

TEI タグで夏目漱石の長篇小説を読む

日 比 嘉 高

1. はじめに—— AI に小説は読めたか

ここ数年、驚異的な速度で進化を遂げた AI（ここでは ChatGPT などの大規模言語モデル）は、文書の要約や翻訳、文章・画像・映像・音楽の生成、知的情報への回答、プログラム・コードの作成、はては人生相談まで、さまざまな役割を果たしつつある。教育の場においてもその影響は大きい。生徒・学生が与えられた課題に対して AI を用いて解答を作成してしまうことが、現在教育現場の大きな問題の一つともなっている。AI の解答の質は、生徒・学生たちの平均的な解答の質を、しばしば上回ってしまうのである。

文学作品の読解ではどうだろうか。私自身、授業の課題において「〇〇の作品を論じよ」という指示を頻繁に出している。この課題に AI は適切に答えるだろうか。次に掲げたのは、私が一昨年度に近代文学研究の入門的授業で出した期末レポートの課題である（文言は調整している）。「この作品は夏目漱石の『草枕』である。この表現を分析せよ。その際、「講義で解説した文学批評理論」や用語から適切な一つを援用すること。ただし援用する理論は、複数を組み合わせてもよい。「講義で解説した文学批評理論」とは、作家論、作品論、テキスト論、メディア論、ナラトロジー、ジェンダー論、空間論、読者論、ポストコロニアル理論である。文字数は 500 字とする」。たとえばこれを、そのまま ChatGPT 5.2 に投げてみる。結果は、次の通りである。

添付の作品は夏目漱石の『草枕』であり、その表現はナラトロジーと空間論の観点から特徴的に把握できる。本作は主人公が旅をする設定を持ちながら、物語的事件や行動の連鎖は極度に抑制され、語りは一貫して内省と観照へと向かう。ナラトロジー的に見ると、一人称語りは出来事を再現するための装置ではなく、自己の感覚や美意識を展開するための媒体として機能している。時間の進行は緩慢で、因果的展開よりも思考の跳躍や反復が優先される。

空間論の観点では、旅によって移動しているにもかかわらず、語りの焦点は物理的空間の変化ではなく、風景を「見る」視線の在り方に集中する。山村という周縁的空間は、近代社会の喧騒から切り離された静止の場として構築され、主人公はそこに身を置くことで、日常的時間や社会的関係から距離を取る。この空間は、移動の結果として到達される場所というより、美を思索するために仮設された舞台である。[…]

出題者の私の意図からすれば、このレポートは私が求めているポイントを押さえていない。このレポートには、具体的な場面、表現の分析がまったく出てこない。文学理論の主要な役割の一つは作品の読解の精緻化・多様化であり、文学理論を使って「読む」ということは、作品の個別の表現が新たな目で「読める」ようになるということが理想である。

そういう目でこの AI の解答を見てみれば、初歩的な概念の理解にもとづいた作品解釈のポイント

が簡潔に整理されている一方で、作品の個別の表現分析はほとんど行われていないことに気づく。この解答が書いているのは一人称の特徴であり、周縁的空間の特徴であるにすぎない。ChatGPT5.2 は、一人称や周縁的空間をめぐる言説の型を保持しており、それらの組み合わせによってこの解答を出力している。その際、「草枕」の本文を取得しているものの、その表現の分析まで行っていない。AI は小説を「読めていない」のである。

コンピュータには、まだ小説が読めない——のだろうか。現代の AI はまだ文学研究者が満足するほどには読めない。しかし突き放すのは早計である。今私が行ったような、ほぼ丸投げ——つまり自立的にコンピュータに読解させるようなやり方では、まだ適切な出力は得られない。だが、いまや私たちのもとにはさまざまな情報科学が作り出した解析のツールや方法が存在する。「読める」か「読めない」かは、やり方による。適切な課題、適切な方法を示せば、コンピュータは興味深い読解の切り口を差し出してくる可能性を有している。その回答は、作品の新しい「読み」の一部を構成しうる。それが本論考で私が主張したいことである。私はそうしたコンピュータとともに読解を行うための補助的なツールとして、今回、TEI タグを用いる。

2. TEI とは何か、それを利用して文学研究に何ができるか

TEI (Text Encoding Initiative) とは、人文学研究のために策定されたテキストマークアップ——本文テキストに定められた形式によってタグ記号を付与していくこと——の国際標準である。文学作品や歴史文書などのテキストを XML 形式にのっとって記述し、章や段落といった構造情報だけでなく、登場人物、地名、注釈、異本や版の差異、解釈のキーワードなどといった意味的・解釈的情報を付加していく記法の規格である (石田友梨ほか編 2022)。

デジタル・ヒューマニティーズの文脈では、TEI は単純なテキストの電子化を超えて、分析・注釈・可視化・他資料との連結を可能にする基盤技術として用いられてきている。たとえば、人物やキーワードにタグを付すことで出現頻度や関係性を計量的に扱うことができ、また異なる作品や資料を横断的に比較することも容易になる。地名にタグを振れば、地図をはじめとした地理データとの連結も可能になる。

TEI は、テキストを単に読みやすく整形して表示するだけではなく、目的に応じて操作したり、別の情報を付加したり統合したり、何らかの角度から分析しうる対象へと転換するための基盤を提供する。TEI を利用することにより、分析者の前には計量的手法と解釈的読解とを統合していく道が開けるのである。

2024 年から私は、山岸郁子、嚴仁卿とともに、Japan Past & Present (JPP) の支援を受けて、「デジタル漱石：全長篇小説の TEI タグ付けデータセット開発とその応用 Digital Soseki: Development and Applications of a TEI-Tagged Dataset for his Novels」のプロジェクトを開始した。これは、夏目漱石の全長篇小説に対し、TEI によるタグを付与し、データセット「デジタル漱石：全長篇小説 TEI タグ付け本文データセット」として公開しようというプロジェクトである¹⁾。同時にこのプロジェクトは、同データセットを用い、空間分析やキャラクターネットワーク分析、モチーフの時系列的変容の考察などといった解析技法、分析手法の開発を行うことも目指している。本論文は、その成果の一部となっている。

表 1 デジタル漱石：全長篇小説 TEI タグ付け本文データセット 主要タグと役割

タグ名	属性名	用途	備考
div	type	本文の区切り	type="chapter" (章)、type="section" (節) など
head	なし	見出し	章番号や作品名の見出し
persName	type, ref	人名	type="chr" / type="non_chr" で虚構/非虚構を区別。 ref は人物 ID 参照
placeName	なし	地名・場所名	土地、場所などの名称
said	who, toWhom	発話	発話者 who と発話の向け先 toWhom を示す
quote	なし	引用	他の作品や著作、慣用句などの引用表現
seg	ana	意味解釈	読解の切り口を概念語で示す (例：#body, #emotion, #family など)

