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

小説を読んだ後に「作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。この客観的に示す事が難しい「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察する。文体印象を形成する要因は様々考えられるが、本論文では①「読点」②「読みでの文字数」と、そこから生じる③「リズム」に着目する。読点により単文を分割し、その読点間の文字数を訓読みで測定する。この読み文字数の傾向から、作者の執筆中に頭の中で想起されやすい長さの特徴を分析する。また、単文における読点数の推移をリズムとみなすことで、文章全体のリズムを考察する。本研究では、かかる、普通に読書をしていただけでは気付きにくい要因が読者に影響を及ぼし、「ーっぽい」といった曖昧な作者の文体印象が形成されると考える。従って①ー③について分析、検討した。

1.はじめに

一般に文体印象は、読者の過去・日常の経験に起因した個人的要因に影響される面を有し、そこに客観的な指標を導入することは困難である。従って本論文では個人的影響は扱わず、作者の”筆癖”のような、文章に内在する一定のパターンを抽出し、その影響による印象形成を考える。そのためテキストデータを数量化し、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分析し印象解明を試みる。

計量文献学における従来研究では「句読点」「文字数」「文章表現方法」「多頻出言語」などを文章に内在するパターンとして用い、新聞や雑誌記事などの文体の特徴付け[5]、作者未定の歴史文献等における作者特定の試み[1,2,3,4]などの研究が多数なされてきた。しかしこれらの研究では、作者の文体から形成される文体印象についてはどれもまだ言及されておらず、印象分析の方法も提案されていない。そこで本研究では、計量文献学の成果を応用した作者の文体印象の特徴付け方法を提案する。

2.方法

①読点

作者毎に特徴が現れるとの報告がある読点[5]に着目する。読点を利用した数量化を行う事は、作者の文体印象に影響を及ぼす要素を得る可能性が高いといえる。加えて、小説を読む過程において、読者は無意識のうちに句読点で読む流れが止められ、その影響を受けている。例えば読点が多い場合だと＜読み辛い＞と感じ、逆に読点がない場合は＜意味が解り辛い、疲れる＞と感じる場合が多いが、これは読点の打ち方により文体印象が形成されていると言える。以上の観点から読点に着目した計測を行う。

まず、研究に用いた作品（表 1）の、句点から次の句点まで（以下より単文とよぶ）における読点の打ち方を測定し、n 個の読点を持つ単文について

$$\text{読点数比率}=\frac{\text{n個の読点を持つ単文総数}}{\text{分析した単文総数}}$$

（式 1）

を調査した。

本論文では、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、長さを作る区切りとする。この区切りは、単文の間を一時的に区切る読点とは異なり、新しく単文を作るものと考え、句点と同様とした。以下、まとめて句点と呼ぶことにする。

②読みでの文字数

羊博士は黙り込んで、

つつらのような白い眉毛を指でこすった。

文字数＝9、18

読み文字数＝13、21

通常、執筆するにあたり、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、文章となる。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長さをもっている。小説のような文章を記述する際この長さが、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、作者独自の判断により句読点を打つと考えられる。一般に漢字混じりの文字数と読みでの文字数では、差がある場合が多いため、句読点間の文字数をそのまま漢字混じりの文字数で計測するのではなく、訓読みで数えた時の「読みの文字数」として計測し、作者が頭の中で想起した言葉の長さを抽出する。これは文体印象を分析するうえで必要なことである。以上より読点を利用し、句読点間の読み文字数の測定を行う。尚、テキス

トデータを読みで表示するにあたり形態素解析器 MeCab[6]を用いた。

③リズム

今、単文に用いられる読点数は、前の単文に使用した読点数に何らかの影響を受けていると考える。例えば読点数 0 の単文の後には倍以上の読点数をもつ単文が来る傾向がみられた場合、文体印象は＜躍動的だ＞となる可能性もある。このように、前の単文との読点数の推移は文体印象と関連性があると考えられる。また、読書中の、無意識のうちにこのリズムから影響を受けている可能性がある。そこで、まず単文中の読点数を、小説の初めから終わりまで順番に並べ前後とのペアリングを行う。この作業により、読点の推移を得る。次に、このペアリングの種類と頻度を測定する。次に、あるペアリングは全体のペアリングの種類の中でどれだけ出現する可能性があるか割合を算出する。これにより、読点数の推移確率を得る。この推移確率を考察することで文体印象を考える。

