起する場合

	赤ク V　合計：146 例	赤ニ V　合計：12 例
色彩の変化を引き起こす意味を含意する動詞	染める 85　彩る 3　塗装する 3　染色する 1　着色する 1	染める 3　彩る 3　着色する 1　色付けする 1
色彩の変化を引き起こす意味を含意しない動詞	塗る 32　変える 3　腫らす 3　熱する 3　熟す 2　焼く 2　付ける 1(以下同じ)　仕上げる　括る　塗り替える	塗る 2　　焼く 1
生産動詞	書く 2　描き込む 1　記す 1	描く 1

4.3.2 「主体動作動詞」と共起する場合

表 4「主体動作動詞」と共起する場合

	赤ク V　合計：101 例	赤ニ V　合計：7 例
発光・光の受けを含意する動詞：	光る 26　燃える 21　輝く 20 照らす 7　点滅する 5　発色する 3　日焼けする 2 照り映える 2　点灯する 2　発光する 1　晒す 1　てかる 1　紅潮する 1　差す 1	光る 2　輝く 1　きらめく 1　照り映える 1
発光・光の受けを含意しない動詞：	選択する 2　流れる 1　高潮する 1　揺れる 1　見る 1　見せる 1　滴る 1	縁取る 2

4.3.3 「主体変化動詞」と共起する場合

表 5「主体変化動詞」と共起する場合

	赤ク V　合計：405 例	**赤ニ V　合計：45 例**
単純に変化を意味する動詞：	変化する 6　変わる 5	変わる 18　変化する 4
色彩の変化を含意する動詞：	染まる 144　色付く 30　濁る 9　変色する 2　紅葉する 1　黒ずむ 1	染まる 9　色付く 6　紅葉する 1
色彩の変化を含意しない動詞：	腫れる 50　爛れる 19　充血する 11　熟れる 7　焼ける 6　錆びる 5　残る 4　実る 4　写る 4　表示する 4　盛り上がる 4　熾る 4　潤む 4　膨らむ 4　咲く 3　滲む 3　浮き上がる 3　濡れる 3　覆う 2　茹で上がる　映える　浮かび上がる　腐る　広がる　荒れる　熟する　焦	見える 5　熟れる 2

	げる　揺らめく　目立つ　育つ　気触れる　ささくれる　灼熱する　1　萌える　弾ける　泣き腫らす　汚れる　ぷつぷつする　むくれ上がる　禿げる　茹だる　浮き出る　粘つく　尖る　膨れる　負う　縮れる　ささくれ立つ　透き通る　上気する　突き出る　ぶつぶつする　はためく　連なる　かじかむ　しぶく　膨れ上がる　湿潤する　帯びる　残す　淀む　仕上がる　彫る　流れ出る　つやつやする　保つ　括弧る　変身する　析出する	

次の節では、「赤ク V」と「赤ニ V」の棲み分けについて考える。

5.　「赤ク V」と「赤ニ V」の棲み分け

5.1　「補語」の場合

動詞「なる」と共起する場合、「赤ニ V」の用例が少なく 26 例である。そのうち、「交通信号」(18 例)　「化学実験」(3 例)「果実の色」(2 例)「傷跡の色」(1 例)「赤字」(1 例)である。

(12)　信号が赤になった。(安倍晴明)

(13)　青いリトマス試験紙が酸性では赤に、赤いリトマス試験紙がアルカリ性では青になる。(幸福の革命)

「赤ニ」の使用環境を考えてみると、まず、「赤ニ」が使用される文脈は、色が離散的に存在し、色が互いに対立しあう概念を表す場面であるといえる。

信号灯の場合、色は「青」・「黄色」・「赤」のように離散的に存在しており、「前進可」・「注意」・「前進不可」といった対立しあう概念を表している。リトマス試験紙の場合は「赤」・「青」は「酸性」か「アルカリ性」かという対立しあう概念を表すものとして離散的に存在していると言える。このように見ると、「赤くなる」に比べて、「赤になる」は色の対立によって表される意味概念の対立のある文脈に現れることが特徴的である。ゆえに、「赤になる」は色で表される対立した概念から、一つの色に焦点を当てることで概念を特定するという文脈で使用されるといってよいだろう。

そのほか、「赤字」の例が 1 文ある。この場面では、結果の色の変化で「欠損になる」ということを表し、メトニミー的な表現とも言えるだろう。

(14)　駅をつくることによって私ども収支がはじけるわけでございますが、その結果が赤になって出ればこれは収支が悪化したわけでありまして……(国会会議録)

一方、「する」と共起する場合、上の三例の文脈のいずれにおいても、色彩の選択肢の存在が読み取られ、「赤にする」を「赤」という選択肢を選ぶというように解釈することが可能である。

(15) 例えば、文章を入力した後で、『○○』という文字は全て赤にする、とか『以上』は『以　上』にして右詰にするとか……(Yahoo!知恵袋)

(16) 全体的に血のイメージを持たせたかったので赤にしてみました。(Yahoo!ブログ)

(17) ―(髪の毛の色の話)「黒く染めて、短くしないと退学だって」―「…いいのかよ。自慢の髪なのにステージで目立つようにって、赤にしたんじゃなかったのか。」(Bad boys!)

5.2 「修飾語」の場合

「補語」の場合と同じように、「修飾語」となる場合においても「赤ニ」は色の対立によって表される意味概念の対立のある文脈に現れやすい。

動詞「変わる」と共起する 11 例のうちの 10 例は信号の色であり、この場合の「赤」は単に色を表しているというよりは、メトニミー的に「赤信号」を表しているともいえるだろう。

(18) 警戒信号が赤に変わった瞬間、だしぬけに扉が開いた。(オバケヤシキ)

そのほか、「赤ニ」が使用される文脈には「色の多様性に言及する場合」「色変化の始まりがある場合」「色が詳細な説明を受ける場合」が特徴的である。いずれの文脈において、「赤ニ」の使用は文法的な義務に近い。

・「色の多様性に言及する場合」22 例

(19) 金や銀や青や赤にきらめく虹のつぶの中で、むすめは手をあわせました。(花にすむ馬)

「色の並列」が含まれる文脈に現れる「赤ニ」の使用は文法的な義務といえる。しかし、このような「赤ニ」の用法は、同時に現れる色のバリエーションを表現できない「赤ク」を機能的に補足しているように見られる。例えば、以下の例は「赤ク」が「青く」と並んで「変化させる」を修飾する例であるが、「赤く青く V」は同時に現れる色のバリエーションの表現として読み取られず、事象において順序的・反復的に現れる色を表している。

(20) 梅本は顔色を赤く青く変化させながら山際を送り、エレベーターの前でしどろもどろに挨拶していた。(御堂筋殺人事件)

・「色変化の始まりがある場合」7 例

(21) 頭部のランプが青から赤に変化した。(まぼろし曲馬団)

(22) 初芝選手のリストバンドがこれまでの黒から赤に変わってたんですがどうかしたんでしょうか？(Yahoo!知恵袋)

カラ格項は色変化の始まりを表している。これに似たような「赤ク V」の例がある。カラ格項の後ろに必ずニ格項がそれに呼応するとも言えないようである。

(23) 実は緑色から赤く変わる。(Yahoo!ブログ)

・「色が詳細な説明を受ける場合」9 例

(24) 青い瞳がルビーのような赤に変化する。(キスは殺しの始まり)

(25) 見とれるようなあざやかな赤に変わり、なんともいえない風味がある。(イングランド田園讃歌)

「色が詳細な説明を受ける場合」は「赤ニ」に連体修飾成分がかかり規定されている。「赤ク」よりはより具体的に事象に関わる色を説明している。。このような修飾語に対して具体的な説明を成すために、名詞としての「赤ニ」の品詞的性質が利用されている。

しかし、修飾語において対を成すものはごく稀である。対を成さない語からなる修飾語は

どのように具体的な説明を受けているかは興味深いが、それについての議論は別稿に譲る。

6. まとめ

本稿は連用修飾語に見られる「A ク」・「N ニ」というパラレルに注目して、「赤ク」・「赤ニ」を例として取り上げ、大規模コーパスデータを用いて棲み分けの分析を試みた。考察の結果を以下のようにまとめておく。

① 通常、動詞の補語と修飾語の位置に現れるのは形容詞「赤ク」であり、「赤ニ」の使用は稀である

② 「赤ニ」は色によって表される意味概念の対立のある文脈に現れやすい。

③ 以下の三つの場合において、「赤ニ」の使用は文法的な面において義務的といってよいが、事象を表す機能の面から「赤ク」の補足を行っている。

「色の多様性に言及する場合」(例：赤と金に染める)

「色変化の始まりがある場合」(例：ランプが青から赤に変化する)

「色が詳細な説明を受ける場合」(例：あざやかな赤に変わる)

言語資料

現代日本語書き言葉均衡コーパス(データバージョン 1.1)

検索アプリケーション　中納言(ver.2.4)

参考文献

奥田靖雄(1977)　「アスペクトの研究をめぐって――金田一的段階」『国語国文』8 宮城教育大学［松本泰丈編（1978）『日本語研究の方法』に所収、pp.203-220.］

井本亮(2013)　「『現代日本語書き言葉均衡コーパス』にみられる副詞的修飾関係「赤ク V」について」『商学論集』82(1)福島大学経済学会　pp.1-19

工藤真由美(1995)　『アスペクト・テンス体系とテクスト―現代日本語の時間の表現―』ひつじ書房

新川忠(1979)　「「副詞と動詞とのくみあわせ」試論」『言語の研究』　言語学研究会編 むぎ書房　pp.173-202

新川忠(1996)　「副詞の意味と機能―結果副詞をめぐって」『ことばの科学』 7 言語学研究会[編] むぎ書房　pp.61-80

橋本四郎(1975)　「修飾 ― 連体と連用」『日本語と日本語教育　文法編』　文化庁　pp.141-192

早津恵美子(2010)　「連用修飾語の解体--再構築にむけて」『国文学 : 解釈と鑑賞』 75 ぎょうせい　pp.60-68

矢澤真人(1985)　「情態修飾成分と〈シテイル〉の意味」『日本語学』第二巻 明治書院　pp.63-80

矢澤真人(1992)　「格格の階層と修飾の階層」『文藝言語研究　言語篇』21 筑波大学文芸・言語学系　pp.53-70

矢澤真人(2000)　「副詞的修飾の諸相」『文の骨格』　岩波書店　pp.189-233

連体助詞の「ノ」と文体の関係

森　秀明（東北大学文学研究科）

Relationship between Literary Style and Modifying Particle "no"

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

要旨

名詞の頻度と文体には強い関連性があり，硬い文体や客観的な文体ほど名詞の頻度が高い。一方，連体助詞の「ノ」は名詞の頻度に連動して増減する。それでは，ノと文体の関係はどうだろう。硬い文体は難易度が高い傾向があるため，名詞の増加以上にノが増加するのだろうか。本研究では BCCWJ 図書館書籍に文体指標をつけた国立国語研究所（2015）を利用し，文体の違いによる名詞とノの回帰直線を調査した。回帰直線は外れ値に弱く，この除去が分析のカギとなる。図書館書籍では固有名詞や数詞が列挙されるサンプルが存在し，これらが外れ値となっている。そこで文書構造タグの<figureBlock>と<list>が存在するサンプルを除き，普通名詞と普通名詞に接続するノに絞って回帰直線を調査した。その結果，文体による変化は見られなかった。ノは，普通名詞に連動して増減するだけで，その頻度に文体の影響はない。ノの頻度は，人間の意志や個性とはほとんど無関係に増減している可能性がある。

１．研究の目的と先行研究

日本語のテキストで使用されている品詞の構成比率には一定の規則性が存在し，名詞比率に連動して動詞や形容詞類の割合が規則的に変化することが知られている。樺島（1955）は現代語の延べ語数を使用した品詞構成比率（図 1）を，大野（1956）は古典文学の異なり語数を使用した品詞構成比率（図 2）を分析し，これを明らかにした。

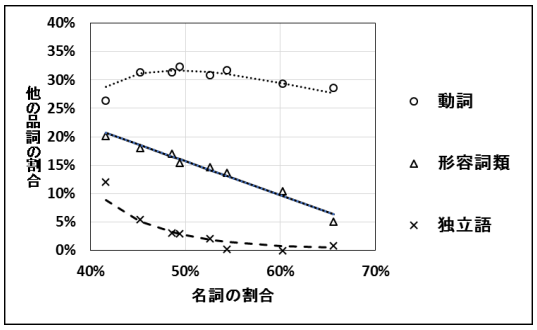


図 1：樺島（1955）第一表に基づく散布図

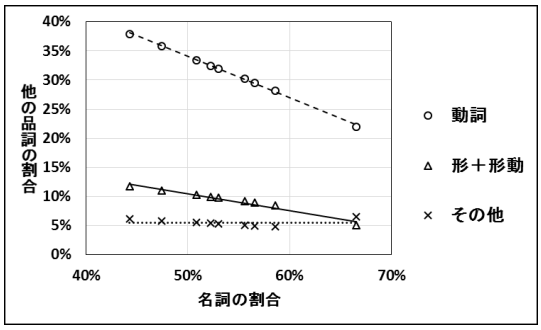


図 2：大野（1956）第七表に基づく散布図

図 1，2 に見られる規則性を定式化した数式は「樺島の法則」や「大野・水谷の法則」と呼ばれ，計量的な言語研究における重要な発見と位置づけられてきた。ただしこれらの研究では名詞と付属語の関係は不明なままであったため，発表者は BCCWJ・固定長・長単位データを使用してこれを調査し，昨年度の言語資源活用ワークショップで報告した（森，2017）。品詞比率を調査する際，図鑑などのサンプルでは名詞が多数列挙され，もはや文章とは呼べ

ないテキストも存在したため，名詞比率 45%未満等の条件で絞り込んだサンプルを仮に「一般的な日本語テキスト」と定義してこれを分析に使用した。主な調査結果は次の通りである。

(1) 助動詞と名詞には，強い相関がある。
(2) 助詞と名詞には，相関関係がない（日本語のテキストで助詞比率はほぼ一定である）。
(3) 連体助詞と名詞，接続助詞と名詞には中程度の正と負の相関関係があるが，これらを「語と語を結合する助詞」と考えて頻度を合計すると，この合計数と名詞との相関が低くなる（日本語のテキストで結合助詞比率はほぼ一定である）[1]。
(4) 格助詞や係助詞などと名詞には正と負の相関があるが，結合助詞以外の助詞の全てを「格関係に関わる助詞」と考えて頻度を合計すると，この合計数と名詞との相関が低くなる（日本語のテキストで格関係助詞比率はほぼ一定である）。

表 1　品詞頻度の積率相関行列：BCCWJ 図書館書籍 SC の一般文書　n =10,385

	名詞	普通名詞	動詞	その他	助動詞	助詞	結合助詞	格関係	格助詞	係助詞	終助詞	準体助詞	副助詞	接続助詞	連体助詞
普通名詞	0.814														
動詞	−0.533	−0.423													
その他	−0.545	−0.410	−0.003												
助動詞	−0.722	−0.668	0.222	0.231											
助詞	0.024	0.089	−0.106	−0.326	−0.381										
結合助詞	0.340	0.340	−0.250	−0.259	−0.426	0.429									
格関係助詞	−0.261	−0.195	0.103	−0.112	−0.028	0.652	−0.405								
格助詞	0.362	0.415	−0.007	−0.480	−0.419	0.400	0.036	0.374							
係助詞	−0.198	−0.222	−0.118	0.087	0.150	0.223	−0.190	0.385	−0.249						
終助詞	−0.517	−0.580	0.230	0.321	0.365	−0.003	−0.325	0.270	−0.514	0.005					
準体助詞	−0.441	−0.428	0.133	0.208	0.286	0.171	−0.245	0.379	−0.238	0.119	0.425				
副助詞	−0.193	−0.045	0.028	0.157	0.019	0.184	−0.159	0.320	−0.227	−0.030	0.159	0.145			
接続助詞	−0.549	−0.486	0.546	0.199	0.199	0.156	0.155	0.027	−0.285	0.007	0.349	0.225	0.098		
連体助詞	0.650	0.610	−0.570	−0.355	−0.499	0.274	0.771	−0.370	0.215	−0.170	−0.508	−0.359	−0.202	−0.509	
普通名詞＋ノ	0.579	0.708	−0.456	−0.305	−0.472	0.223	0.686	−0.350	0.196	−0.193	−0.477	−0.342	−0.132	−0.434	0.877

表 2　品詞頻度の積率相関行列：BCCWJ 新聞 SC の一般文書　n =1,353

	名詞	普通名詞	動詞	その他	助動詞	助詞	結合助詞	格関係	格助詞	係助詞	終助詞	準体助詞	副助詞	接続助詞	連体助詞
普通名詞	0.631														
動詞	−0.560	−0.346													
その他	−0.673	−0.470	0.110												
助動詞	−0.714	−0.664	0.318	0.456											
助詞	−0.028	0.278	−0.098	−0.304	−0.486										
結合助詞	0.050	0.215	−0.228	−0.053	−0.218	0.373									
格関係助詞	−0.063	0.139	0.055	−0.277	−0.351	0.774	−0.299								
格助詞	0.228	0.252	0.097	−0.534	−0.425	0.526	−0.168	0.655							
係助詞	−0.442	−0.407	0.150	0.432	0.382	−0.132	−0.083	−0.079	−0.396						
終助詞	−0.232	−0.182	−0.094	0.230	0.068	0.204	−0.160	0.319	−0.277	0.056					
準体助詞	−0.422	−0.296	0.127	0.338	0.264	0.078	−0.100	0.148	−0.199	0.358	0.175				
副助詞	0.053	0.304	−0.073	−0.118	−0.237	0.347	−0.018	0.369	−0.028	−0.019	−0.122	−0.026			
接続助詞	−0.517	−0.457	0.340	0.451	0.386	−0.142	0.172	−0.264	−0.424	0.403	0.108	0.232	−0.105		
連体助詞	0.363	0.475	−0.415	−0.325	−0.434	0.423	0.796	−0.108	0.109	−0.323	−0.210	−0.233	0.048	−0.460	
普通名詞＋ノ	0.215	0.566	−0.260	−0.168	−0.335	0.330	0.675	−0.121	−0.018	−0.199	−0.154	−0.131	0.111	−0.299	0.792

このことから，日本語は文体や叙述内容に関わらず，文章を書く際には全語数の約 1/3 を

[1] 表 1 の図書館書籍 SC の場合、名詞と結合助詞の相関は.340 で弱い相関があるため、「日本語のテキストで結合助詞比率はほぼ一定である」とは言えないが、表 2 の新聞 SC では.050 と相関がないため、仮にこのように考える。格関係助詞も同じである。

助詞に使用し，助詞の 1/3 は語と語の結合に，残りの 2/3 は格関係に使用するシステムを持っていると考えられる。

この発表に対しては大きくまとめて次の 2 点の指摘を頂戴した。

（5）これだけでは，この現象がどのような言語学的意味を持っているのか分からない。
（6）全データを使用していない分析結果を，日本語全体に一般化することはできないのではないか。

本研究ではこれを受けて，この問題をさらに検討する。

日本語における品詞比率の問題は，文体との関連で論じられてきた。樺島（1955：386）では「名詞の百分率をもって，文章の特性を計る尺度となし得る」とされ，「N の増加は話し言葉的なものから書きことばへと向かつている」，「感情の表現をなすものから關係の表現をなすものへと，N が増す」（p.387）などの特徴が指摘されている。

しかし，どのような文章においても助詞の比率が一定であるということは，助詞はそのような文体や叙述内容とは無関係に使用されていることを示唆している。ただし，助詞全体の比率は一定であったが，例えば連体助詞には名詞と正の相関が，接続助詞には負の相関があった。名詞に対して相関があるという点では動詞や助動詞などと同じである。このため，対象を連体助詞ノに絞り，文体の違いによってノがどのように使用されているかを調査することにする。

BCCWJ 図書館書籍サブコーパス（以下図書 SC と略す）には，10,551 文書を人手で判断して文体情報を付与した国立国語研究所 (2015)『BCCWJ 図書館サブコーパスの文体情報』[2]が存在し，その詳細は柏野（2013）で紹介されている。ここでは図書 SC のサンプルをテキスト構造が単純なもの（例：章 節構造）と，テキスト構造・紙面形式などの点で文体の評定値をつけるのになじまないもの（全体の約 2 割）に分け，前者には「専門度，客観度，硬度，くだけ度，語りかけ性度」といった評定値が，後者には「対談，Q&A 形式，図解，用語解説」等の分類情報が付与されている。本研究ではこの文体指標を利用し，文体の違いによってノの使用に変化があるかどうかを調査する。

硬い文体と軟らかい文体，客観的な文体と主観的な文体などでは，名詞の頻度が異なることが知られている。硬い文体や客観的な文体ほど名詞の頻度が高く，凝縮的な文体になっている（樺島，1955）。その一方で，連体助詞ノは，名詞の頻度に連動して増減することが知られている（森，2017）。それでは，ノと文体の関係はどうであろうか。硬い文体や客観的な文体では名詞の頻度が高いため，ノの頻度が高くなるのは当然だが，これらの文体のテキストは複雑で難易度の高い内容を記述していることが多いことから，名詞が多くなった以上にノの頻度が高くなるのであろうか。

これを調査するには，サンプルごとの名詞の頻度を X 軸（説明変数），ノの頻度を Y 軸（目的変数）とする散布図に回帰直線を描き入れて，この傾きや切片が硬いテキストと軟らかいテキストによって異なるのかどうかを観察すればよい。回帰直線の傾きや切片が異なるなら，ノの頻度は文体によって使い分けられているし，これが同じであれば文体による使い分けはないと考えられる。

[2] http://pj.ninjal.ac.jp/corpus_center/anno/の「サンプルに対する文体指標（sty）」で，BCCWJ_LB_Stylistics-1.0.zip のファイルが公開されている。

ノの使用が文体とは無関係に行われているとすれば，日本語のテキストにおいて助詞の比率が一定である理由も分かりやすくなる。すなわち，助詞は文体や叙述内容に関わらず，語の結合と格関係の表示に一定数を必要とするシステムで，そこに人間の意志や個性はほとんど介在していない。日本語文法では助詞・助動詞を付属語とか機能語という扱いで同列に扱ってきたが，真に日本語の機能をつかさどっている品詞は助詞であり，助動詞は名詞に連動して増減する点において，動詞や形容詞と同じ分類に入る品詞である。このことは，助動詞が文体や叙述内容に深くかかわる品詞であることを示唆している。

（6）のデータ選択の問題は，コーパス言語学において古くから論じられてきた問題で，基本的には分析に適さないデータは除くべきだと考えられる。国立国語研究所（2015）でも文体の評定値をつけるのが難しかったサンプルが 2 割ほど存在することが指摘されている。これらは，「対談，Q&A 形式，図解，用語解説」等のサンプルで，文体分析に使いたくともその性質が文体評定に適さないため，評定がつけられなかったサンプルである。前回発表でも国立国語研究所（2015）に従って分析を行おうと試みたが，なお分析に適さないと考えられるサンプルが存在したため，名詞比率 45%未満という基準を設けた。しかし，このような一律に足切りをする基準では，恣意的なデータ選択の印象を免れないため，今回は，図書 SC を構築する際に付与された文書構造の情報に基づいてサンプルの選択を行い，データ選択の違いによって分析結果にどのような影響が出るのかについても考察する。

２．分析データ

２.１　使用するコーパスとデータの種類

分析には図書 SC の固定長・長単位データを使用する。BCCWJ では形態素解析用辞書 UniDic と長単位解析器 Comainu によって品詞情報が付与されている。UniDic の品詞体系は基本的に学校文法の体系に近いが，形容動詞はその語幹を「形状詞」として認定され，活用語尾は助動詞に分類されている。また長単位では複合名詞を 1 語に認定するほか，複合助詞，複合助動詞を一語として認定している。本研究では格助詞ノを連体助詞として格助詞から分離して分類する以外，品詞の認定は UniDic の品詞体系に従った。また本研究では品詞を類別して分析する際，基本的に山崎（2014）の類別基準を参考にしたが，品詞比率が大きい名詞，動詞，助詞，助動詞以外は一括して「その他」として扱った[3]。また格助詞や係助詞と言った助詞の下位分類を中分類，それらを合計した助詞全体を大分類と呼ぶ。

２.２　データの絞り込み

図 3 は，図書 SC の 10,551 サンプルについて品詞比率を求め，横軸を名詞比率，縦軸を助詞比率にして描いた散布図である。本研究では何も絞り込みを行わないデータをフルデータと呼ぶ。

図 3 では名詞比率 40%までは楕円形で，そこから下に向かう尾がついているような形をしている。図 4 は国立国語研究所（2015）の文体情報を使用し，柏野（2013）で「文体判断が単純にいかないもの」と判断された 1,758 サンプルを除いた上で図 3 と同様に描いた

[3] 名詞：名詞・代名詞・接尾辞－名詞的，動詞：動詞，接尾辞－動詞的，助詞：助詞，助動詞：助動詞，その他：長単位語数表（BCCWJ_WC_LUW_v10.xlsx）の語数（記号等除外・固定長）から上記の品詞数を除いたもの。山崎（2014）では名詞に「記号」を含めるが，本研究では「その他」の品詞数の算出に長単位語数表（記号等除外・固定長）を使用したため，名詞に「記号」は含めなかった。

散布図である。本研究ではこれを章節構造データと呼ぶ。

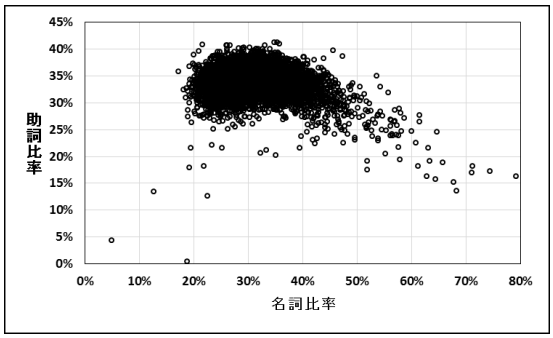


図 3 名詞比率と助詞比率の散布図：
図書 SC フルデータ，N=10,551

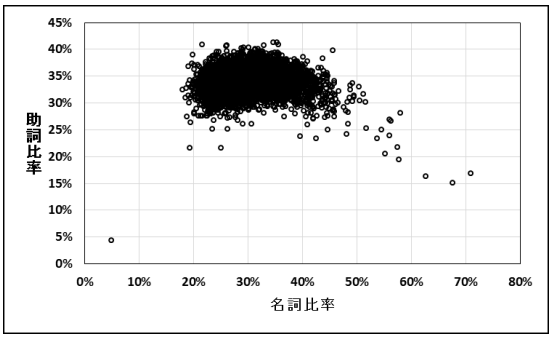


図 4 名詞比率と助詞比率の散布図：
図書 SC 章節構造データ，N=8,792

「文体判断が単純にいかないもの」は図解，コマ割などが多用される「視覚表現多用系」，用語解説，見本・カタログ形式などの「データベースやリスト系」，対談，インタビューなどの「対話系」など 11 の観点から分類されているサンプルで，「テキスト構造・紙面形式に特徴をもつもの」である。次の（7）は「視覚表現多用系」，（8）は「データベースやリスト系」の文書の一部である。

（7）アリのなかまクロオオアリアリ科■働きアリ７～十三ｍｍ■４～十月　全国■里山■成虫・幼虫●日本では最大のアリ働きアリ女王アリ←ムネアカオオアリアリ科■働きアリ８～十二ｍｍ■５～十月■北・本・四．九■里山■成虫・幼虫●クロオオアリに似るが胸が赤い（BCCWJ サンプル ID：LBqn_00015，実著者不明，『昆虫』，名詞比率 50.9%，助詞比率 27.5%）

（8）今後，世界遺産条約の締約が期待される中東の国々アラブ首長国連邦Ｕｎｉｔｅｄ　Ａｒａｂ　Ｅｍｉｒａｔｅｓ面積　八万三千六百ｋｍ２人口　二百五十八万人主要言語　アラビア語首都　アブダビ通貨　ディルハム民族　アラブ人宗教　イスラム教（BCCWJ サンプル ID：LBo5_00063，実著者不明，『世界遺産ガイド』，名詞比率 71.4%，助詞比率 18.2%）

（7），（8）の文書では助詞の数に比べ名詞の数が著しく多い。その理由はこれらの文書に名詞句の列挙が多く含まれるからである。これらの「文体判断が単純にいかないもの」を除くと，図 4 のように尾の部分がかなり少なくなる。それでもまだ図 4 では名詞比率 45%までの楕円形の塊と尾に分かれているように見える。

次に図 4 の尾の部分のサンプルを観察する。（9）は図 4 で最も名詞比率が高いサンプル（10）は名詞比率 44.5%のサンプル（11）は名詞比率が最も少ないサンプルである。

（9）また，高速十号線（新宿区付近～練馬区付近），同内環状線（墨田区付近～新宿区付近同十一号線（葛飾区付近～市川市付近），同晴海線（江東区付近～千代田区付近），同磯子線（横浜市南区付近～同市磯子区付近），同２号線（延伸），第二東京湾岸道路，都心新宿線及び首都高速道路４号線の機能強化について計画を進める。（BCCWJ サン

プル ID：LBg6_0001，実著者不明，『首都圏白書』，名詞比率 70.9%，助詞比率 17.0%）

（10）宗室は有爵と無爵があり、爵位は次の十四等に別れる。　親王、世子、多羅郡王、長子、多羅貝勒、固山貝子、鎮国公、輔国公、不入八分鎮国公、不入八分輔国公、一・二・三等鎮国将軍、一・二・三等輔国将軍、一・二・三等奉国将軍、奉恩将軍。　（BCCWJ サンプル ID：LBi9_00142，高陽（著）永沢道雄・鈴木隆康（訳）『西太后』，名詞比率 44.5%，助詞比率 35.2%）

（11）２、無政府主義派（イ）共産主義ノ主張ハ基礎ヲ社会大衆ニ置キ、巧ミニ之レヲ誘致シテ民衆的革命ヲ目的トスルニ反シ、無政府主義ハ権力ヲ否定シ、暴力革命ヲ高調スル点ニ於テ今次ノ如キ突発事変ニ際シテハ警戒ノ必要寧ロ前者ヨリ以上必要トスルモノアリ。（BCCWJ サンプル ID：LBs2_00005，松尾尊兌，『世界史としての関東大震災』，名詞比率 4.8%，助詞比率 4.4%，その他比率 87.1%）

（9）は柏野（2013）で「文体判断が単純にいかないもの」には認定されていないが，道路の名前が列挙されており，一般的なテキストとは見なしにくい。（10）も後半は名詞の列挙で一般的な文章になっていない。(11) は名詞がたくさん出現しているが，名詞比率は 4.8% となっている。その理由はほとんどの品詞を「カタカナ文」というカテゴリで解析されているためで，うまく形態素解析できていないと考えられる。本研究の目的は文体の違いによって名詞とノの回帰直線に変化があるかどうかを調査することにあるため，ノが出現する余地なく名詞が列挙されているサンプルを含めて分析する意義は低いと考えられる。

前回の発表では名詞の列挙を含む文を少なくする目的で名詞比率を 45%未満に絞り込み，解析ミスと考えられる「カタカナ文」を多く含む文書を少なくする目的でその他比率は 30% 未満に絞り込んだ。その上で，この「名詞比率 45%未満・その他比率 30%未満」のサンプルを仮に「一般的な日本語テキスト」と定義してこれを分析に使用した。しかし，このような絞り込みでは恣意的なデータ選択を行っている印象はぬぐえない。そこで本研究では，BCCWJ にタグ付けされている文書構造の情報を利用してサンプルの絞り込みを行う。

文書構造タグは，原資料を電子的なテキストに変換するに当たって，元々の資料が持っていた構造を復元できるように付与された情報である（詳しくは，山口，2014；西部・大島・間淵・小林ほか，2011；山口・高田・北村・間淵，2011 を参照のこと）。図 5 は図表からサンプルを取得する際につけられた文書構造タグの例である。上の段は原資料の表が，下の段はそれを電子化したデータが表示されている。

図 5 の左の表は，上段に表のキャプションがあり，下に項目と数字の表がある。下段では先頭に<figureBlock>, 最後に</figureBlock>というタグが付与されている。<figureBlock>とは，「図表・写真・絵などの要素と，それに付随する文書要素をまとめた要素を表す。」（山口・高田・北村・間淵，2011:82）。つまり<figureBlock>のタグが付いていると，その文書には図表・写真・絵などの要素が含まれていることを意味している。左の表は，数字が主体であるため，サンプリングされたのは表のキャプションのみである。

これに対し，右の図は言語情報が主体であるため，表の中のリストがデータとして採取されている。この時つけられているのが<list>というタグである。<list>は「箇条書きなど，列挙された文書要素の集まりを表す。」（山口・高田・北村・間淵，2011:91）。つまり<list>のタグが付いていると，その文書には名詞の列挙が含まれる可能性が高くなる。

〔PB24_00304：『生命倫理とこころのケア』〕

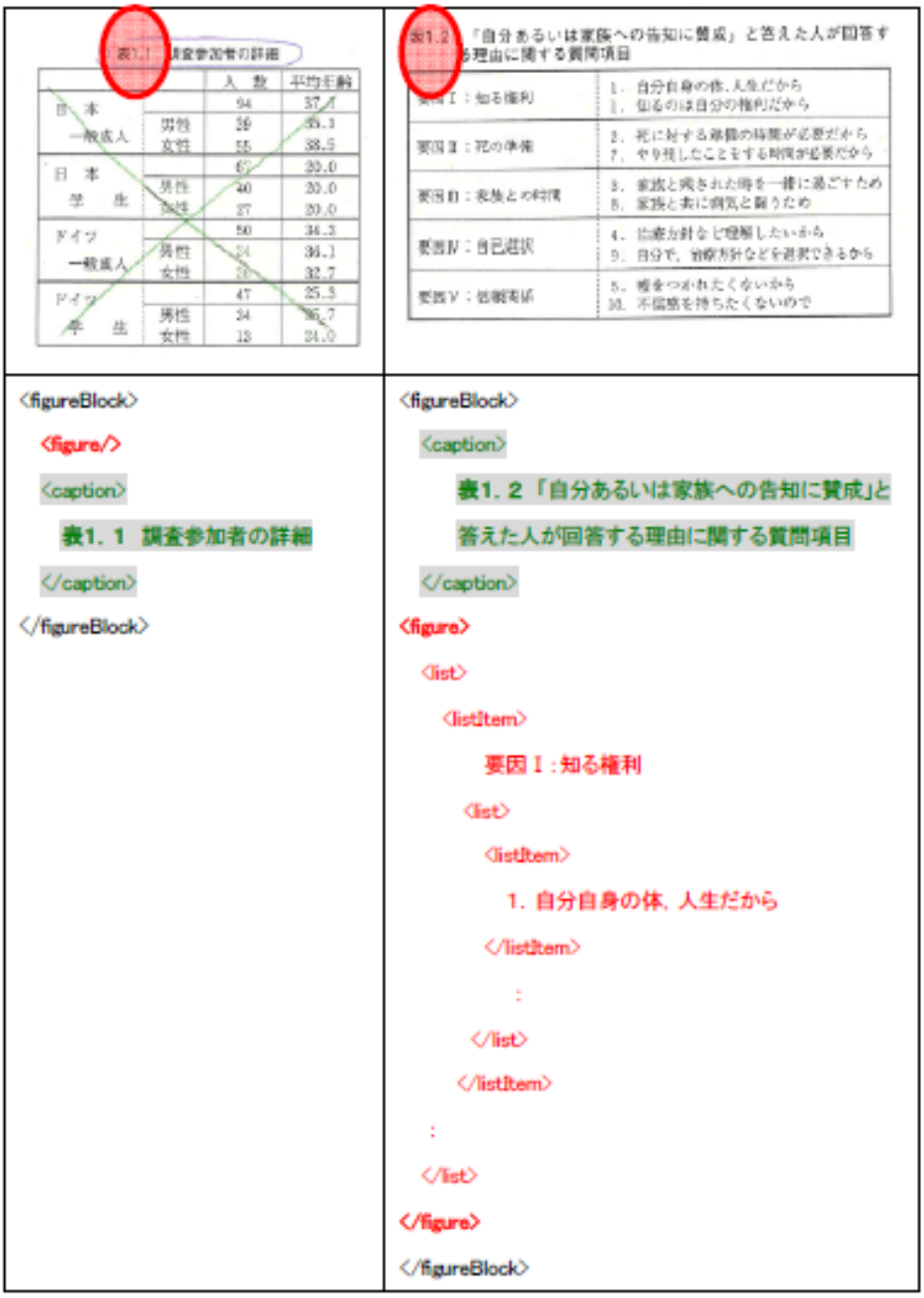


表1.1 調査参加者の詳細

		人数	平均年齢
日本 一般成人		94	37.4
	男性	39	35.1
	女性	55	38.5
日本 学生		67	20.0
	男性	40	20.0
	女性	27	20.0
ドイツ 一般成人		50	34.3
	男性	24	36.1
	女性	26	32.7
ドイツ 学生		47	25.3
	男性	34	26.7
	女性	13	24.0

表1.2 「自分あるいは家族への告知に賛成」と答えた人が回答する理由に関する質問項目

要因Ⅰ：知る権利	1．自分自身の体，人生だから 1．伝えるのは自分の権利だから
要因Ⅱ：死の準備	2．死に対する準備の時間が必要だから 7．やり残したことをする時間が必要だから
要因Ⅲ：家族との時間	3．家族と残された時を一緒に過ごすため 8．家族と共に病気と闘うため
要因Ⅳ：自己選択	4．治療方針など理解したいから 9．自分で，治療方針などを選択できるから
要因Ⅴ：情報獲得	5．嘘をつかれたくないから 10．不信感を持ちたくないので

```
<figureBlock>
  <figure/>
  <caption>
    表1．1　調査参加者の詳細
  </caption>
</figureBlock>
```

```
<figureBlock>
  <caption>
    表1．2「自分あるいは家族への告知に賛成」と
    答えた人が回答する理由に関する質問項目
  </caption>
<figure>
  <list>
    <listItem>
      要因Ⅰ：知る権利
      <list>
        <listItem>
          1．自分自身の体，人生だから
        </listItem>
          :
      </list>
    </listItem>
  :
  </list>
</figure>
</figureBlock>
```

図 5　図表からのサンプリングの例（西部・大島・間淵・小林ほか，2011:271 より引用）

本研究では，<figureBlock>と<list>のタグが含まれている文書は，名詞の頻度を使用した文体分析には適さない文書と判断し，これを除いた文書で分析を行う。本研究ではこれを選抜データと呼ぶ。この基準によって除かれる文書数は 2,266 文書で，残存率は 78.5%（8,283 文書）である。また選抜データと国立国語研究所（2015）の章節構造データの基準を同時に適用した際に除かれる文書数は 3,362 文書で，残存率は 68.1%（7,189 文書）である。本研究ではこれを二重選抜データと呼ぶ。

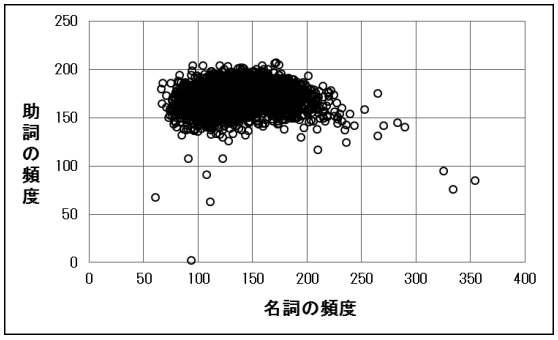


図 6 名詞頻度と助詞頻度の散布図：
図書 SC 選抜データ，：N=8,283

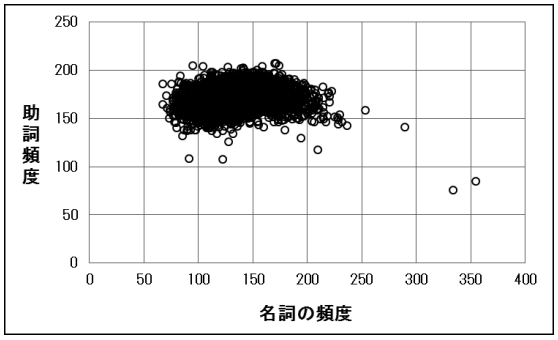


図 7 名詞頻度と助詞頻度の散布図：
図書 SC 二重選抜データ，：N=7,189

図 6 は<figureBlock>と<list>のタグが含まれている文書を除いた選抜データ，図 7 はここからさらに国立国語研究所（2015）の章節構造データのみを残した二重選抜データである。図 3 のフルデータから図 4 の章節構造データ，図 6 の選抜データ，図 7 の二重選抜データと削除数を増やすと，外れ値と思われるサンプルがより多く除かれていくことが確認できる。しかし，図 7 の二重選抜データでもなお外れ値と思われるサンプルが若干残っている。

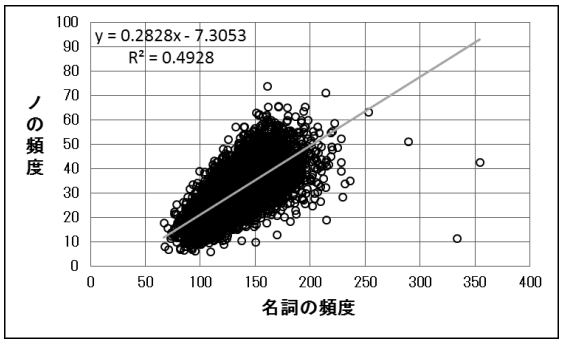


図 8　名詞頻度とノの頻度の散布図
図書 SC 二重選抜データ，：N=7,189

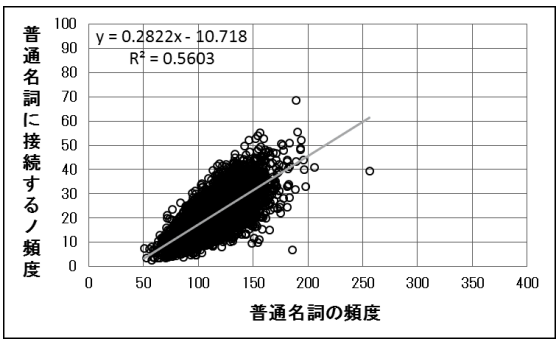


図 9　普通名詞頻度と普通名詞に接続するノの頻度の散布図，図書 SC 二重選抜データ，：N=7,189

図 8 は名詞とノの散布図で，二重選抜データを使用しているものの，このままでは，依然として外れ値の影響を受けることが考えられる。図 8 で名詞頻度が多いサンプルを観察すると，なお，固有名詞や数詞の列挙が残っていることが分かる。図 8 で最も名詞が多いサンプルは，先に挙げた例文（9）の『首都圏白書』で，2 番目に多いものが（12）の『立川飛行場物語』である。これらは，道路の路線名や町名などの固有名詞が列挙されているため，名詞数が多くなっている。

（12）大正十年の東京府電話帳を見ると、立川で五十三本の電話がひかれていたことがわかりますが、町名と氏名は次のようになっています。　＞零番＝立川郵便局―公衆通話用及び電報託送用　＞１番＝立川郵便局―一般事務用　＞　２番＝岩崎輝彌―子安農園立川分園―上古新田　＞　３番＝野沢源次郎―貿易商―下和田　＞　４番＝馬場福太郎―旅館―停車場前　＞　５番＝園部五郎吉―糸繭商―停車場前　＞　６番＝内藤九一―米穀商―停車場前　＞　７番＝旗野留五郎―雑貨商―停車場前　＞

８番＝村野安五郎―肥料商―停車場前　＞９番＝和知平三郎―雑貨商―停車場前＞十番＝中村久之助―料亭―停車場前　＞（BCCWJ サンプル ID：LBb3_00039，三田鶴吉，『立川飛行場物語』，500 語当たりの名詞頻度：333.8 語，500 語当たりのノ頻度：11.3 語）

そこで図 9 のように名詞を普通名詞に絞り，ノも普通名詞に接続するものに絞ると，概ね外れ値に影響されない状態になる。よって，分析に使用するデータは基本的に二重選抜データにし，調査対象は普通名詞と普通名詞に接続するノとすることにする。二重選抜データにおける普通名詞の頻度は 867,737 語で，全名詞の 79.1%，普通名詞とそれに接続するノの頻度は 162,769 語で，全ノの 70.5%，全名詞に接続するノの 80.3%になる。

３　分析結果と考察

３．１　外れ値と回帰直線の関係

本来はフルデータを使用し，全名詞と全ノの回帰直線を観察するのが望ましい調査である。しかし，本研究ではフルデータを 7 割弱に絞り込み，調査対象も普通名詞とそれに接続するノに限定することにした。本節ではなぜこのような絞り込みを行う必要があるのかについて，改めて説明する。

図 10 は，専門性や難易度でノの使用が変化するかどうかを調査するため，書籍の流通管理のために付与されている日本図書コード（C コード）の「教養・専門」と「児童」の区分を使用し，フルデータで普通名詞とそれに接続するノの散布図と回帰直線を描いた図，図 11 は，選抜データで同様の内容を描いた図である。

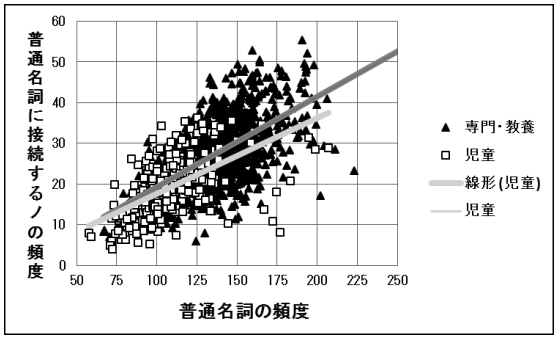


図 10 普通名詞とノの散布図・フルデータ

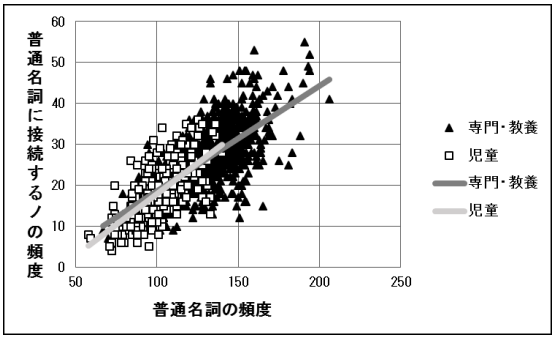


図 11 普通名詞とノの散布図・選抜データ

フルデータを使用した図 10 では，「児童」より「専門・教養」の傾きが急で，専門性の高い文体では，「児童」よりノが使用される割合が高いと判断される。しかし，選抜データを使用した図 11 では，回帰直線は一致し，専門性の高さによってノが使用される割合は変わらないと判断される。つまり，分析目的に不向きな文書を除けば，難易度や専門性が高いからと言ってノが多用されるわけではなく，専門性の高い書籍でも，児童向けの書籍でもノの使用傾向は一定であると考えられる。

選抜データに絞り込むために分析から除外した文書には，先に用例を示した（7）『昆虫』，（8）『世界遺産ガイド』が含まれており，（7）は「児童」，（8）は「教養」に分類されている。表 3 は，「児童」のフルデータ 387 文書から，60 の文書を除いた中で，普通名詞の数が多い文書 top10 のリストである。この第 1 位が用例（7）の『昆虫』である。これ以外の文

書も書名を見ると，図鑑，辞典，スポーツの解説書など，章節構造を持ったテキストとは明らかに異なる構造を持ったテキストであることが分かる。

表 3：「児童」から除いた普通名詞の多い文書 top10

ID	書名	普通名詞	ノ
LBqn_00015	昆虫	207	29
LBmn_00029	蛾蝶記	199	29
LBln_00025	道ばたの食べられる山野草	195	31
LBkn_00001	見てわかるルアーフィッシング	184	33
LBhn_00007	植物記	183	21
LBgn_00032	漢字事典五年生	177	8
LBpn_00009	服部幸應のはて・なぜ・どうしてたべものクイズ	174	18
LBnn_00001	バスケットボール	172	11
LBgn_00015	漢字事典四年生	167	14
LBdn_00020	New野球テクニック	166	35

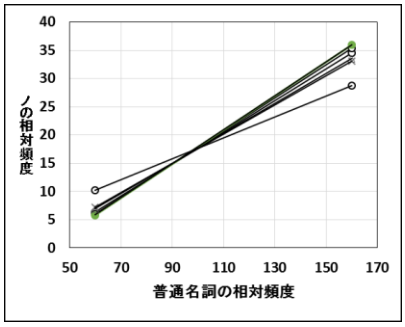


図 12：文書削除数別回帰直線

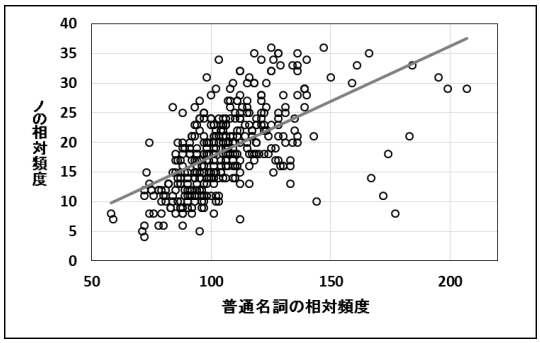


図 13 「児童」の散布図・フルデータ

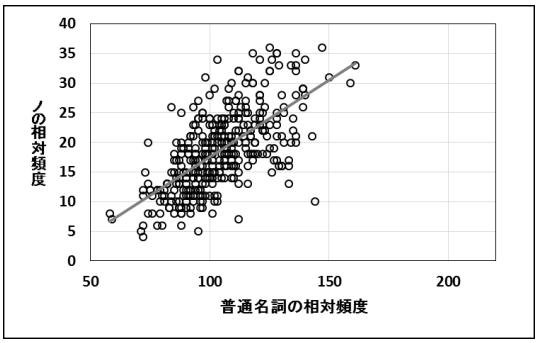


図 14 「児童」の散布図・10 文書削除

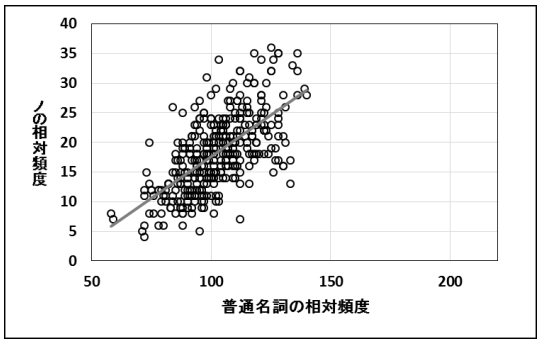


図 15 「児童」の散布図・30 文書削除

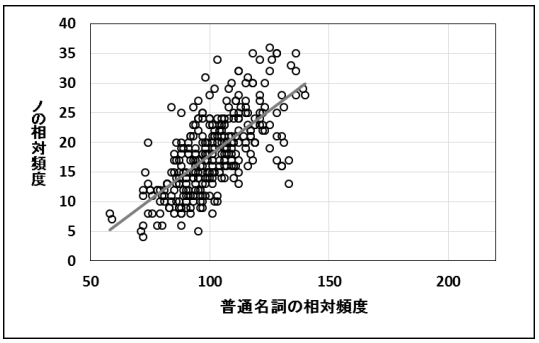


図 16 「児童」の散布図・選抜データ

ただし，<figureBlock>と<list>のタグが含まれている 60 文書をすべて除く必要があったかどうかの判断は難しい。図 12 はフルデータから普通名詞が多い順に特殊文書を 10 文書ずつ除いていった回帰直線 7 本の比較である。最も傾きが低いのがフルデータ，その上の直線がそこから表 3 の 10 文書を除いた時の回帰直線である。変化が激しいのは初めの 10 文書を除いた場合だけで，60 文書まで除く必要はなかったという考え方もできるかもしれない。

図 13 はフルデータで，明らかに外れ値と思われる文書が図の右側に散らばっている。図 14 はこれから 10 文書を除いた散布図，図 15 は 30 文書除いた散布図，図 16 は 60 除いた散布図で，これが選抜データとなる。散布図で確認しても図 14～図 16 の違いはごくわずかで

ある。しかし，文書を削除する基準をどこで線引きするかは難しく，恣意的なデータ操作を避けるためには分かりやすい基準に従うのが妥当だと思われる。

３．２　他の文体指標の結果と考察

「難易度や専門性」の違いによる普通名詞とノの回帰分析に続き，他の文体指標を使った調査の結果を示す。文体指標は，国立国語研究所（2015）を利用し，「硬度」「くだけ度」「客観度」「語りかけ性度」及び，章節構造データには分類されていない話し言葉の「対話系」と，それ以外の文書で，普通名詞とノの回帰直線がどのように異なるかを観察した。これらの指標は 2 段階～5 段階に区分されているが，図 17～図 21 では対極に位置する指標のみを使用している。分析データは図 19 のみ選抜データ，それ以外は二重選抜データを使用した。

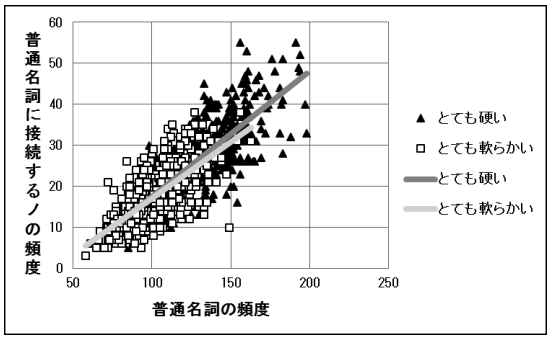


図 17 普通名詞とノの散布図・硬度

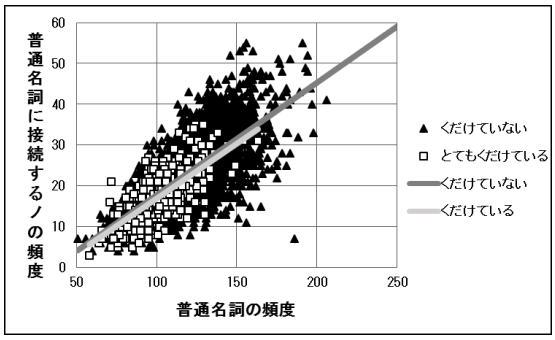


図 18 普通名詞とノの散布図・くだけ度

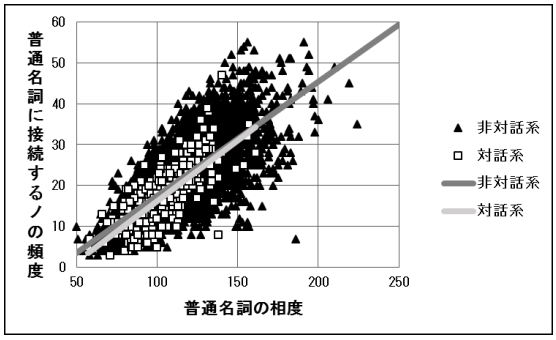


図 19 普通名詞とノの散布図・対話・非対話

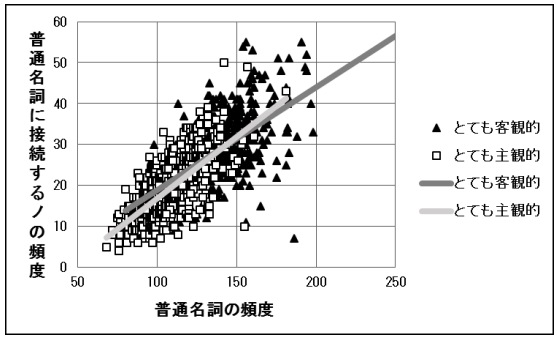


図 20 普通名詞とノの散布図・客観度

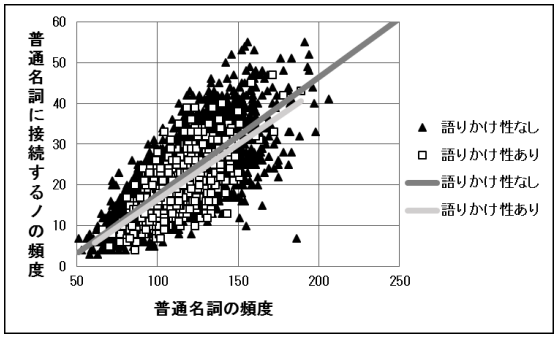


図 21 普通名詞とノの散布図・語りかけ性

表 4：文体別回帰式の係数と R^2

	傾き	切片	R^2
専門・教養	0.257	−7.122	.474
児童	0.301	−12.239	.451
とても硬い	0.307	−13.278	.530
とても軟らかい	0.274	−10.296	.475
くだけていない	0.276	−9.810	.520
とてもくだけている	0.268	−9.883	.465
非対話系	0.278	−10.304	.552
対話系	0.295	−13.197	.549
客観的	0.251	−6.249	.386
主観的	0.302	−13.337	.454
語りかけ性なし	0.287	−11.066	.584
語りかけ性あり	0.272	−10.809	.479

注：t 検定の結果すべての係数は 5%水準で有意

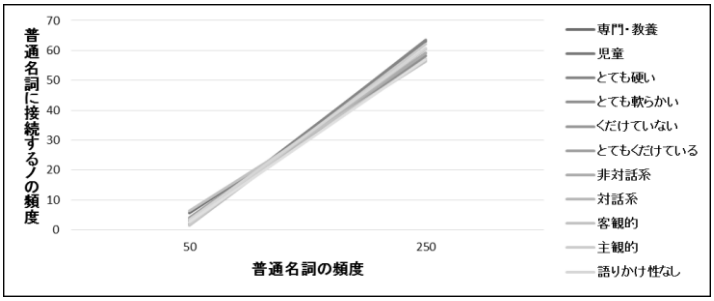


図 22　文体別回帰直線比較

図 17～図 21，及び表 2 を見ると，これらの指標の回帰式はほぼ一致し，文体指標や話し言葉・書き言葉（「対話・非対話」）で普通名詞とノの関連性は変化しないことが示唆された。文体によって普通名詞の頻度が特徴的な分布の違いを見せる一方，どのような文体であっても普通名詞の頻度が同じなら，使用されるノの頻度はほぼ同じになる。つまりノによって連体修飾節を作る述べ方は，難易度や専門性だけでなく，話し言葉や書き言葉，文章の硬軟，くだけ度，客観性，語りかけ性などには影響されない。これらの文体の違いは，執筆者の個性や執筆の意図の違いによって生じていると考えるのが自然である。このため，どのような文体であっても，普通名詞の頻度が決まればほぼ機械的にノの頻度が決まるという現象は，ノの使用に人間の個性や意志が介在する余地が小さいことを示唆していると考えられる。

４．まとめと今後の課題

本研究では BCCWJ 図書 SC のサンプルに文体指標をつけた国立国語研究所（2015）を利用し，「専門性」「硬度」「くだけ度」「客観度」「語りかけ性度」などの文体の違いによって，連体助詞ノの使われ方が異なるかどうかを調査した。

これを調査するには，テキストごとの名詞の頻度を X 軸（説明変数），ノの頻度を Y 軸（目的変数）とする散布図に回帰直線を描き入れて，この傾きや切片が文体の違いによって異なるのかどうかを観察すればよい。回帰直線の傾きや切片が同じであれば文体による使い分けはないと考えられる。

ただし，回帰直線は外れ値の影響を強く受けるため，恣意的ではない基準で，できるだけ外れ値を減らす方法を検討した。図書 SC のサンプルを観察すると，図表が含まれている文書や固有名詞・数詞が多用されている文書で，名詞の列挙が頻出する例が見られた。このため，文書構造タグの，<figureBlock>と<list>のタグがついている文書を除き，普通名詞と普通名詞に接続するノの頻度に絞って分析することで，外れ値の影響を受けにくい分析が行えると考えた。

<figureBlock>と<list>のタグがついている文書を除いた選抜データを使用し，国立国語研究所（2015）の文体指標を用いて分析すると，さらに対象となる文書が絞り込まれる。本研究ではこれを二重選抜データと呼ぶ。この二重選抜データを使用して，普通名詞と普通名詞に接続するノの頻度の回帰直線を描くと，文体の違いによって回帰直線の傾きや切片が異なることはなかった。

文体の違いは，執筆者の個性や執筆の意図の違いによって生じていると考えられる。このため，どのような文体であっても，名詞の頻度が決まればある程度機械的にノの頻度が決まるという現象は，ノの使用に人間の個性や意志が介在する余地が小さいことを示唆していると考えられる。

本研究ではコーパスの全データを使用せず，できるだけ外れ値が含まれないような基準を模索した。回帰分析において外れ値を除くことは重要だが，どのような方法を取れば，必要最小限のデータを除くことができるのか，今後さらに工夫していく必要がある。また，接続助詞やその他の助詞も文体に関係なく増減するのか，それを調査するためにはどのようなデータ選択を行う必要があるのか，これらを検討しながら調査を進めていくことが今後の課題である。

文　献

大野晋（1956）「基本語彙に関する二三の研究」『国語学』24，pp.34-46.

樺島忠夫（1955）「類別した品詞の比率に見られる規則性」『国語国文』24（6），pp.385-387.

柏野和佳子（2013）「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』Vol.4 No.1，pp.43-53.

国立国語研究所（2015）『BCCWJ 図書館サブコーパスの文体情報』（第 1 版）.（http：//pj.ninjal.ac.jp/corpus_center/anno/よりダウンロード可能）

森秀明（2017）「一般的な日本語テキストにおける助詞比率の規則性」『言語資源活用ワークショップ 2017 発表論文集』，国立国語研究所，pp.9-22.

西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也（2011）『『現代日本語書き言葉均衡コーパス』における電子化テキストの構築』，特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-03）.

山口昌也（2014）「第 3 章　文書構造の電子化」山崎誠（編）『講座日本語コーパス　2.書き言葉コーパス　設計と構築』朝倉書店，pp.45-67.

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる（2011）『『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2』，特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-04）.

山崎誠（2014）「言語単位と文の長さが品詞比率に与える影響」『第 5 回コーパス日本語学ワークショップ予稿集』国立国語研究所，pp.233-242.

「日本語日常会話コーパス」への談話行為アノテーションの試み：タグ選択が困難な事例に焦点を当てて

居關 友里子（国立国語研究所）
門田 圭祐（早稲田大学）
伝 康晴（千葉大学・国立国語研究所）

Dialog Act Annotation for the *Corpus of Everyday Japanese Conversation*

Yuriko Iseki (National Institute for Japanese Language and Linguistics)
Keisuke Kadota (Waseda University)
Yasuharu Den (Chiba University, National Institute for Japanese Language and Linguistics)

要旨

本研究では日常生活の中に生じた，具体的な文脈の中に埋め込まれた会話を扱った「日本語日常会話コーパス（CEJC）」に対する談話行為アノテーションの試みについて報告を行う。現在試行中の枠組みについて紹介した上で，実際のアノテーション作業の中で見出された談話行為の判断が困難な事例を示し，その要因について CEJC の特性を参照しながら議論する。

１．はじめに

会話は発話を通して行われる行為が一つ一つ積み重なることで形成される。これらの行為やその配列は，当該の会話がどのような性質のものなのか，どのような活動として参与者に体験されているのかを知る手がかりとなり得る(Mehan, 1979 ほか)。本研究はこれらを明らかにするための研究資源である，会話内で行われている行為の情報「談話行為情報」を，コーパスデータに付与する試みについて報告する。

近年日本語を扱ったコーパスの整備が進められ，ここには日本語で行われた会話を扱ったものも存在する(Koiso et al., 2016)。その多くは会話を収録するために，つまり研究のために生じた会話を収めたものである。その一方で，私たちが普段体験している会話の多くは，日常生活の中に存在する具体的な文脈の中に埋め込まれ自発的に生じた会話であり，日常生活に結びついた言語使用や社会的活動の組み立て，それらの多様性について明らかにするためには，こういった普段行われている様々な会話の収集および分析が不可欠である。このような動機に基づき，現在国立国語研究所では「日本語日常会話コーパス（CEJC）」の構築が進められている(Koiso et al., 2018)。おしゃべり，会議をはじめ，食事やスポーツ，場所の移動などといった話すこと以外の活動に従事しながらの会話など，さまざまな場面で行われた会話を対象に，会話場面を録画・録音したデータ，発話の転記，形態論情報をはじめとしたアノテーションデータの提供を予定している。ここに含まれる予定のものの一つが談話行為情報である。先に述べたように，談話行為情報は日常で体験される会話を構成する行為を研究するための資源として活用されることが期待され，談話行為情報付与の枠組みについてこれまでテストアノテーションと枠組みの修正を繰り返し，検討を重ねてきた(居關ほか，2017)。本研究ではこのうち特に CEJC が実際に取り扱う会話データへのアノテ

ーション作業に注目し，見出された問題点，検討を要する点について報告を行う。

２．先行研究と本研究の位置づけ

談話行為情報を付与する枠組みについては先行研究で多く扱われてきており，用いられているタグセットや作業者間の一致率，各データにおけるタグの使用分布などが報告されている(Carletta et al., 1996; Alexandersson et al., 1997; Allen and Core, 1997; Jurafsky et al., 1997; Bunt, 2009)。近年では談話行為に関する国際規格も提示され(Bunt et al., 2012; ISO 24617-2, 2012)，枠組みが整備されつつある。ISO 24617-2 をくだけた会話に対して適用した例(Fang et al., 2012)や日本語会話に対して適用した例もあり(平岡ほか，2013)，本研究もこれを一部援用する。

アノテーションの対象である CEJC は先述の通り対面で行われた会話を中心に扱っており，参与者たちは視覚的資源を用いることが可能である。ここでは発話だけでなく身体動作によるやり取りも行われ，日常生活で体験されている様々な非言語活動，例えば手作業や場所の移動，食事などがしばしば会話と並行して行われている。参与者は二名以上であり多い場合は十名以上が同じ場でやり取りに関わっている場面もある。ISO24617-2(2012)は非言語的やり取りや多人数の参与する会話も適用対象に含んではいるものの，こういった会話の特徴が談話行為情報の付与に際しどのような影響を生じ得るかについては具体的なデータに対するアノテーション作業を通して確認する必要がある。以下では本研究が用いる談話行為アノテーションの枠組みについて紹介した上で（3 節），日常生活の具体的な文脈に生じた会話に対するアノテーションの試みについて，特にアノテーションが困難であった点，検討が必要とされる点を中心に報告を行う（４節)。

３．アノテーションの枠組み

３.１．特徴

本研究の付与する談話行為タグは，当該発話で行われている行為やそれらの連なりに関する情報を提供する。この情報はコーパス利用者がそこで行われている相互行為の詳細な分析を行うための事例収集に際し，特定の談話行為を担う発話や，その連なりのパターンを拾い出すために用いることを想定している。そのため当該のやり取りに関する詳細な分析に直接用いるのではなく，当該のやり取りの骨組みとなる特徴が安定的に表示されることを目指している。

本枠組みは国際規格 ISO24617-2 を参考にしている。ISO24617-2 は様々なレベルに関わる多数のタグを備えている。会話は何らかのタスクを遂行するだけでなく，参与者が互いの発話をどのように受け取りそれに反応していくかといったことが常に判断されながら行われる多面的な活動である(Allwood, 1977)。これを受け，情報を多層的に付与し，同一発話箇所が複数の談話機能を担う「多機能性」を反映することが可能な枠組みとなっている。また発話順番は談話機能を付与するに際しては長く複雑な構造を持つとし，この複雑さに対応した情報を付与できるよう，タグの適用部分について談話機能の意味的な単位を自由に指定することができる。こういった特徴は当該データにおけるやり取りをより正確に捉えたい場合に非常に有益であると考えられる。その一方でアノテーション作業には膨大なコストが予想される。

これに対し本枠組みは先述の通り，事例を拾い出すために必要となる一定の情報を，安定的に提供することを目指している。タグセットは大きくなるとアノテーションの安定性を

低下させ，その一方で実際に用いられるタグはある程度限られるとされる(Popescu-Belis, 2008)。そこで本枠組みでは ISO24617-2 で細分化されているタグの一部はより汎用的なタグでまとめ，出現が少ないものについてはタグセットから外すこととした。また多機能性については生じている機能を限なくアノテーションに反映するのではなく，制限を設け限定的に反映可能とした。

以上を踏まえ本枠組みで主に扱うのは（１）基本的な談話機能と談話機能間の局所的な関係，および（２）談話の展開やより大局的な行為連鎖に関わる情報である。この両者を反映できる枠組みとして，アノテーションを二段階に分ける方法を用いる。第一段階では（１）の情報を， ISO24617-2 で設けられた主要な談話機能タグを用いて表示し，これに上乗せする第二段階で（２）の情報を，新たなタグを導入し表示する。また談話機能毎に個別に単位化を行うことはせず，CEJC における転記や統語情報付与の過程で採用している単位である「長い発話単位(Japanese Discourse Research Initiative, 2014)」を使用し，この単位毎に先述の談話機能を付与する。

３.２．付与規則

本枠組みが付与する談話行為情報は各発話が担っている談話機能の情報，および関係を結ぶ発話間の関係を示す依存関係情報である。

まず談話機能情報について，多機能性を一部反映するため情報を二段階に分けて付与する(居關・伝, 2018)。一つ目は（１）意味・語用論レベルタグ（第一レベルタグ）であり，当該会話にある目的達成や活動の進行に向けた振る舞いとその受け入れ，また社会的付き合いに関わる振る舞いに対して付与する。二つ目は（２）相互行為レベルタグ（第二レベルタグ）であり，相互行為の展開に関する情報や，第一レベルタグで付与した談話機能を大局的視点から拡張させた情報について付与する。一つの発話単位に対し第一レベルタグを一種類付与した上で（必須），第二レベルタグを任意で一種類まで選択する。笑いなどの非言語音や身体的振る舞いはタグ付与の対象外とするが，前後の談話行為の判断には適宜利用する。

第一レベルタグに属する談話機能は大きく三種類に下位分類される。タスクの進行に関わる振る舞い（タスク系），先行発話に対する自己の注意・知覚・理解・評価などに関する振る舞い（フィードバック系），社会的な関係の構築・維持に関わる儀礼的振る舞い（社会的付き合い管理系）である[1]。タグセットを表１に示す。

第二レベルタグでは，談話の開始や終了を示す発話に関する情報，何らかの行為が行われるに際し，特定の行為の準備から当該の行為に至るまでの過程を示す情報，あるいは理解や聞き取りの問題に対処するやり取りの情報など，相互行為の進行や展開に関する談話機能情報を付与する（表 2)。

これらの談話機能タグは，当該単位でなされている発話が談話においてどのような機能を果たしているのか，単位の末尾部分まで聞いた時点で最も妥当な機能について可能な限り特定的なタグを選択することとした。

[1] 三種類のいずれの談話機能にも該当しない場合は，この他に存在する「その他」から選択する。

表 1　第一レベルタグセット

タスク系	情報提供／独り言／情報要求／確認要求／返答としての情報・確認提供／依頼系（依頼・指示・命令・提案・勧誘）／依頼系への対処／申し出／申し出への対処／注意獲得／注意獲得への対処
フィードバック（FB）系	FB 肯定／FB 否定／FB 補完
社会的付き合い管理系	挨拶／謝罪／謝罪への対処／感謝／感謝への対処
その他	判断不可能／該当なし

表 2　第二レベルタグセット

準備系	準備の準備／準備／準備系への対処／準備が投射する本体
回収系	回収／回収への対処
修復系	修復開始／修復操作
談話構造化系	談話開始準備／談話開始／談話開始対処／談話終了準備／談話終了／談話終了対処

続いて依存関係情報を付与する。先述の談話機能タグを付与した上で，各発話が他の話者によってなされた発話との間に特定の関係を結ぶ場合にのみ付与する。結びつきの種類を示す依存関係タグを選択し，関係を結ぶ発話番号を指定する。依存関係タグは二種類がある。一つは「予測的依存関係」であり，「情報要求-返答」や「挨拶-挨拶」など，ある発話が次に特定の行為を要求するタイプの発話間の関係を示す。もう一つは「遡及的依存関係」であり，フィードバック応答に相当する発話とその応答の源となった発話との間に結ばれる関係を示す。ここにはあいづちの他に，発話の聞き取りや理解に何らかの問題が生じたことを示すような発話（修復開始: Schegloff, Jefferson & Sacks, 1977）も含まれる。

４．CEJC データへのアノテーションに際しての困難点

3 節で提示した枠組みを用い，CEJC で収録した多様な会話場面に対しアノテーションを試行した（18 場面，計 451 分）。第一レベルタグは第一著者あるいは第二著者が付与した。第二レベルタグは第一著者が単独で付与し，第一レベルタグのチェックを合わせて行った。これらの作業の際に見出された談話機能の判断に検討を要する例について，その理由と考えられる CEJC の特性を四つ取り上げ，これに沿って見ていく。

４.１.「多人数会話」に起因するもの

参与人数が多い場合，宛先の判断が困難な発話が多く見られた。特に親しい者同士が自由な雰囲気で話す場面では同時発話や会話の分裂(Egbert, 1997)がしばしば生じており，いずれの発話に対する反応か（依存関係情報）を特定しにくかった。また発話の宛先情報は談話機能の判断にも影響し得る。例えば発話の宛先となる参与者が発話者よりも知識のある者として想定される場合は「情報要求」としてタグ付与される発話が，発話者の方が知識のある者として想定される場合「情報提供」として聞かれ得る。このように発話の宛先の情報が直接談話機能の判断に関わる場合もある。視線の方向や移動のタイミングが参照できる場合はこの情報を用いたが不確かな事例が多く，このような場合には，いずれの先行発話の末尾付近や焦点要素(高梨ほか, 2010)付近で発話が開始されているのかといった発話のタイミ

ングを主に参照した。その他，発話内容や発話スタイル（敬体・常体の区別）などといった情報も複合的に判断材料に用いることで，宛先判断の情報を補いタグ付与を行った。

４.２.「マルチモダリティの使用」に起因するもの

発話と身体動作が組み合わさって生じている場合，両者がほぼ同時に生じ対応する機能を果たしている場合には特に問題なかったものの，生起するタイミングや担う談話機能にずれがある場合，談話機能の判断が身体動作に引っ張られやすく，発話それ自体の機能を判断しにくい場合があった。

【例 1】T001_002(1120.887)カラーボックスを二人で組み立てている[2]

01	A	で次こっちだ((板を持ち上げながら))	依頼系
02	B →	これは((A の動きに合わせて板を持ち上げる))	?
03	A	うん	FB 肯定
04	A	そうして	FB 肯定
05	B	ひっくり返すよ	依頼系
06	A	うんうん((板を持ち直す))	依頼系への対処

例 1 は A と B が一つのカラーボックスを組み立てている際のやり取りである。二人は同じ枠板に手を掛けながら 01 行目以降のやり取りを行っている。A は 01 行目で B と二人で支えていた板を持ち上げながら「で次こっちだ」と発話している。これは板を持ち上げる（そしてひっくり返す）提案として聞かれる（タグ候補 : [依頼系]）。注目したいのは続く 02 行目の B による発話である。02 行目「これは」という発話は，発話のみ聞くとどのような談話機能を果たしているのか明確でない。そこでこの時の B の身体動作を参照してみると，A が 01 行目で行った提案に応じるように板に手を添えている。このような場合に 02 行目の発話が担う談話機能を，発話ではなく身体動作で示されている行為である「提案に応じる」（タグ候補 : [依頼系への対処]）とすることが妥当なのかについては検討する必要がある。

【例 2】T007_017(204.469)会議中に参与者の一人 A が別の参与者 B の手元にお菓子を提供する

(01)	A	((複数あるお菓子を一つずつ机に置く))	(付与対象外)
02	B	あーすいません	申し出への対処
03	A →	((お菓子を置き終わる))はい	?
04	B	ありがとうございまーす	感謝

また例 2 は会議中にお菓子の受け渡しが行われている箇所である。お菓子を机に置き終えた A による 03 行目の発話「はい」は，発話冒頭が顕著に高い音で発話されている。この談話機能について，01 行目のお菓子を差し出す身体動作と合わせて申し出を果たしていると解釈することもできるかもしれないが，実質的に申し出を行っているのは 01 行目の身体

[2] 書き起こしは左列から順に「発話番号」,「発話者」,「第一レベルタグ」(必要に応じて更に「依存関係タグ」,「依存先発話番号」)を示す。一重括弧内の発話番号は，本枠組みでは本来発話番号を与えられない，言語的振る舞いを伴わない身体動作に便宜的に番号を与えたものである。また身体動作など映像に現れていた振る舞いに関する注記は，第一レベルタグ列の二重括弧内に示す。なお第二レベルタグについては本稿では省略する。

的振る舞いである。03 行目の発話はむしろ，申し出が終わったことを指標しているとみなすほうが適切かもしれない。このように身体的振る舞いがそこでのやり取りに強く関連している場合，発話それ自体が担う談話機能の判断に困難が生じやすかった。なお ISO24617-2(2012)は身体的振る舞いの談話行為情報についても発話と同様の基準で単位化を行いアノテーションするとしているが，具体的な適用事例は少なく，上述のような問題にどう対処していくのかについては独自に検討する必要がある。

４.３.「活動への関与の多様性」に起因するもの

収録された場面において参与者は，多くの場合会話以外の活動にも関与している。運転や食事，手作業，読書などがその例である。発話者，発話の受け手，あるいはその両者がこういった活動に従事している際の発話には，その発話が他者に向けられたものであるのか否か，判断が困難であるものが多数見出される。例 3 はその一例であり，携帯電話を操作しながら発話がなされている箇所である。参与者二人は各々の携帯電話で旅行に関するウェブページを閲覧しており，B は検索画面に出発日や泊数を入力している。B は発話を繰り返しているが，A は反応を返していない。

【例 3】K001_014(326.897)参与者二人それぞれが自身の携帯を操作し旅行サイトを見ている

01	B	現地出発日がー?	?
02	B	ついたちでいいの?	?
03	B	ん?	?
04	B	一 二 三 四 五	?
05	B	ついたちでいいのか	?
06	B	ん?	?
07	B	三十一でいいのか?	?
08	B	現地出発日? ((A に視線を向ける))	情報要求?

ここで行われている発話が他者に向けて産出されているか否かは，当該発話の談話機能の判断に直接関係する。実験室における自然会話の収録ではさほど観察されない一方で今回多く観察されたのは「独り言」と聞かれ得る発話である。これは述べ立て（タグ候補：[情報提供]）や質問（タグ候補：[情報要求]），特定の発話に対する反応（タグ候補：[FB 肯定]）と連続的なものであるため，出現の度にこれらの可能性を考慮しつつ談話機能を判断する必要がある。

次に挙げる例 4 も，発話者やその受け手になり得る者が当該会話以外の活動に関与している状況で交わされたやり取りの例である。一つの机に母親と息子二人の三人が着席している。S は手元のワークブックに向かって宿題に取り組みながら，問題に出てくる数字，また計算結果を小声で発話している。この発話は自身の独り言と聞くことができる一方で，向かい側に着席している母親の反応を伺う様子（02 行目）も観察されるため，何らかの他者志向を認めることができるかもしれない。

【例 4】T003_001(91.264)宿題をやっている S を正面からで母親 M が見ている

01	S	一十百千万((ワークブックの数字の桁数を数えながら))	?
02	S	二万五千分((発話末でちらと M に目線を向ける))	?

なお会話外活動に従事していない場合でも，参与人数が多い会話では同様の判断を求められる場合があり，これも相対的に参与の度合いが弱い者が生じやすくなることに起因すると考えられる。通常，他者意識の有無の判断に用いるのは発話者の身体の向き，視線の向きや移動のタイミングであるが(Goodwin, 1981; Heath, 1986 など)，会話以外の活動に従事している場合はその対象物に身体的志向が向けられていることが多くこれらの要因は判断に用いにくかった。このような場合には声の大きさや韻律操作の行われ方なども参照し，これらの要因を総合的に用い談話機能を判断した。

４.４.「環境への埋め込み」に起因するもの

CEJC の収録は実生活の営まれる現場において行われ，様々な環境に取り巻かれながら会話が行われている。この周囲の環境の特徴もアノテーションの難易に関わる要因の一つとして挙げることができる。アノテーションが難しかったのは，周囲の環境が頻繁に変化する場合，あるいは変化しない場合でも，環境内に参照され得る事物が多く置かれている場合である。例えば屋外での歩行場面では，協力者たちの周囲の景色や車の往来などといった環境が常に変化し，新しい情報がその都度もたらされる。あるいは団らんの際部屋に置かれたテレビなども，新しい情報を次々と提示する。こういった情報に対し協力者が反応を示す際，参照した対象物と考えられる事物は必ずしも言語化されず，また即時的な反応を示す場合は特に発話が短い（「ああ」「え」など)。このような発話は言語外の文脈を参照する度合いが高くなる一方で，当該発話が参照するものの候補が環境の中に多くあるため特定しにくく，その結果として談話機能や依存先の判断が難しくなっていると考えられる。これは一発話が長めであり言語的やり取りを安定的に参照できる「語り」などが行われている箇所のアノテーションが比較的容易であることと対照的である(居關ほか，2017)。

【例 5】C001_005 (165.807)道端の草花などについて話しながら二人で散歩している

01	B	もっとあそこの下の色が綺麗なのがいっぱいあるのよ	情報提供		
02	A	うん うん	FB 肯定	遡及的依存関係	01
03	B	散歩してると	情報提供		
04	B →	あらー	?	?	?

【例 6】T001_002(821.103)一つのカラーボックスを二人で組み立てる

01	A	今度の溝はこっちの溝に入れないといけない	依頼系		
02	B	溝を?	確認要求	遡及的依存関係	01
03	B	はーい	依頼系への対処	予測的依存関係	01
04	B →	えっ	?	?	?
05	A →	あー	?	?	?
06	A	じゃあ　これを	依頼系	?	?

例 5 は草花や建物などといった目に入ってきた光景について話しながら二人で散歩している場面である。03 行目まででは，01 行目以前で見つけた花について B が説明を加えている。注目したいのはこの直後，04 行目でなされた何らかの気付きを示す発話である。この発話がどのような依存関係をどの発話と結ぶのか，あるいはいずれの発話とも結ばないのかについて判断するために，04 行目の発話が何に反応してなされたものかに関する情報が

必要となるが，その候補は環境中に多く存在している。

また二人でカラーボックスの組立作業を行っている例 6 でも同様に，当該発話が参照するものが判断しにくい。例 6 で A と B の二人は同じ一つの板を手に持ち，これを別の板の溝に差し込もうとする（01–03 行目）。そして直後の 04 行目で何らかの問題が生じていることが示される。この問題が 03 行目までのやり取りに関わる問題であるのか，その後環境の中で生じた問題なのかは 04 行目の発話に付与するタグを左右し得る。

また 4.2 節で挙げた要因にも共通して関わる問題として，タグ付与対象外である環境や身体動作が会話の流れに絡む場合，ここでの依存関係情報は発話のみで完結し得る依存関係のようには付与できず，その構造は崩れた形で付与されることがあり得る。例 2'はその一例である。通常の行為の連なり方では，行為を開始するもの（例：[申し出]）がそれに対する反応（例：[申し出への対処]）に先行する。これに対し，例 2'の 01 行目のような非言語的振る舞いによる働きかけがなされた場合，現時点で用いている枠組みではこれらの行為間の関係を正確にタグに反映することができない。01 行目は発話を伴わない身体動作であるためタグ付与対象外であり，そのためここで付与されたタグのみを参照した場合やり取りは申し出への対処から始まっていることになる。

【例 2'】T007_017(204.469)会議中に参与者の一人 A が別の参与者 B の手元にお菓子を提供する

(01)	A	((複数あるお菓子を一つずつ机に置く))	(付与対象外)		
02	B →	あーすいません	申し出への対処	予測的依存関係？	？
03	A	((お菓子を置き終わる))はい	？		
04	B	ありがとうございまーす	感謝		

５．おわりに

本研究では，日常生活の具体的な文脈の中に生じた会話を扱うコーパスである CEJC データに対する談話行為アノテーションの試みについて，使用する枠組みの概要および実際のアノテーション作業から生じた問題点の報告を行った。見出された問題点は CEJC データの持つ特性「多人数会話」,「マルチモダリティの使用」,「活動への関与の多様性」,「環境への埋め込み」を軸に提示した。これらの特性は，参与人数や収録場所，話題，行われる活動などを統制するのではなく，CEJC が実生活の文脈の中で行われているやり取りの解明に向け，幅広い会話場面を柔軟に取り扱うことを目指し構築中のコーパスであることに結びついたものであると考えられる。4 節で取り上げたような，機械的にタグを割り当てていくことが難しいこれらのやり取りは，私たちが普段の相互行為の中で実際に体験し対処しているものである。こういったデータに対するアノテーションの枠組みを整備していくこと，またアノテーション作業を積み上げることを通して，日常生活を構成する言語行動の解明に繋げたい。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究（略称「日常会話コーパス」）」の研究成果を報告したものである。

文　献

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Maier, E., Schmitz, N. R. B., and Siegel, M. (1997). *Dialogue acts in VERBMOBIL-2. Verbmobil report 226*, DFKI Saarbrücken.

Allen, J. and Core, M. (1997). *Draft of DAMSL: Dialogue act markup in several layers.* Department of Computer Science, University of Rochester.

Allwood, J. (1997). A critical look at speech act theory. In Dahl, O. (ed.) *Logic, pragmatics and grammar*. Lund: Studentlitteratur. pp. 53-69.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *Proceedings of the AAMAS 2009 Workshop Towards a standard markup language for embodied dialogue acts (EDAML 2009)*, pp. 13-23, Budapest.

Bunt, H., Alexandersson, J., Choe, J.W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue act annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 430-437, Istanbul, Turkey.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. (1996). HCRC dialogue structure coding manual. Tech. rep. HCRC/TR-82, Human Communication Research Centre, University of Edinburgh.

Egbert, M. (1997). Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language and Social Interaction* 30, pp.1-51.

Fang, A. C., Cao, J., Bunt, H., and Liu, X. (2012). The annotation of the Switchboard corpus with the new ISO standard for dialogue act analysis. In *Proceedings of the 8th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pp. 13-18, Pisa, Italy.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic Press.

Heath, C.(1986). *Body movement and speech in medical interaction*. Cambridge: Cambridge University Press.

平岡拓也・Neubig, G.・Sakti, S.・戸田智基・中村哲 (2013)「説得対話コーパスの構築と分析」『情報処理学会研究報告』2013-SLP-99, pp. 41-46.

居關友里子・第十早織・伝康晴・小磯花絵 (2017)「日常会話コーパスのための談話行為タグの設計」『言語処理学会第 23 回年次大会発表論文集』pp.104-107.

居關友里子・伝康晴 (2018)「二段階発話連鎖アノテーション：意味・語用論と相互行為」シンポジウム「日常会話コーパス」Ⅲ，国立国語研究所.

ISO 24617-2 (2012). *Language resource management— Semantic annotation framework (SemAF)— Part 2: Dialogue acts.*

Japanese Discourse Research Initiative. (2014). 発話単位ラベリングマニュアル version 2.0. http://www.jdri.org/ resources/manuals/uu-doc-2.0.pdf.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Draft 13.* Institute of Cognitive Science Technical Report 97-02. University of Colorado, Boulder. https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y. and Usuda, Y. (2018). Construction of the *Corpus of Everyday Japanese Conversation*: An interim report. In *Proceedings of the 11th International Conference on Language Resources and Evaluation,* pp.4259-4264, Miyazaki, Japan.

Koiso, H., Tanaka, Y., Watanabe, R., and Den, Y. (2016). A large-scale corpus of everyday Japanese conversation: On methodology for recording naturally occurring conversations. In *Proceedings of*

LREC 2016 Workshop on Casual Talk among Humans and Machines, pp. 9-12, Portoroz, Slovenia.

Mehan, H. (1979). *Learning lessons: The social organization of classroom instruction*. Harvard University Press, Cambridge, Massachusetts.

Popescu-Belis, A. (2008). Dimensionality of dialogue act tagsets: An empirical analysis of large corpora. *Language Resources and Evaluation* 42:1, pp.99-107.

Schegloff, E. A., Jefferson, G. & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language* 53:2, pp.361-382.

高梨克也・常志強・河原達也 (2010)「聞き手の興味・関心を示すあいづちの生起する会話文脈の分析」『人工知能学会研究会資料』SIG-SLUD-A903, pp. 25-30.

「『了解』は使わないように」「了解です！」

髙橋圭子・東泉裕子・佐藤万里（フリーランス）

‘Don’t Say “Roger”.’ ‘Roger-*desu*!’

Keiko Takahashi, Yuko Higashiizumi, Mari Sato (freelancers)

要旨

近年,ビジネスマナーに関する書籍やウェブ上において,「了解」は上から目線の言葉で失礼なので使わないほうがよい,とする記述が少なからず見られる。本発表では,各種コーパスの用例,辞書やマナー本の記述などを調査し, (1)応答詞としての「了解」とその派生形式, (2)「了解は失礼」説,のそれぞれについて,出現と広がりのさまを探る。

１．はじめに

現代日本語においては,情報伝達や指示・依頼などへの応答として,「了解」という語が使われることがある。例えば,『現代日本語書き言葉均衡コーパス(Balanced Corpus of Contemporary Written Japanese: BCCWJ)』においては,次のような用例がある[1]。

(1) 「おい,来たぞ。門を開けてやれ」彼は下にいる同僚にハンディ・トーキーで呼び掛けた。「了解。でも八分の遅刻で助かったな。これが三十分も遅れたりしたら…」ハンディ・トーキーから声が返った。下もやきもきしていたのだろう。

【出典】 BCCWJ サンプル ID：LBb9_00066 図書館・書籍
中島渉(著) 『サザンクロス流れて』1987 年

(2) 「(略)ここから先の事務的な手続については,労務部と直に詰めて貰いたい」小暮は素っ気なく答えた。確かに後任の人事,ワーキング・ビザの取得,引き継ぎ事務などを考えれば,事はそう簡単に運ばない。「了解しました,本日は,恩地前委員長の帰国をご確約下さり,有難うございました」沢泉は,恩地の帰国が記録に残るように,重ねて云った。

【出典】 BCCWJ サンプル ID：OB5X_00112 特定目的・ベストセラー
山崎豊子(著) 『沈まぬ太陽 2(アフリカ篇 下)』1999 年

(3) 「そうか・・・惑星,あんまり無茶するなよ！」「了解です！」俺は興奮する気持ちと焦る気持ちを抑えながら冷静に走ることを心がけた。

【出典】 BCCWJ サンプル ID：OY14_44275 特定目的・ブログ 2008 年

本発表では,これらを「了解」の応答詞用法と呼ぶ。詳細は後述するが,BCCWJ その他のコーパスには,「了解」の応答詞用法の形式として「了解」「了解した」「了解しました」「了解いたしました」「了解です」などが認められる。

一方,ビジネスマナーに関する書籍やウェブにおいては,「了解」は失礼なので使わないほうがよい,と記述するものがある。例えば,次のようなものである。

[1] 以下,『BCCWJ』の用例には，順に，サンプル ID，レジスター，著者 (書籍)，書名，出版年を記す。また,下線は発表者による。

(4) 「了解しました」はビジネス上あまりよくありません。誤解されがちですが,“了解”は敬語ではありません。ていねいな意味は含まれていないので気をつけましょう。(野村 2011, p.49)

(5) 「了解」には「理解して承認すること」という意味もあります。「承認」には上から下という印象があるため,人によっては「了解しました」に違和感を持つ人もいます。（梅津 2013, p.36）

(6) 決して相手を傷つけたり,不愉快な思いをさせようとして使った言葉じゃないのに,結果として失礼になってしまった。こんな悲劇を防ぐためにも,言っていい言葉,言ってはいけない言葉について勉強しておきましょう。
■上司から指示や命令を受けたとき
上司から「これやっといて」などと指示や命令を受けたとき,なんと言って返事してますか？ありがちなのが「分かりました」「了解です」の言葉です。
何がいけないの？ちゃんと「分かった」と返事してるじゃない？と思うかもしれませんが,実はこれらは同僚や部下に対して使う丁寧語。上司に対して使ってしまうと失礼に当たるのです。正解は「かしこまりました」「承知致しました」という謙譲語なので,間違えないようにしましょう。
【出典】http://news.livedoor.com/article/detail/7986822/ 2013 年

本発表では,このような見方を「了解は失礼」説と呼ぶ。他方,これに異を唱える論もある。例えば,飯間(2016)は(7)のように述べている。

(7) 「了解いたしました」は失礼なことばではない
「了解」ということばは失礼である――と言われるようになったのは,ここ 10 年ほどのことです。誤解に基づくのですが,日本語関係の一般書で無責任にそう説明するものが増えました。もともと,人に分かったと意思表示をするときには,「言われなくても分かってるよ」というような語感が伴いやすい。そこで,なるべく丁寧な,へりくだった表現が必要です。(略)「了解」は,「分かる」の漢語表現にすぎず,特に敬意はありません。「了解しました」は「分かりました」とほぼ同等で,もう少し敬意がほしいのは確か。「了解です」だけなら,なるほど失礼です。これは「納得です」「承知です」などがぞんざいなのと同様です。しかしながら,丁寧かつへりくだった「いたしました」をつけて「了解いたしました」と言えば,何ら失礼ではなく,敬語として十分です。(略)
【出典】https://twitter.com/IIMA_Hiroaki/status/741895366618927104/photo/1/ 2016 年

本発表では, コーパス,辞書,書籍などを調査し,「了解」の応答詞用法,「了解は失礼」説およびそれに対する反論について,出現と広がりの様相を明らかにすることを目指す。

２．先行研究

「了解」の語義については,中山(2013, 2014)が,『太陽』『新潮文庫の 100 冊』『現代日本語書き言葉均衡コーパス(BCCWJ)』の各コーパスを用いて,明治期の「理解」から現代の「理解＋承認」という意味に変化してきたこと,「了解」には「諒解」「領解」「領会」という表記もあるが意味の相違には関わらないこと,などを明らかにしている。本発表も中山に従い表記の相違は不問とする。ただし,中山の調査範囲では応答詞用法はごく少数である。

「了解」の応答詞用法については,後述の(8c)にも指摘のあるように,無線通信の用語に関わる。無線通信においては,「受信した（Received)」「受領（Reception)」を表す“R”の音標アルファベット(phonetic alphabet)[2]として“Roger”が用いられており,日本語ではこれに「了解」があてられている。これは,『電波法無線局運用規則』(1950 年 11 月 30 日電波管理委員会規則第 17 号)第十四条の規定する別表第四号にも記載されている[3]。

「了解は失礼」説については,国立国語研究所 HP「よくある『ことば』の質問」(2012)にも「『了解しました。』は敬意表現にならないか」という質問が寄せられている。その回答をまとめると(8)のようになる。

(8) a. 「了」「解」ともに「よくわかる・さとる」という意味で,「了解」は類似の動詞を重ねた言葉である。待遇の要素は含まれていない。
b. 日本語での漢語の用法には,短くてすむ,抽象度が増す,その場での「あらたまり」が増す,などという副次的効果がある。
c. 電話や無線などの通信,業務でのやり取りなどで,相手の言ったことが分かった,という証拠に短く「了解」と返事をすればよい,という用法がある。
d. 国語辞典の中には,相手からの指示・命令への返答としての使い方や,ぶっきらぼうで敬意が不足という注を記述しているものもある。
e. 本来待遇の意味を含まない「了解しました」が「ぶっきらぼう」に感じられる理由は理屈ではっきりとは説明はつかない。おそらく無線通信や業務・作業の際の返答がなにかしら影響を及ぼしているかもしれない,といった程度で,明確には分からない。
f. 「わかりました」「かしこまりました」「承りました」など,もっとふさわしい言葉がいろいろあるのだから,それに譲ればよいのではないか。

また,真鍋(2014) ,菊池(2016)は「了解は失礼」説の出現時期を多くのマナー本やウェブ記事から検討している。真鍋(2014)の推定は 2008 年頃である。菊池(2016)は,2009～2010 年頃のマナー本がきっかけとなり 2011 年頃からウェブ上で広がった,これはウェブメディアブームの時期であり,普通使われているこの言葉は実は間違いであるといった逆説的でアクセス数を集めやすい記事が量産された,その一つがこの「了解は失礼」説である,と推定している。

３．コーパスにおける「了解」の応答詞用法

表 1 は,各種コーパスにおける「了解」の応答詞用法を形式ごとにまとめたものである。コーパスの略称は次のとおりである。

- 【CHJ】: 国立国語研究所『日本語歴史コーパス(Corpus of Historical Japanese)』
- 【BCCWJ】: 国立国語研究所『現代書き言葉均衡コーパス(Balanced Corpus of Contemporary Written Japanese)』
- 【名大】: 国立国語研究所『名大会話コーパス』
- 【青空】: 国立国語研究所『青空文庫パッケージ』(20180401)

[2] 電話や無線通信で文字を伝えるときに聞き間違いを防ぐため使われる,各文字を示す単語。その一覧を音標コード(phonetic code)または通話表という。欧文では A, B, C にそれぞれ Alfa, Bravo, Charlie, などがあてられている。(JapanKnowledge 所収『デジタル大辞泉』より)

[3] ただし,電波法の前身である『無線電信法』(1915 年 11 月 1 日施行)には「了解」の語は見られない。

- 【新潮】: 新潮社(1995)『新潮文庫の 100 冊』
- 【戦時】: 現代日本語研究会(2004)『戦時中の話しことば』
- 【職場】: 現代日本語研究会(2011)『女性のことば・男性のことば(職場編)』
- 【日常】: 現代日本語研究会(2016)『日常会話のことば』

表 1　各コーパスにおける「了解」の応答詞用法

コーパス	CHJ	青空	戦時	新潮	職場	名大	BCCWJ	日常
データ収録年代[4]	1874-1925	1891-1990	1936-1955	1894-1995	1993-2000	2001-2003	1976-2008	2011-2014
了解	0	7	0	2	1	3	116	0
了解した	0	1	0	0	0	0	12	0
了解しました	0	0	0	0	0	0	27	1
了解いたしました	0	0	0	0	0	0	5	0
了解です	0	0	0	0	0	0	13	0
了解だ	0	0	0	0	0	0	1	0
その他	0	0	0	0	0	0	0	0
計	0	8	0	2	1	3	174	1

【CHJ】においては,「了解」436 例のすべてが明治・大正(雑誌)であったが,応答詞用法は見出せなかった。また,戦時中のラジオドラマ台本をコーパス化した【戦時】においても,「了解」の応答詞用法はなかった。

今回の調査範囲における「了解」の応答詞用法の初出は,【青空】における海野十三「宇宙戦隊」(『海軍』1944 年 5 月～1945 年 3 月)の 5 例である。そのうち,はじめの 2 例を(9), (10)に示す。

(9)　「機長」帆村が上を向いて叫んだ。「おう」山岸中尉が答える。「高度二万メートルを突破しました」「はい,了解」

【出典】 青空 作品 ID：3367 海野十三(著) 「宇宙戦隊」1944～1945 年

(10) 高度を二万九千まであげてみたが,異変はさらに起らない。そこで望月大尉は,「高度二万八千に戻り,水平飛行で偵察を継続するぞ」と,山岸中尉に知らせた。「了解」

【出典】 青空 作品 ID：3367 海野十三(著) 「宇宙戦隊」1944～1945 年

この 2 例の話し手はともに二号艇長の山岸中尉であり,聞き手は(9)は部下の帆村, (10)は一号艇長の望月大尉である。戦隊の上下の序列と関係なく,応答詞「了解」は用いられている[5]。

なお,海野十三(1897−1949)は『日本大百科全書』(JapanKnowledge 所収)によれば日本にお

[4] 【青空】の底本初版発行年は 1891-1990,入力に使用された版は 1891-2002 年である。また,【新潮】については,今後,雑誌初出年,単行本初版,文庫初版の各年を精査し,修正していく必要がある。

[5] 【青空】における「了解」の応答詞用法は他に,江戸川乱歩「サーカスの怪人」(『少年クラブ』1957 年), 江戸川乱歩「奇面城の秘密」(『少年クラブ』1958 年 1 月号～12 月号)に各 1 例がある。なお,「奇面城の秘密」における表記は「りょうかい」である。また,「了解した」1 例は海野十三『海底都市』(日本正学館(冒険少年文庫) 1948 年) である。

ける SF の先駆者の一人であり,早稲田大学理工学部から逓信省電気試験所勤務を経て作家生活に入ったという。

【新潮】においては「了解」104 例のうち応答詞用法は 2 例,井上ひさし(1970)『ブンとフン』および赤川次郎(1984)『女社長に乾杯！』各 1 例である。(11)は前者の例で,警察庁長官と手下の悪魔の会話である。

(11) 「ブンが,フン先生の家にあらわれたのだ。さ,はやく行け！」「わかってます。で,フン先生というひとの家の所番地は？」「市川市のはずれに下総の国分寺という有名なお寺がある。そのお寺の裏側の畑の中の一軒家だ」「了解。いってきまーす！」
【出典】 新潮 井上ひさし(著) 『ブンとフン』1970 年

【BCCWJ】においては,「了解」1308 例のうち応答詞用法は 174 例である。表 2 はその形式をサブコーパスごとにまとめたものである。

表 2　BCCWJ における「了解」の応答詞用法

	出版		図書館	特定目的				計
	新聞	書籍	書籍	ﾍﾞｽﾄｾﾗｰ	ブログ	知恵袋	国会	
データ収録年代	2001-2005	2001-2005	1986-2005	1976-2005	2008	2005	1976-2005	
了解	2	42	42	11	19	0	0	116
了解した	0	2	4	1	4	1	0	12
了解しました	0	9	6	5	3	4	0	27
了解いたしました	0	0	1	0	2	1	1	5
了解です	0	1	0	0	10	2	0	13
了解だ	0	0	0	0	1	0	0	1
計	2	54	53	17	39	8	1	174

表 1・表 2 から次のような傾向が読み取れる。「了解」の応答詞用法は「了解」単独を基本としつつ,「了解した」「了解しました」「了解いたしました」などバリエーションが増加していく。特に 2008 年の【BCCWJ】ブログにおける「了解です」の増加が注目される。また,【BCCWJ】ブログには(12)のように「だ」がつく形式もある。

(12) んで,こっちでのバイトがまだあるらしく,その時は住む場所が無いので泊めてくれとのことですw　まぁかなりお世話になっているのでね,それぐらい了解だぜ。近所付き合いが多かった奴がもうあの部屋にはいないのだなと思うと,何ともいえない思いです。
【出典】 BCCWJ サンプル ID：OY14_47231 特定目的・ブログ 2008 年

ここまでの調査をまとめると,「了解」の応答詞用法は【青空】・【新潮】・【BCCWJ】「出版・書籍」「図書館・書籍」「特定目的・ベストセラー」といったフィクションや,【BCCWJ】ブログのようなウェブメディアに多く見られる。詳細は紙幅の都合上省略するが,前者のフィクションでの用例は SF・戦記物・推理ミステリ・ファンタジー・ライトノベルなどと呼ばれるジャンルに目立ち,対面会話のほか,無線や電話などの通信機器を媒体とする例も多

い。また,「了解」単独の応答詞用法は上下関係に関わらず用いられている.

４．国語辞典における「了解」の記述

国語辞典における,「了解」の項の応答詞用法および待遇面に関する記述を表 3 にまとめる[6]。語義は,中山(2013, 2014)に倣い「意味 A：理解」と「意味 B：理解＋承認」に分け,さらに応答詞用法を別項とした。哲学用語としての記述は省略した。

表 3　国語辞典における「了解」の記述

辞典	版	年	意味 A：理解	意味 B：理解＋承認	応答詞用法
国語大辞典言泉	初	1986	①よく理解すること。また、理解して承認すること。命令や指令の伝達の際、「わかりました」の意で用いられる。諒解。「『ただちに現場へ急行して下さい』『了解』」		
日本国語大辞典	2	2000-2002	①はっきりとよくわかること。よく理解すること。また,理解して承認すること。		(なし)
明鏡国語辞典	初	2002	①物事の意味・内容・事情などを理解すること。「話しを聞くや否やその意味をーした」「両者間には暗黙のーがある」	②理解した上で承認すること。承知・承諾すること。「雑誌社に転載のーを求める」「その件についてはー済みだ」	③無線通信で,通信内容を確かに受け取ったことを表す語。
新明解国語辞典	6 7	2004 2011	①事の内容や事情が∆分かる(分かって納得する)。「ーを∆得る(とりつける・求める)」／「ー事項5」	②思いやって事情などを容認すること。	運用：①は,相手からの指示・命令に対して納得したことを表す返事として用いられることがある。例,「『救援隊ただちに出動せよ』『了解』」
大辞林	3	2006	①事情を思いやって納得すること。理解すること。のみこむこと。了承。領解。領会。 「事情をーする」 「 ーできない」		② 無線などの通信で, 通信内容を受け取ったことを表す語。 「『ただちに行動を開始せよ』『ー』」
広辞苑	6 7	2008 2018	①さとること。会得すること。また,理解して認めること。諒解。「ーを求める」「暗黙のー」		(なし)
学研現代新国語辞典	4 5	2009 2012	①物事の筋道・事情などを,よく理解して承認すること。「上司のーを得る」		② 〔無線通信などの対話で〕「分かった」「聞こえた」の意で使う語。
明鏡国語辞典	2	2010	①物事の意味・内容・事情などを理解すること。「話しを聞くや否やその意味をーした」「両者間には暗黙のーがある」	②理解した上で承認すること。承知・承諾すること。「雑誌社に転載のーを求める」「その件についてはー済みだ」表現：近年目上の人の依頼・希望・命令などを	③無線通信で,通信内容を確かに受け取ったことを表す語。「『こちら本部,どうぞ』『ー,こちら現場,どうそ』」

6　応答詞用法または待遇面に関する記述のある辞典（新明解・大辞林など）については,その出現の版以降を記載した。そのような記述のない辞典（日国・広辞苑など）については最新版のみを記載した。

				承諾する意に使う向きもあるが，慣用になじまない(ぶっきらぼうで,敬意が不足)。「分かりました」「承知しました」のほか,「承りました」「かしこまりました」などを使いたい。	
大辞泉	2	2012	①物事の内容や事情を理解して承認すること。了承。「―が成り立つ」「来信の内容を―する」[用法]了解・理解――「彼は友の言う意味をすぐに了解(理解)した」「その辺の事情は了解(理解)している」など,意味がわかる,のみ込むの意では,相通じて用いられる。◇「了解」には,相手の考えや事情をわかった上で,それを認める意がある。「暗黙の了解を得る」「お申し越しの件を了解しました」◇「理解」は,意味や意図を正しくわかる意が中心となる。「文章を理解する」「何を言っているのか理解できない」◇「了解できない」は,意味はわかるが承認できないの意になり,「理解できない」は単に意味がわからないの意になる。◇類似の語「了承」は「了解」とほぼ同じに使うが,「了解」よりも承認する意が強い。「上司の了承を得る」「双方とも大筋で了承した」	(なし)	
例解新国語辞典	8	2012	事情や理由などを,よくわかってみとめること。用例：了解をえる。了解をもとめる。暗黙の了解がある。「現地へ急行せよ」「了解」。類：了承。注意：目上の人から言われたことに対しては,「了解しました」でなく「承知しました(いたしました)」とこたえる。		
三省堂国語辞典	7	2014	相手の行っている内容を理解(して承知)すること。「－をもとめる・－できない・－しました［『－です』とも］」		

５．マナー本における「了解」の記述

東京都内の公共図書館において,「敬語」「マナー」「ビジネスメール」などで検索してヒットした書籍 319 冊の「了解」の応答詞用法に関する記述を調査した。調査結果を表 4 にまとめる。

表４　調査対象書籍と「了解」の記述

発行年	記述なし	非「了解は失礼」説	「了解は失礼」説	計
～2002	49	0	0	49
2003	6	2	0	8
2004	6	0	1	7
2005	10	3	1	14
2006	16	2	3	21
2007	18	2	0	20
2008	16	6	4	26

2009	19	2	2	23
2010	19	4	4	27
2011	6	1	5	12
2012	16	2	7	25
2013	13	4	10	27
2014	13	0	3	16
2015	8	0	5	13
2016	7	0	2	9
2017	12	0	6	18
2018	1	0	3	4
計	235	28	56	319

今回の調査範囲では,2002 年までは「了解」に関する記述は見当たらない。2003 年以降数年間は,「了解」の使用を良い例として紹介する“非「了解は失礼」説”と,「承知しました」「かしこまりました」などに言い換えるべきであるとする“「了解は失礼」説”が拮抗する。そして,“「了解は失礼」説”の優位が確定するのが 2011 年であり,これは菊地(2016)の指摘するウェブメディア隆盛期と軌を一にする。そして 2014 年以降, “非「了解は失礼」説”は姿を消す。

「了解は失礼」説の根拠は,(13)のようにまとめられる[7]。

(13) a. 相手への尊敬の意がない,丁寧さが不足,など敬語として不適切だから：22 冊。
b. 上から目線の言葉だから：12 冊。うち a.d.各 1 冊,c.4 冊と重複。
c. 理解し承認するという意味だから：7 冊。うち b.4 冊,f.1 冊と重複。
d. 軍隊・警察のイメージ：3 冊。
e. 事務的だから：3 冊。すべて a.と重複。
f. 簡略で軽い感じだから：3 冊。うち a 2 冊,c 1 冊と重複。
g. カジュアル・くだけた表現だから：2 冊。
h. メール語だから：2 冊。

最も多く挙げられているのは, (13a)「了解」が敬語でないという点であり,これが(13e)事務的, (13f)簡略・軽い,という受けとめ方にもつながっている。目上の相手に対しては,より重厚な敬語表現が適切である,とする主張のようである。第 1 節で引用した(4)は(13a)の, (5)は(13b)および(13c)の例である。

(13d)の軍隊・警察のイメージは,3 節でみた SF の戦隊ものや推理ミステリでの多用と関連するかもしれない。加えて,アニメやゲームなどにおいても同様の傾向が指摘できそうであるが,論証は今後の課題である。ただし,軍隊・戦隊もの,警察もののすべてに「了解」が多用されているわけではない。【戦時】には軍隊生活を描いた作品もあるが用例は見出せない。また,目視の調査であるが,松本(1971) ,鎌田(1989) ,手塚(2007)にも該当例はない。松本(1971)

[7] 1 冊で複数の根拠を挙げるものもあれば,根拠が示されていないものもあるので,合計数は「失礼」説 56 冊とは一致しない。

は刑事を主人公とする推理小説で,雑誌初出は 1957-58 年である。手塚(2007)は戦時下を描いた作品集で,収録作品 7 本の初出は 1968-79 年である。鎌田(1989)は刑事ドラマのシナリオ集で,収録作品 8 本の放映日は 1973-79 年である。

(13g)カジュアル・くだけた表現というイメージの由来は明らかでないが, (13f)簡略・軽い, (13h)メール語という点とあるいは関連するかもしれない。直接「失礼」である根拠とした 2 冊を含め,「了解」が若者のメールに多用されていることに触れている書籍は少なくない。また,今回の調査範囲の書籍には言及はないが,若者のメールや SNS では「了解」の短縮語である「り(ょ)」なども多用され,顔文字やスタンプなどもあるという。今後の調査課題である。

６．暫定的考察と課題

「了解」の応答詞用法と無線用語「了解」の関わりはまず間違いないだろう。現代より相当劣悪であった通信状況のもとでは,待遇的配慮よりも,緊急かつ重要な指令が無事受信されたことを手短に伝える効率性が優先されたのは当然である。Brown & Levinson(1987)のポライトネス理論においても,「オン・レコードであからさまに(Bald on record)」言語行為が行われる場面として,重大な緊急事態や,チャネル・ノイズなどが原因で最大限の効率を優先しなければならないようなケースを挙げている(田中監訳 2011, pp.124-127)。

やがて,電話や対面会話などにも「了解」が用いられ,効率性だけでなく待遇的配慮も必要とされるようになり,「了解しました」「了解いたしました」などが出現した。これらとの比較により,単独の「了解」が「乱暴でぶっきらぼう」(『明鏡国語辞典第二版』2010, 梅津 2012 など）という印象を与えるようになってしまったのかもしれない。ただし,これら派生形式の用例数は「了解」単独の用法を越えるには至っていない。

一方,「了解です」は用例数をかなり伸ばしている。これは,「です」の勢力拡大(井上 1998, 1999, 2007),ニュースやブログなどにおける「動名詞＋です」の多用(田中 2012, 鈴木 2010, 2011, 2012),日本語の名詞指向性(新屋 2014)などと関わる現象の一端ではないかと思われる。

ただし,「です」は幼稚・舌足らずとの印象を免れていない。文化庁(1950)で認められた形容詞述語文にも違和感の声は消えていない。「なるほどですね」という相づちにも疑問の声があがっている。「了解です」への抵抗感が,「了解は失礼」説の出現の一因となった可能性もあるのではないか。

以上は,今回の調査結果を踏まえた暫定的考察であるが,まだ論証できていない推測の部分もある。今後,ウェブや SNS の用例,アニメ,ゲームなど,さらに調査を進めたい。

謝　辞

無線用語についてご教示くださった日本アマチュア無線連盟,ウェブや SNS についてご教示くださったインフォーマント諸氏に感謝申し上げます。

文　献

Brown, Penelope and S. C. Levinson (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press. 田中典子監訳(2011).『ポライトネス：言語使用における,ある普遍現象』研究社

飯間浩明(2016).「『了解いたしました』は失礼なことばではない」https://twitter.com/IIMA_Hiroaki/status/741895366618927104/photo/1

国立国語研究所「よくある『ことば』の質問」(2012).「『了解しました。』は敬意表現になら

ないか」　http://pj.ninjal.ac.jp/QandA/vocabulary/ryokai/
井上史雄(1998).『日本語ウォッチング』岩波新書
井上史雄(1999).『敬語はこわくない－最新用例と基礎知識』講談社現代新書
井上史雄(2017).『新・敬語論：なぜ「乱れる」のか』NHK 出版
梅津正樹(2012).『敬語のレッスン』創元社
梅津正樹(2013).『知らずに使っている実は非常識な日本語』アスコム
鎌田敏夫(1989).『ボスー太陽にほえろ！傑作選』立風書房
菊池良(2016).「『了解しました』より『承知しました』が適切とされる理由と、その普及過程について」https://liginc.co.jp/246919
現代日本語研究会(2004).『戦時中の話しことば』ひつじ書房
現代日本語研究会(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(2016).『談話資料　日常生活のことば』ひつじ書房
新潮社(1995).『CD-ROM 版　新潮文庫の 100 冊』
新屋映子(2014).『日本語の名詞指向性の研究』ひつじ書房
鈴木智美(2010).「ニュース報道における『{動名詞(VN)／感動詞相当句} +です』文について－『現地を緊急取材です』『老舗料亭に問題発覚です』－」『東京外国語大学留学生日本語教育論集』第 36 号, pp.57-70.
(http://repository.tufs.ac.jp/bitstream/10108/57678/2/jlc036005_ful.pdf よりダウンロード可能)
鈴木智美(2011).「ブログ等に見られる『{動名詞(VN)／感動詞相当句} +です』文について－『～に感謝です』『～をよろしくです』の意味・機能－」『東京外国語大学留学生日本語教育論集』第 37 号, pp.15-28.
(http://repository.tufs.ac.jp/bitstream/10108/63375/2/jlc037002_ful.pdf よりダウンロード可能)
鈴木智美(2012).「ニュース報道およびブログ等に見られる『～です』文の意味・機能～『～を徹底取材です』『～に期待です』『～をよろしくです』～」『東京外国語大学論集』第 84 号, pp.341-357.
(http://repository.tufs.ac.jp/bitstream/10108/70857/2/acs084017_ful.pdf よりダウンロード可能)
田中伊式(2012).「ニュース報道における『名詞＋です』表現について～『イチロー選手が電撃移籍です』『尖閣諸島で新たな動きです』～」NHK 放送文化研究所『放送研究と調査』62:10, pp.16-29.
(http://www.nhk.or.jp/bunken/summary/research/report/2012_10/20121002.pdf よりダウンロード可能)
手塚治虫(2007).『「戦争漫画」傑作選』祥伝社
中山健一(2013).「「了解」の意味の変遷 —19 世紀末から現代にかけて—」『国立国語研究所第 3 回コーパス日本語学ワークショップ予稿集』pp.169-178.
(https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_22.pdf よりダウンロード可能)
中山健一(2014).「漢語「了解」の意味変化—太陽コーパスの分析を中心に—」『茨城キリスト教大学紀要』48, pp.1-14.
(https://ic.repo.nii.ac.jp/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=353&item_no=1&page_id=25&block_id=38 よりダウンロード可能)
野村恵理奈(2011).『たった 5 秒で相手の心をつかむ一言の力』大和書房
文化庁(1950).『これからの敬語』

http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kakuki/01/tosin06/index.html
松本清張(1971).『点と線』新潮文庫
真鍋宏史(2014).「了解は失礼か?」
http://d.hatena.ne.jp/takeda25/20140209/1391937000

辞　典

見坊豪紀・市川孝・飛田良文・山崎誠・飯間浩明・塩田雄大(編) (2014)『三省堂国語辞典　第七版』三省堂
北原保雄(編) (2002)『明鏡国語辞典　初版』大修館書店
北原保雄(編) (2010)『明鏡国語辞典　第二版』大修館書店
金田一春彦・金田一秀穂(編) (2009)『学研現代新国語辞典　第四版』学習研究社
金田一春彦・金田一秀穂(編) (2012)『学研現代新国語辞典　第五版』学習研究社
尚学図書(編) (1986)『国語大辞典　言泉』小学館
林四郎(監修) (2012)『例解新国語辞典　第八版』三省堂
松村明(編) (2006)『大辞林　第三版』三省堂
松村明(監修)(2012)『大辞泉　第二版』　小学館
日本国語大辞典第二版編集委員会 (2000-02) 『日本国語大辞典　第二版』小学館
新村出(編)(2008)『広辞苑　第六版』岩波書店
新村出(編) (2018)『広辞苑　第七版』岩波書店
山田忠雄・柴田武・倉持保男・山田明雄・酒井憲二(編) (2004)『新明解国語辞典　第六版』三省堂
山田忠雄・柴田武・酒井憲二・倉持保男・山田明雄・上野善道・井島正博・笹原宏之(編) (2011)『新明解国語辞典　第七版』三省堂
JapanKnowledge　https://japanknowledge.com/

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「言語データベースとソフトウエア」
http://www2.ninjal.ac.jp/lrc/index.php?%B8%C0%B8%EC%A5%C7%A1%BC%A5%BF%A5%D9%A1%BC%A5%B9%A4%C8%A5%BD%A5%D5%A5%C8%A5%A6%A5%A7%A5%A2
総務省「電波法無線局運用規則」
http://www.tele.soumu.go.jp/horei/reiki_honbun/a723930001.html
電気通信主任技術者総合情報「無線電信法」
http://asaseno.aki.gs/houki/musendenshinhou.html

NWJCにおける敬語使用とレジスターとの関係

金 賢眞（大阪大学大学院文学研究科）

Relationship between Use of Honorific and Register in NWJC

Hyunjin Kim (Graduate School of Letters, Osaka University)

要旨

発表者は、現在、NWJC を用いた敬語研究を志向し、その一環として、謙譲表現「ご～する」を尊敬用法に用いた「誤用例」の調査・分析を進めている。WEB コーパスは、この種の「誤用例」を収集するのに好適な言語資料だが、残念ながら、NWJC にはレジスターの情報が付与されていない。敬語の使用および誤用にはレジスターによる差があると考えられ、たとえば、ブログなどでは敬語の使用は多くないが、誤用が現れやすい傾向があり、それに対して、企業や公的機関のホームページなどでは敬語が多く使用されるものの、誤用は少ないと予想される。しかし、こうした予想を NWJC によって直ちに確認することはできない。そこで、本発表では、NWJC から得た「ご～する」の使用例を仮設したレジスターに分類し、動詞別の使用頻度や動詞ごとの誤用率がレジスターによってどのように異なるかを統計的に分析して、NWJC における敬語使用とレジスターとの関係を具体的に検討する。

１．はじめに

発表者は現在、国語研日本語ウェブコーパス（以下、NWJC）を使用し、謙譲表現「ご～する」の動詞別の使用頻度、そして、「ご～する」という表現が謙譲語Iであるという性質から考えたとき、その使用が「正用」であるか、それともいわゆる「誤用」であるか、その使用状況を調査・分析している。ウェブコーパスは、この種の「誤用」も含まれるデータを収集するに好適な言語資料であるが、残念ながら、NWJC にはレジスターの情報は付与されていない。

敬語の使用およびいわゆる「誤用」にはレジスターによる差があると推測される。たとえば、サービス業を含む企業や公的機関のホームページなどでは敬語も多く使用され、その性質上、誤用は少ないと予想される。一方、日記や個人同士のやり取りが多いと思われるブログなどでは、そのくだけた雰囲気から、そもそも敬語の使用自体が少なく、使用された場合には「誤用」が現れやすいと考えられる。しかし、NWJC で確認できる情報は URL のみで、こうした予想を NWJC によって直ちに確認することはできない。

しかし、動詞別の「敬語の誤用」を調査している以上、資料のレジスターを把握することは必要である。もし、レジスターによる差が明確に存在し、NWJC が偶然その特徴的なレジスターの資料を多く含んでいるだけであれば、それは調査全体の結果の信用を損なうためである。また、仮にそのような差が存在するとしても、その事実を知らずにデータを利用しているか、それともそれを理解した上で利用しているかでは結果の解釈が大きく異なる。

そこで、本発表では、NWJC から得た「ご～する」の使用例を仮設したレジスターに分類し、動詞別の使用頻度や動詞ごとの誤用率がレジスターによってどのように異なるかを統計的に分析し、NWJC における敬語使用とレジスターとの関係を検討する。具体的には、NWJC におい

て確認できる唯一の情報である URL を利用し、その「ドメイン」を分析することによって、NWJC というデータのレジスターの特徴を把握することを試みる。

２．研究の枠組み

２.１　謙譲表現「ご～する」の正用と誤用

現代の「ご～する」の規範的用法として規定されるのは、「謙譲語I」の用法のみである。「謙譲語I」用法は大前提として行為の主体は必ず自分自身、もしくは、話し手がウチの人間として捉えているもので、「客体」との何らかの関係をもつ使用でなければならない。この何らかの関係としては二つがあげられる。一つは（1）のように、その動詞がヲ格、ニ格などに客体をとる、「客体」に直接関わる行為である。

（1） ↓今月コチラを申し込みされた方を**ご招待し**ます

もう一つの関係は、前述したものとは異なり、文そのものには客体が形式として明示的には現れないが、「客体のために」行われる行為で、（2）のように、その行為の結果が客体に恩恵を与える行為である。

（2） その人がもっともお似合いになる「白」を**ご提案し**ます

蒲谷（1992）は前者の使用をI類、後者の使用をII類と説明している。以下の記述では蒲谷（1992）のI類とII類という用語を援用したい。

I類の中には、一般的に動詞そのものはI類に分類されない行為であっても、（3）と（4）のように特定の使用場面においては客体に直接関わりをもつ用例もある。

（3） 色々と**ご注文し**ましたが、納車までスムーズにご対応頂けて有難うございます

（4） 今後も**ご利用し**たいと思いますので宜しくお願いします

（3）と（4）は文には客体が形式として明示的には記されず省略されているが、共通して「（相手の）商品・サービス」をヲ格の補語によって取る表現である。これらは客体と直接、関係し、補語となる対象を高めている使用であるため、「謙譲語I」として解釈できるものである。

一方、実際使用されている「ご～する」の用例には上述のような「謙譲語I」としての用法ではなく、以下のように、相手・相手側、もしくは、謙って表現すべきでない相手がする行為について使用されている例が目立つ。

（5） 『カード決済』を**ご利用する**事はできません

(6) 中国産の低価たばこは取り扱っておりませんのでご安心してください

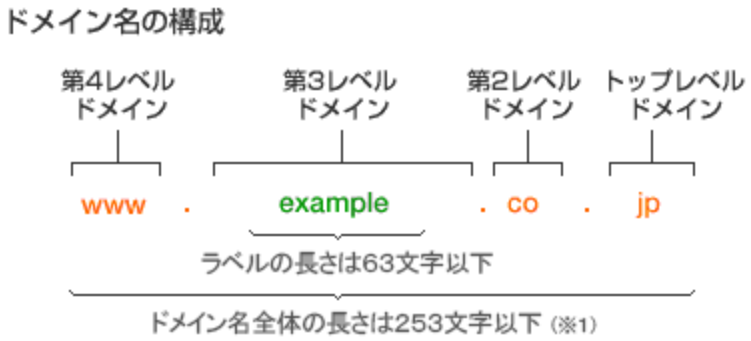
(7) ロコミで広まって 100 万人がご利用しています

これらは厳密に言うとそれぞれレベルは異なるが、全て「謙譲語I」としての使用ではないという点においては共通している。

本発表のデータでは、(1) から (4) までのような用例を「正用」として認め、(5) から (7) までのような用例は「誤用」として分類する。

2.2 ドメインとは

ドメインとは、簡略に説明するとインターネット上で使用されている住所のことである。具体的には、「www.example.co.jp」のようなものであり、「.」で区切られている各部分は「ラベル」と呼ばれる。ドメイン名を構成するラベルの中で、最も右側に位置するラベルを「トップレベルドメイン」と称し、以下左へ順に「第 2 レベルドメイン」「第 3 レベルドメイン」などのように続く。これを図式化すると、<図 1>のようになる。



<図 1>ドメイン名の構造
（出典：一般社団法人日本ネットワークインフォメーションセンター（JPNIC））

この中で、登録されているウェブページの性質を区別する重要な基準となり得るドメインはトップレベルである。ドメイン名をトップレベルドメイン（TLD）で分類すると、分類別トップレベルドメイン（gTLD：genericTLD）と国コードトップレベルドメイン（ccTLD：country code TLD）に大きく分けることができる。gTLD は TLD だけでそのドメインの用途と登録対象が確認できるが、ccTLD の場合、その用途と登録対象を確認するためには、TLD だけでなく、第 2 レベルドメインまで確認する必要がある。

そこで、本発表では、基本的には TLD を基準にドメインを分類し、ccTLD である JP ドメイン名に関しては、第 2 レベルまでを分類基準とする。各ドメインの詳細な説明に関しては紙面の都合上、割愛するが、必要に応じて随時説明を加えていきたい。なお、その場合、各ドメインの詳細の説明は株式会社レジストリサービス（JPRS）の「JP ドメイン名の種類」の説明ページ（https://jprs.jp/about/jp-dom/spec/）を参考に記述する。

2.3 コレスポンデンス分析

コレスポンデンス分析は対応分析とも呼ばれる分析の手法で、1960 年代にフランスの Jean-Paul Benzecri によって提唱され、1970 年代以降に広まった比較的新しいものである。小林（2010）では、コレスポンデンス分析について、以下のように説明している。

> コレスポンデンス分析は、データ表の行や列に含まれる情報を少数の成分（コレスポンデンス分析では次元［dimension］と呼ぶ）に圧縮し、それらの関係を散布図上に布置することで、視覚的なデータの俯瞰を可能にする。[1]

コレスポンデンス分析は、アイテムごとのカテゴリ間に潜む複雑な関係性の分析に有用な手法であり、本発表においても、ドメインと使用される各動詞と誤用率の関係性を明らかにする方法として最も有効な方法であると思われる。従って、本発表では、分析のための統計手法としてこのコレスポンデンス分析を利用する。

2.4 先行研究

田野村（2012）は現代日本語書き言葉均衡コーパス（以下、BCCWJ）に収められた特定目的サブコーパスという言語資料の特性について分析しており、同論考の中で、BCCWJ の利用者は BCCWJ が複雑な内部構成を有することと、BCCWJ が従来の日本語研究にはあまり使われてこなかった新種のデータを含んでいることに注意しなければならないことを指摘した。

本発表で分析する NWJC は従来の日本語研究においてあまり使用されてこなかった新種のデータとしての性質が強い。そのため、NWJC を利用するときには、BCCWJ 以上に資料の特性を把握することが必要になると思われる。しかし、それにも関わらず、NWJC はレジスターが分類されていないため、その複雑な内部構成を『梵天』によっては直ちに確認することはできない。そこで、本発表では「ご～する」という限られた言語表現からではあるが、NWJC から収集された結果から誤用率を求め、「ご～する」の誤用率における NWJC のレジスターとデータの特性を考察したい。

また、田野村（2012）は、BCCWJ に格納されている Yahoo!ブログのデータから、ブログという媒体が言語研究の資料としてどのような性質であるかについて分析し、その結果、ブログは一般に抱かれている「個人が毎日あるいは気の向いたときに近況やエッセイを書いてインターネット上に公開する」という印象とは異なり、現状としてはその少なからぬ部分を各種の宣伝目的の記事と外部からの引用による記事が占めていることを指摘した。

NWJC による「ご～する」の用例収集の結果からブログのデータを分類しても、その傾向は同様であり、実際、必ずしもブログだからと言って、個人が気の向くままに書いているものとは限らないことが確認できた。自分でホームページを作成する技術を持たない個人経営の小規模店舗が、広報の必要性とホームページ作成を依頼する経費を勘案し、手軽に作成・管理できるブログをホームページ代わりに活用している例が多い印象であり、本来のブログの用途と言

[1] 小林雄一郎（2010）245 ページより抜粋

われる、完全な個人の近況報告などはむしろ少ないように思われた。また、実際にはもう一つの用途として、既にホームページを持っている企業や団体であっても客との心的距離を縮める、もしくは、その内容更新の手軽さから、簡単な情報発信のためにブログを活用することもあり、単純に「ブログ」というだけでは、その性質は定義できない。本考察では、その点も踏まえ、NWJC のデータの性質について総合的に考えていきたい。

３．調査概要

３.１　調査目的

本発表は資料としての NWJC の性質を把握することをその目的とする。具体的には、「ご～する」という特定の敬語表現の調査結果を基に、NWJC の中のデータにはどのようなレジスターの特徴が見られ、各ドメインをどのように分類することができるかについて考察する。

３.２　調査資料

本発表では国語研日本語ウェブコーパス(以下、NWJC)を調査資料とする。NWJC は国立国語研究所が 2017 年に公開したウェブを母集団とする 100 億語規模のコーパスで、2014 年 10 月から 12 月までの間に収集したデータを格納している。

３.３　調査対象

本来「ご～する」など、「ご」の付く敬語形式は、一般に「お＋和語動詞語幹＋する」など、「お」の付く形式も含まれるが、「お＋和語動詞語幹」については「お使い」のように、「お＋和語動詞語幹」の組み合わせで一語として定着しているものや、「お掃除する」のように、美化と謙譲の区別が曖昧になっているものも含まれるため、「ご～する」の誤用とはその性質が異なる可能性が考えられる。従って、本発表では、「お～する」は対象外とし、「ご＋漢語サ変動詞語幹＋する」のみを分析の対象としている。

その中でも、NWJC において「ご～する」形式を多く取っている動詞を頻度数順に上位 30 語まで選定した。その 30 語を以下に示す。動詞の前の数字は NWJC における「ご～する」形式での頻度数の順位を表しているものである。

「01．紹介」「02．用意」「03．案内」「04．提供」「05．報告」「06．説明」
「07．挨拶」「08．提案」「09．連絡」「10．利用」「11．相談」「12．購入」
「13．招待」「14．来店」「15．披露」「16．安心」「17．協力」「18．奉仕」
「19．対面」「20．確認」「21．使用」「22．参加」「23．注文」「24．訪問」
「25．指導」「26．理解」「27．注意」「28．心配」「29．満足」「30．予約」

次に、30 語の動詞をそれぞれ「ご～する」形式に入れて NWJC から得られた用例をダウンロードし、その中からランダムで各 400 例ずつを選定した。そして、用例の文脈を見ながら、そのデータを正用と誤用に分類した。その過程の中で、30 語の内、使用の傾向が偏っている、コーパスから得られる結果のみでは前後の文脈や人物同士の関係の把握ができないなどの理由によって分析対象として不適切であると判断した「18．奉仕」「19．対面」の 2 語は分析対象

から除外し、最終的には総計 28 語が対象となった。そして、分類の結果から、各動詞の「ご～する」の誤用率を求め、その結果を今回の分析の対象とする。

３.４　調査方法

まず、調査対象をまとめ、各動詞の正用と誤用の実数をドメイン別に再集計した上で、各動詞と各ドメインの誤用率のクロス表を作成する。そして、その結果を基に統計解析ソフトである IBM SPSS Statistics によりコレスポンデンス分析を行い、その結果について考察する。

ただし、クロス表内に欠損値が多くては正確な分析の妨げになるため、次のような補正処理を行った。第一に、補正として欠損地が多い動詞を削除した。その基準としては、他のドメインと明白に性質が異なるドメインに現れる動詞を残すことにした。特に、公的機関のドメインであり、「公的なテクスト」としての性質が強いという性質から一番規範意識が高いと思われるドメイン go.jp の結果を最大限残すべく、go.jp で使用されていない動詞を削除した。その結果、「01．紹介」「02．用意」「03．案内」「05．報告」「06．説明」「08．提案」「09．連絡」「10．利用」「11．相談」「15．披露」「17．協力」「20．確認」「25．指導」「26．理解」の 14 語が残った。そこから、更に教育機関という明確な性質をもつドメイン ac.jp において 1 例も使用されていない「05．報告」「08．提案」「15．披露」「20．確認」の 4 語を削除した。残りの 10 語の内、「01．紹介」「02．用意」「03．案内」の 3 語においてはほとんど全てのドメインが誤用率 0%、「10．利用」「26．理解」においてはほとんど全てのドメインが誤用率 100%という偏った傾向をみせていた。コレスポンデンス分析では、このようにドメインによって動詞の誤用率に差がみられないものを分析の対象として含めてしまうと、動詞とドメインの誤用率の関係を説明する際に信頼性が大きく損なわれる。従って、上記の 5 語を削除して、ドメインによって誤用率の傾向が異なっている 5 語、「06．説明」「09．連絡」「11．相談」「17．協力」「25．指導」のみを分析対象として残した。

そして、ドメインの中でも、性質が全く同じドメインであるため、まとめずにそのまま分析する意味のないものは一つのカテゴリにまとめた。たとえば、以下の＜表 1＞において、「local」という項目は、.local という純粋なドメインを表しているわけではなく、tokyo.jp、osaka.jp のような地域型 JP ドメイン名をまとめたものである。地域型 JP ドメイン名は詳細な地域によって第 2 レベルドメインこそ異なるが、その性質は全て同様であるため、本研究においてはまとめても特に問題のないものである。地域型 JP ドメイン名は地域を表すドメイン名として、地方公共団体・特別区およびその機関、他の属性型 JP ドメイン名の登録資格を満たす組織、日本に在住する個人、病院が登録できるドメイン[2]である。

また、＜表 1＞には「他の国」という項目も存在するが、これも理屈としては「local」と同様である。まとめる前のデータには ac, at, bz, ca, cc, fm, in, is, md, me, ms, nu, sc, sg, th, to, tv, us, vc, ws のようなドメインが収集されたが、これらのドメインは、全て ccTLD で、日本の jp ドメイン名のようなものである。詳細なドメイン名は異なっても、基本的にはその中に有意な差は

[2] 株式会社レジストリサービス（JPRS）「JP ドメイン名の種類」の説明ページ（https://jprs.jp/about/jp-dom/spec/）を修正・加筆している。（2018 年 07 月 13 日最終確認）

存在しないと考えられるため、これらも「他の国」として一つの項目にまとめている。この「他の国」項目に含まれる ccTLD は基本的には登録資格に特に制限はなく、誰でも自由に登録できるドメインである。

ほかのドメインに関しては、「local」と「他の国」ほどドメイン間の性質が完全に一致するものは見られず、それぞれ程度の差はあれど、異なる特徴を有するため、ドメインそのものを特定基準によって任意にまとめたものはこの二つだけである。

最後に、調査対象の動詞 5 語のうち、2 つ以下の動詞でしか使用されていないドメインを削除した。この過程で 1 語の動詞においてのみ使用されている ed.jp、travel、2 語の動詞においてのみ使用されている lg.jp が除外され、最終的にドメインは ac.jp、biz、co.jp、com、go.jp、info、jp、「local」、ne.jp、net、or.jp、org、「他の国」の 13 項目が分析対象として含まれた。

４．調査結果

まず、分類の結果を以下＜表 1＞に示す。＜表 1＞では「local」と org において「指導」が空欄になっているが、これは総計 1 例も使用されなかったものである。

＜表 1＞動詞・ドメインの使用頻度・誤用・正用・誤用率のクロス表

		ac.jp	biz	co.jp	com	go.jp	info	jp	local	ne.jp	net	or.jp	org	他の国	総計
06.説明	総数	4	11	41	150	19	4	90	8	9	50	3	4	5	398
	誤用	0	0	1	6	2	0	8	0	1	0	0	0	0	18
	正用	4	11	40	144	17	4	82	8	8	50	3	4	5	380
	誤用率%	0	0	2	4	11	0	9	0	11	0	0	0	0	5
09.連絡	総数	3	1	38	210	3	7	81	1	15	29	4	3	2	397
	誤用	1	0	7	16	0	1	11	1	2	5	0	1	1	46
	正用	2	1	31	194	3	6	70	0	13	24	4	2	1	351
	誤用率%	33	0	18	8	0.0	14	14	100	13	17	0	33	50	12
11.相談	総数	1	1	31	160	2	9	121	1	35	27	9	1	2	400
	誤用	1	1	12	73	0	3	55	1	10	12	4	0	1	173
	正用	0	0	19	87	2	6	66	0	25	15	5	1	1	227
	誤用率%	100	100	39	46	0	33	46	100	29	44	44	0	50	43
17.協力	総数	4	6	10	222	2	9	89	2	5	43	5	1	2	400
	誤用	4	5	6	196	0	6	78	1	5	32	5	1	2	341
	正用	0	1	4	26	2	3	11	1	0	11	0	0	0	59
	誤用率%	100	83	60	88	0	67	88	50	100	74	100	100	100	85
25.指導	総数	5	7	28	157	1	23	99		27	38	9		3	397
	誤用	5	6	11	105	1	15	49		11	14	4		1	222
	正用	0	1	17	52	0	8	50		16	24	5		2	175
	誤用率%	100	86	40	67	100	65	50		41	37	44		33	56
総計		30	60	362	1775	37	83	925	21	182	388	61	14	33	3971

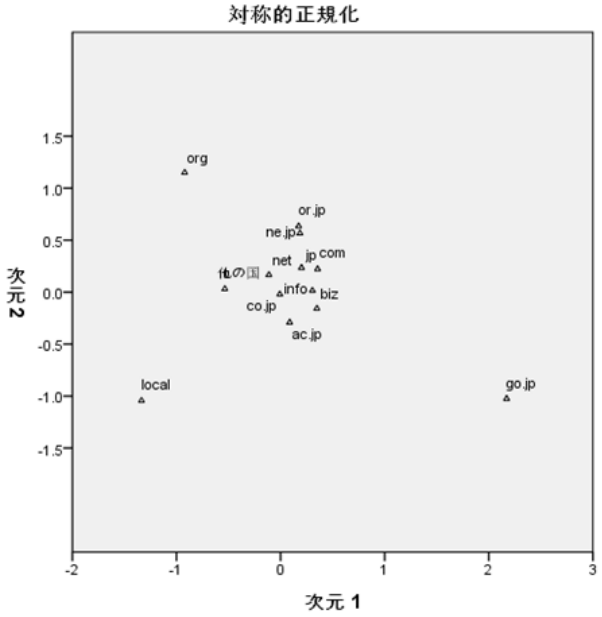
次に、＜表 1＞の結果により求めたコレスポンデンス分析の結果の要約は以下＜表 2＞の通りである。

＜表 2＞コレスポンデンス分析結果の要約

次元	特異値	要約イナーシャ	カイ 2 乗	有意確率	イナーシャの寄与率		信頼特異値	
					説明	累積	標準偏差	相関
								2
1	.473	.224			.586	.586	.013	.214
2	.290	.084			.220	.805	.018	
3	.241	.058			.152	.957		
4	.127	.016			.043	1.000		
要約合計		.382	1008.266	.000[a]	1.000	1.000		

a. 自由度 48

コレスポンデンス分析の結果、4 の次元が抽出され、次元 2 までの累積寄与率は 80.5%であった。以下では、この結果の散布図を基に分析を進める。



＜図 2＞ドメインの行ポイント散布図

＜図 2＞の結果をみると、10 個のドメインが原点に近い中央の方に集中しており、残りの 3 個のドメインは中央に位置するドメインを囲み、次元 1 を底辺とする直角三角形を形作るよう

に位置していることが確認できる。具体的には、次元 1 の正方向の特徴は go.jp によって、負方向の特徴は「local」と org によって特徴づけられ、次元 2 の正方向の特徴は org によって、負方向の特徴は「local」と go.jp によってそれぞれ特徴づけられている。

中央のグループには商業用ドメインである co.jp と biz や、教育機関である ac.jp などのような特徴的なものをはじめ、特徴を限定しにくい com、net などのようなドメインまで、多様なドメインが含まれている。これら、中央のグループは、これまでのように TLD と第 2 レベルドメインだけをみると共通する特徴がまとまらず、解釈が難しいが、一つの共通点がある。それは、このグループに含まれる結果は全体的な URL をみると、or.jp を除いては、全てブログという特殊なものが含まれているということである。具体的には、今回分析の対象となった用例総計 1992 件の中で、URL だけでブログであることが判明するものが 938 件と、ほぼ半数近くを占めており、ac.jp、biz、co.jp、com、info、jp、「local」、ne.jp、net、「他の国」においてブログでの使用が見られた。しかし、上記に見られるように、中央ではないところに布置されている「local」において 1 例、また、URL そのものからは判定しないため、上記の数字には含まれていないが、分析の中で元のリンクに辿って確認した結果、org にもブログでの使用が 1 例確認された。一方、中央に布置されている or.jp においては、ブログの使用は見られない。以上の結果をまとめると、NWJC において収集されるデータの中には、全体的にブログの用例が多く、このブログの用例が何か影響を与えている可能性もあると思われるが、中には例外も存在し、それが具体的にどのように影響しているかまでは今回のデータだけでは解釈できない。

また、ブログとは言え、その用途と性質が一様ではない点、org と「local」においても 1 例ずつではあるがブログの用例が含まれている点などを考えると、単純にブログが特徴的な性質を示すと解釈するのは、多少説得力が弱いように思われる。全体の用例数の割合から考えても、ほとんど全ての用例が含まれている中央グループのドメインがこの原点に近いところに布置されているということに鑑みると、NWJC において収集されるデータは、「ご～する」の誤用においては特徴的な偏りは少ない可能性が高い。

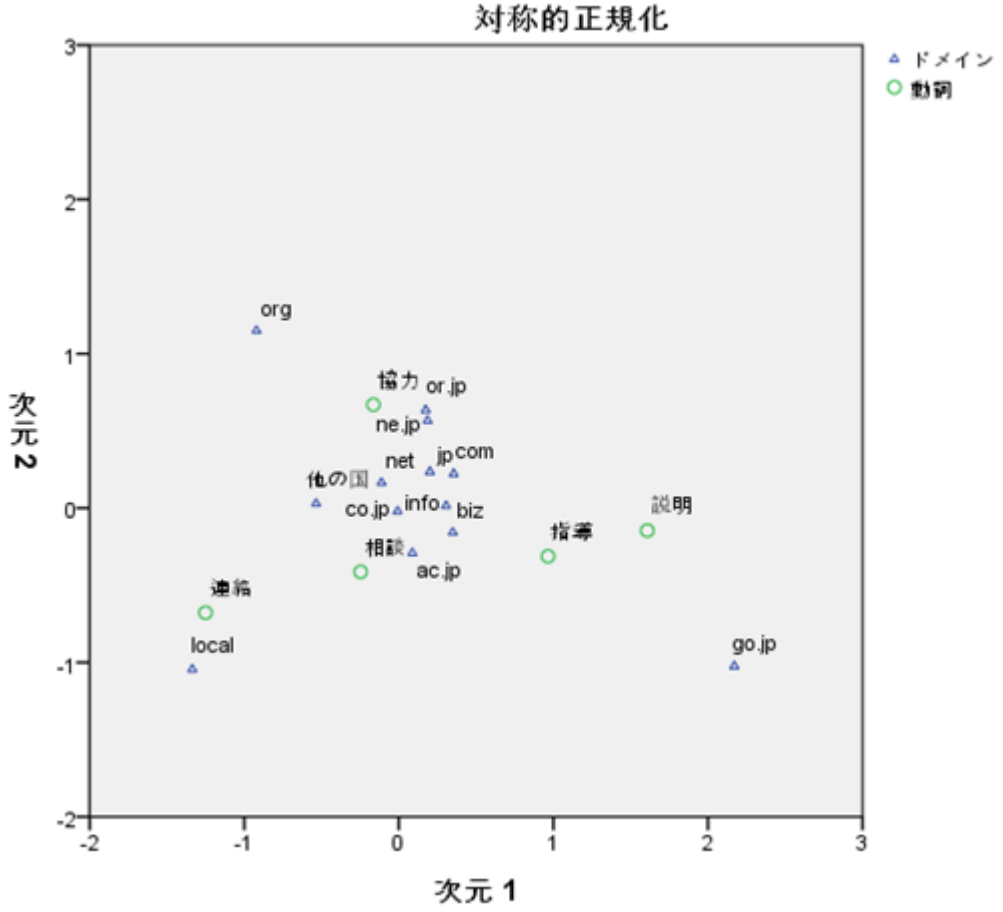
一方、特徴的に現れた三つのドメインは以下のような性質をもつドメインである。以下の記述において、go.jp と「local」のドメインは株式会社レジストリサービス（JPRS）の「JP ドメイン名の種類」のページの説明を参考にしている。1 つ目の go.jp は日本の政府機関や各省庁所管の研究所、特殊法人、独立行政法人が登録できるドメインであり、中でも政府機関は、一つの組織で複数の go.jp ドメイン名を登録できる。2 つ目の「local」（地域型ドメイン）は地域を表すドメイン名として、地方公共団体・特別区およびその機関、他の属性型 JP ドメイン名の登録資格を満たす組織、日本に在住する個人、病院が登録できるドメインである。最後の org は用途としては非営利組織用のドメインであるが、その登録条件には制限がないものである。

まず、次元 1 として、正方向に特徴づけられた go.jp は、公的機関としての性質が最も強いドメインであり、一般的に考えると最も誤用が現れにくい、強い規範意識が働いているドメインである。一方、負方向に特徴づけられた「local」は、go.jpのように政府機関である「特別区およびその機関」のページも含まれるが、go.jp とは異なり、ほかの属性型 JP ドメイン名の登録資格を満たす組織か日本に在住する個人でも登録できるドメインであるため、公的機関としての性質と、その他の性質が混在し　ているドメインである。また、org は非営利組織で、通常の用語の定義から考えると、非営利組織の中に政府組織は含まない。このことから、次元 1

では純粋な公的機関としての性質が強いドメインであるか、それとも、公的機関を全く含まない、もしくは、他の性質が混在し、純粋な公的機関でないドメインであるかが特徴として現れているように思われる。一方、次元 1 においては、org と「local」は近いところに布置されているが、非営利組織の中に一部公共的な活動を行う団体が含まれるため、両者は公共的な団体としての性質が共通している。

次に、次元 2 では org と go.jp、「local」の間に特徴的な差が見られる。次元 2 において、「local」と go.jp に共通する性質で、org とは異なる最も大きな性質は、公的機関、中でも政府機関のページが含まれているか否かであると解釈される。

つまり、次元 1 においても、次元 2 においても、共通して公的機関のもつ性質が大きく作用していると推測される。



＜図 3＞ドメインの行ポイント・動詞の列ポイントの散布図

＜図 3＞は行ポイントである「ドメイン」と、列ポイントである「動詞」の結果を合わせた散布図である。レジスターを分析するためには、本来であれば影響を与え得る要因を全て総合的にみていかなければならない。しかし、今回は「ご～する」形式での動詞の誤用の性質とい

う極限られた現象に注目して分析を行っている上に、13 のドメインに対し、その特徴を解釈するための動詞の数は 5 語だけになっている。従って、本来のコレスポンデンス分析のように最も近くに特徴づけられているものだけを考慮して分析してしまうと、その傾向性が解釈しにくいと想定される。そこで、本来のコレスポンデンス分析の手法とは若干異なるが、動詞に関しては単純に近いところに布置されているものを基準に分析するのではなく、各次元において正方向で現れているか、負方向で現れているかを基準にし、＜図 2＞のドメインだけの結果において特徴的に現れたドメインと関連付けて解釈していきたい。

動詞に関しては、次元 1 では「指導」「説明」が正方向に布置され、「相談」「協力」「連絡」が負方向に布置されている。しかし、この中で「協力」と「相談」は、負方向に布置されてはいるが、原点に近いものであるため、次元 1 においてはそこまでの特徴は示されていない可能性が高いと予想される。一方、次元 2 に関しては、「協力」だけが正方向に布置され、「連絡」「相談」「指導」「説明」は負方向に布置されている。ここでも「説明」と「指導」は 0 に近いところに布置されているため、この 2 動詞に関しては、次元 2 においては有意味な結論は得られない可能性が高い。

ドメインと動詞を合わせて考えると、まず、次元 1 からは、go.jp が正方向に特徴づけられ、同じく正方向に布置されている動詞は「説明」と「指導」の 2 語であった。一方、「local」と org は負方向に布置されており、動詞の中では「協力」「相談」「連絡」が同じく負方向に布置されていた。＜表 1＞のそれぞれの「誤用率」からも確認できるよう、go.jp においては、第一次元において同じく正方向に布置された動詞である「説明」と「指導」においてのみ誤用が現れ、負方向に布置された「協力」「相談」「連絡」においては、1 例の誤用も現れていない。これに対し、負方向に布置された「local」と org は正方向に布置された動詞「説明」「指導」に関しては 1 例の誤用も現れておらず、同じく負方向に布置された動詞「協力」と「連絡」において誤用が現れた。

もう一つの負方向の動詞である「相談」は org に関しては誤用が現れておらず、「local」でのみ誤用が現れる。このように対照的な結果を示す org と go.jp であるが、＜表 1＞の誤用率に照らし合わせて考えると、「相談」に関しては、誤用率が 0%であるという点において両者に共通点が見られ、org とドメインとしては共通する性質を含めていた「local」に関してのみ差がみられる。動詞の解釈で既に述べている通り、＜図 3＞において、「相談」は原点に近い動詞であるため、今回の結果においては、特徴が弱く、解釈が難しい。

「協力」も「相談」同様、第 1 次元においては原点付近に布置されているため、次元 1 に関してはあまり有意味な特徴は見られないと予想される。しかし、次元 2 に関しては、正方向において org と「協力」の間に関連性が見られ、 政府機関を含まない非営利組織で用い誤用が現れやすいという「協力」の誤用率の特徴が説明できると解釈される。

以上をまとめると、次元 1 において、純粋な公的機関のドメインにおいて「説明」と「指導」は誤用が現れやすく、「連絡」は誤用されにくいと思われる。一方、次元 2 においては、「協力」は政府機関を含まない非営利組織において誤用されやすいと解釈される。以下では、これらの解釈について、実例をもって検証する。

まず、動詞の中で最も正方向に布置されている「説明」は go.jp において総計 19 例確認され、その中で 2 例が誤用、17 例が正用で、誤用率は 10.5%であった。以下（8）と（9）は go.jp におけるその誤用例であり、（10）から（12）まではその正用例である。

（8） その上で、この後ちょっと**御説明し**ていただくほうで議論したほうがいいかもしれませんけれども、（後略）（go.jp）

（9） 今、実態につきましては課長から**ご説明し**たとおりでございます（go.jp）

（10） あなたは、こういう事情を私が**御説明し**ても、そうではないというふうに言い張られますか（go.jp）

（11） ○＝＝年金課長　400 万人は幾つかの統計を組み合わせて推計したもので、ちょっと何もなしで説明するのは大変ですので、次回以降で**御説明し**たいと思います（go.jp）

（12） 文章のほうで引き続き説明させていただきますが、ここでご説明させていただく試験につきましては繁殖能に関する指標として産卵率、孵化率、育成率、それとこれら 3 つの指標をかけ合わせた総合繁殖指数というもので**ご説明し**たいと思います〔ママ〕（go.jp）

今回収集された用例の中で go.jp において「説明」が使用されている場面は、全て何かの会議の「議事録」であった。一方、org と「local」に現れている「説明」の正用には、以下のようなものがあった。

（13） 上下水道料金の支払方法について**ご説明し**ています(「local」)

（14） お客様と直に接して**ご説明し**、表情やご反応を見、時にはご質問やご意見をいただけるギャラリー･トークは、私たちにとっても貴重な場なのです(「local」)

（15） ・採用後、1 ヶ月間を研修期間とし、教室の理念や指導について**ご説明し**ます(org)

（16） ※弊社からのご案内メールが正しく届かなかった時はお電話にて**ご連絡し**ます（org）

以上のように、org と「local」の「説明」の用例は、一般向けに発信される公的な文書に使用されているが、議事録のものは見られない。会議の中では、自分側から相手に何かを説明することも、逆に、相手から説明されることも多い。また、「会議」という公的な場面では、敬語を使用しなければならないという意識が強く働き、相手の「説明」という行為に関しても無理に「ご～する」という形式を使用しているため、誤用が現れているように思われる。公的機関の中でも政府機関では、「会議」の「議事録」が一般に公開されることが多い。コレスポン

デンス分析の結果だけでは「純粋な公的機関であるか否か」が次元 1 において大きく影響を与えると解釈したが、さらに具体的には、「純粋な公的機関」の中でも、特に go.jp に含まれている「会議」という場面の、相手が存在する話し言葉としての性質が「説明」において特徴的に現れているように思われる。

次に、正方向に布置された「指導」に関して分析する。「指導」は go.jp において 1 例確認され、その 1 例が誤用で、誤用率 100%となった。その用例を以下（17）に示す。

（17） また、流された人を助けるための「スローバッグ」の投げ方や受取り方などの方法について**ご指導し**ていただきました（go.jp）

「指導」は go.jp、org、「local」の 3 ドメインを合わせても 1 例しか使用されておらず、（17）の用例がその唯一の用例であった。このことから、go.jp、org、「local」のような公的機関や公共的団体のドメインは、他の性質をもつドメインに比べて相対的に、何かを「指導する」という行為は想定されにくいように思われる。（17）の用例は go.jp の「徳島河川国道事務所」で使用された用例で、このページは各種講座などについての開催結果を報告しているものである。基本的には直接他人に何かを「指導」するという行為は想定されにくく、指導するとしても、機関の内部の人が講師を務めるのではなく、外部から講師を招聘することになる可能性が高いことが公的機関の特徴の可能性の一つとして考えられる。

次に、次元 1 において最も負方向に布置されている動詞「連絡」と「local」と org、go.jp の間の関係について考察する。まず、go.jp では、「連絡」は 3 例使用されているが、その 3 例は全て正用であり、誤用率は 0%となっていた。正用の用例は以下、（18）から（20）までのようなものであり、このことから、公的機関は自分側から相手に連絡をするだけで、相手からの「連絡」を求めることは基本的にはないと解釈される。

（18） 審査の結果は 2 月 18 日（水曜）を目処に**ご連絡し**ます（go.jp）

（19） また、募集締切後に受講の可否を**ご連絡し**ます（go.jp）

（20） 参加の可否については、返信はがき又は FAX で**ご連絡し**ます（go.jp）

しかし、実際の用例をみてみると、「local」で現れている「誤用」も公的機関である政府機関の使用であったため、＜図 3＞の結果から解釈したように、特徴的に現れている要因が単純に「純粋な公的機関であるか否か」と解釈するだけでは不十分であることが確認される。「local」では「連絡」が 1 例だけ確認され、その 1 例が誤用であったため、誤用率 100%となっている。その用例は以下の（21）である。

（21） 利用日の 7 日前までに**ご連絡し**て下さい　（「local」）

（21）は小平市のホームページで、他市の保養施設の利用方法を説明しているページにおいて使用されている用例である。同じ政府機関であっても、国全体をまとめる go.jp とは異なり、都道府県と市町村と、比較的狭い範囲に限られるものであるため、基本的にはその地域に居住している居住者のために発信する情報であり、go.jp に比べ、読み手との関係が近いと思われる。このことから、次元 1 は散布図で解釈したように、単純に「純粋な公的機関であるか否か」だけで区別されているわけではなく、国全体をまとめる公的機関全般での使用であるか、それとも、読み手との関係がより近い、狭い範囲での使用であるかによっても特徴づけられる可能性があると思われる。しかし、「local」に現れている「連絡」の用例が正用・誤用に関係なく（21）の 1 例しかなかったため、今後分析データが増えれば、結果が変わる可能性も高い。

最後に、次元 2 において、「協力」と org の間の誤用率の関係性について考察する。「協力」は org において 1 例使用されており、その 1 例が誤用であったため、誤用率 100%となった。org における「協力」の誤用の用例は以下のようなものであった。

（22） 治験が治験実施計画書（治験を安全に行うにあたって、治験を依頼した製薬会社が決めたルール）を遵守(じゅんしゅ)しているかを細かくチェックしながら、治験にご協力していただいている患者さまからの相談（心のケア）や質問に対する対応を行っています (org)

（22）は 2018 年 7 月現在はリンクが切れており、全文の確認はできないが、URL からして病院の案内文の一部であることが確認できる。

これに対し、次元 1 において「org」と同様に負方向に布置されていた「local」では、「協力」は 2 例使用されており、その中の 1 例が誤用として現れ、誤用率 50%となっていた。「local」の「協力」の誤用を（23）に、正用を（24）に示す。

（23） ご来場いただきました皆さんには、アンケート調査にご協力していただきまして、ありがとうございました（「local」）

（24） （前略）あるいは県営名古屋空港がそれに御協力する機会があるのか、そんなことは当然これからしっかりと検討していかなければなりませんし、（後略）（「local」）

（23）は「ちよだ区議会だより」から抽出された用例で、区民大会の結果を報告したものである。

一方、go.jp では、「協力」は 2 例使用され、その 2 例とも正用の用例であった。具体的には、次のような用例が収集された。

（25） 自主共聴組合さまからのご要望により、以下についてご協力します（go.jp）

（26） したがって、当面の施策としては、先般商工委員会で十一時半ぐらいまでやっていただきまして上げていただいたあの法律の適用ということで私ども大蔵省としての立場か

らも精いっぱい御協力することがあればしていくというのが今日の対応策でございます（go.jp）

go.jp においては、誤用は現れなかったが、「local」において誤用として現れている（23）の例が政府機関の使用であるため、公的機関としての性質が「協力」の誤用率に影響しているとは解釈できない。しかし、そもそも次元 2 の寄与率は次元 1 の寄与率の半分にも満たないことが＜表 2＞より確認できる。このことを総合的に考えると、次元 2 において現れた特徴は、今回の用例の実数の少なさによるものであり、実際のところは特徴として説明できない可能性が高いと考えられる。

５．おわりに

本発表ではドメインをレジスターとして仮設し、「ご～する」のドメイン別・動詞別の誤用率についてコレスポンデンス分析を行うことにより、ドメインを NWJC のデータのレジスター類似のものとして利用する可能性を試みた。その結果、今回のデータに限って言うと、NWJC により収集されるドメインにおいて、ドメインと動詞との関係の中で特徴的に現れているものは極一部だけであることが確認された。しかも、これら、特徴の見られないドメインが NWJC により収集されるドメインのほとんど全てを占めており、NWJC はさほど偏りのないデータのように解釈される。「ご～する」の誤用において一部特徴的に現れているものは公的機関での使用のように、全体の日本語の中では極めて限られた部分であるように思われる。つまり、NWJC に含まれているドメイン同士の特徴にそこまで深刻な差は見られず、NWJC を利用して敬語を分析するにあたり、特定ドメインの集中による結果の偏りはあまり現れない可能性が高いように思われる。

今回の結果からは、「ご～する」の誤用においては、go.jp、org、「local」の 3 ドメインにおいて最も顕著な特徴が見られた。そして、その特徴は純粋な公的機関のドメインであるか否か、さらに、純粋な公的機関のドメインであっても、国全体をまとめる公的機関全般での使用であるか、それとも、読み手との関係がより近い、狭い範囲の公的機関での使用であるか、また、議事録が含まれたデータであるか否かによって説明される可能性があることが明らかになった。

しかし、今回の分析に関しては、以下のようないくつかの問題点があることを指摘したい。まず、そもそも対象としている動詞の数が少ないため、ドメイン別の性質を明確に説明しにくいことが問題として挙げられる。次に、各動詞の用例数も 400 と限られているため、そこからさらにドメイン別に分類をすると、ドメイン別の用例数が少なくなりすぎるという点である。そのため、1 例の変動だけでも誤用率が大きく変わり、結果が変動してしまうため、それが統計の結果に大きな影響を与えている。最後に、「local」のように、複数の性質が含まれているドメインの存在と、ブログのように性質の定義が曖昧なものが各ドメインに含まれている点が結果の解釈に悪影響を及ぼしている可能性も存在する。

このことを踏まえて今回の結果を考えると、NWJC において「ご～する」の動詞別誤用率に関する研究にドメインによる結果の偏りの問題はそこまで大きな影響を与えてはいないように思われるが、go.jp のように、一部明確な性質を持つものについては、他とは区別される明確

な特徴がみられる。しかし、今回の調査は、用例数の少なさ、一部ドメインの性質の定義の難しさなどが統計結果に影響を与え、事実を正確に反映していない可能性も考えられる。これについては、データの数を増やし、更にドメインをなるべくカテゴリ化した上で再度分析を行いたい。これを今後の課題とする。

文　献

蒲谷　宏（1992）「『お・ご～する』に関する一考察」『辻村敏樹古希記念　日本語史の諸問題』明治書院，pp.141-157.

金　賢眞（2015）「謙譲表現『ご～する』の誤用―公的テクストにおける実態とその要因―」大阪大学文学研究科修士論文（未公刊）

国立国語研究所コーパス開発センター編（2017）『国語研日本語ウェブコーパス』（2014-4Q データ，梵天バージョン 1.0.0）https://bonten.ninjal.ac.jp/（2018 年 07 月 13 日最終確認）

小林雄一郎（2010）「第 10 章　コレスポンデンス分析；データ間の構造を整理する」石川慎一郎・前田忠彦・山崎誠編『言語研究のための統計学入門』くろしお出版，pp.245-264.

田野村忠温（2012）「BCCWJ に収められた新種の言語資料の特性について :データ重複の諸相とコーパス使用上の注意点」『待兼山論叢. 文化動態論篇』46，pp.59-83.

関連 URL

一般社団法人日本ネットワークインフォメーションセンター（JPNIC）（https://www.nic.ad.jp/ja/）（2018 年 07 月 13 日最終確認）

株式会社レジストリサービス（JPRS）「JP ドメイン名の種類」の説明ページ（https://jprs.jp/about/jp-dom/spec/）（2018 年 07 月 13 日最終確認）

『国語研日本語ウェブコーパス』検索系『梵天』　http://bonten.ninjal.ac.jp/（2018 年 07 月 13 日最終確認）

学校お便り文の高頻出語彙の縦断的研究
―4 年生から 6 年生までの名詞・サ変名詞・動詞の分析―

今村桜子（横浜国立大学大学院）

Longitudinal Study of High Frequency Words in the School Letters
An Analysis of Nouns and Verbs in the Elementary School News Letters from 4th to 6th grades

Eiko Imamura (Yokohama National University)

要旨

首都圏の公立小学校のお便り文(3 年分 712 部)からコーパスを作成し，学年ごと（4 年生から 6 年生）の語彙の違いを分析した。本研究は，学校お便り文に用いられる語を縦断的に観察することで，外国人保護者の日本語支援に役立てることを目的とする。

KH Coder で形態素解析を行ったところ，総語数は 4 年 56,968 語，5 年 106,084 語，6 年 77,167 語。異なり語数は 4 年 7,420 語，5 年 9,935 語，6 年 9,395 語であった。

①頻度グラフにより少数の高頻出語と多数の低頻出語が観察される。高頻出語の学習が次年度以降の読取りに効果的であると考えられる。②品詞ごとの高頻出語を抽出し，4 年生の上位 100 語が 5、6 年生の上位 100 語に含まれる割合を分析した結果，名詞・サ変名詞・動詞で 74%から 83%に上ることが分かった。③サ変名詞「卒業」は、4 年生で 32 回，5 年生で 66 回，6 年生で 104 回（20 位）出現する。6 年生に多いが，高頻出語の学習が他学年の保護者にとっても，学校文化理解や内容スキーマ活性に役立つと示唆される。

１．はじめに

「生活のための日本語：全国調査」によると，生活する外国人に質問した言語行動のうち，「学校や園からの配布物や連絡ノートを読み，必要に応じて準備する」が「日本語でできない」割合は 48.2%である（国立国語研究所 2009）。母親が外国人である場合は，お便り文を同居する夫や夫の両親，子供自身に読んでもらっているケースが多い（富谷他 2011）が，自力で処理対応できないことから，家庭内の発言権が弱かったり，自尊感情が持てなかったりするケースが報告されている(伊藤 2007)。

桑原（2017）は，生活する外国人が生活場面で手にする文書を読むために，ポイントとなる漢字・語彙・表現を精選して提示すべきだと提言している。本研究は，お便り文に用いられる語を縦断的に観察することで，優先的に学ぶべき語彙を抽出し，外国人保護者[1]のお便り文書の読解支援に役立てることを目的とする。

２．先行研究

地引（2013）は，小学校配布物 293 件を形態素解析し，解析した語が旧日本語能力試験の出題基準語彙のどの級の該当語彙であるかレベル判定し，その割合を調査した。更に

[1] 本稿での「外国人保護者」とは，日本で子どもを学校に通わせる保護者であり，日本語非母語話者の方を指す。

「名詞」から頻出語彙 150 語を抽出し，それらが小学校配布物から情報を得るために必要かどうかを，外国人保護者，日本人保護者，教師の 3 者に問う意識調査を行った。その結果，「情報を得るために必要な語彙」として 32 語を抽出した。

李(2017）は，福岡など 4 市から収集した文書を分析し「学校おたよりコーパス」を構築し，総文字数 880,869 のデータを抽出した。また，中国籍の保護者からの「中国と同じ漢字を使っていても，組み合わせによって(意味が)わからなくなる」との意見から，必要度の高い複合名詞を抽出している。次に，抽出した複合名詞の理解度を調査し，「外国に存在しないか，類似したモノ・コト」「母国とは別のモノ・コトを指す場合」「漢字の多義性から生じる誤解」により，お便りの理解が困難になると分析している。さらに，語彙のみを教えるのではなく，学校文化を伝えることの重要性に言及している。

３．研究課題

本稿の課題を以下の３つとした。

①お便り文にはどのような語が用いられるか。学年ごとの総語数と異なり語数を調査し，語彙表と，頻度グラフを作成する。

②どの学年にも用いられる語は何か。品詞ごとの高頻出語を縦断的に観察・分析し，網羅率を調査する。3 学年で共通して用いられる「最頻出語」を抽出する。

③ある学年に特有の語彙[2]が他学年でどのように出現するか。

４．小学校で配布されたお便り文に用いられている語彙

４．１　研究方法

平成 25 年度から 28 年度に首都圏の公立小学校で一人の児童に配布されたお便り文を 1121 部収集した。年度途中からの収集であった平成 25 年度分を除き，1 年分の文書が揃っている平成 26 年度から 28 年度の 3 年度分のお便りを対象とし，このうち学校と PTA から使用許諾を得た文書，合計 712 部を分析対象とした。紙のお便り文をＯＣＲソフト「本格読取４」でデータ化し，KH Coder[3]を利用して形態素解析[4]を実施し，コーパス[5]を構築した。

また、お便り文配布の目的は，学校から家庭へ連絡事項を伝達することであり，受け取った保護者は，その内容に基づいて参加不参加の意思表示をしたり，持ち物の準備をしたりするなどの適切な行動をとることが求められる。つまり，行事名や持ち物名称の意味が理解できるだけではなく，自らのすべき対応を読み取る必要がある。そのためには「提出する」などの「サ変名詞＋する」や「使う」「書く」などの動詞の習得が必要であるとの考えから，本研究では名詞のみでなく，サ変名詞と動詞も分析対象とした。

得られたデータから，品詞別に出現回数を記入した抽出語リストを出力し，それを基に，名詞，サ変名詞，動詞，動詞 B の頻度表を作成した。更に，頻度 1 位から各語までの累積語数が総語数に占める割合を計算した。

４．２　結果と考察

学年毎の語彙数を表 1 に示す。総語数は 4 年生 56,968，5 年生 106,084，6 年生 77,167。異なり語数は 4 年生 7,420，5 年生 9,935，6 年生 9,395 であった。５年生のお便り部数が最も多いためか，品詞ごとの総語数，異なり語数共に５年生の語彙数が多い(表２)。品詞別の総

[2] ある学年に特有の語彙とは，その学年のみが体験するコト・モノの名称のこととする。

[3] KH Coder2.0。樋口耕一氏（立命館大学産業社会学部）によって制作されたテキストマイニング用ソフトウエア。形態素解析ソフト茶筅 2.0，茶筅の辞書 IPADIC2.7，統計ソフト R3.1.0 等が同梱されている。

[4]個人情報保護の観点から「固有名詞（組織，人名，地域)」を排除したのち，記号等も排除した。

[5]本研究で構築したコーパスとは，李他（2012）に依る広義のコーパスであり，筆者の日本語教育実践の根拠とすべく構築され，使用される言語資料のことである。

語数では名詞，サ変名詞，動詞 B，動詞の順に多い。しかし，異なり語数では名詞，サ変名詞，動詞，動詞 B の順となる。動詞 B 上位の「する」「なる」「ある」「できる」等の基本語は，それぞれの出現回数が特に多いためと考えられる。KH Coder ではこれら基本語をひらがな表記の動詞「動詞 B」として別の品詞にしているため，漢字を含む動詞の出現状況が把握しやすい。

表１　４～６年生のお便り部数と語彙数

	4 年生	5 年生	6 年生
お便り部数	212	268	232
総語数	56,968	106,084	77,167
異なり語数	7,420	9,935	9,395

表２　４～６年生のお便り　品詞別出現語彙数

		４年生	５年生	６年生
名詞	総語数	16507	28867	20074
	異なり語数	2422	3298	2951
サ変名詞	総語数	9189	17913	12760
	異なり語数	974	1263	1225
動詞	総語数	4410	9844	6532
	異なり語数	721	967	885
動詞 B	総語数	6319	12866	9573
	異なり語数	463	609	573

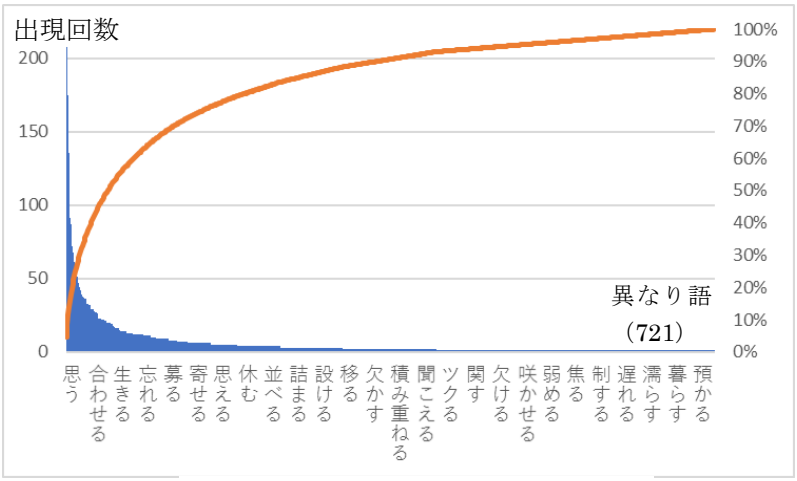


図１　　４年生　動詞　頻度表

縦軸に出現回数，横軸に頻度をとって頻度別の棒グラフを作成すると，L 字型の分布を表す(図 1)。少数の高頻度語と多くの種類の低頻度語[6]で構成されている。例えば，４年生の動詞の総語数は 4410 で異なり語数は 721 であるが，100 位「慣れる」までで総語数に占める語の割合は 66%となり，199 位「見せる」で 80%となる。ここから，頻度の高い語を優先的に学習することで，一定の教育効果を生むと考察する。学習する外国人保護者にとっては急に理解できる語が増えたように感じられ，動機づけが高まる効果があるのではないか。

一方で，順位の低いものを学習しても，実際に受け取るプリントに登場しない確率が増え

[6]ジップ（ZIP）の法則。日本語以外の言語調査でも確認される（萩野・田能村 2011）。

ていく。いくら学んでも上達したという実感が得られにくいという状況が生まれてしまうと予想される。しかしながら，地引(2013)で指摘される通り，頻度の低い語の中にも内容理解や行動につながる重要語が含まれる場合がある。従って，低頻度語については，お便りに出てきたときにその都度辞書を引くなり，周りの助言を得るなどして，意味を理解することが有効だと指導し，大量の語彙を覚えさせようとするのではなく，適切に辞書を使用するストラテジーや，質問をするストラテジーを，合わせて指導することが有効ではないか。

５　縦断的観察による最頻出語の抽出

５．１　研究方法

本稿のコーパスは 3 学年分のお便りの語彙の量と内容が縦断的に観察できる点が特徴である。その点を生かし，課題 1 で得られた頻度表を基に、学年毎の異同を観察し，4 年生の高頻出語上位 100 語のうち，5 年生，6 年生でも上位 100 位に入る語がいくつあるか，品詞（名詞，サ変名詞，動詞，動詞 B[7]）ごとの網羅率を調査した（表 3，図 2）。更に 3 学年全てで 100 位以内に入っている語を「最頻出語」とし，品詞ごとに抽出した。

５．２　結果と考察

4 年生の上位 100 語と同じ言葉は，名詞では 5 年生で 74 語，6 年生で 75 語が上位 100 位に入っている。サ変名詞は 5 年生で 83 語，6 年生で 81 語，動詞は 5 年生で 82 語，6 年生で 79 語。動詞 B については，5 年生で 72 語，6 年生で 74 語が重複していた。平均値は 77.5%である。毎年多くの語が繰り返し学校お便り文に用いられていることが明らかになった。

このような，毎年用いられる語の学習は，次年度以降に保護者がお便り文を読む際の助けになると考えられる。そのため，3 学年全てで上位 100 位に入っている語を最頻出語とし，特に優先して学習する語として提示する。支援現場ではまずその効果の可能性に言及し，動機づけを高めることが肝要であるだろう。

学習順序として，２例提案する。①4 年生の児童をもつ外国人保護者にまず 4 年生の各品詞の上位語を提示し，次年度以降にそれ以外の高頻出語彙を提示する。②最頻出語を，割合の多い名詞、動詞の順に提示する。支援現場で接するそれぞれの保護者に応じて、適切な学習順序を組み立てれば，効率的な語彙の学習が可能になるのではないだろうか。

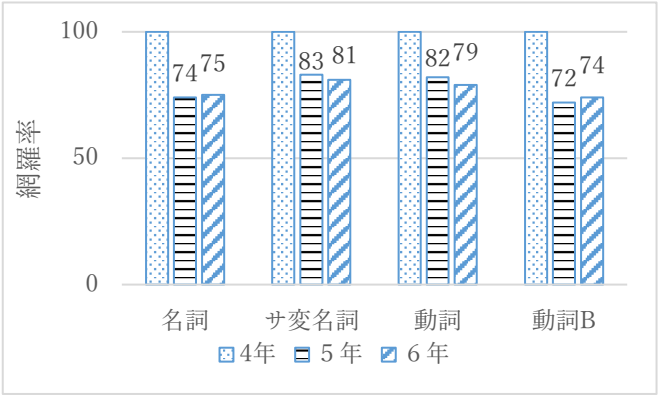


図 2　　4 年生上位 100 語の 5 年生，6 年生での網羅率

品詞ごとの最頻出語と，それぞれの語の 4 年生時の出現回数を表 3-6 に示す。

7 動詞 B は KH Coder 内の品詞名であり，茶筅の出力における品詞名では「動詞－自立(平仮名のみの語)」である。

表３　　名詞　最頻出語

委員 533　牛乳 333　学校 317　小学校 306　先生 281　国語 254　子ども 220 算数 188
学級 174　児童 167　学年 143　体育 143　本部 131　学期 118　クラス 111　運動会 111
パン 109　市立 109　役員 109　大会 95　交通 93　音楽 90　食品 87　年度 85　家庭 84
自分 80　教室 79　行事 76　地区 76　総会 75　地域 74　皆様 72　スープ 69　校外 69
米 68　小麦粉 63　内容 61　場所 58　スポーツ 57 皆さん 56　会員 55　校長 54　情報 54
キャベツ 53　会長 52　自転車 50　遠足 49　野菜 49　用紙 47　氏名 43　社会 40
夏休み 39　市内 38　目標 38　コーン 37　小さじ 35　校庭 34　様子 33　チーズ 32
体育館 31　アンケート 29　個人 29　（62 語）

表４　サ変名詞　最頻出語

活動 308　給食 239　参加 195　お願い 194　指導 189　保護 179　協力 156　予定 114
見学 112　報告 109　授業 101　お知らせ 99　連絡 96　運営 92　懇談 92　担任 91　生活 90
理解 86　提出 84　テスト 80　教育 77　学習 74　協議 70　親睦 68　練習 68　準備 67
成人 66　水泳 63　代表 63　注意 61　発行 60　参観 58　記入 57　会計 56　メール 55
集金 55　開催 52　下校 51　運動 50　確認 49　当番 49　選出 47　パトロール 45　登録 42
配布 41　お手伝い 40　広報 40　意見 37　募集 37　お話 35　試食 35　紹介 35　応援 34
会議 34　卒業 32　電話 32　一緒 31　出席 31　担当 31　利用 31　話 31　企画 30　実施 30
マーク 29　作成 29　体験 29　成長 28　通学 28　希望 27　開始 26　仕事 26　面談 26
監査 25　計画 25　使用 22　（75 語）

表５　動詞　最頻出語

思う 208　行う 175　食べる 136　考える 98　使う 91　見る 87　出る 72　入る 68　待つ 61
入れる 54　作る 53　行く 52　書く 51　向ける 47　聞く 44　始まる 42　終わる 39
出す 39　楽しむ 38　煮る 37　過ごす 36　洗う 36　感じる 33　言う 33　読む 33
決める 32　学ぶ 29　合わせる 29　出来る 29　守る 28　加える 27　知る 27　来る 27
含む 23　申し上げる 23 切る 23　迎える 22　取り組む 22　整える 22　深める 21
増える 21　続く 21　伝える 20　防ぐ 20　遊ぶ 20　話す 20　頑張る 19　残す 19　違う 18
取る 17　引き取る 16　混ぜる 16　走る 16　終える 15　始める 14　支える 14　置く 14
配る 14　教える 13　見守る 13　受ける 13　焼く 13　話し合う 13　開く 12　覚える 12
願う 12　得る 12　分かる 12　歩く 12　忘れる 12　帰る 11　見える 11　乗る 11
付ける 11　変わる 11　立つ 11　合う 10（77 語）

表６　動詞 B　最頻出語

する 2085　なる 531　ある 410　できる 304　いる 244　くる 88　いう 84　いただく 71
やく 65　つく 64　ちる 63　つける 58　こる 54　とる 54　つくる 48　つる 44　みる 44
きる 41　よる 40　あげる 39　わかる 33　くださる 31　こめる 30　もやす 29　かむ 27
ゆでる 26　いただける 25　かく 25　かける 23　ふる 23　やる 23　でる 22　にる 22
たつ 18　がんばる 17　めんじる 17　おく 15　ござる 15　いく 13　かかる 13　ける 13
わる 13　たく 12　もつ 12　さる 11　まとめる 11　もらう 10　もる 10　ひく 9　ねる 8
まつ 8　あう 7　しく 7　しめる 7　かう 6　(55 語)

６．一学年に特有の語の出現状況

６．１　研究方法

課題１において得られた頻度表の名詞，サ変名詞，動詞の上位 120 語までを観察し，ある学年に特有の語彙として「卒業」に着目し，3 学年での出現回数とその順位を比較した。

６．２　研究結果と考察

サ変名詞「卒業」は、4 年生で 32 回，5 年生で 66 回，6 年生で 104 回出現する。「卒業」は 6 年生で体験する行事であることから，6 年生に多いことは予想されたが，他学年でも頻度は低くはなく，上位 100 位以内に入っていることが注目される。「学校だより」等，全学年に向けて配布されるお便りによって，目にするものと考えられる。

ある学年に特有の言葉が他の学年にも高頻度で出現する場合があることが明らかになった。それらの語の持つ文化背景まで学ぶ機会を設けた場合には，高頻度語彙の学習が学校文化という内容スキーマを活性させることに役立つ可能性があると考察する。

表７　「卒業」の学年毎の順位と出現回数

	4 年	5 年	6 年
頻度順位	61 位	59 位	20 位
出現回数	32 回	66 回	104 回

７．おわりに

学校お便り文からコーパスを構築し，縦断的に観察・分析することで，少数の頻度上位語の数が全体に占める割合が多いことから，高頻出語の学習は効果があると考察した。また，3 学年に共通して出現していた高頻出語を「最頻出語」として抽出し，優先的に学習すべき語彙として提案した。更に，一学年に特有の語が他の学年にも高頻度で出現する場合があり，これらの語の学習が内容スキーマを活性させることに役立つ可能性に言及した。

本研究は，生活する外国人の日本語支援をする立場から行った。今後は，本コーパスを教育実践に生かすため，「行事名」や「月別」などでタグ付けし，利便性を高めていくことが課題である。更に，得られた知見を基に，具体的なカリキュラムデザインや教材の作成を進めたい。

文　献

伊藤孝恵(2007)「国際結婚夫婦のコミュニケーションに関する問題背景：外国人妻を中心に」『言語文化と日本語 教育』33 号, pp5-72

荻野綱男　田野村忠温(2011)『講座 IT と日本語教育 5 コーパスの作成と活用』明治書院

桑原陽子(2017)「初級読解教材作成を目指した非漢字系初級学習者の読解」『国立国語研究所論集』13 号, pp127-141

地引愛(2013)「小学校配布物から情報を得るために必要な語彙の探索：使用頻度の高い語彙に注目して」『学習院大学国語国文学会誌』56 号, pp76-92.

富谷玲子・内海由美子・仁科浩美（2012）「子育て場面で外国人保護者が直面する書き言葉の課題　－保育園・幼稚園児の保護者を対象とした調査から―『神奈川大学言語研究』34：pp. 53-71

李暁燕(2017)「外国人保護者に対する日本語支援－小学校配布プリントの特徴および「学校カルチャー語彙」の分析を通じて-」『地球社会総合科学』24,2 号,pp1-12
李在鎬・石川慎一郎・砂川有里子（2012）『日本語教育のためのコーパス調査入門』くろしお出版

関連 URL

独立行政法人国立国語研究所日本語教育基盤情報センター学習項目グループ・評価基準グループ（2009)『生活のための日本語：全国調査」結果報告』＜速報版＞ http://www.ninjal.ac.jp/products/syllabus/research/pdf/seika_sokuhou.pdf)

児童・生徒作文の日本語修辞ユニット分析と教員評価の検討

田中 弥生（東京大学大学院　総合文化研究科）†

Study of Rhetorical Unit Analysis and Teacher Evaluation of Composition Corpus of Japanese Elementary and Junior High School Students

Yayoi Tanaka (Tokyo University)

要旨

本研究は，作文評価への修辞機能と脱文脈化程度という新たな観点の提示を試みるものである。選択体系機能言語理論における英語談話分析手法の一つである修辞ユニット分析の手順を基に，テキスト内で用いられている修辞機能を特定し，脱文脈化程度が高い表現か低い表現か，すなわち事象が一般的なこととして表現されているか，特定の個人的なことやその場のことと表現されているかなどを示す。作文は従来，使用語彙，文字種，語種，文の種類や構造など様々な観点で評価されてきた。本発表では，児童・生徒によって同一のテーマで書かれた作文について，学年の違いによる修辞機能と脱文脈化程度の用い方の特徴を明らかにするとともに，小中学校の教員による評価の高低との連関から，作文評価における新たな観点の提示の可能性を検討する。

１．はじめに

作文はこれまで，使用語彙や，漢字の割合などの文字表記(鈴木ほか 2011，宮城・今田 2015b)，文の種類や構造(石田・森 1985)，言語形式による類型化(笹島 2017)など様々な観点から評価されてきた。

現在，検討を進めている日本語テキスト分析のための手法は，選択体系機能言語理論における英語談話分析手法の一つである修辞ユニット分析（Rhetorical Unit Analysis，以下 RUA）を基にしている。テキスト内のメッセージ（おおむね節）単位で修辞機能を特定し，脱文脈化程度が高い表現か，低い表現か，すなわち事象が一般的なこととして表現されているか，専門的なこととして表現されているか，個人的なことと表現されているかを示すことができる。

岩田(1995)によると，教室授業では情報の“脱文脈化”と“文脈化”という二つの知的作業を子どもたちに課しているという(岩田 1995 : 33)。また，学校における児童の発達に関して，「学校では，その多くの学びが，実生活の状況とは切り離され，脱文脈化された概念的な知識の獲得にむけられる。日常の生活文脈に密着した状況的な思考から，日常の生活文脈から切り離された，抽象的で無性格的な思考に慣れることが重要な課題となるのである」(岩田 1995 : 33)と述べている。学校での授業や経験を重ねることによって，すなわち学年が上がるとともに，脱文脈化表現を使用できるようになることが考えられる。文脈化した（脱文脈化程度の低い）個人的な経験を，一般的な知識として理解し，脱文脈化した表現で示したり，反対に，脱文脈化した一般的な知識を文脈化させて自分自身に引き付けて表現したりする。例えば，一緒に遊んでいる弟がボールを持っていることを描写するのに，「弟は今ボ

† yayoit22@gmail.com

ールを持っている」と書くこともできれば，「人の手は物をつかむことができる」と表現することもできる。学校で脱文脈化された概念的な知識が獲得されるとすれば，脱文脈化程度の高い表現の使用は，子どもの成長を測る一つの観点ととらえることができるだろう。また，これまでは教員が作文を評価する際に明示的にとらえられていなかった脱文脈化程度の観点が，どのように評価と関わっているのかを確認することによって，評価基準の検討も可能になると考えられる。

本研究は，まず，同一のテーマで書かれた小２と中２の作文を分析対象として，学年の違いによる作文における修辞機能と脱文脈化程度の特徴を明らかにする。次に，修辞機能と脱文脈化程度の様相と小中学校の教員による評価との連関を検討する。

以下，２．でRUAの概要を示し，３．で分析方法を述べ，４．分析結果の報告と考察を行い，５．でまとめと今後の課題について述べる。

２．修辞ユニット分析の概要と先行研究

Cloran(1994,1995,1999)によって英語母子会話の分析について提案されたRUAは，テキストの意味単位を特定するための手法(佐野 2010b)だが，その過程においてメッセージ（おおむね節）[1]単位で，発話機能(speech function)，中核要素(central entity)，現象定位(event orientation)の三つを認定することで，修辞機能（rhetorical function）の種類を特定し，その結果として脱文脈化程度（degree of de-contextualization）を知ることができる。談話や文章においてどのような修辞機能が用いられているか，使用されている言語表現は脱文脈化程度が高いか低いか（「いま・ここ・わたし」から遠いか近いか），という観点からの検討が可能になる。RUAは佐野・小磯(2011)によって日本語への適用が検討され，英語と日本語の言語の違いに関わる修正が加えられている。佐野(2010b)は，特定目的の作文指導への利用について，修辞機能と脱文脈化指数の出現状況から作文の専門性の判断が可能であり，専門性に問題がある場合には，中核要素や現象定位の変更の指導によって改善が可能であることを述べている。また，田中(2017)では，日本語非母語話者を読み手と想定して日本語母語話者が作成した自治会加入勧誘のチラシについて，わかりやすさ評定で上位のチラシと下位のチラシでは用いられている修辞機能に違いがあることを述べている。

ある事象について述べる際に，個人的なこととして表現するか，一般的なこととして表現するか，あるいは，当事者として語るか，一般論として語るか，聞き手・読み手の受ける印象は異なるため，目的に応じた使い分けが必要である。RUAはその差を捉えることができる手法であるといえるだろう。

３．分析

３.１　分析対象

本研究の分析対象は，『児童・作文コーパス』(宮城・今田 2015a)の中の「「手」作文コーパス」(阿部ほか 2017)に格納されている作文である。「「手」作文コーパス」は，1992年と2016年に小学生・中学生を対象として同一の条件で調査した作文を収集したものである。同一の国立大学附属小中学校で，「これから『手』という題で作文を書きます。どんなことを書いても自由です。原稿用紙1枚（400字）で書きます」という指示のもと，作文が書かれたという（阿部ほか 2017：237）。本発表では，2016年に書かれた小２と中２の作文から

[1] ３.２.１参照のこと

ランダムにサンプリングされた計 24 件を分析対象とする。表 1 に，作文サンプル数，文の数，1 作文の平均文数，最多・最小の文数を示す。一つの作文内の文の数は最少が 8 文，最多が 20 文で，平均は 13〜14 文程度である。

表 1. 作文数・文の数

	作文サンプル数	文の数	1 作文平均文数	最多文数	最少文数
小 2	6	86	14.3	17	11
中 2	18	223	12.4	20	8
計	24	309	12.9	20	8

3.2 分析手法

RUA では，１．メッセージとその種類を認定し，２．発話機能・中核要素・現象定位を認定し，３．修辞機能の特定と脱文脈化指数の確認を行う。表 2 に示したように，発話機能・中核要素・現象定位の組合せから修辞機能を特定し，脱文脈化指数を確認する。

表 2. 発話行為・中核要素・現象定位からの修辞機能の特定と脱文脈化指数の確認

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

「n/a」は該当なし／背景が薄い灰色の部分が修辞機能の種類／[ ]内は脱文脈化指数
（佐野（2010b）および佐野・小磯（2011）の修辞機能の特定表に脱文脈化指数を合わせて示したもの）

図 1 に，修辞機能と脱文脈化指数をその指数の順番に並べて示した。脱文脈化指数とは，中核要素の here（発話地点との空間的な距離）の程度と現象定位の now（発話時点との時間的な距離）の程度によって，近いものから遠いものまで修辞機能を線上に示した際の順序の指数である。脱文脈化指数[1]は例えば「手を見せてください」のような，コミュニケーションが行われているその場面と時間の「今・ここ」に最も近い表現である。一方，脱文脈化指数[14]は，例えば「人間の手は物をつかむための器官である」のような，そのコミュニケーションの場面・時間とは直接関わりのない，つまり「今・ここ」から最も遠い表現である。

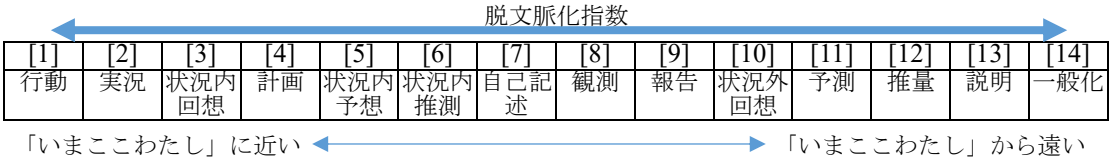


図 1. 修辞機能と脱文脈化指数

以下，分析手順について述べる。

３.２.１　メッセージの認定

まず，分析対象をメッセージ単位に分割し，種類を認定する。図２にメッセージの種類を示す。メッセージは原則としておおむね節であるが，本発表では，日本語における RUA の基準を検討する過程として，句点を区切りと考え，分析単位とする。主部や述部が省略されていると考えられる場合には，補足する。

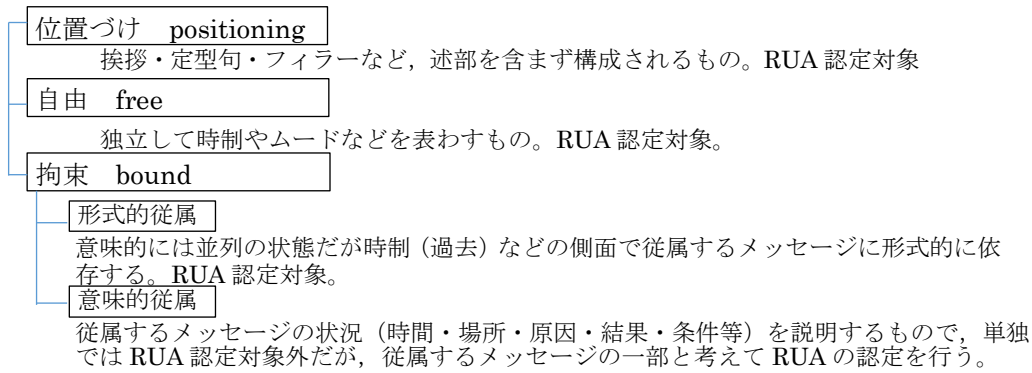


図 2　メッセージの種類

「位置づけ」に分類されるのは，「おはよう」や「ありがとうございます」のような挨拶や定型句，フィラー，述部を含まないものなどである。本研究の分析対象サンプルである(1)の(ii)は，名詞の羅列であり述部などが省略されているわけではないと判断して「位置づけ」に分類した。「自由」に分類されるのは(1)(i)のような単文や(2)の主節部分である。従属節は「拘束」に分類され，その機能から「形式的従属」と「意味的従属」に分けられる[2]。ただし本発表では，上述のように句点を区切りと考え，従属節はいずれも単独での認定の対象とせず，メッセージの種類が「自由」のものを対象とする[3]。なお便宜上本稿では，以降，分析単位を「メッセージ」と呼ぶ。

また，(3)のように「～と思う」や「～と言った」などの引用表現については投射[4]とし，被投射部分「～」が分析対象で，「自由」に分類される。

(1)　(i)母の手は、毎日忙しく働いている。
　　(ii)お弁当を作る手、洗濯物を干す手、パソコンのキーボードを打つ手、私をなでてくれた手。　(2016-8-f-3)[5]

(2)　そういうときはしょうどくをし、ばんそうこうをだいたいはります。(2016-2-m-4)
　　自由

(3)　だから私は人類は自分達の手のおかげで現在地球の頂点にたてていると思う。
　　自由　投射
　　(2016-8-m-4)

[2] 形式的従属と意味的従属の例はここでは省略する。
[3] 主節のみで分析した場合と従属節も含めて分析した場合の比較を今後行う。
[4] 投射については，佐野(2010a)を参照のこと。
[5] 本稿の例文は分析対象コーパス内のサンプルである。括弧内にサンプル ID を示す。

３.２.２　発話機能の認定

分析対象を確認した後，発話機能の認定を行う。発話機能は，「提言 proposal」か「命題 proposition」に分類する（Halliday and Matthiessen 2004）。「提言」は表 3 の(a)の品物・行為の交換（提供あるいは命令）に関するメッセージ，「命題」は(b)の情報の交換（陳述あるいは質問）に関するメッセージが該当する。

表 3　発話機能 (Halliday & Matthiessen 2004: 107)

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

本発表の分析対象データでは，品物や行為をやり取りする「提言」は見られなかった。例えば「手を見せてください」のような，発話機能が「提言」のメッセージがある場合には，この段階で，修辞機能は「行動」，脱文脈化指数は最も低い[1]と認定される。一方，３．２．１で例示した(1)の(i), (2)(3)のような，情報を提供したり，あるいは要求したりする，「命題」と分類されるメッセージについては，この後，中核要素と現象定位の認定を行い，修辞機能と脱文脈化指数を確認する。

３.２.３　中核要素の認定

中核要素は，メッセージの中心となるもので，基本的には主語によって表現される。メッセージの内容の中心がコミュニケーションの場面に存在するか否か，場面とは関わらず存在するものか，によって分類し，メッセージの発信者との空間的距離を示す。照応など前後のメッセージを用いて判断する場合もある。中核要素の分類を図 3 に示す。

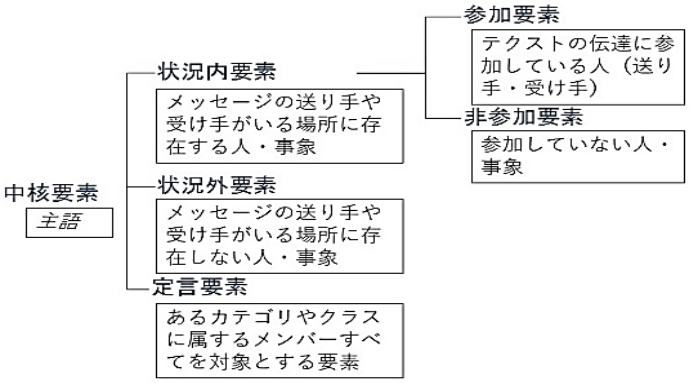


図 3　中核要素の分類(佐野・小磯 2011)

３.２.４　現象定位の認定

現象定位は，メッセージの発信時点とその内容との時間的距離を示す要素で，副詞や述部から判断する。メッセージによって表現されている出来事がすでに起こっていることか，ま

だ起こっていないのかを，メッセージが発信されている時（Time of speaking）を基準とした時間的な位置を特定する。現象定位の分類を図 4 に示す。

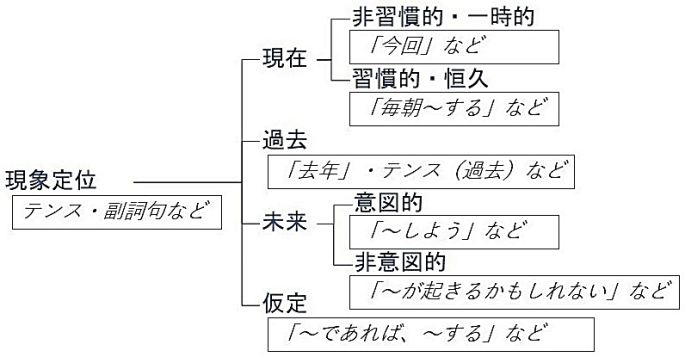


図 4　現象定位の分類（佐野・小磯 2011 を加筆修正）

３.２.５　修辞機能の特定と脱文脈化指数の確認

表２に示したように，発話機能と中核要素と現象定位の組み合わせによって，修辞機能が特定され，脱文脈化指数を確認できる。脱文脈化程度が低いメッセージは，その文脈すなわち作文の書き手自身のことや，存在する時空間のことを述べているのに対し，脱文脈化指数の高いメッセージは，作文の書き手の存在する場や時間から空間的・時間的に離れたことやその場・その瞬間に関わらないことを述べている。

３.３　教員の評価との連関

本研究の分析対象コーパスの作文は，小中学校の教員によって評価が行われている。表４に示したように，評価基準は，「0.総合評価」の他に１から７まで設けられているが，本発表では，小中学校の教員による「0.総合評価」と，修辞機能・脱文脈化程度との間の連関について，確認する（４．２参照)。

表 4. 作文の評価基準　（それぞれ 5 段階評価）

0. 総合評価
1. 題材に関連ある内容を決めて書けている
2. 書こうとする内容についてよく考えて詳しく書いている
3. 学年相当の語彙を書けている
4. 学年相当の漢字が書けている
5. 文末表現に変化をつけて書けている
6. 語と語、文と文を適切な繋がりを作って書けている
7. 自分の考えが伝わるように順序を意識して書けている

４．分析結果と考察

４.１　定量的分析

「手」作文コーパスの小２と中２のデータを分析した結果を表５と図５に示す。「対象外」とは，(1)(ii)に示したような「位置づけ」のメッセージである。

表 5. 分析結果　　単位：件

	対象外	1 行動	2 実況	3 状況内回想	4 計画	5 状況内予想	6 状況内推測	7 自己記述	8 観測	9 報告	10 状況外回想	11 予測	12 推量	13 説明	14 一般化	計
小 2	0	0	7	17	4	0	0	4	5	8	0	1	1	33	6	86
中 2	5	0	38	13	17	3	0	8	5	46	7	2	5	66	8	223
計	5	0	45	30	21	3	0	12	10	54	7	3	6	99	14	309

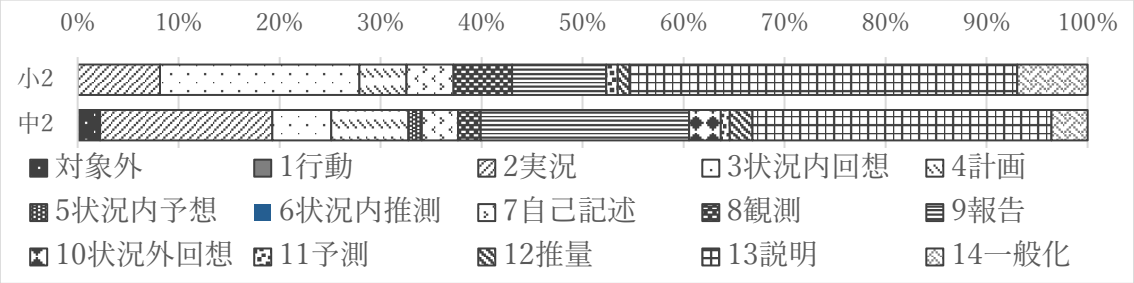


図 5. 小２と中２の修辞機能の比較

小２では,「13 説明」の割合が最も高く，次いで「3 状況内回想」が多く用いられている。「13 説明」とは，(4)のように，手がどういう働きがあるかを示したり，解説したりするもので,「3 状況内回想」とは，(5)のように手にまつわるこれまでの経験を述べるものである。

(4) 手は、いろいろなことにかつやくしてます。　(2016-2-m-4)

(5) わたしの手は、はりがささったり、ぶつけたことがありました。　(2016-2-f-4)

一方中２では「13 説明」「9 報告」「2 実況」の割合が高い。(6)は「説明」の例で，「たくさんの人が」（中核要素＆状況外）と「それになやまされているそうです」（現象定位＆現在；習慣的・恒久）の組み合わせで多汗症の人の悩みについて説明している。(7)は「報告」の例で，省略されている「母の手」（中核要素＆状況外）について，様子を述べている。(8)は「実況」の例で，実態を述べている。

(6) 姉によると多汗しょうの人は手汗がすごくてたくさんの人がそれになやまされてるそうです。　(2016-8-f-5)

(7) (φ＝母の手は)[6]細くて骨張っている小さな手で、今にも折れてしまいそうだ。　(2016-8-f-3)

(8) 私もその一人である。　(2016-8-f-4)

次に，空間的距離と時間的距離の二つの軸から検討する。RUA では，中核要素は，メッセージの発信者である話し手・書き手とメッセージ内容との空間的距離を示し，現象定位は，メッセージが発信された時とメッセージ内容との時間的距離を示している。図６は，小２と中２でそれぞれ用いられている修辞機能の割合をバブルの大きさで示したものである。縦軸は空間的距離で，下方が「ここ・わたし」に近く，上方が遠い。横軸は時間的距離で，左側が「いま・わたし」に近く，右側が遠い。

6 省略されていると考えられる部分は（φ＝ ）で示す。

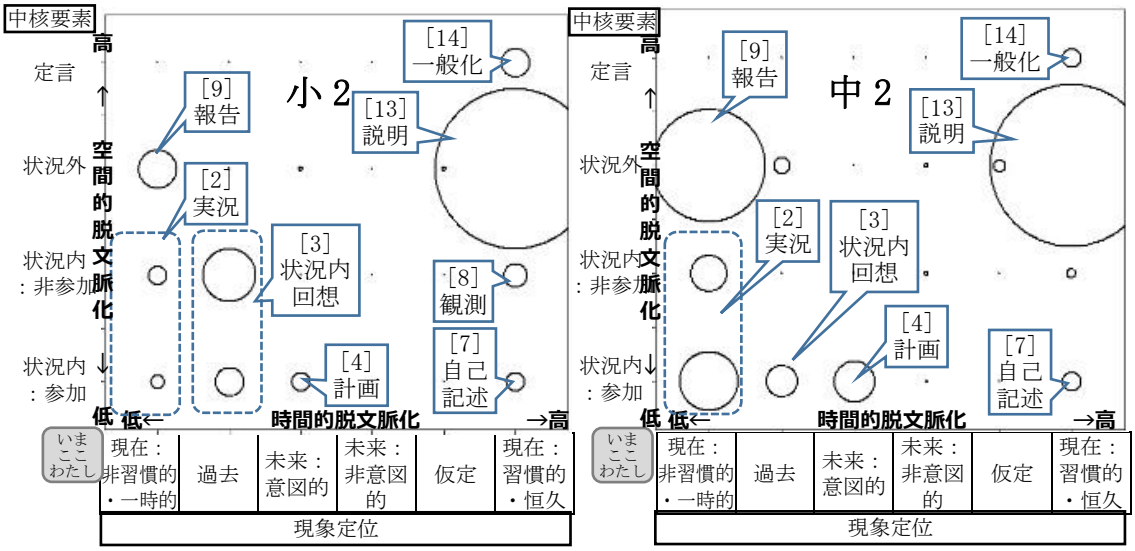


図 6. 空間的距離と時間的距離による比較

縦軸の空間的距離の観点では，小２，中２ともに，下部より上部の方が多く，作文の書き手のいる場所である「ここ・わたし」から遠い，空間的脱文脈化程度の高い表現が多い。一方，横軸の時間的距離の観点からは，小２では「いま・わたし」から遠い右側の時間的脱文脈化程度の高い表現が多く，中２は小２に比較して時間的脱文脈化程度の低い表現も使用している。学校生活の長い中２の作文の方が小２よりも脱文脈化程度の高い表現が現れることが予測されたが，本研究の結果では，小２，中２ともに脱文脈化程度の高い表現が用いられており，中２は様々な脱文脈化程度を使用できるようになることがうかがえた。

４.２　教員評価と修辞機能・脱文脈化程度の連関

本節では，脱文脈化程度の高低と教員による評価との連関を検討する。図７と図８に，中２と小２の各作文の修辞機能の出現頻度と，教員による総合評価の中央値を◇で示した。

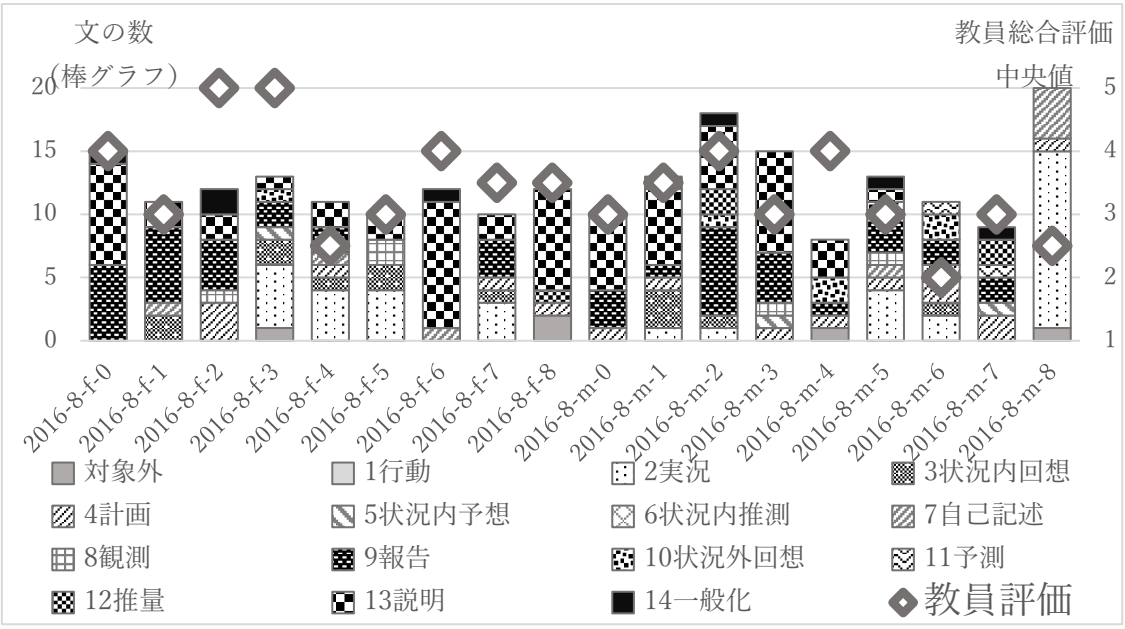


図 7. 中２作文の修辞機能と教員総合評価の中央値

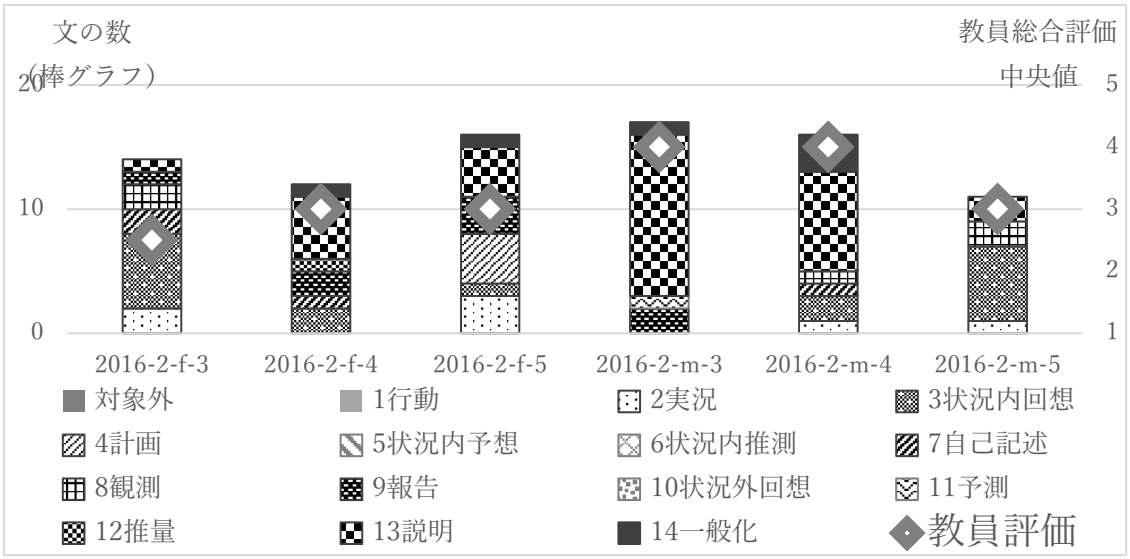


図 8．小２作文の修辞機能と教員総合評価の中央値

図７の 2016-8-m-8（右端）や図８の 2016-2-f-3（左端）のように，脱文脈化程度の低い修辞機能が多く用いられている作文は評価が低く，図７の 2016-8-f-0（左端）や 2016-8-f-2（左から３つ目），図８の 2016-2-m-3，2016-2-m-4（左から４つ目，５つ目）のように，教員総合評価の中央値が 3.5 以上の作文では，いずれも脱文脈化程度の高い修辞機能を用いており，脱文脈化程度の高い修辞機能の使用と教員の評価には連関があることがうかがえる。なお，図７の 2016-8-f-3（左から４つ目）のように脱文脈化程度の低い修辞機能が 5 割を占めていても評価が高い作文もある。

作文内の文の数における脱文脈化程度の高い「14 一般化」と「13 説明」が用いられている割合と教員評価の相関を確認したところ，評定値は r= 0.418 の有意な相関を示した（df=22, p=0.042）。相関の強さは中程度であり，脱文脈化程度の高い表現が用いられていると教員の評価が高くなる傾向があるといえる。また，脱文脈化程度の低い「2 実況」と「3 状況内回想」が用いられている割合と教員評価については，評定値は r=-0.369 で弱い負の相関があるが，有意ではなかった（df = 22, p = 0.07609）。このことから，脱文脈化程度の高い修辞機能の使用は高い評価を得られる可能性があるが，脱文脈化程度の低い修辞機能が低い評価につながるわけではないことがうかがえる。

４.３　作文内の修辞機能・脱文脈化程度の推移

本節では，一つの作文の中での修辞機能・脱文脈化程度の出現と，空間的距離および時間的距離の二つのレベルにおける推移を確認する。

表６では，小２の作文で，教員による総合評価点の中央値が高いものの一つ（2016-2-m-3，図８の左から４つ目）についての，発話機能・中核要素・現象定位，特定される修辞機能と脱文脈化指数を示した。また，図９に当該作文における修辞機能・脱文脈化程度の推移を示した。この作文では最も脱文脈化程度の高い「14 一般化」から始まり，数回，時間的脱文脈化程度の低い「9 報告」「11 予測」が含まれる以外は，ほとんどの文が「13 説明」に集中しており，空間的脱文脈化の低い修辞機能は用いられていない。脱文脈化程度の高い修辞機能が使用されて，教員評価が高い作文の例である。

表 6. 作文の中での修辞機能と脱文脈化指数 小２（サンプル 2016-2-m-3）

	文	発話機能	中核要素	現象定位	修辞機能	脱文脈化指数
1	**手は**、人げんとどうぶつに***あります***[7]。	命題	定言	現在;習慣的・恒久	一般化	14
2	**人げんは**、手がない人も***います***。	命題	状況外	現在;習慣的・恒久	説明	13
3	**手がないひとには**、スプーン、フォークがもてないので***たいへんです***。	命題	状況外	現在;習慣的・恒久	説明	13
4	手があれば(**φ＝人は**)なんでも***もてます***。	命題	状況外	現在;習慣的・恒久	説明	13
5	(**φ＝人は**)ボールやえんぴつ、けしゴム、ふでばこだって、***もてます***。	命題	状況外	現在;習慣的・恒久	説明	13
6	**手は**、もつだけでは、***ありません***。	命題	状況外	現在;習慣的・恒久	説明	13
7	（**φ＝手は**）***ころがしたり、つかんだりできます***。	命題	状況外	現在;習慣的・恒久	説明	13
8	**手は**、あぶないときだって***やくにたちます***。	命題	状況外	現在;習慣的・恒久	説明	13
9	ころんだりしそうになっても、手が先に、じめんい、つけば、きけんなときも手があるので(**φ＝人は**)***顔をぶつけないですむのです***。	命題	状況外	現在;習慣的・恒久	説明	13
10	**手は**、どうぶつとも***ふれあえます***。	命題	状況外	現在;習慣的・恒久	説明	13
11	**手だけであそべることも*あります***。	命題	状況外	現在;非習慣・一時的	報告	09
12	(**φ＝人は**)ジャンケンやアルプスいちまんじゃくで***あそべます***。	命題	状況外	現在;非習慣・一時的	報告	09
13	手があれば、家や水てっぽうなどがつくれるので、**手は**、***やくにたてます***。	命題	状況外	現在;習慣的・恒久	説明	13
14	**手は**、***体の一ぶです***。	命題	状況外	現在;習慣的・恒久	説明	13
15	手には、**ゆびが*あります***。	命題	状況外	現在;習慣的・恒久	説明	13
16	ゆびがあるから、(**φ＝人は**)***手あそびや、どうぐをつくったり、きけんをふせいだりできます***。	命題	状況外	現在;習慣的・恒久	説明	13
17	(**φ＝人は**)これからも手をつかって、***いろいろなことをやってみるのもいいです***。	命題	状況外	未来;意図的	予測	11

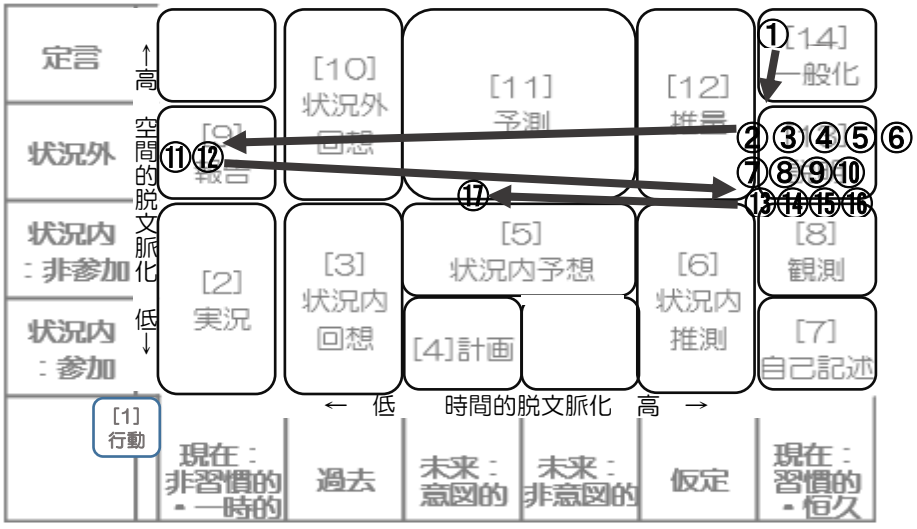


図 9. 作文内での修辞機能・脱文脈化程度の推移(2016-2-m-3)（小２上位）

一方，図１０には，小２で教員による総合評価点の中央値が低いものの修辞機能の推移を示した。

[7] 各文の中の太字ゴシックは中核要素部分，太字イタリックは現象定位部分を示している。

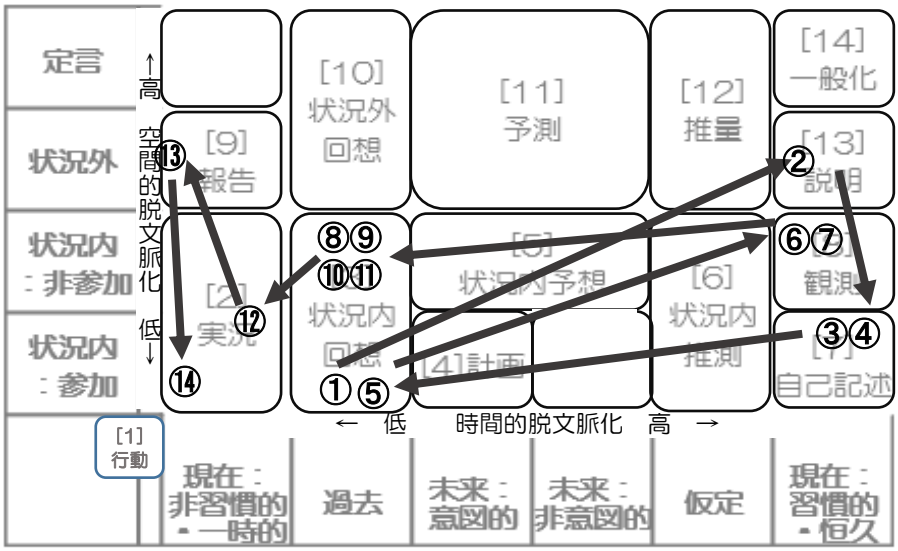


図 10. 作文内での修辞機能・脱文脈化程度の推移(2016-2-f-3)（小２下位）

「13 説明」と「9 報告」も用いられているが，それ以外は空間的脱文脈化程度が低い修辞機能（縦軸の下部）で，特に個人的な経験を述べる「3 状況内回想」が多く用いられている。小２上位の図９と比較すると，空間的脱文脈化程度の使用に違いがあることがわかる。小２の作文については，脱文脈化程度の高いほうが高評価になる可能性がうかがえる。

表７は，中２の作文（４.２のグラフ（図７））の左から４つ目（2016-8-f-3）の分析である。

表 7. 作文の中での修辞機能と脱文脈化指数 中２（サンプル　2016-8-f-3）

	文	発話機能	中核要素	現象定位	修辞機能	脱文脈化指数
1	**私は**母の手が***あまり好きではない***。	命題	状況内;参加	現在;非習慣・一時的	実況	2
2	**（φ＝母の手は）**細くて骨張っている小さな手で、***今にも折れてしまいそうだ***。	命題	状況外	現在;非習慣・一時的	報告	9
3	***私の手の方がよっぽど健康的である。***	命題	状況内;非参加	現在;非習慣・一時的	実況	2
4	しかし何故だか、**（φ＝私は）**母の手を握ると***安心感がある***。	命題	状況内;参加	現在;非習慣・一時的	実況	2
5	こんなに細いのに、**（φ＝私は）**何かに***守られている、という感じがする。***	命題	状況内;参加	現在;非習慣・一時的	実況	2
6	**（φ＝私は）**それが***不思議でならなかった。***	命題	状況内;参加	過去	状況内回想	3
7	***母の手は、毎日忙しく働いている。***	命題	状況外	現在;習慣的・恒久	説明	13
8	お弁当を作る手、洗濯物を干す手、パソコンのキーボードを打つ手、私をなでてくれた手。	対象外				
9	**（φ＝母の手は）**細く小さな手で、数え切れない程たくさんの事を私達に教え、***導いてきてくれた。***	命題	状況外	過去	状況外回想	10
10	その手には**弱さは**みじんも***感じられない。***	命題	状況外	現在;非習慣・一時的	報告	9
11	強い強い母の手なのだということに**（φ＝私は）*気付かされた。***	命題	状況内;参加	過去	状況内回想	3
12	そんな母の手を、**（φ＝私は）**今では少し***自慢に思っている。***	命題	状況内;参加	現在;非習慣・一時的	実況	2
13	自分も将来、**強い母の手を子供に見せられるときが*来るのだろうか。***	命題	状況内;非参加	未来;非意図的	状況内予想	5

図１１に，表７の作文の修辞機能の推移を示した。この作文は「私は母の手があまり好きではない」という，自身の感情を表現する「いま・ここ・わたし」に近い，文脈化した「2

実況」から始まり，空間的・時間的脱文脈化程度を移動しながら展開している。脱文脈化程度の低い修辞機能が 5 割を占めているが教員の評価は高いこの作文からは，脱文脈化程度が高いだけでなく，高低を交えた様々な修辞機能を用いた展開が高く評価される可能性がうかがえる。

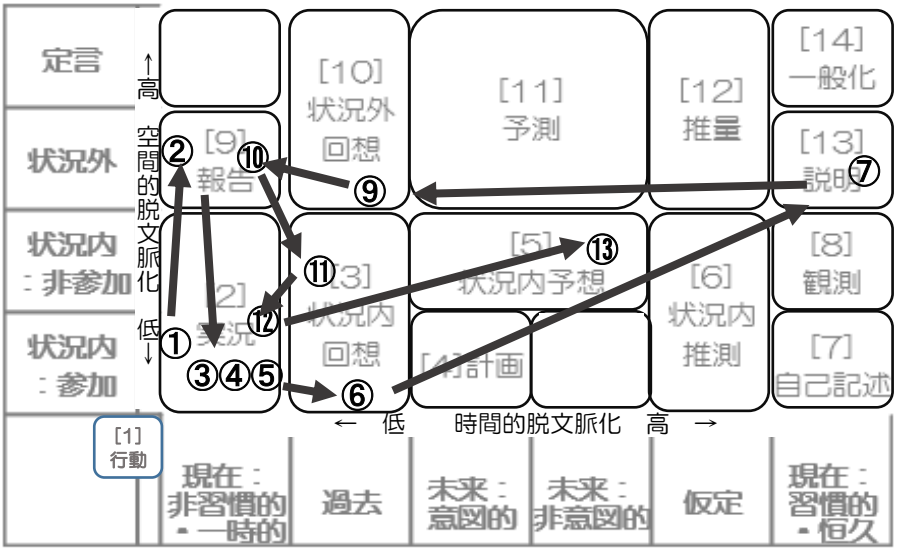


図 11. 作文内での修辞機能・脱文脈化程度の推移(2016-8-f-3)（中２上位）

８番目の文は分析対象外のため「⑧」は欠番である。

一方，中２で教員評価が下位の作文（図７の右から３番目）では，図１２に示すように個人的な経験などを述べる「３状況内回想」から始まり，時間的脱文脈化程度の低い修辞機能に集中している。複数の修辞機能は用いられているが，脱文脈化程度の高い表現が用いられていない作文は評価が低くなる可能性がうかがえる。

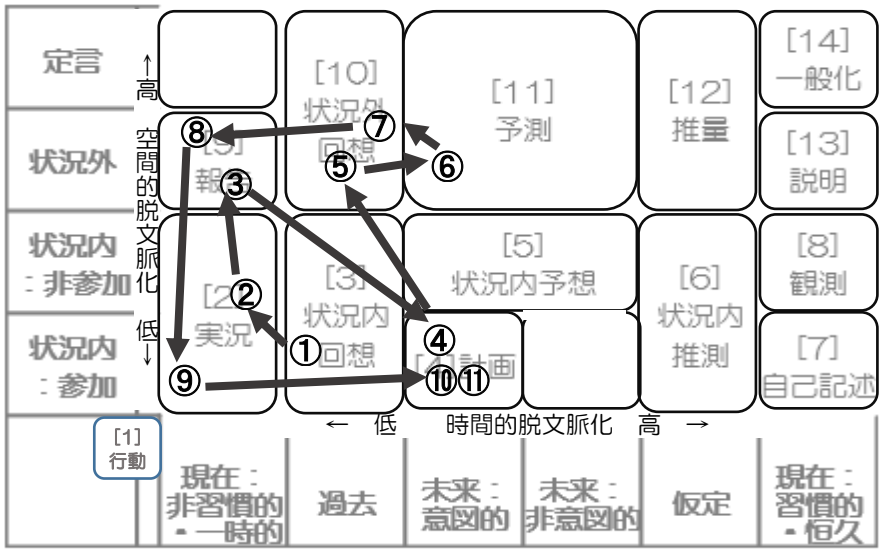


図 12. 作文内での修辞機能・脱文脈化程度の推移(2016-8-m-6)（中２下位）

５．おわりに

本研究は，作文を評価する新たな観点として修辞機能と脱文脈化程度を提示するために，まず，同一のテーマで書かれた小２と中２の作文を分析対象として，作文における学年の違いによる修辞機能と脱文脈化程度の特徴を明らかにし，次に，修辞機能と脱文脈化程度の様相と小中学校の教員による評価との連関を検討した。

分析の結果，学年が上がるにつれ脱文脈化程度の高い表現を用いるようになるという予測とは異なって，小２も高い脱文脈化程度の修辞機能を用いていることが明らかになり，成長とともに，様々な脱文脈化程度の修辞機能を使用できるようになることがうかがえた。

教員評価と脱文脈化程度の連関については，脱文脈化程度の高い修辞機能の使用と高い評価には有意な相関があることが明らかになった。しかし，脱文脈化程度の低い修辞機能の使用が必ずしも低い評価につながるのではなく，様々な修辞機能を用いて高い評価を得ている作文もあり，脱文脈化程度の単純な高低が評価と結び付くわけではないことがうかがえた。

本発表は，「「手」作文コーパス」の一部データについての分析であるが，修辞ユニット分析の手法によって修辞機能及び脱文脈化程度の観点から作文を分析し，作文の様相を確認できることを提示できたと考える。教員評価との連関について，「0.総合評価」以外の 7 つの評価基準も含めた検討は，今後の課題とする。

「意見文」「説明文」など作文の種類によって，教員が望ましいと考える脱文脈化程度の高低は変わる可能性がある。今後，種類の異なる作文についても検証を行い，作文評価や作文指導への有効な活用の提案を検討していく予定である。

謝　辞

本研究の作文データは博報財団「児童教育実践についての研究助成」（助成番号：2016-053）の成果の一部を富山大の宮城氏から提供いただいたものです。また，本研究の一部は科研費基盤(C) (15K02535)によるものです。

文　献

Cloran, C. (1994) Rhetorical Units and Decontextualisation: an Enquiry into some Relations of Context, Meaning and Grammar. Monographs in Systemic Linguistic Linguistics, No.6. Nottingham: Department of English Studies, University of Nottingham.

