Vol.2017-ICS-186 No.7
2017/3/4

係り受け関係の類似性に着目した小説の著者推定

小泉 知夏[1,a)]　菅原 俊治[1,b)]

概要：近年インターネットの普及に伴い，書き手が特定できない文章，いわゆる匿名文章が増加している．そのため匿名性に起因する問題も増えている． 一方で，文書の殆どが電子テキストとなり計算機によって機械的によって処理・分析することができるようになった． 近年の研究では，計算機での分析により文章の書き手をある程度推測することができることを示す研究が多く存在する． 文章自体の分析よる書き手の特定は問題解決のためのひとつの手がかりとなり，発生自体の抑止力ともなりうる．本研究では小説を対象とし，文の構造に現れる特徴から文章の書き手を推定する実験を行った．具体的には，文の構造の最小単位である係り受け関係に注目し，文章がどれだけ似ているかという観点から分析を行った． 結果，係り受け関係が書き手の特徴を表し，書き手を推定する手法となりうることを示した． また，上記の方法が小説というジャンルの文章の著者を推定するのに効果的であることを示した．

Estimation of Novel Authors Using Modification Relations

Koizumi Chinatsu[1,a)]　Sugawara Toshiharu[1,b)]

1. 序論

コンピュータが普及するにつれ，私たちが目にする文書の多くが電子テキストとなった．また，文字認識技術の発達により，これまで使用されていた紙媒体文書の電子化も実用可能となってきた．そして，インターネットが一般的なものとなった昨今，不特定多数の人がインターネット上に文章を投稿している．その中でも SNS や電子掲示板へ投稿される文章など，匿名のものは多い．このメリットとしては，気軽な投稿や意見交換ができることである．反面，投稿者が特定されないが故の問題も増えている．個人や団体への誹謗中傷，なりすましやのっとり，ネット上への犯罪予告などである．

以上のような問題の対策として，これまでの自然言語処理技術を活用して，文章の特徴抽出や類似性の判定の研究・システム開発が進められている．例えば，[1] では，実際に書かれた犯行声明文を対象として，複数の候補者の書いた文章との比較を行い，書き手の特定を試みている．また，多くの大学では学生のレポートにおける剽窃行為に対して，レポート同士の類似性を求めるシステムを導入している．

こうした不正や，書き手特定の精度向上のため，自然言語処理の分野では，文章を書き手によって分類する研究や，文章そのものの特徴抽出に関する研究が注目されている．このような研究の精度が向上すれば，恣意的な中傷やデマの拡散のリアルタイムな抑制，なりすましの発見などが期待できる．

2. 関連研究

多くの研究では，ある特徴量を選択したとき，書き手ごとに文章を分類したり，文章が似ているかどうかを定量化することができる，ということが示されている．例えば [2] では，隣り合う文字の分布を特徴量として近代小説家 8 人の文章の著者推定を行っている．この実験では，3 万字のテキストを用いたときに，94.7%の平均正解率を出していて，著者候補が 8 人である場合，かなりの精度で著者が推定できることを示している．また，[3] では，読点の分布と読点前の文字に着目して実験をおこなっており，3 人の現代作家の作品を著者ごとに分類することに成功している．

[1] 早稲田大学
Waseda Uniersity
[a)] c.koizumi@isl.cs.waseda.ac.jp
[b)] sugawara@waseda.jp

このような記号的なるアプローチがある一方，文法的な特徴量に着目した研究も多く存在する．[4] では，助詞の分布を特徴量としている．また，[5] や [6] のように単語を品詞に置換したときの並びや，品詞とそのままの文字を混合させたとき並びに着目してるものもある．機械的に単語の切れ目や品詞を分析することで，名詞等の単語から文章のトピックを求め，その分布を著者の特徴とする研究も存在する [7].

提案される様々な特徴量の中で，最も有効なものを求めようとする研究も存在する．[8] や [9] では，複数の特徴量の比較，およびその分類法を比較検討しており，文章のジャンルによって，有効性を示すものが異なることを示している．

適切な特徴量を選択し，著者を推定するとき，多くの場合で必ず著者の候補が必要になる．しかし，現実問題で著者の候補を得ることは難しい．そこで，著者の属性を推定することで著者候補を作り出す手助けになる．例えば，[10] では著者の性別を，[11] では，著者の年代推定を試みている．しかしその精度はまだ高くなく，著者の属性を判別するのは著者自身を推定するよりも難しいと言える．

