

# 汎用言語モデルは日本語学習者データに基づく 語彙難易度を予測できるのか

梁震<sup>1</sup> 彭悦<sup>1</sup> 笹尾洋介<sup>2</sup>

<sup>1</sup>京都大学大学院 人間・環境学研究科 <sup>2</sup>京都大学 国際高等教育院

ryo.shin.j85@kyoto-u.jp peng.yue.35z@st.kyoto-u.ac.jp

sasao.yosuke.8n@kyoto-u.ac.jp

## 概要

日本語学習における語彙習得の重要性を踏まえ、本研究では汎用言語モデル GPT-4<sup>i</sup>を用いて日本語の語彙難易度を分析した。具体的には、GPT-4 に語同士の難易度関係を出してもらい、その結果を基に機械学習で語彙難易度を算出する方法を採用した。検証には、中国語母語話者にとっての日本語語彙難易度が判明している語を用いて、本稿の手法で語彙の難易度を推定した。その結果、既存の調査結果の 99% 信頼区間内に収まるデータの割合は 71% と比較的高く、従来の手法よりも迅速かつ広範囲にわたる難易度評価が可能となった。この研究により、日本語教育における効果的な教材開発や教授法の改善に寄与する可能性が示された。

## 1 はじめに

言語の学習において語彙学習は極めて重要であり、特に日本語は、諸言語の中で比較的多くの語を覚える必要のある言語であり、新聞や雑誌を自由に読みこなすには最低 3 万語を理解できないといけなことが指摘されている[1]。膨大な量の語彙をより効率的に学習できるように、過去に様々な基本語彙を選定するための研究調査が行われた。代表的な方法として、言語学専門家が主観的に決定する方法、各種の言語資料の分析を踏まえた（特に使用頻度の高さや使用分野の広さの視点からの）分析に立脚する方法、及びある個人や集団の理解できる語彙についての心理実験の結果に基づく方法があるが、使用頻度の低い語彙の中にも重要な語彙が含まれている可能性も指摘されている[2]。ため、基本語彙の選定や日本語教材作成、語彙指導を行う際に、より客観的かつ実用的な学習対象を選ぶには、3 つ目の方法のよう

に、語彙の理解度（語が表す概念が理解されているかどうかを指しているため、本稿では「難易度」という）を踏まえることが必要であろう。

語彙の難易度に関して、易しい語から難しい語へと徐々に学習することが提唱されているが、学習者によって難易度の順序が異なる可能性も考えられる[3]。また、林[4]も語の難易度は日本語学習者の背景知識に依存し、学習者の母語の違いが考慮されていないことを指摘した。彭他[5]は中国語母語話者を対象に調べた結果、日本語語彙の使用頻度と学習者にとっての語彙難易度とは弱い相関しかないことが示されたため、語彙の難易度推定に関して、使用頻度だけでは日本語語彙の難易度を判断することが難しいことがわかる。

語彙難易度の測定が重要であるのにも関わらず大規模な測定が出来ていない理由としては、語の難易度の計測には大規模な調査実験をする必要があるためである。そのため、語彙指導を行う際には語彙頻度に基づいている場合が多い。語彙頻度に基づいた語彙テストの場合、学習者の母語の違いを十分に考慮していない。例えば、「喜劇」のような漢字由来の語は、使用頻度が低く難易度が高いとされがちであるが、中国語母語話者にとっては漢字の意味から直感的に理解可能であり、実際の難易度は低くなる傾向がある。

このような背景を踏まえ、本研究では、汎用言語モデル GPT-4 を用いた新たな語彙難易度評価方法を提案する。GPT-4 を含む近年の汎用言語モデルは、自然言語で与えた指示に従って様々なタスクをこなす汎用性を獲得してきている、要約・推敲・翻訳、質問応答、対話、算術・論理推論、コーディングといった多様なタスクを実行できる。しかも、個々のタスクの性能は従来の単機能のモデルに比肩するか、

