2K1-1in

# 作家の文体の類似性

## 情報量木カーネルの導入による構文間距離を用いた分析

## Similarity between Auther's Writing Style using Information Tree Kernel

金川絵利子 *1　佐原諒亮 *1　岡留 剛 *1
Eriko KANAGAWA　Ryosuke SAWARA　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

The information tree kernels proposed here give measures of the syntactic similarity of sentences. For two trees, they are defined as the total amount of information contained in their common subtrees, where the amount of information of a common subtree is calculated using the occurrence probability for the subtree. The information tree kernels and the distances defined by the kernels enable us to capture the syntactic similarities and differences in Japanese famous 34 authors' writing styles.

## 1. はじめに

作品に基づく作家の分類や特徴づけは古くから興味を持たれさまざまな研究が行なわれてきた．それらにおいては，作品を特徴づける量として，1) 文書中に含まれる文の長さや読点の数の平均値・単語の出現頻度といった文の表層的な統計量や，2) 読点の直前の格助詞の出現頻度や特定の文節の出現頻度といった構文情報のある側面を表現している量が利用されてきた．(例えば，[前川 95], [金 94], [金 02]). 一方，よく言われる「作家の文体」いう言い回しにおける「文体」という表現は，作品の意味内容や，さらには書かれている媒体さえも含んでいるという主張さえある [山本 14]. しかし，作品を構成する文の表層的統計量と作品の意味内容とのちょうど中間に位置づけられる文の構文構造そのものの違いについてはほとんど議論されてこなかった．それは主に構文構造の違いを数値化する困難さに起因していたと思われる．本研究では，作家の文体を特徴づける重要な要因として作品を構成する文の構文構造に焦点をあてる．

非数値的構造データの類似度を測る尺度としてさまざまなカーネルが提案されてきた．その 1 つに木構造を入力とする木カーネルがあり，言語解析でそれが用いられている [Collins 01]. 例えば，[Mocshitti 06] は，ラベルづけられた文に対し，木カーネルを用いて SVM により文書分類を行なっている．文書中の各文に対し木カーネルを用いて文書分類を行なった場合，単語の違いを無視すると文の句構造が同じであれば，出現頻度が低い部分木と高い部分木で同じ類似度となる．出現頻度が低い同じ構造が 2 つの文書間で出てくればそれらの類似度は高いと考えられ，部分木の出現頻度を反映されるカーネルの構築が課題となる．

本研究は，部分木の出現頻度を反映するカーネルを定義する．そのため，木カーネルに部分木の出現確率を組み込み，木カーネルの概念を拡張するアプローチをとる．なお，本稿では，構文木という用語は，句構造を表現する木や，係り受け関係を表現する木など広い意味での構造木を指す．

連絡先: 氏名：金川 絵利子
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：eriko.k@kwansei.ac.jp

## 2. 関連研究

[Goncalvel 08] は，ポルトガル語で書かれた文書を木カーネルを用いて分類し，構文構造は分類に適さないという結果を得ている．文を木構造に展開したとき構文木の葉は単語となり，その木をカーネルの入力として用いた場合，構文木の骨格の違いよりも単語の違いが強調される結果となる．彼らの分析では，単語を含む構文木を用いており，そのため木構造そのものの違いが反映されていない可能性が高い．

[金 02] は，助詞の n-gram パターンを用いた書き手の識別法を提案しており，一般人の書いた短い文章に対しても高い判別率で書き手の判別が可能であるという結果を得ている．

[太田 09] は，Harmonic Grammar に基づき書き手の識別法を提案している．書き手の文章生成モデルとして Harmonic Gramamar を仮定しており，助詞の出現パターン，読点の打ち方，品詞の出現パターンに着目し，「制約」を作成し，それに対する「重み」計算する．各書き手は，それぞれ異なる制約の重みを持つと仮定し，これを書き手の特徴としている．また，比較的高い正解率で書き手の認識を行なうことに成功している．