Cloran, C. (1995) Defining and Relating Text Segments: Subject and Theme in Discourse, In R. Hasan and P. Fries (eds) On Subject and Theme: From a Discourse Functional Perspective. Amsterdam: Benjamins.

Cloran, C. (1999) Contexts for learning. In Christie, F. (ed.) Pedagogy and the Shaping of Consciousness, London: Cassell, 31-65.

Halliday, M. A. K. and Matthiessen. C.M.I.M. (2004) An Introduction to Functional Grammar (3rd ed.) London: Arnold.

阿部藤子・今田水穂・宗我部義則・冨士原紀絵・松崎史周・宮城信(2017)「児童生徒の「手」作文に於ける経年変化の計量的分析 : 1992 年と 2016 年の作文を比較して」『資源活用ワークショップ発表論文集』2, pp.234-247.　　doi/10.15084/00001478

冨士原紀絵・宮城信・松崎史周(2016).「児童生徒作文の基礎的研究: 児童生徒作文コーパ

スの構築と活用」『お茶の水女子大学子ども学研究紀要』4, pp.9‐20.
http://hdl.handle.net/10083/60027　からダウンロード可能
石田潤・森敏昭(1985)「小学生の文章表現の発達的変化」『広島大学教育学部紀要』1, 33, pp.125-131.
岩田純一(1995)「学校と発達」　岩田純一他（著）『児童の心理学』有斐閣
宮城信・今田水穂（2015a）「『児童・生徒作文コーパス』の設計」,『第 7 回コーパス日本語学ワークショップ予稿集』, pp.223-232, 国立国語研究所
https://www.ninjal.ac.jp/event/specialists/project-meeting/m-2014/jclws07/ からダウンロード可能
宮城信・今田水穂. (2015b). 「作文コーパスを資料に児童・生徒の漢字使用・選択傾向と発達の実態を明らかにする―語彙情報つき作文コーパスの構築と学齢別語彙・漢字使用実態調査―」『漢字・日本語教育研究』5, pp.4‐20.
笹島眞実(2017)「言語形式に基づく児童作文の類型化」『言語資源活用ワークショップ発表論文集』2, pp.290-296.　doi/10.15084/00001530
佐野大樹(2010a)日本語における修辞ユニット分析の方法と手順 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）
http://researchmap.jp/systemists/資料公開/　(RUA の方法と手順 ver.0.1.1)　2018/4/11 閲覧
佐野大樹(2010b)「選択体系機能言語理論を基底とする 特定目的のための作文指導方法について ―修辞ユニットの概念から見たテクストの専門性―」『専門日本語教育研究』12, pp.19-26.　doi.org/10.11448/jtje.12.19
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.　https://www.jasfl.jp/journal/Journal_vol6.pdf からダウンロード可能
鈴木一史・棚橋尚子・河内昭浩(2011)「作文コーパスからみる生徒の使用語彙」『特定領域「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』
pp.343-350.　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/workshop/JC-G-10-02.pdf からダウンロード可能
田中弥生(2017)「日本語非母語話者向け自治会加入勧誘チラシとその作成振り返りコメントの分析―修辞機能と脱文脈化程度の観点から―」『言語情報科学』16, pp.73-88, 東京大学大学院総合文化研究科言語情報科学専攻.　http://hdl.handle.net/2261/00074675 からダウンロード可能

日本語における慣用句の逸脱使用がもつ言語機能 —形容詞の反義語への置き換えを手がかりに—

鈴木 あすみ（東北大学文学研究科）†

Linguistic functions of deviated use of Japanese idioms: focusing on antonyms substitution for the adjective

Asumi Suzuki (Tohoku University)

要旨

「食が細い」という慣用句に対して「食が太い」のような言い方は一般に認められないが、文学作品や広告などにおいて言葉遊びとして用いられることがある。本稿では『国語研日本語ウェブコーパス (NWJC)』を用いた調査から、こうした逸脱的な言い回しが特殊な修辞的技巧をもたない人々にも多く用いられることを示し、その対人的機能について考察する。

１．はじめに

１.１　慣用句の固定性と可変性

慣用句は一般に固定性が高く、語彙的柔軟性を欠くとされている (Liu, 2008; 宮地, 1982)。慣用句に含まれる語をその反義語に置き換えると、慣用句として認められないものになる場合が多い。例えば「厚い」と「薄い」は反義関係にあるが、(1a) に対して (1b) は慣用句として成立せず (西尾, 1985)、慣用句辞書等への掲載も見られない。

(1) a. つらの皮が厚い
　　b. *つらの皮がうすい

(西尾, 1985)

しかし、このような句は慣用句とは認められずとも、いわゆる「言葉遊び」としての使用・理解は可能である。(2) の下線部 a, b に示す三浦梅園の言葉は倫理的な学問観を戯画的・風刺的に表現し、学者にありがちな学識の見せびらかしを指摘するものである (木村, 1973)。

(2) 学問は飯と心得べし。腹にあくの為なり。かけ物などのように人に見せんずる為にはあらず。衣装うつくしくかざり、人に好かれんとするは売女なり。人の見る時躰をなし、人に褒められんとするは歌舞伎ものなり。今の学者はどうやらこの真似するようなり。a. 足の皮はあつきがよし。b. つらの皮はうすきがよし。人もろともに小ざかしく口はきけど行いは女童に見限らる。さるゆえ面の皮あつくなり、足の皮うすくなり、株ふむこと多し。よく心得てつつしむべし。

(三浦梅園『戯示学徒』, 成立年代不明)

鈴木 (2018) では『国語研日本語ウェブコーパス (NWJC)』を用いた調査から、慣用句に含まれる形容詞をその反義語に置き換える用法が実際には多く用いられていることを確認した。(3)、(4) はそれぞれよく知られている慣用句「食が細い」「息が長い」の「細い」「長

† asumi.suzuki1028@gmail.com

い」をその反義語に置き換えた句の用例である。

(3) でも以外と食が太く、モリモリ食べて育ってくれました
【出典】http://44555.blog119.fc2.com/blog-entry-91.html

(4) また、割と長く持つキーワードと、旬と言うか流行の息の短いキーワードを選ぶかでも対応が変わってきます。
【出典】http://affirikouryaku.blog104.fc2.com/blog-date-201001.html,http://affirikouryaku.blog104.fc2.com/blog-category-8.html

１.２　慣用表現およびその変形がもつ機能

イディオムをはじめ、複数の語から成る慣用的な表現は phraseological units (PU) と呼ばれる。Fiedler (2007) によると、PU は装飾的な機能のみでなく、テキストを構成するために重要な機能を果たしており、比喩的な意味を持つ PU は字義通りの意味と併せて洒落を作るために用いられることもある。また、PU にはテキストの中心となってその結束性を高める働きもある (Safina et al., 2015)。

PU には固定的な性質があり、最も典型的な実現形として規準形をもつ (Philip, 2008)。しかし、変形を全く許容しないというわけではなく、修飾語など規準形には無い要素を挿入する (extension)、構成要素の一部を省略する (ellipsis)、構成要素の一部を他の語に置き換える (substitution) などといった変形操作を伴って用いられることもある (Partington, 1996; Safina et al., 2015)。Safina et al. (2015) は雑誌と新聞に基づいた調査から、書き手が PU を変形して用いるのは文章の表現力を高めるためであると指摘している。Safina et al. (2015) によると、PU を変形する用法にはテキストに分析 (analytism) や個人的な評価、芸術的描写、皮肉を付け加える機能がある。

１.３　本稿の目的

前節までに挙げたものを含め、PU の機能に関する先行研究には雑誌や新聞、文学作品、広告などの用例に基づくものが多い。これらは記者やコピーライターなど、高度な修辞技巧を身に付けた人々によるテキストである。しかし、PU の変異形はより日常的な言語使用の中にも現れうる。本稿では『国語研日本語ウェブコーパス (NWJC)』を用いた調査を行い、

①慣用句の逸脱使用はどのような人々がどのような場面で用いるのか

②慣用句の逸脱使用はどのような言語機能をもっているのか

の 2 点を明らかにすることを目的としている。ブログや掲示板、Q&A サイトといったウェブテキストには、

・書き手の多くは特殊な修辞技巧を身に付けた人々ではない

・談話が 1 人の書き手で完結せず、複数人が相互に書き手／受け手になり得る[1]

という特徴がある。このような性質をもつウェブテキストに基づく分析は、PU およびその変異形がもつ対人的な言語機能を明らかにすることにつながる。

[1] 例えば、ブログ本文に対してはコメント、質問に対しては回答、掲示板の書き込みに対してはレスポンスが寄せられることがある。このような場合、それぞれの書き手は異なるのが普通である。

２．調査

２.１　調査方法

本稿では慣用句の一部を反義語に置き換えることで作られる、一般に慣用句とは認められない用法を手がかりに PU の変異形がもつ機能について検討する。どのようなウェブサイトで用いられているか、談話の中でどのように用いられているかという点に着目し、鈴木 (2018) のデータを再分析した。次節にデータ獲得の手順を示す。

２.２　調査対象とコーパスからの用例抽出

予備調査として、まずは形容詞を中心とする慣用句の中で最も典型的な「名詞＋が＋形容詞」(西尾, 1985) を辞書形にもつものを現代言語研究会 (2007)、丹野 (1998)、米川・大谷 (2005) の 3 冊の日本語慣用句辞書から目視で抽出した。それらのうち、反義関係にある形容詞の片方だけが慣用句として用いられる句 (例：食が細い／*太い) について、「*食が太い」のように辞書に掲載が無い方の句 14 種の用例を『国語研日本語ウェブコーパス (NWJC)』から抽出した。検索条件は形容詞をキーとし、その前方 3 語以内に慣用句に含まれる名詞 (例：食) が共起する例を収集した。検索結果総数が 300 を超える場合は全検索結果の中からランダムで抽出した 300 例、それ以下であれば全検索結果を予備調査の範囲 (表 1「分析例」) とした。こうして設定した範囲の中から慣用句的な意味での用例のみを目視で拾い、その粗頻度 (表 1「うち慣用句的用法」) および分析例に対する比率 (表 1「慣用句的用法比率」) の高いものを本調査の対象として選んだ。「頭が柔らかい (考え方が柔軟な様子)」「息が短い (長続きしない様子)」「腰が軽い (気軽に行動を起こす様子)」「食が太い (たくさん食べる体質)」「神経が細い (些細なことでも気にしてしまう性質)」「造詣が浅い (知識が乏しい様子)」である (表 1 星印)。

表 1　予備調査結果

調査対象	検索結果総数	分析例	うち慣用句的用法	慣用句的用法比率 (%)	対象
造詣が浅い	80	80	80	100.00	★
頭が柔らかい	2,319	300	284	94.67	★
食が太い	162	162	145	89.51	★
神経が細い	447	300	246	82.00	★
面の皮が薄い	20	20	14	70.00	
息が短い	142	142	93	65.49	★
態度が小さい	69	69	43	62.32	
腰が軽い	997	300	124	41.33	★
血の気が少ない	38	38	15	39.47	
影が濃い	1,142	300	108	36.00	
足が軽い	2,566	300	9	3.00	
心臓が弱い	2,212	300	7	2.33	
気が少ない	289	289	0	0.00	
耳が近い	223	223	0	0.00	

これら 6 つの句について、「太い食」のようにキーと共起語の順序が逆になる用例も同様

に抽出した。検索結果は全て目視で確認して調査対象に該当しない例[2]を取り除き、最終的に表 2 に示す数の用例を得た。

表 2　NWJC の検索結果における各調査対象の用例数

	頭が柔らかい	神経が細い	腰が軽い	食が太い	息が短い	造詣が浅い	合計
用例総数	4,017	414	377	145	87	72	5,112

２.３　調査

まず「①慣用句の逸脱使用はどのような人々がどのような場面で用いるのか」を明らかにするため、調査対象の句がどのようなウェブサイトで用いられているか、そのレジスターを調査した。レジスターの分類はブログ、掲示板、ニュース、ネット小説などいくつかを思いつく範囲で設定し、それに当てはまらないものが出現した場合は適宜追加した。同様に「②慣用句の逸脱使用はどのような言語機能をもっているのか」についても示唆を得るため、調査対象の句が談話の中でどう用いられているかについて調査した。これらの 2 点についてはいずれも得られた全用例のうち URL が有効なものの元サイトをたどり、目視で調査した。

３．結果

３．１　調査対象の句が現れるレジスター

調査対象とした表現は様々な種類のウェブサイトで確認された (表 3)。特に多いのはブログ、Q&A サイト、掲示板などである。「その他」にまとめたものの中にはネットニュースやメールマガジン、ツイート、会話の書き起こし、行政のサイトなどが含まれる。このようなサイトに見られるテキストの書き手はほとんどが文章の専門家ではなく、取り扱うテーマも日常的なものが多い。

表 3　調査対象の句が用いられるウェブサイトのレジスター

	頭が柔らかい	神経が細い	腰が軽い	食が太い	息が短い	造詣が浅い	合計
ブログ（個人）	2,510	265	263	109	53	51	3,251
ブログ（団体）	105	3	8	2	1	1	120
Q&A サイト	200	19	11	6	4	0	240
掲示板	172	24	9	0	8	3	216
企業サイト	31	1	1	0	0	0	33
ネット小説	17	18	4	3	0	0	42
投稿記事	23	0	3	0	1	1	28
その他	159	6	19	3	8	1	196
合計[3]	3,217	336	318	123	75	57	4,126

[2] 検索条件に一致しても慣用句的な表現でない例も存在する。「この、うどんという日本の国民食、太い麺、強いコシ、胃の中で膨張する」(出典：http://tnc.typepad.jp/main/2006/01/post_ec55.html) などである。

[3] 検索結果のうち URL の有効なもののみを調査対象としたため、表 2 と表 3 では合計数が異なる。

３．２　調査対象の句の用法

調査対象の句には比喩的な意味と文字通りの意味があり、これらをかけて洒落を作る用例 (5) が見られた。また、元の慣用句と対比させることで洒落を作る用例 (6) も確認された。これらを合わせ、洒落を作る用法は全部で 35 例確認された (表 4)。

(5) この、どーでもいい話ができる時間ってとても大切。
ヘッドマッサージより頭が柔らかくなるのです
【出典】http://kekeblog.seesaa.net/article/156066477.html

(6) ぎゃははは〜〜〜！嫁太。５００円そこそこで痩せようとするその神経のほうが図太いで〜〜！先にその神経を細くしたほうがええのとちゃうか〜〜！ぎゃははは〜〜〜！！
【出典】http://estrella-zrx.cocolog-nifty.com/blog/2008/03/post_195a.html

気の利いた表現・ウィットに富んだ表現を作る用例 (7) も見られた。このような用法は数が少なく、書き手も修辞的な文章に慣れ親しんでいると思われる[4]。(7) では「息」を含む慣用句とその変異形がいくつも用いられており、テキストにおいて中心的な役割を果たしている。

(7) 「新国訳大蔵経」という息の長い媒体と「朝日新聞」という息の短い媒体を舞台にしての論争というのは、それこそ息が合いそうにありません
【出典】http://web.kyoto-inet.or.jp/people/kiraya/news0902.html

また、ブログやQ&Aサイト、掲示板などで複数の書き手が談話に参与している場合には、書き手 A の発話に含まれる慣用句を受けて書き手 B が対義語への置き換えがなされた句を用いる例が確認された。(8) ではブログに寄せられたコメント (a) に対して本文の書き手からコメントが返されており、(9) では掲示板への書き込み (a) に対して他者からレスポンス (b) が寄せられている。また、書き手 A の発話に対義語への置き換えがなされた句が含まれており、それを受けて書き手 B が元の慣用句を用いる例も確認された (10)。このように対話の中で反義語への置き換えが用いられる例は 49 例確認された (表 4)。

(8) a. 今年こそ重い腰あげて採寸しようかな
b. いやいや、腰が軽い時で（笑）
【出典】 http://pochiat.blog.fc2.com/blog-entry-845.html

(9) a. 何百人って人間が見てる中であんな戦い方が出来る図太い神経には呆れた
b. >>125 お前の神経か細過ぎワロタ
【出典】 http://ff.doorblog.jp/archives/50285859.html

(10) a. 中学生になってもまだまだごんたまちゃんの頭は柔らかいのね
b. だけどこういう発想って　いつの間にか
なくなっちゃって　頭が固くなってるんですね〜
【出典】 http://sutekinakaisheru.blog75.fc2.com/blog-entry-323.html

[4] (6) は古書店のウェブサイトからの用例である。

表 4　調査対象ごとの洒落および対話における用例数

	頭が 柔らかい	神経が 細い	腰が 軽い	食が 太い	息が 短い	造詣が 浅い	合計
洒落	20	4	10	1	0	0	35
対話	27	4	6	9	2	1	49
合計	47	8	16	10	2	1	84

４．考察

調査の結果から、慣用句に含まれる語を反義語に置き換える用法は特殊な修辞技巧をもたない人々にも日常的に用いられていることが明らかになった。このような表現がもつ機能としては、以下のようなものが考えられる。

①洒落や気の利いた文をつくる機能

修辞的な表現に慣れ親しんだ書き手による技巧的な例 (7) のみならず、変形前の慣用句や文字通りの意味とかけて洒落を作り出す用例 (5, 6) が確認された。この用法はブログなど日常性の高いテキストの中にも現れるものである。これらはいわゆる「上手い事を言う」言い回しであり、表現そのものを際立たせたり、受け手の注意を引いたりする働きがあると考えられる。

②談話参与者間の相互的な機能

ブログや掲示板、Q&A サイトなどにおいて複数の書き手が談話に参与している場合、他者の発言に含まれる元の慣用句を受けて、反義語への置き換えが用いられる例 (8, 9) やその逆 (10) が確認された。このような場合、PU およびその変異形は表現そのものを際立たせて受け手の注意を引く機能の他に、自分が相手の話をきちんと読んでいることを確認する働きも担っていると考えられる。

以上のことから、既存の慣用表現の一部を変形して創り出された新奇な言い回しは、詩的機能や交話的機能 (ヤコブソン, 1984) を担っていることが示唆される。こうした言葉遊び的な用法は狭義の「伝達」ではなく、書き手と受け手の関係構築に主眼を置く言語表現である。

５．おわりに

書籍や新聞、雑誌等は修辞的な文章を書くのに慣れた人々の産出したテキストであり、書き手から受け手に向けて一方向的な伝達が行われることがほとんどである。これに対して、ウェブテキストには様々なレジスターがあり、書き手の属性も多種多様である。中でもブログや掲示板、Q&A サイトなどでは、書き言葉でありながらも複数の参与者による双方向的な対話が行われている場合が多い。今後も一般的な母語話者の日常的なコミュニケーションにおける PU の用法・機能の解明に向け、大規模ウェブコーパスの活用が進むことが期待される。今後は語の置き換え以外の様々な変形も含め調査を進めたい。

文　献

Sabine Fiedler (2007) English Phraseology: A Coursebook. Tubingen: Gunter Narr Verlag.

ロマン・ヤコブソン (1984)『言語とメタ言語』(池上嘉彦、山中桂一訳) 東京：勁草書房.
現代言語研究会 (2007) 『日本語を使いさばく　慣用句の辞典 』東京：あすとろ出版.
木村俊夫 (1973)「三浦梅園の学問観と方法論」『茨城大学教育学部紀要』23，189-196.
国立国語研究所コーパス開発センター編 (2017)『国語研日本語ウェブコーパス』(2014-4Q データ，梵天バージョン 1.0.0). https://bonten.ninjal.ac.jp/, (2018 年 6 月 20 日確認)
Dilin Liu (2008) Idioms: description, comprehension, acquisition, and pedagogy. New York: Routledge.
宮地裕 (1982) 『慣用句の意味と用法』東京：明治書院.
西尾寅弥 (1985)「形容詞慣用句」『日本語学』4 (1)，45-53.
Alan Partington (1998) Patterns and meanings: using corpora for English language research and teaching. Amsterdam; Philadelphia: J. Benjamins Pub.
Gill Philip (2008) Reassessing the canon: 'Fixed' phrases in general reference corpora. In Sylviane Granger and Fanny Meunier (eds.) Phraseology: An interdisciplinary perspective, 95-108, Amsterdam; Philadelphia: J. Benjamins Pub.
Rimma A. Safina, Elena V. Varlamova & Elena A. Tulusina (2015) The Stylistic Potential of the Contextual Usage of Phraseological Units as Hybrid Formations, *Asian Social Science*, 11:19, pp. 64-69.
鈴木あすみ (2018) 「日本語における慣用句の自由度について―形容詞の反義語を手がかりに―」東北大学大学院文学研究科修士論文 (未公開)
丹野顯 (1998) 『意味から引ける慣用句辞典』東京：日本実業出版社.
米川明彦、大谷伊都子 (2005) 『日本語慣用句辞典』東京：東京堂出版.

関連 URL

『国語研日本語ウェブコーパス』検索系『梵天』 http://bonten.ninjal.ac.jp/

日本語歴史コーパスの現代語辞書における未知語義判定システム

田邊 絢（茨城大学大学院理工学研究科）
古宮 嘉那子（茨城大学工学部情報工学科）
浅原 正幸（国立国語研究所コーパス開発センター）
佐々木 稔（茨城大学工学部情報工学科）
新納 浩幸（茨城大学工学部情報工学科）

Detecting Unknown Word Senses in Contemporary Japanese Dictionary from Corpus of Historical Japanese

Aya Tanabe (Ibaraki University)
Kanako Komiya (Ibaraki University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Minoru Sasaki (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要旨

日本語歴史コーパス中の単語には、現代語と同様の意味で扱われている単語と、古語特有の意味を持つ単語がある。本研究では、この現代語にはない古語特有の単語の語義（言葉の意味）を未知語義と定義して、日本語歴史コーパス中から、未知語義を検出するシステムの提案を行う。具体的には、日本語歴史コーパス中の単語を、(1)現代の分類語彙表でその単語の分類番号として登録されている語義をもつ語、(2)現代の分類語彙表にある語義をもつが、現在その語義は、その言葉の語義として分類語彙表は登録されていない語、(3)その語義の定義が現代の分類語彙表にないため、分類番号が振られていない語、の 3 種類にクラス分けする。実験では、各単語について、出現書字形や見出しなどの 8 要素を基本素性として用いた。また、別の日本語歴史コーパスから word2vec を用いて、3 種類の単語の分散表現のベクトル(50 次元、100 次元、200 次元)を作成し、素性として加えた。それぞれ SVM を用いて正解率を比較したところ、日本語歴史コーパス中の未知語義の検出において、単語の分散表現のベクトルが正解率を向上させることが分かった。

１．はじめに

語義曖昧性解消をめぐるアプローチとしては、確率的な言語モデルに基づき、対象単語の前後にある単語の品詞や、各単語同士の共起関係などを特徴として用いて、コーパスから機械学習を行うなどの様々な試みがなされている。古文コーパスでの語義曖昧性解消を行う際に、現代文コーパスでの語義曖昧性解消のタスクをそのまま適用させようとすると、現代語と古語とでの単語の語義の違いから、現代語の分類で分類した語義には当てはまらず、新たに別の正しい語義を付与する必要のある古語の語義が存在する。このような現代文での分類においては未知となる古文の語義を、本研究において未知語義と呼ぶ。

現代文の語義曖昧性解消タスクに関する研究として、(Suzuki et al., (2018))がある。これは、コーパス中の全単語に対して、分類語彙表の類義語と分散表現を利用することで、対象単語とその類義語の周辺単語同士の類似性から語義を予測する手法を取っている。また、(遊佐ら, (2017))では、語義曖昧性解消タスクにおいて分類語彙表を用いることの有効性を示している。また、未知語の検出について、(新納ら, (2012))がある。これは、外れ値検出法を用いて、対象単語の語義が新語義となっている用例を検出する手法を提案している。

古文の語義分類に関して、(宮島ら, (2014))は古文作品における各単語の出現頻度に加えて、すべての単語に国立国語研究所編『分類語彙表』の分類番号を追記し、古語の分類語彙表としてまとめている。また、(小木曽, (2011))の古文用形態素解析辞書の開発に関する研究がある。これは、見出し語に短単位を採用することで、現代語と古語の仮名遣いの違いを考慮し、地の文と会話文とで別に辞書生成することで古文用の形態素解析辞書の解析精度が向上することを示している。

さらに、本研究に関連する研究として、、通時コーパスの構築に関する研究がある。しかし、古文と現代文の違いや、古文同士でも書かれた時代によって構文や語義が大きく異なることなどから、通時コーパスの構築に関しては解決すべき課題が多くある。この通時コーパスの構築に通ずる研究の一部として、まず歴史コーパスにおける語義曖昧性解消を行うために、現代文の語義分類に当てはまらない古文の未知語義を検出し、その種類を分類することを本研究の目的としている。

２． 歴史コーパス

コーパスは、自然言語処理の研究に用いるため、テキストや発話などの文章を構造化し、言語的な情報（品詞、統語構造など）を付与して、大規模に集積しデータベース化した言語資料である。特に古文の文章に対して構造化したものを、歴史コーパスと呼ぶ。また、歴史コーパスにおいては、データ化する過程で文字表記の関係で外字処理、テキストの校訂など表記の置き換えがなされている。歴史コーパスは各語に関して、出現書字形（orthToken）、出現発音形（pronToken）、語彙素読み（reading）、語彙素（lemma）、原文文字列（originalText）、品詞（pos）、古語品詞（sysCType）、活用形（cForm）、語彙素番号（lemmaID）、分類番号（wlsp）の要素からなっている。特に分類番号に関して、ピリオドのあるものは現代語の語義、ピリオドのないものは古語の語義、と区別がされている。

本研究で用いる歴史コーパス中の、現代語にある語義に分類されている語、現代語では別の語が持っている語義に分類されている語、語義分類されていない語（本研究においての未知語義を持つ語）の割合を以下の図１に示す。

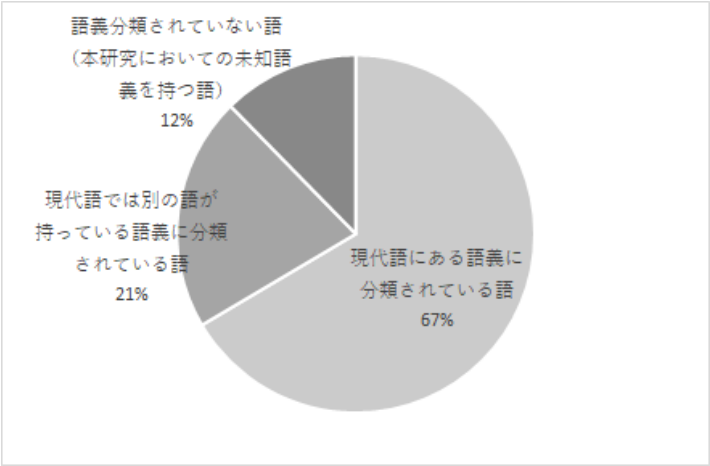


図１　歴史コーパス中の各分類語の割合

３．分散表現を利用した歴史コーパスにおける未知語義判定システム

３.１　原理

歴史コーパス中の語は、語義分類に関して、現代語にある語義に分類されている語、現代語では別の語が持っている語義に分類されている語、語義分類されていない語（本研究においての未知語義を持つ語）の 3 種類に分けられる。本研究ではこれらを分類した。、歴史コ

ーパス中の各語についている、形態素、品詞などの情報を、本実験のための素性として用いて学習データを生成した。また、歴史コーパスの情報に加えて、その語が単義語であるか多義語であるかの情報を素性として用いた。さらに、歴史コーパス中の各語について、別の歴史コーパスを用いて word2vec（Mikolov et al., 2013a, b, c）で生成した各単語の分散表現のベクトルを素性として用いた。生成した各学習データは、LIBLINEAR を用いて 5 分割の交差検定で学習を行う。

３.２　基本素性

本研究の学習データを生成する上での基本素性として、歴史コーパス中の各語について、出現書字形、出現発音形、語彙素などの 9 種類の要素を用いる。これらの要素を基本素性として学習データを生成している。学習データは以下の図 2 のような素性となる。

出現 書字形	出現 発音形	語彙素 読み	語彙素	原文 文字列	品詞	古語 品詞	活用形	語彙素 番号

図 2　学習データの素性

学習データを生成する過程で、歴史コーパス中の各語に関して、素性がすべて同一になる語は一度しか出現しないようにした。これは、歴史コーパスにおいて、特定のある同一単語が何度も出現することが多く、このことにより、LIBLINEAR を用いた 5 分割の交差検定の正解率に少なからず影響を及ぼす可能性が予測されるためである。

３.３　単語の分散表現を用いる手法

本研究では、3.2 節の基本素性に次の 2 つの素性を加えて学習データを生成して実験を行う。

ひとつは、各単語が現代語の語義分類において単義語であるか、多義語であるか、現代語の語義分類に存在しない単語（固有名詞など）であるかの 3 種類を示した素性である。もうひとつは、別の歴史コーパスから、word2vec（以下 w2v）を用いて 50 次元、100 次元、200 次元の分散表現のベクトルを生成し、歴史コーパス中の各単語に素性として付与したものである。

本研究において、歴史コーパスから w2v を用いてベクトルを生成する方法を説明する。まず、使用する歴史コーパスから見出し語を抜き出し、各語をすべて繋げた文章を MeCab を用いて分かち書きする。分かち書きしたデータに w2v を用いて、コーパス中宇の各語について、パラメータを変えて 50 次元、100 次元、200 次元の 3 種類の分散表現のベクトルのデータをそれぞれ得る。これらの分散表現のベクトルを、歴史コーパスから生成した学習データの各語について、ベクトル表現が存在する語のみそのベクトルを素性として追加し、ベクトル表現が存在しない語には零ベクトルを付与する。この学習データを単語の分散表現のベクトルを用いた学習データとした。単語の分散表現のベクトルを付与した学習データの素性を、以下の図 3 に示す。

出現 書字形	出現 発音形	語彙素 読み	語彙 素	原文 文字列	品 詞	古語 品詞	活用 形	語彙素 番号	**単義語か 多義語か**
								＋	**50 次元 or 100 次元 or 200 次元のベクトル**

図 3　分散表現のベクトルを加えた学習データ

以上の 2 種類の学習データに、LIBLINEAR を用いて 5 分割の交差検定を行い、正解率を調査する。

４．分散表現を利用した歴史コーパスにおける未知語義判定システムの実験

４.１　実験のデータ・設定

本実験では、歴史コーパスとして方丈記、竹取物語、虎明本狂言鬼小名、土佐日記、徒然草を使用した。この歴史コーパスの 9 つの要素を素性として、LIBLINEAR に用いるための学習データを生成した。

また、各語に対して w2v を用いた分散表現を得るための別の歴史コーパスとして、方丈記、竹取物語、虎明本狂言、土佐日記、徒然草の 5 つのコーパスをひとつに繋げて用いた。このコーパスから出現書字形の要素を抜き出し、MeCab を用いて分かち書きを行う。分かち書きしたデータに w2v を用いて、50 次元、100 次元、200 次元のベクトル表現のデータをそれぞれ得る。これらのベクトルを、歴史コーパスから生成した学習データの各語について、ベクトル表現が存在する語のみそのベクトルを素性として追加し、ベクトル表現が存在しない語には零ベクトルを付与する。以上のデータから、単語の分散表現を用いた学習データを生成した。

本実験では、word2vec による分散表現を使用した。以下の表 1 のパラメータで学習を行い、分散表現のベクトルのデータを得た。

表 1　w2v の学習パラメータ

C-BoW or skip-gram	-cbow	0
次元数	Size	50, 100, 200
ウィンドウ	-window	5
ネガティブサンプリング数	-negative	0
階層化 1:使用,0:未使用	-hs	1
最低頻度閾値	-sample	1e-3
バイナリデータ化	-binary	1
スレッドの個数	-thread	10

表 1 のパラメータにおいて、次元数の欄の 50、100、200 の 3 種類の数字は、それぞれ 50 次元、100 次元、200 次元に設定してそれぞれ 3 種類のベクトルのデータを得たことを示している。

本実験では、LIBLINEAR を用いて正解率を調査する。LIBLINEAR は、データを線形分離するための、機械学習において広く使用されているライブラリである。分類問題に対して、データに与えられた素性の情報から、線形カーネルを用いることで分類を行い、その正解率を算出する。本実験では、生成した学習データについて、それぞれ 5 分割の交差検定で学習を行い、正解率を調査した。

本手法を評価するにあたって、baseline を図 2 の素性から生成した学習データでの正解率の 77.391%とし、その学習データに、各語が単義語か多義語かの素性を加えた学習データ、さらに分散表現のベクトルを加えた学習データを用いる本手法との正解率を比較し、評価及び考察を行う。

また、w2v を用いて生成したそれぞれ 50 次元、100 次元、200 次元のベクトルを付与した学習データで実験を行い、次元数による正解率の違いについても確認し、評価および考察を行う。

４.２　実験結果

まず、baseline のデータに、各語が現代語において単義語か多義語かを示す素性を付与し

たデータでの実験結果は 79.713%となり、baseline よりもわずかに正解率が上昇した。次に、w2v で生成した 50 次元、100 次元、200 次元の分散表現のベクトルを付与したそれぞれのデータでの実験結果を以下の表 2 に示す。

表 2　ベクトルを付与した学習データの正解率（%）

baseline	単義語か 多義語か	50 次元の ベクトル付与	100 次元の ベクトル付与	200 次元の ベクトル付与
77.391	79.713	80.191	80.396	**80.533**

表 2 から、50 次元のベクトルを付与したデータでは 80.191%、100 次元のベクトルを付与したデータでは 80.396%、200 次元のベクトルを付与したデータでは 80.533%と、付与するベクトルの次元数が大きくなるほど、正解率が上昇することが確認できた。

５　考察

歴史コーパスのデータに w2v を用いて生成した分散表現のベクトルを加えた学習データを用いた実験では、分散表現のベクトルを素性として用いることにより、正解率が上がり、また、ベクトルの次元数が大きいほど、さらに正解率が高くなることが確認できた。これは、歴史コーパスの未知語義の検出に関して、分散表現を用いることの有効性を示している。

本実験では、分散表現を用いたすべてのデータで baseline を上回る正解率を得ることができたが、その上昇率はわずかであった。その要因として、分散表現のベクトルを得るために用いた歴史コーパスは、学習データに用いた歴史コーパスとは違う時代のものも入っており、時代の違いによる語義の違いによって正解率が落ちた可能性が考えられる。そこで、分類タグ付きコーパスの総データ量を増やすことや、また、同時代に編さんされた別の古文コーパスを用いて、より精度の高い分散表現を生成するなど、今後より正解率を上げるための作業および調査がさらに必要であると考える。

また、追加実験として、学習データの素性のうち、先頭三つの主要なの素性（品詞、形態素など）が正解率にどの程度貢献しているのかを実験し、調査した。学習データに用いていた 9 つの基本素性のうち、先頭の主要な素性である、出現書字形、語彙素、品詞の 3 つ素性を用いて、単義語か多義語かの素性、それぞれの次元数の分散表現のベクトルを加えて、先頭三つの素性のみでの学習データを作成した。追加実験の結果を以下の表 3 に示す。

表 3　先頭三つの素性のみの学習データの正解率（%）

	単義語か多義語か	50 次元の ベクトル付与	100 次元の ベクトル付与	200 次元の ベクトル付与
先頭三つの 素性のみ	79.131	79.843	79.843	79.487

表 3 から、先頭三つの素性のみの学習データでの正解率の方が、全ての素性を利用した場合の正解率より下がったことが確認できる。また、50 次元、100 次元の分散表現のベクトルの付与したデータよりも、200 次元の分散表現のベクトルを付与した場合のデータの方が正解率が下がっている。これは、先頭三つの素性の素性のみが必ずしも正解率に大きく寄与しているわけではないことを示している。

６　終わりに

本研究では、歴史コーパスに、単語の分散表現のベクトルを用いて、現代語の分類に当てはまらない未知語義の検出を行い、その種類を分類するシステムを提案した。実験の結果から、歴史コーパスにおける未知語の検出と分類において、分散表現のベクトルを用いること

の有効性、また、ベクトルの次元数が高いほど正解率が上昇することが示された。
また、追加実験の結果から、この未知語を検出するための学習データにおいて先頭三つの素性以外の情報も先頭三つの素性と同様に正解率に貢献していることが確認できた。

謝　辞

本研究は国立国語研究所のプロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「通時コーパスの構築と日本語史研究の新展開」への関連研究の一部として研究成果を報告したものである。

文　献

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou(2018), “All-words Word Sense Disambiguation Using Concept Embeddings”, LREC 2018, no 100,..

遊佐宣彦・佐々木稔・古宮嘉那子・新納浩幸(2017)「分散表現に基づく日本語語義曖昧性解消における類義語と辞書定義文を併用した語義表現の有効性」　言語処理学会第 23 回年次大会発表論文集, pp. 82-85.

新納浩幸・佐々木稔(2012)「外れ値検出手法を利用した新語義の検出」　自然言語処理 19 巻 5 号, pp. 304-327.

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

小木曽智信(2011)「通時コーパスの構築に向けた 古文用形態素解析辞書の開発」　情報処理学会研究報告, pp. 1-4.

Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig(2013). “Linguistic Regularities in Continuous Space Word Representations”, In Proceedings of NAACL 2013, pages 746–751..

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean(2013). “Efficient Estimation of Word Representations in Vector Space”, In Proceedings of ICLR Workshop 2013, pages 1–12..

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean(2013). “Distributed Representations of Words and Phrases and their Compositionality”, In Proceedings of NIPS 2013, pages 1–9..

国立国語研究所（2018）『日本語歴史コーパス』（バージョン 2018.3，中納言バージョン 2.4.2）https://chunagon.ninjal.ac.jp/ (2018 年 4 月 30 日確認)

国立国語研究所 (2017)　分類語彙表－増補改訂版データベース
http://pj.ninjal.ac.jp/corpus_center/goihyo.html (2018 年 1 月 9 日確認)

LIBLINEAR -- A Library for Large Linear Classification
https://www.csie.ntu.edu.tw/~cjlin/liblinear/ (2017 年 12 月 16 日確認)

MeCab: Yet Another Part-of-Speech and Morphological Analyzer(2017)
http://taku910.github.io/mecab/ (2017 年 12 月 16 日確認)

国立国語研究所(編) (2017) 『現代日本語書き言葉均衡コーパス(BCCWJ)』
http://pj.ninjal.ac.jp/corpus_center/bccwj/ (2017 年 12 月 26 日確認)

形態素解析器『Sudachi』のための大規模辞書開発

坂本 美保，川原 典子，久本 空海，髙岡 一馬，内田 佳孝
（株式会社ワークスアプリケーションズ　ワークス徳島人工知能 NLP 研究所）

Large Scale Dictionary Development for Sudachi

Miho Sakamoto, Noriko Kawahara, Sorami Hisamoto,
Kazuma Takaoka, Yoshitaka Uchida
(WAP Tokushima Laboratory of AI and NLP)

要旨

我々は，汎用的な日本語形態素解析器『Sudachi』とその辞書を開発した。本稿では，Sudachi の辞書開発内容について述べる。我々は，まず，UniDic をベースとして，見出し表記，品詞，各種パラメータ等，形態素解析をするための辞書情報を整えた。次に，実用上 UniDic に不足している語句を見出しとして追加した。これには，NEologd から取り込んだ膨大な固有名称も含まれる。さらに，登録見出しについて，アプリケーションが利用しやすい形態素単位の整備，表記のゆれを同一視するための正規化表記の整備等を行い，辞書内容を充実させた。また，形態素解析精度の向上のため，UniDic 由来の見出しについても，弊害となる見出しの抑制や間違いの修正，形態素単位の調整を行った。我々のこれまでの成果は，最新版の辞書ソースに反映し OSS として公開している。

１．はじめに

IT 技術の進展により，近年，産業界において日本語テキスト処理の利用機会はますます増えている。形態素解析はテキスト処理の重要な基盤技術であるが，自由に利用できて，かつ有用な形態素解析リソースは不足している[1]。

商用利用される形態素解析器としては，OSS として公開されている MeCab[2] (Kudo et al.,2004)，kuromoji[3]が大半を占めており，これらで利用可能な辞書としては，IPAdic (Asahara and Matsumoto, 2003)，NAIST Japanese Dictionary[4]，UniDic[5] (Den et al., 2007; Den et al., 2008)，NEologd[6] (Sato et al., 2016; Sato et al., 2017) などがある。しかし，IPAdic，NAIST Japanese Dictionary は，長年メンテナンスされていないため辞書内容が最新でない。

1 http://www.lrec-conf.org/proceedings/lrec2018/pdf/8884.pdf , pp. 1-2.
2 http://taku910.github.io/mecab/
3 https://www.atilika.com/ja/kuromoji/
4 https://ja.osdn.net/projects/naist-jdic/
5 http://unidic.ninjal.ac.jp/
6 https://github.com/neologd

また，UniDic，NEologd は，登録見出しの単位に特徴があり，用途によっては，そのままでは使いにくい。UniDic では，言語の形態論的側面に着目して規定された短単位[7]で見出し登録されている。そのため，たとえば語義を取り扱いたい場合や語彙調査をする場合にはそのままでは不足が生じる。一方，NEologd では，複数の短単位から成る固有表現が一塊で登録されているため，そのまま検索システムで利用すると再現率が低くなる等，支障がある[8]。

我々は，汎用的な辞書として使用できる大規模かつ高品質の辞書データの構築を目指す。

２．Sudachi 辞書の特長

２．１　豊富な語彙

『現代日本語書き言葉均衡コーパス』（BCCWJ）(Maekawa et al., 2014) の形態論情報をアノテーションするために開発された UniDic には，様々なジャンルの語句が斉一な単位で登録されている。見出し数は 75 万語を超える[9]。しかし，UniDic で規定するところの短単位で登録されているため，基本的だと感じる語句が見出し登録されていないことがある。たとえば，「小学校」や「自転車」など日常生活に密着した語句や，「太平洋」「東京都」などの地名，「集英社」「サランラップ」など認知度の高い固有名称も登録されていない。また，「ゆるキャラ」や「スマホ」など，新語への対応も不十分である。

そこで，我々は，新語や固有表現の収集を高頻度で行っている NEologd から，語句を大量に追加した。その際，NEologd の見出しの内部構造に含まれる複合語で，共通して他の見出しにも含まれる複合語（主に接辞付きの語句）についてもあわせて追加した[10]。
表 1 に例を示す。

表 1　NEologd からの見出し追加

NEologd の見出し	Sudachi に登録した見出し	
自転車シェアリング	自転車シェアリング	自転車
自転車通勤	自転車通勤	自転車
不動産登記	不動産登記	不動産
不動産鑑定士	不動産鑑定士	不動産，鑑定士
古民家鑑定士	古民家鑑定士	古民家，鑑定士
古民家再生協会	古民家再生協会	古民家

7　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf
8　http://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/B6-1.pdf
9　我々が採用したバージョンは，unidic-mecab-2.1.2_src.zip
（http://unidic.ninjal.ac.jp/back_number#unidic_cwj）である。
10　接辞付きの語句については，国語辞典から登録したものもある。

これにより，UniDic 由来の短単位語句に加え，NEologd 由来の膨大な固有名称，およびその内部構造である接辞付き語句等を補強し，辞書見出しを充実させることができた（付録 A 参照）。

２．２　３種類の形態素単位

上述のように，登録見出しについては膨大な量を確保できたが，次に，これらを用いた形態素解析結果としてどの長さで形態素認定すべきかを検討しなくてはならない。最適な形態素の長さは，アプリケーションによって異なるからである。たとえば，NEologd 由来の長い単位の語句は，固有表現抽出やテキストマイニングには有利だが，検索システムで使う場合には，短い単位でもインデキシングしないと再現率が低下する等，不都合な面がある。

この問題に対処するため，Sudachi (Takaoka et al., 2018)には，３種類の形態素単位（A 単位（短単位），B 単位（中単位），C 単位（NE 単位））が用意されている。

「A 単位」とは，ほぼ UniDic の短単位規定[11]と同じであるが，次項で述べるように，一部のものについて，さらに短くしたものを「A 単位」としている。「B 単位」とは，「A 単位」に接辞および漢字１文字の名詞[12]が結合したもの，および，複合動詞[13]である。「C 単位」とは，さらに多くの語句が結合したもので，複合名詞や固有名称，慣用句などがこれに相当する。各アプリケーションは，解析時に，これらの中から形態素単位を選択することができる。

我々は，３種類の形態素単位を提供するため，分割情報のアノテーションを行った。分割情報とは，上記３種類の形態素単位の規定に基づき，登録見出しの内部構造を記述し辞書に格納したものである。表 2 に，分割情報とそれに基づいて認定される形態素を示す。

表 2　分割情報と３種類の形態素解析結果

Sudachi の登録見出し	A 単位	B 単位	C 単位
選挙／管理／委員会	選挙｜管理｜委員｜会	選挙｜管理｜委員会	選挙管理委員会
委員／会	委員｜会	委員会	委員会
カンヌ／国際／映画祭	カンヌ｜国際｜映画｜祭	カンヌ｜国際｜映画祭	カンヌ国際映画祭
映画／祭	映画｜祭	映画祭	映画祭

”／”…分割情報として保持している見出しの内部構造の境界，　”｜”…形態素解析結果における形態素境界（以下，同記号を用いる）

11　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf
12　「～案（アン）」「～法（ホウ）」のように様々な名詞の下につく漢字１文字の名詞を指す。
13　「動詞+動詞」で構成されるもののみを「B 単位」とした。

２．２．１　UniDic 由来の見出しの再分割

UniDic では，「和語・漢語は、２最小単位の１次結合体を１短単位とする。」「外来語は、１最小単位を１短単位とする。」[14]という短単位規定に基づいて，見出し登録されている。

しかし和語については，漢語と異なり最小単位[15]が自立語として使用されることが多いため，最小単位で形態素認定した方が実用上都合がよいものがある。たとえば，「夏頃」は，UniDic では規定通り「短単位」だが，検索での利用を想定した場合，「夏」単独でも検索キーワードとして使われる可能性が高い。

そこで，我々は，UniDic 由来の見出しについて，次のようなものに分割情報を付与した。

a)　複合動詞

「動詞＋動詞」で構成される複合動詞については，語彙的複合動詞，統語的複合動詞の別を問わず，基本的に分割情報を付与し単独動詞を「A 単位」，複合動詞を「B 単位」とした（表 3）。

検索システムで利用する場合，たとえば「錆び付く」から「錆びる」が検索できない（あるいはその逆）のは，重大な不具合だからである。ただし，「見積もる」「仕舞う」のように，構成要素である単独動詞の意味が全く継承されていないものは，分割情報を付与せず，この長さを「A 単位」とした。

表 3　複合動詞の分割情報の例

錆び／付く	反り／返る	取り／出す
塗り／分ける	売り／渡す	譲り／受ける

b)　複合名詞

最小単位を用いた別の表現に容易に言い換えられるものや，最小単位に接辞が付いたものは，分割情報を付与し，各構成語を「A 単位」とした（表 4）。

表 4　複合名詞の分割情報の例 1

猫／探し	ゴミ／拾い	湯／洗い
紙／おむつ	仮／住まい	右／ふくらはぎ

14　http://unidic.ninjal.ac.jp/glossary#suw
15　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf , pp. 1-27.

また，表5に挙げるような見出しは，UniDicの短単位規定に基づいて斉一にアノテーションされた結果であるが，コーパス色が強いため，登録されていない他の類型表現と「A単位」での形態素解析結果を合わせるため，分割情報を付与し，各構成語を「A単位」とした。

表5　複合名詞の分割情報の例２

家鴨／柄	二人／組	迷子／牛
ユリ／科	母／山羊	海老／問屋

UniDic由来の見出しの再分割（分割情報付与）はごく一部しか行っていないが，和語見出しについては，まだ再分割する余地があると考えている。
たしかに語構成としては複数の最小単位から成り立っていても，「気持ち」や「靴下」の類まで再分割する必要がないのは明らかであり，UniDicの短単位は，「基準の分かりやすさ，ゆれの少なさという条件を満たしつつ，用例を収集して分析を行うという利用目的にもかなう単位」[16]であるといえる。これに手を入れるということは，基準の不明瞭さやゆれを生み，外部から見て，辞書の開発方針がわかりにくくなる可能性はある。
しかし，我々が分割情報を付与した複合動詞や複合名詞のような語群が，これまでUniDicが検索システム等で使いにくかった一因であることは確かである。UniDic由来の見出しを再分割するにあたって，我々は，できる限りわかりやすい基準を追究している。

２．２．２　NEologd由来の語句の短単位化

NEologd由来の語句は長単位のものが多いため，基本的にすべて分割情報を付与することとしている。ただし，量が膨大なため，まず機械的に分割情報を付与し，それを人手でチェックするという手順をとった。すなわち，NEologd由来の語句を登録せずに作成したSudachi辞書を用いて，NEologd由来の語句を形態素解析し，その形態素解析結果を仮の分割情報とした。これを一つ一つ確認し，間違っていれば修正する，とした。

また，形態素解析結果としては間違っていなくても，我々が付与したい段階的な分割情報でない場合がある。たとえば，「医療費控除」の形態素解析結果が「医療｜費｜控除」の場合，これは正しい形態素解析結果であるが，段階的な形態素単位を提供するためには，「医療費／控除」という分割情報を付与しておく必要がある。さらに「医療／費」の見出し登録も必要である。
こうした確認作業をしながら，分割情報の精度を高めていっている（付録A参照）。

16　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-01.pdf , p.10.

２．３　正規化表記

日本語には，ひらがな，カタカナ，漢字，ローマ字の文字種があり，さらに漢字については，送り仮名，異体字，代用漢字の選択があるなど，非常に複雑な表記体系となっている。たとえば，「アキカン」には，「空き缶，空缶，空き罐，空罐，空きカン，空きかん」等の表記が可能である。

また，外来語の表記については，カタカナで表記する場合，語末の長音のあるなしに加え，原音や原綴りになるべく近く書き表そうとする場合に生じるゆれや，「アフィリエイト（正）ーアフェリエイト（誤）」等の誤用によるゆれもある。また，カタカナで表記せず原綴りをそのまま使用する場合もあり，表記法に多くのゆれが存在する。

テキスト処理において，これら表記のゆれを同一視できるかどうかは，タスクの精度に影響する重要な要素である。我々は，全ての登録見出しに正規化表記の情報を付与し，表記ゆれを同一視できるようにしている。

表６　正規化表記の例

登録見出し	正規化表記
空き缶	空き缶
空缶	空き缶
空き罐	空き缶
空罐	空き缶
空きカン	空き缶
空きかん	空き缶
美術館	美術館
団体戦	団体戦

表6に示すように，それぞれの登録見出しについて，１つの正規化表記の情報を付与している。正規化表記が見出し表記と同じものは，それ自身が正規化表記である。

正規化表記の情報付与に際しては，次の２点が問題となる。一つは，何々を表記ゆれとみなすか，そしてもう一つは，どの表記を正規化表記とするかである。

a）表記ゆれの範囲

我々は，次のようなもの，およびその組み合わせパターンを表記ゆれとしている。

表 7　表記ゆれの範囲

パターン	例
文字種の違い	向日葵-ひまわり-ヒマワリ，燐酸-りん酸-リン酸
漢字の違い （異体字,代用表記,慣用表記）	芸術-藝術，驚歎-驚嘆，徳用-得用
送り仮名の違い	受け付け-受付け-受付
外来語の表記違い	コミュニティー-コミュニティ-コミニティー-community
誤用	シミュレーション-シュミレーション
くだけた言い方	～ちゃあ-～ては

UniDic 由来の見出しについては，同じ語彙素を持つ見出しグループを表記ゆれと見なした。ただし，UniDic では，やや広めに表記ゆれを吸収している部分があったため，表 8 に示すように，新聞等でも書き分けが見られるものは，同じ正規化表記を付与せず，別々の語句とした。

表 8　UniDic の語彙素と異なる正規化表記の採用例

登録見出し	UniDic の語彙素	Sudachi 正規化表記
炊く	焚く	炊く
焚く	焚く	焚く
卸す	下ろす	卸す
下ろす	下ろす	下ろす

また，ひらがな表記については，複数の語句の可能性があるものは，強引にどれかに正規化することはしていない（表 9）。

表 9　多義性のあるひらがな表記の例

登録見出し	UniDic の語彙素	Sudachi 正規化表記
そば	側	そば
そば	岨	そば
そば	蕎麦	そば

b) 正規化表記の選定

正規化表記は表記ゆれを同一視するための情報であり，いわゆる正書法的な観点から表記の選定は行っていない。そのため以下のケースがある。

① よく使われる表記が必ずしも正規化表記でない場合がある。
例）うどん → 饂飩（正規化表記）

② 同じ構成語を含む複合語について，統一的な正規化表記が付与されていない場合がある。
例）イヤフォン → イヤホン（正規化表記）
スマートホン → スマートフォン（正規化表記）

③ 分割情報によって短い単位に分割された場合，各構成語は，元の複合語と異なる正規化表記となる場合がある（表 10）。

表 10 形態素単位により異なる正規化表記の例

登録見出し	「C 単位」の正規化表記	「A 単位」or「B 単位」の正規化表記
取扱説明書	取扱説明書	取り扱い｜説明書
取り扱い説明書	取扱説明書	取り扱い｜説明書
ごまドレッシング	胡麻ドレッシング	ごま｜ドレッシング
胡麻ドレッシング	胡麻ドレッシング	胡麻｜ドレッシング

①，②については，表記ゆれの同一視という観点からは整合しているため，問題はないと認識している。また，③については，検索システムで使う場合，「C 単位」と「A 単位」を併用してもらうことにより精度は確保できると考えている。

３．辞書の精度

３．１ 弊害語の抑制

UniDic 由来の見出しは基本的にすべて採用しているが，形態素解析に弊害となる可能性のあるものは登録を保留とした（付録 A 参照）。たとえば，２文字のカタカナ・ひらがな表記で品詞が記号のものや，１文字の漢字表記で品詞が記号のもの，複合型の数詞等である。これらは，辞書に未登録の語句がブツ切れに形態素解析される弊害や，数詞処理の弊害となるおそれがあるため，登録を見合わせた。

例） アー,記号,一般,*,*,*,*,アー　　うら,記号,一般,*,*,*,*,ウラ
垈,記号,一般,*,*,*,*,ヤ　　六十,名詞,数詞,*,*,*,*,ロクジッ

また，間違いについては適宜修正を行った[17]。

例） 油染みる,動詞,一般,*,*,下一段-マ行,終止形-一般,アブラジミル

→ 「動詞,一般,*,*,上一段-マ行,終止形-一般」（品詞修正）

圧し遺る,動詞,一般,*,*,五段-ラ行,終止形-一般,オシヤル

→ 「圧し遣る」（表記修正）

３．２ 単語コスト

Sudachi が形態素解析をするためのパラメータは，UniDic の単語コストや連接コストを利用している。UniDic 由来の見出しはそのまま値を継承すればよいが，NEologd 由来の語句や内省で追加した語句については，なんらかの方法で単語コストを与える必要がある。そこで，新規追加見出しについては，次のような推定コスト値を付与するとしている。

- 分割情報により内部構造が記述されている見出しについては，それぞれの構成語の単語コスト（から一定の値を引いたもの[18]）および構成語間の連接コストの和を全体の複合語の単語コストとする。
- 分割情報を付与されていない見出しについては，字種や文字数により，特定の値を付与する[19]。

図 1 に例を示す。

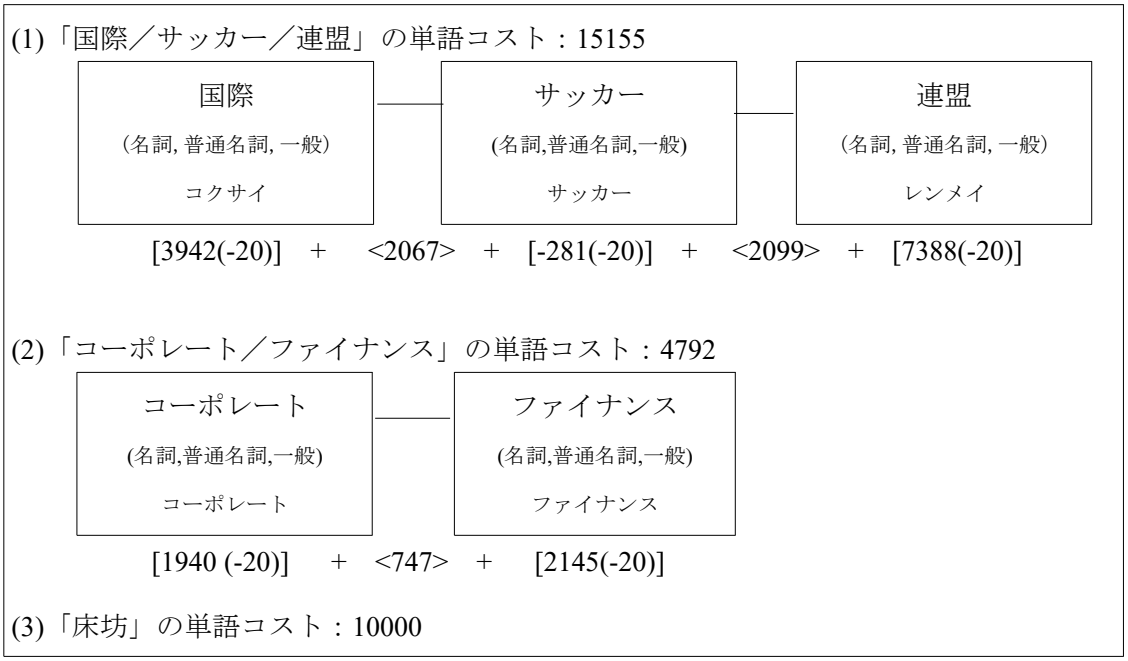


図 1 単語コストの推定 1 [ ] … 単語コスト値 < > … 連接コスト値

17 我々が採用したバージョンは，unidic-mecab-2.1.2_src.zip（http://unidic.ninjal.ac.jp/back_number#unidic_cwj）である。

18 現在は，-20 を適用している。

19 現在は，見出し表記が，アルファベットとスペースのみで構成される場合は"5000"，顔文字は"5000"，記号は"22000"，記号以外で３文字以下の表記は"10000"，記号以外で４文字以上の表記は"15000"，としている。

概ね，この推定コスト値で正しい形態素解析結果を得られているが，新規追加見出しに付与した推定コスト値より，他の登録語の方がコストが低いために，新規追加見出しが誤解析となる場合がある（図2）。

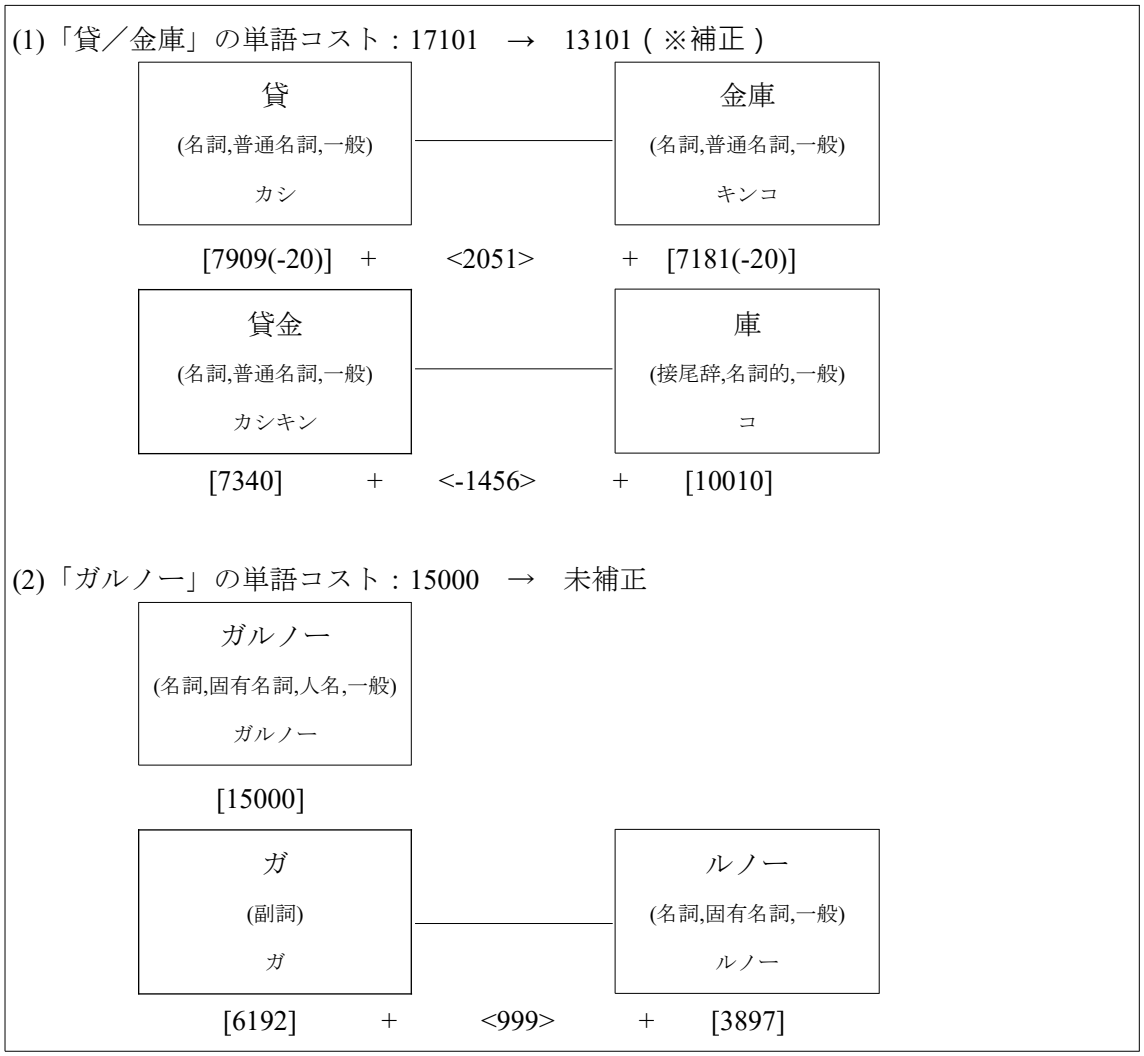


図2 単語コストの推定２ [ ] … 単語コスト値 <> … 連接コスト値

「貸金庫」の例では，「貸｜金庫」の合計コストより「貸金｜庫」の合計コストの方が低いため，見出しの内部構造に基づく推定コスト値 (17101) では，「貸金庫」は形態素解析の第一解にならず，「貸金（カシキン）｜庫（コ）」が第一解となる。これを改修するため，現在は，手動で「貸金庫」の単語コストを適当な値に補正している。
また，「ガルノー」の例では，文字数と文字種により特定の単語コスト値 (15000) を付与しているが，「ガ（副詞）｜ルノー（名詞,固有名詞,一般）」の合計コストの方が低いため，「ガルノー」は形態素解析第一解にならない。

このように，新規追加見出しに付与する単語コストの推定方法には，現状の手法では，根本的な問題があることを認識している。すなわち，追加しようとしている見出しが未登録の状態で，その見出しの内部構造として記述した短単位で正しい形態素解析結果が得られる，あるいは，内部構造によらず特定の値を付与した見出しについては，当該文字列について，他の登録語による形態素解析結果の方がコストが高い，という前提に立っているからである。
新規追加見出しに付与する単語コストの推定方法については，現在，抜本的な見直しを検討中である。

４．メンテナンスの継続

今後も，NEologd から定期的に新語を取り込み辞書の最新性を確保するとともに，機械的なチェック，人手によるチェックを併用しながら辞書内容を拡充，洗練していく。また，既登録語についても，正規化表記や分割情報の付与漏れ，品詞や読みの間違い等，点検が必要な見出しが半数近く存在する。これらの修正も継続して行う。その成果は，随時，OSS として公開していく[20]。”長期にわたる継続的なメンテナンス”[21]は，我々の開発方針の一つである。

５．おわりに

我々は，UniDic をベースに，NEologd から大量の固有名称を登録し，大規模な辞書データを構築した。280 万語を超える登録規模となり，付加情報の整備も着実に進んでいる。

「C 単位」を使えば，「調布の味の素スタジアム」という文字列は，「調布(名詞,固有名詞,地名,一般)｜の(格助詞)｜味の素スタジアム(名詞,固有名詞,一般)」と解析できるし，「ロミオとジュリエットを上演」は，「ロミオとジュリエット(名詞,固有名詞,一般)｜を(格助詞)｜上演(名詞,普通名詞,サ変可能)」と解析できる。形態素解析レベルで，「調布の → 味」や「ロミオと → 上演」のような間違った係り受けの可能性を排除できることが期待できる。しかし，文の構造を解析する上で重要な要素である「として」「にもかかわらず」のような複合辞については，未登録である。また，「役に立つ」「年をとる」のような成句についても，これらを一塊で認識できるデータはない。今後は，こうした連語のデータ構築も進めていきたい。

また，ベースとする UniDic を新しいバージョンに差し替えることも検討中である。さらに，同表記で読みが異なる語句の曖昧性解消や，新規に見出しを追加する際の単語コストの最適化についても課題である。

20 https://github.com/WorksApplications/Sudachi
21 http://www.lrec-conf.org/proceedings/lrec2018/pdf/8884.pdf , p.2.

謝　辞

Sudachi の辞書開発にあたっては，UniDic，NEologd から多くの研究成果を継承している。UniDic の開発に尽力された方々，そして，LINE 株式会社 佐藤敏紀氏をはじめ，NEologd の開発関係者の皆様に感謝申し上げる。また，奈良先端科学技術大学院大学 情報科学研究科教授の松本裕治氏には，Sudachi の形態素単位や正規化表記について，数回にわたる広範な議論の中で適切な助言をいただいた。深くお礼申し上げる。

文　献

工藤拓．Mecab: Yet another part-of-speech and morphological analyzer.

浅原正幸，松本裕治 (2003)『ipadic version 2.7.0 ユーザーズマニュアル』
奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座

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

伝康晴，山田篤，小椋秀樹，小磯花絵，小木曽智信 (2008)
『UniDic version 1.3.9 ユーザーズマニュアル』

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, Yasuharu Den (2014)
"Balanced corpus of contemporary written Japanese." Language Resources and Evaluation, 48:345-371

小椋秀樹，小磯花絵，冨士池優美，宮内佐夜香，小西光，原裕 (2011)『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）』』文部科学省科学研究費特定領域研究「日本語コーパス」データ班

佐藤敏紀，橋本泰一，奥村学 (2016)『単語分かち書き用辞書生成システム NEologd の運用 ―文書分類を例にして ―』情報処理学会 研究報告自然言語処理 2016-NL-229-15

佐藤敏紀，橋本泰一，奥村学 (2017)『単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討』言語処理学会 第 23 回年次大会 発表論文集, pp.875-878

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, Yuji Matsumoto (2018) “Sudachi: a Japanese Tokenizer for Business”

付　録　A

最新版辞書(2018 年 7 月版)の登録見出し数（うち，精査済み数）	2,801,739　(1,339,630)
分割情報を付与した見出し数（うち，UniDic 由来の見出し数）	1,676,735　(　284,335)
UniDic 由来の語句で登録保留としたもの	1,648

英語における前置詞句についての音響分析

于　暁陽（九州大学芸術工学府）
中島　祥好（九州大学芸術工学研究院）
張　一新（九州大学芸術工学府）
岸田　拓也（九州大学芸術工学研究院）
上田　和夫（九州大学芸術工学研究院）

An Acoustic Analysis of Prepositional Phrases in English

Xiaoyang Yu (Graduate School of Design, Kyushu University)
Yoshitaka Nakajima (Dept. Human Science/Research Center for Applied Perceptual Science, Kyushu University)
Yixin Zhang (Graduate School of Design, Kyushu University)
Takuya Kishida (Department of Human Science, Kyushu University)
Kazuo Ueda (Dept. Human Science/Research Center for Applied Perceptual Science, Kyushu University)

要旨

英語における前置詞句のアクセントは、非英語母語話者にとって、学習が難しい点の一つとなる。本研究では、Kishida et al.（2016）が提案した起点移動因子分析という分析手段を用い、前置詞と直後の名詞又は名詞句からなる前置詞句に着目して、前置詞句の音響的特徴および知覚的な役割を調べることを目的とする。前置詞句の役割を明らかにするために、英語母語話者三名（男1名、女2名）が発話したイギリス英語音声データベースを構築し、研究を行った。収録された音声に対し、音素ごとにラベルを付けた。前置詞句を含む対象音声を20臨界帯域に分割して、因子分析を行った。そこから、3因子を抽出した。3300 Hzを超える周波数範囲と密接に関連するhigh factorの因子得点は、前置詞よりも名詞句の方が明らかに高いという結果が得られた。しかし、これ以外の因子得点については、明確な差が出なかった。以上の分析により、3300 Hz以上のhigh factorは名詞句を知覚する際、重要な役割を果たしていることを示唆している。

1. はじめに

英語の前置詞句は、前置詞と直後の（広義の）名詞句からなり、研究に用いた例を挙げてみると、“ John could lend him the latest draft of his work. ”のように、前置詞句における前置詞は“of”、名詞句は“his work”になる。自然な前置詞句アクセントを学習するのは非英語母語話者にとって、難しい点の一つとなる。前置詞は通常、場所・時間・方向などを表し、文章の意味において重要な役割を果している。

冬野 ら（2013）は、心理動詞受動文における前置詞の使い分けについてコーパス調査を行い、日本人英語学習者は受動文として心理動詞を使わない傾向があることを明らかにした。また、望月ら（2016）は、東京外国語大学英語上級学習者コーパスにおける前置詞の誤用類型について調べ、母語移転による影響が強いということを示した。このように、英語の前置詞の使い分け、誤用などについての研究が数多くなされている。しかし、筆者らの把握する範囲では、音響的な観点で英語の前置詞、および前置詞句について行われた研究は見当たらない。前置詞句の音響的な特徴を明らかにすれば、外国語の習得、音声合成など、様々な分野に応用できると考えられる。

2. 研究方法

本研究では、英語における前置詞句の音響的な特徴を調べるために、一つの音節からなる前置詞を含む前置詞句を分析対象とし、Kishida et al. (2016)が提案した起点移動主成分分析によって、英語の前置詞句の音響的特徴、および知覚的な役割を調べる。

Zwicker and Terhardt (1980)は、聴覚系末梢が臨界帯域と呼ばれる狭い周波数帯域(臨界帯域)に分けられるとのモデルを提案した。Ueda and Nakajima (2017)は、音声のパワースペクトルの時間変動を 20 帯域で分割し、1 ms毎にパワー値に対して因子分析を行った。そこで、三つないし四つの因子を取り出して、8つの言語において、共通するパターンをもった因子が現れることを発見した。これは、言語の違いを超える普遍的な音響的特徴が音声に含まれるということを示す。また、Kishida et al. (2016)は、主成分分析を行う際に、分散を計算する起点を無音点に置いて、更に倍音構造の影響を減らすために、スペクトルを平滑化し、分析方法を改良した。そこで、因子数を1から3に増す際に、この因子を用いた雑音駆動音声の明瞭度は 70%まで増加した。これにより、3因子で音声のコミュニケーションについて、研究を行うことが出来ることを示している。

本研究は、Kishida et al. (2016)が改良した起点移動主成分分析を用い、データベースにより男女三人のうちの男性一人の 100 文の音声を 20 臨界帯域に分割し、引き続いて、因子分析を行って、三つの因子を抽出した(Fig.1)。本研究では、300 Hz および 2200 Hz 付近の二つの周波数範囲で大きい因子負荷量をもつ因子を low & mid-high factor と、1100 Hz 付近の周波数範囲において大きい因子負荷量をもつ因子を mid-low factor と、3300 Hz を超える周波数範囲において大きい因子負荷量をもつ因子を high factor と呼ぶこととする。

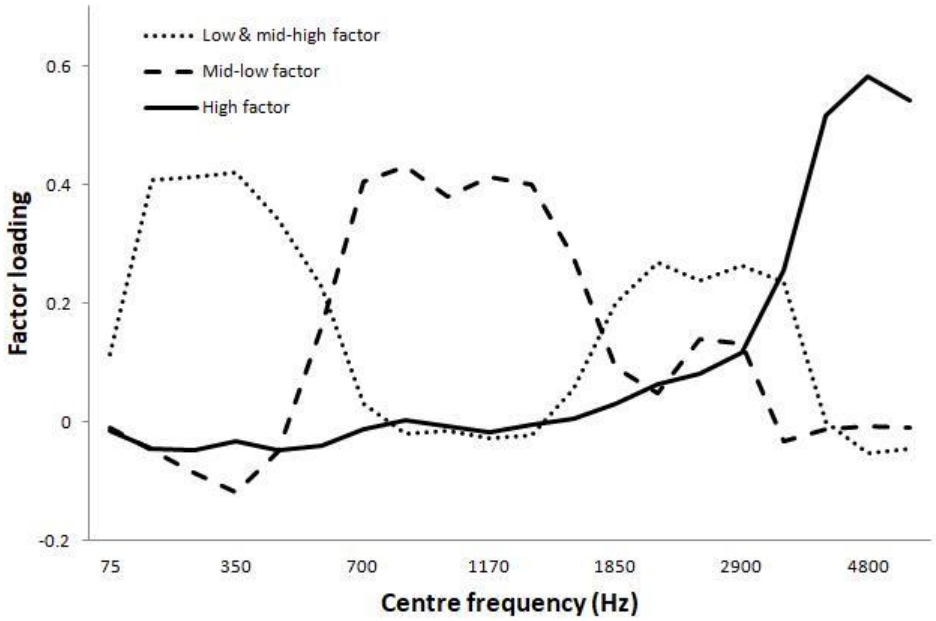


Fig.1 Factor loadings plotted against the center frequency of critical bands.

前置詞句の因子得点を調べるために、分析対象である文の全ての音素にラベルを付けた。これにより、音声波形のどの部分がどの音素に対応するのかが示される。前置詞句における前置詞と名詞句の因子得点の分布を調べるために、ラベルを付けられた前置詞における各音素の時間的中央点の因子得点の平均値と、名詞句における各音素の時間的中央点の因子得点の平均値で

分布図を描いた Fig. 2 (A, B, C, D)。

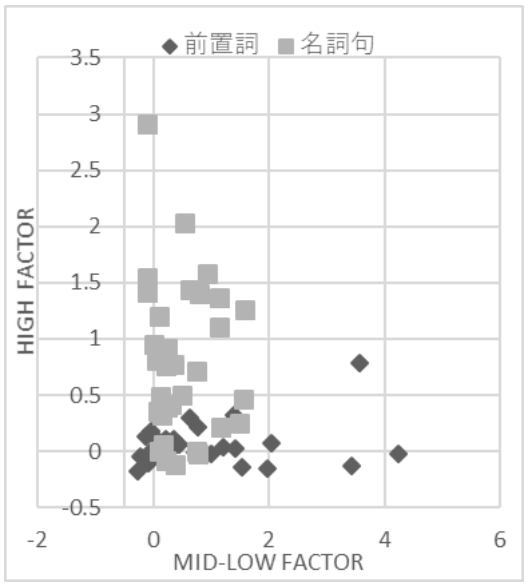


Fig. 2(A)

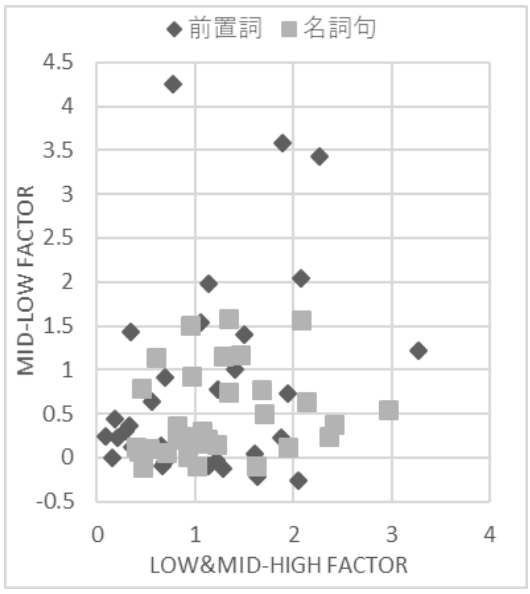


Fig. 2(B)

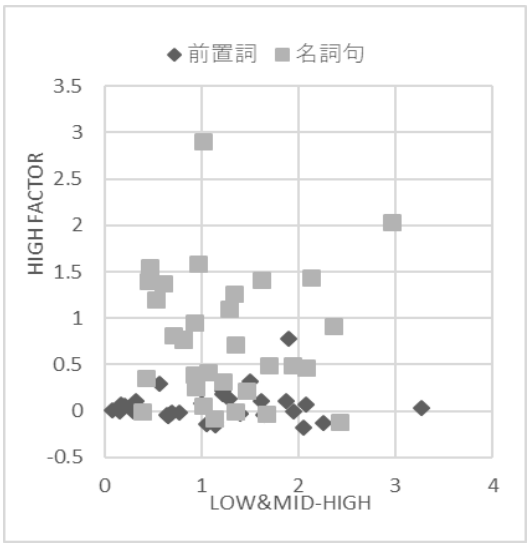


Fig. 2(C)

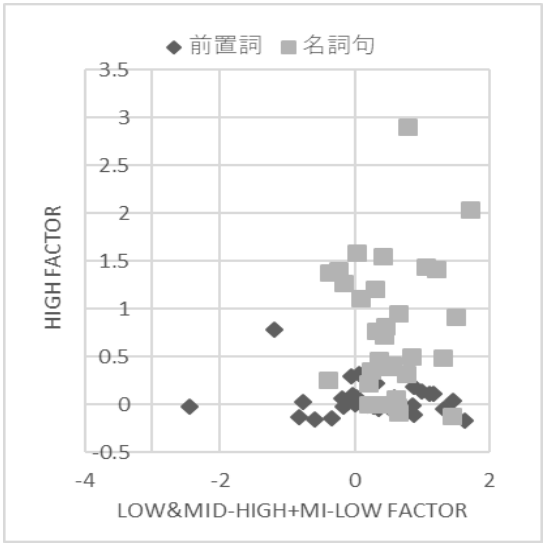


Fig. 2(D)

Fig. 2 (A, B, C, D) Distribution of the prepositions and the noun phrases in the three-dimensional factor space.

続いて、文ごとの前置詞から名詞句までの方向を矢印で描いた Fig.3 (a, b, c, d)。

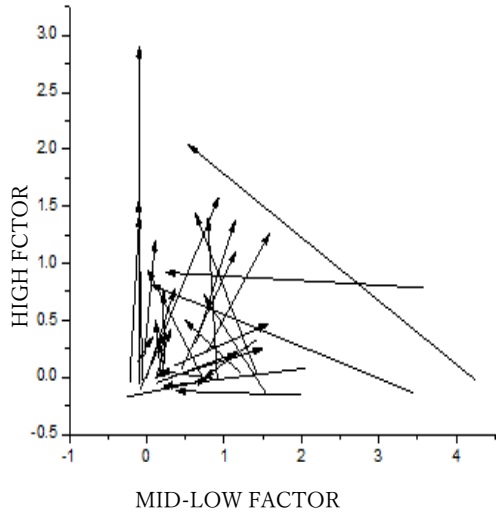


Fig.3(a)

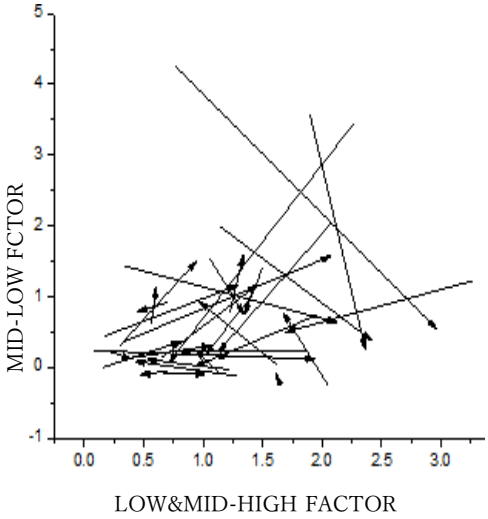


Fig.3(b)

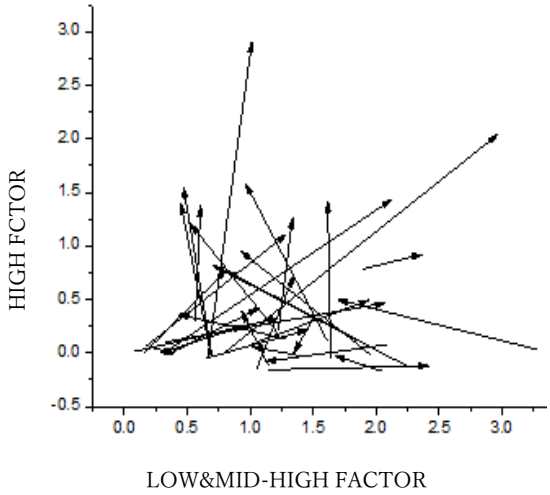


Fig.3(c)

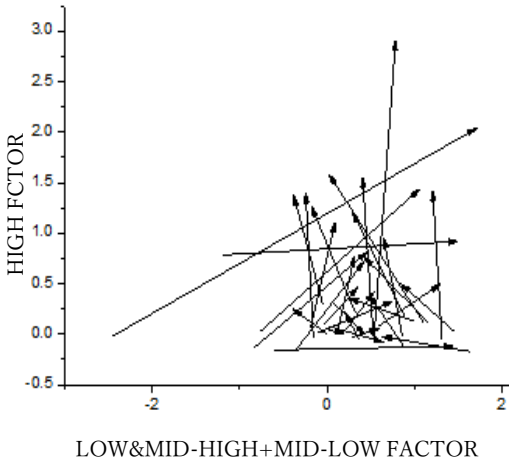


Fig.3(d)

Fig. 3 (a, b, c, d) Tendency from preposition to noun phrase in every sentences.

3. 結果および考察

Fig. 2. A, C, D の前置詞と名詞句の因子得点が high factor において明らかに分散しているが、Fig. 2. B において、前置詞と名詞句の因子得点が low & mid-high factor と mid-low factor において、重なっているため、関連性を得るのが難しかった。そして、分布図 Fig. 2 (A, B, C, D)と同様に、矢印図 Fig. 3 (a, b, c, d) において、Fig. 3.b より、Fig. 3.a, c, d の方が high factor において上を指している傾向があることが分かった。更に、これを検証するために、符号検定（sign test）を行った。100 個の文における前置詞句を含む 32 対の各因子得点について、low & mid-high factor、mid-low factor、および low & mid-high factor + mid-low factor の組み合わせには、有意差が現れなかったが、high factor には有意差が現れた。これは、三因子のうちの high factor には、前置詞におけ

る各音素の時間的中央点の因子得点と、名詞句における各音素の時間的中央点の因子得点には統計的に有意な差があることを示している。従って、3300 Hz 以上の high factor は名詞句を知覚する際、重要な役割を果たしていること可能性が高い。

4. 展開

以上得られた結果を更に検証し、名詞句に含まれる音素の種類による影響を調べるために、前置詞の音素を含む名詞句、又は前置詞に似た発音がある名詞句を用いて文を作り、英語母語話者発音してもらうような実験を計画している。

文の例を挙げると、

The first letter of ‘of ’is o.

They seem to be interested in industrial application.

These lectures were given by bilingual speakers.

追加実験により、前置詞句の音響的な特徴を明らかにすることが期待できる。

謝辞

本研究は、科学研究費補助金（17H06197）の助成を受けた。

文献

Kishida, T., Nakajima, Y., Ueda, K., & Remijn, G. B. (2016). Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech. Frontiers in Psychology, 7:517.

冬野 美晴, 川瀬 義清, 心理動詞受動文における前置詞の使い分けに関するコーパス調査―英語母語話者の用法から見えてくるもの― (2013), 外国語教育メディア学会九州沖縄紀要, 13,.

望月 圭子, ローレンス ニューベリーペイトン, モチヅキ ケイコ他 (2016), 東京外国語大学英語上級学習者コーパス』における前置詞の誤用類型:―日本語母語話者・中国語母語話者英作文の対照, 日本語学習者の母語・地域性をふまえた日本語教育研究 (2), 25-42,.

Nakajima, Y., Ueda, K., Fujimaru, S., Motomura, H., and Ohsaka, Y. (2017). English phonology and an acoustic language universal. Scientific Reports, 7: 46049.

Nakajima, Y., Ueda, K., Remijn, G. B., Yamashita, Y., Kishida, T. (2018), How sonority appears in speech analyses, Acoustical Society & Technology, 39, 179-181.

Zwicker, E., and Terhardt, E. (1980). Analytical expressions for critical-band rate and critical-bandwidth as a function of frequency. Journal of the Acoustical Society of America, 68, 1523-1525.

Ueda, K., and Nakajima, Y. (2017). An acoustic key to eight languages/dialects: Factor analysis of critical-band-filtered speech. Scientific Reports, 7:42468.

副詞の程度性の下位分類の試み
―「あまり・そんなに・それほど・たいして」を例に―

劉 時珍(東京エアトラベル・ホテル専門学校非常勤)†

A trial for the subcategorization of adverbs of degree: Taking ‘amari’, ‘son-nani’, ‘sorehodo’ and ‘taishite’ as its examples

Shizhen Liu (Technos College Tokyo)

要旨

本発表では否定と呼応する 4 つの類義副詞「あまり」「そんなに」「それほど」「たいして」（以下を「4 つの副詞」と呼ぶ）を例に、副詞の程度性について、『現代日本語書き言葉均衡コーパス』（以下を「BCCWJ」と呼ぶ）の調査結果に基づき、先行研究を踏まえた上で、副詞の程度性の下位分類について考察した。

研究方法としては BCCWJ を用い、4 つの副詞を、「たいして」（全数 700 例）を除き 1000 例ずつ無作為抽出し、それぞれの後ろの係り先に注目し、特に形容詞の評価性（『現代形容詞用法辞典』）に基づいて、4 つの副詞の正の評価性と負の評価性[1]を点数化した。

その結果、「あまり」と「たいして」は負の評価性がより強く、「そんなに」は正の評価性がより強く、「それほど」は両方ほぼ同じという結果になった。

以上のことから、「あまり」は負の評価性をより用いるものの、肯定形の「あんまり」という過度の否定により、程度性の部分性と評価性を両方持つことにより数多く使われ、一方、「そんなに」の正の評価性がより強いのは極限の程度を否定するだけの副詞であるためと結論づけた。程度性の下位分類としては「あまり」と「たいして」が同一の、「そんなに」と「それほど」が同一の下位分類に属することが考えられる。

１．はじめに

否定と呼応する類義表現とされる「あまり」「そんなに」「それほど」「たいして」については、今まで、「そんなに～ない」と「あまり～ない」はどちらも「否定を伴って程度の甚だしくないことを表す」とされている（益岡・田窪 1992:142）。この 2 つの副詞の違いに関する先行研究に服部（1994）と小川（2008）があり、記述文法の立場から論述されている。

本発表では、今までの知見を受け入れ、この 4 つの副詞を『現代日本語書き言葉均衡コーパス（BCCWJ）』の中から「たいして」以外の 3 つの副詞をそれぞれ 1000 例無作為抽出し（「たいして」は 1000 例未満なので、700 例を全数調査した）、その上で、4 つの副詞の肯否のジャンルの分布、及び、それぞれの否定形の係り先の形容詞の違いを調べた。

２．ジャンル及び形容詞述語の違い

２.１　BCCWJ における 4 つの副詞のジャンルの分布

BCCWJ の中のデータは主に次の 8 つの分野、『白書』、『雑誌』、『新聞』、『書籍』、『教科書』、『Yahoo！ブログ』、『国会会議録』、『Yahoo！知恵袋』に分け、収集されている。

† liusz77@hotmail.com

1 本発表の「評価性」とはニュアンスという意味を指す。「正の評価性」はプラスのニュアンスであり、「負の評価性」はマイナスのニュアンスである。

本発表では、書籍の中に、フィクションとノンフィクションという大きな境を持ち、文体差の大きいストーリーの描写文（地の文）と会話文からなる文学を一つ独立のジャンル『書籍（文学)』として立て、文学以外の書籍、及び雑誌、新聞、教科書は『書籍（文学以外)』に合併した。その上で、対人的か否かを基準にジャンルを 6 つに分け、『白書』『書籍（文学以外)』『書籍（文学)』『Yahoo！ブログ』(以下『ブログ』と略す)『国会会議録』『Yahoo！知恵袋』(以下『知恵袋』と略す）の順に並べる。この 6 つのジャンルを基準に 4 つの副詞の出現頻度をまとめた結果は表 1 である。

表 1 の結果を見ると、「あまり」の肯定はどのジャンルでも顕著に多いことはなく、『ブログ』で有意に少ない。「あまり」の否定は多い順から『知恵袋』・『ブログ』・『国会会議録』に使われることが分かった。「そんなに」の肯定は『書籍（文学)』が最も多く使われ、次に『知恵袋』に有意に多い結果になっている。「そんなに」の否定は数の違いがあるが、傾向として全く「あまり」と同じ分布になっていることに興味を惹かれる。「それほど」は肯否とも『書籍（文学以外)』に有意に多く、特に否定形は出現回数が多い。「たいして」は『書籍（文学)』に集中していることが分かった。もう一つは 4 つとも白書の用例が少なく、「それほど」が一番多くても 18 例で、次に「あまり」の否定の 11 例である。

表１ 4つの副詞のジャンルの分布

ジャンルの分布	あまり 肯定(115)	あまり 否定(820)	そんなに 肯定(433)	そんなに 否定(560)	それほど 肯定(206)	それほど 否定(706)	たいして 否定(700)	合計
1. 白書	0	11	0	0	0	18	0	29
564万語	0	2	0	0	0	3.2	0	5.2
2. 書籍（文学以外）	35	301	109▽	152▽	113▲	349▲	219▽	1278
6330万語	0.6	4.8	1.7	2.4	1.8	5.5	3.5	16.8
3. 書籍（文学）	41	137▽	174▲	130▽	69	175	263▲	989
2457万語	1.7	5.6	7.1	5.3	2.8	7.1	10.7	29.6
4. ブログ	6▽	122▲	41	99▲	15▽	68▽	79	430
1298万語	0.5	9.4	3.2	7.6	1.2	5.2	6.1	27.1
5. 国会会議録	7	41▲	9	31▲	5	24	10▽	127
560万語	1.3	7.3	1.6	5.5	0.9	4.3	1.8	20.9
6. 知恵袋	26	208▲	100▲	148▲	4▽	72▽	129	687
1202万語	2.2	17.3	8.3	12.3	0.3	6.0	10.7	46.4
総計	115	820	433	560	206	706	700	2840
1億2410万語	6.3	46.4	21.9	33.1	7.0	31.3	32.8	146.0
否定率	87.70%		56.39%		77.41%		99.86%	

$x^2(24)= 337.96$, $p<.01$, Cramer's V = 0.155

「▲」は残差分析の結果、有意に多いもの。「▽」は有意に少ないものをそれぞれを表す。

*各レジスターの全語数で、その行の数字は 100 万語換算した語数である。

表 1 の結果から一つの疑問が浮かんでくる。「あまり」と「そんなに」の否定形のジャンルの分布の傾向はかなり重なっており、「それほど」と「たいして」の否定形は同じ傾向を示している。類似性が高いと言えばそうかもしれないのだが、別の観点からこの 2 つずつの副詞は似ていながらそれぞれの役割を果たしていることが考えられる。

２.２　係り先の形容詞の違い

それで、4 つの副詞の係り先の形容詞について調べた。ジャンルの分布の結果から、見やすいという便宜上からも、「あまり」と「そんなに」の後ろの形容詞述語の違いと、同じ『書籍』の中に集中している「それほど」と「たいして」の形容詞を並べて、表 2、表 3 に結果をまとめる。

表 2，3 の結果から、「あまり」は 820 例の否定形の中に 88 例の形容詞述語があるのに対して、「そんなに」は 560 例の中に 136 例の形容詞述語があることが分かった。

次に、その係り先の形容詞述語の評価性について、『現代形容詞用法辞典』（飛田・浅田 1991）所収の『現代形容詞イメージ一覧』を参考にする。

『現代形容詞用法辞典』には形容詞の評価性のリストがついており、形容詞を「＋＋＋・＋＋・＋・０・－・－－・－－－」の 7 段階に評価性を分けている。本稿では、その 7 段階を相対得点にし、「－－－」を 1 点にし、「＋＋＋」を 7 点になるようにし、1 点間隔で換算した。なお、用法ごとに複数の得点が与えられている場合は、その算術平均（小数点第 2 位を四捨五入）を用いた。このようにして計算した結果は表 2，3 の通りになる。

表 2 の結果から分かるように、「あまり」の形容詞述語はプラスの語が多く（異なり 15 語：のべ 58 語）、マイナスの語が少ない（異なり 7 語：のべ 19 語）。「ゼロ」語も少ない（異なり 4 語：のべ 11 語）。平均点は 4.8 点で、ポジティブなニュアンスという結果である。実際の文は否定文なので、逆にネガティブなニュアンスになる。

(1)食物連鎖の上で問題がないかどうか、これが新しい課題になっています。もっとも、新聞などにはあまり好ましくない材料かもしれませんがね」このように、公的な規制のチェックを受けて進められている研究で。。。　【出典】 BCCWJ サンプル ID：LBo9_00165

(2)「私が思い描く国家は、単純で強固だが、統治しやすい行政組織を持つだろう。あまり複雑ではないバネの力で大きな効果をあげる巨大な機械に似たものになるだろう」
【出典】 BCCWJ サンプル ID：LBp1_00031

(3)疑問だらけの世界です。ｈｔｔｐ：／／ｍｅｅｔｉｎｇｐｏｉｎｔ。ｊｐ／Ｈｏｐｐｙｓｏｆｔ／Ｒ０３；　更にオンラインで変換してくれるサービスもあるようですが、変換する機会があまり多くないのであれば　こちらが手軽でいいかもしれませんって、これ使えるのかしら？？？？　【出典】 BCCWJ サンプル ID：OY04_03622

一方、表 2 の「そんなに」の結果を見ると、マイナスの語が多く（異なり 22 語：のべ 65 語）、プラスの語がやや少なく（異なり 14 語：のべ 41 語）、「ゼロ」語は「あまり」より多い（異なり 9：のべ 30 語）。平均点は 3.9 点である。その結果、否定文になると、ポジティブかニュートラルのニュアンスを感じる文が多い。

(4)ちなみにレベルはＧＯ！ＧＯ！７１８８の「Ｃ７」がちょっと難しい、という程度です（汗）。プリンセスプリンセスには聴かせる、のれるなどいろんな曲があるし、レベル的にもそんなに難しくないので初心者むけだと思います。
【出典】 BCCWJ サンプル ID：OC01_02474

(5)行く経済的余裕もあまりないので、今は独学で勉強を進めています。ただ悩みなのが、数学がそんなに得意ではないということです（大好きなのですが）。とりあえず高 1 の文理選択まで猛勉強して、自分が。。。　【出典】 BCCWJ サンプル ID：OC10_00858

表２　「あまり」と「そんなに」の否定文の形容詞述語の内訳

あまり否定（820）

順位	形容詞	度数	評価性	得点
5	好ましい	4	プラス（異なり15語・のべ58語）	7
10	嬉しい	2		7
10	美味しい	2		7
14	格好いい	1		7
14	楽しい	1		7
14	美しい	1		7
7	上手い	3		6.5
7	面白い	3		6.5
14	可愛い	1		6.3
1	良い	23		5.7
14	安い	1		5.7
5	詳しい	4		5.5
14	親しい	1		5.5
10	強い	2		5.3
3	高い	8		4.4
14	偉い	1		4.4
4	多い	6	ゼロ（異なり4語・のべ11語）	4
7	大きい	3		4
14	近い	1		4
14	長い	1		4
2	好きだ	12	マイナス（異なり7語・のべ19語）	3.8
14	甘い	1		3
14	怖い	1		2.4
10	芳しい	2		2
14	ごつい	1		2
14	塩っ辛い	1		2
14	うるさい	1		1
小計		88	平均点	4.8

そんなに否定（560）

順位	形容詞	度数	評価性	得点
17	格好いい	1	プラス（異なり14語・のべ41語）	7
17	楽しい	1		7
17	嬉しい	1		7
17	明るい	1		7
8	良い	5		5.7
12	安い	4		5.7
17	親しい	1		5.5
8	強い	5		5.3
14	広い	2		5
17	珍しい	1		5
17	濃い	1		4.7
1	高い	16		4.4
17	偉い	1		4.4
17	固い	1		4.2
2	多い	14	ゼロ（異なり9語・のべ30語）	4
8	大きい	5		4
8	長い	5		4
17	でかい	1		4
17	若い	1		4
17	重い	1		4
17	小さい	1		4
17	少ない	1		4
17	芳しい	1		4
4	好きだ	10	マイナス（異なり22語・のべ65語）	3.8
17	暑い	1		3.7
17	熱い	1		3.7
14	短い	2		3.5
17	低い	1		3.5
17	薄い	1		3.5
6	遠い	7		3.3
3	甘い	11		3
17	古い	1		3
17	恥ずかしい	1		3
17	冷たい	1		3
17	鈍い	1		2.7
5	悪い	9		2.5
14	嫌だ	2		2.5
17	苦い	1		2.5
17	恐ろしい	1		2.3
6	難しい	7		2
17	生やさしい	1		2
17	痛い	1		1.8
13	ひどい	3		1.7
17	しんどい	1		1.5
17	汚い	1		1
小計		136	平均点	3.9

表３　「それほど」と「たいして」の否定文の形容詞述語の内訳

それほど否定（707）

順位	形容詞	度数	評価性	得点
14	気持ちいい	3		7
27	喜ばしい	1		7
27	明るい	1		7
16	美味い	2		6.7
27	やさしい	1		6.5
27	可愛い	1		6.3
27	望ましい	1		6
16	良い	2		5.7
27	安い	1	プラス（異なり18語・のべ58語）	5.7
11	親しい	4		5.5
16	詳しい	2		5.5
7	強い	8		5.3
9	広い	5		5
16	珍しい	2		5
16	濃い	2		4.7
27	賢い	1		4.5
2	高い	19		4.4
16	深い	2		4.2
1	多い	30		4
2	大きい	19		4
6	長い	9		4
16	激しい	2	ゼロ（異なり8語・のべ64語）	4
27	細かい	1		4
27	重い	1		4
27	少ない	1		4
27	欲しい	1		4
27	暑い	1		3.7
27	熱い	1		3.7
5	遠い	13		3.3
9	古い	5		3
16	甘い	2		3
27	おかしい	1		3
27	恥ずかしい	1		3
8	悪い	7		2.5
27	寒い	1		2.5
27	寂しい	1		2.4
4	難しい	16		2
27	心細い	1	マイナス（異なり24語・のべ73語）	2
16	痛い	2		1.8
11	ひどい	4		1.7
11	酷い	4		1.7
14	きつい	3		1.5
27	醜い	1		1.5
16	辛い	2		1
16	悲しい	2		1
27	恐ろしい	1		1
27	みっともない	1		1
27	暗い	1		1
27	下手だ	1		1
27	見苦しい	1		1
小計		195	平均点	3.8

たいして否定（700）

順位	形容詞	度数	評価性	得点
21	嬉しい	1		7
7	美味い	6		6.7
2	面白い	10		6.5
10	上手い	4		6.5
10	可愛い	4		6.3
21	ありがたい	1		6
21	綺麗だ	1	プラス（異なり15語・のべ62語）	6
1	良い	11		5.7
21	安い	1		5.7
7	親しい	6		5.5
10	強い	4		5.3
3	広い	8		5
21	珍しい	1		5
21	速い	1		4.5
14	高い	3		4.4
4	大きい	7		4
14	重い	3		4
16	若い	2	ゼロ（異なり6語・のべ17語）	4
16	多い	2		4
16	長い	2		4
21	近い	1		4
4	好きだ	7		3.8
16	忙しい	2		3.5
16	遠い	2		3.3
21	古い	1		3
10	悪い	4		2.5
21	寒い	1	マイナス（異なり11語・のべ32語）	2.5
9	難しい	5		2
21	可哀想だ	1		2
4	痛い	7		1.8
21	ひどい	1		1.7
21	つらい	1		1
小計		111	平均点	4.3

表 3 から、「それほど」はマイナス(異なり 24 語：のべ 73 語)・ニュートラル(異なり 8 語：のべ 64 語)・プラス(異なり 18 語：のべ 58 語)の語数はほぼ均等の状態であり、「たいして」はプラスの語が多く(15：62 語)、次に、マイナスの語(11：32 語)で、ニュートラルの語が少ない(6：17 語)。「あまり」の傾向と同様という結果になっていることが分かる。

３．副詞の程度性の下位分類の考察

今まで、程度副詞の下位分類として、「純粋程度」と「量程度」に分けられることが多い。その例として純粋程度には「とても・大変（に）」などが、量程度には「かなり・相当」などが挙げられる（森山 1985:61、仁田 2002：169, 180）。一方、工藤(1983)と渡辺(1990)では程度副詞の「程度性」と「評価性」も論じられている。本稿の用いる「評価性」は単に「ニュアンス」の意味を表し、先行研究の定義と異なるのだが、工藤(1983)では、程度副詞は 2 重性格を持つと述べられている。その 2 重性格とは、いわゆる陳述的に肯定・平叙の叙法と関わって評価性を持ちつつ、事柄的には形容詞と組み合わさって程度限定性を持つことを言う。渡辺(1990)では、程度副詞の「評価性」が「非評価系の程度副詞」と「評価系の程度副詞」に分けられ、「評価系」は主観的な内面の価値尺度に基づく品定めと言われる。また、田和(2011)は今までの先行研究を分かりやすくまとめ、結合させたものである。

先行研究の知見を踏まえ、今回の調査結果に基づくと、副詞の程度性の中には、評価性をプラスとマイナスに分けると、「負の評価性」をより多く用いる程度副詞と「正の評価性」をより多く用いる程度副詞があると考えられる。

４．おわりに

本発表では 4 つの副詞のジャンルと係り先の形容詞の違いを調べた。今後の課題として、4 つの副詞の動詞述語、モダリティなどの違いを調べ、論を深める。

謝　辞

本研究は一橋大学言語社会研究科博士後期課程在籍中に、指導教員石黒圭先生を初め、ゼミの皆様からたくさん貴重なご指摘をいただき、時間がかかったものの、ようやく雛形になってきたものである。記してずっときちんと言えなかった感謝の意を表します。

文　献

小川典子（2008）「そんなに～ない」と「あんまり～ない」における程度の基準について」『日本語学会 2008 年度春季大会予稿集』pp.103-110、日本語学会
工藤浩（1983）「程度副詞をめぐって」渡辺実編『副用語の研究』pp.176-198、明治書院
仁田義雄（2002）『副詞的表現の諸相』pp. 169, 180、くろしお出版
服部匡（1994）「アマリ～ナイとサホド（ソレホド）～ナイ」『日本語日本文学』6、pp.1-21、同志社女子大学
飛田良文・浅田秀子（1991）『現代形容詞用法辞典』
益岡隆志・田窪行則（1992）『基礎日本語文法（改訂版）』p.142、くろしお出版
森山卓郎（1985）「程度副詞と動詞句」『国文学会誌』第 20 号、pp.60-65
渡辺実（1990）「程度副詞の体系」『国文学論集』23

関連 URL

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

『日本語日常会話コーパス』構築における Praat の利用

西川 賢哉 (国立国語研究所コーパス開発センター) *

Utilization of Praat in the Development of the Corpus of Everyday Japanese Conversation

Ken'ya NISHIKAWA (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所で構築を進めている『日本語日常会話コーパス』(CEJC) のアノテーション作業（書き起こし・短単位情報付与作業）を支援するために，無償の音声分析ソフトウェア Praat を利用したツールをいくつか開発した：(i)［Praat 起動］必要な情報（ファイル名・時刻情報等）が記された Emacs バッファ，あるいは形態論情報修正ツール「大納言」の検索結果画面から Praat を起動し，転記情報とともに当該箇所を表示するツール，(ii)［転記保存］Praat TextGridEditor 上で変更した転記を，CEJC 転記ファイル（タブ区切り形式）に上書き保存するツール，(iii)［メモ］TextGridEdior 上で選択された区間にある転記情報を，その他必要な情報（ファイル名・時刻情報等）とともにクリップボートにコピーするツール，(iv)［別音声聴取］当該会話に参加している別の話者の音声ファイルを追加で開くツール，など。これらのツールを用いることで，音声聴取をはじめとする，話し言葉コーパス構築に不可欠な作業が簡単な操作で行なえるようになり，作業の効率化および精度の向上が期待できる。

1. はじめに

コーパス開発センターでは，音声コーパス構築における作業者の負担軽減や作業の効率化を目指し，作業支援手法の開発を進めている。本稿では，無償の音声分析ソフトウェア Praat (Boersma & Weenink 2018) を利用したアノテーション（書き起こし・短単位情報付与作業）支援ツールを紹介する(1)。

2. 『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation; CEJC)

本稿で紹介するツールは，現在のところ，『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation; 以下 CEJC) 構築作業で使用されている(2)。ツールの紹介に先立ち，

* nishikawa[AT]ninjal.ac.jp

(1) 本稿では Praat の機能についてはほとんど触れない。Praat を基礎からわかりやすく解説したものとして北原・田嶋・田中 (2017)，話し言葉コーパスの構築・分析の観点から Praat の機能を簡潔に紹介したものとして西川 (2015) を参照されたい。なお，以下に紹介するツールにおいては，Praat を外部から操作するプログラム `sendpraat` を使用している。同プログラムは，Praat 公式サイト内で配布されているが，目立たない場所に置かれているため (`http://www.fon.hum.uva.nl/praat/sendpraat.html`)，非常に有益なものにもかかわらず，広く知られているわけではないと思われる。`sendpraat` については，Praat の Help（あるいは，`http://www.fon.hum.uva.nl/praat/manual/Scripting_8_2__The_sendpraat_program.html`）を参照。

(2) 内部で CEJC に特有の処理も行なっているが，できるだけ（最小限の修正を施すだけで）他のコーパスに対しても使用できるよう配慮しつつツールを作成した。

CEJC について必要な範囲で簡単に触れておく。

2.1 収録

日常生活において自然に生じる活動に埋め込まれた多様な会話を収録するために，調査協力者にビデオカメラや IC レコーダーなどの収録機材を 2-3 か月間ほど貸し出し，日常生活における多様な場面での会話を自ら収録してもらう。プロジェクトメンバーは収録場面に立ち会わない。IC レコーダーは会話者全員が装着する。個々の発話に加え，会話全体を録音するために，別の IC レコーダを中央に配置する。したがって，一つの会話に対し，複数の音声ファイルが存在することになる。同時に動画も収録している。収録についての詳細は，田中他 (2018) を参照。

2.2 アノテーション

収録した音声に対し，図 1 に示すようなアノテーションを施す。そこに示されている通り，「コア」と呼ばれるサブセットに対しては，より詳細な情報を人手で付与する。

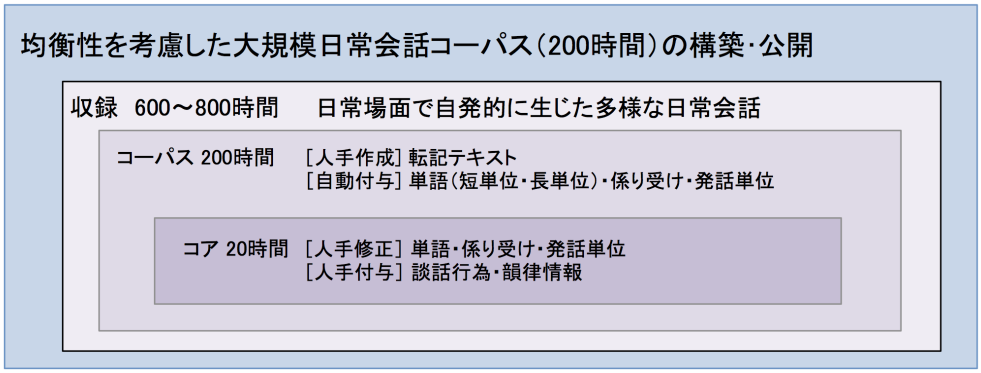


図 1 CEJC のアノテーション（プロジェクトサイトより引用）

以下，転記テキストと単語（短単位）について簡単に述べる。

2.2.1 転記テキスト

映像分析ソフトウェア ELAN[(3)]や Praat を用いて，音声を書き起こす。作業上は，図 2 に示される通り，タブ区切りのテキストファイル (tab-separated values; tsv) で管理されている。1 行は転記単位と呼ばれる単位で区切られており，発話の開始時間と終了時間が割り当てられている。転記テキストには必要に応じて各種タグが付与される。転記テキストについての詳細は，臼田他 (2018) を参照。

2.2.2 単語（短単位）

形態素解析器 MeCab（工藤他 2004)[(4)]と形態素解析用辞書 UniDic[(5)]を用いて，転記テキストを短単位解析したのち，形態論情報管理ツール「大納言」（小木曽・中村 2014）を用いて形態論情報を修正する（図 3)。短単位の規定については，小椋他 (2011) を参照。

(3) http://tla.mpi.nl/tools/tla-tools/elan/
(4) http://taku910.github.io/mecab/
(5) http://unidic.ninjal.ac.jp/

fileID	speakerID	startTime	endTime	pause	text
T004_003	IC01	23.733	24.403	2.716	いいよ いいよ (D ＃＃)。
T004_003	IC02	23.851	24.172	2.09	うん。
T004_003	IC02	26.262	26.947	1	でかいんだよ。
T004_003	IC01	27.119	27.302	1.798	うん。
T004_003	IC02	27.947	28.506	0.002	だから。
T004_003	IC02	28.508	28.99	23.849	あれが。
T004_003	IC01	29.1	29.59	0.097	そうだね。
T004_003	IC01	29.687	30.603	14.262	(W デシ｜出し) にくいんだ。
T004_003	IC03	38.577	39.252	1.109	あー。
T004_003	IC03	40.361	41.516	0.196	雲取も:。
T004_003	IC03	41.712	41.968	0.736	(D イ)
T004_003	IC03	42.704	44.705	0.541	一組だけ外人のご一行みたいの
T004_003	IC01	44.865	45.601	0.899	えー。
T004_003	IC03	45.246	45.935	2.32	帰る時。

図 2　転記テキスト例（タブ区切り）

図 3　形態論情報管理ツール「大納言」

3.　Praat を用いたアノテーション支援ツール

CEJC アノテーション作業を支援するためにこれまでに開発したツールを紹介する。

3.1　Praat 起動 (1): Emacs から

もっとも基本的なツールとして，必要な情報（ファイル名・時刻情報等）が記されたテキストから，Praat を起動し，さらに転記情報とともに当該箇所を表示するツールを作成した。

CEJC 構築作業では，テキストエディタとして Emacs を使用しているため，Emacs Lisp で実装した。この機能は，Emacs 初期化ファイル (.emacs あるいは init.el) で定義してある特定のキー（例えば `C-c C-c C-f`）により実行される。

2.1 節に述べた通り，CEJC では一つの会話に対して音声は複数存在するが，このツールでは起動元のテキストに記されている話者情報を参照し，その話者の IC レコーダーで収録された音声を Praat で開くようにしてある。また，このツールで Praat を起動すると，音声だけでなく，転記も同時に表示される。2.2.1 節に述べた通り，転記ファイルはタブ区切りのテキストファイルで管理されているが，このツールが実行されると，そのタブ区切りファイルから動的に（その場で）TextGrid ファイル（Praat アノテーション形式）が生成され，それが Praat で開かれる。

本ツールは，単なる音声再生機能と比べて，

- 音声だけでなく，波形やスペクトログラムも参照することができる
- Praat TextGridEditor 上で区間を選択し直すことで，特定の部分だけを，繰り返し再生することができる

といった利点がある。

このツールでは，オリジナルの転記テキストからも Praat を起動することができる。ただし，このツールを実行した時点で，TextGrid のほうがマスターデータとなるので，転記ファイルを開いたバッファは自動的に書き込み禁止とするようにしてある。

3.2 Praat 起動 (2):「大納言」から

上と同様の機能を形態論情報修正ツール「大納言」にも実装した。その結果，「大納言」における短単位検索結果画面からも Praat を起動できるようになった。実行方法は，対象とするレコードの「ファイル名」のセルをダブルクリックするだけである。話し言葉コーパス構築作業においては，短単位解析結果から音を聴取したいというケースは，意外に多い。

3.3 転記保存

Praat で表示される転記に誤りが発見された場合，Praat 上で修正を施しファイルに保存できれば便利だが，単純に Praat の保存機能を使うと，TextGrid 形式（Praat のアノテーション形式）でファイルが保存されてしまう。そこで，変更した転記（Praat では TextGrid オブジェクト）を，CEJC 転記テキストの形式（タブ区切り形式；図 2 参照）で上書き保存するツールを作成した(6)。これにより，作業者はわざわざ転記ファイルに戻る必要がなく，Praat 上で自由に転記を修正できる。

3.4 メモ：クリップボードにコピー

転記で対処不明な箇所があった場合など，メモを取っておき，作業者間でその個所を共有したい，といったケースがある。そのとき，そのメモから 3.1 節に述べたツールを用いて，Praat で当該箇所を表示できれば便利である。そこで，TextGridEdior 上で選択された区間にある転記情報を，その他必要な情報（ファイル名・時刻情報等）とともにクリップボートにコピーす

(6) このツールを導入したことにより，CEJC 構築作業において，TextGrid ファイルを管理する必要がなくなった。

るツールを作成した。このツールを実行後，Praat からテキストエディタ等（例えば Emacs）に移動し，ペーストを実行すれば，ファイル名などとともに当該転記が張り付けられる。

3.5 別音声聴取

CEJC のように，複数の話者が参加している会話の音声をアノテーションしている際，別の話者の（同じ個所の）音声を聴取したくなる場合がある。例えば，Praat で IC01 の音声を聞いている最中に，IC02 の音声を聞きたい，といった具合である。そこで，当該会話に参加している別の話者の音声ファイルを追加で開くツールを作成した。このツールを実行すると，別の TextGridEditor が起動するが，転記は同じものが開かれるので，どちらの TextGridEditor でも転記の修正が可能である。

4. おわりに

CEJC アノテーション支援ツールを紹介した。これらのツールを用いることで，音声聴取をはじめとする，話し言葉コーパス構築に不可欠な作業が，簡単な操作で行なえるようになり，作業の効率化および精度の向上が期待できる。

ここに紹介したツールのほかにも，Praat から，そこで選択されている区間の動画を再生するツールなど，追加のツールを現在作成中である。作業者のフィードバックを得ながら，より便利なツールの開発を進めたい。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「大規模日常会話コーパスに基づく話し言葉の多角的研究」の成果である。形態論情報修正ツール「大納言」への Praat 起動機能実装にあたり、中村壮範氏（国立国語研究所コーパス開発センター）の協力を得た。記して感謝する。

文　献

Boersma, Paul and Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/

北原真冬・田嶋圭一・田中邦佳 (2017) 『音声学を学ぶ人のための Praat 入門』ひつじ書房.

工藤拓・山本薫・松本裕治（2004）「Conditional Random Fields を用いた日本語形態素解析」『情報処理学会研究報告自然言語処理（NL）』47, pp. 89-96.

小木曽智信・中村壮範 (2014)「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」『自然言語処理』21 巻 2 号, pp. 301–332.

西川賢哉 (2015)「音声分析ソフトウェア「Praat」」小磯花絵（編）『話し言葉コーパス：設計と構築』朝倉書店, pp. 152–167.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011)「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下）」特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-05-02）（http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf よりダウンロード可能）

田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2018)「『日本語日常会話コーパス』の構築：会話収録法に着目して」『国立国語研究所論集』14, pp. 275–292.
臼田泰如・川端良子・西川賢哉・石本祐一・小磯花絵 (2018)「『日本語日常会話コーパス』における転記の基準と作成手法」『国立国語研究所論集』15, pp. 177–193.

関連 URL

「大規模日常会話コーパスに基づく話し言葉の多角的研究」プロジェクトサイト
http://pj.ninjal.ac.jp/conversation/

多様な研究分野に利用可能な超高精細・高精度手話言語データベースの開発

長嶋祐二 (工学院大学) *, 原大介 (豊田工業大学) , 堀内靖雄 (千葉大学)
酒向慎司 (名古屋工業大学) , 渡辺桂子 (工学院大学) , 菊澤律子 (民族学博物館)
加藤直人 (NHK 放送技術研究所) , 市川熹 (千葉大学/工学院大学)

Development of the Super High-Definition and High-Precision Japanese Sign Language Database Available for Various Research Fields

Yuji Nagashima (Kogakuin University)
Daisuke Hara (Toyota Technological Institute)
Yasuo Horiuchi (Chiba University)
Shinji Sako (Nagoya Institute of Technology)
Keiko Watanabe (Kogakuin University)
Rituko Kikusawa (National Museum of Ethnology)
Naoto Katho (NHK STRL)
Akira Ichikawa(Chiba University/ Kogakuin University)

要旨

手話は言語であるにもかかわらず、音声言語と比べて言語学、工学を含む関連諸分野での研究が進んでいない。本稿では、各個分野における手話研究および学際研究の推進を目的とした、様々な分野の研究者が共通に利用できる汎用的な日本手話の語彙データベース作成について報告する。

言語学者の望むデータ形式と、工学や認知科学の分野で望むデータの形式は異なることが予想される。多分野での利用を可能にするためには、分析や解析内容に応じて手話の多視点の画像、3 次元動作データ、深度画像など様々なデータ形式を含むことが望まれる。さらに、時間軸上で同期したこれらのデータを、各分析者が得意とするデータ形式で解析することを可能にする。データベース上の様々な形式データを同期解析できるアノテーション支援システムも開発する予定である。これにより、様々な視点からの同一手話の解析が可能となり、手話言語に関する新たな知見が得られることが期待できる。

1. はじめに

2006 年 12 月、「障害者の権利に関する条約 」が第 61 回国連総会で採択され、2008 年に 5 月に発効した。この条約の第二条では、「言語とは音声言語及び手話その他の形態の非音声言

* nagasima@cc.kogakuin.ac.jp

語をいう」と定義されている。さらに、第二十一条　表現及び意見の自由並びに情報の利用の機会では、「手話の使用を認め及び促進すること」と、謳われている。日本は、2007 年 9 月 28 日にこの条約に署名し、国内法の整備・改革を行い、2016 年 4 月 1 日に「障害者基本法」の最終改正が施行された。この様な現状の中、全日本ろうあ連盟では「手話言語法」の制定を目指し、活動が行われている。手話言語条例を成立させている自治体は、13 県、75 市、9 町の 97 自治体となっている (2017 年 4 月 1 日現在)。また、国に「手話言語法」の制定を求める意見書は、2016 年 3 月には、全国（47 都道府県・東京 23 区・1,718 市町村）で採択されている [1]。さらに、全国手話言語市区長会には、463 の首長が参加している (2018 年 6 月現在)。手話は、聴覚障害者のコミュニケーション手段の一つであり、音声言語とは異なる文法体系をもつ独立した対話型の自然言語である。手話を構成する要素は、手指動作と非手指動作である。手指動作は、手型、提示される位置、掌方向、および手の大局的な運動により表現される。非手指動作は、視線、頷き、表情、口形など手指動作以外の要素であり、マルチモーダルな機能を表現している。手話の手指動作は、両手で語を構成したり、利き手と非利き手とが独立てそれぞれ語を形成したりする。このように、手話は、音声言語と異なり、複数の調動器官により、線状的にも非線状的にも語を構成する複雑さが存在する。手話は言語であるにもかかわらず、音声言語と比べて言語学、工学を含む関連諸分野での研究が進んでいない。この原因の 1 つは、言語学者や工学者など様々な分野の研究者が共通に利用できる汎用的なデータベースが存在しないためである。しかし、言語学者の望むデータ形式と工学、認知科学の分野で望むデータの形式は異なることが予想される。多くの研究分野で手話を統一的に研究するには、同じ手話動作を各研究者のニーズに合ったデータ形式で提供することが望ましい。同一の手話動作を様々な視点や手法で分析し俯瞰することは、新たな知見を得る機会を増大させる可能性をもつ。

本報告では、手話語彙のデータベースの構築方法について検討い、2017 年度の結果について述べる。

2. データベースへの収録データ形式

日本語の音声や言語データは、国立情報学研究所において、音声資源コンソーシアム (SRC: Speech Resources Consortium) が設立され、日本語の研究の発展に寄与している。音声データは、時間軸方向の 1 次元データであり、様々な解析手法が提案され実用化されている。ビックデータ解析により、新たな音声認識技術が飛躍的に進展している。

一方、手話は複数の調動器官によって語が形成されるとされているが、その音素の構造すらはっきり定義されていない。手話の弁別的特徴や音素、形態素の詳細な分析のためには、手指動作や非手指動作の詳細な分析が必要と考える。音声データが時間軸方向の 1 次元データであったのに対し、手話のデータは時間軸方向の空間的な広がりをもつ 3 次元データである。音声のように、手話の動作データを数値的に解析を行うことで、新たな知見を得られる可能性が高い。しかし、音声に比較して次元数が多くなり、その複雑度はかなり高いと予測される。

さらに、手話のデータは、各研究機関により独自に集められ公開されているものはない。手話を分析するために必要な、3 次元空間上の手指や非手指の構成要素の数値データは公開されていない。このため、手話の数値的な分析には、どの程度の空間・時間分解能のデータが必要

かも不明である。各研究機関によって収録される手話のデータ形式は、動画像映像が多いと考えられる。しかし、利用目的、分析や解析手法が異なるため、カメラ台数や撮影方法、解像度など様々であり、共通に利用することは困難と考えられる。

高精度に手話動作を分析したり高品位な手話 CG 生成したりするには、高精度な動作の 3 次元計測を必要とする。そしてもし、同期して高精度な 3 次元手話動作データと 2 次元の手話映像が存在すれば、非手指動作を含めた手話の認識や動作分析において、3 次元動作がどのように 2 次元に縮退され時間軸方向へ進行しているのかの解析が可能となり、新たな手話理解・認識のための方法論を得ることが可能となると考えられる。

そこで、本データベース構築では、どの程度の空間・時間分解能のデータが必要かも分析できるように、現時点で可能な最高水準の精度の手話動作収録手法とデータ形式について検討を行なう。

2.1 3 次元動作データ

3 次元空間的かつ時間的に高精度にデータを計測する方法は、コストを考えなければ光学式モーションキャプチャ (以下、 MoCap とする) である [2]。文献 [2] の計測では、東映ツークン研究所において、1600 万画素のモーションカメラ 42 台を用いて $2 \times 2 \times 2$ m^3 を計測しているため、空間分解能は 0.5mm を、時間分解能は 120fps を実現している。撮影で用いた再帰性反射マーカは、手型と顔表情を高精度に計測のため直径 3mm を用いている。そこで、本データベースもこの 3 次元計測環境を用いる。これにより、手指動作ならびに非手指 動作の構成要素の詳細な解析を期待できる。

なお、表 1 に、再帰性反射マーカの情報を示す。

表 1　再帰性反射マーカ情報

body Region	Retro-reflective Markers	
	Diameter [mm]	Number
Face	3	33
Hand	3	24×2
Others	10	31
Total Number of Markers		112

2.2 映像データ

手話画像認識や対話分析では、より高解像度のカメラが望まれる。画像計測では、最低 2 台以上のカメラを必要としている。そこで、撮影カメラ構成として、画面解像度はフル HD の 1920×1080pixel、あるいは 4K の 3840×2160 pixel を、時間分解能は 60fps を 3 台用いる。

2.3 深度 (距離) データ

最近、ToF(Time of Flight) 方式により、安価でかつ比較的高精度に距離を計測可能なセンサが普及している。手話認識でもこの方式を用いる研究機関が多くなっている [3],[4]。しかし、時間分解能は最大 $30fps$ となっている。この深度計測システムでは、赤外線映像と通常の映像を同時記録できるメリットがあるので、データベースに収録する。

2.4 異種データの同期収録

本データベースでは、様々な分野での利用とその解析データを統一的に扱うため、2.1, 2.2, 2.3 で得られる手話 データの同期収録を目指す。同一の手話動作を様々な時間分解能や空間分解能で観察したり、分析したりすることの意義は非常に大きいと考える。

3. 語彙の選定方法と言語資料提供者

3.1 語彙の選定方法

紙媒体で出版されている手話の辞書は多く存在する。この中で、最も収録語彙数の多い辞書は、全日本ろうあ連盟から発行されている「日本語－手話辞典」である [5]。この辞書は、日本語語彙数にして約 6、000 語を収録している。また、比較的規模の大きい手話文データベースには、NHK の E テレの手話ニュースからの手話文データベースがある。この手話文データベースには、約 130,000 文、総単語数約 3,036,000 語 (異なり語数で約 76,000 語) となっている (2018 年 3 月時点)[6]。

そこで、提案する DB への収録語彙の選定では、NTT データベースシリーズ「日本語の語彙特性」第 9 巻「 単語親密度　増補版 」[7] と、国立国語研究所・ 情報通信研究機構（旧通信総合研究所）・ 東京工業大学 が共同開発した日本語の自発音声を大量にあつめて多くの研究用情報を付加した話し言葉研究用のデータベース「日本語話し言葉コーパス」[8]、および NHK の E テレの手話ニュースからの手話文データベース用いた。単語親密度では音声親密度の高いもの、日本語話し言葉コーパスと手話文データベースでは出現頻度の高い語彙から、選定候補語彙としている。そして、選定語彙候補の中から、「日本語-手話辞典」に掲載されている語彙を最終的にデータベース収録語彙と決める。ただし、「日本語-手話辞典」に掲載されていなくても、日常よく使われる手話単語と判断される場合には、その語彙も収録候補とする。

2020 年までに、約 5〜6 千語彙の抽出を目指す。2017 年には、テスト的に日本語ラベル数で 400 語彙の抽出作業を行った。そして、この 400 語彙に対して、異動作同義語を含めて 525 語彙の手話動作の確定作業を行った。収録する手話動作形の確定作業は、筆者と研究協力者のろう者の手話母語者、CODA の手話母語者との既存の辞書分析作業と話し合いで決定した。

3.2 言語資料提供者

構築を目指している DB は、最終年度に公開を予定している。そこで、手話語彙の収録は、DB のより広い応用を考えて、男性と女性の各 1 名で行う。言語資料提供者を選定するための条件は、

- 手話母語者の家系の手話母語者
- 撮影にある程度慣れている
- 手話の読み取りがしやすい
- 撮影した映像の公開を許諾する

とした。この DB を構築するプロジェクトの研究協力者の手話母語者の面接などを行うことで、男性 M(38 歳) と女性 K(39 歳) に決定した。

4. 手話語彙の収録方法

構築する DB では、2.1, 2.2, 2.3 での検討に従って、MoCap による 120 fps の超高精度な 3 次元動作、60 fps の Full HD の高精細の映像、30 fps の深度センサからの映像の 3 種類の異なるフレームレートのデータを 2.4 の目的により、同期させて収録することを目指す。この目的達成のため撮影は、専門の知識とノウハウを持つ東映のツークン研究所のモーションキャプチャスタジオで行うことにした。

2017 年に撮影した時の撮影機材の構成と同期の概念図を図 1 に示す。MoCap カメラには、Vicon の 1,600 万画素の T160 と V16 の合計 42 台を用いている。DB に収録するデータ形式は、C3D データ、BVH ファイル形式である。C3D データは、MoCap 用の再帰性反射マーカ点の座標と角度データである。BVH ファイル形式は、Biovision 社によって開発された BioVision Hierarchy によるフォーマット形式である。BVH ファイル内には再帰性反射マーカの点の位置情報は無く、モデル情報が記録されており、その内容は主に HIERARCHY 部と MOTION 部に分けられる。HIERARCHY 部には、キャラクタのスケルトン階層構造が定義されている。MOTION 部には、階層構造中の関節 (JOINT) に対しての位置や回転の値がオイラー角表示で記述されている。

Full HD の映像は、camcorder#01～#03 のカメラ 3 台により正面・左側・右側から撮影した。さらに、図 1 の camcorder #04 は、参考映像として 120 fps の Full HD の映像も収録している。この映像の再生速度は、30 fps となりスーパースローとなっている。Kinect #01 の深度センサには、Kinect v2 を用い赤外線画像と深度画像の収録を行っている。カメラやセン

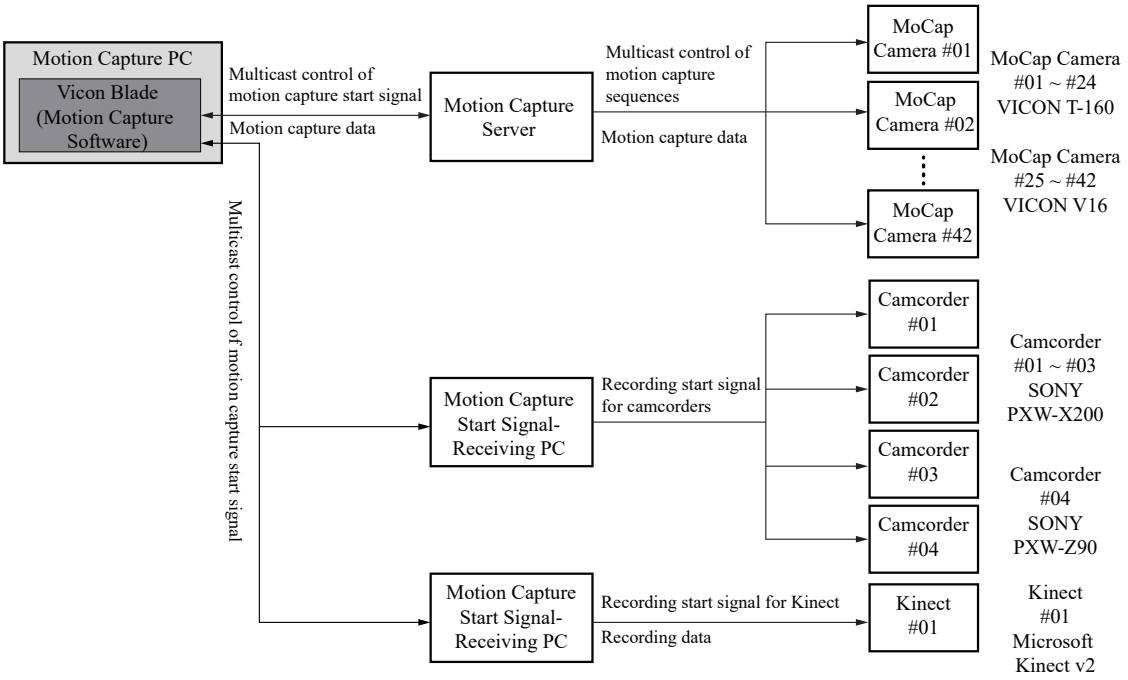


図 1　撮影機材の構成と同期の概念図

サの配置図を図 2 に示す。図 2 の長さと角度の詳細な値を表 2 に示す。また、2017 年撮影時の各カメラとセンサの仕様を表 3 に示す。図 3 に、東映東京撮影所モーションキャプチャスタジオでの DB 収録のためのスタジオの外観を示す。

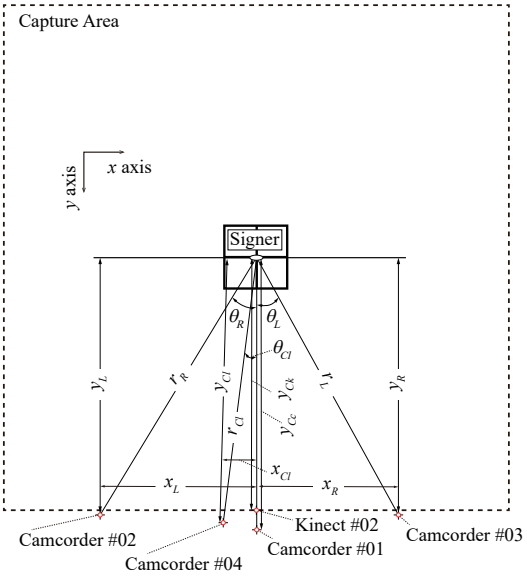


図 2　カメラとセンサの配置図

表 2　図 2 の長さと角度の詳細な値

Equipment	angle [degree]	distance from signer [m]			
		direct distance	y axis	x axis	height
Camcorder#01	0.0	3.27	$y_{Cc} = 3.27$	0.00	1.21
Camcorder#02	$\theta_L = 30.5$	$r_L = 3.41$	$y_L = 2.94$	$x_L = 1.73$	1.20
Camcorder#03	$\theta_R = 28.8$	$r_R = 3.40$	$y_R = 2.98$	$x_L = 1.63$	1.20
Camcorder#04	$\theta_{Cl} = 7.4$	$r_{Cl} = 3.17$	$y_{Cl} = 3.14$	$x_{Cl} = 0.41$	1.28
Kinect #01	0.0	2.51	$y_{Ck} = 2.51$	0.00	1.08

図 3　データ収録のスタジオの外観

表 3　2017 年撮影時の各カメラとセンサの仕様

Optical Motion Capture	
Model Number	VICON T160 (V16)
Frame rate	120 fps
Number of effective pixels	4,704 × 3,456
Number of camera	42
Number of markers	112
Camcorder	
Model Number	SONY PWX-X200
Frame rate	60 fps
Number of effective pixels	1,920 × 1,080
Format	MPEG-4 AVC/H.264
Number of camera	3
Super slow camcorder	
Model Number	SONY PWX-Z90
Frame rate	120 fps
Number of effective pixels	1,920 × 1,080
Format	XAVC
Number of camera	1
Depth sensor	
Model Number	Kinect One Sensor
Resolution	512 × 512
Horizontal field of view	70 degrees
Vertical field of view	60 degrees
Frame rate	30 fps
Number of sensor	1

5.　ビュワーの開発

前節で同期撮影されたデータは異なるフレームレートが混在している。3 次元データや 3 次元アニメーションなど複数の素材を描画することから、広くゲームの世界で利用されている Unity[9] により、同期再生可能なビュアーの開発を行う。ビュアーの開発方針は、任意の 4 種類のデータを同期して再生することである。2017 年度に開発した、ビュアーの主な機能を以下に示す。

- 画面を最大 4 分割して収録されている任意のデータを同期再生
- BVH ファイル形式データによる 3DCG の描画
- C3D データによるマーカ点の描画
- MoCap データは任意の視点と視野角で描画
- MoCap データの描画背景は任意データで可能

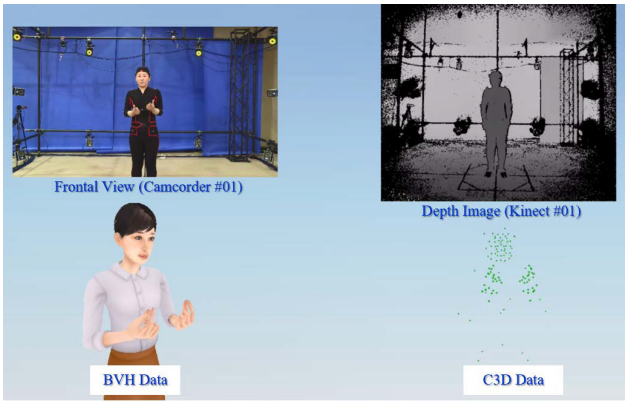


図 4　ビュアーによる各種データの同期再生画面 (手話単語 oNAJI(SAME))

- BVH ファイル形式描画には男性 2 モデル、女性 3 モデルからの選択
- 再生画面の録画 (動画キャプチャ) 機能

図 4 に、手話単語 oNAJI(SAME) の BVH データ (3DCG)、C3D データ (ドットデータ)、正面映像、Kinect センサからの深度データのビュアーによる同期された再生画面を示す。

しかし、Unity の機能の限界から、Full HD 動画と 3 次元動作データとの同期再生では、動画の再生が遅延する問題が発生した。そこで、この問題を解決するため、新たなビュアーの開発を開始している。

6. 今後の課題

本報告では、構築している多用途 DB の構築概念を述べた。構築する DB では、同期撮影されているため異なるフレームレートのデータを時間軸上で分析者の得意とするデータ形式で同期解析を可能としている。言語学、工学などの学際分野で利用可能な手話の言語、認識や生成など工学テータとして、コミュニケーションのモダリティ解析、ビッグデータ解析の礎になるなど様々な分野で利用可能なデータとする。これにより、様々な研究者の視点から同一手話が解析可能となり新たな知見が得られることが期待できる。

今後は、DB の拡張に向けた収録語彙の選定並びに、DB 上の様々な形式データを同期解析できるアノテーション支援システムも開発する予定である。構築する DB は、国立情報学研究所の情報学研究データリポジトリへの手話言語資料の登録も視野に入れて行う。

謝　辞

本研究の一部は JSPS 科研費 17H06114 の助成を受けたものです。

参考文献

[1] 長嶋祐二, 加藤直人, 山内結子, 河野純大：手話コミュニケーションのための情報保障技術, 電子情報通信学誌, Vol.101, No,1, pp.66–72, 2017.
[2] 渡辺桂子, 長嶋祐二：手話形態素辞書作成のための情報入力支援システム, 電子情報通信学会論文誌 D, Vol.J100-D, No.3, pp.298–309, 2017.
[3] Mika Hatano, Shinji Sako and Tadashi Kitamura：Contour-based Hand Pose Recognition for Sign Language Recognition, Proc. of 6th Workshop on Speech and Language Processing for Assistive Technologies, Sep. 2015.
[4] 古谷佳大, 堀内靖雄, 川本一彦, 下元正義, 眞崎浩一, 黒岩眞吾, 鈴木広一：“手話認識における位置・動き特徴量の検討”, 電子情報通信学会論文誌 D, Vol.J99-D, No.1, pp.90–92, 2016.
[5] 日本手話研究所 (編), 日本語–手話辞典, 全日本ろうあ連盟, 1999.
[6] 加藤直人, 内田翼, 東真希子, 梅田修一：ニュースを対象にした手話マルチメディアコーパスの構築, 言語資源活用ワークショップ (LRW2018), 2018.
[7] 天野成昭, 笠原 要, 近藤公久 (編著)：日本語の語彙特性 第 9 巻–単語親密度 増補版–, 三省堂, 2008.
[8] 日本語話し言葉コーパス, http://pj.ninjal.ac.jp/corpus_center/csj/ (2018 年 7 月 27 日参照).
[9] https://unity3d.com/ (2018 年 7 月 27 日参照).

英語における頭子音連結の多変量解析

張 一新 (九州大学芸術工学府)
中島 祥好 (九州大学芸術工学研究院)
于 暁陽 (九州大学芸術工学府)
岸田 拓也 (九州大学芸術工学研究院)
上田 和夫 (九州大学芸術工学研究院)
Sophia Arndt (School of Psychology, National University of Ireland, Galway)
Mark A. Elliott (School of Psychology, National University of Ireland, Galway)

Multivariate Acoustic Analysis of Initial Consonant Clusters in English

Yixin Zhang (Graduate School of Design, Kyushu University)
Yoshitaka Nakajima (Dept. Human Science/Research Center for Applied Perceptual Science, Kyushu University)
Xiaoyang Yu (Graduate School of Design, Kyushu University)
Takuya Kishida (Department of Human Science, Kyushu University)
Kazuo Ueda (Dept. Human Science/Research Center for Applied Perceptual Science, Kyushu University)
Sophia Arndt (School of Psychology, National University of Ireland, Galway)
Mark A. Elliott (School of Psychology, National University of Ireland Galway)

要旨

英語学習者にとって、子音連結を適切に発音・知覚することは大切である。本研究では、英語母語話者が発話した英語音声を録音し、その音響的特徴の分析を行った。英語音節の頭子音連結や、頭子音とその次の母音をスペクトル変化の観点から調べ、録音した音声のスペクトル変化を因子分析した。得られた３因子から鳴音性と密接に関連する２つの因子が取り出された。一方の因子得点が高い場合、もう一方の因子得点は 0 に近く、因子空間におけるその散布図は L 字型のように分布した。単語ごとの頭子音連結から母音に移る際の因子得点はこの L 字型分布に沿って変化することが分かった。この際に、頭子音連結は L 型の中央部に分布し、最も点が密集したのは、角の点に当たる原点の近くだった。第一子音から第二子音の間で、因子得点は統計的に有意に変化するが、第二子音から母音の間では因子得点に有意差はなかった。子音-子音-母音連結(CCV 連結)の各音素に鳴音性と特に相関の高い mid-low factor の特徴がよく反映されていることが分かった。

１．はじめに

英語を外国語とする学習者にとって、子音と子音、子音と母音の組合せを正しく把握することが大切である。英語において子音のみが連続する現象は「子音連結」と呼ばれる。Nakajima et al. (2017) はイギリス英語音声に対して、因子分析を行って３つの因子を取り出し、各音素を三次元の因子得点空間で表した。因子得点空間内で音素が曲線状に並び、さらにその音素の並びは鳴音性の高低の順になることを見つけた。Spencer (1996) による鳴音性という尺度では、母音、渡り音、流音、鼻音、摩擦音・破擦音、破裂音の順に鳴音性が低くなるとしている。音節の始まりは、鳴音性が低い音素から高い音素へつながる、鳴音性連続原理 (sonority sequencing principle; Rahilly, 2016) に従うことが知られている。

この原理に従って、語頭で子音連結からその次の母音に至るスペクトル変化に注目し、鳴音性との関係を探ることを目的とする。

２多変量解析

２.１　音声

２.１.１　分析対象音声

アイルランド英語母語話者の 24 歳から 30 歳までの男性 1 名、女性 2 名、計 3 名の音声を録音した。録音の内容は“The ATR English Database”というデータベースに用いられている 200 文を選択した。200 文について、話者は普段の会話の速さで発音した。

２.１.２　録音条件

録音は暗騒音が 30 dBA 程度の防音室で行った。まず、防音室の暗騒音が話者の発話に影響するかどうかを確認するために、騒音計を設置した。全話者が発音する際の音圧レベルは 65 dBA 以上であり、暗騒音のレベルよりも十分に高かった。話者の口元から 15 cm の距離に設置した録音機(TASCAM, DR-07)を用いてサンプリング周波数 44.1 kHz、量子化ビット数 16 ビットという条件で録音を行った。

２.２因子分析

Nakajima et al. (2017) において示された鳴音性と関連する因子を取り出し、その因子得点を音素ごとに観察するために、Kishida et al. (2016) の方法を用いて、男性 1 名の英語音声のパワースペクトル変化に対して因子分析を行った。Kishida et al. (2016) の分析法が Nakajima et al. (2017) の分析法と異なる点は、パワースペクトルの平滑化にケプストラム分析が用いられる点と、因子分析で導かれる部分空間の起点をパワーが零となる点に修正された、起点移動主成分分析が用いられている点である。本研究では、鳴音性と子音の関係を見る観点から Kishida らの方法を採用した。

３．結果と考察

因子分析により、Fig. 1 に示すような特徴をもつ３つの因子が得られた。得られた３つの因子は先行研究で得られた因子と共通する特徴があった。つまり、Nakajima et al. (2017)と同様に約 1100 Hz 付近の中帯域に大きい因子負荷量をもつ因子(mid-low factor)、約 3300 Hz 以上の高帯域に大きい因子負荷量を持つ因子(high factor)、そして因子の約 500 Hz 以下と約 1700~3300 Hz の間の 2 帯域で大きい因子負荷量をもつ因子(low & mid-high factor)の３つが得られた。

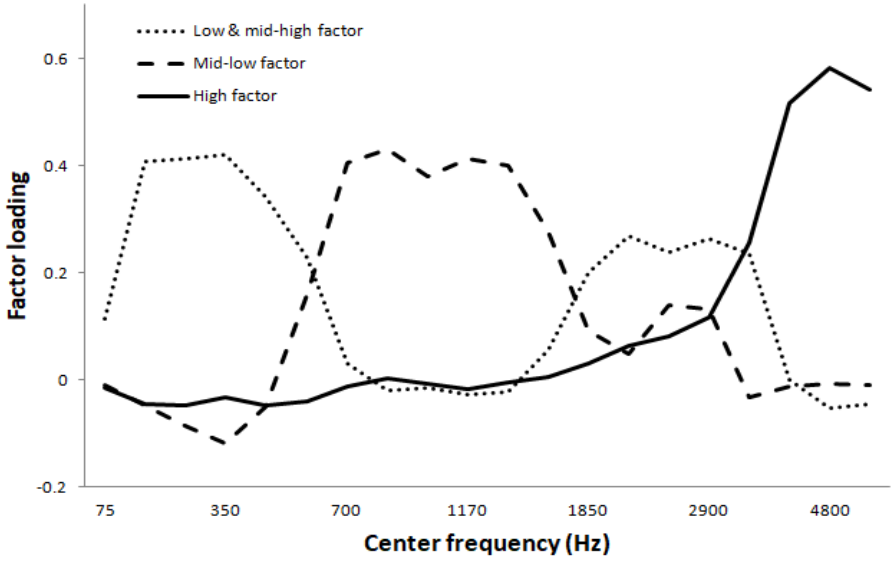


Fig.1 因子分析により、得られた 3 因子の特徴

文頭に現れる 30 の子音-子音-母音連結(CCV 連結)について因子得点を分析した。CCV 連結の各音素にどの因子の特徴が反映されているかを見るために、CCV 連結を音素ごとに分け、音素の時間的中央点における因子得点を因子空間内にプロットした（Fig. 2）。母音は low & mid-high factor、mid-low factor ともに因子得点が大きく、子音はともに小さい傾向があることが Fig. 2-2 から分かる。high factor の因子得点は low & mid-high factor と mid-low factor の因子得点が低い時に高くなる傾向にあった(Fig. 2-1, 2-3)。mid-low factor と high factor の 2 つの因子について、一方の因子得点が高い場合、もう一方の因子得点は 0 に近く、因子空間上で L 字型に分布した(Fig. 2-3)。以上の因子得点の分布の仕方は、Nakajima et al. (2017) における分布と一致している。

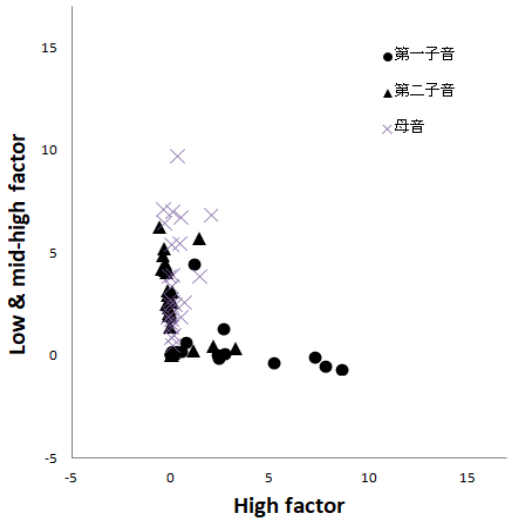


Fig. 2-1

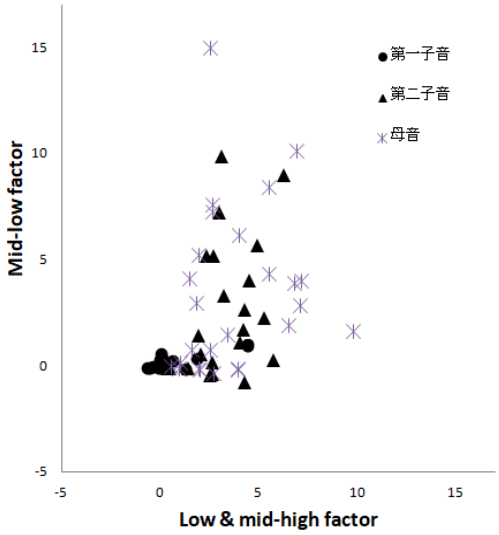


Fig. 2-2

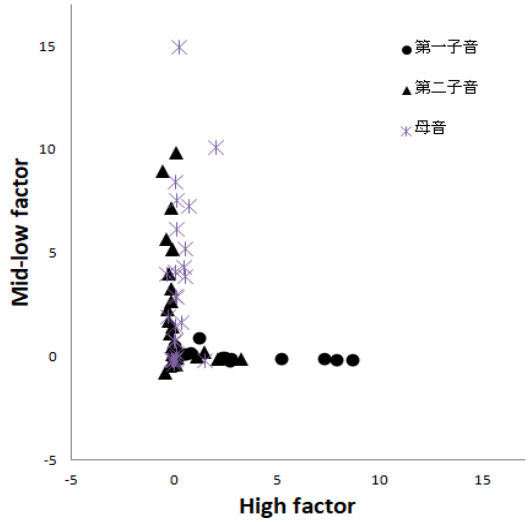


Fig. 2-3

Fig.2　英語母語話者の男性 1 名について、各音素の時間的な中央点での因子得点

また、この因子空間上で、30 の CCV 連結のつながりのそれぞれについて第一子音、第二子音、母音の順に線でつなぐことで各 CCV 連結の発話の際の因子得点の動きを観察した(Fig. 3)。第一子音から第二子音、そして母音に移行する際に、mid-low factor の因子得点が高くなることが分かった(Fig. 3-1)。頭子音連結(第一子音と第二子音)は、因子空間の原点に近い

L 字型の角の付近から始まるか、その付近で終わるかのいずれかである(Fig. 3-2)。第二子音から母音に移行する際に、mid-low factor の因子得点も高くなる傾向があった(Fig. 3-3)。high factor の因子得点が大きく変化するのは、第一子音から第二子音に移行するときだけである。また、頭子音連結(第一子音と第二子音)は L 字型の角に接しており、それは因子空間の原点付近であった。

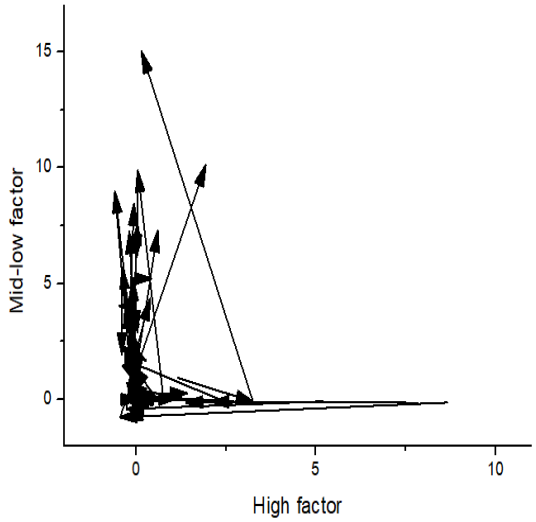


Fig. 3-1

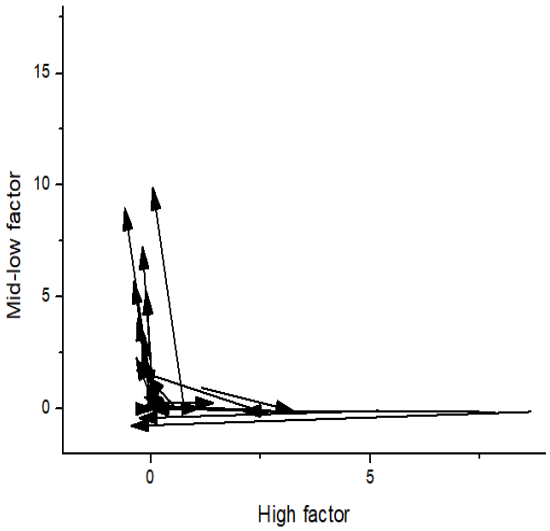


Fig. 3-2

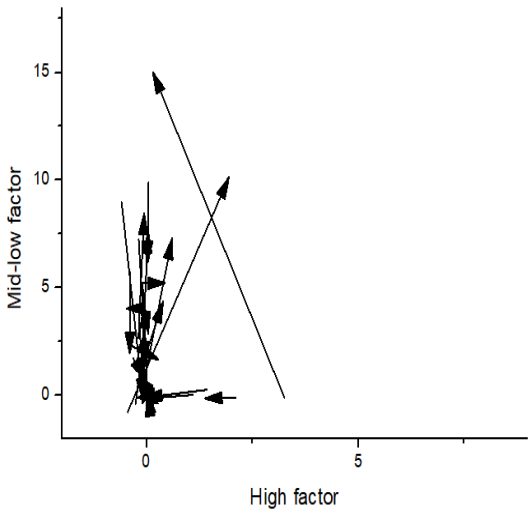


Fig. 3-3

Fig. 3 英語母語話者の男性 1 名について、子音 1 、子音 2 と母音の時間的な中央点での因子得点の動きを矢印で示した。

各因子の因子得点が一組の連結における音素の間で統計的に有意に変化するかどうかを確かめるために、符号検定(有意水準 5%)を行った。low & mid-high factor の因子得点は、第一子音、第二子音の間では有意に大きくなり、第二子音、母音の間では有意差はなかった。mid-low factor の因子得点は、第一子音、第二子音の間では有意に大きくなり、第二子音、母音の間では有意差はなかった。high factor の因子得点は、第一子音、第二子音の間では有意に小さくなり、第二子音、母音の間では有意差はなかった。

第一子音と第二子音の因子得点の変化において、明確な結論が得られた。第二子音、母

音の間では有意差について、統計的に有意となるかどうか、比率の差の大きさ以外に、標本数にも依存している。以上得られた結果を更に検証し、現れた子音-子音-母音音節(CCV連結)を 60 個に増やす予定である。子音連結については、/s/のない子音連結、/sl/, /sw/, /sm/, /sn/の子音連結、/sp/, /st/, /sk/の子音連結、子音が三つの連結(最初は必ず/s/)という四つの種類ごとにそれぞれ線でつなぎ観察することを追加分析として行う予定である。

謝 辞

本研究は、科学研究費補助金（17H06197）の助成を受けた。

文 献

de Saussure, F. (1959). Course in general linguistics (Baskin, W. Trans.). *New York: Philosophical Library.[JL]*.

Kishida, T., Nakajima, Y., Ueda, K., & Remijn, G. B. (2016). Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech. *Frontiers in Psychology*, 7:517.

Nakajima, Y., Ueda, K., Fujimaru, S., Motomura, H., & Ohsaka, Y. (2017). English phonology and an acoustic language universal. *Scientific Reports, 7:46049.*

Rahilly, J. (2016). Sonority in natural language: A review. In M. J. Ball & N. Müller (Eds.), *Challenging Sonority: Cross-Linguistic Evidence*. South Yorkshire, UK: Equinox Publishing Ltd.

Spencer, A (1996). *Phonology: Theory and Description*. Oxford: Blackwell.

Ueda, K., &Nakajima, Y. (2017). An acoustic key to eight languages/ dialects: Factor analyses of critical-band-filtered speech. *Scientific Reports, 7:42468.*

UD Japanese-BCCWJ の構築と分析

大村 舞 (国立国語研究所コーパス開発センター) *
浅原 正幸 (国立国語研究所コーパス開発センター)

Construction and Analysis of UD Japanese-BCCWJ

Mai Omura (National Institute for Japanese Language and Linguistics)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

自然言語処理の分野では多言語かつ言語横断的な言語研究が盛んに取り組まれている。その言語横断的な言語研究の取り組みとして Universal Dependencies (UD) がある。本論文では、日本語のコーパスである UD Japanese-BCCWJ について紹介をする。UD Japanese-BCCWJ は現代日本語書き言葉均衡コーパス (BCCWJ) に付随する係り受け情報などを組み合わせて、UD へと変換、構築した BCCWJ の Universal Dependencie である。これは日本語の UD の中でも 1980 文章、57,256 文、約 126 万単語を含む最大規模また複数のレジスターを内包したデータセットである。UD Japanese-BCCWJ の特徴について説明する。また UD Japanese-BCCWJ の構築手順について説明し、現状における問題点について議論する。

1. はじめに

Universal Dependencies (以下 UD) (Zeman et al. 2017) とは、多言語で一貫した構文構造とタグセットを定義し、言語間での共通した依存構造タグ付きコーパスを提供することを目的としたプロジェクト及びそのコーパス、枠組みのことを指す。我々は日本語版 UD を設計する活動として、日本語コーパスに対する品詞体系、ラベル付き依存構造の定義の策定、その Github 上での文書化と、参照用のコーパスの作成に着手している。

2018 年 7 月現在日本語版 UD では表 1 のように 5 種類の UD が公開されている (この表は文献 (Asahara et al. 2018) を参照して作成した)。日本語ウィキペディアから構築した **UD Japanese-GSD**、他言語間パラレルコーパスから構築された **UD Japanese-PUD** (Zeman et al. 2017)、Kaede treebank (Tanaka and Nagata 2013) から変換して構築した **UD Japanese-KTC** (Tanaka et al. 2016)、さらに「日本語歴史コーパス明治・大正編 I 雑誌 (CHJ) (Ogiso et al. 2017)」から構築した **UD Japanese-Modern** (Omura et al. 2017)、そして本稿で説明する **UD Japanese-BCCWJ** が公開済みである。

本稿ではこの UD 日本語版設計の活動の一環として、現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa et al. 2014) に基いて構築された日本語 UD コーパス **UD Japanese-BCCWJ** について紹介する。UD Japanese-BCCWJ は他の日本語版 UD コーパスよりも大規模

* mai-om@ninjal.ac.jp

表 1　公開されている UD Japanese の一覧 (2018 年 7 月執筆時点)。

ツリーバンク	単語数	バージョン	Copyright	媒体
UD Japanese-BCCWJ	**1273k**	v2.2	内容分離	新聞、書籍、雑誌、ブログ etc.
UD Japanese-KTC	189k	v1.2	内容分離	新聞
UD Japanese-GSD	186k	v2.1	CC-BY-NC-SA	ウィキペディア
UD Japanese-PUD	26k	v2.1	CC-BY-SA	ウィキペディアのパラレルコーパス
UD Japanese-Modern	14k	v2.2	CC-BY-NC-SA	19 世紀の雑誌 (Ogiso et al. 2017)

で、また UD 上で公開されているコーパスの中でも、2 番目に大規模でかつ [(1)]、表 2 で示すような 6 種類のドメインのテキストで構成されたコーパスである。

本稿では UD Japanese-BCCWJ の構築、つまり、BCCWJ から UD の統語構造に変換する手順について説明していく。図 1 に BCCWJ の係り受け構造から UD の単語間係り受け構造に変換する手順の概略を示す。BCCWJ と UD には、品詞体系の違い、係り受け構造と単語間係り受け構造といった違いがある。そのため、これらの違いを考慮して変換する必要がある。そのためには BCCWJ に収録されている形態論情報のみではなく、係り受け構造や、並列構造の情報 (Asahara and Matsumoto 2016)、述語項構造情報 (植田ほか 2015) などを用いる必要がある。

日本語版 UD のプロジェクトでは BCCWJ から UD への変換を行ったことで、UD Japanese-BCCWJ を構築した。そして UD Japanese-BCCWJ や他の日本語版 UD を比較することで、日本語における統語構造と UD における統語表現の違いを比較、評価し、それらの結果についてプロジェクト内で議論を行っている。その結果を対外報告することで、UD プロジェクトに UD のフレームワークについて提言し、日本語版 UD のフレームワークの検討・改善に取り組んでいる。そこで本稿では UD Japanese-BCCWJ において問題となった点も取り上げていく。

2.　日本語における統語構造データと Universal Dependencies

表 2 に日本語版 UD の一覧を示している。現在、UD Japanese-BCCWJ を加えたことで、日本語版 UD は全 UD 内でも 2 番目に大規模な UD コーパスとなっている。公開されているコーパスとして **UD Japanese-KTC** (Tanaka et al. 2016)、**UD Japanese-GSD**、**UD Japanese-PUD** (Zeman et al. 2017)、**UD Japanese-Modern** (Omura et al. 2017) が存在する。これらの方針としては、既存の日本語統語データを用い、UD のフォーマットに自動変換することで低コストで日本語版 UD の構築を実現している。

UD 以外の、存在している日本語の統語構造コーパスには、京都大学テキストコーパス (Kurohashi and Nagao 2003) 、日本語係り受けコーパス (Mori et al. 2014)、Kaede treebank (Tanaka and Nagata 2013) などが存在する。これらのコーパスに共通していることとして、日本語の文節係り受け構造を元にして構築されていることが挙げられる。文節係り受け構造では、文節と

(1) 2018 年 7 月現在 http://universaldependencies.org/ 調べ。最大規模のコーパスはチェコ語の UD Czech-PDT である。

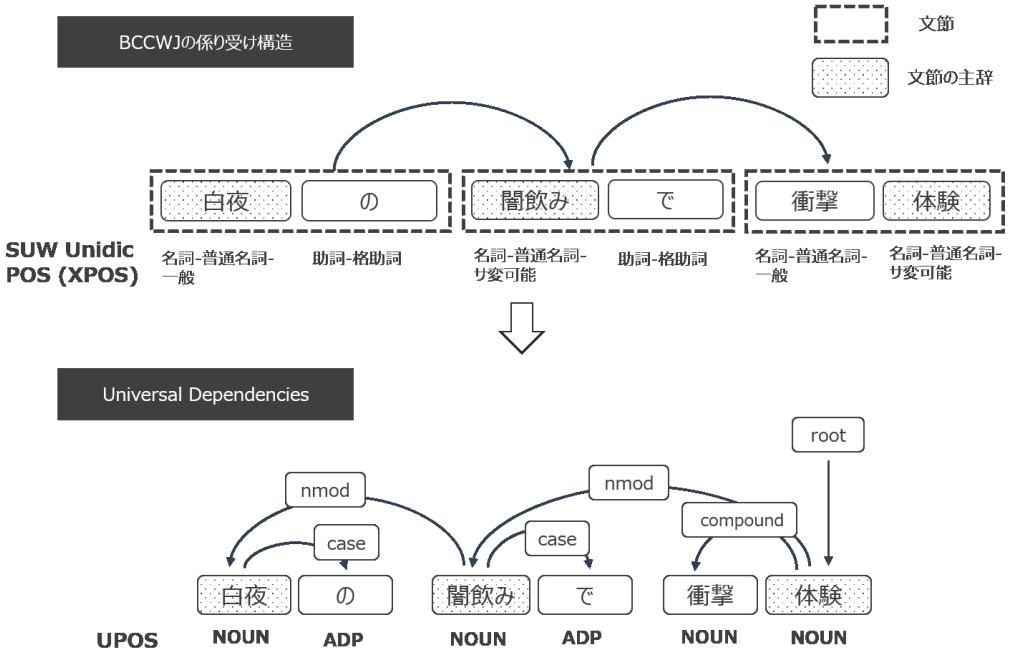


図 1　BCCWJ から UD Japanese-BCCWJ への変換の概要 (サンプルは PB_00001 から)。上の例が BCCWJ、下の例が UD Japanese-BCCWJ を表現している。

いう単語のグループ[(2)]を構成し、文節間の係り関係を記述する形で表現された統語構造であり、図 1 の上部図のような統語構造を持っている。UD Japanese-BCCWJ の基となる現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa et al. 2014) においてもこのような係り受け構造で統語構造を表現している。

一方 Universal Dependencies (UD) では、語順が自由な言語も含めて言語横断的に共通化した体系を確立するために、句構造を考慮せず、すべての構文構造を単語間の係り関係とその係り関係のラベルで表現する。異なる言語間で係り受け構造解析器の性能比較を行うだけでなく、言語学的に類型論的な分析が可能にすべく言語横断的な設計を目指している。そのため図 1 の下部図のような、内容語間の係り受け構造を中心とした表現を採用している。

3.　現代日本語書き言葉均衡コーパス (BCCWJ)

現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa et al. 2014) は、1 億 430 万語のデータを格納した、現在、日本語について入手可能な唯一の均衡コーパスである。サンプルの幅についても、書籍全般、雑誌全般、新聞、白書、ブログ、 ネット掲示板、教科書、法律などいった多領域のジャンル [(3)]が収録されている。

すべての収録サンプルは自動形態素解析によって言語単位、品詞付与が施されている。それぞれのサンプルは 2 種類の形態素、短単位 (Short Unit Word, SUW) と長単位 (Long Unit Word,

(2)　例えば図 1 の場合「白夜/の」「闇のみ/で」「衝撃/体験」が文節である。

(3) BCCWJ においてはこれをレジスターと呼んでいるがここでは他言語の UD とも比較するためジャンルという語で統一する。

表 2　BCCWJ のコアデータに収録されているレジスターの略称一覧

略称	説明
OC	Yahoo!知恵袋
OW	白書
OY	Yahoo!ブログ
PB	書籍
PM	雑誌
PN	新聞

表 3　BCCWJ コアデータのジャンルの分布。略称は表 2 を参照のこと。

ジャンル		OC	OW	OY	PB	PM	PN	合計
文章数	train	421	45	214	58	63	286	1,087
	dev	259	9	129	13	12	27	449
	test	258	8	128	12	11	27	444
	total	938	62	471	83	86	340	1,980
文数	train	2,838	4,456	3,278	7,196	9,546	13,487	40,801
	dev	1,650	780	1,920	1,131	1,510	1,436	8,427
	test	1,619	589	1,722	1,351	1,486	1,114	7,881
	total	6,107	5,825	6,920	9,678	12,542	16,037	57,109
単語数	train	50,415	168,909	51,310	174,394	177,947	300,786	923,761
	dev	29,961	31,471	32,164	27,315	30,328	29,528	180,767
	test	29,624	26,421	28,485	29,612	28,183	26,434	168,759
	total	110,000	226,801	111,959	231,321	236,458	356,748	1,273,287

LUW) という言語単位で解析されてそれぞれ公開されている。短単位は日本語の形態的側面に着目した規定した単位であり、語種ごとに規定した最小単位の線形結合に基づき定義されている。長単位は日本語の構文的な機能に着目して規定した単位であり、文節の構成要素ともなっている。

さらにこれらのデータに対して、BCCWJ の中 1% のサンプルは人手によって解析の誤りを修正されている。この修正されたデータを「コアデータ」と呼ぶ。BCCWJ のコアデータは 1980 文書、57,256 文が収録されており、UD Japanese-BCCWJ はこのコアデータを元に変換している。表 2 に BCCWJ のコアデータに収録されているジャンルの略称の一覧を示し、表 3 に BCCWJ のコアデータの統計を示す。

BCCWJ ではさらに、文節レベルの係り受け構造の情報を BCCWJ-DepPara (Asahara and Matsumoto 2016) で提供している。BCCWJ-DepPara には文節という単語単位のレイヤー情報、文節同士の係り関係の情報、単語間の並列関係の情報などが収録されている。また、BCCWJ-PAS (植田ほか 2015) によって、述語に対する格関係情報を記述した述語項構造という情報も提供されている。述語項構造は UD 関係ラベルを付与する際に参照している。UD Japanese-BCCWJ では形態素の情報、係り受け構造、述語構造などの情報を用いて UD への変換を試みている。

魚フライを食べたかもしれないペルシャ猫
"the Persian cat that may have eaten fried fish"

SUW	魚 NOUN *fish*	フライ NOUN *fry*	を ADP *−ACC*	食べ VERB *eat*	た AUX *−PAST*	か PART	も ADP	しれ VERB *know*	ない AUX *−NEG*	ペルシャ PROPN *Persia*	猫 NOUN *cat*
LUW	魚フライ NOUN *fried fish*		を ADP *−ACC*	食べ VERB *eat*	た AUX *−PAST*	かもしれない AUX *may*				ペルシャ猫 NOUN *Persia cat*	
bunsetsu	魚フライを			食べたかもしれない						ペルシャ猫	

図 2　短単位 (SUW)、長単位 (LUW)、文節の違いを表した例。

4.　BCCWJ から UD への変換手順

図 1 で分かるとおり、BCCWJ と UD の統語構造には違いがある。ひとつは、BCCWJ で使われている品詞体系 UniDic (伝ほか 2007) と UD で採用されている品詞体系 Universal POS(UPOS) (Petrov et al. 2012) とで異なるという点である。そして、BCCWJ は文節係り受けという文節単位の係り受け構造を採用しているのに対し、UD の統語構造は単語間の係り受け構造が要求されている。そして、UD では単語間に 37 種類ものある Universal Dependency Relations (ここでは依存関係ラベルと呼ぶ) という係り関係のラベルを付与する必要があるが、BCCWJ で用いる係り受け構造の情報にはここまで厳密に設定されていない [(4)]。そのため、これらの違いを考慮して変換する必要がある。本稿では、以下の手順で自動的に変換を試みた。

1. 単語単位を認定する。
2. UniDic の品詞体系 UPOS に変換する。
3. 文節係り受け構造を単語間依存構造に変換する。
4. 依存関係ラベルを付与する

それぞれの手順について、以降の節で説明する。

4.1　単語単位の認定

日本語は英語と異なり、単語の区切りが明示的に示されているわけではない。そのため、日本語版 UD における単語を決める必要がある。UD のガイドラインによると「統語的な単語 (syntactic words)」を単語として認定することが求められている。

前述のとおり、BCCWJ には短単位と長単位という言語単位が制定されている。また長単位を組み合わせた文節という単位も制定されている。文節は係り受け構造の単語単位にもなっている。そこで短単位、長単位、文節いずれかあるいはそれらを組み合わせた言語単位を UD で求められている単語とすることにした。図 2 に短単位、長単位、文節の例をあげている。単語認定について考えると、図 2 を例にした場合、例えば「魚フライを」という句は、短単位は「魚/フライ/を」の 3 つの単語に長単位は「魚フライ/を」という 2 つの単語に、そして文節は「魚フライを」という 1 つの単語となる。例から分かるように、短単位、長単位、文節には「短

(4) UD Japanese-BCCWJ で用いる文節係り受け構造の情報 BCCWJ-DepPara には、単語同士係り関係にあるか、並列構造にあるかなどの情報が付与されている。

表 4　Universal PoS version 2.0 (UPOS) の変換規則の一部。さらに具体的なものは (大村・浅原 2017) にも掲載している。

短単位の品詞	短単位基本形	長単位の用例	UPOS
^形容詞-非自立可能		形容詞-一般	AUX
^形容詞-非自立可能		助動詞	ADJ
^名詞-普通名詞-サ変可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-サ変可能		動詞-一般	VERB
^連体詞	^[こそあど此其彼] の		DET
^連体詞	^[こそあど此其彼]		PRON
^動詞-非自立可能	為る		AUX
^動詞			VERB
^名詞-固有名詞			PROPN
^名詞-普通名詞-副詞可能		副詞	ADV
^名詞-普通名詞-副詞可能			NOUN
接頭辞			NOUN
接尾辞			NOUN[(5)]

単位 <= 長単位 <= 文節」という階層関係があることがわかる。また後の 4.2 節でも述べるとおり、短単位と長単位ではそれぞれ異なった品詞体系を持っている。

UD Japanese-BCCWJ では短単位を統語的な単語として認定することにした。これは BCCWJ においては最小で基本的な言語単位、品詞体系を有している。ただし、後の節で説明するとおり、長単位のほうが求められている統語的な単語として、あるいは他言語比較の観点からして合っている可能性が高い。詳しくは 6.1 節にて説明する。

4.2　品詞の変換

UD では品詞体系として Universal PoS version 2.0 (UPOS) (Petrov et al. 2012) が採用されている。これらは多くの言語を定義するための 17 種類の品詞が制定されている。日本語版 UD でもこの UPOS を付与するために、BCCWJ で採用されている UniDic (伝ほか 2007) 品詞体系という品詞から UPOS に変換することで品詞の変換を実現する。

前述したとおり、この UniDic の品詞体系は短単位、長単位で異なっている。BCCWJ における短単位では**語彙主義**的な可能性に基づく品詞体系を採用している。例えば「名詞-普通名詞-副詞可能」は「名詞」用法も「副詞」用法もある語彙であることを意味する。長単位では文脈に基づいてこの用法の曖昧性を解消する**用法主義**に基づく品詞を規定している。さらに短単位に対して、長単位を参照して長単位形態論情報として「用法」の情報が付与されている。短単位を単語として採用したため、品詞体系も短単位の語彙主義的な可能性に基づく品詞体系を採用する。

しかし、UD の品詞体系の標準にあわせる、あるいは他言語同士の比較をするという観点からすると長単位の用法主義に基づく品詞が求められる。例えば「する」を付与することで動詞化する「名詞-普通名詞-サ変可能」という品詞、「な」を付与することで形容詞化する「名詞-普通名詞-形状詞可能」という品詞が短単位の品詞体系には存在する。しかし、長単位の品詞体系であった場合、長単位は動詞であれば「XX する」のような言語単位が 1 つで構成され、これは確実に「動詞」であることが確定する。

(5) 日本語における接尾辞の品詞体系には「接尾辞」と書かれていても機能的なものから名詞的なものと幅があるため一概に NOUN を付与するのには議論の余地がある。現状 NOUN を付与することとする。

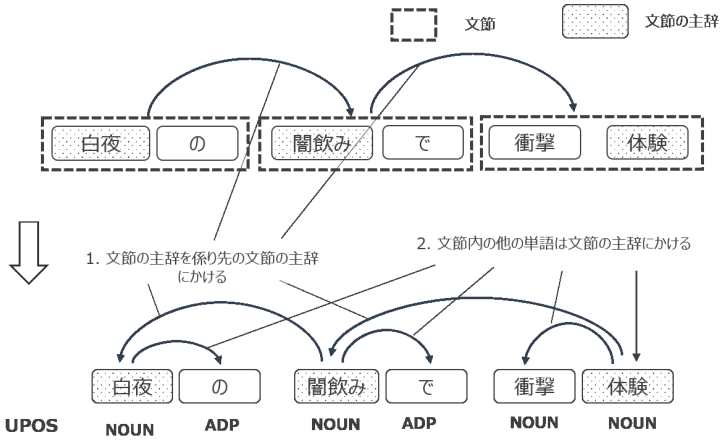


図 3　文節係り受けから UD の単語間係り受けへの変換の概略図

表 4 に UniDic 短単位の品詞体系から UPOS へ変換する規則を示す。表 4 で示している変換規則は短単位の品詞体系に基づいて変換しており、6.1 節で議論する通り、長単位で採用されている用法主義に基づく品詞体系を採用した場合さらに規則は単純になる。しかし、いくつかの理由により現状では用法主義に基づく品詞体系は採用していないものの、いずれ公開予定である。理由についても 6.1 節で説明する。

4.3　統語構造の変換

UD における単語間依存構造を得るために、日本語の統語構造である文節間係り受け構造を用いて変換する。BCCWJ には BCCWJ-DepPara (Asahara and Matsumoto 2016) という文節間係り受け構造・並列構造の情報が提供されている。BCCWJ-DepPara には文節の情報、係り受け関係の情報が収録されている。

BCCWJ の文節係り受け構造から UD の単語間係り受け構造に変換するために、文節間の係り関係のみではなく、それ以外の単語間でも係り情報を加える必要がある。BCCWJ-DepPara には文節の他にも「文節の主辞」(図 3 の網掛け部分) が設定されている。そこで図 3 のように、1. 文節の主辞同士でまず係り関係を結び、そして、2. それ以外の文節内単語に関しては文節の主辞にかける、という手順で文節係り構造から単語間係り構造に変換する。このとき、日本語の係り受け構造の場合、矢印は「係り元」から「係り先」にかかるような向きで表現するが、UD の場合矢印の向きが逆、つまり「係り先」から「係り元」に矢印が向く図になることに注意すること(6)。

日本語において文節の主辞は、図 3 の「衝撃/体験」の「体験」のように、文節の主辞は右側に置かれやすい傾向にある。これは日本語においては、主体となる名詞句は右側におき、補助的な要素は左側に置かれやすいからである。同様に日本語における文節間の係り関係は「左から右に」にかかりやすい。一方で、英語などの言語の場合「右から左」に向かう係り関係が存

(6) UD の単語間係り受け構造の図表現が「係り先」から「係り元」の方向になるだけで、後述のフォーマットのとおり、係り元の単語について、係り先を記述する形 (列 HEAD 参照) になっている。

表 5　依存関係ラベルの付与規則の一部。簡略的に書かれており実際の実装ではより詳細に設定されている。さらに具体的なものは (大村・浅原 2017) にも掲載している。ただし全ては掲載されていない。

ラベル付与ルール	ラベル
その係り元単語は係り先がなく (文末の文節である) でさらに文節の主辞である	root
その係り元単語は UPOSNUMMOD を持っている。	nummod
その係り元単語は UPOSADV を持っている	advmod
係り先単語は VERB を持っており、格助詞「が」が文節内にある	nsubj
係り先単語は VERB を持っており、格助詞「を」が文節内にある	obj
その係り元単語は UPOSVERB 持っており、その係り先単語は UPOSVERB を持っており、文節をまたがっている	aux
その係り元単語は UPOSVERB 持っており、その係り先単語は UPOSVERB を持っており、文節内の関係である	compound

在する場合がある。例えば並列表現の場合は、左に係り先をおいた表現を採用している。この違いが日本語版 UD における並列構造に影響を与えていることを 6.2 節にて議論する。

BCCWJ-DepPara には係り受け構造の情報や並列構造の情報は含まれているものの、UD で定義するように指定されている依存関係ラベル (Marneffe et al. 2014) のような詳細な係り関係の情報は含まれていない。依存関係ラベルには、例えば nsubj、obj、iobj、amod のような係り関係を定義するラベルが存在している(7)。そのため BCCWJ から用いることのできる情報などを利用して、単語間の係り関係に依存関係ラベルを付与する必要がある。表 5 に依存関係ラベルの付与規則の例をあげる。係り先単語について、文節の情報、格情報あるいは並列関係の情報などのを組み合わせることで依存関係ラベルを付与している。

nsubj、obj などのような統語構造の項は、格助詞などが (いわゆる助詞「が」「を」「に」など) 付与されているか否かで依存関係ラベルを付与する。UD の方針としては、あくまで統語構造を表現するものであるため、助詞の標識がある場合は、格標識に基づいて依存関係ラベルを付与する。しかし、日本語は英語とは異なり、必ずしも格標識「が」や「は」「を」などが文上の主体を表しているとは限らない。例えば「は」は通常であれば「私は学校に行く」と言ったとおり「私」が nsubj であるようにラベルを付与することができる。しかし、「象は鼻が長い」といった文の場合、「象」は Topic marker であるため、nsubj を付与すべきかどうかは不明瞭である。また「3 時に公園に行く」といったような文章だった場合、「に」という格助詞が衝突してしまう。この場合、BCCWJ-PAS (植田ほか 2015) の述語構造情報を参照する必要がある (8)。

なお現在のルールでは、 csubj、advcl、acl といった節に関するラベルを付与することができない。なぜならば、英語と比較して日本語は節かどうかの境界が曖昧だからである。節にかんしては 6.3 節にて議論をする。将来、この節の同定に関しても検討する必要がある。

BCCWJ-DepPara にはさらに、並列構造の情報が含まれており、並列の情報を用いて並列の情報 cc や conj を付与することになる。しかし、この並列構造情報を用いても、UD において

(7) 具体的な依存関係ラベルは http://universaldependencies.org/u/dep/index.html 参照。
(8) 日本語版 UD における格標識に関しては Asahara et al. (2018) の 3.4 節にて問題点を議論している。

```
# sent_id = OC01_00001-1
# text = 詰め将棋の本を買ってきました。
1 詰め 詰める VERB 動詞-一般 _ 2 aux _ BunsetuPosition=B|JPYomi=ツメル|BunsetuPositionType=CONT|SpaceAfter=No
2 将棋 将棋 NOUN 名詞-普通名詞-一般 _ 4 nmod _ BunsetuPosition=I|JPYomi=ショウギ|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
3 の の ADP 助詞-格助詞 _ 2 case _ BunsetuPosition=I|JPYomi=ノ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
4 本 本 NOUN 名詞-普通名詞-一般 _ 6 obj _ BunsetuPosition=B|JPYomi=ホン|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
5 を を ADP 助詞-格助詞 _ 4 case _ BunsetuPosition=I|JPYomi=ヲ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
6 買っ 買う VERB 動詞-一般 _ 8 advcl _ BunsetuPosition=B|JPYomi=カウ|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
7 て て SCONJ 助詞-接続助詞 _ 6 mark _ BunsetuPosition=I|JPYomi=テ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
8 き 来る VERB 動詞-非自立可能 _ 0 root _ BunsetuPosition=B|JPYomi=クル|BunsetuPositionType=ROOT|SpaceAfter=No
9 まし ます AUX 助動詞 _ 8 aux _ BunsetuPosition=I|JPYomi=マス|BunsetuPositionType=FUNC|SpaceAfter=No
10 た た AUX 助動詞 _ 8 aux _ BunsetuPosition=I|JPYomi=タ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
11 。 。 PUNCT 補助記号-句点 _ 8 punct _ BunsetuPosition=I|JPYomi=_|BunsetuPositionType=CONT|SpaceAfter=No

# sent_id = OC01_00001-2
# text = 駒と盤は持っていません。
1 駒 駒 NOUN 名詞-普通名詞-一般 _ 3 nmod _ BunsetuPosition=B|JPYomi=コマ|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
2 と と ADP 助詞-格助詞 _ 1 case _ BunsetuPosition=I|JPYomi=ト|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
3 盤 盤 NOUN 名詞-普通名詞-一般 _ 5 iobj _ BunsetuPosition=B|JPYomi=バン|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
4 は は ADP 助詞-係助詞 _ 3 case _ BunsetuPosition=I|JPYomi=ハ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
5 持っ 持つ VERB 動詞-一般 _ 0 root _ BunsetuPosition=B|JPYomi=モツ|BunsetuPositionType=ROOT|SpaceAfter=No
6 て て SCONJ 助詞-接続助詞 _ 5 mark _ BunsetuPosition=I|JPYomi=テ|BunsetuPositionType=FUNC|SpaceAfter=No
7 い 居る AUX 動詞-非自立可能 _ 5 aux _ BunsetuPosition=I|JPYomi=イル|BunsetuPositionType=FUNC|SpaceAfter=No
8 ませ ます AUX 助動詞 _ 5 aux _ BunsetuPosition=I|JPYomi=マス|BunsetuPositionType=FUNC|SpaceAfter=No
9 ん ず AUX 助動詞 Polarity=Neg 5 aux _ BunsetuPosition=I|JPYomi=ズ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
10 。 。 PUNCT 補助記号-句点 _ 5 punct _ BunsetuPosition=I|JPYomi=_|BunsetuPositionType=CONT|SpaceAfter=No
    ....
```

図 4　BCCWJ の UD サンプル (OC01_00001)。上記のようにタブ区切りのテキストファイルになる。

表 6　CoNLL-U 形式の各列の説明

列	フィールド名	説明
1	ID	1-origin の ID (ROOT が 0)
2	FORM	書字形出現形
3	LEMMA	語彙素読みをローマ字にしたもの
4	UPOSTAG	品詞 Universal POS
5	XPOSTAG	品詞 BCCWJ の短単位品詞
6	FEATS	その他品詞情報 (“|” で OR を表現、順不同)
7	HEAD	係り先 ID
8	DEPREL	依存関係ラベル
9	DEPS	Secondary Dependency (List, Head-deprel pairs)
10	MISC	その他 (表 7 参照)

表 7　UD Japanese-BCCWJ における MISC フィールドの項目の一覧

ラベル	説明
BunsetuBILabel	文節の開始か中間かを表現 (B=開始、I=中間)。
BunsetuPositionType	文節の種類
LUWBILabel	長単位の開始か中間かを表現 (B=開始、I=中間)。
LUWPOS	UniDic 長単位品詞体系

解決できない点が存在する。この問題は 6.2 節で議論する。

4.4　フォーマット

以上の節で説明した通りの手順を経て、UD Japanese-BCCWJ は図 4 のようなフォーマットに変換される。このフォーマットはタブ区切りの UTF-8 の文字コードでエンコードされた CoNLL-X フォーマットに基づいている。それぞれの項目については表 6 に説明している。

UD では MISC フィールドを用いることで、さまざまな情報を付与させることができる。そのため、統語構造の情報として重要と思われる情報、長単位の情報、文節の情報を付与させる予定である [9]。表 7 に UD Japanese-BCCWJ の MISC フィールドで付与される情報の項目について説明している。

[9] 現行で公開されているバージョンでは付与されていないが、開発版には付与する予定である。

表 8　単語間係り受け解析の結果 (評価指標 UAS)。

train \ test	OC	OW	OY	PB	PM	PN	all.
OC	89.70	81.99	88.46	87.93	88.45	87.21	**90.49**
OW	80.21	**88.62**	78.08	83.66	84.74	84.95	88.55
OY	86.35	79.54	86.15	84.62	85.67	84.66	**88.21**
PB	89.23	86.23	88.34	**91.56**	90.91	90.63	91.48
PM	87.28	85.57	86.64	89.65	**89.74**	89.32	89.67
PN	86.40	87.66	85.88	88.65	89.31	**91.20**	90.83
all.	86.64	84.84	85.71	87.74	88.18	88.00	**89.89**

5. ジャンルごとの係り受け構造解析

UD Japanese-BCCWJ では 6 種類ものジャンルについて比較的大規模な量の UD が提供される。他の UD でも複数のジャンル収録されて UD も公開されているが、ある程度の量、数千文単位で収録されているものは少ない。UD Japanese-BCCWJ のデータの規模について検討するために、実験として単語間係り受けの解析結果を示すことにする。本稿では形態素解析の結果は示さない。理由としては、既存の形態素解析 (例えば MeCab(Kudo et al. 2004)) を用いて UniDic 品詞体系に品詞を付与することが可能であり、さらに前述のとおり、Unidc 品詞体系から UPOS に変換するのは規則ベースで簡単に変換することができるからである。

単語間係り受け解析を行うツールとして UDPipe (Straka and Straková 2017) を用いた。UDPipe では UD コーパスを元にモデルを構築、解析結果を出力できるツールである。さらに構築したモデルを用いて、単語分割、タグ付け、見出語認定、そして係り受け解析を行うことができる。係り受け解析には Parsito (Straka et al. 2015) という手法が採用されており、これはニューラルネットワークを用いた手法である。使用した UDPipe のバージョンは 1.2.1-devel を使い、オプションはつけずにトレーニング、評価を行った。実際に用いた訓練、テストデータの量は表 3 に示した通りである。評価指標としては Unlabeled attachment score (UAS) を用いた。UAS は係り元単語の係り先が合っているかを計算し、その正解割合を出したものである。

表 8 に結果を示す。表の列はそのジャンルのみで構築したモデルを表しており、行がテストに用いたジャンルのデータを表現しており、'all' はすべてのデータを使った場合を表している。つまり表示されている値は、列のジャンルで訓練したモデルに対して行のテストデータで評価した結果を表現している。

表 8 をみてわかるとおり、OW、PB、PM、PN といった 200,000 単語以上収録されているジャンルにおいては、同一のジャンルのモデルで評価した結果が評価が最も高い。一方、量が比較的少ない OC、OY(100,000 単語程度のもの) はすべてのデータで学習したものの精度が高くなっていることが分かる。そのため、必ずしも大規模な文章量があれば精度が良くなるというわけではなく、ある程度規模があれば、同一のジャンルでトレーニングしたモデルの方が精度がよくなる、といった結果を確認することができた。UD Japanese-BCCWJ を用いることでこのように、量による違い、ジャンルによる違いでの比較を行うことができることが分かる。

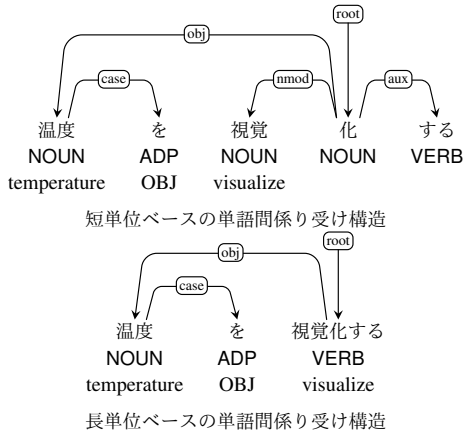


図 5　短単位と長単位の品詞体系による違いの例

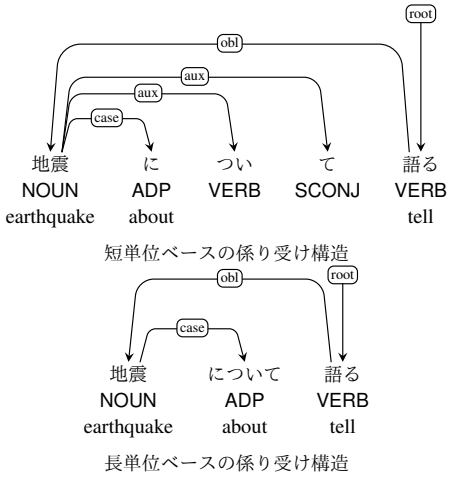


図 6　短単位と長単位における複単語表現の違い

6. 議論

この節では UD Japanese-BCCWJ において構築した際に検討する必要ができた内容などについて議論する。本稿では UD Japanese-BCCWJ を中心に説明してあり、日本語版 UD について全体的な議論については文献 (Asahara et al. 2018) で議論している。

6.1 単語認定単位について

UD における単語単位の認定は日本語版 UD において議論すべき問題のひとつである。前述の通り、BCCWJ で用いることができる単語単位には短単位、長単位、文節が存在する。現行の UD Japanese-BCCWJ では短単位を採用している。UD プロジェクトにおける単語とは、「統語的な単語 (syntactic word)」であると規定されている。UD Japanese-BCCWJ では短単位を採用しているものの、この統語的単語としては短単位よりも長単位の方が近いと考えられる。

例えば、短単位と長単位では品詞体系が異なり、これは長単位の方が syntactic word に合っている可能性がある。図 5 は短単位の場合と長単位の場合で UD にしたときの例である。短単位の場合、「可視/化/する」という語が 3 単語に分かれてしまい、それぞれ、NOUN、NOUN、VERB と UPOS をバラバラに与えられる。そのため、「可視化する」というフレーズが動詞であるかどうかを表現するのに係り関係を細かく設定する必要がでてくる。一方で長単位の場合、これは「可視化する」というひとつの単語になり、長単位は用法主義に基づく品詞であるため、「動詞」であると品詞体系からも確定する。

さらに図 6 のように、複単語表現「について」という表現も、短単位の場合は 3 つの単語で構成される一方で、長単位であればひとつにまとまってくれるため、機能語と名詞句との関係も簡素に表現できる。このように、元々長単位の品詞は構文に基づいて構成されているのもあり、UD の「統語的な単語」にあっていると考えられる。

しかし、現状は短単位を UD Japanese-BCCWJ では採用している。ひとつは長単位を厳密に解析できるツールがないこと、もうひとつの理由としては、複合表現の中でも、必ずし

も UD に合うような「統語的な単語」でない可能性があるためである。今後長単位で UD Japanese-BCCWJ を構築することで、これらの問題について検討する必要があるだろう。

6.2 並列構造

並列構造も UD、特に日本語や韓国語などで問題になっている。理由は 2 つあり、1 つ目の理由としては、日本語は主辞を右側に置く言語であるのに対して、英語は主辞となる句を左に置く言語であるため、並列構造のルールに反してしまう、という点である。2 つ目の理由として、例えば conj は名詞句の並列の並列を表現しており、名詞並列句であるか否かを考えなくてはいけないものの、UD Japanese-BCCWJ の場合、名詞並列句であるか、動詞並列句であるかの情報がない、という点である。

例えば、「と」という接続表現がある。 基本的には、英語でいう “with” の意味合いだと考えられるだろう。この with の意味合いの場合、UD では図 7 の上記の例のように nmod を付与する。しかし、必ずしもこの「と」が “with” の意味合いであるとは限らない。例えば、図 7 の中間の例のような「パンとジャム」の場合、「パンに（つける）ジャム」という意味合いが考えられるため、この「と」という接続表現は「with」の意味合いと考え nmod になると考えられる。一方で「パンとごはん」の場合、「ごはん」と「パン」とを並列に並べているだけである、と考えられるためこれは並列表現であるとみなし conj でつなげるべきである。しかしこの区別をするための情報は BCCWJ において付与されていないため、nmod でつなぐ表現であるのか、conj でつなぐ表現であるのかの区別が難しい。

また、前述のとおり日本語は「左から右にかかる」右主辞傾向の言語である。一方で英語、UD における基準では「右から左にかかる」左主辞傾向の言語である。そのため、UD の規定に従うならば図 7 の中間の例のような表現にする必要がある。しかし、現状の手順では図 7 の下部の図のような表現になってしまい、UD の規定に反してしまう。そのため左主辞の構造への変換という手順が必要となり、実直に実装することが難しいと言えるだろう。

6.3 節 (Clause)

UD の依存関係ラベルでは単語と句、節を分けるようにデザインされている。しかし、日本語では、単語、句、節との境界が曖昧である。なぜならば、日本語の文には主語も含めて、必ずしも明示的な格要素を書く必要がないためである。

図 8 に日本語における節と形容詞節の例をあげる。図 8 の上の例は名詞主題がついた形容詞節である。しかし、下の例は形容詞は修飾しているのか、叙述的であるのかが断定できない。なぜならば、日本語では、名詞叙述形容詞の名詞主題は省略できるからである。図 8 の一番下の例の場合、おそらく「しっぽ」などが補われると考えられるが、全体的に赤い猫である可能性もあるだろう。いずれであるかは、文脈から判断するしかない。このように、単純な修飾か、形容詞節であるかの区別は現状難しいため、すべての名詞句につく形容詞には acl を付与している。

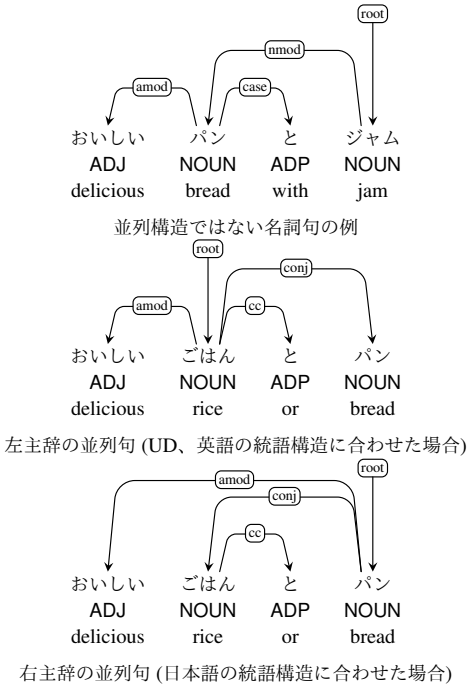


図 7　日本語における名詞句の並列構造

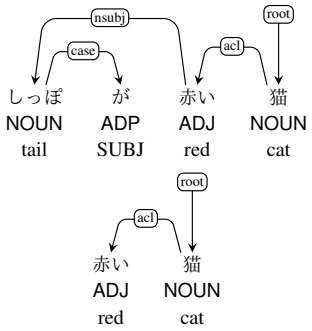


図 8　日本語における節と句の違い

7. まとめと今後の展望

本稿では現代日本語書き言葉均衡コーパス (BCCWJ) から Universal Dependencies(UD) のフレームワークに変換した UD Japanese-BCCWJ を構築した。そして、BCCWJ と UD の違いに触れ、その構築手順や特徴について説明した。UD Japanese-BCCWJ は 2018 年 4 月に公開されている [(10)]。

しかし本稿で議論したように、UD Japanese-BCCWJ あるいは日本語版 UD において検討しなくてはいけない問題点が存在する。例えば単語の単位認定が短単位であるのは UD の統語的な単語単位としてふさわしいとは言い難いため、長単位などの別の単語単位のコーパスも用意する必要があるだろう。

それぞれの日本語版 UD では、基としているコーパスが異なるために、品詞体系などの違いから、ルールがそれぞれ異なってしまっている。例えば、UD Japanese-KTC は句構造ツリーバンクから構築されており、BCCWJ の係り受け構造から変換されたものではない。そこで、今後は日本語 UD において、なるべく同一のルールで構築できるように、UD Japanese-BCCWJ で用いたルールに従って構築できるように調整を行いたいと考えている。これにより日本語 UD 間でのコーパスの差異を減らすことができると考えられる。

謝　辞

(10) http://universaldependencies.org/ にて UD Japanese-BCCWJ として配布されている。また BCCWJ の中納言アカウントを持っている場合、https://bccwj-data.ninjal.ac.jp/mdl にて変換済みのデータをダウンロードすることができる。

本研究（の一部）は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。

文　献

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Hěctor Martínez Alonso, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li (2017). “CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki (2018). “Universal Dependencies Version 2 for Japanese.” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1824–1831. Miyazaki, Japan.

