Vol.2020-CH-122 No.3
2020/2/1

『源氏物語』及びその補作における特徴語句抽出の試み

土山玄[1]

概要：本研究では『源氏物語』の補作として『山路の露』及び『雲隠六帖』を採り上げる。これらの補作と称される作品と『源氏物語』との間において出現率が顕著に相違する語句、すなわち特徴語句の抽出を行う。特徴語句の抽出では TF-IDF やカイ二乗統計量、あるいはランダムフォレストなどの機械学習の手法が用いられることが多い。まず、本研究ではこれらの抽出手法を概観し、それぞれの手法の特徴について検討を加える。次いで、『源氏物語』と補作の特徴語句の抽出結果に基づき、『山路の露』及び『雲隠六帖』の単語の出現傾向の特徴について検討する。

キーワード：源氏物語　計量文献学　特徴語の抽出　マハラノビス距離

Extracting Feature Words Between “The Tale of Genji” and Apocryphal Texts

Gen TSUCHIYAMA[†1]

Keywords: The Tale of Genji, Stylometry, Feature words, Mahalanobis' generalized distance

1. はじめに

『源氏物語』は平安時代に紫式部の手によって執筆されたとする全 54 巻から構成される長編物語であり、後世の文学へも大きな影響を与えた。実際に、『源氏物語』の第 54 巻「夢浮橋」以降の続編が後人の手によって創作されており、そのような作品は補作と称されることが多い。このような補作の中で『山路の露』及び『雲隠六帖』は特に著名である。

そこで、本研究では『源氏物語』とその補作である『山路の露』と『雲隠六帖』を研究対象として取り上げ、補作である『山路の露』と『源氏物語』との間、『雲隠六帖』と『源氏物語』との間において出現傾向が顕著に相違する単語を用いて抽出する。

2 つの文献、あるいは 2 つの文学作品の間に出現傾向が顕著に異なる単語は特徴語と称され、このような特徴語を抽出する手法は多数ある。近年では 2 つの群に分析対象を分類する際に変数重要度を計算できる決定木、決定木のアンサンブル学習であるランダムフォレストなどのデータサイエンスの手法が用いられることが多い。また、TF-IDF やカイ二乗検定における統計量を利用することで特徴語の抽出を行うこともある。しかし、本研究では線形判別分析において用いられるマハラノビス距離を用いて特徴語の抽出を行った。これらの特徴語の抽出方法については後述する。

2. データ

『源氏物語』の写本系統は青表紙本系、河内本系、別本と 3 系統に大別される。本研究では、青表紙本系の大島本を主な底本とする『源氏物語語彙用例総索引 自立語編』[1]及び『源氏物語語彙用例総索引 付属語編』[2]を電子化したデータベースを分析に利用した。この『源氏物語語彙用例総索引』は『源氏物語』の本文すべてについて、形態素解析を行ったものである。なお、形態素解析については、『源氏物語大成 索引篇』[3]の単語認定基準に準拠している。

補作の資料は次の通りである。分析に使用した『山路の露』の本文は『日本古典全書源氏物語 7』[4]に付されたものである。これは玖山九條稙通（1507-1594）の自筆本を池田亀鑑が校定した。次に、『雲隠六帖』は『源氏物語の研究』[5]に付されたものを底本とした。『源氏物語の研究』によれば、これは某家に所蔵される近世中期に書写された写本の本文を翻刻したものとされる。分析において用いるテキストデータは、これら文献における各作品の本文を『源氏物語大成 索引篇』の単語認定基準に準拠し、単語分割されたものである。すなわち、本研究において分析に使用する『源氏物語』及び補作 2 作品のデータベースは同一の基準によって単語認定がなされている。

3. 特徴語の抽出

特徴語の抽出方法は主に 2 つに大別される。外的基準ありの方法と外的基準なしの方法である。外的基準ありの手法は主に機械学習における分類器が用いられ、決定木やランダムフォレストがこれに該当する。ランダムフォレストは特に分類精度が高く、様々な分野で用いられる手法である。しかし、ランダムフォレストなどの手法は大規模デー

1 お茶の水女子大学文理融合 AI・データサイエンスセンター
Center for Interdisciplinary AI and Data Science, Ochanomizu University

タ、すなわちビッグデータを分析することを想定した分析手法であり、本研究で採り上げるような対象数の少ないデータに対しては必ずしも有効であるとは言えない。補作の1作品対『源氏物語』54巻となるため、意味のある分析ができない可能性がある。

