国立国語研究所学術情報リポジトリ

Design and Construction of the Showa Speech Corpus

メタデータ	言語: jpn 出版者: 公開日: 2022-01-21 キーワード (Ja): キーワード (En): 作成者: 丸山, 岳彦, 小磯, 花絵, 西川, 賢哉, MARUYAMA, Takehiko, KOISO, Hanae, NISHIKAWA, Ken'ya メールアドレス: 所属:
URL	https://doi.org/10.15084/00003522

国立国語研究所論集 (NINJAL Research Papers) 22: 197–221 (2022)
ISSN 2186-1358

『昭和話し言葉コーパス』の設計と構築

丸山岳彦[a]　小磯花絵[b]　西川賢哉[b]

[a] 専修大学／国立国語研究所 研究系 音声言語研究領域 客員教授
[b] 国立国語研究所 研究系 音声言語研究領域

要旨

国立国語研究所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」では，2016 年度より『昭和話し言葉コーパス』（SSC: Showa Speech Corpus）の構築を進めてきた。2021 年 3 月にその構築作業が完了し，コーパス検索アプリケーション「中納言」で一般公開を開始した。『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群を再編成し，現代の技術で話し言葉コーパスとして整備したものである。過去の音源を現代の技術でコーパス化したという点において，日本語では従来存在しなかったタイプのコーパスであると言える。また，現代の話し言葉コーパスと連結し，比較・対照することによって，話し言葉の経年変化を探るための「通時音声コーパス」として利用できる点で，画期的である。本稿では，今回構築した『昭和話し言葉コーパス』について，そこに収録されている録音資料群の出自や当時の国立国語研究所の状況，コーパス構築の過程とアノテーション，さらに予備的な分析結果について述べる *。

キーワード：『昭和話し言葉コーパス』（SSC），『談話語の実態』，アノテーション，話し言葉の経年変化，通時音声コーパス

1. はじめに

2000 年代に入って以降，国立国語研究所においてさまざまな日本語コーパスの整備が進められている。2004 年に公開された『日本語話し言葉コーパス』（CSJ）を皮切りとして，『現代日本語書き言葉均衡コーパス』（BCCWJ），『日本語歴史コーパス』（CHJ），『日本語日常会話コーパス モニター公開版』（CEJC），『日本語諸方言コーパス モニター公開版』（COJADS），『多言語母語の日本語学習者横断コーパス』（I-JAS）など，多様なコーパス群が開発されてきており，コーパス検索アプリケーション「中納言」の上で一般に公開されている。組織的に開発された大規模な日本語コーパスが相次いで公開されたことにより，分析者は自分の利用目的に合ったコーパスを選択できるようになった。多様な日本語コーパスの開発・公開と，それに基づく分析範囲の拡大という流れは，今後もしばらく続くものと思われる。

このような状況の中で，筆者らが 2016 年から構築を進めてきた。『昭和話し言葉コーパス』（SSC: Showa Speech Corpus）が完成し，2021 年 3 月，コーパス検索アプリケーション「中納言」

* 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」（プロジェクトリーダー：小磯花絵），および科研費基盤研究（B）「『昭和話し言葉コーパス』の構築による話し言葉の経年変化に関する実証的研究」（16H03426）によるものである。また，本稿の一部は丸山（2015, 2016, 2019, 2020），丸山他（2021）で報告してきた内容を取りまとめ，大幅な加筆・修正を加えたものである。

で一般に公開された。これは，1950 年代から 1970 年代にかけて国立国語研究所において作成された古い録音資料群を再編成し，各種のアノテーションを施した，約 44 時間分の話し言葉コーパスである。過去の録音資料を収集してコーパス化したという点において，『昭和話し言葉コーパス』は，日本語では従来存在しなかったタイプのコーパスであると言える。

古い録音資料を大量に文字化して日本語研究に利用した例としては，相澤・金澤（2016）が挙げられる。相澤・金澤（2016）では，大正期から昭和前期にかけて「SP レコード」に記録された演説音声（約 18.5 時間分）が文字化され，多角的な観点から分析が行われている。しかしながら，話し言葉コーパスとしてのアノテーション・構造化がされていないこと，音声データ・文字化データが一般に公開されていないこと，などの問題点がある [1]。

今から約 60 年前の日本人は，どのように話をしていたのだろうか。そこから現代に至る過程で，日本語の話し言葉にはどのような経年変化が生じたのだろうか。これらの点をコーパス言語学的な手法によって明らかにするためには，過去から現代に至る複数の録音資料を時系列に並べた「通時音声コーパス」が必要になる。今回完成した『昭和話し言葉コーパス』を，『日本語話し言葉コーパス』『日本語日常会話コーパス』など現代の話し言葉コーパス群と連結することにより，日本語では初めての「通時音声コーパス」が実現できることとなった。

そこで本稿では，今回我々が構築した『昭和話し言葉コーパス』について，そこに収録されている録音資料群の背景（2 節），コーパスの設計と構築（3 節），さらに予備的な分析結果（4 節）について述べる。

2.『昭和話し言葉コーパス』の背景

本節では，『昭和話し言葉コーパス』に収録された 1950 年代から 1970 年代の録音資料群について，当時の国立国語研究所の状況や録音資料の収集方法などを示し，その出自を解説する。

2.1 1950 年代の国立国語研究所における話し言葉研究

1952 年，国立国語研究所において共通語の話し言葉研究を目的とした「第 1 研究室」が設置された。1955 年に「話しことば研究室」と改称されるこの研究室で実施されたのは，日常談話や独話を大量に録音して書き起こし，そこに見られる韻律・語彙・文法などについて定量的な観点から分析するという，記述的な研究であった。その最初の研究成果は，1955 年の研究報告書『談話語の実態』として刊行されている。以下では，1950 年代前半の研究所年報や，『談話語の実態』の記述に基づいて，当時の研究方法を概観してみよう。

録音作業に際して，最初に行われたのが，どのような種類の日常談話を集めるか，その選定方針の検討であった。これについては，以下の記述および図 1 がある。

> 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

[1] SP レコード音源の一部は，国立国会図書館のウェブサイト「歴史的音源」で聴くことができる。https://rekion.dl.ndl.go.jp/

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。（『昭和 27 年度 国立国語研究所年報 4』p.6）

Reel No.	略称	録音状態の否	地区 下町	山手	周辺	場所 家庭	近隣	学校	職場	公共施設	性 男男	女男	男女	女女	年齢 若若	壮若	若壮	壮壮	教養 義義	義専	専義	専専	相手 1人	2～4人	5～7人	8人以上	既知 未知
3	I 夫妻	可		×		×							×	×	×					×			×					×
7	T家雑談	可		×		×					×	×	×	×			×	×			×	×		×				×
67	N家座談	可			×	×					×		×	×		×	×	×	×		×			×			×	×
86	トタン屋	可			×		×				×						×	×			×	×		×			×	
76	じいさんばあさん	可			×	×					×	×	×	×		×					×			×			×	×
93	魚屋雑談	可			×		×						×	×	×						×	×		×				×
97	U 氏談	可			×	×					×			×		×				×				×			×	×
61	学生座談 1	可			×		×				×				×					×					×			×
66	井戸端	可			×		×					×					×	×	×	×	×	×			×			×
98	女の会	可			×		×					×						×	×	×	×	×		×				×
2	女子学生	可			×			×				×			×				×	×					×			×
59	松根屋	可	×							×	×					×				×		×			×			×
57	高島屋	可	×							×	×		×	×		×	×	×			×	×				×	×		
100	物語美髪	可	×							×	×	×					×	×	×				×	×				×
図5	絵画館おばさん	可		×						×		×	×			×			×		×	×			×		×	×
25	床屋	稍可		×						×	×						×	×	×		×	×	×					×
51	一研雑談	可		×						×	×		×	×			×	×		×				×				×
64	三醫女工	稍可			×					×		×			×				×					×				×
62	学生座談 2	稍可			×		×				×		×	×	×		×	×		×				×				×

図 1　録音資料一覧表（『昭和 27 年度 国立国語研究所年報 4』p.8，一部）

ここから見て取れるのは，できるだけ多様な場面から日常談話を収集し，さまざまな言語的特徴を内包した分析用資料を作ろうとする姿勢である。図 1 を見ると，地区・場所・性・年齢・教養・相手という大分類の下に複数の項目が設けられ，その範囲を広くカバーするように録音資料が収集されていることが分かる。異なる言語的特徴を持つ対象を複数の基準によって区分し，対象全体の多様性をできるだけサンプルに反映させようとする方法論は，現代でもそのまま通用するものと言える。

図 2 は，1955 年の『国立国語研究所 要覧』に掲載された録音作業の風景である。写真を見る限り，録音作業に使用されたのは東京通信工業株式会社（現ソニー株式会社）の可搬式オープンリール型録音機「M-2 型」（通称「デンスケ」）のようである。録音に使用されたのは 80 巻のオープンリールテープで，約 40 時間分に相当する。当時としては非常に大規模な資料だったと思われる。

