# 〈共同研究プロジェクト紹介〉萌芽・発掘型 : テキストにおける語彙の分布と文章構造 テキストにおける語彙の分布と文章構造


| | |
|---|---|
| 著者 | 山崎 誠 |
| 雑誌名 | 国語研プロジェクトレビュー |
| 巻 | 4 |
| 号 | 1 |
| ページ | 54-60 |
| 発行年 | 2013-06 |
| URL | http://doi.org/10.15084/00000731 |

国語研プロジェクトレビュー
NINJAL Project Review　Vol.4 No.1 pp.54–60（June 2013）



# テキストにおける語彙の分布と文章構造

Distribution of Vocabulary and Sentence Structures in Texts

**山崎 誠**（YAMAZAKI Makoto）

## 1. はじめに

これまでの語彙の計量的研究はひとまとまりのテキストに出現する語彙全体を扱うことが多く，テキストの内部構造についてはほとんど検討されてこなかった。山崎（1983）では文章における語彙の分布がその文章の構造に関係していることが指摘されているが，その後の日本では同種の研究は行われず，Youmans（1991），Hoey（1991），Matthiessen（2002）など海外における研究の進展が目立っている。

本プロジェクトでは，これまで十分でなかった，個々のテキスト内で語彙がどのように使用され，どう推移していくかという立場からの研究を中心に行った。分析の観点としては，内容語，機能語，及び，文章構成の 3 つである。以下それぞれについて説明する。

## 2. 内容語について

### 2.1 語彙的結束性の測定

テキストにおける内容語の分布は Halliday & Hasan（1976）で提唱された語彙的結束性の概念を利用する。語彙的結束性とは，文中の語とテキスト内の他の文中の語との意味的関係であり，両者の間にテキストを理解する上で一貫性のある解釈が成り立つものである。例えば，同一語の繰り返しや類義語，上位語の使用などが語彙的結束性の例である。

山崎（2012b）では，テキストを物理的段落に区切り，それらの間の類似度を測ることによって，語彙的結束性の有り様を観察することを試みた。

表 1 は『現代日本語書き言葉均衡コーパス』のうち，白書のサンプル（サンプル ID：OW6X_00000（平成 16 年度文部科学白書））をもとに，すべての段落間の類似度を示したものである[1]。類似度は水谷（1980）の D（非対称類似度）を用いた。計測は短単位であり，品詞が空白，補助記号，助詞・助動詞であるものを除外した。これらは語彙的結束性を表しているわけではないからである。同様に，語彙的結束性への貢献が低い語群についても除外し

[1] 縦の系列の段落が横の系列の段落に対してとる共起語率の表。例えば，P3 の P2 に対する共起語率は 0.5686。この値は P2 への P3 からの共起語率と解することもできる。太字は当該段落から他の段落への共起語率のうちもっとも値が高いもの。P4 の段落を例にとると，P4 の横の列（0.5, 0.3, 0.45, 0.25, 0.25, 0.25, 0.4）の中でいちばん高い値の 0.5 になる。下線は他の段落から当該段落への共起語のうちもっとも値が高いもの。P5 の段落を例にとると，P5 の縦の列（0.2821, 0.25, 0.3333, 0.25, 0.2593, 0.1786, 0.2105）の中でいちばん高い値の 0.3333 になる。

た。この語群の候補として田中（1973）の「無性格語」を利用した。無性格語とは，田中によれば「これらの単語は，どんな文章にも現われるようなものであって，ある特定な文章や文献の性格とか特徴とかを反映することは，ほとんどない。いわば無性格な語群であろう。」（田中 1973：157）とされるものである。田中（1973）では 108 語が無性格語としてリストアップされている。本稿ではこの無性格語を短単位での実現形に合わせて適宜修正して用いた。また，無性格語の趣旨を汲み，リストに上がっていない数詞についても無性格語として処理した。

