コロケーションとシンタクス
—形容詞と名詞のコロケーションを対象に—

スルダノヴィッチ・イレーナ（国立国語研究所・リュブリャーナ大学）†

Collocation and Syntax: Adjective and Noun Collocations

Irena Srdanović (University of Ljubljana/ National Institute for Japanese Language and Linguistics)

１．はじめに

コロケーション研究は、その始まりから、統計的な面から見た単位と単位の組み合わせに焦点をあてていたが、近年の研究ではシンタクスおよびシンタクスを超えて考えるまでの方法へのシフトが見られる（Grefenstette 1992、Stefanowitsch&Gries 2003、Tanomura 2010、Seretan 2011）。統計的な面から見たコロケーション研究は、語と語の間のスパンを 3〜5 語に設定し、単語と単語の組み合わせの強さを何種類の統計値によって計算する傾向が良く見られる（Hunston 2002）。このアプローチは、現在幅広く使われているツールにおいても、5 単語以内のコロケーションスパン、または、2 単語に絞ったコロケーション抽出などにおいて影響を与えた。それ以外にも、コンコーダンサーにおける用法･意味･パターンの観察方法が多く用いられてきた。一方、このような研究アプローチにシンタクスの面からコロケーションを観察するアプローチを加えることによって、コロケーションデータの取り出し方を更に精密なものにすることができる。

本研究では、「形容詞（連体形）＋名詞」のコロケーションを対象にして、名詞を修飾する形容詞の組み合わせ以外に、そのコロケーションがどのような構成になるか、およびそれぞれの形容詞の用法にどのような傾向があるかを現代日本語のコーパスを利用しながら検討する。従来の研究では、文中の形容詞の機能について多くの指摘があった（鈴木 1972、西尾 1972、高橋 1998、八亀 2008）。これらの研究では主として、形容詞の機能を、叙述用法、連体用法、連用用法の三つに分けて、どの機能が中心的であるかについて議論してきた。さらに、形容詞 136 語を対象にした形容詞辞書 IPAL の研究を紹介した橋本･青山（1992）および宮島（1993）が挙げられる。その研究では、三つの用法のうち、形容詞によってその用法があるかどうかを明確にし、用法のある語形についても、その用法の量的な面では偏りがあるということが述べられている。さらに、大規模データを基にしたジャンル別の分析がある。例えば、形容詞の用法分布･語義分布についての考察（姜 2012）、連体形でしか、または連用形でしか利用されない形容詞（小川他 2008）、形容詞述語のタイプと述語になりやすい語となりにくい語（前川 2012）などである。

本稿では、文節・文章における連体形の形容詞の機能およびそれに関して形容詞ごとの特集の傾向を調べる。以下のような問題点を対象にして、形容詞と名詞のコロケーションの構成を検討する。

統計的に見たコロケーション		**シンタクスを考慮に入れたコロケーション**
高い　建物	……	高い　建物
高い　国	……	教育水準　が　高い　国
高い　国	……	インフレ率　の　高い　国
高い　人	……	コミュニケーション能力　の　高い　人
[形容詞（連体形）＋　名詞]		*[文節・文章における「形容詞（連体形）＋　名詞」]*

† irena.srdanovic@ff.uni-lj.si

統計的な方法だけで形容詞と名詞のコロケーションを抽出すると、形容詞単独で名詞を修飾するコロケーション（「高い建物」）と、不十分だと考えられるコロケーション（「高い国」「高い人」）が同じように扱われている。その区別ができるようにするため、それぞれの形容詞の傾向を把握した上、「形容詞（連体形）＋名詞」のコロケーションを抽出するためのコーパスクエリシンタクス（CQS）のルールを改良する。

２．「形容詞（連体形）＋名詞」の分析およびそれぞれの形容詞の振る舞い

本章では、高･中・低頻度の形容詞（各 3 語）を選び、それぞれの形容詞と名詞のコロケーションデータのコンコーダンスを JpTenTen コーパス（Pomikálek&Suchomel 2012、スルダノヴィッチ他 2013）からランダムな 100 例を取り出し、形容詞の前後文脈まで含めて分析した。分析対象の形容詞は高頻度の「高い」「多い」「寒い」、中頻度の「青い」「甘い」「親しい」、低頻度の「痛々しい」「甘辛い」「野太い」。「手早い」という低頻度の形容詞は、最初に分析対象にランダムに選んだが、連体形+名詞の用法はないので、その代わりに「野太い」にした。

