Vol.2021-HCI-195 No.3
Vol.2021-UBI-72 No.3
2021/11/30

# 多様なリーダビリティ尺度を整理して提示するUI

江原　遥[1,a)]

**概要**：外国語テキストの読みやすさ（リーダビリティ）を自動的に判定する自動リーダビリティ判定タスクにおいては、様々なリーダビリティ尺度が用いられている。複数の尺度を照らし合わせることで初めて見えてくる性質もあるかと思われるが、複数のリーダビリティ尺度をそのまま知恵字するだけでは、外国語学習者にとってそうした学習に有用な情報を得ることは難しい。本研究では、これらを整理してユーザに提示できるUIを提案する。例えば、テキスト全体の文脈を考慮して難度を決める時に重要な語と語自体の難しさの両方の尺度が考えられるが、これらをどのように提示するかや、自己申告式の語彙テストを用いて、語を知っていると誤解しやすそうな語をどのように提示するか、といった具体的な応用をかんがみ、様々な応用で有用な提示方法を考察し、議論する。

## 1. はじめに

学習支援において、テキストがどれだけ読みやすいかを測るリーダビリティ判定は学習用テキスト推薦などに応用を持つ重要な課題である。この「リーダビリティ」として、既存には多数の判定手法（尺度）が提案されてきた（[1], [2], [3] 他2節に詳述）。これらの多様なリーダビリティ測定手法のうち、学習支援・教育の観点からは、どの手法を用いるのがよいのだろうか？

学習支援や教育の目的でリーダビリティ判定手法を考えた時には、**理想的には次の性質が満たされる事が望ましい**と考えられる。1つは実際の教育の場でのテキストの「難しさ」を反映している測定手法が望ましいという事である。例えば、学習者への学習用テキストの推薦のためにテキストの難しさを知ることを目的にリーダビリティ判定を行うのであれば、実際に教員が行ったテキストの難しさの判定をもとにリーダビリティ判定を行うのがよいだろう。

次に、リーダビリティ判定がされたときに、その判定の根拠として、学習支援や教育で用いる学習単位に沿った根拠が提示されることが望ましい。例えば、テキスト中のどの単語が、この場面ではどの程度難しく、そうした単語がどれぐらい多いので、このテキストはこれぐらい難しい、といった判定ができると、その後の具体的な学習計画や教育の計画を立てやすく、有用であると考えられる。このように、単にテキスト全体を見てリーダビリティのスコアが数値で示されるだけではなく、「単語」に代表される教育や学習の場面においても用いられている単位で、テキスト中の難しい箇所が示されれば、教育上有用であろう。言い換えれば、リーダビリティ判定を機械学習における判別問題としてとらえた時、**判別器が、教育上通常用いられる単位で判別結果の根拠を説明できる「説明性」（Explainability）を有していることが望ましい**と捉えられる。

本稿では、特に外国語学習者の学習支援におけるリーダビリティ判定問題に注目して、様々な既存手法を整理したうえで、上記の2つの性質を持った識別器をそれぞれ提案する。前者については、語学教師によるリーダビリティ判定結果を模倣して、所与のテキストをできるだけ正確にリーダビリティ判定する「教師ありリーダビリティ判定」の手法である。後者については、語学教師のリーダビリティ判定結果を全く使わず、単語テスト結果データだけからリーダビリティ判定する「教師なしリーダビリティ判定」の手法である。この手法は、単語テスト結果データを用いているので、例えばテキストを難しくしている要因である単語を出力するなど、教育上重要な説明性を持つ。

実際に語学教師によるリーダビリティ判定結果データセットを用いたところ、前者の精度が後者より高かったものの、後者は教師なし設定で最高性能を達成した。また、後者の手法は、具体的に各テキストで難しい単語が得られるなど、教育上有用な説明性を有していた。

そして、これらの説明性を可視化し、入力テキストの文脈を考慮した難しさと、入力テキストに含まれる語彙のみの難しさを提示できる手法を提案する。

[1] 東京学芸大学
Tokyo Gakugei University, 4-1-1 Nukuikita-machi, Koganei-shi, Tokyo, 184–8501, Japan
[a)] ehara@u-gakugei.ac.jp

| 訓練 | 手法 | 研究の例 | 可能な性能評価指標 |
|---|---|---|---|
| 教師あり | ラベル予測 | [1], [2], [3], [4], [5] | 識別精度、順位相関 |
| 教師なし | ランキング学習 | [6] | 順位相関 |
| | 回帰スコア | [7] | |
| | 言語モデルスコア | [8] | |
| | 個人化リーダビリティ | [9], [10], [11], [12], [13] | |

