『子ども版日本語日常会話コーパス』の構築

著者	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
雑誌名	言語資源ワークショップ発表論文集
巻	1
ページ	103-108
発行年	2023
URL	http://doi.org/10.15084/00003729

『子ども版日本語日常会話コーパス』の構築

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（国立国語研究所）
滕越（東京大学大学院総合文化研究科／国立国語研究所）・西川賢哉（国立国語研究所）

Construction of CEJC-Child

Hanae Koiso, Haruka Amatani, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino,
Yoshiko Kawabata, Yayoi Tanaka (NINJAL)
Yue Teng (The University of Tokyo / NINJAL), Ken'ya Nishikawa (NINJAL)

要旨

2022 年 3 月に公開した『日本語日常会話コーパス』(CEJC) は、成人中心のコーパスであり、未成年者、とくに 10 歳未満の子どもの会話はあまり含まれてい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、子どもを中心とする多様な場面・相手との会話を含む映像付きコーパスを新たに開発し、成人中心の CEJC と接続させることにより、コミュニケーションを含む言語の発達・変化の過程を、子どもから高齢者まで多世代に渡り実証的に研究できる基盤を構築することを目指している。発表では、新たに構築する子ども版の日常会話コーパスの設計や収録状況について報告する。

1. はじめに

これまで乳幼児を含む子どもの言語発達に関する研究が数多く行われてきたが、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。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、多様な話者による多様な場面の日常会話 200 時間をバランスよく集めた『日本語日常会話コーパス』（CEJC）を開発し 2022 年 3 月に一般公開したが、このコーパス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。乳幼児から高齢者までの多世代を対象とする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には、子どもを対象とする会話コーパスが不可欠である。

子どものデータの記録と共有化については、これまで CHILDES[(1)]などを中心に積極的に進められてきた（宮田 2004）。また幼児とその両親の自然発話 500 時間以上を集めた大規模な『NTT 乳幼児音声データベース』も 2008 年に公開された[(2)]。

乳幼児の言語発達において養育者、特に母親の子どもに対する影響が強いことから、既存の

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

コーパス・データベースでは家庭での母子間会話が対象とされることが多かった。しかし成長するにつれ、友達や幼稚園の先生との会話といったように、多様な場面・相手との会話が言語の発達に深く影響するようになる。そのため、子どもの成長とともに広がる多様な場面・相手との会話を含むコーパスの構築が求められている。

こうした状況を受け、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、子どもを中心とする映像付き会話コーパスを新たに開発し、成人中心の CEJC と接続させることにより、コミュニケーションを含む言語の発達・変化の過程を、子どもから高齢者まで多世代に渡り実証的に研究できる基盤を構築することを目指して活動している。

本プロジェクトで構築するコーパスは幾つかあるが、その中で最も力を入れているのは、子どもを中心とする多様な場面・相手との日常会話を経年的に収録した『子ども版日本語日常会話コーパス（CEJC–Child）』（仮）である。そのために、2019 年度より収録の準備を進めてきた。本発表では、コーパスの設計や収録状況について報告する。

2. コーパスの基本設計

2.1 設計方針

成人中心の CEJC と合わせて研究できるようにするために、設計方針も CEJC を踏襲し、次の方針のもとでコーパスを構築することとした。

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。
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。
3. 音声データだけでなく映像データも記録・公開する。

このような方針のもと、2019 年度より順次、会話収録を進めてきたが、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。これについては 3 節で状況を報告する。

2.2 会話の収録

■収録法　日常場面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、3 節で述べる調査協力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・話者との会話を収録してもらうこととした。自然な会話を記録するため、調査者が収録に介在しないという点も CEJC と同じである。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。

