『日本語日常会話コーパス』語彙表・語数表 ver.2020.03　解説

2020.3.18 作成

１．データの概要

本データは 2018 年 12 月に公開した『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation，以下 CEJC）のモニター公開版の『中納言』（2.4.5 データバージョン 2018.12） に基づく語彙表及び語数表である。

CEJC の完成時には約 200 時間分の日常会話を収録することになるが，それに先だってモニター公開版（約 50 時間の会話が収録）がリリースされている。モニター公開版の概要は，web サイト[1]のほか，次を参照のこと。

> 小磯花絵・天谷晴香・石本祐一・伊關真理子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉（2019），『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），国立国語研究所[2]

今回公開する語彙表・語数表は，その約 50 時間の会話データを対象に頻度 1 までの見出し語を対象に作成したものである。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），国立国語研究所[3]

CEJC モニター公開版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，次のとおりである。

◆会話の属性◆

形式：

- 「雑談」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」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

場面：

- 会話の行われている場所が「自宅」「職場」「学校」か，公共や商業の「施設」，「交通機関」か，あるいは，それら以外の「屋内」か「屋外」か

◆話者の属性◆

- 「性別」と「年齢」。

---

1 https://www2.ninjal.ac.jp/conversation/cejc-monitor.html
2 https://www2.ninjal.ac.jp/conversation/report/report03.pdf
3 https://www2.ninjal.ac.jp/conversation/report/report04.pdf

- 「年齢」は 5 歳刻み。なお，10 歳刻みを基本に 10 代以下と 70 代以上はまとめて示した集計を行っている語彙表も作成している。

２．語彙素の集計方法

(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)(1)で得られた見出し語の集合から以下の条件に該当するものを除外した。
  1) 品詞に「空白」「補助記号」「記号」の文字列を含むもの。
  2) 語彙素が空(null)のもの（この場合，語彙素読みも同時に空になっている）。
  3) ただし，品詞「言いよどみ」は除外していない。
(3) 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

３．１　CEJC 語彙表
・ファイル名：1_cejc_frequencylist_token.tsv （xlsx 形式もあり: 2_cejc_frequencylist_token.xlsx）
・14,418 行，UTF8，タブ区切り。
・第 1 行目は見出し。2 行目以降がデータである。各行には以下の表 1 に示す 278 の項目が並んでいる。
・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。
・Excel 等のソフトに読み込んで，目的の列で並べ替えれば，属性別の語彙表を得ることができる。

表 1　語彙表の各項目

番号	見出し	備考
1	rank	CEJC 全体の順位
2	語彙素読み	語彙素読み
3	語彙素	語彙素
4	品詞	品詞
5	語彙素細分類	語彙素細分類
6	語種	語種
7	frequency	CEJC 全体の頻度
8	pmw	CEJC 全体での 100 万語当たりの頻度
9	会議会合_rank	会議会合の順位
10	会議会合_frequency	会議会合の頻度
11	会議会合_pmw	会議会合全体での 100 万語当たりの頻度
12	雑談_rank	雑談の順位
13	雑談_frequency	雑談の頻度
14	雑談_pmw	雑談全体での 100 万語当たりの頻度
15	用談相談_rank	用談相談の順位

16	用談相談_frequency	用談相談の頻度
17	用談相談_pmw	用談相談全体での 100 万語当たりの頻度
18	屋外_rank	屋外の順位
19	屋外_frequency	屋外の頻度
20	屋外_pmw	屋外全体での 100 万語当たりの頻度
21	学校_rank	学校の順位
22	学校_frequency	学校の頻度
23	学校_pmw	学校全体での 100 万語当たりの頻度
24	交通機関_rank	交通機関の順位
25	交通機関_frequency	交通機関の頻度
26	交通機関_pmw	交通機関全体での 100 万語当たりの頻度
27	施設_rank	施設の順位
28	施設_frequency	施設の頻度
29	施設_pmw	施設全体での 100 万語当たりの頻度
30	自宅_rank	自宅の順位
31	自宅_frequency	自宅の頻度
32	自宅_pmw	自宅全体での 100 万語当たりの頻度
33	室内_rank	室内の順位
34	室内_frequency	室内の頻度
35	室内_pmw	室内全体での 100 万語当たりの頻度
36	職場_rank	職場の順位
37	職場_frequency	職場の頻度
38	職場_pmw	職場全体での 100 万語当たりの頻度
39	男性_rank	男性の順位
40	男性_frequency	男性の頻度
41	男性_pmw	男性全体での 100 万語当たりの頻度
42	女性_rank	女性の順位
43	女性_frequency	女性の頻度
44	女性_pmw	女性全体での 100 万語当たりの頻度
45	0-4 歳_rank	0-4 歳の順位
46	0-4 歳_frequency	0-4 歳の頻度
47	0-4 歳_pmw	0-4 歳全体での 100 万語当たりの頻度
48	5-9 歳_rank	5-9 歳の順位
49	5-9 歳_frequency	5-9 歳の頻度
50	5-9 歳_pmw	5-9 歳全体での 100 万語当たりの頻度
51	10-14 歳_rank	10-14 歳の順位

