『日本語日常会話コーパス』モニター版の設計・評価・予備的分析

著者	小磯 花絵, 天谷 晴香, 居關 友里子, 臼田 泰如, 柏野 和佳子, 川端 良子, 田中 弥生, 伝 康晴, 西川 賢哉
雑誌名	国立国語研究所論集
号	18
ページ	17-33
発行年	2020-01
URL	http://doi.org/10.15084/00002540

国立国語研究所論集 (NINJAL Research Papers) 18: 17–33(2020)
ISSN: 2186-134X print/2186-1358 online

『日本語日常会話コーパス』モニター版の設計・評価・予備的分析

小磯花絵[a]　天谷晴香[b]　居關友里子[b]　臼田泰如[b]　柏野和佳子[a]
川端良子[b]　田中弥生[b]　伝　康晴[c]　西川賢哉[d]

[a] 国立国語研究所 音声言語研究領域
[b] 国立国語研究所 音声言語研究領域 非常勤研究員
[c] 千葉大学／国立国語研究所 音声言語研究領域 客員教授
[d] 国立国語研究所 コーパス開発センター 非常勤研究員

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，『日本語日常会話コーパス』（CEJC）の構築を進めている。CEJC は，日常会話の多様性を捉え自然な会話行動が観察できるよう，様々な種類の会話をバランスよく収めることを目標に掲げている。2021 年度末に予定している本公開に先立ち，コーパスの利用可能性や問題などを把握するために，目標とする 200 時間のうち 50 時間の会話データについて，2018 年 12 月にモニター公開を開始した。本稿ではまず，コーパスの設計について，会話の収録法，データの公開方針，調査協力者の内訳，コーパスの規模や構成などの観点から概観する。次に，収録されているデータが設計通りバランスがとれているかを，話者と会話の両面から検証する。最後に，コーパスを用いた予備的分析を通して，CEJC モニター版を活用した研究の可能性を示す *。

キーワード：日本語日常会話コーパス，コーパス構築，コーパス評価，日常会話の特徴

1. はじめに

これまで種々の話し言葉コーパスが構築・公開されてきたが，その多くは特定の場面や話者層に偏っており，日常生活の中で私たちがどのような言語行動をとっているかを調査することは難しいという問題があった。そこで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」（小磯 2017）では，さまざまな種類の日常会話 200 時間をバランスよく収録した『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（小磯ほか 2017）。CEJC は，（1）日常場面の中で当事者たち自身の動機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・公開し会話行動を総体的に解明するための研究環境を提供することを目指している。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。そのため，会話をいかに収録するか，それをどのような方針のもとで整備・公開するかなど，検討すべき課題も多く，その取り組みを田中ほか（2018），臼田ほか（2018），小磯・伝（2018）などで報告してきた。

* 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」（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。

200 時間の会話コーパスの本公開は 2021 年度末に予定しているが，コーパスの利用可能性や問題などを把握するために，このうち 50 時間の会話を対象とするモニター公開を 2018 年 12 月 4 日に開始した（以下，CEJC モニター版）。本稿では，CEJC モニター版の設計について概説した上で（2 節），収録されているデータが設計通りバランスがとれているかを，話者と会話の両面から検証する（3 節）。また CEJC モニター版を用いることで，どのような研究の可能性が開けるかを，コーパスを用いた予備的分析を通して具体的に示す（4 節）。

2. コーパスの設計

2.1 会話の収録法

CEJC では，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収集した（田中ほか 2018）。個人密着法は，性別・年齢の点から均衡性を考慮して選別された調査協力者（以下，協力者）に収録機材を 3 ヶ月ほど貸し出し，できるだけ多様な場面，多様な話者との会話を 13 〜 15 時間程度収録してもらうという収録法である。この中から，会話や話者のバランス，データの質や倫理的・法的な問題，話者の希望などを考慮し，コーパスに格納するデータを選別した。

コーパス全体の設計としては，40 名の協力者（男・女× 20 代・30 代・40 代・50 代・60 代以上×各 4 名）を対象とするが，CEJC モニター版では，2018 年 3 月末の時点で収録・第 1 次文字化作業・フォローアップインタビューを終了した協力者の中から，性別・年齢などのバランスを考慮して，公開対象とする協力者を 20 名選んだ。協力者の内訳については 2.3 節で述べる。

個人密着法では次のように会話の収録を行った。映像については，360 度撮影可能な Kodak PIXPRO SP360 4k を会話の場の中央に 1 台配置して話者を中心に撮影すると同時に，GoPro Hero3+ を 2 台設置して話者や会話の状況を俯瞰的に記録した（図 1）。収録の状況等により，1 台あるいは 2 台のカメラでの撮影となることもあった。また散歩などの移動の際には，話者のうち 1 名が Panasonic HX-A500 1 台を装着して収録した。会話音声については，会話の場の中央に置いた IC レコーダーにより会話全体の音声を収録すると同時に，話者ごとの音声をより明瞭に記録するために，各話者が装着した IC レコーダーによって個々人の声を中心に録音した。個人密着法に基づく収録の詳細については田中ほか（2018）を参照されたい。

2.2 収録データの選定と公開方針

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。しかしながら，話し言葉の場合，実際にどのようなレジスター的広がりがあるかを把握すること自体が重要な課題である。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，会話行動調査を実施した。この調査では，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，などをたずねた（小磯ほか 2016）。CEJC では，この調査結果を一つの目安として格納するデータの選定を進めた（小磯ほか 2017）。3.2 節では，この調査結果と比較しながら CEJC モニター

図 1　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，右の二つの映像は GoPro Hero3+ で録画したもの。論文掲載用に話者の顔にボカシの処理を加えている。

版のバランスについて評価する。

本コーパスは，実際の日常場面の会話を映像・音声データまで含めて公開するが，その中には公開の承諾を得ていない第三者の顔やテレビなどの著作物の写り込みなどが多く見られる。そのため，これまでに収録した多様な会話データをもとに具体的な問題を洗い出し，その対応について，肖像権や個人情報保護，著作権などの観点から，知財関連を専門とする弁護士とも相談を重ね，データの公開方針を定めた。モニター版もこの方針に従ってデータを整備した。具体的な方針については小磯・伝（2018）を参照されたい。

