学習者作文における形態素解析の課題と対応

第10章｜学習者作文における形態素解析の課題と対応

日本語学習者の作文に品詞等の形態論情報が付与されていれば、その作文の特徴を量的に分析することが可能となる。しかし、学習者の作文には母語話者の作文と異なり、誤用や母語由来の表現が含まれているため、単に既存の形態素解析器を用いるだけでは誤解析が多発する傾向にある。本章では、『日本語学習者縦断作文コーパス（W-CoLeJa）』構築の事例を通じ、誤解析の発生要因を明らかにし、実用的な対応手法の1つである「ユーザ辞書構築」について述べる。

KEYWORD
｜**学習者コーパス**｜**形態論情報**｜**誤解析**｜**誤用**｜**ユーザ辞書**｜

□1　はじめに

日本語教育現場で働く教師なら、学習者の言語能力の成長を正確に把握したいと思うことがあるだろう。「この学習者は前回の作文と比べて語彙が豊富になったのか」「使える文法項目は増えているのか」といった疑問に対し、従来は教師の経験と直感に頼ることが多かった。しかし、このような主観的な評価だけでは、学習者の言語変化を体系的に分析することには限界がある。また、多数の学習者のデータを扱う場合、人手による分析では時間的制約もあるため、大量のテキストデータから言語的特徴を効率的に抽出することは困難である。このような状況に対する解決策の一つが**形態素解析**という技術である。形態素解析とは、専用の解析ツール（形態素解析器）を用いて、文章を**形態素**と呼ばれる意味を持つ最小の単位に自動的に分割し、それぞれに品詞や活用形などの詳細な情報を付与する処理のことである。たとえば、「学生が図書館に行きました」という文は、次のように解析される。

学生（学生：名詞）｜が（が：助詞）｜図書（図書：名詞）｜館（館：接尾辞）｜に（に：助詞）｜行き（行く：動詞）｜まし（ます：助動詞）｜た（た：助動詞）

形態素解析を行えば、PCを用いて大量の作文データから特定の語彙や文法項目を瞬時に抽出し、統計的に分析することが可能になる。このように大量のテキストや発話を収集してデータベース化した言語資料を**コーパス**と呼ぶ。コーパスを活用することで、個別の事例では見えてこない全体的な傾向や規則性を発見することができる。

では、このような形態素解析をコーパスに適用することで、具体的に何ができるのだろうか。まず、語彙や文法の使用状況を定量的に把握できる。語彙については、使用頻度や多様性を数値化でき、異なるレベルの学習者間で比較できる。文法面では、助詞の使用パターンや動詞活用の正確性を分析することにより、習得状況や困難点を明らかにできる。さらに、同一学習者の時系列データを分析することで日本語能力の成長過程を追跡でき、母語の異なる学習者グループ間の比較を通して、語彙・文法使用の差異や母語の影響の現れ方を実証的に検討することができる。

こうした量的分析により、語彙指導では、学習者が実際によく使用する語彙と、使用頻度の低い語彙を区別することで、より効率的な指導順序を決定でき、文法指導でも、学習者が頻繁に誤る項目を客観的に特定し、重点的な指導を行うことが可能となる。

標準的な日本語の表記法や語彙使用パターンに基づいて開発された形態素解析器は、『現代日本語書き言葉均衡コーパス（BCCWJ）』[1]のように母語話者が産出した一般的な書籍や新聞記事などを対象とする場合、非常に高い精度を達成しており、実用的なレベルにある。しかし、こうした解析器は学習者コーパスを解析する際には、精度が大きく低下するという重要な問題がある。学習者の作文には、語彙や文法の誤用、表記の揺れ、ひらがな表記の多用など、標準的な日本語テキストでは見られない様々な現象が含まれており（迫田 2020; 横野 2024）、これらが形態素解析器にとって「想定外」の入力となるためである。その結果、正確な処理が困難となる。たとえば、学習者が「うれしかった」を「うれしじゃった」と誤って記述した場合、形態素解析器は本来一つの形容詞の活用形として扱うべき表現を、意味的に関連のない複数の語として認識してしまうことがある。また、「容量」をひらがなで「ようりょう」と書いた場合、「よう」と「りょう」という2つの独立した語として誤って分割される可能性がある。

このような背景から、本章では『日本語学習者縦断作文コーパス（W-CoLeJa）』の構築過程で得られた経験を基に、学習者コーパスにおける形態素解析の課題

1| https://clrd.ninjal.ac.jp/bccwj/

と実用的な対応策について詳しく解説する。特に、技術的な専門知識がなくても理解できるよう、具体的な事例を用いながら、学習者の作文が形態素解析の際にどんな問題を起こすのか、また、どうすればその問題を解決できるのか、問題点の整理とその解決方法を説明する。

