

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」
進捗報告

小磯 花絵
(国立国語研究所)

プロジェクトの概要

- ❖ 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 (音声言語研究領域)

- ❖ 研究期間: 2016年4月～2022年3月 (6年間)

- ❖ 研究概要:

日常場面で自発的に生じる会話約200時間を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。

四つの研究班

会話コーパス構築班(班長:小磯)

日常場面で自発的に生じた多様な日常会話約200時間を収録した日常会話コーパスを構築・公開



研究利用

構築支援・コーパス評価



レジスター班 (班長:山崎)

語彙・文法・韻律などに着目して、話し言葉のレジスター的多様性を研究

相互行為班 (班長:伝)

会話相互作用の中で文法が果たす役割を、英語会話などとの比較を通して研究

経年変化班 (班長:丸山)

1950年代の音声資料を用い、文法や韻律などの観点から話し言葉の経年変化を研究

四つの研究班

会話コーパス構築班(班長:小磯)

日常場面で自発的に生じた多様な日常会話約200時間を収録した日常会話コーパスを構築・公開



研究利用

構築支援・コーパス評価



レジスター班 (班長:山崎)

語彙・文法・韻律などに着目して、話し言葉のレジスター的多様性を研究

BCCWJ拡張
発話者情報
シナリオデータ

既存会話データの整備
名大会話コーパス

相互行為班 (班長:伝)

会話相互作用の中で文法が果たす役割を、英語会話などとの比較を通して研究

会話コーパス構築支援
設計・収録・転記基準
談話行為アノテーション

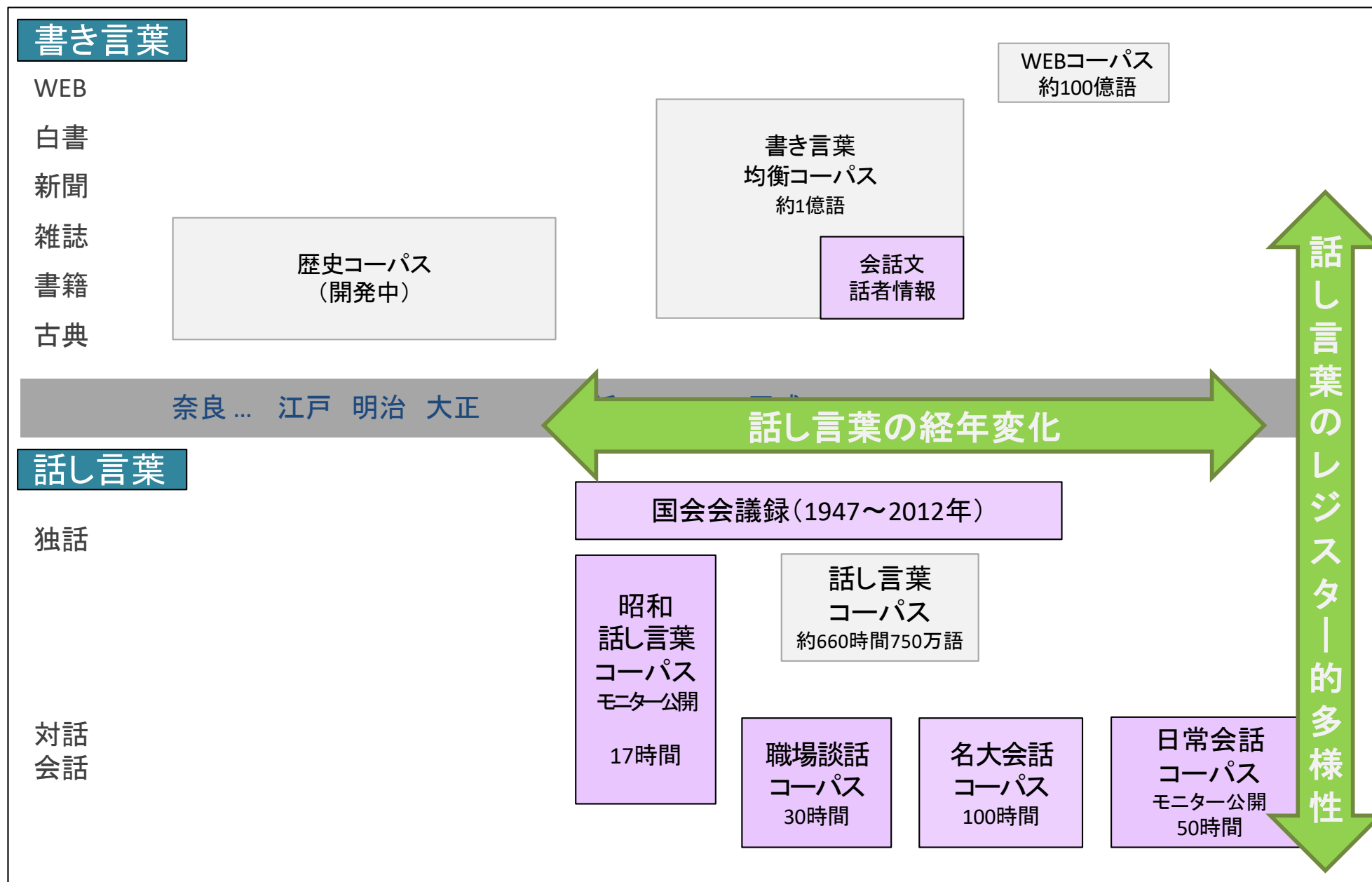
経年変化班 (班長:丸山)

