外国語学習者にとって意外な用例の難しさ測定のための語彙テスト・データセット構築

江原遥 [1]
[1] 東京学芸大学　教育学部
ehara@u-gakugei.ac.jp

概要

外国語学習者（第二言語学習者）が初期に学ぶ語の多くは高頻度語であり，よく使われる用例とは異なる意味で使われる用例がある．こうした学習者にとって意外な用例は，よく使われる用例より難しいと思われるが，どの程度難しいのであろうか？このような文脈を考慮した用例ごとの語の難しさ測定には，近年急速に発達している文脈化単語埋め込みなど，文脈の意味を捉える深層学習技術の活用が容易に考えられる．しかし，これを評価するために，同一の学習者にある語の複数の用例についての設問に回答させる語彙テストやその結果のデータセットが，既存にはほとんど存在しない．そこで，本研究では，語の用例ごとの難しさ評価のためのデータセットを構築し，項目反応理論に基づき評価する．

1　はじめに

外国語学習において，語の習得は，外国語学習の大部分の時間を占める．その数が膨大であるためか，外国語学習における語の習得においては，語彙量が主に注目されてきた．例えば，有名なテキスト中の語の 95%〜98%以上を（述べ語数で）知らないと，テキストを十分に理解する事が難しい [1] といった応用言語学分野の実験結果では，テキスト中の語をカウントする際に，語の多義性は考慮していない．これは，テキスト中の語の語義を自動的に推定する事が技術的に難しかったためと考えられる．

一方で，外国語学習において初期に学ぶ単語の多くが高頻度語であり，高頻度語は多義性を持つことが多い．語彙量計測などの目的で用いられる語彙テストのうち，一般的な多肢選択式の試験として広く用いられている Vocabulary Size Test [2] では，語義を限定する工夫が施されている．具体的には，表 1 に示すように，設問で語だけを提示するのではなく，文中に語を入れ込み，その語の意味と最も近い選択肢を選択させる形式になっている．さらに，文法的な証拠から適切な選択肢を選ぶことができないよう，どの選択肢も，文中の指定部分とそのまま入れ替えても，文法的には正しい文になるように作成されている．

語彙テスト結果データ，（どの学習者がどの設問に正答/誤答したか，というデータ）から，学習者の能力値や語の難しさを推定するモデルとしては，項目反応理論 (Item Response Theory) [3] が教育学や心理学の分野で広く使われている．しかし，項目反応理論は設問の意味的情報を一切用いずに，純粋に被験者の正答/誤答の情報だけから語の難しさを推定するモデルである．このため，設問が少しでも変更された場合，再度，語彙テストを行わなければ結果がわからない．例えば，語の用例ごとに外国語学習者にとっての難しさを測定するためには，各語について語の用例の分，適切な作問を行う必要がある．例えば，1 万語の各語の用例についてて，適切な多肢選択式の作問を行えば，設問数は数万にもなり，こうした問題の全てに学習者に回答してもらう事は明らかに非現実的である．近年では BERT(bidirectional encoder representations from transformers)[4] をはじめとする，大規模な母語話者コーパスで事前学習を行う事で文脈を捉えた意味処理を行う手法が盛んである．そこで，こうした手法を用いて，設問の用例と頻出用例との近さを計測する事で，語の難しさを計測する手法が容易に考えられる．しかし，こうした発展性のある手法を評価するためには，まずは，評価のために信頼できるデータセットが必要になる．現状では，著者の知る限りそうしたデータセットは整備されていない．

そこで，本稿では，語の用例ごとの難しさ測定のために，英語を第二言語として学ぶ学習者 (English-as-a-Second-Language Learners) を対象とした

語彙テストを構築した．この語彙テストで，語の意外な用例の設問については，複数の英語母語話者・大学英語教員に問題として成立していることを確認した．次に，この語彙テストを用いてクラウドソーシングを用いて，学習者の被験者反応実験では，実際に項目反応理論を用いて，被験者反応のみを用いて，語の通常の用例と，意外と思われる用例の間の難しさの差を計測した．項目反応理論を用いて分析した結果，語の通常の用例の方が，統計的有意に難しく，良問とみなせる度合い（識別力）も高いことを定量的に示した．本研究で作成されたデータセットは，今後[1)]で公開する予定である．

