社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (42)
2007－BIO－11 (42)
2007／12／21

小説における文体印象解析の試み

望月 朝香　　鈴木 泰博
名古屋大学大学院情報科学研究科複雑系科学専攻

要旨

文章を読んだ後に「著者らしい雰囲気のする文章だった」と感じる経験はよくある。この目に見えない「雰囲気」、即ち作者特有の文体印象について、小説のテキストデータを数量化し分析することで考える。文体印象を醸し出す要素は様々考えられるが、本研究では①「句読点の分布」②「読みでの文字数」③「文章表現方法・多頻出言語」の3項目に着目する。句読点の使い方に著者特有の文構造が、読みでの文字数では頭の中のリズム、文章表現方法・多頻出言語から著者の愛用語や単語使用方法の癖が現れ、文体印象の要素となるためである。本論文では作者の特徴の抽出に相当する①と②について行う。かかる特徴付けを行い文体や作家の雰囲気を数量化することで、雰囲気を用いた文献検索を可能にし、また作者不詳の歴史文献の作者特定へ応用することが可能である。

1.はじめに

文体印象解析の従来の研究では①「句読点の分布」、②「読みでの文字数」、③「文章表現方法・多頻出言語」（以下では文体印象属性とよぶ）のうち①③を用い、新聞や雑誌記事等の文体の特徴付け[4]、作者未定の歴史文献における作者特定の試み[1,2,3]はあるが、さらに雰囲気の特徴付け・分類に着目し、かつ②を考慮した研究はまだ知られていない。

一般に、「雰囲気」は各々の読者にとっての内省的なものであり、また作品内容の影響も大きいため（例えば、ホラーとメルフェン）、そこに客観的な指標を導入することは困難である。しかし、読者は様々な文体印象属性の影響から作品や作家への印象を醸成している面は否めない。そこで、読者への内省的影響や作品論を検討するのではなく、作家の“筆癖”のような、文章に内在する一定のパターン、に着目することで雰囲気の数量化を試みる。作品論的な要素（読者への内省を喚起する）を省いても作者特有の雰囲気は消えずに残る例として、幅広いジャンルで小説を書く作家から一貫した同じ印象を受ける事がある。また、月日を重ねるにつれ思想やテーマが変わっていく作者の小説にも、一貫した雰囲気がある。これらは、文体から受ける印象も雰囲気の要因になっている。

また文体においても、同一の作家でも変化し易く、その変化は顕著にわかる。しかし変化しているに関わらず、以前と変わらずに受ける同じ雰囲気がある。かかる、普通に読んでいただけでは気付かない文体に内在する構造について本研究では①句読点の分布、②「読み」での文字数、③文章表現方法・多頻出言語、の3項目に着目する。これらに着目した理由は、それは句読点の使い方に著者特有の文構造、読みでの文字数では頭の中のリズム、多頻出言語から著者の愛用語や単語使用方法の癖が現れ、文体印象の要素となり得ると考えるからである。

2.方法

本研究では先に挙げた文体印象属性のうちまず①と②について検討する。

①構造（句読点の分布）

小説を読む過程において、読者は意図・意識せずして句読点により僅かな時間、読む流れが止められていると考える。これは、読点（「、」）が多いと読み辛いと感じ、逆に読点がない長文は意味も理解し難く疲れる、などといった経験からも確かである（そのばあ“読み辛い”、“疲れる”が雰囲気に相当）。句読点は文法的な文構造上で打たれる場合と、作者の判断に左右され打たれる場合とにわけられるが、本研究では後者の場合に着目する。それは、句点(「。」)を大きな一区切りと捉え、読点(「、」)はこの大きな一区切りを更に僅かな時間によって区切っていると考えるからである。この区切りの打ち方に作者の特徴があるのかを調べ、その構造を調査する。

②リズム（「読み」での文字数）

通常、文章はまず頭の中で語句として想起される。この語句が連なり、文となる。頭の中の語句は文字に置き換える前であり、書き手が好む自由なリズム（長さ）をもっている。小説のような文章を記述する際このリズムが、論文や新聞の文章のような‘型’の制約を受けないため比較的柔軟に書き出される。作者はこれらの過程で独自の判断により読点を打つ。