Takaaki Tanaka, and Masaaki Nagata (2013). “Constructing a Practical Constituent Parser from a Japanese Treebank with Function Labels.” *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL’2013)*, pp. 108–118. Seattle, Washington, USA.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto (2016). “Universal Dependencies for Japanese.” *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1651–1658.

Toshinobu Ogiso, Asuko Kondo, Yoko Mabuchi, and Noriko Hattori (2017). “Construction of the ‘Corpus of Historical Japanese: Meiji-Taisho Series I - Magazines’.” *Proceedings of the 2017 Conference of Digital Humanities (DH2017)*. Montrèal, Canada.

Mai Omura, Yuta Takahashi, and Masayuki Asahara (2017). “Universal Dependency for Modern Japanese.” *Proceedings of the 7th Conference of Japanese Association for Digital Humanities (JADH2017)*, pp. 34–36.

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

48:2, pp. 345–371.

Masayuki Asahara, and Yuji Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annotation Treebank on the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58. Osaka, Japan.

植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸 (2015).「『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション」 第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214.

Sadao Kurohashi, and Makoto Nagao (2003). *Building a Japanese Parsed Corpus – while Improving the Parsing System.*, Chap. 14 pp. 249–260. *Treebanks: Building and Using Parsed Corpora.*: Springer, Dordrecht.

Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada (2014). “A Japanese Word Dependency Corpus.” *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 753–758. Reykjavik, Iceland.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007).『コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用』 国書刊行会, pp. 101–123.

Slav Petrov, Dipanjan Das, and Ryan McDonald (2012). “A universal part-of-speech tagset.” *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC2012)*, pp. 2089–2096.

大村舞・浅原正幸 (2017).「現代日本語書き言葉均衡コーパスの Universal Dependencies」 言語資源活用ワークショップ発表論文集, pp. 133–143.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning (2014). “Universal Stanford Dependencies: A cross-linguistic typology.” *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 4585–4592. Reykjavik, Iceland.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto (2004). “Applying conditional random fields to Japanese morphological analysis.” *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Barcelona, Spain.

Milan Straka, and Jana Straková (2017). “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe.” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99. Vancouver, Canada.

Milan Straka, Jan Hajič, Jana Straková, and Jan Hajič jr. (2015). “Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle.” *Proceedings of Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*.

LINE[1]データベースの設計と属性情報付与の現状について

宮嵜 由美（国立国語研究所音声言語研究領域）

Fundamental Planning of LINE Database and Participant's Information

Yumi Miyazaki (National Institute for Japanese Language and Linguistics)

要旨

本稿では，現在構築中の「LINE データベース」の設計と現状について，①データ収集方法，②データ提供者と参加者の属性，③研究用データベースとしての加工を，具体例とともに報告した。2016 年から収集を始めた本 LINE データベースへの協力者は，2018 年 6 月時点で延べ 183 名，約 35,800 行[2]のデータである。

１．はじめに

本稿では，2016 年 4 月から 2018 年 6 月現在まで，筆者が収集した LINE のデータのデータベース構築の基本設計を紹介する。LINE は，文字メッセージの送受信の他に，通話，テレビ電話などの機能も有しており，従来のコミュニケーションツールでの言語生活の多くの部分を兼務する。研究対象としても，再現性を確保したデータベース構築の為，また機械的な検索も可能となるデータベース構築の為にどのような手続きが必要であるか，提案する。

２．データ提供の依頼手順と収集方法

２.１　データ提供の依頼手順

LINE データベース構築にあたり，まず，筆者の担当する大学の講義受講生数人にデータ提供依頼を行った。そのうち，データ提供を許諾した数人から，さらに提供者の紹介を受ける方法でデータ提供者を募った。この方法により，「ある人物が形成するコミュニケーションネットワーク」と「言語生活」の一端をうかがい知ることができると考えた。

なお，本稿では，実際に筆者とコンタクトを取りデータ提供を行った，直接のデータ提供元となる協力者を「データ提供者」と呼び，提供されたデータに登場する参加者については，「参加者」と呼ぶ。「参加者」のデータ提供許諾と匿名加工の範囲については，データ提供者から参加者に個別に確認してもらった。

知り合いの知り合い同士が知り合い，という事もあり，同一人物同士がそれぞれの LINE コミュニティにおいて，例えば二者間の場合と，三者以上が同じ画面（以下，トークルーム）でやり取りに参加する LINE 内（以下，グループ LINE）ではどのように互いを待遇しているか，違いは生じているのか，いないのか，その様子を観察する事もできる。

例えば，図１に示したデータ提供者 A の例をみると，データ提供者 A が参加するグループ LINE の構成員の一人とは，別のトークルームでの LINE のやりとりを行っているというケースがみられる。さらには，収集したデータの一部には提供者 A の母親との LINE デ

---

[1] 「LINE」は株式会社 LINE の商標，または登録商標です。

[2] 本稿では改行をもって１行とし，データも改行毎に１セルに入力されている。

miyazakiyumi@gmail.com

ータを含むが，その母親は息子との LINE の他，母親がひとりの女性として参加している友人 A との LINE データ提供があるというケースが存在する。それぞれが，それぞれの場（トークルーム）内で振舞いが変わるか，変わらないのか，という視点からの分析も可能である。

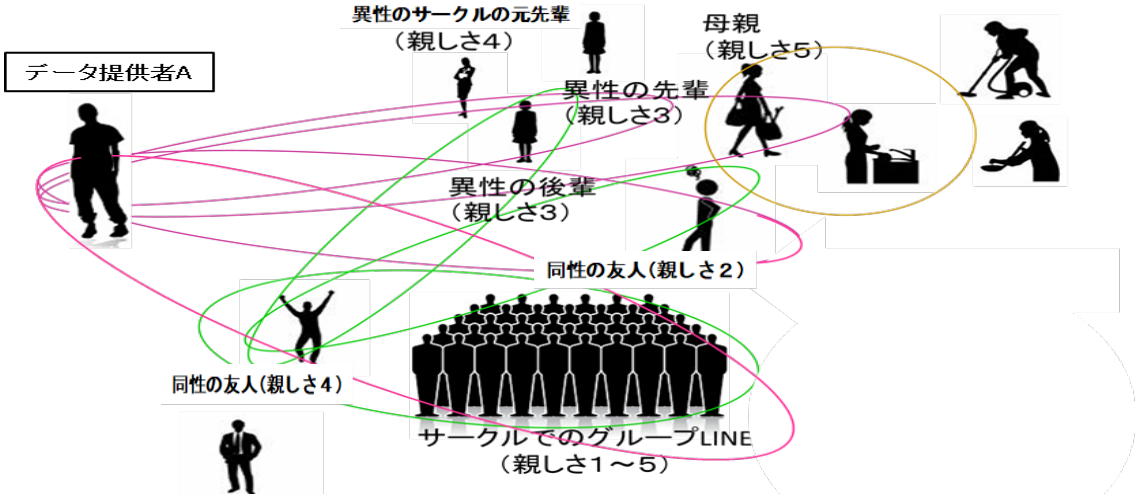


図１：LINE データ上のコミュニケーションネットワークの例（データ提供者 A の場合）

２.２　データ収集方法

LINE のコミュニケーションツールとしての特徴については三宅（2018）に詳しいが，今回データベースを整備するにあたり検討すべき点として，受信媒体の違いで文字化けが発生しないスタンプ[3]，画像，動画，通話機能などが文字メッセージとともに同じトークルーム内で連続性を持って送受信できるその様子を，検索可能な状態にどう再現するかという点にある。

収集したデータにも，スタンプのみの会話や，文字メッセージの送受信から突然音声通話に変わる，待ち合わせの場所は「位置情報」システムを使って教え合う，メッセージの送受信にどれだけの時間がかかったかといった既読・未読に関わる問題など，文字メッセージ以外の要素の再現すべき点は多々あり，それらも機械的な検索ができる限り可能となるよう考慮する必要がある。

これらの問題を勘案し，基本的には，LINE のデータテキスト化機能を使い，提供者任意の部分のみの①テキストデータを提供してもらうよう依頼した。さらに該当する部分の画面の②スクリーンショット画像を同時に提供するよう依頼した。

２.３　データ提供者の属性（2016 年 4 月～2018 年 6 月末時点）

【フェイスシート】

フェイスシートの質問項目は以下となっている。

Ａ．データ提供者自身に関する項目

１）名前・LINE 登録名・性別・年齢・職業・LINE 使用歴・出身地

Ｂ．LINE データ参加者に関する項目

１）名前・性別・年齢・職業・関係性・親しさ（データ提供者による記入）

[3] スタンプとは「LINE」で使用されるステッカーのような画像データを指す。

２）グループ LINE の場合はその「グループ名」と参加者全員の「LINE 登録名」

B, １）に示した親しさの判定については，データ提供者の視点から，５段階評価（とても親しい～全く親しくない）で評価されたものである。相手をどう認識し，待遇するかという問題であるが，この親しさの尺度以外にも，提供者と参加者の関係性の情報（サークルの同期や年上であっても職場の同僚など）も確認しており，この尺度はあくまで相手の待遇に関わる一つの要因として捉えていただきたい。

対面以外では携帯メールが主なコミュニケーションツールであった時代と同様（宮嵜：2004），親しい友人としか LINE はしない，という意見も聞かれた。しかし，実際にデータ収集を行ってみると，グループ LINE と呼ばれる複数人が同じトークルームに参加する場合などには，親しくないと判断される人物とのやり取りも少なくない（後述のグラフ１参照）。また，二者間の LINE でも，親しくない相手とのやり取りも行われており，今回のデータに含まれる。

２．４　データ提供者・参加者の属性：属性の流動性

2018 年６月末時点で整備が進んでいるデータ提供者・参加者の年齢と性別は表１の通りである。前述の通り，起点としたデータ提供者が友人同士という場合もあり，LINE コミュニティの形成にも重複がみられる。つまり，同一人物が複数の LINE に参加している場合があり，表１は延べ人数を示す。年齢は提供された LINE データ送受信時のものである。

さらに，データ提供者・参加者の職業と延べ人数を表２に示す。職業についても，LINE 送受信時のものである。同一人物が，それぞれの LINE データ送信時に別の職業に移行したケースがあり，表１の合計とは一致しない。

グラフ１に示したのは，データ提供者からみた参加者との親しさである。データ提供者の判断による協力者との親しさであり，データ提供者が異なると，同一人物であっても，データ提供者の評価によっては親しさが異なる場合もある。

表１ データ提供者・参加者の年齢と性別（人）

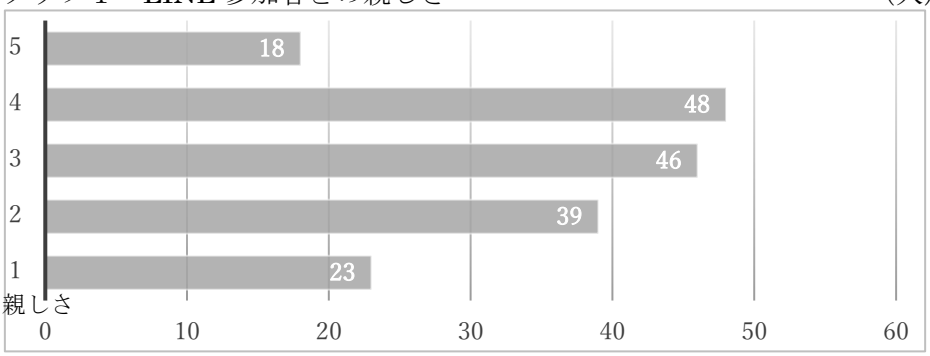
	男性	女性	合計
19 歳	6	4	10
20 歳	28	32	60
21 歳	30	20	50
22 歳	20	8	28
23 歳	5	4	9
24 歳	1	1	2
25 歳	1	0	1
54 歳	0	1	1
55 歳	0	5	5
56 歳	0	1	1
57 歳	0	5	5
58 歳	0	10	10
59 歳	0	1	1

データ送受信時の年齢は，筆者が大学生へのデータ提供依頼開始時に起点としたことから，現時点では大学生学部生に該当する年齢と，その親の世代の，大きく分けて２つの年齢層が存在するデータとなった。

表２ データ提供者・参加者の職業 （人）

学生	137
会社員	16
主婦／パート	9
アルバイト	8
専業主婦	6
介護士	3
教員	2
主婦／歯科衛生士	1
会社役員	1
教会牧師	1

グラフ１ LINE 参加者との親しさ （人）



３．研究用データベースとしての加工について

本調査では，LINE に備わっている「トーク履歴の送信」機能を使用し，文字列情報を忠実に再現できる収集方法を取った[4]。そこから研究用として匿名性やデータベースとしての検索性を考慮し，以下の加工方針に沿った xlsx 形式のデータベースを作成した。

また，スタンプや画像等の参照の為，本文と対応するトークルーム画像に対し，個人や場所等を特定できる部分を匿名加工した画面がポップアップされるようにした。

図２に匿名加工後のデータベースの例を示し，以下，入力規則について述べる。

３．１ データに付与される属性情報

＜データ属性＞

当該のトークルームの参加者の構成を示す。

① 二者間 LINE，10 代・20 代女性同士の例：ＦＦ １２００３

[4] 提供されたデータの一部には，画像データのみの提供もあり，手作業でのテキスト化を行ったデータも部分的に含んでいる。

② グループ LINE，50 代女性同士の例：G　F F　５５　００１
（実際のデータに空白は含まない）

左から，提供者が男性の場合には「M」女性には「F」を付与。冒頭に G が付く場合はグループ LINE である事を示し，次に提供者の性別・協力者の性別（グループ LINE の場合は参加者の性別構成）を示す。

数字の一桁目は提供者，二桁目は参加者の，それぞれ提供された LINE を送受信した時期の年齢の最小値と最大値を示す。下三桁はデータベース全体におけるトークルームの通番を示す。

よって，①の場合は，10 代女性と 20 代女性間のトーク履歴であることを示し，②の場合は，50 代女性間のグループラインのトーク履歴であることを示す。

図２ LINE データベース加工例

	データ属性	通番	管理ID	性別	年齢	職業	関係	親しさ	送信日	送信時間	吹き出し内行数	本文
2	FF12003	1	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:37		<Lf003>ですお久しぶりです
3	FF12003	2	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:37		ブロックしないでね(ノω`)
4	FF12003	3	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:41		<Lf003>ーーーーーー
5	FF12003	4	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:41	1	<人名>から勝手にもらいました～
6	FF12003	5	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:41	2	連絡もせずに消えててごめんね。
7	FF12003	6	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:41	3	わたしのLINEはあんまり回さないでもらえると嬉しいです(*_*)
8	FF12003	7	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:42		やっぱ早いね既読
9	FF12003	8	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		うん、わかった、大丈夫だよ！
10	FF12003	9	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		ちょーど目覚まし止めた(笑)
11	FF12003	10	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:42		おはよう[絵文字]
12	FF12003	11	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		[スタンプ]<ギュー･･>
13	FF12003	12	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		おはよう
14	FF12003	13	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:42		本当ごめんね
15	FF12003	14	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:43		大丈夫だよ(笑)
16	FF12003	15	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:43		<人名>から<Rf001>と<人名>が心配してるって聞いて、申し訳なくなった
17	FF12003	16	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:44		LINEは最近始めた！

＜通番＞
同トークルームにおいて，時系列で並べた場合の本文の通し番号を示す。

＜管理 ID＞
本文の送信者の管理番号を示す。
① 提供者 A の例：Rf001　　② 参加者 B の例：Lf002

①の例はそれぞれ，左から，トークルーム右側[5]「R」の吹き出し，女性，提供者通番を示す。
②の例はそれぞれ，左から，トークルーム左側「L」の吹き出し，女性，参加者通番を示す。

＜性別[6]＞
男性には１，女性には２を付与。

[5] 図２を参照されたい。
[6] 性別，年齢，職業については，公開を希望しない場合もあり，その場合は X が付与される。

＜年齢＞
本文送受信時の年齢を示す。（年齢の情報はあるが，生年月日の情報がない場合は年度で判断）

＜職業＞
本文送受信時の職業を示す。ただし，表２に示したように筆者がコーディングした。

＜関係＞
データ提供者が判断した参加者との関係を示す。
① 提供者 C（Rf004）の例：（Lf003 の）サークルの後輩
② 参加者 D（Lf003）の例：（Rf004 の）サークルの先輩

＜親しさ＞
データ提供者が判断した参加者との親しさを示す。親しさの判定については，データ提供者の視点から，５段階評価（とても親しい～全く親しくない）で評価されたものである。尺度の捉え方については，前述の【フェイスシート】に詳しい。

＜送信日＞
本文の送信日を示す。

＜送信時間＞
本文の送信時間を示す。

＜吹き出し内行数＞
LINE 本文は，「吹き出し」画像内に提示され送られる。どの文字列，絵記号までをひとつの吹き出し内に収めるかは，送信者の任意による。
ひとつの吹き出し内に複数の改行がある場合，改行毎にセルを分け，出現順に番号を付与。改行が存在しない場合は何も入力しない。

３．２本文セル内の加工規則

＜本文セル＞
以下，付与する記号類はすべて全角とする。
① １吹き出し内，１行を１セルに入力。
② １吹き出し内に複数の改行による文字列が入力されている場合は，改行毎に１行下に入力。
③ １吹き出し内に複数の改行による文字列が入力されている場合は，「吹き出し内行数」列に改行毎に通し番号を付与。改行がない場合は何も記入しない。
④ 登場した人物や場所，施設名称については，それぞれ全角＜＞で囲い，＜人名＞，＜地名＞，＜施設名＞とする。
⑤ 元データ画像に「▷」マークが付与され，送信者の位置情報が示されている場合は，［位置情報］[7]と記入。

[7] 図２を参照されたい。

⑥ [8]日時が不明の場合は，文頭に▲を付与し　▲不明瞭　と記載。

⑦ 本文が不明瞭の場合は，文頭に▲を付与し　▲不明瞭　と記載。
一部推測できる場合は，文頭に▲を付与し，全角「」で括り本文を入力する。
例：▲「おお！！」

絵文字

① 絵文字については，半角[]で括り，[絵文字]と入力。
例：　きいてないよー  ！！　→　きいてないよー[絵文字]！！

顔文字

① 顔文字（記号の組み合わせによるもの）については，できる限りそのまま入力する。

② 再現が難しい場合は，半角[]で括り，[顔文字保留]と入力。

スタンプ

① スタンプは，半角[]で括り，[スタンプ]と入力。１スタンプ毎に１セルに入力。

② スタンプに文字列が付与されている場合は，[スタンプ]の後に文字列のみ，全角＜＞で括り文字列を入力する。　例：[スタンプ]＜さすけねー＞

③ 動くスタンプと確認できた場合は，[スタンプ]の前に，全角【】で括った【動】を入力する。記号は全角。
例：【動】[スタンプ]
例：【動】[スタンプ]＜おはようございます！＞

④ 「  」マークが付与された音の出るスタンプの場合は，[スタンプ]の前に，全角【】で括った【音】を入力する。記号は全角。どのような音が出ているかわかる場合は，[スタンプ]記号の後に＜＞で括り文字列を入力する。音声が確認できない場合は，＜＞内に＜音声不明＞と入力する。
例：【音】[スタンプ]＜さすけねー＞
例：【音】[スタンプ]＜音声不明＞

文字フォント（LINE 特有の文字スタンプ）

①  　のような，LINE 特有の絵文字フォントの場合，開始部[文字スタンプＳ]と終了部[文字スタンプＥ]で括る。ブラケットは半角，アルファベットともに全角。

② 文字スタンプが１吹き出し内の文中の一部にある場合
例：でさー，[文字スタンプＳ]ありえない[文字スタンプＥ]わけ！

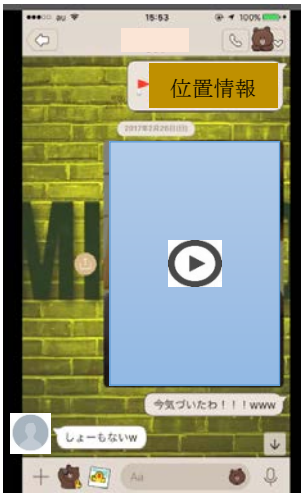
写真

写真やスクリーンショットの画像の場合，半角 [] で括り，[写真]と入力する。写真毎に１セルに入力。

8 ＜本文セル＞⑤, ⑥, 顔文字②については，画像ファイルのみ提供されたデータに付与したものである。

動画

写真に ▶ が付与されている場合は動画である。その場合半角［］で括り，［動画］と入力。動画毎に１セルに入力。



Rm025	20:26		おけー
Rm025	21:28		着いた
Lm026	21:35		どこいるの？
Lm026	21:36		☎ 不在着信
Rm025	21:50		☎ 通話時間 0:54
Rm025	21:50	1	[位置情報]
Rm025		2	
Rm025	18:26		[動画]
Rm025	18:26		今気づいたわ！！！www
Lm026	20:51		しょーもないw
Rm025	15:51		今夜地元で＜人名＞さんと飲むんだけど来る？？
Lm026	16:14		いくわ
Lm026	16:14		何時から？
Rm025	16:15		まだ決めてないー
Rm025	16:15		前から話してた、＜地名＞のホルモンの店

図３　動画，通話，不在着信，位置情報通知の処理例

通話

① 通話が行われた場合は「☎」マークを挿入する。半角の空白を挿入後，同セル内に通話時間を記入。

② 不在着信の場合は，「☎」マークを挿入し，半角の空白を挿入後，不在着信と記入。

３．３文字列以外の情報参照のためのポップアップ機能

図４　ポップアップ画像の加工例

トークルーム画像データは，図２，３，４に示すように個人情報に匿名化の処理を施し，ポップアップ画像で参照できるようにした。

検索機能を使用する際は xlsx 画面の本文列を参照し，スタンプや絵文字等，文字列以外の情報については該当部分の画像番号をクリックすることで，ポップアップ画面が別途立ち上がり参照できる（図２）。

その他画面左脇に表示される参加者のアイコン，写真や動画画面，人名等個人が特定できる情報，位置情報なども加工した。人名等は，既に管理 ID が付与されている人物の場合はその番号が付与され，それ以外の人物には＜人名＞が付与される。

４．おわりに

本稿では，2018 年 6 月末時点で収集したデータと，研究用データベース加工の際に付与した属性情報やデータにおける個人情報の加工の方針とともに，データの概要を紹介した。

本稿報告の通り，現時点で整備が行われているのは，主に 20 代の大学生を中心とした LINE コミュニティのデータであるが，現在 30 代，40 代の提供者とそのコミュニティにおける LINE のデータ収集を行っている。本発表で得た助言を元に引き続きデータ整備を行う。

謝　辞

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

文　献

岡本能里子(2016)「雑談のビジュアルコミュニケーションーLINE チャットの分析を通して」村田和代・井出里咲子『雑談の美学』pp.213-236,ひつじ書房

加藤安彦(2007)「ケータイメイルにおける顔文字と記号の出現頻度とその関係ーケータイメイルコーパスの紹介とともにー」『専修国文』81 巻，pp.1-17，専修大学日本語日本文学会

三宅和子(2018)「LINE の中の「方言」ー場と関係性を熟成する言語資源ー」小林隆編『コミュニケーションの方言学』第 14 章 pp.319-337, ひつじ書房

宮嵜由美(2004)『場面における言語行動のストラテジーの考察 ー携帯メールを中心にー』東京都立大学人文科学研究科国文学専攻修士論文

宮嵜由美(2015)「LINE を用いた依頼場面における送受信者の言語行動ー表現の担う機能と構造に着目して」西尾純二他編『言語メディアと日本語生活の研究』pp.5-20,大阪府立大学人間社会学部／大学院人間社会学研究科

宮嵜由美(2017) 「LINE を使用した依頼：2 者間・3 者間での受け手のフォローと共話性」第 21 回「ひと・ことばフォーラムーSNS の教育・研究の可能性についてー」ひと・ことば研究会

水谷信子(2001)「あいづちとポーズの心理学」『言語』第 30 巻第 7 号 pp.47-51,大修館書店

『日本語歴史コーパス(CHJ)』の教育利用の実践報告 ― 高校の古典の授業における活用例 ―

宮城 信(富山大学 人間発達科学部)*
江口 遼至(金沢高校)

Practice Report on Educational Use of " the Corpus of Historical Japanese (CHJ)" : Examples of Japanese Classical Literature Classes for High School Students

Shin Miyagi (University of Toyama)
Ryoji Eguchi (Kanazawa High School)

要旨

本稿は、日本語歴史コーパス(CHJ)を活用した学校現場での実践報告である。CHJ を学校現場で利用するためには、様々な制約がある。一方で CHJ の教育利用は始まったばかりであり、電子教科書の普及や教室でのインターネット環境の整備が進みつつある現在、CHJ は質的量的に見ても教材・資料としての価値は高く、今後様々な場面での活用が期待される言語資源である。ここでは高等学校で CHJ を活用した古典の授業の実践報告を行い、その利点と今後の課題について言及する。

１．はじめに

近年学校教育における ICT 機器を活用した探索的な授業の試みがなされている。いくつもの学校で理科や社会科等の授業実践が公開され、認知度が高まっている。ほとんどの学校では十分なインターネット環境がないが、現在段階的に整備が進み、利用できる端末の数も確保されつつある。今後、各教科の授業の中で ICT 機器が教材、または教科書として利用されるようになるのはもはや時間の問題であろう。

言うまでもなく ICT 環境の整備・導入によって授業のあり方は大きく変化していく。もっとも導入が遅れている教科の一つである国語科の授業においてもそれは同様である。翻って国語科の授業で利用できるコンテンツに着目してみると、電子教科書を除けばめぼしいものがないというのが論者の率直な感想である。その点、CHJ はすでに公開され基本無償で利用可能である点、古典作品の資料として真正性が高い資料である点、訓練次第で中高生でもどうにか検索できるインターフェイスをもつ点で、現在有望なコンテンツとして期待されるものの一つである。

一方で、いかに素晴らしい資料であっても、それを新規の教材として導入するには授業時数が不足しているというのが現場の認識であろう。また、授業内容は十分に成熟しているため他のものに簡単には置き換えにくいと思われる。特に古典の学習についてはその傾向が強い。例えば文学研究の成果である専門の辞典や本文索引等が教育の現場で活用されることは管見の限り見受けられない。このような現状に鑑みてみると、単純に利便性が高く先端的であるという理由だけで ICT 環境や機器を導入しても新しい形態の授業に移行することは難しいのではないかと思われる。現場で求められている資料は、研究者の必要としているそれとは異なることを認識し、現場に即した資料の開発とそれを活用した実践方法の指南を発信していかなければ、国語科の授業を大きく変えることは不可能である。

* miyagi@edu.u-toyama.ac.jp

２．国語科の授業における CHJ 機器の活用

CHJ を活用した教材による言語活動や検索結果に基づき作成された資料を利用して学習を行うためには、教室で利用可能な、学習者の数に応じたある程度の台数の ICT 機器があることが望ましい(標準的な学習であれば、4，5 名に 1 台程度)。本研究では CHJ 利用デバイスにタブレット端末を採用した。赤堀(2015)でも指摘されるように、「タブレット端末は、形や仕組みはパソコンに近いが、人との関わりの観点からは、パソコンよりも紙に近い。(中略)タブレット端末は、直接に指でタッチする。」(：p.13)という特性を有している。保守的な国語科の学習(特に古典の授業では、これまでの学習方法を堅持することにアイデンティティーを感じているようにさえ思う)において、例えば端末室でデータの検索を行うような ICT 機器の導入は相応の障害があるように感じている。一方、タブレット端末であれば、教科書やノートと一緒に机上に並べて置くことができる。また現状では ICT 機器は授業中のスポット的な使用に留まっており、比較的小型のタブレット端末であれば必要に応じて授業の途中で自由に出し入れができるという点も利点であると考えられる。

３．高校における CHJ を活用した古典の授業の可能性

中学校で古典作品に触れることがあっても、本格的な古典の学習が始まるのは高校 1 年生の「国語総合」の授業からである。したがって古文の資料やデータに意味を見出すことが期待されるのは、学習を進めて古典知識などをそれなりに習得した高校 2 年生以降と考えられる。ただし、学習の入門期である高校 1 年生であっても、課題を現代的な内容と関連付けて設定したり、現代人の思考に結びつけたりすることによって、学習を成立させることは可能である。よって、CHJ の新規の教育的活用法を模索する本研究では、導入のハードルを低くするために、あえて入門期の高校 1 年生を対象として、CHJ を教材に用いた古典の授業を提案・実践した。

４．アクティブ・ラーニングを意識した古典の授業

４．１　古典におけるアクティブ・ラーニング学習の課題

1 節で現在の古典の授業が硬直化していることを指摘した。しかしながら、指導要領の転換期を迎え、今後古典の授業においてもアクティブ・ラーニング学習化の要求は確実に高まってくると推察される(もちろん全ての授業をアクティブ・ラーニング学習化する必要はない。ただ様々な単元において試行しておくことは必要である)。

河添編(2018)では、「アクティブ・ラーニング型授業」と「アクティブ・ラーニング(学習：論者註)」を区別すべきと主張する。後者が目指すべき学びの形で、学習者が主語(学びの主体)となるべきものであると述べている。その実現のためには「授業中に話す、教える、評価するといった行動を学習者に引き渡す」(：p.17)ことが重要であるとする。すなわち課題を立てる、調査する、吟味するという学習活動を学習者に委ねるということである。

仮にこの指摘に従い、それらを学習者に引き渡したとして、古典の授業ではおそらく多くの学習者は一歩も前に進むことはできない。文学的作品に関する知識、古典常識、古典作品に関する直感のいずれもが不足しており、何をどうすれば良いのか、皆目見当も付かないからである。古典におけるアクティブ・ラーニング学習は思いのほか難易度が高い。

一方で、何かを「取り敢えずやってみる」ということであれば、主体的な学習も可能かもしれない。本研究で考える「取り敢えずやってみる」ことを可能とするツール(教材)が、CHJ ということになる。ただし CHJ を現場で有効に活用するためには、様々な準備と先にも述べたように実践の指南が必要であると考えられる。以下、論者らが試みた CHJ を活用した古典のアクティブ・ラーニング学習について紹介する。

４．２　気になる言葉を調べる

多くの場合、中学校段階の古典の学習では、教師が本文から指定して重要語句(内容の解釈や授業の展開に欠かせない語句)を取り上げて解説する。高校の古典の学習では、それ以外にも本文を読み込む中で気になった語について辞書等を利用して学習者が自ら調べることも多い。その際、辞書の語釈が調べたい箇所を説明するには不十分であれば、用例に頼るしかないのであるが、例解古語辞典であっても、紙幅の都合もある紙の辞書では用例はさほど多くは示されず、調べている箇所そのものが用例として上がっていることも少なくない。

これに対して、CHJ は適切に検索することができれば、いくらでも用例を手軽に採集することができる用例の宝庫である。コーパス等 ICT 活用の最大の利点は、表計算ソフトなどを利用して、多様で大量の情報の収集、整理が容易に行えることにある。第二の古語辞典と考え、従来型の古語辞典と並行利用することによって、古典作品に対する理解をより深めることができる(特にタブレット端末を利用すれば、授業中の出し入れも自由で電子辞書感覚で利用できる)。

４．３　作中の言葉から作者のものの見方や考え方を探る

高校の古典の授業では、一通り読解した後、本文中から作者のものの見方や考え方を表した箇所や語句を探しだすという学習が散見される。学習者の思考の流れを教師がコントロールし易いという点では安心感のある学習である。石塚(2008)が「作品の知識をもたせることは否定しないけれども、それに終始する古典教育であってはならない」(: p.3)と指摘するように、時には学習者の興味関心に従って、教師の掌から飛び出して作品と向き合う学習活動も必要である(4．1 節の河添 2018 の指摘も同様にとらえられる)。しかしながら、本文だけでは新たな疑問を発見するには短すぎるし、また、その本文もそれまでの学習で教師によってあらかた掘り尽くされている。本文以外にそれを求めようとして、辞書や文学全集を利用したとしても、残念ながら多くの学習者にとってはそれは相当高難易度の学習活動と言えよう。

そこでこの問題の解決のためにはコーパスを利用した学習が最適であることに気付く。コーパスで検索することによって、対象が作品全体に一気に拡大するため、教師が意図した範囲内に収まらないこともある。また CHJ を利用した調べ学習では、検索が容易であるため、以下の本稿での実践のように特定の答えを想定せずに様々な語句を手当たり次第検索(「取り敢えずやってみる」である)した結果から、作者がどのような人物であったのかを想像するといった探索的学習も可能である。その結果、予想外の新たな状況が浮かび上がることもある(教育活動であるので、ここでは敢えてその真偽は問わないことにする)。

このように、グループ学習にこだわらず、また既習の学習方法にとらわれることなく、得られた結果を批判的に解釈してどのように消化していくかが目指すべき高校生段階での古典のアクティブ・ラーニング学習と言えよう。

５．授業実践の概要

５．１　授業計画

本稿で報告する授業実践は以下のように実施された。

５．１．１　実践協力校

本実授業践授は、平成 30 年 3 月 26 日に私立金沢高等学校(石川県金沢市)で実施した。授業参加者は高校一年生の特進クラスの生徒で、授業者は同校の江口遼至教諭、論者(宮城)は CHJ の利用補助などを行う授業補助者として実践授業に参加した。また授業は公開授業の形式で実施され、後に述べるように多くの参観者があった。

５．１．２　授業準備

・CHJ 利用アカウントの用意

CHJ を利用するためには、利用者が個人で利用アカウント申請をする必要がある。今回は投げ込み教材的な実践であるので、実践協力校の国語科教員にも依頼して、複数のアカウントを確保した。学習の形態にもよるが、一クラス 35 人前後として、4,5 人グループ毎に 1 台程度の端末が用意できれば ICT の利用環境としては十分である。

・ICT 環境と利用端末の用意

今回の実践協力校である金沢高等学校は、すでに教室にインターネット環境が整備され、学習者が個人のタブレット端末を所有しており、基本的な操作は習得済みであったため CHJ 導入時の問題は少なかった(なお、先に述べたように教室で CHJ を利用するための端末としてはタブレット型が最適である)。

・資料の用意

教科書(『新探求国語総合(古典編)』(桐原書店)の他に、本授業で使用するワークシートと検索する語を見つけるための『読んで見て覚える重要古文単語 315』(桐原書店)を準備させた。

５．２　実践報告

本稿の授業実践は、「国語総合」の古典の授業として、3 時間で実施された。教材は「徒然草」である。「徒然草」は比較的多くの章段を学習済みで、筆者の考え方を自分の考えと比較することが容易であると考えたからである。なお、文中の( )内は前掲の学習活動のおよその時間を示している。

［第 1 時］

「徒然草」の復習と CHJ を利用するための準備にあてた。

［第 2 時］

本実践の主要部は本時である。簡単に前時での課題「作者のものの見方や考え方について」の確認と検索方法の復習をした(5 分)。

本時の目標は、CHJ を利用して、『徒然草』の使用語彙から、「作者のものの見方や考え方を探る」という課題である。このクラスはこれまで「徒然草」の八九段(奥山に、猫又といふものありて)や、第二三六段(丹波に出雲といふ所あり)などの章段を学習してきている。それも踏まえて本時では「徒然草」全体を対象として形容詞や形容動詞の使用状況(使用頻度の偏り)に着目して、課題に取り組むことにした。学習者が主体的に検索する語を決めて、「取り敢えずやってみる」という調べ学習である。調査協力校では学習者一人一人がタブレット端末を所有しているが、基本的な作業は 4 人グループに分かれて行い教科書や単語帳を利用して調べ学習を行った(20 分：次頁 写真 1)。

CHJ では、検索結果は、現代語訳のない本文のみが表示される。検索結果の誤読を防ぐために、必要に応じて web で現代語訳の紹介サイト(誤訳の少ない確認に適切なサイトを事前に教師が紹介)を参照することも推奨した。検索結果を語別に使用頻度と代表的な用例をワークシートに記入して整理させた。

写真１　タブレット端末を利用した調べ学習

教科書や単語帳で当たりを付け、検索にはタブレット端末を用い、紙のワークシートに記録するという、新旧の学習材を並行利用した学習活動で、ICT 機器の導入において現在もっとも抵抗が少ない授業形態であると言える。また、教育効果も期待でき、赤堀(2015)では複数の異なるメディアを並行利用することによって学習効果も高まると指摘している。

写真２　タブレット端末とワークシートの並行利用

クラス全員のワークシートの内容を集計すると、検索された形容詞・形容動詞は全部で 54 語あり、検索した人数が多い語は以下の通りであった(5 分)。

(12 名) : くちおし、(11 名) : こころにくし、(10 名) : いみじ、めでたし、(8 名) : をかし、(7 名) : あさまし、(6 名) : わろし、(5 名)なまめかし、むつかし、(4 名) : ゆゆし、(3 名) : あはれ、あやし、うるはし、けし、さかし、まさなし、むげなり(以上、上位 17 語)

本時の目標である「作者のものの見方や考え方を探る」という課題解決のために、多くの学習者が検索した上位 17 語のうち、ここでは下線の 8 語がネガティブな意味の語と解釈できる。もちろんこれらの語がすべて作者である兼好法師のもののとらえ方を反映しているわけではないが、これまでの学習で学習者らがとらえた作者像の一端を表しているように思う。その後、ワークシートを基に、グループ内で整理した語から、作者のどのようなものの見方や考え方、人物像が見えてくるかを話合い交流した(10 分)。

交流時にワークシートの記録を参照したり、タブレット端末上での検索結果を示したり、学習者がもっとも有効であると考える方法で意見の交流を行っていた。他教科の授業でタ

ブレット端末を利用した経験を持つ学習者であったので、比較的容易にメディアの使い分けを行えたのではないかと推察する。

写真３　タブレット端末での交流

この段階でも、学習者から「ものごとを二極化してとらえる傾向がある」「理想の人物像がとても高いためほぼ全員に対して物足りなさを感じている」といった、前時までの授業では指摘のなかった、新たな作者像が提出されており、CHJ の活用によって学習者らが新たな視点を提供できる可能性が示唆された。

授業のまとめとしてタブレット端末の google フォームを利用して振り返りを入力させ、即座に集計してスクリーンに写し出し、「User Local テキストマイニングツール」(https://textmining.userlocal.jp/)を利用して、提出された意見をビジュアル的に関連付けてクラス内で共有した(5 分)。

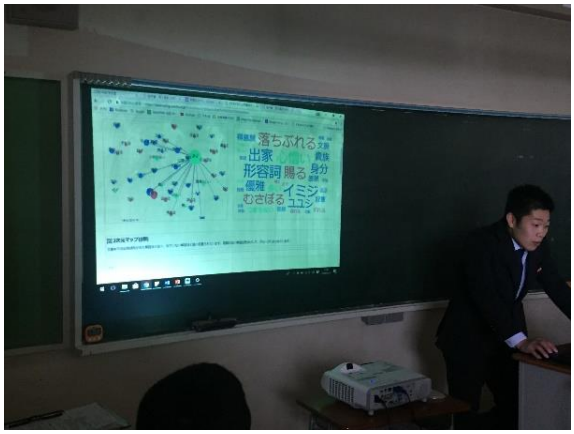


写真４　テキストマイニングツールでの共有

時間が許せば、プロジェクタで投影された内容を基に、グループ内で話し合う交流の機会を設け、他人の様々な意見に触れることで、作品に対する理解を一層深められると考えられる。本稿の実践授業では時間の制約から交流の場を割愛せざるを得なかった。

［第 3 時］

図書館に移動して実施した。まず、前時の復習として、第 2 時に google フォームで共有した振り返りをプリントにまとめ配布した。それを読んで、自分と異なる意見やよく分からない意見を見つけ、なぜそうなるのかを書いた本人に確認するという交流を行った。

必要があれば CHJ を利用して追加で調査した。調べたい内容によっては、CHJ より現代語訳や頭註がついた『新編古典文学全集』(小学館)で当該の章段に当たる方が良い場合もあるので、図書館の資料にも当たるよう促した。その際、図書館司書に調べ学習の補助を依頼した。本実践のような調べ学習では、CHJ だけで完結することはなく、文学全集等も含めた他の媒体と相互に活用することで理解を深めていくべきである。

５．３　教室の状況

本研究のような外部の研究者が協力して実践された ICT を活用する国語科の授業は、ICT を活用した授業を多数実施している金沢高校(現在文科省から研究指定校の認定を受けている)においても新しい試みであった。そのため多くの参観者を迎えることができた。日頃から各教科で ICT を活用した授業に取り組んでいる教師らにとっても ICT を活用した国語科の授業は関心の高いものであることが伺える。

写真５　参観者で賑わう授業(第２時)

６．今後の課題

CHJ は、当面のところ用例検索用の第二の古語辞典として、または、調べ学習用の支援ツールとして、スポット的な発展的な古典の学習の中に組み込んでいければ十分である。いずれ改善されるであろうが、多くの教室では未だインターネット環境が十分ではないし、学習用コンテンツとしてみても現行の CHJ ではまだ十分には現場の要求を満たしていないからである。一方で、全てではないにしてもアクティブ・ラーニング学習への転換が求められている現状に鑑みると、論者はいずれ間違いなく CHJ が古典の学習において不可欠な学習材(または支援ツール)になると確信している。それに向けて、今後継続的に多くの実践者が継続的に CHJ を活用した実践を積み重ねていくことが肝要である。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「古文教育に資する、コーパスを用いた教材の開発と学習指導法の研究」による成果の一部である。

文　献

赤堀侃司(2015)『タブレット教材の作り方とクラス内反転授業』、Jam House
石塚修(2008)「伝統的な言語文化の尊重に向けて」『月刊国語教育研究』440、日本国語教育学会
河添房江(編) (2018)『アクティブ・ラーニング時代の古典教育　小・中・高・大の授業づくり』、東京学芸大学出版会
明治書院(編) (2016)『高等学校国語科 授業実践報告集 アクティブ・ラーニング編』、明治書院

双方向 LSTM による分類語彙表番号を語義とした all-words WSD

新納浩幸（茨城大学工学部情報工学科）*
鈴木類（茨城大学大学院理工学研究科情報工学専攻）†
古宮嘉那子（茨城大学工学部情報工学科）‡

All-words WSD with WLSP number as a Sense Label Using a Bidirectional LSTM

Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)
Rui Suzuki (Ibaraki University, Major in Computer and Information Sciences)
Kanako Komiya (Ibaraki University, Department of Computer and Information Sciences)

要旨

語義曖昧性解消は意味解析の重要な要素技術であるが，実際のシステムに利用されることは少ない．これは現状の語義曖昧性解消が主として教師あり学習のアプローチをとっているため，対象単語が限定されてしまうからである．我々は対象単語を限定しない all-words WSD システムを，大規模な語義タグ付きコーパスと双方向 LSTM を用いて構築した．構築できたシステムは，入力されたテキスト内の全ての単語にその語義を高精度に付与できる．本論文では構築したシステムを紹介し，その有用性を文書分類タスクと語義のクラスタリングから示す．

1. はじめに

語義曖昧性解消は意味解析の根幹の処理でありながら，そのシステムが現実のアプリケーションで広く利用されているとは言いがたい．これは現状の語義曖昧性解消が，主として，教師あり学習のアプローチをとっているため，対象単語が限定されてしまうことが大きな原因である．そのため対象単語を限定せず，入力文内の単語全てにその語義を付与する all-words WSD (Word Sense Disambiguation) の重要性が古くから指摘されてきた (Navigli (2009))．しかし全ての多義語に対して語義タグ付きの用例データベースを構築するコストは膨大であるため，単純に通常の WSD の手法を拡張するだけでは実現できない．辞書の用例文などを用いた知識ベースの手法や教師なし学習の手法を用いる試みもあるが，多くの場合，MFS (Most Frequent Sense) と同等以下の精度しか得られず，実用的な精度とは言えない (Kulkarni et al. (2010))．

一方，語義タグ付きコーパスの開発は並行して行われており，昨年，国立国語研究所から BCCWJ に対する分類語彙表番号アノテーションデータ (Kato et al. (2017)) が公開された．

* hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
† 17nm709g@vc.ibaraki.ac.jp
‡ kanako.komiya.nlp@vc.ibaraki.ac.jp

これを語義タグ付きの用例データベースと見なせば，all-words WSD は実現できるが，全ての多義語に対する用例データベースとしては小規模であり，個々の多義語の WSD を解くという直接的な手法は利用できない．ここでは all-words WSD を各単語に語義（ラベル）を与える系列ラベリング問題としてみなし，双方向 LSTM (Long-ShortTerm Memory) を用いて all-words WSD システムを構築した．

本論文では構築したシステムの詳細を述べる．また，その有用性を示すために 2 つの実験を行う．1 つは文書分類のタスクである．通常，文書分類は文書を bag of words のモデルを使って素性ベクトルに変換するが，本システムを利用して素性ベクトルに語義の素性を加えることができる．語義の素性を追加することで識別精度が向上する．

もう 1 つは語義のクラスタリングである．本システムを利用してコーパスを語義列のコーパスに変換できる．さらに得られた語義列のコーパスを word2vec (Mikolov et al. (2013) Mikolov et al. (2013)) にかけることにより，語義の分散表現が得られる．この語義の分散表現を利用して，語義のクラスタリングを行う．本システムの語義は分類語彙表番号であり，分類語彙表が概念辞書であるため，得られたクラスタリング結果と分類語彙表を比較でき，主観的に構築されている分類語彙表の考察が行える．

2. 関連研究

all-words WSD はドメインを限定すれば通常の教師あり学習も可能である．実際に SemEval で行われた all-words WSD のタスクでも，いくつかの教師あり学習によるシステムが参加している．ただし教師あり学習はその拡張性に問題がある．

教師あり学習手法を用いない場合，all-words WSD の手法は知識ベースの手法か教師なし学習手法に分類できる (Kulkarni et al. (2010))．知識ベースの手法として，古典的には Lesk の手法 (Lesk (1986)) がある．これは対象単語の周辺の単語集合と，対象単語の各語義の定義文中に現れる単語集合との重なり度合いを調べ，その度合いの大きい語義を選択するというものである．ただし一般に知識ベースの手法は語義の頻度の情報を利用していないために，精度が低いという問題がある．

教師なし学習手法には様々なタイプのものが存在するが (Yarowsky (1995), Izquierdo-Beviá et al. (2006), Zhong and Ng (2009))，近年は，語義列の生成モデルを定義し，ある種のヒューリスティックを導入することでプレーンなコーパスから生成モデルのパラメータを推定する手法が採られている (Boyd-Graber et al. (2007), Tanigaki et al. (2013, 2015), Komiya et al. (2015))．教師なし学習手法は知識に基づく手法よりも精度は高いが，実用的な精度とは言えない．さらに教師なし学習手法による all-words WSD システムは，通常の WSD システムとは異なる入出力となっていることにも注意すべきである．通常の WSD システムの入力は WSD の対象単語を含む文であり，出力はその対象単語の語義である．一方，教師なし学習手法による all-words WSD システムの入力はコーパスである．入力コーパス中のすべての単語に語義を付与する．しかし新たに対象単語を含む文を単独で入力しても，その対象単語に語義を付与することはできない．

また近年は語義の分散表現を求めることで all-words WSD を実現することも試みられてい

る (Neelakantan et al. (2014), Chen et al. (2014))．単語 w に対する i 番目の語義の分散表現 s_i とする．w の周辺文脈を分散表現 v で表現し，s_i と v の類似度を測り，最も類似度の高い i に対する語義を識別結果とすることで，WSD が行える．この手法は知識ベースの手法の一種であり，この手法も精度が低いという問題がある．一般に MFS よりも高い精度は得られない．そのため逆に分散表現を用いて，MFS を推定する方法も研究されている (Bhingardive et al. (2015))．

一方，Hatori は all-words WSD を通常の WSD の対象単語を拡張する形ではなく，all-words WSD を系列ラベリング問題として見なして解いている (Hatori et al. (2008))．ただし系列ラベリング問題を解く手法として CRF を用いているために，対象単語に対する語義付き用例が少ない場合に，統一した枠組みでの処理ができない．このような問題に対処するために，Shinnou は CRF ではなく点推定を用いた日本語 all-words WSD システム KyWSD (Shinnou et al. (2017)) を開発したが，単語切り自体もシステム内の学習で行っており，一般の単語切りとは異なる単位を用いているので，その点で実用的ではない(1)．

また本システムと同じく分類語彙表番号を語義とした日本語 all-words WSD の研究として (Suzuki et al. (2018)) がある．そこでは対象単語とその類義語の周辺単語の分散表現の距離を計算する教師なし学習のアプローチで語義曖昧性解消を行っている．この研究も教師なしであるため新たな入力文に対して語義を付与することができず，実用的ではない．

以上の観点から本システムの特徴を述べると，本システムは all-words WSD を系列ラベリング問題として見なして，双方向 LSTM を用いて解いている．そのためその精度は高く，語義のラベル数が多い場合や，ある単語について用例が少ない場合でも，同一の枠組みで処理が行える．また教師なし学習ではないために，新たな入力文に対しても語義を付与することができる．

3. 双方向 LSTM による系列ラベリング

all-words WSD は系列ラベリング問題とみなすことができる．系列ラベリング問題をニューラルネットワークで扱う場合，リカレントニューラルネットワーク（以下，RNN）を使用する．RNN では時刻 t の中間層の内容を時刻 $t-1$ の入力に使い状態を保持しながら学習することで時系列に対応しているが，長い系列データを学習する際に勾配の消失が起こる場合がある．LSTM は RNN の一種種であり，入力，出力，忘却の 3 つのゲート処理を取り入れることで勾配の消失を防ぎ長い系列データにも対応することができる．また RNN は時系列データを扱うものであり，自然言語処理では文や文書の単語列を時系列データとみなして使っている．そのため通常，注目している時刻 t 以降の単語も利用できるため，データを逆方向からも解析できる．順方向の LSTM と逆方向の LSTM を同時に用いて，時刻 t での出力を求めるのが双方向 LSTM である (図 1 参照)．

モデルで利用した訓練データは BCCWJ に対する分類語彙表番号アノテーションデータである．このデータは 340,879 トークン, 19,432 タイプ（記号空白を含む），918 種類の語義で構

(1) 用言の語幹と語尾を分離する形で単語切りが行われている．

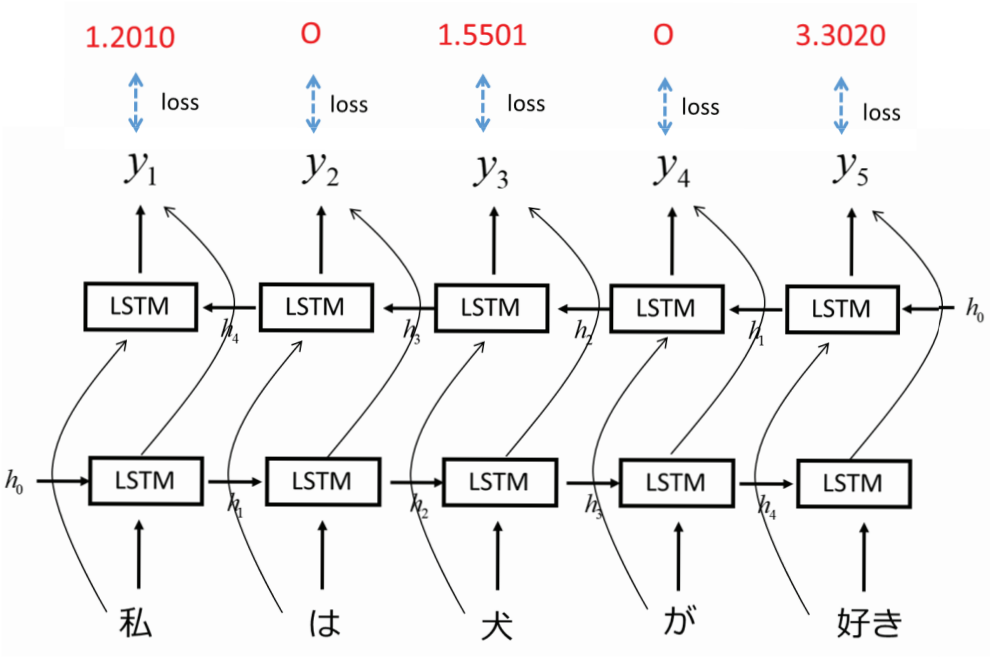


図 1　双方向 LSTM による語義の付与

成されている．このデータのうち 1 割をテストデータのために取り除いておき，残りの 9 割で双方向 LSTM の学習を行った．LSTM は 2 階層を用い，単語から分散表現に変換する部分は学習を行わずに既存の日本語分散表現データである nwjc2vec (Shinnou et al. (2017)) を用いた．epoch は 20 回で終了させた．

4.　システム概要

本研究で作成したシステムは先に構築した双方向 LSTM のモデルを核とするが，形態素解析や語義の候補の算出部分は別の資源を利用している．これらを組み合わせて本システムの all-words WSD が実現できている．システムへの入力はプレーンなテキストであり，出力はそのテキストが形態素解析され自立語に語義（分類語彙表番号）が付与されたものである．分類語彙表番号を持たない助詞などの単語には語義として ‘O’ というラベルを付与している．本システムの実行例を以下に示す．

システムの入力から出力までのステップは以下のとおりである．

step.1 単語切りと品詞付け

単語切りには，辞書として unidic-cwj-2.3.0 を用いた形態素解析器 MeCab を使用した．

step.2 語義候補の取り出し

各単語の語義候補の取り出しには国語研より公開されている分類語彙表番号-Unidic 語彙素番号対応表 wlsp2unidic を用いた．wlsp2unidic は分類語彙表番号から Unidic 語彙素番号を求めることも，Unidic 語彙素番号から分類語彙表番号を求めることもできる．本研究では分類語彙表番号を語義としているため，MeCab で出力される各単語の語彙素番号をもとに

```
> cat sample.txt
学問に基づいた本物の技術を創出できるのはやはり大学だと実感し始めていたからである.

> ./allwordswsd sample.txt
学問-名詞/1.3074 に-助詞/0 基づい-動詞/2.1110 た-助動詞/0 本物-名詞/1.1040
の-助詞/0 技術-名詞/1.3421 を-助詞/0 創出-名詞/1.3200 できる-動詞/2.1220
の-助詞/1.1000 は-助詞/0 やはり-副詞/4.3120 大学-名詞/1.2630 だ-助動詞/0
と-助詞/0 実感-名詞/1.3001 し-動詞/2.3430 始め-動詞/2.1502 て-助詞/0
い-動詞/2.1200 た-助動詞/0 から-助詞/0 で-助動詞/0 ある-動詞/2.1200
. -補助記号/0
```

図 2　all-word WSD の実行例

wlsp2unidic を用いて語義候補となる分類番号を取り出した.

step.3 分散表現への変換

各単語は分散表現に変換されて LSTM へ入力される．分散表現は LSTM の学習時に同時学習することもできるが，ここでは既存の分散表現 nwjc2vec (Shinnou et al. (2017)) を利用した.

step.4 LSTM による語義の確率算出

システムを構築する際の訓練データは 918 種類の語義で構成されている．したがって LSTM によって出力されるのは入力したテキストの各単語に対する 919 個の語義（語義なしを含む）の確率となる.

step.5 語義候補との調整と出力

システムが出力するのは，入力テキストの各単語に語義を付与したものである．**step.2** で語義候補が 1 つだった場合はその語義を付与し，語義候補が複数あった場合はその中で最も確率が高いものを **step.4** から求め付与する.

5. 精度測定の実験

システムの精度測定の実験について述べる．本実験では，分類語彙表番号アノテーションデータの 9 割を訓練データとして双方向 LSTM で学習し，1 割をテストデータとした．テストデータには 1,736 種類の多義語が含まれており，多義語の総数は 7,137 である．また，多義語の総語義数は 19,320 で平均語義数は 2.71 である．ただし，ここでは Section 4 の **step.5** の出力の調整は行わず，LSTM の出力の中で最も確率の高い語義を付与している．実験の結果，すべての単語の正解率は 0.870 (0.768)，多義語の正解率は 0.737 となった．多義語の平

均語義数が 2.71 なので，語義をランダムで選択した場合の正解率は 0.369 である．MFS によるすべての単語の正解率は 0.827 であった．このため理論的に本システムの多義語の正解率は MFS 以上の値であることから，本システムの精度は高いと考えられる．

また，ここで求めた全単語の正解率の内訳は，単義語の正解率 0.786，多義語の正解率 0.737 であるが，これは LSTM の出力である．つまり全語義の中から最も確率の高いものを出力していることに注意したい．実際のシステムが語義を出力する際は，対象単語の候補の語義のうち最も確率の高いものを出力すればよい．この点を考慮に入れた場合には正解率は更に高くなると予想できる．

6. 文書分類への応用

本章では本論文で構築した all-words WSD システムの応用として文書分類を行う．

文書分類は文書を bag of words のモデルを用いて素性ベクトルに変換し，その上で SVM などの機械学習手法を利用して解決する．bag of words のモデルでは，通常，素性として自立語を用いるが，all-words WSD システムを利用することで，その自立語の語義も素性として利用できる．

実験データは以下で公開されている Amazon Dataset の日本語文書を利用した．

`https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus-webis-cls-10/#webis-download`

上記データは感情分析データであり，3 つの領域 (‘books (B)’, ‘DVD (D)’ 及び ‘music (M)’) からなっている．また各領域に訓練データとテストデータがそれぞれ 2,000 文書存在する．つまり計 12,000 の 文書が存在する．この 12,000 文書に対して形態素解析を行い，自立語を素性として，各文書を素性ベクトルで表し，Naive Bayes により識別した (表 1 の BOW)．また上記 12,000 文書に対して all-words WSD を行い，自立語の他に自立語の語義も素性として加えて，各文書を素性ベクトルで表し，Naive Bayes により識別した (表 1 の BOW+sense)．実験の結果を表 1 に示す．

Domain	BOW	BOW+sense
(B) books	0.6820	0.6765
(D) DVD	0.7265	0.7420
(M) music	0.7645	0.7725
Average	0.7243	0.7303

表 1　文書分類の実験結果

更に領域適応の実験も行った．先でのデータでは 3 つの領域（B,D,M）があるので，領域シフトのその組み合わせとして 6 通りがある．それぞれの領域シフトにおいてソース領域の訓練データにより Naive Bayes の分類器を作成し，それをターゲット領域のテストデータの正解率

で評価した．ソース領域の訓練データでは素性として自立語のみのもの（BOW) と自立語の他に自立語の語義を加えてたもの（BOW+sense) を各々試し，比較する．結果を表 2 に示す．

DA	BOW	BOW+sense
B → D	0.6775	0.6825
B → M	0.6500	0.6610
D → B	0.6620	0.6715
D → M	0.6975	0.6945
M → B	0.6460	0.6455
M → D	0.6910	0.6925
Average	0.6707	0.6746

表 2　文書分類の領域適応

以上の実験から文書分類では自立語だけではなくその語義も素性として加える効果が確認できた．

7.　語義のクラスタリングへの応用

本章では本論文で構築した all-words WSD システムの応用として，分類語彙表番号の分散表現を構築し，語義のクラスタリングを行う．

7.1　語義の分散表現の構築

語義のクラスタリングを行うために，まず語義の分散表現を構築した．手順は以下のとおりである．

最初に毎日新聞 ’93 年から ’99 年の記事からランダムに 30 万文を取り出し，これを分散表現構築のためのコーパスとした．次に本システムを用いてこのコーパス中の全単語に分類語彙表番号を付与した．次にシステムの出力から分類語彙表番号の分かち書きを作成した．このとき語義が O （つまり自立語ではない）の場合，つまりその単語が自立語ではない場合は，語義ではなく単語表記とした．

分類語彙表番号の分かち書きは 7,937,723 トークン，12,698 タイプから構成されている．このように分かち書きされたテキストに対して word2vec (Mikolov et al. (2013) Mikolov et al. (2013)) を用いて分類語彙表番号の分散表現を構築した．次元数は 100 とした．用いた word2vec の学習のパラメータを表 3 に示す．

7.2　語義のクラスタリング

前章で構築した語義の分散表現のうち，名詞に対応する語義だけを対象にしてクラスタリングを行った．対象とした語義は 509 種類である．利用したクラスタリング手法は Word 法である．得られたデンドログラムを図 4 に示す．

次にこのクラスタリング結果を評価する．分類語彙表は概念辞書であるために，語義が階層構造（木構造）の中で定義されている．そのため同一親ノードまでの木の深さにより，擬似的

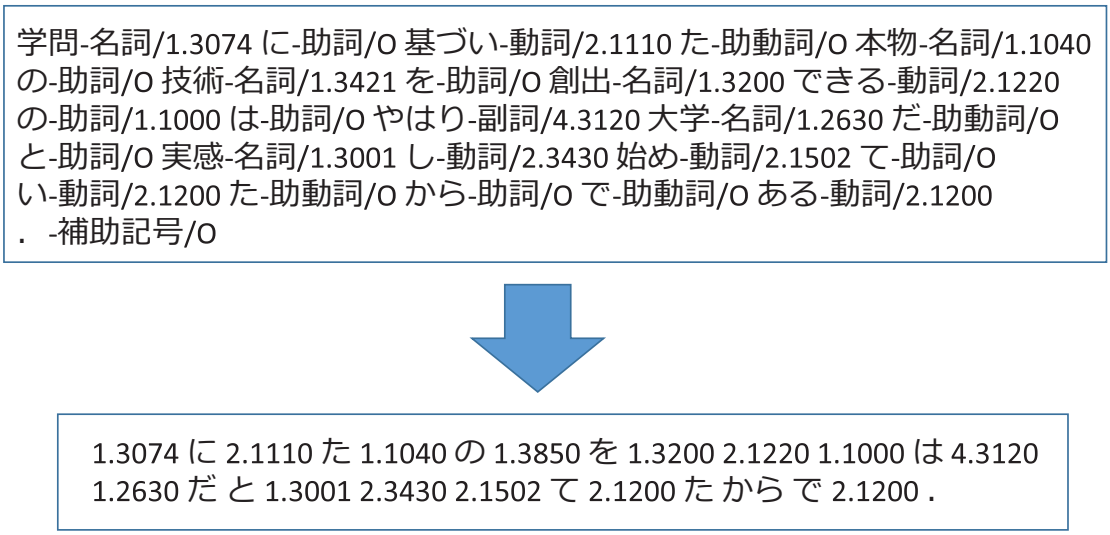


図 3　語義の分かち書き

CBOW or skip-gram	-cbow	1
Dimensionality	-size	100
# of surrounding words	-window	5
# of negative samples	-negative	5
Hierarchical softmax	-hs	0
Mini. sample threshold	-sample	1e-3
# of iterations	-iter	5

表 3　word2vec のパラメータ値

に語義間の距離を測ることができる．具体的には語義番号の桁数の位置が木の階層を表し，その桁数の数値がその階層でのクラスを表している．例えば語義番号 ‘1.1344’ と ‘1.1310’ では ‘1.13**’ の部分が共通の親ノードに対応するため，距離は 3 となる．‘1.1611’ と ‘1.1612’ では ‘1.161*’ の部分が共通の親ノードとに対応するため，距離は 2 となる．このように語義間に距離を定義すると，クラスタ内の距離が定義できる．クラスタ内の距離とはクラスタ内の各要素の間の距離の平均である．このクラスタ内の距離の平均をクラスタリング結果の評価値と見なせる．

先にクラスタリングして得られたデンドログラムを ‘cut’ することで，50 個のクラスタに分割した．各クラスタについてクラスタ内の距離を調べ，その平均を取ると 3.054 となった．クラスタリング対象となった 509 個の語義の集合からランダムに 10 個選びそれを 1 つのクラスタと考えて，クラスタ内の距離を調べる実験を 20 回行い，その平均を取ると 3.716 であった．このことから本論文で得た語義のクラスタリング結果は妥当性があると考えられる．

また参考としてクラスタ内の距離が最も小さかったクラスタは以下である．その距離は

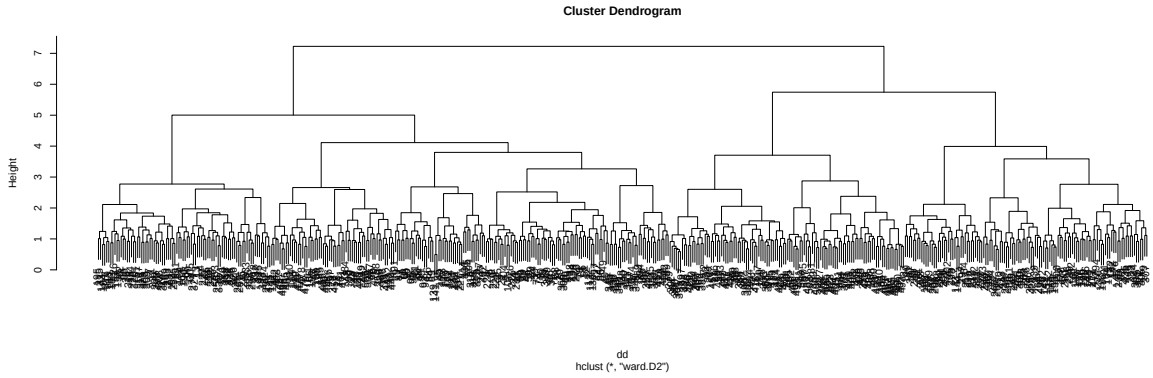


図 4　語義のクラスタリング結果

1.733 であった.

```
1.3101, 1.3105, 1.3110, 1.3111, 1.3112, 1.3150
```

またクラスタ内の距離が最も大きかったクラスタは以下である．その距離は 3.821 であった.

```
1.1100, 1.1623, 1.2030, 1.3047, 1.3300, 1.3360, 1.4580, 1.5000
```

これらの点から分類語彙表は概念の階層構造を考察できる.

8.　おわりに

本論文では分類語彙表番号を語義とした all-words WSD システムを作成した．訓練データには BCCWJ に対する分類語彙表番号アノテーションデータを使用し，学習には双方向 LSTM を用いた．構築したシステムは高い識別精度を示した．また構築したシステムの応用として文書分類と語義のクラスタリングを行うことで，本システムの有用性を示した.

謝　辞

本研究（の一部）は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。

文　献

Roberto Navigli (2009). “Word sense disambiguation: A survey.” *ACM Computing Surveys (CSUR)*, 41:2, p. 10.

Anup Kulkarni, Mitesh M Khapra, Saurabh Sohoney, and Pushpak Bhattacharyya (2010).

"CFILT: Resource conscious approaches for all-words domain specific WSD." *SemEval-2010*, pp. 421–426.

Sachi Kato, Masayuki Asahara, and Makoto Yamazaki (2017). "Annotation of 'Word List by Semantic Principles' information on 'Balanced Corpus of Contemporary Written Japanese'." *Processing of NLP 2017*, pp. 306–309 (In Japanese).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space." *ICLR Workshop paper*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*, pp. 3111–3119.

Michael Lesk (1986). "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." *the 5th annual international conference on Systems documentation*, pp. 24–26.

David Yarowsky (1995). "Unsupervised word sense disambiguation rivaling supervised methods." *ACL-95*, pp. 189–196.

Rubén Izquierdo-Beviá, Lorenza Moreno-Monteagudo, Borja Navarro, and Armando Suárez (2006). "Spanish all-words semantic class disambiguation using Cast3LB corpus." *MICAI 2006: Advances in Artificial Intelligence*, pp. 879–888.

Zhi Zhong, and Hwee Tou Ng (2009). "Word Sense Disambiguation for All Words without Hard Labor." *IJCAI-2009*, pp. 1616–1622.

Jordan L Boyd-Graber, David M Blei, and Xiaojin Zhu (2007). "A Topic Model for Word Sense Disambiguation." *EMNLP-CoNLL-2007*, pp. 1024–1033.

Koichi Tanigaki, Mitsuteru Shiba, Tatsuji Munaka, and Yoshinori Sagisaka (2013). "Density Maximization in Context-Sense Metric Space for All-words WSD." *ACL-2013*, pp. 884–893.

Koichi Tanigaki, Shuichi Tokumoto, Tatsuji Munaka, and Yoshinori Sagisaka (2015). "Hierarchical Bayesian word sense disambiguation for mapping context space to sense space (in Japanese)." *IPSJ SIG on NLP*, pp. NL–220–5.

Kanako Komiya, Yuto Sasaki, Hajime Morita, Hiroyuki Shinnou, Minoru Sasaki, and Yoshiyuki Kotani (2015). "Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation." *PACLIC-29*, pp. 35–43.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum (2014). "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space." *EMNLP-2014*, pp. 1059–1069.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun (2014). "A Unified Model for Word Sense Representation and Disambiguation." *EMNLP-2014*, pp. 1025–1035.

Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Harichandra Redkar, and Pushpak Bhattacharyya (2015). "Unsupervised Most Frequent Sense Detection us-

ing Word Embeddings." *HLT-NAACL 2015*, pp. 1238–1243.

Jun Hatori, Yusuke Miyao, and Jun'ichi Tsujii (2008). "Word Sense Disambiguation for All Words using Tree-Structured Conditional Random Fields." *COLING-2008*, pp. 43–46.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori (2017). "Japanese all-words WSD system using the Kyoto Text Analysis ToolKit." *PACLIC-31*, No.11.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou (2018). "All-words Word Sense Disambiguation Using Concept Embeddings." *LREC-2018*.

Hiroyuki Shinnou, Masayuki Asahara, and Minoru Sasaki Kanako Komiya (2017). "nwjc2vec: Word Embedding Data Constructed from NINJAL Web Japanese Corpus (In Japanese)." *Natural Language Processing*, 24:5, pp. 705–720.

『キングコーパス』の構築と活用

髙橋　雄太（明治大学大学院・日本学術振興会）†

Construction and Practical Use of “King Corpus”

Yuta Takahashi (Meiji University / Japan Society for the Promotion of Science)

要旨

本発表は，昭和期の雑誌『キング』を資料として構築した『キングコーパス』の設計と活用についてである。国立国語研究所の明治・大正期の『太陽』のコーパスに続く資料として，大衆雑誌『キング』を選定し，1933 年と 1941 年でコーパスを構築した。『キングコーパス』の設計については，資料の選定方法や，字数，語数，記事数，著者数といった基礎的な規模を示し，また文体・記事ジャンル・品詞分布の観点から『太陽』との連続性について検討した。『キングコーパス』の活用については，和語の表記を例に，「当用漢字音訓表」との比較を行った。「当用漢字音訓表」採用の和語の動詞を対象に，各語の使用表記の一致度を計測した。その結果，経年変化で一致度が高くなる傾向があることがわかり，「当用漢字音訓表」の設計の背景に，近代における用字法の変化があることが明らかになった。

１．はじめに

近年，近代語の雑誌コーパスが構築されたことにより，量的・通時的に言語を分析することが可能になり，明治期から大正期にかけての言語の変化の様相が，詳らかに解明されるようになった。稿者はこれまで，特に『太陽』のコーパスを用いて和語の表記の変遷を研究し，明治期から大正期にかけて，「1 語複数表記」から「1 語 1 表記」に向かうこと(髙橋 2016)や，複数表記語においては「1 義複数表記」から「1 義 1 表記」に向かう(髙橋 2015)など，表記が「揺れ」から「安定」に向かうことを明らかにしてきた。しかしながら，研究を進める上では，『太陽』の最終年である 1925(大正 14)年では，言語の変化が完了しない事例が少なからず確認された。

図 1 には，「タノム」という語の各表記の使用記事率の推移を示した。「タノム」には近代では「頼」と「恃」の表記が用いられ，経年変化によって「恃」の使用が減少していく過程が観察できる。しかしながら，1925 年の段階では，第 2 表記である「恃」も 8.1%使用されており，1925 年以降の調査が望まれた。

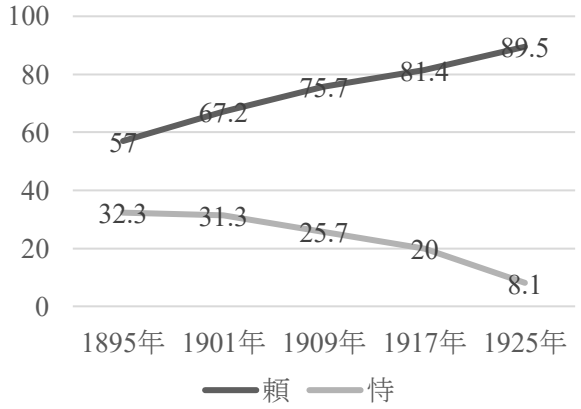


図1　タノムの各表記の使用記事率

† yutati.h@gmail.com

また, 表記の歴史における重大な転換点の一つと考えられる 1946 年の「当用漢字表」に至るまでの変遷や, 歴史区分としての狭義の「近代」の通時的な分析など, 近代語の表記の研究においては, 昭和前期の調査の必要性が非常に高いと考えられる。

『太陽』をはじめとした近代語のコーパスとの通時的分析を, 計量的に, また網羅的に行う上では, 昭和前期のコーパスを構築することが望ましいと考える。そこで, 本研究では, 『太陽』に続く資料として, 昭和前期の資料を選定し, 独自にコーパスを構築した。以下 2 節では, 資料の選定方法及びその設計について記し,3 節では構築したコーパスと『太陽』を合わせ用いて, 和語の表記の調査を行う。

２.『キングコーパス』の構築

２.１　資料の選定

近代語のコーパスの構築対象として『太陽』が選定された理由としては, 『太陽』が「分量の多さ」, 「ジャンルの広さ」, 「執筆陣の多彩さ」, 「読者層の厚さ」を持ち, 近代語の資料として代表性を持つことが指摘されている(田中 2012)。そのことを受けて, 『太陽』に続く資料として, 多様な記事ジャンルを擁する**①総合雑誌**であることを条件とした。次に, 昭和前期(1945 年まで)を通じて観察することの出来る**②長期刊行**の雑誌であり, かつ月刊誌など継続して刊行されて**③分量が多い**資料に絞った。また, この時代の雑誌のほとんどにあてはまることだが, 『太陽』が大正末期には口語記事が 95%以上を占めていたことを受けて, **④口語記事が中心である**ことを条件とした。

以上の①～④の条件を満たす資料としては,

『キング』(旧：大日本雄辨講談社 / 現：講談社刊行, 1925 年~1957 年)
『改造』(改造社刊行, 1919 年～1955 年)
『中央公論』(旧：反省会 / 現：中央公論新社刊行, 1887 年～現在)

の 3 誌[1]があげられる。この 3 誌からさらに 1 誌を選定する上では, 『太陽』が後年になるほど総ルビ化していったことを受けて, **⑤総ルビ**(『キング』は総ルビ, 『改造』と『中央公論』はパラルビ)であること, その他, **⑥戦時下の言語統制の影響の少なさ**(『改造』と『中央公論』は横浜事件をきっかけに 1944 年に軍部から廃刊勧告を受けた後, 1946 年に両誌とも復刊), **⑦当時最も読まれた**ことを重視して, 『キング』を構築対象として選定した。

⑤の総ルビについては, 特にルビがないと読み分けのできない事例において, 整備がしやすいという利点もある。⑥については, 『キング』も題名が敵性語であることから, 1943 年から 1945 年にかけて『キング 改題　富士』とされる(1946 年以降は『キング』に戻される)が, 内部の記事構成などに大きな変化はなく, 『改造』や『中央公論』に比して, 言論の統制の影響は少なかったと考えられる。⑦については, 2.2 で述べる。

以上の①～⑦の条件から, 総合的に判断して, 『キング』を『太陽』に続く資料として選定し, コーパスを構築することとした。

２.２　雑誌『キング』の概要

『キング』は, 1925(大正 14)年から 1957(昭和 32)年にかけて大日本雄辨講談社(現講談社)より刊行された, 総合雑誌である。『キング』が創刊される以前の大正期には, 大日本雄辯会講談社からは『雄辯』(政治), 『面白倶楽部』『少年倶楽部』『少女倶楽部』(児童向け), 『婦人くらぶ』(家庭, 女性向け)などが発行され, 特定の層にターゲットを絞る雑誌の「細

[1] 同条件を満たす資料としては『文芸春秋』もあげられるが, 同誌は 1930 年代までは文学色が非常に強く, ①総合雑誌としてのジャンルの不偏性が薄いことが危惧されたため, 候補から除外した。

部化」が進められていた。しかしながら，上記の『キング』の方針のように，万人に読まれる雑誌を目指し，1920 年に『雄辯』を総合雑誌化した『現代』を創刊，その流れを受けて，少年少女向け，女性向けの記事も含んだ『キング』の刊行に至った(佐藤 2002)。

創刊号には，『キング』の編集方針について，以下の(1)の文言が載せられた。近代化に伴い，書物が一部の学のある人間のためのものから，国民誰しもが読むことのできる媒体に変容したことを受けて，年齢や身分に関係なく読むことができる雑誌を目指していたことが分かる。全ページにルビが施されているのも，読者層の拡大を目指した方針の一つであると考えられる。『キング』創刊号には同様にキャッチフレーズとして「一家一冊」「面白くてためになる」「国民大衆雑誌」と謳われており，雑誌では史上初めて 100 万部の売り上げを達成している。創刊号の発行部数 74 万部は，大正 14 年の『中央公論』の 8 万部，『主婦の友』24 万部(日本近代文学館 1977)と比較しても，群を抜いて多く，また，1928 年 11 月号の最高発行部数 150 万部は，雑誌誌上最高であり(社史編纂委員会 1959)，『キング』はこの時代もっとも広く読まれた雑誌であると考えられる。その時代に最もよく読まれた雑誌という点では，既存の『国民之友』と『太陽』にも共通し(近藤 2014)，両誌との連続性があると考えられる。

(1)「職業・階級・貧富貴賤の差別なく，老若男女，知識あるものも，知識なきものも,翕然として茲に集まり，限りなき興味を以て耽読しつつある間に，自ずから高尚なる気品と，堅固なる道念とを涵養せられ，一世是に由ってその風を改むるに至らんこと」

『キング』には全年齢層が楽しめるように記事のジャンルが幅広く取られており，教養的なものでは文学，哲学，科学，伝記，教訓，経済，軍事など，娯楽的なものでは大衆小説，講談，落語，童話，踊り，替え歌などがあり，漫画やなぞなぞといった子供向けのページも盛り込まれている。著者層も非常に幅広く，北原白秋や菊池寛，佐々木味津三といった時代を代表する作家から，近衛文麿や齋藤実をはじめとした政治家，その他にも軍人，実業家，漫画家，翻訳家，医者，教師，噺家，俳優，主婦，など多岐にわたっている。この点からも，『キング』はジャンルや著者層において，多様性が担保されているといえるだろう。

２．３　『キングコーパス』の設計

２．３．１　規模

近代語の雑誌コーパスが，『国民之友』(1887/1888 年)，『太陽』(1895 年, 1901 年, 1909 年, 1917 年, 1925 年)のようにおおよそ 8 年おきに構築されていることを受け，『太陽』の最終年の 1925 年から 8 年おきに, 1933 年(昭和 8 年)と 1941 年(昭和 16 年)を構築対象とした。規模としては, 35 万語～40 万語を想定し, 1933 年は 2 月・6 月号, 1941 年は 2 月・6 月・10 月号[2]の全文(表紙・目次・奥付・図表・広告，及び新刊紹介や懸賞の当選者発表記事など固有

[2] 1933 年の 2 冊, 1941 年を 3 冊としたのは，戦時下の 1941 年に紙の配給が激減したことを受け，雑誌のサイズ・ページ数を削減した(講談社八十年史編集委員会 1990)ため，年によってテキストの分量が大きく変わるためである。1933 年は 1 冊あたり 600 ページ前後，1941 年は 300 ページ前後である。

名詞の羅列が中心で本文が少ない記事は除く)を電子テキスト化した。以下の表 1 には,『太陽』,『キング』のそれぞれの年次別の文字数, 語数[3], 記事数, 著者数を示した。

表 1　『太陽』と『キング』の年次と規模

雑誌名	年	文字数	語数	記事数	著者数
『太陽』	1895	3340782 字	202 万語	722 記事	1000 名超
	1901	3239029 字	197 万語	624 記事	
	1909	3102795 字	187 万語	746 記事	
	1917	2960783 字	180 万語	527 記事	
	1925	3423588 字	203 万語	1063 記事	
『キング』	1933	645773 字	37 万語	155 記事	123 名超
	1941	684350 字	41 万語	152 記事	110 名超

(各情報は,『太陽』は国立国語研究所 2005・服部ほか 2016 より)

『キングコーパス』は年次およそ 40 万語で,『太陽』の各年次からは, 約 20%程度の規模となる。この規模の差は非常に大きく, 特に頻度が低い周辺的な語ほど,『キングコーパス』には出現しにくいことが考えられるが, ある程度の頻度がある中～高頻度語を対象とすれば, コーパスとしての不偏性は担保されると考えられる。

２．３．２　記事ジャンル

次に, 図書館での書籍の分類に用いられる「日本十進分類法(NDC)」を利用して,『太陽』(国立国語研究所 2005)と『キング』の記事ジャンルの記事数を表 2 に示す。それぞれの( )内には, 比率(%)を示した。

表 2　『太陽』と『キング』の年次別ジャンル一覧

NDC	1895 年	1901 年	1909 年	1917 年	1925 年	1933 年	1941 年
0 総記	22(3)	9(2)	46(7)	46(7)	218(21)	9(6)	0(0)
1 哲学	45(6)	46(8)	13(2)	8(2)	12(1)	15(10)	8(5)
2 歴史	78(11)	60(10)	34(6)	10(2)	83(9)	9(6)	6(4)
3 社会科学	201(29)	248(42)	288(47)	195(42)	191(23)	11(7)	45(30)
4 自然科学	36(5)	28(5)	34(5)	14(3)	33(5)	1(1)	5(3)
5 技術.工学	41(6)	35(7)	18(3)	16(4)	70(9)	4(3)	7(5)
6 産業	51(7)	51(9)	20(28)	6(2)	38(5)	0(0)	2(0)
7 芸術.美術	60(9)	14(2)	3(0)	46(11)	74(8)	40(26)	21(14)
8 言語	12(2)	4(1)	3(0)	2(0)	2(0)	1(1)	0(0)
9 文学	153(22)	97(16)	133(22)	77(17)	147(18)	65(42)	58(38)
計	699	592	617	467	868	155	152

『キング』と『太陽』を比較すると,『キング』の方が「文学」の比率が大きい。これは 1 記事 500 字に満たない小エッセイや句集のような記事が多いことも寄与していると考えられるが, 文学色の強かった媒体であったといえる。『太陽』で最も比率の高かった「社会科学」は,『キング』では, 1941 年は, 政治・経済・軍事など幅広い記事が収録されているが, 1933 年では 1941 年に比して軍事や戦争にまつわる記事がなく, 比率が低くなったと考

[3] 国立国語研究所の短単位規程による。『キング』については, Web 茶まめによる形態素解析による結果である。

えられる。『キング』1933 年では，「社会科学」の比率が低い代わりに，「芸術.美術」の比率が高いが，絵画などに加えて特に演劇・スポーツ・娯楽など生活に密接な内容の記事が多いことが寄与していると考えられる。また，1925 年は「総記」の比率が極めて高くなっているが，国立国語研究所(2005)によると，これには新刊紹介の記事が多数含まれており，『キングコーパス』では新刊紹介を除外したことが，比率の差に表われたのだと思われる。

以上観察したとおり，『キング』1933 年は「社会科学」の比率が「文学」より小さい点で『太陽』とは不連続であるといえ，『キングコーパス』を用いた調査においては，この相違点を把握した上で分析する必要がある。ただし，幅広いジャンルの記事を収録しており，ジャンルの不偏性という点は担保されているといえるだろう。

２．３．３　文体

文体についても検討する。近代語コーパスでは，文末辞にしたがって，「なり/たり/あり/けり/り/つ/ぬ/き/べし」をとれば文語，「です/ます/ござる/である/だ/た」をとれば口語と認定されている。「中納言」ではその他に「韻文」と「混在」を認めているが，本稿では国立国語研究所(2005)にしたがい，どちらがより中心的であるかによっていずれかに判別した。表 3 には，年次別の文体別記事数と比率を示した。

表 3　『太陽』と『キング』の年次別文体別記事数

記事	1895 年	1901 年	1909 年	1917 年	1925 年	1933 年	1941 年
総記事数	729	635	652	504	889	156	152
口語記事	39(5%)	168(27%)	376(58%)	355(70%)	835(94%)	153(98%)	150(99%)
文語記事	689(95%)	467(74%)	267(41%)	129(26%)	52(6%)	3(2%)	2(1%)

『太陽』では明治期から大正期にかけて，文語記事が減少し，口語記事が増加する。『キング』はその流れの上にあるといえ，韻文(北原白秋「丈夫の唄」，句集「近作玉什」など)を除いて全て口語で書かれている。このことから，『太陽』をはじめとした明治大正期のテキストの口語化の延長上に，『キング』もあるといえるだろう。

２．３．４　品詞分布

最後に，品詞の比率を確認する。図 2 には，『太陽』と『キング』のそれぞれの品詞分布を示した。

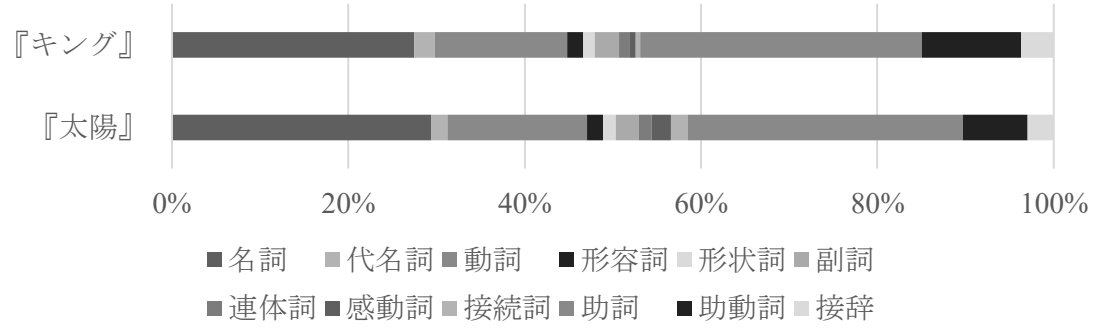


図 2　『太陽』と『キング』の品詞分布

名詞・助詞・動詞・助動詞の順に比率が大きくその他が小さい点で共通している。『キング』の方が助動詞の比率がやや大きいが，全体として品詞分布は連続しており，大きな差異はないと考えられる。

２．４　『キングコーパス』構築のまとめ

以上,2 節では,『太陽』に続く資料の選定方法と, 選定した『キング』を基に構築した『キングコーパス』の規模の設計を示した。2.3.2 からは, 記事ジャンル・文体・品詞分布の観点から『太陽』と『キング』の連続性を検討した。構築した『キングコーパス』は, ジャンルの多様性, 執筆陣の多彩さ, 文体, 品詞分布で連続性を確認し,『太陽』と並び用いるに耐えうるデータとなったと考える。一方で, コーパスの規模は『太陽』に比べて大きく劣り, 記事ジャンルの構成比率にも差異が見受けられた。分析対象の選定や用例分析の際には, この点を踏まえて調査する必要があることも確認した。

３.『キングコーパス』の活用

３．１　「当用漢字音訓表」採用訓との一致度の変遷

3 節では,『太陽』と『キングコーパス』を用いて, 構築の従来の目的であった 1946 年の「当用漢字表」との対比を行う。

「当用漢字表」は,1946 年に出された国語政策であり, 公的文書や教科書, 新聞, 出版物などにおいて使用できる文字の制限を目的としている。1948 年の「当用漢字別表」, 文字の音訓を制限した「当用漢字音訓表」, 字体を制限した 1949 年の「当用漢字字体表」,1973 年の「当用漢字改定音訓表」の一連の告示をまとめて「当用漢字」と呼ばれ,1981 年の「常用漢字表」の告示と共に廃止された。

本研究では,1946 年の「当用漢字表」での採用字, 及び 1948 年の「当用漢字音訓表」での採用訓と, 近代雑誌コーパスにおける和語の表記の実態が, どの程度一致し, また変遷したかを分析する。

３．２　調査方法

３．２．１　語区分の設定

近代語のコーパスにおける語の区分は「語彙素」に依っており, 語彙素は電子化辞書「UniDic」の区分に従って認定されている。「UniDic」の語区分は,『日本国語大辞典』を含む複数の辞書の見出し区分と, 現代語のコーパスの用例を参考に区分がなされている(小椋ほか 2011)。例えば,「UniDic」ではナラウという語を「習う」と「倣う」の 2 語にわけている。これは, ヲ格をとるかニ格をとるかによって意味が明確に分けられるための区分と思われるが, 近代では「古い習慣に習う」といったように, 格成分が必ずしも一致せず, また意味と表記の対応関係にも揺れが認められるという問題点が指摘できる。このことから, 近代語のコーパスを用いた研究では, 近代語の研究に合った区分の設定が望まれる。そこで, 本稿では,『日本国語大辞典』の見出しの区分に従って語を区分し, ナラウの例では, 同一見出しとしているので, 単一の語ナラウとして扱う。

３．２．２　調査対象語の選定

年次 40 万語規模の『キングコーパス』において, 1 年に 20 件程度の頻度が確保できるように,100 万語あたりの相対頻度 50 以上の語,『太陽』と『キングコーパス』を合わせた統合コーパス(約 1050 万語)における頻度 466 以上の語を対象とする。この条件に該当する和語の自立語は 650 語あり, 本稿ではそのうち最も語数の多い動詞 267 語を対象とする。このうち, ①「入る(ハイル/イル)」や「出る(デル/イデル)」などルビがないと読み分けができない 19 語, ②連用形がいずれの語か判別できない「借り(カリル/カル)」, ③誤解析のパターン・数が突出して多く修正ができない「為る」, ④複合語の「出来る」の 22 語とを除外した 245 語を, また, 本稿では,「当用漢字音訓表」で訓が採用されていない 22 語についても除外し,223 語を対象とする。

３．２．３　表記対応一致レベルの判定

223 語はそれぞれ，「当用漢字音訓表」で単一表記語，「当用漢字音訓表」で複数表記語に分けられる[4]。表 4 には，それぞれの語数と比率，いくつかの語例を示した。

表 4　分類別の語数と比率

表記実態	語数(比率)	語例
単一表記	194 語(78.9%)	仰ぐ，遊ぶ，与える，当たる，預かる，集まる，集める，当てる，怪しむ，誤る，争う，改める，有る，歩く，言う，生きる，行く，急ぐ，至る，入れる，伺う…
複数表記	29 語(11.8%)	合う，上げる，表わす，現われる，歌う，打つ，生まれる，犯す，送る，起こす，起こる，収める，押す，顧みる，変える，変わる，聞く，差す，立てる，付く…

これら 223 語の表記の対応が，どの程度一致しているのかを各年次で判定するために，「当用漢字音訓表」にて各語の表記として採用されていない表記の使用記事率を利用する。冒頭に挙げたタノムを例に判定する。各表記の使用記事率は，その語が出現した記事数を 100%とした時の，各表記の出現記事数から算出した。以下の表 5 にはタノムの表記別の使用記事率を示したが，このうち「当用漢字音訓表」にてタノムの表記として採用されていない「恃」の使用記事率に注目する。この数値を x とすると，「x=0」の場合 A，「0<x≦5」の場合 B，「5<x≦20」の場合 C，「10<x」の場合 D といったように，4 つのレベルに判定する。

表 5　タノムの表記別使用記事率と表記対応一致レベルの推移

表記	1895	1901	1909	1917	1925	1933	1941	当用
「頼」	57.0%	67.2%	75.7%	81.5%	89.5%	90.5%	92.9%	○
「恃」	32.3%	31.3%	25.7%	20.0%	8.1%	0%	3.6%	×
判定	D	D	D	D	C	A	B	

タノムの場合，1895 年から徐々に「恃」が使用されなくなり，特に 1933 年以降はタノムにはほぼ「頼」しか使用しなくなったことがわかる。このような判定を 223 語全てに行う。本研究では，A または B であれば「当用漢字音訓表」と一致，C または D であれば「当用漢字音訓表」と不一致と考える。

３．３　調査結果

『太陽』と『キングコーパス』の年次別に，動詞 223 語につき 3.2.3 で判定した表記対応一致レベルの比率を示すと，図 3 の通りになる。

図 3 にみる通り，A+B の比率が 1895 年から 1941 年にかけて 20 ポイントほど上昇し，後年になるほど「当用漢字音訓表」との一致度が上昇していると考えられる。特に 1917 年から 1925 年の上昇が大きく，大正末期頃から，「当用漢字音訓表」の用字法により近い用字法になったといえるだろう。

また，レベル別に注目すると，D レベルも 1925 年を境に大きく減少し，D レベルから ABC レベルに移行する語が多いことがわかる。反対に，A レベルも増していく傾向にある

[4] 「当用漢字音訓表」では，動詞の自他，あるいは動詞形・形容詞形・形容動詞形・名詞系など複数の形のうち一つを示し，原則的に，他の形にも用いてよい旨が記載されている。例えば，「苦」には「ク | くるしい | にがい」とあるが「くるしむ」にも対応し，「変」に「ヘン | かわる」とあるが「かえる」にも対応していると考える。

ことが指摘できる。1933 年や 1941 年で特に A レベルが多いのは，コーパスの規模が小さいために，逸脱した限定的な表記が出にくく，B レベルが少なくなったことによるものであると考えられるが，一方で A レベル増加の延長線上にあることは間違いないだろう。

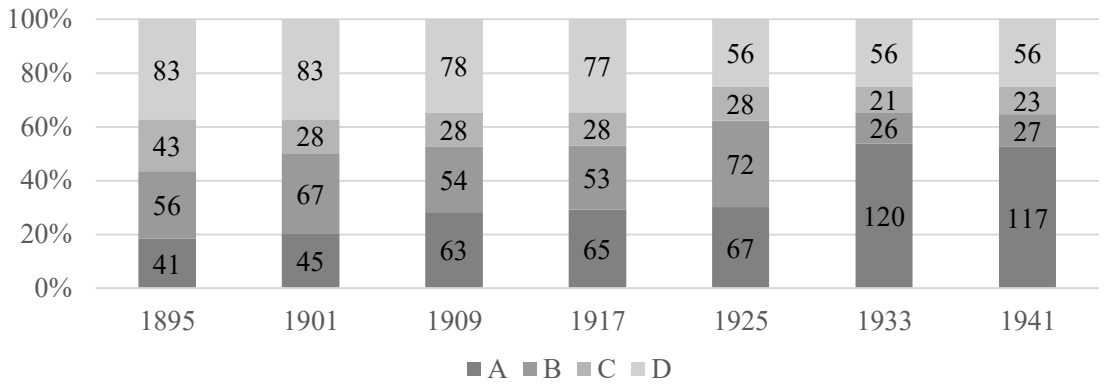


図3　年次別表記対応一致レベルの推移

この結果から，和語の表記は，近代において経年変化で「当用漢字表」ならびに「当用漢字音訓表」の用字法に近づいていく，言い換えると，「当用漢字表」の設計の背景に，それまでの近代における用字法の変化があったことが窺える。最後に，1941 年の段階でも D レベルにある 56 語を観察すると，表 6 のように，4 つのタイプに分類できる。②や③のタイプにも多く語が属していることを考えると，この調査の発展として，「当用漢字別表」や「常用漢字表」との対比や，1945 年以降の調査が望まれるだろう。

表 6　不一致語のタイプ

概要	語例
①意味・頻度などで第一表記と第二表記が主要と限定の関係にあるタイプ(19 語)	逢う，窺う，射つ，捺す，斬る，喰う，応える，覚る，棲む，経つ，喰べる，吐(ツ)く，称える，啼く，呑む，貼る，護る，赦す，亙る，嗤う
②「常用漢字表」で訓が採用されるタイプ(12 語)	挙げる，在る，聴く，指す，闘う，就く，務める，倣う，分かれる
③一致に向かう(表記の淘汰)途中段階にあるタイプ(17 語)	与る，蒐める，検める，到る，吝しむ，怖れる，随う，棄てる，斃れる，援ける，為す，睡る，遺す，初まる，以(モ)つ，依る
④上記 3 分類の当てはまらない(7 語)	云う，唄う，較べる，乞う，蒙る，捉える，儲ける

４．おわりに

本稿では，『太陽』に続く昭和期の調査資料として，総合雑誌『キング』を選定し，その設計と活用について記した。2 節では，資料の選定方法とコーパスの設計について検討し，コーパスの規模や記事ジャンルの構成比率に差異はあるものの，ジャンルや著者の多様性・文体・品詞分布においては『太陽』との連続性のあることが確認できた。3 節では，『太陽』と『キングコーパス』を併用して，近代雑誌コーパスと「当用漢字音訓表」にお

ける和語の動詞の用字法の一致度の推移を調査し，明治期から昭和期にかけて経年変化で，用字法の一致度が高くなり，「当用漢字音訓表」の設計の背景に近代における用字法の変化があることが明らかになった。今後の課題としては，「常用漢字表」や「当用漢字別表」との対比，「当用漢字表」採用基準との関係，「当用漢字音訓表」不採用語の用字法の調査が挙げられる。

コーパスの構築については，規模の拡張や XML タグの整備など課題は尽きないが，現状のコーパスの規模や記事ジャンルの構成比率に注意しつつ，精力的に『キングコーパス』を活用していきたい。

謝　辞

本稿は,日本学術振興会特別研究員奨励費 17J03579「近代における和語の表記の変遷」(代表：高橋雄太)，および，国立国語研究所機関拠点型基幹研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」(プロジェクトリーダー：小木曽智信)の成果の一部である。

また，本稿で取り上げた『キングコーパス』の構築に際して，明治大学の田中牧郎先生，国立国語研究所の小木曽智信先生，間淵洋子さん，近藤明日子さんをはじめに多くの方々に助言・助力を賜った。記して感謝申し上げる。

文　献

小椋秀樹，小磯花絵，冨士池優美ほか(2011)『『現代日本語書き言葉均衡コーパス』形態論情報規程集　第 4 版　(下)』国立国語研究所.

講談社八十年史編集委員会(1990)『クロニック講談社の 80 年』講談社.

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

近藤明日子(2014)「『国民之友コーパス』解説書 第 1.1 版」2018 年 7 月 23 日確認.
<http://pj.ninjal.ac.jp/corpus_center/cmj/doc/kokumin_manual_v1_1.pdf>

佐藤卓己(2002) 『『キング』の時代：国民大衆雑誌の公共性』岩波書店.

社史編纂委員会(1959) 『講談社の歩んだ五十年　昭和編』講談社.

髙橋雄太(2015)「『太陽コーパス』における和語動詞「あう」の用字法」『第 7 回コーパス日本語学ワークショップ予稿集』国立国語研究所, pp195-202.

髙橋雄太(2016)「近代における和語の表記の変遷　―複数表記から単一表記へ―」『国際日本学研究論集』第 4 号 ,明治大学大学院, pp.37-48.

田中牧郎(2012)「近代語コーパスにおける資料選定の考え方」『近代語コーパス設計のための文献言語研究　成果報告書』国立国語研究所, pp.13-26.

日本近代文学館(1977) 『日本近代文学大事典　第 5 巻　新聞・雑誌』講談社.

服部紀子，間淵洋子，近藤明日子，小木曽智信(2016)「『日本語歴史コーパス　明治・大正編 I雑誌』ver.1.0 の公開」『日本語学会 2016 年度秋季大会予稿集』日本語学会, pp.157-162.

文化庁「当用漢字音訓表」2018 年 7 月 23 日確認.
<http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kakuki/syusen/tosin04/index.html>

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/

Web 茶まめ　http://chamame.ninjal.ac.jp/

『明六雑誌』『東洋学芸雑誌』の特徴語から見る明治前期書き言葉の語彙特性

近藤 明日子（国立国語研究所コーパス開発センター）†

The Characteristic of the Written Japanese Vocabulary in the Early Meiji era: An Analysis of the Specialized Vocabulary of "Meiroku Zasshi" and "Toyo Gakugei Zasshi"

KONDO Asuko (National Institute for Japanese Language and Linguistics)

要旨

明治前期の語彙の特性を明らかにすることを目的として、明治前期の書き言葉を代表する資料『明六雑誌』『東洋学芸雑誌』と明治中期以降の書き言葉を代表する資料『国民之友』『太陽』との語彙の頻度を比較し、『明六雑誌』『東洋学芸雑誌』に有意に高頻度な語（特徴語）を抽出し、その特性を考察した。その結果、①文語体・漢文訓読文由来の語、②一字漢語、③新しい事物・概念を表すための新語で後に別語に置き換わった語、④新しい事物・概念を表すための新語で後に事物・概念の衰退とともに衰退した語、⑤『明六雑誌』『東洋学芸雑誌』で特にとりあげられた話題と関連する語、という主に 5 種の類型の特性を有することが明らかになった。

１．はじめに

日本語史において、明治・大正期は約 60 年という短期間に書き言葉が急激に変化した時期として特徴づけられる。開国による新しい事物・概念の流入による新漢語の発生等により語彙も大きく変化した。また、言文一致の進展に伴う文語体から口語体への文体の転換は語彙にも大きな変化をもたらした。

本研究では、明治・大正期の始まりにあたる明治 0 年代から 10 年代（以下、「明治前期」と呼ぶ）の書き言葉の語彙について、それ以降の時期の書き言葉の語彙と比較して、その特性を考察する。使用する資料は、明治前期の書き言葉の資料『明六雑誌』『東洋学芸雑誌』と、その比較対象とする明治 20 年代から大正期の書き言葉の資料『国民之友』『太陽』である。これらの資料はすべてコーパス化されており、その形態論情報を利用して、明治前期の資料の語彙の頻度とそれ以降の年代の資料の語彙の頻度を算出、計量的に比較し、明治前期の資料に有意に高頻度な語（特徴語）を抽出する。そしてその類型の考察を通じて、明治前期の語彙の特性を明らかにしていく。

２．分析方法

２.１　使用する資料

本研究では明治前期の書き言葉を代表する資料として『明六雑誌』と『東洋学芸雑誌』を使用する。前者の『明六雑誌』は、明治 6（1873）年に学術啓蒙を目的に結成された明六社の機関誌で、明治 7（1874）～8（1875）年に 1 号から 43 号まで発行された。明六社社友 16 名によって書かれた、政治・経済・社会・教育・言語等の幅広いジャンルの 155 の

† kondo@ninjal.ac.jp

論説が掲載されている。明治 0 年代における書き言葉の代表性を相当程度担保した資料と見なし得る。『明六雑誌』は国立国語研究所（2017）『日本語歴史コーパス 明治・大正編Ⅰ雑誌』（短単位データ 1.1）（以下、「明治・大正編 I 雑誌」と呼ぶ）に 1～43 号の全号が収録されており、本研究ではこのデータを使用する。後者の『東洋学芸雑誌』は、明治 14（1881）年から昭和 5（1930）年にかけて 567 冊刊行された学術総合雑誌である。自然科学の啓蒙を目的としながら、特に創刊当初は文芸等の他のジャンルの記事も掲載し、多くの読者を得た[1]。本研究では、創刊当初の 1～15 号（初版、明治 14～15 年刊）に着目し、49 名の著者・訳者[2]による幅広いジャンルの計 130 記事[3]を掲載した、明治 10 年代における書き言葉の代表性を担保した資料と捉え、使用する。『東洋学芸雑誌』1～15 号（以下、単に『東洋学芸雑誌』と呼ぶ）は現在、国立国語研究所でコーパス化が進められており[4]、その 2018 年 7 月時点でのデータを使用した[5]。このデータの仕様は「明治・大正編 I 雑誌」に準拠しており、『明六雑誌』のデータと『東洋学芸雑誌』のデータを併せて分析することが可能である[6]。

これらの資料から明治前期の書き言葉の語彙特性を明らかにするためには、比較対象となる資料が必要である。本研究では「明治・大正編 I 雑誌」に収録された『国民之友』明治 20（1877）・21（1878）年刊行の 1～36 号および『太陽』明治 28（1895）・明治 34（1901）・明治 42（1909）・大正 6（1917）・大正 14（1925）年刊行の計 60 号分[7]のデータを使用する。「明治・大正編 I 雑誌」は「明治・大正時代の書き言葉を広く見渡し、年代ごとの変遷をたどれるように、各年代を代表する雑誌から一定の年数ごとの刊行分を収録し」[8]ているコーパスであり、その中の『国民之友』『太陽』のデータは、明治 20 年代から大正期（以下、「明治中期以降」と呼ぶ）の書き言葉の代表性を担保した資料として使用することができる。

各コーパスのデータ[9]の中から、非文芸ジャンルのサンプル[10]を調査対象として選定し、その地の文を調査対象とした[11]。よって、本研究は論説・報道等の非文芸ジャンルの文章の

---

1 『改訂新版　世界大百科事典』「東洋学芸雑誌」項、
https://japanknowledge.com/psnl/display/?lid=102005214300
2 翻訳記事の場合、原著者ではなく訳者を数えた。
3 漢文体の記事は除く。
4 将来的に「明治・大正編 I 雑誌」の一部として公開される予定である。
5 構築中のコーパスのため、以下の分析では公開版とは異なる値が算出される場合がある。
6 「明治・大正編 I 雑誌」のデータ仕様については間淵・近藤・服部（2017）を参照のこと。
7 「明治・大正編 I 雑誌」に収録されている『太陽』巻号の詳細は
http://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html を参照のこと。
8 http://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html
9 コーパスデータは国立国語研究所のデータベースに格納されたものを使用し、コーパスの言語量や各語の粗頻度を算出した。
10 サンプルとは、コーパス収録対象として選定されたひとまとまりのテキストの範囲を指す。本研究で使用するコーパスでは、雑誌に掲載された各記事を単位としてサンプルが定められており、サンプル＝記事と捉えてほぼ支障はない。「明治・大正編 I 雑誌」では各サンプルにジャンル情報として「文芸」または「非文芸」の分類が付与されており、「非文芸」のサンプルを選定した。なお、コーパスには雑誌本体の構造要素をあらわすテキスト（誌名・欄名等）を収録するサンプル（サンプル ID が「000」で終わるもの）があるが、本研究ではこのサンプルは対象外とした。
11 サンプル中にある他文献からの引用部分や登場人物の発話部分は、サンプルの地の文とは位相を異にするため、調査対象外とした。また、漢文体や外国語で書かれた部分も調査対象外とした。

語彙について考察するものであり、小説・戯曲・詩歌等の文芸ジャンルの文章の語彙については言及しない。この非文芸ジャンルのサンプルの地の文の延べ語数・異なり語数を文語・口語の文体別に示したものが表 1 である。語数のカウントでは、コーパスの採用する「短単位」を 1 語とし、記号類・未知語は除いた。また、異なり語数は、コーパスの短単位の形態論情報のうち「語彙素・語彙素読み・語彙素細分類・語種・品詞」の 5 要素が一致するものを同語と見なし、カウントを行った。

表 1　調査対象資料の言語量

資料	文語		口語		全体	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
明治前期	334,605	19,875	8,164	1,652	342,769	20,215
明六雑誌	157,483	12,201	8,164	1,652	165,647	12,692
東洋学芸雑誌	177,122	13,038	0	0	177,122	13,038
明治中期以降	4,274,873	65,445	3,771,357	56,161	8,046,230	81,151
国民之友	846,992	29,207	6,893	1,075	853,885	29,295
太陽	3,427,881	59,753	3,764,464	56,114	7,192,345	76,761

２.３　特徴語の抽出

表 1 にあげた明治前期資料に出現する 20,215 語（異なり）について、明治中期以降の資料での出現頻度と比較して、有意に高頻度であることの程度（特徴度）を指数化し、特徴度の上位の語を明治前期の特徴語として抽出する。語の頻度の高低の程度を計る統計的指標は複数あるが[12]、本研究では対数尤度比（log-likelihood ratio）を使用する。対数尤度比による特徴語抽出は、英語学で一定の評価を得ており（石川 2008、p.99）、日本語史研究においても、古典作品の特徴語の抽出（宮島・近藤 2011）や古典作品内部の文体別特徴語の抽出（小木曽 2015）等に利用されている。対数尤度比やそれを補正した特徴度の算出方法は宮島・近藤（2011）に拠り、以下のように行った。

対数尤度比=2(alna+blnb+clnc+dlnd-(a+b)ln(a+b)-(a+c)ln(a+c)-(b+d)ln(b+d)
-(c+d)ln(c+d)+(a+b+c+d)ln(a+b+c+d))[13]

a: 対象資料での語 W の度数
b: 参照資料での語 W の度数
c: 対象資料の延べ語数-a
d: 参照資料の延べ語数-b

本研究では明治前期の資料（『明六雑誌』『東洋学芸雑誌』）が対象資料、明治中期以降の資料（『国民之友』『太陽』）が参照資料となる。語 W の対象資料での相対頻度が参照資料での相対頻度より低い場合、対数尤度比に-1 を乗じる補正を行い、その値を語 W の特徴度とする。特徴度は、語 W の対象資料での相対頻度と参照資料での相対頻度が等しい場合 0 となり、対象資料での相対頻度が参照資料での相対頻度より大きい場合は正の値、対象資料

[12] 内山・中條・山本他（2005）には、8 種の単独指標と単独指標を複数組み合わせた複合指標 1 種の計 9 種の指標があげられている。

[13] ln は自然対数を表す。a または b が 0 の場合、alna または blnb を 0 と見なし計算する。

での相対頻度が参照資料での相対頻度より小さい場合は負の値となる。正の値の場合、その値が大きければ大きいほど参照資料に比べ対象資料で高頻度に出現する程度が高く、負の値の場合、その値が小さければ小さいほど参照資料に比べ対象資料で低頻度に出現する程度が高いことを示す。

算出した特徴度の降順上位 100 語を表 2（稿末）に示す。表中、語彙素読み・語彙素細分類は省略した語がある。また品詞は大分類のみを示した。この明治前期において特に特徴的な 100 語を考察し、この時期の語彙の特性を明らかにする。

３．特徴語の類型から見る明治前期の語彙特性

表 2 にあげた明治前期の特徴語を検討した結果、以下の①～⑤の 5 種の類型におよそ分類できた。明治前期の特徴語は、明治中期以降との比較から抽出されたものであるから、「明治前期では使用されたが明治中期以降に衰退し使用されなくなった語」が主に含まれる。それに該当する類型が①～④である。他に、「語の衰退以外の要因で明治前期に高頻度になった語」も含まれており、それに該当する類型が⑤である。そして①～⑤にあてはまらない語を⑥として分類した。特に①～⑤の類型の性質が明治前期の語彙の特性を表すものとなる。以下、①～⑥の類型とそこに所属する語を詳しく見ていく。

①　文語体・漢文訓読文由来の語

表 1 での文語・口語別の語数から分かるように、明治前期は文語体が主流で口語体がほとんど見られなかったものが、明治中期以降の特に『太陽』において文語体から口語体への転換が起き、明治中期以降では全体のおよそ半分が口語体なる。このことから、明治中期以降と比較して抽出された明治前期の特徴語のなかには、文語体に特徴的な語が多く見出され、これを第一に分類した。

典型的なものとして、文語助動詞「なり」「ず」「べし」「む」「非ず」「ごとし」、係助詞「や」、接続助詞「雖も」「ども」「に」がある。また、接続助詞「ば」のうち確定条件用法は口語体では使用されない文語体特有のものとしてあげられる。この他に、文語体の中でも漢文訓読体に特徴的な語句で使用される語が多く見られる。山田（1935）・吉田他（編）（2001）・松崎（2006）に挙げられている漢文訓読に特徴的な語句で使用される語として、名詞「曰く」「故」「所」「所以」、代名詞「此れ」、動詞「然る」「持つ」（～ヲ以テ）「得る」（可能表現「～ヲ得」）「用いる」「欲する」、形容詞「如此し」、副詞「蓋し」「凡そ」「豈」「必ず」、接続詞「夫れ」「即ち」「而して」「且つ」、副助詞「のみ」があげられる。また、先にあげた助動詞・助詞のうち「べし」「非ず」「ごとし」「雖も」も漢文訓読に特徴的な語である。また名詞「方今」・代名詞「余」「小子」も漢文訓読由来の語と見なせると考えられる[14]。

松崎（2006）では明治期の理科教科書を用いて、文語体の実用的文章での漢文訓読的性格を有する語の出現状況を調査し、刊行年の古いものにより顕著に出現することを明らかにしている。本研究から、雑誌の非文芸ジャンルの文章でも、刊行年の古い明治前期のものに漢文訓読由来の語が多く使用される傾向があることが明らかになった。

[14] この他、「者」「其の」「又」「皆」「至る」や格助詞「を」（～ヲ以テ、～ヲ得）も漢文訓読由来の語法により特徴語となったものと推測するが、なお検討したい。

②　一字漢語

次に分類するのは一字漢語である。「説」「理」「学」「害」「法」「書」「葉」「意」「論」「利」「言」「異」、また一字漢語サ変動詞「論ずる」「変ずる」「概する」があげられる。一字漢語については、田中（2013、pp.106-107、174-177）で明治中期から大正期、大正期から昭和期にかけてそれぞれ衰退する語の一類型として指摘されている。本研究から、明治前期から明治中期以降にかけても同様の傾向が見出されたことになる。これらの語のうち、「理」「学」「葉」（ページの意）「意」「利」「言」「異」「変ずる」「概する」は、現代語ではほとんど使用されず、別語に置き換わったと考えられる語であり、その衰退への道程が明治中期以降にはじまっていたことが示唆される。また、語義の一部が別語に置き換わった語としては、「法」（方法の意の場合）と「書」（書物の意の場合）がある。残る「説」「論」「論ずる」は現代語でも普通に使用され、別語への置き換えが想定できないものである。すべて論説に関わる語であることから、『明六雑誌』『東洋学芸雑誌』のサンプルのほとんどが論説文であるのに対し、『国民之友』『太陽』では報道文等の論説文以外のサンプルが増加するという雑誌の性格の差に由来して『明六雑誌』『東洋学芸雑誌』の特徴語になったもので、語の衰退とは直接関係しないものであるかもしれない。

③　新しい事物・概念を表すための新語で、後に別語に置き換わった語

次に分類するのは、明治中期以降別語に置き換わり衰退した語である。②でとりあげた一字漢語の中にも同じ類型に属するものがあったが、ここでは一字漢語以外の語をとりあげる。語種別にあげると和語「民」、漢語「人民」「理学」「駁者」「教門」「行星」「原質」「交易」「交際」「正金」、外来語「リバティー」がある。後に置き換わる別語として考えられるのは、「民」「人民」は「国民」、「理学」は「哲学」「自然科学」、「教門」は「宗教」、「行星」は「惑星」、「原質」は「要素」、「正金」は「正貨」、「リバティー」は「自由」である[15]。また、外国との商取引を指す場合の「交易」が「貿易」に、外国との付き合いを指す場合の「交際」が「外交」にというように、語義の一部が別語に置き換わった場合も見られる。ここにとりあげた語の多くは新しく輸入された事物・概念を表す新漢語であり、明治前期までに使用がはじまりながら、明治中期以降には定着せず短期間で衰退した語ということになる。新漢語のめまぐるしい消長の一端を垣間見せるものと言えるだろう。なかでも「駁者」（反論する者の意）は『日本国語大辞典　第二版』にも掲載されておらず、『東洋学芸雑誌』の 1 サンプル（6 号「東京經濟雜誌に答」）のみに出現する一過性の最たる語である。また、「リバティー」（実際の出現形は「リバーチイ」「リベルチー」「リベルテイ」「リボルチー」と多様）はその概念そのものを解説するサンプルで使用されており、まだ概念自体が新しくそれを表す新語も定着していない時期には、原語のカタカナ表記語を使用する段階があったことを象徴する語である。また、和語「民」は国民を表すために使用されたもので、ここから新しい概念を既存の和語で表したものの定着せず別の漢語（この場合「国民」等）に置き換わる場合もあったことが分かる。

④　新しい事物・概念を表すための新語で、後に事物・概念の衰退とともに衰退した語

次にとりあげるのは、指し示す事物・概念が衰退するとともにそれを表す語も衰退した

15 「諂諛」はこびへつらうことの意で、置き換えの別語は「追従」「媚び」等であろうか。なお考えたい。

と考えられるもので、「開化」「民選」「開明」があげられる。「開化」「開明」は明治前期の文明開化を表す語であり、「民選」は自由民権運動の出発点である明治 7 年の「民選議院設立建白書」をきっかけに明治前期重要な語となったが、それぞれ明治中期以降は別語に置き換わることもなく概念とともに語も衰退したものである。

⑤　『明六雑誌』『東洋学芸雑誌』で特にとりあげられた話題と関連する語

次に分類するのは、『明六雑誌』『東洋学芸雑誌』で特にとりあげられた話題と関連して使用された語で、語形や概念の衰退とは関係がないと考えられる語である。「化学」「紙幣」「化合」「水素」「ミル」「自由」「物」(「化合物」「有機物」等で使用)「原子」「三宝」「議院」「炭素」「泉」(「硫黄泉」「昇騰泉」等で使用)「肉桂」(「肉桂酸」「肉桂油」で使用)「有機」「酸素」「分子」「動脈」「尭舜」「権利」「同権」「ＳＯ」「ニトロ」があげられる。これらの多くは自然科学分野の用語であり、自然科学の啓蒙を目的とした『東洋学芸雑誌』の扱う話題の特徴を捉えた語群と言える。また「ミル」(『自由論』の著者)「自由」「三宝」(人生の幸福を追究するための三要件を「三つの宝」として言う語)「議院」「権利」「同権」は『明六雑誌』に主に出現し、『明六雑誌』の扱う主な話題を象徴する語群と言える。「紙幣」は『明六雑誌』『東洋学芸雑誌』でそれぞれ数号にわたる連載記事で主題となり、「尭舜」は『東洋学芸雑誌』の 2 サンプル（9 号「尭舜は孔教の偶像なる所以を論ず」、13 号「牧都宇氏に答ふ」。どちらも著者は井上圓了）で主題となり、そこで集中的に使用されたために特徴語となったものである。

⑥　その他

最後にとりあげるのは、①～⑤には分類できなかったその他の語である。

「異（コト)」は形容動詞「異なり」や副詞「異に」の形で使用されるもので、それぞれ口語体では動詞「異なる」や副詞「殊に」という別語としてコーパス上は扱われるもので、コーパスの同語異語判別の規定に拠り特徴語となったものである。

「知る」「国」「人」「成す」は特徴語となった背景が不明な語で、今後の更なる考察が必要な語である。

４．おわりに

以上、明治前期の書き言葉の資料『明六雑誌』『東洋学芸雑誌』を、明治中期から大正期にかけての書き言葉の資料『国民之友』『太陽』と比較して、明治前期の特徴語を抽出した。そして、特徴語を主な 5 種の類型に分類し、各類型の性質こそが明治前期の語彙の特性を表すものとして考察した。このように、隣接する時期の語彙との比較から抽出した特徴語の考察は、当該資料に代表される時期の語彙の特性を捉えるのに有効な方法の一つと考えられる。なお、本研究で見出された明治前期の語彙特性は、明治前期と比較して明治中期以降に頻度が低くなった語に焦点をあてることで導き出されたもので、特性の一部を明らかにしたにすぎない。明治前期の語彙特性の全体像を把握するためには、今後、明治前期以前の時期の書き言葉を代表する資料との比較から抽出される特徴語の考察を行い、語彙特性の別の側面を補完していく必要がある。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」、MEXT 科研費 JP15H01883「日本語歴史コーパスの多層的拡張による精密化とその活用」による研究成果の一部である。

文　献

石川慎一郎（2008）『英語コーパスと言語教育』大修館書店.

内山将夫・中條清美・山本英子・井佐原均（2004）「英語教育のための分野特徴単語の選定尺度の比較」『自然言語処理』11:3, pp.165-197.

小木曽智信（2015）「中古和文における文体別の特徴語」『コーパスと日本語史研究』ひつじ書房, pp.93-117.

国立国語研究所（2017）『日本語歴史コーパス　明治・大正編 I 雑誌』（短単位データ 1.1）http://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html

田中牧郎（2013）『近代書き言葉はこうしてできた』岩波書店.

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 Personal」https://japanknowledge.com/personal/ に拠る.

平凡社（編）（2014）『改訂新版　世界大百科事典（第 6 版）』平凡社. ネットアドバンス社提供サービス「ジャパンナレッジ Personal」https://japanknowledge.com/personal/ に拠る.

松崎安子（2006）「明治期の文語文の類型―小学校理科教書を対象として―」『文化』70:1-2, pp.92-105.

間淵洋子・近藤明日子・服部紀子（2017）「『日本語歴史コーパス 明治・大正編 I 雑誌』（短単位 Ver.1.1）テキストの凡例と「中納言」表示項目について」http://pj.ninjal.ac.jp/corpus_center/chj/doc/abstract-meiji-taisho-201703.pdf

宮島達夫・近藤明日子（2011）「古典作品の特徴語」『計量国語学』28:3, pp.94-105.

山田孝雄（1935）『漢文の訓讀によりて傳へられたる語法』宝文館出版. 復刻版 1979 に拠る.

吉田金彦・築島裕・石塚晴通・月本雅幸（編）（2001）『訓点語辞典』東京堂出版.

表 2　『明六雑誌』『東洋学芸雑誌』の特徴語 100 語

順位	語彙素（語彙素読み）［語彙素細分類］	語種	品詞	明治前期粗頻度	明治中期以降粗頻度	特徴度
1	なり[断定]	和	助動詞	10,305	127,977	3308.5
2	者(モノ)	和	名詞	2,078	13,359	2213.2
3	を	和	助詞	19,730	336,296	1824.5
4	此れ(コレ)	和	代名詞	4,619	53,896	1732.0
5	ず	和	助動詞	6,655	95,266	1341.6
6	然る(シカル)	和	動詞	1,490	11,271	1273.4
7	説(セツ)	漢	名詞	547	2,102	976.0
8	曰く(イワク)	和	名詞	671	3,318	954.5
9	其の(ソノ)	和	連体詞	4,852	70,252	933.5
10	べし	和	助動詞	3,185	42,055	839.9
11	理(リ)	漢	名詞	207	185	797.2
12	人民(ジンミン)	漢	名詞	547	2,747	766.1
13	持つ(モツ)	和	動詞	1,955	23,748	656.2
14	理学(リガク)	漢	名詞	203	293	651.6
15	民(タミ)	和	名詞	232	461	638.7
16	又(マタ)	和	副詞	1,090	10,683	599.9
17	雖も(イエドモ)	和	助詞	694	5,331	579.2
18	や	和	助詞	1,282	13,887	571.0
19	開化(カイカ)	漢	名詞	154	169	552.0
20	夫れ(ソレ)	和	接続詞	301	1,169	532.3
21	故(ユエ)	和	名詞	874	8,223	520.2
22	論ずる(ロンズル)	混	動詞	372	1,872	519.8
23	余(ヨ)	漢	代名詞	546	3,915	502.8
24	ども	和	助詞	823	7,656	500.2
25	む	和	助動詞	1,837	24,117	491.8
26	蓋し(ケダシ)	和	副詞	411	2,444	479.6
27	即ち(スナワチ)	和	接続詞	943	9,743	465.7
28	化学(カガク)	漢	名詞	209	624	450.3
29	紙幣(シヘイ)	漢	名詞	150	265	438.4
30	化合(カゴウ)	漢	名詞	110	92	432.8
31	学(ガク)	漢	名詞	142	264	404.6
32	水素(スイソ)	漢	名詞	103	89	401.0
33	皆(ミナ)	和	名詞	492	3,881	394.4
34	ば	和	助詞	2,802	43,821	388.9
35	害(ガイ)	漢	名詞	170	466	387.7
36	民選(ミンセン)	漢	名詞	89	64	366.5
37	至る(イタル)	和	動詞	1,048	12,532	366.2
38	開明(カイメイ)	漢	名詞	88	65	359.6
39	駁者(バクシャ)	漢	名詞	50	0	319.8
40	得る(エル)	和	動詞	1,334	18,267	311.2
41	ミル[Mill]	固	名詞	73	50	304.9
42	に	和	助詞	874	10,497	301.8
43	自由(ジユウ)	漢	名詞	366	2,833	301.8
44	所(トコロ)	和	名詞	1,507	21,711	293.3
45	物(ブツ)	漢	接尾辞	264	1,678	284.4
46	原子(ゲンシ)	漢	名詞	59	22	284.4
47	三宝(サンポウ)	漢	名詞	53	11	281.1
48	知る(シル)	和	動詞	726	8,402	275.9
49	議院(ギイン)	漢	名詞	130	397	275.7
50	教門(キョウモン)	漢	名詞	43	0	275.0
51	如此し(カクノゴトシ)	和	形容詞	108	250	273.2
52	異(コト)	和	名詞	368	3,055	272.1
53	人(ヒト)	和	名詞	963	12,465	268.8
54	行星(コウセイ)	漢	名詞	43	1	265.5
55	而して(シカシテ)	和	接続詞	732	8,672	262.0
56	用いる(モチイル)	和	動詞	353	2,958	257.1
57	凡そ(オヨソ)	和	副詞	224	1,374	252.0
58	豈(アニ)	和	副詞	216	1,286	251.6
59	法(ホウ)	漢	名詞	303	2,335	251.5
60	書(ショ)	漢	名詞	189	1,006	249.2
61	原質(ゲンシツ)	漢	名詞	50	19	240.1
62	葉(ヨウ)	漢	名詞	68	79	238.5
63	炭素(タンソ)	漢	名詞	81	141	238.5
64	泉(セン)	漢	接尾辞	57	39	238.1
65	相(アイ)	和	接頭辞	397	3,826	225.5
66	肉桂(ニッキ)	漢	名詞	37	2	221.0
67	有機(ユウキ)	漢	名詞	73	120	220.9
68	酸素(サンソ)	漢	名詞	62	70	219.8
69	分子(ブンシ)	漢	名詞	112	383	218.9
70	のみ	和	助詞	746	9,551	215.1
71	必ず(カナラズ)	和	副詞	372	3,574	212.6
72	意(イ)	漢	名詞	209	1,400	210.7
73	論(ロン)	漢	名詞	313	2,755	209.9
74	動脈(ドウミャク)	漢	名詞	52	41	208.4
75	方今(ホウコン)	漢	名詞	90	246	205.6
76	尭舜(ギョウシュン)	固	名詞	48	30	205.5
77	変ずる(ヘンズル)	混	動詞	124	523	204.4
78	権利(ケンリ)	漢	名詞	161	894	202.9
79	利(リ)	漢	名詞	139	682	199.1
80	欲する(ホッスル)	和	動詞	299	2,651	198.0
81	非ず(アラズ)	和	助動詞	113	448	196.4
82	同権(ドウケン)	漢	名詞	41	16	195.9
83	SO(エスオー)	記号	名詞	30	0	191.9
84	交易(コウエキ)	漢	名詞	51	49	191.7
85	国(クニ)	和	名詞	675	8,699	190.8
86	所以(ユエン)	和	名詞	229	1,765	190.0
87	小子(ショウシ)	漢	代名詞	31	1	189.4
88	学術(ガクジュツ)	漢	名詞	126	590	188.8
89	成す(ナス)	和	動詞	916	13,038	186.3
90	ごとし	和	助動詞	1,659	27,135	184.5
91	概する(ガイスル)	混	動詞	34	6	184.1
92	言(ゲン)	漢	名詞	180	1,196	183.4
93	諮諏(テンユ)	漢	名詞	37	13	180.4
94	交際(コウサイ)	漢	名詞	107	441	180.0
95	且つ(カツ)	和	接続詞	321	3,115	179.9
96	ニトロ[nitro]	外	名詞	31	3	178.2
97	正金(ショウキン)	漢	名詞	64	131	173.4
98	異(イ)	漢	名詞	30	3	172.0
98	リバティー[liberty]	外	名詞	30	3	172.0
100	理(コトワリ)	和	名詞	110	495	171.1

「飲み倒す」とはどういう意味なのか ーGoogle 検索を利用した日本語の低頻度複合動詞の分析ー*

徐 敏徹 (立命館大学大学院言語教育情報研究科) †

What Is the Meaning of 'Nomi-Taosu'? : Analysis of Japanese Low-Frequency Compound Verbs by Using Google Search

Mincheol Seo (Graduate School of Language Education and Information Science, Ritsumeikan University)

要旨

コーパスという用語の定義には、おおむね「大規模」という単語が登場する。しかし、そのような（大規模な）コーパスであっても、日常生活における使用頻度の低い言葉に関しては、そこから有用な情報を得ることが難しい。

本研究では、意味記述が不十分だと考えられる日本語の低頻度語彙的複合動詞を取り上げ、Google の検索エンジンとクローラーを利用し、用例を網羅的に収集した。このような方法は、従来困難であった低頻度語彙の用例分析を可能とする。

本稿では、低頻度複合動詞である「飲み倒す」を取り上げ、その特徴を記述し、前項ないし後項動詞が共通している「飲み尽くす」「飲み潰す」「踏み倒す」との比較分析を行った。分析結果、「飲み倒す」は「酒を飲んでその代金を払わないままにする」という本来の意味よりも、「たくさん飲む」という派生的な意味での使用が顕著であることが明らかになった。また、「飲み倒す」と最も類似性が高い複合動詞は「飲み尽くす」であることがわかった。

1. はじめに

大勢の人が無意識のうちに使用する言葉には、ときおり一定の規則が見られる。そのような一定の規則に基づいて、言葉の意味や用法を詳細に記述したものを我々は「辞書（もしくは辞典）」と呼ぶ。しかし、日常生活においてそれほど使用頻度が高くない言葉に関しては、辞書における意味記述が十分だとは言い難い。その一因として挙げられるのが、分析可能な用例数の不足である。言葉が実際の場面でどのように使われているのかを把握するのが困難であるため、辞書における意味記述も不十分になるのである。本稿では、使用頻度が低い言葉の意味記述をより充実させるためには、大規模な言語資料を用いることが重要であることを示す。その手始めとして、日本語の低頻度語彙的複合動詞に焦点を当てて分析を行う。

---

* 本稿は平成 29 年 12 月 13 日、立命館大学文学部に提出した卒業論文を加筆修正したものである。

† lt0715ix {at} ed.ritsumei.ac.jp

2. 先行研究

日本語の複合動詞に関する研究は、影山 (1993) 以降、主に理論言語学的な観点から活発に行われている。一方、複合動詞の実例を分析した研究はそれほど多くない。影山 (2013) からは、複合動詞分析における実例の重要性を垣間見ることができる。影山 (1993) では複合動詞を統語的複合動詞（「食べ終える」「走り続ける」等）と、語彙的複合動詞（「繰り返す」「思い出す」等）に分けている。また、語彙的複合動詞における前項動詞（以下 V1）と後項動詞（以下 V2）の間に見られる抽象的な機能関係を並列関係、右側主要部の関係、補文関係に分類している。そのうち、補文関係にあたる語彙的複合動詞（「見逃す」「聞き漏らす」等）はその数が少なく、「例外的であろうと何となく想定されていた (影山 2013:p.8)」。しかし、実際に使われた語彙的複合動詞の用例を集めると、補文関係に当てはめるしかないと思われる例が予想以上に多いことから、従来の語彙的複合動詞を主題関係複合動詞とアスペクト複合動詞に再編成したのである。とりわけ、語彙的な複合動詞の V1 が反復形を許すことを証明するために、「ウェブから収集したもの (影山 2013:p.32)」を証拠として挙げている点は注目に値する。

普段さほど見聞きしない語に関しては、その語の前後にどのような語が共起するのか、母語話者の内省による判断だけでは限界がある (滝沢 2007, 2010)。このとき、有用な道具となるのがコーパスである。コーパスとは、コンピュータで利用することができる言語研究のための大規模なデータのことで、実際に使われた用例がその言語の在り方を正しく反映するように、組織的に収集して公開したものを指す (前川 2009)。このようなデータの活用は、母語話者の直観に基づく従来の言語研究では成し得なかった隠れていた言語現象をあぶり出し、とりわけ、辞書編纂において言葉のより精緻な記述を可能としている (田野村 2010, 滝沢 2017)。しかし、コーパスであっても限界はあり、低頻度語彙に関しては検索結果がないか、あるとしてもわずかな量に過ぎない。ここで代案となるのがウェブ検索である。ウェブ検索を利用すると、既存のコーパスからは得られない用例を集めることができる。このことは、岡島 (1997) をはじめ、多くの研究者が指摘している点でもある (橋本 2007, 杉村 2007, 荻野・田野村 2011, 杉村 2014, 荻野 2014, 滝沢 2017, Taylor 2017)。しかしながら、それを活用した低頻度複合動詞の研究は、筆者の知る限り見当たらない。

そこで、本研究では、現代日本語における低頻度語彙的複合動詞の用例をウェブ検索で収集し、どのような語が複合動詞とよく共起しているのか分析することを目的とした。ウェブページを大規模なコーパスと見なし、そこから複合動詞の用例を収集・分析することで、複合動詞の意味記述を現在より充実させることが期待される。とりわけ、低頻度複合動詞は用例数の不足により、共起語分析が等閑視されて来たが、本研究を行うことによって初めて、それが実現可能になると考えられる。また、収集した用例に対し、一定の規程に従って再利用可能な形で整理を行う。このような作業は、集めた用例を今後の複合動詞研究にも活用可能にするであろう。

本研究でウェブ検索の対象とした複合動詞は、頻度調査の結果得られた低頻度複合動詞 617 語のうち 10 語である。しかし、本稿では紙面の都合上、その 10 語のうち、さらに範囲を絞り、「飲み倒す」だけを取り上げることにする。以下は、本研究における研究課題である。

1. 「飲み倒す」がヲ格を取るとき、目的語としてはどのような名詞が現れているのか。
2. 「飲み倒す」から見られる特徴は、前項ないし後項動詞が共通している他の複合動詞「飲み尽くす」「飲み潰す」「踏み倒す」からは見られないのか。

本稿の流れとしては、次の第 3 章で低頻度語彙的複合動詞の選別方法、検索エンジンやクローラーを用いた用例収集および整理方法について述べる。また、どのような過程を経て「飲み倒す」が研究対象として選ばれたのかも、そこで触れることにする。次に、第 4 章では「飲み倒す」をはじめ、「飲み倒す」と前項ないし後項動詞が共通している「飲み尽くす」「飲み潰す」「踏み倒す」の分析結果を報告する。なお、第 5 章では考察をし、最後の第 6 章で本稿の結論を述べる。

3. 研究方法

以下第 3 章では、最初に今回分析対象となった低頻度複合動詞をどのように選定したのかについて述べる。その後、Google からの用例収集方法およびデータの整理方法を簡潔に説明する。

3.1 低頻度語彙的複合動詞の選別

3.1.1 複合動詞レキシコン

国立国語研究所（以下国語研）では、『複合動詞レキシコン（国際版 ver.1.10）』というデータベースを公開し、語彙的複合動詞（以下では便宜上、単に複合動詞と略す）約 2,700 語に関する情報を提供している。ここから、低頻度複合動詞を選別するためには、『複合動詞レキシコン（以下レキシコン）』にある約 2,700 語の頻度情報が必要である。しかし、『レキシコン』上では、複合動詞の頻度情報を網羅的に確認する術がない(1)。ゆえに、何らかの基準を設けて頻度調査を行い、複合動詞の頻度表を作成する必要がある。そこで利用したのが BCCWJ の語彙表である。

3.1.2 BCCWJ の語彙表

国語研のコーパス開発センターでは、「『現代日本語書き言葉均衡コーパス（The Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ）』語彙表」を公開している。本稿では、「BCCWJ 主要コーパス語彙表（Version 1.0）」を使用し、複合動詞の頻度調査を行った。この語彙表を使用した理由は、後述する出版・図書館サブコーパスにおける固定長サンプ

(1) 『レキシコン』の各見出し語には、国立国語研究所・Lago 言語研究所が開発した BCCWJ 検索システムである NINJAL-LWP for BCCWJ（以下 NLB）へのリンクがあり、そこから見出し語の頻度情報が確認できる。しかし、そのような方法で 2,700 語以上の頻度を得るためには、その分退屈な作業を繰り返さなければならない。クローラーを用いると簡単にできる作業ではあるが、筆者はその当時、プログラミングについての知識がなかった。

ルの度数を取得するのが容易だったからである[2]。 今回の頻度調査では、出版 (書籍・雑誌・新聞) サブコーパスと、図書館 (書籍) サブコーパスのみを対象としている。BCCWJ には他にも多様なサブコーパスがあるが、その中で最も均衡的なものは出版・図書館サブコーパスの固定長サンプルである (田野村 2014:pp.121-122)。『「現代日本語書き言葉均衡コーパス」利用の手引』には、固定長サンプルが「母集団（＝推計された総文字数）からの抽出比が明確である点で、基本語彙表や漢字表の作成、語彙・文字調査など、統計的な言語研究に向く（p.30)」と明記されている。このようなバランスの取れたサブコーパスの固定長サンプルを使えば、頻度調査における偏りをなるべく抑制できると考えられる。

3.1.3 動詞の頻度表作成

頻度表を作成するために、「BCCWJ 主要コーパス語彙表」を基にし、『レキシコン』にある複合動詞約 2,700 語の頻度調査を行った。詳細な頻度表の作成過程、同音異義語や同字異音語の取り扱いなどは、次の URL から確認することができる (`http://bit.ly/2uEfC1G`)。得られた複合動詞の頻度表から、低頻度複合動詞（以下では「低頻度」という言葉を、「使用頻度が 0 に等しい複合動詞」を指すときに使用する）だけを取り出すと 617 語になる。このままでは数が多すぎるので、Excel を利用し、各々の複合動詞に乱数を付与してから昇順に並び替えた。その後、上位の 10 語まで絞り込んだ結果、以下の語が得られた。

(1) 飲み倒す、打ち仰ぐ、聞きっぱなす、困り切る、裂け広がる、使い慣れる、言い消す、咳き入る、微笑み掛ける、飲み交わす

「飲み倒す」が本稿の分析対象となった理由は、ここからわかるように、単純に最も上に位置していたからに過ぎない。

3.2 用例収集

3.2.1 用例検索および収集日

「飲み倒す」の用例検索および収集は、2017 年 9 月 14 日から 15 日にかけて行った。時間帯は検索エンジンに対する負荷を減らすために、午前零時（日本標準時）を過ぎてからにした。なお、「飲み倒す」と前項動詞が共通している「飲み潰す」、「飲み尽くす」の用例検索および収集は、それぞれ 2017 年 10 月 31 日、同年 11 月 3 日に行った。

3.2.2 検索エンジンの選択

複合動詞を検索する前に、まずは検索エンジンを選択しなければならない。代表的な検索エンジンとして、以下のようなものが挙げられる。

(2) a. Google（`https://www.google.co.jp`）

[2] 「BCCWJ 主要コーパス語彙表」ではなく、「BCCWJ 語彙表 (全体) : 短単位語彙表データ」を使用してもよい。実際に、両語彙表を使用して複合動詞の頻度調査を行った結果、高頻度複合動詞においては最大 73 回の差が出た。たとえば、「繰り返す」の使用頻度は「BCCWJ 主要コーパス語彙表」で 1,658 回、「BCCWJ 語彙表 (全体) : 短単位語彙表データ」で 1,585 回となっている。しかし、すでに述べたように本研究では低頻度複合動詞だけを扱っている。ゆえに、高頻度複合動詞におけるこのような頻度の差が本研究に影響を与えるとは考え難い。

b. Yahoo!（https://www.yahoo.co.jp）

c. goo（https://www.goo.ne.jp）

本研究で使用した検索エンジンは、比較的に用例の数が多かった (2a) の Google である。ウェブ検索による日本語研究の先駆的研究とも言える荻野 (2014) は、検索エンジンの不安定さに言及し、比較的に安定的なヒット件数が得られる goo の使用を促している。しかし、本研究で検索エンジンを利用する主な目的は、複合動詞の実例を収集することである。また、2017 年 10 月 31 日現在、goo はヒット件数を表示していないため、ヒット件数を用いて研究することができない状態である。なお、同じ語句を Google と goo で検索してみると、前者の方が検索結果の件数が多い。たとえば、「飲み倒す」を後述するフレーズ検索（3.2.5 参照）すると、得られる用例の件数は Google で 239 件、Yahoo!は 238 件ヒットするのに対し、goo では約 120 件である（検索日：2017 年 10 月 31 日）。約120 件と書いたのは、すでに述べたように goo ではヒット件数を表示しないため、検索結果を直接数え上げたからである。一方、Google と Yahoo!を検索エンジンとして比較するのは、あまり意味がない。なぜなら、2010 年の秋から、Yahoo!は Google の検索エンジンと検索連動型広告配信システムを採用しているからである(3)。実際に、上述した「飲み倒す」のフレーズ検索結果では 1 件の差しか見られていない。

3.2.3 複合動詞の表記に関する問題

複合動詞を検索する前に、もう一つ考えなければならないのは、複合動詞の書字形である。文部科学省の「送り仮名の付け方（昭和 48 年内閣告示第 2 号）」には、複合の語の送り仮名の付け方として、以下の本則を設けている。

(3) 複合の語（通則 7 を適用する語を除く。）の送り仮名は、その複合の語を書き表す漢字の、それぞれの音訓を用いた単独の語の送り仮名の付け方による。

【出典】 文部科学省『送り仮名の付け方』

たとえば、単独の語「思う」と「出す」の複合語は、それぞれの送り仮名の付け方を使用し、「思い出す」と表記する。しかし、(3) のような本則はあくまでも仮名遣いのよりどころであり、それ以外の表記も当然可能である。一例として、「繰り返す」は「繰り返す」という表記が最もよく用いられている。しかし、「繰返す」や「繰りかえす」、もしくは「くり返す」、さらにはすべてをひらがなで「くりかえす」と書く場合もある。その証拠として、次ページの表 1 に BCCWJ における「繰り返す」の書字形頻度を示す(4)。

一方、「思い付く」の場合は「思いつく」の方が一般的であろう。このような複合動詞の表記のゆれを考慮し、Google で検索する際には多様な複合動詞の書字形を用いている。最終的に、Google で検索した複合動詞およびその書字形は以下の通りである。

(4) a. 飲み倒す、飲みたおす、のみたおす

b. 打ち仰ぐ、打仰ぐ、打ちあおぐ、うち仰ぐ、うちあおぐ

c. 聞きっぱなす、ききっぱなす

(3) Yahoo! Japan - プレスリリース（https://about.yahoo.co.jp/pr/release/2010/0727a.html）

(4) 検索には NLB を利用した（検索日:2017 年 10 月 9 日）。

表 1　BCCWJ における「繰り返す」の書字形頻度

書字形	頻度	割合
繰り返す	6,445	80%
くり返す	816	10%
くりかえす	551	7%
繰返す	143	2%
繰りかえす	80	1%

d. 困り切る、困切る、困りきる、こまりきる
e. 裂け広がる、裂広がる
f. 言い消す、言消す、言いけす、いい消す、いいけす
g. 咳き入る、咳入る、咳きいる、せき入る、せきいる
h. 飲み交わす、飲交わす、飲交す、飲みかわす、のみ交わす、のみ交す、のみかわす
i. 使い慣れる、使いなれる、使慣れる、つかい慣れる、つかいなれる
j. 微笑み掛ける、微笑みかける、微笑掛ける、微笑かける、ほほえみ掛ける、ほほえみかける

なお、「のみ倒す、ききっ放す、こまり切る」は検索対象から取り除いた。その理由は、分析に使用可能な用例がほとんど、あるいは、まったくなかったからである。

3.2.4　用例収集のためのクローラー

検索エンジンで複合動詞を検索し、用例を収集する作業は、すべてクローラーが自動的に行った。クローラー (crawler) とは、「ウェブ上を常に這いまわって、新しい情報とウェブページをかき集め、中央のリポジトリに蓄えるようにする (Langville and Meyer 2009:pp.15-16)」仮想的なロボットである。クローラー作成にはプログラミング言語 python を使用した。ソースコードは、次の URL から確認できる (`http://bit.ly/2uEfC1G`)。

3.2.5　複合動詞の活用語尾に関する問題

クローラーが Google で複合動詞を検索する際には、フレーズ検索を基本としている。フレーズ検索とは、検索語句をダブルクオーテーションで囲んで検索する方法である。Google ではこのような検索方法をフレーズ検索と呼んでいる (Dornfest et al. 2007)。検索エンジンを利用して言語研究を行う場合、フレーズ検索は必須とも言える (荻野 2014)。

クローラーは、複合動詞が五段活用する場合、語尾を変換しながらフレーズ検索を行う。くわえて、複合動詞がテ形になる場合の音便「っ」「ん」「い」も付け加えて検索する。つまり、一つの複合動詞の書字形に対して計 6 回の検索を行うことになる (語尾が「す」で終わる場合は 5 回)。たとえば、「困り切る」の場合、検索語句は以下のようになる。

(5) 困り切ら、困り切り、困り切る、困り切れ、困り切ろ、困り切っ

なお、3.2.3 で述べたように「困り切る」の他の書字形として「困切る」「困りきる」「こまりき

る」もある。結果的には、語彙素「困り切る」に対して計 24 回の検索をすることになる。一方、下一段活用動詞に対しては、書字形ごとに計 7 回の検索を行った。たとえば、「使い慣れる」の場合、以下のような検索を行っている。

(6) 使い慣れ､､使い慣れな、使い慣れる、使い慣れれ、使い慣れろ、使い慣れよ、使い慣れて

「使い慣れ､」には読点（､）がついているが、検索エンジンは一部の演算子を除き、基本的には記号を無視して処理する。にもかかわらず、読点を挿入している理由は、単に語尾が変換することを示すためである。

3.2.6 データベースの整理および形態素解析

収集した用例のうち、不適切な例や重複して登場する例などは削除した。以下、どのような基準に従って用例を削除したのか、その基準の一部を示す。

(7) a. 性的な表現が含まれている場合
b. 有益な情報を導き出せない場合（やはり全花火必須だな 飲み倒さ...）
c. 動詞と動詞の間に記号が含まれている場合（…なくてもいい。消さない。）

その後､形態素解析器 MeCab（mecab of 0.996）と電子化辞書 UniDic(unidic-mecabver. 2.1.2) を使い、形態素解析を行った。詳細は補足資料を参照されたい (`http://bit.ly/2uEfC1G`)。

3.3 筑波ウェブコーパスの活用

本稿では既存のコーパスとして、筑波ウェブコーパス（Tsukuba Web Corpus: TWC）を、検索ツールとしては国立国語研究所・Lago 言語研究所が開発した「NINJAL-LWP for TWC（以下 NLT）」を利用した。本稿では、低頻度複合動詞「飲み倒す」と前項ないし後項動詞が共通している他の複合動詞も分析している。その際、複合動詞の用例が既存のコーパスに一定量以上ある場合、それを活用するようにした。すでに触れたように、本稿でウェブ検索を活用している理由は、既存のコーパスから低頻度複合動詞の用例を探しても、わずかな数しかないか皆無だったからである。しかし、「踏み倒す」の場合は、今回の頻度調査によると使用頻度が 18 回となっており、低頻度複合動詞にはならない。実際、TWC における「踏み倒す」の使用頻度は 595 回となっている。

4. 分析結果

本章では、低頻度複合動詞 617 語を昇順で並び替えた結果から最も上に位置していた「飲み倒す」を取り上げ、分析結果を報告する。また、「飲み倒す」の特徴を探るために前項ないし後項動詞が共通している複合動詞、「飲み尽くす」「飲み潰す」「踏み倒す」との比較分析も行う。

4.1 「飲み倒す」とは

まず、そもそも「飲み倒す」とはどういう意味なのかを確認しておきたい。「飲み倒す」には辞書に記載されている本来の意味と、V2 の「倒す」によって生み出される派生的な意味とが

ある。『日本国語大辞典（第二版、以下日国）』では「飲み倒す」を以下のように記述しており、これは「飲み倒す」の本来の意味にあたる。

1. 酒を飲んでその代金を払わないままにする。飲みつぶす。
2. 酒を飲んで財産をつぶす。飲みつぶす。

【出典】 日本国語大辞典

『レキシコン』でも「飲み倒す」を「酒を飲んで、代金を倒す（代金を払わない）」と記述し、その例文として (8) を示している。

(8) 彼は酒を飲み倒した.

【出典】 複合動詞レキシコン

また、以下の (9) は今回集めた用例の中で「飲み倒す」を上述した本来の意味で使用している例である（下線部筆者）。

(9) a. てめえのような奴は、おおかた年中、その手で飲み屋を飲み倒しているのだろう（吉川英治『宮本武蔵』）。
b. 人を殺し、酒屋を飲みたおし、その尻尾は童学草舎へ持って行けなどという乱暴者から、そういわれてはたまらない（吉川英治『三国志』）。
c. 酒場を飲み倒したり、女を強奪したり、人を恐喝するなどもっての外じゃ（金史良『天馬』）。

【出典】 WWW（Google ブックス）

「飲み倒す」の目的語としては「飲み屋」「酒屋」「酒場」等が現れており、これらは「酒を供する店」として分類することができる。ところで、収集した用例の中には、(9) のような「酒代を払わない」という意味ではなく、「たくさん飲む」や「十分に飲む」という意味での使用が目立っている。これは「飲み倒す」の派生的な意味にあたる。以下の (10) にその例を挙げる。

(10) a. 有楽町～銀座で朝まで飲み倒してみた
b. 健康にいいと言われるものを片っ端から試し、サプリメントを飲み倒すようになるかもしれない。
c. 何と飲み放題で、何の気兼ねもなく、飲み倒していただけます……しかし高いお酒を飲み倒されると、お店の裏で店長が静かに泣いていることをお忘れなく

【出典】 WWW

このように「倒す」が複合動詞の V2 に位置すると、「倒す」が持つ語彙的な意味が希薄になり、V1 が表す行為や動作の反復強調を示す場合がある (長谷部 2012, 影山 2013)。

4.2 用例分析結果

「飲み倒す」がどういう意味を持っているのかを確認したところで、以下では「飲み倒す」をはじめ、「飲み尽くす」「飲み潰す」「踏み倒す」の用例分析結果を報告する。

4.2.1 「飲み倒す」

分析に用いた「飲み倒す」の用例数は 924 例、使用頻度は 964 回である。用例数と使用頻度が異なっている理由は、一つの用例の中に複合動詞が複数回使われている例があるからである。まず、「～を飲み倒す」に前方共起する名詞を次の表 2 に示す。

表 2 「～を飲み倒す」に前方共起する名詞（上位 5 語、頻度降順）

名詞	助詞	複合動詞	頻度
酒	を	飲み倒す	21
ビール			19
ワイン			10
焼酎			3
夜			3

「飲み倒す」がヲ格を取るとき、目的語位置に現れる名詞の延べ語数は 114 語、異なり語数は 59 語である。表 2 を見ると、酒が 21 回で最も頻度が高く、次にビールが 19 回、ワインが 10 回、焼酎が 3 回現れており、これらの名詞だけで延べ語数全体の 46% を占めている。表に載っていない飲み物まで合計すると使用頻度は 86 回となり、延べ語数の 75% 以上を占めることになる。4.1 で触れたように、「飲み倒す」は本来の意味ではなく、ほとんどが「たくさん飲む」という派生的な意味で使われていることがわかる。その分、「飲み倒す」の目的語として「飲料」が占める割合は高い。それでは、「飲み倒す」と前項ないし後項動詞が共通している「飲み尽くす」「飲み潰す」「踏み倒す」からもこのような傾向が見られるのだろうか。

4.2.2 「飲み尽くす」

「飲み尽くす」は統語的複合動詞であるが、本稿では便宜上、単に複合動詞と呼ぶことにする。「飲み尽くす」の用例数は 1,336 例、使用頻度は 1,369 回である。日国では「飲み尽くす」を以下のように記述している。

余すところなく飲む。すっかり飲む。

【出典】 日本国語大辞典

以下の表 3 は、「～を飲み尽くす」に前方共起する頻度 10 以上の名詞である。

表 3 「～を飲み尽くす」に前方共起する名詞（頻度 10 以上、頻度降順）

名詞	助詞	複合動詞	頻度
ビール	を	飲み尽くす	101
酒			51
ワイン			48
水			13
スープ			13
全て			12

「飲み尽くす」がヲ格を取るとき、目的語位置に現れる名詞の延べ語数は 522 語、異なり語数は 196 語である。「〜を飲み尽くす」は「〜を飲み倒す」と同じく、「飲料」に属する名詞の出現が目立っている。ビール（101 回）をはじめ、酒（51 回）、ワイン（48 回）、水（13 回）、スープ（13 回）が現れており、これらの名詞だけで延べ語数の 43% を占めている。一方、「飲み尽くす」の用例からは以下の (11) のように、飲み物の「種類」に言及している例も相当数見られている。

(11) a. 年末は夜を徹して100 種類の日本酒を飲み尽くせ！
b. これまで日本全国の牛乳 150 種類を飲み尽くしたという、牛乳マニア……
c. 1000 種類以上の海外のクラフトビールはあらかた飲み尽くしてしまったので

【出典】 WWW

このような用例は他にも 53 例あり、「飲み尽くす」の全体用例に占める割合は 3.97% である。これは「飲み尽くす」の特徴的な用法だと言える。「飲み倒す」にもこの類の用例はあるが、6 例（0.65%）しかない。「飲み潰す」からは 3 例（0.79%）見られている。

4.2.3 「飲み潰す」

「飲み潰す」の用例数は 379 例、使用頻度は 389 回である。日国では「飲み潰す」を以下のように説明している。

1. 酒を飲んでむなしく日を暮らす。
2. 飲酒にふけって財産をすっかりなくす。飲みたおす。
3. 「のみたおす（飲倒）(1)」に同じ。

【出典】 日本国語大辞典

『レキシコン』ではこの複合動詞を「酒を飲んで（酒代で）財産をなくしてしまう」と記述し、次の用例を示している。

(12) 彼は財産を飲み潰した.

【出典】 複合動詞レキシコン

「飲み潰す」と前方共起する名詞には、酒（6 回）、相手（5 回）、身代（4 回）等が現れている。「飲み倒す」と同様、「飲み潰す」の目的語としても「飲料」が出現してはいるが、その割合は高くない。一方、次ページの表 4 に示したように、「飲み潰す」は助動詞「れる」との共起が目立っている(5)。

4.2.4 「踏み倒す」

「踏み倒す」の用例は、3.3 で述べたように、TWC を利用して検索した。TWC における「踏み倒す」の出現頻度は 595 回である。日国では「踏み倒す」を以下のように記述している。

1. 踏みつけて倒す。足げにして倒す。

(5) 「飲み潰される」が受身を表す用例数は 96 例（使用頻度は 97 回）で、全体用例の 25.33% を占めている。それに比べ、「飲み倒される」は 19 例で 2.06%、「飲み尽くされる」は 57 例で 4.27% である。この点については、本研究の研究課題と直接的に関係しないため、これ以上詳しくは立ち入らないことにする。

表 4　「飲み潰す」と後方共起する助動詞（上位 5 語、頻度降順）

複合動詞	助動詞	頻度
飲み潰さ	れる	97
飲み潰し	た	36
飲み潰さ	ない	9
飲み潰し	てる	4
飲み潰さ	なく	4

2. 代金や借金などを支払わないですませる。他人に損害を与える。また、他人の体面や信用などを傷つける。

【出典】 日本国語大辞典

また、『レキシコン』では「代金や借金を意図的に払わずに終わらせる」と記述し、次の例を示している。

(13) 男は、飲み代を踏み倒した.

【出典】 複合動詞レキシコン

ここで「踏み倒す」と「飲み倒す」の「倒す」に注目してみたい。もし、両者における「倒す」が同じ働き、つまり「動作の反復や強調」の意味を持っているのならば、「踏み倒す」を「たくさん踏む」「十分に踏む」「繰り返して踏む」、もしくは「踏むという動作の強調」と解釈できるはずである。しかし、TWC の用例を観察すると、そのような意味だと考えられる用例は以下の (14) くらいしか見当たらない。

(14) 数百メートル先に見えるゴール目がけ、53-11 という重いギアを全力で踏み倒す！

【出典】 TWC、下線部筆者

このように、「踏み倒す」は「飲み倒す」や「飲み潰す」から見られる派生的な意味での解釈が一般的ではないことがわかる。一方、以下の (15) は、「踏み倒す」を「酒を飲んで代金を払わないままにする」という意味で使用している例である。

(15) a. ……酒代を踏み倒したエピソードなどが悪意を持って書かれている。
b. 外で酒を飲んで遊んでも、金は払わず踏み倒していたに違いない。
c. 飲み代を集金するため旧社屋にやってきたクラブのママが、誰もいないので踏み倒されたと思い……

【出典】 TWC、下線部筆者

「踏み倒す」と「飲み倒す」は、「酒代を払わない」という意味は共通しているが、決定的な相違点がある。それは、ヲ格の補語として「踏み倒す」は「金銭」、「飲み倒す」は「酒を供する店」、もしくは「飲料」に関する語を必要とする点である。一例として、(15a) の「踏み倒す」を「飲み倒す」に置き換えると、次のようになる。

(16) ……酒代を飲み倒したエピソードなどが悪意を持って書かれている。

この例が非文だとは言い切れない。しかし、本稿での分析結果に基づくと、「酒代」は「飲み倒す」の目的語としてふさわしくない。4.2.1 では「飲み倒す」がヲ格を取るとき、目的語位置に

は酒、ビール、ワイン、焼酎等、主に飲み物が来ると述べた。むろん、飲み物ではない名詞も現れてはいるが、その中で以下の表 5 にある名詞、つまり、借金、料金、金、～費等は現れていない。反対に、「踏み倒す」がヲ格を取るときの目的語には、酒、ビール、ワイン等の飲み物は現れない。

表 5　「～を踏み倒す」に前方共起する名詞（NLT を基に作成、頻度 5 以上、頻度降順）

名詞	助詞	複合動詞	頻度
借金	を	踏み倒す	75
料金			12
金			10
費			9
債務			7
代金			6
代			6
お金			6

5. 考察

分析結果を基に、「飲み倒す」と、「飲み尽くす」「飲み潰す」「踏み倒す」の関係を下の図 1 のようにまとめることができる。

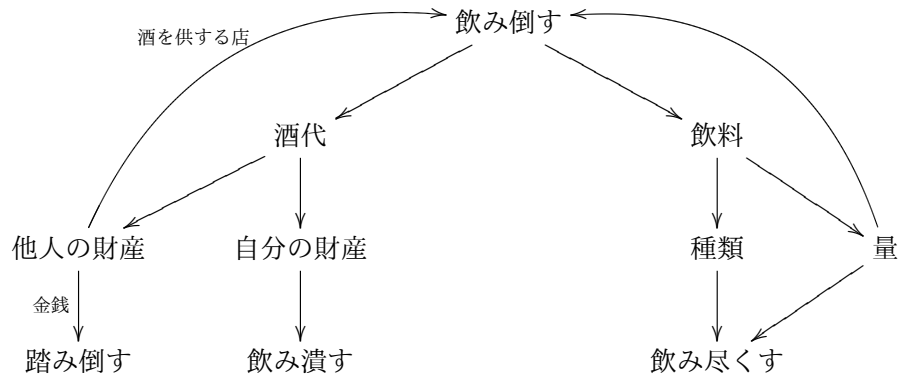


図 1　「飲み倒す」の意味分化による複合動詞の選択

複合動詞「飲み倒す」は、「酒代」に関わる（本来の）意味と、「飲料」に関わる（派生的な）意味を持っている。まず前者を見ると、それが「他人の財産」と関係があるのか、それとも「自分の財産」と関係があるのかによって枝分かれする。もし、「他人の財産」と関係があり、かつ、それの指示対象が金銭であれば「踏み倒す」が選択される。一方、「他人の財産」と関係があり、かつ、それが酒を供する店を指している場合は「飲み倒す」を使う。今度は、酒代が

「自分の財産」と関係がある場合、「飲み潰す」が使われる。次に、「飲み倒す」の意味が「飲料」に関わる場合は、「種類」と「量」に分かれる。飲み物の「種類」に関して言及するときには「飲み倒す」ではなく、「飲み尽くす」を使う。一方、「飲み物の量、（一般的に）とりわけ大量」を表す場合は、「飲み倒す」または「飲み尽くす」のどちらも使用される傾向がある。

複合動詞の数は、語彙的複合動詞に限っても 2,000 語以上ある。普段よく使う複合動詞の数は、この膨大な量のうち一部に過ぎないであろう。馴染みの薄い複合動詞の意味を記述する場合、母語話者の内省だけでは限界がある。このような問題点を克服するため、本研究では Google 検索を利用し、低頻度複合動詞の用例を収集した。

本稿では第 2 章で述べてように、研究課題を二つ設定した。それに対する答えは次の通りである。「飲み倒す」がヲ格を取るとき、目的語としては「飲料（酒、ビール、ワイン等）」が最もよく現れている（研究課題 1）。「飲み倒す」の場合、「酒代を払わない」という（語彙的複合動詞の）本来の意味ではなく、「たくさん飲む」という派生的な意味でよく使われている。このことは、「飲み倒す」がヲ格を取るとき、目的語として飲料が 75% 以上を占めていることからも明らかである。また、このような「飲み倒す」の特徴は、「飲み尽くす」から最も顕著に見られている。一方、「飲み潰す」からはこのような傾向があまり見られず、「踏み倒す」からは見られない（研究課題 2）。したがって、「飲み倒す」を「たくさん飲む」という意味で使うのであれば、それを「飲み尽くす」に置き換えても問題はない。

複合動詞の実証的研究は、主に V2 と結合する V1 に焦点を当てる傾向がある。日本語における複合語の主要部は右側であることを考えると、このような傾向は当然のことかもしれない。しかし、コーパスやウェブなどの大規模な言語資料を使えば、複合動詞の内部結合にとどまらず、そこから少し離れて外部を観察することもできる。なかんずく、ウェブ検索は低頻度語彙に関しても、そのような言語研究を可能とする。それゆえ、従来のままでは不十分であった低頻度複合動詞の意味記述をも充実させることができるのである。

本研究では、頻度調査において一部の同音異義語を区別した。場合によっては、考察対象から外した複合動詞もある。しかし、低頻度複合動詞を探ることに主眼を置いた以上、このような作業はもっと慎重に行わなければならない。たとえば、「取り上げる」と同音異義語である「撮り上げる」の BCCWJ における使用頻度は 0 である。つまり、本研究の指針によると、この語は低頻度複合動詞に当てはまるのである。この矛盾に気づいたのは、卒業論文提出を一か月控えていた 2017 年 11 月 13 日だった（頻度調査を行ったのは 2017 年 2 月ごろである）。今後、低頻度複合動詞の研究を続ける際には、削除した複合動詞を綿密に観察し、複合動詞の頻度表を修正する必要がある。

用例を集める際には、重複用例の収集を防ぐためにクローラーが URL を参照している。しかし、それだけでは限界があり、最終的にはすべての用例を目視で確認しなければならない。しかしながら、研究者個人が膨大な数の用例を一々覚えながら、重複する用例を取り除くことは現実的に不可能である。ゆえに、本稿で利用した用例にもいくつかの重複用例が紛れ込んでいる可能性がある。

6. おわりに

辞書で「飲み倒す」を引くと「酒を飲んでその代金を払わないままにする」と書いてある。しかし、Google で「飲み倒す」を検索し、用例を分析した結果、ほとんどの人が辞書には書いていない「たくさん飲む」という意味でこの言葉を使用していることがわかった。「飲み倒す」と前項ないし後項動詞が共通している複合動詞であっても、互いには微妙な意味上の違いがある。このような低頻度語彙の意味を把握するためには、用例の分析が重要になってくる。日常生活の中での使用が限られているからこそ、ひたすら辞書を眺めたり、考え込むだけでは、その単語がどのような意味で使われているのか気づきにくい。言葉は変化する生き物だと言われる。そのような言葉が、我々の生活の中でどのような振る舞いをしているのか、まんべんなく目を配り、その特性をなるべく詳細に記述することは、言語の本質を究明するためにも欠かせない作業であると考えられる。

第 2 章で述べたように、本来は「飲み倒す」だけではなく、他の低頻度複合動詞の分析も行う予定であった。今後は、選別した他の複合動詞に関しても分析を行い、低頻度だからこそ普段認識できなかった隠れていた言語現象を明らかにしたい。

文　献

Amy N. Langville, and C. D. Meyer (2009)『Google PageRank の数理 : 最強検索エンジンのランキング手法を求めて』岩野和生・黒川利明・黒川洋訳, 共立出版.

岡島昭浩 (1997)「インタ-ネットで調べる (特集 ことばを調べる)」日本語学, 16:12, pp. 52–59.

荻野綱男 (2014)『ウェブ検索による日本語研究』朝倉書店.

荻野綱男・田野村忠温編 (2011)『コーパスとしてのウェブ』明治書院.

影山太郎 (1993)『文法と語形成』ひつじ書房.

影山太郎編 (2013)『複合動詞研究の最先端 : 謎の解明に向けて』ひつじ書房.

杉村泰 (2007)「コーパスを利用した複合動詞の類義分析–インターネット検索エンジンの利用」言葉と文化, (8), pp. 289–304.
(http://hdl.handle.net/2237/8349 よりダウンロード可能)

杉村泰 (2014)「コーパスを利用した複合動詞「V1-抜く」と「V1-抜ける」の意味分析」言語文化論集, 35:2, pp. 55–68.
(http://hdl.handle.net/2237/19706 よりダウンロード可能)

John R. Taylor (2017)『メンタル・コーパス : 母語話者の頭の中には何があるのか』西村義樹ほか編訳, くろしお出版.

滝沢直宏 (2007)「巨大データの必要性–言語の周辺的・慣習的側面を探るために (特集インターネットと言語研究–情報を選り分け、活用するために)」月刊言語, 36:7, pp. 34–41.

滝沢直宏 (2010)「シンポジウム周辺部を記述するための大規模コーパスの利用–その方法と留意点」英語語法文法研究, (17), pp. 23–37.

滝沢直宏 (2017)『ことばの実際 2 コーパスと英文法』内田聖二・八木克正・安井泉編, 研究社.

田野村忠温 (2010)「日本語コーパスとコロケーション–辞書記述への応用の可能性 (特集 コーパスを活用した言語研究 (1))」言語研究, (138), pp. 1–23.
(http://www.ls-japan.org/modules/documents/LSJpapers/journals/138_tanomura.pdf よりダウンロード可能)

田野村忠温 (2014)「BCCWJ の資料的特性—コーパス理解の重要性」石井正彦ほか『コーパスと日本語学』田野村忠温編, pp. 119-151, 朝倉書店.

橋本和佳 (2007)「名詞とそれを修飾する形容詞の関係 (特集コロケーション)」日本語学, 26:12, pp. 38–46.

長谷部郁子 (2012)「語彙的アスペクト動詞としての『〜倒す』について」「日本語レキシコン」共同研究発表会（複合動詞特集、2012 年 9 月 24 日、於：東北大学).
(http://pj.ninjal.ac.jp/lexicon/長谷部（2012-09-24）.pdf よりダウンロード可能)

前川喜久雄 (2009)「導入 コーパスとは何か (特集 日本語研究とコーパス)」国文学：解釈と鑑賞, 74:1, pp. 6–14.

Rael Dornfest, Paul Bausch, and Tara Calishain (2007)『Google Hacks：プロが使うテクニック&ツール 100 選（第 3 版)』石川隼輔ほか訳, オライリー・ジャパン, オーム社 (発売).

関連 URL

国立国語研究所コーパス開発センター『現代日本語書き言葉均衡コーパス語彙表』http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

国立国語研究所コーパス開発センター『「現代日本語書き言葉均衡コーパス」利用の手引 (第 1.1 版)』http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual_03.pdf

国立国語研究所『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『複合動詞レキシコン』http://vvlexicon.ninjal.ac.jp

国立国語研究所・Lago 言語研究所『NINJAL-LWP for BCCWJ』http://nlb.ninjal.ac.jp

小学館『日本国語大辞典（第二版)』https://japanknowledge.com/contents/nikkoku/

筑波大学・国立国語研究所・Lago 言語研究所『NINJAL-LWP for TWC』http://nlt.tsukuba.lagoinst.info

文部科学省『送り仮名の付け方』http://www.mext.go.jp/b_menu/hakusho/nc/k19730618001/k19730618001.html

先天性全盲ろう児の音声言語訓練長期記録の分析状況及び保存活動

菊池英明†[1]、市川　熹[2,1,3]、岡本　明[4]、長嶋祐二[3]、藤本浩志[1]、引田秋生[5]
([1]早大、[2]千葉大、[3]工学院大、[4]筑波技大、[5]元山梨県立盲学校)

Long-Term Record of Spoken Language Training for Congenital Deaf-blind Children and its Preservation Activities

H. KIKUCHI (WASEDA Univ.), A. ICHIKAWA (CHIBA Univ.), A. Okamoto (Tsukuba Univ. of Tech.),
Y. NAGASHIMA (KOGAKUIN Univ.), H. FUJIMOTO (WASEDA Univ.),
and A. HIKITA (Yamanashi Prefectural School for the Visually Impaired)

要旨

山梨県立盲学校での先天性全盲ろう児に対する音声言語獲得訓練と生活指導に関する数万点に及ぶ、1950（昭和 25）年からの長期時系列的多角的記録と教材資料などが残されている[1]~[4]。梅津八三東大教授が指導し、一貫して進めてきた盲人の認知行動・心理の研究の知見をベースに，先天盲ろう児への教育という未知の課題に対して取り組んだ科学的研究の実践過程記録である。言語獲得が極めて困難な先天性盲ろう児に対する数万件の実践記録群は、おそらく世界で唯一の極めて貴重な資料であり、盲ろう児当事者から表出された点字や録音資料からは、学習の進行程度を直接見ることが期待される。言語獲得プロセスの解明や盲ろう児教育に重要な示唆が得られるであろう。しかし最も質の悪い時代の紙や録音テープ等に記録され劣化が著しいため、現在電子化保存と「データベース開発」(DB 化)を進めている。DB 化後は山梨県立盲学校に移管、公開する計画である。訓練記録、訓練経緯、同校での分析状況および発音訓練用の木製口模型などの教材、現状の同保存活動等を紹介する。

１．先天性全盲ろう児教育海外事例と本事例

先天全盲ろう児の教育の試みは世界的には 19 世紀後半ころから行なわれ始めたようである。米国のヘレン・ケラー（1880～1968)，ローラ・ブリッジマン（1829～1889)、ロバート・スミスダス(1925～2014)、ソ連（現ロシア）のオリガ・イワノヴァ・スコロホドワ（1911～1982）などの例が報告されているが[5]~[8]、何れも逸話的あるいは理念的で、具体的データが示されておらず、その後の検証が困難である。

一方本報告で取り上げる資料は、長期にわる複数名の公私の日々の具体的状況に関する多角的側面の記録と教材であり、方法論の検証も期待できる貴重な事例である。

２．山梨盲における先天盲ろう教育[1]

２.１　教育の開始

1948（昭和 23）年、山梨県立盲唖学校（当時）の堀江貞尚校長は、県下の未就学の盲児、ろう児の実態調査を行なった結果、その中に盲ろう二重障害児がいた。堀江はその盲ろう T 男（5 歳）の教育を決意した。

その後、校長は三上鷹麿へと変わったが、堀江の考えを理解し盲ろう児への教育に熱意を傾けた。やがて横浜から S 子（7 歳）が加わった。日常の基本的習慣づけから始められ、教員・寮母たちの献身的な努力により、歯磨きや手洗い、食事，着替えなども自分でできるよ

† kikuchi@waseda.jp

うになった。

次に言語獲得の訓練が行われている。点字と物との結び付けの訓練が、彼らが好きな飴や菓子と「あめ」「かし」と点字で打ったカードを工夫して、忍耐強く続けられたが、のどや唇を触らせて言葉を読み取らせる「触話法」もほとんど進まず、教員たちの焦りと諦めは強まっていった。堀江は「盲児への教育経験から、盲ろうであっても比較的簡単に言語シンボルを理解させることができると考えていたが、楽天的空想に過ぎなかった」と述懐している。

２.２　梅津八三らによる教育研究実践

1951（昭和 26）年、山梨盲に訪ねた梅津八三東大教授が盲ろう児二人に出会い、その半年後には教諭たちの試行錯誤の結果、約 30 種類の身振りサインでの交信が周囲の人との間にでき、日常の行動をコントロールできるようになっていた。

三上と梅津は、複雑多様な内容に対応できる「言語行動」を獲得することを目標とし、1952（昭和 27）年には「盲聾教育研究会」を発足させている。研究会メンバーは頻繁に山梨盲を訪れ、夏休みには合宿を、また盲ろう児を自宅に泊まらせてともに生活、科学的知見に基づく教育方法が模索・実践された。この教育実践の状況を克明に記録し，授業ノートや指導記録、盲ろう児の日記（点字など）、往復書簡（同）などのデータ類すべてを残している。

梅津はいきなり点字の訓練をするのではなく、物の形に対する概念形成の訓練から始めている。板に三角形、正方形、円形の穴が開いていて、それにはまる形の板をはめ込む「形態板」作業や、点の位置の識別の訓練を、次に実物と点字の対応付け訓練を行い、物の名前の点字カードと、名前の点字を貼った物を比べて選ばせた。物と点字の間に対応関係が成立するようになり、単語は徐々に増え、動詞なども学習、文章を組上げて高単位の交信ができるようになった（約２年）。

つづいて音声による発信の訓練を行っている。まず口の形をつくることを教え、次に意図的に息を出すことや、声帯を緊張変化させることへと順に進み、この 3 つを統合して訓練し、徐々に言葉が出せるようになった（約１年）。

ローマ字指文字による言葉の受発信も訓練され、さらに算数、社会などの教科の学習も行なわれた。これらは点字学習、発声の矯正などもあわせてＴ男、Ｓ子や、その後入学した 2 名の盲ろう児が転校や卒業するまで続けられた。

３．教育実践の記録・資料

３．１　山梨盲での保管・整理

彼らの卒業後，1971（昭和 46）年に盲ろう学級は閉鎖され、当時担当だった志村太喜弥教諭は国立特殊教育総合研究所（現国立特別支援教育総合研究所、以下特総研）に転出、盲ろう児教育は特総研付属の国立久里浜養護学校に引き継がれ、資料も特総研に移管、梅津とともに整理が行われた。梅津はこの資料・データ類に一つずつ克明にカードを作成し、番号を付けて整理した。資料はその後、中澤恵江研究員（現横浜訓盲学院学院長、全国盲ろう教育研究会会長）らによってさらに整理され、保管されていた。

その後 2007（平成 19）年に山梨盲から数名の教員が特総研を訪れて、集中的な整理作業が行われ、山梨盲で開かれた『盲ろう唖』教材・資料展などを経て、2011（平成 23）年、全ての資料は山梨盲へ戻された。

表１　盲ろう児教育実践資料群

1-1 盲ろう児からの点字資料
1-2 盲ろう児の日記類
1-3 指導記録類
1-4 成績、学級日誌、寄宿舎日誌
1-5 盲ろう児との往復書簡
2-1 概念形成学習教材（立体模型）
2-2 記号操作学習教材
2-3 教科学習教材
2-4 発声・口形教材
3-1 指導研究報告書
3-2 指導系統図
3-3-1 学校生活・日常生活映像

山梨盲では「盲ろう教育研究委員会」を設立、膨大な資料を保管し、また梅津が作成した教材の整理カードと教材実物との照合、資料の電子データ化などに取り組み、年報発行や資料展開催などの広報活動にも努めてきている。

３．２　保管されている教材・資料

保存されている資料群[1]~[4]を表１に、資料例や教材例を写真１～６に示す。

写真１～３は資料の現状の一部で、写真３からは資料の劣化状況が判る。写真４～６は音声訓練計画や教材、当事者からの録音音声テープなどの例である。録音テープも劣化が懸念されるため再生をひかえている。

写真１　保管されている資料の例

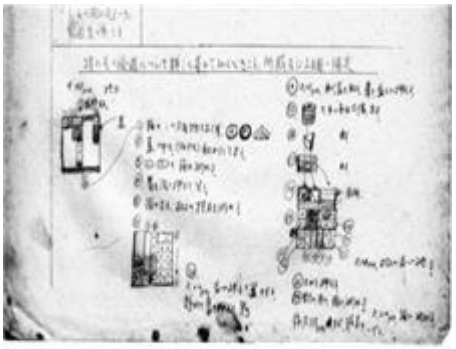


写真２　寮母による寄宿舎の日誌の一部

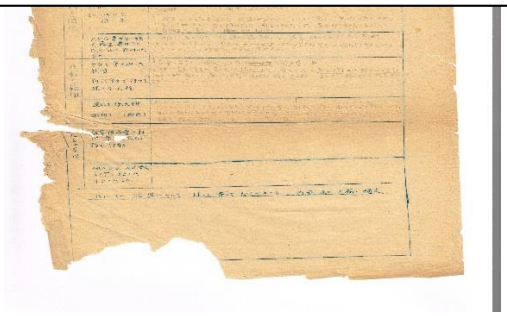


写真３　破損が危ぶまれる手書き記録の

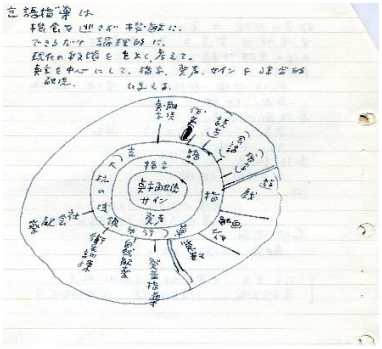


写真４　発声訓練の計画検討例

写真５　発声口形モデル

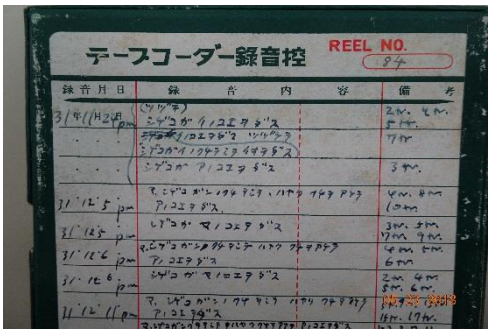


写真６　発音訓練音声記録例
「い」の声が出たという記録など

4. 資料の分析

4.1 梅津の論文

梅津の論文には，一貫して進めてきた盲人の認知行動・心理の研究の知見をベースに，先天盲ろう児への教育という未知の課題に対して取り組んだ科学的研究の実践過程が述べられている。これらには盲ろう児への働きかけとその結果については詳細に述べられている[9][10]。

4.2 盲ろう教育研究委員会による分析

盲ろう教育研究委員会により一部の資料については分析が進められ（写真７)、いくつかの報告書にまとめられている。

これらには，日本で初めての先天盲ろう児教育に関わった方々の献身的な取り組みの様子や，教育実践・研究の経過などがまとめられていて，歴史的にもまた研究資料としても貴重な文献である[11]~[23]。

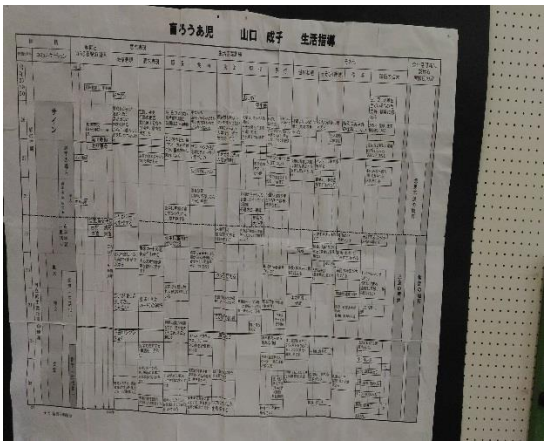


写真７　訓練の流れの分析例

しかし，物の概念，物の名前の理解が日常の個々の指導場面において，いつ，どう発現してきたかなど具体的なポイントについては、手作業による多量な資料を相互に関係付ける分析には限界があり、今後に残された課題である。各資料を電子化し関連付ける DB システムの開発が不可欠である。

5. 資料電子化と保管

5.1 電子化と DB 化の試み

山梨県立盲学校の「盲ろう教育研究委員会」ではこれまで前述のように全体の資料リストを作成、いくつかの資料の電子化を進めてきた。

しかし電子化は手を付けた段階にあり、以下のように電子化と DB 化の支援を進めている。本年度から本格的に電子化を開始しているが、資金調達と作業量の制約もあり、以下のように多くの方々の協力を得ているが、完成までには今後数年を要するものと思われる。

なお資料は旧仮名時代のものであるため、資料は点字も墨字も旧仮名遣いか誤字が混在している。しかし読み取りデータはそれ自体が時代と当事者の状況を反映したものであり、生データとしてそのまま記載する。表記(旧仮名)と音声訓練などの音声の違いも、学習への影響などの分析に有効な情報を提供するであろう。

5.2 電子化プロジェクトとその活動

上記の経験に基づき、新たに体制を作ることを検討、いくつかの大学と研究機関の研究者による「先天盲ろう児教育資料電子化プロジェクト」（以下、プロジェクト）が発足した。

プロジェクトの支援の基本方針

・あくまでも山梨盲の意向に沿った支援、協力であること（プロジェクトのステアリングメンバーは山梨盲の盲ろう教育研究委員会から「外部調査研究員」の委嘱を受け、連携を密にして進めている)。
・電子化資料の著作権・所有権を山梨盲に集約させること。
・資料番号などは特総研と山梨盲で付与したものを変更しない。
・データ構造はこれまでの山梨盲での分析手続きと研究結果に整合性があること。

・活動のベースとして、科研費、民間の助成金などの獲得に努める。
・資料の劣化状況にかんがみ、文化財などの取り扱いの経験があり、技術力の高い機関に電子化を依頼すること。
・DB はリレーショナルデータベース(RDB)とすること。
・DB は、多くの研究者が利用できるように公開を計画すること。
・プロジェクトには言語獲得プロセスの探求および視聴覚障害児の言語獲得教育手法の開発を担当するグループを置き、その検討結果を上記 DB システムの開発にも寄与すること。

5.3 山梨盲ろう教育資料電子化事業実行委員会

しかし全資料の電子化には専門技術が必要であり、プロジェクトの資金は現段階では大幅に不足する見通しのため、別途盲学校関係者を中心に「山梨盲ろう教育資料電子化事業実行委員会」を設立、プロジェクトと並行して募金活動を行い、分担して電子化を進める体制を作っている。

謝　辞

本資料の作成と保存に今日まで継続的に取り組まれてきた中澤恵江先生はじめ特総研での関係者、白倉明美先生（現山梨県立ふじざくら支援学校教頭）はじめ山梨県立盲学校および同校盲ろう教育研究委員会などの関係者の皆様、電子化保存とDB化を進めているプロジェクトおよび山梨盲ろう教育資料電子化実行委員会のメンバーの方々に感謝いたします。

本研究の一部はJSPS科研費 JP18HP7002の助成を受けたものです。

文　献

[1]岡本 明,“先天盲ろう児教育の夜明けー山梨県立盲学校における実践記録ー”, ノーマライゼーション, pp.36-38, Aug. 2012
[2]文部省初等中等教育局特殊教育課, 山梨県立盲学校における盲聾教育に関する研究ー文部省指定実験学校報告書ー, 1970
[3]中澤恵江編, 心理学梅津八三の仕事, 第 1 巻～第 3 巻, 春風社, 2000
[4]『山梨県立盲学校盲ろう教育研究委員会年報』第 3 号, 山梨県立盲学校, 2014
[5]William Wade, The Blind-Deaf: A Monograph, HECKER BROTHERS, 1901
[6]メシチェリャコフ, 盲聾唖児教育ー三重苦に光をー, 坂本市郎訳, ナウカ, 1984
[7]広瀬信雄著, 盲ろうあ児教育のパイオニア・サカリャンスキーの記録, 文芸社, 2014
[8]広瀬信雄編著, もう一人の奇跡の人～「オリガ・I・スコロホードワ」の生涯, 新読書社, 2012
[9]梅津八三,“盲ろう児の言語行動の形成”, 精神薄弱児研究, 昭和 52 年 10 月号, 1977
[10]梅津八三, 重複障害児との相互輔生行動体制と信号系活動, 東京大学出版会, 1997
[11]堀江貞尚,“ろう盲(二重障害)児”, 東北大学教育学部研究年報, 昭和 28 年, 1953
[12]藤口透吾, 0 学級の子供たち, 誠心書房, 1956
[13]山梨県立盲学校, 昭和 36 年度文部省指定実験重複障害児研究報告書―盲聾唖教育の研究―, 1962
[14]山梨県立盲学校編, 盲ろうあ教育, 山梨県立盲学校, 1965
[15]志村太喜弥, 重度・重複障害児の教育盲ろう児の指導実践に学ぶ, コレール社, 1989
[16]冨田和子, ある盲聾児の初期指導, 山梨県立盲学校, 1997
[17]ヴァスクレセーニエ編集部編, 広瀬信雄訳, みえる・きこえる 指先の世界, 新読書社, 1977
[18]1997 山梨県立盲学校, 創立 80 周年(盲ろう教育開始 50 周年)記念誌, 1999
[19]山梨県立盲学校, 創立 90 周年記念誌, 2008
[20]山梨県立盲学校, 山梨県立盲学校創立 90 周年記念事業盲学校講話シリーズ講話集, 2008
[21]山梨県立盲学校, 山梨県立盲学校盲ろう教育研究委員会年報第 1 号, 2011
[22]山梨県立盲学校, 山梨県立盲学校盲ろう教育研究委員会年報第 2 号, 2012
[23]白倉明美, 岡本 明, 盲ろう教育・福祉 ～山梨県立盲学校での先天盲ろう児教育～, 日本盲教育史研究会第 3 回研究会, 2014

『BCCWJ 図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性

馬場 俊臣（北海道教育大学札幌校）

The Possibility of Studies of Stylistic Features of Words using "Writing Style Annotation for the Library Subcorpus of the BCCWJ"

Toshiomi Baba (Hokkaido University of Education, Sapporo Campus)

要旨

本稿では、『BCCWJ 図書館サブコーパスの文体情報』のアノテーションデータを利用し、『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」内の語(語彙素)を対象として各語の文体差を数値化する試みを行い、この試みが文体研究において有効性・可能性があることを示す。まず、『文体情報』の専門度、硬度などの文体に関する 5 種類の指標別に、各語の平均値を算出する方法を示す。次に、各語の平均値については専門度、客観度、硬度、くだけ度の 4 指標は相互に強い相関があること、品詞別では感動詞と接続詞の 4 指標の平均値とそのばらつきは他の品詞に比べて特異であり品詞の特徴が表れていること、語種別では和語、漢語、外来語の特徴の違いが平均値の違いに表れていることなどの全体的傾向を示す。さらに、各語の平均値が、語の文体差に関する内省判断と強い相関があることを先行研究の調査結果と比較して示す。

１．はじめに

本稿では、『BCCWJ 図書館サブコーパスの文体情報』[1](以下、『文体情報』)のアノテーションデータを利用して、語の文体差を数値化する試みを行い、文体研究[2]における有効性・可能性を検討する。

『文体情報』は、後述のように『現代日本語書き言葉均衡コーパス』[3](以下、BCCWJ)内の「図書館サブコーパス」の全サンプルに対する文体情報のアノテーションデータである。このアノテーションデータのうちの「専門度」「客観度」「硬度」「くだけ度」「語りかけ性度」(以下、「5 指標」)の値を利用して、「図書館サブコーパス」に出現する全語彙素[4](以下、「語」)の 5 指標のそれぞれの平均値を算出し、この平均値が各語の文体差を表すものと仮定して、その有効性・可能性を検討する。

語の文体差に関しては、「文体的特徴からする単語の分類は、連続的であり、程度の差によるものである。」(宮島 1977：873)と指摘されているように「連続的」である。ただし、実用的には「離散的」に段階差を設定して扱われることが多い[5]。本稿では、語の文体差を「連続的」な姿のまま捉えようとする試みでもある。

---

[1] 国立国語研究所(2015)。

[2] 本稿では、「文体」を、硬さ・柔らかさの違い、書き言葉・話し言葉の違いなどの「類型的な文体」に限定する。

[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/ 参照。

[4] 後述のように、本稿では「長単位」の語彙素のみを対象とする。

[5] 井上(2009)、柏野(2016)など参照。

以下、2 節で、『文体情報』の概要を紹介するとともに、5 指標の平均値の算出方法を示す。その上で、本稿で検討対象とする語の範囲を示す。3 節で、5 種類の文体指標の平均値の全体的傾向・特徴を、5 指標の相互の相関、品詞別の特徴、語種別の特徴の観点から検討する。4 節で、各語の文体指標の平均値がどの程度、語の文体差に関する内省判断と一致するかを先行研究の調査結果を利用しながら検討する。5 節で、全体のまとめと今後の展望を述べる。

２．各語の 5 指標の平均値

２.１　『BCCWJ 図書館サブコーパスの文体情報』について

まず、『BCCWJ 図書館サブコーパスの文体情報』の概要を紹介する[6]。

『文体情報』は、BCCWJ に収録されている「図書館サブコーパス」の書籍サンプル(10,551 サンプル)を対象として、「内容・表現の文体的特徴を表す分類指標」及び「形式・内容・表現に文体判断が単純にいかない特徴をもつものの分類指標」を付与したデータである。「内容・表現の文体的特徴を表す分類指標」には、「対象読者に想定される読解レベル(難易度)」に関わる「専門度」、「テキストの作成意図」に関わる「客観度」、「さまざまな文体情報」に関わる「硬度」「くだけ度」「語りかけ性度」の 5 種類の文体的特徴を表す分類指標(5 指標)が設けられている。なお、「さまざまな文体情報」のうち、「硬度」「くだけ度」は「形式性、親疎性を問う」指標であり、「語りかけ性度」は「口語性を問う」指標である。これらの 5 指標は、それぞれ「言語データ構築経験有のおおよそ 20～50 代の女性、延べ 9 名。」の作業者によって付与され、それぞれ次の 3 段階～5 段階のいずれかの段階が付与されている。

(a)専門度　1 専門家向き、2 やや専門的な一般向き、3 一般向き、4 中高生向き、5 小学生・幼児向き

(b)客観度　1 とても客観的、2 どちらかといえば客観的、3 どちらかといえば主観的、4 とても主観的

(c)硬度　1 とても硬い、2 どちらかといえば硬い、3 どちらかといえば軟らかい、4 とても軟らかい

(d)くだけ度　1 とてもくだけている、2 どちらかといえばくだけている、3 くだけていない

(e)語りかけ性度　1 とても語りかけ性がある、2 どちらかといえば語りかけ性がある、3 特に語りかけ性はない

5 指標の平均値の算出に当たっては、この各指標の段階の番号を数値として扱った。

なお、5 指標の付与対象となったサンプル数は、「図書館サブコーパス」全 10,551 サンプルのうちの、8,887 サンプル[7]である。

---

[6] 『文体情報』に関する説明は、国立国語研究所(2015)の添付文書「概要」及び柏野(2013)に基づく。

[7] 柏野(2013)には 8,887 サンプルと記載されているが、『文体情報』内のデータでは、専門度、硬度、くだけ度、語りかけ性度が付与されているのはそれぞれ 8,821 サンプル、客観度が付与されているのは 5,901 サンプルである。

２.２　平均値の算出方法

『文体情報』の対象である「図書館サブコーパス」内の長単位[8]の全語彙素(語)を対象として、それぞれの語が用いられている全サンプルの専門度、客観度、硬度、くだけ度、語りかけ性度別に平均値を求め、その値を、その語の「専門度平均値」「客観度平均値」「硬度平均値」「くだけ度平均値」「語りかけ性度平均値」とする。

ただし、ある語が同一のサンプルに 2 回以上用いられている場合、そのまま平均値を求めるとそのサンプルの指標の値の影響が相対的に強く出てしまうため、同一のサンプルに 2 回以上用いられている場合は 1 回用いられているものとして平均値を求めた[9]。

実際の算出作業は、データベース管理システム MySql を用い、次の手順で行った。

① 下準備として、『文体情報』アノテーションデータ[10]の「専門度」「客観度」「硬度」「くだけ度」「語りかけ性度」の各列を指標の段階の数値だけに置き替えた[11]アノテーションデータを作成した。

② BCCWJ の DVD 版公開データ(BCCWJ-DVD 版 Version 1.1)の長単位データ[12]及び①で作成した修正版『文体情報』アノテーションデータのそれぞれをインポートしたテーブルを作成した。

③ BCCWJ 長単位データの各行(レコード)(用いられているすべての語彙素)を、サンプル ID[13]の一致する『文体情報』アノテーションデータと結合したテーブル(「延べサンプル方式テーブル」)を作成した。行(レコード)数は 30,273,796 行[14]である。

④ 異なり語数を確認するために、この「延べサンプル方式テーブル」に「語彙素」「語彙素読み」「品詞」「語種」[15]を結合した列(「品詞語彙素等」)を挿入したテーブル(「延べ語テーブル」)を作成し、この「延べ語テーブル」から「品詞語彙素等」が同一の

---

8 「長単位」は「複合語を把握する」ことができ「サンプルの言語的特徴の解明に適した」単位である(国立国語研究所コーパス開発センター2015)とされている。

9 接続詞のみを対象として「硬度」「くだけ度」の平均値を扱った馬場(2018)では、ある語が同一のサンプルに 2 回以上用いられている場合そのまま平均値を求める方式を「延べサンプル方式」、2 回以上用いられている場合でも 1 回用いられているものとして平均値を求める方式を「異なりサンプル方式」と呼んでいる。「異なりサンプル方式」である本稿の算出方法を具体例で示しておく。仮に「御昼」という語が、サンプル A(専門度 1、硬度 1)(使用頻度 3 回)、サンプル B(専門度 3、硬度 3)(使用頻度 1 回)、サンプル C(専門度 3、硬度 2)(使用頻度 2 回)の 3 サンプルで用いられていれば、専門度平均値は「(1+3+3)÷3＝2.3333」、硬度平均値は「(1+3+2)÷3＝2.0000」となる。

10 LB_all.csv。

11 元データは、例えば「3 一般向き」のように指標の段階及び選択肢表現が入っている。これを「3」のように数値のみに置き換えた。この作業は Excel を用いて行った。

12 Disk2(NumTrans 版)の「TSV_LUW_NT」(長単位データ)の LB.zip(「図書館サブコーパス」)のデータを用いた。

13 『文体情報』アノテーションデータの列名は「SampleID」である。

14 「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」及び「BCCWJ 品詞構成表(Version 1.1)」では、長単位の延べ語数は 25,031,768 語である。「品詞」フィールドに「URL、カタカナ文、方言、未知語、漢文、空白、英単語、補助記号、言いよどみ、記号」のタグが付けられている(本来の「品詞」以外の)計 5,242,027 語は、この語彙表の長単位の延べ語数からは除かれている。「BCCWJ 品詞構成表(Version 1.1)」と比べると本稿のデータは「名詞」が 3 語多い。なお、語数に 2 語の差がある理由については不明である。

15 「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」及び「BCCWJ 品詞構成表(Version 1.1)」では、長単位は「語彙素」「語彙素読み」「品詞」「語種」の 4 つの組を用いて同一の見出し語を特定している。これに倣った。

行(レコード)の重複を削除したテーブルを作成した。行(レコード)数は 821,510 語[16]である。

⑤　④の「延べ語テーブル」から、「サンプル ID」及び「品詞語彙素等」が同一の行(レコード)の重複を削除したテーブル(「異なりサンプル方式テーブル」)を作成した。行(レコード)数は 7,600,397 行である。

⑥　この「異なりサンプル方式テーブル」に基づき、(本稿での)すべての異なり語 821,510 語それぞれの専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値及びそれぞれの標準偏差の一覧表を作成した。

表 1 に、語別の 5 指標ごとの平均値のリストの一部を示す。

表 1　語別の 5 指標ごとの平均値のリスト(一部)

語彙素	品詞	専門度頻度	専門度平均値	専門度標準偏差	客観度頻度	客観度平均値	客観度標準偏差	硬度頻度	硬度平均値	硬度標準偏差	くだけ度頻度	くだけ度平均値	くだけ度標準偏差	語りかけ性度頻度	語りかけ性度平均値	語りかけ性度標準偏差
だ	助動詞	8821	2.9747	0.5909	5901	2.3965	0.9239	8821	2.5912	0.7349	8821	2.5871	0.5913	8821	2.6548	0.6442
は	助詞-係助	8821	2.9747	0.5909	5901	2.3965	0.9239	8821	2.5912	0.7349	8821	2.5871	0.5913	8821	2.6548	0.6442
が	助詞-格助	8821	2.9747	0.5909	5901	2.3965	0.9239	8821	2.5912	0.7349	8821	2.5871	0.5913	8821	2.6548	0.6442
に	助詞-格助	8821	2.9747	0.5909	5901	2.3965	0.9239	8821	2.5912	0.7349	8821	2.5871	0.5913	8821	2.6548	0.6442
の	助詞-格助	8821	2.9747	0.5909	5901	2.3965	0.9239	8821	2.5912	0.7349	8821	2.5871	0.5913	8821	2.6548	0.6442

さて、この「語別の 5 指標ごとの平均値のリスト」の異なり語(行数)は 821,510 語であるが、「品詞」フィールドに「URL、カタカナ文、方言、未知語、漢文、空白、英単語、補助記号、言いよどみ、記号」のタグが付けられている語が 492 語含まれている。これらは本来の「語」ではないため、これを除いた 821,018 語が「図書館サブコーパス」の異なり語数となる。さらに、この 821,018 語のうち 101,209 語は 5 指標の付与対象サンプルに出現していない。この 101,209 語を除く 719,809 語が 5 指標の付与対象サンプルに出現している異なり語である。

この 719,809 語のうち、本稿では、専門度、硬度、くだけ度、語りかけ性度の 4 指標の付与対象サンプル数が 100 以上[17]の 7,877 語をこれ以降の分析対象とする。

参考までに、表 2 にこの 7,877 語の品詞別語数を示す。

表 2　本稿で分析対象とする 7,877 語の品詞別語数

品詞	語数	品詞	語数	品詞	語数
名詞	3,967	形容詞	194	感動詞	59
動詞	2,294	助詞	137	接続詞	48
形状詞	524	助動詞	86	連体詞	29
副詞	456	代名詞	79	接尾辞	4

16 「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」及び「BCCWJ 品詞構成表(Version 1.1)」では、長単位の異なり語数は 821,025 語である。「品詞」フィールドに「URL、カタカナ文、方言、未知語、漢文、空白、英単語、補助記号、言いよどみ、記号」のタグが付けられている(本来の「品詞」以外の)計 492 語は、この語彙表の長単位の異なり語数からは除かれている。「BCCWJ 品詞構成表(Version 1.1)」と比べると本稿のデータは「名詞」が 7 語少ない。

17 専門度、硬度、くだけ度、語りかけ性度の 4 指標が付与されているサンプルは一致するため、この 4 指標の付与対象サンプル数はどの語も同じである。しかし、この 4 指標が付与されていても客観度が付与されていないサンプルがあるため、客観度の付与対象サンプル数は若干少なくなる。客観度の付与対象サンプル数が 100 未満の語も、以下の分析対象に含まれている。なお、「100 以上」としたのはある程度の大きさのサンプル数を確保するためであり、特に理論的根拠はない。

３．5 指標の平均値の全体的傾向・特徴

３.１　5 指標の相互の相関

5 指標の平均値の全体的傾向・特徴を見るために、5 指標の各平均値の相互の相関、品詞別の特徴、語種別の特徴を分析する。

まず、語別の 5 指標の各平均値の相互の相関分析を行う。

図 1 は、語別の 5 指標ごとの平均値を二組ずつセットにし、それぞれの散布図、相関係数、無相関検定結果[18]を示したものである。

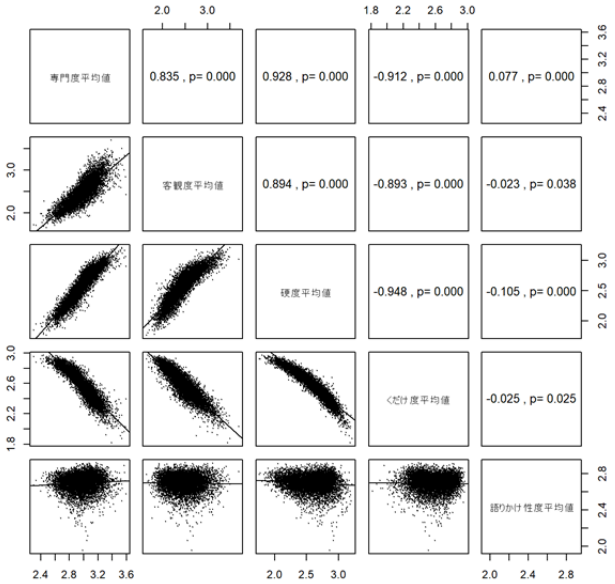


図 1　語別 5 指標各平均値相互の散布図、相関係数、無相関検定結果

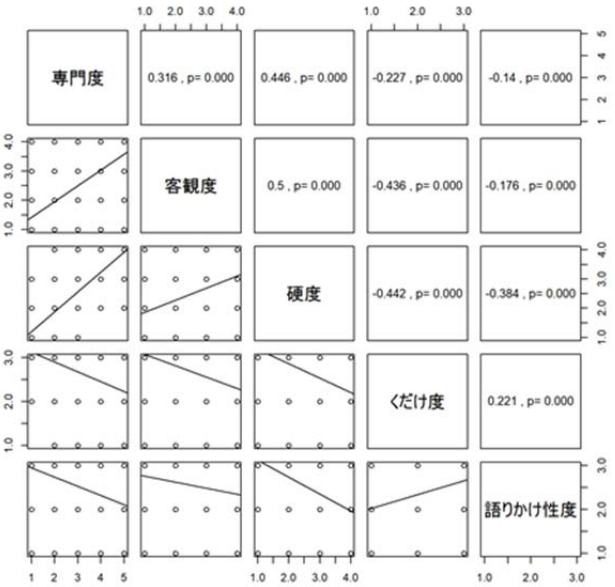


図 2　『文体情報』5 指標相互の散布図、相関係数、無相関検定結果

[18] 統計 R(ver.3.4.3) cor.test 使用。ピアソンの積率相関係数を求めた。相関係数、無相関検定結果は、小数点第 4 位以下四捨五入。

図 1 のとおり、専門度、客観度、硬度、くだけ度の 4 指標の平均値は相互に強い相関があるのに対し、これら 4 指標と語りかけ性度とはいずれも相関がない。

『文体情報』では、もともと 5 指標には相関があるのであろうか。5 指標がともに付与されている 5,901 サンプルの 5 指標の段階をそのまま用いて相関分析を行った。図 2 は、『文体情報』の 5 指標を二組ずつセットにし、それぞれの散布図、相関係数、無相関検定結果[19]を示したものである。図 2 のとおり、専門度、客観度、硬度、くだけ度の 4 指標は相互に中程度の相関ないし弱い相関があるのに対し、語りかけ性度は専門度・客観度とは相関はなく、硬度・くだけ度とは弱い相関がある。

このように、もともと『文体情報』の専門度、客観度、硬度、くだけ度の 4 指標は相互にある程度の相関はあるが、ある語がどのような「内容・表現の文体的特徴」を持つサンプルに使われやすいかという、語別の 5 指標の平均値を比べた本稿の分析方法では、専門度、客観度、硬度、くだけ度の 4 指標で表される文体的特徴は共通性が高いことが示されていると考えられる。

３.２　品詞別の特徴

本節では、品詞別に 5 指標の平均値の特徴を分析する。

図 3 は、品詞別に、5 指標ごとの全体の平均値を図示したものである。

図 4 は、品詞別に、硬度のみの語別平均値の分布を示した箱ひげ図である。

図 5 は、品詞別に、5 指標ごとの語別平均値の標準偏差を図示したものである。

図 3 を見ると、専門度、客観度、硬度、くだけ度の 4 指標で、品詞によって若干の傾向の違いがあることが分かる。特に、感動詞は他の品詞に比べて専門度、客観度、硬度の全体平均値が高く、くだけ度の全体平均値が低くなっており[20]、異なった傾向にあることが分かる[21]。感動詞は会話で使われやすいということが表されているものと思われる。図 4 は語別の硬度平均値の分布のみを、品詞別に示したものであるが、感動詞は他の品詞と分布が異なっている。図は示さないが、客観度、硬度、くだけ度についても同様である。

次に、図 5 を検討するが、接尾辞は対象とする語が 4 語で少ないためここでの検討対象からは除く。図 5 のとおり、専門度、客観度、硬度、くだけ度の 4 指標で、接続詞の語別平均値の標準偏差が他の品詞に比べて最も大きい。これら 4 指標で語別平均値のばらつきが大きいということであり、接続詞は語の違いによる文体差が他の品詞に比べて大きいこ

---

[19] 統計 R(ver.3.4.3) cor.test 使用。段階の値をそのまま用いスピアマンの順位相関係数を求めた。なお、ピアソンの積率相関係数であっても傾向は同じである。相関係数、無相関検定結果は、小数点第 4 位以下四捨五入。

[20] 専門度、客観度、硬度は、それぞれより専門的、客観的、硬いほど数値が低くなる。くだけ度は、よりくだけているほど数値が低くなる。数値が逆方向になることに注意が必要である。

[21] 5 指標のそれぞれについて、品詞を群とする等分散性の検定(Bartlett 検定)を行った結果、5 指標すべてにおいて $p < 0.001$ で各群の分散が等しくないと判断された(専門度 $\chi^2 = 75.922$、客観度 $\chi^2 = 42.742$、硬度 $\chi^2 = 84.655$、くだけ度 $\chi^2 = 51.265$、語りかけ性度 $\chi^2 = 121.45$)。そのため、5 指標のそれぞれについて、品詞を群とする Kruskal-Wallis 検定を行った。その結果、5 指標すべてにおいて群の効果は $p < 0.001$ で有意であった(専門度 $\chi^2 = 458.61$、客観度 $\chi^2 = 432.9$、硬度 $\chi^2 = 430.99$、くだけ度 $\chi^2 = 481.02$、語りかけ性度 $\chi^2 = 131.69$)。多重比較(Steel-Dwas 法)を行った結果、専門度、客観度、硬度、くだけ度の 4 指標に関しては、感動詞と(接尾辞を除く)他の各品詞との間で $p < 0.001$ で有意差があった。なお、感動詞の次に専門度等の全体平均値が高い代名詞は、(接尾辞を除くと)客観度(代名詞と副詞との組み合わせ)、硬度(代名詞と接続詞との組み合わせ)で $p < 0.05$ で有意差のない組み合わせがあった。

とが表されている。一方、この 4 指標で、感動詞が他の品詞に比べて標準偏差が最も小さい。感動詞は、(図 3、図 4 の結果と合わせると)専門度、客観度、硬度の値が高いテキスト(サンプル)、くだけ度の値が低いテキスト(サンプル)に集中して使われており、語の違いによる文体差が他の品詞に比べて小さいことが表されている。