様々な特徴量のうち，特に文法的なものに対する研究は多い．一方で，それらの殆どは品詞に焦点をあてていて，文自体の構造には着目していない．文の構造に着目した研究として，[12]，[13] があげられる．文の構造は，品詞に加えて修飾・被修飾の関係を加味した特徴量であり，文を書く際に著者の意思が強く表れる可能性は高い．文は，文節という単位に区切ることができ，ある文節は別の文節を修飾する，という形で文を形づくっている．この関係を基に，文から木を作成することができる．同一著者が書いた文は，平均して文の構文木の類似性が高いことを，これらの研究では示している．

文章には，書き手の特徴が表れている．この書き手らしさを表す特徴を特徴量と呼び，これを活用して著者推定がなされている．[12]，[13] では，構文構造に著者の特徴が表れるかについてのみ焦点を当てているため，複雑な計算を行い類似性を評価している．しかし 1 節で述べたような問題の解決を目的とする場合，簡単な計算である程度の精度を保つ方が重要と言える．

そこで本研究では，構文構造のなかでも単純な構造にのみ焦点を当てて，文章の類似性を評価可能か評価する．また，相対的な値の大小より，著者候補の中から正しい著者を推定可能かを評価する．具体的には，文の係り受け関係に着目しその類似性を定量化する．比較のため，対象テキストは小説とし，[12]，[13] に登場する Sub tree，Subset tree に着目した類似性の定量化も行う．また，係り受け関係に着目する手法が単純かつ高速であることを示すため，処理速度の計測を行う．

表 1 ルール 1

記号	a	c	d	j	n	r	v	x
品詞	形容詞	接続詞	副詞	形容動詞	名詞	連体詞	動詞	感動詞

表 2 ルール 2

記号	説明
adt	連用テ形
ccr	ナガラ
cd1	順接仮定バ
cd2	順接仮定ト
fol	ニシタガイ
cn1	逆接仮定テモ
cn2	逆接確定ガ
inc	ウチニ
nom	準体助詞ノ
pps	目的を表すニ
qut	引用を表すト
rsn	原因を表すカラ
sml	トキ

3. 準備

3.1 テキストの前処理

構文構造そのものだけに着目するため，実験の前に文章にいくつかの処理を施し純粋な文章のみを抽出する．本節ではその詳細について述べる．

3.1.1 クリーニング

青空文庫から得たテキストデータは，作品名や作家名，ルビなどが混在しておりこのまま使用できない． したがって以下のクリーニングを行い図 1 から図 2 のようなテキストへ変換する．

(1) 本文のみを取り出す
(2) ルビ, 空白, 改行を削除する
(3) 鍵括弧を地の文へ変換する

ただしスペースの問題上，小説本文の途中を省略して表示する．

3.1.2 縮約的還元

構文構造自体に注目するために，単語の影響を排除する．そこで文に含まれる単語を品詞ごとに記号に置き換える．これを縮約的還元と呼ぶ ([12], [13])．縮約的還元ルールを以下に示す．

ルール 1. 品詞ごとに記号に変換する．品詞と記号の対応を表 1 に示す．

ルール 2. 名詞に断定の助動詞「だ」が続く場合，n に連続して cop という記号を付与し ncop と表す．

ルール 3. 動詞に続く助動詞の種類によって，v に連続して以下の記号を付与する．具体的な記号を表 2 に示す．

飴だま
新美南吉

【テキスト中に現れる記号について】

《》：ルビ
（例）わたし舟《ぶね》

　春のあたたかい日のこと、わたし舟《ぶね》にふたりの小さな子どもをつれた女の旅人《たびびと》がのりました。
　舟《ふね》が出ようとすると、
「おオい、ちょっとまってくれ。」
～略～
　そして、
「そオれ。」
とふたりの子どもにわけてやりました。
　それから、またもとのところにかえって、こっくりこっくりねむりはじめました。

底本：「ごんぎつね　新美南吉童話作品集 1」てのり文庫、大日本図書
　　1988（昭和 63）年 7 月 8 日第 1 刷発行
底本の親本：「校定　新美南吉全集」大日本図書
入力：めいこ
校正：鈴木厚司、もりみつじゅんじ
2003 年 9 月 29 日作成
青空文庫作成ファイル：
このファイルは、インターネットの図書館、青空文庫（http://www.aozora.gr.jp/）で作られました。入力、校正、制作にあたったのは、ボランティアの皆さんです。

