CN-004

# 外国語多読学習支援のための教材推薦と教材生成の融合
# Combining Text Recommendation and Generation for Extensive Reading

江原 遥[1)]
Yo Ehara

## 1 はじめに

外国語多読学習支援のために，教材中のテキスト情報から教材内容を考慮して教材の難度等を推定し，学習者に合った個別最適な教材を推薦したい．学習支援システム等のために，こうした教材推薦を行うには，学習者が項目に回答できるかを予測し，学習者に適応した水準の項目（設問，教材）を選び出すタスクが基本となる．学習者が項目に回答した履歴のデータがあれば，教育心理学などで能力や難しさのモデル化に多用される項目反応理論 (Item Response Theory，以下 IRT) [1] を用いて，学習者の能力と項目の難しさを推定し，学習者の反応予測を行う事ができる．IRT に基づくモデルは通常，学習者の回答パターンにのみ依存し，項目（設問）が自然文で書かれていても文意を理解しない．自然言語処理においては，近年，Transformer モデルに代表される大規模言語モデルが自然文理解で高い性能を示している [2]．従って，設問文の理解に，これらの大規模言語モデルを用いたい．しかし，これらの言語モデルは，通常，言語を入力とするため，学習者ごとに異なった判定を行うことができず，学習者に適応的な予測を行うことが難しい．大規模言語モデルを用いて設問文の文意を考慮した学習者適応を行うには工夫を要する．

本研究では，まず，教材推薦のために，大規模言語モデルを用いて，設問文を考慮して学習者反応を予測（シミュレート）する簡便な方法を提案する．次に，複数の学習者のシミュレーション結果を統合して，大規模言語モデルによる予測モデルを通じ，ある学習者集団が個々の設問を解いた場合に，正答者数分布を求める手法を提案する．これにより，言語モデルからの設問の難しさ等，教育上，設問を解釈する上で重要な値が取得できる．学習者が設問に正答できる確率が求めれば，例えば正答確率が 50%を少し超えた難しさの問題，すなわち，学習者がかろうじて解ける水準の問題を教材として推薦する方法が，教育効果の高い教材推薦の方法として考えられる．「かろうじて解ける水準の問題が教育効果が高い」という考え方は，例えば語彙学習であれば，“Krashen's i+1”[3] 等の法則で，広く知られている．

さらに，本研究では，学習者に適合する適切な難しさの教材がない場合に，教材自体を言語モデルによって生成してしまう事で，教材推薦と教材生成を融合する事を目指す．例えば，適切な難度の多読教材がない場合には，ChatGPT（`https://openai.com/blog/chatgpt`）をはじめとする外部の大規模言語モデル (Large Language Models, LLMs) を用いて，その場で多読用のテキストを生成してしまうことができる．そこで，学習者反応予測を用いて，学習者が知らないと判定された単語を含む教材を生成する実験を，合わせて行った．

本研究の貢献は，教材推薦と教材生成を融合させた個別最適な学習者支援システムの例とその有効性を，外国語の多読学習支援を題材に，実データを用いた実験を通じて示したことである．

本稿の構成は次のとおりである．まずは，本節以降，語学学習者の語彙テストの結果から，現在では標準的な，BERT（Bidirectional Encoder Representations from Transformers[2]）などの Transformer モデルに基づく，文単位で学習者が所与の文を可読かどうか判定する識別器の構成方法を説明する．これにより，内容を考慮した文単位での個別可読性予測が可能である事や，その精度を簡単に示す．文単位の可読性予測ができれば，少し難しめの文を教材として与えることができる．

ただし，外国語多読学習において，どのような教材を推薦すればどのような教育効果が得られるのかを記録したデータセットは，知る限り，存在しない．が，応用言語学の分野では，どのような学習者にどのようなテキストを推薦すると，どの程度の新しい語彙が獲得されるかということについて，シミュレーション実験を行う方法が存在する．この実験について 8 節で述べる．

また，現在では，多読教材の中に存在しないテキストについては，ChatGPT 等の生成 AI を用いることで生成できる．指定したトピックで，指定した単語を必ず含む多読教材を生成するように指示することで，教材生成を行うシミュレーション実験を 9 節に示す．

文単位の可読性の評価用データセットについて，3 節に述べる．学習支援システムのために，典型的な語義の知識状態から，非典型的な（意外な）語義の知識状態を予測する課題についての評価用データセットを作成した．1 つの語について，典型的な語義で使われている文と意外な語義で使われている文を用意・作問し，クラウドソーシング上でデータ収集を行った（表 1，表 2）．設問は，複数の英語母語話者の確認の取れたものを用いた．IRT を用いて典型的/意外での設問の困難度等の分析を行い，学習者反応データ上でも，意外な語義の方が典型的な語義より難しい事を示す．作成したデータセット上で，典型的な語義のテスト反応から意外な語義への反応をどの程度予測できるか評価する（5 節）．

学習者反応の予測では，大別して 2 種類の手法を比較した．まず，教育心理学などで能力や難しさのモデル化に多用される，設問文の文脈を考慮しない，前述の IRT [1] を用いた手法である．次に，大規模な母語話者コーパスを事前学習に用いることで設問文の文脈を考慮する事ができる Transformer モデルの手法（BERT など）に基づく提案手法である．前述のように，Transformer モデルは，能力の考慮など，学習者によって異なる結果を予測する仕組みを通常持たない．本研究では，Transformer モデルを学習者反応予測問題に適用する手法をあわせて提案し，その予測性能が IRT による手法より高いことを意味する．また，IRT の利点は学習者の能力値等を合わせて推定できる解釈性にあるが，Transformer モデルから IRT で推定した能力値とよく相関する値を抽出する手

1） 東京学芸大学，Tokyo Gakugei University.

