2013年度 卒業論文

中古日記文学の計量国語学的検討 –『和泉式部日記』と『更級日記』を題材に–

Metric study on the diaries in the Heian period – Study on “Izumishikibu nikki” and “Sarashina nikki”

太刀岡 勇気
Yuuki Tachioka

2013年12月10日

日本大学 通信教育部 文理学部 文学専攻(国文学)
Department of Japanese Literature, the College of Humanities and Sciences, Distance Learning Division, The Nihon University.

目 次

序章　はじめに

本研究の背景

近年のコンピュータ科学の進展に伴って、人文科学の分野でも自然言語処理の分野で用いられてきた計量的な手法 [1, 2] によって、文献資料や文学作品を分析する研究がおこなわれている [3–5]。古来より、計量的な分析の試みは行われていたが [6]、日本語においてそれが可能になったのは形態素解析器の性能向上によるところが大きい [7]。一般によく知られているのは、テキストの分類と著者同定 (authorship attribution) に計量的手法を用いた例である [5, 8–15]。単語 [10] や句読点の使用法 [15] から著者性をとらえる研究が多い。近年では文学理論との融合を狙った文学的な研究 [16]、心理学・哲学的な考察を試みた研究 [17] から、学生のレポートの類似性判定 [18]、日本語の難易度判定 [19] といった実用的な利用まで様々な応用がみられる。科学的方法に則り客観的な事実から判断することができ、主張に一般性を持たせられるのが最大の特長であり、異なる文献間での客観的比較が簡単に行えるようになった。文学的な観点としては、人間の「内省」に頼った主観的な読みではなく、客観的で網羅的な読みに応用することが期待されている [20]。そのためには、恣意的な部位のみを取り出し主張するのでなく、ある程度網羅的に検討する必要がある [1] 。

研究の立場としては、文字面に注目して通時的にコーパスを追いかけるものや、タグ付けされた品詞情報を用いるものがある。前者の立場の研究は、国語学的な立場からの研究が多く、ある特定のキーワードに注目したものとしては、古くは室町時代語に関しての検討 [22] があげられる。現代語に対しては、新聞コーパスを対象とした研究 [23]、Web を対象とした研究 [24] 等がある。文献 [25] では、和歌と散文をひらがなに開いたときの n-gram を比較することで、手作業では気づかない新たな引き歌の発見に計量分析を役立てている。後者の立場の研究からは、より多くの指標が得られる。例えば、仏典などの中国の漢字文献を対象とした研究 [26]、英語の文献を対象にした研究 [14] が挙げられる。孤立語である中国語や英語では形態素解析の必要がなく、かなり機械的に品詞情報などのタグ付けが行えるため、後者の立場での研究が進んでいる。日本語においても、従来より品詞数や平均文長などから計量的な分析は行われてきた。ところが日本語は膠着語であるため、品詞の

[1] 文献 [21] に計量テキスト分析の 4 つの原則として以下のようなものがあげられていて興味深い。

(1) All quantitative models of language are wrong –but some are useful.

(2) Quantitative methods for text amplify resources and augment humans.

(3) There is no globally best method for automated text analysis.

(4) Validate, Validate, Validate.

ここにある通り、計量分析はあくまでも人間の読みをサポートするためのものである。

タグ付けが難しく、このような研究は長らく人手によって行われてきたため[2]、手間がかかることから、あまり盛んではなかった。近年、コンピュータを用いることでデータベースへのタグ付けと解析の効率が向上した。特に精度の高い形態素解析器 [28] が開発されたことに伴い、品詞情報のタグ付けが機械的に行えるようになったことが大きい。例えば近代文学作品において、形態素解析結果を利用した研究 [11] がある。

従来の研究における問題点は、主に 3 つあると考えられる。1 つ目は解析手法とデータベースの問題である。このような研究において最も重要なのは、データベースの整備と解析手法 (指標) の選択である。手法の選択においては、対象の性質をよく理解した上で有効な方法を選ぶべきであって、作品内容の理解が前提となっていることは言うまでもない。このような手法を取る場合、内容を理解せずに分析を行っても何らかの結果は出るため、そのような研究が多くなっているように思われる。例えば文献 [11] では、青空文庫 [29] から入手した近現代文学のデータベースを「茶筅 [30]」でタグ付けし、形態素解析の誤りを一切考慮せずに分析を行っている。このような分析では、人間の読みに近づけるとは思えない。まずは分析の前段階として、対象とするテキストが統一的な基準でタグ付けされていることが必要である。多くの分析が、古典文学大系等の校訂済み本文を用いて行われているのは問題があると思われる。校訂は複数の写本を元に編者の主観的判断によってなされるため、これによって分析を進めたのでは編者のバイアスが混入することは避けられない。ここではできるだけ本文に近い形でデータベースを作成し、必要な場合には写本の影印にも当たりながら分析をする。また品詞情報といった機械的に算出できる指標を用いた研究が多くなっているのが現状である。品詞情報は確かに重要な情報ではあるがこのような客観的指標だけで人間の記録物の特性をすべて明らかにすることはできないと思われる。形容詞の分類をして頻度分析をするといった従来の主観的尺度に基づく計量分析にも、評価すべき点は多いと思われる。ここでは従来からの主観的手法のよい点も取り入れながら分析を進めることにする。1.4 にも示す通り、形態素解析結果の安易な利用は問題であると考えられる。

2 つ目は対象とする作品の偏りである。これらの研究が対象とする文学作品は、近現代文学 [31] が中心であり、それ以外は近世の西鶴ですらあまり検討されていない [32] のが現状である。データベースの整備に手間がかかるため、これらの研究はある一部の作品 (データベースが整備されているもの) に偏っているのである[3]。中古作品においても、『源氏物語』や『宇津保物語』といった物語文学 [5, 33] や『古今集』といった和歌文学 [34] が中心であり、中古において物語文学と比肩する日記文学に関する検討は見られない。そこでここでは、中古の日記文学の代表格である『和泉式部日記』と『更級日記』を題材に計量的な分析を行い、その特質を明らかにすることを「**目的 1**」とする。

3 つ目は異本の問題がまったく考慮されていないことである。古典文学は著者による原本がほとんど存在せず、現状利用できるのはほとんどが何度も書写を重ねられてきた写本であり、異なる写本 (異本) が残されている [35]。これは近現代文学ではそれほど問題とならないが、近代以前は書写者にオリジナルを尊重する意識がそれほど高くなかったため[4]、改変や創作が行われており、原本を推定するのが困難であるものも多い。実際、書誌学的

[2] 例えば『源氏物語』の著者同定をめぐる一連の研究 [5, 27] では人手で品詞タグを修正した『源氏物語』のデータベースを利用している。

[3] 実際、近代文学作品に関する研究は、青空文庫のデータベースをそのまま用いたものが多い。

[4] 仏典などの書写はきわめて正確に行われている [36]

な検討により、『和泉式部日記』は書写時期が最も古いと考えられる「三条西家本」が最もよい底本とされるが、それだけでは不足であることが国文学的な読みの立場から指摘されている [37]。また誤写などもあり [38]、一つの本だけで本文は同定できない。このような問題がありながらも、従来はなんらかの翻刻を底本に分析が行われている。これには上述の編者によるバイアスが存在する問題がある。計量的な分析手法は異なる作品を区別するような手法を提案しているが、異本のばらつきが異なるテクスト間のばらつきよりも十分小さくなければ、そのような指標はあまり役に立たないといえる。そこでここでは『和泉式部日記』の 4 つの異本を対象に、それらの関係性を明らかにすることを「**目的 2**」とする。

本研究の目的

本研究の目的は前節でも述べたとおり、前節で述べた問題点を考慮しながら、日記文学の特質をあきらかにすることに加え、異本によるばらつきと作品の違いによるばらつきを定量的に分析することにある。そこでまず「目的 1」として、他本間の比較をする。ここでは同程度の分量からなる『和泉式部日記』と『更級日記』の比較を行う。つぎに、「目的 2」として同本 (『和泉式部日記』) 内での異本の比較を行う。これにより、異本によるばらつきと、作品の違いの程度がどの程度指標に反映されるかを検討する。

本研究の構成

以上を研究の目的とした本論文の構成は以下の通りである。分析手法に関して 1 章で述べる。また用いた分析指標に関しても 2 章で述べる。3 章で、目的 1 に対応して、他本間の比較として『和泉式部日記』と『更級日記』の比較を行い、同本 (『和泉式部日記』) 内での異本との比較と合わせて検討する。『和泉式部日記』の系統づけに関して、既往の文献学的なアプローチと本論文で提案する手法の比較も行う。

第1章　計量分析手法

本研究では、複数の異本を一つのテキストから生成可能な独自の仕様を定義し、それに対して計量的手法を用いて分析を行う。本章では分析手法に関して述べる。本研究で用いた分析手法は一般性があるため、他の作品の分析にもそのまま用いることができる。計量的な分析手法は、文献 [39–41] にまとめられている。

計量的な分析手法を通して、実際の作品の分析を行うためには作品の特徴を表すと考えられるなんらかの統計量を得る必要がある。文献 [41] では「内的証拠による比較」として「表記の特徴」、「文字・単語列などの共起関係」、「構造分析」に分類されている。本論ではこのうち「表記の特徴」と「文字・単語列などの共起関係」に分類される特徴量を使う。本研究の分析に用いた統計量に関しては、次章を参照されたい。

1.1　異本を生成するためのデータベース

異本間には類似性があるので、それらのテキストを扱うのに別々のデータベースとして管理するのは非効率的であり、異なる部分のみを明示的に保持するのが望ましい。そこでここでは 1 つのテキストから複数の異本が生成可能な独自の仕様を以下のように定義した。

`\d{[t1]` 対象テキスト 1`[t2]` 対象テキスト 2`@`底本`}`

ある底本に対して、異なる箇所のみを上記のようにマークアップすることで複数異本が 1 つのデータベースとして管理可能である。例えば本文が [t1] は AA,[t2] は BB, 底本では CC のように書かれていた場合には、

```
\d{[t1]AA[t2]BB@CC}
```

のように書く。@以降の部分を取り出すと底本になり、[t1] 以降で {[か@} が現れるまでの部分を取り出すと t1 となる。

1.2　n-gram 分析

文章を分析する上で基本となるのは n-gram である。n-gram とは、文字あるいは単語の連鎖を確率で表すものである。n-gram 分析を単語で行うためには、あらかじめ次に述べる形態素解析により文を形態素列に分割しておく必要がある。例えば「日本」の後にくる単語を大量のデータベースを使って調査すると、「大学」が続く確率は、「高校」が続く確率よりも高いことが予想される。大量の文章からこの確率を学習すれば日本語の用いられかたが明らかとなり、次にくる単語の予測を行うモデルとして文字認識や音声認識など音声

表 1.1　3-gram の一例 (『和泉式部日記』(三条西家本))

確率値	3-gram
0.222	こと も あら
0.173	こと も あり
0.158	こと も ある
0.148	こと も いかに
0.139	こと も いで
0.134	こと も かろがろしう
0.146	こと も きき
0.176	こと も きこえ
0.134	こと も たまさか
0.143	こと も のたまは
0.143	こと も め

表 1.2　3-gram の一例 (『和泉式部日記』(応永本))

確率値	3-gram
0.173	こと も あれ
0.153	こと も いかに
0.153	こと も いで
0.153	こと も かかる
0.148	こと も かろがろしき
0.152	こと も きき
0.227	こと も なし
0.186	こと も のたまはせ

言語処理には不可欠の技術である。これとは違ってある対象となる文章から n-gram の確率を学習することで、その文章の癖を学習することもできる。

例えば、「こと-も-*」という 3-gram の単語連鎖 (アスタリスクは任意の単語を表す) を『和泉式部日記』(三条西家本) で学習した場合の確率値を表 1.1 に示す。「こと-も-あら」が他の 3-gram に比べて出現頻度が高いことが分かる。

同じく『和泉式部日記』の応永本で学習した場合の確率値を、表 1.2 に示す。このように同じ作品でありながら 3-gram の傾向は幾分異なることが分かる。1-gram では両者にそれほど差は出ないが、3-gram になると出現頻度が低く空間がスパースであるため両者を区別することができる。

これに対して、『更級日記』(定家本) で学習した場合の確率値を表 1.3 に示す。当然のことではあるが、表 1.1 と表 1.2 の差異に比べて、表 1.3 のそれは相当に異なる。

このように両者の傾向は異なっている。この異なり具合をうまく定量的に評価すること

表 1.3 3-gram の一例 (『更級日記』(定家本))

0.180	こと も **あはれ**
0.147	こと も **うち**
0.322	こと も **え**
0.142	こと も **おかしく**
0.205	こと も **なき**
0.243	こと も **なく**
0.135	こと も **わすれ**
0.183	こと も **思**

ができれば、各本間の関係性をとらえることができると考えられる [42–44]。このことから評価データにある n-gram が、学習データに存在しないことが容易に想定できる。出現しない n-gram に関しては、このモデルを用いると確率が 0 になってしまう。しかしながら実際には確率 0 ということはあり得ないのでなんらかの確率値を予測して与える必要がある。特に 3-gram のように高次の n-gram は単語連鎖がスパースになるため、評価データにある n-gram が学習データに存在しない可能性が高い。そこで、それより低位の n-gram を用いて確率値の推定が行われる。例えば「扉-を-開ける」がコーパスにあり、「扉-を-閉める」がコーパスになかったとしても、両者の確率値は同じようであることが予測される。そこで「扉-を-閉める」の 3-gram 連鎖の確率値は、「扉」「を」「閉める」の 1-gram、「扉-を」「を-閉める」の 2-gram のうち存在するものの確率値を平滑化して推定する。これを back-off という。n-gram モデルを作る際には、上述の確率値に加えて、コーパスからあらかじめ back-off 係数を学習しておく必要がある。

n-gram 分析を行う際には、「かな漢字交じり」で行うものと、すべて「ひらがな」で行うものがある。中国の漢字文献の分析の際には、一文字ずつ分かれていて表記の揺れも少ないため、漢字の文字単位での n-gram の分析がよくおこなわれる。日本語の場合は、同一本文中であっても、かな漢字の揺れが発生するため、何を分析したいかにもよるが、かな漢字交じり文を扱うと問題を生じることもある。特に和歌の n-gram 分析ではすべてひらがなに直してから、分析することが多い [3, 42] これは和歌特有の掛詞の問題を考慮するためでもある。掛詞はたとえ表の意味を元に漢字で表記されていたとしても、裏の意味で用いる場合には異なる漢字をあてなければならないため、漢字かな交じりでは分析が困難である。ただし古典語には濁点の概念がないために、清音と濁音を区別することはできない。文字単位の n-gram 分析器としては、morogram [45] が有用であり、本研究でもひらがな単位の n-gram の分析には morogram を用いた。このプログラムは文献 [46] のアルゴリズムを実装していて高速であり、頻度 1 やユニグラムの算出ができることが特長である。

一方、かな漢字交じりは、文章の書き手の特性・時代背景を考慮して分析することができるという利点もある。どちらも用途による特性があり、優劣は決め難いために本研究では併用して分析を実行した。

原文	ゆめよりもはかなき世のなかをなげき								
形態素解析		ゆめ	より	も	はかなき	世	の	なか	を	なげき
	読み	ユメ	ヨリ	モ	ハカナキ	ヨ	ノ	ナカ	オ	ナゲキ
	語彙素	夢	より	も	儚い	世	の	中	を	嘆く
	品詞	名詞	助詞	助詞	形容詞	名詞	助詞	名詞	助詞	動詞

図 1.1 形態素解析の例.