表 1 は、文節・文章における「形容詞（連体形）＋名詞」の構成タイプを形容詞ごとに示す。構成タイプは以下のように分けられる。

- 形容詞単独で名詞を修飾するもの
  （[Ai+N]、例えば「寒い季節」「高い評価」「甘い考え」など）

- 連体修飾節の述語の機能を持つ形容詞で、節が名詞を修飾するもの
  （「地震危険度の高い地域」「雨の多い国」「砂抜きが甘い店」）
  それらは、[N が Ai]N、[N も Ai]N、[N の Ai]N、N は...Ai]N、[N は Ai]N に分類して分析した。ここは、「の」「が」が最もよく使われている助詞であるが、「は」「も」もたまに見られる。時々、かかわる「名詞+助詞」と形容詞の間に副詞および他の単位が現れる。複数の単位の場合には、[N は...Ai]N のように示した。

- 所有の「の」＋形容詞で名詞を修飾するもの
  （[Poss の+Ai+N]、例えば「タレの甘辛い味」「男性の野太い声」「女の子と男の子との間の親しい関係」）

- 名詞＋形容詞という構造をもった複合形容詞で名詞を修飾するもの
  （[N+Ai]+N、例えば「香り高いコーヒー」「誇(ほこ)り高い N」「テンション高い人」）。このタイプは、複合形容詞ではなく、単純に連体修飾の「が」の省略と考えられる場合もある。

合計で見ると、最も高い頻度で現れている構成が形容詞単独で名詞を修飾するタイプであり、続いて述語の機能を持つ形容詞が多い。さらに、各形容詞の構成を量的に見ると、形容詞の振る舞いに偏りがあることが明らかになった。「多い」という高頻度の形容詞は述語の機能の出現は非常に高く、90%を超えている。「高い」という形容詞は、連体形の形容詞が連体修飾節の述語であるケースが全体のおよそ半数を占めていた（「質の高いサービス」）。または、形容詞の前に名詞が付く言語表現が 5 例あった（「香り高いコーヒー」）。「多い」と「高い」は、述語の機能を持つ形容詞として代表であり、量の意味の領域を表す他の形容詞には同じような傾向があることが考えられるが、それを確認するためにさらに数多くの形容詞の用法を検討する必要がある。一方、「青い」「甘辛い」「野太い」「甘い」「痛々しい」「寒い」は、連体形の形容詞が連体修飾節の述語になるケースがないか少ないが、形

容詞によって 7%から 17%まで所有関係の「の」の用法が多く見られる（「日本の青い空」、「男の野太い声」など）。

表 1　各形容詞によって文節における「連体形の形容詞+名詞」の構成の傾向

構成			高頻度の形容詞の例			中頻度の形容詞の例			低頻度の形容詞の例			合計
			多い	高い	寒い	青い	甘い	親しい	痛々しい	甘辛い	野太い	
形容詞（連体形）+名詞	単独[形＋名]	[Ai+N]	7	39	90	82	86	93	83	62	85	627
	述語の機能を持つ形容詞	[Nが…Ai]N		2								2
		[NがAi]N	62	9	1	1	7		6		1	87
		[NもAi]N	9	1								10
		[NのAi]N	18	39								57
		[Nは…Ai]N		3								3
		[NはAi]N	2		1							3
		合計	91	54	2	1	7	0	6	0	1	162
	所有関係	[Possの+Ai+N]		2	8	17	7	7	11	12	11	75
	複合形容詞	[N+Ai]+N	2	5								7
合計			100	100	100	100	100	100	100	74	97	871

さらに、「高い」を例にして、BCCWJ と JpTenTen からランダムに取り出した 100 例ずつのデータには「高い（連体形）+名詞」の構成における違いを調べた。表 2 に示したように、BCCWJ における単独の[Ai+N]のほうが 1 割で多く、JpTenTen における述語の機能を持つ形容詞のほうが多いという傾向が見られる。述語の機能を持つ形容詞の用法のうち、BCCWJ における[N が Ai]N のほうが JpTenTen より多く、JpTenTen における[N の Ai]N のほうが BCCWJ より若干多いと分かった。BCCWJ のレジスターデータを検索した結果、[Ai+N]・[N が Ai]N・[N の Ai]N は、書籍で最も現れていることが分かった。

表 2 文節における「高い（連体形）+名詞」の構成の傾向

構成		BCCWJ	JpTenTen
単独[形＋名]	[Ai+N]	51	39
述語の機能を持つ形容詞	[N が…Ai]N	1	2
	[N が Ai]N	14	9
	[N も Ai]N		1
	[N の Ai]N	28	39
	[N は...Ai]N	1	3
	[N は Ai]N	1	
	合計	45	54
所有関係	[Poss の+Ai+N]	1	2
複合形容詞	[N+Ai]+N	3	5

