NINJAL
National Institute for Japanese Language and Linguistics

『日本語日常会話コーパス』の構築

小磯　花絵（国語研究所）

2016年9月1日
シンポジウム「日常会話コーパス」I

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

主要な日本語会話コーパス（日本語母語話者主対象）

コーパス名	規模	概要	音声
名大会話コーパス	161名100時間	親しい者同士の雑談	無
女性のことば・職場編 男性のことば・職場編	各21名	職場のフォーマル・インフォーマルな場面の自然談話	無
BTSによる多言語話し言葉コーパス	249会話70時間	友人同士の雑談、教師学生面談会話、電話会話など	一部
CALL HOME Japanese	120会話20時間	アメリカ在住日本人と国内の家族・友人との電話会話	有
CallFriend Japanese	31会話	アメリカ在住の日本人同士の電話会話	有
さくらコーパス	18会話	大学生の会話（話題指定）	有
千葉大学3人会話コーパス	12会話2時間	大学生の会話（話題指定）	有

- 収録のために集められた状況での会話を対象とするものも多い
- 親近者同士の雑談，電話会話，キャンパスでの会話など，話者や場面などに偏りが見られる
- 音声データを提供していないものが多い、映像データを提供するものはほとんどない
- 形態論情報などの基本的なアノテーションが提供されないものが多い

本プロジェクトで目指す会話コーパス

- 多様な話者・場面を扱う
  → ①コーパス設計
- 日常場面において、当事者たちの動機・目的に基づき自発的に生じたリアルな活動を記録
  → ②収録法
- 映像データ・音声データを記録・公開する
  → ③映像・音声データの収録方法
- 形態論情報などの基本的なアノテーションを施す
  → ④アノテーション

NINJAL
National Institute for Japanese Language and Linguistics

『日本語日常会話コーパス』
Corpus of Everyday Japanese Conversation, CEJC

均衡性を考慮した大規模日常会話コーパス（200時間）の構築・公開

収録　600～800時間　　日常場面で自発的に生じた多様な日常会話

コーパス 200時間　［人手作成］転記テキスト
［自動付与］単語（短単位・長単位）・係り受け・発話単位

コア 20時間　［人手修正］単語・係り受け・発話単位
［人手付与］談話行為・韻律情報

NINJAL
National Institute for Japanese Language and Linguistics

コーパス設計

NINJAL National Institute for Japanese Language and Linguistics

予備研究：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）

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

調査実施状況

- 応募人数：266人
- 調査返送人数：243人（729日分，3日/1人）
- 内訳

性別・年代

	20代	30代	40代	50代	60代以上	計
女性	25	25	25	24	25	124
男性	23	22	25	24	25	119
計	48	47	50	48	50	243

世帯員数

1人	2人	3人	4人	5人	6人	7人
35	69	63	55	14	6	1

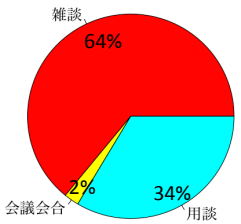
職業

職業	人数
会社員・公務員など	99
自営業	12
パート・アルバイト	32
学生	36
専業主婦	33
無職・定年退職者	18
その他	13

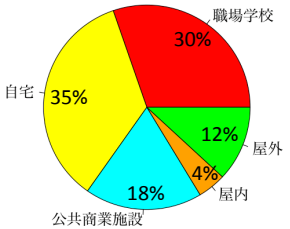
NINJAL National Institute for Japanese Language and Linguistics