1950年代の音声資料を用い、文法や韻律などの観点から話し言葉の経年変化を研究

昭和話し言葉コーパス
1960前後・50時間

国会会議録
1947-2012年

本プロジェクトで整備・構築するコーパス



コーパス名	略 称	個別検索	包括的検索	備 考
書き言葉 現代日本語書き言葉均衡コーパス 中納言版	BCCWJ	✓	✓	従来より利用いただいている BCCWJ のデータです（ コーパスの紹介ページ ）。こちらのページから BCCWJ アノテーションデータをダウンロードできます。梵天版は こちら
書き言葉 国語研日本語ウェブコーパス 中納言版	NWJC	準備中	✓	梵天版は こちら
話し言葉 日本語話し言葉コーパス	CSJ	✓	✓	コーパスの紹介ページ
話し言葉 日本語日常会話コーパス モニター公開版	CEJC	✓	✓	コーパスの紹介ページ
話し言葉 昭和話し言葉コーパス モニター公開版	SSC	✓	✓	コーパスの紹介ページ
話し言葉 名大会話コーパス	NUCC	✓	✓	コーパスの紹介ページ
話し言葉 現日研・職場談話コーパス	CWPC	✓	✓	コーパスの紹介ページ
通 時 日本語歴史コーパス	CHJ	✓	✓	コーパスの紹介ページ
方 言 日本語諸方言コーパス モニター公開版	COJADS	✓	✓	コーパスの紹介ページ
日本語学習者 多言語母語の日本語学習者横断コーパス	I-JAS	✓	✓	I-JAS に関する詳細な情報は「 I-JAS 関連資料 」をご参照ください。プレインテキスト・音声・作文は「 データ配布 」からダウンロードできます。



本プロジェクトで整備・構築するコーパス

2021.3

書き言葉

WEB

白書

新聞

雑誌

書籍

古典

歴史コーパス
(開発中)

書き言葉
均衡コーパス
約1億語

会話文
話者情報

会話文の話者情報
(性別・年齢層)
データ追加公開

奈良 ... 江戸 明治 大正

昭和

平成

話し言葉

独話

国会会議録(1947~2012年)

