現代のテクスト分析
その理論と実装

石井 大地

【研究協力】
京都大学情報学研究科 知能情報学専攻 博士課程 長谷川 嵩矩
ＣＲＵＮＣＨＥＲＳ株式会社

○はじめに

近年のコンピューターによるデータ解析の技術・精度の発展は、多種多様のテクストを様々な角度から分析することを可能にしている。既にマーケティングリサーチや、インターネット上における CGM（Consumer Generated Media＝消費者生成メディア）の分析には、こうした技術が大々的に取り入れられている。本論では、筆者を中心とするグループが研究している日本語の文学テクストの解析手法について、その基礎的な理論と実装の手順を概説する。

人間の感性や直感だけでは把握できないような莫大な情報を扱うことで、それまでは気づくことのできなかった作品の特徴、あるいは作品間の連関といったものを知ることができるようになる。このことが、文学研究の幅を広げ、創作的才能についての新たな知見をもたらすことが期待される。

本論では、そうしたテクストの解析技術を使って、主に作品間の連関をみるための手法を概説するが、ここに提示された手法を理解・応用すれば、作品内における特徴を抽出し、比較する多種多様なシステムを構想し、設計することができるようになるはずだ。

なお本論で扱うテクスト解析の技術的な参考資料としては、『Rによるテクストマイニング入門』（石田本広、2008、森北出版）に詳しい。

○「テクスト」と「分析」

テクストとは、語とその配置によって形作られる情報のことである。テクストの意味内容とは、第一義的には、どの言葉を使うか、そしてその言葉をどう配置するかによって決定される。扱う語を変えると意味が変わり、その配置を変えれば意味が変わる。語とその配置が同じなら、ひとまず同じ意味を表すと言える。

もちろん、テクストにおける意味の決定も、厳密に言えば、テクストの外にある様々な社会的文脈、読み手の考えを反映して様変わりする。またどの一つの言葉をとっても、その意味は時代に応じて変化し得るので、固定的な意味が常に担保されるわけではない。とはいえ、テクスト自体が改変されない限り、語とその配置という情報そのものが変わることはない。テクストを分析するということは、このデータセットについて分析をするということである。

分析する、という用語は、その字の通り、「複雑な事柄を要素や成分に分け、構成を明らかにする」、つまり与えられたデータを分解・整理することで、そのデータの持つ意味や意義をより明快に理解しようとする試みである。これを情報という観点から言い直せば、複雑で大量の情報を、ある指標に基づいて要約し、そこから何らかの意味や意義を見出すことこそ、分析行為に他ならない。

テクストを分析するとは、語とその配置という情報を、適切に分解・再整理（＝要約）し、そのテクストの持つ意味や意義を明らかにすることである。それが人間の手によるものであれ、コンピューターを用いるものであれ、テクスト分析とは、要するにテクストの情報をいかに意味のある形に要約・整理するかが問われるものである。

とはいえ、この要約という作業が行きすぎると、文章の理解が大雑把になり、意味のないものとなりがちである。たとえば私たちは普段、何かの本を取って読んだあと、「面白い／つまらない」とか、「泣けた／泣けなかった」といったような短絡的な評価で済ませてしまうことが少なくない。こうした評価も、作品に対する一種の「要約」作業に他ならないのだが、しかしこういった大雑把な判断は、少なくとも研究的意義のある批評・鑑賞とは言わない。ある基準に従って作品から得られる情報を要約して見出したものがその程度の二者択一的判断であっては、その作品の何がどう良いのか、悪いのか、まったく分からないからだ。

もちろんこれは極端な例だが、研究者による分析においても、同様の恐れがないとは言えない。きわめて複雑な文学作品の内容を、過度に単純化したり、あまりに分かりやすい指標に置き換えて論じてしまうと、肝心の細部が見落とされる可能性が高い。

コンピューターを用いることの利点は、人間では到底不可能なほどの大量のデータを高速に処理できることである。人間ならば大雑把にしか把握できないことを、より細かく、正確に、高速に把握することで、人間の認識では摑みきれなかった特徴を見て取ることにこそ、コンピューターを使う意味がある。つまり、情報の要約の度合いを人間による精読よりも大幅に引き下げることで、作品の細部をより精緻に検証することができるのである。

○日本語テクスト解析の試み

語とその配置に注目したテクスト解析は、語と語とのあいだにスペースの入っている欧米言語などでは比較的容易に行うことができるが、日本語でこれを行う場合、まずは与えられた文章を語に分解する必要がある（後述するが、この作業を「形態素解析」と呼ぶ）。昨今では MeCab 等のオープンソースの形態素解析ソフトウェアが登場したことで、誰でも一般的な目的においては十分な精度を持った解析作業ができるようになった。

こうした技術的基盤の上で、人文学の分野でも技術の応用は始まりつつある。一例として、井上靖、谷崎潤一郎、中島敦、三島由紀夫らの作品における句読点の使い方を調べた研究がある[1]。この研究では、助詞と読点の組み合わせの頻度を計測することで書き手の癖を判別し、どのテクストが誰によって書かれたものかを判定できることが分かっている。実際、形態素解析とその結果を応用した統計処理を行えば、こうした解析を行うことは難しいことではない。そしてそれを実践するためには家庭用 PC が一台あれば足りる。

だがそうした状況であるにも拘わらず、いま現在、コンピューターを用いた日本語テクスト解析が人文学研究に重要な影響を与えているとは言いがたい。その理由の一つは、研究者のもっぱらの関心である「文学の面白さ」についての解析が十分に行われてこなかったことにある。これまでに行われてきた研究の多くは、「句読点」や「特定のキーワード」など、テクストの持つ多様な特徴のうちの、ごく一部の特徴だけが、それも決まって作品の面白さに関係のなさそうな要素ばかりがなぜか注目されてきた。

そもそも、書き方の句読点の打ち方の癖をいくら判別しても、それで文学の持つ創造性や批評性について何らかの重要な知見が得られるとは考えにくい。また作中に出てくるキーワードを数える、といったような単純な方法で作家や作品の持つ意義を評価することは不可能だろう。これでは、先の例で挙げたような、「面白い／つまらない」といった単純な印象批評と、やっていることの水準があまり変わらないようにさえ思えてくる。

