2022.09 公開

『日本語日常会話コーパス』短単位語彙表・語数表 ver.2022.09 解説

1. データの概要

本データは、『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation、以下CEJC)の『中納言』(2.6.0 データバージョン 2022.03)に収録されている200時間分のデータ(およそ2016 年 4 月から 2020 年にかけての収録)に基づいて、頻度 1 までの見出し語を対象にして作成した語彙表と語数表である。

CEJC の詳細は下記文献を適宜参照のこと。なお、本データを利用した研究成果を発表する場合、『『日本語日常会話コーパス』短単位語彙表・語数表 ver.2022.09』を利用した成果であること、及び、以下に記す文献を 1 件以上、明記すること。

- 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・渡邊友香 「『日本語日常会話コーパス』の設計と特徴」『言語処理学会第 28 回年次大会発表論文集』pp.2008-2012, 2022.3.
- 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・渡邊友香 『『日本語日常会話コーパス』設計・構築・特徴』(国語研究所「日常会話コーパス」プロジェクト報告書 6)2022.3.
- Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yuka Watanabe, and Yasuyuki Usuda. Design and Evaluation of the *Corpus of Everyday Japanese Conversation*, *Proceedings of LREC2022*, pp.5587-5594, 2022.6.

なお、2018 年度版(50 時間分の収録データ)に基づいて作成した語彙表・語数表の基本的な統計情報は、次の報告書に公開している。適宜参照されたい。

- 大村舞・柏野和佳子・山崎誠(2020),『日本語日常会話コーパス』 モニター公開版の語彙(国立国語研究所「日常会話コーパス」プロジェクト報告書 4)

CEJC には様々な属性情報が付与されているが、語彙表・語数表で取り上げる主な属性は次のとおり。

◆会話の属性◆

形式:

- 「雑談」は、会話の目的や話題などがあらかじめ定められていない会話。
- 「用談・相談」は、会話の目的はある程度決まっているが時間や場所などは定められていない会話。
- 「会議・会合」は、「用談・相談」とは異なり時間や場所などが定められている会話。
- 「授業・レッスン」は、授業や習い事の際の会話。

場所:

- 会話の行われている場所が「自宅」「職場」「学校」か、「公共商業施設」「交通機関」か、あるいは、それら以外の「室内」「屋外」か。

◆話者の属性◆

- 「性別」と「年齢」。
- 「年齢」は 5 歳きざみ。

2. 語彙素・語数表の集計方法

(1) 語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定した。
(2) CEJC は誤解析を含む。そのため、本語彙表・語数表のデータも同様にエラーを含んでいる。
(3) CEJC 語彙表に収められた語の頻度の総計は、2,419,171 である。『中納言』の検索結果で表示される「記号・補助記号・空白を除いた検索対象語数」と一致する。

3. CEJC 語彙表

・pmw(100 万語当たりの頻度)は、小数点以下第 7 位まで示した。
・同順位の語があった場合は、語彙素読み、語彙素、品詞の順に文字コード昇順で並べた。

3.1 CEJC 語彙表
・ファイル名:2_cejc_frequencylist_suw_token.tsv
(xlsx 形式もあり: 3_cejc_frequencylist_suw_token.xlsx)
・29,535 行, UTF8, タブ区切り。
・語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定したものである。
・第 1 行目は見出し。2 行目以降がデータ。各行には「1_CEJC 語彙表の項目一覧.xlsx」のシート「CEJC 語彙表の項目」に示す 311 の項目が並んでいる。

3.2 語形別の CEJC 語彙表
・ファイル名:4_cejc_frequencylist_gokei.xlsx
・31,300 行。
・語彙表では語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定するのに対し、語彙素、語彙素読み、品詞、語彙素細分類、語種、語形の 6 つの組で見出し語を特定したものである。
・第 1 行目は見出し。2 行目以降がデータ。各行には「1_CEJC 語彙表の項目一覧.xlsx」のシート「語形別の CEJC 語彙表の項目」に示すとおり、3.1 の語彙表に「語形」を加えた 312 の項目が並んでいる。

