

国立国語研究所学術情報リポジトリ

会話コーパスの設計と構築

メタデータ	言語: Japanese 出版者: 研究社 公開日: 2026-07-24 キーワード: 作成者: 小磯, 花絵 メールアドレス: 所属: 国立国語研究所
URL	https://doi.org/10.15084/0002000721

This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0
International License.



第1章 | 会話コーパスの設計と構築

本章では、会話コーパスの設計と構築に焦点を当てる。1節では、均衡性を志向した話し言葉コーパスを設計するにあたり、これまでどのような方法が採用されてきたかを紹介する。2節では、幾つかの日本語会話コーパスを取り上げ、特に設計面に着目して紹介する。3節と4節では、『日本語日常会話コーパス』を取り上げ、構築法を紹介するとともに、その量的特徴を概観する。

□ 1 会話コーパスの設計

コーパスとは、実際に使用された書き言葉や話し言葉を対象に、研究目的に応じて定めた基準に従い、**代表性**やバランスを考慮して収集し、コンピュータで処理可能な形に電子化した上で、研究に必要な種々の情報を付与したデータベースである。

ここで、「代表的に、あるいはバランスよく収集する」方法として一般的に採用されるのは、研究目的に基づき設定された調査対象の**母集団**から、**ランダムサンプリング**（無作為抽出）などの方法により**標本**を抽出する手法である。ランダムサンプリングには様々な手法が存在する。例えば、母集団全体から無作為に抽出する単純無作為抽出法、母集団をいくつかの層に分け、各層の比率に応じてランダムに抽出する層別抽出法、母集団を集落（クラスター）に分けていくつかの集落を無作為に抽出し、それらの集落全体を対象とする集落抽出法などが挙げられる。

例えば、『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese, BCCWJ）に含まれる出版サブコーパスの書籍に関しては、層別抽出法が採用された。具体的には、日本十進分類法（NDC）の「0. 総記」から「9. 文学」および「n. 記録なし」の11分類毎に、総文字数を推定して比率を算出し、対象とする2001年から2005年までの5年間を加味して55の層を設けた上で、それぞれからランダムに資料を抽出するという方法である（丸山・柏野 2014）。

では、話し言葉の場合はどうであろうか。まず第1に、母集団の設定そのものが困難であるという問題に直面する。話し言葉においては、私たちが日常的にどのような種類の話し言葉をどの程度発話しているかを正確に把握すること自体が非常に難しい¹。また、話し言葉においては、どのような「層」を設けて収集設計に活かすべきかという点についても検討が求められる。話者の性別や年齢といった属性であれば、人口推計などの統計資料を活用することが可能であるが、会話の場面や話題、話し手・聞き手の関係性といった要素をどのように収集設計に反映させるべきか、更にどのような要素を考慮すべきかについても、慎重な検討が必要となる。

そこで本節では、**均衡性**を志向した話し言葉コーパスを設計するにあたり、これまでどのような方法が採用されてきたかについて、4つの話し言葉コーパス・調査を取り上げて具体的に紹介する。なお、均衡性という概念の定義は研究者によって異なるが(山崎・前川 2014)、本章においては「できるだけ多数の言語変種をカバーする」という意味で用いる。話し言葉コーパスでは、独話か会話か、話者の性差・年齢差・地域差、話者間の関係性、場面や話題、対面か非対面などの観点から、「積極的に多数の変種をカバーして、対象言語の全体像を把握」(前川 2013) することを目指して設計することを意味する。

1.1 Survey of English Usage

コーパス言語学の黎明期における代表的な試みとして、University College LondonにおいてRandolph Quirkらにより1959年から開始された**Survey of English Usage**を紹介する。このプロジェクトは、イギリス英語の文法を精密に記述することを目的とし、多様なレジスターにわたる書き言葉および話し言葉を、それぞれ50万語、計100万語収集するものであった。図1に、話し言葉の設計を示す。

話し言葉は対話と独話に大別される。対話は更に、プライベートな会話と、放送されたインタビューやパネルディスカッションといった公開討論に分けられ、プライベートな会話は、対面での会話と電話会話に細分されている。一方、独話に関しては、説教や政治演説など、原稿が用意された演説と、実況解説や議会討論でのスピーチなど原稿のない自発的な演説とに分けられている。このように、会話か独話か、私的か公的か、対面か非対面か、自発的か否か、といっ

1| 書き言葉も、例えば私的な手紙やメモ書きなどまで含めてその総体を把握することはできないという意味では同様の問題がある。

た複数の観点から話し言葉の層を設定し、各層からサンプルを選定することによって、多様な変種の話し言葉を収集しようとしていたことが分かる。

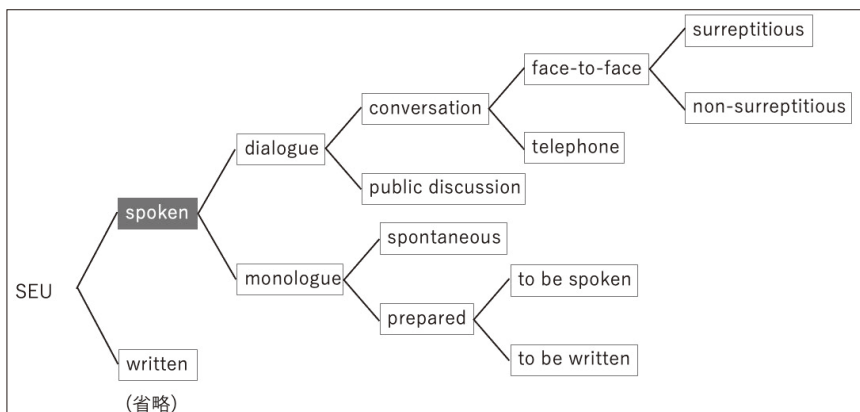


図1 Survey of English Usageにおける話し言葉パートの設計

なお、当時は紙のカードを用いて編纂作業が行われたが、その後、話し言葉の部分については、Lund UniversityのJan Svartvikらの主導のもと開始されたSurvey of Spoken Englishにおいて電子化が進められ、1990年に**London-Lund Corpus of Spoken English**として公開された (Svartvik 1990)。

1.2 国立国語研究所における話し言葉の調査研究

日本では、1952年から国立国語研究所において、共通語の話し言葉に関する調査研究が開始された。この研究は、談話語の実態を明らかにすることを目的とし、1年次には東京における日常の談話が録音され、2年次には比較の対象として、ニュース、ニュース解説、座談会、落語、講義、劇などが収録された (国立国語研究所 1955)。調査の主たる対象である日常の談話がどのように収集されたのかについて、報告書には、「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「同時に、話し手・聞き手の性・年齢・教養・相手 (の人数、未知既知)・地域などにもなるべく片寄りのすくな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」とある (国立国語研究所 1955: 2)。図2に、1952年に国立国語研究所で収録された日常談話の一覧を示す。

Reel No.	略 称	録音 状態 の否	地 区	場 所	性				年 齢			教 養			相 手				未既 知
			下 山 周 町 手 辺	家 近 学 職 庭 隣 校 場	公 共 施 設	男 女 男 女	男 女 男 女	若 壯 若 壯	若 壯 若 壯	専 専 専 専	専 専 専 専	1 2 5 8 人 人 人 人	4 7 以上						
3	I 夫 妻	可	×	×			×	×				×			×			×	
7	T家雑談	可	×	×			×	×	×		×	×			×			×	
67	N家座談	可	×	×			×	×	×		×	×			×			×	
86	トクン屋	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 1 座 談	可	×	×			×		×			×			×			×	
66	井 戸 端	可	×	×			×			×	×	×	×	×	×			×	
98	友 の 会	可	×	×			×			×		×	×	×	×			×	
2	女 子 生	可	×	×			×		×			×	×		×			×	
59	松 根 屋	可	×		×		×		×			×	×		×			×	
57	高 島 屋	可	×		×		×	×	×	×		×	×			×		×	
100	柳 橋 美 髪	可	×		×		×	×	×		×	×			×	×		×	
5	絵 画 館 おばさん	可	×		×		×	×	×		×	×	×	×	×			×	
25	床 屋	稍可	×		×		×			×	×	×	×	×	×			×	
51	一研雑談	可	×		×		×	×	×	×		×			×			×	
64	三 縣 女 工	稍可	×		×		×		×			×			×			×	
62	学 生 2 座 談	稍可	×	×			×	×	×	×		×			×			×	
104	三 美 容 室	稍可	×		×		×			×	×		×	×	×			×	
15	結婚申込	可	×		×		×	×	×	×	×	×	×	×	×			×	
87	職安男子	可	×		×		×			×	×		×	×	×			×	
88	職安女子	可	×		×		×	×	×	×			×	×	×			×	
79	無尽の会	可	×	×			×	×	×	×		×				×		×	
70	女 子 大 室 事 務	可	×		×		×	×	×	×			×	×		×		×	

図2 国立国語研究所で1952年に収録された日常談話の一覧(国立国語研究所 1955, p.3)

このように、話者の属性や打合せ、談話の場面の多様性といった観点から、日常談話の広がりをはできるだけ捉えるよう工夫されており、サンプルの収集においてもその多様性の確保が意識されていたことが分かる。なお、この調査も当時はカード形式で編纂されたが、日常談話のデータについては後年、『昭和話し言葉コーパス』(丸山ほか 2022)の一部として再編された。国立国語研究所のオンライン検索システム「中納言」で利用できるほか、同サイトから音声・転記・形態論情報・メタデータなどをダウンロードすることもできる。

1.3 British National Corpus

1990年代に入ると、コンピュータ技術の進展に伴い、大規模な電子化コーパスが構築・公開されるようになった。ここでは、話し言葉を含む代表的なコーパスとして**British National Corpus, BNC**を取り上げる。BNCは全体が約1億語から構成されており、そのうち1割に相当する約1000万語が話し言葉に充てられている。BNCにおける話し言葉のパートは、大きく分けて、日常会話を記録したものと、公的な場面における会話や独話を記録したものの2種に分類される (Aston & Burnard 1998)。

