Design and Construction of the Corpus of Everyday Japanese Conversation

国立国語研究所論集 (NINJAL Research Papers) 24: 153–168 (2023)
ISSN 2186-1358

『日本語日常会話コーパス』設計と構築

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[c]　田中弥生[b]　伝　康晴[d]　西川賢哉[b]　渡邊友香[e]

[a] 国立国語研究所 研究系
[b] 国立国語研究所 研究系 非常勤研究員
[c] 国立国語研究所 言語資源開発センター
[d] 千葉大学
[e] 国立国語研究所 研究系 技術補佐員

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，2022 年 3 月に最終公開した。CEJC は，（1）日常生活で実際に交わされる会話を対象とすること，（2）多様な場面における多様な話者による会話をバランスよく格納すること，（3）映像まで含めて公開することを特徴とする。日常会話を対象とする映像付き大規模コーパスの構築は世界的に見ても新しい取り組みである。コーパスの規模は，200 時間，577 会話，240 万語，延べ話者数 1675 人である。本稿では，コーパスの設計・構築について，会話の収録法や収録機器，コーパスの基本構成，公開する音声・映像データのフォーマット，転記テキスト，各種アノテーション等などの観点から概観した上で，収録データのバランスについて検証する *。

キーワード：音声コーパス，日常会話，コーパス設計，アノテーション

1. はじめに

国立国語研究所ではこれまで，『日本語話し言葉コーパス』や『現代日本語書き言葉均衡コーパス』，『国語研日本語ウェブコーパス』，『日本語歴史コーパス』など，大規模なコーパスの構築・公開を進めてきた。特に現代日本語の書き言葉の全体像を把握するために構築された『現代日本語書き言葉均衡コーパス』の公開により，辞書編纂や自然言語処理，言語教育での活用など，基礎研究に留まらない研究の広がりを見せている。一方，話し言葉については，『日本語話し言葉コーパス』の公開により，話し言葉の言語学的・音声学的な研究や音声情報処理研究を支える基盤は整えられたと言えよう。しかし，『日本語話し言葉コーパス』は，独話を主対象とするコーパスであり，日常生活の中で交わされる会話は含まれていない。我々は日常生活の中でどのような言葉を使い，人といかなる仕組みでコミュニケーションしているのか，また日常場面でのさまざまな活動を言葉や身体を用いていかに組織化しているのかなど，問うべき課題は多い。こうした研究を支える基盤として，実際の日常会話場面を対象とした大規模な会話コーパスの構築が不

* 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」（プロジェクトリーダー：小磯花絵）の研究成果である。言語処理学会第 28 回年次大会の発表（小磯ほか 2022b）を基にまとめたものである。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。

可欠である。

海外では，Quirk により 1959 年に開始された The Survey of English Usage 計画において，書き言葉だけでなく話し言葉が大規模に収録され，それに基づく記述文法書が作成されている。その後も，British National Corpus や Bank of English，The Santa Barbara Corpus of Spoken American English など，会話を含む話し言葉を収録した大規模なコーパスが数多く構築され，それに基づく語彙・文法・談話研究やレジスター研究，会話コミュニケーション研究などが推進されてきた。

日本においても，1950 年代から国語研究所において日常会話を含む話し言葉の収録とそれに基づく実証的な話し言葉研究が始まり，『談話語の実態』（国立国語研究所 1955）や『話しことばの文型』（国立国語研究所 1960，1963）といった研究報告書がまとめられた。『談話語の実態』は，The Survey of English Usage 計画が始まる 4 年前に刊行されており，世界的に見ても先進的取り組みであったと言える。しかし残念ながら，収集された音声資料は当時一般には公開されず[1]，その後も日本国内において長らく話し言葉コーパスの構築・公開は行われてこなかった。1990 年以降，種々の話し言葉コーパスが公開されるようになったが（表 1），大学生などの若者や親しい者同士の会話，職場での会話，電話会話といったように，特定の場面や話者層に偏ったもの，また収録のために集まって雑談してもらうなど作られた場面の会話を対象とするものが多かった。更に音声データを公開していないコーパスも少なからずあり，映像データまで含むコーパスはほとんど存在しない。しかし我々が普段用いる言語の実態を調べるには，実際の日常場面の会話を対象とするコーパスが不可欠である。また非言語行動まで含め総合的にコミュニケーション行動の仕組みを明らかにするには，映像データまで含めたコーパスが求められる。

このような状況を受け，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，さまざまなタイプの日常会話 200 時間をバランス良く収録した『日本語日常会話コーパス』（*Corpus of Everyday Japanese Conversation*，以下 CEJC）の構築を進めてきた。2018 年 12 月には，CEJC の利用可能性や問題などを把握するために，CEJC 全体 200 時間のうち 50 時間の会話をモニター公開（小磯・天谷ほか 2019，2020），また 2020 年 3 月には 50 時間を追加公開し，幅広い分野の研究に活用されてきた。そしてプロジェクトの最終年度にあたる 2022 年 3 月に CEJC 全体 200 時間を本公開した（小磯ほか 2022a，2022b）。

本稿ではまず，CEJC 本公開版の設計・構築について概観した上で，設計通り多様な種類の会話をバランスよく収録できているかを検証する。

[1] 当時収録した話し言葉データは大変貴重であることから，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」において再整備し，『昭和話し言葉コーパス』として 2022 年 3 月に一般公開した（丸山ほか 2022）。