その一方で、TF-IDFやカイ二乗検定は外的基準なしの方法となる。外的基準がないデータセットの場合、群が想定されていないため群間における出現傾向が顕著に相違する特徴語を抽出することはできない。そのため、分析対象となる個体別に特徴語を求めることになる。まず、TF-IDFについて概観する。TFはTerm Frequencyの頭文字であり、ある文献における単語の出現率を意味する。IDFはInverse Document Frequencyの頭文字であり、逆文書頻度とも称される。TF及びIDFは以下の数式で求められ、TF-IDFはTFとIDFの積である。

$$TF = \frac{テキストXにおける単語wの頻度}{テキストXの述べ語数}$$

$$IDF = \log \frac{テキスト数}{単語wを含むテキスト数}$$

古典文学作品を対象にTF-IDFを求めると、テキスト数と単語wを含むテキスト数がおよそ同じ値になることが多く、次に、TF-IDFは単語の出現率と大きく乖離しない値になることが多い。

次に、カイ二乗検定を用いる方法である。これは表頭をテキストXとそれ以外のテキスト、表側を単語wとそれ以外の単語とした2行2列のクロス集計表を作り、そのクロス集計表に対しカイ二乗検定を行い、その結果として求められたp値を特徴語の抽出に用いる方法である。すなわち、p値が小さいほど特徴語であると言える。この方法では2行2列のクロス集計表を用いることから、テキストX以外のテキストにおける出現頻度の分散が考慮されないという問題がある。

本研究では『源氏物語』と補作という外的基準があるが、『源氏物語』の対象数が54となるのに対し、補作の対象数が1であるため、外的基準ありの方法を用いることができず、外的基準なしの方法を用いる必要がある。先に触れたように、本研究ではTF-IDFやカイ二乗検定ではなく。マハラノビス距離を用いて特徴語の抽出を行う。マハラノビス距離は線形判別分析において、個体と群の距離を測る際に用いられる。マハラノビス距離の特徴として、対象となる群の平均値だけではなく、分散も用いる点がある。

4. 分析

4.1 『源氏物語』と『山路の露』の特徴語

本研究では品詞別に特徴語の抽出を行った。分析に用いた品詞は名詞、代名詞、動詞、形容詞、形容動詞、副詞、連体詞、助詞、助動詞の9品詞である。本稿では名詞、助詞、助動詞において抽出された特徴語について報告する。名詞は異なり語数が5000語を超え、また『源氏物語』の54巻に『山路の露』を加えた55の対象において、ごく少数の対象でのみ出現する単語も存在することから本研究では出現頻度上位100語を採り上げ、各単語の出現率について『源氏物語』全54巻と『山路の露』とのマハラノビス距離を求めた。助詞、助動詞についてはすべての単語を対象とし、『源氏物語』全54巻と『山路の露』とのマハラノビス距離を求めた。

まず、表1は名詞の特徴語10語である。これら10語はマハのビス距離の絶対値を求め、降順に整列した結果である。また、表2は助詞の、表3は助動詞の特徴語である。

表1　『山路の露』における名詞の特徴語

単語	マハラノビス距離
ユメ/名詞	24.179
アマギミ/名詞	22.501
ソデ/名詞	11.544
サマ/名詞	8.527
コロ/名詞	6.592
ノチ/名詞	5.813
カタ/名詞	4.965
ウコン/名詞	4.947
アハレ/名詞	4.331
ココチ/名詞	3.408

表2　『山路の露』における助詞の特徴語

単語	マハラノビス距離
ドモ/助詞	162.277
ヘ/助詞	25.938
ダニ/助詞	8.253
カ/助詞	6.531
ツツ/助詞	4.855
ガ/助詞	4.267
ド/助詞	3.535
トモ/助詞	2.718
ナガラ/助詞	2.427
デ/助詞	2.412

表3　『山路の露』における助動詞の特徴語

単語	マハラノビス距離
キ/助動詞	3.571
リ/助動詞	1.341
ラル/助動詞	1.269
ケム/助動詞	1.113
ツ/助動詞	0.780
ナリ/助動詞	0.670
ジ/助動詞	0.502
ベシ/助動詞	0.469
ス/助動詞	0.447
ヌ/助動詞	0.336