<sup>i</sup> Open AI GPT4 Model  
(<https://openai.com/research/gpt-4>)

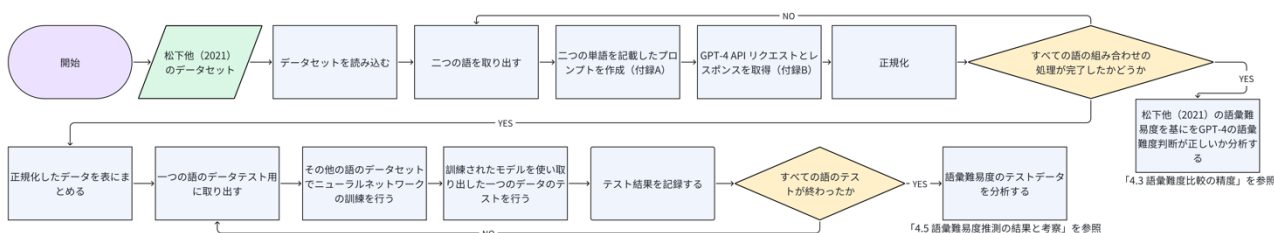


図 1 研究ワークフロー

場合によっては超えるという報告も出てきている [6].GPT-4 のトレーニングデータセットは、インターネット上の広範囲なテキストから構成されている。データセットの中には多種多様な言語の母語話者のデータがあるため、学習者の母語別の日本語難易度の分析も期待できる。

## 2 関連研究

日本語語彙の難易度判定の問題を解決するために、水谷他[3]は教科書コーパス語彙表を単語難易度の基準データとし、また、単語難易度と相関が高いと考えられる 3 種の手がかりを素性として用い、機械学習によって語の難易度推定を試みた。しかし、日本語教育語彙表で難易度の基準を設定することが妥当であるか不明である。また、この方法では学習者の母語の影響は考慮されていない。それに対して、林[4]は、中国語を母語とする日本語学習者を対象に、日本語の辞書と中国語の辞書における語義の違いや漢字表記の違いに基づいた難易度の推定方法を提案した。しかし、この研究では中国語母語話者を対象とした場合のみ議論しており、母語の異なる学習者への応用までは至ることは難しいため、より汎用的な方法を見つけることが期待される。この 2 つの研究を踏まえて、より応用範囲の広い、かつ、学習者の母語を考慮した難易度推定の方法の開発は重要であると言える。

語彙の難易度について、研究により定義が分かれているが、一般に語が表す概念の難しさとされる場合 [3] が多い。本研究で用いる 125 語のデータセットの難易度はラッシュ分析により算出したものである。ラッシュ分析における項目難易度推定は標本に依存せず、平均項目難易度を 0 として計算されているため、算出された数値は相対的な値であるが、任意の 2 語間の難易度の比較は可能であり、本稿での研究には十分に利用できるデータと言える。

## 3 研究目的と方法

本研究の目的は、汎用言語モデルは日本語の語彙難易度を理解できるかどうかを究明することによって、様々な言語を母語とする日本語学習者にとっての語彙難易度推定の方法を提案することである。具体的には、汎用言語モデルに一部の語彙の難易度関係を学習させた上で出力された語彙難易度をもとに作成した機械学習モデルの予測値が、実際の語彙難易度とどれくらい一致しているかを調べる。

本稿では、汎用言語モデルによる語と語の難易度比較の論理値 (-1 もしくは 1 のデータ) をさらにニューラルネットワーク機械学習で語彙難易度を推定する手法を採用する。

検証のために、この研究では、松下他[7]で実施された語彙テストの分析結果の一部を用いている [4]。松下他[7]は、日本国内外にいる、様々な言語を母語とする日本語学習者を対象に、VSTRJ-50K ともう一つの漢字テスト KCT を実施した後、さらに、ラッシュ分析により各問に関連している語の平均項目難易度を算出したが、本研究で使用したのは、中国語母語話者 201 名の VSTRJ-50K の結果による 125 語の平均項目難易度であった。