たとえば、「授業で集めた作文、PCで分析してみたいけど、何から始めれば…」「形態素解析にかけてみたら、なんだか変なところで区切られてしまった…」「たくさんの作文から表現の傾向を調べたいのに、どう整理すればいいんだろう…」といった声は、現場でしばしば聞かれる。本章は、このような問題に直面する日本語教師や研究者、学部生・大学院生を主な読者として想定している。ただし、対象はそれに限られない。たとえば、日本語そのものに関心を持ち、自分の書いた文章を分析してみたいと考える人も含まれるだろう。文章を形態素解析器にかけてみれば、自身の語彙や表現の使い方の傾向が見えたり、時間を追って比較することで文章の変化や成長を確認できたりする。こうした場合に有効なのが、小規模なコーパスを自ら構築することである。ここでいう「小規模なコーパス」とは、研究機関が公開する大規模コーパスとは異なり、個人や小グループが特定の目的のために数十から数百件程度の作文を電子化し、検索や集計ができるように整備したデータベースを指す。本章で紹介する具体的な事例や対応策は、そのような小規模なデータを実際に処理・活用しようとする際に役立つものである。一方で、形態論情報の設計や付与基準の策定、誤解析要因の体系的分析、修正作業の効率化や自動化といった課題は、学習者コーパスを本格的な言語資源として整備するために不可欠である。こうした、より専門的・方法論的な議論については、本章と同じ筆者が執筆した朱ほか（2026予定）に譲ることとする。そこでは主に大規模な学習者コーパスを設計・開発する研究者や技術者を想定しているが、学習者データを精緻に分析するために、コーパス内部の構造や付与基準を深く理解する必要がある研究者にとっても、有用な指針となるだろう。本章が学習者データ活用のための実践的な入口であるとすれば、朱ほか（2026予定）は言語資源の基盤整備に関わる専門的な探究を進めるための道しるべとなる。

なお、本章で報告する作業は、国立国語研究所第4期機関拠点型基幹研究プロジェクト「日本語学習者の作文の縦断コーパス研究」の一環として実施された。本プロジェクトは、海外の大学で日本語をゼロから学び始めた学習者の4年間にわたる作文能力の発達過程を明らかにすることを目的とし、日本語作文とその母語訳、執筆メモ、執筆過程などを収録したW-CoLeJaを構築している。

本章で扱うのは、このコーパスにおける形態論情報修正作業（形態素解析結果の人手修正と誤解析パターンの分析、自動処理の実装）である。また、プロジェクトリーダーである石黒が本作業の全体方針を立案し、自然言語処理研究者である横野と相談しつつ、李・朱が形態素解析結果の人手修正と誤解析パターンの分析を、工藤が自動処理の実装を担当した。本章の執筆は主として李が行い、各担当者による検討結果を統合して報告する。

□2　学習者作文における誤解析の実態

実際に学習者の作文に形態素解析[2]を適用すると、どのような問題が起きるのだろうか。W-CoLeJaには2025年時点で7,939本の作文が含まれているが、そのうち約半数の4,137本について、日本語学・日本語教育学を専門とする作業者が形態素解析の結果をチェックした。その結果、学習者の作文には大きく分けて3つのパターンで形態素解析の誤り、すなわち誤解析が起きることがわかった。

2.1　辞書未登録語による誤解析

第一の問題は、形態素解析器の辞書に登録されていない言葉が含まれることである。形態素解析器は、あらかじめ用意された辞書に基づいて文章を解析する。辞書未登録語の問題は母語話者のテキストでも生じるが、学習者の作文では特に顕著となる。その理由は、学習者が自国の文化や人物について言及する際に、日本語の辞書には収録されていない固有名詞を頻繁に使用するためである。

例（1a）は、「ins」と「sns」が辞書未登録のため、固有名詞としてではなく文字ごとに分割されて解析された事例である。特に「ins」は中国語圏でInstagramの略称として用いられるが、日本語では一般的でなく、学習者が母語の略称を持ち込んだ例と考えられる。以下、（a）は誤解析例、（b）は適切な解析結果を示すものとする。

（例 1）[3] 辞書未登録の英字略称による誤解析の事例

（1a）そうなれば、｜ i（I：記号）｜ n（N：記号）｜ s（S：記号）｜また｜ s（S：

2｜ 形態素解析には、後述の3.1節と同一の設定（MeCab [ver.0.996] およびUniDic [ver.202302]）を用いた。

3｜ 本章では、学習者作文における形態素境界の表示方法として、システム出力時の境界を「｜」、修正後の形態素境界アノテーションを「／」で示す。また、（ ）内には「語彙素：品詞」の形式で語の情報を記すこととする。