３．シンタクスを考慮に入れたコロケーション抽出

現在使われている多くのコーパス検索システムは、統計的な方法に頼り、構文的情報を直接扱うことができない。本研究で利用するスケッチエンジンというツール（Kilgarriff 2004）は、シンタクスを考えたパターンの規則化を採用していることで、いわば第 4 世代のコンコーダンスツールであると言える（McEnery&Hardie 2012）[1]。Gahl（1998）によって提案された「corpus query syntax (コーパス検索シンタクス)」を実装し、主に品詞と正規表現を活用したルールによってコロケーション抽出ができる。日本語のコロケーションが抽出できるために、日本語の文法を考えたルールを品詞、正規表現、活用形などで作成した（Srdanović ら 2008、スルダノヴィッチら 2008、2013）。

「形容詞+名詞」のコロケーションを取り出すために、まず、以下のルール 1 を利用した。このルールは、2 単位を取り出すためのルールであり、2 は、連体形の形容詞（「無い」を除外）を取り出し、1 は、名詞を取り出す(数詞を除外)。

ルール 1

```
*DUAL
=modifier_Ai/modifies_N
    2: [tag="Ai.*" & word!="ない|無い" & infl_form="Attr.*"] [tag="Pref"]? 1:[tag="N.*" &
                                                                    tag!="N.num"]
```

ルール 2 は、2 章に紹介した分析を行った上で、文節における「形容詞（連体形）+名詞」の構成を考慮に入れて改善したルールである。このルールにより、述語の役割を持つ連体形の形容詞を除外することが可能になった。述語の役割の場合には、「の」と「が」がよく使われるので、「の」と「が」が形容詞の前にない例文だけを取り出す。[2]　このルールでは、名詞と所有の「の」が形容詞＋名詞の前に来る場合は（例えば「朝の寒い空気」）、その出現はその「形容詞＋名詞」コロケーションには見られない。それでもなお、所有の「の」がないコロケーションのケースのほうが圧倒的に多いため、高頻度のコロケーション結果には差異がなく、中･低頻度のデータにも差異が少ないという結果を得た。

ルール 2

```
*DUAL
=modifier_Ai/modifies_N
        2: [tag="Ai.*" & word!="ない|無い" & infl_form="Attr.*"] [tag="Pref"]?
1:[tag="N.*" & tag!="N.num"] within! [word="が|の" | tag="N.*"] [tag="Ai.*" & word!="ない|
無い" & infl_form="Attr.*"] [tag="Pref"]? [tag="N.*" & tag!="N.num"]
```

以上のルール 1 と同じように、ここに挙げたのは 2 単位を取り出すためのルールである。2 は、連体形の形容詞（「無い」を除外）を取り出し、1 は、名詞を取り出す(数詞を除外)。

---

[1] コロケーションにおけるパターンおよびシンタクスの重要性が明らかになる中で、それによく対応しているツールがスケッチエンジンである、ということは McEnery&Hardie (2012: 257, 129, 43-45)により指摘されている。

[2] 「が/の＋「名詞」＋連体形の形容詞＋名詞」を扱うルールは別のルールにしたため、このようなコロケーションデータも抽出できる。

ただし、この形式の前に「が・の・名詞」が無い場合に限るという点はルール 1 と違う点である。

表 3 は、ルール 1 と 2 によって「高い」という形容詞を例にして、取り出せるコロケーションデータを示す。コロケーションタイプは、modifies_N である。ルール 1 とルール 2 によって抽出したコロケーションリストの違いをピンク色でマークした。「高い」が連体修飾節の述語の機能を持ち、連体修飾節が名詞を修飾する例がルール 1 のリストに見られる（表 3 の左側）。例えば、「背が高い人、コミュニケーション能力が高い人、給料が高い人」、「質が高い作品」などがある。一方、ルール 2 のリスト（表 3 の右側）にはその例がないため、単独で形容詞+名詞の例が多い（「高い精度」、「高いハードル」など）。

