国立国語研究所学術情報リポジトリ

話し言葉コーパスの構築と公開

著者	小磯 花絵
雑誌名	電子情報通信学会誌
巻	102
号	6
ページ	554-557
発行年	2019-06
URL	http://doi.org/10.15084/00003012

話し言葉コーパスの構築と公開

Construction and Publication of Spoken Corpus

小磯花絵

1. 話し言葉データの蓄積・文字化データの公開

話し言葉の研究は内省が効かないこともあり，音声を録音・文字化して蓄積し，それを研究に活用するという方法論は古くからとられてきた．例えば国立国語研究所では，東京における日常談話の収録を 1952 年に開始している．比較のために収集されたニュースや講義などの独話と合わせ，音声を文字化した上で語・文節・文・イントネーションなどの情報を付与し，日常談話の特徴を定量的に解明する研究が進められた[1]．1960 年，1963 年には，拡張した話し言葉データに基づき，話し言葉の総合文型を目指す報告書が刊行されており[2,3]，話し言葉データに基づく記述研究から理論研究への展開が図られた．正に現在で言うところのコーパスに基づく研究だが，残念ながら収集された音声データや文字化資料・各種研究用付加情報は公開されることはなかった．この時代，人文系の研究のために集められた多くの話し言葉のデータが，同じ運命をたどったものと思われる．

人文系の研究のために集められた話し言葉データが公開されるようになったのは 1990 年代に入ってからである．計算機の普及が背景にあったことは間違いない．

例えば「女性のことば・職場編」[4]には，1993 年に集められた職場での朝・会議・休憩時の女性の談話の文字化資料（電子版）が，10 編の研究論文とともに収められている．その 4 年後には，1999～2000 年の男性を対象とする職場談話をまとめた「男性のことば・職場編」[5]も刊行され，現在は両者合本の形で再公開されている．また 2000 年代前半には約 100 時間の雑談を収めた「名大会話コーパス」が公開された[6]．現在公開されている会話コーパスとしては最大規模のものであり，日本語学や日本語教育の分野などでこのコーパスを用いた研究が数多くなされている．

しかし残念ながら，職場談話・名大会話コーパス共に音声データは含まれていない．話し言葉を総合的に研究するには，音声は欠かすことのできない重要なデータであるが，職場談話や雑談に含まれる膨大な個人情報の問題もあり，音声の公開に至らなかったケースが多い．

2.「日本語話し言葉コーパス」

音声まで含めコーパスとして公開する動きは，「千葉大学地図課題コーパス」[7]や「日本語話し言葉コーパス」[8]，「千葉大学 3 人会話コーパス」[9]など，文理融合のプロジェクトを中心に進められてきた．ここでは筆者が構築に携った「日本語話し言葉コーパス」を取り上げ，その設計について簡単に紹介する．

「日本語話し言葉コーパス」（CSJ）は，自発性の高い独話を中心とする 660 時間規模のコーパスであり，2004 年に公開された．音声認識や要約などの情報工学と，言語学など人文系の研究の推進を目指し，国立国語研究所，情報通信研究機構（旧通信総合研究所），東京工業大学が共同して構築したコーパスである．

種々の学会における実際の研究発表（学会講演）と，一般話者による主に個人的な体験談等に関する 10～15 分程度のスピーチ（模擬講演）がその中心を占めるが，比較のためにインタビュー・課題試行対話などの会話データや朗読音声も若干含まれている．

660 時間という規模も他の話し言葉コーパスを圧倒するが，人文学の研究からすると，CSJ の最大の特徴はコアと呼ばれる約 45 時間のデータ範囲に集中して付与されている豊富なアノテーション情報であろう．図 1 に，コーパス全体及びコアに対して付されたアノテーション情報を示す．形態論（単語）情報や係り受け情報などの統語・形態論的な情報だけでなく，分節音情報や韻律情報，また発話に対する聞き手の主観的印象を尺度化して

小磯花絵　国立国語研究所音声言語研究領域
E-mail koiso@ninjal.ac.jp
Hanae KOISO, Nonmember (Spoken Language Division, The National Institute for Japanese Language and Linguistics, Tachikawa-shi, 191-8561 Japan).
電子情報通信学会誌 Vol.102 No.6 pp.554-557 2019 年 6 月
©電子情報通信学会 2019