1.3 形態素解析

形態素解析とはごく単純に言えば、文章を品詞分解して単語列に分割する技術であり、n-gram 分析と合わせて自然言語処理の必須の技術である。英語などの分かち書きが普通の言語では、特にこの処理は不要であるが、日本語などの膠着語では n-gram 分析の前処理としてこの形態素解析が必要となる[1]。上述の n-gram 分析に品詞の情報を加えることで、「東京 (名詞)+行き (名詞)」と「行き (動詞)+ます (助動詞)」の「行き」を区別して扱うことができる。

1.4 形態素解析の問題点

1.4.1 粒度

形態素解析は、字面の文字単位で行うのが普通である。本来形態素解析の目的から考えれば、なるべく少ない構成要素で (還元的に) 形態素解析を行うのが望ましいと考えられる。しかし本当に文字単位でよいのだろうか。拍単位で行えば、より細かい単位で分析することができる。すなわち文章の原子は何なのかという問題である。文献 [47] においても言語の最小単位を何にするかの問題が扱われている。古典語を分析の対象とする場合には、さらに難しい問題を孕むことになる。

例えば以下のように、茶まめでは文字単位では分析が難しい例が見られた。「我身」を茶まめに掛けると、「我 (ワレ:代名詞)+身 (ミ:名詞)」のように誤った形態素解析結果が得られる。これは中古 UniDic が「わが」を連体詞と考えていないためである。確かに、校訂されて「我が身」となっていれば、「我 (ワレ:代名詞)+が (ガ:助詞-格助詞)+身 (ミ:名詞)」のように読みは誤っているものの、品詞上は正しい形態素解析結果となる。「我身」は一語の名詞として扱うこともできる (実際『旺文社古語辞典』では一語の名詞としている)。ところが「我身の上」は「我身+の+上」ではないので、新しい名詞として「我身の上」とすると使用頻度の低い名詞が増えてしまうという問題がある。「我」を「連体詞」とすれば、「我 (ワガ:連体詞)+身 (ミ:名詞)」「我 (ワガ:連体詞)+身 (ミ:名詞)+の+上」のように正しく形態素解析できるが、「我が身」に対しても「我が (ワガ:連体詞)+身 (ミ:名詞)」と解釈しなければ一貫性が失われる。この問題は、一文字に一形態素を割り当てる現在の形態素解

[1] 英語でも品詞の推定の際には同様の問題が起こりうる。

析の限界を表している。例えば、読みに直した文字列「わがみ」に対して形態素解析を行えば、「わが」を連体詞としなくても、上述の正しい結果が得られる。いずれにしても「なでふ」のような熟したものに関してはこれ以上細かくわけることは不可能であるので、本論では「わが」を連体詞としておいた。

校訂済み本文であれば、このような点を考慮して送り仮名を決めているのであまり問題にならないが、これは送り仮名の揺れが大きい古典語のオリジナルテキストを解析する際には大きな問題となる。いずれにしても、最小単位を文字としている以上は、多かれ少なかれ還元主義の限界を露呈することになる。

1.4.2 連濁

つぎに、連濁の問題がある。「木の葉」であれば「木 (コ:名詞)+の (ノ:格助詞)+葉 (ハ:名詞)」とするのは何の問題もないと考えられる。しかし「紅葉葉」の 3 文字目の「葉」は「バ」と読まれるがこれを「葉 (バ:名詞)」あるいは「葉 (バ:接尾辞)」と登録するのがよいか、「紅葉葉」を一単語として登録するのが良いかは問題である。「葉 (バ:名詞)」として登録するのは、単独で「葉 (バ)」と読まれることは無いので抵抗がある。「葉 (バ:接尾辞)」とした場合には「落ち葉」も「落ち (オチ:動詞)+葉 (バ:接尾辞)」としなければならないのだろうか。本論では連濁の起こっているものは一つの単語として扱うことにした。(「紅葉葉」は一つの名詞とした。)

1.4.3 掛詞

また、掛詞の問題もある。和歌に関しては、表で解釈するか、裏で解釈するかが問題である。「みるめ」と書いてあったときに「見る目」とするか「海松布」とするかという問題である。二通りで解釈しておくという方法もあるが、いつでも二通りの解釈が可能なわけでもない。「あふみち」で「逢ふ (あふ)+道 (みち)」と「近江路 (あふみぢ)」のように濁点の違いで表記不能なものもある。本論では表の意味を主体にとっておくこととし、濁点の有無で意味が変わる場合には濁点をつけない方の意味を優先した。

1.4.4 一貫性

中古 Unidic では、「して (接続詞)」を「す (動詞)+て (接続助詞)」とするなど、還元主義的な部分も見られる一方で、「動詞+す・さす」で表される使役動詞は別に項を立てるなど、あまり一貫した基準であるとは言えない。どの粒度で分析するかに関しては一貫性が必要であるので、この点に関しては改善が必要であるが、ここは修正しなかった。さらに問題なのは、例えば「宣ふ」と「宣はす」で「のたまわ (ノタマウ:動詞)+せ (ス:助動詞)+ず (ズ:助動詞)」と「のたまはせ (ノタマワス:動詞)+ず (ズ:助動詞)」と「は」と「わ」を替えただけで異なる分析結果が出てくることである。これは前者が主に近世のコーパスから学習したもので、後者が中古のコーパスから学習したものであるためと考えられる。翻刻では後者の方で統一されているが、実際の原本では両方の表記があり得るので、これに関

しては統一しておいた。このように一貫性のある一つに解釈を定めるのがとても難しい。また「ものから」のように「もの＋から」の結合で品詞が変化 (名詞から接続詞) するものもある。元の意味を失っていると考えられる品詞変化に関しては、変化後の品詞を使うことで対応した。

1.4.5　複合名詞・動詞

複合名詞・動詞は複合することで元の名詞・動詞とは意味が異なり、一語として考えるか二語としてとらえるかは議論がある [48, 49]。複合動詞をどの程度認めるかは大きな問題であり、文体と品詞構成比率を整理した文献 [50] でも、複合動詞を認めるかどうかで指標に大きな差がでることを示している。例えば、「世の中」は、文脈によっては「世+の+中」(世間) ではなく「世の中」(男女の仲) として解釈すべきである。同様に「見知る」は「見る+知る」でもよいかもしれないが、「思ひ立つ」(決意する) は「思ふ+立つ」(考えて出発する) ではない。

ただし、複合動詞中に係助詞が挿入されることがあることはよく知られており、「思ひ立つ」を一語とした場合には、「思ひも立たず」の解釈はどうなるのかという問題がある。「おぼし立つ」と「思ふ」の部分が尊敬語化したときにこれを別の動詞とするかという問題もある。「思ひ」を名詞と考える見方もあるが、「名詞+動詞」の場合は「波立つ」のように連濁が起こるのに「動詞+動詞」の場合は連濁が起こらないという問題も指摘されている。本論では、「思ひ立つ」は一語として扱ったが、「思ひも立たず」「おぼし立つ」は複合語として扱った。この部分に関してはより詳細な検討が必要となろう。複合語では現代語に対しても盛んに研究が続けられている [51–53]。文献 [54] では、前項動詞が自由な統語的複合動詞と語彙的複合動詞に複合動詞を分類しているが、古典語においてもこのような区別が役に立つかもしれない。

1.4.6　表記の揺れ

古典語では、送り仮名の省略が非常に多い。例えば、「思ふ」は「思」と書かれ「思ふ」「思ひ」「思へ」など活用語尾は省かれることが非常に多いため、辞書に、「思」に「おもふ」「おもひ」「おもへ」など複数の読みを持たせておく必要がある。「宣う」も「のたまふ」「のたまう」「の給ふ」「の給う」「の給」など様々な表記があり得る。「お」と「を」の揺れも多い。「おうな (老女)」と「をうな (女性)」のように意味の違いを意図して書き分けられているものもあるので、一概にまとめて扱うことはできない形態素解析の学習の際には、このような表記の揺れを考慮して学習する必要がある。今回は一つずつ人手で修正することで対応した。

このように古典語の分析は表記が多様性に富んでいたり、一つの語に複数の意味を担わせていたりするため、現代語よりも格段に問題は複雑である。

第2章　計量分析指標

本研究では1章に記した分析手法により、様々な分析を行う。その際に、文章を文字の連鎖とみなせば、文字上の分析指標を用いることができる (2.1節)。形態素解析を行った上で、それに基づく指標には文体の分析指標 (2.2節) と頻度統計の指標 (2.4節) を用いることができる。

2.1　文字上の分析指標

2.1.1　漢字率

新日本古典文学大系等の校訂済み本文は漢字が現代的な基準で見て適当になるように校訂されているが、中古の本文はそれに比べるとかなが圧倒的に多い。今回なるべく忠実な本文を使うことを意図しているので、漢字率も指標として用いることができる。また異本が存在する場合、漢字率は元の本文の影響を少なからず受けると思われるので、それらの異本間での漢字の使用率を算出することで、当該本文を特徴づける量とすることができる。[55] でも漢字の含有率を指標としており[1]、[56] では用字法に書写意識が現れるとしている。

2.1.2　文字の相違率 (Levenshtein 距離による計量)

文字の相違率を判断するために Levenshtein 距離 (編集距離) を用いた。置換 (substitution)、挿入 (insertion)、削除 (deletion) の3つの手順により、任意の文字列の間で変換が可能であるが、Levenshtein 距離はそのような変換を可能にする手順のうちの最小回数として与えられる。これはある文字列を他の文字列に変換するのにかかるコストを距離として用いたものである。Levenshtein 距離は

(1) 動的計画法に基づくアルゴリズムで高速に計算できる。

(2) コストを自分で決めることができるため、その際誤りやすい文字間のペナルティーを考慮することができる [57]。(本研究では、「ん」と「む」に本質的な違いを認めず、それらの間の距離は0とした。)

[1] [55] では、「字母」と「改行位置」も特徴量として使っているが、字母は見た目や個人性の影響を大きくうけ、改行位置は紙や字の大きさから定まる物理的制約の影響があるため、漢字率よりはばらつきが大きくなると考え、本研究では採用しなかった。

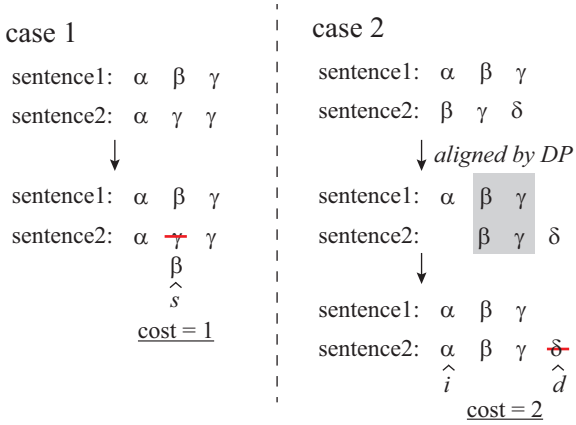

図 2.1　Levenshtein 距離の計算例.

という特長があるため、検索の分野でよく用いられる [2] 。このようにして算出された距離行列からデンドログラムを算出した。

Levenshtein 距離の計算例を図 2.1 に示す。ここで 4 つの記号 (α,β,γ,δ) があると仮定し、文 2(sentence 2) を文 1(sentence 1) へ変換することを考える。ここで 'case 1' では、中央の β と γ が異なっているだけであるので、これは文 2 の γ を文 1 に置換する (substitute) により実現できる。'case 2' では、一見文 1 と文 2 ではすべての記号列が異なっているように見えるため、3 回の置換が必要であるかのように考えられるが、β と γ は両者に共通しているので、この部分を整列させることで α の挿入 (insert) と δ の削除 (delete) により、2 回の手続きで文の変換を実現できる。この方が手順が小さく Levenshtein 距離は上述のようにこのような変換を可能にする手続きの中で最小の回数として与えられるので、この場合の Levenshtein 距離は 2 となる。この整列には動的計画法 (dynamic programming: DP) を用いることで高速に計算できる。ただし長い文章全部に対して DP をかけると計算の時間がかかり、またずれてしまうこともあるので、文単位程度であらかじめ簡単に整列させておくことが望ましい。

ただし古典語の文章は表記のゆれが大きいため、この方法にも問題点はある。例えば、「行き給」と「いきたまふ」は全く同じ内容を示しているが、Levenshtein 距離は 4 となる。これに対して「行き給」と「き (来) 給」は意味が反対であるが、Levenshtein 距離は 1 となる。また「き (来) たまふ」と「き (着) たまふ」は Levenshtein 距離は 0 であるが、意味は異なる。このように表記のゆれがあった場合には、Levenshtein 距離が本文の内容の乖離度を代表しない場合が考えられる。

[2] さらに文字の入れ替わりを考慮した Jaro-Winkler distance[58] も提案されており、聖書の異本をとらえるのに適用されている [59]。ただしこれは主に活字に多く起きる現象で、手書きの古典文学の場合には入れ替わりの問題は起こりにくいと考えられる。このような指標は文献 [60] にまとめられている。

2.1.3 文字単位の n-gram

文字列を記号列の連鎖とみなして n-gram を作ることが行われる。これは中国語の分析でよく用いられる手法である [61, 62]。日本語においても和歌の分析にはひらがな単位の n-gram が用いられることがある [63]。これは掛詞を考慮すると形態素境界が一意に定まらないことがあり得るからである。これは形態素解析が不要であるため文字上の分析指標に分類できる。ただし、漢字に複数の読み方がある場合は注意が必要である。

2.2 文体の分析指標

文体を分析に、名詞の比率など抽象化された指標が有効であることはよく知られている[3]。文献 [31] には文体を分析するための指標として、9 つの指標があげられている。しかしながら、接続詞率 (接続詞を持つ文の割合) 等のいくつかの指標[4] は、古典語の分析においてはほとんど意味をなさない。そこで、ここでは古典の分析にも適用可能な以下の 5 つの指標を用いた。