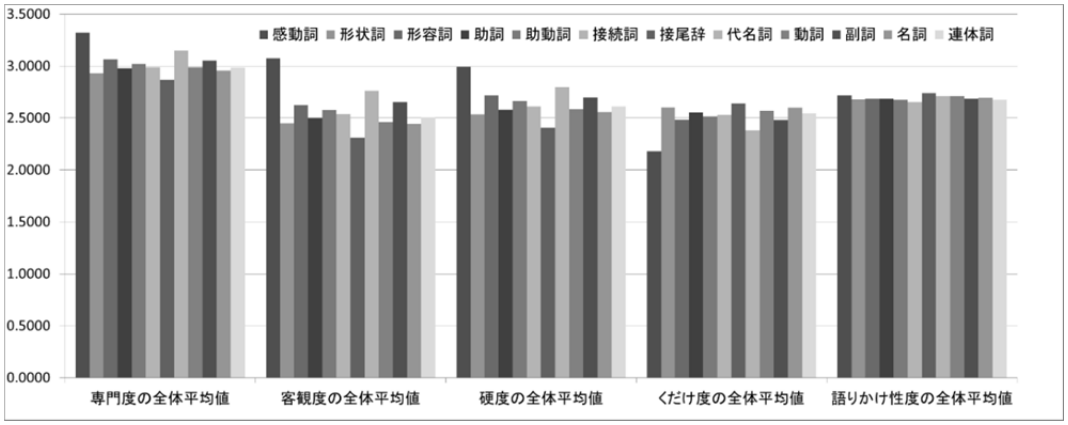


図 3　品詞別の 5 指標ごとの全体平均値

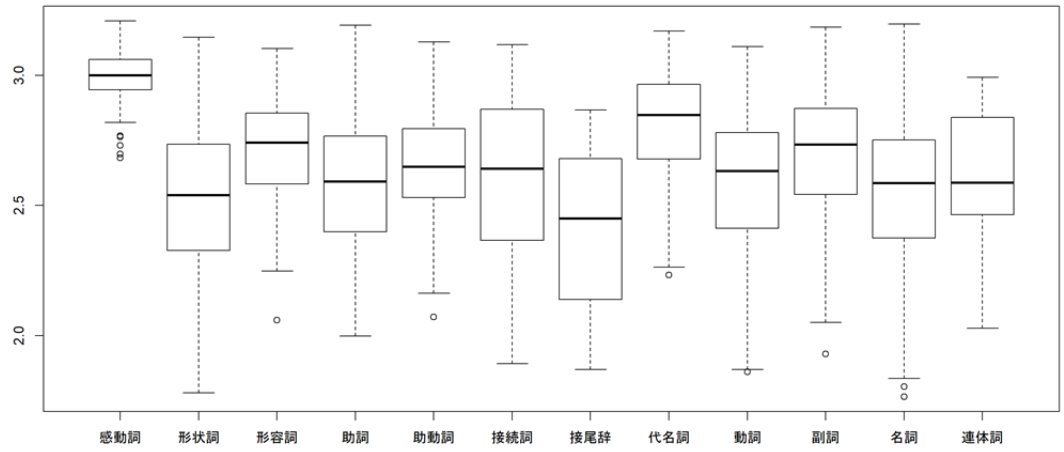


図 4　品詞別の語別硬度平均値の分布を示す箱ひげ図

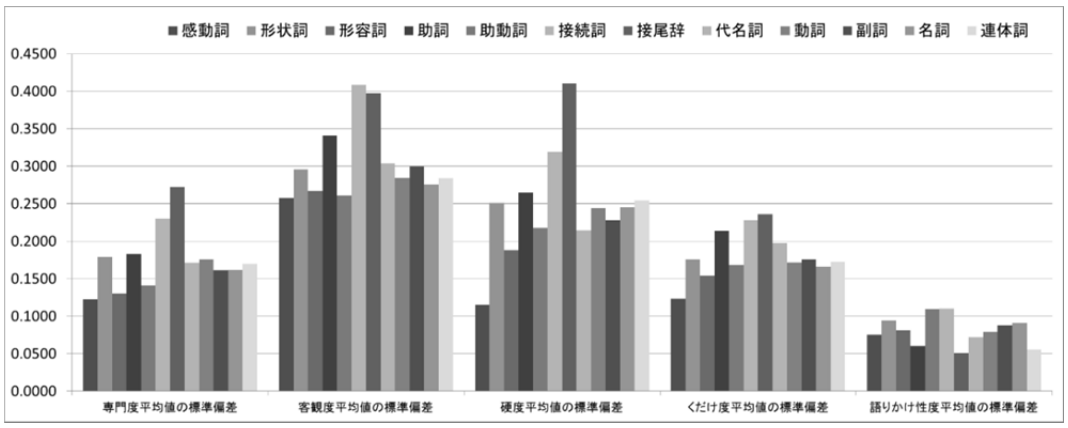


図 5　品詞別の 5 指標ごとの語別平均値の標準偏差

３.３　語種別の特徴

本節では、語種別に 5 指標の平均値の特徴を分析する。

図 6 は、語種別に、5 指標ごとの全体の平均値を図示したものである。

図 7 は、語種別に、硬度のみの語別平均値の分布を示した箱ひげ図である。

図 8 は、語種別に、5 指標ごとの語別平均値の標準偏差を図示したものである。

和語、漢語、外来語の 3 種類に絞って分析していく。

図 6 を見ると、専門度、客観度、硬度、くだけ度の 4 指標で、語種によって傾向の違いがあることが分かる。和語と外来語は、専門度、客観度、硬度の全体平均値が高く、くだけ度の全体平均値が低くなっている[22]。和語と外来語は漢語に比べて、硬くないくだけた文体のテキストで使われやすいということが表されている。図 7 は語別の硬度平均値の分布のみを語種別に示したものであるが、和語と外来語は漢語よりも硬度平均値の分布が全体的に高くなっている。図は示さないが、専門度、客観度、くだけ度についても分布の傾向は同じである。

次に、図 8 であるが、和語、漢語、外来語について、専門度、客観度、硬度、くだけ度の 4 指標に共通した特徴を指摘することはできない。専門度、客観度、硬度の 3 指標に限れば、漢語の語別平均値の標準偏差が最も大きい。すなわち、語別平均値のばらつきが大きいということであり、漢語は和語や外来語に比べて文体差が大きい傾向にあるということが表されていると見られる。

参考として、表 3 に、和語、漢語、外来語の硬度平均値の上位と下位の各 20 語(昇順)を示す。

[22] 5 指標のそれぞれについて、語種を群とする等分散性の検定(Bartlett 検定)を行った結果、5 指標すべてにおいて $p < 0.001$ で各群の分散が等しくないと判断された(専門度 $\chi^2 = 123.89$、客観度 $\chi^2 = 67.549$、硬度 $\chi^2 = 231.48$、くだけ度 $\chi^2 = 31.422$、語りかけ性度 $\chi^2 = 89.552$)。そのため、5 指標のそれぞれについて、語種を群とする Kruskal-Wallis 検定を行った。その結果、5 指標すべてにおいて群の効果は $p < 0.001$ で有意であった(専門度 $\chi^2 = 1846.9$、客観度 $\chi^2 = 1238.7$、硬度 $\chi^2 = 1622.3$、くだけ度 $\chi^2 = 1611.8$、語りかけ性度 $\chi^2 = 110.08$)。多重比較(Steel-Dwas 法)を行った結果、専門度、客観度、硬度、くだけ度、語りかけ性度の 5 指標で、和語と漢語との間及び外来語と漢語の間で $p < 0.001$ で有意差があった。また、客観度、硬度、くだけ度の 3 指標で、和語と外来語の間で $p < 0.05$ で有意差がなかった。全体として、和語と漢語との間及び外来語と漢語との間で有意差があり、和語と外来語との間では有意差がないという傾向が見られた。

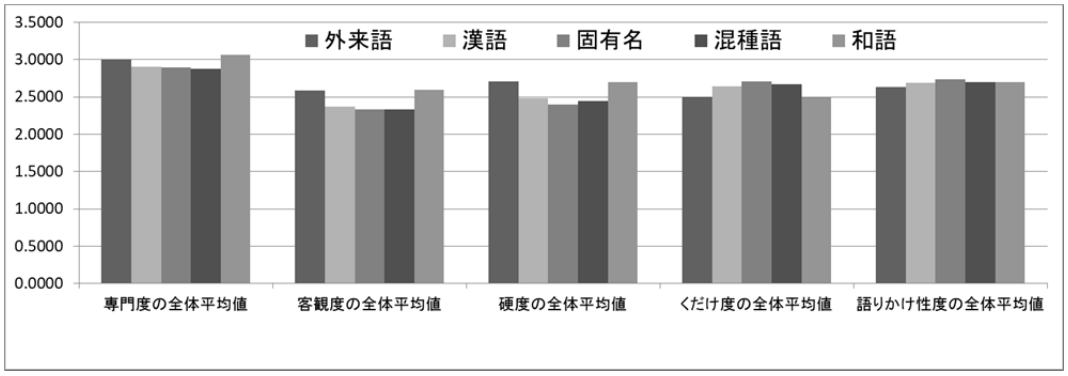


図 6　語種別の 5 指標ごとの全体平均値

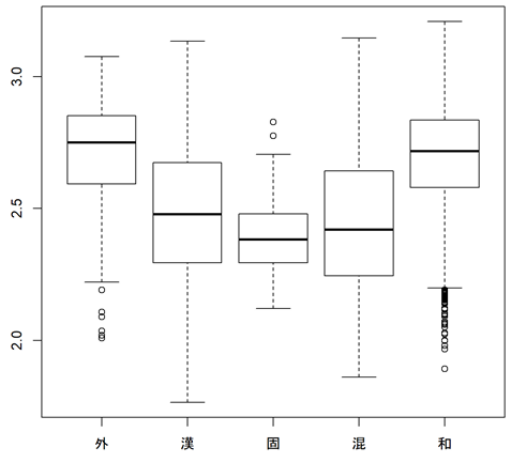


図 7　語種別の語別硬度平均値の分布を示す箱ひげ図

(外:外来語、漢:漢語、固:固有名、混:混種語、和:和語)

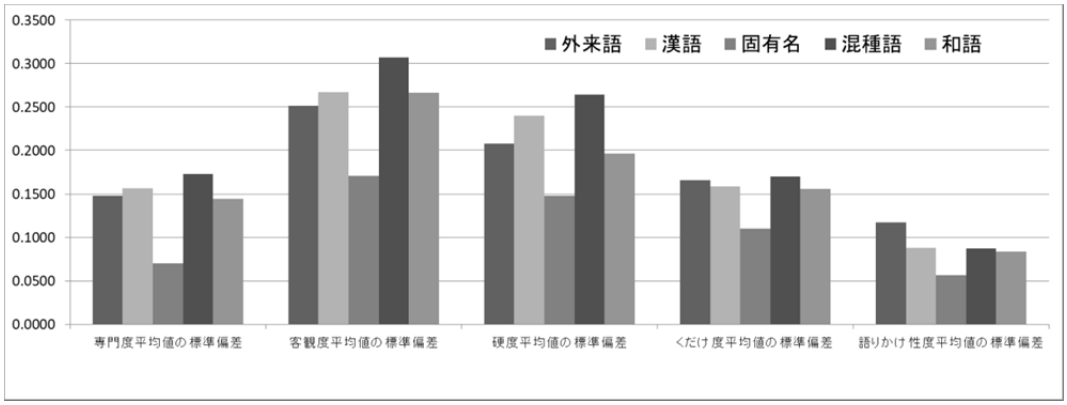


図 8　語種別の 5 指標ごとの語別平均値の標準偏差

表 3　和語、漢語、外来語の硬度平均値の上位と下位の各 20 語(昇順)

語彙素	語種	硬度平均値	語彙素	語種	硬度平均値	語彙素	語種	硬度平均値
並びに	和語	1.8929	事項	漢語	1.7658	オブ	外来語	2.0088
見直し	和語	1.9667	広範	漢語	1.7801	イン	外来語	2.0186
枠組み	和語	1.9807	要件	漢語	1.8033	アンド	外来語	2.0351
とともに	和語	1.9980	増大	漢語	1.8358	イデオロギー	外来語	2.0893
定め	和語	2.0000	形成	漢語	1.8462	デ	外来語	2.1081
及び	和語	2.0244	適用	漢語	1.8492	ザ	外来語	2.1912
のみならず	和語	2.0279	成立	漢語	1.8561	メカニズム	外来語	2.2208
基づく	和語	2.0282	規定	漢語	1.8600	ニーズ	外来語	2.2214
其れ故	和語	2.0502	等(接尾辞)	漢語	1.8703	アプローチ	外来語	2.2276
著しい	和語	2.0591	条約	漢語	1.8807	コスト	外来語	2.2423
をめぐる	和語	2.0635	紛争	漢語	1.8889	プロセス	外来語	2.2576
異	和語	2.0650	上記	漢語	1.8987	スローガン	外来語	2.2700
こととなる	和語	2.0714	中核	漢語	1.8992	ネットワーク	外来語	2.3333
基	和語	2.0909	利害	漢語	1.9000	プロジェクト	外来語	2.3333
にわたり	和語	2.0964	移行	漢語	1.9027	メディア	外来語	2.3352
担い手	和語	2.0991	実施	漢語	1.9034	システム	外来語	2.3354
における	和語	2.1022	体系	漢語	1.9118	ピーク	外来語	2.3643
押し進める	和語	2.1159	輸出	漢語	1.9252	キーワード	外来語	2.3661
欠く	和語	2.1167	従来	漢語	1.9290	シナリオ	外来語	2.3679
盛り込む	和語	2.1168	契機	漢語	1.9305	モデル	外来語	2.3681
:	:	:	:	:	:	:	:	:
で(助詞)	和語	3.1193	元気(名詞)	漢語	2.9588	バッグ	外来語	2.9510
すっ(副詞)	和語	3.1250	魔法	漢語	2.9611	トイレ	外来語	2.9518
御風呂(名詞)	和語	3.1262	洗濯	漢語	2.9649	コップ	外来語	2.9560
御昼(名詞)	和語	3.1275	暢気	漢語	2.9690	チーズ	外来語	2.9619
ちゃう(助動詞)	和語	3.1284	元気(形状詞)	漢語	2.9778	ポケット	外来語	2.9724
ふん(感動詞)	和語	3.1341	結構	漢語	2.9799	キッチン	外来語	2.9737
ふうん(感動詞)	和語	3.1390	頂戴	漢語	2.9870	ベランダ	外来語	2.9741
ずーっと(副詞)	和語	3.1397	二匹	漢語	2.9902	カップル	外来語	2.9758
すうっと(副詞)	和語	3.1441	去年	漢語	3.0000	ドレス	外来語	2.9778
こら(感動詞)	和語	3.1463	一杯(副詞)	漢語	3.0034	オーケー	外来語	2.9802
や(形状詞)	和語	3.1471	御主人	漢語	3.0044	デート	外来語	3.0000
ううん(感動詞)	和語	3.1572	変	漢語	3.0061	クリスマス	外来語	3.0075
お(感動詞)	和語	3.1681	一生懸命(形状詞)	漢語	3.0080	スカート	外来語	3.0110
私達(代名詞)	和語	3.1707	餓鬼	漢語	3.0254	ケーキ	外来語	3.0179
とっても(副詞)	和語	3.1741	御免	漢語	3.0455	バケツ	外来語	3.0190
じゃん(助詞)	和語	3.1818	御飯	漢語	3.0633	ピンク	外来語	3.0361
思いっ切り(副詞)	和語	3.1852	内緒	漢語	3.0672	プレゼント	外来語	3.0368
ねん(助詞)	和語	3.1923	一生懸命(副詞)	漢語	3.0683	ママ	外来語	3.0382
御ばあちゃん(名詞)	和語	3.1972	本当	漢語	3.1148	キス	外来語	3.0435
ん(感動詞)	和語	3.2089	一杯(名詞)	漢語	3.1346	パパ	外来語	3.0765

４．内省判断に基づく語の文体差との比較

４.１　柏野(2016)の文体 4 段階との比較

本節では、各語の文体指標の平均値(「硬度平均値」の結果を主に示す)がどの程度、語の文体差に関する内省判断と一致するかを先行研究の調査結果と比較しながら検討する。

比較する先行研究は、語の文体差に関する多数の文献の記述に基づいて主に接続詞や副詞などの文体差をまとめた柏野(2016)及び語の文体差に関するアンケート調査結果を示している井上(2013)である。

まず、柏野(2016)の「使用目安の分類」(4 段階)との比較を行う。

柏野(2016)は、大学教育や日本語教育における学術的文章作成に役立てるために、「作文技術に関する文献」及び「書き言葉と話し言葉の相互関係に関する文献」に記載された様々な語の文体に関する記述を広く調査し、「書き言葉的」「話し言葉的」として示されている文体差のある語や表現を抽出して一覧にして示し、学術的文章作成の際の「使用目安の分類」(4 段階)を行っている。この「使用目安の分類」(4 段階)は、内省判断に基づく語の文体

差に関するこれまでの文献の知見を総合したものであり極めて有意義なものである。5 指標の語別平均値とこの「使用目安の分類」(4 段階)とがどの程度一致するかを見ることによって、5 指標の語別平均値の妥当性を見ることができる。

「使用目安の分類」(4 段階)は次の a～d のように分類されている(柏野 2016:38)。この a～d をそれぞれ 1～4 の値に置き替えて比較を行う。

a. 「話し言葉的」な語と比較的はっきり位置づけられるため、学術的文章には避けるべき語
b. 「書き言葉的」な語ともいえるが、学術的文章では避けた方が望ましい語
c. 「書き言葉的」な語と比較的はっきり位置づけられるが、学術的文章の文脈・内容によっては使用に注意が必要な語
d. 「書き言葉的」な語として学術的文章での使用に特に問題のない語

比較の対象とする語は、柏野(2016)の「表 1　a.接続の「書き言葉的・話し言葉的」な語」及び「表 2　b.副詞と c.文末の「書き言葉的・話し言葉的」な語」に記載されている語のうち、接続詞と副詞(BCCWJ の長単位の接続詞と副詞のみ)の計 51 語である。

図 9 に 51 語の硬度平均値と柏野の 1(=a)～4(=d)の段階(「柏野 4 段階」)の散布図を示す。

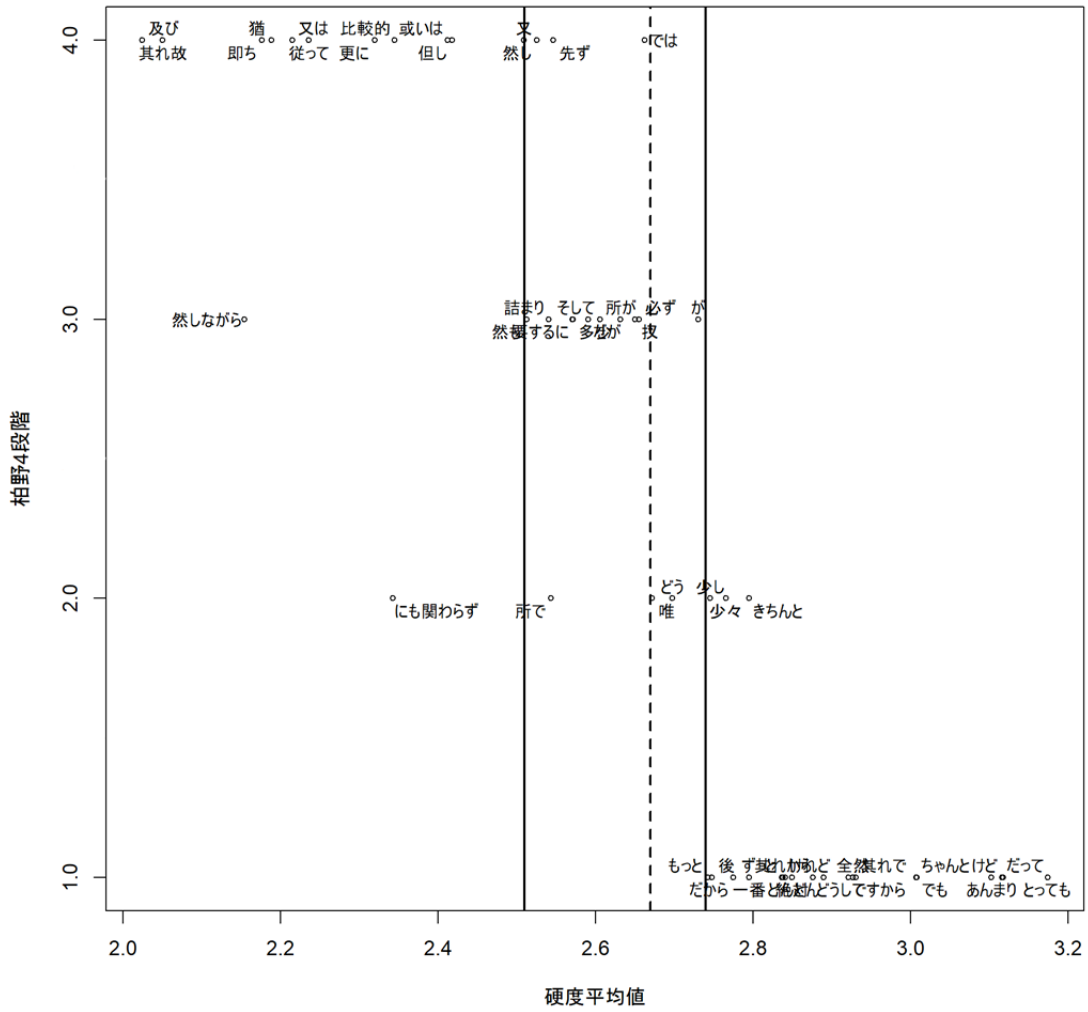


図 9　硬度平均値と柏野 4 段階との散布図(接続詞・副詞 51 語)

硬度平均値と柏野 4 段階の相関分析を行った。スピアマンの順位相関係数は r_s = -0.886（p < 0.001）であり、硬度平均値と柏野 4 段階とには強い相関がある[23]。

図 9 の散布図に基づいて、硬度平均値と柏野 4 段階との語ごとの対応を見る。

「然し、先ず、では」「然しながら、が」「にも関わらず、所で、少し、少々、きちんと」の 10 語を除くと、残りの 41 語は次のとおり対応している。ただし、この硬度平均値による区分けは一応の目安を示しただけのものである。柏野の 4(=d)の段階と 1(=a)の段階とは明確に区分けすることができる。しかし、3(=c)の段階と 2(=b)の段階との区分けは可能なのか、また、3(=c)の段階と 4(=d)の段階、2(=b)の段階の 1(=a)の段階の区分けの基準をどこに設けるのかという点は慎重に検討する必要がある[24]。

柏野の 4(=d)の段階：硬度平均値 2.5100 未満
柏野の 3(=c)の段階：硬度平均値 2.5100 以上 2.6700 未満
柏野の 2(=b)の段階：硬度平均値 2.6700 以上 2.7400 未満
柏野の 1(=a)の段階：硬度平均値 2.7400 以上

このように、硬度平均値は、内省判断による語の文体差とある程度一致しており、硬度平均値(専門度平均値、客観度平均値、くだけ度平均値)を、語の文体差の目安として用いることは可能である。

４.２　井上(2013)の「アンケート文体値」との比較

次に、井上(2013)のアンケート調査結果と比較する。

井上(2013)は、調査対象語(64 語)が「単語の文体 5 分類案」のどの分類に位置付けられるかの判断を求める調査を計 381 名に対して行い、その結果に基づいて、「アンケート調査により求めた単語の文体値(アンケート文体値)」[25]を示している。「単語の文体 5 分類案」とは、「レベル 1　卑俗体」「レベル 2　口頭体」「レベル 3　汎用体」「レベル 4　書記体」「レベル 5　文章体」の 5 分類である。

図 10 に、調査対象語のうちの 43 語[26]の硬度平均値と「アンケート文体値」との散布図を示す。

硬度平均値と「アンケート文体値」の相関分析を行った。ピアソンの積率相関係数は r= -0.817（p < 0.001）であり、硬度平均値と「アンケート文体値」とには強い相関がある[27]。

---

[23] 統計 R(ver.3.4.3) cor.test 使用。ちなみに、ピアソンの積率相関係数は r = -0.837（ p < 0.001）である。専門度平均値(r_s = -0.827, p < 0.001)、客観度平均値(r_s = -878, p < 0.001)、くだけ度平均値(r_s = 0.859, p < 0.001)の場合もそれぞれ強い相関があり同様の結果であった。語りかけ性度平均値については、スピアマンの順位相関係数は r_s = 0.193（ p = 0.175）、ピアソンの積率相関係数は r = 0.254（ p = 0.072 ）であり、有意な相関があるとは言えない。

[24] 散布図は示さないが、専門度平均値、客観度平均値、くだけ度平均値の場合も、ここで指摘した 3(=c)の段階と 2(=b)の段階との区分けの困難さが表れている。

[25] 「調査語の各文体レベルの判断人数にそのレベル値を乗じ、総和を求めた後、判断人数で除し」(井上 2013:303)て求めた平均値である。なお、井上(2013)は、「アンケート文体値」と「コーパス調査により求めた文体値(コーパス文体値)」とを比較し「コーパス文体値の有効性について検討」しているが、本稿では、「アンケート文体値」のみを利用する。

[26] 「アンケート文体値」が示されている 64 語のうち、本稿で対象とする語(語彙素)と対応させることのできるのは 43 語である。

[27] 統計 R(ver.3.4.3) cor.test 使用。なお、専門度平均値(r = -0.796, p < 0.001)、客観度平均値(r = -749, p < 0.001)、くだけ度平均値(r = 0.797, p < 0.001)の場合もそれぞれ強い相関があり同様の結果であっ

図 10 のとおり、「少々」など若干の語がやや外れた位置にあるが、語の文体差の内省判断を求めるアンケート調査の結果と硬度平均値とはある程度対応しており、やはり、硬度平均値(専門度平均値、客観度平均値、くだけ度平均値)を、語の文体差の目安として用いることは可能である。

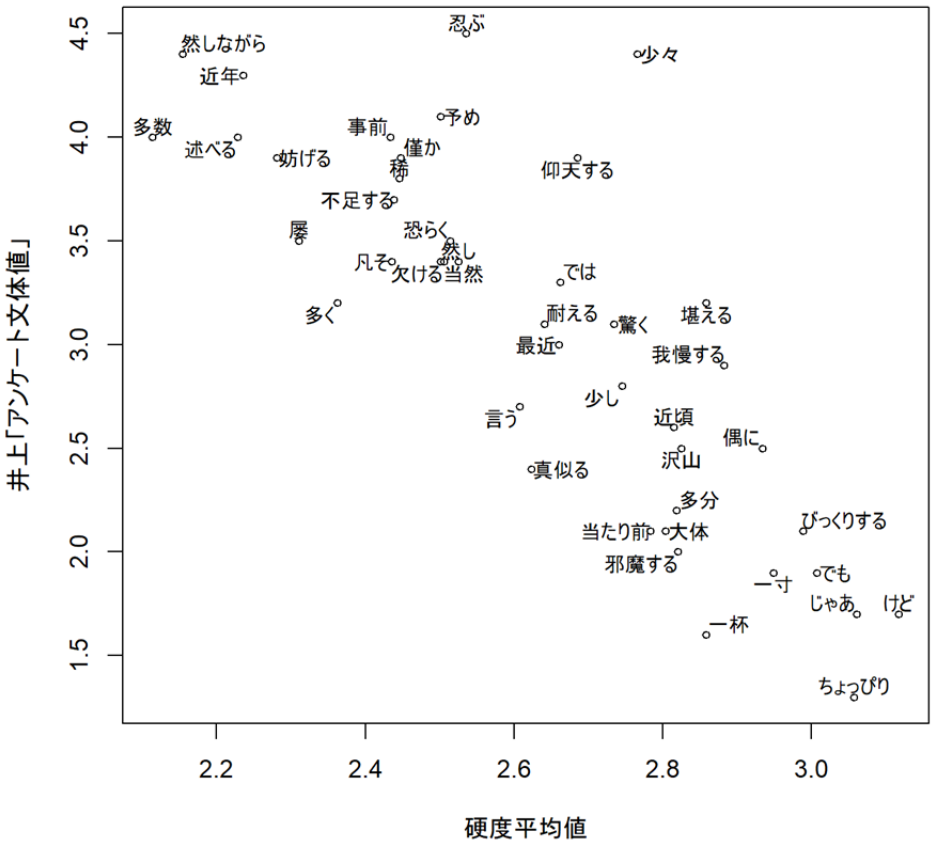


図 10　硬度平均値と井上「アンケート文体値」との散布図

５．おわりに

本稿では、『現代日本語書き言葉均衡コーパス』(BCCWJ)内の「図書館サブコーパス」のサンプルに対して文体情報を付与した『BCCWJ 図書館サブコーパスの文体情報』のアノテーションデータを利用し、「図書館サブコーパス」内の語(語彙素)を対象として各語の文体差を数値化する試みを行い、この試みが文体研究において有効性・可能性があることを示した。

本稿の概略は次のとおりである。

まず、『文体情報』の専門度、硬度などの文体に関する 5 種類のアノテーションデータを利用して、「図書館サブコーパス」内の記号類等を除く異なり語 719,809 語それぞれの専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値を算出する方法を示した。本稿では、この 719,809 語のうち、文体指標の付与対象サンプル数が 100 以上の 7,877 語を対象として種々の観点から分析を行った。

た。語りかけ性度平均値については $r = 0.210$（ $p = 0.176$ ）であり、有意な相関があるとは言えない。

次に、語別の 5 指標の各平均値の相互の相関分析を行い、専門度、客観度、硬度、くだけ度の 4 指標の平均値は相互に強い相関があるのに対し、これら 4 指標の平均値と語りかけ性度の平均値とはいずれも相関がないことを示した。ある語が使われるサンプルの「内容・表現の文体的特徴」は専門度、客観度、硬度、くだけ度の 4 指標では共通性が高いと考えられる。品詞別の分析では、感動詞と接続詞は 4 指標の平均値とそのばらつきが他の品詞に比べて特異であり、感動詞は会話のような硬くないくだけた文体のテキストに偏って使われやすいことや接続詞は語の違いによる文体差が他の品詞に比べて大きいことという特徴が表されていることを示した。語種別の分析では、和語・外来語と漢語とは 4 指標の平均値の傾向が異なっており、和語と外来語は漢語に比べて相対的に硬くないくだけた文体のテキストで使われやすいという特徴が表されていることを示した。

さらに、各語の文体指標の平均値が、語の文体差に関する内省判断と強い相関があることを、柏野(2016)の学術的文章作成の際の「使用目安の分類」(4 段階)と井上(2013)のアンケート調査に基づく語の「アンケート文体値」と比較して示した。硬度等の 4 指標の平均値は内省判断による語の文体差とある程度一致しており、硬度等の 4 指標の平均値を語の文体差の目安として用いることが可能であることを示した。

さて、本稿の硬度平均値などの 4 指標の平均値は、語の文体差を「連続的」に捉えることができるものであり、さらに、文体指標付与対象サンプル数 100 以上に限っても 7,877 語という多数の語が対象となっている。今後、この平均値を用いて、語の文体差に関する従来の知見を検証するとともに新たな知見を得るための様々な分析を行いたい。品詞別の詳細な分析、語種別の詳細な分析、あるいは複合辞の文体差などさまざまな観点からの興味ある課題が多数ある。国語教育や日本語教育への応用も課題として挙げられる。

ただし、4 指標の平均値は相互に強い相関があったが、各平均値が果たして同一の性質の文体差を表しているものなのかなど、その性質の検討も必要である。話し言葉・書き言葉、硬・軟、フォーマル・インフォーマルなど文体差は多次元的に捉えることができるが、それとの整合性も検討する必要がある。また、「書き言葉らしさ」「話し言葉らしさ」の観点から語の文体差を一次元の連続的なものとして数値化した「語彙密度平均値」(佐野ほか 2009、佐野 2009、佐野 2016)がある。4 指標の平均値とこの「語彙密度平均値」との異同を検討する必要がある。さらに、平均値以外にも、最頻値や標準偏差などの値に着目して、その特徴や活用の可能性についても検討する必要がある[28]。

このように、今後、『BCCWJ 図書館サブコーパスの文体情報』を利用した様々な研究の進展が期待される。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」「コーパスアノテーションの基礎研究」による成果の一部である『BCCWJ 図書館サブコーパスの文体情報』を利用して行われたものである。この成果を利用させていただいたことに感謝申し上げる。また、柏野和佳子氏から、『BCCWJ 図書館サブコーパスの文体情報』に関する貴重な情報をいただくとともに、氏のご論考をお送りいただいた。記し

[28] 『文体情報』を用いた研究としては、語の文体差に関わる柏野ほか(2014)などの一連の研究、各指標に関する統計的分析を加えた浅原ほか(2014,2015)や浅原・加藤(2015)の研究、特に「語りかけ性度」に関わる加藤ほか(2014)などの一連の研究などがある。

て感謝申し上げる。
(本研究は JSPS 科研費 JP16K02715 の助成を受けたものである。)

文　献

浅原正幸・加藤祥(2015)「文体指標を特徴づける係り受け部分木の抽出」『第 8 回コーパス日本語学ワークショップ予稿集』,国立国語研究所言語資源研究系・コーパス開発センター,pp.171-178.

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』,国立国語研究所言語資源研究系・コーパス開発センター,pp.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』,国立国語研究所言語資源研究系・コーパス開発センター,pp.7-16.

井上次夫(2009)「論説文における語の文体の適切性について」『日本語教育』(141),日本語教育学会,pp.57-67.

井上次夫(2013)「単語の文体判断について(3)―話しことばと書きことば―」『全国大学国語教育学会　国語科教育研究　第 125 回広島大会研究発表要旨集』,全国大学国語教育学会,pp.303-306.

柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4(1),国立国語研究所,pp.43-53.

柏野和佳子(2016)「学術的文章作成時に留意すべき「書き言葉的」「話し言葉的」な語の分類」『計量国語学会第六十回大会予稿集』,計量国語学会,pp.37-42.

柏野和佳子・立花幸子・保田祥・飯田龍・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織・椿本弥生・沼田寛(2012)「書籍テキストへの文体情報付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『第 2 回コーパス日本語学ワークショップ予稿集』,国立国語研究所言語資源研究系・コーパス開発センター,pp.155-164.

柏野和佳子・中村壮範(2014)「BCCWJ 図書館サブコーパスの文体情報検索ツールによるテキスト分析」『第 5 回コーパス日本語学ワークショップ予稿集』国立国語研究所言語資源研究系・コーパス開発センター,pp.171-180.

加藤祥・柏野和佳子・立花幸子・丸山岳彦(2014)「語りかける書きことばの表現」『国立国語研究所論集』(8),国立国語研究所,pp.85-108.

国立国語研究所(2015)『BCCWJ 図書館サブコーパスの文体情報』(第 1 版)(http://pj.ninjal.ac.jp/corpus_center/anno/) (BCCWJ_LB_Stylistics-1.0.zip).

国立国語研究所コーパス開発センター(2015)『『現代日本語書き言葉均衡コーパス』利用の手引　第 1.1 版』(http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html).

佐野大樹(2009)「「話し言葉らしさ・書き言葉らしさ」の計測―語彙密度の日本語への適用性の検証―」『機能言語学研究』5,日本機能言語学会,pp.89-102.

佐野大樹(2016)「語彙密度から見た語彙シラバス」森篤嗣(編)『ニーズを踏まえた語彙シラバス』,くろしお出版,pp.79-93.

佐野大樹・丸山岳彦・山崎誠・柏野和佳子・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子(2009)『語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み』(特定領域研究「日本語コーパス」平成 20 年度研究成果報告書),文部科学省科学研究費特

定領域研究「日本語コーパス」データ班.
馬場俊臣(2018)「接続詞の文体差の計量的分析の試み―『BCCWJ 図書館サブコーパスの文体情報』を用いて―」『北海道教育大学紀要　人文科学・社会科学編』69(1),北海道教育大学,pp.1-14.
宮島達夫(1977)「単語の文体的特徴」松村明教授還暦記念会(編)『松村明教授還暦記念　国語学と国語史』,明治書院,pp.871-903.

関連 URL

「『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説」 http://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu.html (BCCWJ 語彙表解説_1.1.pdf)
国立国語研究所(2015)『BCCWJ 図書館サブコーパスの文体情報』(第 1 版) http://pj.ninjal.ac.jp/corpus_center/anno/ (BCCWJ_LB_Stylistics-1.0.zip)
国立国語研究所共同研究プロジェクト「コーパスアノテーションの基礎研究」成果物配布サイト http://pj.ninjal.ac.jp/corpus_center/anno/
「BCCWJ 品詞構成表(Version 1.1)」 http://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu.html (BCCWJ_frequencylist_pos_ver1_1.zip)

脚本テキストに基づくコーパス文体論の可能性
―テレビドラマ脚本に注目して―

松下 晶子（専修大学大学院文学研究科）*
丸山 岳彦（専修大学文学部 / 国立国語研究所）

Possibility of corpus stylistics with scenario corpora

Shoko Matsushita (Graduate School of Letters, Senshu University)
Takehiko Maruyama (Senshu University / NINJAL)

要旨

現在、「日本脚本アーカイブズ推進コンソーシアム」により、1950 年代以降のテレビドラマの脚本を収集し、それらを体系的に保存・アーカイブ化する活動が進められている。脚本は、「話されることを前提とした書き言葉」という点で特徴的な書き言葉であるが、これまでの言語研究の中で顧みられることは少なかった。収集した脚本をコーパス化して定量的に分析することにより、新たな言語学的利用の可能性が開かれると考えられる。そこで本発表では、脚本のテキスト化・コーパス化を試験的に実施した経緯を述べ、そのデータを使ってどのような言語研究が可能になるかについて論じる。故市川森一氏による、1970 年代から 2010 年代までの脚本、32 作品をテキスト化し、パイロットスタディを実施した。このような分析は、近現代における言語の短期的な変化の研究、ある作家の作品に関するコーパス文体論的研究などにつながると考えられる。

1. はじめに

一口に「書き言葉」と言っても、その内実は一様ではない。実際の書き言葉は、ある特定のレジスター（言語使用域）の中で実現するものであり、レジスターの違いに応じて、多種多様な様式・スタイルを持つ。すなわち、書き言葉には、レジスターごとに多様な変異が存在する。

2011 年に公開された『現代日本語書き言葉均衡コーパス』（BCCWJ）は、書籍・雑誌・新聞・白書・教科書・広報紙・Yahoo!知恵袋・Yahoo!ブログ・韻文・法律・国会会議録という 11 種類のメディアから抽出した書き言葉のサンプルを収録しており、公開以来、メディアの違いに応じた言語変種（バラエティ）の分析が盛んに進められている。一方、BCCWJ に収録されなかったタイプの書き言葉も、多く存在する。本稿で取り上げる「脚本」も、その一つである。脚本は、これまで言語研究の中で分析対象データとして扱われることの少なかったメディアであり、その言語的な特徴の分析は、遠藤ほか (2014) を除き、ほとんど進んでいない。「話されることを前提とした書き言葉」という点で、脚本はユニークな性格を持つ書き言葉であるが、では、「脚本コーパス」を作った場合、言語分析の観点から、どのように利用できるであろうか。

本稿では、脚本をコーパスとして利用することにより、どのような言語学的分析が可能かに

* arukumatsushita@gmail.com

ついて論じる。現在、「日本脚本アーカイブズ推進コンソーシアム」によって実施されているテレビドラマの脚本をアーカイブ化する活動を紹介し、その中で筆者らが進めている脚本のテキスト化・コーパス化の経緯について述べる。さらに、そのデータを用いてどのような言語研究が可能になるかについて、パイロットスタディの結果を交えながら論じる。そこから、近現代における言語の短期的な変化の研究や、ある脚本家の作品を対象とした「コーパス文体論」の研究などに展開する可能性について指摘する。

2. 「日本脚本アーカイブズ推進コンソーシアム」の活動

一般社団法人「日本脚本アーカイブズ推進コンソーシアム（NKAC）」(1) は、2012 年 6 月に設立されて以降、テレビ・ラジオ放送（特にテレビドラマ）の脚本を収集し、アーカイブとして体系的に保存する活動を継続している。テレビドラマの脚本は出版物ではないため、図書館で収集・保存されることは原則的になく、制作後は放送局や番組関係者によって「保管」されることが一般的である。このため、体系的な収集・管理ができず、古い時代のものは散逸して失われてしまう可能性が高い。その一方、特に 1980 年代以前のテレビ放送は録画テープを重ね録りして使用していたため、当時の映像・音声自体が放送局に残っていないことも多い。

過去のテレビドラマの詳細を知る手掛かりとして、脚本を収集してアーカイブ化すれば、日本のテレビ放送史における放送文化研究やドラマ研究の分析対象データとして役立てることができる。また、脚本として書かれた日本語が、テレビ放送の歴史の中でどのように変化してきたのかを探るための言語資料として考えれば、言語学的な研究にとっても極めて有用であろう。

NKAC の活動により、2017 年までに、1930 年代から 2010 年代までの脚本、約 8 万点が集められた。これらは現在、国立国会図書館や川崎市民ミュージアムなどに分置されている。また、国立国会図書館に納められた脚本約 27,000 点のうち、3,000 点を超える分がデジタル化され、全国の図書館に配信されている。さらに、収集された脚本はデータベースに登録され、「脚本データベース(2) 」として、ウェブ上で公開されている。ここでは、ドラマタイトル、作家、放送局、放送日、キャストなどの情報を検索することができる(3)。なお、このデータベース化事業は、文化庁委託事業「文化関係資料のアーカイブ構築に関する調査研究～放送番組の脚本・台本のアーカイブ構築に向けた調査研究～」などによる成果である。

また、市川森一、永六輔、藤本義一といった脚本家については、その脚本作品のリストや本文（の画像）、脚本家の足跡や関係者のインタビューなどが、特設サイト上にまとめられている(4) (5) (6)。

2017 年度からは、科研費基盤 (B)「「脚本クロニクル」サイト構築とその教育活用および国際発信」もスタートし、その研究・教育への利用の模索も開始された。現在は、脚本の収集・データベース化と、その研究利用という両面から、活動を継続している。

---

(1) 「日本脚本アーカイブズ推進コンソーシアム」https://www.nkac.jp/
(2) 「脚本・台本の総合一覧 脚本データベース」http://db.nkac.or.jp/
(3) ただし、脚本の本文にあたる部分の閲覧・検索はできない。
(4) 「市川森一の世界」http://ichikawa.nkac.or.jp/
(5) 「永六輔バーチャル記念館」http://eirokusuke.nkac.or.jp/
(6) 「藤本義一アーカイブ」http://fujimotogiichi.nkac.or.jp/

3. 「市川森一 脚本コーパス」の構築

現在、科研費基盤 (B)「「脚本クロニクル」サイト構築とその教育活用および国際発信」における活動の一環として、「脚本」をテキストデータ化し、脚本の言語学的利用という観点からその利用可能性を探っている。以下では、これまでにパイロット的に作成している「市川森一脚本コーパス」について述べる。

故市川森一氏（1941～2011）は、かつて日本放送作家協会理事を務め、後の NKAC の発足に尽力した脚本家である。1960 年代に『快獣ブースカ』や『ウルトラマン A』などの子ども番組で活動を開始し、1970 年代以降は『傷だらけの天使』『黄金の日日』『花の乱』『港町純情シネマ』『淋しいのはお前だけじゃない』『異人たちとの夏』『幽婚』など、数々のテレビドラマ作品を手掛け、多くの受賞作品を生み出した。2003 年には国会の総務委員会において脚本アーカイブズ活動の必要性を提言し、その活動が、2012 年の NAKC の設立につながっている。

NKAC では、「デジタル脚本アーカイブズのトライアル」として、市川森一氏によるドラマ脚本やその関連書類（手書き原稿、創作ノート）をデジタル化し、「市川森一の世界」としてウェブ上に公開している。この延長線上にある活動として、今回、市川森一による脚本をテキスト化し、「市川森一 脚本コーパス」を作成した（今のところ、一般公開の予定はない）。

今回、コーパス化したのは、以下の 3 冊に含まれる脚本のテキストである。

1. 『市川森一 センチメンタルドラマ集』（1983 年、映人社） 14 作品収録
2. 『市川森一 メランコリックドラマ集』（1986 年、映人社） 10 作品収録
3. 『市川森一 メメント・モリドラマ集』（2012 年、映人社） 8 作品収録

全ページを OCR で読み込み、人手で校正したうえで、テキストファイルを作成し、さらに Mecab + unidic-cwj-2.3.0 で形態素解析を実施した。また、セリフとト書きの部分を区別するタグを付与して、両者を区別できるようにした。

次に、3 冊に含まれる脚本（32 作品）を放送年の順に並べ、全体を以下のように区分した。それぞれの総語数（句読点含む）と、各期に含まれる作品タイトルを、以下に示す。

第 1 期： 1969 年～1979 年 放送 （13 作品、165,042 語）

「仮面の墓場」「祭りのあとに」「林で書いた詩」「冬の時刻表」「夢に吹く風」「みどりもふかき」「紙コップのコーヒー」「夢のながれ」「霧の日の童話」「幻のぶどう園」「バースデイ・カード」「ラスト・ダンス」「露玉の首飾り」

第 2 期： 1980 年～1989 年 放送 （11 作品、137,411 語）

「春のささやき」「チャップリン暗殺計画」「蝶の鼓」「いもうと」「夢の指環」「鬼の恋舟」「受胎の森」「星の旅人たち」「途中下車」「ただ一度の人生」「中国服の女」

第 3 期： 1998 年～2011 年 放送 （8 作品、122,329 語）

「幽婚」「ここではない何処か」「乳房」「風の盆から」「銀河鉄道に乗って」「月の光」「旅する夫婦」「蝶々さん」

1 作品に含まれる語数は、最小で 7,780 語、最大で 39,086 語、平均で 13,274 語であった。

4. 分析

以下では、パイロットスタディとして、「市川森一 脚本コーパス」を分析した結果について示す。分析の観点は、(1) 終助詞の分布、(2) 二人称代名詞の使用、という 2 点である。

4.1 終助詞の分布

はじめに、脚本コーパスの中に現れた終助詞の分布について分析を行なう。

まず、表 1 は、第 1 期～第 3 期の各期に現れた終助詞の総数と出現率をまとめたものである。この表からは、時代を追うごとに、終助詞の出現率が減っていることが読み取れる。

表 1　各期における終助詞の総数と出現率

	総語数	総終助詞数	終助詞率
第 1 期	165,042	4,014	2.43%
第 2 期	137,411	2,776	2.02%
第 3 期	122,329	2,142	1.75%

次に、図 1 は、各期の終助詞のうち上位 5 つを抽出し、終助詞の総数に占める割合をグラフにしたものである。

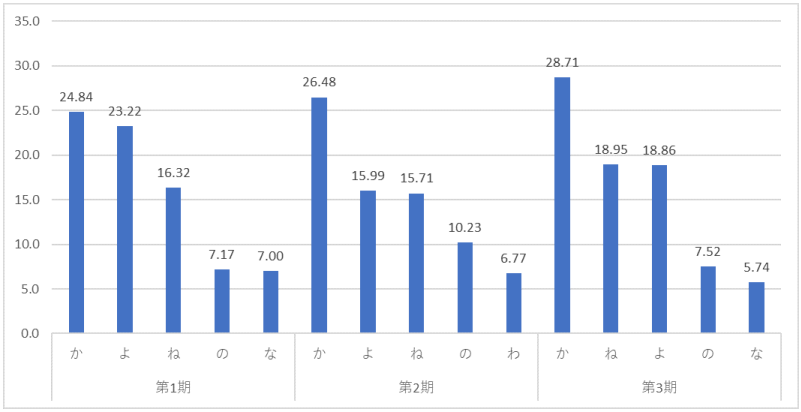


図 1　各期の終助詞（上位 5 位）の総終助詞数に占める比率

先にも述べた通り、時代を追うごとに終助詞の数は減少しているが、出現している終助詞の形式は、期によって異なる。どの期でも一番多いのは、「か」であるが、2 位と 3 位に注目すると、第 1 期では「よ」「ね」の登場回数は差が開いているが、第 2 期になるとその差が縮まり、第 3 期になるとわずかながら「ね」が「よ」を上回っている。

次に、終助詞の分布を「登場人物の性別」という観点から見てみよう。なお、今回の調査では、発話数が上位 5 位の中に入り、かつ 20 以上の発話をしている登場人物のセリフに限定して分析している。

図 2 は、女性登場人物が用いる終助詞の上位 5 つを抽出し、総終助詞数に対する比率をグラフにしたものである。ここではどの期でも登場する終助詞の種類は一緒だが、その順位が違う

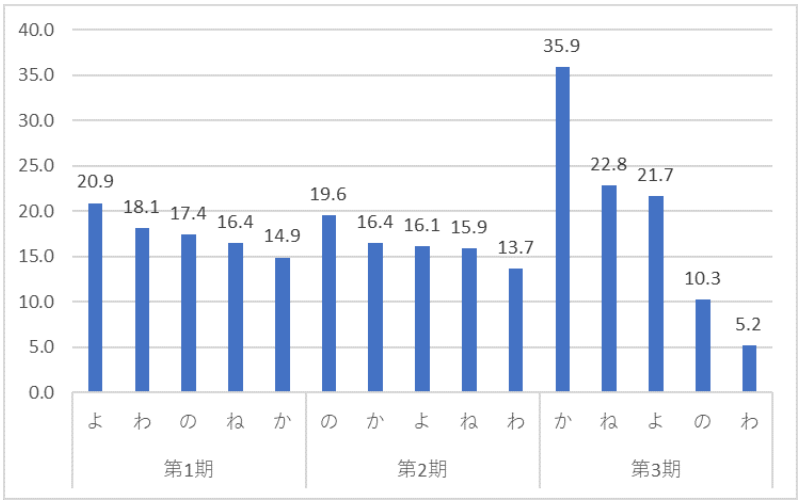


図 2　総終助詞数に対する各期の終助詞（上位 5 位）の比率（女性）

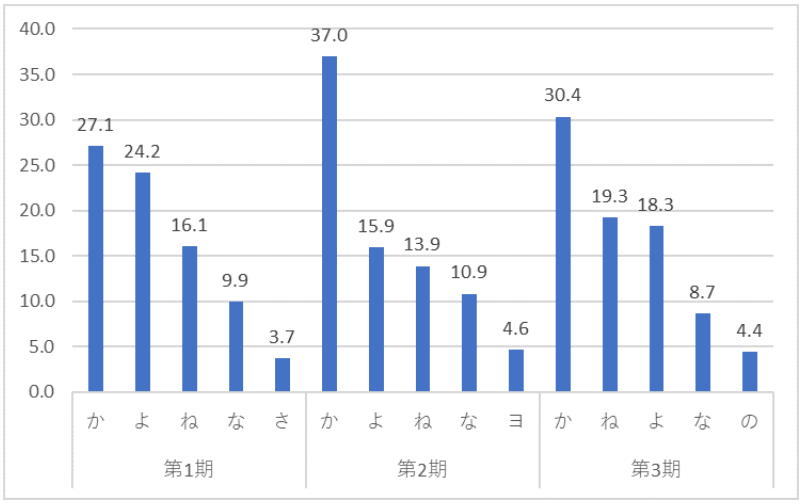


図 3　総終助詞数に対する各期の終助詞（上位 5 位）の比率（男性）

点が特徴的である。その中でも、特に女性らしい終助詞とされている「わ」は、第 2 期からその順位を下げ、上位 5 位に入ってはいるものの、その割合は大きく減少している。以下、各期の女性登場人物が用いる終助詞「わ」を含む文を挙げる。

(1)　「判ったわ。わざわざどうも（と席へ）」（第 1 期）

(2)　「懐しかったわ。ちっとも変わってなくて」（第 2 期）

(3)　「？……まだ、お義父さまにもご挨拶してないわ。」（第 3 期）

「女性らしさ」を表す終助詞として「わ」が用いられていることは、どの期でも変わっていない。しかし、割合が減少しているという事実から考えられることとして、女性らしさとして「わ」を用いる割合が減っていったというだけでなく、「女性」を象徴する役割語として「わ」を用いる割合も減っていった、という点が挙げられるのではないだろうか。この現象の要因は、当時の時代背景によるものなのか、脚本家自身考え方によるものなのか、現段階では分からない。今後、既存のコーパスと照らし合わせたり、脚本家自身の時代ごとの境遇を照合したりと

いった作業が必要である。

一方、図 3 は男性の登場人物が用いる終助詞の割合である。男性は、第 3 期で「ね」と「よ」が逆転したこと、どの期でも 5 位に異なる終助詞が挙げられている点が特徴的である。ここでは、第 1 期の男性の登場人物が用いた、終助詞「さ」を含む例を見ておこう。

(4) 「テニヲハが抜けてるンだ。ウエルニッケのいわゆる言語促迫。クルーベリン学派の術語では、支離滅裂というヤツさ」

(5) 「女を喰いもんにしていた二枚目の芋学生さ」

ここから、発話者の気取った発言に対し終助詞「さ」が用いられている印象を受ける。特に第 1 期には気取った発言をする性格の男性が多く登場する傾向があり、このために「さ」が上位に位置づけられたと考えられる。

なお、終助詞がカタカナで表記される場合が目についた。その比率を、表 2 に挙げる。

表 2　終助詞がカタカナで表記される割合

	第 1 期	第 2 期	第 3 期
ヨ	5.8% (57)	22.6% (130)	0.5% (3)
ネ	7.0% (49)	5.4% (25)	5.1% (22)
ワ	1.8% (4)	19.7% (46)	2.7% (2)
ナ	4.1% (12)	10.9% (20)	0.0% (0)

（カッコ内はひらがな終助詞とカタカナ終助詞の総数）

表 2 は、ひらがな終助詞とカタカナ終助詞の総数に対する、各期のカタカナ終助詞率をまとめたものである。ここでは、ひらがな終助詞の上位と重なっている「ヨ」「ネ」「ワ」「ナ」を抽出した。これを、1 万語あたりの調整頻度としてグラフ化すると、図 4 になる。

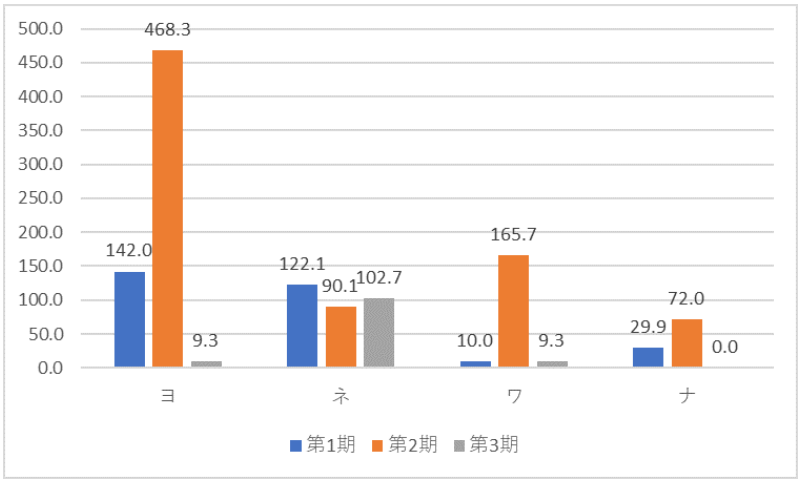


図 4　カタカナで表記された終助詞の頻度（1 万語あたり）

「ネ」「ナ」には目立った差はないが、「ヨ」「ワ」に関しては期によってかなりの差があることが、グラフから読み取れる。以下、カタカナで終助詞が用いられている例を挙げる。

(6) 「A さん、なに訊かれたのヨ」

(6) は女性の発話であるが、ひらがなの「よ」に比べて、上昇調のようなニュアンスが加わり、単なる「質問」に加えて、からかいや冷やかしのような意味合いが感じられる。

(7) 「全て、銭次第ですヨ」

(7) は男性の発話であるが、ひらがなの「よ」の場合に比べて、ここでは悪だくみや誘惑といったマイナスの言い含みを感じやすくなるように思われる。

このように、カタカナの終助詞を用いることで、ひらがなを用いる時とは印象が変わり、発話者の心情やニュアンスを豊富に表現する効果が感じられる。その他の種類の終助詞について、その表記の違いと用法・効果について分析することは、今後の課題の一つである。

4.2 二人称代名詞の使用

以下では、二人称代名詞の「あなた」「あんた」について分析してみよう。図 5 は、各期における「あなた」「あんた」の出現数を、1 万語あたりの調整頻度としてグラフ化したものである。

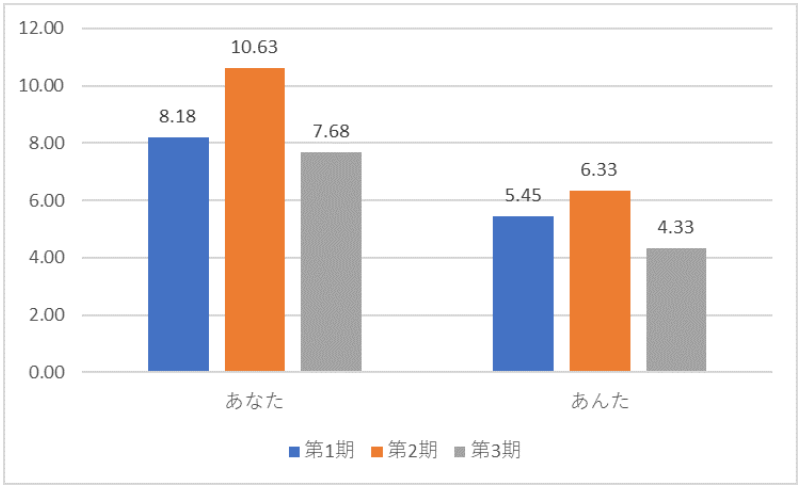


図 5 各期における「あなた」「あんた」の出現数（1 万語あたり）

第 1 期での「あなた」「あんた」の数が、それぞれが第 2 期で増えて、第 3 期で第 1 期を少し下回る数になっている。これには、何か要因があるのだろうか。

ここで、「あなた」「あんた」の出現数（粗頻度）を、登場人物の性別という観点から調査してみよう。ここでも、終助詞の場合と同様に、発話数の合計が各作品で上位 5 位までの人物で、かつ 20 以上の発話をしている人物に絞ったデータを使用する。

表 3 性別ごとの「あなた」「あんた」の出現数

	あなた		あんた	
	女性	男性	女性	男性
第 1 期	55	58	10	69
第 2 期	66	64	46	15
第 3 期	51	15	22	27

図 6 で特徴的なのは、それぞれ男女内でグラフが逆の形をしていることである。すなわち、

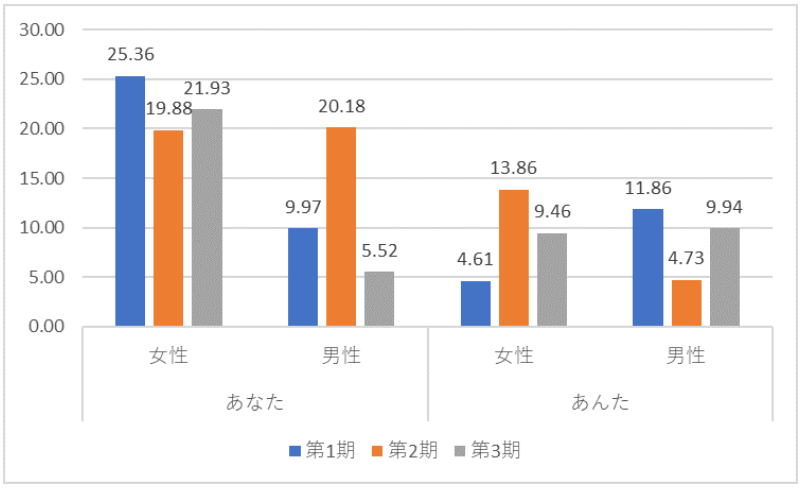


図 6　性別ごとの「あなた」「あんた」の出現数（1 万語あたり）

「あなた」と「あんた」の使用において、男女間で相補的な関係になっていることが分かる。この点については、今後、登場人物の男女それぞれがどのような相手に向かって「あなた」「あんた」を使っているのか、「あなた」「あんた」それぞれを使う登場人物がどのような性格なのかといった、具体的な場面分析を合わせて行なうことで、その要因を分析することができると考えられる。

5.　考察：「脚本」が言語研究に対して持つ可能性

ここまでの分析結果を受けて、脚本コーパスを言語研究に用いることの意義と可能性について、若干の考察を加えておく。脚本コーパスを言語研究に用いることの意義として、以下では 2 点に分けて整理しておこう。

1 点目は、「話すことを前提として書かれた言葉」が持つ言語的特徴を明らかにすることである。脚本で書かれたセリフの部分は、話すことを前提として書かれた書き言葉である。これは、実際の話し言葉とは大きく異なる（例えば、脚本には非流暢性がほとんど現れない）。では、同じ「書かれた話し言葉」である「小説の会話部分」に現れる言葉とは、同じだろうか、違うだろうか。

脚本には、例えば以下のように、「…」のみで構成されるセリフが多く見られる。

(8) 骨の折れる音が、場内に反響する。棒立ちのまま、息をのむ、犬尾、山口、ヨーコ。
静寂……。
犬尾「……」
ヨーコ「……」
山口「……」
ジュン「……」

これらは登場人物の沈黙を表すものであるが、これほど多くの「…」の連続は、小説にはまず出現しない。このような表現は、映像（沈黙する登場人物たち）を基盤とする脚本ならではのものであろう。今後は、BCCWJ の小説に現れる発話部分を取り出し、脚本のセリフと定量

的に比較・分析することで、脚本の言語的特徴を明らかにすることが課題の一つとなる。

2 点目は、個人の脚本家の作品群を集めることで、「コーパス文体論」の研究への道が拓かれるのではないか、という点である。1 人の文学作家の手による作品群を集めて、例えば「夏目漱石コーパス」や「太宰治コーパス」を作成し、両者の文体的な違いや時代ごとの特徴を定量的に比較・分析していく、という研究手法は、今後、大きな可能性を持つと考えられる。Stubbs(2014) はコンピュータを用いた文学テクストの量的分析の可能性について論じているが、これはすなわち、「コーパス文体論」の研究の方向性を示唆するものである。英語を対象としたコーパス言語学では、そのような試みがすでにいくつも発表されており（Studer 2012, Mahlberg 2015 など）、今後は日本語でもコーパス文体論的な研究が望まれる。

脚本を分析対象とする場合にも、同じことが言えると思われる。複数の脚本家による脚本を時代ごとに収集してコーパス化することにより、個人の作風の変化を探ったり、現代語の短期的な変化の影響を分析したり、異なる脚本家と文体を比較したりするような研究が可能になるだろう。今回パイロット的に作成した「市川森一 脚本コーパス」はそのような方向に向かう取り組みであり、「脚本コーパスに基づく文体論」につながるものと考えている。

6. おわりに

以上、本稿では、脚本のテキスト化・コーパス化を試験的に実施した経緯を述べ、そのデータを使ってどのような言語研究が可能になるかについて示した。「市川森一氏 脚本コーパス」を試作し、そのデータを用いたパイロットスタディとして、終助詞と二人称代名詞「あなた」「あんた」の分布について見た。

終助詞の調査では、性別や年代、表記といった観点からその分布を見た。特に、カタカナで表記される終助詞は、それが用いられている状況やそれを用いた登場人物の性格といった面から、カタカナ終助詞が持つ役割を分類できそうである。それらをひらがな終助詞と比較することにより、脚本の文体的な特徴を示す一例として考えることができると思われる。

また、二人称代名詞「あなた」「あんた」の調査では、第 2 期が特徴的な動きをしていることが判明した。これには、時代が影響しているのか、脚本家が置かれた当時の状況が影響しているのか、考えられる要因は多くある。今後、既存の書き言葉コーパス・話し言葉コーパスと照合しながら調査を進めることで、データに裏打ちされた脚本文体論の研究を進めることができるだろう。

脚本コーパスの量をさらに増やし、定量的な分析を実施する環境整備を行なうこともまた、今後の課題である。これにより、コーパスに基づく文体論研究の一例として、より精度を増した形で、脚本の言語学的な分析を進めたい。

謝　辞

本研究は、科研費基盤 (B)「「脚本クロニクル」サイト構築とその教育活用および国際発信」（17H02598）、科研費基盤 (B)「「昭和話し言葉コーパス」の構築による話し言葉の経年変化に関する実証的研究」（16H03426）、および国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」によるものである。

文　献

Mahlberg, Michaela (2015) *Corpus Stylistics and Dickens's Fiction.* Oxford: Routledge.

Stubbs, Michael（2014）Quantitative methods in literary linguistics. Peter Stockwell, P. and Whiteley, S. (eds.) *The Cambridge Handbook of Stylistics.* 46–62. Cambridge: Cambridge University Press.

Studer, Patrick (2012) *Historical Corpus Stylistics: Media, Technology and Change.* Continuum.

遠藤織枝・木村拓・桜井隆・鈴木智映子・早川治子・安田敏朗 (2004)『戦時中の話しことば ラジオドラマ台本から』ひつじ書房.

『UniDic』を活用した語構造情報付与の試み ―『日本語歴史コーパス』を対象に―

村山 実和子（国立国語研究所 言語変化研究領域）†

Annotation of Word Structures in Japanese using "UniDic" : A Study of the Corpus of Historical Japanese

Miwako Murayama (National Institute for Japanese Language and Linguistics)

要旨

本研究は『日本語歴史コーパス』に出現する合成語に対し，その内部構造に関する情報を新たに追加することで，日本語の語形成研究に使用可能なデータの構築をめざすものである。その方法として，各種コーパスに紐付いた解析用辞書「UniDic」の見出し語に対して，構成語情報を付与することを試みる。その設計方針と有用性を述べるとともに，現状の課題について報告する。

1. はじめに

国立国語研究所（以下，国語研）では，上代から近代までの日本語を通時的に研究するための基礎資料として，『日本語歴史コーパス（以下，CHJ）』（国立国語研究所 2016）の構築を進めている。すべてのテキストを斉一な単位で分割し，詳細な形態論情報を付与している点が特長である。ただし，その単位認定は基本的に現代語のコーパスに準拠するものであり，時代によっては「語」の認定に揺れが生じる場合もある。例えば，ある語の連続を複合語とみなすかどうか，現代語では内省可能なものでも，古代語ではその認定は難しい。そのため，CHJ では現状，奈良～鎌倉時代編までのコーパスでは複合動詞を認定しないが，室町時代編以降は現代語の規程に照らして複合動詞を認めるというように，時代ごとに異なる対処を行っている[1]。以下，用例中の「｜」は単位の境界を表しており，(1a)の［聞き入れ］が四段動詞「聞く」の連用形＋下二段動詞「入れる」の未然形，と分割されるのに対し，(1b)の［聞き入れ］は，下二段動詞「聞き入れる」の未然形として処理される。

(1) a. ｜御前｜近き｜人｜など｜の｜けしき｜ばみ｜言ふ｜を｜も｜聞き｜入れ｜ず｜
【出典】CHJ サンプル ID：20-枕草 1001_00120『枕草子』
b. ｜蛇｜情強｜に｜し｜て｜少し｜も｜聞き入れ（qiqiire）｜なんだ｜に｜因っ｜て｜
【出典】CHJ サンプル ID：40-天伊 1593_00067『天草版伊曾保物語』

また，単位認定の規程上（→3.2 節），前項または後項の独立性や語種によって，形態的に同一のものであっても，単独で扱う場合と，語の一部として扱う場合がある点にも注意が必要である。これは接頭・接尾語の類にしばしば見られるものである。例に挙げた「めく」は UniDic では単独の要素（品詞：接尾語-動詞的）として扱われている。しかし，上接語が象徴語など拘束的な要素である場合には，語の内部要素として処理される。

† m-murayama@ninjal.ac.jp

[1]『日本語歴史コーパス平安時代編』および『同　鎌倉時代編』形態論情報規程集の記述に拠る

(2) a. ｜鳥｜の｜声｜など｜も｜こと｜の｜外｜に｜春｜めき｜て｜
【出典】CHJ サンプル ID：30-徒然 1336_01019 『徒然草』
b. ｜濡れ｜たる｜やう｜なる｜葉｜の｜上｜に｜きらめき｜たる｜こそ｜
【出典】CHJ サンプル ID：30-徒然 1336_01137 『徒然草』

上代から近代まで幅広い時代の資料を検索対象とする CHJ においては，形態論情報付与のために一定の基準を設けることは運用上必要不可欠であるといえる。ただしそれは現代語の状況に立脚したものであるため，基本ルールから逸脱するものは時代ごと、あるいは語ごとに例外的な基準を設けることで対応している（それは各時代の規程集に明文化される）。そのような構築上の背景に留意することが，コーパス利用の前提となることは言うまでもない。しかしながら，このように場合によって単位が異なるケースをみると，語構成要素の情報（以下，構成語情報とする）を付与することで，データが扱いやすくなる場合もあるように思う。また，公開済のコーパスをベースとして構成語情報を追加することによって，時代・資料・品詞など様々な条件下で，前項・後項のバリエーションや出現状況の調査が可能になることが期待される。日本語の歴史的な語形成を考察する手がかりとして有用な情報であるといえよう。本発表では，CHJ に出現する語を対象に，構成語情報を付与する方法と，その有効性を検討し，現状の課題についても報告する。

２．関連研究

古代日本語の形容詞・形容動詞の語構造に着目し，計量的な分析を行うものとして，村田菜穂子氏の一連の研究が挙げられる。古代語に加え，中～近世期の資料に関しても調査が進められており，語構造に関する情報を付した語彙表が順次報告されている。合成語（複合語・派生語）の変遷を知るうえでも，歴史資料に見える語の分析方針を検討するうえでも重要な資料であるといえる。それらの語彙表には，出典，用例数，活用型，語構造についての情報が付されるが，あくまで一覧化したものであり，各語が用いられる構文の環境や，用法などをただちに参照できるものではない。公開中のコーパスのテキストに対して，各語の構成語情報を付与することができれば，データの汎用性，および再現性はより高くなることが期待される。

また動詞に関しては，オンラインデータベース「複合動詞レキシコン[2]」（国立国語研究所，2015）が公開されている。複合語自体の検索のみならず，前項・後項それぞれに検索でき，『現代日本語書き言葉均衡コーパス（BCCWJ）[3]』の例文と関連づけるなど，利便性の高いデータ提供を行っている。ただし，このデータベースに収録される語彙は「現在の日本語で一般的に使われている複合動詞」（2,790 語）が対象であり，「古語・古典語」は対象外であることが明記されている。

さて，本研究では，コーパスのテキストそのものではなく，コーパスと関連付けられた電子化辞書「UniDic」に，別途，語構造に関する情報を持たせることを計画している。この手法は，淺尾（2017）で提案されており，フリーライセンスで提供している UniDic の内容をもとに 199,098 項目に対して語構成情報を付与する内容が示されている。その研究結果を反

[2] http://vvlexicon.ninjal.ac.jp/

[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/

映した検索ツールが試験公開されている[4]が，そのように品詞によらず，網羅的に日本語の語構成情報を付与したデータベースは他に類をみないものである。ただし，使用したデータのバージョンから，CHJ で追加された項目は含まれていないものと見られる。本研究では，以上の研究を参照しながら，歴史資料に出現する語に対する構成語情報の付与について検討を行う。

３．対象とするデータの概要

3.1 『日本語歴史コーパス（CHJ）』

本研究では，CHJ を対象として対象語の収集，分析，情報の付与を行う。表 1 に，CHJ に収録されているコーパスの一覧と延べ語数を示す（2018 年 7 月現在）[5]。

表 1　調査対象

サブコーパス	収録作品	延べ語数（短単位）
奈良時代編	Ⅰ万葉集	98,499
平安時代編	（仮名文学作品）	856,682
鎌倉時代編	Ⅰ説話・随筆／Ⅱ日記・紀行	821,010
室町時代編	Ⅰ狂言／Ⅱキリシタン資料	358,419
江戸時代編	Ⅰ洒落本	204,519
明治・大正編	Ⅰ雑誌	12523,750

CHJ は日本語の通時的研究のための基礎資料として整備が進められている。2018 年 3 月に，『室町時代編IIキリシタン資料』『江戸時代編I洒落本』のデータが加わり，部分的ではあるが，上代から近代まで一本化した資料を検索対象として扱えるようになった。これらのコーパスは，先述のとおり国語研の規定する斉一な単位（短単位）によってテキストを分割し，各単位に対して，形態論情報が付与されている。その形態素解析は，国語研が整備している電子化辞書 UniDic を利用して行われる。コーパス本文の短単位に付与された形態論情報と，UniDic に立項された見出し語の情報は，国語研の形態論情報データベースによって関連づけられている（小木曽・中村 2011）。したがって，UniDic の見出し語（語彙素）に対して，その構成語情報を付与することができれば，各語がどのコーパスにどのように出現するかについても容易に参照可能となる。

3.2 UniDic の見出し語について

UniDic の特長として、以下の 2 点が挙げられる（小椋ほか 2011）。

（ア）一語の認定基準がわかりやすく、判断の揺れが少ない「短単位」を見出し語として採用する。

（イ）表記や語形の違いに関わらず、同じ語であれば同一の見出しを与えるという方針をとり、語を階層化した形で登録している。階層構造の最上位を「語彙素」（辞書の見出しに相当）とし，語彙素＞語形（語形の違いを区別する層）＞書字形（表記の違いを区別する層）という階層を設ける。

「短単位」は，用例収集を目的とし，言語の形態論的側面に注目して規定された言語単位である。短単位の認定にあたっては，まず「最小単位」が規定される。その上で，文節

[4] http://asaokitan.net/jmorph/

[5]「語彙統計：バージョン 2017.3」（http://pj.ninjal.ac.jp/corpus_center/chj/201703.html）の数値に拠る。記号や空白はのぞく。万葉集，キリシタン資料，洒落本の語数は公表前のため，稿者の調査に拠る数値を示す。

の範囲内で短単位の規程に基づいて結合させる（又は結合させない）ことにより，短単位が認定される。この「最小単位」は、和語・漢語・外来語・記号・人名・地名の種類によって以下のように分類されている。その多くは、本研究で扱う語構成要素となりうることから、形態素境界や構成要素の判断をするにあたって、適宜参照する。

表 2　最小単位の分類（小椋ほか 2011、上 pp.7）

分類		例
一般		和語：豊か　大　雨 … 漢語：国語　研究　所 … 外来語：コール　センター　オレンジ …
数		一　二　十　百　千 …
その他	付属要素	接頭的要素：相　御　各 … 接尾的要素：兼ねる　がたい　的 …
	助詞・助動詞	だ　ます　か　から　て　の …
	人名・地名	星野　仙一　大阪　六甲 …
	記号	A　B　ω　イ　ロ　ア　JR …

4．作業方針と今後の課題

4.1　構成語情報の作成と UniDic との連携

本研究で目指す構成語情報付与のイメージを，図 1 に示した。構成語情報として，以下の情報を必須項目とする。

(ア) 語彙素 ID
(イ) 分類｛複合／派生／不明｝
(ウ) 連番（前項・後項等の位置を定めるもの）
(エ) 構成語_語彙素 ID（各構成要素の語彙素 ID）

UniDic に登録された語を一意に同定することのできる「語彙素 ID」の情報を利用することで，その語の内部構造に関する情報を入力する。図のように，各要素が UniDic に登録されている語であれば，その情報にリンクできるよう，語彙素 ID を入力する。語彙素 ID を利用して，UniDic の情報と対応づけることによって，UniDic に紐付いたコーパスの用例も適宜参照することが可能になる。

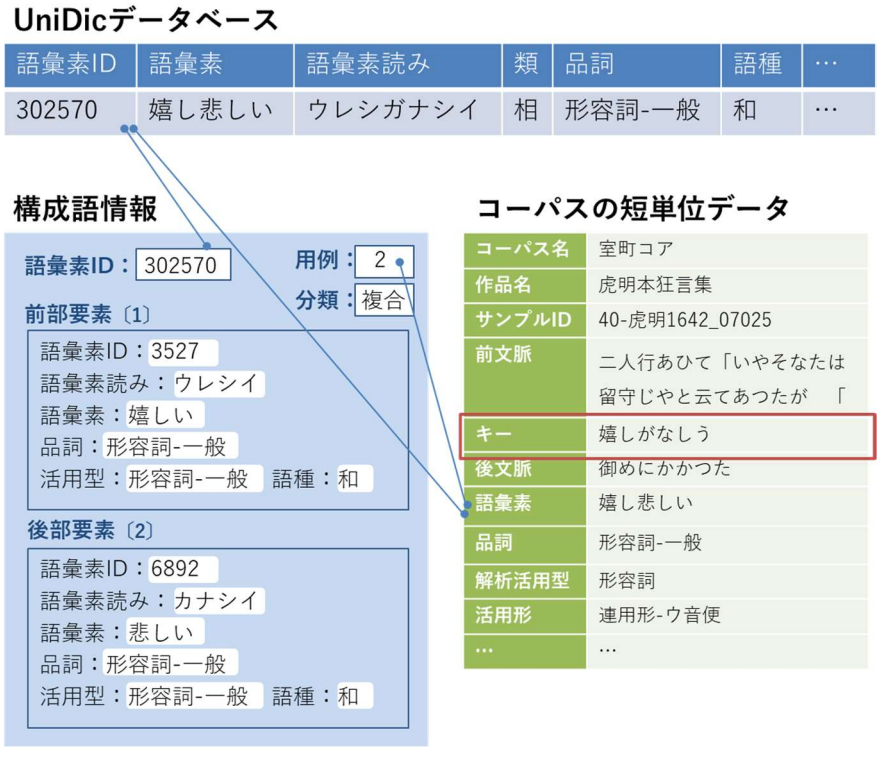


図 1　　UniDic・コーパスデータと対応付けた構成語情報のイメージ

なお，UniDic の編集を行うためのツールである「UniDic Explorer」には，語彙素に付与可能な情報の一つとして，「構成語情報」の入力欄が用意されている[6]（短単位で複合語となる語に対し，それを構成する複数の語彙素 ID が入力可能）。この機能を生かして UniDic に直接入力することも検討したが，その場合，各構成要素が UniDic の見出し語として立項されていることが条件となる。そのため，短単位以下の語を構成要素として付したい場合，UniDic の規定にそぐわない形式を登録するか，あるいは空欄とすることになる。そこで本研究では，CHJ に出現する語を抽出したうえで，必要な情報を入力した別表を作成し，対応づけることとした。本研究で作成したデータのうち，反映できるものについては，UniDic に還元することも視野に入れている[7]。なお、現在，試行として CHJ に出現する形容詞を対象としてデータの入力を進めている（表 3）。その作業内容の一部を、表 4 に示す（この別表と UniDic を対応付けたデータベースに関しては当日のポスターで紹介する）。

表 3　CHJ 室町時代編・江戸時代編に出現する動詞・形容詞の数

		キリシタン	狂言	洒落本
動詞	延べ	21,730	47,121	32,082
	異なり	1,959	1,868	2,167
形容詞	延べ	2,145	5,325	5,156
	異なり	221	238	347

表 4　UniDic と対応付ける構成語情報の入力例

コーパス	語彙素ID	語彙素	語彙素読み	類	語種	最小単位	分類	連番	語彙素ID（構成要素）	類	語彙素	語彙素読み
江戸	268920	甲斐無い	カイナイ	相	和	カイ/ナイ	複合	1	5615	体	甲斐	カイ
江戸	268920	甲斐無い	カイナイ	相	和	カイ/ナイ	複合	2	27442	相	無い	ナイ
江戸	290140	掛かりがましい	カカリガマシイ	相	和	カカリ/ガマシイ	派生	1	6016	用	掛かる	カカル
江戸	290140	掛かりがましい	カカリガマシイ	相	和	カカリ/ガマシイ	派生	2	8140	接尾	がましい	ガマシイ
江戸	93415	限り無い	カギリナイ	相	和	カギリ/ナイ	複合	1	6117	体	限り	カギリ
江戸	93415	限り無い	カギリナイ	相	和	カギリ/ナイ	複合	2	27442	相	無い	ナイ
江戸	6631	堅苦しい	カタクルシイ	相	和	カタ/クルシイ	派生	1	6603	固い	カタイ	カタイ
江戸	6631	堅苦しい	カタクルシイ	相	和	カタ/クルシイ	派生	2	保留	–	–	–
江戸	8361	聞き苦しい	キキグルシイ	相	和	キキ/グルシイ	複合	1	8399	用	聞く	キク
江戸	8361	聞き苦しい	キキグルシイ	相	和	キキ/グルシイ	複合	2	10523	相	苦しい	クルシイ

4.2　作業上の課題

現在，上記のように抽出した語彙に対して分類・情報付与の作業を進めているが，その過程で課題となった事例についていくつか紹介する。

語の認定に揺れが生じる例として，「モノ-」（名詞「物」から）、「ウス（ウソ）-」（形容詞「薄い」の語幹から）のように，一概にその品詞を確定できないものが存する。

(3)　a. もの おもはしき その人は嶋の内なる大見やに。名高き若づめ大角とて
【出典】CHJ サンプル ID：52-洒落 1826_01026『色深狭睡夢』

b. うそ 甘い（vfoamai）物を食らうた上なれば，何かは良からう
【出典】CHJ サンプル ID：40-天伊 1593_00002『天草版伊曽保物語』

また，個別の例ではあるが，語源俗解などにより，発生時とその後の時代とで，異なる語として認識されるものも存する。例として、現代語における「形容詞語幹＋クルシイ」は，中

[6] 複合動詞について試験的に入力されているものもあるが、基本的に未入力の状態であり、現時点では公開対象にない情報である。

[7] 語彙素 ID を使用し、連番によって要素の位置を定める考え方はこの「構成語情報」に拠っている。データを還元する可能性があることから、共通の仕様となるよう調整した。ただし、UniDic Explorer では複合語を対象としているため、複合・派生等の分類は行なわれない。

世～近世にかけては「クロシイ」という接尾辞による派生語であったが、近世後期以降、「～苦しい」と認識されるようになった形式である（村山 2012）。このような場合、共時的には接尾辞による派生語とみなしたいが、近現代の例では「苦しい」とするのが望ましい。

(4) a. なんじやいなかたくろしいその羽織もぬぎなんせんか
【出典】CHJ サンプル ID：52-洒落 1757_01005『聖遊廓』
b. 『まあ、そんな固苦しいことを言はないだつて、いゝでせう。』
【出典】CHJ サンプル ID：60M 婦倶 1925_12020『女人群像』

(5)の例はやや特殊なパターンではあるが、(4)の場合は、複合動詞の後項についても同様に、動詞と見るか接尾辞と見るか、判断が割れるものが現れうる。これらの形式については，先行研究も参照しながらなるべくその時代に即した分類を心がけたいものの，時代や語によって分類を変えるよりも，同形式として収集できるほうが結果として望ましいものと考える。したがって，まずは一定の基準にしたがって分類しておき，構成語情報には特記事項を加えるか，または注意が必要な形式としてリストアップすることで対処に代えたい。

５．おわりに

本稿では、『日本語歴史コーパス』に出現する語に対し、語構成要素に関する情報を付与する試みについて紹介した。淺尾（2017）でも述べられているが、やはり最大の課題は、語構造をどのように認定するかということになる。特に歴史的な資料を扱う上では，その語がどのように発生したかを明らかにすることには限界がある。一方で現代語の視点から，いかなる要素の結合であるかを分析することには，余計な解釈を付け足してしまう恐 れもあるといえよう[8]。とりわけ派生語については，何を接辞とみなすか，明確な基準が必要である。実作業を通して得られた課題について検討しつつ，一定の基準を策定し，複合・派生語の研究に資するデータとなるよう引き続き整備を進めていく。

謝　辞

本研究は JSPS 科研費 17K13471「派生・複合情報を付与した歴史コーパスによる語形成の歴史的研究」による成果の一部である。また，データベースの関連付けにあたって，中村壮範氏（国立国語研究所，コーパス開発センター）にご協力いただいた。記して感謝申し上げる。

文　献

淺尾仁彦（2017）.「日本語語構成情報データベースの構築」『言語資源活用ワークショップ 2016 発表論文集』，pp.120-125

池上尚（2016）.「中古語複合形容詞［名詞＋評価形容詞］の一語性」『国語語彙史の研究』35，pp.39-55

小木曽智信・中村壮範（2011）．『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版』国立国語研究所（特定領域研究「日本語コーパス」平成 22 年度研究成果報告書）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011). 『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下)』，特定領域研究「日本語コーパス」

[8] 語の認定には認識の差も関わる。前川・村田（2015）では、コーパスを使った語彙研究の有用性を示すとともに、品詞性や複合度については作成者や使用者の間で差が生じることが指摘されている。

平成 22 年度研究成果報告書 http://pj.ninjal.ac.jp/corpus center/bccwj/doc/report/JC-D-10-05-02.pdf.
国立国語研究所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』形態論情報規定集,　http://pj.ninjal.ac.jp/corpus_center/chj/doc/morph-heian-2016.pdf よりダウンロード可能
国立国語研究所コーパス開発センター（鴻野知暁）編（2017）.『日本語歴史コーパス　鎌倉時代編』短単位規定集 Ver.1.0, http://pj.ninjal.ac.jp/corpus_center/chj/doc/morph_kamakura_v1_0.pdf よりダウンロード可能
国立国語研究所(編) (2018).『日本語歴史コーパス』, (バージョン 2018.03,中納言バージョン 2.2.1) https://chunagon.ninjal.ac.jp/
前川武・村田菜穂子（2015）.「索引とコーパスを利用した形容詞語彙の採取について」『国語語彙史の研究』34 , pp.227-241
村田菜穂子（2005）.『形容詞・形容動詞の語彙論的研究』, 和泉書院.
村田菜穂子（2015）.「中古形容詞に見られる複合的方式についての一考察」『国語語彙史の研究』34 , pp.91-109
村山実和子（2012）.「接尾辞クロシイ考」『日本語の研究』8-4 , pp.16-30

関連 URL

コーパス検索アプリケーション『中納言』	https://chunagon.ninjal.ac.jp/
『日本語歴史コーパス』	http://pj.ninjal.ac.jp/corpus_center/chj/
『UniDic』	http://unidic.ninjal.ac.jp/

Twitter で使われる「深い」の意味 ―「強い」「すごい」と比較して―

加藤恵梨（大手前大学現代社会学部）†
山下紗苗（明石工業高等専門学校）
上泰（明石工業高等専門学校）

The Meaning of fukai Used on Twitter: Compared with tsuyoi and sugoi

Eri Kato (Otemae University)
Sanae Yamashita (National Institute of Technology, Akashi College)
Yasushi Kami (National Institute of Technology, Akashi College)

要旨

近年、Twitter で「うれしみ」などの感情を表す「み名詞」が多用されており、それらは「深い」や「強い」などの形容詞と共起することが多い。本研究では、Twitter で使われる「感情を表す『み名詞』＋が深い」がどのような状況で使われ、どのような意味を表すのかについて分析する。また、類義表現である「感情を表す『み名詞』＋が強い」「『み名詞』＋がすごい」が表す意味についても分析し、それらの意味の違いを明らかにする。分析の結果、「感情を表す『み名詞』＋が深い」は「ある感情の積み重なりにより、その感情の程度が高くなっているさま」と「感情の程度が普通の程度を超えているさま」という二つの意味を表すことを述べた。

１．はじめに

形容詞などの語幹に付いて名詞に変える働きを持つ接尾辞に「－さ」と「－み」がある。「－み」は「対象から把握される主観的な状態や、感情・感覚を、総体的・全一体的な状態概念として表す」という機能を持つものである(森田 1989)。「－み」は「－さ」に比べて生産性が低く、「－み」の付き得る形容詞は限られていると述べられてきた[1]。しかし、現在インターネット上で投稿されている文章において、従来「－み」が接続しなかった形容詞、多くの名詞、動詞連用形に「－み」が接続する現象が起きている(宇野 2015)。また、それらの新しい「み名詞」は「ある」「を感じる」や「が深い」といった表現と共起する割合が高いことが指摘されている(水野 2017)。

本稿は、インターネット上のコミュニケーションツールである Twitter において、感情を表す「み名詞」が「が深い」と共に使われた場合、どのような意味を表すのかについて分析する。また、類義表現である「感情を表す『み名詞』が強い」「感情を表す『み名詞』がすごい」についても分析し、それぞれがどのような意味を表すのかについて明らかにする。

２．「深い」について

２．１　先行研究の記述

次元形容詞「深い」がどのような意味を表すのかを確認すると、國廣(1970)は「深い」は

† erikato@otemae.ac.jp

[1] 森田(1989)は「－み」の付き得る形容詞は「赤み、明るみ、暖かみ、厚み、甘み、ありがたみ、痛み、うまみ、おかしみ、重み、おもしろみ、辛み、悲しみ、臭み、苦しみ、渋み、酸っぱみ、すごみ、楽しみ、強み、懐かしみ、苦み、憎しみ、深み、丸み、柔らかみ、弱み」といった限られた語だけであると述べている。

「基準面から物の内部に向かって入りこんで行く隔たりが標準値より大きい」という意味を表すと述べている。

また、「深い」が使われる意味領域について、西尾 (1972) は、視覚的なもの（「闇」「森」「色」）、時（「秋」「冬」）、関係（「仲」「交際」）、感情（「愛」「悲しみ」）、知的作用（「理解」「意味」）、その他（「罪」「混乱」）という 6 つを挙げている。

さらに、本稿が対象とする「深い」が「感情」の領域に使われた場合についての記述を見ると、小出(2000)は「深い愛」「深い悲しみ」などの表現について「この種の感情は『心』の中にあって、『心』の深い所にあるものほど度合いが強いという認識があるからであろう」と述べている。

２．２　先行研究の記述の検討

西尾は「深い」が使われる意味領域を 6 つ挙げている。1 節で述べたように、現在、インターネット上で投稿されている文章において、「深い」は「み名詞」という感情の領域の表現と使われることが多い。「深い」は以前から感情領域の表現と共に使われることが多かったのろうか。その点について『現代日本語書き言葉均衡コーパス(BCCWJ)』で確認する必要がある。

また、小出は「深い」が感情表現と共に使われた場合、「度合いが強い」ということを表すと述べている。そうであるならば、「強い」や「すごい」が感情表現と共に使われた場合も「深い」と同じような意味を表すと考えられる。では、それらにはどのような意味の違いがあるのだろうか。それらの違いについて明らかにする必要がある。

３．BCCWJ における「～が深い」「～が強い」「～がすごい」に前接する表現

「～が深い」及び「～が深い」の類義表現である「～が強い」「～がすごい」にどのような語が前接するのかを『現代日本語書き言葉均衡コーパス(BCCWJ)』で調べると、表 1 から表 3 のような結果が得られた。

表 1　「～が深い」に前接する表現（全 890 例）

	前接語	出現数
1	関係	66
2	奥	53
3	造詣	26
4	関わり	20
5	雪	17
6	根	16
7	縁	13
7	彫り	13
9	水深	9
9	欲	9

表 1 から分かるように、「～が深い」は「関係」「関わり」「縁」が前接する割合が高く、西尾の言う「関係」の表現と良く使われている。また、「～が深い」に前接する「み名詞」は、「悲しみ」が 4 例、「親しみ」が 1 例のみであった。このことから、BCCWJ において「～が深い」は感情表現が前接する割合が低いと言える。

表 2　「～が強い」に前接する表現（全 6435 例）

	前接語	出現数
1	傾向	317
2	風	277
3	ほう	222
4	印象	118
5	イメージ	107
6	気持	105
7	力	104
8	可能性	100
9	思い	87
10	気	78

表 2 から分かるように、「～が強い」は「傾向」や「ほう」といった抽象的な表現が前接し、ある物事から感じられることの程度について述べるときに使われている。また、「～が強い」に前接する「み名詞」について見ると、「痛み」が 21 例、「甘み」が 10 例、「青み」「かゆみ」が 4 例、「憎しみ」「渋み」「えぐみ」が 2 例、「恨み」「苦み」が 1 例であった。このことから、「『み名詞』＋が強い」は五感によって感じ取ったことの程度について述べるときに多く使われていると言うことができる。

表 3　「～がすごい」に前接する表現（全 788 例）

	前接語	出現数
1	の	64
2	ところ	23
3	ほう	13
3	何	13
5	これ	11
6	音	9
7	それ	7
7	こと	7
9	自分	5
10	た	4

表 3 から、「～がすごい」は特定の事物やある部分を表す表現が前接し、それらの程度について述べるときに多く使われていると言うことができる。また、「～がすごい」に前接する「み名詞」は見られなかった。

以上のように、BCCWJ では「～が深い」だけではなく、「～が強い」「～がすごい」においても感情表現及び「み名詞」と共に使われている割合が低い。このことから、「～が深い」は以前から感情領域の表現と共に使われることが多かったわけではないことが分かる。

４．調査方法

Twitter において「感情を表す『み名詞』＋が深い」及び「感情を表す『み名詞』＋が強い」「感情を表す『み名詞』＋がすごい」がどのような意味を表すかを分析するにあたり、次のような調査方法を用いた。

まず、Twitter の検索機能を利用して「かなしみ」「くるしみ」「つらみ」「うれしみ」「たの

しみ」に関する使用例を抽出した。次に、それらの「み名詞」と「深い」「強い」「すごい」の共起例を収集するため、それらの「み名詞」に「が深い」「が強い」「がすごい」が連接しているツイートのみを収集した。なお、ツイートの収集には Twitter API を使用しており、文字単位での完全一致に該当するもののみが抽出されている。

対象期間は 2018 年 7 月 9 日～18 日（Twitter の仕様による）、合計件数は 1281 であり、その内訳はかなしみ 98、くるしみ 1、つらみ 768、うれしみ 400、たのしみ 13 である。

５．Twitter における「『み名詞』が深い」の意味

はじめに「『み名詞』が深い」について分析する。

５.１　意味１：ある感情の積み重なりにより、その感情の程度が高くなっているさま

(1) 最近変わったもの食べると必ずお腹壊すようになってかなしみが深い　（2018-07-15）
(2) 最近ゆんたろさんと会えてないのかなしみが深い（2018-07-10）
(3) 毎日あついしセミは初めて見たしつらみが深いわ（2018-07-18）

まず (1)は、変わったものを食べると必ずお腹を壊すことで、かなしみを感じるような経験が重なり、かなしみを感じる程度が高くなっているさまを「かなしみが深い」と表している。続いて(2)は、最近ある人（ゆんたろさん）に会えないことで、かなしみの感情の程度が高くなっているさまを「かなしみが深い」と表している。さらに(3)は、毎日暑いといった日々の夏の厳しい暑さによって、つらいという感情の程度が高くなっているさまを「つらみが深い」と表している。

以上から、「『み名詞』が深い」の一つ目の意味は、「ある感情の積み重なりにより、その感情の程度が高くなっているさま」と記述することができる。

５.２　意味２：感情の程度が普通の程度を超えているさま

(4) 今日学校でブレーカー落ちてクーラーなしの中午前部活やってて 2、3 年は午後部活なしなのに 1 年は部活ありで途中からブレーカーが回復してクーラーついたはいいもののあまりに理不尽すぎるつらみが深い（2018-07-17）
(5) ふぉろーしてる配信者さんが好きすぎてつらみが深いの極み（2018-07-18）
(6) つらみが深い　好きな人に嫌われるってこの世で１番辛い（2018-07-18）
(7) 就活も何とか納得のいく終わりかたできて、さらにやっと地獄の日々から解放されてうれしみが深いです（2018-07-12）

(4)は「あまりに理不尽すぎる」ことに「つらみが深い」と感じているように、つらいと感じる程度が非常に高いさまを「つらみが深い」と表している。続いて(5)は、「好きすぎてつらみが深い」とあるように、ある人が好きすぎてつらいという感情が増し、つらいと感じる程度が普通の程度を超えているさまを「つらみが深い」と表している。また(6)は、好きな人に嫌われ、「この世で１番辛い」という思いを「つらみが深い」と述べている。さらに(7)は、「やっと地獄の日々から解放されて」とあるように、非常にうれしいことを「うれしみが深い」と述べている。

以上から、「『み名詞』が深い」の二つ目の意味は、「感情の程度が普通の程度を超えているさま」と記述することができる。

６．Twitter における「『み名詞』が強い」の意味

続いて「『み名詞』が強い」について分析する。

６．１　意味１：予想・期待と現実が異なり、ある感情の衝撃の大きさを感じているさま

(8) うあー、遊びの約束かな？楽しみにしてた分かなしみが強いよね...！（2018-07-14）
(9) 色々な方がネップリしてて良いな！思いつつ私の地域にセブン無くてかなしみが強い（2018-07-09）
(10) 職場の方のお嬢さんから可愛い手作りカードホルダーをいただいた。うれしみが強いのでお返しにラデュレ買った（2018-07-11）

まず(8)は「楽しみにしてた分かなしみが強い」とあるように、期待と現実が異なることで感じる、かなしみの衝撃の大きさを「かなしみが強い」と表現している。続いて(9)は、自分の住む地域にセブンイレブンがあったらいいと思っているが、現実には無いというように、期待と現実が異なることで、かなしみの衝撃の大きさを「かなしみが強い」と述べている。さらに(10)は、予想外にある人からプレゼントをもらったことで感じたうれしいという感情の大きさを「うれしみが強い」と表現している。

以上から、「『み名詞』が強い」の一つ目の意味は、「予想・期待と現実が異なり、ある感情の衝撃の大きさを感じているさま」と記述することができる。

６．２　意味２：ある感情で胸が締め付けられるように感じるさま

(11) 脚が短いせいで喧嘩に勝てないマンチカン、かわいいよりもかなしみがつよい　胸がギュってなる（2018-07-11）
(12) やっぱりでぃーさんの兼さん最高。泣くほど可愛いムリ。堀川くんも可愛いつらみが強い（2018-07-17）
(13) 田中圭誕生祭 2018　かわいみのある、はるたんありがとうございます！久々の公式の更新うれしみが強い（2018-07-10）

まず(11)は「胸がギュってなる」とあるように、かなしみの感情で胸が締め付けられるように感じるさまを「かなしみがつよい」と表している。続いて(12)はある人について「泣くほど可愛い」と感じ、それによりつらいという感情で胸が締め付けられるように感じるさまを「つらみが強い」と表している。さらに(13)は、久々の公式更新が非常にうれしいさまを「うれしみが強い」と述べている。

以上から、「『み名詞』が強い」の二つ目の意味は、「ある感情で胸が締め付けられるように感じるさま」と記述することができる。

７．Twitter における「『み名詞』がすごい」の意味：何らかの衝撃を受け、ある感情で気持ちが満たされているさま

最後に、「『み名詞』がすごい」について分析する。

(14) しくしく、自分も全部落ちた～(´；ω；｀)かなしみがすごい（2018-07-14）
(15) お腹が痛すぎてかなしみがすごい（2018-07-11）
(16) いまうちに CD を聴ける媒体がないことに気づいてかなしみがすごい（2018-07-11）
(17) まいふぁーざーがアコギ買ってくれるうれしみがすごい（2018-07-14）

まず(14)は、全部試験に落ち、「かなしみ」の感情で気持ちが満たされているさまを「かな

しみがすごい」と表している。続いて(15)は、お腹が痛すぎることで「かなしみがすごい」と述べている。また(16)は、家に CD を聴ける媒体がないことに気づいたことで、「かなしみがすごい」と述べている。さらに(17)は父親がアコギを買ってくれることになり、「うれしみがすごい」と表現している。このように、(14)から(17)は、何らかの衝撃を受け、発話時にある感情で気持ちが満たされているさまを表している。

以上から、「『み名詞』がすごい」の意味は、「何らかの衝撃を受け、ある感情で気持ちが満たされているさま」と記述することができる。

８．おわりに

本稿では、インターネット上のコミュニケーションツールである Twitter において、感情を表す「み名詞」が「が深い」と共に使われた場合、どのような意味を表すのかについて分析した。また、類義表現である「感情を表す『み名詞』が強い」「感情を表す『み名詞』がすごい」についても分析し、それぞれがどのような意味を表すのかについて明らかにした。今回の分析では、感情を表す「み名詞」を「かなしみ」「くるしみ」「つらみ」「うれしみ」「たのしみ」に限定したが、今後はその他の「み名詞」についても調査・分析を行いたいと考えている。

文　献

宇野和(2015).「Twitter における『新しいミ形』」『国文』123, pp.106-94.
國廣哲彌 (1970).「日本語次元形容詞の体系」『言語の科学』2, pp.13-26.
小出慶一 (2000). 「次元形容詞の空間的用法と非空間的用法」『群馬県立女子大学紀要』21, pp.1-13.
西尾寅弥 (1972). 『形容詞の意味・用法の記述的研究』秀英出版
水野みのり (2017). 「ネット集団語における接尾辞『－み』の語基拡張」『思言：東京外国語大学記述言語学論集』第 13 号, pp.167-174.
森田良行 (1989). 『基礎日本語辞典』角川学芸出版

関連 URL

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

日本語の二重目的語構文の基本語順について

浅原 正幸 (国立国語研究所) *
南部 智史 (モナシュ大学)
佐野 真一郎 (慶應義塾大学)

Predicting Japanese Word Order in Double Object Constructions

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Satoshi Nambu (Monash University)
Shin-Ichiro Sano (Keio University)

要旨

本稿では日本語の二重目的語構文の基本語順について予測する統計モデルについて議論する。『現代日本語書き言葉均衡コーパス』コアデータに係り受け構造・述語項構造・共参照情報を悉皆付与したデータから、二重目的語構文を抽出し、格要素と述語要素に分類語彙表番号を付与したうえで、ベイジアン線形混合モデルにより分析を行った。結果、名詞句の情報構造の効果として知られている旧情報が新情報よりも先行する現象と、モーラ数が多いものが少ないものに先行する現象が確認された。分類語彙表番号による効果は、今回の分析では確認されなかった。

1. はじめに

日本語は語順が自由な言語である。日本語の語順に影響を与える影響について、主に計算言語学の分野で調査されてきた (Yamashita and Kondo 2011, Orita 2017)。一つの良く知られている知見 'long-before-short' (Yamashita and Chang 2001) として、かき混ぜにより長い名詞句が短い名詞句よりも前に傾向がある。本稿では、そのなかで日本語の二重目的語構文に注目する。二重目的語構文でニ格名詞句 (1) とヲ格名詞句 (2) のどちらが先行するのかを検討する：

(1) 太郎が 花子に 本を あげた

(2) 太郎が 本を 花子に あげた

日本語においてはどちらの語順も可能であるため、理論言語学においては何が正規語順かについて、ある語順は他の語順から派生されているという仮説に基づいて議論されてきた (Hoji 1985, Miyagawa 1997, Matsuoka 2003)。本稿では、単にコーパス中の二重目的語構文の頻度

* masayu-a@ninjal.ac.jp

表 1　先行研究との比較

	(Sasano and Okumura 2016)	(Orita 2017)	本研究
コーパス	ウェブコーパス	NAIST テキストコーパス	BCCWJ-PAS BCCWJ-DepPara
ジャンル	ウェブ	新聞記事	新聞記事, 書籍, 雑誌, 白書 ブログ, Q/A サイト
対象	ガ-ニ-ヲ-述語	ガ-ヲ-述語	ガ-ニ-ヲ-述語
文書数	n/a	2,929	1,980
文数	100 億語	38,384	57,225
文型数	648 types × 350,000 samples	3,103 tokens	584 tokens
分析対象	verb types	syntactic priming, NP length, given-new, and animacy	NP length, and given-new
分析方法	線形回帰・ NPMI	ロジスティック回帰 (glm)	ベイジアン線形混合モデル (rstan)

を数えることにより正規語順について議論するのではなく、理論言語学などの先行研究で言及されている様々な要因を考慮した統計分析を行う。この目的のために、語順に影響を与える要因を考慮したベイジアン線形混合モデルを用いて分析を行う。

'long-before-short' 以外の要因として、文脈中の名詞句の情報状態が語順に影響を与えるという理論的な枠組がある (Lambrecht 1994, Vallduví and Engdahl 1996)。この枠組では、ある名詞句が文脈中で情報の状態としてどのような機能をもつか、情報の新旧・トピック（話題）・フォーカス（焦点）などの観点を導入する。この情報の状態が語順を決める基本的な要因の一つであるということを、次の二つの理由に基づいて仮定する：(1) 日本語の文中、談話に既出の要素は、談話の未出の要素に先行する (Kuno 1978, 2004, Nakagawa 2016)，(2) 日本語の文中、フォーカス（焦点）もしくは新情報は述語の直前に出現する傾向にある (Kuno 1978, Kim 1988, Ishihara 2001, Vermeulen 2012)。一般的な日本語の語順に関するこれらの 2 つの主張に基づいて、二重目的語構文について以下のような仮説を検討する。

(3)　仮説:
二重目的語構文において、談話に既出の要素は他の要素より左に位置する傾向にあり、談話に未出の要素は他の要素より右に位置する傾向にある。

先行研究で提案されている 'long-before-short' 関連の要因と名詞句の情報状態を考慮したうえで、日本語二重目的語構文の語順を推測する統計モデルを構築する。本研究の新規性として、述語項構造と共参照が人手でアノテーションされたものを使うことで、代名詞（既出）か否か（未出）のような二分的な分析でではなく、先行文脈に名詞句の指示対象が出現している回数を含めて考慮する点があげられる。

2.　先行研究

表 1 は、コーパスに基づく日本語の語順の分析に関する直近の先行研究について示す。

Sasano and Okumura (2016) は日本語の二重目的語構文の単文レベルの正規語順について、大規模なウェブコーパスを用いて「ガ-ニ-ヲ-述語」と「ガ-ヲ-ニ-述語」のどちらが優勢かにつ

いて調査した。形態素解析器 JUMAN と係り受け・格解析器 KNP により自動解析した 100 億文を準備し、統語的な曖昧性がない部分木を取り出して分析資料を準備した。この資料に対し、動詞タイプ (SHOW 型 or PASS 型) を要因とした語順の分析を線形回帰と相互情報量 (normalized pointwise mutual information) に基づいて分析した。彼らのモデルでは共参照のような文間の関係を調査していない。

Orita (2017) は、述語項構造と共参照情報が人手で付与された NAIST テキストコーパスのアノテーションを用いて、直接目的語と主語の語順のかき混ぜを予測する統計モデルを構築した。彼女の調査では、プライミング・名詞句の長さ・有生性・情報状態（既出／未出）の効果を調査した。頻度主義的な統計分析（単純なロジスティック回帰）では、主語と目的語の語順について、情報状態に関する効果は観察されなかった。

本研究では、共参照情報が二重目的語構文の語順に影響を与える要因であると想定して、人手による係り受け・述語項構造・共参照を重ね合わせたデータを用いた、ベイジアン線形混合モデルに基づく調査を実施する。

3. 実験

3.1 データ：BCCWJ-PAS

本研究では、『現代日本語書き言葉均衡コーパス』(以下、BCCWJ) (Maekawa et al. 2014) に対する述語項構造と共参照のアノテーション BCCWJ-PAS を研究資料として用いる。このアノテーション基準は NAIST テキストコーパス (Iida et al. 2007) のものに準じる。このデータに、文節係り受けアノテーション BCCWJ-DepPara (Asahara and Matsumoto 2016) を重ね合わせることで、直接係り受け関係がある 主語（ガ）・直接目的語（ヲ）・間接目的語（ニ）・述語の 4 つ組を抽出する。ゼロ照応や格交替の事例を排除した結果、57,225 文から 584 例の 4 つ組を抽出した。

図 1 に Yahoo! 知恵袋サンプル (OC09_04653) の例を示す。表層文字列は、我々の語順を評価する際の相対距離の基本単位である文節単位に区切られている。相対距離は 4 つ組のなかから以下の 6 対を評価する：ガ-述語 ($\mathrm{dist}^{subj}_{pred}$), ヲ-述語 ($\mathrm{dist}^{dobj}_{pred}$), ニ-述語 ($\mathrm{dist}^{iobj}_{pred}$), ガ-ニ ($\mathrm{dist}^{subj}_{iobj}$), ガ-ヲ ($\mathrm{dist}^{subj}_{dobj}$), ニ-ヲ ($\mathrm{dist}^{iobj}_{dobj}$)。距離は対の左要素と右要素の文節に基づく距離を評価する。例えば、図 1 において、$\mathrm{dist}^{subj}_{pred}$ は「彼女が」と「使います」の距離を 4 とする。

日本語の語順の傾向として ‘long-before-short’ の効果を調査するために、項名詞句の長さを統計モデルの固定効果として導入する。項名詞句の長さは、BCCWJ-PAS でラベルづけされている項名詞句の右端を最右要素とし、係り受け木上で項名詞句を根とした場合の部分木の左端を最左要素とし、この 2 要素間の発音形のモーラ数に基づきガ格モーラ数 (N^{subj}_{mora}), ヲ格モーラ数 (N^{dobj}_{mora}), ニ格モーラ数 (N^{iobj}_{mora}) を定義する。例えば、図 1 において N^{subj}_{mora} は「その彼女が（ソノカノジョガ）」のモーラ数 6 と定義する。なお、部分係り受け木の最大スパンに基づくために、項名詞句の長さは 2 文節以上に対して定義する場合もある。

さらに、given-new ordering を確認するために、先行文脈における共参照要素数を固定効果に入れる。BCCWJ-PAS のアノテーションから得られたガ格名詞句・ヲ格名詞句・ニ格名詞

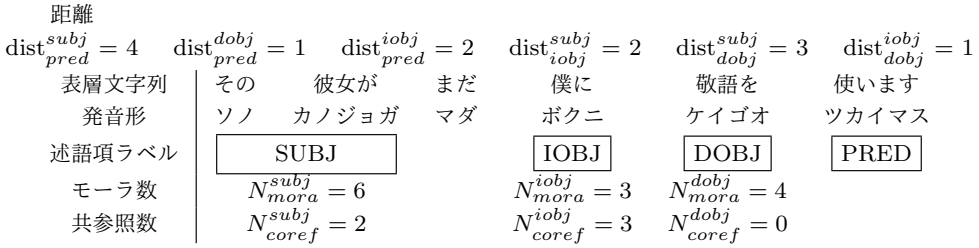


図 1　BCCWJ Yahoo! 知恵袋サンプルの例: (OC09_04653)

表 2　基礎統計

	min	1Q	med	mean	3Q	max
$\mathrm{dist}^{subj}_{pred}$	1.0	4.0	5.0	5.8	7.0	23.0
$\mathrm{dist}^{dobj}_{pred}$	1.0	1.0	1.0	1.7	2.0	13.0
$\mathrm{dist}^{iobj}_{pred}$	1.0	1.0	2.0	2.3	3.0	17.0
$\mathrm{dist}^{subj}_{iobj}$	-14.0	1.0	3.0	3.5	5.0	21.0
$\mathrm{dist}^{subj}_{dobj}$	-10.0	2.0	3.0	4.1	5.0	22.0
$\mathrm{dist}^{iobj}_{dobj}$	-12.0	-1.0	1.0	0.6	2.0	16.0
N^{subj}_{mora}	2.0	4.0	5.0	6.5	8.0	32.0
N^{dobj}_{mora}	2.0	3.0	4.0	5.3	6.0	37.0
N^{iobj}_{mora}	2.0	4.0	5.0	6.1	7.0	52.0
N^{subj}_{coref}	0.0	0.0	1.0	6.9	6.0	105.0
N^{dobj}_{coref}	0.0	0.0	0.0	0.5	0.0	44.0
N^{iobj}_{coref}	0.0	0.0	0.0	3.1	1.0	99.0

句の共参照数を $N^{subj}_{coref}, N^{dobj}_{coref}, N^{iobj}_{coref}$ として定義する。表 2 に、距離・モーラ数・共参照数の基礎統計量を示す。

3.2　統計処理

ベイジアン線形混合モデル (Sorensen et al. 2016) (BLMM) に基づき、2 つの項の間の距離もしくは項と述語の間の距離を評価する。具体的には、次の式に基づき統計モデルを作成する：

$$\begin{aligned}\mathrm{dist}^{left}_{right} &\sim \mathrm{Normal}(\mu,\sigma)\\ \mu &\leftarrow \alpha+\beta^{subj}_{mora}\cdot N^{subj}_{mora}+\beta^{subj}_{coref}\cdot N^{subj}_{coref}\\ &\quad +\beta^{dobj}_{mora}\cdot N^{dobj}_{mora}+\beta^{dobj}_{coref}\cdot N^{dobj}_{coref}\\ &\quad +\beta^{iobj}_{mora}\cdot N^{iobj}_{mora}+\beta^{iobj}_{coref}\cdot N^{iobj}_{coref}.\end{aligned}$$

ここで $\mathrm{dist}^{left}_{right}$ は left 要素と right 要素の文節を単位とした距離を意味する。例えば

$\text{dist}_{iobj}^{subj}$ は、主語 subj (left) と間接目的語 iobj(right) の間の文節に基づく距離 (隣接は 1) を表す。左右が反対の場合には負の値を持つ。これを、平均値 μ と標準偏差 σ の正規分布によりモデル化する。平均値 μ は切片 α と、2 タイプの変数の線形式により定義する。1 つ目のタイプの変数 $N_{mora}^{subj}, N_{mora}^{dobj}, N_{mora}^{iobj}$ は、主語・直接目的語・間接目的語のモーラ数に対するものである。それぞれ 2 文節以上の場合には、文節境界を越えてモーラ数を数える。2 つ目のタイプの変数 $N_{coref}^{subj}, N_{coref}^{dobj}, N_{coref}^{iobj}$ は、主語・直接目的語・間接目的語の共参照先行詞数に対するものである。β_b^a は、変数 N_b^a に対する傾きを表す。

これを rstan パッケージを用いて推定する。warmup 後のイテレーションを 2000 回に設定し、4 回シミュレーションを実施した。全てのモデルは収束した。

4. 結果と考察

4.1 結果

表 3 距離の推定結果

距離	α	β_{mora}^{subj}	β_{mora}^{dobj}	β_{mora}^{iobj}	β_{coref}^{subj}	β_{coref}^{dobj}	β_{coref}^{iobj}	σ
$\text{dist}_{pred}^{subj}$	4.814***	0.146***	-0.031	0.040	0.002	-0.056	-0.009	3.323
	(0.375)	(0.040)	(0.042)	(0.032)	(0.011)	(0.043)	(0.016)	(0.100)
$\text{dist}_{pred}^{dobj}$	1.593***	-0.009	0.061***	-0.032**	-0.001	0.037**	-0.005	1.072
	(0.128)	(0.013)	(0.014)	(0.011)	(0.004)	(0.014)	(0.005)	(0.032)
$\text{dist}_{pred}^{iobj}$	2.100**	-0.022	-0.056**	0.112***	-0.018***	-0.045	0.037***	1.861
	(0.217)	(0.022)	(0.023)	(0.018)	(0.006)	(0.024)	(0.009)	(0.055)
$\text{dist}_{iobj}^{subj}$	2.668***	0.171***	0.026	0.071**	0.020	-0.011	-0.046**	3.577
	(0.420)	(0.043)	(0.045)	(0.035)	(0.012)	(0.047)	(0.017)	(0.108)
$\text{dist}_{dobj}^{subj}$	3.205***	0.155***	-0.092**	0.072**	0.003	-0.094**	-0.004	3.452
	(0.404)	(0.041)	(0.043)	(0.034)	(0.012)	(0.046)	(0.017)	(0.103)
$\text{dist}_{dobj}^{iobj}$	0.502	-0.013	-0.117***	0.143***	-0.017**	-0.081**	0.041***	2.436
	(0.287)	(0.029)	(0.030)	(0.024)	(0.008)	(0.033)	(0.011)	(0.071)

** > ±2SD, *** > ±3SD

表 3 に BLMM により推定されたパラメータ値を示す。値は平均値と標準偏差（カッコ内）による。

まず、主語と述語間の距離 ($\text{dist}_{pred}^{subj}$) は、主語のモーラ数にのみ影響を受ける。主語のモーラ数が多いほどその述語との距離が長くなる傾向が見られた。

直接目的語と述語間の距離 ($\text{dist}_{pred}^{dobj}$) は、直接目的語のモーラ数・共参照先行詞数と、間接目的語のモーラ数の影響を受ける。i) 直接目的語のモーラ数が多いほど、述語からの距離が遠くなる、ii) 直接目的語の共参照先行詞数が多いほど、述語からの距離が遠くなる、iii) 間接目的語のモーラ数が多いほど、直接目的語と述語との距離が近くなる傾向が見られた。

間接目的語と述語間の距離 ($\text{dist}_{pred}^{iobj}$) は、間接目的語のモーラ数・共参照先行詞数と直接目的語のモーラ数・共参照先行詞数の影響を受ける。i) 間接目的語のモーラ数が多いほど、述語からの距離が遠くなる、ii) 間接目的語の共参照先行詞数が多いほど、述語からの距離が遠くなる、iii) 直接目的語のモーラ数が多いほど、間接目的語と述語との距離が近くなる、iv) 直接目的語の共参照先行詞数が多いほど、間接目的語と述語との距離が近くなる傾向が見られた。

二つの項名詞句間の距離 ($\text{dist}_{iobj}^{subj}$, $\text{dist}_{dobj}^{subj}$, and $\text{dist}_{dobj}^{iobj}$) も項名詞句ー述語間の距離と同じ傾向がある。しかしながら、項名詞句のモーラ数は項の長さ（構成する文節数）と相関があるため、最左項名詞句と最右項名詞句距離（例えば、主語と直接目的語間）が、間にある項名詞句（間接目的語）のモーラ数の影響を受ける。

4.2 考察

結果より、二重目的語構文において、主語は直接目的語や間接目的語より先行する傾向がわかった。間接目的語は、直接目的語より先行するが、有意ではなかった (p=0.09)。

共参照先行詞数に対する推定された係数 ($\text{dist}_{pred}^{dobj}$ に対する N_{coref}^{dobj} や、$\text{dist}_{pred}^{iobj}$ に対する N_{coref}^{iobj}) をみると、直接目的語と間接目的語間の語順に対して 'given-new ordering' の仮説を支持していることがわかる。共参照先行詞の数が多い目的語は、述語からの距離が遠くなる傾向がみられる。

モーラ数に対する推定された係数 ($\text{dist}_{pred}^{subj}$ に対する N_{mora}^{subj} や、$\text{dist}_{pred}^{dobj}$ に対する N_{mora}^{dobj} や、$\text{dist}_{pred}^{iobj}$ に対する N_{mora}^{iobj}) をみると、二重目的語構文の全ての項名詞句が 'long-before-short' の仮説を支持していることがわかる。ある目的語と述語間の距離に対して、もう一つの目的語のモーラ数に対する推定された係数が負の値であること ($\text{dist}_{pred}^{iobj}$ に対する N_{mora}^{dobj} や $\text{dist}_{pred}^{dobj}$ に対する N_{mora}^{iobj}) から、長い目的語が他の目的語に先行する傾向が見られた。

5. おわりに

本稿ではベイジアン線形混合モデルを用いて述語項構造・共参照アノテーションデータを分析することにより日本語の二重目的語構文の語順について検討した。結果、直接目的語と間接目的語の間に 'given-new ordering' の傾向が見られることがわかった。さらに全ての項名詞句について、'long-before-short' の傾向が確認された。

今後、名詞句の有生性や述語動詞の分類が二重目的語構文の語順に与える影響について分析する。このために項名詞句と述語動詞に対する分類語彙表番号付与の作業を進めている Kokuritsukokugokenkyusho (1964)。

謝　辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP15K12888, JP17H00917, JP18H05521 によるものです。

文　献

Hiroko Yamashita, and Tadahisa Kondo (2011). "Linguistic Constraints and Long-before-short Tendency." *IEICE Technocal report (TL)*:TL2011-19, pp. 61–65.

Naho Orita (2017). "Predicting Japanese scrambling in the wild." *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, pp. 41–45. Valencia, Spain: Association for Computational Linguistics.

Hiroko Yamashita, and Franklin Chang (2001). ""Long Before Short" Preference in the Production of a Head-final Language." *Cognition*, 81:2, pp. B45–B55.

Hajime Hoji (1985). “Logical Form Constraints and Configurational Structures in Japanese.” Unpublished doctoral dissertation, University of Washington.

Shigeru Miyagawa (1997). “Against Optional Scrambling.” *Linguistic Inquiry*, 28, pp. 1–26.

Mikinari Matsuoka (2003). “Two Types of ditransitive constructions in Japanese.” *Journal of East Asian Linguistics*, 12, pp. 171–203.

Knud Lambrecht (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Vol. 71. *Cambridge Studies in Linguistics.*: Cambridge University Press.

Enric Vallduví, and Elisabet Engdahl (1996). “The linguistic realization of information packaging.” *Linguistics*, 34:3, pp. 459–520.

Susumu Kuno (1978). *Danwa no bunpoo [Grammar of discourse]*. Tokyo: Taishukan Shoten.

Susumu Kuno (2004). “Empathy and direct discourse perspectives.” Lawrence Horn, and Gregory Ward (Eds.), *The handbook of pragmatics.*: Oxford: Blackwell. pp. 315–343.

Natsuko Nakagawa (2016). “Information Structure in Spoken Japanese: Particles, word order, and intonation.” Unpublished doctoral dissertation, Kyoto University.

Alan Hyun-Oak Kim (1988). “Preverbal focusing and type XXIII languages.” Jessica Wirth Michael Hammond, Edith A. Moravcsik (Ed.), *Studies in syntactic typology.*: Amsterdam: John Benjamins. pp. 147–169.

Shin-ichiro Ishihara“Stress, focus, and scrambling in Japanese.” Ora Matushansky Elena Guerzoni (Ed.), *MITWPL 39.*: Cambridge, MA: MITWPL. pp. 142–175.

Reiko Vermeulen (2012). “The information structure of Japanese.” Renate Musan Manfred Krifka (Ed.), *The expression of information structure.*: Berlin: De Gruyter Mouton. pp. 187–216.

Ryohei Sasano, and Manabu Okumura (2016). “A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2244. Berlin, Germany: Association for Computational Linguistics.

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

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto (2007). “Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations.” *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139. Prague, Czech Republic: Association for Computational Linguistics.

Masayuki Asahara, and Yuji Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annota-

tion Treebank on the 'Balanced Corpus of Contemporary Written Japanese'." *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 49–58. Osaka, Japan: The COLING 2016 Organizing Committee.

Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth (2016). "Bayesian Linear Mixed Models using Stan: A tutorial for psychologists, linguists, and cognitive scientists." *Quantitative Methods for Psychology*, 12:3, pp. 175–200.

国立国語研究所 (編) (1964).『分類語彙表』　秀英出版.

比喩指標としての「感じる」−文法形式と比喩の関係−

菊地礼（中央大学大学院文学研究科）

“Kanjiru” As Marker of Metaphor

Relation between Grammar and Metaphor

Rei Kikuchi (Graduate School of Letters, Chuo University)

要旨

本発表は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（BCCWJ）を用いて収集した比喩表現データを分析・考察する。中村（1977）『比喩表現の理論と分類』によれば直喩の指標は 7 類 82 種 359 号と多岐にわたる。しかし，直喩の典型である「よう」以外の分析はなされていない。本発表ではコーパスを用いた網羅的な用例収集を行い，分析に耐える量を確保する。その一例を本発表は動詞「感じる」によって示す。「感じる」は「A ガ B ヲ」「A ヲ B ト」「A ヲ B デ」等の 10 の構文を作るが，「A ニ B ヲ」「A ヲ B ニ」等の 8 つの構文で比喩を表わすことが可能である。しかし、直喩と認定できる例はその中から限定される。これは「感じる」が比喩指標として機能することが例外的事例であることを意味する。モダリティ形式としての文法化が比喩指標には求められるが、「感じる」は特定の構文環境においてのみ不完全ながら文法化を果たし、比喩指標と同様の機能を得る。

1. 問題点と目的

直喩は、「A ハ B ノヨウニ C ダ」等の形式によって比喩を表現する。「ヨウニ」が比喩を明示（鍋島 2016)するものであるとして、その明示性をもって隠喩と区別される。その明示を担い、直喩を特徴付ける形式が比喩指標である。比喩指標の範囲を検討したものとして中村（1977）が挙げられる。中村（1977）は、動詞から接辞に至るまで 441 種（82 種 359 号）の語を比喩指標として認定した。しかし、比喩指標の典型である「ヨウ」（木下 2003, 小松原 2016）以外の指標については、後の研究において検討されることは少ない。比喩指標の認定基準が明確でないためである。

比喩指標の分析が進展しない主たる原因は用例数の不足である。中村（1977）でも文学作品 50 篇から収集しているが、比喩指標の典型例「ヨウ」は 50 篇 3875 例出現するのに対し、D-1 類の「感じる」は 3 作品 4 例しか収集されていない。本発表では、現代日本語書き言葉均衡コーパス（以下 BCCWJ）を用いた網羅的な直喩表現の収集によって、分析に耐える量を確保する。これにより、表現形式の類型化、比喩用法の記述を可能にする。

本発表は、「感じる」を対象とする。第一に、非比喩用法と比喩用法、それぞれの構文的な差異を明らかにする。比喩になり得る構文、なり得ない構文を選別する。第二に、「感じる」を比喩指標として用いている例を抽出し、「感じる」が比喩指標として機能する条件を明らかにする。

2. 用例収集

菊地ほか（2018）は分類語彙表番号を付与したBCCWJを用いて「感じる」を含む比喩文の収集を行ったものである。そこで得られた用例を本発表の分析に用いる。収集の概要は次のとおりである。

2.1. 比喩指標

中村（1977）の比喩指標要素をキーとして用いる。比喩指標要素は次の【表1】のようにDからSの7類82種359号に分類される。本発表ではその内D類1種に属する「感じる」を対象とする。

【表 1】:比喩指標要素の種類と数

類	種数	号数
D（動詞）	18	83
F（副詞）	19	86
J（助詞）	8	30
K（形容詞・助動詞）	13	49
M（名詞）	8	53
R（連体詞・接頭辞）	5	12
S（接尾辞）	11	46
合計	82	359

2.2. コード付与

また、用例の収集に先立って、これらの比喩指標要素に分類語彙表番号の付与を行った。「感じる」には「2.3001」が与えられる。コード付与によって、中村（1977）に登録された比喩指標要素と同じ意味に分類される語もキーに含めることが可能となる。

2.3. 対象資料

用例収集対象には、分類語彙表番号を付与したBCCWJ （Maekawa et. al. 2014）の約27万語（2017年10月時点）を用いた。これらのデータから、語彙素「感じる」を含む用例と、「感じる」と同じ分類語彙表番号（.3001）を有する短単位・長単位（「感じ」「感覚」「実感」など）を含む用例を検索・収集した。この結果、比喩指標要素＜感じる＞を含む指標比喩用例の候補として、「感じる」289例、同意味の短単位424例、長単位78例を得た。あわせて指標比喩候補791例である。

2.4. 比喩判定

この 791 例について用例の比喩指標要素の前後文脈を確認し、指標比喩であるか判定を行った。＜感じる＞指標を含む比喩用例は31例であった。　うち、「感じる」は21例、「感じる」と同じ分類語彙表番号を有する用例は「感じ」5例、「感覚」3例、「実感」1例、「感ずる」1例という分布となった。中村（1977）における4例を超えるものであり、出現する構文の記述の妥当性を高めることができる。このように出現率は高いとはいえないが、コーパスを用いることにより従来の手作業による収集以上の例を集めることができる。

本発表は格や構文という文法的側面に着目するために「感じる」21 例を用いる。

3. 分析

本発表は、比喩を次のように規定する。第一は、二つの事物・事柄が事実否定的な関係にあることであり、第二にその結び付きによりイメージが具体的になる場合である。「感じる」の場合、前項に来る語が通常感覚することのできないものであり、喩えられる事物にイメージが付与された例を比喩用法として認定した。「寂しさ」のような感情や「痛み」のように実際に感覚可能なものを前項に持つ例は非比喩用法とした。

3.1. 「感じる」の構文と意味

非比喩用法における「感じる」の用例を格関係に基づいて整理すると、次の 10 構文にまとめられる。

【表 2】: 非比喩用法における「感じる」の取る構文[1]

	類型	用例数[2]	類型例
①	＜主体＞ガ＜対象＞ヲ	120	３０代で独身の男の人ガ寂しさヲ感じる
②	＜主体＞ガ＜対象＞ト	43	全ての人ガ「理想の病院」ト感じる
③	＜対象＞ニ＜思考＞ヲ	36	結婚ニ「精神的なやすらぎ」ヲ感じる
④	＜主体＞ガ	24	
⑤	＜対象＞ガ＜思考＞ト	21	このような商品をお客様に提案することガ罪にさえなるト感じる
⑥	＜主体＞ガ＜対象＞ニ	10	私ガひとりぼっちニ感じる
⑦	＜対象＞ガ＜思考＞ニ	6	家事ガ負担ニ感じる
⑧	＜対象＞ヲ＜主体＞ニ	4	温かな愛に満ちた腕ヲ体ニ感じる
⑨	＜対象＞ヲ＜思考＞ト	2	家ヲ新しいト感じる
⑩	＜主体＞ニ＜対象＞ガ＜思考＞ト	2	国民ニ科学技術や科学者等ガ身近な存在ト感じる

ガ格が感覚主体を表わす場合（①②④⑥）、主体の心理・感情を感覚対象として取り、それをヲ格やニ格が目的語や補語として表わしている。しかし④は目的語や補語を取らず、主体の内部感覚を示す自動詞的用法である[3]。

[1] 構文を整理するに当たり、中村（1977）における類型化の方法を用いた。これにより受動・使役・可能などのヴォイス付加形式や修飾関係の表現は基本形の格によって整理した。また、「感じる」の前項が形容詞・形容動詞・助動詞の場合、表現として適格となる格助詞に置換した。

[2] 用例は 2.3. にて得た「感じる」289 例から比喩用法 21 例を除いた 268 例である。

[3] 次のように性的な快感を得ている例などが挙げられる。

彼氏が私のことを遊んでる女なんじゃないかと疑っているようです。実際遊んでる理由に上記の原因はありません。すべて感じている結果です。それを無知な彼氏に教えてやりたいのですがどうしたら理解できるでしょうか？

(OC15_00007-890)

(1)　３０代で独身の男の人が寂しさを感じるときってどんなときですか。映画館とかで並んで入館を待つとき・・・・二列で並んでと言われるけど・・・・自分だけが、一列の時あとは、花火大会を一人で見に行く時・・・・ディズニーランドに行きたくても一人では行けない時（…）

(OC09_04688-160)

(2)　日本でも従来の病院の概念を覆すような病院らしくない病院ができ始めている。次の特集では、具体的な取り組みを紹介したい。もちろん、すべての人が「理想の病院」と感じるわけではないだろう。だが、かなりの人が「いい」という評判の病院なのである。

(PM11_00322-31030)

(3)　硫化水素で命を絶つ若者が多いですね。この世で生きることは苦しいのだろうと、頭ではわかります。たくさん人がいてもひとりぼっちに感じる冷たい日々を過ごしている人もいるかもしれません。

(OY14_02015-1700)

(1) は「３０代で独身の男の人」が映画館に並ぶときといった特定の状況下で「寂しさ」を抱くことを述べた文である。主体内部に生起した感情を「感じる」によって受けている。(2) は「わけではない」と否定形を伴うが、従来の病院とは異なる取り組みを行う病院に対して「理想の病院」であるとの判断を下すものである。(3) は自殺する若者が抱く生き辛さを語った文である。周りに人がいても、主体の主観の中では孤独感を抱いていることを「ひとりぼっち」によって表わし、それを「感じる」によって受ける。

(1) (2) (3) は、話者の内面に生じる感情をヲ格やニ格で受ける (1) (3) と、ある対象についての主体の判断をト格で示す (2) に分けることができる。また、次の (4) のようにヲ格に「痛み」のような感覚対象を五感により知覚する例もある。

(4)　玲子さんには話していないが、正弘さんも自分の体の変化に戸惑っているという。急に白髪が増え始め、歩く時にひざの痛みを感じるようになった。

(PN2c_00006-7380)

ガ格が感覚対象を表わす場合 (⑤⑦⑩)、その対象に対して主体が抱く思考や印象を「感じる」によって受ける文となる。それは感覚対象をニ格・ヲ格で表わす場合も同様である (③⑧⑨)。

(5)　男が払うのが「当たり前」みたいな連中やっぱりそういう考えの子は男性に頼るというか、金・金・・・・ブランド・・・とねだる女が多いと感じます将来を考えたら、結婚はオススメしないですよ自分で稼いで買おうという子でないと、ダンナさんの稼ぎにブゥブゥ文句ばかり言う嫁になります

(OC09_04681-1700)

(6)　「家事が負担に感じるから」にいずれか１つ以上回答した人の割合。

(OW6X_00069-81430)

(5) は男性に対して金銭やプレゼントを期待する女性についての印象を述べた文である。

「女性」は「感じる」の対象でありガ格で示され、それに対して主体は「多い」と判断を下している。(6) は、生活に関するアンケートの設問の一つである。対象である「家事」はガ格を、それに対する判断である「負担」はニ格で示されている。(5) (6) ともに対象に対する判断を示している点で (2) に近い用法となる。

これらをもとに「感じる」の語彙的意味を記述するとすれば次のようになる。主要な意味項目は≪≫で括った。

(7) 「感じる」は主体の≪内部≫に生起した感情・感覚を表わす。感覚を表わす場合、≪身体的≫な知覚を表わす。また、ある対象に対して抱いた主体の≪判断≫を表わすこともできる。

格関係は【表 2】に示したように複数に渡るが、そこで示されている意味は、感情・対象・判断のいずれをどのように知覚したかが重要となることを (7) は示す。これをもとにして「感じる」が取る構文の意味パターンを次のようにまとめることができる。

(8) ー ① ＜感覚主体＞の内部に感情・感覚が生じる
② ＜感覚主体＞が＜感覚対象＞を五感によって知覚する
③ ＜感覚対象＞に対する判断を示す。

主体をニ格によって表わす場合もある (⑧⑩) が、⑧は主体の外部に存在する事物を自身の皮膚感覚によって知覚する点で (8) ②に、⑩は感覚対象に対して「身近」と主体の心理的な把握がなされるものであり (8) ③と同様のタイプの構文として見なすことができる。

それぞれのパターンにおける感覚主体と対象、判断の関係は次のように図示される。

【表 3】:「感じる」における主体・対象・判断

	構文	主体	対象	判断
①	＜感覚主体＞の内部に感情・感覚が生じる	○	×	×
②	＜感覚主体＞が＜感覚対象＞を五感によって知覚する	○	○	×
③	＜感覚対象＞に対する＜判断＞	○	○	○

構文としての明示よりも、その構文を成立させるために必要な要素を○で示した。このような構文の持つ意味的パターンのうち、比喩用法として用いられるのはいずれであるか、またその原因となる要素を次節にて検討する。

3.2. 比喩用法における「感じる」

比喩用法において「感じる」を用いる場合、次のように８つの構文を形成する。

【表 4】: 比喩用法における「感じる」の取る構文

	構文	用例数	類型例
①	＜対象＞ニ＜思考＞ヲ	7	日差しニ質量ヲ感じる
②	＜主体＞ガ＜対象＞ヲ	5	人ガ時の流れヲ感じる
③	＜主体＞ガ＜対象＞ト	3	人ガ心を洗われたト感じる
④	＜対象＞ガ＜思考＞ト	2	会議ガ「大切な心の引継ぎ」であるト感じる
⑤	＜対象＞ヲ＜思考＞ニ	1	色とりどりの草花ヲセピア色ニ感じる
⑥	＜場＞カラ＜対象＞ヲ	1	表情カラ透明な情熱ヲ感じる
⑦	＜主体＞ガ＜対象＞ヲ＜場＞デ	1	ユーザーが何を考えているかヲ肌デ感じる
⑧	＜対象＞ト＜対象＞ガ＜思考＞ト	1	数学ト少女マンガのイラストガ同じくらい美しいト感じる

非比喩用法との顕著な差異は【表 2】④「＜主体＞ガ感じる」のような目的語を取らない構文が存在しないことである。これは「感じる」が比喩用法となる場合、目的語を最低限の要素として要求することを意味する。これにより【表 3】①が比喩となり得ない。

⑥⑦は、対象が存在する場や感覚する部位をカラ格、デ格で表わしているが「＜対象＞ヲ」感じる点では②と同じ構造を持つ。⑧もまたト格により事物が並列されるが「＜対象＞ガ＜思考＞ト」と同構造である。残る①②③④⑤は、感覚対象に比喩的な思考・印象を抱く①④⑤、主体が比喩的に表現された対象を感じる②、比喩的な判断を表わす③のタイプに分かれる。それぞれの用例は次のようになる。

①→（9）そのイチローの「自分自身のスタイル」を見ていると、ストレッチなどの所作の手順やグラブとバットの取り扱いも含めた型とリズムに、日本古来の武術の修業者、更には茶道など日本文化のさまざまな分野の型と呼吸に通ずるイメージを感じてしまうのである。

（PN4g_00009-16800）

②→（10）ゴツゴツとした黒い地面が広がり、ところどころにまっ黒にこげた木の切れはしが落ちている。火口だった場所は、すり鉢のように深くえぐれている。けれど、今はそこに緑の木がのび、時の流れを感じさせる。

（PB1n_00024 - 6750）

③→（11）奥多摩のことを書くのはもうこのへんにしておこうかね。さすがは奥多摩です。いろいろな光景に心を洗われたと感じております。

（OY14_03713-520）

（9）は、野球選手イチローのプレイスタイルを見て、それを刺激とした連想から日本の伝統文化のイメージを想起する。イチローのプレイと日本の伝統文化は属するカテゴリーが異なるものであるが、後者と前者を関係付けることによりイチローが伝統文化に一脈続

くという新たな観点を喚起する点で比喩となる。被喩辞[4]「イチローのスタイル」と喩辞「日本文化のさまざまな分野の型と呼吸に通ずるイメージ」が相互比較されることにより比喩性が生じる。しかし、「イメージ」の具体性を扱うものと解釈することも可能であり、境界的な例である。

(10) は噴火が起きてから数年後の火口の情景とそれに対する話者の心理を描いている。噴火により多くの木々が焼失した場所に木が生長し始めており、その情景に時間の経過を感じとっている。「時の流れ」は慣用的であるが「流れ」により時という抽象体を流体という具体物へと転化する。その結果として川の流れに身を置くようにして時の経過を感じ取るように表現する。

(11) は、奥多摩に観光に行き、その豊かな自然に癒されたことを書いたものである。癒されてすっきりしたことを「心を洗われた」と表現している。「洗う」によって抽象物である「心」を洗うことができるものとして実体化する点で比喩となる。そのような比喩的判断を「感じる」によって受ける。

(10) (11) はヲ・ト格の部分に比喩が生じている。しかし、(9) は「感じる」の前のヲ格部分だけを見ても比喩が生じているとは言い難い。これはその前のニ格部分の事柄と相互的に比較することによって両者の結び付きが比喩であることに気づくものである。このように主体・対象・判断のいずれを明示するか、そしてそれをどのような格関係のもとに置くかにより比喩の生じ方に差が生れる。この差異は【表 3】にて示した「感じる」の取る構文の意味パターンのうち、②と③の対立となる。

直喩は被喩辞と喩辞の比較による相互作用が重要である（李 1990）ことを踏まえると、【表 3】②を用いる表現は感覚対象が喩辞によって表現され被喩辞が明示されていない表現となり、直喩として認めることができない。これらの用例のうち、直喩として認定できるのは【表 3】③のように対象に対して主体の判断が下されたものである。それに照らし合わせると結果として【表 4】①④⑤⑧の 11 例が「感じる」が比喩指標として用いられた直喩の候補となる。また、そこから (9) のように比喩かどうかの境界的な例を除くと①「＜対象＞ニ＜思考＞ヲ感じる」の内の 2 例となった。

4. 考察

ここまでの分析により「感じる」の持つ構文の中から直喩となり得るものを確定した。確例は 2 例であり、これは「感じる」を比喩指標として認定することの難しさを意味する。比喩指標として認定するよりも、「＜対象＞ニ＜思考＞ヲ感じる」のような特定の構文環境に置かれた場合に例外的に比喩指標に似た機能を果たすと捉えたほうが実態に即する。中村 (1977) における出現数の少なさもこの点から理解される必要がある。以下は、例外的ながらも比喩指標と同様の機能を果たし得る条件を考察する。

4.1. 比喩指標の条件

比喩指標は直喩のマーカーとして比喩性の明示を担う。比喩指標の典型例である「よう」は命題に対する話者認識を表すモダリティ形式であり、命題がある条件を満たした場合、比

[4] 「被喩辞」は喩えられる事物・事柄を表す言葉。「喩辞」は喩えるために引き合いに出された事物・事柄を表す言葉。

喩指標として機能し得る。そこで、「よう」が受ける命題がいかなる性質を持つものであるかに着目して比喩指標となる条件を考察する。

(12)　燃えるような赤いもみじ　(せきしろ『たとえる技術』)
(13)　肌が雪のように白い

(12) は「もみじ」の赤さを描写するために喩辞「燃える」を用いている。「燃える」と「もみじ」は「よう」を用いないと次の (14) のように非比喩として解釈可能なものとなり、比喩的修飾には「よう」が必須となる。

(14)　燃えるもみじ

これは「もみじ」を現実に燃やすことが可能な故であり、そのように事実として燃焼しているという解釈を排除するために事実否定的な話者認識を「よう」によって表示する。これは「よう」の取る命題の真偽判断からいえば、偽であることを表わす。「よう」は推量用法の場合、真偽未確認の命題を取るが、ここでは命題が偽であるという認識を表す。

(13) は被喩辞「肌」の白さが喩辞「雪」を引き合いに出すことによって表現される。「肌」が「雪」によって組成されることはありえず、両者の関係は事実として成り立たない。その点で命題は偽である。しかし、事実においては結び付かない「肌」と「雪」であるが、「白い」という点に着目することにより比較することが可能となる。その比較を通して、「雪」の白さのイメージが「肌」に付与される。このように「肌」と「雪」という事実として結び付かない二物が有意味な関係を有する命題を「よう」は取る。

このように命題が偽でありかつ、構成する二つの事物が一定の根拠をもとにして有意味に結び付いている場合、「よう」は比喩指標として機能する。

(15) ― ①　偽の命題を取る。
　　②　二つの事物・事柄の間に何らかの関係性がある。

4.2. 比喩指標としての「感じる」

「感じる」は例外的ではあるが「＜対象＞ニ＜思考＞ヲ感じる」といった特定の構文に置かれた場合、文法機能として (15) ①②のような命題を取ることができる。それが (16) (17) となる。両例ともに使役態によって表現されているが、通常形に戻した場合、「日差しニ質量ヲ感じる」「恋愛運ニもろさヲ感じる」となる。

(16)　梅雨はすでに明け、九州地方は一気に夏模様である。質量を感じさせる強い日差しがビルや街路樹に降り注ぎ、道路に濃い陰影を作っていた。
(PB43_00054-1540)

(17)　どことなく“もろさ”を感じさせる恋愛運。カップルの人は、彼に傷つけられることが多いかも。
(PM31_00254-68750)

（16）は、「日差し」に「質量」を感覚することはなく、その点で真の情報とはなりえない。本来感覚できないものを前項に置くことにより、取り立てる情報が偽であることを保証する。偽の情報を取るが、「日差し」と「質量」の結び付きは理解不可能なものとはならない。「日差し」がいかなる強さを持つのかをただ「強い」と比喩を用いずに修飾するよりも効果的に説明することができる。それは「日差し」と「質量」を結ぶための共通項が存在し、それを根拠にして両者を比較することにより有意味に結び付くためである。ここで共通項となるのは皮膚感覚である。日差しの暑さは皮膚感覚によって知覚するが、実体としては認知できない。そこで何らかの物質が肌に接しているかのように表現することで、言い換えれば「質量」により「日差し」を実体化することにより程度の強調を行う。しかし、これは一方で（7）で示した「感じる」の持つ語彙的意味である≪身体的≫な感覚を表しており、実際に感覚しているという表現となる。語彙的意味が存在していることにより「よう」のような文法機能を主に担う形式に比べるとモダリティ形式化の度合は弱い。それは、この表現が「よう」や「みたい」に置換することができないことからも裏付けられる〔(18)（19)〕。

(*[5]18）質量のような強い日差し
(*19）質量みたいな強い日差し

語彙的意味が残存することにより、用いられる場面が（16）のように身体感覚を表す場面に制限される。文法形式化しきれず、使用できる場面に制限がかかることが比喩指標として機能することを困難にしていることを意味する。

（18）は、「恋愛運」がどのような様態にあるかを、それが実際には持たない属性である「もろさ」によって表現する点で比喩となる。「恋愛運」は実体ではない以上、実体物の属性である「もろさ」を持つとするのは偽の情報となる。しかし、伝達における効果としては「よくない恋愛運」のように表現するよりも、当該の恋愛運がどのような属性を持っているのか分かりやすいものとなる。両者ともに些細な接触で傷がつくものであり、そのような共通項をもとにして「もろさ」という属性が付与され、かつ「恋愛運」を実体化する。この例は（16）のように身体感覚を表すものではないが、ある事柄に対しての思考・印象を≪判断≫している点では「感じる」の語彙的意味が残存しており、(16）と同様にモダリティ形式化しきれていない。しかし（16）（17）ともに（15）①②を満たすために比喩指標に近い用法となっている。

このように「感じる」は「＜対象＞ニ＜思考＞ヲ感じる」という特定の格関係のもと、対象と思考の関係が通常ならば成立しない偽の情報であること、感覚対象と思考が共通項によって有意味に結ぶという命題となっている場合、比喩指標と同様の機能を果たす。しかし、語彙的意味が残存することにより、モダリティ形式化が阻まれ、「よう」のような用法の広さを持たないとして結論付けることができる。

[5] *は非文を表わす。

5. おわりに

本発表は「感じる」の構文を整理し、そこから比喩用法になり得る構文を選別した。それをもとに、直喩として認めることのできる構文を確定し、比喩指標として用いられるための条件を考察した。まとめるならば次のようになる。

① 「＜対象＞ニ＜思考＞ヲ感じる」という特定の格関係において直喩と認定することができる
② 偽の情報であり、かつ二物が有意味な関係にある命題を受けることが必要となる。
③ 「感じる」の語彙的意味が比喩指標としての形式化を阻む。

この結果は中村（1977）における比喩指標の認定に対する再考を求める。他の比喩指標についても同様に用例を収集し、構文の整理や意味の分析を施すことにより、それぞれの語が比喩指標となるための条件が明らかになる。このようなアプローチは直喩の有する形式の類型化などの比喩研究への寄与とともに、文法研究と比喩研究を接続しようとするものである。

参考文献

加藤祥・浅原正幸・山崎誠（2017）「『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーション」，言語処理学会第 23 回年次大会発表論文集，306-309.
菊地礼・加藤祥・浅原正幸（2018）「「感じる」を指標とするメタファー用例の収集とその分析」（メタファー研究会、発表資料）
木下りか（2003）「直喩形式と類似性　―ヨウダとニテイル―」『大手前大学人文科学部論集』4 号, pp153 - 164.
小松原哲太（2016）『レトリックと意味の創造性』京都大学学術出版会.
せきしろ（2016）『たとえる技術』文響社.
中村明（1977）『比喩表現の理論と分類』国立国語研究所報告 57.
鍋島弘治朗（2016）『メタファーと身体性』ひつじ書房.
K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka and Y. Den, (2014) “Balanced corpus of contemporary written Japanese”, Language Resources and Evaluation, 48:2, 345-371.
李徳奉（1990）「比喩の意味における喩辞と被喩辞の相互関係について」『日本語と日本文学』13 号, pp. 37 - 49.

日本語 Wikipedia を用いた慣用句の構成性の数値化

岡田優也 (関西学院大学大学院)

Numerical Representations of the Compositionality of Idioms in Japanese Wikipedia

Yuya Okada (Kwansei Gakuin University)

要旨

本研究は、日本語慣用句の構成性の度合いを数値化する手法を用いて、慣用句と一般連語句の差異について調査するものである。構成性の数値化は、日本語 Wikipedia の全記事からなるコーパスをもとに、調査対象とするひとつの慣用句について構成的連語句分散表現ベクトルと非構成的連語句分散表現ベクトルの 2 つのベクトルをそれぞれ獲得し、この 2 つのベクトル間のコサイン類似度を計算することによって行う。構成的連語句分散表現ベクトルとは、連語句の構成要素である単語の分散表現ベクトルから加算的に計算されるベクトルである。一方、非構成的連語句分散表現ベクトルとは、連語句を内部構造のないひとまとまりの表現だと捉え、コーパスから直接的に獲得されるベクトルである。獲得した 2 つのベクトルのコサイン類似度が慣用句の構成性の度合いを反映するものであることを主張するため、一般連語句についても同じ手法により構成性の度合いの数値化を行い、慣用句について獲得された数値との間に統計的に有意な差が生じることを示す。

１．はじめに

これまで、慣用句の句全体としての意味は、慣用句を構成する単語の意味の総和と一致しないという性質があることが指摘されてきた (石田 2015、宮地 1982)。

(1) つくばに来る (一般連語句) / 頭に来る (慣用句)
(2) (カップに) 熱湯を注ぐ (一般連語句) / 目を注ぐ (慣用句) (石田 2004)

上記 (1)、(2) の例では、「頭に来る」、「目を注ぐ」[1]という慣用句の全体としての意味が、その構成要素である単語の意味からは推測することができない。一方、「つくばに来る」、「熱湯を注ぐ」という一般連語句[2]の全体としての意味は、構成要素である単語の意味を知る人ならば理解することができる。本研究では、この慣用句の性質を非構成性と呼ぶことにする。

その一方で、この慣用句の非構成性という性質は絶対的なものではないことも指摘されている。宮地 (1991) は、慣用句を連語的慣用句と比喩的慣用句の下位分類に大きく分け、連語的慣用句は「相対的に言って一般連語句より語と語との結びつきが強いもの」であり、比喩的慣用句は、さらに「句全体として派生的・比喩的意味を持つ」ものであると定義している。

---

[1] 慣用句の漢字による表記とひらがなによる表記については、引用部および引用部について言及する場合を除き、後述する『用例でわかる慣用句辞典改定第 2 版』の表記に合わせた。
[2] 一般連語句とは、二つ以上の単語がほぼ自由に結合してできる句のことを表す (石田 2015)。

(3) 嘘をつく、風邪をひく、汗をかく、雨があがる
(4) 頭にくる、口が重い、お茶をにごす、顔がひろい (宮地 1991: 69-70)

上記 (3) に、連語的慣用句の例を示した。連語的慣用句においては、「嘘」、「風邪」、「汗」、「雨」などの慣用句の構成要素となる単語が通常の意味のまま全体としての意味の一部となっている。一方、(4) の比喩的慣用句の例においては、「頭」と「くる」や「口」と「重い」などの単語が通常の意味では、慣用句の全体としての意味に寄与していない。このように、慣用句は、ある程度構成的な表現である連語的慣用句と構成的でない比喩的慣用句に分類することができる。

この連語句に対する分類は、慣用句とそれ以外の表現との違いや慣用句の中でも意味的な性質に違いがあることを考察するために重要な指針となると考えられる。しかし、一般連語句と連語的慣用句、連語的慣用句と比喩的慣用句の間の境界線は、必ずしもはっきりしない場合がある (宮地 1991)。本研究では、慣用句の非構成性という段階的な性質をよりうまく捉えるために、自然言語処理分野において提案された慣用句の構成性の度合いを数値的に表現する手法を応用し、日本語の慣用句に適用する方法を提案する。

２．先行研究

２.１　日本語学分野における慣用句研究から

石田（2015）は、慣用句の性質として、一般連語句よりも構成要素が固定していること(形式的固定性)、一般連語句よりも文法的な制約が強いこと（統語的固定性)、句全体の意味が句を構成する個々の単語の意味の積み重ねと一致しないこと（意味的固定性）の 3 つをあげている[3]。また、慣用句の意味的固定性は、慣用句の性質の一つである統語的固定性に反映されると主張している。なぜなら、ある慣用句の統語的固定性が高いということは、その慣用句の構成要素が句の中で解釈可能な意味を表していない、もしくは通常の意味から変化している可能性があるため、その慣用句の意味的固定性が高くなると考えられるからである。

たとえば、「目を向ける」という慣用句は、名詞句への転換（「鈴木へ向けた目には…」)、連体修飾語の付加（「恨めしそうな目を向けた」)、連用修飾語の挿入（「目をくるりと信夫に向けた」）などの統語的な操作を容認するため、統語的固定性が低いと判断できる。一方、「頭に来る」などの慣用句は、このような統語的な操作を許さず、統語的固定性が高い慣用句であると考えられる。したがって、「頭に来る」という慣用句は、「目を向ける」という慣用句と比較して、統語的固定性が高いため、意味的固定性も比較的高い慣用句であると判断することができる（石田 2004)。

これらの石田（2004, 2015）の研究は、対象となる慣用句に対して統語的な操作が可能かという明確な基準によって、客観的に判断することが難しい慣用句の意味的固定性を推定しているという点で評価できる。しかし、この手法では個々の慣用句に対していくつもの統語テストを適用し、慣用句の統語的固定性を母語話者の直感によって判断する必要がある。そのため、数多くの慣用句に対して、意味的固定性を測定することは難しい。

２.２　自然言語処理分野における慣用句研究から

慣用句に関する研究は、自然言語処理の分野においても、複合表現（Multi Word Expressions）研究の一部として盛んに行われている。Constant ら（2017）のまとめによる

[3] 石田（2015）における意味的固定性の概念は、本研究でいう非構成性とほぼ一致するものであると考えられる。

と、自然言語処理で行われるさまざまなタスクにおいて、複合表現をうまく処理するために、複合表現として処理すべき連語句の発見（MWE Discovery）と文中に出現する連語句が、複合表現としての意味と字義どおりの意味のどちらの意味で使用されているのかの判定（MWE Identification）という 2 つの課題が設定されている。この 2 つの課題のうち、特に複合表現の発見手法が本研究と関係するため、ここで紹介する。

複合表現の発見の手法の中には、複合表現の意味が非構成的であるという性質を利用するものがある。複合表現においては、構成要素となる単語の意味の積み重ねと全体としての意味が一致しないため、複合表現全体の意味的表象とその構成要素の意味的表象に類似関係が認められない。この全体としての意味と構成要素の意味との関係から個々の連語句が複合表現であるかの判断をくだすのである。たとえば、英語の“hot dog”という表現について考えてみると、“hot dog”という表現の全体としての意味と構成要素である“hot”と“dog”という 2 つの単語の意味の間に、何ら明示的な類似点を指摘することはできない。したがって、“hot dog”という表現は、複合表現であるとみなすことができる。

Salehi ら（2015）は、単語の分散表現ベクトル化手法を使い、複合表現における構成性の度合いを数値的に予測する手法を提案している。まず、調査対象となる複合表現およびその構成要素となる単語の分散表現ベクトルを獲得する。その後、以下の（5）、（6）の式によって、複合表現のベクトルと構成要素のベクトルの類似度を計算する。

(5) $comp_1(MWE) = \alpha sim(MWE, C_1) + (1 - \alpha) sim(MWE, C_2)$

ここで、MWE は複合表現の分散表現ベクトルを、C_1、C_2 はそれぞれ複合表現を構成する単語の分散表現ベクトルを表す。α は、0 から 1 までの実数値をとる重みパラメータである。類似度の測定には、コサイン類似度が用いられている。この計算式により、複合表現全体の分散表現ベクトルとその構成要素の単語の分散表現ベクトルの類似度が測定される。

(6) $comp_2(MWE) = sim(MWE, C_1 + C_2)$

一方、(6) の計算式では、複合表現の構成要素である単語の分散表現ベクトルをあらかじめ足し合わせた上で、複合表現全体の分散表現ベクトルとの類似度を測定している。類似度の測定には、(5) と同様にコサイン類似度が用いられている。

2 つの計算式 (5)、(6) は、どちらも複合表現の構成性の度合いを数値化するものであるが、前提となる考え方には、若干の違いが存在する。(5) の計算式では、複合表現と構成要素の単語のベクトルについての類似度を評価している。つまり、対象となる複合表現が非構成的な表現であるならば、その全体としての意味と構成要素の意味の間に類似性がみられないであろうということが仮定されている。一方、(6) の計算式では、構成要素の単語のベクトルを足し合わせることにより、複合表現のベクトルを MWE とは別にもう一つ作成し、両者の類似度を評価している。よってこの計算式では、複合表現が非構成的な表現であるならば、その構成要素同士の意味を足し合わせたとしても、全体としての意味に近いものを得ることはできないということを仮定していると考えられる。

Salehi らの手法は、単語の分散表現ベクトル化という、人手による処理を必要としない手法を利用しているため、計算対象となる表現が増えたとしてもそれほどコストが増加しないという利点がある。一方、Salehi らの研究では、得られた構成性の度合いを人手で作成されたデータセットと比較することによって評価しているため、適切なデータセットが存在しない言語について得られた構成性の度合いを評価するのは難しい。

３．手法

ここでは、Salehi らの手法を参考にして、本研究が提案する日本語の慣用句の構成性の度合いを数値化する方法を説明する。図 1 は、この手法の概要を示したものである。

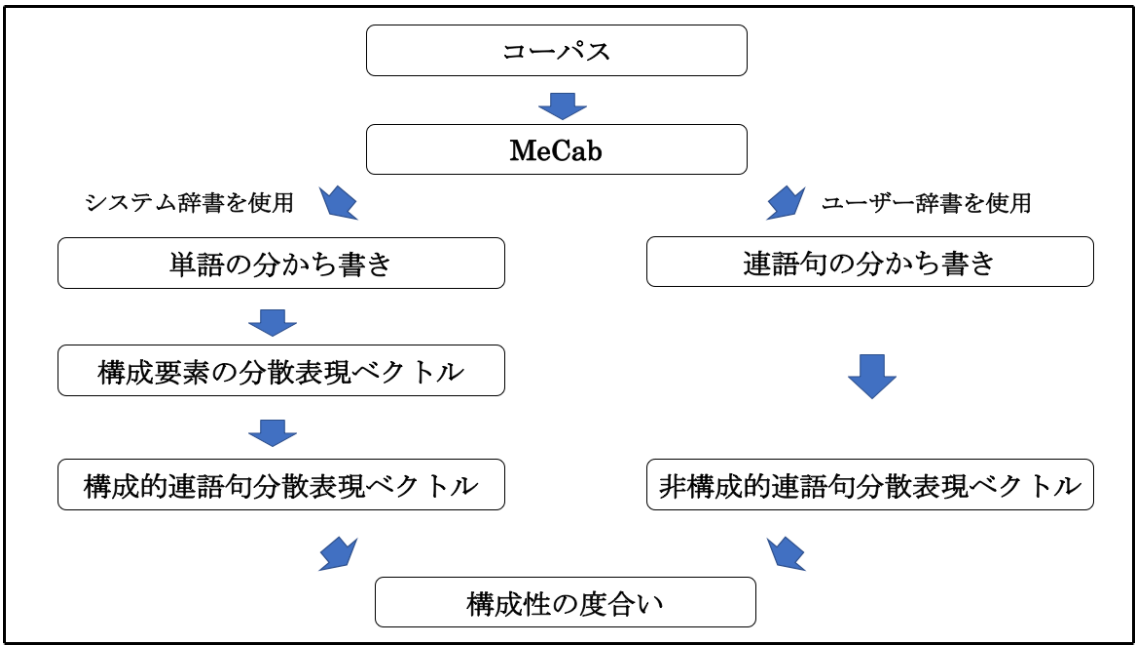


図 1：手法の概要

まず、日本語 Wikipedia のデータからなるコーパスを形態素解析器 MeCab によって、分かち書きする。この際、MeCab のデフォルトのシステム辞書である IPA 辞書と共に本研究で調査対象となる連語句について記述したユーザー辞書を使用した。このユーザー辞書を使用することによって、調査対象の連語句が構成要素に分解されず、一つの単語と同等の扱いを受けることになる。

次に、分かち書きデータから Mikolov ら（2013）によって作成された word2vec ツール[4]を使用し、単語および調査対象である連語句の分散表現ベクトルを獲得する。続いて、連語句の構成要素である単語の分散表現ベクトルをすべて足し合わせ、連語句の分散表現ベクトルを計算する。本研究では、この構成要素のベクトルの総和によって計算されるベクトルを構成的連語句分散表現ベクトルと呼ぶことにする。一方、ユーザー辞書を用いて分かち書きされた連語句については、ひとまとまりの表現として、連語句の分散表現ベクトルが与えられる。ここでは、この連語句を一つの単語のように扱い、直接与えられるベクトルを非構成的連語句分散表現ベクトルと呼ぶことにする。

こうして獲得された連語句に対する構成的連語句分散表現ベクトルと非構成的連語句分散表現ベクトルのコサイン類似度を計算することによって、連語句の構成性の度合いを数値的に推定する[5]。構成的連語句分散表現ベクトルは、連語句の構成要素である単語の総和によって獲得されるため、構成性の原理を部分的に仮定したベクトルであると考えられる[6]。

---

[4] word2vec は、gensim (3.4.0) から利用した。

[5] この手法は、先行研究で言及した Salehi ら（2015）の手法において（6）の式を使用したものにおおよそ相当する。

[6] ベクトル同士の単純な加算では、語順や句構造上の違いを捉えることができないため、完全

一方、非構成的連語句分散表現ベクトルは、構成性の原理を仮定せずに獲得することのできるベクトルである。したがって、構成性の原理を部分的に仮定するベクトルと構成性の原理を仮定しないベクトルの 2 つのベクトル表現におけるコサイン類似度を計算することにより、連語句がどの程度構成性の原理に従う表現かを推定することができると考えられる。

４．実験

４.１　実験設定

４.１.１　調査資料

連語句をベクトル化するためのコーパスには、日本語 Wikipedia のデータを利用した。2018 年 4 月 1 日時点の日本語 Wikipedia のダンプデータ[7]から、wikiextractor[8]によって、テキストデータを取り出したのち、記事タイトルの削除、半角および全角の丸括弧内の補足説明の削除、余分な空行の削除を行った。

テキストデータの分かち書きには、形態素解析器の MeCab を利用した。分かち書きの際、表層形を原形に変換する処理を入れたため、活用のある形態素については、基本形で出力されている。

また解析用の辞書には、デフォルトのシステム辞書に加えて、本研究において実験対象とした連語句について記述したユーザー辞書を使用した。表 1 に「立つ」の IPA 辞書の記述例、表 2 に「腹が立つ」のユーザー辞書の記述例を示す[9]。

表 1：IPA 辞書の記述例

表層形	左文脈 ID	右文脈 ID	原形
立つ	738	738	立つ
立た	740	740	立つ
立ち	743	743	立つ
立て	736	736	立つ

表 2：ユーザー辞書の記述例

表層形	左文脈 ID	右文脈 ID	原形
腹が立つ	1285	738	腹が立つ
腹が立た	1285	740	腹が立つ
腹が立ち	1285	743	腹が立つ
腹が立て	1285	736	腹が立つ

ユーザー辞書の記述では、IPA 辞書と対応する表層形および原形の表現の前に「腹が」という表現が追加されている。左文脈 ID には、すべての表層形に対し 1285 という一般名詞を表す数字が与えられ、右文脈 ID には、対応する表層形と同じ数字が割り振られている。ここで、左文脈 ID とは、対象となる単語を左から見た場合の文脈 ID を示し、右文脈 ID と

に構成性の原理に基づくベクトルとみなすことはできない。

[7] https://dumps.wikimedia.org/jawiki/

[8] https://github.com/attardi/wikiextractor

[9] 実際の IPA 辞書、ユーザー辞書には品詞情報や読みなどの詳細な記述が含まれているが、紙面の制約上、ここでは議論に必要な部分のみを記載してある。

は、対象となる単語を右から見た場合の文脈 ID を示す。したがって、本研究のユーザー辞書では、実験対象となる連語句に対して左から見たときは一般名詞として、右から見たときは動詞の活用形となるように記述したことになる。

４.１.２　実験対象となる連語句

ここでは、実験対象とした連語句の選択方法について説明する。慣用句は、『用例でわかる慣用句辞典改定第 2 版』に収録されている身体の部位を使った慣用句について、名詞＋格助詞＋動詞の 3 つの形態素によって構成される動詞慣用句のみを調査対象とした。身体の部位を表す名詞表現は、日本語の身体表現として代表的と考えられる以下の 25 の名詞に限定し、格助詞については、「が」、「を」、「に」の 3 つを使用した。この段階で選び出された慣用句の数は、323 である。

(7)　「頭」、「顔」、「口」、「首」、「舌」、「歯」、「鼻」、「額」、「頬」、「眉」、「耳」、「目」、「足」、「腕」、「肩」、「体」、「腰」、「尻」、「手」、「肌」、「腹」、「膝」、「身」、「胸」、「指」

比較対象となる一般連語句は、慣用句を選択した際に基準とした 25 の名詞表現と 3 つの格助詞の組み合わせによって構成される 75 のペアに対し、それぞれのペアと共に用いられる動詞について使用頻度[10]の高いものを順に 5 つ抽出し、名詞と格助詞のペアに結合することによって作成した[11]。この段階で選び出された一般連語句の数は、375 である。

また、word2vec を使用する際、最低出現頻度数を 10 と設定したため、選び出された 323 の慣用句と 375 の一般連語句において、日本語 Wikipedia コーパスの中でトークン頻度[12]が 10 よりも低い表現については排除されている。その結果、180 個の慣用句と 135 個の一般連語句が調査の対象となった。

４.１.３　実験方法

調査対象となる 180 の慣用句と 135 の一般連語句のそれぞれについて、3 章で記述した方法によって、構成性の度合いを数値化する。獲得した数値データが、慣用句と一般連語句の構成性の度合いを反映したものであるかを判断するために、対応なしの t 検定を行う。慣用句の特性が非構成性であるということを考えると、慣用句に対して得られた構成性の度合いは、一般連語句に対して得られた数値よりも統計的に有意に低いと予測される。

４.２　実験結果

表 3 に慣用句と一般連語句それぞれに対して得られた構成性の度合いの記述統計量を示す。

---

10 使用頻度の測定には、現代日本語書き言葉均衡コーパスを使用した。

11 出来上がった名詞+格助詞+動詞の組み合わせに慣用句が含まれないようにするため、実験対象の慣用句のリストと照合して、基本形が同じものについては除外した。またこの際、慣用句における動詞の漢字表記とひらがな表記の両方を一般連語句の集合から除外するように注意を払った。

12 この際、慣用句と一般連語句の両方で動詞の活用形については頻度集計に含まれているが、対象となる連語句の間に修飾要素が挿入される、構成要素の移動が起きるなどの統語的な変形が起きている場合に関しては、頻度集計に含まれていない。

表 3：記述統計量

記述統計量	個数	平均	標準偏差	最低値	中央値	最大値
慣用句	180	0.183	0.105	-0.052	0.176	0.515
一般連語句	135	0.279	0.125	0.019	0.265	0.577

慣用句の方が、一般連語句よりも平均値が約 0.096 低いことがわかった。一方、慣用句の最大値は 0.515 と一般連語句の平均よりも大きく、一般連語句の最低値も 0.019 と慣用句の平均値よりも小さな値が出ている。したがって、慣用句と一般連語句の構成性の度合いの分布は、中心となる位置に差はあるものの互いに重なりあったものになると考えられる。

この慣用句と一般連語句における構成性の度合いが統計的に有意なものであるかを判断するために、対応なしの t 検定を行った。その結果、t 値 7.448 ($p < 0.05$) で統計的に有意であることがわかった。また、効果量を測定したところ Hedges の g 値が 0.848 で大きな効果があることがわかった[13]。

４．３　考察

実験結果より、本手法により慣用句に対して与えられる値の平均は、一般連語句に対して与えられる値の平均よりも有意に低いことがわかった。これは、本手法により構成性の度合いを数値化することができているならば、予測できるものであり、本手法を支持するものであると考えられる。

一方、慣用句に与えられた値と一般連語句に与えられた値には、互いに重なりあう部分があり、一般連語句の平均を上回る値が与えられた慣用句と慣用句の平均を下回る値が与えられた一般連語句が一定数存在することがわかった。これは、あるしきい値によって慣用句と一般連語句を明確に区別することが難しいことを示している。また、このことは、宮地 (1991) などで述べられている慣用句と一般連語句との境界線は必ずしもはっきりしたものではないという慣用句と一般連語句との間の段階的な性質を捉えている可能性もある。

表 4 に予測に反して高い値が与えられた慣用句を上位 10、表 5 に一般連語句について予測に反して低い値が与えられたものを上位 10 示す。

表 4：高い値が与えられた慣用句

慣用句	構成性の度合い
身に付ける	0.515
首を切る	0.433
目に入る	0.430
手が届く	0.415
胸に秘める	0.411
足を入れる	0.399
手を借りる	0.387
尻を叩く	0.377
目が覚める	0.371
手を入れる	0.365

表 5：低い値が与えられた一般連語句

一般連語句	構成性の度合い
顔をする	0.019
口をつぐむ	0.027
体を起こす	0.090
目にかかる	0.096
腹に据える	0.099
首が回る	0.111
耳を持つ	0.116
頭が上がる	0.121
眉をひそめる	0.122
顔をしかめる	0.122

[13] 効果量の測定は、Kline (2004; 101-2) の計算式をもとに計算した。

表 4 の慣用句の多くは、字義どおりの意味と慣用句としての意味の両方の解釈ができる多義的な慣用句である。たとえば、「身に付ける」は (8) のように字義どおりの意味で使用できるのと同時に、(9) のように慣用句としての意味でも使用可能である。

(8) 昨日買ったアクセサリーを身に付けて、映画に行った。
(9) 長年の訓練の結果、立派な芸を身に付けた。

word2vec を使用したベクトル化では、一つの単語に対して一つのベクトル表現を与えるため、単語の多義的な構造をうまく取り扱うことができない。そのため、多義的な表現については、複数の意味の影響を受けたベクトルが獲得されると考えられる。基本的に字義どおりの意味では、構成性の原理が成り立つため、両方の解釈が可能な慣用句についてはその影響を受け、高い値が与えられた可能性がある。

それに対して、「胸に秘める」、「手を借りる」などの慣用句は、字義どおりの解釈を許さないと考えられるため、興味深い。「胸に秘める」という慣用句は、「心に秘める」という字義どおりに解釈できる表現で言い換えることができるため、「秘める」の部分については字義どおりの意味が残っていると考えられる。また、「手を借りる」についても、「助けを借りる」と言い換えることができるため、「借りる」の部分は、字義どおりの意味が残っている可能性がある。したがって、慣用句であっても比較的構成性の度合いが高いものについては、高い値が与えられている可能性がある。

一方、表 5 の一般連語句の中には慣用句と考えられる表現が混入している。これは、本研究で利用した慣用句のリストに含まれていない慣用句が存在したこと（口をつぐむなど)、格助詞の違いなどの文法的な違いをうまくマッチングできていなかったこと[14](目にかかるなど) 等が原因であると考えられる。また、一般に否定の表現とともに使われる慣用句の一部が取り上げられている可能性もある (頭が上がる / 頭が上がらない)。この点については、本研究における実験で慣用句と一般連語句の振り分けが十分でなかったことを示すものであるが、慣用句の構成性の度合いを数値化するという研究手法自体を否定するものではないと考えられる。

５．まとめと今後の課題

本稿では、複合表現の構成性を数値化する手法を日本語の慣用句に適用する方法を示し、慣用句と一般連語句とで得られる値の比較を行った。その結果、慣用句と一般連語句では得られた値の平均に統計的に有意な差があることがわかった。また、慣用句と一般連語句のそれぞれで得られた値は、重なりあう部分も多く、慣用句と一般連語句の間の段階的な性質を部分的に捉えている可能性が示唆された。この可能性を検証するためには、どのような慣用句に大きな値が与えられやすいか、どのような一般連語句に小さな値が与えられやすいかの詳細な調査が必要である。また、手法としては多義的な慣用句をどのように扱うかに注意をはらう必要がある。これらの点については、今後の課題としたい。

また、本稿では慣用句と一般連語句の間にどのような差が生じるかを研究対象としたが、先行研究では慣用句という大きな分類の中でも、構成性の度合いに差があることが指摘されている。本研究においても、慣用句の構成要素の意味が部分的に全体としての意味に寄与

14 たとえば、本研究で用意した慣用句のリストには「目を掛ける」という慣用句は存在したが、「目に掛ける」という助詞違いの慣用句は存在しなかったため、そのためこの「目に掛ける」という慣用句を機械的に一般連語句のリストから除外することができなかった。

していると考えられる慣用句（胸に秘める、手を借りるなど）には比較的高い値が与えられている。提案手法によって、異なる慣用句における構成性の度合いの違いをどの程度適切に表現することができているのかについては、今後の研究課題としたい。

参考文献

Constant, M., G. Eryiğit, J. Monti, L. Van Der Plas, C. Ramisch, M. Rosner, & A. Todirascu (2017). Multiword expression processing: A survey. *Computational Linguistics* 43(4), 837-92.

Kline, R. B. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. American Psychological Association.

Maekawa, K., M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, & Y. Den (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48(2), 345-371.

Mikolov, T., K. Chen, G. Corrado, & J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781.

Salehi, B., P. Cook & T. Baldwin (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 977-83.

石田プリシラ（2004）「動詞慣用句の意味的固定性を計る方法―統語的操作を手段として―」『国語学』 55: 42-56 日本語学会.

石田プリシラ（2015）『言語学から見た日本語と英語の慣用句』開拓社.

宮地 裕（1982）「慣用句解説」『慣用句の意味と用法』 237-265 明治書院.

宮地 裕（1991）「慣用句の意味」『「ことば」シリーズ 34 言葉の意味』 65-76 凡人社.

「ＸＸ(と)」、「ＸＸな」、「ＸＸしい」の構造・文法機能 ―畳語による生産性について―

陳祥（筑波大学　人文社会科学研究科博士後期課程）†

The Productivity of Reduplication in Japanese : Form,The grammatical function

Hiang　Chen
（University of Tsukuba /Graduate School of Humanities and Social Sciences Doctoral Program）

要旨

日本語の重複表現の形式は様々であるが、本稿で取り扱うのは「畳語」と呼ばれ、２つの語基の重複によって形成される語である。日本語には畳語が数多く、名詞、動詞など様々な品詞からなる。畳語には生産性が限られているが、「色々な」のように後ろに「-な」、「細々しい」のように後ろに「-しい」などを付け加えることによって、新しい語として形成することが可能である。今回は畳語表現である「ＸＸと」、「ＸＸな」、「ＸＸしい」という３種類を研究対象とする。データとして『現代日本語書き言葉均衡コーパス(BCCWJ)』を使い、３種類それぞれの共起関係や文法的振る舞いを明らかにする。また、構造上においては「色々な、*色々しい、色々(と)」、「*軽々な、軽々しい、軽々(と)」のように、３種類における「ＸＸ」の使い分けがあると見られ、３種類それぞれの構造が異なることを課題として分析した。例えば、派生元の品詞から見ると、「ＸＸと」の派生元である「Ｘ」の品詞は、名詞あるいは形容動詞、動詞、形容詞、副詞であるものがあり、バリエーションが一番豊かであることが分かった。また、BCCWJ を使用し、副詞としての畳語は形容動詞としても反復形容詞としても使われている用例が見られる。それは、副詞としての畳語は「ＸＸと」と接続することが最も多いが、「ＸＸな」や「ＸＸしい」の用例も見られることから、使用用法のうち自由度が最も高いと考えられる。今回の考察から、畳語による生産性の理解の助けになることを期待している。

１．はじめに

日本語には、同一の文字・語根・語を重ねる畳語と呼ばれる語群が存在する。この中には擬音・擬態語を重ねたものも含まれるが、最も一般的なものは「人々」「時々」といった類である（石川 2017)。反復される品詞成分を研究している玉村(1975)では、「人々、山々」のような名詞反復、「われわれ、そこそこ」のような代名詞反復、「恐る恐る、走り走り」のような動詞反復、「たかだか、長々しい」のような形容詞反復などの構造パターンがあると示している。また、これらの語は副詞として使われることが多いと述べている。副詞的に使用される畳語の後接構造を考察する研究は黄(2009) [1]が挙げられている。副詞として動詞を修飾する場合は「と」が 65%で「に」は６％に過ぎないと述べている。副詞的に使用される畳語以外に、「色々な、様々な」のように形容動詞として使われることも、「長々し

† sabrina9632@gmail.com

[1] 黄(2009)は、畳語型のオノマトペを研究対象として、それらの後接条件を調査した。

い、若々しい」のように反復形容詞として使われることもある。よって、畳語はこのように後ろに「-な」、「-しい」などを付け加えることによって、新しい語として形成され、バリエーションが豊かであると言える。しかし、畳語はどのような後続条件が存在するか、その構造に対応する意味範疇や共起関係や文法的振る舞いは十分に明らかにされていない。そこで、本研究では現代日本語において「ＸＸと」、「ＸＸな」、「ＸＸしい」を研究対象として取り上げ、対応する共起関係や文法的振る舞いを、『現代日本語書き言葉均衡コーパス(以下“BCCWJ”と称する)』を用いて分析する。３種類における「ＸＸ」の使い分けがあると見られ、最終的には、畳語による生産性の理解の助けになることが望ましい。

２．先行研究

畳語の構造・意味を分析した研究としては飯田(2005)、禹(2015)、石川(2017)が挙げられる。

２.１　飯田(2005)

飯田(2005)では、「状態という持続的な内容」をもつ形容詞が、合成語のうえで重複関係になる要因を考察するため、形容詞性構成要素からなる重複形容詞の構成要素を中心に、構成要素の意味特徴及び重複形容詞の意味用法を調査した。

その結果、構成要素の意味に「数えられる」対象・範囲等が含まれていれば、重複関係が成立する傾向があること、及び重複関係は多(回)数性を表すことがわかった。この多(回)数性は、場合によっては強調表現として考えられる。

２.２　禹(2015)

禹(2015)では、村上春樹の小説作品である『1Q84』Book-１・Book-２・Book-３の３冊に対象を絞り、現代日本語における畳語の諸機能について考察した。

先行研究では、日本語の畳語の機能について＜複数＞、＜反復＞、＜強調＞の三つの意味用法があるが、それらは名詞は＜複数＞、動詞は＜反復＞、形容詞は＜強調＞を表わすと記されている。しかし、『岩波 国語辞典[2]』において日本語の名詞の畳語に見られた＜複数＞と＜反復＞とに関する見解の相違ははっきりしない。そのため、禹(2015)では具体例を手がかりにして、言語を話す主体との関連を取り入れ、３つの意味用法を考察した。

結果から、名詞の畳語形を＜複数＞として捉えたり、＜反復＞として捉えたりするという経験的事実に関わる重要な要因は、＜時間＞という次元の介入の有無に求められることが分かった。また、＜強調＞は、外界に対する話者の積極的な捉え方の反映であると解し、さらにそれには「時空数量の強調」と「様態の強調」が存在することを明らかにしている。

２.３　石川(2017)

石川(2017)では、X々型畳語に対象を絞り、現代日本語の書き言葉コーパス資料を使い、高頻度語形を特定した上で、構造（品詞成分、反復要素同一性、モーラ数、後接構造)、使用（時代影響、ジャンル影響)、意味の３つの点から調査を行った。

構造に関する結果ては、反復される要素の品詞は、名詞が他の品詞（形容詞、副詞など）より多いこと、後接助詞については形容詞（句の一部）として名詞を修飾する事例が最も多く、主格、副詞用法がそれに次ぐこと、副詞用法の中では、最も多い順からは助詞を介在させない形「φ」、「に」、「と」などが明らかになった。そして、意味機能に関しては、複数（42%）＞反復（25%）＞個別（18%）＞強調（5%）＝増加（5%）＝語調調整（5%）となり、複数の意味が最も多いこと、複数の中では多数が単純複数より多いことと述べて

[2] 西尾実・岩淵悦太郎・水谷静夫（2009)『岩波国語辞典　第 7 版』岩波書店

いる。

３．調査概要

３.１　調査資料

畳語の品詞はバリエーションが豊かで、名詞や副詞や形状詞など様々である。しかし、品詞が異なっても、重複構造が一致するのは畳語の特徴だと言える。BCCWJ の検索機能は３つあり、それぞれ短単位検索、長単位検索、文字列検索である。現段階では BCCWJ において「Ｘ々」の重複構造を設定し、畳語を網羅することができていない。従って、共起関係と文法機能を分析する前に、辞書を使い手作業で畳語の数を把握することにした。

具体的には、辞書において、「ＸＸと」、「ＸＸな」構造形式をとっている畳語を抽出し、擬音語と擬態語のようなオノマトペも含む。そして、「ＸＸしい」構造形式をとっている反復形容詞はジャパンナレッジ Lib から抽出され、その条件設定が次節で詳しく述べる。

抽出した研究対象である「ＸＸと」、「ＸＸな」、「ＸＸしい」を畳語リストに作成した。今回は「ＸＸと」、「ＸＸな」、「ＸＸしい」からそれぞれ使用頻度が最上位５語を相互で比較しながら、構造・文法機能を考察する。

３.２　調査対象

まず、畳語である「色々」は副詞用法とする「色々と」の形と、形容動詞の用法とする「色々な」の形を持つのに対し、畳語である「生き生き」は副詞用法とする「生き生きと」の形のみを持っている。「ＸＸと」と「ＸＸな」は単純語として抽出することが困難であり、後接関係が確認できない状態である。そして、「若々しい」を分解し、畳語である「若々」としては存在せず、単純語として認定し、短単位検索で検索することが可能である。

従って、今回は「ＸＸと」と「ＸＸな」を同一の抽出方法を用い、辞書から畳語を抽出したうえで、BCCWJ から後接構造のバリエーションを確認する抽出手法を取る。「若々しい」は「ジャパンナレッジ Lib」を使い、条件を設定したうえで反復形容詞を抽出する。辞書の抽出やコーパスの条件設定やデータ化など詳しくは次節で述べる。

３.２.１ 辞書からの抽出

本研究では、日常生活に必要な語をはじめ、科学技術・情報・医学などの最新語が収載されている『旺文社国語辞典　第十一版』を用い、現代日本語における畳語である「ＸＸ」の数を抽出する。国語辞典としては初めて、多義語にその語の核となる意味〈中心義〉が付されており、大学生、高校生向けだけではなく、一般向けでであることで、『旺文社国語辞典　第十一版』を選んだ。

今回は漢字と仮名で表示されている「ＸＸ」の構造を持つ 287 語を抽出した。各語の品詞や意味などを含め、リスト化する。次に、BCCWJ を使い、現代日本語の書き言葉として使用されている語を抽出する。

３.２.２ コーパスからの抽出

3.2.1 の辞書からの抽出は現代語をベースに行われているが、本研究の目的としている畳語の構造と意味と文法機能を分析するためには、数多くの用例から分析を行うことが求められる。そこで、本研究では国立国語研究所から公開されている BCCWJ を使用する。BCCWJ は、書籍全般、新聞、ネット掲示板、教科書などのジャンルにまたがって１億 430 万語のデータを対象とした書き言葉均衡コーパスである。現在、日本語について入手可能な唯一の均衡コーパスだと言われている。このように多様な言語資料の性質を持ち、近代語の資料として代表性を有すると考える。

『旺文社国語辞典　第十一版』から抽出した 287 語のうち、14 語[3]はヒットされていない

[3] BCCWJ では、えんえん(炎炎)、しょしょ(処処)、そぼそぼ、たんたん、ちゃくちゃく(嫡嫡)、ちんちん、ちんちん(沈沈)、てきてき(滴滴)、なおなお(尚尚)、はらはら、ひょうひょう、ぼうぼう(某某)、れんれん(連連)、れんれん(漣漣)の畳語が見つからない。

ため、調査対象は合計 273 語となる。273 語の使用用法と使用頻度を収集するため、BCCWJ の文字列検索と短単位検索を利用する。具体的な手順は下記の通りである。「ＸＸ」の後接構造、語彙素の表記、品詞などを確認するため、文字列検索では短単位、長単位の区切りを意識せずにコーパス全体のデータを検索した。検索した結果は 273 語の「ＸＸ」のうち、形容動詞として使われる「ＸＸな」は 14 語であり、副詞として使われる「ＸＸと」は 209 語である。

次に、短単位検索を利用し、確認した語彙素表記の「ＸＸ」を入力し、キー後方に現れる１語の距離を条件と設定した。「ＸＸと」の場合は図１のように、後方共起キーからのところが「語彙素が『と』」かつ「品詞の中分類が『助詞-格助詞』」を指定し、各畳語の使用頻度を検索する。

図１　「ＸＸと」の抽出方法

「ＸＸな」の場合は図２のように、後方共起キーからのところが「語彙素が『だ』」かつ「品詞の小分類が『連体形-一般』」を指定し、各畳語の使用頻度を検索する。

図２　「ＸＸな」の抽出方法

抽出した「ＸＸと」、「ＸＸな」の使用頻度を調査し、最上位５語を分析対象とし、結果は表１と表２にまとめる。

表１　使用頻度が高い５語「ＸＸと」

	平仮名	漢字	数
1	いろいろと	色々と	2820
2	つぎつぎと	次々と	2281
3	どうどうと	堂々と	1201
4	いきいきと	生き生き・活き活きと	849
5	ひとびとと	人々と	675

表１に示すように、副詞として使われる「ＸＸと」を使用頻度の高い順に並べると、「色々と」、「次々と」、「堂々と」、「生き生き・活き活きと」、「人々と」となった。

表２　使用頻度が高い５語「ＸＸな」

	平仮名	漢字	数
1	さまざまな	様々な	13455
2	いろいろな	色々な	9714
3	さんざんな	散々な	94
4	しゅじゅな	種々な	49
5	しんしんな	津々な	14

表２に示すように、形容動詞として使われる「ＸＸな」を使用頻度の高い順に並べると、「様々な」、「色々な」、「散々な」、「種々な」、「津々な」となった。

表１と表２から分かるように、畳語である「色々」は副詞と形容動詞と両方として使われる。副詞として使われる「次々」、「堂々」などの畳語は形容動詞として使われるかは疑問であり、使用されているかを検証する必要がある。同様に、形容動詞として使われる「様々」、「散々」などの畳語は副詞として使われるかは疑問であり、使用されているかを検証する必要がある。よって、本研究は畳語による生産性が自由ではなく、構造に偏りがあると考える。

３．２．３　ジャパンナレッジLibからの抽出

「ジャパンナレッジLIB」とは、約50種類の辞事典、叢書、雑誌が検索できる国内最大級のデジタル辞書・事典サイトである。基本検索以外に、詳細（個別）検索も可能であり、各辞事典に応じたオリジナルな絞り込み機能（ファセット）を加えることによって、細かく考察することができる。

今回は現代日本語における「ＸＸしい」を網羅するため、詳細検索のコンテンツのところ「デジタル大辞泉」、『日本国語大辞典　第二版』を用いる。そして、辞書の見出し語の検索方法では、後方一致検索である「しい」、かつ全文検索である「形容詞」という条件を設定する。抽出した語のうち、あいきょうらしい(愛嬌らしい)、あやしい(怪しい)などのように「ＸＸしい」構造を持たない語を除き、調査対象になるのは101語である。抽出した「ＸＸしい」の使用頻度を調査し、最上位５語を分析対象とし、結果は表３にまとめる。

表３　使用頻度が高い５語「ＸＸしい」

	平仮名	漢字	数
1	ばかばかしい	馬鹿馬鹿しい	493
2	なまなましい	生々しい	467
3	すがすがしい	清々しい	373
4	わかわかしい	若々しい	349
5	みずみずしい	瑞々しい・水々しい	333

表３に示すように、反復形容詞として使われる「ＸＸしい」使用頻度の高い順に並べると、「馬鹿馬鹿しい」、「生々しい」、「清々しい」、「若々しい」、「瑞々しい・水々しい」となった。

４．結果と考察

BCCWJ を使用し、「ＸＸと」、「ＸＸな」、「ＸＸしい」という３種類それぞれの共起関係や文法的振る舞いを明らかにする。また、構造上においては「色々な、*色々しい、色々(と)」、「*軽々な、軽々しい、軽々(と)」のように、３種類における「ＸＸ」の使い分けがあると見られ、３種類の成立をグルーピングするかも課題として分析した。以上の考察から、畳語による生産性の理解の助けになることを期待している。

４．１「ＸＸと」

副詞用法として使用される「ＸＸと」の構造や文法機能などの特徴を明らかにするため、まず、どのような畳語が格助詞「と」と後続するか、畳語の派生元である「Ｘ」の品詞から考察してみる。以下では、現代日本語の「ＸＸと」を派生元の品詞別に５つのカテゴリーに分類している。

A.　派生元が名詞あるいは形容動詞であるもの。例：多多と(多)、嫌々と(嫌) …。
B.　派生元が動詞であるもの。例：生き生きと(生きる)、追い追いと(追う)…。
C.　派生元が形容詞であるもの。例：薄々と(薄し)、長々と(長し)…。
D.　派生元が副詞であるもの。例：すらすらと(すらり)、ちらちらと(ちらり)… 。

畳語の派生元である「Ｘ」の品詞は、名詞あるいは形容動詞であるもの、動詞であるもの、形容詞であるもの、副詞であるものが見られる。これらの畳語は副詞として使われる用法が使用されるかを検証するため、BCCWJ を用い、それぞれ形容動詞の「ＸＸな」と形容詞の「ＸＸしい」の用法があるかを調べる。

まず、畳語の後接は「と」を「な」に変え、「色々な」、「次々な」、「堂々な」、「生き生き・活き活きな」、「人々な」の用例を検索してみる。「色々な」は最も使用され、「堂々な」と「人々[4]」は１件しかなく、「次々な」、「生き生きな・活き活きな」の用例は１件もないことが分かった。以下では「堂々な」と「人々な」それぞれの１例を挙げる。

(1)カン・ドンウォンは『威風堂々(な)彼女』(十七．七%) が最高のヒット作。
Yahoo!ブログ(2008)
(2)補助的な仕事をする女の子正社員。すぐ結婚して辞めちゃう人々(な)イメージ。
Yahoo!知恵袋(2005)

次に、畳語の後接は「と」を「しい」に変えてみる。辞書では「諄々(くどくど)」、「細々

[4] 「人々な」を検索したところ 32 例であるが、「やはり島国の人々(な)のだ。」のような連体修飾の用法ではない例が多い。32 例のうち、31 例は連体修飾用法から除く。

(こまごま)」、「ずかずか」は「副詞」として記述されているが、BCCWJ の実際の用例である「くどくどしい」のような反復形容詞としての用法が見出される。今回は、「諄々(くどくど)しい」、「細々(こまごま)しい」、「ずかずかしい」、「どくどくしい」、「長々(ながなが)しい」、「ぶくぶくしい」、「麗々(れいれい)しい」という７語の反復形容詞が使われていることが分かった。

４．２「ＸＸな」

形容動詞として使われる「ＸＸな」の構造や文法機能などの特徴を明らかするため、まず、どのような畳語が連体活用形「だ」と後続するか、畳語の派生元である「X」の品詞から考察してみる。以下では、現代日本語の「ＸＸな」を派生元の品詞別に３つのカテゴリーに分類している。

A. 派生元が名詞あるいは形容動詞であるもの。例：色々な(色)、様々な(様) …。
B. 派生元が形容詞であるもの。例：熱々な(熱し)、深々な(深し)…。
C. 派生元が副詞であるもの。例：さらさらな(さらり)、もぞもぞな(もぞり)…。

畳語の派生元である「X」の品詞は、名詞あるいは形容動詞であるもの、形容詞であるもの、副詞であるものが見られる。副詞と異なり、形容動詞として使われる畳語は動詞からなるものが見られなかった。そこで、形容動詞として使われる用法がより使用されたかを予測する。以下では、BCCWJ を用い、形容動詞として使われる「ＸＸ」は副詞の用例「ＸＸと」または形容詞の用法「ＸＸしい」を持つかを調べてみる。

まず、形容動詞として使われる「ＸＸ」は副詞の用例「ＸＸと」との関連について述べる。辞書では、形容動詞として使われる「様様(さまざま)」、「深深(しんしん)」などの品詞が「形動タリ」あるいは「形動ダ」のみとして記述されているが、BCCWJ の実際の用例である「様様と」のような副詞的な用法が見出される。他にも、「色々(いろいろ)」、「散々(さんざん)」、「種々(しゅじゅ)」、「隆々(りゅうりゅう)」、「ぐちゃぐちゃ」、「さらさら」、「すかすか」、「ほくほく」、「もぞもぞ」は形容動詞として「ＸＸな」、副詞として「ＸＸと」の構造を両方持っている。つまり、実例から畳語後ろに「-な」を付け加えることによって、新しい語・用法として形成することが可能であり、生産性がある。

形容動詞と副詞両方として使われる「色々」を除き、「様々」、「散々」、「種々」の構造及び文法機能は使用されているかを検証してみる。結果としては、「種々と」(4 例)、「様々と」(３例)、「散々と」(２例)が見出される。以下ではそれぞれ１例を挙げる。

(3)この裏が、ほら、シーボルト先生も、種々(と)異国の花をお育てになられた、お花畠ですのよ。

栗田勇(2001)『漂民』

(4)千代子は様々(と)政男の悪口を言ひ、自分の立場を擁護しているのだった。

沢憲一郎(1994)『二人の千代子』

(5)なぜかといえば、物件にさんざん(と)文句をつけたあげく、大したものを買わないケチな香港人と違って

莫邦富(1995)『十二億人市場を狙え』

(6)はっきりとお断り申して居きます。種々(と)お話もつきませんが、この辺で幕に致しま志う。

北村銀太郎(2001)『聞書き・寄席末広亭』

次に、形容動詞として使われる「ＸＸ」は形容詞の用法「ＸＸしい」との関連について述べる。BCCWJ を検索してみると、同じ畳語構造を持つ場合には、「ＸＸな」と「ＸＸしい」を両方に持つ語は１例もない。以上の考察から、副詞としての用法「ＸＸと」に比べ、形容動詞として使われる用法「ＸＸな」がより使用されやすいと考えられる。

４．３「ＸＸしい」

反復形容詞として使われる「ＸＸしい」の構造や文法機能などの特徴を明らかするため、荒川(2006)の研究が挙げられる。荒川(2006)では、現代日本語の重複形容詞を派生元の品詞別に３つのカテゴリーに分類している。

A. 派生元が名詞あるいは形容動詞であるもの。例:初々しい(ういういしい)、清々しい(すがすがしい) …。
B. 派生元が動詞であるもの。例:忌々しい(いまいましい)、おどろおどろしい(おどろく) …。
C. 派生元が形容詞であるもの。例:痛々しい(痛し)、軽々しい(軽し)…。

畳語の派生元である「Ｘ」の品詞は、名詞あるいは形容動詞であるもの、動詞であるもの、副詞であるものが見られる。これらの畳語は反復形容詞として使われる用法が使用されるかを検証するため、BCCWJ を用い、それぞれ副詞の「ＸＸと」と形容動詞の「ＸＸな」の用法があるかを調べる。

まず、上位５語の反復形容詞である「馬鹿馬鹿しい」、「生々しい」、「清々しい」、「若々しい」、「瑞々しい・水々しい」の後接は「しい」を助動詞の「と」に変えてみると、「清々(せいせい)と」と「生々と」の使用が見られる。

(7)牢囚の身にあって、ベルは闇の中にいた。清々(と)した、静かな闇だ。
冲方丁(2000)『ばいばい、アース』
(8)男の子はひとり車体の柱を握つて、その生々(と)した眼で野の中を見続けた。
横光利一(2002)『編年体大正文学全集』

次に、畳語の後接は「しい」を「な」に変えてみる。BCCWJ では、同じ畳語構造を持つ場合には、「ＸＸしい」と「ＸＸな」を両方に持つ語は１例もないことが分かった。一般的には、形容詞として使われる「～い」か「～な」どちらかの形で捉え、それで反復形容詞「ＸＸしい」は形容動詞「ＸＸな」の形も持つのが困難であると考える。

５．おわりに

以上、辞書とコーパスを用いた考察の結果、畳語の構造を中心に派生元の品詞だけではなく、後接の用法「ＸＸと・ＸＸな・ＸＸしい」が絡み合って存在していることが明らかになった。派生元の品詞から見ると、「ＸＸと」の派生元である「Ｘ」の品詞は、名詞あるいは形容動詞、動詞、形容詞、副詞であるものがあり、バリエーションが一番豊かであることが分かった。また、BCCWJ を使用し、副詞としての畳語は形容動詞としても反復形容詞としても使われている用例が見られる。それは、副詞としての畳語は「ＸＸと」と接続することが最も多いが、「ＸＸな」や「ＸＸしい」の用例も見られることから、使用用法のうち自由度が最も高いと考えられる。

本調査を通し、畳語には生産性が限られているが、後ろに「-な」、「-しい」などを付け加えることによって、新しい語として形成することが可能である。今回は３種類それぞれの最高使用頻度の５語のみを研究対象としたが、今後の課題としては、対象または用例数が少ないため、さらに今回の傾向性を検証する必要がある。また、意味について詳しく言及するには至っていないため、３種類それぞれ元の畳語・派生元との関わりについて論じる。

謝　辞

本研究は、日頃から私の研究指針と研究方法について多大なご指導およびご助言をいただいた小野正樹教授に深く感謝の意を表します。本研究で参考になった畳語の語彙表の作成に際し、国立国語研究所の浅原氏にご協力頂いた。記して感謝申し上げる。

文　献

荒川洋平（2006）「認知意味論に基づく重複形容詞の分析」『高見澤孟先生古希記念論文集』, pp. 71-91, 凡人社.

飯田寿子(2005)「形容詞性構成要素からなる重複形容詞について―構成要素の特質をめぐって―」『国語学研究』44, pp. 80-92.

石川慎一郎(2017)「X々型畳語の構造・使用・意味特性―「現代日本語書き言葉均衡コーパス」を用いた計量的調査―」『統計数理研究所共同研究レポート』373/374, pp. 55-74.

小野尚之(2015)「構文的重複語形成-『女の子女の子した女』をめぐって」由本陽子、小野尚之（編）『語彙意味論の新たな可能性を探って』, pp. 463-489, 開拓社.

黄慧(2009)「日本語のオノマトペに後続する助詞について:『と』および『に』をめぐって」『コーパスに基づく言語学研究報告』1, pp. 267-285.

晋栄和(1995)「現代語畳語形容詞の語構造について:『転成』との関連をめぐって」『東北大学文学部日本語学科論集』5, pp. 49-60.

玉村文郎(1975)「和語は造語力が弱いか」波多野完治・西尾寅弥（編）『現代日本語の単語と文字』, pp. 121-146. 汐文社

禹昊穎(2015)「畳語の諸機能」『学習院大学人文科学論集』24, pp. 25-57.