表 2 デジタル漱石：全長篇小説 TEI タグ付け本文データセット 解釈タグ一覧

学術	芸術	身体・服飾	身分・階級	色	消費
#academic	#art	#body	#class	#color	#consumption
飲食	経済・消費	感情	イベント・事件	家庭・家族	恋愛
#eating	#economics	#emotion	#event	#family	#love
メディア	性格・気質	教育	思潮	旅・移動・乗り物	未定義
#media	#personality	#school	#trendThought	#trip	#uncategorized

上述のように TEI は国際的に規格が定められたタグを用いるが、同時にタグを拡張する方法も定めている。「デジタル漱石」プロジェクトでは、TEI にもとづいた表 1 のような主要タグを用いているが、既定のタグの type 属性や ana 属性を工夫するなどもしている。登場人物名は <persName> タグによって括られるが、これは type 属性によって二分し、登場人物名を示す chr と、固有名詞を表す non_chr に区分した。また、分析の角度を用意するため、表 2 の上列に示した鍵概念を用意し、<seg> タグの ana 属性（下列）においてこれを記述した。こうした解釈のための <seg> タグを、本論文では解釈タグと呼ぶ。これらの鍵概念は『夏目漱石事典』（平岡敏夫ほか編 2000）の項目を踏まえて選定されている。つまり、漱石文学においてどのような鍵概念が重要なのかを『漱石事典』の構成を元に判断したわけである。これは、既存の漱石研究の蓄積を、データ分析と接続したいという意図によるものである。

TEI によるタグ付けを行った漱石の長篇テキストは、さまざまな分析の取りかかりが埋め込まれた状態となる。これから先の分析者のアイデア次第で、いくつもの解析法を受け止められるのだ。付け加えれば、本データセットはクリエイティブ・コモンズ・ライセンス CC BY-SA、つまり「原作者のクレジット（氏名、作品タイトルなど）を表示し、改変した場合には元の作品と同じ CC ライセンス（このライセンス）で公開することを主な条件に、営利目的での二次利用も許可される CC ライセンス」²⁾ で公表されるため、後の分析者や研究者が、新しいタグや属性を付加していくこともできる。

データセット「デジタル漱石：全長篇小説 TEI タグ付け本文データセット」の特徴については、今回は詳細な記述を割愛する。本論文の目的は、データセットの紹介ではなく、それを用いて何ができるかの概観にある。冒頭の議論に立ち返れば、自立的には文学作品を「読めない」状態にあるコンピュータに、読解の取りかかりを与えることで、「読める」ようにする、あるいは人間の読解の支援を行えるようにすることが、この研究の発想の根幹となる問題意識である。

コンピュータを利用して人文学に関わる資料を解析したり、視覚化したりしようとするデジタル・ヒューマニティーズ (DH) には、一つの課題がある。それは、資料の質的な分析に、どこまで踏み込めるかの問題である。コンピュータの解析速度は人間の読解スピードとは比較にならないほど速く、短時間のうちに大量の資料を読み込み、指示した答えを抽出することができる。国立国会図書館が提供するデジタル・テキストをはじめ、さまざまなデジタル・アーカイブが整備されつつある今、コンピュータが有する量的な資料分析の力は、研究にとって計り知れないインパクトを持つ。フランコ・モレッティ (モレッティ 2016) はこうした量的分析の方向性に、「遠読」という名を与えた。

その一方、文学研究の世界では、作品の表現を精緻に読み込む質的な分析——いわゆる「精読」が重視されてきた。本論考が取り上げる夏目漱石のような著名作家の著名作品においては、一つの作品についての研究論文が数百本存在することも珍しくない。この精読にどこまでコンピュータの力を利用できるのか、というのが目下の文学系 DH の課題の一つである。

橋本健広の「精読と遠読の接続——伝統的な文学研究にデジタル・ヒューマニティーズの技法をどう取り入れるか」(橋本 2024) は、精読と遠読に関わる研究文献を広く渉猟し、その歴史的経緯から最新の動向まで見渡した論考である。橋本は、精読の試みが「その始まりから、一方で遠読的な読みの方向性を志向しており、また精読の方法自体においてもテキストの意味の多層性との格闘であった」(p.81) こと、DH の研究領域の傾向として「計量的手法の探求を志向し、[...] 大量のデータから特定の特徴や傾向を選び出す研究が多い、一方でミクロ的な精読を志向する興味も確かに存在」(p.84) したことを指摘する。その上で、「全体から傾向を見出す遠読の手法は、一部のテキストの意味の不確定さを読み取る精読と、方向性が異なっている」(p.87) としながら、その両者を接続する道を探っている。橋本にならうというならば、本研究は TEI タグを使うことで、精読と遠読を接続しようという試みとも言えるだろう。

以下、本論前半においては、漱石の長篇小説を対象に、複数作品の比較分析の試みを行う。後半では、「草枕」を対象として、個別の作品の精読に取り組む。

3. 遠読——複数作品の比較分析

3-1 解釈タグで複数作品を比較する

手はじめに、TEI の各作品のテキスト内で、どの解釈タグがどれくらい使われているのかを出現頻度によって見てみる (なお、以降の分析において「未分類 uncategorized」タグはすべて割愛した)。頻度は1万字あたりの解釈タグの出現回数によって算出している。対象とした作品は、「こころ」「坊っちゃん」「抗夫」「草枕」である。結果のグラフ (図1) は、作品毎の内容や表現上の特徴を視覚化しているものと捉えられる。

