『日本語日常会話コーパス』モニター公開版の構築

著者	小磯 花絵
雑誌名	計量国語学
巻	32
号	2
ページ	133-142
発行年	2019-06
URL	http://id.nii.ac.jp/1328/00003055/

doi: 10.24701/mathling.32.2_133

計量国語学 32 巻 2 号（2019 年 9 月） pp.133-142.

研究資料

『日本語日常会話コーパス』モニター公開版の構築

小磯 花絵（国立国語研究所）

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，自宅での家族との会話や飲食店での友人との会話，職場での同僚との会合，散策時の会話など，日常生活における多様な場面の会話を，映像まで含めて収録・公開するものであり，世界的に見ても極めて新しい試みである．最終的には 200 時間規模のコーパスとして 2021 年度末に公開する予定であるが，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 時間のデータについて 2018 年 12 月にモニター公開を開始した．本稿では CEJC モニター公開版の設計・構成やそれを用いた研究の可能性について概説する．

キーワード：会話コーパス，コーパス設計，アノテーション

1．はじめに

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」（2016～2021 年度）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』（*Corpus of Everyday Japanese Conversation*, CEJC）の構築を進めている．

プロジェクトの開始当時，国立国語研究所では，多様なレジスターの書き言葉をバランスよく集めた『現代日本語書き言葉均衡コーパス』や，講演を中心とする『日本語話し言葉コーパス』など，大規模なコーパスを構築・公開しており，オンライン検索システム「中納言」で比較的容易に研究に利用できる環境が整っていた．しかし日常会話を対象とするコーパスは存在しておらず，日常会話の言葉を書き言葉や講演などと比較し研究できる環境作りが望まれていた．

研究所の外に目を向けた場合，日本語の会話を対象とするコーパスはある程度公開されていたが（表 1），若者や親近者同士の雑談，電話会話，職場会話といったように，話者や会話形式，場面などに偏りが見られた．また収録のためにわざわざ集まって雑談してもらったものなど，実際の日常生活で交わされる会話ではなく作られた場面での会話も少なくなかった．そのような中で『談話資料日常生活のことば』（現代日本語研究会ほか 2016）は日常会話を対象とする数少ないデータベースであるが，残念ながらテキストのみの公開で音声を聞くことができない．音声が提供されないという問題は他のコーパスにも見られ

る．100 時間という大量の雑談を対象とする『名大会話コーパス』（藤村ほか 2011）はこれまで多くの研究で利用されてきたが，残念ながら音声にアクセスできないため，文字化されたテキストに基づく分析に限定されてしまう．また全体的に見て映像まで提供するコーパスはごく一部に限られる．

表 1：主要な日本語の会話コーパス

コーパス名	規模	概要	メディア
名大会話コーパス	129 会話 100 時間	親しい者同士の雑談	無
BTSJ 日本語自然会話コーパス	333 会話 79 時間	友人同士の雑談, 教師学生面談会話，電話会話など	音声（一部）
Sakura コーパス	18 会話 7.5 時間	大学生の会話	映像
千葉大学 3 人会話コーパス	12 会話 2 時間	大学生の友人同士の会話	音声
CALL HOME Japanese	120 会話 20 時間	アメリカ在住日本人と国内の家族・友人との電話会話	音声
女性のことば 職場編 男性のことば 職場編	49 会話 9 時間 62 会話 12 時間	職場のフォーマル・インフォーマルな場面の自然談話	無
談話資料 日常会話のことば	96 会話 18 時間	日常生活の会話	無

そこで本プロジェクトでは，1）日常場面で自然に生じる会話を対象とすること，2）多様な話者による多様な場面の会話をバランスよく集めること，3）音声・映像を含めて公開し，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，を目標に，2016 年度より 200 時間規模の日常会話コーパスの構築に着手した．コーパスの本公開は 2021 年度末の予定だが，コーパスの利用可能性を把握し今後の構築に活かすために，2018 年 12 月に 50 時間分の会話を対象とするモニター公開を開始した（以下 CEJC モニター公開版[1]）．本稿では CEJC モニター公開版の設計・構成やそれを用いた研究の可能性について概説する．

