

語彙の豊富さの応用可能性

植 田 麦

要旨

本研究は、語彙の豊富さを中心とした変数を用いて、有価証券報告書・小説・所信表明演説の3種類のテキストを分析するものである。研究の主な目的は以下の2点である。

1. 稿者が先に研究した際の語彙の豊富さの値を他のサンプルと比較し、相対的位置を確認する。
2. 語彙の豊富さの応用可能性を示す。

研究では、語彙の豊富さ (s)・名詞率・MVR・平均文長を変数として使用した。考察の結果、以下の知見が得られた。

- 有価証券報告書の語彙の豊富さ (s) は約 .850 で、これは学生のレポートと同程度である。
- 小説では、作家によって文体の特徴にちがいがあり、特に江戸川乱歩の作品では低年齢層向けと非低年齢層向けで明確な差異がみられた。
- 所信表明演説では、時代が下ると MVR が低下し、より理解しやすい文体に変化している。

本研究では、語彙の豊富さを中心とした複数の変数の組み合わせによる分析可能性を提示した。

はじめに

稿者はこれまでに、いくつかの研究において語彙の豊富さを指標のひとつとして用いてきた。たとえば、大学1年生および2年生を主たる対象としたアカデミックライティングの授業において課したレポートについて、その教育効果

の計測を試みた研究（植田麦：2021 および植田：2022）では、語彙の豊富さの指標 s （注 1）を用いた。当該研究では、受講当初に執筆したレポートに対して受講から一定期間を経たのちに執筆したレポートでは、 s の減少傾向が観測された。しかし、これは語彙が貧弱になったのではなく、アカデミックライティングとして不適切な語彙がふり落とされた結果であることが明らかとなった。いってみれば、トレーニングの結果、余分なものがそぎ落とされたのである。

以上は一例ではあるが、語彙の豊富さは研究上さまざまな利用が可能である。しかし、サンプルを横断する研究は、存外少ない。たとえば、上に示した稿者の研究では、語彙の豊富さ（ s ）はおおむね 850 程度であった。これはたしかに「担当するアカデミックライティングの授業」という枠内においては意味をもつ数字ではある。しかしそれが他の種類のサンプルに比して大きいのか小さいのかはわからない。そのため、本稿における第一の研究目的として、語彙の豊富さの比較、就中、大学生のレポートにおけるそれについて相対的位置の確認を目指す。

第二の研究目的として、語彙の豊富さの応用可能性の提示を設定する。この目的に基づき、各変数についての相関を確認し、主成分分析を行う。これは、各変数の情報を圧縮することにより、対象とするサンプルの特徴を観測するためである。

後述するように、語彙の豊富さを用いた先行研究では、一定の枠内のサンプルにおいて数値を算定し、比較検討する。もちろんそれらの研究には十分な意義が認められるのであるが、数値そのものは必ずしも提示されない。また、仮に数値が示されていたとしても、すべての先行研究において常に同じ算定方法で求められたものではない。よって、先行研究に示された数値を活用したい場合は、自らの研究においても参考とする研究と同じ指標を使用しなければならない。ある指標を使用する場合、それをなぜ使うのか、いかに使うのかは、研究そのものの構想と大きく関わる。とすれば、参考とする研究の構想に自分の研究を近似させていくことにもなりかねない。

また、複数ある語彙の豊富さの指標を比較検討する研究もある。これらも学術的価値は高く、稿者自身も少なからず恩恵を受けている。ただ、やはり「数値自体にどんな意味があるのか、どんな価値があるのか」を必ずしも明確に語るものではない。

このような問題意識から、本稿では、有価証券報告書・小説・演説を対象とした語彙の豊富さの計測結果を資料として公開する。さらに、それらに加えて

名詞率・MVR・平均文長を示す（本稿末尾）。

1 先行研究

1-1 語彙の豊富さ

TTR は、語彙の豊富さの指標として最も広く知られたものであろう。日本語による研究では、管見の限りでは樺島忠夫・寿岳章子（1965）による使用が最初とみられる（ただし、TTR ではなく NKR——延べ語数・数え語数比率——として提示されている）。しかし、延べ語数の大きく異なるサンプルを比較するとき、TTR は妥当な指標となりえない。そのため、多くの代替指標が提案されている。

語彙の豊富さについての研究は、指標の比較検討を主としたものと、指標を利用したものがある。前者としては、鄭弯弯・金明哲（2018）が各指標を網羅している。当該論文では TTR を含めた 11 指標について、日本語・中国語・英語のテキストについての比較検討を行い、「文章の長さと言語の依存性が最も低い指標は s である」と結論づけている。先述のとおり稿者自身も鄭・金（2018）に従い、s を語彙の豊富さの指標として利用している（植田：2021 等）。

他の研究成果としては、今田水穂による一連のものが挙げられる。一例として今田（2021）をみると、当該研究では児童作文を対象として 24 種の指標を用いている。また、浅石卓真（2017）はテキストの特徴を示す各種指標について概略を示している。その中で語彙の豊富さについても紹介しており、TTR・D・K 特性値が利用されてきた状況を指摘している。

指標を利用した研究としては、金明哲によるものが多い。劉雪琴・金明哲（2017）は、宇野浩二の休養期間前後の作品を対象に文体の変化を計測している。そのひとつとして、語彙の豊富さ（TTR、K 特性値、Sichel の S 値）を用いている。また、柳燁佳・金明哲（2019）では TTR を分析の指標のひとつとして用いて、菊池寛「妖妻記」の前半と後半の文体について考察している。その他の研究者によるものとしては、鈴木崇史・影浦峯（2011）が Simpson の D と延べ語数を調整した TTR を用いて、田島ますみ・深田淳・佐藤尚子・玉岡賀津雄（2009）は Carroll の修正 TTR を用いて研究を行っている。また、TTR を使用した研究（大川慎：2019、富永愛：2015）もある。このように、語彙の豊富さを用いた研究は多い。

1-2 名詞率・MVR

名詞率・MVRは、樺島・寿岳（1965）が「要約的表現」と「描写的表現」とを区別する際に提示した指標である。当該論文では第一に名詞（N）、第二に動詞（V）、第三に形容詞・形容動詞・副詞・連体詞（M）、第四に接続詞・感動詞（I）を区分し、それらの比率を分析に使用している。このうち第四のIについては「比率は極めて小さく（中略）品詞比率は無視してかまわない」とされる。名詞率は[品詞全体に占める名詞の割合]である。MVR(Modifier-Verb Ratio)は $[M/V \times 100]$ である。つまり、形容詞等と動詞を比較したとき、前者の比率が大きければ求める値は大きくなり、後者が大きければ値は小さくなる。これらの指標については、