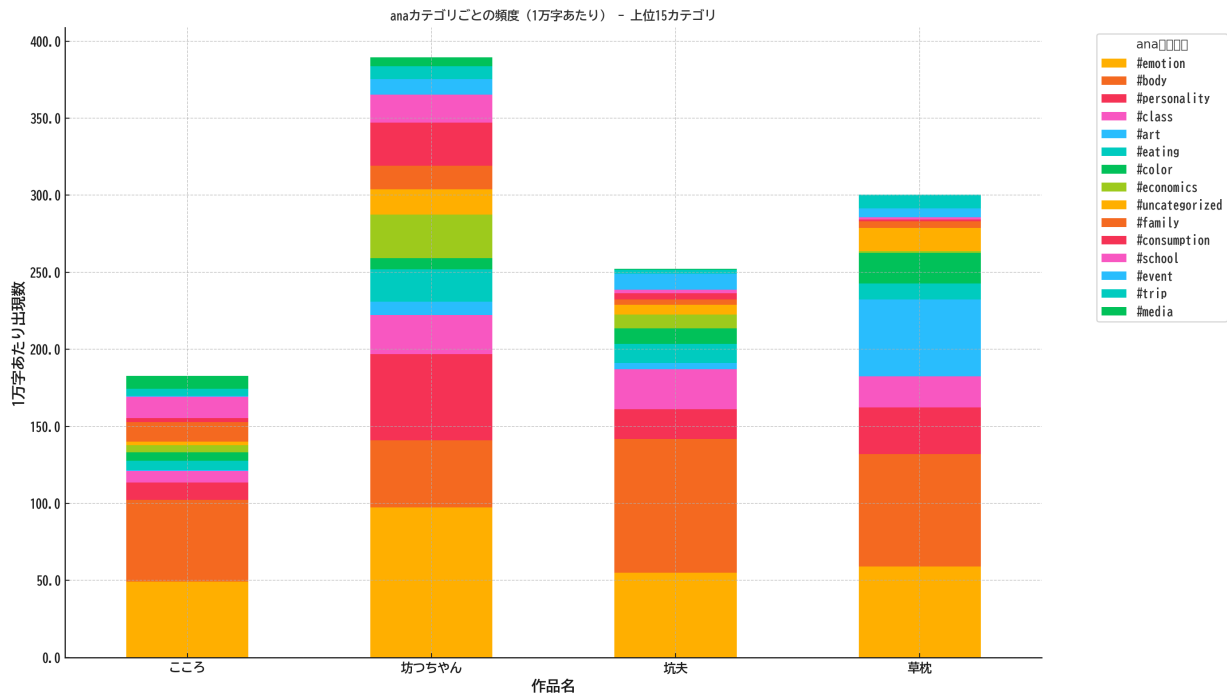


図1 「こころ」「坊つちゃん」「抗夫」「草枕」1万字あたりの解釈タグ出現頻度

まず全体としてみてみれば、ほぼすべての解釈タグにおいて、「坊つちゃん」が他の三作に比べて突出して高い出現頻度を示している点が目を引く。とりわけ「感情 #emotion」「身体・服飾 #body」「性格・人格 #personality」「階級 #class」といった、人間の感情・身体についてのタグ、性格についてのタグ、社会的位置づけに関わるタグが分厚く積み上がっており、この作品が人物の行動や感情の反応、社会的対立に関心を寄せている様子が定量的に裏付けられている。

これに対して「こころ」は、全体のタグ頻度が最も低く、感情や身体のタグは一定量見られるものの、「イベント・出来事 #event」や「旅・移動・乗り物 #trip」、「メディア #media」といった外在的・出来事的なタグや、移動や通信に関わるタグは少ない。もっとも、「こころ」には先生の長い手紙が含まれていることは周知の通りだが、本文中における「手紙」の語の頻度は低いということだろう。「こころ」は事件の連鎖よりも、心理の省察や登場人物の対話によって構成される小説であり、同一の感情状態や関係性が長く引き延ばされるために、単位文字数あたりのタグ出現が抑制される傾向にあると言えそうだ。

「抗夫」は、その中間的な位置にあり、#emotion と #body の比重が比較的高い一方で、#trip や #event に関わる層が厚く、身体、移動、労働経験といった要素がテキストの運動を支えていることが読み取れる。「草枕」は全体の頻度では「坊つちゃん」ほど高くないものの、#art（芸術・美術）、#color（色）といった芸術や視覚に関わるタグ、また #media（メディア）に関するタグが相対的に多いという特徴をもっている。以上、図1のグラフにおいては、漱石の四つの長篇の内容面、表現面での特質の違いが、1万字あたりの解釈タグ出現頻度という指標を通じて可視化されているということを見て来た。

次は、どのような解釈タグがどの作品に現れているかについて、12作品をヒートマップで示してみよう。図2は「長篇12作品における解釈タグのヒートマップ（割合）」である。

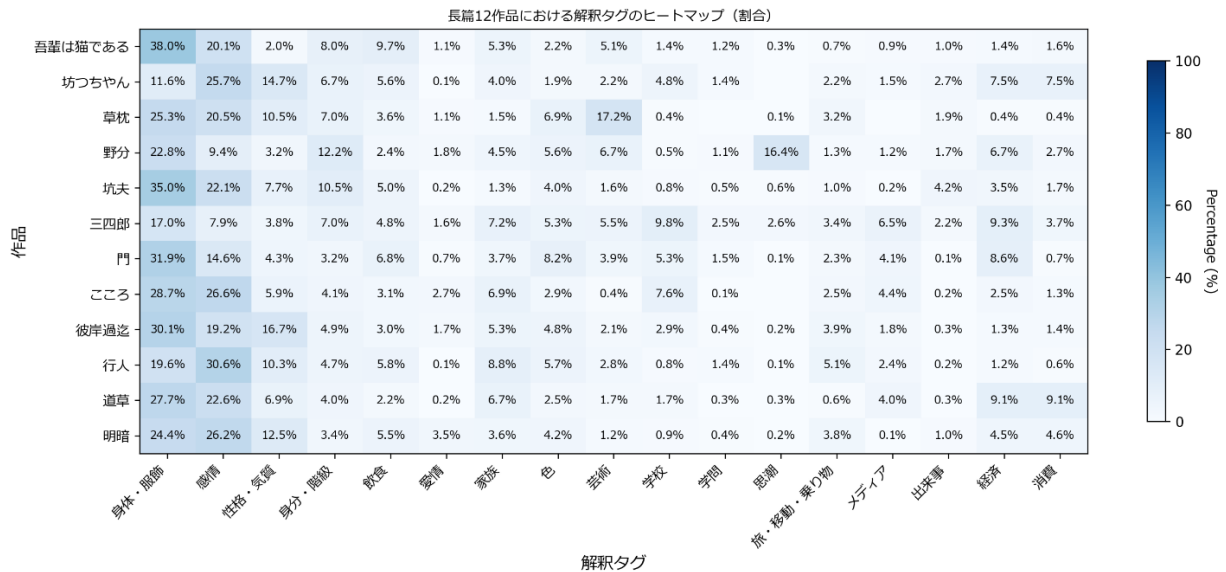


図2 長篇 12 作品における解釈タグのヒートマップ (割合)

全体的な傾向についてまず把握すると、ほぼ全作品において「身体・服飾」「感情」が強く、漱石長篇が基本的に身体表現や知覚・内面描写を厚く書く小説だ、ということが確認できる。とりわけ「吾輩は猫である」は「身体・服飾」「感情」が突出して高いことがわかる。これは猫の挙動を語りか追いかけているという面が影響しているのかもしれない。また「坑夫」「彼岸過迄」「道草」「明暗」も「身体・服飾」の解釈タグが多いことが見て取れる。

