『日常会話コーパス』プロジェクト ： コーパスに基づく話し言葉の多角的研究

著者	小磯 花絵
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	114-119
発行年	2017
URL	http://doi.org/10.15084/00001464

『日常会話コーパス』プロジェクト —コーパスに基づく話し言葉の多角的研究—

小磯 花絵（国立国語研究所研究系音声言語研究領域）*

Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation

Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所では，2016 年 4 月から「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」プロジェクトを開始した。このプロジェクトでは，さまざまなタイプの日常会話 200 時間をバランス良く収録した大規模な日常会話コーパスを構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性を，レジスター・相互行為・経年変化の観点から多角的に解明することを目指す。本発表では，プロジェクト全体で推進する研究，およびそのために整備・公開する複数の言語資源の全体像について触れた上で，本プロジェクトの中核を占める『日本語日常会話コーパス』を取り上げ，コーパスの設計について報告する。

1. はじめに

日常会話は社会生活の基盤であり，日常の話し言葉の特徴や仕組み，日常生活を円滑にするための会話コミュニケーションの有様を解明することが求められている。こうした研究を支えるものとして日常会話を広く収集したコーパスの構築は急務である。

海外では，Quirk により 1959 年に開始された The Survey of English Usage 計画において，書き言葉だけでなく話し言葉が大規模に収録され，それに基づく記述文法書が作成されている。その後も，British National Corpus (BNC) や Bank of English，The Santa Barbara Corpus of Spoken American English など，会話を含む話し言葉を収録した大規模なコーパスが数多く構築され，言語学的な研究だけでなく，会話コミュニケーション研究など多様な研究が推進されている。

日本においても，1950 年代から国語研究所において日常会話を含む話し言葉の収録とそれに基づく実証的な話し言葉研究が始まり，『談話語の実態』(国立国語研究所 1955) や『話しことばの文型 (1)(2)』(国立国語研究所 1960, 1963) といった研究報告書がまとめられた。『談話語の実態』は，The Survey of English Usage 計画が始まる 4 年前に刊行されており，先進的な研究であったと言える。しかし残念ながら，収集された音声資料は公開されず，またその後も日本国内においては長らく話し言葉コーパスの構築・公開は行われてこなかった。1990 年代以降，種々の話し言葉コーパスが公開されるようになったが，特定の場面や話者層に偏った

* koiso@ninjal.ac.jp

ものが多く，BNC のように均衡性に配慮して設計されたコーパスは作られてこなかった。

このような状況を受け，国立国語研究所では，2016 年度より機関拠点型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」(2016〜2021 年度）を開始した。このプロジェクトは，さまざまなタイプの日常会話 200 時間をバランス良く収録した大規模な日常会話コーパスを構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性を，「レジスター」「経年変化」「相互行為」の観点から多角的に解明することを目指すものである。

本稿では，プロジェクト全体で進める研究，およびその推進のために整備・公開する複数の言語資源について概説した上で，本プロジェクトの中核を占める『日本語日常会話コーパス』を取り上げ，コーパスの設計について報告する。

2. プロジェクトの構成

本プロジェクトでは，日常会話コーパスの構築を主導する班と，コーパスを用いて研究を推進し研究の観点からコーパスを評価する三つの研究班を組織している。

コーパス構築班　さまざまなタイプの日常会話 200 時間をバランス良く納めた大規模なコーパス『日本語日常会話コーパス』の構築・公開を主導する（班長：小磯花絵)。

レジスター班　日常会話に加え，講演などの独話や小説などの会話文をも含む多様なレジスターの話し言葉を比較し，語彙・文法・韻律などの特性を探る（班長：山崎誠)。

相互行為班　会話相互行為の中で文法が果たす役割やその特性・構造を，英語など日本語以外の会話との比較を通して総合的に分析する（班長：伝康晴)。

経年変化班　昭和期の話し言葉と現代の話し言葉を，アクセント・韻律・語彙・文法などの観点から比較し，話し言葉の経年変化過程を実証的に解明する（班長：丸山岳彦)。

3. プロジェクトで整備・公開する言語資源

国立国語研究所では，『日本語話し言葉コーパス（CSJ)』や『現代日本語書き言葉均衡コーパス（BCCWJ)』，『国語研日本語ウェブコーパス（NWJC)』など，大規模なコーパスの構築・公開を進めてきた（図 1)。特に，現代日本語の書き言葉の全体像を把握するために構築された 1 億語からなる BCCWJ やウェブを母集団とする 100 億語規模の NWJC の公開により，多様なレジスターを考慮した現代日本語書き言葉の研究をコーパス言語学の手法に基づき研究する環境が整備され，辞書編纂への活用や日本語学習者・日本語教師の利用など，基礎研究に留まらない広がりを見せている。

