日本語学習者のための名詞と修飾語のコロケーション検索プログラムの開発とその使用例

中溝朋子（山口大学留学生センター）
坂井美恵子（大分大学国際教育研究センター）
金森由美（大分大学国際教育研究センター）
大岩幸太郎（大分大学教育福祉科学部）
刈谷丈治（山口大学名誉教授）

Search Program for the Collocation of Nouns and Their Modification for Japanese Learners: Development and Some Application Examples

Tomoko Nakamizo (International Student Center, Yamaguchi University)
Mieko Sakai (Center for International Education and Research, Oita University)
Yumi Kanamori (Center for International Education and Research, Oita University)
Koutarou Ooiwa (Faculty of Education and Welfare Science, Oita University)
Joji Kariya (Professor Emeritus at Yamaguchi University)

１．はじめに

本研究では日本語学習者用に、『現代日本語書き言葉均衡コーパス DVD 版公開データ』(2011)（以下、BCCWJ と略す）を用いて、名詞を中心語とし、修飾語を共起語とするコロケーションの共起頻度とダイス係数を算出するプログラムを開発している。本発表ではその概要と具体的な検索の結果、およびそこから日本語学習者に示せる例について述べる。本プログラムの特徴は、連体修飾表現として共起する「修飾語」について品詞を問わず頻度を一括して計算し、名詞とのダイス係数、および共起頻度順によって提示することができるという点である。これにより、学習者が類義語名詞の違いを修飾語の違いという点から理解し、また作文などの際に、より適切な連体修飾表現を選択するための支援を行いたいと考える。以下、本プログラムについて具体的に述べる。

２．日本語教育におけるコロケーション、および本プログラムの意義

近年、日本語教育におけるコロケーション習得の重要性は多く指摘され、コロケーションに特化された教材やコロケーションが検索できる web サイトも多く開発されている（神田他 2011、山口他 2012 など）。本研究では、コロケーションの中でも修飾語と名詞のコロケーションに注目し、BCCWJ 内で名詞を修飾する語の使用実態を明らかにするためのプログラムを開発している。

例えば日本語学習者にとって「興味」の強さを表現したいときに、「強い興味」なのか「深い興味」なのか、また「イベントが大きい」という意味を表現したいときに「大きいイベント」なのか、「大きなイベント」なのか、「一大イベント」なのかなど、様々な選択肢の中でどれを選ぶかは判断が難しいと考えられる。コロケーションを検索する web サイトとしては、既に NINJAL-LWP for BCCWJ(NLB)や「日本語作文支援システム『なつめ』」など

があるが[1]、本検索プログラムは後述するような特徴を持つ、修飾語と名詞のコロケーションに特化したものである。将来的には、学習者自身が検索できるようインターフェイスを作成し、現在開発中のコロケーション習得教材とリンクさせることで、コロケーションの習得にも結び付けられるよう計画している。以下、３．で本プログラムの概要について述べる。

３．プログラム概要

３.１．　データ

データには、BCCWJ(2011)の SUW（短単位）可変長データを使用している。

３.２．　分析方法

３．２．１　分析の手順

本プログラムの最も重要な機能は、上記データの文の構成単語列に対する語彙素 and/or 品詞の正規表現による検索機能である。品詞は前方部分だけ指定すれば良い。分析は、まず文中から対象となる部分を品詞の正規表現で検索し、その後集計等の処理を行う。具体的な検索、および分析の手順は、以下の通りである。

(1)検索したい修飾語に応じて任意に「修飾語パターン」「中心語パターン」「共起パターン」を定義し、それに基づき、全サブコーパス（もしくは任意のサブコーパス）について検索を行う。
(2)修飾語、中心語、共起の頻度を計算し、共起頻度表に修飾語頻度、中心語頻度を取り込み頻度表にする。
(3)サブコーパス毎の頻度表を結合し、全頻度表とする。
(4)サブコーパス語数を使って頻度から１億語あたりの相対頻度とダイス係数を計算する。

以下、３．２．２で本研究において定義したそれぞれのパターンについて説明する。

３．２．２．　本研究における「中心語パターン」、「修飾語パターン」、「共起パターン」

本研究では「中心語パターン」と「修飾語パターン」は、表１のように定義した[2]。なお、表中の品詞分類は、BCCWJ で用いられている品詞分類の名称を用いている。

表１　本プログラムで用いた修飾語と中心語のパターン

修飾語パターン		中心語パターン	共起パターンの例
A	<接頭辞>	**<名詞>＋<接尾辞>*** (<動詞以外または文末>)	<一大>**<決心>** <新>**<委員><長>**

