単語親密度の再調査と過去のデータとの比較

藤田 早苗　　小林 哲生
NTT コミュニケーション科学基礎研究所

{sanae.fujita.zc, tessei.kobayashi.ga}@hco.ntt.co.jp

1 はじめに

単語親密度とは，語のなじみ深さを数値化したものである．天野らが NTT データベースシリーズ「日本語の語彙特性」[1] で発表した単語親密度のデータは，語彙力の調査 [2] や単語認知に関する実験心理学的研究 [3]，また脳機能障害 [4] や言語発達障害の医療分野の研究 [5] など，さまざまな用途で使用されている．

天野らは，18 歳から 29 歳までの 40 人を評価者として 7 段階の評定を行い，約 7 万語，第 9 巻を合わせると 10 万語以上の単語親密度を調査している．天野らの調査では，評価者は全員実験室に集められ，丁寧なインストラクションと事前の訓練がなされた．さらに事前の訓練や事後評価で信頼度の低い評価者のデータは取り除かれており，評価の質の担保がなされている．また，天野らは文字だけを評価者に提示した場合の文字単語親密度と，音声も同時に聞かせた場合の文字音声単語親密度，音声だけを聞かせた場合の音声単語親密度の調査も行っている．大変有益なデータだが，天野らの親密度の評定実験の時期は 1995 年 9 月～1997 年 7 月と 20 年以上前である[1]．そのため，親密度の数値自体が現代とは変化している可能性があり，また，スマートフォンやコンビニ，インターネットのような語も含まれていないという問題がある．

浅原 [6] は，分類語彙表 [7] の見出し語に対してクラウドソーシングを用いて，「知っている」「書く」「読む」「話す」「聞く」の 5 項目について，5 段階の評定を行っている．各語につき少なくとも 16 人以上で評価を行っているが，評価者の統制やスクリーニングは行っていない．

本研究では，評価者の統制や事前・事後のスクリーニングも実施し，クラウドソーシングによる信頼度の高い文字単語親密度 (以下，親密度) の調査方法を提案，大規模調査を実施し，過去の調査結果と比較する．

本研究の貢献は以下の通りである．

(1) クラウドソーシングによる信頼度の高い親密度評定方法を提案 (2 章)
(2) 新語を含む大規模な調査を実施 (163,017 語，3 章)
(3) 親密度の経年変化を調査 (3 章)

[1]第 9 巻 (第 1 巻に含まれない約 3 万語) の調査時期は 2002 年

2 単語親密度調査方法

信頼度の高いデータをクラウドソーシングで集め，かつ，天野ら [1] とできるだけ条件をそろえるため，事前・事後のスクリーニングを実施する．

手順の概要は次の通りである．

1. 事前スクリーニングによる評価者の選定
   漢検 2 級[2]以上か，百羅漢 60 で 54 点以上．
2. インストラクション (図 1) を提示
3. セット単位で評価を実施
   500 語（あるいは 1000 語）のサブセット 2 つで 1 セット．サブセット間で約 5%の語は重複させる．評価は 7 段階．1-7 のうち 2 つ以上使っていない値がある場合や，同じ評価値ばかり付与された場合，アラートが出て評価が保存できない．
4. 事後スクリーニング
   半分以上同じ評価をつけていたり，重複させた語の相関係数が 0.5 より低い評価者を除外する．事後スクリーニングを通った評価者には他のセットの評価も依頼する．

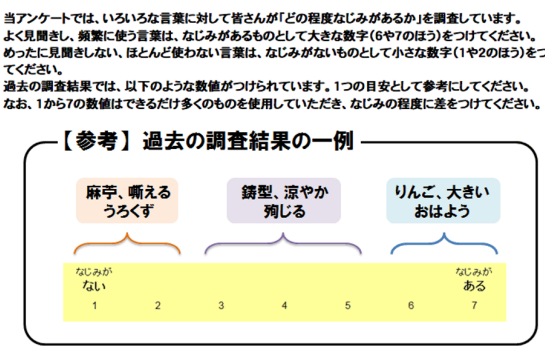

図 1: インストラクション

[2]日本漢字能力検定協会の HP (https://www.kanken.or.jp/kanken/outline/degree.html) によると，2 級は高校卒業・大学・一般程度であり，常用漢字はすべて読み書き活用できるレベル