話し言葉については，CSJ の構築・公開により，話し言葉の言語学的・音声学的な研究や音声情報処理研究を支える基盤は整えられたと言えよう。しかし，CSJ は独話を主対象とするコーパスであり，日常生活の中で交わされる会話は含まれていない。我々は日常生活の中でどのような言葉を使い，人といかなる仕組みでコミュニケーションしているのか，また日常場面でのさまざまな活動を言葉や身体を用いていかに組織化しているのかなど，問うべき課題は多い。こうした研究を支える基盤として，実際の日常会話場面を対象とした大規模な会話コーパスの構築が求められている。本プロジェクトにおいてコーパス構築班が主導する『日本語日常

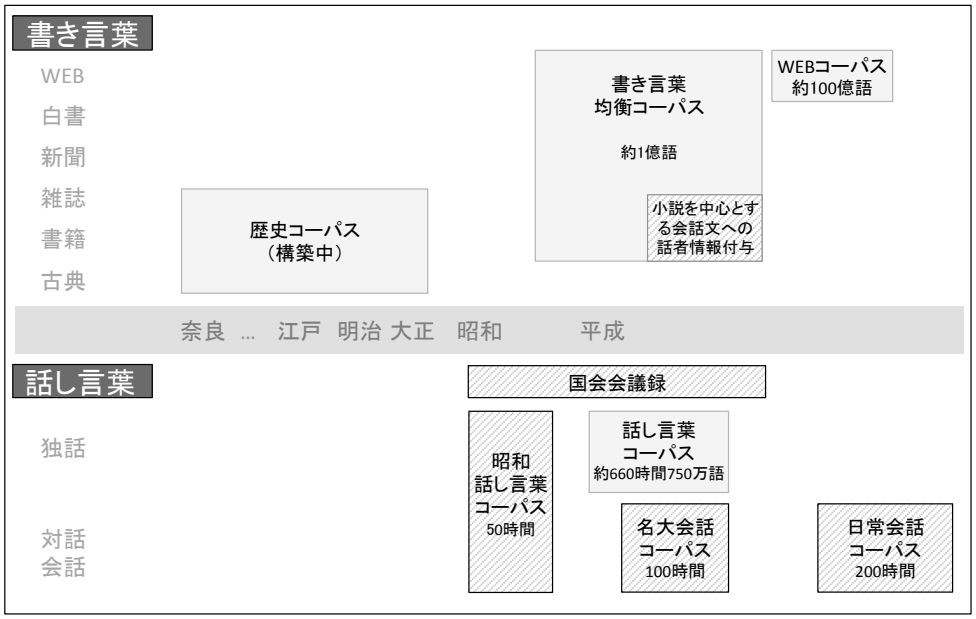

図 1　国立国語研究所で公開・構築中の主たる言語資源（斜線は本プロジェクトで構築する言語資源）

会話コーパス』は，まさにこうした状況を受けて計画したものである。

また書き言葉については，機関拠点型基幹研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」が構築を進める『日本語歴史コーパス』により，書き言葉の通時的な変化を研究する基盤も整いつつある。海外では，2006 年に The Diachronic Corpus of Present-Day Spoken English（DCPSE）という，The Survey of English Usage で 1960 年代後半から 1990 年代前半に録音されたイギリス英語の話し言葉を集めた通時コーパス（約 88 万語）が公開された(1)。こうした話し言葉の経年変化研究の基盤を整えるべく，本プロジェクトでは，経年変化班が中心となり，次の二つの言語資源の構築を計画している。

『昭和話し言葉コーパス』　先に言及した『談話語の実態』および『話しことばの文型 (1)(2)』のために 1950 年代から 1960 年代に録音された日常談話を対象にデータを整備し，『昭和話し言葉コーパス』として一般公開する (丸山 2016)。規模は会話・独話各 25 時間，計 50 時間を予定している。

