『昭和話し言葉コーパス』の構築(4):メタデータの設計とデータの公開

丸山 岳彦
(専修大学・国立国語研究所)

シンポジウム「日常会話コーパス」VI
2020年3月4日(木)

1

構成

1. 経緯
   - 「経年変化班」のねらい
   - これまでの活動の経緯
2. 『昭和話し言葉コーパス』
   - 収録音声の例
   - コーパス化の手順と方針
   - 量的な構成
   - メタデータの設計
3. 分析
   - 終助詞の分布
4. 展望
   - 「通時音声コーパス」と連結可能性

2

「経年変化班」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
- → 「話し言葉の経年変化」に関する実証的な研究

3

これまでの活動の経緯

●2016年4月『昭和話し言葉コーパス』設計・構築 開始
●2017年9月 国際シンポジウム 開催(国語研)
●2019年3月『モニター公開版』(中納言)公開開始
●2019年3月『モニター公開データ 第1版』配布開始
●2020年1月『モニター公開データ 第2版』配布開始
●2021年3月『昭和話し言葉コーパス』完成・公開(中納言)
●2021年度 『昭和話し言葉コーパス』 配布開始(予定)

4

5

『昭和話し言葉コーパス』(SSC)

- 1952～1974年に国語研で作られた録音資料を再編
  - ◆独話　50資料　17時間　180,664語　50人
  - ◆会話　73資料　27時間　348,458語　344人
- 2020年度末に一般公開
  - ◆中納言
  - ◆データ配信（音声、転記テキスト、形態論情報、検索環境、メタデータ）

6

SSC 収録音声：独話

1955年3月　国立国語研究所新庁舎開き式典
所長挨拶ならびに経過報告 / 祝辞

1955年3月　国立国語研究所新庁舎開き記念講演会
現代の敬語意識 / 三つの語彙調査

1957年9月　全国国語科指導主事研修講座
話し言葉の表現意図 / 方言調査法 / 言語能力の発達 / 助詞・助動詞 / 文型 / 国語教育 / 新聞文章研究法

1959年3月　国立国語研究所創立10周年記念祝賀式
所長挨拶 / 祝辞 / 来賓挨拶

1959年3月　国立国語研究所創立10周年記念講演会
明治初期の書きことば / 現代語の標準 / 話しことばの文法 / これからの日本語

1965年12月　第17回国立国語研究所創立記念講演会

1966年12月　第18回国立国語研究所創立記念講演会

1967年12月　第19回国立国語研究所創立記念講演会

1968年3月　見坊氏退官記念講演

1969年2月　国立国語研究所創立20周年記念講演会
あいさつ－研究所と語彙研究－ / 語彙調査と基本語彙 / 形容詞の意味の特質

1969年12月　第21回国立国語研究所創立記念講演会

1972年12月　第24回国立国語研究所創立記念日

1974年6月　国立国語研究所研究棟落成式典
所長挨拶 / 来賓挨拶 / 祝辞

7

SSC 収録音声の例：独話

1959年3月6日　時枝誠記「祝辞」
（国立国語研究所 創立10周年記念祝賀式）

で どういうことが 考えられてきたかと 申しますというと おー 言語を そういう 流動の 姿に おいてでなくて えー 静止した 姿に おいて えー 文字であるとか 仮名遣いであるとか っていう ことのみが 問題視されて そこに えー 国語問題の 解決の おー 糸口を 見出そうとし また そこに 基礎的な あー 研究が あると こういうふうに いー 考えてきたと 思うんであります

8

SSC 収録音声：会話

会議会合
　組合団交

用談相談
　その電気ごたつは安全ですか / タクシー苦情 / 魚屋 / 研究室の電話

雑談
　3人の女性 / 3人の青年 / A美髪店 / I家雑談 / K教育委員会雑談 / K高校生 / N家雑談 / S女子大事務室 / S女子大生 / T社応接室 / U家雑談 / U夫妻 / Y理髪店 / ジイサン・バアサン / トタン屋 / 一研雑談 / 下町家族 / 絵画館のおばさん / 鎌倉主婦 / 魚屋小僧 / 劇団員雑談 / 三鷹学生 / 三鷹分室 / 歯科大学生 / 女性雑談 / 少年工員 / 接客用語について / 扇子屋 / 浅草噺 / 男女学生座談 / 麻布主婦 / 面接録音調査 / 友の会 / 養老院

