ID	KK290607
種別	論文
タイトル	中古日記文学の計量国語学的分析と異本間の関係性の客観分析 —『和泉式部日記』と『更級日記』を題材に—
Title	Metric Japanese study on diary literature in the Heian period and objective measurement on relationship between variants: Analyses on "*Izumishikibu nikki*" and "*Sarashina nikki*"
著者	太刀岡　勇気
Author	TACHIOKA Yuuki
掲載号	29巻6号
発行日	2014年9月20日
開始ページ	187
終了ページ	210
著作権者	計量国語学会

計量国語学 29 巻 6 号（2014 年 9 月）pp.187-210.

論文

中古日記文学の計量国語学的分析と異本間の関係性の客観分析

―『和泉式部日記』と『更級日記』を題材に―

太刀岡 勇気（三菱電機株式会社）

要旨

計量国語学的分析では，文章をいくつかの計量指標に基づき分析する．このような方法は主に，現代語の分析に使われ，著者同定などに成果を挙げている．しかしながらこの方法を古典文学作品に適用する際には，異本の問題が生じる．原本が残っていないことが通例の古典文学作品には異本が多く，これが時には同一著者のものとは思えないほどの文章の相違を伴うからである．本稿では，編集距離とパープレキシティーを用いることで，異本間の関係性を定量的に表す方法の有効性を示す．提案法が従来の計量指標の主成分分析による分類法に比べて，文献学の分野での知見とよりよい一致を示すことを，中古日記文学の代表的な作品である『和泉式部日記』を用いて検証する．さらに同一作品中の異本間の差異が，他作品との差異に比べて十分小さいことを，『更級日記』との比較を通じて示す．

キーワード：異本，文体分析，編集距離，N グラム，パープレキシティー

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（北 1999）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（近藤 2000; 金 2000; 村上 2004）．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（Brinegar 1963; 村上 2004; Uzuner and Katz 2005; Mingzhe and Minghu 2012）である．単語（Uzuner and Katz 2005）や句読点の使用法（Mingzhe and Minghu 2012）から著者性をとらえる研究が多い．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る．これにより，テキストの客観的比較を簡便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して，人間の読みを支援することが期待されている（近藤 2001; Grimmer and Stewart 2013）．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．例えば，近藤，近藤（2001b）では，和歌と散文をひらがなの連鎖とみなしたときの n-gram を比較することで，手作業では気づきにくい新たな引き歌の発見に計量分析を役立てている．

このような文字列だけの情報に加えて品詞情報を用いることで，文字面だけからは得ら

れない豊富な情報が得られる．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，ある程度機械的に品詞情報のタグ付けが行えるため，このような立場での研究が進んでいる．日本語においても，従前より品詞数などの計量的な分析は行われてきた（宮島 1970）が，日本語は膠着語であるため品詞のタグ付けが難しく，このような研究は主に人手で整備された索引を利用して行われてきた．そのため索引による基準の違いが問題となっていた．近年，統計モデルを用いた精度の高い形態素解析器が開発され，品詞情報のタグ付けが機械的に行えるようになった．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（金，村上 2007）がある．

計量国語学的な分析を古典文学作品に適用する際には，データベースと解析手法が問題となる．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．本稿では，誤りを含む形態素解析結果をそのまま利用するのではなく，できうる限り誤りを積極的に修正することを前提とする．古典文学作品を扱うためには，異本の問題を無視するわけにはいかない．古典文学作品は，著者による原本はほとんど残っておらず，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）が存在する（堀川 2010）．これは近現代ではさほど問題とならないが，近代以前は書写者に原本を尊重する意識がそれほど高くなかったため，誤写や自由な改変・創作が行われており，原本の推定が困難であるものも多い．この観点からは，多くの分析が校訂済み本文に基づ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．計量的な分析手法は異なる作品を区別する手法を提案しているが，それには同一作品中の異本のばらつきが，作品が異なることのばらつきよりも十分小さいことが前提条件となる．

ところでこれらの研究が対象とする文学作品は，一部データベースが整備されている作品に偏っており，中古の日記文学に関しては検討が見られない．そこで本稿では『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 1」とする．また，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，その特徴を抽出することを「目的 2」とする．分析手法と分析指標に関して，2 章・3 章でそれぞれ述べた後[1]，4 章で，目的 1 に対応して，『和泉式部日記』の系統づけに関して，既往の文献学的知見と本稿で提案する手法によって得られた結果の比較を行う．合わせて，目的 2 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』）内での異本による差異と作品の違いによる差異の比較を合わせて検討する．

2．計量分析手法

本稿では，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．本節では分析手法に関して述べる．ここで用いた分析手法は一般的なものであり，他の作品の分析にも用いることができる．

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

1　形態素解析結果を修正する際に問題に感じた点は 2.3 節にまとめた．

られるなんらかの統計量を得る必要がある．例えば師（2011）は「内的証拠による比較」を，「表記の特徴」，「文字・単語列などの共起関係」，「構造分析」に分類している．本稿ではこのうち「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扱う．具体的な統計量に関しては，次節の分析条件を参照されたい．

