『日本語日常会話コーパス』の設計と特徴

小磯 花絵 [1]　天谷 晴香 [1]　石本 祐一 [1]　居關 友里子 [1]　臼田 泰如 [1]　柏野 和佳子 [1]
川端 良子 [1]　田中 弥生 [1]　伝 康晴 [2]　西川 賢哉 [1]　渡邊 友香 [1]

[1] 国立国語研究所　[2] 千葉大学

{koiso,h-amatani,yishi,iseki,usuda,waka,kawabata}@ninjal.ac.jp
{yayoi,nishikawa,yuwatanabe}@ninjal.ac.jp　den@chiba-u.jp

概要

本稿では 2022 年 3 月に一般公開する 200 時間規模の『日本語日常会話コーパス』（CEJC）の設計と特徴について報告する．CEJC は多様な場面・多様な話者による現実の日常会話をバランスよく格納し，映像まで含めて公開する点に特徴のあるコーパスである．CEJC 全体には音声・映像・転記・形態論情報（長短二種）が，うち 20 時間については更に係り受け情報，談話行為情報，韻律情報が提供される．

1　はじめに

2016 年度より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『日本語日常会話コーパス』（**Corpus of Everyday Japanese Conversation**, CEJC）の構築を進めてきた [1]．CEJC は，(1) 日常生活で交わされる会話を対象とすること，(2) 多様な場面・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．大規模な日常会話を映像まで含めて公開するのは世界的に見ても新しい取り組みである．2018 年に 50 時間のデータをモニター公開し [2, 3]，広く研究に活用されてきたが，2022 年 3 月に 200 時間全体を本公開する．2 節と 3 節で CEJC 本公開版の設計と特徴をそれぞれ報告する．

2　CEJC の設計

2.1　収録法

多様な会話をバランスよく収録するために，British National Corpus の話し言葉のパートの収録法 [4, 5] を参考に次の 2 つの方法で会話を収録した．

個人密着法　性別・年齢をバランスさせた協力者 40 名に，できるだけ多様な場面・話者との会話を収録してもらった．CEJC 200 時間のうち約 185 時間をこの収録法で収集した（表 1）．

特定場面法　個人密着法で収録された会話のバランスを検証し，不足する会話の種別を特定した上で [6]，その不足を補うために，仕事中の会議会合を約 10 時間，中高生の雑談・打合せ等を約 5 時間，計 15 時間をこの収録法で収集した．

表 1　調査協力者の属性と収録データの規模

年齢	男性			女性		
	職業	会話数	時間	職業	会話数	時間
20代	学生	10	4.3h	学生	14	4.4h
	学生	10	4.2h	学生	21	6.0h
	先生	14	5.5h	会社員等	15	4.2h
	先生	17	3.7h	会社員等	8	4.0h
30代	会社員等	12	3.1h	会社員等	12	5.0h
	会社員等	11	4.7h	自由業	22	4.8h
	自由業	11	5.6h	自由業	17	5.4h
	会社員等	14	4.6h	専業主婦	12	5.6h
40代	会社員等	10	3.6h	会社員等	9	4.5h
	会社員等	11	3.9h	パートタイム	12	5.0h
	先生	23	5.0h	パートタイム	10	4.8h
	自由業	13	4.8h	自営業	17	4.4h
50代	会社員等	9	4.6h	会社員等	14	4.2h
	会社員等	17	6.0h	会社員等	12	4.5h
	先生	9	4.2h	自営業	11	4.6h
	先生	10	4.2h	自由業	12	4.6h
60歳以上	定年退職	14	5.8h	専業主婦	13	5.1h
	定年退職	13	4.6h	会社員等	12	4.2h
	自由業	18	4.8h	自営業	14	4.4h
	先生	17	4.3h	自由業	13	4.3h
計		263	91.5h		270	94.0h