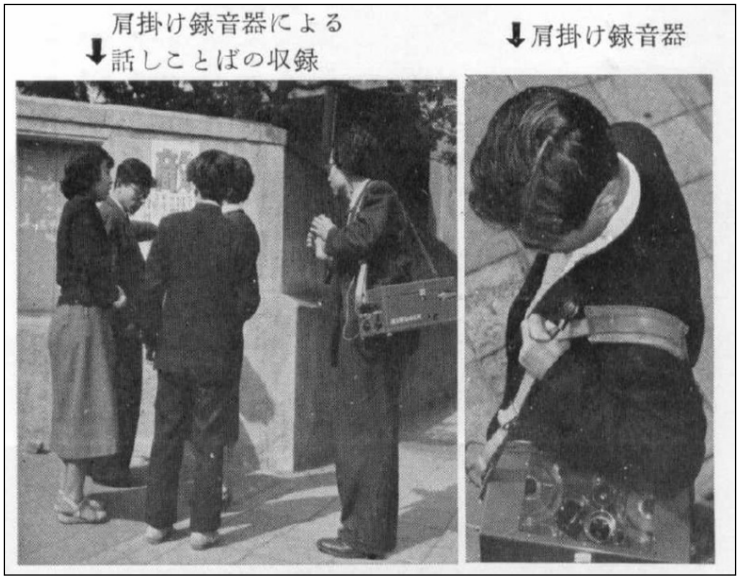

図 2 「話しことばの収録」（『国立国語研究所 要覧』1955 年）

さらに，日常談話と比較するための資料として，講義，講演，ニュース，ニュース解説などが録音された。これらは独話の資料として位置付けられ，会話資料との比較・分析に用いられた。

録音された音声は，機械速記「ソクタイプ」によりローマ字で文字化され，カードに転記された上で，語・文節・文の境界が人手で認定された。さらに後年，各文が 1 枚のカードにカタカナで転写され，文節境界，構文情報，イントネーションなどの情報が書き込まれた。その分析結果をまとめたものが，1955 年に刊行された『談話語の実態』である。この報告書の中では，「イントネーション」「文・文節・語の長さ」「文の構造」「語の種類・使用度数・用法」といった分析項目が立てられ，定量的な分析結果が示されている。文長などの分析には 18 巻分（10,118 文）の日常談話が，語の種類の分析には 20 巻分（83,620 語）の録音資料が，それぞれ用いられた。さらに，各項目の分析では，発話者の性別や年齢，発話場面などの違いによってイントネーション・語彙・文法がどのように異なる分布を示すかが記述されている。

また，『談話語の実態』に続く話し言葉研究の成果として，『話しことばの文型（1）―対話資料による研究―』（1960 年），『話しことばの文型（2）―独話資料による研究―』（1963 年）が刊行されている。ここでは，話し言葉の「総合文型」を明らかにすることを目的として，より多くの録音資料をもとに詳細な文法記述が行われた。この『話しことばの文型（1）（2）』では，観察されたさまざまな言語現象を定量的に分析・記述するだけでなく，そこで見出された文の構造を一般化・抽象化し，日本語の文法構造を説明するモデルを構築する段階にまで至っている。これは後年，南（1974, 1993）で広く知られることになる，南不二男による「文の階層構造モデル」（通称「南モデル」）の原型となった。現代日本語文法研究における基礎的な知見の一つとして知られる南モデルが，話し言葉コーパスの観察の中から生まれてきたという事実は，コーパスが一般

言語学的研究に極めてうまく作用した例として記憶されるべきだろう（丸山 2014）。

2.2 1950 年代における話し言葉研究に対する評価

ここまで見てきたように，1952 年に開始された国立国語研究所による一連の話し言葉研究は，現代におけるコーパス言語学の研究プロセスとほぼ同様の手続きで行われていたと言える。複数の基準によってバランスよく録音資料を収集する作業や，音声を転記した結果に対して語･文節･文の境界をマークしたり，イントネーションの型を付与したりする作業は，現代のコーパス言語学で言うところのサンプリング作業・アノテーション作業に等しい。発話者の年齢や性別，発話場面などによって対象を区分し，そこに見られる言語変異を定量的に捉えようとする方法論は，後年の社会言語学を先取りしていたとも言える。

そもそも，言語分析のために録音資料を大量かつ組織的に収集するという作業自体，世界的に見ても極めて早い時期の試みであったと思われる。イギリスの UCL（University College London）において，Randolph Quirk らによって “Survey of English Usage”（SEU）計画が開始されるのは 1959 年のことである。Quirk らはその後，イギリス英語の書き言葉と話し言葉とを 50 万語ずつ集めた 100 万語のコーパスを作成し，コーパス言語学の端緒を開いたとされるが，1955 年刊行の『談話語の実態』はそれよりもはるかに早い。1950 年代に開始された国立国語研究所における話し言葉研究は，コーパスに基づく話し言葉研究の源流として位置づけられるものであると言える。

一方，惜しまれるのは，作成された録音資料を再利用･一般公開するという考え方がなく，せっかくの録音資料が研究者間で共有されることがなかったという点である。当時，調査のために収集された言語資料は，分析結果の報告書が出版されると，倉庫に入れられることが通例であったという（宮島 2007）。1950 年代から 60 年代にかけて，国立国語研究所では書き言葉の語彙調査も盛んに行われていたが，そこで作成された膨大な言語資料（用例カード）も，国立国語研究所（1972a, b）など少数の例外を除けば，再利用はされなかった。「話しことば研究室」で作成された録音資料群もまた，保管されている録音資料の一覧が作成されたり（国立国語研究所 1969），1980 年代後半に「文脈付き用語索引」としてローマ字･カタカナで転記された KWIC 形式のデータがマイクロフィッシュ化されたり（国立国語研究所 1990）したものの，その中身が具体的な分析対象として扱われることはなく，長い間「お蔵入り」の状態であった。

2.3『昭和話し言葉コーパス』の構築計画

1999 年，国立国語研究所を中心に『日本語話し言葉コーパス』（CSJ: Corpus of Spontaneous Japanese）の構築が開始され，約 661 時間・752 万語という大規模な話し言葉コーパスが 2004 年に公開された（国立国語研究所 2006）。さらに 2016 年，『日本語日常会話コーパス』（CEJC: Corpus of Everyday Japanese Conversation）の構築が開始された。これは，さまざまな話者・場面による約 200 時間分の日常会話をビデオ付きで収録したコーパスであり，2021 年度末の完成・公開を目指して構築作業が進められている（小磯他 2017）。CEJC が完成すれば，CSJ とともに，

会話・独話の広い範囲をカバーする現代日本語の話し言葉コーパスが揃うことになり，コーパスに基づく話し言葉研究のさらなる拡大が期待される。

これに対して，古い時代の録音資料を集めて話し言葉コーパスとして整備すれば，過去における話し言葉の実態を分析するための歴史コーパスを得ることができる。さらにそのようなコーパスを現代日本語の話し言葉コーパス群（CEJC・CSJ）と連結すれば，話し言葉がどのように変化したかを探るための「通時音声コーパス」が実現できることになる（丸山 2015）。異なる時代の録音資料を集めて編纂した通時音声コーパスは，世界的に見ても実践例が少なく，イギリスの “DCPSE”（Diachronic Corpus of Present-day Spoken English）やフランスの “ESLO”（Enquêtes Sociolinguistiques à Orléans）など，ごく少数しかない。

そこで筆者らは，かつての「話しことば研究室」において作成された録音資料群に着目した。『談話語の実態』『話しことばの文型（1）（2）』における分析対象データとして集められた録音資料群は，現代の視点から見れば約 60 年前の話し言葉の音声データであり，貴重な研究資料になることは間違いない。前述の通り，当時の録音資料群（特に日常会話）はサンプリングの過程を経ており，データの多様性も確保されている。これらを新たに話し言葉コーパスとして整備して，CEJC・CSJ と連結すれば，通時音声コーパスとしての運用が可能になる。

以上のような動機に基づいて，過去の録音資料群を収集・再編し，新しい話し言葉コーパスを構築することを計画した。このコーパスを『昭和話し言葉コーパス』（SSC: Showa Speech Corpus）と名付け，2016 年度より構築作業を開始した。

3.『昭和話し言葉コーパス』の設計と構築

本節では，『昭和話し言葉コーパス』（SSC）の設計および構築の過程について述べる。実際の構築作業は，（1）過去の録音資料の収集，（2）転記テキストの作成，（3）アノテーション（発話単位の認定，転記テキストに対する各種タグの付与，形態素解析による形態論情報の付与），という順序で進め，これと並行して（4）メタデータの設計と付与を行った。以下では，各作業の内容について述べる。

3.1 過去の録音資料の収集

SSC の構築にまず必要となるのは，1950 年代以降，オープンリールテープに録音されていた録音資料群を収集し，デジタル化する作業である。幸いなことに，1990 年代以降，国立国語研究所情報資料部門（当時）において，オープンリールテープに記録されていた過去の録音資料群を DAT に複製し，デジタル化する作業が進められていた。当時の作業記録を見ると，すでにテープが劣化していて音声が聴取できなかったり，再生中にテープが切れてその場で修復したりするなど，大きな困難が伴ったようである。DAT に複製された音声データは，さらに WAV ファイルに変換され，所内に保存されていた[2]。

