大妻女子大学紀要 ― 文系 ―　No. 42,　2010年 3 月

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 ――

中　尾　桂　子

あらまし

一般に，インターネット上の記述は，話しことばと書きことばの中間であると言われることが多いが，そうなのか。さらに，文体差は，品詞の使用状況に現れるものなのか。本稿では，樺島・寿岳（1965）が文体分析に用いた品詞構成率 MVR に基づき，メールでの自己紹介文，話しことば，書きことば，3 の種のテキストを比較した。その結果，記述媒体の違いにより，差が明確になったことから，インターネット上の記述は従来の話しことばでも書きことばでもなく，新しい文体的特徴を持つ可能性，さらに，文体比較の指標として品詞情報の有効性が確認できた。

キーワード：メールテキスト，話しことば，書きことば，品詞構成率，相関

1. はじめに

一般に，インターネットやメール上で記述されることばは，話しことば的書きことばだと言われる。この考えは，インターネットという世界が，2 次元の紙面よりはインタラクティブで，しかし，話しことばほどには即時的な世界ではないということによる。

確かに，インターネットの世界でやり取りされることばは，文字を媒体とするコミュニケーションである。また，文字言語の使い方とは異なり，短文で倒置や省略も多く，多分に口語的でもあり，視覚的な情報が加味される場合もある。このため，話しことばと書きことばとその中間的なものと位置づけたくなるものであるだろう。

しかし，インターネットでやりとりされるテキストは，果たして本当に両者の中間的な存在だと位置づけられるものか，または，それは，なにをもってそう位置づけられるのだろうか。また，どちらかの性質を強くもつのであれば，それは，話しことばと書きことばのどちらの性質だろうか。

この問題を整理することは，テキスト属性を明らかにすることにつながり，かつ，記述の際のスタンスが決まる根拠を把握するものでもあると言えるだろう。また，文章表現やアカデミックライティングの指導の際に，評価，目的のポイントを明示的に区別する視点を再考し，整理することにつながると期待できる。

そこで，本稿では，テキストとしての性質を調べるため，3 種のテキストを比較し，その性質を分析する。インターネットでやり取りされるテキストとして，今回は，メールでの自己紹介文を利用する，また，会話テキストとしては，話しことばを文字起しした名大コーパス 2008 年度版を，さらに，書きことばとして，口語的な記述スタイルであるとされる大学生の記述である意見文と，会話を多用すると考えられる小説を加えて分析し，3 種類，計 4 つのテキスト差を検証してみる。なお，テキストの分析には，従来，様々な方法が用いられてきたが，本稿では，樺島・寿岳（1965）

の文体比較法に基づき，品詞構成比率 MVR 値を利用する。

2. 統計手法を用いた日本語研究の動向

2.1. 現　状

昨今，自然言語処理分野の飛躍的な発展に伴い，自然言語処理と計算言語学の間の学際的な研究が増えてきた。それは，テキストを電子的データとして扱う研究分野として，自然言語処理系の工学的な研究と，大量言語データとしてのコーパスに基づく教育や文体研究とに，大きくわけて考えられていたものが結合しはじめ，全体として，再度細分化されていることを表しているとも考えられる。しかし，中尾（2009）でみたように，従来の計算言語学や計量言語学（国語学）の分野で培われてきた国内独自発達の統計的手法と，自然言語処理の他の分野での統計的手法には，系統の異なりとでも言うような違いがみられ，言語分析に用いられる統計手法は，概念上同じものであっても，年代分野の違いで術語が異なる。また，その利用の立場を明確にはしておらず，根拠が不明な場合も多い。たとえば，計量国語学系の相関分析を行っている興味深い研究に馬場他（2000）があるが，そこでは，ピアソンの積率相関係数，スピアマンの順位相関係数，ケンドールの順位相関係数を用いた相関分析 3 種のいずれをなぜ使うか書かれていない。もちろん，当該分野での自明のことで，読む人が読めばわかることなのかもしれないが，学際的な研究が増えてきた昨今では，特定の分野に関わらず，先行研究として関心を寄せられる場合もあることから，門外漢にも立場を明示してもらえるとうれしい。

このような現状は，国立国語研究所が行った語彙調査結果の 1964 年の発表で，日本語の定量的研究の方法と理論（計量語彙論）が確立したと広く認識されるようになり（伊，2000），それ以降，日本語定量化研究（この場合は計量語彙論）における術語の定義やそれが示す概念が半世紀近く修正されずに使われている（山崎，2009）という指摘からも推測できる，検証自体の不活性化が一因でもあるだろう。

現在のように統計をテキスト分析に利用するのに多方向からの異なった流れが存在している状態では，それぞれの交差部分がどこにあるか，また，日本語テキスト研究で利用される統計的手法にはどのようなものが多いかということを後行の研究者は確認する必要がある。そして，統計的検証法自体が当該の言語現象を分析するのに妥当な方法であるかどうかを見直す研究を繰り返し行うことも，日本語テキストを用いた研究の今後の発展に寄与するものと考えられることから，統計手法の工夫や手法の開示を行う規範的な姿勢づくりを徹底させることが，環境づくりとして必要だと言えよう。

2 節では，日本語テキスト分析において統計的手法がどのような捉え方であったか，いくつかの研究やケーススタディーを取り上げ，定量的研究の変数や指標，利用する統計についての概観を試み，本稿のテキスト分析に用いる手法の位置づけとする。

2.2. 日本語テキスト分析が行われる研究分野

日本語テキストを用いた定量的研究の分野，ならびに，先行研究における統計的手法についてごく簡略的に整理してみる。

日本語テキストを計量的に分析する研究は，題材が偶発的に日本語テキストであったという研究から，日本語そのものを対象とするものまで多様なものがある。題材が偶然日本語であったという研究も含めると，社会言語学，計量文献学，計量行動学，心理学，経済学など，多彩な分野が関係

する。日本語そのものを対象とする研究は，文学における文体論研究や言語学，従来の国語学（現在の日本語学だが，2004 年以前の研究は，以下，国語学の分野と言う）となる。

日本語そのものを対象とする日本語テキスト分析は，①ある現象がテキスト個別のものであるか見るものと，②言語現象一般の性質を示すものであるか見るものとの 2 つに分けられる。前者の①に該当する先行研究としては，安本（1963）や村上（2004）の行った計量文献学や，樺島（1963），大野（1956）など国語・国文学という分野での文体論があげられるが，これらの研究において，統計は，作者推定，作者心情の推移，成立時期や推移の過程推定などにおける複数のテキスト間の検定や分類に利用されている。後者②に該当する分野としては，主に，計量国語学，計量言語学，コーパス言語学となり，語彙を量的に調べた分布分析から概念上の構造を探る量的記述が範疇に入る。こちらの場合，国語・国文学の分野の研究と，国語に対する基本調査の分野での発展が目覚しく，古くは，水谷（1977），安本（1985），国立国語研究所の一連の調査や，昨今の様々なコーパスベースの研究があげられる。

目的ではなく，分野という点で見ると，数理言語学という分野で言う計量言語学，狭義の数理言語学，計算言語学や，英語学や英語教育の分野で盛んなコーパス言語学に分けられる。計量言語学は，言語現象の 1 つとして語彙を量的に調べるもので，数理言語学という分野の 1 つと位置づけられている（伊藤，2002）。計量言語学は，基本的には語彙を量的に見るものであるが，それには，文体を数値化して統計分析するものや，言語の年代を統計的に見るもの，言語行動など社会言語学の分野の研究分析において統計や量的記述を行うものも含まれる。また，数理言語学には，形式意味論や文法研究における形式性，記号論を扱う狭義の数理言語学と，計算言語学が含まれ，計算語学は自然言語処理の分野での言語研究を指すというように，細分化される。

本節では，テキストの規模にかかわらず，計量的に分析する手法事態を概観するため，厳密な意味では区別せず，日本語テキスト分析ということばで，上記を網羅的に捉える。

2.3. 文体分析・語彙量調査における統計的手法

テキストの性質を語彙の統計量で記述する文体分析や語彙量によるテキスト分析では，概ね，以下のように統計値を利用する。

① 検討観点（変数）と，調査用の指標を決める
② 指標の実数を数える
③ 比較時，テキストの規模が異なる場合，調整（観測された実数値を百分率や千分率に計算しなおしたり），標準化する
④ 計量結果に基づいて，指標間の適合度や差異の程度などを検定する（分散，有意水準，推定値など，分析のための数値を計算する）
⑤ 該当テキストの性質を見るために，他のテキストと比較して差を見る
⑥ 最後に，計算値に基づいて比較した結果から，検討する観点，すなわち，変数について判断を下す

しかし，妥当性や関連性を検討したり，概観を端的に捉えるためには，様々な数学的，または，統計的計算法が工夫される必要がある。語彙全体，または，あるテキストに特徴的に出現する語彙に特化して，その頻度数を元に，テキスト間の（語彙同士の）共通度，類似度，集中度，不均等度（偏り具合）といったものを示すのであるが，このときに計算される指数は，ケーススタディーを通して実証的に考案した結果，確立されてきたものである。これらのうち，計算結果の安定性が高いものは，汎用的に用いられることになり，結果，さらに発展，進化を繰り返し，固有名詞化した

呼称を持ったものとなるに至る。それらが，いくつかの統計上の計算方法，すなわち，テキストの統計的分析のための手法となっている。

日本で行われてきた語彙の計量調査においても，文体分析のケーススタディーを通して培われてきたモデルがある。たとえば，樺島・寿岳（1965）の MVR（Modifying words and Verb Ratio: MVR: 筆者推測）という品詞比率の分布を調べるモデルや，国立国語研究所（1983）の語彙調査などで培われた計量語彙論のための「水谷モデル」などである（水谷，1983 に集約）。

2.4. 関連性の分析について

統計の分野では，関係性について明らかにする分析全般を相関分析と言うが，「相関分析」という個別の分析手法が存在するわけではなく（内田他，2003），本節でも「相関分析」のみを指す意味としてではなく，広義の意味での「関連性」について概観する。

関連性は，基本的には，ある事象と別の事象との間で比較し，それぞれの事象に共通するなんらかの事柄，たとえば，頻度等といった数値の大小により，判断される。この比較の際，関連を見るための共通項は，データの性質や形態，また，何を比較するかという観点によって異なるため，相関分析で関連性を見るためのポイント，すなわち，指標が，少なからず存在することになる（内田他，2003）。言語現象を取り扱った相関分析は，特徴的に使用される単語やその程度，文長，といった着眼点，すなわち，指標に基づき，何らかの観点や検証テーマを変数として取り上げ，テキスト間の差異を調べるということでテキスト間の関係を比較し，2 種類以上のテキスト間に関連があるかどうかについて納得できる分析結果を出すのに利用されてきた。

一般に，統計的分析の初歩段階では，基本統計量に基づいてデータ形態を概観し，次いで，相関係数や相関比を求めて判断される。言語現象の場合も基本的には同じであるが，この段階では，相関表や図，総関係数で関係がある（強弱）とわかっても，因果関係の有無は確認できないため，さらに進めて，テキストや言語現象間の関係の方向性やつながりの強さといった観点から関連性について明らかにする場合が多い。その場合は，比較する観点，すなわち，変量を複数にすることで，どのような事項，すなわち，因子が，両方の関係の強弱により影響を与えているかということを見ていくのであるが，一般に，統計的手法として，重回帰分析，判別分析，主成分分析，因子分析，クラスター分析と呼ばれるものになり，これらはまとめて，多変量解析と呼ばれる。