- 要約的表現：名詞率・大
- 描写的表現：名詞率・小
- ありさま>動き： $M > V$
- ありさま<動き： $M < V$

と指摘され、さらに名詞率とMVRを組み合わせると、

- 名詞率が大きくMVRが小さい：要約的
- 名詞率が小さくMVRが大きい：ありさま描写的
- 名詞率が小さくMVRが小さい：動き描写的

とされる。

樺島・寿岳（1965）以降、名詞率とMVRは多くの研究に利用されている。井関龍太・菊池理紗・望月正哉・福田由紀・石黒圭（2022）は、青空文庫所収テキストを読んだ読者にアンケートを行い、指標が読者のイメージ喚起におよぼす影響を分析している。深澤克朗・沢登千恵子（2018）は中世の勅撰和歌集を対象として、大川孔明（2020）は平安・鎌倉期の文学作品を対象として分析している。

名詞率とMVRを用いた研究では、必ずしも樺島・寿岳（1965）が提示した「要約的／描写的」「ありさま描写的／動き描写的」といった分類を用いておらず、指標を変数のひとつとして利用してきた。名詞率・MVRはその提案者の想定を超えて、まさに文体を科学的（統計的）に処理する有効な指標として機能している。

2 データの分析 1 サンプル全体

データの検討にあたり、まず、全般の概要を示したのち、有価証券報告書・小説・演説について個別の分析を示す。一連の分析により、語彙の豊富さの応用可能性を提示したい。

2-1 概要

サンプルサイズは 282 件、有価証券報告書が 100 件、小説が 100 件、所信表明演説が 82 件である。

有価証券報告書は 2024 年 4 月 18 日時点で、日本国内株式の時価総額のランキング 1 位から 100 位までの企業のもをを対象とする。具体的には、2023 年度の有価証券報告書のうち「第 2【事業の状況】」を用いた。有価証券報告書を使用したのは、第一に異なる業種の企業（書き手）であっても書式が統一されており、対象として適切であると考えたことによる。第二に機能的な内容であるため、小説・演説と比較するのに適していると判断されたためである。

有価証券報告書は、ファイルが PDF であるため、全文をエディタに貼り付けたあと、正規表現を使用して改行を削除した。そのため、たとえば、

2002 年 8 月 中国第一汽車集团有限公司と、中国における自動車の共同事業に関する基本合意書を締結

2004 年 6 月 中国において乗用車を生産・販売するため、広州汽車集团股份有限公司との間で合弁契約を締結 (トヨタ自動車)

のように本来なら異なる情報（文）であっても、「2002 年 8 月 中国第一汽車集团有限公司と、中国における自動車の共同事業に関する基本合意書を締結 2004 年 6 月 中国において乗用車を生産・販売するため、広州汽車集团股份有限公司との間で合弁契約を締結」のようにひとつの文として扱われてしまう箇所もある。しかしながら、サンプリングの結果、影響は軽微であると判断された。

小説は芥川龍之介・岡本かの子・坂口安吾・太宰治・夏目漱石・宮本百合子・山本周五郎・吉川英治および江戸川乱歩の 9 人を対象とし、乱歩を除く 8 人はそれぞれ 10 作品ずつを採用した。乱歩については、低年齢層向け作品と非低年齢層向け作品を 10 作品ずつ採用した。データの採取にあたっては R の aozora パッケージを使用し、ルビ等、分析に不要なノイズを削除した（石田基広：2017）。

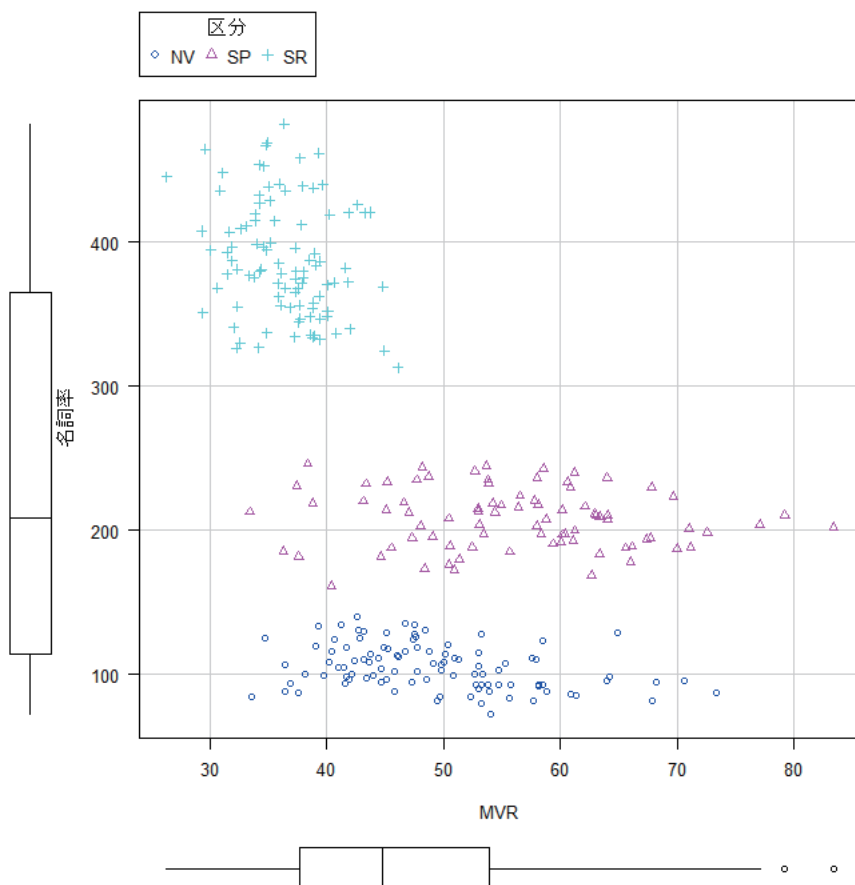
所信表明演説は石田（2017）を参考に、所信表明演説コーパス（注 2）所収のものを利用した。対象となるのは第 18 回の吉田茂から第 187 回の安倍晋三まで、全 82 回分である。発言者によってサンプルサイズに差がある。最多は佐藤栄作の 13 回、最少は宇野宗佑・羽田孜・福田康夫・麻生太郎・鳩山由紀夫の 1 回である。

形態素解析にはKH Coderを用いた。形態素解析器はMeCabを使用した。強制抽出は行っていない。よって、語は基本的に短単位で抽出されている。分析対象とした品詞は、名詞・動詞・形容詞・形容動詞・副詞である。連体詞については、中尾桂子(2010)・大川(2020)等に従い、用いていない。統計解析にはEZRを使用した。ただし、データの整備にあたってはExcelで処理を行っている。