表 1　主要な日本語の会話コーパス

コーパス名	規模	概要	音声・映像
名大会話コーパス [2]	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 2022 年 3 月 NCRB 連動版 [3]	474 会話，118.5 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）
日本語話題別会話コーパス J-TOCC[4]	1800 会話，150 時間	120 組の親しい大学生同士が 15 種類の話題別に 5 分ずつ会話	無
CABank Japanese Sakura Corpus[5]	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 [6]	12 会話，2 時間	大学生の友人同士の会話	音声
CallFriend Japanese[7]	60 会話	アメリカ在住の日本人同士の電話会話	音声
談話資料 日常生活のことば（現代日本語研究会ほか 2016）	96 会話，18 時間	日常生活の会話	無
女性のことば・職場編 男性のことば・職場編 [8]	111 会話，21 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

2. コーパスの設計・構築

2.1 設計の方針

CEJC の設計方針は次の 3 点である。日常場面の会話を 200 時間という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と言える。

（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。
（2）　できるだけ多様な場面における多様な話者との会話をバランスよく集める。
（3）　音声や転記テキストだけでなく映像まで含めて収録・公開する。

2.2 会話の収録

2.2.1 二つの収録法：個人密着法・特定場面法

前節で言及したように，CEJC が対象とするのは日常生活の中で自然に生じる会話である。こうした会話をバランスよく収録するために，次の二つの方法でデータを収録した。

個人密着法：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）40 名（男女× 20 代・30 代・40 代・50 代・60 歳以上×各 4 名）に収録機材等を 3 ヶ月ほど貸し

[2] https://mmsrv.ninjal.ac.jp/nucc/
[3] https://ninjal-usamilab.info/btsj_corpus/
[4] http://nakamata.info/database/
[5] https://ca.talkbank.org/access/Sakura.html
[6] http://research.nii.ac.jp/src/Chiba3Party.html
[7] https://catalog.ldc.upenn.edu/LDC96S53
[8] 現代日本語研究会（2011）の付属談話資料として公開されているほか，『現日研・職場談話コーパス』（https://www2.ninjal.ac.jp/conversation/shokuba.html）として整備され，検索システム「中納言」（https://chunagon.ninjal.ac.jp/）で公開されている。

出し，協力者を中心に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った[9]。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，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協力者 1 人あたり 4 〜 6 時間，計 185 時間の会話を選別した。収録データの選定方針については 2.3.1 節で，協力者の詳細については 2.5.1 節で述べる。また個人密着法の詳細については田中ほか（2017，2018）を参照されたい。

特定場面法：個人密着法で収録した会話の場面や話者のバランスを検証し，個人密着法では収録が難しい場面を調査者が主体となり収録した。具体的には，(1) 仕事に関する会議会合，および，(2) 未成年者の会話が少なかったことから（小磯・天谷ほか 2020），これらを増補すべく，仕事中の会議会合を約 10 時間，中高生の雑談・講義・打合せ等を約 5 時間，計 15 時間をこの収録法で収集した。対象の詳細は 2.5.2 節で述べる。

2.2.2 収録機材

会話の収録に際しては，以下の対面会話収録を基本としつつ，移動時や電話等の状況に応じて異なる機材を用いた。収録に用いた機材を表 2 に示す。なお，この表に示す設定は収録時に採用したものであり，公開する音声・映像データの設定ではない。公開データの仕様については 2.4.2 節で言及する。収録機材の詳細は田中ほか（2017，2018）を参照のこと。

対面会話収録（基本収録）：対面会話の収録には原則として表 2（1）の機材を用いた。収録の様子および機材を図 1 に示す。中央 IC レコーダーは，会話の場の中央に 1 台配置し会話全体の音声を録音した。また個々の話者の音声を鮮明に収録するために，各話者は個人 IC レコーダーをフォルダーに入れ首から下げて収録した。PIXPRO SP360 4k は 360 度撮影可能なカメラで，会話の中央に 1 台配置し会話全体を撮影，GoPro Hero3+ は広角のカメラで，原則 2 台を会話の場の脇に配置し会話を俯瞰的に撮影した。なお収録の状況や失敗等により一部の機材のみを用いて収録することもあった。

徒歩移動時の収録：散歩などの徒歩移動時は，移動の際の周りの状況や会話に登場する事物を把握することを目的に，表 2（2）に示すウェアラブルカメラを用い，話者のうち 1 名が装着して撮影した。また中央 IC レコーダーは用いず個人 IC レコーダーのみとした。

車内収録：車で移動中の会話を収録する場合には，表 2（3）に示すように，原則として個人 IC レコーダーおよび車内に固定した GoPro 1 〜 2 台を用いた。

電話収録：電話会話の収録については，2 人の音声を分離して録音する設定を当初試みたが，録音に不具合が生じたため，途中から電話をスピーカーフォンにし中央 IC レコーダーで収録する方法に切り替えた。原則としてカメラは用いなかった。

[9] 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，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た。

特定場面法での収録：特定場面法では室内での収録が中心であったため，原則として表 2（1）の機材を用いたが，大人数の会議会合場面では必要に応じて（5）のカメラも追加で使用した。