表 1: 文字単語親密度調査方法の比較

項目	日本語の語彙特性第 1 巻 (第 9 巻)[1] [1]	浅原 [6]	本調査
調査方法	実験室	クラウドソーシング (Yahoo!)	クラウドソーシング (マクロミル)
調査時期	1995.9.–1997.7. (2002.2-3.)	2018.11.	2018.9.–2019.12.
対象語の出典	新明解国語辞典第四版 (学研国語大辞典第二版から追加)	分類語彙表 [7]	日本語の語彙特性 1, 9, 幼児語彙発達 DB[9], BCCWJ[10], NTT 絵本 DB[11]
対象語数	76,945 (32,443)	100,830	163,017
評価の観点	どの程度なじみがあるか	どの程度 知っている/書く/読む/話す/聞くか	どの程度なじみがあるか
評価尺度	1 - 7	1 - 5	1 - 7
親密度の計算方法	平均値	ベイジアン線形混合モデル	平均値
語ごとの評価者数 事後スクリーニング通過	40 人 32 人 (35 人)	16 人以上 -	52 人 平均 23.5 人
異なり評価者数 事後スクリーニング通過	40 人 32 人 (35 人)	3,392 人 -	4,141 人 1,475 人
評価者毎の評価語数 (平均, 最大, 最小)	全語	-	2,160, 25,005, 1,000
評価者の年齢 平均, 標準偏差, 最大, 最小	18 歳以上 30 歳未満 23.0, 2.75, 29, 18	20 歳以上 -	18 歳以上 35 歳以下 29.0, 4.38 , 35, 18
男女比	1:1	-	平均 1:1.3[3]
事前スクリーニング方法	百羅漢[2][8] で 60 点以上 (100 点満点中)	-	漢検 2 級以上か，百羅漢 60[4] で 54 点以上 (60 点満点中) [5]
事後スクリーニング方法	ポストテスト 500 語の評価値の Peason の相関係数が 0.5 以上	-	各テストセット毎に重複させた 5%の語の評価値の相関係数が 0.5 以上
備考	音声/文字音声単語親密度，心像性等も付与		

[1]( ) 内は第 9 巻の値. [2]100 問の漢字の読みテスト. [3]出来る限り 1:1 に近くなるようにしたが，条件を満たすモニタは圧倒的に男性が少なく，1 : 1 の評価者を確保できなかったセットもある. [4]百羅漢 [8] から識別力が高い順に 60 語を選んだ. [5]当初，調査費用削減のため漢検 2 級を条件としたが，評価者が不足したため百羅漢 60 を復活採用した.

調査対象語は，日本語の語彙特性第 1 巻，第 9 巻 [1]（以下，第 1 巻，第 9 巻），幼児語彙発達データベース [9] の全収録語，さらに，BCCWJ[10] と，NTT 絵本データベース [11] の形態素解析結果から固有名詞，フィラー，助詞，助動詞，非自立語，接続詞，接頭辞，接尾辞のみの語を除き，頻度順に約 5 万語を抽出した.

表 1 に，日本語の語彙特性 [1]，浅原 [6]，本調査の比較を示す．本調査では，事前のスクリーニングを実施したうえ，システム上同じ評価ばかり付与できないように制限をかけたが，それでも事後スクリーニングを通過した評価者は異なりで 35.6%のみだった．実験室に評価者を集めた日本語の語彙特性 [1] では通過率が 80% であることを考えると，クラウドソーシングでの通過率は非常に低い．しかも，繰り返し評価を依頼した良好な評価者であっても，数度目のセットでいきなり重複語の相関係数が非常に低くなって継続評価の依頼を打ち切った例もある．浅原 [6] はスクリーニングを実施していないが，本調査結果を見る限りある程度のスクリーニングは実施した方が信頼度の高いデータを得られるのではないかと考えている.

