『昭和話し言葉コーパス』の設計・構築と分析 (2)：コーパスの構成とメタデータの設計

丸山岳彦
専修大学・国立国語研究所
maruyama@isc.senshu-u.ac.jp

西川賢哉
国立国語研究所
nishikawa@ninjal.ac.jp

田嶋明日香
国立国語研究所
tajima-a@ninjal.ac.jp

小磯花絵
国立国語研究所
koiso@ninjal.ac.jp

1 はじめに

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」では、2016 年度より『昭和話し言葉コーパス』（Showa Speech Corpus：SSC）の構築を進めてきた。『昭和話し言葉コーパス』は、1950 年代から 1970 年代にかけて国立国語研究所で作成された録音資料（独話・会話）を再編し、話し言葉コーパスとして整備したものである。これまで、2 回のモニター公開を経て、2021 年 3 月にコーパス全体が完成し、Web 上のコーパス検索アプリケーション「中納言」で本公開される。そこで本発表では、『昭和話し言葉コーパス』の設計と構築について、特にコーパスの構成とメタデータの設計を中心に述べる。

2 『昭和話し言葉コーパス』開発の経緯

『昭和話し言葉コーパス』は、1952 年以降、国立国語研究所で作成されてきた録音資料群を収集し、再編した音声コーパスである。1952 年から 1970 年代の初頭にかけて録音された自発的な日本語音声（独話・会話）約 44 時間分を収録しており、20 世紀半ばから現代に至る過程において、日本語の話し言葉がどのように変化してきたかを探るための言語資源として活用することができる [1, 2, 3]。

1948 年に設立された国立国語研究所では、1952 年度に話し言葉を専門的に分析する「第 1 研究室」を設置して以降、さまざまな場面における日常会話や、講演会での講演、挨拶・祝辞などの独話を録音してきた。可搬型のオープンリール型録音機（通称デンスケ、おそらく M1 型）をさまざまな場所に持ち込み、市井の人々の日常会話音声や、国立国語研究所員の講演音声を録音したことが、当時の報告書や『国立国語研究所年報』などに記録されている。特に日常会話の録音作業では、「地区・場所・性・年齢・教養・相手」という条件を設定し、各条件をなるべく広くカバーするような収集方針が取られた。当時の年報には、以下のような記述がある [4]。

> 日常の談話が多く得られる場合として、衣食住・社交等の生活機能と家庭・近隣・職場・市町村などの生活環境との切点から具体的な談話の場面を収集し、また、性・年齢・教養・相手（の数、未知既知）・地域などになるべく片寄りの少いことを目安として、調査地点・調査対象・調査場面の予定表を作成した。（p.6）

言語研究を目的として、日常の多様な場面・多様な話し手から偏りなく会話音声をサンプリングしたこの試みは、世界的に見ても極めて早い時期に行われた話し言葉研究の実践例であったと言ってよい。録音された音声資料は転記され（図 1）、さらにカード式のデータベースとして整理され（図 2）、分析に使用された。この分析結果は、『談話語の実態』（1955 年）、『話しことばの文型 (1)(2)』（1960、1963 年）などの報告書にまとめられている。

当時の音声を録音したオープンリールテープは、国立国語研究所の資料庫に保存されていたが、1990 年代以降、音源を DAT テープにダビングし、デジタル化する作業が進められてきた。この音声データを収集・再編し、『昭和話し言葉コーパス』として整備することを着想したのが 2013 年ごろであり、2016 年度から本格的に構築を開始した。