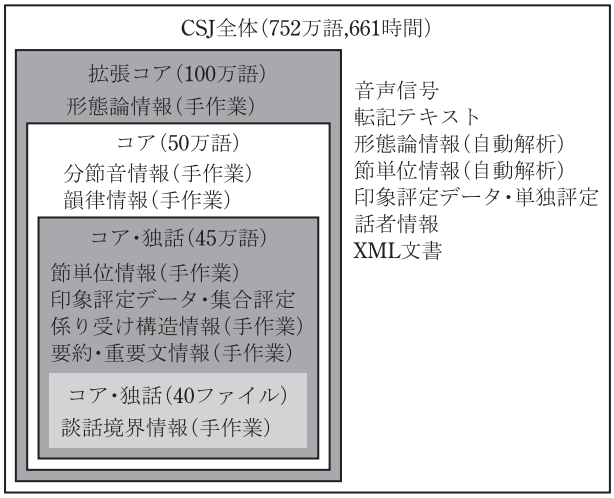

図1 「日本語話し言葉コーパス」のアノテーション情報

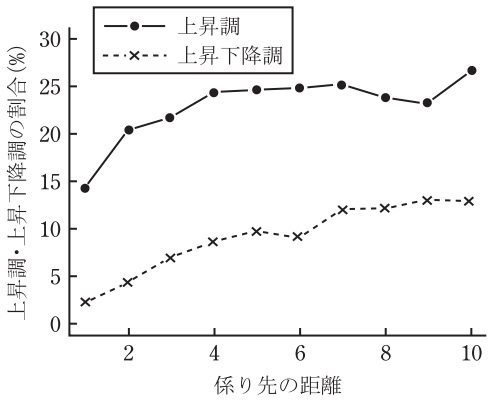

図2 係り先の距離ごとの上昇調・上昇下降調の出現率

表現した印象評定データなど，多岐にわたるアノテーションが含まれている．

こうしたアノテーションは，相互に関連付けることで多様な研究が可能となる．CSJ では，個々のアノテーションファイルに加え，各種アノテーションを階層関係により統合した XML ファイルが提供されている．しかし統語・形態論関連の情報と音声・韻律関連の情報は必ずしも階層的表現に適しているとは言えず，利便性に問題があった．そこで 2012 年には多層的アノテーションの表現方法として主流となっているスタンドオフ形式による表現を採用した RDB も構築・公開している．

複数のアノテーションを活用した人文系の研究を一例紹介する．韻律・節単位・係り受け情報を利用することで，句末に上昇を伴う音調（上昇調や上昇下降調など）が，切れ目の大きな節により多く見られること，また「漱石の小説」のように直後の文節に係るとき（距離 1）よりも，「漱石の新聞に連載された小説」（距離 3）のように先の文節に係る方が，当該文節（この場合は「漱石の」）の句末に上昇成分がより多く見られること（図 2）が分かっている[10]．こうした研究は，CSJ のように豊富なアノテーションがあることで実現した研究である．

3. 日常会話・映像データの公開

音声データ・アノテーション情報を公開し人文学の話し言葉研究を大きく推進した CSJ だが，独話が中心であり，会話としてはインタビューや課題指向対話などが若干数含まれるだけである．人文学の研究課題の一つは，日常生活を営む中で我々がいかなる言語行動を取っているかを明らかにすることである．これまでにも会話コーパスは少なからず構築・公開されているが，雑談を集めたコーパスも含めその多くは収録のために集められた状況での会話を扱っている．

そこで国立国語研究所では，筆者が携る共同研究プロジェクトにおいて，200 時間規模の「日本語日常会話コーパス」（CEJC）の構築を進めており，その一部に当たる 50 時間の会話を映像を含めて 2018 年 12 月にモニタ公開したところである[11]．本章では CEJC の設計とモニタ公開データの特徴について紹介する．

3.1 設計

CEJC では，①日常場面で自然に生じる会話を対象とすること，②多様な場面・多様な話者の会話をバランス良く集めること，③音声だけでなく映像まで含めて収録・公開すること，④研究に必要な各種アノテーション情報を施すことによって，会話行動を多角的に解明するための研究環境を提供することを目指している．本節ではこれらに焦点を当て CEJC の設計を概観する．

3.1.1 CEJC の収録法

①，②を実現するために，性別・年齢などのバランスを考慮して選別された協力者 40 名に収録機材を貸し出し，協力者の日常生活の中で自然に生じる会話を協力者自身に記録してもらうという方法を中心に会話を集めている．その際，許諾が得られる範囲で，できるだけ多様な場面・人との会話を収録するよう依頼している．例えば，30 代専業主婦の場合，家族との会話であっても，食事場面や子供の宿題を見る場面，旅行先での散策場面，夫の家族を混じえた場面の会話などを収録している．

図 3 に 30 代男性が収録した妻・義母との会話場面の映像サンプルを示す．このサンプルにあるように，最大 3 台のカメラで会話を収録し，個々の映像及び図 3 に示すような合成した映像を公開する．音声については，原則として全ての話者が IC レコーダを装着して当該話者の音声を中心に記録すると同時に，会話の場の中央に配置した IC レコーダで会話全体の音声を記録しており，全ての音源を公開する．

