様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ５ 日現在

機関番号：３６３０１
研究種目：研究活動スタート支援
研究期間：2020～2021
課題番号：２０Ｋ２１９７７

研究課題名（和文）学習者コーパスと自然言語処理技術を活用した包括的な誤用分析の試み

研究課題名（英文）An attempt at a comprehensive error analysis using learner corpus and natural language processing techniques

研究代表者
西村　嘉人（NISHIMURA, Yoshito）

松山大学・経済学部・講師

研究者番号：００８８２４３２

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では、英語学習者の英作文を対象に、自然言語処理の技術を活用することで、大規模かつ包括的な誤用分析を行った。具体的には、サンプルサイズの大きい学習者コーパスに収録されている英作文を使用し、どのような熟達度の英語学習者が、どのような誤用を犯しやすいのかを体系的に記述することを目的にした。さらに、文法的誤用及び語彙的誤用とCommunicative Adequacyの関係も探求し、これらの関係性について考察した。

研究成果の学術的意義や社会的意義
熟達度が高い英語学習者が犯しやすい誤用項目と熟達度が低い英語学習者が犯しやすい誤用項目を特定したり、熟達度に関わらずどのような項目において英語学習者が誤用を犯しやすいのかを特定したり、また、どのような誤用が意思伝達を阻害したり、しなかったりするのかを特定することで、効果的な学習の指導や教材開発へ貢献することが可能となる。

研究成果の概要（英文）：The objective of the current study was to address a comprehensive error analysis by employing natural language processing techniques (ERRANT) and a learner corpus (ETS Corpus of Non-Native Written English) to detect and classify an L2 learner's all possible syntactic and lexical errors and by coding communicative adequacy to each sentence, and thereby find out which errors L2 learners are more likely to make, what error categories they make, and whether the higher the proficiency level the higher the communicative adequacy.

研究分野：応用言語学

キーワード：誤用分析　誤用検出　学習者コーパス　ライティング　Communicative Adequacy

## １．研究開始当初の背景

学習者の誤りを分析することは、第二言語習得研究では重要な研究分野の一つである。誤用分析には、エラータグを付与した学習者コーパスが活用されることが多い。それと同時に、サンプルサイズが極めて限定されていたり、エラータグ付けに一貫性がなかったり、生起するエラーが限定されていたりするといった問題も存在する。また、従来の誤用分析では、特定の文法及び語法のみを対象にした研究が多く、包括的な誤用分析の結果に基づいた体系的な誤用傾向の記述が蓄積されているとは言い難い。

誤用分析において障壁となるのは、エラータグの付与である。自然言語処理の技術革新が目覚ましい現代においても、学習者の英作文を入力するとエラーを検出し、誤りを訂正した正確な文が自動的に出力されるような、誤用検出から訂正までをすべて自動化したプログラムは開発されていない。しかしながら、英語学習者の作文に対応した英語母語話者による添削文が付与されていれば、それら二文を比較することで、自動的にエラーを検出することが可能なプログラムは既に開発されている。それは、ERRANT（Bryant et al., 2017; Felice et al., 2016）である。ERRANT は、学習者の英作文と英語母語話者の添削文をもとに、文法的及び語彙的誤用を自動的に検出し分類できる（精度は 90%以上）。つまり、ERRANT を活用することで、エラータグ付与にかかる莫大な人的及び時間的コストを削減することが可能となる。このように、第二言語習得研究において、誤用分析を行うための必要なコストは、低減しつつある。また、自然言語処理の技術革新により、ユーザーフレンドリーな構文解析装置や誤用検出プログラムは、無償で手に入る時代に突入している。このような時代背景は、多大な人的・時間的コストの削減が期待でき、誤用分析のような研究を加速させる。

## ２．研究の目的

本研究では、英語学習者の英作文を対象に、自然言語処理の技術を活用することで、大規模かつ包括的な誤用分析を行う。具体的には、英作文において、どのような熟達度の学習者が、どのような誤用を犯しやすいのかを体系的に記述することを目的とする。これまでの研究では、活用できるサンプル数の制限や、分析を行うための人的資源及び技術的な問題が原因で、包括的な誤用分析を行うために乗り越えなければならない障壁が少なからず存在していた。しかしながら、AI を活用した自動採点システムの開発や誤用の自動検出システムの開発など、自然言語処理分野の急速的な技術革新によって、これまでの障壁が緩和されつつある。また、世界最大規模の学習者コーパス（ETS Corpus）の整備も同時に進んでおり、包括的な誤用分析を行うための土壌は既に完成しつつある。本研究では、ETS Corpus と学習者の英作文から文法的及び語彙的誤用を自動的に検出することが可能な自然言語処理プログラムを活用する。これにより、英語学習者の英作文における誤用傾向について、大規模な実証データに裏付けられた頑健かつ体系的な記述が可能となる。本研究を遂行する上で、誤用分析における人的・時間的コストの削減に取り組み、誤用分析の自動化手順を確立する。

