『日本語日常会話コーパス』モニター公開版の設計と特徴

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 *

* 国立国語研究所 † 千葉大学

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」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（小磯ほか 2017）。CEJC は，多様な場面・多様な話者による日常会話を，映像まで含めて記録・公開することを目指すものであり，世界的に見ても新しく画期的な取り組みである。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を把握するために，2018 年 12 月に 50 時間分の会話データをモニター公開した（以下，CEJC モニター公開版）。本稿では，CEJC モニター公開版の設計とその特徴について報告する。

2 CEJC モニター公開版の設計

2.1 コーパスの規模

CEJC では個人密着法を中心に会話を収集している（田中ほか 2018）。個人密着法は，性別・年齢などの観点からバランスを考慮して選別された協力者に収録機器等を 3ヶ月ほど貸し出し，日常生活で生じる会話を協力者自身に記録してもらうという収録法である。

CEJC モニター公開版では，2018 年 3 月の段階で収録・一次転記が終了した協力者の中から，性別・年齢のバランスを考慮して 20 名を選んだ（内訳は 2.2 節参照）。モニター公開版には，1 人平均 2.5 時間，計 50.3 時間，セッション数 116 [(1)]，会話数 126，延べ話者数 390 名，異なり話者数 237 名[(2)]，計 609,327 語（短単位）[(3)]の会話が含まれている。

[(1)]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，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・法的な観点などから問題のある部分をカットした結果，1 つのセッションが複数の会話に分かれることもある。

[(2)]集計では，当該会話に主として参加した者に限定し，店員など一時的に会話に加わった者は対象外とした。

[(3)]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。

2.2 調査協力者の内訳

表 1 に，CEJC モニター公開版が対象とする協力者 20 名の属性と，対象とする収録セッション数，会話数，会話時間，語数（短単位数）の情報を示す。収録スケジュールの都合で，40 代女性が 1 名多く 60 代女性が 1 名少ないものの，それ以外は性別・年代をバランスさせ，各スロット 2 名となっている。

2.3 データの仕様

CEJC モニター公開版は，映像・音声・転記テキスト・短単位情報・メタ情報・検索システム（映像再生機能付き）を含むハードディスクでの公開[(4)]と，短単位情報での検索と文字列検索が可能なオンライン検索システム「中納言」での公開がある。以下ではハードディスク版のデータの仕様について概説する。なお「中納言」で提供するのは短単位情報とメタ情報であり，ハードディスク版のサブセットである。

■ 映像・音声データ

会話の映像・音声の収録に用いた機材と公開のファイル形式を表 2 に示す。映像について，室内などの基本収録では，原則最大 2 種・計最大 3 台のカメラを用いて話者や会話の場の映像を中心に収録した。個々の映像だけでなく複数映像を 1 つに合成した映像も作成した（図 1 参照）。散歩など移動時の収録には 1 台のカメラを用い，周囲の様子などを中心に記録した。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中心に記録すると同時に，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。収録の詳細については田中ほか (2018) を参照されたい。

■ 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などにより区切った

[(4)]本公開では，長単位情報や係り受け情報などのアノテーションについても提供する予定である（小磯ほか 2016）。

表 1: 調査協力者の属性，対象とする収録セッション数・会話数・会話時間・語数

	男性					女性				
年代	職業・職種等	収録数	会話数	時間	語数	職業・職種等	収録数	会話数	時間	語数
20 代	大学生	5	5	2.2h	34,216	大学生	7	7	2.6h	31,645
	大学院生	5	5	2.5h	33,870	大学生	5	10	2.6h	23,817
30 代	自営業・自由業	4	4	2.8h	29,296	会社員・公務員等	5	6	2.7h	28,526
	会社員・公務員等	6	6	2.1h	31,239	専業主婦	7	7	2.8h	35,887
	会社員・公務員等	4	5	2.1h	23,081	会社員・公務員等	5	5	2.6h	27,193
40 代	自営業・自由業	6	6	2.4h	27,523	パートタイム	6	6	2.6h	33,408
						パートタイム	6	6	2.6h	31,709
50 代	会社員・公務員等	7	7	2.4h	26,750	会社員・公務員等	7	7	2.2h	22,825
	会社員・公務員等	4	4	2.6h	25,140	自営業・自由業	6	6	2.7h	32,303
60 代	その他 (非常勤講師)	9	9	2.1h	28,850	専業主婦	6	7	2.7h	34,728
以上	定年退職	6	8	3.0h	47,321					