また、縦軸方向は作品名が発表順に並んでいるが、「坑夫」と「三四郎」の間に、断絶があるように見える。横軸方向は解釈タグが並べてあるが、左1/3ぐらいは身体や感情、性格、飲食、愛情、家族など、個人に関わるタグが並べてあり、真ん中の1/3ぐらいは色、芸術、学校、思潮など、抽象的で高踏的な領域に関わるタグを配置した。右の1/3は、旅やメディア、経済、消費などの社会的な領域に関わるタグを置いている。「三四郎」より下の漱石の中期・後期作品においては、ヒートマップの横軸真ん中ぐらい——つまり抽象的・高踏的タグ——と、右側——つまり社会的領域のタグ——が、幅広く色が濃くなっていることが見て取れる。その一方、「坑夫」より上の前期作品においては、色のつき方はまばらである。大局的に見て言えば、作品によって集中して現れる解釈タグが異なっていた前期に対し、中期・後期においては、作品間の差は減る傾向にあり、かつ解釈タグもそれほど偏りのない、比較的幅広い領域のタグが出現する傾向にあると言ってよさそうである。さらに中期・後期のようすをより詳しく見てみれば、「家庭」の割合の高さや、「学校」が中期に固まっていること、「経済」が最晩年の「道草」「明暗」に多いことなど、興味深い特徴も観察できる。

一方、横軸方向に見た場合には、作品ごとにどのような解釈タグの割合が多いのかという配分の特徴が見えてくる。たとえば、「野分」は「感情」のタグが少ない一方で、「思潮」のタグが多い。「草枕」は「芸術」が相対的に強く、「坑夫」は「身体・服飾」や「身分・階級」「出来事」のタグが相対的に多いが、それ以外のタグはおしなべて少数であるという特徴も読み取れる。

このように、解釈タグを用いたヒートマップは、作品毎の特徴や、漱石の作品史上の変化を視覚的に表現してくれる。

3-2 会話数から複数作品を比較する

次は、会話に着目してみる。TEIには<said>というタグがあり、文字通りこれに囲われた文字列が会話・発話であることを示す。このタグには「誰が」「誰に」言ったのかを注記できる属性が付与されているため、発話主体と聞き手とを明確にしたテキストの構造化が行える。「デジタル漱石：全長篇小説 TEI タグ付け本文データセット」では、漱石の全長篇作品におけるすべての対話について<said>タグを付しており、その意義は大きいと考えている。

最初のグラフは、「長篇 12 作品の会話数の比較（総文字数との割合）」（図 3）である。これは、文字通り、作品中における会話（＝発話）の数を数え、作品の総文字数に対するその割合を算出したものである。これにより、その作品の中に占める会話の比率、重みが測れることになる。

見てすぐに気がつくのが、「野分」の割合が突出しているということである。割合は 1.17% である。白井道也の演説が含まれていることも想起されるが、ここは文字数ではなく会話・発話の数を数えているために、長広舌であるかどうかは問題となっていない。「野分」は、漱石の長篇の中では抜きん出て会話の回数が多い作品であり、登場人物同士の対話や議論、応酬が物語の大きな駆動力となっている可能性が指摘できる。

逆に「坑夫」が極端に低い（約 0.36%）ことも際立っている。一人称の語りで進行するこの作品では、主人公の内面独白や鉱山の観察描写が主調をなしており、対話や発話の数は非常に低くなっている。結果として作品はモノログの色調を帯びることになる。「坑夫」の特徴をよく捉えた数字になっているといえるだろう。

グラフの増減を眺めると、ヒートマップの結果と同様に、前期の作品においては会話の比率が作品によって大きく変動していることが観察できる。「三四郎」までその傾向が続く。一方、「門」で発話数がぐっと下がった後は、堅調に増加していくならかな上昇基調が観察できる。つまり長篇作品史の前半において、漱石は人物の対話を描出の仕方に関して、さまざまな形態を試みていたのに対し、後半では対話の使い方にそれほど揺れはなくなっていたという可能性が考えられる。また付け加えれば、最終作の「明暗」に近づくに従って会話数がどんどん増えていることの意味も考えてみたくなる。晩年に向けて漱石の物語は対話劇の色を濃くしていったようなのである。

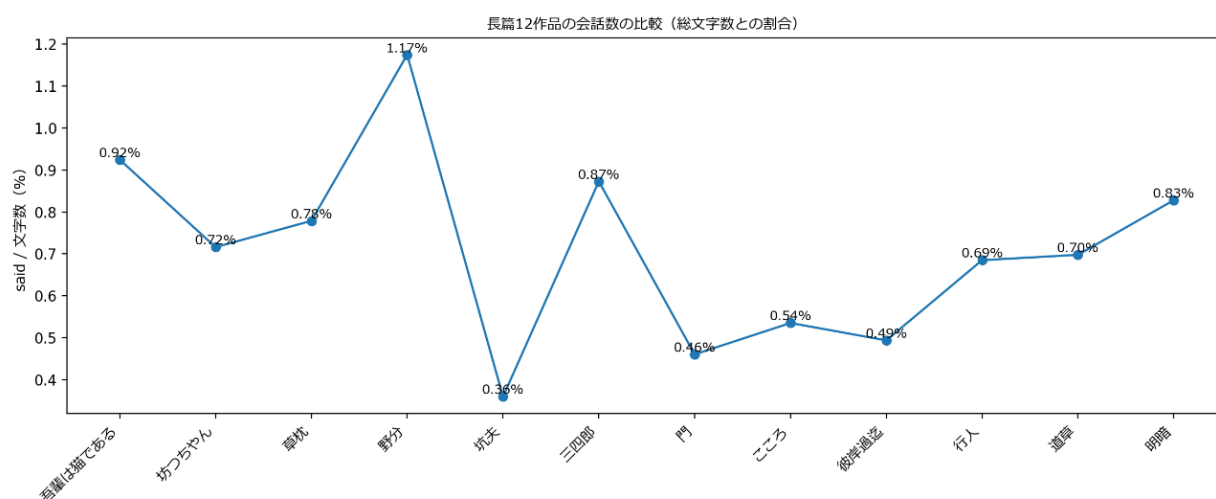


図 3 長篇 12 作品の会話数の比較（総文字数との割合）

最後に、「三四郎」の位置がやはり面白い。「それから」の分析が今回含まれていないため、保留は必要ではあるが、「三四郎」「それから」「門」という三部作とされることが多く、つい後半のグループに入れたくなるが、ヒートマップの結果を見ても、会話数の結果を見ても、「坑夫」から「三四郎」にかけて、特徴的な値が続いており、「門」以降の安定的なようすとは一線を画しているように見える。「三四郎」が前期と後期の端境に立つ作品であることが、量的な分析から見えてくる。

さて、図3を検討してきたが、このグラフは会話数を数えており、会話の長さ（文字数）は見えていなかった。本データセットではすべての発話が<said>タグで囲われているために、文字数を数えることができる。これを踏まえて、12長篇を対象にその推移を見てみよう（図4）。図4では、発話1件当たりの文字数について、平均値と中央値を示している³⁾。

まずどの作品においても、中央値より平均値がかなり大きいということが見て取れる。平均値は極端な値に影響されるため、このグラフの場合に即して言えば、少数の長い発話があるために、平均値が押し上げられているということになる。対して中央値は、大小いずれも極端な値の影響は受けにくいので、その作品において頻出する長さの会話文字数がつかめる。中央値は12～24文字の範囲に収まっており、漱石の長篇における典型的な発話は多くの作品においてこの範囲であるとわかる。とくに「門」から「行人」までは中央値が18か19文字と安定しているのが目立つ。