表 3 ルール 1 とルール 2 で取り出すコロケーションリストの例（「高い＋名詞」）

jpTenTen, ルール 1
頻度 3842021
modifies_N 1301151

物	52241	値段	5164
所	47269	買い物	4936
評価	34341	品質	4611
事	33885	効果	4484
方	20634	水準	4441
人	18208	信頼	4359
位置	17129	建物	4100
為	15942	サービス	3903
レベル	15793	筈	3836
技術	11889	上	3835
金	10532	状態	3741
声	9533	日本	3644
場所	8679	地域	3403
訳	6671	製品	3301
山	6557	そう	3207
人気	6436	数値	3142
気	6059	ビル	3131
場合	5930	値	3101
作品	5865	価格	3084
商品	5819	天井	3009
音	5566	国	2925
確率	5546	クオリティー	2917
時	5291	デザイン	2896
奴	5228		

jpTenTen, ルール 2
頻度 3842021
modifies_N 585081

所	33988	気	2806
評価	30911	数値	2662
物	23027	値	2629
レベル	13553	価格	2599
位置	13548	天井	2515
金	9636	クオリティー	2509
事	9148	精度	2419
方	8805	筈	2387
技術	8745	支持	2354
声	7709	ハードル	2300
人気	5308	目標	2299
山	5268	ビル	2292
確率	5001	次元	2275
場所	4834	奴	2099
値段	4575	給料	2094
買い物	4467	壁	2063
音	4441	安全	2044
効果	4099	パフォーマンス	2031
為	4067	耐久	1994
信頼	3777	能力	1969
水準	3738	金額	1966
品質	3680	場合	1942
訳	3196	性能	1823
建物	2970		

更に、ルール 2 で取り出したデータの制度を検討した結果、「が」「の」が形容詞から離れたところに現れる場合、「は」「も」が使われている場合だけが残され、単独の形容詞と名詞のコロケーションデータとして取り出してしまう例が 5%である。というのは、95%の制度で結果が取り出せることである。例えば、そのような「高い」の例は、以下のようなものである。

- 最新リマスタリングで揃えられているので 、資料的価値は高いはず。
- そのフォルダというラベルの付いた写真、メディア、または可能性が最も高いデータを見つけることができます。
- その効果はもちろん、成功率がとても高いことでも定評があります。
- 最後の戦いも質は高いものの、あっさり目 。
- 頭の回転が速く、人への関心が強く、社会へのアンテナも高い人です。

なお、NINJAL-LWP for BCCWJ というシステムは、コロケーションや文法的振る舞いの情報を抽出するために、係り受け関係のアノテーションを付与した。本研究で扱った「形容詞+名詞」のコロケーションの結果をみる(表 4)と、係り受け関係でも、連体修飾節の述語の機能を持つ形容詞と単独の「形容詞+名詞」のコロケーションの区別ができない。例えば、「高い」を検索すると、名詞のコロケーションには、「高い国」、「高い男」、「高い[人名]」などの不十分だと考えられるコロケーションが表示される。しかし、「の＋名詞＋形容詞」、「が＋名詞＋形容詞」のコロケーションも別のタイプとして抽出できる。

表 4　NINJAL-LWP for BCCWJ の「高い＋名詞」の結果の一部

名詞	頻	名詞	頻度
の	867	【数字】	109
もの	659	割合	107
ところ	572	場合	98
こと	529	声	94
伸び	238	比率	90
評価	236	地位	88
ん	219	とき	81
人	217	場所	79
水準	206	技術	77
ほう	186	地域	75
【地域】	163	【人名】	73
【一般】	159	値段	64
ため	145	わけ	58
位置	140	数値	58
レベル	134	男	55
山	133	国	54
よう	126	物	53

４．まとめ

本研究では、コーパスから取り出した「形容詞（連体形）＋名詞」のコロケーションを対象にして、単独名詞を修飾する形容詞の組み合わせ以外に、そのコロケーションが文節・文章においてどのような構成になるかを検討した。従来の研究は、形容詞の機能を終止用法、連体用法、連用用法の三つに分けたが、連体形の形容詞の機能を更に分けて、検討する必要があるということが本研究で明らかになった。本研究で、9 語の形容詞の用法を

検討した結果、文節・文章における「形容詞（連体形）＋名詞」の構成に以下のようなものがあった。

- 形容詞単独で名詞を修飾する（「寒い季節」）
- 連体修飾節の述語として名詞を修飾する（雨の多い国」）
- 所有の「の」＋形容詞で名詞を修飾する（「タレの甘辛い味」）
- 名詞＋形容詞という構造をもった複合形容詞で名詞を修飾する（「香り高いコーヒー」）このタイプは、連体修飾の「が」の省略と考えられる場合もある。

さらに、各形容詞を比較した結果、それぞれの形容詞の振る舞いの違いがあることが明らかになった。例えば、「高い」という高頻度の形容詞は、連体形の形容詞が連体修飾節の述語であるケースが全体のおよそ半数を占めていた（「質の高いサービス」）。また、形容詞の前に名詞が付く言語表現が 5 例あった（「香り高いコーヒー」）。一方、「青い」は、連体形の形容詞が連体修飾節の述語になるケースが 1 例しかなく、所有の「の」の用法が多く見られる（「日本の青い空」）。本研究では、高･中･小頻度の 3 語ずつの形容詞を対象にしたが、今後はさまざまな形容詞の意味領域を把握するために、数多くの形容詞を検討することが望ましい。

