文章読解中における類義語学習を目的とした質問または説明ベースのマイクロタスクの効果検証

澤野 令 † 香取 浩紀 ‡ 矢谷 浩司 §
東京大学 † 東京大学 ‡ 東京大学 §

1 はじめに

第二言語学習で最も一般的な語彙の誤用問題の一つは, 類義語に関するものである. 不適切な語彙の使用は, 誤解や社会的に不利な結果を引き起こす可能性があるため, 類義語間の意味の違いを学習することは重要である.

一般的に, 読書や会話を通じて文脈から副次的に単語の意味を学習することは, 偶発的な語彙学習と呼ばれる. 偶発的な語彙学習の手法やその効果については, これまでさまざまな研究が行われており, 類義語間の意味の違いの学習に関しても, この方法を活用できると考えられる.

しかし, これらの研究の多くは, 一度きりや数回といった, 少ない回数や短期間での学習効果の検証にとどまっている. また, LLM (大規模言語モデル) の登場により, 学習コンテンツの生成や, フィードバックの提供など, 従来より柔軟かつ多様なタスクの設計が容易になった. 先行研究によると, AI や LLM による情報提示の方法を工夫することで, ユーザーの思考方法や理解度に影響を与えることができると示されているが, これらの知見を基に偶発的な語彙学習のタスクを設計した場合の, 学習効果や認知負荷に関する十分な調査が行われていない.

本研究では, AI や LLM による介入方法に関する先行研究の知見を基に, 文章読解中に登場する単語とその類義語の意味の違いに関して, 提示された説明文を理解する「説明モード」のタスクと, 質問に対して回答する「質問モード」のタスクの 2 種類を設計した. これらのタスクについて, それぞれの学習効果と認知負荷の長期的な変化を, 30 日間の実験を通して被験者内比較で評価した. 本研究の貢献は, 以下の 2 点である.

- 「説明モード」と「質問モード」の学習効果の比較
- 2 つのモードの認知負荷の長期的な変化の比較

2 ユーザスタディ

2.1 システムの概要

実験タスク用の Web サイトを開くと, 学習対象単語がハイライトされた 200 字程度の英文が表示される. 英文を読みながらハイライトされた単語をクリックすると, その単語と類義語 (1 個) の意味の違いに関する「説明モード」または「質問モード」の学習タスクが表示される. 「説明モード」では, あらかじめ GPT4-o によって生成された, その単語と類義語の意味の違いの説明文 (300-400 字) が表示される. 「質問モード」では, その単語と類義語の意味の違いを問う質問文と回答入力欄が表示される. 回答を入力すると, あらかじめ用意された模範回答 (説明モードの説明文と同一) を基に, GPT4-o が生成したフィードバック文が表示され, その後模範回答が表示される.

各タスクを完了すると, 完了したタスクの認知負荷を測定するための NASA-RTLX のモーダルが表示される. タスク期間中は, 毎日異なる英文が表示され, 英文内の全てのタスクを完了すると, 1 日の作業が終了する. 各学習対象の単語の学習タスクは, タスク期間を通じてそれぞれ 5 回ずつ登場する. また, 各参加者ごとに 36 個の学習対象の単語と類義語のペアを 18 個ずつランダムに二群に分け, 一方は常に「説明モード」, もう一方は「質問モード」のタスクが表示される.

2.2 ユーザスタディの流れ

実験参加者は, 本学の倫理審査申請の承認後，Crowd-Works を通じて募集した. 募集条件は, 18 歳以上 79 歳以下で, CEFR (ヨーロッパ言語共通参照枠) B1 以上の英語力を有していることとした. ユーザスタディは, 事前テスト, タスク期間 (30 日), 事後テストの順番で進められ, 事後テストまで完了した参加者は 22 人であった. またその中から無効と判断されたデータを除外し, 最終的に 20 人の参加者のデータに対して分析を行った. 謝礼として, 事前・事後テストそれぞれにつき 600 円, タスク 1 日につき 100 円を支払った.

事前・事後テストでは, Google フォーム上で 36 個の学習対象の単語とその類義語間の意味の違いを, 短文形式で記述する問題に回答してもらった. 事前・事後テストの採点基準は, あらかじめ生成した単語と類義語の意味の違いの説明文 (タスク期間中に使用したものと同一) を基に設定した. 具体的には, 研究実施者が説明文から意味の異なる要素を抽出し, その要素が回答にどれだけ含まれているかを基準に採点を行った. その後, 各参加者ごとに, 「説明モード」に属するペアの回答の正答率,「質問モード」に属するペアの回答の正答率を算出した. 正答率は, 各モードの回答全体で含まれていた要素数を, 各モードの回答全体で含まれるべき要素数 (模範回答全体で含まれる要素数) で割った値とした.

Investigating the Effect of Questioning-Based and Explaining-Based Microtasks for Learning Synonyms During Reading Comprehension

† Rei Sawano, The University of Tokyo
‡ Hiroki Katori, The University of Tokyo
§ Koji Yatani, The University of Tokyo

3 結果

3.1 学習モード間の学習効果の比較

事前テストと事後テストの正答率の変化に対し, 学習モードを変数とした対応のある t 検定を行った結果, モード間に有意差が確認された ($t(19) = -2.65$, $p = 0.0157$). また効果量 (Cohen's d) は -0.242, 平均値の差は 0.038 となり,「質問モード」の方がわずかに高い結果になった.