表 2　IC レコーダー・カメラ

（1）対面会話収録（基本収録）

音声	中央 IC レコーダー Sony ICD-SX1000	1 台	収録時の設定：リニア PCM，16bit，44.1kHz，ステレオ マイク：内蔵マイク（広指向性，感度は auto に設定）
	個人 IC レコーダー Sony ICD-SX734	人数分	収録時の設定：リニア PCM，16bit，44.1kHz，ステレオ マイク：内蔵マイク（狭指向性，感度は事前調査で定めたレベルに固定）
映像	Kodak PIXPRO SP360 4k	1 台	収録時の設定：1440 × 1440，59.94fps
	GoPro Hero3+	2 台	収録時の設定：1920 × 1080，59.94fps

（2）徒歩移動時の収録

音声	個人 IC レコーダー	人数分	機材・設定は対面会話収録（基本収録）に同じ
映像	Panasonic HX-A500	1 台	収録時の設定：920 × 1080，29.97fps

（3）車内収録

音声	個人 IC レコーダー	人数分	機材・設定は対面会話収録（基本収録）に同じ
映像	GoPro Hero3+	1 〜 2 台	収録時の設定：1920 × 1080，59.94fps

（4）電話収録

音声	中央 IC レコーダー	1 台	機材・設定は対面会話収録（基本収録）に同じ

（5）特定場面法での収録：原則（1）の機材を用いたが，必要に応じて以下のカメラを追加で使用

映像	Sony HDR-CX675	1 〜 2 台	収録時の設定：1920 × 1080，30fps

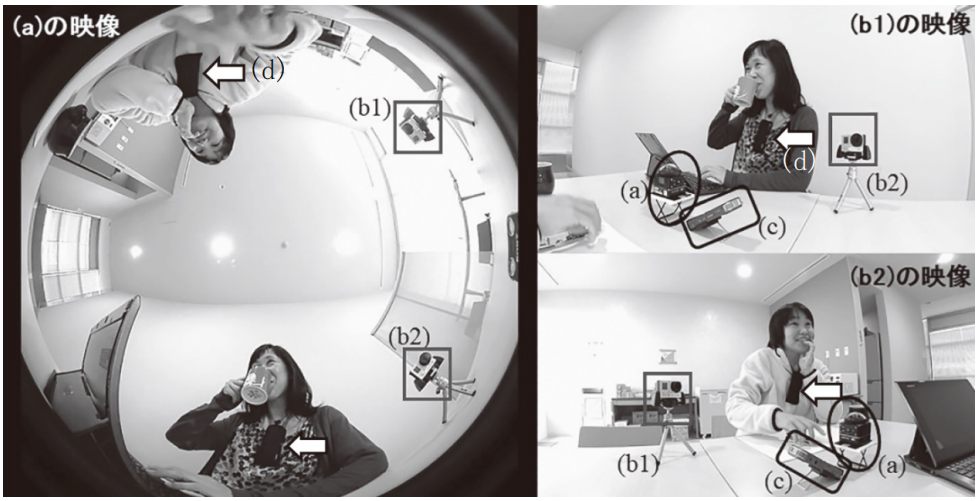

図 1　対面会話収録（基本収録）の様子
（a）PIXPRO SP360，（b）GoPro，（c）中央 IC レコーダー，（d）個人 IC レコーダー

2.3 データの選定・公開の方針

2.3.1 選定方針

我々の言語生活を正確に記述しその本質を解明するには，日常の言語生活の幅広いレジスターをカバーするようサンプルを選定することが求められる。『現代日本語書き言葉均衡コーパス』では，書き言葉の生産，流通，受容の各過程が書き言葉の実態を捉える上で重要とした上で，出版データと図書館収蔵図書を母集団としたランダムサンプリングを行い，生産実態と流通実態を反映したサブコーパスを設計した（Maekawa et al. 2014）。しかし話し言葉の場合，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，プロジェクトの開始に先立ち，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，2014 年から 2015 年にかけて会話行動調査を実施した。この調査では，243 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 ではこのうち，特に会話の種類（雑談，用談相談，会議会合，授業レッスン）や会話の話者数，会話の行われた場所（自宅，職場学校，交通機関など），会話中に行われた活動（家事雑事, 仕事学業, 食事など）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 節では，この調査結果と比較しながら CEJC のバランスを検証する。

2.3.2 公開方針

CEJC 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著作物の写り込みなどが多く見られる。そこで実際に収録した会話データをもとに具体的な問題を洗い出し, 肖像権や個人情報保護, 著作権などの観点からデータの公開方針を定めた。この方針に従い，公開可能なデータの選定や映像・音声・転記テキストのマスキングなどを実施した。方針の詳細については小磯・伝（2018）を参照されたい。

2.4 コーパスの構成

2.4.1 基本構成

表 3 に示す通り，CEJC 全体に対して，映像・音声データ，転記テキスト，短単位情報（人手修正），長単位情報（自動解析），会話・話者に関するメタ情報を提供する。また個人密着法で収録した会話の中から 20 時間を選別して「コア」データセットを作成し，人手修正・付与した複数のアノテーションを提供する。これらのデータを，(1) 有償版（全データ対象）と (2) オンライン検索システム『中納言』版（無償利用・提供データに制限あり）として公開する。