2.2　コーパスの構成

CEJC の構成を図 1 に示す．

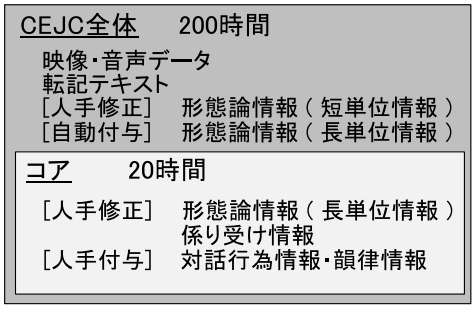

図 1　CEJC の構成

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2　収録に用いた機材と公開時の映像・音声ファイルの形式

		機種名・台数	公開時のファイル形式
映像	室内などでの基本収録	Kodak PIXPRO SP360 4k・最大 1 台	mp4, H264, 1440×1440, 29.97fps
		GoPro Hero3+・最大 2 台	mp4, H264, 1280×720, 29.97fps
	特定場面法の一部で使用	Sony ICD–SX1000・2 台	mp4, H264, 1280×720, 29.97fps
	移動時の収録	Panasonic HX-A500・1 台	mp4, H264, 1280×720, 29.97fps
	複数映像の合成 [∗1]	—	mp4, H264, 1360×720[∗2], 29.97fps
音声	個々人の音声	Sony ICD-SX734・話者数分	リニア PCM, 16bit, 16kHz, モノラル
	会話全体の音声	Sony ICD-SX1000・1 台	リニア PCM, 16bit, 16kHz, ステレオ
	複数音源の合成 [∗3]	—	リニア PCM, 16bit, 16kHz, ステレオ

∗1 基本収録で複数の映像ソースがある場合，1 つの映像データとして合成したものも公開（図 2 参照）
∗2 基本構成 3 台の場合．他の構成の場合には異なることがある．
∗3 会話全体の音声データに問題がある場合，個々人の音声を合成した音源を公開

200 時間に対して，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）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とし，人手修正・付与した複数のアノテーションを提供する．規模は，200 時間全体について，会話数 577，延べ話者数 1675 名，異なり話者数 862 名，約 240 万語（短単位），コア 20 時間について，会話数 52，延べ話者数 169 名，異なり話者数 135 名，約 25 万語である．

2.3　映像・音声データ

収録に用いた機材と公開時のファイル形式を表 2 に示す．図 2 にあるように，原則最大 3 台のカメラを用いて収録した．これら複数の映像を 1 つに合成した映像も作成して提供する．移動時の収録には 1 台のカメラを用い，周囲の様子などを中心に記録した．電話会話などで映像がないこともある．音声については，各話者が身に付けた IC レコーダーにより当該話者の音声を記録すると同時に，会話の場の中心に置いた IC レコーダーで会話全体の音声を記録した．会話全体の音声データに問題がある場合，個々人の音声を合成した音源を作成して公開する．映像・音声データについては収録の状況等により欠損もある．収録の詳細は [7] を参照のこと．

図 2　映像データの例．左の映像は PIXPRO SP360 で，右上下の映像は GoPro 2 台で撮影したもの．

2.4　転記テキスト

転記テキストは，発話単位 [8] と転記単位（発話単位を知覚可能なポーズなどにより区切った単位）という 2 種類の単位を採用し，ELAN / Praat を活用し映像・音声を参照しながら作成した．漢字仮名ま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する [9]．発音の情報については次節「形態論情報」を参照のこと．

2.5　アノテーション

形態論情報　形態論情報として短単位情報と長単位情報を提供する [10, 11]．短単位情報は，転記を対象に形態素解析器 MeCab と形態素解析用辞書 UniDic で解析した上で，コーパス全体を対象に人手修正した．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：「(W ジューリョー|柔道)」であれば「ジューリョー」）の情報も提供する．長単位情報は，人手修正された短単位情報を基本に自動解析した上で，コア 20 時間分を人手修正した．

