『日本語日常会話コーパス』バランスの検証と研究の可能性

小磯　花絵

国立国語研究所

1

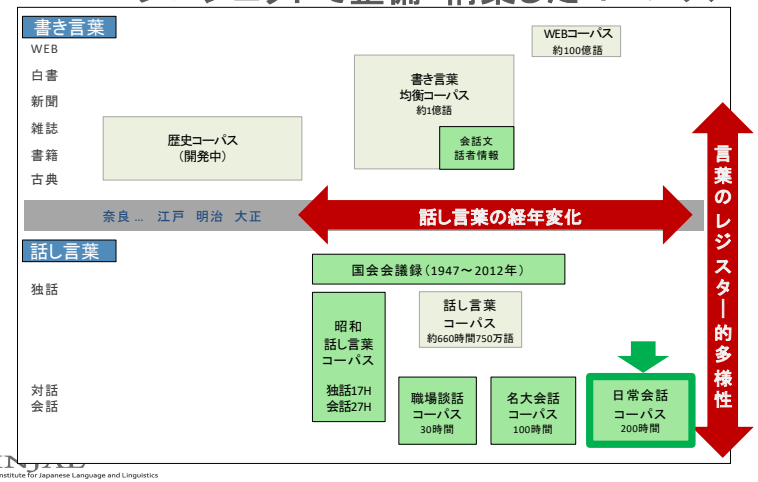
プロジェクトの概要

- 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。

2

プロジェクトで整備・構築したコーパス



3

『日本語日常会話コーパス』設計と構築

4

『日本語日常会話コーパス』の特徴

❖多様な場面・多様な話者の会話を記録

❖日常場面で自然に生じるリアルな活動を記録

❖ 音声データ・文字テキストに加えて映像データも公開

NINJAL National Institute for Japanese Language and Linguistics

5

日常会話コーパスの収録法

■ 個人密着法　　185時間

- ✓ 性別・年齢の観点からバランスを考慮して選別された協力者に収録依頼
- ✓ （男女×年齢5世代×各4-5人=40-50人、職業偏らないよう配慮）
- ✓ 機材機器等を2-3か月ほど貸し出し，協力者の日常生活で自発的に生じるリアルな会話を記録（1協力者あたり平均約15–18時間収録）
- ✓ コーパス構成比や倫理的問題等を考慮してコーパスに含める会話を選別
  - 1協力者あたり約4-5時間を選別，計160-200時間（目安）

■ 特定場面法　　15時間

個人密着法では収録の難しい場面

- ✓職場での会議・会合
- ✓未成年者（中高生）中心の会話

NINJAL National Institute for Japanese Language and Linguistics

6

個人密着法　調査協力者の内訳

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	学生	4.2h	T009	学生	6.0h
	T006	学生	4.3h	K003	大学生	4.4h
	T022	先生	3.7h	K009	会社員・公務員等	4.2h
	K007	先生	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	3.6h	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	4.8h
	T019	先生	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	先生	4.2h	K008	自営業・自由業	4.6h
	T018	先生	4.2h	K011	会社員・公務員等	4.5h
60歳～	T013	先生	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

NIN National Institute

7

映像の収録（基本）

Kodak PIXPRO SP360 4K

会話者の中心に360度撮影可能なカメラを配置

GoPro Hero3+

会話を俯瞰的に記録するカメラを1～2台配置

NI National Inst

8

音声の収録

話者ごとにICレコーダーを首から下げた
フォルダーに入れて装着し、
当該話者の音声を中心に収録

個人ICレコーダー

Sony ICD-SX734

Sony ICD-SX1000

中央ICレコーダー

会話の場の中央に置いて
会話全体を収録

NINJAL National Institute for Japanese Language and Linguistics

9

データ規模

時間	200時間
会話数	577会話
話者数 (延べ)	1675名
話者数 (異なり)	862名
語数	240万語

NINJAL National Institute for Japanese Language and Linguistics

10

『日本語日常会話コーパス』バランスの検証

NINJAL National Institute for Japanese Language and Linguistics

11

会話行動調査

- 目的：日常会話の多様性を明らかにし，それに立脚して多様な日常会話をバランスよく納めたコーパスを設計
- 実施時期：H26.11〜H27.2
- 対象：243人
  （年齢・性別バランス）
- 調査日：平日2日・休日1日
  （計3日／1人）
- 総会話数：9272会話