記号）｜ n（N：記号）｜ s（S：記号）」に写真をアップロードする前に、写真を加工しなければなりません。
（1b）そうなれば、／ ins（INS：固有名詞）／また／ sns（SNS：固有名詞）／に写真をアップロードする前に、写真を加工しなければなりません。
（CCK004-A3-03-09）

同様に、「この英雄の名前は钟南山である」という文でも、中国の医師の人名「钟南山」が「钟南（未知語）｜山（名詞）」という2つの語に誤って分割されている。

（例 2）辞書未登録の人名による誤解析の事例
（2a）この英雄の名前は｜钟南（unk[4]：名詞）｜山（山：名詞）｜です。
（2b）この英雄の名前は／钟（钟：人名 - 姓）／南山（南山：人名 - 名）／です。
（CCC014-A2-01-02）

この問題は各国の学習者が自国の文化について書くときによく起こる。韓国語母語話者では「BTS」「BLACKPINK」といったK-POPグループ名、ベトナム語話者では「スビン・ホアンソン」「ホーおじさん」といった人物名、タイ語話者では「ソンクラーン」といった祭典名など、各文化圏に特有の表現が頻繁に登場する。これらの多くは日本語の標準的な辞書に収録されておらず、適切に解析されない。さらに現代的な製品名やサービス名も問題となる。「アイフォン（iPhone）」「ウィーチャット（WeChat）」「エアーポッズ（AirPods）」は英語表記が一般的であるが、日本語表記された場合には誤解析の原因となる。背景には、新しい語彙や語形の出現に辞書更新が追いつかないということが考えられる。

2.2　ひらがな表記の多用による誤解析

第二の問題は、ひらがなで書かれた言葉の処理である。学習者は、適切な漢字を知らない場合、忘れてしまった場合、確信が持てない場合など、様々な理由でひらがな表記を選択する。しかし、ひらがなが続くと、形態素解析器は語の境界を特定することが困難になる場合がある。「写真をとるのは早くて、かんたんで、時間やデータようりょうが少しかかるから」という文では、「容量」をひらがなで「ようりょう」と書いたため、「様（接尾辞）｜リョウ（名詞）」という2つの別々の語に分割されてしまっている。

4｜“unk”は形態素解析において未知語と判定された語を表す。

(3) ひらがな表記による誤解析の事例
(3a) 写真をとるのは早くて、かんたんで、時間やデータ｜よう（様：接尾辞）｜りょう（リョウ：名詞）｜が少しかかりますから。
(3b) 写真をとるのは早くて、かんたんで、時間やデータ／ようりょう（容量：名詞）／が少しかかりますから。 （VVA025-A3-01-03）

次のような用例もある。「ちちははははプレゼントをありますという」という文では、学習者が「父母は」という意味で書いたが、ひらがな表記のため形態素解析器は「ちち（副詞）｜ははは（感動詞）」という解析結果を出してしまった。このような場合、「父（名詞）」「母（名詞）」「は（助詞）」といった個別の語が正しく認識されず、文全体の語彙分析が不正確になる。

(4) ひらがな表記の連続による誤解析の事例
(4a) ちち（ちち：副詞）｜ははは（ははは：感動詞）｜プレゼントをありますという。
(4b) 父（父：名詞）／母（母：名詞）／は（は：助詞）／プレゼントをありますという。 （CCK029-A1-01-01）

2.3 誤用による誤解析

第三の問題は、学習者による語彙や文法の間違いそのものである。学習者が間違った活用をしたり、表記を間違えたりすると、辞書にない語形ができてしまい、これが形態素解析の間違いを引き起こす。「一番うれしじゃった」という文では、学習者が「うれしかった」を「うれしじゃった」と間違って活用したため、形態素解析器は「うれ（代名詞）｜為る（動詞）｜ちゃう（助動詞）｜た（助動詞）」という複数の語として分析してしまった。本来は「嬉しい（形容詞）｜た（助動詞）」として扱うべき表現である。

(5) 活用の誤りによる誤解析の事例
(5a) 一番｜うれ（うれ：代名詞）｜し（為る：動詞）｜じゃっ（ちゃう：助動詞）｜た（た：助動詞）｜の（の：助詞）｜です（助動詞）。
(5b) 一番／うれしかっ（嬉しい：形容詞）／た（た：助動詞）／の（の：助詞）／です（助動詞）。 （KKA010-A1-01-01）

表記の間違いも問題となる。「皆さんにとって嬉しいプレセントは何ですか」という文では、「プレゼント」を「プレセント」と書き間違えたため、「プレ-pre（名詞）｜セント-saint（名詞）」という2つの語に分割されている。