[2] 音声データの収集には，熊谷康夫氏，井上文子氏，磯部よし子氏らの助力を得た。記して感謝申し上げる。

デジタル化された音声データを集めてみたところ,『談話語の実態』『話しことばの文型 (1) (2)』のために作成された録音資料群や，その後 1970 年代まで断続的に作成されていた録音資料（特に創立記念式典での祝辞や挨拶，記念講演会での講演など），約 40 時間分の会話（電話会話を含む），約 25 時間分の独話を集めることができた。このほかに，録音レベルが低く声が聞き取れない音声データや，テープの劣化のためかノイズだらけの音声データもあったが，これらは研究利用に耐えないと判断し，SSC への収録は断念した。

収集できた録音資料のうち，できる限り会話の多様性を確保した会話と独話をそれぞれ 25 時間ずつ選び出し，合計 50 時間分の音声データを収録することにした。ただし，国立国語研究所外の専門家が学術的な講演をしている独話資料（約 8 時間分）は，著作権保護の観点から，公開対象には含めないこととした。

3.2 転記テキストの作成

次に必要となるのは，音声の転記作業である。まず，国立国語研究所の中央資料庫に保管されている『談話語の実態』関連資料を調査したところ，1950 年代当時の転記テキスト（紙媒体）をいくつか発見することができた。図 3 に例を示す。

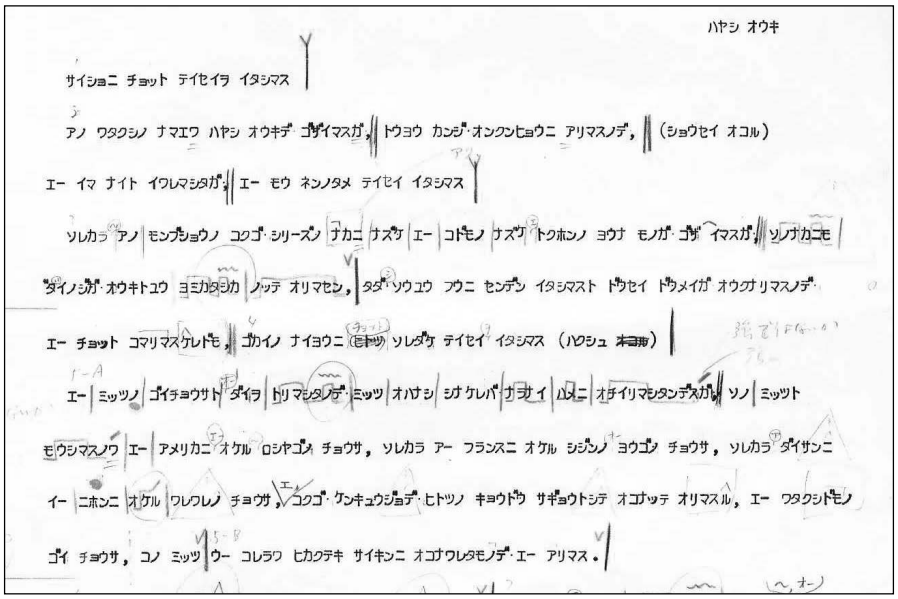

図 3　1950 年代の転記テキスト（1955 年 3 月録音，林大「三つの語彙調査」）

これらの転記テキストを再利用できるかどうかを検討したが，すべての音声データに対応する転記テキストを見つけることができず，また転記の正確性に問題があることも判明したため（後述），総合的な作業効率を考慮して，新たに転記テキストを作成し直すことにした。

転記テキストを新規に作成するに当たり，以下のような基準で第一次作業を行った。

- 転記作業者が音声を聞いて，可能な限り忠実に音声を転記する。
- フィラーや語の言いさし，非語彙的な引き延ばし，笑いや咳なども転記の対象とする。
- 会話データでは，話者ごとに ID を付与して各話者の発話を分けて転記する。
- 転記テキストは，漢字かな交じり文で表記する。
- 「私（ワタシ・アタシ・ワタクシ）」「白（ヒロ）」のように，漢字に複数の読みがある場合や標準的な読みとは異なる発音は，気づいた範囲でその読みも併せて転記する。

このうち，漢字かな交じり文による転記の過程で表記のゆれが生じた場合，3.3 節で述べるアノテーション（形態素解析）の段階で，可能な限り表記を統一した。語の言いさし，非語彙的な引き延ばし，笑いや咳などについても，アノテーションの段階でそれぞれ特別なタグを付与することにした。

実際に転記テキストの作成を開始すると，以下の二つの問題が頻繁に生じた。

（1）音声が不明瞭で聞き取りが困難な場合
（2）発話者の割り当てが困難な場合

（1）は，録音された音量が小さかったり，複数人の発話が重複したりして，発話内容が全く聞き取れない（あるいは聞き取りに自信がない）場合である。これらは，転記作業者が繰り返し音声を聞くことによって発話内容が同定できたケースもあったが，どうしても聞き取れなかった箇所・聞き取りに自信がない箇所には，「聴取不能」「聞き取りに自信がない」ことを表すタグを挿入することで対処した。

（2）は，特に多人数会話の場合に，どの話者がどの発話をしたのかが同定できないという場合である。例えば，「絵画館のおばさん」という録音資料（1952 年 9 月録音）は，5 人の「神宮外苑絵画館掃除婦」の雑談を録音したものであるが，壮年層の女性 5 人が入れ替わりで（または各自が勝手に）しゃべり，そこに複数人による相槌が重複して打たれるので，どの発話がどの話者によるものかがまったく同定できない事例が多く生じていた。CSJ や CEJC では話し手がヘッドセットマイクや IC レコーダーを個別に装着しているため，個々の話し手による発話がクリアに録音されているが，1950 年代当時は 1 本のマイクだけで録音していたため，このような問題が生じるのは不可避的と言わざるを得ない。最終的には，転記作業者が最も適切だと判断した発話者に当該の発話を割り当てることで対処した。

なお，1950 年代当時の転記テキストが見つかった分について，実際の音声と照合してその内容を確認したところ，今回の作業で聞き取れなかった音声は当時も聞き取れていなかったらしく，実際の発話内容が大幅に省略された形で転記されているケースが散見された。この点について，『昭和 28 年度国立国語研究所年報 5』には，以下の記述がある（p.17）。当時も聴取不能の問題には対処できておらず，前述の通り，転記の正確性に問題があることを認識していたようである。

一般に，資料の中には聴取困難あるいは不能の個所が挿入されており，この部分は分析の対象としなかったが，この聴取不能の個所および聴取不能の発言に，話しことばの大きな問題が含まれていると考えられる。そういうものも何等かの方法でつきとめるべきであった。

3.3 アノテーション

次に，作成した転記テキストに対して実施したアノテーション（発話単位の認定，転記テキストに対するタグの付与，形態素解析による形態論情報の付与）について述べる。

3.3.1 発話単位の認定・転記テキストに対するタグ付与

国立国語研究所内において作業班（11 名の作業者と 2 名の監督者）を組織し，転記テキストと音声とを照合しながら，転記テキストの確認・修正作業を行った。この際，音声分析用ソフト Praat を用いて，200 ミリ秒以上のポーズおよびその末尾の文法形式に従って「発話単位」の種類を認定し，さらに各単位の時間情報（開始時刻・終了時刻）を付与した。

発話単位とは，一連の発話を統語的・意味的なまとまりを備えた範囲に切り分けた単位のことであり，CEJC・CSJ にも付与されている。SSC では，発話中にポーズが生じた個所が文末表現に準じる言い切りの形になっている場合には句点を，「〜ですが」「〜けれども」など切れ目の大きい従属節境界になっている場合には読点を，それ以外でおよそ 200 ミリ秒以上のポーズが空いている場合には記号を付けずに発話を区切ることにより，発話単位の認定・分類を行った。これは，CSJ の構築時に実施した節単位認定（絶対境界，強境界，弱境界の区別）に相当するものである（丸山・高梨・内元 2006）。

発話単位の認定作業と同時に，CEJC の転記作業で採用されている仕様に準じて，表 1 に示す各種タグを転記テキストに付与した。なお，会話音声の中に個人情報（人名，住所，電話番号など）が含まれる場合には，該当する範囲の音声をマスキングし，転記テキストは「(R ＊＊＊)」のように対応するモーラ数にあわせて「＊」で置換することで伏字化の処理を行った。

