クラウドソーシングを用いた語彙テスト結果データセット作成

江原　遥
産業技術総合研究所　人工知能研究センター
y-ehara@aist.go.jp

1　はじめに

語学学習者支援の場面においては，語学学習者が習得していない事柄を把握することが肝要である．なぜなら，語学学習者は，習得済みの事柄に対する支援よりも未習得の事柄に対する支援を必要としているからである．これは，語彙学習の支援においても同様であり，語学学習者が習得していない語を把握することが，語学学習者に対する支援の第一歩となる．過去，工学的な支援の目的で，実際に語学学習者が未習得の語を，語学学習者に大きな負担をかけずに検出する研究が行われてきた [7, 6, 8, 5, 4]．こうした研究には，実際に，語学学習者が学習済みの語を記録したデータセットが欠かせない．すなわち，語学学習者に語彙テストを実施し，そのテスト結果を収集したデータセットが必要となる．

こうした，語彙テストの結果にデータセットは過去に存在するものの，一般公開されてダウンロード可能な形になっているデータセットは少ない．その主たる理由は，語彙テストが学校の教室など教育の過程で行われるものであって，そもそも外部に公開することが難しいからであると推察される．また，教室で収集されたデータセットは，対象となる語学学習者の年齢層や学習歴が，限定されてしまう傾向があるという問題もあり，幅広い語学学習者を対象とした工学的なシステムを作成する目的に適しているとはいいがたい．

そこで，本研究では，一般公開可能な英語学習者の語彙テスト結果データセットをクラウドソーシングを用いて作成し，公開し[1]，その詳細について本稿で述べる．クラウドソーシングでは，インターネット上の不特定多数のユーザに対して有償で語彙テストを受けてもらい，データを作成した．提案するデータセットでは，100 人の英語学習者に，100 問からなる単語テストに回答してもらうことで作成した．作成にかかった費用は 5 万円未満であった．作成したデータセットがテストデータの信頼性について統計的実験を行い有望な結果を得た．

本研究の貢献は，下記のとおりである．

- 一般公開可能な語彙テスト結果データを作成したこと
- クラウドソーシングにより，低価格で同種のデータセットが作成可能であることを示したこと
- 作成したデータセットの信頼性について有望な結果を得たこと（クロンバックのアルファ係数 0.91）

2　関連研究

語学学習者の語彙テストについての研究は多くあるが，一連の研究の中で一般にデータセットまで公開されているものは，著者の知る限りない．語彙テストについての研究では，多くは，百人程度から数百人程度の参加者に語彙テストを受けてもらい，その語彙テストの結果の統計的性質が報告されている．文献 [2] では，多肢選択式や自己申告式など，3 種のテスト形式の比較がされており，どのテスト形式でも語学学習者の語彙能力を高い信頼性で計測できることを確認している．文献 [2] の研究では，54 問からなる語彙テストに 167 人の大学生が回答したとあるが，論文中からはこのデータセットがいっぱい公開されているようには読み取れなかった．近年，1,000 人の語学学習者に対して数百語からなる語彙テストを実施することによって得たテスト結果についての報告 [14] もあったが，このデータも一般公開されていない．

自然言語処理の研究では，語学学習者が書いたテキストを集めた，学習者コーパス形式のデータを用いた研究が多い．学習者コーパスを用いた研究他，自然言語処理分野における語学学習者研究は，この文献 [15] によくまとめられている．学習者コーパス形式のデータは，学習者が作文を行うときに用いた語彙（生産語彙）の情報を含む重要な言語資源ではある．しかし，語彙テストと大きく異なり，語学学習者によって作文に用いた語が異なるため，1 語あたりの語学学習者の

[1] http://vocabularyprediction.com/ にて公開．利用の際に引用することが望ましい文献情報も同ページに記載した．

反応数が少なくなる問題がある．例えば，100 人の語学学習者に 100 問（1 問 1 語彙）からなる語彙テストを実施した場合には，1 語あたり 100 人の語学学習者の反応が得られるが，同人数の語学学習者に 100 語からなる作文を書いてもらった場合では，ほとんどの語について，その語を用いた語学学習者の数は 10 人を下回ることが容易に予想される．言い換えると，行を語学学習者，列を語彙とする行列の形式でデータを表現した場合，学習者コーパス形式のデータは，通常，反応が記録されている行列内の要素が非常に少なく，この行列が疎になるということである．

直近では，2018 年 1 月に Duolingo の研究者らが中心となって，第二言語獲得のモデリングのためのデータセット (Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling, SLAM）が公開され，関連する Shared Task が始まった [13]．このデータセットは対象とする語学学習者が数十万人と非常に大規模である上，語学学習者の経時的な変化が記録されており，非常に興味深い公開データセットではある．しかし，著者が独自にこのデータセットを簡単に確認した結果では，前述の行列形式でデータを表現した場合，行列中の全要素数の約 0.7%程度であり，やはり疎になる問題を抱えているように見える[2]．

