

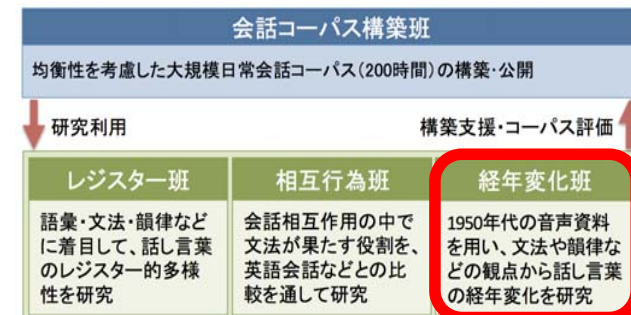
『昭和話し言葉コーパス』の構築

丸山 岳彦（専修大学）

2016年 9月 1日（木）
シンポジウム「日常会話コーパス」Ⅰ
（国立国語研究所）

研究体制

「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」



研究体制

- 大規模日常会話コーパス：経年変化班
- 科研基盤B：「昭和話し言葉コーパス」の構築（2016~2020）

相澤 正夫（国語研）	服部 匡（同志社女子大）
佐野真一郎（慶應大）	丸山岳彦（専修大）
塩田 雄大（NHK文研）	山口昌也（国語研）
高田三枝子（愛知学院大）	山崎 誠（国語研）

研究のねらい

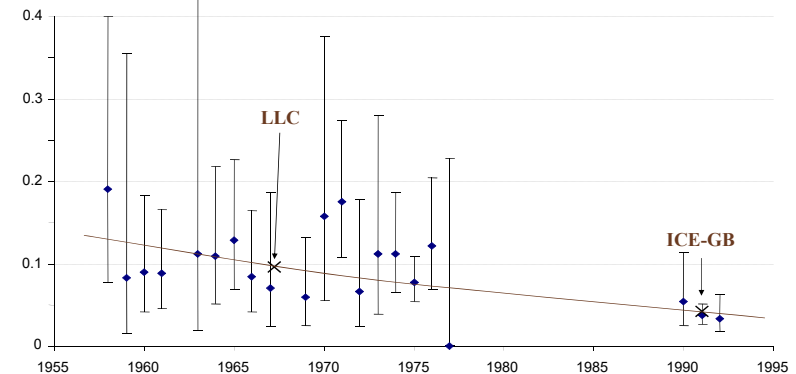
- 古い時代の録音資料を、コーパスとして編集する。
 - ◆ 音声 = 過去の話し言葉を知るための第一次資料
 - ◆ 対象となるのは、20世紀以降に録音された音声
 - ◆ 現存する音声資料の調査・収集が必要
- 現代から見た、過去の話し言葉の特徴を探る。
 - ◆ 発音、アクセント、イントネーション
 - ◆ 文法、談話
 → 「話しことばの経年変化」に関する実証的な研究

関連研究

- DCPSE (Diachronic Corpus of Present-day Spoken English) University College London cf. Survey of English Usage
 - ◆ イギリス英語の自然談話を収録した通時コーパス
 - ◆ London-Lund Corpus から40万語 (1958年～1977年)
 - ◆ ICE-GB から40万語 (1990年～1992年)
- Aarts et.al (2014)
 - ◆ 助動詞 must, may, shallの使用が大幅に減少、would, could, should も減少、can, willが増加
 - ◆ 現在進行形の比率が増加



“shall” の減少 (DCPSE)



『昭和話し言葉コーパス』

- 昭和期の録音資料を収集し、段階的に整備
 - ◆ できるだけ日常談話を集めたい
 - ⇕
 - ◆ 現存する録音資料 (特に日常談話) は限られている
- 昭和20年後半～40年代前半
 - ◆ 国立国語研究所「話しことば研究室」が収集した日常談話資料
 - 『談話語の実態』 (1955)
 - 『話しことばの文型 (1) (2)』 (1960, 1963)