Praat 上で発話単位を認定し，時間情報を付与した結果の例を図 4 に示す。

表 1　転記テキストで用いるタグの一覧（臼田他（2018）p.182 一部改変）

タグ	概要	使用例
1. 非語彙的な発音の変化や言いよどみに関わるもの		
:	非語彙的な母音の引き延ばし	お盆 : で，あの :
%	非語彙的な音の詰まり	い % くら，小さく % て
(W)	一時的な発音エラー	(W タブ | たぶん)，(W カ | から)
(D)	語の言いさし	(D ビ) ビール，(D イシ) 一合
2. 韻律・パラ言語的情報に関わるもの		
(L)	笑いが生じている箇所	(L)，(L はい)
(S)	歌いながらの発話	(S ブランコ)，(S フ フ フーン)
＜＞	発音に類する行為	＜咳＞，＜喃語＞
3. 聞き取り等の判断の信頼性に関わるもの		
(U)	聞き取りに自信がない箇所	(U 多少)，(U 電車ん乗って)
(X)	語が不明な箇所，聴取不能	(X ＃) ですよ，(X ＃) の様子
4. 転記テキストの可読性や内容理解の補助に関わるもの		
(K)	漢字表記できない箇所	(K マ : ル | 丸) く，(K イ % チ | 一) 番
(M)	音や言葉のメタ的な引用	(M えー) とかいうふうな言葉
(O)	外国語・方言	(O ブエナス ノーチエス)
5. 発話単位・転記単位に関わるもの		
。	発話単位末（絶対境界）	何を話しますか。，共通の話題を。
、	発話単位末（強境界）	六年たちましたけれども、，思いますが、
6. 形態素解析のための作業上のもの		
(Y)	複数の読み方が存在する発音	(Y ワタクシ | 私)，(Y サンジッ | 三十)
7. その他，個人情報保護やコメントに関わるもの		
(R)	個人情報などを伏字化した箇所	(R ＊＊) ちゃんの (R ＊＊＊) 先生が
@	転記テキストに対するコメント	(M 学生)。@「ク」は無声母音
＜＞	その他の注記	＜録音途切れ＞，＜飴屋の太鼓の音＞

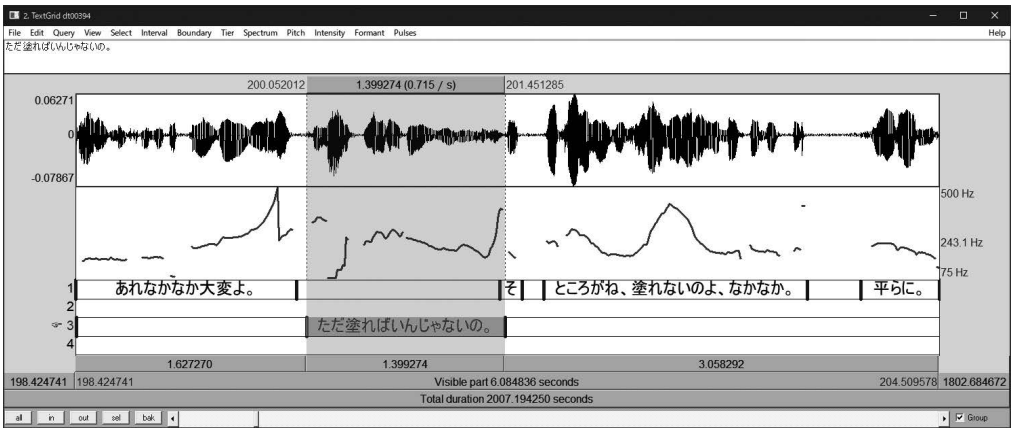

図 4　Praat による発話単位の認定と時間情報の付与

3.3.2 形態論情報の付与

こうして作成された転記テキストに対して，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。解析結果は国立国語研究所内のサーバーに格納し，7 人の作業者による解析結果のチェック，および誤解析の修正・確認作業を実施した上で，各種情報をタブ区切りで格納した形態論情報データベースの形に整備した。

3.4 メタデータの設計と付与

録音資料を分析する際，その音声の録音年月日，録音場所，発話者の属性（性別，年齢，職業，出身地など），発話状況などの「メタデータ」を参照できるようになっていることは必須の条件である。すなわち，話し言葉コーパスにはできるだけ詳しくメタデータが付与されていることが求められる。SSC の場合，問題は，過去の録音資料についてどれだけのメタデータが得られるか，という点であった。

このうち独話の録音資料については，当時の国立国語研究所の所員が講演をしたり，著名な関係者が祝辞を述べたりしているケースが大半であるので，当時の年報や記録写真などを手掛かりにして，詳細な情報を明らかにすることができた。この結果，すべての独話データに対して，録音年月日，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）などをメタデータとして付与することができた。

問題は，会話の録音資料である。前述の通り，1952 年から開始された録音作業は，当時の市井の人々の会話を録音したものが大半である。録音作業の現場において，録音内容に関する詳細な情報が記録されていたのか，発話者に関する情報（フェイスシート）が収集されていたのかすら，当初は全く分からなかった。

ところが，国立国語研究所の中央資料庫に保存されている当時の関連資料群を探索した結果，図5のようなメモ書きを複数発見することができた[3]。上段はオープンリールの箱に入っていたメモ書き，下段は未整理の状態で封筒に入れられていた雑多な資料群の中から見つけたものである。

[3] 中央資料庫で当時の資料群を探すよう助言してくださったのは，前川喜久雄氏である。記して感謝申し上げる。

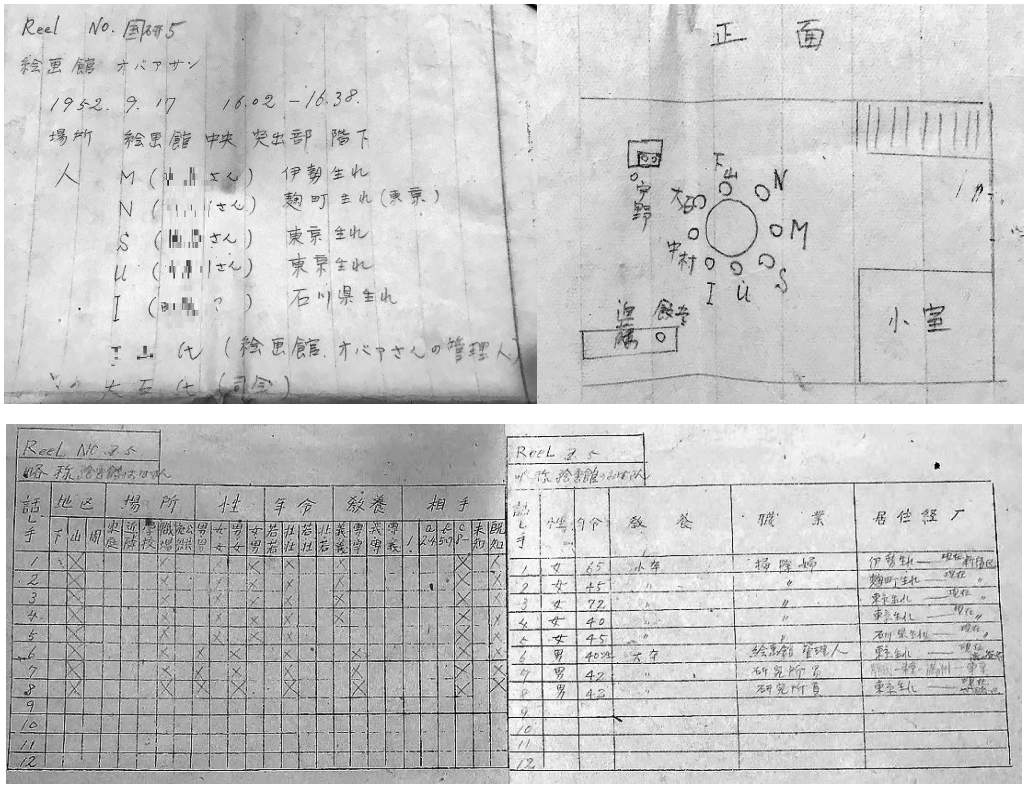

図 5　中央資料庫に保存されていた録音時のメモ（1952 年 9 月録音「絵画館のおばさん」）

図 5 上段のメモ書きからは，録音年月日と場所，録音に参加した発話者の苗字（図中では伏字にしてある）と出身地，録音時の状況を知ることができる。下段の左側は，おそらく，図 1 の一覧をまとめる際の原本となった資料で，録音作業の現場で取っていたメモであろう。下段の右側のメモには，各発話者の属性（性，年齢，教養，職業，居住歴）が記載されている。このようなメモ書きの存在は，録音時においてかなり詳細にメタデータ（フェイスシート）が記録されていたことを裏付けるものである。

断片的に残されたこれらの情報を丁寧に収集・整理し，会話音声のメタデータの整備を進めた結果，当初の見通しよりもはるかに詳細な情報を付与することができた。手掛かりが得られず，最後まで話者属性（年齢，居住地，出身地など）が付与できなかった分については「不明」とし，追加情報として音声を入念に聴取した作業者が推測値（「壮年層」など）を付与した。

上記の情報を整理し，SSC のメタデータを設計･整備した。ただし，会話と独話とで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。SSC（会話・独話）のメタデータに含まれる情報を，表 2 に示す。これらのうち特に会話については，CEJC に付与されたメタデータとの連結可能性を考慮して設計した。CEJC を同じ条件で検索し，直接比較・対照できるようにすることを意図したものである。