2.3 調査協力者

モニター版に含まれる 20 名の協力者に関する属性およびデータの規模（対象とする収録セッション・会話の数[1]，会話時間，語数[2]）を表 1 に示す。性別・年齢（20 代・30 代・40 代・50 代・60 代以上）をバランスさせ，各層 2 名ずつとなるよう選別したが，収録スケジュールの都合から，女性については 40 代が 3 名，60 代以上が 1 名となっている。職業については，性別・年齢のように統制はしていないが，可能な範囲で多様性を持たせるように選んだ。結果，会社員・公務員

[1] 協力者が収録したデータの中には，会話の途中から記録されているものや，会話の途中で終わっているものも少なくない。そのため，協力者が 1 回に収録したもの（これを「収録セッション」と呼ぶ）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た。倫理的・法的な問題や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，一つの収録セッションのデータが複数の会話に分かれることもある。

[2] 協力者以外の話者も含む，会話中の全ての語数（短単位数）。語数を算出するにあたり，固有名などで伏せ字としたもの，語彙等不明で品詞情報が付けられなかったもの，品詞が記号あるいは歌（ハミングなど）のものは除いた。

等 7 名（うち 1 名は会社経営者），自営業･自由業 3 名，パートタイム 2 名，その他（非常勤講師）1 名，学生 4 名，専業主婦・定年退職 3 名となっている。

表 1　協力者 20 名の属性，対象とする収録セッション数（セ数）・会話数・会話時間・語数

年齢	男性					女性				
	職業	セ数	会話数	時間	語数	職業	セ数	会話数	時間	語数
20 代	大学生	5	5	2.2 h	34,216	大学生	7	7	2.6 h	31,645
	大学院生	5	5	2.5 h	33,870	大学生	5	10	2.6 h	23,817
30 代	自営業・自由業	4	4	2.8 h	29,296	会社員・公務員等	5	6	2.7 h	28,526
	会社員・公務員等	6	6	2.1 h	31,239	専業主婦	7	7	2.8 h	35,887
40 代	会社員・公務員等	4	5	2.1 h	23,081	会社員・公務員等	5	5	2.6 h	27,193
	自営業・自由業	6	6	2.4 h	27,523	パートタイム	6	6	2.6 h	33,408
						パートタイム	6	6	2.6 h	31,709
50 代	会社員・公務員等	7	7	2.4 h	26,750	会社員・公務員等	7	7	2.2 h	22,825
	会社員・公務員等	4	4	2.6 h	25,140	自営業・自由業	6	6	2.7 h	32,303
60 代以上	その他	9	9	2.1 h	28,850	専業主婦	6	7	2.7 h	34,728
	定年退職	6	8	3.0 h	47,321					
計		56	59	24.2 h	307,286		60	67	26.1 h	302,041
						総計	116	126	50.3 h	609,327

2.4 コーパスの規模と構成

CEJC モニター版では，（1）50 時間の会話の映像・音声データなどを収めたハードディスクでの公開（ハードディスク版）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」での公開（中納言版）を行っている。それぞれ提供するデータの内訳を表 2 に示す。

表 2　CEJC モニター版が提供するデータの種類

データ種別	ハードディスク	中納言
映像・音声データ	○	×
転記テキスト	○	×
短単位情報	○	○
話者・会話に関するメタ情報	○	△ *
検索システム	○	○

* 備考情報など一部を除く

以下では提供するデータの仕様について概説する。詳細については各項目で挙げる参考文献を

参照されたい。

■映像・音声データ　2.1 節で述べた機器を用いて収録した映像については，図 1 にあるような一つ以上の映像ソースを合成した映像と，個別の映像ソースを公開している。音声についても収録した全ての音源を公開しているが，収録の失敗等により全ての話者の音声が揃わないこともある。また会話全体の音声を記録した音源に問題がある場合には，各話者の音声を合成した音源を提供する。公開している映像・音声データのフォーマットについては小磯ほか（2019）の 2.1 節を参照のこと。

■転記テキスト　転記テキストは，（1）話し手と聞き手が行為や情報を交換する際の基本単位として定義される統語的・談話的・相互行為的なまとまりをもった発話単位[3]（JDRI 2017）と，（2）発話単位を知覚可能なポーズなどによって更に細かく切り，音声との対応を細かく取れるよう設定した転記単位，の 2 種類の単位ごとに区切ったファイルを提供する。転記テキストは，2 種類の単位（発話単位・転記単位）ごとに，CSV 形式，EAF 形式（映像解析ソフトウェア ELAN[4] 用），TextGrid 形式（音声分析ソフトウェア Praat[5] 用）の 3 種類の形式を用意している。発音エラーや非語彙的な母音の延伸などを表現するために，『日本語話し言葉コーパス』や『千葉大学 3 人会話コーパス』の転記の仕様を参考に設計した一連のタグを転記テキストに付与している。転記テキストの詳細については臼田ほか（2018）・小磯ほか（2019）2.2 節を参照のこと。

■短単位情報　モニター版では，長短 2 種類の形態論情報のうち短単位情報を提供する（小椋 2014）。短単位情報は，転記テキストを対象に形態素解析器 MeCab（Kudo et al. 2004）と形態素解析用辞書 UniDic（伝ほか 2007）を用いて自動解析した上で，人手による修正を加えた。品詞体系については『現代日本語書き言葉均衡コーパス』に準拠しているが，CEJC は話し言葉であることから，（1）言いよどみ（「ワ 私」などの語の言いさし），（2）歌（ハミングなどで歌っている箇所），（3）伏せ字（個人情報等のうち，仮名（かめい）ではなく「＊」で伏せ字化した箇所），（4）形態論情報付与対象外（発話内容が全く理解できず語が想定できない場合）の 4 種類を新たに設定した。CEJC モニター版の短単位情報の詳細については，小磯ほか（2019）2.3 節を参照のこと。

■話者・会話に関するメタ情報　モニター版では，（1）会話に関するメタ情報（話者数・会話の形式・場所・活動・話者間の関係性・備考情報）と，（2）話者に関するメタ情報（年齢・性別・職業・協力者から見た関係性・備考情報）を提供する。詳細については小磯ほか（2019）2.4 節を参照のこと。

