国立国語研究所『中納言』で公開中の話し言葉コーパスの概要

国立国語研究所『中納言』で公開中の話し言葉コーパスの概要

柏野　和佳子

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 NINJAL

『中納言』について

『中納言』の概要(約5分) https://youtu.be/AovMD2N4Qag
「コーパス検索アプリケーション『中納言』音声配信」(約5分)
https://youtu.be/XJTPLMj2yIs
「まとめて検索『KOTONOHA』」(約5分半)
https://youtu.be/CEdGQwIWa28

話し言葉コーパス

『中納言』で公開中のコーパス

『日本語話し言葉コーパス』（CSJ）①

- 国立国語研究所・情報通信研究機構（旧通信総合研究所）・東京工業大学が共同開発
- 1999年構築開始，2004年6月公開
- 約661時間，約753万語（短単位）を収録
- 質・量ともに世界最高水準の話し言葉データベース

音声のタイプ	短単位数	長単位数
学会講演	3,279,364	2,654,823
模擬講演	3,605,729	3,115,302
その他の講演	282,728	239,989
朗読と再朗読	207,478	172,216
対話	149,826	131,544
全体	7,525,125	6,313,874

https://pj.ninjal.ac.jp/corpus_center/csj/

『日本語話し言葉コーパス』（CSJ）②

- **学会講演**
  - 理工学，人文，社会の3 領域におよぶ種々の学会における研究発表のライブ録音。講演時間は10 分から25 分程度が大半。
  - 学会講演の多くをしめる理工学系の学会では，男性の大学院生であることが多い。
  - 発話スタイルは，あらたまり度が高い。

- **摸擬講演**
  - できるだけ年齢と性別のバランスをとった一般話者による，日常的話題についての講演。（例えば「人生で一番嬉しかったこと」「人生で一番悲しかったこと」「私の住んでいる街」等）
  - 1 講演の長さは10～15 分程度。聞き手は3，4 名。
  - 発話スタイルは，学会講演よりもくだけている。

『日本語話し言葉コーパス』(CSJ)③

- **その他の講演**
  - 研究機関が一般聴衆を対象に企画した連続講演会の講演音声。対象は歴史や民俗学など。
  - 国語研究所が一般聴衆むけに開催した講演会の講演音声，および国語研究所員を聴衆とした識者による講演（1 講演のみ）。
  - 専門学校における日本語教師養成関係の講義音声。

- **朗読**
  - 自発的な音声の特徴を明らかにする2種類の比較データ。
  - 「朗読音声」は模擬講演話者の一部が，書き言葉のテキストを朗読したもの。
  - 「再朗読音声」は，学会講演ないし模擬講演として収録された音声の転記テキストを同一の話者が朗読した音声。フィラーや言い直しも朗読の対象。
  - 再朗読の話者は，学会講演話者から選ばれた10 名と模擬講演話者から選ばれた6 名の合計16名。

『名大会話コーパス』(NUCC)

- 日本語母語話者の100時間分の129件の会話(雑談)を収録して,文字化したコーパス。
- 科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」(平成13年度～15年度　研究代表者　大曽美恵子)により作成された。

- 名古屋近辺で録音されたデータが最も多いが、東京近辺、北海道、新潟で録音されたものもある。
- 共通語による会話が大半を占めるが、方言も使われている。
- 参加者の年代は様々で10代～90代までと幅広い。女性の方が多い。
- 日本語教育関係者、言語研究者が多いので、日本語のメタ言語的な使い方が多い。
- 親しい者同士の雑談が多いが、初対面同士、研究メンバー同士の会話も一部入っている。先輩―後輩の会話もある。
- 話題を一切制限していない雑談であるが、参加者は録音していることを知らされていた。

https://mmsrv.ninjal.ac.jp/nucc/

『現日研・職場談話コーパス』(CWPC)

- 現代日本語研究会が作成した，首都圏の有職女性19名（20代～50代）と，を調査協力者として，首都圏の有職男性21名（20代～50代）の職場での自然談話を文字起こししたテキストを元にしたコーパスである。
- その元となっているテキストは，『合本　女性のことば・男性のことば（職場編）』（現代日本語研究会編，2011年，ひつじ書房）の付録CD-ROMに収録されている。
- 1990年代にいち早く行われた先駆的な試みで，職場での会話を調査協力者自身に録音してもらい，自然な談話を収録するという方法で得られた，たいへん画期的なものであると評価されているデータである。

