現代日本語におけるコロケーション:検出と分析

STRAFELLA Elga Laura（奈良先端科学技術大学院大学）[†1]
林部 祐太（奈良先端科学技術大学院大学）[†2]
松本 裕治（奈良先端科学技術大学院大学）[†3]

Detection and Analysis of Collocations in Contemporary Japanese

Elga Laura Strafella (Nara Institute of Science and Technology)
Yuta Hayashibe (Nara Institute of Science and Technology)
Yuji Matsumoto (Nara Institute of Science and Technology)

1 はじめに

本研究ではコーパスから現代日本語におけるコロケーションを検出し、それらの構文パターンと意味を分析することを目標としている。本稿では、まずコロケーションの定義を行い、「制限コロケーション」・「比喩的イディオム」・「真性イディオム」の違いを簡単に検討する（2 章）。次に、コロケーション検出の研究でよく用いられる代表的な指標を取り上げ、その特徴を述べる（3 章）。そして、それらの違いを定量的に区別するために我々が行ったアンケート調査について説明し（4 章）、アンケート結果を分析する（5 章）。最後に、まとめと今後の方針について述べる（6 章）。

2 コロケーションとは

「コロケーション」とは「連結語句」「連語」「語の配列」なども言われるが、要するに「ある単語と単語のよく使われる組み合わせ」のことである。例えば、“辞書” という単語では、「辞書を引く」「辞書で調べる」「分厚い辞書」とは言えるが,「辞書を読む」「太い辞書」とは通常言わない。このような自然な単語の組み合わせは、そのまま覚える必要があり、日本語だけでなく、全ての自然言語における共起現象の一つである。外国語学習者はできるだけ “正しく、自然な” 言語を話すために, その言語の “正しく、自然な” 組み合わせを使わなければならない。コロケーションは全体的な意味が各単語の意味の組み合わせと比喩的に違ってくることが多く、母国語を参考したり、文字通りに翻訳したりすると “奇妙な表現” になる可能性が高い。

本研究ではコロケーションを単語の結び付きの強さに応じて「制限コロケーション」・「比喩的イディオム」・「真性イディオム」の 3 段階で区別する.

2.1 制限コロケーション

制限コロケーションとは、「傘をさす」のように単語の自然な組み合わせのことである。「傘を開く」や「傘を開ける」は「傘をさす」と意味的には、ほとんど変わらないが、母語話者が聞くと、不自然な表現である。ただし、イディオムとは異なり、その組み合わせから特別な意味は生じない。

[†1]elga-s@is.naist.jp
[†2]yuta-h@is.naist.jp
[†3]matsu@is.naist.jp

表 1: 統計指標の例

フレーズ	頻度	相互情報量	T スコア	ダイス係数
腹が立つ	285509	4.71668	529.552	0.170419
歯が立つ	54035	3.13341	222.328	0.0332873
足が出る	19443	-1.44844	-454.072	0.00172974
目が出る	14238	-2.69391	-1645.38	0.000943862
腕が立つ	4845	0.216424	13.5477	0.00239027
首が据わる	4167	4.3758	63.7468	0.00351143
目が据わる	3727	2.57955	56.4299	0.000587353
背が立つ	729	-1.1192	-55.6667	0.000458049
胴が据わる	8	1.55885	2.47487	0.000161713

2.2 比喩的イディオムと真性イディオム

二語以上の単語が固く結びつき、それぞれの単語とは異なる意味を持つものをイディオムとよぶ。ただし、ある程度元々の単語から全体の意味が推測できる場合とそうでない場合がある。

個々の語の意味から構成的に理解できないものを「真性イディオム」と呼ぶ。たとえば、「腹が立つ」や「腹を立てる」は、両方とも “怒る” という意味を指し、それぞれの単語の意味を知っていても、全体の意味を推測できない。このように単語が固く結びついた表現は, 日本語学習者にとって一つずつ覚えるしかない。

一方ある程度元々の単語から全体の意味が推測できる場合、「比喩的イディオム」と呼ぶ。たとえば、「足を伸ばす」とは、2 つの意味を持ち、一つは「足をうーんと伸ばす」という物理的な動作を指し、もう一つは個々の単語の意味から少し離れた「遠出する」という比喩的な意味である。このような結合句は「制限コロケーション」と「真性イディオム」との中間の段階にある。

3 よく用いられる統計指標

最近の言語コーパス研究では、複数の統計指標に基づく共起語検出に対応しているものが多い。しかし、どの統計値を用いれば良いのかは, はっきりとした結論は得られていない。本稿では、一般にコロケーション研究に広く用いられる, 単純頻度・ダイス係数・MI スコア・T スコアという 4 種類の指標を紹介する。アンケート（後述）で使用したフレーズを用いて、表 1 に各指標値の例を示す。

以下, 中心語 A の頻度を f_A, 共起語 B の頻度を f_B, コーパスでの総語数を W と表記する.