そこで筆者らは、「書き手の癖」「キーワードの出現頻度」といった部分的な特徴を抽出するのではなく、テクストの持つ多様な特徴を、（ある程度の要約・縮約は施すものの）全体的に取り込み、解析可能なデータに変換することができないかと考えた（＝要約の度合いを引き下げる）。文体や、内容や、そこに込められた感情、といった多様な特徴をまるごと含み、なおかつ解析が可能な定量的データを手にすることができれば、そのデータを起点として、「文学の面白さ」といった抽象的な評価に迫るような研究が可能になるからである。

○基本的なコンセプト＝ベクトル化

テクストの多様な特徴をまるごと含むためには、語とその配置、というテクストの構造

をそのままモデル化する必要がある。だが、これはもちろん、テクストデータをそのままの形で扱うということを意味しない。ここで第一に理解しておくべき事実は、デジタル化されたテクストのデータは、それ自体では定量的に解析できないということだ。

たとえば「あ」という文字はコンピューターの内部では実際には16進数で「E38182」という数値で表現されている（UTF-8 エンコーディングの場合）。同様にして「い」は「E38184」であり、「町」という文字は「E794BA」である。それぞれの文字に割り当てられた数値自体は、ただ単に特定の文字を示すだけであり、その数字自体には意味はない。「い」＝E38184 は、「あ」＝E38182 より２大きい数だが、これは「い」が「あ」より「大きい文字」であることを意味しない。距離や温度を示す数値は意味のある数値だから、それらを定量的に扱う、つまり大小や多寡を比較したり、演算したりできる。だが文字コードの数値は便宜的なもので、その数値自体の比較や演算ができない。「語」や「文」も、「文字」が集まってできるわけだから、当然、そのままの形では定量的な比較ができない。

テクストのデジタルデータをそのままの形で扱っている限り、分かるのは「ある文字と別の文字が一致するか否か」だけである（＝定性分析）。データ量が少なく手軽に扱えそうなテクストデータだが、その分析が意外に難しい主要な理由がここにある。

従って、何か意味のあるテクスト解析を行おうとする際には、この定性的データの集合体を、比較可能な定量的データに変換する作業が必要不可欠である。

また、語の「配置」についても、作品の長さが様々な異なる状況で、そのまま比べることも難しい。ある語が 10 番目に出てきた、100 番目に出てきた、ということをそのまま比較してもあまり意味はないことは容易に理解される。

そこで筆者を含む研究グループが考えたのは、語とその配置を、語の出現頻度、語の組み合わせの出現頻度という２種類の連なりとして超高次元のベクトルデータで表現するという方法である。言語情報が語とその配置でできているならば、どの語がどのくらい使われているかを見れば語について知ることができ、そしてそれぞれの語がどういった語と組み合わされてどのくらい使われているかをみれば、ある程度、語の配置に関する情報も摑むことができる。後者は、語の出てくる順番というデータに比べると要約の度合いが高いものだが、それでも語の配置について語るデータとしては非常に有用で、なおかつ解析ソフトウェアで容易に扱うことができるため、研究に活用するには良い指標だろう。

文章は、文中の語の出現頻度情報、そして語と語の組み合わせの出現頻度情報に置き換えられることで、超高次元のベクトルデータとなる。ここで使う「頻度」という値は、「文字コード」とは異なり、定量的に比較・計算可能な数値である。このやり方で文学作品のテクストデータを定量化すれば、それを起点とした作品内の解析、あるいは作品間の解析ができる。

それでは、作品をベクトルデータに変換するためにはどういった作業が必要になるだろ

うか？　研究の目的により、その細かい設定は変える必要があるが、大まかに言えば以下に述べる５つのステップが必要となる。

１．テクストの準備
２．形態素解析
３．フィルタリングルールの設定
４．頻度表・共起頻度表の作成
５．データの合成と定量分析

以降、それぞれについて解説していこう。

○１：テクストの準備

まずもって、文学的評価の高い作品は、大半が紙の書籍として販売されており、解析のためのテクストがデジタルデータとして手に入らないことが多い。著作権の切れた過去の作品のうち一部は、青空文庫[2]等のウェブサイトで公開されているが、重要な作家の重要な作品の多くのテクストは紙に印字されている。

ここで、青空文庫掲載作品についてのみ解析をするのであれば、データも揃っているので研究を素早く進められることは確かであり、またそのような研究を行っているグループも多い。だが、だからといって著作権が切れるまで優れた現代作家の作品を解析しないまま過ごすのでは、一体何のための研究なのか、ということになりかねない。どのみち、いま現在の私たちの興味を反映した研究をするためには、いま現在読まれている作品をも含めて解析を進めなくてはならない。作者の没後五十年を待つというやり方で、素晴しい研究ができるわけではないのだ。

だとすれば、市販の紙書籍からデジタルデータを作成するという労苦も厭うわけにはいかない。そこで筆者のグループは、市販されている紙書籍をスキャニングし、得られた画像データを OCR（光学文字認識）プログラムにかけてテキストを生成。そこで得られたデータに残る誤認識文字等を修正するプログラムを作って適用し、また目視でのデータチェックも合わせ、あらかたの誤字を取り除く作業を行った。

そうはいっても、今回取り扱った小説作品は、数万字から数十万字に及ぶ分量がある。筆者らのグループでも相当な時間とコストをかけて対応にはあたり、ぱっと読んだときに誤字脱字が目につかない程度にはデータを整えている。それでも投入できるリソースには限界がある。従って、スキャニング・OCR・修正プログラム・目視チェックで得られたテクストデータが、一文字の誤りもない完璧なデータである保証はないことをあらかじめ断

っておく。

○２：形態素解析

形態素とは、意味の最小の単位のことである。日本語では語と語の間にスペースがないため、与えられた文章を、まずは形態素に分割することで扱いやすくする。この作業を形態素解析という。形態素解析のためには、あらかじめ辞書データを用意する必要がある。

形態素解析のツールも、またそのための辞書データも、既にオープンソースプロジェクトとなって日本中の研究者・企業に利用されている。筆者のグループも、オープンソースの形態素解析ソフトウェア「MeCab（和布蕪）」と、そのために作られたIPADIC辞書データを用いて研究を行っている。