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

『名大会話コーパス』(data)と『現日研・職場談話コーパス』(M/F)の用例

会話ID	前文脈	キー	後文脈
data077	この子は、E短の子だよ。あっ、そうなんだー、	微妙	、微妙。うんそういうのばっかり。
M06Q031	まー、驚くことが多いって話よー。	微妙	なニュアンスで教えてくれてー。
data072	なかなか時間がないんだよね。ねーあたしもだよ。	やばー	い。あっ、TOEICさ、こないだあったけど、
M12Q031	#あの、これやうめーや、ちょっと、	やばい	んじゃない。#このランチメニュー。
data011	5級だったしね、一番最初受けたの。	まじ	?6年のときに5級。
M21Q011	3時までさー、ずっーとしゃべってて。#	まじ	でー。#まじ、もーあたし、その前の日とかに、
data046	うーん。無理。だから、なんで。とにかく	無理	。それは、そういうことしたら、
M21K011	#夜は	無理	っす、平日の夜は無理っす。
data056	うんどこで見たの。	てか、	あの、日本に来たときの。
M12Q101	#うん。#いや、	てか、	自動なんだよー、もー。
data103	んー、かっこいい。＊＊＊。	すごい	(かわいい)、この絵。
F11Q011	#すごい、なんだっけそれ。#	すごい	(おいしい)やつ。
data065	あ、君は日本文学専攻か、ふーん、とか言って。	受ける	ー。うーん話しかけやすい雰囲気なんじゃん。
F15K011	#だって、みんな、	うける	ものねー、あれ、すごく。#うけますねー。
data085	今日、さむ。な、何となく寒そう、	みたいな?	うん。雪が降ったときとか、
F15Q011	#でしょ#でビール、がーん、みたいな	みたいな。	#もっとほかのお母さまの意見も聞いた方が###。

『日本語日常会話コーパス』
（CEJC）
現在構築中，モニター版を公開中

『日本語日常会話コーパス』は，国立国語研究所の機関拠点型基幹研究プロジェクト「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）の研究成果です。

http://pj.ninjal.ac.jp/conversation/

「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」(約10分半)