図 1　クリーニング前

春のあたたかい日のこと、わたし舟にふたりの小さな子どもをつれた女の旅人がのりました。舟が出ようとすると、おオい、ちょっとまってくれ。
～略～
そして、そオれ。とふたりの子どもにわけてやりました。それから、またもとのところにかえって、こっくりこっくりねむりはじめました。

図 2　クリーニング後

ルール 4. 動詞，形容詞，形容動詞，断定の助動詞「だ」が連用形に活用される場合，連続して adv という記号を付与する．例えば，「動かない」は「vadv ナイ」と表す．

ルール 5. 以上のルールに当てはまらない文字はカタカナでそのまま表記する．

表 3　文 1 の品詞

単語	品詞	適応ルール	縮約的還元後の表記
その	連体詞	1	r
作家	名詞	1	n
の	助詞	5	ノ
日常	名詞	1	n
の	助詞	5	ノ
生活	名詞	1	n
が	助詞	5	ガ
そのまま	副詞	1	d
作品	名詞	1	n
に	助詞	5	ニ
も	助詞	5	モ
あらわれ	動詞	1,4	vadv
て	助詞	1	テ
居り	動詞	1,4	vadv
ます	助動詞	5	マス

以上のルールを適用し縮約的還元を行う．たとえば「その作家の日常の生活が、そのまま作品にもあらわれて居ります。」という文を用いて例を示す．この文 1 を単語ごとに区切ると「その/作家/の/日常/の/生活/が/、/そのまま/作品/にも/あらわれ/て/居り/ます/。」となる．各単語について，品詞，適応ルールと縮約的還元後の表記はそれぞれ，表 3 のようになる．表 3 をもとに縮約的還元を行う．例えば，動詞「あらわれ」を変換する際，次に助詞「て」が続いているためこの動詞は連用形に活用される．したがって，ルール 1 とルール 4 を用いると「vadv」となる．動詞「居り」も同様である．これらを繰り返して得られた変換後の文は「r n ノ n ノ n ガ 、 d n ニモ vadv テ vadv マス 。」となる．

3.2　構文木

3.1 節で示した下処理を施したテキストから 1 文を取り出し，構文木を作成する．構文木作成にあたり，必要な係り受け解析は 3.2.1 節で説明する．構文木の作成については 3.2.2 節で説明する．

3.2.1　係り受け解析

構文木作成には，形態素解析と係り受け解析は必須である．文章は大きいものから順に，文章，段落，文，文節，単語という単位で分割できる．形態素解析とは文章を単語ごとに区切り, 単語に品詞タグを付与するものである．また係り受け解析とは文節同士の修飾・非修飾の関係を分析するものである．係り受け解析は形態素解析を基に行われる．本研究では形態素解析のツールとして MeCab[*1]，係り受け解析のツールとして CaboCha[*2]を使用する．図 3 に CaboCha の実行例を示す．この文の係り受け関係は，「その→作家の」，「作家の→日常の」，「日常の→生活が、」，「生

[*1] http://taku910.github.io/mecab/
[*2] https://taku910.github.io/cabocha/

```
その作家の日常の生活が、そのまま作品にもあらわれて居ります。
その 連体詞,*,*,*,*,*, その, ソノ, ソノ
作家 名詞, 一般,*,*,*,*, 作家, サッカ, サッカ
の 助詞, 連体化,*,*,*,*, の, ノ, ノ
日常 名詞, 一般,*,*,*,*, 日常, ニチジョウ, ニチジョー
の 助詞, 連体化,*,*,*,*, の, ノ, ノ
生活 名詞, サ変接続,*,*,*,*, 生活, セイカツ, セイカツ
が 助詞, 格助詞, 一般,*,*,*, が, ガ, ガ
、 記号, 読点,*,*,*,*,、 ,、 ,、
そのまま 副詞, 一般,*,*,*,*, そのまま, ソノママ, ソノママ
作品 名詞, 一般,*,*,*,*, 作品, サクヒン, サクヒン
に 助詞, 格助詞, 一般,*,*,*, に, ニ, ニ
も 助詞, 係助詞,*,*,*,*, も, モ, モ
あらわれ 動詞, 自立,*,*, 一段, 連用形, あらわれる, アラワレ, アラワレ
て 助詞, 接続助詞,*,*,*,*, て, テ, テ
居り 動詞, 自立,*,*, 五段・ラ行, 連用形, 居る, オリ, オリ
ます 助動詞,*,*,*, 特殊・マス, 基本形, ます, マス, マス
。 記号, その-D
作家の-D
日常の-D
生活が、-------D
そのまま---D |
作品にも-D |
あらわれて-D
居ります。
EOS
```