昭和
話し言葉
コーパス
モニター公開

17時間

話し言葉
コーパス
約660時間750万語

職場談話
コーパス
30時間

名大会話
コーパス
100時間

日常会話
コーパス
モニター公開
50時間

対話
会話

レジスター	☐ コア	☐ 非コア
出版・新聞	☐	☐
出版・雑誌	☐	☐
出版・書籍	☐	☐
図書館・書籍	☐	☒
特定目的・白書	☐	☐
特定目的・ベストセラー	☐	☐
特定目的・知恵袋	☐	☐
特定目的・ブログ	☐	☐
特定目的・法律	☐	☐
特定目的・国会会議録	☐	☐
特定目的・広報紙	☐	☐
特定目的・教科書	☐	☐
特定目的・韻文	☐	☐

図書館サブコーパス
NDC9(文学)のうち
小説については、
会話文に話者情報
が付いている

ジャンル

上でチェックを入れたレジスターに対し、さらに詳細なジャンルを指定できます。

☒ 図書館・書籍

☐ 日本十進分類法 (NDC) ☐ 0 総記 ☐ 1 哲学 ☐ 2 歴史 ☐ 3 社会科学 ☐ 4 自然科学 ☐ 5 技術・工学 ☐ 6 産業 ☐ 7 芸術・美術 ☐ 8 言語 ☒ 9 文学 ☐ 分類なし

サンプル ID	前文脈	キ ー	後文脈	品 詞	書名/出典	話者名	話者 性別	話者 年齢層
LBi9_00210	とうてい考えられないがねえ」 # 「島岡専務は、わたくしを 憎んでいるんです」	わ	「...だから、そんなとんでもない 嘘を...」 # 「憎む? # それ はどういう	助詞- 終助 詞	黒白の囃	関口友子	女	成年 層
LBi9_00062	やっぱり...こういう格好はちょ っと...」 # 「何言ってるの よ。 # すごく似合ってる」	わ	「よ。 # ねえ、みんな?」 # 「ド リーンがうっとりと言う。 # 職業 柄、というわけ	助詞- 終助 詞	黒炎の貴公子	ドリーン	女	若年 層
LBt9_00271	てからの方がいいんじゃない い? # 「何も入っていないかも しれない」	わ	「よ」 # 「もし空なら、あなた はこの先も情報なんて寄越さな いでしょう。 # 同じ	助詞- 終助 詞	黒の狙撃手	M	女	成年 層
LBk9_00004	はアラウカーノ族の勇敢な血 が入っているんだろう?」 # 「そうだった」	わ	「。 # でも、ゴローに誤解されな く嬉しい。 # 話してよかった。 。 # ...今夜	助詞- 終助 詞	黒いコンドル	サンドラ	女	成年 層
LBn9_00254	」 # 「エリスはほろ苦く微笑っ ている。 # 「ダイヤのこと? # 「ええー大好きだ」	わ	「」 # 「いつか、持ち切れない ほど君にあげよう」 # 「突拍子 もない申し出	助詞- 終助 詞	鴛の驕り	ジーン・ケ アリ	女	成年 層
LBn9_00254	皮肉の混じった声を出した。 # 「料金も高い」 # 「正当な 報酬だ」	わ	「一世紀の発明王の権利を守るの だから」 # 「権利ビジネス、 か」	助詞- 終助 詞	鴛の驕り	ジーン・ケ アリ	女	成年 層
LBg9_00157	贅沢って言うのよ。 # 「(新 宿)へ来てから、もう二十人 以上殺した」	わ	「。 # それからーそれから...」 # 「富士子は両手で顔を覆った。 # 瘡の	助詞- 終助 詞	魔界都市ブル ース	富士子	女	成年 層

本プロジェクトで整備・構築するコーパス

書き言葉

WEB

白書

新聞

雑誌

書籍

古典

WEBコーパス
約100億語

書き言葉
均衡コーパス
約1億語

歴史コーパス
(開発中)

会話文
話者情報

2021.3

昭和

平成

1950～70年代に
国語研究所で収録した
独話17時間
会話25時間
本公開

話し言葉

独話

国会会議録(1947～2012年)