52	10-14 歳_frequency	10-14 歳の頻度
53	10-14 歳_pmw	10-14 歳全体での 100 万語当たりの頻度
54	15-19 歳_rank	15-19 歳の順位
55	15-19 歳_frequency	15-19 歳の頻度
56	15-19 歳_pmw	15-19 歳全体での 100 万語当たりの頻度
57	20-24 歳_rank	20-24 歳の順位
58	20-24 歳_frequency	20-24 歳の頻度
59	20-24 歳_pmw	20-24 歳全体での 100 万語当たりの頻度
60	25-29 歳_rank	25-29 歳の順位
61	25-29 歳_frequency	25-29 歳の頻度
62	25-29 歳_pmw	25-29 歳全体での 100 万語当たりの頻度
63	30-34 歳_rank	30-34 歳の順位
64	30-34 歳_frequency	30-34 歳の頻度
65	30-34 歳_pmw	30-34 歳全体での 100 万語当たりの頻度
66	35-39 歳_rank	35-39 歳の順位
67	35-39 歳_frequency	35-39 歳の頻度
68	35-39 歳_pmw	35-39 歳全体での 100 万語当たりの頻度
69	40-44 歳_rank	40-44 歳の順位
70	40-44 歳_frequency	40-44 歳の頻度
71	40-44 歳_pmw	40-44 歳全体での 100 万語当たりの頻度
72	45-49 歳_rank	45-49 歳の順位
73	45-49 歳_frequency	45-49 歳の頻度
74	45-49 歳_pmw	45-49 歳全体での 100 万語当たりの頻度
75	50-54 歳_rank	50-54 歳の順位
76	50-54 歳_frequency	50-54 歳の頻度
77	50-54 歳_pmw	50-54 歳全体での 100 万語当たりの頻度
78	55-59 歳_rank	55-59 歳の順位
79	55-59 歳_frequency	55-59 歳の頻度
80	55-59 歳_pmw	55-59 歳全体での 100 万語当たりの頻度
81	60-64 歳_rank	60-64 歳の順位
82	60-64 歳_frequency	60-64 歳の頻度
83	60-64 歳_pmw	60-64 歳全体での 100 万語当たりの頻度
84	65-69 歳_rank	65-69 歳の順位
85	65-69 歳_frequency	65-69 歳の頻度
86	65-69 歳_pmw	65-69 歳全体での 100 万語当たりの頻度
87	70-74 歳_rank	70-74 歳の順位

88	70-74 歳_frequency	70-74 歳の頻度
89	70-74 歳_pmw	70-74 歳全体での 100 万語当たりの頻度
90	75-79 歳_rank	75-79 歳の順位
91	75-79 歳_frequency	75-79 歳の頻度
92	75-79 歳_pmw	75-79 歳全体での 100 万語当たりの頻度
93	80-84 歳_rank	80-84 歳の順位
94	80-84 歳_frequency	80-84 歳の頻度
95	80-84 歳_pmw	80-84 歳全体での 100 万語当たりの頻度
96	85-89 歳_rank	85-89 歳の順位
97	85-89 歳_frequency	85-89 歳の頻度
98	85-89 歳_pmw	85-89 歳全体での 100 万語当たりの頻度
99	90-94 歳_rank	90-94 歳の順位
100	90-94 歳_frequency	90-94 歳の頻度
101	90-94 歳_pmw	90-94 歳全体での 100 万語当たりの頻度
102	雑談_屋外_rank	雑談_屋外の順位
103	雑談_屋外_frequency	雑談_屋外の頻度
104	雑談_屋外_pmw	雑談_屋外全体での 100 万語当たりの頻度
105	雑談_学校_rank	雑談_学校の順位
106	雑談_学校_frequency	雑談_学校の頻度
107	雑談_学校_pmw	雑談_学校全体での 100 万語当たりの頻度
108	雑談_交通機関_rank	雑談_交通機関の順位
109	雑談_交通機関_frequency	雑談_交通機関の頻度
110	雑談_交通機関_pmw	雑談_交通機関全体での 100 万語当たりの頻度
111	雑談_施設_rank	雑談_施設の順位
112	雑談_施設_frequency	雑談_施設の頻度
113	雑談_施設_pmw	雑談_施設全体での 100 万語当たりの頻度
114	雑談_自宅_rank	雑談_自宅の順位
115	雑談_自宅_frequency	雑談_自宅の頻度
116	雑談_自宅_pmw	雑談_自宅全体での 100 万語当たりの頻度
117	雑談_室内_rank	雑談_室内の順位
118	雑談_室内_frequency	雑談_室内の頻度
119	雑談_室内_pmw	雑談_室内全体での 100 万語当たりの頻度
120	雑談_職場_rank	雑談_職場の順位
121	雑談_職場_frequency	雑談_職場の頻度
122	雑談_職場_pmw	雑談_職場全体での 100 万語当たりの頻度
123	用談相談_屋外_rank	用談相談_屋外の順位