2.1 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 }

底本と異なる箇所のみを上記のように明示的に示すことで，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}

のように書くことにする．@ 以降の部分を取り出すと底本になり，[t1] 以降で「{[」もしくは「@}」が現れるまでの部分を取り出すと，t1 となる．底本と異本が同一の場合には，異本の記述を省略する．

2.2 n-gram 分析

文章を分析する上で基本となるのは，文字あるいは単語の連鎖を確率で表した n-gram である．日本語のような膠着語に対して単語の n-gram 分析を行うには，次に述べる形態素解析により文を形態素列に分割する必要がある．例えば「高速」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，「道路」が続く確率は，「自動車」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ると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使える．別の観点としては，対象の文章から n-gram の確率を計算することで，その文章の癖を見出すこともできる．

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』の写本の一つである「三条西家本」で学習した場合の確率値を表 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 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，高次になるにつれ，出現頻度が低く空間がスパースになるため，両者を区別するのに有用である．

『更級日記』「定家本」で学習した場合の確率値を，表 3 に示す．当然のことではあるが，表 1 と表 2 の差異に比べて，表 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ので，この異なり具合をうまく定量的に評価することができれば，書き手による違いを表現できると考えられる（近藤，近藤 2001a; 谷本 2001; 土山，村上 2012）．

評価データにある n-gram が，学習データに存在しなかった場合には，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ため，なんらかの確率値を予測して与える必要がある．特に高次の 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ので，それより低次の 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」

表 1：3-gram の例（『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	こともたまは
0.143	こともめ

表 2：3-gram の例（『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	こともたまはせ

表 3：3-gram の例（『更級日記』定家本）.

確率値	3-gram
0.18	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と予測される．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 back-off 係数を学習する．

n-gram 分析を行う際には，「漢字かな交じり」で行うものと，すべて「かな」で行うものがある．中国の漢字文献の分析には，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，同一の本文であっても，漢字 / かなの揺れが発生するため，漢字かな交じり文を扱うと問題を生じることもある．このため，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（近藤 2000; 近藤，近藤 2001a）．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ことがあるため，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，すべて「ひらがな」で表したとしても，清音と濁音を区別することはできない．一方，漢字かな交じりには，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，優劣は決め難いが，本稿では写本の性質をより反映すると考えられる漢字かな交じりで分析した．

2.3 形態素解析の（古典語解析上の）問題点

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る．英語などの分かち書きされている言語ではこれは不要であるが，日本語などの膠着語では n-gram 分析の前処理として形態素解析が必要となる．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」を区別できる．ただし形態素解析を正書法の整っていない古典文学作品（特に校訂されていない本文）に適用すると，以下に述べるようないくつかの問題がみられた．

2.3.1 粒度

形態素解析は，字面の文字単位で行うのが普通であり，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[2]．しかし本当に文字単位で十分なのだろうか．音素や音節単位で行えば，より細かい単位で分析することができる．伊藤（2002）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[3]．

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）」のような誤った形態素解析結果が得られる．これは中古 UniDic が「わが」を連語とし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）+

2　形態素の認定基準には，主なものに長単位と短単位によるものがある．本稿では UniDic が採用している短単位を基本としている．短単位はゆれが少ないといわれているが，全くないということはない．あまりに形態素の単位を短くしすぎると，実用的でないためである．実際，短単位の形態素解析器であっても，かなり長いものも登録してある．たとえば UniDic では「木造」は 1 形態素だが，「レンガ造」は「レンガ + 造」となる．「木造」も「木 + 造」とすることもできるが，そうはしていない．本節での論点は，長単位・短単位といった単位のいずれかを用いるのがよいということを主張することを意図するのではなく，それだけでは一意に解釈するのが難しい場合があり，それが古典語の場合には現代語よりも問題になりやすいということを問題提起するところにある．

3　形態素解析器の単純な誤り（既存の短単位の枠組みで容易に修正できるもの）に関しては，人手で修正した．以下に述べる問題は，その修正に際して解決が困難であった問題である．

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[4]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「連語」とすれば，「我（ワガ：連語）＋身（ミ：名詞）」「我（ワガ：連語）＋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：連語）＋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，1 文字に 1 形態素を割り当てる現在の形態素解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」を連語としなくても，上述の正しい結果が得られる[5]．本稿では「わが」を連語として扱った．

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，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問題となる．いずれにせよ，最小単位を文字としていることによる限界が存在する．

2.3.2 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．2 通りの形態素解析結果を示すこともできるが，いつでも 2 通りの解釈が可能なわけでもない．「あふみち」は，「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，ひらがなでも一意には表現できない．本稿では，表の意味を優先して形態素解析を行った．

2.3.3 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，「動詞＋す・さす」で表される使役動詞は別に項を立てるなど，一貫していない点も見られる．本来，どの粒度で分析するかに関しては一貫性が必要であるが，どちらが正しいとは一概に言えないので，この部分に関しては中古 UniDic と同様の方針を取ることにした．また問題なのは，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．一般的に中古作品を校訂する際には，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．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．

