学位論文要約

題目：テキストにおける語彙的結束性の計量的研究
氏名：山崎　誠

要約：
本研究は、日本語のテキストを対象として語彙的結束性（lexical cohesion）が語彙の量的な側面にどのように現れるかを明らかにすることを目的とする。具体的なリサーチ・クエスチョンは以下の 3 点である。
１．語彙的結束性はテキスト全体の語彙の計量的特性とどのような関係にあるか。
２．語彙的結束性はテキスト中の語彙の分布にどのような形で現れるか。
３．語彙的結束性はテキストの構造にどのように関係しているか。

第 1 章では、概念規定、先行研究、研究手法、データについて概括した。
まず、概念規定として、Halliday & Hasan（1976）で提唱された結束性（cohesion）の概念の一部である語彙的結束性について検討し、テキストをまとめあげる意味で重要なもう一つの概念である一貫性（coherence）との関係を位置付けた。
先行研究としては日本語学における文献を中心に研究の進展を記述した。従来、日本語学・国語学の中で語彙的結束性と同等あるいは類似の概念は「反復」という名称で扱われることが多く、主としてテキストの構造分析の方法論として 1980 年代ごろから登場している。それと同時に言語教育、とくに英語教育における作文の評価指標として語彙的結束性が使われること、および、自然言語処理の分野では語彙的結束性を利用したテキスト分析が行われていることを概観した。
本研究では、語彙的結束性がテキストにおいて現れる際の計量的特性に着目し、マクロな観点からテキストの一般的特性あるいは個別のテキストの特徴を見いだすことを主眼とする。本研究で扱う情報は出現する語の数であったり、テキスト中に出現する同一語の間の距離であったり、それ自体としては言語の本来的機能であるコミュニケーションには係わらない情報であるが、それらの情報を利用してテキストの持つ普遍的な性質や個別的な性質の解明につなげることを目標とした。
本研究で主として使用するデータは、国立国語研究所を中心として開発された『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese、略称 BCCWJ）である。このコーパスを利用する理由は、データの規模が大きい（短単位で約 1 億語）ため、分析に必要な事例を収集しやすいこと、多様なレジスター（言語変種）を含むため、そのレジスター間の比較ができること、データが公開されており、第三者による検証が可能なことの 3 点である。

第 2 章ではテキスト全体の計量的特徴が語彙的結束性にどのように係わるかを示した。
第 1 節では、従来語彙の計量的指標としてとらえられてきた異なり語数の延べ語数に対する比、すなわちテキストにおける 1 語当たりの平均使用度数（n/k 値）からそのテキストの持つ特徴（特に文章論的な観点から見た特徴）をとらえた。その結果、平均使用度数の高いテキスト及び低いテキストにはそれぞれ文体的な特徴があることを指摘した。

第 2 節では、共起語集合（キーとなる語の前あるいは後ろの特定の位置に出現する語の集合）という考えを用いて、BCCWJ においてコロケーションが現れる様子を計量的な指標の観察から記述した。

第 3 節では、語彙の豊かさを表す指標として知られている TTR（Type/Token Ratio）の値がテキスト中で使用されているどの品詞の影響を受けているか計量的に調査した。語数をほぼ一定にしたデータをコーパスから抽出し、当該テキスト全体の TTR と各品詞の TTR との相関を調べたところ、名詞類との相関が一番高く、動詞類、形容詞類がそれに次ぐことが分かった。また、名詞類の中でも普通名詞との相関が高く、TTR の値は普通名詞の使用に大きく左右されることが確認された。

第 3 章ではテキストにおける語の分布から語彙的結束性がどのように働いているか、その一面を明らかにした。

第 1 節では、テキストにおける見出し語の出現間隔の分布とレジスターとの関係を考察した。その結果、高頻度語（主に機能語）の出現間隔は、BCCWJ のレジスターによって違いがあることが分かった。また、テキストにおける出現間隔の平均は、レジスターによる違いは見られないが、出現間隔の総個数がレジスターにより違いがあった。

第 2 節では、多義語を構成する個々の意味が、テキストにおいてどのような出現状況を示すか、使用頻度の高い名詞、動詞、形容詞の多義語 12 語を選び、テキスト中での意味の分布を調べたものである。結論として次の 3 点の結果を得た。

１．多義語がテキスト中で 2 回以上使われる際、同じ語義で使われることが多い。ただし、例外もあり、必ずしも強い制約とは言えない。

２．同じ語義で使われる多義語の間の出現間隔は異なる語義で使われる同一の多義語の出現間隔よりも短い。

３．多義語のうちのある語義に対する類義・対義関係を作る語のうち出現間隔が短いものが認められた。これらの現象は語彙的結束性がひとまとまりのテキストに対して働いていることの現れであろうと推測される.

