対話映像音声資料の収録とデータ化の方法

対話映像音声資料の収録とデータ化の方法

小磯 花絵　　前川 喜久雄
(E-mail: koiso,kikuo@kokken.go.jp)
言語行動研究部 第2研究室

要旨：現在, さまざまな分野で対話への関心が高まるなか, 対話研究を支えるものとして, また相互の研究の接点として, 対話データの重要性が認識されるようになってきた. しかしながら, 高品質の対話データを多角的な検討が可能となる程度の量収集しようとすると, さまざまな問題にぶつかる. そのため対話データを作成するための技術や経験, 知識を蓄積することがもとめられている. このような動向を背景に, 本研究室では現在, 対話コーパスを構築すると同時に, コーパスを正確かつ効率的に構築するための手続きや作業環境の検討を行っている. そこで今回の研究室公開では, 実際の対話の収録と書き起こし作業に焦点をあて, その方法について紹介する.

キーワード： 対話コーパス, 自発的発話, 対話収録, 転記作業

1　はじめに

近年, 言語学や音声学, 心理学, 音声工学など, さまざまな分野で対話への関心が高まりつつある. これをうけて, 研究の出発点となる対話の資料, 「対話コーパス」の構築が求められるようになってきた [1]. 対話コーパスとは, 対話の音声データや書き起こしテキストを基本に, そのほか対話状況の映像資料や, 韻律, 統語, 談話情報などのラベルをも含む資料全体を指す. このような対話コーパスを利用することで, たとえば話者の交替やあいづちの生起にどのような韻律・統語が関与しているかを分析することが可能となるのである [2].

このように対話研究においてコーパスが果たす役割は非常に大きくその利用が望まれるものの, 対話コーパスの構築にはさまざまな問題が存在しており, 構築作業は容易とは言えない. そのため対話コーパスを効率的に作成するための技術や経験, 知識の共有化が重要な課題となっている.

このような流れをうけ, 本研究室では対話コーパスを構築すると同時に, コーパスを正確かつ効率的に構築するための手続きや作業環境の検討を行っている. 今回の研究室公開では, 収録してきた対話データを紹介すると同時に, コーパスを構築するための方法や作業環境についても紹介する.

2　対話の収録

2.1　どのような対話を収録するか

2.1.1　課題指向的対話の有効性

人は朝おきてから夜ねるまで, いろいろな人といろいろな場所で会話をしている. たとえば家での家族との会話や学校での友達との会話など, 会話をしない日はないと言っていいだろう.

そのため, 街にテープレコーダーを持っていけば膨大な対話を容易に収録することができる. しかし, 研究の目的によってはこのようなデータはあまり役に立たないことがある. というのは, 家の中や街のなかでは目覚しや, ドアを閉める音, 車の音など非常に多くの音が常に鳴っているからである. 対話の音声的な側面に興味をもつ研究者にはこのような「ノイズ」の混入したデータはあまり役に立たないのである. そのため音声に関心のある研究者は, 防音室のように外の音が遮断された部屋で対話を収録することが多い. このような状況で収録された音声はノイズが少なく, またヘッドフォンマイクなどを利用することで複数の話者の音声を分離して録音することができるため, さまざまな音声分析が可能となるのである.

しかし, もし我々が防音室に入れられ, ヘッド

フォンマイクをして相手と何でもいいから会話をしろと言われたら，いつも通り会話ができるだろうか．緊張して何をはなしていいかわからず，そらぞらしい会話となるのが関の山である．そのため，防音室を利用して対話を収録する場合には何か課題やゲームなどを利用することが多い．2人(あるいは3人以上)の被験者に課題やゲームをやってもらい，課題を遂行するなかでお互いにかわされた対話を収録するのである．たとえば経路の描かれた地図をもつ人が相手にそのルートを教えるという地図を利用した課題[3]や，2人でクロスワードを解くといったゲーム[4]などを利用して，実際に対話の収録が行われている．

このような種類の対話は「課題指向的対話」と呼ばれる．防音室のような実験室環境では，こういった課題やゲームをする方が，単に日常的な会話をしてもらうよりもずっと盛り上がることが多く，結果としてより自然な対話が得られるのである．