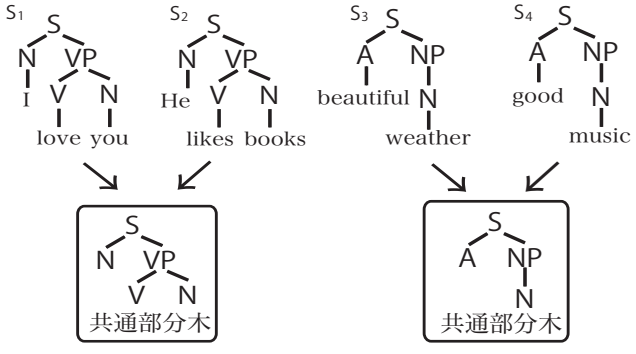
## 3. 情報量木カーネルと文間距離

木カーネルは，木構造データに対して定義されたカーネルである．与えられた 2 つの文の構文木に共通する部分木の個数を数え上げカーネル値する．一般に，カーネルは，引数である 2 つの対象間の類似度を表す一つの指標である．木カーネルも 2 つの木のある類似度を表現するが，木に含まれるすべての部分木を対等に扱うため，共通部分木の数という意味での類似度になっている．

ここで，一般的にはあまり用いられることがない独特の構文を共通に持つ言い回しをしばしば使う 2 人の作家を考えよう．この共通の構文をこの 2 人以外の作家はあまり用いないということは，この構文は，2 人の作家の文体を特徴付ける一つの重要な要因であるといえる．しかし，木カーネルを直接構文の類似度として用いたのでは，このような特徴を浮かび上がらせることはできない．具体例で示そう．図 1 の Tiny Einglish における文 $s_1$ と $s_2$ のカーネル値と，文 $s_3$ と $s_4$ のそれは等しい．一方，$s_1$ と $s_2$ の構文に含まれる部分木の生成確率は，$s_3$ と $s_4$ のそれに比べると大きく，$s_1$ と $s_2$ の構文はごく普通に

文書中に現れるが，$s_3$ と $s_4$ の構文は比較的まれに使われる構文と言える．

*Tiny English*
$s_1$：I love you.　$s_2$：he likes books.
$s_3$：beautiful weather.　$s_4$：good music.
句構造規則：生成確率
S → N VP：0.8　S → A NP：0.2
VP → V N：1.0　NP → N：1.0
A → beautiful：0.5　A → good：0.5
N → I：0.2　N → you：0.2
N → He：0.2　N → books：0.2
N → weather：0.1　N → music：0.1
V → love：0.5　V → likes：0.5




(1)　(2)

図 1: Tiny Einglish の文法と生成文．(1) Tiny Einglish における文 $s_1$ と $s_2$ のそれぞれの構文木と共通部分木．(2) Tiny Einglish における文 $s_3$ と $s_4$ のそれぞれの構文木と共通部分木．

この Tiny English では，$s_1$ と $s_2$ の木カーネル値と $s_3$ と $s_4$ のそれは両者とも 3 であるのに対し，$s_1$ と $s_2$ の共通部分木の生成確率は 0.8 であり，$s_3$ と $s_4$ のそれは 0.2 で前者と大きく異なる．

そこで構文中の部分木の生成確率を考慮したカーネル，すなわち，情報量木カーネルを提案する．以下では，各エッジにその生成確率が付与された構文木（生成確率付き構文木）であり，1 つの木におけるそれぞれの部分木の生成は独立であると仮定する．2 つの文 1 と文 2 のそれぞれの構文木を $T_1, T_2$ とし，$N_1$を $T_1$のノードの集合, $N_2$を $T_2$のノードの集合 とする．$T_1$ と $T_2$ が与えられたとき，$T_1$ と $T_2$ の**情報量木カーネル**を以下のように定義する．

$$K_I(T_1,T_2) = \sum_i \lambda^{\mathrm{size}(i)} h_i(T_1) h_i(T_2)(-\log p_i)$$
$$= \sum_{n_1 \in N_1} \sum_{n_2 \in N_2} \sum_i \lambda^{\mathrm{size}(i)} I_i(n_1) I_i(n_2)(-\log p_i)$$
$$= \sum_{n_1 \in N_1} \sum_{n_2 \in N_2} \Phi(n_1, n_2).$$

ただし，

1. ($n_1$をルートとする部分木) ≠ ($n_2$をルートとする部分木) のとき
$$\Phi(n_1, n_2) = 0,$$
2. ($n_1$をルートとする部分木) ≠ ($n_2$をルートとする部分木) で, かつ 終端記号の前であれば
$$\Phi(n_1, n_2) = \lambda(-\log p_i),$$
3. それ以外のとき (Subset Trees の場合)
$$\Phi(n_1, n_2) = \lambda(2^{nc(n_1)}(-\log p_i) + \sum_{j=1}^{nc(n_1)} 2^{nc(n_1)-1} \Phi(ch(n_1, j), ch(n_2, j))),$$