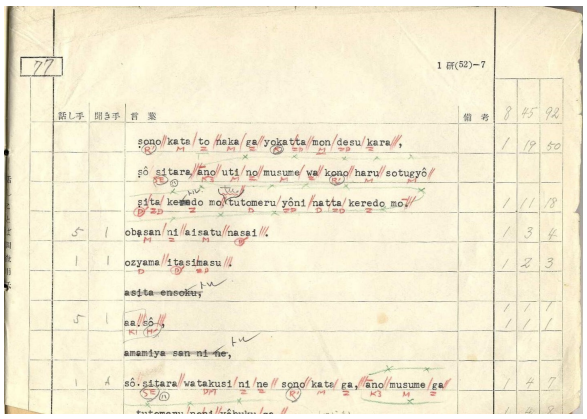

図 1　録音音声の転記

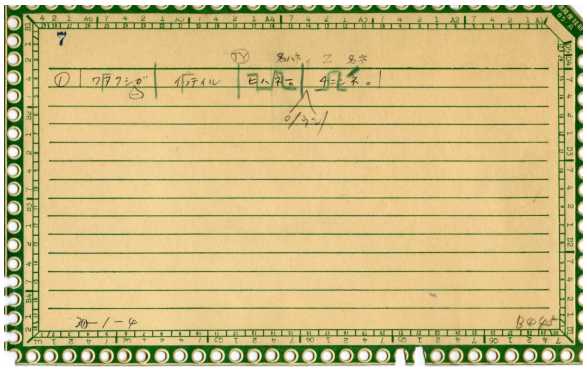

図 2　カード式データベース

3　『昭和話し言葉コーパス』の構成

以下では、『昭和話し言葉コーパス』（以下、SSC と略記する）の構成について述べる。

3.1　音声の種類

SSC の収録音声は、「会話」「独話」に分かれる。

3.1.1　会話

1952 年度に「第 1 研究室」が設置された際、「現代話し言葉の実態と性質を明らかにする。材料としては主として日常談話による」という研究項目が定められた [4]。この方針に基づき、日常談話の録音作業が本格的に開始されたのが、1952 年 9 月である。上述のように、会話参与者の性・年齢・教養・相手（の数、未知既知）・地域などを考慮して、バランスの取れたサンプリングが企図されたようである。これら初期の録音資料群は、1955 年の報告書『談話語の実態』での分析に用いられた。また、1955 年以降の録音資料群は、1960 年の報告書『話しことばの文型 (1) 対話資料による研究』に用いられた。

3.1.2　独話

SSC に収録された独話音声は、「国立国語研究所新庁舎開き記念講演会（1955 年）」「国立国語研究所創立 10 周年記念祝賀式（1959 年）」「国立国語研究所創立 20 周年記念講演会（1969 年）」など、当時の国立国語研究所における講演会・講座や式典・祝賀式などで録音した講義、挨拶・祝辞が主なものである。1963 年の報告書『話しことばの文型 (2) 独話資料による研究』でも分析に利用されている。

SSC に収録された会話音声・独話音声の一覧を、稿末の「SSC 収録音声一覧」に掲載する。

3.2　収録時期

それぞれの収録時期は、以下の通りである。

会話:　1952 年 3 月（録音日不明）～1969 年 12 月 6 日
独話:　1955 年 3 月 26 日～1974 年 6 月 5 日

3.3　転記テキスト

転記テキストは、新規に作成した。当時の転記テキストを再利用することも検討したが、聞き取り困難な箇所が書き起こされていないケースや、数十秒の範囲で転記が欠落しているケースも見られたことから、新規に作成する方がよいと判断した。録音レベルが極めて低い音声や、ノイズが多く混入している音声など、録音状態の悪い音声データについては、SSC への収録対象から除外した。

過去の音声資料を新規に転記する作業には、多くの困難が生じた。ノイズの混入により音声が聞き取りづらい場合だけでなく、現代ではあまり使われない言い回しが特定できない場合も多かった。以下はその例である（これらは転記者たちの調査と努力により、特定することができた）。

1. 次があったら●●●ということを、常に考えているわけでありますが
2. これらを、え、●●●まするために、これも、おー、西尾所長から先ほど、
3. 文化庁でも、おー、その当時、次の、お、国語課長の国松君が●●●●●

また、当時の収録機器はマイク 1 本であるため、特に多人数会話で発話の重複が生じていると、発話内容の聞き取り、発話者の同定が極めて困難であった。これらの場合には、「聴取不能」「発話者不明」と注釈を付与せざるを得なかった。