**表 1** リーダビリティ判定の手法の問題設定の分類。教師あり/教師なしは、各テキストの難度の正解ラベルを用いるか否かを表す。

## 2. 自動リーダビリティ判定

前述の様々な手法は**表 1** のようにまとめられる。ラベル予測は、現在の自然言語処理では典型的な問題設定であり、リーダビリティ判定を、教師あり多値識別問題に帰着させる。代表的な研究としては、文献 [1], [2], [3], [4] など多くの既存研究が挙げられる。

ランキング学習は、文献 [6], [14] が詳しい。この論文では、所与の 1 テキストに対してテキストの難しさのラベルを予測する教師あり識別の問題設定ではなく、テキストの集合を入力として、これらのテキストを難しさの順番に並び替える「教師ありランキング学習」の問題に帰着させている。

英語のリーダビリティ判定の古典的な研究として、テキストの難しさの段階（テキストが用いられている学年など）に対して、テキスト中の単語の平均長などの回帰式を用いた研究がある。Flesch-Kincaid Grade Level (FKGL) [15] や、SMOG grade[16]、Coleman-Liau index [17] などがこれにあたる。日本語については文献 [18] が詳しい。

言語モデルは、直近で発表された論文 [8] で用いられている手法である。所与のテキストに対して、言語モデルのパープレキシティなどを用いた指標を計算する。個人化リーダビリティについては次節にまとめる。

## 3. 個人化リーダビリティ

応用言語学分野では、読み手となる学習者が所与のテキストを読めるかどうかを判定する研究が盛んである。この問題設定は、応用言語学分野では 1980 年代からある古典的な問題設定である [9], [10]。

この設定では、まず、読み手となる外国語学習者が事前に語彙テスト（単語テスト）を受けているものとする。そして、その語彙テストの結果を用いて、所与のテキスト中の知っている単語（既知語）を推定し、そこから既知語率を計算し、既知語率が閾値を超えた場合に、学習者がテキストを「読める」と判断する [9], [10], [11]。個人化リーダビリティは、簡単に言えば、「テキスト中で知らない単語の比率が多ければ、テキストは読めないはずだ」という直観に基づく手法である。

単純にはこの通りだが、語彙テストから個々の学習者の既知語をどのように推定するのか、また、既知語率が閾値を超えた場合にテキストが「読める」と判定する事の妥当性の 2 点について、詳述する。

### 3.1 既知語判定

語彙テストの結果から既知語を推定する点については、理想的には、テキスト中に現れそうなその言語の全ての語種について、学習者が知っているかどうか、学習者をテストする事が望ましいが、これには学習者に膨大な負担がかかり、非現実的である。現実的な方法として、高々数百語程度の語彙テストを、数十分ほど行ってもらい、このテスト結果を利用して、語彙テストに含まれない語を各学習者が知っているかどうかを推定する方法がとられている。例えば、文献 [11] では、100 語からなるテストを考案している。

語彙テストの結果から、語彙テストに含まれない語を各学習者が知っているか推定する手法の 1 つとして、単純に、語彙量 (vocabulary size) を用いた方法が挙げられる [19]。すなわち、全ての学習者が、British National Corpus などの均衡コーパス中の頻度順に語を学習することを仮定し、頻度の高い順に、推定された語彙量番目までの語は全て知っており、それより頻度の低い語については全て知らないと推定することで、既知語判定を行っている。この既知語判定問題については、機械学習の観点からは、語彙テストの結果を訓練データとして、語と学習者が与えられたときに学習者が語を知っているか否かを判定する、単純な二値識別の問題として定式化できる [20], [21]。この 2 値識別の問題に対して、半教師あり学習や能動学習を用いて精度向上した研究が文献 [22] である。また、既知語判定問題の標準的なデータセットについては、筆者が以前作成している [23]。

既知語判定問題は、テキスト中の知らない単語を発見する Personalized Complex Word Identification タスクの一種ともみなせ、テキスト簡単化の個人化などにも応用されている [24]。

### 3.2 既知語の閾値

学習者がテキストを「読める」既知語率の閾値については、95%または 98%の値が用いられることが多い。英語の既知語率と、テキストが「読める」閾値の関係性の検証については、文献 [25] が代表的である。具体的には、イスラエルの大学入試問題の英語の読解問題で、読み手が合格水準に達している場合に、その読解問題のテキストが「読める」と定義している。

また、既知語率の閾値については、既知語判定問題の識別器が返す、「ある語が既知語である確率」を用いて、所与のテキストの「既知語率の確率分布」を計算し、既知語率

の閾値の解釈性を保ったまま性能向上させる手法を、著者は過去に提案している [26]。

### 3.3 問題設定の違い

このように、個人化リーダビリティは、自然言語処理の典型的なリーダビリティ判定の評価用データセットとは、「リーダビリティ」の信頼性をどこに依拠するかの点で異なっている。自然言語処理の典型的なリーダビリティ判定の評価用データセットは、前述の OneStopEnglish コーパス [27] がそうであったように、基本的には語学教師などで構成される、テキストに対して正解ラベルを付与した「アノテータ」に依拠したリーダビリティである。つまり、「リーダビリティ」と言いながらも、実際に語学学習者がテキストを「読める」かどうかについては直接測定しておらず、その点はアノテータとなる語学教師の判断に依拠している訳である。