---

[3]『日本語話し言葉コーパス』で採用した，主節に対する独立性の高い節を基本とする「節単位」を踏襲しつつ，会話に特有な話者交替などの相互行為的観点を加えて定義した単位のこと（丸山 2015）。
[4] https://tla.mpi.nl/tools/tla-tools/elan/
[5] http://www.fon.hum.uva.nl/praat/

■検索システム　同梱する全文検索システム「ひまわり」（山口・田中 2005）では，転記テキストを対象に文字列や単語での検索ができるほか，簡単な集計などを行うことも可能である。また，観察支援システム FishWatchr（Yamaguchi 2018）の機能も統合しており，「ひまわり」で検索した箇所の映像を簡単に閲覧することができる。詳細については小磯ほか（2019）の 2.5 節を参照のこと。

3. バランスの検証

3.1 話者の年齢・性別の観点から

本節では，性別と年齢の観点から CEJC モニター版に含まれる話者のバランスを検証する。モニター版が対象とする 116 の収録セッションに含まれる話者は，延べ 390 名，異なり 237 名である[6]。表 3 に，性別・年齢ごとの話者数・発話時間・語数の情報を示す。発話時間とは，当該話者が実際に発話した時間を転記テキストの情報を利用して算出したものである。

表 3　性別・年齢ごとの話者数・発話時間・語数

年齢	男性				女性				計			
	延べ話者数	異なり話者数	発話時間	語数（千語）	延べ話者数	異なり話者数	発話時間	語数（千語）	延べ話者数	異なり話者数	発話時間	語数（千語）
～ 10 歳	9	4	0.5 h	4.8	3	2	0.2 h	1.8	12	6	0.6 h	6.5
10 代	19	8	1.3 h	19.3	4	4	0.3 h	3.5	23	12	1.6 h	22.8
20 代	31	20	4.0 h	60.3	28	14	3.1 h	41.2	59	34	7.0 h	101.5
30 代	23	15	2.8 h	37.2	37	19	5.1 h	64.4	60	34	7.8 h	101.6
40 代	30	16	3.3 h	44.4	51	31	7.6 h	97.0	81	47	10.8 h	141.4
50 代	25	14	2.3 h	32.1	44	30	5.9 h	81.7	69	44	8.2 h	113.8
60 代	19	11	1.6 h	23.5	24	17	3.2 h	38.5	43	28	4.8 h	62.0
70 代	23	17	2.5 h	36.1	9	7	0.9 h	10.8	32	24	3.4 h	46.9
80 代	4	3	0.2 h	1.9	4	3	0.4 h	4.6	8	6	0.6 h	6.5
90 代	0	0	0 h	0	2	1	0.2 h	2.2	2	1	0.2 h	2.2
不明	0	0	0 h	0	1	1	0.2 h	2.0	1	1	0.2 h	2.0
計	183	108	18.4 h	259.4	207	129	26.9 h	347.7	390	237	45.3 h	607.1

CEJC では，協力者の性別・年齢をバランスさせることにより，多様な世代の話者の会話を収集できるよう設計した。表 3 を見てみると，男性と女性の延べ話者数は 183 名と 207 名でありバランスがとれている。また年齢についても，20 代，30 代，40 代，50 代，60 ～ 70 代についていずれも 60 ～ 80 名程度となっている。ただし，前節で述べた通り，収録スケジュールの都合で 40 代女性の協力者が他より 1 名多く 60 代女性が 1 名少なかったことから，女性の分布を見る限り，40 ～ 50 代が 60 ～ 70 代と比べて多い。本公開のデータでは協力者の年齢・性別を完全にバラン

[6] 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。こうした話者まで含めると，延べ 423 名である。

スをとる予定であり，こうした話者の年齢の偏りは補正されるものと考えられる。

日常会話を扱った他のコーパス・データベースと話者の性別・年齢の分布を比較する（図 2）。ここでは『名大会話コーパス』（藤村ほか 2011）と『談話資料 日常生活のことば』（現代日本語研究会 2016）を取り上げる。

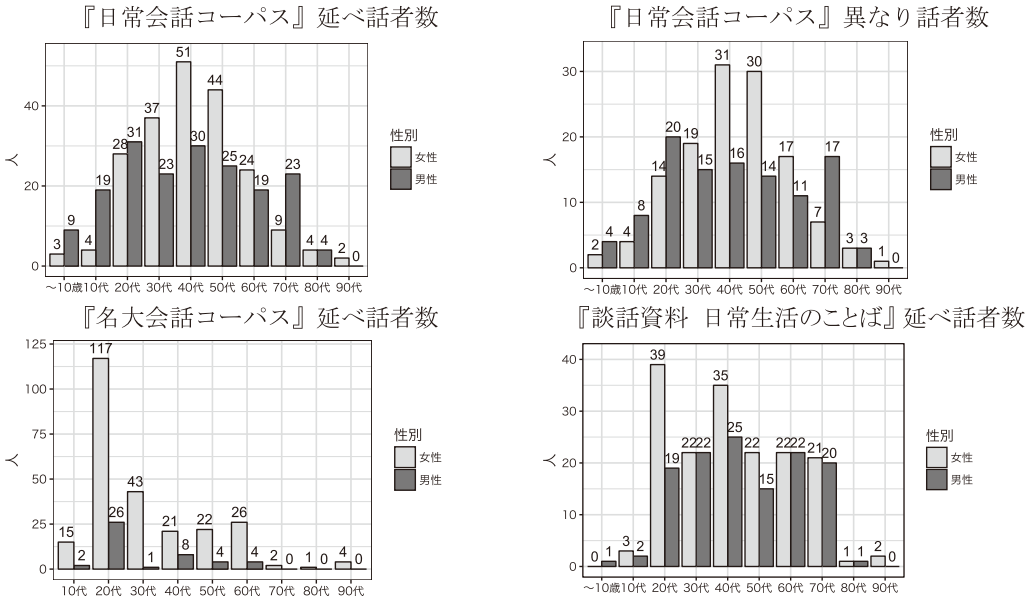

図 2　会話者の性別・年齢の内訳（人）

『名大会話コーパス』は，129 会話，100 時間の雑談を収めた大規模なコーパスである。日本語の雑談を集めたコーパスとしては最も規模の大きなものであり，多くの研究に活用されている。性別・年齢の分布を見てみると，話者 296 名中，85% に当たる 251 名が女性であり，また約半数にあたる 143 名が 20 代であるなど，話者の性別・年齢に大きな偏りが見られる。