多変量解析には，上記の他にもいくつかあるが，言語現象の分析での利用が少ないようである。それは，変数設定と指標設定の際に言語の性質上設定できないものがあることや，言語というものの分析が，どこまで集めても言語の母体には近づかないのであるから，必ず，母体となる母数を推測するという前提のもと，統計的解析が進められるということによる。つまり，母体の推測を前提としながらも，暗黙的にそこは回避して考えることが多く，推測的に検証することはあまりない。このため，ごく限られた手法で比較観点の関係を見るのみとなるのであろう。

内田他（2003）や 2.3. 節でも述べたように，関係の分析は，データと目的の数だけ，知恵と工夫が必要とされ，その手法がいくつも示されるということにつながる。言語現象の分析目的に合致する範囲ということになるのかもしれないが，可能性を検証していくのも必要であろう。もちろん，言語分析における統計手法のうち，汎用的なものが繰り返し利用される場合，類似の先行研究の手法に倣い，分析し，納得する結果を結論付けるということが繰り返されるが，それにより，目的やデータを考慮せず，汎用的なモデルで分析し，結果検証に対する納得を得ようとする場合もある。次節で先行研究の例を紹介しながら，日本語テキストの分析で行われる統計的手法をごく簡単に概観するのであるが，各分野別の歴史的な経緯と代表的な統計手法の用いられ方について先行研究を

あげてまとめ，その中で相関係数，回帰分析，因子分析といった手法とテキスト分析との関係を整理する。

2.5. 言語分析に利用される統計

では，2.4.節で述べた研究目的や分野別に，言語現象はどのように統計的に分析されているのか。

まず，狭義の数理言語学であるが，言語を一種の形式的体系として扱う形式意味論や，理論言語学が含まれる。言語を数学記号に置き換えて計算し，計算結果，すなわち，計算による証明に基づき理論化しようとするものである。ここでは集合理論や代数などの数学的計算が行われるが，その規則化や検証に統計的な手法を用いるわけではない。

次に，計算言語学であるが，ここで利用される統計手法は，情報検索時の検索対象（重要語と呼ばれる）や，ある概念を特徴づける一連の語群抽出に利用される。また，自然言語処理システムの構文解析時にも利用されている。機械翻訳や音声翻訳，ロボット製作を目的とする場合，自然言語処理技術の向上が必要であるが，統計はこれら工学的なシステム開発のために，自然言語の，語彙的概念，語彙ネットワーク，係り受け，共起傾向を探り，自然言語に近いものを再構成する過程で利用される。

計算言語学における統計は，より高精度な構文解析や抽出を志向するものの，手法自体を特に意識はしていないように見える。中尾（2007）でも利用を試みた，北他（2002）の残差 IDF やエントロピーを応用した統計手法が工夫され，より正確で簡便なものが常に求められているが，特に，テキストを分析するための統計手法の工夫には差がないようである。ただし，自然言語処理の技術を応用する実証的文法研究や，語彙の定量化といった学際的な分野が発展しつつあり，この方面では，統計的手法が用いられ，その利用手法についての分析も行われ，モデル化が進められている（李・井佐原，2005）。

計算言語学の応用による計量語彙論，ならびに，コーパス言語学での統計手法を見ると，検定，相関分析における同様の計算を利用することが多い。それは，下準備や利用ツールにおもねる部分が大きいことによると考えられる。コーパス言語学では，言語現象の定量化において，語彙的な面から計測するために語の単位を決めて分割するなどといった，一定の下準備が必要になるため，分析の前段階の処理を自動化する目的で開発されたコンコーダンサーというシステムを利用することが多い。下準備とは，語彙数，文数，1 行中の単語数などの実測値とその標準化値，並びに，平均や中央値といった語彙の基本統計量を明らかにするとともに，連接関係を目視するための KWIC インデックスを利用した共起語の概観やその傾向を数値化するための n-gram 連接の統計量などを指す。日本語テキストが処理できるコンコーダンサーは少ないが，表音文字言語で利用するコンコーダンサーには，たとえば，AntConc や Word Smith Tool, TXTANA などがあり，これらには，定量化の際の計算方式が選択できるように，複数の計算が組み込まれている。これら下処理の関係で，同様の統計手法を利用することにつながっているのだろう。

コーパス言語学という言い方で一くくりにするものの，応用分野は広く，学際的なものも多い。これまでは，統計的手法を用いる意図としては共通する点が多いものの，手法の違いを分析的に捉えて応用しようとするよりは，先行研究を踏襲するのみで統計手法自体の検証はさほど分析的ではない場合も多かったが，学際的な研究が増えた結果，他分野の手法を通して，客観視しようとする視点が起こり，統計手法の選別自体が研究目的になることも多くなっている。たとえば，特定の現象が一般的な現象かどうかについて見るような場合，ある特定のテキストと，母集団となる言語全般とを比較するとして，相関係数を求めることや対数尤度比検定などを行うこと，また，データの

性質や比較対象の違いを考慮して，母集団がないノンパラメトリックな場合や母集団を推測するパラメトリックな検定ををを行うこと，そして，そのために，相関係数では，スピアマンの順位相関係数や，ピアソンの相関係数などを弁別的に用いたり，検定でも，Z 検定，G 検定，F 検定，ピアソンの χ^2 検定などを，区別して用いたりして，手法の意味を吟味して区別するようになってきていることなどである。

最近のコーパス言語学では，データとなるテキストの位置づけや検定目的に応じて統計手法が選別され，どの計算式を使うかについては，それぞれの研究者の工夫点となると受け取られている。この選択という行為が，より的確に目的となる指標から変数を読み取るために焦点化の方法を工夫するということにつながり，英語学や英語教育額で盛んなコーパス言語学的統計計算の工夫につながっている（石川，2008 等）。

一方，計量言語学では，言語現象を統計的に分析し，言語現象から理論や法則を帰納的に導くことが，一応の大前提とされているが，そこへ至るまでの過程として，ケーススタディーが報告されることも多い。ここでも統計量による分析が行われるが，統計量の計算方式は，コーパス言語学でコンコーダンサーに組み込まれているような検定や相関分析に関するいくつかの計算が対比的に利用されている。ただし，日本語学や日本語教育学における語彙量の定量化研究は，コーパス言語学が台頭する以前から日本で行われてきた流れがあり（山崎，2009），60 年代以降の大規模語彙調査を先導してきた水谷（1983）に見られるような計量語彙論が，確立，完成したという意識が一般化していることから，語彙の基本的な統計量とその利用法や指標として計量される対象語句の検証，それらを判断するために利用された統計的検証法自体を工夫しようという意識はそれほど高くないようである。しかしながら，その一方で，荻野（2002）が指摘するように，従来の方法より，どこか斬新な手法を常に探し，以前の方法を検証することなく，常に新しい手法の応用とその新手法利用に対する賛同を求める風潮がある。ある特定の分析モデルが実証できれば，それを繰り返し，別の類似言語現象に当てはめて分析を繰り返すが（在，2002，2004-6），それを同一人物が繰り返すだけでなく，他者も積極的に検証しあうという慣習は，ごく一部の限られた範囲でしか行われていないようである。

以上を踏まえながら，次節では，本節の目的である，日本語テキストデータの統計的分析手法を比較し，言語現象における関連性判断のための統計的手法を考察するが，以下，取り扱う論文は，入手が比較的簡便なものに限定されていることを断っておきたい。

2.6. 日本語教育での相関分析の手法とその対象

2.6.1. テキスト分析例 1 ── 計量言語学・計量語彙論における統計手法 ──

計量語彙論は，国立国語研究所の語彙調査の経過とともに相前後して発展してきたと考えられる。この国立国語研究所の大規模な語彙調査は，母集団である日本語というものの性質を，限られたテキストから推測することによって標本を抽出するという考えで進められている。最初に語を特定し，その後，語の定義に基づいて分割したあと，語ごとに頻度を計測していくのであるが，この過程でも，それぞれの段階で，統計的検証を行いながら進められていた。

統計的手法としては，最初にデータである対象テキストの代表値や散布度を求め，次いで，個別の事象を検討しつつ，標本を抽出するために，推定，検定，相関分析が行われているが，その計算方法は，日本語の性質を検討した計量国語学の分野での手法に応じて，日本語に合う方法として検証済みだとされている。

国立国語研究所の語彙調査は，計量言語学における語彙論と，計量国語学の流れを練り上げるよ

うな流れで発展したが，計量語彙論や計量国語学の分野での研究とは，性質が異なる。語彙調査は，語の単位認定における詳細な分析と，膨大な作業と工夫が行われたが，それは，標本抽出という目的に特化されている。一方の計量言語学，計量語彙論では，計量国語学会の系統で，言語，心理，数学，社会学，工学の分野における研究手法の公開的応用の場として統計的手法の研究やモデル化が行われていた（伊藤，2002）のであるから，両者の関係は深いが同じものとは位置づけられない。国立国語研究所の語彙調査や，その統計的手法は，水谷（1983），ならびに，『現代雑誌九十種の用語用字』分冊⑶に詳細にまとめられており，その質量ともに多いことから，ここでは扱わず，そちらを参照いただきたい。

言語の文法的現象を計量的に分析する計量言語学の分野は昨今，自然言語処理技術の発展とともに，新たな局面を迎えているが，計算言語学との学際的な研究も進んでいる。また，従来の計量語彙論での基本手法の問題点を踏まえ，さらに，計算処理に，認知言語学的視点など，外部の言語理論を変数や因子に取り入れる手法を提案する研究が見られる（李，2002，2004，2006）。統計的手法を用いることで従来の文法分析に奥行きが出た研究である。

2.6.2. テキスト分析例 1 ── 文体論における統計的手法 ──

日本語学における日本語の計量分析は，語彙調査を中心に見ると 50 年代から盛んであったと言う（丸山・田野村 2000，山崎 2000）が，同じく日本語・日本文学にかかわる文体論での計量的な分析も，同時期から盛んであった。個別の研究では数が多いことから，代表的な研究者の名前だけをあげると，安本美典，波多野完治，宮島達夫，大野晋，村上征勝，小池栄治等があげられる。この他にも多くが文体分析において計量的な手法を利用している。

文学における文体論で，統計的手法を用い，指標モデルを考案して利用している初期の代表として，樺島・寿岳（1965）の「文体の統計的観察」があげられる。小林（2005）が樺島・寿岳（1965）を指して「分析項目が多岐にわたり，かつ，項目のバランスがよく有意性を保っているので，安定した結果を得やすい」としているように，計量的文体分析を行う場合に引用されることの多い論文である。

文体論では，あるテキストに特異に多い特徴語や，品詞構成で，比率という観点から分析が行われることが多いため，ここでは，樺島・寿岳（1965）の手法を紹介しながら，文体論の分野における語彙の計測と標準化の方法を確認する。

計量的な文体分析における樺島・寿岳（1965）の目的は，主観的な印象を客観評価することであった。そして，理想的な文体把握方法というのは質的分析点を数量化したものであるとするが，定義が困難であるとして，質的分析点を加工後，数量化することでより理想的な方法に近づこうという考え方で研究している。また，計量語彙論では，実際には，作品を単に統計的に記述する立場の分析が多いと憂い，数える部分をはっきり定義すること，ならびに，定義や計量にぶれを生じさせないことを第一に考えて計測，データ化を行っている。

樺島・寿岳（1965）は，文体を統計的に観察するための指標モデルを考案し，それに基づき，テキスト内の指標同士を検証して文体分析に応用している。樺島・寿岳（1965）の『文体の統計的観察』では，短編小説 100 編の各作品から無作為に 80 文ずつ抽出し，そのテキストに対して 10 項目の指標の使用頻度を計量した後，その 10 項目の指標に基づいて短編小説 100 作品を比較する。そして，それぞれの差から作家の文体分析状況を考察しているが，そのときの指標は，名詞の比率，MVR（形，形動，副，連体/動詞数×100），指示詞の比率，字音語の比率，文の長さ，接続詞を持つ文の比率，引用文の比率，現在止めの文の比率，色彩語の比率，表情語の比率といった 10 種