特徴的な値を持つ作品に目を向ければ、「吾輩は猫である」は平均会話文字数が45.15であるのに対し、中央値は24字となっていて、その差が最大である。典型的な会話の長さも全長篇の中では最長の部類であることは確かだが、中央値が20を超える作品も珍しくないために、突出しているとまでは言えない。つまり、「吾輩は猫である」の会話は、普段の発話の中に、長大な発話が頻繁に混ざっていることが推定される。苦沙弥郎に集まる迷亭、寒月、東風らの長広舌や講釈が思い起こされる結果である。

「坑夫」と「三四郎」の中央値が非常に低いことも目に付く。それぞれ12文字、14文字である。これは、作品内の多くの発話が他作品と比べて一段短いことを示している。会話が基本的に短いやり取りで進行するという特徴があるわけである。

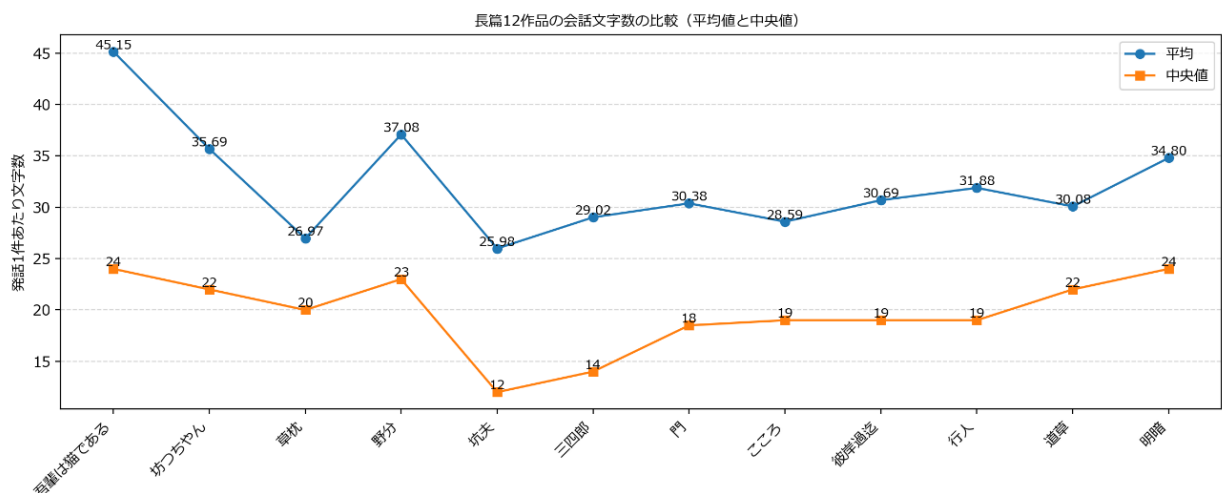


図4 長篇12作品の会話文字数の比較（平均値と中央値）

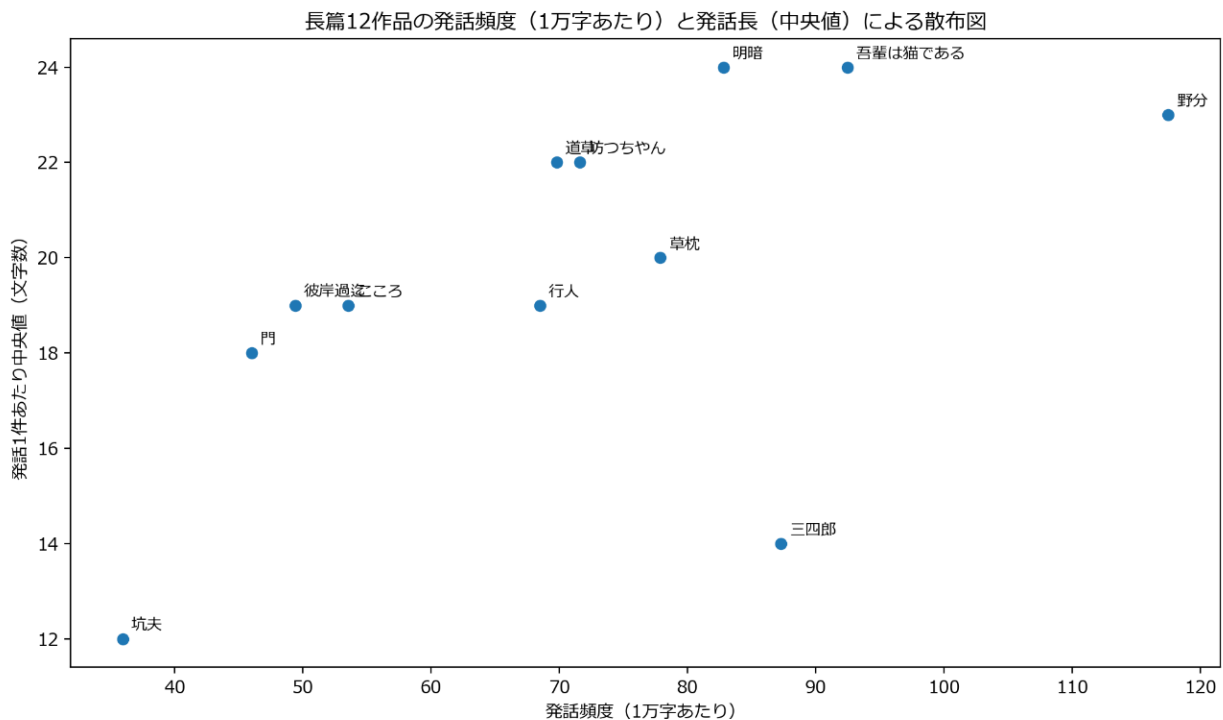


図5 長篇12作品の発話頻度（1万字あたり）と発話長（中央値）による散布図

後半にいくにしたがって会話数が伸びていくことは図4の分析で見たところだが、会話の文字数も伸びている。「明暗」の中央値は24であり、「吾輩は猫である」や「野分」と並ぶ長さである。平均値も35.80で、「吾輩は猫である」との差は大きいものの、「坊つちやん」「野分」と並んでいる。津田もお延もお秀も吉川夫人も、よくしゃべる人物というわけである。

本節の最後に、図3と図4の両方を考え合わせてみよう。図5は、発話1件当たりの文字数を縦軸に、一作品内の発話頻度（1万字あたり）を横軸に取った散布図である。この図からは、「野分」「坑夫」「三四郎」が、他の作品群の列から離れていることが見て取れる。「野分」は発話の頻度が高く、発話の文字数も多い。一方、「坑夫」は発話の頻度も低く、発話の文字数も少ない。ここでも面白いのは「三四郎」で、発話の文字数が少ないわりに、発話の頻度はそこそこ多い。他の漱石の長篇とは発話の文字数と発話の頻度の比率が逆になっていて、そのために「三四郎」だけが列から外れて下部に位置しているように見えるのである。