ここで，$h_i(T)$ は，すべての木に 1 から番号をつけたとして，$i$ 番目の部分木が木 $T$ に出現する回数である．$p_i$ は $i$ 番目の部分木の生成確率であり，生成確率の低い部分木に対して大きいカーネル値を与えるために驚き度合いである情報量を用いている．$I_i(n)$ はノード $n$ をもつ部分木中に $i$ 番目の部分木が，存在するとき 1，それ以外 0 となる指示関数である．$nc(n)$ はノード $n$ が持つ子ノード数を表し，$ch(n, j)$ はノード $n$ を持つ部分木の $j$ 番目の子ノードを示す．また，$size(i)$ は $i$ 番目の部分木を生成するために適用した生成規則数で，$\lambda$ は $0 < \lambda \leq 1$ を満たし，木の大きさに対する依存度を低くする効果を持つパラメータである．

$T_1$ と $T_2$ の情報量木カーネル値は共通する部分木の情報量の和となり，生成確率が低い共通の部分木が 2 つの木で出現するほど値は大きくなる．先に挙げた Tiny English における $s_1$ と $s_2$ の情報量木カーネル値は 0.644 bit であるのに対し，$s_3$ と $s_4$ のそれは 4.644 bit と約 7 倍となる．

情報量木カーネル値を用いて文 $s_1$ と $s_2$ の距離を定義しよう．文 $s$ の構文木を $T(s)$ とする．このとき文 $s_1$ と $s_2$ 間の構文距離を以下で定義する．

$$d(T(s_1), T(s_2)) = \sqrt{K_I(T(s_1), T(s_1)) + K_I(T(s_2), T(s_2)) - 2K_I(T(s_1), T(s_2))}$$

ここで，$K_I(T_1, T_2)$ は構文木 $T_1$ と $T_2$ の情報量木カーネル値である．

## 4. 評価

文の構文を表現する方法には，句構造文法によるものや，係り受け解析によるもの・意味論的構造も考慮した LFG や HPSG など様々ある．本研究では，純粋に構文構造の違いに焦点を当てるため，句構造と係り受け構造とで文の構造を表現する．しかし，現在のところ，さまざまな作家の作品を構文解析できるだけ十分に強力で一般的な日本語句構造文法は存在しない．そのため，本研究では，係り受け構造に着目して作家の文を分析する．

### 4.1 前処理

まず，分析対象とする各文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は削除し，また,「」内の会話文は「」を削除し会話文の本文のみを使用する．その他の記号に対しては原文通り使用した．

1 文ずつ Cabocha[工藤 02] を用いて形態素解析と係り受け解析を行なった．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，骨格としての構文構造の類似性に注目するため，用いられる単語の違いによるカーネル値への影響は極力排除したい．そのため，Cabocha の形態素解析をもとに，[長谷川 94] に基づいて，単語を，品詞と形態素情報を表す記号に還元的に置換した．例えば，(私は 音楽を 聴きながら， 大好きな 本を 読んだ．) の還元的縮約は，(n は n を vccr j n を v) となる．ただし，n，vccr，j，v はそれぞれ名詞,「ながら」が語尾に付いた動詞，形容詞，動詞を表す非終端記号である．

### 4.2 出現確率

生成確率 (出現確率) として本研究では，1) 出現回数に基づく相対頻度と，2) Cabocha のスコアによる「相対頻度」を用いた．毎日新聞 3 年間分 (2010 年から 2012 年) と，NHK の NEWS WEB60 日分 (2014 年 7 月 20 日から 2014 年 7 月 27 日と，2014 年 9 月 12 日から 2014 年 11 月 3 日)，さらに青空文庫の中から比較的作品数の多い 34 作家の 5,909 作品から成るコーパス (文数 5,511,696，文節数 42,024,675，句読点を除く単語数 109,929,329，単語の種類 327,487) から係り受け

の出現確率を計算した．すなわちまず，コーパス中のすべての文に対して，Cabocha で形態素解析を行ないさらに還元的縮約を行なう．Cabocha の係り受け解析の結果から，係り元の品詞と係り先の品詞などによるすべての種類の係り受けを列挙し，そのおのおのの出現回数と Cabocha の総スコアを求める．

