Possibilities and Problems in Compiling a Diachronic Speech Corpus of Japanese

「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―

丸山 岳彦（専修大学 / 国立国語研究所）*

Possibilities and Problems in Compiling a Diachronic Speech Corpus of Japanese

Takehiko Maruyama (Senshu University / NINJAL)

要旨

本稿では、話し言葉の経年変化を知るための「通時音声コーパス」が持つ可能性と問題点について述べる。過去の録音資料を大量に収集し、別の時代の録音資料と比較することにより、話し言葉がどのように変化してきたかを実証的に明らかにすることができる。本稿では、「通時音声コーパス」が言語研究に果たす可能性について論じた上で、現在構築している『昭和話し言葉コーパス』を例として、過去の録音資料をコーパス化する際に生じる作業上の問題点について、具体例を交えながら論じる。さらに、『昭和話し言葉コーパス』に見られるいくつかの言語現象を取り上げ、過去の録音資料をコーパス化して分析することの意義について述べる。

1. はじめに

2000 年代に入って以降、国立国語研究所が中心となり日本語コーパスの整備が進められた結果、現在では多様な日本語コーパスが利用できるようになった。現代日本語の書き言葉コーパス（BCCWJ）、話し言葉コーパス（CSJ、CEJC）、歴史コーパス（CHJ）、ウェブコーパス（NWJC）、学習者コーパス（I-JAS）、方言コーパス（COJADS）など、多様な日本語コーパスがオンラインで利用できる環境が整い、その拡張も続けられている。また、日本語コーパスの充実とともに、コーパスに基づく日本語研究・日本語情報処理研究も質量ともに活況を呈しており、この動向は今後もしばらく続くものと考えられる。

この状況において考えるべき課題の一つは、今後のコーパス開発がどのような方向に向かうべきか、ということであろう。これまでになかったタイプの日本語コーパスを設計・開発し、その分析を進めることによって、新しい日本語研究の可能性を検討することが必要と思われる。

そこで本稿では、話し言葉の経年変化を知るための「通時音声コーパス」について論じる。過去の録音資料をコーパスとして整備し、時代ごとに比較・分析することにより、話し言葉が変化してきた過程を知るための言語資源とすることができる。現在、我々が構築を進めている『昭和話し言葉コーパス』を例として取り上げ、その作業上の問題点や、分析の事例を示すことで、「通時音声コーパス」の可能性について述べる。

* maruyama@isc.senshu-u.ac.jp

2. 「通時音声コーパス」の定義と事例

2.1 「通時音声コーパス」とは何か

ここでは、「通時コーパス」を「一定量を備えた言語資料を時代ごとに整備し、相互に比較・分析することで、言語の経年変化を実証的に明らかにすることを目的としたコーパス」と定義しておく。英語では、1991 年に公開された “Helsinki Corpus” (The Diachronic Part of the Helsinki Corpus of English Texts) [(1)]以降、多くの通時コーパスが作成・公開されており、さらに文学作品のコーパス化により、コーパス文体論（corpus stylistics）の研究へと発展している（Mahlberg (2015), McIntyre and Walker (2019) など）。日本でも、2013 年以降『日本語歴史コーパス（CHJ）』の整備・公開が段階的に進められており、2019 年 3 月の時点で、奈良時代から昭和時代までのテキスト、1,862 万語が中納言で検索できるようになっている。

さて、通時コーパスに収録される対象は、どの言語でも、通常、書き言葉である。これは書き言葉が記録・保存に適したメディアであることによる。日本語の場合、書き言葉の資料によっておよそ 1300 年ほどの歴史を遡ることができ、その間に日本語がどのように変化してきたかを探ることができる。一方、清水 (2014) によると、年代が特定できるもっとも古い日本語の録音資料は、1900 年にパリ万国博覧会で録音された音声である。そもそも録音機が発明されたのが 19 世紀半ばであり、それが広く普及したのは実質的に 20 世紀に入ってからであることを考えると、話し言葉の資料は、せいぜい過去 100 年程度の歴史しか遡れないことになる。

それでもなお、過去の録音資料を収集し、コーパス化して時代ごとに並べることによって、「一定量を備えた録音資料を時代ごとに整備し、相互に比較・分析することで、言語の経年変化を実証的に明らかにすることを目的としたコーパス」を構築することができるはずである。そのようなコーパスを、ここでは「通時音声コーパス」と呼ぶことにする。

