自然言語処理 137－1
（2000．6．1）

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

松浦 司* 金田 康正**

*東京大学大学院理学系研究科情報科学専攻
*東京大学情報基盤センタースーパーコンピューティング部門

概要

本稿では、文章中のn-gram分布状況を著者の特徴量として、文章の著者を推定する手法を提案する。文章中におけるn-gram出現確率分布関数間の非類似度に基づいて著者推定を行うが、非類似度は提案関数*dissim*の他、Tankardの手法、ダイヴァージェンス、およびクロスエントロピーを用いてそれぞれ計算し、4関数の著者判別精度を比較した。1-gramから10-gram分布を特徴量とし、日本近代作家8人の92作品を対象とする著者推定実験結果について報告する。本手法は文章に関する付加的な情報を全く必要とせず、形態素解析などを要求しない。また特定の言語および文章の性質を利用しないため、多くの言語・テキストにそのまま適用可能であることが期待できる。

Authorship Detection of Sentences by 8 Japanese Modern Authors via N-gram Distribution

Tsukasa Matsuura* and Yasumasa Kanada**

*Department of Information Science, Faculty of Science,
Graduate School of the University of Tokyo
**Information Technology Center, Computer Centre Division, the University of Tokyo

Abstract

We propose a method for authorship detection based comparisons between n-gram distributions in sentences. The authors are detected via dissimilarity between probability distribution functions of n-grams in sentences. We have compared four functions to measure the dissimilarity, i.e. *dissim* (proposed function), Tankard's method, divergence and cross entropy. We report the experiments where the 92 works in total by 8 Japanese modern authors are analyzed via from 1-gram to 10-gram distribution. Our method requires no additional information on texts, i.e. no preliminary analyses. All the machine-readable texts can be attributed by the same method.

1 はじめに

本稿では、文字 n-gram[1]の出現確率分布を手がかりに文章の著者を判別する手法を提案する。本手法では文字を単位として分析を行い、形態素情報や構文情報を必要としないため、実験対象テキストに対して形態素解析などの処理を前もって施す必要がない。一語あたりの文字数や [2] 一文あたりの語数 [5] を特徴量として著者推定を行う手法もこれら高次の情報は要求しないが、文章中における単語間の区切れが明確であることを前提としているので、日本語や中国語など単語分かち書きが難しい言語で書かれた文章に直接適用することはできない。また対象言語に応じて分析方法を調整する必要がなく、機械可読文書全てを同一の方法で処理できることも提案手法の大きな利点である。

従来の著者推定に関する研究では、書き手の癖の現れそうな特徴量を人間が経験に基づいて仮定するもの [3] [1] [6] [7] [9] が多かったが、本手法はそうした人手による調整を要求しない。更に本手法を用いて著者を特徴づける文字列 (著者特徴 n-gram) を抽出することができる [8] ので、どの様な点に書き手の癖が現れるかを著者特徴 n-gram に基づいて求めることも可能であると考える。本研究は、従来人間が経験から構築していた分析の観点を、自動的に提供することを目標の一つとしている。

研究報告 [8] では 3-gram の出現確率分布に注目し、同手法によって芥川龍之介および菊池寛による小説文の著者判別を試みる実験を報告した。本稿ではより多くの書き手、即ち日本近代作家 8 人による文章の著者判別を試みた実験結果を報告する。分析対象とするテキストの量も増大させ、総計 3,001,650 文字の文章を用いて実験を行った。

[1]本稿では n-gram は長さ n 文字の文字列を指す。提案手法は n-gram モデルに基づくものではないのでご注意願いたい。

2 n-gram 分布による著者推定

文字に着目して著者推定を行うにあたり、文字同士の隣接状況を分析に反映させるために文章の n-gram 分布を使用する。n-gram 分布とは、n 個の文字が隣接して生じる文字の共起関係、即ち n-gram の出現確率を記録したものである。文字同士の隣接状況を無視し、一文字一文字の出現確率を記録したい場合には、n の値を 1 にすればよい。

n-gram 出現確率分布関数間の非類似度を文章間で計算し、互いに非類似度が低い文章群ほど同一の著者によって書かれた可能性が高いものと考える。n-gram 出現確率分布関数間の非類似度 *dissim* を以下のように定義する。