124	用談相談_屋外_frequency	用談相談_屋外の頻度
125	用談相談_屋外_pmw	用談相談_屋外全体での 100 万語当たりの頻度
126	用談相談_学校_rank	用談相談_学校の順位
127	用談相談_学校_frequency	用談相談_学校の頻度
128	用談相談_学校_pmw	用談相談_学校全体での 100 万語当たりの頻度
129	用談相談_交通機関_rank	用談相談_交通機関の順位
130	用談相談_交通機関_frequency	用談相談_交通機関の頻度
131	用談相談_交通機関_pmw	用談相談_交通機関全体での 100 万語当たりの頻度
132	用談相談_施設_rank	用談相談_施設の順位
133	用談相談_施設_frequency	用談相談_施設の頻度
134	用談相談_施設_pmw	用談相談_施設全体での 100 万語当たりの頻度
135	用談相談_自宅_rank	用談相談_自宅の順位
136	用談相談_自宅_frequency	用談相談_自宅の頻度
137	用談相談_自宅_pmw	用談相談_自宅全体での 100 万語当たりの頻度
138	用談相談_室内_rank	用談相談_室内の順位
139	用談相談_室内_frequency	用談相談_室内の頻度
140	用談相談_室内_pmw	用談相談_室内全体での 100 万語当たりの頻度
141	用談相談_職場_rank	用談相談_職場の順位
142	用談相談_職場_frequency	用談相談_職場の頻度
143	用談相談_職場_pmw	用談相談_職場全体での 100 万語当たりの頻度
144	会議会合_屋外_rank	会議会合_屋外の順位
145	会議会合_屋外_frequency	会議会合_屋外の頻度
146	会議会合_屋外_pmw	会議会合_屋外全体での 100 万語当たりの頻度
147	会議会合_学校_rank	会議会合_学校の順位
148	会議会合_学校_frequency	会議会合_学校の頻度
149	会議会合_学校_pmw	会議会合_学校全体での 100 万語当たりの頻度
150	会議会合_交通機関_rank	会議会合_交通機関の順位
151	会議会合_交通機関_frequency	会議会合_交通機関の頻度
152	会議会合_交通機関_pmw	会議会合_交通機関全体での 100 万語当たりの頻度
153	会議会合_施設_rank	会議会合_施設の順位
154	会議会合_施設_frequency	会議会合_施設の頻度
155	会議会合_施設_pmw	会議会合_施設全体での 100 万語当たりの頻度
156	会議会合_自宅_rank	会議会合_自宅の順位
157	会議会合_自宅_frequency	会議会合_自宅の頻度
158	会議会合_自宅_pmw	会議会合_自宅全体での 100 万語当たりの頻度
159	会議会合_室内_rank	会議会合_室内の順位

160	会議会合_室内_frequency	会議会合_室内の頻度
161	会議会合_室内_pmw	会議会合_室内全体での 100 万語当たりの頻度
162	会議会合_職場_rank	会議会合_職場の順位
163	会議会合_職場_frequency	会議会合_職場の頻度
164	会議会合_職場_pmw	会議会合_職場全体での 100 万語当たりの頻度
165	男性_0-4 歳_rank	男性_0-4 歳の順位
166	男性_0-4 歳_frequency	男性_0-4 歳の頻度
167	男性_0-4 歳_pmw	男性_0-4 歳全体での 100 万語当たりの頻度
168	男性_5-9 歳_rank	男性_5-9 歳の順位
169	男性_5-9 歳_frequency	男性_5-9 歳の頻度
170	男性_5-9 歳_pmw	男性_5-9 歳全体での 100 万語当たりの頻度
171	男性_10-14 歳_rank	男性_10-14 歳の順位
172	男性_10-14 歳_frequency	男性_10-14 歳の頻度
173	男性_10-14 歳_pmw	男性_10-14 歳全体での 100 万語当たりの頻度
174	男性_15-19 歳_rank	男性_15-19 歳の順位
175	男性_15-19 歳_frequency	男性_15-19 歳の頻度
176	男性_15-19 歳_pmw	男性_15-19 歳全体での 100 万語当たりの頻度
177	男性_20-24 歳_rank	男性_20-24 歳の順位
178	男性_20-24 歳_frequency	男性_20-24 歳の頻度
179	男性_20-24 歳_pmw	男性_20-24 歳全体での 100 万語当たりの頻度
180	男性_25-29 歳_rank	男性_25-29 歳の順位
181	男性_25-29 歳_frequency	男性_25-29 歳の頻度
182	男性_25-29 歳_pmw	男性_25-29 歳全体での 100 万語当たりの頻度
183	男性_30-34 歳_rank	男性_30-34 歳の順位
184	男性_30-34 歳_frequency	男性_30-34 歳の頻度
185	男性_30-34 歳_pmw	男性_30-34 歳全体での 100 万語当たりの頻度
186	男性_35-39 歳_rank	男性_35-39 歳の順位
187	男性_35-39 歳_frequency	男性_35-39 歳の頻度
188	男性_35-39 歳_pmw	男性_35-39 歳全体での 100 万語当たりの頻度
189	男性_40-44 歳_rank	男性_40-44 歳の順位
190	男性_40-44 歳_frequency	男性_40-44 歳の頻度
191	男性_40-44 歳_pmw	男性_40-44 歳全体での 100 万語当たりの頻度
192	男性_45-49 歳_rank	男性_45-49 歳の順位
193	男性_45-49 歳_frequency	男性_45-49 歳の頻度
194	男性_45-49 歳_pmw	男性_45-49 歳全体での 100 万語当たりの頻度
195	男性_50-54 歳_rank	男性_50-54 歳の順位

