『日本語日常会話コーパス』構築状況と予備的分析

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 †,*

* 国立国語研究所 † 千葉大学

1 はじめに

国立国語研究所では，H28 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」を開始し，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『日本語日常会話コーパス (Corpus of Everyday Japanese Conversation, CEJC)』の構築を進めている。CEJC の特徴は，(1) 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る会話を対象とすること，(2) 話者や場面などの観点からバランスよく集めること，(3) 転記・音声だけでなく映像まで含めて収録・公開することである (小磯ほか, 2016)。

日常場面で自然に生じる会話をバランスよく集めるため，CEJC では，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（協力者）自身に日常会話を収録してもらう方法（個人密着法）を主軸に会話を収集している (田中ほか, 2017)。

本稿では，これまで個人密着法に基づき収録した会話を対象に，会話の種類や話者の属性の観点からデータのバランスについて検証する。また収録した会話データの特徴についても報告する。

2 収録会話のバランスの検証

調査計画 個人密着法では，首都圏に在住の協力者 40 名（男女 × 20 代・30 代・40 代・50 代・60 代以上 × 各 4 名）に日常会話 15〜18 時間程度の会話を収録してもらう。その上で，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，各名約 4〜5 時間分の会話，計 180 時間程度を選定する計画である。個人密着法による会話の種類を調査し，個人密着法では収集の難しい種類の会話については，調査者が主体となり収録する特定場面法で補う。

収録状況 H29 年 12 月末現在，28 名が調査を終え，そのうち 20 名についてはコーパスに格納する会話を選定した。これは個人密着法に基づくデータの半分に相当する。このデータの特性を見ることで，CEJC における個人密着法のデータ全体の傾向を予測することができ，また今後のデータ選定の方針を見直すことが可能となる。選定を終えた協力者 20 名の内訳を表 1 に示す。データの規模は，94 時間，210 会話，延べ話者数 783 名，異なり話者数 424 名である。このデータを対象に，会話の種類や話者の属性の観点からデータのバランスを検証する。なお，小磯ほか (2016) では 10 名の協力者のデータを対象に，会話の件数の観点からバランスを見たが，今回はデータを倍に増やし，会話の件数と時間の観点からバランスを検証する。

表 1: 選定を終えた協力者 20 名の性別・年齢の内訳

	20 代	30 代	40 代	50 代	60 代以上	計
男性	2 名	2 名	2 名	2 名	2 名	10 名
女性	2 名	2 名	3 名	2 名	1 名	10 名

会話の属性に基づく検証 会話の属性についてはコーパス設計のために実施した会話行動調査の結果と比較する。この調査は日常会話の実態をとらえてコーパス設計に活かすために実施したものである。1 人あたり平日 2 日・休日 1 日に行った全会話を対象に，会話の形式や会話が行われた場所，会話中に行われていた活動，話者数などを調査した (小磯ほか, 2016)。

協力者 20 名のデータを対象に，会話形式，場所，活動，話者数についてその内訳を求めた。行動調査の結果と合わせて図 1 に示す。上段は会話の件数で見た場合の，下段は会話の時間で見た場合の結果である。

会話形式の結果から見る。会話の件数で見た場合，行動調査より用談が若干少なめだが，時間で見ると用談はむしろ多く，会議会合が少ない。会議会合は他と比べ平均して 1 会話が長いが，コーパスに格納するデータは 1 協力者あたり 4〜5 時間であり，この中で多様な会話を含めるため，1 会話の長さに上限を設けている。そのことが影響したと考えられる。このような制限による偏りが若干見られるものの，形式の観点