法も提案する．本研究の内容は，教育データマイニングのトップ国際会議である Educational Data Mining 2022 の short paper, poster paper に採択された（査読付き）ものの発展である [4, 5]．また，9 節の関連論文として，AIED 2023 の Innovation & Industry Track に採択された論文が挙げられる．

## 2　関連研究

### 2.1　外国語学習支援の学習者反応データセット

本研究では，設問文を考慮した学習者反応予測を行いたい．そのためには，設問文の文意を考慮することが重要であるような設定で試験を行い，その結果を記録したデータセットが必要となる．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 [6] 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 [7] 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，各学習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，学習者を行，設問を列とし，学習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，学習者の設問に対する回答内容から，設問の難しさや学習者の能力値を推定を目標とするが，この推定のためには，各学習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，文中の語に対する学習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，[7] 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，学習者が難しいと感じた語が記録されているだけであり，学習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．

### 2.2　提案手法の関連研究

教育 AI などの分野で本研究のタスク設定に似たタスクとして，Knowledge Tracing[8] が挙げられる．Knowledge Tracing のタスクは，本研究とは目的が異なり，学習支援システムのログなどを対象にしたモデルであり，学習者が設問を解いた順序や時間の情報が与えられている設定であることが特徴である．この状況で，学習者の知識や能力の成長をモデル化するものである．そのため，Knowledge Tracing は，本研究とは逆に，設問に関しては人手のメタ情報（どの設問がどのようなスキルを試験するものであるのかなど）が与えられている設定であることも多い．通常，質がよくわからない設問は，学習支援システムに追加されないからである．また，このように，設問に関して潤沢な情報が得られる事が想定されているため，設問のテキスト情報は，殆どの場合利用されず，テキスト情報を人間が解釈したメタ情報を利用する設定になっている．設問のテキスト情報を利用する手法は，設問間関係を考慮する [9] など僅かである．近年にも，BERT を用いた教育応用が提案されているが [10, 11, 12]，これらの研究では個人化学習者支援については扱われていない．また，近年では，Knowledge Tracing を拡張し，正解不正解だけではなく，プログラムのコードなどを学習者が回答する場合のモデルが提案されている [13]．しかし，これも時間情報と紐づいた設定で，長期に収集されたデータを対象にしたものである．

一方，本研究の場合は，時間情報と紐づいておらず，１回の試験のデータでも適用することができる．そのため，システムを使い始めたばかりのユーザに対しても，教材推薦を行うことができる．

また，著者の知る限り，応用言語学の分野においても，外国語学習者を対象に，ある語の意外な意味／典型的な意味の 2 種類を同時に試験したデータセットを作成し，項目反応理論を用いて各意味の難しさを試験結果データから客観的に推定・分析した研究は見当たらない [14, 15, 16].

**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 [17]. データセットはクラウドソーシングサービス Lancers[1)] 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 名の学習者（被験者）から回答があった．以後，用語の統一のため，被験者という語は用いず，学習者という語を用いる．Lancers の作業者は大部分日本語母語話者であるため，このデータセット中の学習者の母語は，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 [17] と同様に，Vocabulary Size Test（VST）[18] を用いた．ただし，VST は 100 問からなるのに対して，[17] 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，学習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，学習者は，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．

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．学習者は，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

1) 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

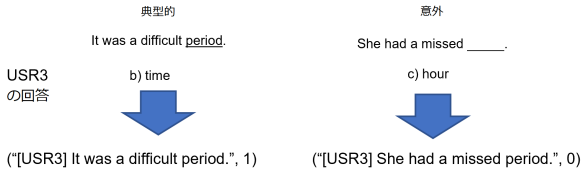


図 1　学習者トークンの導入．

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

項目反応理論のモデルについて，説明する．学習者の数を $J$，設問（項目, item）の数を $I$ とする．簡単のため，学習者の添字（index）と学習者，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，学習者 $j$ 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，2 パラメター—モデル (2PLM) では，学習者 $j$ が項目 $i$ に正答する確率を次の式でモデル化する．

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

ここで，$\sigma$ 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．$\sigma$ 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．式 1 において，$\theta_j$ は能力パラメタ（ability parameter）と呼ばれ，学習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．式 1 より，$\theta_j$ が $d_i$ を上回る時，学習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，$\theta_j - d_i$ が正答確率と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，学習者 $j$ 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．識別力が高い設問は，直観的には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点で「ひっかけ問題」のような性質を持たず，性質の良い設問であることを示している．

なお，項目反応理論には，多肢選択式の設問で分からなくても選択肢を無作為に選んで正答出来てしまう確率を考慮する 3 パラメータモデル（3PLM）が存在するものの，今回のデータセットの被験者数（学習者数）では，被験者数が少なすぎてパラメタ推定が不安定であるという報告 [19] があるため，よりパラメタ数の少ない 2PL モデルを用いた．