2 関連研究

本研究で提示するデータセットの必要性を既存研究との関連から説明する．1 つは，語学学習アプリ Duolingo 上の設問に対する回答データを用いた SLAM データセット [5] である．もう 1 つは，多数の語学学習者に対して，文中のわからない語をアノテーションさせた複雑単語推定（Complex Word Identification, CWI）のデータセット [6] である．これらのデータセットと本研究で提示するデータセットの違いとして，各被験者は多くある設問のうちのごく一部にしか回答していないという点が挙げられる．言い換えると，被験者を行，設問を列とし，被験者の設問に対する回答内容を要素とする行列を考えた場合，これらのデータセットでは行列が疎になっている．項目反応理論は，被験者の設問に対する回答内容から，設問の難しさや被験者の能力値を推定を目標とするが，この推定のためには，各被験者がほぼ全ての設問に回答している形式のデータセットであることが望ましい．また，どちらのデータセットでも，文中の語に対する被験者の回答が記録されてはいるものの，設問について，今回のデータセットのような語の通常の用例と，意外と思われる用例といったようなアノテーションはされていない．さらに，[6] を含む CWI のデータセットでは，一般に，提示された文に対して，被験者が難しいと感じた語が記録されているだけであり，被験者が実際にその語の意味を適切に理解しているかテストを通じた確認はしていない．すなわち，意味は理解できたが難しいと感じてアノテーションした場合もあれば，単純に意味が分からなかった場合も含まれる．

1） http://yoehara.com

表 1 実際の設問例

It was a difficult period.
a) question
b) time
c) thing to do
d) book

3 語彙テスト作成・データセット

語彙テスト作成・データセット作成は，著者が過去に語彙テスト結果データセット作成時の設定に準じて行った [7]. データセットはクラウドソーシングサービス Lancers[2)]から，2021 年 1 月に収集した．英語学習にある程度興味がある学習者を集めるため，過去に TOEIC を受験したことがある学習者のみ語彙テストを受けられると明記して，データを収集した．その結果，235 名の被験者から回答があった．Lancers の作業者は大部分日本語母語話者であるため，学習者の母語は，大部分日本語を母語とするものと思われる．

まず，通常の語彙テストとしては，文献 [7] と同様に，Vocabulary Size Test （VST）[2] を用いた．ただし，VST は 100 問からなるのに対して，[7] では，低頻度語に関する設問では，Lancers 上のどの学習者もほとんどチャンスレートしか回答できていなかったことから，被験者の負担感を減らし的確な回答を集めやすくするため，低頻度語 30 問を削った．すなわち，残り 70 問を通常の語彙テストとして用いた．この設問例を表 1 に示す．文中の単語に下線が引かれてあり，被験者は，この単語と交換した際に元の文と意味が最も近くなる選択肢を選ぶように求められる．この際，文法的から選択肢を絞れてしまわないように，選択肢は下線部と文字通り置き換えても正文となるように作られている．例えば表 1 であれば，複数形の選択肢が内容に配慮されている．

一方，学習者にとって意外である思われる用例については，著者が作問し，英語母語話者を含む静岡理工科大学の教員複数名に問題として成立しているか確認を取る方法で，作成した．この際，表 1 と同様の形式にして，“period” という単語について 2 つの設問がある事が分かってしまうと，意外な語義については通常の語義以外の選択肢を選ぶことで，選択肢を絞り，意味を知らなくても回答できてしまう．そこで，本研究では，次の 2 つの工夫を行った．

1. 意外な語義を問う設問については，下線部の意

2） https://lancers.co.jp/

表 2 意外な意味を問う設問例

She had a missed _____.
a) time
b) period
c) hour
d) duration

味について問う形にはせず，空欄を埋める形式の問題とした．これにより，意外な語義については正答を知らなければ，どの語についての設問であるのかもわからないようにした．