日常会話の収集に際しては、性別、年齢 (5区分: 15~24歳、25~34歳、35~44歳、45~59歳、60歳以上)、社会階層 (4区分)、地域 (3区分) のバランスを考慮して124人の協力者が選定された。選定された協力者には、約1週間、レコーダーを携帯して日常の会話を録音してもらうという手法が採用された。この方法により、会話の場面や相手の多様性を自然に確保することが意図されている。話し言葉全体のうち約420万語がこの手法により収録された。

一方で、この方法では日常会話中心のデータ収集となることから、多様な話し言葉を確保するために、会合や講義、セミナー、ラジオ番組といった公的な場での話し言葉も収録された。具体的には、表1に示す4つの領域が定められ、それぞれ独話と対話、計610万語がコーパスに集められた。

表1 公的な場の話し言葉の4領域

教育・教養	独話	講義、講演、実習、ニュース解説など
	対話	授業での先生・生徒間のやり取りなど
ビジネス	独話	企業発表、デモンストレーションなど
	対話	商用の打合せ、インタビューなど
団体	独話	政治演説、説法など
	対話	議会審議など
レジャー	独話	スポーツ実況中継、テーブルスピーチなど
	対話	放送の対談番組、クラブミーティングなど

1.4 日本語日常会話コーパス

最後に取り上げるのは、2016年から構築が開始され、2022年に公開された200時間規模の『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC)である。国立国語研究所において構築された独話中心の

『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese, CSJ)との差別化を図るために、日常会話に特化した構成となっている。

CEJCの最大の特徴は、日常会話の実態を的確に捉え、多様な会話をバランスよく収録するための設計指標とするために、構築に先立ち**会話行動調査**を実施している点にある。調査では、起床から就寝までに行った全ての会話について、「いつ」「どこで」「誰と」「何をしながら」「どのような種類の会話を」「どれくらいの長さで」行ったかを尋ねた(小磯ほか 2016)。この調査結果を参考にしながら、コーパスに収める会話を選定した。このように日常生活において生起する会話の分布を把握した上で、偏りのないデータ収集を目指した点に、CEJCの設計上の工夫が見られる。なお、CEJCについては3節と4節で詳述する。

□2 日本語の会話コーパス

前節では、話し言葉を対象とする4つのコーパス・調査を取り上げ、話し言葉がどのように収集されているかという点に焦点を当てながら、その設計の特徴を見てきた。これらはいずれも、できるだけ幅広い場面や多様な話者の話し言葉をバランスよく収集することを目指し、収録方法やサンプルの選定方法に様々な工夫が凝らされていることが確認された。

こうした幅広い場面・多様な話者の話し言葉の収集を実現するには、ある程度の規模が必要となる。実際、公開されている多くの話し言葉コーパスは、特定の場面や話者属性に偏っている傾向が見られる。ただし、それは必ずしも問題であるとは限らない。研究の目的によっては、特定の場面や特定の話者に限定されたデータがむしろ有用となることもあるためである。

したがって、コーパスを構築する際には、どのような目的に基づき、どのような設計方針で構成されているかを明確にすることが重要である。また、既存のコーパスを研究に利用する際にも、その設計を十分に理解した上で、自身の研究目的に照らして適切かどうかを慎重に検討する必要がある。

本節では、こうした観点に基づき、日本語会話を対象とするいくつかのコーパスについて、特にその設計面に注目しながら紹介する。

2.1 現日研・職場談話コーパス

『現日研・職場談話コーパス』は、2018年に国立国語研究所のオンライン検索システム「中納言」で公開されたコーパスだが、その元になったデータは、『女性のことば・職場編』『男性のことば・職場編』(現代日本語研究会 1998, 2002)²

の付属CD-ROMに収められた文字化テキストである。現代日本語研究会が中心となり収集・整備された。音声データは公開されていない。

『女性のことば・職場編』では、1993年に首都圏の有職女性19人（20～50代）が、『男性のことば・職場編』では、1999年から2000年にかけて同じく首都圏の有職男性21人（20～50代）が、それぞれの職場で自然談話を録音している。いずれも録音の対象は、(1) 朝、出勤後の1時間、(2) 会議・打合せ中の1時間、(3) 休憩時間の1時間であり、それぞれから内容のまとまりがある約10分の談話が切り出され、文字化されている。その結果、『女性のことば』には9.2時間分、154人の話者による会話が、『男性のことば』には12.1時間分、215人の話者による会話が収録されている。

この調査の特徴は、職場という限定された場面においても、フォーマル・インフォーマルを含む多様な会話状況や、年齢・職種における話者の多様性を意識して設計されている点にある。例えば、会議や打合せなどの公式な場面だけでなく、休憩時間の雑談や、来訪者とのやり取りなども対象としており、話者間の関係性やフォーマリティの違いが比較可能なよう工夫されている。また調査方法としては、協力者自身に録音を依頼する形式が採用されており、この点は1節で紹介したBNCやCEJCと共通する。協力者は年代毎に偏りが生じないように配慮されており、20代から50代までの各年代で男女とも3～6人が選ばれている。結果として、収録された会話データは、当時の職場における年齢や性別の構成をある程度反映しており、特定の層に偏った資料とはなっていない。年齢別・性別の話者数の分布を見ると（図3）、30代から60代にかけて年齢が上がるにつれて話者数が減少する傾向が見られるが、これは実際の職場における年齢構成のピラミッド型分布を反映していると考えられる。また、20代を除いて男性話者数が女性よりも多いのは、1990年代当時の職場において男性の比率が高かったことによると推測される。

2) 現在、これらの書籍内容・CD-ROMのデータはまとめられ『合本 女性のことば・男性のことば（職場編）』（現代日本語研究会 2011）として刊行されている。

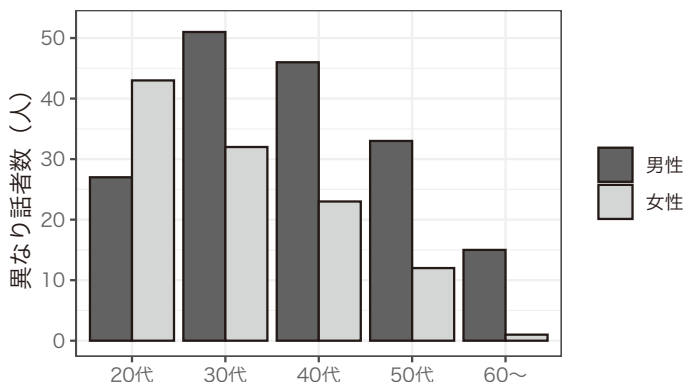


図3 『現日研・職場談話コーパス』話者の性別・年齢の内訳

2.2 名大会話コーパス

『名大会話コーパス』は、2001年から2003年にかけて収録された、約100時間分、計129本の自然会話が収められたコーパスである（藤村ほか 2011）。音声データは公開されていない。当時、コーパスを用いた言語研究が活発化していたが、利用可能な資源は書き言葉に偏っており、話し言葉、特に自然会話を対象とするコーパスは限られていた。本コーパスは、このような背景のもと、限られた条件の中で可能な限り多様な自然会話を収録することを目指して構築されたものである。

収録は、名古屋大学で日本語学・日本語教育を専攻する大学院生が主に担当したが、地域的な偏りを避けるため、研究プロジェクトの研究代表者や研究分担者なども収録に携わった。本コーパスに含まれる話者198人のうち44%は大学のある愛知県の在住者であったが、残る56%は国内16都道府県および韓国在住者で構成されており、地域的な広がりも確保されている。

また収録対象とする会話は2人会話を基本とした（全体の74%）。動画は記録されておらず音声のみの収録であったため、文字化に際して話者を識別しやすくするための制約であった。しかし、性別や声質の差により話者の識別が可能と判断される場合には3人から4人の会話も収録し、できるだけ会話人数の多様性を確保するよう配慮されている。

会話の内容は、友人や同級生、家族など、親しい関係にある者同士のインフォーマルな会話が多くを占めるが、初対面同士の会話や仕事関係の会話など、フォーマルなものも含まれている。一方で、話者の性別および年齢については、

20代女性が突出して多く、全体として女性話者の比率が高い(図4)。これは、大学院生が主体となってデータを収集したという背景に起因すると考えられる。ただし、10代から90代までの幅広い年齢層が含まれており、多様な話者属性を意識した収集が行われたことも確認できる。

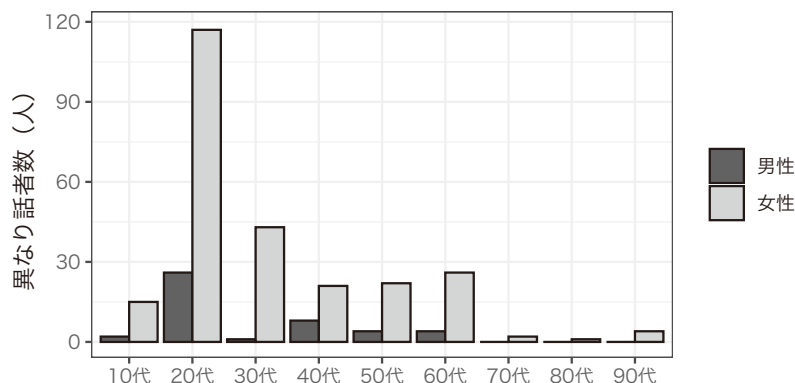


図4 『名大会話コーパス』話者の性別・年齢の内訳

『名大会話コーパス』は2009年に国立国語研究所に移管され、現在は研究所のウェブサイトを通じて文字化テキストとメタ情報が公開されている。また形態論情報(短単位情報)が付与されたデータは、コーパス検索アプリケーション「中納言」において検索が可能であり、更に全文検索システム「ひまわり」のパッケージも併せて提供されている。

2.3 BTSJ 1000人日本語自然会話コーパス

『BTSJ 1000人日本語自然会話コーパス』(以下、BTSJコーパス)は、基本的な文字化の原則BTSJ (Basic Transcription System for Japanese) に基づいて文字化された日本語の自然会話データと、対応する音声を取録したコーパスである。本コーパスは、人間の相互作用としての言語使用を語用論的に分析することを目的として開発された。データは、段階的に拡張を重ねながら整備され、2023年には514会話、約127時間、延べ1000人以上の話者を含む大規模なコーパスとして公開された。

