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)(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(名詞)+の(格助詞)+葉(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(名詞)」あるいは「葉(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(名詞)」とするのは、単独で「葉」と読まれることは無いので抵抗がある。「葉(接尾辞)」とした場合には「落ち葉」も「落ち(動詞)+葉(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+道」と「近江路」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。

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

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 perplexity PP 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。

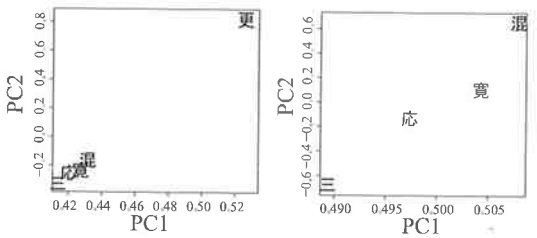

図1 分析指標の主成分分析結果

図2 Levenshtein距離に基づくデンドログラム

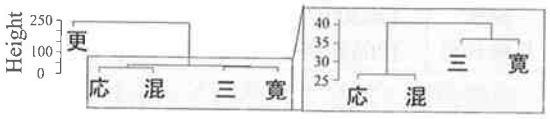

図3 Perplexityに基づくデンドログラム

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.