辞　書

山口明穂、和田利政、池田和臣(2013)『旺文社国語辞典　第十一版』旺文社

国立国語研究所編(2004)『分類語彙表　増補改訂版』東京:大日本図書

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/

コーパス検索ツール『NINJAL-LWP for BCCWJ (NLB)』　http://nlb.ninjal.ac.jp/

オンライン辞書・事典サイト『ジャパンナレッジLIB』　https://japanknowledge.com

ニュースを対象にした手話マルチメディアコーパスの構築

加藤 直人（NHK 放送技術研究所）†
内田 翼 （NHK 放送技術研究所）
東 真希子（NHK 放送技術研究所）
梅田 修一（NHK 放送技術研究所）

Multimedia Corpus of Japanese Sign Language for TV News

Naoto Kato (NHK Science and Technology Laboratories)
Tsubasa Uchida (NHK Science and Technology Laboratories)
Makiko Azuma (NHK Science and Technology Laboratories)
Shuichi Umeda (NHK Science and Technology Laboratories)

要旨

ＮＨＫ・Ｅテレで放送されている手話ニュースをデータベース化した，手話マルチメディアコーパスについて述べる．ニュースは構文の逸脱も少ないため，手話の言語研究をする上で比較的扱いやすい対象である．しかし，ニュースを対象にした大規模な手話コーパスはない．我々が現在構築している手話マルチメディアコーパスは，2018 年 3 月末現在で，約 15 万文（延語数 3,036,000 語，異なり語数 76,000 語）を有し，手話コーパスでは大規模なものである．手話では手や指の動作である手指動作とともに，それ以外の動作である非手指動作も重要な情報を持つことが指摘されているので，コーパスの一部には代表的な非手指動作である顔情報（顔表情と口型）も書き起こしている．本稿では，顔情報に関する統計的な分析結果についても報告する．

１．はじめに

手話ニュースを対象にした手話マルチメディアコーパスの構築を進めている[1][2]．この最大の目的は，日本語から手話への機械翻訳の研究に資するためである．手話は，先天的あるいは幼少時に失聴した聾者にとって母語であり，日本語より理解しやすい．しかし，手話による情報提供はまだ少ない．これは，日本語から手話への“翻訳”という作業を伴うからである．手話は日本語とは異なる言語であるため，英語への翻訳のように手話への翻訳が必要となる．その一つの解決方法が機械翻訳である．現在の機械翻訳の研究はコーパスベースの手法が主流となっている．コーパスベース機械翻訳では，２言語間の大規模の対訳コーパスから機械学習することで翻訳知識を自動獲得しているので，翻訳対象としている言語の知識にあまり立ち入ることなく，システムを開発できるという利点がある．日本手話は文法などがまだ十分に解明されていないので，機械翻訳にはコーパスベース手法が適していると考えた．実際にこれまで構築してきた日本語と手話の対訳コーパスを利用して，日本語テキストから手話 CG（Computer Graphics）へ自動的に機械翻訳するシステムを開発した[3][4][5]．翻訳精度向上のためには対訳コーパスの拡大が必要となるので，現在もコーパス構築を継続している．

本稿では，ＮＨＫの手話ニュースをデータベース化した，手話マルチメディアコーパスについて述べる．手話マルチメディアコーパスは日本語書き起こし，手話書き起こし，手話映像から構成される．さらに，手話書き起こしの一部のデータでは，非手指動作の一つである顔情報（顔表情と口型）も書き起こしている．本稿では，顔情報に関して統計的な分析結果についても報告する．

† katou.n-ga@nhk.or.jp

２．手話マルチメディアメディアコーパス

２.１　ＮＨＫ手話ニュース

手話マルチメディアコーパスが対象としているのはＮＨＫ・Ｅテレで放送されている，手話ニュース，手話ニュース 845，週間手話ニュースの３つの手話ニュース番組である（以下では，３つの番組をまとめて「手話ニュース」と呼ぶ）．これらの番組にはアナウンサーの音声とルビ付きの字幕[1]も付いている．手話ニュースでは，１番組に１人～２人の手話キャスターが手話でニュースを伝えるとともに，同時にアナウンサーによる日本語音声や字幕でも情報を伝える．また，手話ニュースでは，VTR の取材映像が流れているときには手話が付かず音声と字幕だけであるため，手話文数は多くない．表 1 に手話マルチメディアコーパスが対象としている手話ニュースの情報を示す．

表 1 手話マルチメディアコーパスの対象とした番組

番組名	放送時間	手話文数（平均）[2]
手話ニュース	月曜～金曜　13:00-13:05（ 5 分） 土曜・日曜　19:55-20:00（ 5 分）	16.7
手話ニュース 845	月曜～金曜　20:45-21:00（15 分）	35.6
週間手話ニュース	土曜　11:40-12:00（20 分）	49.0

手話ニュースは毎日放送されているので，データ収集は容易である．また，ニュースを対象としているため，会話とは異なり構文の逸脱も少ないと考えた．更に，放送するまでに複数人のチェックが入るので，そこで使われる言語表現の普遍性は高いと考えた．また，最新の話題も扱われているので，新しい語彙も出現する．手話ニュースには手話キャスターが総勢十数人ほどおり，これまで聾者，CODA（Children of Deaf Adults）[3]，手話通訳士が担当していたが，現在は聾者が圧倒的に多い．

手話マルチメディアコーパスの構築は 2009 年 4 月放送分の手話ニュースから開始した．データベース化にあたっては，作業者は手話ニュースを見て，人手で文単位に日本語書き起こし，手話映像の切り出し，手話書き起こしを行う．日本語書き起こしと手話映像の切り出しは日々行っているが，手話動作の書き起こしは古いデータから順に行っている．2018 年 3 月末現在，手話動作の書き起こしは 2017 年 1 月までの手話ニュースで完了している．その結果，2009 年 4 月～2017 年 1 月までの日本語と手話の対訳文ペアは約 15 万文に達し，手話単語数は延べ約 3,036,000 語，異なりで約 76,000 語に至っている．手話文は総勢 18 人の手話キャスターによるものであり， 61%が聾者（12 人）， 13%が CODA （2 人），26%が手話通訳士（4 人）によるものである．

２.２　コーパス構築

手話マルチメディアコーパスでは，1 文ごとに日本語，映像，手話をまとめ，データベース化している．日本語はアナウンサーが話した言葉をそのまま書き起こしている．手話は単語ごとに書き起こし，その表記には日本語ラベル[4]による gloss 表記を使用している．例え

[1] ここでいう「字幕」はオープンキャプションのことである．台詞を文字化しているクローズドキャプションではない．
[2] 2009 年 4 月に放送された手話ニュースの平均である．
[3] 聾者の親を持つ健聴者．
[4] 日本語ラベルとは，意味的に日本語単語を使い，日本語と区別できる表記法（例えば，括弧{}）を付けて手話単語を表記する方法．

ば，手話単語“{温度}”は意味的に近い日本語単語「温度」を括弧{}で括って表している．gloss 表記は単語ごとに表記されるので機械翻訳で扱いやすい．映像は手話の一文ごとに切り出している．

手話書き起こしは，キャスターの手話映像を見て日本語ラベルを人手で付けるという作業になる．日本語ラベルは，全日本聾唖(ろうあ)連盟が発行している「日本語－手話辞典」[6][7]の定義にしたがった．

手話には音声言語にはない独特の言語的特徴があり，手話書き起こしの際に問題となる．手話の言語的な特徴の代表的な例をいくつか挙げる．

（ⅰ）手話の文末を特定することが難しい．

日本語の音声言語では文末を表す助動詞があるため比較的文末を特定しやすいが，手話には助動詞のような表現がほとんどないため文末を特定することが難しい．

（ⅱ）手話では手指動作と非手指動作を使って言語表現を行う．

手指動作とは手や指の動きである．一方，非手指動作とはそれ以外の身体の動きである．代表的な非手指動作に頷き，顔情報（顔表情，口型）がある．非手指動作には言語の意味や文法的な役割を担っているものがある．例えば，日本語の「厳しい冬」は，手話では手指動作で「冬」を表現し，同時に，厳しい顔表情（非手指動作）をして「厳しい冬」を表現する場合がある．

（ⅲ）右手と左手で別々の語彙を表すことがある．

例えば，「５人」を手話では左手で「５」，右手で「人」を表して表現する．

（ⅳ）手話には固定語彙（Frozen Lexicon）の他に，その変形がある．

例えば，手話で日本語の「座る」は固定語彙であり右手（利き手）１つで表すが，左手でも同じ手型を同時にするという『変形』を行うことで「2 人で並んで座る」という意味になる．

このような特徴は手話を書き起こす際に問題となる．これらの問題を全て解決してから手話を書き起こすことは困難であるため，現在は第１次近似として手話書き起こしを以下のようにしている．

（ⅰ）手話ニュースでは手話を行わないときには両手を前に重ねて置く．

この位置をホームポジションと呼ぶ．手話の１文はホームポジションからホームポジションまでの動作と定義した．なお，手話の１文は必ずしも日本語の１文には対応しない．

（ⅱ）手話の書き起こしは手指動作を基本とし，非手指動作の書き起こしは当初頷きに限定した．頷きは比較的容易に特定できる非手指動作であり，言語的にも句や文の境界として重要なためである．しかし，手話翻訳の精度向上には顔情報が不可欠であることがわかったため[8]，一部のデータに対しては顔情報の書き起こしを始めている．

（ⅲ）左手と右手で別々の語彙を表すときは１つの語彙として記述した．

例えば，手話キャスターが左手で「５」，右手で「人」の動作をして「５人」と表現した場合には，“Ｌ：{５} Ｒ：{人}”と記述した．ここで，Ｌは左手の動作を，Ｒは右手の動作を表す．

（ⅳ）固定語彙で記述できない手話はその動作を日本語で説明し，“[ ]”を付けた．

例えば，手話キャスターが固定語彙にはない「ミサイルが飛来する」動作をしたときには，“[ミサイルが飛来する様子]”と記述した．また，固定語彙においても説明が必要な場合には，例えば，手話キャスターが “{みんな}”という手指動作を前後の単語とスムーズにつながるように変形させた場合には，“{みんな} [変形]”のように記述した．

2.3　手話マルチメディアコーパスの表示システム

手話ニュースコーパスの管理・検索を行う表示システムを開発した（図 1）．表示システムの画面は検索キーワード入力部，ニュース情報出力部，映像出力部，対訳出力部から成る．

検索キーワード入力部で検索したい単語（キーワード）を入力する．図 1 の例では「インフルエンザ」と入力している．

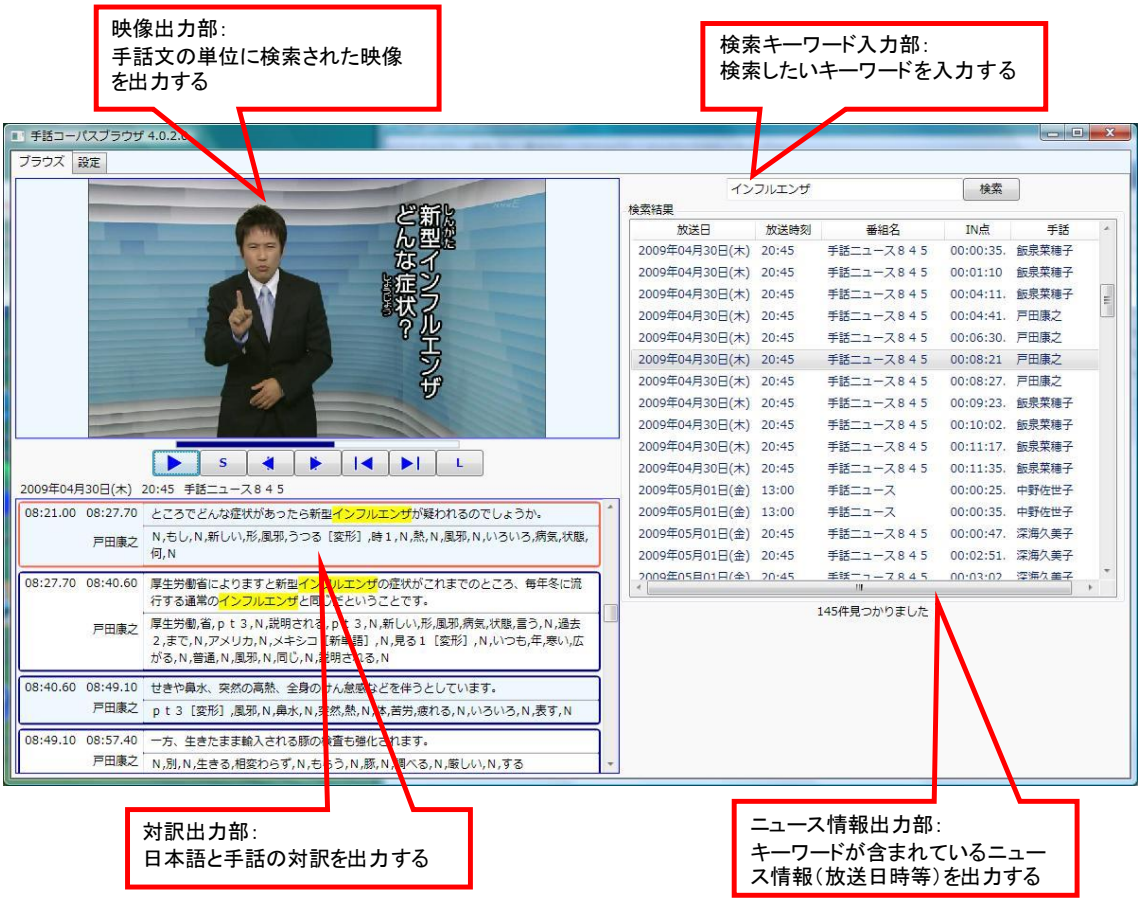


図 1 手話マルチメディアコーパスの表示システム（参考文献[2]の 1 図から）

ニュース情報出力部ではキーワードが含まれている文のニュース情報を出力する．ニュース情報には，その番組が放送された日時，番組名，その文の開始時刻と手話キャスターの名前が表示される．このような情報を提示することで，手話キャスターの違いによって手話がどのように異なるかなどを分析することが可能となる．

映像出力部ではキーワードを含む手話映像を 1 文単位で出力する． 1 文単位の出力なので映像の中からキーワードに対応する映像を探す必要があるが，映像を探しやすくするために映像をスローやコマ送りで再生できるようにしている．また，システムは画面拡大機能を有しており，顔の表情や指の動きなど細かい部分を拡大して見ることもできる．

対訳出力部ではアナウンサーが話している日本語とその手話の書き起こしを対訳の形で出力する．ただし，手話書き起こしでは表示の煩わしさを避けるために，｛｝を省略している．出力された対訳を参照することで，手話書き起こしを見るだけで，どのような手話単語が使われているのかを調べることができる．

開発した表示システムは自動翻訳の研究用としてだけでなく実用的な利用も考えられる．例えば，過去のニュースで用いられた手話を検索できる翻訳メモリー（Translation Memory）としての利用である．ニュースでは固有名詞や専門用語がよく出てくるが，これらの手話翻訳を決めるのには多大な労力が必要である．開発したシステムで過去の対訳用例を検索し活用することで，翻訳作業の効率化が期待できる．実際，手話マルチメディアコーパスを手話ニュースの現場からもアクセスできるようにしたところ，大変好評である．

３．顔情報

手話では手指動作とともに顔情報が情報伝達のために重要である．顔情報には大きく分けて顔表情と口型があるが，顔表情はその書き起こし基準を定義することが非常に難しいこともあり，口型とともに書き起こしの対象としていなかった．しかし，翻訳システムの精度向上には顔情報が不可欠なことがわかったため，コーパスの一部に対して顔情報の書き起こしを開始した．顔情報を含む書き起こしの例を図２に示す．

日本語：気温が30度を超える真夏日になったところも出ています
手話：{温度}{30}{温度が上がる}{本番}{暑い1}[NMM]{日}{変わる1}{場所}{ある1} N
口型：キオン サンジュウ NULL マ ナツ ビ NULL NULL パ NULL

図２ 日本語，手話（＋顔表情），口型の書き起こしの例

図２の手話書き起こしをみると，手話単語“{暑い１}”に顔表情が付いていることを表す記号“[NMM]”が追加されている．また，口型の行が追加され，例えば，「キオン」のように口型が書き起こされている．

今回，2013年９月から2014 年２月までの手話ニュースを対象に顔情報を書き起こして，統計分析を行った．その手話キャスターは聾者が 11 名，CODA が１名，手話通訳士が３名の計 15 名である．次節以降では，顔表情と口型に対する分析結果について述べる．

３．１　顔表情

顔表情を書き起こすためには，あらかじめ顔表情のラベルを定義しておく必要がある．しかし，日本手話の顔表情は，網羅的な分析が未だ行われていない．したがって，顔表情を書き起こすためにあらかじめ顔表情のラベルを定義しておくことは難しい．さらに顔表情には強弱もあるため，どの部分に顔の表情があるのかを判断することも難しい．そこで，現在，顔表情は，顔表情が明確に出ている判断できる単語やわたり[5]部分のみを書き起こしの対象として顔表情の記号“[NMM]”を付与した[6]．例えば，手話単語“{暑い１}”に明確な顔表情が付いている場合には，顔表情の記号“[NMM]”を手話単語“{暑い１}”に付加して，“{暑い１}[NMM]”と書き起こした．このような顔情報の書き起こし作業は，CODA１名，聾者１名の計２名で行っている．

2013 年９月から 2014 年２月までの手話ニュースに対して，顔表情の書き起こしを行った．その顔表情の書き起こしデータに対して，まず，手話キャスター別に顔表情（[NMM]）が付いた文の頻度を計算した．その結果を表２に示す．

[5] 手話単語から手話単語への遷移動作．
[6] NMM(Non-Manual Movement)

表 2 手話キャスター別の顔表情（[NMM]）の割合

手話キャスター	顔表情の割合
聾者	
A（女性）	25.7 (141/548)
B（女性）	6.9 (45/652)
C（女性）	9.4 (17/181)
D（女性）	4.5 (35/776)
E（女性）	13.2 (67/488)
F（男性）	11.8 (61/517)
G（男性）	12.3 (119/971)
H（男性）	18.2 (97/534)
I（男性）	8.6 (64/743)
J（男性）	11.4 (56/490)
K（男性）	22.0 (124/564)
CODA	
L（女性）	4.1 (38/916)
手話通訳士	
M（女性）	3.3 (17/181)
N（女性）	3.2 (35/776)
O（女性）	3.4 (67/488)

表 2 を見ると，やはり聾者のほうが CODA や手話通訳士よりも顔表情が付いている割合が高いことがわかる．しかし，聾者だけに焦点をあててみると，顔表情がついている割合が 4.5%～25.7%と手話キャスターによって大きな開きがあることがわかる．

次に手話キャスターの属性にかかわらず，単語ごとに顔表情が付いている頻度を計算した．単語の上位 10 語を表 3 に示す．ここで，“pt3”は指さしを表す．

表 3 顔表情（[NMM]）が付いた手話単語上位 10 語

頻度	手話単語
80	とても
74	雪
67	pt3
59	何
54	寒い
47	強力
39	風2
34	混乱
26	意味
22	最高

表 3 を見ると，“とても”，“強力”，“最高”のような程度を表す語が多いことがわかる．手話では，手指動作だけでなく顔表情でも程度を強調することにより，意味を伝わりやすくしているのであろう．

３．２　口型

まず，手話における口型（mouth pattern）について説明する．口型とは口の動きではあるが，手話通訳士が手話をしながら日本語を話すという口の動きではない．

口型には音声言語由来のマウジング（mouthing）と手話独自のマウスジェスチャー（mouth gesture）がある[9][10][11]．マウジングとは音声言語（本稿では日本語）借用の口の動きである．名詞と一緒に表出される場合が多い．名詞の中でも固有名詞の場合にはマウジングが付きやすい．例えば，日本語の固有名詞「九州」を手話で表出する際には手指動作とともに「キュウシュウ」と口を動かす．あるいは手話単語の同音異義語を区別する場合にマウジングを使う．例えば，日本語の「南」，「暑い」，「うちわ」は，手話では手話単語{暑い１}の一語で表すので手指動作（うちわで扇ぐような動作）では日本語のその３つの意味を区別できないが，「ミナミ」，「アツイ」，「ウチワ」と口を動かすことによりその意味の区

別をしている．一方，マウスジェスチャーは音声言語とは無関係な口の動きである．動詞と一緒に表出される場合が多く，副詞的意味やアスペクトを付加する役割がある．例えば，マウスジェスチャー“パ”が手話単語“{言う}”とともに用いられると，日本語の「言った」という意味になる．

2013年9月から2014 年2月までの手話ニュースに対して口型を書き起こした．ただし，口型の書き起こしは聾者の手話キャスターのみを対象としている．口型の書き起こしは顔情報の書き起こしと同じ作業者である，CODA１名，聾者１名の計２名で行っている．後者の聾者も読話能力が高いので，両者とも口型の認識力は高く，口型の書き起こし作業には問題ないと考えている．

その結果，日本語，手話，口型の３つがセットとなったものは6,369 文，手話単語総数は100,270 語が得られた．

口型の分析は「口型のみ」，「手話－口型」，「日本語－手話－口型」の３つの観点から行った．ただし，手話ニュースコーパスでは，N（頷き）や首振りなど非手指動作の一部も手話単語と同様に書き起こしているが，今回の分析からは除いた．以下ではそれぞれについて述べる．

（１）口型のみ

まず，単純に口型のみで頻度を計算した．表 4 にその上位 10 位を示す．ここで，口型の「NULL」は口型が付かない場合を表す．

表 4 口型の上位 10 位

頻度	口型
38869	NULL
2448	mm
2107	パ
923	オ
774	ハ
593	アル
558	カイ
542	トウ
503	キョウ
494	ノ

表 5 手話キャスター（聾者のみ）別の口型の割合

手話キャスター	口型の割合
A（女性）	57.7 (4286/7427)
B（女性）	55.6 (6876/12369)
C（女性）	64.1 (2017/3145)
D（女性）	71.9 (7791/10831)
E（女性）	57.7 (4559/7897)
F（男性）	47.5 (4077/8590)
G（男性）	61.7 (9311/15099)
H（男性）	67.6 (5588/8263)
I（男性）	60.6 (6933/11445)
J（男性）	68.9 (5192/7545)
K（男性）	63.0 (5291/8404)

表 4 を見ると，NULL の出現頻度が 38869 回と 2 位以下を大きく離しており，口型が付かない手話単語が多いことがわかる．しかしながら，手話単語全体の割合で見ると，口型が付く手話単語の割合は 61.2%（1－38869/100270）と，口型が付く手話単語のほうが多い．さらに，口型使用の個人差を比較するために，手話キャスターごとに口型が付く手話単語の割合を計算した．その結果を表 5 に示す．表 5 を見ると，平均値（61.2%）より若干低い手話キャスター（F，47.5%）や若干高い手話キャスター（D，71.9%）はいるものの，平均値からあまり離れてはいない．これらのことから，手話では口型もまた情報伝達の手段としてよく使われていることが確認できる．

また，上位には「mm」，「パ」，「オ」などのマウスジェスチャーが多いことがわかる．例えば，マウスジェスチャー「mm」は口を突き出してとがらした状態であり，動詞と一緒に使われ「問題なく」などの副詞的意味を動詞に付加する[12]．一方，その次に続く「アル」，「トウ」，「キョウ」などはマウジングである．例えば，「トウ」は一つには日本語名詞「党」の借用であろう．

（２）手話－口型

次に，手話－口型の 2 つ組で頻度を計算した．表 6 にその上位 10 位を示す．ここで，“ｐｔ３”は指さしを表わす.

表 6 手話－口型の頻度（全体）

頻度	手話	口型
7076	pt3	NULL
1059	{終わる}	NULL
691	{何}	NULL
605	{終わる}	パ
568	{ある1}	NULL
544	{ある1}	アル
527	{いろいろ}	NULL
461	{説明}	NULL
446	{夢2}	NULL
445	{場所}	NULL

表 7 {終わる}－口型の頻度（口型が付く場合）

頻度	手話	口型
605	{終わる}	パ
29	{終わる}	mm
19	{終わる}	オワリ
15	{終わる}	タ
10	{終わる}	オワル
9	{終わる}	シタ
8	{終わる}	ッタ
4	{終わる}	オワッタ
2	{終わる}	オワ
1	{終わる}	パパパ

表 6 を見ると，上位には口型がつかない場合（NULL）が多い．手話単語“{終わる}”の口型はマウスジェスチャーであり，手話単語“{ある１}”の口型はマウジングである．手話単語“{終わる}”は，表 6 を見ると，口型が付く場合（頻度 605 回，“パ”）と付かない場合（頻度 1,059 回，NULL）がある．そこで，手話単語“{終わる}”に伴う他の口型をみるために，頻度を計算した．その結果を表 7 に示す．表 7 を見ると，手話単語“{終わる}”の口型には，マウスジェスチャーである“パ”や“ｍｍ”がある一方で，マウジングの場合もある．そのマウジングも手話単語の日本語ラベルから派生したとみられる “オワリ”，“オワル”，“オワッタ”，“オワ”や，元の日本語文から派生したとみられる“タ”，“シタ”があることは興味深い．口型を付ける場合でもさまざまな観点から考える必要があろう．今度は逆に口型“パ”に伴われる手話単語に着目するため，手話単語－口型“パ”の頻度を計算した．結果を表 8 に示す．口型“パ”は完了を表わす助動詞的な役目があるため手話単語も動詞となることが予想されるが，表 8 を見ると，いずれも動詞的な手話単語に付いているのが確認できる．

表 8 手話－口型“パ”の頻度

頻度	手話	口型
605	{終わる}	パ
332	{た}	パ
88	{言う}	パ
53	{明るい1}	パ
43	{起きる2}	パ
37	{決める1}	パ
36	{発見}	パ
32	{爆発}	パ
19	{飛び出す}	パ
18	{た}[右手のみ]	パ

表 9 {九州}－口型の頻度

頻度	手話	口型
33	{九州1}	キュウシュウ
16	{九州2}	キュウシュウ
3	{九州}[左]	キュウシュウ
1	{九州}[新単語・左下]	キュウシュウノ
1	{九州}[新単語・左]	キュウシュウ
1	{九州1}	ク
1	{九州1}	キュウシュウモ
1	{九州1}	キュウシュウマデ
1	{九州1}	キュウシュウノ
1	{九州1}	NULL

ニュースには固有名詞が頻出するが，手話ではマウジングが付く場合が多い傾向にある．そこで，実際の傾向を調べるために固有名詞の手話単語“{九州}”と口型の頻度を計算した．

結果を表 9 に示す．表 9 を見ると，マウジングが付かなかったのは 1 つだけであり，その他の場合はマウジングが付いていた．マウジングには“キュウシュウ”が多かったが，“ク”のみの場合や，“ノ”，“モ”，“マデ”といった日本語助詞のマウジングが付いている場合もあった．

（３）日本語－手話－口型

最後に日本語－手話－口型の 3 つ組で頻度を計算した．そのためには，日本語－手話の単語対応付けをする必要がある．その対応付けは我々が開発した単語のアライメントアルゴリズム[13]によって自動的に行った．ここでは簡単に説明する．

我々のアルゴリズムでは，茶筅[14]で形態素解析をしたのち，表層的に一致する日本語単語と手話単語は前処理で対応付けておき，前処理で対応付けられなかった日本語単語と手話単語を対象としてＥＭアルゴリズム[15]によって自動対応付けをする．ただし，複合的な意味をもつ日本語単語はあらかじめ文字単位に分割するという工夫をしている．

自動アライメントにより日本語－手話の単語対応付けをし，（２）節の結果と合わせることにより日本語－手話－口型の 3 つ組を得た．その頻度計算した結果を表 10 に示す．

表 10 日本語－手話－口型の頻度

頻度	日本語(品詞)		手話	口型
1193	に	助詞-格助詞-一般	pt3	NULL
1080	た	助動詞	{終わる}	NULL
750	て	助詞-接続助詞	pt3	NULL
621	NULL	NULL	pt3	NULL
582	を	助詞-格助詞-一般	pt3	NULL
469	が	助詞-格助詞-一般	pt3	NULL
397	た	助動詞	{終わる}	パ
374	、	記号-読点	pt3	NULL
356	など	助詞-副助詞	{いろいろ}	NULL
341	きょう	名詞-副詞可能	{今1}	キョウ

表 10 を見ると，日本語－手話－口型の 3 つ組で頻度が高いのは助詞や助動詞であり，口型を伴わない場合が多い．例えば，助詞では“にｰｐｔ３－NULL”，助動詞では“たー{終わる}－NULL”である．また，日本語の助動詞「た」が手話や口型の違いから，次の 4 つの場合に手話翻訳されていた．

“たー{終わる}－NULL”
“たー{終わる}－パ”
“たー{た}－パ”
“たー{た}－NULL”

日本語の助動詞「た」を手話に翻訳する際には，手話単語の訳語選択（{終わる} or {た}）や口型選択（パ or NULL）が必要であろう．

次に日本語の名詞と口型の関係をみるために頻度統計をとった．ただし，口型が付いている場合だけを対象とし，一般名詞の場合と固有名詞の場合に分けた．一般名詞であるか固有名詞であるかは形態素解析の結果で判断している．一般名詞の場合を表 11 に，固有名詞の場合を表 12 に示す．

表 11 一般名詞－手話－口型の頻度

頻度	日本語(品詞)		手話	口型
318	手話	名詞-一般	{手話}	シュワ
282	ニュース	名詞-一般	{ニュース}	ニュース
134	会	名詞-一般	{会}	カイ
99	雪	名詞-一般	{雪}	ユキ
93	天気	名詞-一般	{空}	テンキ
86	党	名詞-一般	{党}	トウ
84	案	名詞-一般	{案}	アン

表 12 固有名詞－手話－口型の頻度

頻度	日本語(品詞)		手話	口型
116	東京	名詞-固有名詞-地域-一般	{東京}	トウキョウ
108	アメリカ	名詞-固有名詞-地域-国	{アメリカ}	アメリカ
105	庁	名詞-固有名詞-組織	{庁}	チョウ
101	日本	名詞-固有名詞-地域-国	{日本}	ニッポン
101	安倍	名詞-固有名詞-人名-姓	アベ[指文字]	アベ
92	北	名詞-固有名詞-地域-一般	{北}	ホク

表 11 をみると，一般名詞でも口型はマウジングであることが確認できる．例えば，日本語「天気」の手話訳は“{空}”であるが，口型は手話の日本語ラベルの読みである“ソラ”ではなく，日本語借用の「テンキ」である．出現頻度が 3 回以上[7]の一般名詞では約 73%の単語に口型が付いていた．

表 12 を見ると，やはり固有名詞でも口型はマウジングであることが確認できる．また，出現頻度が 3 回以上の固有名詞では約 89%の単語に口型が付いていた．これは一般名詞の場合（約 73%）より高い値であり，固有名詞には名詞以上に口型が必要であることがわかる．

4．おわりに

ＮＨＫ・Ｅテレで放送されている手話ニュースをデータベース化した，手話マルチメディアコーパスについて述べた．さらに，手話マルチメディアコーパスでは，そのコーパスの一部に顔情報（顔表情と口型）も書き起こしを行い，その統計的な分析結果についても述べた．今回の統計分析の結果から，従来から口型について定性的に分析されてきた結果が，我々のコーパスでも確認できた．また，日本語を口型付手話に翻訳する際に考慮すべき知見を得ることができた．しかしながら，日本語－手話－口型の 3 つ組に対して統計をとるにはまだコーパスサイズが小さいという問題がある．今後もさらにコーパスを拡大していきたい．

我々は手話 CG 翻訳の研究をしており，その精度向上をめざしている．特に，手話 CG のわかりやすさ，自然さを向上させるには顔情報（表情や口型）の付加が欠かせない．表情に関しては部分的に取り組んでおり，主に感情を表す 6 つの表情を手話 CG に付加した[16]．一方，口型については今回の知見を活かして研究を進めていきたい．手話ニュースコーパスを使って，どの単語に口型を入れればよいのか，どのような口型を入れればよいのかを分析する．また，口型を手話 CG で表現する研究も進めている．

[7]日本語－手話－口型の 3 つ組で低頻度の場合には，日本語の形態素解析の誤りや日本語－手話の単語対応付けの誤りが多いと考えられるため，出現頻度 3 回未満の 3 つ組は除外した．

文　献

[1] 加藤直人，“手話ニュースコーパスの構築,”言語処理学会年次大会，PA2-5，pp.494-497，2010.

[2] 加藤直人，“手話における言語資源の研究動向,” NHK 技研 R&D, No.139, pp.10-19, 2013.

[3] 加藤直人，“金子浩之，井上誠喜，梅田修一，比留間伸行，長嶋祐二，“用例利用による日本語－手話 CG 翻訳システム,”電子情報通信学会 HCG シンポジウム 2011, I-1, 2011.

[4] 加藤直人，“日本語テキストから手話 CG への翻訳技術,”NHK 技研 R&D, No.134, pp.45-52, 2012.

[5] 加藤直人，宮崎太郎，井上誠喜，梅田修一，清水俊宏，比留間伸行，長嶋祐二，“気象ニュースを対象とした手話 CG 翻訳システム,” 電子情報通信学会 HCG シンポジウム 2013，pp.433-438, 2013.

[6] 米川明彦（監修），日本手話研究所（編），“日本語-手話辞典,”（財）全日本聾唖連盟出版局, 2006.

[7] 米川明彦（監修），日本手話研究所（編），“新 日本語-手話辞典,”（財）全日本聾唖連盟出版局, 2011.

[8] 加藤直人，宮崎太郎，井上誠喜，金子浩之，比留間伸行，長嶋祐二，“気象情報を対象にした手話 CG 翻訳システムの開発とその評価,”電子情報通信学会論文誌 D, Vol.J100-D, No.2 p.217-229, 2017.

[9] Boyes Braem, Penny and Rachel Sutton-Spence(eds), “The Hands are the Head of the Mouth. The Mouth as Articulator in Sign Languages,”Signum, 2001.

[10] SLLing-Net［手話文法研究室］
http://slling.net/resources/glossary.htm#mouth_gesture

[11] 坊農真弓，“日本手話会話におけるマウジングと言い直し,” 電子情報通信学会福祉工学研究会 WIT2009-57，pp.13-18, 2015.

[12] 手習教室（第 4 講）
http://blue.ribbon.to/~korokan/jsl/4-2.html

[13] 加藤直人，宮崎太郎，“日本語-手話の単語アライメントによる非手指動作の検出,”言語処理学会年次大会，A1-1，p.286-289，2014.

[14] 茶筌
http://chasen.naist.jp/hiki/ChaSen/

[15] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation,”Computational Linguistics Vol.19,No.2, pp.263-311, 1993.

[16] 加藤直人，宮崎太郎，井上誠喜，梅田修一，東真希子，比留間伸行，長嶋祐二，“顔表情を部分的に挿入した手話 CG 翻訳システム,” 電子情報通信学会 HCG シンポジウム 2014，pp.33-38, 2014.

ベイズモデルによる方言音声共通語化過程の分析

前川 喜久雄（国立国語研究所 音声言語研究領域・コーパス開発センター）†

Bayesian Modeling of the Process of Dialect Standardization

Kikuo Maekawa (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所が山形県鶴岡市で収集した共通語化調査データのうち第 1〜3 回調査の音声項目データを用いて、方言音声共通語化過程の統計モデルを構築した。既に報告した第 1 回調査データと同様、第 2 回・第 3 回調査データも二項分布に基づくロジスティック回帰モデルを適用するには分散が大きすぎる（過分散状態）。そのため、ベルヌーイ分布の成功確率が種々の要因によって変動するベイズモデルを考案した。7 種のモデルの性能を F 値・平均予測誤差・WAIC の三者で評価した結果、回帰直線の切片が話者と語彙の要因によって変動し、傾きが語彙の要因によって変動するモデルが最良モデルとなった。このモデルの F 値は 0.95 に達しており、強い説明力を有している。さらにこのモデルにおける話者の個体性情報を「性別・言語形成地域・教育歴」の情報で置換したモデルを評価したところ、第 2・第 3 回調査データについては、最良モデルとほぼ同等の性能を発揮するものの第 1 回調査については性能がかなり低下することが判明した。

１．はじめに

国立国語研究所が 1950 年以来、ほぼ 20 年間隔で 4 回実施してきた山形県鶴岡市における社会言語学的調査（以下鶴岡調査と呼ぶ）は、方言の共通語化過程をリアルタイムで追跡したデータとして有名である（国語研 1953, 1974, 2007）。2017 年 5 月には鶴岡調査データのうち第 1〜3 回調査の音韻項目（36 項目）のデータベースが一般公開された。現在、このデータはエラー修正を施した最新版が公開されており、誰でも自由に利用することができるオープンデータとなっている（関連 URL 参照）。

以下では、一般公開されたデータを用いて、鶴岡における共通語化を統計的にモデル化することを試みる。鶴岡調査データのうち第 1 回調査の音声項目については既に報告済みであるが（前川 2017）、本稿では、第 2 回・第 3 回調査データの音声項目に分析を施し、3 回の調査で把握された共通語化プロセスの異同について議論する。

２．第１回調査音声項目の分析結果

最初に前川(2017)（以下では前報と呼ぶことにする）の成果をまとめておく。前報における成果のひとつは、第 1 回調査音韻項目データは、二値データ（方言 vs 共通語）ではあっても、二項分布には従っていないことの発見であった。音声項目は 36 個の質問からなり、質問に対する被調査者（話者）の回答は原則として共通語（0）か方言か(1)の二値データとして記録されている。しかしこれを成功確率 p、試行回数 N=36 の二項分布から生成されたデータとみなして、話者の年齢による回帰モデルを構築しても予測精度は F 値で 0.57 程度にしか達しない。その原因はデータに観察される分散が理論値 N*p*(1-p)に較べて大きすぎる（過分散）ことにあり、話者の年代別にみると、最小でも 7 倍以上、最大で 10 倍近い分散が観察された。このように顕著な過分散が生じる原因は、p の値が年齢以外の様々な要因

† kikuo@ninjal.ac.jp

の影響を受けて変動するためと考えられた。

前報では、そのようなデータに対応しうるモデルとして、個々の調査項目の成功（共通語形）と失敗（方言形）をベルヌーイ分布で予測する回帰モデルを作成した。説明変数としては、話者の年齢に加えて、音韻クラス、調査項目、話者の個体差をとりあげ、これらの変数が回帰式の切片ないし傾きに影響を及ぼすか否かを様々に組み合わせたモデルを比較検討した。最終的に選択されたのは、話者ごと・調査項目ごとに切片が変化し調査項目ごとに傾きが変化するモデルであり、F 値、平均予測誤差、WAIC のいずれに関しても最良の値を示した。このモデルの F 値は 0.823 であった。

3．今回の分析

今回の分析では、第 2 回・第 3 回調査データの音韻項目に対して前報と同様の分析を施し、その結果を比較検討する。また前回の検討では対象外とした社会的属性の影響を検討するために、話者の個体差の要因を、話者の性別、言語形成地域、教育歴の 3 要因でどの程度まで代替できるかという問題もあわせて検討する。ただしその前に第 2 回・第 3 回調査データの性格を第 1 回と比較する形で把握しておく必要がある。

3．1　第 2 回・第 3 回調査データの過分散指数

今回分析対象とするデータは、鶴岡市民を母集団として無作為抽出されたデータであり、鶴岡調査関係の文献では経年調査データと呼ばれるタイプのデータである。図 1 に第 1〜3 回の調査データの生年代ごとの分布状態をバイオリンプロットで示す。縦軸が共通語化得点（0-36）、横軸が 10 年刻みの生年代（1，2…は 10 代、20 代を表す）である。また図中の丸印は平均値を示している。

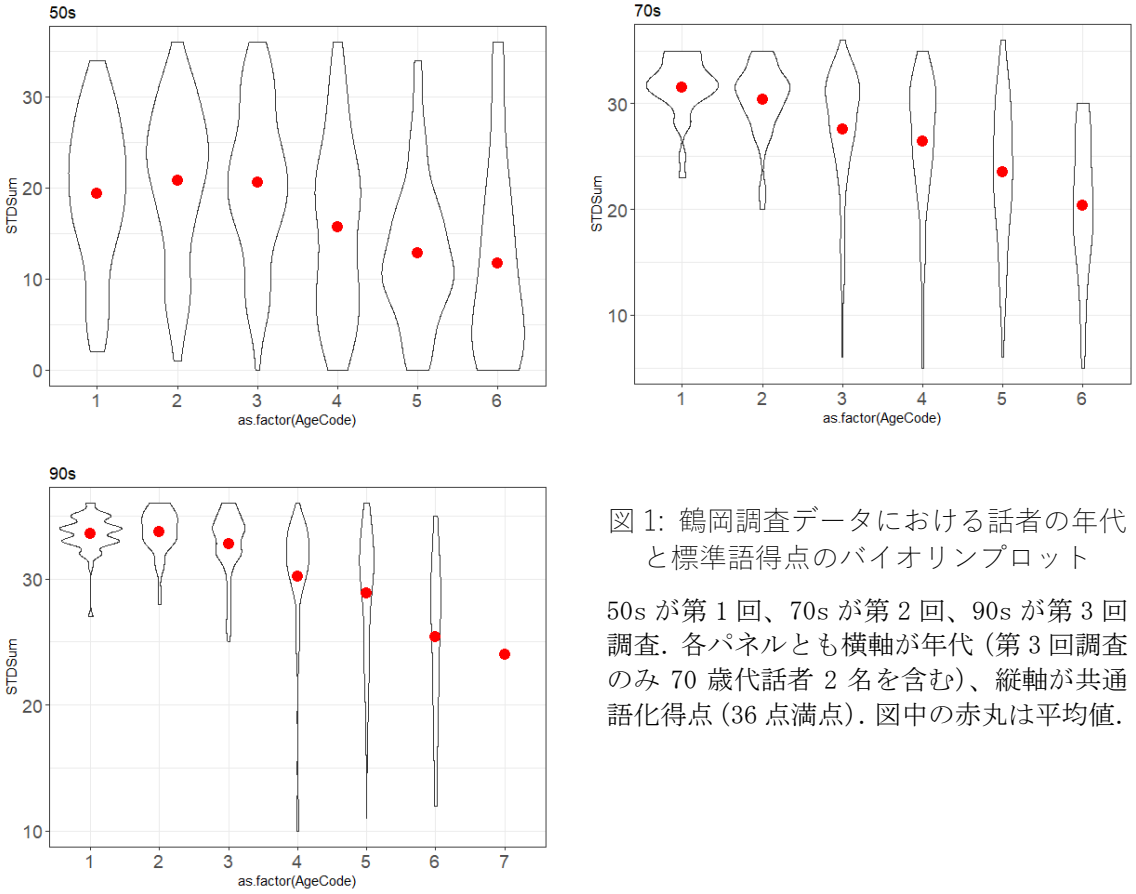


図 1: 鶴岡調査データにおける話者の年代と標準語得点のバイオリンプロット

50s が第 1 回、70s が第 2 回、90s が第 3 回調査．各パネルとも横軸が年代（第 3 回調査のみ 70 歳代話者 2 名を含む）、縦軸が共通語化得点（36 点満点）．図中の赤丸は平均値．

第 1 回調査の分布に比較すると、第 2・第 3 回調査データの分布は特に若年層において分布域を収斂させる傾向をみせてはいるものの、まだ多くの年代において縦軸の全域にわたってサンプル（つまり話者）が分布している。

表 1 は図 1 と対応させる形でデータの過分散指数を計算した結果である。過分散指数は、観察された分散÷分散の理論値で計算される値であり、分散の理論値は N×p×(1-p)で計算する。試行回数 N は 36 であり、成功確率 p の値には観察されたデータから計算した確率を充てた。表 1 をみると過分散状態を完全に脱しているのは第 3 回調査の 10 代だけであることがわかる。

表 1. 調査別・年代別の過分散指数と話者数

	第 1 回調査		第 2 回調査		第 3 回調査	
年代	過分散指数	話者数	過分散指数	話者数	過分散指数	話者数
10 代	7.54	57	1.95	31	0.99	45
20 代	7.94	95	2.31	68	1.45	52
30 代	8.16	131	5.33	99	2.22	86
40 代	10.07	97	5.53	86	5.84	74
50 代	7.33	7	5.60	70	4.80	74
60 代	14.11	39	4.80	47	5.02	*68

*70 代 2 名を含む

３．２　音韻クラスと調査項目

データが過分散状態に陥る原因は共通語化の成功確率 p が何らかの要因で変動することにある。同一年代に属する話者の個体差が大きな要因であるが、それ以外に、前報で確認できたように、音韻のクラスやさらには同一クラス内の調査語彙（調査項目）ごとに p が変動している可能性もある。この点を確認したのが図 2 である。

図 2 の各パネルは鶴岡調査音韻項目を構成する 7 種の音韻クラス（「アクセント」「中舌化」「i と e」「唇音化」「口蓋化」「前鼻音化」「有声化」）別に、各クラスに含まれる調査項目の平均共通語化率の調査毎の推移を比較している。全体的傾向としては第 1 回から第 2 回へ、また第 2 回から第 3 回へと進むにつれ、平均共通語化率が上昇する。しかし、第 3 回調査の段階においても、音韻クラス間の格差は解消されていない。「i と e」「唇音化」「口蓋化」「前鼻音化」「有声化」では、共通語化の平均値が 0.9 前後に達しているが、「中舌化」では 0.8 程度、「アクセント」では 0.5 以下である。

クラス内での調査語彙間の差も、全体としては調査が進むにつれて減少している。しかし、アクセントのように調査項目間の差が拡大したクラスもあることは注目に値する。

結論として、第 2 回・第 3 回調査データのモデリングにおいても、すべての調査項目に対して同一の共通語化確率を当てはめる（つまり二項分布を想定する）ことには無理がある。そこで前報と同じく、共通語化得点を予測するのではなく、ひとつひとつの調査項目における回答が共通語形（1）か非共通語形（0）かを予測するベルヌーイ分布に基づく回帰モデルを採用し、共通語化の成功確率 p に音韻クラスや調査項目が影響を及ぼすモデルを考案することにした。この場合、予測すべきサンプルの数は理論上は表 1 の話者数を調査項目数(36)倍したものになるが、話者によって無回答などの欠損値が存在することがあるので、若干減少する。

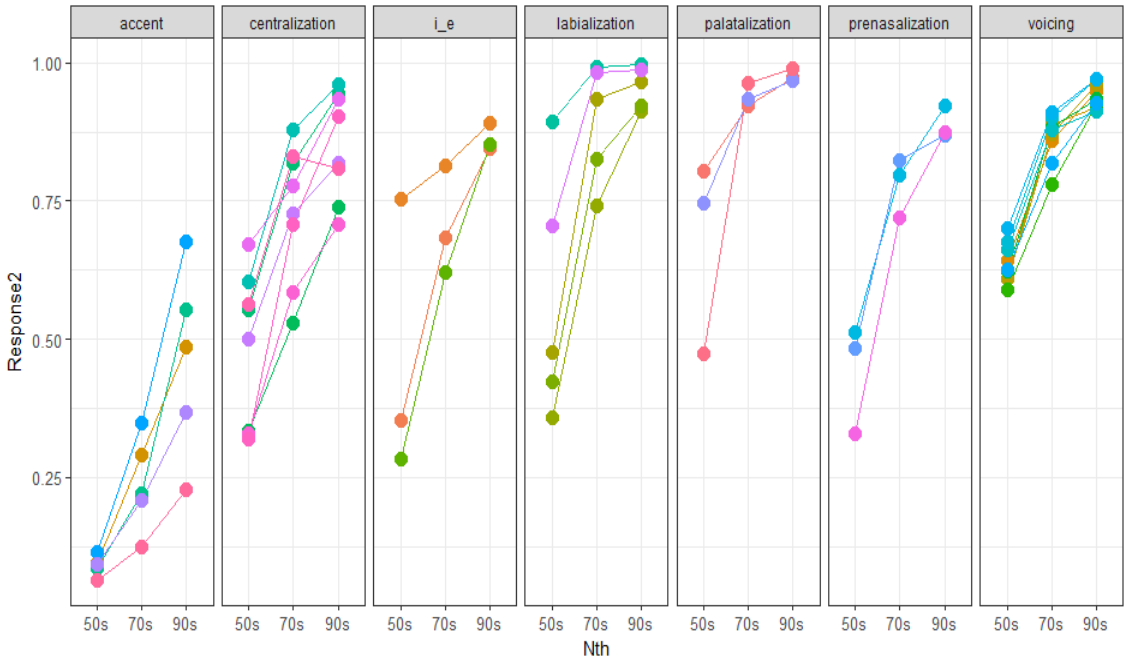


図 2: 音韻クラスおよび調査語彙による平均共通語化率の調査毎の推移
各パネルが音韻クラス（左から「アクセント」「中舌化」「i と e」「唇音化」「口蓋化」「鼻音化」「有声化」）の別、パネル内の線が調査項目の別を示す.

4．統計モデリング

4．1　ベイズ回帰モデルによる予測

前節の分析結果を踏まえ、今回の分析でも前報と同様、複数のベイズ回帰モデルを考案して比較することにした。本節で検討の対象とする 7 種のモデルの特徴を表 2 にまとめた。Stan 言語による回帰モデルのプログラム（後掲する図 3-5 参照）では、i 番目のサンプルに関するベルヌーイ分布の成功確率 q[i]が inv_logit()関数によって決定される（図 3 の 21 行、図 4 の 31 行、図 5 の 35 行参照）。その際、inv_logi()関数の引数は、`a[i]+ b[i]*Age[i]`のような話者の年齢（`Age[]`）の一次式の形をしている。表 2 で切片、傾きと呼んでいるのは、この一次式の切片（`a[]`）と傾き（`b[]`）のことである。

表 2 のモデル 1～4 は前報におけるモデル 1～4 と同一であり、表 2 のモデル 6, 7 はそれぞれ前報におけるモデル 5, 6 と同一である。表 2 のモデル 5 は話者の個体差単独での影響をより詳しく検討するために、今回新たに追加したモデルである。

これらのモデルのパラメータはベイズ統計の手法によって推定し、モデルの実装には Stan 言語を利用した。図 3 に表 2 のモデル 5 にあたる回帰モデルのベイズ推定用 Stan プログラムを示す。これ以外のモデルの推定に用いた Stan プログラムは前報に掲載されているので参照されたい。

図 3 も含め、今回利用したベイズモデルでは推定すべきパラメータのレンジを指定しているだけで、事前分布は指定していない。このような場合、Stan は無情報事前分布として一様分布ないし非常に大きな分散をもつ正規分布を採用する。Stan プログラムは R 言語からライブラリ Rstan を介して実行した。MCMC による事後分布のシミュレーションの実行条件に関しては、チェイン数を 3 に固定し、iteration は 2000～4000、warmup は 1000～2000、

thinning は 1〜3 の範囲で、モデルの収束状態を観察しながら調整した。収束の判定条件としては rhat < 1.1 を採用した。これは Stan による分析で広く採用されている判定条件である（松浦 2016）

表 2：7 種の回帰モデル

モデル	特　徴
1	一次式の切片も傾きも一定のモデル（ベースライン）
2	音韻クラスごとに一次式の切片と傾きの両方が変化するモデル
3	語彙ごとに一次式の切片と傾きの両方が変化するモデル
4	話者ごとに一次式の切片だけが変化するモデル。傾きは固定
5	話者ごとに一次式の切片と傾きの両方が変化するモデル（2018.01.19 に追加）
6	話者ごとに一次式の切片が変化し語彙ごとに傾きが変動するモデル
7	話者ごと・語彙ごとに切片が変化し語彙ごとに傾きが変化するモデル

```
01:  #BernLogitRegHie4_waic.stan by KM
02:  #話者によって切片と傾きが変化する Bernoulli 階層ロジスティック回帰
03:  data {
04:    int I;                          //データ総数
05:    int Nsbj;                       //話者数
06:    int<lower=14,upper=70> Age[I];  //話者の年齢
07:    int<lower=0, upper=1> Y[I];     // i 番目のデータの共通語化状態(1 ないし 0)
08:    int<lower=1, upper=Nsbj> Subject[I];   // 話者の個体識別番号
09:  }
10: parameters {
11:  real<lower=-5, upper=5> as[Nsbj];
12:  real<lower=-0.2, upper=0.2> bs[Nsbj];
13: }
14: transformed parameters {
15:   real a[I];
16:   real b[I];
17:   real<lower=0, upper=1> q[I];
18:   for (i in 1:I) {
19:     a[i] = as[Subject[i]];
20:     b[i] = bs[Subject[i]];
21:     q[i] = inv_logit(a[i] + b[i]*Age[i]);
22:   }
23:  }
24: model {
25: for (i in 1:I){
26:   Y[i] ~ bernoulli(q[i]);
27:   }
28: }
29: generated quantities {
30:   real y_pred[I];
31:   real log_lik[I];
32:   for (i in 1:I){
33:     y_pred[i] = bernoulli_rng(q[i]);
34:     log_lik[i] = bernoulli_log(Y[i], q[i]);  //WAIC のために対数尤度を保存
35:   }
36: }
```

図 3: ベイズ推定による回帰分析の Stan プログラム（表 2 のモデル 5）

モデルによる予測の評価指標も前報と同様、平均予測誤差・F 値・WAIC の 3 種を用いる。平均予測誤差はモデルによって予測された 36×話者数個のサンプルの値（0 か 1）と実際の

観測値との差の絶対値の平均である。平均予測誤差は理解しやすいが、予測すべきデータが均等に分布していない（例えば大部分が 0 で 1 が僅かであるなど）場合、適切な指標となりがたい。

F 値はこの欠点を補うために考案された指標であり、モデルの予測値の適合率(例えば 1 と予測したもののうち実際に 1 であったものの割合）と再現率(例えば実際に 1 であったもののうち 1 と予測されたものの割合）の調和平均として定義される。

最後に WAIC は AIC を非正規分布に基づくモデルにも適用できるよう拡張した新しい情報量基準である。WAIC による判定は交差検定(cross validation)と漸近等価であると考えられている（関連 URL 参照）。

表 3: 第 1 回調査データの予測

モデル	平均予測誤差	F 値	WAIC
1（ベースライン）	0.422	0.568	24121
2（音韻クラス～切片・傾き）	0.419	0.676	21503
3（語彙～切片・傾き）	0.298	0.708	20318
4（話者～切片）	0.284	0.714	20154
5（話者～切片・傾き）	0.283	0.716	20156
6（話者～切片、語彙～傾き）	0.179	0.821	15267
7（話者・語彙～切片、語彙～傾き）	0.175	0.823	14684

表 4: 第 2 回調査データの予測

モデル	平均予測誤差	F 値	WAIC
1（ベースライン）	0.261	0.850	15923
2（音韻クラス～切片・傾き）	0.185	0.882	12463
3（語彙～切片・傾き）	0.178	0.885	11886
4（話者～切片）	0.230	0.858	14982
5（話者～切片・傾き）	0.229	0.858	14984
6（話者～切片、語彙～傾き）	0.143	0.907	10591
7（話者・語彙～切片、語彙～傾き）	0.131	0.914	9993

表 5: 第 3 回調査データの予測

モデル	平均予測誤差	F 値	WAIC
1（ベースライン）	0.151	0.919	11428
2（音韻クラス～切片・傾き）	0.128	0.928	8720
3（語彙～切片・傾き）	0.120	0.932	8318
4（話者～切片）	0.141	0.922	10908
5（話者～切片・傾き）	0.142	0.921	10899
6（話者～切片、語彙～傾き）	0.093	0.947	7440
7（話者・語彙～切片、語彙～傾き）	0.088	0.949	6960

表 2 の各回帰モデルのパラメータを図 3 のようなベイズモデルで推定し、その予測性能を評価した結果を調査別に表 3-5 として示す。どの調査においても、またどの評価指標に関しても、モデル 7 が最良と判定されている。[1] この結果は、前報における第 1 回調査データ

[1] 平均予測誤差は小さいほど、F 値は 1.0 に近いほど、そして WAIC は小さいほどモデルの性能が良いと判断する。

のモデリングの結果とも一致している。

次に調査間の差に注目すると、第 1 回よりも第 2 回、そして第 2 回よりも第 3 回調査データの方が、より高い精度で予測できていることがわかる。このような結果は、表 1 に示した過分散指標から予想されるところではあるのだが、第 3 回調査データ（表 5）の場合は、ベースラインモデルの F 値が既に 0.9 を上回っており、最良モデルと判定されたモデル 7 以外のモデルも実質的に高い性能を発揮していることが注目される。

４．２　話者の個人差の性別・出生地・教育歴による代替

表 3-5 の結果は、共通語化の要因として、話者の個体差が語彙の個体差（調査項目の差）と同程度に予測に貢献していることを示していた。[2] 本研究で話者の個体差をモデルに含める場合、M 名の話者は共通語化に関して M 通りの異なる状態をとりうると考えている（図 3 のプログラムでは、19-21 行でその関係が実装されている）。しかしこのように文字通りの意味での個体差を想定することは、鶴岡調査を含む従来の社会言語学の分析では稀であり、年齢・性別・出身地・教育等の諸属性の集合によって話者の個体差を代替していることが多い。以下ではこのような代替の有効性を鶴岡調査のデータを用いて定量的に検討する。話者の年齢以外の属性として、性別・言語形成地域・教育歴の影響を評価する。

この目的のために、新たに 2 個の回帰モデルを考案した。表 2 のモデル 6（話者が切片にだけ影響し、語彙が傾きにだけ影響するモデル）をベースラインとして、話者の個体差（純粋な個体差）を「話者の性別＋言語形成地域」の組み合わせで代替するモデル（これをモデル 8 とする）と、「話者の性別＋言語形成地域＋教育歴」で代替するモデル（モデル 9）の 2 種類である。

これらの回帰モデルの Stan プログラムを図 4, 5 として示す。図 4 の 31 行では、一次式の切片を int1[i]と int2[i]の二つの分離しており、前者が話者の性別（図 4 の 28 行参照）、後者が言語形成地域（同 29 行参照）による影響を被るモデルとなっている。同じく図 5 では一次式の切片を 3 個に分割し（図 5 の 35 行）、それぞれが性別、言語形成地域、教育歴による影響を被るモデル（同 31-33 行）となっている。

```
01: # BernLogitRegHie10_waic.stan
02: # 話者(Subject2)を性別と言語形成地情報でどれだけ代替できるかの検討
03: # 話者に替えて、性別と言語形成期によって切片が変化し、
04: # Item2 をハイパーパラメータとして傾きが変化する Bernoulli ロジスティック回帰
05: # by KM  2018.01.25
06: data {
07:   int I;
08:   int Nitm;
09:   int Nsbj;
10:   int <lower=1, upper=2> Sex[I];
11:   int <lower=1, upper=3> BP[I];
12:   int<lower=1, upper=36> Item[I];
13:   int<lower=14,upper=70> Age[I];
14:   int<lower=0, upper=1> Y[I];
15:   int<lower=1, upper=Nsbj> Subject[I];
16: }
17: parameters {
18:   real<lower=1,upper=2> asx[2];
19:   real<lower=-1,upper=2> abp[3];
```

[2] ベースライン（モデル 1）とモデル 3 ないしモデル 5 の相違を比較せよ。

```
20:   real<lower=-0.2,upper=0.1> bs[Nitm];
21: }
22: transformed parameters {
23: real<lower=-2, upper=2> int1[I];
24: real<lower=-2, upper=2> int2[I];
25: real<lower=-0.5, upper=0.5> b[I];
26: real<lower=0, upper=1> q[I];
27: for (i in 1:I) {
28:   int1[i] = asx[Sex[i]];
29:   int2[i] = abp[BP[i]];
30:   b[i] = bs[Item[i]];
31:   q[i] = inv_logit(int1[i] + int2[i] + b[i]*Age[i]);
32: }
33: }
34: model {
35: for (i in 1:I){
36:   Y[i] ~ bernoulli(q[i]);
37:   }
38: }
39: generated quantities {
40:   real y_pred[I];
41:   real log_lik[I];
42:   for (i in 1:I){
43:     y_pred[i] = bernoulli_rng(q[i]);
44:     log_lik[i] = bernoulli_log(Y[i], q[i]);
45:   }
46: }
```

図 4: モデル 8 の Stan プログラム

```
01: # BernLogitRegHie11_waic.stan
02: # 話者に替えて、性別と言語形成期と教育歴をハイパーパラメータとして
03: # 切片が変化し、Item2 をハイパーパラメータとして傾きが変化する
04: # Bernoulli 階層ロジスティック回帰
05: # by KM  2018.01.25
06: data {
07:   int I;
08:   int Nitm;
09:   int Nsbj;
10:   int <lower=1, upper=2> Sex[I];
11:   int <lower=1, upper=3> BP[I];
12:   int <lower=1, upper=9> Edu[I];
13:   int<lower=1, upper=36> Item[I];
14:   int<lower=14,upper=70> Age[I];
15:   int<lower=0, upper=1> Y[I];
16:   int<lower=1, upper=Nsbj> Subject[I];
17: }
18: parameters {
19:   real<lower=1,upper=2> asx[2];
20:   real<lower=-1,upper=2> abp[3];
21:   real<lower=-2,upper=2> aed[8];
22:   real<lower=-0.2,upper=0.1> bs[Nitm];
23: }
24: transformed parameters {
25:   real<lower=-2, upper=2> int1[I];
26:   real<lower=-2, upper=2> int2[I];
27:   real<lower=-2, upper=2> int3[i];
28:   real<lower=-0.5, upper=0.5> b[I];
29:   real<lower=0, upper=1> q[I];
30:   for (i in 1:I) {
31:     int1[i] = asx[Sex[i]];
```

```
32:     int2[i] = abp[BP[i]];
33:     int3[i] = aed[Edu[i]];
34:     b[i] = bs[Item[i]];
35:     q[i] = inv_logit(int1[i] + int2[i] + int3[i] + b[i]*Age[i]);
36:   }
37: }
38: model {
39: for (i in 1:I){
40:   Y[i] ~ bernoulli(q[i]);
41:   }
42: }
43: generated quantities {
44:   real y_pred[I];
45:   real log_lik[I];
46:   for (i in 1:I){
47:     y_pred[i] = bernoulli_rng(q[i]);
48:     log_lik[i] = bernoulli_log(Y[i], q[i]);
49:   }
50: }
```

図 5: モデル 9 の Stan プログラム

表 6-8 に各属性に関する話者数の分布を示す。

表 6: 話者の性別の分布

	第 1 回調査	第 2 回調査	第 3 回調査
男	212	228	317
女	284	280	396

表 7: 話者の言語形成地域の分布

	第 1 回調査	第 2 回調査	第 3 回調査
鶴岡市	338	318	456
山形県内	117	139	178
山形県外	38	49	75
不明	3	2	4

表 8: 話者の教育歴の分布

	第 1 回調査	第 2 回調査	第 3 回調査
小学校	145	61	37
中学校	184	194	211
高校	83	197	275
専門学校	8	22	37
旧制高校	5	1	8
大学	12	23	66
その他	31	6	70
無し	24	1	--
無回答	4	3	9

表 9 にモデル 8, 9 による予測精度の評価結果を示す。表 9 からは、話者の純粋な個体差を話者の社会言語学的な属性で代替したモデルの予測精度はベースラインに劣ることが分かる。つまり、社会言語学的な属性は話者の個体性を十全には把握できていない。

ただし、代替による劣化の程度は調査によって異なっている。図 6 は表 9 の F 値だけを

調査間で比較したグラフである。第 1 回調査データにおいては、モデル 8, 9 はベースラインであるモデル 6 から顕著に劣化しているが、第 2 回調査では劣化幅が減少し、第 3 回調査では一層減少していることがわかる。

表 9: モデル 6, 8, 9 による予測の精度

調査	モデル	予測誤差	F 値	WAIC
第 1 回	6	0.179	0.821	15267
	8	0.289	0.724	19571
	9	0.268	0.737	18805
第 2 回	6	0.143	0.907	10591
	8	0.184	0.884	12043
	9	0.178	0.887	11665
第 3 回	6	0.093	0.947	7440
	8	0.122	0.931	8477
	9	0.120	0.932	8262

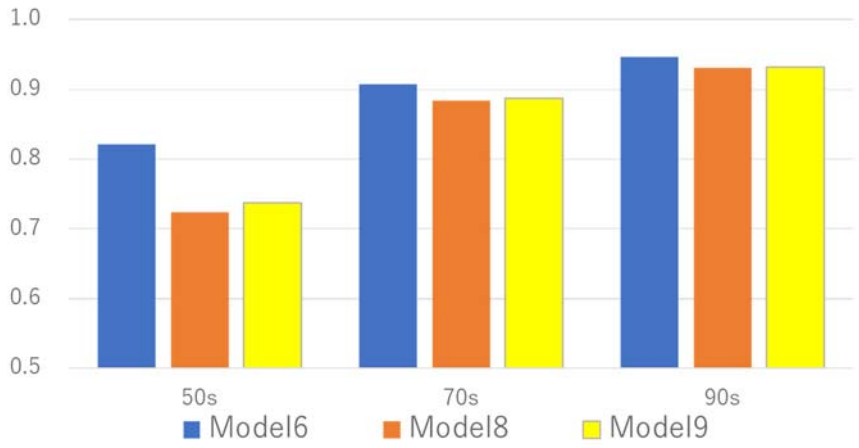


図 6: モデル 6, 8, 9 の F 値の 3 回の調査を通しての比較

5. 議論と結論

今回の分析では、まず第 2 回・第 3 回調査データも、二項分布によって生成されたとみなすと、過分散状態に陥ること、従って、前報と同じベルヌーイ分布に基づく回帰モデルでの分析が望ましいことを確認した。

その方針に従って種々の回帰モデルを比較検討したところ、前報で最良モデルとして採用したモデルが第 2 回・
第 3 回調査データにおいても最良の予測をもたらすことが判明した。「話者ごと・語彙ごとに切片が変化し語彙ごとに傾きが変化する」モデルである。

一方、第 1 回調査データと第 2 回・第 3 回調査データの間には明らかな差も認められた。最も単純なベースライン回帰モデル（切片も傾きも固定で、話者の年齢だけを変数とするモデル）による予測の F 値は、第 1 回調査では 0.57 にとどまっているのに対して、第 2 回調査では 0.85 を、また第 3 回調査では 0.9 を上回っている（表 4, 5 参照）。これからわかるように、第 1 回調査に比べると第 2 回・第 3 回調査では年齢による予測の有効性が大きく向上している。前報では、話者の年齢が言語変化（共通語化）の要因として最も重要であると

いう社会言語学上の仮定は無批判には受け入れられないと考えたが、共通語化が進展した段階ではこの仮定はある程度まで有効になることが確認できた。

また、年齢以外の話者の個体差を、性別・言語形成地域・教育歴で代替したモデル 9 による予測結果をみても、第 1 回調査データと第 2 回・第 3 回調査データとの間には質的な相違が認められた。第 1 回調査では代替によって顕著な劣化が生じるが、第 2 回・第 3 回では劣化の幅が小さい。この点においても、従来の社会言語学的分析は共通語化がある程度まで進展したデータに対してよりよく適合しているといえる。

しかし、以上を換言すれば、共通語化の初期状態は、従来の社会言語学的分析手段によっては十分に把握しきれないということである。初期段階にある言語変化の分析においては、話者と調査対象語彙の個体差に関して格別の配慮が必要となることを今回の分析は示している、というのが本稿の結論である。

本研究の次のステップとしては、今回は独立して分析した第 1 回から第 3 回までのデータを統一的に分析するための統計モデルの開発が挙げられる。

謝　辞

鶴岡調査の話者と調査者、ならびに鶴岡調査データベースの公開に尽力された国立国語研究所の関係者に感謝します。本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度)の成果です。

文　献

国立国語研究所.『地域社会の言語生活―鶴岡市における実態調査―』（国立国語研究所報告 5）秀英出版, 1953. doi/10.15084/00001214

国立国語研究所.『地域社会の言語生活―鶴岡市における 20 年前との比較―』（国立国語研究所報告 52）秀英出版, 1974. doi/10.15084/00001251

国立国語研究所.『地域社会の言語生活―鶴岡における 20 年間隔 3 回の継続調査―』国立国語研究所, 2007.

前川喜久雄「鶴岡市共通語化調査データの確率論的再検討」言語資源活用ワークショップ 2017 発表論文集, pp.163-180, 2017.09.06. doi/10.15084/00001517

松浦健太郎『Stan と R でベイズ統計モデリング』共立出版, 2016.

関連 URL

鶴岡調査データベース：http://www2.ninjal.ac.jp/longitudinal/tsuruoka.html

WAIC：http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/waic2011.html

『日本語日常会話コーパス』活用環境の構築

山口昌也 (国立国語研究所音声言語領域) †

Development of an Environment to Make Use of "Corpus of Everyday Japanese Conversation"

Masaya YAMAGUCHI (Spoken Language Division, NINJAL) †

要旨

本発表では，『日本語日常会話コーパス』を活用するための環境構築について述べる。『日本語日常会話コーパス』は動画・音声，転記テキストを含み，転記テキストには形態素解析結果などの言語学的な情報がアノテーションされている。本発表で提案する活用環境は，全文検索システム『ひまわり』と観察支援システム FishWatchr を統合することにより実現した。本環境を用いることにより，次のことが可能になる。(1)『ひまわり』で転記テキストを全文・単語検索し，当該位置の映像を FishWatchr で閲覧すること，(2)FishWatchr 上で動画再生位置に簡易なアノテーション（二つのユーザ定義ラベル，自由テキストを記述可能）を付与すること，(3)FishWatchr 上で転記テキストを表形式で表示し，選択した転記テキスト位置の動画を再生すること。また，動画の再生と同期させて転記テキストをスクロール表示すること。

1 はじめに

本稿では，現在，国立国語研究所で構築中の『日本語日常会話コーパス』（以後，CEJC）(小磯花絵ほか, 2017) を活用する環境の構築について述べる。CEJC は日常場面で自発的に生じた多様な日常会話を収録したコーパスで，ビデオデータ，転記テキストを含み，転記テキストには単語情報，会話情報，話者情報などの言語学的な情報がアノテーションされている。

このように，CEJC は一次資料のビデオデータに対して，さまざまなデータがアノテーションされている。これらのデータを扱うには，既存のツールを利用することができる。例えば，映像の閲覧や分析には，メディアプレーヤー (例：VLC[1])，ビデオアノテーションシステム (例：ELAN[2] (Brugman and Russel, 2004))，音声分析であれば，Praat[3](Boersma and Weenink, 2001) などの音声分析ソフトウェア，転記テキストに対する全文検索や単語検索に対しては，KHCoder[4](樋口耕一, 2003) や『ひまわり』[5](山口昌也・田中牧郎, 2005) などである。

その一方で，CEJC に含まれるデータを効率的に利用するには，複数の種類のデータを統合して利用する環境が必要である。例えば，転記テキストを検索したとき，検索結果の当該シーンや，話者・会話データなどの発話状況を迅速に参照できれば，効率的な分析が可能になる。

そこで，本稿では，複数のツールを組み合わせて，CEJC を有効に活用できる利用環境を構築する。活用環境の設計にあたっては，上で挙げた例のように，転記テキストを検索し，一次資料であるビデオ

† http://www2.ninjal.ac.jp/masaya

[1] https://www.videolan.org/vlc/

[2] https://tla.mpi.nl/tools/tla-tools/elan/

[3] http://www.fon.hum.uva.nl/praat/

[4] http://khc.sourceforge.net/

[5] http://www2.ninjal.ac.jp/lrc/index.php?himawari

データを参照したり，ビデオデータに簡易なアノテーションを行う利用形態を想定する。また，本環境は，CEJC 公開時に同梱することを想定しているため，容易に利用できることを目指す。

以上の背景から，本環境は，全文検索システム『ひまわり』と観察支援システム FishWatchr とを組み合わせて，構築する。『ひまわり』は，XML でアノテーションされたテキストに対するコンコーダンサで，全文検索・単語検索，検索文字列の KWIC 表示，アノテーション内容の表示・集計をすることができる。2004 年から一般公開され，これまでに『太陽コーパス』[6]『日本語話し言葉コーパス』[7]などの検索システムとして同梱されてきた実績がある。

一方，FishWatchr はディスカッション練習やプレゼンテーション練習などの協同型教育活動の観察を支援するためのシステムである。学習者が利用することを前提としたシステムであり，ELAN のように複雑なアノテーションを行うことはできないが，特定のシーンに対して，アノテーション専用ボタンで容易にアノテーションすることが可能である。

この後の本稿の構成は，次のようになっている。まず，次節では，本環境の設計を行うために，基本的な利用形態を定めた上で，必要とされる機能を示す。3 節では，本環境を構築するための方法として，CEJC のデータを『ひまわり』と FishWatchr にインポートする方法を示す。さらに，4 節で実現した環境での実行例を示し，5 節でまとめを述べる。

2　活用環境の設計

2.1　CEJC のデータ構成

前述のとおり，CEJC には複数の種類のデータが含まれている。ここでは，本環境の設計を行う前に，CEJC のデータ構成（図 1）について説明しておく。なお，ここで述べるデータ構成は，本環境に関連する部分のみであり，全データ構成については，小磯花絵ほか (2017) などを参照されたい。

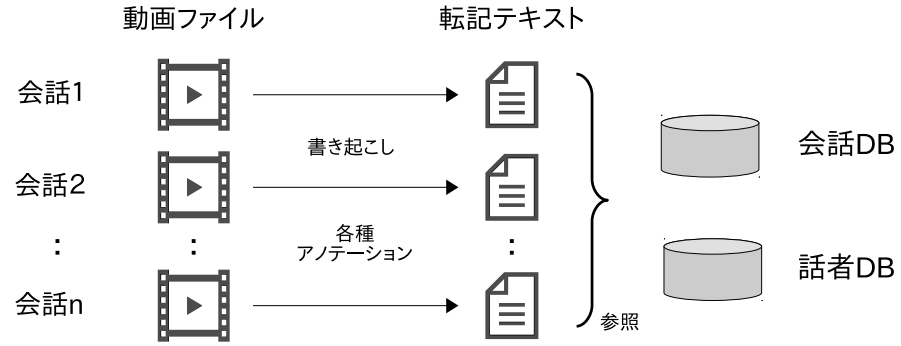


図 1　CEJC のデータ構成

このように，各会話はビデオファイルとして格納される。会話内の発話は書き起こされ，転記ファイルに記述される。発話には，話者名，ビデオファイルにおける時間情報，単語の情報（短単位），言いさしや言い誤りなどの情報がアノテーションされる。会話や発話者の詳細情報は，データベースとしてまとめられており，適宜参照できるようになっている。

[6] http://pj.ninjal.ac.jp/corpus_center/cmj/taiyou/
[7] http://pj.ninjal.ac.jp/corpus_center/csj/

2.2 基本的な利用形態と機能

本稿では，図 2 のような利用形態を想定する。この図のとおり，転記テキストに対する全文検索，もしくは，単語検索を行い，その結果を KWIC 表示するのが最も基本的な利用方法である (図 2 ①②)。この際，個々の検索結果には，KWIC キーに関連する発話者，会話などの情報も併記する。

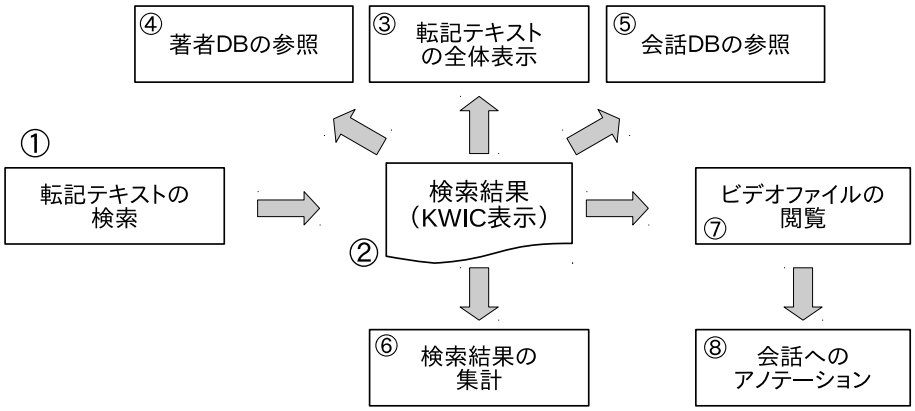


図 2 想定する基本的な利用形態

この検索結果を起点として，さまざまな処理を行う。処理は，(1) 転記テキスト関連情報の取得 (図 2 ③④⑤)，(2) 検索結果の集計 (図 2 ⑥)，(3) ビデオファイル閲覧 (図 2 ⑦⑧)，の三つに分けられる。

■転記テキスト関連情報の取得 検索結果の KWIC 表示は，表示範囲が限定されるため，転記テキスト全体を閲覧できるようにする。また，詳細な話者情報や会話情報を個々の検索結果に併記するのは表示量が多くなりすぎるので，必要に応じて，閲覧したい検索結果を選択し，話者 DB，会話 DB を参照できるようにする。

■検索結果の集計 大量の検索結果が得られた場合，その結果を集約したり，統計的な分析を支援することができるようにする。例えば，検索結果から発話者別の頻度を求めたり，検索文字列の調整頻度を計算するために，会話ごとの単語数を収集するといった処理である。

■ビデオファイルの閲覧 検索結果の当該シーンのビデオを転記テキストと並行して，閲覧できるようにする。また，閲覧しているビデオファイル中のシーンを指定して，ラベルやコメントを付与するなどの簡易的なアノテーションができるようにする。

3 活用環境の実現

3.1 全体構成

本稿で提案する CEJC 活用環境の全体構成を図 3 に示す。

活用環境は，大きく分けて，『ひまわり』と FishWatchr の二つのシステムから構成される。図 2 の利用形態と対応させると，『ひまわり』は①〜⑥を担当し，FishWatchr は⑦⑧を担当する。ただし，転記テキストの全体表示（③）には，Web ブラウザを用いる。

CEJC に含まれるデータのうち，『ひまわり』側で扱うのは，転記テキスト，会話データベース，話者データベースである。転記テキストは，『ひまわり』用のコーパスファイルに変換する。その際，転記テキストにアノテーションされている単語情報は，XML タグとして記述される。詳細は，この後で述

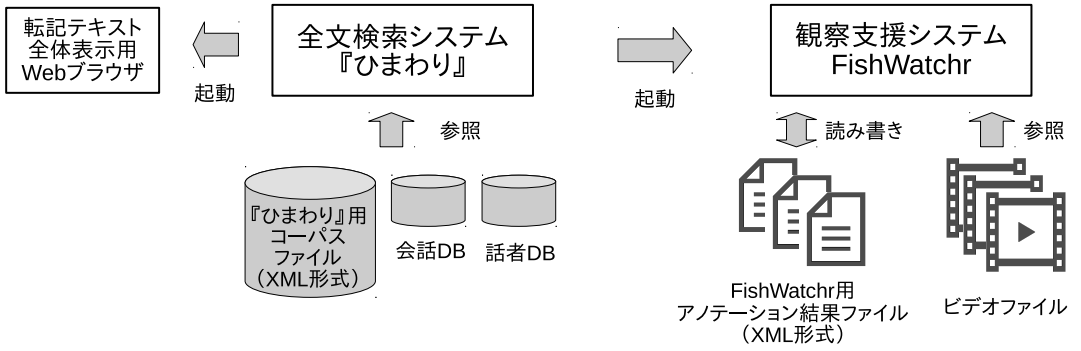


図 3　活用環境の全体構成

べる。会話データベース，話者データベースは，『ひまわり』内部のデータベース（書誌情報などを格納するために設計されたもの）のデータベースに格納される。

一方，FishWatchr 側で扱うのは，ビデオファイルと転記テキストである。転記テキストは，会話ごとに FishWatchr 用のアノテーション結果ファイルに変換され，ビデオとの連動表示（4.2 節参照）を行うために利用される。このデータファイルには，FishWatchr で新規に行ったアノテーションも追記される形で，保存される。

この後の節では，CEJC のデータが『ひまわり』，FishWatchr にどのようにインポートされるかを詳しく説明する。なお，図 2 で示した各種の処理については，両システムの機能を組み合わせて利用しているため，動作例を 4 節で示すにとどめる。

3.2　『ひまわり』への CEJC データのインポート

図 4 は，CEJC の CSV 形式の転記テキストの例である。1 行 1 発話で記述され，発話の開始・終了時間，話者名が付与されている。発話部分には，独自の形式のタグが付与されている。例えば，2 行目の（F あの:）は F タグでフィラーであることを示している。4 行目の（W ナ|なん）は W タグで言い間違いを修正している (|の前が修正前，後が修正後である)。なお，転記テキストには，単語情報もアノテーションされているが，表記が複雑になるため，ここでは除外している。

startTime	endTime	speaker	text
381.949	383.086	IC02_美沙	なんかね（F あの）
383.175	385.061	IC02_美沙	えーとねー ちょっと待ってね:。
384.607	386.226	IC01_玲子	（W ナ|なん）かもうすごい未知の世界。
386.209	386.825	IC02_美沙	でしょ:。

図 4　転記テキストの例 (C001_002 から一部引用)

『ひまわり』にインポートする際は，これらを XML 形式で記述する。図 5 は，図 4 の転記テキストを XML 形式に変換した結果である。なお，紙面上の見やすさの関係上，転記テキストの 4 行目の発話のみ示した。また，適宜改行を挿入するとともに，タグの属性は説明に必要なもののみ記述している。

発話は u タグでマークアップされ，1 会話分の発話が cecj タグでマークアップされる。どちらのタグも属性を持ち，cejc タグの name 属性値は会話の ID の役割を果たす。u タグの startTime，endTime 属性は，ビデオファイルにおける発話の開始時間・終了時間を表し，発話をビデオファイルと関連付ける。speaker，speakerID 属性は，発話者名，発話者 ID である。speakerID 属性値は，発話 DB を参照

する際のキーとなる。単語情報 (短単位) は，s タグでマークアップされる。s タグの p，l，t 属性はぞれぞれ品詞，「語彙素」，「タグなし出現形」を保持する。

```
<cejc name="C001_002">
                    :
<u startTime="384.607" endTime="386.226" speaker="IC02_玲子" speakerID="C001">
 <s p="代名詞" l="何" t="なん">(W ナ|なん)</s>
 <s p="助詞-副助詞" l="か" t="か">か</s>
 <s p="副詞" l="もう" t="もう">もう</s>
 <s p="形容詞-一般" l="凄い" t="すごい">すごい</s>
                    :
 <s p="名詞-普通名詞" l="世界" t="世界">世界。</s>
</u>
                    :
</cejc name="C001_002">
```

図 5　『ひまわり』へのインポート例

『ひまわり』は，XML タグを無視して全文検索する。したがって，全文検索時は，CEJC の独自タグを考慮しつつ，検索文字列を指定する必要がある。独自タグを除外して検索したい場合は，単語検索で「タグなし出現形」を検索すればよい。前後 2 単語だが，前後の文脈を指定できる。

3.3　FishWatchr への CEJC データのインポート

FishWatchr では，『ひまわり』と異なり，一つの転記ファイル（会話）が一つの FishWatchr 用のアノテーション結果ファイルとしてインポートされる。形式は，XML である。FishWatchr のアノテーションは，時間的範囲を持たない，特定の 1 シーンに対して行われる。そのため，一つの発話はその開始時間を基点とするアノテーションとして記述される。

図 6 は，図 4 の転記ファイルを FishWatchr 用のアノテーション結果ファイルに変換した結果である。comment_list タグは，1 会話分のアノテーション結果を表し，media_file 属性にビデオファイル名を格納している。発話は comment タグでマークアップする。commnet タグの commenter，comment_time 属性は，それぞれ発話者，ビデオファイルにおける発話開始時間を保持する。さらに，comment_type 属性にはユーザ定義のラベル，aux 属性には自由記述のコメントが格納される。

```
<comment_list media_file="C001_002_MIX.mp4">
 <comment commenter="IC02_美沙" comment_type="" aux="" comment_time="381949">
    なんかね (F あの)</comment>
 <comment commenter="IC02_美沙" comment_type="" aux="" comment_time="383175">
    えーとねー ちょっと待ってね:。</comment>
 <comment commenter="IC01_玲子" comment_type="" aux="" comment_time="384607">
    (W ナ|なん) かもうすごい未知の世界。</comment>
 <comment commenter="IC02_美沙" comment_type="" aux="" comment_time="386209">
    でしょ:。</comment>
</comment_list>
```

図 6　FishWatchr へのインポート例

インポート後の初期状態では，転記テキストからインポートした発話のみがアノテーションされている状態だが，後述するように，ユーザはビデオを参照しながら，任意のシーンに対してアノテーションを追加することができる。

4　実現結果

4.1　『ひまわり』

まず，『ひまわり』で CEJC を検索した結果を図 7 に示す。検索結果には，画面左から，検索文字列「会話」に対する KWIC，会話 ID，話者関連情報，発話情報，単語情報が含まれる。

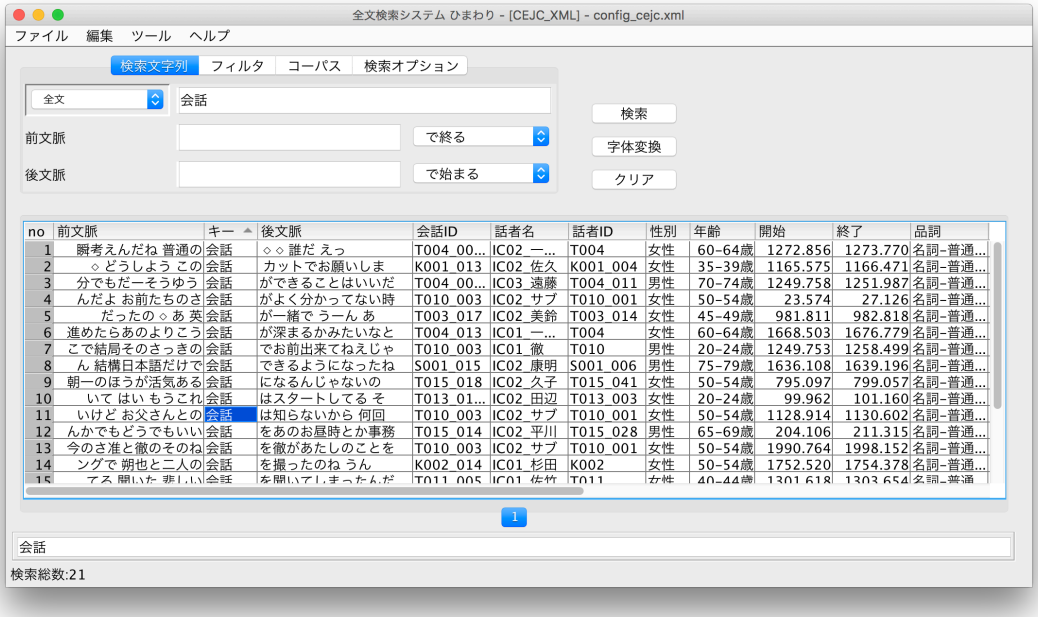


図 7　『ひまわり』による CEJC の検索例

このうち，KWIC 部分のセルをダブルクリックすると，図 8 のように，当該の転記テキスト全体が Web ブラウザで表示される。また，「会話 ID」「話者 ID」列のセルをダブルクリックすると，それぞれ会話 DB，話者 DB から検索された情報が表示される（図 9 は話者情報）。なお，それぞれのデータベースの内容は一覧することも可能である。これら以外の列をダブルクリックした場合は，FishWatchr が起動し，当該シーンがビデオ再生される。

検索結果の分析には，『ひまわり』の分析支援機能を利用する。ここでは，単語「です」の会話ごとの調整頻度を求めるのに必要なデータを収集してみる。図 10 左が「です」を単語検索し，会話ごとの出現頻度を集計した結果である。図 10 中央は s タグを会話別に集計し，会話別の総単語数を求めた結果である。さらに，図 10 右は，『ひまわり』の集計結果連結機能を用いて，二つの結果を連結した結果である。この結果を Excel などの表計算ソフトウェアにコピー＆ペーストすれば，例えば，会話ごとの調整頻度を計算することができる。

4.2　FishWatchr

ここでは，『ひまわり』から起動した FishWatchr でビデオファイルを閲覧し，特定のシーンにコメントをつけてみる。

図 8　転記テキストの全体表示

図 9　話者情報の表示

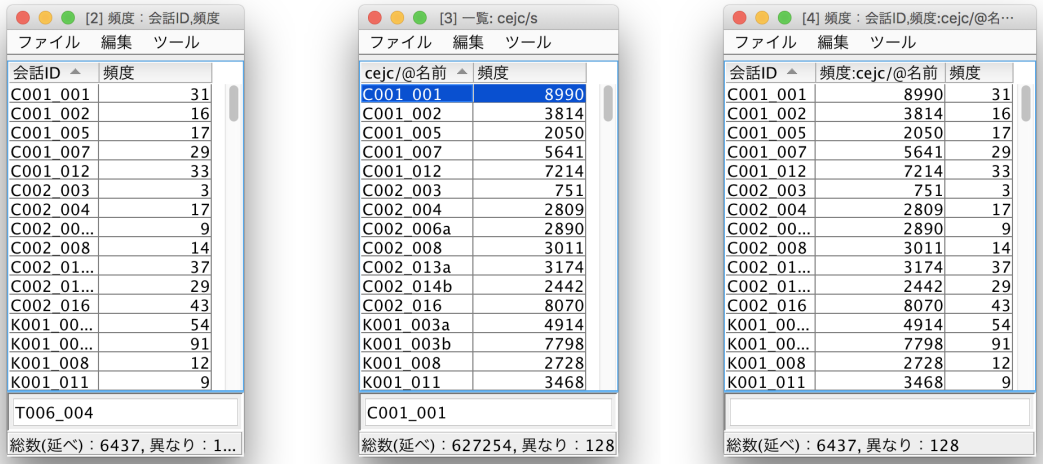


図 10　調整頻度を計測するためのデータ収集（「です」の出現頻度，会話ごとの総単語数，両者の結合結果）

図 11 は，FishWatchr にアノテーション結果ファイルを読み込んだ例である。ウィンドウ右上が会話のビデオである。下部の表はアノテーション表であり，1 アノテーション（つまり 1 発話）1 行で表示される。ウィンドウ左上はアノテーション表を時系列にプロットした図である。

アノテーション表の各アノテーションには発話の開始時間，発話者名，会話 ID，発話の転記テキストが含まれる。このうち，発話の開始時間，転記テキストは変更できないように設定されている。そのため，発話に対してコメントしたい場合は，最右列の「補助情報」欄を用いる。

アノテーション表の表示はビデオの再生と連動してスクロール表示させることもできる。また，閲覧したい行をダブルクリックすると，当該のシーンが再生される。

画面最下部の二つのボタン (「ラベル 1」「ラベル 2」) はアノテーション専用のボタンである。押下すると，ビデオの再生位置にアノテーションが付与される。ボタンのラベルはユーザが 8 個まで定義可能である。

5　おわりに

本稿では，CEJC を有効に活用するための環境の構築方法として，全文検索システム『ひまわり』と観察支援システム FishWatchr を組み合わせて用いる方法を提案し，その実現結果を示した。今回構築したシステムは，本年度に予定されている CEJC のモニタ公開版にも同梱される予定である。

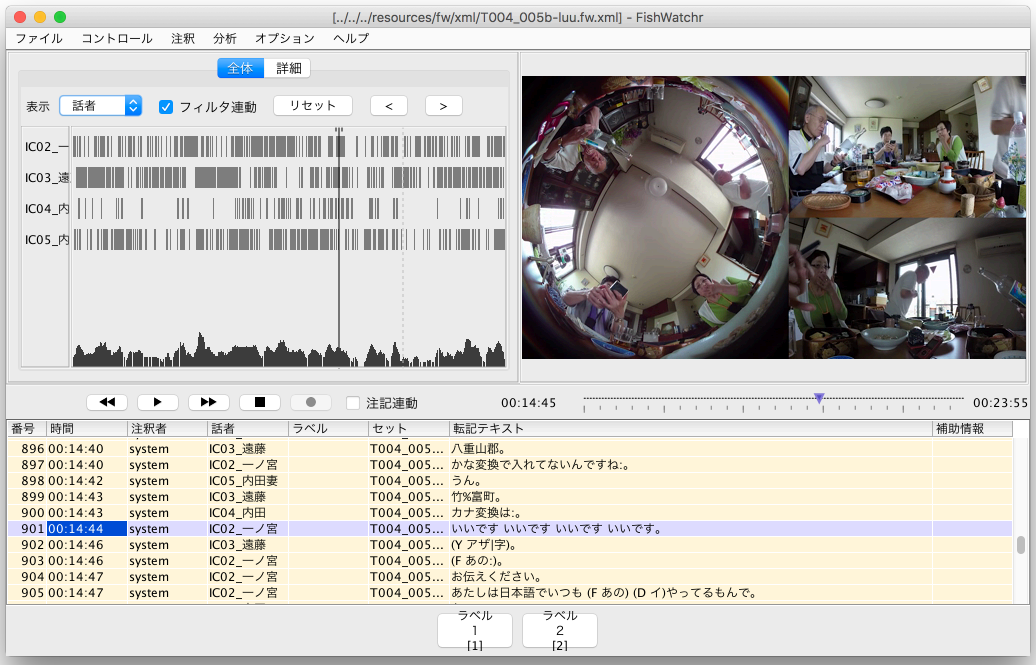


図 11　FishWatchr の動作例

謝　辞

本研究は，国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」，および，科研費基盤研究 (B)『「昭和話し言葉コーパス」の構築による話し言葉の経年変化に関する実証的研究』の一環で行われたものである。本環境を設計するにあたり，国立国語研究所の川端良子氏から貴重なご意見をいただいた。また，国立国語研究所の西川賢哉氏，小磯花絵氏には，データの作成・利用方法に関して情報を提供していただいた。深く感謝いたします。

文　献

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』構築」　言語処理学会第 24 回年次大会発表論文集, pp. 775–778.

H. Brugman, and A. Russel (2004). “Annotating Multimedia/ Multi-modal resources with ELAN.” *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.*

Paul Boersma, and David Weenink (2001). “PRAAT, a system for doing phonetics by computer.”, 5, pp. 341–345.

樋口耕一 (2003).「コンピュータ・コーディングの実践—漱石『こころ』を用いたチュートリアル—」, 24, pp. 193–214.

山口昌也・田中牧郎 (2005).「構造化された言語資料に対する全文検索システムの設計と実現」　自然言語処理, 12:4, pp. 55–77.

「よい子」って誰？
–政策ニュース映画のナレーション表現に関する研究の一環として–

春木 良且(フェリス女学院大学 国際交流学部)†
田中 弥生(東京大学大学院 総合文化研究科)

Who is a 'YOI-KO'?
As a Part of the Research on Narration Representation of Municipal Newsreel

Yoshikatsu Haruki (Ferris University) †
Yayoi Tanaka (University of Tokyo)

要旨

本研究では、昭和 20 年代後半から 30 年代に掛けて、各自治体で制作された、地域の復興を記録した行政映画（政策ニュース映画）のうち、神奈川県川崎市分を取り上げる。ニュース映画は、主に映像とナレーションから構成されているが、ここではナレーション中で、特に地域の子供を表現するときに多用される「よい子」という表現に着目した。テーマによって、全くその表現が使われないものも多く存在する。どういう子供が「よい子」なのか、よい子ではない子はなぜ、よい子とは呼ばれないのか、コンテンツとナレーションの相関から傾向を洗い出して行く。

１．政策ニュース映画の概要とナレーションの意義

１.１　政策ニュース映画とは

テレビが一般化する以前、映画媒体によってニュースが製作され、映画館などで本編の上映前などに流されていた。戦後になって、主に復興の記録として、様々な制作主体によって盛んに製作されたが、特に各自治体によって地域の広報としてつくられたものを、「政策ニュース映画（municipal newsreel）」と総称する。

政策ニュース映画は、主に昭和 27（1952）年のサンフランシスコ講和条約の発効前後から製作され始めており、昭和 30 年代後半から 40 年代に掛けて、東京オリンピック前後あたりが全盛期だったと言えるだろう。テレビの普及によって、映画が完全に娯楽のものとなって、報道の役割を失ってから、ニュース映画自体は衰退して行った。政策ニュースも、それに合わせるように、自治体のテレビ番組や Web ニュース等に移行して行った。尚、現在見ることが出来る、戦後最初期のものは、昭和 23（1948）年度の茨城県政ニュースであり、最後のものは、平成 19（2007）年の神奈川県のものである。

製作主体は、広域普通地方公共団体（都道府県）単位のものが多く、基礎的地方公共団体（主に市区）レベルでは、県単位で作成されたものの一環として、管理されているものと、市レベルで独自に製作されたものがある。市区レベルの全体像は把握できなかったが、現在ネット等で公開されているもので言えば、神奈川県横浜市、川崎市、山形県山形市、愛知県一宮市、刈谷市、岡山県岡山市、静岡県浜松市などのものがある。どちらも県単位での製作がなされていた地域であり、神奈川、岡山、静岡などでは、製作会社も同じである。また愛知県一宮市、刈谷市は、県とは別に独自に製作している。また東京都の特別区でも、杉並区や台東区、練馬区で作成された記録がある。

† haruki@ferris.ac.jp

図1には、各地域の政策ニュース映画保存概況を示すが、戦争時に空襲被害が激しかった地域に、多くの記録が残されているという点から、当初は復興の記録としての性格が強かったことが推定される。

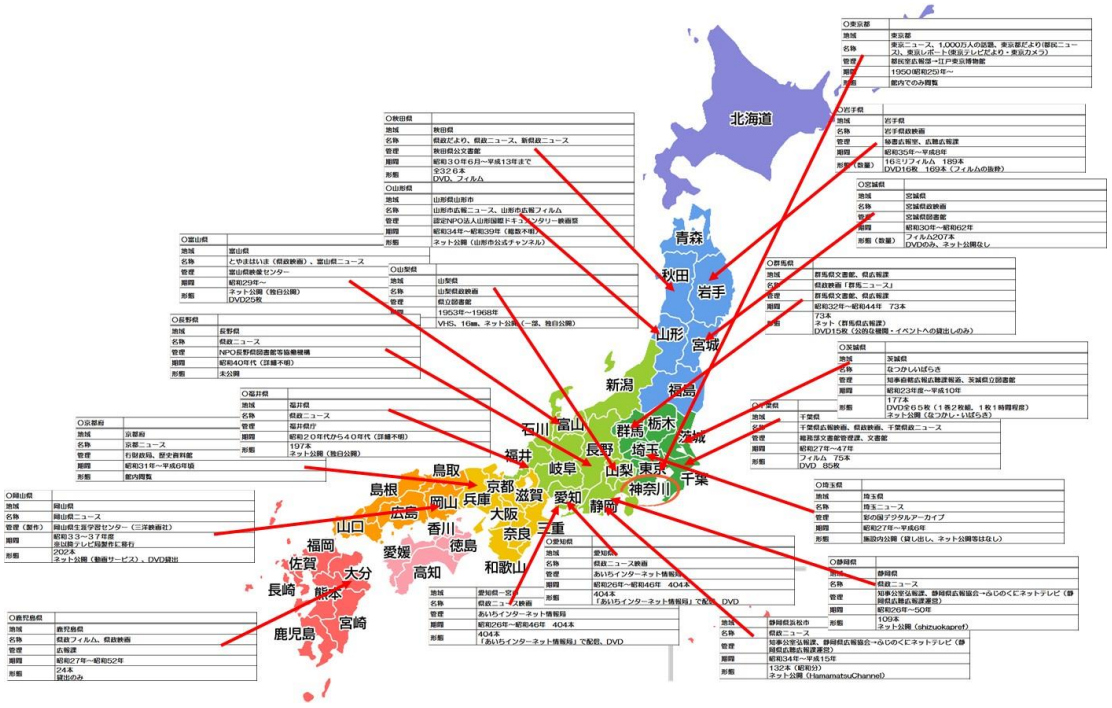


図 1　政策ニュース映画保存概況

これらの政策ニュース映画は、多く複写されて配布され、映画館や公民館などで上映された。しかし公文書ではなく行政刊行物の一種として扱われており、管理規定が明確でない等の理由から、散逸してしまったものが多々ある。また現存するものも、媒体フィルムの劣化やデジタル処理がなされていないなど、史料管理として問題を抱えているものも多い。

その存在自体が殆ど知られていないために、政策ニュース映画に関する研究は、殆どなされていないのが現状である。政策ニュース映画に記録されているものは、高度成長期を挟んだ、日本が大きく変化を遂げてきた時代の、地域の姿や市民生活などであり、公文書と個人の記録の間を繋ぐ、ミッシングリンクのような存在であると言えよう。

報告者らは、神奈川ニュース映画協会が作成し公開した神奈川県政ニュース映画のうち、川崎市の委託によるものについて、昭和 27 年から平成 19 年までの全 719 本を分析する機会を得た[1]神奈川県では、社団法人神奈川ニュース映画協会という団体によって、昭和 26(1951)年頃から、横浜、川崎、厚木などの県内諸地域の復興が記録されて行った。同団体は、昭和 25(1950)年に設立され、神奈川県や、横浜市、川崎市など、県内の公共団体の施策と事業を PR するニュース映画や記録映画、教育映画などを数多く製作してきたが、平成 19(2007)年に役割を終え解散している。正式な記録はないが、最後に映画館で流されたニュース映画は、この神奈川ニュース映画協会のものだと言われている。

神奈川ニュース映画協会の解散にともない、多くのニュース映画が、横浜市、川崎市に移管された。川崎市の委託による分は、川崎市民ミュージアムがデジタル化等を行い、権利処理等も終えて、

[1] 本数は、計 719 件（ニュース No.29～1347）、1 件あたり約 100 秒（30 秒～119 秒）。合計再生時間：19 時間 30 分 34 秒。

現在は川崎市市民文化局が管理を行っている。また各映像は、「川崎市映像アーカイブ」としてYouTube 等で公開されており、市政関係イベントの素材としても利用されている。

これら市政ニュース映画は、昭和戦後史を記録した貴重な資料ではあるが、主に川崎市の政策内容を説明するもので、余り一般に訴求するものではない。特に、高度経済成長前の昭和 20 年代の動画は、衛生観念や人権意識などが大きく異なった時代のものでもあり、現代の感覚からは不愉快な映像や表現、題材も含まれている。

しかし、港湾部の発展やインフラの整備に合わせ、都市部へ人口が集中し、産業集積が進むことで、日本の工業技術を支えた京浜工業地帯が成立していくプロセスが記録されているという意味で、他の自治体には類を見ないほど重要な史料である。市政を記録したニュース映画は、この時代の地域や社会変化を捉えるための恰好の素材であり、敗戦後から高度成長期に至る産業資本主義社会の成立プロセスをうかがい知ることができる。既に戦後 70 年以上、高度成長期からも 60 年が過ぎた現在では、ここに記録された市民の暮らしは、十分に民俗学的な関心の対象であり、都市部における経済発展をトリガーとした、産業民俗学的な観点から分析を行っている。

１.２　市政ニュース映画の構成

前述のように、政策ニュース映画は、各広域自治体で製作されたが、地域や年代によって若干の違いがあるものの、1つのテーマに関して1，2分程度に完結したニュースが、複数本集まって 1 本となり、一回に上映された。全体では概ね、10 分から 15 分ほどの長さで、映画館では、新聞社系の制作会社による全国版のニュース映画などと併映されたようである。

川崎市政ニュース映画は、この 1 テーマごとに分割してデジタル化、管理されている。各ニュースは、タイトル画面のあとすぐ本編が始まり、ナレーションによって進行する。ニュース映画に包含されている情報は、映像そのものと、言語情報の 2 種類である。各ニュースには、それぞれタイトルがついているが、映像サイズが昭和 36 年分からビスタ化されて以降、タイトルが縦書きから横書きになり、若干文字数が多くなる傾向にある。言語情報としては他に、コンテンツ中に含まれるものとして、ナレーション、テロップ、さらに被写体中の看板や貼り紙など、いわゆる文字景観などが含まれる。ナレーション以外の音声は、基本的に BGM や実況音だけであり、音声情報としては市長の発言などごくわずかである。その基本的な構成は、最後まで変わらない。若干のテロップが付加するものもあるが、政党や設備の固有名などごくわずかである。また短時間に多くの内容を扱っているので、通常の映画やドキュメンタリーよりはシーンの転換やコマ割りがかなり細かいのが特徴である。そのため、映像に対してかなり多くのナレーションが付加されている。

図2には、報告者の作成による動画の構成表例を示す。昭和 27 年 7 月 24 日付の「川崎市政 28 周年」と題されたもので、30 秒程度のものである。川崎市政ニュース映画の中でもごく初期のもので、オープニングを含めて八つほどのシーンから構成されており、各々が 2、3 秒で目まぐるしく移り変わる。初期のものは、総時間が短く、ナレーションも比較的少ないものが多いが、本ニュースも、80 文字ほどで、全動画の中で最も少ない。

ニュース映画という性格上、映像化されているものに対しては、ほぼシーン毎にナレーションが付加されている。昭和 2、30 年代のものは、被写対象の変化が激しいことに加え、映像自体が劣化していることもあり、このナレーションが動画の分析においては重要な情報源となっている。本ニュース映画に関しては、川崎市民ミュージアムによって、ナレーションの書き起こしが行われている。

メタ情報

ニュースタイトル	川崎市制２８周年	※川崎市データより
公開日	昭和27年07月24日	※川崎市データより
ファイル名		
目録番号	S27-3	※年度-通し番号
時間	0:00:30	※[h]:mm:ss

コンテンツ情報

場所	全区	※川崎市データより
地名・キーワード	市制記念	※川崎市データより

構成表

NO	カット	カット映像	タイム	コメント・ナレーション	スーパー	コンテンツ	タグ
1	オープニング				川崎市政二十八周年 川崎		商店街
2	トロリーバス		0:00:04				旧市庁舎 トロリーバス
3	工業地タオ		0:00:06	市政28周年を迎えた川崎市では、7月1日その記念行事を行いました。			京浜工業地帯
4	川崎競輪		0:00:09				競輪 川崎競輪
						川崎競輪場	
5	川崎球場		0:00:13	人口37万、無軌道電車も走る、スマートな港湾港都としての、目覚ましい発展ぶりを、喜び合う一日でした。			野球 川崎球場
6	不明		0:00:16				
7	市庁舎 パレード		0:00:19				旧市庁舎 ボーイスカウト
8	エンド		0:00:30				花火大会

図 2　構成表例

春木他(2017)では、このナレーションに着目し、ニュース映像のコンテンツ分析を試みた。このように、市政ニュース映画においては、ナレーションを中心とした言語表現は、映像表現の説明情報、あるいはメタ情報として機能しており、コーパス化することで、動画に対するタグとして機能させることも可能であると考えられる。本研究では、言語情報に着目することで、動画の分析を試み、被写体となっている当時の社会状況の理解を目指している(図3)。

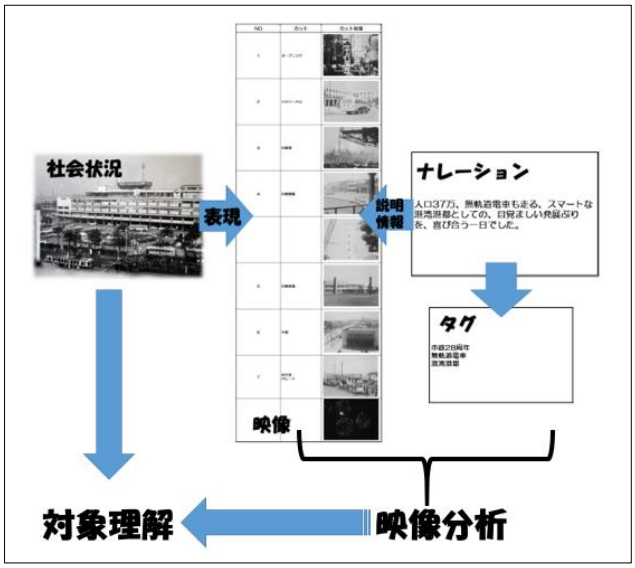


図 3　映像分析と対象理解

２．政策ニュース映画とコンテンツ

２.１　行政課題としての社会的弱者

政策ニュース映画は、地域の行政施策の記録でもあり、高齢者や障碍者、さらに子供などに対する教育、福祉系の政策は重要なテーマとしてしばしば取り上げられている。川崎市の場合、工業地帯として発展してきたことや、長く環境汚染などの問題を抱えてきたことなどから、経済至上主義的な印象があり、福祉行政のイメージが希薄なのは否定できない。しかし、ことこの市政ニュース映画を見る限り、他の地域よりはるかにきめ細かく、社会福祉等への対処がなされて来ていることがわかる。

川崎市は、昭和 20 年代には数多く存在していた、失業者やいわゆる戦争未亡人などに向けた失対事業や、授産所、障碍者施設、老人施設など、さらに障碍者教育、また集団就職などで都市部に移住してきた青少年などに対する成人教育など、いわゆる社会的弱者等に対して、意外なほど、ハード、ソフトを合わせた施策を行ってきたということが、政策ニュース映画を通して得た、新たな発見であった。

本研究では、それらの中で、特に子供についてフォーカスを当てる。教育や社会福祉を中心とした子供に対する施策は行政側としても大きな関心があるため、政策ニュース映画の題材として、子供は特に多く取り上げられている題材である。

子供に関しては、まず前提として、団塊の世代を中心とした、戦後のベビーブームの結果としての子供人口の多さを指摘する。昭和31（1956）年から始まって行く日本の高度経済成長を下支えした、多くの若手労働者の存在を「人口ボーナス」と総称するが、それらの層が子供時代だった頃の記録が、昭和 20 年代からのニュース映画には残されている。団塊の世代、戦後のベビーブーマーであるこれらの子供たちが、数年後に、労働力人口として、日本の高度成長の後期を支えていくことになる。

例えば、昭和 32 年 7 月 17 日「みんなで体操」には、ラジオ体操のために小学校の校庭に近隣の人々が 5000 人も集まった様子が記録されており、さらに昭和 32 年 5 月 15 日「こどもの日」では、「約2万人の良い子たちが、手に手に日の丸をかざして市内の目抜き通りを行進」などとナレーションと共にその様子が映される（図4）。決してフィクションでは表現できないような、凄まじいばかりの子供の多さを通して、人口ボーナスの実際とその後の高度経済成長の予兆を垣間見ることができ

ると言えるだろう。

図 4 人口ボーナス

2.2 こどもの表現

前述のように、ニュース映画の映像に対して、ナレーションは説明情報としての機能を持つため、ナレーションの分析によって、当時の政策の背後にある様々な事情や人々の価値観などを理解することが可能である。ここでは特に、一連のこども関連のニュース映画における、こどもに対する形容表現に着目する。

現代の感覚として、若干違和感を覚える表現に、ニュース映画では多用されている「良い子」「よい子」という呼称がある。主にナレーションとして多用されているが、タイトルとしても以下の 4 本で使われている。

1961/5/23 良い子の交通訓練
1965/5/25 よい子へのおくりもの
1968/8/27 良い子の願い交通安全
1969/5/25 よい子へのおくりもの

これらのうち、1965/5/25「よい子へのおくりもの」以外は、すべて交通関係のテーマである。また1969/5/25「よい子へのおくりもの」は、本編中のナレーションにも「良い子」という表現が使われている。これはおそらく、現代のテレビニュースなどでは、まず聞くことができないような表現だろう。全717 本中、ナレーション中で用いられている 12 本から、20 か所の用例を KWIC 形式で、表1に示す。

表 1「良い子」の KWIC

公開年月	表題	KWIC		
1954/1/27	羽根つき大会	豆戦士がにぎやかな観衆の応援に励まされ、羽根を打っての大熱戦。良い	子	たちは寒さを忘れて楽しい一日を過ごしました。
1954/8/18	夏の子供達	また、山あいの清流では…こうして、良い	子	たちは心身の鍛錬に真夏の太陽をいっぱいに浴びて楽しそうでした。
1954/9/15	小学校での気象観測	に余念のない児童たちがいます。それは、川崎市の西生田小学校の良い	子	たちが行っている気象観測です。昭和２３年に理科の実習で始められた
1955/8/17	夏二題・林間学校	子供たちにとって夏休みこそ大自然に親しむ絶好の機会と、川崎小学校の良い	子	たちは、都塵を避けて雄大な丹沢山麓・青根村小学校で林間学校を
1957/5/15	こどもの日	て、第二会場の市民会館に向かいました。ここでは、日本の良い	子	たちから外国のお友達に武者人形などの贈り物があり、続いて、各国の
		の日を祝う盛沢山の催しが各地で見られ、まず、約２万人の良い	子	たちが手に手に日の丸をかざして市内の目抜き通りを行進。続いて、市民
		で全市の子供会大会を開きました。さすがにこの日ばかりは大勢の良い	子	たちで広い会館も膨れ上がらんばかりの盛況ぶりでした。また、その近く
1959/8/25	緑陰の子供たち	川崎市では図書館のおじさんたちが毎日のように巡回しています。良い	子	たちも大喜びでお手伝いです。涼しい木陰のそこここで、子供たちの夢は
1961/5/23	良い子の交通訓練	全校揃って交通訓練に大変熱心です。毎週月曜日には、全校800名の良い	子	たちが、先生お手製の信号機の下で訓練に励んでいます。自転車
1965/5/25	よい子へのおくりもの	５月５日は子供の日。川崎市では、市内の良い	子	たちを向ヶ丘遊園地に招待しました。この催しは毎年行われているもの
		花と新緑で覆われた美しい園内は、終日、明るい顔の良い	子	たちでにぎわいました。また、大師、小田の両公園に、子供の夢を
1966/2/22	もうすぐ一年生	でも入学のための健康診断が市内の各小学校で行われました。良い	子	たちは、お母さんたちの見守る中で、緊張のうちにも嬉しさは隠せない
1968/8/27	良い子の願い交通安全	このほど、川崎市の産業文化会館で「ママとよい	子	の交通安全会議」が開かれました。会場に集まった良い子とお母さんたち
		ママとよい子の交通安全会議」が開かれました。会場に集まった良い	子	とお母さんたちは、市長や警察の人のお話を聞いて、事故の増えて
		の注意が大切なことを改めて心に誓い合いました。その後、お母さんと良い	子	の代表が意見を発表しましたが、良い子の代表、ヨナガクミコちゃんは、
		た。その後、お母さんと良い子の代表が意見を発表しましたが、良い	子	の代表、ヨナガクミコちゃんは、（声）この良い子たちの願いを無駄にし
		発表しましたが、良い子の代表、ヨナガクミコちゃんは、（声）この良い	子	たちの願いを無駄にしないよう、みんなで交通安全に努力いたしましょう。
1968/10/22	熱意でみのった児童公園	中、除幕・開園式を行いました。この遊園地は、何とかして良い	子	たちに遊び場を贈りたいという地元商会の熱意と役所の協力が実り、開園の
1971/11/23	わたしはデコちゃん	「デコちゃん号」と命名された機関車の前には、良い	子	たちが大勢つめかけ、この名前をつけて教育長から記念品を受けるお友達
		なんと、地球を３５周もしたのです。こうして現役を退き、川崎のよい	子	たちの生きた教材として活躍を始めるデコちゃん号です。

初出は、昭和 29(1964)年であり、昭和49(1971)年が最後であるが、タイトルとして使われている昭和 40(1965)年「よい子へのおくりもの」では、ナレーション中で全く使われていないため、「良い子」という表現は、実質的には昭和 30 年代までだったと考えていいだろう。

さらに用例として特筆すべき点として、1954/5/19「市政の民主化と広報委員会」中で、広報委員会の活動を示す行政関連資料として、「川崎時報」が映るが(図 5)、その見出しに「よい子に正しい●●(不明)を」とあり、行政資料でも使われていたことが推定できる。またその次のシーンに、1954 年 2 号と記載のある「市政グラフ」が映るが、残念ながらこれらも行政関連資料であり、筆者の調査の範囲ではあるが、どちらも発見できなかった。

図 5 川崎時報の見出し

注目されるのは、子供がテーマであるにもかかわらず、「良い子」という形容がなされていないケースが少なからずあるという点である。ナレーション中で、「良い子」「よい子」ではなく、「子供」、「こども」という表現が使われている例は、政策ニュース映画中には非常に多く見受けられる。一体、「良い子」とそうではない「子供」の違いは、どこにあるのだろうか。さらにその分化が生まれる理由はどこにあるのであろうか。

例えば、こどもの日をテーマにした、昭和 28(1953)年 5 月の「子供の日」のナレーション全文は以下である。

> *「5月5日はこどもの日。今年はうららかに晴れた五月晴れにこいのぼりの数もぐんと増え、子供たちを中心の催し物が至る所に見られました。この日、川崎市では、家庭に恵まれない孤児たちにもこの喜びを与えようと、市長さんがたくさんのおみやげを持って孤児院を訪れ不幸な子供たちを喜ばせました。また、ある孤児院の子供たちは市長室に招かれ、市長さんから激励の言葉を受けましたが、食べ盛りの子供たちをもっとも感激させたのは、美味しいごちそうのようでした。」*

家庭に恵まれない孤児、不幸な子供たちといったストレートな表現があるが、ここでは「良い子」とは表現されていない。

また昭和 33(1958)年 5 月 27 日の「カメラルポ　母子寮」でも子供が登場するが、こちらにも「良

い子」という表現はない。

> *「疲れて仕事から帰ったあと、内職をしなければならない母を助けて炊事の手伝いをする子供たち。どんなに家庭的に恵まれていなくても、苦しみや寂しさにくじけずいつも明るく生きてゆく母と子供たちです。」*

ここでも「家庭的に恵まれていない」などの形容があるが、単に子供としか表現されていない。昭和 35(1960)年 04 月 26 日「施設の子供たち」でも、共働きや母子寮、いわゆる孤児院が取り上げられているが、やはりそちらにも「良い子」という表現はない。ほかに、聾学校をテーマにした、昭和29(1954)年 2 月 24 日「聾学校の子供達」でも、「耳が聞こえない不幸な子供達を教育する、川崎市立ろう学校は、昭和26年に誕生して…子供達は楽しく勉強に励んでいます。」と、「良い子」とは表現されていない。

おそらく「良い子」という形容には、何かハンディキャップや様々な事情を抱えていないといった価値観があることが推測される。多分に感覚的な使い分けがなされているようであり、厳密に論理的な区別はできないだろうが、行政関連資料として製作されたものであるがゆえに、言葉の選び方にも、何らかの政策的な意図や判断が含まれていると考えられる。以降には、「良い子」とそうではない「子供」の用例を元に、両者の属性を明らかにする。

表2は、昭和 40 年代までのナレーション中での「子供」の用例を、KWIC 形式で整理したものである。

表 2「子供」の KWIC

公開年月	表題	KWIC		
1953/5/21	子供の日	市長さんがたくさんのおみやげを持って孤児院を訪れ不幸な	子供	たちを喜ばせました。
		また、ある孤児院の	子供	たちは市長室に招かれ、市長さんから激励の言葉を受けましたが、
		食べ盛りの	子供	たちをもっとも感激させたのは、美味しいごちそうのようでした。
1954/2/24	聾学校の子供達・	耳が聞こえない不幸な	子供	達を教育する、川崎市立ろう学校は昭和２６年に誕生して、
		先生の鬼神にも及ばぬ努力で、	子供	達は楽しく勉強に励んでいます。
1955/4/20	保育園の一日	児童福祉法に基づいて、家庭の都合で十分に面倒を見られぬ	子供	や、悪い干渉によって
		悪い干渉によってその影響が心配される	子供	などを預かる保育園が、市内の２４か所に設けられています。
		ここに預けられた	子供	たちは、一日の仕事が終わってお母さんが迎えに来るまで、
		保母さんはお母さん役となり、また、お姉さん役となって、毎日、優しく	子供	達の面倒を見ています。
		こうした保育園には、市内の約１８００人もの	子供	が温かい愛の手で育まれていますが、
1955/12/20	計量まつり	使い古しや壊れたものがたくさん集められて、火にくべられました。そして、	子供	達は引き換えに新しいものさしをもらって、大喜びでした。
1958/5/27	カメラルボ母子寮	寮の一日は、	子供	たちに心を残しながら母親が仕事に出勤していくことから始まります。
		仕事に出勤していくことから始まります。	子供	にしてみれば病気になっても看護してくれる母がいない、
		内職をしなければならない母を助けて炊事の手伝いをする	子供	たち。どんなに家庭的に恵まれていなくても、
		苦しみや寂しさにくじけずいつも明るく生きてゆく母と	子供	たちです。
1958/6/24	カメラルボろう学校を訪ねて	川崎市では、このほど、上小田中に、耳の不自由な	子供	たちのろう学校を新築しました。
1959/10/27	赤い羽根いろいろ	市内木月にある新日本学院を訪ね、親のない気の毒な	子供	たちと一緒に遊びました。
		最後に訓盲院を訪ねましたが、	子供	たちの演奏に胸を打たれそれぞれ大きな感銘を受けるとともに、募金の意義を再
1960/1/26	川崎市に簡易保育所	川崎市では、このほど、高津地区の下作延に県下でも初の試みとして、日雇い労務者の	子供	を専門に預かる簡易保育所を新設しました。
		約９０㎡の建物に満１歳の乳児を含めておよそ４０人の	子供	たちを月わずか１００円で預かるというものです。
1960/4/26	施設の子供たち	両親が働いている	子供	たちを預かる大師保育園では、６１人の園児が元気で遊んでいます。
1962/4/24	創造する幼児たち	市内の保育園では、	子供	たちを明るく伸び伸びと育てるため、昨年から絵や工作で芸術教育を行っていま
		与えられた道具を使って、	子供	たちはいろいろなものを作り出します。
		いろいろなものを作り出します。	子供	の感情、家庭環境などが、制作態度や作品を通して観察することができて、
		制作態度や作品を通して観察することができて、	子供	たちの生活指導に大いに役立っています。
1962/6/26	子供へ３つのおくりもの	京浜第二国道沿いにある南河原公園に	子供	たちのための立派な遊び場が誕生しました。
		外側を船のようにかたどったこの遊び場は、色もきれいに塗り分けられており、	子供	たちが自由に変化のある遊びができるように、いろいろな工夫が凝らされていま
		自由奔放に遊べるこの遊び場は、早くも	子供	たちの人気の的となっています。
		高津地区の橘保育園は新しい設備を備えており、	子供	たちは元気に遊び、生活指導を受けています。
1963/5/28	楽しく子供遊園会	5月１２日、向ヶ丘遊園地を一日	子供	たちに開放して、川崎市子供遊園会が開かれました。
		朝早くから集まった	子供	たちは、絵本やお菓子のプレゼントに大喜び。
		園内は一日中、	子供	たちの明るい笑顔と元気な声があふれていました。
1965/8/24	子らすこやかに	この配本所は、毎週２回、３時から７時まで開かれていますが、	子供	たちには夏休みの読書、宿題の整理などに役立つと人気を呼んでいます。
		一方、夏休みの	子供	たちの校外指導として、市では、教育委員会と警察とが協力して、
		教育委員会と警察とが協力して、	子供	たちに間違いのないよう、夜の街をパトロールしています。
		繁華街などで遅くまで遊んでいる	子供	たちは早く家に帰るよう補導します。
		こうして	子供	たちが夏休みを健やかに過ごせるよう、愛のパトロールカーは巡回を続けていま
1965/10/26	留守家庭児に勉強部屋	３時にはおやつも出て、食べ盛りの	子供	たちは大喜び。安心して働けますと、父兄たちにも好評です。
1966/2/22	もうすぐ一年生	児童の身体検査がそれぞれの学校で行われました。健やかに成長した	子供	たち。付き添いのお母さんたちの目が喜びを隠し切れません。

両者の違いを明らかにするために、その用例を、元となっている①ニュース映画のテーマと、②ナレーション語彙の特徴語、③「よい子」「子供」を形容する語、の 3 点から整理した。どのような文脈で、どういう意味合いとして、その言葉が使われるのか、どういう子供たちが良い子となり、またならないのかを明らかにする目的である。

「良い子」に関しては、全 12件のうち、学校教育、特に小学校での学びに纏わるものが 6 件、公園など地域設備に関するものが 3 件、交通安全教育に関するものが 2 件となっている。

「良い子」ではない「子供」に関しては、抽出した全15件のうち、教育をテーマにしたものが11件あるが、それらは聾学校、盲学校、養護学校など当時の「特殊教育」、現在の「特別支援教育」にあたるものが殆どであり、孤児院（現児童養護施設）2、聾学校2、保育園2、簡易保育所（現児童福祉施設）などが含まれている。通常の学校教育に関しては 1 件だけだが、テーマとしてはいわゆる鍵っ子を対象とした学内の保育施設（現学童クラブ）に関するものであり、これも特殊教育の文脈のものと言っていいだろう。

特徴語に関しては、形態素解析を行ったうえで、「良い子」と「子供」について、特に差異が大きかった語に関して、頻度を求めた。詳細は表3に示すが、ここで見るように、かなり明確に語の頻度に違いが出ている。「良い子」に関しては、友達、先生、全校、大勢、日ごろ、楽しい、訓練、大会、熱意、応援などの語が複数の頻度があったのに対して、「子供」においては、これらは軒並み出現していない。

逆に、「子供」で頻度が高い、母、勉強、家庭、指導、生活、毎日などは、「良い子」では低く、愛、感謝、環境、激励、心配、不幸、母親、さらに、施設、自由、設備、福祉などは、0である。

表 3「良い子」「子供」での語の頻度

	学校	保育	母	勉強	家庭	指導	募金	生活	毎日	夏休み	孤児	施設	自由	設備	福祉	母子	遊び場	安全	交通	愛	感謝
良い子	9	1	2	1	0	1	0	1	2	5	0	0	0	0	0	0	3	4	6	0	0
子供	13	12	11	7	5	5	4	4	4	4	3	3	3	3	3	3	3	3	3	2	2
	激励	心配	不幸	母親	楽しく	先生	気の毒	共稼ぎ	図書	市民	会館	公園	廊下	行進	全校	大勢	日ごろ	楽しい	訓練	大会	熱意
良い子	0	0	0	0	1	4	0	0	1	2	4	4	1	2	2	2	2	3	3	3	3
子供	2	2	2	2	2	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0

最後に、「良い子」「子供」それぞれの形容語であるが、表1で示したように、「良い子」では、「川崎市の西生田小学校の」、「川崎小学校の」、「日本の」、「全校800名の」、「市内の」などのように、学校を中心とした特定の組織に所属しているといった属性が示されている。

それに対して、「子供」の場合、「不幸な」、「耳が聞こえない不幸な」、「耳の不自由な」といった障碍や、「家庭の都合で十分に面倒を見られぬ」、「悪い干渉によってその影響が心配される」、「親のない気の毒な」、「両親が働いている」、「内職をしなければならない母を助けて炊事の手伝いをする」、「繁華街などで遅くまで遊んでいる」など、家庭、生活環境など、個々の事情に属するような、否定的要因が使われている。中には、「食べ盛りの」という表現もあるが、孤児院や困窮家庭の文脈などで使われているため、食事がままならないような事情を示すような表現であることにも注目される。唯一、肯定的なものとして使われている「健やかに成長した」という表現は、昭和 41（1966）年のもので、既に「良い子」が使われなくなってきた時点のものであり、なおかつ、親から見た子供に対する感情の表現となっているため、他の「子供」に対する形容語とは、若干意味合いが違っていると言えよう。

なおこれらは、この時代の価値観に基づいた表現であり、現代においては、人権等の観点から不適切な表現が含まれていることは付記しておく。

これら一連の「良い子」「こども」にまつわる表現から、大まかに「良い子」の特徴を、属性（デモグラフィック変数）、嗜好、関心など（サイコグラフィック変数）に基づき、抽出すると以下のようになるだろう（図6）。

良い子：

小学校で通常教育を受けている、両親がいる、（共働き家庭）鍵っ子ではない、先生の熱心な指導を受けている、学校では勉学のほかに、課外活動に勤しんでいる、全校の行事や市主催の行事

には積極的に参加する、夏休みの林間学校などを楽しんでいる、交通安全が一番気になっている

図 6 良い子の属性

これに対して、「良い子」ではない「子供」は、障害を持つ、鍵っ子、日雇い労務者の親、十分に面倒を見てもらえない、貧困である、繁華街などで夜遅くまで過ごす、日々の生活が関心事項であるなど、「良い子」とは全く異なった属性を持つ姿が示されている。

特に、前述のように、母、家庭、さらに生活、毎日という語が頻出しているのに加え、愛、感謝、激励、心配、不幸など、学業より日々の生活そのものが関心事項であることが推測される。また「良い子」での頻出語である、友達、先生、全校といった語が含まれないので、「良い子」が存在する場所である学校とは違った場所での日常が想定されている(図7)。

図 7「良い子」 ではない子供の属性

３．ペルソナとしての「良い子」

このように、属性から見ていくと、「良い子」の像は明確であり、その像に適合しない属性を持った子供たちが、「良い子」とは表現されなかったものと結論付けられるであろう。そのため、「良い子」ではない「子供」の属性は、社会関係や個人のもの、環境など様々な観点からのものが含まれており、一概にその姿を定義することは難しい。

そもそも政策ニュース映画は、行政側による、政策の遂行結果の広報と、これからの施策に関す

る問題提起の、大きく2つの目的を持っている。昭和 20 年代は、主に復興に向けた前者が中心だったが、復興の達成とともに、経済成長に政策的な関心が移行して行き、後者が重要になって行く。すなわち、行政課題を可視化することで、市民の理解を得るという目的で政策ニュース映画が作成されるようになっていったわけである。例えば、戦災の影響による具体的な施設の処理に纏わる最後のニュースは、川崎大空襲による被害を受けた、川崎市消防署の建て替えを取り上げた、昭和31(1956)年 3 月 21 日「望楼にニュー・スタイル」が最後である。以降、戦争に関しては、慰霊祭が取り上げられる程度となっていく。

前述のように、昭和 2, 30 年代は、戦後のベビーブーマーを中心とした、子供に対する施策が、重要な政策課題となっていた。当時の出生率と幼児死亡数を見ると、高い確率で子供の死亡があったし、昭和 20 年代では、未だ戦災孤児などの問題もあったはずであり、事情を抱えてない子供のほうが少ない状況だったと言えよう。特に政策的な観点から言えば、子供に対しては教育と、社会福祉と2つの観点からの施策が考えられる。そうした環境下で、「良い子」と表現される子供の姿は、あくまでも教育行政の対象であって、福祉行政の対象ではない。言い換えると、「良い子」ではない子供は、こと政策ニュース映画で取り上げられること自体が、何らかの課題を抱えているため、どちらかと言えば、福祉行政の対象だと言えるのではないだろうか。

川崎市は、集団就職などで工業地帯に流入してきた若年労働者に対し、いち早く成人学校を開講するなど、教育、福祉行政に対しては、先進的で熱心な地域である。こうした多くの行政課題の中で、通常の学校教育を受けている、福祉行政の対象となるような問題を抱えていない層が「良い子」と総称されているとするならば、それは行政活動における、ある種の「ペルソナ」として捉えることもできるのではないだろうか。それによって、それ以外の子供たちが抱えている課題などを強調するといった、広報、広聴上の効果があると言える。

近年、政策課題を市民レベルで検討するために、マーケティングの領域で多用されていたペルソナを用いることが、しばしば試みられるようになってきている。特にシビックテックの文脈などで、ICT やオープンデータなどの活用と共にデザイン志向の一つの試みとして行われることが多々ある。

こうした一連の動きの先駆的なものとして、この政策ニュース映画における子供や老人などの描かれ方を捉えることも可能ではないだろうか。その意味では、政策ニュース映画は単なる市民に対する広報手段だけではなく、行政側の自己評価や目標設定として、さらに市民からの広聴事項としても機能していたのではないかと推定される。「良い子」とは、政策側から見た、子供のペルソナの一つであった、本稿ではそのように結論付ける(図8)。

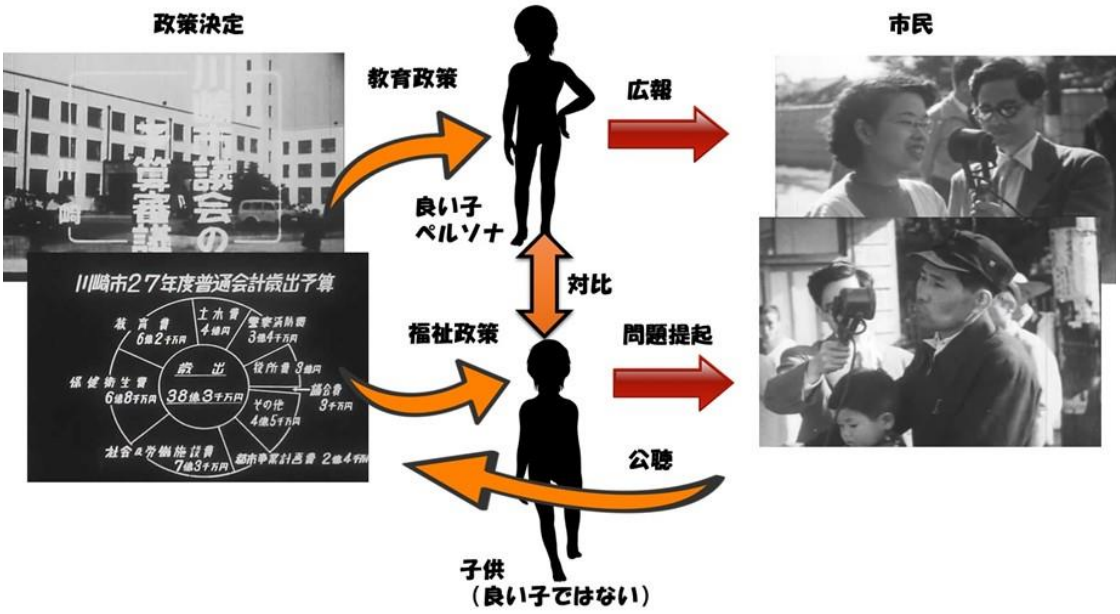


図 8　ペルソナとしての「良い子」と政策の広報、広聴

こうした政策課題を、ニュース映画中において、ある種のペルソナ的な存在を設定して可視化し、強調しながら問題提起をしていったものの例には、この他に老人の問題も指摘できる。老人の問題は、戦前は家制度を前提とした相続制度による家督相続者が担うものとされたが、戦後、家制度が廃止され、家督相続が均分相続に変わったことから、老人は「扶養」すべき存在になり、社会共通の関心事項となって行く。

昭和 28 年 09 月 16 日の「老人の楽園」は、養老院の開設に関するニュースであるが、「溝口に新設された川崎市立養老院は恵楽園と名付けられ、身寄りのない老人を収容しています。」とのナレーションが入る。「身寄りのない」という属性が強調されているが、これは家督を含む概念である、戦前の老人対策を継承した、老人像であると言えよう。この時点では、老人政策は、この「身寄りのない老人」に向けたものだったと言える。

昭和 34(1959)年に国民年金法が制定され、同年 4 月からの老人福祉年金の支給が開始する。ここから、老人は「家」ではなく、政策上の課題となって行く。昭和 33 年 09 月 23 日「としよりの日」では、「お寿司屋さんが腕によりをかけてのプレゼント。恵まれぬ100名のお年寄りたちは大喜び。」となり属性が「恵まれぬ」という形で抽象度が上がっている。ここでは、「身寄り」以外に経済面なども含んだ属性で政策対象の老人が規定されており、老人政策の変化が読み取れる。

以降、昭和 36(1961)年 4 月に軽費老人ホーム国庫補助が認められ、また昭和 38(1963)年には老人福祉法が制定される。これによって、老人福祉施設として、生活保護法の養老院を継承した養護老人ホーム、特別養護老人ホーム、そして軽費老人ホームが認められるようになって行く。この高齢者週間を取り上げたニュース映画は、以降、これを入れて 7 本製作されるが、そのいずれにも、老人を形容する語句は付加しない。

昭和 35 年 09 月 27 日　としよりの日
昭和 36 年 09 月 26 日　としよりの日
昭和 38 年 09 月 24 日　としよりの日

昭和 40 年 09 月 28 日　としよりの日
昭和 42 年 09 月 26 日　いつまでもお元気で−敬老の日ー
昭和 45 年 09 月 22 日　敬老の日

この時点において、政策の対象となるべき老人は、「身寄りのない」→「恵まれない」といった条件から、すべての「老人」へと変化していったと言えるであろう。

人口問題研究所の指摘によると、人口の高齢化が社会的に認識され始めたのは、1960 年代半ば以降のことで、1968（昭和 43）年 9 月の国民生活審議会の「深刻化するこれからの老人問題」は、年金、福祉、保健、就労、住宅対策をあげた。同月、全国社会福祉協議会は、「居宅ねたきり老人実態調査」を発表した。これを契機として、「ねたきり老人問題が社会問題となり、脳卒中などの医療対策と介護問題が課題となった。」とある。昭和 47（1972）年の中央社会福祉審議会老人福祉専門分科会による、「老人ホームのあり方に関する中間意見」での、「老人ホームを「収容の場」から「生活の場」へと転換させる必要性の指摘」は、ここにおいて家制度が完全に終了した。こうした一連の老人政策の変化は、ニュース映画で記述、表現されている老人像と明確に対応するものである。

政策ニュース映画は、広報としての役割が中心ではあったが、特に行政側が認識をしている行政課題を記述したものという側面もある。「良い子」とは、それを可視化するための、政策的なペルソナとして生み出されたものであるとするならば、老人の例で見るように、様々な政策課題において、対象の言語化、可視化をするといった手法は多用されたものと思われる。政策ニュース映画研究における、ナレーションの語彙分析の役割は決して小さくないということが、改めて明らかになったことを付記しておく。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」によるものである。また、政策ニュース映画の概況調査は、平成 29 年度放送文化基金助成事業の一環として行った。

文　献

春木 良且・田中 弥生・田村 寛之(2017)「川崎市市政ニュース映画のナレーション分析を用いた映像理解の試み ： 市民アーカイブズ構築のための枠組みとして」『言語資源活用ワークショップ発表論文集　２』, pp.239 – 251.
国民生活審議会調査部会老人問題小委員会 (著), 経済企画庁国民生活局 (編)（1968）「深刻化するこれからの老人問題―国民生活審議会調査部会老人問題小委員会報告 (1968 年) 」
中央社会福祉審議会老人福祉専門分科会「老人ホームのあり方に関する中間意見」（1972）
http://www.ipss.go.jp/publication/j/shiryou/no.13/data/shiryou/syakaifukushi/66.pdf
人口問題研究所「ワーキングペーパー」　http://www.ipss.go.jp/pr-ad/j/soshiki/ipss_j2017.pdf

関連 URL

川崎市映像アーカイブ　https://www.kawasaki-movie-archive.com/
『政策ニュース映画研究』　https://www.facebook.com/municipalnews/

敬語接頭辞異形「お～」「ご～」両者の用例のある語について

服部 匡（同志社女子大学）

Words Occurring with Both Variants of Honorific Prefix: *o-* and *go-*

Tadasu Hattori (Doshisha Women's College of Liberal Arts)

要旨

「お～」と「ご～」のどちらも伴う語の存在が知られているが、コーパスでの網羅的調査が従来なかった。青空文庫・新聞記事データベース・自作ウェブコーパスを用いた探索により、多数の語を発見した。各コーパスでの両者の使用傾向について、いくつかの観点からの観察を示す。

１．はじめに

敬語接頭辞オ/ゴの選択はある程度は語基の語種に依存し、漢語にはゴ、和語にはオがつくのが規範的パターンであるが、「お電話・お約束」「ごゆっくり・ごもっとも」のように双方向の例外があり、また、「返事」のように、オもゴも用いられる語があると言われる。しかし、実際の使用についての網羅的調査が行われていない。

本研究では、青空文庫・新聞記事・自作ウェブコーパス[1]を用い、オとゴの両方の用例のある語を探索した。

２．先行研究の概要

松下(1930)は、「「御丈夫」「御立派」「御試験」など「お」「ご」兩方にいふものも多少あるが、其れも前後の關係で大體は分かる。何となれば「お」は平易な語で「ご」は莊重な語であるからである」と述べている。

柴田(1957)は、アクセント辞典から抽出した 4,830 語について、オ/ゴをつけるのがおかしいかを 18 名に問う調査を行い、特にオ/ゴ許容率の高い 50 語のうち次の 14 語で、許容する回答がオ/ゴの両者にあったという。

(1) 馳走, 主人, 病気, 誕生, 都合, 祝儀, 出席, 援助, 商売, 食事, 仏前, 焼香, 高名, 気分

『明鏡国語辞典』編集部による 2004 年のアンケート[2]で、調査した 30 語のうち、「どちらかが圧倒的に優勢というのでなく、ゆれが見られるもの」は次の 6 語である。

(2) 誕生, 返事, 相伴, 入り用, 礼状, 葬儀

また、「どちらでもよい（場面によって使い分ける）」との回答の比率が比較的多かったのは、次の 5 語である。

(3) 返事, 誕生, 予算, 相伴, 礼状（比率が高い順）

井上(1999)は、「以前は『ご』が付いていた漢語に最近『お』が付きはじめた」例として「入学」「受験」「葬儀」を挙げている。また、規範的観点からの書物であるが、奥秋(2007)

[1] 国語研日本語ウェブコーパスは、公開 N-gram に「お」で始まる 2gram が欠落しているなどの不備があり、網羅的調査に利用できないが、特定用例の観察には利用した。浅原正幸氏によれば今後も N-gram の改善の予定はないとのことである（2018.6.6 付）。

[2] http://www.taishukan.co.jp:80/meikyo/0404/0404_top.html（現在は公開停止）。北原(2004)にも一部が掲載されている。

がテレビ番組での発話で、次の 15 語の「不適当な」オの使用例を報告している。

(4) 臨席, 出産, 注意, 遺族, 冥福, 自宅, 加入, 紹介, 遺体, 家族, 出発, 勝手(に), 返送, 両親, 担当

新聞・雑誌の語彙調査(国語研究所)に基づく分析を示した田中(1972)は、両形とも出現する語は「都合」「利息」程度しか発見できなかったという。

両形の選好について、大石(1975)は、女性はオを好む傾向があると言い、「通知」「返事」の例を挙げる。また、菊地(1994)は、「ご返事」「お返事」では、前者が尊敬語・謙譲語Ａ、後者が美化語（たとえば幼稚園言葉）という使い分けの傾向が認められるという。

３．コーパスでの出現状況

使用したコーパスは次の通りで、どれもテキストデータである。

『青空文庫全』(2007)DVD 収録作品[3]　　約 172MB　著者数 327 作品数 6,367
読売新聞記事データ集 1987-2014 年　　約 6.5GB
自作ウェブコーパス　2010 年 1 月構築　約 100GB

原データと、それを UniDic+MeCab で解析したデータを用いて探索した。青空文庫については、必要な場合、現在公開の xhtml ファイルによりルビも参照している。

なお、3 つのコーパスは総データ量が異なり、また、調査に当たり、語を選別した基準(出現回数等)も異なるので、発見された語数の比較には意味がない。

以下でいう「{お/ご}～」の出現数やその比率とは、平仮名表記されている例の中での数値である。「御」と漢字表記される例は、読み(「お・おん・ご」など)が決められないため対象から除外している。従って、すべての例の中での書き手の意図する音形の比率は不明である。

探索にあたっては、オ/ゴを伴う形式の敬語上の機能(尊敬語・美化語など)や統語的性質は問わない。また「お勝手(=台所)」/「お勝手に（=随意に)」のように明白に別語とみなしうる場合を除いては、意味的区分も行っていない。

同語の異表記は可能な限り統合した(例：返事・返辞・へんじ)。また、次のように読みの確定できない用例のある語は対象から省いたが、不徹底である。

(5) 名代(なだい・みょうだい) 微行(ちょうこう・しのび) 供物(くもつ・そなえもの) 入用(にゅうよう・いりよう) 両親(りょうしん・ふたおや)

３.１． 青空文庫の場合

「{お/ご}～」両者の用例のある形式は、少なくとも 173 に昇る。それらを「オ使用者数 /(オ使用者数+ゴ使用者数)」によって分類すると、次のようになる[4]。その形で始まる複合形式も含んだ数値である(例：ご家来衆、お誕生日)。この点は以下同じ。全例目視確認した。

(6) 80%以上のもの

愛嬌, 医師, 加減, 勘定, 客来, 綺麗, 景物, 元気, 言伝, 行列, 講義, 沙汰, 支度, 慈悲, 手配, 女中, 焼香, 丈夫, 食事, 政治, 台所, 大切, 大名, 誕生, 茶屋, 念仏, 奉行, 模様, 夕飯(ゆうはん/ゆうめし?), 立派, 料理, 牢

(7) 20%以下のもの

ゆっくり, ゆるり, 挨拶, 安心, 案内, 遠慮, 機嫌, 近所, 苦心, 苦労, 婚礼, 災難, 持参, 自身, 自分, 冗談, 心配, 先祖, 相談, 注文, 亭主, 都合, 披露, 秘蔵, 病気, 満足, 無事, 迷惑, 厄介, 用意, 立腹,

(8) その他（両形使用者数の計が５未満のものを含む)

[3] 生年の分かる日本語著者・翻訳者に限定し、同一人の異名による重複を省いたもの。
[4] あくまで青空文庫収録テキストに関する調査であり用例もそれによる。厳密には、信頼できるテキストでの確認が必要である。

衣裳, 遺骨, 家中, 家来, 会釈, 戒名, 活発, 看病, 癇癖, 奇特, 帰国, 気性, 気分, 祈祷, 記憶, 窮屈, 吟味, 決意, 検死, 検分, 見物, 後室, 公儀, 公方, 差配, 座所, 参詣, 散歩, 仕官, 支配, 時世, 次男, 自慢, 舎弟, 社参, 寿命, 修行, 住職, 祝儀, 出家, 出仕, 出馬, 出立, 書面, 助力, 勝手(に等), 商売, 将軍, 上申, 上人, 上達, 城下, 城中, 城内, 新規, 新造, 身代, 身分, 進物, 陣屋, 征伐, 政道, 接待, 説法, 詮議, 造作, 他言, 多分, 多忙, 打擲, 対面, 退屈, 大層, 堪能, 茶寮, 寵愛, 調度, 直参, 定紋, 登城, 同行, 同道, 道理, 得心, 内室, 難儀, 年始, 能面, 配下, 番士, 番所, 贔屓, 非番, 病人, 不在, 返事, 勉強, 法事, 本寺, 冥加, 名物, 面前, 遊山, 様子, 用談, 来客, 利用, 路地, 浪人, 牢屋

同一著者の両形使用

次のように、同一著者が同一語に「お～」「ご～」の両方の形を用いている場合がある[5]。

(9) 「おい君、お父さんは近頃どうしたね。相変らずお丈夫かね」(夏目漱石 1867- 明暗)

(10) 「あなたは大分ご丈夫(じょうぶ)のようですな」(同 坊ちゃん)

(11) そのうちに、弟のお機嫌をとるために、あなたの著書を弟から借りて読み、(太宰治 1909 - 斜陽)

(12) 先生のご機嫌をとろうと思って、先生の座談はとても面白い、ちょっと筆記させていただきます、と (同 黄村先生言行録)

(13) 仲哀天皇は、ある年、ご自身で熊襲をお征伐におくだりになり、筑前の香椎の宮というお宮におとどまりになっていらっしゃいました。(鈴木三重吉 1882- 古事記物語[6])

(14) お社をお作りになって、今度のご征伐についていちいちお指図をしてくださった、底筒男命以下三人の神さまを、この国の氏神さまにお祀りになった後 (同上)

著者ごとに、両方の形を用いている語の数を示すと少なくとも次の数になる。

(15) 佐々木味津三 63, 国枝史郎 7, 鈴木三重吉 6, 宮沢賢治 3, 倉田百三 2, 太宰治 2, 菊池寛 2, 夏目漱石 2, 宮本百合子 1, 岡本綺堂 1 ,楠山正雄 1, 葛西善蔵 1, 辻村もと子 1

中でも、佐々木味津三(時代小説家、1896～1934)は、(16)の 63 形式に対して「{お/ご}～」の両形を用いているが、それらは、想像による “疑似江戸語”の形を含む疑いがある(そうしたことはオ/ゴ以外にもあるかもしれない)。国枝史郎も同傾向が見える[7]。

(16) 案内, 奇特, 気性, 祈祷, 記憶, 吟味, 検分, 見物, 後室, 公儀, 公方, 行列, 差配, 沙汰, 座所, 災難, 仕官, 支配, 慈悲, 持参, 時世, 自身, 自分, 社参, 住職, 祝儀, 出仕, 出馬, 書面, 助力, 将軍, 上人, 城下, 城中, 城内, 身分, 進物, 陣屋, 詮議, 相談, 多忙, 大名, 堪能, 茶屋, 茶寮, 寵愛, 直参, 難儀, 年始, 能面, 配下, 番士, 番所, 披露, 秘蔵, 奉行, 本寺, 名物, 面前, 遊山, 様子, 牢, 牢屋

佐々木が同一作品中に二つの形を混用している例を挙げる。

(17) かりにもお将軍家お秘蔵と名のつく品なんですから、お箱の結構壮麗はいうまでもないことなので、 (右門捕物帖 明月一夜騒動)

(18) 上さまご秘蔵のご名宝が紛失いたしたとあっては捨ておかれませぬゆえ、いかにもお力となりましょう (同上)

各著者が「{お/ご}～」の形で用いる語数のうちで両者用いる語の比率が知れるとよいが、

[5] 漱石の「お機嫌」は、大阪人の「さいなら、お機嫌よう」という発話を写したもの(行人)であるように、必ずしも著者自身の言い方でないと考えられるものも含む。

[6] 子供向けの書物ということが両形使用に関係するかもしれない。

[7] もっとも、「来客」という語では、佐々木はオのみ 3 回、国枝はゴのみ 4 回用いている。

求めていない。

３.２． 現代の新聞の場合

「{お/ご}～」あわせて 10 回以上出現する形式のうち両者の用例のあるものは、少なくとも 40 ある。それらにおける「{お/ご}～」の内訳は次のようになる[8]。全例目視確認した。

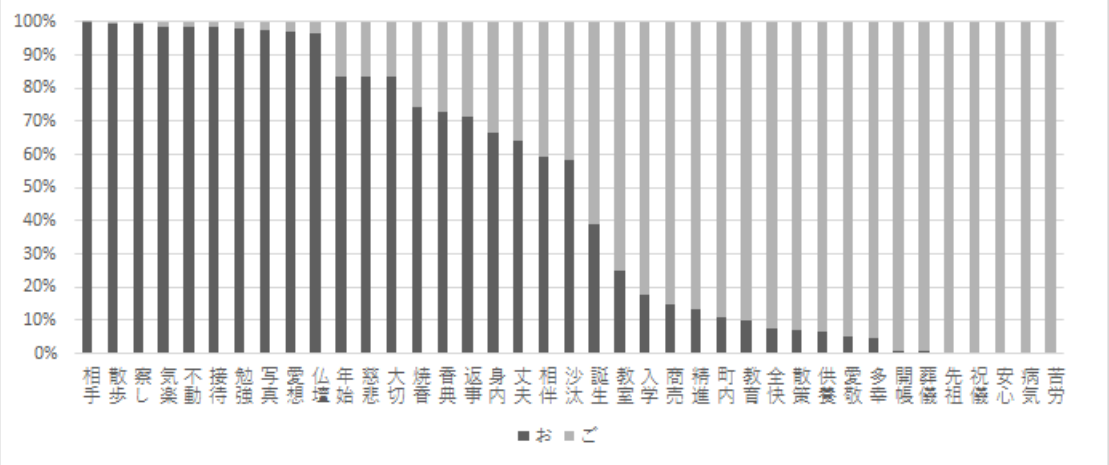


図１　「お～」と「ご～」の比率

左右両端近くの語では片方の形が 100%のように見えるが、実際には他方の形が少数ある。「少なくとも一方の用例がある語」の総数は不明である。

仏事・慶弔事関連、教育・学習関連など、ある程度共通点のある語群が見られる。

二字漢語動詞の尊敬語形式

当該用例のある「誕生・入学・散歩・散策」で、尊敬語「{お/ご}～になる」の形(諸活用形を含む)の例のすべてはゴを伴っている(例の数は順に 8, 2, 1, 1)。例えば「散歩」では、全体としては次のように「お」に大きく偏っているが、「{お/ご}散歩になる」の形だけを取るとゴの１例がすべてである。なお、ウェブコーパスでは語によりオの例も見られる(次節)。

	お	ご
散歩	343	1

(19) 先帝陛下（昭和天皇）は日曜にお子様を連れてご散歩になり、英国大使館に向いた土手の上から町の様子をご覧になった。(読売 1990.01.22)

(20) 昨年１２月に敬宮愛子さまがご誕生になったことから「親子の情愛」をテーマに決め、香川県内の羊牧場で観察、構想を練った。(読売 2002.12.07)

３.３． ウェブコーパスの場合

自作ウェブコーパスにおいて、「{お/ご}～」あわせて 100 回以上[9]出現する形式で、少ない方の形に 1%以上の使用があるものは、少なくとも 131 ある。それらにおける「{お/ご}～」の内訳は次のようになる。もっとも、ウェブコーパスはサンプリングによるものではなく、

[8] 次のような言及例を含む。内容は興味深いものである。

(i) 道警によると、詐欺サイトでは、添えられた文言が「お安心の上お求め下さい」などと不自然な表記になっていたり、日本語では使わない記号が残っているケースが目立つ。外国人がインターネット上の自動翻訳サービスを利用して日本語サイトを作っているためらしい。(読売 2014.12.22)

[9] ゴミが多いため、当該語が次の要素の直後にある例のみを数えた。
は・も・て・が・を・の・に・で・へ・から・にて・。・、・，・「・『

またそれ以前に、データには内容の重複や多くのゴミなどの問題があり(語ごとの観察により、特にゴミが多い語は省いているが)、数値の厳密な比較は意味がない。

(21) オの比率が 90%以上のもの (高い順)

奉行, 写経, 言いつけ, 仲人, 達者, 愛想, 餞別, 計らい, 命じ, 加減, 生憎, 給仕, 指図, 仏壇, 灯明, 誕生, 貴族, 盛ん, 受験, 信じ, ついで, 名刺, 追従, 察し, 精霊, 訴え, 年始, 仏像, 戒壇, 返事, 位牌, 身内, 加持, 衣装, 香典

(22) ゴの比率が 90%以上のもの (高い順)

用事, 次男, 返礼, 本堂, 真影, 拝読, 落胤, 立派, 決意, 住職, 散策, 加護, 朝食, 亭主, 法事, 用立て, 披露, 趣味, 陽気, 出現, 母堂, 健康, 講話, 見物, 遺言, 新造, 真言, 会食, 伝言, 休憩, 開帳, 納車, 寺院, 祝儀, 支配, 慈愛, 配送, 縁日, 入学, 気分, 預金, 身分, 講義, 町内, 多幸, 結納, 機会, 命日, 火葬, 仏前, 試用, 昼食, 必要, 遺骨, 供養

(23) その他 (オの比率の高い順)

大切, 遊戯, 修理, 作品, 慈悲, 説, 聴聞, 口座, 日記, 上人, 便利, 使者, 焼香, 案じ, 教化, 手配, 接待, 導師, 覚悟, 勝手(に等), 城下, 商品, 説法, 参詣, 沙汰, 相伴, 商売, 夕食, 人数, 祝詞, 葬儀, 精進, ミサ, 料金, 会釈, 供花, 祝辞, 名義, 講師, 病気

オとゴとで意味や用い方の相違が明白なものもある。例えば「遊技(遊戯)」では「お～」の大部分は幼稚園などでの活動やそれに類したものを指す一方、「ご～」はほとんどがパチンコ店などから顧客に向けたもので、「ご遊技頂けます」「ご遊技をお楽しみ下さい」などが典型的である。なお「講師」は、オかゴかを問わず、ほぼ、宗教関係の文脈のようである。

一字漢語動詞連用形

「命じ・信じ・察し・案じ」で、オ/ゴ両形の使用が見られる。「～{になる/ください/遊ばす/なさる/いたす/申す/申し上げる/なさいます}」などの尊敬語/謙譲語動詞句を形成する例が大部分であるが、「察し」に関しては「～{だ/の通り}」などの用い方もある。

これらの動詞は先頭要素が字音形態素であるが、一語化してアクセントなどの面で和語単純動詞に近い性質を持つ。そうしたことが両用傾向に関連すると思われる。

二字漢語動詞の尊敬語形式

尊敬語動詞句「{お/ご}誕生になる」(諸活用形を含む)でのオ/ゴの比率は次のようである。

	お	ご
誕生	11	58

新聞の場合とは異なってオにも一定数の用例がある。

(24) 毎度、ご来店まことにありがとうございます。本日 9/6（火）の秋篠宮紀子さまに親王殿下がお誕生になった慶事を祝しまして、お買い上げの皆様全員に、お米とお塩をプレゼントさせていただきます。

(25) クリスマスは、私たちの救い主イエス・キリストがお誕生になったことを喜ぶときです。

「朝食・昼食・夕食」「口座」など :顧客に向けた使用

3 種類の「～食」のどれにも、オとゴの両者の形が見られるが、出現傾向はかなり異なる。ゴの用例比率は、朝食 (98.8%) ＞ 昼食 (91.1%) ＞ 夕食 (75%) である。

大まかな観察の限りでは、3 語とも、飲食を供する施設や関係者による顧客向けの使用で特にゴの比率が高い。(26)はその例である。夕食は(27)(28)のように一般人のブログなどで話題になりやすい。ちなみに「食事」ではオの比率が 99.96%、「会食」では 2%である。

(26) まず、せっかく楽しみにお越し頂いた今回のご夕食にご満足頂けず、申し訳なく存じます。

(27) 今宵のお夕食は、京都・宮津から届いたばかりのブリでしゃぶしゃぶをしようとい

うことになった。

(28) 今日は１９：３０にみなさんで集まってお夕食しましょって約束になってる。

なお、「口座」のような語では、顧客向けであっても、オの方が多い。低頻度のため計数対象外であるが、「通帳」「利息」など、金融機関で用いられる語に同類が多いようである。

(29) 商品解説に無き、破れ・書き込み等ありました際には御返却いただき、お客様のお口座にご返金申し上げます。

「手配」「納車」「配送」は、客のために行う動作を指す例が多い語であるが、(先頭要素の)語種を反映して、「手配」はオの方が多く後の 2 語はゴの方が多い。

４．おわりに

オ/ゴ両者の用例のある語を３つのコーパスから報告し、いくつかの観察を示した。話者の知識において、個々の語にオ/ゴどちらかが結びつけられているわけでは必ずしもないと思われる。さらに、オとゴの機能差・使用レジスターの差異、統語パターンとの関連などの問題が残されている。

文　献

井上史雄（1999）『敬語はこわくない』 講談社.
大石初太郎（1975）『敬語』 筑摩書房.
奥秋義信（2007）『勘違い敬語の事典―型で見分ける誤用の敬語』 東京堂出版.
菊地康人 (1994) 『敬語』 角川書店.
北原保雄（2004）『問題な日本語―どこがおかしい? 何がおかしい?』 大修館書店.
柴田武 (1957) 「「お」の付く語・付かない語」 『言語生活』 70, pp. 40-49.
田中章夫 (1972) 「「オ」のつくことば・「ゴ」のつくことば」 『國文学 解釈と鑑賞』 5 月臨時増刊号, pp. 40-45.
松下大三郎（1930）『増補改訂 標準日本口語法』 勉誠社.

撥音（の解析）は機械（UniDic）にとっても簡単ではなかったんだ！ ―BCCWJ を中心に―

劉　志偉（埼玉大学）†

/N/ is not easy for UniDic as well

take BCCWJ as an example

要旨

日本語の撥音は種々雑多であるゆえ、日本語学習者にとっては学習しにくい項目である。本発表では、BCCWJ の非コアデータも視野に入れて、撥音の解析に関しては解析精度が98%に到底及ばないことを提示するとともに、具体的に「一般名詞」「オノマトペ」「漢語副詞」「漢字読み」「慣用句」「近畿方言」「呼称」「古典」「語尾」「固有名詞」「ぞんざい表現」「駄洒落」「同音異語」「動詞連用」「特定」「入力ミス」「話し言葉」「表記仮名」「表記仮名遣い」「表記漢字」「フィラー」「複合語」「（近畿以外）方言」「略語」「若者表記」「若者言葉」等の単純誤解析が多いことを明らかにする。

１．はじめに

劉（2018）では、日本語の特殊拍の一つである撥音が学習者にとって難しいことについて述べられている。現代語に限って考えても、日本語の撥音は実に種々雑多である。例えば、話し言葉には「君んち」「嫌んなる」「そんで」といった、くだけた言い方があるのに対し、書き言葉では「割れんばかりの拍手」「いざ行かん」「触れなば落ちん」等固い表現が挙げられる。また、用言の活用に関しては、いわゆる標準語においてだけでも「わかんない」「謝んなさい」「飛べんの」「かもしんない」のようにラ行音が撥音化する場合がある。さらに近畿方言の「食べんで」「行きまんねん」等も考え合わせると、教科書では「飛ぶ」のテ形「飛んで」またはタ形「飛んだ」しか習わない日本語学習者にとって撥音が極めて難解である。

一方、コーパスを用いてデータを収集する際、解析器による誤解析のうち、撥音に関するものがとりわけ多いことに気づかされる。解析器も言わば日本語を学習する存在と見なすことができる。そこで、本稿では『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を手がかりに、解析システム（UniDic）にとってどのような撥音の判定が難しいか、また日本語学習者にとっての難点との異同について考察を行う。

２．検索条件

筆者本来の目的は動詞または助動詞に後続する撥音を抽出することにあった。従って、「キー」の箇所に順次に「品詞」→「中分類」→「動詞－一般」（交替で「動詞－非自立可能」）を選択し、「後方共起条件 1」の箇所に順次に「活用形」→「小分類」→「未然形－撥音便」（交替で「連用形－撥音便」「終止形－撥音便」「 終止形－撥音便」）を設定した上で、検索ツール中納言 2.4 を用いてデータバージョン 1.1 のデータを抽出した。「動詞－一般」と「動詞－非自立可能」がそれぞれ後続する 4 種類の撥音便と組み合わせると、計 8 個のファイルのデータを収集した。

† di82zhi@yahoo.co.jp

なお、BCCWJ の解析精度については、山崎（2013）で以下のように述べられている。

> BCCWJ の形態論情報はその大半をプログラムで自動的に付与している。１億語というデータを全部人手でチェックすることは現実的ではないためである。形態論情報の精度は約 98%である（コアデータでは約 99%）。したがって、平均して 100 語に１語の解析エラーがあることになる。エラーの種類は、言語単位の区切りが違っているもの、品詞が違っているもの、読みが違っているもの等である。（115 頁）

３．結果

３.１　各ファイルの誤解析の割合

入手した 8 個のファイルをそれぞれ目視で用例を確認し、「キー」の箇所の情報を「語彙素」及び「語彙素読み」と照らし合わせて、「誤解析」と思われる数及びパーセンテージを表１に示した 。

表 1 BCCWJ における各ファイル誤解析の割合

no.	判別不可用例	近畿方言以外の方言	誤解析（キー）	誤解析（%）	後件誤解析（非撥音）	考察可能な対象例	各ファイル用例数
1	0	45	645	9.02%	4	6458	7152
2	0	0	3	5.45%	0	52	55
3	1	8	207	5.45%	9	3572	3797
4	0	3	81	9.62%	5	753	842
5	2	43	423	9.89%	3	3807	4278
6	0	0	7	36.84%	0	12	19
7	0	5	194	10.91%	22	1558	1779
8	0	2	12	6.73%	1	164	179
合計	3	106	1572	8.64%	44	16376	18101

３.２　誤解析のタイプ

誤解析の内実を明らかにすべく、本稿では BCCWJ で抽出した撥音の誤解析（計 1572 例）に対して下位区分を行った。UniDic を学習者に見立てて、間違った解析をもたらした理由に基づき 、表 2 のようなタグ付けをした。

表 2　誤解析の区分一覧

誤解析区分	用例数	誤解析区分	用例数	誤解析区分	用例数
呼称（人名を含む）	305	若者表記	39	同音異語	14
表記漢字	231	入力ミス	31	動詞連用	13
固有名詞	184	漢語副詞	28	一般名詞	12
表記仮名	138	方言	27	同字異訓	11
近畿方言	125	フィラー	27	語尾	7
漢字読み	106	古典	21	若者言葉	7
オノマトペ	76	複合語	18	表記仮名遣い	7
複合要素	66	特定	16	駄洒落	6
話し言葉	53	同音異語	14	総計	1572

４．考察

４.１　「キー」の誤解析

検索条件については 2 節で述べたように、「キー」の箇所に「動詞」（一般／非自立可能）を置き、「後方共起」（後文脈）に撥音諸形を後続させた。本節では「キー」が誤解析にな

っている場合、「実際の語」 を提示すると同時に、「区分」の箇所に誤解析をもたらした理由も示した。表 3 を参照されたい。

表 3　誤解析の諸タイプの代表例

no.	前文脈	キー	後文脈	実際の語	区分	語彙	語彙素読	サンプル ID
1	十日はソウルに滞在する生活が続いた。#夜遊びは韓国にいても変わらない。# 「やくざは夜ひとりでは	寝ん	(もん)でっせ」# 取り引きの仲間が笑いながら遊びに誘った。#「やくざは金や。#金が力や」#先輩たちから	寝る	異語異訓	休む	ヤスム	PB12_00144
2	したら商標権を侵害されたと訴えたりしないのですか？#そんな事言ったら「かっぷぬーどる」や「カッパ海老	せ	(ん)」そっくりのお菓子で違いはハングルだけと言うのも有ります。#もちろん本家日本企業は全く関係有りません。#もともと	海老せん	一般名詞	為る	スル	OC05_02403
3	ありげに見つめ合った。# メグレはそのつるはしをもってキャビンにもどった。#それから一時間以上のあいだ、憲兵は	どし	(ん)、どしんという鈍い音を聞きつづけた。#「ねえ、きみ…」# ふたたびメグレは甲板の昇降口から顔を	どしん	オノマトペ	度する	ドスル	LBl9_00192
4	冷凍おにぎりを解凍してお茶漬け状にして頂く ちっちゃくて可愛いおにぎりだが、よく出来てる#昼下	たぶ	(ン)ナポリタンを頂いてしまいそう（胃の調子が悪くて気分が悪いのが解消し、復活してきた） ＞	多分	漢語副詞	食べる	タベル	OY14_54020
5	年生まれの人たちは、西暦何年には何%存命である”という資料ってありますか？#平均余命（へ	いき	(ん)よめい）とは、ある年齢の人々が、その後何年生きられるかという期待値のことである。#生命	平均余命	漢字読み	行く	イク	OC09_10652
6	貰います。#すみ# 中村はん、ほんまにもう色々と…#うめ# いややわ。#そんなに言われたらうち居る所がの	うなり	(まん)がな。#ホホホ…そらそうとな会長はん、いつぞやお話した、ホラ、大阪に…#通仁# 布教所を作る	無くなり（ます）	近畿方言	唸る	ウナル	LBo9_00027
7	買うぞ！#TOD2買うぞ！#なんでTODはないんだばか！#がんばれテイルズ超がんばれ！#そういえば	なり	(たん)から聞いたけど坊ちゃん、マンガでてるんだって…？#買うしかないじゃないですかぁっ（ダンッ！）#さっそく	なりたん	呼称（人名を含む）	成る	ナル	OY14_36367
8	兄ちゃんに言うたとばい。#あん時どがんかしとったら」# 「もうよかて。#うちは兄さんに感謝こそ	すれ	(恨ん)どることなんかこれっぽっちもなかとやっけん。#それより兄さんの言うごとそこで働いてみようかね。#仲居さんなら	（こそ）＋する	古典	擦れる	スレル	PB39_00182
9	えらくのんびりしてるんでやすねえ」# 「いやね。#これはちょいとおまえさんには分りにくい楽しみでね」# 「そう	で	(やすか)…」# 目吉は、少し不満そうな色を浮べたものの、いくつかの腑に落ちぬことを整理し	やすか	語尾	出る	デル	LBh9_00140
10	教育の面もありました。# 一説では、遊女は客をだます狐で、それも尾のない狐だから「	尾い	(らん)」だとか。# 傾城は美人の別称で、中国の故事からきています。# 漢の李延年が帝	要らぬ	駄洒落	付く	ツク	LBa3_00020
11	泣いて、学校から帰ってきた途端力が抜けて泣いてで・・・・。#俺はどんだけ泣いたら気が	すめ	(ん)orz でもこうやって毎日泣いていってるとさ、日に日に涙は少なくなっていってるような気	済む	同音異語	住む	スム	OY14_28469
12	良かったのだと思います。#楽器屋さんによってはいろいろ吹き比べも出来るので好みの音色とお値段を照らし合わせて	選ら	(ん)でみてはいかがでしょう？#もしかしたらクランポン以外にも素敵な楽器と出会えるかもしれませんよ。	選ぶ	同字異訓	選る	エル	OC01_02482
13	。#食わせる物がなくって屋根ェ上げて風ェ食らわしとこうとな、餓鬼に着せるものがなくってアンベラへ	くる	(ン)どこうと俺の勝手だい…何を言ってやん。#他人の財政ィ立ち入りやがらァ…。#嫌なら俺ァひとり	包む	動詞連用	来る	クル	PB29_00172
14	ま」と出る状態です。#直しかたを教えてください#ALTキーを押しながら、カタカナひらがなキーを押す。#くらすちみちらかかち	しい	(とん)ら#ほらなおったでしょ	くらすちみちらかかちしいとんら	特定	為る	スル	OC02_07788
15	こちらを見ている。#それでも僕にはかすかな震えが伝わってくるんだ。#ほら、池に小石を投げ込んだら	さ	(ざ)彼が立つだろう？# あんな感じでね。#僕は歩調を落として彼女の顔に自分の顔を近づけた。#距離	さざ波	入力ミス	為る	スル	LBi9_00023
16	偉大なマナのイメージが崩壊しちゃいそうだから。#だって自我消えるかも知んないジャン！#一生懸命育てたのに！#	つか	(どん)だけ不運なの！？#（アレンの自我が消えるなんて誰も言ってないから#本業に戻ってるラビ様#キャー――	つか	話し言葉	付く	ツク	OY14_12410
17	セリエAは「せりえあー」と呼ぶのに、なぜACミランは「えーしー	み	(らん)」と言うのですか？#セリエAの正式名称は Campionato Italiano del Calcio Serie-A で、Serie	ACミラン	表記仮名	見る	ミル	OC06_03990
18	そのわけを聞いたところが、軍艦に乗り甲板に起つてゐる時の練習なのださうであつた。#休憩時間に	し	(やがん)だり、凭れたりした者は罰せられる。	しゃがむ	表記仮名遣い	為る	スル	LBa9_00077
19	をやっているわけですね。#いわゆる志布志湾波見港の公有水面埋め立てに関連しての東串良町漁協総会の有効性	いか	(ん)、こういうことでありますが、この県議会等での議論、県当局のとっている態度、これらを含め	如何	表記漢字	行く	イク	OM21_00010
20	昌史。#彼はステーションキッズという事務所で大江千里のマネージメントを担当している強者だ。#いたって冷静にうけ流す。# 「	あら	(ん)、かわいいお店」# 「わああ、いい。#いいわあ」# 普段は男まさりにサブでマネージメントしているヒロミちゃんで	あらん	フィラー	有る	アル	LBg7_00053
21	の安らぎが破れる。#昔の飲食は空腹をみたせば足りた。#それを今では林を焼き池をさらえ、生物を	切りこま	(ざ)いているではないか。# 抱朴子が言う、# 物事は現在行き過ぎがあるからといって、すべて止めて	切り細裂く	複合語	切り込む	キリコム	LBe1_00021
22	NOVAキッズのCMですが、我が家では大うけです。#「I am エーと student」#「えーと	はいら	(ん)よ。」#というのです。#お宅ではうけているCM何かありますか。#杉田かおるさんが出ている「	要らぬ	複合要素	入る	ハイル	OC01_07191
23	はらちがあがないべ。#困ったごどだなあ～。#一関の観光にとっては大打撃だじゃな。#追い討ちを	かげ	(でん)のがこのガソリン高と光熱水費の値上がりだじゃ。#な～んか写真の内容ど載せでいる話がかみあわないがら、	かける	方言	嗅ぐ	カグ	OY11_01706
24	の荒しさんが 死ねると信じていた あふぉなアタシに 教えてくれたんで、（けんか腰で	すまそ	(ん))#いびきを かいてた 彼を起こした。#「ごめん」と 言った。# 「切ったか？」	すむ	若者言葉	澄ます	スマス	OY07_00095
25		こ	(ン)ばンわン♪つー#今日は雨だったから#珍しく1日家にいた～笑#まあタ方はぶらぶらしたけどねw	こんばんは	若者表記	来る	クル	OY14_50842
26	授業。#めんどくさい。#気分のらない。#まあ、授業に気分がのる日なんて無いけどね・・・。#とりあえず、今見てる	はがれ	(ん)1期全部見終わってから学校行きたい。#見る時期間違ったかな・・・・。#夏休みまで待って・・・	ハガレン	固有名詞	剥がれる	ハガレル	OY14_52453

４.２　名大会話コーパスとの比較

コーパスの解説では「機械的に形態素解析を行い，一部手修正を行った後，結果をタグ付けしてい」ると記されている。すべてがコアデータではないということになるが、撥音に限ってみると、表４の通りほぼコアデータに匹敵する解析精度に達している。

表４　名大会話コーパスにおける各ファイル誤解析の割合

no.	判別不可用例	近畿方言以外の方言	誤解析（キー）	誤解析（%）	後件誤解析（非撥音）	考察可能な対象例	各ファイル用例数
1	1	3	1	0.08%	0	1225	1230
2	0	0	0	0.00%	0	13	13
3	0	0	0	0.00%	0	241	241
4	0	0	0	0.00%	0	13	13
5	2	1	8	1.60%	0	490	501
6	0	0	0	0.00%	0	11	11
7	0	0	0	0.00%	0	135	135
8	0	0	0	0.00%	0	2	2
合計	3	4	9	0.42%	0	2130	2146

５．結びにかえて

解析システムは人間ではないが、言語を学ぶという意味では人間と同じく日本語学習者と見なすことができる。劉（2018）で示した、学習者にとって撥音に関する学習が難しいとされる箇所と比較すると、いわゆる標準語における話し言葉（話し言葉／若者表記）、書き言葉（古典）、準標準語（近畿方言／方言）等が共通して難しいということが言えよう。また、劉（2018）では考察対象としていなかったが、「駄洒落」と「特定」は日本語学習者にとっても判定が難しいタイプであると思われる。ただし、全体的に言えば、日本語学習者に比べ、解析システムが難しいと感じる種類の方が圧倒的に多いと見なすことができる。

謝　辞

本研究は基盤研究（C)「中国語話者から見たニア・ネイティブレベルを目指すための語彙に関する総合的研究」（16K02818）の助成を受けた成果の一部である。また、調査ではBCCWJ と名大会話コーパスを利用させて頂いた。開発関係者の皆様に謝意を申し上げる。

文　献

『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版（第５章　形態論情報）
（http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html）2018 年 7 月 23 日最終確認
名大会話コーパス（全文検索システム「ひまわり」）
（https://mmsrv.ninjal.ac.jp/nucc/）2018 年 7 月 23 日最終確認
山崎誠(2013)「コーパスでできること２－BCCWJ を例に－」『日本語学』32-14、pp.104-116、明治書院
劉志偉（2016）「学習者の視点から見た「準標準語」文法項目について」『武蔵野大学日本文学研究所紀要』3、pp.53-69、武蔵野大学日本文学研究所
劉志偉(2018)「日本語教育の立場から垣間見たラ行音撥音化―日本語学習者の視点から―」『埼玉大学紀要（教養学部）』54-1、頁数未定、埼玉大学教養学部

『現代日本語書き言葉均衡コーパス』書籍サンプルに対する NDC 記号拡張アノテーションと NDC 形式区分を用いた「随筆」の文体分析

加藤 祥（国立国語研究所コーパス開発センター）†
櫻井 芽衣子（日本工業大学）
森山 奈々美（津田塾大学大学院）
浅原 正幸（国立国語研究所コーパス開発センター）

Refinement of NDC annotation on the Balanced Corpus of Contemporary Written Japanese and writing style analysis of essays based on the NDC auxiliary table

Sachi Kato (National Institute for Japanese Language and Linguistics)
Meiko Sakurai (Nippon Institute of Technology)
Nanami Moriyama (Tsuda University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

現在,『現代日本語書き言葉均衡コーパス』に含まれる書籍サンプルに付与された日本十進分類法（NDC）分類記号に，補助分類を拡張する作業を進めている。国立国会図書館の NDC 情報（8 版・9 版）を参照し，人手によって補助分類の確認と追加を行う。本発表は，現在までに作業の完了した図書館サブコーパス 10,551 サンプルについて，情報付与作業方法とその結果を報告する。本作業により，たとえば形式区分を利用し，ジャンルの分散する「随筆」「理論」「研究法」などのカテゴリで BCCWJ サンプルを分類することが可能となる。そこで，付与した形式区分「随筆」サンプル群を例とし，語彙特徴から文体的な傾向を調査した。さらに，柏野（2015）の文体指標を用い，「随筆」の文体特徴として考えられてきた「主観的」で「軟らかく」「くだけている」傾向などを確認する。

１．はじめに

あるテキストの文体的な特徴を分析するときには，著者に関する情報のほか，新聞記事か書籍か, あるいは Web 上のブログの文章かなど, ジャンルの違いに着目することが多い。文体分析にコーパスを活用し，ジャンルごとの文体特徴を明らかにすることが期待される。現在,『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）では，サブコーパスを指定し，新聞，雑誌，書籍，Web ブログなどのテキスト属性の分類が可能である。さらに，書籍は日本十進分類法（NDC）分類記号による主題の分類や，図書分類コード（C コード）による販売対象と発行形態の分類が付与されている。しかし，ジャンル分類は主に媒体と内容に基づき，書籍の形式については分類されていない。すなわち，現状，芸能人やアスリート，料理人などによる随筆は，その内容からそれぞれ芸能や産業などに分類されるため，文体分析対象の「随筆」は適切に収集し難い。反対に，「自然科学」ジャンルの典型例であろう「方法論」の文体を調査したい場合では，随筆の混在が分析結果へ影響を及ぼすともいえる。そのため，BCCWJ 公開後にも，たとえば「随筆」を調査対象とした研究は多々見られるが（立川 2014，高崎 2012 など），いずれも雑誌の「随筆」特集などから独自に収集したデータを用いた分析であり，大規模データの調査は難しい状態にあると考えられ

† yasuda-s@ninjal.ac.jp

る。そこで，BCCWJ に付与された NDC 記号を拡張し，下位分類を用いて BCCWJ サンプルを「随筆」や「理論」などのジャンルでの分類も可能とすることを目指す。本稿は，現在までに付与の完了した図書館サブコーパス 10,551 サンプルについて，情報付与作業方法とその結果を報告する。また，実際に形式区分「随筆」を用い，「随筆」サンプル群の語彙的・文体的な傾向を調査する。

２．BCCWJ 書籍サンプルに対する NDC 記号アノテーション

２.１　アノテーション対象

BCCWJ には，出版・書籍（PB：7,482 サンプル），図書館・書籍（LB：10,551 サンプル），特定目的・ベストセラー（OB：405 サンプル）の 3 種類の書籍サブコーパスがある。本研究は，これらすべての書籍サンプルに対し，現在付与されている NDC 分類記号（第一次区分：類目表・第二次区分：綱目表・第三次区分：要目表）に，下位区分（小項目，補助表[1]の形式区分・地理区分など）があれば，該当する番号を付与する。また，本稿で報告する図書館サブコーパスから開始し，順次アノテーションを進める。

BCCWJ 書籍サンプルの NDC 分類記号としては，(1)(2)(3)(4)のような 3 桁が付与されており，主題による分類が可能となっている。「少納言」や「中納言」による検索では，NDC の類目を用いたジャンル指定も可能である。

⑴サンプル ID：LBl9_00056『伊達政宗』……913
9：文学（類目），91：日本文学（綱目），913：小説・物語（要目）

⑵サンプル ID：LBl2_00027『縄文人・弥生人 101 の謎』……210
2：歴史（類目），21：日本史（綱目）

⑶サンプル ID：LBq5_00062『いい音が聴きたい』……547
5：技術（類目），54：電気工学（綱目），547：通信工学・電気通信（要目）

⑷サンプル ID：LBg4_00014『お天気博士の季節へのラブレター』……451
4：自然科学（類目），45：地学（綱目），451：気象学（要目）

本アノテーション作業により，既存の 3 桁に「.」以降の番号（下位区分）が追加される場合（(1)’(2)’(3)’(4)’に例示する）がある。すなわち，(1)では文学共通区分で時代情報が追加され，(2)では歴史の小項目が追加されるというように，さらに詳細な書籍サンプルの分類が可能となる。また，(3)(4)のように，形式区分が追加される場合は，内容ではない「事典」「随筆」などの分類が可能となる。

⑴’ サンプル ID：LBl9_00056『伊達政宗』……913.6
913（日本文学小説）.6（文学共通区分（明治以降））

⑵’ サンプル ID：LBl2_00027『縄文人・弥生人 101 の謎』……210.025

[1] NDC 新訂 9 版では，6 区分（形式区分・地理区分・海洋区分・言語区分・言語共通区分・文学共通区分）が一般補助表にあたり，類の一部分に固有補助表（細区分表）がある。なお，新訂 10 版（2017 年以降）では言語共通区分・文学共通区分が固有補助表となったが，国立国会図書館サーチ API の付与済み NDC 情報に依拠する。

210（日本史）.025（小項目（考古学））

(3)’サンプル ID：LBq5_00062『いい音が聴きたい』……547.033
547（通信工学・電気通信）.033（形式区分（事典））

(4)’サンプル ID：LBg4_00014『お天気博士の季節へのラブレター』……451.049
451（気象学）.049（形式区分（随筆））

２.２　アノテーション方法

BCCWJ の書籍サンプルについて，国立国会図書館の NDC 情報（8 版・9 版が付与されている）を参照した。各サンプルの候補となる書籍情報を収集し，人手により BCCWJ サンプルの書籍タイトル・著者・出版社・発行年を確認し，補助分類（(1)’(2)’(3)’(4)’に見られる「.」以降の番号）があれば追加を行う。なお，データの確認には，国立国会図書館サーチ API（http://iss.ndl.go.jp/information/api/）を用いた。また，BCCWJ 構築時に NDC 分類番号が確認できず，3 桁の NDC 番号が付与されていないサンプルについても，国立国会図書館データで該当書籍に NDC 情報が付与されている場合は，新規に番号を取得することとした。本稿で報告する LB サンプルに対しては，作業者 2 名が NDC 情報付与作業を行った。

２.２　LB サンプルへのアノテーション結果

LB サンプル（総数：10,551 サンプル）への情報付与結果を表 1 に示す。3 桁の番号が付与されたサンプルは 10,092 サンプルあり，「分類なし」となっていた番号のないサンプルが 459 サンプルあった。本作業により，8,690 サンプルに補助分類が追加され，83%のサンプルにおいて NDC 番号が拡張された。番号の追加がなかったサンプルは，「9X3」（要目：小説）が 18%を占めるほか，「X04」（要目：論文集）が 5%など，国立国会図書館データにおいて補助分類の付与がない書籍である。

また，BCCWJ 構築当時に「分類なし」であったサンプルのうち，410 サンプルにも NDC 番号を付与することができた。なお，NDC 番号の確認できなかった書籍（49 サンプル分）は，いずれも国立国会図書館では雑誌扱い（ムックや特集版など）とされていたものである。また，現在の BCCWJ とは番号の異なる書籍が 10 サンプル分見つかった。これらの違いの原因には，NDC の版の違いや，後の修正などが考えられる。

表 1. LB サンプルへの NDC 番号拡張アノテーション結果（いずれもサンプル数）

LB サンプル	番号付サンプル	拡張サンプル	新規番号付与	番号違い
10,551	10,092	8,690	410	10

３．NDC 情報を用いた随筆サンプルの抽出

本稿では，拡張した NDC 番号の補助区分を用い，特に文体研究に活用が可能と考えられる「随筆」サンプルを BCCWJ の LB から抽出することを試みる。

３.１　NDC による「随筆」の抽出

NDC では，文学者が書いた随筆，あるいは多数主題を扱った随筆が「9X4」（文学）に分類されるが，特定主題を扱った随筆・エッセイは「.049」の形式区分が付与される。ゆえに，BCCWJ サンプルから「随筆」テキストを収集するためには，既存の「9X4」のほか，形式区分「.049」の付与されたサンプルを抽出する必要がある。LB において「9X4」は 300 サンプルある。

３.２　LB サンプルの形式区分による「随筆」の抽出

まず，LB サンプルに付与された形式区分について整理しておく。表 2 は，本作業によって付与された NDC 補助分類の形式区分が，どのように分布しているのか整理したものである。形式区分は，全サンプルの約 10%に付与されていた。

表 2 では，「.04」が 3.6%を占めている。「.04」は「論文集」であるが，このうち，「.049」は「特定主題随筆」にあたる。「特定主題随筆」は，各ジャンル（内容による分類：NDC の類目にあたる）に分散した随筆・エッセイである。よって，「.049」の付与されたサンプルを抽出すれば，料理人やアスリートなどによって記された随筆についても調査対象とすることが可能になる。「.049」が付与されたサンプルは 81 サンプルあった。

表 2．LB サンプルの NDC 形式区分

形式区分	サンプル数	LB サンプル全体における割合
.01 理論．哲学	87	0.8%
.02 歴史的・地理的論述	275	2.6%
.03 参考図書	61	0.6%
.04 論文集．法論集．講演集．会議録	382	3.6%
.05 逐次刊行物	62	0.6%
.06 団体	51	0.5%
.07 研究法．指導法．教育	116	1.1%
.08 叢書．全集．選集	35	0.3%
計	1069	10.1%

４．随筆の特徴語彙

本節では随筆に特有な語彙を調査する。具体的には，随筆とそれ以外の 2 群（A 群：随筆と B 群：それ以外）に分け，それぞれの群における語彙素の頻度をもとに，どちらに偏っているかを対数尤度比（log-likelihood ratio: LLR）により数値化し，調査を行う。LLR は，コーパス言語学で特徴語彙を取り出すために用いられる指標で，次式によって定義する：

$$\mathbf{LLR}(\mathbf{w}) = \mathbf{2}\ (\mathbf{a}\ \log_e \boldsymbol{a} + \mathbf{b}\ \log_e \boldsymbol{b} + \mathbf{c}\ \log_e \boldsymbol{c} + \mathbf{d}\ \log_e \boldsymbol{d} - (\mathbf{a}+\mathbf{b})\ \log_e(\boldsymbol{a}+\boldsymbol{b}) - (\mathbf{a}+\mathbf{c})\ \log_e(\boldsymbol{a}+\boldsymbol{c}) - (\mathbf{b}+\mathbf{d})\ \log_e(\boldsymbol{b}+\boldsymbol{d}) - (\mathbf{c}+\mathbf{d})\ \log_e(\boldsymbol{c}+\boldsymbol{d}) + (\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d})\ \log_e(\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}+\boldsymbol{d}))$$

ここで a:A 群に出現する語彙素 w の出現頻度，b: B 群に出現する語彙素 w の出現頻度，c: A 群の延べ語数 – a ，d: B 群の延べ語数とする。LLR (w) 自体は偏りしか評価しないために，どちらの群に偏っているかを示さない。 この問題を扱うために，w の A 群における使用率 (a / a+c) が，B 群における使用率 (b / b+d)よりも小さい時に -1 を乗ずる（これを修正 LLR と呼ぶ)。

A 群を LB サンプル内の随筆(9X4 および.49)とし，B 群を随筆以外とした場合の修正 LLR 上位語（記号や固有名詞を除く）を表 3 に示す。随筆（A 群）の特徴的な語彙として，一人称代名詞の「私」や動詞の「思う」をはじめ，「ね」「か」のような読み手に語りかける終助詞や敬体の「です」，接続詞として「けれど」のようにくだけた語などが得られていることがわかる。表外でも，「僕（331.66)」「自分（288.42)」などの一人称に関する語，「好き（290.88)」「面白い（273.11)」のような評価に関する語が上位語として散見される。これらの語彙は，随筆の文体的な特徴として，主観性や語りかけ性，硬度やくだけ度などと関わる可能性が考えられる。また，「って」「と」「言う」のように引用に関わる語も見つかっている。随筆における引用は，客観的な論拠というよりも，一般論や同意を求めるため

の表現と考えられるため，専門性などとの関わりが考えられる。次節では，文体指標を用いた検証を行いたい。

表 3. LB サンプルの随筆(9X4 および .049) の特徴語彙

特徴語彙	修正 LLR
私	1939.13
言う	1259.50
ね	1223.40
です	1216.10
けれど	901.93
だ	853.68
笑い	832.26
書く	820.35
小説	816.71
って	712.22
も	704.44
か	678.31
と	598.10
思う	584.64

なお, 本稿の作業として新たに取得された「.049」分類の随筆と, 文学の類目にあたる「9X4」分類の随筆の異同を確かめておく。

表 4. LB サンプルの.049 と 9X4 の特徴語彙

.049 の特徴語彙	修正 LLR	9X4 の特徴語彙	修正 LLR
ます	757.11	私	-199.26
相続	510.34	た	-189.80
ストレス	443.08	の	-155.67
上司	407.72	小説	-145.90
検索	357.06	書く	-110.47
会社	307.81	男	-95.19
です	281.23	文学	-89.86
相手	266.17	女	-74.67
退職	261.88	家	-68.38
プレゼンテーション	249.08	カレー	-68.17
為る	232.21	作品	-65.78
クレーム	225.68	有る	-64.10

表 4 に，各々の分類の特徴語彙を示す。随筆サンプルの中でも，様々な類目に分散する「.049」と文学類目に分類される「9X4」では，それぞれ特徴語に違いが見られた。「9X4」では，随筆一般に特に多く見られた「私」のほか，「男」「女」「家」のような名詞，「小説」「書く」「文学」のような文学類目ゆえの語が特徴的に現れている。これに対し，「.049」では，「ます」「です」が特徴的であり，そのほかには表に見る「相続」「ストレス」「上司」などをはじめ，表外でも上位語には「教育（104.67)」や「企業（98.05)」，「妃（93.39)」，

「ウイルス（70.89）」,「副腎（63.99）」など，様々なジャンルの語彙であると推測される内容語が見られる。すなわち，内容語のほかで特に敬体が特徴的な文体だといえる。反対に,「9X4」分類では，敬体が特徴ではない。「9X4」分類の分析では，文学類目としての偏りや文学類目の特徴語彙が取得されるが,「.049」として各類目に分散していた「随筆」テキストを加えることにより, 敬体のような特徴語が取得できたと考えられる。次節でも,「随筆」としての総計に加え,「.049」と「9X4」の異同についても確かめておく。

５．随筆の文体

一般に，随筆には特徴的な文体傾向が現れると考えられている。ジャンル別の調査を行う際,「随筆」が着目されることは多い。しかし，これまで「随筆」の文体傾向について，大規模かつ主題横断的な調査は困難であった。そこで，BCCWJ の LB に含まれる全随筆サンプルについて，柏野（2013）の示す文体指標を参照し，随筆の文体傾向分析を行う。

国立国語研究所（2015）では，BCCWJ の LB について人手で文体分類を行い，全てのサンプルに以下の情報を付与している。

(a) 専門度：
1 専門家向き/2 やや専門的な一般向き/3 一般向き/4 中高生向き/5 小学生・幼児向き

(b) 客観度：
1 とても客観的/2 どちらかといえば客観的/3 どちらかといえば主観的/4 とても主観的

(c) 硬度：
1 とても硬い/2 どちらかといえば硬い/3 どちらかといえば軟らかい/4 とても軟らかい

(d) くだけ度：
1 とてもくだけている/2 どちらかといえばくだけている/3 くだけていない

(e) 語りかけ性度：
1 とても語りかけ性がある/2 どちらかといえば語りかけ性がある/3 特に語りかけ性はない

以下，(a)から(e)の 5 つの指標について，随筆サンプルと LB サンプル全体を対照し，随筆の文体に特徴が見られるのかを検証する。なお，表の各分布合計は総計と合致していないが，サンプルにより，文体指標の付与されていない場合がある（国立国語研究所，2015 の分類②にあたる場合,「客観度」は読み手に小説と判断されたサンプルに付与されていないなど）ことによる。

５.１　随筆の専門度

表 5 に専門度分布を示す。随筆サンプルの 8 割程度が「一般向き」と判定されており，随筆の対象読者は，概ね「一般」であると考えられる。

前節で見た特徴語彙として, 敬体が現れていた（前節の表 3･表 4 参照）ことは, 特に「.049」において，類目によっては「やや専門的な一般向き」のような内容であったとしても,「です」「ます」を使用することによって（表 4 参照）やや専門度をやわらげ,「一般向き」のテキストであるという印象を読み手に与えることに役立っている可能性がある。

表 5. LB における専門度分布

	1 専門家向き	2 やや専門的な一般向き	3 一般向き	4 中高生向き	5 小学生・幼児向き	総計
9X4	0	8	246	1	0	300
	0.0%	2.7%	82.0%	0.3%	0.0%	
.049	0	2	56	0	0	81
	0.0%	2.8%	78.9%	0.0%	0.0%	
随筆計	0	10	302	1	0	381
	0.0%	2.7%	81.4%	0.3%	0.0%	
LB 全体	141	929	7065	384	302	10551
	1.3%	8.8%	67.0%	3.6%	2.9%	

５.２　随筆の客観度

表 6 に客観度分布を示す。いわゆる随筆は，主観的であることが予想される。そして，本稿の調査結果を見ても，「とても主観的」が半数近く，「どちらかといえば主観的」をあわせた主観的傾向は，7 割程度に見られる。このことは，特徴語彙として「思う」や評価語彙（前節表 3）が現れていたこととも関連性があろう。但し，各ジャンルに分散していた（形式分類「.049」）随筆においては，書籍全体と同程度の「どちらかといえば客観的」なサンプルも得られている。「.049」と「9X4」の語彙を比較した際，「私」は「9X4」にのみ最も特徴的な語として現れていた（前節表 4 参照）。「.049」の随筆は，内容としては各ジャンルに分類された特定主題であるため，多様な主題を扱う「9X4」分類よりも「客観的」と読み取られる可能性がある。

表 6. LB における客観度分布

	1 とても客観的	2 どちらかといえば客観的	3 どちらかといえば主観的	4 とても主観的	総計
9X4	1	27	57	148	300
	0.3%	9.0%	19.0%	49.3%	
.049	0	17	16	25	81
	0.0%	23.9%	22.5%	35.2%	
随筆計	1	44	73	173	381
	0.3%	11.9%	19.7%	46.6%	
LB 全体	950	2523	1566	862	10551
	9.0%	23.9%	14.8%	8.2%	

５.３　随筆の硬度

表 7 に硬度分布を示す。書籍全体よりも「とても軟らかい」と判断されたサンプルの割合の高いことがわかる。「どちらかといえば軟らかい」割合では大差がないが，読み手が極端に「軟らかい」という印象を受けるテキストが，随筆の文章には高い割合で出現する可能性が考えられる。なお，特徴語彙（前節表 3）からテキストの硬軟の印象判定への影響は見えにくいが，敬体や「けれど」「って」のようなくだけた語の混在と，「ね」「か」のよう

な読み手への働きかけ，「思う」のような主観性などが組み合わさることで，「軟らかい」印象を与える可能性は考えられよう。

表 7. LB における硬度分布

	1 とても硬い	2 どちらかといえば硬い	3 どちらかといえば軟らかい	4 とても軟らかい	総計
9X4	1	59	164	31	300
	0.3%	19.7%	54.7%	10.3%	
.049	0	10	34	14	81
	0.0%	14.1%	47.9%	19.7%	
随筆計	1	69	198	45	381
	0.3%	18.6%	53.4%	12.1%	
LB 全体	619	3065	4440	697	10551
	5.9%	29.0%	42.1%	6.6%	

５.４　随筆のくだけ度

表 8 にくだけ度分布を示す。「とてもくだけている」「どちらかといえばくだけている」ともに，書籍全体よりも高い割合が明らかとなった。特徴語彙としても「けれど」「って」のような話しことば的と考えられる表現が見られていた（前節表 4）。反対に，「くだけていない」随筆は，随筆全体の約三分の一に留まる。なお，この傾向は，「.049」でも同様であるため，内容別のジャンル（類目）検索を行う際には，随筆の文章の影響によって，期待しない「くだけ」た用例が得られる可能性が考えられる。

表 8. LB におけるくだけ度分布

	1 とてもくだけている	2 どちらかといえばくだけている	3 くだけていない	総計
9X4	37	115	103	300
	12.3%	38.3%	34.3%	
.049	13	23	22	81
	18.3%	32.4%	31.0%	
随筆計	50	138	125	381
	13.5%	37.2%	33.7%	
LB 全体	473	2696	5652	10551
	4.5%	25.6%	53.6%	

５.５　随筆の語りかけ性度

文章であっても，読み手が語りかけるような感じを受けるテキストは，随筆のようなテキストに現れる特徴であると考えられてきた。しかし，「語りかけ性度」に着目した調査では，いわゆるハウツー本のような教示的内容を含む書籍テキスト全般において，語りかけ性があると判断される傾向があるとわかった（加藤ほか，2014）。随筆と語りかけ性度には関連性が見られるのだろうか。特徴語彙として，「ね」「か」のような直接的な語りかけと考えられる終助詞が得られた（前節表 3）ことからも，随筆は語りかけ性度が非常に高いのではないかという期待があろう。

本調査の結果，随筆では，書籍全般よりも「とても語りかけ性がある」「どちらかといえ

ば語りかけ性がある」ともに，いくらか高い割合が示された（表 9）。もっとも，LB は小説（9X3）が全体の約 3 割（2,932 サンプル）を占めるため，小説作中における作者の顔出しや一人称小説の地の文などの影響が考えられ，大差とは言い難い。また，随筆であっても「特に語りかけ性はない」が半数以上を占め，語りかけるような文体が随筆に特有であるとも言えまい。「.049」で「語りかけ性度」が若干高い割合となるのは，各分野における著名人などの特定主題の随筆に対し，読み手が教示を受けるような印象を持った可能性も考えられる。

表 9. LB における語りかけ性度分布

	1 とても語りかけ性がある	2 どちらかといえば語りかけ性がある	3 特に語りかけ性はない	総計
9X4	31	54	170	300
	10.3%	18.0%	56.7%	
.049	11	14	33	81
	15.5%	19.7%	46.5%	
随筆計	42	68	203	381
	11.3%	18.3%	54.7%	
LB 全体	833	1379	6609	10551
	7.9%	13.1%	62.6%	

５.５　随筆の文体特徴

文体指標との対照から，随筆（「9X4」「.049」）の文体は以下のような傾向が確認された。

① 一般向き
② 主観的傾向
③ 極端に軟らかい印象を受けるテキストが含まれる場合がある
④ くだけたテキストの割合が高い傾向
⑤ 語りかけるテキストの割合が高いとまでは言い難い

以上により，随筆の読者対象層は広く，読者に「主観的」かつ「くだけ」た印象を与える場合が多いといえる。一般的な「随筆」に期待されると考えられる文体特徴が検証できた。また，「随筆」には，とくに軟らかい印象を与えるテキストも含まれるという可能性も見られた。「随筆」を調査の対象とするときには，これらの特徴の関わる言語現象（4 節参照）が得やすい可能性がある。反対に，このような「随筆」がジャンル（NDC の類目）内に分散していることで，あるジャンルを調査対象とするにあたり，これらの特徴が影響を及ぼす可能性も考えられる。

６．まとめ

本研究の進める作業により，BCCWJ の書籍サンプルを，現状の内容分類（NDC の類目・綱目・要目）の小項目や形式などによって，詳細あるいは異なる基準で分類することが可能となる。本稿では，現在までに完了した LB の情報付与結果（NDC の下位区分「形式区分」）を用い，「随筆」サンプルを抽出し，大規模かつジャンル横断的な「随筆」の語彙特

徴と文体特徴の分析を試みた。文学ジャンルにとどまらないテキストを分析することで，「随筆」の特徴語彙や文体傾向が取得できた。同様に「形式区分」を用いることで，本稿で試行した「随筆」のほか，「論文集」「理論」，「研究法」「伝記」などの特定分類の抽出や分析を行うことも可能である。また，書籍の各ジャンル（NDC の類目別）の下位区分（小項目や細区分）を利用すれば，時代や地域などによる分類をはじめ，詳細に絞り込んだサンプルテキストの抽出が可能となる。PB と OB への情報付与も進め，BCCWJ 書籍サンプルのさらなる活用を図りたい。さらに，図書館情報学におけるデータの活用についても検討したい。

謝　辞

本研究は，国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」，科研費基盤(C)「文体分析を目的としたコーパスの文書情報拡張及びその利用」による。

文　献

国立国語研究所(2015) 『BCCWJ 図書館サブコーパスの文体情報』（第 1 版）

加藤祥・柏野和佳子・立花幸子・丸山岳彦(2014)「語りかける書きことばの表現」『国立国語研究所論集』8, pp.85-108

立川 和美(2014)「文章と談話における引用表現 : 随筆と雑談・相談を例として」『流通經濟大學論集』49(1), pp.31-47.

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

柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4(1), pp.43-53.

高崎 みどり(2012)「美味を意味する語の使用と性差 :「おいしい」を中心に」『お茶の水女子大学人文科学研究』8, pp.55-68.

日本図書館協会分類委員会(1995)『日本十進分類法新訂 9 版』日本図書館協会

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/

国立国会図書館サーチ　http://iss.ndl.go.jp/

『現代日本語書き言葉均衡コーパス』に対する名詞述語文アノテーション

今田 水穂 (文部科学省) *

Copular Sentence Annotation on 'Balanced Corpus of Contemporary Written Japanese'

Mizuho Imada (MEXT)

要旨

「現代日本語書き言葉均衡コーパス」コアデータに対して名詞述語文に関する文法情報付与を行った。付与したラベルの概数は名詞述語文の主語述語 13700 組、名詞述語の連体修飾節 3200、機能表現などの周辺的ラベル 4600 である。主語は名詞主語とノ節主語に分類し、前者は is_a など述語との意味関係を、後者は分裂文焦点に対する文法関係を付与した。述語は通常の名詞述語の他、「X は Y になる」のような補語も若干数付与した。周辺表現ラベルは「する」が省略された漢語動詞、名詞述語由来の機能表現、述語が省略された節などを含む。本稿では、データの設計と既存の述語項構造データとの違い、構築したデータの計量的概観について説明し、本データが名詞述語文の文法研究における諸問題とどのように関係するかについて論じる。

1. はじめに

日本語の名詞述語文研究では、現象の観点からは措定と指定の区別 (三上 1953)、ウナギ文 (奥津 1978)、述語名詞の文末表現化 (新屋 1989, 角田 2011) などが、理論的な観点からは名詞の指示性 (西山 2003) や文の情報構造 (砂川 2005) などが主要な話題として取り上げられてきた。主語と述語の意味関係という観点から見ると、多くの研究では is_a 関係や is_the 関係を表す文を名詞述語文の中心的事例と見なしており、ウナギ文などそれ以外の関係を表す文はしばしば名詞述語文の周辺的事例として扱われてきた。

表 1　be の多義性

文	意味	文	意味
吾輩は猫である	吾輩 is_a 猫	きゅうりは緑色だ	きゅうり is 緑色
田中が幹事だ	田中 is_the 幹事	子豚は 3 匹だ	子豚 amounts_to 3 匹
山田は愉快な性格だ	山田 has 愉快な性格	僕はウナギだ	僕 eats ウナギ

* imadamizuho.ac@google.com

しかしながら、名詞述語文にはウナギ文のように文脈に依存して関係が決定するようなものから、事物の属性や数量を叙述するような文脈にそれほど依存しなくても解釈が可能なものまで、is_a や is_the 以外の様々な関係を表す文があり、文の意味解釈を考える上で be の多義性を解決することは不可欠の課題である。

名詞述語文の意味を構成する主要な要素の 1 つは名詞の意味である。「猫」や「緑色」は形式意味論では 1 項述語だが、存在論的タイプが異なる。「幹事」は「誰」と「何」の 2 項を取る 2 項述語であり、「誰」をガ格、「何」をノ格で取る。別の 2 項述語には「高さ」があり、対象と数値を項に取るが、統語的な具現化は一様ではなく、「東京タワーの高さは 333m だ」「東京タワーは 333m の高さだ」などがある。名詞の項構造に関する最近の研究としては、庵 (2007)、竹内 (2015) などがある。

もう 1 つの主要な要素は名詞の意味を組み合わせて文全体の意味を作るための形式的な規則である。この規則は「東京タワーの高さは 333m だ」「東京タワーは 333m の高さだ」「東京タワーは 333m だ」などの異なる統語構造を共通の意味解釈へと結びつける必要があり、標準的な形式意味論の演算規則では必ずしも十分ではない。生成語彙論 (Pustejovsky 1998) は意味演算規則を拡張する有力な提案の 1 つであり、概念と概念の関係に関する知識 (クオリア構造) が句の意味形成において積極的な役割を果たす。形式意味論を用いた名詞述語文の最近の研究には郡司 (2015, 2016) が、生成語彙論を用いた研究には今田 (2012) がある。

本研究では、これらの意味論的研究を進めるための言語資源の整備を目的として、「現代日本語書き言葉均衡コーパス」(BCCWJ) コアデータに対する名詞述語文の述語項構造付与を行っている。同データに対する述語項構造データとしては既に BCCWJ-PAS(小町・飯田 2011) があるが、名詞述語は特にコピュラが省略された場合に述語か否かの判別が難しく、BCCWJ-PAS と本データの判定には異同がある。現段階において本研究が付与している述語項構造ラベルは、名詞述語文の主語と述語の関係を記述するものと、名詞述語が他の節の項に相当する場合にその関係を記述するものである。述語名詞が他の節の項に相当する場合とは、「肉を食べたのはトラです」のような分裂文と、「トラは肉を食べる動物です」のような連体修飾節である。他に「～は～が～だ」構文や属格のアノテーションも行いたいが、現時点では着手できていない。

意味論的情報の付与は整備したデータを用いた次の段階の研究課題だが、試験的に分裂文以外の名詞述語文について主語と述語の意味関係を is_a、is_the、has などのラベルで付与した。これらはより複雑な意味論的情報を付与するための予備的分類としての用途の他に、主語と述語の存在論的タイプが一致しない名詞述語文の数量的内訳の調査や、措定文、指定文などの名詞述語文類型についての記述的研究のための用例収集などの用途を想定している。また、名詞述語やコピュラ由来の機能表現など、名詞述語文のアノテーションの過程で見つかった周辺的表現についてもラベルを付与した。以下では、構築したデータの設計について説明し、付与したラベルの数量的内訳の報告と、記述的および理論的研究のための利活用についての展望を述べる。

2. データの設計

2.1 名詞述語文ラベル

名詞述語文に対して、主語と述語の関係を示す述語項構造ラベル (図 1 上段) と、節主語 (ノ節のみ) または連体修飾節と述語の関係を示す述語項構造ラベル (図 1 下段) を付与した(1)。

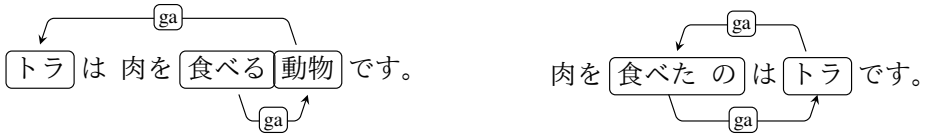


図 1　アノテーション対象とする述語項構造

ただし本データは名詞述語文のアノテーションを目的としているため、全ての情報を名詞述語を主要部とする形式に集約した。そのため、節主語や連体修飾節については、一般的な述語項構造ラベルとは矢印の向きが逆になっている。「〜は〜が〜だ」構文のように主語が複数ある場合や、連体修飾節が複数ある場合には、述語名詞に最も近いものに対してのみラベルを付与した。

図 2　本データにおけるアノテーション設計

この設計では、名詞述語文に付与するラベルは名詞述語ラベル (npred、ncomp)、主語ラベル (nsubj、csubj)、連体修飾節ラベル (acl) の 3 種類に分類される。名詞述語ラベルには npred と ncomp の 2 種類があり、npred は通常の名詞述語を表す。ncomp は「X が Y になる」「X を Y とする」など Y が節末の述語ではなく補語の位置に生起する場合で、実質的に X と Y の間に主語と述語の関係が成り立つ。

図 3　名詞述語

主語ラベルは nsubj と csubj の 2 種類がある。nsubj は名詞主語である。主語は述語に直接かかっているとは限らず、連体修飾節の被修飾語や先行文脈中の名詞句などの場合がある。テ

(1) データの構築は、既存のアノテーションデータ (BCCWJ、BCCWJ-PAS、および BCCWJ-CBL (丸山 2013) を参照してアノテーション対象文字列の候補をリストアップし、人手で修正する方法で実施した。なお、本データで使用するラベル名の一部は Universal Dependencies (http://universaldependencies.org) を参考にした。

キスト中に主語に相当する文字列がない場合 (外界参照など) は、BCCWJ-PAS の仕様に倣い、1 人称 (exo1)、2 人称 (exo2)、一般 (exog)、節参照 (ana_cla) のいずれかのラベルを付与した。

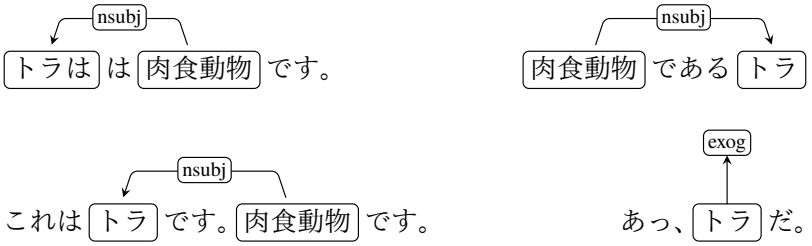


図 4　名詞主語

nsubj の下位分類として、主語と述語の関係を次の 8 種に分類し、いずれかのラベルを付与した。is_a と is_the は主語と述語の存在論的タイプが一致する。それ以外のラベルのうち、is_sth_like は隠喩に相当し、他は広義の換喩に相当する。

表 2　意味関係ラベル

ラベル	説明
is_a	主語が述語の下位クラスまたはインスタンス
is_the	主語と述語が同一
has	主語と述語が所有関係 (「X の Y」)
has_prop_of	主語と述語が所有関係 (「X の Y」と言いにくいもの)
means	定義文など
is_sth_like	比喩表現など
amounts_to	数量の叙述
other	その他

csubj は節主語である。ただし、csubj は分裂文を区別することが主な目的のため「の」節にのみ付与した。

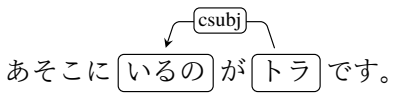


図 5　節主語

csubj の下位分類として、節主語の述語と名詞述語の文法関係を分類し、以下のいずれかのラベルを付与した。

表 3　文法関係ラベル

ラベル	説明	例
ga	ガ格	笹を食べるのはパンダだ (パンダが笹を食べる)
o	ヲ格	パンダが食べるのは笹だ (パンダが笹を食べる)
ni	ニ格	パンダがいるのは上野だ (上野にパンダがいる)
time	時間表現	パンダが来たのは先月だ (先月パンダが来た)
mod	その他の格と修飾語	パンダが来たのは中国からだ (中国からパンダが来た)
_mod	間接的修飾語	爪が鋭いのはパンダだ (パンダの爪は鋭い)
ext	外の関係	パンダが笹を食べるのは事実だ

連体修飾節ラベルは acl の 1 種類のみである。動詞、形容詞、名詞述語などの連体形の他、名詞述語の連体形に相当すると考えられる名詞 +「の」も連体修飾節として扱った。

図 6　連体修飾節

acl の下位分類として、連体修飾節の述語と名詞述語の文法関係を分類し、ラベルを付与した。ラベルは csubj の下位分類と同様だが、日本語記述文法研究会 (2008) に倣い、外の関係を次の 3 種類に分類した。

表 4　文法関係ラベル (acl 限定)

ラベル	説明	例
e1_内容補充	外の関係	パンダが笹を食べる事実
e2_相対名詞		パンダが笹を食べた後
e3_付随名詞		パンダが笹を食べた結果

2.2　機能表現ラベル等

機能表現ラベル等は、名詞やコピュラ由来の機能表現など、名詞述語文に関係する周辺的表現をマークアップしたものである。

表 5　機能表現ラベル等

ラベル	説明
vpred	動詞述語 (「する」が省略されたサ変動詞)
xpred	その他の述語
naux	名詞由来助動詞 (「はずだ」など)
nmark	名詞由来機能表現 (「中で」など)
cmark	コピュラ由来の機能表現 (「だが」など)
func	その他の機能表現 (「によって」など)
orphan	述語が省略された節末名詞句 (「トラに願いを。」)
concat	2 つ以上の文節が連結して 1 つの文節のようになっているもの
ambig	名詞述語か動詞述語か特定できないもの (「いざ調査開始！」)

vpred と xpred は名詞述語以外の述語である。vpred は名詞述語と同形の動詞述語であり、サ変名詞に後接する「する」が省略されたものである。xpred はその他の述語で 3 例あるが、いずれもイレギュラーな例であり、今後廃止する可能性がある。naux、nmark、cmark、func は名詞やコピュラに由来する機能表現である。func はそれ以外の機能表現であり、多くは格助詞 + 動詞の形式の複合助詞である。orphan は述語の省略された名詞句に付与した(2)。concat は元データである BCCWJ において、複数の文節に分割すべきものが単一の文節にまとめられている場合に付与した。ambig はラベルの判別が不可能なものに付与した。多くは、サ変名詞が述語として使用されているが、名詞に後接する「だ」ないし「する」が省略されており、かつ「だ」「する」のいずれも付加可能なため、npred か vpred かの判断ができないものである。

naux、nmark、cmark が名詞やコピュラに由来する周辺的表現であるのに対して、func、orphan、concat はいずれも名詞述語と直接関係しない言語要素だが、共通点がある。もともと名詞述語は「名詞 + で」と「ある」の 2 つの文節が結合して 1 つの文節になったものである(3)。「だ」「です」などは「である」と異なり 2 つの部分に分割することができないが、名詞述語が名詞句としての機能と述語としての機能を併せ持った文節であることは同様である。func は「を」「に」「と」などの格助詞と動詞が結合して機能語化したものであり、これも元は名詞句と述語が融合した文節と考えられる。orphan は述語を欠く名詞句だが、省略された述語はテキスト中に実体を持たないので、アノテーション仕様上は名詞句にゼロ形式の述語が包摂されたものとして扱う方法が考えられる。concat の一部も、名詞句と述語が結合して 1 つの文節になったものである。

(2) 本データの orphan は原則として文末の名詞句に限られるが、類似する構造は文中に生起することもある。1 つは「犬が肉を、猫が魚を食べた」のような部分並列構造である。部分並列構造をアノテーションしたコーパスはいくつか存在するが、コーパスによって扱いは異なる。浅原 (2013) など参照。もう 1 つは「～を目標に頑張る」のような副詞節で、「～を」のかかり先は「目標に」だが、活用する述語を欠く。

(3) 中学校の国語教科書では「だ」「です」は助動詞の一覧に含まれるが「である」は含まれないことが一般的である。この場合、「である」は「名詞 + で」と補助動詞「ある」の 2 つの文節に分割される。

表 6　結合された文節

ラベル	例		
npred	猫で + ある	→	猫である
func	猫に + よって	→	猫によって
orphan	願いを + ∅。	→	願いを。
concat	関係が + ない	→	関係がない

本研究は自動処理によりアノテーション対象の候補をリストアップし、人手で修正するという手続きを取った。これらの言語表現がリストアップされたのは名詞文節と述語文節の結合という構造的類似性によるものである。本研究の本来の目的からは外れる表現だが、構造的類似性という観点からは名詞述語の近傍の研究対象と考えることができるため、特にラベルを付与して残した。

3. データの計量的概観

3.1 概要

BCCWJ コアデータの書籍、雑誌、新聞、白書の 4 レジスタ (571 サンプル) に対してアノテーションを行い、延べ 35293 のラベルを付与した。ラベルの内訳を表 7 に示す。

表 7　ラベル数

ラベル			要素数	小計	合計
名詞述語文ラベル	述語	npred	13487	13730	30673
		ncomp	243		
	主語	nsubj	12617	13730	
		csubj	1113		
	連体修飾節	acl	3213	3213	
機能表現ラベル等	述語	vpred	1516	1519	4620
		xpred	3		
	機能表現	naux	806	1792	
		nmark	241		
		cmark	206		
		func	539		
	その他	orphan	744	1309	
		concat	493		
		ambig	72		

BCCWJ-PAS[(4)]との差異を確認する。ここでは、BCCWJ の「名詞」または「代名詞」の長単位単語のうち、長単位を構成する短単位のいずれかに PAS の述語タグが付与されているものを PAS における名詞述語と見なす。この数え方では、PAS の名詞述語は書籍、雑誌、新聞、白書の 4 レジスタで 11512 ある。本データの npred ラベルとの重なりを確認すると、本データと PAS の両方で名詞述語とするもの 8266 例、本データのみ名詞述語とするもの 5221 例、PAS のみ名詞述語とするもの 3246 例だった。PAS のみ名詞述語とするもの 3246 例について、本データで付与したタグの内訳を次の図に示す。nsubj、csubj、acl は npred ないし ncomp に付随するタグなので対応関係から除外する。remove はいずれのタグにも該当しないと見なしてタグを付与しなかったものである。

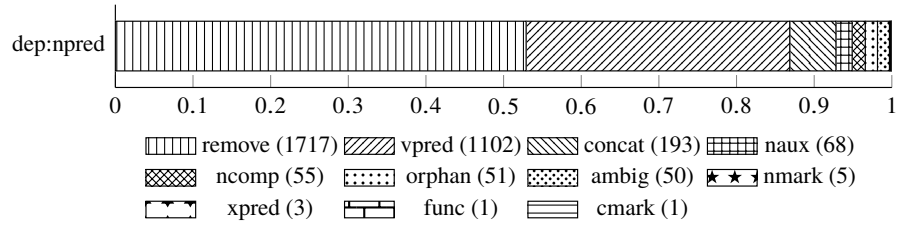


図 7　BCCWJ-PAS における名詞述語の本データにおける扱い

3246 例のうち約 5 割の 1717 例は、本データでは名詞述語文ではないと見なし、アノテーション対象から除外している。これらの多くは文全体が名詞句であるなど、文末に名詞が生起するが述語として機能していないものである。また、約 3 割の 1102 例は本データでは動詞述語 (vpred) としてアノテーションした。これらは文末にサ変名詞が生起しているが、「だ」ではなく「する」が省略されていると判断できるものである。

3.2　名詞述語文ラベル

3.2.1　名詞述語・補語

名詞述語は補語も含めて約 13700 語ある。このうち補語は 243 語と少ないが、今回のアノテーションは述語を中心に実施したので、補語は悉皆的に付与できていない。補語に後節する述語は「なる」が 141 例と最も多く、次いで「する」19 例、「いう」13 例などがあり、その他には「認める」「位置付ける」「期待する」「思う」などが見られる。名詞述語に準ずる補語については、今後、悉皆的なアノテーションを検討したい。

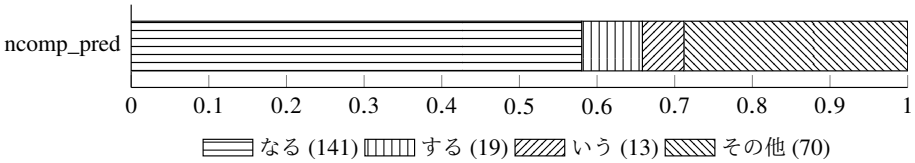


図 8　ncomp に後接する述語

(4) BCCWJ-DepParaPAS-3.3.0_1.2.0 を使用した。

3.2.2 名詞主語

主語は約 13800 語あるが、そのうち名詞主語は約 12700 語、節主語は約 1100 語である。ここで言う節主語は「の」節に限られるが、主語の大部分が名詞主語であることが分かる。

名詞主語の下位分類を確認する。名詞述語文は is_a や is_the を表すものが規範的であり、それ以外のものは周辺的である (特に語用論的な解釈を要求されるものがウナギ文と呼ばれる) と見なされる傾向がある。しかし実際には、is_a と is_the は合計しても名詞述語文全体の半分程度であり、それ以外の主語と述語の存在論的タイプが一致しないような名詞述語文が大きな割合を占めることが分かる。

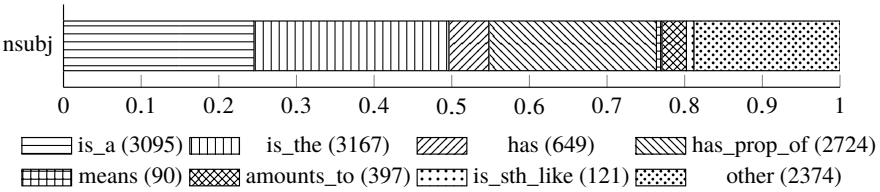


図 9　nsubj における意味関係ラベルの内訳

3.2.3 節主語

節主語の下位分類 (名詞述語との文法関係) の内訳を示す。時間表現も含めると、節主語内の要素が名詞述語として焦点化された分裂文とみなすことができるものが過半数を超えるが、分裂文ではない節主語も 4 割以上を占める。

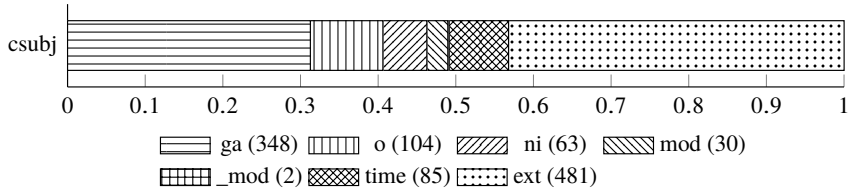


図 10　csubj における文法関係ラベルの内訳

分裂文ではない節主語が、どのような名詞述語を取るか確認する。「こと」「もの」などの形式名詞の他、「特徴」「狙い」「目的」など命題的情報を内容として持つ概念を表す名詞が多く、その内容を「〜のが特徴だ」のように節主語が具体的に示す。

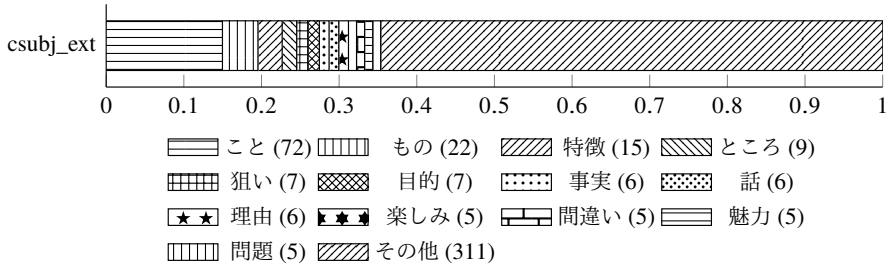


図 11　csubj(ext) における名詞述語の内訳

3.2.4 連体修飾節

連体修飾節は約 3200 例あり、名詞述語の約 23% が連体修飾節を持つ。連体修飾節の下位分類 (名詞述語との文法関係) の内訳を示す。節主語と比べると ga の割合が大きいことが分かる。また、節主語ではほとんど見られなかった_mod も、連体修飾節では比較的多く見られる。外の関係の連帯修飾節は、ほとんどが内容補充連体修飾節だった。

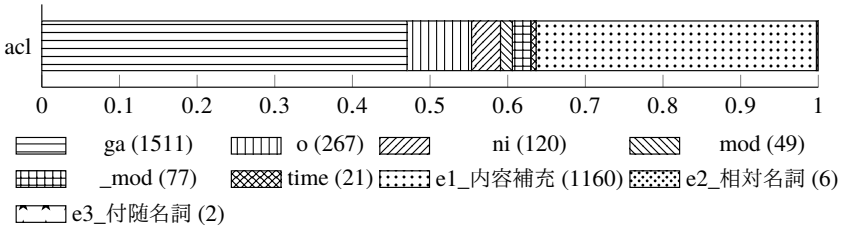


図 12　acl における文法関係ラベルの内訳

内容補充連体修飾節を取る名詞述語の内訳を確認する。分裂文ではない節主語の場合と同様、名詞述語は命題的情報を内容として持つ概念を表す名詞が多い。しかし個別の名詞を見ると「予定」「疑い」「方針」などが高頻度であり、節主語の場合とは語彙が異なる。

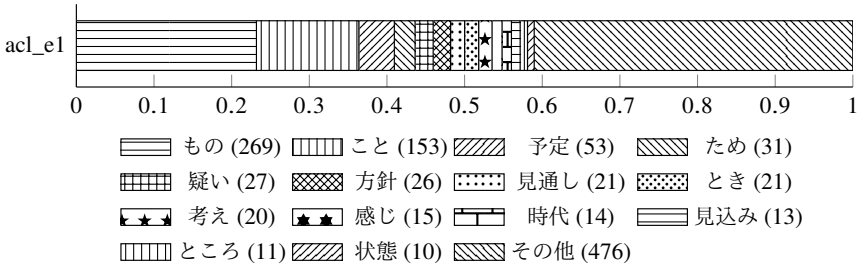


図 13　acl(e1) における名詞述語の内訳

3.3 機能表現ラベル

名詞述語文ラベル以外のラベルのうち、付与対象となる表現がある程度限られる naux、nmark、cmark、func の 4 種類の機能表現ラベルについて、主要な表現の内訳を示す。naux は名詞由来の助動詞相当表現で、ほとんどが形式名詞 +「だ」である。主なものを以下に示す。表記上の変種 (「はず」「筈」や「だ」「です」「である」など) はまとめて集計した (この節の他の図も同様)。

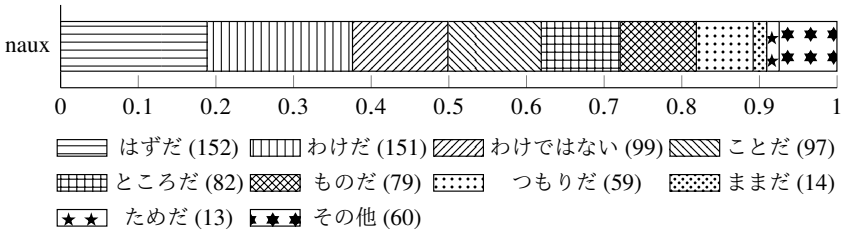


図 14　naux の内訳

nmark は名詞由来の助動詞以外の機能表現である。多くは形式名詞か、または形式名詞 + 「で」「に」の形態で、連用的な従属節を作る。主なものを以下に示す。

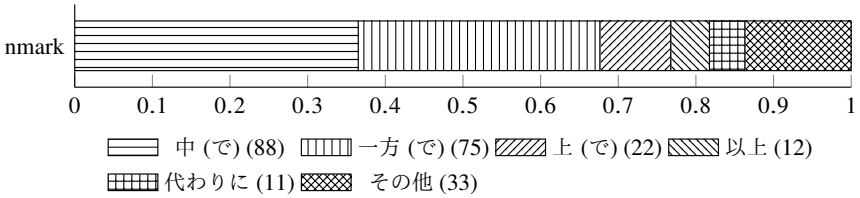


図 15　nmark の内訳

cmark は「だ」などのコピュラに由来する機能表現である。多くは、提題ないし取り立ての機能を持つもの (「だが」「だって」「だと」など) か、並列の機能を持つもの (「～だ～だ」「～だの～だの」「～だとか～だとか」) のいずれかに分類できる。主なものを以下に示す。

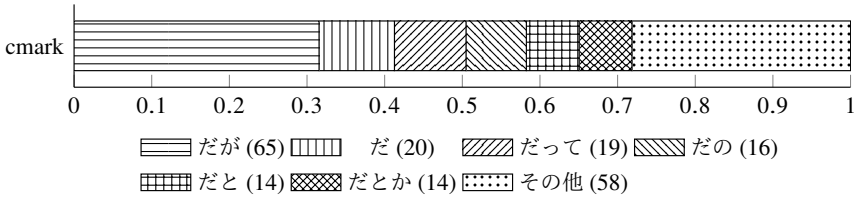


図 16　cmark の内訳

func はその他の機能表現であり、多くは格助詞 + 動詞に由来する複合助詞である。主な例を以下に示す。naux、nmark、cmark などが名詞やコピュラを由来とする名詞述語文の周辺的表現であるのに対して、func は基本的に名詞述語文とは関係がない。これらの表現が今回のアノテーションの副産物としてタグ付与された背景については、2.2 節で述べた通りである。

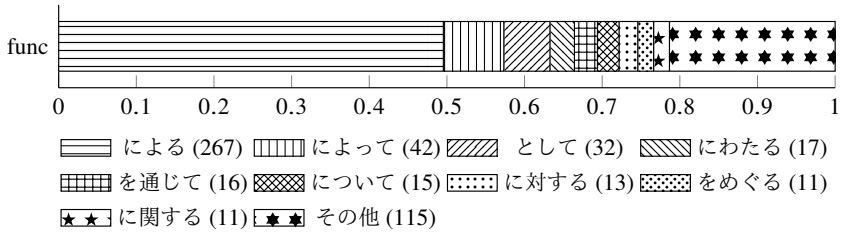


図 17　func の内訳

4. データの利活用

4.1 記述研究との接点

本データの意味関係ラベルはより複雑な意味論的情報を付与するための予備的分類として付与したものだが、一方で名詞述語文研究における主要な研究課題のために役立つことを念頭において設計している。また節主語や連体修飾節に付与した文法関係ラベルは BCCWJ-DepPara には含まれないガ、ヲ、ニ以外の文法関係もサポートしており、文法研究のために役立つ。個別の研究課題について、本データのどのラベルが関係するかについてまとめる。

■措定と指定　is_a と is_the は包摂関係と同一関係を区別するために設定したラベルであり、措定文と指定文の区別とある程度対応することを想定している。措定文、指定文の他に同一性文などの分類を設ける場合には、これも is_the に含まれる。これらの記述的分類は、名詞句の指示性 (西山 2003)、主題や焦点などの情報構造 (砂川 2005)、あるいは「指定する」「同定する」といった発話機能によって区別されるものであり、包摂関係や同一関係のような集合論的関係のみで区別されるものではない。しかし、より詳細な分析のための大まかな前処理としては役立つ。

■ウナギ文　is_a と is_the 以外の関係を表す名詞述語文は、しばしばウナギ文と総称される広範な領域に押し込められて、十分な整理が与えられてこなかった。本データでは、この領域に対して has、has_prop_of、amounts_to、other という下位区分を割り当てている。has は主語と述語の間に属格関係が成り立つ場合であり、述語が部分 (「彼はおかしな顔だ」)、属性 (「彼はおかしな性格だ」)、命題的概念 (「彼は～する予定だ」) などを表す文に適用することを想定している。他の 3 つは属格関係が成り立たない場合であり、has_prop_of は属性、amounts_to は数量、other はその他の様々な事物や概念との関係を記述することを想定している。

■隠喩文　is_sth_like は「人生は旅だ (のようなものだ)」などの比喩表現に付与するために設定したラベルである。ウナギ文がしばしば換喩と関連付けて論じられるのに対して、この構文は直喩や隠喩と関連する。特に隠喩は、形態的には通常の名詞述語文との区別がつかず、換喩と同様に意味論的な情報の付与が不可欠である。

■文末名詞文　文末名詞文は「～は～する予定だ」のように、名詞述語が文法化して文末表現のようになった文である。名詞述語は連体修飾節を持ち、名詞述語の文末表現化によって連体

修飾節が主節化する。連体修飾節が主節化するためには、主節主語が連体修飾節述語の項であり、かつ主節述語が項ではない (削除可能である) 必要があり (今田 2017)、従って連体修飾節は多くの場合、外の関係の連体修飾節である。本データでは acl の下位分類として文法関係ラベルを付与しており、このラベルは外の関係も含む。BCCWJ-DepPara の述語項構造データと組み合わせることで、文末名詞文のような複雑な構造的条件を満たす構文を効率的に抽出することができる。

■分裂文 「のは」「のが」で表示される主語は、形態的特徴から容易に抽出することが可能だが、本データでは csubj ラベルを付与し、さらに下位分類として文法関係ラベルを付与した。「のは」を主語とする名詞述語文には分裂文[(5)]とそれ以外のものがあるが、両者の判別のために文法関係ラベルが役立つ。また、同定文・提示文 (西山 2003) は主語が「のが」で表示される文を多く含む。

■定義文 means は「～とは～だ」のような定義文を記述するために設定したラベルである。主語は「とは」「というのは」で表示されるが、「は」の場合もある。述語は「～のことだ」の形式が多いが、「～という意味だ」にも means ラベルを付与した。分裂文と同様、形態的特徴から抽出しやすい構文だが、通常の名詞述語文と同形のものもあるため means ラベルが役立つ。定義文は、主語のメタ言語性、定義文と同定文の機能的類似性、定義文と分裂文の統語的類似性などの研究課題がある。

4.2 理論研究との接点

Jackendoff (2002) は文の意味構造をいくつかの層に分割して記述する。この枠組みを使うと、名詞述語文の意味構造は次のように記述することができる。

音韻/統語構造	吾輩 $_1$ は猫 $_2$ である
記述層	be(x_1, y_2)
指示層	1
情報構造層	Topic_1

本研究の意味分類ラベルは記述層の情報を記述するものであり、be の下位分類に相当する。従来の文法研究で問題とされてきた名詞述語文の意味論的特徴のいくつかについては、この構造の別の層で記述される。例えば、この構造の指示層は「吾輩」が指示的で「猫」が叙述的であることを表し、情報構造層は「吾輩」が主題であることを表す。

本データでは be の下位分類を 8 種類のラベルで表現したが、実際には be の解釈はより多様である。体系的で豊かな意味記述のために、意味関係ラベルを組織化されたより大規模な意味データベースで置き換えることを考えてみることにしよう。SUMO(Suggested Upper Merged Ontology)(Niles and Pease 2001) は述語論理をベースとしたオントロジー体系である。本データのラベルの少なくとも一部は、SUMO の述語で次のように置き換えることができる。

(5) 正確には英語の分裂文 “It is . . . that . . .” ではなく擬似分裂文 “What . . . is . . .” に相当する。

表 8　SUMO 表現

本データ	SUMO
x is_a y	instance(x, y) | subclass(x, y)
x is_the y	equal(x, y)
x has_prop_of y	attribute(x, y)
x amounts_to y	measure(x, y)
x means y	containsInformation(x, y)

しかし理論的により興味深いことは、これらの述語を be の下位分類として割り当てることよりも、様々な構文を横断して観察される同義性を一般的に記述するのに役立つということである。次の図は、BCCWJ の実例の意味解釈を SUMO 述語のネットワークで表現したものである(6)。

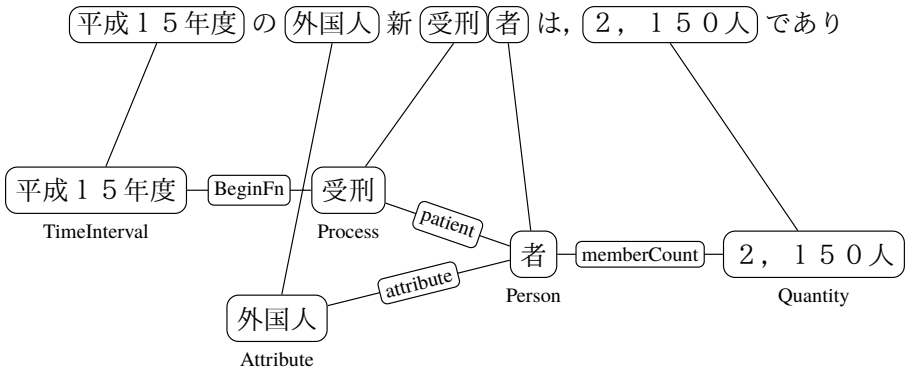


図 18　意味ネットワーク

同じ意味構成は「平成１５年の新受刑者は２，１５０人が外国人だ」「外国人新受刑者は平成１５年が２，１５０人だ」など、様々な構文で表現される。これらの構文が共有する意味情報は、統語規則、語彙規則、インターフェイス規則などによってそれぞれの構文にエンコードされるが、どの部門に多くの仕事を与えるかは理論によって異なる(7)。

統語と意味のギャップをどの部門で解消する立場を取るにせよ、運用論的な観点からは我々

(6) 簡単のために、SUMO 表現は簡略化してある。例えば「平成１５年度」と「受刑」の関係は、厳密には and(instance(x, TimeInterval), instance(y, Process), during(BeginFn(y), x)) と表現される。また、memberCount は実際には Person ではなく Collection のインスタンスを項として持つ。

(7) 主流派生成文法は統語論に多くの仕事を与える傾向があるが、別の理論、例えば Sympler Syntax(Culicover and Jackendoff 2009) では統語論の仕事を最小にしてインターフェイスや意味解釈に多くの仕事を割り当てる。名詞述語文研究では、同義性の問題は、特に指定文とウナギ文において中心的な話題の 1 つになっている。指定文については、多くの研究者が「幹事は田中だ」と「田中が幹事だ」ないし「幹事なのは田中だ」との同義性を共通の基底構造からの統語的派生によって説明することを提案している (西山 1985, 西垣内 2016)。ウナギ文については、「僕はウナギだ」をより整形的な構造からの派生として説明する多くの説明が提案されており (奥津 1978)、ウナギ文と整形文の間の同義性問題として捉えることができる。一方で、仁田 (1980) のようにウナギ文の解釈を統語論ではなく意味論の問題として扱う立場もある。

がいかにして表層的な言語表現からその意味解釈へ適切に到達できるかを説明することがより重要な課題となる[8]。標準的な方法の 1 つは形式意味論的な計算に基づく解釈の解決であり、郡司 (2015, 2016) は固有名詞、1 項述語、2 項述語に相当する名詞を含む名詞述語文の意味を形式意味論の手法で分析している。生成語彙論 (Pustejovsky 1998) は、標準的な計算プロセスから外れる多様な意味の生成を形式的に記述するために役立つ。今田 (2012) は、ウナギ文を含むいくつかのタイプの名詞述語文について、意味構成の手続きを生成語彙論を用いて分析している。

本データの意味関係ラベルは be の下位分類という浅いレベルの意味情報に相当するが、意味の理論は言語をいかにして図 18 のような深いレベルの意味情報と結び付け、同義性や推論の問題を解決するかを目標とすべきである。統語と意味の対応を研究することは、言語の複雑さを理論のどの部門に割り当てるかを研究することでもあり、また組み合わせのシステムが言語や認知のどの問題を処理するべきかを明らかにすることにも繋がる。

4.3 まとめ

BCCWJ に対する名詞述語文アノテーションの概要について説明し、記述研究や理論研究との繋がりについて論じた。このデータは名詞述語文に特化した述語項構造情報と意味関係情報、および名詞述語文の周辺的な機能表現に関する情報を含み、各種の文法研究に応用することが期待できる。一方で、本研究で付与した意味関係ラベルは粒度の荒い分類であり、文法研究の様々な目的を満たすためには意味構造の他の層の情報や、より深い解釈レベルの意味との結び付きを分析する必要がある。今後、データの利活用や情報の拡充のための研究を進めたい。

付録: 物理フォーマット

本文とラベル情報を分離したスタンドオフ形式の XML である。s は文、w はラベルを表す。

```
<sample>
  <s start="0">トラはネコ科の肉食動物であるはずだ。</s>
  <w id="0" type="npred" text="肉食動物" start="7" end="11"/>
  <w id="1" type="nsubj" text="トラ" start="0" end="2" head="0"
  ↪  rel="is_a"/>
  <w id="2" type="acl" text="ネコ科" start="3" end="6" head="0" rel="ga"/>
  <w id="3" type="naux" text="はずだ" start="14" end="17"/>
</sample>
```

各ノードの属性は以下の通りである。w 要素の属性のうち、text、start、end 属性はテキスト内に実体を持つノードにのみ付与される。target 属性は type が nsubj、csubj、acl で、テキスト内に実体を持たない (外界照応の) ノードにのみ付与される。head 属性は type が nsubj、csubj、

[8] 「僕はウナギだ」を「僕の注文はウナギだ」のような別の構造に還元するアプローチは、それ自体では意味解釈の説明としての効力を持たない。このような考え方は、統語と意味のインターフェースを簡潔にするためには役立つが、我々がいかにして与えられた入力から「僕の注文はウナギだ」という整形的な構造を復元するかという問題はそのまま残される。

acl のノードにのみ付与される。

表 9　各要素の属性

ノード	属性	説明
s	start	文頭位置 (サンプル頭からの文字数)
w	id	ラベル ID
	type	ラベルの種類 (npred | ncomp | . . .)
	text	ラベル付与範囲の文字列 (外界照応以外)
	start	ラベル開始位置 (サンプル頭からの文字数)
	end	ラベル終了位置 (サンプル頭からの文字数)
	target	外界照応タイプ (exo1 | exo2 | exog | ana_cla)。type = nsubj | csubj | acl のみ。
	head	述語ラベル ID。type = nsubj | csubj | acl のみ。

謝　辞

本研究は JSPS 科研費 17H00009 の助成を受けたものです。

文　献

浅原正幸 (2013) 「係り受けアノテーション基準の比較」,『第 3 回コーパス日本語学ワークショップ予稿集』, 81–90 頁.