① どんな会話か

<自由にメモしてください> レストランで1人でランチ中、旅行について友人と電話で相談

② いつ （1つ選択）

☐ 午前 ☑ 午後 ☐ 夜（午後6時頃〜）

③ どのくらい （1つ選択）

☐ 5分未満 ☐ 5〜15分 ☑ 15〜30分 ☐ 30分〜1時間

☐ 1〜2時間 ☐ 2〜5時間 ☐ 5〜10時間 ☐ 10時間以上

④ どこで （1つ選択）

☐ 自宅 ☐ 職場・学校 ☑ 公共商業施設 ☐ 交通機関

☐ それ以外の屋内 ☐ それ以外の屋外

⑤ だれと （あてはまるものそれぞれに人数を記入）

家族：＿人 親戚：＿人 先生・生徒：＿人

仕事・学業関係：＿人 公共商業関係：＿人

友人・知人：1人 顔見知り・見知らぬ人：＿人

⑥ 何をしながら （1つ選択）

☑ 食事 ☐ 家事・雑事 ☐ 身周りの用事 ☐ 療養

☐ 仕事・学業 ☐ 業務外・課外活動 ☐ 社会参加

☐ レジャー活動 ☐ 付き合い ☐ 移動 ☐ 休息

⑦ どんな種類 （1つ選択）

☐ 雑談 ☑ 用談・相談 ☐ 会議・会合 ☐ 授業・レッスン・講演

⑧ その他 （あてはまるものすべて選択）

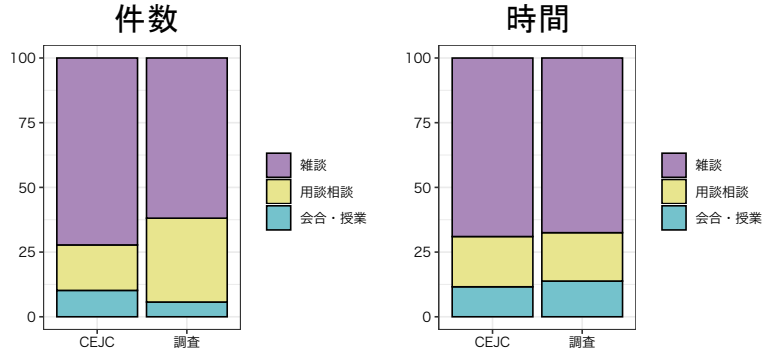
☑ 電話・スカイプなどの遠隔での音声・映像会話

☐ 外国人を含む会話 ☐ 外国語を含む会話

NINJAL National Institute for Japanese Language and Linguistics

12

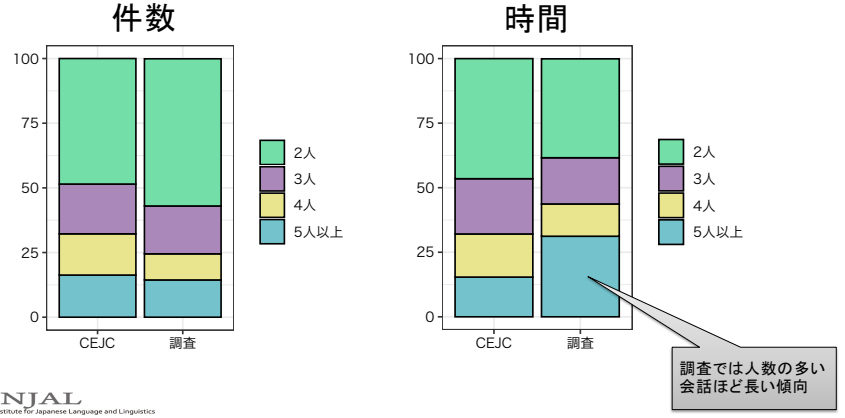
行動調査との比較：会話形式



NINJAL National Institute for Japanese Language and Linguistics

13

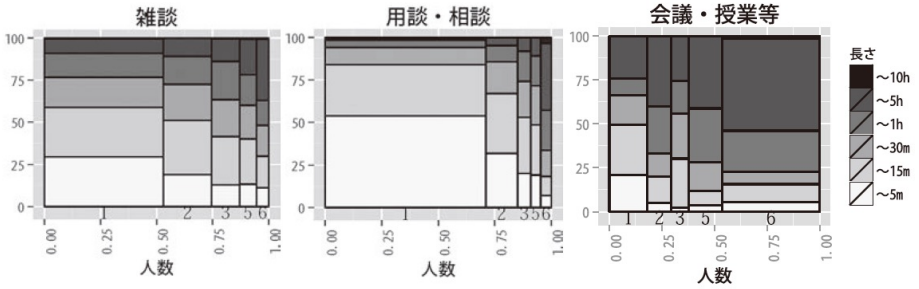