第 3 節では、テキストにおいて多義語の語義がどのような出現傾向を示すかを『現代日本語書き言葉均衡コーパス』をデータとして調査したものである。

第 3 節では、テキストに出現する多義的な名詞の意味が特定の 1 つの意味で用いられやすいことを観察した。普通名詞を観察した結果、テキスト中では約 7 割～8 割の多義語が特定の意味でのみ使用されていることが分かった。その例外となっているのは、「事（こと）、物（もの）、訳（わけ）、為（ため）、所（ところ）、筈（はず）」など、意味が形式化して機能語に近い用法を持つ語が多く、ほとんどが文法的な意味での使用に関わるものであった。このことを通じて、文法的な意味は語彙的結束性に関与する度合いが低いことを指摘した。

第 4 章では、語彙的結束性がテキストの構造にどのように関係するのかを明らかにした。

第 1 節では、単純な指標である共起語率を用いてテキストの語彙的結束性の度合いを観察した。その結果、法律・白書・国会会議録のように語彙的結束性の高いテキストと、新聞・ベストセラー・雑誌のように語彙的結束性の低いテキストがあることが分かった。NDC（日本十進分類法）別に観察したデータでは、文学の語彙的結束性が低いという結果になった。これは文学に会話文

が多く、その会話が 1 段落と認定されているというデータの特徴の現れである。また、テキスト中の共起語率の推移をみることにより文章のセグメンテーションへの応用が考えられることを示した。

第 2 節では段落間の非対称的類似度を利用して、テキストの語彙的結束性のようすを概観した。今回扱ったデータは白書のサンプル 1 つのみであったが、すべての段落間の組み合わせを観察することにより、どの段落とどの段落とが関係が深いのかが分かり、テキスト全体の語彙的結束性の一端を伺うことができた。また、隣接した段落以外にも語彙的結束性の高い段落があり、それらの関係を利用したテキストの構成の分析への発展の可能性を示唆した。

本節で利用した「無性格語」のリストは雑誌九十種調査の結果から作られたもので、異なるレジスターの分析に耐えるかどうかは検証が必要である。例えばリストには固有名詞「日本」が含まれているが、白書の分析には「日本」は重要な話題として必要な語であり、必ずしも無性格とは言えないだろう。また、今回使用したサンプルについては無性格語を排除しなくてもほとんど同じ結果であったが、どのような場合にこのリストが有効かは確認が必要である。

第 3 節では、語彙的結束性の典型的な現象である、同語の繰り返しに基づく、用語類似語を利用して、話題の展開を測る尺度を提案し、テキスト中の意味段落を切り出す試みを行った。提案した尺度は一定の精度（閾値 0.2 程度）で有効な意味段落の抽出に成功した。しかし、この方法には次のような問題がある。算出の方法として意味情報のほとんどを捨象しているため、話題の展開を測る尺度とは言うものの、話題が新しいものに変わったかどうかしか示すことができない。また、文章論で議論される、「順説、逆説、累加、対比」などの関係は分析できない。これは客観的な測定のため、意味情報を積極的に使用しなかったことの代償である。

第 5 章では、本研究全体のまとめと課題について記述した。

総括として、以上の考察を通じて、日本語のテキストにおいて語彙的結束性という現象を定量的に捉えることのできるいくつかの側面を具体的に明らかにした。また、語彙的結束性は、テキストにおける普遍的な性質として現れるだけでなく、個々のテキストにおいてはそのテキストの個性を表す属性としても現れることも示した。

本研究の課題は大きく 2 つに分けられる。1 つは手法について、もう 1 つはデータについてである。

本研究で用いた手法は平均使用度数や TTR、語の出現間隔、語彙の類似度など、計量言語学・コーパス言語学における基本的な尺度であった。自然言語処理では、特徴語の分析において tf-idf の利用がさかんである。他分野で開発された尺度の活用も盛んである。言語学でも、生物学的多様性を表す Simpson の D が言語分析にも使われるようになったり、文献学で論文の評価を表すために開発された h-index（h 指数）が語彙の計量的分析にも応用されている。このような新たな尺度は新たな観点からの分析につながり、今後それらの手法を取り入れた分析が急務となる。

データについての問題点は、本研究で利用したデータは基本的にまとまったテキストの一部であって、完結したテキストではないという点である。一貫性と語彙的結束性の関係を把握するためには最初から最後までの完結したテキストが必要である。残念ながら BCCWJ には著作権の関係もあり、そのようなテキストはほとんど含まれていない。

最後に質的研究と量的研究の相対化・融合について述べる。本研究が主眼を置いたのはもっぱら量的な研究の方であった。それを質的な研究に転換していく方法の確立が必要である。また、

人文系の研究にありがちなことだが、内省や観察により得られた結果や事実が個々に独立しており、言語研究の体系の中での位置付けが不明確なものが多い。それらの事実を体系的に整理し、研究の方向性を明らかにした上で研究を行う必要があるだろう。

参照文献

Halliday, M.A.K. and Hasan, R.（1976）Cohesion in English. Longman.（邦訳『テクストはどのように構成されるか』、大修館書店、1997 年）

以上