調査結果①

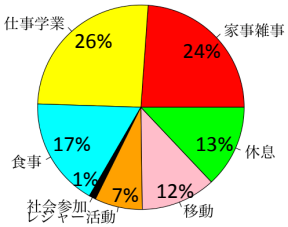
形式



場所

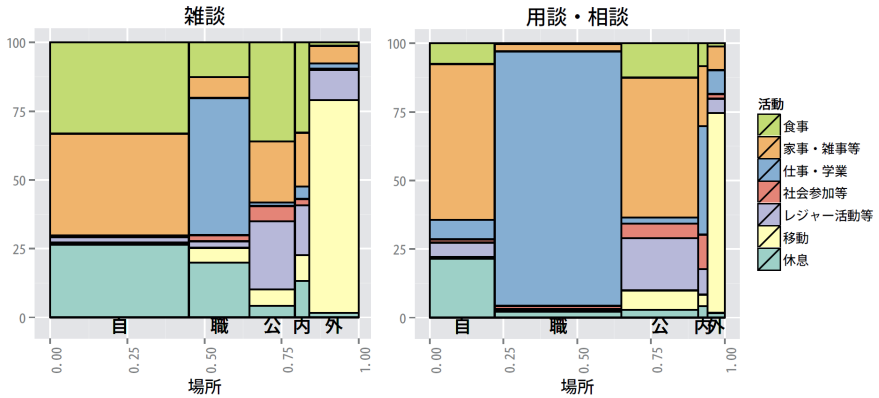


活動



NINJAL National Institute for Japanese Language and Linguistics

調査結果②



自:自宅, **職**:職場・学校, **公**:公共商業施設, **内**:それ以外の屋内, **外**:それ以外の屋外・交通機関

NINJAL National Institute for Japanese Language and Linguistics

調査結果に基づくコーパス設計方針

- 会話の形式の構成比を参考に全体の構成比を決定

  （例） 雑談:用談相談:会議会合 ＝ 6割 : 3～3.5割 : 0.5～1割

  （時間数）6.5～7割 : 2割 : 1～1.5割

- 場所と活動の観点から，ある程度出現するカテゴリーを対象に構成比を決定

		形式								
		雑談			用談・相談			会議・授業等		
活動＼場所		私的	公的	それ以外	私的	公的	それ以外	私的	公的	それ以外
件数	食事・休息	30%	5%	10%	5%		5%	5%		
	仕事・学業	15%	10%	5%	15%	40%	20%	0%	70%	10%
	その他	5%		20%	0%	0%	15%	0%	0%	15%
長さ	食事・休息	30%	5%	20%	10%		5%	5%		
	仕事・学業	10%	5%	5%	10%	50%	15%	0%	70%	10%
	その他	5%		20%	0%	0%	10%	0%	0%	15%

➡ この比率を遵守するのではなく、大きな偏りが生じないよう、一つの指針とする

収録法

NINJAL National Institute for Japanese Language and Linguistics

British National Corpus の Spoken Part

- 規模:全体1億語のうち1000万語
- 設計
  - 人口構成別に収集:日常的な会話の記録（420万語）
    - 成人（16歳以上）124名
    - 話者の属性（性別・年齢5区分・社会階層4区分・地域12区分）でバランス
  - 文脈別に収集:公的な場での会話や独話（615万語）

教育・教養	独話	講義、講演、実習、ニュース、など
	対話	授業での先生・生徒間のやりとり、など
ビジネス	独話	企業発表、デモンストレーション、など
	対話	商用の打合せ、インタビュー、など
団体	独話	政治演説、説法、など
	対話	議会審議、など
レジャー	独話	スポーツ実況中継、テーブルスピーチ、など
	対話	放送の対談番組、クラブミーティング、など

NINJAL National Institute for Japanese Language and Linguistics

日常会話コーパスの収録方法

■ 個人密着法

- ✓ 性別・年齢など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（首都圏在住者、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　（本プロジェクト外で収録の可能性あり）