それでは複数の人で行う課題ならば何でもよいかというと，そう簡単でもない．当然のことながら，課題は相手と対話をしながら進めるタイプのものでなければならない．たとえば，2人でクロスワードをとくという課題の場合，それぞれ黙ったまま自分で勝手に答えを出し，最後に答えをつきあわせるということも考えられる．この場合，対話は課題の最後にしか生じないことになる[1]．このように，利用する課題によって対話の量や質がかなりかわるため，課題は慎重に選択しなければならないのである．

2.1.2 カロリータスクとは

本研究室ではいくつかの課題を考案し，試験的に対話の収録を行っているが，比較的面白い対話がみられた課題として，「カロリータスク」がある．

カロリータスクとは，2人の被験者がクッキーやパイ，チョコレートなど，10個のお菓子の写真が載っている用紙をそれぞれ持ち，その写真をみながらお互いに相談してカロリーの高そうな順にお菓子を並べるという課題である．

[1] 中里他(1994)では，被験者間での発話のやりとりを活発化させるために，タテとヨコの鍵を1人ずつ別々に持たせるという工夫をしている．

お菓子は市販されているごく一般的なものである．そのため，被験者はお菓子の写真やパッケージの説明，過去の経験などにもとづきながら，2人でどれがカロリーが高そうかを相談する．2人は離れて座っており，お互いの用紙がみえないようになっている．また2人が見ているお菓子のリストは次の3つの点で異なっている．

- 用紙に載っている10個のお菓子のうち9個は一緒で，残り1つが被験者によって異なる．(つまり2人あわせて11種類のお菓子がリストされている．)
- 用紙の中でお菓子が印刷されている位置が異なる．
- 一方の用紙にはパッケージの写真(お菓子の絵・写真入り)だけが，もう一方の用紙にはパッケージの写真に加えて，「バターがチェリーの風味を花咲かせる××クッキー」といった宣伝文句や，「口当たりのソフトさ，厚さのバランスよし」といった簡単なコメントが付いている．

2.1.3 カロリータスク対話の特徴

カロリータスクの特徴として，以下3点が挙げられる．

第1の特徴は，2人のもっている情報が必ずしも一致していないため，課題を遂行するためにお互いに情報を交換しなければならず，必然的に対話が生じるという点である．課題は相手との対話を要求するタイプのものでなければならないことを先に指摘した．カロリータスクでは，お菓子のリストに偏りをもたせているため課題の遂行にあたってお互いに情報を交換する必要があり，必然的に対話が生じる．

第2の特徴は，一方の被験者だけが情報をもち相手に何かを教示をする，といったような役割の非対称性が存在しないため，発話の量や質に関する均一性が期待できるという点である．何かを教示するという課題の場合，教示する側が発話をする量が多くなるといったように，発話の量に関して偏りが生じるだけでなく，たとえば教示側の発話に命令的な発話が多発したり，また被教示者にはあいづちが多く出現したりといった，発話の質的な側面にも偏りが生じやすい．一方カロリータスクではこのよ