2.2 先行研究

海外ではすでに「通時音声コーパス」の構築と分析を実践した研究事例がある。ロンドン大学の Survey of English Usage から公開されている “DCPSE” （The Diachronic Corpus of Present-Day Spoken English)[(2)]は、London-Lund Corpus に収められた 1960 年代後半から 1980 年代前半の録音資料と、ICE-GB に収められた 1990 年代前半の録音資料、それぞれ 40 万語分ずつを連結した通時音声コーパスである。Aarts et al. (2013) では、イギリス英語の話し言葉の経年変化について、DCPSE に基づいて実証的に分析した結果が示されている。

また、フランスのオルレアン大学で構築されている “ESLO”（Enquête sociolinguistique à Orléans)[(3)]は、1968 年から 1971 年にかけてオルレアンで録音された約 300 時間、450 万語分の録音資料群（ESLO1）と、2008 年に再度オルレアンで ESLO1 と対照可能な形で録音された約 400 時間、600 万語分の録音資料群（ESLO2）から構成される通時音声コーパスである。

さらに、イギリス北部の Tyneside における英語音声を収めた “DECTE” （Diachronic

(1) http://clu.uni.no/icame/manuals/HC/INDEX.HTM
(2) https://www.ucl.ac.uk/english-usage/projects/dcpse/
(3) http://eslo.huma-num.fr/

Electronic Corpus of Tyneside English)[4]は、1960 年代後半に Tyneside Linguistic Survey（TLS）で作成された約 40 時間分の録音資料、および 1994 年に Phonological Variation and Change in Contemporary Spoken English （PVC）で作成された 18 時間分の録音資料を合体させたものである[5]。

これらはいずれも、1960 年代後半に作成された録音資料と、比較的近年に作成された録音資料の組み合わせによって「通時音声コーパス」を構成しているという点で共通している。世界的な潮流として、過去の録音資料をいわば「発掘」し、現代の録音資料と組み合わせることによって「通時音声コーパス」を構築する動きが進んでいると見ることができるだろう。

2.3 日本語における「通時音声コーパス」の可能性

歴史的な録音資料を実際に聴取できる形で話し言葉コーパスが整備されていれば、当時の発音・アクセント・イントネーションだけでなく、発話速度や非流暢性など、従来は分析する手段がなかった言語現象まで、広く研究の射程に収めることができる。また、書き言葉（文献資料）には反映されない当時の言葉づかいを直接的に観察・収集できる可能性も高く、音声・音韻、語彙、語法・文法など、多くの研究にとって有用な研究資源になる可能性がある。問題は、「通時音声コーパス」に収録する録音資料をどのように集めるか、という点である。

日本における過去の録音資料のうち、一定量が入手できるものとして、『岡田コレクション』が挙げられる。これは、1915 年から 1946 年にかけて SP レコードに録音された、政治演説や講演などを中心とする約 18.5 時間分の録音資料群である。相澤・金澤 (2016) には、『岡田コレクション』に収録された録音資料をさまざまな角度から分析した論文が収められている。

『岡田コレクション』よりも後の時代における録音資料として、国立国語研究所内が 1950 年代から作成を開始した録音資料群を挙げることができる。1952 年に設置された「第 1 研究室（1955 年に「話しことば研究室」に改称）」では、1952 年以降、多種多様な場面・話者による日常会話を録音し、約 40 時間分の録音資料を作成して、1955 年には報告書『談話語の実態』として分析結果を公表している。言語研究を目的として、日常談話を体系的にサンプリング・収集して分析したこの研究は、世界的に見ても極めて早い試みだったと言える。コンピュータもない時代、たった 3 年の間に、調査の設計・録音・転記・分析・出版まで終えているのは、まさに驚異的と言わざるを得ない。

さて、筆者が 1950 年代当時の国語研で作成されていた録音資料を「発掘」し、話し言葉コーパスとして再編することを着想したのは、2012 年ごろのことであった。すでにデジタルデータ化されていた過去の録音資料群から、独話 25 時間分、会話 25 時間分の録音資料を収集し、『昭和話し言葉コーパス』と称して整備を進めて、一般に公開することを計画した。現在、2016 年に開始した科研費基盤研究 (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」、および国立国語研究所音声言語研究領域におけるコーパス開発の一環として、『昭和話し言葉コーパス』の構築を進めている。

