

Bayesian Linear Mixed Model を用いた単語習得時期推定

浅原 正幸 国立国語研究所・総合研究大学院大学 masayu-a@ninjal.ac.jp

概要

本研究では『分類語彙表』全項目について単語習得時期を大規模クラウドソーシング調査で収集し、Bayesian Linear Mixed Model でモデリングを行った。同データは <https://github.com/masayu-a/WLSP-SchoolGrade> にて公開した。

1 はじめに

著者は、『分類語彙表』[1]に対する単語親密度の調査を進めてきた[2, 3]ほか、機能表現辞書『つつじ』[4]についての単語親密度調査[5]を進めてきた。同様の親密度調査はNTT CS 研でも進められている[6, 7, 8, 9]。さらに近年 L1, L2-学習者の日本語語彙難易度の推定も進められている[10, 11]。水谷ら[10]は形態素解析器 JUMAN++ の単語辞書に掲載されている約 26000 語を対象にクラウドソーシング調査を行い、漢検難易度・日本語能力試験の級・単語頻度・NTT 親密度データなどの対照を行った。

本研究では、『分類語彙表』の全語彙を対象に単語習得時期の大規模質問紙調査を行った。調査においては、各語について「知る」「聞いてわかる」「読んでわかる」「話す」「書く」ができるようになった時期を7択の選択肢から選択させた。モデリングはアンケートの選択肢を習得年齢 (age of acquisition: AoA) に写像したうえで、日本語 L1 学習者の単語習得傾向を年齢に基づく対数正規分布と仮定して、Bayesian Linear Mixed Model (BLMM) によるモデリングを行った。さらに『分類語彙表』の語義情報と、各習得時期における親密度の被覆率から、どの時期にどのような概念を習得する傾向にあるかを明らかにした。

2 単語習得時期の質問紙調査

『分類語彙表』の全語彙 100,827 語を対象に、「知る」(KNOW)・「聞いてわかる」(LISTEN)・「読んでわかる」(READ)・「話す」(SPEAK)・「書く」(WRITE) ができるようになった時期を、Yahoo! クラウドソーシングによる質問紙調査により収集した。習得

時期の選択肢として「小学校入学以前」「小学 1-2 年生」「小学 3-4 年生」「小学 5-6 年生」「中学生」「中学卒業後」「知らない・××できない」の7択の粒度とした。付録の図 6 に調査画面の一部を示す。読み書きについては、その表記を問わないという注釈をつけた。2024 年 10 月 16 日 23:00 に開始し、10 月 19 日 20:40 に終了した。1 語あたり 20 人分のデータを収集し、延べ 13,940 人が参加した。15 語回答ごとに 5 円相当の謝礼を支払い、全調査で 1344400 円要した。本調査に際して、国語研の研究倫理審査を受けている。

3 BLMM を用いたモデリング

本節ではアンケートで得られたデータを統計モデリングすることを考える。アンケートで得られるデータは順序尺度である。また実験協力者毎に習得時期の揺れがあることが想定される。ここで、(1) 順序尺度を想定習得年齢に写像してモデリングを行う、(2) BLMM を用いたモデリングを行う、(3) 得られた想定習得年齢を再度順序尺度に写像するなどの工夫を行う。これらの工夫は値域が異なる名義尺度 (順序尺度) のスムージングを目的としている。

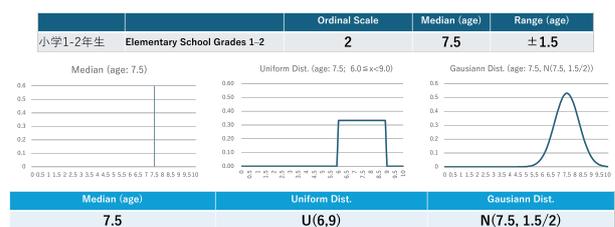


図 1 「小学 1-2 年生」の年齢割り当て

まず、(1) 調査選択肢を想定習得年齢に写像する。表 1 に割り当てをまとめる。中央値と値域は当該学年の可能な年齢とし、中学卒業後は中央値 18、値域 ± 3.0 とした。「知らない」は中央値 128、値域 ± 0.0 とした。サンプリングの手法として (1a) 順序尺度をそのまま使う、(1b) 常に中央値を使う、(1c) 以下の一様分布を仮定してサンプリングを行う、(1d) 以下の正規分布 (平均: 中央値、2SD で値域を被覆) を仮定してサンプリングを行うの 4 種類を検討した。