著者が過去に収集し公開したデータセット [7][3]は，こうしたデータセットのうち，公開されている数少ないものの 1 つである．このデータセットでは，12,000 語について，16 人の語学学習者に自己申告式で語を知っている度合いを回答してもらった．本研究で作成したデータセットとの違いは，[7] のデータセットが自己申告式であるのに対して本稿で扱うデータセットは多肢選択式で収集した点，また，データセット中で対象とした語学学習者の数がこのデータセットでは 16 人であったのに対し，本稿で提案するデータセットでは 100 人であることである．

3 データセット

本稿におけるデータセット作成は，主たる目的として，言語支援システムの作成や評価に用いるという工学的な目的を想定している．こうしたシステムの多くは，不特定多数の語学学習者によって使用されることが想定されている．この点は，語学教師が教室で語学学習者に対して試験を施す場合と大きく異なる．例えば，前者の不特定多数の語学学習者を対象とする設定では，幅広い年齢層など多様な背景の語学学習者を対象とする必要があることが想定されるが，後者の教室の設定では，語学学習者は，年齢層や居住地区などの属性が限定されていることが容易に予想される．

幅広い背景を持つ作業者を対象にしたデータセットを，低額な予算で作成するため，Web 上で不特定多数の作業者に有償で作業してもらう仕組みであるクラウドソーシングを用いた．クラウドソーシングサービスとしては，Lancers[4]を用い，データ取得を行った時期は 2016 年 1 月から同年 3 月である．同社は日本の会社であるため，作業者の大半は日本語を母語とする．語彙テストとしては，Vocabulary Size Test (VST)[11] を用いた．このテストは，英語学習者の受容語彙サイズを測定する目的に適した試験であり，独立の著者による研究で追試されている [2]．試験は 100 語からなり，30 分程度で回答できるように設計されている[5]．100 人の作業者に，1 人あたり 383 円を払った．実際の問題は，例えば，次のようである．

microphone: Please use the <microphone>.

a machine for making food hot
b machine that makes sounds louder
c machine that makes things look bigger
d small telephone that can be carried around

このテスト結果をほかのテスト結果と比較するために，作業者は，Test of English for International Communication (TOEIC) テスト受験者に限定し，TOEIC テストの合計点数 (リスニング点数とリーディング点数の和) も作業者から収集した．ただし，十分な作業者数を確保するため，TOEIC のテストを受験した日付は限定せず，TOEIC 合計点数は自己申告式とし，TOEIC 受験や点数を証明する書類なども必須としなかった．また，TOEIC テストはリスニングとリーディングの 2 つに分かれて点数がつくが，この点数の内訳については記載を必須とせず，合計点数のみ記載必須とした．TOEIC テストは受験時間が約 2 時間であるのに対して，利用した Vocabulary Size Test は約 30 分であるため，後者から測定した能力値を用いて前者の点数をうまく予測することができれば，Vocabulary Size Test が短時間での英語力のスクリーニングの目的で利用するにも有望であるといえる．

[2] “data_en_es.tar.gz’z 中の “en_es.slam.20171218.train” を調べた．語は Python 言語の NLTK ライブラリに含まれる PorterStemmer で stemming を行った．246,229 人× 1,552 語の要素を持つ行列に対し，反応が記録されている要素数は 2,622,958 件であった．

[3] http://yoehara.com/esl-vocabulary-dataset/ にて公開

[4] https://www.lancers.jp/

[5] https://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

4 テスト理論を用いた分析

4.1 記法

作業者が前述のような問題に回答する場合を考える．この時，心理統計の分野では，前述のような 1 問を “項目” と呼ぶことが多い．作業者の集合を I とし，項目の集合を J とする．各作業者 $i \in I$ が各項目 $j \in J$ に回答するとき，各項目に対する反応を y_{ij} とする．作業者 i が項目 j に正答した場合 $y_{ij} = 1$，誤答した場合 $y_{ij} = 0$ とする．

4.2 クロンバックのアルファ係数

クロンバックのアルファ係数は，テストの信頼性の尺度の一つである．これは，直感的には，反応データがどの程度一次元の特性（すなわち，能力）を反映しているとみなせるか，を示す尺度である．また，クロンバックのアルファ係数は，内的整合性の尺度であるとも解釈できる．項目集合が，すべて同じ “能力” といえるような 1 次元の尺度を測定しているのであれば，例えば，項目集合をランダムに 2 組に分割したときに，片方の組で似た反応パターンを示した作業者群は，もう片方の組での反応パターンも似ているはずである．この度合いを測定しているのがクロンバックのアルファ係数といえる．