図 3 CaboCha の実行例の一部

表 4 文章 1 のエッヂとノード

ノード	エッジ
r	r → n ノ
n ノ	n ノ→ n ノ
n ノ	n ノ→ n ガ
n ガ	n ガ→ vadv テ
d	d → vadv テ
n ニモ	n ニモ→ vadv テ
vadv テ	vadv テ→ vadv マス
vadv マス	

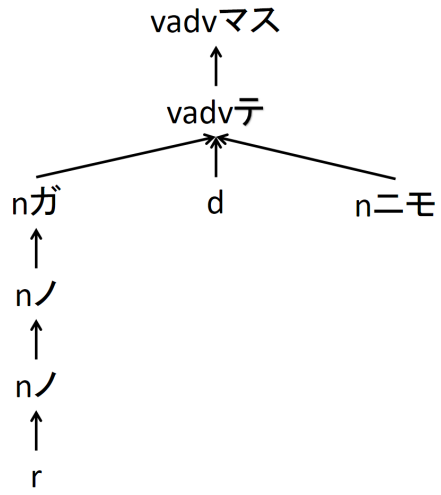

図 4 文章 1 の構文木

活が、→あらわれて」,「そのまま→あらわれて」,「作品にも→あらわれて」,「あらわれて→居ります。」となることが分かる．

3.2.2 構文木の作成

本研究で使用する構文木は，以下に説明する，狭義の構文木を指す．本研究で用いる構文木は一文に現れる文節をノードとし，係り受け関係を用いてエッジを作成する．係り受けには向きが存在するため，構文木のエッジにも向きが存在する．したがって，生成される構文木は根となるノードが決められた根付き木，かつ，エッジが葉から根に向かう有向木となる．例えば，文 1 「その作家の日常の生活が、そのまま作品にもあらわれて居ります。」を構文木に変換する場合，まず，3.1.2 節で述べた縮約的還元を行う．この構文木のエッヂとノードは表 4 のようになる．文 1 の構文木を図 4 に示す．

4. 文の類似性の評価

3.2.2 節で述べた方法を用いて構文木を作成し，2 つ構文木の類似性を求める．本研究では，2 つの構文木の同じ構造を数え上げることで類似性を定量化する．定量化の方法に必要な概念として，ST，SST の説明を 4.1 節，ME の説明を 4.2 節で行う．これらを用いた定量化について 4.3 で説明する．

4.1 ST と SST

SST (Subset Tree) と ST (Sub Tree) は，木構造の一部であり，それ自身も木構造となるものを指す [14]．その中で，木構造の任意のノードがそれ以下の全子ノードとともに構成するものを ST，一部の子ノードとともに構成するものを SST と呼ぶ．また，それぞれの集合を STs，SSTs と呼ぶ．文 1 を例にとり，図 5 に STs，図 6 に SST の一部分を示す．図 5 より，文 1 の構文木には ST が 5 個り，SST は図 6 のものも含めて 37 個ある．

4.2 ME

構文木においてふたつのノードとそれをつなぐ 1 本のノードで表現される構文木の一部を係り受けエッジ (以下 ME) と呼ぶ．文 1 のうち ME にあたる構造を図 7 に示す．ME は，ST や SST と比較してサイズが小さく，類似性の定量化に要する時間が少ない．