『談話資料 日常生活のことば』（以下，談話資料）は，96 会話，約 18 時間の日常生活の会話を集めたデータベースである。『談話資料』には，首都圏に在住，あるいは言語形成期を首都圏で過ごした 20 代から 70 代の各世代の男女 2 ～ 3 名，および 20 代の学生男女各 2 名，計 31 名の協力者に収録してもらった，日常生活の中で生じる 3 場面（3 名のみ 4 場面）の会話が収められている。話者の性別・年齢を見てみると，20 代から 70 代にかけて男女ともバランスよく分布していることが分かる。収録の方法が CEJC に非常に類似しているが，協力者の性別･年齢を統制し，またできるだけ異なる場面を収録してもらうことによって，収録される話者の性別・年齢をある程度バランスさせることが可能となることが分かる。

一方, いずれのコーパスにおいても不足しているのが未成年者である。CEJC も『談話資料』も，協力者は 20 代以上の成人に限定している。そのため，未成年者の収録数は必然的に少なくなる。また未成年者は，話者数だけでなく発話時間や語数の少なさも目立つ。会話全体の時間に対して

実際に発話している時間の割合を調べると，10 歳未満と 10 代の未成年者はいずれも成人の半分程度であり，単位時間あたりの発話量が少ない。本公開のコーパスでは，こうした偏りを補正するために，特定場面法と呼ばれる別の収録法[7]で未成年者の発話を補填する予定である。

3.2 会話の形式・話者数・活動の観点から

2.2 節で述べたように，CEJC では，小磯ほか（2016）で報告した会話行動調査の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。そこでモニター公開データを対象に，会話の形式，会話の話者数，活動の内訳を求め，会話行動調査の結果（以下「調査」）と比較することで，CEJC モニター版に含まれる会話のバランスを検証する。会話の形式と話者数については 3.1 節で取り上げた『談話資料』でも情報が提供されていることから，併せて比較する。

■ 会話の形式　会話の形式（雑談／用談相談／会議会合）に関する結果を図 3 に示す。左は会話の件数で見た場合の割合，右は会話の時間で見た場合の割合である。結果から，件数・時間ともに CEJC では雑談が約 7 割を占めており，「調査」より若干多いものの概ねバランスがとれていることが分かる。会議会合は件数で見ると「調査」より多いが，時間で見ると少ない傾向が見られる。CEJC では多様性を確保するためにコーパスに含める会話を最大 1 時間としているのに対し，実際の会議会合は 1 時間を越える長いものが多いことが影響している。談話資料（図では談話）は，件数・時間いずれも雑談が 9 割近くを占めているが，用談相談（用談）や会議会合（会議）も 5% 強ずつ含まれており，雑談だけでなく用談や会議なども含めるよう設計したことが分かる。

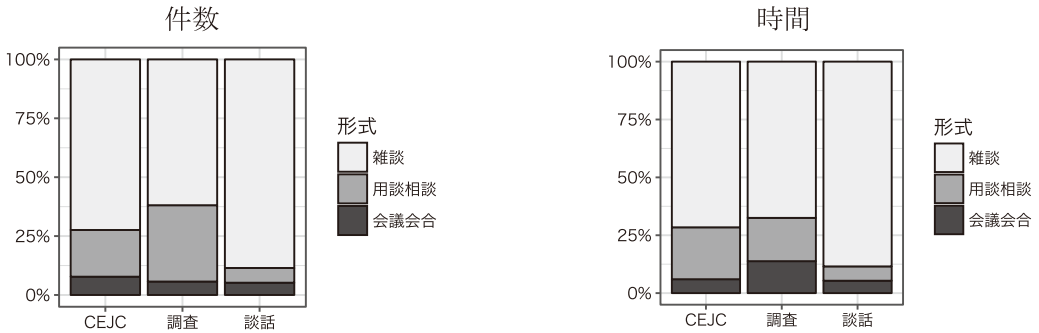

図 3　会話の形式：CEJC モニター公開データ・会話行動調査・談話資料の比較

■ 話者数　会話を構成する話者の数は会話の構造や展開に影響を与えるため，会話の多様性を確保するには話者数のバランスも視野に入れておく必要がある。話者数に関する結果を図 4 に示す。話者数が 2 人の場合，件数で見ると CEJC の方が「調査」より少ないが，時間で見ると「調

[7] 未成年者同士の会話や職場での会議など，2.1 節で記した個人密着法に基づく収録法では収集が難しい場面の会話を，調査者が主体となり調整して収録する方法。調査者は介在するが，日常場面で自然に生じる会話を対象とする。

査」よりやや多い傾向が見られる。一方，5 人以上の会話については逆の傾向が見られる。このように件数と時間との間で若干の偏りはあるものの[8]，人数に関してもある程度バランスよくデータが集められていることが分かる。談話資料は，2 人の会話が 65% 強と多めだが，3 人以上の会話も含まれており，人数の多様性も確保されたデータとなっている。

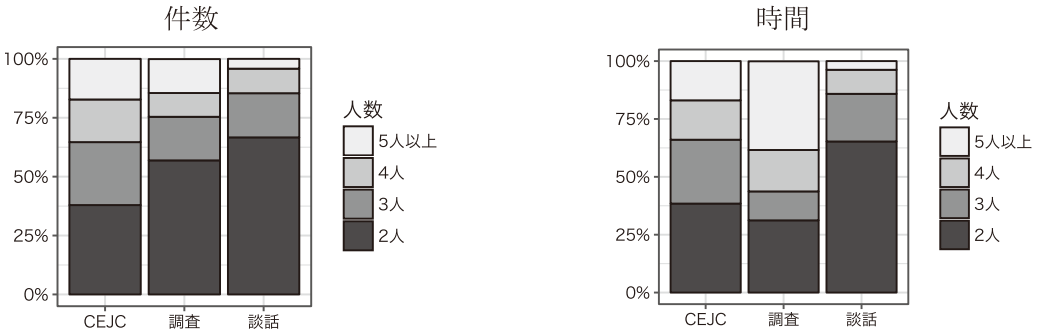

図 4　話者数：CEJC モニター公開データ・会話行動調査・談話資料の比較

■活動　活動に関する結果を図 5 に示す。CEJC モニター版では，自宅での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，「調査」と比べると，家事雑事・仕事・学業中の会話がかなり少なく，友人との付き合いといった私的活動が多い傾向が見られる。個人密着法では協力者が主体となり会話を収録することから，職場や学校などでの仕事・学業中の会話の収録は難しく，家族との会話を除くと，公共商業施設での友人との私的活動が必然的に多くなる。不足する種類の会話については，今後，特定場面法で補填する。