類の比率である。

樺島・寿岳（1965）モデルの特徴は，名詞比率と，他品詞の比率との関係で記述の文体を予測できるという点にある。また，名詞以外の品詞構成率を MVR という独自の指標モデルで表すことである。この MVR（形，形動，副，連体/動詞数×100）値の大小を見て文体を推測するのであるが，MVR の値が大きいということは，動詞以外の自立語（品詞）が多く，様態記述中心の文章ということになり，MVR 値が小さければ，動詞が多く，動的な記述が中心の文章ということになるとして，これを用いることで，数値データで客観的に簡略して文体が捉えられるというのである。

これは，名詞が品詞比率の代表値として捉えられることを検証し，名詞と MVR 値を利用することによってテキストの性質を推測する指標にできることを確かめた結果によるものであるが（樺島，1963），この樺島（1963）の品詞構成比率がとる分布は，水谷（1977）の改訂でより明確になっている（伊藤，2002）。名詞とそれ以外の品詞との関係から，テキストの品詞構成に基づき，記述文体を推定するという手法である。

図 1 に，小説 100 作品における MVR と名詞比率で品詞構成率の分布を表す。縦軸に MVR 値，横軸に自立語中の名詞の比率（%）を取っている。樺島・寿岳（1965）はこのような分布を見て，動きの多い文体かありさま中心の描写文体かについての読者側の心的印象を追確認した分析を行って，描写の分類を行おうとしている。コーパスを用いて行う計量的な文体研究でも，指標の実測値を計上するところから始めるが，樺島・寿岳（1965）はその方法を明確にしていない。当時の単語認定は，国立国語研究所の研究に準じるものであることが多く，暗黙の了解があるのかもしれない。

また，樺島らは，語彙の実測値に対して標準化を行うということをせず，テキストをあらかじめ平均化することや，分析するための指標を抽象化するなどの方法で分析を進めている。テキストデータは，出典先から同数ずつをランダムに集めてくるため，既に，均一なデータとされているとして，特に，実測値を調整する必要がないとすることによるのだろう。

以上のように，計量語彙論の分野では，語彙ベースでの文体研究への応用などで，樺島・寿岳（1965）の MVR や樺島（1955）や大野（1956）の品詞構成比率の分布法則といった，指標モデルや分析モデルが数多く開発されている。これら日本語の平均的な品詞構成比率などの計量語彙論的研究で培われた分布法則等は，水谷静夫により，検証，修正を加えられ，より抽象度の高いモデルへと改訂され今日の基礎知識や定説へとつながっているものが多い。

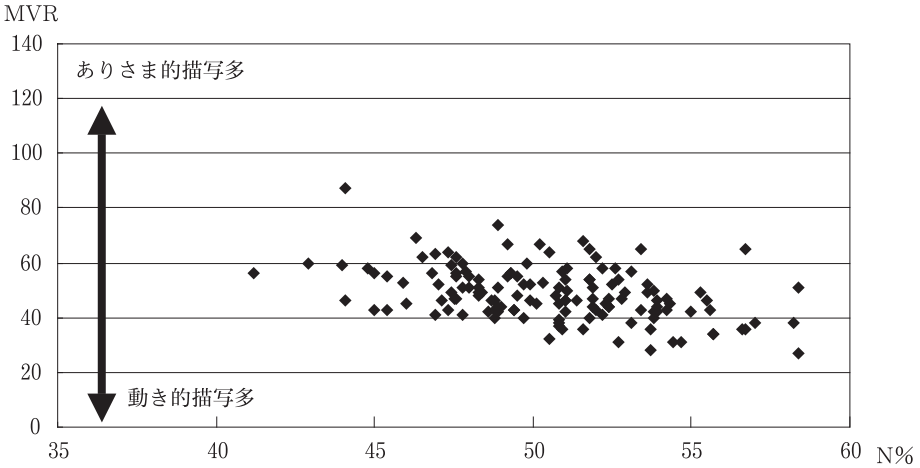

図 1　樺島・寿岳の MVR を利用した 100 小説の描写文体の比較（1965）

ただ，それが，後の計量言語学やコーパス言語学における統計的手法の検証や改訂へとつながったようすはない。「国産」の統計的手法は計量国語学の分野で検証，追認が繰り返され，基礎知識として定着する完成度が高いものとなっているが，今日，同様の検証や，文体分析を行うのに，これらの手法が利用されず，今日の計量語彙論的研究はコーパス言語学や計算言語学の潮流に沿っている。国産とでも言うべき統計的研究は，欧米のコーパス言語学における統計的手法やその検証方法へと，関心点を含めて推移している。

その理由として，60 年代の文体の統計的分析の手法が，今日の文体，計量国語学系の研究にとっては自明の理として統計量のごく基本的なものという位置づけになったということが考えられるが，もう一つ，自然言語処理技術の発展に伴い，日本語における統計手法やその検証判定への関心が薄れ，従来の計量国語学での統計手法と昨今利用される統計手法の間の乖離を生んだこと，さらに，同様の統計的手法だけでは，新たなことがわからなくなったということが同時期に重なったことが考えられるであろう。

もちろん，計量的な文体研究は，現在でも数多く行われているが，語彙頻度の実数を統計的に標準化して分析，比較することは少なく，計測した実数を如何に扱ったかについてはそれほど配慮しないことも多い（小林，2005，小池，2005）。それは，文体論の目的が主観評価の論理的な説明にあり，分析観点によってはコーパスを用いず，用例を集めてその頻度の多少を見ることで分析できる場合も多いということ，ならびに，歴史的に，かつて充分議論されたという意識があること，さらに，使い古された手法だけでは不明な点を明らかにすることができなくなった段階に至ったということ，そして，統計的手法で分析が可能な範囲を超えた研究が主流となっていることによるのだろう。そして，これが，日本語の文体論の歴史的な流れと現状を表す状態ということなのだろう。

2.6.3. テキスト分析例 3 ── コーパス言語学・計算言語学における統計手法：相関・検定 ──

コーパス言語学では，基本的に，個別のコーパスの特徴を見る場合，他のコーパスと比較し，差が見られた点が特徴だとする流れで行われる。これは，計量語彙論，文体論，コーパス言語学と研究スタンスや分野が異なっていても，統計的に行うという観点からすれば，共通することで，扱う対象が言語である以上，母語全数調査が不可能なため，参照できる母体がない場合の統計的な考え方に基づいている。したがって，言語研究で用いられる統計手法が類似してくるわけである。

コーパス言語学におけるコーパス間の比較では，目的に応じて，ある特定の観点（指標）の出現や分布を二つのコーパス間で比較する場合もあれば，いくつかの観点（指標）を複数のコーパス間で比較する場合もある。また，比較時には，差があるか，それは絶対的な差か，偶然起こりうる範囲の差か，偶然には起こりえない程度の「意味」のある差，すなわち，有意差か，という具合に，「差」の様相が重要となる。このとき，有意差があるかどうかについて見るために，有意差検定を行い，テキスト間の相違や指標間の差について，その差が偶然に起こり得ないもの，すなわち，差があるということを確かめる。対象コーパスデータ，比較する目的（変数），観察点（指標）が得られたら，指標の実測値を 2 項表に整理し，差があるかないか（仮説）を確認するために，ボーダーライン（期待値）を設定する。その後，対象コーパス間の指標同士の相関係数を求め，有意な差の有無を見る。

小林（1997）は，宮島（1970）の「古典対照語い表」を利用して，宮島が古典テキストの類似具合を相関係数を用いて確認した研究を追認した。さらに，品詞別に相関を調べ，宮島（1970）の研究を精緻化し，テキスト相関の類似度を品詞別に見る意義を示している。その際，宮島の最初の手法では，相関係数が非常に高かったが，それを「語彙数が多いために互いに 0 となる負の相関によ

るものだとして質的データに変換する方法で客観性を出している。このように，コーパス言語学の基本は，相関関係の強弱をどのような観点を指標に行うかという点が工夫するところである。

統計的手法の中の検定は，小規模なコーパスを用いた差の有無に対してよく行われるが，それは，小規模のデータでは特に，僅差が大きな意味を持つため，有意差を厳密に区別する場合が多いことによる。

村上（2005）では，大学留学生，または，予備教育の留学生に対する作文試験や課題などの評価において，書く能力を念頭において成績をつける場合，また，合格基準に至る能力か否かを測る場合，単一の型の文章を書くだけでは能力が測れないことを示している。作文の評価では，評価者間の差が大きいこと，さらに，評価者が評価しているのは「正確さ」や「多様性」,「段落」,「文」といた技術的形式的で正誤判断の付けやすいものに限られ,「文体」や「文のわかりやすさ」,「内容」といった観点に対しては，いずれの評価者も考慮していないことが，評価者と評価の観点との相関係数を求めることであきらかにしている。

教育分野における研究では，これまで，主観的な評価が多く，心象を客観視するという姿勢は少なかったが，Lee（2006）のように，日本語教育学の分野でも，ごく基本的な手法として有意差検定が利用されることが増えている。Lee は，作文の能力測定を，複雑さ，正確さ，流暢さの 3 点を日本語に合わせて検討を加えて指標にし，同一テーマで記述した留学生と日本人大学生の作文を検定し，両者がその 3 指標に基づいて異質であることを明らかにした。そして，論の立て方を文章構成パターンとして 7 タイプに分類し，両者の異なりに対する心象を形に表している。

また，昨今，コーパス言語学の分野計算言語学の分野との境界が薄れているが，工学的である計算言語学の分野での研究テーマが自然言語の教育的，言語学的観点により近づいた研究が増えている。近藤・松吉・佐藤（2006）はテキストの難易度推定システムを構築しているが，それは小中高大学生の教科書 111 冊から 1,167 サンプル 728,002 字のコーパスを用いて，それぞれを比較し，テキストの難易度調査を行った結果に基づく。そこでは，英語学における難易度算定公式に準拠した日本語の難易度算定方式を検討し，難易度推定フレームワークを作成して教科書コーパスで実証的に検証している。

基本的には，難易度の推定には，ある確率論的モデルを仮定しているときに，その観測データが得られる確率を指す尤度，または，手持ちの観測データであるパラメータ値が得られる確率を示す最尤推定により，推定を行っている。テキストに対して 13 段階の難易度クラスを設定し，この 13 個の尤度を求めて比較することで，難易度を決定していく。これに加えて，工学的価値を高める処理として，生起確率に対して，確率分布を調整するためのガウス関数の利用，ならびに，尤度の多項式回帰により，僅差のテキストレベルを明確に補正するという方法を用いている。尤度比検定まではコーパス言語学的分析手法といえるが，推定と確率の分布調整は，標本抽出ではともかく，少なくとも，現在のコーパス言語学の分野で行われるテキスト間比較では利用しないだろう。

しかし，計算言語学の分野からコーパス言語学的な分析を行うもので工学的研究ではあるが，教育に応用するための読解テキストの判定といった教材作成の面でも有益である。今後の分野境界における学際的な研究は，その手法と考え方において応用の可能性が高く，興味深いものになると考えられる。

2.7. テキスト分析法の例 4 ── 言語研究・教育分野における統計手法：因子分析・回帰分析 ──

社会調査や言語研究における内省，インタビュー，アンケートなどにおいては，頻度や傾向といった数量調査の結果が，いかなる要因によって決まるのかを特定することが多い。日本語のテキスト

