通時音声コーパスの可能性：

『昭和話し言葉コーパス』の構築と分析

丸山 岳彦
（専修大学・国立国語研究所）

2018年 7月 18日（水）
外語大語学研究所　定例研究会

今日の内容

1. コーパス言語学とは何か
   a. なぜコーパスが必要か
   b. 話し言葉コーパスの構築
2. 話し言葉コーパスの歴史
   a. 20世紀前半の録音資料
   b. 1950年代の音声コーパス
   c. Survey of English Usage
   d. 2000年代の音声コーパス
3. 『昭和話し言葉コーパス』
   a. 研究目的、構築体制
   b. 収集した録音資料の例
   c. イントネーションの変化
   d. 文法形式の変化
   e. 今後の課題
4. 通時音声コーパスの可能性
   a. 通時音声コーパスとは何か
   b. 共時的視点と通時的視点

1. コーパス言語学とは何か

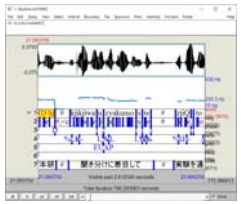
コーパス言語学とは何か

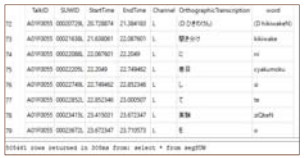
- コーパスとは何か？
  - 実際の書き言葉・話し言葉の例を大量に収集してデジタル化し，コンピュータで検索できるように構造化した言語資料
  - “a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis”. (Francis 1982)
  - 書き言葉コーパス、話し言葉コーパス、歴史コーパス、学習者コーパス、方言コーパス、映像付きコーパス…
- コーパス言語学とは何か？
  - コーパスを用いて言語を研究する方法論

話し言葉コーパスの構築

●話し言葉コーパスの構築手順
1. 音声の録音
2. 転記テキストの作成
3. アノテーション（研究用情報の付与）

5	つまり　＆ツマリ
6	長調ですとか　＆チョーチョーデストカ
7	短調の　＆タンチョーノ
8	変化の　＆ヘンカノ
9	(D ひきわけん)　＆(D ヒキワケン)
10	0012 00021.611-00023.045 L:
11	聞き分けに　＆キキワケニ
12	着目して　＆チャクモクシテ
13	0013 00023.384-00027.184 L:
14	実験を通して　＆ジッケンオトーシテ
15	(F えー)　＆(F エー)
16	知見を　＆チケンオ
17	得ましたので　＆エマシタノデ





話し言葉コーパスの条件

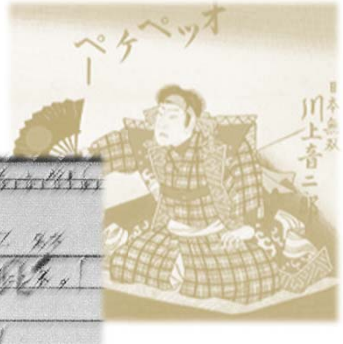
●朗読音声ではない、自発的な話し言葉を収録していること

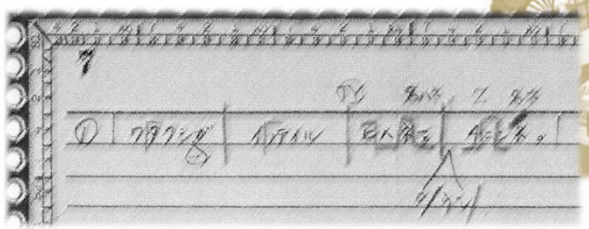
●多様な話者・場面による話し言葉を収録していること

●定量的な分析に耐える量を備えていること

●種々の研究用情報がアノテーションされていること
- ◆転記テキスト・時間情報
- ◆発話単位
- ◆韻律情報、形態論情報、統語構造情報、談話構造情報…
- ◆メタデータ（録音年月日、話者情報、場面情報…）