3 調査結果と過去のデータとの比較

過去（第 1 巻，第 9 巻）の調査と本調査における単語親密度のヒストグラムを図 2 に示す．ヒストグラムの形を比べると，第 1 巻では親密度 5.5 あたりに最大の山があり，第 9 巻では親密度 2.9 あたりに山がある．これは，第 9 巻では第 1 巻で調査していない語のみを調査したため，親密度の高い語の多くが調査対象外だったためだと考えられる．本調査のヒストグラムで最も高い山は，親密度 1.9 あたりだが，これも新規に追加した語は，親密度が低い語が多かったのだと考えられる.

比較のため第 1 巻と同じ語のみの結果を図 3 に示す．図 3 では，比較的なだらかな二瘤の山になってい

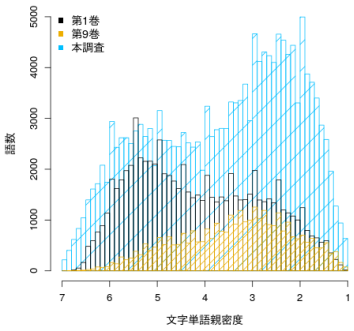

図 2: ヒストグラム:全調査

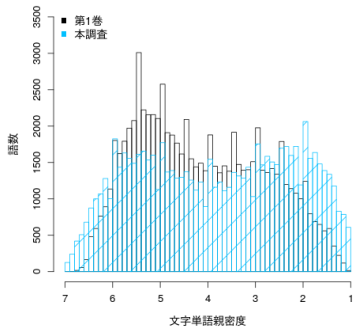

図 3: ヒストグラム：第 1 巻と本調査で同じ語のみ

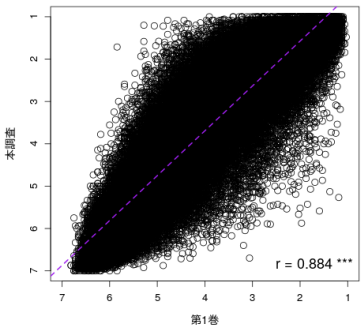

図 4: 散布図

る．また，第 1, 9 巻の両調査では親密度が 1 や 7 となった語はなかったが，本調査では親密度が 1 や 7，つまり，評価者が全員 1 をつけたり，7 をつけた語も相当数あることが分かる．なお，親密度が 7 となった語は，スマートフォン，インターネット，日曜日，玉ねぎなど 41 語であり，親密度が 1 となった語は，優婆夷 (ウバイ)，ベデカー，のんこのしゃあ，素袷わせなど 260 語だった．

第 1 巻と同じ語に対する本調査の親密度の散布図を図 4 に示す．第 1 巻の調査時期は約 20 年前だが，両者の相関係数は 0.884 であり，両者には強い相関がある．本調査ではクラウドソーシングを利用したが，事前・事後のスクリーニング，インストラクションなどを実施することで，実験室での調査と同様，信頼度の高い親密度が得られたと言える．また，20 年経過しても，多くの語では親密度に大きな変化がないことがわかる．

一方で，大きく単語親密度が変化した語も存在する．第 1 巻と本調査の親密度の差が 3 以上あった語を表 2, 3 に示す．表 2 は親密度が上がった語，表 3 は下がった語である．表 2, 3 には，読みも掲載したが，評価者には読みは提示していない．

親密度が大きく上がった語には，アナフィラキシーやレギンス，マニフェストなどが含まれる．一方，親密度が大きく下がった語には，プリンスメロンや生テープ，ミリバールなどが含まれる．例えば，ミリバールの親密度が大きく下がったのは，気圧の単位がミリバールからヘクトパスカル[3]に変わったためだと思われる．変更は 1992 年からであるため，第 1 巻の調査時にもすでにヘクトパスカルに変更されているが，調査時 (1995-1997) に 18 歳から 29 歳なので，まだよく記憶に残っていたと思われる．一方で本調査の評価者は，単位の切り替え前後に生まれているために親密度が低いのだと考えられる．過去の調査も本調査も評価者の年齢を統制したため，こうした差を顕著に捉える事が出来たと考えられる．

