『子ども版日本語日常会話コーパス』モニター版の概要

『子ども版日本語日常会話コーパス』モニター版の概要

小磯花絵（国立国語研究所）*・石本祐一（ものつくり大学／国立国語研究所）・居關友里子
江口典子・柏野和佳子・川端良子・田中真理子・田中弥生・西川賢哉（国立国語研究所）

Overview of the Monitor Version of CEJC-Child

Hanae Koiso (NINJAL), Yuichi Ishimoto (Institute of Technologists / NINJAL),
Yuriko Iseki, Noriko Eguchi, Wakako Kashino, Yoshiko Kawabata,
Mariko Tanaka, Yayoi Tanaka, Ken'ya Nishikawa (NINJAL)

要旨

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」（2022〜2027 年度）では、2022 年に公開した『日本語日常会話コーパス』（CEJC）で不足する子どものデータを補充するために、子どもを中心とする多様な場面・相手との会話を対象とする『子ども版日本語日常会話コーパス』（CEJC-Child）の構築を 2022 年度から進めている。収録対象は 8 世帯 12 名の子どもであり 100 時間規模のコーパスを構築する。このうち 50 時間のデータを対象に 2024 年度中にモニター公開する予定である。本稿では、会話の収録法とコーパスの構築状況について説明した上で、CEJC-Child モニター版の特徴について報告する。

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、日常会話 200 時間を対象とする『日本語日常会話コーパス』（CEJC）を開発し、2022 年 3 月に一般公開した（小磯ほか 2023a）。CEJC は、(1) 日常場面の中で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・公開することを特徴としており、多様な分野の研究で活用されている。しかし CEJC は成人の調査協力者を中心に会話を収集したため、未成年者、特に 10 歳未満の子どもの数がかなり少ないという問題がある。

これまでにも、CHILDES（宮田 2004）[1]や『NTT 乳幼児音声データベース』[2] など、乳幼児を対象とするコーパス・データベースが数多く構築・公開されてきた。これらのコーパス・データベースでは、乳幼児の言語発達において養育者の影響が強いことから、家庭での会話が対象とされることが多かった。しかし、子どもの成長とともに、多様な場面、多様な相手との会話がコミュニケーション行動の発達に深く影響するようになる。そのため、場面や話者の多様性を重視したコーパスの構築も求められている。また、発達研究は乳幼児に限られるものではなく、学童期、青年期、成人初期、壮年期、老年期など、多世代に渡り見ていく視点も不可欠である。成人を中心とする CEJC に、子ども中心のコーパスを新たに構築することによって、乳

* koiso@ninjal.ac.jp
1) https://childes.talkbank.org/
2) http://research.nii.ac.jp/src/INFANT.html

幼児から高齢者までの多世代を対象にコーパス言語学的手法に基づく話し言葉研究を広く推進することが可能となる。

そこで、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」（2022～2027 年度）では、子ども中心の会話コーパス 100 時間を収めた『子ども版日本語日常会話コーパス』（CEJC-Child）の構築を進めてきた。CEJC-Child は、CEJC と同様、1) 収録のために集められた状況での会話ではなく、日常場面の中で自然に生じる会話を対象とすること、2) 子どもの成長とともに広がる多様な場面における多様な話者との会話をできるだけ収録すること、3) 音声データだけでなく映像データも記録・公開することを目指している。

CEJC-Child の本公開は 2026 年度に予定しているが、コーパスの利用可能性や問題などを把握し今後の構築に活かすために、50 時間分の会話を 2024 年度中にモニター公開することを予定しており、現在、公開の準備を進めている。モニター版では、50 時間分の会話の映像・音声データ、転記テキスト、形態論情報（短単位情報）、メタ情報、関連文書やツールなどを対象とする公開と、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」での公開を予定している。本稿では、会話の収録法、コーパスの構築状況、およびモニター版対象データの特徴について報告する。

2. 会話の収録法

日常生活の中で自然に生じる会話を収録するために、CEJC の個人密着法に準拠した収録法を採用した。具体的には、表 1 に示す調査協力世帯 8 世帯に収録機材等一式を貸し出し、調査対象とする子どもを中心とする多様な場面、多様な話者との会話を収録してもらった。自然な会話を記録するため、調査者は収録に介在しなかった。協力世帯には、できるだけ毎月 1 時間程度、1～4 年程度の中長期に渡り収録してもらうよう依頼した。

対象とする子ども（収録開始時点で未就学児のもの）の属性は表 1 に示す通りである。協力世帯 Y001 の Y001_017 と Y005 の Y005_007 は、調査開始後に生まれ収録会話の一部に参加しているが、データ数も少ないことから調査対象児 12 名には入れていない。特に子どもが小さいうちは家庭での収録が中心となるため、兄弟のいる世帯や家庭がバイリンガル環境など、できるだけ家族構成などに多様性を持たせるようにしている。また、1～4 年程度に渡る調査の中で、対象とする子どもの月齢の多様性を確保するため、収録開始時期の月齢についても幅を持たせている。なお協力世帯 Y001 の父親は CEJC の調査協力者であり（協力者 ID:T001）、対象児出産前の母親とともに CEJC の会話に参加している。

