『日本語日常会話コーパス』収録の進捗状況

著者	田中 弥生, 柏野 和佳子, 角田 ゆかり, 伝 康晴, 小磯 花絵
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	248-257
発行年	2017
URL	http://doi.org/10.15084/00001479

『日本語日常会話コーパス』収録の進捗状況

田中 弥生（国立国語研究所 音声言語研究領域 ／ 東京大学大学院 総合文化研究科）†
柏野 和佳子（国立国語研究所 音声言語研究領域）
角田 ゆかり（国立国語研究所 音声言語研究領域）
伝 康晴（千葉大学文学部・国立国語研究所 音声言語研究領域）
小磯 花絵（国立国語研究所 音声言語研究領域）

Construction of "Corpus of Everyday Japanese Conversation" : Progress Report of Recording Naturally Occurring Conversations

Yayoi Tanaka (National Institute for Japanese Language and Linguistics / The University of Tokyo)
Wakako Kashino (National Institute for Japanese Language and Linguistics)
Yukari Sumida (National Institute for Japanese Language and Linguistics)
Yasuharu Den (Chiba University / National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

2016 年度から構築が始まった「大規模日常会話コーパス」プロジェクトによる『日本語日常会話コーパス』の収録手続きの概要と進捗状況について報告する。本プロジェクトでは，日常場面の中で自然に生じた会話を対象とする。そのため，性別・年代などの点からバランスを考慮して調査協力者を選別し，収録機材等を 2〜3 カ月程度貸し出して調査協力者自身に日常会話を収録してもらう方法を採用している。本発表では，こうして定めた収録方法の概要を述べるとともに，これまでに終了した 13 名の調査協力者による約 200 時間の収録について進捗状況や生じた問題などを報告する。

１．はじめに

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」（プロジェクトリーダー：小磯花絵）では，日常場面で自発的に生じた会話約 200 時間を収録した大規模なコーパス『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC)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・相互行為・経年変化の観点から多角的に解明することを目指している。本稿は，『日本語日常会話コーパス』構築における収録の方法を概説し，進捗状況について報告する。２節で本コーパスの概要を，３節で現在行っている収録方法を説明し，４節でこれまでの収録状況の報告を行い，５節で今後の予定について述べる。

２．コーパスの概要

日常場面での会話を収録するためには，収録場面を人工的に設定するのではなく，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)を対象としなければならない。日常の言語生活を反映したコーパスの設計に際し，我々が普段どのような場面で会話しているか実態を知るため，本コーパスの構築に先立

† yayoi@ninjal.ac.jp

って，2015 年度に「会話行動調査」を実施した。この調査結果を参考に，多様な会話をバランスよく収めたコーパスを構築する。調査の詳細は小磯他(2016)を参照されたい。

理想的には，起床してから就寝までの，自宅や職場，店舗，屋外，交通機関など，さまざまな場所で生じる会話が対象となる。このような多様な日常場面での会話を収録するために，British National Corpus(BNC)の spoken part の収録法(Burnard and Aston 1998)を参考に，以下の 2 つの収録法を採用することとした。

個人密着法　性別･年代などの点からバランスを考慮して選別された調査協力者（以下，協力者）に収録機材等を一定期間貸し出し，協力者自身に会話参加者（以下，会話者）との日常会話を収録してもらう方法。プロジェクトメンバー（以下，調査者）は原則として介在しない。

特定場面法　職場での会合や店舗での店員とのやりとりなど，個人密着法では技術的・倫理的に収録が難しいと思われる場面を特定し，調査者が主体となり収録する方法。調査者は介在するが，日常場面の中で自然に生じる会話を対象とする。

本プロジェクトではまず個人密着法に基づき収録調査を開始した。構成は以下の通りである。なお，コーパス設計の詳細は小磯他(2017)を参照されたい。

■調査期間：　2016 年 4 月～2018 年度（予定）。
■協力者の属性：首都圏（東京都，神奈川県，埼玉県，千葉県）に在住の 20 代以上の男女。出身地や生育地域の制限は設けていない。
■協力者の人数：約 40～50 名。
20 代，30 代，40 代，50 代，60 代以上の男女，それぞれ 4-5 名を予定。協力者は個人情報を取り扱うなど重い責任が生じることから，未成年者を含めないこととした[1]。
■収録時間：　協力者 1 名あたり 15～18 時間。
■コーパスへの採録時間：協力者 1 名あたり約 4～5 時間。40～50 名で合計 160～200 時間。