表 1 に示したように、名詞では「ユメ」「ソデ」「サマ」という単語が偏って『山路の露』に頻出していると言える。

他方、助詞においては表 2 に示したように「ドモ」「ヘ」「ダニ」が偏って『山路の露』に頻出していることが明らかになった。特に、「ドモ」はマハノビス距離が大きく、『山路の露』にとって重要な特徴語であると考えられる。表 3 に示したように助動詞においては「キ」「リ」「ラル」などが特徴語であると指摘できるが助詞に比べるとマハラノビス距離が総じて小さくなっている。

表 4　『雲隠六帖』における名詞の特徴語

単語	マハラノビス距離
ユメ/名詞	26.360
ホトケ/名詞	19.823
ムカシ/名詞	8.725
コノヨ/名詞	6.818
ココロ/名詞	6.242
ヨ/名詞	5.032
ホド/名詞	4.617
ミチ/名詞	3.786
カタ/名詞	3.570
アリサマ/名詞	3.528

表 5　『雲隠六帖』における助詞の特徴語

単語	マハラノビス距離
ヘ/助詞	11.693
ドモ/助詞	9.741
ナド/助詞	8.384
ド/助詞	7.995
ソ/助詞	7.664
マデ/助詞	7.622
ゾ/助詞	4.928
バ/助詞	4.017
ナガラ/助詞	3.633
デ/助詞	3.204

表 6　『雲隠六帖』における助動詞の特徴語

単語	マハラノビス距離
タリ/助動詞	11.945
キ/助動詞	7.299
ケリ/助動詞	7.076
メリ/助動詞	5.035
ズ/助動詞	3.776
ラル/助動詞	0.918
ベシ/助動詞	0.736
ヌ/助動詞	0.712
ナリ/助動詞	0.540
リ/助動詞	0.493

4.2 『源氏物語』と『雲隠六帖』の特徴語

次いで、『源氏物語』と『雲隠六帖』との間における特徴語の抽出においても同様に 9 品詞を対象として分析を行った。ここにおいても名詞、助詞、助動詞の分析結果について報告する。表 4 は『源氏物語』と『雲隠六帖』との間における名詞の特徴語である。『山路の露』と同様に『雲隠六帖』においても「ユメ」が『源氏物語』に比べて多用されている。

次に、表 5 は助詞の特徴語、表 6 は助動詞の特徴語である。『山路の露』の特徴語の抽出結果では、上位の特徴語はどの単語も『山路の露』に多用されていた。しかし、表 5 では「ナド」及び「ド」が、表 6 においてはマハラノビス距離が最大となった「タリ」、そして「メリ」が『雲隠六帖』におよそ用いられていないことが明らかになった。

5. おわりに

古典文学作品を対象とした特徴語の抽出に、本研究ではマハラノビス距離を用いた。従来のカイ二乗検定では分散を考慮して特徴語を抽出することは難しかったが、マハラノビス距離を用いることで、データの分散を考慮した特徴語の抽出が可能になったと考えられる。

また、『源氏物語』と『山路の露』及び『雲隠六帖』という補作 2 作品を対象として特徴語を抽出した結果、名詞ではどちらの補作においても「ユメ」という単語が共通して特徴語となることが明らかになった。次いで、『山路の露』と『源氏物語』との間では『山路の露』に偏って頻出する単語が認められたが、その一方で『雲隠六帖』と『源氏物語』との間においては『雲隠六帖』におよそ用いられない単語が特徴語として抽出された。

このように、特徴語を抽出することによって、作品間の文章の特徴を計量的に可視化することができると考えられる。今後は、このような特徴語を用いたより発展的な統計解析が行われることが期待される。

謝辞　本研究は JSPS 科研費 19K20627 の助成を受けたものです。

参考文献

[1] 上田英代, 村上征勝, 今西祐一郎, 樺島忠夫, 上田裕一. (1994). 『源氏物語語彙用例総索引 自立語編』, 勉誠出版.
[2] 上田英代, 村上征勝, 今西祐一郎, 樺島忠夫, 上田裕一, 藤田真理. (1996). 『源氏物語語彙用例総索引 付属語編』, 勉誠出版.
[3] 池田亀鑑, (1985). 『源氏物語大成 索引篇』, 中央公論社.
[4] 池田亀鑑. (1955). 『日本古典全書 源氏物語 7』, 朝日新聞社.
[5] 長谷川和子. (1957). 『源氏物語の研究』, 東宝書房.
[6] 土山玄, 村上征勝. (2011). 「源氏物語と宇津保物語における語の使用傾向について」, 人文科学とコンピュータシンポジウム論文集, 情報処理学会, Vol. 2011, pp. 125-132.
[7] 金明哲. (2009). 『テキストデータの統計科学入門』, 岩波書店.