## 5　学習者反応予測による評価

### 5.1　IRT による学習者反応予測

語の意外と思われる語義の難しさを典型的な語義の難しさで代替してしまうと，学習者が設問に正答/誤答するかを IRT で予測する際，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 人の学習者を 135 人と 100 人に分ける（図 2）．意外と思われる語義の設問群（12 問）のパラメタについては前者の 135 人の学習者反応だけから，典型的な語義の設問群（70 問）のパラメタについては 235 人全員の学習者反応で推定する．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．式 1 より，推定された学習者の能力値 $\theta_j$, 語義の困難度 $d_i$ を用い，$\theta_j > d_i$ であれば学習者 $j$ 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，意外と思われる語義の 12 問の困難度パラメタを直接用いた場合と，対応する語の典型的な語義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．予測精度 (accuracy) の結果を表 3 に記す．その結果，直接用いた場合の予測精度は 64.4%，典型的な語義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．この結果から，学習者反応の予測における，語の語義ごとに困難度を推定することの重要性がわかる．より直接的に言い換えれば，この結果は，語の意外な用例の難しさを，語の典型的な用例の難しさで置き換えると，学習者反応予測の精度が著しく低下することを示唆している．

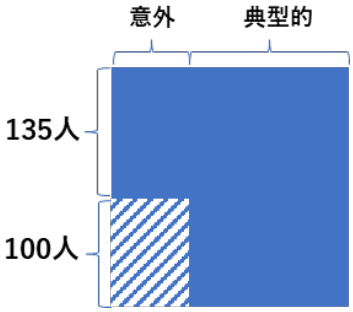


図 2　実験設定．青く塗られた部分がパラメタ推定に使われる訓練データ．斜線部が性能比較に用いられるテストデータ．

表 3　図 2 斜線部の予測精度 (accuracy)．

| 手法 | 精度 |
|---|---|
| IRT (能力 - 235 人から推定した典型的な語義の困難度) | 0.544 |
| IRT (能力 - 135 人から推定した意外な語義の困難度) | 0.644 |
| 提案手法 (bert-large-cased) | 0.674 (**) |
| 提案手法 (bert-base-cased) | **0.688** (**) |
| 提案手法 (bert-base-uncased) | 0.655 |
| 提案手法 (roberta-base) | 0.681 (**) |
| 提案手法 (albert-base-cased) | 0.671 (*) |

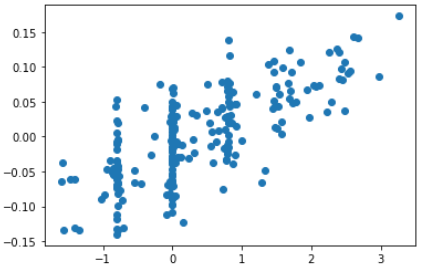


図 3　IRT の能力パラメタ（横軸，pyirt によって算出）と，学習者トークンの単語埋め込みベクトルの第一主成分得点（縦軸）．

### 5.2 Transformer モデルと IRT の性能比較

IRT を用いた手法は，学習者反応のみに依存し，設問文の意味などは全く考慮されていない．では，設問文の意味をも考慮した学習者反応予測を行うと，学習者反応のみを用いた IRT の手法より高精度に予測できるのだろうか？大規模言語モデルのうち，自然言語処理で文意を考慮した予測手法として近年多用される，Bidirectional Encoder Representations from Transformers (BERT)[2] に代表される Transformer モデルと IRT の予測性能を比較した．

Transformer モデルは近年の深層転移学習による大規模言語モデルの代表的な手法であり，大量のラベルなしデータからの事前学習 (pre-training) と，ラベル付きデータを用いた微調整 (fine-tuning) という 2 種類の学習からなる．事前学習では，大量のラベルなしコーパスを用いて，当該言語の基本的な構造を学習し，入力文の言語としての自然さを計算可能にする．この過程は計算量が非常に大きいが，様々なタスクに対して汎用的に用いることができる．そこで，通常，事前学習は，**bert-large-cased** 等の，英語版 Wikipedia 等を用いて訓練された **transformers**（`https://github.com/huggingface/transformers/`）の**事前学習済モデル**を用いる．事前学習済モデルの詳細情報，例えば事前学習に用いたコーパスなどの情報は `https://huggingface.co/models` に記載されている．多くのモデルは英語版 Wikipedia を使用している．

後段の微調整 (fine-tuning) では，実際に，目的とするタスクに合わせて，事前学習済モデルを追加訓練する．本研究のタスクにおいては，ラベルは，IRT 同様，学習者が正答する場合 1，誤答する場合を 0 とする 2 値判別問題である．事前学習済モデルに設問文と学習者の両方を入力し，微調整を行いたいが，通常，大規模言語モデルの微調整では言語しか入力として扱えないため，学習者の情報を入力することができない．そこで，後述の方法で，この問題を解決する．

### 5.3 Transformer モデル上の個人化判別

Transformer モデルを個人化判別に対応させる手法は，自然言語処理の言語教育応用の目的では著者の知る限り知られていない．ただし，Transformer モデルに特殊なトークン（語）を加えて微調整を行い，様々な問題設定に対応させる手法は知られており，ライブラリ上で特殊なトークンを加える機能が用意されている．本研究では，この機能を利用することで，学習者に対応するトークン（**学習者トークン**）を作り，これを入力系列の最初に置くことによって判別を行う手法を提案する（図 1）．例えば，学習者 ID が 3 番の学習者を表すトークン “[USR3]” を導入し，“[USR3] It was a difficult period.” が入力であれば，3 番の学習者が “It was a difficult period.” という文から成る設問に正答するか否かを予測する問題に帰着させる．入力文はそのままで，入力文の前に，単純に学習者トークンが挿入されている点に注意されたい．導入するトークン数は学習者数と同数である．Transformer では各トークンに対して，その語としての機能を表現する**単語埋め込みベクトル**があるので，学習者トークンに対しても埋め込みベクトルが作られる．