1950年代の国語研「話しことば研究室」

- 昭和27年、「第一研究室」が設置される
 - ◆ 共通語の話しことば研究が、方言研究から分離
 - ◆ 中村通夫、大石初太郎、飯豊毅一、宇野義方、進藤咲子



モーター・レコーダーによるラジオの話しことばの共同分析



ソナ・ストレッチャーによる音声資料の聞き取り



ソナ・グラフによる実験の状況

今回収集した録音資料の例（対話）

- 3人の女性
- 九段高校生
- 石野家雑談
- 職安男子部
- 絵画館のおばさん
- 荒井美髪店
- ジイサン・バアサン
- 大修館応接室
- 三鷹学生
- トタン屋
- 養老院
- 一研雑談
- 組合団交
- 麻布主婦
- 少年工員
- 魚屋小僧



当時の作業風景

64ファイル、約37.5時間
約264,000字

今回収集した録音資料の例（独話）

- 助詞・助動詞
- 講演4人
- 国語講義
- 日本語のアクセントほか
- 国研新庁舎開き式典
- 国立国語研究所10周年記念式典
- 新庁舎開き記念講演会
- 創立10周年記念講演会
- 創立20周年記念講演会



林四郎氏 講演（1968年）

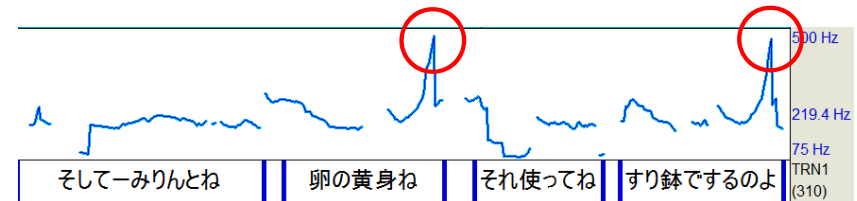
33ファイル、約28.7時間
約146,000字

『昭和話し言葉コーパス』の構築手順

1. 録音資料の（再）収集（50時間分）
※ 録音状態のよいものを選別
2. 転記テキストを新規に作成
3. 転記テキストに時間情報を付与
4. 各種アノテーションの実施
 1. フィラー、語断片
 2. 発話単位、節境界ラベル
 3. 形態論情報
 4. 韻律情報

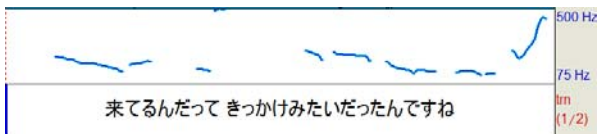
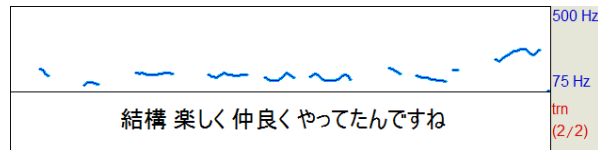
分析1：上昇イントネーション

- 句末・発話末の急激な上昇イントネーション
- 例：「三人の女性」（1957年2月録音）



分析1：上昇イントネーション

- 句末・発話末の急激な上昇イントネーション (CSJ)



分析1：上昇イントネーション

- 現在も「句末の急激な上昇調」は存在する？
 - ◆ 上品な高齢の女性が少し気取って話すような場面
- 1950年代に20代の女性 → 現在は80代
 - ◆ 現在でも当時のイントネーションを保持？
 - ◆ 「おばあちゃんがしゃべってるみたい」

話し言葉に現れるイントネーションの経年変化

分析2：「まする」

- 非常に予算の窮屈な、あー、時代でありまするから、
(山本有三：1887年生)
- ラジオニュースの書き方というような本を見ますると、
(波多野完治：1905年生)
- 20万語を収載すると書いてありまするけれども、
(林大：1913年生)