VSTRJ-50K テストの詳細は以下の通りである。まず、佐藤他[8]によると、このテストは、松下[9]の「日本語を読むための語彙データベース」の使用頻度順位をもとに、上位 50000 語のうち、凡そ 400 語毎に 1 語を選ぶように、125 語を無作為抽出した上で、1 語につき 1 問で語の意味についての選択問題が出題されている。また、このテストは筆記試験 (時間制限 40 分以内) であり、各問において、目標語自体、その目標語を使った 1 つの非定義文、及び 4 つの意味が日本語文字により提示され、受験者が回答時に 4 つの選択肢の中から一つを選ぶ形となっている。ただし、「ぜんぜんわからないときは答えないください」との指示も示されている。それから、評価について、1 問 1 点で計算された得点 (満点は 125 点)

×400 で各受験者の語彙サイズを推計できるが、松下[7]のラッシュ分析で用いられたのは、各受験者の得点の素点であった。この研究で得られた 125 語の難易度を含むデータセットを用いる。

汎用言語モデルに直接語の難易度に関する推定を問い合わせる方法を試みたが大まかな数値しか出力されなかった。この方法では得られた結果は一般的な推測に留まり、実用的な精度に欠けていた。

そこで図 1 で示された手順のように、研究の方針を変更した。まず、項目難度が判明している 125 語のデータセットを、(1) プロンプト (付録 A) に記載する 25 語、(2) 単語間難易度比較用 100 語、の 2 つに分割した。まず、GPT-4 にプロンプトに記載されている情報を学習させることで、語の難易度に影響を与える具体的な要因を推定させることが容易となる。残りの 100 語については、2 つずつの組み合わせの語彙難易度の比較し、相対的な難易度を評価する方法が採用された。語彙難易度の比較データは、後続のニューラルネットワーク機械学習による語彙難易度の分析用データとした。

## 4 実施と結果

### 4.1. プロンプトとリクエスト

GPT-4 を含む汎用言語モデルは対話形式で入力することができる。その際に、適切な文章 (プロンプト) を用いることで期待する返答を得ることができる。プロンプトは、富嶋[10]のプロンプト紹介を参考に作成した。

本研究では、具体的な 125 語をデータセットから無作為に選出した 25 語を含むプロンプトを構築し、GPT-4 に提供した。これにより、汎用言語モデルが中国語母語話者にとって難易度に影響する要因を判断するための情報を学習させることができる。これは、言語処理タスクにおける GPT-4 の適応性を利用し、より複雑な言語理解の課題への適用させるためである。残りの 100 語を 2 語ずつ難易度の比較をさせた。そのため、プロンプトは組み合わせ論的に 4,950 通りある。これらのプロンプトとそれに対する GPT-4 の反応の具体的な例は、付録 A 及び付録 B にて詳述される。これらの 4,950 のプロンプトを Python 3.11.7

を用いて gpt-4-1106-preview モデルの API ヘリクエストを行った。

### 4.2 レスポンスの正規化

GPT-4 から得られる日本語のレスポンス (付録 B) は文章であり、正答率の統計分析および後続の語彙難易度算出のため、論理値へと変換する必要性が生じる。具体的には、比較される二つの語に関して、【単語 A】の難易度が【単語 B】よりも高いと判断された場合には -1 を割り当て、それ以外に 1 を割り当てる。