行動調査との比較：話者数



NINJAL National Institute for Japanese Language and Linguistics

14

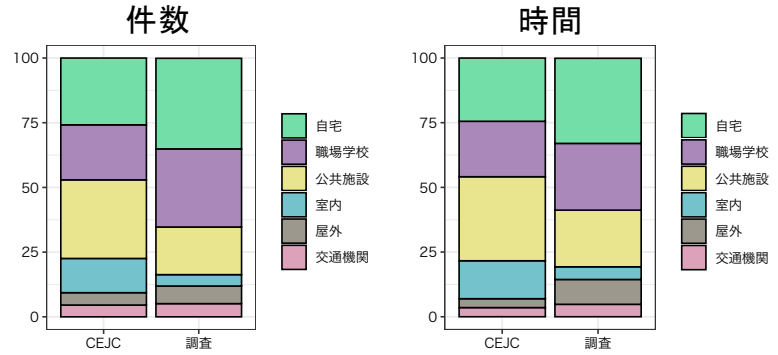
調査結果：会話の話者数と会話時間との関係



NINJAL National Institute for Japanese Language and Linguistics

15

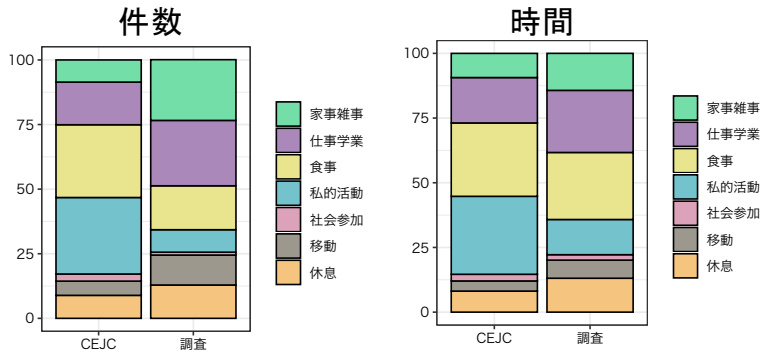
行動調査との比較：場所



NINJAL National Institute for Japanese Language and Linguistics

16

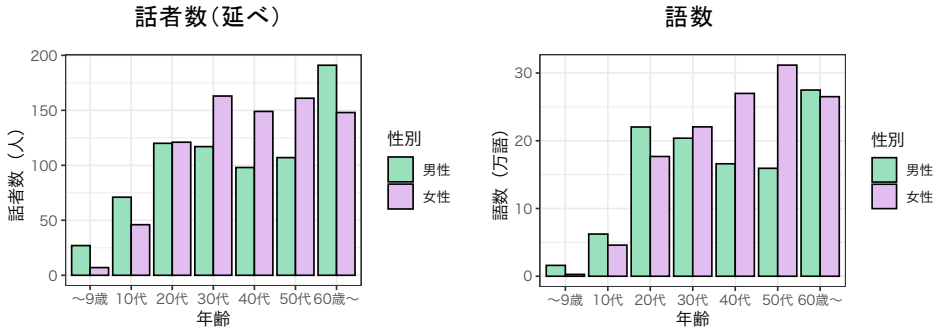
行動調査との比較：活動



NINJAL National Institute for Japanese Language and Linguistics

17

性別・年齢別の話者数・語数の分布



NINJAL National Institute for Japanese Language and Linguistics

18

『日本語日常会話コーパス』を活用した研究の可能性

2つの研究事例の紹介を通して

NINJAL National Institute for Japanese Language and Linguistics

19

日本語日常会話コーパスを活用した研究の可能性

① 多様な話者・場面の会話を収めたコーパスを活用した研究の可能性
- ➢スピーチレベル（丁寧体・普通体の選択）

② 書き言葉・話し言葉を含む多様なレジスターのコーパスを活用した研究の可能性
- ➢ 並列節を導く接続助詞「が」「けれども」