2．『日本語日常会話コーパス』モニター公開版の概要

2.1　コーパス全体の設計

モニター公開版は本公開版のサブセットであるため，はじめに CEJC 全体の設計について概観する（小磯ほか 2017）．

多様な話者による多様な場面の会話をバランスよく集めるために，年齢と性別の観点からバランスをとった 40 名の協力者（男女 × 20 代・30 代・40 代・50 代・60 代以上 × 各 4 名）に収録を依頼し，できるだけ多様な場面の会話を収録してもらうという方法を採用した．自然に生じる会話を対象とするため，研究者は収録に立ち会わない．一人あたり 15 時間ほど収録してもらい，その中から会話の種類や話者のバランスなどを考慮して 4〜6 時間程度の会話を公開データとして選別する．また未成年者の会話や職場での会議などこの手法では収録が難しい種類の会話については，これとは別の手法で収録し，コーパス全体としてできるだけ多様な話者・会話をバランスよく含むようにする計画である．

[1] https://www2.ninjal.ac.jp/conversation/cejc-monitor.html

図 1：会話の映像データの例（論文掲載用に話者の顔にボカシ処理をしている）

図 1 に，協力者が実際に収録した会話の映像の例を示す．図にあるように，協力者は原則として 3 台のカメラを用いて会話の映像を記録する．また音声についても，会話の場の中心に設置する IC レコーダーで会話全体の音声を録音すると同時に，個々の話者が装着する IC レコーダーで各話者の音声をより明瞭に録音する．

収録した音声に基づき転記テキスト（臼田ほか 2018）を人手で作成した上で，短単位情報・長単位情報（小椋 2014）を自動で付与し，短単位情報については全て人手で修正する．また全体 200 時間のうち 20 時間については，高精度かつ多様なアノテーションを付与するデータセットと位置付け，長単位情報を人手で修正すると同時に，文節間の係り受け情報や談話行為情報（居關ほか 2017），韻律情報（五十嵐 2015）を新たに人手で付与する．こうしたアノテーションを，200 時間の会話の映像・音声データ，転記テキスト，会話や話者に関するメタ情報と合わせ，2021 年度末に公開する予定である．

2.2　モニター公開版の公開方式・公開データの種別

200 時間から構成される CEJC のうち 50 時間の会話を対象に，2018 年 12 月にモニター公開を開始した（小磯ほか 2019）．CEJC モニター公開版では，(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 モニター公開版が提供するデータの種類

データ種別	ハードディスク版	中納言版
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	○	○

2.3　協力者の構成

表 3 にモニター公開対象とする協力者 20 名の情報を示す．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各層 2 名ずつとなっている．職業についても，会社員・公務員等 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，できるだけ多様性を持たせている．

表 3：協力者の属性，対象とする会話数と会話時間

年代	男性			女性		
	職業	会話数	時間	職業	会話数	時間
20 代	大学生	5	2.2h	大学生	7	2.6h
	大学院生	5	2.5h	大学生	10	2.6h
30 代	自営業・自由業	4	2.8h	会社員・公務員等	6	2.7h
	会社員・公務員等	6	2.1h	専業主婦	7	2.8h
40 代	会社員・公務員等	5	2.1h	会社員・公務員等	5	2.6h
	自営業・自由業	6	2.4h	パートイタイム	6	2.6h
				パートイタイム	6	2.6h
50 代	会社員・公務員等	7	2.4h	会社員・公務員等	7	2.2h
	会社員・公務員等	4	2.6h	自営業・自由業	6	2.7h
60 代以上	その他	9	2.1h	専業主婦	7	2.7h
	定年退職	8	3.0h			

2.4　会話の内訳

本節では，会話の形式，会話中の活動，会話の話者数の観点から，CEJC モニター公開版に含まれる会話の内訳（会話数）を示す．

それぞれの内訳を，予備調査として実施した会話行動調査の結果と合わせて図 2 に示す．この行動調査は，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て CEJC の設計に活かすために実施したものである．成人約 250 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，などをたずねている（小磯ほか 2016）．

図 2 左に示す会話の形式の内訳から，CEJC モニター公開版には，雑談だけでなく，用談相談や会議会合も少なからず含まれていることが分かる．ただし，協力者によっては収録・公開の許諾が得られる範囲が雑談に偏ってしまうこともあり，行動調査では雑談が 60%であるのに対して CEJC モニター公開版では 72%と，やや雑談が多い．

図 2 中央に示す会話中に行っている活動の内訳を見ると，食事の場面や友達とのつきあいといった私的活動が多いものの，収録・公開の許諾が得られにくい中で，料理や家具組み立てといった家事雑事の場面や取引先との打合せといった仕事の場面なども，ある一定数，収められていることが分かる．