2.2.1 名詞の比率

文章に含まれる名詞の割合が文章の性質を表すことが古くから知られている。文献 [67] には「サマリー的な文章ほど名詞の比率が大きい」ことが、以下のように述べられている。「一般に言語表現において、事件の筋道を綜合して述べようとする場合には、事柄の関係に叙述の重点がおかれ、何が、何を、何になどを明らかにする骨格的表現となる。そしてこれによって名詞の比率が大きくなり、他の品詞の比率が減少することが見られる。」名詞率は式 (2.2) に示すように名詞率を式 (2.1) で求められる自立語数で除して求める。

$$\text{自立語数} = \text{全単語} - \text{助詞数} - \text{助動詞数} \tag{2.1}$$

$$\text{名詞率} = \frac{\text{名詞数}}{\text{自立語数}} \times 100[\%] \tag{2.2}$$

2.2.2 Modifier Verb Ratio(MVR)

MVR は式 (2.3) により求められる。MVR とは、「形容詞・形容動詞・副詞・連体詞」(Modifier) の合計数を「動詞」(Verb) で除した比率 (Ratio) を表す。この指標は値が高いほど「ありさま描写的」、低いほど「動き描写的」である [68] といわれる。

$$MVR = \frac{\text{形容詞・形容動詞+副詞+連体詞}}{\text{動詞数}} \times 100[\%] \tag{2.3}$$

[3] 文献 [64, 65] には近現代の文学作品を対象に文体の分析指標を用いて因子分析を行った例が載せられている。

[4] 具体的には「字音語の比率、接続詞をもつ文の比率、現在どめの文の比率、色彩語の比率 (‰)、表情語の比率 (‰)」は古典語の分析には適していない指標であると考えられる。現代語の分析においては接続詞は論理展開を示す重要な指標になりうる [66] が、古典語においてはほとんど使われないためである。

2.2.3　指示詞の比率

文中に含まれる指示詞の割合を式 (2.4) により求める。指示詞は適切に使われていれば、文章の冗長性を減らし、読みやすくすることに貢献するが、使いすぎは文章の文脈依存性を高め、理解を難しくする。

$$\text{指示詞率} = \frac{\text{指示詞数}}{\text{自立語数}} \times 100[\%] \tag{2.4}$$

2.2.4　文の長さ (自立語数)

文の長さを式 (2.5) により求める。古典語の文章は現代語の文章に比べて、一文の長さが長いことが特徴である。わかりやすい文章を書くためには短い文で簡潔な文章を書くことが求められる。長い文が一概に読みにくいとは言えないが、長い文がわかりにくいことが多いのもまた事実である。近現代の日本語の文章の文長を調べた研究には、[69, 70] がある。

$$\text{文長} = \frac{\text{自立語数}}{\text{全文数}} \text{ [語/文]} \tag{2.5}$$

2.2.5　引用文の比率

古典文学にこの指標を厳密に適用することは、引用文をどこまでにするかにかかるので難しい問題を孕んでいるが、ここではそれほど厳密に考えず、和歌、会話、心情表現に該当する箇所を引用部分として特定している。近現代文の文章に対してはカッコで囲まれている部分の文字数を数えるが、古典にはカッコは付されていないため、引用会話部分であると思われる部分を判断してカッコを付して引用率を算出した。(多くの場合、「と・など」等で受けている箇所を引用部としている。) 全体の文章に占める引用や会話部分の割合を式 (2.6) により求める。

また心情表現に関しても、直接表現と間接表現の 2 通りが考えられ[5] 、どちらを使うかに作者の特徴が現れると考えられる。ここでは前者は心情表現を直接的に表している箇所であるとして、式 (2.7) により求めた。

$$\text{引用率} = \frac{\text{引用・会話文字数}}{\text{全文字数}} \times 100[\%] \tag{2.6}$$

$$\text{心情率} = \frac{\text{心情表現文字数}}{\text{全文字数}} \times 100[\%] \tag{2.7}$$

2.2.6　各種品詞の比率

これらの指標に加えて、代名詞、形容詞、形状詞、副詞、動詞の比率を指標に加えた。

[5] 直接表現の例としては、「「あさまし」とおぼゆ」が、間接表現の例としては「あさましうおぼゆ」があげられる

2.2.7 語種の比率

語種は和・漢・外来・混種の 4 つがあるが、今回の分析では外来語はないので和・漢・混種の 3 種類に関してその出現頻度を比較した。上記、漢字率と文体の分析指標、各種品詞の比率、語種の比率の計 16 種類の指標を用いて文体の分析を行った。

2.3 n-gram 分析の類似性 (Perplexity の利用)

n-gram 分析の類似性を指標として使うことができる [10]。ユニグラムの分析結果は汎用的に異なるテキスト間で比較可能である。学習するテキストを一つ選び、形態素解析済みテキストに対してバイ/トライグラムモデルを作成し [6] 、それ以外を評価テキストとして式 (2.8) で表される perplexityPP を評価した。

$$PP = \frac{1}{P(w_1, \ldots, w_n)^{\frac{1}{N}}} \tag{2.8}$$

ここで $P()$ は単語列 $w_1, \ldots, w_N$ が観測される確率で、PP はその相乗平均の逆数である。perplexity は次にくる単語が等確率と考えたときの予測される単語数の平均を表し、平均的な現代日本語の文章では数百程度の perplexity になることが多い。容易に n-gram モデルで予想可能なテキストに対しては perplexity は低くなることから、テキストの類似性が定量的に評価できると考えられる。

Perplexity の概念を図 2.2 を用いて説明する。Levenshtein 距離のときの説明と同様に、記号は α,β,γ,δ の 4 つとする。perplexity の計算には評価テキストに対応する n-gram モデルが必要である。評価データにおいて、α の次に β が来るときの perplexity を計算する。'case 1' は、モデルを作るための学習データにおいて、α の後に β,γ,δ が等確率で現れた場合とする。この時それぞれの連鎖の確率は 1/3(=0.33..) である。perplexity は式 2.8 に示すように、連鎖確率の逆数であるので、perplexity(=PP) は 3 となる。'case 2' は、モデルを作るための学習データにおいて、α の後に β,γ が等確率で現れた場合とする。この時それぞれの連鎖の確率は 1/2(=0.5) であり、perplexity は 2 となる。学習データにおいて α-β の連鎖は 'case 1' よりも 'case 2' の方が起こる確率が高かったため、perplexity は小さくなったことが分かる。これは 'case 2' の方が、n-gram モデルを生成モデルと考えた場合、α の後に β が来やすいことを示しており、直観とも一致する。'case 3' は 'case 2' と同様、学習データには α-β,α-γ の連鎖しか見られなかった場合であるが、α-β の方が α-γ の 3 倍起こりやすかったとする。その場合、確率は図に示した通りになり、α-β の perplexity は 3/4 の逆数の 4/3(=1.33..) となる。これは 'case 2' に比べてもより α-β の連鎖が起こりやすいという直観と一致している。

2.4 頻度統計の分析指標

頻度統計が文章の分類に有効であることはよく知られている。単語間の頻度統計を用いて語彙・文章の類似性を判定する試み [18, 73] が行われている。ここでも頻度統計を用い

[6] n-gram モデルを学習するフリーのツールとして palmkit[71] や srilm[72] がある。本論文では srilm を用いた。

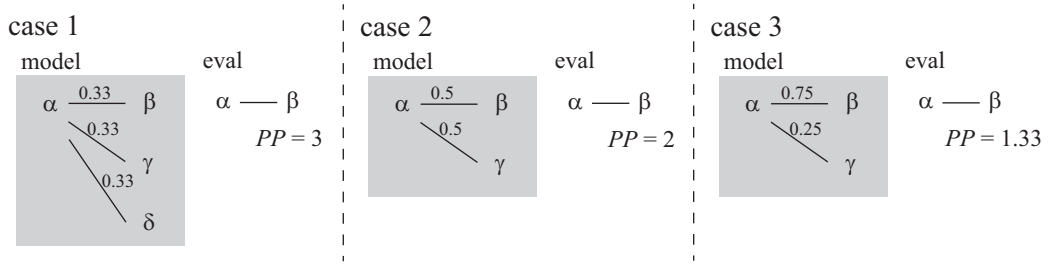

図 2.2　Perplexity の計算例.

て、語の使われ方等を分析する。ここでは語の出現順を無視する Bag-of-words の手法を用いて検討した。これによって n-gram モデルよりも柔軟に語と語の間の弱い共起関係を測ることができる。例えば、和歌における特定の単語同士の共起関係を検討することで、時代ごとに語と語の共起関係が異なることが示されている [34]。離れた場所にある単語同士の共起関係を探る場合には n-gram モデルよりも Bag-of-words の手法の方が有効である。

2.4.1　cosine 類似度

Bag of words は、各単語ごとの頻度ベクトル $\mathbf{h}_i = (w_1, w_2, \ldots, w_N)$ を各本ごとに求める。ここで i は各本のインデックス $(1 \leq i \leq I)$ であり、w_n は各単語の頻度である。ただし活用語はすべて原型を用いて集計することにする。N は対象 I テキストすべてに現れる単語の上限であり、当該テキストに見られない単語の頻度は 0 とすることにした。そして、このようなベクトル間の類似度を測るのには、cosine 類似度がよくつかわれる。テキスト i と j の間の cosine 類似度 c は内積の公式を用いて

$$c = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{|\mathbf{h}_i||\mathbf{h}_j|} \tag{2.9}$$

によって求められる。$\top$ は転置を表す。ベクトルの Euclid 距離は

$$|\mathbf{h}_i| = \sqrt{\mathbf{h}_i^\top \mathbf{h}_i} = \sqrt{w_1^2 + w_2^2 + \cdots + w_N^2} \tag{2.10}$$

で求められる。

2.4.2　助動詞出現頻度相関

品詞の中でもどのような助動詞を使うかはその文章の特徴を表すとされ、従来より様々な研究がおこなわれており、様々な文献を比較した研究もみられる [74]。ここでも助動詞別に検討を行った。

第3章　『和泉式部日記』4異本間の関係性と『更級日記』との比較

この章では『和泉式部日記』の特性を明らかにするために、同程度の分量である『更級日記』との比較を通じて計量的分析を行う。

3.1　底本について

3.1.1　『和泉式部日記』あるいは『和泉式部物語』

『和泉式部日記』の原本は残念ながら現存していないか見つかっておらず、主に表 3.1 に示す四つの系統に分別されている。

表 3.1　『和泉式部日記』の代表的な 4 異本

「三」	三条西家本系統
「寛」	寛元本系統
「応」	応永本系統
「混」	混成本系統

上述の 4 つの異本のうち、三条西家本系統のみが室町時代の書写であることが知られており、寛元本系統・応永本系統・混成本系統は江戸時代の書写である [1] 。三条西家本の祖本が最も古いと考えられていることもあって、各種翻刻テキスト [76–78] の最も一般的な底本となっている [2] 。ここでは 4 種の異本の代表的なものを底本にして、それらの関係性を探ることにする。ちなみに『和泉式部日記』は三条西家本のみ『和泉式部日記』と題がつけられているが、そのほかの系統の本には『和泉式部物語』とあり、江戸時代の目録等を見てもこの題の方がよく知られていたようである。

文献 [80] は、三条西家本の本文に他本の異同を対校しながら翻刻したものである。本研究では、この文献をもとに 1.1 節にあげた形式のデータベースを作成した [3] 。底本の電子

[1] 書誌情報などは [75] を参照されたい。各写本間の関係性は [75](p. 170 の図) にまとめられている。
[2] 例えば [79] では、異同箇所の比較から「三条西本を善本と考え」ている。
[3] この文献では、

(1) 宮内庁書陵部蔵 伝**三条西**実隆筆本

(2) 吉田幸一氏蔵 **寛元**奥書本

データは [81] にあるものを用いた [4] 。なお、このデータには、誤りが散見されたので参考文献 [76, 77, 80, 82] により修正した。写本・版本のテキスト分析が自動で認識できるようになれば [5] 、これらの作業もより簡単に網羅的に行えるようになると考えられる。その後、濁点 [6] 、句読点を付与し、会話文の箇所と心情を直接的に表している箇所 (主に「と」「など」で受けている部分) を特定した [7] 。その際には、[76, 77] 等、先学の解釈を参考にした。

例えば『和泉式部日記』のはじめの部分が、このデータベースでどのように記述されるか説明する。この部分を対校すると、

(1) ゆめよりもはかなき世のなかをなげきわびつつあかしくらすほどに、(三)

(2) ゆめよりもはかなきよのなかをなげきわびつつあかしくらすほどにはかなくて、(寛)

(3) 夢よりもはかなき世の中をなげきつつあかしくらすほどにはかなくて、(応, 混)

のようになる。これを

```
\d{[応混] 夢@ゆめ}よりもはかなき\d{[寛] よの [応混] 世@世の}\d{[寛応混] 中@なか}を
なげき\d{[応混]@わび}つゝあかしくらすほどに\d{[寛応混] はかなくて@}、
```

のように表す。異なる箇所を \d{} でくくった範囲内に書き記す。

3.1.2 『更級日記』

『更級日記』にはいくつかの異本が知られているが、近代に入って諸本に共通の錯簡が、御物本の補修の際の綴じ間違いにより生じたことが明らかになり [93]、諸本は御物本を共通の祖とすることが明らかになった。御物本は定家晩年の書とされる。本報では底本には最も一般的な「御物 (定家) 本」を用いた。電子データは [94, 95] を参考にし、文献 [76, 96–98] を参照して適切な修正を加えた。

---

(3) 京都大学蔵 **応永**廿一年奥書本

(4) 群書類従所収 流布本 (**混成**本)

を元にしている。

[4] 該当ページの中には、文献 [76] を参照したとの表示があった。

[5] 写本、古文書、版本を対象にいくつかの研究例 [83–88] が報告されている。中古の写本は、連綿がよくおこるので、文字境界のゆれにロバストで手書き文字認識に用いられる隠れマルコフモデルによるアプローチ [89, 90] が、アラビア語のように続け字を基本とする言語の文字認識 [91] に使われていることから有効であると考えられる。

[6] 濁点を付与して解析するかしないかは、前述のように特に和歌の掛詞において問題になる。一意に濁点を付与することができない場合が存在するからである。しかしながら濁点を付与しないと検索性が低下し、形態素解析を正しく行うことができないという問題があるので、ここでは濁点を付与したテキストを対象に検討した。この妥当性に関しては別途検討されるべきであろう。なお濁点の付与にはかなりの手間がかかる。また踊り字の場合に繰り返し箇所が清音か濁音かを判断する必要があるのは近代以前の文章を解析する場合に特有の問題であろう。濁点の自動付与に関する研究も行われている [92]。