2.3.4 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（金田一 1953; 関 1958）．複合動詞をどの程度認め

4　実際『旺文社古語辞典』では一語の名詞としている．
5　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．

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（大野 1956）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，「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．本稿では，「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．これに関してはより詳細な検討が必要となろう [6]．

2.3.5 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」の揺れも多いが，「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，一概にまとめることはできない．形態素解析モデルの学習の際に，このような表記の揺れを考慮する必要があるが，今回は一つずつ人手で修正した．

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．

3． 計量分析指標

本章では，2 章に記した分析手法により分析を行う際に使う指標について述べる．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（3．1 節）．形態素解析結果が得られる場合には，文体の分析指標（3．2 節）と頻度統計の指標（3．4 節）が利用できる．

3.1 文字上の分析指標

3.1.1 漢字率

「新日本古典文学大系」や「新編日本古典文学全集」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．本稿ではできうる限り忠実な本文を使っているので，漢字率を指標として用いることができる．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．徳永（1995）では，これを「用字法と書写意識」という観点で考察している．斎藤（2011）でも，漢字の含有率を指標としている [7]．

3.1.2 文字の相違率（編集距離による計量）

文字の相違率を判断するために，編集距離（Levenshtein 距離）を用いた．任意の文字

6　影山 (1993) 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．

列の間は，置換（substitution），挿入（insertion），削除（deletion）の 3 つの手順により変換が可能であり，編集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，距離として解釈したものである．編集距離は，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）コストを恣意的に設定することで，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 [8]（師 2007）という特長があるため，検索の分野でよく用いられる [9]．算出された距離行列を，デンドログラムで表現することで関係性を可視化した．編集距離の計算例を図 1 に示す．ここで文は 4 つの記号（α，β，γ，δ）からなると仮定し，文 2（sentence 2）を文 1（sentence 1）へ変換することを考える．「case 1」は，中央の β と γ が異なるのみであるので，文 2 の γ を文 1 に置換する（substitute）により変換できる．「case 2」は，一見文 1 と文 2 ですべての記号列が異なっており，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れば，α の挿入（insert）と δ の削除（delete）により，2 回の手続きで変換できる．編集距離は変換を実現する手続きに必要な最小の回数として与えられるので，この場合は 2 となる．整列には動的計画法を用いることで高速に計算できるが，長い文章同士を動的計画法で整列させると計算の時間がかかり，また大きくずれてしまうこともあるので，文単位程度の比較的短い単位であらかじめペアを作る必要がある [10]．

ただし古典語の文章は表記のゆれが大きく，ひらがなが多いため，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，編集距離は 4 となる．一方で，「き（来）たまふ」と「き（着）たまふ」は，編集距離は 0 であるが，意味は異なる．このように表記のゆれにより，編集距離が本文の内容の乖離度を代表しない可能性がある．

3.2 文体の分析指標

文体の分析に，名詞の比率など抽象化された定量的な指標が有効であることはよく知られている [11]．小林（2005）では，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 [12] は，古典語の分析においてはほとんど意味をなさないので，ここでは古典の分析にも適用可能な以下の 3．2．1 から 3．2．5 までの 5 つの指標を用いた．これに，自立語率と 3．2．6 から 3．2．8 までの 3 種類（9 つ）の指標を加えた計 15 種類の指標を用いた．

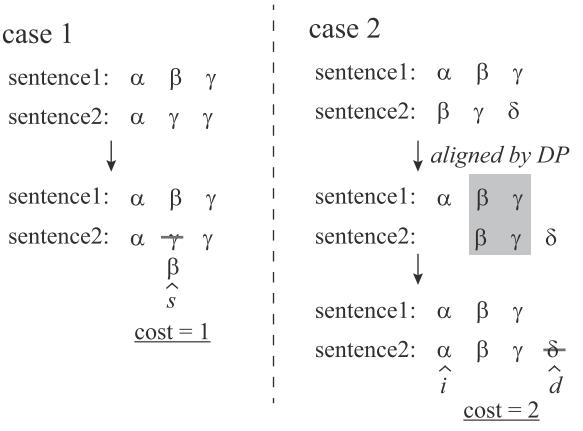
7　同文献では，さらに，「字母」と「改行位置」も特徴量として使っているが，字母は見た目や個人性の影響を大きくうけ，改行位置は紙や字の大きさから定まる物理的制約の影響があるため，漢字率よりはばらつきが大きくなると考え，本稿では採用しなかった．

8　本稿では，「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．

9　さらに文字の入れ替わりを考慮した Jaro-Winkler distance が，聖書の異本をとらえるのに適用されている (Miyake 2013)．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 (Cohen, Ravikumar, and Fienberg 2003) にまとめられている．

10　ここでは文を単位として処理をしているが，ある程度アライメントが整っている単位であれば何でもよく，たとえば段落単位でも構わない．句読点は編集距離の計算対象としていないので，文認定の結果が分析結果に影響を与えることはほとんどない．たとえば，写本 1 で「A．B．C」となっているところが，写本 2 では「A，B．C」となっていても，分析結果に影響を与えることはない．

図 1：編集距離の計算例.



