規範的な日本語日常対話コーパスの設計

赤間 怜奈 [1,2]　磯部 順子 [2,1]　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,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日本語日常対話コーパスの開発に取り組んでいる．本稿では，コーパスの概要とその構築方法を紹介する．実際に本コーパスの一部として収集した小規模対話データの特性を，複数の観点から既存の対話コーパスと比較分析し，その結果を報告する．

1　はじめに

対話システムの性能についての議論は，一般論として，人間の主観的な「良さ」に基づいておこなわれる．近年は，複数システムの出力応答について「どちらが良いか」を人間が比較評価し，その勝敗によって性能を議論する枠組みがひとつの主流となっている．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は着実に性能が改善されているが [1, 2, 3]，その一方で，その性能改善の要因や依然として残る技術的な課題についてはほとんどわかっていない状況といえる．この要因のひとつとして，そもそも対話は言語の表現の自由度が非常に高いものであるために（基本語彙以外の語句の出現や，基本語順からの逸脱も頻繁に起こる），一定の基準や特定の正解を軸にしたエラー分析が困難となっていることが考えられる．

そこで我々は，基本語彙や正しい語順の使用を可能な限り優先した規範的な言語表現で構成される新たな対話コーパスとして，**日本語日常対話（Japanese Daily-Dialogue; JDD）コーパス**の開発に取り組んでいる．[1)] ここで扱う規範的な対話は，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が失われているが，その分，ある種形式的で簡潔な問題設定となっているため，対話を対象とした意味的・統語的言語理解の分析はしやすいものになっていると考えられる．また，本コーパスに収録する対話の作成や書式の正規化は，全て人手でおこなわれているため，ノイズが少なく計算機上でも処理がしやすい．表 1 にコーパス中の対話例を示す．

本稿では，日本語日常対話コーパスの概要を紹介し，その構築手順を説明する．その上で，実際にコーパスの一部として収集した小規模対話データについて，これが望ましい規範的な性質を有していることを複数の観点から確認する．

表 1　作成した対話データの例（トピック「旅行」）

A: 卒業旅行は、イタリアに行こうと思います。
B: それは、楽しそうです。イタリアのどちらに行かれるご予定ですか？
A: ローマとヴェネツィア、フィレンツェです．もう予約してきました。
B: それは、良いですね。私も以前フィレンツェに行ったことがありますが、食事がとても美味しかったことが印象に残っています。
A: はい。フィレンツェは、それが楽しみで行くことにしました。
B: ローマとヴェネツィアも、史跡と美術館巡りがとても楽しいと聞いたことがあります。
A: はい。実は私、大学でイタリア美術を研究していたので、その勉強を兼ねて行くことにしたのです。
B: そうだったのですね。そうしましたら、ますます楽しみですね。
A: はい、ありがとうございます。

2　日本語日常対話コーパス

2.1　概要

日本語日常対話コーパスは，書き言葉を対象に，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，実際の日常生活で使用するには少しの不自然さがあるものの，道徳的な内容かつ

1）　完成した日本語日常対話コーパスは，主に研究用途として公開することを予定している．

表 2　作成された対話の統計情報

トピック	対話数	発話数	トークン数	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	204	1,227	18,497	6.01	90.67	15.07
学校	202	1,251	18,033	6.19	89.27	14.41
旅行	200	1,425	21,554	7.13	107.77	15.13
健康	200	1,235	19,720	6.18	98.60	15.97
娯楽	200	1,294	18,638	6.47	93.19	14.40
全体	1,006	6,432	96,442	6.39	95.87	14.99

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ような対話表現に近い．すべての対話はふたりの話者 A，B が交互に発話をおこなう形式で，基本的には対話の始まりと終わりが設計されている．ひとつの対話は 4 以上の発話，ひとつの発話は 1 以上の文で構成されている．各対話には，トピック情報が付加されている．将来的には，対話単位での難易度，発話単位での発話行為タグや感情タグ等の付加情報も追加する予定である．

2.2　構築手順

日本語日常対話コーパスの構築は，すべての工程を人手でおこなう．構築手順は以下の通りである：

1. コーパスに含める対話のトピックを選定
2. 各トピックに該当する対話を作成
3. 書式（表記揺れ，常用漢字等）を正規化
4. 発話行為タグ等の付加情報をアノテーション