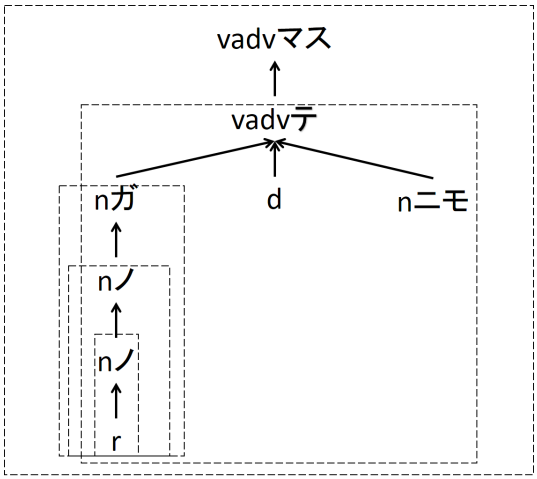

図 5　文 1 の ST

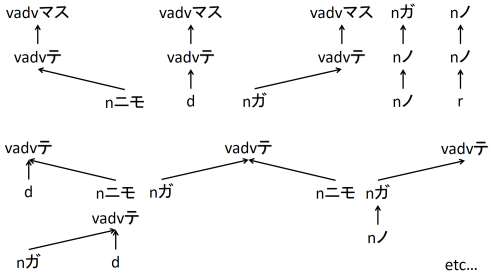

図 6　文 1 の SST の一例

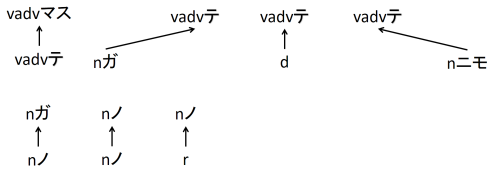

図 7　文 1 の ME

4.3　文の類似性の定量化

ふたつの文が何らかの観点から類似していれば，それは同一の著者が書いたものである可能性が高い．本研究では，その判断のための手法として構文構造に着目する．文は，品詞を持つ単語を意味が通るように組み合わせたものと解釈でき，その組み合わせ方は様々である．この組み合わせ方が似ているとき，構文構造が似ていると考える．文の類似性を定量化するため，本研究では 2 文から生成される構文木に共通する ST，SST，ME の数をそれぞれ求める．以下に定義を示す．

ある文 V から生成された構造木を T とし，それに含まれる ST の集合を U_T，SST の集合を U'_T，ME の集合を U''_T とする．V_1，V_2 から生成される構文木をそれぞれ T_1，T_2 とし，共通する ST，SST，ME の数をそれぞれ $S(T_1,T_2)$，$S'(T_1,T_2)$，$S''(T_1,T_2)$ とすると，それぞれ以下の式で表される．

表 5　使用テキストの作家と作品数

作家	作品数
芥川龍之介	102
太宰治	125
国枝史郎	61
宮本百合子	138
宮沢賢治	90
夏目漱石	27
新実南吉	45
小川未明	273
田中貢太郎	47
岡本かの子	176

$$S(T_1,T_2)=\sum_{t\in U_{T_1}} I(t,T_2) \tag{1}$$

$$S'(T_1,T_2)=\sum_{t\in U'_{T_1}} I(t,T_2) \tag{2}$$

$$S''(T_1,T_2)=\sum_{t\in U''_{T_1}} I(t,T_2) \tag{3}$$

ここで $I(t,T_2)$ は，部分木 t が木 T_2 に現れるとき 1，それ以外は 0 の値を返す関数である．本研究では対象テキストから指定の文数取り出し (n 文とする)，総当たりで S，S'，S'' を求める．求めた n^2 個の値の総和をそれぞれ，ST 構文類似度，SST 構文類似度，ME 構文類似度と呼ぶことにする．

5.　実験

同著者同士の構文類似度が異著者同士よりも高いとき，その手法が著者を推定する手法として有効と言える．本節では，使用するテキストについて 5.1 節で， 実験の流れについて 5.2 節で述べる．また，各実験ごとに 5.3 節，5.4 節で説明する．

5.1　使用テキストと文の選択

今回使用するテキストは青空文庫の小説ページ[*3]に掲載がある作家のうち，比較的作品数の多い著者 10 人の作品を用いた．著者と小説作品数を表 5 に示す．

また，本実験では著者の推定を目的とするため，同著者同士の構文類似度も算出する． このとき，全く同じ文同士の値が大きくなるのは自明である． したがって，たまたま同じ文章が選択されてしまうことを防ぐため，比較するテキストデータに同じ文が入らないようにする．また，文節の数が多くなれば，構文類似度が大きくなる可能性も大きくなる．したがって，文節数の影響を減らすため，文節数が 5 から 10 の文を取り出す．この文節数は，今回使用したテキストのうち，占める割合が多く，かつ文節数が少なすぎない範囲を考慮して選択した．

[*3] http://yozora.kazumi386.org/9/1/ndc913.html

5.2 実験の流れ

各構文類似度を求める実験の流れを示す．取り出す文は実験 1 では 100 文，実験 2 では 20 文，40 文，60 文，80 文，100 文を取り出して実験を行う．ここでは，n 文と示す．また，簡単のために著者名を A から J で表す．

step1．各著者 n 文取り出す．
step2．著者 A から n 文取り出す．このとき，step1 で取り出した文は選択しない
step3．著者 A から J と著者 A の各構文類似度を求める．
step4．求めた値の中で最大値をとるものを 100 として，他の値にもスケーリングを行う
step5．step2 から step4 を著者 B から J についても同様に行う．

