様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月７日現在

機関番号：６２６１８
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５２０４２９
研究課題名（和文）
日本語のコロケーションを記述するための統計指標のコーパスによる検証
研究課題名（英文）Validation of Statistical Scale Describing Collocation in Japanese

研究代表者
山崎　誠（YAMAZAKI MAKOTO）
大学共同利用機関法人人間文化研究機構国立国語研究所・言語資源研究系・准教授
研究者番号：３０１８２４８９

研究成果の概要（和文）：
本研究では，以下の２つについて考察を行った。(1)典型的なコロケーションにおける係り受けの分布の調査。「を＋動詞」の直前にその動詞に係る名詞句が来る確率は約 97%，「名詞＋を」の直後にその名詞句の係る動詞が来る確率は約 90%であった。形容詞・形容動詞と名詞句については，前後 5 語までにおける係り受けが約 98%であった。(2)慣用句を使った文章ジャンルの判別。「手」を含む慣用句 74 項目を指標として「人文科学」「社会科学」「自然科学」という 3 つのジャンルの判別を多変量解析法によって行った結果，5 つの慣用句によって高率でジャンルが判別されることが分かった。

研究成果の概要（英文）：
In this study, we examined the following two points. (1) Research on the distribution of modifying relation in the typical collocation. It was found that the probability of occurring of modifying noun phrase just before “wo+verb” phrase was 97%, while the probability of occurring of modified verb just after “noun+wo” phrase was 90%. As for adjectives and adjective verbs(keiyo doshi), there was 98% of occurrence of a modifying/modified phrase within five words of the modified/modifying phrase. (2)Discrimination of text genre by idioms. Result of multivariable analysis, we found that five idioms including the word “hand” can discriminate three text genres: humanities, social studies and natural sciences.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,700,000	510,000	2,210,000
２００９年度	800,000	240,000	1,040,000
２０１０年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：人文学
科研費の分科・細目：言語学・日本語学
キーワード：コロケーション，コーパス，推移確率，統計指標，共起

１．研究開始当初の背景

英語研究においては，コロケーションに関する統計指標が多数提案され，辞書編纂や英語教育において実用化されている。例えば，「粗頻度 (Raw Frequency)」,「共起頻度比」,「T スコア（T-score)」,「相互情報量（MI Score／Mutual Information Score)」, 対数尤度比（Log-likelihood Score／G Score／G2

Score）等である（石川 2006）。

しかし，日本語では，これらの指標を具体的に適用した例がほとんどない。その理由としては次のような事情が考えられる。

(1)日本語のコーパスが整備されていない。テキストファイル形式のデータは新聞記事等を中心に入手しやすくなってきたが，単語に分割され品詞等のタグが付けられたデータが少ない。

(2)研究者側の事情として，コーパスの使い方が用例の発見及び引用という利用形態にとどまっており，統計的な観点での分析になかなか入っていけない。

近年，この状況は以下のように急速に改善されつつある。

上記(1)の背景には，日本における語彙の統計的調査の歴史が関係していると推測される。日本語の語彙の統計調査は，国立国語研究所の用語用字調査が中心的な存在であったが，そこでは語の単純頻度（使用率）や各ジャンルにおける分布が主な関心事であり，語と語の関係については，ほとんど顧みられてこなかったという事情がある。

しかし，国立国語研究所が平成 18 年度から開始した「現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Japanese)」のプロジェクトによって，現代日本語のコーパスの整備は大幅に進むことは間違いない。このプロジェクトでは，現代日本語の様々なジャンルからバランスのとれたテキストを集め，統一された言語単位に分割し，品詞情報等を付けたデータを公開する。

上記(2)に関しては，コーパスを利用した研究の拡大が挙げられる。2000 年代に入り，現代語を扱った論文でコーパスを使った統計的な手法を用いているものが 3 割程度あり，従来の用例提示を主とする利用の仕方以外の方法も広がってきていることが指摘されている（山崎 2007）。