（6）表記の誤りによる誤解析の事例
（6a）皆さんにとって嬉しい｜プレ（プレ -pre：名詞）｜セント（セント -saint：名詞）｜は何ですか。
（6b）皆さんにとって嬉しい／プレゼント（プレゼント -present：名詞）／は何ですか。（TTA003-A1-02-04）

2.4　誤解析の影響

これまで見てきたように、W-CoLeJaにおける誤解析は、主に「辞書未登録語」「ひらがな表記の多用」「誤用」という3つの要因から生じることが確認された。これらの間違いは、データ分析において看過できない影響を与える。形態素解析の誤りが最も直接的に影響するのは、語彙使用の実態把握である。学習者がどのような語彙をどの程度使用しているかという基本的な情報が、解析の誤りによって歪められてしまう。

学習者が「うれしかったプレゼント」について書いた作文を例に考えてみよう。もし「プレゼント」が「プレ」と「セント」に誤って分割されると、学習者が「プレゼント」という語彙を使ったことを正しく記録できない。代わりに、「プレ」と「セント」という2つの語が記録される。これでは、学習者が実際にどのような語彙を使っているかを正確に把握することができない。同じように、「容量」が「よう」と「りょう」に分かれてしまえば、学習者がこの語彙を使いこなしていることがデータに反映されない。

このような間違いが積み重なると、語彙数の計算も不正確になる。本来100語の語彙を使った作文が、誤解析により異なる語彙数として計算されてしまう。逆にいえば、複数の語がひとまとまりとして誤認識されれば、実際より少ない語彙数として記録される。これでは、学習者の語彙力を正しく評価することはできない。

文法の分析においても、誤解析の影響は無視できない。「うれしじゃった」という学習者の誤用が「うれ（代名詞）｜し（動詞）｜じゃっ（助動詞）｜た（助動詞）」として解析されてしまうと、そこには形容詞の活用で困難を抱えているという重要な情報が隠れてしまう。本来なら「この学習者は形容詞の過去形に

課題がある」という教育的に価値のある発見があるはずなのに、それが見えなくなってしまう。助詞の使い方、動詞の活用、文の構造など、文法のあらゆる側面で同様の問題が起こりうる。

このように、形態素解析の誤りは、学習者の言語能力を正確に把握し、効果的な教育方法を開発するための分析そのものに影響を与える問題である。

□3　問題を解決するための工夫

上述した問題をどのように解決すればよいだろうか。W-CoLeJaでは、限られた時間と人手という制約の中で、形態素解析結果の誤りを人手で修正する作業を、より効率的に進める方法を検討した。作文の総数は7,939本であり、人手による一つ一つの確認・修正作業自体は不可能ではないが、すべてを同じ精度で処理すると膨大な時間を要する。そこで、現実的な作業効率を考慮し、段階的なアプローチを採用した。

3.1　形態論情報修正作業のワークフロー

W-CoLeJaの形態論情報整備の工程を図1に簡潔に示す。

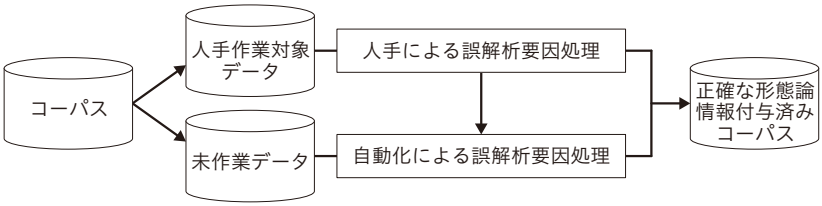

図1　W-CoLeJaの形態論情報整備のワークフロー

このアプローチの基本的な考え方は、全体の一部を分析することで得られた知見を、残りの部分に効率的に適用するというものである。まず、全体の約半数の作文を「人手作業対象データ」として選定し、これらに対して既存の形態素解析器[5]による初回解析を実施した。この初回解析により、各作文に対して形態素の境界、品詞情報、語彙素などの基本的な形態論情報が自動的に付与さ

5| 具体的には、UniDic辞書（ver. 202302）を用いた形態素解析器MeCab（ver. 0.996）を使用した。UniDic辞書は、国立国語研究所が提供する電子化辞書であり、規定された斉一な言語単位（短単位）と階層的見出し構造に基づいて開発されている。