BTSJコーパスの特徴は、会話グループ毎に収集の目的や会話条件が統制されている点にある。これにより、研究目的に応じて適切な会話グループを選択

し、話者の年齢・性別などの社会的属性や話者間の関係性といった要因を踏まえた分析が可能となっている。例えば、会話グループ20では、女性同士の初対面の雑談という条件のもと、日本語母語話者同士の会話と、日本語母語話者と日本語学習者との接触場面の会話が収録されており、母語話者間と接触場面との比較ができる。このようにBTSJコーパスは、多様な社会的条件に基づいた談話研究を可能にする設計となっている。

BTSJコーパスに含まれる文字化テキストおよび音声ファイルは、国立国語研究所のウェブサイトからダウンロードすることができる。BTSJコーパスの一部の会話グループには動画データも含まれているが、これらの動画はダウンロードすることはできず、『自然会話リソースバンク』において視聴する形式で提供されている。

2.4 日本語話題別会話コーパス

『日本語話題別会話コーパス』(Japanese Topic-Oriented Conversation Corpus, J-TOCC) は、15の異なる話題について、120組の話者ペアが各話題につき5分間会話したデータを収録した会話コーパスである(中俣ほか 2021; 中俣 2023)。録音は2018年から2019年にかけて実施された。音声データは公開されていない。

J-TOCCは、日本語教育において、話題を軸とした学習活動や教材開発を支援することを目的として構築された。語彙、文法、談話ストラテジーといった言語的要素に対して話題が及ぼす影響を検討可能にするため、話者の属性や関係性、会話時間といった諸条件を統一している点に大きな特徴がある。

例えば、話者は全て日本語を母語とする20歳以上の大学学部生であり、会話ペアは互いに親しい関係にある者同士で構成されている。また、性別の組合せ(男性同士・女性同士・男女)においてもバランスを取り、性別が談話に与える影響を分析可能な設計となっている。更に、地域的なバランスにも配慮し、東日本・西日本の両地域から均等な時間分のデータが収集されている。

話題については、山内(2013)に掲載されている100の話題の中から、山内(2018)が整理した親密度・必要度3段階を参考に、「身の回りの話題」11件と「社会に関わる話題」4件が選ばれた。「身の回りの話題」については、大学生にとっても身近であり、初級日本語学習者向けとして適切であることなどの観点から、親密度・必要度が低いものを除いた上で、「食べること」「ファッション」「旅行」「スポーツ」「マンガ・ゲーム」「家事」「学校」「スマートフォン」などが選ばれた。

また、話題の多様性を確保するため、「社会に関わる話題」として、やや複雑で身近でないテーマも含む4つの話題「夢・将来設計」「マナー」「住環境」「日本の未来」が加えられた。「社会に関わる話題」は、親密度・必要度の低いものから選ばれている。

15の話題それぞれに、山内(2018)による親密度・必要度と具体度(具体名詞の割合の多寡による分類)の情報がまとめられている。また各話者が、それぞれの話題についてどれだけ詳しいか、あるいはどれだけ自信を持って話すことができたか、という観点から5段階で自己評価した話題精通度の情報も提供されている。そのため、話題の親密度・必要度や具体度、話者の話題への精通度が言語使用に与える影響を分析することもできる。

□3 『日本語日常会話コーパス』の構築

CEJCの特徴は、(1)多様な場面における多様な話者による会話をバランスよく収集していること、(2)動画まで含めて収録・公開していること、(3)多様なアノテーションを整備・公開していることである。CEJCの構築に関する一般的な流れについては小磯ほか(2022, 2023)などで詳細に紹介した。そこで本節では、上記(1)～(3)のような特徴を持つコーパスを構築するにあたり、特に検討や調整を要した項目に重点を置き、CEJCの構築過程を準備研究の段階から振り返りつつ紹介する。

3.1 『日本語日常会話コーパス』構築の準備

3.1.1 動画を含むコーパス公開に向けた法的・倫理的問題の検討

CEJCの構築にあたって最初に直面した課題の1つは、日常生活における会話を動画付きで収録・公開する際に、どのような法的・倫理的問題が生じ得るかを明確化することであった。特に問題となり得るのは、公共商業施設や屋外などでの収録において、撮影や公開に同意していない第三者の容貌が写り込む場合の対応である。このような第三者の映像に対しては、モザイクなどのマスキング加工を施すことも1つの対応策となるが、その処理には多大な手間を要し、かつ研究資料としての価値が損なわれる懸念もある。そのため、過剰な編集はできる限り避け、法的・倫理的観点から問題がないとされる範囲を見極めた上で、必要最小限の対応に留めるべきであると考えた。この問題は弁護士と相談の上、主に肖像権の観点から検討した。肖像権の侵害に該当するか否かは、次のような複数の要素を総合的に勘案して判断される。

撮影の目的 公共性・公益性が低いほど侵害とみなされやすい。

撮影の場所 公道や商業施設などの公共的な空間よりも、自宅などの私的空間における撮影の方が侵害と判断されやすい。

撮影された被撮影者の活動内容 日常的な行動に比べ、センシティブな場面（例：職務質問、病院でのやり取りなど）は侵害とみなされやすい。

撮影の様態 隠し撮りなど不自然な方法による撮影は侵害と判断されやすい。

撮影の必要性 目的に対して撮影が不必要と判断される場合、侵害の可能性が高まる。

公開の手段 不特定多数への拡散性が高い手段での公開は侵害と判断されやすい。

被撮影者の社会的地位 公人や有名人よりも、一般人の方が侵害と判断されやすい。

これらの要素を踏まえ、CEJCの収録場面に即して整理すると、(1) 撮影の目的：日常会話コーパスの構築・公開という公共性の高い目的のために、(2) 撮影の場所：公道や公園、公共商業施設など公の場所において、(3) 被撮影者の活動内容：飲食や歩行など日常的な行動をしているところを、(4) 撮影の様態：隠し撮りなどではなく通常の方法で、(5) 撮影の必要性：日常生活における会話の記録のために必要となる範囲を収録し、(6) 公開の手段：拡散可能性の高い公開方法ではなく国立国語研究所と契約した場合にのみ提供する場合は、肖像権の非侵害と判断される可能性が高いと考えられる。

一方で、私的空間における第三者の写り込みや、未成年者・社会的弱者とされる人物が写り込む場面、あるいはセンシティブな行動が含まれるケースについては、個別に対応を検討し、例えば動画へのマスキング加工や、データの非公開措置をとる必要がある。

後述の通り、CEJCでは主として一般の**調査協力者**（以下、協力者）に収録を依頼する計画であったため、収録実施にあたっては、(1) 私的空間での収録にあたっては必ず関係者の許諾を得ること、(2) 公共の場での撮影では隠し撮りとならないよう収録中であることが周囲に明確に分かるようにし、必要に応じて収録の目的を説明すること、(3) カメラの位置や向きを調整し、不要に第三者が大きく写り込まないよう配慮すること、などの注意点を事前に明示して共有することとした。

3.1.2 会話行動調査の実施

1.4節でも言及したように、多様な場面における多様な話者による会話をバランスよく収集するために、我々が日常生活の中で、どのような種類の会話を、どの程度行っているのかを把握する必要があると考えた。そこで予備調査として、日本放送協会（2011）による国民生活時間調査などの既存調査を参考にしつつ、独自の**会話行動調査**を計画し、2014年から2015年にかけて実施した（小磯ほか 2016）。調査は、首都圏在住の243人を対象に行われた。調査期間は平日2日、休日1日とし、起床から就寝までに行った全ての会話について、「いつ」「どこで」「誰と」「何をしながら」「どのような種類の会話を」「どれくらいの長さで」行ったかを回答してもらった（図5）。

① どんな会話か			
<自由にメモしてください>			
自宅で夕食を食べながら家族と雑談			
② いつ (1つ選択)			
☐ 午前	☐ 午後	☒ 夜(午後6時頃～)	
③ どのくらい (1つ選択)			
☐ 5分未満	☐ 5～15分	☐ 15～30分	☒ 30分～1時間
☐ 1～2時間	☐ 2～5時間	☐ 5～10時間	☐ 10時間以上
④ どこで (1つ選択)			
☒ 自宅	☐ 職場・学校	☐ 公共商業施設	☐ 交通機関
☐ それ以外の屋内		☐ それ以外の屋外	
⑤ だれと (あてはまるものをそれぞれに人数を記入)			
家族: <u>3</u> 人	親戚: <u> </u> 人	先生・生徒: <u> </u> 人	
仕事・学業関係: <u> </u> 人	公共商業関係: <u> </u> 人		
友人・知人: <u> </u> 人	顔見知り・見知らぬ人: <u> </u> 人		
⑥ 何をしながら (1つ選択)			
☒ 食事	☐ 家事・雑事	☐ 身周りの用事	☐ 療養
☐ 仕事・学業	☐ 業務外・課外活動	☐ 社会参加	
☐ レジャー活動	☐ 付き合い	☐ 移動	☐ 休息
⑦ どんな種類 (1つ選択)			
☒ 雑談	☐ 用談・相談	☐ 会議・会合	☐ 授業・レッスン・講演
⑧ その他 (あてはまるものすべて選択)			
☐ 電話・スカイプなどの遠隔での音声・映像会話			
☐ 外国人を含む会話		☐ 外国語を含む会話	

図5 調査票の記入様式

得られた結果は、収録すべき会話場面の想定、収録機材の選定、収録手順の策定など、構築準備に関わる多岐にわたる判断の根拠として活用した。また CEJCに格納するデータの多様性とバランスを確保するための指針としても活

用された。調査の詳細については小磯ほか（2016）を参照されたい。

3.1.3 収録法の策定

CEJCでは、場面および話者の多様性を確保するための方法として、**個人密着法**と呼ばれる収録方式を基本とすることとした。この方式は、1.3節で紹介したBNCの収録法を参考に検討されたものであり、性別と年齢のバランスを考慮して選定された協力者に収録機材を一定期間貸与し、その期間中に日常生活の中で自然に生じる多様な場面・多様な話者との会話を記録してもらうというものである。収録された会話のうち、コーパスに格納するデータは、収録の質と倫理的・法的観点からの適切性、場面や話者の多様性、更に協力者の希望などを考慮して選定を行う。また個人密着法によって一定量の会話を収録した後、会話行動調査の結果を参照しながら、収録された場面や話者属性のバランスを検証し、不足する場面や話者属性を特定する。その上で、調査者が主体となって必要な会話を追加で収録することとした。この補完的な収録方式を**特定場面法**と呼ぶ。

