日本語日常対話コーパスの構築

赤間 怜奈 [1,2]　磯部 順子 [2]　鈴木 潤 [1,2]　乾 健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所
{akama, jun.suzuki, kentaro.inui}@tohoku.ac.jp, yoriko.isobe@riken.jp

概要

規範的な日本語表現で構成される高品質な日本語日常対話コーパスを構築した．本稿では，構築したコーパスの概要とその構築方法を説明する．既存対話コーパスとの比較分析により構築したコーパスの特長を明らかにした上で，「規範的」という性質を持つ言語資源の利点について述べる．

1　はじめに

昨今の対話システムは，出力内容に着目した総合的な主観評価の上で着実な性能が改善がなされている [1, 2, 3]．次におこなうべきは，その性能改善の要因の理解や依然として残る技術的な課題の把握であるが，対話で用いられる言語表現は自由度が非常に高いという性質を持つため（基本語彙以外の語の出現や，基本語順からの逸脱は常である），一定の基準や特定の正解を軸にした分析では，性能改善のための本質的な要因を特定することが困難である．

そこで我々は，新たに，基本語彙の使用と基本語順の遵守を可能な限り優先した規範的な言語表現で構成されている対話コーパス**日本語日常対話コーパス（Japanese Daily Dialogue）**を構築した．[1)] 本コーパスに収録する収録する対話の作成や調整をすべて人手でおこない，ノイズが非常に少なく計算機上での処理がしやすい，かつ，品質が十分に保証されている信頼できる対話データを獲得した．規範的な対話は，実際の対話データ（たとえば，音声対話の書き起こしや SNS）が持つ人間のリアルな言語活動の表出という特長を失うが，その代わり，ある種形式的で簡潔な問題設定を対話の枠組みで実現しているといえる．規範的な対話を対象とした対話の意味的・統語的言語理解に関する分析は，一般的な対話，とりわけリアルな対話を対象とする場合に比べて，比較的容易に実行できる可能性がある．

本稿では，日本語日常対話コーパスの構築手順を説明し，データの規模や特長を報告する．その上で，本コーパスが有する規範的な性質の利点を，具体的な事例とともに紹介する．

1）　2023 年 3 月公開予定．

表 1　対話データの例．トピック「旅行」．

A: おはようございます。高原の朝は冷えますね。
B: おはようございます。本当ですね。羽織るものが欲しいです。
A: 朝食の前に散歩でもいかがですか？
B: 良いですね。どこを歩きましょうか？
A: 湖の周りを歩きましょう。林道の先に湖があるそうですよ。
B: それでは、湖を一周しましょう。
A: 一周するのにどのくらい時間がかかるでしょうか？
B: ロッジのオーナーに聞いてみましょう。

2　日本語日常対話コーパス

本稿で我々が構築する日本語日常対話コーパスは，規範的な対話を収録した言語資源である．ここでいう規範的な対話とは，道徳的な内容かつ正しく丁寧な表現で書き表されている対話のことを指す．直感的には，初等から中等教育レベルの言語学習用教材で用いられる表現に近い．くだけた表現が頻発する実際の日常会話で使用するには不自然かもしれないが，基礎的な語彙かつ基本的な語順を可能な限り優先した言語表現である．

収録されている対話の例を表 1 に示す．すべての対話は話者 A，B が交互に発話をする形式であり，基本的には，対話単位でひとまとまりとなるように対話の始まりと終わりが設計されている．1 対話は 4 発話以上，1 発話は 1 文以上で構成されている．日本語日常対話コーパスの統計情報を表 2 に示す．[2)] 本コーパスには，日本語の日常会話に馴染み深いと考えられている 5 つのトピックに関する対話が収録されている．対話数は，1 トピックあたり 1,000 以

2）　トークン分割には，日本語形態素解析機 MeCab と日本語形態素解析辞書 mecab-ipadic-NEologd を用いた．

表 2　日本語日常対話コーパスの統計情報.