表 1 調査選択肢に対する想定習得年齢の割り当て

| | Ordinal Scale | Median (age) | Range (age) | Uniform Dist. | Gaussian Dist. |
|-----------|---------------|--------------|-------------|---------------|--------------------------|
| 小学校入学以前 | 1 | 3.5 | ± 3.5 | $U(0, 7)$ | $N(3.5, \frac{3.5}{2})$ |
| 小学 1-2 年生 | 2 | 7.5 | ± 1.5 | $U(6, 9)$ | $N(7.5, \frac{1.5}{2})$ |
| 小学 3-4 年生 | 3 | 9.5 | ± 1.5 | $U(8, 11)$ | $N(9.5, \frac{1.5}{2})$ |
| 小学 5-6 年生 | 4 | 11.5 | ± 1.5 | $U(10, 13)$ | $N(11.5, \frac{1.5}{2})$ |
| 中学生 | 5 | 14 | ± 2.0 | $U(12, 16)$ | $N(14, \frac{2.0}{2})$ |
| 中学卒業後 | 6 | 18 | ± 3.0 | $U(15, 21)$ | $N(18, \frac{3.0}{2})$ |
| 知らない | 7 | 128 | 0 | 128 | 128 |

図 1 に「小学 1-2 年生」の年齢割り当ての例を示す。一様分布（最小値 a, 最大値 b）:

$$U(a, b) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

正規分布（平均 μ , 標準偏差 σ ）:

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

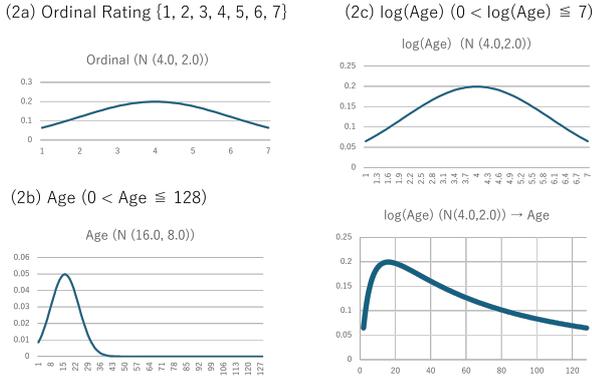


図 2 統計モデル

次に、(2) BLMM を用いたモデリングを行う。(2a) 順序尺度の正規分布 $N(4.0, 2.0)$ 、(2b) 写像した年齢の $N(16.0, 8.0)$ 、(2c) 写像した対数年齢の $N(4.0, 2.0)$ の 3 つのモデルを試行した。図 2 に各モデルを示す。実験協力者については、(2a)、(2c) には $N(0.0, 1.0)$ を、(2b) には $N(0.0, 4.0)$ を、ランダム効果としてモデル化した。各モデルについて、3 回モデル化を行った平均の値を求めた (iter=500, chains=3, warmup=50)。

表 2 推定想定習得年齢の再割り当て

| 区分 | Range (age) |
|-----------|-------------------------------|
| 小学校入学以前 | $0 < \text{age} < 6.5$ |
| 小学 1-2 年生 | $6.5 \leq \text{age} < 8.5$ |
| 小学 3-4 年生 | $8.5 \leq \text{age} < 10.5$ |
| 小学 5-6 年生 | $10.5 \leq \text{age} < 12.5$ |
| 中学生 | $12.5 \leq \text{age} < 15.5$ |
| 中学卒業後 | $15.5 \leq \text{age} < 22.5$ |
| 知らない | $22.5 \leq \text{age}$ |

(2b)(2c) のモデルについては表 2 に基づいて順序尺度に再割り当てを行った。表 3 に (1d) 正規分布で

表 3 モデルの集計結果:(1d: 正規分布)(2c: 対数正規分布)

| 区分 | KNOW | LI | RE | SP | WR |
|-----------|-------|-------|-------|-------|-------|
| 小 入学以前 | 7397 | 7127 | 5259 | 5153 | 1915 |
| 小学 1-2 年生 | 12626 | 12399 | 12490 | 7610 | 4499 |
| 小学 3-4 年生 | 18937 | 18646 | 19204 | 10809 | 7215 |
| 小学 5-6 年生 | 14402 | 14654 | 15190 | 11073 | 8727 |
| 中学生 | 10645 | 10835 | 11250 | 12757 | 12831 |
| 中学卒業後 | 11375 | 11544 | 11786 | 15484 | 20226 |
| 知らない | 25448 | 25625 | 25651 | 37944 | 45417 |