3.データ

表 1.作品一覧

著者	作品名(年代)
森鴎外 (1862-1922)	「舞姫」(1890)
	「オタ・セクスアリス」(1909)
	「青年」(1910)
	「阿部一族」(1913)
	「山椒大夫」(1915)
	「高瀬舟」(1916)
夏目漱石 (1867-1916)	「幻影の盾」(1905)
	「草枕」(1906)
	「虞美人草」(1908)
	「夢十夜」(1908)
	「こころ」(1914)
	「硝子戸の中」(1915)
芥川龍之介 (1892-1927)	「羅生門」(1915)
	「鼻」(1916)
	「蜘蛛の糸」(1918)
	「地獄変」(1918)
	「邪宗門」(1918)
	「河童」(1927)
宮沢賢治 (1896-1933)	「注文の多い料理店」(1921)
	「よだかの星」
	「セロ弾きのゴーシュ」
	「北守将軍と三人兄弟の医者」
	「氷河鼠の毛皮」(1923)
	「銀河鉄道の夜」(1927)
中島敦 (1909-1942)	「山月記」(1942)
	「文字禍」
	「李陵」(1942)
	「光と風と夢」
	「弟子」
	「狐憑」
太宰治 (1909-1948)	「思ひ出」(1933)
	「玩具」(1935)
	「富嶽百景」(1939)
	「走れメロス」(1940)
	「雪の夜の話」(1944)
	「人間失格」(1948)

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。

4.実験と結果

①読点

単文における読点の数を数え、頻度を調査した。

図 1、図 2 では、縦軸が読点数比率、横軸が単文における読点の数(（式１）における n)に対応する。以上の分析により森鴎外・夏目漱石・宮沢賢治（図３）と（図４）がそれぞれ類似した分布をすることが示された。つまり、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 と単文中に打つ読点数がばらつく傾向がみられた。

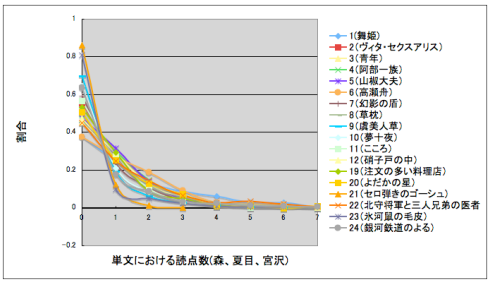

図 1. 森・夏目・宮沢

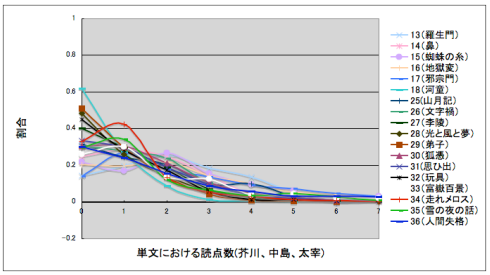

図 2. 芥川・中島・太宰

②読みでの文字数

次に、読みに変換した各テキストにおいて、句読点で区切りその間の読み文字数を抽出し、割合を調査した。解析したデータには長編と短編の作品が混合しているため、長編（図３）と短編（図４）に分けて分布を比較した。これは、例えば長編の場合は読み文字数が長くなる、といった小説の長さに影響を受けるのかを調べるためである。結果、どちらも 15 文字前後の頻度を頂点とした山形の形状と、５文字前後を頂点とし右側にテールが見られる形状が得られ、顕著な違い

はみられなかった。これにより読みの流れを区切る間隔は短編・長編といった小説の長さには左右されないことが確認された。

作家毎に傾向を調査してみると、作森鴎外（図５）と夏目漱石（図６）、芥川龍之介（図７）と宮沢賢治（図８）、中島敦（図９）と太宰治（図１０）のそれぞれが類似した傾向をもつことが示された。　また、作者自身の作品間の形状は、他の小説の形状と比較すると、非常に類似している。

