# 文章の統計的研究
## —芥川賞作家と直木賞作家の作品を中心として—

2006MI088 黒部由利絵　2006MI192 恒川紗帆
指導教員：木村美善

## 1　はじめに

計量文献学とは，文学作品や哲学書，歴史書などの分析を計量的な手法を用いて行う研究分野である．作家の文章を計量化して，作品の真贋の判断や作者の癖や特徴を読み取るなどの研究が行われている．私たちは，文章の句読点の数や付ける位置，文体の特性などを分析することによって，作者を割り出すことや作品の謎を解くことにたいへん興味をもった．そこで，芥川賞作家の作品と直木賞作家の作品からデータを集め，作家の癖や特徴，違いや共通点などについて調べることを卒業研究のテーマに選ぶことにした．

## 2　データについて

芥川賞作家 15 人と直木賞作家 15 人の計 30 人を分析対象とし，それぞれの作品を 3 作品ずつ計 90 作品を取りあげることにした．また，各受賞者 15 人のうち 1990 年以降の作家を現代作家，それ以前の作家を非現代作家として，前者 7 人，後者 8 人の計 15 人とした（[1] 参照）．

芥川賞受賞者の現代作家として，吉田修一，荻野アンナ，川上弘美，辻仁成，藤沢周，綿矢りさ，絲山秋子を選び，非現代作家として，井上靖，石原慎太郎，北杜夫，津村節子，村上龍，宮本輝，高樹のぶ子，松本清張を選んだ．直木賞受賞者の現代作家として，宮部みゆき，唯川恵，石田衣良，江国香織，熊谷達也，朱川湊人，東野圭吾を選び，非現代作家として，城山三朗，司馬遼太郎，黒岩重吾，五木寛之，井上ひさし，林真理子，逢坂剛，白石一郎を選んだ（図では，各作者の苗字とローマ数字で示してある）．

読点前の文字の使用率として，各作品から 800 字ずつを 5 箇所選び，計 4000 字の中で読点の前に使われている文字を「と」「て」「は」「が」「で」「に」「ら」「も」「し」「を」「り」「の」「く」「時」「か」「ば」「た」「い」「後」「ず」「れ」「き」「る」「う」「その他」の 25 文字の場合で数え，それぞれの使用率を調べた．文体の特性として，読点前の文字の使用率で選び出した 4000 字のうち，「直喩」「声喩」「色彩語」「人格語」「文の長さ」「句点」「読点」「漢字」「現在止」「過去止」「不定止」の 11 項目の数を数えた．因子分析では，直喩，声喩，色彩語，人格語，文の長さ，句点，読点，漢字の 8 項目（文体 1）と文の長さ，過去止，現在止，不定止の 4 項目（文体 2）に分けて分析を行った（[4]，[5] 参照）．

## 3　分析方法

分析には因子分析，クラスター分析，多次元尺度法を用いた（[2]，[3] 参照）．

## 4　読点前の文字の使用率（非現代作家）

### 4.1　因子分析

主成分分析の結果，因子数を 4 とすることが妥当であるという結果を得た，寄与率を上げるために「と」「に」「も」「り」「の」「ば」「ず」「れ」「き」「る」の 10 変数を減らして分析を行った．第 4 因子までの累積寄与率は 53.0%であった．

- 第 1 因子は「時」「か」「た」「その他」の因子負荷量が正に大きく，「が」「で」の因子負荷量が負に大きい．『主語や名詞の後以外の位置で読点を多く用いる因子』と名づける．
- 第 2 因子は「を」「後」「う」の因子負荷量が正に大きく，「で」「く」の因子負荷量が負に大きい．『仮定文を用いて文を区切り，内容を表現する因子』と名づける．
- 第 3 因子は「ら」「た」「い」の因子負荷量が正に大きく，「が」の因子負荷量が負に大きい．『動詞の後に読点を用いて，文を区切る因子』と名づける．
- 第 4 因子は「て」「く」「その他」の因子負荷量が正に大きく，「は」「し」の因子負荷量が負に大きい．『否定の文を用いて内容を表現する因子』と名づける．

図 1，2，3 は，第 1 因子から第 4 因子までのバイプロット図である．

外れ値である村上龍の作品は，仮定文や否定文を用いて文章を表現する特徴があると考えられる．また，村上龍と井上ひさしの作品が同じ位置にプロットされていることから，2 作品は，書き方が似ており，特に「い」の後に読点を用いる傾向があると言える．