収録調査の流れは CEJC に準じて行った（田中ほか 2018）。協力世帯への依頼内容は次の通りである。

1. 参加者に対して収録調査の趣旨やデータ公開の方法などについて説明
2. 参加者に対してデータ収録・公開に関する同意書への署名を依頼
3. 参加者に対してフェイスシート（話者の性別や出身地など）への記入を依頼
4. 収録の日時や使用機材、参加者等の情報を記録
5. カメラ・IC レコーダーを用いた会話の収録（毎月 1 時間程度）
6. 収録データの調査者への提出（3～6 ヶ月に 1 回程度）

調査者が収録に介在しないため、親戚や友人、知人など、会話に新たに参加する人（参加者）に対して収録調査の趣旨を説明し、データ収録・公開に関する同意をとってもらうなど、協力世帯が行う内容は多岐に渡る。収録時間は毎月 1 時間程度としたが、子育て世帯の多忙さもあり、毎月の収録を強く求めることはしなかった。

表 1　調査協力世帯・調査対象児の基本情報

世帯 ID	話者 ID	性別	収録開始時の月齢	収録期間	同居家族	備考
Y001	Y001_000 (Y001_017)*	女 (女)	2 歳 6 ヶ月 (0 歳 0 ケ月)	55 ヶ月 (31 ヶ月)	父・母	父は CEJC の調査協力者
Y002	Y002_000	女	5 歳 8 ヶ月	34 ヶ月	父・母	日韓バイリンガル
Y005	Y005_000 (Y005_007)*	女 (男)	1 歳 7 ヶ月 (0 歳 2 ヶ月)	35 ヶ月 (19 ヶ月)	父・母	日中バイリンガル
Y006	Y006_000 Y006_004	女 男	1 歳 6 ヶ月 6 歳 6 ヶ月	36 ヶ月	父・母 姉 (小学生)	
Y008	Y008_000	女	0 歳 9 ヶ月	18 ヶ月	父・母	
Y009	Y009_003 Y009_000	女 男	1 歳 2 ヶ月 4 歳 2 ヶ月	24 ヶ月	父・母	
Y010	Y010_000 Y010_003	男 男	0 歳 10 ヶ月 3 歳 8 ヶ月	41 ヶ月	父・母	
Y011	Y011_001 Y011_000	男 男	1 歳 6 ヶ月 4 歳 11 ヶ月	14 ヶ月	父・母	

∗ 調査開始後に誕生

収録には、原則として表 2 に示すカメラ 2 台、IC レコーダー 1 台を用いた（基本収録）[3)]。

表 2　基本収録で用いた機材と設定

	品名	設定	使用台数
映像	ZOOM Q2n-4K	1920 × 1080, 30fps	2
音声	SONY ICD-SX1000	リニア PCM 44.1kHz, 16bit	1

CEJC では、音声収録として、会話全体を録るために IC レコーダー 1 台を、各話者の音声を録るために人数分の IC レコーダーを使用したが、乳幼児に IC レコーダーを持たせることが難しかったことから、収録機材は上記の通り最小限に留めた。このように IC レコーダーを 1 台に減らしたため、カメラについては、音声を高精度に収録できる小型の機器を選択した。採用した ZOOM Q2n-4K は、非圧縮で高いサンプリング周波数の音声を収録することができる（設定：リニア PCM 48kHz, 24bit）。IC レコーダーによる録音音声に問題がある場合にはカメラで収音した音声も提供する。なお、機材設置の準備が間に合わない場合や、屋外での収録のためにこれらの機材を持ち出すことが難しい場合には、必要に応じてスマートフォンなど容易に利用できる機材を用いて収録してもよいこととした。協力世帯 Y001 にはスマートフォンで収録した会話が含まれている。収録の様子を図 1 に示す。

3) 協力世帯 Y001・Y002 については、収録を進めながら相談して機材を確定させた。そのため、CEJC で利用した GoPro や SP360 で収録したものが一部含まれている。また協力世帯 Y006 は家の間取りなどの都合でカメラ 3 台を用いて収録した。

図 1　収録した会話の映像の例（左は基本収録の機材で、右は iPhone で収録した映像）

3.　コーパス構築状況

2024 年 8 月 5 日現在、収録は全て終え、コーパスに格納する 100 時間分の会話を選定した。今年度公開予定のモニター版 50 時間については、協力世帯へのヒアリングを実施し、発話内容などが不明な箇所や公開の可否等に関する確認を行った。