個人密着法での収録をある程度進めた段階で，不足する種類の会話を補うために，特定場面法を実施する。コーパス全体で 200 時間の規模を目指す。本稿では，すでに調査が進んでいる個人密着法の収録方法と，収録状況を述べる。

３．収録方法

本節では，個人密着法による収録方法について述べる。上述のとおり，この収録法では，研究者は収録場面に立ち会わず，収録に伴う一連の作業を協力者自身に担当してもらう必

[1] 協力者が集める会話の中に未成年者が含まれることはある。しかし個人密着法では，必ず協力者（つまり成人）が加わる会話のみが対象となるため，未成年者のみにより構成される会話がこの方法で収録されることはない。個人密着法で収録したデータの会話者属性の性質を調査し，仮に未成年者が少ないなどの偏りが見られる場合には，特定場面法などで補うことも検討する。

要がある。そのため，収録調査の手続きや関連資料などを入念に検討して定めた。なお，田中他(2017)でその詳細を述べたため，本節では概要を記すにとどめる。

３．１　協力者の募集方法

主に調査者の伝手により協力者を集めた。属性が偏らないよう，年代・性別の他，職業の有無も考慮した。候補となる人に，協力者募集のチラシあるいはプロジェクトのホームページにて概要を確認してもらった後，30 分～1 時間程度調査についての詳細な説明を行って，意思を確認したうえで，協力者を決定した。

３．２　協力者に依頼する作業

協力者が行う作業は，以下の通りである。

① 会話の収録（録画・録音）
② 会話者への調査内容及び公開方法の説明
③ 会話者への同意書への署名の依頼
④ 会話状況（日時，使用機材，配置など）の記録
⑤ 会話者の属性（性別，出身地など）に関するメタ情報収集のための会話者へのフェイスシート記入の依頼
⑥ 自宅等での機材や書類の管理
⑦ 定期的なデータ提出
⑧ メールや電話などでの調査者とのやりとり
⑨ 各種打合せ（収録開始時の一連の調査方法についての説明・調査終了時のフォローアップインタビュー，いずれも 3 時間程度）

一連の調査協力に対し，協力者に謝金 12 万円を支払う。謝金の額は，テスト収録に基づき作業量を試算した上で確定した。調査者が介在せずに上記の作業を行ってもらうため，マニュアルは詳細に整備し，機材については協力者の負担が最も少ない形での設定とした。また，会話者の同意書やメタ情報を適切に収集できるよう，同意書やフェイスシートの様式も改良を重ねた。同意書及びメタ情報の収集については田中他(2017)を参照されたい。協力者については，個人情報取り扱いも含めたガイドライン（複製の禁止，調査で得た個人情報を調査以外に用いない，データ保管の安全性の確保など）を作成し，調査開始時に説明の上，同意書に署名を得ている。

自然に発生する日常会話を収録するため，協力者には，収録のために人を集めるのでなく，日常の自宅での家事や食事，収録とは関係なく設定された会食や打ち合わせなどの場面に機材を持ち込み，会話者の同意を得たうえで収録するよう求めた。収録場所や場面，会話の相手などに関して，多少のバリエーションがあることが望ましいことも伝えた。

３．３　収録調査の流れ

収録調査期間（機材貸し出し期間）は基本的には 2 カ月程度，最大で 3 カ月とした。大まかな流れは以下のとおりである。一度にすべての収録を行うのではなく，4～5 回に分けて収録してもらい，随時調査者が確認し，フィードバックを行う。特に第一次収録終了後は必ずフィードバックを待ってから第二次収録を開始してもらうこととした。詳細については田中ほか（2017）を参照されたい。

① 機材等送付（調査者から協力者へ）

② 収録開始時打ち合わせ
③ 第一次収録
④ フィードバック（調査者から協力者へ）
⑤ 第二次収録
⑥ 第三次収録
⑦ 第四次収録 （必要に応じて第五次収録）
⑧ 機材等返却（協力者から調査者へ）
⑨ 調査終了時打ち合わせ

３．４ 収録機器

協力者は必ずしもカメラなどの機械類の取り扱いが得意なわけではなく，また自宅外に機材を持ち出すこともあるので，シンプルな設定，短時間での設営，簡単な操作，軽量といった制約のもとで収録方法と収録機器を検討し，次のような方法で収録を実施している。