4 まとめと今後の課題

本稿では，クラウドソーシングによる信頼度の高い単語親密度調査方法を提案した．日本語の語彙特性第 1 巻，9 巻の両方の語彙だけでなく，幼児語彙発達データベースや，コーパスからも調査対象語を抽出し，約 16 万 3 千語という大規模な調査を実施し，分析結果を報告した．20 年以上前に実験室で調査した結果である日本語の語彙特性第 1 巻と比較し，相関係数 0.884 という強い相関があり，多くの語では親密度は大きく変化しているわけではないことを示した．一方で，親密度が大きく下がった語（プリンスメロンやミリバールなど）や，親密度が大きく上がった語（アナフィラキシーやレギンスなど）もあることも示した．

本稿では，日本語の語彙特性と同様，評価者のスクリーニングを実施して，良好な評価者のみによる評定平均を単語親密度として用いた．今後は，平均だけでなく分散も考慮したり，浅原 [6] と同様ベイジアン線形混合モデルを適用した場合との比較も行いたい．

また本調査では，辞書からだけでなく，幼児が初期に覚える言葉や，コーパスでの高頻度語を調査対象に加えることで，これまで調査対象となってこなかった幼児語や新語，カタカナ語も多く調査した．今後，本調査結果の語彙数調査 [12] への反映や，語彙数推定テストの作成，幼児の語彙発達との関係調査に取り組みたい．

[3]ヘクトパスカルは第 1 巻では調査されておらず，第 9 巻では 4.941, 本調査では 5 だった．

表 2: 単語親密度が 3 以上上がった語

語	読み	第 1 巻	本調査	差
進捗	シンチョク	2.375	5.464	-3.089
乖離	カイリ	2.094	5.192	-3.098
匍匐	ホフク	1.531	4.63	-3.099
膵臓	スイゾウ	2.688	5.889	-3.201
熱中症	ネッチュウショウ	3.188	6.4	-3.212
鱧	ハモ	1.688	4.909	-3.221
重篤	ジュウトク	2.281	5.538	-3.257
シルバーウィーク	シルバーウィーク	2.625	5.9	-3.275
ネグレクト	ネグレクト	1.781	5.071	-3.29
デング熱	デングネツ	1.531	4.864	-3.333
眩暈	ゲンウン	2.000	5.519	-3.519
欅	ケヤキ	1.844	5.429	-3.585
ボールド	ボールド	2.031	5.667	-3.636
付箋	フセン	3.062	6.7	-3.638
オノマトペ	オノマトペ	1.375	5.13	-3.755
マニフェスト	マニフェスト	1.969	5.875	-3.906
レギンス	レギンス	1.594	5.586	-3.992
アナフィラキシー	アナフィラキシー	1.188	5.267	-4.079