係り受け情報　コアを対象に，発話単位を範囲に文節間の係り受け関係の情報を自動で付与した上で人手修正した．BCCWJ-DepPara の基準 [12] に準じ，通常の係り受けの "D"，フィラーや言いよどみなど係り先が決められないものの"F" などのラベルを付与した [13]．

談話行為情報　コアを対象に，ISO 24617-2 [14] をベースに日常会話用に整備した基準に基づき，発話単位ごとに人手で付与した [15]．各発話が担う談話機能の情報，および関係を結ぶ発話間の関係を示す依存関係情報が提供される．前者は意味的・語用論

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

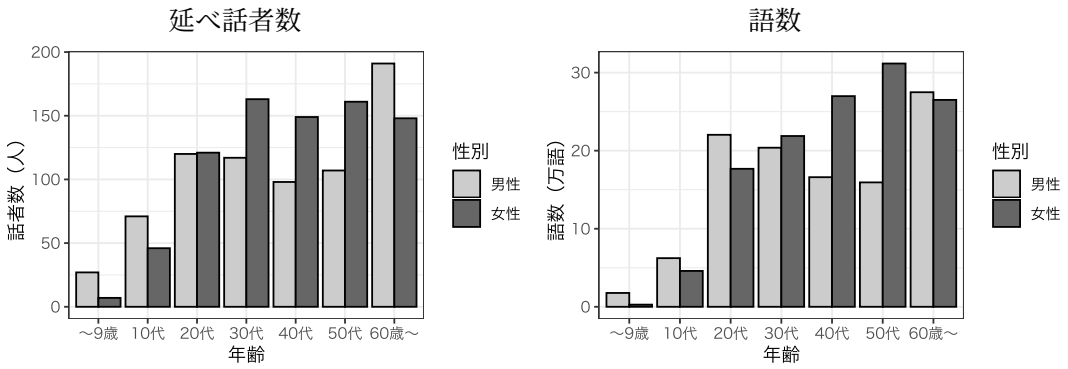

図 3　性別・年齢ごとの延べ話者数と語数（短単位）の分布

的レベルの情報（例：質問，応答，注意獲得，感謝）と，インタラクションレベルの情報（例：修復，会話の開始・終結）からなる．

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 [16] の簡略版に準拠して韻律情報を付与した [17]．アクセント句・イントネーション句の境界情報や，句末の音調などが提供される．

会話・話者に関するメタ情報　会話に関するメタ情報として，会話形式，話者数，会話が行われた場所，会話中の活動，話者間の関係性，備考情報が，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），備考情報が提供される．

検索システム　全文検索システム「ひまわり」が同梱されるほか，オンライン検索システム「中納言」での検索環境（音声再生機能付き）も提供される．「ひまわり」では観察支援システム FishWatchr を統合することで，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 [18]．

2.6　データ公開方針

映像・音声・転記テキストの公開に際し，同意書に記した条件に基づき，個人情報等の観点から次の通り加工した．話者の名前，所属組織名，自宅・所属組織の住所・電話番号，マイナンバーなどの個人識別符号，および本人が公開を希望しない箇所は，転記テキストで仮名あるいは「＊」で伏せ字化し，該当箇所の音声をビープ音で置換した．映像については，収録・公開の同意を得た話者については顔にボカシなどの処理は加えずに公開する．ただし，名札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などが写り込んだ場合については，肖像権や個人情報保護法などを参考に法的・倫理的な観点から問題を整理した上で公開方針を定め，必要と判断した箇所にボカシ処理を加えた．公開方針の詳細は [19] を参照のこと．

3　CEJC の特徴

3.1　話者の属性

100 時間の会話を対象とする『名大会話コーパス』は，現在公開されている日本語母語話者の会話コーパスの中では最も規模の大きなものだが，話者の大半が女性であり約半数が 20 代と，話者の性別・年齢に強い偏りが見られる [3]．CEJC はこうした偏りが生じないよう，2.1 節で言及した通り協力者の選定や収録法などを工夫した．そこで本節では，CEJC に収められている話者の属性（性別・年齢）に偏りがないかを検証する．