3.1.4 収録機器の選定

収録機器の選定にあたっては、調査者および一般の協力者による複数回のテスト収録を通じて、操作性や携帯性、収録品質などの観点から実用性を検証した上で、最適な機器構成を確定した。なお、ここで言及する機器や技術的設定に関する基本的事項は第2章に詳述されているので、必要に応じて参照されたい。またCEJCで実際に使用した機器や設定の詳細については田中ほか（2018a, b）を参照のこと。

■**基本要件** 機材選定にあたっては次の3点を重視した。(1) 幅広い世代の一般の協力者が容易に操作できること、(2) 屋外での使用も想定し、携帯性に優れた軽量の機材であること、(3) 話者の視界に入る位置に配置しても圧迫感を与えないこと。

■**録画機材の選定** 動画収録には、複数のウェアラブルカメラを組合せて使用する方法を採用した。主構成として、広い視野角を有し様々なアタッチメントが利用可能なカメラ（GoPro Hero3+）2台を用いて会話場면을俯瞰的に撮影するとともに、360度撮影可能なカメラ（Kodak PIXPRO SP360 4K）を会話の場の中心に設置し、話者全員の動画を網羅的に記録できるようにした（図6）。更に、

会話行動調査の結果、散歩など移動を伴う場面の会話も少なくないことが明らかになったため、腕などに装着可能なカメラ（Panasonic HX-A500）も導入した。なお、特定場面法では調査者が主体となって収録を行うため、機材操作や携帯性に関する制約が緩和される。そこで、大人数による会議などの収録に際してはデジタルビデオカメラ（Sony HDR-CX675）も併用した。

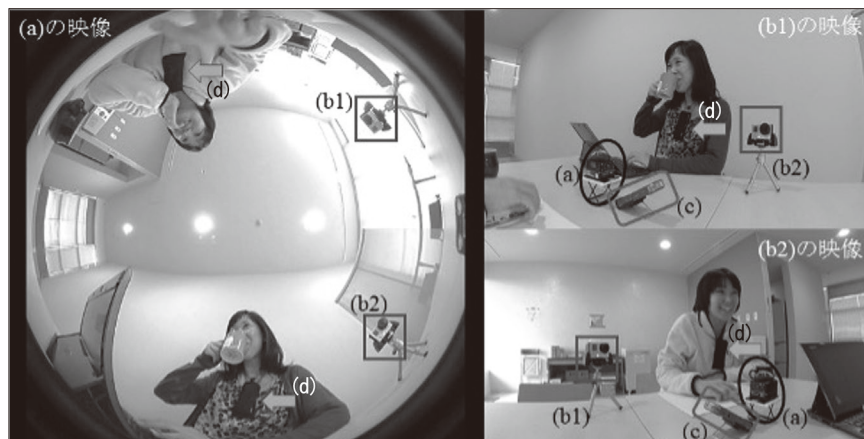


図6 収録機材：左の映像は(a)の360度撮影可能なカメラで、右上と右下の映像はそれぞれ(b1)と(b2)のカメラで撮影。また(c)は会話全体を、(d)は個人の音声を記録するためのICレコーダー。

■録音機材の選定 音声収録においては、会話全体の音声と各話者の発話の両方を明瞭に記録することとした。多チャンネル録音機器やワイヤレスマイクの使用も検討したが、操作性・携帯性を考慮し、通常よく使用されるICレコーダーを採用した。会話の場の中心に置いたICレコーダー（Sony ICD-SX1000）で会話全体の音声を記録すると同時に、各話者が身に付けたICレコーダー（Sony ICD-SX734）により当該話者の音声をより明瞭に記録することとした。

後者については、ICレコーダーをフォルダーに入れ、首から下げる方式を採用した。ICレコーダーは、当該話者の音声を明瞭に収録するために単一指向性の内蔵マイクを備え、かつ、音声分析に耐え得る音質を確保するために無圧縮のリニアPCM方式を選択できる機種とした。会話行動調査の結果、話者が2人から6人の会話が全体の90%以上を占めていたことから、ICレコーダーは最大6台まで同時使用できるよう準備することとした。

3.1.5 収録手順・補助資料の策定

事前に実施したテスト収録の結果を踏まえ、個人密着法による収録手順を整理した。個人密着法の最大の特徴は、収録に関する一連の作業を協力者自身が担う点にある。そのため協力者には、(1) 音声・動画の収録、(2) 会話参加者に対する調査内容・データ公開方法などの説明、(3) **同意書**への署名依頼、(4) **フェイスシート**の記入依頼、(5) 収録状況(日時、使用機材、機材配置など)の記録といった、多岐にわたる作業を依頼する必要がある。(2)～(4)については次節で言及する。

これらの作業を円滑に進めるため、機材の使用方法や設定手順、収録作業の流れをまとめた手引きを作成した。手引きは図を多用して直感的に分かりやすいものとなるよう心掛けた。また詳細な情報を記した手引きとは別に、収録時に携帯できるよう小型の手引きも作成した。

収録状況を記録するための記録ノートにも工夫を加えた(第2章、図7参照)。このノートには、会話の概要や日時、機材の種類、話者や機材の配置などの記録欄を設けるとともに、見開きの反対ページには、機器の電源や録画ボタンの位置といった基本的事項に加えて、「カメラのレンズキャップを外す」「録音開始後にホールド機能を有効にする」など、操作上見落としやすい項目を記載し、協力者が手順を確実に踏めるよう配慮した。各話者が装着するICレコーダーについても、レコーダーと話者の対応関係の混乱を避けるために、レコーダーやフォルダーに6色の識別色を割り当て、記録ノートにも対応する色を記載した。これにより、誰がどのレコーダーを使用したかを視覚的に確認でき、対応関係の取り違いを防ぐようにした。

これらの資料や手順は、本調査開始後も随時見直しを行い、必要に応じて更新を加えた。なお、個人密着法による収録のために作成した一連の資料は田中ほか(2018b)の付録に収められている。

3.1.6 研究倫理審査

人を対象とする研究では、調査参加者の権利を保護し倫理的・法的問題を回避するために、事前に**研究倫理審査**を受け、**インフォームド・コンセント**を適切に得る必要がある。特にCEJCは、日常生活で交わされる会話を動画まで含めて収録・公開するため、コーパスの構想段階から研究倫理審査を想定し、同意取得の手続きやデータの管理・加工・公開方針、リスクの想定とその軽減策などを具体的に検討した。CEJCでは、調査参加者に渡す文書として、**説明文書**、

説明文書の補助資料（フライヤー）、同意書・同意撤回書、フェイスシートを準備した。

■**説明文書** 説明文書は、調査の目的や方法、参加の任意性、同意撤回の権利、研究参加に伴う危害の可能性の有無とその内容、公開の方法、個人情報の取り扱い、調査者の連絡先など、必要事項を過不足なく記載する必要がある。CEJCのように動画、音声、転記テキスト、各種アノテーション、話者や会話に関するメタ情報などを公開する場合は、何をどのように加工して公開するかを具体的に明示する必要がある。CEJCの説明文書では次のように記載した。

収録した会話の映像・音声・文字化資料、研究用の情報、フェイスシートにご記入いただいた生年（5年刻み）、そこから求めた年齢（5歳刻み）、性別などの情報、協力者が記録した会話状況に関する情報は、個人情報などの扱いに配慮した上で、次の2つの方法で公開することを予定しています。

【**オフライン公開**】 国立国語研究所が定める利用目的に同意し契約書を交わした者に対し、ハードディスクなどの記録媒体を通じて、オフラインで公開します。その際、名前などの個人が特定できる箇所は、その音声が聞こえないように加工し、文字化資料では、その箇所を仮名（かめい）や伏せ字に置き換えるなどの処理をします。また、ご本人が公開を希望しない箇所については、音声・文字化資料を同様に加工します。

【**オンライン公開**】 ウェブなどオンラインで公開する場合、「オフライン公開」に記した音声・テキストの加工に加え、顔の一部にぼかしを加えるなど、個人が特定できないように映像も加工します。また、会話の短いシーン毎の映像・音声の視聴に留め、会話全体を通してオンラインで閲覧できないようにします。

■**説明文書の補助資料（フライヤー）** 協力者が調査の主旨や公開方法を会話参加者などに説明する際の負担を軽減するために、文字主体の説明文書・同意書に加えて、説明用のフライヤーを用意した（図7）。このフライヤーには、調査の内容や目的、意義などを簡潔に示すとともに、音声・動画・文字化テキストのそれぞれについて、名前などの個人情報が含まれた際にどのような加工が施されるかを、図や写真を用いて直感的に把握できるよう工夫した。更に、より詳細な情報が必要な場合に備えて、国立国語研究所の調査担当者の連絡先を明

記し、紙媒体の配布に加え、プロジェクトのウェブサイト上でも同内容を閲覧できるようにした。

国立国語研究所 日常会話 調査研究

■ どんな研究？

国語研究所では、さまざまな会話場面の映像と音声を収録し、日常の会話場面で私たちがどのような言葉遣いや振る舞いしているかを調べています。収録した映像・音声データは、会話のデータベースとして登録・公開することを予定しています。

■ 何が分かるの？

- ◆ 場面や相手によって言葉の使用はどのように変わる？
- ◆ 書き言葉と比べて会話の言葉はどのような特徴がある？
- ◆ 昔の言葉と比べて言葉の使い方や声の抑揚などはどのように変化している？
- ◆ 複雑やジェスチャーなどの非言語行動はどのような役割を果たしている？

■ 何の役に立つの？

- ◆ 会話の言葉遣いや振る舞いの特徴を調べることは、素人に話し言葉のしくみの理解にとどまらず、発話を適切にコントロールし対人関係などの理解につながるります。
- ◆ 動画や音声による言葉の使い方の違いなどが分かることで、外国人が日本語を学ぶときの助けになります。
- ◆ 今、人間と会話をするロボットが出てきました。私たち人間の日常会話を数多く納めたデータベースは、人間と会話するロボットの性能向上につながります。
- ◆ 昔の会話を参照して当時の生活や文化を知ることができるように、後世の人々が21世紀初頭の日本人の生活や文化を知るための貴重な記録となります。