コアは個人密着法で収録した会話の中から選定した。多様性を確保するために，原則として協力者のうち 20 代，30 代，40 代，50 代，60 歳以上の男女各 2 名計 20 名が集めた会話の中から，一人あたり平均 1 時間（40 〜 80 分）の会話を選定した。ただし作業の進捗の都合で，30 代の女性が 3 名，60 歳以上の女性が 1 名となった。

表 3　提供するデータの一覧

	コーパス全体	コア	「中納言」版の対象
音声データ	○	○	検索箇所前後の再生
映像データ	○	○	×
転記テキスト	○	○	×
形態論情報（短単位情報）	○	○	○
形態論情報（長単位情報）	○	○	○
係り受け情報	×	○	×
談話行為情報	×	○	×
韻律情報	×	○	×
会話・話者に関するメタ情報	○	○	△（備考等除く）

CEJC のデータの規模は表 4 に示す通りである [10]。

表 4　データの規模

	コーパス全体	うち「コア」
時間数	200 時間	20 時間
セッション数	461 セッション	52 セッション
会話数	577 会話	52 会話
延べ話者数	1675 名	169 名
異なり話者数	862 名	135 名
語数（短単位）	約 240 万語	約 25 万語

2.4.2 音声・映像データ

2.2.2 節で言及したように，中央 IC レコーダーで収録した会話全体の音声データと，個人 IC レコーダーで収録した話者ごとの音声データを公開する。ただし，中央 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，個人 IC レコーダーで収録した各話者の音声を合成した音声を作成した。公開する音声データのファイル形式はいずれもリニア PCM，16bit，16kHz，中央 IC レコーダーと合成音声はステレオ，個人 IC はモノラルである。

映像データについては，個々の映像に加え，それらを合成した映像（図 1 参照）を公開する。映像データのファイル形式は，いずれも mp4，コーデック H.264，フレームレート 29.97fps である。

2.4.3 転記テキスト

転記テキストは，発話単位（JDRI 2017）と転記単位（発話単位を，知覚可能なポーズあるいは笑いなど異なる音種によって細かく切ったもの）という 2 種類の単位を採用し，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

[10] ここで「セッション」とは協力者が 1 回に収録したもの，「会話」とはそこからある程度のまとまりをもった範囲を切り出したものを指す。倫理的・法的な問題や協力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，一つのセッションが複数の会話に分かれることもある。

成した。漢字仮名交じり表記を基本とし，言いさしや言い間違い，非語彙的な母音の延伸，笑いなどを表すタグによって会話に生じる諸現象を表現している（臼田ほか 2018）。発音の情報については次項「形態論情報」を参照のこと。またコアについては，発話の重複箇所の明示など，情報をより詳細化した転記テキストも合わせて提供する。

2.4.4 アノテーション

■ **形態論情報（短単位情報・長単位情報）** CEJC では，長短 2 種類の形態論情報を提供する。短単位情報は，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，200 時間全体を人手修正した。長単位は長単位解析器 Comainu を用いて自動解析した上でコアに限定して人手修正した。品詞については，原則として『現代日本語書き言葉均衡コーパス』の体系を踏襲したが，CEJC は話し言葉であることから，「言いよどみ」や「歌」（ハミングなどで歌っている箇所），「伏せ字」などの品詞を設けた。また，語彙素・語形が一意に同定できない語（例：色紙「シキシ／イロガミ」）は，音を聴取した上で特定した。転記のタグを利用して得られる言い間違いを含む実際の発音（例：「国語」の発音「コクゴ」「コクゴー」「クゴ」「コクゲゴ」など）の情報も提供する。

■ **係り受け情報** コアを対象に，発話単位（2.4.3 節参照）を範囲に，文節間の係り受け関係の情報を付与した。BCCWJ-DepPara（浅原・松本 2018）の基準に準じ，（1）通常の係り受け相当の “D”，（2）文境界相当の “Z”，（3）係り受けを付与するうえで後続文節と連結する “B”，（4）係り先が決められない “F” の四つのラベルを用いる（浅原・若狭 2022）。

■ **談話行為情報** コアを対象に，ISO 24617-2（ISO 24617-2 2012）をベースに日常会話用に整備した基準に基づき，発話単位を対象に人手で付与した（Iseki et al. 2019）。基本的な談話機能（例：情報提供・申し出・挨拶・謝罪）を中心に，談話の展開や会話の調整に関わる情報，発話間の依存関係に関する情報も付与した。

■ **韻律情報** コアに含まれる 157 名の主たる話者（店員など一時的に会話に参加するものを除く）のうち，方言の使用状況や音声の質を考慮して 151 名を選別した上で，『日本語話し言葉コーパス』用に整備した X-JToBI（五十嵐ほか 2006）の簡略版に準拠して韻律情報を付与した（小磯・菊池・山田 2020）。アクセント句・イントネーション句の境界情報や，句末の音調情報などが提供される。

■ **メタ情報** 会話に関するメタ情報として，収録年，話者数，会話形式，会話が行われた場所，会話中の活動，話者間の関係性などの情報を，また話者に関するメタ情報として，年齢（5 歳刻み），性別，出身地（都道府県，外国の場合は国），居住地（同），職業，協力者からみた関係性（個人密着法のみ）などを提供する。

2.4.5 検索システム

オンライン検索システム「中納言」（音声再生機能付き）を無償利用できるほか，有償版では全文検索システム「ひまわり」が提供される。観察支援システム FishWatchr 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視聴することができる（山口 2018）。