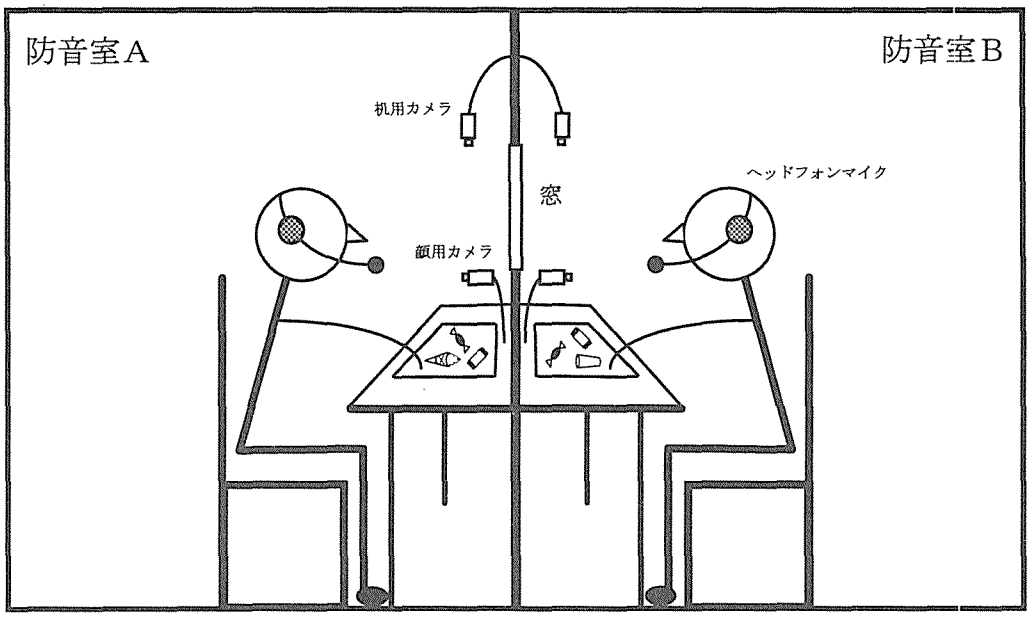

図 1: 対話の収録状況

うな役割差がないため, 発話の量や質に関して被験者間でバランスのとれた対話が期待できる.

第 3 の特徴は, 市販されている一般的なお菓子を利用しているため, 過去の経験や自身の嗜好といった課題以外の話題が出現しやすく対話が多様化するという点である. 課題指向的対話は大抵課題に関係する話題を中心に対話が進められるため, 対話は全体的に単調なものとなりがちである. カロリータスクでは過去にそのお菓子を食べたという経験そのものが課題の遂行に有用であることから, 自然と話題に多様性が生じることが期待される.

実際に収録されたカロリータスク対話の抜粋を表 1 に示す. 話者 L さんと R さんの発話のバランスもよく, またあいづちなども適度に挿入され, お互い活発に発話が交換されていることがわかる.

2.2 収録実験

次に実際の収録について簡単に説明する. 音響的な品質を保証するために対話の収録は防音室を利用して行なわれた. 図 1に収録の状況を示す.

防音室は壁で 2 つの小部屋に分離されており, 相手の上半身が見える程度の窓 (幅 80cm × 高さ 60cm) が壁の中央にとりつけられている. 被験者はこの窓ごしに机をならべ向かいあって座り, ヘッドフォンマイクを利用して対話をした[2]. 両者の音声は完全に分離されており, DAT に 2 チャンネル独立で録音された.

机の上にはお菓子の写真の載った用紙, 回答用紙, 筆記用具, そしてカメラが置かれた (図 1参照). カメラは被験者の顔より少し低い位置に設置され, 顔を中心とした上半身が記録された. また窓の上から下方向に向かってカメラが設置され, 机の上や卓上での被験者の手の動きなどが記録された. 被験者のじゃまにならないように小型のカメラが用意された. 以上各被験者ごとに 2 台ずつ, 計 4 台のカメラによる映像はそれぞれ独立にビデオテープに記録された. またこれら 4 画面を 1 画面に同期合成した画像も同時に記録された.

被験者には対話を収録するという本来の目的は伏せられ, 簡単なゲームをしてもらうということのみ伝えられた. 被験者として何組かの 2 人組 (いずれも知り合い同士) が選ばれた[3]. 両者はカロリー

[2]相手の声はヘッドフォンを通じて聞こえてくるため, 対面での対話よりは, むしろ電話での対話に近い.

[3]今後初対面の 2 人組の対話も含めて収録することを検討している.

タスクに関する簡単な説明を受けたのち課題を開始した.

3 書き起こし作業

3.1 転記テキストに求められるもの

対話音声を書き起こした転記テキストは, 統語や韻律, 談話, 身体動作などに関する情報を付与することができるという点で, 対話研究に欠かすことのできない重要な資料である. しかし, 転記テキストはあくまでも音声情報や画像情報の一部を記述したものにすぎず, 文字データに変換することによって失われる情報は非常に多い[4]. そのため, 分析を行う際には転記テキストから音声・画像情報を容易に参照できることが望まれる. このような相互参照を実現するための1つの方法は, 「発話」や「文」といった転記テキストの基本となる単位ごとにその単位の開始時刻と終了時刻を記録し, この時間情報を参照して転記テキストから音声・画像情報にアクセスするというものである.