３．４．１ 基本収録

基本的な収録の機器を表 1 に示す。また，2 名が対面する収録の基本的な配置で撮影される映像を図 1 に示す。表 1 の各機器を図 1 内に記号で表示した。

表 1 基本収録機材

	品名	設定	基本使用台数	特徴	図 1 内記号
映像	Kodak PIXPRO SP360 4K	1440×1440, 60fps	1	360 度撮影可能なカメラ。会話者たちの中央に配置。	(a)
	GoPro Hero3+	1920×1080, 60fps	2	170 度の視野角を持つカメラ。会話者を俯瞰的に記録。	(b1)(b2)
音声	IC レコーダー Sony ICD-SX734	リニア PCM, 44.1kHz,16bit	最大 6	会話者ごとにフォルダーに入れて首から下げる。	矢印
	IC レコーダー Sony ICD-SX1000	リニア PCM, 44.1kHz,16bit	1	会話全体を録音。中央に配置。	(c)

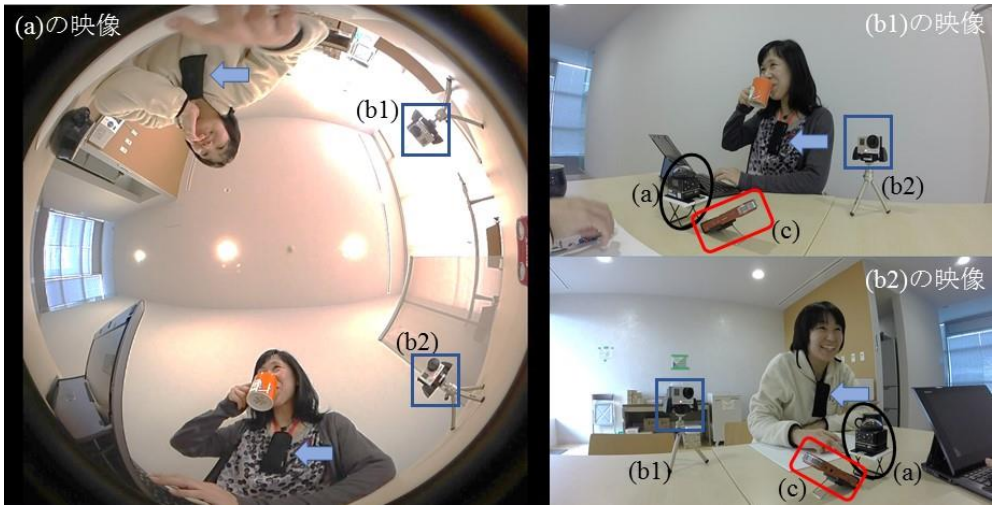


図 1 対面 2 名会話時の映像

３．４．２　基本収録以外の機材

基本収録以外の機器について，表 2 に示す。

表 2　基本収録以外の機器等

	品名	数量	特徴
移動時収録	Panasonic HX-A500	1	ウェラブルカメラ。散歩や散策，外出先への移動の際に，1 名が頭に装着。音声は基本収録の会話者ごとの IC レコーダーを使用。
車内収録	車載用アクセサリ	適宜	カメラは基本収録 GoPro を主に使用。会話者ごとの IC レコーダーも使用。
電話会話収録	小型マイクロフォン	1	スマートフォンと IC レコーダーを接続するコードをあらかじめ接続した小型マイクロフォンを使用。

３．５　マニュアル（手引き）

収録調査が問題なく進められるよう，マニュアルを用意した。調査の進め方や，機材の取り扱い方法，データの提出のタイミングや方法などを記載した『会話収録の手引き』の他，具体的な機器の操作については，別冊で『会話収録の手引き―基本収録編―』『会話収録の手引き―移動編―』『会話収録の手引き―電話編―』を作成した。

４．　収録状況の報告

上述の表 1 に示した機材を 6 セット用意し，2016 年 4 月から順次調査を開始してきた。約 10 カ月が経過し，表 3 に示したように，合計 13 名の収録調査者が調査を完了し，6 名が現在調査中である。

表 3　協力者の属性（2017 年 1 月 24 日現在）

年代	男	女	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	自営自由業　(調査中)	自営自由業　(終了)	2 人
60 代以上	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