### 4.3 語彙難易度比較の正答率

100 語の 2 つずつの組み合わせは 4,950 通りある。本研究では全て【単語 B】の方が松下他[7]のデータセットで語彙難易度が高いものを選んだため、「1」が正答である。統計結果は、「1」が 3732、「-1」が 1218 であった。そのため、GPT-4 の正答率の割合は 75.39% である。これは、GPT-4 は比較的高い割合で語彙難易度の関係性を理解していると言える。正規化したデータから見るに、語彙判断を誤った語に偏りがある。例えば、「完徳」という語の語彙難易度は容易であると判断されているが実際は容易とは言えない。このような誤りが生じた理由は、プロンプトに記載した語の難易度を習得した結果、漢語は中国語母語話者にとって容易であると判断されたからだ。リスパンスから推測できるが、実際中国語ではあまり使用されていない漢語が存在する。プロンプトに記載した語数が少なく、以上のような特徴を把握することができなかつたと考えられる。

### 4.4 機械学習による語彙難易度の推定

前節では GPT-4 による語彙難易度比較の真理値の分析を行った。本稿の目的は、汎用言語モデルによる語彙難易度分析であるため、さらに、前節の真理値データセットを用いた語彙難易度の推測を行う。

それぞれの語に 100 個の真理値のデータがあるため、回帰手法を用いれば連続値である語彙難易度を予測することが可能となる。本研究では、ニューラルネットワークを用いて語彙難易度の推定を行った。使用したネットワークは AutoKeras<sup>ii</sup>によって見つけ

<sup>ii</sup> Developed by DATA LAB at Texas A&M University, collaborating with keras-team for version 1.0 and above. (<https://autokeras.com>)



られたものであり、100 個の真理値データを入力として受け取り、連続値である語彙難易度を出力する。このモデルは、以下の構造を持っている。入力層 (Input Layer) は 100 次元のベクトルを受け取り、カテゴリデータのエンコーディング (Multicategory Encoding) と正規化 (Normalization) を行った後、2 層の隠れ層 (Dense) を経て、最終的に 1 次元の出力を行う (Regression Head)。このモデルには 4,522 個のパラメータがあり、このうち 4,321 個が訓練可能である。モデルの検証には、Leave-One-Out (LOO) クロスバリデーションを採用した。この手法では、データセットの各サンプルに対して、そのサンプルをテストデータとし、残りのすべてを訓練データとして使用する。これにより、モデルの汎化能力を評価することができる。

#### 4.5 語彙難易度推測の結果と考察

ニューラルネットワーク機械学習の結果は以下の通りである。語彙難易度の予測結果を示す図 2 において、横軸は語の No.、縦軸は語の難易度である実際の難易度値 (緑色の点) と予測値 (赤色の点) は、全体的な傾向において相関が見られる。95%信頼区間 (青色の破線) と 99%信頼区間 (赤色の破線) 内に予測値が収まっている割合は、それぞれ 61% と 71% と計算された。これは、予測モデルが語彙難易度を概ね正確に推定できていることを示唆している。

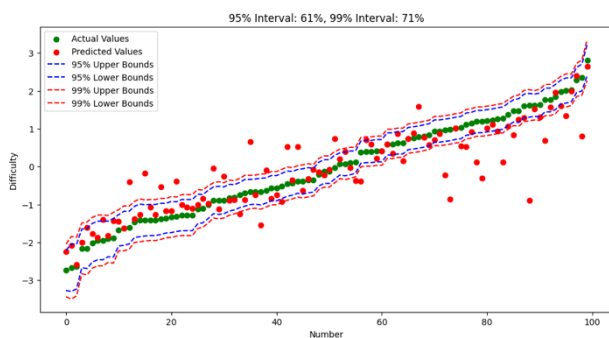


図 2 語彙難易度予測結果

しかし、予測値と実際の難易度値との間にいくつかの外れ値が存在することも観察された。特に、語彙難易度が低いと予測されたにもかかわらず、実際には難易度が高いと評価されている語がいくつか存在する。これは、モデルが特定の語の特性を十分に捉えきれていないことが原因と考えられる。また、モデルの訓練に使用されたデータセットの多様性が不十分である可能性も指摘できる。