2. 通常の語義についての設問を先に行ってしまうと，そこで出てきた単語と同じ語が正答であろう，という推測ができてしまう．そこで，意外な語義についての設問群を最初に行い，通常の語義についての設問群に移動したら，意外な語義についての設問群には戻れないようにした．

この 2 つの工夫を施した実際の設問例が表 2 である．“period” には通常の「期間」の他に「生理」という意味があり，これを問うている．被験者は，70 問の通常の用例の語彙テストの前に，表 2 のような設問を 13 問解くように求められる．ただし，先に解く表 2 の形式の選択肢が，表 1 の形式の問題に影響していないかどうかを後で確認できるよう，意外な語義ではあるが，通常の語義の設問群の側に対応する設問がない設問を 1 問設けた．これにより，対応する問題は 12 問となる．

4 項目反応理論

項目反応理論モデルについて，簡単に説明する．被験者の数を J，設問（項目, item）の数を I とする．簡単のため，被験者の添字（index）と被験者，項目の添字と項目を同一視する．例えば，i 番目の項目を，単に i と書くことにする。y_{ij} は，被験者 j が項目 i に正答するとき 1，誤答であるとき 0 であるとする．試験結果データ $\{y_{ij}|i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}\}$ が与えられたとき，2PL モデルでは，被験者 j が項目 i に正答する確率を次の式でモデル化する．

$$P(y_{ij} = 1|i, j) = \sigma(a_i(\theta_j - d_i)) \tag{1}$$

ここで，σ は $\sigma(x) = \frac{1}{1+\exp(-x)}$ で定義されるロジスティックシグモイド関数である．σ は $(0, 1)$ を値域とする単調増加関数であり，$\sigma(0) = 0.5$ である．実数を $(0, 1)$ の範囲に射影し，確率として扱うために用いられている．(1) において，θ_j は能力パラメタ（ability parameter）と呼ばれ，被験者の能力を表すパラメタである．d_i は困難度パラメタ（difficulty parameter）と呼ばれ，項目の難しさを表すパラメタである．(1) より，θ_j が d_i を上回る時，被験者が正答する確率が誤答確率より高くなる．$a_i > 0$ は，通常，正の値を取り，識別力パラメタ（discrimination parameter）と呼ばれる．この値が大きいほど，$\theta_j - d_i$ が正答確率/誤答確率に大きく影響するようになる．$\theta_j - d_i$ を用いて，被験者 j が設問 i に正答するか否かが見分けやすくなる事を表しているため，「識別力」と呼ばれる．より直観的には設問 i が，能力値が高い学習者と低い学習者を正確に見分けられるという意味で良問であることを示している．

5 実験

項目反応理論の困難度・識別力の各パラメタを求めるには，pyirt [3] を用いた．これは，周辺化最尤推定（Marginalized Maximum Likelihood Estimation）により項目反応理論を行うライブラリである．前述のデータセットに対して，2PL モデルを用いて困難度と識別力パラメタを求めた．表 1 と表 2 のように，設問のペアが 12 組ある．通常の用例，学習者にとって意外と思われる用例の困難度パラメタを，それぞれ横軸，縦軸に表し，横軸と縦軸の縮尺・範囲を同一にプロットし図 1 に示した．各点は語を表す．

困難度の比較 図 1 の左下から右上まで，点線で対角線を示した．図 1 の横軸・縦軸とも，困難度パラメタの値であり，この値が大きいほど難しいと判定される．そのため，この対角線より左上にある点は，通常の用例の困難度より，学習者にとって意外と思われる用例の困難度の方が，語彙テスト結果データからも学習者にとって回答が難しいと判定された語ということになる．今回は設問数が少ないので，図 1 の結果が偶然得られた可能性がどの程度あるか検証するため，横軸の値の列と縦軸の値の列で統計的検定を行った．Wilcoxon 検定の結果，縦軸の値の列が統計的有意に横軸の値の列より大きかった（$p < 0.01$）．すなわち，縦軸の設問群の方が横軸の設問群より難しかった事が示唆される．