3.2.1 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．樺島（1961）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）に示すように，名詞数を式（1）で求められる自立語数で除して求める．

$$\text{自立語数} = \text{全単語数} - \text{助詞数} - \text{助動詞数} \tag{1}$$

$$\text{名詞率} = \frac{\text{名詞数}}{\text{自立語数}} \times 100[\%] \tag{2}$$

3.2.2 Modifier Verb Ratio（MVR）

MVR は，式（3）により求められる．MVR とは，「形容詞・形容動詞・副詞・連語」（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（小野，田中，持尾 2007）といわれる．

$$MVR = \frac{\text{形容詞・形容動詞数} + \text{副詞数} + \text{連語数}}{\text{動詞数}} \times 100[\%] \tag{3}$$

3.2.3 指示詞の比率

文中に含まれる指示詞の割合を，式（4）により求める．指示詞は適切に使われていれ

11　文献（川崎 1967; 安本，本多 1981）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．

12 「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．例えば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（村田 2007）が，古典語においてはほとんど使われない．

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．

$$\text{指示詞率} = \frac{\text{指示詞数}}{\text{自立語数}} \times 100[\%] \tag{4}$$

3.2.4 文の長さ

文の長さを式（5）により求める[13]．文の長さには作品の特徴が現れる（樺島 1953）．『伊勢物語』などの歌物語は極端に一文が短く，『源氏物語』などの女流物語文学や日記文学は一般に長い．

$$\text{文長} = \frac{\text{自立語数}}{\text{全文数}} [\text{語} / \text{文}] \tag{5}$$

3.2.5 引用文の比率

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，中古の文章にはカッコは付されていないため，引用部分であると思われる部分にカッコを付して引用率を算出した．古典文学にこの指標を厳密に適用することは，引用文の認定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[14]としている．全体の文章に占める引用部分の割合を式（6）により求める．

$$\text{引用率} = \frac{\text{引用・会話・和歌文字数}}{\text{全文字数}} \times 100[\%] \tag{6}$$

3.2.6 心情表現の比率

引用率と関連するが，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[15]，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別に特定し，式（7）により求めた．

$$\text{心情率} = \frac{\text{心情表現文字数}}{\text{全文字数}} \times 100[\%] \tag{7}$$

3.2.7 各種品詞の比率

名詞以外にも，代名詞・形容詞・形状詞・副詞・動詞の比率を指標に加えた．

3.2.8 語種の比率

語種は和語・漢語・外来語・混種語の 4 つがあるが，今回の分析では外来語は出現しないので，和・漢・混種の 3 種類に関してその出現頻度を比較した．

3.3 n-gram 分析の類似性（パープレキシティーの利用）

2.2 に述べた通り，n-gram の類似性を指標として使うことができる（Uzuner and

13　文の認定は先学の基準に従った．鈴木（1957）を基本とした．
14　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．
15　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．

Katz2005）．形態素解析済みテキストに対して，学習するテキストを一つ選び 3-gram モデルを作成し，それ以外を評価テキストとして式（8）で表されるパープレキシティー PP を評価した．n-gram モデルの学習と評価には，SRILM[16] を用いた．

$$PP = \frac{1}{P(w_1,\ldots,w_n)^{\frac{1}{N}}} \tag{8}$$

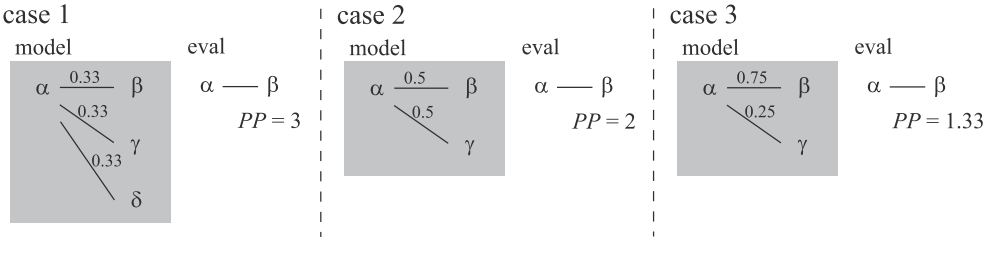
ここで P（）は，単語列 $w_1,\cdots,w_N$ が観測される確率で，PP はその相乗平均の逆数である．パープレキシティーは，次にくる単語が等確率と考えたときに，予測される単語数の平均に対応する．n-gram モデルで容易に予想可能なテキストに対しては，パープレキシティーは低くなることから，テキストの類似性が定量的に評価できると考えられる．