将来的には、『岡田コレクション』や『昭和話し言葉コーパス』に続く新たな録音資料を収集

[4] https://research.ncl.ac.uk/necte/
[5] DECTE については、Allen et al. (2007) に詳細な記述がある。

してコーパス化し、それらを時代ごとに並べて連結することによって、20 世紀における日本語の話し言葉を総合的に見渡すための「通時音声コーパス」として拡張することを構想している。そのようなコーパスは、『日本語歴史コーパス』と同様、各時代の個別コーパスをモジュールとして組み込む「スーパーコーパス」として実現されることになるだろう。そのような構想の下、現時点でやるべきことは、対象となり得る録音資料の調査・選定と、それらをコーパス化する際に問題となる点をあらかじめ洗い出しておくことである。次節では、後者の例として、『昭和話し言葉コーパス』の構築過程で生じた問題について、具体例を交えながら述べる。

3. 『昭和話し言葉コーパス』の設計・構築と問題点

本節では、『昭和話し言葉コーパス』の設計および構築の過程を紹介する。さらに、実際の構築過程でどのような問題が生じているかについて、具体例を交えて述べる。

3.1 『昭和話し言葉コーパス』の設計と構築

『昭和話し言葉コーパス』は、前述のとおり、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を再編し、コーパス化しようとするものである。独話 25 時間、会話 25 時間、計 50 時間のデータを整備し、2020 年度以降に公開することを予定している。

実際の構築作業は、(1) 音声データの収集、(2) 転記テキストの作成、(3) 転記テキストと時間情報のアライメント、(4) 形態論情報の付与、という順序で進めており、これと並行してメタデータの設計・付与を行っている。

2018 年度末には、それまでに作業が完了した 17 時間分の独話データについて、モニター公開を開始した。1955 年から 1974 年にかけて録音された 50 講演（約 17 時間、異なりで 33 人の話者）の独話データ（音声ファイル、転記テキスト、TextGrid ファイル、全文検索システム「ひまわり」による検索環境を含む）が DVD に収録されており、稿末の URL から申し込めば、研究に利用していただくことが可能である。

独話データの例として、1959 年 3 月 7 日に行われた「国立国語研究所創立 10 周年記念」の「祝賀式」および「記念講演会」で録音された講演の一覧を表 1 に挙げる。

表 1 『昭和話し言葉コーパス』に収録された独話（講演）の例

祝賀式	記念講演会「現代語の発展のために」
西尾実「所長挨拶」31.4 分	西尾実「あいさつ」8.4 分
祝辞：橋本龍伍（文部大臣）3 分	山田巌「明治初期の書きことば」61.2 分
祝辞：兼重寛九郎（日本学術会議会長）2.5 分	林大「現代語の標準」48.3 分
祝辞：関口隆克（国立教育研究所長）3.6 分	大石初太郎「話しことばの文法」39.9 分
祝辞：時枝誠記（国語学会代表理事）13.9 分	岩淵悦太郎「これからの日本語」25.5 分
祝辞：山本有三（日本芸術院会員）13 分	
祝辞：土岐善麿（国語研評議員会長）4.8 分	
祝辞：安倍能成（文部大臣）7.9 分	
祝辞：片山哲（衆議院議員）2.3 分	

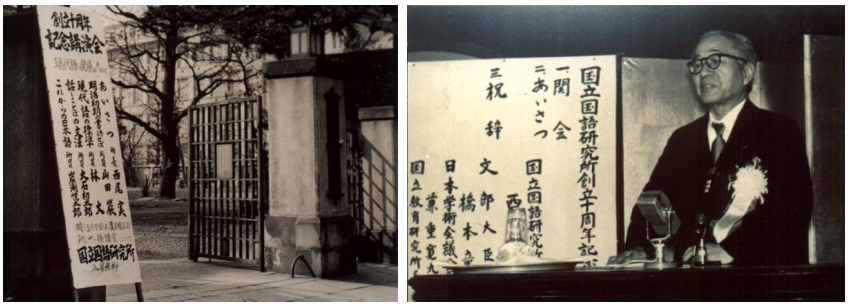

図 1　「国立国語研究所創立 10 周年記念祝賀式」の様子（1959 年 3 月 7 日、西尾実所長による挨拶）

3.2　会話データの設計と再現

1952 年から録音が開始された会話の録音資料は、そもそも言語調査のために設計・収集された資料群であるという点に特徴がある。1952 年『昭和 27 年度 国立国語研究所年報 4』には、以下のような記述や、図 2 のような録音資料の一覧が掲載されている。