図3　30代男性が収録した妻・義母との会話の映像サンプル　　掲載用に話者の顔にぼかしを入れている．話者が首に下げたフォルダにICレコーダが入っており，当該話者の音声を中心に記録．

3.1.2　データ公開方針の策定

映像・音声データの公開については，慎重な検討が必要となる．会話の収録・公開に関する同意書では，音声・映像・転記テキスト等に記録された名前・自宅住所等・所属組織の名称・住所等は伏せるが，話者の顔にぼかし処理は加えないこととしている．この条件に全ての話者が同意した場合に収録がなされる．よって同意を得た話者の扱いについてはこの通り対応すればよいということになるが，実際はそう単純でもない．複数の情報を連結することで，あるいは自宅外観などの映像から，個人や自宅（付近）が特定され兼ねないケースもある．協力者は複数の会話を収録するためその危険性が高い．そのためヒアリングを通して協力者にどの情報まで出してよいか，どの情報は伏せるべきかを確認している．

また，公開の同意を得ていない第三者の顔や声，テレビ画面やBGMなどの著作物の写り込みの扱いも問題となる．準備研究の段階から，知財関連を専門とする法律家と肖像権や個人情報保護，著作権などの観点から相談を重ね，データの収録・整備・公開の方針の大枠を定めた．その上で，実際の収録データを元に具体的な問題を洗い出し，対応策を検討して方針をまとめた．ここで定めた公開方針やその判断に至る根拠などについては，早い段階から一般に公開している[12]．これまで，個人情報等の扱いの問題で音声データの公開に至らなかったコーパスが多いことを考えると，こうした知見を蓄積・共有することは，人文学データのオープン化を促進する上で極めて重要である．

3.1.3　アノテーション

CSJと同様，全体に対して自動解析を主とするアノテーションを施すと同時に，コーパスの一部（コア20時間）に対しては詳細なアノテーション情報を人手を介して高精度に付与する．全体に対し自動で付与するのは，短単位情報，長単位情報，文節情報，係り受け情報である．談話行為情報と韻律情報などはコアにのみ人手で付与する．またメタ情報として，会話や話者，話者間の関係性などの情報も公開する．

3.2　CEJCモニタ公開版の概要

コーパスの利用可能性や問題などを把握するために，全体の四分の一に相当する50時間分の会話データのモニタ公開を2018年12月に開始した[11]．本節ではこのデータセットについて概説する．コーパス全体200時間の規模などもここから推定できる．

公開方式としては，①映像・音声データ，転記，短単位，各種メタ情報，検索ツールを収めたハードディスクでの公開と，②形態論情報（短単位情報）をオンラインで検索できる「中納言」（国語研究所が提供する各種コーパス用の検索システム）がある．

CEJC全体では40名の協力者に収録を依頼しているが，モニタ版では半分に相当する20名を対象に，平均2.5時間，計50時間の会話が対象となっている．協力者は，5世代（20代，30代，40代，50代，60代以上）×男女×各2名ずつである．ただし収録の進捗の都合で，40代女性が3名，60代以上の女性が1名となっている．

話者数は，延べ392名，異なり237名である．ここからコーパス全体としては延べ約1,600名，異なり約950名程度となることが推定される．延べ人数の場合の年齢・性別の内訳を図4に示す．40代50代の女性が多めで60代以上の女性が少なめなのは，正にこの世代の女性協力者の偏りによる．CEJC全体では協力者の年齢・

