『日本語日常会話コーパス』モニター公開版の語彙

山崎 誠，大村 舞
人間文化研究機構国立国語研究所
yamazaki@ninjal.ac.jp，mai-om@ninjal.ac.jp

1 はじめに

国立国語研究所では，2018 年 12 月に『日本語日常会話コーパス』のモニター公開を開始した．このコーパスは，同所の機関拠点型基幹研究プロジェクトの 1 つである「大規模日常会話に基づく話し言葉の多角的研究[1]」(プロジェクトリーダー：小磯花絵)で構築しているものである．本稿では，このモニター公開版の語彙から見た日常会話の特徴を同じ話し言葉のコーパスである，『名大会話コーパス』『職場談話コーパス』『日本語話し言葉コーパス』などと比べることにより明らかにする．とくに話者の属性や会話の属性からみた語彙の量的な観察を行う．

2 『日本語日常会話コーパス』

『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation, CEJC）は，「さまざまなタイプの日常会話 200 時間をバランス良く納めた大規模な」コーパスであり，「調査者は立ち会わず、生活の中で生じる会話を会話者自身に収録してもらうことで、日常会話をより自然な形で記録する点に特色があ」る（以上，「」内はプロジェクト HP より引用）

2018 年 12 月に同コーパスのモニター公開版（以降，CEJC モニター版と呼ぶ）約 50 時間が公開された．このモニター公開版には映像・音声・転記テキストなどが含まれている．また，WEB 上の検索ツール「中納言」でもモニター公開版を検索できるようになった．「中納言」での公開に合わせて，モニター公開版の語彙表・語形表を公開することになっている．本発表では，CEJC モニター版を使って，日常会話の語彙的な特徴を明らかにする．

3 語彙量

3.1 全体の語彙量

表 1 は CEJC モニター版をはじめとする，話し言葉のコーパスの延べ語数と異なり語数である[2]．今回公開した CEJC モニター版は『名大会話コーパス』の約半分の延べ語数であるが，会話数は『名大会話コーパス』とほぼ同じ数である．CEJC モニター版の時間数が約 50 時間，『名大会話コーパス』の時間数が約 100 時間であることを考えると，CEJC モニター版には 1 会話当たりの時間数が相対的に短い会話が収録されていることが分かる．

表 1　各コーパスの語数と会話数

コーパス	延べ語数	異なり語数	会話数
CEJC モニター版	609327	14417	126
名大会話コーパス	1129271	18186	129
職場談話コーパス	183884	7189	1324
CSJ 学会講演	3255168	23172	987
CSJ 模擬講演	3592080	32682	1715

図 1，2 は 126 会話の延べ語数と異なり語数の分布である．図 1，2 から延べ語数，異なり語数ともに若干右に歪んだ分布であることが分かる（それぞれの歪度は 2.90 と 2.45）

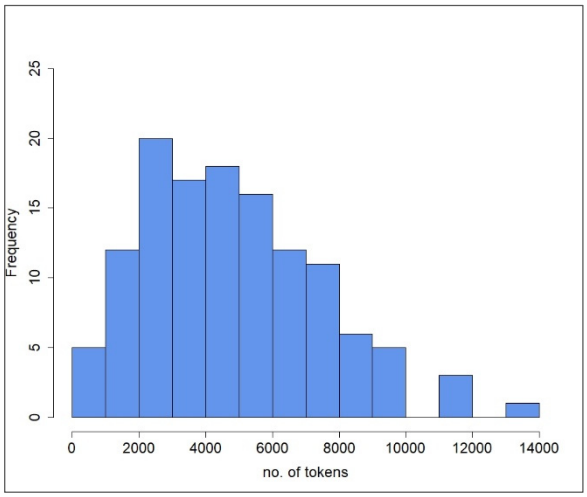

図 1　CEJC モニター公開版の延べ語数の分布

[1] プロジェクトの詳細は，https://pj.ninjal.ac.jp/conversation/corpus.html　を参照されたい．また，モニター公開版の概要については，小磯他（2019）を参照されたい．

[2] 語数はいずれも短単位で品詞が「空白」「補助記号」「記号」を除いたものである．CEJC モニター公開版の場合は，「伏せ字」「形態論情報付与対象外」「歌」「記号」を除いたものである．