本実験では，複数ある著者候補から文章の書き手を選択する形式をとる． したがって，同著者同士から算出された値と，異著者同士から算出された値を比較する必要がある．今回評価の基準としている値は試行によって平均値が変化するため，結果の平均をとるにあたりスケーリング (step4) を行う．

5.3 実験 1

実験 1 では，各著者の作品中からランダムに 100 文取り出し，各構文類似度を算出する．

5.4 実験 2

実験 2 では，取り出す文の数を 20 文，40 文，60 文，80 文，100 文と変化させて，テキスト量の増加による各構文類似度，実行時間のそれぞれの変化を比較する．

6. 実験結果

6.1 実験 1

各著者の作品中からランダムに 100 文取り出し，各構文類似度を求める．結果を表 6〜8 に示す．行内で構文類似度が上位 2 位以内の数値を枠で囲っている．

同一著者同士の値が行で上位 2 以内に入っている人数はそれぞれ，3 人，7 人，8 人である．また全手法に共通して，岡本はどの著者との値も高く，岡本よりは減るが田中も同様である．反対に太宰，国枝は自分自身を含めたどの著者と比較しても値が低い．加えて，岡本の値に着目すると，ST 構文類似度に比べ SST，ME 構文類似度が最大値をとなる値が減っている．また，SST から ME にかけては，最大値となる値は殆ど変化がないが，最大値を出している値自体が減少している．

6.2 実験 2

作家 10 人の小説から取り出す文を 20 文から 100 文まで，20 文ごとに増やして各構文類似度を求める．同一著者同士の値が行内で上位 2 位以内である著者数をグラフにしたものを図 8 に，求めるのに要した時間を図 9 に示す．20 文から 100 文までの行内順位 2 以内の著者数に関して平均を求めると，ST，SST，ME はそれぞれ 7 人，5.8 人，3.4 人となっている．この情報から著者を推定する場合を考えると，テキスト量の変化によってほとんど精度は変化しない．手法ごとに比較すると，精度の高さは ME ＞ SST ＞ ST である．また，実行時間に着目すると，時間の短い順に ME，SST，ＳＴであり，テキストが 100 文のとき，ST は ME の約 2.4 倍，SST は ME の約 1.9 倍の実行時間を要する．

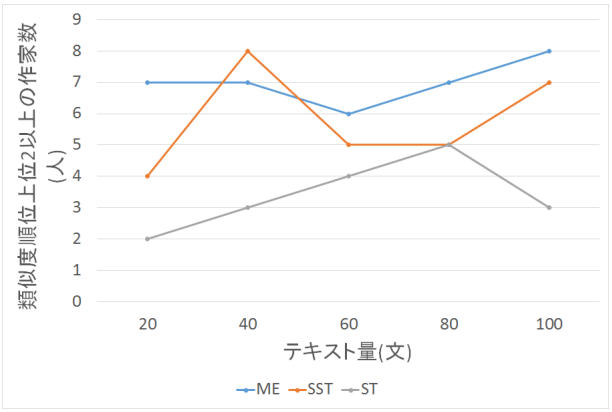

図 8　各手法における構文類似度上位 2 位の著者数

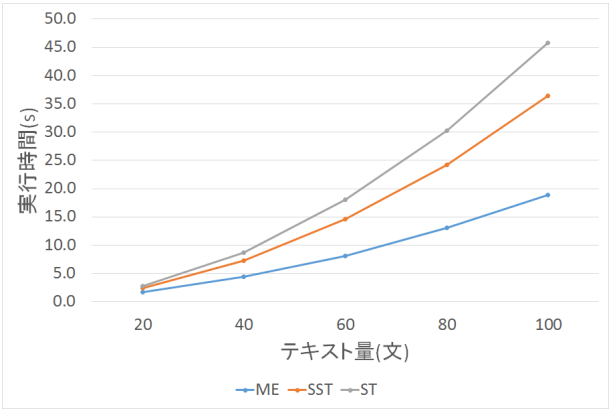

図 9　各手法における実行時間

7. 考察

7.1 実験 1

同一著者同士の値が行で上位 2 以内に入っている人数は ME と SST が多く，ST はこのふたつに比べて少ない．これは SST や ME よりも ST の方が，部分構造としての大きさが比較的大きいことが起因してると考えられる．これにより，違う著者同士で偶然同じＳＴを持つ文が表れたときに，数値が大きく影響を受けている．また，各手法で共通して岡本，田中の列では値が大きく，反対に太宰，国枝の列では小さい．他者との値の大小については広く用いられる構文を使うかどうかにより，自身との値の大小について