分析2：「まする」

- 国語の問題というのは難しいんでありますから、
(山本有三：1887年生)
- えー、選挙の、おー、放送を聞いておりますと、
(波多野完治：1905年生)
- 非常に、あの、時代的な差もありますけれども、
(林大：1913年生)

同一話者による文法形式のゆれ

分析2：「まする」

	ます	まする
会話	2,668 (100.0%)	0 (0.0%)
独話	5,280 (98.1%)	104 (1.9%)
山本有三	103 (84.4%)	19 (15.6%)
波多野 完治	109 (97.3%)	3 (2.7%)
林 大	155 (96.3%)	6 (3.7%)

分析2：「まする」

- 「まする」の出現条件
 - ◆ 演説や講演など、改まった発話スタイルの独話
 - ◆ 比較的少数の話し手による？
 cf. 国会会議録では現代でも観察可（服部 2011）
- 「まする」の消失
 - ◆ 時代が進むにつれて若い世代で「まする」が消失

話し言葉に現れる文法形式の変化

分析3：「ざんす」

- それであの晩、怖ござんしたよ。
「絵画館のおばさん」（1952年9月録音）
- 何でもたくさんたまったからよござんすね。
「麻布主婦(1)」（1957年5月録音）
- ああ、そうざんすか。
「PTA 戦後の国語教育(1)」（1957年7月録音）

話し言葉に現れる文法形式の消失

問題（1）：不明瞭な音声への対処

- 小さい音量・ノイズなどで、発話内容が聞き取れない
例：「男女青年雑談」（1953年？録音）

2：みんな●、経営者、そして儲かる、28年のね、●の方々には日本の国のためにね、こういうことのために発展してきて、力を尽くしておられる●ならいいけど、民間のほうが先へ進んじゃってね。●。立ち遅れてるっていうのは、●。

5：●

2：●

5：そういうのと比較されたのよく出てるわね。2倍になったり。

問題（２）：発話の重複による問題

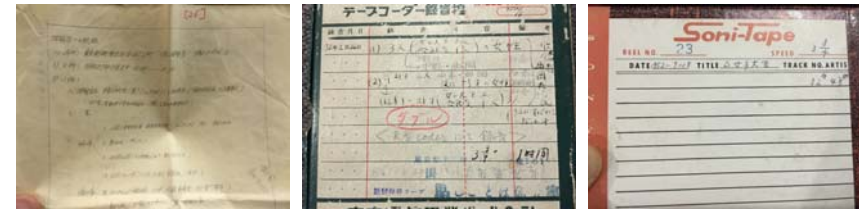
- 発話の重複で聞き取れない、発話者の割り当てが困難

例：「絵画館のおばさん」（1952年9月録音）

- | | |
|--|--|
| 7：今、もう、こちらや、こちらへ、
みんななんか、もう、し、知り
抜いてるけど。 | 7：若い人のおっしゃることもね。
1：おばあさんですからね。（笑）
2：こちら長いんですか。 |
| 6：ううん、とんでもないこと。
私なんか、何にも分かりません。 | 7：なんでも、その、★キョウ★。
1：ええ。 |
| 7：何でも。若い人、いるから。あ
んたんとこいったら、みんな●。 | 7：親ですから聞いてますが、私ど
もでは、また、みんな、もう。
1：ざっと30年ばかり。 |
| ?：●。 | |

問題（３）：メタデータの不足

- 一部は、発話者・発話場面などの情報が分からない



詳細なメモが
残されている場合

箱の裏にメモが
残されている場合

ほとんど情報が
ない場合

当面の課題

- 転記テキスト（第一次作業）の修正
 - ◆「聞き取れない箇所」への対処、発話者の割り当て
- 転記テキストに対する時間情報の付与
- メタデータの整備
 - ◆発話者の属性（性別・年齢）
 - ◆録音時期
- パイロット的な分析（音声・音韻・語彙・文法）

作業工程：科研 基盤(B)