**表 1　段落間の類似度**

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | 平均 |
|---|---|---|---|---|---|---|---|---|---|
| P1 | | **0.5128** | 0.4103 | 0.4359 | 0.2821 | 0.4615 | 0.2564 | 0.3077 | 0.3810 |
| P2 | 0.3906 | | **0.5156** | 0.1719 | 0.25 | 0.3125 | 0.0781 | 0.2344 | 0.2790 |
| P3 | 0.4118 | **0.5686** | | 0.3529 | 0.3333 | 0.2549 | 0.1765 | 0.3333 | 0.3473 |
| P4 | **0.5** | 0.3 | 0.45 | | 0.25 | 0.25 | 0.25 | 0.4 | 0.3429 |
| P5 | 0.12 | **0.18** | 0.16 | 0.08 | | 0.12 | 0.04 | 0.12 | 0.1171 |
| P6 | **0.4815** | 0.3333 | 0.2963 | 0.1852 | 0.2593 | | 0.2963 | 0.3333 | 0.3122 |
| P7 | 0.2857 | 0.1429 | 0.25 | 0.25 | 0.1786 | **0.3214** | | 0.2143 | 0.2347 |
| P8 | 0.2368 | 0.2895 | **0.3158** | 0.1579 | 0.2105 | 0.2632 | 0.1053 | | 0.2256 |
| 平均 | 0.3466 | 0.3324 | 0.3426 | 0.2334 | 0.2520 | 0.2834 | 0.1718 | 0.2776 | 0.280 |

表 1 によると，ある段落 a から他の段落 b への類似度において，対象とする相手方の段落 b との類似度が最も高い段落（表の太字のセル）がほとんど第 1 段落（P1）～第 3 段落（P3）に集中しており，このサンプルは前方の段落に依存する傾向があることが見て取れる。

また，本研究の目的の一つとして，テキストにおける語彙の分布の状況を視角化して示し，それをもとに結束性等の性質を分析することがある。図 1 は，同上のサンプルを構成する八

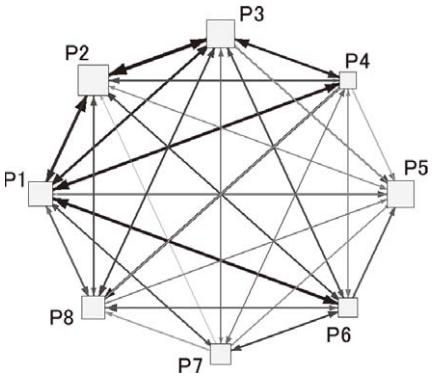


**図 1　段落間の類似度に基づくグラフ**

つの段落について，類似度の多寡を矢印の太さで表したものである。この例では，すべての段落が線で結ばれることから結束性の高いサンプルであると言える。このように，類似度の値及びグラフの次数などを使ってテキストの結束性を表すことができる。

## 2.2　語彙的結束性と文章構成

山崎（2012a）では，白書のサンプル OW1X_00000（昭和 54 年版経済白書）を利用して，隣接する段落の類似度を測定し，直前の段落への類似度よりも直後の段落への類似度が上回っている場合，その段落が意味的な切れ目である可能性があること示した（図 2 参照）。

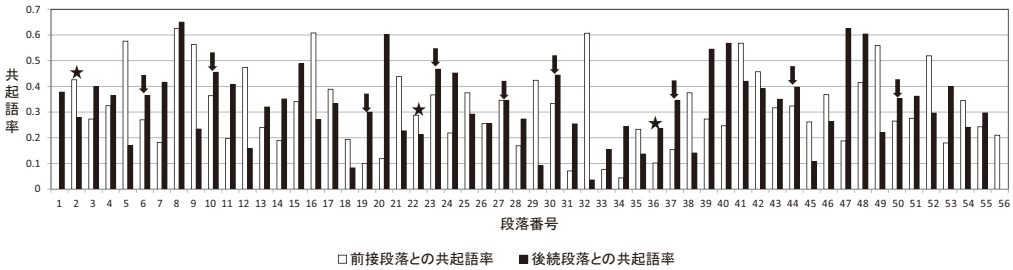


**図 2　文章中の共起語率の推移**

図 2 で★を付けた 3 箇所は大きな節が開始する箇所，下向きの矢印を付した 9 箇所はその節の中で小見出しが立っている箇所である。矢印の部分における後続段落との共起語率（右側の黒い棒）と前接段落との共起語率（左側の白い棒）とを比べてみると，9 箇所のうち 8 箇所が後続段落との共起語率が前接段落との共起語率を上回っている（残りの 1 箇所は同じ値）。このことは，新規の内容になった最初の段落は，新しい話題を展開させるためその次の段落との結束性が高くなっていると言えるのではないだろうか。