3.4　アノテーション

音声ファイルと転記テキストが準備できた段階で、(1) 時間情報の付与、(2) 形態素解析、という 2 種のアノテーションを実施した。(1) については、Praat 上で発話単位を区切り、各単位に時間情報を付与した（図 3）。(2) については、形態素解析用辞書「現代話し言葉 UniDic（ver.3.0.1.1）」と MeCab（ver.0.996）により形態素解析を実施した。

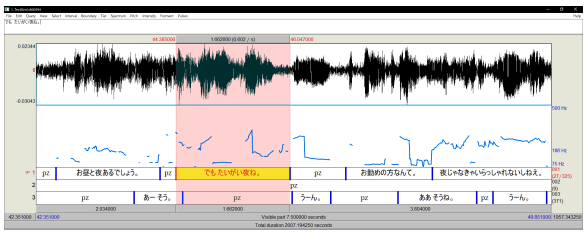

図 3　時間情報の付与

また、会話音声の中で個人情報（人名、住所、電話番号など）が含まれるケースがあった。これらについては、該当する範囲の音声をマスキングし、転記テキストは伏字化の処理を行った。

3.5　データサイズ

SSC のデータサイズについて、表 1 に示す（最終的な公開時には数値が変動する可能性がある）。なお「話者数」は延べ人数を表す。

表 1　SSC のデータサイズ

種別	ファイル数	総時間数	総語数	話者数
会話	73	約 27 時間	353,061 語	344 人
独話	50	約 17 時間	180,675 語	50 人

3.6　メタデータ

通常、音声コーパスには「メタデータ」が付与され、発話者や発話場面に関する情報などが提供される。例えば、『日本語日常会話コーパス』（CEJC）では、発話者の年齢、性別、出身地、居住地、職業、話者間の関係性や、会話場面の具体的な説明などがメタデータとして記録されており、当該の会話データの詳細が分かるように設計されている [5]。

SSC でも、可能な限り詳細なメタデータをコーパスに付与することにした。かつ、特に会話音声については、CEJC におけるメタデータとできる限り項目を揃えることにより、将来的に SSC と CEJC を連結して検索し、比較できるようにした。メタデータの設計と構築については、次節で述べる。

4　メタデータの設計

4.1　メタデータの復元

SSC に含まれる録音資料のうち、独話については、当時の国語研所員による講義音声や、関係者による祝辞のケースが大半である。これらについては、『国立国語研究所年報』や昔の記録写真などから当時の録音状況を推定した結果、すべての録音資料について、録音日、録音場所、発話者情報（氏名、性別、当時の年齢、生年、出身地、職業、肩書）、発話状況（講演のイベント名、講演タイトル）などを特定することができた。

一方、会話については、録音状況の割り出しが非常に困難であった。当初は、最初に入手した音声データに付されていた録音資料の名称しか手掛かりがなかったため、詳細なメタデータの付与は不可能であると思われた。しかしながら、国立国語研究所の中央資料庫に保存されている当時の資料群を探索し、当時の資料群を隅から隅まで吟味した結果、オープンリールの箱に記載されたメモ書き（図 4）や、当時のフェイスシートと思われるメモ（図 5）を発見することができた。