一方、個人化リーダビリティは、前述のように、学習者がテキストを「読める」か否かについて、読解問題を通じた検証に基づいているため、学習者がテキストを「読める」かどうかを直接的に計測して検証されてはいる。ただし、語彙テストからリーダビリティの判定に至るまでに、学習者の既知語の推定と、学習者の既知語率と学習者がテキストを「読める」か否かの推定の 2 つの推定が入っている。このように、複数の不確実な推定のプロセスが入っているにも関わらず、応用言語学分野で個人化リーダビリティが広く使われている理由は、おそらく、既知語率が解釈しやすい概念であること、また、既知語率の閾値が比較的狭い範囲（95%〜98%）で判定できることが服須の研究で示されていることが、貢献していると思われる。その背後には、「テキスト中で知らずに意味を推測しながら読める単語の量には認知的な限界があり、その限界はテキストによって大きく変わらないだろう」という直観があるものと思われる。

### 3.4 個人化リーダビリティを用いた教師なし自動リーダビリティ判定

本節では、前節で説明した個人化リーダビリティ判定器を用いて、自然言語処理分野で一般的な自動リーダビリティ判定器を作成する手法について詳述する。個人化リーダビリティでは、まず、個々の外国語学習者が知っている語彙を推定する必要がある。これには、100 単語程度の語彙テスト [11] の結果を分析し、この 100 単語以外の単語について、学習者が各単語を知っているかどうかを判定する手法が用いられる。このようにして推定された学習者が知っている語彙から、語彙テストを受けた学習者がテキストを読めるかどうかを判定する [19]。この際には、学習者がテキスト中の 95%〜98% 程度の単語を知っていれば学習者がテキストを読めるとする応用言語学上の知見を用いることが行われている。前述のように、学習者が事前に語彙テストを受けなければならないという設定のためか、応用言語学分野の外では、この手法はあまり用いられていない。

[23] では、100 問の語彙テストについて、クラウドソーシング上で集めた 100 人の被験者の回答が収められている。この語彙テストデータセットは、もちろん、リーダビリティを判定するテキストとは全く関係のないものである。この中で、最も標準的な語彙力を持つ学習者にとっての個人化リーダビリティ判定を、一般的なリーダビリティとして算出する。

語彙テストの分析には項目反応理論 [28] の考え方を応用したモデリングを用いる。これは、語彙テストのようなテストの各設問に対して、被験者が正答/誤答したという結果のデータセットから、被験者の能力値と各設問の難しさを同時に推定する心理モデルである。これは、機械学習の用語を用いれば、本質的には単純な 2 値ロジスティック回帰モデルと同等である。

$\mathcal{V}$ を語彙の集合とし、$\mathcal{L}$ を学習者の集合とする。$z_{v,l} \in \{0,1\}$ を、学習者 $l \in \mathcal{L}$ が語 $v \in \mathcal{V}$ に正答したかどうかとする。$z_{v,l}=1$ であれば、正答、$z_{v,l}=0$ であれば誤答とする。$z_{v,l}=1$ であることは、学習者 $l$ が単語 $v$ を知っていることを示唆する。

次に、$\{z_{v,l}\}$ を訓練データとして、次のモデルを学習する。

$$p(z=1|v,l) = \mathrm{sigmoid}(a_l - d_v) \tag{1}$$

(1) で、$a_l$ は学習者 $l$ の能力パラメタ、$d_v$ は単語 $v$ の難しさパラメタである。また、sigmoid はロジスティックシグモイド関数であり、$\mathrm{sigmoid}(x) = \frac{1}{1+\exp(-x)}$ で定義される。

ロジスティックシグモイド関数は、ニューラル識別モデルで用いられる softmax 関数の 2 値版であり、$(0,1)$ の範囲での単調増加関数である。$\mathrm{sigmoid}(0) = \frac{1}{1+1} = \frac{1}{2}$ であるので、学習者の能力パラメタ $a_l$ が単語の難しさパラメタ $d_v$ より大きければ、学習者が単語を知っている確率が 1/2 を超える。このように、学習者の能力と単語の難しさを同じ尺度で比較できるのが、項目反応理論の大きな特徴の 1 つである。