### 5.4 文単位の可読性判別器

**重要な点として，提案手法では，文中のどの語についての設問であるかという情報や，誤答選択肢の情報は与えていない**．すなわち，提案手法の判別器は，表 1 のどの単語に下線が引かれているかや，表 1 や表 2 の正答以外の選択肢の情報を用いない．提案手法は，単純に正解となる文を入力とし，これを学習者が理解できるか否かを判別する判別器を構成している，と解釈できる．すなわち，**提案手法は，文単位で，学習者個別の可読性判別器を構成していることに相当する**．これにより，提案手法は，表 1 と表 2 という仔細の異なる 2 種類の多肢選択式の問題に対応できる．このように，提案手法の適用範囲を広くとることができる．応用言語学では，テキスト中の 95%以上の語を知らないとテキストが理解できない事が知られている [14]．文単位の場合，学習者が文中の 1 語でもわからなければ正答できないケースが大半であるため，文中の全ての語の語義を知っている事と正解となる文を理解できるかは，多くの場合，同一視できる．

Transformer モデルのその他の実験設定は，多用される設定を用いた．判別には **transformers** ライブラリの **AutoModelForSequenceClassification** を用いた．微調整には Adam 法 [20] を用い，バッチサイズは 32 とした．

Transformer モデルを用いた結果を，表 3 に示す．*は IRT の最高性能と比較して Wilcoxon 検定で統計的有意であることを表し，**は $p < 0.01$，*は $p < 0.05$ を表す．また提案手法の（）内は用いた事前学習済モデル名である．表 3 では，まず，学習者トークンを導入した提案手法が，IRT を用いた従来手法より高い性能を達成していることが分かる．この実験結果は，設問文の意味を考慮する事で，IRT より高精度な判別が行えることを示している．

次に，“roberta-base” は cased（大文字・小文字を区別するモデル）であるのに対し，“albert-base-v2” は uncased（大文字・小文字を区別しないモデル）である．この結果から，良い精度を得るためには “cased”，すなわち，大文字と小文字を区別して扱うモデルでなければならないことが示唆される．この理由は，次のように推察される．この実験環境では，各質問は短い文から構成されているため，モデルは大文字で始まる文の開始を認識する必要があるためであろう．

さらに，表 3 では，**bert-base-cased** が最も高い性能を示した．より大きな事前学習済モデルである **bert-large-cased** よりも **bert-base-cased** が高い性能を示した理由として次のことが考えられる．学習者特性を表す学習者トークンの単語埋め込みベクトルは，今回作成した比較的小さい訓練データで訓練しているため，小さいモデルの方が微調整（fine-tuning）に適していた可能性がある．

## 6 解釈性―学習者トークンからの能力値抽出

IRT は，学習者の能力パラメタを持つことにより，学習者の特性について解釈しやすい．一方，Transformer モデルでは，学習者の特性は学習者トークンに対する単語埋め込みベクトルという多次元の形で表現されており，そのままでは直感的な解釈が難しい．しかし，Transformer モデルは個人化判別問題で高精度を達成しているので，学習者トークンの単語埋め込みベクトルの中に能力値の情報が含まれていると考えられる．

微調整後の **bert-large-cased** の場合の学習者トークンに対する単語埋め込みベクトルのみを集めた．すなわち，学習者の人数分の単語埋め込みベクトルの集合がある．このベクトル集合に対して主成分分析を行い，その第一主成分得点と IRT の能力値パラメタを比較した

（図 3）．各点は学習者を表す．IRT の能力値パラメタの算出には，Python の **pyirt** ライブラリを用いた．両者は相関係数 0.72 という強い相関を示した（$p < 0.01$）．これにより，提案手法を用いた場合でも，能力値は学習者トークンの第一主成分得点として容易に抽出できることが分かった．これにより，提案手法は文意を考慮することにより IRT より高い精度を達成しながら，IRT と同様に「能力値を取り出せる」という高い解釈性を持つことが示された．

図 3 では，縦に筋が入っているように見える部分がある．これは，pyirt の内部で使われている IRT のパラメタ推定アルゴリズムの性質で，横軸の学習者の能力値パラメタの推定の際，能力に大きな差がない能力値パラメタは 1 つの値にまとめられる性質があるため，横軸が同じ値を取る学習者が存在するためである．確認のため，同じデータを，pyirt とはプログラミング言語も異なる全く独立の実装である R 言語の **ltm** パッケージも用いて推定した．これは，教育心理学の標準的な教科書で使用されているソフトウェアである [21]．相関係数は 0.72 で，やはり統計的有意性を示した ($p < 0.01$)．

また，第二以降の主成分得点には能力値との統計的有意な相関は見られなかった．

## 7　設問の難しさや識別力の抽出法

ここまでは微調整済の BERT モデルから学習者の能力値を抽出する方法であったが，さらに，設問の難しさや識別力に相当する値を抽出する方法を提案する．方法の概略を示す．BERT は被験者が設問文が指定されれば，その被験者がその設問に正答できるかどうかだけでなく，その確率値も予測として出力できる．ある設問に着目し，全被験者がその設問を解いた時の正答できる確率を BERT に出力させ，ここからその設問の正答者数の確率分布を計算する．被験者間の独立性を仮定すると，数学的には，成功確率が互いに異なる独立なベルヌーイ分布の和の分布であるポアソン 2 項分布を計算する事に相当する．この時，その設問の正答者数の確率分布の平均を設問の難易度，分散を識別力のような設問の良さと解釈する事が可能になる．