『国会会議録』ひまわり検索版　国立国会図書館の許諾を得た上で，『国会会議録検索システム』に収録されている国会の会議録のうち 1947 年から 2012 年に開催された衆議院・参議院の本会議・予算委員会を対象に，全文検索システム『ひまわり』で検索できるよう整備する。話し言葉の経年変化研究を効率的に行えるよう，発言者の生年や肩書などの情報も付与し，2016 年 12 月に一般公開した。実際に経年変化研究などで利用しながら，データやシステムの改良を進める予定である。

また，日常会話を含む多様なレジスターの話し言葉を対象に，語彙や文法，韻律などの特性を分析するために，レジスター班を中心に次の三つの言語資源の整備を計画している。

(1) http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm

『名大会話コーパス』　科学研究費基盤研究 (B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」(2001〜2003 年度，研究代表者：大曽美恵子）で作成された，120 会話，計 100 時間の雑談を納めたコーパスである。収録時期は CSJ と重なる。現在は国立国語研究所に移管され，研究所のホームページから転記テキストがダウンロードできるようになっている。本プロジェクトでは，転記テキストを対象に，形態素解析用辞書 UniDic と形態素解析器 MeCab を用いて形態論情報（短単位）を自動付与し[(2)]，メタ情報として発話者の属性（性別・年齢・出生地など）と会話の情報（収録日・収録場所など）を整理した上で，オンライン検索システム『中納言』および全文検索システム『ひまわり』にて 2016 年 12 月に一般公開した (柏野ほか 2017) [(3)]。

BCCWJ **会話文発話者情報**　BCCWJ の書籍における会話文を特定し，発話者の属性情報（話者名・性別・年代）を付与する。2019 年度の公開を目指して作業を進めている (宮嵜ほか 2017) 。

『女性のことば・職場編』『男性のことば・職場編』　『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) として公開されている，職場会話の転記テキストを対象に形態論情報を付与し，全文検索システム『ひまわり』に搭載した。このデータについては，権利関係の都合で一般公開はできないが，出版社の許諾を得た上で，当該書籍を購入したプロジェクトメンバーに限定して利用している。

このように本プロジェクトでは，話し言葉の経年変化やレジスター的特性の研究を支える言語資源の開発を積極的に進め，権利関係の許す範囲で公開する。その際，UniDic に基づく短単位情報の付与や『中納言』・『ひまわり』での公開といったように，できるだけ同一の基準・同一の検索環境を整えることで，複数の言語資源にまたがる分析の利便性を高める。

4. 『日本語日常会話コーパス』の概要

本節では，プロジェクトの中核を占める『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC）を取り上げ，コーパスの設計について報告する。なお，小磯ほか (2017) でその詳細を述べたため，ここでは概要を記すに留める。

■基本方針　日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話を対象とする。幅広いレジスターをカバーするようサンプルを選ぶ。普段われわれがどのような種類の会話を行っているかを調査し (小磯ほか 2016)，その結果を参考に多様な種類の会話を納めたコーパスを構築する。

■規模　コーパスに納める会話の総時間を 200 時間に定める。これまでに収録・転記したデータから試算すると，全体で 200 万語程度になると予想される。

■収録法　日常会話をバランスよく収録するために，首都圏に在住の協力者 40〜50 人（男女 × 20 代・30 代・40 代・50 代・60 代以上 × 各 4〜5 人）に収録機材等を貸し出し，協力者自

(2) 形態論情報の一部については人手で修正を加えている。
(3) 『中納言』での公開については，国立国語研究所コーパス開発センターと共同で実施した。

表 1　プロジェクトで構築する言語資源の公開予定時期

言語資源	公開予定時期	補足
『女性のことば・男性のことば—職場編—』ひまわり版	2016 年 6 月（既公開）	プロジェクト内部限定
『国会会議録』ひまわり版	2016 年 12 月（既公開）	
『名大会話コーパス』中納言版・ひまわり版	2016 年 12 月（既公開）	
『日本語日常会話コーパス』50 時間分	2018 年度	モニター公開
『昭和話し言葉コーパス』テキストのみ	2018 年度	モニター公開
『現代日本語書き言葉均衡コーパス』話者情報	2019 年度	BCCWJ 中納言版を拡張
『昭和話し言葉コーパス』50 時間	2020 年度	本公開
『日本語日常会話コーパス』200J 時間	2021 年度	本公開