従って、①の読点の分布から、読点間の読み文字での長さをリズムとみなすと、作者特有のリズムを抽出する事が出来る。歴史文献

研究では文の長さを考える事は非常に多い。この時、文の長さとは漢字まじりの文字数で数えられたものである。句読点も一文字と数えている事も多い。しかしここで問題にしたい事は、単純に漢字まじりの文字を数えるのでは、作者の頭の中のリズムを抽出する事が出来ないという事である。何故ならば、漢字は「読み」での文字数と異なるからである。頭の中のリズムを抽出するならば、読み文字で数えるべきである。また、句読点は文章における区切り文字であり、実際に読むことはしない。寧ろ、頭の中でも間として読みの流れを区切っている。そこで本研究では句読点や鍵括弧等と言った区切り文字は数えず、そのまま読み文字でのリズムを区切る仕切りとして扱った。これにより、作者の頭の中のリズムを抽出できる。例えば、

山路を登りながら、こう考えた。

i）文字数＝14　　ii）読み文字数＝11、7

となる。日本の短歌や俳句にあるように、5,7,5,7,7　といったリズムは文字数ではなく、読み文字のリズムである。このことからも、小説の文章中にも作者のリズムによって組み立て等てられている個所があると考える事は有効であると考える。

2.1 データ

表１　用いた作品一覧

著者	作品名(年代)	全文の文字数	小説の長さ
森鷗外 (1862-1922)	「舞姫」(1890)	14215	短編
	「キタ(ハイタ)・セクスアリス」(1909)	53494	長編
	「青年」(1910)	113414	長編
	「阿部一族」(1913)	26537	短編
	「山椒大夫」(1915)	20065	短編
	「高瀬舟」(1916)	8114	短編
夏目漱石 (1867-1916)	「幻影の盾」(1905)	20694	短編
	「草枕」(1906)	76378	長編
	「虞美人草」(1908)	130478	長編
	「夢十夜」(1908)	14622	短編
	「こころ」(1914)	150545	長編
	「硝子戸の中」(1915)	49322	長編
芥川龍之介 (1892-1927)	「羅生門」(1915)	50096	長編
	「鼻」(1916)	5264	短編
	「蜘蛛の糸」(1918)	2834	短編
	「地獄変」(1918)	24339	短編
	「邪宗門」(1918)	38637	長編
	「河童」(1927)	34731	長編
宮沢賢治 (1896-1933)	「注文の多い料理店」(1921)	4738	短編
	「よだかの星」	4699	短編
	「セロ弾きのゴーシュ」	10863	短編
	「雪渡り」	9196	短編
	「氷河鼠の毛皮」(1923)	5594	短編
	「銀河鉄道の夜」(1927)	35490	長編
中島敦 (1909-1942)	「山月記」(1942)	5229	短編
	「文字禍」	6066	短編
	「李陵」(1942)	31886	短編
	「光と風と夢」	72089	長編
	「弟子」	22257	長編
	「狐憑」	4419	短編
太宰治 (1909-1948)	「思ひ出」(1933)	28611	短編
	「玩具」(1938)	4014	短編
	「富嶽百景」(1939)	13227	短編
	「走れメロス」(1940)	8649	短編
	「雪の夜の話」(1944)	4063	短編
	「人間失格」(1948)	65497	長編

実験の処理をコンピュータ上で扱うため、小説は電子図書館青空文庫のテキストデータを用いた。本研究では６名の著者による各６作品、計36作品を対象として実験を行った(表１)。作品の選択条件として、作品の書かれた時代が偏ることのないよう、ある作者の生存した時代で必ず他の比較対象作者がいることを条件とした。また、作者によって得意とする小説の長さ（長編または短編）が異なるが、各作者の長編と短編が含まれるように、かつ、書かれた時期が前期・中期・後期のものとして含まれるよう考慮した。

3.結果

①構造（句読点の分布）

一文における読点（「、」）の数を数え、頻度を調査した。ここで、一文とは句点（「。」）から次の句点までのことである。ただし、会話文や思考を表す際に使用される括弧などの囲み記号は、読書をする際にその部分で僅かな間流れを止められるとみなし、リズムを作る区切りとする。この区切りは、読点とは異なり新しい一文を作ると考え、句点と同様とした。以下、まとめて句点と呼ぶことにする。