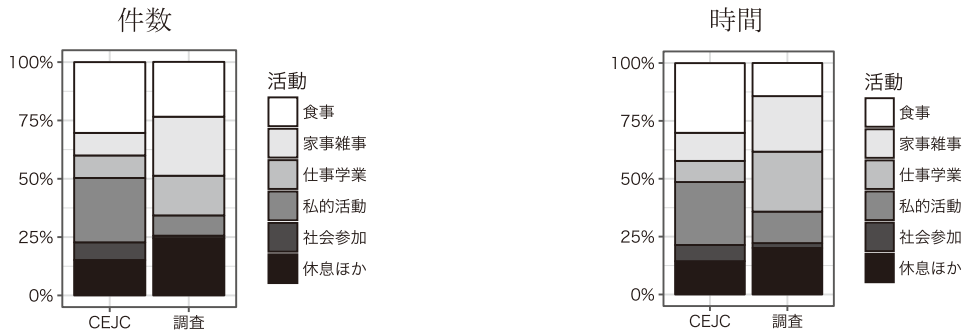

図 5　活動：CEJC モニター公開データと会話行動調査の比較

[8] 会話行動調査から，話者数が多くなるほど会話時間は長くなる傾向にあることが，またごく短い会話は 2 人会話に多く見られることが分かっている（小磯ほか 2016）。CEJC ではデータを上限 1 時間に設定していること，また協力者に短い会話をわざわざ収録してもらうことは難しいことから，件数と時間の間に必然的に差が生じることになる。

4.『日本語日常会話コーパス』モニター版を用いた研究の可能性

前節では，話者の性別や年齢，会話の形式などの観点から，CEJC モニター版に含まれる話者や会話が比較的バランスよく分布していることを示した。本節では，本コーパスを用いることでどのような研究の可能性が開けるかを，コーパスを用いた予備的分析を通して見ていく。

4.1 並列節を導く接続助詞「けれども」類・「が」の出現傾向

並列節を導く接続助詞の「けれども」には,「けれど」「けども」「けど」などの表現のバリエーションがある。これらの表現を本稿では「けれども」類と称す。丸山（2014a）は，現代日本語の多様なレジスターの書き言葉をバランスよく収録した『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）と，講演などの独話を中心とする『日本語話し言葉コーパス』（以下 CSJ）を用い，「けれども」類の分布を調べたところ，BCCWJ に含まれる Yahoo! 知恵袋や Yahoo! ブログなどのくだけた文体で書かれた書き言葉では「けど」の使用が極めて多いこと，話し言葉では改まったスタイルからくだけたスタイルに移行するにつれ「けれども」の使用が少なくなることを指摘している。しかし丸山が分析した当時，国語研究所が提供するコーパスの中に日常会話を含むものはなかったことから，日常会話でどのような分布を示すかは明らかになっていない。そこで本節では，BCCWJ，CSJ に CEJC モニター版を加え，「けれども」類の各表現の分布を比較する。

CSJ については，学会発表などを中心とする「学会講演」と，一般の話者による個人的な体験談などを集めた「模擬講演」を分析に用いる。また BCCWJ からは，「行政白書」「新聞」「雑誌」「Yahoo! ブログ」「国会会議録」を取り上げる。これらのサブカテゴリーをここではレジスターと呼ぶ。「けれども」類に，同じく並列節を導く接続助詞「が」を加え，レジスターごとに分布を求めた。結果を表 4 と図 6 に示す[9]。

表 4　レジスターごとに見た接続助詞「けれども」類と「が」の調整頻度（100 万語あたり）と割合

レジスター	が	けれども	けれど	けども	けど
白書	18015（100%）	2（0.0%）	1（0.0%）	0（0%）	0（0%）
新聞	27975（94.1%）	73（0.2%）	402（1.4%）	0（0%）	1292（4.3%）
雑誌	27483（76.5%）	321（0.9%）	1492（4.2%）	18（0.1%）	6620（18.4%）
ブログ	45552（65.4%）	255（0.4%）	1777（2.6%）	304（0.4%）	21753（31.2%）
国会	49437（59.5%）	33402（40.2%）	90（0.1%）	47（0.1%）	102（0.1%）
学会	51226（51.9%）	25278（25.6%）	880（0.9%）	14662（14.9%）	6606（6.7%）
模擬	35697（27.3%）	35846（27.4%）	4460（3.4%）	20553（15.7%）	34113（26.1%）
会話	1186（1.8%）	774（1.2%）	214（0.3%）	1252（1.9%）	62609（94.8%）

結果を見る前に，ここで取り上げたレジスターのスタイルについて言及しておく。CSJ に付与されている印象評定データの結果から，学会講演よりも模擬講演の方がくだけた発話スタイルで

[9] 丸山（2014a）の分析では BCCWJ，CSJ ともにコアと呼ばれるデータセットを用いているのに対し，本稿ではコア以外のデータも含めて分析したことから，若干値が異なる。

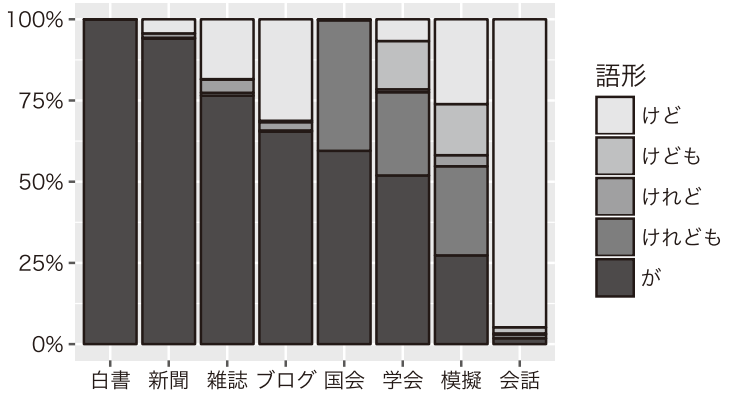

図 6　レジスターごとに見た並列節を導く接続助詞「けれども」類と「が」の分布

あることが指摘されている（籠宮ほか 2007）。また BCCWJ を用いた分析から，新聞よりも白書の方がより改まったスタイルであることも分かっている（小磯ほか 2008）。新聞には一般の記事だけでなくコラムなども含まれることが影響しているためである。また三宅（2005）などにより，インターネット上の言葉は話し言葉に近いことも指摘されている。こうしたことを念頭に置いて結果を見てみよう。