◆◆◆◆◆ お問い合わせ先 ◆◆◆◆◆

この調査に関してご質問やご相談などありましたら、下記までお問い合わせください。

担当：国立国語研究所 日常会話コーパスグループ
メールアドレス： corpus@nins.ac.jp
電話：03-5861-3131（直通）
代表 小嶋花緒

プロジェクトのWebサイト
<http://gj.ninjal.ac.jp/conversation/corpus.html>



◇ 音声・テキスト

個人情報（お名前・所属組織名など）やご本人が公開を望まない箇所は、次のように加工します。

- ▶ 音声：聞こえないように加工
- ▶ 文字化テキスト：仮名（かめい）や伏せ字（××）に置き換える

仮名に置き換えた場合の例
音声は聞こえないように加工

◇ 映像

国語研究所と契約を交わした者に限定して公開する場合、会話参加者の顔にぼかしなどの加工は加えません。研究者など限られた利用にとどめます。

上記以外の方法で公開する場合、左の図のように、全ての会話参加者の顔にぼかし処理を施します。日本語学習者や学校の先生などの幅広い利用が見込まれます。

なお、会話を全体ではなく短いシーンごとの映像・音声の公開に留めます。

実際の会話の例

あつ（山本さん） 醤油、その棚に置いてあるんだけど、ちょっと取ってくれる？ あと、ついでに（太田さん）に醤油皿もお願ひ

公開用に加工したテキストの例

あつ（山本さん） 醤油、その棚に置いてあるんだけど、ちょっと取ってくれる？ あと、ついでに（太田さん）に醤油皿もお願ひ

■ どうやってデータを公開するの？

図7 会話参加者に対する説明用のフライヤー

■同意書・同意撤回書 同意書は、次の通り、調査参加およびデータ公開への同意を兼ねた書式とした。その際、説明文書とは別に、改めて公開するデータの範囲、加工や公開方法に言及し、誤った理解での同意を避けるよう配慮した。

私は、〇〇氏から、「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」に関する説明を受け、この調査への参加、および、この調査において記録された私の映像・音声・文字化資料、研究用情報、フェイスシートに記入した情報、会話状況情報等の公開について、以下の条件のもとに同意します。

1. 国立国語研究所が定める研究教育利用・商業利用（統計情報利用）の条件に同意し契約を交わした者に対して公開する際には、データに以下の処理をほどこす。

- 私の名前、学校や会社など私が所属する組織の名称、自宅・所属組織の住所・電話番号の音声は聞こえないように加工し、文字化資料においても、仮名や伏せ字に置き換えるなどの処理をする。

●私が公開を望まない箇所の音声・文字化資料も同様に加工する。

2. 第1項以外の方法での公開については、上記に加え、顔の一部にボカシを加えるなど、私個人が特定できないように映像を加工する。また、会話全体ではなく短いシーン毎の映像・音声の公開に留める。

同意撤回書は、調査参加者が後から自由意思により同意を取り消せることを保障するために、必ず用意する必要がある。

■フェイスシート 話者の性別・出身地・職業などを記入してもらう調査票である。個人情報も多く含むため、研究に必要かつ十分な情報に絞り、過剰な取得は避けるよう心掛ける必要がある。CEJCでは、動画を含む日常会話が公開されることから、言語研究に必要な情報と、調査参加者を保護するために避けるべき情報との接点で、何を収集するかを慎重に検討した。例えば、出身地や居住歴などは言語研究に有用であるため情報を収集することとしたが、詳細には尋ねず都道府県レベルに留めることとした。また協力者には相手との関係性について尋ねたが、親疎関係まで踏み込んで質問することは控えた。

以上、CEJCの研究倫理審査に向けた準備について説明した。調査参加者の信頼を得て協力を仰ぐには、倫理的配慮が研究活動の中心に据えられていなければならない。説明の明確化、適切な同意の取得、情報の安全な管理、適切なデータの加工など、全てが揃って初めて、安心して活用できる研究データが生まれることを心に留めておく必要がある。なお、倫理審査で用いた上記の書類は、田中ほか(2018b)で確認できる。

3.2 『日本語日常会話コーパス』構築の流れ

3.2.1 会話の収録とバランスの検証

■個人密着法に基づく収録 プロジェクトの予算および期間の制約を踏まえ、個人密着法による調査の協力者は40人とした。性別・年齢層の点から均衡性を考慮して選別された首都圏に在住する40人の協力者(男女×20代・30代・40代・50代・60歳以上×各4人)に収録を依頼し、協力者1人あたり10～15時間、収録してもらった。協力者の選定にあたっては、広く公募する方法は採用せず、知己からの紹介を通じて選んだ。協力者に動画も含めて日常会話を収録してもらう必要があり、責任を持って調査に取り組める信頼性の高い人物であることが重要だったためである。

収録調査に先立ち、協力者との打合せを行った。打合せでは、収録の手順や収録機器の操作方法などの説明に加え、図8にあるような会話サンプルを提示しながら、自身の生活を振り返ってどのような場面が収録可能かを具体的に相談した。この手順を踏むことによって、収録対象とする場面や話者の多様性を確保することが可能となった。また、収録の場面をある程度想定することができるため、車にカメラを固定するためのアタッチメントなど、必要となる収録用の補助器具を準備することも可能となった。

例えば私（女性／会社員／両親と暮らしている）なら・・・

- | | |
|-----------------|-------------------|
| ・ 平日、自宅で家族との夕食 | ・ 職場でのランチ |
| ・ 家族と週末の予定の相談 | ・ 仕事帰りに同僚とご飯 |
| ・ 母とお菓子作りをしながら | ・ 休日に友人とカフェでおしゃべり |
| ・ 母にスマホの使い方を教える | ・ 友人宅にお泊り |
| ・ 父とパソコンのセットアップ | ・ 習い事の手芸教室で |
| ・ 職場での打合せ | ・ 行きつけのネイルサロンで |

例えば私（男性／会社員／妻と小学生の子供と暮らす）なら・・・

- | | |
|------------------|-----------------------|
| ・ 自宅でゆっくり食事&団らん | ・ 職場での休憩中 |
| ・ 子供の勉強を見ながら | ・ 同僚との飲み会 |
| ・ 子供とボードゲームをしながら | ・ 出張で新幹線移動中に同僚と |
| ・ 家族で家具を組み立てながら | ・ マンションの理事会 |
| ・ 妻と家族旅行の相談 | ・ なじみの電器屋さんとエアコン修理の相談 |

図8 事前打合せで提示した会話場面の例（一部抜粋）

個人密着法に基づき全体で約500時間の会話を収録した。その上で、動画・音声のデータの質や倫理的・法的な問題、場面や話者のバランス、協力者の希望などを考慮し、コーパスに格納するデータとして全体で185時間（協力者1人あたり3～6時間）を選んだ。

■モニター版の構築とバランスの検証 200時間の会話を対象とする本公開に先立ち、50時間分のデータを対象にモニター版を公開した（小磯ほか 2020a）。モニター版の構築にあたっては、個人密着法の収録状況を踏まえつつ、性別・

年齢のバランスに配慮して公開対象とする協力者20人を選定した。その上で会話行動調査の結果などを参照してバランスを検証した。その結果、仕事や学業中の会話の割合が調査結果と比べて少ないこと、また20歳未満の未成年者が成人と比べて著しく少ないことなどが明らかになった(小磯ほか 2020a)。

■**特定場面法に基づく収録** モニター版を対象とする検証結果を踏まえ、特定場面法として、仕事中の会議会合を10時間、中高生の雑談・打合せ・講義などを5時間、計15時間を調査者が主体となり収録した。当初の計画では、個人密着法を160～180時間程度とし、残りを特定場面法で補う予定であった。しかし、特定場面法による収録期間が新型コロナウイルス感染症の流行期と重なり収録が困難であったため、特定場面法によって収録されたデータは約15時間に留まり、個人密着法で収録した会話を当初の計画より多めに採用することとした。

3.2.2 「コア」の設定

CEJC200時間のうち、1割に相当する20時間分の会話を**コア**と定め、多様なアノテーションを高精度に施すこととした。コアはモニター版の中から優先的に選定したが、その際、会話形式や話者の属性のバランスを考慮するとともに、協力者毎の会話数・時間のばらつきを抑えて多様性を持たせるようにした。また、後述の通りコアに対して**X-JToBI**に基づく**韻律情報**を付与することから、ノイズレベルが著しく高いものや、**X-JToBI**が対象とする東京方言から大きく逸脱した発話が多くを占めるものは対象外とした。結果、モニター版に含まれる協力者20人から関西出身の1人を除き、かわりにモニター版以外の協力者1人を含め、20人の協力者による20時間の会話をコアとして選定した。コアの詳細は小磯ほか(2020b)を参照されたい。

3.2.3 転記・アノテーション

収録した動画・音声データを対象に、第2章4節に示すような形式変換や同期処理を行った上で、**転記テキスト**の作成や**アノテーション**付与を実施した。

■**転記テキスト** **ELAN**(第6章)や**Praat**(第7章)上で音声波形や動画を参照しつつ、統語的・談話的・相互行為的なまとまりを持った**発話単位**や、それを知覚可能な無音で区切った転記単位を認定しながら、発話内容を漢字仮名交じりで表記した。言いよどみや言い間違い、音声的・韻律的特徴など、文字だけでは表現しきれない様々な情報を補足するために、20種類以上の**転記用タグ**を導

入して転記テキストに付与した。転記の詳細は第3章を参照のこと。

■**形態論情報** 短単位と長単位という2種類の粒度の形態論情報を付与した。短単位は、発話単位を入力単位とし、形態素解析器MeCabと電子化辞書UniDicを用いて自動解析した上で、200時間全体を人手で修正した(西川・渡邊 2020)。長単位は、短単位を元に、長単位解析器Comainuを用いて自動解析した上で、コアに限定して人手で修正した。形態論情報については第5章1節を参照のこと。

■**文節係り受け情報** 文節間の係り受け関係³⁾は、コアを対象に、発話単位を範囲に文節係り受け解析器CaboChaにより自動付与した上で、人手で修正した(浅原・若狭 2022)。BCCWJに対する文節係り受け情報(浅原・松本 2018)の基準に準じ、通常の係り受け相当、文境界相当、後続文節との連結、係り先不定をそれぞれ示す4つのラベルが適用された。