以上の結果から、提案したニューラルネットワークモデルは、語彙難易度の推定において有効であると結論づけることができるが、モデルの精度をさらに高めるためには、追加の特徴量の導入やデータセットの拡充が必要であると言える。

## 5 おわりに

本研究では、汎用言語モデル GPT-4 を用いた日本語語彙の難易度分析方法を提案し、その有効性を検証した。日本語の語彙学習は、学習者のコミュニケーション能力や理解度の向上に不可欠であり、特に語彙難易度は学習過程において重要な役割を果たす。語彙の難易度測定は、適切な言語テストの出題や教材開発に不可欠であるが、従来の方法では大規模な測定が困難であった。

本研究では、GPT-4 を使用して単語間の難易度関係を出力させ、その結果を基にニューラルネットワーク機械学習で語彙の難易度を算出する手法を採用した。この手法は、松下他[7]の研究で使用された中国語母語話者にとっての日本語語彙難易度のデータセットを用いて検証され、その結果、既存の調査結果の 99% 信頼区間内に収まるデータの割合は 71% と比較的高いことが示された。これは、従来手法よりも迅速かつ広範囲にわたる難易度評価が可能であることを意味する。本研究では、中国語母語話者にとっての語彙難易度に焦点を当てたが、松下他[7]で実施された語彙テストの分析結果のデータセットに含まれている韓国語母語話者にとっての語彙習得難易度の推定も可能である。また、その他の母語話者にとっての語彙難易度についても、妥当性・信頼性を検証した語彙テストを作成すれば同様に本研究の方法を用いることが可能である。また、今後の課題として、本研究では 25 語をプロンプトに記載して汎用言語モデルに学習させたが、学習の語数の違いがどの程度精度に影響するのか検証する予定である。

## 謝辞

この研究において、佐藤ほか[8]で開発された語彙サイズテストの問題、および松下ほか[7]での実施結果のデータ（ラッシュ項目難度推定値データ）をご提供、また、利用をご許諾くださった千葉大学佐藤尚子先生を代表とする研究チームの先生方に深く感謝いたします。

## 参考文献

- [1] 坪根由香里, 鈴木理子, 阪本史代, 神谷道夫. 学習者から見た効果的な語彙の指導法・学習法アンケート結果より一. 小出記念日本語教育研究会論文集, pp. 107-128, 2001.
- [2] 佐藤浩史, 笠原要, 金杉友子, 天野成昭. 単語親密度に基づく基本語彙の選定. 人工知能学会論文誌, 19(6), pp. 502-510, 2004.
- [3] 水谷勇介, 河原大輔, 黒橋禎夫. 日本語単語の難易度推定の試み. 言語処理学会第 24 回年次大会発表論文集, pp. 670-673, 2018.
- [4] 林妙玉, 白井清昭. 中国語母語話者を対象とした日本語単語の難易度推定. 情報処理学会第 84 回全国大会, 2, pp.75-76, 2022.
- [5] 彭悦, 梁震, 笹尾洋介. 日中バイリンガルの音声版日本語語彙サイズテストの開発と検証. 日本語教育, pp. 93-108, 2023.
- [6] 乾健太郎. ChatGPT の出現は自然言語処理の専門家に何を問いかけているか. 自然言語処理, 30(2) pp. 274-274
- [7] 松下達彦, 佐藤尚子, 笹尾洋介, 田島ますみ, 橋本美香. 第二言語としての日本語語彙量と漢字力—第一言語と学習期間の影響—. 日本語教育, 178, pp. 139-153, 2021.
- [8] 佐藤尚子, 田島ますみ, 橋本美香, 松下達彦, 笹尾洋介. 使用頻度に基づく日本語語彙サイズテストの開発—50,000 語レベルまでの測定の試み—. 千葉大学国際教養学研究, 1, pp.15-25, 2017.
- [9] 松下達彦. 日本語を読むための語彙データベース（オンライン）（2016年11月6日）