2.5 データの内訳

2.5.1 個人密着法による収録データの内訳

個人密着法では，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 歳以上×各 4 名）に収録を依頼し，コーパスに格納するデータとして 185 時間の会話を選定した。表 5 に，協力者の属性，対象とする収録・会話の数，会話時間[11]，語数（短単位数）[12] の情報を示す。

表 5　個人密着法の協力者の属性，対象の収録セッション数（セ数）・会話数・会話時間・語数

年代	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	学生	8	10	4.3	49,059	学生	9	14	4.4	47,918
	学生	8	10	4.2	64,433	学生	16	21	6.0	68,083
	先生	14	14	5.5	48,266	会社員等	13	15	4.2	59,720
	先生	10	17	3.7	33,804	会社員等	8	8	4.0	56,620
30 代	会社員等	6	12	3.1	30,873	会社員等	11	12	5.0	52,038
	会社員等	11	11	4.7	64,451	自由業	15	22	4.8	53,156
	自由業	11	11	5.6	58,240	自由業	15	17	5.4	58,198
	会社員等	13	14	4.6	82,450	専業主婦	12	12	5.6	68,581
40 代	会社員等	8	10	3.6	39,732	会社員等	9	9	4.5	49,450
	会社員等	8	11	3.9	38,310	パートタイム	12	12	5.0	66,925
	先生	15	23	5.0	56,877	パートタイム	10	10	4.8	57,758
	自由業	9	13	4.8	57,789	自営業	11	17	4.4	55,050
50 代	会社員等	9	9	4.6	50,350	会社員等	10	14	4.2	51,675
	会社員等	15	17	6.0	69,213	会社員等	9	12	4.5	51,399
	先生	8	9	4.2	56,859	自営業	10	11	4.6	55,375
	先生	9	10	4.2	48,891	自由業	10	12	4.6	43,165
60 歳以上	定年退職	10	14	5.8	84,299	専業主婦	11	13	5.1	66,449
	定年退職	12	13	4.6	65,008	会社員等	9	12	4.2	53,324
	自由業	12	18	4.8	48,590	自営業	12	14	4.4	59,341
	先生	15	17	4.3	60,376	自由業	11	13	4.3	52,427
計		211	263	91.5	1,107,870		223	270	93.8	1,126,652
						総計	434	533	185.3	2,234,522

2.5.2 特定場面法による収録データの内訳

特定場面法では，表 6 に示すように，仕事中の会議会合を約 10 時間と，中高生の雑談・講義・打合せ等約 5 時間を収集した。

[11] 表 5・表 6 の「時間」について，小数点以下第二位を四捨五入しているため，内訳の計と総計とが合わないことがある。なお，単位は「h（時間）」である。
[12] 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。

表 6　特定場面法による会話の種別と収録セッション数（セ数）・会話数・会話時間・語数

種類	概要	セ数	会話数	時間	語数
仕事中の会議会合	研究機関での広報に関する会議	4	9	2.8	29,337
	歯科医院での定例会合	10	16	1.9	26,278
	撮影に関する会合	3	4	2.3	25,605
	学会関係の講習会企画の会合	1	3	2.3	27,228
	事務所での書籍出版に関する会合	1	2	0.8	11,128
高校生の会話	高校生の部活の反省会	1	1	0.5	9,787
	高校生同士の雑談	1	1	1.1	9,515
	高校生同士の雑談・母子間の雑談	2	3	1.4	14,962
	高校生同士のオンラインでの雑談	1	2	0.3	2,501
中学生の会話	講義での中学生と教員との会話・その後の雑談	2	2	0.3	2,670
	講義での中学生と教員との会話	1	1	1.3	19,145
	計	27	44	14.9	178,156

3. 収録データのバランスの検証

3.1 会話の属性の観点から

2.3.1 節でも言及したように，CEJC では会話行動調査の結果を一つの目安として CEJC のデータの選定を行った。そこで本節では，会話の形式，会話の話者数，会話が行われた場所，会話中に行われた活動を対象にその内訳を求め，調査結果と比較することで，CEJC に含まれる会話のバランスを検証する。

3.1.1 会話の形式

会話の形式（雑談／用談相談／会合・授業）に関する結果を図 2 に示す。図の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の比較である。

会話の件数で見ると，CEJC では行動調査よりも雑談が少し多く用談相談は少ない傾向を示しているが，会話時間で見るとバランスよく収録できていることが分かる。行動調査から現実の日常生活では 5 分未満の短い用談相談が多く，必然的に会話の件数に対し総時間は短くなる傾向にあり，そのことが用談相談の件数と時間の差に現れていると考えられる。また会合・授業は件数で見ると行動調査より多いが，時間で見ると若干少ない。コーパスに含める会話は多様性を確保するために最大 1 時間という制限を設けているが，実際の会合・授業には 1 時間を越える長い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。