性別・年齢ごとに見た延べ話者数と語数の分布を図 3 に示す．CEJC の大半を占める個人密着法では，成人の調査協力者を中心に，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．図 3 から，20 代以上の成人について，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，いずれの世代の男女とも約 100 人以上の話者，15 万以上の語を含んでおり，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．そこで未成年者のデータ拡充のために，2022 年度より，子どもを主対象とする映像付きコーパスを構築するプロジェクトを新たに開

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

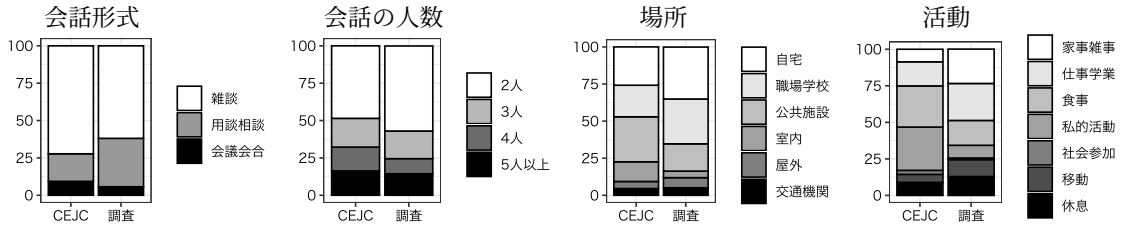

図 4　会話形式・会話の人数・場所・活動に関する CEJC と行動調査の比較：会話件数で見た場合

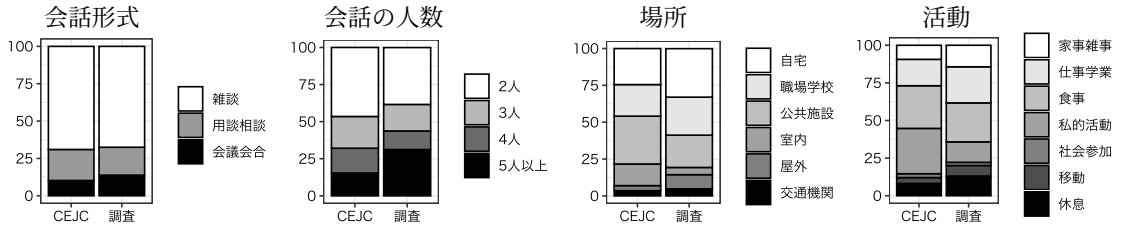

図 5　会話形式・会話の人数・場所・活動に関する CEJC と行動調査の比較：会話時間で見た場合

始する予定である [20].

3.2　会話の属性

CEJC の構築に先立ち，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，会話行動調査を実施した [21]．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．本節ではこの調査結果と比較しながら CEJC のバランスについて検証する．

図 4・5 に，会話形式・場所・活動・会話の人数に関する CEJC と行動調査の分布を示す．図 4 は会話の件数で見た場合，図 5 は会話の総時間で見た場合の比較である．なお活動については 1 つの会話に複数付きうるため重複して算出している．

会話形式　図 4 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だが，図 5 の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(1)．

会話の人数　会話形式とは逆に，会話の人数については，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 は行動調査より 5 人以上の会話の時間が短い傾向が見られる．[21] 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が，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，収録会話を上限約 1 時間としている．こうした選定基準が 5 人以上の会話時間の抑制につながっている．

場所・活動　場所と活動について，CEJC では行動調査よりも，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．2.1 節で述べたように，個人密着法での収録状況を検証し，職場での仕事中の会話が行動調査よりも少ない傾向にあったことから，特定場面法において仕事中の会議会合約 10 時間を収録した．この増補によりかなりの改善が見られたが，行動調査と同水準となるには致らなかった．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．

このように行動調査と比べて若干の差はあるものの，会話の形式，会話の人数，場所，活動の観点から CEJC は総じて多様な会話をある程度バランスよく格納していると言うことができるだろう．

4　おわりに

本稿では 2022 年 3 月に公開する CEJC の設計と特徴について報告した．オンライン検索システム「中納言」での無償公開と，映像・音声・転記・アノテーション等を含むコーパス全体の有償公開を行う．詳細については以下を参照されたい．

https://www2.ninjal.ac.jp/conversation/cejc.html

(1)　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．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞　本研究は国語研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」の成果を報告したものである．会話収録にご参加くださった皆さまに感謝します．

参考文献

[1] 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. 『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, 2017. https://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/B5-4.pdf.

