2021.08.26
オンライン検索システム「中納言」講習会

『日本語日常会話コーパス モニター公開版』概説

国立国語研究所 言語変化研究領域　山崎誠

概要

- 1.『日本語日常会話コーパス』(CEJC)
- 2. CEJCモニター公開版の概要
- 3. データについて

『日本日常会話コーパス』(CEJC)

- 話者や場面などを考慮してさまざまなタイプの日常会話200時間をバランスよく納めたコーパス。
- 2016年度構築開始，2021年度完成予定。
- 略称は，CEJC（Corpus of Everyday Japanese Conversation）

1.1 CEJCの特徴

- さまざまな場面における自然な日常会話をバランスよく収めたコーパス。音声だけでなく，映像も公開。映像付きの大規模日常会話コーパスは世界初の試み。

- さまざまな場面における会話を収集するために，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 40名の調査協力者の方々に機材機器を3か月ほど貸し出し，協力者の日常生活で自然に生じる会話を協力者自身に記録してもらっている。

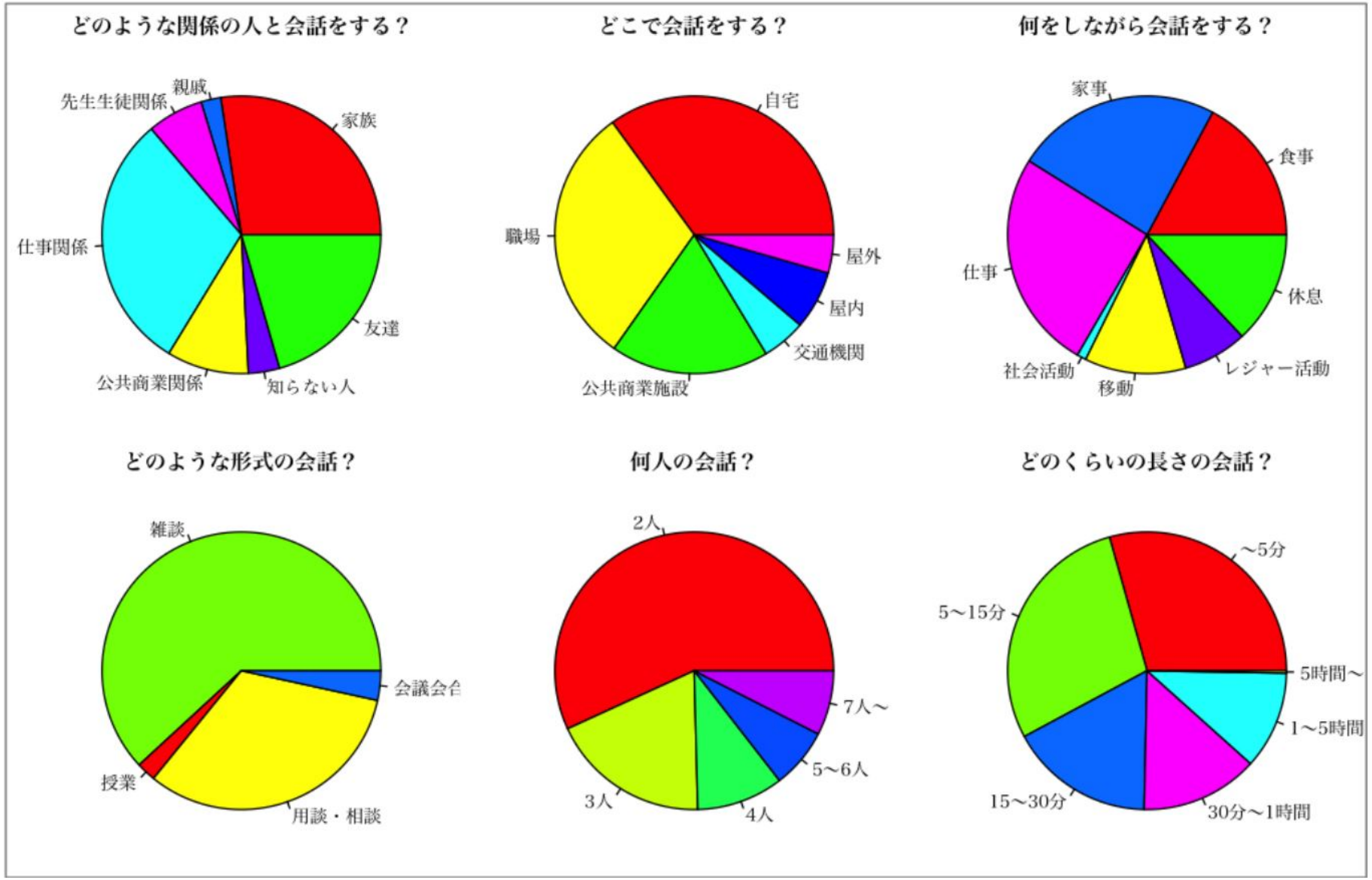
- （参考）
- https://www2.ninjal.ac.jp/conversation/cejc.html