コーパスに格納する 100 時間の会話について、CEJC と同様の基準（臼田ほか 2018）で転記テキストを作成しており、8 割程度が終了している。転記テキストに対し、2 種類の形態論情報（短単位・長単位）を自動で付与し、短単位については全体を人手修正する。また 1 割に相当する 10 時間を「コア」データセットと定め、長単位情報を人手で修正するほか、係り受け情報なども付与する予定である。短単位の人手修正については 6 割程度を終え、長単位の人手修正を開始したところである。構築は順調に進んでおり、計画通り 100 時間全体を 2026 年度中に本公開できる見通しである。

4.　モニター版のデータの特徴

モニター版として現在整備を進めているデータの規模（暫定値）は、52.6 時間、セッション数 157[4]、会話数 226、総語数（短単位）約 37 万語[5]、延べ話者数 518 名、異なり話者数 49 名[6]である。また調査対象児の収録月数、就学状況、セッション数、会話時間、実発話時間、語数の情報を表 3 に示す。

4) 協力者が 1 回に収録したものを「セッション」、セッション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「会話」と称す。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。

5) 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、品詞が歌（ハミングなど）であるものは除いた。なお喃語は品詞を「喃語」とし語数に含めた。

6) 店員など一時的に参加した話者は除いた。これらを含めると、延べ話者数 540 名、異なり話者数 61 名である。

表 3　調査対象児ごとの収録月数・就学状況・セッション数・会話時間・実発話時間・語数

話者 ID	収録月数	就学状況	セッション数	会話時間 (分)	実発話時間 (分)	語数
Y001_000	27 ヶ月	就園前～幼稚園年中	38	493	121	17,100
Y002_000	34 ヶ月	幼稚園年長～小学 3 年生	21	367	138	23,700
Y005_000	23 ヶ月	就園前～幼稚園年少	19	499	86	11,000
Y006_000	18 ヶ月	就園前	19	561	49	5,700
Y006_004	18 ヶ月	幼稚園年長～小学 2 年生	19	558	59	9,500
Y008_000	18 ヶ月	就園前	14	352	17	1,400
Y009_003	16 ヶ月	保育園	24	422	24	2,600
Y009_000	16 ヶ月	保育園	22	393	71	9,600
Y010_000	14 ヶ月	就園前	20	381	24	2,000
Y010_003	14 ヶ月	幼稚園年少～年中	22	445	103	12,800

2018 年に公開した CEJC のモニター版は、50 時間、総語数 617 万語、延べ話者数 390 名、異なり話者数 237 名である（小磯ほか 2020）。両者とも同じ 50 時間であるが、異なり話者数は CEJC が 237 名であるのに対し CEJC-Child は 49 名とかなり少ない。CEJC モニター版は調査協力者 20 名が集めた会話を対象としているのに対し、CEJC-Child モニター版は協力世帯 8 世帯と限られていること、成人中心の CEJC に比べ乳幼児も含む CEJC-Child では活動の幅が成人ほど多くないこと、また収録の多くが新型コロナウイルスの影響により外出や他者との接触を控えていた時期に重なっていたこと[7)]が影響していると考えられる。

表 4 に、調査対象児から見た関係ごとの異なり話者数、セッション数、会話時間、実発話時間、語数の情報を示す。

表 4　調査対象児から見た関係ごとの異なり話者数・セッション数・会話時間・実発話時間・語数

調査対象児から見た関係性	異なり話者数	セッション数	会話時間 (分)	実発話時間 (分)	語数
本人（調査対象児）	10	157	4,472	692	95,300
父	7	88	1,757	346	76,700
母	7	144	2,871	717	140,800
兄弟姉妹*	3	29	772	97	16,600
祖父母	7	21	262	77	15,200
その他の親戚	4	5	54	14	2,900
友だち	5	6	143	33	4,500
知人	8	10	229	75	13,800
店員等	10	9	167	2	600

∗ 調査開始後に誕生した 2 名と協力世帯 Y006 の小学生の姉。1 世帯 2 名の調査対象児を含めると話者に兄弟姉妹の関係を含むものは 157 セッション中 71 セッション。

上述の通り新型コロナウイルスの影響により、同居している家族との会話が多く、家族以外は全体の 1 割程度に留まるが、その中には、祖父母やはとこなどの親戚や、友だち、知人（例：

7) 多様な場面における多様な話者との会話を収録するという方針のもと、2019 年から会話収録を進めてきたが、小磯ほか（2023b）でも報告した通り、新型コロナウイルス感染拡大の影響により、2020 年以降、対象とする場面や話者に偏りが見られるようになった。モニター版が対象とする 157 のセッションのうち、収録年が 2019 年のものは 21 件、2020 年は 69 件、2021 年は 66 件、2022 年は 1 件である。