なお，本研究ではひとまず手順 2 までを小規模に実施し，規範的な特性を持った対話データが取得可能であることを確かめる．本試行で取得する対話データを，日本語日常対話コーパス v0（JDD v0）と呼ぶ．JDD v0 の構築方法の詳細を次節で説明する．

2.3　各手順の詳細

手順 1: トピックの選定　本コーパスで取り扱うトピックは，既存の他言語資源と日本語会話に関する既存研究を参考に，基礎的な日本語対話を広く含むように慎重に選定した．基本的には，近年の対話研究で人気のある英語日常対話コーパス DailyDialog [4] で採用されているトピックを参考にする．[2] 一方で，我々の目的は基礎的な日本語対話が収録されたコーパスを構築することにあるため，選定するトピックは，日本の文化的特性によく調和するもの，かつ，高度な専門知識等を必要とせず多くの人にとって容易対話を展開できるものであることが望ましい．そこで，日本語教育学の知見を参考に日本語の会話によく馴染むトピックを優先的に採用することを考える．本研究では，山内らによる話題の難易度分類を参考にした [5]．これは，橋本ら [6] が作成した日本語会話の 100 の話題について，各話題に関連する語に対しての人間の主観的な身近さや会話内での需要に基づき，それらの難易度を 4 つのレベルに分類したものである．たとえば，「財政（Finance）」は英語では日常会話における一般的な話題のひとつであるが，山内らの分類によると，日本語では難易度が高い話題（難易度レベルが上から 2 番目）とされている．これらを参考に，最終的に，日常生活，学校，旅行，健康，娯楽の 5 つを日本語日常対話コーパスのトピックとして採用した．

手順 2: 対話の作成　対話の作成は，日本語を母国語とする 51 名の作業者によっておこなわれた．各トピックにつき 8〜11 名が割り当てられ，作業者 1 名がひとつの対話（つまり，話者 A，B による一連の発話系列）を作成した．作業者には，できるだけたくさんの一般的な語彙と多様なモダリティを含んだ 4 発話以上からなる対話を，正しく丁寧な日本語表現で作成するよう指示した．作業者への教示と，提示した参考資料の詳細を，付録 A に示す．さらに，対話データ全体の品質を担保するために，作業者が作成したすべての対話について作業者とは別の 5 名による品質チェックがおこなわれた．

2.4　JDD v0 の基本統計

上記の手順により，本研究で収集した対話に関する統計情報を，表 2 に示す．[3] 統計値の算出に際し，トークン分割には日本語形態素解析機 MeCab [4]

2）　Ordinary Life, School Life, Culture & Education, Attitude & Emotion, Relationship, Tourism, Health, Work, Politics, Finance の 10 種のトピックが採用されている．

3）　付録 B に，発話長とトークン長の分布と対話例を示す．

4）　https://taku910.github.io/mecab/

表 3　各日本語対話コーパスの統計情報と語彙的特長

コーパス	発話数	トークン数 (N)	1 発話あたり 平均トークン長	語彙数 (V)	TTR (%)	Herdan's C
JDD v0	6,432	96,442	14.99	6,530	6.77	**0.7654**
Business Scene Dialogue	24,171	298,124	12.33	11,991	4.02	0.7451
JEmpatheticDialogues	80,000	1,211,366	15.14	24,506	2.02	0.7215
JPersonaChat	61,793	1,471,949	23.82	20,033	1.36	0.6974
Opensubtitles	3,170,155	19,997,429	6.31	150,606	0.75	0.7092
Twitter	3,157,896	39,832,298	12.61	364,955	0.92	0.7319

と日本語形態素解析辞書 mecab-ipadic-NEologd [5] を用いた．すべてのトピックでそれぞれ 200 以上，全体で 1,006 の対話が得られた．発話単位で計数すると，6,432 発話となる．トピック毎の対話の傾向を観察すると，1 対話あたりの発話長については，旅行トピックが他の 4 つに比べて若干大きい値であった．また，1 発話あたりのトークン長については，トピック間でのばらつきはあまり認められなかった．なお，完成版の日本語日常対話コーパスは，この 2〜10 倍のデータサイズとする予定である．