3.3 書字形別の CEJC 語彙表
・ファイル名:5_cejc_frequencylist_shozikei.xlsx
・32,344 行。
・語彙表では語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定するのに対し, 語彙素, 語彙素読み、品詞、語彙素細分類、語種、書字形の 6 つの組で見出し語を特定したものである。
・第 1 行目は見出し。2 行目以降がデータ。各行には「1_CEJC 語彙表の項目一覧.xlsx」のシート「書字形別の CEJC 語彙表の項目」に示すとおり、3.1 の語彙表に「書字形」を加えた 312 の項目が並んでいる。

3.4 発音形別の CEJC 語彙表

・ファイル名：6_cejc_frequencylist_hatuonkei.xlsx

・40,281 行。

・語彙表では語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定するのに対し、語彙素、語彙素読み、品詞、語彙素細分類、語種、発音形出現形の 6 つの組で見出し語を特定したものである。

・第 1 行目は見出し。2 行目以降がデータ。各行には「1_CEJC 語彙表の項目一覧.xlsx」のシート「発音形別の CEJC 語彙表の項目」に示すとおり、3.1 の語彙表に「発音形出現形」を加えた 312 の項目が並んでいる。

4. CEJC 語数表

4.1 CEJC 語数表

・ファイル名：7_cejc_wc.xlsx

・2,030 行。

・会話 ID、話者別の延べ語数表。。

表 1 語数表の項目

番号	見出し	備考
1	会話 ID	収録された範囲からまとまりをもって切り出された「会話」に与えられる固有の ID 補足）通常、「セッション ID」と同じだが、問題のある箇所等を除いた結果、複数に分割された場合は枝番がつく 例）C001_001 ：協力者 C001 の 1 番目の収録セッションの会話（枝番がないためセッション ID と同じ） 例）T002_011a ：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
2	コア	コアデータに該当する場合「1」,該当しない場合「0」
3	セッション ID	1 回の収録に与えられる固有の ID 例）C001_001 ：協力者 C001 の 1 番目の収録セッション
4	会話概要	会話の概要
5	収録年	会話を収録した年（西暦），整数
6	会話時間	会話の時間（分），整数
7	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），整数
8	話者間の関係性	当該セッションの主たる話者間の関係 補足）組合せで複数の関係性が割当てられることがある 値）家族/親戚/友人知人/同僚/元同僚/仕事関係/先生生徒/サービス場面関係

9	形式	主たる会話の形式 補足）会合の合間に雑談が交じるなど複数の形式が関わるものもあるが主たる形式を一つ認定 値）雑談/用談・相談/会議・会合/授業・レッスン
10	場所	会話が行われた場所 値）自宅/職場（-,相手の職場）/学校/公共商業施設（飲食店,商業施設,公共施設,宿泊施設,医療福祉施設,スクール）/交通機関（車,電車）/室内（家族宅,実家,親類宅,友人知人宅,別荘）/屋外
11	活動	何をしながら会話をしていたか 補足）食事しながら仕事、食事したあと仕事、といったような場合には複数（最大２つ）の活動を割当て 値）仕事/学業/家事雑事/身の周りの用事/食事/付き合い/レジャー活動/社会参加/課外活動/休息/移動/その他
12	話者 ID	話者を一意に同定する固有の ID 例）C001 …調査協力者の場合 例）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
13	話者ラベル	話者に与えられた仮名（カメイ）の化された名前 例）高橋, 修司, パパ
14	年齢	収録当時の年齢（5 歳刻み）, 欠損値(NA)あり 例）0–4 歳, 15–19 歳, 30–34 歳, 90–94 歳
15	性別	話者の性別 値）男性/女性
16	出身地	話者の出身地, 都道府県レベル（外国の場合には国レベル）, 欠損値(NA)あり 例）北海道, 宮城県, 東京都, 京都府, 中国
17	居住地	話者の現在の居住地, 都道府県レベル（外国の場合には国レベル）, 欠損値(NA)あり 例）北海道, 宮城県, 東京都, 京都府, 中国 補足）個人密着法は首都圏在住の協力者に依頼しているため, 会話に参加する話者も首都圏居住者が多い
18	職業	話者の職業, 欠損値(NA)あり 値）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(-,非常勤講師)/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（-,農業従事,バレエ講師,シルバー人材センター）