表 2 『昭和話し言葉コーパス』に付与されたメタデータの一覧

会話	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
独話	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

3.5『昭和話し言葉コーパス』に収録された録音資料群の内訳

以下では，SSC に収録された録音資料の内訳を示す。録音資料の一覧については，稿末の付録も参照していただきたい。

3.5.1 データサイズと発話者の属性

まず, SSC のデータサイズを表 3 に示す。「総語数」は形態素解析を実施した結果のうち「品詞」に「記号」「補助記号」「空白」を含むものを除外した語数，「話者数」は延べ人数を表す [4]。

次に, SSC に含まれる録音資料の発話者の年齢分布を示す。図 6 は会話データ, 図 7 は独話データの分布である。縦軸は延べ人数，横軸は年齢を表す。会話データの発話者は，15 ～ 19 歳から 50 ～ 54 歳を中心に，全世代にわたって幅広く分布している。正確な年齢が判明しなかった発話者（「不明」）が全体の約 3 分の 1 を占めるが，この大半は「壮年層（20 代から 50 代）」と推測された。全体的には男性の方が女性よりも多い傾向にある。一方，独話データの話者のうち女性は 1 名だけであり，残りは全て男性であった。会話データに比べると，年齢層は高めであると言える。

表 3 SSC のデータサイズ

種別	ファイル数	総時間数	話者数（男性 / 女性）	総語数（男性 / 女性）
会話データ	73	約 27 時間	342 人（190 人 /150 人）	344,547 語（217,462 語 /126,976 語）
独話データ	50	約 17 時間	50 人（49 人 /1 人）	180,212 語（177,596 語 /2,616 語）
合計	123	約 44 時間	392 人（239 人 /151 人）	524,759 語（395,058 語 /129,592 語）

[4] 会話データには，性別不明の話者 2 名の発話（109 語）が含まれている。

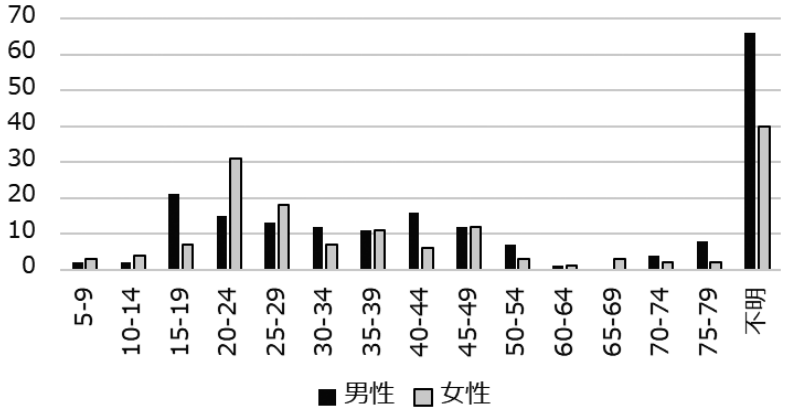

図 6　発話者の年齢分布（会話データ）

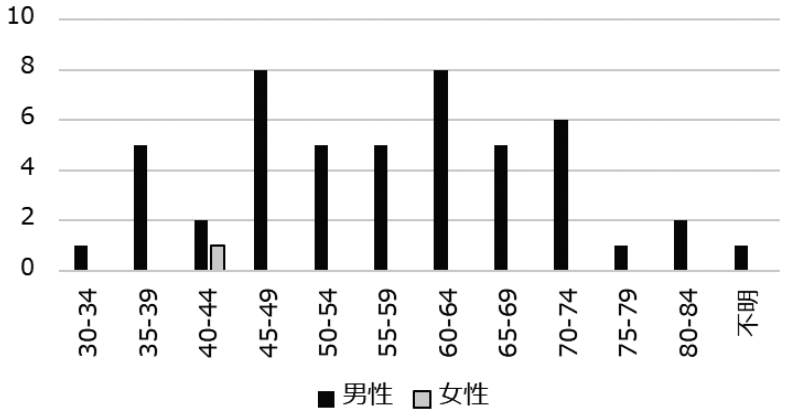

図 7　発話者の年齢分布（独話データ）

3.5.2 会話データのサイズ

SSC の会話データには, 1952 年から 1969 年にかけて録音された 73 会話（約 27 時間, 延べ 342 人, 異なりで 240 人の話者）を収録している。会話データに含まれるファイル数，収録時間，話者の性別ごとの総語数を収録年ごとに示すと，表 4 のようになる。

表 4　SSC 会話データのサイズ

収録年	ファイル数	収録時間	総語数 男性	総語数 女性
1952 年	21	12.0 時間	80,600 語	63,264 語
1953 年	3	1.7 時間	8,426 語	6,986 語
1955 年	3	1.6 時間	23,013 語	3,158 語
1957 年	33	9.0 時間	80,231 語	48,714 語
1958 年	2	0.6 時間	4,538 語	1,529 語
1960 年	6	1.6 時間	16,435 語	842 語
1969 年	5	0.5 時間	4,219 語	2,483 語

表 4 で収録語数が顕著に多くなっているのは，『談話語の実態』のための録音作業を開始した 1952 年，および『話しことばの文型（1)』のための録音作業を開始した 1957 年である。その他の年にも散発的に録音された資料があるが，1961 年以降は録音資料がほとんど存在していない。

このうち，確認できる中で最も古いものは，1952 年 3 月に作成された「一研雑談」という録音資料である。「第 1 研究室」に録音機が導入されたのを機に，当時そこに在籍していた研究員 4 人が試験的に座談を録音してみたものであろう。次いで，1952 年 6 月に「三鷹分室」という録音資料が作成されている。当時の国立国語研究所が分室として使用していた旧山本有三邸（現・三鷹市山本有三記念館）で録音されたものであり，これも言わば「身内」の録音である。1952 年 7 月以降は，図 1 に示したように，さまざまな場所に出かけて行って録音を実施している。録音作業の試行と本番，という当時の手順をうかがい知ることができる。

発話者の男女比では，男性の方が人数が多く，発話している語数も多い。発話者数と総語数で平均を取ると，男性で 1,144.5 語 / 人，女性で 846.5 語 / 人となっている。

3.5.3 独話データのサイズ

SSC の独話データには，1955 年から 1974 年にかけて録音された 50 講演（約 17 時間，延べ 50 人，異なりで 33 人の話者）を収録している。独話データに含まれるファイル数，収録時間，話者の性別ごとの総語数を収録年ごとに示すと，表 5 のようになる。

収録語数が顕著に多くなっているのは，「国立国語研究所創立 10 周年記念講演会・祝賀式」（1959 年），「全国国語科指導主事研修講座」（1957 年），「国立国語研究所創立 20 周年記念講演会・祝賀式」（1969 年），「国立国語研究所新庁舎開き記念講演会・式典」（1955 年）などがあった年で，ここに録音資料が集中している。一方，録音資料が極端に少ない年や，録音資料が存在しない年もある。特に 1960 年代の前半は，録音資料が存在しない。

発話者の男女比は，前述のように男性が 49 人（異なりで 32 人），女性が 1 人であり，大きく偏っている。講演や司会，所長挨拶など，総語数の 90% 以上を国立国語研究所員が担当していることから，当時の研究所員の男女比が影響したものと考えられる。

表 5　SSC 独話データのサイズ

収録年	ファイル数	収録時間	総語数 男性	総語数 女性
1955 年	13	2.4 時間	22,397 語	
1957 年	7	3.9 時間	39,364 語	2,616 語
1959 年	14	4.4 時間	45,024 語	
1965 年	1	0.03 時間	325 語	
1966 年	1	0.3 時間	3,303 語	
1967 年	2	0.2 時間	2,261 語	
1968 年	1	0.9 時間	9,356 語	
1969 年	4	3.3 時間	37,800 語	
1972 年	1	0.9 時間	11,261 語	
1974 年	6	0.6 時間	6,505 語	

これら独話の録音資料は，1952 年に始まった会話の録音資料に遅れて，その比較資料として収集が始まったものであり，報告書等を読む限り，会話の録音作業で見られたような緻密なサンプリングが施されてはいないようである。これらの独話の資料は，当時の研究（特に 1963 年の報告書『話しことばの文型（2）―独話資料による研究―』）で使われていたものもあるが，総体的には，「たまたま録音する機会があった講義や講演，挨拶，祝辞など」という位置付けになると考えられる。

3.6『昭和話し言葉コーパス』の構成に関する評価

最後に，SSC の会話データ・独話データの構成に対して，量的・質的な側面から評価を加えておく。特にメタデータの中に付与された「形式」および話し手の性別に着目し，その量的な分布を CEJC・CSJ と比較することで，通時音声コーパスとしての連結可能性について論じる。

はじめに，SSC の量的な側面を考えてみよう。通時音声コーパスの先行例として先に挙げた DCPSE には，1960 年代後半から 1980 年代にかけての話し言葉が 40 万語，1990 年代前半の話し言葉が 40 万語，合計 80 万語が収録されている（うち独話は 13 万語）。これに対して，SSC は約 34 万語の会話，約 18 万語の独話を収録しており，さらに CEJC・CSJ と連結して利用することを考えれば，DCPSE を上回るデータ量が確保されていると言える。