9

SSC 収録音声の例：会話

1952年9月　「三鷹学生(2)」

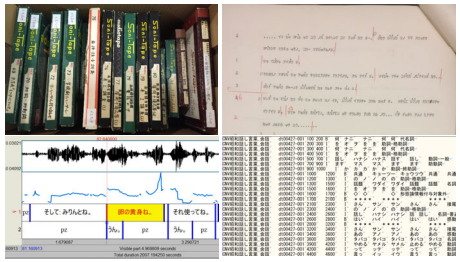
1-3: 15-24歳（学生）4: 40-44歳（男性） 5: 35-39歳（女性）

1: いやあ スキーなんか スキーは 危ないらしいですよ
3: どうして
1: いや あの よく あの 木 木の 切り株なんかが あちこち に ある
1: それに ぶつかると 足が こう 反対に なっちゃうんだ
2: 足は そんな ないよ
3: 危なくないですよ
5: 切り株の ある ところへ 行かないでよ
1: ああ そうです
2: だ 特にね そりゃ 上手ん ならなきゃいけないですよ
2: だ こう 平地で 滑るんだよ 普通は
3: そうそうそう
2: 初めはね
3: うん
2: 少し 上手に なれば 山だってね 滑るけども
4: 山滑りは やらないんだよ
2: 山滑りまで 行くようん なれば
4: 畑あたりで やるんだよ
2: やや 緩い 傾斜でね すーっ と じゃ それ じゃあ 平気
5: 写真が あるんですよね スキーに 行ったときの 写真
5: だけどね その後で すぐ 転んだんですって 一番 初めね
2: まず 一通り ス スタイルを そろえるだけでも 大変ですね スキーは
5: そうですよね
2: まず スタイル そろえて 写真を 撮るんだよ

10

コーパス化の手順と方針

●コーパス化の手順
- ◆当時の録音資料を収集
- ◆新規に転記テキストを作成
- ◆時間情報を付与
- ◆形態論情報を付与
- ◆メタデータを作成



●個人情報・著作権の処理
- ◆独話：国語研の所員以外の学者による講演は収録を見送り。ただし、祝辞・挨拶などは収録
- ◆会話：個人名・固有名は伏字化

11

量的な構成 (1)

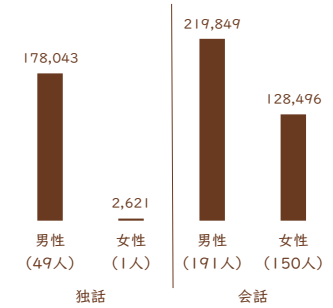
●『昭和話し言葉コーパス』(SSC)

タイプ	録音数	総時間	総語数	話者数（男/女）
独話	50資料	17時間	180,664語	50人（49/1）
会話	73資料	27時間	348,458語	343人（191/150）
計	123資料	44時間	529,122語	

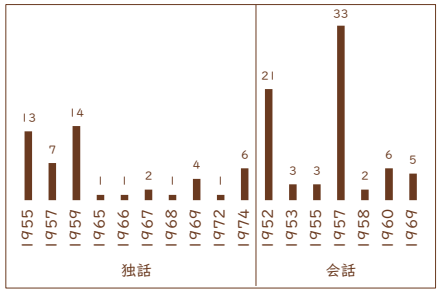
12

量的な構成 (2)

