学会記事

Non Peer-Reviewed

# 『日本語日常会話コーパス』本公開版の構築

小磯　花絵†

## 1　はじめに

これまで多くの会話コーパスが構築・公開されてきたが（表 1），大学生などの若者や親しい者同士の会話，職場での会話，電話会話といったように，対象とする話者や場面などに偏りが見られる．また実際の日常場面の会話ではなく，収録のために集まって雑談してもらうといったように，作られた場面の会話も多い．更に音声データを公開していないコーパスも少なからずあり，映像データまで含むコーパスはほとんど存在しない．しかし我々が普段用いる言語の実態を調べるには，実際の日常場面の会話を対象とするコーパスが不可欠である．また非言語行動まで含め総合的にコミュニケーション行動の仕組みを明らかにするには，映像データまで含めたコーパスが求められる．

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「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，実際の日常場面を対象に，多様な場面・

| コーパス名 | 規模 | 概要 | メディア |
|---|---|---|---|
| 名大会話コーパス | 129 会話 100 時間 | 親しい者同士の雑談 | 無 |
| BTS による多言語話し言葉コーパス 2021 年度版 | 446 会話 112.5 時間 | 友人同士の雑談，教師学生面談会話，電話会話など | 音声（一部） |
| 日本語話題別会話コーパス J–TOCC | 1800 会話 150 時間 | 120 組の親しい大学生同士が 15 種類の話題別に 5 分ずつ会話 | 無 |
| CABank 日本語 Sakura コーパス | 18 会話 7.5 時間 | 大学生の会話 | 映像 |
| 千葉大学 3 人会話コーパス | 12 会話 2 時間 | 大学生の友人同士の会話 | 音声 |
| CALL HOME Japanese | 120 会話 20 時間 | アメリカ在住日本人と国内の家族・友人との電話会話 | 音声 |
| CallFriend Japanese | 60 会話 | アメリカ在住の日本人同士の電話会話 | 音声 |
| 談話資料 日常生活のことば | 96 会話 18 時間 | 日常生活の会話 | 無 |
| 女性のことば・職場編<br>男性のことば・職場編 | 111 会話 21 時間 | 職場のフォーマル・インフォーマルな場面の自然談話 | 無 |

**表 1**　主要な日本語の会話コーパス

† 国立国語研究所

話者による会話 200 時間を含む『日本語日常会話コーパス』(The Corpus of Everyday Japanese Conversation, 以下 CEJC) の構築を進めてきた．CEJC は音声だけでなく映像まで含めて公開するが，日常場面の会話を 200 時間という規模で映像まで含めて公開するというのは，世界的に見ても新しい取り組みである．そのため，いかに自然な会話を収録するか，また倫理的・法的にどのような問題があるかなど (小磯，伝 2018)，多くの問題と向き合いながらコーパスの構築に取り組んできた．2018 年 12 月に 50 時間分の会話を対象とする試験公開を行い (小磯 他 2020a)，プロジェクトの最終年度にあたる 2022 年 3 月末に 200 時間全体を本公開する．本稿では CEJC 本公開版について紹介する．

## 2 CEJC における会話の収録

多様な場面・話者による会話を対象とするために，CEJC では個人密着法と特定場面法に基づき会話を収集した．個人密着法では，年齢・性別のバランスをとった 40 名の調査協力者（20 代・30 代・40 代・50 代・60 歳以上の男女，各 4 名）にカメラや IC レコーダーなどの収録機材を貸し出し，できるだけ多様な場面・話者との会話約 15 時間を収録してもらった．その上で，会話の種類や話者のバランスなどを考慮して，1 協力者あたり約 4〜6 時間，計 185 時間のデータを選別した．また個人密着法で収集した 50 時間の会話を調査したところ，職場での仕事中の会話や未成年者による会話などが十分に収集できていないことから (小磯 他 2020a)，こうした不足する場面や話者の会話については，調査者が主体となり会話を収録する特定場面法で 15 時間のデータを収録した．

協力者が実際に収録した会話の映像の例を図 1 に示す．

**図 1** 映像データの例：夫婦で料理をしている場面．左の映像は 360 度撮影可能な Kodak PIXPRO SP360 を会話の場の中央に配置し話者を中心に撮影したもの．右・上下の映像は GoPro Hero3+を 2 台設置し話者や会話の状況を俯瞰的に記録したもの．論文掲載用に話者の顔にボカシ処理をしている．

図 1 にあるように，原則として 3 台のカメラを用いて会話の映像を記録した．また音声についても，会話の場の中心に設置する IC レコーダーで会話全体の音声を録音すると同時に，個々の話者が装着する IC レコーダーにより各話者の音声をより明瞭に録音した．このように，多様な研究ニーズに応えるために複数の映像・音声データを提供する点も，従来の会話コーパスにはない特徴と言える．