トピック	対話数	発話数	トークン数	語彙数	1 対話あたり 平均発話長	1 対話あたり 平均トークン長	1 発話あたり 平均トークン長
日常生活	1,070	8,462	131,879	7,585	7.91	123.25	15.58
学校	1,058	8,197	138,569	7,063	7.75	130.97	16.90
旅行	1,021	8,459	138,898	7,787	8.29	136.04	16.42
健康	1,061	8,344	137,691	6,990	7.86	129.77	16.50
娯楽	1,051	8,318	136,683	7,717	7.91	130.05	16.43
全体	5,261	41,780	683,720	19,384	7.94	129.96	16.36

表 3　表現の正規化による各値の変化.

	発話数	トークン数	文字数	語彙数
正規化前	41,811	642,523	1,088,080	19,964
正規化後	41,780	683,720	1,153,394	19,384
	(−0.1%)	(+6.4%)	(+6.0%)	(−2.9%)

上，コーパス全体で 5,261 である．発話単位で計数すると，41,780 発話である．1 対話あたりの平均発話長は，旅行トピックが他より若干大きい値だが，全体として 8 前後である．

3　構築手順

日本語日常対話コーパスの構築手順は以下の通りである．これらすべての工程は，人間の作業者によっておこなわれた．各工程の詳細は後述する．

1. コーパスに含める対話のトピックの選定
2. 各トピックに関連する対話の作成
3. 倫理的または道徳的に不適切な表現の除去
4. 表記の正規化

トピックの選定　基礎的な日本語対話を広く収録するためには，選定するトピックは，多くの日本語話者が高度な専門知識を必要とせず容易に対話を展開できるものであることが望ましい．本研究では，基本的には英語日常対話コーパス DailyDialog [4] が採用したトピックを参考にしつつ，日本語教育学分野の日本語会話における話題の難易度分類 [5, 6] を参照することで，日本語話者にとって馴染み深く日本の文化的特性に調和するトピックを優先的に採用するという方針をとり，最終的には，「日常生活」「学校」「旅行」「健康」「娯楽」の 5 つを採用した．

対話の作成　対話の作成は，日本語を母国語とする合計 59 人の作業者によっておこなわれた．各作業者は，与えられたトピックに関連する対話，すなわち仮想の話者 A, B による一連の発話系列を作成した．作業者には，可能な限り多くの一般的な語を含めること，文法的に正しく丁寧な日本語表現を用いることなどを指示した．作業の品質を担保するために，対話作成者とは異なる 1 人以上の作業者が，作成された対話が与えられた指示を満たしていること，および，対話として成立していることを確認した．作成された対話は，5,454 対話であった．

不適切な表現の除去　倫理的あるいは道徳的な問題を含む対話や了解性の低い対話は，コーパスの規範性および品質を損なう可能性がある．本研究では，国や報道機関が定める言語表現に関するガイドライン[3)]を参考に以下を「不適切な表現」と定義し，該当する対話はコーパスに収録しないようにした．

- 差別や偏見を含む
- 特定の属性を誹謗中傷あるいは攻撃する
- 第三者に不利益を与えうる事実誤認がある
- 倫理的あるいは道徳的でない話題を扱う
- 意味不明な発話内容や会話展開を含む

具体的には，作成されたすべての対話について，上述の項目に該当する可能性があるかを著者らが人手で確認し，該当する場合は，対話全体あるいは問題のある一部箇所を削除する，または，問題のある一部箇所を適切な表現に書き改める作業をおこなった．削除された対話の例を付録 A 表 7, 8 に示す．

表記の正規化　対話が規範的であるためには，文法的および意味的に正しく一貫性のある表現で表記されている必要がある．不適切な表現を除去し終えたすべての対話について，合計 3 人の校正者が校正作業および校閲作業をおこなった．表記は，原則として記者ハンドブック [7] に従った．誤字脱字の修正や用字の統一などの一般的な校正作業に加え，くだけた話し言葉表現（略語や俗語，撥音便化など）を丁寧な書き言葉表現に変換すること，文法的に適切な文となるよう省略された語（助詞など）を補完

3)　「公用文作成の考え方」について（https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/93650001_01.html，文化庁），NHK 放送ガイドライン 2020 改訂版（https://www.nhk.or.jp/info/pr/bc-guideline/，日本放送協会）など．

表 4　各日本語対話コーパスの統計情報と特長.