3.1 単純頻度

コロケーションを抽出する上で、もっとも単純な指標は、単純頻度である。しかし、より詳細な分析を行おうとすると、単純な頻度だけで判断するのは危険である。もともと頻度の低い語であれば、その語との共起頻度も自然に低くなり, 逆に、もともと多く出現する語であれば、その語との共起頻度が高くなる.

そのため、中心語と共起語の結合形の単純頻度 (共起頻度) だけでなく、 複数の指標を組み合わせて、それぞれが示す共起度を比べる必要がある。なお, 以下共起頻度を f_{AB} と表記

する.

3.2 ダイス係数

ダイス係数は、中心語頻度と共起語頻度の関係だけで 2 語のコロケーション強度を計測する尺度である。共起頻度を中心語頻度と共起語頻度の和で割って 2 倍した値である。式は次のようになる。

$$D = 2 \times \frac{f_{AB}}{f_A + f_B}$$

3.3 相互情報量

相互情報量は、ある語と共起語の統計的な独立性を示す指標である。ただし、頻度が低い語について敏感に反応し過ぎるという欠点をもっている。式は次のようになる。

$$I = \log_2 \frac{f_{AB} \times W}{f_A \times f_B}$$

3.4 T スコア

T スコアは、2 つの語の共起関係の統計的有意性を図り、共起の程度が偶然による確率を超えていると、どのくらいの確かさで言えるかを示す指標である。式は次のようになる。

$$T = \frac{\left(f_{AB} - \frac{f_A \times f_B}{W}\right)}{\sqrt{f_{AB}}}$$

4 アンケート調査

コロケーション研究においてよく使用される何種類の指標のうち、どの指標が有効であるかを調べるためにアンケート調査を行った。

4.1 アンケートの概要

アンケートには「身体名詞-[がをに]-動詞」の組のうち、比較的共起頻度の高いフレーズ 78 個を用いた。頻度は「日本語係り受けコーパス」(JDC) より計算した。JDC とは、約 1 億ウェブページからなる日本語ウェブコーパス 2010(NWC2010) より日本語係り受け解析システム CaboCha を用いて、助詞を介した語と語の係り受けを抽出したコーパスである。

アンケートは、当大学の日本語母語話者 (21 人）と留学生 (日本語能力試験の 1・2 級の 15 人) を対象に実施した。

4.2 母語話者に対する質問

母語話者には、各フレーズの名詞と動詞の結びつきの強さを、自由結合（コロケーションではないフレーズ)・コロケーション・比喩的イディオム・真性イディオムの 4 段階のうち、ど

表 2:「足が出る」に対する母語話者の回答の例

語義	DIC	POW	FREQ	ユーザ番号
0	4	2	2	8
1	2	4	2	8
2	1	4	1	8

表 3: 留学生に対するアンケートから得られたデータの例

フレーズ	意味 1	意味 2	意味 3	名詞の意味	動詞の意味	推測した意味	ユーザ番号
腹が立つ	to get angry			stomach	to stand		2
歯が立たない	be not able to do	too difficult to realize or to do	too hard to chew	teeth	to stand (negative)		60
背が立つ				back	to stand	to get higher	2

の段階に有るのかを回答してもらった。そして、与えられたフレーズの意味を思い付いた順に記入してもらった。

「足が出る」という 2 つの意味を持つフレーズについて母語話者から得られた回答の一部を表 2 に示す。各項目は、それぞれ次の質問に対する回答である。

- **DIC**: 辞書を使わなくても, この意味が出てきましたか?
  1:辞書を引いて初めて意味を知った、2:辞書を引くと意味を思い出した、3:辞書を引かなくても分かった、4:辞書には載っていない意味だった
- **POW**: 名詞と動詞の結びつきの強さはどのくらいですか?どうしても分からない場合は「不明」を選択して下さい.
  2:自由結合、3:コロケーション、4:比喩的イディオム、5:真性イディオム、-2:不明
- **FREQ**: このフレーズが 100 回使われたとして, この意味を表現するために, 使われるのは概ね何%くらいだと思いますか?
  (注 1) このフレーズがよく使われるかどうかという質問ではありません.
  (注 2) このフレーズには意味が 1 つしか無いならば,80%〜を選択してください.
  1:〜29%、2:30%〜、3:50%〜、4:80%〜、