識別力の比較 識別力についても，図 1 と同様にプロットし，図 3 に示した．識別力は，直観的には，高いほど，その問題で（他の問題で推定される）能力値が高い学習者と低い学習者を分けることができるという意味で，良問である度合いを表す．学習者にとって意外と思われる用例は，能力値が高い学習者でも知らないことがあり，低い学習者でも知って

3） https://github.com/17zuoye/pyirt

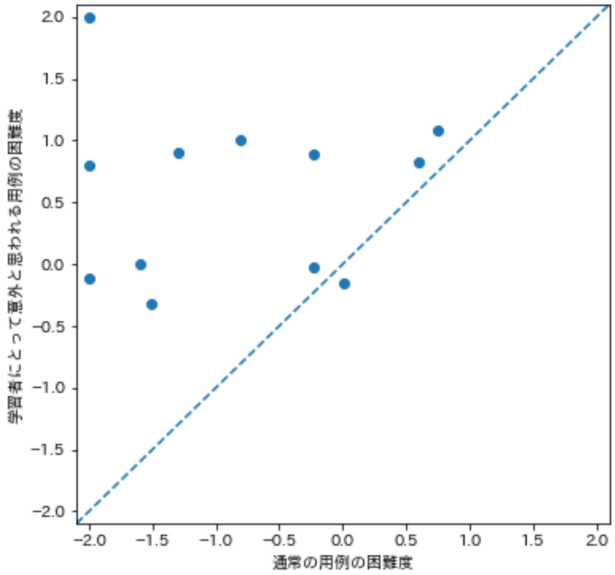

図 1　各語の，通常の用例の困難度（横軸）と学習者にとって意外と思われる用例の困難度（縦軸）のプロット．各点は各語を表す．

いることがあるため，通常の用例よりも識別力が低いと予想される．全ての語について，通常の用例の方が，意外と思われる用例よりも識別力が高いと推定されている．この結果も，Wilcoxon 検定の結果，統計的に有意であった ($p < 0.01$)．識別力のプロットについては紙面の都合のため付録に記す．

困難度と識別力のプロット　識別力は，直観的には，他の設問で能力値が高いと推定された学習者が簡単な問題に誤答してしまう，また，他の設問で能力値が低いと推定された学習者が難しい問題に正答してしまう場合に低下する．今回の設定では，前者のケースはあまり見られないが，後者のケースで回答が分からない学習者がとりあえず選んだ選択肢に正答してしまう事はあるので，困難度の高い説問ほど，識別力が低く出ることが予想される．困難度の高い問題の識別力を向上させる 1 つの方法としては，選択肢に「わからない」や未回答を許すという方法が考えられる．しかし，クラウドソーシング上でこの方法を取ると，ほぼ全ての設問に対して「わからない」と回答するケースなどがあるため，今回はこの方法は取らなかった．全ての語についての困難度パラメタと識別力パラメタのプロットを示した（図 2）．図 2 から，困難度が増加するにつれて，識別力が減少していく傾向が見て取れる．困難度パラメタと識別力の間の困難度パラメタと識別力パラメタの間のスピアマンの順位相関係数は-0.739（$p < 0.01$）で，「強い相関」が認められた．