2-2 分析

樺島・寿岳(1965)に従い、MVRと名詞率の散布図(図2-1)を示す。サンプルは有価証券報告書(SR)・小説(NV)・所信表明演説(SP)で区分している。

図2-1 サンプル全体の名詞率・MVR 散布図



機能的な文書である有価証券報告書は全般に高・名詞率／低・MVRであり、小説は名詞率が低く、MVRは高いものから低いものまで様々である。所信表明演説は、名詞率は有価証券報告書と小説の中間に位置し、MVRにはばらつきがある。演説は機能性が求められつつも、単なる説明ではなく情動性が必要とされるため、このような分布がみられるのではないか。

各区分ごとの指標についての平均値・標準偏差・中央値は、以下のとおりである（表 2-1、2-2、2-3）。

表 2-1 有価証券報告書

	s	名詞率	MVR	平均文長
平均値	.852	388.705	36.395	96.191
標準偏差	.013	39.612	3.769	12.389
中央値	.852	381.805	36.370	92.940

表 2-2 小説

	s	名詞率	MVR	平均文長
平均値	.898	105.303	49.068	42.721
標準偏差	.019	15.485	8.086	16.789
中央値	.899	103.576	47.773	38.702

表 2-3 演説

	s	名詞率	MVR	平均文長
平均値	.906	207.913	56.142	64.526
標準偏差	.011	2.374	1.078	13.280
中央値	.905	208.925	56.497	65.142

3 データの分析 2 有価証券報告書

図 3-1 有価証券報告書の散布図行列

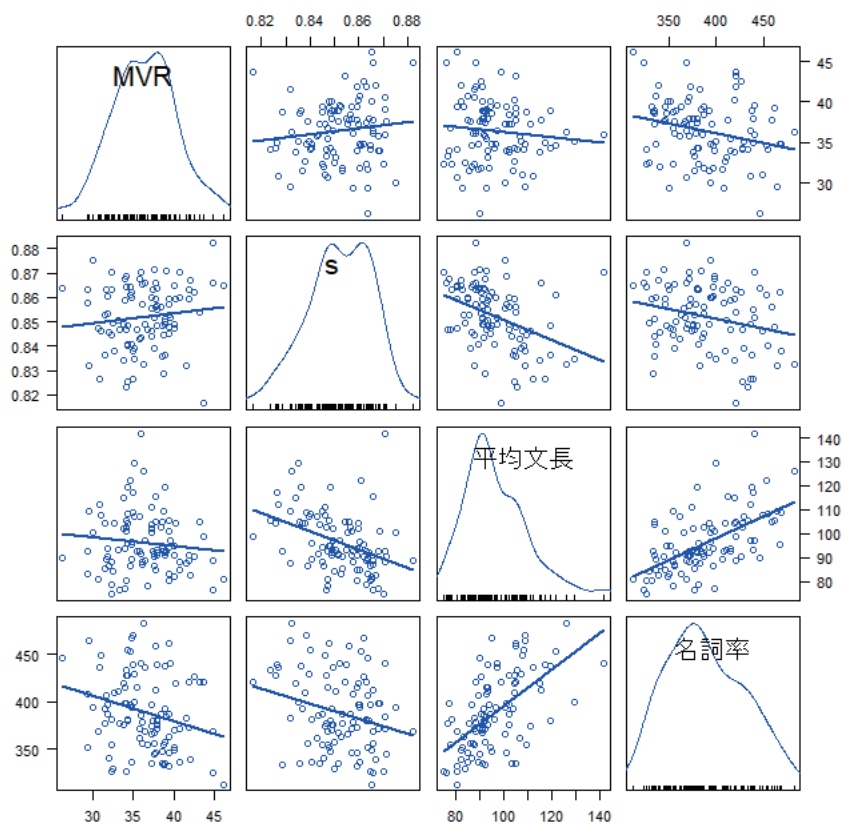


表 3-1 有価証券報告書の相関係数行列

SR	MVR	s	平均文長	名詞率
MVR	1.000	.139	-.092	-.245
s		1.000	-.450	-.208
平均文長		***	1.000	.615
名詞率	*	*	***	1.000

p-values <.001 *** <.01 ** <.05 *

散布図行列（図 3-1、曲線はカーネル密度推定値、以下同じ）・相関係数行列（表 3-1、スピアマンの相関係数、以下同じ）からはいくつかの組み合わせに相関をみることができる。 p 値が 5% 水準を下回るものについてみると、名詞率と平均文長に中程度の正の相関、名詞率と MVR および s に弱い負の相関、平均文長と s に中程度の負の相関がある。

表 3-2 有価証券報告書の主成分分析

	PC1	PC2	PC3	PC4
固有値	1.929	0.941	0.764	0.367
標準偏差	1.389	0.970	0.874	0.605
寄与率	.482	.235	.191	.092
累積寄与率	.482	.717	.908	1.000
主成分負荷量	PC1	PC2	PC3	PC4
MVR	.294	.920	.183	.186
s	.456	-.265	.822	-.216
平均文長	-.600	.287	.240	-.707
名詞率	-.588	-.039	.483	.648

主成分分析の結果は以上（表 3-2）のとおりである。第二主成分（PC2）までの累積寄与率が 71.7% であるため、分析は第二主成分までで行う。第一主成分（PC1）は s が正、平均文長と名詞率が負である。短文で要約的でなく、語彙がやや豊富な変数といってよいだろう。第二主成分は MVR が正であるため、ありさま描写的とみることができる。

試みに、企業のウェブサイトに掲示されている「会社概要」「企業概要」にある設立年について、戦前（A）・戦後 1990 年代まで（B）・戦後 2000 年以降（C）で層別した散布図が図 3-2 である。C について、やや第一主成分が負に位置しているが、大きな差というほどではない。業種・業態や企業文化が各変数に影響を与えているかもしれないが、本稿においては紙幅の都合から可能性の提示にとどめる。