サンプリング (2c) 対数正規分布でモデル化した集計結果を示す。付録表 13, 14, 15, 16, 17, 18 に示す他のモデルと比べて適切にスムージングできており、以下ではこの結果について検証を行う。

4 構築したデータの検証

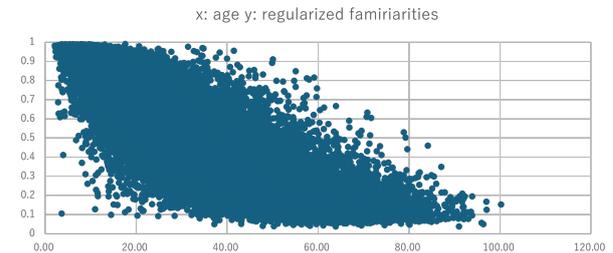


図 3 親密度と想定習得年齢の対照

まず、親密度データ WLSP-Familiarity (Version 4.0)¹⁾ [2, 3] との対照を行う。親密度情報を 0-1 区間に正規化したものと推定した想定習得年齢をプロットしたグラフを図 3 に示す。Spearman の順位相関係数が -0.808 (p 値 < 2.2e-16) であることから、親密度と想定習得年齢と強い逆相関があることがわかる。

次に、頻度データ ndIngramdata²⁾ の tosho-all (図書) と zasshi-all (雑誌) の頻度との相関を検証した。toshos-all について頻度が得られた 96643 語については、Spearman の順位相関係数が -0.236 (p 値 <

1) <https://github.com/masayu-a/WLSP-Familiarity>: 2018 年から 2023 年の 18,686,785 データポイントを再推定したもの。
2) <https://github.com/ndl-lab/ndIngramdata>: tosho-all は図書資料約 97 万点における総出現頻度が 4 以上の 1gram から 5gram までの度数、zasshi-all は雑誌資料約 132 万点における総出現頻度が 4 以上の 1gram から 5gram までの度数。

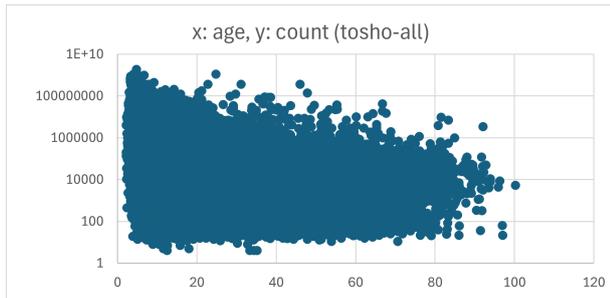


図4 頻度（図書）と想定習得年齢の対照

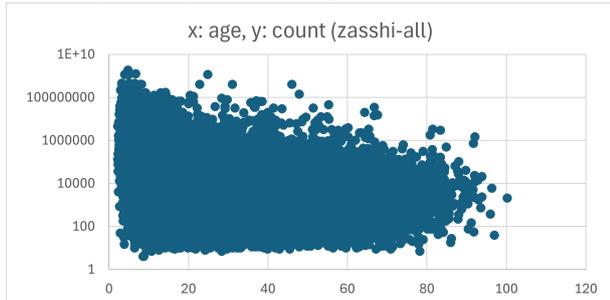


図5 頻度（雑誌）と想定習得年齢の対照

2.2e-16)であることから、弱い逆相関があることがわかる。zasshi-allについて頻度が得られた96990語については、Spearmanの順位相関係数(Rho)が-0.376(p値<2.2e-16)であることから、弱い逆相関があることがわかる。この相関分析の結果から、頻度データから想定習得年齢を推定のは難しいことが示唆される。散布図を図4,5に示す(縦軸の頻度は対数を取った)。

表4 「知っている」推定習得年齢が小さい語

| 語 | 類 | 部門 | 中項目 | 推定習得年齢 |
|-------|---|----|-----|--------|
| さようなら | 他 | | 挨拶語 | 2.15 |
| みんな | 相 | 関係 | 量 | 2.19 |
| おなか | 体 | 自然 | 身体 | 2.23 |
| おしっこ | 体 | 自然 | 身体 | 2.25 |
| きゅうり | 体 | 自然 | 植物 | 2.25 |