昭和
話し言葉
コーパス
本公開

独話17H
会話25H

話し言葉
コーパス
約660時間750万語

対話
会話

職場談話
コーパス
30時間

名大会話
コーパス
100時間

日常会話
コーパス
モニター公開
50時間

資料 ID	会 話 ／ 独 話	前文脈	キ ー ウ ー ド	後文脈	品 詞	録 音 年	形 式	場 所	年 齢	性 別	職 業
C53_03_BT ▶	会話	で食べればいいんじゃないですか # #Iした # わたしは一本 # あ三十円出るI	わ	Iね # 三十円 # うーん # 一人じゃオIおっきすぎるIからIな # あれあれはあーもっと小さいのIちょうだい	助詞-終助詞	1953年8月	用談相談	職場	NA	女性	魚屋店主
C54_01_CT ▶	会話	一番意外だったのIねーI組 # 何Iあなた # ーI組もI全然I静かなIんだ # おとなしかったI	わ	Iね # それでね # うん # もうIねあんなIねもうIどってIんばったIんなIクラスIがさIもうI静かにIね	助詞-終助詞	1955年3月	雑談	学校	15-19歳	女性	高校生
C57_10_CT ▶	会話	んだね # ちょっとIこれIご飯IにしようIか # ちょっとIこれIいつものI森I雅之IのI感じIしない	わ	Iね # うーん # まあIこれIいやIらしいIそうIね # え # 母IとI子IのI窓 # あIこれIあのI読売I新聞I一時I出	助詞-終助詞	1957年5月	雑談	個人宅	45-49歳	女性	主婦
C54_01_CT ▶	会話	Iはある # それI先生Iが悪い # そうIいうIことI # Iでしょ # #IかなりIあとIで # それIはI # IであるI	わ	Iよ # いいI心I掛けIだIなIとI言うIようIなIうちIだけ # ーI組I二I組I三I組IはIものIすごくI勉強	助詞-終助詞	1955年3月	雑談	学校	15-19歳	女性	高校生
C57_25_CT ▶	会話	それでI振り出しIで # へえー # そいIであIのI女I店員IがI着物I着IてIさ # 大変IなI騒ぎIだったI	わ	Iよ # ああー # ねI赤I前IダイI赤I前I掛けI締めIてIさ # 何 # ほいIであIそこIでIいらっしやIいIつつIてIね	助詞-終助詞	1957年6月	雑談	個人宅	50-54歳	女性	主婦
C54_01_CT ▶	会話	ね # だIってI自分IたちIのI月給IをI自分IたちIでI決めるIんIだIもんIね # そう # そうIね # #I	わ	IねIあんなIのIね # 勝手IにIね # 欲しいIだけI取IっIちゃIう # 随分I勝手IだIねIみんなIね # うーん # 全然 # なん	助詞-終助詞	1955年3月	雑談	学校	15-19歳	女性	高校生
C57_12_CT ▶	会話	うちIの子 # うんIそうIよIそうIですIよIね # あらIあらIあらIあら # そう # だIからIもうI熱演IだI	わ	IこのIかIたIたちIは # ロンドIもIなさIったIの # ほー # ええ # なIかなIかIねIモーIツアルトIのIソナタIこなせるIようIこんなIお	助詞-終助詞	1957年6月	雑談	個人宅	40-44歳	女性	主婦

本プロジェクトで整備・構築するコーパス

書き言葉

WEB

白書

新聞

雑誌

書籍

古典

歴史コーパス
(開発中)

書き言葉
均衡コーパス
約1億語

会話文
話者情報

WEBコーパス
約100億語

2021.2

奈良 ... 江戸 明治 大正

昭和

平成

中納言

100時間の日常会話
モニター公開(50時間増補)