すなわち、本研究の目的は、学習者コーパスの英作文に対して英語母語話者による添削文を付与し、ERRANT を活用することで誤用の検出と分類を自動化し、その自動化手順を確立した上で、大規模な実証データに基づいた英語学習者の誤用傾向を体系的に記述することである。また、学習者の英作文一文ごとに、英語母語話者及び英語学習者双方による主観的な理解度を付与することで、単に誤用傾向を分析するだけでなく、Communicative Adequacy の観点からも英作文を評価し、どのような誤用がコミュニケーションを阻害しうるか、あるいは阻害しないのかを明らかにすることを目的とする。

## ３．研究の方法

本研究は、学習者コーパスと自然言語処理技術を活用した包括的な誤用分析を試みた。データは、ETS Corpus of Non-Native Written English を使用した。ETS Corpus は、英語学習者の英作文データが 12,100 ファイル収録されている世界最大級の学習者コーパスである。英語学習者の母語は、日本語や韓国語、中国語、スペイン語など合計で 11 に達する。ETS Corpus は、それぞれの母語を持つ英語学習者から、1,100 ファイルずつ収録されている。本研究では、日本語を母語とする英語学習者 240 ファイル分を分析対象とした。また、英作文の内容に応じて三段階の評定（Low、Medium、High）もファイルごとに付与されている。本研究では、この評定を熟達度の代理指標とした。

文法的及び語彙的な誤用頻度を算出するために、学習者の英作文に対応するようにパラレルに英語母語話者による添削文の付与を行うことで、ERRANT での分析を行った。添削文の付与は、テキストファイルを専用のフォーマットに予め整形したものを使用した。添削文を付与する際には、学習者が書いた文の統語構造を維持したまま最小限の修正を行うように指示を行った。学習者の英作文と添削文を比較することで、Replacement、Missing、Unnecessary の 3 つのエラー

カテゴリーごとに文法的及び語彙的誤用を算出した。

また、先に述べた Communicative Adequacy の評定については、学習者の英作文を添削する英語母語話者と、対象となる英作文を執筆していない英語学習者による評定の 2 つの評定を付与した。評定は、先行研究をもとに、0 から 3 の 4 段階評価とした。また、これら評定は Weighted Clause Ratio のような重み付けを分析の際に行った。英語母語話者及び英語学習者による Communicative Adequacy の評定基準は、以下の表のとおりである。

表 1. Communicative Adequacy の評定基準

| **Rating** | **Quality of original sentence** |
|---|---|
| 0 | Original sentence is well written, and no revision required. |
| 1 | Meaning of original sentence is clear, and minimal editing is required. |
| 2 | Meaning of original sentence is somewhat understandable, and some level of editing is required to convey the meaning. |
| 3 | Meaning of original sentence is unclear or original sentence requires extensive syntactic changes and you've input *NG* under **%NTV.** |

## ４．研究成果

### （１）全体的な傾向

全体的な傾向としては、学習者の熟達度が高いと誤用の数が必ずしも少なくないことが明らかになった。熟達度の高い学習者は、複雑な構文や単語を使用する傾向にあり、むしろ他の学習者よりも誤りの数が多い場合があることが明らかになった。反対に、熟達度が低い学習者は比較的簡単な構文や単語を使用したり、文自体の長さが短かったりするため、誤りの数が熟達度の高い学習者よりも少なくなる学習者も存在した。Communicative Adequacy の観点では、英語母語話者と英語学習者による評定の相関は比較的高いことが明らかになった。同じ母語をもつ英語学習者による評定では、0 や 1 が付与されている文であっても、英語母語話者による評定では 2 や 3 が付与される場合も少ないながらも存在した。

### （２）文法的誤用及び語彙的誤用と評定の関係について

評定の程度に関わらず、文法的な誤用と評定には顕著な関係は見られなかった。もちろん、文法的誤用が多い英作文は評定が低くなる傾向にはあるが、文法的誤用が多い英作文であってもある程度語数があり、段落構成が明確な場合は、三段階のうち最低評価を得る場合は少なかった。英作文を評価する場合、内容に重点を置くのか、構造及び論理展開に重点を置くのか、文法及び語彙面の使用法に重点を置くのか、語数に重点を置くのかによっても評価が割れるため、一概の結果を一般化することはできない。語彙的誤用と評定の関係についても文法的誤用と評定の関係と同様の傾向が見られた。

### （３）文法的誤用及び語彙的誤用と Communicative Adequacy の関係について

先に述べたように、英語母語話者の評定と英語学習者による評定の関係は、似たような傾向が示された。すなわち、英語母語話者が Communicative Adequacy が高いと評価した英作文は、英語学習者も同様に高い評定を付与している。また、その逆も然り。ただし、面白いことに、英語母語話者の評定が低い場合であっても、英語学習者がその母語特有の背景知識を駆使して文意を読み取り Communicative Adequacy を高く評定したケースも存在した。本研究のサンプルサイズでは、到底一般化できるほどのデータが得られていないので、今後の課題として日本語以外のも他の母語でもこのような現象が見られるのか探求していく必要がある。

## 参考文献


Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In R. Barzilay & M-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 793–806). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1

Felice, M., Bryant, C., & Briscoe, T. (2016). Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 825–835). The COLING 2016 Organizing Committee.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|