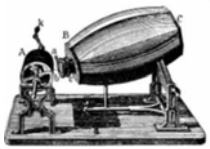
2. 話し言葉コーパスの歴史





20世紀前半の録音資料

●録音された最古の人の声
- ◆「月の光」
- ◆1860年4月9日 録音



●録音された最古の日本人の音声
- ◆1900年 7月 パリ万博での録音
- ◆岩間くに（新橋料亭の女将）
- ◆川上音二郎一座「オッペケペー節」

20世紀前半の録音資料

●国立国会図書館「歴史的音源（れきおん）」
- ◆民謡、歌舞伎、落語、政治演説、放送…
- ◆ウェブサイトから一部を聴取可能

●「岡田コレクション」
- ◆35,000枚の「SPレコード」
- ◆合計12時間、212の政治演説など
- ◆1915年～1946年の録音

『岡田コレクション』の例

●大隈重信（1838-1922）の演説音声
- ◆1915年3月2日 録音
- ◆「憲政ニ於ケル世論ノ勢力」

帝国議会は解散されました。
今まさに、旬月の後に選挙が行なわれて、
今全国は選挙の競争が盛んに起こっておる
時でありますんであります。

1950年代の音声コーパス

●国立国語研究所における話し言葉研究
- ◆1948年設立
- ◆1952年「話しことば研究室」を開設

モニター・レシーバーによる
ラジオのことばの共同分析

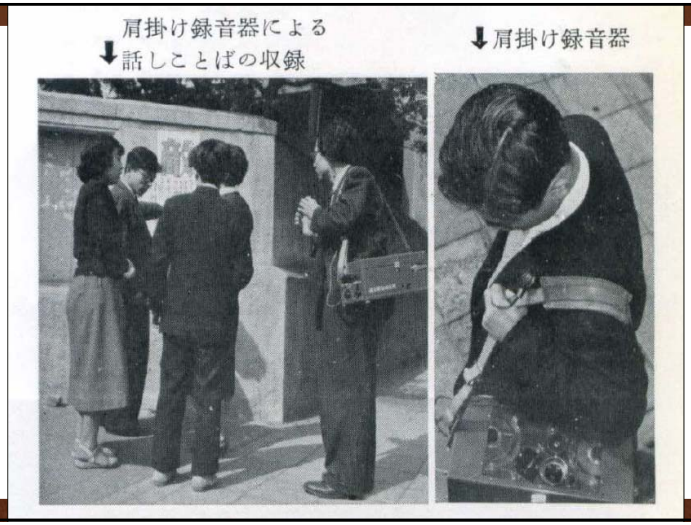
ソナ・ストレッチャーに
よる音声資料の聞き取り

ソナ・グラフによる実験の状況

1950年代の音声コーパス

●オープンリールレコーダーで、日常談話を収録
- ◆60巻の録音テープ（約30時間分）に日常談話の音声を録音
- ◆さまざまな発話場面を偏りのないように収集（サンプリング）

日常の談話が多く得られる場合として，衣食住・社交等の生活機能と家庭・近隣・職場・市町村などの生活環境との切点から，具体的な談話の場面を収集して一覧表を作り，同時に，話し手・聞き手の性・年齢・教養・相手（の人数，未知既知）・地域などにもなるべく片寄りのすくないことを目安として，調査地点・調査対象・調査場面の予定表を作成した。