196	男性_50-54 歳_frequency	男性_50-54 歳の頻度
197	男性_50-54 歳_pmw	男性_50-54 歳全体での 100 万語当たりの頻度
198	男性_55-59 歳_rank	男性_55-59 歳の順位
199	男性_55-59 歳_frequency	男性_55-59 歳の頻度
200	男性_55-59 歳_pmw	男性_55-59 歳全体での 100 万語当たりの頻度
201	男性_60-64 歳_rank	男性_60-64 歳の順位
202	男性_60-64 歳_frequency	男性_60-64 歳の頻度
203	男性_60-64 歳_pmw	男性_60-64 歳全体での 100 万語当たりの頻度
204	男性_65-69 歳_rank	男性_65-69 歳の順位
205	男性_65-69 歳_frequency	男性_65-69 歳の頻度
206	男性_65-69 歳_pmw	男性_65-69 歳全体での 100 万語当たりの頻度
207	男性_70-74 歳_rank	男性_70-74 歳の順位
208	男性_70-74 歳_frequency	男性_70-74 歳の頻度
209	男性_70-74 歳_pmw	男性_70-74 歳全体での 100 万語当たりの頻度
210	男性_75-79 歳_rank	男性_75-79 歳の順位
211	男性_75-79 歳_frequency	男性_75-79 歳の頻度
212	男性_75-79 歳_pmw	男性_75-79 歳全体での 100 万語当たりの頻度
213	男性_80-84 歳_rank	男性_80-84 歳の順位
214	男性_80-84 歳_frequency	男性_80-84 歳の頻度
215	男性_80-84 歳_pmw	男性_80-84 歳全体での 100 万語当たりの頻度
216	男性_85-89 歳_rank	男性_85-89 歳の順位
217	男性_85-89 歳_frequency	男性_85-89 歳の頻度
218	男性_85-89 歳_pmw	男性_85-89 歳全体での 100 万語当たりの頻度
219	男性_90-94 歳_rank	男性_90-94 歳の順位
220	男性_90-94 歳_frequency	男性_90-94 歳の頻度
221	男性_90-94 歳_pmw	男性_90-94 歳全体での 100 万語当たりの頻度
222	女性_0-4 歳_rank	女性_0-4 歳の順位
223	女性_0-4 歳_frequency	女性_0-4 歳の頻度
224	女性_0-4 歳_pmw	女性_0-4 歳全体での 100 万語当たりの頻度
225	女性_5-9 歳_rank	女性_5-9 歳の順位
226	女性_5-9 歳_frequency	女性_5-9 歳の頻度
227	女性_5-9 歳_pmw	女性_5-9 歳全体での 100 万語当たりの頻度
228	女性_10-14 歳_rank	女性_10-14 歳の順位
229	女性_10-14 歳_frequency	女性_10-14 歳の頻度
230	女性_10-14 歳_pmw	女性_10-14 歳全体での 100 万語当たりの頻度
231	女性_15-19 歳_rank	女性_15-19 歳の順位

232	女性_15-19 歳_frequency	女性_15-19 歳の頻度
233	女性_15-19 歳_pmw	女性_15-19 歳全体での 100 万語当たりの頻度
234	女性_20-24 歳_rank	女性_20-24 歳の順位
235	女性_20-24 歳_frequency	女性_20-24 歳の頻度
236	女性_20-24 歳_pmw	女性_20-24 歳全体での 100 万語当たりの頻度
237	女性_25-29 歳_rank	女性_25-29 歳の順位
238	女性_25-29 歳_frequency	女性_25-29 歳の頻度
239	女性_25-29 歳_pmw	女性_25-29 歳全体での 100 万語当たりの頻度
240	女性_30-34 歳_rank	女性_30-34 歳の順位
241	女性_30-34 歳_frequency	女性_30-34 歳の頻度
242	女性_30-34 歳_pmw	女性_30-34 歳全体での 100 万語当たりの頻度
243	女性_35-39 歳_rank	女性_35-39 歳の順位
244	女性_35-39 歳_frequency	女性_35-39 歳の頻度
245	女性_35-39 歳_pmw	女性_35-39 歳全体での 100 万語当たりの頻度
246	女性_40-44 歳_rank	女性_40-44 歳の順位
247	女性_40-44 歳_frequency	女性_40-44 歳の頻度
248	女性_40-44 歳_pmw	女性_40-44 歳全体での 100 万語当たりの頻度
249	女性_45-49 歳_rank	女性_45-49 歳の順位
250	女性_45-49 歳_frequency	女性_45-49 歳の頻度
251	女性_45-49 歳_pmw	女性_45-49 歳全体での 100 万語当たりの頻度
252	女性_50-54 歳_rank	女性_50-54 歳の順位
253	女性_50-54 歳_frequency	女性_50-54 歳の頻度
254	女性_50-54 歳_pmw	女性_50-54 歳全体での 100 万語当たりの頻度
255	女性_55-59 歳_rank	女性_55-59 歳の順位
256	女性_55-59 歳_frequency	女性_55-59 歳の頻度
257	女性_55-59 歳_pmw	女性_55-59 歳全体での 100 万語当たりの頻度
258	女性_60-64 歳_rank	女性_60-64 歳の順位
259	女性_60-64 歳_frequency	女性_60-64 歳の頻度
260	女性_60-64 歳_pmw	女性_60-64 歳全体での 100 万語当たりの頻度
261	女性_65-69 歳_rank	女性_65-69 歳の順位
262	女性_65-69 歳_frequency	女性_65-69 歳の頻度
263	女性_65-69 歳_pmw	女性_65-69 歳全体での 100 万語当たりの頻度
264	女性_70-74 歳_rank	女性_70-74 歳の順位
265	女性_70-74 歳_frequency	女性_70-74 歳の頻度
266	女性_70-74 歳_pmw	女性_70-74 歳全体での 100 万語当たりの頻度
267	女性_75-79 歳_rank	女性_75-79 歳の順位

