3G4-OS-05b-4

作家の構文の類似性

木カーネルを用いた構文間距離による文体の分析

Similarity between Auther's Writing Style using Tree Kernel

佐原諒亮[*1] 金川絵利子[*1] 岡留 剛[*1]
Ryosuke SAWARA Eriko KANAGAWA Takeshi OKADOME

[*1]関西学院大学大学院理工学研究科
Graduate School of Science and Engineering, Kwansei Gakuin University

Focusing on the dependency structure of sentences, this study compares authors' writing styles by introducing a distance between dependency trees defined using tree kernels. It discusses the similarities or differences among the dependency structure of sentences that consist of novels of 31 Japanese authors.

1. はじめに

作家の文や文章の特徴づけには，文の長さや句読点の間隔・用いる品詞などがよく用いられる．一方,「作家の文体」に焦点をあてた場合，文の句構造や係り受け構造が重要となる．本研究では，係り受け構造に着目して解析を行ない，作家間の文構造の類似度に着目する．文の句構造の類似度を測る指標として木カーネル [1] が存在する．本研究では，木カーネルを用いて構文間の距離を定義し，文体の分析を行なう．文章中の文の順序も重要な特徴となり得るが，文構造に焦点をあてるために，今回は文の順序は考慮せず bag of sentences に基づいて解析を行なう．

2. 木カーネル

木カーネルは，2 つの木構造データ間の共通している構造として，部分木を用いるカーネルであり，共通する部分木の個数を数えることで値が決定される．部分木の定義は以下である．

- 少なくとも 1 個以上の子を持つ任意のノードを選んで，このノード（部分木の根とする）と，子孫ノードの組み合わせで得られる木である．
- 部分木の根以外のあるノードが木に含まれる場合，その兄弟ノードもすべて木に含む．

木カーネルは，木構造 T_1，T_2 に対して，以下の式で定義される．

$$K(T_1, T_2) = \langle \boldsymbol{\phi}(T_1), \boldsymbol{\phi}(T_2) \rangle = \sum_{S \in \tau} \phi_S(T_1)\phi_S(T_2),$$

ここで，S は部分木である．τ はすべての固有木の集合で，また，$\phi_S(T)$ は，木 T が S を部分木として含むときは 1，含まないときは 0 となる．これにより，T_1 と T_2 の共通の部分木の数え上げを実現している．

連絡先: 氏名：佐原 諒亮
所属：関西学院大学大学院理工学研究科
住所：〒 669-1337 兵庫県三田市学園 2-1
メールアドレス：ryosukesawara@kwansei.ac.jp

3. 木カーネルを用いた 2 文間の距離

木カーネルを用いて，文 s_1 と s_2 の距離を表現する．2 つの文 s_1 と s_2 のそれぞれの構文木が T_1，T_2 のとき，文 s_1 と s_2 間の距離を以下のように定義する．

$$d(s_1, s_2) = \sqrt{K(T_1, T_1) + K(T_2, T_2) - 2K(T_1, T_2)},$$

ここで，$K(T_1, T_2)$ は構文木 T_1，T_2 の木カーネルの値を示す．カーネルはヒルベルト空間での内積として示されることから，d は距離であることを証明することができる．

4. 評価

文の構文を表現する方法として，句構造文法や係り受け構造によるものがある．本研究ではこの 2 つの構造に着目して解析を行ないたい．しかしながら，現在のところ，さまざまな作家の作品を構文解析するための強力で一般的な日本語句構造文法が存在しない．従って，本研究は係り受け構造のみに絞り，作家の文を解析する．

4.1 前処理

評価を行なうに当たり，用いる文書のクリーニングを行なう．すなわち，半角・全角スペースなどの空白文字を削除し，会話文は「」を削除した文を使用する．その他の記号に関しては原文通り使用した．

また，用いたテキストに対して係り受け解析と形態素解析を行なうため，係り受け解析器の CaboCha と形態素解析器の Mecab を用いた．2 つの文の木カーネル値は葉である単語に大きく依存する．本研究では，構文構造の類似度に焦点を当て，用いられている単語の違いは極力排除したいため，各単語を品詞と形態素情報を表す記号に還元的に縮約したコーパスを用いる．例えば,「僕は学校まで走る」という文の還元的縮約は,「n は n まで v」となる．ただし，この場合の n と v は，それぞれ名詞と動詞を表している．