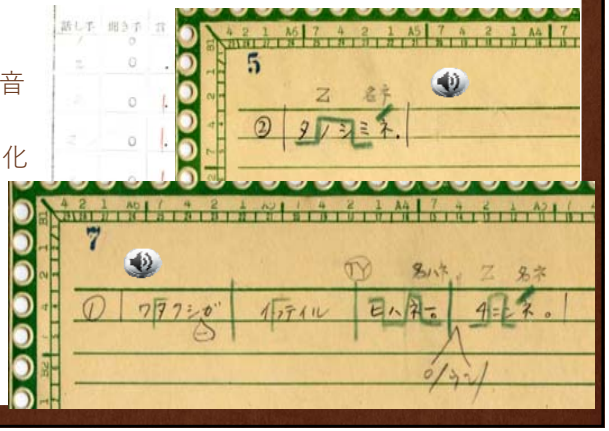


当時のオープンリールテープ

- 現在は中央資料庫に収蔵、大半はデジタル変換済み
- メモ書きがある場合、貴重な資料（メタデータ）になる

当時の録音資料

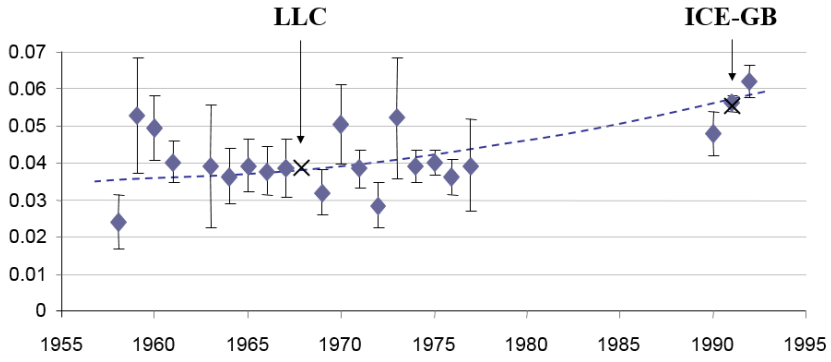
- 「三人の女性」
  - ◆1957年2月26日録音
  - ◆転記テキスト
  - ◆発話ごとにカード化
  - ◆アノテーション
    - ➢アクセント
    - ➢韻律
    - ➢統語情報



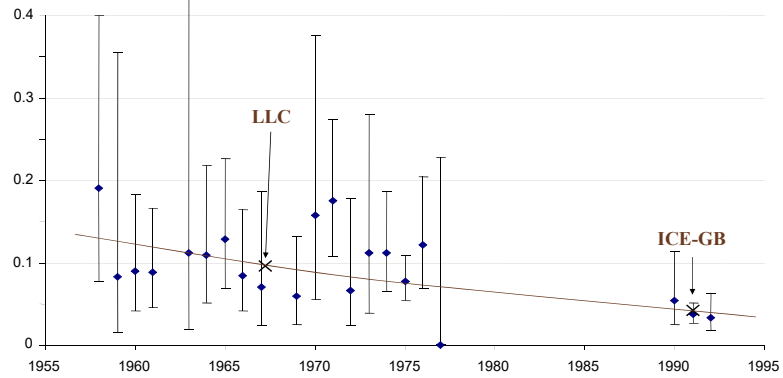
Survey of English Usage

- 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

現在進行形の増加（DCPSE）



“shall” の減少 （DCPSE）



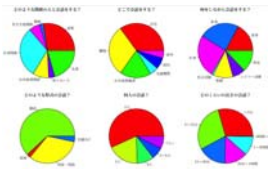
2000年代の音声コーパス

- 『日本語話し言葉コーパス』CSJ
  - ◆Corpus of Spontaneous Japanese

- ●2004年公開。651時間、752万語の自発音声を収録。
  - ◆90%は独話（学会講演・模擬講演）、10%は対話、再朗読など
- ●豊富なアノテーション
  - ◆転記テキスト、形態論情報、節境界ラベル、統語構造情報、分節音ラベル、韻律ラベル、話者情報…
- ●開発目的
  - ◆音声認識技術の精度向上、自発音声の言語学的研究

2000年代の音声コーパス

- 『日本語日常会話コーパス』（CEJC）
  - ◆Corpus of Everyday Japanese Conversation
- ●2016年 構築開始。2022年、200時間分を公開予定。
- ●さまざまな場面における日常会話をビデオ収録。
  - ◆家族、友人、同僚、知らない人…
  - ◆自宅、職場、レストラン、交通機関…
  - ◆食事、打ち合わせ、団らん、旅行…
- ●アノテーション
  - ◆転記テキスト、形態論情報、発話単位、