長さ n の文字列 (n-gram) を x で表す。x_i を分析対象とする言語の文字とすると、x は $\mathrm{x}=\{x_1x_2\cdots x_n|\forall i.x_i\in\chi\}$ と表される。文章 P 及び Q の n-gram 分布がそれぞれ確率分布関数 $P(\mathrm{x})$ 及び $Q(\mathrm{x})$ で表されるとき、文章 P,Q 双方に出現する n-gram(共通 n-gram) の集合を C とし $C=\{\mathrm{x}|P(\mathrm{x})Q(\mathrm{x})\neq 0\}$ と定義する。

文章 P,Q 間の類似度を以下の $dissim(P,Q)$ によって表す。但し $\mathrm{card}(C)$ は集合 C の要素数を表す。

$$dissim(P,Q)=\frac{1}{\mathrm{card}(C)}\sum_{\mathrm{x}\in C}\left|\log\frac{P(\mathrm{x})}{Q(\mathrm{x})}\right| \qquad (1)$$

$P(\mathrm{x})$ と $Q(\mathrm{x})$ とが全く同じであるときに限って *dissim* の値は 0 となる。また、*dissim* は $P(\mathrm{x})$ と $Q(\mathrm{x})$ とに関して対称である。猶、*dissim* は研究報告 [8] で提案した類似度関数 *sim* と同一のものであるが、$P(\mathrm{x})$ と $Q(\mathrm{x})$ との間の相違が大きいほど大きな値を示す関数であり、文章間の similarity よりは寧ろ dissimilarity を表す尺度であるので、直感的に理解しやすい名称に改めた。

著者推定のための確率分布関数間類似度尺度としては、他に Tankard により以下の関数が提案されている [4]。

$$Tankard(P,Q)=\sum_{\mathrm{x}}|P(\mathrm{x})-Q(\mathrm{x})| \qquad (2)$$

本稿では *dissim*、Tankard の手法、ダイヴァージェンス (Kullback-Leibler 情報量)、及びクロスエントロピーを用いて文章間非類似度を計算し、それぞれの非類似度に基づいて著者判別を試みた実験について報告する。ダイヴァージェンス及びクロスエントロピーは、確率分布関数間の相違を表す代表的な関数として比較対象に選択した。但し、フロアリングやスムージングに対応する処理は行わない。ダイヴァージェンスとクロスエントロピーとの定義を以下に示す。式 (7) で示されている通り、ダイヴァージェンスとクロスエントロピーとは独立なものではないが、別個の非類似度評価関数として扱う。

$$D(P,Q) = \sum_{\mathrm{x}\in C} P(\mathrm{x}) \log(P(\mathrm{x})/Q(\mathrm{x})) \tag{3}$$

$$\begin{aligned} H_C(P,Q) &= \sum_{\mathrm{x}\in C} -P(\mathrm{x}) \log Q(\mathrm{x}) && (4)\\ &= \sum_{\mathrm{x}\in C} -P(\mathrm{x}) \log P(\mathrm{x}) && (5)\\ &\quad + \sum_{\mathrm{x}\in C} P(\mathrm{x}) \log \frac{P(\mathrm{x})}{Q(\mathrm{x})} && (6)\\ &= H(P(\mathrm{x})) + D(P(\mathrm{x}), Q(\mathrm{x})) && (7) \end{aligned}$$

3 著者推定実験対象

実験にした文章とその著者とを付表に示す。分析した文章数は総計 92 であり、文章中に含まれる総文字数は 3,001,650 文字である。書き手の数は 8 人であり、いずれも明治から昭和初期にかけての日本近代作家である。執筆時期が互いに近く、かつ十分な量の機械可読文章が入手できる書き手としてこの 8 人を選択した。文章は全て青空文庫[2]によって公開されているものを使用した。