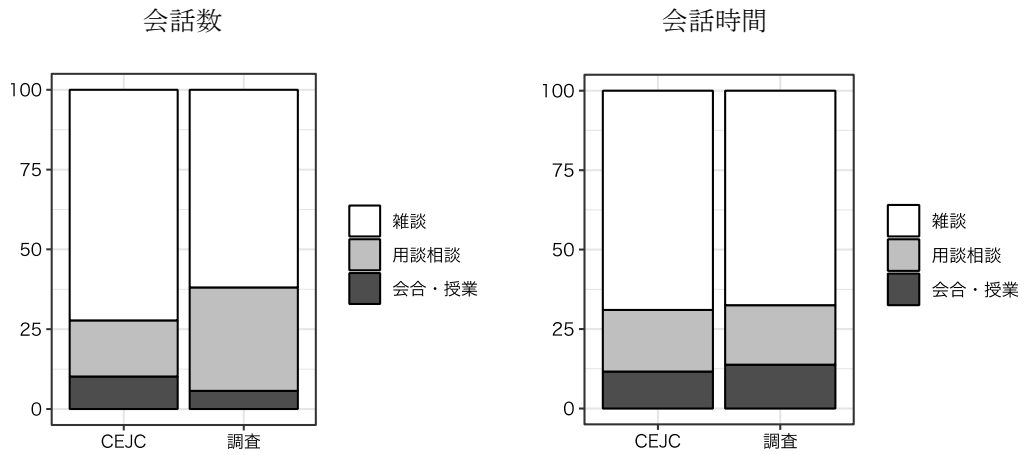

図 2　会話の形式：CEJC と会話行動調査の比較

3.1.2 会話の話者数

会話の話者数ごとの結果を図 3 に示す。前節の会話形式とは逆に，会話の話者数については，件数で見るとバランスよく収録できているが，時間で見ると CEJC の方が行動調査よりも 5 人以上の会話の時間が少ない傾向が見られる。行動調査の結果から，5 人以上の場合，1 〜 5 時間の長い会話が多く含まれることが分かっている（小磯ほか 2016）。しかし CEJC では多くの会話を収録しバリエーションを確保するために，1 会話の上限を約 1 時間としており，この選定基準が 5 人以上の会話時間の抑制につながったものと見られる。