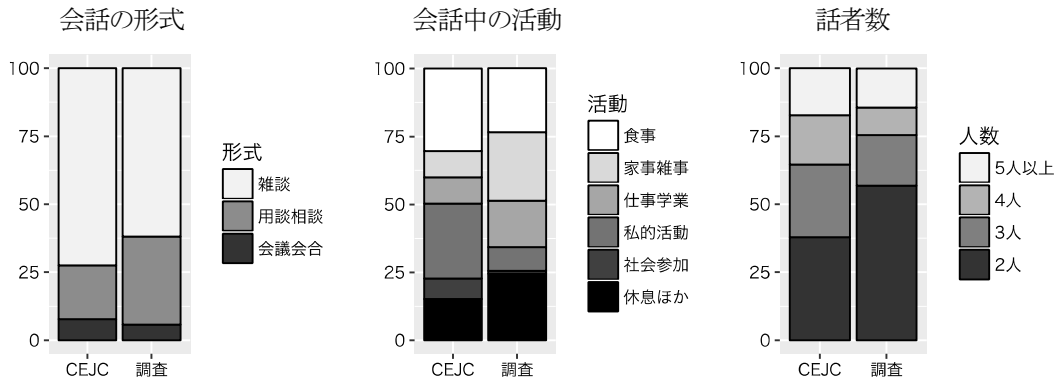

図 2：会話の形式・会話中の活動・話者数の内訳：モニター公開版と行動調査の比較

図 2 右に会話中の話者数の内訳を示す．話者数によって会話の構造などが変わりうるため，CEJC では話者数についてもできるだけ多様性を持たせるよう心掛けている．行動調査では 2 人会話が約 6 割を占めていたが，多様性を確保するため，3 人会話，4 人会話を少し多めに含めている．

2.5　検索システム

2.5.1　全文検索システム「ひまわり」

ハードディスク版には，転記テキストを対象に文字列や単語での検索ができる全文検索システム「ひまわり」が同梱されている．図 3 は，短単位「やっぱり」を検索した結果の画面である．短単位情報や前後文脈のほか，話者や会話に関するメタ情報が出力される．また簡単な集計などを行うこともできる．

この検索システムには，観察支援システム FishWatchr の機能が統合されており，検索した箇所や転記テキストの任意の位置の映像を簡単に閲覧することができる．このように，検索結果からすぐに該当箇所の映像データを閲覧できることによって，研究の可能性は格段に広がるものと考えられる．

2.5.2　オンライン検索システム「中納言」

CEJC モニター公開版は，国立国語研究所が提供するさまざまなコーパスをオンラインで検索できる「中納言」でも利用できる[2]．話者や会話に関するメタ情報なども合わせて表示される点は，ハードディスク版に同梱されている全文検索システム「ひまわり」と同じだが，動画の閲覧はできない[3]．このようにオンライン検索システム「中納言」は機能の点では制限されるが，現代日本語書き言葉均衡コーパスや日本語話し言葉コーパスなど，「中納言」で検索可能な多様なコーパスを，同種の条件で容易に検索し比較できるといったメリットもある．