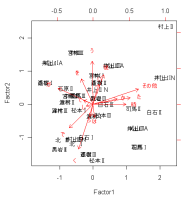


図 1　第 1・第 2 因子

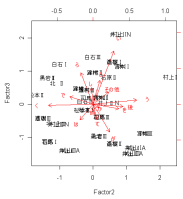


図 2　第 2・第 3 因子

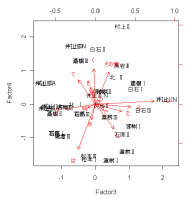


図 3　第 3・第 4 因子

### 4.2　クラスター分析

因子分析と同じ変数を用いてクラスター分析を行った結果，図 4 のようなデンドログラムが得られた．デンドログラムの左から，第 1 群，第 2 群，第 3 群，第 4 群と 4 つの群に分けて意味付けを行った．

- 第 1 群は，『「が」の後に読点を多く用いる群』と名づける．

- 第 2 群は,『「が」の後に読点をあまり用いず,「で」の後に多く用いる群』と名づける.
- 第 3 群は,『「が」「で」「は」の後にあまり読点を用いない群』と名づける.
- 第 4 群は,『「が」「で」の後にあまり読点を用いず,「は」の後に多く用いる群』と名づける.

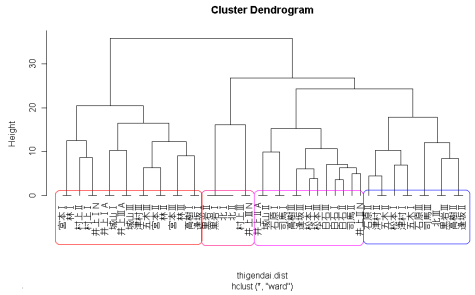


図 4　デンドログラム（読点・非現代作家）

### 4.3　多次元尺度法

相関係数を用い非計量ＭＤＳ関数で，第 3 次元までの分析を行った．ストレス値は 0.0905 であった．グラフから以下のことが読みとれる．

第 1 次元は「は」を表し，第 2 次元は「が」を表している．よって第 1 次元と第 2 次元のグラフは右上にあるものほど,「は」と「が」の後に読点を用いる割合が大きい作品と言える．第 3 次元は「し」を表している．よって第 2 次元と第 3 次元のグラフは右上にあるものほど,「が」と「し」の後に読点を用いる割合が大きい作品と言える．第 3 次元と第 1 次元のグラフは右上にあるものほど,「し」と「は」の後に読点を用いる割合が大きい作品と言える．

多次元尺度法の結果，井上靖と城山三郎の作品が「は」「が」の後に読点をよく用いることや，松本清張，津村節子，五木寛之の作品は「が」の後には少なく「は」の後にはよく読点を用いることが分かる．これは，因子分析やクラスター分析の結果とほぼ一致する．

## 5　読点前の文字の使用率（現代作家）

### 5.1　因子分析

主成分分析の結果，因子数を 3 とすることが妥当であるという結果を得た，非現代作家と同様に「で」「し」「を」「の」「た」「い」「き」「る」「う」の 9 変数を減らして分析を行った．第 4 因子までの累積寄与率は 50.3%であった．

- 第 1 因子は「て」「は」「が」「に」「ら」の因子負荷量が正に大きく,「ず」の因子負荷量が負に大きい.『読点を多く用いる因子』と名づける.
- 第 2 因子は「と」「て」「に」「も」「か」「その他」の因子負荷量が正に大きく,「後」「ず」「れ」の因子負荷量が負に大きい.『文章の背景の表現に関する語に読点を用いる因子』と名づける.
- 第 3 因子は「く」「ば」の因子負荷量が正に大きく,「り」の因子負荷量が負に大きい.『仮定文を用いて文を区切り，内容を表現する因子』と名づける.