従って、作者が小説を書くときに、頭の中で想起する言葉の長さには、自身特有のものと、例えば森と夏目が類似しているといった大まかな分類が出来ることが明らかになった。

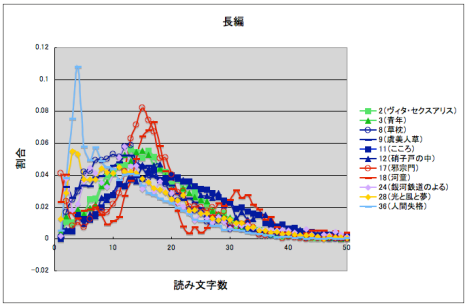

図 3.長編

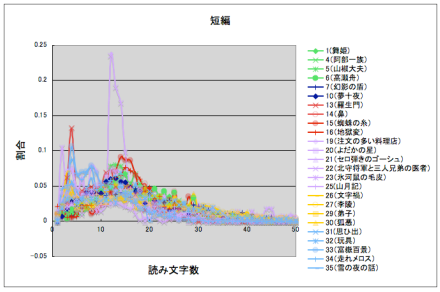

図 4.短編

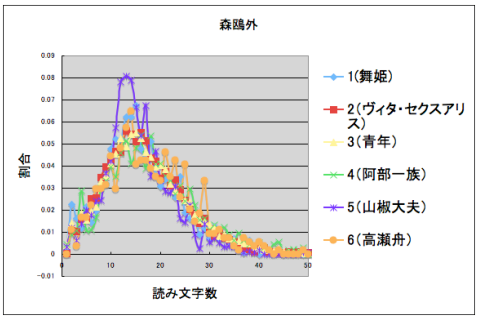

図 5.森鴎外

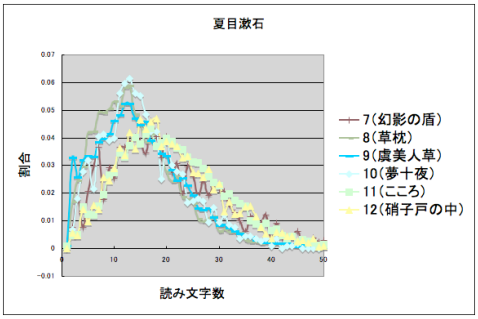

図 6.夏目漱石

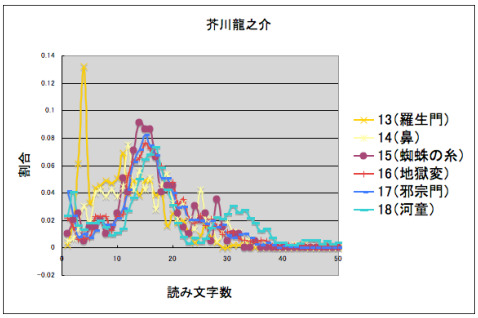

図 7.芥川龍之介

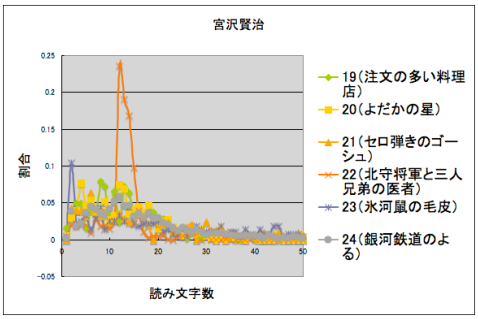

図 8.宮沢賢治

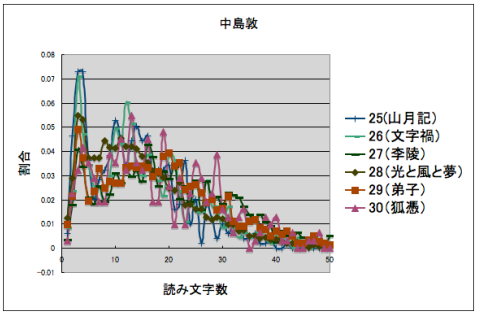

図 9.中島敦

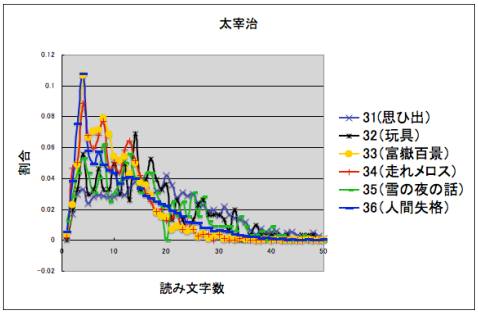

図 10.太宰治

③リズム

図 11 に､36 作品の推移確率を示す｡見方を説明する。左から、右に向かって順に作品が並んでいる。順序は表 1 の上から下への順序に対応している。つまり一番左列が森鴎外の「舞姫」の情報であり、右端が太宰治の「人間失格」の情報となっている。また、上から下にむかって読点の推移を表している。つまり、一番上の行は読点 0 の単文のあとに読点 0 の単文が続く。この推移を 0-0 と表記すると、２行目は 0-1, ３行目は 0-2・・・と、読点 0 の単文の後に続く読点数が 0ー6 までを表している。7 段目は 1-0、1-1、1-2 となっており、読点 1 の後に続く単文の読点数を 0ー6 まで表している。このようにして、0-0 から 6-6 までの 49 通りの推移における、各作品の、確率を表している。推移確率が 0.001 より低い場合は白、0.1 以上 0.2 以下は青、0.2 より高い場合は赤で表表示している。結果、一般的に読点 0 の単文のあとには、同じく読点 0 の単文がくることが多い。ここでは数値を表示していないため確認出来ないが、夏目は 0-0 の確率が他と比較して非常に高かった｡虞美人草に至っては 0.499 となっている。