■収録の流れ　収録調査の流れについても、原則として CEJC に準ずることとした（田中ほか 2018）。協力世帯に依頼することの概要は次の通りである。

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）

先述の通り、調査者は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、その内容は多岐に渡る。

新たな参加者がいる場合には、上記の 1〜3 を行う必要がある。収録の都度、4 と 5 を行う。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を考慮し、厳密に徹底するよう依頼することはしなかった。6 については、録音・録画データのほか、4 の収録の記録、及び、2 の同意書、3 のフェイスシート（新規参加者がいる場合）をまとめて提出してもらう。

■収録機材　収録には、原則として表 1 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 1　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、映像収録用にカメラを 3 台、また音声収録用に会話全体を録るための IC レコーダー 1 台と各話者の音声を録るための人数分の IC レコーダーを使用するといったように、多くの機材を用いる収録方法を採用したが、子育て世帯の負担を軽減するため、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）。1 台の IC レコーダーと 2 台のカメラを分散して配置することによって、例えば部屋の中で遊びながら移動するような場合などであっても、各話者の音声をある程度収音することが可能となる。

また、表 1 の機材を基本としつつも、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。

収録の様子を図 1 に示す。

2.3　コーパスの規模

コーパスの規模（目標）は 100 時間である。収録したデータの中から、会話や話者のバランス、データの質や倫理的・法的な問題などを考慮し、コーパスに格納するデータを選別する。

(3) 収録を進めながら、協力世帯から意見をもらい、機材を確定させた。そのため、GoPro など異なる機材で収録したものが一部含まれている。また家の間取りなどの都合により、カメラを 3 台用いる世帯もある。

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）

これまでに収録したデータから、コーパスの規模は、230 セッション、340 会話[4]、延べ話者数 900 人、総語数 120 万語程度になると予想される。

2.4　転記テキスト・アノテーション

コーパスに格納する 100 時間の会話について、原則 CEJC と同様の基準で転記テキストを作成する（臼田ほか 2018）。転記テキスト作成の際、発話単位（JDRI 2017）も認定する。その上で、2 種類の形態論情報（短単位・長単位）を自動で付与する。また 1 割に相当する 10 時間を「コア」データセットと定め、形態論情報を人手で修正するほか、人手で談話行為情報（Iseki, et al. 2019）も付与する計画である。

3.　収録状況

表 2 に示す 8 世帯 14 名の子どもを対象に収録を進めている。子ども会話の収録では、特に子どもが小さいうちは家庭での収録が中心となる。そのため、兄弟のいる世帯や家庭がバイリンガル環境の場合など、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 年程度にわたり経年的に調査する予定であるが、その中で対象とする子どもの年齢の多様

(4) 協力者が 1 回に収録したものを「セッション」と、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」と称する。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。

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。

表 2　対象児・協力世帯の情報（2022 年 8 月現在）

世帯 ID	性別	収録開始時点の月齢	収録開始時期	収録期間 (予定)	同居家族	備考
Y001	女	2 歳　6 ヶ月	2019 年 6 月	48 ヶ月	父・母	
	女	0 歳　1 ケ月	2021 年 8 月	22 ヶ月		
Y002	女	5 歳　7 ヶ月	2019 年 6 月	36 ヶ月	父・母	日韓バイリンガル
Y005	女	1 歳　7 ヶ月	2020 年 1 月	36 ヶ月	父・母	日中バイリンガル
	男	0 歳　2 ヶ月	2021 年 5 月	20 ヶ月		
Y006	男	6 歳　6 ヶ月	2020 年 1 月	36 ヶ月	父・母・	
	女	1 歳　6 ヶ月			姉 (小学生)	
Y008*	女	0 歳　9 ヶ月	2020 年 6 月	18 ヶ月	父・母	
Y009*	男	4 歳　2 ヶ月	2020 年 6 月	25 ヶ月	父・母	
	女	1 歳　2 ヶ月				
Y010	男	3 歳　8 ヶ月	2020 年 9 月	30 ヶ月	父・母	
	男	0 歳 10 ヶ月				
Y011*	男	4 歳 11 ヶ月	2021 年 2 月	14 ヶ月	父・母	
	男	1 歳　6 ヶ月				