[2] https://chunagon.ninjal.ac.jp

[3] 2019 年度中に，検索箇所の音声を一部視聴できる機能を付ける予定である．

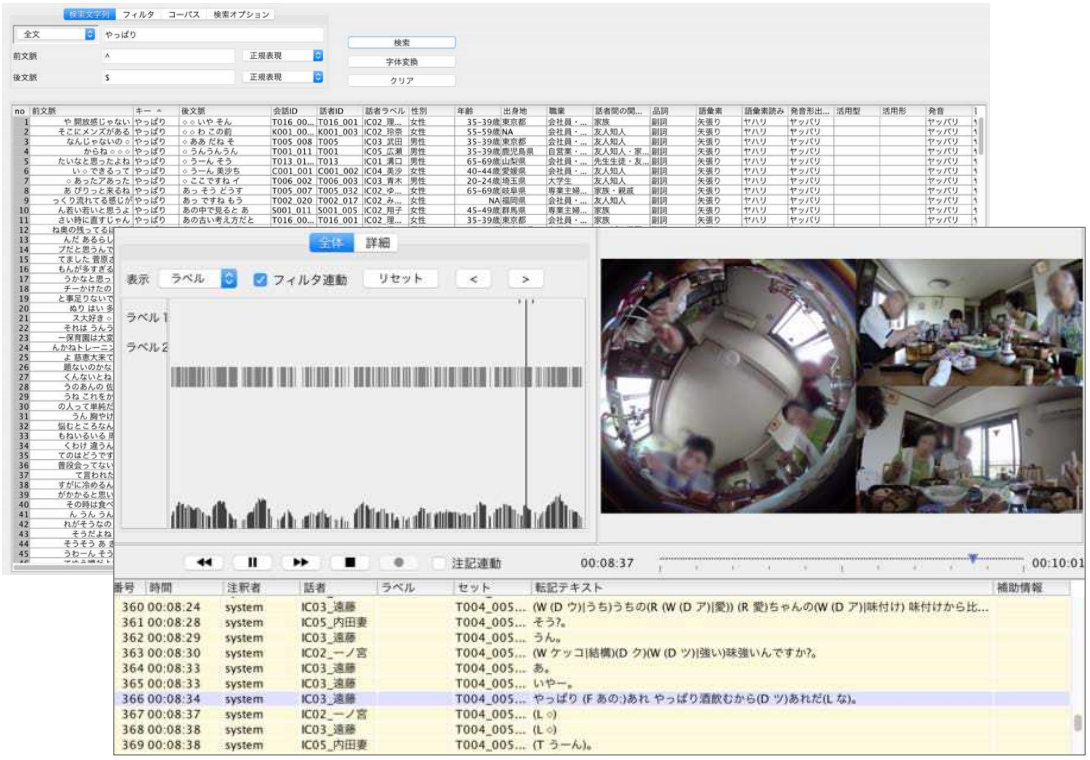

図 3：「ひまわり」の検索画面と，検索箇所の転記テキスト・映像の閲覧画面

3．コーパスを用いた研究の可能性

3.1　複数のコーパスに基づくレジスター間の比較：縮約形を例に

1 節で述べたように，これまで国立国語研究所では，新聞や雑誌，ブログなど多様なレジスターの書き言葉をバランスよく集めた『現代日本語書き言葉均衡コーパス』(BCCWJ)や，講演を中心とする『日本語話し言葉コーパス』(CSJ) など，大規模なコーパスを公開してきた．これに日常会話を対象とする CEJC モニター公開版が加わることで，多様なレジスターの言葉を，同一の言語単位・品詞体系のもとに整備されたコーパスを使って容易に比較することができるようになった．ここではその一例として，「見ている」や「読んでいる」などの「動詞＋て＋いる」と，「見てる」や「読んでる」のようなその縮約形に着目し，レジスターごとにどの程度の割合で縮約形が現れるかを比較する．2.5.2 節に記した「中納言」を用いてデータを抽出し，結果をまとめた（図 4）．

BCCWJ に含まれる書き言葉から見てみよう．行政白書のようなスタイルの高いレジスターの文書では縮約形は一切出現しないのに対し，会話文を含む小説では少しずつ見られるようになり，話し言葉に近いとされるブログでは縮約形が 40%近くを占めていることが分かる．CSJ が対象とする講演を見ると，学会講演のようにスタイルの高い話し言葉ではブログよりも少なく 20%強しか見られないのに対し，個人的な体験談などを集めた模擬講演では約半数が縮約形である．一方，日常会話では，縮約形が 98%と，ほぼ縮約された形でしか発話されていないことが分かる．詳細は示さないが，雑談，用談相談，会議会合といった会話の形式別に見ても，また年齢別に見ても，縮約形の出現率に差はほとんど見られない．日常会話では縮約形がかなり定着した形として用いられていることが分かる．