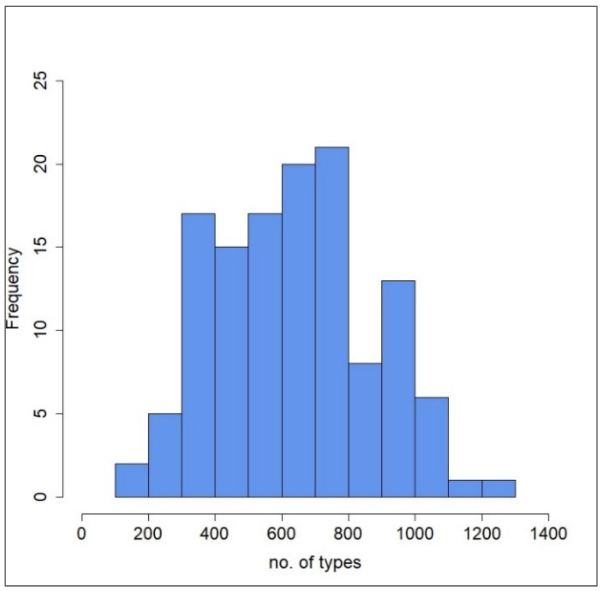

図 2　CEJC モニター公開版の異なり語数の分布

3.2　品詞と語種

表 2 は CEJC モニター公開版の品詞の数とその割合である．山崎（2017：281）では，CEJC も含め，他の話し言葉コーパスにおける品詞の割合（延べ語数）と比較し，CEJC の特徴として，形容詞，副詞，感動詞の割合が多く，連体詞の割合が低いと指摘しているが，その傾向は今回も同様と認められる．

表 2　CEJC モニター公開版の品詞

POS	延べ語数	割合	異なり語数	割合
名詞	106228	0.174	10745	0.745
代名詞	23579	0.039	46	0.003
動詞	64225	0.105	1401	0.097
形容詞	18644	0.031	227	0.016
形状詞	7282	0.012	340	0.024
副詞	37297	0.061	923	0.064
連体詞	4721	0.008	23	0.002
接続詞	4279	0.007	18	0.001
感動詞	64297	0.106	202	0.014
助詞	178749	0.293	85	0.006
助動詞	76091	0.125	38	0.003
接頭辞	3755	0.006	67	0.005
接尾辞	11540	0.019	301	0.021
言いよどみ	8640	0.014	1	0.000
計	609327	1.000	14417	1.000

表 3 に CEJC モニター公開版の語種を，表 4 に比較のために名大会話コーパスの語種を掲げた．若干の違いはあるもの，延べ・異なりともに同じような分布を示している．

表 3　CEJC モニター公開版の語種

語種	延べ語数	割合	異なり語数	割合
和語	517894	0.850	5472	0.380
漢語	56713	0.093	4455	0.309
外来語	12264	0.020	1956	0.136
混種語	4366	0.007	489	0.034
固有名詞	9288	0.015	1996	0.138
記号	162	0.000	48	0.003
その他	8640	0.014	1	0.000
合計	609327	1.000	14417	1.000

表 4　名大会話コーパスの語種

語種	延べ語数	割合	異なり語数	割合
和語	981760	0.869	6798	0.380
漢語	108986	0.096	5716	0.319
外来語	18621	0.016	2694	0.151
混種語	7271	0.006	536	0.030
固有名詞	11522	0.010	2040	0.114
記号	574	0.001	113	0.006
計	1129982	1.000	17897	1.000

3.3 会話の種類

次に，会話の種類による語彙の量を見る．会話の種類としては，形式，場所，活動の 3 つがあり，モニター版ではそれぞれ，以下のような属性を持つ．

形式：会議会合，雑談，用談相談
場所：屋外，学校，交通機関_車，施設_医療福祉施設，施設_飲食店，施設_宿泊施設，施設_商業施設，自宅，室内_実家，室内_親類宅，室内_知人宅，職場
活動：レジャー活動，レジャー活動・移動，移動，家事雑事，家事雑事・食事，学業，休息，仕事，社会参加，社会参加・食事，食事，身の回りの用事，付き合い，付き合い・食事

これらの会話の種類別に延べ語数を集計したのが表 5〜7 である．表 5 の会話の形式を見ると，延べ語数の約 7 割[3]を占めるのが「雑談」である．雑談は会話数でも約 7 割を占めている．延べ語数を会話数で割