NINJAL National Institute for Japanese Language and Linguistics

20

分析方法

- 分析対象：述語を動詞・形容詞とする69130発話（従属節は除く）
- 丁寧体・普通体の別 ：助動詞「です」「ます」の有無により分類
  待遇レベルの低さから以下は対象外
  - 確認要求表現としても使われる「でしょ(う)」（中北2000, 佐竹2016）
  - 口語表現・母親語「(っ)す」（行くっす）、「(っ)しょ」（やるっしょ）、「でしゅ・でちゅ」（いいでちゅか）
- 終助詞の有無 （対象：ね・よ・ぞ・ぜ）
- 丁寧さ・対人のモダリティーにもとづくスピーチレベルの分類

① 丁寧体・終助詞なし
② 丁寧体・終助詞あり
③ 普通体・終助詞なし
④ 普通体・終助詞あり

高 ↕ 低 スピーチレベル

丁寧体に終助詞「ね」「よ」が付くと親しみは増すが丁寧さの度合は低下（益岡 1991）

終助詞を用いた普通体は待遇レベルが低いと認識されやすい（佐藤・福島 1998）

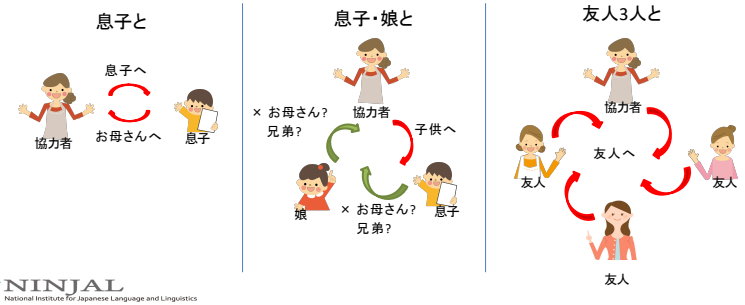
各会話の話者ごとに各タイプの比率を算出して分析に利用

21

分析方法

相手との関係性の認定に関する補足

発話の受け手（誰に向けて発話したのか）の情報が記されていないため
会話者間の関係性の情報から自動的に特定できる場合に限定



NINJAL National Institute for Japanese Language and Linguistics

22

関係性別 データ数

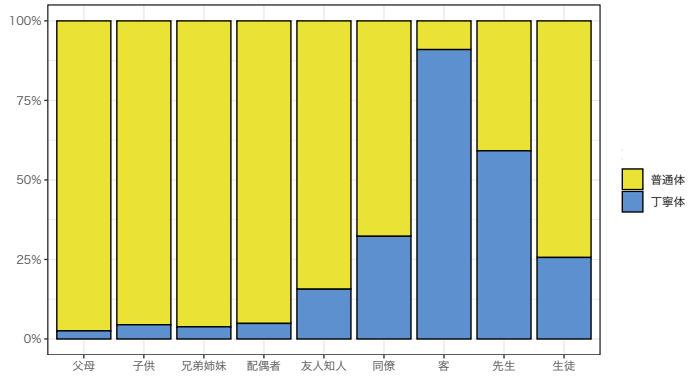
関係性	人数	発話数
父母	50	4840
子供	34	3310
兄弟姉妹	10	1659
配偶者	33	3846
友人知人	153	37285