268	女性_75-79 歳_frequency	女性_75-79 歳の頻度
269	女性_75-79 歳_pmw	女性_75-79 歳全体での 100 万語当たりの頻度
270	女性_80-84 歳_rank	女性_80-84 歳の順位
271	女性_80-84 歳_frequency	女性_80-84 歳の頻度
272	女性_80-84 歳_pmw	女性_80-84 歳全体での 100 万語当たりの頻度
273	女性_85-89 歳_rank	女性_85-89 歳の順位
274	女性_85-89 歳_frequency	女性_85-89 歳の頻度
275	女性_85-89 歳_pmw	女性_85-89 歳全体での 100 万語当たりの頻度
276	女性_90-94 歳_rank	女性_90-94 歳の順位
277	女性_90-94 歳_frequency	女性_90-94 歳の頻度
278	女性_90-94 歳_pmw	女性_90-94 歳全体での 100 万語当たりの頻度

３．２　書字形別の CEJC 語彙表

・ファイル名：3_cejc_frequencylist_shozikei.xlsx
・15,497 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 行目以降がデータである。各行には語彙表と同じ項目に加え，「書字形」を加えた 279 の項目が並んでいる。

３．３　発音形別の CEJC 語彙表

・ファイル名：4_cejc_frequencylist_hatuonkei.xlsx
・18,351 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 行目以降がデータである。各行には語彙表と同じ項目に加え，「発音形出現形」を加えた 279 の項目が並んでいる。

３．４　性別と年齢_年齢まとめと年齢不明付記版の CEJC 語彙表

・ファイル名：5_CEJC_語彙表_性別と年齢_年齢まとめと年齢不明付記版.xlsx
・14,418 行。
・第 1 行目は見出し。2 行目以降がデータである。各行には語彙表と同じく，基本項目（語彙素等）と性別・年齢（5 歳きざみ）に関する項目のほか，新たに年齢まとめ（10 代以下と，70 代以上は基本 10 歳きざみ）と，性別の年齢不明項目を加えた 305 の項目が並んでいる。

表 2　語彙表の各項目（色付きの項目は，本語彙表にのみ付記したもの）

番号	見出し	備考
1	rank	CEJC 全体の順位
2	語彙素読み	語彙素読み
3	語彙素	語彙素
4	品詞	品詞
5	語彙素細分類	語彙素細分類
6	語種	語種
7	frequency	CEJC 全体の頻度
8	pmw	CEJC 全体での 100 万語当たりの頻度
9	男性_rank	男性の順位
10	男性_frequency	男性の頻度
11	男性_pmw	男性全体での 100 万語当たりの頻度
12	女性_rank	女性の順位
13	女性_frequency	女性の頻度
14	女性_pmw	女性全体での 100 万語当たりの頻度
15	0-19 歳_rank	0-19 歳の順位
16	0-19 歳_frequency	0-19 歳の頻度
17	0-19 歳_pmw	0-19 歳全体での 100 万語当たりの頻度
18	20-29 歳_rank	20-29 歳の順位
19	20-29 歳_frequency	20-29 歳の頻度
20	20-29 歳_pmw	20-29 歳全体での 100 万語当たりの頻度
21	30-39 歳_rank	30-39 歳の順位
22	30-39 歳_frequency	30-39 歳の頻度
23	30-39 歳_pmw	30-39 歳全体での 100 万語当たりの頻度
24	40-49 歳_rank	40-49 歳の順位
25	40-49 歳_frequency	40-49 歳の頻度
26	40-49 歳_pmw	40-49 歳全体での 100 万語当たりの頻度
27	50-59 歳_rank	50-59 歳の順位
28	50-59 歳_frequency	50-59 歳の頻度
29	50-59 歳_pmw	50-59 歳全体での 100 万語当たりの頻度
30	60-69 歳_rank	60-69 歳の順位
31	60-69 歳_frequency	60-69 歳の頻度
32	60-69 歳_pmw	60-69 歳全体での 100 万語当たりの頻度
33	70-94 歳_rank	70-94 歳の順位
34	70-94 歳_frequency	70-94 歳の頻度
35	70-94 歳_pmw	70-94 歳全体での 100 万語当たりの頻度

36	0-4 歳_rank	0-4 歳の順位
37	0-4 歳_frequency	0-4 歳の頻度
38	0-4 歳_pmw	0-4 歳全体での 100 万語当たりの頻度
39	5-9 歳_rank	5-9 歳の順位
40	5-9 歳_frequency	5-9 歳の頻度
41	5-9 歳_pmw	5-9 歳全体での 100 万語当たりの頻度
42	10-14 歳_rank	10-14 歳の順位
43	10-14 歳_frequency	10-14 歳の頻度
44	10-14 歳_pmw	10-14 歳全体での 100 万語当たりの頻度
45	15-19 歳_rank	15-19 歳の順位
46	15-19 歳_frequency	15-19 歳の頻度
47	15-19 歳_pmw	15-19 歳全体での 100 万語当たりの頻度
48	20-24 歳_rank	20-24 歳の順位
49	20-24 歳_frequency	20-24 歳の頻度
50	20-24 歳_pmw	20-24 歳全体での 100 万語当たりの頻度
51	25-29 歳_rank	25-29 歳の順位
52	25-29 歳_frequency	25-29 歳の頻度
53	25-29 歳_pmw	25-29 歳全体での 100 万語当たりの頻度
54	30-34 歳_rank	30-34 歳の順位
55	30-34 歳_frequency	30-34 歳の頻度
56	30-34 歳_pmw	30-34 歳全体での 100 万語当たりの頻度
57	35-39 歳_rank	35-39 歳の順位
58	35-39 歳_frequency	35-39 歳の頻度
59	35-39 歳_pmw	35-39 歳全体での 100 万語当たりの頻度
60	40-44 歳_rank	40-44 歳の順位
61	40-44 歳_frequency	40-44 歳の頻度
62	40-44 歳_pmw	40-44 歳全体での 100 万語当たりの頻度
63	45-49 歳_rank	45-49 歳の順位
64	45-49 歳_frequency	45-49 歳の頻度
65	45-49 歳_pmw	45-49 歳全体での 100 万語当たりの頻度
66	50-54 歳_rank	50-54 歳の順位
67	50-54 歳_frequency	50-54 歳の頻度
68	50-54 歳_pmw	50-54 歳全体での 100 万語当たりの頻度
69	55-59 歳_rank	55-59 歳の順位
70	55-59 歳_frequency	55-59 歳の頻度
71	55-59 歳_pmw	55-59 歳全体での 100 万語当たりの頻度
72	60-64 歳_rank	60-64 歳の順位