被験者数を $N$ 人とし，学習者の添字を $n$ とする（被験者の中から特定の被験者を選び $N$ と $J$ が異なる設定もあり得るので，違う文字を用いた）．項目数を $I$ 個とし，項目の添字を $i$ とする．学習データ上で予測器を微調整した後，予測器は学習者 $n$ が項目 $i$ に正しく回答する確率を出力することができる．この確率を $BERTProb(n,i)$ と表記する．簡単のために，ここからは設問 $i$ に焦点を当てる．$BERTProb(n,i)$ を使って，N 人のうち，質問 $i$ に正答する者の確率分布を求めたい．そこで $BERTProb(n,i)$ の確率で 1，そうでなければ 0 となるベルヌーイ分布に従う確率変数 $A_n$ を $A_n \sim Bernoulli(BERTProb(n,i))$ と定義する．ここで，簡単のため，これらの確率変数 $\{A_1, \ldots, A_n\}$ は互いに独立であると仮定する．学習者について和をとり，項目 $i$ の全 $N$ 人の中での正答者数の確率分布は次のように書ける．

$$A_i = \sum_{n=1}^{N} A_n \tag{2}$$

式 2 は互いに独立なベルヌーイ分布の和であり，ポアソン 2 項分布と呼ばれる[2)]．この分布の計算は，動的計画法を用いて計算可能である．この動的計画法の詳細は [22, 5] を参照されたい．

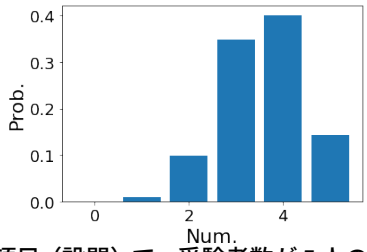


**図 4　ある項目（設問）で，受験者数が 5 人のときに予測される正答者数の分布．**

$A_i$ は確率分布なので，平均と分散を計算できる．$A_i$ は，全 $N$ 人のうち，項目 $i$ の正答者数である事に注意すると，$A_i$ の平均は，問題 $i$ の難易度を表していると解釈できる．また，$A_i$ の分散は，問題 $i$ の正解者数を予測のためのエラーバーと解釈できる．同じような難しさの設問の中では，分散が最も小さい，つまり，正答者数の予測がつきやすい問題と考えられる．$A_i$ の分散は，項目反応理論における「識別力」に似た性質を持つ指標である．項目反応理論の識別力は，項目が能力の高い被験者と低い被験者を識別する力を表す．直感的には，能力が本当は高い被験者が間違えてしまうような確率の少ない，ひっかけ問題でない度合いを表しているとも考えられる．$A_i$ の分散は，直感的には，項目反応理論の識別力の大小関係を逆にした「ひっかけ問題である度合」と解釈できるが，項目反応理論はモデルが固定されているのに対し，$A_i$ の分散は予測器 $BERTProb(n,i)$ の確率値さえわかればどのような予測器を用いても計算できるので，深層転移学習のような複雑な手法を用いた場合でも計算できる．

横軸に $A_i$ の分散，縦軸に $A_i$ の平均値をとることでリスク・リターンプロットを作成できる．まず，どの程度の難しさの設問を選びたいかを決めて縦軸の値に注目し，次に同程度の難しさの問題の中で横軸の値が最も小さいもの（最も左にあるもの）を選ぶことで，特定の難易度の問題のうち，学習者反応を予測しやすい性質の良い問題を選択可能である．この図の最も左にあるリスクが小さい点を結んだ線を「効率的フロンティア」といい，経済分野で多用される概念である [5]．

図 4 と図 5 に，ある項目（設問）$i$ について，受験者数がそれぞれ 5 人，100 人である場合の分布を描いた．（5 人については，ランダムに受験者を選んだ）図 5 には，難しい項目と簡単な項目の 2 つを選んで正答者数の分布を描いて重ねたグラフを示した．実際の正答者数は，図 5 の左側が 31 人，右側が 56 人であった．図 4 から，受験者数が少ないときでも，非対称な分布の形が計算できている事が分かる．また，図 6 に，リスク-リターンプロットを描いた．各点は前述の 12 問の設問であり，破線は効率的フロンティアである．全 12 問のうち，効率的フロンティア上の問題を選ぶことで，5 問の性質の良い問題をさらに選び出せている事が分かる．