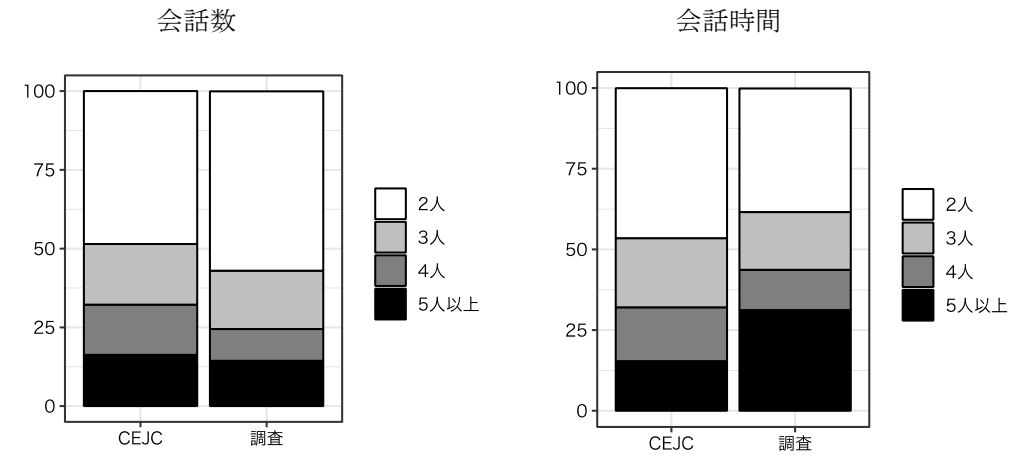

図 3　会話の話者数：CEJC と会話行動調査の比較

3.1.3 会話の行われた場所・会話中に行われた活動

会話の行われた場所（「場所」）と会話中に行われた活動（「活動」）に関する結果を図 4・図 5

に示す。なお活動については一つの会話に対し二つのカテゴリーを当てることもあるが，その場合は二重にカウントしている。

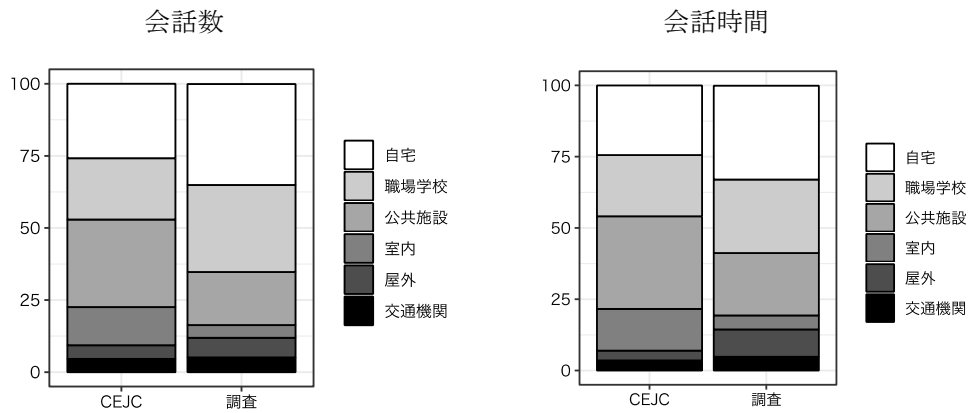

図 4　会話の行われた場所：CEJC と会話行動調査の比較

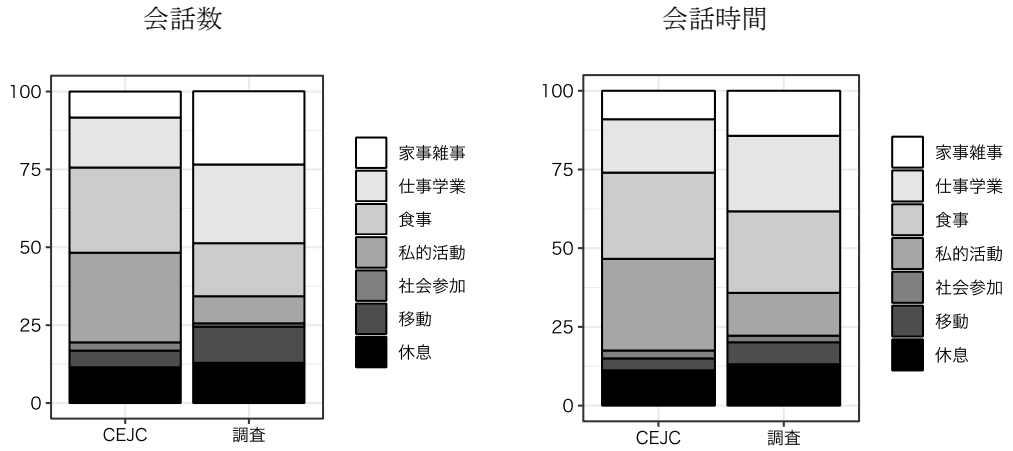

図 5　会話中に行われた活動：CEJC と会話行動調査の比較

CEJC では，自宅での食事の準備や片付け，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ものの，行動調査と比べると，自宅・職場での家事雑事・仕事中の会話が若干少なく，飲食店などの商業公共施設や友人宅・実家などの室内での私的活動（友人との付き合いやレジャー活動，課外活動等）中の会話が多い傾向が見られる。なお，自宅での会話数・会話時間が行動調査よりも少ない傾向が見られるが，これはコーパス中の会話のバリエーションを増やすために，あえて自宅での会話を減らしたことによる。

3.2 話者の属性の観点から

表 7 に性別・年代別にみた話者数・発話時間・語数の情報を示す。ここで発話時間とは，当該話者が実際に発話した時間を，転記テキストの発話単位の継続時間長から算出したものである。また図 6 に，性別・年代ごとの話者数の分布を示す。

表 7　年齢・性別ごとの話者数・発話時間・語数（千語）

	男性				女性				計			
年齢	延べ話者数	異なり話者数	発話時間	語数	延べ話者数	異なり話者数	発話時間	語数	延べ話者数	異なり話者数	発話時間	語数
～ 9 歳	27	11	1.7	16	7	5	0.3	3	34	16	2.0	19
10 代	71	33	4.6	62	46	31	3.3	46	117	64	7.9	108
20 代	123	56	15.7	220	123	45	12.7	177	246	101	28.4	397
30 代	121	60	14.8	204	163	55	16.9	220	284	115	31.8	424
40 代	100	41	11.7	166	150	88	21.3	270	250	129	33.0	436
50 代	108	57	11.5	159	164	77	23.4	312	272	134	34.9	471
60 代	119	49	12.7	172	100	47	14.8	185	219	96	27.6	358
70 代	55	37	6.3	87	30	19	3.8	49	85	56	10.0	136
80 代	15	11	1.0	13	11	7	1.7	20	26	18	2.8	33
90 代	2	1	0.2	2	9	6	0.9	10	11	7	1.0	13
不明	57	56	0.4	6	74	70	1.0	12	131	126	1.4	17
計	798	412	80.6	1108	877	450	100.0	1304	1675	862	180.7	2413

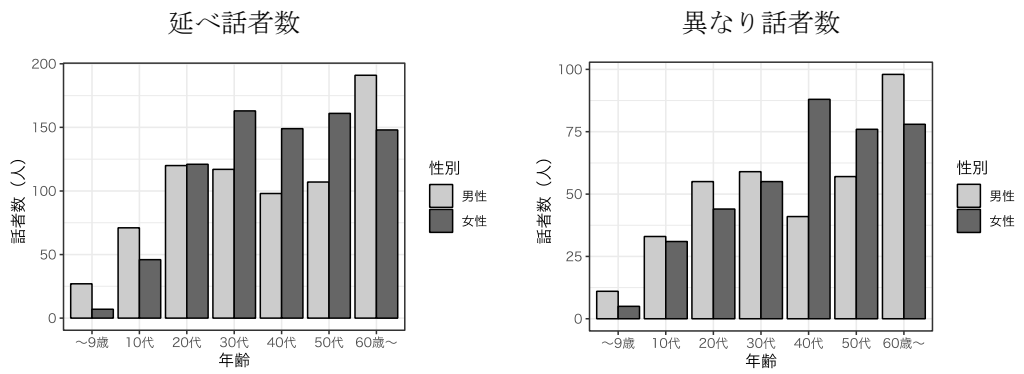

図 6　会話者の性別・年齢の内訳（人）

CEJC の大半を占める個人密着法では，成人の協力者を中心に，友人や同僚，家族などとの会話を収録しているため，必然的に協力者と同世代の話者が多く含まれることになる。結果から，40 代・50 代の男性が若干少なく女性が多いなどの多少の違いは見られるものの，20 代から 60 代までの成人についてはいずれの世代の男女とも延べ話者数 100 人以上，語数 15 万以上と，概ねバランスよく収録できていることが分かる。一方，未成年者については成人と比べて数が少ない。個人密着法が成人中心の収録法であるため，特定場面法により中高生を対象に友達同士の雑談や部活動の打合せなど 5 時間弱の会話を補ったことから，少なくとも 10 代についてはある程