まず「が」と「けれども」類の割合に着目して結果を見る。図 6 から，全体的に書き言葉では話し言葉よりも「が」の使用が多いことが分かる。書き言葉の中を見ると，改まったスタイルで書かれる傾向の強い白書では「けれども」類はほとんど見られず「が」が圧倒的に用いられているのに対し，新聞，雑誌，ブログの順に「が」が少なくなり「けれども」類が増える。一方，話し言葉では，国会会議録，学会講演，模擬講演の順に「が」が少なくなり，最もくだけたスタイルと考えられる日常会話では「が」の使用はほとんど見られない。このように，改まったスタイルからくだけたスタイルになるほど「が」が減り「けれども」類が増える傾向が見られる。丸山（2014b）は，学会講演では模擬講演よりも「が」が多く，逆に「けれども」類は少ないことから，改まったスタイルでは「が」がより好まれることを指摘している。今回の分析から，丸山の指摘する傾向が，白書のような改まったスタイルの書き言葉から日常会話のようなくだけた話し言葉までの幅広いレジスターにおいて明瞭に観察されることが分かる。

次に「けれども」類の内訳を見る。丸山（2014a）は話し言葉では改まったスタイルからくだけたスタイルに移行するにつれ「けれども」の使用が少なくなることを指摘しているが，日常会話では指摘の通り「けれども」はわずか 1% であり「けど」が 95% と圧倒的に多いことが分かる。

「が」の結果と合わせると，(1) 話し言葉では，国会などのかなり改まった場では「が」や「けれども」が中心でそれ以外の形はほとんど現れない，(2) くだけたスタイルになるほど「が」や「けれども」が少なくなり，「けど」が多くなる，(3) 日常場面のようにかなりくだけた場では「けど」が中心でそれ以外の形はほとんど現れない，とまとめることができる。

4.2 副詞「やはり」の語形の出現傾向

副詞「やはり」には,「やはり」の他に, 話し言葉で多く用いられるとされる「やっぱり」や「やっぱし」,「やっぱ」などの語形のバリエーションがある。田中（2004）は，CSJ の紹介の中で，学会講演では「やはり」が，模擬講演では「やっぱり」が多いことを指摘し，両講演の違いを改まり度の高低の尺度とするならば, CSJ を用いることで「形態間の文体的特徴のレベルを序列化し，相互関係を計測することができる」（p. 80–81）としている。しかし前節でも見たように CSJ の講演は日常会話と比べると発話の改まり度は高いため,「やっぱり」から転じた「やっぱし」や，最もくだけた語形と考えられる「り」の脱落した「やっぱ」はほとんど出現していない。スタイルの影響を見るには，レジスターを幅広く設定する必要がある。そこで，前節の分析と同じように，CSJ に BCCWJ と CEJC モニター版を加え,「やはり」類の語形の選択とレジスターとの関係について見てみる。結果を表 5 と図 7 に示す。

表 5　レジスターごとに見た副詞「やはり」類の調整頻度（100 万語あたり）と割合

レジスター	やはり		やっぱり		やっぱし		やっぱ	
白書	8	（100.0%）	0	（0%）	0	（0%）	0	（0%）
新聞	53	（72.0%）	20	（27.0%）	0	（0%）	1	（1.0%）
雑誌	128	（50.0%）	116	（45.2%）	1	（0.2%）	12	（4.6%）
ブログ	210	（34.3%）	312	（50.9%）	4	（0.6%）	88	（14.3%）
国会	986	（82.8%）	204	（17.1%）	0	（0%）	0	（0%）
学会	374	（63.8%）	197	（33.6%）	0	（0%）	15	（2.6%）
模擬	918	（31.1%）	1705	（57.8%）	35	（1.2%）	290	（9.8%）
会話	0	（0%）	865	（54.2%）	15	（0.9%）	717	（44.9%）

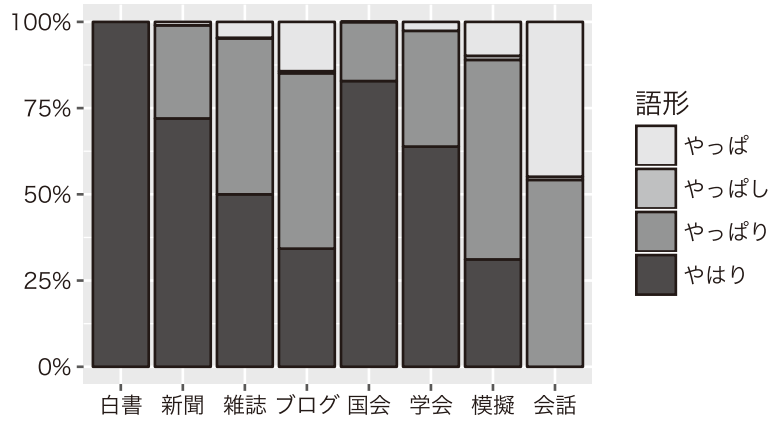

図 7　レジスターごとに見た副詞「やはり」の分布

書き言葉を見ると，最も改まり度の高い白書では「やはり」しか見られないのに対し，くだけたスタイルになるにつれ「やはり」が減少し,「やっぱり」を中心に「やっぱ」も含めて増える

傾向が見られる。話し言葉でも同様に，改まり度の最も高い国会で「やはり」が多用され，徐々に「やはり」は減少し，「やっぱり」と「やっぱ」が増える。特に日常会話では「やはり」は一切見られず，「やっぱり」と「やっぱ」がほぼ半々となる。このように，書き言葉，話し言葉ともに，「やはり」類の語形の選択にスタイルの影響が強く見られることが分かる。