この形態素解析ソフトウェアを用いると、与えられた文章をすべて語（厳密には語ではなく形態素だが、実用上の差異はほぼないので、これからは形態素のことも「語」と呼ぶこととする）に分割し、それぞれの語の品詞や活用の情報とともに出力することができる。なお、この形態素解析プログラムは、辞書にない言葉でも文章の形から自動で品詞を判定してくれる上、辞書データを自分たちで編集できるため、使う過程で解析の精度を高めてゆくことが可能だ。

さて、この形態素解析で得られた出力データを活用すれば、語ごとの出現回数や頻度を集計できる。また、語にはそれぞれ登場した順番も記録されるので、そのデータを用いることで、語と語との組み合わせごとの出現回数や頻度も計測できる（これを共起分析という）。共起データは、「私＋は」「あなた＋を」といった、名詞と助詞の組み合わせや、「助け＋られる」といった動詞＋助動詞の組み合わせ、あるいは「青い＋花」といった形容詞＋名詞の組み合わせなど、文章表現の特徴や意味的なつながりを測るのに便利だ。

また特定の語に対してのみ共起分析を行うことで、作中の特定の要素に着目した解析もできる。たとえば、主人公の名前が「春子」だったとき、春子と組み合わせて使われる＝もっとも接近した位置で使われている形容詞を抽出していけば、文中で春子がどのような人物として描かれているかを知る手がかりになる。このような品詞を絞った共起分析は、語の配置についての定量分析のためには特に優れた方法となる。

この基本的な技術をベースに、それをどのように解析を役立てるかが、技術研究の重要なポイントとなる。

○３：フィルタリングルールの設定

形態素解析では、記号や空白も含む文中のすべての情報が解析される。だが文学作品の内容を判定するのに、括弧や改行、空白の数を数えるのが効果的であるとは思えない。作

品の情報を全体として把握するにしても、不必要なデータや、重要度の低いデータをある程度取り除いた方が、むしろ作品の特徴を見やすいということがある。

また、ニュース記事を解析するのと、小説を解析するのとでは、解析時にどこに力点を置くかは少しずつ異なってくる。ニュース記事であれば、「いつどこで何が起きたか」を伝えることが優先されるため、場所や事柄を示す名詞が重要になると考えられる。一方で、小説の場合、名詞を解析すると登場人物の名前や代名詞の数が非常に多くなってしまうので、こうしたものを省くような処理をした方が適切なことも多い。また感情表現が作品の印象を決定づけるので、形容詞、形容動詞、動詞など、状態や動作を示す言葉についての検証は優先順位が高くなる。

こうした基準を、プログラムによる解析時にフィルタリングルールとして設定することで、意味のある情報を適切に導きやすくなる。さらにこうしたテクストの解析は、わずかな解析対象データの増加が計算量を飛躍的に増大させることが多いので、不必要な情報を解析対象から外すことで、計算にかかる時間、必要となるコンピューター資源を大幅に減らすことができるメリットもある。

私たちの研究グループは、小説を中心に解析を行うために、大きく次の２つのフィルタリングルールを設定している。

【ルール１】多数の小説作品のテクストを形態素解析し、得られた単語セットをから、解析専用の辞書（形態素解析用の辞書とは異なる）を作成し、この辞書内の語について、出現頻度や共起頻度をみる。これによって、小説によく使われる表現を重視した解析が可能になる。

【ルール２】解析専用辞書のなる語のうち、出現頻度があまりに低い語は解析対象外とする。これによって誤字脱字を省き、かつ作品間の些末な差異情報を省くことで、作品間の比較をしやすくする。

○４：頻度表・共起頻度表の作成

形態素解析の結果と、あらかじめ用意した解析専用辞書の情報をつきあわせて、語ごとの出現頻度と、語の組み合わせセットの出現頻度を測定する。たとえば、ある作品Ａについて、このような形で結果が得られることになる。

作品Ａ＝[0.0030411, 0.00005115, 0, 0.00080654, 0, 0, 0.00000298, ……]

この数値の羅列が、語ごとの出現頻度のベクトルデータである。この数字の並びは、解

析専用辞書に出て来る語の順番に一致する。たとえば解析専用辞書が、

青い　赤い　緑　動く　彼　私　……

という並び順なら、「青い」という形容詞が使われる頻度が 0.0030411 だということになる。これは全体を 1 としたときの単語の出現頻度であり、より正確に言えば、あるテクストにおける語の確率分布である。

単語の組み合わせについても、

青い-花　私-の　動く-た　フランス-へ　……

といったような組み合わせを無数に作り、それぞれについて、上に述べたようなベクトルデータを作ることができる。

このベクトルデータは、解析専用辞書にあるすべての語、およびそれらの語の組み合わせセットを次元として含む超高次元ベクトルデータである。その次元数は、フィルタリングルールの設定方法によっても変わるが、語の頻度に関してでも数千の単位、その組み合わせとなると、数百万～数千万の単位になり得る。あまりに計算量が多くなると、いかに最新のコンピューターでも解析が十分な速度で行えないため、特に語の組み合わせについては、解析に用いる品詞の組み合わせを限定するなどして、次元数を抑える工夫も要所要所で必要となる。

○５：データの合成と定量分析

このようにして、テクストを、頻度表と共起頻度表から２つのベクトルデータに置き換えておけば、テクスト間の関係を分析するのは非常に容易なこととなる。ここでは基本的な解析として、それぞれの作品がどの程度似通っているかを判定してみよう。

ベクトルとは、向きと長さを持った量のことだ。数千次元から数万次元のベクトル、というと話が難しそうだが、その性質は二次元や三次元の場合と同じである。従って次元数が多いといっても、筆者らが行おうとしている解析方法の原理を考える上では、二次元や三次元の比喩を使えばかんたんに理解できるため、これからはその比喩で説明を進めてみたい。

まず、ここに作品Ａ、作品Ｂ、作品Ｃの３作品があり、それぞれについて頻度表ベクトルを作成したとする。仮にその値を、

作品Ａ＝[0.3, 0.1, 0.6]
作品Ｂ＝[0, 0.1, 0.9]
作品Ｃ＝[0.2, 0.2, 0.6]

としよう。ここで、それぞれの作品同士がどの程度似通っているかをみるには、それぞれの頻度表ベクトルの差の絶対値を取れば良い。言い換えるなら、各ベクトルの終点間の距離こそが、作品間の「距離」となる。