コーパス	発話数	トークン数 N	語彙数 V	1 発話あたり平均トークン長	多様性 C	親密性 S_F	可読性 S_R
JDailyDialogue	41,780	683,720	19,384	16.36	**0.7348**	**5.74 ± 0.88**	**4.66 ± 0.08**
Business Scene Dialogue	24,171	298,124	11,991	12.33	**0.7451**	**5.78 ± 0.90**	4.21 ± 0.19
JEmpatheticDialogues	80,000	1,211,366	24,506	15.14	0.7215	5.72 ± 0.84	4.56 ± 0.16
JPersonaChat	61,793	1,471,949	20,033	23.82	0.6974	5.73 ± 0.86	**4.84 ± 0.12**
Opensubtitles	3,170,155	19,997,429	150,606	6.31	0.7092	4.85 ± 1.32	3.78 ± 0.55
Twitter	3,157,896	39,832,298	364,955	12.61	0.7319	4.58 ± 1.40	2.88 ± 0.39

すること，対話内容に矛盾があれば指摘あるいは矛盾を解消するよう修正することなど，本研究特有の作業も依頼した．作業例を付録 A の表 9 に示す．表 3 に，校正者が正規化をする前と後の対話データの統計量を示す．正規化後は，用字用語の統一の効果か語彙数が約 3%減少した一方，省略が補完されたためかトークン数と文字数は約 6%増加した．

4　分析：コーパスの特長

既存コーパスとの比較により日本語日常対話コーパスの特長を明らかにする．以下に設定の概要を示す．詳細は著者らの過去論文 [8] を参照されたい．

4.1　設定

比較対象　比較対象として次の 5 つのコーパスを採用した: Business Scene Dialogue [9]， JPersonaChat, JEmpatheticDialogues [10]，Opensubtitles [11]，Twitter（長澤ら [12] の前処理を適用）．これらはすべて，日本語日常対話コーパス同様，人によって書かれた雑談対話を収録したコーパスである．

分析の観点と尺度　語彙的多様性，語彙親密性，可読性の 3 つの観点で，コーパスの特長を分析した．語彙的多様性を測定する尺度として，Herdan の C [13, 14] 用いる．[4] C の値は，総トークン数 N と語彙数 V を用いて以下で算出される:

$$C = \log V \,/\, \log N \,. \tag{1}$$

語彙親密性は，単語親密度データベース [16, 17] を用いて，親密性スコア S_F を以下の式で算出した:

$$S_F = \frac{1}{|\mathcal{D}|} \sum_{v \in \mathcal{D}} fam(v) \,. \tag{2}$$

ここで，$fam(\cdot)$ は，データベースに存在する語 $v \in \mathcal{D}$ についてその単語親密度を返す関数である．[5]

4）　一般的によく利用される TTR (Type-Token Ratio) [15] はデータサイズ N の影響を受けやすい．それを軽減するように標準化された尺度が C である．

5）　値の範囲は $1 \le S_F \le 7$ であり，大きいほど親密性が高い．

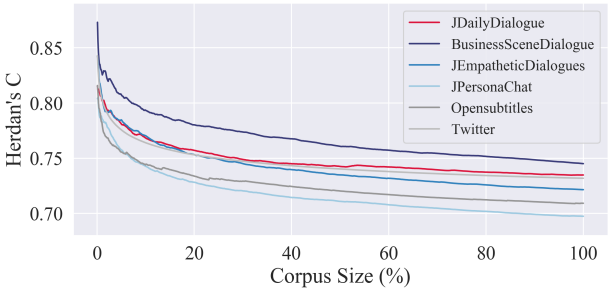

図 1　コーパスサイズと語彙の Herdan's C の関係

可読性の算出には，日本語文章難易度判別システム jReadability[6] を用いた [18]．コーパスから無作為に抽出した 100 発話リーダビリティスコアの測定を 5 回繰り返し，平均値を可読性スコア S_R とした．

4.2　分析結果と考察

表 4 に，各コーパスの基本的な統計情報と，各観点についてそれぞれの尺度で算出した値を示す．

多様性　日本語日常対話コーパスの語彙的多様性を表す C の値は，Business Scene Dialogue に次いで，2 番目に大きい値であった．図 1 の曲線は，各コーパスの x % ($0 \le x \le 100$) に該当する数の発話で算出された C の値の変化を表す．日本語日常対話コーパスは，x の増加に伴う C の減少が小さいことから，多様な語彙が豊富に含まれていることがわかる．