れる。その後、日本語学・日本語教育学を専門とする複数の作業者が、この初回解析結果をチェックし、誤解析が発見された箇所に対して分析情報を記録した。「人手作業対象データ」の分析結果から誤解析の主要パターンを抽出し、自動修正のためのパターンとユーザ辞書を構築する。構築されたパターンと辞書を用いて、残余の未作業データを処理する。最後に、人手修正と自動修正を経たデータを統合し、改良されたユーザ辞書とパターンを用いて再度形態素解析を実施する。この過程により、元の解析結果と比較して誤解析を減らした学習者コーパスが構築される。なお、本章の目的は、技術的な専門知識がなくても形態素解析の精度を向上できる方法の紹介でもあるため、自動修正により誤解析パターンを処理する方法に関する詳細については割愛する。「自動修正」に関心のある方は、朱ほか（2026予定）を参照していただきたい。

人手による分析では、以下の8つの項目を設定し、それぞれの誤解析について体系的に情報を収集した。

［1］**【誤解析チェック】**形態素の境界が正しく認識されていない箇所を特定する。該当する行に「1」を記入し、後の作業で重点的にチェックすべき箇所を明確にする。

［2］**【修正案】**間違った解析結果を正しく直した形を記録する。たとえば「プレ｜セント」と誤って分かれた箇所には「プレゼント」という正しい語形を記入し、本来あるべき解析結果を示す。

［3］**【結合ID】**分かれてしまった語をまとめるための番号を付与する。「て｜つ｜じん」のように3つに分かれた「哲人」には、すべて同じ番号（例：10）を割り当てて、これらが本来一つの語であることを示す。番号は10、20、30のように10刻みで付与し、作業ミスを防ぐ。

［4］**【分割】**1つにまとまってしまった語を複数の語に分ける必要がある場合を示す。本来は複数の語に分かれるべきなのに、一語として誤認識された箇所に「TRUE」と記入する。

［5］**【辞書追加候補】**辞書に載っていない語で、今後辞書に追加すべきものを示す。学習者の母国の人名や地名、新しい製品名など、「unk（未知語）」として表示された語や、不適切に分割された固有名詞に「TRUE」を記入する。

［6］**【母語表現】**学習者が自分の母語をそのまま使っている箇所を分類する。文中に「钟南山」のように母語の語彙が混じっている場合は「本文」、漢詩や諺など長い引用の場合は「引用」と記入する。

[7]**【保留中】**意味がよくわからず、後で確認が必要な箇所を示す。文脈だけでは学習者の意図が判断できず、母語で書かれた作文と照らし合わせて意味を確認する必要がある箇所に「TRUE」を記入する。

[8]**【備考】**作業中に気づいた重要な情報を自由に記録する。辞書追加候補の語がどんな意味なのか、保留中の箇所でどんな問題があるのかなど、後の作業に役立つメモを記入する。

実際の作業例を図2に示す。図2では、初回の形態素解析によって自動的に付与された基本情報（書字形、語彙素、語彙素読み、品詞等）に加えて、人手による8項目の分析結果が記録されている様子を確認できる。

誤解析チェック	修正案	結合ID	分割	辞書追加候補	母語表現	保留中	備考	書字形	語彙素	語彙素読み	品詞大分類
								アルベール	アルベール	アルベール	名詞
								・	・		補助記号
								カミュ	カミュ-Camus	カミュ	名詞
								は	は	ハ	助詞
								小説	小説	ショウセツ	名詞
								家	家	カ	接尾辞
								も	も	モ	助詞
1	哲人	4650						てつ	鉄	テツ	名詞
1	哲人	4650						じん	塵	ジン	接尾辞
								か	か	カ	助詞
								も	も	モ	助詞
								しれ	知れる	シレル	動詞
								ない	ない	ナイ	助動詞
								。	。		補助記号

図2　人手による誤解析要因処理の例①（FFA040-A2-01-02）

学習者が「私の好きな有名人」というテーマでカミュについて書いた作文で、「哲人」という語をひらがなで「てつじん」と書いたとする。ところが形態素解析器は、この「てつじん」を「てつ（鉄：名詞）｜じん（塵：接尾辞）」という本来意図した語とは異なる2つの語として分割してしまう。このような間違いに気づいた作業者は、まず前後の文脈を読んで学習者の意図を推測する。「小説家も鉄の塵かもしれない」という解釈も考えられるが、文脈から判断すると（必要に応じて対訳作文を参考にする）、カミュについて論じている流れで「小説家も○○かもしれない」と書いているので、学習者は「哲人（哲学者）」と書きたかったのではないかと推測できる。そこで、Web上で形態素解析が行えるWebサービス「Web茶まめ」[6]を使って「哲人」を含む文を試してみる。「アルベール・カミュは小説家も哲人かもしれない」と入力すると、「哲人」が一つの語として認識されることが確認できる。こうして「てつじん」は本来「哲人」という一語であることが確定したので、作業者は以下のように記録する。まず「てつ」