話し言葉

独話

国会会議録(1947

昭和
話し言葉
コーパス
本公開

独話17H
会話25H

話し
コーパス
約660時間750万語

職場談話
コーパス
30時間

名大会話
コーパス
100時間

日常会話
コーパス
モニター公開
100時間

対話
会話

検索対象の選択

公開時期

☐ 2018年度版 ☐ 2020年度版

会話 ID	前文脈	キ ー	後文脈	品 詞	公 開 時 期	形 式	年 齢	性 別
C002_013a ▶	見て話しをするよりほかないんだ	わ	# うーん # そうね # んね # だから行こうね # もう	助詞- 終助詞	2018年 度版	雑談	85- 89歳	女性
S002_004 ▶	んですよ # あ # あのねたぶんねあれだ	わ	# 俺電車ん中で飯を食う文化が	助詞- 終助詞	2020年 度版	雑談	35- 39歳	男性
T004_010b ▶	レンジ # そう # それで終わってああもう終わった	わ	って感じで子供で # うん # うん # あたしま だ作っ	助詞- 終助詞	2020年 度版	雑談	55- 59歳	女性
K004_013 ▶	をどうするか # あっ # これはもう入れた	わ	# なんかへーしか言っていないわ # あー # これ	助詞- 終助詞	2018年 度版	会議 会合	50- 54歳	女性
S002_014 ▶	うんうん # いいじゃん # 巖島 # 巖島神社すげえよか った	わ	# 行ったことなかったの # 初めて行った # 修 学	助詞- 終助詞	2020年 度版	雑談	35- 39歳	男性
T004_010b ▶	あー # うん # すごい # でもそれは子供も偉い	わ	# うん # うーん # 大したもんだ # それはもう ね # 行き	助詞- 終助詞	2020年 度版	雑談	55- 59歳	女性
T002_015 ▶	時にやり放題やっちゃって酒は飲む	わ	ジャンクに走るわベビートイん時なんか毎 日	助詞- 終助詞	2018年 度版	雑談	40- 44歳	女性
T005_011b ▶	そこらにいたってしゃあないしね # じゃあふらふら する	わ	！と思ったよね # # # # # # # # # # # # # そっかイイ	助詞- 終助詞	2020年 度版	用談 相談	45- 49歳	男性

『日本語日常会話コーパス』 本公開に向けて

年代	男性			女性		
	協力者ID	職業・職種	時間	協力者ID	職業・職種	時間
20代	T010	大学生	4.2h	T009	大学生	6.0h
	T006	大学院生	4.3h	K003	大学生	4.4h
	T022	会社員・公務員等	3.7h	K009	会社員・公務員等	4.2h
	K007	会社員・公務員等	5.5h	K013	会社員・公務員等	4.0h
30代	T001	自営業・自由業	5.6h	K001	会社員・公務員等	5.0h
	T005	会社員・公務員等	4.6h	T003	専業主婦	5.6h
	S002	会社員・公務員等	4.7h	K005	自営業・自由業	5.4h
	K012	会社員・公務員等	3.1h	T008	自営業・自由業	4.8h
40代	T016	会社員・公務員等	3.7h	C001	会社員・公務員等	4.5h
	T002	自営業・自由業	4.8h	T011	パート・アルバイト	4.9h
	T019	会社員・公務員等	3.9h	K004	パート・アルバイト	5.0h
	T020	会社員・公務員等	5.0h	T014	自営業・自由業	4.4h
50代	T015	会社員・公務員等	6.0h	C002	会社員・公務員等	4.2h
	S001	会社員・公務員等	4.6h	K002	自営業・自由業	4.6h
	T024	会社員・公務員等	4.3h	K008	会社員・公務員等	4.6h
	T018	自営業・自由業	4.2h	K011	会社員・公務員等	4.5h
60歳 ～	T013	非常勤講師	4.3h	T004	専業主婦	5.1h
	T007	定年退職	5.8h	K006	自営業・自由業	4.4h
	K010	会社員・公務員等	4.8h	T017	会社員・公務員等	4.2h
	T023	定年退職	4.6h	T021	自営業・自由業	4.3h

残り15時間: 不足する場面・話者層の収録

- 仕事の会議・会合
- 未成年者(中学生・高校生)
 - ✓ レクチャー・会合
 - ✓ 雑談

データ規模

会話時間	200時間
会話数	460会話
話者数(延べ)	1498人
話者数(異なり)	775人
語数	223万語

年齢・性別の分布（延べ）

