# 日本語テキストの計量的分析における統計手法



大妻国文　第41号　2010年3月

# 日本語テキストの計量的分析における統計手法

中　尾　桂　子

**あらまし**：日本語，日本語教育の分野での計量的なテキスト分析に利用される統計的手法を概観し，主に使われている統計手法を調べた。古い時代には，相関と類似性を判断するためのものとして，相関分析，回帰分析，数量化Ⅲ類，Ⅳ類が用いられており，その他は，独自に検証を繰り返して確証を得た日本テキスト分析用の独自の手法が多かった。昨今では，いわゆる，統計手法らしい相関，検定を中心に，単回帰分析や判別分析といった関係分析の手法や，因子分析，主成分分析，コレスポンデンス分析，クラスター分析，数量化Ⅲ類といった類似性判別のための分析が利用されることもあり，言語データに対する統計的手法の可能性を見るような研究も増えている。しかし，研究の多くは，言語分析が主眼のため，統計手法の選別とその根拠についての検証には紙面が割かれないことも多く，利用根拠が明示的でない研究も多かった。追認可能な形で，研究結果をまとめていくような姿勢が望まれたことから，統計が言語分析におけるごく一般的な基本手順の1つとして捉える姿勢や，言語分析におけるある種の利用規定のようなものが必要なのではないかと考えられた。

**キーワード**：日本語テキスト　計量的分析　統計手法

## Simple overview of statistical technique in Japanese text analysis:

**Summary**: This paper aims to provide a simple overview of the statistical methods for the text analysis in the field of Japanese and Japanese-language education. For the first stage the methods such as "correlation analysis", "regression analysis", "discriminant analysis", and "Hayashi's quantification method type Ⅲ" were used to judge the correlation and the similarity. But recently, the research that pursues the possibility of the statistical technique to the language data has increased. And the

technique of the relationship analysis such as "correlation coefficient", the test for "significant difference", "factor analysis", "principle component", "correspondence analysis", "cluster analysis", and the analysis for the similarity distinction of "Hayashi's quantification method type II" or "Hayashi's quantification method type III" might be used centering on the correlation and the authorization of a statistical technique. Through this overview, it found that it is very important to state the reason of the choice on the statistical methods, and it should be established the rule when a statistical technique was used as one of the very general procedures in the linguistic analysis.

**Keywords**: Japanese text data, measuring analysis, statistical method

## 1. はじめに

日本語のテキスト分析について，丸山・田野村（2007）は，電子化された大量テキストを用いて研究する「コーパス日本語学」の流れを，その研究環境に応じて3期に分類している。第1期は国立国語研究所（以下，国研）主導の大規模調査等が盛んに行われた時期で，第2期はコンピュータの普及に伴う個人研究が開始された時期である。第1期と第2期の違いは，国研主導の大規模調査であるか，または，ある特定のコーパスについて個人が分析できるようになったかにあり，第3期は，個人でも大規模調査が可能な環境を国立国語研究所主導で大規模均衡コーパスを整備することにより可能ならしめる時期であると言う。

丸山・田能村（2007）で3期に分類されたコーパス研究環境は，そのまま，日本語の計量的分析における統計手法の違いをも指す。丸山・田能村（2007）の第1期，すなわち，国研主導で大規模調査を行っていた時代は，世界的にも同種の研究としては早い時期のものであったことから，国研の厳密なサンプリングと厳密な統計手法の検証をベースに行われた日本独自の研究が多く，日本語の言語特性を表すモデルや，計量的手法のモデル化が数多く行われていた。これら第1期の手法検証の結果が，先行研究として，コーパス第2期にも受け継がれるが，実は，第2期は前半と後半に分けられる。そして，第1期から第

2期前半までと，第2期後半以降の研究に利用される統計手法には性質の違いといった点で，乖離が見られる。

第2期の始めの頃の日本語テキストを用いたコーパス研究での統計手法は，第1期の国研の手法を踏襲して行うものが多く，第1期の流れを汲む研究では，統計手法の選別は自明の理として利用根拠を断らずとも，第1期の名残から，結果と分析が行われることが多かった。第2期のはじめ頃までは，日本語研究における統計分析では，国立国語研究所の手法と概念に準じて研究すればよいという意識が定着していたこと，また，同時に，他の手段は未検証であるとして排他されやすかった土壌のあったことが考えられるが，それは，1964年の語彙調査結果の発表で，日本語の定量的研究の方法と理論（計量語彙論）が確立したとされ，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きたと言う山崎（2009）の指摘からも推察される。

一方，国外では1964年のBROWNコーパスを皮切りに，イギリスやアメリカで，大規模コーパスが作成され，コーパス研究が盛んになり，計量的に有意な関係を検証するために統計的手法を利用することが一般的になっていた。海外でのコーパス研究で用いられる統計手法は，経済界，教育，心理，医学や社会学など，他分野でも利用される一般的な統計手法の中で，言語データに見合った手法が選別され，その根拠とともに踏襲されていた。