個人密着法では収録の難しい場面

- ✓ 職場での会議・会合
- ✓ 店舗での接客場面、など

NINJAL National Institute for Japanese Language and Linguistics

映像・音声データの収録方法

NINJAL National Institute for Japanese Language and Linguistics

個人密着法による収録要件

- ✓ 会話に参加する人（収録協力者）自身に収録
- ✓ 屋内だけでなく屋外を含む多様な日常場面の会話が対象

⇒ 小型・軽量で簡単に操作できる収録機材

- ➢ 会話を含む活動を詳細に観察・理解できるデータ

⇒ 音声だけでなく映像の記録が不可欠

NINJAL National Institute for Japanese Language and Linguistics

映像の収録機器：屋内の場合

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

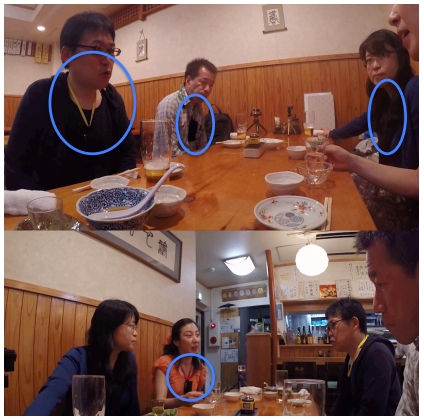
映像の収録機器：屋外の場合

ウェアラブルカメラ Panasonic HX-A500 を利用
（収録協力者のみ装着し会話の状況を撮影）

NINJAL
National Institute for Japanese Language and Linguistics

音声データの収録機器：屋内の場合

- 会話者ごとにICレコーダーを首から下げたフォルダーに入れて装着

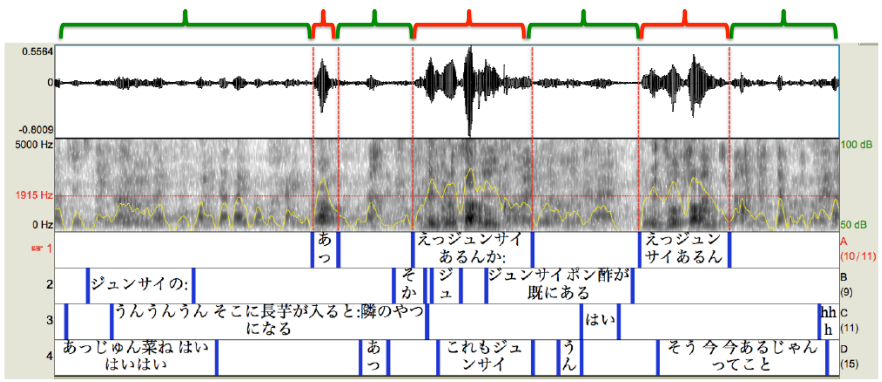


Sony ICD-SX734

あご下
約15cm

NINJAL
National Institute for Japanese Language and Linguistics

ほかの話者の声　話者Aの声



NINJAL
National Institute for Japanese Language and Linguistics

同意書・フェイスシート

調査協力者に、会話に参加する方に対して、以下のことも担当していただく

①収録の趣旨などを説明
②収録・公開に関する同意を得る（同意書に署名）
③フェイスシート（誕生日・性別・出生地・居住歴・職業）に記入してもらう

NINJAL
National Institute for Japanese Language and Linguistics

アノテーション

NINJAL
National Institute for Japanese Language and Linguistics

アノテーションの種類

- 転記テキスト
- 発話単位情報
- 形態論情報（短単位・長単位）
- 係り受け情報

原則、自動アノテーション
コア（20時間程度）は人手修正

- 談話行為情報（相互行為班と共同）
- 韻律情報

コア（の一部）に人手付与

NINJAL
National Institute for Japanese Language and Linguistics

転記の概要

- 転記作業はELANを用いて動画と音声を参照しながら行う
- 転記単位ごとに開始・終了時間を認定し、音声波形に割り付けながら転記する
- 転記単位は次のいずれかで分割する
  - 知覚可能な休止がある場合
  - 発話単位の切れ目がある場合
  - 異なる音種（言語音・単独の笑い・泣き・歌）が続く場合
- 原則として発話内容を漢字仮名じりで表記する
- 転記単位のうち発話単位の切れ目に句点「。」を記す
- 各種タグを用いて会話に生じる現象・発音を表現する

NINJAL
National Institute for Japanese Language and Linguistics

会話コーパス構築班の転記で用いるタグ

タグ	概要	使用例
:	非語彙的な母音の引き伸ばし	すご：い，デー：タ
%	非語彙的な音の詰まり	す%ごい，解%析
?	疑問上昇調	行きます?，コップ?
(D)	語の言いさし	(D こ)明日から
(W)	言い誤り・発音の怠け等の一時的な発音エラー	(W こえ|これ)，(W ぎーつ|技術)
(L)	笑いが生じている箇所	(L)，これ(L なんですけど)
(C)	泣きが生じている箇所	(C)，(C なにが)
(S)	歌が生じている箇 所	(S)，(S ふるさと)，(S ヘイヘイホー)
(E)	口になにかを含んだ状態で発話をしている箇所	(E あそこすごかった:)
(M)	表記原則から外れた方法で表記せざるを得ない場合	すごい を(M すっごい)と発音する
(K)	何らかの原因で漢字表記できない箇所	(K あ:ま:|甘)い，(K り%つ|律)
(Y)	漢字表記の一般的な読みと発音が異なる箇所	(Y ぜつ|舌)，(Y ぎょく|玉)
(U)	聞き取りや語の判断に自信がない箇所	(U 外国/外交)，(U な###)
(R)	個人情報に関わる仮名処理 候補	(R 佐藤)さん，(R 国語研究所)
。	発話単位末	うん。，やったけど。，食べます。

※相互行為班と共同し，上記以外のタグ（発話の重複開始のタイミングなど）の付与も検討

NINJAL
National Institute for Japanese Language and Linguistics

進捗状況

NINJAL National Institute for Japanese Language and Linguistics

調査協力者

2016.8.29現在

年齢	男性		女性	
20	学生	調査中	学生	調査中
20	学生	今年度予定	学生	今年度予定
30	有職者	調査終了	専業主婦	調査終了
30	有職者	調査中	有職者	調査終了
30			有職者	今年度予定
30			有職者	今年度予定
40	有職者	調査終了	有職者	調査終了
40	有職者	今年度予定	有職者	今年度予定
40			有職者	今年度予定
50			有職者	調査終了
60	有職者	調査中	ボランティア活動	調査終了
60	有職者	今年度予定		