表 3-3 は有価証券報告書の s ・名詞率・MVR・平均文長について、平均値・標準偏差・中央値をまとめたものである。

いずれも平均値と中央値の差は小さく、標準偏差も大きくない。有価証券報告書は書式が固定されていることもあり、書き方の揺らぎが小さいと考えられ

図 3-2 有価証券報告書の主成分分析散布図

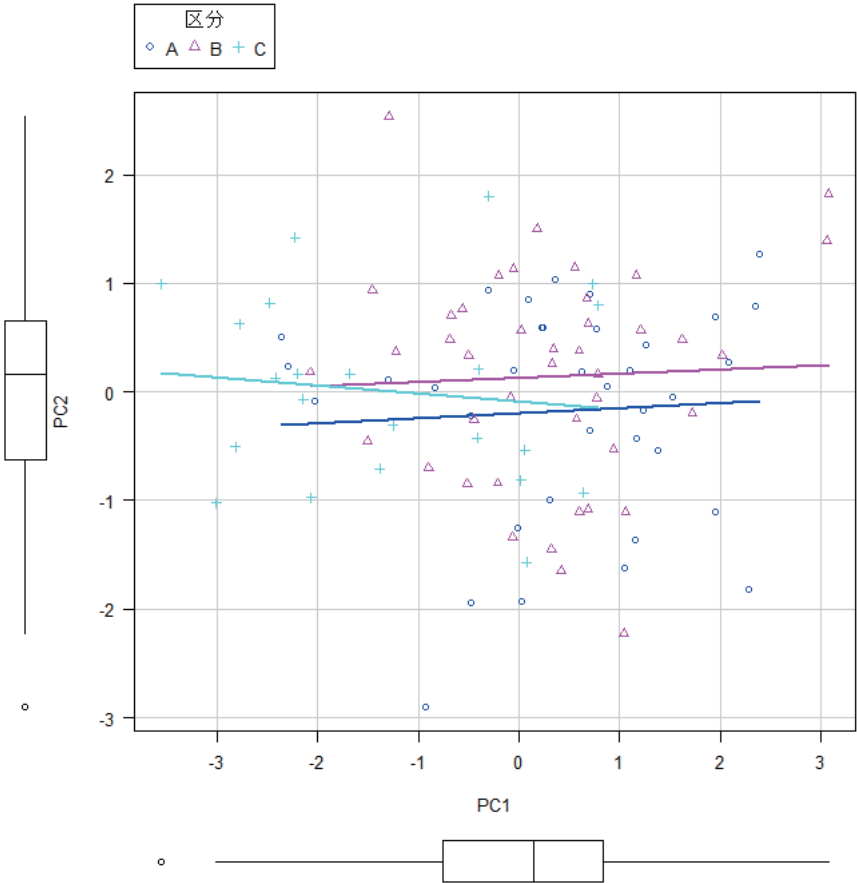


表 3-3 有価証券報告書の平均値・標準偏差・中央値

	s	名詞率	MVR	平均文長
平均値	.852	388.705	36.395	96.191
標準偏差	.013	39.612	3.769	12.389
中央値	.852	381.805	36.370	92.940

る。なお、有価証券報告書の語彙の豊富さ（s）は、稿者が勤務先で担当する学生に課すレポートのそれとほぼ同じである。本稿執筆の目的のひとつは、テキストの種類と語彙の豊富さとの関係の解明であった。レポートや有価証券報

告書といった機能的文書の場合、語彙の豊富さ (s) は .850 をひとつの基準とすることができるのではないかと。ただし、機能的文書のすべてが当てはまるわけではない。たとえば、新聞の社説などの場合、s は .900 前後である。また、形態素解析のしかたによって数値は変動する。本稿では KH Coder（形態素解析器は MeCab）を用いたが、形態素解析器および辞書に何を用いるか、またどの語をいかなる品詞として認定するかによって、数値は変動しうる。