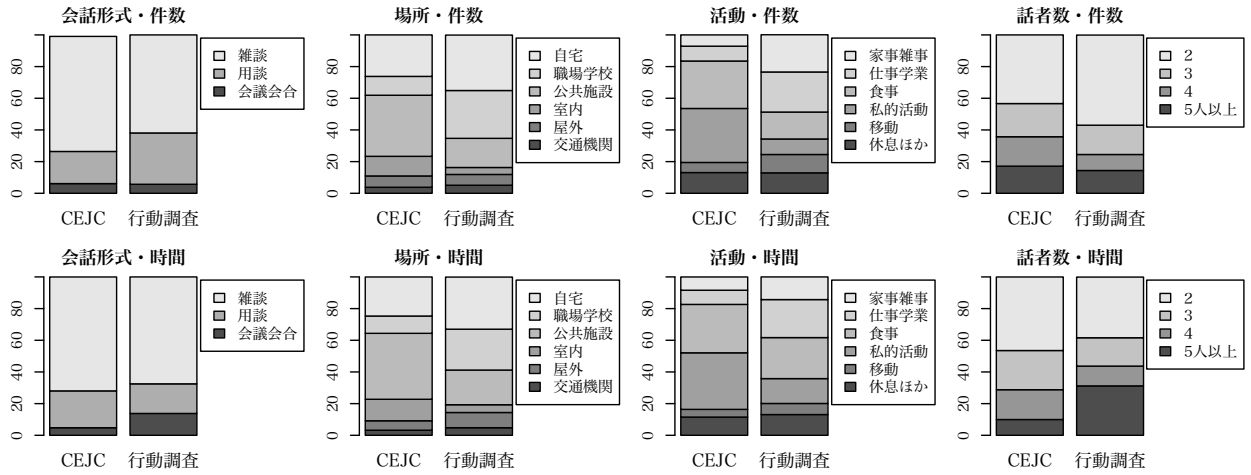

図 1: CEJC 格納予定データ（協力者 20 人分）と会話行動調査における会話の形式・場所・活動・話者数の比率（上段：会話件数，下段：会話時間）

表 2: CEJC 格納予定データの話者の職業の内訳

職業	会社員	自営業	パート	専業主婦	無職	就学前	小学生	中学生	高校生	大学生	その他	計
異なり	250	95	67	98	41	10	27	15	2	95	23	723
延べ	135	50	30	61	28	6	6	6	1	34	7	364

からは概ねバランスよくデータが選定できていると言える。この傾向は話者数についても見られる。時間で見た場合，5 人以上の会話が行動調査より少い傾向が見られるが，人数の多い会話は会議など長めの会話が多く，1 会話あたりの時間を制限したためである。

次に場所について見る。小磯ほか (2016) では，自宅や職場学校での会話数が調査よりも少なく，公共商業施設の会話が多く見られることを指摘した。この結果を受け，その後のデータ選定ではバランスをとるよう心がけた。結果，自宅の会話はかなり改善された。公共商業施設の多さはまだ目立つが，前回報告時からの大幅な改善が見られる。一方，職場学校については全く改善が見られない。個人密着法という収録法の限界であり，特定場面法で補う必要がある。

活動について，小磯ほか (2016) では，家事・雑事や仕事学業が少なく，付き合いやレジャー活動などの私的活動が多いことを指摘した。家事・雑事は，会話の件数では改善は見られないが，時間で見ると件数ほどの偏りは見られない。これは私的活動についても同様である。一方，仕事学業は全く改善が見られない。場所の場合と同様，特定場面法での調整が必要である。

話者の属性に基づく検証　話者の年齢・職業については若干偏りが見られる。職業別に見た話者数を表 2 に示す。小磯ほか (2016) において，未成年が少ないこと，特に中学生が異なりで 1 人，高校生は 0 人であることを指摘した。協力者は成人に限定しているため，未成年者のデータは集めにくい。そこで，家族に未成年者，特に中高生を含む人に積極的に調査協力をお願いした。表 2 にあるように，中学生は異なりで 6 名まで増えたが，高校生は依然 1 名である。今後の重点課題である。

3　CEJC に見る日常会話の多様性

前節で取り上げた協力者 20 名の会話のうち，H29 年 12 月 26 日までに UniDic+Mecab で形態論情報を付与し人手修正を進めているデータを用いて，CEJC に含まれる会話の特徴を調べ，CEJC が多様な会話を収めているかを見ていく。分析に用いるデータは 81 会話，38 時間，延べ話者数 275 名，異なり話者数 159 名である。

3.1　発話量

CEJC の日常会話および CSJ の独話（学会講演・模擬講演）と自由会話を対象に，1 分あたりの語数（以下，語数/分）を会話・講演ごとに算出した（図 2）。図から，日常会話では語数/分の分散が CSJ の独話・会話よりも大きいことが分かる。CSJ の場合，講演や会話をするためにその場に参与しているため，一定時間以上発話しないという状況は生じにくい。一方，日常会話では，例えば旅行の相談などのように会話をす