■**談話行為情報** ISO24617-2 (2012) をベースに日常会話用に整備した基準に基づき、コアを対象に、談話行為情報を人手で付与した(Iseki, *et al.* 2019)。発話単位毎に、意味的・語用論的レベルの情報(例: 質問、応答、注意獲得、感謝)や、インタラクションレベルの情報(例: 修復、会話の開始・終結)などが付与された。談話行為情報については第4章を参照のこと。

■**韻律情報** コアに含まれる157人の主たる話者(一時的な会話参加者を除く)のうち、方言の使用状況や音質を考慮して151人を選別した上で、CSJ用に整備したX-JToBI(五十嵐ほか 2006)の簡略版に準拠して韻律情報を付与した(小磯ほか 2020b)。韻律情報については第7章1節を参照のこと。

3.2.4 データ公開方針の策定

動画・音声データや転記テキストを公開するにあたり、収録・公開の承諾を得ていない第三者の容貌や音声記録に含まれる場合の対応が、重要な課題となった。そのため、収録データが一定量に達した段階で、具体的な事例を整理し、専門の弁護士との協議を経て公開方針を策定した。主な方針は以下の通りである。詳細については小磯・伝(2018)を参照されたい。

●データ公開の基本方針は、収録前に話者と交わした同意書の条件(3.1.6節参

3) 2022年3月のCEJC本公開後に、Universal Dependenciesに基づく長単位・短単位を言語単位とする係り受け情報も提供された(大村ほか 2025)。

照)に準拠する。話者の名前、所属組織名、自宅・所属組織の住所・電話番号については、動画・音声にはマスキング加工を施し、転記テキストでは仮名化・伏せ字化を行う。また、会話中に登場する一般人の実名や、ヒアリングにおいて協力者が非公開を希望した内容についても同様の対応を行う。

- 第三者の容貌が動画に写り込んだ場合には、肖像権の観点から総合的に判断する(3.1.1節参照)。例えば公の場で一般的な行動をしている場合は公開対象とするが、私的またはセンシティブな場面、あるいは社会的弱者が写っている場合は、マスキング加工を施すか当該範囲を非公開とする。
- 第三者の発話が収録音声に混ざり込んだ場合、プライバシー権などの観点から総合的に判断する。例えば第三者が一時的に会話に参加する場合、一般的なやり取りであれば転記対象とするが、私的な内容の場合は非公開とする。また周囲の第三者の発話が背景音として大きく記録された場合、私的な内容を多分に含む箇所を公開対象から外すなどの措置を講じる。

なお、CEJCにおける転記テキストの仮名化・伏せ字化の対応については第3章4節を、また音声や動画に対する一般的なマスキング加工の方法については第2章4.3節を参照のこと。

3.2.5 検索システム・リレーショナルデータベースの構築

転記テキストや各種アノテーションは、CSV形式のほか、アノテーションツールELANや音声分析ソフトウェアPraat用のファイルなどで提供している。これらに加え、次の方法でも公開している。

コーパス検索アプリケーション「中納言」 国立国語研究所が提供するコーパス検索アプリケーション「中納言」で、短単位・長単位を用いた高度な検索や文字列による検索ができる。詳細は第5章を参照のこと。

全文検索システム「ひまわり」 言語研究用に設計された全文検索システム「ひまわり」で、短単位・長単位を用いた検索や文字列検索ができる。また観察支援システムFishWatchrが統合されており、検索した箇所の動画を視聴することもできる(山口2018)。

リレーショナルデータベース 形態論情報(短単位・長単位)、係り受け情報、談話行為情報、韻律情報を相互に関連付けて表現したリレーショナルデータベースCEJC-RDBを提供している。詳細は第8章を参照のこと。

□ 4 『日本語日常会話コーパス』の量的概観

本節では、CEJCに収録されたデータの量的特徴を概観し、コーパス活用にあたって留意すべき点を整理する。

CEJC全体の規模は、200時間、461セッション、577会話⁴、約240万語⁵、異なり話者数862人、延べ話者数1675人である。収録法別の規模およびコア・非コアの規模も合わせて表2に示す。

表2 CEJCにおける会話時間・セッション数・会話数・語数・話者数

	CEJC全体	個人密着法	特定場面法	コア	非コア
時間	200	185	15	180	20
セッション数	461	434	27	43	418
会話数	577	533	44	525	52
語数(万語)	240	223	17	215	25
異なり話者数	862	80	782	135	817
延べ話者数	1675	1511	164	145	1530

4.1 話者毎のデータ規模

CEJCは話者毎にデータの規模が大きく異なる。語数を指標に具体的に見てみよう。話者毎に総語数を算出し、語数の規模を10段階に分けた上で、該当する人数を個人密着法と特定場面法に分けて求めた。結果を表3に示す。個人密着法については協力者から見た関係性毎の内訳を示す。

4| セッションとは1回の収録で記録した範囲を、会話とは収録された範囲からまとまりを持って切り出された範囲を指す。1つのセッションが複数の会話に分かれることがある。

5| 語数(短単位数)を算出するにあたり、固有名などで伏せ字としたもの、語彙など不明で品詞情報が付けられなかったもの、品詞が記号あるいは歌(ハミングなど)のものは除いた。

表3 1人あたりの総語数と該当話者数

総語数/1人	個人密着法							特定 場面法	計
	本人	家族 親戚	友人 知人	仕事	先生 生徒	店員 客等	ほか		
0語（発話なし）	0	0	0	2	0	0	0	5	7
10語未満	0	3	2	6	0	19	2	1	33
50語未満	0	2	5	8	1	44	2	3	65
100語未満	0	4	7	7	2	24	0	3	47
500語未満	0	22	32	25	1	19	2	19	120
1,000語未満	0	19	42	30	2	3	3	10	109
3,000語未満	0	47	125	58	9	14	3	21	277
5,000語未満	0	27	34	19	6	2	2	6	96
10,000語未満	2	27	16	4	4	1	0	8	62
10,000語以上	38	4	0	0	0	0	0	4	46

協力者を中心に会話を収集する個人密着法では、必然的に協力者の語数が多くなる。実際、協力者（本人）は、40人中38人が10,000語以上、残る2人も5,000～10,000語と、語数が多い。そのため、協力者の発話に注目し、場面や相手によっていかに言葉を使い分けているかを調べる事が可能となる。

収録語数が少ない話者の傾向も見ておこう。7人については発話が一切ないが、このうち6人は人数の多い会議の参加者である。50語未満を見ると、会議の参加者のほか、飲食店の店員など会話に一時的に加わる人が多くを占める。このように収録語数が人によって大きく異なるため、分析の際には注意が必要となる。例えば、個人を無視して全体をまとめて集計すると、協力者など語数の多い話者の傾向が結果に強く影響する。それを避けるために、話者毎に集計した上で全体の傾向を見たり、混合効果モデルなど話者毎の違いを考慮した分析をするといった工夫が求められる。

4.2 話者の属性毎のデータ規模

4.2.1 性別・年齢層

図9に、性別・年齢層毎の延べ話者数と語数の分布を示す。

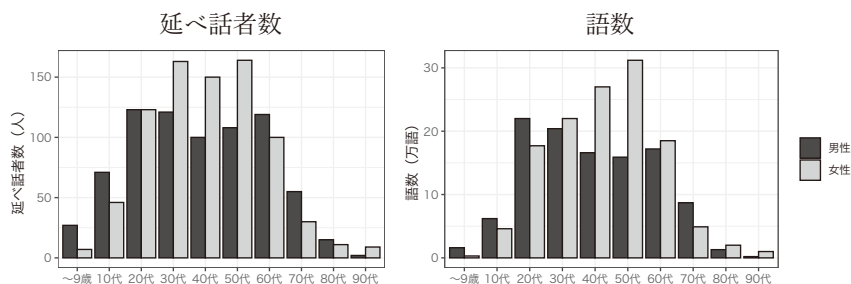


図9 性別・年齢層毎の延べ話者数と語数

CEJCの大半を占める個人密着法では、成人の協力者を中心に友人や同僚、家族などとの会話を収録しているため、必然的に協力者と同世代の話者が多く含まれることになる。図9から、40代・50代の男性が若干少なく女性が多いなどの多少の違いは見られるものの、20代から60代までの成人についてはいずれの世代の男女とも延べ話者数100人以上、15万以上の語を含んでおり、概ねバランスよく収録できていることが分かる。70代、80代、90代となるにつれ人数は減る。また20歳未満についても20～60代と比べて人数がかなり少ない。世代間や年齢層間の比較を行う際には注意する必要がある。

4.2.2 出身地・居住地

個人密着法では首都圏に在住する協力者40人に収録を依頼したため、また特定場面法でも歯科医院での会合を除き首都圏で収録したため、収録対象となる話者の居住地は首都圏に集中している。そのことも影響し、出身地も首都圏が多く、出身地不明を除く異なり話者数726人のうち東京都出身が全体の41%、首都圏出身が全体の64%を占めている。また首都圏以外に異なり話者数が20人以上あるのは、大阪府、長野県、岐阜県に限られる。このように、CEJCは出身地の違いが言語に与える影響を定量的に比較するには適したコーパスとは言えない。一方で、話者の出身地が首都圏に集中していることから、方言差が生じやすい言語現象を扱う際には、出身地を東京都や首都圏に限定しても、一定規模のデータを確保することが可能である。

4.3 話者間の関係性毎のデータ規模

CEJCでは話者間の関係性の情報として、(1) 会話中の全ての話者間の関係性

の組合せ、(2) 一意に特定できる場合に限定した話し手から見た聞き手との関係性、(3) 協力者から見た会話相手との関係性の3種類を提供している。

4.3.1 会話中の全ての話者間の関係性の組合せ

会話中の全ての話者間の関係性を、家族、親戚、友人知人、同僚、元同僚、仕事関係(取引先など)、先生生徒、サービス場面関係(店員と客など)の水準で整理した。例えば図10のケースAは全ての話者間の関係性が家族と1つに定まるが、ケースBは友人知人・先生生徒といったように、複数の関係性の組合せで表現される。この関係性の情報は中納言の検索結果にも表示される。