分析対象とした全 92 作品は、72 本の小説、9 本のエッセイ、5 本の書簡形式文章, 3 本の戯曲、2 本の日記、一本の談話からなる。これらの作品は旧仮名遣いおよび旧字体を用いて執筆されたものと成立時期から推定されるが、うち 67 作品は現代仮名遣いおよび新字体に改められている。11 作品は文語体で記述されており、国木田独歩と菊池寛とによる文章に関しては、文語体によるものと口語体によるものとが混在している。著者の特徴を把握するためには仮名遣いおよび字体が統一されている文章を分析することが望ましいが、著者推定手法としては、文章のジャンル、仮名遣い、字体、および文体の違いに対して頑健である方が有用である。

作品の文章には作品名や著者名、章名は含めないが、改行や空白はそれぞれ一文字とみなす。但し、空行はデジタル化の際に整形のために挿入されたものとして予め削除してある。また改行後の空白は段落を改める意味を持つが、直前の改行によって段落の区切りは示されており冗長であるのでこれも削除した。JIS X 0208 で表現できない文字は外字の存在を表すタグと置き換える。外字タグはそれぞれ一文字として計算する。英文字の大文字と小文字とは区別するが、1 バイト文字の英文字と 2 バイト文字の英文字とは同一のものとみなす。Tankard は大文字・小文字の違いを無視した英文字 26 種のみに着目しているが [4]、Tankard の手法を用いる場合も上記の文字セットを分析対象とする。

比較するテキスト間で文字数が著しく異なると *dissim* の信頼性が低下し、また短い文章から著者の癖を把握することは難しいと考えられるので、非類似度を計算するテキストの長さは全て 3 万字とした。この文字数は、最低約 2 万 2 千文字の文章の著者判別が可能であった実験 [8] に基づいて決定した。3 万字より長い作品は先頭の 3 万字を取り出して一つの比較用テキストとする。但し 6 万字より長い作品であっても、一作品から複数の比較用テキストを作成することは行わない。これは、同一作品中の文章は作家によって意図的に文体が統一されている可能性があり、また同一のテーマに沿って書かれたものであるので、著者の

[2] http://www.aozora.gr.jp/main.html
ボランティアの手による WWW 上の電子図書館。著作権が消滅した文学作品を中心に無償公開を行っている。

癖以外の要因によって互いの非類似度が低くなると考えられるからである。3万字より短い作品は、3万字に達するまでランダムな順に繋ぎ合わせて使用する。この場合も一作品中の文章が複数の比較用テキストに含まれることは禁止する。また作品の繋ぎ合わせ方を変えて全部で10の比較用テキストセットを作成し、それぞれにおいて著者推定精度を測定した。

分析するテキストの長さと著者判別精度との関係を観察するため、1万字から2万7千5百字の長さのテキストを用いて著者を判別することも試みた。これらの短いテキストは、前述の3万字の比較用テキストの先頭から当該文字数を取り出して作成した。

4 著者推定の精度

著者判別は、比較用テキストセットの中から比較基準テキストを一つ選択し、この基準テキストと他のテキストとの間の非類似度を計算することで行う。基準テキストの著者、即ち基準著者によって書かれたテキストと基準テキストとの間の非類似度の最大値が、異著者によって書かれたテキストと基準テキストとの間の非類似度の最小値よりも小さい場合を成功例と定義する。成功例においては、基準テキストとの非類似度の小さい順にテキストをソートすると、基準著者によるテキストが先頭から連続して並ぶ。また、基準著者によるテキストの非類似度の最大値よりも小さい非類似度を示すテキストは、基準著者の手によるものと推定できる。更に、基準テキストとの非類似度が最小のテキストが基準著者によるものである場合を部分成功例と定義する。部分成功例においては、非類似度最小のテキストの著者推定のみが可能である。部分成功例は成功例に含まれる。