4.2 実験

木カーネルを用いて実験を行なった．本研究の実験では，青空文庫から比較的作品数の多い 31 作家を選んだ．木カーネルを用い作家の比較をする際，文の数をて一致させる必要があるため，各作家の全作品からランダムに 100 文抽出し，木カーネル値の総当たり平均と，2 文間の距離の平均から，2 作家間の距離を求める．結果はこれを 10 回行なったものの平均を用

いている．また，木カーネルの計算は Moschitti のプログラム [3] を用いて計算した．パラメータ λ の値は，デフォルトの 0.4 とした．これは，他の値を用いて実験を行なった場合でも，作家間の関係性に大きな差が見られなかったためである．この際，Subset Trees (SSTs) を用いる方法と，Sub Trees （STs）を用いる方法とがあり，両者について解析を行なった．両者の結果においても大きな相違がなったため，今回は STs による解析について述べる．各作家ごとの木カーネル値の総当たり平均値と 2 作家間の距離を表にまとめた（表 1，表 2）．

表 1: 木カーネルを用いた SSTs（Subset Trees）の代表 5 作家間のカーネル値

	芥川	太宰	宮沢	夏目	新美
芥川	3.60	3.50	2.43	3.74	3.15
太宰	3.50	3.66	2.42	3.77	3.25
宮沢	2.43	2.42	1.72	2.59	2.18
夏目	3.74	3.77	2.59	4.08	3.40
新美	3.15	3.25	2.18	3.40	3.00

表 2: 木カーネルを用いた SSTs（Subset Trees）の代表 5 作家間の距離

	芥川	太宰	宮沢	夏目	新美
芥川	0.00	0.71	0.84	0.66	0.73
太宰	0.71	0.00	0.86	0.66	0.62
宮沢	0.84	0.86	0.00	0.93	0.73
夏目	0.66	0.66	0.93	0.00	0.71
新美	0.73	0.62	0.73	0.71	0.00

木カーネルによる値を用いた 2 作家間の文体的な特徴による距離をもとにバネモデル [4] を作成し，各作家間の関係を可視化した（図 1 ）．

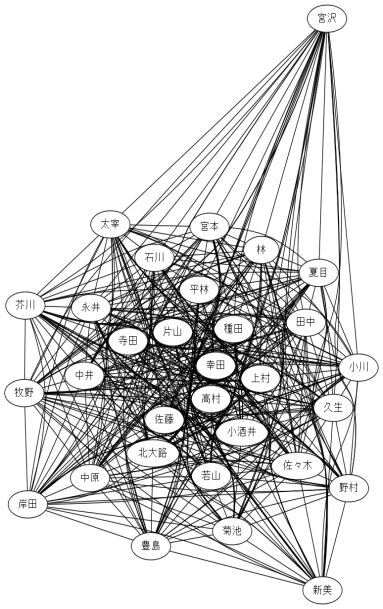

図 1: 木カーネルの値を用いた作家の文体間の平均「距離」によるバネモデル

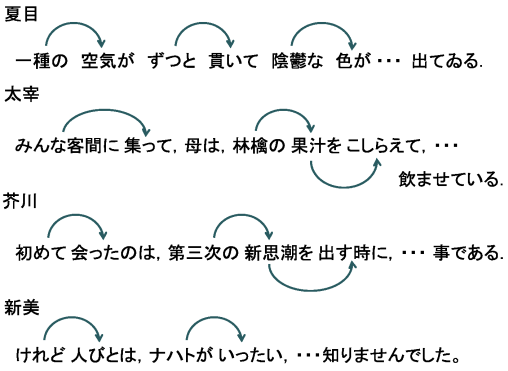

図 2: 各作家の 1 文の係り受け関係

5. 議論

表 1 からわかったことに加え，木カーネルを用いて行なった分析をもとに議論を行なう．今回は比較的文の長さの平均に差のない，芥川，太宰，夏目，新美に主に着目し，最後に宮沢についてと，ほかの特徴づけとの本研究による評価の比較を行なう．

5.1 夏目について

表 1 から，夏目が自分自身と比較した場合，カーネル値の総当たり平均値が一番大きくなっていることが見て取れる．従って，夏目は用いる係り受けの種類がほかの作家に比べて少なく，執筆する文の種類が比較的少なくなるのではないかと考えられる．実際に高いカーネル値を算出しやすい文について分析を行なった．今回は以下にその中の 3 文を例に挙げる．

