『日本語日常会話コーパス』の構築

小磯 花絵 *　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　† 千葉大学文学部

1 はじめに

国立国語研究所では，『日本語話し言葉コーパス』（CSJ）や『現代日本語書き言葉均衡コーパス』（BCCWJ），『国語研日本語ウェブコーパス』（NWJC）など，大規模なコーパスの構築・公開を進めてきた。特に，現代日本語の書き言葉の全体像を把握するために構築された 1 億語からなる BCCWJ やウェブを母集団とする 100 億語規模の NWJC の公開により，多様なレジスターを考慮した現代日本語書き言葉の研究をコーパス言語学の手法に基づき研究する環境が整備され，辞書編纂への活用や日本語学習者・日本語教師の利用など，基礎研究に留まらない広がりを見せている。

話し言葉については，CSJ の構築・公開により，話し言葉の言語学的・音声学的な研究や音声情報処理研究を支える基盤は整えられたと言えよう。しかし，CSJ は独話を主対象とするコーパスであり，日常生活の中で交わされる会話は含まれていない。我々は日常生活の中でどのような言葉を使い，人といかなる仕組みでコミュニケーションしているのか，また日常場面でのさまざまな活動を言葉や身体を用いていかに組織化しているのかなど，問うべき課題は多い。こうした研究を支える基盤として，実際の日常会話場面を対象とした大規模な会話コーパスの構築が求められている。

このような状況を受け，国立国語研究所では，今年度より機関拠点型基幹研究プロジェクト「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」（平成 28〜33 年度）を開始した[*1]。このプロジェクトは，さまざまなタイプの日常会話 200 時間をバランス良く収録した大規模な日常会話コーパスを構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性を，「レジスター」「経年変化」「相互行為」の観点から多角的に解明することを目指すものである。

本稿では，プロジェクトで構築する日常会話コーパスの基本設計および構築状況について報告する。

2 コーパスの基本設計

本節では，プロジェクトで構築する『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation, CEJC）の基本設計について概説する。

2.1 会話の種類の内訳

我々の日常会話行動を正確に記述し，その本質を解明するには，日常会話の幅広いレジスターをカバーするようサンプルを選ぶことが求められる。しかし話し言葉の場合，実際にどのようなレジスター的な広がりがあるかを把握すること自体，重要な課題である。

そこで，日本語母語話者が日常的に交わす会話の実態をとらえてコーパス設計に活かすために，平成 24 年度に予備調査として，成人約 250 人を対象に，起床から就寝までの間に行った全ての会話について，いつ，どこで，誰と，何をしながら，どのような種類の会話を行ったか，などを問う会話行動調査を実施した。調査結果を踏まえ，会話の種類の内訳の目安を，会話の形式・活動・場所の観点から求めた（調査の詳細は小磯ほか (2016) を参照のこと。また 3 節でも部分的に調査結果に触れる）。コーパスを構築するにあたり，調査で求めた比率を厳密に遵守するのではなく，この値を一つの目安にデータを選定することによって，多様な種類の会話を納めたコーパスを構築できると考えている。

2.2 会話の収録法

CEJC が対象とするのは，収録のために集められた状況での会話ではなく，日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話（naturally occurring conversation）である。こうした会話をバランスよく収録するために，次の二つの方法でデータを

[*1] http://pj.ninjal.ac.jp/conversation/

表 1　コーパスの規模

時間（目標値）	200 時間
語数（推定値）	200 万語
会話数（推定値）	400 会話
延べ話者数（推定値）	1200 人
異なり話者数（推定値）	600 人

収録する。

個人密着法　性別・年代の点から均衡性を考慮して選別されたインフォーマント（以下，協力者）に収録機材等を貸し出し，協力者自身に日常会話を収録してもらう方法。調査者は収録に介在しない。

特定場面法　職場での会議や接客場面の会話など，個人密着法では収録が難しいと思われる場面を，調査者が主体となり収録する方法。調査者は介在するが，日常場面で自然に生じる会話を対象とする。

現在は個人密着法に基づく収録を進めている。会話収録の詳細については田中ほか (2017) を参照のこと。

2.3　コーパスの規模・構成

2.3.1　コーパスの規模

本プロジェクトで構築するコーパスの規模は，200 時間を目標とする。これまでに収録・転記したデータにもとづき，コーパスの総語数，会話数，および会話者数（延べ・異なり）の推定値を求めた。表 1 にまとめて示す。