4. 遠読から精読へ——解釈タグと読む「草枕」

4-1 分析手法の検討

ここからは、作品を一つに絞り込んで、解釈タグを用いた量的な解析を、作品の精読へと接続させていく試みを行う。扱うのは「草枕」である。TEIタグを用いることで、これまで人間による読解では明確になってこなかった何らかの着眼点がえられないだろうか。

作成した「草枕」のTEIファイルを材料として、ChatGPTと応答を重ねたところ、たとえば次のような提案がえられた。【中心性の「差分」「加速度」を見る】、【発話ネットワーク全体の「密

度」「直径」「モジュラリティ」の推移】、【発話「方向性」ベクトルの推移】、【発話の「二者関係」の強度の変化】、【新規登場人物・退出人物のイベント分析】、【章ごとの「テーマ変調」】、【発話の「感情変調」】、【発話ネットワークの「コミュニティ遷移」】。可能性がありそうなものも、実現が難しそうなものも、実現してもあまりよい成果がえられそうにないものもあるが、まずは、「中心性」に着目する方法を試みる。具体的には、「草枕」の TEI データから発話ネットワークを構築し、章ごとの会話構造を 4 種類の中心性（次数中心性、媒介中心性、近接中心性、核構造中心性）で分析するというものである⁴⁾。これにより、物語が進行していく中で現れる「特異点」をあぶりだしたいと意図したわけである。

ただし結果を簡略に述べれば、得られたグラフは登場人物間の対話の様態を反映した複線の折れ線グラフとなった。会話に加わる登場人物の入れ替わりや、発話の頻度、会話の中心人物などを視覚化する興味深いものではあり、グラフだけではなくアニメーション化も行えたためにそれなりの成果はえられた。しかしながら、作品分析を行う研究論文という本考察の趣旨からすれば、「草枕」を人の目で読んだときの内容把握を超えるものとはいえず、新規性・発見性の面で不満が残った。長篇に入れたとはいえ、「草枕」は長大な物語ではなく、また登場人物数も少ない。より複雑で、要素が豊かな作品で試みれば、成果は出そうである。

4-2 解釈タグの数量から解析する

別のアプローチとして、解釈タグを用いた分析を試みることにした。目的は、会話タグを利用した中心性の分析と同じく、物語の「特異点」を数値から発見するということにある。図 6 に示したのは、「草枕」の TEI ファイルに含まれる解釈タグの数を章ごとに数え、折れ線グラフとして表したものである。

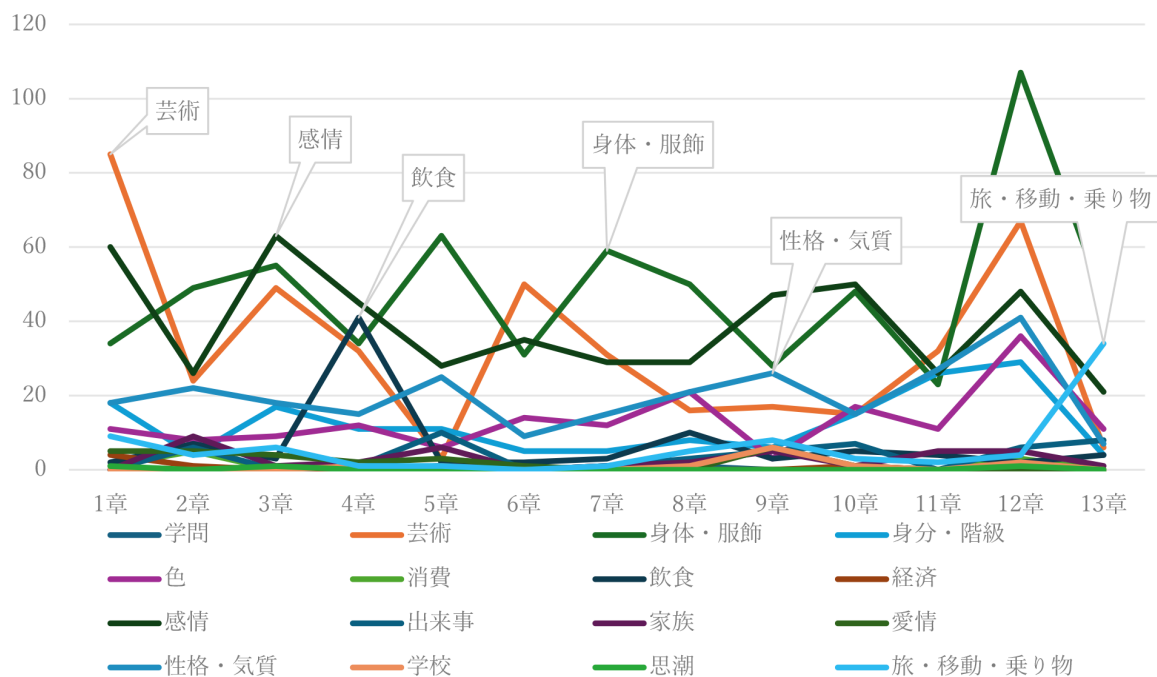


図 6 「草枕」における解釈タグの章ごとの出現数

全体的な傾向を見ると、12章において解釈タグが突出して多いことがわかる。「身体・服飾」「芸術」「感情」をはじめ、複数のタグがここで多数出現しており、総数を押し上げている。結末近いこの章は、余が宿の窓や、散歩に出た木槿の中から、那美のようすを観察したり彼女のようすを品評したり、那美と夫との短いやりとりの場を目撃する場面である。那美の行動に目を引くものが多い上に、これまでの那美についての画工としての余の観察が踏まえられ集成され、その結果が一人称語りで高密度に語られていることから、さまざまな種類の解釈タグが増えていると考えられる。12章は、本作品における那美をめぐる解釈の切り口が集中的に現れていた章であり、作品解釈の鍵となる章であることが示唆されているだろう。

個別のタグに目を向ければ、「身体・服飾」タグが波打ちながらも平均して多く、全体を通して考えたときには最多数のタグとなっている。「草枕」が、身体表現を一つの語りの軸にしていることがここから言えそうである。「感情」タグも数が多い。ただし5～8章では継続して低めの数になっているという特徴がある。「芸術」は、多寡の起伏が大きいという特徴がある。1、3、6、12章に大きな山があるが、こうした章においては余の画工としての観察や考えが前面に出ている章だといえるだろう。つまり余の画工としての観察は、間欠的に現れるという特徴が、ここから指摘できそうである。