Culicover, Peter W and Ray Jackendoff (2009) *Simpler Syntax*, Oxford: Oxford University Press, OCLC: 874574508.

郡司隆男 (2015) 「日本語のコピュラ文の形式意味論的分析」,『トークス= Theoretical and applied linguistics at Kobe Shoin : 神戸松蔭女子学院大学研究紀要言語科学研究所篇』, 第 18 号, 13–24 頁, 3 月.

——— (2016) 「項を 2 つとる名詞コピュラ文の形式意味論的分析」,『トークス= Theoretical and applied linguistics at Kobe Shoin : 神戸松蔭女子学院大学研究紀要言語科学研究所篇』, 第 19 号, 17–28 頁, 3 月.

今田水穂 (2012) 「名詞述語文の生成語彙論的解釈」,『文藝言語研究. 言語篇』, 第 61 号, 83–101 頁.

——— (2017)「外の関係の連体修飾節を伴う名詞述語について」,『言語資源活用ワークショップ 2017 発表論文集』, 74–83 頁.

庵功雄 (2007) 『日本語におけるテキストの結束性の研究』, くろしお出版.

Jackendoff, Ray S. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution*: Oxford University Press.

日本語記述文法研究会（編） (2008) 『現代日本語文法 6 複文』, くろしお出版.

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

丸山岳彦 (2013) 「BCCWJ に対する節境界ラベルのアノテーション」,『言語処理学会第 19 回年次大会発表論文集』, 154–157 頁.

三上章 (1953) 『現代語法序説: シンタクスの試み』, 刀江書院.

Niles, Ian and Adam Pease (2001) "Towards a Standard Upper Ontology," in *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, pp. 2–9.

西垣内泰介 (2016) 「「指定文」および関連する構文の構造と派生」,『言語研究』, 第 150 巻, 137–171 頁.

西山佑司 (1985) 「措定文, 指定文, 同定文の区別をめぐって」,『慶應義塾大学言語文化研究所紀要』, 第 17 巻, 135–165 頁.

——— (2003) 『日本語名詞句の意味論と語用論: 指示的名詞句と非指示的名詞句』, ひつじ書房.

仁田義雄 (1980) 『語彙論的統語論』, 明治書院.

奥津敬一郎 (1978) 『「ボクハウナギダ」の文法 : ダとノ』, くろしお出版.

Pustejovsky, James (1998) *The Generative Lexicon*: MIT Press.

新屋映子 (1989) 「"文末名詞"について」,『国語学』, 第 159 号, p88–75 頁.

砂川有里子 (2005) 『文法と談話の接点: 日本語の談話における主題展開機能の研究』, くろしお出版.

竹内孔一 (2015) 「名詞の項構造データの構築」,『第 8 回コーパス日本語学ワークショップ予稿集』, 233–236 頁.

角田太作 (2011) 「人魚構文 : 日本語学から一般言語学への貢献」,『国立国語研究所論集』, 第 1 巻, 53–75 頁.

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/

クラウドソーシング発注文書におけるレジビリティの量的分析

岩崎 拓也（国立国語研究所理論・対照研究領域／一橋大学大学院言語社会研究科）†
井上 雄太（一橋大学大学院言語社会研究科）††

Quantitative Analysis of Legibility in Crowdsourcing Purchase Orders

Takuya Iwasaki (NINJAL / Hitotsubashi University Graduate School of Language and Society)
Yuta Inoue (Hitotsubashi University Graduate School of Language and Society)

要旨

インターネットを介して業務を遂行するクラウドソーシングにおいて，まず重要になるのが可読性である。可読性には，文法の難易度にかかわるリーダビリティと文章の見やすさにかかわるレジビリティがある。可読性を上げるためには，わかりやすい文法や表現を使用するだけではなく，見やすい文書にしなければならない。レジビリティを上げる要素として考えられるのは，句読点やカッコといった記号，文と文の間の空白行の挿入などであるが，これらについて計量的な観点からの分析は多くない。本研究では，クラウドソーシングで実際に使われた発注文書のデータベースを元に，記号や空白行に焦点を当て，その多寡と応募数との相関の分析を行った。その結果，記号は hourly 型の支払い形式，改行は hourly 型と task 型の支払い形式，空白行は task 型の支払い形式の間において弱い相関が認められた。

１．はじめに

クラウドソーシングは，「高速」「低コスト」が利点ではあるものの，ワーカー（受注者）への対応の難しさがある。受注者を段階的に育成する方法も提案されてはいるが（芦川・川村・大須賀 2017），単発的な発注である場合は育成ができず，効果が得られにくいことも考えられる。そこで，重要になるのは発注文書である。内容が明瞭でわかりやすい発注文書であれば，より多くのワーカーが応募してくる可能性が高くなると考えられる。実際に，清水・中川（2015）では，発注文書内の質問が曖昧であるという問題を挙げている。このことは，発注者に対してどうすれば文書がわかりやすくなるかという説明が求められていることを示している。本研究では，クラウドソーシングの発注文書の本文（以下，発注文書）を対象として分析を行う。クラウドソーシングにおいて，ワーカーは Web サイトにアクセスした後，タイトル一覧を読み，自らが得意な（または作業したい）案件をクリックし，発注文書を読む。これを繰り返し行い，最も自分に合う案件に応募する。この際，本文の内容が見にくく，内容がわかりにくい場合は応募する可能性が低くなると考えられる。通常，発注文書を読みやすくする要素としては，簡潔な表現を用いることや内容を箇条書きにすることなどが挙げられる。また，箇条書きにする際や言及している内容を階層化する際に，記号の使用や空白行の挿入も行われる。このような要素は，いずれも書き手（クラウドソーシングにおいては発注者）が読み手（受注者）に対して，より目に止まるように見せたり，何らかの注釈を示したりするために使用されているものだと考えられる。実際に，有光ほか

† i.was.aki[at]hotmail.co.jp
†† inoue.yta[at]gmail.com

（2014）では，かぎカッコによる効果の検証を調査しており，調査の結果，読み手はほとんどの文においてかぎカッコがある文の方がわかりやすい文章であると感じていることを明らかにしている。しかしながら，記号による文字装飾の有無と応募数の相関，発注文書で使用される記号の種類とその使用傾向については計量的な立場からの分析は行われておらず，その実態は明らかにされていない。

２．先行研究

クラウドソーシング発注文書データベースの構築については，井上（2018）が初期段階での報告を行っている。井上（2018）では，形態素解析を行ったデータベースにおける基礎統計量（タイトルの形態素数と品詞別頻度，本文と件名の形態素数と品詞別頻度）について報告している。また，クラウドソーシングにおける文書の見やすさについての先行研究として，岩崎（2018），佐野（2018）がある。岩崎（2018）では，構築中のクラウドソーシング発注文書データベースにおける発注文書の依頼タイトルに焦点をあて，その中で使用されている記号の分析を行っている。その結果，タイトルの 94.6%において，何らかの記号が使用されていることが報告されている。また，記号の使用には，年代差，性差などの偏りも存在している可能性が示唆された。佐野（2018）では，1000 件の発注文書内にある見出しを，使用されている記号を手がかりに分析している。その結果，見出しにおいては，記号の使用頻度が高く，その中でも墨付カッコ（【 】）の使用が多く，見た目にインパクトを与えていることを指摘している。

３．本研究の目的

本研究では，上述した先行研究を踏まえ，「クラウドソーシング発注文書データベース（仮称）」（以下，CSDB）の構築方法の詳細を説明し，データの特徴に沿った傾向を挙げる。その上で，次の 3 点を明らかにすることを試みる。

発注文書において，

（1）使用されている記号の種類と多寡。

（2）記号の多寡と応募数の相関。

（3）空白行の挿入と応募数の相関。

これらが明らかになれば，発注文書の読みやすい（または見やすい）書き方の傾向がわかり，発注文書と応募数の相関が明らかになれば，ワーカーが興味を持つ発注文書の一端が集合知的に明らかになると考えられる。

４．クラウドソーシング発注文書データベースの構築（井上）

以下では，CSDB の元となったデータの概要を取り上げた後，DB の構築方法の詳細と仕様・本文の形態素解析結果の基本統計量を用い DB の全体像について解説する。

４.１　元となるデータについて

以下では，本研究で作成したデータベースの元となったデータの概要について解説する。データベースの元となったのは，日本最大級の総合型クラウドソーシングサイトを運営するクラウドワークス社の発注データ 1 ヶ月分（2017 年 8 月）28,896 件である。クラウドワークス社より提供された発注データは，案件ごとに id が割り振られ，基礎的な情報として件名・本文，付加的な情報として支払い形式・案件カテゴリ・期間・予算・応募数・ユーザ

ーによる評価等が含まれている。

本研究ではクラウドソーシングにおける業務の流れを大きく分ける仕事形式・支払い形式に着目する。支払い形式はテーブル上にて以下の 4 カテゴリに分けられる。

（4）hourly 型：時間単価制。EC サイトの運営など長期間のものに多い。

（5）fixed_price 型：固定報酬制。記事制作など作業量の決まったものに多い。

（6）competition 型：応募採用者のみに支払われる形式。ロゴマークの募集など。

（7）task 型：固定報酬制。アンケート・レビューなど短時間で終わるものに多い。

また，これらの 4 カテゴリは，サイト上では（4）（5）がプロジェクト形式，（6）がコンペ形式，（7）がタスク形式という 3 つの仕事形式でまとめられている。それぞれの形式の発注件数は以下の表 1 のとおりであり，fixed_price 型と task 型が全体の 96.2%を占めている。

表 1　CSDB における仕事形式（支払い形式）別の発注件数とその割合

仕事形式	支払い形式	発注件数	割合
プロジェクト形式	hourly 型（時間単価制）	787	2.72%
	fixed_price 型（固定報酬制）	13897	48.09%
コンペ形式	competition 型	322	1.11%
タスク形式	task 型	13890	48.07%
合計	-	28896	-

４.２　データベースの構築方法

CSDB では提供された発注データと紐付ける形で，発注文の件名・本文に対し形態素解析[1]を行った。形態素解析エンジンには MeCab（ver 0.996）を利用し，解析用辞書には unidic-cwj-2.2.0 を用いた。解析結果の整形には Python 3.6.4 を利用し，csv 形式にて出力した。

形態素解析の結果から，書字形・語彙素・語彙素読み・品詞（細分類まで）・文字種を抽出した後，書字形の文字数・一致用文字列を追加した。一致用文字列としては，語彙素・語彙素読み・品詞（細分類まで）を結合したものを採用した。これは既存の形態素解析による形態素 id のみでは，辛い（からい／つらい）最中（さいちゅう／もなか）といった同一の語彙素が語彙素読みによって複数の意味を持つ場合に対応できないことによる。

解析結果に対し，件名内の各形態素には形態素 id・案件 id を，本文内の各形態素には行 id[2]・形態素 id・案件 id を付与することで，CSDB 内にて各形態素を一意に呼び出すことが可能となっている。また，共通の案件 id を持つことにより解析結果を閲覧数などの提供データ内の付加情報と紐付けた分析も可能である。本文内の改行情報は，本文全体の末尾にあるもののみを削除し，空白行の情報を取り出しやすいよう配慮した。

なお，クラウドワークスの発注文作成画面では，本文内の文字列に対し文字サイズ・文字色・ボールドによる装飾を指定でき，これらは html タグによって記載される。提供された

[1] 正確には形態素ではなく短単位による解析であるが，通例に従い以降の文章では短単位での解析を形態素解析として扱う。

[2] 文単位ではなく行単位に id を付けたことは，対象となる本文に箇条書きなど末尾がないものが多数出現することによる。当 DB における行単位は，改行コードによって区切られたものとし，表示環境に大きく依存する自動的な文字送りに関しては対象としない。

データ内にて本文に html タグを含む発注文は 18,433 件，含まない発注文は 10,463 件であった。装飾用の html タグは形態素解析の際に誤解析の元となるため，前処理の段階で削除した。html による文字装飾を CSDB に組み込む点に関しては，今後の課題とした。また，html タグ付きの本文に関しては，ブラウザ表示時には適用されない改行タグ以外の改行情報を削除し，改行タグのみを改行情報として認定した。前処理段階にて改行は LF（¥n）に，文字コードは UTF-8 に統一した。

４.３　データベース上における基本統計量

以下では，CSDB 内での発注文本文の支払い形式別の形態素数について概観する。まず，発注文書における形態素解析による品詞大分類ごとの総数を形式別に示したものを表 2 に示す。また，表 3 は発注文書 1 件ごとの形態素数の平均である。

表 2　CSDB における支払い形式別の形態素数（品詞大分類）

品詞大分類	hourly 型	fixed_price 型	competition 型	task 型	全体
改行	19311	815439	20472	478717	1333939
感動詞	237	8929	145	3327	12638
記号	4039	110774	11661	83346	209820
空白	1705	88203	3200	51644	144752
形状詞	1962	87455	2460	50304	142181
形容詞	995	43315	1497	30289	76096
助詞	31374	1459333	39529	1002781	2533017
助動詞	10005	533388	13054	370627	927074
接続詞	212	12257	395	9570	22434
接頭辞	3142	216411	3845	110514	333912
接尾辞	4168	154069	5171	96417	259825
代名詞	692	28970	648	16731	47041
動詞	17612	830531	21472	578518	1448133
副詞	820	49007	1513	33564	84904
補助記号	29636	1160583	37991	868297	2096507
名詞	67027	2570670	79549	1841809	4559055
連体詞	609	25805	670	19638	46722
合計	193546	8195139	243272	5646093	14278050

表 3　発注文書 1 件あたりの平均形態素数（支払い形式別）

品詞大分類	hourly 型	fixed_price 型	competition 型	task 型	全体
改行	24.54	58.68	63.58	34.46	46.16
感動詞	0.3	0.64	0.45	0.24	0.44
記号	5.13	7.97	36.21	6	7.26
空白	2.17	6.35	9.94	3.72	5.01
形状詞	2.49	6.29	7.64	3.62	4.92
形容詞	1.26	3.12	4.65	2.18	2.63
助詞	39.87	105.01	122.76	72.19	87.66
助動詞	12.71	38.38	40.54	26.68	32.08
接続詞	0.27	0.88	1.23	0.69	0.78
接頭辞	3.99	15.57	11.94	7.96	11.56
接尾辞	5.3	11.09	16.06	6.94	8.99
代名詞	0.88	2.08	2.01	1.2	1.63
動詞	22.38	59.76	66.68	41.65	50.12
副詞	1.04	3.53	4.7	2.42	2.94
補助記号	37.66	83.51	117.98	62.51	72.55
名詞	85.17	184.98	247.05	132.6	157.77
連体詞	0.77	1.86	2.08	1.41	1.62
合計	245.93	589.71	755.5	406.49	494.12

次に，発注文書 1 件あたりの改行を含む形態素数の基本統計量を表 4 に示す。4.1 で確認したように，作業内容に対する指示を比較的多く必要としない task 型・competition 型は，形態素数が少ない値で分布している。また，発注後のやりとりをすることが可能なため，業務内容を発注文の中でどこまで説明するかが発注者次第となる hourly 型・fixed_price 型では，形態素数のばらつきが大きいことが見て取れる。

表 4　発注文書 1 件あたりの形態素数（支払い形式別）

	hourly 型	fixed_price 型	competition 型	task 型	全体
Max.	1589	2297	1758	2220	2297
3rd Qu.	913	807	374	571	724
Mean	601.1	589.7	309.1	406.5	494.1
Median	501	562	276	353	438
1st Qu.	296.5	319	144	191	248
Min.	4	1	2	3	1
SD	381.1	333.1	242.3	280.4	321.7
Var	145233.9	110980.6	58721.1	78598.4	103508.5

５．クラウドソーシング発注文書データベースにおける記号と空白行（岩崎）

以下では，CSDB において使用されている記号の種類と多寡をそれぞれ取り上げ，その傾向を明らかにする。また，空白行の挿入が 1 文書あたりどのくらいの割合で挿入されているかといった傾向についても明らかにする。

５.１　分析対象

クラウドワークスにおいて，仕事の依頼形式は複数存在している。今回は，「プロジェクト形式（hourly 型），（fixed_price 型）」と「コンペ形式（competition 型）」，「タスク形式（task 型）全ての仕事形式を対象とし，支払い形式ごとにどのように記号と改行，空白行が挿入されているかを分析していく。

５.２　発注文書における記号の種類と多寡

以下の表 5 に各支払い形式において使用されていた記号とその使用頻度（上位 10 語）を挙げる。それぞれの型に出現した全記号とその頻度については，本論文の最後に示す。まず，使用されている記号の種類を見てみると，hourly 型においては 70 種類，fixed_price 型においては 155 種類，competition 型においては 101 種類，task 型においては 134 種類の記号の使用が確認された。同データベースを使用した発注文書のタイトルにおいては，121 種類の記号が使用されており（岩崎 2018），発注文書においてはタイトルよりも記号の種類が多いことがわかった。なお，それぞれの形式に使用されていた記号の総頻度は，hourly 型が 29,636，fixed_price 型が 1,160,554，competition 型が 37,991，task 型は 868,297 であった。

表 5　発注文書で使用されていた記号の種類と頻度（上位 10 語）

	hourly 型			fixed_price 型			competition 型			task 型		
	語彙素	頻度	割合	語彙素	頻度	割合	語彙素	頻度	割合	語彙素	頻度	割合
1	、	5178	17.47%	。	209953	18.09%	。	5808	15.29%	。	164034	18.89%
2	・	4825	16.28%	、	194088	16.72%	、	5774	15.20%	、	128230	14.77%
3	。	4101	13.84%	・	148402	12.79%	・	4419	11.63%	・	103096	11.87%
4	）	1838	6.20%	）	52860	4.55%	/	3267	8.60%	※	38618	4.45%
5	（	1823	6.15%	（	51431	4.43%	.	2298	6.05%	＝	36925	4.25%
6	/	1354	4.57%	__	42895	3.70%	:	1959	5.16%	【	34057	3.92%
7	:	1266	4.27%	!	37845	3.26%	）	1889	4.97%	】	34042	3.92%
8	【	1082	3.65%	:	36201	3.12%	（	1887	4.97%	）	32696	3.77%
9	】	1082	3.65%	■	32294	2.78%	▽	1188	3.13%	（	31561	3.63%
10	～	946	3.19%	/	31673	2.73%	※	1181	3.11%	/	27378	3.15%

#網掛け箇所は，全ての形式で 10 位以内だった記号である。

（8）☆★☆こんな方は必見☆★☆

・面白い記事を読むのが好きな人

・１日２～３時間自由な時間がある人

・ネットサーフィンが好きな人

（job_offer_id: 1547354）

（9）[形式]
以下の記事のように小見出しと段落を付けてください
http://tsuufuu.com/gout-woman

[掲載サイトと参考サイト]
http://tsuufuu.com/gout-obesity-urolith
http://tsuufuu.com/gout-measures

（job_offer_id: 1725117）

次に，使用されていた記号について見ていく。全ての形式において使用されていた記号は，句読点と中点（・），カッコ，スラッシュ（/）であった。中でも，句読点と中点（・）が多く使用されていることがわかる。これらの記号のうち，句読点は文を書く際に必須の記号である。また，中点は（8）のように箇条書きをする際の頭に打たれることが多く，発注文書を書く際に見やすさとわかりやすさを工夫する上で用いられる記号であると言える。スラッシュは（9）のように成果物のイメージや参考となる記事やサイトの URL の一部であると考えられる。文書による伝達を行う必要があるクラウドソーシングにおいて，発注者は受注者に発注者が求める成果物を URL のリンク先を介してあらかじめ提示している。これらの記号は，いずれもクラウドソーシングの発注文書において大きな役割を果たしていると考えられる。

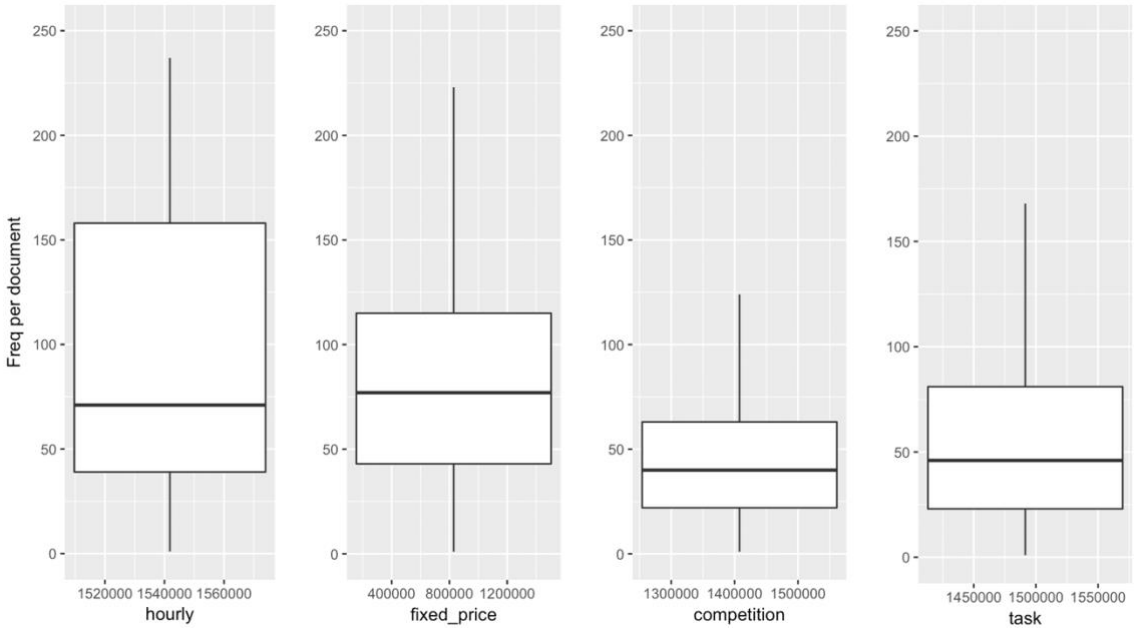


図 1　発注文書 1 件あたりの記号の数

表 6　発注文書 1 件あたりの記号の数

	hourly 型	fixed_price 型	competition 型	task 型
Max.	237	604	386	755
3rd Qu.	158	115	63	82
Mean	92.32	83.65	48.33	62.57
Median	71	77	40.5	47
1st Qu.	39	44	22	23
Min.	1	1	1	1
SD	64.7	52.52	38.74	56.33
Var	4186.19	2758.66	1500.42	3172.92

図 1 と表 6 は，各仕事形式の発注文書 1 件あたりの記号の使用頻度をまとめた boxplot と基本統計量である[3]。図 1 を見ると，hourly 型の分散が非常に大きく，同じ「プロジェクト形式」であっても，支払い方法の違いにより，記号の使用量に差があることがわかる。また，competition 型は分散が非常に小さく，平均も他の形式に比べて小さいことがわかる。「コンペ形式」は，デザイン系の発注が多く，発注文書 1 件あたりの平均形態素数も competition 型が最も少なく（competition 型:309.1, hourly 型:601.1，fixed_price 型:589.7，task 型:406.5），competition 型の発注文書は最低限の指定しか書いていないことが多いと考えられる。そのため，記号の使用も少なくなっていると考えられる。

各形式における記号の使用と応募数（クラウドワークスでは「提案人数」と表記）に相関があるかを分析した。その結果，hourly 型にのみ弱い相関が認められた（r= 0.2550893）。

５.３　発注文書の長さ（改行数）

次に，発注文書の長さについて見ていく。ここでの長さは文書において改行した回数を指す。通常，クラウドソーシングにおける発注文書は，Web ページ上で閲覧するものである。Web ページ上では一度に表示される文章量が限られているため，文章が多く改行されている場合はそれだけ多くページをスクロールしなければならない。しかし，文章が短すぎると，十分な説明を行うことが難しくなり，受注者（読み手）にとっては，わかりにくい（応募しにくい）発注文書となってしまう。

[3] boxplot の外れ値は除去して表示している（以下，全ての boxplot も同様である）。

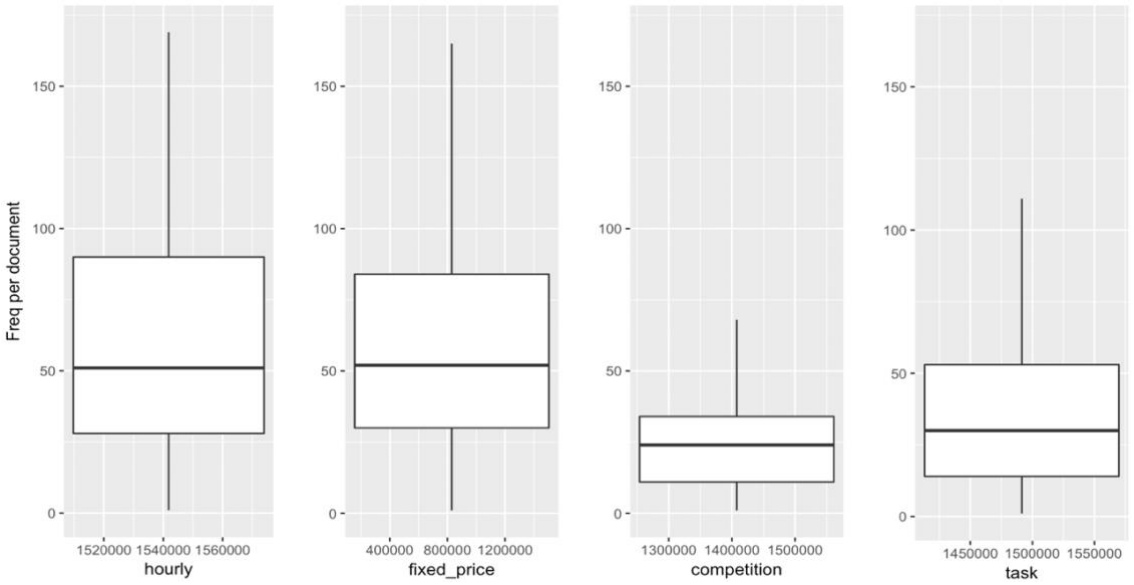


図 2　発注文書 1 件あたりの長さ（改行数）

表 7　発注文書 1 件あたりの長さ（改行数）

	hourly 型	fixed_price 型	competition 型	task 型
Max.	169	324	142	187
3rd Qu.	90	86	34	53
Mean	60.97	59.68	27	35
Median	51	53	24	30
1st Qu.	28	30	22	14
Min.	1	1	1	1
SD	41.36	37.28	20.74	26.04
Var	1710.88	1390.17	429.97	677.94

図 2 と表 7 は，各形式の発注文書 1 件ごとの長さ（改行数）をまとめた boxplot と基本統計量である。前節で見た記号とは異なり，「プロジェクト形式」である hourly 型と fixed_price 型は類似した分布であることがわかる。competition 型と task 型については，改行が少なく短い発注文書がほとんどであることがわかる。文章量という視点から分析を行う場合は，支払い形式ではなく，仕事形式の単位で分析を行うことが望ましいと考えられる。

（10）先月の実績をご報告下さい。

（job_offer_id: 1507893）

（11）エクセルファイルの件です

（job_offer_id: 1558011）

また，表 7 を見ると，最小値が 1 となっているが，これは（10）や（11）のように本文が 1 行のみの発注文書である。ただし，このような例は，前回の発注の続きであったり，特定の受注者に対しての発注であったりするなど，通常の発注とは異なる発注であると考えら

れる。各形式における改行の使用と応募数に相関があるかを分析した。その結果，hourly 型（r= 0.2227566）と task 型（r= 0.2316384）に弱い相関が認められた。

５.４　発注文書における空白行の数

最後に，発注文書内で挿入された空白行について分析を行う。空白行が挿入されている発注文書を計算したところ, hourly 型 322 件中 311 件(93.48%), fixed_price 型 13897 件中 13608 件 (97.92%), competition 型 787 件中 734 件 (93.48%), task 型 13890 件中 12984 件 (93.48%) と全ての形式で 90%以上の割合で空白行が挿入されていた。次の図 3 は，発注文書 1 件あたりの空白行の数をまとめた boxplot である。記号や改行の数と比較すると，hourly 型において今までの傾向とは異なり，分散が小さいことがわかる。一方で，同じ「プロジェクト形式」の fixed_price 型は hourly 型より分散が大きい。つまり，hourly 型では空白行の挿入に一定の傾向が存在する可能性がある。各形式における空白行の使用と応募数に相関があるかを分析した。その結果，task 型にのみ弱い相関が認められた（r=0.211738)。

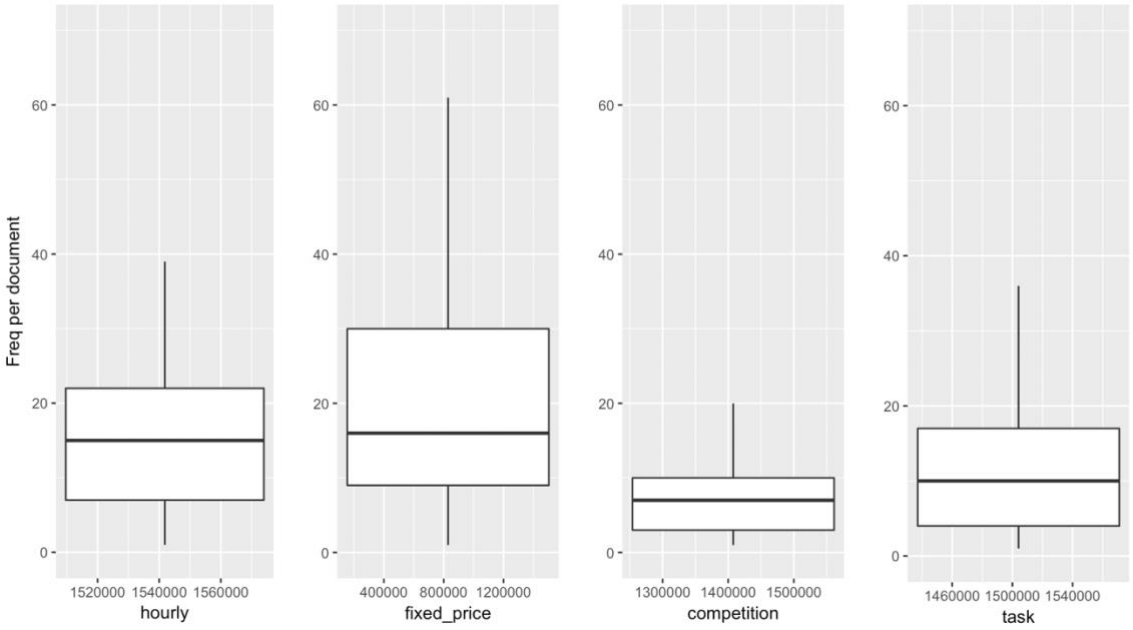


図 3　発注文書 1 件あたりの空白行の数

表 8　発注文書 1 件あたりの空白行の数

	hourly 型	fixed_price 型	competition 型	task 型
Max.	75	225	71	99
3rd Qu.	22	32.25	10	17
Mean	18.91	22.82	8.101	12.23
Median	15	16	7	10
1st Qu.	7	9	3	4
Min.	1	1	1	1
SD	16.32	20.4	6.97	10.46
Var	16.32	416.03	48.55	109.38

６．まとめと今後の課題（井上・岩崎）

以上，本研究では，インターネット上において「クラウドソーシング発注文書データベース（仮称)」の構築方法の詳細を説明し，データの特徴に沿った傾向を挙げた。その上で，発注文書において，使用されている記号の種類と多寡，発注文書の長さと空白行の挿入数を調査し，各変数と応募数の相関を見た。その結果，記号は hourly 型，改行は hourly 型と task 型，空白行は task 型との間において，それぞれ弱い相関が認められた。

今後の課題を以下に挙げる。現時点での CSDB はクラウドワークス社の 1 ヶ月分の発注データを用いて構築されているが，将来的に大規模化が予定されている。これに伴うデータの肥大化と処理時の付加の増大に対応するため，本文の現在形態素解析結果のテーブルを分割し，他のデータと連携しやすい形の中間テーブルを設計する必要がある。また，発注文や件名と応募数・評価などとの相関を分析する際の精度を高めるため，単価・期間・サイト内での案件を目立たせるために使用されるオプション機能の有無などを加味して標準化したコアデータの作成についても検討する必要がある。

今回は発注文書における記号の種類と多寡のみを明らかにしたが，これらの記号が発注文書内でどのような機能を持っているかを分析する必要がある。また，空白行については，発注文書 1 件あたりの数のみを明らかにしたが，実際に発注文書を読んでみると，連続して空白行が挿入される場合もある。そのため，どのぐらいの空白行が連続して挿入されるのかといった点も考慮する必要がある。また，空白行の挿入は，書き手の段落意識と関係があると思われる。この点についても分析して行く必要がある。

謝　辞

本研究は，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」（プロジェクトリーダー: 石黒圭）の研究成果の一部であり，JSPS 科研費 17K18504 挑戦的研究（萌芽）助成を受けたものである。

文　献

有光隆・八木秀次・呉志強・李在勲（2014）「読みやすいテキストと日本語の文章構造に関する一考察」，『工学教育』, 62:2, pp. 2_51-2_56.

芦川将之・川村隆浩・大須賀昭彦（2017）「クラウドソーシングワーカーの段階的育成方法の提案」，『人工知能学会論文誌』, 32:3, pp. B-G81_1-13.

井上雄太（2018）「発注文書の分析に有効なデータベースの構築」『第 20 回日本テレワーク学会研究発表大会』発表資料，千葉商科大学, 2018 年 7 月 8 日.

岩崎拓也（2018）「発注文書で目を惹く記号・顔文字の使い方」『第 20 回日本テレワーク学会研究発表大会予稿集』, pp.169-172.

佐野彩子（2018）「内容が一目でわかる発注文書の見出し」『第 20 回日本テレワーク学会研究発表大会』発表資料，千葉商科大学, 2018 年 7 月 8 日.

清水伸幸・中川雅史（2015）「クラウドソーシングの現状と可能性 : 2. マイクロタスク型クラウドソーシングの現状と課題 -実際の運用の知見から-」，『情報処理』, 56:9 pp.886-890.

資料 1　hourly 型の発注文書で使用されていた記号の種類と頻度

	語彙素	頻度	割合（%）
1	、	5178	17.47%
2	・	4825	16.28%
3	。	4101	13.84%
4	）	1838	6.20%
5	（	1823	6.15%
6	/	1354	4.57%
7	:	1266	4.27%
8	【	1082	3.65%
9	】	1082	3.65%
10	～	946	3.19%
11	●	882	2.98%
12	.	617	2.08%
13	※	544	1.84%
14	!	477	1.61%
15	■	375	1.27%
16	▽	330	1.11%
17	「	321	1.08%
18	」	321	1.08%
19	＝	267	0.90%
20	▼	264	0.89%
21	★	233	0.79%
22	__	143	0.48%
23	『	127	0.43%
24	』	127	0.43%
25	∝	114	0.38%
26	ー	90	0.30%
27	?	84	0.28%
28	◆	76	0.26%
29	&	73	0.25%
30	＞	73	0.25%
31	＜	72	0.24%
32	×	71	0.24%
33	~	45	0.15%
34	○	41	0.14%
35	♪	32	0.11%
36	［	31	0.10%
37	］	31	0.10%
38	…	29	0.10%
39	っ	26	0.09%
40	↓	25	0.08%
41	＋	23	0.08%
42	☆	19	0.06%
43	→	15	0.05%
44	◇	15	0.05%
45	,	14	0.05%
46	＊	13	0.04%
47	^	12	0.04%
48	／	11	0.04%
49	＃	11	0.04%
50	%	8	0.03%
51	《	6	0.02%
52	》	6	0.02%
53	◎	6	0.02%
54	－	5	0.02%
55	@	5	0.02%
56	-	4	0.01%
57	⇒	4	0.01%
58	÷	4	0.01%
59	unk	4	0.01%
60	^ ^	2	0.01%
61	□	2	0.01%
62	⑥	2	0.01%
63	⑦	2	0.01%
64	‥	1	0.003%
65	°	1	0.003%
66	<U+2661>	1	0.003%
67	<U+2666>	1	0.003%
68	◯	1	0.003%
69	〇	1	0.003%
70	㇒	1	0.003%

資料 2　fixed_price 型の発注文書で使用されていた記号の種類と頻度

	語彙素	頻度	割合（%）
1	。	209953	18.09%
2	、	194088	16.72%
3	・	148402	12.79%
4	）	52860	4.55%
5	（	51431	4.43%
6	__	42895	3.70%
7	!	37845	3.26%
8	:	36201	3.12%
9	■	32294	2.78%
10	/	31673	2.73%
11	※	30002	2.59%
12	.	25207	2.17%
13	【	24474	2.11%
14	】	24469	2.11%
15	＝	22090	1.90%
16	「	20633	1.78%
17	」	20587	1.77%
18	～	15631	1.35%
19	♪	11141	0.96%
20	☆	8780	0.76%
21	ー	7082	0.61%
22	▽	7010	0.60%
23	＊	6984	0.60%
24	★	6561	0.57%
25	●	6544	0.56%
26	?	6415	0.55%
27	○	6155	0.53%
28	▼	6148	0.53%
29	↓	4545	0.39%
30	◆	3943	0.34%
31	―	3019	0.26%
32	^	2978	0.26%

33	%	2942	0.25%
34	→	2925	0.25%
35	×	2638	0.23%
36	『	2481	0.21%
37	』	2474	0.21%
38	＞	2254	0.19%
39	~	2228	0.19%
40	＜	2101	0.18%
41	］	2082	0.18%
42	［	2081	0.18%
43	＋	2058	0.18%
44	&	1987	0.17%
45	…	1846	0.16%
46	／	1844	0.16%
47	＾＾	1525	0.13%
48	―	1279	0.11%
49	⇒	1266	0.11%
50	○	1263	0.11%
51	@	1137	0.10%
52	◎	1037	0.09%
53	◯	848	0.07%
54	<U+2661>	781	0.07%
55	⑦	780	0.07%
56	□	731	0.06%
57	＼	695	0.06%
58	②	692	0.06%
59	③	688	0.06%
60	,	534	0.05%
61	unk	527	0.05%
62	¥	488	0.04%
63	《	480	0.04%
64	》	478	0.04%
65	‥	460	0.04%
66	―	407	0.04%
67	④	328	0.03%
68	”	325	0.03%
69	<U+2013>	294	0.03%
70	°	264	0.02%
71	↑	253	0.02%
72	；	251	0.02%
73	⑥	239	0.02%
74	-	236	0.02%
75	―	232	0.02%
76	＃	225	0.02%
77	◇	200	0.02%
78	<U+266B>	187	0.02%
79	｜	139	0.01%
80	△	107	0.01%
81	≫	93	0.01%
82	≪	85	0.01%
83	“	82	0.01%
84	<U+203C>	78	0.01%
85	っ	70	0.01%
86	←	59	0.01%
87	（＾＾）	55	0.005%
88	<U+2669>	49	0.004%
89	⑧	46	0.004%
90	<U+2022>	40	0.003%

91	⑨	28	0.002%
92	〔	27	0.002%
93	〕	27	0.002%
94	´	27	0.002%
95	｀	25	0.002%
96	<U+266C>	25	0.002%
97	<U+25B6>	24	0.002%
98	〈	22	0.002%
99	〉	22	0.002%
100	⑩	22	0.002%
101	▲	20	0.002%
102	☆彡	20	0.002%
103	∇	16	0.001%
104	<U+21E8>	16	0.001%
105	ㇴ	16	0.001%
106	｛	14	0.001%
107	｝	14	0.001%
108	<U+25B7>	13	0.001%
109	$	13	0.001%
110	（＾＿＾）	12	0.001%
111	÷	11	0.001%
112	<U+21E9>	11	0.001%
113	∀	10	0.001%
114	⑤	10	0.001%
115	⑪	8	0.001%
116	ぃ	8	0.001%
117	〃	6	0.001%
118	<U+25C9>	6	0.001%
119	<U+2666>	6	0.001%
120	±	5	0.0004%
121	<U+2049>	5	0.0004%
122	<U+25E6>	5	0.0004%
123	⑫	5	0.0004%
124	㎡	5	0.0004%
125	＾＾；	4	0.0003%
126	<U+2665>	4	0.0003%
127	├	4	0.0003%
128	〒	4	0.0003%
129	ヽ	4	0.0003%
130	⑬	4	0.0003%
131	-	3	0.0003%
132	（＾＾	3	0.0003%
133	≦	3	0.0003%
134	≧	3	0.0003%
135	’	2	0.0002%
136	（株）	2	0.0002%
137	°	2	0.0002%
138	⇔	2	0.0002%
139	<U+21C4>	2	0.0002%
140	<U+24F5>	2	0.0002%
141	①	2	0.0002%
142	ぇ	2	0.0002%
143	〟	1	0.0001%
144	〟	1	0.0001%
145	<U+3016>	1	0.0001%

146	<U+3017>	1	0.0001%
147	<U+3018>	1	0.0001%
148	<U+3019>	1	0.0001%
149	<U+FF5F>	1	0.0001%
150	<U+FF60>	1	0.0001%

151	＞＜	1	0.0001%
152	√	1	0.0001%
153	┌	1	0.0001%
154	♯	1	0.0001%
155	m	1	0.0001%

資料 3　competition 型の発注文書で使用されていた記号の種類と頻度

	語彙素	頻度	割合（%）
1	。	5808	15.29%
2	、	5774	15.20%
3	・	4419	11.63%
4	/	3267	8.60%
5	.	2298	6.05%
6	:	1959	5.16%
7	)	1889	4.97%
8	（	1887	4.97%
9	▽	1188	3.13%
10	※	1181	3.11%
11	■	1172	3.08%
12	「	813	2.14%
13	」	812	2.14%
14	【	626	1.65%
15	】	618	1.63%
16	＝	545	1.43%
17	%	543	1.43%
18	！	434	1.14%
19	～	330	0.87%
20	―	240	0.63%
21	×	235	0.62%
22	&	172	0.45%
23	？	158	0.42%
24	●	103	0.27%
25	』	101	0.27%
26	『	100	0.26%
27	＋	94	0.25%
28	◆	91	0.24%
29	○	84	0.22%
30	→	81	0.21%
31	▼	78	0.21%
32	ー	77	0.20%
33	#	57	0.15%
34	＜	54	0.14%
35	＞	54	0.14%
36	★	52	0.14%
37	”	50	0.13%
38	＊	48	0.13%
39	◎	42	0.11%
40	／	37	0.10%
41	］	35	0.09%
42	［	32	0.08%
43	…	29	0.08%
44	⇒	24	0.06%
45	㎝	23	0.06%

46	~	22	0.06%
47	〇	22	0.06%
48	unk	20	0.05%
49	□	19	0.05%
50	¥	16	0.04%
51	◇	15	0.04%
52	《	11	0.03%
53	》	11	0.03%
54	↓	10	0.03%
55	“	9	0.02%
56	←	9	0.02%
57	▲	7	0.02%
58	☆	7	0.02%
59	㎜	7	0.02%
60	@	6	0.02%
61	＾＾	6	0.02%
62	,	5	0.01%
63	〃	5	0.01%
64	-	4	0.01%
65	;	4	0.01%
66	＾	4	0.01%
67	♪	4	0.01%
68	≪	3	0.01%
69	≫	3	0.01%
70	◯	3	0.01%
71	〒	3	0.01%
72	㎡	3	0.01%
73	っ	3	0.01%
74	㈱	2	0.01%
75	｛	2	0.01%
76	｝	2	0.01%
77	↑	2	0.01%
78	<U+25B6>	2	0.01%
79	<U+2777>	2	0.01%
80	④	2	0.01%
81	⑥	2	0.01%
82	―	1	0.003%
83	（株）	1	0.003%
84	〈	1	0.003%
85	〉	1	0.003%
86	〔	1	0.003%
87	〕	1	0.003%
88	＼	1	0.003%
89	｀	1	0.003%
90	<U+203C>	1	0.003%
91	<U+21E8>	1	0.003%
92	<U+25B7>	1	0.003%

93	<U+25C9>	1	0.003%
94	<U+2661>	1	0.003%
95	<U+2776>	1	0.003%
96	≒	1	0.003%
97	△	1	0.003%
98	☆彡	1	0.003%
99	⑦	1	0.003%
100	⑧	1	0.003%
101	ｉ	1	0.003%

資料 4　task 型の発注文書で使用されていた記号の種類と頻度

	語彙素	頻度	割合（%）
1	。	164034	18.89%
2	、	128230	14.77%
3	・	103096	11.87%
4	※	38618	4.45%
5	＝	36925	4.25%
6	【	34057	3.92%
7	】	34042	3.92%
8	）	32696	3.77%
9	（	31561	3.63%
10	/	27378	3.15%
11	.	24875	2.86%
12	」	22155	2.55%
13	「	21937	2.53%
14	：	21826	2.51%
15	■	17737	2.04%
16	～	16433	1.89%
17	！	13135	1.51%
18	ー	12611	1.45%
19	？	6501	0.75%
20	＊	4717	0.54%
21	￣	4586	0.53%
22	＿	4583	0.53%
23	○	4192	0.48%
24	％	3768	0.43%
25	▽	3762	0.43%
26	…	3393	0.39%
27	★	3245	0.37%
28	＞	3124	0.36%
29	＜	3088	0.36%
30	●	2957	0.34%
31	◆	2877	0.33%
32	＆	2669	0.31%
33	◯	2496	0.29%
34	『	2186	0.25%
35	』	2183	0.25%
36	△	2044	0.24%
37	―	1929	0.22%
38	▼	1889	0.22%
39	☆	1698	0.20%
40	→	1509	0.17%
41	○	1471	0.17%
42	♪	1455	0.17%
43	＾	1358	0.16%
44	,	1194	0.14%
45	⇒	1128	0.13%
46	×	1106	0.13%
47	↓	994	0.11%
48	＋	841	0.10%
49	｜	710	0.08%
50	］	652	0.08%
51	［	635	0.07%
52	”	621	0.07%
53	◎	495	0.06%
54	~	478	0.06%
55	<U+2022>	403	0.05%
56	◇	394	0.05%
57	□	257	0.03%
58	≪	249	0.03%
59	≫	249	0.03%
60	├	233	0.03%
61	←	222	0.03%
62	／	210	0.02%
63	↑	205	0.02%
64	▲	194	0.02%
65	（＾＾）	192	0.02%
66	⑥	168	0.02%
67	＾＾	142	0.02%
68	＃	117	0.01%
69	＠	108	0.01%
70	④	108	0.01%
71	“	90	0.01%
72	unk	88	0.01%
73	⑦	58	0.01%
74	っ	53	0.01%
75	｀	46	0.01%
76	´	46	0.01%
77	≪	42	0.01%
78	≫	42	0.01%
79	⑧	42	0.01%
80	ゞ	35	0.004%
81	①	31	0.004%
82	<U+2661>	28	0.003%
83	ぇ	25	0.003%
84	〈	24	0.003%
85	〉	24	0.003%
86	⑨	23	0.003%
87	<U+2666>	20	0.002%
88	―	18	0.002%
89	；	15	0.002%
90	＼	14	0.002%
91	⑩	14	0.002%
92	｛	12	0.001%
93	｝	12	0.001%

94	∞	9	0.001%
95	②	9	0.001%
96	<U+2665>	8	0.001%
97	<U+3016>	8	0.001%
98	<U+3017>	8	0.001%
99	⑤	8	0.001%
100	∀	7	0.001%
101	⑪	7	0.001%
102	―	6	0.001%
103	-	6	0.001%
104	〔	6	0.001%
105	〕	6	0.001%
106	<U+2013>	6	0.001%
107	<U+21E8>	6	0.001%
108	③	6	0.001%
109	<U+2669>	5	0.001%
110	々	4	0.001%
111	ぃ	4	0.001%
112	う	4	0.001%
113	（・∀・）	3	0.000%
114	<U+25B6>	3	0.0003%
115	<U+2776>	3	0.0003%

116	<U+2777>	3	0.0003%
117	<U+2778>	3	0.0003%
118	<U+2779>	3	0.0003%
119	<U+277A>	3	0.0003%
120	‘	2	0.0002%
121	＾＾；	2	0.0002%
122	<U+216A>	2	0.0002%
123	＞＜	2	0.0002%
124	ㇴ	2	0.0002%
125	‥	1	0.0001%
126	’	1	0.0001%
127	＾▽＾	1	0.0001%
128	°	1	0.0001%
129	±	1	0.0001%
130	<U+203C>	1	0.0001%
131	＞＜）	1	0.0001%
132	〒	1	0.0001%
133	ゞ	1	0.0001%
134	⑫	1	0.0001%

話し言葉における代名詞「あれ」の用法の分布

山崎 誠（国立国語研究所研究系言語変化研究領域）†

Usage Distribution of Pronoun *Are* in Spoken Japanese

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

代名詞「あれ」には，主に現場文脈指示，言語文脈指示，記憶文脈指示の 3 つの用法がある（日本語記述文法研究会 2009）。本発表では，話し言葉におけるこれらの用法の分布を調査し，その特徴を報告するものである。とくに，記憶文脈指示における指示対象が文脈中に述べられなかったり，指示対象が前文脈でなく後文脈に出てきたりするような用法について着目する。使用したコーパスは，『日本語話し言葉コーパス・学会講演』『同・模擬講演』『名大会話コーパス』である。これら 3 つのコーパスからそれぞれランダムに抜き出した 200 例の代名詞「あれ」の観察では，およそ，どのコーパスにおいても言語文脈指示と記憶文脈指示の 2 用法で大半を占めるということが分かった。記憶文脈指示に特徴的な構文では「あれなんだけど，〜」という従属節を構成するもの多かった。また，『名大会話コーパス』や『CSJ 模擬講演』で顕著なのは，「#あのーあれですか、カンツオーネで何が好きなんですか、一番好きな歌。」（NUC，data028，152830）のように実際に言いたいことの前置き的に使う場合が 10%前後見られたことである。

１．はじめに

本稿は話し言葉における代名詞「あれ」の用法についてその分布をコーパスを用いて明らかにすることを目的とする。なぜ話し言葉を対象にするのかというと，代名詞「あれ」は話し言葉に多く現れるからである。表 1 に示したように，書き言葉のコーパスよりも話し言葉のコーパスのほうが概ね相対頻度[1]が高い。また，表 2 に示したように『現代日本語書き言葉均衡コーパス』（BCCWJ）の各レジスター別に相対頻度を見ると，法律，白書，広報誌などの硬い，あるいは，改まった書き言葉では代名詞「あれ」の相対頻度が低く，ブログ，韻

表 1 各コーパスにおける代名詞「あれ」の頻度

コーパス	粗頻度	相対頻度（PMW）
現代日本語書き言葉均衡コーパス	12,413	118.32
日本語話し言葉コーパス・学会講演	228	69.11
日本語話し言葉コーパス・模擬講演	1,032	283.77
日本語話し言葉コーパス・対話[2]	113	750.84
名大会話コーパス	2,438	2,153.77
多言語母語話者の日本語学習者コーパス・日本語母語話者・対話[3]	200	984.74

† yamazaki [at] ninjal.ac.jp

1 相対頻度は，記号・補助記号・空白を除いた語数で算出した。

2 『中納言』における検索対象のうち，「対話・学会」「対話・模擬」「対話・課題」「対話・自由」を選択した場合。国立国語研究所（2006：3）によると，それぞれ「学会講演インタビュー」「模擬講演インタビュー」「課題指向対話」「自由対話」に当たる。

3 迫田（2016：30）によると，「できるだけ自然な日本語会話の流れを尊重し」た，半構造化インタビューである。

文，知恵袋，ベストセラーなど，軟らかい，あるいは，くだけた書き言葉では「あれ」の相対頻度が高いことが分かる。なお，ベストセラーにおける「あれ」の相対頻度が高いのは，小説の会話文が多く含まれていることによるものと思われる[4]。

表 2 『現代日本語書き言葉均衡コーパス』の各レジスターにおける代名詞「あれ」の頻度

レジスター	粗頻度	総語数	相対頻度（PMW）
特定目的・法律	0	1,079,146	0.00
特定目的・白書	4	4,882,812	0.82
特定目的・広報誌	16	3,755,161	4.26
特定目的・教科書	24	928,447	25.85
出版・新聞	44	1,370,233	32.11
出版・雑誌	376	4,444,492	84.60
出版・書籍	2799	28,552,283	98.03
特定目的・国会会議録	524	5,102,469	102.70
図書館・書籍	4473	30,377,863	147.25
特定目的・ブログ	1610	10,194,143	157.93
特定目的・韻文	36	225,273	159.81
特定目的・知恵袋	1717	10,256,877	167.40
特定目的・ベストセラー	790	3,742,261	211.10

以上見てきたように，代名詞「あれ」は話し言葉に多く現れるが，その用法がどのような分布を示すかは管見のかぎり見当たらない。本稿では，話し言葉のコーパスが充実してきたことから，これらを使用して代名詞「あれ」の用法分布を記述することを試みる。また，併せて，話し言葉でよく用いられる用法があれば，それについても指摘したい。

２．「あれ」の用法

代名詞「あれ」の用法は，『新明解国語辞典第七版』[5]によると，以下のようになっている。

①話し手・聞き手から離れて存在し、両者が共に認め得る事物自体を指す語。「ーは何だろう／ーを見てごらん／ー〔=あそこに見えるの〕が国立劇場だ／ー〔=あそこ〕に見えるは」
②すでに話題になるなどして、話し手・聞き手が共に意識している事柄を指す語。「ー〔=あの件〕はその後どうなりましたか／ー〔=あの問題〕を先に片づけてしまおう／山田君はー〔=君も知っている通りの状態〕で意外にしっかりしているんだよ」⇒ これ・それ・どれ：こそあど
③はっきり口にしたくないこと、ちょっと忘れたこと、うまく言えないことなどの代りに用いる語。「今ごろ申し上げるなんてー〔=a 恥ずかしい。b 申し訳ない〕ですが／代金はーでしたら〔=都合が悪かったら〕後でも構いません」
④〔「ー…これ…」の形で〕一つに限ることなく、いろいろの物や事柄に及ぶことを表わす。「ーが欲しいこれが欲しいと、だだをこねる／ーもしなければ、これもしなけれ

[4] 特定目的・ベストセラーの NDC 9 番台（文学）の「あれ」の粗頻度は 518，相対頻度（PMW）は 236.74 である。これは文学以外の粗頻度 272，相対頻度（PMW）175.11 の約 1.35 倍である。

[5] 三省堂 Web Dictionary による。

ばと考えるだけで頭が痛くなる」⇒ あれこれ・あれやこれや

日本語記述文法研究会（2009：16-17）の分類基準によると，①は現場文脈指示，②は言語文脈指示，③は記憶文脈指示におおよそ相当する[6]。④は慣用表現と考える。他の国語辞典における分類もほぼ同様なことから，本稿ではこれらの分類を用いて分析を行う。

３．データ・方法

本研究で使用するデータは以下のとおりである。（　）内は本稿で使う略称である。
『日本語話し言葉コーパス』（CSJ[7]）
学会講演（CSJ_APS）
模擬講演（CSJ_SPS）
『名大会話コーパス』（名大，NUC[8]）
なお，本研究で利用する言語単位はすべて短単位である。

CSJ 学会講演，CSJ 模擬講演，名大会話コーパスのそれぞれから，中納言で語彙素「彼れ」を検索し，ヒットした結果からランダムに 200 件を選び，それらに対して上述の①～④の用法を付与した。

上記の用法の認定方法であるが，①の現場指示の用法は，以下の例文（1）のように，現場指示の用法を引用の形で語ったものを含む。ただし，例文を読み上げているようなものは除外した。

（1）当時は友達が持っていたテレビアニメのキャラクターだが付いた既製品が本当に羨ましく眩しく見えてああたしも<u>あれ</u>が欲しいとか<u>あれ</u>を買ってほしいという風に何度も何度もねだった覚えがあります（CSJ 模擬講演，講演 ID：S11F1157，開始位置：12500，下線は筆者，以下同じ[9]。）

②の言語文脈指示と③記憶文脈指示との区別は曖昧になる可能性があることから，「あれ」の前文脈に，その指示対象ないしは指示対象とみなすことのできる言語表現が現れていれば，②言語文脈指示とし，そうでなければ③とした[10]。したがって，（2）（3）のように指示対象が「あれ」の直後に現れている場合は③とした。すなわち，（2）の「あれ」は，後文脈の「医学エビデンスベーストメディスン」に対応し，（3）の「あれ」は，後文脈の「膨れるん」に対応していると考える[11]。

（2）#それからまエービＥＢＭというのはその一<u>あれ</u>ですね#根拠に基づいた医学エビデン

---

[6] 厳密には『新明解国語辞典第七版』の③と日本語記述文法研究会（2009：34-36）の記憶文脈指示との対応にはずれがある。③は記憶文脈指示のうち，「思い出せない場合」に近いものと思われる。同書には，「話し手が指示対象を思い出せず，聞き手がその指示対象について知っていると思われるときには，ア系の指示表現が用いられる。」（日本語記述文法研究会 2009：35）としている。

[7] 『中納言』に収録されているデータ（データバージョン 2018.01）。

[8] 『中納言』に収録されているデータ（データバージョン 2018.02）。

[9] 以下，出典の示し方は，講演 ID（名大会話コーパスの場合は，会話 ID）と開始位置のペアで示す。

[10] 後方参照を認めれば例文（2）（3）は②言語文脈指示になる。

[11] 「あれ」の部分を後文脈の対応する語句と置き換えても違和感がない（文脈上自然なつながりになる）ことを作業基準として分類を行った。

スベーストメディスンですね（CSJ_APS，A07M0185，8630）
（3）#このドアもちょっと固め。#ああ、あの、固めって、雨が降って湿気あると下が、あのあれじゃない、膨れるんじゃない、それ。（NUC，data118，3370）

また，（4）のように，前文脈をさらにたどれば，指示対象が現れるのかもしれないが，前文脈 300 語までに指示対象が現れなければ③とした。

（4）#でもう一つですね行って感じたのは空気が奇麗だっていうことです#川崎の空気と比較しますと全然違います#まーまたさっきのあれじゃないですが#プラネタリウムで星星って言ってましたけども#実は星が奇麗に見えるんです#東京よりはですよ#川崎よりは見えるんです#（CSJ 模擬講演，講演 ID：S03M1133，開始位置：15540）

コーパスの検索結果からでは判断がつかない例は，「不明」とした。また，若干の誤解析が見られたが，これは分析対象から外した。

4．結果

表 3 に各コーパスにおける「あれ」の用法の分布を示した。() 内の割合（百分率）は，除外の例を除いた合計に対するものである。いずれのコーパスでも，言語文脈指示と記憶文脈指示が多く，両者はほぼ同じかあるいは，言語文脈指示がやや多いかという分布になっている。それ以外の用法はほとんど見られなかった。

表 3　各コーパスにおける「あれ」の用法の分布（頻度と割合）

用法	CSJ 学会講演		CSJ 模擬講演		名大会話コーパス	
現場文脈指示	5	(2.6%)	5	(2.5%)	0	(0%)
言語文脈指示	85	(44.5%)	109	(54.5%)	109	(55.6%)
記憶文脈指示	89	(46.6%)	74	(37%)	83	(42.3%)
慣用表現	4	(2.1%)	6	(3%)	1	(0.5%)
不明	8	(4.2%)	6	(3%)	3	(1.5%)
除外	9		-		4	
計	200	(100.0%)	200	(100.0%)	200	(99.9%)

4.1　記憶文脈指示に特徴的な用法

記憶文脈指示の用法にはいくつか特徴的なものが見られた。量的に多かったのは，（5）～（7）に示すような，「あれ」＋「だ/です」＋接続助詞という表現である。

（5）#でそいからもう一個の方の話はちょっと時間がないのであれですが#おんえっと一細かい話は省略しますけれど（CSJ_APS，A01M0958，93070）
（6）#まーあの一人に紹介されたりしていやこのドイツ人の彼よりもこちらのフィリピン人のコーチの方がいいだろうと思ってあのーそういうのも変えられましたね#そういうのもまず日本的なそのしがらみとかだったらいや多少あれだけど#もうちょっとだから我慢しちゃおうっていう風になってたかもしれないんですけど#そういう風にあのんドライに割り切れた私にびっくりというところもありました（CSJ_SPS，S01F1659，30210）
（7）#９時、１０時、すごい早い飲み会だよね。#無理だもん。#だって、起きてられへんもん。#まあ、あたしもね、次の日あるし、あれだから、早い方がいい、早くていいんだけど。（NUC，data005，18220）

これらの表現は，『新明解国語辞典第七版』の記述「③はっきり口にしたくないこと、ちょっと忘れたこと、うまく言えないことなどの代りに用いる語。」に当てはまる例と考えられる。これらは婉曲的用法と見なすことができる。一方，形はそれに似ているが，異なるタイプの用法も見受けられる。以下の（8）～（11）に示す。

（8）#でそれが現在はもうワントゥーワンなりますと#あいつを殺そうという風なあのーおーあれですね#湾岸戦争のピンポイントですね#あそれになってきつつある訳なんですけども（SPS_APS，A07M0896，60230）
（9）#これまでえー色々なイベントあったけど#どれも鳴かず飛ばずだったけど#そん中で一番こうかがやしか輝かしかったお思い出がこの格技大会その思い出は僕の中でこう一生輝き続けています#で後日談がありまして#んでその格技大会のえー打ち上げの時にあれですね#好きな女の子とこう懇ろになるようなチャンスもあってそれもあってこう格技大会僕の中で凄い特別な地位を占めている訳です#（CSJ_SPS，S01M0877，36600）
（10）#最近、超寒くない？#寒い。#ねー、超初歩的な会話になるね。#なんかかしこまっちゃうよね。#ほんとだよ。#なんで？#えー#これ、コートねー、あれなんだよー、お取りおきしたんさー。#あっ、そー、何色買った？（NUC，data072，3420）
（11）映画見てても。#プレスリーなんかいまだに人気ある。#あー。#移民したい第１の国でしょ、アメリカがね。#そう#うーん#昔は＊＊＊なんつってばかにしてたんだけど。#あのーあれですか、カンツオーネで何が好きなんですか、一番好きな歌。#私ですか。#えー、今まで歌った中ではー「アネマ・エ・コーレ」が一番。（NUC，data028，152830）

これらの例は，「はっきり口にしたくないこと」とは異なると考えられる。なぜかと言うと，「あれ」の直後にその内容が明示的に示されているからである。これらの「あれですね」「あれなんだよー」「あれですか」は，直後に述べることの前置き的に使われていることが分かる。その意味では，先取りした言語文脈指示と見なすこともできる。このような例は，CSJ 学会講演で 8 例（4.2%），CSJ 模擬講演で 18 例（9%），名大会話コーパスで 24 例（12.2%）見られた。この用法については，堀口（1997：104-105）に「すぐ後でくわしく言うつもりのモノやコトの代用」という指摘がある。

4.2 「あれする」

「あれ」にサ変動詞「する」を付けた「あれする」という用法がいくつか見られた（CSJ 学会講演 3 例，CSJ 模擬講演 5 例，名大会話コーパス 7 例）。これらはすべて記憶文脈指示であり，ほとんどが指示対象が文脈中に現れない例であった。以下に例を示す。

（12）#このダウンシフトのところはあのーえーそうですねまーそういうことですからちょっとここで切るしかないですね#あのーわじゃここからはまー後ということにですねあのさせていただくことにえーいたします#えーとちょっとま時間がですねあのーあれしてしまってですねあのちょっとまとまりのないお話になってしまいまして#恐縮ですけれども#えーっと時間が迫ってるっていうことですのでえーっと午後の方に回さしていただきたいと思います#（CSJ_SPS，A13F0984，174140）
（13）#父も本当にその時はショックだったんでしょうね#腑抜けになってしまったんですよ#んー何か元の最初の頃のお父さんのようにね静かな人になってしまってん幾ら私が一生懸命にあー父の世話をしてあげてもやっぱり癒されなかったんでしょうね#そんなことをあれしてるうちに平成六年に再びまた脳梗塞で倒れちゃったんです（CSJ_SPS，S05F1047，29180）
（14）#壊れないよ、こんなの。#これすごい頑丈。#うん。#これ、しっかりあれしてるから。#頑丈そう。#これ紙もほら、まだ、しわに、の紙じゃない、ほら。#うん#ただあれだからね。#これは送ったって壊れないよ。（NUC，data038，77190）

（12）は，「時間があれして」の指示対象とみなすことのできる，「時間が迫ってる」が後文脈中にあるが，（13）の「そんなことをあれしてる」（14）の「しっかりあれしてる」の指示対象は，文脈中に見当たらない。

５．まとめと今後の課題

本稿では話し言葉でよく使われる代名詞「あれ」の用法の分布を概観した。『CSJ 学会講演』『CSJ 模擬講演』『名大会話コーパス』の観察結果からは，話し言葉における代名詞「あれ」は，言語文脈指示と記憶文脈指示の 2 用法が中心であることが分かった。記憶文脈指示の中では，指示対象が文脈中にない用法（明示的な言い方を避ける婉曲的用法），および指示対象が直後の後文脈に現れ，いわば前置き的に使われている用法が特徴的な用法として挙げられる。

今後の課題としては，前文脈にどのような語句が現れるか，その特徴を明らかにすること，また，書き言葉との比較および，用法の歴史的な経緯，また，学習者の日本語[12]との比較が考えられる。例えば，話し言葉における代名詞「あれ」の相対頻度の高さは，『日本語歴史コーパス』（CHJ）の明治・大正期のデータにも認められる（表 4 参照）ことから，いつごろからこの傾向が始まったのかも今後の課題となろう。

表 4 『日本語歴史コーパス』（CHJ）明治・大正期における代名詞「あれ」の頻度

文種	粗頻度	相対頻度（PMW）
口語・会話文	697	797.35
口語・地の文ほか	434	83.13

謝　辞

本研究は，国立国語研究所のプロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー・小磯花絵）および日本学術振興会・科学研究費補助金「会話文への発話者情報の付与によるコーパスの拡張」（JP15H03212）による成果に基づいて行われている。

文　献

国立国語研究所（2006）国立国語研究所報告 124『日本語話し言葉コーパスの構築法』

迫田久美子（編）（2016）『海外連携による日本語学習者コーパスの構築―研究と構築の有機的な繋がりに基づいて―I-JAS 構築に関する最終報告書』

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

堀口純子（1997）『日本語教育と会話分析』くろしお出版

[12] 多言語母語話者の日本語学習者コーパス・学習者・対話における代名詞「あれ」の粗頻度は 279，相対頻度（PMW）は，200.16 で，日本語母語話者の場合と比べてかなり低くなっている。

現職教員による児童・生徒作文の評価基準の分析

宮城 信(富山大学 人間発達科学部)*
浅原 正幸(国立国語研究所 コーパス開発センター)†
今田 水穂(文部科学省 初等中等教育局)‡

Analysis of Evaluation Method of Teachers of Compositions Written by Elementary and Junior High School Students

Shin Miyagi (University of Toyama)
Mizuho Imada (Ministry of Education, Culture, Sports, Science and Technology)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

論者らの研究グループは、児童生徒の作文を収集して作文コーパスを構築して児童・生徒の作文能力の発達に関する研究を進めている。本研究は、それを利用した研究成果の一部である。本研究の探求課題は以下の２点である。

・1992 年と 2016 年の作文の現職教員による評価の違いの分析
・現職教員による作文評価の基準の抽出と分析

近年現職教員から「最近の子ども達の文章作成能力は低下した」という声をよく耳にする。この直感に妥当性はあるのか、またその違いはどのような点にあるのかというのが第一の課題である。また、学校現職で作文を評価する機会は多いが、その基準は必ずしも明示的ではない。現職教員の協力を得て、評価の実際を明らかにしようというのが第二の課題である。1992 年と 2016 年にできるだけ条件を揃えて作成した作文を収集して、現職教員に評価してもらい、その評価結果から評価間の関係を抽出し、統計的に分析した。その結果に基づいて、①1992 年と 2016 年の作文で評価に差がでるのか、②現職教員は作文のどのような点を評価する傾向があるのか、という点を明らかにした。

１．ここまでの研究成果と本研究の課題と意義

論者らの研究グループは、児童生徒の作文コーパスを構築して、文章能力の学年別発達の分析を進めている(宮城・今田 2015)。1992 年に作成された「手」という題の作文と条件を揃えた 2016 年に作成された作文を基にコーパスを構築し、児童らの文章作成能力の 1992 年と 2016 年の 24 年間の経年変化の語や文章量の計量的調査を行った(宮城他 2017)。その結果、24 年間の変化は小さく、いくつかの点で 2016 年の作文の方がむしろ数値的に良い状況にあることが確認された(宮城他 2017)。

しかしながら一方で、近年現場でよく耳にする「最近の子ども達の文章作成能力は低下した」という現職教員の感覚があり、先の調査結果と齟齬を生じている事態となっている。まず明らかにしなくてはならないのは、教員らはどのような観点に基づき作文を評価してい

* miyagi@edu.u-toyama.ac.jp
† masayu-a@ninjal.ac.jp
‡ Imadamizuho.ac@gmail.com

るのかという問題である。さらにその評価の違いが 1992 年と 2016 年の作文評価の違いとして抽出できるかという問題である。そこで本研究では、両年の作文に計量的な違い以外の差異があるという仮説を立てて、現職教員に収集した作文を評価してもらい、評価の基準や違いの抽出を試みることにした。現職教員による評価基準を明らかにすることは今後の作文指導改善に直結すると考えられ、現場での評価の透明性の確保にも貢献することが期待される。また、教員がおぼろげながら感じている文章作成能力の違いとは何か、にも迫ることができる。このような課題の解明に貢献できれば、本研究の成果は文章論的研究だけではなく、国語科教育においても重要な意味をもつことになる。

２．作文コーパスの概要

本研究で調査資料とした 1992 年と 2016 年の作文コーパスは同一校の児童・生徒が同一の条件で作成した、時間的・文体的な幅を有したもので、これらの資料によって児童らの作文能力を解明する縦・横断的な調査が可能になる。作文調査の概要は以下の通りである。

2016 年「手」作文の調査概要

協力校：国立大学附属小学校・中学校、全 9 学年(grade1-9)

調査時期：2016 年春期(6～7 月頃)

調査時間：小学校 40 分、中学校 45 分

作文課題：「手」(小学校低学年では「て」と板書した。)

調査方法：国語の授業で作文を作成する。専用の用紙(400 字詰め原稿用紙)を用意し、調査に使用する。調査に当たっては「これから「手」という題で作文を書きます。自由に書いてください。」とだけ指示して、作成させた。辞書などの使用は不可、個別の質問に対しても返答しない。

備考：調査時に作文を 400 字で作成するよう指示した(追加の用紙は配布しないことを伝達した)が、何人かの児童・生徒が裏面や空欄に文章を書き続けていた。これらの余剰な文章も作文コーパスに収録してある。

作文コーパスのおよその規模は以下の通りである(表 1 の grade1-9 は、小 1～中 3 に対応している)。

表 1　サンプル数

	grade	1	2	3	4	5	6	7	8	9	sum
1992	class	1	1	1	1	-	1	1	1	1	8
	sample	40	40	40	40	-	40	40	40	40	320
2016	class	3	3	3	3	4	4	4	4	4	32
	sample	101	101	100	98	106	126	115	112	120	979

※網掛け部が本調査の評価作文の抽出元

電子化は研究協力者の協力を得て実施した。児童・生徒の作文には誤字・脱字、不適切な句読点、文の不整が多数含まれている。資料の保存状況としては、1992 年の資料は調査作文本文が散逸し、ワープロ打ちされたコピーが残されているのみである。さらに小学校 5 年生(grade 5)が欠損している。

対照資料となる 2016 年の電子化作業は、当時の調査参加者に確認して 1992 年調査時にテキスト化した基準に従い、誤字・脱字に関しては間違いが明らかな場合は正しく修正した。文の不整や句読点についてはそのまま電子化を行った。データの単語数については、「手」コーパス 1992 がおよそ 5 万語、同 2016 がおよそ 23 万語程度である(宮城他 2017)。

３．現職の教員による作文評価

３.１　調査概要

本研究の調査は、現職の教員による作文の主観評価の調査によって行われ、総合評価 1 項目と観点別評価 7 項目で実施された。評価する作文は、小学校 2 年生、同 4 年生、同 6 年生から各 12 編ずつ、また中学校 2 年生から 36 編ずつ、小学校中学校供に計 36 編ずつを選ぶことにした。評価者である調査協力教員は学年に応じて現職の小学校教員と同中学、高等学校の国語科教員に依頼することにし、それぞれ都合 36 編の作文について評価を行うことにした。評価者の有効回答数は、小学校 44 名、中学校 15 名、高校 3 名で計 62 名であった。評価作業は 2015 年 10 月から 2016 年 4 月頃までに行われた。現職教員に依頼しているため、作業場所や時間は指定せず、休日などの空き時間を利用して実施している。1 編当たりの評価時間や辞書や関係書籍の参照など特に制限を設けていない。日頃の作文評価を意識して同様に実施するように依頼した。なお、評価基準のぶれを防ぐため、長期休暇中など時間が取れるときにまとめて作業するように指示した。調査の作業時間は、調査依頼者らが試験的に実施しておよそ 4，5 時間の作業であることを確認した。評価者は、地域によって偏りがあるが、北海道から九州まで広範囲の教員に依頼することができた。

３.２　評価作文のサンプリング方法

作文の抽出に際しては、第一段階で文章量の多寡に極端な偏りがないものを抽出し(ほとんどの作文が抽出された)、その後、1992 年の作文と 2016 年の作文が同数になるように、また男子が書いた作文と女子が書いたい作文が同数になるように無作為に抽出した。各作文には氏名性別などの個人情報は付記されていない。調査作文はテキスト化したものを同一のフォーマットで配布した。また念のため本文中にある個人名など個人情報調に関わる箇所は伏せ字にする処理を行った。極端に文章量が少ないと内容にかかわらず、それだけで低く評価されてしまうこと危惧したためである。また調査年度による評価の違い、性別による評価の違いの有無を確かめるためにそれぞれ同数の作文を評価資料として抽出している。なお評価作文には、作成者の学年のみが示されており、内容以外、評価に影響を与える可能性のある要素を極力排除した。最終的に評価者は学年の情報のみを頼りに当該学年の児童・生徒として評価するよう依頼した。試みに小学校 2 年生の評価作文の内訳を示せば以下の通りである。

表２　小学校 2 年生の評価作文内訳

	男子	女子	小計
1992	3	3	6
2016	3	3	6
小計	6	6	12

３.３　調査用紙と調査順

フェイスシートと評価表の項目は以下の通り。本調査の主な記入用紙(いずれも A4 版)は以下の 3 点である。調査依頼時に評価作文などと一緒に封筒に同封して配布した。

(フェイスシート)

・教師歴
・担当したことがある学年
・よく文章を書かせていますか
・作文を書くときによくする指導は何ですか
・作文を書くときに、子ども達で交流をしますか。それはどの段階ですか
・子ども達の文章を、確認したり、評価を伝えたり具体的な指導をしたりしますか
・作文指導に関連して、子ども達のどのような力が伸びて欲しいと思いますか
・作文指導で心がけていること、困っていることは何ですか

フェイスシートへの記入によって、評価者がおよそどのような教育歴の持ち主か、日常的にどのような指導を行っているかを知ることができる。今後の分析で、教育歴・指導法と評価基準の関連性についても分析の対象に含める予定である。

(評価表 1（評価項目：総合評価))

0. 総合評価

総合評価は作文を総花的に 5 件法で評価する調査である。この評価が現場の作文評価ではもっとも標準的な手法であるとともに、総合的な評価である故に評価基準が明確にされていないという問題点も内包している。

(評価表 2（評価項目：観点別評価))

1. 題材に関連ある内容を決めて書けている
2. 書こうとする内容についてよく考えて詳しく書いている
3. 学年相当の語彙を書けている
4. 学年相当の漢字が書けている
5. 文末表現に変化をつけて書けている
6. 語と語、文と文を適切な繋がりを作って書けている
7. 自分の考えが伝わるように順序を意識して書けている

観点別評価は作文を指定された 7 つの観点から 5 件法で評価する調査である。

調査に際して、評価者の現職教員に以下の①～④の手順で行うよう指示した。

①フェイスシートに記入する
②作文を読んで総合評価を行う(評価表 1)
③作文を読んで 7 つの観点別評価を行う(評価表 2)
④調査完了報告および調査協力同意書への記入

特に③の観点を先に知ると②の評価に影響がでる可能性があるので、そこは厳格に作業順を守るように複数回確認を行った。

評価の過程で作業順に混乱が生じないように調査用紙を作業順に並べて封入する、観点別評価(調査票 2)を別の封筒に封入して、表に「総合評価終了後に開封する」というような注意書きをするなどの配慮を行った。

評価に協力した現職教員には本研究の意義と調査結果の利用範囲などに関して、個人が特定できる情報は開示しない、個人の評価に関して正否や妥当性への言及はしないなど個人情報の保護について十分な配慮があることを説明し、研究の趣旨を理解した上で調査への参加の許諾を得ている。ただし調査の目的の関係で 1992 年と 2016 年の作文が混在していることに関しては調査終了後に通達した。調査後は、評価表に整理番号を附して管理している。また、別表を作成して整理番号を附したフェイスシートとの関連付けを行っている。本調査では、関係者でも個人の特定ができないようになっている。なお本研究の実施に当たり、代表者が事前に勤務校の研究倫理審査委員会の審査を経ている。

４．結果と考察

探求課題に即して、調査結果を２つの観点から分析した。一つ目は、1992 年と 2016 年の調査年の違いによって作文の評価に差があるかどうか、書き手の児童・生徒の性別による作文の評価に差があるかどうか、学年間の差があるかどうか、である。それぞれについて評価項目別に分析した(４.１節)。二つ目は、総合評価に７つの観点別評価の差が影響しているかどうかである。それぞれの項目間の関係を学年別に分析を行った(４.２節)。

４.１　調査結果(調査年差、性別)

全体像を把握するために、各調査項目の調査年の差を比較し、以下の図１を得た。

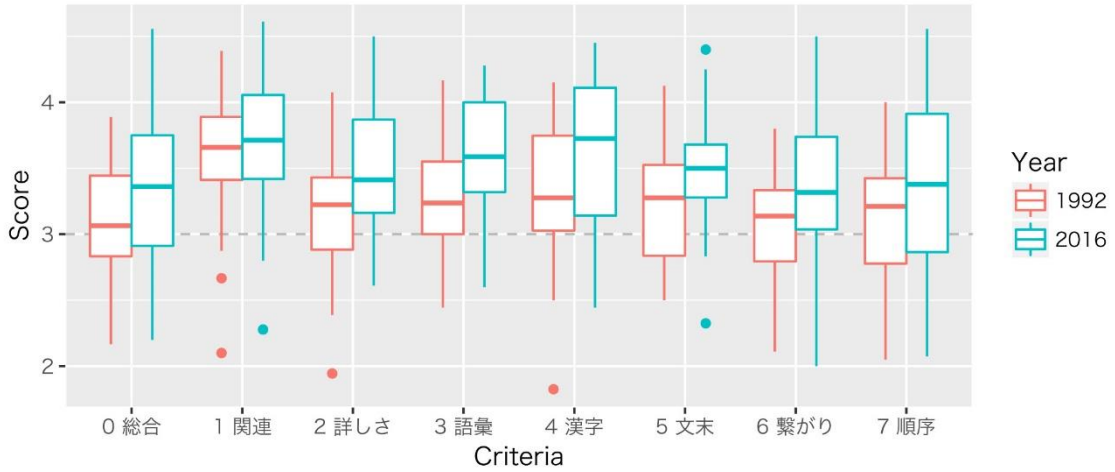


図１　調査年別の作文評価

すべての評価項目において 2016 年の作文の方が 1992 年の作文より評価が高いこと分かる。全体的な傾向は変わらないものの、この分析モデルでは評価者ごとの評価傾向の差(高めの評価を付け易い、逆に低めの評価を付け易いといった差)をそのまま反映している。また本研究では探求課題として、性別や学年差についてもあげているので、以下これらの要素を統一的に分析することを試みる。

調査結果の分析には、一般線形混合モデルによるモデル化を用い、以下調査項目別に調査結果の調査年差・性別・学年差に注目して分析した。

(分析方法)
一般線形化混合モデル：Score~ Year + Gender + Grade + (1 |Teacher_ID)
固定要因：
Year: 調査年 (1992 or 2016 因子化)
Gender: 性別 (男子 or 女子 因子化)
Grade: 学年 (2,4,6,8 因子化)
ランダム要因：Teacher_ID: 評価教員の個人差を吸収

４.１.１ [0. 総合評価]
総合評価の結果を分析し、以下の表３を得た。

表３ ［0. 総合評価］

調査年=2016	0.200 (0.041) (p<0.01)
性別=男子	-0.191 (0.041) (p<0.01)
学年	0.036 (0.041) (p<0.01)
切片	3.055 (0.099) (p<0.01)
観測数	2088
対数尤度	-2874.357

表３から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.２ [1. 題材に関連ある内容を決めて書けている] (題材)
題材の結果を分析し、以下の表４を得た。

表４ ［1. 題材］

調査年=2016	0.125 (0.040) (p<0.01)
性別=男子	-0.076 (0.040) (p<0.1)
学年	0.087 (0.014) (p<0.01)
切片	3.139 (0.110) (p<0.01)
観測数	2088
対数尤度	-2852.835

表４から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また学年間の差も有意(p<0.01)であった。性別の差は有意傾向(p<0.1)に留まり、差が見られなかった。

４.１.３ [2. 書こうとする内容についてよく考えて詳しく書いている] (内容)
内容の結果を分析し、以下の表５を得た。

表５　[2. 内容]

調査年=2016	0.300 (0.042) (p<0.01)
性別=男子	-0.160 (0.042) (p<0.01)
学年	0.061 (0.014) (p<0.01)
切片	2.944 (0.111) (p<0.01)
観測数	2088
対数尤度	-2926.983

表５から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.４　[3. 学年相当の語彙を書けている] (語彙)

語彙の結果を分析し、以下の表６を得た。

表６　[3. 語彙]

調査年=2016	0.284 (0.034) (p<0.01)
性別=男子	-0.170 (0.034) (p<0.01)
学年	0.081 (0.012) (p<0.01)
切片	2.939 (0.098) (p<0.01)
観測数	2088
対数尤度	-2539.225

表６から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.５　[4. 学年相当の漢字が書けている] (漢字)

漢字の結果を分析し、以下の表７を得た。

表７　[4. 漢字]

調査年=2016	0.214 (0.036) (p<0.01)
性別=男子	-0.375 (0.036) (p<0.01)
学年	0.137 (0.012) (p<0.01)
切片	2.830 (0.108) (p<0.01)
観測数	2088
対数尤度	-2668.497

表７から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.６　[5. 文末表現に変化をつけて書けている] (文末表現)

文末表現の結果を分析し、以下の表８を得た。

表８　[5. 文末表現]

調査年=2016	0.261 (0.035) (p<0.01)
性別=男子	-0.259 (0.035) (p<0.01)
学年	0.079 (0.012) (p<0.01)
切片	2.943 (0.097) (p<0.01)
観測数	2088
対数尤度	-2583.702

表８から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.７　[6. 語と語、文と文を適切な繋がりを作って書けている] (接続表現)

接続表現の結果を分析し、以下の表９を得た。

表９　[6. 接続表現]

調査年=2016	0.246 (0.038) (p<0.01)
性別=男子	-0.223 (0.038) (p<0.01)
学年	0.070 (0.013) (p<0.01)
切片	2.854 (0.100) (p<0.01)
観測数	2088
対数尤度	-2765.583

表９から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.８　[7. 自分の考えが伝わるように順序を意識して書けている] (順序)

順序の結果を分析し、以下の表１０を得た。

表１０　[7. 順序]

調査年=2016	0.198 (0.043) (p<0.01)
性別=男子	-0.273 (0.043) (p<0.01)
学年	0.101 (0.015) (p<0.01)
切片	2.714 (0.113) (p<0.01)
観測数	2088
対数尤度	-3018.741

表１０から、調査年の差は有意(p<0.01)で、2016 年の作文の方が評価が高かった。また性別の差も有意(p<0.01)で、女子の作文の方が評価が高かった。学年間の差も有意(p<0.01)であった。

４.１.９　まとめ

ここまでの分析をまとめて、以下の表１１のように整理した。

表１１　調査年差、性別

調査項目	調査年	性別	学年
0. 総合評価	1992<<<2016	男子<<<女子	***
1. 題材に関連ある内容を決めて書けている	1992<<<2016	男子<　女子	***
2. 書こうとする内容についてよく考えて詳しく書いている	1992<<<2016	男子<<<女子	***
3. 学年相当の語彙を書けている	1992<<<2016	男子<<<女子	***
4. 学年相当の漢字が書けている	1992<<<2016	男子<<<女子	***
5. 文末表現に変化をつけて書けている	1992<<<2016	男子<<<女子	***
6. 語と語、文と文を適切な繋がりを作って書けている	1992<<<2016	男子<<<女子	***
7. 自分の考えが伝わるように順序を意識して書けている	1992<<<2016	男子<<<女子	***

ここまでの分析から分かったことをまとめると、以下のようになる。

・すべての調査項目で、2016 年の作文の方が 1992 年の作文よりも有意に高く評価されている。
・題材に関連する内容が書けていること([1. 題材])を除くすべての調査項目で、女子の作文の方が男子の作文よりも有意に高く評価されている。
・すべての調査項目の評価で、学年間の差が有意である。

４.１.１０　検討すべき文章量と評価の関連

評価作文の文章の語数(token)を詳しく見てみると、以下の図 2 のように分布している。

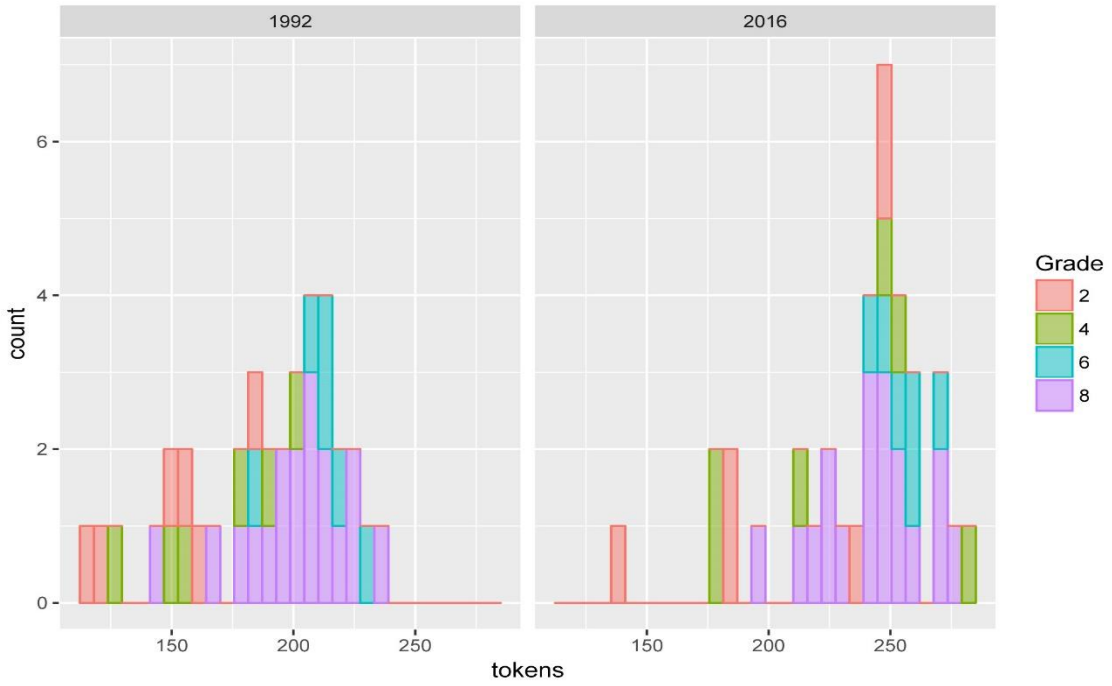


図２　評価作文の文章量

抽出時に文章量に大きな違いがないようにある程度の調整は行っているが、語数を比較すると 2016 年の作文方がばらつきが大きく、全体的に 1992 年の作文より語数が多いことが確認できる。

次に文章量と各評価項目との関係を整理して、次の図３を得た。

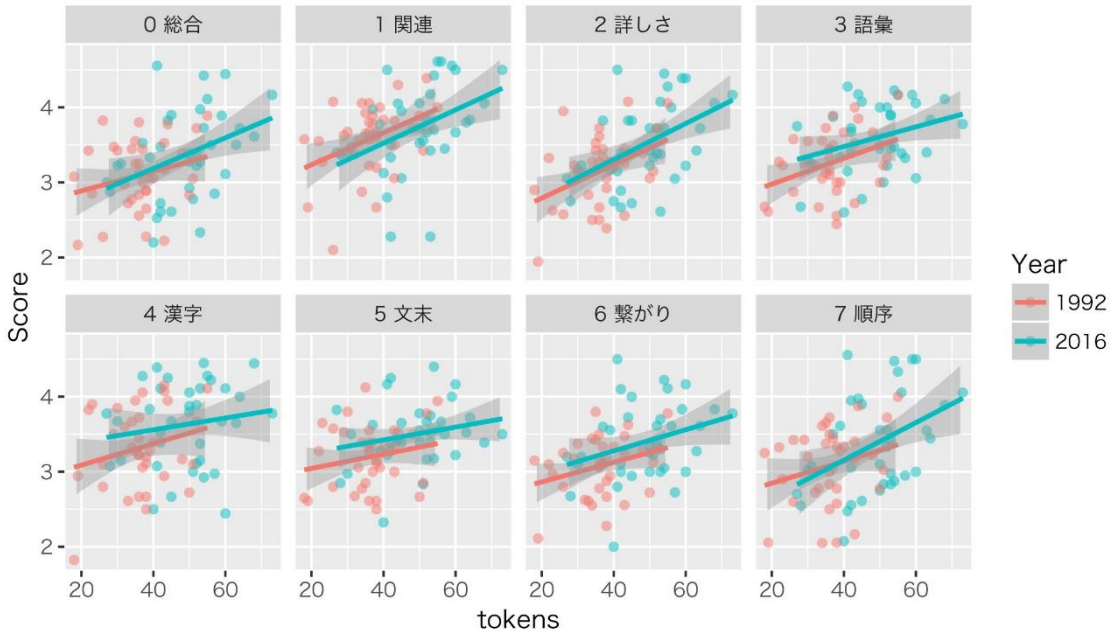


図３　文章量と評価の関連性

どの評価項目においても基本的に右肩上がりで語数が増えるほど評価が高くなる傾向がある。さらに図２で見たように、1992 年の作文より 2016 年の作文の方が語数が多い傾向にある。4.1.9 節で述べた全体的に 2016 年の方が評価が高いという傾向とこの語数の差は相関があるように見える。

一方で、語数が多いことと漢字が書けていると評価されること([4.漢字])、適切な繋がりを作って書けていること([6.繋がり])などいくつかの評価項目が直接的に語数の多寡の影響を受けるかは説明が難しいように思う。見通しとしては、語数が多いと評価が高いということではなく、評価が高い文章を書ける書き手は、語数の多い文章を書く傾向がある(長文を書くことができる)ということの反映ではないかと考えている。ここまでの分析ではこの点について明確な答えを出す用意がないので、ここでは調査年差の結果に疑問が残ることを指摘するに留め、今後語数を調整したデータの比較を行うなどして、条件を整えた分析を実施したい。

４.２　調査結果(総合評価と観点別評価)

総合評価の結果を分析し、以下の表１２を得た。

表１２　評価項目間の相関

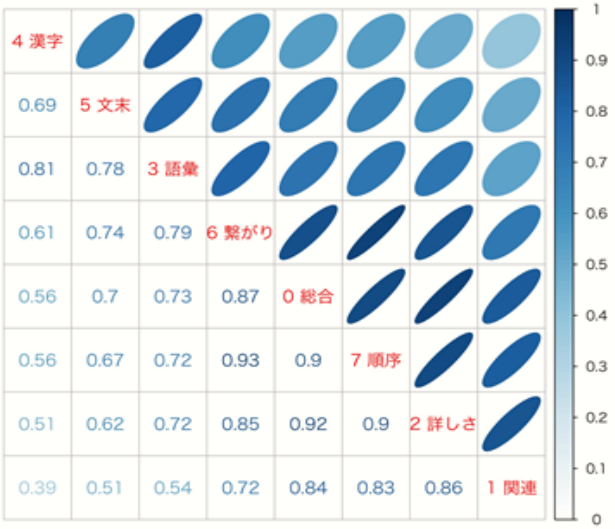


表１２から、相対評価と観点別評価の項目間の相関関係を確認すると、すべての観点別評価と相関があることが分かる。特に、内容に関する項目である[1.題材]や[2.詳しさ]、作文の構成に関する項目である[6.繋がり][7.順序]との相関が高かった。語彙に関する項目である[3.語彙]、[4.漢字]、[5.文末表現]も相関があると認められるが(ほぼ 0.7 以上)、それらに比べるとやや低い値であった。全体で見た場合、項目間の相関関係の優劣を認めることは難しい。よって、以下、学年別に総合評価と観点別評価の相関を分析する。

４.２.１　調査結果(調査年差、性別)

調査結果の分析には、一般線形混合モデルによるモデル化を用い、以下学年別に調査結果の総合評価と７つの観点別評価と関係に注目して分析を行った。

(分析方法)

一般線形化混合モデル：

Score~Score_1+Score_2+Score_3+Score_4+Score_5+Score_6+Score7+(1|Teacher_ID)

固定要因：観点別評価 7 項目

ランダム要因：Teacher_ID: 評価教員の個人差を吸収

４.２.２　小学２年生

小学２年生の評価の結果を分析し、以下の表１３を得た。

表１３　総合評価と観点別評価(小学２年生)

調査項目	
1. 題材に関連ある内容を決めて書けている	0.303 (0.039) (p<0.01)
2. 書こうとする内容についてよく考えて詳しく書いている	0.192 (0.046) (p<0.01)
3. 学年相当の語彙を書けている	0.099 (0.049) (p<0.05)
4. 学年相当の漢字が書けている	0.065 (0.038) (p<0.1)
5. 文末表現に変化をつけて書けている	0.082 (0.046) (p<0.1)

6. 語と語、文と文を適切な繋がりを作って書けている	0.020(0.049)
7. 自分の考えが伝わるように順序を意識して書けている	0.214(0.046) (p<0.01)

表１３から、小学校２年生の作文評価では、[1.題材]、[2.詳しく]、[7.順序](p<0.01)、[3.語彙](p<0.05)が有意に高かった。一方、[4.漢字]、[5.文末表現]、[6.繋がり]は有意ではなかった。

４.２.３　小学４年生

小学４年生の評価の結果を分析し、以下の表１４を得た。

表１４　総合評価と観点別評価(小学４年生)

調査項目	
1. 題材に関連ある内容を決めて書けている	0.230(0.046) (p<0.01)
2. 書こうとする内容についてよく考えて詳しく書いている	0.203(0.044) (p<0.01)
3. 学年相当の語彙を書けている	0.065(0.052)
4. 学年相当の漢字が書けている	0.023(0.044)
5. 文末表現に変化をつけて書けている	0.139(0.042) (p<0.01)
6. 語と語、文と文を適切な繋がりを作って書けている	0.143(0.049) (p<0.01)
7. 自分の考えが伝わるように順序を意識して書けている	0.139(0.047) (p<0.01)

表１４から、小学校４年生の作文評価では、[1.題材]、[2.詳しく]、[5.文末表現]、[6.繋がり]、[7.順序]が有意 (p<0.01)に高かった。一方、[3.語彙]、[4.漢字]は有意ではなかった。

４.２.４　小学６年生

小学６年生の評価の結果を分析し、以下の表１５を得た。

表１５　総合評価と観点別評価(小学６年生)

調査項目	
1. 題材に関連ある内容を決めて書けている	0.173(0.049) (p<0.01)
2. 書こうとする内容についてよく考えて詳しく書いている	0.254(0.050) (p<0.01)
3. 学年相当の語彙を書けている	0.125(0.051) (p<0.05)
4. 学年相当の漢字が書けている	0.045(0.043)
5. 文末表現に変化をつけて書けている	0.002(0.046)
6. 語と語、文と文を適切な繋がりを作って書けている	0.153(0.051) (p<0.01)
7. 自分の考えが伝わるように順序を意識して書けている	0.209(0.048) (p<0.01)

表１５から、小学校６年生の作文評価では、[1.題材]、[2.詳しく]、[6.繋がり]、[7.順序](p<0.01)、[3.語彙](p<0.05)が有意に高かった。一方、[4.漢字]、[5.文末表現]は有意ではなかった。

４.２.５　中学２年生

中学２年生の評価の結果を分析し、以下の表１６を得た。

表１６　総合評価と観点別評価(中学２年生)

調査項目	
1. 題材に関連ある内容を決めて書けている	0.152(0.030)(p<0.01)
2. 書こうとする内容についてよく考えて詳しく書いている	0.255(0.034)(p<0.01)
3. 学年相当の語彙を書けている	0.071(0.039)(p<0.1)
4. 学年相当の漢字が書けている	0.024(0.035)
5. 文末表現に変化をつけて書けている	0.146(0.033)(p<0.01)
6. 語と語、文と文を適切な繋がりを作って書けている	0.111(0.035)(p<0.01)
7. 自分の考えが伝わるように順序を意識して書けている	0.218(0.035)(p<0.01)

表１６から、中学２年生の作文評価では、[1.題材]、[2.詳しく]、 [5.文末表現]、 [6.繋がり] 、[7.順序]が有意 (p<0.01)に高かった。一方、[3.語彙]、 [4.漢字]は有意ではなかった。

４.２.６　まとめ

ここまでの分析をまとめて、以下の表１７のようになる。

表１７　総合評価と観点別評価(まとめ)

調査項目	小 2	小 4	小 6	中 2
1. 題材に関連ある内容を決めて書けている	0.303***	0.230***	0.173***	0.152***
2. 書こうとする内容についてよく考えて詳しく書いている	0.192***	0.203***	0.254***	0.255***
3. 学年相当の語彙を書けている	0.099**	0.065	0.125**	0.071*
4. 学年相当の漢字が書けている	0.065*	0.023	0.045	0.024
5. 文末表現に変化をつけて書けている	0.082*	0.139***	0.002	0.146***
6. 語と語、文と文を適切な繋がりを作って書けている	0.020	0.143***	0.153***	0.111***
7. 自分の考えが伝わるように順序を意識して書けている	0.214***	0.139***	0.209***	0.218***

ここまでの分析から分かったことをまとめると、以下のようになる。

・全学年で題材に関連する内容が書けていること([1.題材])、詳しく書けていること([2.詳しく])、順序を意識して書けていること(、[7.順序])は総合評価に有効である。
・中学年以降は適切な繋がりを作って書けていること([6.繋がり])が総合評価に有効である。
・学年別では漢字が書けていること([4.漢字])は総合評価に有効ではない。
・全体的な傾向として語彙に関わる項目([3.語彙] [4.漢字] [5.文末表現])は総合評価に影響しにくい。

５． 本研究のまとめ

ここまでの考察をまとめると以下のようになる。

・作文評価においてもほとんどの評価項目で、1992 年より 2016 年の作文の方が評価が高い。(ただし、2016 年の作文の方が語数が多く、影響を与えている可能性を排除できない。)
・作文評価においてもほぼすべての評価項目で、男子より女子の作文の方が評価が高い。
・すべての評価項目で学年毎に作文評価に差がある。
・全体で見るとすべての観点別評価が総合評価に有効である。
・学年別に見ると漢字が書けていることは総合評価に有効ではない。
・中学年以降は適切な繋がりを作って書けていることが総合評価に有効である。
・すべての学年において題材に関係ある内容、詳しく書いている、順序を意識して書いているが総合評価に有効である。

既に述べたように、本研究では「1992 年と 2016 年の作文に計量的な違い以外の差異がある(1992 年の方が評価が高い)」という仮説を立て、①1992 年と 2016 年の作文で評価に差がでるのか、②現職教員は作文のどのような点を評価する傾向があるのか、という探求課題の解明を目指した。分析した結果、①については、評価に差が見られたが仮説に反して現職教員の評価においても 2016 年の作文の方が評価が高くなった。今後これが調査協力校以外でどの程度一般的な傾向なのかを明らかにしていかなければならない。また②については、全体的にはすべての観点別評価が総合評価に影響するが、学年別に見ると関連する観点別評価と比較的関連が弱い観点別評価があることが分かった、この結果に基づいて現職教員の評価の基準を抽出し、ある程度の客観化ができたと考える。観点別評価の精度を高めるために、事後アンケートの実施や評価作文の書き込みなどを参照して具体的な評価ポイントについて分析を深めていく必要がある。(次節も参照のこと。)

６． 今後の課題

本研究で考察できなかった点について言及しておく。本研究の分析の基本となる資料は調査票の評価(5 件法)に基づく分析である。本調査では、３．３節でも示したようにフェイスシートにも回答を依頼しているので、作文評価と関連付けて、教師歴や授業スタイルとの関係を考察する必要がある。若手(教師歴５年未満とベテラン教師歴２０年以上で評価の傾向に違いがあるのではないかと考えるが、その検証も実施したい。また、書き手の個人情報保護の観点から評価作文も回収したが、評価者の現職教員らが、評価作文中に気になる語や不正と判断した表現など様々な書き込みをしていることに気付いた。この書き込みを分析すれば、評価者の評価がどのような点に着目してなされたかを抽出することができそうである。これは評価者の作業の直接的な痕跡であり、引いては作文評価モデルの構築に貢献できる真正性の高い資料となるので、今後の課題の中でも優先的に取り上げたいものの一つである。ここまでの課題は本研究の不備というよりも、本研究の成果を承けた今後の発展的な研究課題となる。

さらに、その先に見えてくる探求課題として、次のようなものについても研究を進めていくことによって、文章論的研究や国語科教育(特に作文指導)への成果が十分に期待できると考えている。

・他の分析方法との連携(例えば、脱文脈化指標との連携分析)
・総合評価に有効な観点に基づく評価モデルの構築
・総合評価に有効な内容モデルの構築(効果的な文章構成とは何か？)
・総合評価に有効な語彙、接続表現、文末表現の抽出

謝　辞

本研究は、「博報財団第 11 回児童教育実践についての研究助成《継続助成》「現職との協働による児童・生徒作文能力の経年変化に関する発展的研究」(2018-2019 年度、研究代表者：宮城信、助成番号：2016-053 継)」による成果の一部である。記して感謝申し上げます。

文　献

宮城信他(2017)「現職との協働による児童・生徒の作文能力の経年変化に関する研究」(ことばのこえII)、博報財団第 11 回児童教育実践についての研究助成　研究成果報告書

宮城信・今田水穂(2015)「『児童・生徒作文コーパス』の設計」『第 7 回コーパス日本語学ワークショップ予稿集』223-232.(https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no7_papers/JCLWorkshop_No7_web.pdf)

関連 URL

作文を支援する語彙文法的事項に関する研究プロジェクト(https://sites.google.com/site/sakubunshienpropject/)

スペイン語における前置詞句の数・定性

―7 前置詞のクラスタリング ―

喜多田 敏嵩 (東京外国語大学大学院)

Definiteness and Number in the Spanish Prepositional Phrases: Cluster Analysis of Seven Prepositions

Toshitaka Kitada (Tokyo University of Foreign Studies)

要旨

スペイン語学において、名詞句の数・定性は活発な議論が展開されてきた分野であるが、前置詞句の項として生起する名詞句に焦点を当てた研究は少なく、部分的な記述がわずかに存在するのみである。本稿は、スペイン語において最頻出の前置詞である de および、de と可換な用法を有する 6 前置詞 con, desde, en, para, por, sobre の計 7 つを分析対象として、前置詞句の数・定性に関する Fernández Ramírez (1986) および Bosque (1996) の空間指示性による分類的記述の妥当性を検証するものである。データの収集には、コーパス検索ツール Sketch Engine において公開されている 100 億語規模の均衡ウェブコーパス esTenTen [2011, Eu＋Am] を使用し、算出された 7 前置詞に後続する普通名詞の単数形生起頻度と、限定詞との共起頻度を 2 変数としたクラスタリングを実行した。その結果、これら 7 つの前置詞は {con, de, por} {desde, en, para, sobre} の 2 群に分類され、Fernández Ramírez (1986) の挙げる分類基準と、en を con, de, por と同じクラスターに位置付ける Bosque (1996) の記述に検討の余地があることが分かった。

0. はじめに

名詞句の数・定性は、スペイン語学において活発な議論が行われてきた分野であり、スペイン語と日本語の両方で多くの先行研究が存在する。しかし、その大半は、文における統語機能を果たす名詞句に焦点を当てたものであり、前置詞句内をはじめとする、統語的要請を受けにくい位置にある名詞句の数・定性を体系的に記述した研究は、筆者の管見の及ぶ限り存在しない。本稿は、de および de と意味の重なりを有する 6 つの前置詞 con, desde, en, para, por, sobre の計 7 つについて、後続する普通名詞の数・定性を変数とした階層的クラスター分析を実施し、Fernández Ramírez (1986) と Bosque (1996) による前置詞句の数・定性に関する部分的記述の不透明性を指摘するものである。

1. 準備

本稿で分析対象とする構造は、 <前置詞＋名詞句> という語列により構成される前置詞句である。ここでは、英語を参照点としながら、スペイン語の前置詞句を、その構成素に基づいて記述しておきたい。

1.1. スペイン語の名詞句

スペイン語における名詞句は、名詞や代名詞を主要部として構成されるが、名詞を主要部とした場合、付接部には限定詞や形容詞が生起し、寺崎 (1998: 54-57) によれば、一般的

に以下のような構成をとる。

表 1. スペイン語における名詞句の構成

(限定詞)	(修飾語)	名詞	(修飾語)
el ART.DEF.M.SG 'the child'		muchacho child.M.SG	
los ART.DEF.M.PL 'the classic authors'		autores author.M.PL	clásicos classic.M.PL
su[1] 3.POSS.SG 'his/her/their last trip'	último last.M.SG	viaje trip.M.SG	

1.2. 名詞の語彙素性

名詞句の主要部を果たす普通名詞については、英語と同様に、その語彙素性による分類が可能である。スペイン語における普通名詞の分類については、その素性の複合性から一元的な分類が困難であることが Bosque (1999) や RAE and ASALE (2009) により述べられているが、Bosque (1999) の詳細な分類によれば、普通名詞の語彙的特徴を記述する素性として、可算性 (非連続性)、数量化可能性、個別性、有形性の 4 つがあり、これらの素性を典型的に備えた普通名詞として、以下のような語が挙げられている。

表 2. スペイン語における普通名詞の語彙素性とその例

可算/不可算	可算的	árbol (木) , mesa (机)
	不可算的	agua (水), vino (ワイン), plata (銀),
数量化可能性	数量化可能	libros (本)
	数量化不可能	celos (嫉妬), ganas (欲求)
個別/集合	個別的	casa (家), árbol (木)
	集合的	familia (家族), arboleada (木々)
有形/無形	有形的	例示なし[2]
	無形的	verdor (活力), temor (恐怖),

本稿では、これら全てを包括した普通名詞を取り扱い、固有名詞を分析の対象外とする。

[1] スペイン語において、3 人称単数 (彼、彼女、それ) と 3 人称複数 (彼ら、彼女ら、それら) の所有形容詞の間には形態上の区別がなく、その解釈は、他の要素との結束性に委ねられる。
[2] 特別な例示がないのは、有形/無形が他の素性のように明確なものではないことによる。

1.3. 前置詞句

寺崎 (1998: 54-57; 97-98) によれば、前置詞句は、前置詞を主要部、後続する名詞句を被制語とすることで構成されており、el reloj *de oro* (the watch of gold) のように名詞句の後置修飾語として、または (1) のように形容詞句の修飾語として機能する。いずれの場合も、前置詞に後続する名詞句は主要部ではない点に注目されたい。

(1) Aquella reacción era ajena *a su carácter*.

Aquel-la	reacción	era	ajen-a	a	su	carácter.
that.DEM-F.SG	reaction	COP.3SG.IND.IPFV	far-F.SG	from	3.POSS	character.

'that reaction was far from his character'

あの態度は彼の性格とは相容れないものだった。

また、文の構成素としては、(2) 直接補語 (3) 間接補語 (4) 斜格補語[3] (5) 付加語の機能を果たす。以下の例 (2)-(5) が上記の各機能と対応している。

(2) Trataron muy mal *a sus clientes*.

Trat-aron	muy	mal	a	su-s	cliente-s.
attend-3PL.IND.PST	very	badly	to	3PL.POSS-PL	client-PL

'they attended very badly to their clients'

彼らは顧客に非常に悪い扱いをする。

(3) Enseñé el camino *a un extranjero*.

Enseñ-é	el	camino	a	un	extranjer-o.
Tell-1SG.IND.PST	ART.DEF.M.SG	way	to	ART.INDEF.M.SG	foreigner-M.SG

'I told the way to a foreigner'

私は外国人に道を教えた。

(4) No disponía *de bastante dinero para comprarlo*.

No	dispon-ía	de	bastante	dinero	para	comprar=lo.
NEG	dispose-1SG.IND.IPFV	of	enough	money	to	buy.INF= it.ACC.3SG.M

'I didn't have enough money to buy it'

私はそれを買うのに十分なお金を持っていなかった。

(5) Mi hermano y yo dormimos *en la misma habitación*.

Mi	hermano	y	yo	dorm-imos	en	la	mism-a	habitación.
1SG.POSS	brother	and	I	sleep-1PL.IND.PST	in	ART.DEF.F.SG	same-F.SG	room

'my brother and I slept in the same room'

兄と私は同じ部屋で寝た。

[3] complemento de régimen preposicional のことであり、前置詞補語という訳も確認される。

このように、前置詞句は様々な統語機能を果たすが、前置詞句内の名詞句は主要部となることがなく、項として構造を支えている。

1.4. スペイン語の統語的特徴

本稿に関連するスペイン語の主要な統語的特徴は次の 2 点に大別される。

1 点目は、修飾語句の後置である。英語では、形容詞をはじめとする修飾語句は通常名詞に前置されるのに対して、スペイン語では名詞に後置させるのが通例である。したがって、「危険な動物たち」という名詞句は、両言語で次のように表現されるが、名詞「動物たち」に対する形容詞「危険な」の位置が鏡像的なものとなる。

日本語	英語	スペイン語
危険な動物たち	dangerous animals	animales peligrosos

2 点目は、限定詞あるいは量化表現を伴わない名詞句は文頭への生起および任意の句の主要部を構成しにくいというものである。例えば、「ライオンは危険な動物である」という総称文の主語ライオンは、英語では (6) のように、無冠詞複数形で表すことが可能であるが、スペイン語では (7) にあるように、定冠詞の付与が義務的である。

(6) Lions are dangerous animals.

(7) Los leones son animales peligrosos.

Los	leon-es	son	animal-es	peligros-os.
ART.DEF.M.PL	lion-PL	COP.3PL.IND.PRES	animal-PL	dangerous-M.PL

'the lions are dangerous animals'

その他にも、両言語間には、文の語順の制約などの特筆すべき差異が認められるが、句のレベルでは、無冠詞単数形が句の主要部を務めることが稀であり、数・定性の付与が一般的に必要である点が、本稿に最も関連するスペイン語の言語的特徴である。

1.5. 名詞句と前置詞句の構成と機能

スペイン語学において、主要部を果たす名詞句の数・定性に関しては様々な見地からの分析が活発に行われているのに対して、前置詞句内をはじめとする、主要部を担わない名詞句の数・定性に関しては、部分的な記述がわずかに散見されるのみである。しかしながら、主要部ではない名詞句は、文の統語的要請を受けにくい位置を占めていることなどから、主要部を果たす名詞句と類似した傾向の数・定性を見せる確証があるとは言いがたい。したがって本稿では、前置詞句内の名詞句に関する数・定性の記述の必要性を出発点として、部分的に存在する過去の論考にもとづきながら、前置詞句の数・定性を体系的に記述するための基盤となる考察を行っていく。

2. 先行研究

2.1. Fernández Ramírez (1986)

管見の及ぶ限りにおいて、前置詞句における冠詞の生起に関する最も詳細な記述を行っているのが Fernández Ramírez (1986) である[4] 。Fernández Ramírez (1986: 166-169) は、明確な空間指示機能を有する前置詞は無冠詞単数形を後続させにくいとし、tras, detrás de, junto a, encima, desde, debajo, hasta, sobre, dentro の 9 つの前置詞 (句) を挙げている。

2.2. Bosque (1996)

スペイン語の無冠詞名詞句に関する詳細な記述を行っている Bosque (1996: 53-55) は、無冠詞名詞単数形が前置詞に後続する構造の容認度の低さに触れながら、実際に無冠詞名詞単数形が項となって形成される前置詞句の特性に関する記述を行っている。その中で、前述の Fernández Ramírez (1986) の分類基準を参照しながら、con, de, en, por の 4 つを、無冠詞名詞単数形を後続させやすい前置詞であるとしている。

2.3. 問題の所在

ここまで、本稿の分析に有用な 2 つの先行研究を挙げたが、これらの論考が、en をうまく分類しえない記述となっている点を問題として指摘したい。Fernández Ramírez (1986) は、明確な空間指示性の有無により前置詞の分類を試みていたが、この性質が確固とした指標であるとは言いがたい。また、Bosque (1996) の挙げる 4 つの前置詞 con, de, en, por において、en は英語の at, in, on に相当する前置詞であり、空間的意味を有している。したがって、空間指示性の強弱を尺度にした分類において、en が弱者に分類されるというのは、スペイン語学習者としての経験的印象に反するものであり、前置詞に後続する名詞句の数・定性を分類する尺度としての空間指示性の強弱は、その適性に疑問が残る。

本稿では、コーパス調査を通じて、Bosque (1996) の挙げる 4 前置詞が、後続名詞句の数・定性について同様の分布を見せるかを検証し、先行研究の挙げる分類尺度の不透明性を考察する。

3. 分析手法

3.1. 分析対象の限定

本稿で分析対象とするのは、普通名詞が限定詞や複数語尾を伴って前置詞に後続する語列であるが、分析にあたって各品詞の定義を決めておかねばならない。普通名詞については、前述のとおり、固有名詞ではない名詞の類とし、限定詞については、寺崎 (1998) の定義にしたがうことにする。寺崎 (1998: 75) によれば、スペイン語の限定詞には、冠詞、指示形容詞、所有形容詞、関係形容詞、疑問形容詞、数詞、不定形容詞の 7 つが該当するが、本稿では、4.1. で後述する理由から、関係形容詞と数詞を除いた 5 つを限定詞とする。

また、前章で挙げた先行研究では、意味の重なりが考慮されていない多数の前置詞が記述されている。そこで本稿では、López (1972) の記述に依拠しながら、スペイン語において最頻出の前置詞 de と、de に関連のある前置詞のみを分析対象とする。López (1972) は、スペイン語における前置詞の意味の対立・中和について図式などを用いた詳細な分析を行っており、その中で de との意味の中和を有し、可換な場合がある前置詞として、con, desde,

[4] Solís García (2011:111) 参照。

en, para, por, sobre の 6 つを挙げている。なお、de を含めたこれら 7 つの前置詞は、英語では、主に以下のような前置詞に該当する。

表 3. 7 前置詞の英語対照表

con	de	desde	en	para	por	sobre
with	of	from since	in on	for to	because of by	over about

そして 6 前置詞は、以下のような場合に de と置換可能である[5]と述べられている。

表 4. de と置換可能な例

日本語	英語	de を使う場合	可換な例
雪に覆われた	covered with snow	cubierto de nieve	cubierto por la nieve
マドリードから来る	come from Madrid	viene de Madrid	viene desde Madrid
1 つのティーセット	a tea set	un juego de té	un juego para té
赤く塗られている	painted red	pintado de rojo	pintado en rojo
人生について話す	to talk about life	hablar de la vida	hablar sobre la vida
グラス 1 杯の水	a cup of water	un vaso de agua	un vaso con agua

前述の Bosque (1996) が挙げる 4 前置詞 con, de, en, por は、これら 7 つの前置詞 con, de, desde, en, para, por, sobre に包含されている。Bosque (1996) の挙げていない 3 つの前置詞のうち、desde, sobre は Fernández Ramírez (1986) が空間指示性を持つものとして挙げており、記述のない para についても、その空間的意味に加えて、pro/per + a、すなわち pro あるいは per に a が付随して形成されたものであるとする通時的見解[6]に留意すれば、後続する被制語の名詞句に数・定性を求めやすい a[7] と類似した分布を見せることが予想される。したがって、López (1972) の挙げる 7 前置詞は、後続名詞句に数・定性を求めにくい con, de, por と、求めやすい desde, para, sobre, そして両者に属する可能性がある en により構成されていることから、本テーマに適した分析対象であると言える。

3.2. 分析の手順

本分析の手順は、以下のとおりである。まず、既存のスペイン語コーパスから、7 前置詞に後続する名詞句に関する数・定性の頻度を収集し、その後データを標準化することで、各前置詞に関する変数を導出する。そして、求めた変数を用いて 7 前置詞に関する階層的クラスター分析を実施し、形成されたクラスターを考察する。なお、変数間の非類似度の算出にはユークリッド距離を、クラスター間の距離算出にはウォード法を使用する。また、分析には統計ソフト R を使用する。

[5] 表 4 は López (1972) の挙げる例を引用したものであるが、一部修正して掲載している。
[6] para の形成に関しては、pro + ad > pora > para とする見解 (RAE, 2014) と、per + ad > (par+ ad) pora > para (Coromines and Pascual, 1985) とする見解があるが、いずれの形も a の要素 (ad) を有している。本稿では、para の形成に関する議論に関しては態度を保留し、a を含んだ前置詞であるという見解の一致のみを援用する。
[7] 前置詞 a に後続する名詞句の数・定性の明確性は Bosque (1996: 78-79) などで言及されている。

3.3. 使用コーパスの詳細

本分析では、esTenTen [2011, Eu + Am] というコーパスを使用する。esTenTen [2011, Eu + Am] は、コーパス検索ツール Sketch Engine にて公開されているコーパスであり、2011 年に公開されていたスペイン語圏 19 か国のウェブページをもとに構築されたスペイン語均衡ウェブコーパスである。総収録語数は約 95 億語となっており、筆者の管見の及ぶ限り、約 450 億語を収録する通時コーパス Google Books に次ぐ、最大規模のスペイン語コーパスである。スペイン語には大別して、スペインで用いられるヨーロッパスペイン語と、中南米で用いられるアメリカスペイン語があるが、このコーパスを分析に用いる理由は、スペイン・中南米両地域のデータを豊富に含んでいる点と、CQL (Corpus Query Language) を用いた柔軟な検索が可能である点による[8]。

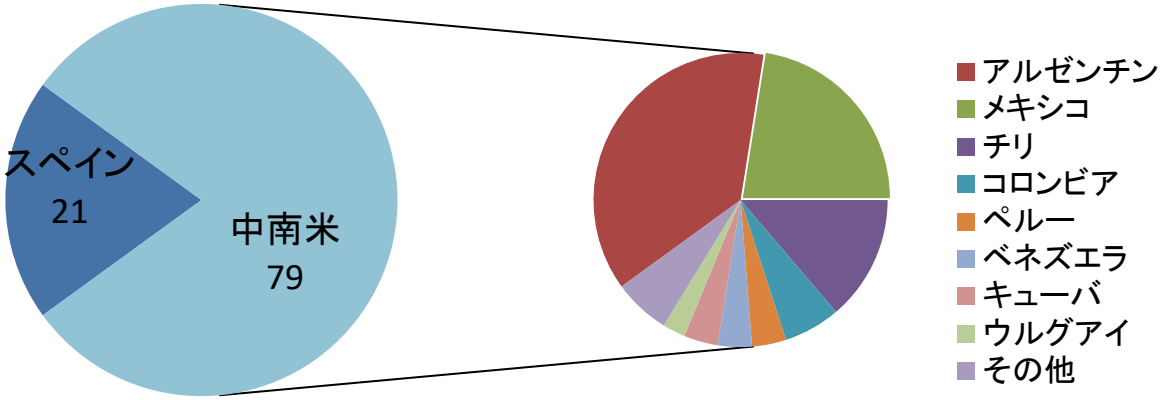


図 1. esTenTen [2011, Eu + Am] の構成国内訳

4. 分析

4.1. データ収集

コーパスからデータとして収集したいのは、限定詞・複数語尾を伴った名詞句が 7 前置詞に後続する語列、すなわち con NP, de NP, desde NP, en NP, para NP, por NP, sobre NP である。しかし、前述のとおり、スペイン語には後置修飾句・節が頻繁に生起するため、後続する語列に何らかの制限を設けないと、NP の数・定性を、前置詞か NP のいずれかに求めることができなくなる可能性がある。したがって、抽出する語列を「前置詞＋(限定詞)＋普通名詞＋ピリオド」という構造に限定して検索式を作成した。以下がその詳細である。

品詞タグ

普通名詞	普通名詞単数形	限定詞[9]	ピリオド
NC	NC.S	D	Fp

[8] 詳細は Kilgarriff et al. (2014) や Kilgarriff and Renau (2013) 参照。

[9] コーパスの限定詞タグ D には数詞と関係形容詞が含まれていない。しかし、数詞は他の限定詞との共起が唯一可能である点から、純粋な限定詞ではないと言えないため、分析から除外した。関係形容詞についても、直後にピリオドが来る語列には表れにくい語であるため、分析対象外とした。

検索式 1 : [word="前置詞"][tag="NC.*"][tag="Fp"]
抽出例: con discapacidad. (障害を抱えて), para niños. (子供のために)
検索式 2 : [word="前置詞"] [tag="D.*"]?[tag="NC.S*"][tag="Fp"]
抽出例: desde el exterior. (外から), sobre la salud. (健康に関して)

検索式 1 は、普通名詞が限定詞を伴わず 7 前置詞に後続し、直後にピリオドが来る語列を抽出する式であり、これをもとに高頻度上位 100 の普通名詞の総用例数に占める、単数形用例数の相対頻度 α を導出した。また、検索式 2 は、普通名詞単数形が限定詞を任意で伴って前置詞に後続したのちにピリオドが生起する語列を抽出する式であり、これにより、高頻度上位 100 の普通名詞単数形の総用例数に占める、限定詞を伴わない用例数の相対頻度 β を導出した。なお、普通名詞のタグ NC/ NC.S で抽出された語が以下に該当した場合は、ノイズとして除外した。

A) 固有名詞: internet, iPhone など
B) 代名詞: ti (前置詞格代名詞 2 人称単数形) など
C) 副詞: atrás (後ろへ), mañana (明日) など
D) 月を表す名詞: enero (1 月), febrero (2 月) など
E) 規範的でない、あるいは正書法に即していない表現: tod@s (みんな), dia[10] (日) など
F) 複合名詞のため、形態上単複の区別がつかないもの: portaaviones (空母) など
G) 名詞の語彙素性が数性を帯びているもの: tijeras (ハサミ), pantalones (ズボン) など
H) 使用が特定の地域に集中している語: camote (サツマイモ), ómnibus (バス) など
I) その他: URL など

以上の操作から導出された 2 変数 α, β と、α, β を標準化した Z_α, Z_β は次のとおりである。

表 5. α, β および Z_α, Z_β の数値[11]一覧

	α	β	Z_α	Z_β
con	0.91	0.44	0.66	0.78
de	0.78	0.55	0.11	1.27
desde	0.77	0.06	0.03	-0.92
en	0.97	0.22	0.98	-0.20
para	0.28	0.02	-2.18	-1.09
por	0.94	0.55	0.81	1.24
sobre	0.67	0.02	-0.41	-1.08

4.2. 数性・定性の無相関検定

本稿では、7 前置詞をケース、Z_α, Z_β を変数としたケースクラスター分析を実施するわけであるが、その前に 2 変数 Z_α, Z_β の母集団である、前置詞に後続する名詞句の数性と定性

[10] tod@s は todos (全員) の男女両形を総称する非規範的な表現である。dia には、正書法上必要なアクセントが付されていない。
[11] 数値は小数点第 3 位を四捨五入して表記している。

における相関の有無を確認しておきたい。帰無仮説 H_0 を「母相関係数は 0 である」、対立仮説 H_1 を「母相関係数は 0 ではない」として t 検定を行うと、2 変数 Z_α, Z_β の相関係数 r は 0.61 となり、検定統計量 T は 1.72 となる。これは自由度 5 の t 分布にしたがうので、検定量 T は有意水準 5 % の臨界値 2.57 を下回る。よって、帰無仮説 H_0 は棄却されず、数・定性の相関は有意なものであるとは言えない。したがって、本分析では、これらの標本である Z_α, Z_β を独立した 2 変数であるとして、クラスタリングに使用する。

4.3. クラスター分析

以上をふまえ、統計ソフト R を使用して、7 前置詞をケース、Z_α, Z_β を変数とする階層的ケースクラスター分析を実行した。ユークリッド距離にて算出したケース間の距離を示した非類似度表、2 変数を軸とした平面上に占める各ケースの座標を示した散布図と、ウォード法を用いたクラスタリングの樹形図は以下のとおりである。分析の結果、7 前置詞から {con, de, por} {desde, en, para, sobre} の 2 つのクラスターが形成された。

表 6. 非類似度表

	con	de	desde	en	para	por	sobre
con							
de	0.73						
desde	1.82	2.19					
en	1.03	1.70	1.20				
para	3.41	3.29	2.21	3.28			
por	0.48	0.70	2.30	1.45	3.79		
sobre	2.15	2.41	0.47	1.65	1.77	2.62	

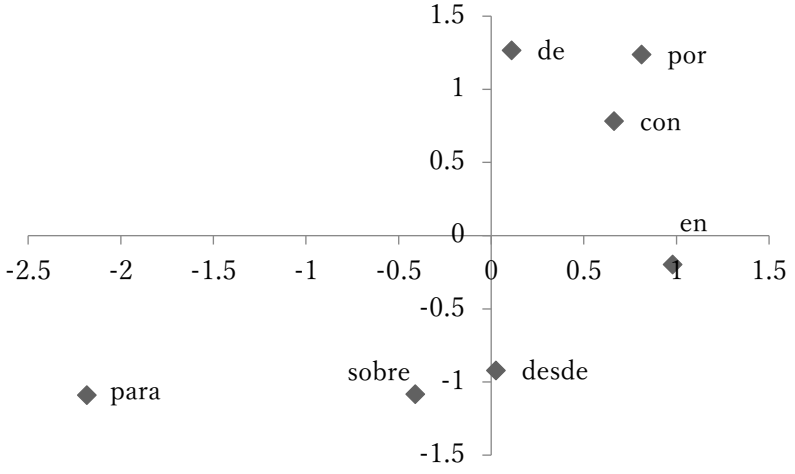


図 2. 散布図

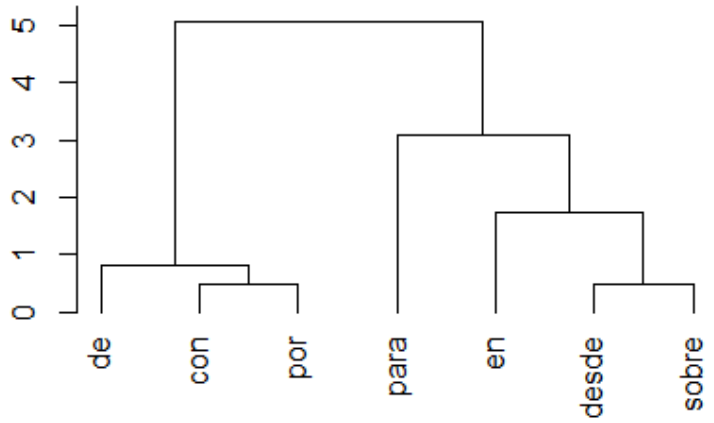


図 3. デンドログラム

4.4. 分析結果の考察

クラスタリングの結果、7 前置詞は {con, de, por} {desde, en, para, sobre} の 2 群に分類され、Bosque (1996) の挙げる 4 前置詞は同様の分布を見せず、2.3. にて挙げた事項が、検討の余地を残している問題であることが分かった。

しかし、散布図における 7 前置詞の座標を確認してみると、en が {con, de, por} と {desde, para, sobre} の中間に位置していることが分かる。デンドログラムでは、en が早い段階で後者のクラスターに組み入れられているが、これは para の孤立が関与しているものであり、en と {con, de, por} の分布上の類似性は否定できるものではなさそうである。したがって今後は、en の数・定性に留意しながら、空間指示性を補いうる何らかの指標を考案することで、前置詞句の数・定性に関する複合的な分類記述を試みるのが賢明であると言える。先行研究の唱える空間指示性は、各前置詞の意味に関する尺度であったが、統語・形態に関しても、それぞれ以下の尺度を想定できる。統語的尺度に関して、con, de, en, por は、関係代名詞と共起する際に定冠詞の付与が任意となる性質を共有している[12]。たとえば、以下の (8) では、下線部が a la que (to the which) ではなく a que (to which) となって、定冠詞単数女性形 la が省略されている。

(8) Esta es la persona a que me refería antes.

Est-a	es	la	persona	a	que	me	refería	antes.
this-F.SG	COP.3SG.IND.PRES	ART.DEF.F.SG	person	to	REL	REFL.1SG	refer-1SG.IND.IPFV	before

‘this is the person that I referred to before’

また、形態レベルでは、単音節性が共通の性質として挙げられる。3.1. において para の

[12] RAE and ASALE (2009: §44.2e) 参照。本稿では扱っていないが、a も含まれる。

成り立ちを考察したが、para をはじめとする複音節の前置詞の中には、複数の語が組み合わさって現在の形となり、結果として、より具体的な意味を有しているものもある。たとえば、desde はラテン語の前置詞句 de ex de の縮約形であり (Coromines and Pascual, 1985) 、複音節性と意味の具体性の相関も検討すべき尺度であると言える。

5. むすびにかえて

本稿では、これまでに体系的な分析が行われていない、スペイン語における前置詞句の数・定性について、Fernández Ramírez (1986) および Bosque (1996) が用いた空間指示機能の強弱が、分類の指標として適当であるかを、クラスター分析を通して考察した。その結果、7 つの前置詞 con, de, desde, en, para, por, sobre は {con, de, por} {desde, en, para, sobre} の 2 群に分類された。これにより、Bosque (1996) が後続名詞句に数・定性を求めにくい前置詞として挙げる con, de, en, por の分布の不一致が確認され、分類基準としての空間指示性の不透明性が明らかになった。これに伴い、本稿では、先行研究がとった意味的分類を補いうる形態・統語的分類基準の想定も試みた。今後の研究では、前置詞句の数・定性を、形態・統語・意味の 3 つの視座から分析し、複合的な体系記述を行っていきたい。

最後に、本稿で明らかになった 2 点の研究課題を挙げる。1 点目は、コロケーションとの親和性に着目した調査である。本稿のコーパス調査で得られた無冠詞の前置詞句の中には、イディオムとして学習書などで見られるものが相当数存在した。ゆえに、コロケーションの生産性と無冠詞名詞句の後続頻度の関連性を考察していくことが求められる。2 点目は、地域的有意差の検定である。本稿では、語彙以外に地域的差異の影響が表れることはないという立場をとったが、前置詞句の数・定性に関する体系的記述には、地域的有意差の有無を分析していくことも必要である。

謝辞

本稿は、日本ロマンス語学会第 56 回大会 (2018 年 5 月 12 日 於 京都大学) にて行った発表「スペイン語における前置詞後続名詞の数・定性―名詞の現働化による 7 前置詞のクラスタリング―」に一部修正を加えたものである。

略号一覧

-	inflexion	DEM	demonstrative	PL	plural
=	clitic boundary	F	feminine	POSS	possessive
1	first person	IND	indicative	PRES	present
3	third person	INDEF	indefinite	PST	past
ACC	accusative	INF	infinitive	REFL	reflexive
ART	article	IPFV	imperfective	REL	relative
COP	copula	M	masculine	SG	singular
DEF	definite	NEG	negative		

参考文献

Bosque Ignacio (ed.) (1996): *El sustantivo sin determinación*, Visor Libros.

Coromines Joan and Pascual José Antonio (1985): *Diccionario crítico etimológico castellano e hispánico,* Gredos.

Fernández Ramírez Salvador (1986): *Gramática española* (2.ª ed.), vol.3.2: *El pronombre*, Arco-Libros.

藤田健 (2011)「フランス語とスペイン語における不定冠詞の分布について」『北海道言語文化研究』北海道言語研究会, 9 号, pp. 1-22. (https://muroran-it.repo.nii.ac.jp/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=5375&item_no=1&page_id=13&block_id=21 よりダウンロード可能)

Kilgarriff Adam, Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel and Suchomel Vít (2014): "The Sketch Engine: ten years on", *Lexicography*, 1, pp. 7-36. (https://www.sketch.engine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf よりダウンロード可能)

Kilgarriff Adam and Renau Irene (2013): "esTenTen, a vast web corpus of Peninsular and American Spanish", *Procedia-Social and Behavioral Sciences*, 95, pp. 12-19. (https://www.sketchengine.eu/wp-content/uploads/esTenTen_web_corpus_of_Peninsular_and_American_Spanish_2013.pdf よりダウンロード可能)

López María Luisa (1972): *Problemas y métodos en el análisis de preposiciones*, Gredos.

RAE (Real Academia Española) (2014): *Diccionario de la lengua española* (23.ª ed.), (http://www.rae.es/rae.html よりアクセス可能 2018 年 8 月 27 日確認)

RAE and ASALE (Real Academia Española and Asociación de Academias de la Lengua Española) (2009): *Nueva gramática de la lengua española,* Espasa Calpe.

Solís García Inmaculada (2011): *El concepto de referencia y su utilidad en la didáctica del español como lengua extranjera*, Tesis doctoral, Universidad de Oviedo. (https://www.tdx.cat/bitstream/handle/10803/79991/UOV0090ISol%C3%ADs%20Garc%C3%ADa.pdf?sequence=6&isAllowed=y よりダウンロード可能)

寺﨑英樹 (1998)『スペイン語文法の構造』大学書林.

使用コーパス

Sketch engine. esTenTen [2011, Eu + Am] (9,497,213,009 palabras): (https://www.sketchengine.eu / よりアクセス可能 2018 年 9 月 9 日確認)

マルチアクティビティに伴う発話の分類：遂行発話と雑談

天谷 晴香（国立国語研究所音声言語研究領域）†

Utterances accompanied with multiactivity: Executional speech and small talk

Haruka Amatani (National Institute for Japanese Language and Linguistics)

要旨

日常的に会話はマルチアクティビティ(複合活動) の中で行われることが多い．他の活動を主として行う場面での会話は，雑談に加えて，活動の遂行に関わる発話が見られる．雑談も活動に関連して行われることが多く，全くその場の物事と関わらない発話は少ない．その場の物事に関した発話とその場にない物事に関する発話の区分は談話研究において重要である．Chafe (1994) は意識の宛先をもとにこれらの言語表現の分類を提示した.話し言葉においては話し手と聴き手が会話スペースを共有することがほとんどであるため，書き言葉に比べて場に即した言語活動が行われやすい．本発表では参加者が共同で料理活動を行いながらの会話データを扱う．料理行程を進捗させるための発話と料理の進捗に関わらない発話を分類する．

１．はじめに

日常会話は他の活動に伴って行われることが多い．会話内容は活動内容に影響される．また会話の種類は，活動に必要な会話と活動に関わらない雑談のような会話がある．

例えば，二人以上で共同で料理をする時，協力してひとつの作業を進める場面では特に，互いに発話によって作業の進行速度やタイミングを調整する必要がある．難度の高い共同作業–例えば一方がフライパンで材料に火入れをしている所に他方が調味料を入れるなど–の際には発話による調整も増えるだろう．一方で，一人で材料をかき混ぜるような単純作業のような場面では料理行程に関わらない発話も出やすいと考えられる．

マルチアクティビティ研究は，複数の活動がひとつのマルチアクティビティとして活動間の境界が際立つことなく実行される方法を分析するものである(Mondada 2011 他). 本稿は，活動に伴う発話をその性質によって分類し，分類された発話が活動の全体構造の中でどのような位置に発現するか探るものである．

２．発話の分類

日常の活動に伴う発話を分類するため，本研究では二つの指標を検討する．「その場の事物への関与」と「活動遂行への関与」である．

「その場の事物への関与」は活動の有無に関わらず一般的な会話の分類として有効な指標である．２．１節で詳細を見る．「活動遂行への関与」は，活動に伴う会話を特徴づけるを指標である．この指標には段階性を設ける．これらについては２．２節で論じる．

† h-amatani@ninjal.ac.jp

２．１．その場の事物への関与

その場の事物・状況に関係した発話であるか否かをひとつの指標とする．その場の事物への関わりの有無が会話の性質を測るひとつの指標にたりうることについて直接的な議論は少ないが周辺的に言及している研究は多い．[1]これは言語使用者にとって直感的な指標と言えるだろう．

話し手の意識の在りかを軸に Chafe(1994)はその場の事象に関わる発話と関わらない発話について分類の可能性について論じた．[2] Chafe は，言葉の受け取り手と距離のある書き手と異なり，話し言葉では発話者が発話の受け取り手と場を共有しているため，その場の事物に関する発話が多く見られると推測した．しかし彼の夕食会のデータにおいてはその場の事物に関わらない発話が多く見られている．

２．２．活動遂行への関与

活動遂行に関わる発話は，関与の度合いに段階がある．段階を捉えるために，まず関与度の強さ・弱さの両極にあたる発話の性質を考える．

マルチアクティビティ研究において会話活動を独立に扱う立場と他の活動の一部と捉える立場があるが，会話活動の独立性の段階について議論されることは少ない．独立性の度合いを明確にすることがマルチアクティビティをより詳細に理解する手立てとなる．会話活動の独立性の段階は，本稿における発話の活動遂行への関与の度合いに相補的である．

活動遂行に必要な発話として，活動の指示をする発話と，指示を実行するために相談・協議する発話が考えられる．

活動遂行の手順には二人以上で行う活動の場合，参加者間の関係によっていくつか異なるパターンが考えられる．一つには，参加者の一人が活動の手順に熟知しており活動を主導していくパターンがある．また他には，活動の手順をよく知らない参加者が集まって相談しながら遂行していくパターンが考えられる．いずれのパターンにおいても，指示発話と相談の発話は遂行に必要である．

指示の発話としては, 例えば天谷(2017)における一方の参加者が他方に化粧行為をする際の「眉毛を描きます」と声をかけるような発話がある．また相談・協議の発話としては同じデータから「このカクッとなってるのはそういう風に生えてるん？」と聞くような自身の現状認識を伝えるような発話が含まれる．

３．データ

現在構築中の『大規模日常会話コーパス』から，日常の料理場面（お菓子作り）のビデオデータを分析の対象とした．参加者は２名で，母親と中学生の息子である．ビデオは全体で約 35 分で，分析対象とした断片はビデオ開始から約 6 分から約 8 分の 2 分間である．

４．分析と考察

２節で示した発話の分類の指標を用いて発話データを分析した．対象の断片に見られた発話のうち，あいづち表現と感動詞表現は以下の分類から排除した．

分析対象の単位は統語的に文を構成するひとまとまりとした．『大規模日常会話コーパス』

---

[1]例えば，子供の指差しについて実験を行った So et al. (2010)では，So et al. (2009)との実験結果の違いを子供と大人の差だけでなく, 物語り(displaced story telling)とその場の会話(here-and-now conversation)の違いに求める記述をしている．

[2]Chafe は前者を「直接モード immediate mode」，後者を「置換モード displaced mode」の発話とそれぞれ呼ぶ．置換モードにより未来や過去の出来事についての言語表現ができる．

における発話単位がおよそこれに相当する．ただし，コーパスにおける発話単位では倒置文はふたつの発話単位とするが，本分析では統語的に文を構成する単位を対象とするために倒置文はひとつの単位として扱った．

４．１．発話の分類：その場の事物への関与

Chafe が推測したように発話全体の多くがその場の事物に関したものであった．分類の対象となった発話文 59 文中 45 文(76.27%)がその場の事物に関する発話であった．これは Chafe の夕食会のデータとは割合が反対だが，活動場面では夕食会のように会話がその集まりの主たる目的となる場面と異なり，その場で活動を遂行させる目的が主であるためと考えられる．

その場にない事物をトピックにした発話は 59 文中 14 文(23.72%)と限られた数であった．14 文中 2 文が断片直後に起きた事例で，12 文が断片の後半に起きた事例である．

その場の事物に関わらない発話事例と活動の構造との関係を見てみると，どちらも料理の行程と行程の間に起きている．一つ目は断片の開始直後に起きたもので，次の行程を開始するためにレシピを読み上げる発話である．料理活動においてレシピを読み上げる行為は，より一般的な発話機能として捉えると「計画」に当たるものである．二つ目の事例では，話の受け手側は別の行程に着手しているが，トピックの導入者はひとつの行程を終え片付けをしながらトピックを導入し発話している．この時のトピックは，料理用のはかりを買いに行く店についてであった．母親が「今度料理用のはかり買おうかな.」「最近いいお店見つけたから.」とトピックを導入したものである．

４．２．発話の分類：活動遂行への関与

活動遂行への関与について，活動の遂行に関わらない発話と，遂行に関わる発話に分類した．さらに遂行に関わる発話を２．２節で論じたように指示発話と相談・協議発話に分類した．それらに当てはまらない発話も見られた．

活動の遂行に関わらない発話は全体 59 文中 21 文(35.59%)であった．遂行に関わらない発話もその場の事物と関係しており，データの中では息子が母親に材料のチョコレートをすこし食べていいか聞く「一つかじっていいべ.」などの発話が見られた．

活動の遂行に関わる発話は 59 文中 38 文(64.4%)であった．

活動の遂行に関わる発話のひとつのカテゴリーである指示発話は 59 文中 17 文(28.81%)であった．指示発話の例としてはレシピの読み上げや「余ったらこの辺に置いといて.」というような材料の扱いに関する指示が見られた．

また相談・協議発話は 59 文中 18 文(30.5%)であった．発話内容は，料理に使用するバターの分量をはかる際に発された「三十グラムってこんななの？」や，バターを器に塗りつける行程の途中に見られた「お母さん，かなり塗ったと思うけど．俺.」のように参加者が違いに自身が行っている行程の状態を相手に報告するものであった．

さらに遂行に関わる発話のうち指示や協議に分類されないものが 3 文(5.08%)見られた．これらは息子がバターを器に塗る際に「塗ーり.」と発話したものであった．このような発話は感動詞的にも捉えられるが，ここでは作業の状態を報告しているものとして相談・協議発話に準ずる説明発話として分類した．

４．３．二つの指標の交わり

その場の事物への関与と活動遂行への関与，二つの指標を用いることでより詳細な発話の分類が可能になった部分がある．

活動に関与する発話は通常，その場の事物に関わるものと考えられる．これには４．１節で報告した，料理用のはかりを買う店についてのトピック発話群が相当する．一方で，活動の遂行に深く関与しながらその場の事物に関する発話でなかったものが，レシピの読

み上げである．レシピの読み上げは，上で述べた通り，一般的な機能として「計画」の発話としての機能を持っており, Chafe がその場の事物に関わらない発話の特徴として述べた「未来」に関する発話に相当するものであると考える．レシピの読み上げ発話は，「未来」に行う行程あるいは「未来」に出来上がるものについて述べた発話であるため，その場の事物に関わったものではない．

料理用のはかりについてのトピック発話とレシピの読み上げ発話は同じようにその場の事物に関わらない発話であるが，活動の遂行への関与の指標を取り入れることで一方は活動の遂行に関わらずもう一方は活動の遂行に関わっており，マルチアクティビティの中で異なる機能を持った発話として捉えることができた．

また，その場の事物に関わらない相談・協議の発話として「量れって，うち，はかりない.」という発話が見られた．これはそのしばらく後に現れる「今度料理用のはかり買おうかな.」のトピックの導入に繋がるものである．今回の指標を用いると「量れって，うち，はかりない.」はその場の事物に関わらない活動の遂行に関わる相談・協議の発話として分類され，「今度料理用のはかり買おうかな.」はその場の事物に関わらない活動の遂行に関わらないトピック発話として分類される．このように複数の指標によってトピックの移行のグラデーションを捉えることが可能になる．

５．おわりに

本稿では日常の活動に伴う発話の分類を，「その場の事物への関与」と「活動の遂行への関与」を軸に試みた．「その場の事物への関与」については，Chafe が述べたように関与度の高い発話が多く見られた．またその場にないものを話題とした発話は活動の行程の間に発現した．「活動の遂行への関与」については段階性を捉える試みとして，活動の遂行に関わる発話をその性質によって指示発話、相談・協議発話、説明発話に分類した．また，二つの指標を用いることでより多角的にマルチアクティビティに伴う発話の分類を行えることを示唆した．

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」による成果を利用して行われたものである。

文　献

天谷晴香 (2017). 「他者への化粧行為と雑談」.『日本認知科学会第 34 回大会発表論文集』.

Wallace Chafe (1994). *Discourse, Consciousness, and Time*. Chicago: The University of Chicago Press.

Lorenza Mondada (2011). The organization of concurrent courses of action in surgical demonstrations. In Jurgen Streek, Charles Goodwin, and Curtis LeBaron (eds), *Embodied Interaction: Language and Body in the Material World,* pp. 207-226. Cambridge: Cambridge University Press.

Wing Chee So, Ozlem Ece Demir, and Susan Goldin-Meadow (2010). When speech is ambiguous geture steps in: Sensitivity to discourse-pragmatic principles in early childhood. *Applied Psycholinguistics* 31-1, 209-224.

Wing Chee So, Sotaro Kita, and Susan Goldin-Meadow (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science* 33, 115-125.

コーパスに基づく字順転倒漢語の網羅的把握の試み

間淵 洋子（国立国語研究所 言語変化研究領域）†

A Corpus Based Study of Sino-Japanese Words with the Reversed Order of Characters in Modern Japanese

MABUCHI, Yoko (National Institute for Japanese Language and Linguistics)

要旨

近代に特徴的に多く見られる「華麗」と「麗華」のように字順が逆転した漢語対を，近代語のコーパスを用いて網羅的に抽出し，対となる各語の使用頻度が 10 以上の約 400 の漢語対を選定した。これら約 400 語について，現代語においても対をなして残っているか，どちらかが淘汰され 1 語に集約されたか，あるいは両語共に淘汰されたかといった，使用状況の変化を調査し，対となる各語間の意味的関係性や近代における使用頻度と，現代での使用状況とに関わりがあるかを検討した。その結果，両語の意味関係において異なりがあるものは両語が併存する傾向，また，使用頻度の高い語形は現代語に残存する傾向を見出すことができた。一方で，辞書的意味においてほぼ同義と判断される漢語対においても，多くの漢語対で両語が残存していた。これらの対は，①一方が極めて限定的に現代語コーパスで用いられているだけで，実際には他方の優勢な語形にほぼ統一化されており，語の淘汰の過渡期と見られるものや，②辞書における語義には大差が認められないものの，一方の用法が限定・固定化されており（連語，文法機能，特殊な意味・文脈等），用法の分化が明らかなために併存しているものが多く，近代漢語の変化の方向性は，１義１語を志向して語が淘汰されていると位置づけることができた。

１．はじめに

近年，様々な時代の日本語を対象とした大規模なデータベースが整備されつつある。特に，語彙を対象とした研究には欠かせない単語情報の付与されたコーパスによって，これまでは難しかった，日本語の語彙の全体像を実証的に捉えることが可能になり，これらを用いた通時的な語彙研究も行われるようになっている（田中 2016 など）。

発表者はこれまで，これらの言語資源を活用することで，近代と現代との間に見られる漢語の差異の実態や変化について分析・検討を行ってきた（間淵 2016a，間淵 2016b，間淵 2017a，間淵 2017b，間淵 2018）。これらの研究は，これまで多く指摘されてきた近代漢語の特異性（池上 1984，武部 1981，田島 1998a，今野 2012 など）について，大規模データを活かした計量的手法により大局的な観点からその実態を見渡し，変化の方向性とその背景を明らかにすることを目的に行ってきたものである。その結果が示唆するのは，近代漢語の語彙・表記・語法に見られる変化は，いずれも多様性が次第に淘汰され画一化されるという方向性を持っており，その背景原理が「言語運用の合理化」「語と意味の１対１対応」にある，ということであった。

本研究では，これらの変化の方向性と背景原理を裏打ちすると思われる，さらなる事象の一つとして，「華麗」と「麗華」，「遊戯」と「戯遊」のように，字順が逆転した 2 字漢語対（以下「字順転倒漢語」と呼ぶ）を取り上げ，コーパスを用いて実在するペアの抽

† mabuchi@ninjal.ac.jp

出と，出現実態を明らかにする。これらの語については，鈴木(1979, 1986)，田島(1998b, 1998c)等，多くの語を取り上げた研究がなされているが，その結論は，①近代初期に多くの字順転倒漢語が認められること，②同義・同用法の漢語対はどちらかの語が衰退・淘汰の傾向にあること，③後年（現代）まで併用される対は，意味や位相の分化が起きて使い分けがされていることなどである。ここでは，この指摘について計量的に実証することを目指すとともに，近代における使用実態の変遷にも着目し，これまで発表者が明らかにしてきた，表記や語法の画一化の様相や，特に画一化が進行する時期などと共通点が見られるかを観察する。

２．研究方法

２.１　コーパス

本研究では，近代における字順転倒漢語の実態をできる限り網羅的に捉え，また，それらが現代語においてどのように用いられているか把握するために，近代語と現代語のコーパスを用いた調査を行う。両時代のコーパスにおける出現状況を比較することで，字順転倒漢語の通時的な変化を見出すことができるはずである。対象とするコーパスの概要を表１に示す。

表１　調査対象コーパスの語彙量

時代	資料	出版年	語数（万）
近代	明六雑誌	1874-5	18
	国民之友	1887-8	101
	太陽	1895	202
		1901	197
		1909	187
		1917	180
		1925	203
	女学雑誌	189-45	59
	女学世界	1909	52
	婦人倶楽部	1925	54
	全体	1874-1925	1,253
現代	BCCWJ(出版 SC)	2001,2005	1,234

近代語のコーパスとしては,『日本語歴史コーパス　明治・大正時代編 I 雑誌（以下,「CHJ 明治大正雑誌」または「CHJ」と略記)』（収録語数約 1,253 万語，自立語数約 754 万語）を用いる。

これと対照する現代語のコーパスとしては,『現代日本語書き言葉均衡コーパス（以下,「BCCWJ」と表記)』の出版サブコーパス（以下,「SC」と表記）のうち 2001 年，2005 年の発行分の可変長サンプルを用いる（記号等を除く収録語数約 1,234 万語，助詞・助動詞を除く自立語数約 751 万語)。BCCWJ には，出版 SC のほかに，現代における言葉の流通実態を捉えるのに適した図書館 SC，個別の研究目的に沿うデータを集めた特定目的 SC があるが，本研究では，比較する近代のコーパスが雑誌のみであるため，逐次性の観点から共通性の高い資料として，雑誌や新聞を含む出版 SC を対象とする。出版 SC より 2001 年分と 2005 年分の 2 カ年のみを用いたのは，近代語コーパスと語数を概ね同様になるよう

に調整するためである。

２.２　調査対象語の抽出

近代で使用の見られる字順転倒漢語対を，以下の手順で抽出した。

1) 「CHJ 明治大正雑誌」「BCCWJ」から網羅的に 2 字漢語を抽出し，両コーパスにおけるすべての 2 字漢語の語彙表を作成（間淵 2017b を参照）
2) 「CHJ 明治大正雑誌」に出現するすべての漢語（＝近代漢語）に対する字順転倒文字列をエクセルの文字列操作関数等により作成
   例 : 漢語 A「漢語」→「=right(対象語のセル)&left(対象語のセル)」→転倒文字列 B「語漢」
3) 近代漢語の語彙表（頻度降順のもの）に，手順 2 で作成した転倒文字列 B が含まれるかをエクセルの MATCH 関数等で評価し，字順転倒漢語対を抽出
   例：「=match(転倒文字列セル,近代漢語の列範囲,0)
   →戻り値が「#N/A」(該当なし)のものは，字順転倒漢語対ではない
4) 字順転倒漢語対の重複（語 A-語 B の対と語 B-語 A の対）を解消
   例：=if(手順 3 の戻り値-ROW()>0,true,false)
   →戻り値 true のみを残す。近代で優勢（より高頻度）の語 A と劣勢（より低頻度）の語 B の字順転倒漢語対ができる。

上記手順により得られた，近代に見られる字順転倒漢語対は，2,239 対であった。このうち，両語の使用頻度が 10 以上のものを今回の分析対象とし，粗頻度 10 未満の語を含む対を除外した。さらに，「規定（キテイ）」と「定規（ジョウギ）」のように，字順転倒により漢字音の読みが異なるものについても対象外とした。これにより，分析対象として残ったのは，395 対 790 語となった。

２.３　分析の観点

先行研究で用いられる分析のための枠組みはおよそ以下 4 点にまとめられる。

1) 意味　対となる語同士の意味の重なり。同義か，類義か，異義か等。
2) 語法　各語の品詞的用法。名詞か，動詞か，形容詞か，副詞か。動詞の場合，自動詞か他動詞か等。
3) 語構成　各語を構成する漢字同士の関係。同義・類義または対義の字の並立関係か，述語－目的語，修飾－被修飾といった文法的関係か等。
4) 使用頻度　各語の頻度と勢力関係。いずれも低頻度・高頻度か，頻度に偏りがあるか等。

本研究では，近現代における漢語語彙の変化の把握を主たる研究目的とするため，近代で特異に多く存在した字順転倒漢語対が現代では減少した，という語彙変化との相関において，より重要な観点と見られる 1)意味と 4)使用頻度を中心に調査分析を行う。これら２項が，近代と現代における使用頻度を比較し得られる字順転倒漢語対の変化パターン（両語残存，両語消失，１語残存１語消失等）に，どのように関連するかを検討していくこととする。

３　調査結果

３.１　近現代間に見る使用状況の変化

まず，字順転倒漢語対が近代から現代にかけてどのように変化したか，あるいはしなかったかを把握するために，各語の使用状況（使用の有無）により以下の通り 4 分類した。

AB 併存：近代優勢漢語 A，近代劣勢漢語 B 共に，現代でも一定の（2 例以上）使用が認められる
A 残存 B 消失：近代優勢漢語 A のみ，現代での使用が認められる
B 残存 A 消失：近代劣勢漢語 B のみ，現代での使用が認められる
AB 消失：近代優勢漢語 A，近代劣勢漢語 B 共に，現代では（ほぼ）使用が認められない

分類結果を表 2，図 1 に示す。

表 2　現代での使用状況に基づく分類

分類	ペア数	字順転倒漢字対の例（A 粗頻度上位 25 対）
AB 併存	207	政治/治政,社会/会社,国民/民国,国家/家国,事実/実事,人民/民人,外国/国外,議会/会議,多数/数多,法律/律法,文明/明文,女子/子女,平和/和平,同一/一同,議論/論議,権利/利権,感情/情感,科学/学科,中心/心中,統一/一統,運命/命運,習慣/慣習,便利/利便,決議/議決,少年/年少
A 残存 B 消失	135	主人/人主,旅行/行旅,制限/限制,範囲/囲範,生存/存生,簡単/単簡,平生/生平,人士/士人,古今/今古,往来/来往,長官/官長,気風/風気,英仏/仏英,富豪/豪富,抵抗/抗抵,制裁/裁制,善良/良善,野蛮/蛮野,支度/度支,権威/威権,理事/事理,結論/論結,風習/習風,良好/好良,風流/流風
B 残存 A 消失	23	売淫/淫売,人才/才人,退隠/隠退,著大/大著,国中/中国,上進/進上,桂月/月桂,懐抱/抱懐,威武/武威,知了/了知,人前/前人,狂熱/熱狂,書信/信書,政法/法政,航通/通航,天則/則天,愛他/他愛,編中/中編,林学/学林,皇上/上皇,風通/通風,稿本/本稿,険峻/峻険
AB 消失	30	英独/独英,露仏/仏露,属僚/僚属,学政/政学,振作/作振,米独/独米,論定/定論,明月/月明,精励/励精,差等/等差,想美/美想,郡市/市郡,治平/平治,婢僕/僕婢,国君/君国,才学/学才,慈仁/仁慈,深甚/甚深,同公/公同,容儀/儀容,厭倦/倦厭,衆多/多衆,軍中/中軍,火熱/熱火,省減/減省

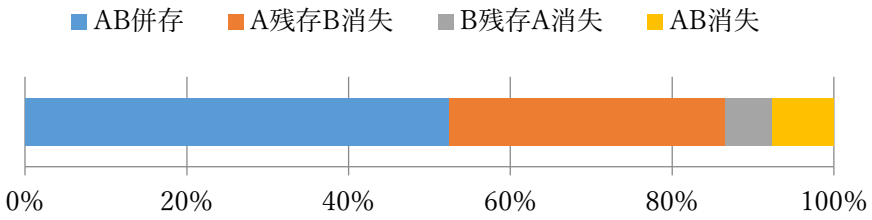


図 1　字順転倒漢字対の現代での使用状況割合

近代で見られた字順転倒漢字対は，半数以上が現代でも残存しているが，40%程度の対で片方の語が淘汰されており，そのほとんどは近代において劣勢の（使用頻度がより低い）語が消失している。

３.２　使用頻度と語彙レベル

近代における使用頻度の多寡については，まず使用状況分類ごとに平均値を求めた。結

果を以下に示す。

表 3　使用状況と使用頻度

	AB 併存	B 消失	A 消失	AB 消失
語 A 平均	479.1	147.3	47.6	38.7
語 B 平均	88.5	22.3	22.0	16.9

さらに，田中（2011）の「語彙レベル」の枠組みを参考に，語のカバー率（累積使用率）により以下の通りレベル分類を行った[1]。

表 4　語彙レベルの分類基準値

レベル	カバー率	頻度範囲	調整頻度範囲
A	60%	1,125-358,407	150.51-47,951.51
B	80%	214-1,122	28.63-150.11
C	90%	60-213	8.03-28.50
D	95%	23-59	3.08-7.89
E	99%	4-22	0.54-2.94
F	100%	1-3	0.13-0.40

次に，レベル AB を高頻度，CD を中頻度，EF を低頻度として，対をなす語 A－語 B（両者の頻度は A＞B）の頻度組み合わせを「高－高」「高－中」「高－低」「中－中」「中－低」「低－低」の 6 分類とし，前節に示した使用変化の 4 分類とのクロス集計を行った。結果を表 5 に示す。

表 5　漢語対の頻度と使用状況変化との相関

	AB 併存	A 残 B 消	B 残 A 消	AB 消失	総計
高－高	10				10
高－中	55	9			64
高－低	19	15			34
中－中	65	29	11	5	110
中－低	52	70	5	15	142
低－低	6	12	7	10	35
総計	**207**	**135**	**23**	**30**	**395**

漢語対の頻度と使用状況変化との関係を見ると，AB どちらも近代において中高頻度の対は併存する語が多いが，近代で B が低頻度のペアは，B が淘汰される傾向が強い。

３.３　両語の意味的関係性

対となる語同士の意味的関係性については，田島(1998b)の枠組みに倣い表 6 に示す 4 分類とした。具体の漢語対を分類する際には，『日本国語大辞典　第 2 版』[2]の記述に基づき

[1] 分類手法の詳細は，間淵 2017a を参照のこと。

[2] 田島(1998b)では，現代語の中型辞典を語義分類の判定に用いており，第 1 義が現代語においてより基本的な語義であることを意図したものとするが，本研究では，近代における

判断した。なお，辞書に掲載のない語を含む場合は，意味関係を「不明」とした。
先に 3.1 節において示した使用変化と意味関係分類のクロス集計結果を表 7 に示す。

表 6　両語の意味関係分類

分類	意味関係	本研究における分類基準
I	ほぼ同義	意味記述に字順転倒語が用いられており，内容がほぼ重なるもの
II(1)	類義（重なり大）	両語の第 1 義において，Iの基準に当てはまるが，両語あるいはどちらかの語が第２義以下の異なる語義を持つもの
II(2)	類義（重なり小）	両語あるいはどちらかの語が第２義以下の異なる語義を持ち，第２義以下の語義において字順転倒語と同義関係にあるもの
III	異義	I，II(1) ，II(2)に当てはまらないもの

表 7　意味関係と使用変化の相関

	AB 併存	B 消失	A 消失	AB 消失	総計
Ⅰ	62(42%)	64(43%)	10(7%)	12(8%)	148(100%)
Ⅱ(1)	43(57%)	27(36%)	1(1%)	4(5%)	75(100%)
Ⅱ(2)	26(63%)	11(27%)	4(10%)		41(100%)
Ⅲ	74(60%)	31(25%)	8(7%)	10(8%)	123(100%)
不明	2(25%)	2(25%)		4(50%)	8(100%)
総計	207	135	23	30	395

類義・異義のものは対となる両語が現代でも併存する割合が高く，同義のものは劣勢の語 B が淘汰される割合が，他よりもやや高い。

分析対象とする 395 の字順転倒漢語対に関して，上記分類を適用した結果，Ⅰ，Ⅱ(1)，Ⅱ(2)に分類された 264 対を表 7 に掲載する。なお，田島(1998b)に言及のある対（以下,「田島リスト」）の場合，その分類結果をそのまま利用することとし，これに含まれない対（以下「新規」）と別に示す。

同義性に着目したいため，原義から派生義の順に配置されている『日本国語大辞典　第 2 版』を用いることとした。

表 8　意味関係による漢語対の分類

分類	田島リスト		新規	
Ⅰ 148 対	70 対	多数/数多,平和/和平,議論/論議,統一/一統,運命/命運,旅行/行旅,便利/利便,生存/存生,簡単/単簡,消費/費消,国内/内国,君主/主君,苦痛/痛苦,評論/論評,途中/中途,製作/作製,治療/療治,分配/配分,苦労/労苦,事変/変事,祖先/先祖,産物/物産,真正/正真,練習/習練,合併/併合,現出/出現,木材/材木,明言/言明,圧制/制圧,見識/識見,愛情/情愛,剰余/余剰,売淫/淫売,関連/連関,名声/声名,争闘/闘争,生死/死生,熱情/情熱,人才/才人,食糧/糧食,襲来/来襲,音声/声音,争論/論争,誕生/生誕,継承/承継,微細/細微,補填/填補,修学/学修,上奏/奏上,忠誠/誠忠,武勇/勇武,略奪/奪略,変転/転変,平衡/衡平,降下/下降,威武/武威,造営/営造,遊歴/歴遊,保留/留保,敵対/対敵,社寺/寺社,戦敗/敗戦,綱紀/紀綱,書信/信書,教示/示教,婢僕/僕婢,認容/容認,下垂/垂下,険峻/峻険,補修/修補	78 対	国民/民国,人民/民人,制限/限制,平生/生平,英仏/仏英,富豪/豪富,抵抗/抗抵,民衆/衆民,善良/良善,野蛮/蛮野,風習/習風,良好/好良,英米/米英,商工/工商,熟練/練熟,滅亡/亡滅,復興/興復,鋼鉄/鉄鋼,見聞/聞見,認識/識認,抑制/制抑,西南/南西,英独/独英,忍耐/耐忍,覚醒/醒覚,辛苦/苦辛,異同/同異,歩行/行歩,露仏/仏露,展開/開展,過大/大過,横暴/暴横,西北/北西,退隠/隠退,詩文/文詩,約定/定約,東南/南東,感興/興感,属僚/僚属,抑圧/圧抑,憂鬱/鬱憂,源泉/泉源,従順/順従,福利/利福,該当/当該,絶滅/滅絶,畏敬/敬畏,喜悦/悦喜,米独/独米,調整/整調,興隆/隆興,財貨/貨財,婚約/約婚,独仏/仏独,知了/了知,銃砲/砲銃,精励/励精,正邪/邪正,随伴/伴随,狂熱/熱狂,艦船/船艦,治平/平治,毛髪/髪毛,往古/古往,弱小/小弱,景勝/勝景,航通/通航,慈仁/仁慈,深甚/甚深,成育/育成,愛他/他愛,独露/露独,厭倦/倦厭,上座/座上,累積/積累,省減/減省,液汁/汁液,絶佳/佳絶
Ⅱ (1) 75 対	30	感情/情感,習慣/慣習,往来/来往,始終/終始,大同/同大,回転/転回,趣旨/旨趣,結論/論結,光栄/栄光,奇怪/怪奇,実情/情実,接触/触接,時日/日時,趣意/意趣,退廃/廃退,礼拝/拝礼,天上/上天,夜半/半夜,弁論/論弁,途方/方途,隔離/離隔,威勢/勢威,情欲/欲情,図絵/絵図,親和/和親,由縁/縁由,離別/別離,氷結/結氷,協和/和協,心痛/痛心	45	政治/治政,社会/会社,国家/家国,事実/実事,範囲/囲範,決議/議決,少年/年少,古今/今古,証明/明証,産出/出産,権威/威権,階段/段階,一斉/斉一,性情/情性,詔勅/勅詔,舞踏/踏舞,海陸/陸海,文法/法文,衆人/人衆,女児/児女,山海/海山,着実/実着,水分/分水,慰安/安慰,土壌/壌土,知覚/覚知,勲功/功勲,線条/条線,与党/党与,富国/国富,上進/進上,水流/流水,説話/話説,年期/期年,胃腸/腸胃,東北/北東,英和/和英,差等/等差,上面/面上,花月/月花,動地/地動,容儀/儀容,深浅/浅深,霜雪/雪霜,煙火/火煙
Ⅱ (2) 41 対	15	議会/会議,進行/行進,接近/近接,所在/在所,色彩/彩色,愛敬/敬愛,局部/部局,居住/住居,感触/触感,望遠/遠望,懐抱/抱懐,近親/親近,異変/変異,石碑/碑石,交情/情交	26	法律/律法,女子/子女,同一/一同,中心/心中,人士/士人,現実/実現,出発/発出,長官/官長,気風/風気,格別/別格,長所/所長,理事/事理,風流/流風,出世/世出,根性/性根,女工/工女,害毒/毒害,理法/法理,春陽/陽春,地平/平地,国中/中国,桂月/月桂,教法/法教,急性/性急,政法/法政,末端/端末

（備考）上記に含まないⅢ異義 123 対，不明 8 対の内訳は以下の通り。

Ⅲ(123 対)　外国/国外,文明/明文,権利/利権,主人/人主,科学/学科,大事/事大,発揮/揮発,議院/院議,内部/部内,本国/国本,作家/家作,馬車/車馬,年来/来年,人道/道人,理学/学理,理論/論理,空中/中空,関税/税関,製鉄/鉄製,動機/機動,国王/王国,室内/内室,上陸/陸上,制裁/裁制,軍国/国軍,祖父/父祖,支度/度支,実質/質実,牛乳/乳牛,部下/下部,気運/運気,生長/長生,客観/観客,名家/家名,同会/会同,数字/字数,女王/王女,義理/理義,人工/工人,女皇/皇女,席上/上席,国立/立国,質素/素質,理性/性理,年中/中年,法王/王法,素養/養素,人家/家人,下流/流下,名人/人名,事故/故事,著名/名著,人知/知人,

文人/人文,風光/光風,礼儀/儀礼,前面/面前,経常/常経,中隊/隊中,波長/長波,毒蛇/蛇毒,所要/要所,父君/君父,害虫/虫害,道中/中道,日中/中日,水上/上水,来朝/朝来,商会/会商,年末/末年,著大/大著,学政/政学,同党/党同,応対/対応,大老/老大,砲火/火砲,制服/服制,字音/音字,限定/定限,分権/権分,論定/定論,別種/種別,明月/月明,名山/山名,船客/客船,人前/前人,手工/工手,主教/教主,城外/外城,気鋭/鋭気,老中/中老,中腹/腹中,本院/院本,国君/君国,柱石/石柱,形象/象形,才学/学才,名文/文名,機転/転機,理非/非理,道士/士道,天則/則天,高座/座高,部面/面部,逸散/散逸,番茶/茶番,編中/中編,衆多/多衆,席次/次席,林学/学林,皇上/上皇,軍中/中軍,公主/主公,火熱/熱火,風通/通風,所好/好所,稿本/本稿,民生/生民,霊山/山霊,経蔵/蔵経,流俗/俗流,超出/出超,府城/城府

不明(8 対)　大国/国大,振作/作振,満天/天満,想美/美想,郡市/市郡,同公/公同,兵部/部兵,中小/小中

４．考察

４.１　意味関係と頻度・使用実態変化との関連性

本研究の意図は，近代における漢語語彙変化（主に漢語語彙の縮小，表記や語法の画一化）の方向性が，「言語運用の合理化」「語と意味の１対１対応」であることを補強する事例として，字順転倒漢語対の減少の実態を捉えることである。この方向性から見ると，意味の重なりの度合いが高い対では，どちらかの語形（多くは勢力の弱い語形）が淘汰され，そうでない対は併存することになるはずである。意味関係と，近代での頻度・現代での使用実態変化との間の関係性についてまとめれば，以下のようになる。

表 9　意味関係と頻度・使用実態変化における関係性

	理想的	想定外
Ⅰ	B 消失, AB 低頻度消失	AB 併存, A 中高頻度消失
Ⅱ(1)	AB 併存, AB 低頻度消失	A･B 中高頻度消失
Ⅱ(2)	AB 併存, AB 低頻度消失	A･B 中高頻度消失
Ⅲ	AB 併存, AB 低頻度消失	A･B 中高頻度消失

3 章で見た調査結果からは，これに当てはまる理想的な組合せが多く見られるものの，実際には想定外の組合せも少なからず生じている。

そこで，ここではその要因を探るべく，近代・現代における個々の用例を分析し，用法の実態を確認してみたい。

まず，辞書的には意味の同一性が高いと思われる対にもかかわらず，現代においても併存している対について見てみよう。

４.２　衰退の過渡期にあると思われる漢語対

意味関係がⅠにもかかわらず A 語 B 語の両語が併存している語には，近代でも使用頻度の差があり B 語が少なく，現代において更に使用頻度を減らして極めて低頻度でのみ用いられる語がある。「統一」「一統」,「苦痛」「痛苦」,「治療」「療治」「苦労」「労苦」「練習」「習練」などがこれに当たる。

「苦痛」「痛苦」を例に検討する。両語の使用状況を，近代については出版年による年次別，現代についてはレジスター別に集計すると以下のようになる。

表 10　「苦痛」「痛苦」の使用実態

	近代							現代		
	1875	1887	1895	1901	1909	1917	1925	書籍	雑誌	新聞
苦痛	1	45	83	79	127	109	90	204	11	12
痛苦		5	23	10	8	3	5	3		

「痛苦」は，1887 年から用例が見られ 1895 年にやや使用割合が多くなるが，近代においては，いずれの年においても優勢語形「苦痛」が圧倒的に高頻度である。また，現代では書籍に以下 3 例が見られるのみであり，文学的ジャンルかつ現代においては生年の早い著者による使用である。ここから，「苦痛」「痛苦」の字順転倒漢語対は，優勢語形「苦痛」にほぼ統一され，「痛苦」は今後，生年の早い著者による使用が見られなくなる段階で，淘汰されるものと思われ，衰退の過渡期であると位置づけることができる。

(1) 自分が打ちのめされる痛苦をあじわいながら学んでゆくことで成立します。
BCCWJ・PB19_00305,大江健三郎(1930 年代生)『鎖国してはならない』2001 年，9 文学
(2) 本物の指令体が、お互を痛苦と悦楽の階段の極点へ、一気に昇らせる。
BCCWJ・PB19_00316,田久保英夫(1920 年代生)『仮装』2001 年，9 文学
(3) 黒田自身の痛苦な反省にふまえて語りかけられていることを実感させられた。
BCCWJ・PB52_00283,吉川文夫(1930 年代生)『今のぼくは二十七歳』2005 年，2 歴史

なお，近代では優勢であった A 語が現代では衰退し，B 語が優勢になっている「争闘」「闘争」，「熱情」「情熱」，「争論」「論争」，「社寺」「寺社」，「戦敗」「敗戦」，「成育」「育成」などの漢字対もあるが，これらも現代で劣勢の A 語が衰退する過渡期であり，いずれ B 語にほぼ集約されていくものと考えられる。

４.３　意味や用法に分化が見られる漢語対

同様に，意味関係が I にもかかわらず A 語 B 語の両語が併存している語には，現代で両語がいずれも勢力を保って用いられている漢語対も多く見られ，「国民」「民国」，「平和」「和平」，「議論」「論議」，「運命」「命運」，「便利」「利便」，「国内」「内国」，「評論」「論評」，「途中」「中途」，「製作」「作製」，「分配」「配分」，「木材」「材木」，「関連」「連関」，「継承」「承継」，「該当」「当該」などの語がこれに当たる。

「便利」と「利便」を例に検討する。近代においては，例 4，5 に見るように，ほぼ同様の意味で用いられている「便利」「利便」は，現代では，A 語「便利」がほぼ形容詞用法に固定化され用いられ，一方の B 語「利便」は「利便性」の形に固定化して用いられている。

(4) 次に鑛山、鐵道等に及ぼし頻りに其便利なるを主張すれども
CHJ・60M 太陽 1895_04044,『太陽』1895 年 4 号
(5) 其後各地に鐵道開通するに從つて、其利便なるを認識し、
CHJ・60M 太陽 1925_05067,『太陽』1925 年 5 号
(6) 新幹線ができて、所要時間も半分に短縮され、便利になったとは聞いていたが、
BCCWJ 出版・書籍・PB50_00035，高石きづた『マーガレットの花』2005 年
(7) 線路などの維持費が安い路面電車に転換し、利便性を上げて存続することを決めた。
BCCWJ 出版・新聞・PN5a_00007，朝日新聞 2005 年 2 月 4 日朝刊

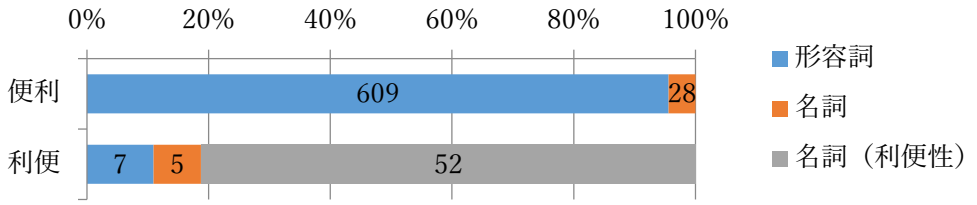


図 2　「便利」「利便」の現代における用法分布

これらの語の多くは，近代においてほとんど差異がなく用いることができたものが，現代においては用法（連語，語法，意味，位相等）が分化することで，共に生き残った対のケースと捉えることができる。

５．まとめ

近代に特異に多く見られた字順転倒漢語対は，現代では半分程度が併存状態を保っているが，残りは淘汰されていることが分かった。対となる漢語が併存するか，どちらか一方を残し他方が淘汰されるかは，両語の意味的関係性による部分が大きく，類義関係または異議の関係にある対では 60%程度で併存状態が保たれているものの，同義関係にある対は40%程度が併存し，40%は一方が淘汰されていた。

これまで発表者が研究を行ってきた，漢語の語彙や表記・語法における近現代間の通時的変化の方向性は，概ね「１義１語」を志向した漢語語彙の淘汰収束であると見られたが，同義関係にある別語形の併存が 40%程度見られる点は，これに反するように思われる。しかし，個々の語の使用実態を見ると，現代語で同義異形語が存在しているように見えるものも，淘汰収束の過渡期として残る両語併存，あるいは，用法の分化による類義・異義異形語であることを示唆することができた。

今回個々の語の変遷を取り上げることができなかった，同義（あるいは類義・異義）の漢語対が淘汰される過程についても，今後明らかにし，これまで発表者が明らかにしてきた異表記の淘汰や語法の画一化の時期と，字順転倒漢語対の淘汰の時期との関連についても，更に研究を進めたい。

謝　辞

本研究は国立国語研究所のプロジェクト「通時コーパスの構築と日本語史研究の新展開」による成果の一部である。

文　献

浅野敏彦(1997)『国語史のなかの漢語』和泉書院
荒川清秀(1997)『近代日中学術用語の形成と伝播: 地理学用語を中心に』白帝社
池上禎造(1984)『漢語研究の構想』岩波書店
今野真二(2012)『百年前の日本語―書きことばが揺れた時代 (岩波新書)』岩波書店
鈴木丹士郎(1979)「二字漢語の字序について」押見虎三二教授御退官記念事業会編『国語表現論叢（押見虎三二教授御退官記念論集）』明治図書出版，pp.239-245
鈴木丹士郎(1981)「「抵抗」と「抗抵」」国語語彙史研究会編『国語語彙史の研究　二』和泉

書院，pp.237-254
鈴木丹士郎(1986)「二字漢語の字順についての問題」『国語論究１　語彙の研究』明治書院，pp.278-300
高野繁男(2004)『近代漢語の研究: 日本語の造語法・訳語法』明治書院
田島優(1998a)『近代漢字表記語の研究』和泉書院
田島優(1998b)「字順の相反する二字漢語」『近代漢字表記語の研究』和泉書院，pp.316-339
田島優(1998c)「『新説　八十日間世界一周』における字順の相反する二字漢語」『近代漢字表記語の研究』和泉書院，pp.340-374
田中牧郎 (2016) 「第 8 章　語種」斎藤倫明編『日本語語彙論 I （講座　言語研究の革新と継承１）』ひつじ書房，pp.241-274
間淵洋子(2016a)「近現代漢語におけるサ変動詞用法の変化: 形態論情報付きコーパスを用いて」『国際日本学研究論集』4，pp.17-35
間淵洋子(2016b)「近代二字漢語における同語異表記の実態と変化: 形態論情報付きコーパスを用いて」『計量国語学』30(6) ，pp.257-274
間淵洋子(2017a)「近代雑誌コーパスにおける漢語語彙の特徴: BCCWJ との比較から」『国立国語研究所論集』13，pp.143-166
間淵洋子(2017b)「近代漢語の品詞性に見る多様性の画一化: 形容詞用法を中心に」『言語資源活用ワークショップ発表論文集』2，pp.93-106
間淵洋子(2018)『近代漢語における表記・語法の多様性とその変化に関する計量的研究: 現代語確立期にみる言語変化の様相と背景』明治大学大学院国際日本学研究科 2017 年度学位請求論文
吉川明日香(2005)「字順の相反する二字漢語: 「掠奪―奪掠」「現出―出現」について」国立国語研究所編『雑誌『太陽』による確立期現代語の研究: 『太陽コーパス』研究論文集』博文館新社，pp.143-155

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/
（CHJ : 中納言 2.4.2 データバージョン 2018.03, BCCWJ : 中納言 2.4 データバージョン 1.1）

実践医療用語の語構成要素抽出の試み

内山 清子（湘南工科大学工学部コンピュータ応用学科）†
岡照晃（国立国語研究所コーパス開発センター）
東条佳奈（目白大学社会学部社会情報学科）
小野正子（西南女学院大学保健福祉学部）
山崎誠（国立国語研究所言語資源研究系）
相良かおる（西南女学院大学保健福祉学部）

Extracting of Word Constituents contained in Medical Terms

Kiyoko Uchiyama (Dept. of Applied Computer Sciences, Sonan Institute of Technology)
Teruaki Oka (Dept. Corpus Studies, NINJAL)
Kana Tojo(Faculty of Studies on Contemporary Society, Mejiro University)
Masako Ono(Faculty of Health and Welfare, Seinan jo Gakuin University)
Makoto Yamazaki (Dept. Corpus Studies, NINJAL)
Sagara Kaoru (Faculty of Health and Welfare, Seinan jo Gakuin University)

要旨

医療現場で用いられる電子カルテなどの記録文書（医療記録）に専門用語としての医療用語が大量に含まれている。医療記録に記載された言語情報を正確に理解・活用するためにはこれらの医療用語の理解が必要となる。

医療記録に含まれる語には、複数の語からなる複合語や臨時一語も多く、これらは、病名、身体の部位名、処置名、薬剤名等、様々な用語から構成されている。しかし、現在はこの語構成要素の組み合わせのパターンや語構成要素間の関係などが曖昧である。

そこで、本研究では複数の語からなる実践医療用語の語構成要素の抽出を試みた。語構成要素の条件を独自で定義した後、ComJisyoV5、と今後公開予定の V6 の登録候補語に対象として、MecabMeCab 0.996 と UniDic-cwj-2.2.0 を利用して形態素解析を行った。分割された単語の品詞情報を手がかりにして、単一単位となり得る品詞列を抽出した。次に抽出した候補リスト以外に語構成要素となる品詞列があるかについて検討を行った。

１．はじめに

医療記録には、専門用語に加え、略語や隠語など独特な表現が含まれる。この記録データを電子カルテシステムの普及により、施設内での共有や、大量の医療記録データを二次利用する研究なども増加してきている。しかし、実際に医療記録に記載された言語情報を正確に理解・活用するためには、専門用語を含む複合語や臨時一語などの用語を適切に抽出することが不可欠である。本研究では、複数の語からなる実践医療用語に特化した形態素解析辞書の構築を目的とし、実践医療用語を構成している語構成要素を抽出することを試みた。

† uchiyama@sc.shonan-it.ac.jp

日本語学において、複合語の構造を明らかにする「語構成論」について、斎藤(1996)は語構成要素から語が成立するまでの一連のプロセスの内容を明らかにする研究分野として定義している。斎藤(2004)では、従来の語構成論では複合語が語構成論の対象となっているのに対して、単純語も複合語も同じく語構成論の対象となるという立場に立っている。また、単語化、語構成要素、語構成要素間の関係の 3 つの観点から語構成論を論じている。

本研究における語構成要素とは、専門用語を理解する上で分割できる最小単位の語に相当するものであると考えている。つまりすでに辞書に登録されている 2 文字や 3 文字の名詞や動詞などの単語だけでなく、複数の単語や形態素が結合した単位も含んでいる。この語構成要素を確定することで、語構成要素の意味が理解できていれば複数の語構成要素からなる臨時一語や複合語などの専門用語を理解するための支援になる。今回は従来の形態素解析で処理した結果をどの程度活用できるかを検討するために分析を行う。以下、2 章において分析対象としたデータについての説明を行い。3 章に抽出方法、4 章に抽出結果を述べ、その後考察と今後の課題について記述する。

２．対象データ（ComJisyoV5）

医療従事者用の臨床記録文書（看護記録、プログレスノート、医療経過記録など）を解析するための支援として 2008 年から形態素解析辞書 ComeJisyo を作成し、2013 年 11 月に ComeJisyoV5（登録語数 77,760 語）を公開している。本研究では、文字長が 2 文字以上の用語を対象とし、非公開の研究用見出し語データ 52,974 語と ComeJisyo V5-1 および公開予定の ComeJisyoV6 の登録語の併せて 109,721 語で重複している 31,162 語を抽出し、本研究における語構成要素候補とした。この候補は 2 文字から 22 文字から成る医療用語が含まれているため、語構成要素と複数の語構成要素を組み合わせた複合語になっている。この用語のうち、複合語がどのような語構成要素から成り立っているのかを調べるために、まず語構成要素自体を独自に定義し、抽出することが必要となる。

３．抽出方法

まず、語構成要素の候補となりうる条件を設定した。語構成要素の候補とならないものとして臨時一語を含む複合語の認定条件をあげた上で、その条件に該当しない、かつ機械的に抽出することが可能な条件を考えた。

石井(2007)によると、「臨時一語の認定条件」は、1)複数の単語が臨時的に結びついたもの、2)複合語、3)もとの単語列に復元することができるものとしている。反対に、臨時一語と判定されないものは、①固有名、②組織名・役職名、③ときの表現、④地名、⑤数量に関する表現としている[1]。このように臨時一語の認定条件に該当しないものが語構成要素の候補になると定義し、形態素解析結果から品詞を指定して抜き出す条件を考えた。

語構成要素候補リスト 31,162 語に含まれる用語を対象として MecabMeCab 0.996 と UniDic-cwj-2.2.0 で形態素解析を行い、品詞を付与した。

表 1 に形態素結果語の語構成要素リストの構成要素数を示す。

表 1 構成要素数

構成要素数	該当単語数
1	4062
2	8927
3	7789
4	5637
5	2878
6	1178
7	453
8	168
9	51
10	14
11	4
12	1

表 1 の通り、構成要素が 1 つからなる用語は 4062 語含まれており、この品詞は記号が 1 語と形状詞が 12 語、後は全て名詞となっていた。このことから形状詞や名詞は単独で語構成要素の条件として設定することが妥当であると判断した。そこで、「臨時一語の認定条件」や臨時一語と判定されないものを参考にしながら、語構成要素候補として以下の品詞列の条件を設定した。

①品詞列の条件
単一語 :「名詞」または「形状詞」のもの
二語以上 :「名詞」＋接尾辞」、「形状詞」＋「接尾辞」の並びになっているもの、「記号のみの組み合わせ」
「名詞」＋「名詞」の 2 文字
②①以外で文字数 3 文字以下のもの

このように①と②に該当する用語を抽出し、語構成要素として認定可能かどうかを分析していく。

４．抽出結果

3 章の抽出ルールに従って語構成要素候補を抽出した結果を表 2 に示す。

表 2　抽出ルール別該当数

	抽出条件	13,849
①	名詞	4062
	形状詞	12
	名詞｜接尾辞	1418
	形状詞｜接尾辞	16
	名詞｜名詞　の 2 文字	74
	記号のみ	188
②	3 文字以下	1299
	条件①と②の合計	7327

表 2 に示した通り、構成要素の品詞として名詞がもっとも多く含まれていた。また、「名詞＋接尾辞」の組み合わせも多く出現していたことや、記号の連続も専門用語の一部を構成しているものがあるなど、条件として①は適切であったと考えられる。

次に①の条件を満たさない 3 文字以下の単語であるが、①の条件を除いた後に分析すると名詞、形状詞、記号を含む用語が全て対象外となってしまう。残った単語を見てみると感動詞、助詞、助動詞、動詞、副詞など解析誤りを含むものが多かった。そこで、再度 2 文字以上 3 文字以下の用語には元々どのような品詞列が含まれているかを調べた。

表３　文字数 3 以下の用語の品詞列

品詞列	用語数	品詞列	用語数
名詞/名詞	1012	記号/接尾辞	16
名詞/名詞/接尾辞	55	名詞/記号	11
名詞/接尾辞/接尾辞	34	記号/名詞	10
接頭辞/名詞/接尾辞	30	名詞/名詞/名詞	10
名詞/接尾辞/名詞	29	形状詞/名詞	8

表 2 や 3 で分かる通り、「名詞＋接尾辞」の品詞列がもっとも多いことから抽出ルールにこの条件を入れた妥当性が確認できる。文字数 3 文字以下という条件をつけると 2 文字で「名詞＋名詞」や「接頭辞＋名詞」の用語は一つの語構成要素とすべきものを多く含んでいる。3 文字で「一側性」などの「名詞＋名詞＋接尾辞」、「一次的」「一次性」などの「名詞＋接尾辞＋接尾辞」、「両価性」などの「接頭辞＋名詞＋接尾辞」といった品詞列はこれも一つの語構成要素と認定しても良いものである。このように 3 文字以下の用語から長単位の品詞列をまず確定させてから、残りの用語について分析をするのが適切な手順であったのではないかと考察する。

同様の問題として、名詞の中には一文字単語である「群」「型」などが 888 語も含まれていたが、これらを単独で語構成要素とすることは難しい。

５．考察

ここまで、医療用語の語構成要素を選定するにあたって、まず品詞列をてがかりとした条件を設定して抽出を試みた。その結果として「名詞＋接尾辞」「形状詞＋接尾辞」が語構成要素候補としてもっとも適した条件であった。一方で、文字数と品詞列の対応をさせることによって、条件を細かく設定すればより効率的に抽出できるのではないかと考えられる。その中で、抽出条件の優先順位も考慮する必要がある。たとえば、まず長単位の品詞列を適応させ、その語構成要素を一単位と設定し、次に大雑把な品詞列の条件に当てはまるものを抽出していく手順を考えていくべきである。

また、分析している中で、出現頻度や連接確率などを考慮にすればより効率的な抽出が可能になるのではないかという結論に達した。各語構成要素が用語全体のどの位置で出現しやすいかという分析も試みたが、語構成要素が確定していない中で分析することが難しく、統計的にまとめることができなかった。しかし、確実に用語の頭や終わりに出現する語構成要素や、複数の語構成要素が組み合わさって長い用語を構成する場合には、その順番もある程度規則性があることなどが見て取れた。

６．まとめと今後の課題

本研究は、医療現場で作成される医療記録に記載された言語情報を正確に理解・活用するために必要となる、専門用語を含む複合語や臨時一語などの用語を構成している語構成

要素を適切に抽出することを試みた。形態素解析結果の品詞列を手がかりとして、「名詞＋接尾辞」「形状詞＋接尾辞」「名詞」単独、「形状詞」単独、複数の「記号」列からなる用語を分析し、語構成要素と認定するに適切な条件であることを確認した。

今回は、抽出条件を中心に検討していたため、手順や、条件の優先順位を整理することができなかった。今後は抽出条件と手順を明確にし、更に語構成要素間の連接頻度や結合関係などを分析していきたい。

謝　辞

本研究は JSPS 科研費(18H03499)の助成を受けたものです。

文　献

石井正彦(2007)，現代日本語の複合語形成論，ひつじ研究叢書，ひつじ書房.

斎藤倫明(1996)，現代日本語の語構成論的研究−語における形と意味−，日本語研究叢書，ひつじ書房.

斎藤倫明(2004)，語彙論的語構成論，ひつじ研究叢書，ひつじ書房.