表4に「知っている」の推定習得年齢が小さい語の例を示す。なお、より小さい推定習得年齢として「聞くことができる」1.84歳の「お母さん(1.2120: 体-主体-家族-親・先祖)」がある。付録の表19,20,21,22,23に他の学齢で習得される語の代表例を示す。

次に教育基本語彙の各データとの対照を行った。「教育基本語彙の基本的研究」[13]には、7種の教育基本語彙³⁾が含まれている。阪本教育基本語彙は国

3) 阪本教育基本語彙[12]・新版本教育基本語彙[14]・田中教育基本語彙[15]・池原教育基本語彙[16]・児言研教育基本語彙[17]・中央教育基本語彙[18,19]・国語研教育基本語彙[20]の7種。中央教育基本語彙・国語研教育基本語彙は『分類語彙表』初版[21]に基づいて構成されており、同書には『分類

語辞典『言林』[22]をもとに語彙を選定し、国語教育の専門家により3段階の判定を得点化している。『国語学大辞典』では「理解語彙としては少なすぎるし、表現語彙としては多すぎる」という批判がある。A(小学校第1~第3学年)・B(小学校第4~第6学年)・C(中学校の学習段階に分けられるとともに、それらに優先順位(数字)が付与されている。表5に、阪本教育基本語彙[12]と本データ(「知っている」)との対照比較結果を示す。おおむね対角要素に度数が集まっていることから良いモデルになっていると考える。2つのデータの齟齬として「あぜ」(阪本A1・本データ中学生)・「おぶさる」(阪本A1・本データ中学卒業後)のような、近年使われなくなった語もあれば、「朝」(阪本A1・本データ中学生)のように本データの問題点も確認された。

また、『分類語彙表』に想定習得時期を付与したことにより、『分類語彙表』の各概念の語彙がどの時期にどの程度習得されるかを明らかにすることができる。表6に「1.10 体-関係-事柄(KNOW)」の習得時期の分布を示す。分布は単語の頻度のほか、正規化した単語親密度の合計によってもモデル化できる。親密度を用いることにより、あまり知られない(習得されない)語の影響を考慮したモデル化が可能になると考える。

各概念で親密度に基づく被覆率(親密度被覆率)を習得時期毎に集計し、そのパーセンテージが高いもの順に5つずつ並べた表を表7,8,9,10,11,12に示す。

「小学校入学以前」に主に習得される概念では、呼び掛け(38.59%)や感動(35.38%)など、感情やコミュニケーションに関する語彙が習得されている。「小学1-2年生」では、真偽(35.28%)や力(32.57%)、空間(30.19%)など、関係や空間に関連する基礎的な概念の習得が進む。「小学3-4年生」では、生物(47.19%)や身体(44.97%)など、自然や身体に関連する概念がの被覆率が高い。「小学5-6年生」では、交わり(25.70%)、空間(25.23%)、天地(24.63%)など、社会的・自然的な要素に関する概念が習得される。「中学校」では、機関(18.37%)や待遇(17.68%)など、より高度な社会的な概念や活動に関連する語彙が習得される。「中学卒業後」では、言語(14.57%)や経済(14.18%)など、大人

語彙表』初版の語彙に対する7種の教育基本語彙の対照表が収録されている。しかしながら、本研究で利用した『分類語彙表』増補改訂版とはかならずしも1対1対応しないために、一部の語彙が「未定義」となる。

表5 学年別分類データの対比 (行：阪本教育基本語彙 [12]・列：本データ [KNOW])