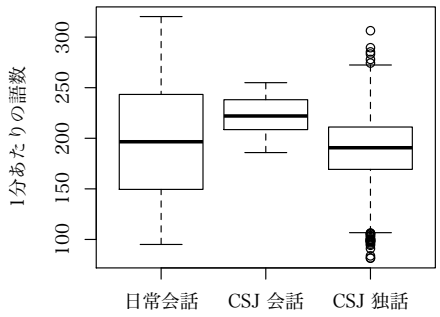

図 2: 1 分あたりの語数－日常会話と CSJ の独話･会話

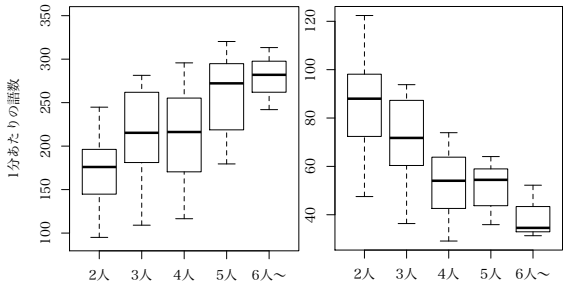

図 3: 会話の話者数ごとにみた 1 分あたりの語数

ることが期待される場面もあるが，例えば料理を一緒に作ったり散歩をする際などは必ずしも常に会話することが求められるわけではない。語数/分だけを見ても，CEJC が多様な日常場面における会話行動をとらえていることが分かる。CSJ の会話のように，会話することを目的に集められたデータと異なる点である。

図 3 の左「会話全体」は，CEJC の日常会話を対象に 1 分あたりの語数を会話を構成する話者数ごとにプロットしたものである。また 1 人あたりの語数を算出し，図の右「1 人あたり」に示す。1 人あたりの語数を見ると，話者数が多くなるにつれ語数/分が少なくなる傾向が見られる。通常，会話では複数の話者が同時に発話することは少ないため，話者数が多くなれば 1 人あたりの語数は当然減る。興味深いのは，会話全体で見た場合に話者数が多くなるほど語数/分は増える傾向にある点である[1]。これは，話者数が多いほど，ある一時点においていずれかの参与者が発話をしている可能性が高くなるためと考えられる。また，4 人以上の会話では会話が 2 つ以上に分裂することもあるため，その影響も考えられる。前節で見たように CEJC では会話の話者数についても偏らないよう配慮しているが，その結果，多様な会話の構造をとらえていると言えよう。

3.2 敬体の使用率

ここでは，参与者同士の関係性によって話体がどのように使い分けられているかを概観する。図 4 に，相手との関係性（相手がどの関係に当たる人か）ごとにスピーチレベルの割合を求めた結果を示す。スピーチレベルについては，まず「です」「ます」「ございます」を伴うか否かに応じて敬体と常体を区別する。また敬体の場合，終助詞が付与されるほうがスピーチレベルは低いことが指摘されているため (伊集院, 2004; 佐竹 , 2016)，終助詞（ね，よ，な，さ，わ，の）の使用の有無についても区別する。なお，集計の対象とした発話は，述語を動詞あるいは形容詞とする文である。

まず敬体・常体の使用傾向を概観する。図 4 から，相手が客の場合，敬体使用率が極めて高く，取引先がそれに次ぐ傾向が見られる。このように仕事の相手としてやりとりをする相手に対してはスピーチレベルが高くなることが分かる。いわゆる親疎や上下関係がスピーチレベルに関係していると考えられる事例も見られる。たとえば家族・友人と知人を比較すると，相対的に親の関係である家族や友人に対する方が知人よりも敬体使用率は高い。また先輩･同僚と後輩，および，先生と生徒には，それぞれ上下の関係があると考えられるが，いずれもこの順に敬体使用率は下がる。

終助詞の有無についても傾向を見てみよう。全体的に見て終助詞の使用は常体での発話により多く，どの関係性においても敬体で終助詞が付与される割合は常体より低い傾向にある。ただし先生と生徒については敬体における終助詞使用率が高い。今回対象とした先生・生徒間の会話は 4 会話のみで，いずれも授業などの改まった場のものではなく，くだけた場での雑談であった。このことが相対的にスピーチレベルの低い終助詞を伴う敬体の使用につながったと考えられる。

次に，特定の協力者 2 名に限定してそれぞれ傾向を見てみる（図 5）。協力者 A（30 代男性専門職，図 5 の左）については，上述の分析において家族としてまとめて分類した会話を，妻のみの場合と義母（妻の母）も同席する場合に分けた。協力者 A の場合，妻や後輩に対する常体の使用が，それ以外の相手よりも際立って多い傾向が見られる。ごく近しい関係，また

[1] 図は省略するが，語数ではなく発話数で見ても同様の傾向を示す。人数が増えるとあいづちなどの聞き手行動が増えるため，その影響も考えられるが，感動詞類を除いても同じである。