https://youtu.be/HksF4QtbrI8

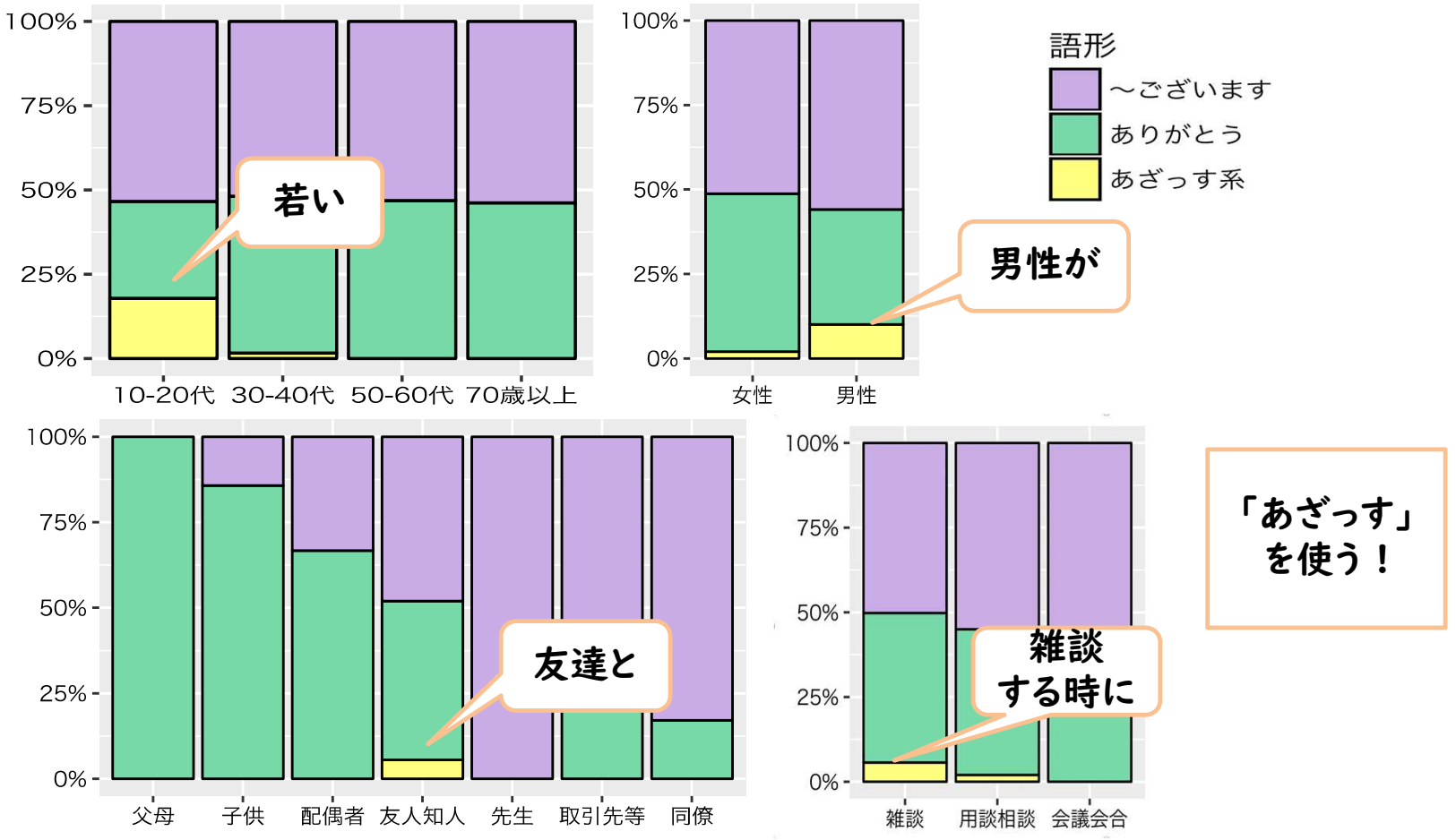
- 一般の協力者に依頼し，日常に生じる自然な会話200時間を記録
- 映像・音声を含めて収録・公開
- 多様な場面・話者を対象