## 8　多読学習のための教材推薦シミュレーション

外国語のテキストの多読によって外国語の語を学習することを「付随的学習」と良い，単語帳などで語を覚え

2) https://en.wikipedia.org/wiki/Poisson_binomial_distribution

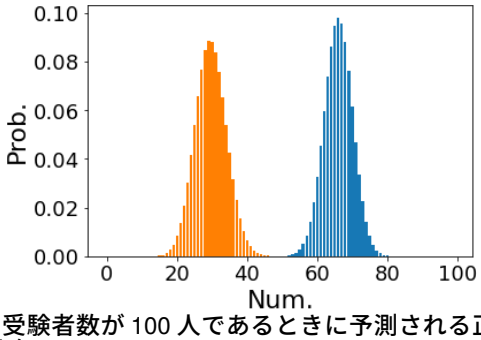


**図 5　受験者数が 100 人であるときに予測される正答者数の分布．**

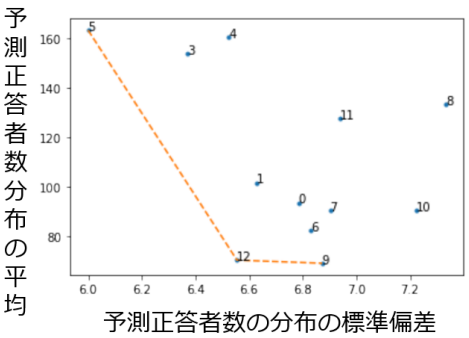


**図 6　各設問について，学習者の総数が 100 である場合のリスク（横軸，各設問の予測される正解者数の分布の分散）とリターン（縦軸，各設問の予測される正解者数の分布の平均）をプロットしたものである．各点は設問を表し，各点の番号は設問番号である．縦軸が同程度の値であれば，分散が小さい設問（図中左側の設問）が識別力が高い可能性が高いとされた．**

る「意図的学習」と対比される．応用言語学では，多読学習を通じて語彙習得について，１）テキストに含まれる語のうち述べ語数で 95%以上を知らないとテキストを読めないこと，２）テキストが読めれば知らない 5%の語が付随的学習で獲得される可能性が高いことが知られている [14]．そこで，この結果を用いて，各学習者が各テキストを読んだときに獲得される語彙量を計算し，教材推薦を行うシミュレーション実験をした．語彙テスト結果については，著者がクラウドソーシングを用いて過去に公開したデータがある [17]．具体的には，クラウドソーシング上の学習者（TOEIC の受験経験があるものに限定）100 人に，100 問からなる語彙サイズ計測用の単語テスト Vocabulary Size Test (VST)[18] を受けてもらった結果のデータセットである．VST は多肢選択型のテストであり，**英文中に埋め込まれた**単語の言い換えとして適切な選択肢を 4 つの選択肢から選択するテストである．

単語テストの結果を使って，各学習者が所与の単語を知っているかどうかを判別する確率的識別器を作成し，この確率値を用いて分布を描いた．識別器は，この実験では単純なロジスティック回帰を用いて構成した．特徴量としては，COCA コーパスの頻度，British National Corpus (BNC) コーパスの頻度を用いた．ただし，頻度は − log( 頻度 ) の形に直して特徴量として用いた．テキストとしては，Brown Corpus を用いた．テキスト長さが実験結果に影響しないように，Brown Corpus の 500 件の各テキストのうち，先頭から 300 語を切り出し，実験に用いた．この 500 件のテキストから，ある学習者の付随的学習に適したテキストを選択することが，我々の目的である．付随的学習はテキストを読める必要があるので，

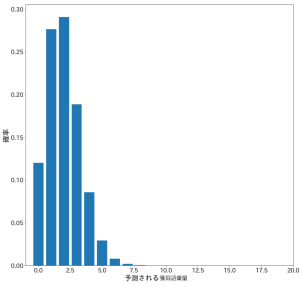

**図 7　予測される獲得語彙量の分布の例．**

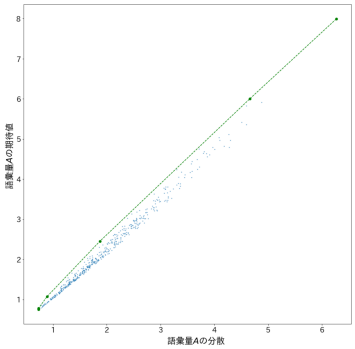

**図 8　各テキストの獲得語彙量分布の期待値と分散．**

好成績の学習者に起こりやすい．まずは，最も成績の良かった学習者（VST で 96 問正解）を対象に実験を行った．図 7 にこの学習者があるテキストを読んだ時の獲得語彙量の分布を 1 つ示す．図 7 より，獲得語彙量の分布には幅があり，単純に期待値が高いテキストを選べばよいわけではないことがわかる．

各テキストを読む際に予測される獲得語彙量の期待値と分散を同時に考慮するため図 8 に図示した．ここでの平均・分散は式 2 と同様，ポアソン二項分布を用いれば良い．各点は Brown Corpus の各テキストである．獲得語彙量を学習者にとっての利得と考えると，獲得語彙量の期待値が同じであれば，できるだけ獲得語彙量の分散が少ないテキストを選択する方が，確実に語彙を増やせるので，学習者にとっては得となる．すなわち，図 8 の左上部分が，この学習者にとって最も効率的に付随的学習を行える文書群である．このように，図 8 の縦軸は利得，横軸の獲得語彙量の分散は**リスク**とみなせ，図 8 も，リスクとリターンの関係図とみなせる．

図 8 のようなリスクとリターンの関係図においては，左上部分が最も低リスクで利得を増やせる選択であり，この部分が前述の**効率的フロンティア**とみなせる．図 8 では，500 件あった選択肢の中から，凸包を用いて効率的フロンティアに属するテキスト 5 件が選択された．すなわち，学習者の付随的学習に適したテキストを 1/100 に絞ることができた．この 5 件の中でどのテキストを選択するかは，学習者がどの程度のリスクをとって語彙量を増やしたいかに依存するため，学習者に任せる方法が一案である．