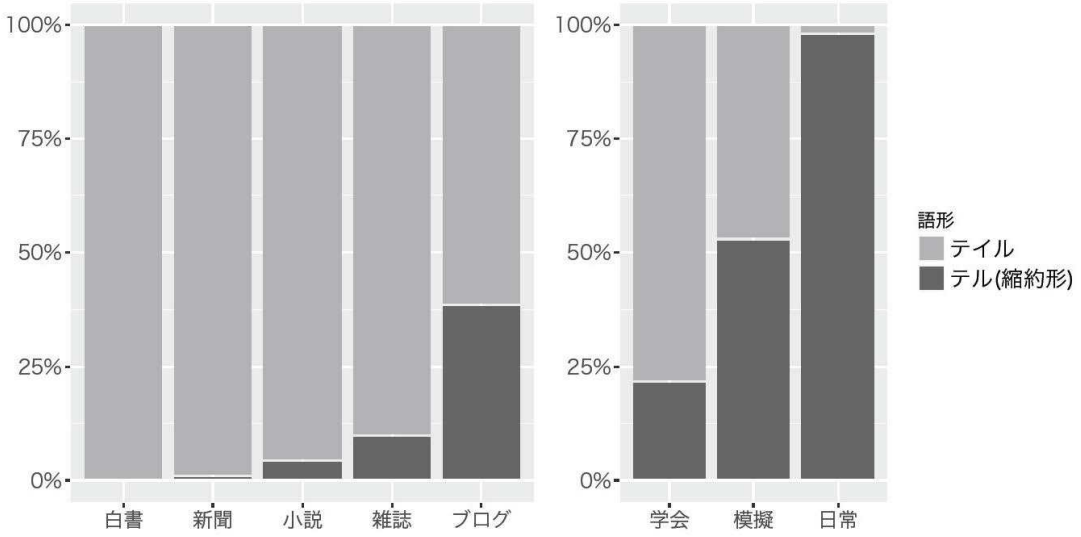

図 4：レジスターごとに見た「動詞＋て＋いる」の縮約形・非縮約形の出現率

3.2　会話の相手や場面などに応じたことばの使い分け：丁寧体・普通体を例に

前節では，日常会話をひとつのレジスターとしてまとめ，書き言葉や講演と比べた．しかし日常会話の中でも，会話の相手や場面などに応じたことばの使い分けがある．CEJC は多様な話者・多様な場面の会話を対象とすることから，こうした分析に適したコーパスであると考えられる．そこで本節では，会話の相手や場面などに応じた使い分けが顕著に現れる例として丁寧体・普通体の選択を取り上げる．

分析対象は述語を動詞・形容詞とする発話である．従属節で終わる中途発話文は分析から除いた．また相手との関係性が一意に特定できる話者の発話を対象とした．図 5 に，話し手から見た聞き手の関係性ごとの丁寧体・普通体の出現率を示す．

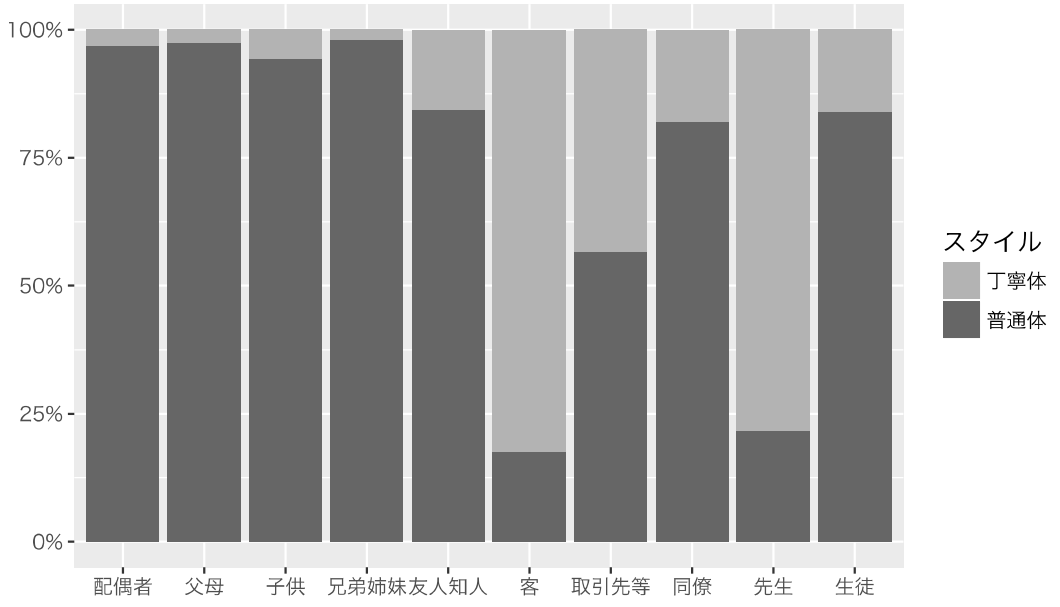

図 5：聞き手の関係性ごとに見た丁寧体・普通体の出現率

図 5 から，家族に対してはほぼ普通体しか用いないのに対し，相手が客や先生の場合には高い確率で丁寧体を用いていることが分かる．一方，友人知人については，丁寧体出現率の分散が大きく，丁寧体・普通体の選択にその他の要因が影響している可能性がある．そこで友人知人の場合に限定し，会話の形式として雑談と用談相談・会議会合に分けた上で，上下関係別に丁寧体・普通体の出現率を調べてみると（図 6），いずれの形式においても，「年下＜同世代＜年上」の順に丁寧体が多く用いられ，また雑談より用談相談・会議会合の方が丁寧体を多く用いていることが分かる．