次に，SSC の質的な側面について考えてみたい。まず，SSC の会話データでは，全ての会話が「雑談」「用談相談」「会議会合」という三つの「形式」に分類されている。これは CEJC の設計に準じたものであり，「雑談」は「会話の目的や話題などがあらかじめ定められていない会話」，「用談･相談」は「会話の目的はある程度決まっているが時間や場所などは定められていない会話」，「会議・会合」は「時間や場所などが定められている会話」をそれぞれ指す（小磯他 2016）。SSC と CEJC（『日本語日常会話コーパス モニター公開版（データバージョン 2021.03）』）について，形式および発話者の性別ごとに総語数を求めて示すと，図 8 のようになる。

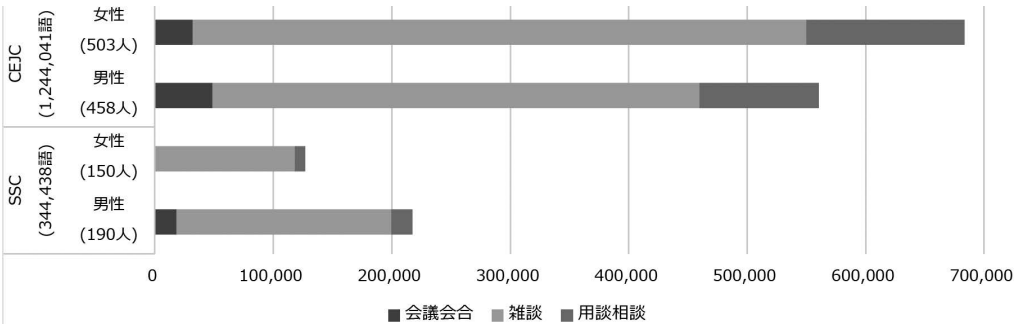

図 8　会話データの形式・性別ごとの総語数（SSC 会話・CEJC）

SSC と CEJC の両方で，「雑談」が大半を占めていることが分かる[5]。「用談相談」の割合が

[5] CEJC における雑談の比率は，CEJC の設計時に実施した「会話行動調査」（現代日本人が日常どのような

CEJC の側に多い, SSC には女性の「会議会合」がないなど, 比率の上で若干の相違はあるものの, 形式の分布は似通っている。比較対象とするデータごとに調整頻度を適切に算出すれば，SSC・CEJC を連結して比較することは問題ないと考えてよい[6]。

次に，SSC の独話データは「講演」「司会」「挨拶・祝辞」という三つの形式に分類されている。「講演」は創立記念講演会などにおける学術的な講演，「司会」は講演会などの司会進行を務める司会の発話，そして「挨拶・祝辞」は記念式典などにおける所長や来賓の挨拶，祝辞などが該当する。一方，CSJ では「学会講演」「模擬講演」と呼ばれる 2 種類の独話音声が中心となっており[7]，SSC の形式はこれに準拠しているわけではない。両者の対応を考えると，SSC の「講演」と CSJ の「学会講演」が録音資料の性質として対応するものの，SSC の「司会」「挨拶・祝辞」に対応する形式は，CSJ 側に存在しない。逆に，CSJ の「模擬講演」に対応する形式は，SSC 側に存在しない。

さらに，SSC の「講演」と，CSJ の「学会講演」のサイズを比較すると，表 6 のようになる。下段の数値は，「(男性 / 女性)」の内訳を表す。前述のように，SSC「講演」における女性話者の割合は CSJ「学会講演」と比べて圧倒的に少ない。すなわち，両者はともに学術的な講演である点で共通するものの，話し手の年齢分布という点においては大きく異なっていることになる。

表 6　SSC「講演」と CSJ「学会講演」の比較

	話者数	収録時間	総語数
SSC 講演	50 人 （49 人 /1 人）	11.2 時間 （11 時間 /0.2 時間）	12.2 万語 （11.9 万語 /0.3 万語）
CSJ 学会講演	987 人 （814 人 /173 人）	275.0 時間 （224.2 時間 /50.8 時間）	329.9 万語 （270.8 万語 /59.1 万語）

SSC のように，過去の録音資料を集めてコーパス化する際，録音資料を新規に獲得することができない場合には，このようなデータ構成上の不均衡が生じることはやむを得ない。設計が異なる複数のコーパスを連結して比較する時には，そこに含まれるデータ構成の実態をきちんと理解した上で，分析対象とする集合（形式，性別，年齢など）ごとに調整頻度を算出して比較するなどの対処が重要である。

このような不均衡を解消するためには, データの拡充が必要となるだろう。例えば, SSC にあって CSJ にない「挨拶・祝辞」を新規に録音して話し言葉コーパスを構築すれば，SSC との対応が取れることになる。逆に，過去の録音資料をさらに発見（発掘）し，SSC に追加していくことができれば，より幅広いレジスター（言語使用域）を備えた通時音声コーパスを実現すること

種類の会話をどの程度行っているかの調査）において雑談が全体の 6 割を占めるという結果が得られたことから設定したものである（小磯他 2016）。

[6] 調整頻度の算出に必要な語数表は，SSC のウェブサイトから入手できる。

[7] CSJ には,「独話」(学会講演，模擬講演，その他),「対話」(学会講演インタビュー，模擬講演インタビュー，課題指向対話，自由対話),「朗読」(朗読，再朗読）という三つのタイプとその下位分類が設定されている。このうち，学会講演と模擬講演で全時間数の 90% を占める。学会講演は「種々の学会における研究発表のライブ録音」，模擬講演は「日常的話題についての講演」である（国立国語研究所 2006）。その他のタイプについては，ここでは触れないこととする。

ができる。これらについては，将来的な課題としておく。

3.7『昭和話し言葉コーパス』の一般公開

2016 年度に構築を開始した SSC は，5 年間の作業期間を経て，2021 年度末に完成させる計画を立てた。2019 年 3 月には，それまでに作業を終えた独話データを『昭和話し言葉コーパス モニター公開版』としてまとめ，オンライン検索アプリケーション「中納言」で公開した。これと同時に，音声データ・転記テキスト・時間情報付き転記テキスト（TextGrid 形式）・メタデータ・全文検索システム「ひまわり」による検索環境を同梱した DVD（『昭和話し言葉コーパス モニター公開版 DVD』）を作成し，研究利用に限定して希望者に配布した。2021 年 3 月には，コーパス全体が計画通りに完成し，「中納言」で一般に公開した。これにより，形態論情報を利用して約 60 年前の音声を検索し，その前後の範囲の音声を聴取できる環境が整った。

2021 年度中には，SSC の全データ，すなわち，音声データ・転記テキスト・時間情報付き転記テキスト（TextGrid 形式）・形態論情報データ・メタデータ・全文検索システム「ひまわり」による検索環境などをまとめて，国立国語研究所のサーバー上から配信を開始する予定である。

4.『昭和話し言葉コーパス』の予備的分析

本節では，SSC の予備的な分析結果の例を示す。以下，(1) 句末・発話末の急激な上昇イントネーション，(2) 一人称代名詞「私」の発音のバリエーション，(3) 助動詞「ます」と「まする」のゆれ，(4) 助動詞「ます」「た」と助動詞「です」の連接，という四つの分析例を示す。

4.1 句末・発話末の急激な上昇イントネーション

まず，「句末・発話末の急激な上昇イントネーション」について見てみよう。SSC の会話データには，特に女性の発話の中で，図 9 のような急激な上昇イントネーションが現れることがある。

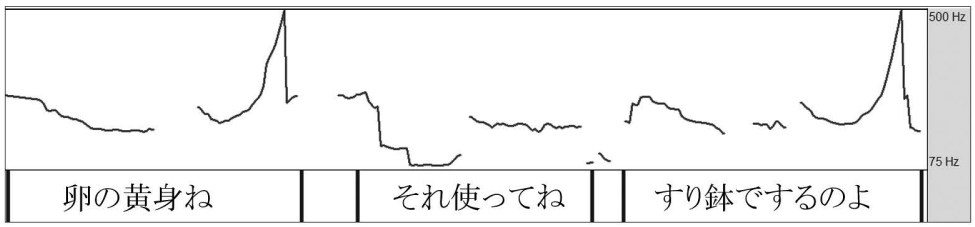

図 9　句末・発話末の急激な上昇イントネーション

これは，1957 年 2 月に録音された「三人の女性」という資料に出現した，「(そしてー みりんとね，) 卵の黄身ね，それ使ってね，すり鉢でするのよ」という 20 代半ばの女性の発話である。図のピッチ曲線を見ると，「黄身ね」の「ね」，「するのよ」の「よ」，すなわち，一部の句末・発話末において，ピッチが急激に上昇していることが分かる。この上昇イントネーションは，無論，聞き手に対する質問や疑問を表すものではない。

このような句末・発話末における急激な上昇イントネーションは，小津安二郎作品をはじめとする 1950 年代の邦画において，特に女性の台詞の中でよく観察されるように思われる。このようなイントネーションの型が SSC の中でも散見されるという事実からすると，当時の女性の間で多用されていたイントネーションであったことが推察される。