- 私は田舎の客が嫌いだった．
- 私は子供の時から彼らの席に侍するのを心苦しく感じていた．
- それで私はただあまり仰山だからとばかり主張した．

この 3 文を見てみると，主語が明確であり，かつ文頭に近い部分に来ていることがわかる．夏目は主語の位置が比較的文頭に固定されており，これにより係り受けの種類が少なくなったためにカーネル値がほかの作家と比較したときに大きくなったと考えられる．

また，表 2 から，夏目は宮沢とは距離が比較的遠いものの，残りの 3 作家とは他の値と比べて小さな値となっており，どの作家とも似た文を書くのではないかという仮説を立てることができる．そこで，夏目，太宰，芥川，新美について，各作家間でカーネル値を算出したときに大きな値を取った文について分析してみる．図 2 はその中から各作家 1 文ずつ抽出し，その 1 文の係り受けを表した図である．図 2 を見てみると，各作家とも比較的簡単な係り受けによる句をつなげることで，1 文を生成しているということが見て取れる．すなわち，どの作家と似た文というよりは，どの作家も句単位で見てみると，1 つ 1 つの構造はそれほど複雑ではないため，共通する部分木が多くなり，結果として距離が小さくなったと言える．

5.2 太宰について

次に太宰について議論する．表 1 から，太宰は夏目よりも自分自身と比較した場合，カーネル値の総当たり平均値が少々小さくなっていることがわかる．つまり，夏目と比較すると，

より用いる係り受け構造の種類が多いのではないかと考えられる．夏目と同様，実際に高いカーネル値を算出しやすい文について分析を行なった．以下にその中の 3 文を例に挙げる．

- そのときから，私は，いままでの私でなくなりました．
- 父が死んだ事を知ってから，自分はいよいよ腑抜けたようになりました．
- あなた，とお呼びしていいのか，先生，とお呼びすべきか，私は，たいへん迷って居ります．

この 3 文を見てみると，夏目と同様に主語は明確に書かれているが，夏目と異なり，主語が文中に現れていることがわかる．従って，主語の場所がばらつくことで，夏目と比較して係り受けの種類が増え，執筆する文の種類もそれに伴い増えるため，カーネル値が小さくなったと考えられる．

また，表 2 を見てみると，他の作家と比較した場合，太宰は新美と距離が近いことがうかがえる．そこで，ここでは太宰と新美で木カーネル値を算出した場合，大きな値を取ったものを 2 ずつ以下に挙げる．

- みんな客間に集って，母は，林檎の果汁をこしらえて，五人の子供に飲ませている．(太宰)
- 語学の勉強と称して，和文対訳のドイルのものを買って来て，和文のところばかり読んでいる．(太宰)
- けれど，あまりどなりちらしたので，体がふるえるとみえて，二，三べん自転車に乗りそこね，それからうまくのって，行ってしまいました．(新美)
- つまり，良寛さんといふりつぱなお坊さんがじつさいにゐて，その人の書き残したものや，その人について書かれたものもいろいろあつて私はそれらのものを土台にして書けばよかつたのです．(新美)

これらの文を見てみると，「て」や「して」を多く用いていることがわかる．太宰と新美はこのように「て」·「して」を用いて句を並列にしているという傾向があるということが分析でわかった．

5.3 芥川について

芥川についても議論をしたい．表 1 を見てみると，芥川は夏目や太宰と比較して木カーネルの値が小さくなってることがわかる．つまり，夏目と太宰と比較すると，より用いる係り受け構造の種類が多いのではないかと考えられる．今回も実際に高いカーネル値を算出しやすい文について分析を行なった．以下はその中から 3 文抽出したものである．

- その日も電燈のともり出した時分，中村はあるカフェの隅に彼の友だちと話していた．
- 事によると彼等が読んだのも，僕の持つてゐる詩集のやうに，印刷の拙い本だつたかも知れない．
- 若しみづから甘んじて永久の眠りにはひることが出来れば，我々自身の為に幸福でないまでも平和であるには違ひない．

これらの文を見てみると，最後の文のように芥川は主語が存在しない場合が他作家よりも多く，さらに太宰と同じように主語が文中に現れることがわかる．このことから，太宰よりもさらに用いる係り受け構造が多くなり，その結果さまざまな文を著すためにカーネル値が小さくなったと考えられる．