親密性　日本語日常対話コーパスの親密性スコア S_F は，Business Scene Dialogue に次いで 2 番目に大きい値であった．[7] なお，JEmpatheticDialogues や JPersonaChat も，ほぼ同程度の値であった．スコア S_F の最大値が 7.0 であることから，これらのコーパスが示した 5.7〜8 という値は十分に高いものといえる．したがって，日本語日常対話コーパスを含むこれらのコーパスには，日本語話者にとって馴染み深い一般的な表現が多数含まれていることがわかる．

可読性　日本語日常対話コーパスの可読性スコア S_R は，JPersonaChat に次いで 2 番目に大きい値で

6）　https://jreadability.net/.

7）　付録 B に，語彙親密度の頻度分布を示す．

表 5 各コーパスの語彙のうち，既存資源の語彙でカバーされる語の割合（%）．なお，言語モデル (*) については，いずれも語彙の分割単位が異なること（コーパスは単語単位，言語モデルはサブワード単位）に注意が必要である．

コーパス	NTT 語彙 DB	教育基本語彙	fastText	BERT*	GPT-NeoX*
JDailyDialogue	64.05	42.00	82.04	39.27	30.56
Business Scene Dialogue	65.68	46.30	87.95	49.86	38.36
JEmpatheticDialogues	62.56	38.11	84.15	33.54	27.96
JPersonaChat	60.49	37.47	84.70	35.17	29.85
Opensubtitles	29.79	10.25	66.88	12.05	10.04
Twitter	16.67	4.44	50.16	5.56	4.95

表 6 各コーパスに含まれる差別語・不快語と出現頻度．

コーパス	種類数	出現頻度
JDailyDialogue	4 / 170	9 (0.001%)
Business Scene Dialogue	2 / 170	3 (0.001%)
JEmpatheticDialogues	19 / 170	72 (0.006%)
JPersonaChat	15 / 170	112 (0.008%)
Opensubtitles	77 / 170	3,962 (0.020%)
Twitter	80 / 170	17,814 (0.045%)

あった．これらのコーパスには，多くの人にとって読解しやすい比較的平易な表現で記述された対話が，数多く含まれていることがわかる．

総評 以上より，日本語日常対話コーパスは，既存コーパスに匹敵するまたは上回る語彙的多様性，親密性，可読性をバランスよく兼ね備えていることがわかる．一般的で馴染み深い多様な語彙を豊富に含み，正しい日本語表現で記述された平易で読みやすい対話が収録されていることがわかる．

5 議論

5.1 規範的であることの利点

既存資源の活用 表 5 は，各コーパスの語彙のうち，既存資源（辞書 [17, 19]，単語ベクトル [20]，言語モデル[8)9)]）の語彙に登録されている語，つまり，既存資源で分析可能な語彙の割合を示す．[10)] 日本語日常対話コーパスは，いずれの資源においてもこの値が比較的大きい．一貫して正しく表記された語彙を多く含む日本語日常対話コーパスは，その規範的な性質ゆえに未知語が生じにくく，比較的，既存資源を活用しやすいデータであるといえる．

低攻撃性 表 6 は，記者ハンドブック [7] に「差別語・不快語」として掲載されている 170 語のが，各コーパスにどの程度含まれているかとその出現頻度を示す．[11)] 日本語日常対話コーパスは，語彙に含まれている差別語・不快語の種類数，それらのコーパス全体における出現頻度ともに小さい値であった．構築時に非倫理的あるいは非道徳的な表現を積極的に排除していることもあり，日本語日常対話コーパスには攻撃的な表現がほとんど出現しない．

5.2 他言語対話コーパスとの関係

国内外の雑談対話研究において，DailyDialog [4]，Persona-Chat[21]，EmpatheticDialogues[22] は近年のデファクトスタンダードといえるデータセットである．我々が構築した日本語日常対話コーパスは，教材的（規範的）な表現で書かれたマルチターンの高品質な日常対話という特長を持つ DailyDialog の日本語版と見做すこともできる．杉山ら [10] が公開している JPersona-Chat，JEmpatheticDialogues と合わせると，日本語日常対話コーパスの公開によってデファクトスタンダードの日本語版対話データが揃ったことになり，日本語を含む言語横断的な対話研研究のため土台が整備されたといえる．