本稿では，このうちすでに調査が終了した 13 名の収録状況について報告する。なお，収録されたデータのうちコーパスに格納するデータ（全体の約 3〜4 分の 1）については，小磯他(2017)を参照されたい。

４．１　収録回数と収録時間

13 名の年代別にみた収録回数と収録時間の内訳は表 4 の通りである。20 代と 40 代の一人当たり収録時間が少ないのは，それぞれ 1 名ずつ，調査期間中に転職などの理由によって，予定していた収録ができず調査期間を終えたことによる。なお，３．３で，収録調査期間（開始時打ち合わせから最終収録日まで）は基本的に 2 カ月と述べたが，最も短い協力者で 46 日，最も長い協力者で 91 日，平均すると 67 日であった。

表 4　年代別収録回数及び収録時間（調査終了分）

年代	協力者人数	収録回数	一人当たり収録回数	収録時間（時間：分）	一人当たり収録時間（時間：分）
20 代	3	51	17	42:43	14:14
30 代	5	103	20.6	86:02	17:12
40 代	2	33	16.5	25:26	12:43
50 代	1	19	19	17:02	17:02
60 代以上	2	34	17	31:27	15:43
計	13	240	18.5	202:40	15:35

４．２　収録形態

３．４で述べたように，調査には，大きく分けて，基本収録，移動時収録，さらに電話会話と車内での収録という 4 つの形態がある。図 2 に，収録形態ごとの収録回数を示す。

基本収録では，３．４．１で述べたように， SP360，GoPro（2 台），中央に配置する IC レコーダー, 会話者一人一人が首からかける IC レコーダーの，4 種類の機材を使用する。すべての機材を使用した収録（フルセット）は基本収録の約 6 割を占めている。しかし，何らかの事情により，すべての機材を使用できていない場合もある。図 3 のように，2 名が並んで座る場合，正面に 1 台の GoPro を配置することで図 4 のように 2 名が十分に撮影できるため, 1 台の