表 6　ST 構文類似度

	芥川龍之介	太宰治	国枝史郎	宮本百合子	宮沢賢治	夏目漱石	新実南吉	小川未明	田中貢太郎	岡本かの子
芥川龍之介	97.2	65.3	62.7	66.9	63.2	75.4	76.2	77.0	83.6	86.2
太宰治	90.0	71.5	68.0	68.4	71.4	75.1	80.5	82.1	91.7	88.9
国枝史郎	86.1	67.3	77.1	69.9	71.9	81.8	80.9	80.1	88.0	95.0
宮本百合子	91.2	73.1	75.1	73.1	72.2	81.2	81.4	81.3	88.3	88.5
宮沢賢治	83.4	72.9	71.1	67.2	76.2	76.8	79.7	81.1	91.2	90.4
夏目漱石	93.6	67.7	73.2	72.5	68.5	83.2	85.1	87.9	86.6	91.6
新実南吉	82.1	67.2	67.2	67.7	65.6	80.4	82.7	81.8	84.2	93.9
小川未明	84.6	71.5	72.1	69.1	70.1	79.6	85.0	94.2	82.8	96.5
田中貢太郎	90.1	72.1	68.4	70.1	74.7	76.0	80.0	78.8	88.2	88.9
岡本かの子	83.7	65.4	69.4	64.3	66.5	78.6	81.0	81.6	82.3	93.6

表 7　SST 構文類似度

	芥川龍之介	太宰治	国枝史郎	宮本百合子	宮沢賢治	夏目漱石	新実南吉	小川未明	田中貢太郎	岡本かの子
芥川龍之介	86.8	64.4	60.9	76.6	56.3	86.2	53.0	54.0	83.2	98.7
太宰治	71.8	67.2	62.2	78.9	64.4	84.3	57.8	57.4	89.2	99.1
国枝史郎	67.9	61.4	63.5	76.8	55.9	79.9	51.7	51.1	80.9	100.0
宮本百合子	66.2	61.4	60.0	75.9	49.4	79.9	41.4	42.3	81.2	100.0
宮沢賢治	59.0	60.0	52.1	59.3	98.5	63.7	82.5	83.6	76.6	78.5
夏目漱石	71.7	62.4	58.8	75.7	50.3	90.1	44.1	48.0	82.0	98.8
新実南吉	60.1	59.1	52.6	54.6	91.0	61.4	91.6	96.8	71.3	75.0
小川未明	57.1	55.0	48.0	52.2	84.1	61.8	89.8	97.2	65.3	68.9
田中貢太郎	65.9	61.3	57.4	73.4	56.7	78.3	49.2	47.9	83.1	97.7
岡本かの子	56.4	50.3	51.4	65.2	42.9	68.1	36.9	36.4	70.9	100.0

表 8　ME 構文類似度

	芥川龍之介	太宰治	国枝史郎	宮本百合子	宮沢賢治	夏目漱石	新実南吉	小川未明	田中貢太郎	岡本かの子
芥川龍之介	86.9	60.0	55.5	73.9	49.3	76.0	52.3	51.7	76.3	96.6
太宰治	79.6	69.1	61.1	79.5	63.0	81.2	60.6	62.2	83.8	95.3
国枝史郎	76.1	60.6	61.9	75.0	52.3	74.5	54.4	48.9	77.2	98.2
宮本百合子	73.9	59.0	57.2	81.7	48.1	75.6	47.1	45.0	76.8	95.6
宮沢賢治	65.6	62.7	47.7	61.1	92.9	66.1	83.2	88.5	74.6	78.2
夏目漱石	79.0	62.2	60.4	81.8	53.4	83.1	54.1	53.5	78.0	95.7
新実南吉	67.7	58.5	52.4	63.2	85.4	65.1	88.9	89.8	70.9	84.8
小川未明	61.0	55.9	45.9	55.2	81.5	58.9	88.5	98.9	62.3	72.0
田中貢太郎	77.6	65.5	59.4	77.8	55.4	75.6	52.8	52.0	80.6	95.3
岡本かの子	65.5	53.4	51.0	68.2	44.4	66.6	45.2	42.7	71.0	100.0