すべての章において言えるというわけではないが、「身体・服飾」と「芸術」のタグを比べて観察したときに、両者が逆向きの増減を示す傾向があることも興味深い。1、5、6、7章において、一方のタグ数が上昇しているときに、もう一方のタグが下降している。10章もわずかではあるが、似た傾向がある。身体表現と芸術のトピックは、一致しても不思議ではないはずだが、「草枕」という作品においては、そうではないようだ。

さて、最後に注目したいのが「旅・移動・乗り物」のタグ（ここでは「旅タグ」と略記する）である。旅タグは、ほぼ全章にわたって低い出現数で、その数は0～8回である。ところが13章だけ34と桁違いに跳ね上がる。割合で見ても、13章中の全タグ合計数が142であるうち、旅のタグは34（約24%）という高い割合になっている。しかも、折れ線グラフが明らかに表現しているように、旅タグは、他のタグの動向との比較においても特異な動きを示している。ここまで数が多かった主要な各タグは、12章で上がった後、13章で大きく数を低下させている。比較的目立たないマイナーなタグは、微減、微増のままである。そのなかで旅タグだけが、大きく上昇しているのである。

「草枕」の読者にはあらためて言うまでもないことだが、この小説は旅の物語である。主人公の余が、俗世を離れて鄙びた温泉宿に向かい、そこで滞在する話だ。タイトルそのものが、「草枕」という、旅先の宿や、旅寝、あるいは旅自体を指す言葉を冠している。では、この最終章において旅のタグが急上昇するという構造の意味は、どのように解釈すれば良いだろうか。ここから先が、人の読解の出番である。

4-3 旅タグと読み直す「草枕」

旅タグがないからといって、テキストが旅を描いていないわけではない、ということの確認からはじめよう。物語の冒頭の一章および二章は、那古井に向かう余が、山路を行くようすに費やされている。それはあきらかに、旅の過程の描出である⁵⁾。作中で余が繰り返す言葉に「非人情の旅」という鍵語もある。この語は二章が初出となり、俗世間から離れ、「淵明、王維の詩境を直接に自

然から吸収して、すこしの間でも非人情の天地に逍遙したい」(一章)という余のもくろみに従った那古井滞在を指す表現としてある⁶⁾。

ではなぜ第一章において旅タグが少ないのかと言えば、山路に行く描写はあっても、旅や移動や乗り物に関わる単語が多くは現れないからである。これはタグの付け方の方針の問題とも言えるが、作品のテキストが人の世や詩、目に入る風景、非人情をめぐる思念を語ることに費やされているという特徴によるものだと説明できる。旅を描きながら、通常であれば旅を構成する語彙が、少ないのだ。

このことは、関心の持ち方は異なるが、先行研究でも指摘されてきた。加藤禎行(2000)は「汽車論」を切り口に「草枕」を論じるその冒頭で、「停車場から旅立つ場面なしに山路を登っている」と指摘している(p.28)。余は、東京の人間らしいことが髪結床の主人との会話からうかがわれる。とすれば、彼は那古井への山路をたどる前に、汽車に乗っており、船にも乗っていただろう。彼的那古井への旅は、汽車や停車場の描写から始まってもおかしくはないのだ。平川祐弘(1982)は、漱石がイギリス滞在中に訪れたピトロクリでの経験を論じながら、漱石にとってその土地が「汽車の走らぬ世界であるかのような錯覚」を与える「別乾坤」「桃源郷」であったことの意味を論じた(pp.319-320)。汽車の表象を抑圧することによって、桃源郷を現出させるような志向が、漱石の文業にはあったとされてきたのである⁷⁾。

ただ、十三章における旅タグ単独の急上昇という特異な変化をもとに考える本論考では、作家漱石の志向性は考慮の中に入れない。今回の作業では、タグは作品の中にしかおかれておらず——TEI タグは作品と作品の外をつなぐツールにもなりうるのだが今は措く——、いわば作品表現に関わる構造的な読みを深める方向に進むのが妥当だろう。

私が注目したいのは、十三章の旅が、余の旅ではないという単純な事実である。「草枕」は山路の描写に始まり、別乾坤たるべしと願われる那古井の滞在を描き、川下りの描出で終わるという構造を持つ。だがあらためて確認すれば、終わりの旅は、余自身の旅ではなく、久一の旅立ちであり、那美の夫の出立であったのだ。結末の場面を引用しよう。

車輪が一つ廻れば久一さんは既に吾等が世の人ではない。遠い、遠い世界へ行つて仕舞ふ。其世界では烟硝の臭ひの中で、人が働いて居る。さうして赤いものに滑つて、無暗に転ぶ。空では大きな音がど／＼と云ふ。是からさう云ふ所へ行く久一さんは車のなかに立つて無言の儘、吾々を眺めて居る。吾々を山の中から引き出した久一さんと、引き出された吾々の因果はこゝで切れる。もう既に切れかゝつて居る。車の戸と窓があいて居る丈で、御互の顔が見える丈で、行く人と留まる人の間が六尺許り隔つて居る丈で、因果はもう切れかゝつて居る。(十三章)

日露戦争へ出征する久一、満洲へ向かう那美の夫を乗せた汽車が、動き出す場面である。旅立つ久一と、残る者たちを汽車の扉が隔てていく。「草枕」の語りが、ここで「因果」という言葉を繰り返していることに目を留めよう。

「因果」の語は、実は「草枕」の中で重要な役割を果たしている。一番最初に現れるのは、那美の登場場面である。温泉宿の部屋で寝転び、メレディスの一節を引きながら⁸⁾、余は、昨晚から今

朝にかけての未知の女性——那美——との邂逅にその内容を重ねてみている。「二人の間には、ある因果の細い糸で、この詩にあらわれた境遇の一部分が、事実となって、括りつけられている」と余は言い、「因果もこのくらい糸が細いと苦にはならぬ」と続ける（四章）。

次に「因果」が現れるのは、鏡が池の由来を余が源兵衛から聞く場面である。源兵衛はずっと昔の志保田の嬢様が、旅の虚無僧を見初めてしまった悲劇を語り聞かせる。そこで「因果」が現れるのである——「因果と申しますか、どうしても一所になりたいと云ふて、泣きました」（十章）。昔の嬢様を介して、那美にオフィーリアのイメージが重なる重要な場面である。

こう見てくれば、「草枕」において「因果」の語は、重要な人物同士を結ぶ縁を説明するための鍵語になっていると考えて良いだろう。その事実を確認した上で、久一の出立の場面に戻ろう。「吾々を山の中から引き出した久一さんと、引き出された吾々の因果はこゝで切れる」と余は語った。人間の意志を超える「因果」の働きによって、久一と「吾々」が切り離されるのだ。

「吾々」？ 戦場へ向かう久一を送り出す旅において、「余は無論御招伴に過ぎん」（十三章）のではなかったのか。久一との因果を語りながら図らずも余の口から出た「吾々」の語を、読者はどう読めば良いのか。「残る吾等も否応なしに残らねばならぬ」（十三章）——余はそうも続けるのだ。