付録

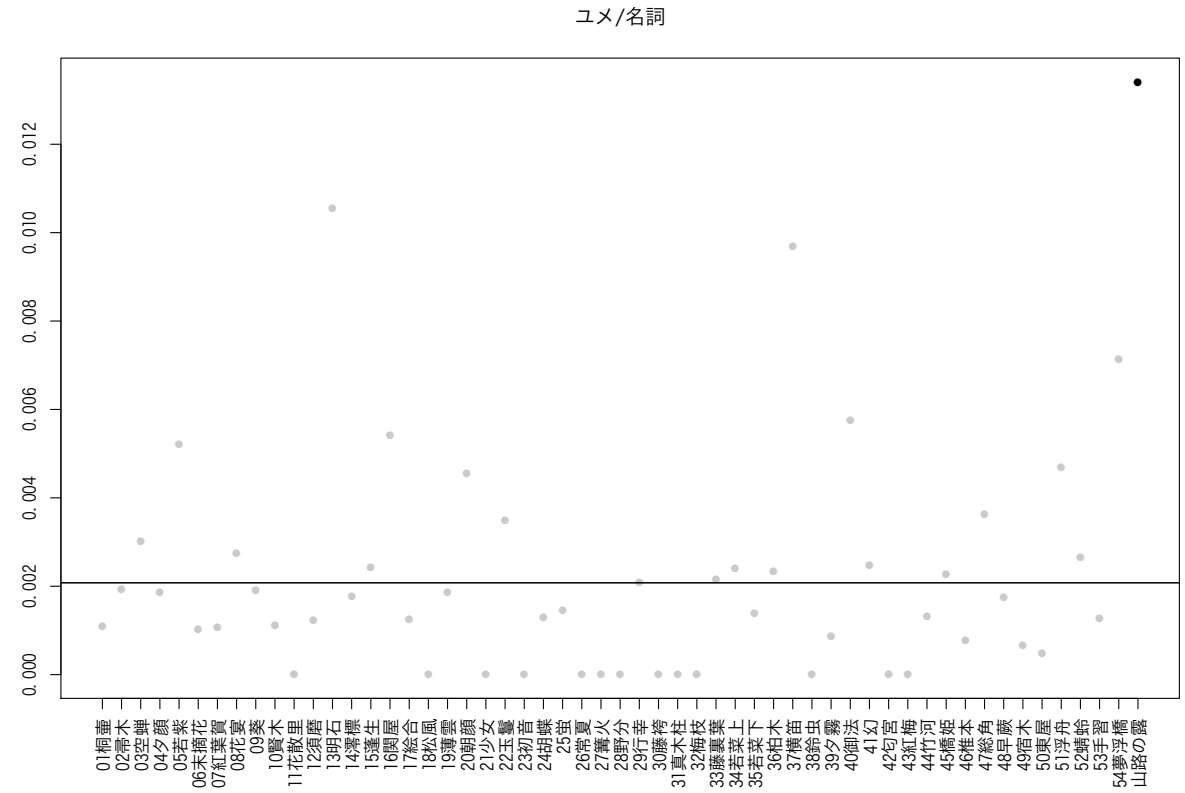

図 1　『源氏物語』と『山路の露』における「ユメ」の出現率