一方，現代の話し言葉でこのような型のイントネーションがどれくらい存在するかは，CSJ・CEJC を詳しく調べてみないと分からない。70 歳代以上の高齢の女性が少し気取って発話しているような状況では，現代でも観察されるように思われるが，少なくとも 20 歳代の若年層の女性の中では使われていないように感じられる。過去 60 年の間に，このようなイントネーションの型が徐々に廃れていったということであろう。

1950 年代から 1970 年代における急激な上昇イントネーションの実態を詳細に分析するためには，SSC に対して新たに韻律情報（特に句末の BPM）をアノテーションし，その数量的な分布を把握する必要がある。この点は将来的な課題としたい。

4.2 一人称代名詞「私」の発音のバリエーション

「私」という一人称代名詞には，「ワタシ」「アタシ」「ワタクシ」「アタクシ」という発音上のバリエーションが存在する。これらの発音の分布は，過去から現在にかけて変化してきているだろうか。SSC の会話データと CEJC（『日本語日常会話コーパス モニター公開版（データバージョン 2021.03）』）を比較して，この点を調べてみよう。

SSC 会話および CEJC から代名詞「私」を検索し，その語形を集計した。出現頻度を男女それぞれ 10 万語あたりの出現数に正規化してグラフ化したものを，図 10，11 に示す。

SSC 会話では一定数が見られていた「ワタクシ」「アタクシ」が，CEJC では男女ともほぼゼロになっている。SSC 会話・CEJC はいずれも日常会話を収録したデータであるが，1950 年代の日常会話では多用されていた「ワタクシ」「アタクシ」という丁寧な形が，現代ではほぼ使われなくなっていることが分かる。

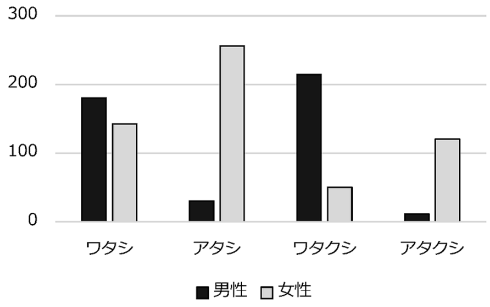

図 10 「私」の発音の分布（SSC 会話）

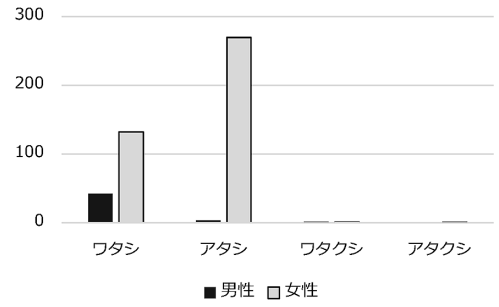

図 11 「私」の発音の分布（CEJC）

また，女性の「アタシ」「ワタシ」の使用数は SSC 会話と CEJC の間でほぼ変わっていないが，男性は「ワタシ」「アタシ」の数は大きく減少している。この点については，「僕」「俺」「うち」

などの一人称代名詞もあわせて検索し，時代ごとに選好される人称代名詞の分布を総合的に考える必要があるだろう。

4.3 助動詞「ます」と「まする」のゆれ

次に，SSC の独話データの中で特徴的に観察される助動詞「ます」と「まする」のゆれについて見る。以下，(1) は 1959 年「国立国語研究所 10 周年祝賀式」における山本有三氏（当時 72 歳）の祝辞，(2) は林大氏（当時 42 歳）の講演からの例である。例文末尾にファイル ID を記す。

(1) a. 非常に予算の窮屈な あー 時代でありまするから （M58_06_SC）
b. 国語の問題 (W ッタ | ったら) 難しいんでありますから （M58_06_SC）
(2) a. 新しい字引が二十万語を収載すると書いてありまするけれども，（M54_12_LT）
b. 非常に え あの 時代的な差もありますけれども，（M54_12_LT）

これらは，同一話者による同じ講演の中で観察された例の組である。いずれも同じ接続助詞（から，けれども）の直前で「ます」と「まする」の両方が用いられている。すなわち，「ます」「まする」という二つの文法形式の間でゆれが生じていることになる。

これに対して，CSJ・CEJC で「まする」の例を検索すると，CSJ に 3 例「まする」の出現が認められたものの，これらは古典の例文を読み上げているものであり，自然発話としての「まする」の例は一切見つからなかった。一方，「国会会議録検索システム」を検索すると，1990 年代までは「まする」の例が一定数見つかるほか，現代でもなお少数の例が見つかる。(3) の a. は 1924 年生まれの話者（村山富市），b. は 1962 年生まれの話者（文部科学官僚）による発話である。

(3) a. 財源を十分積み立てることが望ましい姿でありまするが，（1994 年 12 月）
b. 学校の働き方改革のための取組状況調査によりまると，（2021 年 4 月）

このことから，少なくとも政治的な発言の場面においては，現在でもなお「まする」が少数残存しているものと考えられる。

4.4 動詞のル形・タ形・マス形と助動詞「です」の連接

SSC の会話データで，動詞のル形・タ形・マス形に助動詞「です」が後接するパタンが観察される。「おいしいです」「おいしかったです」のような，形容詞のル形・タ形に「です」が後接する場合は現在では広く許容されているが，以下のように動詞のル形・タ形・マス形に「です」が後接する場合はどうだろうか（さらに「ましたです」「でしたです」というパタンもある）。

(4) a. にかわへくっ付いた : 竹のクシが外れるですね。 （C60_02_CT）
b. 商売は何をなさったですか。 （C52_17_CT）
c. たいがい六時半ちょっとすぎにここへ来ますです。 （C52_08_CT）

d. 怖いこともありましたですよ。 （C52_08_CT）
e. 遊覧はちっとも入りませんでしたですけどね。 （C52_08_CT）

このような例は，現代日本語文法の観点で内省すると，非文として判断されるだろう。ところが，現代の話し言葉コーパスの中にもこのような例が見つかることがある。（5）は CSJ，（6）は CEJC で観察された例である。

（5） a. 絵を 書いたです （S01M0764）
b. 非常に 感じることも ありますですね （S02F0180）
c. 静かで いいところでしたですね （S03M0174）
（6） a. 必ず カラオケ 行くですよ （T015_008）
b. ここに 丸を 付けてくる 人が いるです （T007_009）
c. もっと 大変だったですね （K002_003）

前川（2007）は，「このあたりは雨が降ったです」のような「動詞タ形＋です」の用例が存在することを指摘した上で，「これらの用例が用いられたであろう文脈を想像してみる。すると私などは（中略）非文と断定しにくく感じられてくる。合理化の契機が与えられれば，むしろ適格文に思えてくる（p.21）」と述べている。上記の「ますです」「ましたです」などの例に対しても，前川（2007）の主張は通用するように思われる。文法的な規範意識から逸脱するように見えるこれらの例が，話し言葉の中にたまたま出現した「誤り」なのか，実は体系的に存在している文法的な連接なのか，この点は内省で即断することなく，過去から現在に至る広範な録音資料を収集して，注意深く吟味する必要があるだろう。

5. おわりに

本稿では，1950 年代から 1970 年代にかけて作成された日常会話・独話の録音資料を再編して構築した『昭和話し言葉コーパス』（SSC）について，収録されている録音資料群の出自と内訳，コーパス構築の過程とアノテーション，予備的な分析の結果などについて論じた。

1950 年代に国立国語研究所で実施されていた話し言葉の定量的な研究は，世界的に見ても極めて早い時期に属する先駆的なものであり，現代のコーパス言語学的な視点から見ても高く評価できるものである。当時の録音資料を再編成して構築した『昭和話し言葉コーパス』を，現代の話し言葉コーパス（『日本語話し言葉コーパス』『日本語日常会話コーパス』など）と比較・対照することにより，過去 60 年間で生じた話し言葉の経年変化を，アクセント・イントネーション・語彙・文法など，さまざまな観点から明らかにすることができるだろう。本稿ではイントネーション，語形のバリエーション，文法形式のゆれなどを予備的分析として取り上げたが，話速やポーズ長，句末境界音調（BPM）の実態など，音声的な諸特徴についてより深く分析することを，今後の課題の一つとしておきたい。

さらに将来的な課題として，二つの点を挙げておく。一点目は，録音資料の追加・拡充である。

3.6 節でも述べたように，現状の『昭和話し言葉コーパス』と『日本語話し言葉コーパス』『日本語日常会話コーパス』は，レジスター（言語使用域）が必ずしも一致しておらず，片方にある形式がもう片方にはない，という不均衡が生じている。過去の録音資料にあった形式の話し言葉を新たに録音してコーパス化したり，逆に，より多くの古い録音資料を発見（発掘）して『昭和話し言葉コーパス』に追加したりすることで，より幅広いレジスターを備えた「通時音声コーパス」を実現することができるだろう（丸山 2015）。特に，1970 年代から 1980 年代にかけて国立国語研究所で作成されていた録音資料群を『昭和話し言葉コーパス』に追加することは，技術的には可能であると思われる。

二点目は，異なる設計によって構築された録音資料を統一的にカタログ化するためのメタデータの設計である。各録音資料を体系的に整理し，それらがどのような性格を持つ言語資料であるのかを正確に知るためには，適切なメタデータを設計しておく必要があるだろう。

コーパスの多様化が進む中，今後「通時音声コーパス」という新たなコーパス開発の方向性が拡大し，コーパスに基づく話し言葉の経年変化の研究が進展することを期待したい。

参照文献

相澤正夫・金澤裕之編（2016）『SP 盤演説レコードがひらく日本語研究』東京：笠間書院.