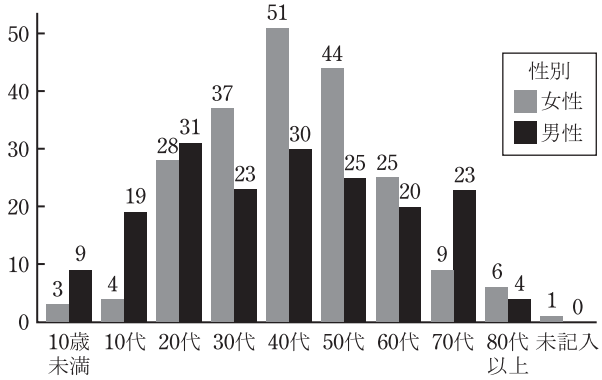

図 4　CEJC モニタ版の話者の年齢・性別の分布

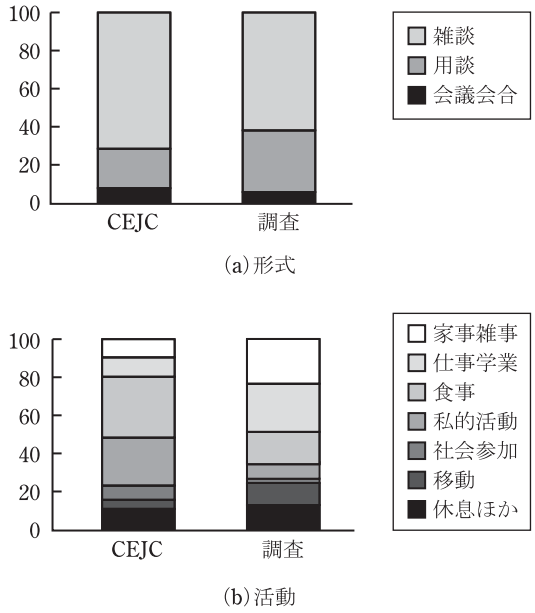

図 5　モニタ版と会話行動調査における会話形式・活動の内訳

性別をバランスさせるため，こうした偏りは解消されることが期待できる．未成年者の数が少ないのは，協力者を成人に限定したためである．

CEJC では多様な会話をバランス良く集めるために，予備調査として我々の普段の会話行動について調査している[13]．図 5 に，モニタ公開版を対象に会話の形式と活動の内訳を求め，行動調査の結果と比較した結果を示す．会話の形式としては，雑談が若干多く用談が少ない傾向ではあるが，調査結果と類似した比率となっている．一方，活動については，コーパスでは仕事中の会話が少ない．職場での収録が難しいことによる．

未成年者の会話，また仕事中の会話といったように，不足する種類の会話については，3.1.1 に示した方法とは別の収録法で補うことも予定している．

CEJC はまだ公開して間もないため研究事例は多くないが，例えば相手との関係性に応じて，同じ人が丁寧体・普通体などの話体をいかに使い分けているか，声の高さなどがいかに変化するか，などの研究が進められている．また指示表現と指差しとの関係など，映像を活用した研究も行われている．

4. 過去に収録された話し言葉データの公開

1. で述べたように，過去に収録された話し言葉データが公開されず埋もれたままになっていることも多い．国語研究所では，研究所の資料庫に眠っているこうした音源や映像をディジタル化して視聴できる環境を整え，現在では内館利用できるようになっている．その一部についてはコーパスとして整備し一般に公開する予定である[14]．また全国各地の昔の方言音声も多い．失われつつある方言音声を公開することは社会的に見ても意義のあることである．1. で取り上げた職場談話データ・名大会話コーパスなどを，関係者の許諾を得た上で再整備し，国語研究所が提供するオンライン検索システム「中納言」で公開するという活動もしている．このように，蓄積されてきた話し言葉を再整備・公開して研究に利活用できる環境を整えることも重要な課題である．

文　　献

（1） 国立国語研究所，談話語の実態，秀英出版，東京，1955.
（2） 国立国語研究所，話し言葉の文型 I，秀英出版，東京，1960.
（3） 国立国語研究所，話し言葉の文型 II，秀英出版，東京，1963.
（4） 現代日本語研究会編，女性のことば・職場編，ひつじ書房，東京，1998.
（5） 現代日本語研究会編，男性のことば・職場編，ひつじ書房，東京，2002.
（6） https://mmsrv.ninjal.ac.jp/nucc/
（7） 堀内靖雄，中野有紀子，小磯花絵，石崎雅人，鈴木浩之，岡田美智男，仲真紀子，土屋　俊，市川　熹，“日本語地図課題対話コーパスの設計と特徴，”人工知能誌，vol. 14, no. 2, pp. 261-272, March 1999.
（8） 国立国語研究所，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，東京，2006.
（9） http://research.nii.ac.jp/src/Chiba3Party.html
（10） 小磯花絵，“日本語自発音声における複合境界音調と統語構造との関係，”音声研究，vol. 18, no. 1, pp. 57-69, April 2014.
（11） https://pj.ninjal.ac.jp/conversation/cejc-monitor.html
（12） 小磯花絵，伝　康晴，“『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―，”国語研論集，no. 15, pp. 75-89, July 2018.
（13） 小磯花絵，土屋智行，渡部涼子，横森大輔，相澤正夫，伝　康晴，“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，”国語研論集，no. 10, pp. 85-106, Jan. 2016.
（14） 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―，”専修大人科研月報，no. 282, pp. 39-55, 2016.

（2019 年 1 月 3 日受付　2019 年 1 月 18 日最終受付）

小磯　花絵（こいそ　はなえ）

1994 千葉大・文・行動科学卒．1996 同大学院修士課程了．1998 奈良先端大情報科学研究科博士後期課程了．同年，国立国語研究所．現在，同音声言語研究領域代表，教授．話し言葉のコーパス研究に従事．編著「話し言葉コーパス―設計と構築―」（朝倉書店）．