第2期に，徐々に，電子データとしての日本語を個別にコーパスとして利用する個人研究者が増えていったこと，また，英語コーパス研究や海外での手法が国内で多く見られるようになった結果，利用される統計手法が，第2期の前半と後半では異なりはじめた。最初，国内のコーパス研究で用いられていた統計手法を利用していたものの，海外でコーパス研究を学び，日本に持ち帰った研究者のスタンダードである「世界的に一般的だ」と言えそうな統計手法に触れた国内の個人研究者は，未検証手段が排他的に受け取られやすい国内のそれまでの研究から離れ，根拠の明確な提示と，世界的に一般的な統計知識を利用することで分析結果が確認しやすい国外のコーパス言語学の流れに沿う手法を用いることが多くなっていったのである。つまり，国内で独自に発展してきた

統計手法の層と海外スタンダード層との差が明確になり，徐々に，海外スタンダードの方が主流になっていったのであるが，そのため，分析目的と検証点は同様でも，利用される統計的手法が異なる，または，術語が異なるということになり，統計手法としては，時期的にも，また，コーパスの種類の上でも，関連するかのように，第1期から第2期前半と，第2期後半の間で手法に乖離が見られるようになったというわけである。

その後，第3期の時代に入った現在では，過去に国語学で培われてきた第1期の統計的法則は既に，ほとんど利用されなくなっており，また，より，研究目的が細分化された結果，他分野で開発，改良された様々な統計的手法が試されるようになり，汎用的なものが再び根拠を明確にせず，踏襲して利用されるようになっている。

本稿は，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，その歴史的な経緯と分野の区別に応じて「一般的」だと考えられる日本語テキスト分析の統計的手法について概観し，日本語研究で，定量的研究の変数や指標，利用する統計についての検証課題について考察する。

## 2．計量的な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，すなわち，目的が日本語の言語的分析以外にある研究としては，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ため，全体像は不明な部分も多い。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)，国研の一連の調査や，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。

コーパス言語学と，計量言語学，狭義の数理言語学，計算言語学の3つを含む数理言語学との違いは，ある言語現象を特定の現象か一般的なものかを判断するのに大量の電子データを参照するか否かということであるが，これらの分野の違いは，どちらかというと，コーパスを利用していることを強調するかしないか，研究者のスタンスがどこにあるかを区別することにあるようで，実際の研究は大差ないように見える。本稿では，テキストの規模にかかわらず，計量的に分析する手法自体を概観するため，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。

## 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，以下のような手順で分析を行い，統計値を利用する。

①検討する観点と，それを調べるために焦点を当てる指標を決める

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ことが多いことから，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以上の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したり，分散，有意水準，推定値など相関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，他のテキストと相対的に比較して差を見る

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っていく。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して定着している。

そのような統計的計算法は，言語テキスト以外でも培われてきたものが多い。統計ソフトなどに組み込まれているものもいくつかあるが，たとえば，SEM (Structural Equation Modeling: SEM) などである。



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。語彙の量的な分布で語彙の構造を分析す

るものとして，水谷（1977）は，法則性が捉えられるものについてだけということで次のようなものをあげている。まず，使用率の分布を見るものとして，「ジップの法則」，修正版の「ジップの第２法則」，マンデルブロの法則，水谷の法則である。これらは，二次関数に当てはめてその語彙の分布量をみるものである。次に，品詞の構成比を見る法則として，大野の法則，樺島の法則と，他に，法則でないが，出自別に見た語種の構成比百分率の出し方などをあげている。そして，語彙における語の関わり合い，すなわち，関連性や類似性を見るためには，水谷独自のブール代数計算や，林の数量化第Ⅲ類，第Ⅳ類をあげている。現在でも言語データに対して一般によく利用される統計的手法となったものは，林の数量化Ⅱ類，Ⅲ類といったもので，類似性や関連性の検証のための統計手法である。使用率や構成比に関する分析モデルは，日本独自のテキスト分析手法で，統計手法としては意識されなくなっている。

## 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い（内田他，2003）。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。この判断の根拠となる係数や計算結果で比較，確認する際，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，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（内田他，2003）。それが統計的分析手法として区別されているのである。

言語現象を取り扱った統計手法は，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，他は言語現象の分析での利用が少ない。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とから，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）も述べているように，関連性をみる分析は，データと目的の数だけ，知恵と工夫が必要とされるため，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，その可能性は常に検証していく必要がある。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って分析し，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，逆に，目的やデータを考慮せず，汎用的なモデルで分析したという事実で結果検証に対する納得を得ようとする場合もある。

以下，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，次章では，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。