逆に，矢印の直前の段落は，あるまとまりの最後の段落を意味する。この部分の後続段落と前接段落の共起語率はどうなっているかというと，9 箇所中 6 箇所で前接段落との共起語率の値のほうが高い。これは一つの例にすぎないが，このような文章中での共起語率の推移を利用して段落のまとまりを自動的に推測することに応用できる可能性がある。

## 2.3　多義語の意味の分布

山崎（2010）では語彙的結束性の特殊な例として多義語を取り上げ，それらがテキストの中の一定のまとまりの範囲では同じ意味で使われる傾向があることを示した。

『現代日本語書き言葉均衡コーパス』の図書館サブコーパス（可変長ファイル）を対象に，「甘い・高い・起きる・乗る（載る）・呼ぶ・式・電話」の 7 語及びその類義語・対義語となる「起こる・降りる・招く・掲載・寝る」の 5 語を対象に，各サンプルにおける出現状況を計量的に調査した。結果を図 3 に示す。

語による違いはあるものの，一つのサンプルの中では同じ意味で使われやすい語があるこ

とが確認された。また，多義語である語が同一テキスト中に複数回出現する場合，直近の語が同じ意味であるか，異なる意味であるかを調査したところ，表 2 の結果を得た。これによると，隣接した語（お互いに同じ多義語）の間の距離は，意味が同じである場合のほうが小さいことが分かる[2]。また表には示さないが，類義語・対義語の場合も異なる意味で用いられた多義語よりも総じて出現間隔が小さいことが分かった。

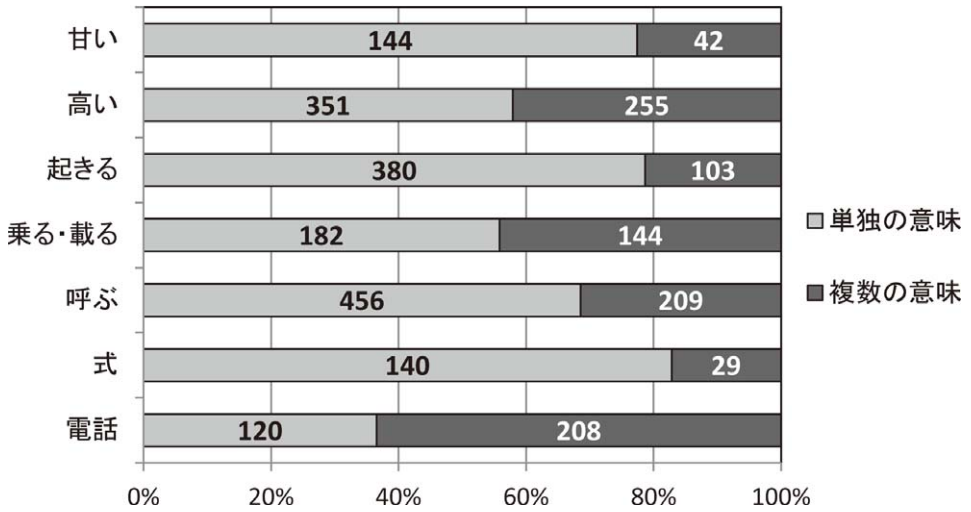


**図 3　同一テキスト中での多義語の意味の現れ方**（表中の数字はサンプル数）

**表 2　多義語の意味の出現間隔**

| | 甘い | 高い | 起きる | 乗る・載る | 呼ぶ | 式 | 電話 |
|---|---|---|---|---|---|---|---|
| 同じ意味 | 562.2 | 633.1 | 745.1 | 652.0 | 715.0 | 338.9 | 349.4 |
| 異なる意味 | 1332.2 | 1033.4 | 1246.7 | 1248.2 | 1053.0 | 1081.5 | 447.9 |