3　分析：既存対話データとの比較

3.1　設定

比較対象　比較対象として，日本語日常対話コーパスと同様，人手で書かれた対話データからなる次の 5 つのコーパスを用いる[6]:

- **Business Scene Dialogue コーパス** [7]: ビジネスシーンにおける会議、交渉、雑談などの対話が収録された日英対訳コーパス．
- **JPersonaChat** [8]: 発話者のペルソナを反映した雑談対話コーパス PersonaChat [9] の日本語版．
- **JEmpatheticDialogues** [8]: 感情的な状況下での発話とそれに対する共感からなる雑談対話コーパス EmpatheticDialogues [10] の日本語版．
- **Opensubtitles** [11]: 日本語映画字幕コーパス．隣接する字幕を発話系列と見做すことによって対話コーパスとして利用できる [12]．
- **Twitter**: 日本語のツイートを収集し，リプライチェーンを発話系列と見做すことによって対話コーパスとして利用できる．本研究では，長澤らの前処理済み Twitter データのうち 2020 年分のデータのみを用いる [13]．

5）https://github.com/neologd/mecab-ipadic-neologd/

6）音声会話の書き起こしやシステムによる生成を含む対話データも存在するが，これらは JDD（人手による書き言葉）とは性質が大きく異なるため比較対象に含めない．

分析の観点と方法　JDD v0 のデータ特性を，(1) 表層上の語彙的特長，(2) 語彙親密度，(3) リーダビリティの 3 つの観点から，既存対話データとの比較を通じて分析する．

まず，語彙的特長については，各コーパスの基本的な統計情報に加えて，表層上の語彙の多様性を TTR (Type-Token Ratio) [14] と，Herdan の C [15, 16] 用いて算出する．これらの指標は，総トークン数 N と語彙数 V を用いてそれぞれ次のように表される:

$$TTR = \frac{V}{N}, \qquad C = \frac{\log V}{\log N}\,. \tag{1}$$

TTR はデータサイズの影響を受けやすく，N の増加に伴い値が小さくなる．これを解決し，サイズの異なるデータ間でも TTR を比較できるように標準化された尺度が C である．

次に，語彙親密度については，単語親密度（令和版）データベース [17, 18] を用いて，各コーパスの語彙親密度スコア S_F を以下の式によって算出する:

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

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．

最後に，リーダビリティについては，日本語文章難易度判別システム jReadability [7] によって算出されるリーダビリティスコアを用いる [19]．このとき，S_R は，平均文長 a，漢語率 b，和語率 c，動詞率 d，助詞率 e を用いて，以下の式によって算出される:

$$\begin{aligned} S_R = 11.724 &- 0.056a - 0.126b \\ &- 0.042c - 0.145d - 0.044e\,. \end{aligned} \tag{3}$$

3.2　分析 1: 語彙的特性

表 3 に，各コーパスに収録されている発話数やトークン数等の基本統計と，算出した TTR，C の値

7）https://jreadability.net/.
システムの入力文字数上限の都合で，各コーパスから無作為に抽出した 500 発話を対象としてスコアを算出した．

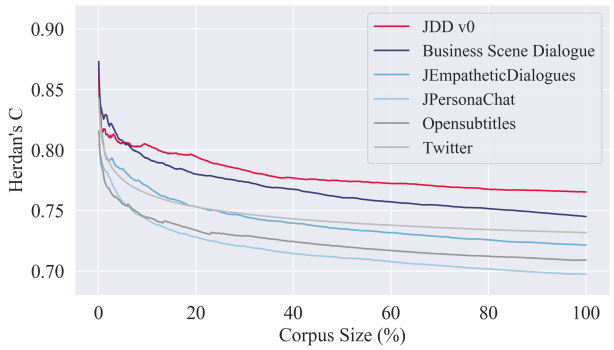

図 1　コーパスサイズと語彙の多様性（Herdan's C）の関係

を示す．まず，1 発話あたりのトークン長について，JDD v0 は 14.99 であり，これは他のいくつかのコーパスと同程度の値であった．分析対象のうち比較的話し言葉に近い対話データである Opensubtitle は，他よりも極端に小さい値であった．次に，各コーパスの表層上の語彙の多様性を表す C の値は，JDD v0 が他よりも大きかった．図 1 の曲線は，各コーパスの x % ($0 \leq x \leq 100$) に該当する発話文で算出された C の値の変化を表す．JDD v0 は x の増加に伴う C の減少が小さく，このことからも JDD v0 には，多様な語彙が豊富に含まれていることがわかる．