[7] 同様の作業を『更級日記』に対しても行った。

3.1.3 形態素解析

品詞のタグ付けは、中古語の形態素解析辞書「中古 UniDic」 [99–101] を “MeCab” [28] と組み合わせた形態素解析エンジン「和文茶まめ」[102] より行った。こちらはそれぞれの異本の本文 (漢字かなまじり) に対して行うことで比較した。ただし形態素解析の誤り (全体の 5%程度) や以下のようにいくつかの問題があったので、人手で修正を加えた。また古典特有の問題として、繰り返し記号の多用があげられる。繰り返し記号 (ゝ等) に関しては、元の字の連続に直してから形態素解析を行った。

3.2 既往研究における『和泉式部日記』の系統づけ

『和泉式部日記』研究のレビューは [103] と [104] に詳しい。国文学・文献学の立場からの検討の古いものとしては [105, 106] がある。

3.2.1 『和泉式部日記』の系統論

文献 [107] により、寛元本系統の本が紹介され、今日の三系統を基にした理論が構築された。同文献では「新出本[8] を基にして、それより応永・三条西両本が出たものと想定することが出来る」と結論付けている。

文献 [106] では、文献を詳細かつ計量的に扱い、校異数の比較により、混成本は「応永本を基にして寛元本の要素をとりいれた」応永本は「寛元本系統に属しながら三条西家本の要素を取り入れて成立した末流本」「三条西家本と寛元本とは別種の系統をなして対等の地位に立ち存在している」という結論を得ている。またこれを補強する形で、寛元本の誤写箇所を詳細に検討し、「寛元本は三条西本に比較して、誤写・誤脱・衍の数がかなり多」く [38, 108]、「寛元本は三条西本、応永本の中間的性格を示す」と述べられている。ただし「寛元本から三条西本と応永本に分岐したということではない」。

これに対して、[109] では、校異数よりも質を重視し、共通誤脱の分析を行い、「三条西家本と寛元本系統が、応永本とは異なる共通の祖本から出た」との別の結論を得ている。[75] もこの説を支持し、「脱文・数詞・官職名等の類について、異同関係を考察」され、特に「数詞の誤写」に注目して、「応永本系は、三条西・寛元両本系とは縁故関係も薄く、遠い」、「応永本は『三条西家本と寛元本とは別種の系統をなして対等の地位に立ち存在している』ことは認められるけれども、応永本系統が『寛元本系統に属しながら三条西家本の要素を取り入れて成立した末流本』といふことにはなりにくいやうである。これはやはり『三条西本と寛元本系統が、応永本系統とは異る共通の祖本 (B 本) から出た』とみる方が蓋然性がある」と結論付けている。

現在は、三条西本が最も善本と考えられているが、[79]、「三条西家本には添削意識や合理化の跡が見出される」[109] との指摘も重要であると考えられる。

文献 [110] では、諸本間に総語彙数に差は少ないことから、非共通語彙の検討を行い、その語彙的関係性を考察している。全体としては、

[8] 著者注:寛元本

(1) 「寛元」と「三条西」は相違が少ない

(2) 「寛元」と「応永」はかなり相違している

(3) 「三条西」と「応永」はかなり相違しているが、「寛元」と「応永」のほうがより遠い

と結論付けている。ところが、「名詞」に関しては、

(1) 「三条西」は「寛元」よりも「応永」に近い

(2) 「三条西」・「応永」はともに「寛元」に遠い

となり、さらに「動詞」に関しては

(1) 「応永」は「寛元」・「三条西」に遠い

(2) 「三条西」は「寛元」よりさらに「応永」に遠い

となる。このように分析する対象によって三様の結論を得ている。これは「非共通語彙のほとんどが1回の使用度数」であり、安定した統計量にならなかったためと考えられる。非共通語彙の発生率は動詞が最も高く、「動詞が最も異同の生じやすい品詞」であると結論付けている。中でも複合動詞に異同が生じやすいことが分かっている。この研究は本文全体を通した統計量を使う必要性を示唆していると考えられる。

3.2.2 『和泉式部日記』の文体

『和泉式部日記』は主人公の女による三人称語りである点 [111] が、通常の日記文学と異なっており、その特異性が注目を集めてきた。実際、「歌物語というにふさわしい作品である」と指摘する説 [112] もある。また著者に関しても様々な議論があり、[113] では、『和泉式部日記』を自作と認めず、『伊勢物語』、『平中物語』、『篁物語』、『多武峯少将物語』と同一の歌物語の系列にある作品であるとした。また俊成の作であると考える説もあり [107]、物議をかもしたが [114]、現在では否定的に考えられている。

文献 [115] では、過去のことを語る形式である文末の語「けり」とその活用形である「ける」「けれ」を「歌物語の文体の特色」とし、『和泉式部日記』ではこれらの使用は歌物語と比べて他の日記文学と同程度に少ないことを示している。また「主観的心情表現、自己告白的表現が随所にあること」の二点を「日記文学の文体の特徴」とし、このことから『和泉式部日記』が「日記文学の文体を持っている」と結論付けている。

[116] では、「けり」の使用が少ないことに加えて、文末の「なむ」の使用が少ないこと (1、2例 (テキストにより異なる) 見える程度) を挙げ、「『伊勢物語』など歌物語多出の「なむ」が『源氏物語』で減ってゆき、語り調子「なむ」や「けり」を多出せず、話し言葉で述べてゆく様式になる」とし、「日記も土佐→蜻蛉→和泉というように同様の経過をたどっている」と述べている。このように『和泉式部日記』の文体の特異性を時代変化に求める説もある。

3.3 結果と考察

3.3.1 文体の分析指標

1 で述べた 16 指標に従い、分析を行った。結果を表 3.2 に示す。ここには合わせて「総文字数」と「総単語数」も示している。

このように本文の規模は、『更級日記』の方が 33-35%大きい程度である。『和泉式部日記』の 4 つの異本間で差異が出ているものとしては、漢字率があげられる。これに対して、『和泉式部日記』と『更級日記』の他作品間の差異を表すものは、引用率・心情率・名詞率があげられる。三条西家本の指標で他本の指標を割った (正規化した) 結果を図 3.1 に示す。これによりこの傾向がよくわかる。『和泉式部日記』は和歌の引用が非常に多く、三人称語りでありながら「女」の心情表現が豊かであることが特徴であるので、それが引用率および心情率の高さに表れているものと思われる。『更級日記』は、名詞率、漢語、固有語の比率が『和泉式部日記』より高い。これは『更級日記』が、事実叙述的であるところからきていると思われる。本来日記は記録的な色彩が強いために事実叙述的なのは当然であるが、『和泉式部日記』の場合は明らかに事実の記録よりも心情の吐露を主眼としている。

異本間では漢字率に大きな差異が現れたのは、写した人の性別・年代などが影響しているのではないかと思われる。MVR には差異がみられなかった。これは同ジャンルであることが原因であると考えられる。

指標が 16 個あり、それぞれの関係がわかりにくいので、主成分分析により主要な変数 2 つを取り出した[9] 。結果を図 3.2 に示す。明らかに、『和泉式部日記』内の異本のばらつきは、それらと『更級日記』との差に比べて著しく多く、作品間の分析をするのには、これらの指標が有効であるといえる。ただし、異本間の差異を分析するほどには、この指標の精度が高くはない可能性がある。例えば、これによると、混成は寛元に近いことになるが、

表 3.2 分析結果

	総文字数	漢字率	平均文長	引用文字率	心情文字率	総形態素数	自立語率	MVR
三条西	20025	7.2%	52.4	47.5%	12.2%	10810	52.6%	40.5%
寛元	19975	8.3%	54.0	46.6%	12.3%	10906	52.4%	39.4%
応永	19840	8.6%	53.3	47.4%	12.4%	10865	52.5%	41.5%
混成	20200	10.7%	54.4	46.4%	12.8%	11186	52.3%	41.6%
更級日記	26546	9.1%	66.7	33.2%	4.1%	14517	55.7%	40.3%

	名詞率	代名詞率	形容詞率	形状詞率	副詞率	動詞率	和語	漢語	固有語	混成語
三条西	37.1%	3.4%	8.0%	1.2%	7.4%	40.9%	98.3%	1.3%	1.0%	0.3%
寛元	37.3%	3.3%	7.9%	1.1%	7.3%	41.1%	98.3%	1.3%	1.0%	0.3%
応永	36.6%	3.2%	7.9%	1.2%	7.8%	40.8%	98.2%	1.4%	1.0%	0.3%
混成	36.7%	3.1%	7.8%	1.1%	8.0%	40.7%	98.1%	1.5%	1.0%	0.3%
更級日記	46.6%	3.6%	7.9%	1.1%	5.1%	35.1%	96.3%	2.2%	1.3%	0.3%

[9] 寄与率は 2 つの変数で 100%であるので、それ以下の変数は無視できることを確認している。

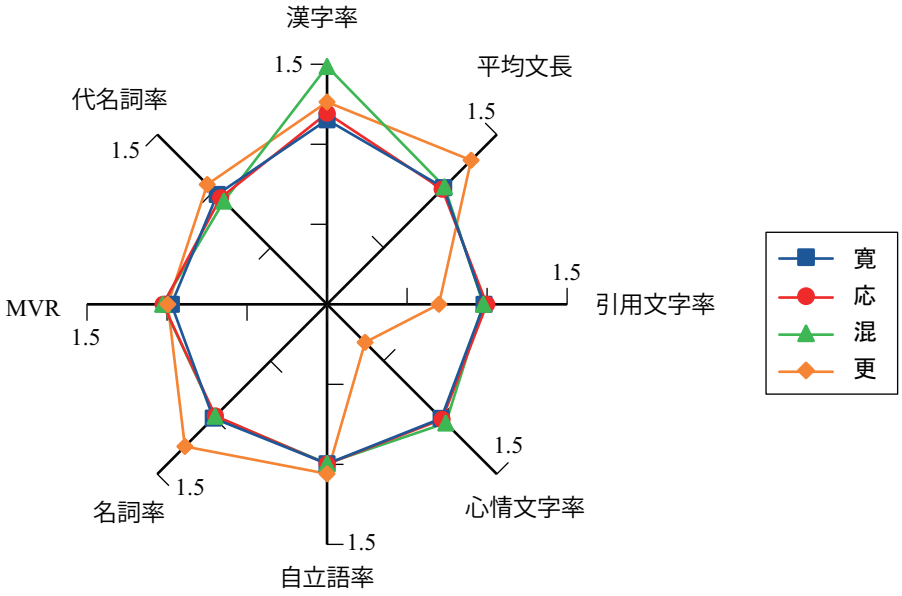

図 3.1　分析指標のレーダーチャート.

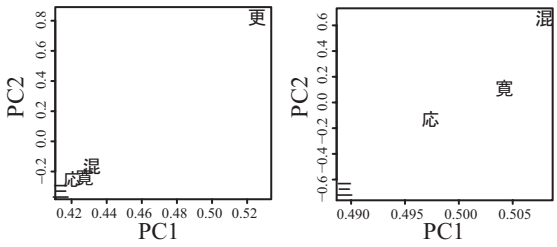

図 3.2　分析指標の主成分分析結果.(主成分 1(PC1) と主成分 2(PC2) を 2 次元平面上にプロット.)

文献学的な観点 (3.2.1 参照) からは混成は応永に近いことが分かっているので、この分析は妥当なものではないと考えられる。

また、各指標間でダイナミックレンジに差があるため、どの本同士がどれほど近いかを定量的に測ることは難しい。例えば、主成分分析した平面上での Euclid 距離は意味を持たない。このように主成分分析には限界がある。

3.3.2　Levenshtein 距離による分析

表 3.3 に『和泉式部日記』4 異本間の Levenshtein 距離を示す。距離は対称性を有するため、右三角成分は省略されている。これだけでは関係性が分かりにくいので、図 3.3 のようにデンドログラムで表す手法がよく使われる。これによると、今回分析した異本は 2 つのグループに分けられることが分かる。混成が応永と近いことは、国語学・文献学的な検

表 3.3 『和泉式部日記』4 異本間の Levenshtein 距離

	三条西	寛元	応永	混成
三条西	0	-	-	-
寛元	4916	0	-	-
応永	5480	5412	0	-
混成	5751	5148	3305	0

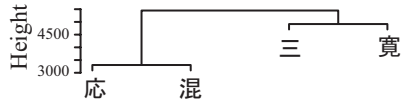

図 3.3 Levenshtein 距離に基づくデンドログラム.

討では一致して述べられているし、[109] や [75] が得ている「三条西家本と寛元本系統が、応永本とは異なる共通の祖本から出た」[109] との結論 (3.2.1 参照) ともこの分析結果は一致するものである。

3.3.3 n-gram の分析

Levenshtein 距離によって異本間の関係性を考察することができるが、この方法は他本 (更級日記) に対しては使えない。また動的計画法も本文全体に対してやると精度が低下するなどの問題があるため、事前にある程度アライメントを整えておく必要があり、それなりに手間がかかる。ひらがなの 1-gram の分析は和歌などには有効だが、本文に関しては全部に読みを付ける手間がかかる割に有意な結果が得られるとは考えにくい。そこで 3-gram を作成しその perplexity を計算する方法による分析を行う。本手法であれば、形態素解析がある程度「正しく」できていれば、3-gram を作るだけでツールを用いて計算できる。表 3.4 に分析結果を、図 3.4 にそのデンドログラムを示す。異本は Levenshtein 距離を用いた場合と同様に分類でき、他作品との比較も行えている。これから異本間のばらつきは他作品に比べて十分小さいことが裏付けられた。

表 3.4 3-gram の perplexity の分析結果 (品詞情報なし)

学習 \ 評価	三条西	寛元	応永	混成	更級
三条西	8.9	35.2	35.6	41.4	198.5
寛元	36.4	9.0	40.0	40.2	205.7
応永	37.2	40.3	9.3	26.7	195.3
混成	44.2	41.1	26.6	9.5	204.1
更級	242.6	241.6	244.3	235.9	9.0

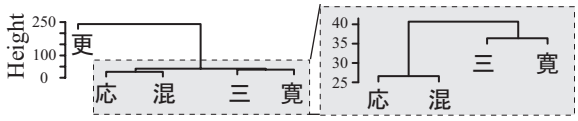

図 3.4 Perplexity に基づくデンドログラム.

表 3.5 和歌中の 1-gram の相関係数 (三条西家本との比較) と出現数

	相関係数	出現数
三条西	1	4490
寛元	0.998	4489
応永	0.998	4451
混成	0.999	4483
和泉式部 (続) 集	0.964	51169
更級日記	0.937	2744

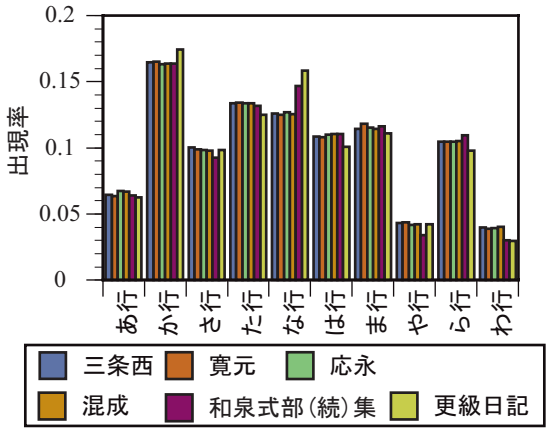

図 3.5 1-gram の各行における出現率の割合.