ある文節の係り先の文節の種類も重要であるが，係り受けの文節間距離も重要な構文情報である．文節間の距離は，隣り合う文節同士で 1 とし，間に $k$ 個の文節が存在する場合を $k+1$ とした．しかし，すべての係り受けの種類と文節間距離を考慮し区別すると，係り受けの種類が多く，ほとんどのもの相対頻度が 0 に近くなる．そのため係り受けの種類ごとに，文節間の距離のグルーピングが必要となる．そのため，距離が小さい係り受けはそのまま独立に，距離が大きくなるほどまとめてグループ化されるようにグループ化する．本研究では，Fibonacci 数列を用いてグループを構成した．すなわち，Fibonacci 数列の各数が 1 つのグループの構成員数となるように，距離 1 のものから順にグループ化する．全係り受けの出現総数と，ある文節間距離を考慮した係り受けの出現回数の比として，相対頻度を計算し，その係り受けに対する出現確率とする．すなわち，文節 a が文節間距離 $r$ の文節 b に係る係り受け $\langle a, b, r\rangle$ の出現確率は以下のように計算する．

$$\langle a, b, r\rangle = \frac{N(\langle a, b, r\rangle)}{\sum_{r'=1,2,\cdots} \sum_{f \in F} \sum_{e \in E_f} N(\langle f, e, r'\rangle)}$$

ここで，$N(\langle a, b, r\rangle)$ は係り受け $\langle a, b, r\rangle$ の出現回数を示す．また，$F$ は全係り受けの係り元の集合，$E_f$ は $f$ を係り元に持つ係り受けの係り先の集合を示す．Cabocha のスコアに基づく「相対頻度」からの出現確率も同様に計算する．

### 4.3 情報量木カーネル値

各文に対して，還元的縮約を行なった確率付き構文木から情報量木カーネルを計算する．情報量木カーネルの実装は，[Mocshitti 06] の木カーネルプログラムを拡張する形で行なった．作成したプログラムの計算量は，2 つの木のノード数の積に比例する．

### 4.4 実験

青空文庫の作家の中から比較的作品数の多い 34 作家で実験を行なった．各作家の全作品からランダムに 100 文抽出し，情報量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．これを 10 回行なったものの平均を結果とした．パラメータ $\lambda$ の値は，木カーネル値を求める [Mocshitti 06] のデフォルトの 0.4 とした．情報量木カーネルの総当たり平均と 2 作家間の距離を，出現相対頻度に基づく情報量を用いる方法と，Cabocha のスコアから求めた情報量を用いる方法のそれぞれに対して，Subset Trees (SSTs) Kernel と SubTrees (STs) Kernel の二種類の実験を行なった．今回はスペースの関係上，著名な 5 作家の芥川龍之介と太宰治・夏目漱石・新美南吉・宮沢賢治について議論を行なう．各作家ごとの SSTs での情報量木カーネルの総当たり平均値と SSTs での作家間の距離を表にまとめた（表 1・表 2）．また，同様の実験を木カーネルでも行ない，各作家ごとの SSTs での木カーネルの総当たり平均値を表 3 にまとめた．

表 1: 出現相対頻度に基づく情報量を用いた SSTs(Subset Trees) の代表 5 作家の情報量木カーネル値の総当たり平均．

| | 芥川龍之介 | 太宰治 | 宮沢賢治 | 夏目漱石 | 新美南吉 |
|---|---|---|---|---|---|
| 芥川龍之介 | 8.793 | 0.166 | 0.124 | 0.196 | 0.136 |
| 太宰治 | 0.166 | 75.014 | 0.097 | 0.144 | 0.101 |
| 宮沢賢治 | 0.124 | 0.097 | 46.640 | 0.107 | 0.083 |
| 夏目漱石 | 0.196 | 0.144 | 0.107 | 5.840 | 0.121 |
| 新美南吉 | 0.136 | 0.101 | 0.083 | 0.121 | 3.089 |

表 2: 出現相対頻度に基づく情報量を用いた SSTs(Subset Trees) の代表 5 作家の距離．

| | 芥川龍之介 | 太宰治 | 宮沢賢治 | 夏目漱石 | 新美南吉 |
|---|---|---|---|---|---|
| 芥川龍之介 | 0.00 | 7.66 | 5.45 | 3.75 | 3.37 |
| 太宰治 | 7.66 | 0.00 | 8.32 | 7.37 | 7.15 |
| 宮沢賢治 | 5.45 | 8.32 | 0.00 | 5.11 | 4.74 |
| 夏目漱石 | 3.75 | 7.37 | 5.11 | 0.00 | 2.93 |
| 新美南吉 | 3.37 | 7.15 | 4.74 | 2.93 | 0.00 |