> 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・市町村などの生活環境との切点（ママ）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.	話者名称	録音状態の可否	地区			場所					性				年齢				教養			相手				
			下町	山手	周辺	家庭	近隣	学校	職場	公共施設	男男	女女	男女	女男	若若	壮壮	若壮	壮若	専専	雑専	雑雑	1人	2-4人	5-7人	8人以上	未知既知
3	1夫婦談	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	友の会	可			×			×				×						×	×	×	× ×		×			×
2	女子学生	可			×				×			×			×					×	×			×		×

図 2　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）

図 2 にある分類情報（地区・場所・性・年齢・教養・相手）は、当時のサンプリングの実施において、全体のバランスを統制するための手掛かりとなった情報である。これらは、当時の設計を反映する情報として、『昭和話し言葉コーパス』のメタデータにもできる限り反映させることを予定している。

しかしながら、『昭和話し言葉コーパス』には、1955 年の『談話語の実態』刊行後に収集された会話も収録されているため、録音資料のすべてに図 2 の分類情報が付与されているわけでは

ない。また、図 2 に記載のある録音資料が手元に残っていない（音声データが見つからない）というケースもある。ここから、当時は担保されていたはずのサンプリングの設計・録音資料のバランスを、そのままの形で復元することは、難しいと言わざるを得ない。過去の言語資料を扱う場合、このようなデータの欠損・欠落が生じることは、不可避的である。

3.3　独話データの設計と再現

『昭和話し言葉コーパス』モニター公開データは、1955 年から 1974 年にかけて録音された 50 講演（約 17 時間、異なりで 33 人の話者）を収録している。収録された録音資料の総時間数を年代ごとに示すと、図 3 のようになる。縦軸は収録時間数、横軸は収録年を表す。

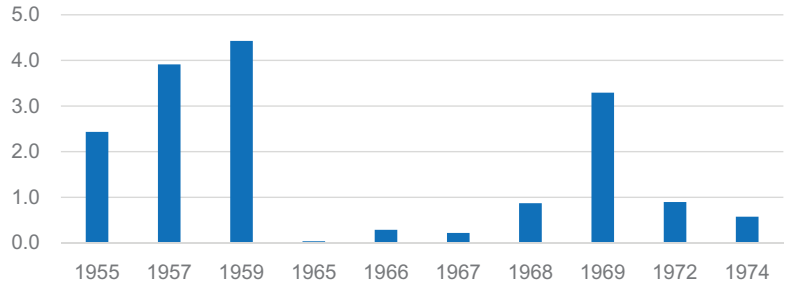

図 3　『昭和話し言葉コーパス』モニター公開データに収録された録音資料の総時間数

収録時間数が大きく盛り上がっているところは、「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、録音資料が存在していない。

そもそも独話の録音資料は、1952 年に始まった会話の録音資料に遅れて収集が始まったものであり、報告書等を読む限り、会話の収集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に見える。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)』）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、祝辞などが残っていたもの」ということになるだろう。また、異なりで 33 名の話者がいるが、このうち女性は、当時国語研所員であった芦沢節の 1 名だけであった。

なお、2004 年に公開された『日本語話し言葉コーパス』（CSJ）には主として「学会講演」と「模擬講演」という 2 種の独話が含まれるが、後者は性差や年齢差のバランスが意図的に統制されているのに対し、前者はバランスが取れていない（特に理系の学会発表には男子大学院生が多く登壇するため、女性のサンプルが少ない）。1950 年代以降の会話と独話の収録作業も、おそらく似たような状況だったのではないかと推測される。

3.4　音声の聞き取りと転記テキストの再現

話し言葉コーパスは、第一次資料としての音声そのものが重要であり、かつ、その正確な転記が付与されていることが前提である。『昭和話し言葉コーパス』では、録音資料の転記作業を新規に実施したが、音声が不明瞭で、発話者が何と言っているのか聞き取れないケースが多

く発生した。以下に例を挙げる。●が聞き取り困難な箇所である。

(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●●●●●

これらは、作業の過程で繰り返し聞くことによって、発話内容が判明する場合もある（上記はそのケースである）。しかしながら、音量が小さかったり、ノイズが混入したりして、発話内容が全く聞き取れない場合も多い。特に多人数会話では発話の重複が頻繁に生じるため、発話内容の聞き取りや発話者の割り当てが極めて困難になるケースが多く発生している。これらはそもそも聞き取れない以上、「聴取不能」として転記テキストに残しておくほかない。