関係性	人数	発話数
同僚	51	10984
客	85	1626
先生	11	668
生徒	16	2114

NINJAL National Institute for Japanese Language and Linguistics

23

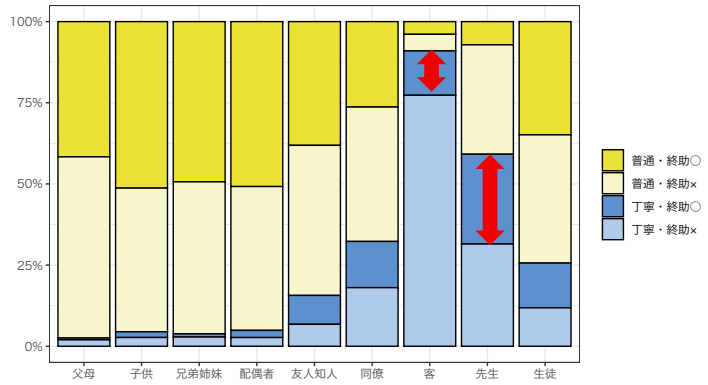
相手との関係性とスピーチレベル（丁寧体・普通体）



NINJAL National Institute for Japanese Language and Linguistics

24

相手との関係性：丁寧体・普通体×終助詞の有無



NINJAL National Institute for Japanese Language and Linguistics

25

店員と客の会話例

撮影禁止

女将が料理を受ける場面 → 丁寧体が主

客３　お母さん すいません。
女将　はい。
客３　揚げパン 二つ:。
女将　はい。
客３　また 切っ 切ってもらって:。
女将　揚げパン 二つ。　中立形(注文繰り返し)
客２　じゃ (D チ)。
女将　二個でいいですか。　丁寧体
客３　二個でいいです。
女将　はい。
客２　チーズスティックもお願いします。
女将　チーズスティック。　中立形(注文繰り返し)
客２　はい。
（略）
客４　すみません。あと 角煮:。
女将　角煮。　中立形(注文繰り返し)
客４　はい。
女将　はい。ありがとうございます:。　丁寧体

ていねい調の中に中立形が現れる場合
前文の一部を繰り返したり否定したりするだけの文が中立形になりやすい（野田 2003）

女将が収録器具を見つけて客に話しかける（非接客場面）→ 普通体に切りかわる

女将　えー。何:これ。
客４　え。
客１　ちょっとね。
客３　もう 当然の 当然の疑問ですよ。
客１　あー。まあ まあ まあ まあ まあ まあ まあ。
客２　確かに 確かに。
客３　当然の疑問ですよね。
女将　びっくり。
大将　さっきから気になっちゃって。
女将　かわいい。これ。

画面キャプチャは行わないようお願いします

26

宿題中の小学生の子供と母親の会話

息子	学習目標。算数を得意になる。だめ?。	
母親	んー?。学習目標ってこっちだよ?。 自分一人でできるようになることだから やっぱ本を読むことじゃない**です**かね。	普通体 丁寧体
息子	算数がトク。二つにする。	
母親	一個でいい。一個で。	普通体
息子	じゃあ毎日本を読むのは（母親：うん）やだ。	
母親	なんで。@笑いながら発話	
息子	一日飛ばすならいい。	
母親	だめだよ。@笑いながら発話 毎日イッ毎日やることが大事なんだから 一日一冊とかにしたら?。	普通体 普通体
息子	じゃあ本を読むようにする?。	
母親	うん。頑張れ。	普通体

NINJAL National Institute for Japanese Language and Linguistics

27

先生と学生との会話例

撮影禁止

大学生の運営メンバー数名と担当教師で運営する高校の講座の企画などについて、担当教師（高校時代の部活の顧問でもある）と相談。

この場面では日程がかわったことを誰が伝えるかについて語っている。

学生	でも 若干変わったじゃない**です**か。	丁寧体 学生
学生	あの 先生の振り返り:とかでなくなったので。	
先生	ああ ああ ああ。それはどっちでも**いい**よ。	普通体 先生 +よ
学生	そうゆうのは 伝えたほうが:。	
先生	うん。ど:っちでも。どっちが**いい**かな。	普通体 先生
学生	まずは その 四人でやってもらうってゆうのも:	
	どっちから伝えたほうが**いい**です**か**ね。	丁寧体 学生 +ね
学生	自分がゆったほうが**いい**です**か**?。	丁寧体 学生
学生	うーん。	
学生	(R 大槻)先生 お願いしてもいい**です**か?。	丁寧体 学生
先生	うん。**いい**よ。	普通体 先生 +よ

NINJAL National Institute for Japanese Language and Linguistics

28

先生と学生との会話例

教員と学生が講座の企画について打ち合わせの途中に話が脱線し，
他のメンバー(女子)の印象について話している

先生	あの子たち いいよ:。
学生	はい。
先生	あ:れ:いいよ。
先生	あれはね:。
学生	でもみんな あれですね。
学生	キャラってゆうとあれかもしれないけど。
学生	違いますよね。
学生	全然 色が違うので:。
先生	ちょっと: 僕はもう(R 真帆)ちゃ:んお気に入り。
学生	いや。でも。わかりますけど。
先生	あの: なん なんかわからないか。
学生	わ:かります わかります。
先生	その 顔。
先生	うーん。ごめんな。
先生	その 美人かどうか って別として かわいらしいです。
学生	かわいいですよね:。