また，表 2 を見てみると，芥川は夏目以外比較的距離が大きくなっていることが見て取れる．このことから，芥川はほかの作家があまり用いない構文をした文を書く可能性があると考えられる．そのため，今回も高いカーネル値を算出しやすい文について分析を行ない，その中から 3 文を抽出した結果を以下に示す．

- 孔雀はまるで扇のやうに，虹色の尾羽根を開いて見せた．
- 事によると彼等が読んだのも，僕の持つてゐる詩集のやうに，印刷の拙い本だつたかも知れない．
- 夏のやうに白鷺が空をかすめて飛ばないのは物足りないけれども，それだけのつぐなひは十分あるやうな気がする．

この 3 文を見てみると，「やうに（ように）」や「やうな（ような）」を用いていることが見受けられる．芥川はこのように「やうに（ように）」や「やうな（ような）」といった比喩表現をほかの作家よりも多く用いる傾向があることか分析よりわかった．

5.4 新美について

また，新美についても議論をする．表 1 から，新美は自分自身との木カーネル値の総あたり平均を求めた時，前記した 3 作家と比べると少々小さめになっていることがわかる．従って，新美はよりたくさんの係り受け構造を用いて文を著すのではないかと考えられる．新美についての特徴的な文について，今まで同様木カーネル値が大きかったものの中から 3 文を例に挙げることにする．

- 物干台のようなわくのついた車をしたてて，それにランプやほやなどをいっぱい吊し，ガラスの触れあう涼しい音をさせながら，巳之助は自分の村や附近の村々へ売りにいった．
- 見ると，舞台の正面のひさしのすぐ下に，大きな，あか土色の蛾がぴったりはりついていました．
- そして，今か今かと，ペテロや，ほかのわかものが，がいせんしてくるのをまっていました．

これらの文を見てみると，1 文目のように読点の間隔が比較的長い文もあれば，2 文目の文のように読点の間隔に差のある文や 3 文目のように読点の間隔が狭い文など，さまざまな文が見受けられる．新美はさまざまな係り受けの距離の句を多用することで，係り受け構造の種類を多くしているため，カーネル値が小さくなり易かったのではないかと考えられる．

表 2 に着目した場合，太宰との距離が近いことは議論したが，太宰は芥川と比較的近い距離となっているにもかかわらず，新美は芥川とは太宰に比べて距離が大きくなっていることが確認できる．つまり，新美と芥川は係り受け構造に着目した場合，共通する係り受け構造を持つ可能性が他作家よりも低いと言える．これはやはり，読点の間隔，つまり句の長さが関係してると考えられる．新美と太宰は読点の間隔の平均に大きな差がなく，その分散も比較的似ている．一方で芥川の比較すると，芥川は新美や太宰と比べた時，読点の間隔が大きい．その点で芥川と比較すると 1 つの句の中で共通する係り受け関係が少なくなってしまうため，距離が大きくなってしまうと考えられる．

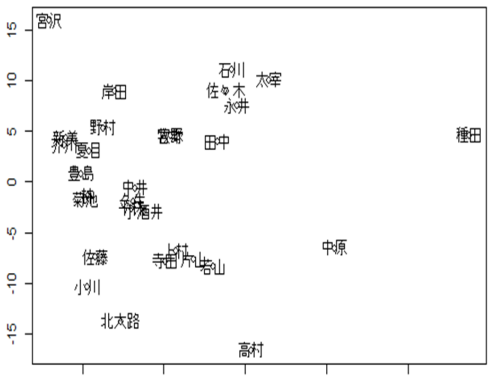

図 3: 非計量多次元尺度構成法による各作家のプロット

5.5 宮沢について

宮沢についても少しだけ触れておく．表 1 からわかる通り，宮沢はほかの作家と比べて自身のカーネル値と他作家とのカーネル値が小さくなっている．これは，宮沢は他作家と比べた際に，比較的短い文を書くために，1 文に含まれる係り受け構造が少なくなってしまったためだと考えられる．宮沢の 1 文と他作家の 1 文を木カーネルで比較したとき，宮沢の 1 文には係り受け構造が少ないため，カーネル値は必然的に小さくなってしまう．従って，文の長さに影響を受けない工夫を施す必要が今後必要である．