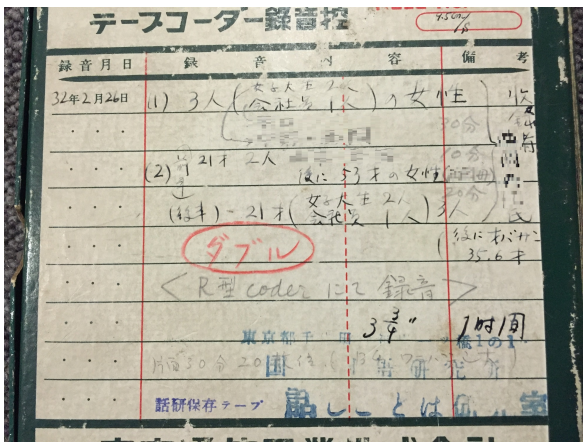

図 4　オープンリールテープの箱に記されたメモ

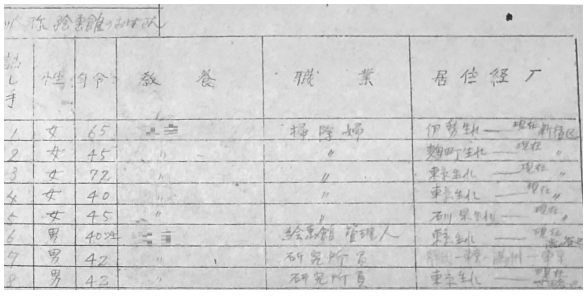

図 5　当時のフェイスシート

図 4 からは、録音年月日と参与者の属性を知ることができる。図 5 のメモは、各発話者の属性（性、年齢、教養、職業、居住歴）が記載されたフェイスシートである。このような断片的に残された情報を丁寧に拾い上げることにより、会話音声のメタデータを整備した結果、当初の見通しからするとはるかに詳細な話者情報を付与することができた。それでも付与できなかった話者属性（年齢、居住地、出身地など）については、推測値を付与する（「壮年層」など）、または「不明」とすることで対処した。

4.2 メタデータの設計

上記のように収集した情報を整理し、SSC のメタデータを設計した。ただし、独話音声と会話音声はメタデータの性質が大きく異なるため、個別に設計することとした。各メタデータに含まれる情報を、以下に示す。

4.2.1 独話音声のメタデータ

- **独話音声データ：**
  ファイル ID、録音年度、音声タイプ、タイトル、話者、話者 ID、話者年齢、イベントタイプ、イベント名、録音年月日、録音場所
- **話者情報データ：**
  話者 ID、話者名、性別、生年、出身地、職業

4.2.2 会話音声のメタデータ

- **会話音声データ：**
  ファイル ID、タイトル、録音年度、会話形式、話者 ID、会話概要、録音場所、話者数
- **話者情報データ：**
  話者 ID、話者名、性別、生年、出身地、居住地、職業、話者の関係性

このうち会話音声のメタデータは、前述の通り、『日本語日常会話コーパス』（CEJC）に付与されたメタデータ（図 6）との連結可能性を考慮して設計した。SSC と CEJC を同じ条件で検索することで、過去 60 年間に日常会話の音声がどのように変化してきたかを比較・対照することができるようになる。

図 6 CEJC のメタデータ

5 おわりに

古い録音資料を収集してコーパス化し、話し言葉の経年変化を実証的に明らかにしようとする試みは、相澤・金澤 (2016) などごく少数の事例しか存在してこなかった [6]。『昭和話し言葉コーパス』は、設計当初から話し言葉の経年変化を探るためのコーパスとしての運用を想定しており、独話音声を『日本語話し言葉コーパス』（CSJ）と比較したり、会話音声を『日本語日常会話コーパス』（CEJC）と比較したりすることによって、独話・会話の話し言葉がどのように変化してきたのかを分析することが可能である。『昭和話し言葉コーパス』の完成によって、「通時音声コーパス」という新しいタイプの日本語コーパスが利用可能になったと言えるだろう。

2020 年度をもって『昭和話し言葉コーパス』の構築作業は完了し、その全体が 3 月に「中納言[1)]」で公開される。利用登録をすれば、『昭和話し言葉コーパス』をオンラインで検索することができ、その音声を聴取することもできる。ぜひご活用いただきたい。

参考文献

[1] 丸山岳彦. 『昭和話し言葉コーパス』の計画と展望—1950 年代の話し言葉研究小史—. 専修大学人文科学研究所月報, Vol. 282, pp. 39–55, 2016. http://doi.org/10.34360/00007004.

[2] 丸山岳彦. 「通時音声コーパス」の可能性と問題点—『昭和話し言葉コーパス』の構築と分析. 言語資源活用ワークショップ 2019 発表論文集, pp. 402–412. 国立国語研究所, 2019. http://doi.org/10.15084/00002592.