## 5．言語分析に利用される統計

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いて理論化しようとする。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汎用されているが，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るが，これらでは，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て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。

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が，それが影響しているのであろう。

コーパス言語学での分析のための下準備で出す数値とは，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，英語等でよく利用される。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。

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が，データの性質や比較対象の違いにより，たとえば，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想定して，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る。検定でも，Z検定，G検定，F検定，ピアソンの$\chi^2$検定などを，区別して用いることができるようになっている。

コンコーダンサーで提示された下準備の結果は，文法規則を一般化したり，第二言語習得過程の状況を母語話者と比較したりするのに応用されるが，言語テキストの分析では，検定や，因子分析，回帰分析，分散分析，主成分分析といった多変量解析が行われることが多い。

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る。そして，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っている。そして，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と考えられる（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て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。

そ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。

ただし，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，本稿の目的である，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。

## 6．日本語テキストでの相関分析の手法とその対象

### 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，国研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。この国研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。

国研の語彙調査は，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。

国研の語彙調査や，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(3)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, 04, 06）。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。

### 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文体を予測できるという点にある。また，もう1つは，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化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。

樺島・寿岳（1965）は，縦軸にMVR値，横軸に自立語中の名詞の比率(%)を取って小説100作品におけるMVRと名詞比率で品詞構成率の分布を視覚的に見ることで，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い，描写の分類を行おうとした。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）ではその方法を明確にしていない。当時の単語認定は，国研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しているのかと思われるが，断りはない。



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。

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。

## 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。

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1167サンプル728002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。

### 6.4. 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 $\alpha$ 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で，ビジネスで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よく行われ，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。

## 7．日本語・日本語教育におけるテキストの統計的分析法と課題

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ことが一般的になってきたが，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，利用方法や呼称が分野によって異なる。



従来の日本語の文体研究の流れにおいては，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で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。

考察目的は同じでも，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。しかし，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もあるが，どちらかといえばより安定した定量化のための工夫が行われることにつながっている方が多い。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，類似性分析の応用として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。商品開発のためのアンケート記述を分析する目的で発達してきたテキストマイニングでは，より簡便さが求められてきた結果マイニングツールの開発と向上により，計量言語学や計算言語学の分野で培われた手法と同様のものをよりたやすく利用して画一的に，主観的判断の分析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，いくつかの注意が必要だが，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。

統計的分析手法は分野の境界ではなく，研究目的による違いで弁別されるものであるはずで，実際，自然言語処理技術など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は必須の技法と位置づけられるようになってきている。しかし，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さらに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く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ている。

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，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。ということであれば，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が，そうすると，統計手法の種類が限定的に一般化しているとはまだ言いきれ

ないことになり，統計手法の利用上の問題点は，ごく一部のものであるということも考えられる。ただし，手法の使用頻度に関係なく，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研究課題であることから，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は違いないだろう。

また，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。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という問題も見られる。やはり，統計手法の使用理由と，追認のための手順説明は，テクニカルな説明を目的とする論文にだけ備えるのではなく，図のキャプションのように，全ての研究に明示されるように気をつけていくことが，結局は，全体の発展に寄与することになると考え，研究者１人１人が散逸させておく責任を追認可能性にて補うように気をつけることを一般化していくべきだろう。

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。さらに言えば，それは，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ことも一因であるのかもしれない。両方とも推測ではあるが，全体的な流れからすると，それほど外れた話でもないだろう。目的が明確だが，統計初心者であるとか，工学的に結果が重要である場合，言語データに統計を用いて概算するのは，現在のコンピュータの上で誰でもが追認できるわかりやすい手法を利用したい。パーソナルコンピュータの性能の向上と，個人ベースの個別コーパス研究が盛んになったことからすれば，自然な流れである。

しかしながら，これは，過去の過ちを再び繰り返しているという問題でもある。丸山・田野村（2007）のコーパス第１期から第３期にわたる日本語テキスト分析紆余曲折を考えれば，研究のために個人が利用する統計手法自体に対し

ての検証法もモデル化されていることが望ましい。そう考えれば，コーパスの変遷について概観する論文は多いが，統計手法に関して，歴史的変遷，または，分野間の相違という観点で，追認や比較を行った検証研究が少ないことは問題である。

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が，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，研究に用いる統計手法が多様化する昨今，できれば，多様な統計手法の交差部分がどこであるか，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整理することは重要である。今後，益々，統計手法の工夫や手法の開示を行う環境づくりが利用規定として求められるようになるだろう。

日本語の計量的研究においては，統計を用いたら，同時に，その手法を検討していくという，統計手法をスケールとして共有する姿勢やその規則化が，現状の課題として考えられるべきではないだろうか。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だろう。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。

本研究は，2008年度統計数理研究所共同研究レポート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』東京都書.
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.
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.
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.
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.
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.
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.
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.
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.