| 分類 | 学年 | 小入学以前 | 小1-2 | 小3-4 | 小5-6 | 中学生 | 中卒業後 | 知らない | 未定義 | 総計 |
|-----|-------|-------|------|------|------|------|------|------|-----|-------|
| A1 | 小1-3年 | 1469 | 841 | 440 | 166 | 121 | 106 | 143 | 27 | 3313 |
| A2 | 小1-3年 | 819 | 872 | 504 | 221 | 128 | 130 | 152 | 48 | 2874 |
| B1 | 小4-6年 | 297 | 914 | 915 | 368 | 161 | 109 | 115 | 9 | 2888 |
| B2 | 小4-6年 | 186 | 695 | 862 | 404 | 230 | 171 | 150 | 33 | 2731 |
| B3 | 小4-6年 | 191 | 477 | 682 | 391 | 246 | 304 | 317 | 48 | 2656 |
| C1 | 中学校 | 35 | 262 | 847 | 661 | 339 | 255 | 175 | 8 | 2582 |
| C2 | 中学校 | 42 | 179 | 649 | 610 | 403 | 334 | 358 | 10 | 2585 |
| C3 | 中学校 | 62 | 152 | 468 | 494 | 433 | 425 | 529 | 9 | 2572 |
| C4 | 中学校 | 52 | 141 | 379 | 403 | 354 | 411 | 773 | 26 | 2539 |
| 未定義 | | 322 | 345 | 442 | 320 | 210 | 205 | 344 | 307 | 2495 |
| 総計 | | 3475 | 4878 | 6188 | 4038 | 2625 | 2450 | 3056 | 525 | 27235 |

表6 分類番号 1.10 体-関係-事柄 (KNOW) の習得時期の分布

| | 小入学以前 | 小学1-2 | 小学3-4 | 小学5-6 | 中学校 | 中学卒業後 (知らない) | 全体 |
|------------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 頻度 (割合) | 31 (7.23%) | 83 (19.35%) | 85 (19.81%) | 54 (12.59%) | 40 (9.32%) | 47 (10.96%) | 89 (20.75%) |
| 親密度合計 (割合) | 28.47 (8.30%) | 78.52 (22.89%) | 80.01 (23.32%) | 49.41 (14.40%) | 35.10 (10.23%) | 35.37 (10.31%) | 36.18 (10.55%) |

表7 「小学校入学以前」に主に習得される概念 (KNOW)

| 類 | 部門 | 中項目 | 分類番号 | 親密度被覆率 |
|---|----|--------|------|--------|
| 他 | | 呼び掛け | 4.32 | 38.59% |
| 他 | | 感動 | 4.30 | 35.38% |
| 用 | 自然 | 天地 | 2.52 | 32.52% |
| 他 | | 挨拶 | 4.33 | 30.43% |
| 他 | | 動物の鳴き声 | 4.50 | 27.69% |

表8 「小学1-2年生」に主に習得される概念 (KNOW)

| 類 | 部門 | 中項目 | 分類番号 | 親密度被覆率 |
|---|----|-----|------|--------|
| 用 | 関係 | 真偽 | 2.10 | 35.28% |
| 用 | 関係 | 力 | 2.14 | 32.57% |
| 相 | 関係 | 空間 | 3.17 | 30.19% |
| 相 | 自然 | 身体 | 3.56 | 29.12% |
| 相 | 自然 | 物質 | 3.51 | 26.65% |

表9 「小学3-4年生」に主に習得される概念 (KNOW)

| 類 | 部門 | 中項目 | 分類番号 | 親密度被覆率 |
|---|----|-----|------|--------|
| 相 | 自然 | 生物 | 3.53 | 47.19% |
| 用 | 自然 | 身体 | 2.56 | 44.97% |
| 相 | 関係 | 空間 | 3.17 | 38.15% |
| 相 | 関係 | 類 | 3.11 | 32.54% |
| 用 | 自然 | 自然 | 2.50 | 30.31% |

表10 「小学5-6年生」に主に習得される概念 (KNOW)

| 類 | 部門 | 中項目 | 分類番号 | 親密度被覆率 |
|---|----|-----|------|--------|
| 相 | 活動 | 交わり | 3.35 | 25.70% |
| 用 | 関係 | 空間 | 2.17 | 25.23% |
| 相 | 自然 | 天地 | 3.52 | 24.63% |
| 用 | 関係 | 様相 | 2.13 | 24.34% |
| 体 | 活動 | 対遇 | 1.36 | 23.99% |

表11 「中学校」に主に習得される概念 (KNOW)

| 類 | 部門 | 中項目 | 分類番号 | 親密度被覆率 |
|---|----|-----|------|--------|
| 体 | 主体 | 機関 | 1.27 | 18.37% |
| 相 | 活動 | 待遇 | 3.36 | 17.68% |
| 相 | 活動 | 言語 | 3.31 | 17.58% |
| 相 | 活動 | 経済 | 3.37 | 17.56% |
| 用 | 関係 | 類 | 2.11 | 15.74% |