3.3.4 文字上の n-gram

和歌をひらがなの連鎖とみなして出現頻度を検討した。1-gram に関して検討する。作者によるかなの使用頻度の違いを考慮するために、『和泉式部集』『和泉式部続集』の和歌も比較対象に含めた。相関係数と出現数を比較した結果を表 3.5 に示す。異本間の相関は非常に高かった。これに対して、『和泉式部 (続) 集』との相関が非常に高かったのは類似歌が入っているのが影響していることは当然考えられるが、これに加えて作者による特徴と考えることができる。実際、『更級日記』との相関係数は低い。各行における出現頻度を図 3.5 に示す。「か行」において『更級日記』が、「な行」において『和泉式部 (続) 集』、『更級

表 3.6 異なり形態素数

テキスト	形態素数 (品詞あり)	形態素数 (品詞なし)
三条西	1110	1085
寛元	1139	1109
応永	1132	1105
混成	1169	1105
更級	1801	1764

表 3.7 Bag of words の cosine 類似度 (品詞情報あり).

	三条西	寛元	応永	混成	更級
三条西	1	-	-	-	-
寛元	0.9984	1	0.9985	0.9980	0.9187
応永	0.9977	0.9985	1	0.9984	0.9181
混成	0.9982	0.9980	0.9984	1	0.9137
更級	0.9147	0.9187	0.9181	0.9137	1

日記』の出現率が高くなっている。これは前者が「か」が多く、後者がそれぞれ「な」と「の」が多かったことによる。

3.3.5 頻度統計の分析

総形態素数は表 3.2 に示したが、表 3.6 には異なり形態素数を取り上げる。品詞情報のあり・なしで区別したがそれほど差がなかった。語彙数は現代語から考えられるよりも当然少ないことが分かる。『和泉式部日記』の語彙に関する研究には、[117] がある。『更級日記』の方が事実の記述に重きを置いているため、固有表現が増加し、異なり形態素数が多くなった。

単語の頻度を集計して、式 2.9 により、cosine 類似度を求めた。品詞情報のありなしで指標に差は見られなかったので、品詞情報ありの場合の cosine 類似度を図 3.7 に示す。他作品間では指標に差異が出ているものの、異本間では 0.001 程度の差異しか見られず、非常に高い類似度を示しており、異本を区別する指標として用いるのは難しいことが分かる。このように単語の頻度だけに基づく指標ではなく、perplexity のように単語間の連接関係を考慮することが重要である。

文体の分析を行うために、助動詞の出現数と頻度 [%] を表 3.8 に示した。『更級日記』の品詞別の考察が文献 [74] にある [10] 。「させる、せる、られる、れる」はそれぞれ「さす、す、らる、る」に当たるが、中古 unidic ではこれらの助動詞の原型を現代語とのつながりを考えてか、前者のように扱っているので、ここでは両方を表記した。文献 [115, 116] でも触れられているように、『和泉式部日記』と『更級日記』において「けり」の使用頻度はそれほど高くない。「き」「けむ」の使用頻度は、『更級日記』の方が 2 倍から 5 倍程度高い。これ

[10] 索引 [118] の利用も効果的である。

表 3.8 助動詞ごとの出現数 (絶対数) と総形態素数で除した助動詞ごとの頻度 [%].

	絶対数					百分率				
	三条西	寛元	応永	混成	更級	三条西	寛元	応永	混成	更級
き	60	72	67	71	162	0.56	0.66	0.62	0.63	1.12
けむ	5	6	5	4	30	0.05	0.06	0.05	0.04	0.21
けり	69	75	68	76	85	0.64	0.69	0.63	0.68	0.59
ごとし	2	2	2	2	3	0.02	0.02	0.02	0.02	0.02
させる (さす)	60	46	58	59	8	0.56	0.42	0.53	0.53	0.06
じ	25	24	26	24	11	0.23	0.22	0.24	0.21	0.08
ず	174	172	172	174	191	1.61	1.58	1.58	1.56	1.32
せる (す)	104	95	86	121	32	0.96	0.87	0.79	1.08	0.22
たり-完了	147	153	157	158	248	1.36	1.40	1.45	1.41	1.71
つ	59	59	59	59	31	0.55	0.54	0.54	0.53	0.21
なり-伝聞	16	15	21	17	26	0.15	0.14	0.19	0.15	0.18
なり-断定	262	263	257	271	393	2.42	2.41	2.37	2.42	2.71
ぬ	134	126	132	138	123	1.24	1.16	1.21	1.23	0.85
べし	53	57	57	59	71	0.49	0.52	0.52	0.53	0.49
まし	17	18	16	17	19	0.16	0.17	0.15	0.15	0.13
まじ	6	7	5	5	3	0.06	0.06	0.05	0.04	0.02
まほし	7	6	7	7	6	0.06	0.06	0.06	0.06	0.04
む	138	143	148	153	118	1.28	1.31	1.36	1.37	0.81
むず	1	2	1	1	0	0.01	0.02	0.01	0.01	0.00
めり	26	25	26	27	14	0.24	0.23	0.24	0.24	0.10
らむ	31	31	33	34	13	0.29	0.28	0.30	0.30	0.09
られる (らる)	18	18	18	16	32	0.17	0.17	0.17	0.14	0.22
り	22	24	18	24	31	0.20	0.22	0.17	0.21	0.21
れる (る)	37	44	40	41	55	0.34	0.40	0.37	0.37	0.38
総形態素数	10810	10906	10865	11186	14517	100	100	100	100	100

に対して、「めり」「らむ」の使用頻度は、『和泉式部日記』のほうが 2 倍から 3 倍程度高い。これは『和泉式部日記』があたかも目前で事象がおこっているかの如く生き生きと描かれているのに対し、『更級日記』が過去を振り返る回想的な視点で描かれているところに起因していると考えられる。その他に目立った際としては、「させる (さす)」「せる (す)」の頻度が『和泉式部日記』の方が高いということがあげられる。どちらの作品も内向的ではあるが、『和泉式部日記』は手紙のやりとりなどを通じて他者とかかわる場面が多く描かれていることと関係があるだろう。『更級日記』には、人との交流の場面はあまり登場せず、自分が体験した出来事を淡々と描くという形式であるので、そのスタイルの違いがここに現れていると考えられる。

終章　まとめ

本研究は中古の日記文学の代表格である『和泉式部日記』と『更級日記』を題材に、『和泉式部日記』の 4 つの異本と『更級日記』の関係性を明らかにすることを目的として、計量的な分析を行った。その結果、異本間の差異を表すものとしては「漢字率」が、他本間の差異を表すものは「引用率・心情率・名詞率・代名詞率」が有効である可能性が示された。日記文学の観点としては、名詞率、心情率が重要であると考えられる。記録的な文学では事実や固有名詞のの記述が中心となるため、必然的に名詞の使用率が高まる。これに対して、内情吐露的な文学では、心情率が高くなる。目的 1 に挙げた、日記文学に特質については、ある程度分析できたと考えられる。確かに、これらの指標はジャンル分けや、作品同士の関係性を探るといった大雑把な分析には有用である。しかしながら、指標には恣意性があり、一般化と定量的な分析が難しく、異本の分析のような細かな分析に使うには、問題がある。実際これらの指標の主成分分析結果は先学の検討結果と一致しなかった。

そこで、自然言語処理の分野で用いられている指標である Levenshtein 距離と perplexity を使って、客観的な評価を行った。これにより、異本間の異なり度を測る指標として、計量分析によく用いられている分析指標の主成分分析よりも、文字列同士の Levenshtein 距離や perplexity が有効であることが分かった。特に perplexity を用いることで、同一作品の異本間の差異と異なる作品間の差異を比較できる。これにより同一作品の異本間の差異は、異なる作品間の差異に比べて小さいことが定量的に確かめられた。また Levenshtein 距離と perplexity どちらを用いた場合でも、異本の分類結果は一致した。ここで 4 つの異本に対する系統の分析結果は先学の解釈と一致した。これにより目的 2 に掲げた異本の分析に関しても有効な手法を提起することができたと考えられる。

今後の課題としては、より多くの作品、異本を分析することや、未知の異本の系統付け等があげられる。また『和泉式部日記』の贈答歌のやりとりを定量的に検討することも、考えられる。今回の分析でいくつかの形態素解析に関する問題点を発見した。特に和歌の分析においては多くの課題を有していると考えられる。「歌物語」と「日記文学」の関連に関しても、興味深い。『土佐日記』『伊勢物語』や『篁物語』等の多くの関連する作品を分析し、その境界に関して考察を加えることも必要になってくると考えられる。

[$40 \times 41 = 1640$ 字/枚 $\times 27$ 枚 $=$400 字詰め原稿用紙 110 枚相当]

参考文献

[1] 北研二. (1999). 確率的言語モデル. 東京大学出版会.

[2] S. Bird，E. Klein，E. Loper，萩原正人他 (訳). (2010). 入門 自然言語処理. オライリー・ジャパン.

[3] 近藤みゆき. (2000). n グラム統計処理を用いた文字列分析による日本古典文学の研究：『古今和歌集』の「ことば」の型と性差. 千葉大学人文研究 人文学部紀要. 29: 187–238.

[4] 金明哲. (2000). 自然言語処理における統計手法を用いた情報処理. 統計数理. 48: 271–287.

[5] 村上征勝. (2004). シェイクスピアは誰ですか? –計量文献学の世界–. 文春新書.

[6] A. Kenny. (1982). The computation of style: An introduction to statistics for students of literature and humanities. Pergamon Press.

[7] 工藤拓，山本薫，松本裕治. (2004). Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告/自然言語処理研究会報告. 2004: 89–96.

[8] C. Brinegar. (1963). Mark Twain and the quintus curtius snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association*. 58: 85–96.

[9] R.D. Peng and N.W. Hengartner. (2002). Quantitative analysis of literary styles. *The American Statistician*. 56: 175–185.

[10] O. Uzuner and B. Katz. (2005). A comparative study of language models for book and author recognition. *in Proceedings International Joint Conference on Natural Language Processing (IJCNLP-05)*.

[11] 金明哲，村上征勝. (2007). ランダムフォレスト法による文章の書き手の同定. 統計数理 (特集「文化を科学する」). 55: 255–268.

[12] M. Koppel, J. Schler, and S. Argamon. (2009). Computational methods in authorship attribution. *JASIST*. 60: 9–26.

[13] E. Stamatatos. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*. 60: 538–556.

[14] 田畑智司. (2012). Dickens と Collins の共著作品への文体統計学的アプローチ. 情報処理学会研究報告. CH-93: 1–7.

[15] J. Mingzhe and J. Minghu. (2012). Text clustering on authorship attribution based on the features of punctuations usage. *in Proceedings International Conference on Signal Processing*. 3: 2175–2178.

[16] A. Roque. (2012). Towards a computational approach to literary text analysis. *in Proceedings Workshop on Computational Linguistics for Literature*. 97–104.

[17] I. Raskovsky, D.F. Slezak, C. Diuk, and G. Cecchi. (2010). The emergence of the modern concept of introspection: A quantitative linguistic analysis. *Proceedings NAACL Young Investigators Workshop*.

[18] 深谷亮，山村毅，工藤博章，松本哲也，竹内義則，大西昇. (2004). 単語の頻度統計を用いた文章の類似性の定量化 : 部分的類似性の考慮. 電子情報通信学会論文誌. J87-D-II: 661–672.

[19] 長谷川優，山村毅. (2011). マハラノビス距離を用いた日本語文章の難易度判定. 電子情報通信学会論文誌. J94-D-9: 1589–1592.

[20] 近藤泰弘. (2001). コンピュータによる文学語学研究にできること –古典語の「内省」を求めて–. 全国大学国語国文学会夏季大会シンポジウム. 1–6.

[21] J. Grimmer and B.M. Stewart. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*. 1–31.

[22] 寿岳章子. (1983). 室町時代語の表現. 清文堂.

[23] 金愛蘭. (2011). 外来語動詞「チェックする」の基本語化–20 世紀後半の通時的新聞コーパスを用いて–. 計量国語学会第 55 回大会. 31–36.

[24] 荻野綱男. (2011). goo ブログ検索から見る男女差と年齢差. 計量国語学会. 25–30.

[25] 近藤泰弘，近藤みゆき. (2001). 平安時代古典語古典文学研究のための n-gram を用いた解析手法. 言語処理学会年次大会発表論文集. 7: 209–212.

[26] 漢字文献情報処理研究会 (編). (2012). 電脳中国語学入門. 好文出版.

[27] 上田英代，上田裕一，村上征勝. (1994). 『源氏物語大成』の品詞情報つきフルテキストデータベースの作成について. 情報知識学会誌. 4: 81–93.

[28] Mecab. http://mecab.sourceforge.net/.

[29] 青空文庫. http://www.aozora.gr.jp/.

[30] 形態素解析システム茶筌. http://chasen-legacy.sourceforge.jp/.

[31] 小林千草. (2005). 文章・文体から入る日本語学. 武蔵野書院.

[32] 上阪彩香，村上征勝. (2011). 西鶴作品の文章分析 –先行研究の計量文献学的検証–. 情報処理学会研究報告 人文科学とコンピュータ. 2011-CH-90: 1–7.

[33] 新井皓士. (1997). 源氏物語・宇治十帖の作者問題: 一つの計量言語学的アプローチ. 一橋論叢. 117: 397–413.

[34] 山元啓史. (2010). ブーリアン演算による歌ことばモデルの解析. 第 16 回公開シンポジウム「人文科学とデータベース」. 37–44.

[35] 堀川貴司. (2010). 書誌学入門 古典籍を見る・知る・読む. 勉誠出版.

[36] 池田亀鑑. (1991). 古典学入門. 岩波文庫.

[37] 森田兼吉. (1996). 『和泉式部日記』は三条西家本だけでは読めない : 『和泉式部日記』三系統論再読・続稿. 日本文学研究 (梅光学院大学). 31: 17–28.

[38] 伊藤博. (1978). 和泉式部日記寛元本の誤写箇所について. 大妻女子大学文学部紀要. 10: 67–76.

[39] 丸山直子. (2003). 数理的研究. 国語学. 212: 62–65.

[40] 師茂樹. (2004). 大規模仏教文献群に対する確率統計的分析の試み. 中国宗教文献研究国際シンポジウム. 357–370.