臼田泰如・川端良子・西川賢哉・石本祐一・小磯花絵（2018）「『日本語日常会話コーパス』における転記の基準と作成手法」『国立国語研究所論集』15: 177–193.

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: 85–106.

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 回年次大会発表論文集』775–778.

国立国語研究所（1955）『談話語の実態』東京：秀英出版.

国立国語研究所（1960）『話しことばの文型（1）―対話資料による研究―』東京：秀英出版.

国立国語研究所（1963）『話しことばの文型（2）―独話資料による研究―』東京：秀英出版.

国立国語研究所（1969）『話しことば研究室録音資料一覧』国立国語研究所話しことば研究室.

国立国語研究所（1972a）『動詞の意味・用法の記述的研究』東京：秀英出版.

国立国語研究所（1972b）『形容詞の意味・用法の記述的研究』東京：秀英出版.

国立国語研究所（1990）『話しことば 文脈付き用語索引（2）『談話語の実態』データ『話しことばの文型』データ『速記叢書講談演説集』データ 解説書』国立国語研究所.

国立国語研究所（2006）『日本語話し言葉コーパスの構築法』国立国語研究所.

前川喜久雄（2007）「コーパス日本語学の可能性―大規模均衡コーパスがもたらすもの―」『日本語科学』22: 13–28.

丸山岳彦（2014）「コーパス言語学・語用論の観点から見た日本語複文研究の動向と課題」『日本語複文構文の研究』385–398. 東京：ひつじ書房.

丸山岳彦（2015）「「通時音声コーパス」は可能か」『第 8 回 コーパス日本語学ワークショップ 予稿集』29–36.

丸山岳彦（2016）「『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』282: 39–55.

丸山岳彦（2019）「「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』402–412. 国立国語研究所.

丸山岳彦（2020）「『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 回年次大会予稿集』629–632. 言語処理学会.

Maruyama, Takehiko (2020) On the possibility of a diachronic speech corpus of Japanese. In: Andrej Bekeš and Irena Srdanović (ed.), *Japanese language from empirical perspective: Corpus-based studies and studies on discourse*. 219–234.

Ljubljana: Znanstvena založba FF.
丸山岳彦・高梨克也・内元清貴（2006）「第 5 章 節単位情報」『日本語話し言葉コーパスの構築法』255–322. 国立国語研究所.
丸山岳彦・西川賢哉・田嶋明日香・小磯花絵（2021）「『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」『言語処理学会第 27 回年次大会予稿集』86–90.
南不二男（1974）『現代日本語の構造』東京：大修館書店.
南不二男（1993）『現代日本文法の輪郭』東京：大修館書店.
宮島達夫（2007）「語彙調査からコーパスへ」『日本語科学』22: 29–46.

付記

『昭和話し言葉コーパス』の設計と構築に携わったメンバーは，以下の通りである。
伊藤優介，臼田泰如，大川惠莉，小野瀬敦也，河本はるか，菊池千尋，小磯花絵，小西光，近藤明日子，十河則子，田嶋明日香，土屋菜穂子，中神裕美子，中村壮範，西川賢哉，藤村寛子，松下晶子，丸山岳彦，森本桂子，山縣智子，山口昌也，劉双戍，渡邊友香（50 音順）

関連 Web サイト

『昭和話し言葉コーパス』https://www2.ninjal.ac.jp/conversation/showaCorpus/
「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」https://www2.ninjal.ac.jp/conversation/
コーパス検索アプリケーション「中納言」https://chunagon.ninjal.ac.jp/ssc/

付録：『昭和話し言葉コーパス』に収録された録音資料一覧

以下，『昭和話し言葉コーパス』に収録された録音資料の一覧を示す。会話データは収録年度ごとに資料名を「 」で囲って示す。独話データは収録されたイベントごとに資料名を「 」で囲み，発話者の氏名とともに示す。( ) 内の数値は収録時間（分数）である。

〈会話データ〉

- 1951 年度：「一研雑談」(28)
- 1952 年度：「三鷹分室」(33),「Y 理髪店」(36),「N 家雑談」(36),「三鷹学生」(68),「接客用語について」(33),「扇子屋」(33),「絵画館のおばさん」(35),「S 女子大生」(33),「I 家雑談」(23),「U 夫妻」(36),「S 女子大事務室」(33),「魚屋小僧」(69),「女性雑談」(33),「トタン屋」(25),「A 美髪店」(33),「ジイサン・バアサン」(69),「友の会」(33),「U 家雑談」(33)
- 1953 年度：「男女学生座談」(66),「魚屋」(35)
- 1954 年度：「K 高校生」(34)
- 1955 年度：「組合団交」(63)
- 1956 年度：「3 人の女性」(67),「劇団員雑談」(85)
- 1957 年度：「面接録音調査」(35),「T 社応接室」(32),「タクシー苦情」(38),「歯科大学生」(33),「麻布主婦」(50),「K 教育委員会雑談」(32),「鎌倉主婦」(33),「研究室の電話」(38),「養老院」(36),「少年工員」(32),「下町家族」(32),「3 人の青年」(34)
- 1960 年度：「浅草噺」(97)
- 1969 年度：「その電気ごたつは安全ですか」(31)

〈独話データ〉

- 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）」(35)
  - ➢ 「祝辞（松村謙三，茅誠司，村上俊亮，土岐善麿，柳田国男）」(16)
- 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）(3)
  - ➢ 「所長挨拶（西尾実）」(10)
  - ➢ 「現代の敬語意識（柴田武）」(26)

- ➢「三つの語彙調査（林大）」(54)
- 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）」(13)
  - ➢「方言調査法（野元菊雄）」(44)
  - ➢「言語能力の発達（芦沢節）」(15)
  - ➢「助詞・助動詞（宮地裕）」(64)
  - ➢「文型（永野賢）」(33)
  - ➢「国語教育（輿水実）」(31)
  - ➢「新聞文章研究法（林四郎）」(34)
- 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）」(31)
  - ➢「祝辞（橋本龍伍，兼重寛九郎，関口隆克，時枝誠記，山本有三）」(37)
  - ➢「来賓挨拶（土岐善麿，安倍能成，片山哲）」(15)
- 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）」(8)
  - ➢「明治初期の書きことば（山田巌）」(61)
  - ➢「現代語の標準（林大）」(48)
  - ➢「話しことばの文法（大石初太郎）」(40)
  - ➢「これからの日本語（岩淵悦太郎）」(25)
- 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）」(2)
- 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）」(17)
- 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）」(11)
  - ➢「講演者紹介（岩淵悦太郎）」(2)
- 見坊氏退官記念講演（1968）：
  - ➢「見坊氏退官記念講演（見坊豪紀）」(52)
- 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）」(59)
  - ➢「語彙調査と基本語彙（林四郎）」(80)
  - ➢「形容詞の意味の特質（西尾寅弥）」(54)
- 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）」(5)
- 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）」(54)
- 国立国語研究所研究棟落成式典（1974）：
  - ➢「所長挨拶（岩淵悦太郎）」(9)
  - ➢「来賓挨拶（奥野誠亮）」(6)
  - ➢「祝辞（藤沢達夫，安達健二，平塚益徳，久松潜一）」(19)

Design and Construction of the Showa Speech Corpus

MARUYAMA Takehiko[a] KOISO Hanae[b] NISHIKAWA Ken'ya[b]

[a]Senshu University/Invited Professor, Spoken Language Division, Research Department, NINJAL
[b]Spoken Language Division, Research Department, NINJAL

Abstract
Construction of the "Showa Speech Corpus" (SSC) began in 2016, and was completed in March 2021 and made available to the public online through the corpus search application *Chunagon*. The SSC consists of a collection of recordings made from the 1950s to the 1970s by the National Institute for Japanese Language and Linguistics. Thus, it is a speech corpus made with modern technology, but with old recordings as its content. The SSC is innovative in that it can be used to explore the changes in spoken language over time (i.e., as a "diachronic speech corpus") by linking, comparing, and contrasting the SSC with modern spoken language corpora such as the Corpus of Spontaneous Japanese (CSJ) and the Corpus of Everyday Japanese Conversation (CEJC). In this paper, we describe the origins of the recorded materials stored in the SSC, the process of corpus construction and annotation, and the results of the preliminary analysis.

Keywords: Showa Speech Corpus (SSC), *Research in the Colloquial Japanese*, annotation, diachronic change of spoken language, diachronic speech corpus