73	60-64 歳_frequency	60-64 歳の頻度
74	60-64 歳_pmw	60-64 歳全体での 100 万語当たりの頻度
75	65-69 歳_rank	65-69 歳の順位
76	65-69 歳_frequency	65-69 歳の頻度
77	65-69 歳_pmw	65-69 歳全体での 100 万語当たりの頻度
78	70-74 歳_rank	70-74 歳の順位
79	70-74 歳_frequency	70-74 歳の頻度
80	70-74 歳_pmw	70-74 歳全体での 100 万語当たりの頻度
81	75-79 歳_rank	75-79 歳の順位
82	75-79 歳_frequency	75-79 歳の頻度
83	75-79 歳_pmw	75-79 歳全体での 100 万語当たりの頻度
84	80-84 歳_rank	80-84 歳の順位
85	80-84 歳_frequency	80-84 歳の頻度
86	80-84 歳_pmw	80-84 歳全体での 100 万語当たりの頻度
87	85-89 歳_rank	85-89 歳の順位
88	85-89 歳_frequency	85-89 歳の頻度
89	85-89 歳_pmw	85-89 歳全体での 100 万語当たりの頻度
90	90-94 歳_rank	90-94 歳の順位
91	90-94 歳_frequency	90-94 歳の頻度
92	90-94 歳_pmw	90-94 歳全体での 100 万語当たりの頻度
93	男性_0-4 歳_rank	男性_0-4 歳の順位
94	男性_0-4 歳_frequency	男性_0-4 歳の頻度
95	男性_0-4 歳_pmw	男性_0-4 歳全体での 100 万語当たりの頻度
96	男性_5-9 歳_rank	男性_5-9 歳の順位
97	男性_5-9 歳_frequency	男性_5-9 歳の頻度
98	男性_5-9 歳_pmw	男性_5-9 歳全体での 100 万語当たりの頻度
99	男性_10-14 歳_rank	男性_10-14 歳の順位
100	男性_10-14 歳_frequency	男性_10-14 歳の頻度
101	男性_10-14 歳_pmw	男性_10-14 歳全体での 100 万語当たりの頻度
102	男性_15-19 歳_rank	男性_15-19 歳の順位
103	男性_15-19 歳_frequency	男性_15-19 歳の頻度
104	男性_15-19 歳_pmw	男性_15-19 歳全体での 100 万語当たりの頻度
105	男性_20-24 歳_rank	男性_20-24 歳の順位
106	男性_20-24 歳_frequency	男性_20-24 歳の頻度
107	男性_20-24 歳_pmw	男性_20-24 歳全体での 100 万語当たりの頻度
108	男性_25-29 歳_rank	男性_25-29 歳の順位
109	男性_25-29 歳_frequency	男性_25-29 歳の頻度

110	男性_25-29 歳_pmw	男性_25-29 歳全体での 100 万語当たりの頻度
111	男性_30-34 歳_rank	男性_30-34 歳の順位
112	男性_30-34 歳_frequency	男性_30-34 歳の頻度
113	男性_30-34 歳_pmw	男性_30-34 歳全体での 100 万語当たりの頻度
114	男性_35-39 歳_rank	男性_35-39 歳の順位
115	男性_35-39 歳_frequency	男性_35-39 歳の頻度
116	男性_35-39 歳_pmw	男性_35-39 歳全体での 100 万語当たりの頻度
117	男性_40-44 歳_rank	男性_40-44 歳の順位
118	男性_40-44 歳_frequency	男性_40-44 歳の頻度
119	男性_40-44 歳_pmw	男性_40-44 歳全体での 100 万語当たりの頻度
120	男性_45-49 歳_rank	男性_45-49 歳の順位
121	男性_45-49 歳_frequency	男性_45-49 歳の頻度
122	男性_45-49 歳_pmw	男性_45-49 歳全体での 100 万語当たりの頻度
123	男性_50-54 歳_rank	男性_50-54 歳の順位
124	男性_50-54 歳_frequency	男性_50-54 歳の頻度
125	男性_50-54 歳_pmw	男性_50-54 歳全体での 100 万語当たりの頻度
126	男性_55-59 歳_rank	男性_55-59 歳の順位
127	男性_55-59 歳_frequency	男性_55-59 歳の頻度
128	男性_55-59 歳_pmw	男性_55-59 歳全体での 100 万語当たりの頻度
129	男性_60-64 歳_rank	男性_60-64 歳の順位
130	男性_60-64 歳_frequency	男性_60-64 歳の頻度
131	男性_60-64 歳_pmw	男性_60-64 歳全体での 100 万語当たりの頻度
132	男性_65-69 歳_rank	男性_65-69 歳の順位
133	男性_65-69 歳_frequency	男性_65-69 歳の頻度
134	男性_65-69 歳_pmw	男性_65-69 歳全体での 100 万語当たりの頻度
135	男性_70-74 歳_rank	男性_70-74 歳の順位
136	男性_70-74 歳_frequency	男性_70-74 歳の頻度
137	男性_70-74 歳_pmw	男性_70-74 歳全体での 100 万語当たりの頻度
138	男性_75-79 歳_rank	男性_75-79 歳の順位
139	男性_75-79 歳_frequency	男性_75-79 歳の頻度
140	男性_75-79 歳_pmw	男性_75-79 歳全体での 100 万語当たりの頻度
141	男性_80-84 歳_rank	男性_80-84 歳の順位
142	男性_80-84 歳_frequency	男性_80-84 歳の頻度
143	男性_80-84 歳_pmw	男性_80-84 歳全体での 100 万語当たりの頻度
144	男性_85-89 歳_rank	男性_85-89 歳の順位
145	男性_85-89 歳_frequency	男性_85-89 歳の頻度
146	男性_85-89 歳_pmw	男性_85-89 歳全体での 100 万語当たりの頻度