全てのテキストをそれぞれ基準テキストとした場合について著者判別実験を行い、全比較例中における著者判別成功例の割合を成功率とし、また部分成功例の割合を部分成功率とする。全比較例数は一つの比較用テキストセットに含まれるテキストの数に等しい。

着目するn-gram分布のnを大きくするほど長い範囲にわたる文字の隣接情報を活用でき、分析精度の向上が期待できる。しかし、分析するテキストの長さに対してnが大きすぎると共通n-gramの数が小さくなり、精度の低下を招く。そこで、1-gramから10-gram分布を特徴量として著者推定精度を測定した。但し、nが6よりも大きくなると共通n-gramが全く抽出できないテキスト対が生じるので、7-gramから10-gram分布を用いた実験はTankardの手法についてのみ行っている。

4.1 評価

10の比較用テキストセットを用いた実験における、*dissim*による著者推定成功率の平均値(図 1)、最高値(図 3)、最低値(図 5)を以下に示す。また図2、図4、および図6はそれぞれTankardの手法を用いた場合の成功率の平均値、最高値、最低値を示している。*dissim*における平均成功率の最高値は95.7%であり、Tankardの手法による最高平均成功率84.2%よりも優れている。また最良成功率に注目すると、*dissim*を用いると最短2万文字のテキストで100%の精度が達成されているのに対し、Tankardの手法では最高でも95.2%の成功率である。最悪成功率に関してはTankardの手法による最高値が85.7%と*dissim*による最高値84.2%を上回っている。しかし精度計測回数は10回であるから、両手法の最悪成功率は殆ど変わらないと言える。

同様に*dissim*とTankardの手法による部分成功率を図7から図12に示す。平均部分成功率(図7および図8)の最高値は、*dissim*が98.0% Tankardの手法が99.0%と両手法に差は殆どない。また最悪部分成功率(図11およ

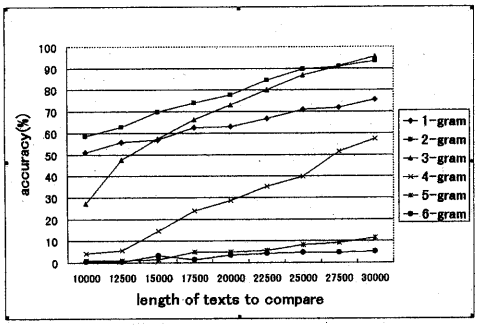

図 1: dissim による著者判別成功率の平均

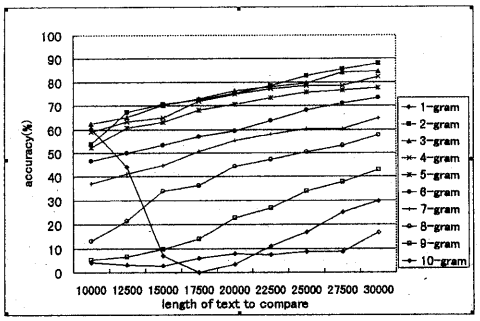

図 2: Tankard の手法による著者判別成功率の平均

び図 12) の最高値は両手法とも同一の値 94.7%を示した。最良部分成功率では *dissim* だけでなく Tankard の手法も精度 100%を記録している。但し、*dissim* は最短 1 万字のテキストに対し 100%の部分成功率を達成できるのに対し、Tankard の手法ではテキストの長さが 1 万 5 千字に達するまで 100%の精度には到達できていない。

全実験を通して、*dissim* を用いる場合は 2-gram および 3-gram 分布が、Tankard の手法を用いる場合には 2-gram から 4-gram 分布が有効であった。テキストの長さが長くなるにつれて精度が向上する傾向があるので、両手法ともより長いテキストを分析した場合にはより良い成績を示す可能性がある。また *dissim* における 4-gram および 5-gram