学習者によって効率的フロンティア曲線は変わるが，多くの学習者に薦められるテキストはあるのだろうか？ 100 人のうち，成績の良い 30 人に対して，同様に各学習者が各テキストを読んだ場合に予測される獲得語彙量の分布から効率的フロンティアを求めた．その結果，10 人以上の効率的フロンティアに選ばれ

たテキストが 7 件あり，最多では 14 人の効率的フロンティアに含まれたテキストがあった．このテキストは，具体的には “In the imagination of the nineteenth century the Greek tragedians and Shakespeare stand side by side , their affinity transcending all the immense contrarieties of historical circumstance , religious belief” で始まり，“contrarieties” という難しい語が一部使われている他は，“belief” など中級学習者なら習得している語で構成されている．この結果から，どの学習者の効率的フロンティアにも比較的安定して含まれるテキストが存在することが分かる．

## 9 教材生成

図 9 は，指定したトピックで，指定した語を含む多読教材の生成システムの画面例を示す．テキストエリアには，ChatGPT で生成されたテキストが表示される．テキストエリアの下にある「Science」という文字列は，学習者が「Science」というトピックを選択したことを示す．この単語をクリックすると，学習者は他のトピックを選択してトピックを変更することもできる．テキストエリアの下には，学習者が読めないと予測される 3 つの単語，“nuance”, “bizarre”, “canary” がリストアップされている．ChatGPT は，学習者のために読むべきテキストを生成した．ChatGPT は，3 つの単語をすべてテキストに含めるよう指示される．また，この 3 つの単語以外の単語は，指定されたトピックの範囲内でできるだけ簡単な語を用いるように ChatGPT に指示する．テキストエリアにフォーカスを移動させると，学習者が知らないと予測される単語が自動的にハイライトされる．学習者はテキストエリア内の単語をクリックすることで，学習者が知らない単語を示すことができる．このクリック情報をもとに，学習者が知らない単語を予測することができる．

このシステムの利用を想定したシミュレーション実験を行った．英語学習者の語彙知識データセットとして，[23] を用いた．このデータセットは，15 人の英語学習者が約 12,000 語の各単語を知っている度合いを 5 段階で自己申告式でアノテーションしたものである．新聞記事などに現れる語は，ほぼ網羅されている．5 段階のうち，語を知っている度合いが最も高い段階のみを学習者が語を知っている場合として，学習者が各単語を知っているかを判別する 2 値判別器を構成した．ただし，[23] のデータセットには学習者が 15 名しかいないため，[23] のデータセットと 43 語が共通する [17] の 100 名のデータセットも合わせて訓練データに用いた．語彙知識の推定には，British National Corpus [24] と Contemporary Corpus of American English [25] の両コーパスにおける単語の対数頻度を特徴量とするロジスティック回帰モデルを用いた．両コーパスは様々なトピックを扱っており，言語能力の評価に利用されている．判別器の構築のその他の設定は，[5] に従った．[5] は，本研究とは異なり，テキスト集合から多読に適したテキストの選出が目的であり，9 節のような多読のためのテキスト生成は行っていない．

判別器を作成した後，各学習者が約 12,000 語から各語を知らない確率を算出した． 学習者が知らない語と判定されたものの中から，確率が 0.5 に最も近い単語を学習対象単語として選択した．これらの単語は，学習者にとって未知の単語の中で最も頻出する可能性が高く，重要な単語であるため，最初に選択された．ChatGPT には，これらの単語を含む学習教材を生成するよう指示した．ChatGPT がテキストを生成したら，生成されたテキスト内に出現する語で，かつ，学習者が知らないと判定された語から，確率が 0.5 に最も近い単語を再び学習対象単語として選択する．ただし，同じ単語が何度も学習対象に選ばれることを防ぐため，以前の試行で選んだことのある単語は，学習対象単語に選ばない．

**表 4 10 回のテキスト生成試行における実験結果．**

| 試行回数 | 1 | 3 | 5 | 7 | 10 |
|---|---|---|---|---|---|
| 既知語の比率 | 0.937 | 0.930 | 0.924 | 0.899 | 0.888 |
| 予測精度 | 0.939 | 0.939 | 0.929 | 0.928 | 0.913 |
| 可読テキスト比率 | 0.466 | 0.600 | 0.307 | 0.363 | 0.090 |

各学習者について，この生成を 10 回繰り返した．生成後，約 12,000 語の語彙データを用いたシミュレーションにより，その読みやすさを測定した．この測定では，学習者がテキストを自然に読み，そこから学習するためには，テキストに含まれる単語の 95%以上を知っている必要があることを示した [14] の結果を利用した．約 12,000 語の語彙は，生成されたテキストに登場するほとんどの単語をカバーしている．この語彙に含まれていない語は，学習者が知らないものとした．トピックについては，学習者が科学，スポーツ，文化の 3 つのトピックから一律にランダムにトピックを選択するケースを想定した．