また，日本語教育では，姫野(2004)のようなコロケーションを収集整理した辞典も刊行され，コロケーションを表す指標を評価する基盤が整ったと言えるだろう。

２．研究の目的

従来もっぱら英語研究で提案されている統計指標を，日本語のコーパスに適用する際に，統計指標の妥当性に影響を与える言語的要因にはどのようなものがあるかを明らかにする。

検証する要因としては，①コーパスの大きさ（語数），②コーパスの種類（ジャンルや文体等），③計る言語単位の長さ，④測定範囲（前後何語までを対象とするか），⑤品詞，⑥コロケーションの型などを考える。これらがどのような条件のもとで測定結果に影響を与えているかを具体的な例によって明らかにする。

例えば，橋本(2007)においては，小説，国語辞書，インターネットを資料として，「顔」に係る形容詞のコロケーションを分析しているが，資料によって現れる形容詞の傾向が異なることが指摘されている。

参考文献

石川慎一郎(2006)「言語コーパスからのコロケーション検出の手法－基礎的統計値についてー」，統計数理研究所共同研究リポート 190「言語コーパス解析における共起語検出のための統計手法の比較研究」，pp.1-14.

橋本和佳(2007)「名詞とそれを修飾する形容詞の関係」，「日本語学」26-12，pp.38-46.

姫野昌子(2004)『日本語表現活用辞典』研究社

山崎誠他(2007)「現代日本語書き言葉均衡コーパスの設計と検索デモンストレーション」『日本語学会 2007 年度秋季大会要旨集』，pp.239-246.

３．研究の方法

(1)係り受けの距離の測定

文における統計量を算出するための基礎データを作成するために『現代日本語書き言葉均衡コーパス』のモニター公開データに含まれる４つのデータ（白書，書籍，Yahoo！知恵袋，国会会議録）及び新聞社説等から「名詞＋助詞＋動詞」の形のコロケーションにおける係り受けの距離を測定し，分布状況を記述する。

(2)文章のジャンルと相関する特徴的表現の抽出

複合動詞や慣用句などの分布を通して文章のジャンルを判別するための指標について（１）のデータ及び学術論文のデータをもとに記述する。

４．研究成果

(1)形容詞連続の名詞修飾

「細く長い道」のような例では，「細く」と「道」とのコロケーションが直接には検出できない。また，「目に余るふるまい」のような慣用句を含んだ表現では，「余る」と「ふるまい」のような例外的なコロケーションが抽出されてしまう。このような，コロケーション測定の目的から外れる事例がどのくらいあり，測定の際にどれくらい影響を与えるかを具体的に検討するため，白書データで実測を行った。結果は形容詞＋形容詞＋名詞の連続は 113 例検出されたが，係り受けが適切でないもの，形容詞+否定辞の「ない」，連用形の「著しく」＋形容詞等，目的に沿わない例を除外すると該当例は 2 例のみであること

が分かった。白書以外のデータでの検証が必要だが，形容詞が 2 つ続いて名詞に係るような例はそれほど多くないことが分かった。

(2)係り受けの位置の調査－動詞の場合

コロケーションがどのくらいの範囲で起きているか,動詞と名詞の関係を取り上げて調査した。利用したデータは,『現代日本語書き言葉均衡コーパス』の「コアデータ」(書籍,白書,新聞,Yahoo!知恵袋)約 80 万語を対象に,「を+動詞」に係る名詞(約 14000 例),「名詞+を」の係る動詞(約 20000 例),「に+動詞」に係る名詞(約 15000 例)について,係りの位置が上記中心語の前後どのくらいの範囲に出現しているかを形態素単位で調査した。その結果,「を＋動詞」は,挿入句や括弧などの記号を除くと殆どの場合,直前に係り元の名詞が出現しているが,「名詞＋を」では,直後に副詞や動詞の取る別の格(「に格」など)に関係する名詞等が現れる場合が約 10%近くあること分かった。このことは名詞を中心としたコロケーションを測定する場合に一定の誤差として影響が出ることを意味している。また,「に＋動詞」では,「について」「において」「における」などの複合辞が多数含まれること及び「すぐに」「実際に」などの副詞を構成する「に」が分離されていることなど,形態素解析に起因する問題を除けば,係り元の名詞はほぼ直前に出現していることが確認された。