普通体 先生 ＋よ
普通体 先生 ＋よ
丁寧体 学生 ＋よね
普通体 先生
丁寧体 学生
丁寧体 先生
丁寧体 学生 ＋よね

NINJAL National Institute for Japanese Language and Linguistics

29

助動詞「です」の縮約形「っす」

- 若い男性が使用
- 丁寧体を使う相手で親しい間柄の相手（先輩など）に使用
- 丁寧体を使わない相手（家族や同級生など）には使用しない
  - 丁寧さをある程度保ちつつ
  - よそよそしさを回避し
  - 親近性を表す

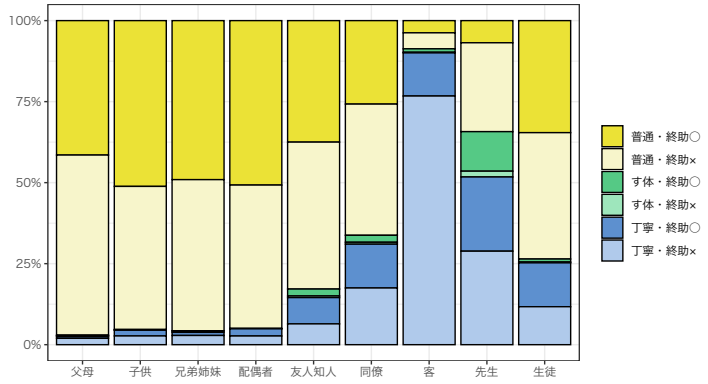
① 丁寧体・終助詞なし
② 丁寧体・終助詞あり
③ す体・終助詞なし
④ す体・終助詞あり
⑤ 普通体・終助詞なし
⑥ 普通体・終助詞あり