同じ方向に向いているベクトル同士は似ている。異なる方向を向いているなら似ていない。原理としては非常に単純である。

作品ＡとＢの距離を距離ＡＢとすると、

距離ＡＢ＝√ (0.3 - 0)2 + (0.1-0.1)2 +(0.6-0.9)2　＝0.42426..... ≒0.424

と計算できる。同様にして、

距離ＢＣ＝√ (0 - 0.2)2 + (0.1-0.2)2 +(0.9-0.6)2　＝0.37416..... ≒0.374
距離ＣＡ＝√ (0.2 - 0.3)2 + (0.2-0.1)2 +(0.6-0.6)2　＝0.14142..... ≒0.141

となる。こうみると、作品ＡとＣは似通っていて、作品Ｂは他の２作品とは似ていない、といったことが分かる。

こういったやり方と同じ原理で、数万次元に及ぶ作品に関するベクトルデータを比較してゆくことで、各作品間の関係を知ることができる（実際には、その「距離」の取り方には様々な計算方法があるが、そうした数学的概念をここで逐次紹介するのは趣旨とずれるので、割愛する）。こうした比較を頻度表ベクトル、共起頻度表ベクトル、あるいは目的に応じて作成した様々なベクトルデータについて行い、その結果をうまく混ぜ合わせたり、単位を揃えるなどして合成すれば、目的に応じて作品を評価する指標を柔軟につくることができる。

筆者らは、頻度表ベクトルと共起頻度表ベクトル、それぞれのデータを比較・計算することで、独自の相関指数を作り、これをある作品と別の作品の似ている度合いを示す指標とした。この方式で、いくつか、現代の作家による、最近の文学賞受賞作とベストセラー作品について、その相関を分析してみた。その全てを掲載することは誌面関係上、不可能であるが、そのうち、頻度表ベクトルに基づく相関指数を以下に示してみよう。なお、ここに記した相関指数は、同一の作品であれば「1」となるように単位を揃えたものなので、数値それ自体には特に意味がなく、ただ値同士の比較だけに意味があることに注意してい

ただきたい。

【頻度表ベクトルに基づく相関指数】

	1Q84 BOOK 1 村上 春樹	1Q84 BOOK 2 村上 春樹	1Q84 BOOK 3 村上 春樹	色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹
1Q84 BOOK 1 村上 春樹	1	0.7389862	0.7300398	0.709936
1Q84 BOOK 2 村上 春樹	0.7389862	1	0.7350917	0.7317618
1Q84 BOOK 3 村上 春樹	0.7300398	0.7350917	1	0.7247963
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.709936	0.7317618	0.7247963	1
スクールアタック・シンドローム 舞城 王太郎	0.5037255	0.508801	0.4944419	0.505972
みんな元気。 舞城 王太郎	0.4900022	0.4925641	0.4794416	0.4912837
煙か土か食い物 舞城 王太郎	0.5240149	0.5143913	0.513298	0.5070399
好き好き大好き超愛してる。 舞城 王太郎	0.4999807	0.5162847	0.5022682	0.5142811
日本沈没 下 小松 左京	0.5203174	0.5214544	0.5220804	0.5038815
日本沈没 上 小松 左京	0.4852381	0.4940277	0.4836232	0.4784511
東京タワー 江國 香織	0.492659	0.4872513	0.4972683	0.4849623
冷静と情熱のあいだ-Rosso 江國 香織	0.4931531	0.4892691	0.4823642	0.4896875
AMEBIC 金原 ひとみ	0.4594775	0.4585138	0.459362	0.4546969
IP／NN 阿部和重傑作集 阿部 和重	0.5399006	0.538962	0.5354256	0.5303347
TUGUMI 吉本 ばなな	0.5073888	0.5004793	0.5087744	0.5109892
イッツ・オンリー・トーク 絲山 秋子	0.4945313	0.4797585	0.4744247	0.4999239
オブ・ザ・ベースボール 円城 塔	0.4264217	0.4399021	0.4360048	0.4355518
がらくた 江國 香織	0.511542	0.4994308	0.5006697	0.5201058
コズミック・ゼロ： 日本絶滅計画 清涼院 流水	0.5379467	0.5382257	0.5437662	0.5270261
ドーン 平野 啓一郎	0.5168985	0.5343517	0.532685	0.5473537
なんとなく、クリスタル 田中 康夫	0.4670798	0.4650912	0.453138	0.4616966
希望の国のエクソダス 村上 龍	0.5645461	0.5720672	0.5795441	0.5826234
虐殺器官 伊藤 計劃	0.5647969	0.5750435	0.5679391	0.5747702
蹴りたい背中 綿矢 りさ	0.5111159	0.488943	0.4834574	0.4854689
重力ピエロ 伊坂 幸太郎	0.5395872	0.5314578	0.537996	0.5348589
人のセックスを笑うな 山崎 ナオコーラ	0.4866948	0.5002271	0.4728178	0.5032122
猛スピードで母は 長嶋 有	0.5193771	0.5130085	0.4966821	0.5195081
優雅で感傷的な日本野球 高橋 源一郎	0.52467	0.5238524	0.5227642	0.5229374
腑抜けども、悲しみの愛を見せろ 本谷 有希子	0.5017125	0.5146143	0.5141759	0.5215004

	スクールアタック・シンドローム 舞城 王太郎	みんな元気。 舞城 王太郎	煙か土か食い物 舞城 王太郎	好き好き大好き超愛してる。 舞城 王太郎
1Q84 BOOK 1 村上 春樹	0.5037255	0.4900022	0.5240149	0.4999807