[2] 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. 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. 2019. https://www2.ninjal.ac.jp/conversation/report/report03.pdf.

[3] 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. 『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, Vol. 18, pp. 17–33, 2020. http://doi.org/10.15084/00002540.

[4] S. Crowdy. The BNC spoken corpus. In G. Leech, G. Myers, and J. Thomas, editors, **Spoken English on computer: Transcription, mark-up and application**, pp. 224–235. Longman, Harlow, U.K., 1995.

[5] Lou Burnard and Guy Aston. **The BNC handbook**. Edinburgh University Press, Edinburgh, U.K., 1998.

[6] Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, and Yasuyuki Usuda. Construction of the Corpus of Everyday Japanese Conversation: An interim report. In **Proceedings of the 11th edition of Language Resources and Evaluation Conference**, pp. 4259–4264, Miyazaki, Japan, 2018. https://aclanthology.org/L18-1672.

[7] 田中弥生, 柏野和佳子, 角田ゆかり, 伝康晴, 小磯花絵. 『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, Vol. 14, pp. 275–292, 2018. http://doi.org/10.15084/00001424.

[8] JDRI. 発話単位ラベリングマニュアル version 2.1, 2017. http://www.jdri.org/resources/manuals/uu-doc-2.1.pdf.

[9] 臼田泰如, 川端良子, 西川賢哉, 石本祐一, 小磯花絵. 『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, Vol. 15, pp. 177–193, 2018. http://doi.org/10.15084/00001602.

[10] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『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）, 2011. http://doi.org/10.15084/00002855.

[11] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『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）, 2011. http://doi.org/10.15084/00002856.

[12] 浅原正幸, 松本裕治. 『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, Vol. 25, No. 4, pp. 331–356, 2018. https://doi.org/10.5715/jnlp.25.331.

[13] 浅原正幸, 若狭絢. 『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, 2022.

[14] ISO 24617-2. Language resource management — semantic annotation framework (SemAF) — Part 2: Dialogue acts, 2012.

[15] Yuriko Iseki, Keisuke Kadota, and Yasuharu Den. Characteristics of everyday conversation derived from the analysis of dialog act annotation. In **Proceedings of the 22nd Conference of the Oriental COCOSDA**, pp. 1–6, 2019. https://ieeexplore.ieee.org/document/9041235.

[16] 五十嵐陽介, 菊池英明, 前川喜久雄. 韻律情報. 日本語話し言葉コーパスの構築法, pp. 347–453. 国立国語研究所, 2006.

[17] 小磯花絵, 菊池英明, 山田高明. 『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, Vol. SIG-SLUD-B903, pp. 34–39, 2020.

[18] 山口昌也. 「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, 2018. http://doi.org/10.15084/00001668.

[19] 小磯花絵, 伝康晴. 『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, Vol. 15, pp. 75–89, 2018. http://doi.org/10.15084/00001597.

[20] 小磯花絵, 居關友里子, 柏野和佳子, 角田ゆかり, 田中弥生, 宮城信. 子どもの会話コーパスの構築に向けて. 言語資源活用ワークショップ発表論文集, Vol. 5, pp. 157–163, 2020. http://doi.org/10.15084/00003155.

[21] 小磯花絵, 土屋智行, 渡部涼子, 横森大輔, 相沢正夫, 伝康晴. 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, Vol. 10, pp. 85–106, 2016. http://doi.org/10.15084/00000810.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).