このようなアプローチにより、確率的な方法で取り出すには不十分だと考えられるコロケーション（例えば、「高い＋コーヒー」、「高い＋サービス」）は、更に正確に取り出せるようになる（例えば、「香り高いコーヒー」、「質の高いサービス」）。今後のコロケーション研究においても、コロケーション分析に統語的アプローチを取り入れつつ、実証的および確率的に、語およびその組み合わせの振る舞いを検討し、記述することが望ましい。確率論的アプローチに統語的アプローチを加えることによって、コロケーションデータの取り出し方を更に精密なものにすることができ、実証的なコーパス分析の有意義さが明確に示される。

謝　辞

本研究は、博報財団第 7 回「日本語海外研究者招聘事業」による研究「日本語教育における語の共起関係」（平成 24～25 年度、受入機関：国立国語研究所、招聘研究員：スルダノヴィッチ・イレーナ）の補助を得ています。

文　献

鈴木重幸（1972）『日本語文法・形態論』むぎ書房
スルダノヴィッチ・イレーナ, 仁科喜久子（2008）「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 国書刊行会, pp.59–80.
スルダノヴィッチイレーナ・スホメルヴィット・小木曽智信・キルガリフアダム (2013)「百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング」『「第 3 回コーパス日本語学ワークショップ」予稿集』 国立国語研究所, pp.229-238.
高橋太郎（1998）「動詞からみた形容詞」『言語』27:3, pp.36-43.
小川典子・李在鎬・横森大輔・土屋智行(2008)コーパス調査による形容詞の連体形と連用形の頻度, ICJLE
西尾寅弥、国立国語研究所報告 44(1972)『形容詞の意味・用法の記述的研究』秀英出版
前川喜久雄(2012)「「形容詞＋です」述語の生起要因についての準備的考察」,第 1 回コーパス日本語学ワークショップ予稿集, pp.211-220.
宮島達夫 (1993)「形容詞の語法と用法」『計量国語学』19:2, pp.94-104.
橋本三奈子・青山文啓（1992）「形容詞の三つの用法：終止，連体，連用」『計量国語学』

18:5, pp.201-214.
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』pp.26-10.
八亀 裕美（2008）『日本語形容詞の記述的研究―類型論的視点から』明治書院
姜 紅(JIANG Hong)(2012)コーパスに基づく多義語「甘い」の意味再分類及び語義分布調査 第 1 回コーパス日本語学ワークショップ予稿集, pp.59-68.
Cantos-Gomez, Pascual and Aquilino, Sánchez (2001) Lexical Constellations: What Collocates Fail to Tell. *International Journal of Corpus Linguistics* 6:2, pp.199–228.
Gahl, Susanne (1998) Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus, ms., ICSI-Berkeley
Grefenstette, John (1992) Use of syntactic context to produce term association lists for text retrieval, *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* , New York: ACM, pp.89-97.
Hunston, Susan (2002) *Corpora in Applied Linguistics*. Cambridge University Press
Kilgarriff, Adam, Rychly, Pavel, Smrž, Pavel & Tugwell, David (2004) The Sketch Engine. *Proceedings of EURALEX*. France: Université de Bretagne. pp.105-116.
McEnery, Tony and Hardie, Andrew (2012) *Corpus Linguistics: Method, Theory and Practice.* Cambridge Textbooks in Linguistics. Cambridge University Press
Pomikálek, Jan, Suchomel, Vít (2012) Efficient Web Crawling for Large Text Corpora. ACL SIGWAC Web as Corpus (at conference WWW)
Seretan, Violeta (2011) *Syntax-Based Collocation Extraction*. Berlin: Springer
Srdanović, Irena, Erjavec Tomaž & Kilgarriff, Adam (2008) A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing*) 15:2, pp.137-159.
Stefanowitsch, Anatol and Gries, Stefan. Th. (2003) Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8:2, 209-243.
Tanomura, Tadaharu (2010) Retrieving collocational information from Japanese corpora: Its methods and the notion of “circumcollocate”. Peter Grzybek et al.(eds.) *Text and Language: Structures, Functions, Interrelations*, pp.213-222.

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/
スケッチエンジンツール Sketch Engine　http://www.sketchengine.co.uk/
中納言コーパス検索アプリケーション https://chunagon.ninjal.ac.jp/search
NINJAL-LWP for BCCWJ http://nlb.ninjal.ac.jp