1Q84 BOOK 2 村上 春樹	0.508801	0.4925641	0.5143913	0.5162847
1Q84 BOOK 3 村上 春樹	0.4944419	0.4794416	0.513298	0.5022682
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.505972	0.4912837	0.5070399	0.5142811
スクールアタック・シンドローム 舞城 王太郎	1	0.6666373	0.618799	0.6191052
みんな元気。 舞城 王太郎	0.6666373	1	0.6023593	0.6347633
煙か土か食い物 舞城 王太郎	0.618799	0.6023593	1	0.6208119
好き好き大好き超愛してる。 舞城 王太郎	0.6491052	0.6347633	0.6208119	1
日本沈没 下 小松 左京	0.4355118	0.4345104	0.4663144	0.4636282
日本沈没 上 小松 左京	0.4394424	0.4379921	0.4461718	0.4409012
東京タワー 江國 香織	0.4969972	0.4699777	0.4501638	0.4937599
冷静と情熱のあいだ-Rosso 江國 香織	0.4485522	0.4255719	0.4187337	0.4507447
AMEBIC 金原 ひとみ	0.5707927	0.5326249	0.5239526	0.5369117
IP／NN 阿部和重傑作集 阿部 和重	0.5114029	0.4926615	0.5196625	0.5307524
TUGUMI 吉本 ばなな	0.5257809	0.5263513	0.4981808	0.5155113
イッツ・オンリー・トーク 絲山 秋子	0.561528	0.5554879	0.5142463	0.5212607
オブ・ザ・ベースボール 円城 塔	0.4039834	0.3959193	0.4396376	0.4367385
がらくた 江國 香織	0.5160458	0.50453	0.4725769	0.4981235
コズミック・ゼロ： 日本絶滅計画 清涼院 流水	0.49738	0.4877427	0.534007	0.5362661
ドーン 平野 啓一郎	0.5041751	0.4772578	0.5273731	0.5137037
なんとなく、クリスタル 田中 康夫	0.49153	0.4737394	0.4957002	0.4906859
希望の国のエクソダス 村上 龍	0.5454727	0.5351162	0.5634784	0.5588732
虐殺器官 伊藤 計劃	0.4988089	0.4887547	0.5086977	0.5288394
蹴りたい背中 綿矢 りさ	0.5338027	0.5158217	0.4861209	0.5274167
重力ピエロ 伊坂 幸太郎	0.5545489	0.5405383	0.5572588	0.5502876
人のセックスを笑うな 山崎 ナオコーラ	0.5409884	0.5446695	0.5069668	0.5422506
猛スピードで母は 長嶋 有	0.5147494	0.5116807	0.4965263	0.5134822
優雅で感傷的な日本野球 高橋 源一郎	0.5339088	0.5208607	0.5427776	0.5501636
腑抜けども、悲しみの愛を見せろ 本谷 有希子	0.5157664	0.4938428	0.5303316	0.5084612

	日本沈没 下 小松 左京	日本沈没 上 小松 左京	東京タワー 江國 香織	冷静と情熱の あいだ-Rosso 江國 香織
1Q84 BOOK 1 村上 春樹	0.5203174	0.4852381	0.492659	0.4931531
1Q84 BOOK 2 村上 春樹	0.5214544	0.4940277	0.4872513	0.4892691
1Q84 BOOK 3 村上 春樹	0.5220804	0.4836232	0.4972683	0.4823642
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.5038815	0.4784511	0.4849623	0.4896875
スクールアタック・シンドローム 舞城 王太郎	0.4355118	0.4394424	0.4969972	0.4485522
みんな元気。 舞城 王太郎	0.4345104	0.4379921	0.4699777	0.4255719
煙か土か食い物 舞城 王太郎	0.4663144	0.4461718	0.4501638	0.4187337
好き好き大好き超愛してる。 舞城 王太郎	0.4636282	0.4409012	0.4937599	0.4507447
日本沈没 下 小松 左京	1	0.6740139	0.4605862	0.4740491
日本沈没 上 小松 左京	0.6740139	1	0.46208	0.4929622
東京タワー 江國 香織	0.4605862	0.46208	1	0.617513
冷静と情熱のあいだ-Rosso 江國 香織	0.4740491	0.4929622	0.617513	1
AMEBIC 金原 ひとみ	0.403369	0.3948391	0.4336996	0.4189898
IP／NN 阿部和重傑作集 阿部 和重	0.4868575	0.4527832	0.4767089	0.4462213

TUGUMI　吉本 ばなな	0.4843597	0.4721204	0.4911367	0.4840674
イッツ・オンリー・トーク 絲山 秋子	0.4206574	0.4075103	0.4938409	0.4611588
オブ・ザ・ベースボール 円城 塔	0.376512	0.3537479	0.3741821	0.3844771
がらくた 江國 香織	0.4693894	0.4566091	0.6307548	0.6260888
コズミック・ゼロ： 日本絶滅計画 清涼院 流水	0.5090624	0.4821687	0.4757475	0.4694634
ドーン 平野 啓一郎	0.4384827	0.4132543	0.4618332	0.4555975
なんとなく、クリスタル　田中 康夫	0.446611	0.4314605	0.4462247	0.4467889
希望の国のエクソダス 村上 龍	0.5112809	0.5206307	0.4986826	0.4614265
虐殺器官　伊藤 計劃	0.5083779	0.4971186	0.4921205	0.4799863
蹴りたい背中　綿矢 りさ	0.4482756	0.4123933	0.4830187	0.4808714
重力ピエロ　伊坂 幸太郎	0.4568884	0.4543096	0.4770056	0.4527502
人のセックスを笑うな　山崎 ナオコーラ	0.4324502	0.4269999	0.485667	0.4811446
猛スピードで母は　長嶋 有	0.4777333	0.4526707	0.4830354	0.4794857
優雅で感傷的な日本野球　高橋 源一郎	0.4772462	0.4782974	0.4837947	0.4623391
腑抜けども、悲しみの愛を見せろ　本谷 有希子	0.4785802	0.4485792	0.462399	0.4598773