表 2: 映像・音声データの収録に用いた機材と公開時のファイル形式

		機種名・台数	公開時のファイル形式
映像	室内などの基本収録	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
	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 1 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も公開

図 1: 合成映像の例。左の映像は Kodak PIXPRO SP360 で，右の映像は GoPro Hero3+ 2 台で録画したもの。論文掲載用に話者の顔にボカシの処理を加えている。

単位）の 2 種類の単位ごとに，CSV ファイル，EAF ファイル（映像解析ソフト ELAN 用），TextGrid ファイル（音声分析ソフト Praat 用）の 3 つのファイル形式で提供する。いずれも会話 ID，話者ラベル，開始・終了時刻，発話内容が単位ごとに記されている。発話内容は漢字仮名まじりでの表記を基本とし，表 3 に示すタグによって会話に生じる諸現象を表現している。

映像・音声・転記テキストの公開に際し，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については，転記テキストで仮名あるいは「∗」で伏せ字化した上で，該当箇所の音声をビープ音で置換した。映像については，収録時に話者と交わした同意書の条件に基き，話者の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，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。公開方針の詳細は小磯・伝（2018）を参照のこと。

■ 短単位情報

短単位情報については，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，人手による修正を加えた。語形・発音形が一意に同定できない語（例：日本「ニホン/ニッポン」）は，音を聴取した上で語形・発音形を修正した。また転記のタグを利用して得られた実際の発音（例：「けれども:」「(W ギーツ|技術)」であれば「ケレドモー」「ギーツ」）の情報を「発音」というフィールドで提供する。

表 3: 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（短単位）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（終助詞の場合を除く）	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 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，雇+われてる
(数字)	発話単位中のポーズ値 (秒)。発話単位で転記を提供する場合のみ	だってあの人でも (0.605) スケボーで通ってたんだよ。

■ 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 夏樹)っぽい，ネパール:(R ＊＊＊＊＊:) に
@	発話に対するコメント	お願いしま:す。@店員への応答

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」やオンライン検索システム「中納言」でも検索することができる。

■ 会話・話者に関するメタ情報

会話 (セッション) に関するメタ情報として，1) 会話形式，2) 話者数，3) 会話が行われた場所，4) 活動（何をしながら会話をしているか），5) 話者間の関係性の 5 種類の情報が提供される。また話者に関しては，1) 年齢（5 歳刻み），2) 性別，3) 出身地（都道府県，外国の場合は国），4) 居住地（同），5) 職業，6) 協力者からみた関係性の 6 種類の情報が提供される。このうち会話に関する「話者間の関係性」と「活動」は複数の値が付与されている。

3 CEJC モニター公開版の特徴

■ 話者の年齢・性別

『名大会話コーパス』(NUCC) は，日本語母語話者を対象とする会話コーパスの中で最も規模の大きなものであるが，話者の 85%が女性であり，48%が 20 代と，話者の性別・年齢に強い偏りが見られる。

前節で延べたように，CEJC はこうした偏りが生じないよう協力者の選定や収録法などを工夫している。CEJC モニター公開版の年齢・性別ごとの延べ話者数の分布を図 2 に示す。また表 4 に話者数・会話時間・語数の情報を示す。延べ話者数で見ると女性 54%，男性 46%であり NUCC のような偏りは見られない。年齢を見ると，成人の男性はバランスよく収録できているが，女性に関しては 40 代・50 代の話者が多く 60 代・70 代が少ない。表 4 から，同じ偏りが会話時間や語数にも見られることが分かる。2.2 節で言及したように，40 代女性の協力者が他より 1 名多く 60 代女性は 1 名少なかったために偏りが生じたと考えられる。本公開では協力者の年齢・性別をバランスさせるため，こうした偏りは補正されるものと考えられる。