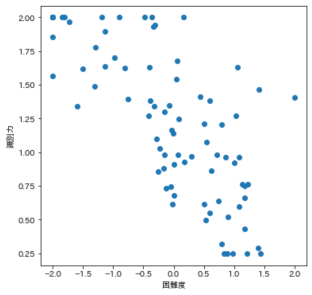

図 2　全語の困難度（横軸）と識別力（縦軸）の関係．

被験者反応予測による評価　現状のように，語の意外と思われる用例の難しさを通常の用例の難しさで代替してしまうと，被験者が設問に正答/誤答するかにどの程度の悪影響があるのだろうか？これを調べるために，次の実験を行った．まず，235 人の被験者を 135 人と 100 人に分ける．意外と思われる用例の設問群（12 問）のパラメタについては前者の 135 人の被験者反応だけから，通常の用例の設問群（70 問）のパラメタについては 235 人全員の被験者反応で推定する．この推定の際には，後者の 100 人 × 12 問，計 1,200 件の回答データは用いていないことに注意されたい．(1) より，推定された被験者の能力値 θ_j, 用例の困難度 d_i を用い，$\theta_j > d_i$ であれば被験者 j が設問 i に正答，そうでなければ誤答と判定できる．設問 i の困難度パラメタとして，意外と思われる用例の 12 問の困難度パラメタを直接用いた場合と，対応する語の通常の用例の困難度パラメタで代替した場合で，この 1,200 件の回答データの予測精度を比較した．その結果，直接用いた場合の予測精度は 64.4%，通常の用例の困難度で代替した場合は 54.4% と，10 ポイントの差が出た．この差は，Wilcoxon 検定で $p < 0.01$ で有意であった．

以上から，被験者反応の予測における，語の用例ごとに困難度を推定することの重要性がわかる．

6　結論

本研究では，外国語語彙学習において，語の通常（典型的）な用例と学習者にとって意外と思われる用例の外国語学習者にとっての難しさの差を，被験者反応から測定可能なデータセットを提案した．このデータセットにより，BERT などの転移学習技術を用いて，文脈に埋め込まれた個々の語の難しさを推定できるモデルを評価する事が可能になった．こうしたモデルの提案が，今後の課題となる．

謝辞

本研究は，科学技術振興機構 ACT-X 研究費 (JPMJAX2006)，ならびに日本学術振興会科学技術研究費補助金（18K18118）の支援を受けています．また，産業技術総合研究所の AI 橋渡しクラウド (ABCI) を使用しています．本データセットの設問が問題として成立しているか確認していただいた，谷口ジョイ先生をはじめとする静岡理工科大学の先生方に深く感謝いたします．

参考文献

[1] I. Nation. How Large a Vocabulary is Needed For Reading and Listening? **Canadian Modern Language Review**, Vol. 63, No. 1, pp. 59–82, October 2006.

[2] David Beglar and Paul Nation. A vocabulary size test. **The Language Teacher**, Vol. 31, No. 7, pp. 9–13, 2007.

[3] Frank B. Baker. **Item Response Theory : Parameter Estimation Techniques, Second Edition**. CRC Press, July 2004.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, 2019.

[5] Burr Settles. Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), 2018.

[6] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. A report on the complex word identification shared task 2018. In **Proc. of BEA**, June 2018.

[7] Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In **Proc. of LREC**, May 2018.

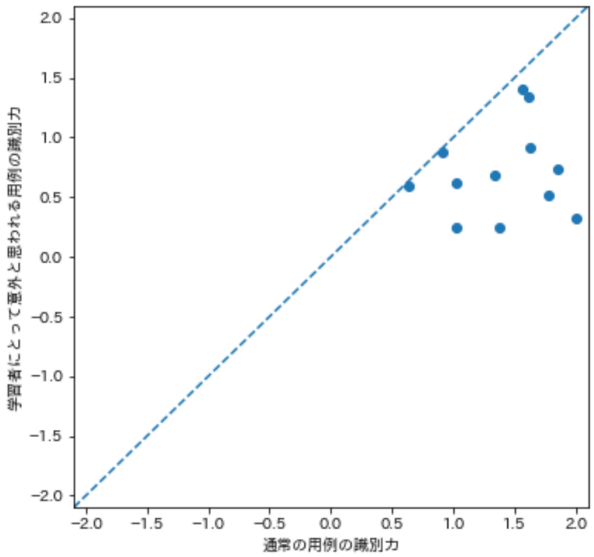

図 3　各語の，通常の用例の識別力（横軸）と学習者にとって意外と思われる用例の識別力（縦軸）のプロット．各点は各語を表す．

A　付録 (Appendix)

付録には，本論に直接関係ないが，参考になる図を示す．図 3 に識別力のプロットを示す．