「やはり」類の語形分布の変動は日常会話の中でも見られる。CEJC に限定した上で，「やはり」類の語形の選択と話者の年齢との関係を見てみよう。10 歳未満で「やはり」類を用いた話者は異なりで 2 名のみであったため対象外とした。結果を図 8 に示す。図から，若い人ほどくだけた語形である「やっぱ」をより多く用いていることが分かる。このように「やはり」類の語形の選択には話者の年齢という要因が大きく関わる [10]。

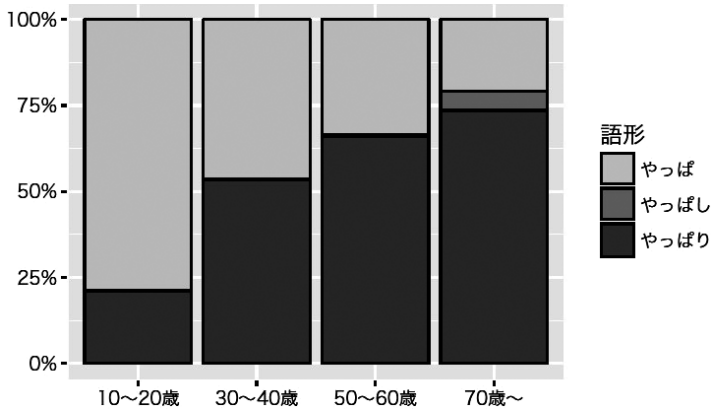

図 8　年齢ごとに見た副詞「やはり」類の分布

4.3 感謝表現「ありがとう」類の出現傾向

前節では「やはり」類の語形の選択に話者の年齢が関わることを見た。本節では，挨拶表現「ありがとう」の幾つかの表現を取り上げ，話者の年齢や性別，会話の形式，年齢上の上下関係，および話し手から見た聞き手の関係性の観点から，「ありがとう」の表現との関係を概観する。「ありがとう」類を，「ございます」が後続する「ありがとうございます」と，後続しない「ありがとう」に分けた。またそれ以外の形として，「あざーす」や「あざっす」「あざます」のような，かなりくだけた表現をまとめて「あざす」系とした。この三つの表現がどのように分布するかを見ていく。なお，上下関係と聞き手の関係性については，提供されるメタ情報から一意に特定できる場合に限定して分析した。結果をまとめて図 9 に示す。

図 9 の「年齢」および「性別」から，いずれの年齢層においても，またいずれの性別において

[10] BCCWJ，CSJ も含め，「やっぱし」は全体的に出現数が少ないが，図から CEJC では 70 歳以上にその使用が集中している傾向が読みとれる。しかしこれは，一人の話者が「やっぱし」を多用し，この世代の「やはり」類の約半数を占めているためである。「やっぱし」を用いた話者の異なりも少ないため，「やっぱし」の使用については，今後データが蓄積された段階で改めて検討したい。

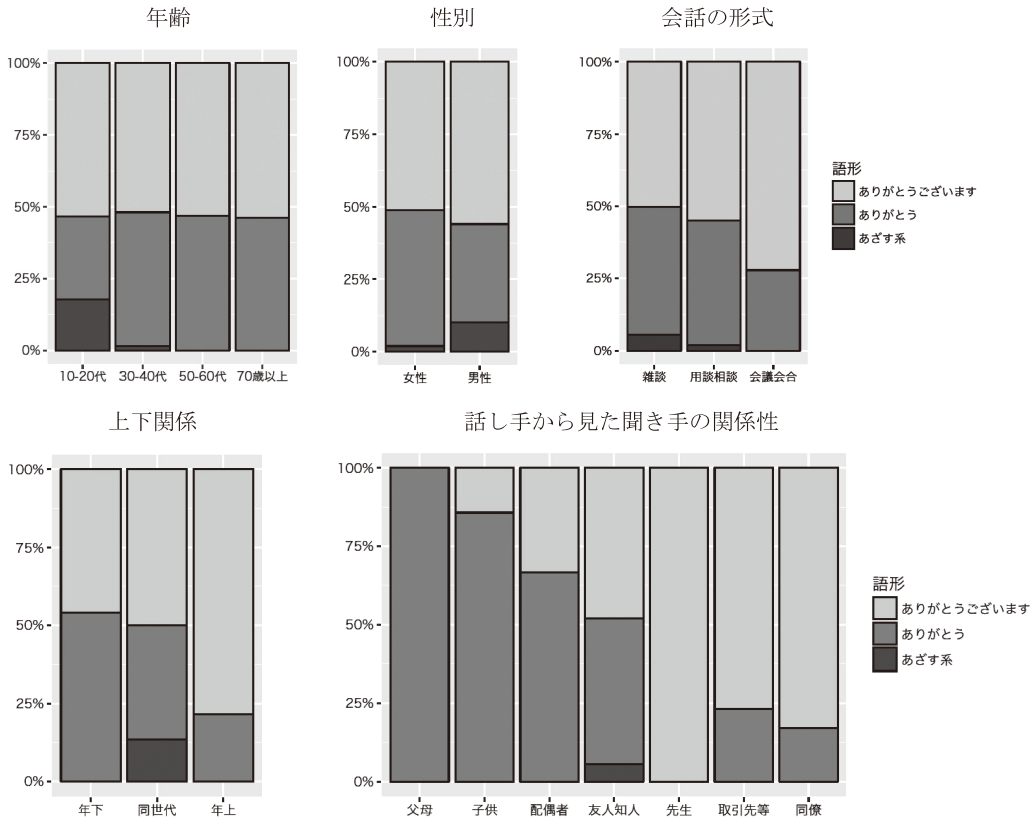

図 9　年齢・性別・会話形式・上下関係・聞き手の関係性ごとに見た「ありがとう」類の分布

も，「ありがとうございます」が全体の半数を占めていることが分かる。違いが見られるのは残る約 50% の「ありがとう」と「あざす」系の内訳である。図から，「あざす」系を用いているのは主に 10 〜 20 代の男性に強く偏っていることが分かる。