分析においても，コーパス言語学の分野や，テキスト特性から文献や筆者を推測するといった計量文献学では，頻度計量の後に，その頻度の特長を示す原因を特定するための統計手法として，因子分析や回帰分析が利用されている。また，昨今の自然言語処理の発展と，利用者の増加により，テキストマイニングツールが利用され，非常に安易に因子分析などの多変量解析が行えるようになっている。ただし，これらでも，統計的には様々な手法があり，計算によっては結果が異なる。また，この因子分析などの手法は，そのデータを概観して心象判断が下せない場合には，結果を有効に利用できないこともある。多変量解析における因子分析という手法は，統計的技術における専門性もさることながら，データに対する専門知識が必要となる。

日本語教育の分野では，テキストから指導と習得の関係を検定し，相関から因果の原因を探ろうとする研究が多いが，統計的手法のヴァリエーションという点から言うと，心理学や応用言語学的見地からの検証研究手法を応用し，工夫を検討する研究も増えてきている。例えば，テキスト分析とは離れるが，玉岡他（2005）の日本語版 Can-Do-Statements のスケール設定の検証がある。

昨今，新たな教育法として，自律型学習を促進する向きが盛んになってきたが，その中の 1 つに，日本語の能力評価や目標設定の基準を示し，自己評価を行うとともに，言語能力を測定するという Can-Do Statements がある。これは，カナダで作成された自己評価型能力測定方式であるが，これが日本語版に改変され，国際交流基金などを落として，日本語評価のスタンダードにしていこうという流れがある。この測定方式では，自己評価を省みながら能力測定を行うための質問紙があり，日本語版として作成するには日本の生活や日本文化に即した達成目標が設定必要となる。このような輸入の調査法や理論を応用する場合，調査紙の翻訳版作成には，レベル分け，目的，スタンダードとして評価される項目，さらには，日本語教育の内容にまで及ぶ問題が隠れており，教育心理学からの示唆を受け，問題点を改善する試みが行われる。

玉岡他（2005）はこの日本語版 Can-Do-Statements のスケール設定を検証しようとして，調査結果の平均，標準偏差，および，斜交プロマックスか移転後の因子パターン行列および因子間相関を求めた。その結果，回答者の日本語能力と質問に対すると回答との相関が高くないことから，自己評価型の質問紙の良さを最大限発揮させるための条件を上げ，質問数，時間効率の良い質問内容などを検討するために，このスケールの信頼性と妥当性を検討した。質問項目として立てられている 180 種の組み合わせ全てについて，$r=.50$ 以上の有意な相関が得られることを確認し，妥当性を検証するためのクロンバックの α 係数がそのほとんどで $\alpha=.9$ を超える極めて高値であることを確かめ，質問紙の因子分析を，最尤法による因子抽出法，すなわち，Kaiser の正規化を伴うプロマックス法による斜交回転で行った。その後，Business Japanese Proficiency Test ビジネスの日本語能力テストの文字，語彙，文法力という項目を妥当性検証項目に加えて相関，標準偏差，因子分析の結果を考察し，それにより，日本語能力が正しく評価されていないことを明らかにした。このときの相関分析において，玉岡らは，相関係数を見るだけでは 2 変数間の関係の有無を調べたに過ぎないとして，言語技能 4 種を説明変数とした重回帰分析（強制投入法およびステップワイズ法）を行ったが，強制投入法式重回帰分析では有意な説明変数とはならないことを確認している。

ここで行われた工夫は，テキスト分析に対するものではないが，アンケート結果に対して行われるものであり，また，教育場面では必要な手法である。教育分野では，アンケートや試験の妥当性と信頼性を確かめることは，教師自身を確かめることになるわけで，言語教育では必須の作業である。また，さらに，高等教育機関においてはその組織の自己評価を行う過程で，授業評価や教員評価が行われる。言語教育における効果と大学評価の間の施策には，目的は同じでもアプローチが異なることが存在する。質問表を用いたアンケートやインタビュー調査，また，評価のためのこれら

の方法で採取された回答は，その妥当性，信頼性の検証を，教員自ら行うことで，長期的な計画やシラバス，コースデザインが組みやすくなるだろう。テキスト分析だけではなく，テキスト分析の結果が正しく反映された授業経営のためにも，ハードにおける検証も含めた形で，相関と回帰分析の利用法，ならびに，その種類の区別の実証的研究が数多く行われることが期待される。

2.8. 日本語・日本語教育におけるテキストの統計的分析法

テキストデータを抽象化し，実測値では見えない差を見出すということで統計的な手法がテキスト分析に用いられるが，統計的計算や手法は似ているものの，目的と着眼点が異なることから，利用方法が分野によって異なるようである。従来の日本語の文体研究の流れにおいては，テキスト特徴詳細化には主観的なものがあり，初めに結論があって，その自論をある程度客観視するために統計が利用されていた。文体記述や文体特長の分類における文体自体の判断が研究者により若干異なるスケールで識別されることからユニークなものとなっていたが，追認しにくい点も否定できない。文体を考察するという目的は同じでも，計量言語学や計量語彙論における研究では，定量化してテキスト特徴を検分し，相違を証明しつつ進められる。あくまでも客観的に記述するために統計手法を使用する。ただし，厳密に定量化を進めようとすると，今度は，言語自体が持つあいまいさにより，完全にはできないことも多い。このジレンマのために，言語の持つあいまいさをないものと仮定することが行われることもあるが，より安定した定量化のための工夫が行われることにもつながっている。文体論と計量言語学，計量語彙論の研究は，ちょうど逆のアプローチで進められるように見える。

さらに，最近は，テキストマイニングによる視覚的な検証が行えるようになっている。テキストマイニングでは，マイニングツールの開発で，計量言語学や計算言語学の分野で培われた手法を利用しながら画一的に，主観的判断を行うことができるようになっている。日本語テキストの簡便な処理が実現されているため，文体論研究や計量的な言語研究でも利用されることが多くなると考えられる。

統計的手法がどのような研究で，どのような目的で利用されているかを見ることにより，これまでの分野の境界が，分野ではなく，研究目的による違いとなっていくことが予想できる。自然言語処理技術やその考え方，ならびに，統計的手法は，テキストを概観しながら特徴を詳細化する分析の流れの中で，必須の技法と位置づけられるようになるのかもしれないが，それには手法としての利用できる範囲や可能性の検証をもっと行う必要があるだろう。

関連性を見るための統計手法は，原因と結果に変数を分けられる手法と分けられない手法に大別できるが，内田他（2003）によると，前者は，重回帰分析，判別分析，正準相関分析となり，後者は，主成分分析，因子分析，クラスター分析，正準相関分析，MT 法となる。しかし，今回，本稿で見た日本語テキストを扱う先行研究では，相関係数，検定，因子分析，回帰分析を利用するものが多かった。

コーパス言語学や計量語彙論の分野の定量的研究では，統計的手法のいずれを利用するかにより，また，どのような統計ツールを利用するかにより，計量結果が影響を受けて分析が異なってくる場合もある。そして，よく利用される計算方法は，統計ソフトに組み込まれているということもあるが，言語というものの性質や分析指標，分析目的による影響を受けるだけでなく，時代背景による研究環境の違いや流行の影響も受けるようである。

計量文献学と言われる分野でも同様の定量化が行われるが，文献の分析，比較のためには，検定だけでなく，因子分析，主成分分析などの統計的手法を用い，テキストの性質やテキスト間の比較

を行う。分野と目的により，若干異なるものの，語彙ベースでのテキスト分析は，相関分析や多変量解析を用いることが多い。

ということは，テキスト分析では，現在でも，ある程度，一般的な統計的分析手法だと考えられるものがあり，一部では，画一的にそれらが利用されることも多いが，その一方で，あまり利用されていない統計手法もあるということになる。ただし，この一般的と考えられている手法は，その手法の良し悪しや可能性をよく判断した上で一般化されたものかどうかはよくわからない。皆と同じ手法を根拠なく利用している向きもあるのではないか。ということで，統計手法と日本語の研究目的の明確な位置づけや分布を整理すること，そして，その上で，一般化するという流れができることが望まれる。

今回，収集した先行研究は，インターネットを経由して，研究機関の論文データベースから入手することが安易なものの中で局所的に調べた。非常に限られた方法で概観したものではあるが，今回の語彙分析に関する限られた範囲で見た限りでも，統計と言いながら，実数を計上し，実数の多少のみで相対比較もなしに結論を出している研究も見られ，相関係数を求めて，差を見比べるという統計手法を用いたり，また，因果を推測したりする解析的手法を用いるものは，限られた範囲の中でもさらに，少なかった。インターネットで入手できる先行研究が，現在発行されている先行研究のある種のサンプル的なものと仮定してみると，統計的手法を十分生かして検定，相関，因果関係分析をするという，統計的な手法を用いた日本語テキストの研究は，まだまだそれほど多くないと言えるということなのかもしれない。したがって，統計手法の種類が限定的に一般化しているとはまだ言えないのかもしれないが，手法毎にどのような目的でテキスト分析ができるかという可能性を探るのも興味深い。

統計手法として言語データの分析に利用できない理由は何か。また，利用手法の検討を試みることを繰り返し，言語研究の統計利用の範囲を明確にしつつ，新手法や新モデルを利用し，それらを相互に検証しあうことが，言語研究の可能性の拡大を試みることになるだろう。そして，限定的な統計手法の利用とは別の話になるが，個人の開発した統計モデル等，統計計算の手法を様々に工夫したものも多いが，荻野（2006）が指摘しているように「やりっぱなし」で捨て置かれる統計的手法の散逸も見られる。もちろん，それは，日本語テキスト分析内容の結果報告と，手法として用いた統計的計算法の違いについての，それぞれに報告する場所が異なっているという，発表分野の区別によるのかもしれない。

しかし，定量的研究において，統計を用い，同時に，その手法を検討していくという，統計を共有する姿勢が，今後の文系研究者が課題として考えるべきものである。そして，それは，できれば，それぞれの研究分野で，手法のセクションを設けて行われることが望ましいのではないか。分野を越えた情報交換や手法比較の検討結果についての報告会の融合が進むことを今後に期待したい。

3. 樺島の品詞構成比率

文学における文体論において，統計的手法を用い，指標モデルを考案して利用している期の代表として，樺島・寿岳（1965）の「文体の統計的観察」がある。小林（2005）が樺島・寿岳（1965）を指して「分析項目が多岐にわたり，かつ，項目のバランスがよく有意性を保っているので，安定した結果を得やすい」としているように，計量的文体分析を行う場合に引用されることの多い論文である。

文体論では，あるテキストに顕著に出現する特徴語や，品詞構成で，比率という観点から分析が

行われることが多い。本節では，樺島・寿岳（1965）の手法を紹介することで文体論の分野における語彙の計測と標準化の方法を追体験し，その可能性の範囲を考察するが，樺島・寿岳（1965）には，語彙構成についての分析手法について詳細には取り上げられていない。

そこで，別途，各テキスト内の語彙構成についての分析を行うが，それには，品詞構成率とは別の方法でより詳細な分析を各テキスト内に対して行う必要がある。今回は，フリーのテキストマイニングシステム KH Coder を利用し，テキストの特徴語や語の出現状況から因子分析を行った結果をもとに，樺島の方法での文体分析とマイニングによる文体分析を行い，樺島・寿岳の追体験の結果と合わせて分析してみる。

3.1. 文体の統計的観察法 MVR と名詞の関係から見た文体調査の意図

計量的な文体分析における樺島・寿岳（1965）の目的は，主観的な印象を客観評価することであった。そして，理想的な文体把握方法というのは質的分析点を数量化したものであるとするが，定義が困難であるとして，質的分析点を加工後，数量化することでより理想的な方法に近づこうという考え方で研究している。