* 世帯 ID に “*” を付けた 3 世帯は 2022 年 8 月現在で収録を終えている

2.1 節で言及したように、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 6 月から会話収録を進めてきたが、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。

図 2 は、2019 年 6 月から 2021 年 12 月末までに収録された会話を対象に、会話の収録場所および会話相手との関係性の割合を年毎に示したものである(5)。図から、2019 年は自宅での収録だけでなく、親類宅など自宅以外の室内や飲食店などの公共商業施設での収録も多く、会話の相手も家族だけでなく親戚や友人知人なども一定数見られたのに対し、2020 年、2021 年は自宅での収録が多くを占め、会話の相手も家族の割合が高くなっていることが分かる。

こうした偏りはあるものの、調査者が収録に介在せず協力世帯に収録を依頼したことが幸いし、収録自体は停滞せずに進めることができた。

なお、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、幼稚園や小学校での会話を対象とするコーパスも別に構築する予定である。CEJC–Child で不足する友達や先生との会話は、これらのコーパスで補填することができると考えている。

4. おわりに

本発表では、現在構築中の『子ども版日本語日常会話コーパス』の設計と収録状況について報告した。成人話者を中心とする CEJC と接続させ、子どもから高齢者まで長期に渡り実証的に研究できる基盤を構築することを目指し、設計や収録法、アノテーション等はできるだけ CEJC を踏襲することとした。収録状況としては、新型コロナウイルスの影響で、2020 年以

(5) あくまで収録されたデータの分布であり、そこから選別してコーパスに格納するデータの分布ではないことに注意する必要がある。

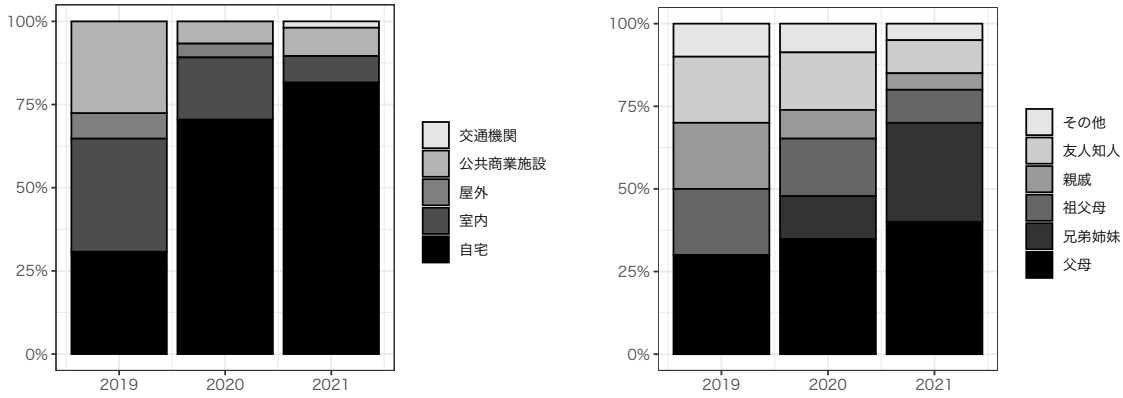

図 2　収録された会話の収録場所・相手との関係性

降、自宅での家族との会話の収録の割合が高くなっているが、協力世帯に収録を依頼していることから、収録自体は順調に進んでいる。2025 年頃の公開を目指して開発を進める。

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 20H01264 による成果である。

参 考 文 献

Yuriko Iseki, Keisuke Kadota, and Yasuharu Den(2019)　“Characteristics of everyday conversation derived from the analysis of dialog act annotation”, *Proceedings of the 22nd Oriental COCOSDA*, pp.16.

JDRI　(2017)『発話単位ラベリングマニュアル』version 2.1, 2017.

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・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 (2022)「『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』pp.2008-2012.

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.