日本語の非流ちょう性
－とぎれと延伸の数量調査から－

佐々木 藍子（国立国語研究所日本語教育研究領域
/東京学芸大学大学院連合学校教育学研究科）†
砂川 有里子（国立国語研究所日本語教育研究領域客員教授）
浅原 正幸（国立国語研究所コーパス開発センター）

Disfluency of Japanese
－Quantity survey of unfilled pause and prolongation－

Aiko Sasaki（National Institute for Japanese Language and Linguistics/
Doctoral Course The United Graduate School of Education Tokyo Gakugei University）
Yuriko Sunakawa（National Institute for Japanese Language and Linguistics）
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

これまで日本語教育では、非流ちょうな発話の指導はほとんど行われてきていない。しかし、実際には日本語母語話者であってもよどみのない流ちょうな発話を行うことはまれであるし、非流ちょうな発話が話し手のストラテジーとして用いられることや、聞き手の理解の促進につながることもある。そこで本研究では、非流ちょう性の要因となる「とぎれ」と「延伸」を取りあげ、「多言語母語の日本語学習者横断コーパス（I-JAS）」に収録されている日本語母語話者データの数量調査を行った。その結果、ストーリーテリング（ST1・ST2）とロールプレイ（RP1・RP2）において、とぎれと延伸ではとぎれのほうが多いが、頻度に男女差がないこと、ST1 と ST2 の間、RP1 と RP2 の間のとぎれと延伸の生起の仕方に差がないこと、および、ストーリーテリング（ST１・ST２）とロールプレイ（RP１・RP2）のタスク間においてとぎれと延伸の生起の仕方に大きな違いがあることが分かった。

１．はじめに

近年、日本語教育の現場でもコミュニケーション能力の向上が重視され、指導法も変わりつつある。しかし、学習者の日本語発話は聞き取りにくいことが往々にしてある。聞き取りにくさの要因には、文法の間違いや発音の悪さなども関係するが、それだけではないと考えられる。

一般的に流ちょうだと感じられる日本語母語話者の発話の中にも非流ちょうな要素が出現することも少なくない。また、よどみなくすらすらと話されると、かえって聞き取りにくさを感じることさえある。そのため、ある程度の非流ちょうさは、何らかの機能を持っているのではないかと思われる。以下の例は、同じタスクを行った日本語母語話者および日本語

† a.sasaki@ninjal.ac.jp

学習者の発話をできる限り忠実に文字化したものである[1]。例１は日本語母語話者の発話で、例２はオーストラリア人日本語学習者の発話である。

（例 1）日本語母語話者

JJJ02：あのアルバイトのシフトなんですけどー、〈はい〉今週三日、〈うんうん〉あのー入れさしてもらってるんですけど、でもできれば、〈ええ、ええ〉それをこれから〈ええ〉週二日にー、変更していただきたいんですけど

（例 2）オーストラリア人日本語学習者

EAU18：あのー、実はねーあの〈うん〉、最近ーなんか、大学がーすごく一忙しくなりましたけど〈うん〉、あのー今はー、たっみつか、三日間のーバイトをしていますがー〈うーん〉、それはーあの、二日間にー、あの変えてもよろしいですっか？

日本語母語話者の発話にも、読点部分が表すように頻繁に発話がとぎれていたり、長音記号で表されるように母音が伸びていたりというような現象が見られるが、日本語学習者の発話では日本語母語話者より、とぎれている部分や引き伸ばしている部分が多いことが分かる。これらのとぎれている部分や母音を引き伸ばしているものは、非流ちょう性のつっかえの一部であり、「とぎれ」、「延伸」と呼ばれている（定延；2016）。日本語母語話者の非流ちょう性のパターンに日本語学習者の非流ちょう性を近づけることができれば、聞き取りにくさが減少し、聞き手に負担をかけない話し方ができるようになるのではないかと思われる。

これまでの非流ちょう性の研究では、とぎれや延伸に関する研究はあまり多くない。その中で、定延（近刊）では、非流ちょうな発話の許容可能性が当該言語の膠着性の高さという文法的な概念に影響される可能性を示している。また、砂川・佐々木（2016）では、I-JASの日本語学習者と日本語母語話者の発話データを分析し、以下のような結果を提示している。ハンガリー語を母語とする学習者はとぎれが多く、韓国語を母語とする学習者は延伸が多い。また、スペイン語母語の学習者のとぎれと延伸の使用頻度は日本語母語話者とほぼ同程度で、ハンガリー語、韓国語を母語とする学習者より少ないが、延伸の方がとぎれより多いことが、日本語母語話者と異なっている。これらの結果からはそれぞれの母語におけるとぎれと延伸の使い方が影響していることが推察され、言語によってとぎれと延伸に異なる選考性が存在している可能性が考えられる。しかし、日本語母語話者の非流ちょう性については、まだ明らかになっていない部分も多い。

そこで、本研究では日本語学習者のコミュニケーション能力養成のための基礎研究として、日本語母語話者の非流ちょう性の実態を探ることを目的とする。具体的には日本語母語話者の発話のデータを用い、（1）とぎれと延伸のどちらが多いか、（2）とぎれと延伸に男女差があるか、（3）タスクの違いによってとぎれと延伸の頻度に差があるか、という 3 つの観点について分析を行い、日本語母語話者の非流ちょう性の一端を明らかにする。

[1] 例文の〈　〉内の発話は対話者のあいづちである。

２．分析

２.１　対象データ

分析対象のデータは、国立国語研究所で公開されている『多言語母語の日本語学習者横断コーパス International corpus of Japanese as a second language』（以下、I-JAS）を使用した。I-JAS は主に 12 の言語それぞれを母語とする日本語学習者の話し言葉と書き言葉のデータが収録されたコーパスであるが、比較コーパスとして日本語母語話者のデータも収録されている。I-JAS は現在も増補中で、完成すると国内外の日本語学習者 1,000 名分、日本語母語話者 50 名分のデータが収録された大規模な学習者コーパスとなる。現在は第三次公開データにより、日本語学習者 610 名分、日本語母語話者 50 名分のデータが公開されている。I-JAS の特徴の 1 つとして多様なタスクのデータが収録されている点が挙げられるが、本研究では第一次および第二次公開データの日本語母語話者 50 名のストーリーテリング（以下、ST）とロールプレイ（以下、RP）を分析に使用した。ST は 4 コマと 5 コマの 2 種類（ST1、ST2）のコマ割漫画を見ながら、調査者に対してそのストーリーを語る独話である。そして、RP は指定された役を提示された内容に沿って演じながら、調査者と対話するタスクである。RP も 2 種類（RP1、RP2）あり、1 つは日本料理店の店長に週 3 日のアルバイトを週 2 に変えてもらうよう依頼するタスク、もう 1 つは日本料理店の店長から、ホールの仕事から調理の仕事に変更することを依頼されるが、それを断るタスクである[2]。

２.２　分析方法

I-JAS の言語データは、コーパス検索アプリケーション『中納言』が使用できるよう、自動の形態素解析が行われ、一部の誤解析やミスについては人手での修正も行われている。形態素解析器は正しい書き言葉に対して処理を行うよう設計されているため、言い淀み、フィラー、無意味語などを含む学習者の発話データでは特に誤解析が生じやすい。そこで、それらの誤解析を防ぐため、I-JAS の言語データには 9 種類のタグが付与されており、『中納言』ではタグによる検索も可能である[3]。

本研究では、タグの 1 つである T タグを分析対象の 1 つとした。T タグとは、自動の形態素解析を行う上で解析の妨げとなる語中の読点や長音記号を削除したり、逆に読点がないため、誤解析を誘発するような箇所に読点を付与したりするための形態素解析用のタグである。この T タグに加え、T タグの付与されていない読点と長音記号を分析対象とし、『I-JAS 中納言』バージョン 2.2.3 短単位データ 20170519 版を使用して検索した。検索したデータのうち、読点がないため誤解析を誘発しそうな箇所に読点を付与する場合の T タグは、検索したデータから除外した。また、あいづちの後にある読点、文末だと思われる箇所や応答詞の後の読点、フィラーの長音も除外し、とぎれの読点と延伸の長音記号に分類し、数量調査を行った。

日本語母語話者の ST と RP について、とぎれと延伸の出現数を集計し、それぞれのデータの総語数に対してとぎれと延伸の出現した割合を算出し、その割合をもとに t 検定および一般化線形モデルを用いた回帰分析を行った。

回帰式は以下の通りである。

rate ~ age + gender + task + feature + (1|subj)

[2] I-JAS について詳しくは迫田編（2016）を参照されたい。
[3] タグについて詳しくは迫田他（2016）を参照されたい。

ここで rate はとぎれや延伸の生起確率で推定する値である。age は年齢、gender は性別（男、女）、task はタスク（RP1, RP2, ST1, ST2）、feature はとぎれか延伸かで、以上を固定要因とした。subj は実験協力者で、ランダム要因とした。

３．結果

３.１　とぎれと延伸の頻度

タスク別および男女別にとぎれと延伸の出現頻度の割合を産出し、図１の箱ひげ図を作成した。また、統計的な有意差の有無を確認するため、一般化線形モデルでも確認した。以下図１の箱ひげ図は、データのばらつきを反映したものであるが、ST においても RP においても、とぎれが延伸に比べてかなり多いことがわかる。また、表１の一般化線形モデルの結果においても、とぎれが延伸より、6.9%多い（p<0.01）ことが分かった。

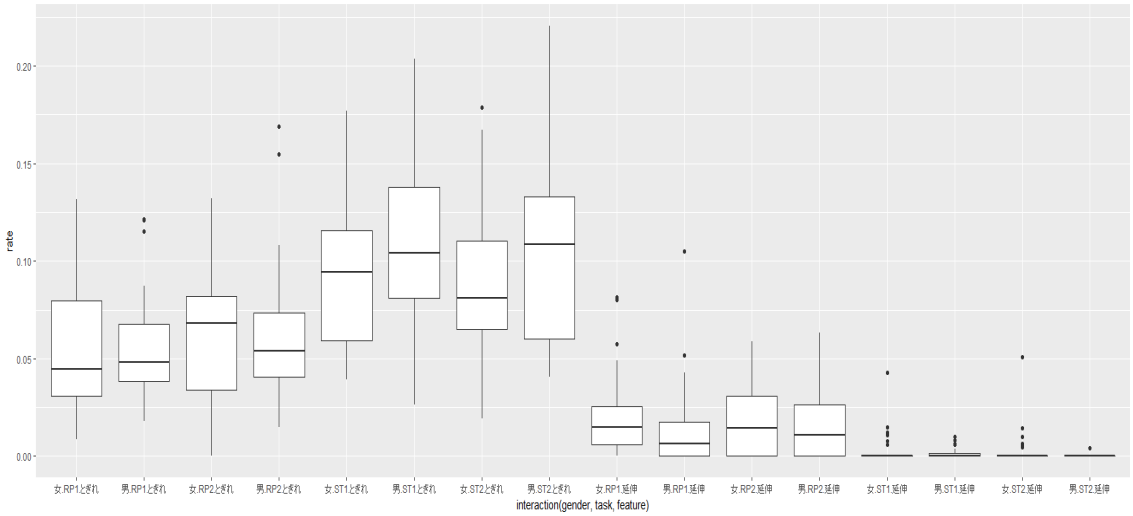


図 1. タスク別のとぎれと延伸の箱ひげ図

表１．とぎれと延伸の割合による一般化線形モデル

変数名			推定値	標準誤差
年齢	1 歳ごと	age	0.0004**	0.0002
性別：男	vs 女	gender 男	0.003	0.004
タスク:RP2	vs RP1	taskRP2	0.002	0.004
タスク:ST1	vs RP1	taskST1	0.014***	0.004
タスク:ST2	vs RP1	taskST2	0.011**	0.004
延伸	vs とぎれ	feature 延伸	-0.069***	0.003
切片			0.056***	0.008

観測数	400 事例
Log Likelihood	766.802
AIC	-1,515.603

Note:

*p<0.1

**p<0.05

***p<0.01

３.２　とぎれと延伸の男女差

次にとぎれと延伸の使用頻度に男女の差が統計的に有意かを確かめるため、有意水準 5%で Welch のｔ検定を行ったところ、とぎれにおいては、ST1 が $t(43)=1.37$, $p=0.18$、ST2 が $t(43)=1.23$, $p=0.22$、RP1 が $t(48)=0.04$, $p=0.97$、RP2 が $t(42)=0.25$, $p=0.81$ であり、有意な差は見られなかった。また、延伸においては、ST1 が $t(33)=-1.03$, $p=0.31$、ST2 が $t(26)=-1.66$, $p=0.11$、RP1 が $t(45)=-0.84$, $p=0.41$、RP2 が $t(47)=-0.40$, $p=0.69$ であり、こちらでも有意差は見られなかった。

３.３　タスクの違いによるとぎれと延伸の頻度差

最後に、タスクによるとぎれと延伸の頻度差の違いを明らかにする前に、同じ種類のタスクである ST 1 と ST2、RP1 と RP2 の間で有意差がないか確認した。これは同じ種類のタスクでも、内容によって違いがないかを確認するものであった。有意水準 5%で Welch のｔ検定を行ったところ、ST1 と ST2 におけるとぎれは $t(97)=0.51$, $p=0.61$、延伸は $t(97)0.49$, $p=0.62$ で、RP1 と RP2 におけるとぎれは $t(97)=0.78$ $p=0.43$、延伸は $t(91)=0.41$, $p=0.68$ であり、有意な差は見られなかった。そのため、ST1 と ST2、RP1 と RP2 においては、とぎれ、延伸共に同じタスク間では頻度にないことが明らかとなった。そこで、タスクの違いによって、とぎれと延伸に有意な差があるかを確認したところ、差があることが分かった。表 1 を見ると、全体としてはとぎれが 6.9%（$p<0.01$）多いことが分かるが、図 1 から分かるように、ST にはとぎれが多く、RP には延伸が多いことが分かった。

４．考察

本研究では、日本語母語話者の非流ちょう性の実態を探ることを目的に 3 つの観点で分析を行った。以下、それぞれの観点ごとに考察する。

(1) とぎれと延伸はどちらが多いか

とぎれは延伸に比べ、有意に多かった。とぎれは非流ちょうなものばかりでなく、発話を流ちょうに進めるためにも必要である。延伸も同様に、非流ちょうな延伸とそうでない延伸が考えられる。本来ならこれらを峻別して、非流ちょうなとぎれと延伸のみを対象とした調査を行う必要があるが、その区別は容易ではない。そのため、今回の調査ではその区別を行わず、流ちょうなとぎれや延伸も含んだ結果を示している。今後はこれらの扱い方も検討していきたい。

(2) とぎれと延伸に男女差があるか

とぎれと延伸については、性別の違いでは有意な差は見られなかった。分析対象の男女の内訳は男性が 23 名、女性が 27 名であった。人数に差はあるが、統計処理を行う際、データごとに発話の総形態素数に対するとぎれと延伸の割合を産出しているため、人数の違いは問題にならない。そのため、とぎれと延伸の使用傾向は男女ともに同じように使用していると考えられる。

(3) タスクの違いによってとぎれと延伸の頻度に差があるか

同じ種類のタスクである ST1 と ST2、RP1 と RP2 のとぎれと延伸の頻度については、有意な差が見らなかった。つまり、ST1 と ST2 のとぎれと延伸の頻度差は見られず、

RP1 と RP2 のとぎれと延伸の頻度差も見られなかった。そのため、ST1 と ST2、RP1 と RP2 をタスクごとにまとめ、ST と RP で比較してみたところ、図 1 から分かるように、ST にはとぎれが多く、RP には延伸が多かった。これは、タスクのタイプの違いによるものだと考えられる。ST は独話であり、RP は対話である。独話は完成された文が使用されることが多いのに対し、対話は中途終了文など、文としての完成を見ない形式が多く使用される。そのため、タスク別のとぎれと延伸の頻度の差は、この影響を受けている可能性が高い。しかし、その他にも、次に述べるようにとぎれと延伸の発話機能の異なりも影響している可能性があるのではないかと考える。

上記 3 つの観点からの分析結果をもとに考察する。ST は、提示された 4 コマと 5 コマのコマ割り漫画のストーリーを調査者に対して話すものである。このタスクはただイラストを見て話せばよいため、一見簡単そうに見える。接続助詞の「て」を多用すれば、1 文で話すことも可能であるが、それだと冗長になり聞き手にとってわかりにくいものになってしまう。このタスクを行う際には、どのような語彙を使用し、どこで文を区切り、きちんと順を追ってストーリー展開を組み立て、最後には結末をどうつけるなど、さまざまなことを考えながら話す必要があるため、とぎれが多くなったのではないかと考える。

一方、RP は依頼や断りをする場面であるが、自身の要望や意向などを相手との関係性や自身の立場を考慮して話さなければならない。さらに、自身の発話に対する相手の反応を伺いつつ、次の発話を考え、自分の要望や意向を伝えていく必要がある。そのため、話し相手に言いにくいことを言う場合や話し相手からの要望を断る場合に延伸を使うことで、ポライトネスを維持しているのではないかと考える。このように非流ちょうな発話はときとしてコミュニケーション方略として用いられていることが考えられる（Lickley(2001)、Bortfeld, et. al.(2001)、伝・渡邉（2009）など）。

５．おわりに

本研究は、日本語学習者のコミュニケーション能力養成のための第一歩として、日本語母語話者の非流ちょう性の一端を探った。そして、ST と RP において、とぎれと延伸ではとぎれのほうが多いが、頻度に男女差がないこと、ST にはとぎれが、RP には延伸が多く使用され、生起の仕方に大きな違いがあることが分かった。

今後はこの結果をどのように指導に活かすかという方法を検討していく必要がある。そのため、まず日本語学習者の非流ちょう性の習得状況を確認したい。その際、日本語能力や母語の違いによって違いがあるか、違いがあるとすればどのような違いがあるかという点にも着目し、分析を進めていきたい。

謝　辞

本研究は国立国語研究所のプロジェクト「多文化共生社会における日本語教育研究」および科研費基盤(A)「海外連携による日本語学習者コーパスの構築－研究と構築の有機的な繋がりに基づいて－」による成果『I-JAS』を利用して行われたものである。

文　献

迫田久美子編（2016）『海外連携による日本語学習者コーパスの構築－研究と構築の有機的な繋がりに基づいて－I-JAS 構築に関する最終報告書（International Corpus of Japanese as

a Second Language)』科研費研究報告書 24251010 代表者迫田久美子
迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016）「NINJAL-多言語母語の日本語学習者横断コーパス International Corpus of Japanese as a Second Language」『国語研プロジェクトレビュー』第 6 巻 3 号, pp. 93-110.
定延利之（2016）「４つの発話モード」庵功雄・佐藤琢三・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.
定延利之（近刊）「言語類型からみた非流ちょう性―膠着語と延伸型続行方式のつっかえ」『社会言語科学会』21-1
砂川有里子・佐々木藍子（2016）「I-JAS を使った非流ちょう性の研究」日本語音声コミュニケーション教育研究会宵山研での配付資料
伝康晴・渡邉美知子（2009）「音声コミュニケーションにおける非流ちょう性の機能」『音声研究』13-1, pp. 53-64.
Bortfeld, H., S. D. Leon, J. E. Bloom, M. F. Schober, & S. E. Brennan (2001) Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech* 44-2, pp. 123-147.
Lickley, Robin J. (2001) Dialogue moves and disfluency rates. In *Proceedings of Disfluency in Spontaneous Speech (DISS)*, pp. 93-96.

関連 URL

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

『日本語日常会話コーパス』モニター公開版の概要

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
（国立国語研究所音声言語研究領域）
西川賢哉（国立国語研究所コーパス開発センター）
伝康晴（千葉大学人文科学研究院／国立国語研究所音声言語研究領域）

Overview of the monitor version of the *Corpus of Everyday Japanese Conversation*

Hanae Koiso, Haruka Amatani, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino,
Yoshiko Kawabata, Yayoi Tanaka, Ken'ya Nishikawa
(National Institute for Japanese Language and Linguistics)
Yasuharu Den (Graduate School of Humanities, Chiba University/
National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」では，200 時間規模の日常会話を収めた『日本語日常会話コーパス』の構築を進めている。このコーパスは，多様な日常場面の会話を，映像まで含めて収録・公開するものであり，世界的に見ても新しい試みである。『日本語日常会話コーパス』の本公開は，プロジェクトの最終年度にあたる 2021 年度を予定してるが，コーパスの利用可能性や問題などを把握し今後の構築に活かすために，50 時間のデータについて 2018 年 12 月にモニター公開することを予定している。そこで本稿では，モニター公開データの仕様や特徴について報告する。

1. はじめに

国立国語研究所では，2016 年度から開始した共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」において，さまざまなタイプの日常会話 200 時間をバランス良く収めた大規模なコーパス『日本語日常会話コーパス』（*Corpus of Everyday Japanese Conversation*, 以下 CEJC）の構築を進めている (小磯 2017)。CEJC の特徴は，(1) 収録のために集められた状況での会話ではなく，日常場面の中で当事者たち自身の動機や目的によって自然に生じる会話を対象とすること，(2) 多様な場面の会話をバランスよく集めること，(3) 音声だけでなく映像まで含めて収録・公開することである (小磯ほか 2017)。特に，日常生活の中で生じる会話を 200 時間の規模で映像まで含めて公開するというのは，世界的に見ても新しい取り組みである。CEJC の本公開は，プロジェクトの最終年度にあたる 2021 年度に行う予定であるが，コーパスの利用可能性や問題などを把握し今後の構築に活かすために，50 時間分の会話を 2018 年 12 月にモニター公開することを予定しており，現在，その準備を進めている。モ

* koiso@ninjal.ac.jp

ニター公開では，(1) 50 時間分の会話の映像・音声データ，転記テキスト，形態論情報（短単位情報），ツールなどを収めたハードディスクでの公開と，(2) 形態論情報（短単位情報）をオンラインで検索できる「中納言」[(1)]での公開を予定している。本稿では，両者に共通するものとして，会話の収録法（2 節），コーパス格納データの選定方針（3 節），及びモニター公開対象データの特徴として調査協力者や会話参加者の属性，会話の種類の内訳（4 節）について報告する。その上で，ハードディスクに同梱するデータとして，映像・音声データ，転記テキスト，短単位情報の仕様についてまとめる（5 節）。

2. 会話の収録法

日常場面の中で当事者たち自身の動機や目的によって自然に生じる会話をバランスよく収録するために，主として個人密着法と呼ぶ収録法で会話を集めている。個人密着法は，日常生活の中で生じる会話を，一般の調査協力者（以下，協力者）自身に収録してもらうという方法である。性別・年代の点から均衡性を考慮して選別された 40 名の協力者（男女 ×20 代・30 代・40 代・50 代・60 代以上 × 各 4 名）に収録機材を 3 ヶ月ほど貸し出し，日常生活における多様な場面の会話を 15 時間程度収録してもらう。この中から，データの質や倫理的・法的な問題，バランス，会話参加者（以下，会話者）の希望などを考慮し，コーパスに格納するデータとして 4〜5 時間の会話を選別する[(2)]。モニター公開で対象とするのは，このうち 20 名の協力者が収録したデータである。協力者の内訳については 4.1 節で述べる。

個人密着法では，調査者は収録に介在しない。そのため，協力者自身に，会話の映像・音声の収録，会話者への調査内容及びデータ公開方法の説明，同意書への署名の依頼，フェイスシート（性別，出身地などの会話者の属性）記入の依頼，会話の収録状況等の記録など，実に多くのことを担当してもらう必要がある。このように収録調査には各種個人情報を扱うなど重い責任が生じることから，協力者は 20 歳以上の成人に限定している。収録法の詳細については田中ほか (2018) を参照されたい。

3. コーパス格納データの選定方針

本節では，コーパスに格納するデータをどのように選定しているかについて述べる。CEJC は，多様な会話をバランスよく集めることを目標に掲げている。そこで，普段われわれがどのような種類の会話をどの程度行っているかの指標を得るために，約 250 人の成人を対象に，起床から就寝までの間に行った全ての会話について，いつ，どこで，誰と，何をしながら，どのような種類の会話を行ったか，などを問う会話行動調査を実施した (小磯ほか 2016)。この調査結果を一つの目安として格納データの選定を進めている (小磯ほか 2017, Koiso et al. 2018)。モニター公開対象についても，構築状況を見ながらできるだけ多様な会話が含まれるように選定した（4.3 節参照)。

個人密着法では，収録を始める前に，機材の設定や会話者への説明，書類の記入などが必要

(1) https://chunagon.ninjal.ac.jp/useraccount/register

(2) 個人密着法では収録が難しいと思われる場面を調査者が主体となり収録する方法として，特定場面法を採用する。約 20 時間をこれに当てる予定である。

表 1　モニター公開対象データの調査協力者の属性，対象とする収録数・会話数，会話時間

	男性				女性			
年代	職業・職種等	収録数	会話数	時間	職業・職種等	収録数	会話数	時間
20 代	大学生	5	5	2.2h	大学生	7	7	2.6h
	大学院生	5	5	2.5h	大学生	5	10	2.6h
30 代	自営業・自由業	4	4	2.8h	会社員・公務員等	5	6	2.7h
	会社員・公務員等	6	6	2.2h	専業主婦	7	7	2.8h
	会社員・公務員等	4	5	2.1h	会社員・公務員等	5	5	2.6h
40 代	自営業・自由業	6	6	2.4h	パート・アルバイト	6	6	2.6h
					パート・アルバイト	6	6	2.6h
50 代	会社員・公務員等	7	7	2.4h	会社員・公務員等	7	7	2.2h
	会社員・公務員等	4	4	2.6h	自営業・自由業	6	6	2.7h
60 代	その他（非常勤講師）	9	9	2.1h	専業主婦	6	7	2.7h
以上	定年退職	6	8	3.0h				

となるため，話が少し進んだところから収録が開始されることもある。また 1 回の収録は最大でも 1 時間程度としており[(3)]，会話の途中で収録が切れることもある。そのため，協力者が収録したものから，ある程度のまとまりをもった範囲を「会話」として切り出し，コーパスに格納するデータを決めている。倫理的・法的な問題や会話者の希望などを考慮し，問題のある部分をカットした結果，一つの収録データが複数の会話に分かれることもある。

4.　モニター公開対象データの特徴

4.1　調査協力者の属性

2018 年 3 月末の時点で，収録調査，コーパス格納データ選定，転記 1 次作業，フォローアップインタビューを全て終了した協力者の中から，バランスを考慮して，モニター公開対象とする協力者を 20 名選んだ。協力者の属性，対象とする収録・会話の数，会話時間の合計を表 1 に示す。収録スケジュールの都合で 40 代の女性が 3 名，60 代以上の女性が 1 名となっているが，それ以外は性別・年代をバランスさせて各層 2 名ずつとした。収録数は全体で 116 回，126 会話，約 50 時間（平均 2.5 時間 / 1 人）である。

4.2　会話者の属性

モニター公開対象となる 116 の収録に含まれる会話者は，延べ 392 名，異なり 237 名である[(4)]。性別・年代ごとの数を図 1 に示す。

男性・女性ともに未成年者の数が少ないが，2 節に記したように，個人密着法に基づく収録調査では重い負担を伴うことから，協力者は成人に限定している。そのため，未成年者の数は他と比べ必然的に少なくなる。また，延べ，異なりともに，女性に関しては 40 代・50 代が多く 60 代・70 代が少ない傾向が見られる。40 代女性の協力者が 3 名と多かったために，同性・

(3) コーパスに格納するのは 1 協力者あたり 4〜5 時間と限られており，バリエーションを確保するために，このような上限を設けている。

(4) データには店員との注文等のやりとりなども含まれるが，多くの場合，店員はメインの会話者ではないため，数には含めていない。店員であっても，長く会話を続ける場合で，収録・公開の同意を得たものについては，その限りでない。そのほか，配偶者との会話の途中で妹と電話で短い会話をしているものがあるが，この場合の妹も数に含めていない。

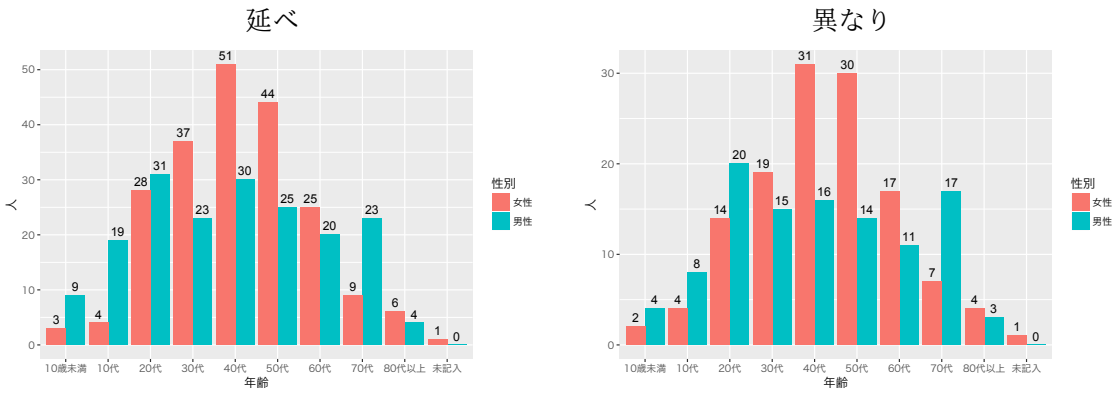


図 1　会話者の性別・年代の内訳（人）

表 2　会話者の職業の内訳（人）

職業	延べ	異なり	職業	延べ	異なり
会社員・公務員等	127	80	高校生	1	1
自営業・自由業	42	28	中学生	9	4
パート・アルバイト	34	19	小学生	15	6
専業主婦	61	42	就学前	6	4
無職・定年退職	27	19	その他	14	4
大学生・大学院生	54	28	未記入	2	2

同世代の人との会話が他と比べて多くなったためと考えられる。CEJC 全体では，協力者の年齢・性別のバランスをとるようにしているが，こうした対応がコーパス全体の質を保証する上で重要と言える。

会話者の職業の内訳を表 2 に示す。未成年者が少ないことと関係するが，高校生・中学生・小学生・就学前の人数が少ない。特に高校生についてはモニター公開対象データでは 1 名のみである。成人については，会社員・公務員等が一番多いものの，それ以外の職業も含めて多様な職業の会話者が含まれている。

協力者から見た相手の会話者との関係を表 3 に示す。家族や友人知人との会話が多く，仕事関係者や学校等の関係者は少ない。公共商業サービス関係の会話者も少ないが，先に注記したように，注文等で会話した店員などは会話者数には含めていない。そのような店員などを含めると，「サービスを提供する人」は 34 名となる。

4.3　会話の種類

3 節で述べたように，CEJC では多様な会話をバランスよく集めるために，会話行動調査を実施した。そこで，モニター公開データを対象に「形式」「会話者数」「活動」「場所」の内訳を求め，行動調査の結果と比較する。両者を合わせて図 2 に示す。図の上段は会話の件数で見た場合の，下段は時間で見た場合の割合の比較である。

「形式」については雑談が約 7 割を占めており，行動調査より若干多いものの，概ねバランスよくデータが選定できていることが分かる。会議・会合は件数で見ると行動調査より多いが，時間で見ると少ない傾向が見られる。CEJC では収録の上限を 1 時間に設定しているのに

表 3　調査協力者から見た会話者との関係の内訳（人）

関係性 1	関係性 2	延べ	異なり
家族親戚	配偶者	30	13
	子供	36	17
	父母	17	12
	義父母	9	7
	自分の兄弟姉妹	7	7
	配偶者の兄弟姉妹	3	3
	自分の祖父母	4	3
	おい・めい	3	3
	その他	5	4
	計	114	69
友人知人		118	108
	計	118	108

関係性 1	関係性 2	延べ	異なり
仕事	職場の上司	1	1
	同僚	7	7
	部下	3	3
	取引先など他社の人	6	6
	その他	9	9
	計	26	26
先生生徒	学校の先生	4	4
	学校の生徒・学生	6	3
	習い事などの生徒	1	1
	計	11	8
公共商業サービス	サービスを提供する人	4	4
	サービスを受ける人	2	2
	計	6	6

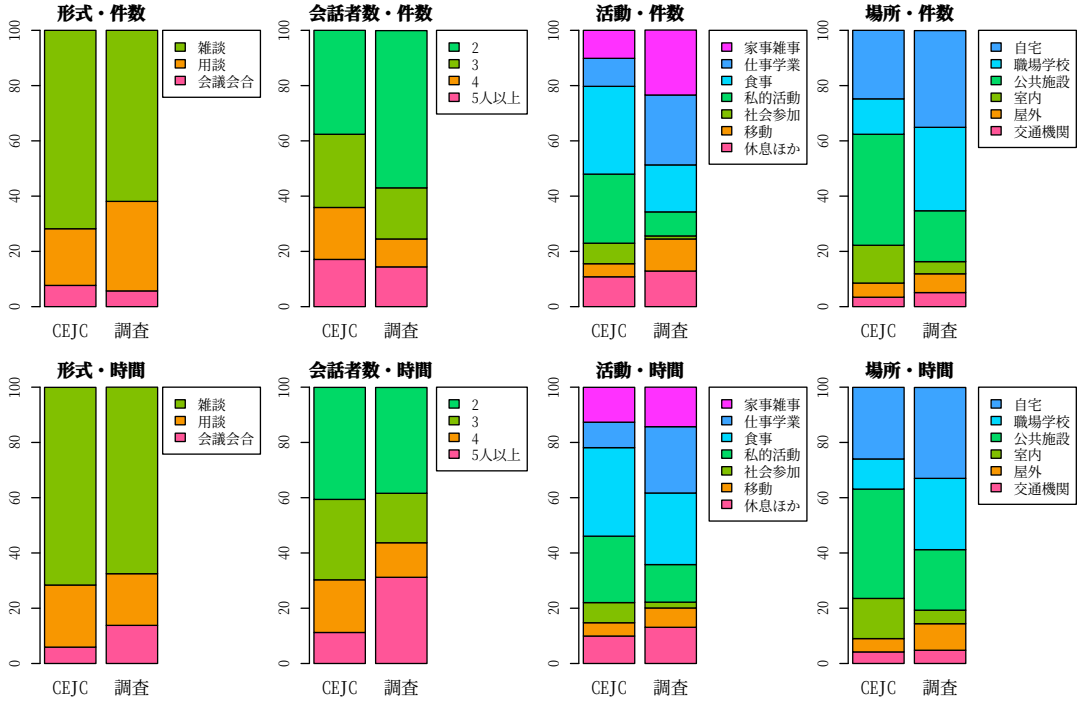


図 2　モニター公開対象データと会話行動調査における「形式」「会話者数」「活動」「場所」の内訳

対し，実際の会議・会合は 1 時間を越えるものが少なからずあることが影響していると考えられる。「会話者数」において，5 人以上の場合に，件数では行動調査と同程度だが時間で見ると少ない傾向が見られるのも，収録の上限を定めた影響と考えられる。

「活動」と「場所」については，料理や棚の組み立てなどの家事雑事，ボランティアなどの社会参加，屋外・交通機関での移動など，多様な場面の会話が収録できているが，行動調査と比べると，職場や学校における仕事・学業中の会話が少ない。個人密着法ではこの種の会話の収録が難しいためであり，今後，特定場面法で補強する予定である。

図 3　基本収録の機材セットで記録した映像の例。左の映像は Kodak PIXPRO SP360 で，右の上下二つの映像は GoPro Hero3+ で録画したもの。論文掲載用に会話者の顔にボカシの処理を加えている。

5.　同梱データの仕様

5.1　映像・音声データ

個人密着法では以下に示す機材を用いて会話を収録している。詳細については田中ほか (2018) を参照されたい。

【基本収録】屋内等での対面会話の収録で主として使用

映像　以下のカメラ 2 種，計最大 3 台（最低 1 台[5]）を使用して会話を録画（図 3 参照）

1. Kodak PIXPRO SP360 4k（1440×1440, 59.94fps）, 最大 1 台。360 度撮影可能なカメラで会話の場の中央に配置。
2. GoPro Hero3+（1920×1080, 59.94ffps），最大 2 台。170 度の視野角を持つカメラで会話者や会話の状況を俯瞰的に記録。

音声　各会話者の音声と会話全体の収録のために以下 2 種の IC レコーダーを使用

1. 個人用 IC レコーダー Sony ICD-SX734（リニア PCM, 16bit，44.1kHz）, 会話者数に応じて 2〜6 台使用[6]。レコーダーをフォルダーに入れ首から下げて収録。マイク設定は「ズームマイク・音声用・ズーム 1」（単一指向性マイクでモノラル録音），感度は事前調査に基づき定めたレベルに固定。
2. 全体用 IC レコーダー Sony ICD-SX1000（リニア PCM, 16bit，44.1kHz）, 1 台。会話の場の中央に配置し会話全体を収録。マイク設定は「ステレオマイク」（IC レコーダー先端左右両端のマイクでステレオ録音），感度は auto に設定。

---

[5] 電話等，非対面での会話の場合はカメラを用いないことが多い。

[6] 会話者が 6 名を越える場合や収録の失敗などの理由で，全ての会話者の音声が個人用レコーダーで収録できていないこともある。原則として貸出は 6 台としているが，協力者の要請により最大 15 台まで使用したことがある。

表 4　公開の映像・音声データのファイル形式

映　像	PIXPRO SP360	mp4, H264, 1440×1440, 29.97fps
	GoPro	mp4, H264, 1280×720, 29.97fps
	HX-A500	mp4, H264, 1280×720, 29.97fps
	合成	mp4, H264, 1360×720, 29.97fps
音　声	ICD-SX734	リニア PCM, 16bit, 16kHz, モノラル
	ICD-SX1000	リニア PCM, 16bit, 16kHz, ステレオ
	合成	リニア PCM, 16bit, 16kHz, ステレオ

【移動時収録】 移動時の会話の収録で主として使用

映像　Panasonic HX-A500（1920×1080, 29.97fps），1 台。散歩などの移動の際に会話者のうち 1 名が頭に装着して映像を収録。

音声　個人用 IC レコーダー Sony ICD-SX734（基本収録と同設定）

複数のカメラによる映像がある場合，図 3 に示すような合成した映像も作成し，個々の映像データと合わせて公開する。また，全体用 IC レコーダーで収録した音声が何らかの理由で公開できない場合[(7)]，あるいは，公開はできるが質に問題がある場合，個人用 IC レコーダーで収録した各会話者の音声を合成した音源を作成して公開する。同梱する映像・音声データのファイル形式を表 4 に示す[(8)]。

映像データの音声については次の通りとする。HX-A500 と合成の映像の場合，全体用 IC レコーダーの音声あるいは合成音声（後者を優先）を用いる。PIXPRO SP360 と GoPro の場合，原則としてそれぞれのカメラで記録した音声を用いるが，音質などに問題がある場合，全体用 IC レコーダーの音声あるいは合成音声（後者を優先）を用いる。

5.2　転記テキスト

図 4 に，モニター公開で提供する転記テキストの例を示す。映像分析ソフトウェア ELAN や音声分析ソフトウェア Praat などを用い，映像・音声を参照しながら人手で作成している。原則として漢字仮名まじりで表記するが，母音の延伸や発音エラーなど会話で生じる現象については，表 5 に示す各種タグを用いて表現する。転記テキストの 1 行は転記単位に相当する。転記単位とは，知覚可能な休止，発話単位の境界，あるいは相互に異なる音種（言語音と笑い，泣き，歌）の境界のいずれかによって区切られる単位である。転記単位ごとに，発話の開始時間と終了時間が割り当てられており，転記テキストから映像・音声データが容易に参照できるようになっている。句点「。」は発話単位の境界を示す。発話単位とは，Japanese Discourse Research Initiative によって策定された「長い発話単位」に相当する (JDRI 2017)。転記テキストの詳細については臼田ほか (2018) を参照されたい。

転記テキストは，2 種類の単位（転記単位・発話単位）ごとに，CSV 形式・EAF 形式（ELAN 用）・TextGrid 形式（Praat 用）で提供する。

(7) 基本収録において録音に失敗した場合，飲食店などでの収録において第三者の会話音声が大きく写り込んでしまうなどの理由により公開すべきではないと判断した場合，及び移動時の収録のように全体用 IC レコーダーでの収録がもともとなされていない場合などが該当する。

(8) 映像については，何らかの事情で収録時の設定が変更されてしまったために，ここに示す値と異なることがある。

開始時間	終了時間	会話者の ID	テキスト
2502.617	2503.920	IC01_一ノ宮	(U この前) 飲み会どこで飲んだの。
2504.661	2505.651	IC03_さとし	えっと 赤坂。
2507.718	2508.495	N10A_酒井	赤坂の
2508.791	2509.744	IC03_さとし	(L)
2509.287	2510.202	N10A_酒井	料亭。
2510.912	2511.480	IC03_さとし	(L いや いや)。
2511.432	2512.185	IC01_一ノ宮	違う違う。
2512.749	2513.451	IC01_一ノ宮	居酒屋。
2513.641	2514.236	IC03_さとし	(W イサカヤ|居酒屋)。
2515.464	2516.201	IC03_さとし	(X フタヘルモ)。
2516.999	2519.648	IC03_さとし	同期の (D ヒ)(D フ) 同期と二人で飲んだぐらいで。
2519.670	2521.473	Z10A_酒井	芸能人もいっぱい歩いてるんじゃないの。
2521.473	2522.070	Z10A_酒井	そうすっと。
2522.235	2522.864	IC03_さとし	んな見る余裕。
2522.869	2526.534	IC03_さとし	もう 仕事終わったら家帰ることしか頭に (L ないです)。
2523.585	2524.039	Z10A_酒井	ね:。
2526.541	2527.636	IC03_さとし	(L)
2530.214	2531.759	IC01_一ノ宮	前ＴＢＳの地下で:
2532.456	2533.398	IC01_一ノ宮	(R (U たか))(W (D サ)|さん) ジュリー見た。@ジュリーを見たのは発話者本人

図 4　転記テキストの例

表 5　転記テキストに用いるタグの一覧

■ 非語彙的な発音の変化や言いよどみに関わるもの

タグ	概要	使用例
:	非語彙的な母音の引き伸ばし	すご:い，けれども:
%	非語彙的な音の詰まり	す % ごい，解 % 析
(W)	言い誤り・発音の怠け等の一時的な発音エラー	(W コエ|これ)，(W ギーツ|技術)
(D)	語の言いさし	(D コ) 明日から

■ 韻律・パラ言語的情報に関わるもの

?	疑問上昇調	行きます?，コップ?
(T)	小さい声で発話している箇所	(T これじゃないのか)
(L)	笑いが生じている箇所，あるいは単独の笑い	これ (L なんですけど), (L)
(C)	泣きながら発話している，あるいは単独の泣き	(C なにが), (C)
(S)	歌いながら発話している，あるいは歌詞を伴わない歌	(S ヘイヘイホー)，(S)
<>	発音に類する行為のうち会話の流れに関わるもの	<舌打ち>，<咳>，<口笛>

■ 聞き取り等の判断の信頼性に関わるもの

(U)	聞き取りや語の判断に自信がない箇所	(U ジャック) に，(U 国産/特産)
(X)	語が不明な箇所	(X フンジン) 中に，(X ＃＃＃)

■ 転記テキストの可読性や内容理解の補助等に関わるもの

(K)	タグ等のために漢字表記できず可読性が落ちる箇所	(K シ:ツ|質) 問，(K ナ % シ|梨)
(M)	音や言葉が言及されており (W) などで対応すると把握しづらい箇所	(M すごい) を (M すっごい) と発音
(O)	一般的に理解が難しい外国語・方言が用いられる箇所	(O ボッソワー)，(O ＃＃＃)
(Y)	漢字表記の一般的な読みと発音が異なる箇所	(Y ゼツ|舌)，(Y センゲン|浅間)
(G)	可読性が低い口語表現	(G 嫌|や)，(G もう|も)
(F)	「あの」「その」類が連体詞ではなくフィラーとして用いられる場合	(F あの)，(F そーの:)

■ 発話単位・転記単位に関わるもの

。	発話単位末	食べます。，やったけど。，うん。
+	1 短単位内の知覚可能な休止により転記単位が分割される場合	す + ごい，神田 + 川

■ その他

(R)	個人情報などに関わる仮名・伏字処理を行った箇所	(R 国語研究所) の (R 佐藤) さん
@	発話に対するコメント	お願いしま:す。@店員への応答

5.3 短単位情報

モニター公開では，長短 2 種類の形態論情報のうち短単位情報 (小椋 2014) を提供する。転記テキストを対象に，形態素解析器 MeCab と形態素解析用辞書 UniDic を用いて自動解析した上で，人手による修正を加えた。自動解析で得られた語形・発音形のうち，語形・発音形が一意に同定できない語（例：一日「イチニチ/ツイタチ」，日本「ニホン/ニッポン」）については，音を聴取しながら語形・発音形の確認・修正を行った。

形態素解析が対象とするのは，転記テキストにおける母音の引き述ばしや音の詰まりのタグをとった形，また言い誤り・発音の怠け等の一時的な発音エラーがある場合はそれを丁寧に発音した場合に生じると想定される語の形である。例えば「けれども:」「く % さい」「(W ギーツ|技術)」であれば，「けれども」「くさい」「技術」が解析の対象となり，発音形にはこうした母音の引き述ばしや音の詰まり，言い誤りの情報は記録されない。そこで，転記テキストから得られる実際の「発音」（この例でいえば「ケレドモー」「クッサイ」「ギーツ」）についても「発音形」とは別に公開する。

短単位情報は，CSV 形式で提供するほか，同梱する全文検索システム「ひまわり」パッケージで検索することができる (山口 2018)。またオンライン検索システム「中納言」でも公開する。

5.4 映像・音声・転記テキストのマスキング処理

会話を収録する際，会話者と交わした同意書では，会話者の名前，所属組織名，自宅・所属組織の住所・電話番号の情報，及び会話者が公開を望まない箇所について，それらが分からないよう会話の音声と文字化資料（転記テキスト）を加工することを定めている。音声は当該箇所をビープ音で置換する処理を施す。転記テキストについては，タグ (R) を付けた上で，仮名あるいは伏字（全角アスタリスク）に置換する処理を施す（図 4 に示す転記テキストの例の最終行を参照）。

同意書において，映像の加工についての条件は付けておらず，原則として会話者の顔にボカシなどの処理を加えずに公開する。ただし，名札や名刺など個人情報を含むものや，収録・公開の同意を得ていない第三者の容貌が映像に写り込むことも少なからずある。そこで，実際に収録された映像・音声データにもとづき法的・倫理的な観点から問題を整理した上で，公開方針を定めた。この方針に従い，必要と判断される箇所について映像にボカシ処理を加える。公開方針の詳細については小磯・伝 (2018) を参照を参照されたい。

6. おわりに

本稿では，『日本語日常会話コーパス』のうち，2018 年度 12 月にモニター公開を予定している 50 時間分のデータの仕様や特徴について報告した(9)。ハードディスク版については，ハードディスク代・郵送代などの実費相当での提供を予定している。モニター公開の最新情報は，以下のページを参照されたい。

http://pj.ninjal.ac.jp/conversation/cejc-monitor.html

(9) 現在，公開に向けて最終的な確認・修正を進めているところであり，ここでの報告と若干異なる可能性がある。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」によるものである。コーパスの収録にご協力・ご参加くださった皆さまに感謝します。

文　献

小磯花絵 (2017).「『日常会話コーパス』プロジェクト—コーパスに基づく話し言葉の多角的研究を目指して—」　言語資源活用ワークショップ 2016 発表論文集, pp. 114–119.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』の構築」　言語処理学会第 23 回年次大会, pp. 775–778.

田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2018).「『日本語日常会話コーパス』の構築 : 会話収録法に着目して」　国立国語研究所論集, 14, pp. 275–292.

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

Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Kenya Nishikawa, Yayoi Tanaka, and Yasuyuki Usuda (2018). “Construction of the Corpus of Everyday Japanese Conversation: An interim report.” *Proceedings of the 11th edition of Language Resources and Evaluation Conference*, pp. 4259–4264.

JDRI (2017).『『発話単位ラベリングマニュアル』 version 2.1』. http://www.jdri.org/open-data/

臼田泰如・川端良子・西川賢也・石本祐一・小磯花絵 (2018).「『日本語日常会話コーパス』における転記の基準と作成手法」　国立国語研究所論集, 15, pp. 177–193.

小椋秀樹 (2014).「形態論情報」　山崎誠 (編)『書き言葉コーパス　設計と構築』　2 巻講座日本語コーパス, 第 4 章　pp. 68–88.

山口昌也 (2018).「「日常会話コーパス」活用環境の構築」　言語資源活用ワークショップ 2018 発表論文集.

小磯花絵・伝康晴 (2018).「『日本語日常会話コーパス』データ公開方針 : 法的・倫理的な観点からの検討を踏まえて」　国立国語研究所論集, 15, pp. 75–89.

日本語学習者属性別の言語行為の対話自動生成への適用に関する一考察

太田博三（放送大学 教養学部）†

A Study on Application of Automatic Dialogue Generation of Language Acts by Japanese Learner's Attribute

HiromitsuOta (The Open University of Japan)

要旨

最近，自然言語処理における対話システムや対話生成が注目されている．チャットボットのコールセンターへの普及により，正確な人間性な対話応答が求められ ている．一方，社会学のエスノメソドロジーや談話分析・会話分析における定性的な相互行為は有益である．そこで，もう一度，国立国語研究所の提供する日本 語学習者会話データコーパスを用いて，効果を検証し対話破綻の傾向や対話生成に適用することを目指した考察である．

１．はじめに

1.1 研究の背景と目的

スマートスピーカーが家庭に普及し，自動運転が実用化されようとしている中，従来から発展し商用化されているロボットの Pepper や各種チャットボット（Chatbot）は，人間と比べて，小さくない乖離があると指摘され，4 年以上が経過している．チャットボットのコールセンター等への導入も 2 年以上経過している．ここでは，制御文による対話応答が第 1 義的に実装され，第 2 義的にディープラーニング（Deep Learning）による運用が，もくされた．しかし，いずれも不完全燃焼に終始している．どちらか一方，もしくは折衷でも，人間に代替する品質に昇華できていないと考え，調べ始めたのが本考察のきっかけである．

次に，エスノメソドロジーや会話分析の勉強会に参加した際，対話システムのような粗い応答では不十分だという趣旨のご発言と勝手ながら解釈した機会があり，追及してみようと考えた．

そこで，現行の対話システムに定性的な視点で考察し，定量的な分析に持ち込むことで，スケール化させ，実用化に結びつける第一歩にしたいというのが目的である．全体的にディープラーニングに適用したらよいか，部分的かも検討したい．

1.2 研究の新規性

本研究の遠い新規性となるが，もっぱら，数量データによるディープラーニングに定性的な要素を取り入れたいという点であるが，本稿では，誰もが入手可能なデータである日本語学習者の会話データを用いることで，統計的な有意性やサンプル数より，日常生活の感覚でわかることを重視したものである．次に，[質問]-[応答]や[申し出]-[受諾/拒否]などの隣接ペアの類型が上記のデータにどのくらいあてはまるかなど，計量化してみた．ここで，実証的な知見が得られれば，話者交代の予測や対話の破綻をしても修復する発話を学習させるなど，次につなげられる．具体的には，隣接ペアの次には，主に以下の 5 つが考えられる．

† 9924658973@campus.ouj.ac.jp

1) Yes/No の応答詞　：あー，うん，えー，そう
2) あいづち　：んー，はい，はー，えー
3) 言いよどみ：んー，あー，えー
4) 呼びかけ　：ね，ねー
5) フィラー　：あのー，そのー，えーと，えっと

今後，これらを分析しシステムに追加することで，更なる対話システムの質的向上につながると考えられる．

1.3　研究の主な手法

基礎集計を中心に行いながら考察する．国立国語研究所の提供している「日本語学習者会話データベース」を用いて集計を行う．隣接ペアは本稿で定義する種類のものに限定し計量化する．次に，それらのペアが全体の会話の促進になっているかなどを考察する．また，その隣接ペアの前後，もしくは直後の発話が修復に向けてのものか，完全に破綻しているが強引に会話を続けたものであるのかも含めて，定性的な判断を行う．

1.4 用いたデータセットについて

国立国語研究所が公開しているコーパスの中の 1 つである「日本語学習者会話データベース」（図 1）を用いる．また KY コーパスも同様の趣旨で作られたものであり，適宜，用いた．1990 年の入管法の改正により，日本の社会状況に応じて，外国人受入れの適切な方策が必要となり，日本語学習を必要とする住民（言語生活者）の需要に見合った言語教育の展開が期待されていた．ACTFL-OPI（全米外国語教育協会認定の面接式口頭能力テスト）を活用し，日本語を用いた自然な会話に限りなく近い対話で構成されている．

図 1．属性別の日本語教育会話データベースの検索画面

表 1-. OPI 能力区分表

区分	OPIレベル	階級	OPI評価
1	超級（Superior）	1	超級
2	上級（Advanced）	2	上級−上
	〃	3	上級−中
	〃	4	上級−下
3	中級（Intermediate）	5	中級−上
	〃	6	中級−中
	〃	7	中級−下
4	初級（Novice）	8	初級−上
	〃	9	初級−中
	〃	10	初級−下

ACTFL(全米外国語協会)による OPI(Oral Proficiency Interview Test)に基づいており，日本語 OPI は 1993 年に発足し，15 年近く経過している．ここでの判断尺度は，次の 4 つに区分されている．

1) 超級（Superior）
2) 上級（Advanced）
3) 中級（Intermediate）
4) 初級（Novice）

これは「日本語学習者会話データベースの利用手引き（平成 22 年 5 月 国立国語研究所）」によれば，言語運用能力は 10 種類の階級に区分されている(表 1)．対話のスクリプトは，インフォーマント（日本語学習者/ データ提供者）とテスター（面接者）とからなり，30 分ほどの対話形式で構成されている．

また上記の 10 段階の OPI レベルや性別，年齢，出身国などを選択することができる．検索条件を設定してダウンロードすると，文字化（一部，音声化）されたスクリプトが入手でき，有用である．

２．先行研究

本考察では，下記の 3 つ区分した．1 つ目は，エスノメソドロジーや会話分析などの社会学である．言語学も多分に含まれている．2 つ目は，対話システムを支える自然言語処理，3 つ目は，深層学習，すなわちディープラーニングである．

２.１　エスノメソドロジー・会話分析

坊農・高梨他(2009)では，隣接ペアとは，[質問]-[応答]の対をなす発話の連鎖を指すものとして，対話システムにおける対話モデルに発話連鎖構造の土台としているとある．さらに，隣接ペアの概念には，[質問]に対し，[応答]がなされなかった場合に，どのような修復連鎖や挿入連鎖構造が生起しながら会話が進行するかを述べている．魏(2015)は「あのー」や「まー」などをフィラーと定義し，発話者が何らかの心的操作を行っている最中に発するもので，場をつなぐ機能を持つ言葉と定義している．多くは「感動詞」や「間投詞」に区分される．このフィラーを使いこなすのも，あいづちなどと同じく，会話をつなぐ言葉として，留意したいと考えている．

２.２　自然言語処理

対話システムに実装される可能性は示している．また，徳永・乾・松本(2005)及び徳永(2014)は，チャット対話の収集からコーパス作成，そしてチャット対話の構造モデルを提案している．このチャット対話の質問や返答などの談話機能を担う構成単位が交換行為である．交換単位は「働きかけ」,「応答」,「補足」の 3 種類に区分され，さらに 2,3 の枝葉に分かれている．また，素性に関する考察は有益であり（表 2),

本研究ではこれらを精緻化することが具体的な目標でもある．素性の組合せと継続関係の同定や再現率は 2 人の場合でも 3 人の場合でも，86%と高く，素性も厳選されている．発言間の結束度は次の式で求めている.< n(名詞), rel(助詞), v(動詞)>の共起確率 P(<n, rel, v>)を求める．この確率 P(<n, rel, v>)は，Probabilistic Latent Semantic Indexing(PLSI)で推定する．単語の共起を潜在的な意味から同時発生とみなす手法である．PLSI における共起確率 P(<n, rel, v>)は次の式で与えられる．

$P(<n, rel, v>) = \Sigma P(<rel, v|z)\ P<rel, v|z)P(n|z)P(z).$

ここで，z は共起の潜在的な意味クラス(隠れクラス)を指しており，パラメータの P(<rel, v|z>, p(n|z), p(z))は EM アルゴリズムで推定している．

表 2 素性一覧徳永・乾・松本(2005)

素性	素性の説明
発言の末尾の表層表現	各発話の末尾が句点，読点，クエスチョンマークであるか否かの2値
CRRuとPREu間の発言時間の差	CRRuとPREu間の発言時間の差が2分以上であるか否かの2値
発言間の結束度	共起確率に基づくCRRuとPREu間とCRRuとNBNu_s間の結束度の強い方を1とする2値
交換行為の対話クラス	対話行為辞典に各交換行為(20種類 ≒ 隣接対)のクラスに分類したもの
交換行為の末尾の表層表現	同一人物の複数読み取れる発話の一番最後の末尾がクエスチョンマークであるか否かの2値
交換行為の発言時間の差	CRRuとPREmの先頭の発話における発話時間の差が5分以上であるか否かの2値

2.3 対話自動生成のディープラーニング

対話応答の自動生成に関しては，ICML Workshop(2015)で Vinyals et al(2015)の Google のチームが NIPS2014 で発表された Sequence to Sequence model を基としている．多層の Long-Short term memory(LSTM)を用いて文章をベクトル化（エンコード）し，別の多層 LSTM を用いてベクトルをデコード（復元）するものである．これは「日本語－英語」間の機械翻訳でよく用いられているアーキテクトであり，従来と比べて，自然な会話を生成するようになった．

Ghazvininejad et al(2018)は，上記のモデルを拡張発展させたものである．会話型だけでなく非会話型データも組み合わせることにより，Seq2seq における Neural Conversation Model を発展させたものである．

2.4 対話システムと会話ユーザーインターフェース

狩野(2017)では，現在に至るまでの対話システムと将来の展望を簡潔にまとめられている．1960 年代に開発された ELIZA や人工無能から，「○」「△」「×」などのアノテーションによるチューリングテストを経て，現代の雑談対話システムの 1 つである 2016 年に発表された論文に基づく Microsoft 社の「りんな」まで網羅している．「りんな」では，発話ペアデータと教師付き機械学習は統計的な対話システムの多くに共通していることが少なくない．

また，対話データを正解データとしてくんれんする強化学習では，状態遷移の訓練になるため，会話の流れを学習することになり，未知の対話に対応することが期待されている．

2.5 国内外での取り組み

対話破綻検出チャレンジ(2015-2017)や DTFC7, NTCIR-14 など，年次でハッカソンのような国際的な大会として開催され，集合知となっている．

3. 基礎集計と分析による考察

日本語学習者会話データベース全体的にデータを見渡してみると，全データは390個ある．インフォーマント(日本語学習者)の属性は，20 代が圧倒的に多く，女性が男性の 2 倍近くおり，大半を占めている．日本語学校生や大学・大学院生が半分を占めている（図 1）．

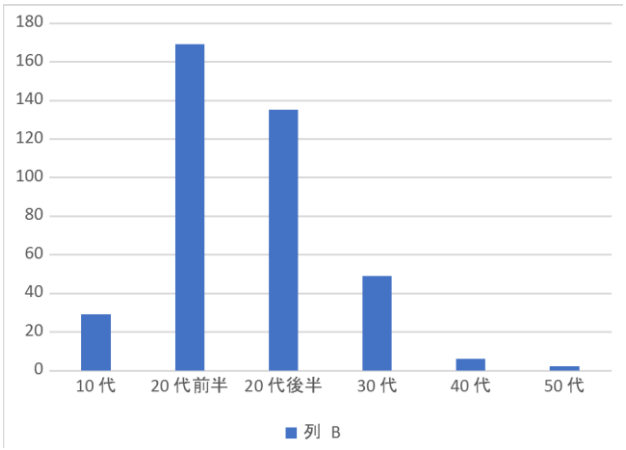


図 3.1 インフォーマントの年代別分布

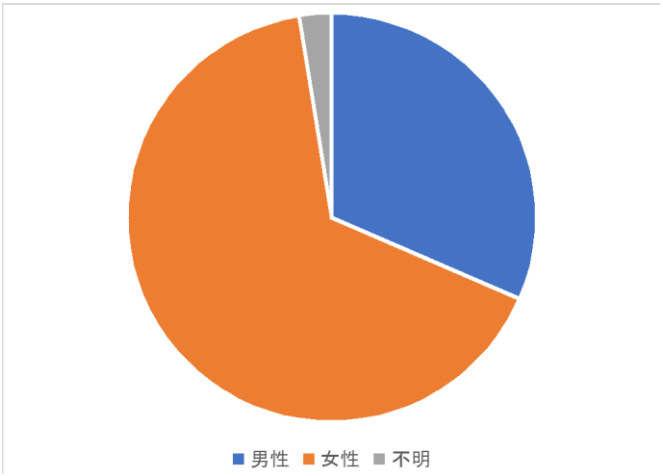


図 3.2 インフォーマントの性別

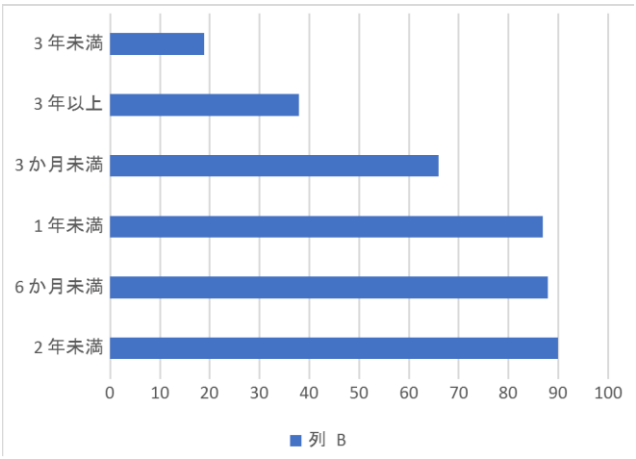


図 3.3 インフォーマントの日本滞在時間

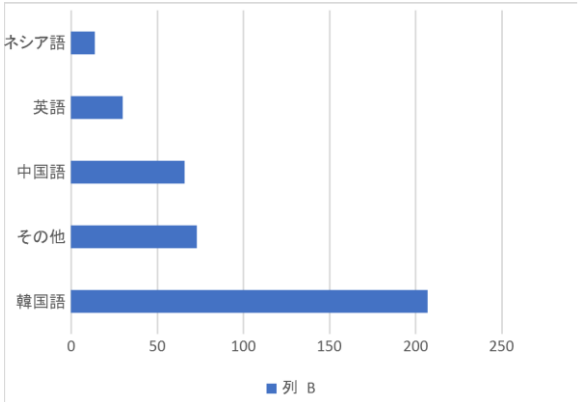


図 3.4 インフォーマントの日本語学習期間

３.１　インフォーマント属性間の比較について

本考察では，自然な対話文の自動生成を目指していることから，次の 2 つを比較考察する．OPI の判断尺度は，国立国語研究所(2010)「日本語学習者会話データベースの利用手引き」に準じ，超級と中級とで比較考察した．超級は人間と仮定し，中級はチャットボットなどの機械としてみた．主な選択した要因は次の 2 つである．

1）流暢さ

2）語用論的能力

超級では，「流暢さ」とは，会話全体がなめらかであること，これに対して，中級では，「流暢さ」とは，つかえることが多く一人で話つづけるのは難しいと定義されている．一方，超級の「語用論的能力」とは，ターンテイキングや間の取り方，相づちなどが巧みにできると定義されているのに対して，中級は，相づちや言い換えなどに成功するのはまれとされている．

3.2 超級と中級のデータについて

中級は年齢が 20 代半ばとまとまって分布しているのに対し，超級は 23 歳から 49 歳とばらつきが大きい．これはデータ数が 9 つと少ないためであるが，出身国と母国語は超級のハンガリーのインフォーマントを除いて，韓国である．滞在期間が大きく異なり，超級は 5-10 年が大半であるのに対し，中級は 3 か月から 6 か月の間に分布している．

表 3.1　超級のインフォーマント属性

	データ番号	OPIレベル	年齢	性別	出身国	母語	職業等	日本滞在期間	日本語学習期間(参考)	日本語能力試験(参考)
1	10	超級	28	女	韓国	韓国語	会社員	5年	7年	－
2	15	超級	26	男	韓国	韓国語	専門学校生	3年	18年	－
3	76	超級	34	女	韓国	韓国語	主婦	5年	4年	－
4	202	超級	35	女	韓国	韓国語	講師	10年	6年	1級
5	230	超級	26	男	中国	中国語	大学院生	5年1ヶ月	7年	－
6	258	超級	43	女	ブルガリア	ブルガリア語	教師	18年	22年？	－
7	323	超級	23	男	韓国	韓国語	大学生	5年	8年	－
8	338	超級	49	男	韓国	韓国語	会社員	22年	2年	1級
9	349	超級	40	女	韓国	韓国語	大学教員	15年	2年～3年	－

表 3.2　中級のインフォーマント属性

	データ番号	OPIレベル	年齢	性別	出身国	母語	職業等	日本滞在期間	日本語学習期間(参考)	日本語能力試験(参考)
1	2	中級-下	22	女	韓国	韓国語	大学生	3ヶ月	11ヶ月	－
2	12	中級-下	25	男	韓国	韓国語	日本語学校生	5ヶ月	5ヶ月	－
3	26	中級-下	27	女	韓国	韓国語	日本語学校生	5ヶ月	18ヶ月	－
4	8	中級-中	24	女	韓国	韓国語	日本語学校生	3ヶ月	7ヶ月	－
5	9	中級-中	25	女	韓国	韓国語	日本語学校生	2ヶ月	8ヶ月	－
6	22	中級-中	28	女	韓国	韓国語	日本語学校生	5ヶ月	1年	－
7	6	中級-上	24	女	韓国	韓国語	日本語学校生	6ヶ月	1年6ヶ月	－
8	7	中級-上	28	女	韓国	韓国語	専門学校生	2年	23ヶ月	－
9	11	中級-上	26	女	韓国	韓国語	日本語学校生	6ヶ月	9ヶ月	－

表 3.3　比較に用いたデータセット数

OPIレベル	母数	使用したデータ数	合計
超級	9	9	9
中級−下	36	3	
中級−中	84	3	9
中級−上	68	3	

3.3 隣接ペアとその計量化の検討

隣接ペアの重要な特性に，第 1 部分(First-Pair-Part: FPP)が産出されると，それに対応する特定の型の第 2 部分(Second-Part-Pair: SPP)の産出が条件的に適切になると前川・小磯他(2015)は言及している．本節では，試みの一環として，形態素解析した後に，同じ語句がでてきたら，その合計の半分として数量化した後に，目視で確認をすることにした(表 3.3.1, 表 3.3.2, 表 3.3.3)．隣接ペアが見出しやすい次の 4 つの品詞に焦点を当てて考察することにした．

1)名詞
2)感動詞
3)間投詞
4)応答詞

表 3.3.1

隣接ペアである例		
‘chiba−1232:514.5590−516.4996		
A1:	選べんだ	←第1部分(FPP)
B2:	選べる	←第2部分(SPP)
A3:	へえ	

表 3.3.2

隣接ペアでない例		
‘chiba−0332:437.2296−441.4541		
C1:	私も動物飼いたいな:	←働きかけ(I)
C2:	植物でもいいや	←働きかけ(I)
A3:	うん	←応答(R)

中級データセット
1) 頻出名詞(上位 10 件)
ん(2254) ー(1717) 笑(649) 私(294) こと(268) 音(249) 人(247) 息(246) 日本(183) お(169) 今(168) , (168) 韓国(167)

1) 頻出詞(上位 10 件)
2)感動詞
はい(2594) あー(944) あ(530) えー(309) そうですね(61) ありがとう(55) はーい(52) え(48) ま(41) ふーん(33) うん(30)

超級データセット
1)名詞

ん(2869) ー(1778) 笑(546) の(376) こと(342) 人(310) それ(292) 日本(261) 今(199) 韓国(184)

2)感動詞
はい(2114) えー(1038) あー(389) あ(273) そうですね(183) ま(159) なるほど(97) え(89) ふん(65) ふーん(54)

3.4 結果の考察
以下の解釈が考えられる.

3.4.1 形態素解析後の中級及び超級の解釈
中級の名詞では,ん(2254) ー(1717) 笑(649) 私(294)などが多く,主観性が見受けられた.その一方で,超級の名詞では,人(310) それ(292) 日本(261) 今(199) 韓国(184)などのようぬ,代名詞や国柄を表す語句が見受けられ,客観性が見受けられた.

また,中級の感動詞では,はい(2594) あー(944) あ(530) えー(309) はーい(52)など,あいまいさが見受けられた.一方で,超級の感動詞では,はい(2114) えー(1038) そうですね(183) ま(159) なるほど(97) など確かな返答が見受けられた.

3.4.2 笑いとフィラー(「あー」)の比較検討
対話の中の笑いは,{笑}で表され,中級が笑(649 に対して,超級は,笑(546)とやや少ない.笑いの機能は,早川(1997)によると,会話のターンを維持する働きがあるとされ,大きく次の 3 つに分類される(表 3.4.1).

1) バランスを取るための笑い
2) 仲間づくりのための笑い
3) ごまかしのための笑い

表 3.4.1 {笑}の談話促進の例

48_日本語学習者会話データベース	
T:	あー,あ,ん海のうえー[上]を通るんですか
I:	通るはいはい通る
T:	通るんですか
I:	はい〈あー〉,あー有名な,ですけど,んおだいばー[お台場]で〈はい〉,んーレインボーブリッジ〈はい〉がみたい〈えーえー〉,な,ほうほう,とうほうがいすありますとうとう
T:	道路
I:	はい
T:	あ道路があります
I:	あります〈はい〉はい{笑}
T:	あーあそうですか〈はい〉んーそれは有名なんですか
I:	はい有名です
T:	あーそう

一方,感動詞-フィラー(表 3.4.2)は,中級では,あー(944)に対して,超級では,あー(389)と 3 分の 1 になっている.フィラーは言いよどみや戸惑いの機能があり,超級の方が少ない結果が出ている.

表 3.4.2　感動詞―フィラーの例

辞書	文境界	書字形（＝表層形）	語彙素	語彙素読み	品詞
現代語話し言葉	I	え	えー	エー	感動詞-フィラー
現代語話し言葉	I	あのー	あの	アノ	感動詞-フィラー
現代語話し言葉	I	あ	あー	アー	感動詞-フィラー
現代語話し言葉	I	あー	あー	アー	感動詞-フィラー
現代語話し言葉	I	えー	えー	エー	感動詞-フィラー
現代語話し言葉	I	えーとー	えーと	エート	感動詞-フィラー
現代語話し言葉	I	あー	あー	アー	感動詞-フィラー
現代語話し言葉	I	えーとー	えーと	エート	感動詞-フィラー
現代語話し言葉	I	あの	あの	アノ	感動詞-フィラー
現代語話し言葉	I	あー	あー	アー	感動詞-フィラー
現代語話し言葉	I	あー	あー	アー	感動詞-フィラー
現代語話し言葉	I	あの	あの	アノ	感動詞-フィラー
現代語話し言葉	I	とー	と	ト	感動詞-フィラー
現代語話し言葉	I	あのー	あの	アノ	感動詞-フィラー
現代語話し言葉	I	あのー	あの	アノ	感動詞-フィラー
現代語話し言葉	I	えー	えー	エー	感動詞-フィラー
現代語話し言葉	I	えー	えー	エー	感動詞-フィラー
現代語話し言葉	I	ん	んー	ンー	感動詞-フィラー
現代語話し言葉	I	えー	えー	エー	感動詞-フィラー
現代語話し言葉	I	えー	えー	エー	感動詞-フィラー
現代語話し言葉	I	えー	えー	エー	感動詞-フィラー
現代語話し言葉	I	んー	んー	ンー	感動詞-フィラー
現代語話し言葉	I	そのー	その	ソノ	感動詞-フィラー
現代語話し言葉	I	んー	んー	ンー	感動詞-フィラー

表 3.4.3　言いよどみ/ 戸惑いのフィラーの例

126_日本語学習者会話データベース
T:　じゃあえーと【姓B】さんは今〈はい〉学校に一来ていま
I:　あー，起きて
T:　んーま1日でもいいですけど〈あー｛笑｝＊まはまだ〉，
I:　じゅじはんーぐら[10時半ぐらい]はい

4. 今後の展望

本研究はチャットボットなどの対話システムを対象としたものであるが，今後は次のような視点で，ロボテックスを対象とした研究につなげたい．秋谷・丹波・久野・山崎他(2007)では，介護ロボットの実現に向けて，介護者と高齢者との相互行為を深く分析したものである．今後はロボットに搭載されることが予見される．相互行為の視点が，より人間的になると考えられ，期待されている．

謝　辞

本研究の一部は，学部時代にマレーシア語や生成変形文法を教えて頂いた正保勇先生（東京外国語大学名誉教授）の雑談の中での教えが大きく影響している。また計量社会科学に関しては，博士課程時代にご指導頂いた聖学院大学大学院の松原望先生（東京大学名誉教授）を想起しながら試行錯誤できた．ここに謝意を表したい．

よび科研費基盤(A)「海外連携による日本語学習者コーパスの構築－研究と構築の有機的な繋がりに基づいて－」による成果『I-JAS』を利用して行われたものである。

文　献

坊農・高梨他(2009)「知の科学 多人数インタラクションの分析手法」3 章, 人工知能学会編集, オーム社

徳永・乾・松本(2005)「チャット対話における発言間の継続関係と応答関係の同定」自然言語処理 言語処理学会

徳永(2004)「チャット対話における発言の継続関係と応答関係の同定」修士論文 奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻

牧野成一他(2001)「ACTFL-OPI 入門」アルク

Vinyals,et al(2015) Quoc Le. A Neural Conversational Model, arXiv

Ghazvininejad(2018) A Knowledge-Grounded Neural Conversation Model. Microsoft

喜連川他(2017) 「暗黙の発話状況を考慮したニューラル対話モデル」. 言語処理学会 第 23 回年次大会 発表論文集

串田, 平本, 林(2017) 「会話分析入門」 勁草書房

船越・東中他(2016) 「対話破綻検出チャレンジにおける対話破綻データと破綻検出結果の分析」 言語処理学会 第 22 回年次大会

東中・船越(2016) Project Next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション 人工知能学会 SLUD

狩野(2017) 「コンピューターに話が通じるか 対話システムの現在」 情報管理 Vol.59 no.10

石崎・伝(2001)「談話と対話」東京大学出版会

魏(2015) 談話におけるフィラー「 ま（一)」の待遇差に関する予備的考察, 山口大学東アジア研究 学術雑誌論文

Smith et.al(2000) Conversation Trees and Threaded Chats, CSCW

秋谷・丹波・久野・山崎他(2007)「介護ロボット開発に向けた高齢者養護施設における相互行為の社会学的分析」電子情報通信学会論文誌 D Vol.J90-D No.3 pp. 798-807

早川(1997) 第 8 章 「笑い」の意図と談話展開機能 「合本 女性の言葉・男性の言葉（職場編)」

関連 URL

国立国語研究所(2009) 「日本語教育ネットワーク」 https://nknet.ninjal.ac.jp/nknet/ndata/opi/

国立国語研究所(2010)「日本語学習者会話データベースの利用手引き」

鎌田･山内「タグ付き KY コーパス」http://jhlee.sakura.ne.jp/kyc/corpus/

対話破綻検出チャレンジ 2015 https://sites.google.com/site/dialoguebreakdowndetection/

DSTC7(2017) Dialog System Technology Challenges http://workshop.colips.org/dstc7/index.html

NTCIR-14(2018-19) - 国立情報学研究所 http://research.nii.ac.jp/ntcir/ntcir-14/index-ja.html

EMCA 研究会 エスノメソドロジー・会話分析とはなにか - http://emca.jp/learn

藤田 自然言語表現の言い換え http://paraphrasing.org/~fujita/paraphrasing-ja.html

『現日研・職場談話コーパス』中納言版公開データの作成

柏野 和佳子（国立国語研究所音声言語研究領域）*
大村 舞（国立国語研究所コーパス開発センター）
西川 賢哉（国立国語研究所コーパス開発センター）
小磯 花絵（国立国語研究所音声言語研究領域）

Supplemental Arrangement for Public Data Available in the Chunagon Versions of "Gen-Nichi-Ken Corpus of Workplace Conversation"

Wakako Kashino (National Institute for Japanese Language and Linguistics)
Mai Omura (National Institute for Japanese Language and Linguistics)
Ken'ya Nishikawa (National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

『現日研・職場談話コーパス』は，現代日本語研究会が作成した，首都圏の有職女性 19 名（20 代～50 代）と，首都圏の有職男性 21 名（20 代～50 代）の職場での自然談話を文字起こししたテキストを元に作成したコーパスである。その元となっている文字化テキストは,『合本　女性のことば・男性のことば（職場編）』（現代日本語研究会編，2011 年，ひつじ書房）の付録 CD－ROM に収録されている。国立国語研究所に提供されたその文字化テキストを MeCab＋UniDic で解析し，オンライン検索システム『中納言』にて『現日研・職場談話コーパス』として公開する。

本発表では,『現日研・職場談話コーパス』の概要と特徴を述べる。

１．はじめに

現在，国立国語研究所の機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）にて，日常会話場面を対象とした大規模な『日本語日常会話コーパス』を構築中である(小磯ほか 2018)。その公開を前に，国語研に提供いただいた既存の会話データをオンライン検索システム『中納言』にて公開するということを進めている。2016 年には，『名大会話コーパス』(藤村ほか 2011)の中納言版(柏野ほか 2017)を公開した。それに続けて，このたび『現日研・職場談話コーパス』の中納言版を 2018 年 8 月より一般公開する運びとなった。

『現日研・職場談話コーパス』は，現代日本語研究会が作成した，首都圏の有職女性 19 名（20 代～50 代）と，首都圏の有職男性 21 名（20 代～50 代）の職場での自然談話を文字起こししたテキストを元に作成したコーパスである。その元となっている文字化テキストは，現代日本語研究会編(2011)の付録 CD－ROM に収録されている。「大規模日常会話コーパスに基づく話し言葉の多角的研究」のプロジェクトにおいて，その文字化テキストを対象に，形態素解析用辞書『UniDic』と形態素解析器『MeCab』を用いて形態論情報

* waka ＠ninjal.ac.jp

（短単位）を自動付与し，メタ情報として発話者の属性（性別・年齢層・職業・出身地など）と会話の情報（場面・場所など）を整理し付与している。

本稿では，『現日研・職場談話コーパス』の概要と特徴を述べる。

２．『現日研・職場談話コーパス』の概要

『現日研・職場談話コーパス』は，以下の 2 つの調査研究(a)と(b)より得た談話の文字化テキストを元に作成したものである。

(a)『女性のことば・職場編』

1993 年 9 月から 10 月にかけて，現代日本語研究会が首都圏の有職女性 19 名（20 代～50 代）を調査協力者として，それぞれの職場での自然談話を録音した。録音方法はレコーダーを首から下げたり，近くに置いたりしてもらい，行った。19 人の職場は皆異なる。職場に着いてからの朝の 1 時間，会議・打ち合わせの 1 時間，休憩の 1 時間を録音したうちから，おのおの 10 分前後のまとまった談話を選択し，文字起こしした。その文字起こしデータを収録した CD－ROM と，それにもとづく研究論文 10 本が『女性のことば・職場編』（現代日本語研究会編）としてひつじ書房から刊行された（現在品切れ，下記合本参照）。

(b)『男性のことば・職場編』

1999 年 10 月から 2000 年 12 月にかけて，現代日本語研究会が首都圏の有職男性 21 名（20 代～50 代）を調査協力者として，それぞれの職場での自然談話を録音した。録音方法は前回と同じである。21 人の職場は皆異なる。前回同様に，職場に着いてからの朝の 1 時間，会議・打ち合わせの 1 時間，休憩の 1 時間を録音したうちから，おのおの 10 分前後のまとまった談話を選択し，文字起こしした。その文字起こしデータを収録した CD－ROM と，それにもとづく研究論文 12 本が『男性のことば・職場編』（現代日本語研究会編）としてひつじ書房から刊行された（現在品切れ，下記合本参照）。

2011 年，上記 2 つの書籍内容・CD-ROM データを合わせ，『合本　女性のことば・男性のことば（職場編）』(現代日本語研究会編 2011)としてひつじ書房から刊行され，現在も販売されている[1]。また，引き続き，日常場面での会話の収集・調査が行われている(現代日本語研究会編 2016)。

現在，国語研にて構築を進めている『日本語日常会話コーパス』は，収録者自身に収録機器を預け，自然な談話を収録してもらっているものである。これと同様に，職場での会話を調査協力者自身に録音してもらい，自然な談話を収録するという方法を，すでに 1990 年代にいち早く行っていたということになる。非常に先駆的な試みであると言える。

そこで得られた談話データは，当時よりたいへん画期的なものであると評価されているデータである。これら談話データを分析し，現代日本語研究会編(2011)では，男性語や女性語と呼ばれる，性別の違いにあわせて使われる言葉や，使われる傾向のある言葉は，現状ではそのような区別がなくなりつつあるということが，明らかにされている。また，いわゆる書き言葉とは異なる話し言葉の実態が様々にとらえられている。

[1] 『現日研・職場談話コーパス』の中納言版は，前後最大 300 文字ずつまでが表示可能である。文脈を確認したい場合などは，この CD－ROM に収録されている元データを参照されたい。

２．１　『現日研・職場談話コーパス』のデータ仕様

２．１．１　ファイル名

元データは，『女性のことば・職場編』，『男性のことば・職場編』，それぞれ一つのテキストファイルで提供されている。本コーパスでは，新たに次のファイル命名規則に基づき分割し，ファイル名を付与している。

例：　F　01 A　01 1
　　(1)　(2) (3) (4) (5)

表１　ファイル名の表すもの

(1)	女性／男性	F,M	F:『女性のことば・職場編』出典データという意味
			M:『男性のことば・職場編』出典データという意味
(2)	協力者コード	01,02,…	元データと同じ調査協力者の識別コード
(3)	場面１	A,K,Q	元データの「朝」「会議」「休憩」の別を示す
(4)	場面２	01,02,…	連番: 新規に付与　[場面１か２が変わる毎に付与]
(5)	場所	1,2,…	連番: 新規に付与　[場所が変わる毎に付与]

２．１．２　メタ情報

元データに付与されている項目から，下記の項目を収録している[2]。

◆会話情報◆
＜場面１＞…「朝」「会議」「休憩」の別。
＜場面２＞…「場面１」の細分類。「挨拶」「院生の指導」「客との対応」「雑談」「仕事」など，69 種あり。
＜調査日＞…調査した年月。ただし，『女性のことば・職場編』出典データのみ。『男性のことば・職場編』出典データはすべて「＊」となっている。1999 年 10 月から 2000 年 12 月の間である。
＜場所＞…会話の場所。ただし，『女性のことば・職場編』出典データのみ。「室内」「廊下」「うなぎ屋」「路上」「店先」「店内」「会社内」「不明」。『男性のことば・職場編』出典データはすべて「＊」となっている。
＜会話参加者数＞…1 人から最大 12 人まで。この数値は元データから算出して新たに会話単位に付与したものである。

◆話者情報◆
＜発話者コード＞…発話者の識別コード。元データでは，発話者コードが，『女性のことば・職場編』では「01Ａ」，『男性のことば・職場編』では「０１Ａ」，のように，数字部分に全角と半角とが用いられている。本コーパスでは，発話者コードはすべて半角にした。また，「M01A」や「F01A」のように，先頭に F（『女性のことば・職場編』出典データという意味）あるいは M（『男性のことば・

[2] 元データには，直前文の話者との関係，相手の情報，相手との関係，職場の規模など，『中納言』には収録してない情報も付与されている。それらは，現代日本語研究会編(2011)を参照されたい。

職場編』出典データという意味）を付与して，元データを区別する。「M」や「F」は，話者の性別を表示するものではなく，元データがどちらの出典のものであるかを区別するものであることに注意が必要。なお，元データの発話者コードに含まれている不明を表す全角のクエスチョンは全角のままになっている（「＊」の使用はない）。また，元データにて，数字とアルファベットの組み合わせではなく，「F03 男」「F04 多」「お客①」「他者(女)」などとなっているものは，そのまま用いている。

＜性別＞…発話者の性別。「男」「女」「？」「＊」が入力されている。『女性のことば・職場編』出典データは不明が「？」であり，『男性のことば・職場編』出典データは不明が「＊」である。両出典データともに，個人が特定できていないものは空白である。

＜年齢層＞…発話者の調査当時の年齢層。『女性のことば・職場編』出典データでは，わかる場合は具体的な年齢が入力されている。9 歳から 60 代まである。『男性のことば・職場編』出典データは 10 年刻みになっている。10 代から 70 代まである。両出典データともに，個人が特定できていないものは空白である。

＜職業＞…発話者の職業。53 種ある。「アルバイタ」と「アルバイター」のゆれなども含んでいる。『女性のことば・職場編』出典データは不明が「？」になっている。両出典データともに，未調査は「＊」であり，個人が特定できていないものは空白である。半角は全角にした。

＜職種＞…発話者の職種。83 種ある。職業と重なるものもある。「？」，「＊」，空白は上に同じ。半角は全角にした。

＜役職＞…発話者の役職。53 種ある。ここにも「アルバイト」と「アルバイター」がある。フェイスシートに役職がないことが明示してある場合は「（なし）」と入力。ただし，「（なし：一般職）」は別にある。また，「無」も別にある。「？」，「＊」，空白は上に同じ。

＜出身＞…発話者の出身都道府県。『女性のことば・職場編』出典データは未調査のため，すべて「＊」となっている。『男性のことば・職場編』出典データも不明は「＊」である。

＜最長居住地＞…発話者の 4 歳～15 歳の最長居住都道府県（≒言語形成地）。『女性のことば・職場編』出典データは未調査のため，すべて「＊」となっている。『男性のことば・職場編』出典データも未調査は「＊」である。

２．１．３　会話データ

下記のとおり，元データと異なる点があることに注意を要する。

- 半角は全角にした。（例：That→Ｔｈａｔ）
- 「［名字］さん」における［名字］のように伏せ字された要素は，全体を一つの単位とし，「伏せ字」という品詞を付与している。
- ＜笑い＞＜間 7 秒＞＜咳ばらい＞＜独り言＞など，元データに付与されている言語情報以外の要素については，除外している。
- 元データにある，上昇「↑」，下降「↓」，発話途中で次の話者の始まった時点の「★」，重なった部分の始まり「→」と終わり「←」は，いずれも除外している。

- 元データにある，疑問下降の「？」と，聞き取り不明の「＃」はそのまま残している。
- 「相づちなどの挿入要素」は，包含する発話から独立させ，本来の発話場所とは異なる位置（原則，直後）に記述している。

上記「相づちなどの挿入要素」の処理について補足する。例えば，下記のような相づち（網掛け部分）を含む例の場合，『中納言』の検索結果では，図 1 のように表示される。

例：F01A021 の[元データ]

F01B　やっぱりさ、（うん　Inf(女)）どちらかってゆうとさ、こうやってぱっと見たときさ、（うん　Inf(女)）目ってこっちを見ていない↑

F01A　うん。

前文脈	キー	後文脈
#やっぱりさ、どちらかってゆうとさ、こうやってぱっと見たときさ、目ってこっちを見てい	ない	#うん#うん#うん。

図 1　F01A021 の相づちの挿入要素のある部分の『中納言』の検索結果

図 1 で前文脈と後文脈にある半角の「#」は発話単位区切り記号である。検索結果の画面では，後文脈では「うん」という発話が 3 度繰り返しているようにしか見えない。しかしながら，『中納言』にある「詳細な文脈情報表示」[3]という機能を使うと，「発話者コード」の欄を見ることにより，最初の二つの「うん」は「Inf(女)」の発話であることがわかるため，相づちの挿入要素らしいとあたりをつけることはできるようにはなっている。

詳細な文脈情報

会話 ID	連番	書字形出現形	語彙素読み	語彙素	語彙素細分類	品詞	活用型	活用形	発音形出現形	語種	原文文字列	発話者コード	性別	年齢層	職業	職種	役職	出身	最長居住地
F01A021	2690	目	メ	目		名詞-普通名詞一般			メ	和	目	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2700	って	ッテ	って		助詞-副助詞			ッテ	和	って	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2710	こっち	コチラ	此方		代名詞			コッチ	和	こっち	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2720	を	ヲ	を		助詞-格助詞			オ	和	を	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2730	見	ミル	見る		動詞-非自立可能	上一段-マ行	連用形-一般	ミ	和	見	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2740	て	テ	て		助詞-接続助詞			テ	和	て	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2750	い	イル	居る		動詞-非自立可能	上一段-ア行	未然形-一般	イ	和	い	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2760	ない	ナイ	ない		助動詞	助動詞-ナイ	終止形-一般	ナイ	和	ない	F01B	女	31	会社員	社長秘書・一般事務	？	＊	＊
F01A021	2770	うん	ウン	うん		感動詞一般			ウン	和	うん	Inf(女)							
F01A021	2780	うん	ウン	うん		感動詞一般			ウン	和	うん	Inf(女)							
F01A021	2790	うん	ウン	うん		感動詞一般			ウン	和	うん	F01A	女	28	会社員	イベント企画開発	無	＊	＊
F01A021	2800	。		。		補助記号-句点				記号	。	F01A	女	28	会社員	イベント企画開発	無	＊	＊

図 2　F01A021 の相づちの挿入要素のある部分の『中納言』の「詳細な文脈情報表示」

[3] 検索結果の「会話 ID」をクリックして表示する。

２．２.『現日研・職場談話コーパス』のデータ量

はじめに，文字化テキストの例(F01A011 の冒頭)を図 3 に示す。

会話 ID：F01A011
調査日：1993 年 10 月　場面１：朝　場面２：電話　場所　室内
会話参加者数：1
発話者コード：F01A　性別：女　年齢層：28　職業：会社員
職種：イベント企画開発　役職：無　出身：＊　最長居住地：＊

はい、お電話代わりました。
はい、お世話になっております。
はい。
はい。
はい、受け取っております。
はい。
ええ。
ええ。
ええ。
はい。
はい。
わかりました。
じゃ、これーはお受けします。

図 3　F01A011 の冒頭の文字化テキスト

上記のような文字化テキストの全体のデータ量は表 2 のとおりである[4]。

表 2 『現日研・職場談話コーパス』の全体

ファイル数	1,324
会話数	22,372
語数(全て)	248,677
語数(記号等除外・全て)	186,906

以下，次のメタ情報のうち主なものについてのデータ量を示す。

会話情報：　場面１，場面２，調査日，場所，会話参加者数
話者情報：　性別，年齢層，職業，職種，出身，最長居住地

[4] 本コーパスには，『女性のことば・職場編』の＜通番＞1,571～1,718 までの 148 行分の会話は含まない。

２．２．１　会話情報①：場面１，場面２

表 3 に場面１の，表 4 に，場面１と２別の語数(記号等除外・全て)の内訳を示す。以降，語数は「記号等除外・全て」のものを示す。

表 3　場面１のファイル数と語数(記号等除外・全て)

場面１	ファイル数	語数
朝	550	52,773
会議	351	68,613
休憩	423	65,520

表 4　場面２と場面１の語数(記号等除外・全て)

場面２	場面１朝	場面１会議	場面１休憩	語数合計
コンピュータの操作方法の相談と説明		1,495		1,495
シャンプー中の応答	260			260
スタッフルームでの雑談	198			198
パソコン操作の指導と相談			1,049	1,049
ブロー中の応答	1,033			1,033
ミーティング・報告	2,339			2,339
レジでの応答	14			14
挨拶	173	24	78	275
挨拶(電話)	211	25	4	240
院生の指導			2,402	2,402
応対	419	214		633
応対（説明）	556			556
会議		8,099		8,099
客との応対	404	23		427
客との対応	48			48
休憩時雑談	9,986	3	10,492	20,481
教師生徒の会話	6			6
業務電話	44			44
検討会		410		410
研究室会議		3,260		3,260
講義	2,187			2,187
雑談	8,144	3,462	23,750	35,356
雑談（パソコン）			451	451
雑談（パソコンの記憶媒体）			505	505
雑談（レストランの食事）			374	374
雑談（交通規制）			271	271
雑談（自転車）			200	200
雑談（朝食）			68	68
雑談（徹夜）			100	100
雑談（転居）			39	39
雑談(電話)	174			174
仕事	200			200
仕事（応対）		993		993
仕事（相談）	321			321
仕事（打合せ）		172		172

場面２	場面１朝	場面１会議	場面１休憩	語数合計
仕事（打合せ？）	234			234
仕事の話	576			576
仕事上の確認			147	147
仕事中の雑談	1,559			1,559
始業前雑談	353			353
指導		62		62
取引先との電話折衝		866		866
出張報告		2,244		2,244
商品管理業務	1,276			1,276
小会議		10,009		10,009
接客と応答			54	54
相談	3,950	359	1,430	5,739
相談(電話)	238			238
打合せ	10,645	23,643	5,124	39,412
打合せ（商談）		1,459		1,459
打合せ（説明）		2,102		2,102
打合せ(電話)	2,668	12	108	2,788
大会議		3,770		3,770
昼食時雑談			17,967	17,967
昼食時雑談・電話			3	3
朝礼	1,791			1,791
電話	692	225	241	1,158
電話（打合せ）	50			50
電話・雑談	21		43	64
電話・打合せ	1,865		620	2,485
電話依頼	91			91
電話引き継ぎ		50		50
電話取り次ぎ		7		7
電話取り次ぎ(電話)		11		11
独り言	32			32
反省会		1,618		1,618
報告		3,946		3,946
《その他》	15	9		24
《不明》		41		41
語数合計	52,773	68,613	65,520	186,906

表 3 に示したとおり，朝，会議，休憩はおおよそ同じくらいのデータ量である。表 4 に示した場面２の分類は多岐にわたっている。そこで，小磯ほか(2016)で用いている会話の形式 4 タイプ＋そのほかという 5 分類に分類しなおした結果の内訳を次の表 5 に示す。

表 5　場面２の 5 分類と場面１の語数(記号等除外・全て)

会話のタイプ	場面１ 朝	場面１ 会議	場面１ 休憩	語数合計
雑談	20,848	3,514	54,302	78,664
用談・相談	17,015	12,547	3,692	33,254
会議・会合	12,670	52,440	5,124	70,234
授業・レッスン・講演	2,193	62	2,402	4,657
そのほか	47	50		97
語数合計	52,773	68,613	65,520	186,906

表 5 でみると，データ量が多いのは，場面１休憩の雑談と，場面１会議の会議・会合であることがわかる。場面１の合計では，雑談と会議・会合がおおよそ同じくらいのデータ量である。フォーマルな場面とそうでない場面の会話がそれぞれ十分な量とれていることがうかがえる。なお，場面２の再分類の内訳は，表 6 のとおりである。

表 6　場面２の 5 分類

場面２	5分類
コンピュータの操作方法の相談と説明	用談・相談
シャンプー中の応答	用談・相談
スタッフルームでの雑談	雑談
パソコン操作の指導と相談	用談・相談
ブロー中の応答	用談・相談
ミーティンク・報告	用談・相談
レジでの応答	用談・相談
挨拶	雑談
挨拶(電話)	雑談
院生の指導	授業・レッスン・講演
応対	用談・相談
応対（説明）	用談・相談
会議	会議・会合
客との応対	用談・相談
客との対応	用談・相談
休憩時雑談	雑談
教師生徒の会話	授業・レッスン・講演
業務電話	用談・相談
検討会	会議・会合
研究室会議	会議・会合
講義	授業・レッスン・講演
雑談	雑談
雑談（パソコン）	雑談
雑談（パソコンの記憶媒体）	雑談
雑談（レストランの食事）	雑談
雑談（交通規制）	雑談
雑談（自転車）	雑談
雑談（朝食）	雑談
雑談（徹夜）	雑談
雑談（転居）	雑談
雑談(電話)	雑談
仕事	用談・相談
仕事（応対）	用談・相談
仕事（相談）	用談・相談
仕事（打合せ）	会議・会合

場面２	5分類
仕事（打合せ？）	会議・会合
仕事の話	用談・相談
仕事上の確認	用談・相談
仕事中の雑談	雑談
始業前雑談	雑談
指導	授業・レッスン・講演
取引先との電話折衝	用談・相談
出張報告	用談・相談
商品管理業務	用談・相談
小会議	会議・会合
接客と応答	用談・相談
相談	用談・相談
相談(電話)	用談・相談
打合せ	会議・会合
打合せ（商談）	会議・会合
打合せ（説明）	用談・相談
打合せ(電話)	用談・相談
大会議	会議・会合
昼食時雑談	雑談
昼食時雑談・電話	雑談
朝礼	会議・会合
電話	用談・相談
電話（打合せ）	雑談
電話・雑談	用談・相談
電話・打合せ	用談・相談
電話依頼	用談・相談
電話引き継ぎ	用談・相談
電話取り次ぎ	用談・相談
電話取り次ぎ(電話)	用談・相談
独り言	そのほか
反省会	会議・会合
報告	用談・相談
《その他》	そのほか
《不明》	そのほか

表 7　電話か否かの語数(記号等除外・全て)

電話か否か	語数
電話	7,217
電話以外	179,689
語数合計	186,906

本コーパスは，電話の会話も多く収録されている。電話全体の語数は，表 7 に示したとおりである。

２．２．２　会話情報②：場所，会話参加者数

次に，表 8 に場所の，表 9 に会話参加者数のファイル数と語数の内訳を示す。職場を中心に収録されたものであるため，室内での会話がほとんどである。会話参加者数は，3 人のものがもっとも多い。

会話参加者数 1 は，先の表 7 に示した電話の会話のほか，下記の表 10 に示したような，「独り言」(a)や，その場に発話相手はいるが，相手が一言も発しない場合の会話に付与されているもの(b)や(c)である。

表 8　場所のファイル数と語数

場所	ファイル数	語数
室内	598	79,723
廊下	35	2,305
うなぎ屋	8	1,468
路上	7	589
店先	6	664
店内	4	104
会社内	3	393
《不明》	2	6
*	661	101,654
合計	1,324	186,906

表 9　会話参加者数のファイル数と語数

会話参加者数	ファイル数	語数
1	136	11,353
2	270	33,446
3	298	49,160
4	216	28,258
5	138	22,690
6	102	16,750
7	35	2,870
8	46	8,938
9	60	10,843
10	9	1,262
12	14	1,336
合計	1,324	186,906

表 10　会話参加者数 1 の電話以外の会話例

項目	元データの通番	ファイル名	会話データ（元データ）
(a)	1982	F05A011	ここらにおいとけばいいんだな。＜独り言＞
(b)	1983	F05A021	今日［名前］ちゃん来たらさあ、これ、これをさ、こっちで、こっちで、うってもらって、ってゆうか、あいだにいっぱいはいるんでねえ、あたしうち始めちゃってもいいんだけど、＃＃＃が来る、あの、あれ、やらなくちゃいけないからとりあえず。＜間＞
(b)	1984	F05A021	これもやんなくちゃいけない、今日、リーダーも。＜間＞
(b)	1985	F05A021	どれをさきにやるかなあ。＜間＞
(c)	2002	F05A041	ともかくあれをやっちまわないと、うんん。＜咳ばらい＞＜間２８秒＞

表 11　12 人の会話の内訳(M05A071)

発話者コード	会話数
M05B	13
M05E	6
M05F	9
M05L	12
M05M	6
M05P	4
M05Q	7
M05R	4
M05S	3
M05T	2
M05U	6
M05W	9
M05 δ	4
M05 ε	17

表 12　10 人の会話の内訳(M18Q011)

発話者コード	会話数
M18A	60
M18B	15
M18C	8
M18D	1
M18E	16
M18G	2
M18I	2
M18J	32
M18K	1
M18L	1

会話者数最多 12 人の会話は，朝の朝礼時の会話である。また，次に多い 10 人の会話は，休憩の雑談時の会話である。発話者別の会話数の内訳は表 11 と表 12 に示すとおりである。12 人の会話に参加しているのは，M05B から M05W の 12 人である。M05δ と M05ε は，その場にいる人であるが特定できなかったため，別のコードが付与されているものである。10 人の会話のうち，M18L は，元データで「@＜笑い　複数＞」となっている会話であり，『中納言』では除外されている。よって，実質は実は 9 人である。

２．２．３　話者情報：性別，年齢層

最後に，表 13 に性別の，表 14 に年齢層別のファイル数と語数の内訳を示す。男性と女性はおおよそ同じくらいのデータ量である。年齢層は，職場を中心とした収録であるため，20 代から 50 代のデータが多く，最も多いのは 30 代である。

表 13　性別のファイル数と語数

性別	ファイル数	語数
男	596	96,657
女	450	86,419
?	6	343
*	53	692
(空白)	219	2,795
合計	1,324	186,906

表 14　年齢層別のファイル数と語数

年齢層	ファイル数	語数
～9	4	57
10代	11	319
20代	256	48,407
30代	292	51,907
40代	265	48,694
50代	152	27,020
60代	32	4,447
70代	6	463
?	84	2,749
*	3	48
(空白)	219	2,795
合計	1,324	186,906

３．『現日研・職場談話コーパス』の特徴

各種語彙表を用いて,『現日研・職場談話コーパス』の話し言葉としての特徴を概観する。

３．１　上位語

同じく会話が収録されているが，すべて雑談である『名大会話コーパス』と，書き言葉の代表として『現代日本語書き言葉均衡コーパス』（以下,『BCCWJ』）の語彙表より，上位語を比較する。比較の結果を表 15 に示す。

表 15　『名大会話コーパス』『現日研・職場談話コーパス』『BCCWJ』の上位語

順位	名大会話コーパス			職場会話コーパス			BCCWJ		
	語彙素読み	語彙素	品詞	語彙素読み	語彙素	品詞	語彙素読み	語彙素	品詞
1	ダ	だ	助動詞	ダ	だ	助動詞	ノ	の	助詞-格助詞
2	ウン	うん	感動詞-一般	ノ	の	助詞-準体助詞	ニ	に	助詞-格助詞
3	タ	た	助動詞	テ	て	助詞-接続助詞	テ	て	助詞-接続助詞
4	テ	て	助詞-接続助詞	ネ	ね	助詞-終助詞	ハ	は	助詞-係助詞
5	ネ	ね	助詞-終助詞	**デス**	**です**	**助動詞**	ダ	だ	助動詞
6	ノ	の	助詞-準体助詞	ノ	の	助詞-格助詞	ヲ	を	助詞-格助詞
7	カ	か	助詞-副助詞	タ	た	助動詞	タ	た	助動詞
8	ト	と	助詞-格助詞	ハ	は	助詞-係助詞	スル	為る	動詞-非自立可能
9	ノ	の	助詞-格助詞	ニ	に	助詞-格助詞	ガ	が	助詞-格助詞
10	モ	も	助詞-係助詞	ト	と	助詞-格助詞	ト	と	助詞-格助詞
11	デ	で	助詞-格助詞	ガ	が	助詞-格助詞	デ	で	助詞-格助詞
12	ガ	が	助詞-格助詞	デ	で	助詞-格助詞	モ	も	助詞-係助詞
13	ニ	に	助詞-格助詞	モ	も	助詞-係助詞	イル	居る	動詞-非自立可能
14	ハ	は	助詞-係助詞	イウ	言う	動詞-一般	マス	ます	助動詞
15	ソウ	そう	副詞	ヨ	よ	助詞-終助詞	ノ	の	助詞-準体助詞
16	イウ	言う	動詞-一般	スル	為る	動詞-非自立可能	アル	有る	動詞-非自立可能
17	ッテ	って	助詞-副助詞	ソウ	そう	副詞	**デス**	**です**	**助動詞**
18	**ナニ**	**何**	**代名詞**	テル	てる	助動詞	イウ	言う	動詞-一般
19	テル	てる	助動詞	ウン	うん	感動詞-一般	コト	事	名詞-普通名詞-一般
20	ヨ	よ	助詞-終助詞	カ	か	助詞-副助詞	ナイ	ない	助動詞

表 15 において赤字が,『名大会話コーパス』と『現日研・職場談話コーパス』にあり,『BCCWJ』にはない。これらは話し言葉としての特徴を表す語と言えるだろう。また,『名大会話コーパス』の 18 位「何」は,『現日研・職場談話コーパス』では 28 位であり,『BCCWJ』では 62 位であることから，これも話し言葉としての特徴を表し得ると考えられる。一方,『BCCWJ』で 17 位である「です」は,『名大会話コーパス』にないが,『現日研・職場談話コーパス』では 5 位と非常に上位に位置する点が目立つ。雑談以外のフォーマルな会話を多く含む『現日研・職場談話コーパス』の特徴を表すと言えそうである。

３．２　LLR（対数尤度比）

『BCCWJ』の書籍の出版サブコーパス（PB）と，特定サブコーパスの一つ yohoo!知恵袋（OC),『名大会話コーパス』，主に学会講演が収録されている『日本語話し言葉コーパス』（CSJ）をベースに，『現日研・職場談話コーパス』の LLR（対数尤度比）により，

特徴語を求めた。表 16 に上位 10 語を示す。感動詞が多く入っていることがわかる。

また，『名大会話コーパス』をベースにした際に 3 位に入った「ゼロ」が目を引く。語彙素読み「ゼロ」の用例は全部で 406 件あり，ほとんどが数字の「０」の用例であった。数字の話が多いというのも，『現日研・職場談話コーパス』の特徴の一つかと推測される。また，書字形「ゼロ」の用例は全部で 18 件あった。そのうち 8 例を表 17 に示す。これらも職場の会話らしい用例と言えそうなものである。

表 16　『現日研・職場談話コーパス』の LLR 上位語

ベース	1	2	3	4	5	6	7	8	9	10
BCCWJ（PB）	ね	うん	あの	はい	てる	って	です	よ	で	そう
BCCWJ（OC）	うん	あの	はい	ね	そう	あっ	で	ええ	えー	ああ
名大会話コーパス	です	ます	ゼロ	はい	えー	あの	えーと	此れ	御早う	が
CSJ	よ	うん	あっ	ああ	はい	まあ	えー	ね	さ	の

表 17　『現日研・職場談話コーパス』の LLR 上位語「ゼロ」の用例

会話 ID	前文脈	キー	後文脈
F03A021	#うん#ここを	ゼロ	にして。#うん、あ、ここに、
F05K011	#お金はかからないわね。#	ゼロ	です。#うん
F06K011	あの、キャリアーとしての経験は	ゼロ	ですからー、
F14Q021	#それで、あのー、実施は	ゼロ	だった人がなん人かいたんですね。
F14Q021	#それーはー。#あの前期まったく	ゼロ	だったんで。#２、３人いるんですけど、
F14Q021	#でなんか［名字］さんもなんかまったく	ゼロ	だとなんかちょっといいわけー、ってゆうか
F14Q021	あの出勤簿とかもまったく空欄になりますよね、ぜ、	ゼロ	とかじゃなくてね。#わかりました。＃＃＃。
M12Q101	#ひさびさに	ゼロ	から組んだけど、＃＃＃＃。

３．３　品詞の分布

続いて，同じく『BCCWJ』の語彙表を用いて，『現日研・職場談話コーパス』と品詞の分布を比較する。それぞれの品詞の分布を図 4 と図 5 に示す。

図 4 と図 5 を比べると，赤の四角形で強調している通り，『現日研・職場談話コーパス』は名詞が少なく，感動詞，副詞，代名詞が多い。これは，『名大会話コーパス』と『BCCWJ』の比較(柏野ほか 2017)と同じ傾向であり，話し言葉としての特徴が表れていると考えらえる。

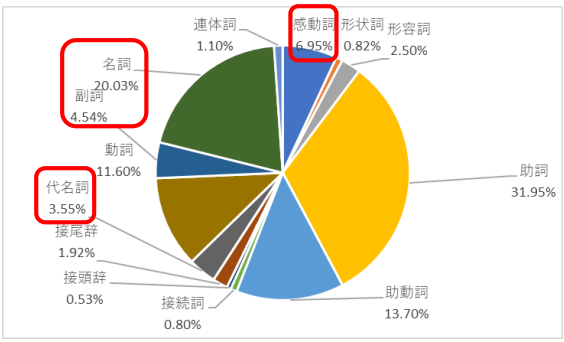


図 4　『現日研・職場談話コーパス』の品詞の分布

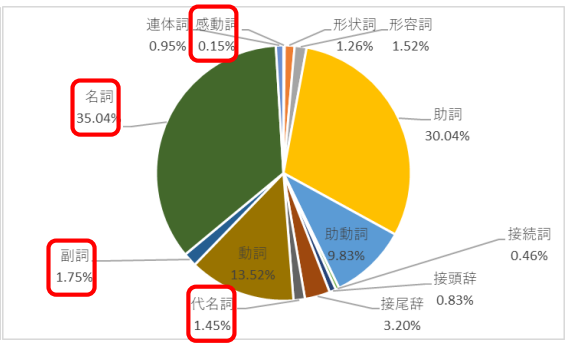


図 5　『BCCWJ』の品詞の分布

３．４　そのほかの特徴語

柏野ほか(2017)では，『BCCWJ』では用例が得にくいが，『名大会話コーパス』で頻出することが期待されるそのほかの特徴語として，表 19 に示す俗語的な用法のある a)〜h)の 8 語を取り上げた。今回，『現日研・職場談話コーパス』について調べた結果と比較して，表 19 にまとめて示す[5]。なお，改めて両コーパスの主な違いを表 18 に示す。

表 18　『現日研・職場談話コーパス』と『名大会話コーパス』の相違点

	『現日研・職場談話コーパス』	『名大会話コーパス』
語数	186,906	1,131,971
収録期間	1993-2000	2001-2003
会話の形式	雑談/用談・相談/会議・会合/ 授業・レッスン・講演/ほか	雑談のみ

表 19　『現日研・職場談話コーパス』と『名大会話コーパス』の用例検索結果

項目	語	『職場』 用例数	『職場』 ＰＭＷ	『名大』 用例数	『名大』 ＰＭＷ	検索方法
a)	微妙	1	5.4	156	109.8	語彙素「微妙」
b)	やば	14	74.9	168	118.2	文字列「やば」
c)	まじ	3	16.1	197	138.6	語彙素「まじ」
d)	無理	33	176.6	273	192.1	語彙素「無理」
e)	てか、	3	16.1	60	42.2	文字列「てか、」
f)	すごい＋形容詞	10	53.5	344	242.0	書字形出現形「すごい」＋形容詞
g)	うける	2	10.7	37	26.0	職場：文字列「うける」 名大：語彙素「受ける」の終止形
h)	みたいな	30	160.5	473	332.8	文字列「みたいな[、。？]」

表 18 に示したとおり，延べ語数は，『現日研・職場談話コーパス』は『名大会話コーパス』の約 7 分の 1 である。そこで，表 19 での用例数の比較は PMW を用いる。a)〜h)の 8 語は話し言葉のなかでも俗語的な言い方であるため，雑談以外の会話が収録されている『現日研・職場談話コーパス』では，全体的に『名大会話コーパス』よりも少ない値になっている。また，少しだけ『現日研・職場談話コーパス』の収録時期が『名大会話コーパス』よりも前になるため，今ある俗語的な用法は，まだ『現日研・職場談話コーパス』にはそうなかったのかもしれない。

たとえば，次の表 20 に実際の用例を示すが，a)「微妙」は，俗語的ではない用法の用例が 1 例あるのみである。俗語的には応答で「それ，びみょー」などと言うのを聞くが，平仮名表記，カタカナ表記でも『現日研・職場談話コーパス』に該当例はなかった。g)「受ける」も同様である。やはり俗語的に応答で「それ，うけるー」などと言うのを聞くが，その該当例はなかった。なお，両コーパスともに，d)「無理」の用例も，現在ほどの俗語的な，応答詞的に用いるような例はみられない。

[5] 表 19 の検索結果数は，当該語の「話し言葉」ならではの用法例の正確な件数ではない。検索もれ，あるいは，別語，別用法の例が少々混じっている。

ただ，それ以外では数は少ないながらも，『名大会話コーパス』と同様に，現在耳にするような俗語的な言い方の用例が得られている。表 20 に示す。なお，会話 ID が「data」から始まるものは『名大会話コーパス』の用例である。

表 20　『名大会話コーパス』と『現日研・職場談話コーパス』の用例

会話 ID	前文脈	キー	後文脈
data077	この子は、Ｅ短の子だよ。あっ、そうなんだー、	微妙	、微妙。うんそういうのばっかり。
M06Q031	まー、驚くことが多いって話よー。	微妙	なニュアンスで教えてくれてー。
data072	なかなか時間がないんだよね。ねーあたしもだよ。	やばー	い。あっ、ＴＯＥＩＣさ、こないだあったけど、
M12Q031	#あの、これやうめーや、ちょっと、	やばい	んじゃない。#このランチメニューー。
data011	５級だったしね、一番最初受けたの。	まじ	？６年のときに５級。
M21Q011	３時までさー、ずっーとしゃべってて。#	まじ	でー。#まじ、もーあたし、その前の日とかに、
data046	うーん。無理。だから、なんで。とにかく	無理	。それは、そういうことしたら、
M21K011	#夜は	無理	っす、平日の夜は無理っす。
data056	うんどこで見たの。	てか、	あの、日本に来たときの。
M12Q101	#うん。#いや、	てか、	自動なんだよー、もー。
data103	んー、かっこいい。＊＊＊。	すごい	(かわいい)、この絵。
F11Q011	#すごい、なんだっけそれ。#	すごい	(おいしい)やつ。
data065	あ、君は日本文学専攻か、ふーん、とか言って。	受ける	ー。うーん話しかけやすい雰囲気なんじゃん。
F15K011	#だって、みんな、	うける	ものねー、あれ、すごく。#うけますねー。
data085	今日、さむ。な、何となく寒そう、	みたいな？	うん。雪が降ったときとか、
F15Q011	#でしょ#でビール、がーん、みたいな	みたいな。	#もっとほかのお母さまの意見も聞いた方が＃＃＃。

遠藤(2011)は，俗語，若者ことば，流行語として次の語の用例が『男性のことば・職場編』にあると指摘している。

やばい，おかまちっく，ぽい（の新しい用法）[6]，まじ・まじで

6 「ぽい」は本来名詞や動詞の連用形につくが，形容詞や動詞の終止的につく新しい用法例を指摘している。「かっこいいっぽい」「コピーしてるっぽい」の例があげられている。

このうち，「やばい」については用例があるということのみ指摘しているが，表 20 に示したとおり，ランチに対して「うめー」「やばい」と発話しており，すでにポジティブの「やばい」の用例が収録されている点は注目に値する。

さらに，「今どきの日本語」については，遠藤編(2018)にくわしく報告されている。

４．おわりに

『現日研・職場談話コーパス』の概要と特徴を述べた。書き言葉コーパスである『BCCWJ』と比べ，雑談を収録した『名大会話コーパス』同様に，「うん」，終助詞の「ね」「か」「よ」や，副詞の「そう」が頻出し，また，感動詞，副詞，代名詞が多く，話し言葉の特徴語の用例が多く得られるコーパスであることを示した。さらに，『名大会話コーパス』が雑談のみであるのに対し，『現日研・職場談話コーパス』は用談・相談，会議・会合，授業・レッスン・講演などの会話を含むため，「です」が頻出する点が『名大会話コーパス』とは異なる特徴であることを示した。そして，LLR を調べることにより，「ゼロ」のような仕事上の会話ならでは現れやすい語があることを示した。また，俗語的なものは，『名大会話コーパス』ほどではないが，ある程度収録されていることも示した。

『中納言』で公開するに際し，貴重な『現日研・職場談話コーパス』のデータが，今後さらにさまざまな研究に活用されることが期待される。

謝　辞

本研究は国立国語研究所の機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）の研究成果を報告したものです。また，オリジナルのデータは，現代日本語研究会による研究成果です(現代日本語研究会編 2011)。遠藤織枝先生，髙﨑みどり先生，髙橋美奈子先生をはじめとする現代日本語研究会のみなさまと，出版元のひつじ書房に感謝申し上げます。そして，『中納言』版の構築と公開に際しては，形態素解析結果の人手修正をはじめ，多くのみなさまにご協力いただきました。みなさまに感謝いたします。

文　献

遠藤織枝(2011)「第２章　男性のことばの文末」現代日本語研究会編『合本　女性のことば・男性のことば（職場編）』pp.33-45, ひつじ書房.

遠藤織枝編(2018)『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.

柏野和佳子・西川賢哉・小磯花絵(2017)「『名大会話コーパス』中納言版・ひまわり版公開データの作成」『言語資源活用ワークショップ 2016 発表論文集』pp.324-335.

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

小磯花絵・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(2018)「『日本語日常会話コーパス』構築状況と予備的分析」『言語処理学会第 24 回年次大会発表論文集』pp.889-892.

藤村逸子・大曽美恵子・大島ディヴィッド義和(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. 43-72, ひつじ書房.

関連 URL

国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」 http://pj.ninjal.ac.jp/conversation/

『現日研・職場談話コーパス』 http://pj.ninjal.ac.jp/conversation/shokuba.html

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

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

『現代日本語書き言葉均衡コーパス』語彙表
http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

日本語オノマトペ共起表現レキシコン JMWEL_onomatopoeic

首藤 公昭（福岡大学名誉教授）
田辺 利文（福岡大学）
高橋 雅仁（久留米工業大学）

A Lexicon of Japanese Onomatopoeic Collocations: JMWEL_onomatopoeic

Kosho Shudo (Fukuoka University, professor emeritus)
Toshifumi Tanabe (Fukuoka University)
Masahito Takahashi (Kurume Institute of Technology)

要旨

オノマトペ（擬音語，擬態語）の豊富さは日本語の特徴の一つとされるが，機械翻訳などの言語処理では十分な対応がなされていない．筆者らが開発した，平仮名べた書き見出し，形態素分かち書き，構文機能，形態・構文構造，内部修飾可否情報，文脈条件などを与えた見出し総数約 34,500 の日本語オノマトペ共起表現レキシコン JMWEL_onomatopoeic の概要を紹介する．本レキシコンは，オノマトペと他語のコロケーション集であり，大規模日本語複単語表現レキシコン JMWEL の部分レキシコンである．

１．はじめに

オノマトペ（擬音語，擬態語）は日本語を母語とする人にとって有用・不可欠な語彙資源であるが，従来の機械翻訳等の自然言語処理 Natural Language Processing, NLP では十分な対応がなされていない．近年，繊細・微妙な人の感覚を的確に表現するオノマトペが注目されており，特定領域の文書に出現するオノマトペを調査する研究（高丸ほか 2014，井上ほか 2017）を始め，オノマトペの意味に関する研究（小宮ほか 2016，福島ほか 2014），オノマトペと他語の共起を調べる研究（玉岡ほか 2011，乙武ほか 2016），オノマトペを手掛かりに文書の種別を判定する研究（渡辺ほか 2015）等々，種々の研究が行われている．

いっぽう，今世紀に入り，日常の言語にはコロケーション，決まり文句，慣用表現等の特異表現が予想外に多種，多量に使われていることが重視されるようになり，NLP分野では複単語表現 Multiword Expression，MWE（Sag et al. 2002, Tanabe et al. 2014），言語学では定型言語 Formulaic Language（Corrigan et al. 2009, Jiang et al. 2007），単語連鎖 Lexical Bundles（Biber et al. 1999），構文文法 Construction Grammar（Fillmore et al. 1988）といった枠組みで種々の研究が進められている．筆者の一人は，1960年代からこの種の日本語複単語表現の総括的なレキシコンJapanese MWE Lexicon, JMWELの開発を進めてきた．（Tanabe et al. 2014）

本稿では，JMWEL の一部をなす日本語オノマトペ共起表現レキシコン JMWEL_onomatopoeic（以下，本レキシコンと記す）の概要を紹介する．

本レキシコンの主な特徴は，

ⅰ　オノマトペと他語の共起表現中にギャップ（内部修飾句）が介在する可能性を記載している

ⅱ　オノマトペ（単体），オノマトペ共起表現中の語彙に漢字・片仮名異表記を与えている

iii　オノマトペの連体，連用，動詞化用法を体系化して記載している
点である．

２．収録表現

本レキシコンの見出しは，

（１） オノマトペ（単体）約 3,000 種[1]

（２） オノマトペと他語が共起した日常よく現れる句（以後，オノマトペ共起表現と記す）約31,500種

であり，「悠々」，「懇々」，「生き生き」のような漢語由来，漢字表記可能なもの，「オヤッ」のような感動詞に分類し得るものを含め，新聞，雑誌等の記事，小説，テレビ，ラジオの放送文から内省によって抽出したものを基本として既存の不特定の辞典類（小野 2007，阿刀田ほか 2004）を使って補強したものである．方言，古語や近年殆ど使われなくなったもの，特定年齢層でのみ使われるものは除外されている[2]．

３．記載情報

本レキシコンは，Microsoft Excel で作成した xlsx ファイルに纏められており，1 行に割り当てた 1 個の見出しに対して，A～I 欄に以下の情報を与えている．例えば，「クンクンと犬が鳴く」という表現に対して与えた情報を A～I 欄の順に列挙すれば以下のようになる．（欄の区切りを・・・で，空データを φ で示す．）

クンクン・・・くんくんといぬがなく・・・くんくん-と-いぬ-が-なく・・・クンクン-と-犬-が-鳴く・・・VP・・・[[Oto]*[[*Nga]*V30]]・・・φ・・・φ・・・音

３．１　オノマトペ：A 欄

見出し表現中に使用されているオノマトペを片仮名表記で与える．オノマトペからその共起表現を検索する際に利用できる．

３．２　見出し：B 欄

オノマトペ（単体），オノマトペ共起表現ともに平仮名べた書きで見出しを与える．同音異義，同音異機能オノマトペは原則として別見出しとする．例えば，「ぱらぱら」は擬音と擬態で別見出し，「こんこん」では，擬音とは別に擬態の多義でも別見出しとする．

３．３　分かち書き：C 欄

オノマトペ共起表現に対し，その分かち書きを平仮名表記上にハイフン「-」で区切って与える．分かち書き単位は，単語，接頭語，接尾語，接頭造語要素，接尾造語要素とし，活用語尾は形容動詞語尾「な」，「に」，「たる」，「と」以外は切り離していない．造語要素とは造語機能を持つ拘束形態素であり，多くの場合，「緊張-感」の「感」のように音読みの一漢字である．

複合語は基本的にアンダースコア「_」で要素語に区切っている．

３．４　異表記：D 欄

[1] オノマトペ単体は複単語表現ではないが，便宜上，本レキシコンに含めている．

[2] 「クルリ」に対する「クルリッ」のような末尾が促音化されたものも原則として独立のオノマトペとみなしている．

オノマトペ（単体），オノマトペ共起表現に対して，漢字，カタカナなど，異表記可能な部分には，C 欄の分かち書きの上で，正規表現類似の形式で選択肢を与える．例えば，「ポッチャリ-と-し-た-(身)体_付き」という D 欄の記載は「ポッチャリ-と-し-た-身体_付き」，「ポッチャリ-と-し-た-体_付き」の可能性，「満(満/々)-たる-自信」は，「満満-たる-自信」，「満々-たる-自信」の可能性を表す．ハイフンやアンダースコアで区切られた C，D 欄の記載から種々の表記形を簡単に生成できる，例えば，C 欄の「からだ_つき」と D 欄から得られた「身体_付き」，「体_付き」から「からだつき」，「身体つき」，「身体付き」，「からだ付き」，「体付き」，「体つき」の 6 通りの表記形が得られる．

３．５　構文的機能：E 欄

収録したオノマトペ（単体）は，形態・構文上の機能により，

1　単純オノマトペ
2　連用オノマトペ（副詞的オノマトペ）
3　接頭オノマトペ
4　接尾オノマトペ
5　名詞性オノマトペ

に分類される．

1 は，格助詞「と」を後接して連用修飾機能を持つものである． 本レキシコンでは，「ころっと」などは，オノマトペ「ころっ」と格助詞「と」の共起表現とみなしている．末尾促音型オノマトペの殆どは1 に分類される．

2 は，そのままでも「と」を後接しても連用修飾機能を持つもの，3, 4 は，他語に連接して造語する機能を持つものである．

表1に本レキシコンにおける1〜5 の分布と例を示す．

E欄には表1の記号が記載されている． 1, 2 の機能は，後述するH欄の情報でより詳細化される．

表 1　採録したオノマトペの分布と例

種別, 記号	見出し数	例
単純オノマトペ，O	1,588	ツルリ，ホワッ，ドッカーン，グネッ，ピューン，ゴロリ，ヒヤッ，ドブン
連用オノマトペ, AdvO	1,155	ドッカリ，フラフラ，ミッチリ，フワフワ，チャリチャリ，ゴリゴリ
接頭オノマトペ，Op	148	ドタ，ジリ，グラ，ゴタ，ソヨ，ビリ
接尾オノマトペ，Os	32	タップリ，タラタラ，モリモリ，ピカ
名詞性オノマトペ，NPO	81	ブツブツ，コリコリ，フリフリ，デコボコ
計	3,004	

いっぽう，本レキシコンに収録しているオノマトペ共起表現には，

ｉ　名詞句
ii　動詞句
iii　形容詞句
iv　形容動詞（語幹）句

v　連用修飾句
vi　連体修飾句
vii　名詞文形式
が有る.
表 2 に本レキシコンにおける i ～viiの分布と表現例を示す.
オノマトペ共起表現の E 欄には表 2 の記号が記載されている.

表 2　収録オノマトペ共起表現の分布と例

種別，記号	見出し数	例
名詞句，　NP	3,688	サッパリ-と-し-た-性格，カリッ-と-し-た-口_当り，ホクホク-し-た-食感，サラサラ-し-た-肌_触り，ギリギリ-の-妥協
動詞句，　VP	21,248	ドタッ-と-音-が-する，鼻-先-に-人参-を-ブラ-(下/提)げる，肌-が-パサパサ-に-乾く，馬-が-ヒィーン-と-嘶く，ニャーン-と-(ネコ/猫)-が-(鳴/啼)く，フッ-と-胸-に-浮かぶ，カツカツ-と-靴音-が-する，クラクラ-と-(眩暈/目眩)-が-する，キリキリ-痛む，(深(深/々)/シンシン)-と-夜-が-(更/深)ける
形容詞句，　AP	485	モチモチ-と-柔らかい，ポンポン-と-威勢-が-良い，ガンガン-と-痛い
形容動詞（語幹）句，　AVP	193	愛嬌-タップリ，フンワリ-と-柔らか，ツンツン-と-無_愛想
連用修飾句，　AdvP	3,721	キョトン-と-し-て，ガラガラ-音-を-立て-て，熟(熟/々)-思う-に，ワイノワイノ-と
連体修飾句，　AdnP	2,187	シドロモドロ-の，グチョグチョ-し-た，ガチガチ-な
名詞文形式など， NPS，INC，OPS	15	英語-が-ペラペラ，予定-が-ビッシリ，収支-が-トントン
計	31,535	

３．６　構文構造と内部修飾可能性表示：F 欄

オノマトペ共起表現に対して C 欄のハイフンによる分かち書きに基づき，係り受け構造を修飾子，被修飾子の対をカッコ[ ]で括って記載する．すなわち，句 α の主辞が句 β の主辞を修飾して出来た句 αβ の構造記述を α，β の構造記述 a, b を使って[ab]と記載する．
ここで，要素単語の構造記述は，以下の英記号とする．

・単純，連用，名詞性オノマトペ：O
・接頭オノマトペ：Op
・接尾オノマトペ：Os
・接頭語：P
・接尾語：S

・接頭造語要素：Q
・接尾造語要素：R
・名詞：N
・動詞：V（未然形V11,V12，連用形V22,V23，終止形V30，連体形V40，仮定形V50，命令形V60）
・形容詞：A（未然形A13，連用形A22，A23，終止形A30，連体形A40，仮定形A50，命令形A60）
・形容動詞（語幹）：K00
・副詞：D
・連体詞：T
・接続詞：C
・機能語及び機能性自立語：活用形も含め英小文字ローマ字綴り

文節内の語接続も便宜上，左2分岐句構造とみなして上記と同様の記述を行っている．
例えば，オノマトペ共起表現「クンクン-と-犬-が-鳴く」の構造記述は「クンクン」＝オノマトペO，「と」＝格助詞to，「犬」＝名詞N，「が」＝格助詞ga，「鳴く」＝動詞終止形V30であることから，[[Oto]*[[*Nga]*V30]] と記載する．図1にその意味する構文木と係り受け構造を示す[3]．

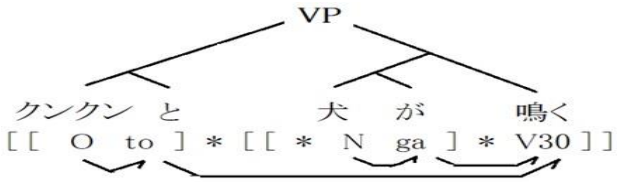


図 1 オノマトペ共起表現「クンクン-と-犬-が-鳴く」の構造記述

この例の如く F 欄の構造記述内には適所にアスタリスク「*」が含まれており，その位置に，直後の句の主辞に対する修飾句が入り得ることを意味している．従って，図 1 の構造記述 [[Oto]*[[*Nga]*V30]] は，例えば，「クンクン-と-朝-から-隣-の-犬-が-寂し-そう-に-鳴く」のような拡張表現の可能性を示している．図 2 にこの模様を示す[4]．

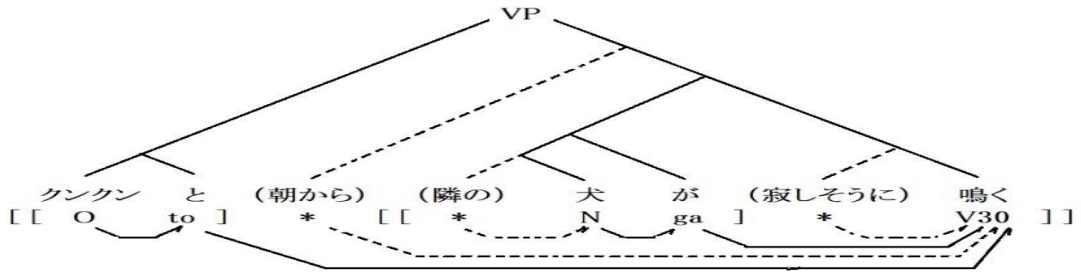


図 2 「クンクン-と-朝-から-隣-の-犬-が-寂し-そう-に-鳴く」の構造記述

JMWELのこのようなギャップ付き構造記述は，定形表現が持ち得る部分的な柔軟性を構文・

[3] この表現に対する 2018 年 6 月時における Google オンライン英訳は "Cunk and dogs cry"であった．
[4] この表現に対する 2018 年 6 月時における Google オンライン英訳は "Cunngun and morning next dog crying lonesome" であった．

意味解析機に反映させるための重要な仕組みである．

３．７　後方文脈条件：G 欄

オノマトペ（単体），オノマトペ共起表現に対し，文末側に呼応する語句がある場合にその情報を与える．例えば，「オチオチと」に対しては，文末側に「休んではいられない」のような否定句が要求されることを<negation>と記す．

３．８　連体化，連用化，動詞化情報：H 欄

オノマトペ（単体）に対し，E欄の構文機能情報を詳細化して与える．オノマトペを連体修飾，連用修飾に使用する場合と動詞化して使用する場合に通常使われる後接語句を以下のように整理した．

・連体修飾：「な」，「の」，「たる」
・連用修飾：「に」，「と」，「ε」
・動詞化：「する」，「になる」，「とする」

ここで，εは空列を表し，オノマトペが後接語句なしで連用修飾できる場合を表す．
例えば，オノマトペ「フラフラ」は，「フラフラの（…状態）」で連体修飾，「フラフラと（…歩く）」，「フラフラ（…歩く）」で連用修飾，「フラフラする」，「フラフラになる」，「フラフラとする」と動詞化すること，「フッ」の場合は，「フッと（…気が付く）」で連用修飾する以外には考えにくいことを，それぞれ，後接する語句集合の三つ組によって{no}-{to, ε}-{suru, ninaru, tosuru}，Φ-{to}-Φ と記載する．Φは空集合を表わす．三つ組のパタンは100種程度である．

３．９　擬音，擬態の別：I 欄

オノマトペ（単体）に対し，擬音，擬態の別を「音」，「態」と記載する．

４．応用例

機械翻訳等の自然言語処理においては，オノマトペをその共起表現やコロケーションとして一括して単位的に捉えることが必要である．例えば，日英機械翻訳において，前出の MWE「クンクンと犬が鳴く」全体を単位的に捉えなければ，的確な英語動詞句 “a dog whines”などを与えることは難しい．これを与えたのち，本レキシコンに記載された構造記述 [[Oto]*[[*Nga]*V30]]] を用いれば，拡張表現「クンクンと朝から隣の犬が寂しそうに鳴く」に対しても適切な訳文，例えば，“The neighbor’s dog sadly whines from the morning”などを生成出来る可能性が生じる．図３にこの翻訳過程のイメージを与える．

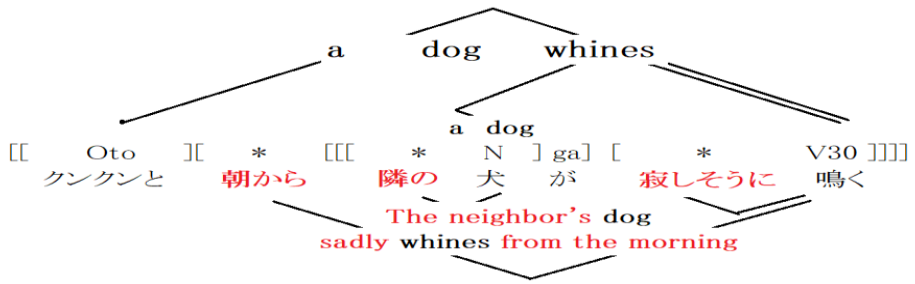


図３　「クンクンと朝から隣の犬が寂しそうに鳴く」に対する英訳
“The neighbor’s dog sadly whines from the morning”を生成する過程のイメージ

５．おわりに

「コロコロと音がする」，「コロコロ転がる」，「話がコロコロ変わる」，「目がコロコロと痛い」，「コロコロと笑う」など，他語との共起によって，「コロコロ」の多義性が低減されるように，語の共起を如何に捉えるかが NLP における基本的な課題である．筆者の一人は 1960 年代に始めた機械翻訳の研究を通して，意味素性や語類による語の共起ルールとは別に表層レベルの相当大規模な語の共起辞書が不可欠であるという認識を得，大規模日本語フレーズレキシコン JMWEL の開発を始めた．JMWEL の基本的な特徴の一つは必ずしも隣接しない語の共起をデータ化している点にある。現在，JMWEL は見出し数 14 万件（異なり）に達し，未だ十分網羅的とは言い切れないが，プロトタイプとしてのレベルには達したのではないかと考えている．

本稿で紹介したオノマトペ共起表現レキシコンは JMWEL のサブレキシコンの一つである．本レキシコン及び JMWEL が今後の日本語処理，機械翻訳，日本語教育・研究に役立てば幸いである[5].

文　献

阿刀田稔子・星野和子 (2004).「擬音語擬態語使い方辞典第2版」 創拓社出版.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (eds.) (1999). “Longman Grammar of Spoken and Written English”, *Harlow: Pearson Education Limited*.

Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen Wheatley (eds.) (2009). “Formulaic Language, vol.1, Distribution and historical change”, *John Benjamins Publishing Company*.

Charles J. Fillmore, Paul Kay and Mary Catherine O’Connor (1988). “Regularity and Idiomaticity Grammatical Construction: The Case of Let Alone” *Language* 64, pp.501-538.

福島弘識・内田ゆず・荒木健治 (2014).「2つの意味を持つオノマトペの意味判別における素性の検討」 言語処理学会第20回年次大会発表論文集, pp.181-184.

井上音々・望月源 (2017).「日本語歌謡曲のオノマトペに関する調査」言語処理学会第23回年次大会発表論文集, pp.963-966.

Nan Jiang and Tatiana M. Nekrasova (2007). “The Processing of Formulaic Sequences by Second Language Speakers”, *The Modern Language Journal*, 91:3, pp.433-445.

小宮嘉那子・佐々木稔・新納浩幸 (2016).「分散表現と文脈ベクトルによるオノマトペの分類の比較」言語処理学会第22回年次大会発表論文集, pp.473-476.

小野正弘 (2007).「日本語オノマトペ辞典」 小学館.

乙武北斗・内田ゆず・高丸圭一・木村奏知 (2016).「表層格に着目したオノマトペ共起語の抽出と分析」 言語処理学会第22回年次大会発表論文集.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger (2002). “Multiword Expressions: A Pain in the Neck for NLP” *Proc. of the 3rd CICLING*, pp.1-15.

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

玉岡賀津雄・木山幸子・宮岡弥生 (2011).「新聞と小説のコーパスにおけるオノマトペと動詞の共起パターン」言語研究139, pp.57-84.

[5] 本レキシコンの利用については関連サイト http://jefi.info を参照されたい.

Toshifumi Tanabe, Masahito Takahashi and Kosho Shudo (2014). “A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing”, *Computer Speech and Language*, 28:6, Elsevier, pp.1317-1339.

渡辺知恵美・中村聡 (2015).「オノマトペロリ：味覚や食感を表すオノマトペによる料理レシピのランキング」人工知能学会論文誌, 30:1, pp.340-352.

関連 URL

日本語処理研究工房ことばの森　　https://jefi.info

語彙多様性指標の可視化と単回帰分析による TTR の補正

今田 水穂 (文部科学省) *

Visualization of Lexical Diversity Indices and Adjustment of TTR by Single Regression Analysis

Mizuho Imada (MEXT)

要旨

語彙多様性を評価する既存の指標には、延べ語数 N と異なり語数 V(N) を入力とするもの、単語別の頻度を入力とするもの、単語列を入力とするものなどがある。本発表では、これらの指標の特徴を整理し、「現代日本語書き言葉均衡コーパス」(BCCWJ) を使用して指標値の分布を可視化する。N と V(N) を入力とする指標のいくつかは、両者の間に冪乗則 $V(N) = aN^b$ を仮定している。TTR は $b = 1$、R は $b = 0.5$ として a を指標値として利用するが、1 では大きすぎ、0.5 では小さすぎる。そこで V(N) と N の対数を単回帰分析して b の最適値を推定し、TTR を補正することを考える。実際には冪乗則は成立しないため、この補正は近似的だが、比較的簡単によい補正を得ることができる。この補正値を他の指標と比較し、テキストサイズが指標値の平均やばらつきに及ぼす影響を評価する。また、BCCWJ の 12 のサブコーパスについて b の値を推定し、一覧で示す。

1. はじめに

テキストの語彙多様性を評価する指標として、タイプ-トークン比 (TTR) が知られる。しかし TTR は延べ語数の影響を受け、テキストが長くなるほど指標値が小さくなる特徴がある。これを補正した指標の 1 つに R があるが、R は補正が強すぎて TTR とは逆に指標値が大きくなる特徴がある。他に多くの指標が提案されているが、計算の容易さ、テキスト長による平均やばらつきの変動など指標ごとに特徴がある。様々な観点から各指標の有用性を検討している最近の研究としては、木村・田中 (2010) や鄭・金 (2018) がある。

本稿では既存の主要な指標を計算に使用する入力データの形態から 3 種類に分類し、それらの特徴を検討する。また、延べ語数 N と異なり語数 $V(N)$ の間にべき乗則 $V(N) = aN^b$ が成り立つという仮定に基づき、両者の対数を単回帰分析して残差 ε を指標値として利用することで TTR を補正する方法を試みる。また「現代日本語書き言葉均衡コーパス」(BCCWJ) を用いて各指標の値を実際に計算し、その分布をグラフで確認するとともに、指標値と延べ語数を単回帰分析することでテキストの長さが各指標の平均や分散に及ぼす影響を評価し、単回帰分析による補正法が平均や分散の変動を受けにくいことを確認する。この補正値は延べ語数と異なり語数から簡単な式で計算することができるが、あらかじめ単回帰分析を行ってサンプル全体の

* imadamizuho.ac@google.com

分布の傾きを示すパラメータ値 b を計算しておく必要がある。そこでレジスタや語の集計単位を様々に変えて同補正法を試み、条件ごとのパラメータ値の一覧を示す。

2. 語彙多様性指標

語彙多様性を表す指標は、(1) 異なり語数と延べ語数を入力とするもの、(2) 単語別の頻度を入力とするもの、(3) 単語列を入力とするものがある。

表 1　データの種類と手法

種類	例	手法
(1) 総語数	{異なり語数: 11966, 延べ語数: 209326}	TTR, R, C, etc.
(2) 単語別頻度	{吾輩: 481, は: 6501, 猫: 237, . . .}	HD-D, Yule's K, etc.
(3) 単語列	{吾輩, は, 猫, である, . . .}	MSTTR, MATTR, etc.

(1) は異なり語数と延べ語数の関係を表す式を仮定して、その式の係数を指標として使うもので、TTR、R(Guiraud 1954)、C(Herdan 1960)、S(Somers 1966)、a^2(Maas 1972)、Uber(Dugast 1979) などがある。(2) は部分集合の特徴量を反復実測の代わりに各語の頻度に基づく確率計算で推定する HD-D(McCarthy and Jarvis 2007)、Yule's K、Simpson's λ、Shannon's H' や、頻度分布の形状を特徴量として利用するものなどがある。(3) は単語列の部分集合の特徴量 (n 語あたりの異なり語数など) を反復実測によって推定するもので、MSTTR、MATTR、voc-d(Mckee et al. 2000)、MTLD、MTLDMA などがある。入力データの情報量は (3) が最も大きく (1) が最も小さいが、その分、計算量も (3) が最も大きく (1) が最も小さい。

(1) のタイプの指標の代表的なものを以下に示す。N は延べ語数、V(N) は N 語あたりの異なり語数である。TTR は異なり語数を延べ語数で割ったものである。R と CTTR は平方根を用いた TTR の変種だが、CTTR は R の定数倍なので実質的に同一の指標である。それ以外のものは対数を用いた TTR の変種である。C、S、k は V(N) および N の対数、あるいは対数の対数を使用する。a^2 は $\log TTR = -a^2(\log N)^2$ と変形することができ、TTR と N の関係式と見なすことができる。Uber は a^2 の逆数なので、実質的に a^2 と同一の指標である。[(1)]

$$TTR = \frac{V(N)}{N} \qquad CTTR = \frac{V(N)}{\sqrt{2N}} \qquad k = \frac{\log V(N)}{\log(\log N)}$$

$$C = \frac{\log V(N)}{\log N} \qquad a^2 = \frac{\log N - \log V(N)}{(\log N)^2}$$

$$R = \frac{V(N)}{\sqrt{N}} \qquad S = \frac{\log(\log V(N))}{\log(\log(N))} \qquad Uber = \frac{(\log N)^2}{\log N - \log V(N)}$$

(1) 鄭・金 (2018) は他に $LN = \frac{1-V(N)^2}{V(N)^2 \log N}$ を挙げている。LN は $LN = -\frac{V(N)^2-1}{V(N)^2}\frac{1}{\log N}$ と変形でき、V(N) がごく小さい値のとき以外は $-\frac{1}{\log N}$ と近似する値になる。本稿では LN は扱わない。

次に、(2) のタイプの指標の代表的なものを以下に示す。n_i は語 w_i の頻度、p_i は語 w_i の生起確率 (n_i/N) である。V(n, N) は長さ N のテキストにおける頻度 n の語の異なり語数である。HDD は延べ語数 N のテキストから無作為に M 語を非復元抽出したときの異なり語数の期待値である。λ と ℓ は Simpson 指数と呼ばれるもので、テキスト中から無作為に 2 語を抽出したとき同じ語である確率に相当し、λ が復元抽出、ℓ が非復元抽出である。K は $K = 10^4 \frac{\sum_{i=1}^{all} n_i(n_i-1)}{N^2}$ と同値であり、N が十分大きければ ℓ の 10^4 倍と近似した結果を返す。H' は Shannon 指数、エントロピーなどと呼ばれるもので、$H' = -\ln\left[\prod_{i=1}^{all} p_i^{n_i}\right]^{\frac{1}{N}}$ と変形することができ、直観的には延べ語数 N のテキストから無作為に N 語を復元抽出したときに元のテキストと同一の単語列が得られる確率、あるいは各語が元のテキストの同じ位置の語と同一である確率の幾何平均と関係がある。m と S は異なり語数 V(N) と頻度 2 の語 (dis legomena) の異なり語数 V(2, N) の比である。Z は異なり語数 V(N)、延べ語数 N、最頻語の頻度 p を関係付ける係数である。m、S、Z は頻度スペクトルの分布形状を語彙多様性指標として利用するものと考えることができる。

$$HDD = \sum_{i=1}^{all}\left(1 - \frac{\binom{N-n_i}{M}}{\binom{N}{M}}\right)$$

$$H' = -\sum_{i=1}^{all} p_i \ln p_i$$

$$\lambda = \sum_{i=1}^{all} p_i^2$$

$$m = \frac{V(N)}{V(2,N)}$$

$$\ell = \frac{\sum_{i=1}^{all} n_i(n_i-1)}{N(N-1)}$$

$$S = \frac{V(2,N)}{V(N)}$$

$$K = 10^4 \frac{\left[\sum_{n=1}^{all} V(n,N)\left(\frac{n}{N}\right)^2\right] - N}{N^2}$$

$$V(N) = \frac{Z}{\log(pZ)}\frac{N}{N-Z}\log\left(\frac{Z}{N}\right)$$

最後に (3) のタイプの指標について説明する。このタイプの指標は、テキストから一定の条件を満たす単語列を抽出する処理を繰り返し、その特徴量の平均などを指標値として使用する。テキストから一定の長さの単語列を抽出したときの異なり語数を特徴量とするもの (MSTTR、MATTR、vocd) と、TTR が一定の値に達するときの単語列の長さを特徴量とするもの (MTLD、MTLD-MA) がある。ただし前者は異なり語数をそのまま特徴量として使うのではなく、MSTTR と MATTR は TTR、vocd は D という指標に換算する。D は長さ n と異なり語数 V(n) の関係を次の式でモデル化したときの係数であり、多数回の測定で得られたデータからフィッティングによって推定される(2)。

$$TTR = \frac{D}{N}\left[\left(1 + 2\frac{N}{D}\right)^{\frac{1}{2}} - 1\right]$$

テキストから一定の条件を満たす単語列を抽出する方法としては、テキストを一定の条件を満たすセグメントに分割して平均を求める方法、一定の条件を満たすウィンドウをテキス

(2) この式は D について $D = \frac{V^2}{2(N-V)}$ と解くことができる。

トの先頭から末尾まで移動させて平均を求める方法、無作為に単語を抽出する方法がある。MSTTR と MTLD はセグメント平均法、MATTR と MTLD-MA は移動平均法、vocd は無作為抽出法である。前述の HDD は、無作為抽出法と同様の計算を実測ではなく理論値として確率的に計算するものである。

3. 単回帰分析による TTR の補正

前節で示した指標のうち、(1) のタイプの指標は V(N) と N から計算される。V(N) と N の関係を理解するために、BCCWJ コアデータを用いて両者の関係を確認する。図 1 の左は横軸を N、縦軸を V(N) とした散布図である。$N > 12000$ の 1 サンプルを外れ値として除外した。N が大きくなるに従って V(N) も大きくなるが、データポイントの分布は直線的ではなく、V(N) と N が比例しているわけではないことが分かる。図 1 の右は同じデータを両対数グラフにプロットしたものである。両対数グラフでは、データポイントが直線的に分布する。

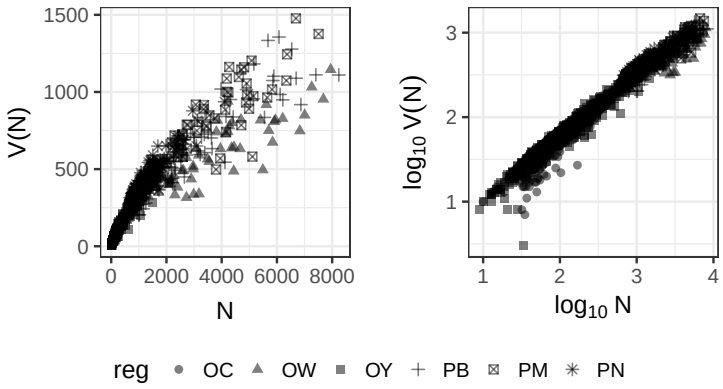


図 1　異なり語数と延べ語数

両対数グラフが直線的に分布するということは、V(N) と N の間にべき乗則 $V(N) = kN^{\beta}$ が成立することを意味する。この関係はヒープスの法則と呼ばれる。ここでは、V(N) と N の関係は次の式で近似できると仮定する。

$$V(N) = 10^{a} \times N^{b}$$

$$\log_{10} V(N) = a + b \times \log_{10} N$$

前節で述べた語彙多様性指標のうち、TTR、R、C は、V と N の間にべき乗則を仮定するモデルと考えることができる。

$$V(N) = TTR \times N \qquad V(N) = R \times N^{0.5} \qquad V(N) = N^{C}$$

これは両対数グラフ上において、TTR、R、C の値が等しいサンプルは、それぞれ直線上に並ぶことを意味する。実際に、TTR、R、C それぞれの上位 5%、下位 5% にあたる値を示す直線を両対数グラフに重ねたものを図 2 に示す。破線は回帰直線である。

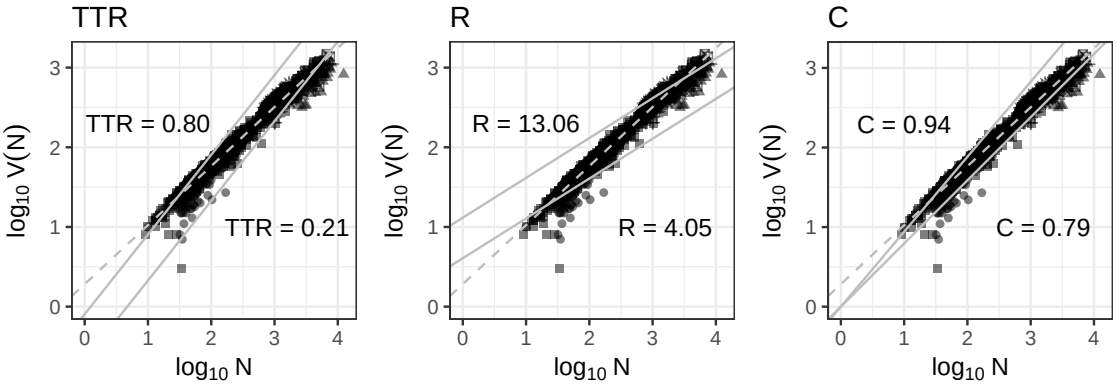


図 2　語の頻度と語彙多様性指標

TTR が等しいデータポイントは傾き 1、R が等しいデータポイントは傾き 0.5 の直線上に並ぶ。しかし実際のデータと比べて TTR は傾きが大きすぎ、R は傾きが小さすぎるため、N が大きくなるほど TTR の実測値は小さくなり、R の実測値は大きくなる。C が等しいデータポイントは、原点を通る傾き C の直線上に並ぶ。しかし実際のデータの分布は原点を通らないため、N が大きくなるほど C の実測値は小さくなる。

そこで実際のデータを単回帰分析して、各データポイントの残差 ε を指標値として利用することを考える。ただし ε をそのまま指標値として使うのではなく、TTR と形式を合わせるために次の式で計算する。

$$ETTR = \frac{V(N)}{N^b}$$

この式は回帰直線の傾き b をパラメータとして、切片 a と残差 ε の和を指標値として利用するもので、$ETTR = 10^{a+\varepsilon}$ である。R が平方根、C が対数を使用するのに対して、この指標は N のべき乗 (exponentiation) を使用するので、指標名は ETTR としておく。b の値はデータを単回帰分析することで推定することができる。BCCWJ コアデータの場合は、$b \approx 0.74$ である。

表 2　単回帰分析

	(Intercept)	log10(N)	R^2	Adj. R^2	Num. obs.	RMSE
log10(V(N))	0.28***	0.74***	0.97	0.97	1980	0.08
	(0.01)	(0.00)				

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

この指標を両対数グラフに重ねた図を以下に示す。TTR、R、C と比べて、データの分布によく当てはまっていることが確認できる。

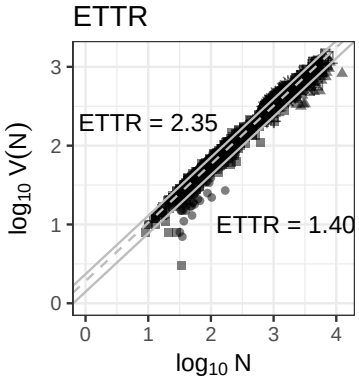


図 3　語の頻度と語彙多様性指標

4. 指標の評価

前節までに挙げた指標がテキストの長さによってどのような影響を受けるか、BCCWJ コアデータを用いて確認する[(3)]。最初に散布図を確認する (図 4)。図の横軸は $\log_{10} N$、縦軸は各指標の値である。CTTR は R の定数倍なので省略した。データポイント数は 1980 だが、一部のデータを外れ値として除外した ($C < 0.7$、$S > 1$、$S < 0.5$、$k < 1$、$a^2 > 0.2$、$U > 100$、$m > 30$、$s < 0.4$)。λ、ℓ、K、Z は値のばらつきが大きいため対数で示した。MSTTR、MATTR、HDD は $N = 42$、MTLD、MTLDMA は $TTR = 0.77$ で計算し、$N \leq 42$ のサンプルを除外した。また MSTTR、MATTR は HDD との比較のため TTR ではなく V(42) に換算して示した。データポイントの形は知恵袋 (OC)、白書 (OW)、ブログ (OY)、書籍 (PB)、雑誌 (PM)、新聞 (PN) の 6 つのレジスタを表す。図中の直線は、回帰直線である。

類似の手法である λ、ℓ、K のうち、ℓ と K はほぼ同様の分布を示しているのに対し、λ は N が小さいときに ℓ や K とは異なる分布を示す。MSTTR、MATTR、HDD は、それぞれ計算の方法が違うものの、似た分布を示す。MTLD と MTLDMA はそれほど似ておらず、MTLD の方がばらつきが大きい。対数関係にある a^2 と Uber、および m と s は、それぞれ対数化した場合に上下対称の分布となる。

次に、N が指標値の平均や分散に及ぼす影響を確認する。平均については、各指標を $\log_{10} N$ で単回帰分析し、その決定係数 R^2 で評価する。分散については、単回帰分析によって得られた残差の絶対値を単回帰分析し、その R^2 で評価する。いずれも R^2 が小さいほど N の影響が小さいと考えられる。結果を図 5 に示す。横軸は指標値の R^2、縦軸は残差の R^2 である。データポイントの形は表 1 で示した各指標の入力種別である。

(3) TTR、R、C、S、a^2、Uber、K、MSTTR、MATTR、HDD、MTLD、MTLDMA は、R の koRpus パッケージを使用して計算した。それ以外の指標は独自に計算した。

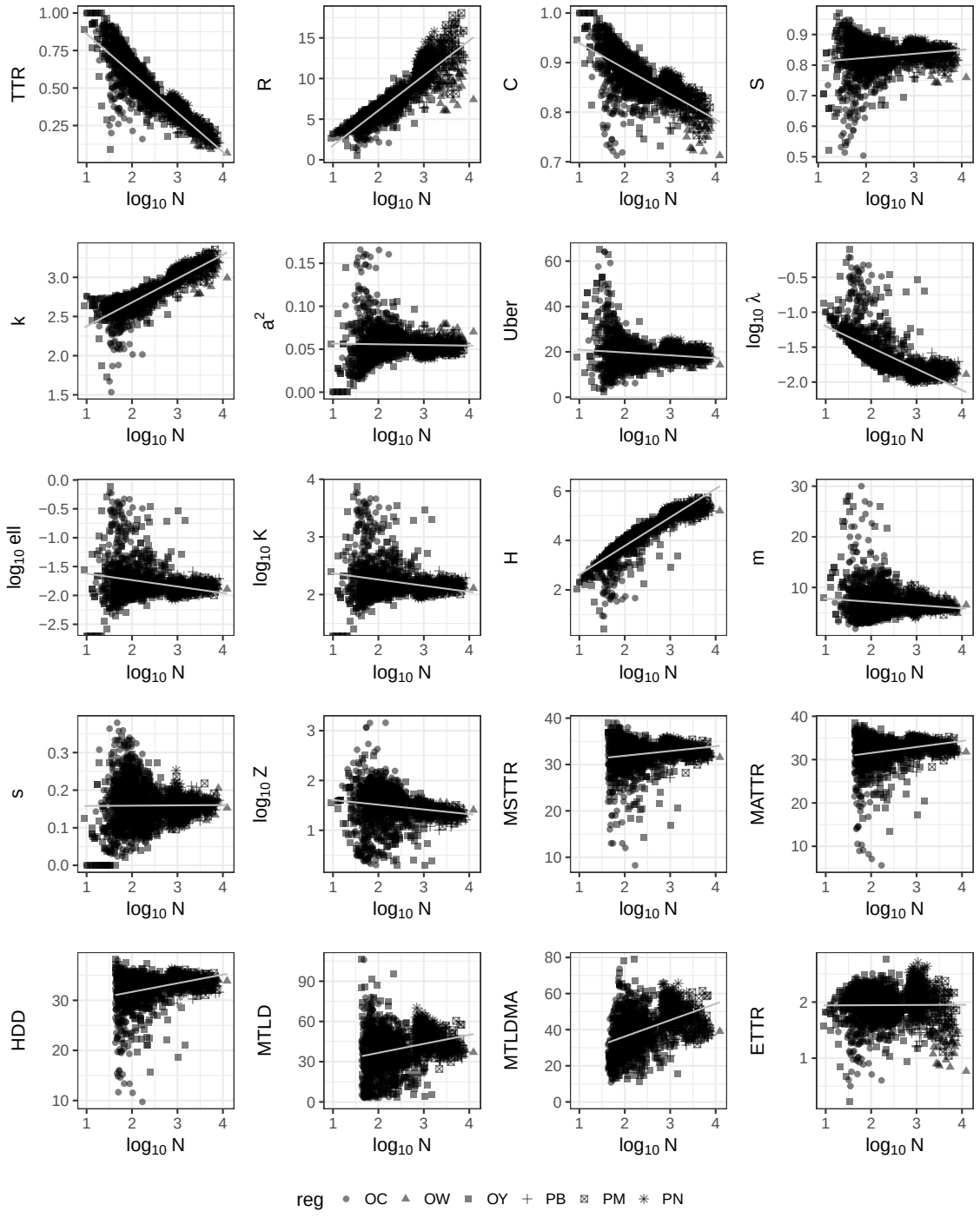


図 4　語の頻度と語彙多様性指標

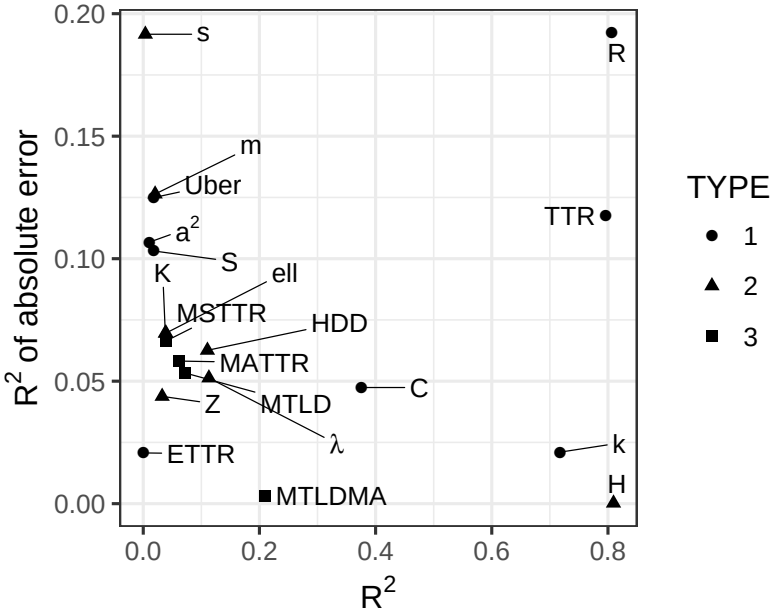


図 5　N が指標値の平均や分散に及ぼす影響

ETTR はどちらの R^2 も小さい値を取り、平均と分散がいずれも N の影響を受けにくいことが分かる。TTR と R はいずれの R^2 も大きな値を取り、N の影響を受けやすい指標だと言える。また R の方がより大きな値を取ることから、TTR より R の方が N の影響が小さいとは言えない。TTR と R 以外では、H や k は平均値が N の影響を強く受けるため、長さの異なるデータの比較には適さないと考えられる。それ以外の指標の多くは、平均が N から受ける影響を小さくすることには成功しているが、分散が N から受ける影響は比較的大きい。図 4 を見ると多くの指標において N が小さいときに指標値のばらつきが大きくなっている。データが小さいときに結果が安定しないことは自然なことではあるが、そのような指標値について N が小さいデータと大きいデータを比較する際には、N が小さいデータの方が極端な値を取りがちであることを考慮する必要がある。

5. 種々の条件における補正値

ETTR はテキストの延べ語数と異なり語数だけで計算することができるが、パラメータ b の値を決定するために単回帰分析を行う必要がある。しかし、あらかじめ b が分かっていれば、その都度単回帰分析を行う必要はない。前々節では、BCCWJ コアデータ (短単位語彙素) において b が 0.74 程度の値になることを確認した。しかし、この値は語の単位やテキストのジャンルといった条件によって変化することが考えられる。そこで本節では、種々の条件下で b がどの程度の値を取るか確認する。コアデータより広い範囲のレジスタを調べるため、BCCWJ 非コアデータの出版サブコーパス (3 レジスタ) と特定目的サブコーパス (9 レジスタ) を調査範囲とした。パラメータとして、次の 3 つのカテゴリーを使用する。

表 3　パラメータ

カテゴリー	値
レジスタ	書籍 (PB) | 雑誌 (PM) | 新聞 (PN) | 白書 (OW) | 教科書 (OT) | 広報誌 (OP) | ベストセラー (OB) | Yahoo!知恵袋 (OC) | Yahoo!ブログ (OY) | 韻文 (OV) | 法律 (OL) | 国会会議録 (OM)
トークン単位	短単位 (suw) | 長単位 (luw)
タイプ単位	語彙素 (lemma) | 書字形基本形 (orthbase) | 書字形出現形 (orth)

これらの組み合わせごとに単回帰分析した結果を表 4 に示す。

表 4　単回帰分析

reg	token	type	a		b		R^2	adj. R^2	Num. obs.	RMSE
OB	luw	lemma	0.3121***	(0.01)	0.7355***	(0.00)	0.96	0.96	1390	0.06
		orthbase	0.2998***	(0.01)	0.7421***	(0.00)	0.96	0.96	1390	0.06
		orth	0.2980***	(0.01)	0.7534***	(0.00)	0.97	0.97	1390	0.05
	suw	lemma	0.3615***	(0.01)	0.7067***	(0.00)	0.95	0.95	1390	0.06
		orthbase	0.3482***	(0.01)	0.7149***	(0.00)	0.95	0.95	1390	0.06
		orth	0.3568***	(0.01)	0.7234***	(0.00)	0.96	0.96	1390	0.06
OC	luw	lemma	0.2039***	(0.00)	0.7845***	(0.00)	0.92	0.92	91445	0.07
		orthbase	0.1940***	(0.00)	0.7920***	(0.00)	0.92	0.92	91445	0.07
		orth	0.1540***	(0.00)	0.8231***	(0.00)	0.92	0.92	91445	0.07
	suw	lemma	0.2737***	(0.00)	0.7432***	(0.00)	0.91	0.91	91445	0.06
		orthbase	0.2599***	(0.00)	0.7538***	(0.00)	0.92	0.92	91445	0.06
		orth	0.2161***	(0.00)	0.7879***	(0.00)	0.92	0.92	91445	0.06
OL	luw	lemma	0.5436***	(0.03)	0.6041***	(0.01)	0.93	0.93	346	0.07
		orthbase	0.5421***	(0.03)	0.6064***	(0.01)	0.93	0.93	346	0.07
		orth	0.5183***	(0.03)	0.6205***	(0.01)	0.94	0.94	346	0.07
	suw	lemma	0.7770***	(0.03)	0.4868***	(0.01)	0.87	0.87	346	0.08
		orthbase	0.7754***	(0.03)	0.4905***	(0.01)	0.87	0.87	346	0.08
		orth	0.7648***	(0.03)	0.5020***	(0.01)	0.88	0.88	346	0.08
OM	luw	lemma	0.3801***	(0.03)	0.7009***	(0.01)	0.98	0.98	159	0.06
		orthbase	0.3704***	(0.03)	0.7045***	(0.01)	0.98	0.98	159	0.06
		orth	0.3401***	(0.03)	0.7203***	(0.01)	0.99	0.99	159	0.06
	suw	lemma	0.6239***	(0.03)	0.5997***	(0.01)	0.98	0.98	159	0.06
		orthbase	0.6099***	(0.03)	0.6057***	(0.01)	0.98	0.98	159	0.06
		orth	0.5828***	(0.03)	0.6213***	(0.01)	0.98	0.98	159	0.06
OP	luw	lemma	0.2099***	(0.02)	0.7973***	(0.01)	0.98	0.98	354	0.03
		orthbase	0.2069***	(0.03)	0.7994***	(0.01)	0.98	0.98	354	0.03
		orth	0.1914***	(0.03)	0.8079***	(0.01)	0.98	0.98	354	0.03
	suw	lemma	0.5775***	(0.04)	0.6548***	(0.01)	0.93	0.93	354	0.04
		orthbase	0.5811***	(0.04)	0.6574***	(0.01)	0.93	0.93	354	0.04
		orth	0.5679***	(0.04)	0.6655***	(0.01)	0.92	0.92	354	0.04
OT	luw	lemma	0.2360***	(0.04)	0.7456***	(0.01)	0.91	0.91	412	0.10
		orthbase	0.2308***	(0.04)	0.7489***	(0.01)	0.91	0.91	412	0.10
		orth	0.2553***	(0.03)	0.7526***	(0.01)	0.92	0.92	412	0.09
	suw	lemma	0.4023***	(0.04)	0.6723***	(0.01)	0.87	0.87	412	0.11

		orthbase	0.3965***	(0.04)	0.6778***	(0.01)	0.88	0.88	412	0.11
		orth	0.4291***	(0.04)	0.6796***	(0.01)	0.89	0.89	412	0.11
OV	luw	lemma	1.3786***	(0.10)	0.4104***	(0.04)	0.35	0.35	252	0.07
		orthbase	1.3459***	(0.10)	0.4255***	(0.04)	0.37	0.37	252	0.07
		orth	1.1158***	(0.08)	0.5139***	(0.03)	0.59	0.58	252	0.05
	suw	lemma	1.6573***	(0.12)	0.3185***	(0.04)	0.20	0.20	252	0.07
		orthbase	1.6221***	(0.12)	0.3356***	(0.04)	0.22	0.21	252	0.07
		orth	1.3047***	(0.10)	0.4529***	(0.03)	0.44	0.43	252	0.06
OW	luw	lemma	0.3842***	(0.02)	0.7167***	(0.01)	0.91	0.91	1500	0.06
		orthbase	0.3800***	(0.02)	0.7188***	(0.01)	0.91	0.91	1500	0.06
		orth	0.3727***	(0.02)	0.7262***	(0.01)	0.92	0.91	1500	0.06
	suw	lemma	0.8427***	(0.03)	0.5361***	(0.01)	0.71	0.71	1500	0.09
		orthbase	0.8525***	(0.03)	0.5369***	(0.01)	0.72	0.72	1500	0.09
		orth	0.8533***	(0.03)	0.5424***	(0.01)	0.73	0.73	1500	0.09
OY	luw	lemma	0.2230***	(0.00)	0.7709***	(0.00)	0.95	0.95	52680	0.08
		orthbase	0.2159***	(0.00)	0.7765***	(0.00)	0.96	0.96	52680	0.08
		orth	0.2014***	(0.00)	0.7922***	(0.00)	0.96	0.96	52680	0.08
	suw	lemma	0.2822***	(0.00)	0.7420***	(0.00)	0.95	0.95	52680	0.08
		orthbase	0.2707***	(0.00)	0.7512***	(0.00)	0.95	0.95	52680	0.08
		orth	0.2558***	(0.00)	0.7674***	(0.00)	0.96	0.96	52680	0.08
PB	luw	lemma	0.3533***	(0.01)	0.7236***	(0.00)	0.89	0.89	10117	0.07
		orthbase	0.3405***	(0.01)	0.7296***	(0.00)	0.89	0.89	10117	0.07
		orth	0.3405***	(0.01)	0.7400***	(0.00)	0.91	0.91	10117	0.07
	suw	lemma	0.5011***	(0.01)	0.6586***	(0.00)	0.81	0.81	10117	0.09
		orthbase	0.4889***	(0.01)	0.6662***	(0.00)	0.81	0.81	10117	0.09
		orth	0.4988***	(0.01)	0.6737***	(0.00)	0.83	0.83	10117	0.09
PM	luw	lemma	0.4526***	(0.02)	0.7092***	(0.00)	0.92	0.92	1996	0.06
		orthbase	0.4389***	(0.02)	0.7152***	(0.00)	0.92	0.92	1996	0.06
		orth	0.4237***	(0.01)	0.7290***	(0.00)	0.94	0.94	1996	0.05
	suw	lemma	0.6380***	(0.02)	0.6362***	(0.01)	0.88	0.88	1996	0.07
		orthbase	0.6272***	(0.02)	0.6437***	(0.01)	0.88	0.88	1996	0.07
		orth	0.6157***	(0.02)	0.6562***	(0.01)	0.90	0.90	1996	0.06
PN	luw	lemma	0.2139***	(0.01)	0.7944***	(0.00)	0.95	0.95	1473	0.03
		orthbase	0.2009***	(0.01)	0.7999***	(0.00)	0.95	0.95	1473	0.03
		orth	0.1816***	(0.01)	0.8126***	(0.00)	0.96	0.96	1473	0.03
	suw	lemma	0.4117***	(0.02)	0.7149***	(0.01)	0.87	0.87	1473	0.04
		orthbase	0.4003***	(0.02)	0.7216***	(0.01)	0.88	0.88	1473	0.04
		orth	0.3937***	(0.02)	0.7298***	(0.01)	0.88	0.88	1473	0.04

表 4 の b の値を ETTR のパラメータとして使用することで、延べ語数の影響を補正した指標値を得ることができる。同時に、これらの値は各レジスタの語彙分布の特徴を示している。他の条件が同じであれば、a が δ 大きくなると V(N) は 10^{δ} 倍になり、b が δ 大きくなると V(N) は N^{δ} 倍になる。N が 10 以上のテキストであれば、b の方が V(N) の値により強く影響する。従って、基本的に b の値が大きいレジスタほど、語彙が多様だと考えられる。luw と suw の値の差が大きいレジスタは、複合語が多いことが予想される。lemma、orthbase、orth の値の差が大きいレジスタは、表記が多様であったり、活用語が豊富であることが予想される。各レジ

スタの語彙の構成は別に調査し検証する必要があるが、本稿では予測の提示に留める。

6. おわりに

TTR などの語彙多様性指標の特徴について検討し、単回帰分析の残差を利用することでテキスト長が指標値に及ぼす影響を補正する方法を示した。この方法はサンプル全体の分布を表すパラメータを単回帰分析によって計算する必要があるが、パラメータが分かっている場合には V(N) と N のみを入力として容易に計算することができ、指標値の平均と分散の変動について他の指標と比べて遜色ない補正を得ることができる。一方で、この手法には問題も存在する。最後に、この手法に関する既知の問題について言及する。

第 1 に、この手法は V(N) と N がべき乗則に従うことを前提とする。厳密には両者の間はべき乗則に従っておらず、V(N) と N の両対数グラフは直線ではなくややカーブする。そのため、N の大きさによってパラメータ b の値は変動する。図 6 は各テキストの先頭から一定の割合の部分単語列を取って単回帰分析したときの b の変動である。OC、OW、OY、PM では N が増加するほど b が減少し、PB と PN は単調に減少はしないが変動している。従って、便宜的には表 4 の b を使用することで TTR や R よりはよい補正を得ることができるが、なるべくよい補正を得るためには分析対象となるサンプルについて単回帰分析を行い、b の値を計算して使用することが望ましい。

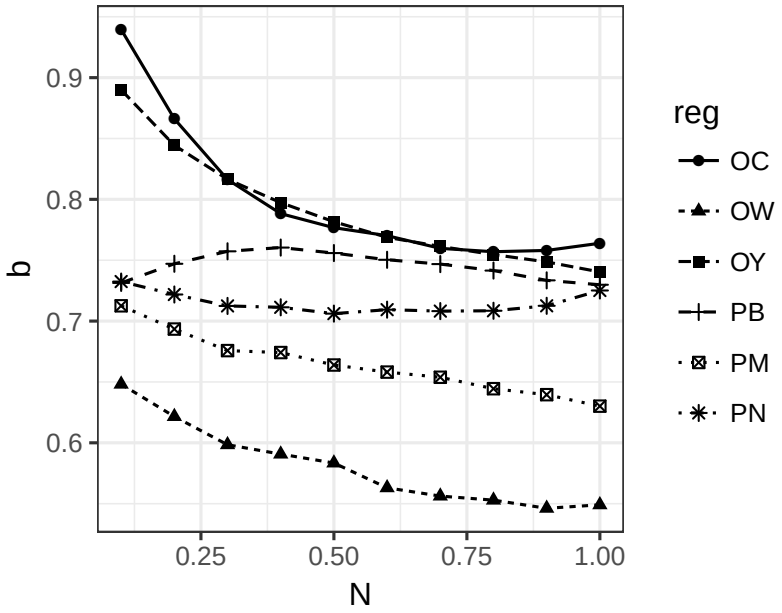


図 6　テキスト長によるパラメータの変動

第 2 に、V(N) と N の値によって計算する全ての手法に共通して言えることだが、語彙の豊かさは必ずしも V(N) と N の値のみから判断できるものではない。例えば、直観的には明らかに語彙の豊かさに差があるように見える小学生と中学生の作文が、語の頻度からみると V(N) と N のいずれも全く等しいということがあり得る。年齢の低い児童が話し言葉の語彙を書き

言葉でも使用するのに対して、年齢が上がるにつれて書き言葉の語彙を習得し使用するようになることを考えると、テキストに現れる語彙数が潜在的な語彙量の総体を正しく反映しているとは必ずしも言えない。語の豊かさについては、語の難易度、レジスタごとの語のふさわしさ、品詞構成比などで評価される語彙密度など、頻度以外の観点からも総合的に判断する必要がある。

文　献

Dugast, Daniel (1979) *Vocabulaire et stylistique*: Slatkine.

Guiraud, P (1954) *Les Charactères Statistiques Du Vocabulaire*: Presses Universitaires de France.

Herdan, Gustav (1960) *Type-Token Mathematics*: Mouton.

Maas, H. D. (1972) "Zusammenhang Zwischen Wortschatzumfang Und Länge Eines Textes," *Zeitschrift für Literaturwissenschaft und Linguistik*, Vol. 8, pp. 73-79.

McCarthy, Philip M. and Scott Jarvis (2007) "Vocd: A Theoretical and Empirical Evaluation," *Language Testing*, Vol. 24, No. 4, pp. 459-488.

Mckee, Gerard, D.D. Malvern, and Brian Richards (2000) "Measuring Vocabulary Diversity Using Dedicated Software," *Literary and Linguistic Computing*, Vol. 15, pp. 323-337.

Somers, H. H. (1966) "Statistical Methods in Literary Analysis," in Leeds, J. ed. *The Computer and Literary Style*: Kent State University Press, pp. 128-140.

木村大翼・田中久美子 (2010) 「文書長に依存しない文書定数」,『言語処理学会第 16 回年次大会発表論文集』, 1090-1093 頁.

鄭弯弯・金明哲 (2018) 「変動係数を用いた語彙の豊富さ指標の比較評価」,『同志社大学ハリス理化学研究報告』, 第 58 巻, 第 4 号, 230-241 頁.

二字漢語を構成する漢字の造語力の変化
—『現代雑誌九十種の用語用字』データと『現代日本語書き言葉均衡コーパス』の比較を通して—

本多 由美子（一橋大学大学院言語社会研究科）†

Changes of Word-Building Ability regarding Component Characters of Two-Character Sino-Japanese Words: A Comparative Analysis on 'Ninety Magazines of Today' and 'Balanced Corpus of Contemporary Written Japanese'

Yumiko Honda (Hitotsubashi University Graduate School of Language and Society)

要旨

宮島(1969)によると，明治時代に比べ現代では二字漢語を構成する漢字の意味はとりにくくなっており，漢語の造語力が弱まった原因の 1 つであるという。このことを宮島(1969)は『郵便報知』と『現代雑誌九十種』のデータを用いて示した。本稿では『現代雑誌九十種』以降の傾向を捉えるために，宮島(1969)と同様の方法で『現代雑誌九十種』のデータと『現代日本語書き言葉均衡コーパス』の「新聞」「知恵袋」とを比較した。比較の観点は二字漢語の構成漢字の「一字漢語の独立用例の有無」と「訓読みによる用例の有無」である。その結果(1)『現代雑誌九十種』以降，「一字漢語」と「訓読み」の用例を有する漢字は減少傾向にあり，(2)『現代雑誌九十種』以降の「訓読み」の減少幅が大きいことが確認された。このことは，部分的，間接的にではあるが，二字漢語を構成する漢字の造語力が低下傾向にあることを示すものと考えられる。

１．はじめに

漢字二字から成る漢語(以下「二字漢語」とする)は明治期に多くの語が作られたが，その生産能力は「明治から現在に近づくにつれて衰えをみせ(『国語学大辞典』「造語力」の項)」ているという指摘がある。現代では二字漢語の造語力が低下していることについて，二字漢語を構成する漢字(以下「構成漢字」と呼ぶ)の意味がとりにくくなっていることが原因の 1 つであると指摘したものに宮島(1969)がある。宮島(1969)は，明治期のデータとして国立国語研究所(1959)の「郵便報知」（以下「郵便報知」とする)，現代のデータとして国立国語研究所(1962，1963，1964)『現代雑誌九十種の用語用字』(以下「雑誌 90 種」とする)のデータを用いて語種別の語数や音読みされた漢字の数など，いくつかの観点から量的に比較を行い，明治期から現代における漢語の位置づけを論じている。その比較の中で，宮島(1969)は構成漢字の用例を量的に捉えることによって，現代では漢字の意味がとりにくくなり，そのことによって新しい二字漢語が作られにくくなったと述べている。

本稿では宮島(1969)の，個々の漢字の意味が漢語の造語に影響を与えるという考えに注目する。そして，二字漢語における漢字の造語力を「ある漢字が他の漢字と組み合わさって新しく二字漢語を形成する能力[1]」と捉え，宮島(1969)と同様の方法で構成漢字の用例を調査することによって，「雑誌 90 種」以降における漢字の造語力の一端を捉えることを試みる。なお，本稿の考察対象は二字漢語のみであるため，本稿の「漢字の造語力」は「二

† nihonda(at)hotmail.com

1 漢字の造語力について，文化審議会(2010:11)に「熟語の構成能力」との記述があり本稿はこれに従う。

字漢語の造語」の範囲に限定する。

２．宮島(1969)による「郵便報知」と「雑誌 90 種」の比較

宮島(1969)は「郵便報知」と「雑誌 90 種」の比較において，構成漢字の「一字漢語としての独立性」と「訓とのつながり」の 2 点について検討している。宮島(1969)によると，明治期よりも現代の方がこの 2 点のいずれについても弱まっているという。

宮島(1969)は，一字漢語とは「漢字一字が表す漢語の要素」で「一字漢語の独立性」を「一字漢語がそれだけで単語として使われる傾向」とし，「門」「線」のように漢字一字で語として用いられるものや，「信ずる」「愛する」のように「する」が付いてサ変動詞を形成する語，「単に」「特に」など副詞的な語を挙げている。また「訓とのつながり」については，和語に当てられる漢字，つまり訓読みに用いられる漢字は意味がとりやすいが，構成漢字の中で訓読みに用いられないものは，構成する要素として意味がとりにくいと述べている。例えば「妊娠」の「妊」や「廉価」の「廉」は，「郵便報知」では「硫酸鉄ハ妊ヲ防ク良品ナリ」や「現金引換へ価廉に品良なるを以て名あり」という用例があり[2]，漢字一字が一字漢語として独立して用いられているのに対し，現代では漢字一字が一語として独立して用いられることがなくなっていることや，「郵便報知」では「委細」の「委」，「検索」の「索」が「委しい(くわしい[3])」「索める(もとめる)」のような和語にあてられ訓読みで用いられているのに対し，現代ではこれらの訓読みは使われなくなっていることなどが挙げられている。

構成漢字が一字漢語として独立して用いられる例(以下「一字漢語の独立用例」とする)の減少は，例えば「脳(ノウ)」「死(シ)」「点(テン)」「線(セン)」のような独立した一字漢語に用いられる漢字が減少していることを意味する。このことは「脳死(ノウシ)」や「点線(テンセン)」のように，これらの一字漢語が直接結びつく方法による造語が行われにくくなること，そしてそれは部分的にではあるが構成漢字の造語力が低下傾向にあることを意味する。「訓読みによる用例」の減少は，訓読みされる漢字が減り，漢字が和語と結びつかなくなることであり，例えば「国外（コクガイ)」という語の意味を理解する際，「国(くに)」「外(そと)」という構成漢字の訓読みと結びつけるように，二字漢語の意味を構成漢字ごとに分解して捉えることが難しくなることを意味する。また，意味がとりにくくなった漢字を使って新しい語は作られにくくなる。宮島(1969)の上述の例「検索」の場合，「索」が和語の表記に用いられなくなることによって「索」の意味がとりにくくなり「検索」の意味を「検」と「索」に分解して捉えることが難しくなるとともに「索」を使った造語もされにくくなるということである。漢字の訓によって二字漢語が直接造語されるわけではないが，漢字の意味がとりにくくなることは，漢字の造語力に間接的に影響を与えると思われる。

宮島(1969)は明治期に新しい二字漢語が作られた際には，漢字一字の意味が明確だったため理解が容易だった語が，漢字一字の意味が明確ではなくなることによって構成漢字に分解して理解することが難しくなったと述べている。このことを量的に検討するために，宮島(1969)は「雑誌 90 種」における高頻度の二字漢語 100 語について，構成漢字が一字漢語として独立して使われることがあるか否か，訓読みに使われることがあるか否かを「郵

[2] 「郵便報知」の用例は宮島(1969)による。

[3] 本稿では，漢字や二字漢語の読み方を示す場合，訓読みを平仮名，音読みを片仮名で示す。

便報知」「雑誌 90 種」から用例を抽出することによって調べている。その上で，各二字漢語がどのような漢字の組み合わせによって形成されているかをまとめ，語を分類している。まず，構成漢字については，一字漢語の独立用例の有無によって「独立・非独立」の 2 種類に分け，また，訓読みに用いられる例(以下「訓読みによる用例」とする)の有無によって「訓あり・訓なし」の 2 種類に分けている。次に，各二字漢語を一字漢語の観点から構成漢字が「両方(2 字とも)独立」「一方が独立」「両方非独立」の 3 種類，訓読みについては「両方訓がある」「一方に訓がある」「両方訓がない」の 3 種類に分けている。宮島(1969:42-43)の結果を表 1，2 に示す。この表は筆者が語数を抜粋しまとめたものである。

表 1「郵便報知」「雑誌 90 種」の一字漢語比較

		郵便報知			計
		両方独立	一方が独立	両方非独立	
雑誌 90 種	両方独立	15	2	1	18
	一方が独立	16	20	4	40
	両方非独立	10	17	15	42
計		41	39	20	100

表 2「郵便報知」「雑誌 90 種」の訓読み比較

		郵便報知			計
		両方訓あり	一方に訓あり	両方訓なし	
雑誌 90 種	両方訓あり	57	5	0	62
	一方に訓あり	13	19	0	32
	両方訓なし	2	3	1	6
計		72	27	1	100

表 1，表 2 ともに単位は語数。宮島(1969:42-43)の抜粋。

「雑誌 90 種」における高頻度の二字漢語 100 語を一字漢語の独立用例で分類したのが表 1，訓読みの用例で分類したのが表 2 である。それぞれの表の左下の色がついている部分と太線で囲まれている部分は筆者がつけた。色がついている部分は，「郵便報知」では用例が確認されたが「雑誌 90 種」では用例が確認されなかった漢字を含む語であり，「郵便報知」から「雑誌 90 種」の間に，二字漢語における構成漢字の意味がとりにくい方に変化した語であると考えられる。反対に右上の太線で囲まれている部分は「郵便報知」では用例が確認されなかったが「雑誌 90 種」では確認された漢字を含む語であり，構成漢字の意味がとりやすい方に変化した語であると捉えることができる。

宮島(1969)も指摘しているように，表 1 の「郵便報知」では「両方独立」，すなわち両方の構成漢字に「一字漢語の独立用例」があった語は 41 語，「一方が独立」39 語，「両方非独立」20 語であったが，「雑誌 90 種」では，それぞれが 18 語，40 語，42 語であり，「両方独立」と「両方非独立」の数字がほぼ逆転している。逆転の内容を見ると，「郵便報知」で「両方が独立」の 41 語のうち，「雑誌 90 種」で「両方独立」の語は 15 語で，残りの 26 語は構成漢字の「一方が独立」(16 語)，あるいは「両方非独立」(10 語)の語である。表 2 においても「郵便報知」では，「両方訓あり」の 72 語のうち，「雑誌 90 種」でも「両方訓あり」の語は 57 語で，残りの 15 語は構成漢字の一方，あるいは両方に訓読みの用例がなかった。表 1，表 2 の結果について宮島(1969)の指摘は次の 2 点である。1 点目は「郵便報知」よりも「雑誌 90 種」の方が，一字漢語が独立して用いられる漢字(表 1)，訓読みで用いられる漢字(表 2)のいずれにおいても用例のある構成漢字の数が減少傾向にある。これは，現代では，二字漢語の要素の意味がとりにくくなっている語が増えたことを意味すること，2 点目は特に一字漢語(43 語)の方が訓読み(18 語)よりも減少幅の大きいことである。

宮島(1969)はこれらの変化が影響し，二字漢語が造語されにくくなってきたと述べている。本稿では，宮島(1969)をもとに構成漢字の用例が減少することは，部分的，間接的にではあるが，二字漢語構成する漢字の造語力が低下する方向に影響を与えるものと考える。

３．研究課題

「郵便報知」は明治 10 年から 11 年(1877-78 年)の 1 年間の新聞の資料，「雑誌 90 種」は昭和 31 年(1956 年)に発行された雑誌の資料をもとにしたデータであり，宮島(1969)は約 80 年間の変化を捉えたものだと言える。では「雑誌 90 種」以降の変化はどのようになっているのだろうか。本稿では，「雑誌 90 種」のデータと『現代日本語書き言葉均衡コーパス』(以下，BCCWJ と略す)の「出版・新聞(コア・非コア，以下『新聞』とする)」および「特別目的・知恵袋(コア・非コア，以下『知恵袋』とする)」を宮島(1969)と同様の方法で比較し，さらに，宮島(1969)の結果と合わせて比較することによって「雑誌 90 種」以降の約 50 年間の漢字の造語力の変化の傾向を捉える資料としたい[4]。表記における国語施策では「雑誌 90 種」と BCCWJ の間には，昭和 47 年(1972 年)に「当用漢字音訓表」，昭和 56 年(1981 年)に「常用漢字表」が告示されている。これらの施策は佐竹(2004)によると，昭和 22 年(1947 年)の「当用漢字表」で定められた制限を見直す「規制の緩和(佐竹 2012:46)」という流れの中での施策であるという。

BCCWJ の「新聞」と「知恵袋」を比較の対象としたのは，新聞は表記の基準が定められており，また多くの人に読まれる媒体であることから，調査時点の統制された規範的な表記のデータとして捉えることができる。一方，「知恵袋」は表記の基準は定められておらず，書き手が自由に表現した表記のデータと捉えることができる。そのため，別々に比較をすることで「雑誌 90 種」以降の傾向が捉えられるのではないかと考えたからである。

本稿では「新聞」と「知恵袋」それぞれのレジスターにおける高頻度の二字漢語 100 語を調査対象とし，「雑誌 90 種」と比較する。

４．調査方法・手順

４．１　宮島(1969)の調査方法

宮島(1969)では，「郵便報知」と「雑誌 90 種」の比較を次のように行っている。まず，「郵便報知」と「雑誌 90 種」の延べ語数に違いがあるため，「雑誌 90 種」の 9 分の 2 の範囲のみを調査対象とすることによって，2 つの語数をそろえている[5]。次に「雑誌 90 種」における高頻度の二字漢語 100 語を調査対象とし，構成漢字の用例の有無をまとめている[6]。「一字漢語の独立用例」については，β 単位にもとづき，漢字一字が一語を表す用例と「信ずる」「愛する」のように一字漢語に「する」が付いて動詞として用いられる例，「単に」「特に」のように副詞的に用いられる例，「訓読みによる用例」については，漢字が訓読みで読まれる例が用例として数えられている。本稿の調査も宮島(1969)と同様の方法で進めた。

[4] BCCWJ の「出版・新聞」は 2001 年から 2005 年に発行された新聞のデータであり，「特別目的・知恵袋」は 2005 年のデータである。

[5] 宮島(1969)では「郵便報知」の延べ語数が 99,384，「雑誌 90 種」が 438,135 であるため，「雑誌 90 種」を 9 分の 2 の範囲にする方法として，「全部で九段階に分かれている調査範囲の、第一、第二段階だけを対象とした。」と書かれている。

[6] 「郵便報知」と「雑誌 90 種」で出現する「一字漢語の独立用例」「訓読みによる用例」をすべて調べた結果の中から，「雑誌 90 種」における高頻度語 100 語の構成漢字の結果を抽出している。

４．２　本稿の調査方法・手順

４．２．１　調査データ

「雑誌 90 種」のデータは国立国語研究所(1997)のフロッピーディスク版（以下，FD 版とする)を用いた。FD 版は国立国語研究所(1962，1963，1964)の調査で採集された β 単位の全語が採録され，リスト化された資料である。項目は語彙素読み順で並べられ，語種，品詞情報と語例が示されている。語例の欄から各語に対する全表記と，それぞれの表記に対する頻度の情報が得られる。例えば「いれる(和語，用の類[7])」の項目は，総頻度 386 であり，表記と頻度は，いれる 36，入る 1，入れる 342，容れる 6，淹れる 1 であることがわかる。

「新聞」「知恵袋」のデータは BCCWJ の「出版・新聞(コア・非コア)」および「特別目的・知恵袋(コア・非コア)」である。「雑誌 90 種」(β 単位)と BCCWJ の「新聞」「知恵袋」(短単位)は延べ語数が異なる[8][9](表 3)。「雑誌 90 種」のサイズに合わせるために BCCWJ の連番を利用した。連番はサンプル内での長単位の並び順を示し，10 きざみで付けられている[10]。本稿では「新聞」は連番の下 2 桁が 30，60，90 のデータ，「知恵袋」は連番が 250 の倍数のデータのみを抽出し，「新聞」の約 33.3%，「知恵袋」の約 4.0%を調査データとした。

表 3　「雑誌 90 種」「新聞」「知恵袋」の延べ語数

	延べ語数
「現代雑誌 90 種[11]」	438,760
「新聞(コア・非コア)」	1,370,233
「知恵袋(コア・非コア)」	10,256,877

４．２．２　調査対象の二字漢語と漢字

「BCCWJ 短単位語彙表」を利用し，「新聞」「知恵袋」それぞれのレジスターにおいて，漢字二字から成る漢語を高頻度順に並べ，上位 100 語とその 100 語を形成する漢字を調査対象とした[12]。各レジスターの語数と漢字数を表 4 に示す。

表 4　調査対象の二字漢語と構成漢字

	延べ		異なり	高頻度の漢字
	語数	漢字数	漢字数	
「新聞」	100	200	151	会 5，民 4，議 3，国 3，対 3，社 3，年 3
「知恵袋」	100	200	167	以 3，意 3，一 3，間 3，分 3
重複	25	50	47	学 2，校 2，時 2

7 品詞として，『分類語彙表』の 4 分類(体の類，用の類，相の類，その他)が示されている。

8 レジスター別の短単位語数および年の情報は「BCCWJ/短単位語数」のページから得た。

9 BCCWJ の短単位は β 単位を元に設計されたものである(山崎編 2014)ため，本稿における二字漢語の調査については β 単位と短単位を同様の単位とみなして扱う。

10 国立国語研究所(2015)第 6 章参照

11 国立国語研究所(1962:314)では，延べ語数は 438,135 語であるが，本稿では調査に使用した FD 版の頻度を合計した数を用いた。

12 数詞(例:二十)と「箇月」は除外した。宮島(1969:41)も同様である

「新聞」は延べ語数 100 語，延べ漢字数 200 字，異なり漢字数 151 字，「知恵袋」は延べ語数 100 語，延べ漢字数 200 字，異なり漢字数 167 字であった。「新聞」と「知恵袋」で重複する語は 25 語，重複する漢字は，47 字(異なり)であった。「新聞」と「知恵袋」で重複する語は 25.0%，重複する漢字は異なりで 30%前後であり，「新聞」と「知恵袋」では高頻度の漢字は異なっている。

４．２．３　手順

本稿で調べるのは「一字漢語の独立用例」の有無と「訓読みによる用例」の有無である。「雑誌 90 種」と「新聞」「知恵袋」について，以下の手順で漢字の用例を抽出し，確認した。抽出の際には，固有名詞，地名，人名は除外した。具体的には，「雑誌 90 種」では，語種が「人」「姓」「姓名」「地名」「名」の項目，「新聞」「知恵袋」では，品詞が「名詞－固有名詞」の語を除外した。「明日(あす)」などの熟字訓や当て字は，漢字に対応する読み方が決められないため，用例には含めなかった。

「雑誌 90 種」では，FD 版のリストの語例の欄を使って調査対象の構成漢字が使われている語を抽出した。語種の情報を参考に「一字漢語」は，漢字一字かつ音読みで用いられている例，「関する」「対する」のように一字漢語に「する」が付いて動詞として用いられる例，「現に」「真に」「確たる」のように副詞的，連体詞的に用いられる例を確認し，また「訓読み」は，漢字が訓読みで読まれている語の例を確認した。

「一字漢語」について，β 単位の一字漢語には，「長官並みの格で」の「格」のような一字漢語が独立して用いられているものや，「代表格」の「格」のような他の語と共に合成語を形成しているもの，「西洋化」の「化」や「多方面」の「多」など接辞として用いられているもの，助数詞が含まれている。本稿で確認するのは一字漢語が独立して用いられた例で，上述の例の中の「長官並みの格で」の「格」のような例のみである。これについて，「雑誌 90 種」のリストでは十分な情報が得られなかったため，「雑誌 90 種」のリストに一字漢語の例がある語については，原文[13]を確認した。

「新聞」「知恵袋」については，コーパス検索アプリケーション『中納言』の短単位検索を用い，書字形と品詞を条件に，調査対象の漢字が含まれている語を抽出した[14]。検索した結果から，「新聞」は連番の下 2 桁が 30，60，90，「知恵袋」は連番が 250 の倍数の結果のみを分析に用いた。書字形と語彙素読みをもとに，一字漢語および訓読みの用例の有無を確認した。一字漢語については，書字形が当該の漢字一字である用例を抽出した後で，「雑誌 90 種」と同様に，用例を目視しながら，前後に名詞が連続しない語[15]を確認した。ただし，前後に数詞が続く場合には，「前 11・0」のように午前の略語の「前」の前後に数詞が続くもののみ用例とみなした。略語については後述する。漢字一字が独立して用いられている例でも，例えば「生」が「セイ」「なま」のいずれであるかなどは短単位検索の結果だけでは十分な情報が得られなかったため，語種の情報は利用しなかった。また，中には前後の文脈を読んでも漢字一字の語が漢語であるか和語であるか判断できない用例もあった

[13] 国立国語研究所(1987)を用いた。この資料はマイクロフィッシュに採集カードが記録されている。

[14] キーの条件式は，例えば「人」の場合は「キー: (書字形 LIKE "%人%" AND NOT 品詞 LIKE "名詞-固有名詞%")」である。

[15]名詞が前接する語で一字漢語の独立用例とみなした語は 1 語「別冊環」で，これは『環(カン)』という雑誌の名前である。

が，それによって今回の調査の「用例の有無」にもとづく分類に影響を与えることはなかった。

４．２．４　略語の扱い

一字漢語の略語の扱いについて，宮島(1969)では詳細に述べられていないが，「一字漢語の独立用例」に「日曜日」を意味する「日」が含まれていること，「安全保障」を略した「安保」は「安全保障」という語の存在を前提とした分割不可能な一語であると記されていることから，本稿では漢字一字で用いられる略語のみを一字漢語の独立用例とし，前述のように前後に「数詞以外の名詞」が連続しない語を用例とした。抽出した略語には，「注も親切」のような助詞が後接する語以外に，「(注)」など前後が補助記号や空白のものも含まれる。略語の用例には「注意」を意味する「注」や，「雑誌 90 種」では法令に用いられる「政(政令の略語)」，「新聞」では「電話番号」の意味の「電」，政党名「自民党」を意味する「自」の用例などが確認された[16]。企業名を略した「トヨタ自」や「三菱電」の「自」「電」は社名の一部とみなし，一字漢語の独立用例とはしなかった。用例の中に「日」が「日曜日」を表わす例と「日本語」を表す例があるように，略語の意味は文脈によって異なる場合がある。また，書籍名に含まれる上下巻を表す「上」は上巻の略語，新聞の見出しの一部に使われていた「ルポ　中」の「中」は「中編」の略語の例とみなしたが，「中」の場合，連続した 3 編(上中下)の 2 番目を表す記号的な使われ方だと捉えることもできる。一字漢語の略語の扱いについては今後の課題としたい。

５．結果と考察

まず，「雑誌 90 種」と「新聞」，次に「雑誌 90 種」と「知恵袋」を比較し本稿の研究課題である「雑誌 90 種」以降の変化を概観する。その後で宮島(1969)の「郵便報知」と「雑誌 90 種」の比較結果を合わせて「郵便報知から雑誌 90 種」の変化と「雑誌 90 種以降」の変化を比較する。

５．１　「雑誌 90 種」と「新聞」の比較(表 5，6)

「雑誌 90 種」と「新聞」の比較結果を表 5，表 6 に示す。表 5 は「一字漢語の独立用例」の有無，表 6 は「訓読みによる用例」の有無による比較である。章末の資料 1，2 に調査対象の全語を入れた表を示す。まず，「一字漢語の独立用例」について表 5 を見ると，「雑誌 90 種」では，「両方が独立」が 32 語，「一方が独立」が 42 語，「両方非独立」が 26 語であ

[16] 本件研究では後述の分析で「雑誌 90 種」と「新聞」，「雑誌 90 種」と「知恵袋」についてそれぞれ比較を行った際，「一字漢語の独立用例」の有無によって二字漢語を分類した。この分類に影響を与える略語，つまり，ある構成漢字に対し「一字漢語の独立用例」が略語のみであった構成漢字と，その漢字を使った略語は，以下の通りである。構成漢字数に大きな違いが見られないことから，本調査の略語の捉え方は本稿の結論には影響しないと判断した。なお，下記の「政 3」の 3 は構成漢字の延べ字数を表す。
【雑誌 90 種と新聞】【雑誌 90 種】延べ 9 字，異なり 7 字，【新聞】延べ 5 字，異なり 5 字，
【雑誌 90 種】改(改進党)，首(首相)，上(上級者)，政 3(政令)，前(午前)，中(中級者，中学生)，明(明治)／【新聞】自(自民党)，上(上巻)，前(午前)，中(中編)，電(電話番号)
【雑誌 90 種と知恵袋】【雑誌 90 種】延べ 6 字，異なり 6 字，【知恵袋】延べ 4 字，異なり 4 字【雑誌 90 種】外(外相)，上(上級者)，前(午前)，注(注意)，日(日曜日)，明(明治)／【知恵袋】注(注意)，日(日本語)，法(法学部)，予(予定)

るのに対し，「新聞」ではそれぞれ 17 語，44 語，39 語であり，「両方独立」は 15 語減り，「一方が独立」と「両方非独立が」それぞれ 2 語と 13 語増えている。また，表 5 の左下の色がついている部分は，「雑誌 90 種」では用例が確認されたが「新聞」では構成漢字の一方，あるいは両方の用例が確認されなかった語数であり，全部で 31 語ある。また，右上の太線で囲まれている部分は，「雑誌 90 種」では構成漢字の用例が確認されなかったが「新聞」では確認された語であり，合わせて 4 語である。左下の色がついている部分の方が語数が多く，用例を有する構成漢字が減る傾向にあることが読み取れる。

次に，表 6 の「訓読みによる用例」を見ると，「雑誌 90 種」では，「両方訓あり」が 73 語，「一方に訓あり」が 23 語，「両方訓なし」が 4 語であるのに対し，「新聞」ではそれぞれ 56 語，35 語，9 語であり，「両方訓あり」が 17 語減り，「一方に訓あり」と「両方訓なし」がそれぞれ 12 語と 5 語増えている。また，色がついている部分が 24 語，太線で囲まれた部分が 2 語であることから，訓読みについても用例を有する構成漢字は減少傾向にあることがわかる。以上のことから，「雑誌 90 種」と「新聞」を比較すると，「一字漢語の独立用例」と「訓読みによる用例」はいずれも減少傾向にあることが確認される。

表 5　「雑誌 90 種」と「新聞」の一字漢語

		雑誌 90 種			計
		両方独立	一方が独立	両方非独立	
新聞	両方独立	16	1	0	17
	一方が独立	15	26	3	44
	両方非独立	1	15	23	39
計		32	42	26	100

表 6　「雑誌 90 種」と「新聞」の訓読み

		雑誌 90 種			計
		両方訓あり	一方に訓あり	両方訓なし	
新聞	両方訓あり	54	2	0	56
	一方に訓あり	19	16	0	35
	両方訓なし	0	5	4	9
計		73	23	4	100

表 5，表 6 とも単位は語数。表 7，表 8 も同じ。

５．２「雑誌 90 種」と「知恵袋」の比較（表 7，8）

「雑誌 90 種」と「知恵袋」を比較した表を表 7，8 に，全語を入れた表を章末資料 3，4 に示す。まず「一字漢語の独立用例」について表 7 を見ると，「雑誌 90 種」では，「両方独立」が 26 語，「一方が独立」が 49 語，「両方非独立」が 25 語であるのに対し，「知恵袋」ではそれぞれ 12 語，46 語，42 語である。「両方独立」は 14 語，「一方が独立」も 3 語減少しているのに対し，「両方非独立」は 17 語増加している。また，左下の色がついている部分は，合わせて 31 語あり，右上の太線で囲まれている部分は合わせて 4 語である。このことから用例を有する構成漢字が減る傾向があることが読み取れる。次に「訓読みによる用例」について表 8 を見ると，「雑誌 90 種」では，「両方訓あり」が 63 語，「一方に訓あり」が 33 語，「両方訓なし」が 4 語であるのに対し，「新聞」ではそれぞれ 44 語，44 語，12 語であり，「両方訓あり」が 19 語減り，「一方に訓あり」と「両方訓なし」がそれぞれ 11 語と 8 語増えている。また，色がついている部分が 29 語，太線で囲まれた部分が 2 語であることから，訓読みについても用例を有する構成漢字は減少傾向にあることがわかる。このように「雑誌 90 種」と「知恵袋」を比較すると「一字漢語の独立用例」と「訓読みによる用例」はいずれも減少傾向にあることが読み取れる。

以上の「雑誌 90 種」と「新聞」，「雑誌 90 種」と「知恵袋」の比較結果から，「雑誌 90 種」以降では「新聞」「知恵袋」いずれにおいても「一字漢語の独立用例」および「訓読み

による用例」の減少傾向が見られ，二字漢語は構成漢字の意味がとりにくい方に変化していることが示唆される。この結果は部分的，間接的にではあるが，「新聞」「知恵袋」の高頻度の二字漢語を構成する漢字の造語力が低下する傾向にあることを表すものであると考える。

表 7　「雑誌 90 種」と「知恵袋」の一字漢語

		雑誌 90 種			計
		両方独立	一方が独立	両方非独立	
知恵袋	両方独立	11	1	0	12
	一方が独立	11	32	3	46
	両方非独立	4	16	22	42
計		26	49	25	100

表 8　「雑誌 90 種」と「知恵袋」の訓読み

		雑誌 90 種			計
		両方訓あり	一方に訓あり	両方訓なし	
知恵袋	両方訓あり	42	2	0	44
	一方に訓あり	21	23	0	44
	両方訓なし	0	8	4	12
計		63	33	4	100

５．３　「郵便報知」と「雑誌 90 種」，「雑誌 90 種」以降の比較

宮島(1969)が行った「郵便報知」と「雑誌 90 種」の比較結果(表 1，2)および「雑誌 90 種」以降の比較結果を合わせ，変化の傾向をさらに検討する。宮島(1969)の調査結果は「雑誌 90 種」の 9 分の 2 の語数で行った調査であるため，「両方独立」や「一方が独立」の語数による単純な比較は難しく，また，宮島(1969)には構成漢字すべての用例の有無について情報が記載されていない。そのため，表 1，2，5~8 での語の分類結果をもとに変化の度合いを比較することにする。

表 1，2，5~8 は，各データの比較において，新しいデータ[17]の方が古いデータよりも構成漢字の用例が減少傾向にあることを示している。しかし，二字漢語において構成漢字の用例が一方のみ変化したのか，両方変化したのかによっても，変化の程度は異なる。そこで，各二字漢語の変化の度合いを構成漢字の用例の有無に変化があった語，具体的には，表 1，2，5~8 の色がつけられている部分および太線で囲まれている部分の語数をまとめ，変化の度合いを表すことによって比較することにした。まとめ方は各比較において，古い方のデータでは構成漢字の用例があったが，新しい方のデータでは用例が確認されなかった場合は，構成漢字の意味がとりにくくなる方向への変化でマイナス方向への変化とし，古い方のデータでは構成漢字の用例が確認されなかったが，新しい方のデータでは確認された場合は，構成漢字の意味がとりやすくなる方向への変化で，プラス方向への変化とした。構成漢字一方のみの変化は－1 または＋1，両方の変化は－2 または＋2 とした(表 9)。

表 9　表 1，2，5~8 の変化のまとめ方(例　一字漢語の独立用例)

		古いデータ		
		両方独立	一方が独立	両方非独立
新しいデータ	両方独立		+1	+2
	一方が独立	-1		+1
	両方非独立	-2	-1	

[17] 例えば「雑誌 90 種」と「新聞」の比較で新しいデータは「新聞」，古いデータは「雑誌 90 種」を指す。

表 9 は「一字漢語の独立用例」の例であるが，「訓読みによる用例」も同様である。集計は，それぞれの欄に当てはまる語数と表 9 の値を掛けて，プラス方向とマイナス方向でそれぞれ集計し，さらにその値を合計した。

図 1，2 に「一字漢語の独立用例」(表 1，5，7)と「訓読みによる用例」(表 2，6，8)の変化の値を集計したものをまとめた。縦軸は変化の度合いを示した値である。まず，プラス方向とマイナス方向の変化の度合いを示し，「計」ではそれの合計を示した。図中の「郵報」は「郵便報知」，「雑 90」は「雑誌 90 種」，「知恵」は「知恵袋」の略である。「郵便報知」と「雑誌 90 種」は「雑誌 90 種」の 9 分の 2 の語数での調査であるため単純に比較できないが，「郵便報知」と「雑誌 90 種」，「雑誌 90 種」と「新聞」，「雑誌 90 種」と「知恵袋」は「一字漢語の独立用例」と「訓読みによる用例」の両比較においてプラス方向よりもマイナス方向への変化が大きい。この結果から，「新聞」「知恵袋」の高頻度の二字漢語を構成する漢字は，意味がとりにくいほうに変化し，このことから部分的，間接的にではあるが，構成漢字の造語力は低下傾向にあると考えられる。

また，「一字漢語の独立用例」と「訓読みによる用例」では変化の傾向が異なることが読み取れる。「一字漢語の独立用例」は，「訓読みによる用例」よりも変化の幅が大きく，図 1 を見ると「郵便報知」と「雑誌 90 種」の間の約 80 年間に「一字漢語の独立用例」は-45 と大きく減少し，「雑誌 90 種」以降の約 50 年間においては「雑誌 90 種」と「新聞」が-28，「雑誌 90 種」と「知恵袋」が-31 であり，「郵便報知」と「雑誌 90 種」の間ほどではないが，減少傾向が続いていることがわかる。「一字漢語の独立用例」が減少していることについて「雑誌 90 種」と「新聞」および「知恵袋」を比較すると，「雑誌 90 種」の中にはやや古い表現のテキストが見られることが理由の 1 つとして考えられる。例えば，「雑誌 90 種」では「学」の用例に「禅僧が，元に留学し，学成って日本に帰る時，師の南洲文藻という禅林に別れの偈を乞うた」という用例が見られた。このようなテキストでは，一字漢語が選ばれやすい可能性があると考える。一字漢語が独立して用いられる場合，「本(ホン)」や「文(ブン)」などのように，身近な普通名詞の語もあるが，上述の「学」や「再燃せしめる因をなした」の「因」，「どの程度まで効を挙げるか」における「効」など，やや硬い文体で用いられやすいと思われる抽象的な意味の一字漢語の用例も見られた。