また，計量語彙論では，実際には，作品を単に統計的に記述する立場の分析が多いと憂い，数える部分を明確に定義すること，ならびに，定義や計量にぶれを生じさせないことを第一に考えて計測，データ化を行っている。

樺島・寿岳（1965）は，文体を統計的に観察するための指標モデルを考案し，それに基づき，テキスト内の指標同士を検証して文体分析に応用している。その方法を概観すると，短編小説 100 編の各作品から無作為に 80 文ずつ抽出し，各々80 文の小規模コーパスを 100 種，合計 8,000 文からなるコーパスを用意し，それに対して 10 項目の指標の使用頻度を計量した後，その 10 項目の指標に基づいて短編小説 100 作品を比較する。そして，それぞれの差から作家の文体分析状況を考察しているが，そのときの指標は，名詞の比率，MVR（形，形動，副，連体/動詞数×100），指示詞の比率，字音語の比率，文の長さ，接続詞を持つ文の比率，引用文の比率，現在止めの文の比率，色彩語の比率，表情語の比率といった 10 種類の比率である。

この MVR（形，形動，副，連体/動詞数×100）値の大小を見て文体を推測するのであるが，MVR の値が大きいということは，動詞以外の自立語（品詞）が多く，様態記述中心の文章ということになり，MVR 値が小さければ，動詞が多く，動的な記述が中心の文章ということになるとして，これを用いることで数値データで客観的に簡略して文体が捉えられるというのである。これは，名詞が品詞比率の代表値として捉えられることを検証し，名詞と MVR 値を利用することによってテキストの性質を推測する指標にできることを確かめた結果によるものであるが（樺島，1963），この樺島（1963）の品詞構成比率がとる分布は，水谷（1977）の改訂でより明確になっている（伊藤，2002）。名詞とそれ以外の品詞の関係から，品詞構成，すなわち，記述文体を推定する手法である。

樺島・寿岳（1965）は，品詞構成から動きの多い文体かありさま中心の描写文体かどうかについての読者側の心的印象を追確認した分析を行って，描写の分類を行おうとしている。樺島・寿岳（1965）モデルの特徴は，名詞比率と，他品詞の比率との関係で記述の文体を予測できるという点にある。また，名詞以外を MVR という独自の指標モデルで表すことである。それは，品詞構成の比率が，日本語特有の語用の性質を持ちながらも，文章の差が顕著に現わすものであると考えられることによる。

3.2. 品詞構成比率と MVR の意味

ここでは，樺島・寿岳（1965）の『文体の科学』に基づいて，文体の定義と統計的観察法を概観する。

樺島・寿岳（1965）は，「文体」の定義が不確立であることから，作家作品の文体的個性を把握するために，独自に，表現特性について整理している。まず，文章を書くときの態度として，「事がらの骨組みだけを書く」か「事がらの細かい部分まで書こうとする」かの 2 つに分け，前者を要約的文章と呼び，後者を描写的文章と呼んでいる。

そして，要約的な文章の代表として，新聞記事，ラジオニュースを上げ，作家の文章にも，「いわゆる 5 W 1 H「いつ（When），どこで（Where），誰が（Who），なぜ（Why），どのように（How）」などをそろえたら新聞記事に近くなるような文章は要約的であると考えてよい」としている。

一方，「文章を読みながら，その内容を映画のシーンやさし絵を見るように想像することができる文章は描写的」だとしている。さらに，「描写的な文章にもいろいろある」として，以下の図 2 のように，表現のあり方を対比的に分類している。

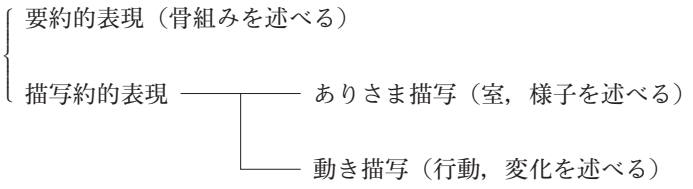

図 2　樺島・寿岳（1965）の表現定義

次に，「ある文章」が図 2 のどの分類に入るかを定めるのは，「我々が読んだときの感じである」が，「見分けを客観的なものさしによって行う」可能性を考えるために，描写的な文章が持つ性質と要約的な文章がもつ性質との間の違いを定量化して，文章の表現がどちらの分類に属するかを推定しようと試みている。

描写的な文章を元の文章よりも文字数が少ない文章に書き改めると，様々に書き換えられるが，共通して，要約的になっていることを上げ，文字数制限による書き直しの前後で生じた変化は，特に品詞の比率，中でも，意味を持つ自立語における名詞，動詞，形容詞類，接続詞類の 4 つの比率で明らかであるとして，その考えを 100 小説の文章の品詞比率を見て，実証し，結果から，平均的なものであることを断りつつも，名詞比率から他の品詞の組の比率の見当が付くとしている。ここでいう形容詞類とは，形容詞，形容動詞，副詞，連体詞のことで，接続詞類とは接続詞，感動詞である。

4 つの品詞グループに分けたが，感動詞と接続詞を合わせたグループの比率は全体の 5%を超えるものではないため，無視し，実際には，名詞と動詞と形容詞類の関係から判断することで十分だとして，3 グループの品詞間の関係と，それをまとめた MVR 値から分析を行う。MVR とは，形容詞類 M の百分率を動詞の百分率で割った値である。この MVR 値と，名詞比率 N%，ならびに，M や V の比率の関係で，文章の品詞比率を推測的に数値化するとともに，表現のあり方を推測するのであるが，M は形容詞類であることから，「ありさま」を表す語群であり，描写がありさま的か動き的かの判断を行う指標となっている。

また，名詞比率 N%が大きい文章は要約的で，逆に，N 率の小さい文章は描写的傾向にあり，MVR が大きい文章はありさま描写的で，小さい文章は動き描写的となる。以上，N と MVR から見た表現の傾向を以下のように記述している。表 1 はこの内容を簡略化してまとめたものである。

・名詞比率 N が大きく，MVR が小さい文章には要約的な文章が多い。
・N が小さく，MVR が大きい文章にはありさま描写的な文章が多い。
・N が小さく，MVR が小さい文章には動き描写的な文章が多い。

樺島・寿岳（1995，p. 35）

確かに，主観的に判断してランダムに採集した 9 作家の作品の品詞比率と傾向を図 2 のようにグラフにして見ると，評価領域に分けられる。樺島・寿岳は境界線を引いて領域を区別しているが，この境界は，読者の主観的判断による心象が反映されるもので，かつ，相対的である。

樺島・寿岳は，このように，例文への主観的印象を計量的に確かめるという方法で，以下のように，文体の分類の観点を 3 種類，表現のあり方を 2 種類に分類し，これらの組み合わせによって，1 つの作品の文体を捉える方法を提案しているが，これらは，質的分析であるとしている。

そして，質的分析による文体把握とは異なる方法として統計的把握の方法をあげ，文体の客観的分析が行えるとしながら，表 3 にあげられる観点を数量化して統計的に扱うことで，文体分析を行おうとしている。ただし，分析時にはテキストのごく平均的な姿しか見られないことを断っている。また，比率で見る場合，ごく小さい比率になった場合，その値に対する判断に疑問が生じやすいことから，統計的特性地の大きさを評価する 5 段階尺度を作成し，表 4 のように整理している。

分析する際の指標として，語彙に関する項目 6 点と文に関する項目 4 点の合計 10 項目について，指標の定義と指標採用の目的を，樺島・寿岳から読み取れた範囲で以下にまとめる（pp. 122–128）。

樺島らがこの 10 項目を指標にしたのは，文体の特異性を個性として採用する場合，平均から外

表 1　N 率と MVR の大小で判断できる表現の傾向

N＞MVR	N＜MVR	小 N/MVR
要約的	描写的 ありさま描写的	動き描写的

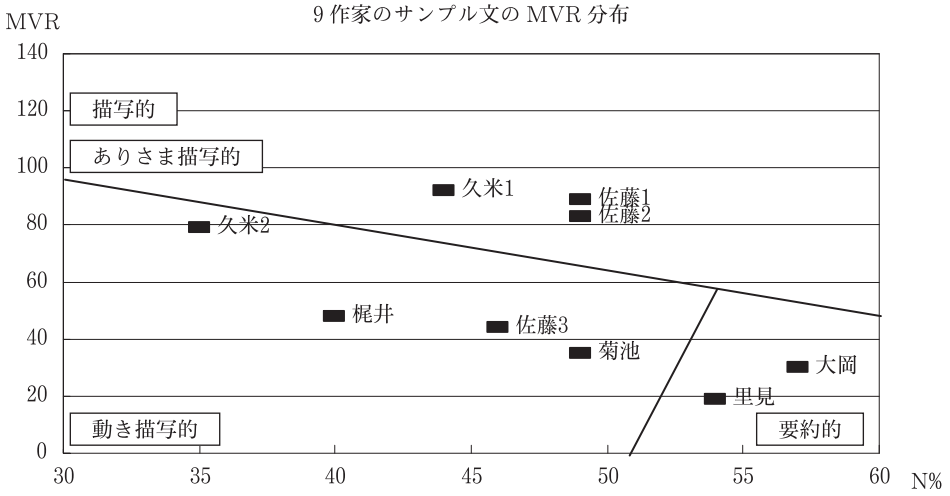

図 2　表現分類のサンプル検証結果

表 2　N%と MVR 値による 9 作品の表現評価

番　号	作　者	評　語	N	MVR
1	大岡昇平	要約的	57	30
2	佐藤春夫 1	（ありさま）描写的	49	89
3	梶井基次郎	動き描写的	40	48
4	里見弴	要約的	54	19
5	久米正雄 1	ありさま描写的	44	92
6	久米正雄 2	動き描写的	35	79
7	佐藤春夫 2	ありさま描写的	49	83
8	佐藤春夫 3	動き描写的	46	44
9	菊池寛	動き描写的	49	35

表 3　表現のあり方

	あり方	タイプ 1	タイプ 2
文体	要約的	ありさま描写的	動き描写的
	散漫的	饒舌的	凝縮的
	説明型	記述型	
表現	明示的	暗示的	
	感化的	通達的	

表 4　統計的特性値判断の 5 段階尺度とそれぞれの例

評　語	極めて小	小	普通	大	極めて大
出現率	10%以下	30%以下		30%以下	10%以下
名詞比率	45	48		54	56
MVR	34	41		55	65
指示詞率	2.1	2.8		5.0	6.0
字音語率	13	16		26	31
文長	7	9		14	18
引用文率	1	8		30	70
接続詞保有率	3	7		21	27

れている点を明らかにするために，文体的特徴を表すと考えられる全ての観点を採用したことによると考えられる。

⑴　名詞の比率・⑵　MVR

樺島・寿岳（1965）は，自立語総数のうちの名詞の比率を求め，ついで，自立語 4 種のうち，名詞以外の比率を抽象化するための指標値，MVR を，次の図 4 のような式で求める。

$$\mathrm{MVR} = \frac{\text{形容詞・形容動詞・副詞・連体詞の数}}{\text{動詞の数}} \times 100$$

図 4　MVR の計算方法

あるテキスト中の名詞の比率と名詞以外の自立語から求めたMVRは，負の相関があるが，これに基づいて，多数のテキストを比較すると，平均的な品詞構成率の文章か否か，また，その記述が要約的か描写的か，かつ，ありさま主体か動き主体かが推測できるとする。