[41] 師茂樹. (2011). 異なる文献間の数理的な比較研究をふり返る. 文字と非文字のアーカイブズ／モデルを使った文献研究. 31–38.

[42] 近藤泰弘，近藤みゆき. (2001). N-gram の手法による言語テキストの分析方法. 漢字文献情報処理研究 第 2 号 特集 2(N-gram が開く世界). 50–55.

[43] 谷本玲大. (2001). 曖昧検索性を持たせた n-gram サーチの手法–『新撰萬葉集』と菅原道真の詩の比較を例に–. 漢字文献情報処理研究 第 2 号 特集 2(N-gram が開く世界). 56–58.

[44] 土山玄，村上征勝. (2012). 語の bigram による『源氏物語』の分類. 人文科学とコンピューターシンポジウム (じんもんこん 2012). 49–54.

[45] morogram. http://morogram.sourceforge.jp/.

[46] M. Nagao and S. Mori. (1994). A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. *In Proceedings of the 15th International Conference on Computational Linguistics*. 611–615.

[47] 伊藤雅光. (2002). 計量言語学入門. 大修館書店.

[48] 金田一春彦. (1953). 国語アクセント史の研究が何に役立つか. 金田一博士古稀記念言語民俗論叢. 329–354. 三省堂.

[49] 関一雄. (1958). 中古中世のいわゆる複合動詞について – 源氏・栄花・宇治拾遺・平家の四作品における –. 国語学. 32: 48–58.

[50] 大野晋. (1956). 基本語彙に関する二三の研究. 国語学. 24: 34–46.

[51] 国立国語研究所影山班プロジェクト「日本語レキシコン」. (2012). 動詞＋動詞型複合動詞・複雑動詞の研究文献一覧 (動詞連用形接続の複合動詞、テ形接続の複雑述語を含む). http://www.ninjal.ac.jp/lexicon/level1/post_2/index.html.

[52] 影山太郎. (2012). 動詞+動詞型複合動詞研究の現状.
http://www.ninjal.ac.jp/lexicon/%E5%BD%B1%E5%B1%B1(2012-09-24).pdf.

[53] 青木博史. (2012). 複合動詞の歴史的変化. 影山プロジェクト「日本語レキシコン」研究発表会. 1–8.

[54] 影山太郎. (1993). 文法と語形成. ひつじ書房.

[55] 斎藤達哉. (2011). 仮名写本における「改行」と「文字使用」. 専修大学人文科学研究所月報. 253: 11–29.

[56] 徳永良次. (1995). 用字法と書写意識. 北海学園大学人文論集. 5: 29–47.

[57] 師茂樹. (2007). 文字オントロジに基づく文字オブジェクト列間の編集距離. CHISE Conference 2005 報告書 & CodeFest 京都 2005 資料集. 1–7.

[58] W. Winkler. (1999). The state of record linkage and current research problems.

[59] M. Miyake. (2013). Different characteristics of variant readings based on comparison of major textual similarity measures. *in Proceedings JADH.*

[60] W.W. Cohen, P. Ravikumar, and S.E. Fienberg. (2003). A comparison of string distance metrics for name-matching tasks. *in Proceedings IJCAI-03 Workshop on Information Integration.* 73–78.

[61] 石井公成. (2001). N-gram 利用の可能性 –仏教文献における異本比較と訳者・作者判定–. 漢字文献情報処理研究 第 2 号 特集 2(N-gram が開く世界). 59–61.

[62] 山田崇仁. (2008). N-gram 方式を利用した漢字文献の分析. 立命館白川静記念東洋文字文化研究所紀要. 1: 1–23.

[63] 竹田正幸，福田智子，南里一郎，山崎真由美，玉村公一. (2007). 和歌データからの類似歌発見. 統計数理 (特集「文化を科学する」). 55: 289–310.

[64] 川崎宏. (1967). 文学作品の因子分析的研究 (i). 長崎大学教養部紀要 人文科学. 1–38.

[65] 安本美典，本多正久. (1981). 因子分析法. 培風館.

[66] 村田年. (2007). 多変量解析による文章の所属ジャンルの判別–論理展開を支える接続語句・助詞相当句を指標として–. 統計数理 (特集「文化を科学する」). 55: 311–326.

[67] 樺島忠夫. (1961). 文体の変異について. 国語国文. 30(11).

[68] 小野望，田中省作，持尾弘司. (2007). 母語学習者コーパスの基礎調査. 筑紫女学園大学・短期大学部人間文化研究所年報.

[69] 樺島忠夫. (1953). 文の長さについて–條件との相関の分析–. 国語学. 15: 21–31.

[70] M. Ishida and K. Ishida. (2007). On distibutions of sentece lengths in Japanse writing. *Glottometrics*. 15: 28–44.

[71] Palmkit. http://palmkit.sourceforge.net/.

[72] Srilm. http://www.speech.sri.com/projects/srilm/.

[73] 宮島達夫. (1970). 古典の品詞統計. 計量国語学. 53.

[74] 宮田和一郎. (1942). 更級日記の語法的研究. 国語文化.

[75] 吉田幸一. (1964). 和泉式部研究一 –和泉式部日記の基礎的研究 –. 古典文庫.

[76] 鈴木知太郎，川口久雄，遠藤嘉基，西下経一. (1957). 日本古典文学大系〈第 20〉土佐日記・かげろふの日記・和泉式部日記・更級日記. 岩波書店.

[77] 近藤みゆき. (2003). 和泉式部日記. 角川文庫.

[78] 清水文雄 (校注). (1981). 和泉式部日記. 岩波文庫.

[79] 大橋清秀. (1991). 和泉式部日記本文の研究. 和泉書院.

[80] 伊藤鉄也 (編). (1991). 四本対照 和泉式部日記 –校異と語彙索引 (古代中世文学資料研究叢書). 和泉書院.

[81] 和泉式部日記 (バージニア大学).
http://etext.lib.virginia.edu/japanese/izumi/shikibu/IzuSanj.html.

[82] 吉田幸一 (編). (1999). 笠間影印叢刊 19 榊原本 和泉式部日記. 笠間書院.

[83] 伊藤鉄也. (1992). 定家文字の自動翻刻. 西日本国語国文学データベース研究会. 1.

[84] 山田奨治. (1995). 高次局所自己相関特徴による古文書かな文字認識. 人文科学とコンピューター. 25-3: 21–30.

[85] 山田奨治，柴山守. (2002). 古文書を対象にした文字認識の研究 (特集 失われゆく情報の復元・保存技術: 人文科学における情報処理 (文献学・データベース共有・史料編纂)). 情報処理. 43: 950–955.

[86] M. Hayashi, S. Nishida, M. Nakata, Q.-W. Ge, and M. Yoshimura. (2008). A method of generating feature graph for handwritten character recognition of Japanese historical documents. *in Proceedings ITC-CSCC*. 305–308.

[87] 耒代誠仁，白井啓一郎，井上聡，久留島典子，馬場基，渡辺晃宏，中川正樹. (2012). シームレスコンピューティングのための古文書字形検索技術. 人文科学とコンピューターシンポジウム (じんもんこん 2012). 85–92.

[88] C. Panichkriangkrai, L. Li, and K. Hachimura. (2013). Interactive system for character segmentation of woodblock-printed japanese historical book images. *in Proceedings Culture and Computing*.

[89] A. Kundu and P. Bahl. (1988). Recognition of handwritten script: A hidden Markov model based approach. *in Proceedings ICASSP*. 2: 928–931.

[90] 嵯峨山茂樹，中井満，下平博. (2000). ストローク HMM に基づくオンライン手書き文字認識方式. 電子情報通信学会技術研究報告. PRMU2000(35): 1–8.

[91] J.H. AlKhateeb, J. Ren, J. Jiang, and H. Al-Muhtaseb. (2011). Offline handwritten Arabic cursive text recognition using hidden Markov models and re-ranking. *Pattern Recognition Letters*. 32: 1081–1088.

[92] 岡照晃，小町守，小木曽智信，松本裕治. (2012). 未整備の歴史的文献への濁点の自動付与アプリケーション. じんもんこん 2012 論文集. 7: 191–198.

[93] 玉井幸助. (1925). 更級日記錯簡考. 育英書院.

[94] 更級日記 (大阪大学). http://www.let.osaka-u.ac.jp/ okajima/sarasina.txt.

[95] 更級日記 (バージニア大学).
http://etext.lib.virginia.edu/japanese/sarashina/SugSara.html.

[96] 西下経一 (校注). (1930). 更級日記. 岩波文庫.

[97] 犬養廉 (編). (1968). 影印本 更級日記. 新典社.

[98] 伊井春樹 (編). (1985). 校注 更級日記. 和泉書院.

[99] 小木曽智信，小椋秀樹，田中牧郎，近藤明日子，伝康晴. (2010). 中古和文を対象とした形態素解析辞書の開発. 情報処理学会研究報告. 2011-CH-85: 1–8.

[100] 小木曽智信. (2011). 通時コーパスの構築に向けた古文用形態素解析辞書の開発. 研究報告 人文科学とコンピュータ（CH）. CH-92: 1–4.

[101] 小椋秀樹，須永哲矢，小木曽智信，近藤明日子，田中牧郎. (2011).「中古和文 unidic」における言語単位の設計. 言語処理学会 第 17 回年次大会 発表論文集.

[102] 和文茶まめ http://www2.ninjal.ac.jp/lrc.

[103] 石原昭平，津本信博. (1986). 主要参考文献解題. 別冊国文学王朝女流日記必携，秋山虔（編）. 學燈社.

[104] 阿部圭一. (1991). 『和泉式部日記』参考文献. 女流日記文学講座 3 和泉式部日記・紫式部日記. 147–158. 勉誠社.

[105] 大橋清秀. (1954). 和泉式部日記成立考. 平安文学研究. 16.

[106] 伊藤博. (1956). 和泉式部日記諸本の系統について. 国語. 4(4).

[107] 川瀬一馬. (1953). 和泉式部日記は藤原俊成の作. 青山学院女子短期大学紀要. 2: 21–52.

[108] 伊藤博. (1981). 和泉式部日記伝本攷. 桜楓社.

[109] 森田兼吉. (1977). 和泉式部日記論攷. 笠間書院.

[110] 竹内美智子. (1986). 平安時代和文の研究. 明治書院.

[111] 織田裕子. (1958). 「和泉式部日記」の作者について. 国語国文. 27(4).

[112] 今井卓爾. (1957). 平安時代日記文学の研究. 明治書院.

[113] 池田亀鑑. (1944). 平安時代文学概説. 八雲書店.

[114] 梅津真理子. (1956). 和泉式部日記の作者について –作者俊成説に対する疑問–. 国語. 4(4).

[115] 大橋清秀. (1961). 和泉式部日記の研究. 初音書房.

[116] 神谷かをる. (1991). 女流日記の文体と機能. 女流日記文学講座 1 女流文学とは何か. 勉誠社.

[117] 竹内美智子. (1963). 『和泉式部日記』の語彙に関する一考察. 国語学. 53: 10–18.

[118] 西端幸雄，木村雅則，志甫由紀恵. (1996). 平安日記文学総合語彙索引. 勉誠社.

発表論文一覧

本研究に関係する発表論文

[1] Y. Tachioka: Objective measurement of variants in classical literature, The International Conference on Culture and Computing (Culture and Computing 2013), Kyoto, pp.202-203, 2013. 9.

[2] 太刀岡勇気: 古典文学における異本間の関係性の客観分析―『和泉式部日記』『更級日記』を題材に―, 計量国語学会講演論文集, 首都大学東京, pp.37-42, 2013. 9.

2013 International Conference on Culture and Computing

Objective measurement of the relationship between variants in classical literature

Yuuki Tachioka
The College of Humanities and Sciences, The Nihon University
Chiyoda-ku, Tokyo, Japan
Email: yuuki_tachioka@yahoo.co.jp

***Abstract*—Stylometrics is a method of analyzing the style of a text using metric features. To apply this to classical literature, it is required that the diversity of variants of the same work is sufficiently smaller than that between different works because for the same work there are many variants, which have been changed from the original form. In this paper, this prerequisite will be confirmed, and the use of the Chinese character ratio, which is affected by the original form, Levenshtein distance, and perplexity is introduced. The experimental results show that the Chinese character ratio is effective for discriminating series of variants and that the Levenshtein distance and perplexity are also effective in addition to principal component analysis of features, which is general in Stylometrics. Especially, by using perplexity, the diversity between variants can be quantitatively compared in different works.**

***Keywords*-Stylometrics; variants of text; principal component analysis; Levenshtein distance; perplexity**

I. Introduction

Stylometrics is a method for analyzing the style of a text using statistical methods [1]. For Japanese these types of research are more difficult than for English because analyses of agglutinative languages, such as Japanese, need part-of-speech (POS) tagging, which divides sentences into words and gives them the POS of the words; but development of the POS tagger enables mechanical POS tagging, and currently the amount of this type of research (e.g., [2]) is increasing.

There are some studies dealing with classical literature but these types of research use reprinted texts. However, to deal with classical literature quantitatively, it is essential to consider variants. There are few original manuscripts written by the authors of classical literature. Most of the texts have been preserved by being transcribed repeatedly, and errors, reformations, and additions change the original contents. The metric analyses are used to discriminate between works, but it is required that a diversity of variants of the same work is sufficiently smaller than that between works. This prerequisite of the research has not been confirmed. This paper focuses on the Japanese classics, but the same problems occur in other languages. The objects of this research are four variants of "Izumishikibu Nikki" with comparison to another work ("Sarashina Nikki"). This paper introduces the use of the Chinese character ratio, which reflects the original usage, Levenshtein distance, and perplexity, and shows that these reflect the degree of diversity of the variants or works.

II. PCA of text features for style analysis

Style analysis uses some features that can discriminate between works effectively. Though general features are shown in IV, here the Chinese character ratio (the ratio of Chinese characters to all characters) is introduced to discriminate between different variants of the same work. Since Japanese sentences consist of Chinese characters and hiragana (Japanese inherent characters), the usage of Chinese characters is affected by the original usage and this ratio characterizes the series of variants. Features' dimensions are reduced by the principal component analysis (PCA) [3].

III. Objective measurement of variants using Levenshtein distance and perplexity

To calculate the diversity of variants, Levenshtein distance is introduced. An arbitrary sentence can be converted to any by three procedures: substitution, insertion, and deletion. The Levenshtein distance is defined as the minimum steps required (i.e., cost) to convert one sentence to another. Dynamic programming (DP) enables fast calculation.

