『子ども版日本語日常会話コーパス』モニター版の構築

小磯 花絵 [1]　石本 祐一 [2,1]　居關 友里子 [1]　江口 典子 [1]　柏野 和佳子 [1]
川端 良子 [1]　田中 真理子 [1]　田中 弥生 [1]　西川 賢哉 [1]

[1] 国立国語研究所　[2] ものつくり大学

{koiso,yishi,iseki,n.eguchi,waka}@ninjal.ac.jp
{kawabata,tmariko,yayoi,nishikawa}@ninjal.ac.jp

概要

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」では、子どもを中心とする多様な場面における多様な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。これは、2022 年 3 月に公開した『日本語日常会話コーパス』で不足する子どものデータを補充するために計画したものである。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が、このうち 50 時間のデータを 2025 年 3 月にモニター公開する。本稿では CEJC-Child モニター版の特徴について報告する。

1　はじめに

2022 年 3 月に一般公開した『日本語日常会話コーパス』（CEJC）[1] は、多様な場面における多様な話者との日常会話をバランスよく収めたコーパスだが、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。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、CEJC で不足する 10 歳未満の子どもを主対象とする会話 100 時間を収めた『子ども版日本語日常会話コーパス』（CEJC-Child）の構築を進めている。

これまでにも、CHILDES [2] [1] や『NTT 乳幼児音声データベース』（INFANT）[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。

そこで CEJC-Child では、家庭での会話だけでなく、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とした。またコミュニケーション行動の分析を可能とするべく、音声データだけでなく映像データも記録・公開する。この設計方針は CEJC と同じであることから、成人を中心とする CEJC と子ども中心の CEJC-Child を合わせることにより、乳幼児から高齢者までの多世代に渡る言葉の変化を分析することも可能となる。

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 時間の会話を 2025 年 3 月にモニター公開する [3]。本稿では CEJC-Child モニター版の概要について報告する。

2　コーパスの設計・構築

2.1　調査協力世帯・調査対象児

調査協力世帯と調査対象児の基本情報を表 1 に示す。収録期間の情報は CEJC-Child 全体である。モニター版については後述の表 3 にまとめる。

表 1 に示す通り、主たる調査対象児は 12 名（収録開始時点で就学前の子ども）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 名には含めていない。

CHILDES や INFANT など、言語発達の研究を志向するコーパス・データベースでは発語前の時期から収録対象とすることが多いが、CEJC-Child では子

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

表 1　調査協力世帯・調査対象児の基本情報（CEJC-Child 全体）

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

どものコミュニケーション行動の研究に資することを目指し、収録開始時点の月齢について幅を持たせるようにした。また、子どもが小さいうちは家族との会話の収録が多くなるため、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にも多様性を持たせている。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や祖母も CEJC の会話に参加しており比較が可能となっている。

2.2　会話の収録

表 1 に示した調査協力世帯 8 世帯に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。これは成人中心の CEJC における個人密着法の収録に準ずるものである [4]。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。

収録には、原則としてカメラ 2 台（ZOOM Q2n-4K, 1920 × 1080, 30fps）[3] と IC レコーダー 1 台（SONY ICD-SX1000, リニア PCM 44.1kHz, 16bit）[4] を用いた（基本収録）。ただし、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。

2.3　転記・アノテーション

収録した会話については、CEJC で採用した転記基準 [5] に準拠して文字化テキストを作成した上で、2 種類の形態論情報（短単位・長単位）を自動で付与し、人手修正をしている。また 1 割に相当する 10 時間を「コア」データセットと定め、係り受け情報なども付与する予定である。

3　モニター版の概要

3.1　コーパスの規模

CEJC-Child 全 100 時間のうち、コーパスの整備状況を考慮し協力世帯 Y011 を除く 7 世帯の収録データから、モニター版として公開する 50 時間の会話を選定した。モニター版の規模を表 2 に示す。参考のために 2018 年に公開した CEJC モニター版 50 時間の情報 [6] も合わせて示す。

表 2　CEJC-Child モニター版のデータ規模

	CEJC-Child	CEJC (参考)
会話時間	52.6 時間	50.0 時間
総語数	37 万語	61 万語
セッション数	157	116
延べ話者数	518	390
異なり話者数	49	237

CEJC-Child と CEJC のモニター版はいずれも会話時間は約 50 時間と同じだが、CEJC-Child の総語数は CEJC の 6 割程度と少ない。後述する通り、子どもは成人よりも発話率や時間あたりの語数が少ないことが関係していると考えられる（表 4）。

また CEJC の異なり話者数は 237 名であるのに対

3）　Y001・Y002 は、先行して収録を開始し相談しながら機材を決めたため、CEJC の収録に利用したカメラ GoPro や SP360 を用いたものが一部含まれている。また Y006 は家の間取りなどの都合でカメラ 3 台を用いている。

4）　CEJC では、各話者の音声を録るために人数分の IC レコーダーも使用したが、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、会話全体を収録する IC レコーダー 1 台に留めた。