1.2 CEJCの設計

- 事前に行われた会話行動調査に基づき設計。
- 日常の言語生活を反映したコーパスを作成するにはため，私たちが普段どのような会話をどの程度行っているのかを調査した。
- 2014年11月から2015年2月にかけ，首都圏在住の成人243名を対象に起床から就寝までの間に行った全ての会話について，いつ，どこで，誰と，何をしながら，どのような種類の会話を，どのくらいの長さ行ったか，などを問う調査をした。

- https://www2.ninjal.ac.jp/conversation/survey.html

1.2 CEJCの設計


どのような関係の人と会話をする？
親戚
先生生徒関係
家族
仕事関係
友達
公共商業関係
知らない人
どこで会話をする？
自宅
職場
屋外
屋内
交通機関
公共商業施設
何をしながら会話をする？
家事
食事
仕事
休息
社会活動
移動
レジャー活動
どのような形式の会話？
雑談
会議会合
授業
用談・相談
何人の会話？
2人
7人～
5～6人
3人
4人
どのくらいの長さの会話？
～5分
5～15分
5時間～
1～5時間
15～30分
30分～1時間

1.2 CEJCの設計

- 参考

- プロジェクト報告書『一日の会話行動に関する調査報告』
- https://www2.ninjal.ac.jp/conversation/report/report01.pdf

- 国立国語研究所論集「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」
- https://repository.ninjal.ac.jp/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=824&item_no=1&page_id=13&block_id=21

2. CEJCモニター公開版の概要

- 2018年12月から公開。2018年度版は50時間。
- 2021年2月に2020年度版（50時間）を公開。合わせて100時間になった。

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

2.1. モニター公開版の公開方法

- オンライン検索システム「中納言」での公開。
  - 形態論情報（短単位情報）での検索と文字列検索ができる。音声・映像の視聴・ダウンロードはできない。該当箇所の音声が聞ける。

- ハードディスクでの公開
  - 映像・音声・転記テキスト・短単位情報・メタ情報・検索システムなどが含まる。
  - 国語研究所と利用契約を結んだ上で利用。
  - コンテンツ利用は無償だが，実費10,000円（税込）が必要。

2.2. モニター公開版の規模

モニター公開版（2018年度版＋2020年度版）	
時間数	100時間
会話数	267会話
セッション数*	224セッション
話者数（延べ）**	961人
話者数（異なり）**	457人

*セッションについては後述。
**話者IDによる集計。

2.2. モニター公開版の規模

	2018年度版	2020年度版
時間数	50時間	50時間
会話数	126会話	141会話
セッション数*	116セッション	121セッション
話者数（延べ）**	463人	498人
話者数（異なり）**	277人	278人

*セッションについては後述。
**話者IDによる集計。

2.3. 会話の属性

- セッションID：協力者が1回の収録で記録した会話のまとまりつけたID。例えば，「C001_001」は，協力者C001の1番目の収録セッションを意味する。

- 会話ID：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切り出したものに付与したID。1つのセッションが複数の会話に分かれることがある。複数に分かれて出来た会話には，「T002_011a」など，a,bで区別する。これは，協力者T002の11番目の収録セッションのうち複数に分割された1つ目の会話を意味する。

- 公開時期：2018年度版，2020年度版。

2.3. 会話の属性

- 会話概要：会話の内容を簡潔に表したもの。

- 会話時間：当該会話の時間。分数で表す。

- 話者数：当該セッションの主たる話者の数（一時的に話に加わった話者などは除く）。

2.3. 会話の属性

- 形式：主たる会話の形式。「雑談，用談相談，会議会合」の3種類。

- 場所：会話が行われた場所。「自宅，職場，学校，屋外，室内，交通機関，施設」など。

- 活動：何をしながら会話をしていたか。「食事，家事雑事，身の回りの用事，仕事，学業，社会参加，レジャー活動，付き合い，移動，休憩」など。2つまで複数付与可。

2.3. 会話の属性

- 話者数：当該セッションの主たる話者の数（一時的に話に加わった話者などは除く）。

- 話者間の関係：主たる話者間の関係。複数の関係性が割当てられることがある。「家族，親戚，友人知人，同僚，仕事関係，先生生徒，サービス場面関係，初対面」など。

2.4. 話者の属性

- 調査協力者（協力者）：会話の収録を主導した人。IDで表す（C001, K004, T015など）。

- 話者ID：話者を一意に同定する固有のID。話者には調査協力者も含む。
  - C001 …調査協力者の場合
  - C001_001 …調査協力者C001が収集した会話に参加した話者の場合