表 3: 代表 5 作家の SSTs(Subset Trees) の木カーネル値．

| | 芥川龍之介 | 太宰治 | 宮沢賢治 | 夏目漱石 | 新美南吉 |
|---|---|---|---|---|---|
| 芥川龍之介 | 4.12 | 3.84 | 2.68 | 4.15 | 3.47 |
| 太宰治 | 3.84 | 4.06 | 2.63 | 4.12 | 3.52 |
| 宮沢賢治 | 2.68 | 2.63 | 1.94 | 2.84 | 2.38 |
| 夏目漱石 | 4.14 | 4.12 | 2.84 | 4.61 | 3.73 |
| 新美南吉 | 3.47 | 3.52 | 2.38 | 3.73 | 3.36 |

## 5. 議論

スペースの関係上構文木を S 式で表現する．例として “私のかばん” という文の係り受け構造木を考える．“私の” は文節間距離 1 の “見た” に係るため，(かばん (私の, 1)) と S 式で表現できる．

表 1 から分かるように，他の作家との類似度を比較した場合，芥川と夏目の情報量木カーネル値が大きい．それゆえ，この二人は一般的に珍しい係り受けを用いた文を多く書くことで，構文的に似ていると推測できる．珍しい係り受けには，その作家らしさ，作家の文体の特徴が含まれていると考えられる．宮沢と新美の情報量木カーネル値が小さいことから，この二人は構文的に似ていない文を書くといえよう．

次に，表 1 での芥川と太宰に着目する．芥川と太宰は，今回使用した青空文庫中の作品に対して，1 文あたりの平均文節数や 1 文あたりの平均単語数の値が類似している．しかし，芥川自身との情報量木カーネル値と，太宰自身のそれは大きく異なる．1 文の長さが長ければ，文に含まれる部分木の個数も一般的には増加するため，文の長さに情報量木カーネルが依存しているのであれば，情報量木カーネル値は大きくなる．しかし，文の長さがほぼ等しい芥川の情報量木カーネル値と，太宰の情報量木カーネル値に差があることから，情報量木カーネルは，ノード数 (文の長さ) への依存が少なく，構文類似度ををとらえている考えられる．

また，表 1 と表 3 の情報量木カーネル値と木カーネル値の差に着目する．情報量木カーネル値は 1 文あたりの共通する部分木の情報量の平均値であり，木カーネル値は共通部分木数を意味する．情報量木カーネルを用いた場合どの作家に対しても，情報量木カーネル値が大きいのは芥川だが，木カーネルを用いた場合では夏目である．これより，共通部分木数は他の作家に対して夏目の方が多いが，情報量を考慮すると芥川の方が値が大きいことが分かる．この推測を裏づける具体例として，芥川の「従つて僕の中の光秀は必ずしも僕の中の紹巴を嘲笑しない」という文と，夏目の「敬太郎はその男と顔を見合せた

時、彼の最後の視線が、自分の足の下に落ちたのを注意した」という文がある．それぞれ太宰の「私は全集の日記の巻を調べてみた」と,「おのれの作品のよしあしをひとにたずねることに就いて自分の作品のよしあしは自分が最もよく知っている」を還元的縮約し構文木で比較すると図 2 のようになる．夏目と太宰の場合，一つの構文木の中に共通部分木である (n の, 1 (n の, 1)) が 2 つあるため，共通部分木数は多くなる．芥川と太宰の場合，共通部分木は (n を, 1 (n の, 1 (n の, 1))) であり，その生成確率は (n を, 1 (n の, 1)) と (n の, 1 (n の, 1)) の積となるためかなり小さくなり情報量木カーネル値は大きくなる．これより，夏目は一文の中に,「私のコップと彼のコップ」のような同じ構造を 2 回使う対照的な文を他の作家に比べ多く書くという特徴があるといえる．また，芥川は「私の友達の麻衣子を探す」のような「麻衣子を探す」だけではなく,「私の」や「友達の」といった細かい描写が他の作家と比べて比較的多いという特徴があると考えられる．