高 ↕ 低 スピーチレベル

NINJAL National Institute for Japanese Language and Linguistics

30

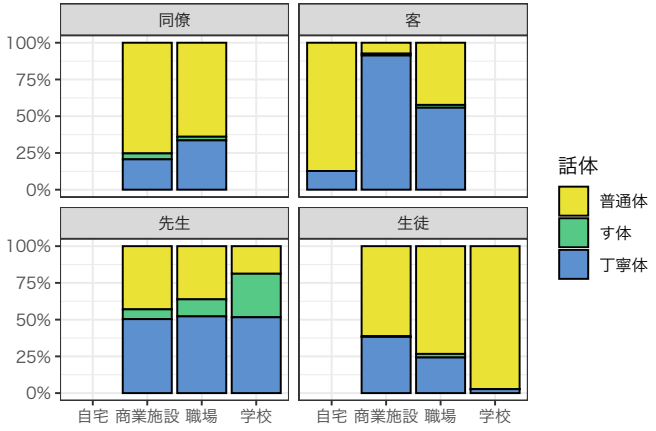
相手との関係性：丁寧体・す体・普通体×終助詞の有無



NINJAL National Institute for Japanese Language and Linguistics

31

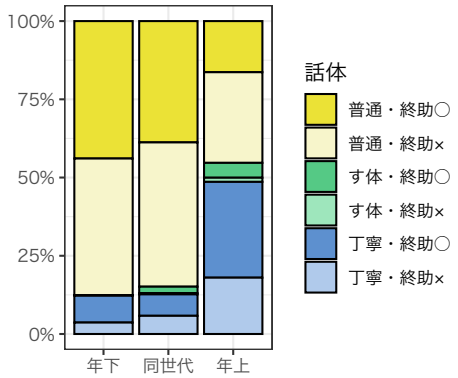
相手との関係性・場所：丁寧体・す体・普通体



NIN National Institute fo

32

友人知人との会話：世代との関係



NINJAL National Institute for Japanese Language and Linguis

33

日本語日常会話コーパスを活用した研究の可能性

① 多様な話者・場面の会話を収めたコーパスを活用した研究の可能性

➢スピーチレベル（丁寧体・普通体の選択）

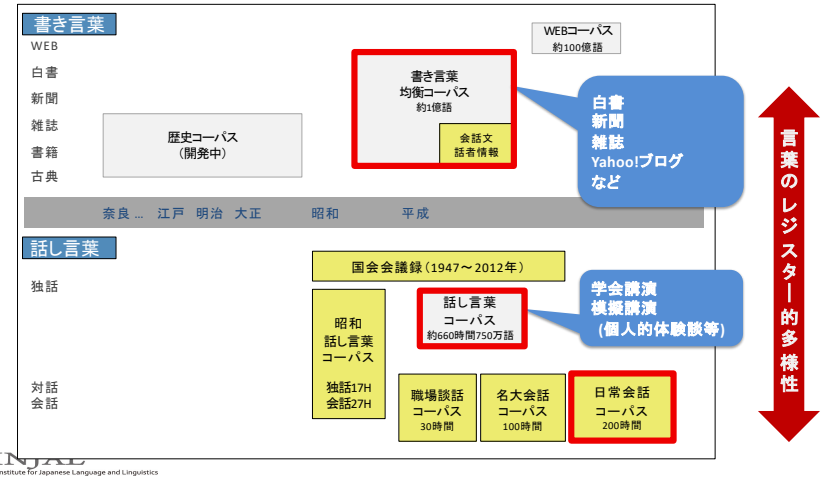
② 書き言葉・話し言葉を含む多様なレジスターのコーパスを活用した研究の可能性

➢ 並列節を導く接続助詞「が」「けれども」

NINJAL National Institute for Japanese Language and Linguistics

34

プロジェクトで整備・構築するコーパス



NINJAL National Institute for Japanese Language and Linguistics

35

「が」と「けれども」類

森田（1980）

✓ 「が」は書き言葉的、「けれども」（類）は話し言葉的

永田・茂木（2007） 意識調査＋CSJの分析

✓ 「が」は書き言葉的、「けど」は話し言葉的

✓ 「が」「けれども」は改まり度が高く、「けど」は改まり度が低い

✓ 講演において「けれど」の使用頻度は低い

丸山（2013） BCCWJ + CSJの分析

✓くだけた文体の書き言葉では「けど」の使用が多い

✓書き言葉では「けども」より「けれど」の方が多く用いられる

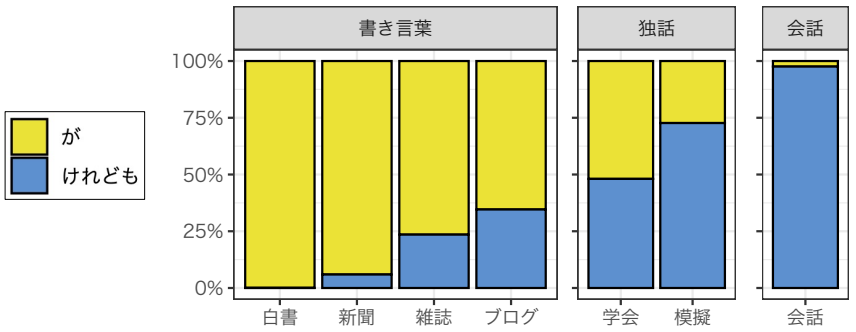
話し言葉では「けれど」より「けども」の方が多く用いられる

NIN National Institute for Japanese Language and Linguistics

36

BCCWJ・CSJ・CEJCにおける「が」「けれども」類の分布

日常会話では
ほぼ「けれども」類