3.3　分析 2: 語彙親密度

表 4 に，各コーパスについて算出した語彙親密度スコア S_F を示す．[8] すべてのコーパスのなかで JDD v0 が最大のスコアを示した．これは，JDD v0 の対話が，他よりも一般的で日本語として馴染み深い表現で構成された基礎的な対話であることを示唆している．また，表中の「分析可能語彙」は，S_F の算出に関与した語の割合，つまり，コーパスの語彙のうちデータベースに登録されている概念として認識された語の割合を示す：$|\mathcal{D}|/V$．この値が大きいほど，計算機での解析も容易な，正しい日本語表現で記述されたデータである可能性が高い．JDD v0 の分析可能語彙は 70.9%で，これは全体における最大値であった．一方で，インターネットスラングや口語的な表現を含む Twitter（16.7%）と Opensubtitle（29.8%）の分析可能語彙の割合は他と比べて極端に小さい値であった．

3.4　分析 3: リーダビリティ

表 5 に，各コーパスについて算出したリーダビリティスコアを示す．JDD v0 は，すべての比較対象を

表 4　語彙親密度に関する分析

コーパス	分析可能語彙 (%)	親密度 S_F
JDD v0	**70.9**	**5.991**
Business Scene Dialogue	65.7	5.777
JEmpatheticDialogues	62.6	5.721
JPersonaChat	60.5	5.730
Opensubtitles	29.8	4.849
Twitter	16.7	4.579

表 5　リーダビリティに関する分析

コーパス	スコア
JDD v0	**4.89**
Business Scene Dialogue	4.20
JEmpatheticDialogues	4.56
JPersonaChat	4.85
Opensubtitles	3.82
Twitter	3.03

上回るスコアを示した．この結果から，JDD v0 の対話データは，僅差で続く JPersonaChat とともに，人間にとって読解しやすい比較的平易な表現で記述されているといえる．

3.5　総評

分析結果を総合すると，既存対話コーパスと比較して，JDD v0 に含まれる対話は，一般的で馴染み深い多様な語彙を豊富に含み（3.2, 3.3 節），正しい日本語表現で記述された（3.3 節）平易で読みやすい（3.4 節）といえる．これらの特性は，我々が対象とする「規範的かつ基礎的な対話」の要件としていずれも望ましいものである．今回取得した JDD v0 について，今後，2.2 節の構築手順に示すように書式の正規化等の処理をおこなうことによって，前述の望ましい特性がさらに顕著に出現するような対話データを獲得できる可能性がある．

4　おわりに

現在我々が開発している規範的な日本語日常対話コーパスについて，その概要と構築方法を説明した．実際のコーパスの一部として収集した約 1,000 対話について，その性質を語彙的特長・語彙親密度・リーダビリティの 3 つの観点から分析し，望ましい規範的な特性を持つ対話を獲得することを確認した．今後は，データ規模の拡大と合わせて，対話の高品質化や付加情報の追加についても取り組む．

8）　付録 C に，語彙親密度の頻度分布を示す．

謝辞

本研究は 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] 博之山内, 直幸橋本. 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] 博之山内, 直幸橋本, 久美子金庭, 由美子田尻. 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] 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.

[8] 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.

[9] 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.

[10] 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.

[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] Oriol Vinyals and Quoc Le. A Neural Conversational Model. In **Proceedings of the 31st International Conference on Machine Learning (ICML) Deep Learning Workshop**, 2015.

[13] 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.

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

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

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

[17] 早苗藤田, 哲生小林. 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.

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

[19] 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.