### 9.1 生成実験結果

多読シミュレーションの実験結果を表 4 に示す．各試行での比較のため，表 4 の既知語や可読性の数値は，各試行で選ばれた学習対象単語を考慮していない．つまり，学習者が各試行で指定された学習対象単語を一切覚えず，最初の状態のままテキストを読んだ場合の既知語や可読性を示した．1 行目は多読の試行回数，つまりシステムが学習者に合わせた文章を生成した回数を示している．表中の全ての値は，15 人の学習者の平均値である．2 行目は，テキスト中で 15 人の学習者が平均して知っている単語の比率を示している．徐々に比率が減少しているのがわかる．この結果は，以前の試行で学習対象単語に選ばれたことのある語は，再び学習対象単語にならないため，徐々に難しい語が学習対象単語に指定されているためと考えられる．3 行目は，各学習者が各語を知っているかの予測精度を示している．徐々に精度が低下していることがわかるが，その減少幅はわずかである．最後に，4 行目は，各試行で 15 名の学習者に合わせた 15 個のテキストのうち，前述の 95%の基準で学習者にとって可読なテキストの比率を示している．試行回数が増えるにつれて，値が減少している．この結果も，以前の試行で学習対象単語として選ばれた語は再び選ばれないため，段々と難しい語が学習対象単語となるためと考えられる．実際には，学習者が各回の語を着実に記憶すれば，可読性の減少はみられなくなると予想される．

## 10 おわりに

本研究では，外国語多読学習のために教材推薦・教材生成の融合のための実験を行った．まずは，BERT 等に学習者を表す単語「学習者トークン」を入れて，語彙テストデータで fine-tuning し文単位で学習者個別に可読性を計測する手法を提案した．さらに，ここから，学習者の能力値や文の難しさ・識別力を抽出する手法も示した．

次に，応用言語学の多読による語彙獲得が起こるケースの知見を利用して，獲得語彙料のシミュレーション

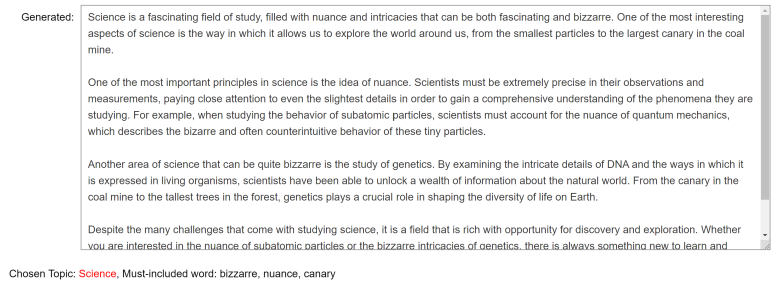


**図 9　教材生成システムの画面例．**

実験を行った．さらに，個別可読性予測器を用いて，ChatGPT で，実際に学習者が学習すべき単語を必ず含むテキストを生成させる実験を行った．実験の結果，試行回数が上がるごとに難しいテキストが生成されており，学習者が知らない語を学習しながら多読することの必要性が示唆された．

今後の課題として，学習者トークンの単語埋め込みから，学習者の能力値よりさらに高度な情報を取り出すことや，実際に推薦・生成した教材による教育効果の詳細な検証等が挙げられる．また，7 節の正答者数（正答比率）予測による難しさ尺度は直感的に理解しやすく，急速に注目を集める生成 AI に自動作問等をさせる際の難しさ指定の尺度としても有用と考えられるので，この方向での実験も今後行いたい．

## 謝辞

本研究は，科学技術振興機構 ACT-X 研究費 (JPMJAX2006) の支援を受けた．

## 参考文献

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

[2] 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.

[3] Robert Patrick. Comprehensible input and krashen's theory. *Journal of Classics Teaching*, Vol. 20, No. 39, pp. 37–44, 2019.

[4] Yo Ehara. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proc. of EDM*, 2022.

[5] Yo Ehara. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In *Proc. of Educational Data Mining (poster paper)*, 2022.

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

[7] 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 *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 66–78, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc., 2015.

[9] Shalini Pandey and Jaideep Srivastava. RKT: Relation-Aware Self-Attention for Knowledge Tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 1205–1214, New York, NY, USA, 2020. Association for Computing Machinery.

[10] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. Classifying math knowledge components via task-adaptive pre-trained bert. In *Proc. of AIED*, pp. 408–419, 2021.

[11] Lele Sha, Mladen Rakovic, Alexander Whitelock-Wainwright, David Carroll, Victoria M Yew, Dragan Gasevic, and Guanliang Chen. Assessing algorithmic fairness in automatic classifiers of educational forum posts. In *Proc. of AIED*, pp. 381–394, 2021.

[12] Shiting Xu, Guowei Xu, Peilei Jia, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. Automatic task requirements writing evaluation via machine reading comprehension. In *Proc. of AIED*, pp. 446–458. Springer, 2021.

[13] Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. Open-ended knowledge tracing for computer science education. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3849–3862, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[14] 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.

[15] Batia Laufer and Geke C. Ravenhorst-Kalovski. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30, April 2010.

[16] I. S. P. Nation and Rob Waring. *Teaching Extensive Reading in Another Language*. Routledge, November 2019. Google-Books-ID: xRu_DwAAQBAJ.

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

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

[19] 荘島宏二郎, 豊田秀樹. テストが複数の出題形式を含むときの項目母数の推定. 教育心理学研究, Vol. 52, pp. 61–70, 2004.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[21] Insu Paek and Ki Cole. *Using R for item response theory model applications*. Routledge, 2019.

[22] Yo Ehara. Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In *Proc. of ICTAI*, pp. 806–814, 2021.

[23] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized Reading Support for Second-language Web Documents. *ACM Trans. Intell. Syst. Technol.*, Vol. 4, No. 2, pp. 31:1–31:19, April 2013.

[24] BNC Consortium. *The British National Corpus*. 2007.

[25] Mark Davies. The corpus of contemporary american english (coca). Available online at `https://www.english-corpora.org/coca/`, 2008.