『日本語日常会話コーパス』

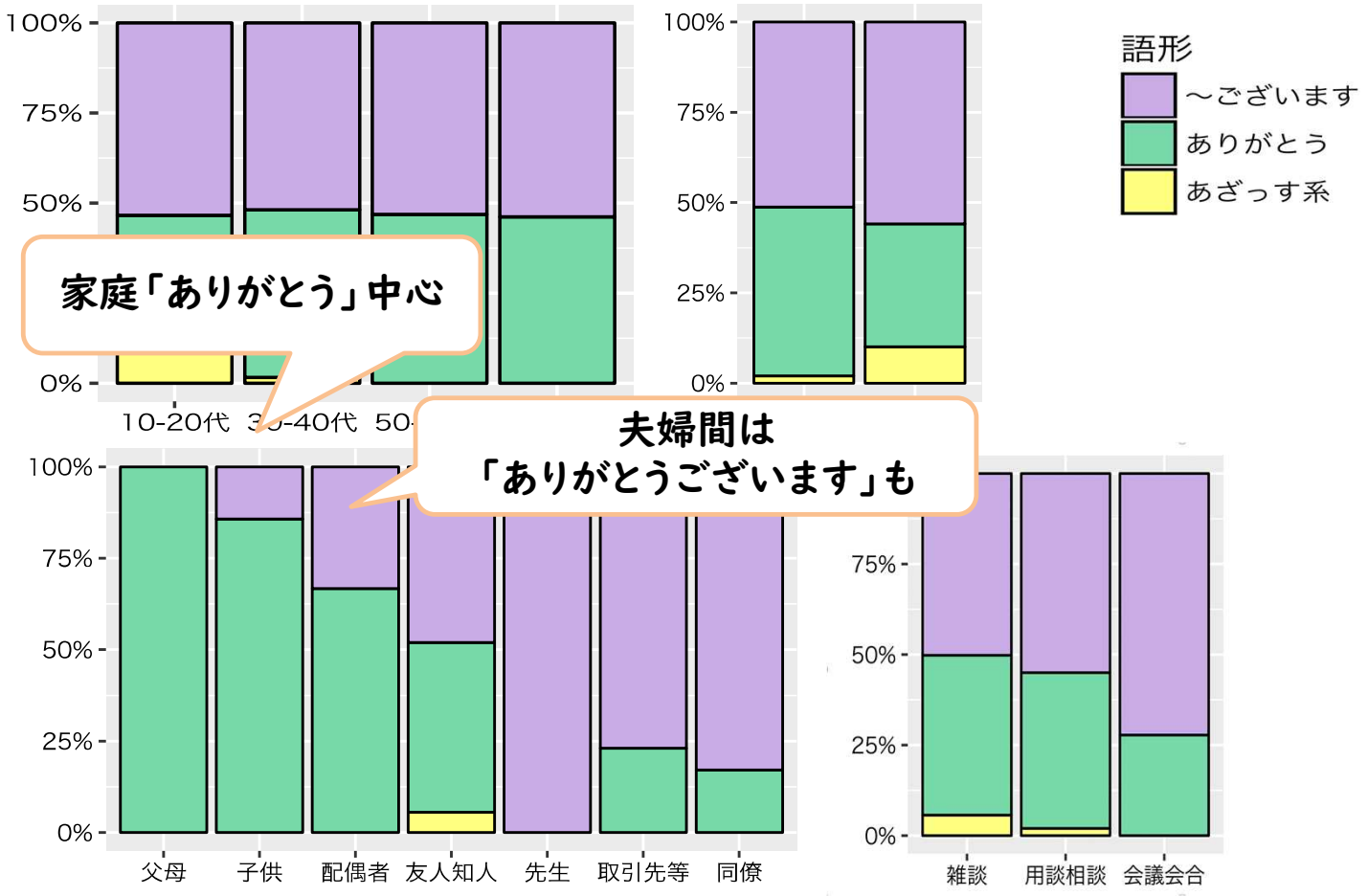
特徴

- 一般の協力者に依頼し、日常に生じる自然な会話200時間を記録
- 映像・音声を含めて収録・公開
- 多様な場面・話者を対象

「ありがとうございます」
「ありがとう」「あざっす!」



「ありがとうございます」
「ありがとう」「あざっす!」