●男女別 語数



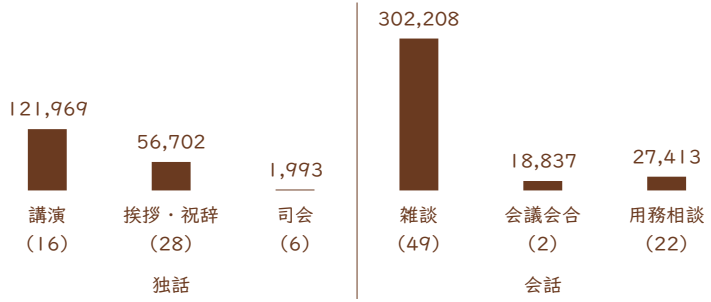
●収録年の分布（録音数）



13

量的な構成 (3)

●形式別 語数



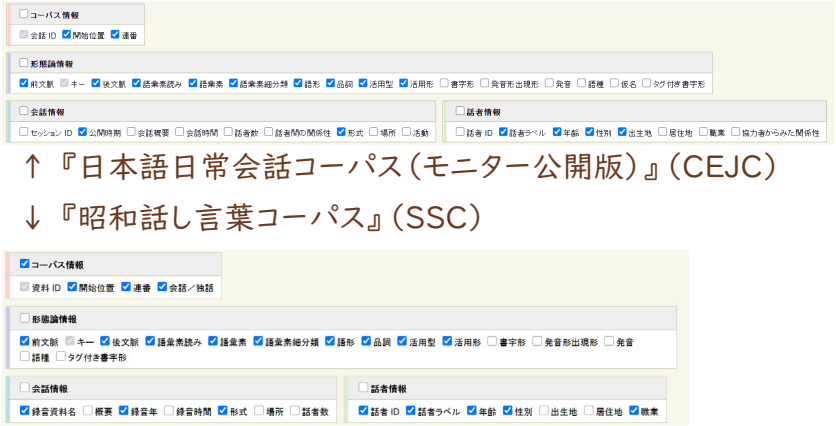
14

メタデータの設計

●SSCメタデータ（中納言版）

◆『日本語日常会話コーパス』(CEJC) に可能な限り準拠

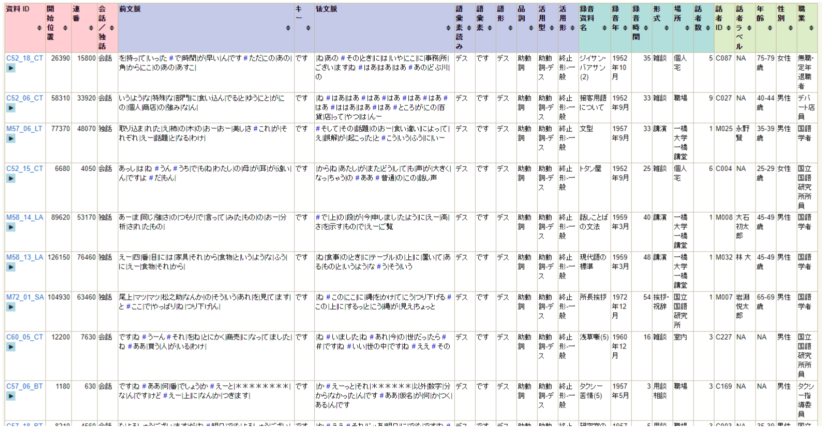
15



↑『日本語日常会話コーパス（モニター公開版）』(CEJC)

↓『昭和話し言葉コーパス』(SSC)

16



17

分析：終助詞の分布

●『昭和話し言葉コーパス』（SSC）の会話データ

●『日本語日常会話コーパス』（CEJC）2018年度版　を比較

	SSC		CEJC	
男性	191人	219,849語	114人	265,201語
女性	150人	128,496語	145人	357,101語
計	341人	348,345語	259人	622,302語