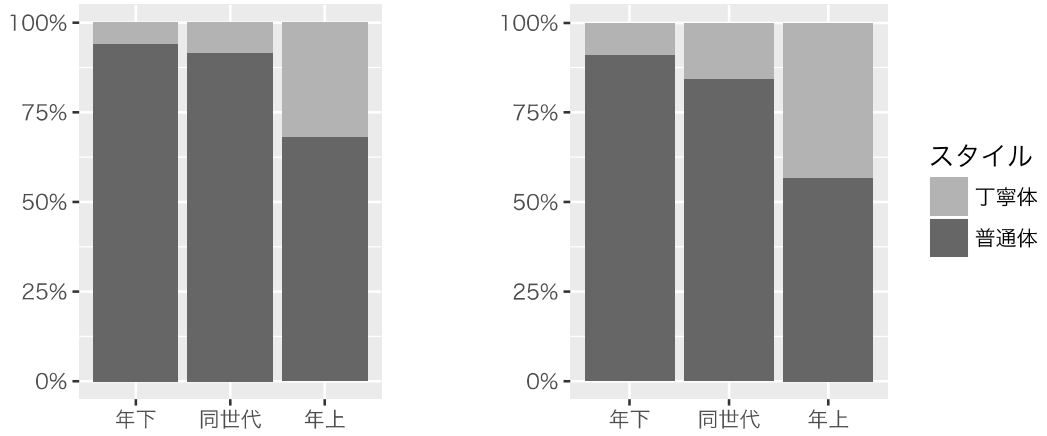

図 6：会話形式・上下関係ごとに見た丁寧体・普通体の出現率

以上見てきた丁寧体・普通体の使い分けは，これまでにも条件を統制した会話データを用いた研究の蓄積があり，目新しい結果ではないだろう．しかし，作られた環境での会話ではなく，まさに日常場面で我々がどのような言葉の使い分けをしているかを，コーパスを用いて定量的に明らかにできるという点は重要である．

4．おわりに

本稿では，CEJC モニター公開版の設計と構成について説明した上で，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，二つの研究事例を通して示した．2018 年 12 月の公開以来，言語学や日本語学だけでなく，日本語教育や情報工学，認知科学など，幅広い分野からの利用申請があった．モニター公開版はコーパスの利用可能性を把握することを目的としている．今後，こうした研究分野で本コーパスがどのように活用されるかを把握し，2021 年度末に予定している本公開に向けて構築を進めていく．

＊謝辞

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

文献

五十嵐陽介（2015）「韻律情報」小磯花絵（編）『話し言葉コーパス 設計と構築』81-100.

朝倉書店.
居關友里子ほか（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』104-107.
臼田泰如ほか（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.
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』68-88.朝倉書店.
現代日本語研究会ほか編（2016）『談話資料 日常生活のことば』ひつじ書房.
小磯花絵ほか（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.
小磯花絵ほか（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.
小磯花絵ほか（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. 国立国語研究所.
https://www2.ninjal.ac.jp/conversation/report/report03.pdf（2019 年 7 月 18 日確認）
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.ひつじ書房.

（2019 年 7 月 18 日受付）

Mathematical Linguistics, Vol.32 No.2 (September 2019) pp.133-142.

Resource

Compilation of the Monitor Version of the *Corpus of Everyday Japanese Conversation*

KOISO Hanae (National Institute for Japanese Language and Linguistics)

Abstract:
We have been constructing the *Corpus of Everyday Japanese Conversation*, CEJC, under the NINJAL collaborative research project since 2016. The main features of the CEJC are i) that we target conversations embedded in naturally occurring activities in daily life; ii) that we collect various kinds of everyday conversations in a balanced manner so as to capture the diversity of everyday conversations and to observe natural conversational behavior; and iii) that we collect and publish not only audio but also video data in order to precisely understand the mechanism of our real-life social behavior. Prior to the publication of the whole corpus scheduled for 2022, we published the monitor version of the CEJC in December 2018. In this article, we first outline the design of the monitor version of the CEJC. Then, we conduct a preliminary analysis, showing possible implications of the corpus.

Keywords: conversation corpus, corpus design, annotation