樺島・寿岳の考えた⑴名詞比率と⑵のMVRの関係で，100小説のテキスト特徴の分布を見てみたのが図5である。左右の特異とされる位置と平均にある作家を明示してみた。

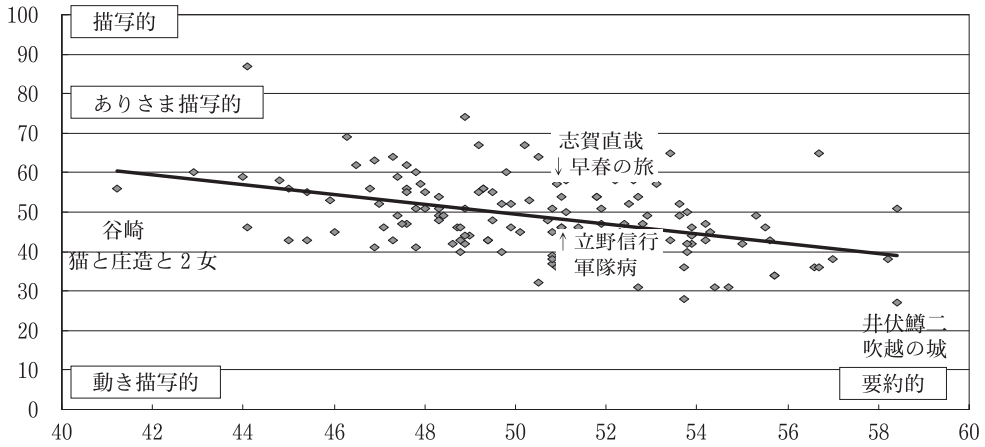

図5　品詞構成に基づく100小説の分布

⑶　指示詞の比率

いわゆるコソア語の自立語数中の比率である。「指示詞は同語の繰り返しをさけ，文と文の関係をつなぐ働きをする」ため，樺島は，この比率が多いと，「文脈への依存度が大きい文章で，冗文的になる」として求めている。そして，指示詞の比率と名詞の比率もかすかな負の相関があると言う。

樺島・寿岳（1965，p.123）では「かすか」の具体的な数値は示されていない。また，コソア語のうち，自立語として数えられるものは限られているが，「この」「その」「あの」等の連体詞と，「これ」「それ」「あれ」「あちら」「こんな」等の指示代名詞，「こう」「そう」等の副詞の区別は行われていない。

一般に，「自立語」は単独で文節を構成できるものとされ，名詞・代名詞・動詞・形容詞・形容動詞・連体詞・副詞・接続詞・感動詞の9種類が上げられることが多い。樺島・寿岳（1965）でも「自立語」には連体詞が含まれているが，感動詞に関しては不明である。

⑷　字音語の比率

漢字の音読み発音の語で，和製漢語も含まれる。文節単位で，全文節中の比率，すなわち，字音語を含む文節比率である。名詞比率が大きい作品は字音語比率が大きいという正の相関が見られると言う。

⑸　文の長さ（自立語数）

1文の自立語数を文の長さとして数え，1テキスト内の1文中の平均自立語数を求め，標本平均値と標本分散をあげる。

自立語数が多い，すなわち，意味のある部分を持つ箇所，すなわち，文が長い場合と言いながら，文の構造は問わないものとする。

⑹　引用文の比率

作品中の「」等のマーカーを用いて，会話や挿入を行う引用が行われる箇所の文字数を，引用文

と引用文以外の地の文から計算している。

$$\text{引用文比率（\%）} = \frac{\text{引用文の文字数}}{\text{（引用文＋地の文）の文字数}} \times 100$$

図 6　引用率の計算式

⑺　接続詞を持つ文の比率

文を相互に関係付ける接続詞の含む文の出現数が全体の文で占める比率のことである。

⑻　現在どめの文の比率

記述からはその定義の詳細が不明である。おそらく，文の終わり方をテンスという観点で見た場合，過去形か現在形，または，テンスフリーの形にわけられるが，このうち，動詞述語文で，現在形で終わる文末を持つ文の比率を数えたと考えられる。

⑼　色彩語の比率

小林英夫が和歌について求めた色彩語調査の結果，日本人の色彩語使用が白，赤，青の 3 つを使用することが安定性を持つ現象であるとしたことを踏まえ，樺島らは，100 小説中における色彩語の使用率を求めた。その結果，4.59±0.44‰（95%信頼区間）であることから，100 小説では，100 語辺りほぼ 4〜5 語の色彩語が現れることを確認し，その順位を示した。この結果は，小林（1953）の近代短歌の色彩調査結果と比較しても，日本人の色彩語使用が安定性のある結果を示すと考えている。おそらく，色彩語の割合がこの 100 小説の平均数値や順位と異なる場合は，文体的特徴と考えられることから，色彩語比率を確認するのが目的であるのではないかと考えられる。ここで上げられている小林英夫の和歌における色彩語調査の出典情報はない。

⑽　表情語の比率

擬声・擬態語のことを，表情語と呼び，これらの比率を求めた。樺島・寿岳は，発生的にはそこに属すが，その意識が薄れたものは対象外とし，逆に，発生的に擬態語ではなさそうであってもその意識を伴うものは対象に含めている。

4. 調査対象（テキスト属性）

一般に，話しことばと書きことばの性質的な違いについては，教育学的，発達心理学的，言語学的など様々な分野で様々な意見があるが，簡単に言うと，話しことばは即時的で，イメージを提示するために単語をそのまま発話するなどで文が短く，社会的な立場の違いを反映しやすい。一方，書きことばは，記録のために，５Ｗ１Ｈに関する内容が盛り込まれ，構文的にも正確さを優先するため，文が長くなる傾向があり，論理性が高く，内容面以外で記述者の社会的な立場の違いを反映することは少ないとされることが多い。

ということは，統語的な観点，品詞的な観点，語用論的な観点からその性質が分けられそうであるが，このうち，どのような品詞が多いかが，最も，差が分かりにくいものではないかと考えられる。差が生じにくいというのは，そこに日本語特有の性質が現れる場所だからであるが，しかし，そのように差が生じにくいところに差が見られるということは，逆に言えば，最も意味が明確な差であるということでもある。したがって，品詞に着目してその差を見ることは，小さな差であっても意味が大きいと言え，話しことばと書きことばの性質の差を見るのには，品詞が良いのではないかと考えられる。

話しことばと書きことばの性質の違いを見るためのテキストとしては，いくつか考えられるが，

インタラクティブなもので，かつ，ある程度，傾向のまとまったものが良いのではないかと考え，今回は，メールの自己紹介文，会話テキスト，小説，作文，小学校教科書5教科6学年分の品詞構成比率を比較してみる。

メールの自己紹介文は，H県の国立大学の情報授業のメール設定の授業で課題として出された文系，理系合わせて120人分の自己紹介の文章である。記述時間と分量の指定は授業により異なるため，等質に条件が指定されているものではない。傾向として共通するのは，記述者の「自己紹介」のためという認識と，テキスト入力，ならびに，メール送信練習のためという目的意識となる。そのため，収集年と学科という条件で計4つのグループ単位で扱う。

話しことばは，「名大会話コーパス」（大曾2003）の2008年改訂版のうち，10代から60代までの女性のデータを用いる。本コーパスは男女合わせた会話コーパスであるが，男性の割合と分布領域に偏りがあるため，20代と4，50代に若干名存在する男性を除き，延べ142人分の女性データのみを利用している。名大コーパスも，メール同様に，年代別グループ単位でまとめて扱う。

書きことばとしては，小学校の教科書の文章と，読者を意識した文章で，かつ，インタラクティブな現象を記述したものを3タイプ選んだ。1つは，読売新聞主催の作文コンクールで優秀賞をとった小学1年生から6年生までの5年分の作文約60人分である。小学生の作文の記述分量は，厳密には同じ分量ではなく，テーマも多様である。また，年齢による差があると想定されることから，低学年と高学年の2つのグループにわける。書きことばの2つ目には，均一的な記述のものを選んだ。それは，H県の大学1年生が学期末に書いた300語程度の意見文84人分をひとまとまりのグループにしたものである。記述時間を約30分とし，分量を300語程度に限定し，テーマをアルバイトの是非に統一している。

そして3つ目は，会話やナレーションをインタラクティブな現象と考え，小説を用いた。ただし，小説は，樺島・寿岳（1965）が提示している巻末の100小説の数値データ（pp.219–224）のみを利用しており，実際の小説の語彙リストは作成していない。小説の年代は，1965年以前のものとなり，若干，現在の小説の会話とは性質が異なるであろうが，小説の会話文に見られる男女や社会的地位を表す位相差を抽象化した表現は，現在の小説にも残るものであるため（金水，2008），会話との比較としては問題ないと考えた。また，小説データのみ，グループではなく，個々の作家1人ずつのMVR値である。1テキストの単位が異なるが，他よりテキスト分量が多いこと，グループ単位と個別単位の違いを確認できるとして，そのまま用いることにした。

電子メール文と大学生作文は研究目的で借用したデータで，一般公開されていないが，それ以外は，入手可能なものである。同様のデータを用いて追認する場合，電子メールの代わりに，インターネットに公開されているブログの記述が利用できるだろう。また，作文は，石川他（2009予定）に添付された巻末データにある作文を利用することができる。名大コーパスと読売新聞の奨学生作文に関しては，研究目的での使用であることを申し沿えて借用依頼をすれば利用させていただける。小説のデータは，樺島・寿岳（1965）の巻末資料として公開されている数値データである。以上，比較するテキストは全部で3種類5タイプである。表6にまとめる。

5. 調査方法

テキスト分析に関するリサーチクエスチョンは，メールの文章は会話や小説などの書きことばとは相違するのか，また，相違するとすれば，どちらに近い品詞構成率を持つのか，というものである。そして，手法に関しては，品詞構成比率の有効性検証である。

表 6　利用データ

電子メール	兵庫県 K 大学の新入生が自己紹介として書いた 10 年間のメールの文章のうち，男女大体同数にあわせて合計 120 人分を利用する。年代の偏りをなくくすために，対極的に 95 年度と 03 年度の文章を利用する。また，同様に学生の専門の偏りに配慮して，文系 2 学科分を利用する。
会　話　文	名古屋大学作成の会話コーパス「名大コーパス」のうち，10 代から 60 代の女性どうしの会話を会話者別にまとめたデータを利用する。10 代 10 人，20 代 66 人，30 代 23 人，40 代 16 人，50 代 18 人，60 代 10 人分の合計 143 人分である。
作　　文	3 つのタイプがある。1 つ目は，読売新聞小学生作文コンクールで優秀賞を取った児童の作文のうち 123 年生という，低学年約 30 人分の作文グループで，3 つ目は小学生の優秀作文 456 年生約 30 人分のグループである。 2 つ目は，H 県の K 大学の学生 60 人分の意見文のグループで，18 歳から 20 歳の学生の記述である。
小　　説	樺島・寿岳の品詞データのみを参照する。今回は樺島・寿岳（1965）が pp. 219–224 であげている作家 100 人の品詞集計結果と，そこからから得られた数値データである。

5.1. 研究の手順

まず，それぞれのテキストの語彙リストを作成し，そして，得られた品詞頻度から，品詞構成比率を求める。品詞構成比率とは，名詞，動詞，形容詞，形容動詞，副詞といった実質語の構成比率であるが，その捉え方には様々な立場があり，ここでは，樺島・寿岳（1965）の手法に倣って，名詞，動詞と，それ以外の形容詞，形容動詞，副詞を合わせた 3 種類に分け，それらを組み合わせて計算する MVR という比率を利用する。この MVR 値を品詞構成比率の代表と考え，この値に対して相関分析を行うものである。樺島（1963）は，文体推測指標として，「名詞＋MVR」の利用を提唱しているが，これは，「名詞が品詞比率の代表値として捉えられることを検証し，名詞と MVR（Modifying words and Verb Ratio: MVR: 筆者推測）という品詞比率の分布を調べることによってテキストの性質を推測する方法」で，名詞と MVR の関係から表現スタイルを推測するものだとしている。MVR は，「形容詞・形容動詞・副詞・連体詞の数」を動詞の数で割り，それに 100 をかけたものであるが，これに，名詞の比率を考え合わせて文体的特徴を考えるというものである。