一方，未成年者，特に女性については，話者数がかなり少ない。個人密着法に基づく収録調査では重い責任を伴うことから協力者は成人に限定している。そのため，未成年者の収録数は他と比べて必然的に少なくなる。特に 10 歳未満の場合，語数の少なさが際立つ。単位時間あたりの発話量が成人と比べ少ないことがうかがえる。

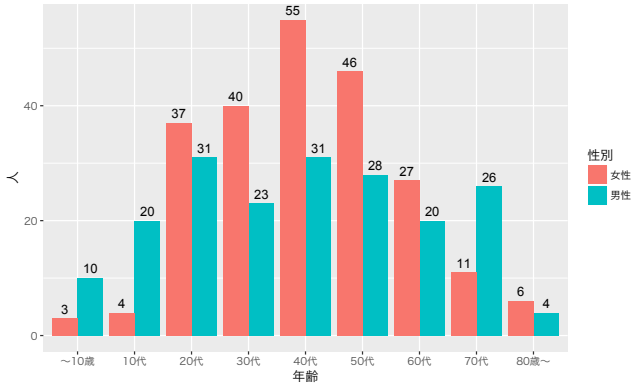

図 2: 性別・年代ごとの延べ話者数の分布

表 4: 年齢・性別ごとの話者数・会話時間・語数

年齢・性別	延べ人数	異り人数	会話時間	語数 (千語)
～10 歳・女	3 (0.7%)	2 (0.8%)	1.2 (0.7%)	1.8 (0.3%)
～10 歳・男	10 (2.4%)	4 (1.7%)	3.6 (2.2%)	4.8 (0.8%)
10 代・女	4 (0.9%)	4 (1.7%)	1.9 (1.1%)	3.5 (0.6%)
10 代・男	20 (4.7%)	8 (3.4%)	8.6 (5.2%)	19.3 (3.2%)
20 代・女	37 (8.8%)	14 (5.9%)	11.8 (7.1%)	41.2 (6.8%)
20 代・男	31 (7.3%)	20 (8.5%)	12.9 (7.8%)	60.3(10.0%)
30 代・女	40 (9.5%)	19 (8.1%)	18.1(10.9%)	64.4(10.6%)
30 代・男	23 (5.5%)	15 (6.4%)	10.8 (6.5%)	37.2 (6.1%)
40 代・女	55(13.0%)	31(13.1%)	23.5(14.2%)	97.0(16.0%)
40 代・男	31 (7.3%)	16 (6.8%)	12.5 (7.6%)	44.4 (7.3%)
50 代・女	46(10.9%)	30(12.7%)	17.0(10.3%)	81.7(13.5%)
50 代・男	28 (6.6%)	14 (5.9%)	9.9 (6.0%)	32.1 (5.3%)
60 代・女	27 (6.4%)	17 (7.2%)	10.7 (6.5%)	38.5 (6.4%)
60 代・男	20 (4.7%)	11 (4.7%)	5.8 (3.5%)	23.5 (3.9%)
70 代・女	11 (2.6%)	7 (3.0%)	4.4 (2.7%)	10.8 (1.8%)
70 代・男	26 (6.2%)	17 (7.2%)	9.6 (5.8%)	36.1 (6.0%)
80 歳～・女	6 (1.4%)	4 (1.7%)	1.5 (0.9%)	6.8 (1.1%)
80 歳～・男	4 (0.9%)	3 (1.3%)	1.7 (1.0%)	1.9 (0.3%)