(1) だけでは、語彙テスト結果データセット中で設問に現れる単語の難易度しか $d_v$ として計算する事ができない。語彙テスト結果データセットで設問とされている以外の単語について $d_v$ を知るためには、$d_v$ をコーパス中の単語頻度などの特徴量から求めればよい。具体的には、次のようにして構成した。

$$d_v = -\sum_{k=1}^{K} w_k \log(\mathrm{freq}_k(v)+1) \tag{2}$$

(2) で、$K$ は使用するコーパスの数、$k$ は $k$ 種類目のコーパスを表し、$\mathrm{freq}_k(v)$ は、$k$ 種類目のコーパス中での単語 $v$ の頻度である。また、$w_k$ は、このコーパスに対する重みパラメタである。(2) で全体に負号がついているのは、一般

に、コーパス中の単語頻度が大きくなるほど単語は簡単になるので、単語の難しさとは逆の尺度であるためである。

パラメタ推定に必要な情報をまとめよう。$\{z_{v,l}\}$ と、コーパスの単語頻度 $\mathrm{freq}_k(v)$ が与えられれば、学習者 $l$ の能力値パラメタ $a_l$ とコーパス $k$ の重みパラメタ $w_k$ が推定できる。(1) と (2) をまとめると、sigmoid 関数内がパラメタに対して線形であるため（つまり、2 種類のパラメタの積から構成される項が存在しないため)、これはロジスティック回帰を使って表現する事ができることがわかる。実際、実験では、Python の機械学習パッケージである scikit-learn [*1]を用いた。scikit-learn は、内部的にはロジスティック回帰の高速実装として有名な LIBLINEAR [*2]を呼び出している。

このようにしてパラメタを求めた後、所与のテキスト $\mathcal{T}$ に対するリーダビリティを判定する。簡単には、最も $a_l$ が標準的な学習者 $l_{\mathrm{avg}}$ を 1 人選び、この学習者がこのテキスト中の各単語を知っている確率をつぎのように求めればよい。ここで、$v \in \mathcal{T}$ は、テキスト中の単語 $v$ を表し、$\mathcal{V}(\mathcal{T})$ はテキスト中で現れる語彙の集合を表す。例えば、テキスト中で 3 種の単語が出現したら、$v \in \mathcal{T}$ では、頻度の回数分、各単語の確率値をかけ合わせるのに対して、$\mathcal{V}(\mathcal{T})$ では、単純に語種の数だけでかけ合わせる。

$$score_{\mathrm{sum}}(\mathcal{T}) = -\log\left(\prod_{v\in\mathcal{T}} p(z=1|v,l_{\mathrm{avg}})\right) \quad (3)$$

$$score_{\mathrm{vsum}}(\mathcal{T}) = -\log\left(\prod_{v\in\mathcal{V}(\mathcal{T})} p(z=1|v,l_{\mathrm{avg}})\right) \quad (4)$$

$$score_{\mathrm{vavg}}(\mathcal{T}) = -\frac{1}{|\mathcal{V}(\mathcal{T})|}\log\left(\prod_{v\in\mathcal{V}(\mathcal{T})} p(z=1|v,l_{\mathrm{avg}})\right) \quad (5)$$

また、$p(z=1|v,l_{\mathrm{avg}})$ が計算できれば、上式に変わり、テキスト中の 95% の単語知っている確率を求め、これをスコアにする方法もある [26]。パラメタ推定の際には、リーダビリティ評価用データセットのテキストの難しさラベルは一切使用しないため、この手法は「教師なし」に分類される。

最後に、以上で述べた個人化リーダビリティ判定においては、語彙テスト結果から、単語の難しさパラメタを求める部分 (2) が本質であることを説明する。説明のため、最も平均的な能力の学習者を 1 人定めて $l_{\mathrm{avg}}$ としたが、(1) では sigmoid 関数は単調増加関数であること、$a_l$ は単純に $d_v$ に足されていることから、実は、どの学習者を選んでも、$a_l$ を固定した時点で、学習者が単語を知っている確率 $p(z=1|v,l$ に寄与するのは $d_v$ だけである。従って、上記の方法は、単語テスト結果を用いて、単語テスト結果とよく相関するような単語の難易度を、コーパス中の単語の頻度 $\mathrm{freq}_k(v)$ から作り出す手法であると捉えられる。

## 4. 実験結果と考察

データセットには、第二言語学習者を対象にしたデータセットとして比較的最近に報告されたものであることから、OneStopEnglish データセットを用いた [3]。このデータセットは、Guardian 誌の記事を語学教師が Elementary, Intermediate, Advanced の 3 種類に書き換えたものである。各レベルには 189 件のテキストがあり、全体では 567 件である。教師あり手法とも比較するため、これを、339 件の訓練データ、114 件の開発データ、114 件のテストデータに分割し、最後のテストデータを用いて性能検証を行う。比較手法は、下記の通りである。