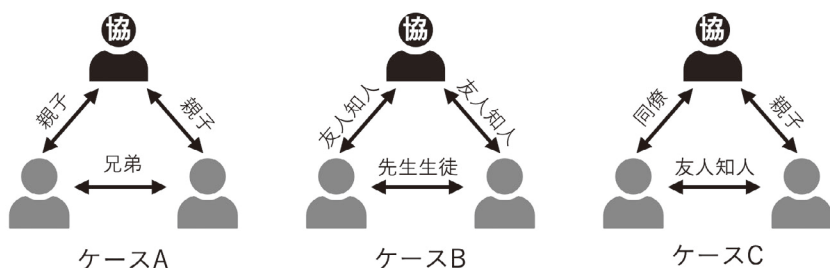


図10 話者間の関係性の例

ケースAのように1つの関係性で表現できるものは、577会話中489件と全体の85%を占める。2つ、3つ、4つの組合せはそれぞれ71件、16件、1件であり、5つ以上の組合せはない。1つの関係性で表現されるもののうち家族と友人知人の会話がいずれも145件、同僚間の会話も50件と少なからずある。定量的分析において、話者間の関係性を統制したり関係性間の比較を行う場合には、家族、友人知人、同僚などを取り上げると一定規模を確保できる。

4.3.2 話し手から見た聞き手との関係性

CEJCでは、話し手から見た聞き手との関係性について、関係が一意に定まる場合に限定してその情報を提供している。表4に示す通り、聞き手との関係性の情報を、大分類、中分類、小分類に分けて整理している。

表4 話し手から見た聞き手との関係性：大分類・中分類・小分類

大分類	中分類	小分類
家族	父母	父、母
	子供	息子、娘
	配偶者	夫、妻
	兄弟姉妹	兄、弟、姉、妹
	祖父母	祖父、祖母
	孫	孫息子、孫娘
家族親戚	義父母	義父、義母
	義理の子供	義息子、義娘
友人知人	-	-
仕事	同僚	-
	取引先など他社の人	-
	サービス提供者	-
	客	-
	その他	-
先生生徒	先生	-
	生徒	-
初対面	-	-

例えば図10のケースAの場合、大分類では、どの話者から見ても聞き手との関係性は家族と一意に定まるが、中分類まで特定できるのは協力者から見た関係性のみである。ケースCのように関係性が全く定まらないこともある。表5に、聞き手との関係性・話し手の年齢層毎の延べ話者数を示す。

表5 聞き手との関係性・話し手の年齢層毎の延べ話者数

話し手から見た聞き手との関係性（中分類）	話し手の年齢									計
	～9歳	10代	20代	30代	40代	50代	60代	70歳～	不明	
父母	5	9	18	7	8	3	0	0	0	50
子供	0	0	0	3	7	12	7	5	0	34
兄弟姉妹	0	0	3	3	2	3	8	2	0	21
配偶者	0	0	0	17	8	13	24	3	1	66
祖父母	0	0	2	1	0	0	1	0	0	4
孫	0	1	0	0	0	0	0	5	0	6
義父母	0	0	0	0	1	1	0	0	0	2
義理の子供（義子）	0	0	0	0	0	0	1	1	0	2
友人知人	0	12	128	90	59	57	56	24	1	427
同僚	0	0	29	71	45	48	25	10	7	235
取引先など他社の人	0	0	0	8	10	13	5	0	1	37
サービス提供者	0	0	1	0	3	2	4	0	3	13
客	0	0	0	4	6	3	1	2	102	118
その他の仕事関係者	0	0	4	14	4	2	8	0	0	32
先生	0	2	6	0	0	1	1	0	1	11
生徒	0	0	1	3	2	4	6	0	0	16
初対面	0	0	0	0	0	0	0	0	1	1
不定	33	93	54	63	95	110	72	70	10	600

中分類で見た場合、話者1675人のうち64%に相当する1075人について関係性が一意に特定できる。大分類で見た場合は87%が特定できる。

なお、本節で取り上げた聞き手との関係性の情報は、コーパス検索アプリケーション「中納言」では提供しておらず、CEJCのウェブサイトからダウンロードすることができる⁶。中納言の結果と合わせて分析する場合、検索結果をダウンロードした上で、中納言のセッションID、話者IDと、関係性データのsessionID、informantIDをキーに結合する必要がある。

4.3.3 協力者から見た会話相手との関係性

CEJCでは、個人密着法で収録した会話について、協力者から見た会話相手

6| https://www2.ninjal.ac.jp/conversation/cejc/data_relation.html

との関係性の情報も提供しており、中納言の検索結果にも表示される。この情報を活用することで、個々の協力者に注目し、相手や場面によってどのように言葉を使い分けしているかを調べるができる。4.1節で言及したように、協力者の語数はほかの人よりも多いことから、こうした分析が可能となる。

4.4 会話の属性毎のデータ規模

会話の属性として、会話形式、話者数、会話された場所、会話中の活動に着目し、会話行動調査の結果と比較しながらデータの規模をまとめる。

4.4.1 会話形式

図11に、会話形式毎の会話数と会話時間の内訳を、会話行動調査の結果と合わせて示す。

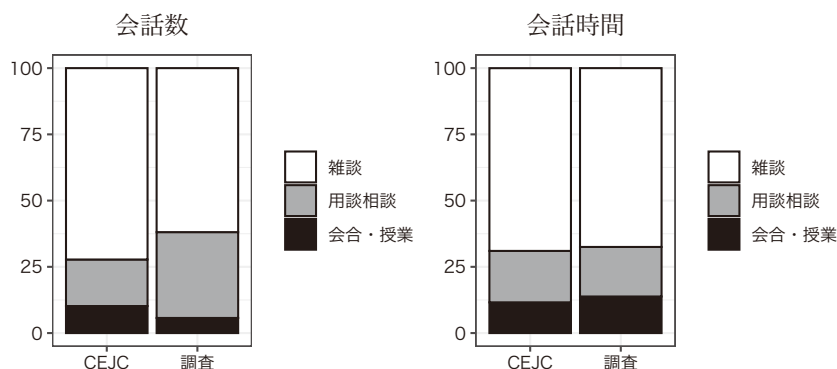


図11 会話形式：CEJCと会話行動調査の比較

会話数、会話時間のいずれにおいても雑談は70%前後を占めているが、用談相談や会合・授業も、会話数で18% (81件)、10% (47件)、時間で見ても20% (39時間)、12% (19時間) と、一定数含まれていることが分かる⁷⁾。

会話数で見ると会話時間を見た方が両者の傾向が類似しているのは、会話形式の種類によって収録時間の長短に特徴があるためである。例えば、会話行動調査から用談相談は5分未満の短いものも多いことが分かっているが、CEJCではこうした短い会話が収録されることは稀であり長めの用談相談が多

7) 会合・授業のうち授業は6会話、4.2時間と少なく、会議・会合の方が多。

いことから、件数は少ないが時間としては会話行動調査と同規模となるように配慮した。会合・授業は逆の傾向を示す。実際の会合・授業は1時間を越える長いものが少なからずあるが、コーパスに含める会話の多様性を確保するために1会話最大1時間程度という制限を設けたため、会話行動調査と同規模の時間となるよう、件数を少し多めに採用した。

4.4.2 会話の話者数

会話の話者数毎の結果を図12に示す。会話数、会話時間のいずれの場合においても、2人会話が50%弱で、3人以上の会話が多く含まれていることが分かる。2節で紹介したように、『名大会話コーパス』や『日本語話題別会話コーパス』など、現在公開されている会話コーパスは2人会話が多い。CEJJCは、会話577件中、4人以上の会話が32%（4人：92件、5人：29件、6人：27件、7人：12件、8人以上：26件）と、人数の多い会話も多く含まれている。また、人数の多い会話は必ずしも会合・授業だけでなく、親族の集まりや友人同士の懇親会などの雑談もある。日常生活の多様な場面の会話が収められていることが分かる。