The basic model that analyzes the sentences is N-gram, which is a chain probability of words or characters. The N-gram model is more flexible than Levenshtein distance because, if contents are totally different, texts can be compared quantitatively by perplexity defined as $P(w_1, \ldots, w_N)^{\frac{1}{N}}$ where P represents a probability of observed N words sequence $w_1, \ldots, w_N$ and perplexity is an inverse of its geometrical mean. Perplexity is related to the number of the next word candidates assuming that the emerging probability is the same. For texts which can be easily estimated by N-gram models, perplexity is low. Using this property, similarity of texts can be quantitatively evaluated.

It is inefficient to deal with variants by respective databases because the variants have a similarity. Here, a data structure which can deal with different texts in a single database is proposed as $\backslash\mathrm{d}\{[\mathrm{t}_i]\mathrm{text}_i@\mathrm{text}_0\}$ where text_i is a different part compared with the base text (text_0). For example, there are four different sentences (AAD, ABD, ABD, ACD) compared with the base sentence (ACD). This part is integrated into the form: $\mathrm{A}\backslash\mathrm{d}\{[\mathrm{t}_1]\mathrm{A}[\mathrm{t}_2\mathrm{t}_3]\mathrm{B}@\mathrm{C}\}\mathrm{D}$. This form also enables fast computation of the Levenshtein distance because the DP alignments are easily arranged.

978-0-7695-5047-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CultureComputing.2013.65

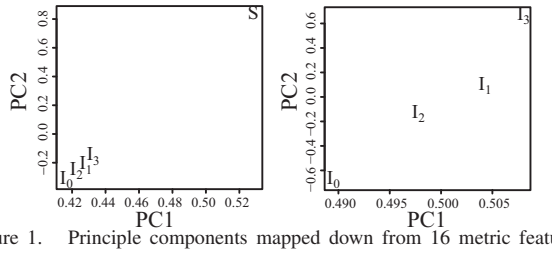

Figure 1. Principle components mapped down from 16 metric features where I_0–I_3 are variants of "Izumishikibu Nikki" and S is "Sarashina Nikki". (left: PCA among I_0–I_3 and S, right: PCA among I_0–I_3)

Table I
LEVENSHTEIN DISTANCE BETWEEN FOUR VARIANTS.

	Sanjo (I_0)	Kangen (I_1)	Ouei (I_2)	Konsei (I_3)
Sanjo	0	4916	5480	5751
Kangen	4916	0	5412	5148
Ouei	5480	5412	0	3305
Konsei	5751	5148	3305	0

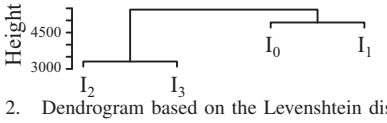

Figure 2. Dendrogram based on the Levenshtein distance.

IV. RELATIONSHIP BETWEEN VARIANTS OF THE SAME WORK WITH COMPARISON TO ANOTHER WORK

To clarify the relationship between variants, this paper focused variants of "Izumishikibu Nikki" and "Sarashina Nikki", which are of similar length. "Izumishikibu Nikki" has four series of variants: Sanjonishi-ke (Sanjo), Kangen, Ouei, and Konsei. Sanjo is the oldest and the most popular. Three other variants were compared with Sanjo after constructing an aforementioned form database with reference to [4]. The base text of "Sarashina Nikki" was Teika-bon, which is the most common. For POS tagging, [5] was used. The tagging errors (up to 5%) were manually modified.

The number of features is 16 as follows: The first four are the sentence length, the quotation ratio (Japanese poem and conversation), the ratio of representing one's feelings directly[1] and additionally the Chinese character ratio; after POS tagging, twelve other features are added to analyze a style of text: modifier-verb ratio[2] and other POS ratios (the ratio of independent word, pronoun, adjective, nominal adjective, adverb, noun, verb, Japanese-origin word, Chinese origin-word, mixed-origin word, and proper noun). According to the text style analyses, the Chinese character ratio (I_0: 7.2%, I_1: 8.3%, I_2: 8.6%, I_3: 10.7%, S: 9.1%) is effective for discriminating between variants where I_0, I_1, I_2, and I_3 denote Sanjo, Kangen, Ouei, and Konsei, respectively, whereas S denotes Sarashina. On the other hand, the quotation ratio (47.5%, 46.6%, 47.4%, 46.4%, 33.2%), the ratio of parts representing one's feelings (12.2%, 12.3%, 12.4%, 12.8%, 4.1%), and the noun ratio (37.1%, 37.3%, 36.6%, 36.7%, 46.6%) are effective for discriminating between works. Principle components obtained by principal component analysis depict the relationship as shown in Fig. 1. Attribution ratio is 100% by these two principal components. PCA can distinguish I_0–I_3 from S, but the relationship among I_0–I_3 is not clear quantitatively because the value range is different at each feature and the Euclid distance between points on the graph is meaningless. This shows the limitation of PCA.

[1]These two features show whether description is direct/indirect-oriented.

[2]The ratio of the number of modifiers (adjective, nominal adjective, adverb, and pronoun adjectival) to the number of verbs: A higher value shows static-oriented and a lower value shows dynamic-oriented.

Table II
PERPLEXITY USING WORD TRI-GRAM OF FOUR VARIANTS OF "IZUMISHIKIBU NIKKI" COMPARED WITH "SARASHINA NIKKI".

train \ eval	Sanjo	Kangen	Ouei	Konsei	Sarashina
Sanjo	8.9	35.2	35.6	41.4	198.5
Kangen	36.4	9.0	40.0	40.2	205.7
Ouei	37.2	40.3	9.3	26.7	195.3
Konsei	44.2	41.1	26.6	9.5	204.1
Sarashina	242.6	241.6	244.3	235.9	9.0

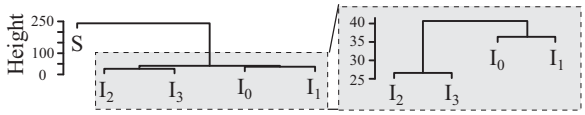

Figure 3. Dendrogram based on perplexity.

Table I shows the Levenshtein distance between four variants of "Izumishikibu Nikki". Fig. 2 shows the relationship between variants. Levenshtein distance clarifies the relationship between variants quantitatively, but this technique cannot be applied to other works (e.g., "Sarashina Nikki"). Prior alignment should be arranged to some extent manually (e.g., per sentence) before DP. If not, mis-alignment would decrease accuracy. On the other hand, Table II and Fig. 3 show the perplexity and its dendrogram using a tri-gram model between variants with the different work. Perplexity calculation only needs a tri-gram model after "correct" POS tags are obtained. One text is selected for training and tri-gram models are constructed by [6], and the other texts are evaluated in terms of perplexity. Dendrograms obtained by using Levenshtein distance and perplexity are similar.

V. CONCLUSION

This paper aims to clarify the relationship between variants of texts. Experiments show that the Chinese character ratio is effective for discriminating between variants and that the Levenshtein distance and perplexity are also effective. Especially, using perplexity, the relationship between variants can be quantitatively compared to different works.

ACKNOWLEDGMENT

The author would like to thank Prof. Ogino at Nihon Univ.

REFERENCES

[1] A. Kenny, *The computation of style*. Pergamon Press, 1982.

[2] M. Jin. and M. Murakami, "Authorship identification using random forests," *Proc of the Inst of Stat Math (J)*, vol. 55, pp. 255–268, 2007.

[3] D. Kaplan, "A computational approach to style in American poetry," *in Proc Int Conf on Data Mining*, pp. 553–558, 2007.

[4] T. Ito, *Izumishikibu Nikki*. Izumi Shoin, 1991.

[5] "http://www2.ninjal.ac.jp/lrc/".

[6] "http://www.speech.sri.com/projects/srilm/".

古典文学における異本間の関係性の客観分析
–『和泉式部日記』『更級日記』を題材に –

太刀岡勇気 (日本大学)

1　はじめに

近年のコンピュータ科学の進展に伴って、人文科学の分野でも自然言語処理の分野で用いられてきた計量的な手法 [1] によって、文献資料や文学作品を分析する研究が行われている [2, 3]。科学的方法に則り客観的な事実から判断でき、主張に一般性を持たせられるのが最大の特長であり、異なる文献の客観的比較が簡単に行えるようになった。研究の立場としては、文字情報に注目するもの [4] と、タグ付けされた品詞情報を用いるもの [5] がある。膠着語である日本語は品詞のタグ付けの手間がかかるため、後者の研究はあまり盛んではなかったが、近年、形態素解析器 [6] を用いて品詞情報を機械的にタグ付けすることで後者の研究も行えるようになった。

しかしながら従来研究には、テキストの取り扱い方に問題があると考えられる。前提として、対象とするテキストが統一的な基準でタグ付けされていることが必要であるのに、多くの分析が、『古典文学大系』等の校訂済み本文を用いて行われている。校訂は複数の写本を元に編者の主観的判断によってなされるため、これでは編者のバイアスが混入してしまう。ここではできるだけ本文に近い形でデータベースを作成し、本文に即した指標も使う。同様に異本の問題がまったく考慮されていない。古典文学は著者による原本がほとんど存在せず、現状利用できるのは何度も書写を重ねられてきた写本であり、異なる写本 (異本) が残されている [7]。これは近現代文学ではそれほど問題とならないが、近代以前は書写者にオリジナルを尊重する意識がそれほど高くなかったため、改変や創作が行われている。また誤写などもあり、一つの本だけで本文は同定できない。計量的な分析手法は異なる作品を区別するような手法を提案しているが、異本のばらつきが異なるテキスト間のばらつきよりも十分小さいことが必要である。

本報では『和泉式部日記』と『更級日記』を題材に[1]計量的な分析を行う。書誌学的にも、『和泉式部日記』は書写時期が最も古く最良本とされる「三条西家本」だけでは不足であることが国文学の立場から指摘されている [8]。そこで『和泉式部日記』の 4 つの異本を対象に、異本間の関係性を明らかにする。加えて、他本間の比較として、同程度の分量からなる『更級日記』との比較を行う。これにより、同一作品内での異本によるばらつきと、作品の違いがどの程度指標に反映されるかを明らかにできる。

2　計量分析手法

2.1　異本を効率的に扱うデータ形式

異本には類似性があるので、それらのテキストを別々に管理するのは非効率的である。ここでは 1 つのテキストから複数の異本が生成可能な独自の仕様を以下のように定義した。

```
\d{[t1] テキスト 1[t2] テキスト 2@底本}
```

ある底本に対して、異なる箇所のみを上記のようにマークアップすることで複数異本が 1 つのデータベースとして管理可能である。

[1] 中古の日記文学については検討が見られない。

三条西家本の本文に他本の異同を対校しながら翻刻した文献 [9] をもとにこの形式のデータベースを作成した。例えば『和泉式部日記』のはじめの部分を対校すると、

1. ゆめよりもはかなき世のなかをなげきわびつゝあかしくらすほどに、(三)
2. ゆめよりもはかなきよのなかをなげきわびつゝあかしくらすほどにはかなくて、(寛)
3. 夢よりもはかなき世の中をなげきつゝあかしくらすほどにはかなくて、(応, 混)

のようになる。これを

```
\d{[応混] 夢@ゆめ}よりもはかなき\d{[寛] よの [応混] 世@世の}\d{[寛応混] 中@なか}をなげき\d{[応混]@わび}つゝあかしくらすほどに\d{[寛応混] はかなくて@}、
```

のように効率的に表すことができる。

2.2 n-gram 分析

文章を分析する基本となるのは文字あるいは単語の連鎖を確率で表す n-gram である。n-gram 分析を単語で行うためには、あらかじめ形態素解析により文を形態素列に分割しておく必要がある。大量の文章からこの確率を学習すれば言葉の用いられかたが明らかとなる。またある対象となる文章から n-gram の確率を学習することで、その文章の癖を学習できる。

n-gram 分析を行う際には、かな漢字交じりで行うものと、すべてひらがなで行うものがある。ただし、同一本文でもかな漢字の揺れがあるため、かな漢字交じり文を扱うと問題を生じることもある[2]。一方、かな漢字交じりは文章の書き手の特性・時代背景を考慮できるという利点もある。本研究ではかな漢字交じりで分析した。

[2]特に和歌の n-gram 分析ではすべてひらがなに直してから、分析することが多い [2]。これは和歌特有の掛詞の問題を考慮するためでもある。

2.3 形態素解析

形態素解析により、文を単語に分割し品詞をタグ付けできる。これは日本語などの膠着語で、n-gram 分析と文体指標を算出するのに必要となる。品詞のタグ付けは、中古語の形態素辞書「中古 UniDic」[10] を「MeCab」[6] と組み合わせた形態素解析器「和文茶まめ」によった。ただし形態素解析の誤り (全体の 5%程度) や以下の問題があるので、人手で修正を加えた。

2.4 古典語に形態素解析を適用する際の問題

形態素を構成する**単位**は、字面の文字とするのが一般的である。しかしこれで充分であろうか [11]。古典語を分析の対象とする場合にはさらに難しい。例えば、「我身」を茶まめに掛けると、「我(ワレ)(代名詞)+身(ミ)(名詞)」のように誤った結果が得られる。これは中古 UniDic が「わが」を連体詞としていないためである。校訂されて「我が身」となっていれば、「我(ワレ)(代名詞)+が(ガ)(助詞-格助詞)+身(ミ)(名詞)」のように、品詞上は正しい結果となる。「我身」は一語の名詞として扱うこともできる[3]が、「我身の上」は「我身+の+上」ではない。新しい名詞として登録すると使用頻度の低い名詞が増えてしまう。「我」を「連体詞」とすれば、「我(ワガ)(連体詞)+身(ミ)(名詞)」「我(ワガ)(連体詞)+身(ミ)(名詞)+の+上」のように正しく形態素解析できるが、「我が身」に対しても「我が(ワガ)(ガ)(連体詞)+身(ミ)(名詞)」としなければ一貫性が失われる。これは、文字単位での形態素解析の限界を表している。例えば、読みの文字列「わがみ」に対して形態素解析を行えば、「わが」を連体詞としなくても、上述の正しい結果が得られる。校訂済み本文であれば、送り仮名を一意に決めているが、送り仮名の揺れが大きいオリジナルテキストを解析する際には大きな問題となる。

[3]『旺文社古語辞典』では一語の名詞としている

つぎに、**連濁**の問題がある。「木の葉」であれば「木(コ)(名詞)+の(ノ)(格助詞)+葉(ハ)(名詞)」とするのは問題ないと考えられる。しかし「紅葉葉」の3文字目は「バ」と読まれるがこれを「葉(バ)(名詞)」あるいは「葉(バ)(接尾辞)」とするのがよいか、「紅葉葉」を一単語とするのが良いかは問題である。「葉(バ)(名詞)」とするのは、単独で「葉(バ)」と読まれることは無いので抵抗がある。「葉(バ)(接尾辞)」とした場合には「落ち葉」も「落ち(オチ)(動詞)+葉(バ)(接尾辞)」とするのだろうか。本論では連濁の起こっているものは一つの単語として扱った。