古典的な自動リーダビリティ判定式については、Python の **readability** パッケージを用いて実装した [*3]。このパッケージには、英語のリーダビリティ判定式として、Flesch-Kincaid Grade Level[15], ARI, Coleman-Liau Index, Flesch Reading Ease, Gunning Fog Index, LIX, SMOG Index [16], Dale-Chall Index が実装されているのでこれを用いた。紙面の都合上、全ての数式をここに表記する事はしない。具体的な式については、脚注に記した **readability** パッケージのプロジェクトページに記載がある。これらの手法は、全てリーダビリティのラベルを用いないので、「教師なし」に分類される。

次に [8] において提案されている、ニューラル言語モデルを用いた教師なし自動リーダビリティ判定について説明する。ニューラル言語モデルについては、事前学習モデル **bert-large-cased-whole-word-masking** を Huggingface の事前学習モデル一覧より取得し、これを用いて計測した各テキストのパープレキシティの平均値をリーダビリティとしたのが **BERTLMavg** である。[8] では BERT を用いた言語モデルとしては **bert-base-uncased** を事前学習モデルに使用したものが用いられているが、**bert-large-cased-whole-word-masking** はこれより大きなモデルである。テキストの文分割については、**nltk** パッケージ[*4]の **sent_tokenize** 関数を用いた。

[8] では、BERT の言語モデルを用いた手法はよい性能をあげられていないが、そのほかの手法は公開されている事前学習モデルを用いておらず、再実装が難しい。そこで、[8] の OneStopEnglish データセットでの最高性能を達成している **TCN RSRS-simple** の結果を、実験結果の表に加えた。ただし、[8] で用いたテストデータが入手できなかったため、この手法は直接の比較が可能ではないため、表中では (*) を

[*1] https://scikit-learn.org/stable/
[*2] https://www.csie.ntu.edu.tw/~cjlin/liblinear/
[*3] https://pypi.org/project/readability/
[*4] nltk.org

用いてそのことを明記した。**TCN RSRS-simple** は、単純に言えば、Temporal Convolutional Network (TCN) を Simplified Wikipedia コーパス上で事前学習させ、さらに、パープレキシティにかわり、Ranked Sentence Readability Score (RSRS) という [8] が独自に定義した指標を用いて判定するものである。**TCN RSRS-simple** のさらなる詳細については [8] を参照されたい。

最後に、**提案手法**が本研究の提案手法である。これは、語彙テストデータセット [23] を用いて、前述のパラメタ推定を行い、(3) を用いて学習者が各単語を知っている確率を自動リーダビリティ判定に用いたものである。コーパスからの単語頻度の特徴量としては、英語教育上広く使われていることから、British National Corpus [*5]と Corpus of Contemporary American English (COCA) [*6]を用いた。さらに、単純に、これらのコーパス頻度を表す特徴量を **BNC**、**COCA** として結果表中に掲載した。

**表 2** に結果を示す。[8] では、スコアとリーダビリティ評価用データセットのラベルとの相関として Pearson's $\rho$ しか用いていないが、これは、スコアの線形性が低いとスコアが下がってしまう事から、順位相関係数として Speraman's $\rho$、Kendall's $\tau$ を用いた。さらに、今回はリーダビリティ評価用データセットでは、一般に、同じ難しさレベルのテキストが多くあるため、同順を多く含むデータセットになっており、同順補正の方法によってスコアが大きく影響を受ける。一般に使われている同順補正は $\tau$-b であり、単に Kendall's $\tau$（ケンドールの順位相関係数）と言った場合、こちらが使用されることが多い。

表 2 の最も左側には、教師なし、教師ありの分類を示した。

最初に、全ての「教師なし」の手法において、提案手法（式 (4)）が全ての尺度で最も良い性能を示した。提案手法は 0.760 と、後述の教師ありの設定で訓練データが少ない場合である spvBERT_half を超える順位相関を達成した。式 (4) は、頻度を考慮して足し合わせる式 (3) よりもよい性能を達成しており、テキスト中の語種数に重要な情報が格納されていることがわかる。このことは、(5) によって、テキスト中の語種数で平均を取り、テキスト中の語種数の情報を反映しないと、著しく精度が下がることからもわかる。

次に、テキスト中の語のうち、難しい語（学習者が知っている確率が低い語）がリーダビリティ判定において重要なのか、簡単な語が重要なのかを確認した。(4) で、テキスト中の難しい語種上位 30 語を削ると、Spearman's $\rho$ が 0.769 まで向上した。一方、簡単な語種上位 30 語を削った場合、0.761 であった。この結果は、テキスト中の難しい語種には、実際に難しい語の他にも、“Redmond” や “Stockholm” といった固有名詞も含まれてしまっており、難しい語種を削ることによって、こうしたノイズが削減されるためであると考えられる。