は文体が変化しにくいかどうかによる．したがって，前の二人は広く用いられる構文を用いて書く傾向にあり，なおかつ自分の文体に変化がないこと，後者の二人はあまり広く用いられない構文を用いて書く傾向にあり，かつ文体が変化しやすい著者であることが推測できる．また，ほかの著者は，広く用いられる構文とそうでない構文を同程度に使用しており，かつ文体の変化が少ない著者であることが推測できる．

7.2　実験 2

図 8 から，各手法の精度はテキスト量よらない．また，40 文の場合を除いて ME ＞ SST ≧ ST となっている．各試行の平均値も 7 人＞ 5.8 人＞ 3.4 人と同様である．これはテキスト量に関わらず ME が最も著者推定に有効な手法であることを示している．また，テキスト量の減少が各手法に大きな影響を与えないことが分かる．理由として，構文上の特徴が少量のテキストにも表れていることが推測される．ただし少量のテキストでは，同著者同士の値が行内順位 2 以内であっても，他の数値との差が小さい．図 9 では，実行時間が最も短い手法が ME であることが示されている．したがって，ME は最も単純かつ高速な手法であると言える．実行時間と精度を合わせて考えると，提案手法である ME は早い時間で同程度以上の精度を保つことが示されている．

8.　まとめと今後の課題

本実験では ME による著者の推定を提案し，[12]，[13] で用いた SST，ST による手法と比較した．結果，短い時間で ST よりも高い精度，また，ＳＳＴと同程度以上の精度での推定ができることを示した．また，著者の使用する構文の変化や個性によっては推特徴を捉えやすい著者，捉えにくい著者がいると言える．以上から，構文構造上の特徴を捉える方法として，ME に焦点を当てる手法が有効であることが分かった．本実験ではＭＥにのみ着目した場合でも，著者を推定が可能な構文上の特徴を捉えることができるか，という点に重きを置いている．そのため，作家数も 10 人と少ない．次の段階としては，著者の候補を増やしても著者の推定が可能な特徴抽出ができるか，より現代に近い時期に書かれた文章ではどうなるか，などが考えられる．1 節で述べた問題を解決するには著者の推定する精度向上が必須であると言える．

参考文献

[1]　財津亘，　金明哲：テキストマイニングを用いた犯罪に関わる文書の筆者識別，日本法科学技術学会誌 (2014).
[2]　松浦司，金田康正：近代日本小説家 8 人による文章の

n-gram 分布を用いた著者判別，情報処理学会研究報告自然言語処理，Vol. 2000, No. 53, pp. 1–8 (2000).
[3] 金明哲：読点から現代作家のクセを検証する，統計数理， Vol. 44, No. 1, pp. 121–125 (1996).
[4] 金明哲：S8-5 助詞の分布に基づいた文章の原著者の認識，日本行動計量学会大会発表論文抄録集，Vol. 24, pp. 144–147 (1996).
[5] 金明哲：品詞のマルコフ遷移の情報を用いた書き手の同定 (工学・ソフトウェア, 第 32 回 日本行動計量学会大会発表一覧)，行動計量学， Vol. 32, No. 1, pp. 101–102 (2005).
[6] 金明哲：文節パターンに基づいた文章の書き手の識別，行動計量学， Vol. 40, No. 1, pp. 17–28 (2013).
[7] 白井匡人，三浦孝夫：LDA を用いた著者推定，*DEIM Forum*, pp. 2–11 (2011).
[8] 金明哲：統合的分類アルゴリズムを用いた文章の書き手の識別，行動計量学，Vol. 41, No. 1, pp. 35–46 (2014).
[9] 三品光平，松田眞一ほか：文章の書き手の同定における分類法の精度比較，アカデミア. 情報理工学編: 南山大学紀要= Academia. Information sciences and engineering: journal of the Nanzan Academic Society， Vol. 13, p. 35 (2013).
[10] 池田大介，南野朋之， 奥村学：blog の著者の性別推定，言語処理学会第 12 回年次大会，pp. 356–359 (2006).
[11] 太田貴久， 増山繁：青空文庫を対象とした書き手の識別とその応用，言語処理学会年次大会発表論文集 (2009).
[12] 金川絵利子，佐原諒亮， 岡留剛：構文構造に着目した文体の類似度の数値化，日本人工知能学会 (2016).
[13] 金川絵利子，佐原諒亮， 岡留剛：情報量木カーネルとそれに基づく作家の構文類似度解析，日本法科学技術学会誌 (2014).
[14] Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees, *European Conference on Machine Learning*, Springer, pp. 318–329 (2006).