表12 「中学卒業後」に主に習得される概念 (KNOW)

| 類 | 部門 | 中項目 | 分類番号 | 中学卒業後 |
|---|----|-----|------|--------|
| 体 | 活動 | 言語 | 1.31 | 14.57% |
| 体 | 主体 | 機関 | 1.27 | 14.55% |
| 体 | 活動 | 経済 | 1.37 | 14.18% |
| 体 | 関係 | 形 | 1.18 | 13.78% |
| 体 | 自然 | 天地 | 1.52 | 13.62% |

として必要な知識や社会的役割を理解するための語彙が習得される。これらの結果から、学年ごとに習得する概念は多様化し、進行に応じて抽象度や複雑さが増していくことがわかる。このように語義、親密度、習得時期を組み合わせることで、語彙の習得時期に関する解像度が向上し、国語教育に役立つ語彙資源の構築が可能となる。

5 おわりに

本稿では、クラウドソーシング調査を用いた語の習得時期の収集とそのモデリングについて検討を

行った。順序尺度であるカテゴリ頻度データをその可能な年齢を想定してサンプリングを行い、あらたに年齢に基づく対数正規分布を語彙習得曲線と仮定してモデリングを行った。これにより適切なスムージング処理を実現した。分析では主に「知っている」の観点のデータに対して、頻度・既存の教育基本語彙との対照を行うとともに、語義に基づく習得被覆率について分析した。今後「聞く」「読む」「話す」「書く」の観点の検証を行うとともに、阪本教育基本語彙以外のデータとの対照を行う。さらに、実際のL1学習者に対する調査を実施する。

謝辞

本研究は JSPS 科研費 JP23K21935 の助成を受けたものです。また国立国語研究所の共同研究プロジェクトによるものです。

参考文献

- [1] 国立国語研究所（編）. 分類語彙表増補改訂版. 大日本図書, 2004.
- [2] 浅原正幸. Bayesian Linear Mixed Model による単語親密度推定と位相情報付与. 自然言語処理, Vol. 27, No. 1, pp. 133–150, 2020.
- [3] 浅原正幸. クラウドソーシングによる単語親密度データの構築 (2021 年版). 言語処理学会第 28 回年次大会発表論文集, 2022.
- [4] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol. 14, No. 5, pp. 123–146, 2007.
- [5] Bocheng Chen and Masayuki Asahara. Word familiarity rate estimation for japanese functional words using a bayesian linear mixed model. In **Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC 37)**, 2023.
- [6] 天野成昭, 近藤公久（編）. 単語親密度, NTT データベースシリーズ 日本語の語彙特性第 1 巻. 三省堂, 1999.
- [7] 天野成昭, 近藤公久（編）. 単語親密度 増補, NTT データベースシリーズ 日本語の語彙特性第 9 巻. 三省堂, 2008.
- [8] 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.
- [9] NTT コミュニケーション科学基礎研究所. NTT 語彙データベース. NTT 印刷株式会社, 2021.
- [10] 水谷勇介, 河原大輔, 黒橋禎夫. クラウドソーシングを用いた習得時期の想起質問に基づく単語難易度データベースの構築. 言語処理学会 第 25 回年次大会 発表論文集, pp. 1503–1506, 3 2019.
- [11] Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. Japanese lexical complexity for non-native readers: A new dataset. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 477–487, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [12] 阪本一郎. 教育基本語彙. 牧書店, 1958.
- [13] 国立国語研究所. 教育基本語彙の基本的研究: 教育基本語彙データベースの作成. 国立国語研究所報告 117. 国立国語研究所, 2001.
- [14] 阪本一郎. 新教育基本語彙. 学芸図書, 1984.
- [15] 田中久直. 教育基本語彙. 新光閣書店, 1956.
- [16] 池原檜雄. 国語教育のための基本語体系. 六月社, 1957.
- [17] 児童言語研究会. 言語要素指導. 明治図書, 1962.
- [18] 中央教育研究所. 学習基本語彙. 中央教育研究所, 1984.
- [19] 中央教育研究所. 学習基本語彙の基礎調査. 中央教育研究所, 1976.
- [20] 国立国語研究所. 日本語教育のための基本語彙調査. 秀英出版, 1984.
- [21] 国立国語研究所. 分類語彙表. 秀英出版, 1964.
- [22] 新村出. 言林. 全国書房, 1952.