「草枕」において、余の旅は閉じられていない。結末の旅は、余の旅ではない。むしろそこで語られているのは、余が「余」ではなく、那美に連なる縁者たちの中で「吾々」として語り出す瞬間である。余は、「吾々」と呼ぶ人びととのあいだに、あらたな因果の糸を結び始めているのである。

5. おわりに——小説を、タグとエクセルと AI とともに読む

以上、TEI タグを用いた漱石長篇テキストの構造化が、どのような研究の展望を開くのか、その事例を示してきた。前半では、量的な把握から作品を理解しようとする遠読的な試みとして、複数作品の比較分析を行った。「吾輩は猫である」や「抗夫」などに、それぞれ特徴的な解析結果が現れることや、作品史を追ったときの変化が会話や発言を括る <said> タグからも観察できることを示した。また「三四郎」が前期と後期の端境に立つ、興味深い作品であることも指摘した。

TEI タグを利用した分析は、作品の精読を行う際の手がかりにもなりえる。とりわけ、今回私たちのプロジェクトが開発した「デジタル漱石：全長篇小説 TEI タグ付け本文データセット」には解釈タグが埋め込んであるため、それを利用したさまざまな考察の道が開けるはずである。本論文後半では、「草枕」を例に、旅タグの挙動が、一つの特異点を示しており、そのことを起点に作品の読み直し——精読に接続できると論じた。もちろん、タグは会話タグや解釈タグだけではない。引用 <quote>、人名 <persName>、地名 <placeName> などのタグも、分析者のアイデア次第で活用できるはずだ。

包み隠さず言えば、「デジタル漱石：全長篇小説 TEI タグ付け本文データセット」の TEI タグは完璧ではない。現時点において、タグの付加／不付加の基準には揺れがあるし、見逃しも多く残っている。作業者が多数であるために、統一性の点でも課題が残っている。だが、こうしたデータセットの不備は、今後徐々に補われていくだろう。まず公開し、有志が訂正していくという情報文化の展開の中で、このデータセットが成長していくことを願っている。

附記

本研究は、「デジタル漱石：全長篇小説の TEI タグ付けデータセット開発とその応用 Digital Soseki: Development and Applications of a TEI-Tagged Dataset for his Novels」プロジェクト（日比嘉高、山岸郁子、嚴仁卿）の一部であり、Japan Past & Present の支援を受けている。プログラム開発に際しては村上英吾氏の、タグ付与作業においては名古屋大学大学院の大学院生諸氏の助力を得た。また TEI ファイルの分析手法、グラフの作成・読解に関して、ChatGPT を用いた検討を行っている。漱石作品の本文は、TEI データセットについては『夏目漱石全集』（第3巻、ちくま文庫、1987年）を、論文中の引用では『漱石全集』（第3巻、岩波書店、1994年）を用いた。

注

- 1) ここでは、全長篇小説として文字数を基準に「吾輩は猫である」「坊っちゃん」「草枕」「野分」「坑夫」「虞美人草」「三四郎」「それから」「門」「こころ」「行人」「彼岸過迄」「道草」「明暗」を選定している。ただし、今回の分析では「虞美人草」「それから」のデータを含めることができていない。「デジタル漱石」プロジェクトの詳細については、次の JPP の頁を参照。また「デジタル漱石」のデータセットについては、JPP のサイト (<https://japanpastandpresent.org/jp/projects/>) の他、日比の Github (<https://github.com/users/Yosh-Hibi/projects/1>) でも公開している。
- 2) 「クリエイティブ・コモンズ・ライセンスとは」 creative commons Japan ウェブサイト。2026年1月6日確認。
<https://creativecommons.jp/licenses/>
- 3) 原則として開始・終端を示すカギ括弧「」も字数に数えている。
- 4) 次数中心性 (degree centrality) は、その章で「どれだけ多くの相手と会話しているか」を示す活動量の指標である。媒介中心性 (betweenness centrality) は、人物が「他者と他者をつなぐ橋渡し」となっている度合いを表す。物語上のキーパーソン／分岐点の指標となる。近接中心性 (closeness centrality) は、ネットワーク全体から見て、その人物がどれだけ「近い位置」にいるかを表す。物語全体の「重心」にどれだけ近いかを示す指標である。核構造中心性 (k-core / coreness) は、ネットワークを「核」と「周辺」に分解したときの、核の“層の深さ”を表す。値が高いほど「その章の会話のコア」に属する人物だと解釈できる。
- 5) ここから、「草枕」は山路に始まり、川下りに終わるという対構造になっているという発想もありうる徐貴淑・河邊聡 (1996) は「初めの「山路」から最後の「川」までで、現実—非現実—現実の図式は閉じることになる」と指摘している。
- 6) この旅をめぐるのは、たとえば佐藤勝 (1972) は「「非人情」の「旅」の試みが俗（俗悪で「軽薄」な二十世紀の現実）を避けるものでなくその中で生き続けるため」のものだとしている (p.71)。また畑有三 (1981) は「人間の現実に直面していく姿勢を画工は選び取る決意を持つ」、そのための過程として「『旅』の経験」が必要だったとしている (p.104)。
- 7) 前田愛 (1990) は、那古井の土地、とりわけそこから吉田への川下りの場面を、桃源郷のイメージと重ねて読んでいる。
- 8) 石崎等 (1977) はメレディスの「シャグパッドの毛剃り」の内容と「草枕」の引用場面とを関連付けて読解している。

主要参考文献

- 石崎 等 (1977) 「『草枕』論の前提」『跡見学園短期大学紀要』13号
- 石田友梨ほか編 (2022) 『人文学のためのテキストデータ構築入門——TEI ガイドラインに準拠した取り組みにむけて』文学通信
- 加藤禎行 (2000) 「『汽車論』の隠喩——夏目漱石「草枕」をめぐる」『日本近代文学』第62集
- 佐藤 勝 (1972) 「『草枕』論」『国語と国文学』49巻4号
- 徐貴淑・河邊聡 (1996) 「境界的空間が持つ意味についての考察——漱石の『草枕』に見る四つの境界的空間」『日本建築学会計画系論文集』61 (485)
- 橋本健広 (2024) 「精読と遠読の接続——伝統的な文学研究にデジタル・ヒューマニティーズの技法をどう取り入れるか」『国際情報学研究』第4号

畑 有三（1981）「草枕の旅」三好行雄ほか編『講座夏目漱石』第2巻、有斐閣

平岡敏夫ほか編（2000）『夏目漱石事典』勉誠出版

平川祐弘（1982）「汽車の走らぬ世界——漱石にとってのピトロクリ」三好行雄ほか編『講座 夏目漱石』第5巻、有斐閣

モレッティ、フランコ（2016）『遠読——「世界文学システム」への挑戦』原著 2013、みすず書房、秋草俊一郎ほか訳

（名古屋大学大学院人文学研究科教授）