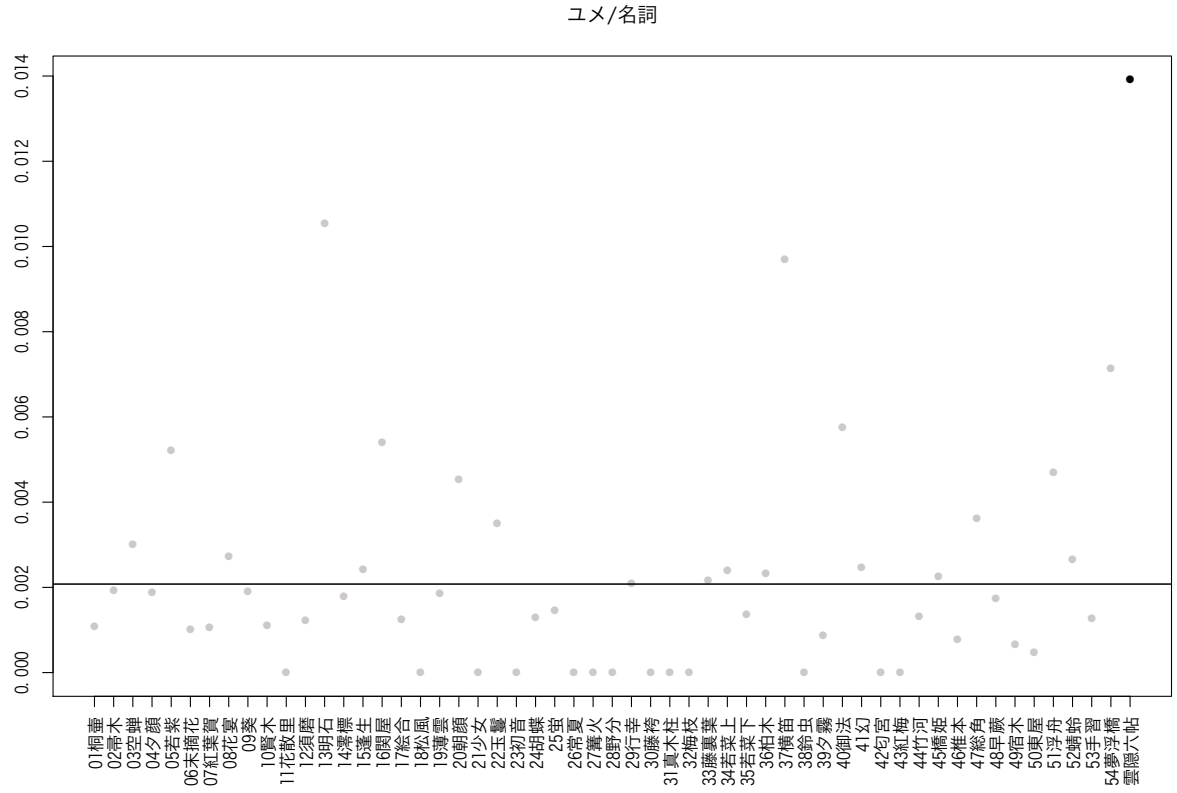

図 2　『源氏物語』と『雲隠六帖』における「ユメ」の出現率

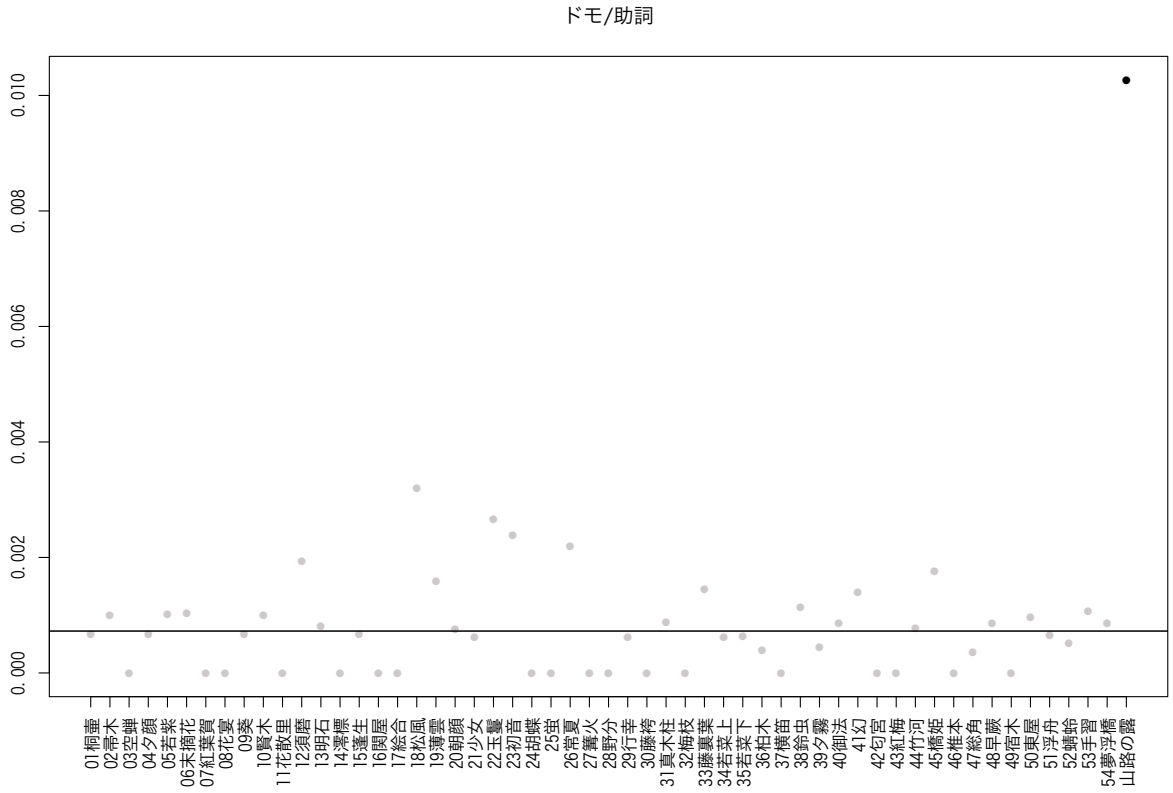