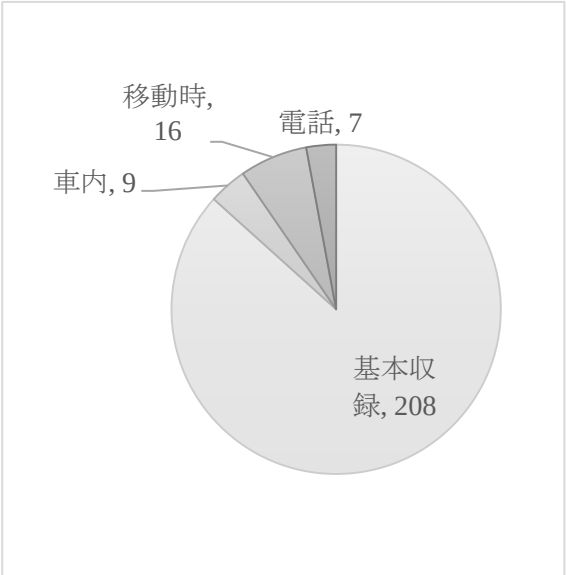


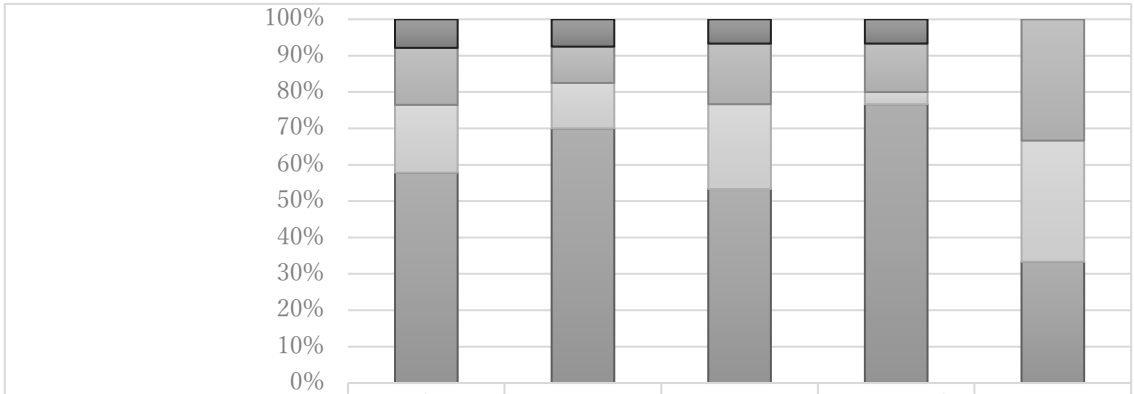
図 2　収録形態ごと収録回数

図 3　2 名並んでの収録

図 4　2 名並んでの収録時 GoPro からの映像

み使用することがある。「GoPro1 台のみフルセット」はこのような状況が含まれ，約 16%を占めている。「GoPro なしフルセット」は SP360 と IC レコーダーでの収録で，約 15%であった。その他に，収録をはじめてすぐに SP360 の電池が切れたためやむを得ず SP360 のない構成で収録を継続したケースや，貸し出している IC レコーダーよりも多い人数での収録になり，やむを得ずレコーダーを付けない人が含まれているケース，IC レコーダーの電源の入れ忘れなどがあった。

収録場所別の機材使用状況を図 5 に示す。「屋外」以外ではフルセットの使用が 5 割を超えている。ある程度落ち着いて収録準備ができることが要因と考えられる。「自宅」「その他の室内」は 7 割程度がフルセット使用である。「その他の室内」は，実家や友人宅など協力者の自宅以外に機材一式を持参して収録するケースが該当する。いずれも，時間的空間的に余裕をもって機材を配置することが可能であるためであろう。レストランや公民館などの「公共商業施設」と「職場学校」では，GoPro1 台のみフルセットと GoPro なしフルセットが合計して 3〜4 割ある。準備の時間が限られたりテーブルやスペースが小さかったりなど，時間的空間的に余裕がない場合が相対的に多いことが考えられる。

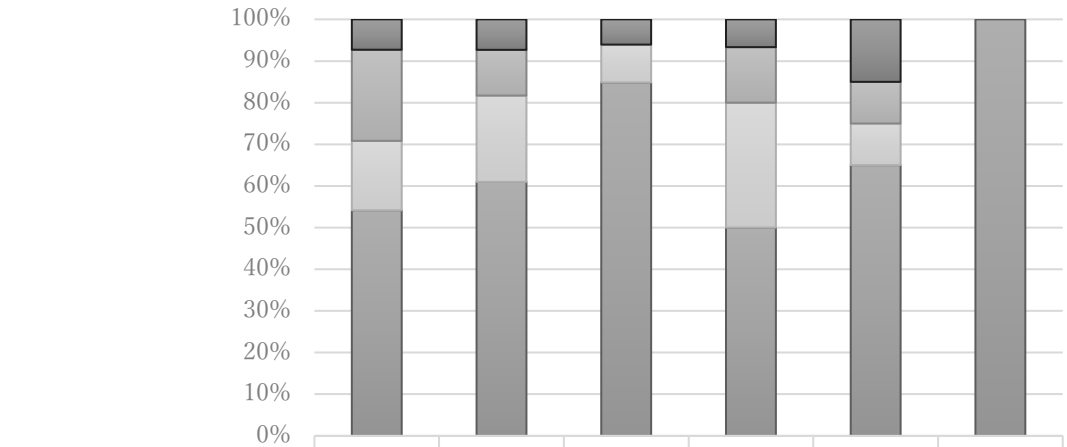


	公共商業施設	自宅	職場学校	その他の室内	屋外
その他	8	3	2	2	0
GoProなしフルセット	16	4	5	4	2
GoPro1台のみフルセット	19	5	7	1	2
フルセット	59	28	16	23	2

図 5　基本収録の収録場所別使用機材

４．３ 活動と使用された機材

次に，活動別の機材使用状況を，図 6 に示す。食事をしながらレジャー活動（付き合いを含む）を行うような忘年会や友人との会食などは「食事」「レジャー活動」の両方に分類しており，1 つの収録に複数の活動が含まれることがあるため，総数は収録合計と一致していない。なお，活動の分類には「移動」を設けているが，ウェラブルカメラ使用の収録を対象とするため，図 6 には含まれていない。



	レジャー活動	食事	休息	仕事学業	社会参加	家事雑事
その他	7	6	2	2	3	0
GoProなしフルセット	21	9	0	4	2	0
GoPro1台のみフルセット	16	17	3	9	2	0
フルセット	52	50	28	15	13	9