6｜ https://chamame.ninjal.ac.jp/

と「じん」の両方に同じ【結合ID】を付与して、これらが本来一つの語であることを示す。そして【修正案】の欄に正しい語形である「哲人」を記入する。こうして記録された情報は、後にユーザ辞書の作成や本文修正のパターン生成に用いられ、形態素解析器による処理精度の向上に役立つ。

辞書	文境界	書字形（＝表層形）	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現形	語種	書字形(基本形)	語形(基本形)	語彙素ID
B		アルベール	アルベール	アルベール	名詞-固有名詞-人名-一般			アルベール	アルベール	固	アルベール	アルベール	1258
I		、	、		補助記号-一般			、	、	記号	、		44
I		カミュ	カミュ-Camus	カミュ	名詞-固有名詞-人名-一般			カミュ	カミュ	固	カミュ	カミュ	7137
I		は	は	ハ	助詞-係助詞			ワ	ハ	和	は	ハ	29321
I		小説	小説	ショウセツ	名詞-普通名詞-一般			ショーセツ	ショウセツ	漢	小説	ショウセツ	16737
I		家	家	カ	接尾辞-名詞的-一般			カ	カ	漢	家	カ	5580
I		も	も	モ	助詞-係助詞			モ	モ	和	も	モ	37562
I		哲人	哲人	テツジン	名詞-普通名詞-一般			テツジン	テツジン	漢	哲人	テツジン	61760
I		か	か	カ	助詞-副助詞			カ	カ	和	か	カ	5568
I		も	も	モ	助詞-係助詞			モ	モ	和	も	モ	37562
I		しれ	知れる	シレル	動詞-一般	下一段-ラ行	未然形-一般	シレ	シレ	和	しれる	シレル	17181
I		ない	ない	ナイ	助動詞	助動詞-ナイ	終止形-一般	ナイ	ナイ	和	ない	ナイ	27438

図3　「Web茶まめ」による形態素解析の出力結果

誤解析チェック	修正案	結合ID	分割	辞書追加候補	母語表現	保留中	備考	書字形	語彙素	語彙素読み	品詞大分類
								例えば	例えば	タトエバ	副詞
								、	、		補助記号
								「	「		補助記号
1	ラ・ペスト	4660		TRUE			固有名詞：作品名	ラ・ペ	ラペ-Rapee	ラペ	名詞
1	ラ・ペスト	4660		TRUE			固有名詞：作品名	スト	スト-strike	スト	名詞
								」	」		補助記号
								と	と	ト	助詞
								いう	言う	イウ	動詞
								本	本	ホン	名詞
								に	に	ニ	助詞
								ペスト	ペスト-pest	ペスト	名詞
								が	が	ガ	助詞
								ある	有る	アル	動詞
								で	で	デ	助詞
								も	も	モ	助詞
								物語	物語	モノガタリ	名詞
								の	の	ノ	助詞
								医者	医者	イシャ	名詞
								は	は	ハ	助詞
								街	街	マチ	名詞
								に	に	ニ	助詞
								のこし	残す	ノコス	動詞
								て	て	テ	助詞
								い	居る	イル	動詞
								ます	ます	マス	助動詞
								。	。		補助記号

図4　人手による誤解析要因処理の例②（FFA040-A2-01-02）

また、同じ学習者の作文に「ラ・ペスト」という表現が登場した例もある。これはカミュの代表作『ペスト（仏：*La Peste*）』のフランス語原題を学習者が音のまま書き写したものである。このように、外国語由来の固有名詞の中には辞書に収録されていない語彙も存在するため、形態素解析器が正しく解析できない場合がある。なお、形態素解析器に語を追加する場合、既存の辞書本体を直接編集するのではなく、別途ユーザ辞書を用意してそこに登録する方法が一般的である。この事例では、「ラ・ペ（名詞）」と「スト（名詞）」のように分割され、誤解析が生じている。そこで作業者は、「ラ・ペスト」を【辞書追加候補】として記録し、「TRUE」を付与する。これは「この語を将来辞書に追加すべき」と

いう印である。さらに【備考】欄には「固有名詞：作品名」のような説明を記入し、後に辞書に登録する際の参考情報とする。

3.2　ユーザ辞書構築による対応

3つの問題のうち、「辞書にない言葉」の問題は、比較的解決しやすいものである。その方法が「ユーザ辞書の構築」である。W-CoLeJaには、既存の辞書にない様々な言葉が登場する。しかし、すべてを辞書に追加すればよいというものではない。追加する言葉を選ぶ際には、慎重な検討が必要である。