このような現象は，語彙的結束性を保つため，意味的に関連のある語が同一の文脈では積極的に繰り返し用いられることの現れと解釈することができる。

なお，この研究に類似した先行研究として，自然言語処理の分野における語義曖昧性解消（word sense disambiguation）という課題において，“one sense per discourse”という作業仮説が Gale ら（1992）により提唱されている。これは経験的事実を明らかにしたものであり，言語学的な裏付けがなされていない。本研究では日本語でも同様の現象が観察されることを指摘するとともにこれを語彙的結束性の現れとして位置付け，言語学的な解釈を与えたところに新規性がある。

[2] 表 2 の数字は，直近の多義語との間の距離を表すが，この数字はデータとして使用した『現代日本語書き言葉均衡コーパス』を「中納言」で検索した結果に含まれる「連番」の値の差で表している。連番は 10 きざみになっているため，隣りあった語の連番の差は 10 となる。したがって，表 2 の数字で表した距離は，2 語の間に含まれる語数を n とすると，10(n＋1)の値で表されるものである。

## 3.　機能語の分布

江田（2012）は，科学書に見られる「ている」の機能を，その長期の進行用法と完了用法に焦点を当てて分析したものである（それぞれ「運動長期」「効力持続」という名称を使う）。

江田は，『CASTEL/J』に所収されている，自然科学入門書 4 冊及び社会科学入門書 4 冊から各 5 万字ずつを選び，計 40 万字のデータを使って分析したものである。

その結果，科学的入門書では「運動長期」と「効力持続」の「ている」は「話題提供」「結論」を表現する機能を持っていることが分かった。具体的には，「運動長期」「効力持続」の「ている」は，ある自然現象あるいは社会現象，研究の結果や研究者の主張を取り上げ，そのことによってその節で何を問題にするかを提示するという話題提供の役割を果たしていた。また，それらは段落末においては，一定の現象や考え方，社会的自然的変化を示すという形で結論を示す役割を担っていた。

江田は，科学的入門書で「運動長期」「効力持続」が話題提供，結論の表示に用いられるのは，「運動長期」が考えや理論あるいは一般的な事象を表すことができるためであること，「効力持続」が先行研究を引用して議論の前提や結論を示していること，「効力持続」の用法では統括主題の存在が，話題と議論，結論を関係させる要素として有効に働いていると考えられることを指摘した。

江田の研究は，これまで文法的な機能だけに着目されてきた形式「ている」の文脈における機能，すなわち，談話構成機能を積極的に指摘したところに新規性がある。

## 4.　学術論文の構造の分析

清水（2011）は，学術論文におけるメタ言語表現の分析である。「国際政治」「英文学研究」「現代経済学の潮流」「考古学研究」「心理臨床学研究」の五つの学会誌から 10 編の論文を選び，そこに現れるメタ言語表現を調査した。

主な結果としては，「話題内容に明示的に言及しているメタ言語表現」は，序論部分で 51.7%，結論部分で 56.9%と多く出現していること，逆に本論部分では，「論文自体への機能を明示しているメタ言語表現」が 26.5%と多かったことを指摘している。また，論文の記述タイプ（論証型論文，検証型論文，複合型論文）の違いよりも学術分野による違いのほうがメタ言語表現の使用傾向に違いが出たとの報告もなされた。

清水の分析は日本語のアカデミックライティングにおいて，留学生が自分の専攻する分野の論文を書く際に適切な情報を提供することに貢献するものである。

●参照文献●

Gale, William A., Kenneth W. Church and David Yarowsky（1992）One sense per discourse. *Proceedings of the workshop on speech and natural language*, 233−237.（held in February 1992, in Harriman, NY.）

江田すみれ（2012）「「ている」の論理的な文章中での使われ方―「効力持続」「長期的な動作継続」を重点にして―」『国立国語研究所論集』2: 19−47.