図 6　基本収録の活動別使用機材

活動別の機材の使用状況をみると，「家事雑事」と「休息」ではフルセットでの収録の比率が高いことがわかる。自宅など落ち着いて準備ができる状況での収録であるためと考えられる。「食事」，「レジャー活動」，「仕事学業」では，GoPro1 台や GoPro なしでの収録が増える。図 3 と図 4 で示したように GoPro1 台で会話者が十分に映る場合や，外食でスペースが狭く置き場所を確保できない場合などが考えられる。「社会参加」は，PTA や地域懇談会，祭りなどの打ち合わせで，スペースの制約で GoPro が置けないケースや，IC レコーダーより多い人数が参加して数が一致しないために「その他」に計上されている場合などがある。

４．４　収録調査において発生した問題

これまで述べたように，協力者自身が収録を行う必要があるため，収録の手続きや各種書類の書式に至るまでかなりの検討を加えてきた。その結果，これまでに大きなトラブルは発生していないが，当初想定していなかった事象も発生しているため，ここに代表的な事例と対応を報告する。こうした事例は，今後同じような収録調査を試みる研究者にとって有益な情報となるだろう。

４．４．１　機材・データについて

■カメラの電池切れ

上述のように，電池が切れてしまったために機材をフルセット使用できなかったケースが比較的多く発生した。特に SP360 は，当初，高解像度 2880×2880 で収録していたが，バッテリーの持続時間が 1 時間程度と短く，収録途中で電池切れとなることがたびたびあった。そのため，7 人目の調査から解像度の設定を 1440×1440 に下げ，バッテリーの持続時間を長くすることによって，収録途中の電池切れをできるだけ防ぐこととした。

■カメラでの静止画の撮影

カメラ（特に GoPro）のデータが動画ではなく静止画となっていたケースが見られた。これらのカメラは，電源ボタンがモード切替ボタンもかねていて，動画や静止画が切り替えられるため，誤って静止画モードで撮影してしまうことがある。動画での録画が開始されたことを，録画中に点滅するランプやカウンターの数字の上昇によって必ず確認するよう，マニュアルに記載したうえで，調査開始時の打ち合わせで強調している。また，調査期間中，数回に分けてデータを提出してもらうため，随時注意を促すことによって同様の問題はなくなっていく。

■IC レコーダー

IC レコーダーは身に着けるため，誤操作（録音レベルの変更など）が起こり得る。そこで協力者には，録音開始後，会話者に配布する前に「ホールド」（誤操作防止状態）の設定にするよう求めている。

また，中央に配置する IC レコーダーは，会話者全員の音声を収録するためのものだが，上部のマイク部分がテレビや居酒屋等での他の客の側を向いている場合，その音声が大きく録音されてしまうことがあり，向きに注意するよう喚起している。

４．４．２　書類について

３．２で述べたように，会話者への調査収録についての同意書とメタ情報収集のためのフェイスシートの記入を，協力者に依頼してもらっているが，指定欄以外への署名や記入もれなどが少なからず生じた。特にレストランなどでの収録の際は，あわただしい状況で，機材の設置を行いながら，調査についての説明と書類記入の依頼をすることになるため，十分な確認ができないものと考えられる。そこで，様式を入念に検討し，例えば，同意書の署名欄を表面に，後日撤回する場合の署名欄を 2 つ折りにした内側の面にするなど変更した結果，これらの問題はかなり解消された。

５．　まとめと今後の予定

本稿では，『日本語日常会話コーパス』構築のために現在行っている個人密着法に基づく収録の概要と現段階での収録状況について報告した。調査者が介在しない状況で，一般の人に，複数の機材を用いた収録や，同意書やフェイスシートの取得など，かなり複雑な作業をお願いしている。しかし，機器の操作が不得意な人も含め，多少の問題はあるものの，何とか無事に調査を終えている。今後，現在調査進行中の 6 名は 2 月から 4 月の間には調査が終了し，2017 年度には 16～18 名の収録調査を依頼する予定である。

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。

文献

Burnard, Lou, and Guy Aston (1998). The BNC Handbook. Edinburgh: Edinburgh University Press. (北村裕(監訳) (2004). 『The BNC Handbook: コーパス言語学への誘い』松柏社,)

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 (2017)「『日本語日常会話コーパス」の構築」『言語処理学会第 23 回年次大会（NLP2017）予稿集』

小磯花絵・土屋智行・渡部涼子・横森大輔・相澤正夫・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・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」『言語処理学会第 23 回年次大会（NLP2017）予稿集』