この方法は非常に有効ではあるものの, 従来のようにカセットを聞きながら書き起こしをするという方法では, 時間情報を正確に記録することも, また作業を効率的におこなうこともできない. そのため, 書き起こし作業を支援するための環境を整備することが求められる. そこで本研究室では, 音声・画像情報とのリンクを可能とする転記テキストを作成すると同時に, 作業を正確かつ効率的に行うための環境の整備を検討している. 以下ではその活動について簡単に紹介する.

3.2 転記上の基本単位

上述したように, 転記テキストと音声・画像情報とのリンクを実現するには, 転記上の基本となる単位を決定し, その単位の開始時刻と終了時刻を効率的に測定する必要がある. このような基本単位として, たとえば「文」や「発話」といった単位が考えらえる. しかし会話においては必ずしも文法的に正しく発話されるわけではなく, 「文」の認定は容易ではない. また「発話」という単位に関しても客観的な基準に欠けており, 作業に利用することは難しい.

そこで本研究室では, 客観的に判定可能な単位として無音区間を利用した単位を採用した. この単位を「間発話単位」と呼ぶ. 間発話単位とは, 1人の発話音声のなかで, ある一定時間以上の無音によって区切られた音声区間のことである. 無音の閾値に関しては検討中であるが, 現時点では100ミリ秒以上の無音で区切られた単位を利用している[5]. 表 1の転記転記テキストは, 100 ミリ秒以上の無音で区切られた間発話単位を利用して書き起こしをしたもので, 1行単位で表現されている.

3.3 作業の流れ

時間情報をもつ転記テキストを正確かつ効率的に作成するために, 以下の手順で作業をしている. 図 2を見ながら作業の手順をおっていく.

まず, 音声データから間発話単位の開始, 終了時点を計算機によって自動的に検出し, 転記テキストのもととなる時間情報テキストを作成する (図 2左上 A 参照.). この時, 間発話単位の境界は波形情報にもとづいて自動的に検出しているため, 会話者の発話音声だけでなく, 机をたたく音や息の音などの発話以外の音 (ノイズ) も同時に検出されてしまう. そのためこのようなノイズ部分を削除する必要がある.

そこで次の段階として, 自動的に検出した間発話単位の開始, 終了時刻を波形上に表示し, その区間の音声を実際に聴いて発話音声かノイズかを判断し, ノイズの場合にはその部分を除去するという作業を行う. たとえば図 2 B には5つの間発話単位 (0010 ～ 0014) があるが, 波形部分をマウスでクリックして1つずつ聴いていくことで, 0012 と 0013 とがノイズであることがわかったとしよう. この時作業者がすることは, これらノイズの部分の境界線をマウスを使って取り除くことだけである. 修正後の図を C に示す. このように, まず波形上で時間情報を修正してから, 新しい時間情報テキストを改めて自動的に作るのである. 図 2 D が修正後の時間情報テキストになる.

[4] この点に関しては, 本冊子中の前川の研究 (「音声によるパラ言語情報の伝達」) を参照して頂きたい.

[5] たとえば千葉大学地図課題コーパス (青野他, 1994) では, 400 ミリ秒以上の無音区間で区切られた発話区間を基本単位として採用している.