表 3: 単語親密度が 3 以上下がった語

語	読み	第 1 巻	本調査	差
プリンスメロン	プリンスメロン	5.844	1.714	4.13
生テープ	ナマテープ	5.281	1.2	4.081
ミリバール	ミリバール	5.062	1.214	3.848
こちこち	コチコチ	5.281	1.583	3.698
旧大陸	キュウタイリク	4.781	1.083	3.698
純毛	ジュンモウ	4.875	1.222	3.653
キーパンチャー	キーパンチャー	5.281	1.81	3.471
軽震	ケイシン	4.625	1.167	3.458
力コブ	チカラコブ	5.344	1.933	3.411
一日増に	イチニチマシニ	4.656	1.25	3.406
特写	トクシャ	4.969	1.579	3.39
海開	ウミビラキ	4.781	1.417	3.364
どんぐり眼	ドングリメ	5.250	1.889	3.361
大足	オオアシ	4.438	1.083	3.355
パリ祭	パリサイ	5.031	1.684	3.347
メイル	メイル	4.562	1.222	3.34
済す	ナス	4.594	1.333	3.261
木戸	キド	4.844	1.6	3.244
増車	ゾウシャ	4.500	1.278	3.222
曲乗り	キョクノリ	4.375	1.167	3.208
ラオチュー	ラオチュー	4.844	1.667	3.177
気病み	キヤミ	4.344	1.167	3.177
旅ガラス	タビガラス	4.406	1.25	3.156
両人	リョウニン	5.062	1.913	3.149
力走	リキソウ	5.469	2.333	3.136
男っ振	オトコップリ	4.625	1.5	3.125
乱脈	ランミャク	4.812	1.714	3.098
ボーゲン	ボーゲン	5.688	2.591	3.097
ロイヤルボックス	ロイヤルボックス	5.250	2.167	3.083
ジルバ	ジルバ	4.781	1.714	3.067
チェッカーフラッグ	チェッカーフラッグ	5.188	2.13	3.058
烈震	レッシン	4.719	1.667	3.052
造	ミヤツコ	4.875	1.833	3.042
コレクトコール	コレクトコール	5.688	2.647	3.041
七難しい	シチムズカシイ	4.438	1.4	3.038
スキャンティー	スキャンティー	4.438	1.4	3.038
紙入	カミイレ	4.500	1.467	3.033
短波	タンパ	5.031	2	3.031
ルンペン	ルンペン	4.875	1.85	3.025
上ぼり	ノボリ	4.938	1.938	3
浅草のり	アサクサノリ	5.000	2	3
教	キョウ	5.000	2	3
ハイミス	ハイミス	4.500	1.5	3

参考文献

[1] 天野 成昭, 近藤 公久. **日本語の語彙特性**. 三省堂, 東京, 1999.

[2] 廣 荻原. 日本人の語彙量 (理解語彙、使用語彙) 調査を行うにあたっての基礎的研究 (日本語学特集). **京都語文**, (21):1–30, 2014.

[3] 水野 りか, 松井 孝雄. 文字の漢字表記語の意味処理に対する構成漢字の影響と処理順序. **心理学研究**, 90(2):201–206, 2019.

[4] 若松 千裕, 石合 純夫, 林 圭輔, 相原 伸子. 語義聾症例における文字言語の理解過程-通常見かけない文字表記語による検討-. **高次脳機能研究** *(*旧 **失語症研究***)*, 36(1):9–19, 2016.

[5] 高橋 三郎. 学齢期の吃音児における語の長さが吃音頻度に及ぼす影響. **音声言語医学**, 59(2):188–193, 2018.

[6] 浅原 正幸. クラウドソーシングによる単語親密度の推定. **言語処理学会第** *25* **回年次大会** *(NLP-2019)*, pp. 45–48, 2019.

[7] 国立国語研究所. **分類語彙表** *CD-ROM (***増補改訂版版***)*. 大日本図書, 2004.

[8] 近藤 公久, 天野 成昭. 百羅漢　～実験参加者の言語能力差の統制のための漢字テスト. *JCSS-TR-69*, 2013.

[9] 小林 哲生, 奥村 優子, 南 泰浩. 語彙チェックリストアプリによる幼児語彙発達データ収集 の試み. **電子情報通信学会技術研究報告**, 115(418):1–6, 2016. (HCS2015-59).

[10] 国立国語研究所 コーパス開発センター. **『現代日本語書き言葉均衡コーパス』利用の手引 第** *1.1* **版**, 2015. `http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/manual/`.

[11] 藤田 早苗, 奥村 優子, 小林 哲生, 服部 正嗣. 絵本と幼児向けの発話に出現する語の多様性比較. **言語処理学会第** *23* **回年次大会** *(NLP-2017)*, pp. 1264–1267, 2018.

[12] 藤田 早苗, 菅原 真悟, 小林 哲生, 新井 庭子, 山田 武士, 新井 紀子. 小学生から高校生に対する語彙数調査と単語親密度との関係分析. **言語処理学会第** *26* **回年次大会** *(NLP-2020)*.