<http://www17408ui.sakura.ne.jp/tatsum/database.html#vdri>

- [10] 富嶋正貴. ChatGPT プロンプト研究所. The English Teachers' Magazine, pp. November-11, 2023

## 付録 A プロンプト

あなたは「日本語教育の専門家」です。  
次の「1. 目的の定義」と「2. プロンプト」を元に、「3. 生成するもの」を「#結果」の形式に合わせて生成してください。その際、「4. 参照物」を参考にしてください。

1. 目的の定義：「二つの単語のうちどちらの単語が中国語母語話者にとって難しいのか知りたい。」
2. プロンプト：「二つの単語のうちどちらの単語が中国語母語話者にとって難しいか考えてください。」
3. 生成するもの：
  - 「4. 参照物」の【難しさリスト】の要因を箇条書きで詳しく分析してください。ただし、【難しさリスト】の単語は易しい単語から難しい単語へと並んでいます。
  - 「4. 参照物」の【条件】にある【単語A】と【単語B】の難しさを分析してください。
  - 「4. 参照物」の【条件】にある【単語A】と【単語B】の難しさを分析した結果を箇条書きで出してください。
  - 最後に【単語A】と【単語B】のうちで【難しい方の単語】を表示してください。ただし、【単語A】と【単語B】のいずれかを必ず選択すること。
4. 参照物：
  - 【難しさリスト】

サテライトなど松下（2021）のデータセット中無作為に25語選出した

【条件】  
【単語A】：  
【単語B】：

#結果：  
(1)要因：  
(2)難しさ分析：  
(3)難しさ箇条書き：  
(4)難しい方の単語：

図 3 プロンプトの内容

## 付録 B GPT-4 の出力

#結果：  
(1)要因：

- 発音の違い：中国語の音韻体系と日本語の発音の間の違いによる難しさ。
- 漢字の使用：漢字が含まれる単語は中国語母語話者にとって理解しやすい可能性があるが、読み方や複合語としての意味で難しさが生じる場合もある。
- 語彙の類似性：中国語と日本語に共通の語彙や類似した語彙が難易度に影響する。
- 文化的な背景：単語が持つ文化的な意味合いや使われ方が難しさに影響を及ぼす。
- 単語形成の規則性：複合語や派生語など、単語がどのように形成されているかによる難しさ。

(2)難しさ分析：

- 【単語A】：童心
  - 発音：「童心」の「童（どう）」「心（しん）」は比較的発音しやすく、中国語の音韻とも類似している。
  - 漢字の使用：「童」と「心」は共に中国語にも存在する基本的な漢字であり、認識しやすい。
  - 語彙の類似性：「童心」は「童年的心态」という意味の共通性が認められる。
  - 文化的な背景：この概念は両言語間で理解されやすいと考えられる。
  - 単語形成の規則性：複合語として直感的に理解しやすい構造をしている。
- 【単語B】：アナゴ
  - 発音：「アナゴ」はカタカナで書かれており、発音が日本語特有のものであるため、中国語母語話者にとっては発音が難しい。
  - 漢字の使用：「アナゴ」は漢字で「穴子」と書かれるが、専門的な単語であり、中国語との共通性が少なく理解が難しい。
  - 語彙の類似性：「穴子」という単語に中国語の類似語彙が見つかりづらいため、新規の単語として覚える必要がある。
  - 文化的な背景：日本独特の食品に関連する単語であるため、中国語話者にはなじみが薄い。
  - 単語形成の規則性：動物の名前ということで類推は可能だが、「穴」+「子」という組み合わせは直感に反するため難しさがある。

(3)難しさ箇条書き：

- 「童心」は発音が容易で漢字も基本的、文化的背景も共通して理解しやすい。
- 「アナゴ」は発音も日本語特有、漢字も専門的、文化的背景の共通性も少なく、全体的な難易度が高い。

(4)難しい方の単語：  
アナゴ

図 4 GPT-4 の出力