2.3.2　個人密着法・特定場面法の構成

個人密着法に従い，首都圏に在住の協力者 40〜50 人（男女 × 20 代・30 代・40 代・50 代・60 代以上 × 各 4〜5 人）を対象に，それぞれ 15〜18 時間程度，計約 600 時間の会話を収録してもらう。収録データの中から，均衡性や倫理的問題，データの質などを考慮し，コーパスに格納・公開するデータとして，各人約 4〜5 時間分の会話，計 160〜200 時間を選定する。また個人密着法による会話の種類を調査し，収集の難しい種類の会話については，特定場面法での収録で補填する。特定場面法での収録対象や分量などは，個人密着法の収録状況を見て判断する。

2.4　コーパスの階層的な構成

コーパスに格納する 200 時間の会話のうち，協力者 20 人，各 2.5 時間，計 50 時間を対象に，平成 30 年度にモニター公開することを予定している。またモニター公開データの中から 20 時間を選定し，「コア」データ

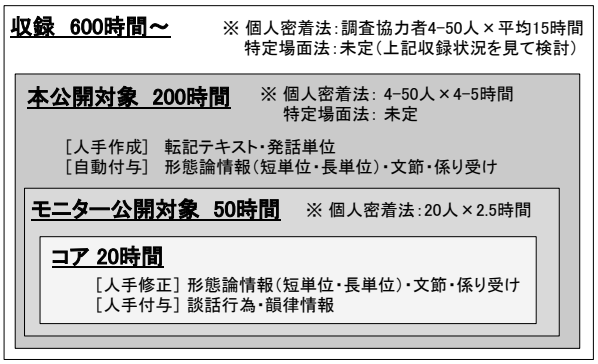

図 1　コーパスの構成

（人手による高精度なアノテーションが付されたデータ範囲）として整備する予定である（図 1 参照）。

2.5　研究用付加情報

会話の映像・音声を収録した上で，次の研究用付加情報（アノテーション）を付与する予定である。

■転記テキスト　ELAN や Praat などを用い，映像・音声を参照しながら，人手で転記テキストを作成する。転記の過程で，語の言いさしや言い誤り，母音や子音の引き延ばしなどに関するタグ付けを行う。タグの設計は，『千葉大学三人会話コーパス』の転記の仕様を参考に定めた。転記の仕様や具体的な作業の流れについては，川端ほか (2017)・臼田ほか (2017) を参照のこと。

■発話単位情報　「長い発話単位」(JDRI, 2017) に準拠して発話単位を認定する。長い発話単位は，話し手と聞き手が行為や情報を交換する際の基本単位に相当し，統語的・談話的・相互行為的な一まとまりとして定義される。転記を作成する段階で，人手で発話単位を同定する。発話単位は後述の自動形態素解析の入力単位となるなど，他の情報付与にも関わる。

■形態論情報（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・長単位情報を付与する。日常会話には語の縮約などくだけた表現が多く見られるため，その使いについては体系的な基準の整備が必要となる。コーパス全体に対して自動で解析したのち，コアについては人手で修正する。短単位は形態素解析器 MeCab と電子化辞書 unidic を用いて解析する。

■文節間の係り受け情報　コーパス全体に対し，発話単位を範囲に文節間の係り受け関係の情報を自動で付与する。コアについては人手で修正する。

■談話行為情報　国際標準化規格 ISO24617-2 に基づき，日常会話用に整備した基準により，コアを対象に発話単位ごとに人手で付与する。現在，基準の確定に向け，『千葉大学三人会話コーパス』を対象とするアノテーションの試行を行っている。詳細は居關ほか (2017) を参照のこと。

■韻律情報　コアのうち，録音状態や方言の度合などに基づき選別した会話を対象に，CSJ 構築の際に整備したラベリングスキーム X-JToBI を簡略化した「簡易版 X-JToBI（仮称）」(五十嵐, 2015) に準拠して人手で付与する。

3　コーパスの構築状況

3.1　会話収録

2016 年 4 月より収録を開始し，2017 年 1 月 10 日現在，13 人が調査を終え，6 人が調査中である。協力者の年齢・性別・職種の内訳を表 2 に示す。