4 データの分析 3 小説

4-1 小説全般

対象となる変数について、散布図行列（図 4-1）および相関係数行列（表 4-1）を確認する。

図 4-1 小説全体の散布図行列

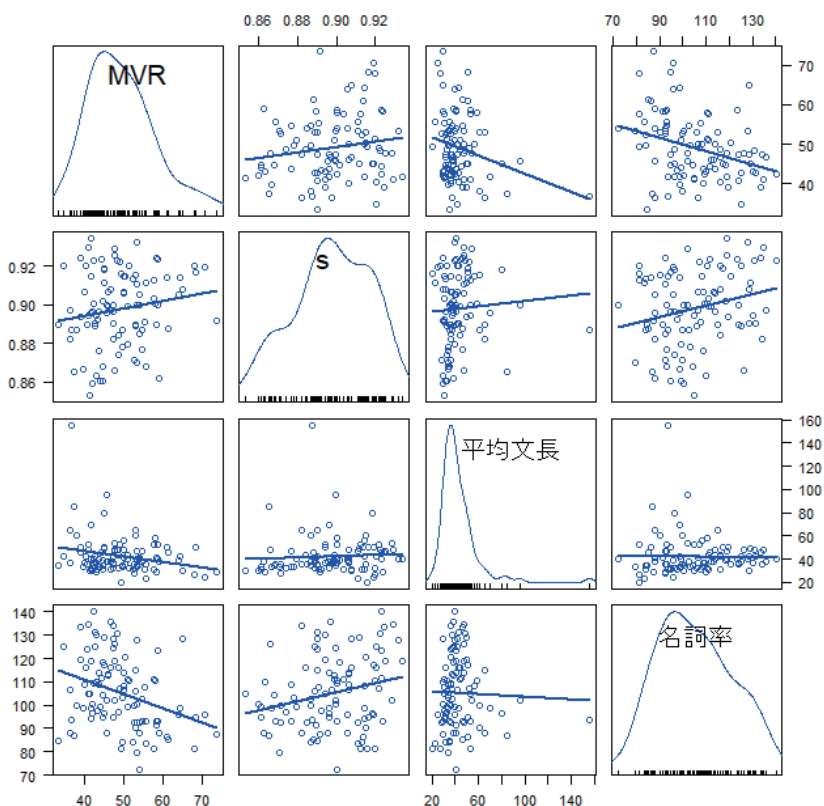


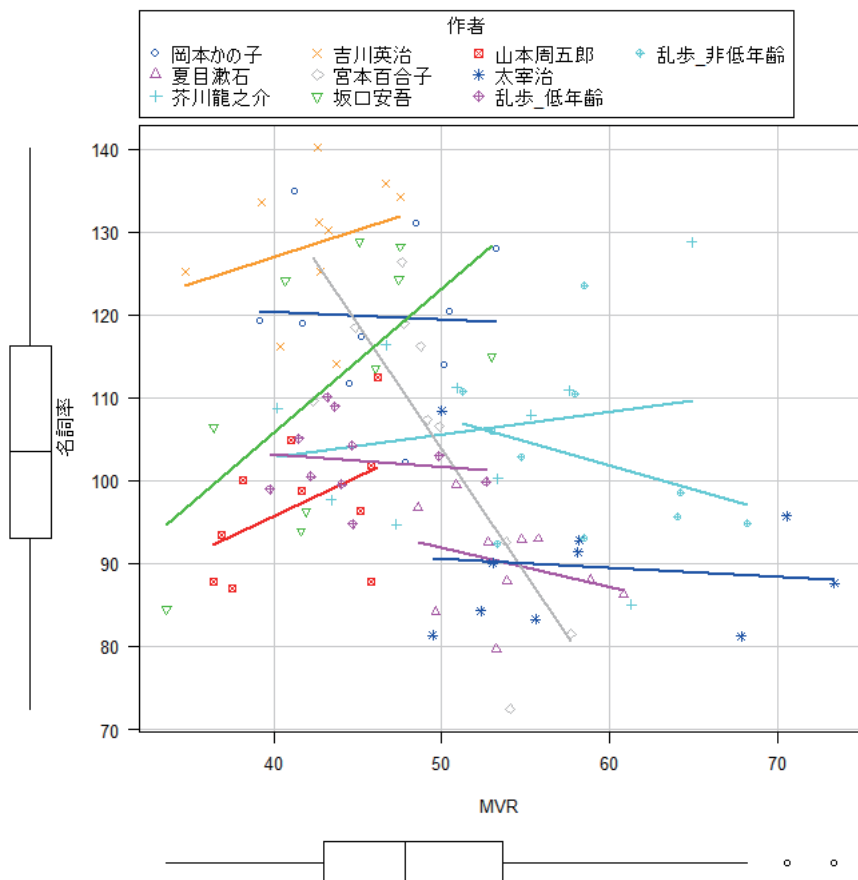
表 4-1 小説全体の相関係数行列

NV	MVR	s	平均文長	名詞率
MVR	1.000	.138	-.130	-.348
s		1.000	.222	.232
平均文長		*	1.000	.150
名詞率	***	*		1.000

p-values <.001 *** <.01 ** <.05 *

散布図行列からもわかるように、名詞率と MVR、s と平均文長、s と名詞率に弱い相関がみられる。いずれも p 値は 5% を下回る。

図 4-2 小説全体の名詞率・MVR 散布図



MVR と名詞率の散布図（図 4-2）からは、高名詞率・低 MVR 群（吉川英治・岡本かの子・坂口安吾）、低名詞率・高 MVR 群（太宰治・夏目漱石・江戸川乱歩非低年齢）、低名詞率・低 MVR 群（山本周五郎・江戸川乱歩低年齢）の配置がみられる。芥川龍之介・宮本百合子は横断的である。

s・名詞率・MVR・平均文長を変数として主成分分析を行ったところ、以下の結果を得た（表 4-2）。

表 4-2 小説の主成分分析

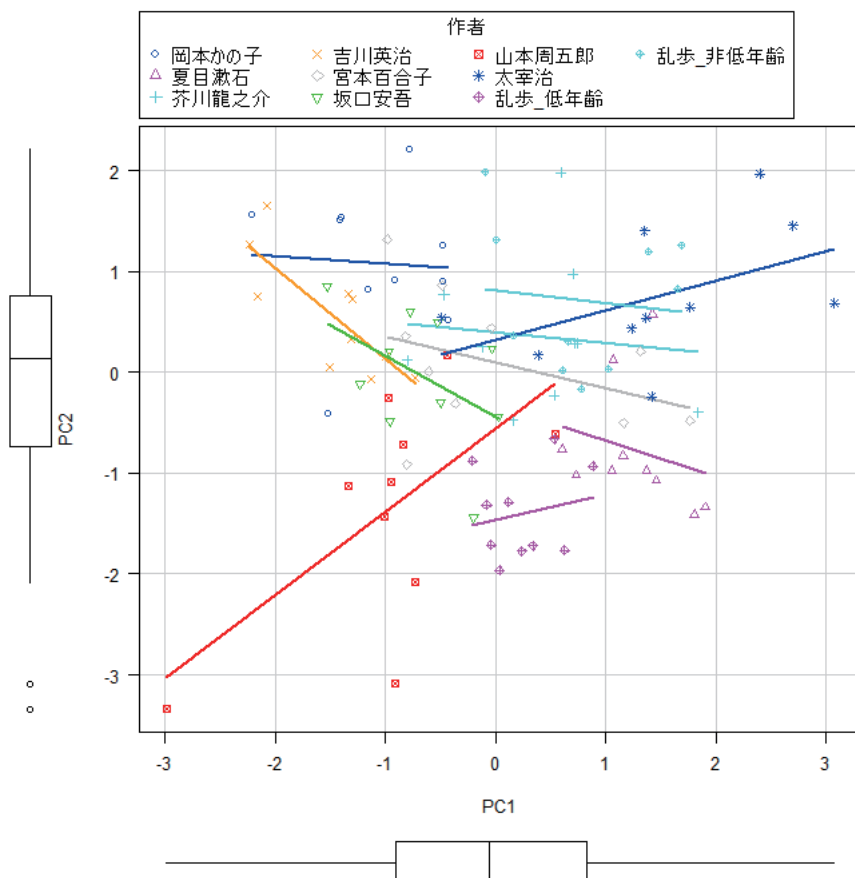
	PC1	PC2	PC3	PC4
固有値	1.397	1.176	0.984	0.442
標準偏差	1.182	1.085	0.992	0.665
寄与率	.349	.294	.246	.111
累積寄与率	.349	.643	.889	1.000
主成分負荷量	PC1	PC2	PC3	PC4
MVR	.670	.382	.205	.603
s	-.159	.779	.402	-.453
平均文長	-.378	-.301	.809	.334
名詞率	-.619	.396	-.375	.565

第一主成分は MVR と名詞率、第二主成分は s、第三主成分は平均文長に特徴的な値をみせる。固有値は第二主成分までが 1 を超え、累積寄与率は 64.3% である。以下、第一主成分と第二主成分の散布図（図 4-3）を示す。

第一主成分と第二主成分の散布図では、第一象限に太宰・芥川・乱歩非低年齢、第二象限に岡本・吉川、第三象限に山本、第四象限に漱石・乱歩低年齢が配置された。宮本は第二象限と第四象限に、安吾は第二象限と第三象限にわたって配置している。