18

分析：終助詞の分布

●全終助詞の出現数

終助詞	SSC	CEJC
男性	898	< 1,510
女性	971	< 1,475

（各層 10000語あたりの出現数に正規化）

終助詞の使用数が増加

19

分析：終助詞の分布

SSC	男性	女性	CEJC	男性	女性
ネ	433	541	ネ	483	580
ヨ	171	150	ヨ	323	281
カ	107	58	カ	205	184
ナ	80	12	ノ	113	121
ノ	30	65	ナ	141	100
サ	42	33	サ	110	99
ワ	4	75	ジャン	47	36
モン	14	10	ッケ	29	24
カシラ	2	15	モン	26	22
ッケ	3	3	ワ	10	13
モノ	2	5	ヤ	7	3
イ	3	0	カシラ	0	4

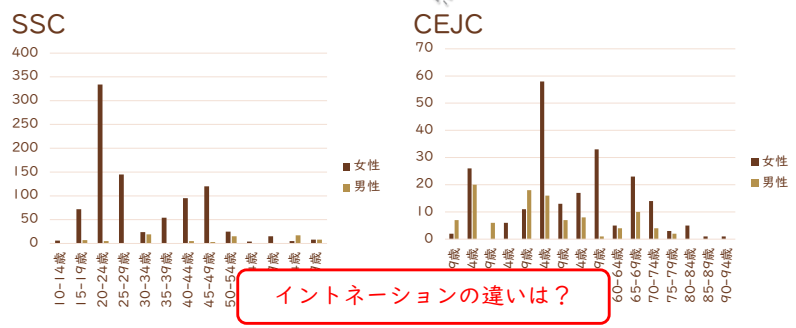
（各層 10000語あたりの出現数に正規化）

終助詞の経年変化

20

分析：終助詞の分布

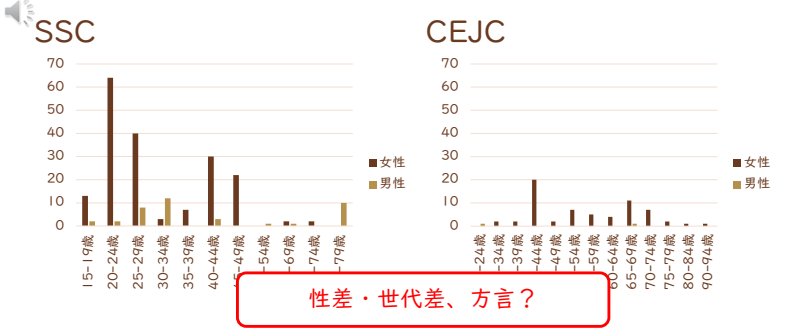
●終助詞「わ」の分布



21

分析：終助詞の分布

●終助詞「かしら」の分布



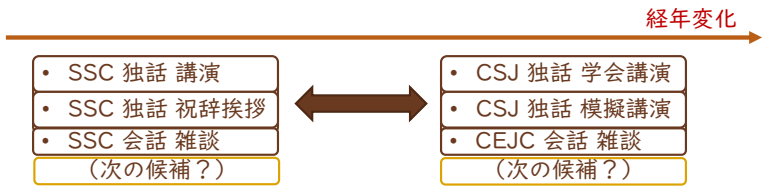
22

「通時音声コーパス」と連結可能性

●「話し言葉の経年変化」を知るためのコーパス群

→ 時代ごとにモジュール化、順次整備を進めて、連結

cf. 日本語歴史コーパス（CHJ）



「連結可能性」 → レジスターの考慮・統一的メタデータの設計

23

謝辞： 本研究は、科研費基盤研究(B)「「昭和話し言葉コーパス」の構築による話し言葉の経年変化に関する実証的研究」（16H03426）、および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」に基づくものです。

現在までに『昭和話し言葉コーパス』構築に携わってきたのは、以下のメンバーです（敬称略 50音順）。

伊藤優介、臼田泰如、大川恵莉、小野瀬敦也、河本はるか、菊池千尋、小磯花絵、近藤明日子、十河則子、田嶋明日香、土屋菜穂子、中神裕美子、中村壮範、西川賢哉、藤村寛子、松下晶子、丸山岳彦、森本桂子、山縣智子、山口昌也、劉双戌、渡邊友香

24