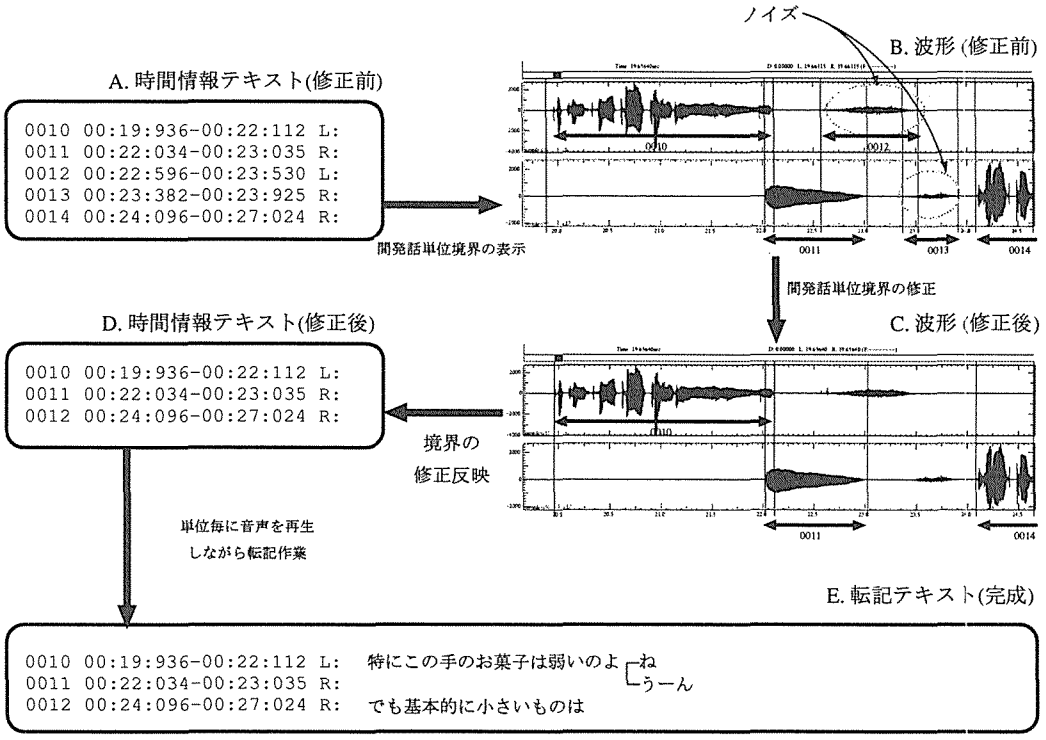

図 2: 書き起こし作業の流れ

最後に, 修正後の時間情報を利用しながら間発話単位の発話音声を順々に聴き, 発話内容を 1 つずつ書き起こしていく (図 2 E).

以上の手続きをふむことで転記テキストが完成する. テープを聴きながら書き起こしをするよりも時間情報などを非常に正確に記述できるうえ, 任意の発話箇所を自在に聴くことができるため, 効率よく転記作業をすることが可能となるのである.

4 おわりに

以上, 本研究室で行っている対話データの構築活動に関して, とくに対話の収録と書き起こし作業に焦点をあて紹介をした. 質・量ともに充実した対話データを構築することが重要な課題とされている現在においては, 構築した対話データだけでなく, データ構築の過程で得られた知見やツールなどをも含めて公開することが望まれる. 本研究室での活動は始まったばかりであり, 今後よりよい対話データを作成するための基準や環境の整備を行う必要があるが, これらの作業過程で得られたデータやツールに関しては, 随時公開していくことを検討している.

参考文献

1. 板橋秀一 (1996). 音声データベース / コーパスとは, 人文学と情報処理 No.12, 6–11.
2. Hanae Koiso, et. al. (in press). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task Dialogues, *Language and Speech*, **41-3,4**.
3. 青野元子他 (1994). 地図課題コーパス (中間報告), 人工知能学会研究会 SIG-SLUD-9402. 25–30.
4. 中里収他 (1994). 協調作業における対話データの収録と分析, 音講論 1-Q-8, 157–8.

表 1: カロリーゲーム対話の転記テキスト (一部)