[3] 表 2 の「割合」は延べ語数の計に対するそれぞれの属性の持つ延べ語数の割合である．以下の表も同じ．

ると 1 会話あたりの延べ語数が算出されるが，その値は，「会議会合」がやや多く，約 5863 語となっている．

表 5　会話形式と延べ語数

形式	延べ語数	会話数	割合
雑談	434541	91	0.713
用談相談	122017	26	0.200
会議会合	52769	9	0.087
計	609327	126	1.00

会話の場所では，施設_飲食店が多く全体の 3 割を占める．その次が自宅で約 2 割を占めている．活動の属性としては，付き合い・食事が一番多く約 2 割を占める．食事は，単独の活動として，また，「家事雑事・食事」「社会参加・食事」としても現れており，それらを合計すると 221268 語となり，全体の約 36% を占める．CEJC モニター公開版の会話の約 1/3 は，食事が関係する会話であることが分かる．

表 6　会話の場所と延べ語数

場所	延べ語数	会話数	割合
自宅	122727	30	0.201
室内_実家	30891	4	0.051
室内_知人宅	40482	7	0.066
室内_親類宅	21122	5	0.035
交通機関_車	25231	4	0.041
屋外	20957	5	0.034
施設_公共施設	42416	6	0.070
施設_医療福祉施設	5480	2	0.009
施設_商業施設	12127	3	0.020
施設_宿泊施設	8560	1	0.014
施設_飲食店	193404	40	0.317
職場	32723	8	0.054
学校	53207	11	0.087
計	609327	126	1.000

表 7　会話時の活動と延べ語数

活動	延べ語数	会話数	割合
家事雑事	29799	7	0.049
家事雑事・食事	23114	3	0.038
身の周りの用事	13717	4	0.023
食事	72775	19	0.119
付き合い	89666	15	0.147
付き合い・食事	116801	25	0.192
仕事	53581	10	0.088
学業	22729	7	0.037
社会参加	57792	9	0.095
社会参加・食事	8578	2	0.014
休息	74587	16	0.122
レジャー活動	10129	2	0.017
レジャー活動・移動	3407	1	0.006
移動	32652	6	0.054
計	609327	126	1.000

3.4 話者の属性

表 8 に話者の性別と延べ語数を示した．人数，延べ語数ともに若干女性が多くなっている．

表 8　性別と延べ語数

性別	延べ語数	人数	割合
女性	349144	255	0.573
男性	260183	208	0.427
計	609327	463	1.000

図 3 は，話者の年代と延べ語数の分布である．20 代から 50 代の成年層が多くを占めるが，60 代，70 代の話者による語数も一定数を占め，各年齢層からまんべんなく会話が採られていることが分かる．

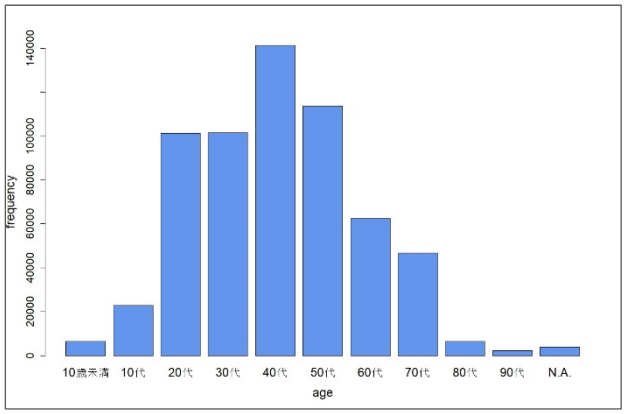

図 3　CEJC モニター公開版の話者の年代と延べ語数

4 語彙表と語形表

CEJC モニター公開版の公開に合わせ，語彙表と語形表を公開する予定である．語彙表は，Web 検索ツール「中納言」で公開している他のコーパスと同様に語彙素による集計で作成したものであり，複数のコーパスの比較に便利である．CEJC モニター公開版では，そのほかに語形表を公開する．語形とは，形態論情報の一つで，語彙素の下に位置し，異語形を区別するレベルである（小木曽 2014 : 101）．例えば，語彙素による集計では語彙素「矢張り」という

語が何回使われたかが分かるだけであるが，語形による集計では,「やはり」「やっぱり」「やっぱし」「やっぱ」のような異語形ごとの出現頻度が分かる．ちなみに，CEJC モニター公開版における語彙素「矢張り」の各語形の頻度は,「やはり 0」「やっぱり 525」「やっぱし 9」「やっぱ 435」である．

5 特徴語

最後に CEJC モニター公開版を使って日常会話における特徴語を見る．特徴語の抽出には LLR（対数尤度比）を利用する（Kilgarriff,1996: 3）.