図 5，6 は第 1 因子から第 3 因子までのバイプロット図である．

外れ値である藤沢周の作品は，第 2 因子に大きく関係しており，文章の背景の表現に関する語に読点を用いるという特徴があると言える．東野圭吾の作品は，2 作品が特に第 3 因子に大きく関係していると考えられ，3 作品が近くにプロットされていることから，仮定文を用いて文を区切り，内容を表現している特徴があると言える．

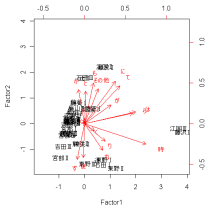


図 5　第 1・第 2 因子

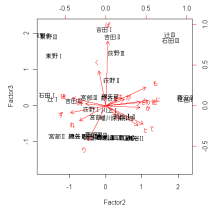


図 6　第 2・第 3 因子

### 5.2　クラスター分析

因子分析と同じ変数を用いてクラスター分析を行った結果，図 7 のようなデンドログラムが得られた．デンドログラムの左から，第 1 群，第 2 群，第 3 群，第 4 群の 4 つの群に分けて意味付けを行った．

- 第 1 群は,『「て」の後に読点をあまり用いないが,「が」の後に読点を多用いる群』と名づける.
- 第 2 群は,『「て」「が」の後に読点をあまり用いない群』と名づける.
- 第 3 群は,『「て」「が」の後に読点を多く用いる群』と名づける.
- 第 4 群は,『「て」の後に読点を多く用い,「が」の後にあまり用いない群』と名づける.

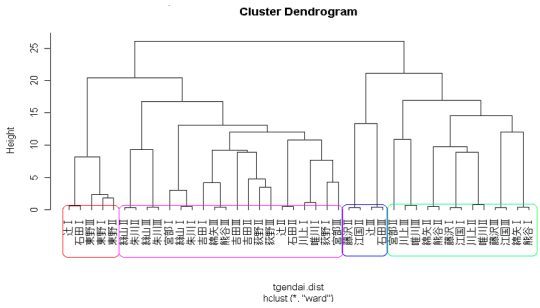


図 7　デンドログラム（読点・現代作家）

### 5.3　多次元尺度法

相関係数を用い非計量ＭＤＳ関数で，第 4 次元までの分析を行った．ストレス値は 0.0929 であった．グラフから以下のことが読みとれる．

第 1 次元は「は」を表し，第 2 次元は「と」を表している．よって第 1 次元と第 2 次元のグラフは右上にあるものほど,「は」と「と」の後に読点を用いる割合が大きい作品と言える．第 3 次元は「て」を表している．よって第 2 次元と第 3 次元のグラフは右上にあるものほど,「と」

と「て」の後に読点を用いる割合が大きい作品と言える．第 4 次元は「その他」を表している．よって第 3 次元と第 4 次元のグラフは右上にあるものほど,「が」と「その他」の後に読点を用いる割合が大きい作品と言える．第 4 次元と第 1 次元のグラフは右上にあるものほど,「その他」と「は」の後に読点を用いる割合が大きい作品と言える．

多次元尺度法の結果，絲山秋子と朱川湊人の作品は「と」「て」の後にあまり読点を用いないことや，川上弘美と唯川恵の作品は「と」「て」の後に読点を多く用いることが分かる．これは，因子分析とクラスター分析の結果とほぼ一致する．

## 6　文体の特性（非現代作家）

### 6.1　因子分析（文体 1）

主成分分析の結果，因子数を 2 とすることが妥当であるという結果を得た．第 2 因子までの累積寄与率は 48.1%であった．

- 第 1 因子は「句点」「文の長さ」に負荷量が大きい．『比較的 1 文が長文な因子』と名づける．
- 第 2 因子は「色彩語」に負荷量が大きい.『幅広い表現方法を使用している因子』と名づける．

図 8 は，第 1 因子と第 2 因子のバイプロット図である．

外れ値である井上ひさしのと五木寛之と司馬遼太郎作品は「変わった表現方法を用いて文章を構成しない作品」である．逢坂剛の作品は「幅広い表現方法を用いている作品」である. 因子分析の結果，芥川賞作品と直木賞作品はそれぞれ近いプロットになった．このことより非現代作家において，芥川賞作家と直木賞作家にはそれぞれ，文章構成に特徴や癖があると考えられる．

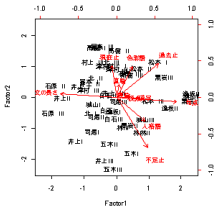

図 8　第 1・第 2 因子

### 6.2　因子分析（文体 2）

文体 2 の 4 項目を用いて主成分分析を行い，因子数を 1 とすることが妥当であるという結果を得た．因子数を 1 とし，寄与率を上げるために「現在止」の変数を減らして分析を行った．第 1 因子の寄与率は 53.2%であった．