19	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（個人密着法による収録の場合）,空白あり 値）本人/家族親戚（夫,妻,父,母,息子,娘,兄,姉,弟,妹,祖父,祖母,婿,嫁,おじ,おば,おい,めい,いとこ,孫娘,孫息子,義父,義母,義兄,義姉,義弟,義妹,その他）/学校の先生/学校の生徒・学生/習い事などの先生/習い事などの生徒/仕事関係者（上司,同僚,部下,元上司,元同僚,取引先など他社の人,その他）/友人知人（-,先輩,同級生,後輩）/サービスを受ける人/サービスを提供する人/その他
20	語数(全て)	整数 ※合計は、『中納言』の検索結果で表示される「検索対象語数」と同じ 2,421,162 語。
21	語数(記号等除外・全て)	整数（品詞に「空白」「補助記号」「記号」の文字列を含むものを除外） ※合計は、『中納言』の検索結果で表示される「記号・補助記号・空白を除いた検索対象語数」と同じ 2,419,171 語。

4.2 形式・場面・年齢・性別・話者間の関係性の語数表

・ファイル名：8_CEJC_語数表_形式と場所と年齢と性別と話者間の関係性.xlsx
・粗頻度から調整頻度を求める場合の参照用の語数表として作成。
・「7_cejc_wc.xlsx」より、形式、場所、年齢（5 歳きざみ、および、10 歳きざみを基本に 10 代以下と 70 代以上はまとめたもの）、性別、話者間の関係性の延べ語数を集計して作成した表を掲載。
・その他の語数表が必要な場合は、各語彙表・語数表をもとに集計されたい。

5. CEJC 品詞構成表

・ファイル名：9_cejc_frequencylist_pos.xlsx
・83 行。
・以下の 4 つの表を収めた。
(1)短単位における品詞の語数（延べ語数）
(2)短単位における品詞の語数（異なり語数）
(3)短単位における品詞の割合（延べ語数）
(4)短単位における品詞の割合（異なり語数）
・いずれの表も第 1 行目は見出し。2 行目以降がデータ。
・列は、形式、場面、性別、年齢、形式と場所、性別と年齢。
・割合（百分率）は小数点以下第 3 位まで示した。

6. CEJC 語種構成表

・ファイル名：10_cejc_frequencylist_wtype.xlsx
・39 行。
・CEJC 品詞構成表と同様に 4 つの表を収めた。表の種類は品詞構成表と同じ。
・いずれの表も第 1 行目は見出し。2 行目以降がデータ。
・列は、形式、場所、性別、年齢、形式と場所、性別と年齢。

・割合（百分率）は小数点以下第 3 位まで示した。

7. 利用上の注意

(1)研究、教育目的であれば無償で自由に利用できる。申し込みの必要はない。
(2)再配布は不可。商業使用（営利目的での利用）は要相談。
(3)論文等に引用する際は出典とバージョンを明記すること。以下に例を示す。
『日本語日常会話コーパス』短単位語彙表・語数表 ver.2022.09
『日本語日常会話コーパス』短単位語彙表 ver.2022.09
『日本語日常会話コーパス』短単位語数表 ver.2022. 09
『日本語日常会話コーパス』短単位品詞構成表 ver.2022. 09
『日本語日常会話コーパス』短単位語種構成表 ver.2022. 09
(4)本データの著作権（編集著作権）は国立国語研究所が有する。
(5)データの瑕疵による損害についてはいかなる場合でも補償しない。
(6)内容の改善のため予告なく更新することがある。

本データに関する問い合わせ先：kotonoha@ninjal.ac.jp （@を半角に変えること）

以上

更新履歴
2020.3.18『日本語日常会話コーパス』語彙表・語数表 ver.2020.03 を作成
2022.1.24『日本語日常会話コーパス』語彙表・語数表 ver.2022.01 を作成
2022.9.6『日本語日常会話コーパス』語彙表・語数表 ver.2022.09 を作成