147	男性_90-94 歳_rank	男性_90-94 歳の順位
148	男性_90-94 歳_frequency	男性_90-94 歳の頻度
149	男性_90-94 歳_pmw	男性_90-94 歳全体での 100 万語当たりの頻度
150	女性_0-4 歳_rank	女性_0-4 歳の順位
151	女性_0-4 歳_frequency	女性_0-4 歳の頻度
152	女性_0-4 歳_pmw	女性_0-4 歳全体での 100 万語当たりの頻度
153	女性_5-9 歳_rank	女性_5-9 歳の順位
154	女性_5-9 歳_frequency	女性_5-9 歳の頻度
155	女性_5-9 歳_pmw	女性_5-9 歳全体での 100 万語当たりの頻度
156	女性_10-14 歳_rank	女性_10-14 歳の順位
157	女性_10-14 歳_frequency	女性_10-14 歳の頻度
158	女性_10-14 歳_pmw	女性_10-14 歳全体での 100 万語当たりの頻度
159	女性_15-19 歳_rank	女性_15-19 歳の順位
160	女性_15-19 歳_frequency	女性_15-19 歳の頻度
161	女性_15-19 歳_pmw	女性_15-19 歳全体での 100 万語当たりの頻度
162	女性_20-24 歳_rank	女性_20-24 歳の順位
163	女性_20-24 歳_frequency	女性_20-24 歳の頻度
164	女性_20-24 歳_pmw	女性_20-24 歳全体での 100 万語当たりの頻度
165	女性_25-29 歳_rank	女性_25-29 歳の順位
166	女性_25-29 歳_frequency	女性_25-29 歳の頻度
167	女性_25-29 歳_pmw	女性_25-29 歳全体での 100 万語当たりの頻度
168	女性_30-34 歳_rank	女性_30-34 歳の順位
169	女性_30-34 歳_frequency	女性_30-34 歳の頻度
170	女性_30-34 歳_pmw	女性_30-34 歳全体での 100 万語当たりの頻度
171	女性_35-39 歳_rank	女性_35-39 歳の順位
172	女性_35-39 歳_frequency	女性_35-39 歳の頻度
173	女性_35-39 歳_pmw	女性_35-39 歳全体での 100 万語当たりの頻度
174	女性_40-44 歳_rank	女性_40-44 歳の順位
175	女性_40-44 歳_frequency	女性_40-44 歳の頻度
176	女性_40-44 歳_pmw	女性_40-44 歳全体での 100 万語当たりの頻度
177	女性_45-49 歳_rank	女性_45-49 歳の順位
178	女性_45-49 歳_frequency	女性_45-49 歳の頻度
179	女性_45-49 歳_pmw	女性_45-49 歳全体での 100 万語当たりの頻度
180	女性_50-54 歳_rank	女性_50-54 歳の順位
181	女性_50-54 歳_frequency	女性_50-54 歳の頻度
182	女性_50-54 歳_pmw	女性_50-54 歳全体での 100 万語当たりの頻度
183	女性_55-59 歳_rank	女性_55-59 歳の順位

184	女性_55-59 歳_frequency	女性_55-59 歳の頻度
185	女性_55-59 歳_pmw	女性_55-59 歳全体での 100 万語当たりの頻度
186	女性_60-64 歳_rank	女性_60-64 歳の順位
187	女性_60-64 歳_frequency	女性_60-64 歳の頻度
188	女性_60-64 歳_pmw	女性_60-64 歳全体での 100 万語当たりの頻度
189	女性_65-69 歳_rank	女性_65-69 歳の順位
190	女性_65-69 歳_frequency	女性_65-69 歳の頻度
191	女性_65-69 歳_pmw	女性_65-69 歳全体での 100 万語当たりの頻度
192	女性_70-74 歳_rank	女性_70-74 歳の順位
193	女性_70-74 歳_frequency	女性_70-74 歳の頻度
194	女性_70-74 歳_pmw	女性_70-74 歳全体での 100 万語当たりの頻度
195	女性_75-79 歳_rank	女性_75-79 歳の順位
196	女性_75-79 歳_frequency	女性_75-79 歳の頻度
197	女性_75-79 歳_pmw	女性_75-79 歳全体での 100 万語当たりの頻度
198	女性_80-84 歳_rank	女性_80-84 歳の順位
199	女性_80-84 歳_frequency	女性_80-84 歳の頻度
200	女性_80-84 歳_pmw	女性_80-84 歳全体での 100 万語当たりの頻度
201	女性_85-89 歳_rank	女性_85-89 歳の順位
202	女性_85-89 歳_frequency	女性_85-89 歳の頻度
203	女性_85-89 歳_pmw	女性_85-89 歳全体での 100 万語当たりの頻度
204	女性_90-94 歳_rank	女性_90-94 歳の順位
205	女性_90-94 歳_frequency	女性_90-94 歳の頻度
206	女性_90-94 歳_pmw	女性_90-94 歳全体での 100 万語当たりの頻度
207	男性_NA_rank	男性年齢不明の順位
208	男性_NA_frequency	男性年齢不明の頻度
209	男性_NA_pmw	男性年齢不明全体での 100 万語当たりの頻度
210	女性_NA_rank	女性年齢不明の順位
211	女性_NA_frequency	女性年齢不明の頻度
212	女性_NA_pmw	女性年齢不明全体での 100 万語当たりの頻度