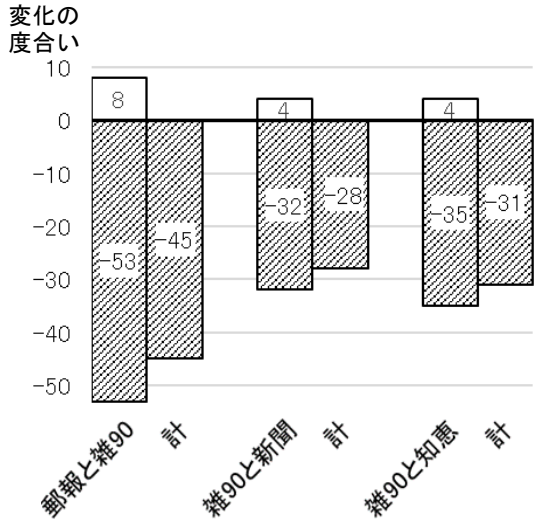


図 1　一字漢語の独立用例による変化の度合い

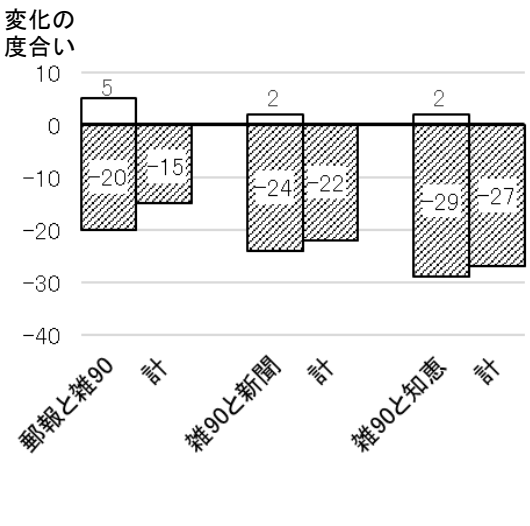


図 2　訓読みの用例による変化の度合い

また，表 5，表 7 を見ると，「雑誌 90 種」と「新聞」，「雑誌 90 種」と「知恵袋」では，レジスターが異なるにもかかわらず，変化の値に大きな違いがないことは興味深い。「新聞」と「知恵袋」は調査対象語において重なる語が 100 語中 25 語（表 4 参照）であり，重複する異なり漢字も 30%程度で，重複しない漢字の方が多い。「新聞」と「知恵袋」を比較するには語数を増やした比較が必要だと思われるため，両者の比較は今後の課題としたい。

図 2 の「訓読みによる用例」の有無の変化については，「一字漢語の独立用例」ほどではないが，「郵便報知」と「雑誌 90 種」の間の約 80 年間においても，「雑誌 90 種」以降の約 50 年間においても減少傾向にあり，構成漢字の意味はとりにくい方に変化していることがわかる。また「郵便報知」と「雑誌 90 種」の間の約 80 年間の減少幅よりも「雑誌 90 種」以降の約 50 年間における減少幅の方が大きいことが変化の特徴として挙げられる。

ここでは「雑誌 90 種」以降の「訓読みによる用例」の減少傾向について，「訓読みの制限の影響」と「レジスターにおける語彙の違い」の 2 点について述べる。図 2 の「雑 90 と新聞」の変化を表す値 2 と-24 は，「雑誌 90 種」では用例がなかったが，「新聞」では用例が確認された構成漢字が延べ 2 字，反対に「雑誌 90 種」では用例があったが，「新聞」では用例が確認されなかった構成漢字が延べ 24 字あったことを示している。図 2 にまとめられた構成漢字がどのような訓読みで読まれているかを表 10 に示す。漢字の後ろの数字は，同じ漢字が複数回用いられたときの回数を示している。例えば，「雑誌 90 種と新聞」の「画 2」は，「映画」と「計画」の 2 語に用いられていることを意味する（章末資料 2 参照）。表 10 における当用漢字表（1947 年）の表外の訓読みは全語が常用漢字表（1981 年）の表外の訓読みでもあった。訓読みを表外（以下「当用・常用外」とする）と表内（以下「当用・常用内」とする）に分けて示す。

表 10　「雑誌 90 種」と「新聞」，「雑誌 90 種」と「知恵袋」における訓読みの有無

	「新聞」あるいは「知恵袋」のみに用例がある訓読み	「雑誌 90 種」に用例があり，「新聞」あるいは「知恵袋」に用例がない訓読み
「雑誌 90 種」と「新聞」	【当用・常用内】 （延べ 2，異なり 2） 授:授かる（さずかる）， 務:務まる（つとまる）	【当用・常用外】（延べ 21 字，異なり 13 字）　活 2:活かす（いかす），容:容れる（いれる），午 2:午（うし，ひる），画 2:画（え），件:件（くだん），術:術（すべ），発 2:発つ（たつ），文:文（ふみ），政 3:政に（まさに），対 3:対う（むかう），以:以て（もって），能:能い（よい），環:環（わ） 【当用・常用内】（延べ 3 字，異なり 2 字）　究:究める（きわめる），業 2:業（わざ）
「雑誌 90 種」と「知恵袋」	【当用・常用外】 （延べ 1，異なり 1） 録:録る（とる） 【当用・常用内】 携:携わる（たずさわる） （延べ 1，異なり 1）	【当用・常用外】（延べ 16 字，異なり 12 字）　購:購う（あがなう），画:画（え），了:了える・了る（おえる・おわる）検:検べる（しらべる），銀:銀（しろがね），質:質ねる（たずねる），番 2:番い（つがい），勉:勉める（つとめる），勿:勿れ（なかれ），対 2:対う（むかう），以 3:以て（もって），能:能い（よい） 【当用・常用内】（延べ 13 字，異なり 11 字）　商:商い（あきない），字:字（あざ），価 2:価（あたい），険:険しい（けわしい），便:便り（たより），説:説く（とく），額:額（ひたい），実:実（み），本 2:本（もと），社:社（やしろ），由:由（よし）

まず，「訓読みの制限の影響」について表 10 を見ると，「雑誌 90 種」に用例があり，「新聞」あるいは「知恵袋」に用例がない訓読みには，「当用・常用外」のものが目立つ。「雑誌 90 種」と「新聞」の欄を見ると，「雑誌 90 種」に用例があり「新聞」に用例がない訓読みの中で「当用・常用外」は延べ 21 字，異なり 13 字であり，「当用・常用内」を加えた延べ 24 字，異なり 15 字のうち，それぞれ 87.5%と 86.7%を占める。同様に「雑誌 90 種」と「知恵袋」では，「雑誌 90 種」に用例があり「知恵袋」に用例がない訓読みの中で「当用・常用外」は延べ 16 字，異なり 12 字であり，「当用・常用内」を加えた延べ 29 字，異なり 23 字のうち，それぞれ 55.2%，52.2%を占める。「雑誌 90 種」は 1956 年発行の雑誌を調査したものであり，当用漢字表が定められた後の調査であるが，表 10 を見ると「雑誌 90 種」では当用漢字表以前の表記慣習が残っており，「新聞」「知恵袋」までの間に当用漢字表による訓読みの制限が定着してきた可能性が考えられる。

次に「レジスターにおける語彙の違い」について，「雑誌 90 種」に用例があり，「新聞」および「知恵袋」では用例がない訓読みは，「新聞」「知恵袋」にその和語が使われていないか，あるいは語としては使われているが当該の漢字が表記に用いられていない可能性がある。このことを確認するために，表 10 において，「雑誌 90 種」に用例があり，「新聞」あるいは「知恵袋」に用例がない訓読みについて，「新聞」「知恵袋」を調べた。『中納言』を用い語彙素読みをキーに検索し[18]，検索結果から「語彙素」の表記と「書字形」を利用して，明らかな同訓異義語を除いた。前後の文脈を詳細に見ていないため大まかな確認ではあるが，表中の下線は「新聞」「知恵袋」では使用例が確認できなかった語である。「雑誌 90 種」と「新聞」の異なり 15 字の訓読み（「当用・常用外」13 字，「当用・常用内 2 字」）のうち，語として出現していなかったものは「ふみ」1 語であった。14 字の訓読みについては，平仮名表記，あるいは他の漢字を用いた表記による使用例が確認された[19]。一方，「雑誌 90 種」と「知恵袋」の異なり 23 字の訓読みのうち，語として出現していなかったものは，12 語（あがなう，あきない，あざ，けわしい，しろがね，たより，つがい，とく，なかれ，ひたい，やしろ，よし）[20]であった。11 字の訓読みについては，平仮名表記，あるいは他の漢字を用いた表記が確認された。ただし，平仮名表記，あるいは他の漢字を用いた表記が確認された訓読みも，本稿における調査対象漢字の訓読みと同じ意味で用いられているか否かについては，用例の詳細な検討が必要である。また，例えば「知恵袋」では，「あがなう」「あきない」の用例は確認されなかったが，同様の意味を表す「購入」「商売」の用例は確認されたことから，「知恵袋」に和語の使用例がないことは，単に話題の違いだけではない可能性もある。

６．まとめと課題

本稿では，「郵便報知」と「雑誌 90 種」の約 80 年の間に二字漢語における構成漢字の意味がとりにくくなってきており，そのことによって新しい二字漢語が作られにくくなったという宮島（1969）の指摘と調査をもとに，「雑誌 90 種」と BCCWJ の「新聞」および「知恵袋」を比較し，「雑誌 90 種」以降の傾向を捉えることを試みた。

18 条件式は，例えば「画（え）」の場合，キー：（語彙素読み="エ" AND 品詞 LIKE "名詞%"）である。

19 例えば，「いれる」には「いれる，入れる」の 2 種類の表記が確認された。

20 「あざ」「たより」「とく」「み」「よし」は用例が「痣」「頼り」「解く，溶く」「身」「良し」の例である。

本稿では，宮島(1969)と同様の方法で「一字漢語の独立用例」の有無と「訓読みによる用例」の有無を「新聞」「知恵袋」における高頻度の二字漢語について調べ，「雑誌 90 種」と「新聞」，「雑誌 90 種」と新聞の比較し，さらに宮島(1969)の結果との比較を行った。その結果，以下の 2 点が確認された。1 点目は「雑誌 90 種」以降も「一字漢語の独立用例」と「訓読みによる用例」を有する構成漢字は減少傾向にあること，2 点目は，明治期からの変化は，「一字漢語の独立用例」の方が「訓読みによる用例」よりも減少幅が大きいが，「訓読みによる用例」は，「雑誌 90 種」以降の方が「雑誌 90 種」以前よりも減少幅がやや大きいことである。このことにより，高頻度の二字漢語については，構成漢字の意味が捉えにくく構成漢字に分解して意味を捉えることが困難になる傾向があることが示唆される。漢字の造語力を検討するには，構成漢字一字の用例による検討のみでは十分ではなく，当該の漢字を含む二字漢語の語数など，構成漢字の実際の使用状況についても考える必要がある。しかし，漢字の意味がとりにくくなることは，それらの漢字を用いた造語が行われにくくなること，すなわち構成漢字の造語力の低下傾向を部分的，間接的に示していると考える。「雑誌 90 種」以降の「一字漢語の独立用例」と「訓読みによる用例」を有する構成漢字が減少傾向にあることの原因としては，「雑誌 90 種」と「新聞」「知恵袋」との文体の違い，訓読みの制限の影響などが見られたが，レジスターによる使用語彙の違いなど，検討課題も確認された。

本調査の課題として，2 点挙げる。1 点目は他のレジスターとの比較である。「雑誌 90 種」と「新聞」，「雑誌 90 種」と「知恵袋」の変化の度合いを比較すると，レジスターが異なるにもかかわらず，変化の値に大きな違いが見られなかったことを述べた。他のレジスターとも比較することによって，変化の傾向をさらに検討できるのではないかと考える。2 点目は，調査対象語の中には，語の意味と語の漢字表記が離れてしまっていると思われるものがある。例えば，「沢山」の「沢」の訓読みの用例は水が流れる沢の例であり，「印象」の「象」の一字漢語の用例は動物の象の用例であった。宮島(1969)は，漢字一字の意味が明確なことが造語や語の理解につながると述べており，本稿では宮島(1969)と同様の基準で調査を行ったが，二字漢語の意味と構成漢字の意味の結びつきの程度も考慮することが必要だと思われる。

謝　辞

本研究を進めるにあたり，山崎誠先生(国立国語研究所)にご指導を賜りました。深く御礼申し上げます。

文　献

国立国語研究所(1959)『明治初期の新聞の用語』国立国語研究所報告 15，秀英出版
(http://db3.ninjal.ac.jp/publication_db/item.php?id=100170015 よりダウンロード可能)
国立国語研究所(1962)「現代雑誌九十種の用語用字 第一分冊：総記および語彙表」国立国語研究所報告 21，秀英出版
(http://db3.ninjal.ac.jp/publication_db/item.php?id=100170021 よりダウンロード可能)
国立国語研究所(1963)「現代雑誌九十種の用語用字 第二分冊：漢字表 」国立国語研究所報告 22，秀英出版
(http://db3.ninjal.ac.jp/publication_db/item.php?id=100170022 よりダウンロード可能)
国立国語研究所(1964)「現代雑誌九十種の用語用字 第三分冊：分析 」国立国語研究所報

告 25，秀英出版
（http://db3.ninjal.ac.jp/publication_db/item.php?id=100170025 よりダウンロード可能）
国立国語研究所（1987）『現代雑誌九十種の用語用字五十音順語彙表・採集カード』国立国語研究所言語処理データ集 3，東京都板橋福祉工場
国立国語研究所（1997）『現代雑誌九十種の用語用字全語彙・表記』[FD 版]国立国語研究所言語処理データ集 7，三省堂
（http://pj.ninjal.ac.jp/corpus_center/archive.html よりダウンロード可能）
国立国語研究所（2015）「形態論情報付きデータ（TSV）」『現代日本語書き言葉均衡コーパス利用の手引 第 1.1 版第 6 章
（http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual_06.pdf よりダウンロード可能）
佐竹秀雄（2012）「漢字と表記」前田富祺，野村雅昭（編）『朝倉漢字講座 2（漢字のはたらき）』普及版，朝倉書店，pp.44-64.
文化審議会（2010）『改定常用漢字表（答申）』文化庁ホームページ
（http://www.bunka.go.jp/seisaku/bunkashingikai/sokai/sokai_10/pdf/kaitei_kanji_toushin.pdf よりダウンロード可能）
宮島達夫（1969）「近代日本語における漢字の位置」『教育国語』16，麦書房，pp.17-44.
山崎誠（編）（2014）『講座日本語コーパス 2 書き言葉コーパス―設計と構築―』朝倉書店

関連 URL

「BCCWJ/短単位語数」
https://maro.ninjal.ac.jp/wiki/index.php?BCCWJ%2F%E7%9F%AD%E5%8D%98%E4%BD%8D%E8%AA%9E%E6%95%B0
国立国語研究所『現代日本語書き言葉均衡コーパス』（通常版，BCCWJ-NT）
（http://pj.ninjal.ac.jp/corpus_center/bccwj/　2018 年 5~7 月利用）
国立国語研究所「『現代日本語書き言葉均衡コーパス』短単位語彙表 」（ver.1.1）
（http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html　2017 年 5 月 24 日取得）

資　料

資料 1　「雑誌 90 種」と「新聞」における「一字漢語の独立用例」の比較(表 5 参照)

		雑誌 90 種		
		両方が独立	一方が独立	両方非独立
新聞	両方が独立	映画，会社，計画，作品，市民，社会，情報，対応，対策，対象，代表，大会，地域，地方，年間，文化(16)	自分(1)	(0)
	一方が独立	可能，会議，会長，現在，社長，住民，制度，政府，説明，大学，中心，年度，発表，民主，利用（15）	以上，環境，関係，関連，期待，共同，国民，五輪，午後，午前，国際，今回，今後，昨年，市場，事件，時間，時代，実施，写真，女性，政権，生活，必要，方針，問題（26）	企業，事業，電話（3）
	両方非独立	首相（1）	改革，開発，学校，技術，議員，協議，経営，経済，結果，政治，選挙，選手，保険，優勝，予定（15）	委員，影響，家族，活動，監督，教育，教授，研究，高校，参加，指摘，支援，施設，事務，出場，女子，世界，全国，組織，調査，販売，容疑，連続（23）

※＿＿は両方に用例があった漢字，＿＿はどちらか一方に用例があった漢字を示す。資料 2～4 も同じ。

資料 2　「雑誌 90 種」と「新聞」における「訓読みによる用例」の比較(表 6 参照)

		雑誌 90 種		
		両方訓あり	一方に訓あり	両方訓なし
新聞	両方訓あり	影響，会社，会長，改革，関係，関連，共同，教育，経営，経済，結果，現在，五輪，国際，国民，今回，今後，作品，参加，市場，市民，指摘，施設，時間，時代，自分，実施，写真，社会，社長，首相，住民，出場，女子，情報，説明，選挙，選手，全国，組織，代表，大会，大学，中心，年間，年度，必要，保険，方針，民主，優勝，予定，利用，連続(54)	教授，事務（2）	(0)
	一方訓あり	以上，映画，開発，活動，環境，企業，技術，計画，研究，午後，午前，事業，事件，政治，生活，対応，発表，文化，容疑（19）	委員，家族，会議，学校，期待，高校，昨年，支援，女性，世界，制度，地方，調査，電話，販売，問題（16）	(0)
	両方訓なし	(0)	可能，政権，政府，対策，対象（5）	監督，議員協議，地域(4)

資料 3　「雑誌 90 種」と「知恵袋」における「一字漢語の独立用例」の比較(表 7 参照)

		雑誌 90 種		
		両方独立	一方が独立	両方非独立
知恵袋	両方独立	一番，金額，銀行，実際，週間，対応，注意，部分，文字，方法，無理（11）	生活（1）	(0)
	一方が独立	一緒，可能，会社，現在，情報，説明，絶対，大学，番号，本当，利用（11）	意見，意味，一般，回答，確認，簡単，関係，基本，経験，削除，使用，時間，時代，自分，質問，写真，出品，商品，人間，先生，多分，入札，必要，表示，病院，普通，毎日，問題，郵便，予定，落札，理由（32）	個人，自動，知恵（3）
	両方非独立	映画，効果，相談，心配（4）	以外，以上，以前，価格，学校，原因，時期，主人，女性，状態，男性，程度，発送，評価，保険，勿論（16）	家族，奇麗，携帯，結構，結婚，検索，購入，高校，最近，最初，参考，終了，設定，沢山，電話，登録，内容，妊娠，不安，勉強，友人，連絡（22）

資料 4　「雑誌 90 種」と「知恵袋」における「訓読みによる用例」の比較(表 8 参照)

		雑誌 90 種		
		両方訓あり	一方に訓あり	両方訓なし
知恵袋	両方訓あり	一緒，回答，確認，関係，結構，原因，現在，効果，最近，最初，削除，参考，使用，時間，時代，自動，自分，写真，主人，出品，情報，心配，人間，生活，設定，先生，多分，大学，沢山，知恵，程度，内容，入札，発送，必要，表示，毎日，友人，予定，落札，利用，連絡（42）	携帯，登録(2)	(0)
	一方訓あり	以外，以上，以前，一番，映画，会社，基本，金額，銀行，購入，質問，実際，終了，商品，説明，絶対，対応，文字，勉強，保険，本当（21）	意見，意味，一般，家族，学校，経験，結婚，高校，個人，時期，週間，女性，相談，男性，注意，電話，病院，不安，普通，部分，方法，無理，問題（23）	(0)
	両方訓なし	(0)	価格，可能，検索，番号，評価，勿論，郵便，理由（8）	簡単，奇麗，状態，妊娠(4)

方言音声に対するテキスト自動アライメントの試み

石本 祐一 (国立国語研究所コーパス開発センター) *

A Study on Automatic Alignment of Utterance Transcription for Japanese Dialect Speech

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)

要旨

音声コーパスの構築にあたり、音声に対する発話・音素・韻律などの各種ラベル付与が作業者の大きな負担となっている。この負担軽減を目的としてラベリングを自動化する試みが行われており、音声認識技術を利用した転記テキストの自動アライメントシステムがすでにコーパス構築の補助として稼働し始めている。しかし、システムの音声認識部を構成する音響モデル・言語モデルが標準語を基に設計されていることから、現在のところは標準語を主とした音声へのシステム利用にとどまっており、標準語とは異なる特性を持ちうる方言音声に対してはシステムの有効性が不明である。そこで本稿では、方言音声に対する転記テキストの自動アライメント性能について調べた結果について報告し、方言音声コーパスの構築におけるテキスト自動アライメントシステムの実用可能性について述べる。

1. はじめに

音声コーパスには様々なラベルが付与されていることが望ましいが、ラベル付与は音声・言語分野の知識を持った作業者による人手作業によるところが大きく、その負荷の高さがコーパス構築における問題となっている。そこで、国立国語研究所コーパス開発センターでは、コーパス構築の負担を軽減するべく、コンピュータによるラベル付与の自動処理の検討を進めている。その中で発話を文字で書き起こしたテキスト（以下、転記テキスト）の音声データへの配置（アライメント）については、人手での修正作業という後処理がある程度求められるものの、ほぼ実用に足ることがわかってきた。これは、音声認識による書き起こしをタイムスタンプ付きで出力する自動字幕作成システム (秋田ほか 2015, 河原ほか 2016) を応用するものであり、音声に加えて転記テキストも入力に用いることで実用的な精度で転記テキストに対応する発話の開始時刻を推定することができる (石本 2017)。現在では実際に、『日本語日常会話コーパス (CEJC)』(小磯ほか 2017) の構築にこのテキスト自動アライメント手法が活用されている。

しかしながら、この手法があらゆる音声コーパス構築においても有効であるかはまだ明確でない。特に、音声認識システムに用いられている音響モデル・言語モデルと異なる性質の音声に対しては音声認識がうまく働かず、アライメントの精度が低下する可能性がある。例えば、上記の自動字幕作成システムは主に『日本語話し言葉コーパス』(国立国語研究所 2006) を基

* yishi@ninjal.ac.jp

表 1　対象データ

ID	地域	話者数	長さ (秒)	発話数	ID	地域	話者数	長さ (秒)	発話数
1	北海道	3	72.667	35	29	奈良	2	105.222	65
3	岩手	2	62.145	33	30	和歌山	3	82.648	59
4	宮城	3	77.027	54	31	鳥取	2	75.704	51
5	秋田	3	62.042	35	32	島根	3	95.392	53
6	山形	4	72.030	40	33	岡山	2	60.893	44
7	福島	3	68.520	23	36	徳島	3	62.405	32
8	茨城	3	60.565	32	37	香川	2	110.048	66
9	栃木	3	62.566	36	38	愛媛	2	60.692	42
12	千葉	3	62.987	54	39	高知	3	72.211	42
15	新潟	4	61.023	36	41	佐賀	3	60.082	37
16	富山	3	63.329	60	42	長崎	2	217.650	88
18	福井	3	80.031	50	43	熊本	2	64.845	35
22	静岡	4	96.930	54	44	大分	3	72.857	35
24	三重	4	66.485	37	45	宮崎	2	80.697	38
25	滋賀	3	75.174	44	46	鹿児島	2	62.904	30
26	京都	4	71.046	46	47_1	沖縄 A	2	67.384	43

に音響モデルおよび言語モデルを構築しているため、いわゆる標準語への適応が中心であり、方言音声に対しては言語モデルが適さなかったり、外国語を母語とする日本語学習者の音声には音響モデルがうまく適合しないことが考えられる。しかし、そのような音声に対してもテキスト自動アライメントシステムが効果的に活用できるのであれば、コーパス構築の負担軽減に大きく貢献できる。

そこで本稿では、現在構築作業が進められている『日本語諸方言コーパス (CJD)』(木部ほか 2017) へのテキスト自動アライメントの導入を考慮し、日本全国の様々な地域の方言音声に対するテキスト自動アライメント性能の検証を行った結果について報告する。

2. 方言音声のテキスト自動アライメント

2.1 対象データ

表 1 に本稿で取り扱う方言音声データの種類および地域ごとのデータ量を示す。これらは、『全国方言談話データベース 日本のふるさとことば集成』(国立国語研究所 2002) に収録されている音声のうち、CJD 構築のために人手で転記テキストのアライメントがすでに行われている部分である。地域によって作業の進捗状況が異なるため、それぞれ 60 秒から 200 秒程度と地域ごとに異なる長さの音声データになっている。なお、音声フォーマットはサンプリング

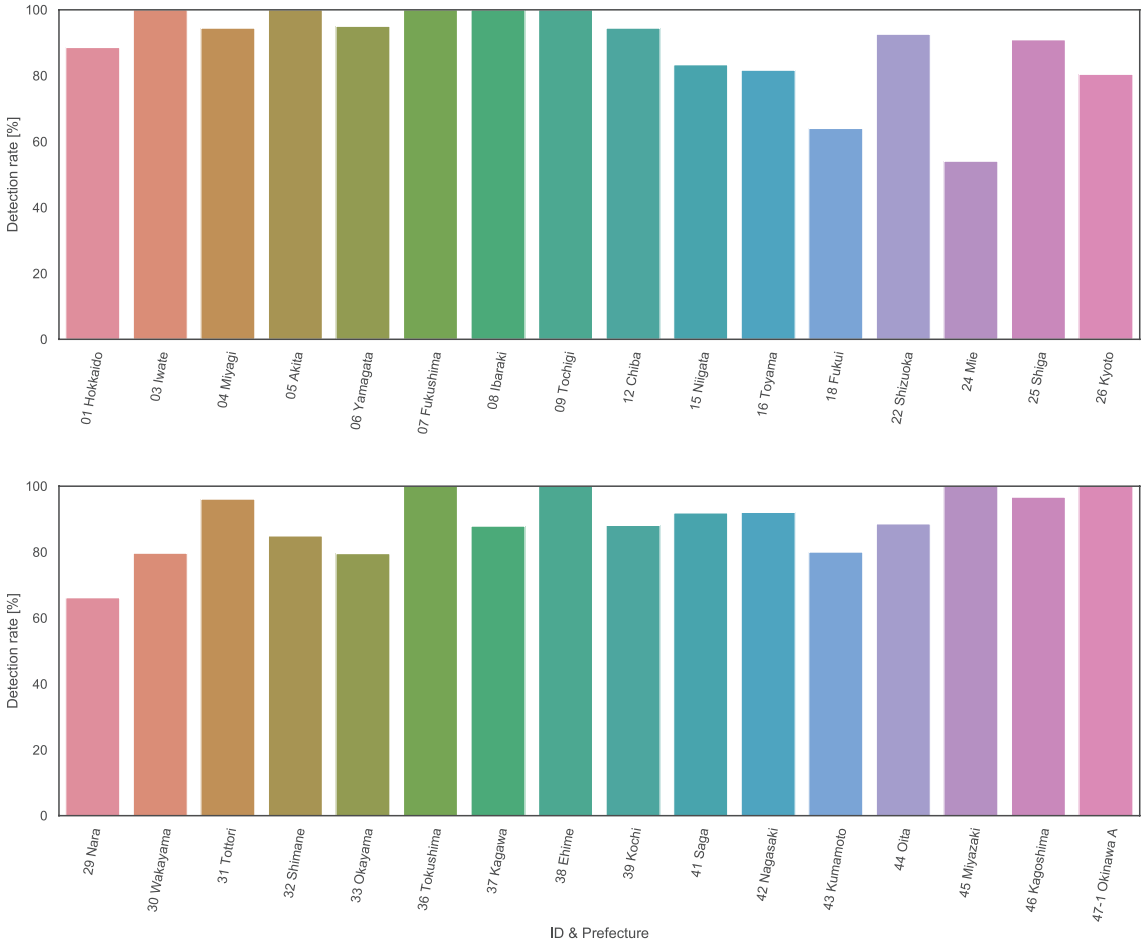


図 1　発話開始時刻推定における検出率

周波数 16 kHz、量子化ビット数 16 bit の PCM である。音声の内容は各地域の方言会話であるが、複数話者の会話が 1 チャンネルに収録されているため発話音声の重複が生じている箇所があり、この重複箇所においては音声認識性能の低下が予想される。また、転記テキストは 200 ms の無音区間を境界とする間休止単位で区切られているため、本稿では間休止単位を発話単位として用いることとする。

2.2　結果

前述の自動字幕作成システムを用いてテキスト自動アライメントの精度を調べる。なお、自動字幕作成システムは字幕用の処理を目的としているため、現時点では発話開始時刻のみの推定となっている。そのため、発話開始時刻に焦点を絞り、音声データおよび転記テキストから各発話の開始時刻を求めるアライメント処理を行った。

2.2.1　検出率

すべての入力テキストに対してアライメントが行われるわけではなく、システムが発話を検出できない場合は、発話位置が推定がされないこともある。各地域のデータに対し発話開始時刻を推定できた発話数の割合（検出率）を図 1 に示す。

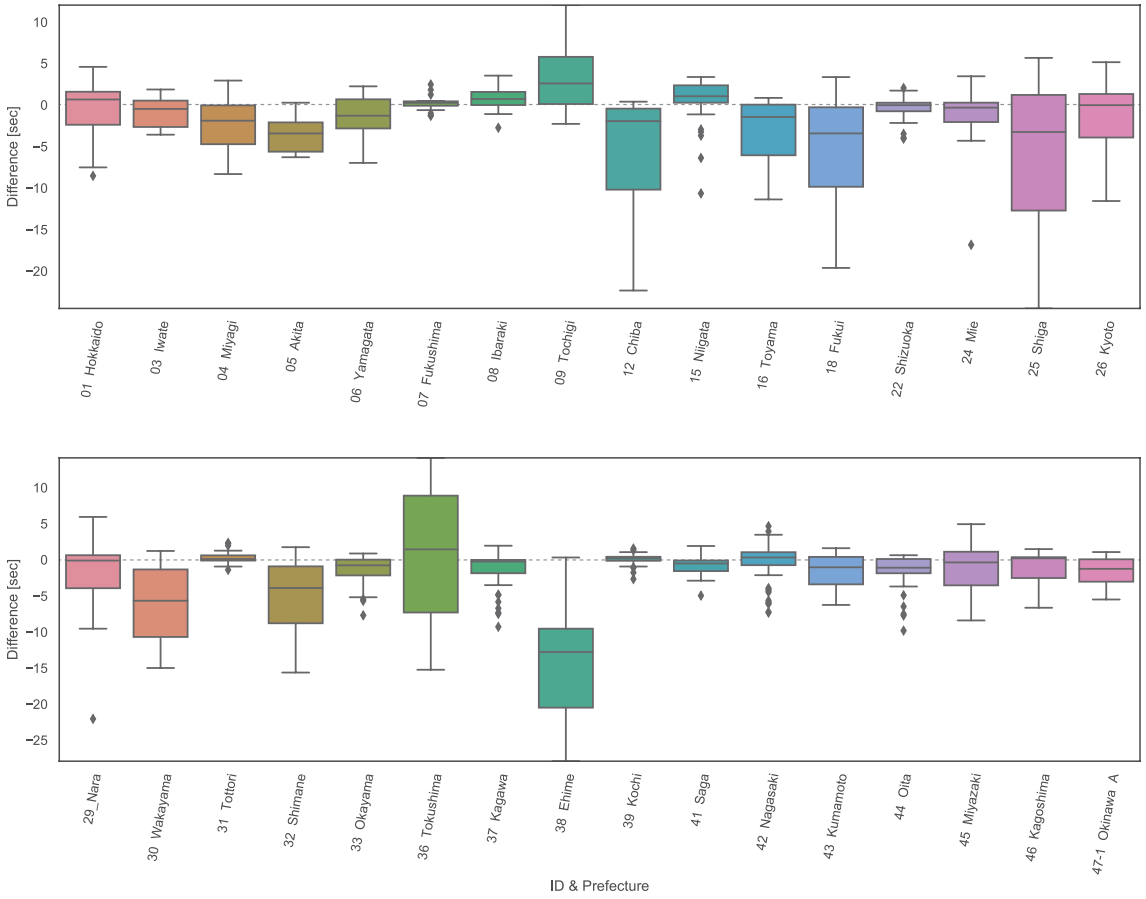


図 2　発話開始時刻の推定誤差

概ね 80% 以上の検出率を示しており、80% 以下の検出率となった地域は、福井、三重、奈良、和歌山、岡山、熊本であった。近畿地方に検出率が低い地域が比較的偏っているようにも思われるが、特に低い福井や三重に隣接する滋賀の検出率は高く、地域の位置と検出率の関係は絶対的なものではない。また、標準語の言語モデルに適合しないように思われる東北地方や九州地方については、予想に反し検出率の大幅な低下は見られなかった。

2.2.2　推定誤差

各地域における正解開始時刻と推定開始時刻との差（推定誤差）の分布を図 2 に示す。なお、人手で行った転記テキストのアライメント結果を正解時刻とみなす。値が負値である場合は推定時刻が正解時刻よりも前になっており、正値である場合は推定時刻が正解時刻よりも遅れていることを意味する。

大きく分類すると、ほとんどの発話の推定誤差が 0 に近い（すなわち正しく推定されている）地域と、一部のみ正しいものの誤りも多い地域、ほとんどの発話で正しい推定ができていない地域の 3 つに分かれている。例えば、北海道、京都、奈良、徳島などは、正しく推定されている発話もあれば大きくずれている発話もあることがわかる。また、宮城、秋田、栃木、千葉、富山、福井、和歌山、島根、愛媛といった地域は推定時刻の多くが正解時刻から大きくずれている。一方で、福島、鳥取、高知といった地域はほとんどの発話に対して正しくアライメ

ントができている。このように推定誤差に関しては、特定の地方に偏って誤差が大きくなるような傾向は見られない。反対に誤差の小ささに着目すると、九州・沖縄地方は他の地方に比べて安定して推定誤差が小さい傾向があるように思われる。

2.3 考察

検出率は低いものの推定誤差が小さい三重のデータを詳細に見ると、ある程度の大きさの暗騒音がデータ全体に入っており、音声の大きさが小さい箇所の検出に失敗している傾向にある。一方、雑音に対して相対的に音声が大きい発話に対しては概ね正確に推定できており、雑音の有無が自動アライメント性能に影響を与えていることがうかがえる。雑音のみが問題であれば、観察された暗騒音はほぼ定常音であるため、雑音除去の前処理を行うことで検出率の改善が期待できる。

また、検出率は高いが推定誤差が大きい秋田や愛媛のデータにおいては、会話の冒頭付近の大きな推定誤差がのちの発話の推定時刻に影響しているようであり、入力に用いた転記テキストの記述順と発話音声の出現順の整合性を重視しすぎた結果として全体的なズレが生じている可能性がある。すなわち、音声に転記テキストを強制的に割り当てる手続きによって本来とは異なる発話音声にテキストが誤って付与されることになり、大きな誤差が生じていると考えられる。音声データの時間が長いほどこのズレは広範囲に影響することが考えられるため、適切な長さに音声データを分割してからアライメントを行うことが必要になるかもしれない。

全体に目を向けると、地方によるアライメント性能の低下は顕著には現れていない。これは標準語との言語的な差異がアライメント性能に大きな影響を与えていないことを示しており、方言音声に対してテキスト自動アライメントが適応できないのではないかという当初の懸念を払拭するものである。もっとも、CEJC を対象にしたアライメントの性能評価では検出率が 98% で推定誤差が概ね ±1 秒程度であった (石本 2017) ことを考えると、まだ無条件でシステムを適応できるものではなく、低い検出率や大きな推定誤差が生じる原因についてさらに分析し明らかにする必要がある。また、誤差が十分に小さいとは言えないことから、後処理において作業者による修正作業は必須であろう。それでも、すべて人手で付与する労力を考慮すると、方言のような標準語とは異なる音声のコーパス構築においてもテキスト自動アライメントによるラベル付与の負担軽減の効果はあり、本システムの活用はコーパス構築の効率化を推し進める仕組みになりうると考える。

3. おわりに

本稿では、音声コーパス構築における負担の軽減を目指して、日本の様々な地域の方言音声に対する転記テキストの自動アライメント性能の検証を行なった。その結果、標準語からの言語的な違いがアライメント性能に影響を与えることはほとんどなく、方言音声に対してもシステムによるテキスト自動アライメントが実現できることが示された。システムの後処理として作業者によるラベル修正作業は必要となるものの、作業者との連携を適切に設計することにより自動アライメントシステムは方言音声のコーパス構築においても有効に活用できると考えられる。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「大規模日常会話コーパスに基づく話し言葉の多角的研究」の支援を受け、国立国語研究所共同研究プロジェクト「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」による成果を利用して行われたものである。

文　献

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

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

石本祐一 (2017).「コーパス構築における発話アライメントの現状」　言語資源活用ワークショップ 2016 発表論文集, pp. 30–37.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』の構築」　言語処理学会第 23 回年次大会発表論文集, pp. 775–778.

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.

木部暢子・佐藤久美子・中西太郎・中澤光平 (2017).「『日本語諸方言コーパス』の構築について」　言語資源活用ワークショップ 2016 発表論文集, pp. 57–68.

国立国語研究所 (2002).『全国方言談話データベース日本のふるさとことば集成』　国書刊行会　1–20 巻.

関連 URL

音声認識を用いた自動字幕作成システム

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

単語の分散表現を用いた領域における出現単語の特徴分析

佐々木 稔（茨城大学工学部情報工学科）†
古宮 嘉那子（茨城大学工学部情報工学科）
新納 浩幸（茨城大学工学部情報工学科）

Semantic Relation Analysis among Domains Using Word Embeddings

Minoru Sasaki (Ibaraki University)
Kanako Komiya (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要旨

自然言語処理では，コーパス中の単語に対して意味的な特徴をベクトルで表現し，様々な自然言語処理タスクにおいて利用することが多い．単語をベクトルで表現することにより，単語間の類似度を計算し，単語や文の違いなどの比較が可能である．これまでの研究では，単語ベクトルを生成するためには，ひとつの大規模な文書データから生成する必要があった．そのため，書籍・新聞・雑誌など，文書の分野による分散表現の比較や特徴分析は行われていなかった．すなわち，分野による単語の類似性や違いは明らかになっていない．そこで，本研究では領域ごとの単語ベクトル生成手法を提案し，各領域における単語の特徴分析を行う．書籍・雑誌・新聞の３つの分野（領域）の日本語のコーパスに対し，指定した対象単語に対して単語ベクトルを作成する．対象単語のベクトルを用いて，対象単語が類似する単語を各領域において抽出し，単語の使用傾向などの分析を行う．実験の結果，同一単語であっても他領域で使われる意味と異なった語義で使用されている単語があることが分かった．加えて，動詞は領域からの依存度が低い，副詞は領域への依存度が高いなど品詞によって領域の依存度が異なるといった傾向があった．また，書籍領域では様々な種類の語句，新聞領域では政治関連語句，雑誌領域においてはカタカナ語が多く登場するなど，類似した単語には領域によって特徴がみられた．

１．はじめに

自然言語処理で単語の特徴を分析する際，文書中の単語をベクトルで表現することがよく行われている(Schnabel et. al 2015)．単語をベクトルで表現することにより，単語の意味的な特徴を表現することや，単語間の類似度を計算することが可能となる．しかし，これまでの研究では，単語ベクトルを生成するためには，ひとつの大規模な文書データから生成する必要があった．そのため，書籍・新聞・雑誌など，文書の分野による分散表現の比較や特徴分析は行われていなかった．すなわち，分野による単語の類似性や違いは明らかになっていない．

そこで，本研究では領域ごとの単語ベクトル生成手法を提案し，各領域における単語の特徴分析を行う．書籍・雑誌・新聞の3つの分野（領域）の日本語のコーパスに対し，領域ごとに単語ベクトルを作成する．単語ベクトルはコーパスから１単語に対し1つのベクトルを与えるword2vecを用いて作成する(Mikolov et. al. 2013)．単語ベクトルを使いて対象単語の分散表現や類似度の比

† minoru.sasaki.01@vc.ibaraki.ac.jp

較を行い，文書の分野（領域）ごとにどのような特徴の違いや，類似性があるかについて分析することを目的とする．

２．分析方法

本節では，領域の違いによる語義の特徴分析手法について述べる．

２.１　各領域の文書から単語集合の抽出

各領域の文書集合に対して，形態素解析を用いて単語に分割する．単語ベクトルを作成するためには，各領域の文書集合を，単語が半角スペースを区切り文字とした単語列へ変換する必要がある．形態素解析を用いて単語列に変換する際，単語の活用形は見出し語に変形する．本稿では，形態素解析ツールとして日本語形態素解析システム「mecab[1]」を使用する．

２.２　対象単語へのタグ付け

形態素解析を行った各領域の文書集合に対し，対象単語に領域タグを付けて区別する．対象単語 w に対し，領域 1 であれば単語の末尾に領域 1 を表すタグを付けた w_1 に変換する．同様に，領域 n 中の w に対しては w_n に変換する．この手順を繰り返し，すべての領域の対象単語 w にタグ付けを行う．その後，タグ付けが行われた各領域の文書集合を 1 つのコーパスとしてまとめ，これを併合コーパスとする．

２.３　単語ベクトルの生成

併合コーパスに対して word2vec[2]の CBOW モデルを用いて，各領域について出現単語の単語ベクトルを作成する．単語ベクトルとは，ひとつの単語をひとつのベクトルで表した知識表現のことである．CBOW モデルは入力層，中間層，出力層からなり，対象単語の周辺単語を入力とし，対象単語を予測するニューラルネットワークである．入力層と出力層は辞書中の単語と一対一に対応したノードから成るベクトルを表している．中間層は指定した数のノードを隠れ変数として与えることができる．入力層と中間層の間には，それらをつなぐ重みがあり，重み行列として表現する．この重み行列を更新することによって，周辺単語から対象単語を予測するように学習を行う．word2vec は文書集合全体で学習した重み行列を用いて各出現単語の単語ベクトルを出力する．

２.４　領域の違いによる単語ベクトルの比較

対象単語の単語ベクトルを求めることができれば，意味的に関連が強い単語は単語ベクトルが近くなることを用いて，それぞれの領域で単語ベクトルの比較を行う．2 つの領域における対象単語の単語ベクトル w_1 と w_2 の類似度が高く，w_1 の類似単語と w_2 の類似単語がほとんど同じであるとき，w_1 と w_2 は 2 つの領域においてほぼ同一の意味で用いられていると推測できる．その領域や単語によって，領域の異なる同一単語の単語ベクトルにどのような特徴や違いが存在するか比較・分析を行う．

３．実験

本節では，単語ベクトルを用いて単語の意味的な用法が領域ごとにどのような違いや相違があるかについて調査するために，日本語文書集合を対象とした実験を行う．

３.１　実験データ

本研究における実験用の文書データには，国立国語研究所が開発した現代日本語書き言

[1] http://taku910.github.io/mecab/
[2] https://code.google.com/p/word2vec/

葉均衡コーパス(BCCWJ)を利用する．BCCWJ は日本語の様々なジャンルの文書を収録した，書き言葉の全体像を把握するために構築されたコーパスである．今回の実験では「新聞」「雑誌」「書籍」の 3 つのジャンルを領域として，これらの領域に含まれる文書データを使用する．

３.２　対象単語

名詞単語である「日本」「市場」「円」「政策」「口紅」「表明」「凹凸」「協力」，動詞である「出す」「見せる」「成る」，形容詞である「早い」「少ない」「可愛い」「憂い」，副詞である「きらきら」「わくわく」の全 19 単語を対象単語として，それぞれの領域で単語ベクトルの比較を行う．それにより，対象単語の意味を調査する．また，同一単語であっても領域による違いが発生しているか，調査を行う．

３.３　実験データの準備

文書集合から word2vec に入力可能な単語列への変換を行う．文書集合に対して mecab を用いて形態素解析を行って単語の基本形に分割する．単語列への変換を行った各領域（書籍・新聞・雑誌）の文書中の対象単語に，書籍領域の「日本」であれば「日本_b」というように，領域ごとにタグ付けを行う．書籍領域の場合は「単語_b」，新聞領域の場合は「単語_n」，雑誌領域の場合は「単語_m」とする．その後，単語ベクトルの類似度算出の際の精度向上のため，各領域の文書データを 10 回繰り返し，これを併合コーパスとする．

３.４　単語ベクトルの生成

2.4 節で説明した word2vec を用いて文章データから単語ベクトルを生成する．word2vec のモデルには，Countinuous Bag-of-Words（CBOW）モデルを利用する．次元数は 300 とし，文脈長は 5 である．反復回数は 20 とする．表 1 に，word2vec の訓練オプションの一覧を示す．

表 1　word2vec で指定した訓練オプション

Skip-gram or CBOW	-cbow	1
次元数	-size	300
文脈長	-window	5
負例サンプリング	-negative	5
階層化ソフトマックス	-hs	0
最低頻度閾値	-sample	1e-3
単語最低出現回数	-min-count	2
反復回数	-iter	20

４．実験結果

各領域に分けた対象単語に対して，word2vec で得られた単語ベクトルが類似する上位 10 単語を表 2 に示す．表 2 の結果を見ると，多くの対象単語において，類似する単語の上位に異なる領域の同一単語が現れた．

まず，各対象単語について類似する単語を分析する．「日本」では，どの領域においても類似単語はおおむね他国名であっが，新聞領域においては聖書の登場人物である「ヤペテ」や島の名前である「トカラ」なども現れた．「市場」では，すべての領域において共通する類似単語は「企業」しか存在しなかった．しかし，各領域の単語ベクトルは互いに類似していた．また，新聞分野では価格に関わる単語が多く類似単語として出現した．「円」では，各領域とも価格の単位として多く使われていることがわかる．また，類似単語に距離や重さなど，数値と共起する他の単位も出現した．「政策」では，別領域の同一単語がどの領域においても最も類似する結果となった．しかし，書籍領域においては主に作戦という意味で用

いられている一方で，新聞・雑誌領域では国の政治情勢に関わる単語が類似単語として多く登場した.「口紅」では，3 領域ともタグの異なる同一単語が類似単語として表れなかった.

表 2:対象単語の単語ベクトルと類似した単語ベクトル

対象単語と領域	単語ベクトルが類似する上位 10 単語
日本_b	日本_m，アメリカ，日本_n，中国，ドイツ， 託い讒ず，エトルリア，アラビア，外国，戦前
日本_n	日本_m，日本_b，中国，韓国，ヤペテ， イラク，トカラ，タイワン，ロシア，外国
日本_m	日本_n，日本_b，フランス，ヨーロッパ，アメリカ， 中国，韓国，ドイツ，写声，ロシア
市場_b	市場_n，企業，システム，競争，取り引き， 化，市場_m，需要，構造，銀行
市場_n	市場_b，市場_m，利下げ，企業，大手， 貿易，下落，じり高，軟調，買戻
市場_m	市場_n，市場_b，経済，デフレ，企業， 大手，株安，資本，グローバル，集権
円_b	円_m，円_n，ドル，トン，￥，キロリットル， グラム，ルピア，キロメートル，カトゥン
円_n	円_m，円_b，ドル，￥，トン， ルピア，カトゥン，株，キロリットル，件
円_m	円_n，円_b，￥，メートル，カトゥン， ドル，バーツ，CC，ルピア，グラム
政策_b	政策_n，政策_m，財政，施策，措置， 緊縮，路線，策，経済，戦略
政策_n	政策_b，政策_m，対中，大綱，戦略， 党内，制裁，対日，機構，構想
政策_m	政策_n，政策_b，対中，枠組み，対米， 税制，対日，国策，大綱，外交
口紅_b	リボン，ピンク，チーク，パウダー，ウール， 同系，裏地，ラメ，ブラウス，コート
口紅_n	ドリトス，衣料，子馬，特売，押麦， 茶粥，プラスチック，後払い，ステンレス，コーンビーフ
口紅_m	前著，パープル，茶系，パール，マフラー， イミテーション，メロー，ラメ，風姿，重修
ライン_b	ライン_m，スクエア，カホウ，ライン_n，シャフト， ハンドル，オン，ライン_b，ルート，エンド
ライン_n	カホウ，勝敗，プル，ライン_b，ディフェンス， スクエア，エイボン，終 盤，CK，パリーグ
ライン_m	シルエット，フレア，ウエスト，ピン，フェース， トーン，ベージュ，タイ ト，ベルト，毛先

戦争_b	戦争_n，戦争_m，侵略，内戦，日露， 戦役，終結，冷戦，占領，敗戦
戦争_n	戦争_b，戦争_m，テロ，イラク，終結， 報復，侵略，内戦，冷戦，蜂起
戦争_m	戦争_n，内戦，戦争_b，チュウトウ，湾岸， 大戦，敗戦，日露，侵略，事変
凹凸_b	紋様，飾り，タペスト リー，取々，白磁， 袋帯，兜，花弁，陶器，水差し
凹凸_n	木目，円錐，中空，新粉，金具， 突起，背面，　色調，生え揃う，太め
凹凸_m	最深，ビジター，聞き 違い，流し台， モーテル，軟式，ポトス，堅木，湾奥，インサイド
表明_b	表明_n，非難，堅持，看過，表明_m， 吐露，容認，　主張，否定，払拭
表明_n	表明_b，言明，堅持，表明_m，了承， 先送り，明言，辞任，首班，会談
表明_m	首班，打倒，宥和，アラファト，施政， 表明_n，辞任，表明_b，蔵相，陸相
協力_b	協力_n，協力_m，連携，支援，賛同， 援助，啓蒙，尽力，提言，友好
協力_n	協力_b，協力_m，連携，支援，交流， 協定，承認，協調，提言，合意
協力_m	協力_b，協力_n，無償，援助，JICA， ソーシャル，有償，アナリスト，依頼，ワーカー
出す_b	出す_m，出す_n，出る，書く，聞く， 入れる，渡す，持つ，張り上げる，流す
出す_n	出す_m，出す_b，出る，引き出す_m，減らす， 入れる，注ぐ，見せる_n，上げる，合わせる
出す_m	出す_n，出す_b，出る，渡す，見せる_m， 入れる，使う，付ける，揃える，残す
成る_b	成る_m，成る_n，有る，言う，分かる， 思う，見える，為る，出来る，考える
成る_n	成る_m，成る_b，入る，繋がる，思う， 言う，為る，出る，陥る，付く
成る_m	成る_n，成る_b，思う，見える，入る， 付く，分かる，出る，変わる，言う
見せる_b	見せる_n，見る，見せる_m，教える，感ずる， 与える，伝える，せる，広げる，引き立てる
見せる_n	見せる_m，見せる_b，見る，面変わり，語る， 決める，見せ付ける，知る，熟す，広げる
見せる_m	見せる_n，見せる_b，見る，感ずる，教える， 付ける，与える，出す_m，覚える，変える

憂い_b	捨身，慈悲，菩提，情，候， ぬ，一揆，少弐，至り，ショウスケ
憂い_n	大意，憂い_m，穢，禾，小猿， エキケン，今様，霊言，舌舐り，スサノオ
憂い_m	ビノバープリー，黄肌，シバコウエン，タド，ヤスミ， コウシカイ，福泉，順大，カサマ，蓚酸
早い_b	早い_n，早い_m，遅い，カヅキ，良い， 起床，ない，寒い，ウタコ，近付く
早い_n	早い_b，早い_m，遅い，カヅキ，暑い， 間に合う，難しい，長い，無理，正確
早い_m	早い_n，早い_b，面映ゆい，長い，遅い， 安い，カヅキ，若い，良い，強い
少ない_b	少ない_n，少ない_m，多い，低い，思い為す， 増える，殆ど，減る，多く，有る
少ない_n	少ない_b，多い，少ない_m，増える，思い為す， せめて，多く，減る，大きい，遅い
少ない_m	少ない_n，少ない_b，多い，遅い，増える， 低い，大きい，思い為す，殆ど，難しい
可愛い_b	可愛い_m，退部，怖い，優しい，優雅， 寂しい，可愛らしい，可愛い_n，嫌，素敵
可愛い_n	可愛い_m，嬉しい，退部，嫌，可愛い_b， 良い，面白い，羨ましい，河馬，幸せ
可愛い_m	キュート，可愛い_b，嫌，優しい，ラブリー， フェミニン，可愛い_n，新 鮮，良い，甘酸っぱい
きらきら_b	文才，藤色，きら，光点，ぽっちり， ヨシウラ，土埃，銀色，ピスタチオ，垂線
きらきら_n	モンテベルディ，巡洋，セルロース，ミマン，ルンメイ， クレゾール，バーデン，尋ね求める，プログレッシブ，暴雨
きらきら_m	ゴージャス，ぱちくり，ぱっちり，ぎょろぎょろ，ほっくり， 潤む，ストーン，羽織物，愛敬，毳立つ
わくわく_b	うんざり，どきどき_m，感激，どきどき_n，狼狽， ぐずぐず，浮き浮き，ときめく，嬉々，苛立つ
わくわく_n	イレン，擽る，町歩き，しゃぎり，RU， フェスタ，触れ合い，むずかる，どきどき_n，宝恵
わくわく_m	どきどき_m，ほろり，　台無し，閉口，居た堪れる， 息継ぎ，おたおた，どぎまぎ，はらはら，まごまご

書籍領域ではファッションに関わる道具として，新聞領域では商品の一つとして，雑誌領域では化粧品として用いられている．「ライン」では，書籍領域では他の領域の同一対象単語と類似度が高かったが，書籍以外の 2 領域では対象単語以外の単語と類似度が高いという結果であった．「凹凸」では，3 領域とも他の領域の同一対象単語が類似単語として全く現れなかった．また，3 領域とも類似単語が共通していなかった．「表明」では，書籍領域と新聞領域が特に類似する結果となった．また，新聞領域と雑誌領域では政治に関わる単語が

類似単語として多く表れた．「協力」は，3 領域ともほとんど同じ意味で用いられ，「援助」「連携」「支援」といった単語が複数の領域で類似単語として出現した．

「出す」は，3 領域ともほとんど同じ意味で用いられる傾向があった．「出る」「入れる」のように 3 領域共に類似単語として表れた語句も存在した．「成る」は 3 領域ともほとんど同じ意味で用いられていた．「見せる」では，3 領域ともほとんど同じ意味で用いられていただけではなく「見る」との類似度も高い結果となった．

「憂い」は，書籍領域では感情表現として，新聞領域ではネガティブなイメージを表す語として用いられていた．雑誌領域においては，「黄肌」や「順大」など，様々なバリエーションの語句が類似単語となった．「早い」は，3 領域ともほとんど同じ意味で用いられていた．また，「遅い」との類似度がどの領域においても高く，時間の単位として多く用いられている．「少ない」は 3 領域ともほとんど同じ意味で用いられていた．「多い」や「増える」など数量を表す単語が，3 領域の類似単語となった．「可愛い」では，書籍領域においては人柄を表す語句が類似単語として多く見られた．新聞領域では，感情表現を表す単語が，雑誌領域では愛らしいという意味の単語が多く見られた．その他，3 領域とも別領域の同一対象単語が類似単語として表れなかった．「きらきら」では，3 領域とも別領域の同一対象単語が類似単語として表れなかった．また，各領域ともカタカナ語の多く類似単語として出現した．「わくわく」では，3 領域とも別領域の同一対象単語が類似単語として表れなかった．また，書籍領域，雑誌領域においては「どきどき」といった畳語が類似していた．

次に，各領域の特徴を述べる．書籍領域に出現した類似単語は，様々な種類の語句が登場していた．新聞領域に出現する類似単語は，政治に関わる単語が他領域に比べ多く登場した．「憂い」「凹凸」「可愛い」など，雑誌領域の類似ベクトルにカタカナ語が他領域と比べ多く出現した．

５．考察

実験結果より，word2vec を用いた実験ではほとんどの単語は，領域の異なる同一対象単語との類似度が最も高くなるか，それに近い結果となった．しかし，全対象単語 19 単語のうち，「口紅」「凹凸」「きらきら」「わくわく」の 4 単語においては，類似する 10 件の単語ベクトルに異なる領域の同一単語は表れず，別領域の同一対象単語との類似度が高い結果となった．例えば，「口紅」では，書籍においては他の化粧品との類似度が高い一方，新聞では広告としての関連語句の類似度が，雑誌では種類や色味などの関連が強い傾向にあった．これは，書籍では数ある化粧品の種類の一つとして，新聞では商品として，雑誌では唇に塗布する化粧品の総称としてというように，各領域において異なる「もの」として扱われていることを示す．動詞である「出す」「見せる」「成る」は，どの領域においてもほぼ同じ意味で用いられていた．動詞の語義は領域への依存度が低いと考えられる．「きらきら」「わくわく」はどちらも副詞であり，この結果から副詞の語義は特に領域に依存すると推測できる．また，「きらきら」と「わくわく」はともに畳語であるが，「わくわく」の類似単語には畳語が多くあらわれたが，「きらきら」にはほとんど現れることはなかった．書籍領域の対象単語に対する類似単語は，様々な種類の語句が出現した．これは，書籍領域では他領域と比べ多様な語句が用いられているからだと考える．また，新聞領域では他領域と比べ政治的意味を持つ単語が，対象単語の類似単語として比較的多く登場した．新聞領域では社会情勢を示す記事が含まれていることが理由であると思われる．「憂い」「凹凸」「可愛い」の 3 単語においては，雑誌領域の類似ベクトルにカタカナ語が他領域と比べ多く出現した．この結果は，雑誌領域では他領域と比べカタカナ語が多く用いられていることを表していると推測できる．

６．結論

本稿では，word2vec を用いて日本語を対象とした単語ベクトルを使用し，領域の違いによる単語の特徴や類似度の分析を行った．実験の結果，同一単語であっても他領域における

意味と異なる意味で使用した単語が存在することがわかった．加えて，動詞は領域の依存度が低い，副詞は領域への依存度が高いなど，品詞によって領域の依存度が異なる傾向があった．また，書籍領域では様々な種類の語句，新聞領域では政治関連語句，雑誌領域においてはカタカナ語が多く登場するなど，類似した単語には領域によって特徴がみられた．この結果は，語義は領域に依存することを示している．

今後は，さらに多くの単語に対し，領域を増やして単語ベクトルの比較や分析を行うことが課題である．今回の実験では，3 領域中の 19 単語を用いることで，各領域における語義の分散表現の特徴を分析した．BCCWJ にある白書など領域を追加し，対象単語を増やすことで各単語の意味の分かれ方や領域による差異がよりわかりやすくなると考えられる．

文　献

Tobias Schnabel, Igor Labutov, David Mimno, Thorsten Joachims (2015). “Evaluation methods for unsupervised word embeddings”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.298-307.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. (2013) "Distributed Representations of Words and Phrases and their Compositionality", Proceedings of the 26th International Conference on Neural Information Processing Systems, pp.3111-3119.

関連 URL

『現代日本語書き言葉均衡コーパス』　http://pj.ninjal.ac.jp/corpus_center/bccwj/

形容詞感動文における曖昧性回避の条件

西内 沙恵（国立国語研究所 研究系 理論・対照研究領域）†

The Condition of Disambiguation in Adjective Exclamatory Sentence

Sae Nishiuchi (National Institute for Japanese Language and Linguistics)

要旨

感動文では，言語的な分脈情報のほか，眼前の情景も意味の推測に大きく寄与するとされる。しかし，「（子供の表情を回想して）この顔のおかしかったこと。」のように，過去の事態も表現しえることから，対象が目の前に実在することは必須要素ではない。本研究は，「青っ！＜*色が青い /**未熟だ＞」や「くさい！＜*嫌なにおいがする /**怪しい＞」のような多義的形容詞による感動文の分析から，先行研究で規定される形式的な文法枠組みに語用論的記述を加えようとするものである。多義語は，共起語や文脈情報などの言語要素が意味選択の大きな手がかりとなるとされる。言語的手がかりがない状態では，多義のいずれの意味も選択される可能性があるのにもかかわらず，眼前性の担保されない状態で出現し，意味の解釈が可能になるのはどのような条件によるのか。実例の観察から出現状況を分析し，意味の選択に語彙の特性と身体性がかかわっていることを論じる。

１．この研究の目的

感動文では，文脈などの言語的手がかりのみならず，眼前の情景なども非言語的な手がかりとして，意味の推測に寄与する（１）。しかし，笹井（2006）の文法枠組みによれば，感動文は過去の事態も表現しえ，対象が目の前に実在する必要が必ずしもない（２）。眼前の光景でなくとも，話し手の頭の中で対象が思い描かれる場合にも感動文は出現し，成立する。また，感動文は，格関係のある共起語を伴わないことがあり，このことから，言語的手がかりがない場合にも，聞き手と情景などの非言語情報の共有がない場合にも，言語表現として成立し，意味特定が可能となる。では，これらの手がかりがない場合に，いかにして意味の特定が担保されるのか。また，その曖昧性が回避される条件とは，どのようなものか。次節以降で感動文の文法枠組みを確認し，実例の出現状況の観察から意味表出の条件を見出す。

(１)（空を見て）青っ！* ＜*色が青い /**未熟だ＞　　（筆者の作例）
(２)（子供の表情を回想して）この顔のおかしかったこと。　　（森本梢子『ママ』）

２．文法枠組み

笹井（2006）は，山田文法に類別される疑問文のジャンルの中で情意を表す形式として扱われてきた「なんと～だろう！」[1]と，喚体句との対応を論じ，感動文の文法形式を整理している。笹井（2006）にまとめられる感動文の文法枠組みを，形容詞に限定して表１に示す。なお，本研究は，笹井（2006）の感動文の文法枠組みを基軸に参考するものであるが，

† snishiuchi@ninjal.ac.jp

1 「なんと」が不定語として機能していないこと，属性概念を持つ語のみならず体言も要求すること，さらに叙述文と比較すると「なんと～だろう」の形式で判断辞の機能が無効化することを根拠に，疑問文としての扱いを見直している。

一部再考が必要と思われる次に述べる二点を表１に反映している。

まず，「なんて」・「なんという」・「なんていう」を，属性概念と体言を要求する点で「なんと」のバリエーションとして扱っている点である。「なんという」・「なんていう」が表１の縦軸２「「こと/の」句的体言」や，４「語幹」の用法を持たないこと，また「なんて」だけが（３）のように５「連体形」用法を有するとして，その出現のし方が統一的でないことが論じられている。笹井（2006）の文法枠組みでは，B 型の表現形式として「なんと」が代表に扱われているが，バリエーションの中では「なんて」が最も広い用法を有する。このことから，本研究では，「なんて」を代表に表１を作成した。次に，A 型の５「連体形」が終止形と同形のため，考察の対象から外している点である。この問題について，B 型の５「連体形」を形容動詞の出現形があることを根拠に，形容詞も連体形の使用が可能なはずとも述べており，論旨が一貫していない。「B 型に形容動詞が終止形で現れることはないため，形容動詞で現れない活用形式が形容詞では現れるとは考えられない」（笹井 2006 : 24）という指摘は，形容詞と形容動詞で出現形式に差異がある限り全ての用法が対応していると断定する根拠が弱いことから，論点先取であろう。この点を考慮し，表１の B-5「なんて[形容詞-連体形]！」形式の出現の可否は，笹井（2006）のものを改変している。文法枠組みにおける活用形の扱いは，本稿では指摘するにとどめ，連体形と考えるべきか終止形と考えるべきか，あるいは品詞別に住み分けがなされていると考えるべきかは，別の機会に議論したい。なお，形容詞連体形が感動文として出現するかどうかという議論は，活用の理論的問題であり，次節で詳述する対象の観察では，終止形も連体形も同形式として統一的に扱っている。

（3）a. なんてかわいいの！　b. *なんとかわいいの！cf. なんと可憐な！　　（筆者の作例）

表１　形容詞感動文の文法枠組み（笹井 2006 を参考に）

	A 型		B 型 文頭「なんて」	
1 逆述語	[形容詞]+[名詞]！	○	なんて[形容詞]+[名詞]！	○
2 「こと/の」句的体言	[形容詞]こと！	○	なんて[形容詞]こと！	○
3 「さ」句的体言	[名詞]の[形容詞-語幹]さ！	○	なんて[形容詞-語幹]さ！	○
4 語幹	[形容詞-語幹]！	○	なんて[形容詞-語幹]！	×
5 連体形	[形容詞-連体形]！	※	なんて[形容詞-連体形]！	※

※笹井（2006）は，形容詞の連体形と終止形が同形のため，考察の対象外としている。

以上，文法的な扱いを見てきた。ここで，感動文の定義を確認しておく。感動文は，ただ話し手の感動が表出される表現形式というだけでなく，「文に示されている情報についての伝達を目的とはしない，話し手の感動を表出する文」（笹井 2006 : 16-17）であると考えられている。笹井（2006）は，感動文の構造の解明に焦点を当てている。本研究では，この基礎的な文法枠組みに，意味の表れ方という語用論的記述を加えることを試みたい。

３．調査

３.１　調査の目的と方法

感動文の語用を探るために，現代語の感動文の文法枠組み（表１）に準じ，実例を調査する。使用実態の調査には，コーパスアプリケーション『中納言』を用いて，『日本語話し言

葉コーパス』（以下，CSJ）・『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）・『名大会話コーパス』（以下，名大）から抽出された実例を観察する。

このとき，感動文の意味表出を分析する観点として，多義の表出の問題に着眼する。この観点から，調査対象の形式が次のように限定される。意味を把握するためには，用法の側面を無視することができないことが国広（1982）で述べられている。多義語が用いられるとき，その語が持つ複数の意味のうちどの意味で用いられているのか，という意味の特定には，格関係のある共起語などが大きな手がかりになる。感動文は，文法的に体言骨子とされ，体言が明示される場合には，被修飾名詞によって形容詞の任意の意味が特定されることが想定される。しかし，表１の４「語幹」と５「連体形」の用法では，形式的に被修飾名詞が明示されない。これらの用法の調査が感動文の意味特定の問題を追求するのに理想的であると考え，本調査では，４「語幹」と５「連体形」を調査対象とする。

さらに，多義語に着眼することで，次のことが期待される。多義の立ち現れ方を分析することは，感動文の運用において言語情報と非言語情報，それぞれの果たす役割，また語そのものに担われる役割を一部解明することにつながる。体言骨子として文法的に被修飾名詞を抱いていても，実際の言語運用では，感動文に被修飾名詞が伴われないこともある。では，そのような運用についてその場ごとに解釈されるものと考える ad hoc 概念構築（Carston 2002）が効いているのだろうか。感動文が「伝達を目的としない」と定義されることから，他者に向けて発せられているものではないことになり，ad hoc 概念構築による解釈は適切ではないということになる。本研究では，解釈過程を追う手順によらず，感動文で曖昧性が回避される条件を整理することを試みる。この整理の必要性は，（４）のように，ある多義語の意味がその場限りの文脈がなくとも曖昧にならないことにも確認される。（４）の使用で立ち現れる意味は，＜嫌なにおいがする＞であり，＜怪しい＞という意味を想定しにくいものと思われる。刑事が一人で思案，独り言つ場面を想定すれば，＜怪しい＞が選択されるかもしれないが，限定的である。このように，被修飾名詞や文脈がない場合でも任意の意味が想定され，意味の混同は起きない。被修飾名詞などの周辺の要素に意味の選択を担わせる多義的形容詞であっても，言語的・非言語的な手がかりなしにある程度意味が特定されるのはなぜだろうか。

（４）くさい！* ＜*嫌なにおいがする /**怪しい＞　　（筆者の作例）

３.２　調査結果

調査の結果，B 型の４「語幹」用法及び５「連体形」用法は抽出されなかった[2]。しかし，B 型の感動文は，書き言葉で多く観察される（笹井 2006）という特性から，後続する文脈で説明的にパラフレーズされていたり，文脈に意味を特定可能にする言語的手がかりがあったりと，意味を曖昧にしておかない工夫がなされている可能性が高いと考えられる。この工夫により，被修飾名詞が明示されず，言語的手がかりがない場合に，いかにして意味の特定が担保されるのかという本研究の疑問の解明にはあまり参考にならない用例が得られることが考えられる。このことから，B 型の用例が得られなくとも，研究の遂行には問題がないと考え，調査を進めた。

節単位・発話単位・文頭と文末に挟まれた形容詞の語幹・終止形・連体形の分布を，形容詞の意味属性別に表２に示す。意味属性の分類は，多くの言語でほかの品詞より形容詞で表されることが多いとされる７分類＜次元＞・＜物体特性＞・＜色彩＞・＜人物特性＞・＜年齢＞・＜価値＞・＜速度＞（Dixon 1977）に則っている。また，日本語に特徴的な用法を考慮し，＜価値＞を＜価値（i）＞と＜価値（ii）＞に類別した。＜価値（i）＞は，「すごい」や意味の向上（amelioration of meaning）を果たした「やばい」など，程度の甚だしさを表すものを分類する。＜価値（ii）＞には，ポジティブかネガティブかいずれかの意味に偏

[2] B 型に４「語幹」用法は存在しないが，念のため検索し，運用がないことを確認した。

るものを含めた。さらに，Dixon（1977）の７分類で＜物体特性＞の下位に分類される「ない」や「多い」などを＜存在＞として類別した。加えて，日本語文法に用法が特異な感情形容詞も，Dixon（1977）では＜人物特性＞に含まれるが，本研究では＜感情＞に区分した。

集計の方針は次の通りである。節単位・発話単位に挟まれた形容詞の語幹・終止形・連体形であっても，不要の意を表す「いい」のような応答のほか，対話者の発話を繰り返すものは除いている。また，副詞や「わ」以外の終助詞を伴うものも，強調の明示や伝達方式の形式化といった機能を多様に果たし，「文に示されている情報についての伝達を目的とはしない，話し手の感動を表出する文」という感動文の定義から一部外れるため，除いている。ぞんざいな「めっちゃ」や終助詞「わ」など，感動文の生産に相性のいい要素も一部あるが，感動文との呼応をまとめる整理が別途必要なものと考え，本調査では扱わない。ほか，形容詞反復発話（大江 2017）は，感動文の一種として数えた。ただし，「めでたしめでたし」のように，形式的には感動文と一致する使用でも，慣習的な意味用法を有するとみなせる場合には，感動文として数えていない。なお，BCCWJ 及び名大の抽出結果に関しては，明らかに感動文として用いられたことを期すために，文末ないし発話単位末に補助記号を含むものを算出した。

表２　意味属性別に見る述べ語数と異なり語数

用法／意味属性	CSJ				BCCWJ				名大			
	語幹		終止・連体		語幹		終止・連体		語幹		終止・連体	
	token	type	token	type	token	type	token	type	token	type	token	type
価値(i)	1	1	12	3	0	0	182	2	0	0	159	3
価値(ii)	0	0	6	3	0	0	224	13	0	0	41	9
感情	1	1	12	5	6	4	449	39	4	4	146	21
存在	0	0	5	1	0	0	51	5	0	0	65	4
人物特性	0	0	3	2	0	0	256	46	0	0	95	10
物体特性	0	0	4	3	0	0	407	58	0	0	104	20
次元	0	0	3	3	0	0	41	15	0	0	35	12
色彩	0	0	0	0	0	0	4	3	0	0	2	1
年齢	0	0	0	0	0	0	2	1	0	0	2	2
速度	0	0	0	0	0	0	8	1	0	0	2	1

CSJ・BCCWJ・名大で一定数抽出された終止形・連体形感動文の述べ語数と異なり語数から，語彙密度指数として C 値（= $\mathrm{Log_e Type} / \mathrm{Log_e Token}$）を求め，意味属性間で比較したものが図１である。

以下で意味属性別に用例とともに感動文の運用傾向を見ていく。なお，「#」は，節単位末であることを表す。また，形容詞感動文が用いられている箇所に下線を引いておく。

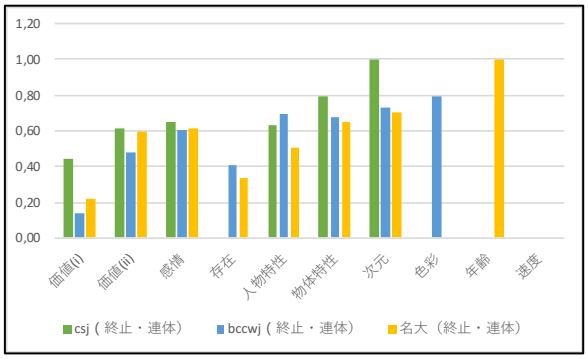


図１　各属性の C 値

傾向１：程度の甚だしさが表される場合（＜価値（i）＞）

形容詞が感動文として用いられるとき，（５）及び（６）のように，「すごい」や「やばい」など，＜価値（i）＞に当たる，程度の甚だしさを表す表現が多く見られた。この場合，その度合いの振り幅が問題となり，ポジティブであったりネガティブであったり，対象のどのような属性を表すかという意味は，解釈の問題とならない。図１に見るように，語彙密度は比較的低く，述べ語数の割に異なり語数が少ないことがわかる。

(5) 全然変わってくるんでしょうしね#うーん#ええ#やっぱりそうなんですか#うーん#えー#凄い#あたし今鳥肌立ちました#
【出典】 CSJ 講演 ID：D01F0030-L（音声タイプ：対話・模擬）
(6) ということに影響してるのではないかという風に考えます#やばい#それからえーと原文の姿をどにぐどのぐらい保持しているかという#
【出典】 CSJ 講演 ID：A03F0072（音声タイプ：独話・学会）

傾向２：＜感情＞が表される場合

（7）から（10）のような＜感情＞に属する形容詞の使用も，頻度としては多かった。ただし，異なり語数も少なくなく，語彙密度は低くない。これらの語彙も，感動文の性質から話し手の感情を表していることが明らかであり，曖昧性の問題が生じない。

(7) ああ、恥ずかし#ああ、恥ずかし#恥ずかしいことが果たしてまだ残っているのかどうか、もう私に 【出典】BCCWJ サンプル ID：LBn9_00134
(8) ちっちゃいエビとかしか扱ったことがないもんね。#そうやなー、そうか。#しんど#生きてんのはちょっとなー。#大変やなー。#うん#か、勘弁してっていう感じ
【出典】 名大 会話 ID：data100
(9) なんかベビーカーに乗ってて、すごい汗かいとったんだね。#うん#怖#うん#うん#昨日とか。#もう今連れ回してるから。#それがなんか夢に出て
【出典】 名大 会話 ID：data116
(10) あの人はさー、見た目重視で今言ったの？#じゃなくて？#おもしろ#見た目重視じゃなくて、全体的に。#全体的に#うん。#すごいおもしろいか
【出典】 名大 会話 ID：data066

傾向３：＜存在＞・＜価値（ii）＞が表される場合

（11）及び（12）は，＜存在＞と＜価値（ii）＞を表す形容詞の感動文である。これらは，語彙密度は低めであった。例のほか，「多い」・「乏しい」・「難しい」・「よろしい」など，いずれも意味論的に意味の分類がなされる多義語だが，存在するかしないか，また価値の高さないし低さを表すものである。これらは，文脈的変容と捉えられることもあるほど多義が接近しており，傾向１と同様に，聞き手は被修飾名詞などの手がかりがなくとも，感動文の意味が特定できる。

(11) 残念。#気を取り直して、中身を探索すると・・・#むむむむむ・・・#無い、なぁ〜んにも、好きなのが入って無いって、言うより、なんじゃこりゃ〜
【出典】 BCCWJ サンプル ID：OY01_02739
(12) #おーーーーーっ！#良いよ！！#麺を啜る。#おーーーっ！#良い！#魚介系出汁が濃厚です・・・醤油ダレの風味が無くなってますね。
【出典】 BCCWJ サンプル ID：OY03_08809

傾向４：典型的な多義的形容詞の場合

（13）及び（14）は，多義語らしい多義的形容詞「甘い」が別義で用いられている例である。(13) は＜糖度が高い＞ことを，(14) は＜考えが足りていない＞ことを意味している。

(13) いないではないか！# 漬物石をどかし、ちょっとなめてみてビックリ。#甘い！# 伯母が一生懸命探しあてたのはグラニュー糖だったのだ。#大慌てで水洗い
【出典】 BCCWJ サンプル ID：LBp5_00027
(14) #そういうわけでまたまたコスモスの写真十点。#これで終わったと思うな！#甘い！#まだまだ続くぞ、コスモス街道！ 【出典】BCCWJ サンプル ID：OY11_03468

傾向３までで，感動文の使用には，そもそも意味の特定が必要ないような程度の大きさや感情を表すものが多く，語彙の特性が大きく偏っていることがわかった。傾向４に見たような典型的な多義語の使用における意味特定について，次節で考察を進めたい。

４．形容詞感動文に見る意味特定の手がかり

本節では，感動文で立ち現れる，ある多義語の意味が曖昧にならない仕組みを考察する。（15）の例について，＜嫌なにおいがする＞という意味が選択され，＜怪しい＞という意味の選択が限定的になることを前節で言及した。また，「甘い」の多義の運用を（13）と（14）に見た。これらの意味表出の条件が身体性と発話の連鎖に見てとれることを次に述べる。

(15) くさい！* ＜*嫌なにおいがする /**怪しい＞　　((4)を再掲)

定延（2002）で，認知者と環境とのインタラクションに生まれる身体性が，文法性にかかわることが論じられている。能動的な認知において，環境と認知者は相互に働きかける関係にあり，環境に認知者が身を置いたり活動したりすると，なんらかの強烈な情報が環境から認知者に返ってくる。この考えにおいて，刺激は，身体性の低いものから高いものまでグラデーションをなしており，認知者が環境に積極的に働きかけなくても，強烈な情報を体感として受け取ることができる。定延（2002）で定義される「身体性」とは，推論や高度な判断を必要とせず，否応なしに感じられやすいことの程度を指す。例えば，（16）と（17）に見るように「痛い」は「赤い」より，「気持ちいい」は「長い」より身体性が高い。これらは，アリサマを表すとき制約となる例である。

(16) a. 私は３回ぐらい痛かったよ。　　b. ?私は３回ぐらい赤かったよ。
(17) a. [マッサージ機の[特強]ボタンを指さして、すでに試した者が]
　　これ押したら、気持ちいいよ～。
　b. [新型テレビの[ワイド]ボタンを指さして、すでに試した者が]
　　??これ押したら、画面が横に長いよ～。（定延 2002：173-174　下線は筆者による）

この概念を参考に，(15)「くさい」の意味選択に立ち戻ると，＜嫌なにおいがする＞ことが＜怪しい＞より身体性が高いことがわかる。＜怪しい＞は，何らかの事実に基づいて推論した結果の判断であるから，身体性は低いといえよう。このことを発話の連鎖の枠組み（筒井 2012）に組み込んだものが表３である。独話で知識・経験が共有されず情報が話し手に完結する場合と，対話で知識・経験が共有される場合では，推論や高度な判断が難しくない。この枠組みに照らして意味の表出を見ると，「くさい」の＜怪しい＞という意味の選択が限定的になることは，予測可能である。

表３　発話の連鎖に照らした感動文の意味表出（筒井 2012 を参考に）

発話の連鎖	知識・経験	情報伝達の方向性と発話タイプ	感動文の意味表出
対話	非共有	要求　→　質問	身体性の高い意味
		提供　→　報告	
	共有	要求・提供　→　共有	身体性の低い意味（推論・高度な判断）
独話	非共有	方向なし　→　独り言	

このように，被修飾名詞が明示されなくとも，身体性の概念と発話の連鎖との掛け合わせから多義語の任意の意味が特定され，意味の混同は起きない。「甘い」の場合，（18a）では，身体性の高い＜糖度が高い＞という意味選択が非共有知識の報告としてなされている。対

して，（18b）では，身体性の低い＜考えが足りない＞という意味選択が，コスモスが満喫される状況が共有された上で示されている。

(18) a. いないではないか！# 漬物石をどかし、ちょっとなめてみてビックリ。#甘い！#伯母が一生懸命探しあてたのはグラニュー糖だったのだ。#大慌てで水洗い
b.#そういうわけでまたまたコスモスの写真十点。#これで終わったと思うな！#甘い！#まだまだ続くぞ、コスモス街道！ （（13）及び（14）を再掲）

５．おわりに

聞き手と情景などの非言語情報が共有されない過去の事態や，被修飾語などの言語情報が示されない語幹用法および終止形・連体形用法でも，感動文は文法的に成立する。しかし，実際の語用では，程度の甚だしさを表す用例や感情形容詞の使用が多く，使用傾向に語彙特性の偏りが見られた。また，典型的な多義語が用いられる場合でも，身体性のかかわりから曖昧性が回避されることを論じた。本研究では，話し手が環境からのインタラクションとしての刺激をどのような動機付けで，感動文という表現形式に表すのかという根本的な問題には立ち入れなかったが，意味の立ち現れ方の条件を提起した。また，感動文の定義である「伝達を目的としない」ことがその生成にどう関わるのか，という聞こえと話し手の聞こえへの意識についてまでは，考察を進められなかった。さらに，曖昧性を回避する条件を提案したが，実験などを通した検証は行っていない。これらを今後の課題に，研究を進めたい。

謝　辞

本稿は，言語資源活用ワークショップ 2018 で発表した内容に加筆修正をしたものです。発表に際し，臼田泰如氏，大江元貴氏，陳祥氏，西川賢哉氏，また多くの方々から重要なご指摘と有益な助言，貴重な情報提供を賜りました。ここに改めましてお礼を申しあげます。なお，いうまでもなく，本論文の不備や誤りはすべて筆者の責任です。

文　献

Carston, R. (2002) *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.
Dixon, R.M.W. (1977) Where have all adjectives gone?, *Studies of Language*. 1, 19-80.
今野弘章 (2012)「イ落ち構文：形と意味のインターフェイスの観点から」『言語研究』141, 5-31.
国広哲弥 (1982)『意味論の方法』東京：大修館書店.
大江元貴 (2017)「「形容詞反復発話」の文法－「怖い怖い。」は「怖い。」と何が違うか－」『日本語学会 2017 年度秋季大会予稿集』
定延利之 (2002)「「インタラクションの文法」に向けて」－現代日本語の擬似エビデンシャル－」『京都大学言語学研究』21, 147-185.
笹井香 (2006)「現代語の感動文の構造－「なんと」型感動文の構造をめぐって－」『日本語 の研究』2(1), 18-34.
清水泰行 (2015)「現代語の形容詞語幹型感動文の構造－「区的体言」の構造と「小節」の構造との対立を中心として－」『言語研究』148, 123-141.
筒井佐代 (2012)『雑談の構造分析』東京：くろしお出版.
八亀裕美 (2003)「形容詞の評価的な意味と形容詞分類」『阪大日本語研究』15, 13-40.

関　連　Ｕ　Ｒ　Ｌ

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

ノンネイティブ日本語教師はコーパスで どのように日本語を調べるか –コーパスを用いた課題の分析から–

清水　まさ子（国際交流基金日本語国際センター）†
木田　真理　（国際交流基金日本語国際センター）

`How Do Non-Native Japanese Teachers Use Corpus? An Analysis through the Task They Did'

Masako Shimizu (The Japan Foundation Japanese-Language Institute, Urawa)
Mari Kida (The Japan Foundation Japanese-Language Institute, Urawa)

要旨

筆者らは将来的に,ノンネイティブ日本語教師(以下,NNT とする)がコーパスを用いて日本語を学んでいく授業を開発しようと考えている。本発表は,このような授業の開発にあたり,実際に NNT がコーパスを用いて興味のある語を調べた際に,彼らがどんな発見をし,どんな所でつまずくのか調べる,いわば授業開発のためのパイロットスタディである。

本調査の結果,NNT はコーパスを使用することによって,対象語の新たな用法やレジスター,またはどんな時に使用するのかについての発見をしていた。しかし,そもそも検索対象語を正しく認識していない場合や,検索結果の解釈の誤りなど,NNT ならではの問題も確認された。

本発表ではこのような調査の結果と考察をもとに,NNT がコーパスを使用する際に踏まえるべき点を示した。

１．はじめに

日本語教育においてコーパスは,教師,学習者両者にとって非常に有効なものであることは言うまでもない。教師や学習者の日々の授業や学習を助けるツールとしてその使用は数多く報告されているが,実際に学習者がコーパスを動かしながら学んでいく,いわゆるデータ駆動型の授業実践は,その有効性が報告されているもかかわらず,日本語教育の分野ではまだ少ない。

発表者らは将来的に,ノンネイティブ日本語教師研修において,研修参加者自らがコーパスを用いて,興味ある語を調べ,考え,分析するという,自律的及び探索的な授業を開発しようとしている。

本発表では,開発途上にある授業実践において,実際に NNT がコーパスを用いて自律的及び探索的に興味ある語を調べた際に,彼らがどんな発見をし,どんな所でつまずくのか調べる,いわば授業開発のためのパイロットスタディである。

２．先行研究

コーパスを用いた授業実践例については,日本語教育以外に国語教育,英語教育等でも報告されている。国語教育では例えば,青空文庫と全文検索システム「ひまわり」を使い,「よう(ようだ・ような・ように,など)」を検索し,結果を分類し表などにまとめて発表する,といった実践が報告されている（鈴木 2015）。また英語教育でも,初級や上級の英語授業におい

†Masako_Shimizu（アットマーク）jpf.go.jp

て,日英パラレルコーパスを用いて学習者自らが学ぶ実践が報告がされている(中條ほか 2006,西垣ほか 2010)。

日本語教育においてコーパスは,教師は例文作りの参照にしたり(中俣 2017),シラバスデザインに役立てたり(庵ほか 2015)と,言うまでもなく幅広く活用されているが,先述したような,実際に学習者がコーパスを動かしながら学んでいくような実践は,まだ少ない。

ただ日本語教育の分野でも,その有効性と可能性は報告されている。寺嶋(2011)は,大学で学ぶ上級学習者を対象に,漢字語彙指導の際に,コーパスと漢字学習支援ツールを用いて,ターゲットとする漢字語について,コロケーション情報やコロケーションを用いた例文を書き出すという実践を報告している。そしてこのようなコーパスを学習者自らが使用する学習は,「オーセンティックな言語に大量に触れること」ができ,「学習者の疑問を解決するための新たなストラテジーにもなる」(p101)と述べている。

このような学習の可能性は,NNT を対象にする研修にとっても有効であると考える。木田・山本(2018)では,海外で日本語を教える NNT は,日本語について疑問が生じても周囲に疑問を解消する人的,物的リソースが少ない場合も多く,自律的に教授上の問題点を解消していかなければならない。そのため,日本語を自ら分析していく力が必要であると述べた。NNT に対するコーパス使用の可能性は,砂川(2012)でも述べられており,母語話者並みの直感を持たない NNT は,コーパスを使用することによって,その不足分を補えるとしている。

３．本発表の目的

先述したように学習者自らがコーパスを使用していく授業実践報告では,その有効性や可能性が示唆されていた。海外で教える NNT に対しての研修でも,このような実践は有効であると考えられるが,今まで NNT に対してこのような実践を行った報告は,管見の限りない。

そこで本発表は,NNT がコーパスを自ら使用して学んでいくようになるには,どのような授業を行えばよいのかを検討するために,まずは簡単な紹介のみでコーパスを用いて調べたいことを自由に検索してみる授業を行う。そして,その授業で提出した課題において,NNT が実際にコーパスを使用して学習していく際に,どのような語を調べ,どのような過程を辿って何を発見し,さらには,どんなことでつまずくのか,といったことを探る。

４．コーパスを用いた探索活動の概要

以下に,１）本発表で調査対象とした NNT の概要,２）調査対象者・調査期間,３）コーパスを用いた授業の内容,４）今回調査対象とする NNT から提出された課題の内容,について述べる。

４.１　NNT の概要

文化庁(2016)によると,国内の日本語学習者数は 217,881 人,日本語教師は 37,962 人であるのに対し,国際交流基金(2015)によると,海外における日本語学習者数は 3,665,024 人,日本語教師は 64,108 人である。海外における日本語学習者と教師数は国内に比べて非常に多く,特筆すべきは,海外における日本語教師は 70%以上が NNT である点である。

NNT の特徴としてはいくつか挙げられるが,その中でも木田(2004)は, NNT は教師であると同時に学習者であるとし,文法の明示的知識や文法指導といった「教師の文法」と,文法学習や文法知識,文法知識の活用といった「学習者の文法」とが混在しているとしている。

４.２　調査対象者・調査期間

調査対象としたのは,国内にある NNT のための研修施設において,若手日本語教師を対象としている「海外日本語教師長期研修」（以下,長期研修）の参加者である。最上位または,上位から 2 番目のクラスで,対象人数は 43 名。調査期間は 2015 年から 2017 年度にかけてである。各年度の研修参加者の平均的な日本語運用力は,日本語能力試験Ｎ2 以上（ACTFL OPI では,上級以上,CEFR 共通参照レベルではＢ1 以上に相当する場合が多い）であり,その国籍は以下である。

インドネシア 2,カンボジア 2,タイ 4,ベトナム 3,モンゴル 2,ハンガリー1,メキシコ 1,フィリピン 2,ミャンマー7,インド 3,スリランカ 1,米国 1,アルゼンチン 2,ブラジル 1,ウクライナ 1,ウズベキスタン 1,ジョージア 1,ブルガリア 3,ロシア 3,トルコ 1,イラン 1　(21 ヶ国)

４.３　コーパスを用いた授業の内容

コーパスを用いた授業は,2015-16 年度においては,日本語科目にあたる「文法演習」の中のテーマの一つとして,2017 年度においては,教授法の「文字・語彙指導」の中で行われた。1 回 3 時間の授業内容は,以下である。なお,日本語コーパスというものの存在について事前に知っていた参加者は,43 名中 2～3 名とわずかであり,大多数は,初めての利用者であった。

・コーパスとは何か(コーパスの定義)
・コーパスの種類（書き言葉コーパス,話し言葉コーパス,学習者コーパス）
・コーパス開発の歴史（アメリカのブラウンコーパス,イギリスの BNC など）
・BCCWJ「少納言」の概要
　少納言の説明画面（http://www.kotonoha.gr.jp/shonagon/）から,「検索にあたっての注意点」「利用法」「検索対象となっているサンプル」「サンプルの長さ」「サンプルの選択基準」などの情報を読み取って理解の確認をする。
・NINJAL-LWP FOR BCCWJ（以下, NLB とする）の概要を,「NLB とは」「使い方」などの説明画面　（http://nlb.ninjal.ac.jp/）から情報を読み取って理解の確認をする。
・言葉の使用実態について調べる以下の練習をクラス全体で行って結果を確認する。
（例：メガネとめがね,表すと表わす,玉子　卵　たまご　タマゴ　や卵焼き　玉子焼きなどの表記の違い）

４.４　今回調査対象とする NNT から提出された課題の内容

上記の授業の後に,以下の課題を出し,数週間後に回収し,これらを今回の調査対象とした。

使用した方を●で示してください：　少納言　　／　　NLB
検索対象語 ＿＿＿＿＿＿
検索方法・問題点：

わかったこと・興味深いと思ったこと：

今回, NNT がどのような語を分析対象として選び,どのような過程を経て何を発見するのか,またどのような点でつまずくのか見るために,以上のような分析自由度の大きい課題にした。

また,研修の全体カリキュラム上,コーパスの使用に関する十分な説明と練習の時間をとれない状況であること,形態素の理解などが伴わない日本語力であること等の理由で,今回はNNT でも利用しやすいと考えられる少納言と NLB に限った。

５．結果

本発表では,4.4 で提出された課題の中から,１）調査動機,２）調査された語,３）調査課程の分析を行う。

５.１　調査動機

語の調査欄に,調査動機を書いた NNT もいた。以下,その動機をいくつかに分けた。大きく分けて,「母語に日本語に該当する意味がないため」「教育上で疑問に思って」「どのような場面,言葉と一緒に使われるのか確認したかったため」「その他」に分類できた。以下に,その例を記す。以下,NNT の例は原文のままである。

【１．母語に日本語に該当する意味がないため】

(1) 「なかなか」はインドネシア語に訳したら,適当な言葉はありませんから,もっと知りたいです。話し手によって,いい意味もあるし,あまりよくない意味もありますから,使いにくいと思います。

(2) 「なんとなく」はいつも聞いている言葉ですが,正しい使い方を知りたいです。インドネシア語に訳したら,適当な言葉はありませんから,もっと知りたいです。

【２．教育上疑問に思って】

(3) この「あいだに」と「あいだは」の違いは学生さんによく聞かれて,自分でもよく迷った文型ですから,探そうと思いました。

(4) 初級の教科書では「動詞辞書形＋つもりです」という形のみで導入されていますが,「○○つもりで」,「○○つもりがない」,「＜動詞過去形＞＋つもり」等の組み合わせも多く使われているように思い,コーパスでそれぞれの使用頻度を調べてみました。

【３．どのような場面,言葉と一緒に使われるのか確認したかったため】

(5) 最近習った四字熟語を NLB で検索し,どのような場面,どんな言葉と一緒に使われていると確認したかったです。

(6) いつもどちらかよく分かりませんから,探してみました。

【４．その他】

(7) あるとき,仕事で英日の通訳をしていたときに英語の「dilemma」を訳せなくて困ったことがありましたが,調べたらそのままカタカナ言葉になっていました。それでいつごろから日本語に入ったか気になって調べました。

(8) お正月だから,カルタについて考えました。BCCWJ の少納言をつかって,「かるた」と「カルタ」を探しました。どの書き方がいちばん多いか知りたかったです。

(9) ニューヨークでは,ハイカラという日本酒バーによく通っているので,言葉として「ハイカラ」が気に入りました。

大きくわけた４つの動機を見ると,NNT ならではの使用動機であるものと,反対に日本語母語話者教師（以下,NT）であっても調査する動機として挙げられるものがあることに気付く。例えば NNT ならではの調査動機としは,「1.母語に日本語に該当する意味がないため」が挙げられる。また,NT,NNT に限らない調査動機としては,「2. 教育上疑問に思って」「3. どのような場面,言葉と一緒に使われるのか確認したかったため」がある。

また,「その他」の調査動機としては,NNT 自身の生活／仕事の中で気になった言葉を挙げている。コーパスを研究目的のものではなく,「ちょっと気になった言葉を調べたいから」といった,ある意味「辞書がわり」のように考えた NNT もいることを示唆している。

５.２　調査された語

NNT が調査対象として選んだ語は,一語のみを調査対象としたものと,複数の語を調査対象とし,用法や頻度を比較しながら調査したものがあった。ここでは,前者を単一調査,後者を複数調査と呼ぶことにする。

単一調査では,以下のような語が選ばれていた。

表１　NNT の単一調査で選ばれた語

表現の種類	実際の調査語
副詞的表現	全然(3),なかなか(2),なんとなく(2),とても(2),よく(2), 一応, まったく,しきりに,せいぜい,いっぺんに,どうせ,ゴロゴロ, ばっちり,あまり,恐る恐る,せめて,ごく,あっけらかん, 大変
名詞的表現	ストラテジー,おおみごころ,めぐみ,連中,恋,縁,ジレンマ, 食いしん坊,坊っちゃん,いずれ,つもり,無邪気,半信半疑
形容詞的表現	やばい,すげえ
連語	に伴って,いわんばかり,たかが,ながらに,ならでは
動詞	食べれる、行く
形状詞	みたい,大丈夫
連体詞	小さな、たっての
その他	無記入など

最も多かったのは,副詞的表現であった。その次に,名詞的表現,形容詞的表現と続いた。

次に,複数調査で選ばれた語を述べる。複数調査は,大きく分けて,「表記の違いを比較したもの」と,「類義語の比較をしたもの」に分けられる。以下に,これらをまとめた表を載せる。

表２　どのような語の表記の仕方を比較しているか

・ぶどう／ブドウ　　・さようなら／さよなら　　・かるた／カルタ ・恐る恐る／おそるおそる　・無線 LAN／無線ラン　　・充分／十分 ・タバコ／煙草／たばこ　　・かっこういい／カッコウいい／格好いい

表３　どのような類義語を比較しているか

どんな比較か	実際の語
名詞的表現間の比較	・レベル／段階／級　・ミルク／牛乳　・つくえ／テーブル ・ストラテジー／方略　・あけおめ／新年　・恋／愛 ・キチン／台所　　・ハイカラ／スマート ・わたし／あたし　・国際／インターナショナル ・普通／通常／普段

動詞の比較	・食べられる／食べれる　・目を閉じる／目をつぶる ・増やす／増える　・食べる／食う　・買う／買い物する ・後悔する／後で後悔する
副詞的表現間の比較	・ふらふら／ぶらぶら　・是非／絶対　・少し／ちょっと ・恐らく／たぶん
連語的表現間の比較	・あまりに／あまりにも　・あいだに／あいだは ・～となる／～になる　・ひょっとして,ひょっとすると,ひょっとしたら／もしかして,もしかすると,もしかしたら
品詞の異なる語を比較	・食べる／食事する／食事　・好きだ／愛している ・要る／必要　・易しい／簡単　・大きい／大きな
その他	・使役受身形とその短縮形 ・使役を使った許可を求める表現　・無記入など

表２,表３を合わせてみると,複数の語を比較する場合,名詞的表現間の語同士が選ばれることが多くなっている。また,同じ品詞同士を比較するのではなく,異なる品詞だが,意味的には類似しているものを比較している場合も見られた。

５.３　調査過程の分析

ここでは,NNT が語を調査する際に,どのようなことを発見し,どのようなことにつまずいているのかについて述べる。

５.３.１　新しい語の用法を発見したケース

NNT が既に知っている語の用法以外の新しい用法を発見するケースが見られた。

(10) 全く+ない形だけではなくて,全く+動詞　全く+形容詞　全く+副詞といろいろ一緒に使う事がわかりました。いろいろ探したところ,全くそのまま,全く知る,全く痛いなどのような使い方もたくさんあって,私にとっては,初めてです。

(11) 「大丈夫」の語を探してみました。日常生活で使うときに,ただ「良い」という意味だけではなくて,「要らない」「問題がない」などの意味もあると気付いきました。それに,「大丈夫かもしれません」の形も少納言にも出ました。この使い方を初めて知りました。

（＊傍線部分,筆者ら加筆）

また,耳にしたことのある語だが,そのレジスターがわかってきた例,また,多様なコロケーションの例から,自分なりに語の使用方法がわかってきた例が見られた。

(12) 「国際」を検索した所 20395 件だったに比べ「インターナショナル」は 585 件あった。最近はインターナショナルという言葉をよく耳にするが実際は国際の使用が多いことがここからわかる。ジャンル別に見て見れば「国際」は法律,産業,世界史,経済,政治,新聞,国会会議録,教科書などに多く載ってあり,「インターナショナル」は主に雑誌,ブログなどの数が多くあった。テレビ番組,コマーシャルでよくインターナショナルという言葉がでて耳に慣れて多く使われている感じがしているが実際は「国際」の言葉の方が重要な場面で多く使われておりまだまだ「強い」と言うことが分かった。

(13) せめては強い意志を表すときに使うとわかってきた。普通,「ほしい」,「動詞＋たい」

と一緒に使われているから。意志だけでなく,反事実的条件文と一緒に出た。具体的に言うと,「ほしかった」,「動詞＋たかった」,「～ばいいのに」と一緒に使われた。

以上のように,教科書や辞書だけでは得られない,語の新しい用法を発見する例となっていることが確認された。

次に問題となるケースの 3 点をあげる。

５.３.２　コーパス利用前に問題があるケース

コーパスを利用する前に,調べようと思っている語自体が誤りの場合が見られた。

(14) 昔は台所,最近はキチンのほうが使われていて,今の率はどうでしょうか。

(15) 「ラブストーリー偶然に」という歌の歌詞に「君があまりきれいだから」という言葉が出てきてちょっと面白いなと思った。普通はあまり＋否定形と教えている。

上記２例とも,正しくは,「キッチン」「あまりに」であるにも関わらず,「キチン」「あまり」をそのままコーパスで調べている。調査対象語の記憶が曖昧な場合,そのまま調べてしまうと,当初調べようと思った言葉の意味とは異なる意味を調べてしまい,それを検索結果だと勘違いしてしまうので注意が必要である。

また調査目的と選択されたコーパスのミスマッチの例も見られた。

(16) NLB で「あけおめ」と「新年」について探しました。どうつかうか確認したかったです。

(17) 「食いしん坊」という表現について,興味を持ったので調べてみたら,84 件しか出てこなかった。（中略）この言葉はサブコーパスごとの書字形分布では書籍や知恵袋広や報紙やブログなどでの使用率が高い,国会会議録や教科書や白書などフォーマルな場面では全然使われていないようだ。（著者注：使用コーパス：NLB）

NLB はそもそも「名詞や動詞などの内容語の共起関係や文法的振る舞いを網羅的に表示できる」（http://nlb.ninjal.ac.jp/）ものであるため,上記のような調査に向いているかどうか疑わしい。

以上の誤りの例から,コーパスを用いた授業においては,事前に検索対象語が正しいかどうか,また,その語を調べるにあたって,どのコーパス使用が適当かどうかを確認する必要があると言える。

５.３.３　検索結果の解釈を誤っているケース

調査対象語が正しく入力できても,その後,検索結果の見方が誤っているため,分析を誤るケースが見られた。

(18) 「たかが」という言葉は多くの場合名詞と一緒に使う（「たかが風邪」「たかが愛」「たかが送料」など）。動詞と一緒に使うこともあるが,少納言や NLB の例文を見てみると,その使い方は「知れる」という動詞に限られているようだ。女性より男性のほうが使う言葉だ。

(19) 調べてみると,「食う」を使う男性が多いと分かった。そして「食う」は現代語だと思ったけど,ずっと前から使われていた。そして,例文を読んだら「食う」は動物などの動作によく使われると気がつきました。

上記２例とも,その調査対象語の使用は男性が多いとしている。しかし,そこまでの情報はコーパスの例文ではわからない。なぜこのような判断をしたのか課題の提出物だけでは分からないが,作者の性別のみを見ている可能性や,会話文の男女の違いを正しく認識していない可能性もある。

また,次のように「タイトル」欄を本のタイトルではなく,言葉の説明だと誤って思っている NNT もいた。

(20) （省略）「恐れ入ります」の説明に「社会人なんだからこれだけは知らねば」,「気持ちがきちんと伝わる話し方」など面白く書かれていた。

少納言の場合は,検索画面は「執筆者」「生年代」「メディア／ジャンル」「性別」「タイトル」などの項目数が多く,NNT にとっては情報を読み取るのが困難である。例文の読み取りだけではなく,今回のように,検索結果の項目自体を誤って認識している場合があることがわかった。

５.３.４　レジスターに関する認識における問題

その語が「話し言葉」か「書き言葉」か,またはフォーマルで使用されるかインフォーマルか等,レジスターに関する認識に問題がある場合が見られた。

(21) 食べれるはまだ話し言葉だとわかった。それから,結果の中で,202 件はブログから来たが,10 件は書籍から来た。その書籍の中で,「食べれる」は登場人物が言ったセリフにしか使われなかった。一番古いのは 1988 年に出版された書籍だった。1 件は報告から来た。この報告中でも人のいった言葉に「食べれる」が出た。

(22) （省略）ウェブでは「テーブル」のほうが多く使われている。たぶん,インフォーマルまたはカジュアルの場面で使いやすいと思う。

前者は,「食べれる」の検索結果中 202 件がブログからの例であり,また書籍の例も「登場人物が言ったセリフ」の中に使用されており,そのため「食べれるはまだ話し言葉」だとわかった,と解釈している。書き言葉均衡コーパスである少納言を使って,話し言葉か書き言葉かを判断することには,限界があるが,話し言葉的かどうかの傾向をつかむことはできる。少納言の中では, ブログは, 話し言葉に近いものであり, 書籍の登場人物のセリフも話し言葉と考えることができる。しかし, 話し言葉だと断定するには, 問題があると言えよう。

また後者の例では,分析者の前提として「ウェブで使用されている語＝インフォーマルまたはカジュアルの場面で使われている」と解釈している点が見られる。

このように,コーパスを使用して分析する際には,語のレジスターに関する知識も必要となってくることが示唆された。

６．考察

今回, NNT のコーパス使用時の一端を見てきたが, NNT に限らず NT がコーパスを使用する際にも見られる事象もあるだろう。例えば,先述したとおり,NT であっても教育上疑問に思ったことを調べたり,ある語の使用場面、共起する言葉などの確認をしたい場合にコーパスを用いたりする場合がある。またレジスターの知識や調査対象語と使用コーパスとのミスマッチも,NT が使用する際にも問題となると考えられる。

逆に,NNT ならではのコーパス使用時の特徴としては,例えば「使用動機」としては,「母語に日本語に該当する意味がない」時にコーパスを利用する場合がある。またコーパスで検索した結果を見る際に,その情報量の多さ故に検索項目を誤って認識したまま分析してしまうケースも見られた。またそもそも検索対象語を正しく認識していない場合もあった。しかし,コーパスを使用することによって,対象語の新たな用法やレジスター,またはどんな時に使用するかといったことがわかったという報告も見られた。これはコーパスでの調査が,砂川(2012)でいうところの「母語話者的直感を補っている」ケースと言える。

このようにNT,NNT 両者に関わる部分もあるが,今回調査した結果,NNT がコーパスを使用する際には,次のような点を踏まえると良いと考えられる。

１）自分が調査したい語の正しい形を知っているか。
２）調査の動機・内容とコーパスの特徴が合致しているかどうか。
３）検索結果の項目を誤って認識させないような項目の見方の基本知識。
４）話し言葉/書き言葉,フォーマル,インフォーマルといった語のレジスターに関する考え方の紹介。
５）文字列検索であるコーパス（少納言）や NLB の特徴を踏まえて,どこまで分析ができるか。

７．まとめ

今回, NNT がコーパスを使いこなせるようになるには,どのような授業を行えばよいのかを検討するために,その基礎資料として,NNT が実際にコーパスを使用して学習していく際の過程を調査動機,調査された言葉,調査過程の分析の３点から見てきた。

コーパスは研究だけではなく,教育にも非常に有効であると考えられてはきたが,日本語教育においては,あまりその報告はなかった。本発表の意義としては,NNT がコーパスを使用する際に何を発見したか,また使用した際の問題点の例をあげ,NNT がコーパスを使用する際の授業に対して,その際に注意するポイントを提示できたことである。また,NNT は,教師であると同時に日本語学習者でもある。従って本発表から得られた示唆は,日本語学習者へも援用できると考えている。

今後は,今回得られた成果をもとに,NNT がコーパスを使って,学習者として自分の日本語力を高めるとともに,教師としても授業準備に役立てられるような授業を開発していくつもりである。

謝　辞

本研究の一部は，JSPS 科研費 JP17K02800 の助成を受けたものです。

文　献

庵功雄・山内博之(編）(2015)『データに基づく文法シラバス』くろしお出版.

木田真理(2004)「外国人日本語教師研修における文法授業のあり方–文法シラバス整備に向けて–」『日本語国際センター紀要』14,pp51-68.
木田真理・山本実佳（2018）「日本語の分析能力を養う中・上級文法授業の試みー外国人日本語教師研修における実践ー」『国際交流基金日本語教育紀要』14,pp35-49.
国際交流基金（2015）『2015 年度　海外日本語教育機関調査』
https://www.jpf.go.jp/j/project/japanese/survey/result/survey15.html(2018 年 7 月 27 日確認）
鈴木一史(2015)「語彙に着目した学習指導」田中牧郎(編)『コーパスと国語教育』,pp36-50.
砂川有里子(2012）「日本語教育へのコーパスの活用に向けて」『日本語教育』150,pp4-18.
寺嶋弘道(2011)「日本語教育におけるコーパスの応用-データ駆動型学習とその実践方法の考察-」『ポリグロシア』20, pp91-103.
中條清美・西垣知佳子・内山将夫・山﨑淳史(2006)「初級英語学習者を対象としたコーパス利用学習の試み」『日本大学生産工学部研究報告 B』39,pp29-50.
中俣尚己(2017)『コーパスから始まる例文作り』くろしお出版.
西垣知佳子・中條清美・木島綾子(2010) 「パラレルコーパスを利用した英語上級者用データ駆動型英語学習の実践の試み」『千葉大学教育学部研究紀要』58,pp279-286.
文化庁（2016）『平成 28 年度国内の日本語教育の概要』
http://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/nihongokyoiku_jittai/h28/(2018 年 7 月 27 日確認）

関連 URL

コーパス検索アプリケーション『少納言』　https://chunagon.ninjal.ac.jp/
『NINJAL-LWP for BCCWJ(NLB)』　http://nlb.ninjal.ac.jp/

『日本語話し言葉コーパス（CSJ）』模擬講演における節頭フィラーの特徴

渡辺 美知子（国立国語研究所）†
是松 優作（東京大学）

Features of Clause-Initial Filled Pauses in Simulated Public Speaking of "The Corpus of Spontaneous Japanese (CSJ)"

Michiko Watanabe (National Institute for Japanese Language and Linguistics)
Yusaku Korematsu (The University of Tokyo)

要旨

文境界，節境界にはフィラーがよく観察される。これは，文や節などの談話の切れ目では，その先の発話内容や表現を考えるのに時間が必要なためと考えられる。本研究では，文頭，節頭でのフィラーの使用に影響する要因を，節頭フィラー全体ならびに頻度の高い「エー」「アノ」「マー」について CSJ コア中の模擬講演を対象に調べた。考察対象とした要因は，①話者の性別，②年齢，③学歴，④講演経験，⑤直前の境界の種類，⑥節中語数，⑦節頭の接続詞の有無，の 7 つである。節頭フィラーの出現確率との間に有意な関連が見られたのは要因①，⑤，⑥のみであった。フィラーの出現確率は男性話者の方が女性話者よりも高く，強い節境界の方が文境界よりも高かった。頻度の高い 3 種類のフィラーを個別に見ると，エーのみ，文境界の方が強い節境界よりも出現確率が高かった。エーは，アノ，マーよりも，深い談話境界で用いられる傾向のあることが明らかになった。

１．はじめに

フィラー，繰り返し，音節の引き伸ばし，言い誤りなどの言い淀みは，自発発話に頻繁に観察される。このような現象は朗読音声には稀なことから，その場で発話内容を考えながら次々と生成していく自発発話に固有の現象と考えられる。文を書く場合は，自分のペースで遂行しながら書くことができる。表現が適切でない場合は消して修正することができる。けれども話し言葉では，一旦発せられた音声は消すことができないし，先を考えるための間が長すぎると，話し相手や聞き手に不審に思われる可能性がある。フィラー，繰り返し，引き伸ばしなどの言い淀みは，不自然に長いポーズを避け，表現や内容を模索中であることを相手に伝えるための一方策と考えることができる。本研究では，これらの言い淀みの中で頻度の高いフィラーの働きに着目する。

認知的，言語的な制約から，メッセージはなんらかのまとまりを持った単位として生成される必要がある。広く受け入れられている発話生成モデルでは，生成プロセスに，発話内容の生成，言語形式の選択，調音という３つのレベルを想定している（Levelt 1989）。文や節や句などの統語的単位は，あるまとまりを持ったメッセージの単位としてもとらえることができる。内容が大きくシフトする談話境界や文境界，節境界ではフィラーの頻度の高いことが報告されている（Swerts 1998, Shriberg 1994）。これは，談話の切れ目や文頭，節頭では，次に何をどのように話すかという，ある程度長いスパンの発話内容の生成が行われるためではないかと考えられる。もしそうであれば，次に伝えようとするメッセージ

† watanabem@ninjal.ac.jp

の内容が豊富なほど，統語的に複雑になり，表現に必要な語彙も増えるため，生成に時間がかかり，フィラーが出現する確率も上がると考えられる。渡辺・外山（2017）では，後続節中の語数が多いほど，節頭のフィラーの出現確率が上がるという仮説（複雑さ仮説）を『日本語話し言葉コーパス（CSJ）』(2016)を用いて検証し，仮説を支持する結果が得られた。また，内容的に切れ目の深い境界ほど，次に何をどう話すかの自由度が高いため，後続発話開始に時間がかかり，フィラーの出現確率が上がるという仮説（境界仮説）を，文境界，深い節境界，浅い節境界の 3 種類の境界を対象に検証した。フィラーの出現確率は，文境界 ＜ 弱境界 ＜ 強境界の順になり，弱境界と強境界の間では支持されたが，文境界と節境界間では支持されなかった。渡辺・外山では英語コーパスとの比較のために，20 代～30 代前半の，大卒以上の学歴を持つ男女各 10 名のみを調査対象とした。そこで，本研究では調査対象を広げ，より多様な話者を対象に，節境界におけるフィラーの使用に影響する要因の分析を行った。フィラーの使用に影響する要因として分析対象としたのは以下の 7 要因である。

① 話者の性別:先行研究では男性話者の方がフィラーの使用率は高い（渡辺・外山 2017）。
② 講演時（1999 年～2002 年）の年齢：年齢がフィラーの使用頻度に影響する可能性があると考えた。
③ 学歴：学歴がフィラーの使用頻度に影響する可能性があると考えた。
④ 講演経験：フィラーはリハーサルによって減少することが報告されている（Goldman-Eisler 1968）。講演経験の違いがフィラーの使用に影響する可能性があると考えた。
⑤ 直前の境界の種類：境界が深いほど，フィラーの使用確率は上がると予測した。
⑥ 節中語数：節中の語数が多いほど，節頭でのフィラーの使用確率は上がると予測した。
⑦ 節頭の接続詞の有無：「で，エー」のような，接続詞とフィラーの共起は珍しくない。接続詞の使用とフィラー使用との間に何らかの関係がある可能性を考えた。

次に，フィラーの中でも高頻度の，「エー」，「アノ」，「マー」の使用に影響する要因の分析を同様に行い，それぞれに固有の特徴や働きがあるかどうかについて考察した。

２．方法

２.１　分析資料

『日本語話し言葉コーパス（CSJ）』(2016) コア中の模擬講演 107 講演を分析対象とした。

２.２　手続き

1 で述べた 7 要因に以下のような水準を設定した。

① 話者の性別：講演者の男女の内訳は，女性 54 講演，男性 53 講演である。
② 講演時（1999 年～2002 年）の年齢： 20 代から 60 代前半まで 5 歳区切りで 9 グループ。人数の内訳は表 1 の通りである。
③ 学歴：中学または高校卒，大学学部卒，修士以上の 3 グループ。人数の内訳は表 2 の通りである。
④ 講演経験：今回が初めて，1~5 回，6 回～10 回，11 回～20 回，21 回以上の 5 グループ。人数の内訳は表 3 の通りである。
⑤ 直前の境界の種類：直前が文境界（絶対境界），強い節境界（強境界），弱い節境界（弱境界）の 3 カテゴリー。境界タイプの分類は CSJ の分類に寄った（国立国語研究所 2006:

p.270)。

⑥ 節中語数：フィラーや語断片は節中語数には含まれていない。また，節頭の接続詞も，2 つの節を繋ぐものと考え，節中語数には含めなかった。

⑦ 節頭の接続詞の有無：節頭に接続詞があるかないかの 2 カテゴリー。

そして，これらの要因を説明変数とし，節頭でフィラーが用いられる確率を目的変数として，ロジスティック回帰分析混合モデルによって推定した。フィラーの使用頻度や用い方には個人差が大きいため, 話者を変量効果として扱った。分析には, R version 3. 5. 0 上で, lme4 パッケージ内の glmer と lmerTest を用いた。

表 1 話者の年代内訳

年齢	人数
20to24	15
25to29	16
30to34	21
35to39	15
40to44	20
50to54	4
55to59	6
60to64	3
65to69	7
計	107

表 2 話者の学歴内訳

学歴	人数
中学卒	1
高校卒	37
学部卒	59
修士以上	10
計	107

表 3 話者の講演経験内訳

講演経験	人数
初めて	66
1～5回	23
6～10回	4
11回～20回	1
21回以上	6
不明	7
計	107

３．結果

3.1　節頭フィラー全体

結果を表 4 に，フィラーの出現確率推定値をプロットした回帰曲線を図 1 に示す。R 言語では，予測変数が因子の場合，変数名を文字順にソートして先頭に来る変数を基準変数とし, その回帰係数値（estimate）を 0 とおくことが慣習となっている。性別要因の「女性」，学歴要因の「中学・高校卒」，境界の種類要因の「弱境界」節頭接続詞要因の「なし」の回帰係数値が 0 となっているのはこのためである。オッズ比はその変数の効果の大きさを表し，値が 1 から離れているほど効果は大きい。7 要因のうち，有意な効果が見られたのは，性別，直前の境界の種類，節中語数の 3 要因だった。即ち，フィラーの使用確率は，女性話者よりも男性話者の方が高く，弱い節境界よりも文境界，文境界よりも強い節境界で高かった。また，節中語数の増加に伴ってフィラーの使用確率も上昇していた。有意な効果のあった要因は, 同じコーパスの 20 代～30 代前半の話者 20 名を対象として調べた，渡辺・外山（2017）の結果と一致している。ただし，境界に関しては，渡辺・外山では，文境界の出現確率が最も低かったのに対し，本研究では弱境界が最も低かった。一方，話者の年齢，学歴，講演経験，節頭接続詞の有無の影響は観察されなかった。

3.2　フィラーの種類別分析

前節と同様の分析を，頻度の高いフィラータイプ「エー」,「アノ」,「マー」について個別に行った。

エー

全ての変数を投入するとモデルが収束しないため，予備的分析から，影響が小さいと思われる，年齢と講演経験の 2 要因を除き，残り 5 要因を説明変数として分析した。分析結

表 4 節頭におけるフィラーの出現確率を推定するロジスティクス回帰分析結果

Variable	Estimate	Std. Error	z value	Pr(>|z|)		オッズ比
(Intercept)	-1.790	0.302	-5.931	3.01e-09	***	0.167
性別						
女性	0					
男性	0.560	0.173	3.238	0.001	**	1.750
年齢	0.004	0.006	0.617	0.538		1.004
学歴						
中学・高校卒	0					
大学学部卒	0.114	0.186	0.613	0.540		1.121
修士以上	0.853	0.513	1.662	0.097	.	2.347
講演経験	-0.019	0.015	-1.215	0.225		0.982
境界の種類						
弱境界	0					
強境界	0.526	0.044	11.881	< 2e-16	***	1.692
文境界	0.464	0.049	9.553	< 2e-16	***	1.590
節中語数	0.012	0.002	4.768	0.000	***	1.012
節頭接続詞						
なし	0					
あり	-0.084	0.053	-1.587	0.113		0.919

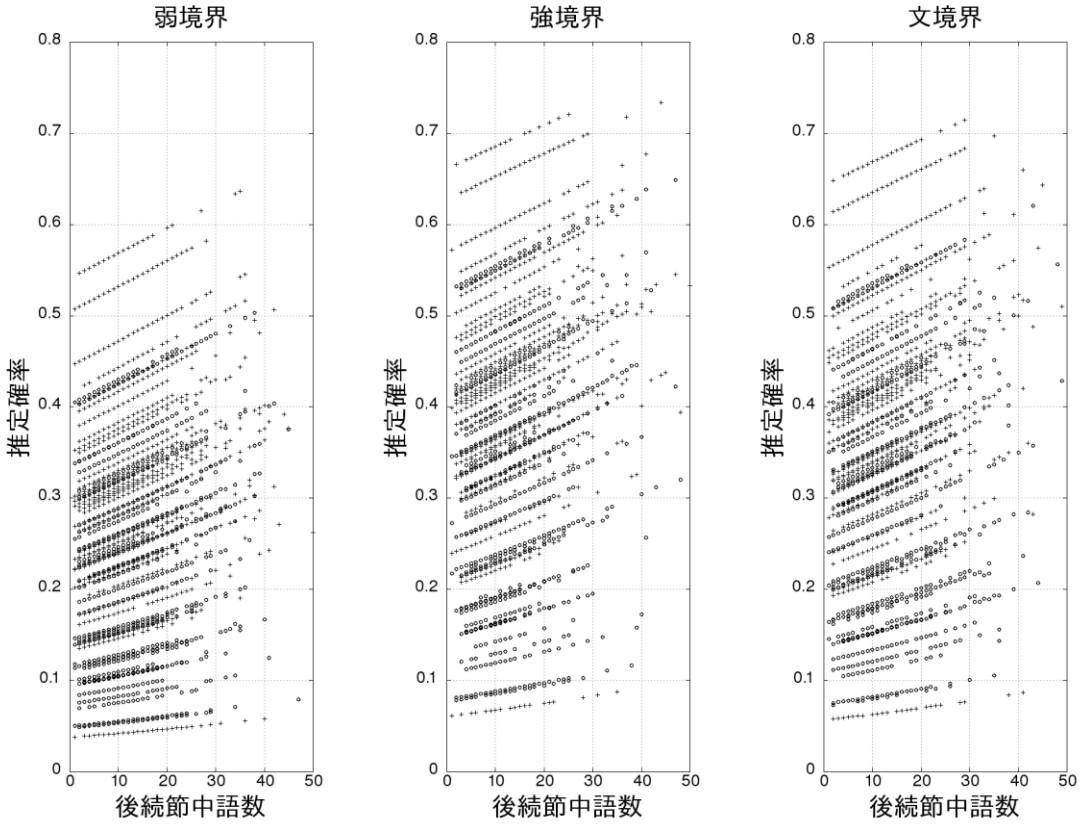


図 1 節頭におけるフィラーの出現確率推定値を示す回帰曲線（境界タイプ別）

果を表 5 に示す。分析対象とした 5 要因全ての効果が有意だった。フィラー全体の傾向同様，エーの使用確率は，女性話者よりも男性話者の方が高かった。節中語数の増加に伴いエーの使用確率が上昇する点もフィラー全体の傾向と一致していた。直前の境界の種類の影響も有意だったが，効果の大きさの順序がフィラー全体の傾向とは異なっていた。効果の大きさは，フィラー全体では，弱境界＜文境界＜強境界 であったのに対し，エーに関しては，弱境界＜強境界＜文境界 であった。つまり，エーの出現確率は境界の深さに対応しており，深い境界ほど高かった。エーは他のフィラーよりも深い談話境界で用いられる傾向のあることが示唆された。

表 4 節頭におけるフィラーの出現確率を推定するロジスティクス回帰分析結果

学歴と節頭接続詞の有無の効果はフィラー全体では現れなかったが，エーの使用確率に関しては有意だった。即ち，大卒の話者の方が中学または高校卒の話者よりも節頭のエーの使用確率は高かった。また，節頭接続詞の使用との間に関係があり，接続詞があるとエーの使用確率は下がる傾向があった。この結果は，「で」などの節頭接続詞とエーは相補的な役割を果たしている可能性のあることを示唆している。即ち，接続詞がフィラー同様，時間稼ぎの働きをしている可能性がある。

表 5 節頭におけるエーの出現確率を推定するロジスティクス回帰分析結果

Variable	Estimate	Std. Error	z value	Pr(>|z|)	オッズ比
(Intercept)	-4.224	0.332	-12.72	< 2e-16 ***	0.015
性別					
女性	0				
男性	1.073	0.327	3.279	0.001 **	2.923
学歴					
中学・高校卒	0				
大学学部卒	0.866	0.359	2.413	0.016 *	2.378
修士以上	0.692	0.639	1.084	0.278	1.998
境界の種類					
弱境界	0				
強境界	0.265	0.070	3.806	0.000 ***	1.304
文境界	0.357	0.071	5.017	0.000 ***	1.429
節中語数	0.013	0.004	3.526	0.000 ***	1.013
節頭接続詞					
なし	0				
あり	-0.207	0.084	-2.465	0.014 *	0.813

アノ

全ての変数を投入するとモデルが収束しないため，予備的分析から，影響が小さいと思われる，学歴と節頭接続詞の有無の 2 要因を除き，残り 5 要因を説明変数として分析した。分析結果を表 6 に示す。フィラー全体で見られた性別の効果が，節頭のアノに関しては有意ではなかった。フィラー全体で見ると，男性話者の方がその使用頻度は高い。しかし，ことアノに関しては，女性話者も男性話者に劣らず頻繁に使用していることになる。一方，フィラー全体では観察されなかった，話者の年齢と講演経験の効果が有意だった。年齢が上がるほど，節頭のアノの使用確率も上昇する傾向があった。一方，講演経験が増えると，節頭でのアノの使用確率は低下した。アノは，話者特性の影響を受けやすいフィラーと考えられる。境界の種類の効果はフィラー全体の傾向と一致しており，弱境界＜文境界＜強境界 の順であった。節中語数の効果もフィラー全体の傾向同様有意で，語数が多いほどアノの出現確率は上昇した。

表 6 節頭におけるアノの出現確率を推定するロジスティクス回帰分析結果

Variable	Estimate	Std. Error	z value	Pr(>|z|)	オッズ比
(Intercept)	-3.995	0.527	-7.574	0.000 ***	0.018
性別					
女性	0				
男性	-0.372	0.343	-1.086	0.277	0.689
年齢	0.024	0.012	2.033	0.042 *	1.025
講演経験	-0.078	0.032	-2.435	0.015 *	0.925
境界の種類					
弱境界	0				
強境界	0.355	0.072	4.93	0.000 ***	1.426
文境界	0.179	0.077	2.336	0.020 *	1.196
節中語数	0.011	0.004	2.462	0.014 *	1.011

マー

7 要因を説明変数とした，節頭のマーのロジスティック回帰分析結果を表 7 に示す。最も顕著な特徴は，フィラー全体で観察された節中語数の効果がマーでは観察されなかったことである。語数が，概念を言語化するプロセスの認知的負荷を反映するものと考えると，マーの使用は言語化の負荷とは別の要因に起因していると考えられ，他のフィラーとは働きの異なることが推測される。

一方，フィラー全体では観察されなかった節頭接続詞の効果が有意だった。エーでも接続詞の効果は有意だったが，エーでは接続詞があると使用確率は低下するのに対し，マーでは上昇した。

表 7 節頭におけるマーの出現確率を推定するロジスティクス回帰分析結果

Variable	Estimate	Std. Error	z value	Pr(>|z|)	オッズ比
(Intercept)	-3.856	0.483	-7.983	0.000 ***	0.021
性別					
女性	0				
男性	1.511	0.274	5.515	0.000 ***	4.532
年齢	0.004	0.010	0.411	0.681	1.004
学歴					
中学・高校卒	0				
大学学部卒	-0.143	0.293	-0.489	0.625	0.867
修士以上	-0.044	0.784	-0.056	0.955	0.957
講演経験	-0.026	0.024	-1.11	0.267	0.974
境界の種類					
弱境界	0				
強境界	0.567	0.066	8.645	< 2e-16 ***	1.763
文境界	0.458	0.074	6.223	0.000 ***	1.581
節中語数	-0.003	0.004	-0.849	0.396	0.997
節頭接続詞					
なし	0				
あり	0.154	0.074	2.092	0.036 *	1.167

他の効果はフィラー全体の傾向と一致していた。性別に関しては，女性話者よりも男性話者の方が使用確率は有意に高かった。オッズ比から，性別の効果は他のフィラーに比べて大きいことがわかる。マーは男性に好まれるフィラーと言えそうである。境界の種類の効果もフィラー全体の傾向と同じで，弱境界＜文境界＜強境界 の順であった。話者の年齢，学歴，講演経験の効果は有意ではなかった。

４．考察

①話者の性別，②年齢，③学歴，④講演経験，⑤直前の境界の種類，⑥節中語数，⑦節頭の接続詞の有無，の 7 要因を説明変数として，節頭フィラーの出現確率をロジスティック回帰分析混合モデルによって推定した。これら 7 要因のうち，①～④は話者属性に関するもので，⑤～⑦は産出された言語の特性に関するものである。話者属性に関しては，フィラー全体では性別の効果のみが有意だった。フィラーを種類別に見ても，アノ以外は男性話者の使用確率が高かった。CSJ と類似した英語コーパスを用いた研究では，節頭フィラーの使用確率に有意な男女差はなかった（Watanabe & Korematsu 2017）。日本語における男女差の原因が生得的なものなのか，文化的なものなのかの追及は今後の課題である。話者の年齢の効果はアノ以外なかった。しかし，表 1 からわかるように，今回用いたコーパスでは 50 代以上の話者の人数が，107 名中 20 名と少ない。さらに，70 代以上の話者は皆無である。年齢層を広げ，年代のバランスを取ることによって，アノ以外にも年齢の効果が表れる可能性はある。学歴の効果はエーにおいてのみ有意で，大卒の話者の使用確率が中学，高校卒の話者よりも高かった。フィラー全体で見ると，修士以上の学歴の話者の使用確率が中学，高校卒の話者よりも高い傾向があった。学歴の高い話者の方が節頭のフィラーの使用確率が高いのかどうか，今後，節中のフィラーの使用確率も考慮して分析を進める必要がある。講演経験はアノにおいてのみ有意で，経験が増えると節頭のアノは減少する傾向があった。文頭，節頭でのアノの多用は，講演に不慣れな印象を聞き手に与えるかもしれない。

講演の言語特性に関する要因では，20 代～30 代前半の話者を対象とした渡辺・外山（2017）同様，境界の種類の効果と節中語数の効果が有意だった。ただし，境界での出現確率は，渡辺・外山では 文境界＜弱境界＜強境界 の順だったのに対し，本研究では 弱境界＜文境界＜強境界 の順だった。この結果の違いは節頭の接続詞の扱いの違いに起因していると考えられる。渡辺・外山では「で，エー」のようにフィラーの前に接続詞がある場合，そのフィラーは節頭のフィラーとは見なさなかったのに対し，本研究では接続詞は節を繋ぐものと考え，このようなケースのフィラーを節頭のフィラーと見なした。文頭には接続詞が用いられることが多いため，文頭のフィラーの確率が上昇し，このような結果になったものと考えられる。このような結果の違いはあるものの，どちらの研究においてもフィラーの出現確率が 文境界＜強境界 である点には変わりがない。文境界は節境界よりも深い境界であるため，続く発話内容生成のための認知的的負荷が大きく，フィラーの出現確率は文境界の方が高いと予測していたが，そのような結果にはならなかった。文境界では節境界よりも一般に長いポーズが置かれる。そのポーズ中に，後続発話内容の生成が可能なため，フィラーの出現が抑えられている可能性がある。

節中の語数はその節によって伝えられるメッセージの豊かさの指標と考えることができる。そして，語数が多いほど内容が複雑で，適切な言語形式の選択には時間がかかると考えられる。節中語数の増加に伴い節頭フィラーの出現確率も増加しており，フィラーが発

話生成のための認知的負荷を反映しているという仮説を支持している。節中語数要因の効果に関しておもしろいのはマーの結果である。マーだけが，この要因の効果が有意ではなかった。つまり，マーの使用には，適切な言語形式選択のための時間確保とは別の要因が働いていることになる。節頭に接続詞があるとマーの出現確率は上昇した。接続詞は談話の切れ目に，副詞のマーは談話を区切ってまとめるような場合によく用いられる。フィラーのマーも，談話標識としてのマーの性質を受け継いでいるのかもしれない。

節頭接続詞の有無要因に関しては，エーとマーにおいて有意な効果があり，接続詞があるとマーの出現確率は上昇し，エーの確率は低下した。節頭接続詞としては「で」の頻度が高く，「で」は「でー」と引き伸ばされることがよくある。「で」はその直後に声門閉鎖が入って母音が続くと「で，エー」となり，「でー」の母音部分とフィラー「エー」との違いは紙一重である。接続詞があるとエーが減少する理由の一つとして，母音の引き伸ばしがエーの代わりをしている可能性が考えられる。

本研究では，節頭におけるフィラー使用全般の特徴だけでなく，頻度の高い３つのフィラータイプの特徴にも光を当てた。今後，講演のサンプル数を増やし，本研究で明らかになった傾向がどこまで一般化できるのかを調べる予定である。

謝　辞

本研究は科研費基盤(C)「後続要素の複雑さが言い淀みの発生に及ぼす影響についての日英語対照研究」（2015～18 年度，課題番号 15K02553）および「日本語と英語のパラレルコーパスを用いた言い淀みの対照言語学的研究」（2018～20 年度，課題番号 18K00559）の助成を受けて行われた。

文　献

Goldman-Eisler, F. (1968) *Psycholinguistics*, London: Academic Press.

国立国語研究所（2006).『日本語話し言葉コーパスの構築法』東京：国立国語研究所. URL: http://pj.ninjal.ac.jp/corpus_center/csj/

Levelt, W., J. M. (1989) *Speaking: From intention to articulation*. The MIT Press: Cambridge, Massachusetts.

Swerts, Marc (1998) “Filled pauses as markers of discourse structure”, *Journal of Pragmatics* 30, 485-496.

Shriberg, Elizabeth (1994) *Preliminaries to a Theory of Speech Disfluencie*, dissertation submitted to UC Berkeley.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.443.7755&rep=rep1&type=pdf, 2016 年 7 月 14 日アクセス

Watanabe, M. & Korematsu, Y. (2017) “Factors affecting clause-initial filler probability in an English monologue corpus”, *Journal of the Phonetic Society of Japan*, Vol. 21 No. 3, pp.24–32.

渡辺美知子・外山翔平 (2017).「『日本語話し言葉コーパス』と対照可能にデザインされた英語話し言葉コーパスにおけるフィラーの分布の特徴」『国立国語研究所論集』12 号, pp.181-203.

『国語研日本語ウェブコーパス』からの新規語彙素獲得の試み

岡 照晃 (国立国語研究所コーパス開発センター) *

An Attempt to Extract New Lemma Candidates from Ninjal Web Japanese Corpus

Teruaki Oka (National Institute for Japanese Language and Linguistics)

要旨

『国語研日本語ウェブコーパス（NWJC）』は、国立国語研究所がこれまで公開してきた『現代日本語書き言葉均衡コーパス（BCCWJ）』や『日本語話し言葉コーパス（CSJ）』と異なり、形態論情報をすべて形態素解析器『MeCab』と『解析用 UniDic』を使って自動付与している。『BCCWJ』や『CSJ』といった既存のコーパスの整備の際には、コーパスアノテーションと同時に、形態論情報のデータベースである『UniDic DB』に新規短単位語彙素を追加していた。そのためコーパス整備と同時に『UniDic DB』も拡張されてきたが、『NWJC』は全自動で構築されたため、新規短単位語彙素の検出と DB への登録が行われておらず、その箇所で自動解析誤りのままとなっている。そこで本研究では、形態素解析を介さず、文字 N-gram の出現頻度と連接頻度の情報から文字 N-gram の分散表現を作成し、『NWJC』から『UniDic DB』に未登録の新規短単位語彙素の候補を列挙する方法について述べる。これにより DB のさらなる拡張が望めるだけでなく、『UniDic DB』のエクスポートデータで作成される『解析用 UniDic』も拡張されるため、それを用いた再解析によって『NWJC』中の誤解析箇所を減らすことにもつながる。

1. はじめに

国立国語研究所コーパス開発センターでは、現在、現代日本語の形態素解析用辞書として、現代書き言葉解析用 UniDic と現代話し言葉解析用 UniDic の 2 つを公開している [(1)]。これらの解析用辞書の語彙は共通であり、1,745,957 の書字形出現形（表層系）と、その活用変化や異語形を束ねた語彙素（辞書の見出し語相当）を 258,550 を含んでいる (表 1)。しかしながら、2013 年以来、（現代語の）解析用 UniDic には新規の短単位登録がなく、辞書の語彙もすでに古くなっている。これは解析用 UniDic の元となる短単位格納 DB、UniDic DB がコーパスアノテーションと同時に拡張されていくものであり、現代語のコーパス整備が現代日本語書き言葉均衡コーパス以来、大規模に行われなかったことによる。そこでコーパス開発センターでは現在、UniDic へ新たに 5,000 の新規語彙素の追加を計画している。その一環として、本研究では、

*teruaki-oka {at} ninjal.ac.jp

(1) http://unidic.ninjal.ac.jp/download/#unidic_bccwj

表 1　解析用 UniDic の語彙の統計。

語彙素数	258,550
書字形出現形数	1,745,957

表 2　NWJC の統計。

No. of URLs	83,992,556
Tokens	3,885,889,575
Types	1,463,142,939
No. of Characters	33,226,333,292

国語研日本ウェブコーパス（NWJC） [Asahara et al., 2014] からの新規語彙素の候補の抽出に取り組みについて述べる。日本語は英語など分かち書きする言語と異なり、単語境界がスペースで明示されない。そのため新規語彙素を文字列中から発見することは難しい。そこで本研究では、単語分割を介さずに文字 N-gram の分散表現ベクトルを学習する sembei [Oshikiri, 2017] というアルゴリズムを採用し、K-neighbour classification によって UniDic に既に登録されている短単位に近い分散表現ベクトルを持った文字 N-gram を新規語彙素の候補として抽出する。本手法により、名詞に関して精度よく新規語彙素の獲得が可能であることが分かった。

2. 国語研日本語ウェブコーパス (NWJC)

国語研日本語ウェブコーパスは、国語研コーパス開発センターで開発された 100 億語規模の日本語のウェブコーパスである。ウェブページの収集には Heritrix crawler [(2)]が使用され、このクローラを 1 億 URL について URL のリストを更新しつつ、3 か月に 1 度動かし、1 年間変化のないウェブページを収集した。クロールされたページは nwc-toolkit-0.0.2 [(3)]によって正規化（HTML タグ削除と NFKC 正規化の後、文へと分割）した。ウェブ上にはコピーされたページも存在し、収集したページの中にもそれは含まれている。そのため Unix の uniq コマンドを使用し、文を延べではなく、異なりにする作業を行い、重複を排除した。NWJC には、2014 年の 10〜12 月 (2014-4Q) に収集されたウェブページのデータが異なり文集合として格納されている。NWJC の統計データを表 2 に示す。

3. 関連研究

日本語の特徴の一つに分かち書きをしないことがあげられる。そのため日本語解析の一番最初のステップは、単語分割、品詞タグ付け、活用推定等を含んだ（日本語）形態素解析処理である。日本語形態素解析では、形態素解析用辞書を用いた手法が主流であり、広く使われているツールとして、CRF [Lafferty et al., 2001] で辞書に登録されている各表層形のコストを学習

[(2)] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix/
[(3)] http://code.google.com/p/nwc-toolkit/

する MeCab [Kudo et al., 2004] [4]がある。

分かち書きされていない生文から新規語彙素の候補を列挙したいと思った場合、もっとも単純な方法は MeCab のような辞書ベースの形態素解析器を利用することである。MeCab をはじめ、辞書ベースの形態素解析器には未知語処理の機能が実装されており、解析結果が「未知語」となった文字列を切り出せば、それを新規語彙素の候補とできる。しかしこの方法の欠点はそもそも未知語が辞書に載っていないため、正しく切り出されないという点にある。また特に今回対象とする Web テキストにはくだけた表現や省略形も多く、それが辞書中の別の短いエントリとマッチして、未知語が未知語と認識されないことも多い。

そこで本研究では、形態素解析器を介さず、文字 N-gram の分散表現ベクトルを学習する sembei アルゴリズム [Oshikiri, 2017] を採用し、新規語彙素の候補を品詞ごとに NWJC から K-neighbour classifier を使って抽出する方法をとる。sembei アルゴリズムについては次節で述べる。本手法は単語の分散表現ベクトルと近傍法を利用したものであるが、本質的には [Mori et al., 1996] の文字ベースの日本語未知語獲得手法の亜種であるといえる。

4. 提案手法： sembei アルゴリズムと K-neighbour classifier を用いた新規語彙素候補の抽出

4.1 sembei アルゴリズムの詳細とパラメータ設定

sembei アルゴリズムのフレームワークでは、まずコーパス中に頻出する文字 N-gram で文をラティスに変換する。これは頻出文字 N-gram に基づく当該文の可能な分割候補を列挙しているとも捉えられるため、辞書ベースの形態素解析で構築される単語ラティスに類似のものである。次に N-gram ラティス上での共起（連接）の統計値を使い、N-gram のベクトルを学習する。

大規模なウェブコーパスである NWJC から頻出する N-gram を獲得するため、sembei では lossy counting アルゴリズムを使って低頻度要素の逐次削除と、頻度の近似値計算を採用している。本手法では、N-gram 長 : N= 1 〜8 を採用し、lossy counting アルゴリズムにより、NWJC から 22,455,810 個、長さ 1〜8 文字の異なり N-gram を獲得した。

sembei アルゴリズムでは、コーパス中の頻出文字 N-gram のみでラティスを構築する。これに対し本手法では、生コーパス中の高頻度 N-gram だけでなく、なるべく均等に N-gram を選択したい。そのため lossy counting アルゴリズムで獲得した全 N-gram の（近似）頻度の分布に基づき、ラティス構築のための N-gram をランダムに選択した。実際には、22,455,810 個の N-gram から 1,150,000 個を頻度の分布に基づいてランダムに抽出し、NWJC 中の文を N-gram ラティスに変換した。

次に、変換した N-gram ラティス上での N-gram 同士の共起（連接）頻度から各 N-gram のベクトルを学習する。Oshikiri (2017) では、negative sampling ありの skip-gram モデル (SGNS-sembei) [5]を採用しているが、ここではもともとの sembei アルゴリズム [6]が採用している

[4] https://taku910.github.io/mecab/
[5] https://github.com/oshikiri/w2v-sembei
[6] https://github.com/shimo-lab/sembei

eigenwords (OSCCA) [Dhillon et al., 2015] を使用した。ここで 1,150,000 個の N-gram 集合を $V = \{v_1, v_2, ..., v_{1150000}\}$ と表す。v_i は各 N-gram を表している。v_i の頻度を $\#(v_i)$ と表記し、v_i と v_j の連接頻度を $\#(v_i, v_j)$ と表記する。C_L は v_i に対する左連接行列で、C_R は v_i に対する右連接行列である。連接頻度のカウントにも、再度 lossy counting アルゴリズムを使用する。

$$C_L := \left(\frac{\#(v_j, v_i)}{\sqrt{\#(v_i)}\sqrt{\Sigma_k \#(v_j, v_k)}} \right)_{i,j} \in \mathbb{R}^{V \times V}$$

$$C_R := \left(\frac{\#(v_i, v_j)}{\sqrt{\#(v_i)}\sqrt{\Sigma_k \#(v_k, v_j)}} \right)_{i,j} \in \mathbb{R}^{V \times V}$$

$$C := [C_L, C_R]$$

ここで OSCCA の分散表現は $G_1^{-1/2}[\mathbf{u}_1, ..., \mathbf{u}_K]$ であり、G_1 は $\#(v_1), \#(v_2), ..., \#(v_V)$ を要素とする V 次対角行列である。$\mathbf{u}_1, ..., \mathbf{u}_K$ は $\sqrt{C}$ の左特異ベクトルの上位 K 件である。$\sqrt{C}$ の要素は $\sqrt{c}$ である。c は C の要素である。$\mathbf{u}_1, ..., \mathbf{u}_K$ を計算するため、randomized SVD [Halko et al., 2011] を使用し、$K = 200$ に設定した。この方法で、1,150,000 個の N-gram から、それぞれ 200 次元の 1,150,000 個のベクトル（分散表現）が作成された。

4.2 N-gram の分散表現からの新規語彙素候補の抽出

N-gram の分散表現から新規語彙素の候補を抽出するため、 Python machine learning toolkit である scikit-learn の K-Neighbour Classifier を使用する。K-Neighbour Classifier のコンストラクタ引数は、*n_neighbours* = 5、*weights* = '*distance*' に設定した。

N-gram($\in V$) がすでに書字形出現形として UniDic に登録されており、それが指定された品詞をただ 1 つを取りうる場合に限り、当該 N-gram の分散表現ベクトルを訓練用の正例とし、それ以外を訓練用の負例とする。また N-gram ($\in V$) が句読点のような記号を中に含む場合も、訓練用の負例とする。訓練用の N-gram を除いた V（UniDic に未登録の表層形）に対し、それが正例（新規語彙素の候補）か負例かを訓練用データで学習した K-neighbour classifier を使い判別する。

5. 結果：抽出された新規語彙素の候補

訓練用データの正例ので指定する品詞に名詞を設定したとき、抽出された結果はほとんどが名詞か名詞句であった。ただし、この抽出結果をそのまま UniDicDB に追加することはできない。UniDic の短単位登録には厳格な規則があり、DB への登録には人手の確認作業が必要になる。人手での確認作業を含めた提案手法を一度実施した結果、約 50 個の新規語彙素候補の獲得に成功した。品詞：名詞と指定して抽出した結果のうち、K-neighbour classifier の確信度（(K-neighbour classifier.pred_proba(X) 関数の出力)0〜1 の値をとる）で 1.0 で正例と判断した結果を表 3 に示す。

これに対し、品詞：動詞を指定した場合、抽出された候補 N-gram のほとんどが副詞 (e.g. たっぷり, たくさん）＋動詞（UniDic DB に登録済み）であった。sembei アルゴリズムは左右に連接する N-gram の情報のみから分散表現ベクトルを作成するため、当該の N-gram である

表 3　正例と予測され抽出された新規語彙素候補の例（名詞）

新規語彙素候補	説明
ＵＦＣ	Ultimate Fighting Championship
Ｗｉ－Ｆｉ	
ＷｉＭＡＸ	
ＹｏｕＴｕｂｅ	
ｉＭａｃ	
ｉＰａｄ	
ｉＰｈｏｎｅ	
ｉＰｏｄ	
ｉＴｕｎｅｓ	
ｐｉｘｉｖ	
ｔｗｉｔｔｅｒ	
のだめ	キャラクター名
ネトゲ	
ホスホニウム	化学用語（Phosphonium）
ホルミル	化学用語（Formyl）
マツコ	人名
ラジコ	サービス名
ルキア	キャラクター名
絵茶	お絵描きチャット
電マ	電動マッサージ
播戸	苗字

v_i の情報は、分散表現ベクトルの学習時に一切使用しないことが原因だと考えられる。そこで例えば、UniDic に既に登録済みの部分文字列を含む N-gram をあらかじめ除外する、もしくは部分文字列として含むという事柄を学習時に素性として追加することが考えられる。

品詞：形容詞を指定したとき、事前の調査で発見していた NWJC 中の新規語彙素「エモい」は獲得されなかった。これは NWJC 中に「エモい」の出現がわずか 500 件しかなく、lossy counting の時点で頻出 N-gram から漏れていたことが原因である。lossy counting のパラメータを調整したところ、「エモい」も頻出 N-gram として残ったが、頻出 N-gram 数が 7 億と非常に巨大なサイズになった。そのため今後は、OSSCA による行列のバッチ計算から SGDN を使ったオンライン学習による分散表現学習に切り替える方針である。

6. おわりに

本稿では、NWJC のような巨大なウェブコーパスから UniDic の新規語彙素候補をどのようすれば獲得できるのかを示した。sembei アルゴリズムを採用し、文字 N-gram の分散表現を学

表 4　正例と予測され抽出された新規語彙素候補の例（動詞）

新規語彙素候補
たくさんし
たくさんつい
たくさん飲
たくさん作っ
たくさん出
たくさん書い
たくさん入れ
たくさん落ち
たっぷり含まれ
たっぷり詰まっ
たっぷり使
たっぷり使っ
たっぷり入っ

習したのち、K-neighbour classification によって文字 N-gram の中から新規語彙素候補を識別した。名詞に関しては新規語彙素の候補が獲得できたが、動詞に対しては副詞を含むといった冗長抽出が起きてしまった。これを避けるには、文字列前方に副詞を含むような N-gram をあらかじめ除外、もしくは負例にするなどの方法が検討される。

謝　辞

本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。

文　献

[Asahara et al., 2014] Asahara, M., Maekawa, K., Imada, M., Kato, S., and Konishi, H. (2014). Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan *Alexandria*, 25(1-2):129–148.

[Bojanowski et al., 2016] Bojanowski, P., Grave, E. Joulin, A and Mikolov, T. (2016). *arXiv* preprint arXiv:1607.04606.

[Dhillon et al., 2015] Dhillon, P. S., Foster, D. P. and Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16: 3035–3078.

[Kono et al., 2015] Kono, T. and Ogiso, T. (2015). Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pages 282–289.

[Mank., 2002] Manku, G. S. and Motwani, R. (2002). Approximate frequency counts over data

streams. *Proceedings of VLDB '02*, pages 346–357.

[Halko et al., 2011] Halko, N., Martinsson, P. G. and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288.

[Kudo et al., 2004] Kudo, T., Yamamoto, K. and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of EMNLP-2004 (the 2004 Conference on Empirical Methods in Natural Language Processing)*, pages 230–237.

[Lafferty et al., 2001] Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pages 282–289.

[Mori et al., 1996] Mori, S. and Nagao, M. (1996). Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 1119–1122.

[Oshikiri, 2017] Oshikiri, T. (2017). Segmentation-Free Word Embedding for Unsegmented Languages. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 778–783.

アクセント音調の諸相とその動態形式

佐藤　大和（東京外国語大学 大学院総合国際学研究院）†

Pitch Characteristics and Their Dynamic Styles Related to Japanese Accents

Hirokazu Sato (Tokyo University of Foreign Studies, Graduate School of Global Studies)

要旨

日本語の規範的なアクセントの型が，自発発話の中でどのような音調動態として実現されているかを明らかにするため，「日本語話し言葉コーパス(CSJ)」における東京方言話者（女性）1名が発話した11分程度の独話音声を用い，約520のアクセント単位の音調分析を行った．アクセント単位の音調は，大きく分類すると卓立型と非卓立型音調があり，同一アクセントであっても両者によって音調形式が異なっている．また，アクセントが知覚される拍（アクセント拍と呼ぶ）におけるピッチの“上昇”“下降”“平坦”等の音調形式が，音響上のアクセント位置と関連すること，特にアクセント位置が後続拍にくる「遅下がり」現象と“上昇”ピッチが密接な関係にあることを示す．アクセントの遅下がりに関しては，フットリズムとの関連にも触れる．更に，アクセント拍以降の音調の降下特性とアクセント単位の音調動態に関しても議論する．

１．はじめに

日本語（共通語）のアクセントの型は，単語を構成する拍の高低の配置，もしくはピッチが高から低へ変化する際の高の拍の位置（アクセント核）によって記述される．これらの型は，拍を意識した比較的丁寧な発音もしくはその内省によって把握される．一方，実際の発話における音調の動的様相では，このような規範的なアクセントの型がそのまま実現されているわけではなく，アクセントの「遅下がり」現象（杉藤 1980，Hasegawa and Hata 1995）のように，アクセントが認知される拍と音響上観測されるアクセント位置との間に乖離が見られる現象のあることも知られている．この研究は，日本語アクセントの型が，実際の連続音声のなかで，どのような動態形式として実現されているか，また逆に，特定のアクセント型をもたらす音調の動的特徴は何かを明らかにすることを目的としている．

また，これまでアクセントの音調パタンの研究は，実験的に計画された単語リストに基づき，単独もしくは一定の埋め込み文の下で読み上げられた音声データに基づいた研究が多かったが，本報告では発話様式を規定しない自発的な発話音声を分析することによって，アクセントに関わる発話のより多様な実現形態を探ることも目的としている．

本報告は，上記目的の研究において得られた現在までの知見についてまとめたものである．

２．実験試料と分析方法

分析に用いた音声資料は，「日本語話し言葉コーパス(CSJ)」における東京方言話者（女性）１名の独話資料（模擬講演）である．発話時間は11分ほどであり，この中でアクセントの

† sato.hirokazu@tufs.ac.jp

ある約 520 個の音声単位に関して分析した.

この話者の発話においては，発話末の終結ピッチ周波数（F0）が 130Hz 程度であることから，この値を基準値とする Semitone(ST)を求め，F0 と ST の双方から音調特性を見ることとした．ST 上では，基準値より 1 オクターブ高い 12 ST が 260 Hz，2 オクターブ高い 24 ST が 520 Hz に相当する．発話データのピッチ範囲は，2 オクターブ，すなわち 520 Hz 以内に収まっている．話し言葉コーパスのデータから，10 msec のフレーム（Frame(FR)）毎に，時間・ピッチ周波数・ピッチ ST・音声セグメント情報等の時系列を取り出し，各音声セグメント（主に母音）におけるピッチ変化率（F0 変化率Δf：Hz／FR，ST 変化率ΔST：ST／FR）を区分内直線近似で求めた.

アクセントに関しては，音声データの聞き取りによってアクセント型の判断を行うとともに，ピッチ周波数特性に基づいてアクセント位置を定めた．前者の判断によるアクセントのある拍を「アクセント拍」，音響特性から設定したアクセントの時間軸上の位置を「アクセント位置」と呼ぶ.「アクセント位置」は，CSJ のドキュメント（五十嵐・菊池・前川(2006)）の記載に準じており，アクセント拍およびこれに後続する拍のピッチ周波数パタンに基づき，上昇ののち下降する特性においてはそのピークを，緩やかな変化から急峻な下降がある場合は下降の開始点を，下降特性のみの場合はその開始時点を「アクセント位置」とした.

また，一つのアクセントを有する音声単位はアクセント句呼ばれるが，ここで分析される音声単位はこれより狭い単位であり，先行する平板アクセントの語や接続助詞，音調上昇を伴い易い副助詞などを除いた，文構成上の基本的単位（文節のコア部分）であって，原則 1 個のアクセントを有する音声区分である．ここではこれを「アクセント単位」と呼ぶことにする.

３．分析結果

今回抽出したアクセント単位の数を，アクセント単位毎に表１に示す．アクセント型は，語頭（句頭）から数えたアクセント拍の位置で示した．以下，これらの抽出例に基づいた分析結果について述べていく.

表１　分析されたアクセント単位数（アクセント型別）

アクセント	１型	２型	３型	４型	５型	６型	その他	計
抽出数	235	96	100	60	23	5	3	522

３．１　二つの発話モード

アクセントがある句や節の音声区分は，音調特性上から見て，大きく分けて二つの発話モードがあることが分かった．一つは，ピッチ周波数の大きな上昇を伴う卓立型音調様式であり，他の一つは主に緩やかな上昇や下降音調を主体とする非卓立型音調様式である．前者は，発話の主要句や焦点の置かれた語を際立たせた句で実現され，句が特に強調された場合にはより顕著な特性となる．語の単独発話の場合も卓立型音調となる傾向があると考えられる．一方，非卓立型音調は，主要句に続く従属句，単調で軽い発話，メリハリのない発話等で見られる．また非卓立型は，結合形式化した後接要素（「～み]たい」など）でも見られる.

卓立型音調は，ピッチ周波数領域がおおよそ 12～24 ST 区間，すなわちピッチの基底値から 1 oct 以上 2 oct までの高い領域で展開するが，非卓立型音調は，主に 0 ～12 ST の 1 oct までの低い領域で展開している.

二つの音調モードの例を図１(1)(2)に示す．図中(1)は，卓立型音調と従属句の非卓立型音調が組み合わさった例である．(2)は，前置き的発話で，非卓立型音調モードのみから成る例となっている．

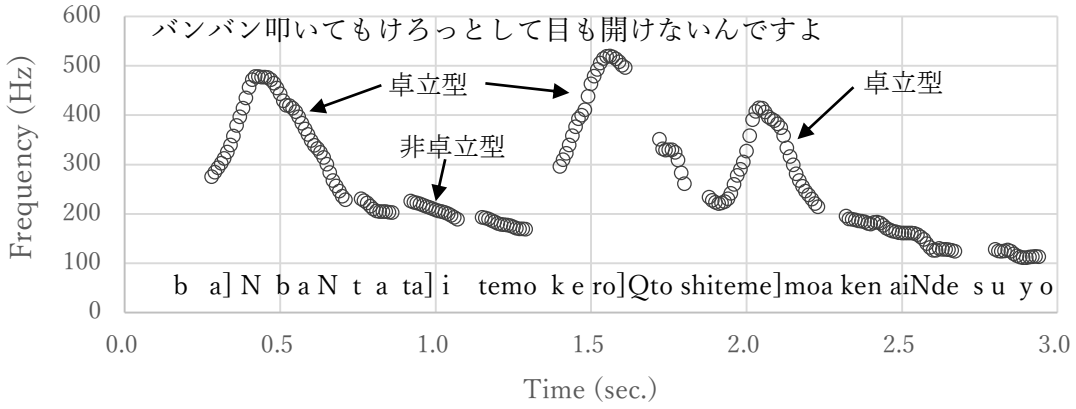


図１（1）卓立型音調モードと非卓立型音調モードの例

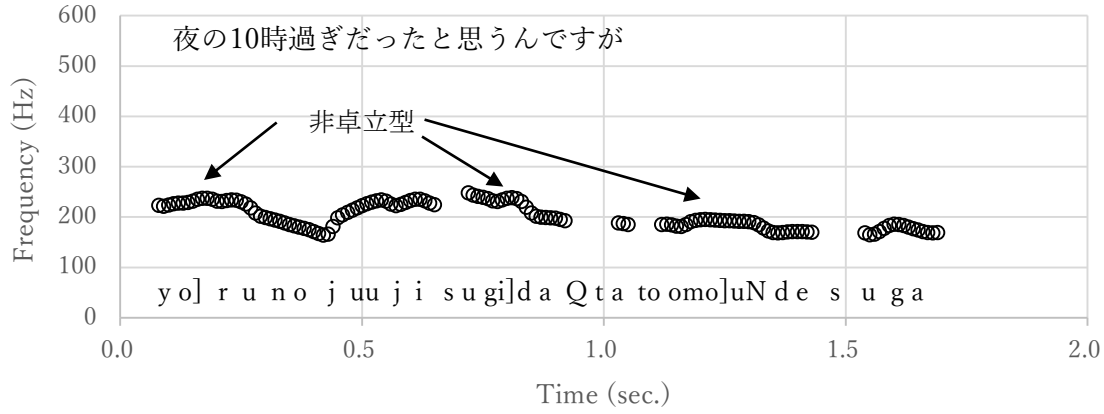


図１（2） 非卓立型音調モードの例

３．２　アクセント拍の音調形式とアクセント位置

前節の図１からも分かるように，(a)の「バ]ンバン」は１型（頭高）アクセントであって，音調的には上昇調で実現されているが，「たた]いても」は２型アクセントで下降調である．つまりアクセントがあるとされる拍の音調は様々であることが予想される．アクセント感覚は，アクセント拍と後続拍との間の高さの相互関係で決まると考えられるが，まずアクセント拍内の音調形式を調べ，アクセント位置との関係を調べた．

アクセント拍の拍内のピッチ周波数を直線近似し，以下の音調形式に分類した．

・下降音調(Falling Pitch: FP) 下降ピッチの音調形式
・平坦音調(Level Pitch: LP) 拍内ピッチの傾きの絶対値が 0.1 ST/FR 未満の場合を平坦のピッチとした．
・上昇音調(Rising Pitch: RP) 上昇ピッチの音調形式
・その他，上昇・下降音調(Rising+Falling Pitch: R&FP)，平坦・下降音調(Level+Falling Pitch: L&FP)なども設定したが，これらは数が少ない．

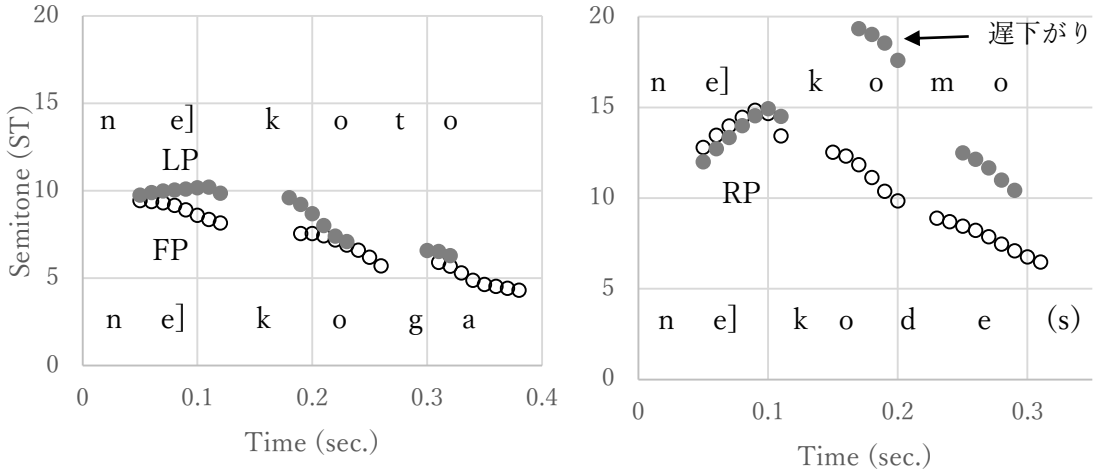


図 2 アクセント拍の各種音調形式
（「猫」を含む各種アクセント単位のピッチパタン）

図 2 に 1 型アクセントの「猫」という語を含む各種音調形式の例を示す．左図の 2 例は，下降(FP)と平坦(LP)音調の例であり，右図の 2 例はともに上昇音調(RP)の例であるが，そのうちの一つは後述するアクセントの遅下がり現象を示すピッチパタンである．アクセント位置（いわゆるアクセント核の位置）は，アクセント拍内の音調形式によって著しく影響される．図 3 は，1 型アクセントにおけるアクセント位置の生起度数を示したものである．横軸は音韻境界から測ったアクセント位置を 10msec 毎のフレーム単位で示している．境界に隣接する直近フレームは±10 msec となる．C1V1 境界はアクセント拍の（子音–母音）境界，V1C2 境界はアクセント拍末境界であり，当該母音と後続子音の境界である．また，C2V2 境界は後続拍内の（子音–母音）境界である．

C1V1 境界近傍にアクセントが分布する音声は，その 90% が下降音調であり，V1C2 境界近傍でアクセントがある音声のうち，72% が上昇音調，27% が平坦音調であった．C2V2 境界以降に分布するアクセントは遅下がりのものであり，72% は上昇音調であった．2 型アクセント，3 型以上のアクセントにおいてもほぼ同様の結果が得られた．3 型以上のアクセント拍では，その殆どが下降と平坦の音調であり，上昇調のものは後接要素が付いた複合形であ

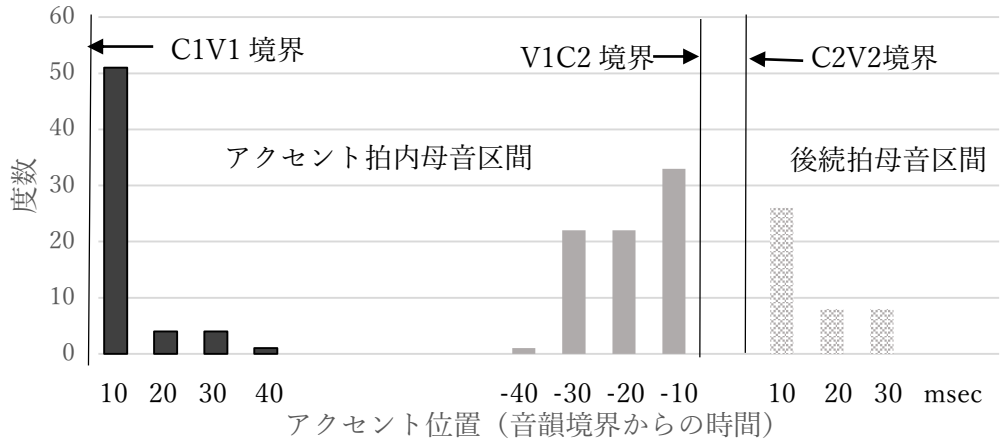


図 3　アクセント位置とその生起度数（1 型アクセントの場合）

って，それが強調されて上昇調となったものが多い（例：「思いま]して」，「そんなこ]と」など）．

以上見たように，アクセント拍の音調形に関して，それが下降調の場合，アクセント位置は当該拍の母音の開始点近傍に分布し，上昇調の場合は当該拍の母音の末尾近傍に分布するか，もしくは後続拍の母音部にまで達する．平坦調の場合は，当該拍末尾近傍にアクセントのくる場合が比較的多い．

３．３　「遅下がり」現象

前節で示したアクセント位置の分布の中で，アクセント拍より遅れて次拍上に観測されるアクセント位置の結果を示した．これがアクセントの遅下がり現象である．「遅下がり」は，81 例で観測された．表２に示すように，1 型（頭高）アクセントが 72.8%と最も多く，そのうちアクセント拍が上昇音調と上昇・下降音調のものが約 80%を占めた．このことからまず第一に，「遅下がり」は，アクセントが上昇音調で実現されることと関連があると考え

表２　アクセント遅下がり生起数（アクセント型別）

	1 型ア	2 型ア	3 型ア以上	計
遅下がり生起数	59（72.8%）	11（13.6%）	11（13.6%）	81

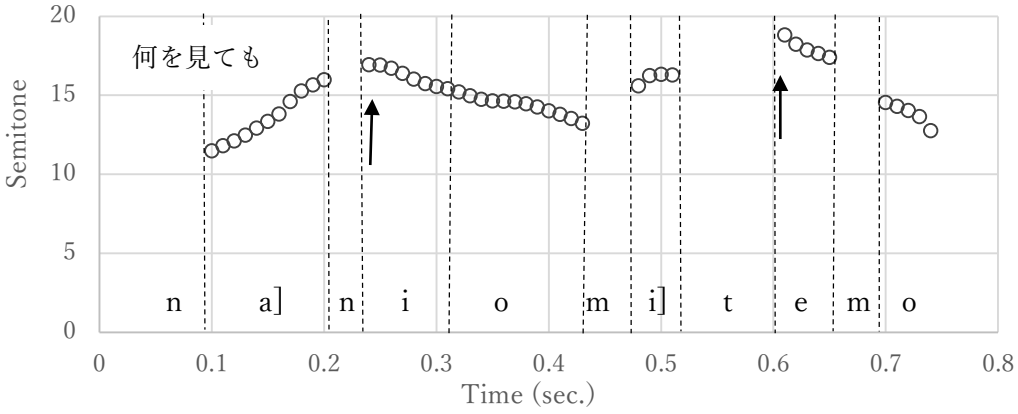


図４　アクセントの「遅下がり」の例

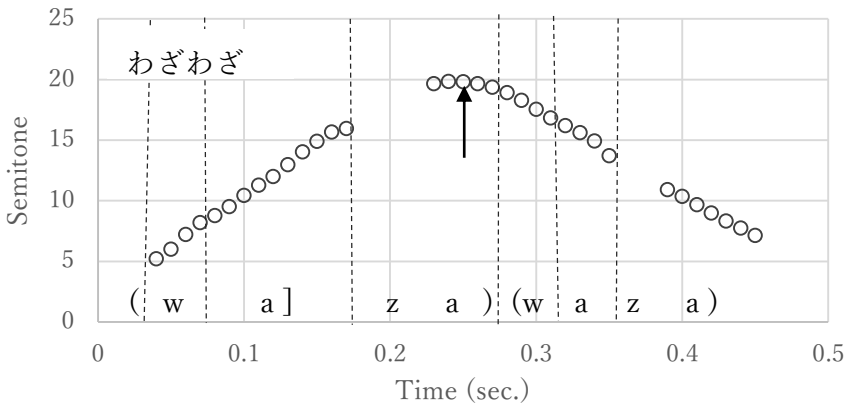


図５　「遅下がり」とフットリズムの関係の例

られる．1 型アクセントでは，第 1 拍目の音調は上昇であることが多く，その音調がアクセント知覚のキーとなっている．さらに第 2 拍目の音調下降は許容範囲の広いことが報告されており，これらが「遅下がり」の一因と考えられる（佐藤 2016, 2017）．卓立型で，かつ強調的発話である場合には，特に遅下がり現象を示すケースが多い．「遅下がり」の例を図 4 に示す（上向き矢印↑がアクセント位置を示す）．

次に，2 拍を 1 単位とするフット（脚）のリズムで発話される場合に，「遅下がり」が生起しやすいことを述べる．長音，撥音などを含む長音節にアクセントがあって上昇音調の場合に，アクセント位置が音節内 2 拍目の長音部や撥音部にくることがあるが，フットリズムの場合には，短音節（軽音節）の連続においてもアクセント位置が次拍にくる「遅下がり」が見られる．図 5 にその一例を示す．これは 4 拍 1 型アクセントの例であるが，（わ]ざ）（わざ）と 2 フットで発音され，アクセントは最初のフットにおけるピッチ上昇で実現されている．このように 2 拍が一つのまとまりとして上昇調で発音されることが「遅下がり」の原因のひとつと考えられる．「遅下がり」に関する詳細は，（佐藤 2018）を参照のこと．

3．4　アクセント拍後の下降特性

一般に，アクセントは高いピッチからの下降によって実現されると考えられているが，どの程度のピッチ降下が生じているのかを調べた．3 拍語 1 型アクセントの場合，アクセント拍が下降調の場合と上昇調の場合の例を図 6 と図 7 に示す．図 6 では，アクセント位置での高さと当該拍末の高さとの関係，図 7 はアクセント位置での高さと次拍末の高さとの関係を示している．大略的に言うと，両図とも直線関係の傾きはおおよそ 1 で，拍末／次拍末ではアクセントの高さより 2 ST 程度降下している．アクセント拍が平坦音調形の場合も含めて，各音調形におけるアクセント位置（10 ST と 20 ST(下降調の場合を除く)）からのピッチ下降量（下降音調の場合は拍内下降，平坦，上昇音調の場合は拍間下降）を表 3 に示す．平坦，上昇音調の場合は，ピッチ降下量はアクセントの高さの高低に関わらず 2〜3 ST 程度である．下降調の場合は当該拍内での降下量であり，平坦，上昇の場合より降下量は小さい値となっている．

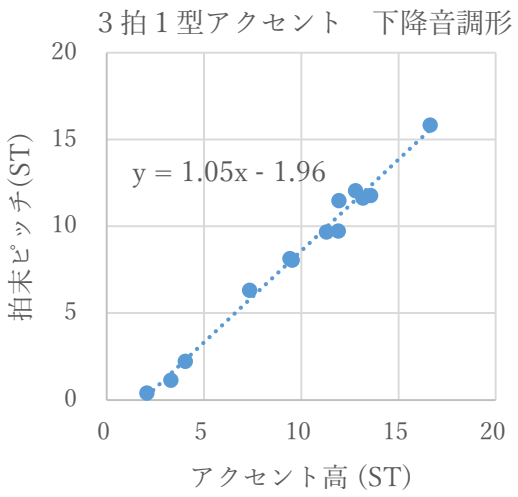


図 6 アクセント位置と拍末の高さの関係

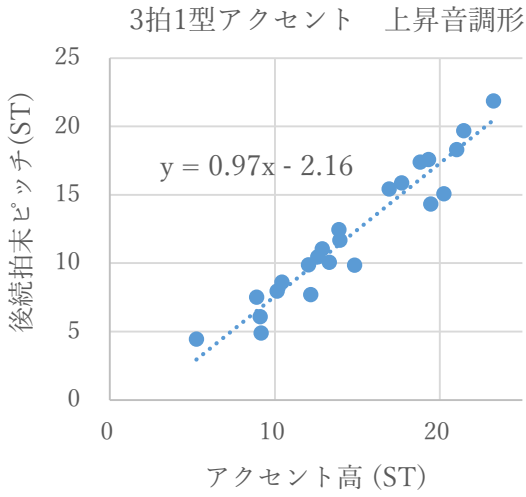


図 7 アクセント位置と次拍末の高さの関係

表３　アクセントの高さからのピッチの拍内下降（下降調）と拍間下降（平坦・上昇）

アクセント拍音調形	下降	平坦		上昇	
アクセントの高さ(ST)	10	10	20	10	20
ピッチ降下量(ST)	1.42	3.39	3.14	2.43	2.71

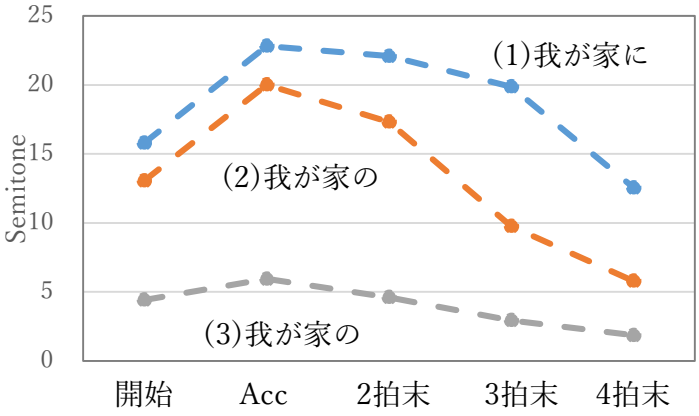


図８　「我が家（に／の）」の音調パタン

図８に４拍１型アクセントの場合の音調パタンの例を示す．図中，（開始）は第 1 拍目の母音開始点，（Acc）はアクセント位置，以降は各拍末におけるピッチのパタンである．(1)の場合は(Acc)から次の拍末までのピッチ降下は極めて小さいが，3 拍から４拍末にかけて大きなピッチ降下がある．(2)では，2 拍目から３拍目にかけての降下が大きい．このような，2 拍目以降の大きな降下特性は，アクセント知覚に直接寄与するというよりも，アクセント単位の終結下降(Coda)的性格があるものと思われるが，今後アクセントとの関係について検討を進める予定である．

３．５　その他観測された事象

最後に，今回の分析で観測された他の事象について紹介したい．表１において，アクセントの型が（その他）のところにある３例である．これらは，音声の聞取りによりアクセントの可能性が二つあり，どちらか一つに決めかねた例である．図９に，「先日のことなんです」の音調パタンを示す．ここで，まず「ことなんです」は，/koto]naNdes/のように/to/にアクセントがあるように聞こえるが，/kotona]Ndesu/のように/na/にアクセントがある

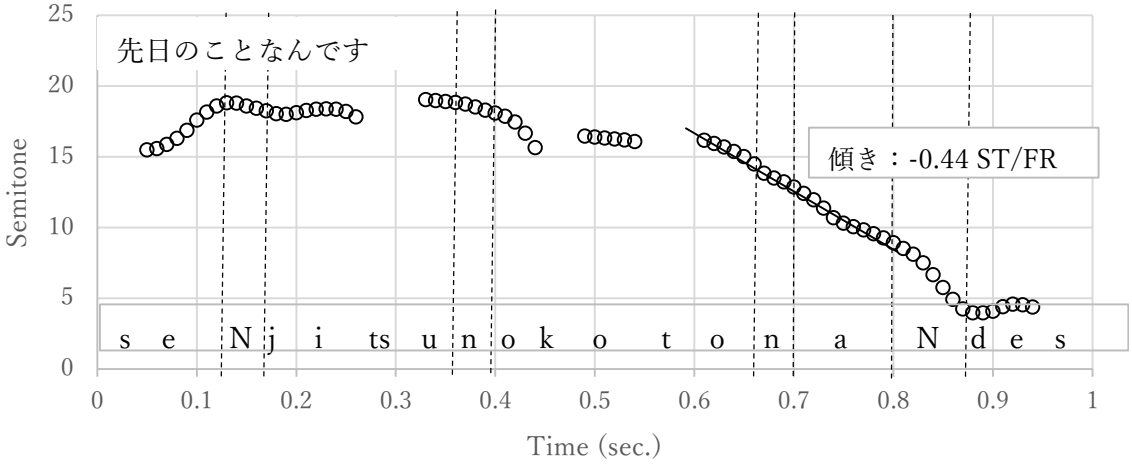


図９　継起ダブルアクセントとその音調パタン

ようにも聞こえる．二つのアクセントが隣り合うため，同時に二つが知覚されることはない．これを仮に継起ダブルアクセントと(sequential double accents)呼んでおく．
音響特性を見ると/to/から/na/にかけて一定の急峻な下降（傾き：-0.44 ST/FR）があり，かつ/N/部ではさらに急峻な下降となる．このため，二つのうちのどちらのアクセント知覚も可能となると考えられる．

「～なんです」という後接要素は単独発話では「な」にアクセントがあるが，ゆっくりとした丁寧な発音では，前部に平板型アクセントの語がくるとこのアクセントが実現する（鼻＋なんです→ /hanana]Ndesu/）．一方，アクセントのある語がくると前要素のアクセントが優先され，後接要素のアクセントは抑圧されてしまうが（花＋なんです→ /hana]naNdesu/），連続発声や自由発話になると，必ずしも抑圧されずにもともとあるアクセント核の性質が顕在化するのではないかと推測される．

４．おわりに

「日本語話し言葉コーパス(CSJ)」を使用して，東京方言話者１名の自発発話音声に見られるアクセント単位の音調パタンの動的側面を分析し，卓立音調と非卓立型音調，アクセント拍の音調形式とアクセント位置の関係，アクセントの「遅下がり」をもたらす要因，アクセント拍およびそれ以降の音調降下特性，継起ダブルアクセント等に関して報告した．

今後は，発話者を増やすなどしてデータの増強を図るとともに，特に，アクセント単位の音節構造，リズム構造と非リズム構造等と音調特性の関係，などに関して研究を進めていく．

謝　辞

本研究は，国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」における共同研究員として実施したものである．

文　献

Hasegawa, Y. and Hata, K. (1995). “The function of f0-peak delay in Japanese”, Proceedings of the 21st Annual Meeting of the Berkeley Linguistics Society, pp.141-151

五十嵐陽介・菊池英明・前川喜久雄(2006).「報告書　日本語話し言葉コーパスの構築法　『第７章　韻律情報』」，URL: pj.ninjal.ac.jp/corpus_center/csj/document.html

佐藤大和(2016).「共通語における動的音調形式とアクセント知覚」，日本音声学会　第 334 回研究例会，2016.12，於：十文字学園女子大学

佐藤大和(2017).「アクセント核のあとピッチの急峻な降下はあるか？−ピッチの動態特性とアクセント知覚−」，3-8-4，日本音響学会 2017 春季研究発表会講演論文集

佐藤大和(2018).「自発発話データから見たアクセントの遅下がり現象」，第 32 回日本音声学会全国大会予稿集，pp.66-71

杉藤美代子(1980).「“おそ下がり”考−動態測定による日本語アクセントの研究」，pp.201-229，徳川宗賢編「論集日本語研究 2　アクセント」，有精堂出版

日本語複単語表現レキシコン JMWEL の概要
－ 動詞性表現を中心に －

首藤 公昭（福岡大学名誉教授）
田辺 利文（福岡大学）
髙橋 雅仁（久留米工業大学）

An Overview of
a Lexicon of Japanese Multiword Expressions: JMWEL

Kosho Shudo (Fukuoka University, professor emeritus)
Toshifumi Tanabe (Fukuoka University)
Masahito Takahashi (Kurume Institute of Technology)

要旨

コロケーション，決まり文句，慣用句，準慣用句などの長単位表現とその派生表現，計約 140,000 の見出しからなり，平仮名べた書き見出しのほかに，形態素分かち書き，構文機能，構文構造，内部修飾可否情報，文脈条件，呼応情報，語釈などを与えた日本語複単語表現レキシコン JMWEL の概要を動詞性表現を中心に紹介する．

１．はじめに

自然言語におけるコロケーション，慣用句，決まり文句など，単語の境界を越えた長単位表現は，従来，計算言語学 Computational Linguistics, CL や自然言語処理 Natural Language Processing, NLP の分野では例外的言語現象とみなされ，必ずしも十分な対応がなされてこなかった．しかし，近年，これらの表現が日常言語でかなり多種類，高頻度で使われていることが改めて認識され，(Sag et al. 2002) がこの種の表現を複単語表現 Multiword Expression, MWE と名付けて NLP における重要性を指摘したのを発端に，それらの表現のコーパスからの自動抽出，レキシコン開発等々，コンピュータ処理に向けた種々の基礎研究が各国で行われるようになって今日に至っている[1].

また，この種の表現は言語の獲得・認知の観点からも重視されるようになり，言語学分野でも定型言語 Formulaic Language (Corrigan et al. 2009, Jiang et al. 2007)，単語連鎖 Lexical Bundles (Biber et al. 1999)，構文文法 Construction Grammar (Fillmore et al. 1988)などの枠組みで，話し言葉，書き言葉の両面から盛んに研究が行われている．

本稿では，筆者の一人が 1960 年代からフレーズベースの日本語処理研究の一環として編纂を進めてきた日本語複単語表現レキシコン Japanese MWE Lexicon，JMWEL の概要と現状を下記の動詞性表現部分レキシコン（１），（２）を中心に紹介する．(Tanabe et al. 2014, 高橋ほか 2018, Shudo et al. 2011, 首藤ほか 2010)　JMWEL の収録見出し数は，異なりで現在 140,000 件程度である．JMWEL の動詞性表現部分レキシコンには，その主要部分として

（1）日本語動詞性複単語表現（1 類）レキシコン：JMWEL_verbal (class1) v3.2 － ガ格，ヲ格，ニ格を介した動詞と名詞のコロケーション集（慣用句等を含む）－，（2）日本語動

[1] 例えば、SIGLEX-MWE(http://multiword.sourceforge.net/)にあるワークショップの一覧をみると 2018 年には LAW-MWE-CxG2018 が開催されている．

詞性複単語表現（2 類）レキシコン：JMWEL_verbal (class2) v3.2 － 述語動詞と種々の語とのコロケーション集（慣用句等を含む）－ が有る．

２．日本語複単語表現レキシコン JMWEL

２．１ JMWEL の特徴

（1）収録表現は，コロケーション，慣用句，準慣用句，決まり文句，格言，諺，一部の複合語，四字熟語，不完全句，挨拶・呼びかけ・応答表現など，日本語における特異表現を幅広くカバーしている．

（2）表現の構文機能，形態・構文構造を与えている．

（3）必ずしも隣接しない語の共起もデータ化している．（例えば，慣用句「手を広げる」には「手を/外国にまで/広げる」など，ギャップの可能性を記載.）

（4）語の長距離呼応をデータ化している．（例えば，「たった一つも/世の中に存在し/ない」など）

（5）不完全慣用表現を収録している．（例えば，「猫に小判」，「ピンからキリまで」など）

２．２ 採録表現

JMWEL では，新聞記事，雑誌記事，小説，随筆，事典・辞書類などの広範な文書から，主として編者の内省により非構成(イディオム)性，および，要素語間の強い共起性のうち少なくとも一方の特異性を持つ単語列を MWE として抽出・収録した．JMWEL の見出し 2,000 個程度をランダムに抽出して調べたところ，約 38%が非構成性を，約 92%が強い単語間共起性を持ち，両方を併せ持つ MWE は 30%程度であった．

２．２．１ 非構成性

要素単語の標準的な機能から表現全体の意味を規則で導くことが難しい表現を非構成性 MWE として収録した．ここでは，単語列 $w_1w_2…w_n$ がまとまった構文・意味・談話上の機能をもち，かつ，$w_1w_2…w_n$ におけるいずれかの単語 w_i $(1≤i≤n)$ をその同義語または類義語 x に置き換えた $w_1w_2…w_{i-1}xw_{i+1}…w_n$ が無意味になるか，全く異なる意味になる，あるいは，不自然になるとき，単語列 $w_1w_2…w_n$ は非構成性 MWE であると近似する[2]．例えば，「赤の他人」は“全く知らない人”の意味では「真紅の他人」に，また「顔を売る」は“アピールする”の意味では「顔を販売する」に置き換えることができないため，非構成性 MWE であるとする．この判断は基本的に内省によっている．例えば，非構成性 MWE には表 1 に示すような種類がある．

表１ 非構成性 MWE の例

種類	例
意味上の非構成性を持つ表現	赤の他人，顔を売る，頭が切れる
形態・構文上での構成性が不備，あるいは不明瞭な表現	とはいえ，ありがとう，お疲れ様
一部の支援動詞構文	批判を加える，計画を立てる
一部の複合語	打ち拉がれる，袋叩き
四字熟語	一生懸命，一心不乱
慣用的な比喩表現	命の限り，血の雨が降る

[2] このような単語の置換不能性がコロケーションのもつ重要な性質の１つであることは（Manning et al. 1999）でも指摘されている．

２.２.２　要素間の強い共起性

表現を構成する単語間で共起性が強い表現を採録した．この種の表現は，構文・意味解析において係り先を優先的に決定して解析の曖昧さを低減する処理や語の出現を予測する種々の処理に有効である．形式的には，単語列 $w_1w_2…w_n$ がまとまった構文・意味・談話上の機能をもち，かつ，$w_1w_2…w_n$ におけるいずれかの単語 w_i $(2≤i≤n)$ について条件付後方出現確率 $pf(w_i|w_1…w_{i-1})$が，あるいは，単語 w_j $(1≤j≤n-1)$ について条件付前方出現確率 $pb(w_j|w_{j+1}…w_n)$ が相対的に高いという確率的な特異性を持つとき，単語列 $w_1w_2…w_n$ は単語間共起性の強い MWE であるとする．例えば，「手を拱く」，「ぐっすり眠る」は，pb(手|を拱く)，pf(眠る|ぐっすり)が大きいと判断して単語間共起性の強い MWE であるとする．この基準は内省によって判断しているが，3.2 で述べる如く，収録結果の妥当性は WEB 上の大量日本語コーパスを用いて統計的に推定されている．単語間共起性の強い MWE には，例えば，表 2 に示すような種類がある．

表２　単語間共起性の強い MWE の例

種類	例
共起性の特に強い表現	風前の灯，ずぶの素人，手を拱く
格言，諺，故事成句の類	急がば回れ，初心忘る可からず，石の上にも三年
擬音，擬態語を伴う表現	ぐっすり眠る，ポッカリと空く，クルクル回る
その他共起性が比較的強いと思われる表現	肩の荷を下ろす，景気が上向く，メリハリの利いた
概念に固有の固定的言い回し	情報検索，女流作家，機械翻訳

２.３　JMWEL の編成

対象表現が多岐にわたるため，JMWEL は，以下のように分割して編集・管理している．以下の 1〜9 は自立語性表現部分レキシコン，10，11 は機能語性表現（複合辞的表現）部分レキシコン，12〜18 はトピック別の部分レキシコンである．19 は現在構築中である．

1. **名詞性複単語表現レキシコン JMWEL_nominal**：
   「無二の親友」，「あれやこれや」，「愚の骨頂」などの約 23,600 表現
2. **動詞性複単語表現（1 類）レキシコン JMWEL_verbal (class 1)**：
   「手を結ぶ」，「意味がある」，「沽券に関わる」など，『名詞』＋「が，を，に」＋『動詞』の形式の句約 35,800 表現
3. **動詞性複単語表現（2 類）レキシコン JMWEL_verbal (class 2)**：
   「骨の髄までしゃぶる」，「ゼロからやりなおす」，「目から鱗が落ちる」など 1 類，3 類以外の動詞的な句約 17,000 表現
4. **動詞性複単語表現（3 類）レキシコン JMWEL_verbal (class 3)**：
   「放り出す」，「飲んだくれる」，「秋めく」などの複合動詞的な句約 3,700 表現
5. **形容詞性複単語表現レキシコン JMWEL_adjective**：
   「頭が痛い」，「性格がきつい」，「途方も無い」などの形容詞句約 5,200 表現
6. **形容動詞性複単語表現レキシコン JMWEL_adjective verbal**：
   「願ったり叶ったり」，「足手纏い」，「判で押した様」などの形容動詞性の句約 2,600 表現
7. **連用修飾複単語表現レキシコン JMWEL_adverbial**：

「思いもよらず」,「気を引き締めて」,「心を鬼にして」などの連用修飾句（副詞的な句）約 16,300 表現

8. 連体修飾複単語表現レキシコン JMWEL_adnominal：
「世に言う」,「筋の通った」,「得も言われぬ」などの連体修飾句（連体詞的な句）約 16,300 表現

9. 談話指標的表現レキシコン JMWEL_discourse marker：
「そうは言っても」,「とはいえ」,「驚くべき事に」など，文頭の談話指標的，文接続詞的，文副詞的な句約 1,300 表現

10. 文末表現（終助詞，助動詞性表現）レキシコン JMWEL_post-predicative：
「～かもしれない」,「～てもよろしい」,「～たところだ」,「～なければなりません」,「～で頂けませんでしょうか」など，話者の態度や相互行為情報，判断情報，テンス，アスペクト，モダリティ，ポラリティ情報等を与える助述（文末）表現，約 4,650 種

11. 関係表現（格助詞，副助詞，接続助詞性表現）レキシコン JMWEL_postpositional：
「～における」,「～のいかんにかかわらず」,「～の甲斐あって」,「～ところの」,「～を励みに」,「～を機に」,「～かの如く」,「～に従って」,「～もそこそこに」などの助詞的表現約 2,700 種

12. 慣用句レキシコン JMWEL_idiom：
「油を売る」,「真っ赤なウソ」,「足が遅い」などの典型的慣用句約 4,900 表現

13. 格言・諺・成句・決まり文句レキシコン JMWEL_proverb/saying/cliché：
「河童の川流れ」,「義を見てせざるは勇無きなり」,「清水の舞台から飛び降りる」などの約 4,000 表現

14. オノマトペ共起表現レキシコン JMWEL _ onomatopoeic：
「グラリ」,「カラカラと」,「ガッツリ食う」などの擬態・擬音語とそれらを伴う典型表現約 34,500 種

15 .四字熟語レキシコン　JMWEL_four character word：
「切磋琢磨」,「支離滅裂」,「魑魅魍魎」などの約 3,500 表現

16. 慣用的不完全句レキシコン　JMWEL_incomplete phrase：
「病は気から」,「棚からボタ餅」,「蟹の甲より年の功」,「石の上にも三年」など，独立してよく使われる，句に纏まらない表現約 470 種

17. クランベリー型表現レキシコン JMWEL_cranberry：
「しがみつく」,「後ろめたい」などのクランベリー形態素(候補)を含む表現約 180 種

18. 呼びかけ・応答・挨拶・独言・間投表現レキシコン JMWEL_call/response/greeting /monologue/interjection：
「参ったなあ」,「どういたしまして」,「あらマア」,「オット」,「本当？」などの約 1,100 表現で，＜驚き＞，＜疑問＞，＜困惑＞など，発話者の感情 27 種と重み付きで対応付けられている

19. 用例文と英訳付き複単語表現レキシコン JMWEL_with J/E sample sentences：
複単語表現約 5,000 に対して用例文とその英訳(案)が記載されている．例えば，慣用句「油を売る」には「彼は勤務中に油を売ってばかりいる．“He is always_messing around_while on his duty”． 彼はよくあの居酒屋で油を売る．“He often_wastes time in idle talk_ at that pub”.」と記載している．

３．動詞性複単語表現レキシコン JMWEL_verbal

動詞性複単語表現（1 類）レキシコン JMWEL_verbal (class1)は，日本語文の最小基本型とも言える　(1)『名詞』+「を」+『動詞』，(2)『名詞』+「が」+『動詞』，(3)『名詞』+「に」+『動詞』 の三つの形式の書き言葉動詞性 MWE 約 35,800 を収録したレキシコンである．（ただし，『サ変名詞』+「を」+「する，遣る，行う，実行する」，『サ変名詞』+「が」+「できる」の形式の表現は一部を除き収録対象外としている．）

いっぽう，動詞性複単語表現（2 類）レキシコン JMWEL_verbal (class2) は，上記 1 類と 3 類動詞性 MWE（複合動詞的表現）を除く書き言葉動詞性 MWE 約 17,000 を収録したレキシコンである．

表現の採録基準は，前述の如く非構成性と要素語間の強い共起性であるが，自由結合句に比較的近いコロケーションから典型的慣用句，典型的決まり文句に亘るかなり広い編集となっている．

３．１　JMWEL_verbal の記載情報

本レキシコンは，Microsoft Excel で作られた xlsx 形式のファイルとして作成されている．xlsx ファイルの各 1 行に 1 表現を対応付け，A～K 欄に各種情報を記載している．例えば，「労に報いる」という表現に対して与えた情報を A～K 欄の順に列挙すれば以下のようになる．（欄の区切りを・・・で，空データを φ で示す．）

class1・・・ろうにむくいる・・・ろう-に-むくいる・・・労-に-報いる・・・VP_a3・・・[*Nni]*V30・・・<adnom. modifier-no>*・・・むくいる・・・報いる・・・φ・・・φ

A～K 欄の情報は，概略，以下の通りである．

A 欄（種別）：動詞性表現の種別を記す．class1：1 類，class2：2 類，class3：3 類

B 欄（見出し）：平仮名ベタ書き見出しを与える．末尾の活用語は終止形（一部，命令形）で収録している．

C 欄（分かち書き）：形態素分かち書きを示す．形態素には単語，接頭語，接尾語，接頭造語要素，接尾造語要素がある．形態素間の区切りはハイフン「-」（明確な区切り）あるいはアンダースコア「_」（弱い区切り）で示している．活用語尾は一部の例外を除き，切り離さない[3]．

D 欄（異表記）：片仮名表記，漢字表記，送り仮名の有無など，表記の多様さを正規表現類似の形式で記載している．例えば，「行(な)う」は「行なう」，「行う」の可能性，「(在/有)る」は「在る」，「有る」の可能性を示す．

E 欄（形態種別）：形態上の種別を VP_α_β の形式にコード化して与える．VP は表現が動詞句（Verb Phrase）であることを意味する．α 部は，例えば，英数字列 a1 で表現が『名詞』＋「を」＋『動詞』の形式であること，d7 で『名詞』＋「が」+『名詞』＋「を」＋『動詞』の形式であることを示す．β 部は表現末尾に助動詞「られる」，「させる」，「ない」，「ぬ」などや形式的自立語，「する」，「ある」などが用いられている場合に，それらを英小文字ローマ字綴り rareru, saseru, nai, nu, suru, aru などで表わす．

[3] 本レキシコンには，形態素解析機との整合を取る際に有効な情報として，ハイフン，アンダースコアで区切った単語候補，複合語候補のリストを添付している．

F 欄（構文構造）：係り受けの修飾子，被修飾子の対を括弧[ ]で括った 2 項句表示で構文構造記述を与える．即ち，句 α（の主辞）が句 β（の主辞）に係って出来た句 αβ の構造記述を，α，β の構造記述 a，b を使って[ab]とする．基本構成要素の構造記述は，自立語を品詞記号で，機能語(および相当語)を英小文字のローマ字綴りで与える． 文節内の語の接続も，便宜上，2 項句構造として記述する．例えば，「顔-を-揃える」の構造記述は[[*Nwo]*V30]とする．ここで，N は「顔」が名詞であること，V30 は「揃える」が終止形の動詞であることを表す品詞記号，wo は「を」が機能語（格助詞）であることを意味する．アスタリスク * は，直後の N「顔」が「元気な顔」のような連体修飾を，V30 の「揃える」が「皆が揃える」のような連用修飾を受ける可能性があることを示している．このようにアスタリスクは後接する句の表現内における独立性を示し，表現中にギャップが生じる可能性がある事を示している． 並列構造は，括弧< >または《 》で，並列される要素は括弧( )で表わす．例えば，「見栄-も-外聞-も-捨てる」の構造記述は [<(Nmo(wo))(Nmo(wo))>*V30] とする．ここで，mo(wo)は，係助詞「も」が深層のヲ格で使われていることを示す．

G 欄（前方文脈条件）：例えば，「目-に-会う」は，「つらい-目-に-会う」のように「目」に対する連体修飾句を必須的に要求するが，修飾句を表現レベルでは特定しにくいので，連体修飾句が文頭側に必須であることを G 欄に<adnom. modifier>*と記載している．

H 欄，I 欄（主動詞部）：収録表現の末尾主動詞部は終止形（一部，命令形）である．終止形以外をカバーするには末尾の主動詞部を活用変化させればよいが，本レキシコンでは，すべての活用形に応じて見出し化しておくことはせず，H，I 欄に末尾主動詞部を抜き出して再録するに留めている． 一般の動詞辞書等を用いれば，必要な活用形はこの情報で容易に導出できる．

J 欄（活用）：一体性の特に強い表現，例えば，格言，諺，決まり文句，一部の古語表現などは，末尾を活用変化させて用いられることは殆どない．この種の表現には J 欄に「活用不要」などの記載をしている．

K 欄（語釈）：慣用句，諺，格言，決まり文句でその意味が難解と思われる表現 500 種程度（1 類の場合）にはユーザーの便宜のため語釈を入れている．

３．２　JMWEL_verbal(class1)の統計的性質

200 億文からなる日本語 WEB コーパスにおける単語 1～7 グラムの出現頻度を求めた Google の大規模データ GSK2007-C（工藤ほか 2007），(以降，GoogleN グラムデータと略記する）との比較によって，JMWEL_verbal(class1)の統計的性質を調べた[4]．詳細は（田辺ほか 2018，Tanabe et al. 2014）に譲るが，主要な調査結果は以下の 2 点である．以下では，対象複単語表現を $w_1w_2w_3$ と記す．ここで，w_1，w_2，w_3 は，それぞれ，『名詞』，『格助詞（「を」，「が」，「に」のいずれか)』，および『動詞』とする．

１．本レキシコン収録表現 $w_1w_2w_3$ の前部分列 w_1w_2 の 14,075 表記の内，10,548 種が GoogleN グラムデータの $w_1w_2w_3$ の w_1w_2 として出現していた．GoogleN グラムデータ上でそれら 10,548 個の w_1w_2 ごとに，続く動詞の出現頻度を求めた結果，本レキシコンにおける動詞 w_3 が GoogleN グラムデータ上で出現頻度第 1 位である場合が 4,983 件であり，当該表現 w_1w_2 の (4,983/10,548)*100=47.24 % で条件付出現確率が最大の動詞 w_3 が選ばれていると推定できた．「ちょっかい-を-出す」，「熱戦-を-繰り広げる」，「アクション-を-起こす」などがこれらに該当する．同様に，第 2 位の場合は 1,495 件で 14.17%，3 位

[4] 調査対象の JMWEL_verbal(class1)は 2010 年時のバージョンである．

は 786 件で 7.45%，4 位は 433 件で 4.11%であった．20 位までの結果をグラフ化して図 1(a)に示す．このことから，本レキシコンに収録されている表現は，条件付確率 $p(w_3|w_1w_2)$の高いものほど多いという傾向が確認された．

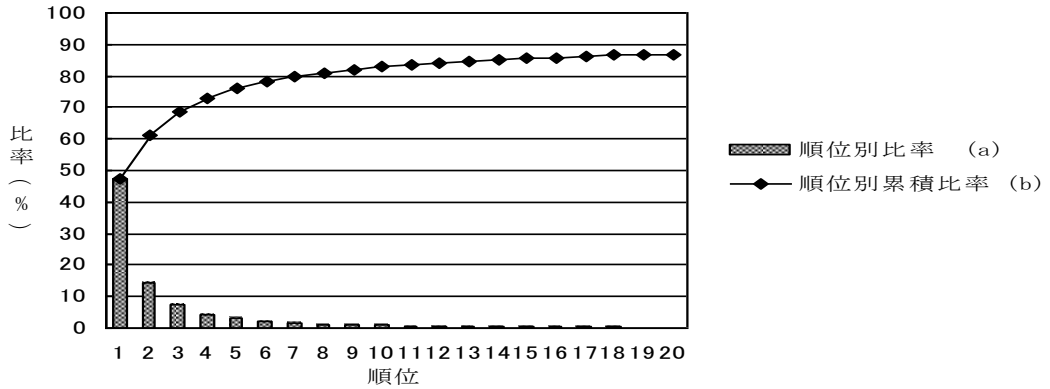


図 1　『名詞』+『格助詞（を，が，に）』+『動詞』型表現の GoogleN グラムデータ
における動詞の出現頻度順位別の動詞採録率(a)，および，順位別の動詞採録累積比率(b)

２．GoogleN グラムデータにおける上記と同じ形式の表現 $w_1w_2w_3$ において，w_1w_2 に続く w_3 に関する正規化エントロピー$H_f(w_3|w_1w_2)$を次式によって求めた．ここで N は動詞 w_3 の種類の数である．

$$H_f(w_3 \mid w_1w_2) = -\left(\sum_{w_3} pf(w_3 \mid w_1w_2) \log pf(w_3 \mid w_1w_2)\right) / \log_2 N$$

次に，GoogleN グラムデータの w_1w_2 を，得られた $H_f(w_3|w_1w_2)$の昇順に並べたうえで 20 区間に分割し，それぞれの区間において本レキシコンの w_1w_2 型表現(計 10,548 件)が含まれる比率を求めた．各区間の含有比率をグラフ化して図 2(a)に，各区間の平均エントロピーを図 2(b)に示す．結果として，本レキシコン w_1w_2 型表現の各区間における含有率は，区間 1 の場合 (1,262/5,542)*100=22.8 %，区間 2 の場合は 22.5%，区間 3 では 20.5%であり，区間 4 以降でも順次低くなっていることが観察された．このことから，本レキシコン収録表現における前部分列 w_1w_2 は，続く動詞 w_3 に関する正規化エントロピー $H_f(w_3|w_1w_2)$が小さいほど，すなわち，後接する動詞部のパープレキシティが小さいものほど多く採録されているという傾向が見られた．本レキシコンにある「墓穴-を（-掘る）」「難色-を（-示す）」「凶弾-に（-倒れる）」などが区間 1(平均エントロピーは 0.27)に含まれていた[5]．エントロピーの大きい表現は解析の曖昧さ低減や予測にあまり有効ではないため，通常の単語単位の処理に任せるのが妥当であると考えており，ほぼ期待された結果である．

基本的には文頭側から文末側へ向かって人の文理解が進むと考えれば，上記の検証は有意味であろうと考えている．また，JMWEL のすべての部分レキシコンにおいて，ほぼ同一基準で表現が採集されているので JMWEL 全体も上記と大差のない傾向を有しているものと考えている．

[5] これらの表現は，いずれも GoogleN グラムデータ上で出現頻度第 1 位であった．

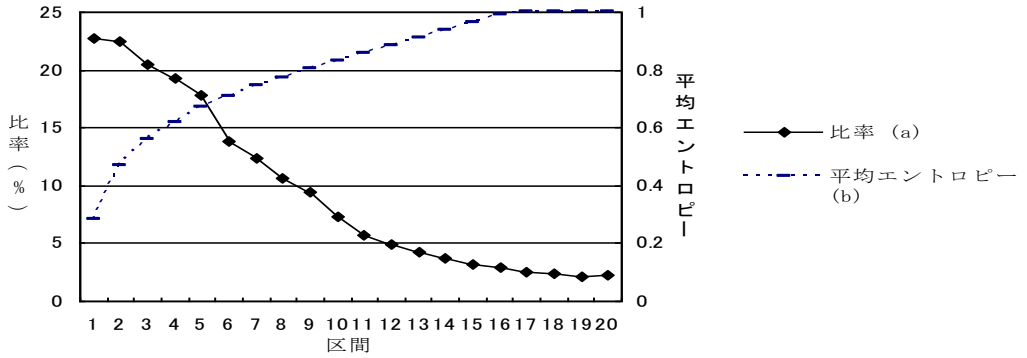


図２　『名詞』＋『格助詞（を，が，に)』 型表現の GoogleN グラムデータにおける後続動詞(正規化) エントロピー区間別採録率(a)，および，各区間の平均エントロピー(b)

４．JMWEL の応用

日本語処理において意味的な纏まりをもつ JMWEL の収録表現を処理単位とすることには大きな利点がある．例えば，JMWEL には，「手に付かず」，「散歩に出る」，「ことにする」という表現が，それぞれ，連用修飾複単語表現，動詞性複単語表現（1 類），文末表現（終助詞，助動詞性表現）として各レキシコンに収録されており，また，それぞれ，連用修飾句 AdvP，動詞句相当表現 VP，助動詞相当表現 Aux であり，構造記述は [[Nni][V12zu]], [[Nni]V30], [[Nni]suru]であることが各レキシコンに記載されている．さらに，「手に付かず」に対しては前方文脈条件としてガ格の後置詞句が必要であることが指定されている．そこで，例えば，入力文「彼は仕事が手に付かず散歩に出ることにした」に対してこれらの情報を用いた一つの構文解析過程のイメージを図 3 に示す．この例のように、MWE を単位的に扱うことで，実質文節数を削減すると同時に文意に近い解析の可能性が高まる[6]．

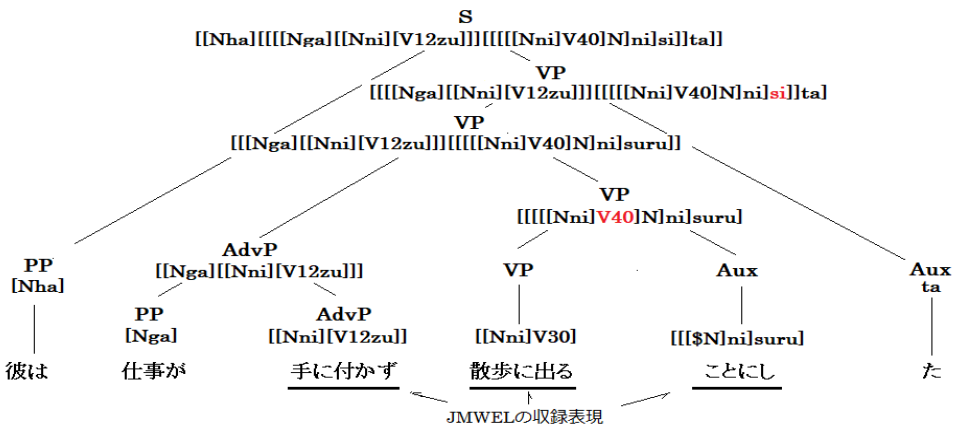


図３　JMWEL による構文解析のイメージ

また，上記のそれぞれの表現に，英訳情報，例えば，“as SUB is unable to get down to doing SUB’s N”，“(to) go out for a walk”，“(to) decide to”を与えたとすると，図 3 の解析にほ

[6] MWE レキシコンを用いる日本語構文解析の手法については，特許第 5379318 号がある．

ぼ平行した形で図 4 のように意味に基づいた日英翻訳が行える可能性が生じる．
その他，日本語の音声認識，日本語による知的対話システム，日本語ワープロの高度化，日本語教育など，JMWEL には広範な応用領域が想定される．

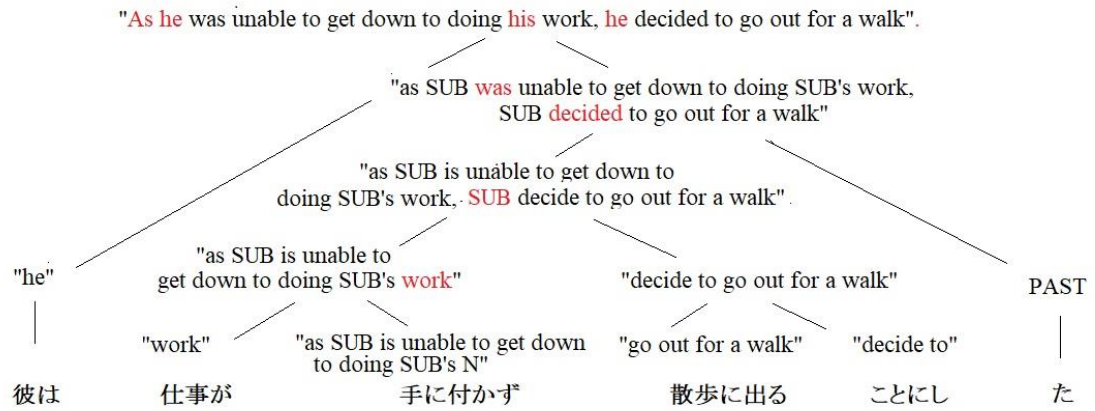


図４　JMWEL による日英機械翻訳のイメージ

５．おわりに

JMWEL は，膨大な日本語の単語 n グラム(2≤n≤18) 集合から，纏まった構文・意味・談話機能を持ち，非構成性，あるいは，要素語間の高い親和性を持つ n グラムだけを掬い取った部分集合の試案である．自然言語の意味処理を試みる際，通常，語類や意味素性による語の共起ルールが作られるが，それでは捉えられない慣用句的表現，決まり文句的表現が思いのほか多く，また，その方法では語の共起度合いの強弱が捉えにくい．本稿で紹介した日本語複単語表現レキシコン JMWEL は．そのような基本認識に基づいて編纂された．
意味の取扱いについては現在なお問題山積であるが，JMWEL は言語表現サイドから改めて意味の問題に切り込むための一次資源として有効ではないかと考えている．機械翻訳のように表層レベルの処理である程度の成果が見込めそうな処理には JMWEL のより直接的な利用が考えられる．
自然言語は言わば言語表現の大海であり，JMWEL の表現収録が十分網羅的であるとは言い切れないが，専門分野，方言を除く書き言葉日本語における特異表現については一つの言語資源プロトタイプとして機能するであろうと考えている[7]．JMWEL がこれからの日本語処理用，日本語研究用の基礎的言語資源として活用され，さらに充実されることを期待したい[8]．

[7] JMWEL は，K. Church の疑問(Church 2011)に対する現時点の一つの回答試案と位置づけられる．
[8] JMWEL の利用については関連サイト http://jefi.info を参照されたい．

文　献

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (eds.) (1999). "Longman Grammar of Spoken and Written English", *Harlow: Pearson Education Limited*.

Kenneth Church (2011). "How Many Multiword Expressions do People Know?", *Proceedings of the MWE workshop(MWE2011)*, ACL, pp.137-144.

Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen Wheatley (eds.) (2009). "Formulaic Language, vol.1, Distribution and historical change", *John Benjamins Publishing Company*.

Charles J. Fillmore, Paul Kay and Mary Catherine O'Connor (1988). "Regularity and Idiomaticity Grammatical Construction: The Case of Let Alone" *Language* 64, pp.501-538.

Nan Jiang and Tatiana M. Nekrasova (2007). "The Processing of Formulaic Sequences by Second Language Speakers", *The Modern Language Journal*, 91:3, pp.433-445.

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

Christopher D. Manning, Hinrich Schűtze (1999). "Foundations of Statistical Natural Language Processing", MIT Press.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger (2002). "Multiword Expressions: A Pain in the Neck for NLP" *Proc. of the 3rd CICLING*, pp.1-15.

Kosho Shudo, Akira Kurahone and Toshifumi Tanabe (2011). "A Comprehensive Dictionary of Multiword Expressions", *Proceedings of the 49th Annual Meeting of the ACL*, pp.169-177.

首藤公昭・田辺利文 (2010). 「日本語の複単語表現辞書：JDMWE」自然言語処理, 17:5, pp.51-74.

高橋雅仁・田辺利文・首藤公昭 (2018). 「日本語複単語表現レキシコン（JMWEL）の概要と現状 －動詞性複単語表現を中心として－」言語処理学会第 24 回年次大会発表論文集, pp.428-431.

Toshifumi Tanabe, Masahito Takahashi and Kosho Shudo (2014). "A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing", *Computer Speech and Language*, 28:6, Elsevier, pp.1317-1339.

田辺利文・高橋雅仁・首藤公昭 (2018). 「日本語動詞性複単語表現(1 類)レキシコンの統計的性質」言語処理学会第 24 回年次大会発表論文集, pp.619-622.

関連 URL

日本語処理研究工房ことばの森　　https://jefi.info

編集後記

今回のワークショップでは、1件の招待講演と64件の一般発表がありました。招待講演をご快諾くださりました慶応義塾大学の吉川正人先生、一般発表に申込をしてくださった方々に感謝いたします。また、優秀発表賞の対象となる発表が14件ありました。その中で、関西学院大学の岡田優也氏の発表「日本語wikipediaを用いた慣用句の構成性の数値化」が受賞しました。複合表現の構成性を単語埋め込みに基づいて数値化するという先進的な研究でした。おめでとうございました。

「コーパスとしてのウェブテキスト活用シンポジウム」を本ワークショップに併設して開催いたしました。本企画は岡照晃さんが中心となり進めていただきました。ご登壇くださりました日本大学の荻野綱男先生、Megagon Labs の林部祐太さん、Insight Tech の三澤賢佑さん、Studio Ousia の山田育矢さんに感謝いたします。

次回は 2019年9月に開催予定です。今回で私は言語資源活用ワークショップの運営から身を引いて、運営を特任助教の石本祐一さん、岡照晃さんに委ねたいと思います。3年間どうもありがとうございました。

国立国語研究所
コーパス開発センター
浅原　正幸