2000年代の音声コーパス

- 『昭和話し言葉コーパス』
- 2016年 構築開始。2020年、50時間分を公開予定。
- 1950～1960年代に国語研で録音された日常談話資料
  - ◆会話：生活場面の様々な日常談話
  - ◆独話：国語研で開催した講演会
- 「話し言葉の経年変化」の研究
  - ◆アクセント、イントネーション
  - ◆語彙、文法、談話構造

3. 『昭和話し言葉コーパス』

『昭和話し言葉コーパス』のねらい

- 古い時代の音声資料を、コーパスとして編纂する。
  - ◆音声データ ＝ 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

『昭和話し言葉コーパス』の構築手順

1. 音声資料の収集（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

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

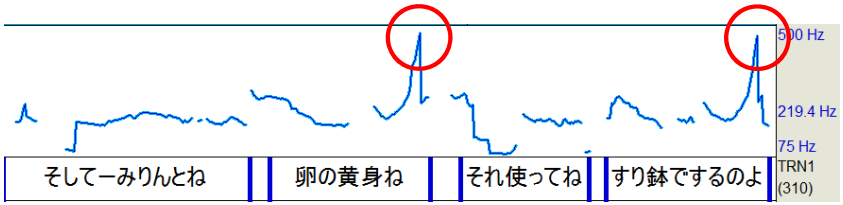
林四郎氏 講演（1968年）

33ファイル、約28.7時間
約146,000字

イントネーションの変化

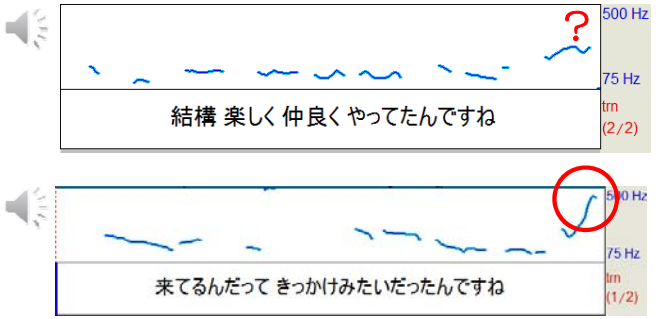
●句末・発話末の急激な上昇イントネーション

「三人の女性」（昭和話し言葉コーパス、1957年録音）



イントネーションの変化

●句末・発話末の上昇イントネーション（CSJ、2000年）



イントネーションの変化

●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

文法形式の変化（１）

● 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、1959年録音）
● ラジオニュースの書き方というような本を見まするとと、
（波多野完治：1905年生、1955年録音）
● 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、1955年録音）

文法形式の変化（１）

● 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、1959年録音）
● えー、選挙の、おー、放送を、おー、聞いておりますと、
（波多野完治：1905年生、1955年録音）
● 非常に、あの、時代的な差もありますけれども、
（林 大：1913年生、1955年録音）

同一話者による文法形式のゆれ

文法形式の変化（１）

	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%)

文法形式の変化（１）

- 「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

文法形式の変化（１）

- 「まする」の出現条件

	岡田コレクション			昭和話し言葉コーパス	CSJコア
	大正期	昭和1ケタ	S10年代		
	23,002語	46,998語	49,070語	358,382語	447,677語
ます	271 86.6%	752 89.8%	903 92.9%	5,280 98.1%	5,604 100%
まする	42 13.4%	85 10.2%	69 7.1%	104 1.9%	0 0%

文法形式の変化（２）

- 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

今後の課題：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。

今後の課題：発話の重複による問題

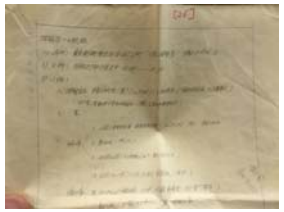
●発話の重複で聞き取れない、発話者の割り当てが困難

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。
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：何でも。若い人、いるから。あんたんとこいったら、みんな●。
？：●。
7：若い人のおっしゃることでもね。
1：おばあさんですからね。（笑）
2：こちら長いんですか。
7：なんでも、その、★キョウ★。
1：ええ。
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。