- 話者ラベル：話者に与えられた仮名（かめい）。

2.4. 話者の属性

- 年齢：収録当時の年齢（5歳刻み）。「0–4歳，15–19歳」など。

- 性別：話者の性別。「女性，男性」

- 出生地：話者の出身地。都道府県レベル（外国の場合には国レベル）で表す。

- 居住地：話者の出身地。都道府県レベル（外国の場合には国レベル）で表す。

2.4. 話者の属性

- 職業：話者の職業。「会社員・役員・公務員・専門職，自営業・自由業，パート・アルバイト，専業主婦・主夫，無職・定年退職，就学前，小学生，中学生，高校生，大学生，大学院生，その他」など。

- 協力者からみた関係性：協力者からみた会話相手との関係。「本人，家族親族，学校の先生，学校の生徒学生，仕事関係者，習い事などの先生，習い事などの生徒，友人知人，サービスを受ける人，サービスを提供する人」など。

2.4. 協力者一覧(一部)

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
S001	50代	男性	会社員・公務員等	4	4	2.6h	0	0	0h
S002	30代	男性	会社員・公務員等	0	0	0h	7	7	4.0h
T001	30代	男性	自営業・自由業	4	4	2.8h	4	4	2.3h

2.6. 会話一覧(一部)

会話ID	会話概要	話者数	時間	形式	場所	活動	話者間の関係性	2020年度追加
C001_001	飲食店で友人4人と誕生日会	5	31	雑談	施設_飲食店	付き合い	友人知人	
C001_002	友人宅に宿泊し起床後に友人と雑談	2	17	雑談	室内_知人宅	休息	友人知人	
C001_003	自宅に泊まりにきた友人と雑談	2	46	雑談	自宅	休息	友人知人	○
C001_004	職場の同僚と飲食店で飲みながら	2	17	雑談	施設_飲食店	付き合い・食事	同僚	○
C001_005	実家から知人宅に向けて母と散歩	2	17	雑談	屋外	移動	家族	
C001_006	友人宅から実家に母と一緒に帰る途中の会話	2	18	雑談	屋外	移動	家族	○
C001_007	実家で両親とニュースを見ながら夕食	3	51	雑談	室内_実家	食事	家族	
C001_012	同僚宅で同僚と食事をしながら	2	39	雑談	室内_知人宅	付き合い・食事	同僚	
C001_013	同僚宅で同僚と飲みながら	2	34	雑談	室内_知人宅	付き合い	同僚	○

2.7. 利用に当たっての注意（個人情報）

- データには多くの個人情報が含まれています。

- 会話者や第三者の個人情報・プライバシー・著作権等を侵害するような扱いは避けて下さい。

- 不明な点は，cejc-monitor [at] ninjal.ac.jp　までお問い合わせ下さい。

2.8. 参考文献

- 研究成果を発表する場合，「『日本語日常会話コーパス』モニター公開版」を利用した成果であることを明記し，以下の以下の文献のいずれかを引用文献として下さい。

- 小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉「『日本語日常会話コーパス』モニター版の設計・評価・予備的分析」『国立国語研究所論集』18, pp.17–33, 2020.1.
- 小磯花絵・天谷晴香・石本祐一・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉『日本語日常会話コーパス』モニター公開版 コーパスの設計と特徴』国立国語研究所日常会話コーパスプロジェクト報告書3, 2019.3.

3. データについて

- コーパスへのアノテーションは基本的に自動的に付与されています。

- 形態素解析の精度は，100%ではなため，コーパスの分析は一定の誤差があることを前提として行う必要があります。

- 比較しようとするデータのサイズが異なる場合，出現した生の頻度ではなく，per million wordなどの相対頻度を使わないと適切な比較にはなりません。

参考URL

- 日本語日常会話コーパス
- https://www2.ninjal.ac.jp/conversation/index.html

- モニター公開版
- https://www2.ninjal.ac.jp/conversation/cejc-monitor.html

- 研究成果
- https://www2.ninjal.ac.jp/conversation/publication.html