次に、提案手法以外の手法との比較を具体的にみていく。

**BERTLMavg** は [8] よりも大きな事前学習モデルを用いてパープレキシティを計測したが、良い結果を示さなかった。これは、パープレキシティが第二言語学習者向けのリーダビリティの尺度として適していないことを示唆する。

**TCN RSRS-simple** は [8] における OneStopEnglish データセット上の最高性能を達成した手法である。[8] においては、性能比較に Pearson's $\rho$ のみが用いられているため、この値だけを表示した。ただし、彼らは同じ OneStopEnglish データセットを用いてはいるが、性能値を算出するために具体的にどのデータをテストデータに用いたのかが公開されていないため、直接の比較は難しく、(*) でこのことを明示した。直接の比較は難しいものの、提案手法は、**TCN RSRS-simple** よりもよい性能を達成できていることがわかる。

また、おもしろいことに、**BNC** と **COCA** の単語頻度については、英語教育の分野では単語の難しさを測る良い指標とされているものの、これら単独ではリーダビリティ評価用データセットのラベルと良い相関が得られなかった。一方、**提案手法**では、前述のように、これらの単語頻度特徴量を (2) を用いて組合せ、語彙テストデータセットに沿う単語難易度を求めている。このことから、複数のコーパスからの単語頻度を組み合わせて、「第二言語学習者にとっての単語の難しさ」をきちんと語彙テストデータから計測することが、自動リーダビリティ判定に重要であることが示唆される。

教師あり学習の手法の結果を示す。**spvBERT** は BertForSequenceClassification 関数を用いてリーダビリティラベルを用いて学習した結果であり、**spvBERT_half** は、訓練データを半分にして同じ学習をした場合である。モデルとしては、前述の **bert-large-cased-whole-word-masking** を用いた。教師データを用いることにより、spvBERT は教師なしである提案手法より高い性能を達成できている。

最後に、表 2 からの教育の観点からの説明性について考察する。spvBERT は、教師あり学習であり、テキスト全体の文脈を見て判別する手法である。一方、提案手法は、教師なし学習ではあるが、単語の難しさについては語彙テストデータセットを用いて正確に求める手法である。提案手法は、単語の難しさについては正確に求めるものの、文脈については見ていない。従って、**spvBERT の性能値と、提案手法の性能値の差が、リーダビリティ判定を平均的な単語難易度だけではなく、文脈を見て行う事による性能向上であると考えることができる。**

[*5] https://www.english-corpora.org/bnc/
[*6] https://www.english-corpora.org/coca/

**表 2** OneStopEnglish データセットでの実験結果・考察

| 教師あり/なし | 手法 | Spearman's $\rho$ | Kendall's $\tau$-b | Pearson's $\rho$ |
|---|---|---|---|---|
| 教師なし | Flesch-Kincaid | 0.324 | 0.253 | 0.359 |
| | ARI | 0.317 | 0.248 | 0.351 |
| | Coleman-Liau | 0.373 | 0.295 | 0.372 |
| | FleschReadingEase | -0.387 | -0.301 | -0.426 |
| | GunningFogIndex | 0.331 | 0.257 | 0.362 |
| | LIX | 0.348 | 0.273 | 0.383 |
| | SMOGIndex | 0.456 | 0.360 | 0.479 |
| | RIX | 0.437 | 0.340 | 0.462 |
| | DaleChallIndex | 0.495 | 0.387 | 0.506 |
| | TCN RSRS-simple | - | - | 0.615(*) |
| | BERTLMavg | -0.220 | -0.173 | -0.040 |
| | BNC | -0.012 | -0.009 | -0.006 |
| | COCA | 0.018 | 0.016 | 0.039 |
| | 式 (3) | 0.730 | 0.592 | 0.715 |
| | 式 (4) | **0.760** | **0.617** | **0.754** |
| | 式 (5) | 0.581 | 0.454 | 0.589 |
| 教師あり | spvBERT_half | 0.751 | 0.729 | 0.747 |
| | spvBERT | 0.866 | 0.856 | 0.864 |

## 5. 文脈を考慮したリーダビリティと語彙のみのリーダビリティを比較できる UI

前節のように、**spvBERT** は、文脈を考慮したリーダビリティ尺度になっており、提案手法は、純粋に語彙のみを考慮した（しかし、その代わりに学習者にとっての語の難易度を正確に考慮した）リーダビリティ尺度になっている。

**spvBERT** が文脈を考慮しているという事は、すなわち、語の並び、構文的な難しさもリーダビリティ尺度に影響してきているという事である。使われている単語はやさしいが、構文的に難しく、適切な構文を解釈するのに時間がかかる文は**ガーデンパス文**（Garden path sentences) と呼ばれ、言語教育や心理学の分野で盛んに研究されている。特に、HCI との関連という事でいえば、ガーデンパス文を読むときの眼球運動に関する研究がある [29]。