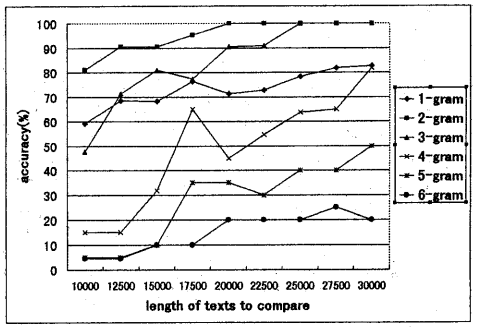

図 3: dissim による著者判別成功率の最高値

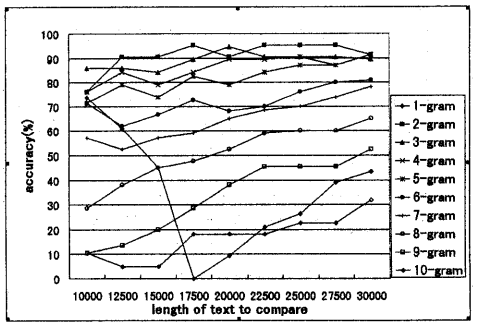

図 4: Tankard の手法による著者判別成功率の最高値

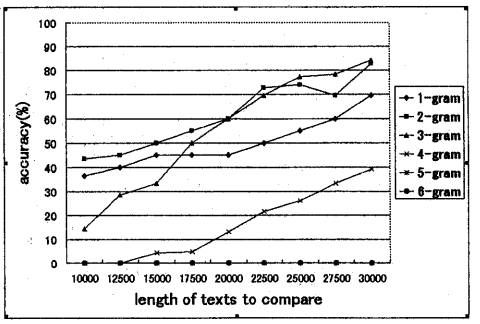

図 5: dissim による著者判別成功率の最低値

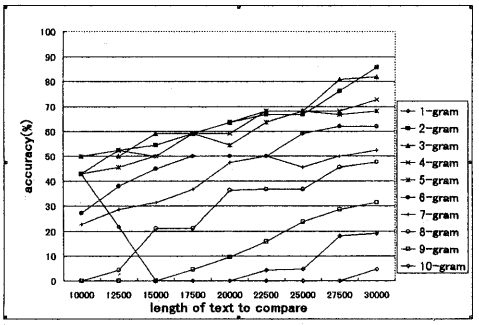

図 6: Tankard の手法による著者判別成功率の最低値

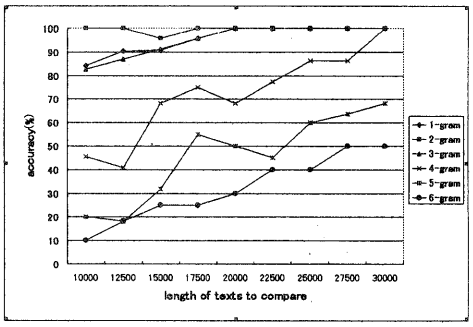

図 9: dissim による著者判別部分成功率の最高値

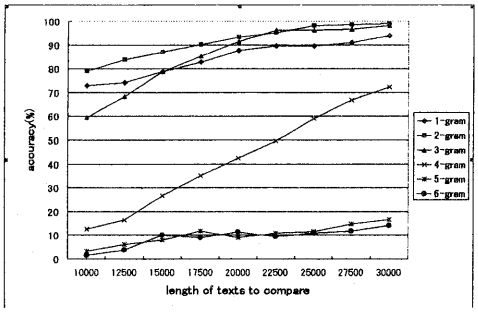

図 7: dissim による著者判別部分成功率の平均

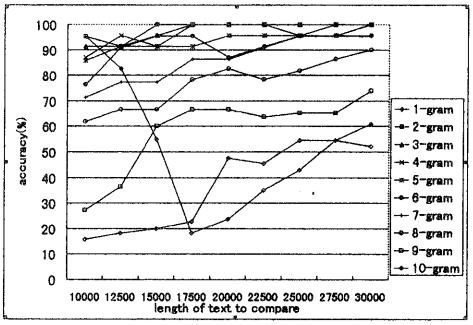

図 10: Tankard の手法による著者判別部分成功率の最高値

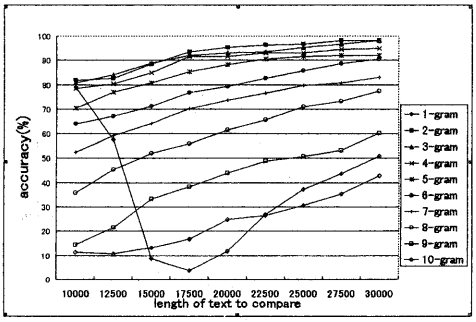

図 8: Tankard の手法による著者判別部分成功率の平均

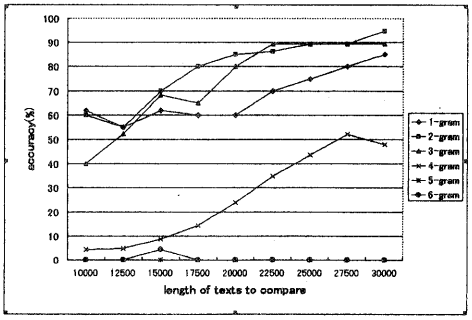

図 11: dissim による著者判別部分成功率の最低値

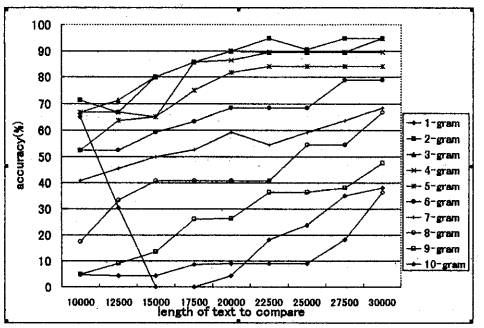

図 12: Tankard の手法による著者判別部分成功率の最低値

分布、Tankard の手法における 5-gram から 10-gram 分布は、対象テキストが長くなった時の精度の伸びが大きいので、長さ 3 万字以上のテキストを分析した場合にはこれらの n-gram 分布が有効になることも考えられる。