上記の記法を用いると，クロンバックのアルファ係数は次のように表せる．

$$\alpha = \frac{|J|}{|J|-1}\left(1 - \frac{\sum_{j=1}^{|J|} P_j(1-P_j)}{\sigma_X^2}\right) \qquad (1)$$

ただし，σ_X は，すべての反応の分散であり，P_j は項目 j の正答率である．

クロンバックのアルファ係数の計算には，R 言語を用いた．提案するデータセットでは，クロンバックのアルファ係数は 0.91 であり，“excellent” と判定された [9, 10, 3]．したがって，提案するデータセットは，高い信頼性を持つといえる．

4.3 項目反応理論

項目反応理論 (Item Response Theory (IRT)) は，項目応答理論とも呼ばれ，テスト結果の分析に広く利用されている理論である [1]．単一の確率モデルを指す名称ではなく，類似の考え方をする多様な確率モデルに対する総称である．ただし，実際には，その中でも特に 2 パラメータモデル (**2PL**)，また，1 パラメータモデル (**1PL**) が特によく用いられている．1PL は Rasch モデルとも呼ばれる [12]．どちらの問題でも，各項目（設問）は確率的に独立であると仮定されており，作業者のある項目に対する反応が別の項目に対する回答に影響することはなく，あくまでその項目自体と作業者の特性だけから回答が決まると仮定されている．これは，直感的には，最初に長い文章題を読ませた後で複数の読解問題に回答するようなケースを考慮せずに，単純化しているといえる[6]．作業者間についても同様に独立性が仮定されており，ある作業者の回答は，ほかの作業者の回答に影響しないと仮定されている．

2PL は，**1PL** の一般化になっているのでまとめて記す．具体的には，作業者 i が項目 j に正答する確率を次の式でモデル化する．ここで，σ はロジスティックシグモイド関数であり，$t \in \mathcal{R}$ に対して $\sigma(t) := \frac{1}{1+\exp(-t)}$ と定義される．

$$P(y_{ij} = 1|i, j) = \sigma\left(a_j(\theta_i - b_j)\right) \qquad (2)$$

式 2 には，項目 j に対して，a_j と b_j という 2 つのパラメータが存在し，作業者 i に対しては θ_i という一つのパラメータが存在する．θ_i は能力パラメタと呼ばれ，作業者 i の能力を表す．b_j は困難度パラメタと呼ばれ，項目 j の難しさを表す．式 2 の $\theta_i - b_j$ という部分に注目してほしい．ロジスティックシグモイド関数は単調増加関数であるので，θ_i の値が大きいほど，作業者 i が項目 j に正答する確率が高くなる．反対に b_j の値が大きいほど，作業者 i が項目 j に正答する確率が低くなる．$\sigma(0) = 0.5$ であるため，作業者 i が項目 j に正答する確率が，誤答する確率より高くなるのは $\theta_i > b_j$ を満たす時であり，逆もまた真である．

a_j は，項目 j の識別力パラメタと呼ばれ，直感的には，問題の品質の良さ，すなわち，良問である度合いを表す指標の 1 つである．a_j は,「項目 j が低能力（θ_i の値が低い）作業者集合と，高能力の作業者集合をどれだけ明確に分離（識別）することができる度合い」を表している．言い換えると，能力値パラメタの値が十分離れた 2 人の作業者がいるとき，a_j が大きくなればなるほど，2 人の間の項目 j の正答確率の差は大きくなる．

式 2 は，視覚的に図示でき，項目特性曲線 (Item Characteristic Curve, ICC) と呼ばれる．この図では，縦軸が確率値，横軸が θ_i を表し，横軸であらわされる値を能力値パラメタとして持つ作業者が項目 j に正答する確率を曲線で表現している．提案するデータセットについても ICC を計算し，図 1 に示す．