パープレキシティーの概念を，図 2 を用いて説明する．編集距離の説明と同様，記号は α，β，γ，δ の 4 つとする．パープレキシティー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 が来るときのパープレキシティーを計算する．「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である．この時それぞれの連鎖の確率は，1/3（=0.33..）である．パープレキシティーは，式（8）に示すように，連鎖確率の逆数であるので，パープレキシティー（=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である．この時それぞれの連鎖の確率は，1/2（=0.5）であり，パープレキシティーは 2 となる．学習データにおいて $\alpha-\beta$ の連鎖確率は「case 1」よりも「case 2」の方が高いため，パープレキシティーは小さいことが分かる．これは n-gram モデルを生成モデルと考えた場合，「case 2」の方が 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 $\alpha-\beta$，$\alpha-\gamma$ の連鎖しか見られなかった場合であるが，$\alpha-\beta$ 方が $\alpha-\gamma$ 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，$\alpha-\beta$ のパープレキシティーは，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 $\alpha-\beta$ の連鎖が起こりやすいという直観と一致している．

3.4 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（宮島達夫 1970; 深谷，山村，工藤，松本，竹内，

図 2：パープレキシティーの計算例．



16 http://www.speech.sri.com/projects/srilm/ より，ダウンロードできる．（2014 年 7 月 15 日確認）

大西 2004）が行われている．ここでも頻度統計を用いて，語の使われ方等を分析する．その際，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．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．

3.4.1 コサイン類似度

Bag of words は，単語ごとの頻度ベクトル $\mathrm{h}_i = (w_1, w_2, \cdots, w_N)$ を，テキストごとに求める．ここで i は各テキストのインデックス（$1 \leq i \leq I$）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により集計する．N は対象の 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 とした．このようなベクトル間の類似度を測るのには，コサイン類似度がよくつかわれる．テキスト i と j の間のコサイン類似度 c は内積の公式を用いて，

$$c = \frac{\mathbf{h}_i^T \mathbf{h}_j}{\left|\mathbf{h}_i\right|\left|\mathbf{h}_j\right|} \tag{9}$$

によって求められる．T は，転置を示す．ベクトルの Euclid 距離は

$$\left|\mathbf{h}_i\right| = \sqrt{\mathbf{h}_i^T \mathbf{h}_j} = \sqrt{w_1^2 + w_2^2 + \ldots + w_N^2} \tag{10}$$

3.4.2 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，従前より様々な研究がおこなわれている（宮田 1942）．ここでも助動詞別に検討を行った．

4．『和泉式部日記』4 異本間の関係性と『更級日記』との比較

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．

4.1 底本について

4.1.1『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いない．主に表 4 に示す 4 系統に分別されている [17]．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（鈴木，川口，遠藤，西下 1957; 近藤 2003; 清水 1981）の最も一般的な底本となっている [18]．ここでは 4 種の異本の代表的なものをもとにして，それらの関係性を探る [19]．

17　書誌情報などは，吉田（1964）を参照されたい．各写本間の関係性は，吉田（1964）の p.170 の図にまとめられている．

18　例えば大橋（1991）では，異同箇所の比較から「三条西本を善本と考え」ている．

19　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．

表 4:『和泉式部日記』の代表的な 4 異本.

略号	系統
「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

伊藤（1991）は，三条西家本を底本に，他本の異同を対校しながら翻刻したものである．ここでは，これをもとに，2.1 節にあげた形式のデータベースを作成した[20]．例えば『和泉式部日記』のはじめの部分

（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中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中をなげきつゝあかしくらすほどにはかなくて，（応 , 混）

を，異なる箇所を ¥d{} でくくった範囲内に書き記すことで，

¥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，縮約して表すことができる．

伊藤（1991）を電子化したのちに，濁点[21]，句読点を付与し，引用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[22]．その際には，文献（鈴木他 1957; 近藤 2003）等，先学の解釈を参考にした．

4.1.2『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（玉井 1925），諸本は御物本を共通の祖とすることが明らかになり，各種テキストの底本に使われている．本稿でも「御物（定家）本」を用いた．

4.2 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」（小木曽，小椋，田中，近藤，伝 2010; 小木曽 2011; 小椋，須永，小木曽，近藤，田中 2011）を「MeCab」（工藤，

20　伊藤（1991）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
（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．

21　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．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．

22　同様の作業を『更級日記』に対しても行った．

山本，松本 2004）と組み合わせた形態素解析エンジン「和文茶まめ」[23] により行った．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，2.3 節に述べたいくつかの問題があったので，人手で修正を加えた．

4.3『和泉式部日記』の系統論

川瀬（1953）により，寛元本系統の本が紹介され，今日の 3 系統を基にした理論が構築された．同文献では「新出本[24] 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．伊藤（1956）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」，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」であり，「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，伊藤（1978,1981）では，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，「寛元本は三条西本，応永本の中間的性格を示す」と述べている[25]．

これに対して，森田（1977）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．吉田（1964）もこれを支持し，「脱文・数詞・官職名等の類について，異同関係を考察」し，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」く，「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．

竹内（1986）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．全体を分析した結果と，名詞，動詞を分析した結果が三様の結論となったと報告している．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．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．なお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」であり，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．

4.4『和泉式部日記』の文体上の特徴

『和泉式部日記』は主人公の女による三人称語りである点（織田 1958）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（今井 1957）もある．また著者に関しても様々な議論があり，池田（1944）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．