A 参考情報

次の単語をいつ知りましたか？
いつ（聞いてわかる・読んでわかる・話す・書く）ようになりましたか？

【ずきずき】（ずきずき）

いつ知りましたか？

小学校入学以前

小学1-2年生

小学3-4年生

小学5-6年生

中学生

中学卒業後

知らない

図6 Yahoo! クラウドソーシング調査画面（一部）

順序尺度をそのまま正規分布 $N(4.0, 2.0)$ でモデル化すると「知らない」が0になる(表13)。

表13 モデルの集計結果:(1a 順序尺度)(2a 正規分布)

| 区分 | KNOW | LISTEN | READ | SPEAK | WRITE |
|---------|-------|--------|-------|-------|-------|
| 小学校入学以前 | 2898 | 2841 | 1814 | 2219 | 728 |
| 小学1-2年生 | 9960 | 9771 | 9384 | 7482 | 5288 |
| 小学3-4年生 | 21218 | 20894 | 21391 | 16214 | 14106 |
| 小学5-6年生 | 30918 | 30939 | 31644 | 27972 | 28214 |
| 中学生 | 27601 | 27959 | 28127 | 32446 | 36671 |
| 中学卒業後 | 8235 | 8426 | 8470 | 14497 | 15823 |
| 知らない | 0 | 0 | 0 | 0 | 0 |

中央値をそのまま正規分布 $N(16.0, 8.0)$ でモデル化すると「小学校入学以前」「知らない」に偏る(表14)。

表14 モデルの集計結果:(1b 中央値)(2b 正規分布)

| 区分 | KNOW | LISTEN | READ | SPEAK | WRITE |
|---------|-------|--------|-------|-------|-------|
| 小学校入学以前 | 63716 | 63424 | 63272 | 44978 | 29806 |
| 小学1-2年生 | 1552 | 1567 | 1614 | 2397 | 3009 |
| 小学3-4年生 | 1371 | 1412 | 1400 | 2141 | 2804 |
| 小学5-6年生 | 1344 | 1355 | 1391 | 1966 | 2689 |
| 中学生 | 1870 | 1814 | 1880 | 2773 | 3875 |
| 中学卒業後 | 3878 | 3978 | 3993 | 5488 | 8083 |
| 知らない | 27099 | 27280 | 27280 | 41087 | 50564 |

一樣分布もしくは正規分布でサンプリングし正規分布 $N(16.0, 8.0)$ でモデル化すると「小学校入学以前」「知らない」が極端に少なくなる(表15,16)。

表15 モデルの集計結果:(1c 一樣分布)(2b 正規分布)

| 区分 | KNOW | LISTEN | READ | SPEAK | WRITE |
|---------|-------|--------|-------|-------|-------|
| 小学校入学以前 | 12 | 12 | 1 | 13 | 0 |
| 小学1-2年生 | 798 | 773 | 290 | 682 | 113 |
| 小学3-4年生 | 5141 | 5037 | 4095 | 7082 | 6632 |
| 小学5-6年生 | 17733 | 17604 | 17389 | 25582 | 30700 |
| 中学生 | 48237 | 48343 | 49351 | 51376 | 55252 |
| 中学卒業後 | 28909 | 29061 | 29704 | 16095 | 8133 |
| 知らない | 0 | 0 | 0 | 0 | 0 |

表16 モデルの集計結果:(1d 正規分布)(2b 正規分布)

| 区分 | KNOW | LISTEN | READ | SPEAK | WRITE |
|---------|-------|--------|-------|-------|-------|
| 小学校入学以前 | 6 | 6 | 0 | 1 | 0 |
| 小学1-2年生 | 759 | 740 | 273 | 681 | 101 |
| 小学3-4年生 | 5073 | 4990 | 4040 | 7028 | 6510 |
| 小学5-6年生 | 17757 | 17606 | 17346 | 25563 | 30700 |
| 中学生 | 48300 | 48239 | 49462 | 51490 | 55395 |
| 中学卒業後 | 28935 | 29249 | 29709 | 16067 | 8124 |
| 知らない | 0 | 0 | 0 | 0 | 0 |

対数正規分布 $N(4.0, 2.0)$ でモデル化すると「小学校入学以前」「知らない」の両端の問題が緩和される(表17,18,3)。中央値を使うか、一樣分布 or 正規分布からサンプリングするかは要検討。