樺島（1963）では名詞と MVR の組み合わせから得られた分類から，様々な文体的特徴を名づけて提唱しているが，本節では，そのうち，名詞と MVR の純粋に数字上の構成比率を利用するのみとする。

5.2. 樺島・寿岳の定義の詳細化と分析項目の選別

本稿では，樺島・寿岳の文体分析を追認するが，基本的な姿勢を踏襲しながらも，利用する指標のいくつかで定義を変更している。

まず，樺島・寿岳の 10 項目であるが，量的な分析だけでは，その意図する質的特徴の予測が困難だと思われるもの，また，定義があいまいで追認の際，計量しづらいもの，そして，その方法で調べた数字で意図する分析が行えるのか疑問があるものがある。疑問点をまとめると，表 7 のようになる。

以上から，本稿での計量的な分類は，名詞と MVR のみを用いたもので行うことにする。また，連体詞や感動詞は，一般に自立語には含まれないこと，さらに，接続詞はその数量から MVR の決定に影響を与えるものではないことを踏まえて，自立語として計上するものは，名詞，動詞，形容

表 7　樺島・寿岳指標の疑問

不明点	項　　目	疑　　問　　詳　　細
定義	⑶指示詞率	ごく小規模の小説コーパスでは増減が明確になると考えられるが，テキスト量が増えれば，指示詞頻度がごくわずかとなるはずである。また，会話コーパスを利用した場合，文脈指示と現場指示は別のものだが，計量的には比較できない。さらに，形態素解析上，連体詞や感動詞の一部は解析ミスが多い。以上から対象外としたい
定義	⑻現代どめ文数率	動詞述語文末のことなのか，連体修飾節の動詞述語の区別はどの程度行われているのか不明
定義 意図	⑼色彩語率	⑶指示詞率と同様に，テキスト量が増えれば，有意な差が見られるか不明。また，形容詞と名詞にまたがる計量となり，数値に重なりが出るのではないか
定義 意図	⑽表情語率	何を表情語とするかの識別定義が不明瞭である。これは副詞の大部分と重なるものではないか
意図	⑷字音語率	音読みの漢字が増えれば名詞が増えるという因果関係は想定できるが，その増減が個人で差があることから増減の因果は関係がないのではないか。音読みの漢字は難易度に関係があるが，これ１つだけでは難易度を判断することはできず，さほど多くの情報は得られないのではないか
意図	⑸文　　長	文の複雑さは自立語数が多いことだけでは判断できない。また，自立語数だけでなく，付属語数との割合で見なければ，長さや複雑さがわかりにくいのではないか。これだけではあまり多くの情報が得られないのではないか
意図	⑹引用率	小説以外では，「」は強調マーカーとなり，意味が異なる。また，小説以外ではその数は少なく，判断材料にならないのではないか
定義	⑺接続詞含有文率	接続詞が，文間のものか文中のものか，接続詞の定義がよくわからない。また，文章構成における質的分析には良いだろうが，量的には，接続詞の内容や使い方がわからないと分析が不十分ではないか

詞，形容動詞，副詞の 5 品詞とし，名詞は，接尾辞系や数詞類を除外したものとする。

5.3. 語彙リスト作成方法

コーパスを用いて行う計量的な文体研究でも，指標の実測値を計上するところから始めるのであるが，その方法について，樺島・寿岳（1965）では明確にされていない。本節での，語彙計量のための下処理は，形態素解析システム『茶筌』を組み込んでいるテキストマイニングシステム KH Coder を利用する。

日本語の文章はデータとして二次利用する場合には，通常，形態素解析を行って単語単位に分けることで，同音語や同形語が多く，分かち書きされていないという日本語の特性を補う下処理が必須である。電子化されたデータを形態素解析で分かち書きし，同じ語の数を数えてソート後，語彙リストに整理するのであるが，今回は，形態素解析システム『茶筌』を組み込んだ KHCoder を利用することで簡便化を諮る。

得られた語彙リストから品詞別の語彙数を集計し，品詞構成比率を計算する。今回は，『茶筌』で品詞タグをつけた語彙のうち，自立語とした名詞，動詞，形容詞，副詞とされる品詞タグを持つ語を選び，全語彙リストを作成している。この後，集計結果を Excel® に読み込み，品詞別に集計して品詞構成比率を計算する。

なお，KH Coder では，品詞を『茶筌』に準拠する。品詞は IPA 辞書に基づで，基本的に学校

文法に準拠している。しかし，形容動詞の一部は形容詞または名詞に分かれて含まれたり，副詞が詳細に細分化されているなど，一部の扱い方が学校文法の品詞とは異なっている。『茶筌』は，品詞が 3 層からなっており，第 1 層で，名詞，動詞，形容詞，副詞といった区別がなされるが，第 2 層で活用や接続に基づき，細分化されている。これを利用して選別することで，名詞は，この第 2 層がサ変接続，一般，代名詞，固有名詞，副詞可能となっているものを指定する。また，第 1 層が名詞であっても，第 2 層が形容動詞語幹の語は，いわゆる，形容動詞として別途計上している。そのため，厳密には，［名詞―形容詞的用法］となるものは名詞ではなく，形容詞でカウントしている。したがって，本稿で形容詞と呼ぶものは，品詞タグの第 2 層が「自立」のものと名詞の形容動詞語幹の語となる。動詞は第 2 層が自立となるもののみを採択し，副詞は区別せず全て採用している。その結果，動詞は第 2 層で自立語に指定されているもののみを採択することになることから，第 2 層が非自立のものは「する」も対象外となっている。接続詞や連体詞は 3. 4. 1. に述べた理由で今回は対象外とする。

また，『茶筌』は形態素毎の出力が標準であるため，たとえば，「アルバイト」を「アル」「バイト」と 2 語で認識することもある。そこで，KH Coder で語彙リストを作成する際，KH Coder に組み込まれた複合語抽出システムを利用して，先に複合語選別を行っておく。

樺島・寿岳（1965）は，語彙の実測値に対して標準化を行うということをせず，テキストをあらかじめ平均化することや，分析するための指標を抽象化するなどの方法で分析を進めている。相関分析では，数値を raw データのまま利用しても同じであることから，いずれのテキストも分量が異なるが，そのジャンル全体から偶然的に抽出されたサンプル相当と考えることにし，改めてサンプル文を抜き出すということはせず，それぞれのテキストを全数調査した結果を利用する。

5. 4. テキストの処理結果

電子メールの文章から得られた品詞数は表 8 のようであった。左側の実数と合計数から比率を計算したものが表の右側にあり，一番端に樺島（1963）の品詞構成比率 MVR 値を出している。会話から得られた品詞内訳は表 9 の通りである。同様に，表 10 に作文をまとめる。なお，今回利用する 6 人 9 冊の小説は名詞，動詞の実数が公開されておらず，名詞比率と MVR のみとなる。また，非公開データである小学校教科書の名詞比率と MVR のみを参照データとして利用し，相関分析に適した形に整理して表 11 のようにまとめた。

6. 結　果

表 11（スペースの関係で折り曲げている）のデータから散布図を作成したものが，図 7 である。すなわち，名詞と MVR の値で作成した相関図である。どのドットが何を指すかがわかりやすいよ

表 8　電子メールの文章の品詞数と構成比

自己紹介	計	名詞	動詞	形・副	N%	V%	M%	MVR
95 コ	3,059	1,979	544	536	65	18	18	99
03 コ	13,136	7,910	2,713	2,513	60	21	19	93
95 地	8,570	5,108	1,833	1,629	60	21	19	89
03 地	13,356	8,117	2,569	2,670	61	19	20	104

表 9　会話の品詞数と構成比

MC	計	名詞	動詞	形・副	N%	V%	M%	MVR
10 代	13,976	5,783	4,720	3,473	41	34	25	73.6
20 代	1 E＋05	44,111	31,692	24,234	44	32	24	76.5
30 代	22,297	9,536	7,134	5,627	43	32	25	78.9
40 代	19,566	9,193	5,965	4,408	47	30	23	73.9
50 代	13,963	7,186	3,911	2,866	51	28	21	73.3
60 代	22,135	10,224	6,865	5,046	46	31	23	73.5

表 10　作文の品詞数と構成比

	計	名詞	動詞	形・副	N%	V%	M%	MVR
大学生意見文	9,538	5,108	3,124	1,306	54	33	14	42
小学低学年 1–3 作文	22,037	10,686	7,811	3,540	48	35	16	45
小学高学年 4–6 作文	928	503	286	139	54	31	15	49

表 11　各データを変数とし，名詞比率と MVR をサンプルとしてみる相関表

	N%	MVR		N%	MVR		N%	MVR
小低作 1–3	48	45	小説（佐 1）	49	89	会 20 代	44	76
小高作 4–6	54	49	小説（梶）	40	48	会 30 代	43	79
大学生作	54	42	小説（里）	54	19	会 40 代	47	74
教国語	45	97	小説（久 1）	44	92	会 50 代	51	73
教算数	48	85	小説（久 2）	35	79	会 60 代	46	74
教理科	47	68	小説（佐 2）	49	83	メール 95 コミ	65	99
教社会	46	132	小説（佐 3）	46	44	メール 03 コミ	60	93
教生活	40	56	小説（菊）	49	35	メール 95 地域	60	89
小説（大）	57	30	会 10 代	41	74	メール 03 地域	61	104

うに後から項目名を追加している。

小説と，会話，作文，教科書，電子メールの文章は，全体で見ると無相関のように見える。しかし，メールの文章を除くと，話しことばと書きことばでは，ゆるやかな，負の相関がありそうに見える。そこで，メール以外のテキストの名詞率と MVR 値の相関係数を求めたところ，−0.4969 となる（四捨五入すると，−0.5)。緩やかな負の相関が確認できる。

また，出現する位置が興味深い。会話，メール，作文は，おおよそまとまった位置に現れるのであるが，メール，社会科の位置は他より外れている。小説は，個別データを参考にあげたもので，会話，教科書，作文の散布傾向を位置づけるために参照するのみのものである。他のテキストと同様のレベルで比較できるものではないが，MVR の散布図の分布領域からみると，小説と会話，作文とは分布領域が重なっており，語彙的性質に大きな違いがないという判断が可能だと考えられる。