## 3　データの規模・内訳

CEJC 本公開版には，461 セッション[1]，577 会話，計 200 時間のデータが含まれている．話者数は延べ 1675 人，異なり 862 人，語数は約 240 万語（短単位）である[2]．

図 2 に，性別・年齢ごとの延べ話者数と語数の分布を示す．CEJC の中核をなす個人密着法では，20 歳以上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たため，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．実際，図 2 を見ると，20 代以上の成人については，40 代・50 代の男性が若干少なく女性が多いなど多少の違いはあるものの，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，未成年者の会話が少ないという予備調査を受け，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ものの，成人と比べると数は少ない．未成年者のデータ拡充のために，子どもを主対象とする映像付きコーパスの構築を新たに進めているところである (小磯他 2020b)．

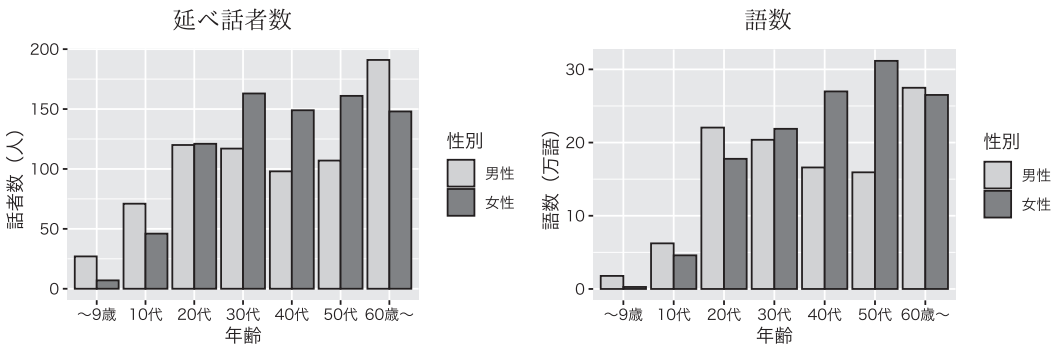


**図 2**　性別・年齢ごとの延べ話者数と語数（短単位）の分布

[1] 協力者が 1 回に収録したもの（「(収録）セッション」）からまとまりをもった範囲を「会話」として切り出した．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] 原稿執筆時点（2021 年 12 月）のデータ整備状況に基づく値．本公開版のデータと若干値が異なる可能性がある．また話者数について，大人数の会議などでは発話を一切していない話者もいるが，こうした話者は集計に含めていない．逆に店員など一時的に会話に参加している者であっても，発話している場合は含む．

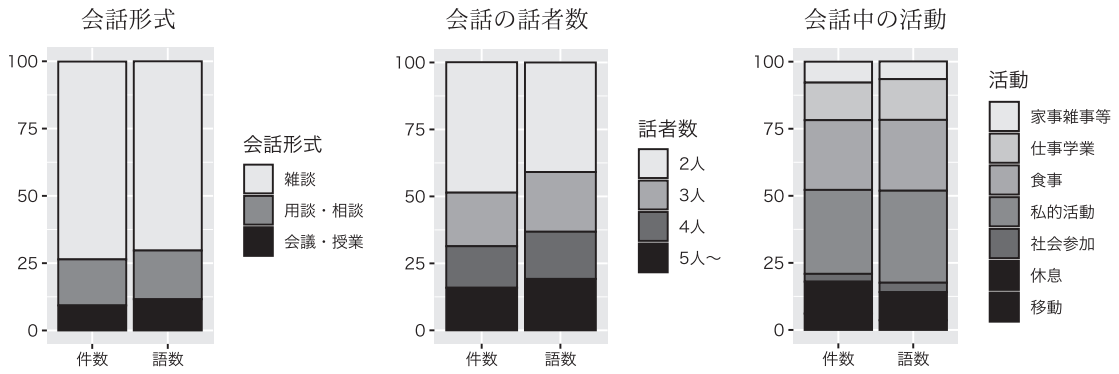


**図 3**　会話形式・会話の話者数・会話中の活動別に見た会話数・語数の割合

また図 3 に，会話形式・会話を構成する話者数・会話中の活動別に見た会話数・語数の割合を示す．会話形式を見ると，会話の件数・語数ともに，雑談が約 75%と多くを占めているが，用談・相談や会議会合・授業レッスンなども一定数含まれていることが分かる[3]．また会話の話者数を見ると，2 人会話が件数の上では約半数と多くを占めているが，5 人以上の多人数会話も件数で約 15%，語数で 20%と，少なからず含まれている．また活動を見ると，食事中の会話や友人との雑談など私的活動中の会話が多いものの，家事雑事や仕事学業中の活動，車や徒歩等での移動中の会話など，多様な状況での会話が含まれていることが分かる．