5.6 他の特徴づけとの比較

最後に文の長さや読点の間隔に着目した分析と，今回の木カーネルを用いて係り受け構造に着目した分析を比較してみる．図 3 は非計量多次元尺度構成法を用いて，文の長さの平均や読点の間隔など，5 つの要素を使用したときに，それを 2 次元上にプロットした図である．この図を見てみると，芥川と夏目は近いところにプロットされており，太宰は 2 人から離れたところにプロットされていることがわかる．文の長さの平均や，読点の間隔に着目した場合，芥川と夏目は似ているが，この 2 人と太宰はあまり似ていないのではないかという結果が得られたことになる．しかしながら今回の木カーネルを用いた係り受け構造に着目した分析では，表 1 からもわかる通り，この 3 作家は比較的似ているのではないかという結果が得られている．従って，文の長さの平均などの表層的な特徴づけに加え，係り受け構造に着目した特徴づけを行なうことで，より作家の特徴づけを明確にできると考えられる．

6. 関連研究

表層的な文字や記号・品詞の並びに基づく作家の特徴づけに関する研究は多くある．例えば，文の長さについて注目している文献として前川 [5] のものがある．この文献では，各作品の文の長さの平均と，文の長さのばらつきを用いて比較を行なった．単語の長さに関して分析を行なったのが金の研究 [6] である．この文献では，単語の長さの分布には著者の特徴が現れるのか，どうすれば著者の特徴がより明確に得られるかについて分析を行なっている．

読点の打ち方について分析を行なった研究が金らの研究 [7][8] である．ここでは，読点の前の文字や読点の前の文字の品詞，読点を打つ間隔に関する情報の有効性を分析した．いずれにおいても各特徴において作家ごとに特徴がみられると述べている．

また，係り受けの距離に関して行なった研究には金の [9] がある．ここでは 3 作家の係り受け距離の分布では，作家の個性が明確には現れないことや，係り受けの距離ごとに読点の打つ頻度には書き手の個性が見られないと述べている．

英語文書においても同様のような研究が行なわれている．例えば，Yule[10] は，作家の文書の文の長さを分析し，その平均値，中央値，四分位範囲が作家ごとに異なるということを述べた．この結果を受けて Yule は,「The Imitation of Chirst」の著者推定を行ない，文の長さの有用性を示した．

また，テキストデータに対するカーネルとしてベクトル空間カーネルが定義されている．これは単語の出現頻度に基づいたカーネルであり，文の句構造の類似性を反映させる際にはほかの手法を考える必要がある．

7. まとめ

本研究では作家間の構文構造の類似度を測るため，木カーネルを用いて構文距離を定義し，それを用いて評価実験を行なった．実験では係り受け解析を行ない，1 コーパス 100 文と制約を設けたコーパスを用いて実験を試みた．その結果，構文的に似ている作家や，似ていない作家を捉えることができた．文の長さによって結果が左右されやすいが，構文的に似ている作家であればカーネルの値が大きくなりやすい，また距離は小さくなりやすいことも確認できた．

参考文献

[Collins 01] Collins, M. and N. Duffy (2001). Convolution kernels for natural language. *Advances in Neural Information Processing Systems 14* [*Neural Information Processing Systems: Natural and Synthetic, NIPS 2001*] , 625-632, MIT Press.

[Haussler 99] Haussler, D. (1999). Convolution Kernels on Discrete Structures. Technical Report, University of Santa Cruz.

[Moschitti] Moschitti, A. TREE KERNELS IN SVM-LIGHT. http://dit.unitn.it/moschitt/.

[Kamada 89] Kamada, T. and S. Kawai (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**, 1, 7-15.

[前川 95] 前川守 (1995). **文章を科学する**，岩波書店.

[金 96] 金明哲 (1996). 日本語における単語の長さと分布と文章の著者， **社会情報**， **5**， 2， 13-21.

[金 94] 金明哲 (1994). 読点の打ち方と著者の文体特徴， **計量国語学**， **19**， 7， 317-330.

[金 93] 金明哲， 樺島忠夫， 村上征勝 (1993). 読点と書き手の個性， **計量国語学**， **18**， 8， 382-391.

[金 96] 金明哲 (1996). 文節の係り受け距離の統計分析，**社会情報**，**5**，2，1-11.

[Yule 39] Yule, G. U. (1939). On sentence-length as a atatistical characterristic of style in prose; with application to two cases of disputed authorship, *Biometrika*, **30**, 3, 363-390.