今後の課題：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

4. 通時音声コーパスの可能性

通時音声コーパスとは何か？

●通時コーパス
- ◆異なる時代を対象とした複数のサブコーパスで構成される
- ◆言語の歴史的変化をコーパスに基づいて分析することが目的
- ◆通常は書き言葉コーパス

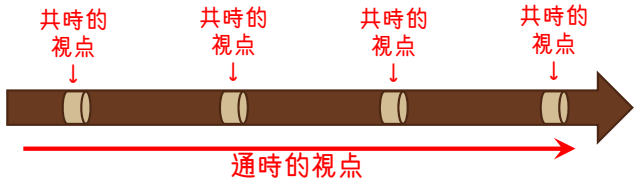
●音声コーパス
- ◆音声ファイルと転記テキストを含む大規模なデータ
- ◆音声学、音韻論、話し言葉の文法、会話分析などで利用
- ◆書き言葉コーパスよりも構築が困難

通時音声コーパスとは何か？

- ●多時代に渡る録音資料の集積
  - ◆ 実際には、1900年以降の独話、1950年以降の会話
- ●話し言葉の通時的な変化を研究することが目的
  - ◆発音、アクセント・イントネーション、文法、意味...
- ● 定量的な分析に耐える、大規模なデータサイズが必要
  - ◆しかし、現存する録音資料は限られている！
- ●豊富なアノテーション（転記、時間情報、形態論情報）メタ情報（録音年月日、場所、話者、属性...）が必要
  - ◆音声学、音韻論、文法論、社会言語学、変異

共時的視点と通時的視点

- ●共時的視点
  - ◆ある特定の時代における言語状態を記述
- ●通時的視点
  - ◆時間の流れにおける言語の変化を記述



共時的視点と通時的視点

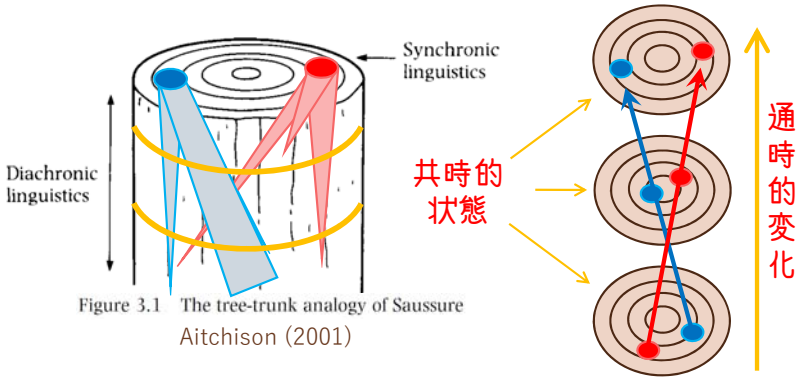


Figure 3.1 The tree-trunk analogy of Saussure

Aitchison (2001)

参考文献

- 相澤正夫・金澤裕之編 (2016)『SP盤演説レコードがひらく日本語研究』笠間書院.
- 小磯花絵編 (2015)『話し言葉コーパス 設計と構築』（講座 日本語コーパス 3）朝倉書店.
- 前川喜久雄編 (2013)『コーパス入門』（講座 日本語コーパス 1）朝倉書店.
- 丸山岳彦 (2014)「現代日本語の多重的な節連鎖構造について —CSJとBCCWJを用いた分析」. 石黒圭・橋本行洋編.『話し言葉と書き言葉の接点』, 93-114. ひつじ書房.
- 丸山岳彦 (2014)「昭和の日常のことば」『日本語学』2014年12月号, 4-15. 明治書院.
- 丸山岳彦 (2015)「「通時音声コーパス」は可能か」『第8回 コーパス日本語学ワークショップ 予稿集』, 29-36. 国立国語研究所.
- 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』第282号, 39-55.