度含まれているが，10 歳未満のデータはかなり限られる。

4. おわりに

本稿では，2022 年 3 月末に公開した CEJC の設計・構築について，会話の収録法や収録機器，コーパスの基本構成，音声・映像データの形式，転記テキスト，各種アノテーションなどの観点から概観した上で，設計通り多様な種類の会話をバランスよく収録できているかを検証した。2018 年 12 月に公開した CEJC モニター版は，言語学・日本語学に留まらず，日本語教育や情報工学，認知科学など，幅広い分野で活用されてきた。200 時間に増補した今回の本公開により，ますます多くの研究や技術開発に利用されることが期待される。

最後に課題を挙げておきたい。3.2 節でも言及したように，CEJC には未成年者，特に 10 歳未満の子どもの会話データがかなり少ないという問題がある。この年齢層の会話データを拡充するため，国立国語研究所共同研究プロジェクト「多世代会話コーパスに基づく話し言葉の総合的研究」を 2022 年度に開始し，子どもを主対象とする映像付きコーパスを構築する。

参照文献

Iseki, Yuriko, Keisuke Kadota and Yasuharu Den (2019) Characteristics of everyday conversation derived from the analysis of dialog act annotation. *Proceedings of the 22nd Conference of the Oriental COCOSDA,* 1–6. https://ieeexplore.ieee.org/document/9041235 (accessed November 2022).

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

Maekawa, Kikuo, 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): 345–371.

浅原正幸・松本裕治（2018）「『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション」『自然言語処理』25(4): 331–356.

浅原正幸・若狭絢（2022）「『日本語日常会話コーパス』に対する係り受け情報アノテーション」『言語処理学会第 28 回年次大会発表論文集』1699–1703.

五十嵐陽介・菊池英明・前川喜久雄（2006）「韻律情報」『日本語話し言葉コーパスの構築法』347–453.

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: 177–193.

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）』東京：ひつじ書房.

現代日本語研究会・遠藤織枝・小林美恵子・佐竹久仁子・橋美奈子編（2016）『談話資料 日常生活のことば』東京：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: 85–106.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉（2017）「『日本語日常会話コーパス』の構築」『言語処理学会第 23 回年次大会発表論文集』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）『プロジェクト報告書 3『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』

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（2020）「『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」『国立国語研究所論集』18: 17–33.

小磯花絵・菊池英明・山田高明（2020）「『日本語日常会話コーパス』への韻律ラベリング―ラベリングの設計と日常会話の韻律の特徴―」『人工知能学会研究会資料』SIG-SLUD-B903: 34–39.

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・渡邊友香（2022a）『プロジェクト報告書 6『日本語日常会話コーパス』―設計・構築・特徴―』
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・渡邊友香（2022b）「『日本語日常会話コーパス』の設計と特徴」『言語処理学会第 28 回年次大会発表論文集』2008–2012.
国立国語研究所（1955）『談話語の実態』（国立国語研究所報告 8）東京：秀英出版.
国立国語研究所（1960）『話しことばの文型（1）―対話資料による研究―』（国立国語研究所報告 18）東京：秀英出版.
国立国語研究所（1963）『話しことばの文型（2）―独話資料による研究―』（国立国語研究所報告 23）東京：秀英出版.
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/（2022 年 9 月 27 日確認）
田中弥生・柏野和佳子・角田ゆかり・臼田泰如・伝康晴・小磯花絵（2017）「『日本語日常会話コーパス』構築における会話収録方法」『言語処理学会第 23 回年次大会発表論文集』481–484.
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.
丸山岳彦・小磯花絵・西川賢哉（2022）「『昭和話し言葉コーパス』の設計と構築」『国立国語研究所論集』22: 197–221.
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』340–347.

関連 Web サイト

国立国語研究所『日本語日常会話コーパス』https://www2.ninjal.ac.jp/conversation/cejc.html（2022 年 9 月 26 日確認）

Design and Construction of the Corpus of Everyday Japanese Conversation

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

[a]Research Department, NINJAL
[b]Adjunct Researcher, Research Department, NINJAL
[c]Center for Language Resource Development, NINJAL
[d]Chiba University
[e]Technical Assistant, Research Department, NINJAL

Abstract
We have constructed the Corpus of Everyday Japanese Conversation (CEJC) and published it in March 2022. The main features of the CEJC include: i) a focus on conversations that occurred naturally in activities of daily life; ii) a balanced collection of everyday conversations that capture their diversity and facilitate the observation of natural, conversational behavior in daily life; and iii) the publication of audio and video data for a better understanding of the mechanism of real-life social behavior. The publication of a large-scale corpus of everyday conversations that includes video data is a new approach. The CEJC contains 200 hours of speech, 577 conversations, approximately 2.4 million words, and 1,675 speakers. In this paper, we describe the process involved in the design and construction of CEJC including the recording method and devices used, structure of the corpus, formats of the audio and video files, transcription, and annotations. We then examine how the conversations in the corpus were selected and compiled in a balanced manner to showcase their variety.

Keywords: spoken corpus, everyday conversation, corpus design, annotation