なお、1950 年代当時の転記テキストを資料庫から探し出し、照合する作業も実施してみたが、当時の段階ですでに聞き取れていなかったらしく、実際の発話内容が大胆に省略された形で転記されている場合が散見された。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。当時も、聴取不能の問題には対処できていなかったことが分かる。

> 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。（p.17）

3.5 メタデータの再現

録音資料を分析する際、そのメタデータを参照することは必須である。その音声の録音日、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、録音資料にはできるだけ詳しいメタデータが付与されていることが望ましい。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、という点である。

このうち独話の資料については、当時の国語研の所員、あるいは関係者が祝辞を述べているケースが大半であるので、『国立国語研究所 年報』や当時の写真などを手掛かりにして、詳細な情報が判明している。『昭和話し言葉コーパス』モニター公開データには、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）などがメタデータとして提供されている。

問題は、会話の資料である。1952 年から開始された録音作業は、当時の「肩掛け録音機（デンスケ）」を担いで街頭に出かけ、市井の人々の会話を録音したものが大半である。『談話語の実態』『話しことばの文型 (1)(2)』といった報告書には図 2 のような一覧が掲載されているものの、録音が実施された当時に、録音時の詳細な情報が残されていたのか、発話者情報（フェイスシート）が取られていたのかすら、当初は全く分からなかった。

ところが、国語研の中央資料庫に保存されている当時の資料群を探索した結果、図 4 のようなメモ書きを発見することができた。上段は当時の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。

上段のメモ書きからは、録音日時と場所、録音に参加した発話者の苗字と出身地、録音時の

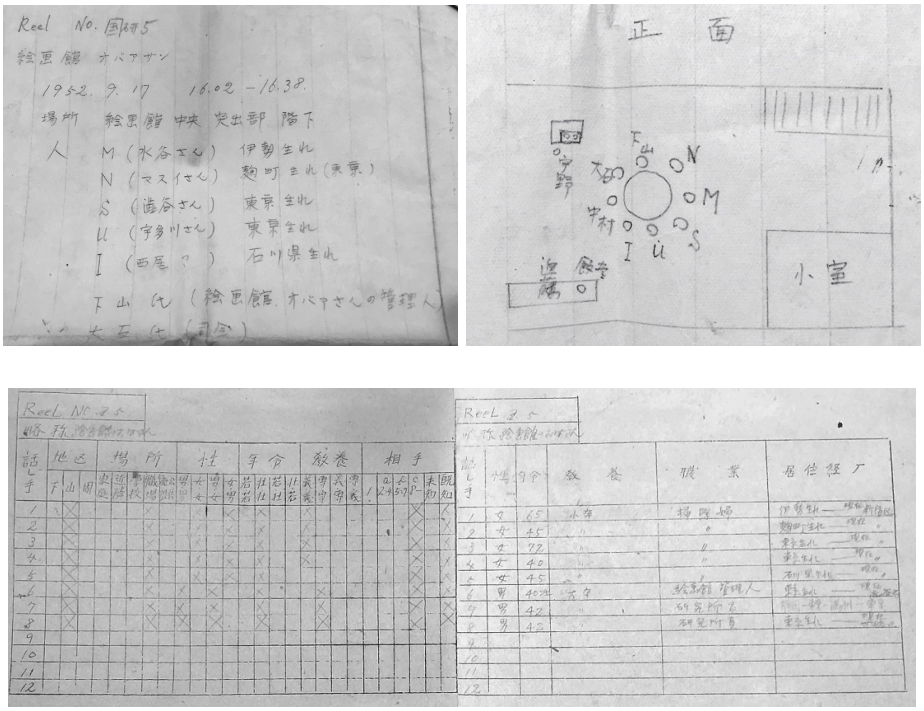

図 4　中央資料庫に保存されていた録音時のメモ（1952 年 9 月 17 日録音「絵画館のおばさん」）

状況を知ることができる。下段の左側は、おそらく、図 2 の一覧をまとめる際の原本となった資料であろう。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）が記載されており、フェイスシートに相当する。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が記録されていたことを裏付けるものだと言える(6)。