なお，外れ値と考えられるメールは，会話や作文，それをカバーする小説のグループとは異なり，

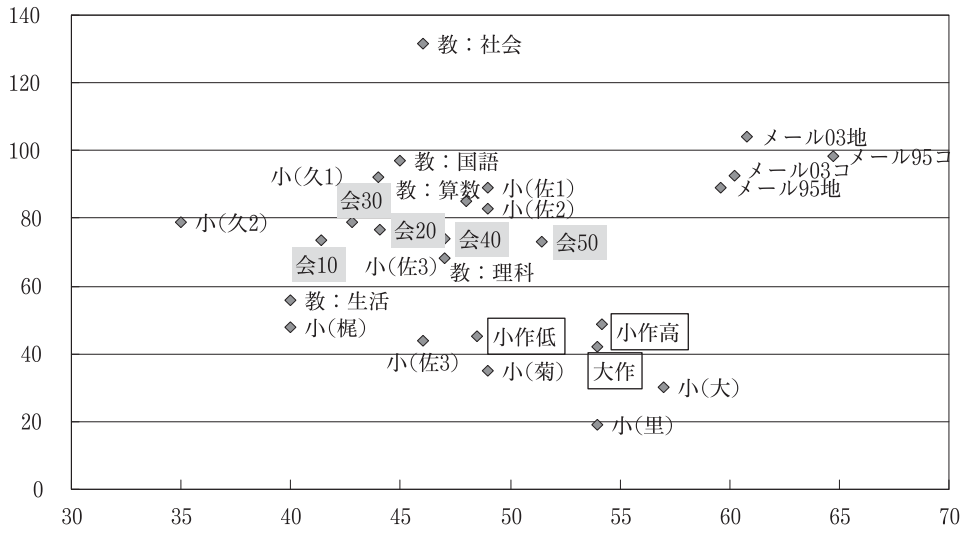

図7　名詞と MVR の相関図

全く別の語彙の性質があると言えそうである。

書きことばの代表を作文や小説，話しことばの代表を会話と考え，さらに，一般な心象にしたがって，メールは会話と書きことばの中間だと予想していたが，メールは，外れ値として除外して考えるべき対象となるほど，全く，話しことばとも，書きことばとも異なる品詞体系を持つ可能性のあることがわかり，メールの文章の日本語としての特異性が明らかになっている。

さらに，会話テキストは，書きことばテキストと比較しても使用語彙における品詞の使用状況に大きな差がないことも明らかになった。

7. 考　察

インターネットやメール上で記述されることばは，話しことば的書きことばだと言われるが，実際には，品詞構成比においては話しことばと書きことばのどちらの性質とも異なることがわかった。

つまり，メールの文章は会話や小説などの書きことばとどちらに近い品詞構成率を持つものではなく，別途，詳細に検討しなければならない，新たな問題を孕む，特異な品詞構成を持つものであると考えられた。

また，60 年代の先行研究，樺島・寿岳（1965）の MVR を応用し，テキストのジャンル別の表現スタイルの差を比較したが，MVR 値を利用して日本語分析や日本語教育分野での応用の可能性を検討した結果，樺島の定義した自立語を利用せずとも，名詞と動詞に加えて，形容詞，副詞のみでも，その効果が発揮されることから，品詞構成率だけでもある程度，文体の分類が行えそうであることが確認できた。

ただし，MVR は，日本語の語彙の性質をよく知った上で提案された分析方法であることがわかり，改めて，日本語にあった統計的手法というものが存在すると確認できたが，どの程度日本語の語彙特性に特化した現象かどうかについては，内外の品詞構成率に関する先行研究をあたり，さらに，検証を進めて確認する必要があるだろう。

今回，フリーのテキストマイニングシステム KH Coder を利用したが，日本語テキストの下準

備の煩雑さを回避して，語彙リストを作成すること，さらに，語彙をグループ分けして語彙構成の大枠を見出すのに，簡便に，かつ，安価に統計的分析が可能になったことを感謝するに至った。

しかし，マイニングシステムは，概要を見るのに良いが，統計的手法の差を十分に理解していない場合，その質的分析に至る過程で，論を検証しながら進むという研究過程が見えにくい。樺島・寿岳の研究手法とは逆の方向だと言えるだろうが，可能性を整理すべく，さらに詳細かつ多様な方法で利用してみる必要がある。

8. おわりに

今回の分析で，話しことばや書きことばは，一般に考えられているように，同じ日本語ではあるが，品詞の使用においても，性質の異なるものであることが伺えた。しかし，一方で，いわゆる，話しことばや書きことばという括りは，文化的，心理的なものであり，言語的には厳密なグループ分けにはならないものである可能性も考えられる。もちろん，今回利用したデータに偏りがあることから，今回の相関分析のみでは一概には言えず，詳細を更に検討していく必要があるのは言うまでもない。

さらに，本稿では品詞構成による文体分析の可能性検証をもう一つの課題としたが，過去の国語学の分析手法は，日本語の特徴に合った分析方法であると，品詞構成率に基づく分析に対しての認識を新たにした。

先行の計量国語学的研究と，昨今のテキスト分析における世界標準型統計手法との差については，両方の良さをどのように生かすかにおいて，さらに検証を重ねる必要があるとも感じる。従来の日本語の文体研究の流れにおいては，テキスト特徴詳細化には主観的なものがあり，初めに結論があって，その自論をある程度客観視するために統計が利用される場合もあった。文体記述や文体特長の分類における文体自体の判断が研究者により若干異なるスケールで識別されることからユニークなものとなっていたが，追認しにくい点も否定できない。

なお，今回，収集した先行研究は，インターネットを経由して，研究機関の論文データベースから入手することが安易なものの中で局所的に調べたため，偏りがあると考えられる。非常に限られた方法で概観したものではあるが，今回の語彙分析に関する限られた範囲で見た限りでも，統計と言いながら，実数を計上し，実数の多少のみで相対比較もなしに結論を出している研究も見られた。相関係数を求めて，差を見比べるという統計手法や，また，因果関係を推測する解析手法を用いるものは，限られた範囲の中でもさらに，少なかった。インターネットネットで入手できる先行研究が，現在発行されている先行研究のある種のサンプル的なものと乱暴に仮定してみると，統計的手法を十分生かして検定，相関，因果関係分析をするという，統計的な手法を用いた日本語テキストの研究は，まだまだそれほど多くないと言えるということなのかもしれない。したがって，統計手法の種類が限定的に一般化しているとはまだ言えないのかもしれないが，手法毎にどのような目的でテキスト分析ができるかという可能性を探るのも興味深い。本稿でみたように統計的手法に基づく日本語テキスト分析の先行研究を見直すとともに，新たな技術の恩恵を合わせて利用していくことは，統計手法に基づく分析が研究の際の基本的な位置づけを決めるための客観性向上に生かされる可能性が確認できたことでもあり，意義深い。今後も手法の検証を重ねたケーススタディーを繰り返し，先の知見をより客観的に利用する方法を考えていく必要がある。

本研究は，2008 年度統計数理研究所共同研究レポート 232 に発表した原稿を加筆修正したものである。

参考文献

石川慎一郎（2008）『英語コーパスと言語教育』大修館書店.
伊藤雅光（2002）『計量言語学入門』大修館書店.
上田博人（1998）『パソコンによる外国語研究（Ⅰ）数値データの処理』くろしお出版.
内田修・菅民郎・高橋信（2003）『文系にもよくわかる多変量解析』東京図書.
大野晋（1956）「基本語彙に関する二三の研究」『国語学』24，pp. 34–46.
荻野綱男（2002）「計量言語学の観点から見た語彙研究」『国語学』Vol. 53, No. 1, pp. 97–115.
要弥由美・小澤伊久美（2008）「統計は怖くない！　図を見てわかる直感的統計分析 ── 論文理解のための構造方程式モデリング（SEM）入門 ──」WEB 版『日本語教育実践研究フォーラム報告』http://wwwsoc.nii.ac.jp/nkg/kenkyu/Forumhoukoku/2008kaname.pdf.
樺島忠夫（1963）『表現論 ── ことばと言語行動』綜芸舎.
樺島忠夫・寿岳章子（1965）『文体の科学』綜芸舎.
北研二・津田和彦・獅々堀正幹（2002）『情報検索アルゴリズム』共立出版.
形態素解析システム『茶筌』；http://chasen.naist.jp/
小池清治・鈴木啓子・松井貴子（2005）『シリーズ日本語探求法 6 文体探求法』朝倉書店.
国立国語研究所（1964）『現代雑誌九十種の用語用字（3）』秀英出版.
小林千草（2005）『文章・文体から入る日本語学』武蔵野書院.
小林英夫（1953）「詩人の感覚 ── とくにハクシューについて」『言葉の心理』宮城音弥編，河出書房.
近藤陽介，松吉俊，佐藤理史（2008）「教科書コーパスを用いた日本語テキストの難易度推定」言語処理学会第 14 回年次大会発表論文集，pp. 1113–1116.
専門用語自動抽出システム（中川裕志，前田朗，小島浩之）；http://gensen.dl.itc.u-tokyo.ac.jp/
玉岡賀津雄・松下達彦・元田静（2005）「日本語版 Can-do Scale はどれくらい正確に日本語能力を測定しうるか：How accurately does a Japanese version of a Can-do Scale measure Japanese language ability?」広島大学留学生教育：Journal of International education, International Student Center, Hiroshima University, Vol. 9, pp. 65–78.
テキストマイニングシステム KH Coder（樋口耕一）；http://khc.sourceforge.net/
中尾桂子（2009）「文体的特徴に基づくテキスト分析の可能性 ── 品詞分析モデル指標 MVR を利用した文体論の言語教育への応用」『コーパス言語研究における量的データ処理のための統計的手法の概観』統計数理研究所共同研究リポート 232, pp. 65–84.
丸山岳彦・田野村忠温（2007）「コーパス言語学の射程」『日本語科学』22，pp. 5–12，国立国語研究所.
水谷静夫（1977）「語彙の量的構造」『岩波講座日本語 9　語彙と意味』岩波書店.
水谷静夫（1983）『朝倉日本語新講座 2　語彙』朝倉書店.
村上京子（2005）「作文評価における文の種類の影響 ── 意見文と説明文の比較 ──」『日本留学試験における記述問題の実施方法と分析観点に関する実証的研究 ── 記述問題の問題形式・量及び評価基準の適正さについて ──』2003・2004 年度文部科学省科学研究費補助金萌芽研究 15652032（研究代表者：村上京子）研究成果報告書.
名大会話コーパス；http://tell.fll.purdue.edu/chakoshi/meidai-chuui.html
馬場由佳・三宅清・馬場充・吉田則夫（2000）「源氏物語における「ずは」の構文的位置づけに関する計量的分析」『計量国語学』22–4，pp. 147–156.
李在鎬（2002）「構文の意味的拡張に基づく第二言語の文法習得 ── コーパスの定量的分析に基づいて ──」，『言語科学論集』（京都大学）No. 8，pp. 99–127.
李在鎬（2004）「助詞「に」の定量的分析への試み:語法研究の新たな手法を求めて」，『日本認知言語学会論文集』No. 4，pp. 55–65.
李在鎬，井佐原均（2005）「統計モデルを用いた助詞「で」の分析」関西言語学会第 30 回記念大会研究発表（関西大学 2005. 6）.
李在鎬，井佐原均（2006）「第二言語獲得における助詞「に」の習得過程の定量的分析」，『計量国語学』第二十五巻四号，pp. 163–180.
李在鎬，黒田航，大谷直輝，井佐原均（2006）「名詞との共起関係に基づく構文の定義」，『認知言語学論文

集』No. 7, pp. 1–10.
山崎誠（2009）「国立国語研究所における諸研究 ── 語彙調査の系譜を中心にして ──」『国文学解釈と鑑賞』第 74 巻 1 号，至文堂.
安本美典（1963）『創作の秘密　作家の性格と心理』誠信書房.
安本美典（1985）『日本語の起源を探る　コンピュータがはかる“やまとことば”成立のモデル』徳間文庫.
読売新聞社 2002–2004「全国小・中学校作文コンクール　小学校一〜三年・四年〜六年」第 52 回–第 54 回.
Bruce Frey，鴨澤眞夫監訳，西沢直木訳 2007『STATISTICS HACKS ── 統計の基本と世界を測るテクニック ──』オライリージャパン.
Lee 凪子（2006）「留学生の書く日本語意見文の分析 ── 日本人学生との比較において ──」『立命館法学』別冊　ことばとそのひろがり(4) pp. 399–412.
Lepton（2008）『Lepton 先生の楽しく学べる統計』ソシム.
IPA 辞書 http://sourceforge.jp/projects/ipadic/
TermExtract；http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr-e.html