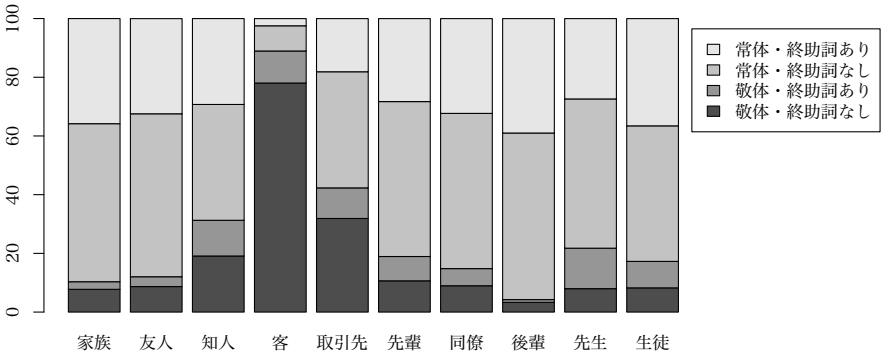

図 4: 相手との関係性とスピーチレベル

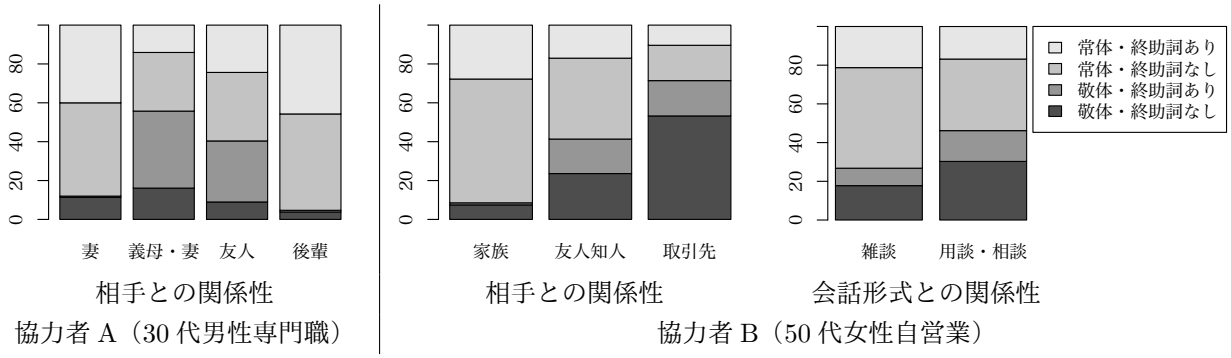

図 5: 相手との関係性・会話形式とスピーチレベル－特定の協力者 2 名に限定した場合－

は自分の方が立場が上になる関係の相手に対しては敬体はあまり用いられないと言える。一方，義母を含む会話では敬体の使用が大幅に増える。興味深いのは終助詞を伴う敬体の使用が多い点である。これによりスピーチレベルを少し落とし，丁寧になりすぎることによる疎遠な印象を避けている可能性が考えられる。協力者 B（50 代女性自営業，図 5 の右）の場合，家族（子供および夫）との会話では常体での発話が大半を占めるが，取引先との会話においては敬体の割合が高い。友人知人はその中間的な傾向を示す。この点は協力者 A の場合と同じである。先に見た通り，場の改まり度も敬体の使用に影響するため，協力者 B については会話形式との関係も調べた。図から，雑談に比べて用談・相談では敬体の使用が多くなる傾向が見られる。改まり度が低いと考えられる雑談に比べて用談・相談では，ある程度は改まった話をする場合が多いと考えられるが，そうした場面の違いを反映した分布を示していると言える。

このように話体の特徴を見るだけでも，CEJC がさまざまな場面での多様な相手との会話を収録しており，それがことばの特徴に反映されていることが分かる。

4　おわりに

本稿では現在構築が進められている『日本語日常会話コーパス』のうち，これまで収録した会話を対象に，会話の種類や話者の属性の観点からデータのバランスを検証した。また整備の進んだデータを用いて会話の言語的な側面の特徴を分析し，本コーパスが多様な日常場面の会話を捉えていることを示した。H30 年度には 50 時間の会話をモニター公開する予定である。

謝辞　本研究は国立国語研究所の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。

参考文献

伊集院郁子 (2004)「母語話者による場面に応じたスピーチスタイルの使い分け――母語場面と接触場面の相違――」　『社会言語科学』, 6:2, pp. 12–26.

小磯花絵ほか (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.

佐竹久仁子 (2016)「日常談話に見られる敬語使用の実態」　現代日本語研究会 (編)『談話資料　日常生活のことば』 東京: ひつじ書房　pp. 191–212.

田中弥生ほか (2017)「『日本語日常会話コーパス』構築における会話収録方法」　『言語処理学会第 23 回年次大会』　pp. 481–484.