## 4　研究用付加情報

CEJC では，200 時間全体に対して転記テキストや短単位情報などの基本的な研究用付加情報（アノテーション）を付与するとともに，「コア」と定める 20 時間のデータ範囲については更に係り受けや談話行為情報なども付与する．CEJC 本公開版で提供する研究用付加情報は次の通りである．

**転記テキスト**　200 時間全体に対し，ELAN・Praat を用いて映像・音声を参照しながら人手で作成．語の言いさしや母音・子音の延伸，笑いなどに関するタグを施す (臼田 他 2018)．

**短単位情報・長単位情報**　BCCWJ の単語・品詞設計に準じて短単位情報 SUW・長単位情報 LUW を付与 (小椋 2014)．SUW は転記テキストを対象に MeCab・UniDic で自動解析した上で，200 時間全体を対象に人手修正．LUW は SUW をベースに自動解析した上で，コアを対象に人手修正．

[3] ただし会議会合・授業レッスンの枠の多くは会議会合であり．授業レッスンは少ない．

**係り受け情報**　コアを対象に，転記テキストで定める発話単位を範囲として文節間の係り受け情報を自動付与した上で人手修正.

**談話行為情報**　コアを対象に，国際標準化規格 ISO24617-2 をベースに日常会話用に整備した基準にもとづき，発話単位ごとに人手で付与 (Iseki et al. 2019).

**韻律情報**　コアを対象に，CSJ に適用したラベリングスキーム X–JToBI(五十嵐 他 2006) を簡略化した簡易版 X–JToBI(小磯 他 2020c) に準拠して人手で付与.

上記情報は，音声・映像データや会話・話者に関するメタ情報などと合わせて有償で公開する．また短単位情報・長単位情報については，オンライン検索システム「中納言」にて無償で公開する.

## 謝　辞

本稿で紹介したコーパスは，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」の研究成果である．コーパスの収録にご協力・ご参加くださった皆さまに深く感謝します.

## 参考文献

五十嵐陽介，菊池英明，前川喜久雄 (2006)．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. [Y. Igarashi et al. (2006). Prosodic Information. Construction of the Corpus of Spontaneous Japanese, pp. 347–453. The National Institute for Japanese Language and Linguistics.].

Iseki, Y., Kadota, K., and Den, Y. (2019). “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.

小磯花絵，伝康晴 (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. [H. Koiso and Y. Den (2018). A Guideline on the Release of the Corpus of Everyday Japanese Conversation: From the Viewpoint of Legal and Ethical Issues. NINJAL Research Papers, 15, pp. 75–89.].

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 (2020a)．『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. [H. Koiso et al. (2020a). Design, Evaluation, and Preliminary Analysis of the Monitor Version of the Corpus of Everyday Japanese Conversation. NINJAL

Research Papers, 18, pp. 17–33.].

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 (2020b). 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, **5**, pp. 157–163. [H. Koiso et al. (2020b). Towards the Construction of a Japanese Children's Corpus. Proceedings of Language Resources Workshop, 5, pp. 157–163.].

小磯花絵，菊池英明，山田高明 (2020c). 『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, **SIG-SLUD-B903**, pp. 34–39. [H. Koiso et al. (2020c). Intonation Labeling for the Corpus of Everyday Japanese Conversation: The Labeling Scheme and Prosodic Features of Everyday Conversation. JSAI Technical Report, SIG-SLUD-B903, pp. 34–39.].

小椋秀樹 (2014). 形態論情報. 書き言葉コーパス：設計と構築, 講座日本語コーパス, 2 巻, pp. 68–88. [H. Ogura (2014). Morphological Information. Corpus of Written Japanese: Its Design and Structure. Series of the Japanese Corpus, Vol. 2, pp. 68–88.].

臼田泰如，川端良子，西川賢也，石本祐一，小磯花絵 (2018). 『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**, pp. 177–193. [Y. Usuda et al. (2018). Criteria and Composition Method of Transcription for the Corpus of Everyday Japanese Conversation. NINJAL Research Papers, 15, pp. 177–193.].

## 略歴

**小磯　花絵**：奈良先端科学技術大学院大学情報科学研究科博士後期課程修了（1998年）．博士（理学）．1998 年国立国語研究所入所．現在，同研究所研究系音声言語研究領域教授．専門は談話分析，コーパス言語学.