上にみた MVR と名詞率の散布図と類似した分布ではあるが、特徴はより明確になったといえるだろう。

図 4-3 小説全般の主成分分析散布図（第一主成分・第二主成分）

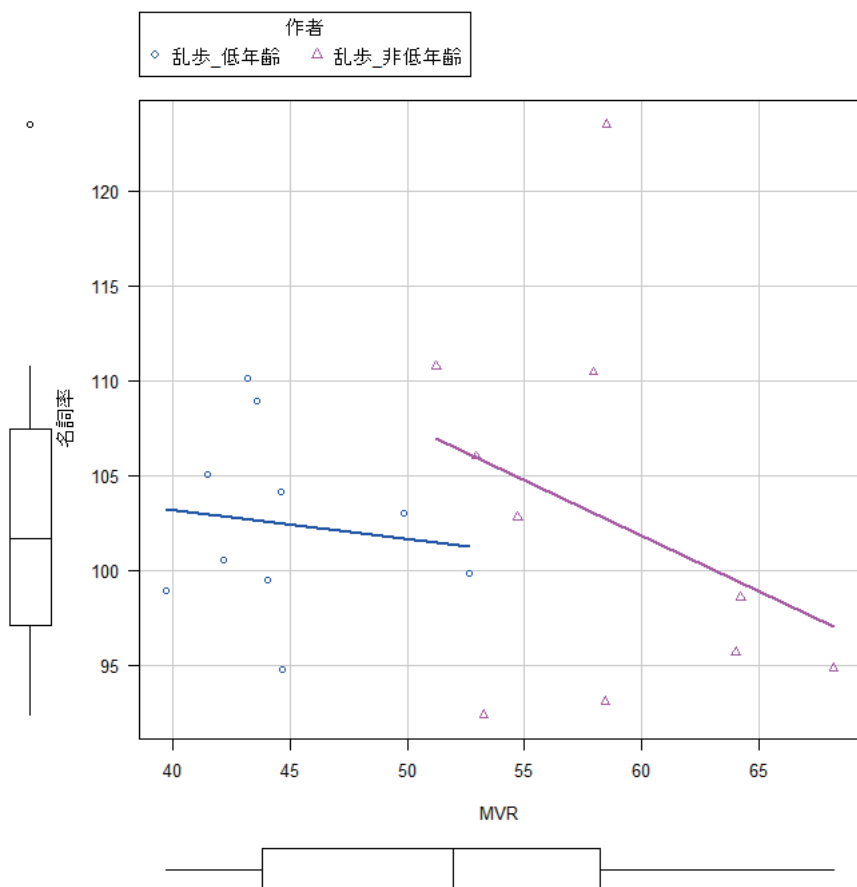


4-2 江戸川乱歩作品

江戸川乱歩の小説は、おおむね低年齢層向け作品と非低年齢層向け作品とに分けられる。先述のとおり、本稿ではそれぞれ10本を対象として分析を試みる。なお、江戸川乱歩作品における低年齢層向け作品と非低年齢層作品との文体比較については、稿者のゼミナールの出身者である前田麻帆氏の卒業論文（「江戸川乱歩の青年小説と児童小説の差異と特徴に関する計量的な考察」、2023年度）に示唆を受けている。

MVRと名詞率とを散布図（図4.4）に配置すると、名詞率には大きな差はないものの、低年齢層向け作品（以下、J）はMVRが低く、非低年齢層向け

図 4-4 江戸川乱歩作品の名詞率・MVR 散布図



作品（以下、NJ）は MVR が高い。樺島・寿岳（1965）は小説における MVR について、大きければ「ありさま描写的」小さければ「動き描写的」と指摘している。また、MVR の高いものほど読者に対する負荷が高いとも述べている。翻って乱歩作品をみれば、J の MVR が低い。つまり対象とするサンプルに限ったことではあるが、低年齢層向け作品は動き描写的であり、読書時の負荷も低くなっているといえる。

語彙の豊富さの指標をみると（表 4-3）、s は平均値・中央値とも NJ が上回っている。s は値が 1 に近いほど語彙が豊富とみなしうるため、J に比して NJ の方が語彙が豊富である（注 3）。これは、先ほどの MVR の検討とあわせて

表 4-3 江戸川乱歩作品における s の平均値・標準偏差・中央値

	J	NJ
平均値	.866	.907
標準偏差	.007	.012
中央値	.865	.908

考えれば、江戸川乱歩作品では低年齢層に対して低負荷が指向されていると指摘しうる。これらの傾向は、江戸川乱歩作品全般に対して行うことで、より詳細かつ厳密な結果が得られるであろう。