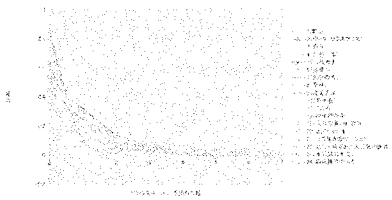

図１．一文の中の読点比率の比較

（森・夏目・宮沢）

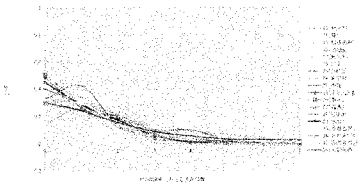

図２．一文の中の読点比率の比較

（芥川・中島・太宰）

36 作品における、一文中の読点数において

読点数比率＝(n 個の読点数を持つ文の総数 / 分析した文の総数)　(式１)

とし、n 個の読点数をもつ文の割合について調査した。図は、縦軸が読点数比率、横軸が一文における読点の数(つまり（式 1）における n)に対応する。分析により森鴎外・夏目漱石・宮沢賢治（図 1）と（図 2）がそれぞれ類似した分布をすることが示された。特定の作品に依らず森鴎外・夏目漱石・宮沢賢治は余り読点を打たないか一つだけ打つ傾向があり、芥川竜之介・中島敦・太宰治は 0 から 3,4 と一文中に打つ読点数がばらつく傾向がみられた。

②リズム（「読み」での文字数）

次に、句読点で区切られた、区切られ文の読み文字数を抽出しヒストグラムを作成し、の読み文字数にリズムに特徴が観られるかを調べた。

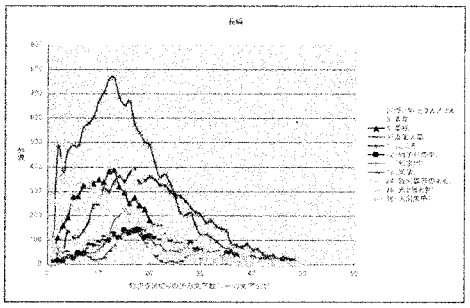

図 3.長編

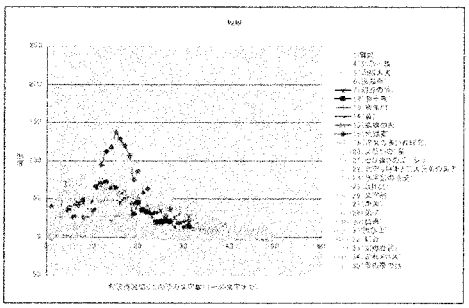

図 4.短編

上図は句読点でくぎられた文を読み文字数により数えた時の、数の頻度分布である。解析したデータには長編と短編の作品が含まれるため（表 1）、長編（図 3）と短編（図 4）に分けて分布を比較してみたが顕著な違いはみられなかった。これより、文章の長さも影響は受けないことが解る。

作家毎に傾向を調査してみるとリズムについては、作品によらず、森鴎外（図 5）と夏目漱石（図 6）、芥川龍之介（図 7）と宮沢賢治（図 8）そして中島敦（図 9）と太宰治（図 10）それぞれが類似した傾向をもつことが示された。