平成 30 年度に予定しているモニター公開は，来年度初旬までに調査を終了する協力者のデータから選定する予定であるため，できるだけ協力者の属性が偏らないよう配慮している。ただし，現時点で 50 代が男女 1 人ずつと他に比べて少ない。来年度初旬までに 50 代の協力者の調査を進める予定である。

調査を終了した 13 人の協力者によって収録された全データの規模は，203 時間，計 240 会話，会話者数（延べ）751 人，会話者数（異なり）301 人である。

3.2　コーパス格納データの選定

コーパスに格納する会話を選定するにあたり，当該協力者が収集する会話（平均 18 会話）の中で，できるだけ多様な種類・場面の会話となるようにする。具体的には，「家族との会話」「友人との会話」「仕事関係者との会話」といったように，大きく 3〜4 種類に分類した上で，その中から，会話の形式，会話が行われる場所や活動，会話者の構成などが偏らないように会話を選定する。例えば家族との会話の場合，自宅での食事場面だけでなく，子どもの宿題を見る場面や親戚を交じえての外食場面なども含めるといったように，バランスをとるよう配慮する。

調査を終了した 13 人のうち 10 人については，コーパスに格納する会話の選定をほぼ終えている。規模はコーパス全体の 1/4 に相当する計 50 時間（1 調査者あたり平均 5 時間）であり，97 会話（1 調査者あたり平均 10 会話），会話者数（延べ）290 人，会話者数（異なり）143 人である。この 50 時間 97 会話を対象に，会話の形式・場所・活動の内訳を求めた。会話行動調査の結果と合わせて図 2 に示す。

表 2　協力者の属性（2017 年 1 月 10 日現在）

	男性	女性	計
20 代	学生（終了） 学生（調査中）	学生（終了） 学生（終了）	4 人
30 代	自営自由業（終了） 自営自由業（終了）	専業主婦（終了） 会社員等（終了） 会社員等（終了）	5 人
40 代	自営自由業（終了） 会社員等（調査中）	会社員等（終了） 自営自由業（調査中） 専業主婦（調査中）	5 人
50 代	自営自由（調査中）	自営自由業（終了）	2 人
60 代〜	無職（終了） 非常勤講師（調査中）	専業主婦（終了）	3 人
計	9 人	10 人	19 人

まず会話の形式について見る。会話行動調査の結果では雑談が約 60% であるのに対し，コーパス格納予定のデータでは約 70% と，少し雑談が多いものの，概ねバランスよくデータが選定できていることが分かる。収録された会話全体では，雑談の比率が 8 割弱と多くなりがちだが，用談・相談や会議・会合・授業・レッスンを積極的に採用することにより，会話行動調査の結果に近い構成となっている。

一方，場所や活動については異なりが見られる。例えば場所については，調査結果に比べ，飲食店や公民館，旅館といった公共商業施設での会話が多く，自宅や職場・学校での会話は少ない。また活動は，調査結果と比べ，レジャー活動・つき会いが多く，家事・雑事や仕事・学業が少ない。

職場・学校での仕事・学業中の会話が少ないのは，個人密着法でこの種の会話が収録しづらいためであり，今後，特定場面法により補強する必要がある。一方，自宅など，収録としては少なからず見られるものの，調査結果より少ない比率となっているものもある。これは次の理由による。自宅の場合，行動調査と同比率で会話を選定すると，自宅での家族との食事中の会話など，同じような種類の会話が多くなりがちである。家族との会話の中で，会話の形式・場所・活動のバランスを考慮すると，結果として自宅での会話が少なくな