また図 9 の「会話の形式」「上下関係」「話し手から見た聞き手の関係性」から，これらの要因が「ありがとうございます」の使用に影響していることが分かる。図から，雑談，用談相談，会議会合と，場の改まり度が高くなるにつれ，「ありがとうございます」の割合が増える傾向が見られる。また相手が年上の場合に「ありがとうございます」を高い割合で用いている。更に聞き手が父母や子供，配偶者など，家族である場合には「ありがとう」が主で「ありがとうございます」は少ないのに対し，先生や取引先，同僚など仕事や学業の場面では「ありがとうございます」が主となる傾向が見られる。友人知人はその中間だが，「あざす」系を用いているのは同世代の友人知人との雑談時であり，家庭での家族との会話や仕事・学業での先生や同僚との会話では見られない。

さほど目新しい結果ではないが，CEJC モニター版が多様な話者・場面の会話を記録しているからこそ，こうした傾向を定量的に示すことが可能となる点は重要であろう。

5. おわりに

本稿では，CEJC モニター版の概要について説明した上で，特に話者の性別や年齢，会話の形式などの観点から，CEJC モニター版に含まれる話者や会話が比較的バランスよく分布していることを示した。その上で，本コーパスを用いることで，どのような研究の可能性が開けるかを，コーパスを用いた予備的分析を通して示した。「日常会話」プロジェクトを開始した時点では，国立国語研究所コーパス開発センターが提供するコーパスは書き言葉に偏っており，話し言葉については独話を主対象とする CSJ のみであった。今回，日常会話を対象とする CEJC をモニター公開することによって，書き言葉・話し言葉を含む多様なレジスターを対象に，言葉の使用傾向を多角的に捉えることができることを，並列節を導く接続助詞「けれども」類・「が」および副詞「やはり」類の分析を通して示した。また CEJC が多様な話者・多様な会話を収録していることによって，話者の年齢や性別，会話の形式，年齢上の上下関係，聞き手の関係性などが言葉の選択に与える影響の分析が可能となることを，感謝表現「ありがとう」類の分析を通して示した。

2021 年度末に 200 時間の会話を対象とする本公開を予定しているが，データ量が 4 倍になることで，分析の可能性が更に広がることが期待される。

参照文献

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: 101–123.

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』43–72. 東京：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』東京：ひつじ書房.

JDRI（2017）『発話単位ラベリングマニュアル』http://www.jdri.org/resources/manuals/uu-doc-2.1.pdf

籠宮隆之・山住賢司・槙洋一・前川喜久雄（2007）「聴取実験に基づく講演音声の印象評定データの構築とその分析」『社会言語科学』9(2): 65–76.

小磯花絵・小木曽智信・小椋秀樹・冨士池優美・宮内佐夜香（2008）「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』192–195.

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: 85–106.

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』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 回年次大会発表論文集』775–778.

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: 75–89.

小磯花絵・天谷晴香・石本祐一・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（2019）『『日本語日常会話コーパス』モニター公開版　コーパスの設計と特徴』プロジェクト報告書 3. https://www2.ninjal.ac.jp/conversation/report/report03.pdf

Kudo, Taku, 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*, 230–237.

丸山岳彦（2014a）「現代日本語の連用節とモダリティ形式の分布―BCCWJ に基づく分析―」益岡隆志・大島資生・橋本修・堀江薫・前田直子・丸山岳彦（編）『日本語複文構文の研究』399–425. 東京：ひつじ書房.

丸山岳彦（2014b）「現代日本語の多重的な節連鎖構造について―CSJ と BCCWJ を用いた分析」石黒圭・橋

本行洋（編）『話し言葉と書き言葉の接点』93–114. 東京：ひつじ書房.
丸山岳彦（2015）「発話の単位」小磯花絵（編）『話し言葉コーパス　設計と構築』54–80. 東京：朝倉書店.
三宅和子（2005）「携帯メールの話しことばと書きことば―電子メディア時代のヴィジュアル・コミュニケーション」三宅和子・岡本能里子・佐藤彰（編）『特集：組み込まれるオーディエンス』234–261. 東京：ひつじ書房.
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』第 4 章, 68–88.
田中牧郎（2004）「新刊・寸刊『日本語話し言葉コーパス』」『日本語学』7 月号 : 80–81.
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: 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: 177–193.
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): 55–77.
Yamaguchi, Masaya (2018) A video annotation system for learners to observe educational activities in Motoko Ueyama. In: Irena Srdanović (ed.) *Digital resources for learning Japanese*, Bononia University Press.

関連 **Web** サイト

『日本語日常会話コーパス』モニター公開のウェブサイト
https://www2.ninjal.ac.jp/conversation/cejc-monitor.html（2019 年 8 月 6 日確認）
『大規模日常会話コーパスに基づく話し言葉の多角的研究』プロジェクトのウェブサイト
https://www2.ninjal.ac.jp/conversation/（2019 年 8 月 6 日確認）

Design, Evaluation, and Preliminary Analysis of the Monitor Version of the *Corpus of Everyday Japanese Conversation*

KOISO Hanae[a]　AMATANI Haruka[b]　ISEKI Yuriko[b]
USUDA Yasuyuki[b]　KASHINO Wakako[a]　KAWABATA Yoshiko[b]
TANAKA Yayoi[b]　DEN Yasuharu[c]　NISHIKAWA Ken'ya[d]

[a]Spoken Language Division, Research Department, NINJAL
[b]Adjunct Researcher, Spoken Language Division, Research Department, NINJAL
[c]Chiba University / Invited Professor, Spoken Language Division, Research Department, NINJAL
[d]Adjunct Researcher, Center for Corpus Development, NINJAL

Abstract
We have been constructing the *Corpus of Everyday Japanese Conversation* (CEJC) under the NINJAL collaborative research project since 2016. The CEJC is designed to contain various kinds of everyday conversations in a balanced manner to capture the diversity of everyday conversations and to observe natural conversational behavior. Prior to the publication of the whole corpus, which scheduled for 2022, we published the monitor version of the CEJC in December 2018. In this paper, we first outlined the design of the monitor version of the CEJC, including recording methods, the release policy of the corpus, corpus size, and annotations. Then, we examined whether the speakers and the conversations in the corpus vary in a balanced manner. Finally, we conducted a preliminary analysis on some linguistic aspects of the monitor version of the CEJC, revealing the possible implications of the corpus.

Key words: *Corpus of Everyday Japanese Conversation*, corpus construction, corpus evaluation, characteristics of everyday conversation