表 5: 形式・話者数・場所・活動ごとの会話（セッション）数・会話時間・語数および行動調査の会話数の比率

	会話数	行動調査	会話時間	語数 (千語)
形式				
雑談	84(72.4%)	(61.9%)	36.0(71.6%)	434.5(71.3%)
用談相談	23(19.8%)	(32.4%)	11.3(22.5%)	122.0(20.0%)
会議会合	9 (7.8%)	(5.7%)	3.0 (6.0%)	52.8 (8.7%)
話者数				
2 人	44(37.9%)	(56.9%)	19.3(38.4%)	188.5(30.9%)
3 人	31(26.7%)	(18.5%)	13.9(27.6%)	179.2(29.4%)
4 人	21(18.1%)	(10.1%)	8.6(17.1%)	103.5(17.0%)
5 人以上	20(17.2%)	(14.4%)	8.5(16.9%)	138.1(22.7%)
場所				
自宅	28(24.1%)	(35.0%)	12.7(25.2%)	122.7(20.1%)
職場学校	16(13.8%)	(30.2%)	6.1(12.1%)	85.9(14.1%)
公共施設	48(41.4%)	(18.4%)	20.6(41.0%)	262.0(43.0%)
室内	15(12.9%)	(4.4%)	6.9(13.7%)	92.5(15.2%)
屋外	5 (4.3%)	(6.8%)	1.9 (3.8%)	21.0 (3.4%)
交通機関	4 (3.4%)	(5.1%)	2.1 (4.2%)	25.2 (4.1%)
活動				
仕事学業	14 (9.7%)	(25.3%)	5.7 (9.1%)	76.3(10.0%)
家事雑事等	14 (9.7%)	(23.5%)	7.6(12.1%)	66.6 (8.8%)
食事	44(30.3%)	(17.0%)	18.9(30.1%)	221.3(29.1%)
社会参加	11 (7.6%)	(1.1%)	4.4 (7.0%)	66.4 (8.7%)
私的活動	40(27.6%)	(8.7%)	17.1(27.3%)	220.0(28.9%)
移動	7 (4.8%)	(11.6%)	3.1 (4.9%)	36.1 (4.7%)
休息ほか	15(10.3%)	(12.9%)	5.9 (9.4%)	74.6 (9.8%)

■ 会話の形式・話者数・場所・活動

表 5 に，会話の形式・話者数・場所・活動ごとの会話数・会話時間・語数の情報を示す。活動は 1 つのセッションに複数付きうるため重複して算出している。参考のために，一般の人を対象に実施した会話行動調査の結果（小磯ほか 2016）も合わせて記す。

会話数の比率を行動調査の結果と比べると，形式と話者数については，若干の差はあるもののほぼ同じ傾向を示しており，バランスよく会話が集められていることが分かる。一方，場所・活動は行動調査との間に差が見られる。場所については，職場学校での会話が少なく，公民館や飲食店などの公共商業施設が多いという傾向が，また活動については，仕事学業や家事雑事等が少なく，食事や社会参加，私的活動が多い傾向が見られる。公共施設などで行われるボランティア・PTA・町内活動等の会議会合も比較的収録されているが，これらは公共商業施設における社会参加の活動に分類されており，公共商業施設や社会参加の多さの一因となっている。職場学校での仕事学業中の会話は個人密着法では収録が難しいため，それを補うためにこの種の会話を積極的にコーパスに採用している。

モニター公開版の語彙の特徴については山崎・大村 (2019) を参照されたい。

4 おわりに

本稿では 2018 年 12 月に公開した『日本語日常会話コーパス』モニター公開版の設計と特徴について概観した。本コーパスは，言語学・日本語学などに留まらず，日本語教育や情報工学，認知科学など，幅広い分野での活用が期待される。CEJC モニター公開版の詳細については以下を参照されたい。

`https://pj.ninjal.ac.jp/conversation/cejc-monitor.html`

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。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

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/`

小磯ほか (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, pp. 85–106.

小磯ほか (2017) 「『日本語日常会話コーパス』の構築」『言語処理学会第 23 回年次大会論文集』pp. 775–778.

小磯・伝 (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.

田中ほか (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.

山崎・大村 (2019) 「『日本語日常会話コーパス』モニター公開版の語彙」『言語処理学会第 25 回年次大会論文集』