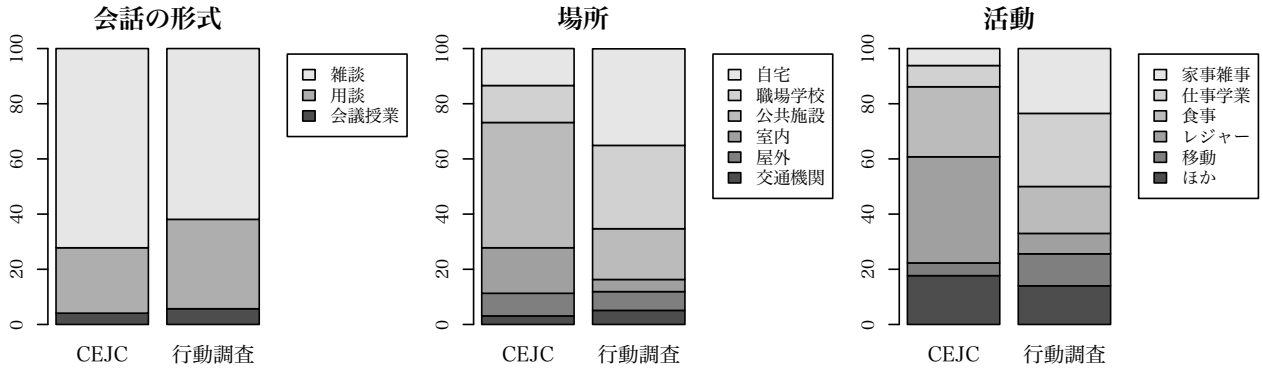

図 2　CEJC への格納予定のデータ（50 時間 97 会話）と会話行動調査における会話の形式・場所・活動の比率

り，逆に公共商業施設など自宅外の会話が多くなる。2.1 節でも言及したように，調査で求めた比率を厳密に遵守するのではなく，偏りがないことを検証するための目安として用いるものであり，今後もこの方針で進める。

会話者の属性についても簡単に見ておく。10 歳未満・10 代の未成年はいずれも全体の 5% 未満と少なく，特に中学生が 1 人（異なり人数），高校生は 0 人である。収録調査は各種個人情報を扱うなど重い負担を伴うことから，協力者は成人に限定している。そのため，未成年者を含む会話の収録の有無は協力者の家族構成に強く依存する。今後，家族に中学生や高校生を含む複数の協力者の収録を予定しているが，仮に会話者の属性情報の強い偏りが改善されない場合には，特定場面法により補うことも検討する。

3.3　アノテーション

転記テキストについては，臼田ほか (2017) の報告にあるように，2016 年 12 月 28 日までに 25.6 時間のデータに対して一次以上の作業を終えており，順調に進んでいる*2。現在，各種タグが付与された転記テキストを対象に，タグを適宜利用しつつ短単位情報を MeCab+UniDic で付与する環境の整備を終えたところである。また転記テキストは，漢字仮名交じりで表記するため，読み（発音）が一意に同定できないこともある。そのため，自動解析の結果得られた発音情報を人手で確認・修正することにより，形態論情報から正確な発音の情報が得られるようにする。こうした確認・修正作業の環境も整ったところであり，今後，本格的に形態論情報の修正作業を進める予定である。

*2 25.6 時間分の転記テキストを解析した結果，総短単位数は 272,760 語（10,647 語/1 時間）であった。表 1 に示したコーパス全体の単語数の推定値は，これに基づき算出したものである。

4　おわりに

本稿では，現在構築中の『日本語日常会話コーパス』の基本設計と構築状況について報告した。また，これまでに収録した会話の内訳について調査し，職場・学校での仕事・学業中の会話や未成年者を含む会話が少ないことが分かった。この結果を踏まえて今後の会話収録の方向性を検討するこで，多様な種類の会話をバランスよく納めたコーパスの構築を目指す。

謝辞 本研究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」の研究成果を報告したものである。コーパスの収録にご協力・ご参加くださった皆さまに感謝します。

参考文献

居關友里子・第十早織・伝康晴・小磯花絵 (2017).「日常会話コーパスのための談話行為タグの設計」　『言語処理学会年次大会発表論文集』.

JDRI (2017).『発話単位ラベリングマニュアル version2.1』. http://www.jdri.org/resources/manuals/uu-doc-2.1.pdf

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵 (2017).「『日常会話コーパス』の転記基準と作業工程」　『言語資源活用ワークショップ 2016 発表論文集』.

小磯花絵・土屋智行・渡部涼子・横森大輔・相沢正夫・伝康晴 (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.

田中弥生・柏野和佳子・角田ゆかり・臼田泰如・伝康晴・小磯花絵 (2017).「『日本語日常会話コーパス』構築における会話収録方法」　『言語処理学会年次大会発表論文集』.

臼田泰如・川端良子・徳永弘子・西川賢哉・小磯花絵 (2017).「『日本語日常会話コーパス』の転記基準と特徴について」　『言語処理学会年次大会発表論文集』.

五十嵐陽介 (2015).「韻律情報」　小磯花絵 (編)『話し言葉コーパス　設計と構築』 東京: 朝倉書店　pp. 81–100.