	AMEBIC　金原 ひとみ	IP／NN 阿部和重傑作集 阿部 和重	TUGUMI　吉本 ばなな	イッツ・オンリー・トーク 絲山 秋子
1Q84 BOOK 1 村上 春樹	0.4594775	0.5399006	0.5073888	0.4945313
1Q84 BOOK 2 村上 春樹	0.4585138	0.538962	0.5004793	0.4797585
1Q84 BOOK 3 村上 春樹	0.459362	0.5354256	0.5087744	0.4744247
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.4546969	0.5303347	0.5109892	0.4999239
スクールアタック・シンドローム 舞城 王太郎	0.5707927	0.5114029	0.5257809	0.561528
みんな元気。 舞城 王太郎	0.5326249	0.4926615	0.5263543	0.5554879
煙か土か食い物　舞城 王太郎	0.5239526	0.5496625	0.4981808	0.5142463
好き好き大好き超愛してる。　舞城 王太郎	0.5369417	0.5307524	0.5155413	0.5212607
日本沈没 下　小松 左京	0.403369	0.4868575	0.4843597	0.4206574
日本沈没 上　小松 左京	0.3948391	0.4527832	0.4721204	0.4075103
東京タワー　江國 香織	0.4336996	0.4767089	0.4911367	0.4938409
冷静と情熱のあいだ-Rosso　江國 香織	0.4189898	0.4462213	0.4840674	0.4611588
AMEBIC　金原 ひとみ	1	0.494499	0.4734561	0.4988936
IP／NN 阿部和重傑作集　阿部 和重	0.494499	1	0.4639531	0.4944065
TUGUMI　吉本 ばなな	0.4734561	0.4639531	1	0.4967773
イッツ・オンリー・トーク 絲山 秋子	0.4988936	0.4944065	0.4967773	1
オブ・ザ・ベースボール 円城 塔	0.4507987	0.4627552	0.3613974	0.3872895
がらくた 江國 香織	0.4844184	0.4852155	0.5209175	0.5038212
コズミック・ゼロ： 日本絶滅計画 清涼院 流水	0.4736391	0.5584067	0.4928517	0.4845504
ドーン 平野 啓一郎	0.4845351	0.540843	0.4658026	0.4907832
なんとなく、クリスタル　田中 康夫	0.4463932	0.4890743	0.4629375	0.472394
希望の国のエクソダス 村上 龍	0.527467	0.5696253	0.516125	0.5098716
虐殺器官　伊藤 計劃	0.4767436	0.5516443	0.4937839	0.4647768
蹴りたい背中　綿矢 りさ	0.4762375	0.451928	0.5297112	0.5190244
重力ピエロ　伊坂 幸太郎	0.5172837	0.5607927	0.508662	0.5177349
人のセックスを笑うな　山崎 ナオコーラ	0.5114338	0.4675545	0.5031269	0.5261135
猛スピードで母は　長嶋 有	0.4766968	0.4839081	0.5307074	0.5403884

優雅で感傷的な日本野球　高橋 源一郎	0.4971	0.541594	0.5110599	0.522537
腑抜けども、悲しみの愛を見せろ　本谷 有希子	0.5011973	0.5183668	0.4751895	0.5053489

	オブ・ザ・ベースボール 円城 塔	がらくた 江國 香織	コズミック・ゼロ：日本絶滅計画 清涼院 流水	ドーン 平野 啓一郎
1Q84 BOOK 1 村上 春樹	0.4264217	0.511542	0.5379467	0.5168985
1Q84 BOOK 2 村上 春樹	0.4399021	0.4994308	0.5382257	0.5343517
1Q84 BOOK 3 村上 春樹	0.4360048	0.5006697	0.5437662	0.532685
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.4355518	0.5201058	0.5270261	0.5473537
スクールアタック・シンドローム 舞城 王太郎	0.4039834	0.5160458	0.49738	0.5041751
みんな元気。 舞城 王太郎	0.3959493	0.50453	0.4877427	0.4772578
煙か土か食い物 舞城 王太郎	0.4396376	0.4725769	0.534007	0.5273731
好き好き大好き超愛してる。 舞城 王太郎	0.4367385	0.4981235	0.5362661	0.5137037
日本沈没 下 小松 左京	0.376512	0.4693894	0.5090624	0.4384827
日本沈没 上 小松 左京	0.3537479	0.4566091	0.4821687	0.4132543
東京タワー 江國 香織	0.3741821	0.6307548	0.4757475	0.4618332
冷静と情熱のあいだ-Rosso 江國 香織	0.3844771	0.6260888	0.4694634	0.4555975
AMEBIC 金原 ひとみ	0.4507987	0.4844184	0.4736391	0.4815351
IP／NN 阿部和重傑作集 阿部 和重	0.4627552	0.4852155	0.5584067	0.540843
TUGUMI 吉本 ばなな	0.3613974	0.5209175	0.4928517	0.4658026
イッツ・オンリー・トーク 絲山 秋子	0.3872895	0.5038212	0.4845504	0.4907832
オブ・ザ・ベースボール 円城 塔	1	0.3878233	0.4505105	0.4493506
がらくた 江國 香織	0.3878233	1	0.4879064	0.4711776
コズミック・ゼロ：日本絶滅計画 清涼院 流水	0.4505105	0.4879064	1	0.5435394
ドーン 平野 啓一郎	0.4493506	0.4711776	0.5435394	1
なんとなく、クリスタル 田中 康夫	0.419791	0.4674103	0.4960619	0.4695019
希望の国のエクソダス 村上 龍	0.4635497	0.5103704	0.5421216	0.5467959
虐殺器官 伊藤 計劃	0.4646949	0.5106019	0.5538069	0.5109843
蹴りたい背中 綿矢 りさ	0.3559914	0.5004229	0.4867673	0.4701528
重力ピエロ 伊坂 幸太郎	0.4302932	0.5055864	0.5261056	0.5387297
人のセックスを笑うな 山崎 ナオコーラ	0.385864	0.4917976	0.4888758	0.4818661
猛スピードで母は 長嶋 有	0.3854717	0.4876587	0.5058935	0.491555
優雅で感傷的な日本野球 高橋 源一郎	0.447146	0.4958225	0.5252635	0.5142791
腑抜けども、悲しみの愛を見せろ 本谷 有希子	0.4362006	0.4863845	0.5262881	0.5302782

	なんとなく、クリスタル 田中 康夫	希望の国のエクソダス 村上 龍	虐殺器官 伊藤 計劃	蹴りたい背中 綿矢 りさ
1Q84 BOOK 1 村上 春樹	0.4670798	0.5645461	0.5647969	0.5111159
1Q84 BOOK 2 村上 春樹	0.4650912	0.5720672	0.5750435	0.488943