そこで、入力されたテキストに対して、**spvBERT** のスコアと、語彙のみを考慮した提案手法によるスコアを同時に表示することにより、構文を考慮した難しさと、語彙のみの難しさの両面でリーダビリティを評価できる UI を提案する。リーダビリティの尺度を新たに提案する論文は多いが、このように、構文を考慮した尺度と語彙のみの尺度の両方を提示することにより、複合的にリーダビリティを可視化しユーザに提示できるシステムは、知る限り本研究が初である。

**図 1** に、可視化の例を示す。各点はテキストである。青い●は OneStopEnglish データセットのテストセットから取ったテキストを表し、橙色の▲はガーデンパス文の文例である。横軸は **spvBERT** によるスコア、縦軸は提案手法による語彙のみのスコアである。具体的には (3) に対して、テキストの長さ間で比較できるように、これをテキストの長さ（単語数）で割った値を表示している。

本研究では、具体的には、**表 3** に示すガーデンパス文を用いた。文例は、`https://en.wikipedia.org/wiki/Garden-path_sentence` 並びに、`https://shublog0203.com/` の「ガーデンパス現象って何【面白いぞ言語の話】」という記事から取得し、日本人名 "Taro" については、これが単語の難しさに影響しないように、"Tom" に変更した。構文を正確に示すには定義など紙面を多く必要とするので、本論文の著者による訳文も加えた。ここから、原文の構文が複雑であることが読み取れる。

図 1 の結果を見ると、ガーデンパス文は、語彙のみで計測した場合には非常にリーダビリティが低い（簡単）と判定されるのに、**spvBERT** は、適切にその難しさを評価できていることが分かる。これは、**spvBERT** が構文的な難しさを考慮しているためであると推察される。最も難しかったのは表 3 に挙げた最初の文例である。

## 6. まとめ

本研究では、学習支援の観点からリーダビリティ判定に求められる機能について考察し、その機能を持ったリーダビリティ判定手法を 2 種類、具体的に提案した。また、構文を考慮したリーダビリティと、語彙のみによるリーダビリティの 2 つを同時に提示することにより、テキストの難しさを視覚的に提示する手法を提案した。ガーデンパス文を用いて、この手法の適切性を定性的に評価した。今後の展望として、構文解析の手法も活用することで、より精緻に語彙・構文・複合の 3 種類の観点からのリーダビリティを提示する仕組みの提案があげられる。自動リーダビリティ判定についてのさらなる詳細は、`http://yoehara.com/readability/`

**表 3** 今回試したガーデンパス文

| 原文 | 訳文（著者による） |
|---|---|
| Tom gave the boy the cat bit a bandage. | トムは、猫が噛んだ少年にバンデージをあげた。 |
| The complex houses married and single soldiers and their families. | 複合体は、結婚しているまたは独身の兵士とその家族に住居を提供した。 |
| The cotton clothing is made of grows in Mississippi. | 服の原料になっている綿花はミシピッピで育った。 |

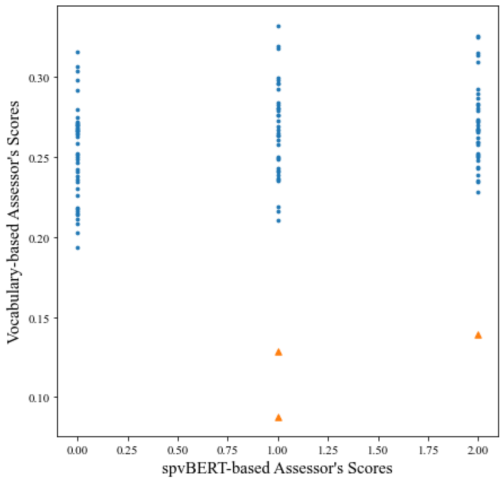


**図 1** リーダビリティの可視化

にまとめる予定である。

## 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）の支援を受けた。

### 参考文献

[1] Feng, L., Jansche, M., Huenerfauth, M. and Elhadad, N.: A Comparison of Features for Automatic Readability Assessment, pp. 276–284 (online), available from ⟨https://www.aclweb.org/anthology/C10-2032⟩ (2010).

[2] Xia, M., Kochmar, E. and Briscoe, T.: Text Readability Assessment for Second Language Learners, *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, San Diego, CA, Association for Computational Linguistics, pp. 12–22 (online), DOI: 10.18653/v1/W16-0502 (2016).

[3] Vajjala, S. and Lučić, I.: OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 297–304 (online), DOI: 10.18653/v1/W18-0535 (2018).

[4] Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M.: Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, New York, Association for Computational Linguistics, pp. 460–467 (online), available from ⟨https://www.aclweb.org/anthology/N07-1058⟩ (2007).