5 データの分析 4 所信表明演説

図 5-1 演説の散布図行列

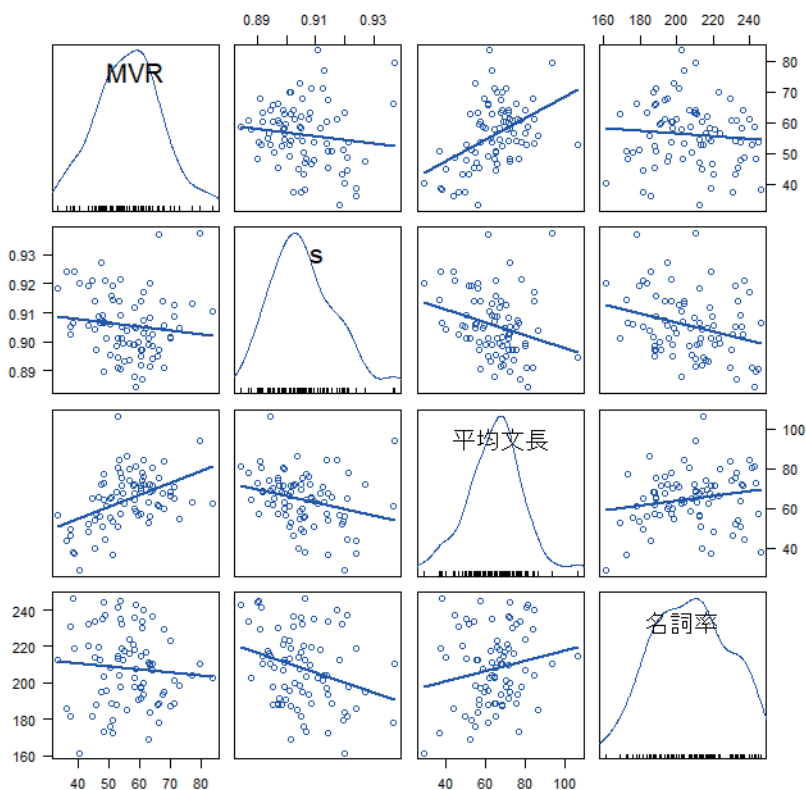


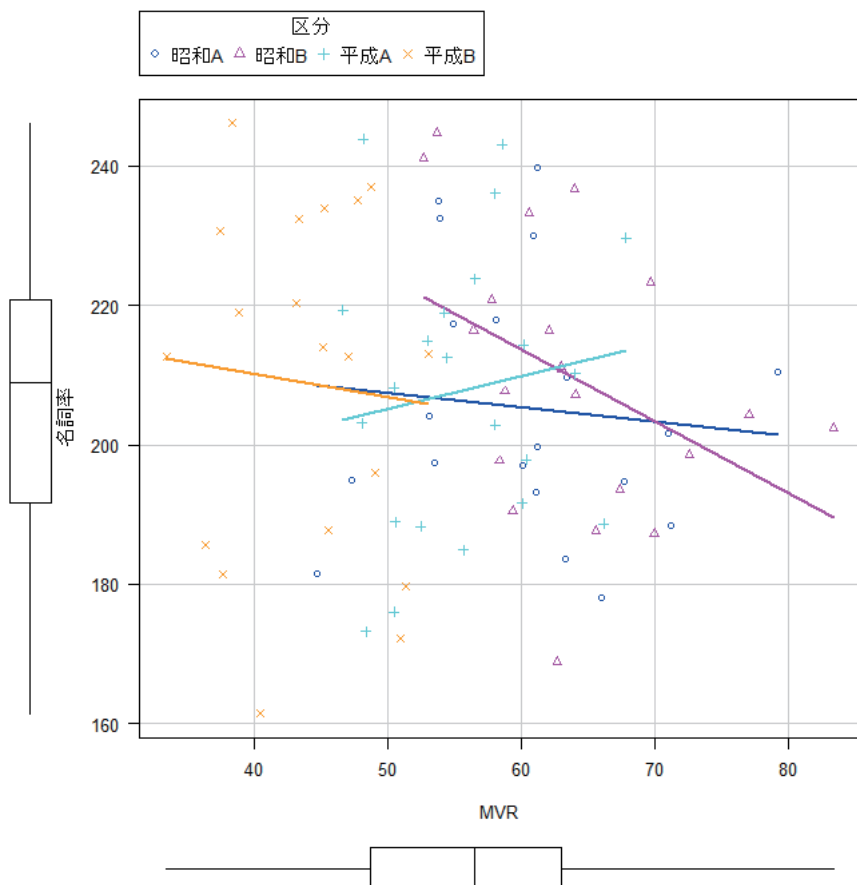
表 5-1 演説の相関係数行列

SP	MVR	s	平均文長	名詞率
MVR	1.000	-.203	.417	-.113
s		1.000	-.322	-.268
平均文長	***	**	1.000	.181
名詞率		*		1.000

p-values <.001 *** <.01 ** <.05 *

散布図行列（図 5-1）および相関係数行列（表 5-1）からは、各変数のあいだにいくつか相関がみられる。 p 値が 5% を下回るものをみると、平均文長と

図 5-2 演説の名詞率・MVR 散布図



MVR とに弱い正の相関、s と平均文長・名詞率に弱い負の相関がみえる。

MVR と名詞率の散布図（図 5-2）については、昭和 A 期（018 回 _ 吉田茂-060 回 _ 佐藤栄作）・昭和 B 期（062 回 _ 佐藤栄作-109 回 _ 中曽根康弘）・平成 A 期（111 回 _ 竹下登-150 回 _ 森喜朗）・平成 B 期（151 回 _ 小泉純一郎-187 回 _ 安倍晋三）を層別すると、明確な偏りがみられる。

いずれも、名詞率にはほぼ偏りが無い。しかし、MVR には明確な傾向を看取することが可能である。昭和 A 期のものと昭和 B 期のものについては差がない。平成 A 期のものは昭和期のものと重複するものの、より低い分布に重なる。平成 B 期は明らかに MVR が低くなっている。先に江戸川乱歩の分析にもみたとおり、MVR が低いと読みやすく、また読者に対して低負荷となる。受け手にとって演説は読むものではなく聞くものではあるが、時代が下るほど、受け手の負荷が軽く理解しやすいものとなっている。また、スピーチライターが演説者に対して低負荷を指向した可能性もある。

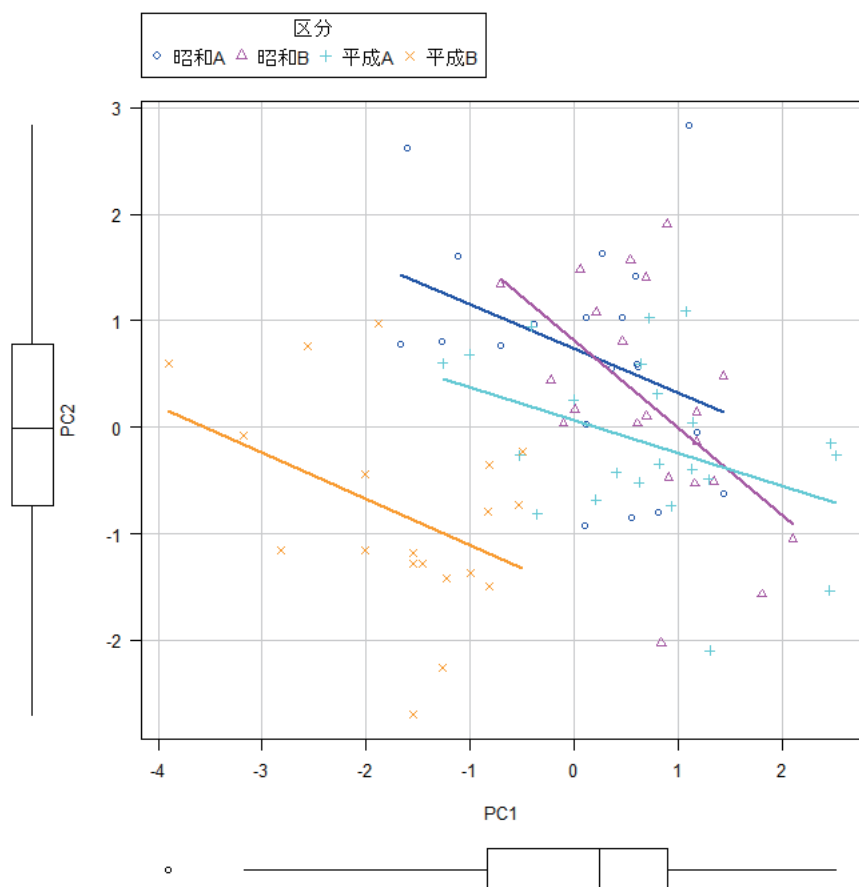
MVR・名詞率・s・平均文長を変数とした主成分分析の結果は、以下の通りである（表 5-2）。