[6]項目反応理論の中には，こうしたケースに考慮したモデルも存在するが，本稿では扱わない．

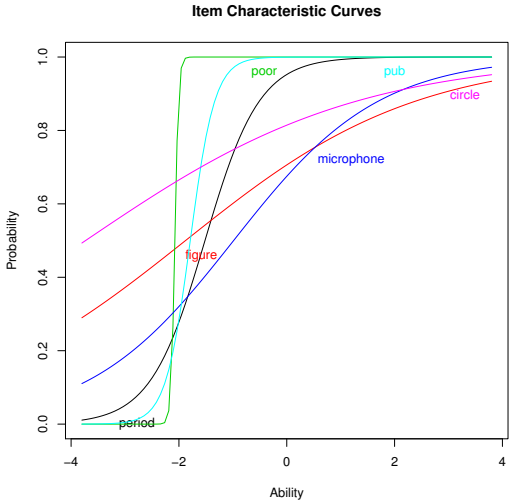

図 1: 提案データセット中の項目特性曲線．縦軸 Probability は確率値，横軸 Ability は能力値パラメタの値．

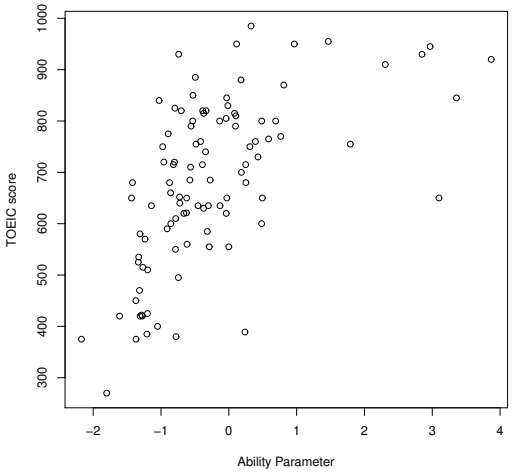

図 2: TOEIC スコア（縦軸）に対する Vocabulary Size Test の結果から求めた能力値パラメタ（横軸）との散布図．

4.4 TOEIC スコアを用いたテストの信頼性評価

前述のように，提案するデータセットでは，作業者は過去に TOEIC テストを受けたことのあるものに限定した．TOEIC テストは，主に日本と韓国で受験者が多い試験で，受験した者の英語力を総合的に測定することが可能とされる．TOEIC テストの受験は 2 時間かかり，受験者に大きな負担を与える．Vocabulary Size Test は 30 分ほどですむので，もしこのデータセットから計算された能力値パラメタから TOEIC のスコアを予測することが可能となれば，受験者の負担を大きく減らすことができる．

図 2 に，各作業者の TOEIC スコアと，Vocabulary Size Test の結果に対して **2PL** で求めた能力値パラメタ θ との相関を示す．能力値パラメタ θ を持つ作業者の TOEIC スコアは $86.50\theta + 703.08$ と表せるという線形回帰が得られた ($p < 0.001$)．したがって，Vocabulary Size Test の結果に対して **2PL** で求めた能力値パラメタの値は，TOEIC スコアの良い推定量になっているといえる．

5 おわりに

本稿では，語彙学習支援システムを作成・評価するなどの工学的な目的に使える，広く入手可能な語彙テスト結果のデータセットを提案した．提案するデータセットは，多肢選択式でクラウドソーシングを用いて作成され，クロンバックのアルファ係数が 0.91，データセットに対して項目反応理論の **2PL** を適用することで求めた困難度パラメタと，TOEIC スコアが $p < 0.001$ で創刊することから高い信頼性を示した．

今回提案したデータセットは，語彙学習以外にも語学学習支援における工学的研究に幅広く利用することが可能であると考えられる．例えば，他の語学学習支援のタスクにおいても，本データセットを併用してマルチタスク学習を行い，性能を向上させるような手法を作成することは可能であろう．今後は，本データセットを幅広く利用していただけるように研究活動を続けていきたい．

謝辞

本研究は JSPS 科研費 15K16059 の助成を受けたものです．

参考文献

[1] Frank B. Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, New York, second edition, 2004.
[2] Brent Culligan. A comparison of three test formats to assess word difficulty. *Language Testing*, Vol. 32, No. 4, pp. 503–520, 2015.
[3] Robert F DeVellis. *Scale development: Theory and applications*, Vol. 26. Sage publications, 2016.
[4] Yo Ehara, Yukino Baba, Masao Utiyama, and Eiichiro Sumita. Assessing translation ability through vocabulary ability assessment. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI*, pp. 3712–3718, 2016.
[5] Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1374–1384, Doha, Qatar, October 2014. Association for Computational Linguistics.
[6] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December 2012.
[7] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized reading support for second-language web documents by collective intelligence. In *Proceedings of the 15th international conference on Intelligent user interfaces (IUI 2010)*, pp. 51–60, Hong Kong, China, 2010. ACM.
[8] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, Vol. 4, No. 2, 2013.
[9] Darren George. *SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e*. Pearson Education India, 2011.
[10] Paul Kline. *Handbook of psychological testing*. Routledge, 2013.
[11] Paul Nation and David Beglar. A vocabulary size test. Vol. 31, No. 7, pp. 9–13, 2007.
[12] Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, 1960.
[13] Burr Settles. Data for the 2018 duolingo shared task on second language acquisition modeling (slam), 2018.
[14] Stuart Webb, Yosuke Sasao, and Oliver Ballance. The updated vocabulary levels test. *ITL - International Journal of Applied Linguistics*, Vol. 168, No. 1, pp. 33–69, 2017.
[15] 永田亮. 語学学習者支援のための自然言語処理. コロナ社, 2017.