表17 モデルの集計結果:(1b 中央値)(2c 対数正規分布)

| 区分 | KNOW | LISTEN | READ | SPEAK | WRITE |
|---------|-------|--------|-------|-------|-------|
| 小学校入学以前 | 6938 | 6719 | 4913 | 4843 | 1714 |
| 小学1-2年生 | 12978 | 12725 | 12774 | 7694 | 4588 |
| 小学3-4年生 | 19200 | 18919 | 19478 | 11029 | 7281 |
| 小学5-6年生 | 14433 | 14594 | 15121 | 11170 | 8807 |
| 中学生 | 10519 | 10808 | 11219 | 12811 | 12942 |
| 中学卒業後 | 11420 | 11530 | 11781 | 15460 | 20248 |
| 知らない | 25342 | 25535 | 25544 | 37823 | 45250 |

表18 モデルの集計結果:(1c 一樣分布)(2c 対数正規分布)

| 区分 | KNOW | LISTEN | READ | SPEAK | WRITE |
|---------|-------|--------|-------|-------|-------|
| 小学校入学以前 | 7958 | 7691 | 5704 | 5635 | 2162 |
| 小学1-2年生 | 12212 | 11912 | 12037 | 7245 | 4445 |
| 小学3-4年生 | 18510 | 18285 | 18844 | 10624 | 7085 |
| 小学5-6年生 | 14431 | 14683 | 15247 | 11031 | 8561 |
| 中学生 | 10727 | 10932 | 11377 | 12712 | 12760 |
| 中学卒業後 | 11477 | 11578 | 11842 | 15535 | 20195 |
| 知らない | 25515 | 25749 | 25779 | 38048 | 45622 |

表19 「小学1-2年生」で習得される語(KNOW) 推定習得年齢7.5歳

| 語 | 類 | 部門 | 中項目 | 分類番号 |
|----------|---|-----|-----|--------|
| 使い切る | 用 | 関係 | 存在 | 2.1250 |
| 雪見 | 体 | 活動 | 生活 | 1.3371 |
| 車両 | 体 | 生産物 | 機械 | 1.4650 |
| くだらない | 相 | 活動 | 経済 | 3.3710 |
| ふにゃふにゃする | 用 | 自然 | 自然 | 2.5060 |

表20 「小学3-4年生」で習得される語(KNOW) 推定習得年齢9.5歳

| 語 | 類 | 部門 | 中項目 | 分類番号 |
|-------|---|-----|-----|--------|
| 花時計 | 体 | 生産物 | 機械 | 1.4640 |
| 持ち味 | 体 | 関係 | 様相 | 1.1302 |
| 午(うま) | 体 | 関係 | 量 | 1.1961 |
| 定時 | 体 | 関係 | 時間 | 1.1611 |
| 食肉 | 体 | 生産物 | 食料 | 1.4323 |

表21 「小学5-6年生」で習得される語(KNOW) 推定習得年齢11.5歳

| 語 | 類 | 部門 | 中項目 | 分類番号 |
|----------|---|----|-----|--------|
| 勝ち星を上げる | 用 | 活動 | 交わり | 2.3570 |
| オーバーランする | 用 | 関係 | 作用 | 2.1524 |
| 侵略する | 用 | 活動 | 交わり | 2.3580 |
| 本意 | 体 | 活動 | 心 | 1.3045 |
| たしなみ | 体 | 活動 | 行為 | 1.3421 |

表22 「中学校」で習得される語(KNOW) 推定習得年齢14.0歳

| 語 | 類 | 部門 | 中項目 | 分類番号 |
|------|---|-----|-----|--------|
| クローブ | 体 | 生産物 | 食料 | 1.4330 |
| 活魚 | 体 | 生産物 | 食料 | 1.4323 |
| あわ粒 | 体 | 自然 | 植物 | 1.5410 |
| 先んずる | 用 | 関係 | 時間 | 2.1650 |
| 斎場 | 体 | 主体 | 社会 | 1.2630 |

表23 「中学卒業後」に習得される語(KNOW) 推定習得年齢18.0歳

| 語 | 類 | 部門 | 中項目 | 分類番号 |
|-------|---|----|-----|--------|
| 実損 | 体 | 活動 | 経済 | 1.3750 |
| 乱発・濫発 | 体 | 活動 | 事業 | 1.3832 |
| - んとこ | 体 | 関係 | 空間 | 1.1700 |
| 愚直 | 相 | 活動 | 行為 | 3.3421 |
| 劇作 | 体 | 活動 | 芸術 | 1.3240 |