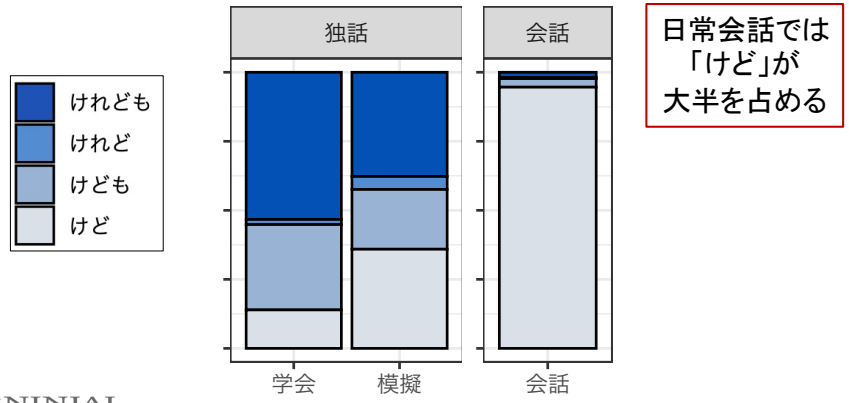


NINJAL National Institute for Japanese Language and Linguistics

37

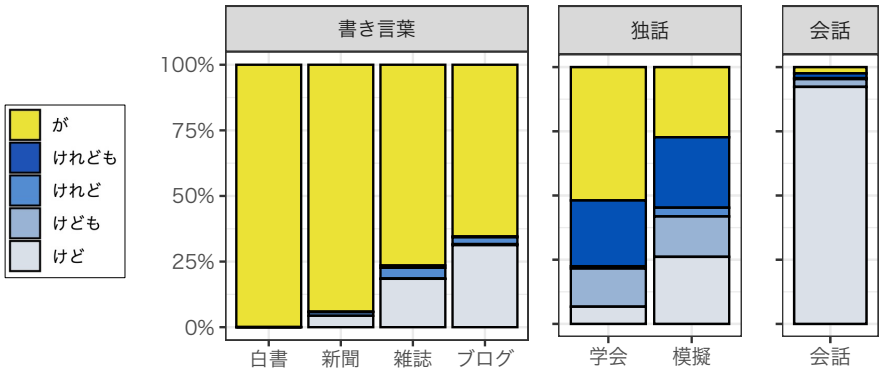
BCCWJ・CSJ・CEJCにおける「けれども」類の内訳

日常会話では
「けど」が
大半を占める



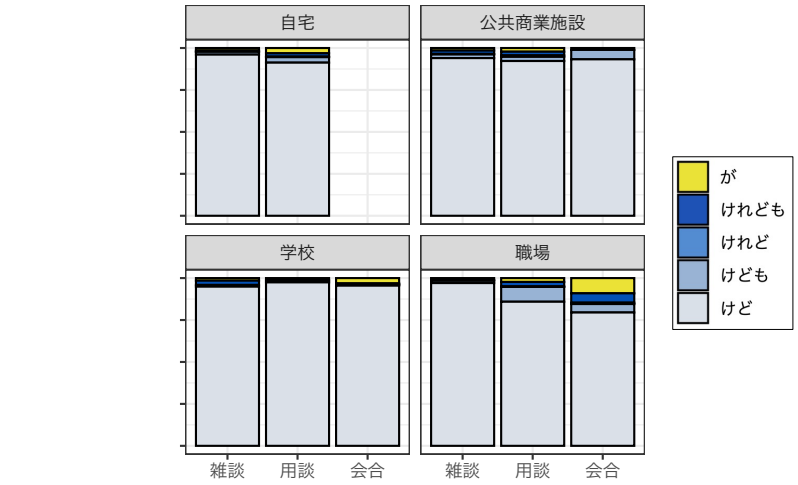
NINJAL National Institute for Japanese Language and Linguistics

38



NINJAL National Institute for Japanese Language and Linguistics

39



NINJAL National Institute for Japanese Language and Linguistics

40

「けれども」類の語形

宮内(2007)

- 江戸後期の資料（文学作品の会話文）
  - 「けれども」類が見られるように
  - 「けど」は見られず「けれど」が多く現れる
- 明治期では「けど」が散見されるように
- 明治期までの対象資料に「けども」は見られない

土井(1969)

- 大正以降では「けれど」より「けど」の使用が優勢に
- 口頭語で「けども」はわずかに散見されるのみ

NINJAL National Institute for Japanese Language and Linguistics

『日本語日常会話コーパス』

公開：2022年3月末 (予定)

情報：以下のページでご案内

https://www2.ninjal.ac.jp/conversation/cejc.html

構築に携ったメンバー

浅原 正幸	天谷 晴香	石本 祐一	居關 友里子	臼田 泰如	大村 舞
柏野 和佳子	門田 圭祐	川端 良子	河村 美雪	菊池 英明	汲田 尚子
小磯 花絵	角田 ゆかり	十河 則子	田中 真理子	田中 弥生	田邊 牧子
寺崎 温子	伝 康晴	徳永 弘子	西川 賢哉	松窪 英子	牟田 浩子
森本 桂子	山縣 智子	山口 昌也	山下 華代	山田 高明	若狭 絢
若松史恵	渡邊 友香	渡部 涼子			

NINJAL National Institute for Japanese Language and Linguistics

謝辞

収録にご協力くださったみなさまに深く感謝します

NINJAL National Institute for Japanese Language and Linguistics