身に日常会話 15〜18 時間程度，計約 600 時間の会話を収録してもらう。この方法を「個人密着法」と呼ぶ。収録データの中から，均衡性や倫理的問題，データの質などを考慮し，コーパスに格納・公開するデータとして，各人約 4〜5 時間分の会話，計 160〜200 時間を選定する。個人密着法による会話の種類を調査し，個人密着法では収集の難しい種類の会話（例えば職場での会議や接客場面の会話など）については，調査者が主体となり収録する「特定場面法」で補う。現在は個人密着法に基づく収録を進めている。

■コーパスの構成　コーパスに格納する 200 時間の会話のうち，協力者 20 人，各 2.5 時間，計 50 時間を対象に，2018 年度にモニター公開することを予定している。またモニター公開データの中から 20 時間を選定し，「コア」データ（人手による高精度なアノテーションが付されたデータ範囲）として整備する予定である。

■研究用付加情報　会話を収録した上で，次の研究用付加情報を付与する予定である。

転記テキスト　川端ほか (2017), 臼田ほか (2017) に記した基準および手続きに基づき，映像・音声を参照しながら人手で転記テキストを作成する。

発話単位情報　「長い発話単位」(JDRI 2017) に準拠して発話単位を人手で認定する。

形態論情報（短単位情報・長単位情報）　BCCWJ の単位・品詞設計に準じて短単位情報・長単位情報を自動で付与した上で，コアについては人手で修正する。

文節間の係り受け情報　発話単位を範囲に文節間の係り受け関係の情報を自動で付与した上で，コアについては人手で修正する。

談話行為情報　国際標準化規格 ISO24617-2 に基づき日常会話用に整備した基準に基づき，コアを対象に人手で付与する。現在，基準の整備を進めている (居關ほか 2017)。

韻律情報　コアのうち，録音状態や方言の度合などを参考に選別した会話を対象に，CSJ 構築の際に整備したラベリングスキーム X-JToBI を簡略化した「簡易版 X-JToBI（仮称）」(五十嵐 2015) に準拠して人手で付与する。

5. おわりに

本プロジェクトでは，『日本語日常会話コーパス』を主軸としつつ，研究を推進する上で必要となる言語資源を各種整備する。これらの言語資源の公開予定時期を表 1 にまとめて示す。本プロジェクトの研究を進めるために，2016 年度は言語資源の整備を集中して進めた。2017 年

度からは，こうした言語資源を活用しながら本格的に研究を推進する。

謝　辞

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」の研究成果を報告したものである。コーパスの収録にご協力・ご参加くださった皆さまに感謝します。

文　献

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.

五十嵐陽介 (2015).「韻律情報」　小磯花絵 (編)『話し言葉コーパス　設計と構築』: 朝倉書店　pp. 81–100.

居關友里子・第十早織・伝康晴・小磯花絵 (2017).「日常会話コーパスのための談話行為タグの設計」　『言語処理学会第 23 回年次大会』.

JDRI (2017).『発話単位ラベリングマニュアル version2.1』.

柏野和佳子・西川賢哉・小磯花絵 (2017).「『名大会話コーパス』中納言版・ひまわり版公開データの作成」　『言語資源活用ワークショップ 2016』.

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵 (2017).「『日常会話コーパス』の転記基準と作業工程」　『言語資源活用ワークショップ 2016』.

小磯花絵・土屋智行・渡部涼子・横森大輔・相沢正夫・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, pp. 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』.

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.

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–』国立国語研究所報告:18: 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–』国立国語研究所報告:23: 秀英出版.

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.

宮嵜由美・柏野和佳子・山崎誠 (2017).「発話文への発話者情報付与の基本設計—『現代日本語書き言葉均衡コーパス』収録の小説を対象に—」　『言語資源活用ワークショップ 2016』.

臼田泰如・川端良子・徳永弘子・西川賢哉・小磯花絵 (2017).「『日本語日常会話コーパス』の転記基準と特徴について」　『言語処理学会第 23 回年次大会』.

関連 URL

『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』プロジェクトのウェブサイト
http://pj.ninjal.ac.jp/conversation/