NINJAL National Institute for Japanese Language and Linguistics

収録状況①

2016.8.29現在

	調査終了(7名)＋調査中(4名)	調査終了(7名)
収録時間*	約140時間	108時間
会話数	169会話	132会話
話者数（延べ）	517人	371人
話者数（異なり）	226人	137人

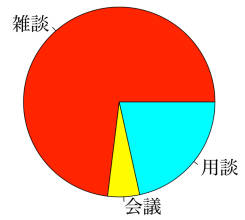
* 収録したメディアファイルから計算。収録準備の時間なども含まれる。

NINJAL National Institute for Japanese Language and Linguistics

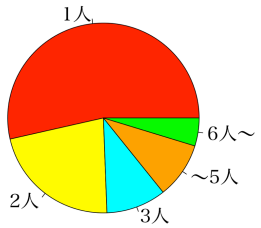
収録状況②

2016.8.29現在

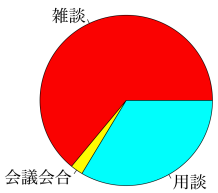
形式

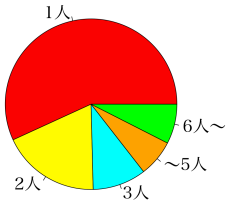


相手人数



※参考：会話行動調査の結果

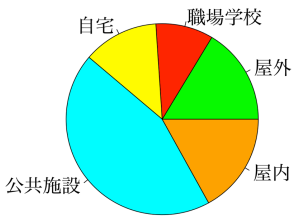




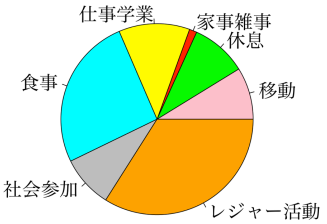
収録状況③

2016.8.29現在

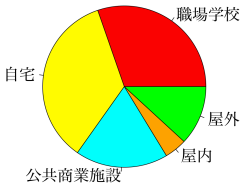
場所

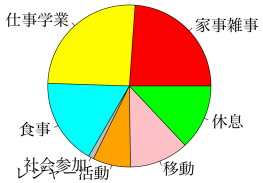


活動



※参考:会話行動調査の結果

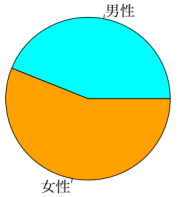




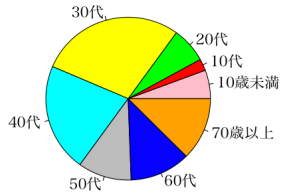
収録状況④

2016.8.29現在

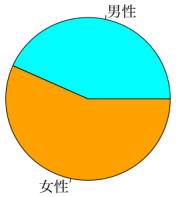
性別・延べ



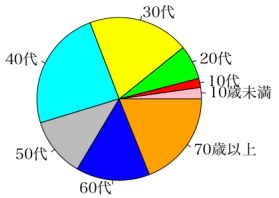
年齢・延べ



性別・異なり



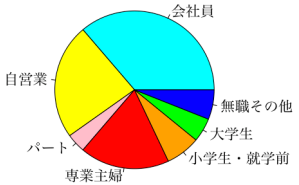
年齢・異なり



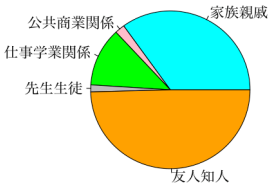
収録状況⑤

2016.8.29現在

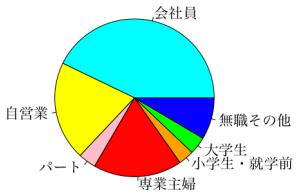
職業・延べ



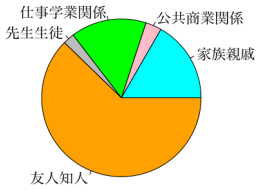
相手との関係・延べ



職業・異なり



相手との関係・異なり



今後の課題・公開予定

- 課題
  - 特定の年代・性別の調査協力者の確保
  - 個人密着法では難しい場面の認定、特定場面法の確立
  - 各種アノテーション基準・手法の確立
  - データ公開に向けた倫理的対応(特に映像データ)の確立

- 公開(予定)
  - 2018年度(プロジェクト3年目)　50時間　モニター公開
  - 2021年度(プロジェクト最終年度)　200時間　本公開

NINJAL
National Institute for Japanese Language and Linguistics