4.3 留学生に対する質問

留学生には、各フレーズの名詞と動詞の意味について答えてもらい、フレーズ全体の意味が分からない場合には、それを推測できるかどうか尋ねた。

表 3 は留学生に対するアンケートから得られたデータの例である。まず、与えられたフレーズに対して最大 3 つまで思いつく意味を列挙してもらい、動詞と名詞の意味も答えてもらった。与えられたフレーズの意味が分からなかった場合、フレーズの意味を推測してもらった。

5 アンケートの結果

本稿では、アンケートに用いた 78 個のフレーズのうち表 1 に示した 9 個のフレーズを使って分析する。

表 4: 複数の意味を持つフレーズに対する母語話者の結び付きの強さの判断の例

フレーズ	意味	真性イディオムと答えた人数	比喩的イディオムと答えた人数
足が出る	予算を超えた支出になる	11	7
	隠しごとが現れる	7	10
目が出る	目玉が飛び出る	3	14
	幸運が巡ってくる	2	13
歯が立たない	固くて噛むことができない	1	7
	自分の力が弱くて対抗や理解ができない	12	10

5.1 一致率と共起頻度

共起頻度の高いフレーズは、母語話者も留学生も回答は似通った。例えば,「腹が立つ」は、母語話者は全員が正確な意味を答え 21 人中 18 人が「真性イディオム」と答えた。一方、共起頻度の低いフレーズは母語話者であってもその意味があまり理解されておらず、コロケーションの強さの判断にばらつきがあった。例えば,「背が立つ」は母語話者は 20 人が辞書を引かないと意味が分からず、8 人が「比喩的イディオム」と答え、9 人が「真性イディオム」と答えた。そのため、回答者の一致率と共起頻度には相関関係があると考えられる。

5.2 複数の意味を持つフレーズ

フレーズには複数の意味を持つものがあった。例えば,「足が出る」には、「布団から足が出る」のように文字通りの意味の他に、「予算を超えた支出になる」や「隠しごとが現れる」という意味がある。その場合は、意味ごとにコロケーションの強さを判断してもらった。その結果の一部を表 4 にまとめた。

表から分かる通り、コロケーションの強さの判断は意味ごとに異なった。しかしながら、3 章で挙げた指標では、意味ごとの分析ができない。

6 まとめ

本稿では、自由結合・コロケーション・イディオムなどの概念を整理して、共起度を測る代表的な指標としてよく使われる単純頻度・ダイス係数・相互情報量・T スコアの 4 つの指標について説明した。

アンケートの結果、共起頻度が極めて高いフレーズは単語間の結び付きの強さの判断の一致率が高かったが、そうでないフレーズは回答者間で判断が割れた。また、複数の意味を持つフレーズは意味ごとに結び付きの強さは異なることがあり、その場合は前述した指標では分析できないことが分かった。

今後は共起に基づく指標だけではなく、フレーズを構成する単語の置き換えに基づく指標を用いることを考えている。コロケーションには、前述したように似たような意味を持つ単語で置き換えると不自然な表現になるという性質がある。そのため、日本語のシソーラス『分類語彙表』を使って、各名詞が共起する動詞の synset を作り、その動詞を同義語と置き換えると、頻度などがどのように変化するか等を探っていきたい。

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」統計数理研究所共同研究レポート 190 巻, pp.225-243

Evert, S. (2004) *The Statistics of Word Co-occurrences: Word Pairs and Collocations*, Ph.D. thesis, University of Stuttgart

金田一秀穂 (2006)『知っておきたい日本語コロケーション辞典：Japanese Collocation Dictionary』学研

庄司香久子 (2010)『日本語言葉のコンビネーション・ハンドブック』（英文版）、講談社

姫野昌子 (2004)『日本語表現活用辞典』、研究社

Seretan, V. (2010) *Syntax-Based Collocation Extraction*. In N. Ide and Véronis J. (eds.) *Text, Speech and Language Technology*, vol.44, Springer Dordrecht Heidelberg London New York

国立国語研究所 (2004) 『分類語彙表』、国立国語研究所資料集 14 巻

関連 URL

日本語ウェブコーパス 2010 http://s-yata.jp/corpus/nwc2010/

日本語係り受けコーパス http://hayashibe.jp/jdc/