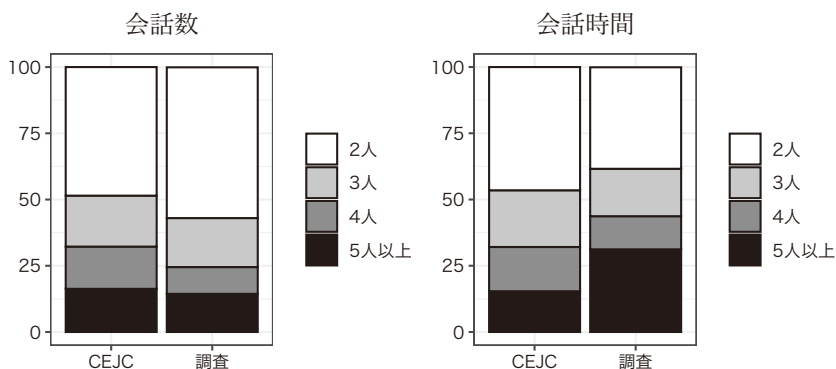


図12 会話の話者数：CEJJCと会話行動調査の比較

4.4.3 会話の行われた場所・会話中に行われた活動

会話の行われた場所と会話中に行われた活動（どのような活動中の会話か）は、相互に関係することから、結果を図13と図14にまとめて示す。

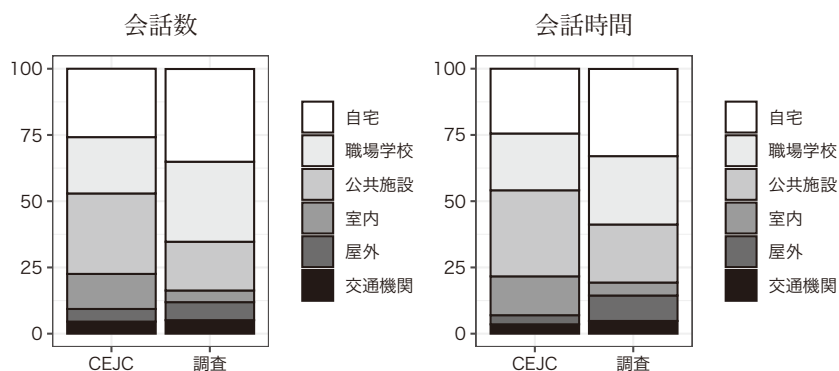


図13 会話の行われた場所：CEJCと会話行動調査の比較

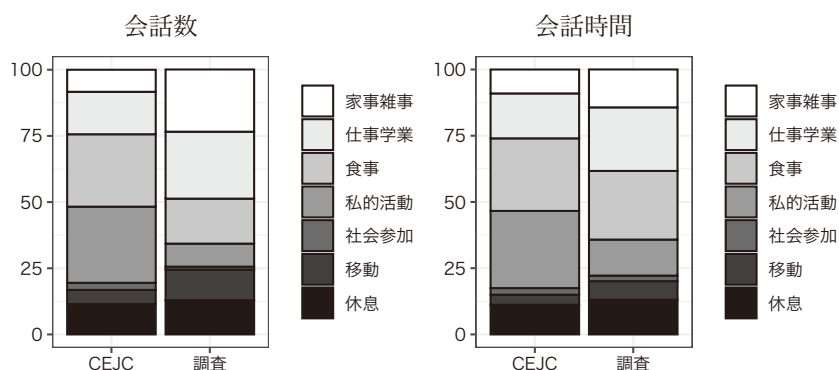


図14 会話中に行われた活動：CEJCと会話行動調査の比較

会話行動調査と比べると、自宅・職場での家事雑事・仕事中の会話が若干少なく、飲食店などの公共商業施設や友人宅・実家などの室内での私的活動中の会話が多い傾向が見られる。先述の通り、個人密着法での収録状況をモニター版で検証し、職場での仕事中の会話が会話行動調査よりも少ない傾向にあったことから、特定場面法において仕事中の会議会合約10時間を収録した。この増補によりかなりの改善が見られたが、会話行動調査と同水準となるには致らなかった。

このように私的活動中の会話が多いものの、食事の準備や片付け、棚の組み立てなどの家事雑事、ボランティアなどの社会参加、旅行などのレジャー活動、

公園など屋外の散策、車や交通機関での移動など、ほかのコーパスにはあまり見られない、日常における多様な場面の会話が収録されている。

□5 おわりに

本章では、話し言葉を対象とする代表的なコーパスを取り上げ、その設計に焦点を当てて紹介してきた。いずれのコーパスも、話者や場面の選定、収録方法などにおいて様々な工夫を凝らしており、話し言葉の実態に迫るための設計思想が随所に見られた。コーパスの目的に応じて、均衡性を指向したものや特定の会話類型に絞ったもの、教育・研究の双方に資する設計を目指したものなど、その多様性は今後のコーパス構築にとっても示唆的である。

その上で、本章後半では『日本語日常会話コーパス』を取り上げ、会話行動調査に基づく収録設計、協力者主体の個人密着法に基づく収録、多層的なアノテーションといった特徴を概観した。CEJCは、日常会話の多様性を捉えた大規模言語資源として、話し言葉研究の今後を拓く基盤となるだろう。

参考文献

- ISO 24617-2 (2012) Language Resource Management — Semantic Annotation Framework (SemAF) — Part 2: Dialogue Acts.
- 浅原正幸・松本裕治 (2018) 『『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション』『自然言語処理』25 (4): 331–356.
- 浅原正幸・若狭絢 (2022) 『『日本語日常会話コーパス』に対する係り受け情報アノテーション』『言語処理学会第28回年次大会予稿集』1699–1703. https://www.anlp.jp/proceedings/annual_meeting/2022/pdf_dir/PT4-2.pdf
- Aston, G. & Burnard, L. (1998) *The BNC Handbook*. Edinburgh University Press.
- 藤村逸子・大曾美恵子・大島ディヴィッド義和 (2011) 「会話コーパスの構築によるコミュニケーション研究」藤村逸子・滝沢直宏 (編) 『言語研究の技法—データの収集と分析—』43–72, ひつじ書房.
- 現代日本語研究会 (編) (1998) 『女性のことば・職場編』ひつじ書房.
- 現代日本語研究会 (編) (2002) 『男性のことば・職場編』ひつじ書房.
- 現代日本語研究会 (編) (2011) 『合本女性のことば・男性のことば(職場編)』ひつじ書房.
- 五十嵐陽介・菊池英明・前川喜久雄 (2006) 「韻律情報」『日本語話し言葉コーパスの構築法 (国立国語研究所報告 124)』347–453. <http://dx.doi.org/10.15084/00001357>
- Iseki, Y., Kadota, K. & Den, Y. (2019) Characteristics of everyday conversation derived from the analysis of dialog act annotation. *Proceedings of the 22nd Oriental COCOSDA*, 1–6.
- 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. <http://dx.doi.org/10.15084/00000810>
- 小磯花絵・伝康晴 (2018) 『『日本語日常会話コーパス』データ公開方針—法的・倫理的な観点からの検討を踏まえて—』『国立国語研究所論集』15: 75–89. <http://dx.doi.org/10.15084/00000810>

[org/10.15084/00001518](http://dx.doi.org/10.15084/00001518)

- 小磯花絵・天谷晴香・居關友里子・白田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2020a)『『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』18: 17-33. <http://dx.doi.org/10.15084/00002540>
- 小磯花絵・菊池英明・山田高明 (2020b)『『日本語日常会話コーパス』への韻律ラベリング・ラベリングの設計と日常会話の韻律の特徴一』『人工知能学会研究会資料』SIG-SLUD-B90: 34-39. http://dx.doi.org/10.11517/jsaislud.88.0_08
- 小磯花絵・天谷晴香・石本祐一・居關友里子・白田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 (2022)『『日本語日常会話コーパス』—設計・構築・特徴一(日常会話コーパスプロジェクト報告書 6)』国立国語研究所. <https://www2.ninjal.ac.jp/conversation/report/report06.pdf>
- 小磯花絵・天谷晴香・居關友里子・白田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 (2023)『『日本語日常会話コーパス』設計と構築』『国立国語研究所論集』24: 153-168. <http://dx.doi.org/10.15084/00003692>
- 国立国語研究所 (1955)『談話語の実態(国立国語研究所報告 8)』秀英出版.
- 前川喜久雄 (2013)「コーパスの存在意義」前川喜久雄(編)『コーパス入門(講座日本語コーパス 1)』1-31, 朝倉書店.
- 丸山岳彦・柏野和佳子 (2014)「サンプリング」山崎誠(編)『書き言葉コーパス—設計と構築一(講座日本語コーパス 2)』22-44, 朝倉書店.
- 丸山岳彦・小磯花絵・西川賢哉 (2022)『『昭和話し言葉コーパス』の設計と構築』『国立国語研究所論集』22: 197-221. <http://dx.doi.org/10.15084/00003522>
- 中俣尚己(編) (2023)『話題別コーパスが拓く日本語教育と日本語学』ひつじ書房.
- 中俣尚己・太田陽子・加藤恵梨・澤田浩子・清水由貴子・森篤嗣 (2021)「日本語話題別会話コーパス J-TOCC」『計量国語学』33 (1): 205-213. http://dx.doi.org/10.24701/mathling.33.1_11
- 日本放送協会 (2011)『2010年国民生活時間調査報告書』NHK放送文化研究所.
- 西川賢哉・渡邊友香 (2020)『『日本語日常会話コーパス』に対する短単位情報付与—作業工程と評価一』『言語資源活用ワークショップ2020発表論文集』324-330. <http://dx.doi.org/10.15084/00003172>
- 大村舞・若狭絢・松田寛・浅原正幸 (2025)「日本語日常会話コーパスの Universal Dependencies: UD_Japanese-CEJC」『自然言語処理』32 (1): 55-90.
- Svartvik, J. (ed.) (1990) *The London-Lund corpus of spoken English: Description and research*. Lund University Press.
- 田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2018a)『『日本語日常会話コーパス』の構築—会話収録法に着目して—』『国立国語研究所論集』14: 275-292. <http://dx.doi.org/10.15084/00001424>
- 田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2018b)『『日本語日常会話コーパス』の構築—個人密着法に基づく会話の収録一(日常会話コーパスプロジェクト報告書 2)』国立国語研究所. <https://www2.ninjal.ac.jp/conversation/report/report02.pdf>
- 山口昌也 (2018)『『日常会話コーパス』活用環境の構築』『言語資源活用ワークショップ 2018 発表論文集』340-347. <http://dx.doi.org/10.15084/00001668>
- 山内博之(編) (2013)『実践日本語教育スタンダード』ひつじ書房.
- 山内博之 (2018)「話題による日本語教育の見取り図」岩田一成(編)『語から始まる教材作り(現場に役立つ日本語教育研究6)』3-16, くろしお出版.

山崎誠・前川喜久雄 (2014)「コーパスの設計」山崎誠 (編)『書き言葉コーパス—設計と構築— (講座日本語コーパス 2)』1-21, 朝倉書店.

関連 URL (2025年7月31日確認)

British National Corpus (BNC)

<http://www.natcorp.ox.ac.uk/>

BTSJ 1000 人日本語自然会話コーパス

<https://mmsrv.ninjal.ac.jp/btsj/corpus.html>

https://isplad.jp/btsj_corpus/

現代日本語書き言葉均衡コーパス (BCCWJ)

<https://clrd.ninjal.ac.jp/bccwj/>

現日研・職場談話コーパス

<https://www2.ninjal.ac.jp/conversation/shokuba.html>

コーパス検索アプリケーション「中納言」

<https://chunagon.ninjal.ac.jp/auth/login>

名大会話コーパス

<https://mmsrv.ninjal.ac.jp/nucc/>

<https://www2.ninjal.ac.jp/conversation/nuc.html>

日本語話し言葉コーパス (CSJ)

<https://clrd.ninjal.ac.jp/csj/>

日本語日常会話コーパス (CEJC)

<https://www2.ninjal.ac.jp/conversation/cejc.html>

日本語話題別会話コーパス (J-TOCC)

<http://nakamata.info/database/>

http://nakamata.info/database/j-tocc_document.pdf

自然会話リソースバンク (NCRB)

<https://isplad.jp/ncrb/>

昭和話し言葉コーパス (SSC)

<https://www2.ninjal.ac.jp/conversation/showaCorpus/>

The Survey of English Usage

<https://www.ucl.ac.uk/english/research/survey-english-usage/>

(小磯花絵)