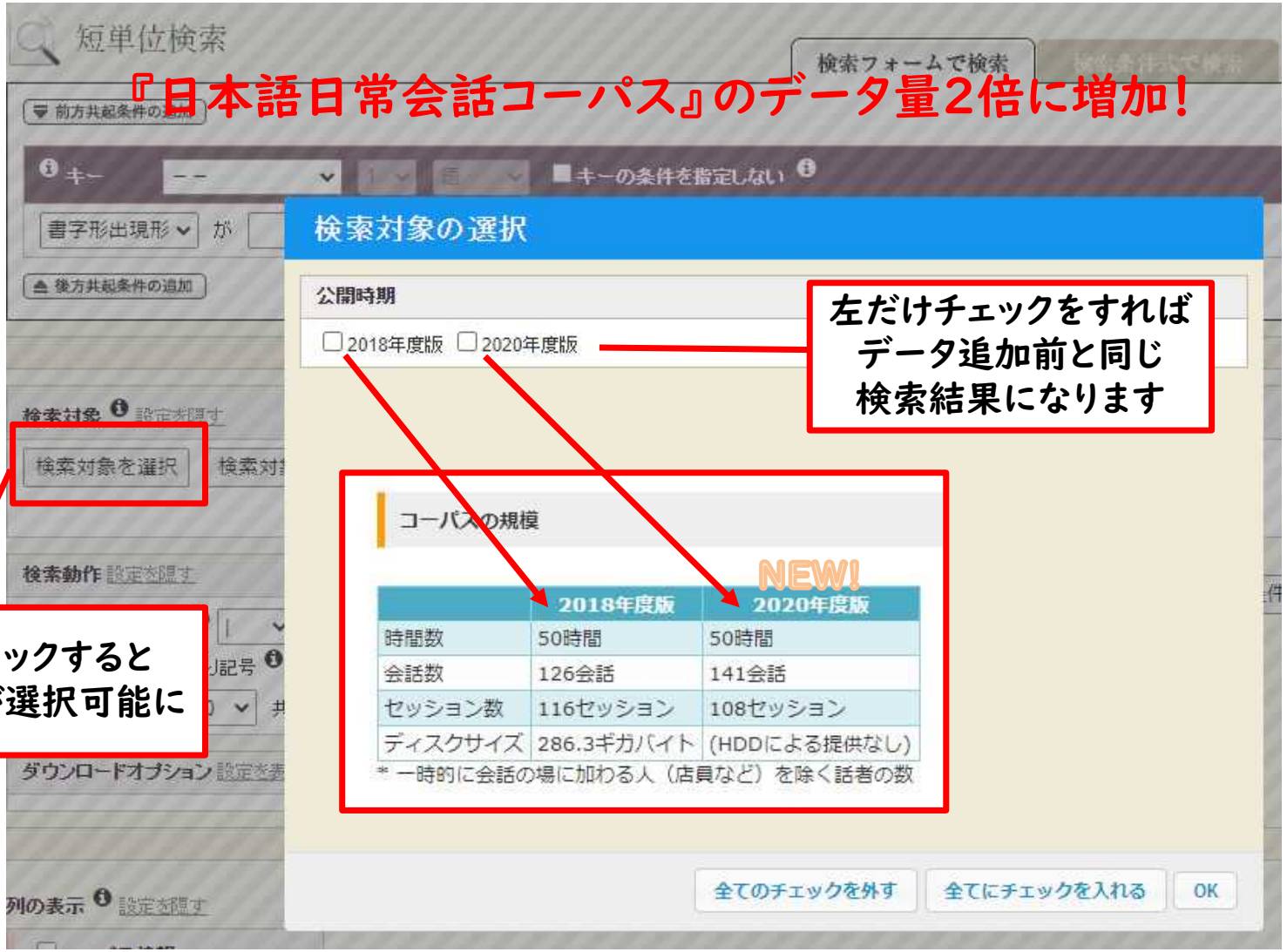


『日本語日常会話コーパス』モニター公開

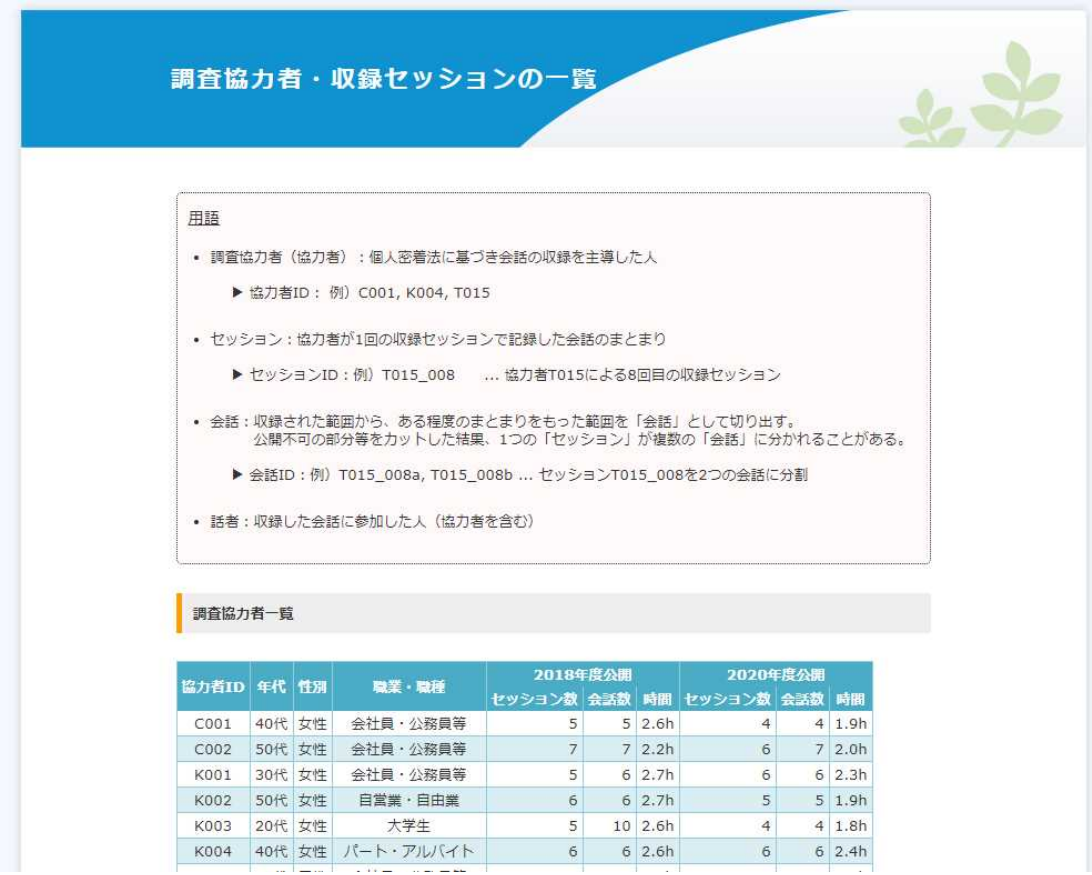
- 2018年度（2018年12月4日）公開開始
- 対象会話：50時間・協力者20名×平均2.5時間
- 公開データ：
  - 映像・音声データ
  - 転記テキスト
  - 短単位（単語）情報