また、芥川は推移の組み合わせが多いことが確認出来る。一般に読点数が多ければ単文の長さは長くなるという、読点数と単文の長さ関係を考えると、芥川は長文と短文の繰り返しも多く、また②の句読点間の長さの結果において、グラフのうねりを考慮すると、文章のリズムが動的であり、単調ではない、＜跳躍的、変化にとんだ、鮮明だ＞といった文体印象を形成する可能性がある。一方、宮沢は読点数が 0 から 2 の間で単文を書く傾向があることがわかる。②の結果と照らし合わせ考えると、一定の読み文字数で、変化の少ないリズムを繰り返す小説が多いと言える。結果＜軽い印象、穏やか、閉静だ＞といった柔らかい印象が形成される可能性がある。

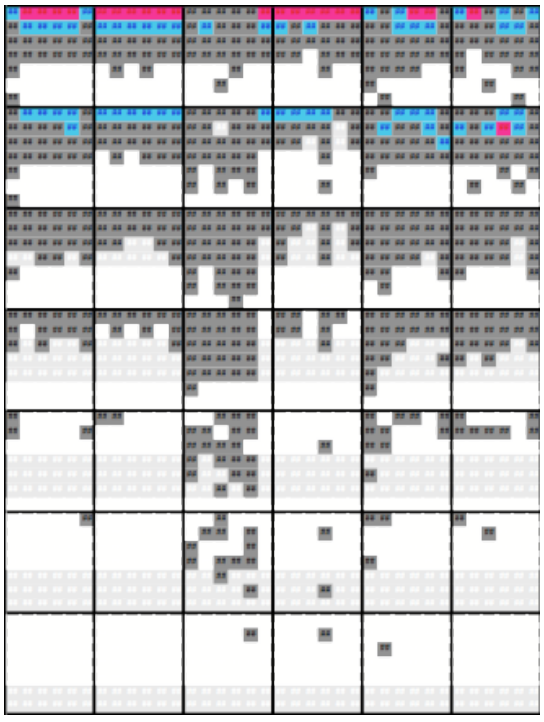

図 11.読点数の推移確率

5.おわりに

本研究では、作者毎に特徴が現れると確証された読点を用い、そこから文体印象を分析する為のテキストの情報抽出に、句読点間の読み文字数と読点の推移という方法を挙げた。

文章をリズムという観点から分析を行い、特に作者により読み文字数に差があることが明らかになった。今後、②で得られる読み文字数においても推移確率を算出し作者のリズムを抽出する。日本の短歌や俳句にあるような　5,7,5,7,7　といった、日本人が心地よいという印象を受ける七五調のリズムは、文字数ではなく読み文字数のリズムである。このことからも、小説の文章中には、作者の好む特有のリズムが存在すると考えられる。その結果と今回の結果をふまえて雰囲気の特定に繋げていきたい。また、今回は結果における心理学的検証は行わなかったが、有用性を確かめるためには実際の印象に結びつくか心理学的な検証も不可欠である。

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]　村上征勝
(http://www.sony.co.jp/Products/SC-HP/cx_pal/vol43/pdf/moving.pdf)
[5]　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)
[6]MeCab: Yet Another Part-of-Speech and Morphological Analyzer
(http://mecab.sourceforge.net/)