図３　『源氏物語』と『山路の露』における「ドモ」の出現率

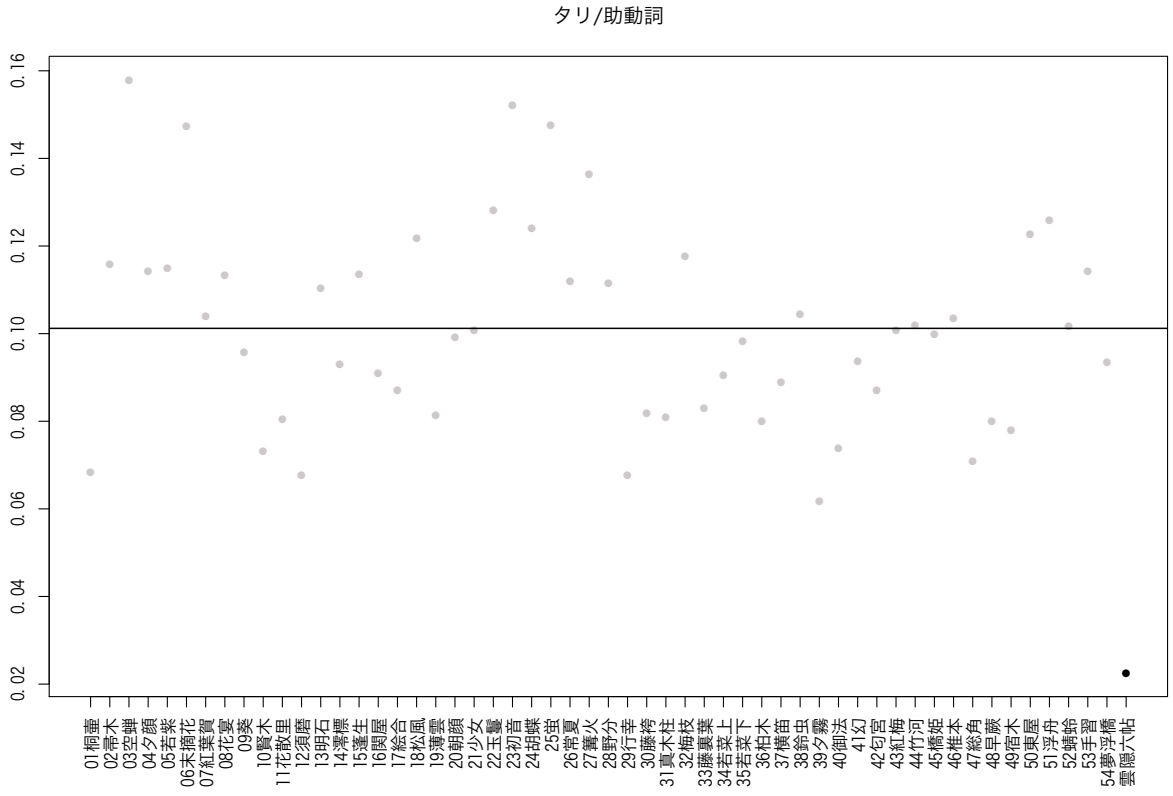

図３　『源氏物語』と『雲隠六帖』における「タリ」の出現率