23　http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%C3%E6%B8%C5%CF%C2%CA%B8UniDic よりダウンロードできる．（2014 年 7 月 15 日確認）

24　著者注 : 寛元本

25　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」．

大橋（1961）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 2 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．

神谷（1991）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．

4.5 結果と考察

4.5.1 文体の分析指標

2 章で述べた漢字率と文体の分析指標の計 16 指標に従い，分析を行った．結果を表 5 と 6 に示す．参考のため，「総文字数」と「総形態素数」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% 程大きい程度である．『和泉式部日記』の 4 異本の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』の間の，作品間の差異を表すものには，引用率・心情率・名詞率があげられる．図 3 には，三条西家本の指標で他本の指標を割った（正規化した）結果を示しており，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．地名などの固有名詞は，明らかに事実叙述的な記述

表 5：文体を表す指標の分析結果．

	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	7.20%	52.4	47.50%	12.20%	10810	52.60%	40.50%
寛元	19975	8.30%	54	46.60%	12.30%	10906	52.40%	39.40%
応永	19840	8.60%	53.3	47.40%	12.40%	10865	52.50%	41.50%
混成	20200	10.70%	54.4	46.40%	12.80%	11186	52.30%	41.60%
更級日記	26546	9.10%	66.7	33.20%	4.10%	14517	55.70%	40.30%

表 6：文体を表す指標の分析結果（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	37.10%	3.40%	8.00%	1.20%	7.40%	40.90%	98.30%	1.30%	1.00%	0.30%
寛元	37.30%	3.30%	7.90%	1.10%	7.30%	41.10%	98.30%	1.30%	1.00%	0.30%
応永	36.60%	3.20%	7.90%	1.20%	7.80%	40.80%	98.20%	1.40%	1.00%	0.30%
混成	36.70%	3.10%	7.80%	1.10%	8.00%	40.70%	98.10%	1.50%	1.00%	0.30%
更級日記	46.60%	3.60%	7.90%	1.10%	5.10%	35.10%	96.30%	2.20%	1.30%	0.30%

の中に現れる．本来，日記は記録的な色彩が強いためにこれは当然であるが，『和泉式部日記』の場合は，事実の客観的な記録よりも心情の吐露を主眼としているため固有名詞の出現数が少ない．異本間で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[26]により主要な変数 2 つを取り出した[27]．結果を図 4 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，文献学的な観点（4．3 節参照）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．

図 3：分析指標のレーダーチャート．

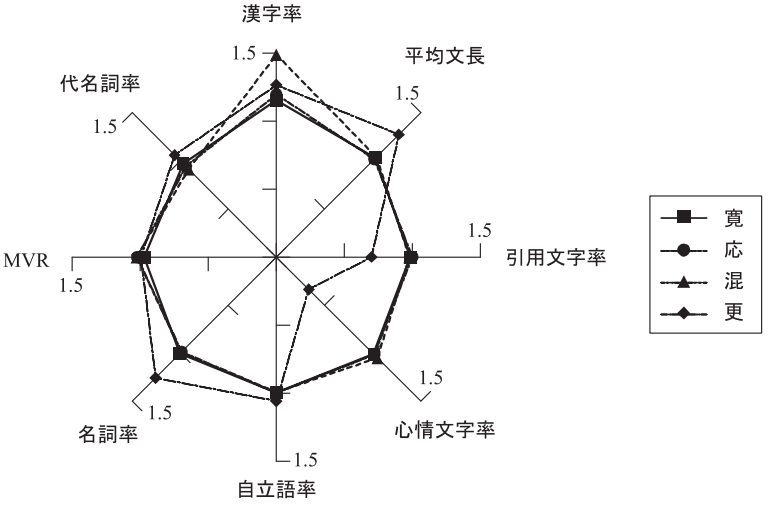
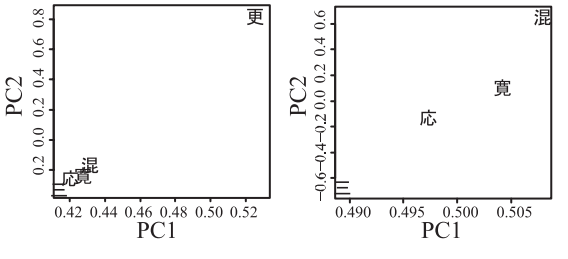


図 4：分析指標の主成分分析結果.（主成分 1(PC1) と主成分 2(PC2) を 2 次元平面上にプロット.）



26　R(http://www.r-project.org/) を用いた.
27　寄与率は 2 つの変数で 100% であるので，それ以外の変数は無視できる.

また，各指標間でダイナミックレンジに差があるため，どの写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．

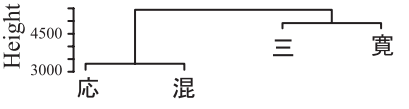
4.5.2 編集距離による分析

表 7 に，『和泉式部日記』4 異本間の編集距離を示す．距離は対称性を有するため，右三角成分は省略した．これだけでは関係性が分かりにくいので，図 5 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・文献学的な検討では一致して述べられているし，森田（1977）や吉田（1964）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」（森田 1977）との結論（4.3 節参照）ともこの分析結果は一致する．

表 7:『和泉式部日記』4 異本間の編集距離.