	男性				女性			
年代	職業・職種等	収録数	会話数	時間	職業・職種等	収録数	会話数	時間
20代	大学生	5	5	2.2h	大学生	7	7	2.6h
	大学院生	5	5	2.5h	大学生	5	10	2.6h
30代	自営業・自由業	4	4	2.8h	会社員・公務員等	5	6	2.7h
	会社員・公務員等	6	6	2.2h	専業主婦	7	7	2.8h
40代	会社員・公務員等	4	5	2.1h	会社員・公務員等	5	5	2.6h
	自営業・自由業	6	6	2.4h	パート・アルバイト	6	6	2.6h
					パート・アルバイト	6	6	2.6h
50代	会社員・公務員等	7	7	2.4h	会社員・公務員等	7	7	2.2h
	会社員・公務員等	4	4	2.6h	自営業・自由業	6	6	2.7h
60代以上	その他（非常勤講師）	9	9	2.1h	専業主婦	6	7	2.7h
	定年退職	6	8	3.0h				

2020年度末（2021年2月17日）に、50時間分が追加され、合計100時間分の転記データの検索が可能に



調査協力者・収録セッションの一覧



調査協力者・収録セッションの一覧

用語

- 調査協力者（協力者）：個人密着法に基づき会話の収録を主導した人
  - ▶ 協力者ID： 例）C001, K004, T015
- セッション：協力者が1回の収録セッションで記録した会話のまとまり
  - ▶ セッションID：例）T015_008　　... 協力者T015による8回目の収録セッション
- 会話：収録された範囲から、ある程度のまとまりをもった範囲を「会話」として切り出す。
  公開不可の部分等をカットした結果、1つの「セッション」が複数の「会話」に分かれることがある。
  - ▶ 会話ID：例）T015_008a, T015_008b ... セッションT015_008を2つの会話に分割
- 話者：収録した会話に参加した人（協力者を含む）

調査協力者一覧

協力者ID	年代	性別	職業・職種	2018年度公開			2020年度公開		
				セッション数	会話数	時間	セッション数	会話数	時間
C001	40代	女性	会社員・公務員等	5	5	2.6h	4	4	1.9h
C002	50代	女性	会社員・公務員等	7	7	2.2h	6	7	2.0h
K001	30代	女性	会社員・公務員等	5	6	2.7h	6	6	2.3h
K002	50代	女性	自営業・自由業	6	6	2.7h	5	5	1.9h
K003	20代	女性	大学生	5	10	2.6h	4	4	1.8h
K004	40代	女性	パート・アルバイト	6	6	2.6h	6	6	2.4h

https://www2.ninjal.ac.jp/conversation/cejc-monitor/convList.html#session

(約16分半)
https://www.youtube.com/watch?v=7fSbEvbTmjc