惜しまれるのは、このようなメタデータが残されている会話データが、一部に限られるということである。当時はすべての録音資料にこれらの情報が付与されていた可能性もあるが、どこかの時点で散逸してしまったのかもしれない。中には録音日しか情報のない録音資料もあり、この場合には現在推測できる範囲で話者情報などを付与するしかない。会話データの全体に対して、当時の詳細なメタデータを再現して付与することは、困難であると言わざるを得ない。

4.　『昭和話し言葉コーパス』で観察される言語現象の例

最後に、『昭和話し言葉コーパス』で観察された言語現象のうち、特に助動詞「です」「ます」の連接を中心に、興味深い文法形式の事例を挙げておく(7)。

4.1　「ますというと」

『昭和話し言葉コーパス』の独話データを聞いていると、「ますというと」という耳慣れない連接が見つかることがある。

(6) 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。記して感謝したい。
(7) 『昭和話し言葉コーパス』の分析事例については、丸山 (2015, 2016, 2018) も参照のこと。

(4) a. どういうことが考えられてきたかと申しますというと、おー、言語を、そういう流動の姿においてでなくて、（時枝誠記 1959 年）
b. そうして、建物のほうと言いますというと、10 年たっても、今もって、木造の貸家住まい。（山本有三 1959 年）
c. 総合雑誌の調査で得られた結果と照らし合わせてみますというと、なかなか、あ、よく選ばれていると思われます。（林大 1959 年）

(4) は、1959 年 3 月 6 日〜7 日に行われた国立国語研究所創立 10 周年祝賀式（時枝、山本）および記念講演会（林）で録音された資料からの例である。この場合の「というと」は引用節ではなく、文脈としては「申しますと、」「言いますと、」「照らし合わせてみますと、」のような条件節トと等価であると考えられる。このような「ますというと」という連接は、少なくとも筆者の語感からすると、非常に違和感のある言い方であるが、『昭和話し言葉コーパス』の独話データの中では、12 例が見つかった。

一方、『日本語話し言葉コーパス』でも、ごく少数ではあるが、類例が見つかった。「ますというと」という用例は、実は現代でも確認することができることになる。

(5) で (F う) それで 見ていきますと言うと (F あの)(F お) かなりの (D (? ん)) ことが 言える (A07M0652)

4.2 「ますのです」

また、『昭和話し言葉コーパス』の会話データの中に、「ますのです」という、これも筆者にとっては不自然と思われる連接が 7 例見つかった。

(6) a. はい。そこに おりますんです よ。（「ジイサン・バアサン」1952 年）
b. 大きな電話局と並んで できてますんです の。（「清水家雑談」1952 年）
c. お帰りになってからの調査も ございますんです けど。（「面接録音調査」1958 年）

ところがこれらの例も、『日本語話し言葉コーパス』で 18 例観察することができる。

(7) a. そいでごみにも色々 ありますんです けども (S11M1380)
b. これはもう既にものに書いて 発表してありますんです が (A06M0046)
c. 朝日が真正面に入る部屋に一人で おりますんです (S08F0486)

(7) の発話者の生年を見ると、a. が 1935-39 年、b. が 1915-19 年、c. が 1930-34 年であり、高齢層（収録時の 2000 年前後の時点で 60 代後半以上）に該当する。発話者の年齢が、「ますんです」の出現に影響している可能性はある。

さらに、国会会議録の中にも「ますのです」の例が複数見つかる。

(8) a. 私はこのように 思いますのです けれども、（山田耕三郎、1981 年）
b. 各省庁間とのまた打ち合わせ等も ございますのです が、（遠藤要、1987 年）
c. 長官の領域に 移っておりますのです から、（西村眞、1999 年）

筆者は、「ますのです」を「不自然な連接」と判断したが、少し前の時代の話し言葉では（あるいは現在もなお）許容されていたということかもしれない。一見不自然に見える連接が、どのような場面でどの程度使用されているのかについては、さらに広範囲のデータから事例を収集して検討する必要がある。

4.3 「ますです」「タ形＋です」

さらに、『昭和話し言葉コーパス』の会話データの中には、「ますです」や「た＋です」という連接が多く見られる。「楽しかったです」のような、形容詞のタ形に「です」が後接する場合は現在では広く許容されていると思われるが、問題は動詞・助動詞のタ形に「です」が後接する場合である。「ますです」が 30 例、形容詞以外の「たです」が 66 例、それぞれ見つかった。