A　対話作成作業に関する参考情報

A.1　作業者への教示

作業者には，日本語学習の教材となるような規範的な日常会話を作成するよう指示した．合わせて，作成される対話は以下の要件を満たすように伝えた: (1) 正しい日本語で書かれていること，(2) 道徳的な内容であること，(3)1 つの対話は最低でも 4 発話 以上で構成すること，(4) 日常的によく使われる語とりわけ話題語を多く出現させること，(5) 自然な会話の流れの中で多様なモダリティを出現させること，(6) 相槌やフィラー等の挿入は必要最低限とすること，(7) 固有名詞に相当する表現は、現実・架空に関わらず具体的に書くこと．

なお，教示が適切であることの確認ならびに作業者の予行練習と疑問点解消の目的で，実際の対話作成作業を開始する前に，計 3 回のプレ作業をおこなっている．

A.2　提示した資料

前節に示す作業者に教示した要件のうち，要件 (4) と (5) については，それぞれの具体的例を記したものを作成し（話題語集，モダリティ集），参考資料として全ての作業者に配布した．話題語集の作成には，分類語彙表増補改訂版データベース (ver.1.0) [9] を参考にした．モダリティ集の作成には，つつじ日本語機能表現辞書 [10] を参考にした．

B　JDD v0 の統計情報: 発話長とトークン長の分布

各トピックにおける対話の発話長（横軸）とトークン長（縦軸）の分布を図 2 に示す．

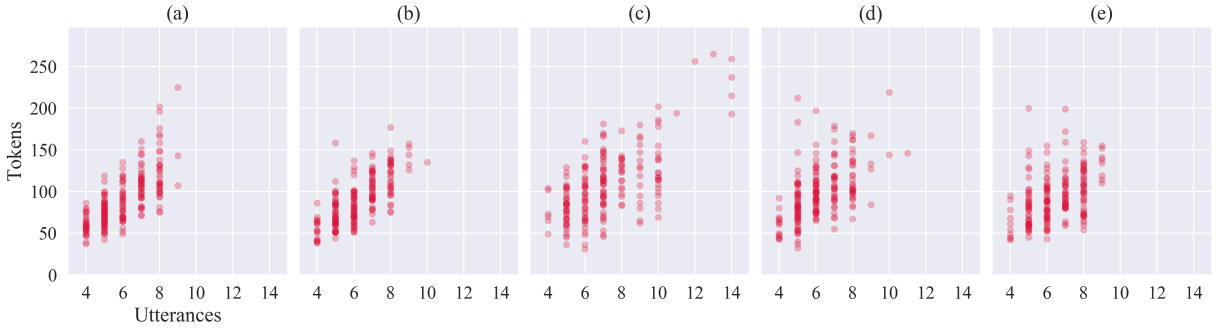

図 2　対話の発話長とトークン長の分布．左から (a) 日常生活，(b) 学校，(c) 旅行，(d) 健康，(e) 娯楽 トピック．

C　各対話データにおける語彙親密度の頻度分布

作成した JDD v0 と，比較対象として用いた 5 種の対話コーパスに含まれる語彙親密度の頻度分布を，図 3 に示す．

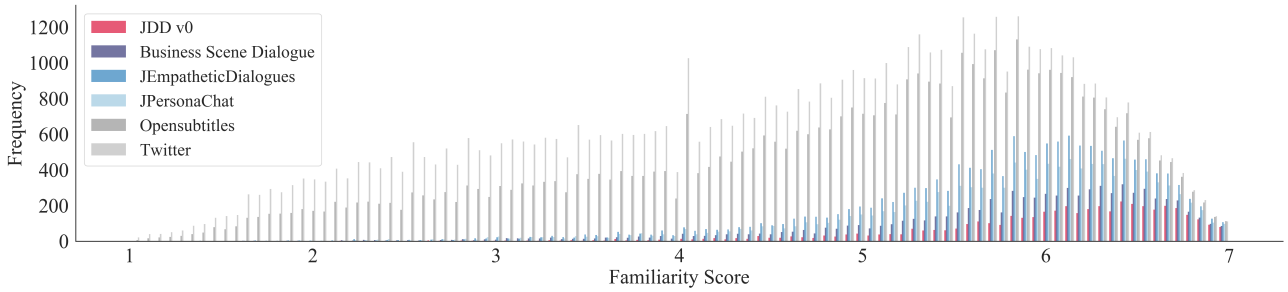

図 3　対話データに含まれる語彙親密度の頻度分布．スコアは，値が大きいほど親密度が高いことを表す．

9）https://github.com/masayu-a/WLSP
10）http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/