[3] 丸山岳彦. 『昭和話し言葉コーパス』の設計・構築と分析. 言語処理学会第 26 回年次大会発表論文集, pp. 629–632. 言語処理学会, 2020. https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/F3-3.pdf.

[4] 国立国語研究所. 昭和 27 年度 国立国語研究所 年報 4. 国立国語研究所, 1952. http://id.nii.ac.jp/1328/00001164/.

[5] 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. 『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, Vol. 18, pp. 17–33, 2020. http://doi.org/10.15084/00002540.

[6] 相澤正夫, 金澤裕之（編）. SP 盤演説レコードがひらく日本語研究. 笠間書院, 2016.

1) https://chunagon.ninjal.ac.jp/

『昭和話し言葉コーパス』収録音声一覧

独話音声

- 国立国語研究所新庁舎開き式典（1955）
  - 所長挨拶ならびに経過報告（西尾実）、祝辞（松村謙三、茅誠司、村上俊亮、土岐善麿、柳田国男）
- 国立国語研究所新庁舎開き記念講演会（1955）
  - 講演者紹介（平井昌夫）、所長挨拶（西尾実）、現代の敬語意識（柴田武）、三つの語彙調査（林大）
- 全国国語科指導主事研修講座（1957）
  - 話し言葉の表現意図について（飯豊毅一）、方言調査法（野元菊雄）、言語能力の発達（芦沢節）、助詞・助動詞（宮地裕）、文型（永野賢）、国語教育（輿水実）、新聞文章研究法（林四郎）
- 国立国語研究所創立 10 周年記念祝賀式（1959）
  - 所長挨拶（西尾実）、祝辞（橋本龍伍、兼重寛九郎、関口隆克、時枝誠記、山本有三）、来賓挨拶（土岐善麿、安倍能成、片山哲）
- 国立国語研究所創立 10 周年記念講演会（1959）
  - 所長挨拶（西尾実）、明治初期の書きことば（山田巌）、現代語の標準（林大）、話しことばの文法（大石初太郎）、これからの日本語（岩淵悦太郎）
- 第 17 回国立国語研究所創立記念講演会（1965）
  - 所長挨拶（岩淵悦太郎）
- 第 18 回国立国語研究所創立記念講演会（1966）
  - 所長挨拶（岩淵悦太郎）
- 第 19 回国立国語研究所創立記念講演会（1967）
  - 所長挨拶（岩淵悦太郎）、講演者紹介（岩淵悦太郎）
- 見坊氏退官記念講演（1968）
  - 見坊氏退官記念講演（見坊豪紀）
- 国立国語研究所創立 20 周年記念講演会（1969）
  - あいさつ－研究所と語彙研究－（岩淵悦太郎）、語彙調査と基本語彙（林四郎）、形容詞の意味の特質（西尾寅弥）
- 第 21 回国立国語研究所創立記念日（1969）
  - 講演者紹介（岩淵悦太郎）
- 第 24 回国立国語研究所創立記念日（1972）
  - 所長挨拶（岩淵悦太郎）
- 国立国語研究所研究棟落成式典（1974）
  - 所長挨拶（岩淵悦太郎）、来賓挨拶（奥野誠亮）、祝辞（藤沢達夫、安達健二、平塚益徳、久松潜一）

会話音声

1951 年度　一研雑談
1952 年度　三鷹分室
Y 理髪店
N 家雑談
三鷹学生
接客用語について
扇子屋
絵画館のおばさん
S 女子大生
I 家雑談
U 夫妻
S 女子大事務室
魚屋小僧
女性雑談
トタン屋
A 美髪店
ジイサン・バアサン
友の会
U 家雑談
1953 年度　男女学生座談
魚屋
1954 年度　K 高校生
1955 年度　組合団交
1956 年度　3 人の女性
劇団員雑談
1957 年度　面接録音調査
T 社応接室
タクシー苦情
歯科大学生
麻布主婦
K 教育委員会雑談
鎌倉主婦
研究室の電話
養老院
少年工員
下町家族
3 人の青年
1960 年度　浅草噺
1969 年度　その電気ごたつは安全ですか