次に，作家間の距離 (表 2) に着目する．どの作家に対しても距離が小さいのは新美である．表 1 から新美は自分自身との情報量木カーネル値が最も低いため，他の作家との距離も近くなると考える．自分自身との情報量木カーネル値が小さいということは，頻出する出現確率の高い係り受け構造を多く使用すると考えられる．今回使用した青空文庫中の作品に対して，1 文あたりの平均文節数や文字数が新美とよく似ている宮沢との比較を考える．新美と宮沢は児童作家であり，芥川や太宰・夏目より比較的短い文を書くとい共通点がある．青空文庫中の新美の全作品で使用している係り受け構造の種類は 19,124 種類であり，宮沢は 36,428 種類である．新美の方が使用する係り受けの種類数が少ない．また，新美と宮沢の係り受け構造の使用回数上位 20 を比較すると，新美はすべて文節間距離 1 の係り受けであるのに対し，宮沢は文節間距離が 4 や 7 のものが存在する．よって，新美は文を構成するのに重要な頻出する，文節間距離の小さい係り受け構造を多く使用し，特徴的な係り受け構造は少ないため他の作家との距離が小さくなったと考えられる．それとは対照的に宮沢は，短い文のなかに様々な構造の係り受けを用いる文を書くと考えられる．

なお，34 作家の類似度をバネモデル [Kamada 89] を用いて可視化した (図 3)．その際,「距離」は情報量木カーネルを用いた作家間の距離を用いた．距離が 100 以上のものは 100 とした．

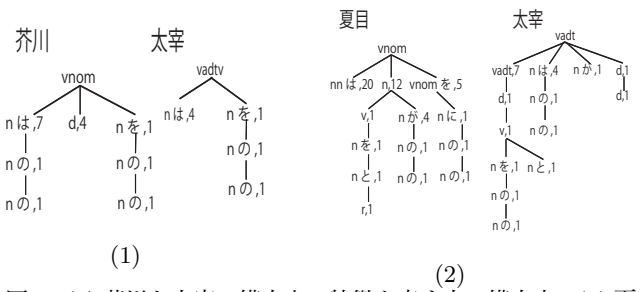



図 2: (1) 芥川と太宰の構文上の特徴を表す文の構文木．(2) 夏目と太宰の構文上の特徴を表す文の構文木．

## 6. おわりに

本研究は，構文情報により作家の文体の類似性を測るため，部分木の出現頻度を反映する情報量木カーネルを定義し，それを用いた 2 文間の構文距離を定義した．また，日本を代表する作家の文を用いた評価実験を行なった．前処理では文書のクリーニングと還元的縮約を行ない，係り受けの出現確率を計算し，情報量木カーネル値を求めた．情報量木カーネル値を用いて 2 作家間の距離を求め，作家の構文的違いをとらえるとこができることを確認した．

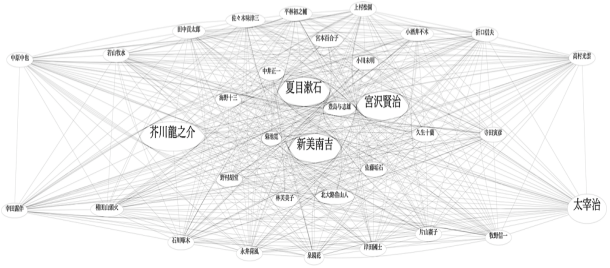



図 3: 出現相対頻度に基づく情報量を用いた SST(subset tree) の情報量木カーネル値を用いた代表 34 作家の距離を示すバネモデルでの表現．

## 参考文献

[前川 95] 前川守 (1995). **文章を科学する**, 岩波書店.

[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 02] 金明哲 (2002). 助詞の n-gram モデルに基づいた書き手の識別, **計量国語学, 23**, 5, 225-239.

[山本 14] 山本貴光 (2014). **文体の科学**. 新潮社.

[Collins 01] Collins, M. and N. Duffy (2001). Convolution kernels for natural language. *In Advances in Neural Information Processing Systems*. 625-632.

[Moschitti 06] Moschitti, M. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. *Proceedings of the 17th European Conference on Machine Learning* (*ECML*2006), 318-329.

[Goncalvel 08] Goncalvel, T. and P. Quaresma (2008). Text classification using tree kernels and linguistic information. *Proceedings of the Seventh International Conference on Machine Learning and Applications* (*ICMLA*'08), 763-768.

[太田 09] 太田貴久, 増山茂 (2009). 青空文庫を対象とした書き手の識別とその応用, **言語処理学会第 15 年次大会発表論文集**, 679-680.

[工藤 02] 工藤拓, 松本裕治 (2002). チャンキングの段階適用による日本語係り受け解析, **情報処理学会論文誌**, **43**, 6, 1834-1842.

[長谷川 94] 長谷川守寿 (1994). 日本語の句構造規則, **筑波応用言語学研究**, **1**, 55-71.

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