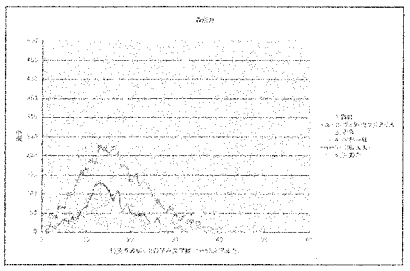

図 5.森鴎外

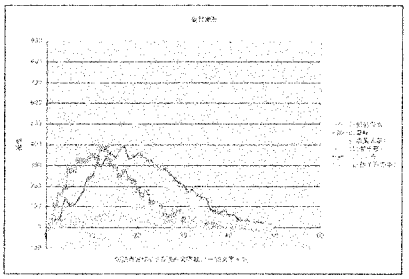

図 6.夏目漱石

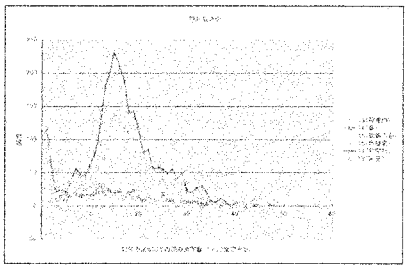

図 7.芥川龍之介

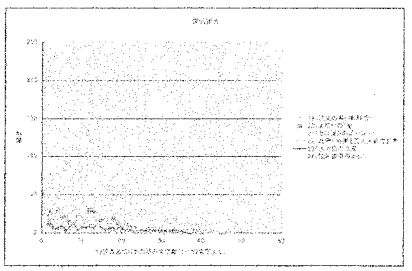

図 8.宮沢賢治

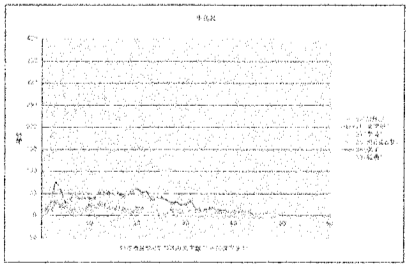

図 9.中島敦

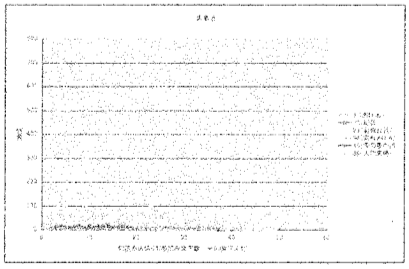

図 10.太宰治

4.おわりに

本研究では文の構造とリズム（文体印象属性①と②）を用い分析を行い、特に作者により読み文字数における特有のリズムがあることが明らかになった。このことから、作者特有の雰囲気を特定する為に読み文字数のリズムを用いることは有効であるということが見出された。今後、この読み文字数のリズムを用いた実験を行い、その結果を用いて雰囲気の特定に繋げていきたい。

今後の課題として以下の①から④について考察したい。

① ③文章表現方法・多頻出言語について、①,②の解析で重要となった語句について切り出し文を抽出し、その構造を調べ、読んだと時に受ける印象を検討する。

② リズムの「組み合わせ」の頻度を調べ、更にそのリズムをもつ文を抽出し、印象を検討する。

③ 形態素解析器 cabocha[5]を用いて文節の切り出しを行い、出現頻度を調べ、そこから印象の検討を行う。

④ 以上の実験をまとめ、雰囲気特定に結びつける手法を考える。

付録

日本語文章の「読み」での文字数抽出方法

手順 1） 青空文庫のテキストデータにはルビや注釈などが含まれている。これらを削除したものを以下「テキストデータ」と呼ぶ。

手順 2） 形態素解析器 Mecab[6]の-O オプションを用いて、テキストデータを全てカナ文字に変換する。これにより文章中の漢字を、読みでの文字数として数えることが可能となる。《問題点として、①Mecab[6]で認識されなかった漢字はそのまま残る、②読みが間違っている（例：「一人」を ‘いちにん’ と出力する）など誤差が含まれる事が挙げられる。誤差率を出すことは今後の課題であるが、ここでは誤差についての問題は扱わない。》

手順 3） カナ化されたテキストデータを、句読点毎に区切り、文字数を数える。

参考文献

[1] 師 茂樹「N グラムモデルとクラスタ分析を用いた漢文古典テキストの比較研究―『般若心経』の異訳の比較を例に」（京都大学大型計算機センター第 69 回研究セミナー「東洋学へのコンピュータ利用」予稿集、2002 年 3 月）

(http://www.ya.sakura.ne.jp/~moro/resources/20020322moro.pdf)

[2] 近藤 泰弘・近藤 みゆき『平安時代古典語古典文学研究ための N-gram を用いた解析手法』（言語情報処理学会第 7 回年次大会『発表論文集』2001）

(http://klab.ri.aoyama.ac.jp/public/paper/20010328.pdf)

[3] 金 明哲『計量文体学からみたテキストマイニング』
(http://www1.doshisha.ac.jp/~mjin/text/2007_05.pdf)

[4] 石田栄美・安形輝・野末道子・久野高志・池内淳・上田修一『文体からみた学術的文献の特徴分析』
(http://www.slis.keio.ac.jp/~ueda/webir/webir041.pdf)

[5] CaboCha/ 南瓜： Yet Another Japanese Dependency Structure Analyzer
(http://chasen.org/~taku/software/cabocha/)

[6]MeCab: Yet Another Part-of-Speech and Morphological Analyzer
(http://mecab.sourceforge.net/)