1Q84 BOOK 3 村上 春樹	0.453138	0.5795441	0.5679391	0.4834574
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.4616966	0.5826234	0.5747702	0.4854689
スクールアタック・シンドローム 舞城 王太郎	0.49153	0.5454727	0.4988089	0.5338027
みんな元気。 舞城 王太郎	0.4737394	0.5351162	0.4887547	0.5158217
煙か土か食い物 舞城 王太郎	0.4957002	0.5634784	0.5086977	0.4861209
好き好き大好き超愛してる。 舞城 王太郎	0.4906859	0.5588732	0.5288394	0.5274167
日本沈没 下 小松 左京	0.446611	0.5112809	0.5083779	0.4482756
日本沈没 上 小松 左京	0.4314605	0.5206307	0.4971186	0.4123933
東京タワー 江國 香織	0.4462247	0.4986826	0.4921205	0.4830187
冷静と情熱のあいだ-Rosso 江國 香織	0.4467889	0.4614265	0.4799863	0.4808714
AMEBIC 金原 ひとみ	0.4463932	0.527467	0.4767436	0.4762375
IP／NN 阿部和重傑作集 阿部 和重	0.4890743	0.5696253	0.5516443	0.451928
TUGUMI 吉本 ばなな	0.4629375	0.516125	0.4937839	0.5297112
イッツ・オンリー・トーク 絲山 秋子	0.472394	0.5098716	0.4647768	0.5190244
オブ・ザ・ベースボール 円城 塔	0.419791	0.4635497	0.4646949	0.3559914
がらくた 江國 香織	0.4674103	0.5103704	0.5106019	0.5004229
コズミック・ゼロ： 日本絶滅計画 清涼院 流水	0.4960619	0.5421216	0.5538069	0.4867673
ドーン 平野 啓一郎	0.4695019	0.5467959	0.5109843	0.4701528
なんとなく、クリスタル 田中 康夫	1	0.5154528	0.4776576	0.4802929
希望の国のエクソダス 村上 龍	0.5154528	1	0.576069	0.4968801
虐殺器官 伊藤 計劃	0.4776576	0.576069	1	0.4673614
蹴りたい背中 綿矢 りさ	0.4802929	0.4968801	0.4673614	1
重力ピエロ 伊坂 幸太郎	0.4457239	0.5582354	0.5345413	0.5047258
人のセックスを笑うな 山崎 ナオコーラ	0.4900177	0.5299132	0.4822454	0.5286052
猛スピードで母は 長嶋 有	0.4810576	0.5341091	0.4924785	0.5204811
優雅で感傷的な日本野球 高橋 源一郎	0.4815762	0.5547952	0.5393121	0.459096
腑抜けども、悲しみの愛を見せろ 本谷 有希子	0.4788242	0.5300874	0.5123339	0.5282869

	重力ピエロ 伊坂 幸太郎	人のセックスを笑うな 山崎 ナオコーラ	猛スピードで母は 長嶋 有	優雅で感傷的な日本野球 高橋 源一郎	腑抜けども、悲しみの愛を見せろ 本谷 有希子
1Q84 BOOK 1 村上 春樹	0.5395872	0.4866948	0.5193771	0.52467	0.5017125
1Q84 BOOK 2 村上 春樹	0.5314578	0.5002271	0.5130085	0.5238524	0.5146143
1Q84 BOOK 3 村上 春樹	0.537996	0.4728178	0.4966821	0.5227642	0.5141759
色彩を持たない多崎つくると、彼の巡礼の年 村上 春樹	0.5348589	0.5032122	0.5195081	0.5229374	0.5215004
スクールアタック・シンドローム 舞城 王太郎	0.5545489	0.5409884	0.5147494	0.5339088	0.5157664
みんな元気。 舞城 王太郎	0.5405383	0.5446695	0.5116807	0.5208607	0.4938428
煙か土か食い物 舞城 王太郎	0.5572588	0.5069668	0.4965263	0.5427776	0.5303316
好き好き大好き超愛してる。 舞城 王太郎	0.5502876	0.5422506	0.5134822	0.5501636	0.5084612
日本沈没 下 小松 左京	0.4568884	0.4324502	0.4777333	0.4772462	0.4785802
日本沈没 上 小松 左京	0.4543096	0.4269999	0.4526707	0.4782974	0.4485792
東京タワー 江國 香織	0.4770056	0.485667	0.4830354	0.4837947	0.462399
冷静と情熱のあいだ-Rosso 江國 香織	0.4527502	0.4811446	0.4794857	0.4623391	0.4598773
AMEBIC 金原 ひとみ	0.5172837	0.5114338	0.4766968	0.4971	0.5011973
IP／NN 阿部和重傑作集 阿部 和重	0.5607927	0.4675545	0.4839081	0.541594	0.5183668

TUGUMI　吉本 ばなな	0.508662	0.5031269	0.5307074	0.5110599	0.4751895
イッツ・オンリー・トーク 絲山 秋子	0.5177349	0.5261135	0.5403884	0.522537	0.5053489
オブ・ザ・ベースボール 円城 塔	0.4302932	0.4355518	0.3854717	0.447146	0.4362006
がらくた 江國 香織	0.5055864	0.4917976	0.4876587	0.4958225	0.4863845
コズミック・ゼロ： 日本絶滅計画 清涼院 流水	0.5261056	0.4888758	0.5058935	0.5252635	0.5262881
ドーン 平野 啓一郎	0.5387297	0.4818661	0.491555	0.5142791	0.5302782
なんとなく、クリスタル　田中 康夫	0.4457239	0.4900177	0.4810576	0.4815762	0.4788242
希望の国のエクソダス 村上 龍	0.5582354	0.5299132	0.5311091	0.5547952	0.5300874
虐殺器官　伊藤 計劃	0.5345413	0.4822454	0.4924785	0.5393121	0.5123339
蹴りたい背中　綿矢 りさ	0.5047258	0.5286052	0.5204811	0.459096	0.5282869
重力ピエロ　伊坂 幸太郎	1	0.4882478	0.512514	0.5155197	0.5518752
人のセックスを笑うな　山崎 ナオコーラ	0.4882478	1	0.5378735	0.5177911	0.5095847
猛スピードで母は　長嶋 有	0.512514	0.5378735	1	0.4935893	0.5077567
優雅で感傷的な日本野球　高橋 源一郎	0.5155197	0.5177911	0.4935893	1	0.5061488
腑抜けども、悲しみの愛を見せろ　本谷 有希子	0.5518752	0.5095847	0.5077567	0.5061488	1

書籍の選択は、基本的に筆者の自宅にあったものを適当に選んだが、それでも、同一著者の本を複数解析する、また同一シリーズの本を複数解析する、いくらか異なるジャンルのものも入れる、といった観点も加味して選定している。

第一に、村上春樹氏、舞城王太郎氏はじめ、同一著者による作品間の相関指数は0.6以上であり、逆に言えば、著者が異なるのに、0.6以上の指数となる作品は一つもない。つまりこの指数は100%の精度で、ある作品がどの著者のものかを判定できることになる。