[5] Fujinuma, Y. and Hagiwara, M.: Semi-Supervised Joint Estimation of Word and Document Readability, *arXiv:2104.13103 [cs]*, (online), available from ⟨http://arxiv.org/abs/2104.13103⟩ (2021). arXiv: 2104.13103.

[6] Tanaka-Ishii, K., Tezuka, S. and Terada, H.: Sorting Texts by Readability, *Computational Linguistics*, Vol. 36, No. 2, pp. 203–227 (online), DOI: 10.1162/coli.09-036-R2-08-050 (2010).

[7] Flesch, J.: Flesch-Kincaid readability formula (1965).

[8] Martinc, M., Pollak, S. and Robnik-Šikonja, M.: Supervised and Unsupervised Neural Approaches to Text Readability, *Computational Linguistics*, Vol. 47, No. 1, pp. 141–179 (2021).

[9] Nation, P.: *Teaching and Learning Vocabulary*, Heinle and Heinle, Boston, MA (1990).

[10] Laufer, B.: What percentage of text-lexis is essential for comprehension, *Special language: From humans thinking to thinking machines*, Vol. 316323 (1989).

[11] Beglar, D. and Nation, P.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).

[12] Ehara, Y., Sato, I., Oiwa, H. and Nakagawa, H.: Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty, *Journal of Information Processing*, Vol. 26, pp. 267–275 (online), DOI: 10.2197/ipsjjip.26.267 (2018).

[13] Lee, J. and Yeung, C. Y.: Personalized Substitution Ranking for Lexical Simplification, *Proceedings of the 12th International Conference on Natural Language Generation*, Tokyo, Japan, Association for Computational Linguistics, pp. 258–267 (online), DOI: 10.18653/v1/W19-8634 (2019).

[14] 佐藤理史：均衡コーパスを規範とするテキスト難易度測定，情報処理学会論文誌，Vol. 52, No. 4, pp. 1777–1789 (2011).

[15] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical report, Naval Technical Training Command Millington TN Research Branch (1975).

[16] Mc Laughlin, G. H.: SMOG grading-a new readability formula, *Journal of reading*, Vol. 12, No. 8, pp. 639–646 (1969).

[17] Coleman, M. and Liau, T. L.: A computer readability formula designed for machine scoring., *Journal of Applied Psychology*, Vol. 60, No. 2, p. 283 (1975).

[18] Hasebe, Y. and Lee, J.-H.: Introducing a readability evaluation system for Japanese language education, *Proceedings of the 6th international conference on computer assisted systems for teaching & learning Japanese*, pp. 19–22 (2015).

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

[20] Ehara, Y., Shimizu, N., Ninomiya, T. and Nakagawa, H.: Personalized Reading Support for Second-language Web Documents by Collective Intelligence, *Proc. of IUI*, IUI '10, ACM, pp. 51–60 (online), available from ⟨http://doi.acm.org/10.1145/1719970.1719978⟩ (2010). event-place: Hong Kong, China.
[21] Ehara, Y., Shimizu, N., Ninomiya, T. and Nakagawa, H.: Personalized Reading Support for Second-language Web Documents, *ACM Trans. Intell. Syst. Technol.*, Vol. 4, No. 2, pp. 31:1–31:19 (online), DOI: 10.1145/2438653.2438666 (2013).
[22] Ehara, Y., Miyao, Y., Oiwa, H., Sato, I. and Nakagawa, H.: Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning, *Proc. of EMNLP*, pp. 1374–1384 (online), DOI: 10.3115/v1/D14-1143 (2014).
[23] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).
[24] Lee, J. and Yeung, C. Y.: Personalizing Lexical Simplification, *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Association for Computational Linguistics, pp. 224–232 (online), available from ⟨https://www.aclweb.org/anthology/C18-1019⟩ (2018).
[25] Laufer, B. and Ravenhorst-Kalovski, G. C.: Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension, *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30 (online), available from ⟨https://eric.ed.gov/?id=EJ887873⟩ (2010).
[26] Ehara, Y.: Uncertainty-Aware Personalized Readability Assessments for Second Language Learners, *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1909–1916 (online), DOI: 10.1109/ICMLA.2019.00307 (2019).
[27] Vajjala, S. and Rama, T.: Experiments with Universal CEFR Classification, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 147–153 (online), DOI: 10.18653/v1/W18-0515 (2018).
[28] Baker, F. B.: *Item Response Theory : Parameter Estimation Techniques, Second Edition*, CRC Press (2004).
[29] 内景太，前田由貴子，　佐藤寛：平易・難解文章が注意欠如・多動症傾向を持つ大学生の眼球運動と文章読解に与える影響―アイトラッキングを用いて―，日本心理学会大会発表論文集日本心理学会第 81 回大会，公益社団法人日本心理学会，pp. 3A–033 (2017).