まず、「どこまでが一つの語か」を決める必要がある。W-CoLeJaでは「短単位」という基準を採用した。たとえば「国立国会図書館」は「国立＋国会＋図書＋館」という4つの短単位に分割される。この基準に従って、学習者の作文に現れたすべての語形や表記の揺れにかかわらず、同一語と判断される表現をすべて異表記として扱い、日本語の公式名を見出し語として設定した。短い語（特に1文字）については慎重に判断した。たとえば、中国語の姓「易」や韓国のアイドル「V」のような語は、他の語の一部（「易しい」「Victory」）になりやすいため、辞書に追加すると逆に他の語の解析を間違える可能性がある。このため、このような語は辞書には登録しない方針とした。

辞書に追加する語には、UniDicに準拠した詳細な情報を付与した（小木曽・中村 2014）。語彙素、語彙素読み、品詞大分類、品詞中分類、品詞小分類という項目について、それぞれ適切な情報を記録する。図5に示すようなユーザ辞書追加リストを作成する。たとえば、韓国のグループ「BTS」は「語彙素：BTS、読み：ビーティーエス、品詞：名詞－固有名詞－組織名」として登録した。中国のアプリ「淘宝（タオバオ）」については、中国語表記と日本語読み「タオバオ」の両方に対応できるよう「語彙素：タオバオ－淘宝、読み：タオバオ、品詞：名詞－固有名詞－一般」として登録した。

書字形出現形	語彙素表記	語彙素読み	語彙素細分類	語種	品詞分類	発音形出現形	短単位メモ
BTS	BTS	ビーティーエス		固	名詞-固有名詞-組織名	ビーティーエス	固有名詞：韓国のアイドルグループ
淘宝	タオバオ-淘宝	タオバオ		固	名詞-固有名詞-一般	タオバオ	固有名詞：アプリ名（中国のショッピングアプリ：タオバオ）
タオバオ	タオバオ-淘宝	タオバオ		固	名詞-固有名詞-一般	タオバオ	固有名詞：アプリ名（中国のショッピングアプリ：タオバオ）
Instagram	Ｉｎｓｔａｇｒａｍ	インスタグラム		固	名詞-固有名詞-一般	インスタグラム	固有名詞：アプリ名（Instagram）
Ｉｎｓ	Ｉｎｓ	インス	Instagram	固	名詞-固有名詞-一般	インスタグラム	固有名詞：アプリ名（Instagramの略称）
鐘	鐘	チョン		固	名詞-固有名詞-人名-姓	チョン	固有名詞：人名（鐘／南山）
南山	南山	ナンシャン		固	名詞-固有名詞-人名-名	ナンシャン	固有名詞：人名（鐘／南山）

図5　ユーザ辞書追加リストの例

3.3　辞書構築の課題

追加語彙を登録したユーザ辞書を導入することで、固有名詞に関する誤解析は一定程度改善されたが、それと同時にいくつかの課題も明らかとなった。その中でも特に難しかったのは、「どこまでを一つの語として扱うべきか」という語の境界の判断である。たとえば、ベトナムの元国家主席「Hồ Chí Minh（ホー・チ・ミン）」という名前を考えてみよう。これは「Hồ」が姓で「Chí Minh」が名前なのか、「Hồ Chí」が姓で「Minh」が名前なのか、あるいは3つすべてを独立した単語として捉えるべきかなど、複数の解釈があり得る。これらを正確に判断するためにはベトナム語の知識が必要であり、言語的な専門性が求められる。また、中国の女優「迪丽热巴（ディリラバ）」のように姓名のうち姓が表記されていない場合もある。このように、世界各国における多様な命名慣習を統一的なルールで処理することは容易ではない。こうした複雑さを踏まえ、W-CoLeJaでは、固有名詞の種類に応じて異なる処理方針を採用した。人名については、言語的知識に基づき、可能な限り姓と名に分割して登録することを原則とした。地名については、細かく分割せず、一語として登録する方針とした。

また、固有名詞の処理をめぐっては、辞書の継続的な更新という課題も浮かび上がった。社会の変化とともに、新しい製品名や人名、サービス名は日々登場し続けている。そのため、ユーザ辞書は一度作成して終わりではなく、学習者の言語使用の変化に応じて継続的に更新・拡充していく必要がある。

□4　パターンマッチングに基づく自動修正

「ひらがな表記」や「学習者の誤用」に起因する誤解析に対しては、パターンマッチングを用いた自動修正を試みた。人手によって修正されたデータに出現したパターンを、未作業のデータにも適用する。具体的には、4,137本の修正済み作文データをもとに、頻出する誤りとその修正例を収集・整理した。たとえば、「りょうし｜ん」という不自然な語の分割が「りょうしん（両親）」に修正された例、「インタ｜ネット」が「インターネット」に訂正された例、「て｜ず｜くり」が「手作り」に統合された例などである。こうした修正パターンを未処理のデータに対して適用し、該当箇所を自動的に置換することで、すべてを人手で修正する必要なく、一定程度の効率化を図ることができた。