表 5-2 演説の主成分分析

	PC1	PC2	PC3	PC4
固有値	1.661	1.182	0.684	0.472
標準偏差	1.289	1.087	0.827	0.687
寄与率	.415	.296	.171	.118
累積寄与率	.415	.711	.882	1.000
主成分負荷量	PC1	PC2	PC3	PC4
MVR	.490	.585	.024	.647
s	-.498	.384	.777	.002
平均文長	.633	.210	.304	-.680
名詞率	.333	-.683	.550	.345

第二主成分までの累積寄与率が 71.1% であるため、分析には第二主成分までを使用する。第一主成分は平均文長と MVR が正、s が負である。語彙を絞った一文の長い、ありさま描写的なテキストといえるだろう。第二主成分は高 MVR・低名詞率であるテキストである。第一主成分と第二主成分について、先ほどと同様の層別をした散布図（図 5-3）を示す。

図 5-3 演説の主成分分析散布図



昭和 A・B 期と平成 A 期が類似した分布を示すのに対し、平成 B 期は第一主成分・第二主成分のいずれも負の象限に配置される。これを主成分に従って考えれば、「語彙は多く文長は短い、要約的で動き描写的な文の集合」といえる。

参考として、以下に各変数の平均値・標準偏差・中央値を示しておく（表 5-3）。

おわりに

以上、対象サンプル 282 件について、語彙の豊富さおよびその他の変数を概

表 5-3 演説の平均値・標準偏差・中央値

	s	名詞率	MVR	平均文長
平均値	.906	207.913	56.142	64.526
標準偏差	.011	20.374	10.078	13.280
中央値	.905	208.925	56.497	65.142

観した。検討の中でもみたとおり、「授業で課したレポートにおける語彙の豊富さと、機能的文書である有価証券報告書の語彙の豊富さがほぼ同一である」という結果が得られた。

そのほかに得られたいくつかの知見も、興味深いものであった。

たとえば、同一作家であっても想定対象の年齢によって書き方を変えている可能性がある。江戸川乱歩の場合、低年齢層対象の作品では MVR が低く、非低年齢層対象の作品では MVR が高い。語彙の豊富さをみても、低年齢層の方が低い。つまり、低年齢層対象の作品は読み手の負担が低い可能性がある。小説家 9 人をみると、夏目漱石・芥川龍之介・太宰治の MVR は高く、吉川英治・山本周五郎のそれは低い。主成分分析では、第一主成分（高・MVR / 低・名詞率）と第二主成分（高・s）が計測された。いずれの主成分も正に配置されるのが太宰治・芥川龍之介・江戸川乱歩（非低年齢）であり、いずれの主成分も負に配置されるのが山本周五郎である。このように、小説の分析においては、名詞率・MVR 等に加えて語彙の豊富さも特性の理解に有用である。

演説においては、時代が降るほど MVR が低くなる。さらに、主成分分析の結果、時代の降った時点での所信表明演説の特徴が「語彙は多く文長は短い、要約的で動き描写的」であることも明らかとなった。

このように、語彙の豊富さは他の変数と組み合わせることで、様々な分析を可能とする。本稿で扱ったサンプルは限定的ではあるが、その応用を示唆することができたものとする。

注記

注 1 語彙の豊富さの指標としては、TTR (Type Token Ratio) が一般的である。これは、異なり語数を述べ語数で除したもの（異なり語数 / 延べ語数）である。TTR は総語数が同程度のサンプルの比較には適している。しかしながら、総語数が大きくなると、延べ語数は大きくなるものの、異なり語数は同じ比率で大きくならない。そのため、TTR を代替する様々な指標が提案されている。s は

そのうちのひとつで、 $[\log(\log(\text{異なり語数})) / \log(\log(\text{延べ語数}))]$ で求める。これは、異なり語数と延べ語数に対して対数を重ねることで、その差を圧縮するものである。

注2 所信表明演説コーパス (<https://github.com/yuukimiyo/GeneralPolicySpeechOfPrimeMinisterOfJapan>)

注3 いずれもサンプルサイズが10と小さいため、U検定等の統計処理は行わない。

参考文献

- 浅石卓真 (2017) 「テキストの特徴を計量する指標の概観」『日本図書館情報学会誌』63-3
- 石田基広 (2017) 『Rによるテキストマイニング入門 (第2版)』森北出版
- 井関龍太・菊池理紗・望月正哉・福田由紀・石黒圭 (2022) 「品詞構成に基づく文体指標は読者の印象とどのように関わるか：MVRと品詞構成率の心理学的検討」『計量国語学』33-7
- 今田水穂 (2021) 「児童作文における語彙多様性の評価」『計量国語学』33-3
- 植田麦 (2021) 「テキストマイニング技術を応用したレポート課題の教育効果測定」『実践國文學』99
- 植田麦 (2022) 「テキストマイニングによるレポート課題の分析 — 文章と構成の観点から —」『実践國文學』101
- 大川孔明 (2020) 「叙述語から見た平安鎌倉時代の文学作品の文体類型」『計量国語学』32-6
- 大川慎 (2019) 「テキストマイニングを利用した観光地イメージの言語間比較に関する試み」『日本情報経営学会誌』39-2
- 樺島忠夫・寿岳章子 (1965) 『文体の科学』綜芸舎
- 鈴木崇史・影浦峯 (2011) 「名詞の分布特徴量を用いた政治テキスト分析」『行動計量学』38-1
- 鄭弯弯・金明哲 (2018) 「変動係数を用いた語彙の豊富さ指標の比較評価」『同志社大学ハリス理化学研究報告』58-4
- 田島ますみ・深田淳・佐藤尚子・玉岡賀津雄 (2009) 「語彙指標数値と文章主観評価の関係：日本人大学生による2種類の書き言葉コーパスを使った実証研究」『中央学院大学人間・自然論叢』29
- 富永愛 (2015) 「いきものがかり・水野良樹と山下穂尊の歌詞に関する文体的特徴分析：計量言語学的手法による」『日本文学』111
- 中尾桂子 (2010) 「品詞構成率に基づくテキスト分析の可能性：メール自己紹介文、小説、作文、名大コーパスの比較から」『大妻女子大学紀要 文系』42
- 深澤克朗・沢登千恵子 (2018) 「後期勅撰和歌集における計量的アプローチ」『情報知識学会誌』28-2

- 柳燁佳・金明哲（2019）「菊池寛「妖妻記」の改題前後の文体分析」『日本行動計量学会大会抄録集』 47
- 劉雪琴・金明哲（2017）「宇野浩二の病氣前後の文体変化に関する計量的分析」『計量国語学』 31-2

本研究において使用したデータは、<https://drive.google.com/file/d/1WziIRGiaFXIcqv8Q3GzuEbXaIvzKIQAx/view?usp=sharing>で公開している。

（うえだ ばく・明治大学教授）