10 の比較用テキストセットを用いた全実験において、ダイヴァージェンスによる成功率の最高値は 39.1%、部分成功率の最高値は 72.3%であり、Tankard の手法および *dissim* よりも著しく著者判別能力が低かった。クロスエントロピーに至っては最高でも、4.55%の成功率、および 5.21%の部分成功率を示し、殆ど著者を推定することができなかった。

5 まとめと今後の課題

近代日本作家 8 人の文章合計 92 篇を対象とする著者推定実験において、*dissim* は平均成功率、最良成功率に関して 4 手法中最良の最高値を示した。特に最良成功率では唯一 100%に到達した。最悪成功率、平均部分成功率、最良部分成功率、および最悪部分成功率の最高値に関しては、*dissim* と Tankard との間に殆ど違いは見られなかった。但し最良部分成功率に関しては、*dissim* の方が Tankard の手法に比べ短いテキストの著者推定において部分成功率 100%に到達した。ダイヴァージェンス、クロスエントロピーの著者推定能力は *dissim* および Tankard の手法よりも著しく低かった。

dissim を用いた場合は 2-gram から 3-gram 分布、Tankard の手法を用いた場合には 2-gram から 4-gram 分布を特徴量として用いるのが効果的であった。提案手法は n-gram モデルに基づくものではないが、n-gram モデルに基づいて確率的言語モデルを構築する場合に一般に有効だとされる n-gram、即ち 2-gram および 3-gram[10] にほぼ符合する。

dissim、Tankard の手法ともに、文章のジャンル、仮名遣い、字体、および文体の違いを超えた著者の特徴を把握することができたと考えられる。また複数の作品をつなぎ合わせて作成したテキストの著者推定も可能であったことから、分析対象の書き手がひとまとまりの文章としては短いものしか残していない場合にも両手法は有効であると考えられる。