一方で、パターンの無条件な適用には慎重を要した。短すぎる語や汎用性の高い形式を含むパターンは、無関係な箇所にも誤って適用される可能性がある。たとえば、「お」を「を」に置き換えるといった処理は、特定の誤用を修正する

意図があっても、適切な語（例：「お母さん」）に誤って適用されるおそれがある。このような誤適用を防ぐため、各パターンについて適用精度を統計的に検証した。すなわち、修正候補が実際に正しい結果となった割合を計算し、一定の基準値を上回るもののみを採用対象とした。

□5　おわりに

本章では、W-CoLeJaの構築事例を通じて、学習者作文における形態素解析の課題と対応手法について検討した。学習者作文に特有の誤解析要因を「辞書未登録語」「ひらがな表記の多用」「誤用」の3つに分類し、それぞれの特徴を分析した上で、「ユーザ辞書構築」による実用的な対応手法を提示した。

本章の成果として、まず学習者データ特有の誤解析要因について体系的な分析を実施した。約4,000本の作文の分析により、従来は個別事例として扱われがちであった誤解析現象をパターン化して理解することが可能となった。これらの分析結果は、先行研究（李 2009; 横野 2024など）が指摘してきた問題を大規模なデータによって実証的に検証したものである。次に、ユーザ辞書による固有名詞対応の有効性を示した。各言語圏に特有の人名・地名・文化語彙を体系的に収録することで、形態素解析の精度向上を確認した。さらに、本章で提示した8項目による人手分析の枠組みは、他の学習者コーパス構築においても参考となる可能性がある。

一方で、いくつかの課題も明らかとなった。語の単位認定において、特に人名処理では統一的な基準設定が困難であることが判明した。多様な言語背景を持つ学習者の人名には、既存の形態素辞書が想定する分割規則では対応が難しいものが多数含まれている。ミドルネームを含む人名や、複数の姓を持つ場合もあり、これらをすべて統一したルールで処理することは困難である。W-CoLeJaでは、こうした複雑さを踏まえつつも、可能な限り姓と名を分割して登録する方針を採用した。また、ひらがなで表記された語については、文脈に応じた意味の選択が必要である。「こうかい」は「公開」「後悔」など複数の語に対応しうるが、その解釈は文脈に大きく依存する。学習者作文には直訳表現もみられ、意味が取りづらい事例も少なくない。W-CoLeJaには学習者の母語作文も収録されており、これを参照することで意味の補完がある程度可能である。ただし、日本語文と母語文が必ずしも対応しているとは限らないため、母語作文は補助的な資料として活用するに留まる。

以上を踏まえると、学習者作文の形態素解析においては、人手による分析と

機械処理の適切な分担が重要である。本章で示した手法が、人手作業の効率化と解析精度の向上の参考となり、学習者の言語使用実態をより正確に把握するための前処理手法の検討に役立てば幸いである。

謝辞

本章で言及した人手による誤解析要因処理に関して示唆に富む助言を頂いた国立国語研究所特任助教本多由美子氏および同プロジェクト非常勤研究員井上雄太氏に謝意を表する。データ処理面では、国立国語研究所共同利用推進センター・言語資源ユニット特任専門職員中村壮範氏より多大な技術的支援を受けたことを記して感謝する。最後に、人手によるデータ作成作業に携わってくださった本プロジェクトのスタッフ各位および科研の関係者、調査協力校の教員・学習者のみなさまのご協力を得た。心より感謝申し上げる。

参考文献

小木曽智信・中村壮範(2014)「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」『自然言語処理』21(2):301–332.

迫田久美子(2020)「I-JAS 誕生の経緯」迫田久美子・石川慎一郎・李在鎬(編)『日本語学習者コーパス I-JAS 入門―研究・教育にどう使うか―』2–13, くろしお出版.

朱雅蘭・李琦・工藤隆弘・横野光(2026予定)「学習者コーパスにおける形態論情報整備に向けた取り組み」『国立国語研究所論集』30(ページ数未定).

横野光(2024)「誤用を含む文の形態素解析の分析とアノテーションの検討」『日本語学会2024年度春季大会発表予稿集』199–202.

李在鎬(2009)「タグ付き日本語学習者コーパスの開発」『計量国語学』27(2):60–72, 計量国語学会.

(李 琦・朱 雅蘭・工藤 隆弘・横野 光)