(9) a. たいがい 6 時半ちょっとすぎにここへ来ますです。（「絵画館のおばさん」1952 年）
b. あれに 654 と書いてあったです。（「タクシー苦情」1957 年）
c. 商売は何をなさったですか。（「ジイサン・バアサン」1952 年）
d. 遊覧はちっとも入りませんでしたですけどね。（「絵画館のおばさん」1952 年）
e. 怖い面もありましたですよ、それでも、へえ。（「絵画館のおばさん」1952 年）

このような例も非文のように見えるが、実は、CSJ でも確認することができる。

(10) a. 表丸四の 右端に 示してありましたですが (A02M0094)
b. そうだ 絵を 書こうというので 絵 絵を 書いたです (S01M0764)
c. これも （D し） 静かで いいところでしたですね (S03M0174)
d. （F えー） 静かに 横たわっていたっすね (S02M0076)

特に (10d) のような例の存在を考えると、特に若年層における日常生活の会話の中ではかなり広範に使われている可能性がある（ただし、『日本語日常会話コーパス モニター公開版』の中で確認されたのは、「持ってきたですか (T013_006)」という 1 例だけであった）。

前川 (2007) は、このような「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」と述べている。上記の「ますというと」「ますのです」「ますです」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形式なのか、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データを収集して、注意深く吟味する必要があるだろう。通時音声コーパスを構築する意義は、このようなところにもある。

5. おわりに

以上、本稿では、「通時音声コーパス」が持つ可能性と問題点について論じてきた。さまざまな種類の日本語コーパスが整備・公開され、研究利用が進んでいる中、これから整備されるべき新たなタイプのコーパスとして、「通時音声コーパス」の可能性を提起した。

ただし、歴史コーパスが一般的に抱える問題と同様、過去の資料を「発掘」して利用する場合、さまざまな問題点や制約が生じることは不可避である。特に今回のように、過去に体系的に収集された研究資料を再利用しようとする場合、当時の設計を復元することができれば望ましいが、データの欠損・欠落が生じたり、聞き取りが困難であったり、当時のメタデータが入手できなかったりと、実際の現場では多くの困難が生じている。これらを現代の視点から補いつつ、できるだけ体系的にデータを整備していくことが求められるだろう。

また、『昭和話し言葉コーパス』の中で観察された、一見すると不自然に思われるような言語現象を指摘した。ところが、現代の話し言葉コーパスを検索すると、少数ながら、同じような例を見つけることができる。過去から現在に至るまでの話し言葉を通時コーパスとして整備し、各形式の動向を分析することによって、これまで気付かれてこなかった現象に光を当てることができる可能性がある。『昭和話し言葉コーパス』は、2020 年度以降に独話データ・会話データを含めて一般公開される予定である。ぜひ研究に利用していただきたい。

謝　辞

本研究は、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）、および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」によるものである。なお、現在までに『昭和話し言葉コーパス』の構築に携わってきたのは、以下のメンバーである（敬称略 50 音順）。
伊藤優介、大川惠莉、小野瀬敦也、河本はるか、菊池千尋、小磯花絵、田嶋明日香、土屋菜穂子、中神裕美子、西川賢哉、藤村寛子、松下晶子、丸山岳彦、山縣智子、山口昌也、劉双戌

文　献

Bas Aarts, Joanne Close, Geoffrey Leech, and Sean Wallis (Eds.) (2013). *The Verb Phrase in English: Investigating recent language change with corpora.*: Cambridge University Press.

Will Allen, Joan Beal, Karen Corrigan, Warren Maguire, and Hermann Moisl (2007). A Linguistic 'Time Capsule': The Newcastle Electronic Corpus of Tyneside English *Creating and Digitizing Language Corpora: Volume 2: Diachronic Databases.* Basingstoke: Palgrave Macmillan pp. 16–48.

Michaela Mahlberg (2015). *Corpus Stylistics and Dickens's Fiction.*: Routledge.

Dan McIntyre, and Brian Walker (2019). *Corpus Stylistics: Theory and Practice.*: Edinburgh University Press.

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』　笠間書院.

清水康行 (編) (2014).『百年前の日本語を聴く』　日本女子大学.

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, pp. 13–28.

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, pp. 29–36.　国立国語研究所.

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, pp. 39–55.

丸山岳彦 (2018).「「通時音声コーパス」から見る話し言葉の経年変化」　日本語文法学会第 19 回大会発表予稿集, pp. 65–72.　日本語文法学会.

関連 URL

「昭和話し言葉コーパス」　https://pj.ninjal.ac.jp/conversation/showaCorpus/