5.1 会話の属性

表 9 は，会話の形式の違いによる特徴語である．上位 20 語を挙げた. 雑談では,「食べる」「美味しい」「飲む」など食事関係の語が上位に来ている.「用談相談」では,「日」「日程」「月」「待つ」などスケジュールに関する語が目立つ.「会議」では,「ルール」「委員」「テーマ」「抑（そもそも）」などが他との違いとして挙げられる．それぞれの会話の形式に現れる感動詞にも違いが見られる. 雑談では,「ふん」「へえ」であり，用談相談では「はい」「よいしょ」，会議会合では,「うん」「えー」「えーと」である.

表 9　会話の形式による特徴語

雑談	用談相談	会議会合
食べる, 美味しい，の（終助詞），ふん（感動詞），さ（助詞），飲む，じゃん（助詞），た（助動詞），父（とう），何（なに），ちゃん（接尾辞），てる（助動詞）俺, パパ, 家（うち），ちび，御前（代名詞），んっ（感動詞），味（あじ），へえ（感動詞）	はい（感動詞），です, 雲（名詞）ます，よいしょ, 日（ひ）, 待つ, 日程, か（終助詞）, 一寸（ちょっと），ノン（固有名詞），グラム, 月（がつ），良い，此れ, 二十，えーと，を（格助詞），ゼミ，次（じ）	うん，えー月（がつ）ルール，です，さん（接尾辞），ポスター, えーとね（終助詞），カラー，冊，名（めい），十（とお）, は（係助詞），九（く）委員，テーマ，抑（そもそも）無し，そう（副詞）

5.2 話者の属性

表 10 は話者の属性による特徴語である. 男性の特徴語に男性的とされる代名詞「俺」「僕」「御前」があるのに対して, 女性の特徴語には代名詞が少なく，上位 20 語には「私（わたし）」1 語である．また，女性的とされてきた終助詞「わ」は 123 位,「かしら」は 1973 位とかなり低い順位となっている.

表 10　話者の性別による特徴語

女性	男性
うん（感動詞），ね（終助詞），そう（副詞），カオ（固有名詞），此れ，私（わたし），雲（くも），あっ（感動詞），センチ（単位），ふん（感動詞）や（助動詞），御（お），本当，此の．園（えん），右（みぎ），さん（接尾辞），ルール，も（係助詞），烏賊	俺（代名詞），ちび，野球，いや（感動詞），アベ（固有名詞, クラシック，出でる，ヨコハマ（地名），君（くん），打つ，ショウマ（固有名詞），御前（おまえ），ケビン（固有名詞），じゃん（終助詞），おお（感動詞），ピザ，次（じ），シーン（場面），サクヤ（固有名詞），僕（代名詞）

6 まとめ

本稿では, CEJC モニター公開版を使って, 日常会話における語彙を量的な面から観察・記述した. 本稿の考察はまだ報告的なものが多く，細かい分析にまでは至らなかった．会話の属性や話者の属性を掛け合わせた考察は今後の課題である.

謝辞

本研究は，国立国語研究所共同研究プロジェクト「大規模日常会話に基づく話し言葉の多角的研究」の成果の一部である。

参考文献

Kilgarriff, Adam(1996) Which words are particularly characteristic of a text? A survey of statistical approaches, Proc. AISB Workshop on Language and Engineering for Document Analysis and Recognition, Sussex, April 1996. 33-40.

小木曽智信（2014）第 5 章　形態素解析,『講座日本語コーパス 2 書き言葉コーパス』朝倉書店.

小磯花絵他（2019 予定）『日本語日常会話コーパス』モニター公開版の設計と特徴，言語処理学会代 25 回大会予稿集.

山崎誠（2017）レジスター・位相の違いによる会話文の語彙的多様性，言語資源活用ワークショップ 2017 発表論文集，pp. 278–289, doi. org/10. 15084/00001529