6 おわりに

規範的な日本語表現で構成される高品質な日本語日常対話コーパスを構築し，その概要と構築手順を紹介した．既存対話コーパスとの比較分析により，構築したコーパスは，既存のものに匹敵するまたは上回る語彙的多様性，親密性，可読性をバランスよく兼ね備えていることを確認した．将来的には，基礎解析系アノテーションの追加や対訳データ化など補完的な新しいデータセットの構築も視野に入れたいが，まずは，高品質な日本語対話データとして公開された日本語日常対話コーパスを多くの方に利用していただけたらうれしい．

8） https://huggingface.co/abeja/gpt-neox-japanese-2.7b
9） https://huggingface.co/cl-tohoku/bert-base-japanese-v2
10） 登録されていない語は，未知語として扱われる．
11） 付録 C に，コーパスに含まれる差別語・不快語を示す．

謝辞

本研究は JSPS 科研費 JP22K17943，JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けたものです.

参考文献

[1] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu Quoc, and V Le. Towards a Human-like Open-Domain Chatbot. In **aiXiv preprint arXiv:2001.09977**, 2020.

[2] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations**, pp. 270–278, 7 2020.

[3] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y. Lan Boureau, and Jason Weston. Recipes for Building an Open-Domain Chatbot. In **Proceedings of 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, pp. 300–325, 2021.

[4] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, Shuzi Niu, and Hong Kong. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In **Proceedings of the The 8th International Joint Conference on Natural Language Processing (IJCNLP)**, pp. 986–995, 2017.

[5] 山内博之, 橋本直幸. 教育語彙表への応用. 有里子砂川（編）, コーパスと日本語教育, 第 2 章, pp. 35–64. 朝倉書店, 2016.

[6] 山内博之, 橋本直幸, 金庭久美子, 田尻由美子. 言語活動・言語素材と話題. 博之山内（編）, 実践日本語教育スタンダード, 第 1 章, pp. 5–525. ひつじ書房, 2013.

[7] 一般社団法人共同通信社. 記者ハンドブック 第 14 版 新聞用字用語集. 2022.

[8] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 規範的な日本語日常対話コーパスの設計. 言語処理学会 第 28 回年次大会 発表論文集, pp. 262–267, 2022.

[9] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the Business Conversation Corpus. In **Proceedings of the 6th Workshop on Asian Translation (WAT)**, pp. 54–61, 2019.

[10] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems. In **aiXiv preprint arXiv:2109.05217**, 2021.

[11] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In **Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)**, pp. 1742–1748, 2018.

[12] 長澤春希, 工藤慧音, 宮脇峻平, 有山知希, 成田風香, 岸波洋介, 佐藤志貴, 乾健太郎. aoba_v2 bot : 多様な応答生成モジュールを統合した雑談対話システム. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, pp. 101–106, 2021.

[13] Gustav Herdan. Type-token mathematics: A textbook of mathmatical linguistics. **Mouton**, Vol. 4, p. 448, 1960.

[14] Gustav Herdan. **Quantitative linguistics**. Butterworth, 1964.

[15] Mildred C Templin. **Certain language skills in children; their development and interrelationships**. University of Minnesota Press, 1957.

[16] 藤田早苗, 小林哲生. 単語親密度の再調査と過去のデータとの比較. 言語処理学会第 26 回年次大会発表論文集, pp. 1037–1040, 2020.

[17] 単語親密度 (令和版). NTT 語彙データベース. NTT 印刷, 2021.

[18] Yoichiro Hasebe and Jae-Ho Lee. Introducing a Readability Evaluation System for Japanese Language Education. **Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J)**, pp. 19–22, 2015.

[19] Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries. **Acta Linguistica Asiatica**, Vol. 2, No. 2, pp. 97–115, 10 2012.

[20] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning Word Vectors for 157 Languages. In **Proceedings of the 11th Edition of its Language Resources and Evaluation Conference (LREC)**, pp. 3483–3487, 2018.

[21] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)**, Vol. 1, pp. 2204–2213, 2018.

[22] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y. Lan Boureau. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 5370–5381, 2019.