また、**掛詞**の問題もある。和歌は、表・裏どちらで解釈するかが問題である[4]。2通りで解釈しておくという方法もあるが、いつでも2通りの解釈が可能なわけでもない。「あふみち」で「逢ふ(アフ)+道(ミチ)」と「近江路(アフミヂ)」のように濁点の違いで表記不能なものもある。本論では表の意味を主体とし、濁点の有無で意味が変わる場合には濁点をつけない方の意味を優先した。

中古 Unidic では、「して (接続詞)」を「す (動詞)+て (接続助詞)」とするなど、還元主義的な部分も見られる一方で、「動詞+す・さす」で表される使役動詞は別に項を立てるなど、あまり一貫していない。どの粒度で分析するかに関しては**一貫性**が必要である。学習コーパスの一貫性もある。例えば「宣はせず」で「のたまわ (ノタマウ:動詞)+せ (ス:助動詞)+ず (ズ:助動詞)」と「のたまはせ (ノタマワス:動詞)+ず (ズ:助動詞)」と「は」と「わ」を替えただけで異なる分析結果となる。これは前者が主に近世のコーパスから学習したもので、後者が中古のコーパスから学習したものであるためと考えられる。翻刻では中古本文に対しては後者で統一されているが、原本には両方の表記があり得る。また「ものから」のように「もの＋から」の結合で品詞が変化 (名詞から接続詞) するものもある。元の意味を失っていると考えられる品詞変化に関しては、変化後の品詞を使った。

複合名詞・動詞は元の名詞・動詞とは意味が異なる。複合動詞を認めるかどうかで文体と品詞構成比率に大きな差がでることが示されている [12]。例えば、「世の中」は、文脈によっては「世+の+中」(世間) ではなく「世の中」(男女の仲) として解釈すべきである。同様に「見知る」は「見る+知る」でもよいかもしれないが、「思ひ立つ」(決意する) は「思ふ+立つ」(考えて出発する) ではない。ただし、複合動詞中に係助詞が挿入されることがあることはよく知られており、「思ひ立つ」を一語とした場合には、「思ひも立たず」の解釈が難しい「おぼし立つ」と「思ふ」の部分が尊敬語化したときに、これを別の動詞とするかという問題もある。本論では、「思ひ立つ」は一語として扱ったが、「思ひも立たず」「おぼし立つ」は複合語とした。

古典語では、**表記の揺れ**が多い。例えば、「思{ふ, ひ, へ}」の活用語尾は省かれるため、「思」に「おもふ」「おもひ」「おもへ」など複数の読みを持たせる必要があり[5]、形態素解析器の学習の際に考慮が必要である。今回は人手で修正した。古典語は表記が多様性に富み、一つの語に複数の意味を担わせることもあるため、現代語よりも格段に問題は複雑である。

3 計量分析指標

3.1 文体の分析指標

文献 [15] にあげられている文体を分析するための9つの指標から、古典の分析にも適用可能

[4]「みるめ」を「見る目」とするか「海松布」とするかで形態素が変わる。

[5]「宣う」も「のたまふ」「のたまう」「の給ふ」「の給う」「の給」の表記がある。「お」と「を」の揺れも多い。

な以下の 5 つの指標を用いた。文章に含まれる**名詞の割合** (式 (1) が文章の性質を表すことが古くから知られている。

$$\text{名詞率} = \frac{\text{名詞数}}{\text{自立語数}} \times 100[\%] \\ \text{自立語数} = \text{全単語} - \text{助詞数} - \text{助動詞数} \quad (1)$$

Modifier Verb Ratio(MVR) は、「形容詞・形容動詞・副詞・連体詞」(Modifier) の合計数を「動詞」(Verb) で除した比率を表す (式 (2))。これは値が高いほど「ありさま描写的」、低いほど「動き描写的」であるとされる。

$$MVR = \frac{\text{形容 (動) 詞+副詞+連体詞}}{\text{動詞数}} \times 100[\%] \quad (2)$$

文中に含まれる**指示詞の割合**を式 (3) により求める。指示詞の適切な使用により、文章の冗長性が減り、可読性が向上する。

$$\text{指示詞率} = \frac{\text{指示詞数}}{\text{自立語数}} \times 100[\%] \quad (3)$$

平均文長を式 (4) により求める。古典語の文章は現代語の文章に比べて、一文の長さが長い。

$$\text{文長} = \frac{\text{自立語数}}{\text{全文数}} \ [\text{語/文}] \quad (4)$$

引用文の比率は、古典文学に厳密に適用するのは難しいが、ここでは、和歌、会話、心情表現に該当する箇所を引用部分とした。全体の文章に占める引用や会話部分の割合を式 (5) により求める。また心情表現に関しても、直接表現 (e.g.「あさまし」とおぼゆ) と間接表現 (e.g. あさましうおぼゆ) の 2 通りが考えられ、どちらを使うかに作者の特徴が現れると考えられる。ここでは前者は心情表現を直接的に表している箇所であるとして、式 (6) により求めた。

$$\text{引用率} = \frac{\text{引用・会話文字数}}{\text{全文字数}} \times 100[\%] \quad (5)$$

$$\text{心情率} = \frac{\text{心情表現文字数}}{\text{全文字数}} \times 100[\%] \quad (6)$$

校訂済み本文は漢字が現代的な基準で見て適当になるように校訂されているが、中古の本文はかなが圧倒的に多い。校訂前の本文に対しては、**漢字率**も指標として用いることができる。異本は元の本文の影響を少なからず受けると思われるので、漢字率を算出することで、当該本文を特徴づける量とすることができる。これに**各種品詞の割合**および**語種**を考慮した 16 種類の指標により評価した。

3.2 Levenshtein 距離および perplexity

文字の相違率を判断するために Levenshtein 距離 (編集距離) を用いた。任意の文字列間は置換、挿入、削除の 3 つの手順により変換できるが、Levenshtein 距離はそのような手順の最小回数として与えられる。これはある文字列を他の文字列に変換するのにかかるコストを距離として用いたもので、動的計画法に基づくアルゴリズムで高速に計算でき、コストを自分で決めることで誤りやすい文字間のペナルティーを考慮することができる [13][6] という特長がある。

n-gram 分析の類似性を指標として使うこともできる。1-gram は汎用的に異なるテキスト間で比較可能である。それ以上の連鎖 (2-/3-gram) に関しては、学習テキストを一つ選び、言語モデルを作成し[7]、それ以外を評価テキストとして式 (7) で表される perplexityPP を評価した。

$$PP = P(w_1, \ldots, w_n)^{\frac{1}{n}} \quad (7)$$

ここで $P()$ は単語列 $w_1, \ldots, w_n$ が観測される確率で、PP はその相乗平均の逆数である。perplexity は次にくる単語が等確率と考えたときの予測される平均単語数を表す。これにより、テキストの類似性が定量的に評価できる。

[6] 「ん」と「む」の間の距離は 0 とした。

[7] srilm[14] によった。

4　『和泉式部日記』4 異本間の関係性と『更級日記』との比較

4.1　底本について

『和泉式部日記』の原本は見つかっておらず、三条西家本系統 (以下「三」と略す)、寛元本系統 (「寛」) ・応永本系統 (「応」) ・混成本系統 (「混」) の 4 系統に分類されている。このうち、三条西家本系統のみが室町時代の書写であることが知られており、それ以外は江戸時代の書写である。三条西家本が最も古いこともあって、各種翻刻テキストの最も一般的な底本となっている。ここでは 4 種の異本の代表的なものを用いて、それらの関係性を探る。『更級日記』(「更」) の底本には「定家本」を用いた。

4.2　結果と考察

4.2.1　文体の分析指標

2 章で述べた指標に従い、分析を行った。結果を表 1 と表 2 に示す。異本間の差異を表すものとしては漢字率が、他本間の差異を表すものは引用率・心情率・名詞率があげられる。『和泉式部日記』は和歌の引用が多く、「女」の心情表現が豊かであることが特徴であるのでそれが表れているものと思われる。異本間では漢字率に大きな差異が現れたのは、写した人の性別・年代などが影響していると思われる。語種は和・漢・固有・混種の 4 種類に関してその出現頻度を比較した。『更級日記』は漢語率が若干高いものの顕著な差は見られなかった。

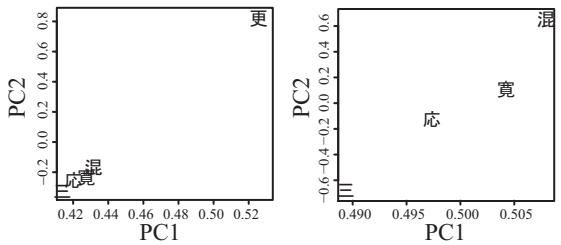

図 1　分析指標の主成分分析結果

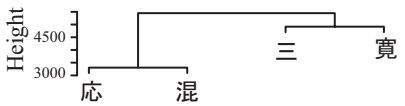

図 2　Levenshtein 距離に基づくデンドログラム

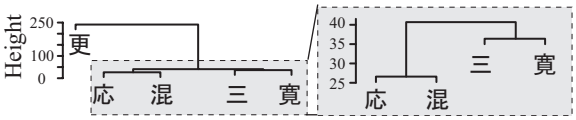

図 3　Perplexity に基づくデンドログラム

指標が 16 個あり、それぞれの関係がわかりにくいので、主成分分析により主要な変数 2 つを取り出した[8]。結果を図 1 に示す。ただしどの本同士がどれほど近いかを定量的に測ることは難しい。各指標間でダイナミックレンジに差があるためである。例えば、主成分分析した平面上での Euclid 距離は意味を持たない。このように主成分分析には限界がある。

4.2.2　Levenshtein 距離、perplexity による分析

Levenshtein 距離を用いて『和泉式部日記』4 異本間を、図 2 のように、今回分析した異本は 2 つのグループに分けられることが分かった。

このように Levenshtein 距離によって異本間の関係性を考察することができるが、他本 (更級日記) に対しては使えない。また本文全体に動的計画法を用いると精度が低下するため、事前の整列が必要で、それなりに手間がかかる。そこで 3-gram を作成しその perplexity を計算した。図 3 にデンドログラムを示す。異本は Levenshtein 距離を用いた場合と同様に分類でき、他作品との比較も行えている。これから異本間のばらつきは他作品に比べて十分小さいことが裏付けられた。本手法であれば、形態素解析ができていれば、3-gram を作るだけで計算できる。

[8] 寄与率は 2 つの変数で 100%である。

表 1　分析結果

	(総文字数)	漢字率	平均文長	引用文字率	心情文字率	(総単語数)	自立語率	MVR
三条西	(20025)	7.2%	52.4	47.5%	12.2%	(10810)	52.6%	40.5%
寛元	(19975)	8.3%	54.0	46.6%	12.3%	(10906)	52.4%	39.4%
応永	(19840)	8.6%	53.3	47.4%	12.4%	(10865)	52.5%	41.5%
混成	(20200)	10.7%	54.4	46.4%	12.8%	(11186)	52.3%	41.6%
更級日記	(26546)	9.1%	66.7	33.2%	4.1%	(14517)	55.7%	40.3%

表 2　分析結果 (続き)

	名詞率	代名詞率	形容詞率	形状詞率	副詞率	動詞率	和語	漢語	固有語	混成語
三条西	37.1%	3.4%	8.0%	1.2%	7.4%	40.9%	98.3%	1.3%	1.0%	0.3%
寛元	37.3%	3.3%	7.9%	1.1%	7.3%	41.1%	98.3%	1.3%	1.0%	0.3%
応永	36.6%	3.2%	7.9%	1.2%	7.8%	40.8%	98.2%	1.4%	1.0%	0.3%
混成	36.7%	3.1%	7.8%	1.1%	8.0%	40.7%	98.1%	1.5%	1.0%	0.3%
更級日記	46.6%	3.6%	7.9%	1.1%	5.1%	35.1%	96.3%	2.2%	1.3%	0.3%

5　まとめ

本研究は中古の日記文学の代表格である『和泉式部日記』と『更級日記』を題材に、『和泉式部日記』の 4 つの異本と『更級日記』の関係性を明らかにすることを目的として、計量的な分析を行った。その結果、異本間の差異を表すものとしては漢字率が、他本間の差異を表すものは引用率・心情率・名詞率・代名詞率が有効である可能性が示された。異本間の異なり度を測る指標として、計量分析によく用いられている分析指標の主成分分析に加えて、文字列同士の Levenshtein 距離や perplexity が有効であることが分かった。特に perplexity を用いることで、同一作品の異本間の差異と異なる作品間の差異を比較できる。これにより同一作品の異本間の差異は、異なる作品間の差異に比べて小さいことが定量的に確かめられた。

謝辞

本研究の遂行に当たっては日本大学文理学部荻野綱男教授および鈴木功眞准教授にご指導いただいた。ここに感謝申し上げる。

参考文献

[1] S. Bird, E. Klein, E. Loper, 萩原正人他 (訳). (2010). 入門 自然言語処理. オライリー・ジャパン.
[2] 近藤みゆき. (2000). n グラム統計処理を用いた文字列分析による日本古典文学の研究. 千葉大学人文研究 人文学部紀要. 29: 187–238.
[3] 金明哲. (2000). 自然言語処理における統計手法を用いた情報処理. 統計数理. 48: 271–287.
[4] 近藤泰弘, 近藤みゆき. (2001). 平安時代古典語古典文学研究のための n-gram を用いた解析手法. 言語処理学会年次大会発表論文集. 7: 209–212.
[5] 金明哲, 村上征勝. (2007). ランダムフォレスト法による文章の書き手の同定. 統計数理. 55: 255–268.
[6] http://mecab.sourceforge.net/.
[7] 堀川貴司. (2010). 書誌学入門 古典籍を見る・知る・読む. 勉誠出版.
[8] 森田兼吉. (1996).『和泉式部日記』は三条西家本だけでは読めない. 日本文学研究. 31: 17–28.
[9] 伊藤 鉄也 (編). (1991). 四本対照 和泉式部日記 –校異と語彙索引. 和泉書院.
[10] 小木曽智信, 小椋秀樹, 田中牧郎, 近藤明日子, 伝康晴. (2010). 中古和文を対象とした形態素解析辞書の開発. 情報処理学会研究報告. CH-85: 1–8.
[11] 伊藤雅光. (2002). 計量言語学入門. 大修館書店.
[12] 大野晋. (1956). 基本語彙に関する二三の研究. 国語学. 24: 34–46.
[13] 師茂樹. (2007). 文字オントロジに基づく文字オブジェクト列間の編集距離. CHISE Conference 2005.
[14] http://www.speech.sri.com/projects/srilm/.
[15] 小林千草. (2005). 文章・文体から入る日本語学. 武蔵野書院.