(3)係り受けの位置の調査－形容詞・形容動詞の場合

コロケーションの範囲を調べるために形容詞，形容動詞の係り受けを調査した。形容詞・形容動詞が名詞に係る場合の距離，及び，形容詞・形容動詞へ名詞が係る場合の両方を調査した。

利用したデータは，『現代日本語書き言葉均衡コーパス』の「コアデータ」（書籍，白書，新聞，Yahoo!知恵袋）約 80 万語である。データ中には形容詞連体形が 3403 例あったが，そのうち形容詞が直接名詞に係る例が 1583 例，（前文脈で）名詞が形容詞に係る例が 838 例であった。形容詞が名詞に係る例のうち，形容詞の直後に名詞が現れる割合は 93.9%，直後の 5 語までては 99.4%であった。逆に名詞が形容詞に係る場合は，直前に現れる割合は 6.8%，2 語前までで 90.8%，5 語前までで 97.6%であった。同様に形容動詞では，直後の 5 語までで 99.9%，直前の 5 語までで 96.2%が出現していることが分かった。このことから，形容詞・形容動詞が名詞に係る場合は，直後の 5 語まででほぼカバーできることが分かった。また，ここでは，連体節中のという限定付きであるが，名詞が形容詞・形容動詞に係る場合もほぼ 5 語以内で約 96%がカバーできることが分かった。

(4)文章のジャンルと相関する特徴的表現の抽出

慣用句を指標として文章ジャンルの判別の可能性を探索した。「手」を含む動詞慣用句，形容詞慣用句 74 項目を指標として用い，『現代日本語書き言葉均衡コーパス』の書籍文章資料を対象として「人文科学系」「社会科学系」「自然科学系」という３つのジャンルの判別を多変量解析法によって行った。その結果，５つの慣用句によって高率でジャンルが判別されることが分かった。

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 村田年・山崎誠,「「手」の慣用句を指標とした文章ジャンルの判別－現代日本語書き言葉均衡コーパスを用いて－」,「日本語と日本語教育」(慶應義塾大学日本語・日本文化教育センター) 39, 2011, pp. 75-88.

② 村田年,「文章と文型８－論文要旨における文型の使用頻度調査－」,「日本語と日本語教育」(慶應義塾大学日本語・日本文化教育センター) 37, 査読有, 2009, pp. 61-92.

〔学会発表〕（計３件）

① 山崎誠,「多義語を構成する意味の使用傾向-品詞と活用形による違い-」，言語処理学会第 17 回年次大会，2011 年 3 月 9 日，豊橋技術科学大学

② 村田年,「『手』の慣用句を指標とした文章の所属ジャンル判別の可能性－現代日本語書き言葉均衡コーパスを用いて－」，テキストにおける語彙の分布と文章構造研究発表会，2011 年 3 月 6 日，国立国語研究所

③ 山崎誠,「テキストにおける語の平均使用度数と文体差」，大規模データ・リンケージ，データマイニングと統計手法，2009 年 10 月 9 日，国立情報学研究所

６．研究組織

(1)研究代表者

山崎　誠（YAMAZAKI MAKOTO）
国立国語研究所・言語資源研究系・准教授
研究者番号：30182489

(2)研究分担者

村田　年（MURATA MINORI）
慶應大学・日本語・日本文化教育センター・教授

研究者番号：50225372

(3) 研究分担者
馬場　康維（BABA YASUMASA）
統計数理研究所・名誉教授
研究者番号：90000215

(4) 研究分担者
橋本　和佳（HASHIMOTO WAKA）
同志社大学・文学部・講師
研究者番号：40511053
（H20→H21）