第 1 因子は「文の長さ」に因子負荷量が大きい.『比較的 1 文ごとが長い因子』と名づける．

### 6.3　クラスター分析

文体の特性のデータを用いてクラスター分析を行った結果，図 9 のようなデンドログラムが得られた．

デンドログラムの左から，第 1 群，第 2 群と 2 つの群に分けて意味付けを行った．

- 第 1 群の作品は人格語が少ない傾向がある．
- 第 2 群の作品は人格語が多い傾向がある．

また第 1 群には芥川賞作家と井上ひさしの作品，第 2 群には直木賞作家が集まっていることから各賞の作家には，人格語の使用率に違いがあると考えられる．

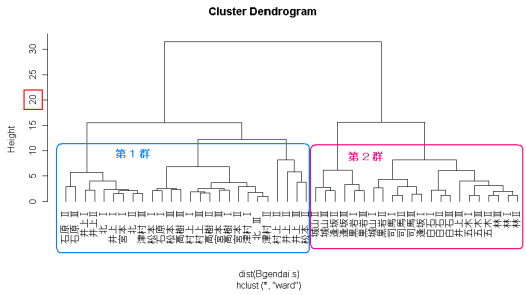


図 9　デンドログラム　（文体・非現代）

### 6.4　多次元尺度法

相関係数を用い非計量ＭＤＳ関数で，第 4 次元までの分析を行った．ストレス値は 0.0706 である．グラフから以下のことが読みとれる．

第 1 次元は「人格語」を表し，第 2 次元は「文の長さ」を表している．よって第 1 次元と第 2 次元のグラフは右上にあるものほど,「人格語」と「文の長さ」が多い作品と言える．第 3 次元は「漢字」を表している．よって第 2 次元と第 3 次元のグラフは右上にあるものほど,「文の長さ」と「漢字」が多い作品と言える．第 4 次元は「直喩」を表している．よって第 3 次元と第 4 次元のグラフは右上にあるものほど,「漢字」と「直喩」が多い作品と言える．第 4 次元と第 1 次元のグラフは右上にあるものほど,「直喩」と「人格語」が多い作品と言える．

多次元尺度法の結果，人格語の頻度が芥川賞作家は少なく，直木賞作家は多いことが分かった．これはクラスター分析の結果と一致する．

## 7　文体の特性（現代作家）

### 7.1　因子分析（文体 1）

主成分分析の結果，因子数を 4 とすることが妥当であるという結果を得た．第 4 因子までの累積寄与率は 66.9%であった．

- 第 1 因子は「句点」「文の長さ」に因子負荷量が大きい.『１文が比較的長い因子』と名づける．
- 第 2 因子は「読点」「漢字」に因子負荷量が大きい.『漢字表現・読点を用いて文章が分かりやすい因子』と名づける．
- 第 3 因子は「声喩」に因子負荷量が大きい.『声喩を使用している因子』と名づける．
- 第 4 因子は「色彩語」「人格語」に因子負荷量が大きい.『人格語と色彩語を用いて表現の幅を広げている因子』と名づける．

図 10，11，12 は，第 1 因子から第 4 因子までのバイプロット図である．

外れ値である井上ひさしの作品は「1 文が短く構成されている作品」，逢坂剛の作品は「幅広い表現方法を用い

ている作品」と言える．

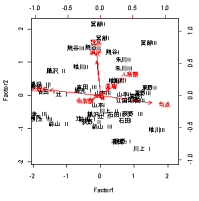

図 10　第 1・第 2 因子

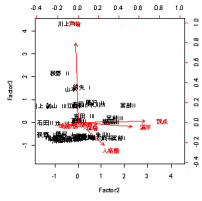

図 11　第 2・第 3 因子

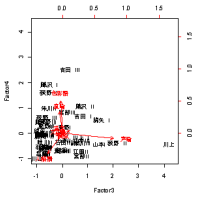

図 12　第 3・第 4 因子

### 7.2　因子分析（文体 2）

非現代作家（文体 2）の因子分析と同様で，主成分分析より因子数を 1 とすることが妥当であるという結果を得た．因子数を 1 とし，寄与率を上げるために「現在止」の変数を減らして分析を行った．第 1 因子の寄与率は 50.2%であった．

第 1 因子は「文の長さ」「不定止」に因子負荷量が大きい.『1 文が長く，その終わり方が時制を含まない因子』と名づける．

### 7.3　クラスター分析

文体の特性のデータを用いてクラスター分析を行った結果，図 13 のようなデンドログラムが得られた．デンドログラムの左から，第 1 群，第 2 群，第 3 群と 3 つの群に分けて意味付けを行った．