A　中間データ

日本語日常対話コーパスを構築する過程で生成されたデータの例を示す．表 7, 8 は，工程 3「倫理的または道徳的に不適切な表現の除去」で除去した対話，表 9 は，工程 4「表記の正規化」で校正者がおこなった修正の例である．

表 7　倫理的あるいは道徳的に不適切な表現を含むとして削除された対話の例．トピックは「学校」．

A: 野木先生、昔と今では子供たちも変わってきていますか。
B: そうです。今の子はみんな真面目になっています。
A: 昔の子は不真面目だということですか？
B: 昔は先生の言うことを聞かない子が多かったので、授業は大変でした。
A: 反抗的な子は確かに今はいません。
B: しかし、静かなだけで何も聞いていない子もいるようです。
A: 昔は話を聞いていない子はすぐわかりましたが、今はそのようなわけにはいきません。
B: 今の方が大変なのかもしれません。

表 8　倫理的あるいは道徳的に不適切な表現を含むとして，一部が削除された対話の例．打ち消し線の部分が除外された箇所．トピックは「娯楽」．

A: 今日はバレンタインデーです。チョコは何個もらえるでしょうか。
B: 100 個はもらえると思います。
A: 本当ですか？
B: 冗談です。
A: 本当は何個くらいもらえる予定ですか？
B: 2 個くらいだと思います。
A: 2 個でももらえれば良いですね。
~~B: 可愛い子からだと嬉しいです。~~

表 9　表記の正規化の例．校正者が削除した表現を打ち消し線，追加した表現を太字で示す．トピックは「学校」．

A: 今週最後の授業が、~~やっと~~ **ようやく** 終わりました。明日は、待ちに待った土曜日です。
B: 今週も大変でしたね。中丸さんは何が一番大変でしたか?
A: 私は ~~一昨日~~ **おとい** の体育の ~~マラソン~~ **授業で行った長距離走** がとても大変でした。安藤さんは、~~どうですか~~ **いかがでしたか**？
B: 私は、今週はそこまで大変ではありませんでした。代わりに来週の小テストが怖いです。 ところで、この後、何か食べに行きませんか?
A: 行きましょう。最近 ~~美味しい~~ **おいしい** スイーツのお店を知ったので、そのお店に行ってみましょう。
B: 良いですね。とても楽しみです。

【校正者コメント】マラソン = 42.195 キロメートル。体育の授業でマラソンをすることは実際にはないかと思い、「ジョギング」に変更しました。「長距離走」などに変更してもいいかもしれません。

B　語彙親密度の頻度分布

各コーパスについて，語彙親密度の頻度分布を図 2 に示す．値が大きいほど親密度が高い．

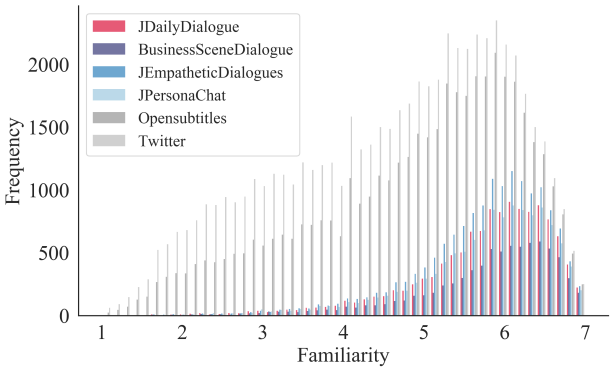

図 2　対話データに含まれる語彙親密度の頻度分布．

C　差別語・不快語の例

本研究で用いた差別語・不快語リストの中で，各コーパスに含まれていた語を以下に示す．Opensubtitles と Twitter は語数が多いため割愛する．

- **JDailyDialogue:**
  { 坊主, めくら, 不治の病, 床屋 }
- **Business Scene Dialogue:**
  { やばい, あんま }
- **JEmpatheticDialogues:**
  { ジプシー, 肌色, 外人, おし, アル中, 盲目的, 色盲, やばい, 名門, スケバン, あんま, 片親, 連れ子, 裸族, 坊主, デカ, 精神病院, 床屋, 浮浪者 }
- **JPersonaChat:**
  { 町医者, あんま, サツ, ジプシー, 肌色, 片親, スキンヘッド, 外人, おし, デカ, 坊主, やばい, 床屋, 片田舎, 浮浪者 }