3.2 学習モード間の認知負荷の長期的な変化の比較

認知負荷の 7 つの指標 (知的・知覚的要求, 身体的要求, タイムプレッシャー, 作業成績, 努力, フラストレーション) に対し, 学習モード (説明モードと質問モード) と学習回数 (1-5 回) の 2 つを説明変数として分析を行った.

「知的・知覚的要求」に関しては, 各群に正規性が確認されたため対応のある二元配置分散分析を行った結果, 学習モードの主効果が有意であった. ($F(1,72) = 9.09$, $p = 0.0074$). また, 平均値の差は 7.380 となり,「質問モード」の方が高いという結果になった. また, 学習回数の主効果も有意であった ($F(4,72) = 6.115$, $p = 0.0003$). また, 交互作用効果は認められなかった ($F(4,72) = 2.004$, $p = 0.103$). その後学習回数に関して, Tukey の HSD 検定で各条件間のペアワイズ比較を行ったが, 有意差は確認されなかった.

「タイムプレッシャー」に関しては, Friedman 検定を行い, 有意差が確認された ($p < 0.0001$). その後, Wilcoxon 符号順位検定を用いてペアワイズ比較を行い, 同一の学習モード間では有意差は確認されなかった. 一方, 学習回数が同じ (1 回および 4-5 回) 場合, 学習モード間で有意差が確認され, 「質問モード」の方が高いという結果になった.

「作業成績」に関しては, 各群に正規性が確認されたため, 対応のある二元配置分散分析を行った結果, 学習モードの主効果が有意であった ($F(1,72) = 16.051$, $p = 0.0008$). また, 平均値の差は 9.522 となり, 「質問モード」の方が高いという結果になった. また, 学習回数の主効果も有意であった ($F(4,72) = 7.982$, $p < 0.0001$). また, 交互作用効果は認められなかった ($F(4,72) = 1.906$, $p = 0.118$). その後, 学習回数に関して, Tukey の HSD 検定で各条件間のペアワイズ比較を行った結果, 有意差は確認されなかった.

「努力」に関しては, Friedman 検定を行い, 有意差が確認された ($p < 0.0001$). その後, Wilcoxon 符号順位検定を用いてペアワイズ比較を行い, 同一の学習モード間では,「説明モード」の 1 回目と 3 回目, 1 回目と 4 回目のみに有意差が見られ, 1 回目の方が高いという結果になった. 一方, 学習回数が同じ (1-5 回) 場合, 学習モード間で有意差が確認され, 「質問モード」の方が高いという結果になった.

「フラストレーション」に関しては, Friedman 検定を行った結果, 有意差が確認された ($p < 0.0001$). その後 Wilcoxon 符号順位検定を用いてペアワイズ比較を行った結果, 同一の学習モード間で有意差は確認されなかった. 一方, 学習回数が同じ (1-5 回) 場合, 学習モード間で有意差が確認され, 「質問モード」の方が高いという結果になった.

4 考察

4.1 学習モード間の学習効果の比較

学習効果に関しては,「質問モード」の方が高いという結果が得られ, 認知負荷が高い方が学習効果を促進するという先行研究 [1] の知見と一致していた. しかしながら, 両者の差は小さいという結果となった. その原因として, 提示された説明文や模範回答の中には, 文面だけでは深い理解を十分に促せないコンテンツが含まれていた可能性が考えられる. このため, アウトプットの有無に関わらず, 学習が一定の段階に達すると, それ以上の学習内容の定着が十分に進まなかった可能性が考えられる.

4.2 学習モード間の認知負荷の長期的な変化の比較

認知負荷に関しては,「努力」を除き, 学習回数を重ねても大きな変化は見られなかった. その原因として, 以下の 2 点が考えられる. 一つ目は, 提示された単語と類義語の違いの説明文や模範回答が, 300-400 字程度と比較的長文であり, 一度に全てを理解したり記憶するのが難しかった可能性である. そのため, 最初は概要を理解し, 回数を重ねると詳細を理解しようとする段階的な学習が行われ, 新たに学習する情報量が比較的一定に保たれた結果, 認知負荷に大きな変動がなかったと考えられる. 二つ目は, 先に述べたように, 説明文の中には文面だけでの理解や記憶が難しい内容が含まれていた可能性である. このため, その部分に関しては学習回数を重ねても理解が進まず, 結果的に認知負荷の変化が抑えられたと考えられる.

学習回数によって変化が見られたのは「説明モード」における「努力」のみであった. この理由として, 参加者が「説明モード」のタスクを繰り返す中で, 説明文のフォーマットに慣れ, 自身にとって効率的な読み方を確立した可能性が考えられる. その結果, 学習内容の情報量や難易度には変化がないが, 努力感のみが低下したと推測される.

また学習モード間の比較では,「知的・知覚的要求」,「タイムプレッシャー」,「努力」,「フラストレーション」に関して,「質問モード」の方が高いという結果となった. これは単に説明文を読むより, アウトプットした方が認知負荷が高くなるという予想通りの結果だった.「作業成績」に関しては, 質問モードの方が高い (つまり悪い) という結果が得られた. これはアウトプットを行うことで, 自分が理解していない点に気づき, 自分の回答や理解度への満足感が低下したためと考えられる. 以上の知見を基に, より効果的な学習タスクの設計が今後の課題として考えられる.

謝辞

実験参加者の方には, 長期間の実験へのご協力, 研究室の皆様からは研究に関連する助言を多数いただきました. この場を借りて, 感謝申し上げます.

参考文献

[1] Hulstijn et al.: *Language learning*, Vol. 51, No. 3, pp. 539–558 (2001).