３．５　分類語彙表番号付与の CEJC 語彙表

・ファイル名：6_CEJC_語彙表_分類語彙表番号付与版.xlsx
・20,114 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，語彙分類番号の 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 行目以降がデータである。各行には語彙表と同じ項目に加え，下記を加えた 287 の項目が並んでいる。

・「分類語彙表番号－UniDic 語彙素番号対応表」を利用して『分類語彙表』の分類番号を付与したものである。該当なしの場合はすべて「-」が挿入されている。
・『分類語彙表』の分類体系に関する情報は，山崎誠・小沼悦（2003）「研究室から：『分類語彙表』増補改訂版について」『国語研の窓』15[4]及び，柏野和佳子（2006）「『分類語彙表』の特徴と位置付け」『日本語科学』19，pp.143-160[5]を参照のこと。

表 3　語彙表に追加した項目

番号	見出し	備考
7	UniDic 語彙素番号	UniDic 語彙素番号
8	語彙分類番号	『分類語彙表』語彙分類番号
9	分類項目_類	『分類語彙表』分類項目_類
10	分類項目_部門	『分類語彙表』分類項目_部門
11	分類項目_中項目	『分類語彙表』分類項目_中項目
12	分類項目_分類項目ラベル	『分類語彙表』分類項目_分類項目ラベル
13	段落番号	『分類語彙表』段落番号
14	小段落番号	『分類語彙表』小段落番号
15	語番号	『分類語彙表』語番号

４．CEJC 語数表

４．１　CEJC 語数表

・ファイル名：7_cejc_wc.xlsx
・会話 ID，話者別の延べ語数表。。

表 4　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	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
3	会話概要	会話の概要
4	会話時間	会話の時間（分），整数

[4] https://kotobaken.jp/mado/15/15-02/
[5] http://doi.org/10.15084/00002158

5	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
6	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／仕事関係／先生生徒／サービス場面関係／初対面
7	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意に認定 値）雑談／用談相談／会議会合
8	場所	会話が行われた場所 値）自宅／職場／学校／屋外／室内（実家／親類宅／知人宅）／交通機関（車）／施設（飲食店／公共施設／商業施設／医療福祉施設／宿泊施設）
9	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）食事／家事雑事／身の周りの用事／仕事／学業／社会参加／レジャー活動／付き合い／移動／休息
10	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
11	話者ラベル	話者に与えられた仮名（カメイ）のラベル 例）高橋，修司，パパ
12	年齢	収録当時の年齢（5 歳刻み），欠損値あり 例）0-4 歳，15-19 歳，30-34 歳，90-94 歳
13	性別	話者の性別 値）男性／女性
14	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），欠損値あり 例）北海道，宮城県，東京都，京都府，中国
15	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

16	職業	話者の職業，欠損値あり 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／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／NA）
17	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 値）本人／家族親族（夫／妻／父／母／息子／娘／兄／姉／弟／妹／祖父／祖母／婿／甥／姪／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／仕事関係者（上司／先輩／同僚／部下／NA）習い事などの先生／習い事などの生徒／友人知人（先輩／同級生／後輩／NA）／サービスを受ける人／サービスを提供する人
18	語数(全て)	整数
19	語数(記号等除外・全て)	整数

4．2　形式・場面・年齢・性別の語数表

・ファイル名：8_CEJC_語数表_形式と場面と年齢と性別.xlsx
・形式，場面，年齢（10 代以下と，70 代以上は基本 10 歳きざみ），性別の延べ語数表。
・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.tsv（xlsx 形式もあり: 10_cejc_frequencylist_pos.xlsx）
・67 行，UTF8，タブ区切り。
・以下の 4 個の表を納めた。
(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 行目以降がデータである。列は語彙表と同じ種類のものが並んでいる。
・品詞の割合（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：11_cejc_frequencylist_wtype.x tsv（xlsx 形式もあり: 12_cejc_frequencylist_wtype.xlsx）
・39 行，UTF8，タブ区切り。
・CEJC 品詞構成表と同様に 4 個の表を納めた。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 行目以降がデータである。列は語彙表と同じ種類のものが並んでいる。
・語種の割合（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，出典とバーションの例を示す。
　『日本語日常会話コーパス』語彙表 ver.2020.03
　『日本語日常会話コーパス』語数表 ver.2020.03
　『日本語日常会話コーパス』品詞構成表 ver.2020.03
　『日本語日常会話コーパス』語種構成表 ver.2020.03
(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