また、同一著者の作品のうち、同一シリーズの作品の相関はさらに強い。村上春樹氏の『1Q84』シリーズ（新潮社、2009）をみると、これは氏の最新作である『色彩を持たない多崎つくると、彼の巡礼の年』（文藝春秋、2013）よりも互いの相関指数が高い。

著者の異なる作品間を比べると、たとえば円城塔氏や小松左京氏の作品は、全体として他の作家との相関指数がきわめて低いことが分かる。これは彼らの作品が独特の表現を持っているであろうことを示唆する。円城氏も小松氏もＳＦジャンルの作品を書いているから、他の文芸作品と異なる傾向を持つのだ、と判断することもできる。しかしその円城氏と小松氏の作品群の相関指数は低く、同じくＳＦの書き手である伊藤計劃氏の『虐殺器官』（早川書房、2007）は、村上春樹氏や村上龍氏の作品と相関指数が高く、他の作家ともとりわけ相関指数が低いわけでもない。従って、円城・小松両氏の作品が、単にＳＦという理由で似ていない、というわけではなさそうである。

そもそも、円城氏の『オブ・ザ・ベースボール』は純文学系の文芸誌・文学界に掲載されたものであるから、ＳＦの書き手でありながらも、純文学の領域で活躍する円城氏の独自性が浮き彫りになったとみることもできる。また小松氏の『日本沈没』（小学館、1973）は発表時期がかなり前のものなので、文章の表現が変わっているのかも知れない。

数値をみていくと、どの作品とどの作品が似ているか、あるいは似ていないかが分かる。

ここに掲げたデータでは、0.55 以上の指数になると似ていると言え、逆に 0.45 を下回ると、傾向が違う作品と言えそうだ。

これは解析の一例であり、共起頻度表に基づく同様の解析結果もあるが、ここではそうした結果を羅列することを趣旨としないので、ここでは割愛する。

○今後の課題・応用の方向性

これまでに述べてきたような解析は、フィルタリングルールの設定によって、いかようにでもその解析の切り口を変化させることができる。作品間の相関を示したのは、あくまでごく一つの解析例に過ぎないことを忘れないでいただきたい。

もちろん、そうした相関分析も、単に２作品間の関係だけを計算して終わり、というわけではなく、様々な応用可能性がある。こうした相関データを用いれば、クラスター分析などの手法によって多数の作品を自動的にグルーピングすることもできる。こうすることで、これまで「ミステリー」とされていた作品が、実のところ「純文学」と呼ばれる作品群に近いことが分かったり、あるいはある「純文学」作品が「ＳＦ」に似ていた、といったことも分かるようになる。

文学研究に関することでは、これまで行われてきた、作家同士の影響関係を調べる研究の際に、そうした影響が実際の作品にどう反映されているかを、作品のテクスト分析と比較によって検証する、といったことも考えられる。

また、高名な文学賞を受賞した作品群を集め、それらのベクトルデータを合算・平均などの形で合成すれば、「文学的評価が高い作品の汎用ベクトル」をつくることができ、そこからあるテクストの文学的評価を予測することもできる。その予測精度をある程度以上に高められれば、文学における「高い評価」が導き出される根拠を、定量的なデータに基づいて知ることができる。

同様のやり方で、インターネット上の、読者による書評やレビューの情報と作品のベクトルデータを合わせることで、ある作品を読者がどのように評価しているか、またどのような傾向の作品に人気が出るかも知ることができるだろう。

もちろん、商業的な応用も可能だ。ベストセラー作品の特徴を抽出することにより、新しく発売する小説のヒットの確率やその度合いを判定したり、どういった文体を持つ作家をデビューさせると将来ヒット作を生み出せるかを予想したりすることが考えられる。

また、これまでインターネットの書籍通販などでは、主に顧客の購買履歴を参考にして、おすすめの作品を推薦するといった機能が使われてきた（これを協調フィルタリングという）。しかし購買履歴を参考にする限り、現時点で既に売れている作品同士の推薦しかできないことになる。となれば、発売したばかりの、新人作家による作品は、そうした推薦

を受けることがなく、売れないまま埋もれてゆくといったことになりかねない。もし、作品のテクスト解析結果に基づいて、「ある作品を読んだ人には、この作品が気に入ってもらえるはずだ」と判定できれば、現時点でまったく売れていない作品であっても、それを気に入るであろう読者のもとに的確に推薦することが可能になる。

そうした無数の応用の可能性が、そのまま、こうした解析技術にとっての課題となる。ここでは大量のテクストデータを定量的に扱うための一般的な理論と技法を紹介したに過ぎないが、本当に大切なのは、こうした考え方や技術を、実際の研究・生活に応用することで、文学作品の面白さに迫り、またそれを社会に広く伝えてゆくことであろう。

○おわりに

コンピューターを用いたこうした研究が、文学に何か危険な影響を与えるのではないか、と考える人がいないわけではない。しかし、実際に手を動かして作業をしてみれば、そうした心配はまったくの杞憂であることが分かる。文学作品は複雑であり、それを単純な解析で一刀両断することはできない。事実、コンピューターを用いた精緻な解析を積み重ねれば積み重ねるほど、むしろこれまで人間によって行われてきた評価や判断の短絡的であったことにこそ思い至る。物事を単純化し、その複雑さを忘却するのは常に人間の方であり、コンピューターは与えられた処理手順をただ高速にこなすことが仕事なのであって、やりようによっては、人間以上に複雑なものを複雑なまま扱うことが可能だ。

文学の奥深さに心打たれて文学研究を志したならば、そしてそうした奥深さを、短絡的な評価や、一面的な見方で否定されたくないのであれば、コンピューターを用いた精緻な分析は一つのきわめて有効な方法となり得ると筆者らは信じる。

注

1 村上征勝『真贋の科学』、朝倉書店、1994 年。

2 青空文庫 (http://www.aozora.gr.jp/)。

Contemporary Text Analysis: The Theory and Implementation

ISHII Daichi
In collaboration with
HASEGAWA Takanori
and CRUNCHERS Inc.

In this paper I introduced an original method to analyze Japanese literary texts combining morphological analysis with vector analysis. In this way each text is replaced to an ultra-high-dimensional vector which represents probability distribution of occurrence and collocation of words. Although the text itself is qualitative, the vector is quantative. This conversion makes it possible to programmably compare a large number of literary texts with realistic time and computing resources.