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）

し、CEJC-Child では 49 名とかなり少ない。CEJC 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 は協力世帯 7 世帯と少ないこと、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広くなく家族との会話が多いこと、収録期間の多くが新型コロナウイルス感染拡大防止のため外出や他者との接触を控えていた期間にあたることなどが影響したと考えられる。

3.2　調査対象児

調査対象児ごとに、対象月齢、就学状況、会話時間、実発話時間（実際に発話した時間）、発話率（会話時間に占める実発話時間の割合）、語数（短単位数）、1 分あたりの語数の情報を表 3 に示す。いずれの子どもも参加している会話時間は 6〜9 時間程度の規模だが、実発話時間や語数にはかなり差が見られる。小さい子どもほど会話時間に占める実発話時間の割合が低く、また実発話時間 1 分あたりの語数も少ない傾向にあるためである。

3.3　モニター版の特徴

表 4 に、調査対象児から見た関係性ごとのデータ規模の情報を示す。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、調査対象児の友だち、両親等の知人なども含まれている。また子ども（調査対象児本人やその兄弟姉妹）の発話率は成人より低く、単位時間あたりの語数も少ないことが確認できる。

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、料理のお手伝いなど）や外食、買い物をしている時の会話なども含まれている。

3.4　公開データ・公開方法

CEJC-Child モニター版では、映像・音声・転記テキスト・短単位情報・メタ情報などを含む公開と、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。

4　おわりに

本稿では、現在構築中の CEJC-Child のうち、2025 年 3 月末に公開予定のモニター版について報告した。子どもの成長とともに広がる多様な場面・多様な話者との会話を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、祖父母や従姉妹、友だちなど、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。モニター版

表 3　CEJC-Child モニター版で提供する調査対象児ごとのデータ規模

話者 ID	対象月齢	就学状況	会話時間 (分)	実発話時間	発話率*	語数	語数/分
Y001_000	2 歳 6 ヶ月〜4 歳 11 ヶ月	就園前〜幼稚園年中	493	121	24.5%	17,100	141
Y002_000	5 歳 8 ヶ月〜8 歳 5 ヶ月	幼稚園年長〜小学 3 年生	367	138	37.6%	23,700	172
Y005_000	1 歳 7 ヶ月〜3 歳 6 ヶ月	就園前〜幼稚園年少	499	86	17.2%	11,000	128
Y006_000	1 歳 6 ヶ月〜3 歳 0 ヶ月	就園前	561	49	8.7%	5,700	116
Y006_004	6 歳 6 ヶ月〜8 歳 1 ヶ月	幼稚園年長〜小学 2 年生	558	59	10.6%	9,500	161
Y008_000	0 歳 9 ヶ月〜2 歳 1 ヶ月	就園前	352	17	4.8%	1,400	82
Y009_003	1 歳 2 ヶ月〜2 歳 6 ヶ月	保育園	422	24	5.7%	2,600	108
Y009_000	4 歳 2 ヶ月〜5 歳 6 ヶ月	保育園	393	71	18.1%	9,600	135
Y010_000	0 歳 10 ヶ月〜2 歳 0 ヶ月	就園前	381	24	6.3%	2,000	83
Y010_003	3 歳 8 ヶ月〜4 歳 10 ヶ月	幼稚園年少〜年中	445	103	23.1%	12,800	124

* 会話時間に占める実発話時間（実際に発話した時間）の割合

表 4　調査対象児から見た関係性ごとのデータ規模

調査対象児から見た関係性	異なり話者数	会話時間 (分)	実発話時間	発話率	語数	語数/分
本人（調査対象児）	10	4,472	692	15.5%	95,300	138
父	7	1,757	346	19.7%	76,700	222
母	7	2,871	717	25.0%	140,800	196
兄弟姉妹*	3	772	97	12.6%	16,600	171
祖父母	7	262	77	29.4%	15,200	197
その他の親戚	4	54	14	25.9%	2,900	207
調査対象児の友だち	5	143	33	23.1%	4,500	136
両親等の知人	8	229	75	32.8%	13,800	184
店員等	10	167	2	1.2%	600	300

* 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。

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

活動	件数	割合	活動	件数	割合	活動	件数	割合
食事・おやつ（自宅）	85	37.6%	料理手伝い	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

50 時間は 2025 年 3 月に、100 時間全体の本公開は 2026 年度末を予定している。

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

[1] 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. 『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, Vol. 24, pp. 153–168, 2023. https://doi.org/10.15084/00003692.

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

[3] 小磯花絵, 石本祐一, 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. 『子ども版日本語日常会話コーパス』モニター版の概要. 2024. https://clrd.ninjal.ac.jp/lrw/lrw2024/o13-paper.pdf.

[4] 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. https://doi.org/10.15084/00001424.

[5] 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, 2020. https://doi.org/10.15084/00001602.

[6] 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. https://doi.org/10.15084/00002540.