0004 00:08:272-00:10:864 R:この中でカロリーの高いものですね
0005 00:11:104-00:11:600 L:うーん
0006 00:12:064-00:14:672 R:うーんいろいろありますねー
0007 00:16:304-00:19:152 L:私あんまりねー ┌お菓子って食べないからねー
0008 00:17:600-00:18:112 R: └うーん
0009 00:19:136-00:19:648 R:うーん
0010 00:19:936-00:22:028 L:特にこの手のお菓子は弱いのよ ┌ね
0011 00:22:034-00:23:035 R: └うーん
0012 00:24:096-00:27:024 R:でも基 ┌本的に小さいものは
0013 00:24:640-00:24:864 L: └じゃ
0014 00:27:424-00:32:256 R:一個あたりっていうことだと小さいものは少ない可能性がありますよね
0015 00:32:608-00:32:848 L:うん
(省略)
0112 03:23:936-03:24:672 L:で 9 個で
0113 03:25:488-03:25:952 L:えー
0114 03:26:352-03:27:424 L:どれからってゆう
0115 03:28:352-03:29:264 L:ことになるのね
0116 03:29:216-03:29:808 R:そうね
0117 03:30:032-03:30:224 L:はい
0118 03:30:976-03:31:792 L:とカロリーの
0119 03:32:576-03:33:728 L:高いほう ┌から
0120 03:33:456-03:34:080 R: └高い
0121 03:34:528-03:35:040 R:ほうから
0122 03:35:200-03:36:896 L: ┌高いほう高い順に {呼吸}
0123 03:35:305-03:36:410 R: └高い順にうん
0124 03:37:872-03:40:000 R:太りそうな順に ┌ってことですね
0125 03:39:232-03:40:288 L: └太りそうな順
0126 03:42:080-03:45:088 R:そ ┌ーするとー
0127 03:42:704-03:45:136 L: └うーんとー
0128 03:47:856-03:51:056 R:なんかまあ大きさにもよるのかもしれないけ ┌ど
0129 03:50:608-03:51:088 L: └うん
0130 03:51:776-03:55:616 R:このユーエンジェルパイって大 ┌きそうな感じがしませんか
0131 03:53:920-03:54:288 L: └うん
0132 03:55:680-03:56:000 L:うん
0133 03:56:720-03:57:472 L:エンジェルパイね
0134 03:57:776-03:58:064 R:んね
0135 03:58:400-03:58:912 L:そうね
0136 03:59:296-03:59:776 R:うん
0137 04:00:976-04:03:072 L:エンジェルパイ夕張メロンだよねこれ
0138 04:03:440-04:03:792 R:そう
0139 04:03:936-04:04:096 L:はい
0140 04:04:192-04:06:304 R:エンジェルパイ夕張メロ ┌ン
0141 04:06:144-04:06:368 L: └うん
0142 04:08:688-04:09:392 L:そうですね
0143 04:09:792-04:10:640 L:これはねー
0144 04:10:608-04:10:896 R:ええ
0145 04:11:184-04:12:283 L:食べたことあるけど
0146 04:12:644-04:13:385 L:おっきい ┌よね
0147 04:13:008-04:13:520 R: └大きい
0148 04:13:645-04:13:891 L:うん
0149 04:14:000-04:14:352 R:うーん
0150 04:15:536-04:16:006 L:でー
0151 04:16:195-04:17:936 L:外側がチョコレートでー +
0152 04:17:936-04:18:208 R:ええ
0153 04:19:084-04:19:808 L:中にー
0154 04:20:560-04:24:048 L:メロンジャムみたいなクリーメロンクリームみたいのが入ってて一
0155 04:24:016-04:24:304 R:うん
0156 04:25:136-04:27:536 L:でー間がクッキーみたいになってっ ┌しょー
0157 04:27:216-04:27:604 R: └うん
0158 04:27:726-04:28:529 R:そうそうそう
0159 04:28:800-04:31:863 L:そういうあれだから結構高い ┌かなー
0160 04:30:816-04:32:155 R: └これ高そうですね
0161 04:32:186-04:32:977 L:う ┌ーん
0162 04:32:325-04:32:514 R: └まあ
0163 04:33:056-04:36:976 R:大きくてクリームチョコレート付き ┌という点でこれが
0164 04:35:856-04:36:096 L: └うん
0165 04:37:408-04:39:616 R:一番 ┌のような気がしますね
0166 04:37:952-04:38:256 L: └うん

(1 列目は発話の番号. 2 列目は発話の開始 / 終了時刻. 3 列目は話者 ID と実際の発話の内容. また発話内の記号 " [" は, そこで両者の発話が時間的に重複していることを示す. 資料の前半 12 行が課題の開始部分の対話, 後半はお互いの用紙のお菓子の確認をおえてカロリーの高そうなお菓子を選びはじめた部分である.)