両親の友人）なども含まれている。

会話が行われた場所と活動の内訳を表 5 と表 6 に示す。場所についてはやはり自宅が多いが、祖父母宅や飲食店、児童館、親の職場、知人宅など自宅以外の場所で行われた会話も少なからず含まれている。また活動については、食事やおやつを食べている時や、遊んでいる時の会話が多いが、そのほかにも、料理をしている時（親と一緒にお菓子作り、料理のお手伝い）や外食、買い物をしている時の会話なども含まれている。

表 5　会話の場所

場所	件数	割合	場所	件数	割合	場所	件数	割合
自宅	167	(73.9%)	児童館*	5	(2.2%)	小売店	1	(0.4%)
祖父母宅	24	(10.6%)	親の職場	5	(2.2%)	屋外	1	(0.4%)
飲食店	19	(8.4%)	知人宅	4	(1.8%)			

* キッズルーム 1 件を含む

表 6　会話中の活動

場所	件数	割合	場所	件数	割合	場所	件数	割合
食事・おやつ	85	(37.6%)	料理	7	(3.5%)	その他*	6	(2.7%)
遊び	110	(48.7%)	外食	3	(1.3%)			
団欒	14	(6.2%)	買い物	1	(0.4%)			

* 勉強中、着物の試着、年賀状の返事書き、プレゼント開封等

年齢層別に見た語数の分布を図 2 に示す。図から、調査対象児の大半を占める 10 歳未満、および、その両親の多くを占める 30 代の語数が多いことが分かる。10 歳未満で女児の語数が男児より多いのは、コーパス全体では男児 6 名、女児 7 名と性別のバランスをとっているものの、モニター公開では男児 4 名、女児 7 名と女児の方が多いこと、また男児 4 名はいずれも兄弟姉妹がいるが女児 7 名のうち 2 名は兄弟姉妹がなく、相対的に子ども 1 人あたりの発話の機会が多いことなどが影響していると考えられる。また 30 代で女性が多いのは、世帯にもよるが母親との会話を収録する機会が父親より多いためである。

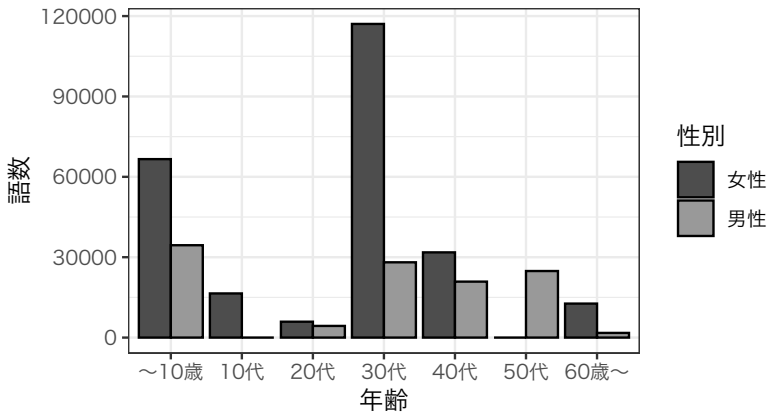

図 2　年齢層別に見た語数の分布

5. おわりに

本稿では、現在構築中の CEJC-Child の収録法と構築状況について説明した上で、2024 年度末にモニター公開を予定している 50 時間分のデータの特徴について報告した。

成人中心の CEJC と同じく、子どもの成長とともに広がる多様な場面における多様な話者との会話をできるだけ収録することを目標の 1 つに掲げたが、新型コロナウイルスの影響で自宅での家族との会話の割合が高くなった。しかし、祖父母や友だち、両親等の知人や店員など、家族以外の会話も一定数含まれている。また家族との会話についても、食事中、遊びながら、親と一緒にお料理しながらといったように、多様な活動を収めることができた。

会話の収録はすべて終了し、転記テキストの作成や形態論情報などのアノテーションも順調に進んでいる。2024 年度末の 50 時間を対象とするモニター公開、および、2026 年度末の 100 時間全体を対象とする本公開に向けて開発を進める。

謝　辞

本研究は、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」および科研費 23K25327 の成果である。

参 考 文 献

臼田泰如・川端良子・西川賢哉・石本祐一・小磯花絵 (2020)　「『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 号, pp.177-193.

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, pp.17–33.

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・渡邊友香 (2023a)「『日本語日常会話コーパス』設計と特徴」『国立国語研究所論集』24, pp.153-168

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・滕越・西川賢哉 (2023b) 「『子ども版日本語日常会話コーパス』の構築」『言語資源ワークショップ発表論文集』1, pp.103–108.

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.

宮田 Susanne 編 (2004)　『今日から使える発話データベース CHILDES 入門』ひつじ書房.