- 第 1 群の作品は現在止が多い傾向がある．
- 第 2 群の作品は現在止と漢字が少ない傾向がある．
- 第 3 群の作品は現在止は少なく，漢字が多い傾向がある．

また，第 1 群には芥川賞作家，第 2 群と第 3 群には直木賞作家と川上弘美，辻仁成，吉田修一が集まっていることから，各賞の作家には文章の終わりに違いがあると考えられる．

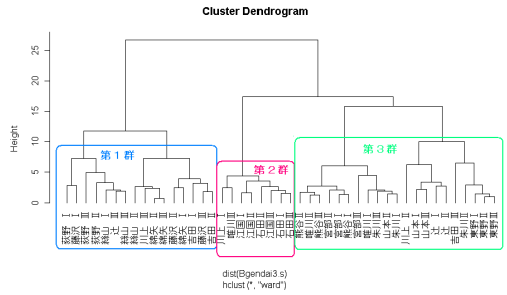


図 13　デンドログラム（文体・現代作家）

### 7.4　多次元尺度法

相関係数を用い非計量ＭＤＳ関数で，第 4 次元までの分析を行った．ストレス値は 0.0679 であった．グラフから以下のことが読みとれる．

第 1 次元は「文の長さ」を表し，第 2 次元は「漢字」を表している．よって第 1 次元と第 2 次元のグラフは右上にあるものほど,「文の長さ」と「漢字」が多い作品と言える．第 3 次元は「不定止」を表している．よって第 2 次元と第 3 次元のグラフは右上にあるものほど,「漢字」と「不定止」が多い作品と言える．第 4 次元は「人格語」を表している．よって第 3 次元と第 4 次元のグラフは右上にあるものほど,「不定止」と「人格語」が多い作品と言える．第 4 次元と第 1 次元のグラフは右上にあるものほど,「人格語」と「文の長さ」が多い作品と言える．

多次元尺度法の結果，文の長さが芥川賞作家は長く，直木賞作家が短いことが分かった．

## 8　芥川賞と直木賞の違い

読点前の文字の使用率からは，各賞の違いをはっきり見つけることが出来なかった．しかし，バイプロットから，非現代作家の方が現代作家より特徴が表れていることがわかる．また，現代作家は，編集者によって編集ががれているため書き方が全体的に似ているのではないかと考えられる．

文体の特性からは，非現代作家においては芥川賞作家は幅広い表現で書かれている傾向があり，人格語が少ない．直木賞作品は淡白な表現で書かれている傾向があり，人格語が多いことが分かった．現代作家においては，芥川賞作品は淡白な表現で書かれている傾向があり 1 文が長いことが分かった．直木賞作品は幅広い表現で書かれている傾向があり，1 文が短いことが分かった．

## 9　おわりに

今回の研究では，芥川賞や直木賞での作家の違いや共通点についてはっきりとした特徴を見つけることは出来なかった．しかし，それぞれの作者において，読点を付ける位置や文章の書き方についての特徴や傾向が多少はあることが分かった．また，非現代作家と現代作家では，多少は違いを見つけることができた．もっとデータを増やしいろいろな方法で分析を行うことで，良い結果を得られるのではないかと思う．

この研究を通して，計量文献学についての興味がより深まり，これから本や雑誌，新聞記事などの文章を読むときには，読点の付け方や文体について見ながら読むと面白いのではないかと思った．そして分析を行う際にも，方法を工夫をすることを学ぶことができた．たくさんのことを学ぶことができ，この研究テーマを選んで良かったと思う．

## 参考文献


[1] 藤本祐之：卒業論文『統計的方法による文体の分析』．南山大学経営学部情報管理学科，1995.

[2] 金明哲：『Rによるデータサイエンス』■データの解析の基礎から最新手法まで■．森北出版株式会社，東京，2007.

[3] 金明哲・中村永友：R で学ぶデータサイエンス 2『多次元データ解析法』共立出版，東京，2009.

[4] 村上正勝：【行動計量学シリーズ】6『真贋の科学　計量文献学入門ー』．朝倉書店，1994.

[5] 村上正勝：＜データの科学＞5『文化を計る　文化計量学序説ー』．朝倉書店，2002.