	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 5: 編集距離に基づくデンドログラム.



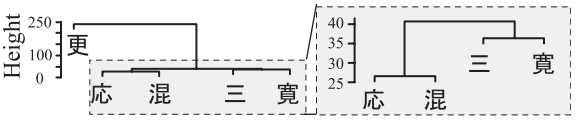
4.5.3 n-gram の分析

編集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，この方法は他作品（ここでは『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適用すると精度が低下するため，事前にある程度（文単位程度）整列させておく必要があり，それなりに手間がかかる．近藤，近藤（2001b）にあるような，ひらがな単位の（音節）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ため，漢字かな交じりの（単語）3-gram を作成し，そのパープレキシティー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 モデルを構築するだけで計算できる．表 8 に分析結果を，図 6 にそのデンドログラムを示す．異本の分類結果は編集距離の場合と同様であり，加えて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作品間のそれに比べて十分小さいことが確かめられた．

表 8：3-gram のパープレキシティーの分析結果（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 6: パープレキシティーに基づくデンドログラム.



4.5.4 頻度統計の分析

総形態素数は表 5 に示したが，表 9 には異なり形態素数を取り上げる [28]．品詞情報のあり・なしで区別したが，両者には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.

表 9: 各テキストにおける異なり形態素数.

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

単語の頻度を集計して，式（9）により，コサイン類似度を求めた．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コサイン類似度を表 10 に示す．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 程度の差異しかなく，非常に高い類似度を示している．ゆえに異本を区別する指標としては不適当で，この目的のためには単語の頻度だけではなく，パープレキシティーのように単語間の連接関係を考慮する必要があることが示された.

文体の分析を行うために，助動詞の出現数と頻度 [%] を表 11 に示した [29]．『更級日記』の品詞別の考察が宮田（1942）にあり [30]，大橋（1961）や神谷（1991）でも触れられているように，『伊勢物語』などの歌物語で特徴的な「けり」の使用頻度は，『和泉式部日記』と『更級日記』においてはそれほど高くない．「き」「けむ」の使用頻度は，『更級日記』

28 『和泉式部日記』の語彙に関する研究には，竹内（1963）がある.

29 「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.

30 索引（西端，木村，志甫 1996）も利用できる.

の方が 2 倍から 5 倍程度高いのに対して，「めり」「らむ」の使用頻度は，『和泉式部日記』の方が 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いえる．その他に目立った差異としては，「させる（さす）」「せる（す）」の頻度が『和泉式部日記』の方が高いことがあげられる．両作品と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のに対して，『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のだろう．

表 10: Bag of words のコサイン類似度（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	–	–	–
応永	0.9977	0.9985	1	–	–
混成	0.9982	0.998	0.9984	1	–
更級	0.9147	0.9187	0.9181	0.9137	1

表 11: 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					頻度(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	1.4	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

めり	26	25	26	27	14	0.24	0.23	0.24	0.24	0.1
らむ	31	31	33	34	13	0.29	0.28	0.3	0.3	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	0.2	0.22	0.17	0.21	0.21
れる(る)	37	44	40	41	55	0.34	0.4	0.37	0.37	0.38
合計	10810	10906	10865	11186	14517	100	100	100	100	100

5．まとめ

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．記録的な文学では，事実や固有名詞の記述が中心となるため必然的に名詞の使用が増え，内情吐露的な文学では，心情の直接表現が増すと考えられるからである．これにより日記文学の特質に関して，ある程度分析を加えることができた．ただし，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が，指標に恣意性があり，一般化と定量的な分析が難しいので，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，従来の計量分析によく用いられているこれらの指標の主成分分析結果による分類結果は，文献学の先学による知見と一致しなかった．

次に，自然言語処理の分野で用いられている指標である編集距離とパープレキシティーを使って，客観的な評価を行った．これより，異本間の異なり度を測る指標としては，分析指標の主成分分析よりも，文字列間の編集距離やパープレキシティーが有効であることが分かった．編集距離とパープレキシティーどちらを用いた場合でも，4 つの異本に対する系統の分析結果は，文献学の先学の知見と一致した．これにより異本の分析に関しても，有効な手法を提案できたと考えられる．特にパープレキシティーを用いることで，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，結果として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相当に小さいことが定量的に確かめられた．

後者の分析手法が有効であった理由に関して私見を述べる．異本の分類は，基本的には差異の部分に注目して行う．前者の主成分分析は，本文全体からただ一つ得られた「〜率」といった指標をもとにしている．これは本文全体の特徴を抽出するのには役立つが，これらの指標を算出した時点で，局所的な差異は失われてしまっているので，文献学的な検討には向かないと考えられる．これに対して，後者の編集距離やパープレキシティーは，主に局所的な差異に着目した分析手法なので，文献学での検討と同じような点に注目することができているのではないかと考えられる．

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．「歌物語」と「日記文学」の関連も，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であると考えられる．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おり，通常の解析とは異なるアプロー

チが必要になると考えられる．

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．

文献

Brinegar, C. (1963) Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship, Journal of the American Statistical Association, 58: 85–96.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003) A Comparison of String Distance Metrics for Name-Matching Tasks, in Proceedings IJCAI-03 Workshop on Information Integration, 73–78.