今後はより多くの書き手による文章を対象に実験を行い、一人の書き手による文章間の非類似度の分布を求めると共に、他言語にも提案手法を適用し、言語による著者の癖の現れ方の異同について考察する予定である。またより長い文章を分析する場合には、より高い推定精度が達成できる可能性や、より n の大きい n-gram 分布が有効になる可能性があり、より大量の文章を扱う実験も行う必要がある。

付表 実験に使用した文章一覧

実験に使用した文章を作家別に示す。括弧内は作家の生没年であり、作家名は生年の順に並べた。

- 国木田独歩 (1871 – 1908) : 源おじ、牛肉と馬鈴薯、非凡なる凡人、少年の悲哀、恋を恋する人、武蔵野、怠惰屋の弟子入り、石清虚、酒中日記、たき火、運命論者

- 岡本綺堂 (1872 – 1939) : 化け銀杏、弁天娘、菊人形の昔、狐と僧、帯取りの池、お照の父、津の国屋、柳原堤の女、幽霊の観世物
- 樋口一葉 (1872 – 1896) : 十三夜、にごりえ、大つごもり、たけくらべ、うつせみ、わかれ道、ゆく雲
- 有島武郎 (1878 – 1923) : 小さき者へ、二つの道、片信、卑怯者、広津氏に答う、一房の葡萄、火事とポチ、小作人への告別、水野仙子氏の作品について、溺れかけた兄妹、宣言一つ、想片、私の父と母
- 菊池寛 (1888 – 1948) : 青木の出京、入れ札、勲章を貰う話、身投げ救助業、M侯爵と写真師、無名作家の日記、大島が出来る話、恩讐の彼方に、勝負事、出世、忠直卿行状記、父帰る、藤十郎の恋、若杉裁判長、ゼラール中尉
- 水野仙子 (1888 – 1919) : 響、輝ける朝、神樂阪の半襟、道—ある妻の手紙—、女、四十餘日、嘘をつく日
- 芥川龍之介 (1892 – 1927) : あばばばば、アグニの神、秋、あの頃の自分の事、或阿呆の一生、或敵打の話、或旧友へ送る手記、或日の大石内蔵助、浅草公園—或シナリオ—)、一塊の土
- 梶井基次郎 (1901 – 1932) : 愛撫、ある崖上の感情、ある心の風景、泥濘、冬の蠅、冬の日、筧の話、過古、器楽的幻覚、Kの昇天—或はKの溺死、交尾、檸檬、のんきな患者、路上、桜の樹の下には、雪後、城のある町にて、蒼穹、闇の絵巻、橡の花——或る私信——

参考文献

[1] ADAMS, L., et al. The Popular Critical View of the Isaiah Problem in Light of Statistical Style Analysis, *Computer Studies in the Humanities and Verbal Behavior*, **7(3-4)** (1973), 149–157.

[2] MENDELHALL, T. A Mechanical Solution of a Literary Problem, *Popular Science Monthly*, **60**, 2 (1901), 97–105.

[3] MOSTELLER, F., et al. Inference in an Authorship Problem, *Journal of the American Statistical Association.*, **58**, 302 (1963), 275–309.

[4] TANKARD, J. The Literary Detective, *BYTE*, **11**, 2 (1986), 231–189.

[5] YULE, G. On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to two Cases of Disputed Authorship, *Biometrika*, **30** (1939), 363–390.

[6] 安本美典 文体を決める3つの因子, 言語, **23**, 2 (1994), 22–29.

[7] 金明哲 助詞の分布に基づいた日記の書き手の識別, 計量国語学, **20**, 8 (1997), 357–367.

[8] 松浦司ら n-gram分布を用いた近代日本語小説文の著者推定, 情報処理学会自然言語処理研究会報告, **99**, 95(NL134) (1999), 31–38.

[9] 村上征勝ら 源氏物語の助動詞の計量分析, 情報処理学会論文誌, **40**, 3 (1999), 774–782.

[10] 北研二 確率的言語モデル, 東京大学出版会 (1999).