Halliday, M.A.K. and R. Hasan（1976）*Cohesion in English*. Longman.［M. A. K. ハリデイ，ルカイヤ・ハサ

ン（著），安藤貞雄・多田保行・永田龍男・中川憲・高口圭転（訳）（1997）『テクストはどのように構成されるか』東京：ひつじ書房.］

Hoey, Michael（1991）*Patterns of lexis in text*. Oxford: Oxford University Press.

Matthiessen, Christian M.I.M（2002）Lexicogrammar in discourse development: Logogenetic patterns of wording. In: Guowen Huang and Zongyan Wang（eds.）*Discourse and language functions*, 91−127. Shanghai: Foreign Language Teaching and Research Press.

水谷静夫（1980）「用語類似度による歌謡曲仕訳―『湯の町エレジー』『上海帰りのリル』及びその周辺―」『計量国語学』12（4）: 145−161.

清水まさ子（2011）「学術論文におけるメタ言語表現の機能と使用状況について」『異文化コミュニケーションのための日本語教育 2』（2011 世界日本語教育研究大会大会予稿集）86−87.

田中章夫（1973）「自動抄録処理におけるキー・ワードの性格」国立国語研究所『電子計算機による国語研究 V』141−184. 東京：秀英出版.

山崎誠（1983）「文章の話題の展開を計る尺度―用語類似度 D の 1 利用法―」『計量国語学』13（8）: 346−360.

山崎誠（2010）「テキストにおける多義語の意味実現の傾向」『計量国語学会第 54 回大会予稿集』25−30.

山崎誠（2012a）「共起語率の分布からみるテキストの語彙的特徴」『第 1 回コーパス日本語学ワークショップ予稿集』221−226.

山崎誠（2012b）「段落間の類似度を利用したテクストの結束性の測定」『第 2 回コーパス日本語学ワークショップ予稿集』291−298.

Youmans, G.（1991）A new tool for discourse analysis: The vocabulary-management profile. *Language* 67（4）: 763−789.

**《要旨》** 本稿では，萌芽・発掘型共同研究プロジェクト「テキストにおける語彙の分布と文章構造」の成果の一部をとりあげて報告する。具体的には，段落間の共起語率を利用した語彙的結束性の分析とその可視化，テキスト中の共起語率の変化が文章構造の把握に利用できること，「ている」の持つ談話構成機能の分析，学術論文におけるメタ言語表現の出現傾向について論じた。

**Abstract:** This paper reports the results of the NINJAL collaborative research project “Distribution of vocabulary and sentence structures in texts.” We introduce a quantitative analysis of lexical cohesion using co-ccurrence rate between paragraphs and a visualization of this analysis. Changes in the co-occurrence rate within a text can be used to study the structure of that text. We also discuss the discourse function of “*te-iru*” and the distribution of metalinguistic expressions in academic articles.

**山崎　誠**（やまざき・まこと）

国立国語研究所言語資源研究系准教授。文学修士。国立国語研究所研究員，同室長，同領域長等を経て，2009 年 10 月より現職。

主な著書・論文：『複合辞研究の現在』（共編著, 和泉書院, 2006），「代表性を有する現代日本語書籍コーパスの構築」（『人工知能学会誌』24(5)，2009），『言語研究のための統計入門』（共著，くろしお出版，2010），*A frequency dictionary of Japanese*（共編著，Routledge，2013）.

社会活動：計量国語学会理事，言語処理学会理事.

**萌芽・発掘型共同研究プロジェクト「テキストにおける語彙の分布と文章構造」**
**プロジェクトリーダー**　山崎 誠
（国立国語研究所　言語資源研究系　准教授）

**プロジェクトの概要**

本研究は，『現代日本語書き言葉均衡コーパス』に収録されているひとまとまりの完結したテキスト等を対象にして，語（内容語及び機能語）の出現状況と時間軸に沿って展開される文章の流れとを有機的に関連付けて動的に捉える観点を提案し，計量語彙論に基づく定量的手法と，語の出現状況及び文章構造・文章展開のモデル化とを通して，より実証的な文章論を開拓するものである。本研究では，テキストにおける語彙の量的構造と文章構造及び当該テキストの持つ特性（表現意図，ジャンル，文体等）との相関を調査・分析し，語彙に内包された文章構成機能を明らかにする。