Grimmer, J. and Stewart, B. M. (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, Political Analysis, 1–31.

Mingzhe, J. and Minghu, J. (2012) Text Clustering on Authorship Attribution Based on the Features of Punctuations Usage, in Proceedings International Conference on Signal Processing, 3: 2175–2178.

Miyake, M. (2013) Different Characteristics of Variant Readings Based on Comparison of Major Textual Similarity Measures, in Proceedings Japanese Association for Digital Humanities (JADH).

Uzuner, O. and Katz, B. (2005) A Comparative Study of Language Models for Book and Author Recognition, in Proceedings International Joint Conference on Natural Language Processing (IJCNLP).

池田亀鑑 (1944)『平安時代文学概説』．八雲書店.

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.

伊藤博 (1956)「和泉式部日記諸本の系統について」『国語』, 4(4).

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.

伊藤博 (1981)『和泉式部日記伝本攷』．桜楓社.

伊藤雅光 (2002)『計量言語学入門』．大修館書店.

今井卓爾 (1957)『平安時代日記文学の研究』．明治書院.

大野晋 (1956)「基本語彙に関する二三の研究」『国語学』, 24: 34–46.

大橋清秀 (1961)『和泉式部日記の研究』．初音書房.

大橋清秀 (1991)『和泉式部日記本文の研究』．和泉書院.

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 (人文科学とコンピュータ)』, CH-85: 1–8.

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」『情報処理学会研究報告 (人文科学とコンピュータ)』, CH-92: 1–4.

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』, 312–315.
織田裕子 (1958)「「和泉式部日記」の作者について」『国語国文』, 27(4).
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』, 27–36.
影山太郎 (1993)『文法と語形成』. ひつじ書房.
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.
樺島忠夫 (1961)「文体の変異について」『国語国文』, 30(11).
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.
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.
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.
北研二 (1999)『確率的言語モデル』. 東京大学出版会.
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.
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.
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.
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)』, 89–96.
小林千草 (2005)『文章・文体から入る日本語学』. 武蔵野書院.
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.
近藤みゆき (2003)『和泉式部日記』. 角川文庫.
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.
近藤泰弘，近藤みゆき (2001a)「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』, 2: 50–55.
近藤泰弘，近藤みゆき (2001b)「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」『言語処理学会第 7 回年次大会発表論文集』, 209–212.
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.
清水文雄 (校注) (1981)『和泉式部日記』. 岩波文庫.
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.
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.
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.
竹内美智子 (1986)『平安時代和文の研究』. 明治書院.
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』, 2: 56–58.
玉井幸助 (1925)『更級日記錯簡考』. 育英書院.
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.
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.
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.
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.
堀川貴司 (2010)『書誌学入門古典籍を見る・知る・読む』. 勉誠出版.
宮島達夫 (1970)「古典の品詞統計」『計量国語学』, 53: 1–8.
宮田和一郎 (1942)「更級日記の語法的研究」『国語文化』.
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.
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』, 55: 311–326.
森田兼吉 (1977)『和泉式部日記論攷』. 笠間書院.
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」『CHISEConference 2005 報告書 & CodeFest 京都 2005 資料集』, 1–7.
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.
安本美典, 本多正久 (1981)『因子分析法』. 培風館.
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.

（2013 年 12 月 17 日受付, 2014 年 5 月 30 日再受付）

Mathematical Linguistics, Vol.29 No.6 (September 2014) pp.187-210.

Paper

Metric Japanese Study on Diary Literature in the *Heian* Period and Objective Measurement on Relationship between Variants

Analyses on "*Izumishikibu nikki*" and "*Sarashina nikki*"

Yuuki Tachioka (Mitsubishi Electric Corporation)

Abstract: Stylometrics analyzes the style of texts based on some metric features. These methods have been mainly applied to modern Japanese texts, and shown its effectiveness especially for authorship attribution. However, when these methods are applied to classical literature texts, existence of variants for the same work causes problems because there are many variants for them, which rarely have an original text, and sometimes these variants are greatly different from the original one. This paper validates a method that represents a relationship between variants quantitatively, using edit distance or perplexity. Experiments on "*Izumishikibu nikki*", which is one of the most popular diary works in the *Heian* period, shows that the proposed method has a better correspondence to the results shown in the previous bibliographical studies, compared to the conventional principal component analysis using multiple metric features. Furthermore, comparison with "*Sarashina nikki*", which is another diary work in the *Heian* period, confirms that the difference between variants for the same work is much smaller than that between different works.

Keywords: Variants, Stylometrics, Edit distance, N-gram model, Perplexity