[1]NINJAL-LWP for BCCWJ(NLB)　http://nlb.ninjal.ac.jp/
「日本語作文支援システム『なつめ』」　http://hinoki.ryu.titech.ac.jp/natsume/
[2]本研究では、従来コロケーションには含まれない「接頭辞」と「名詞」の組み合わせも検索の対象としている。その理由は、接頭辞の中には形容詞や連体詞などの修飾語と同様の意味を表すものがあるためで（例：「一大」と「大きい」など）、日本語学習者が修飾語を選ぶ際には、これらの接頭辞も修飾語の中のひとつの選択肢となり得ると考えたためである。

B	<連体詞>	**<接頭辞>*<名詞>+ <接尾辞>*** (<動詞以外 または文末>)	<大きな>**<変化>**
C	<形容詞><形容詞-非自立可能>?		<青い>**<空>** <少なく><ない>**<被害>**
D1	<接頭辞>*<形状詞>		<最><重要>**<課題>**
D2	<接頭辞>* <形状詞><助動詞>?<形容詞-非自立可能>?		<不><可能><な>**<計画>** <堂々><たる>**<体格>** <格好><良い>**<御><姿>**
E1	<接頭辞>*<名詞>＋<接尾辞>*<助詞>+		**<外国><人><の><新><委員><長>** <御><隣り><さん><の>**<騒音>**
E2	<接頭辞>*<名詞>＋<形容詞>		<思慮><深い>**<行動>**

・中心語はゴシック、修飾語は明朝で記述

＊は、「ゼロもしくは１以上」を表す

？は、「ゼロもしくは１」を表す

＋は、「１以上」を表す

「修飾語パターン」は、中心語である名詞を修飾する表現について、品詞を問わずできるだけ網羅的に検索できるよう、これまでAからE2のパターンを考えた。「中心語パターン」は、名詞が単独で出現する場合も、名詞に接頭辞が前接する場合にも同じ中心語として扱われ、中心語頻度および修飾語との共起頻度を計算できるよう考慮した[3]。

これらの「修飾語パターン」と「中心語パターン」をそれぞれ共起語と中心語とみなし、Aは<接頭辞>が<名詞>と共起する場合を、A以外は、接頭辞の有無を問わず<名詞>と共起する場合を「共起パターン」として検索した。

検索する正規表現のコマンドでの表現は、表２のとおりである。

表２　各「共起パターン」の正規表現

検索対象	文法	
接頭辞	A	[:"接頭辞"] [#] [(:"名詞")+] [(:"接尾辞")*](""-:"動詞"|$)
接頭辞以外（形容詞、形状詞等）	B	([:"連体詞"] [#]
	C	| [:"形容詞" (:"形容詞-非自立可能")?] [#]
	D1	| [(:"接頭辞")* :"形状詞"] [#]
	D2	| [(:"接頭辞")* :"形状詞" :"助動詞" (:"形容詞-非自立可能")?] [#]
	E1	| [(:"接頭辞")* (:"名詞")+] [(:"接尾辞")* (:"助詞")+]
	E2	| [(:"接頭辞")* (:"名詞")+] [(:"形容詞")*]
		)[(:"接頭辞")* (:"名詞")+] [(:"接尾辞")*] (""-:"動詞"|$)

これらの正規表現を用いたfindPatternコマンドでの検索結果は各語の語彙素、品詞、活用形と、該当部分が表示されている文のtsv形式ファイルである。検索コマンド中で[]で囲まれた部分に一致する語が出力される。[#]は、出力のカラム数を合わせるために空白カラムを挿入する指定である。検索結果２例を表３に示す。

[3]接頭辞の場合と異なり、名詞に接尾辞や名詞が後接する場合は、これらが後接しない場合と別の中心語として検索される。例えば本稿の４．で行う「興味」「関心」の検索では、「御興味」「御関心」の「興味」「関心」は、それぞれ接頭辞のない「興味」「関心」と同じ中心語として計算されるが、「関心事」「興味本位」などは別の中心語として扱われる。

表３　findPattern を用いて検索した結果例　（上段は D1、下段は C の検索例）

検索結果例　（「^」は TAB）
|重要|だ|^ |形状詞-一般|助動詞|^ ||連体形-一般|^ |^ |^ | ^|^役割|^ |名詞-普通名詞-一般|^ ||^ |^ |^ | ^問題は、“冤罪を創作する”ためにマスコミ・メディアが<<<重要な>>><<<>>><<<役割>>><<<>>>を意図してはたす、という新しい状況が派生したことである。
|良い|無い| ^|形容詞-非自立可能|形容詞-非自立可能| ^|連用形-一般|連体形-一般|^ |^ |^ |^ |日|^ |名詞-普通名詞-副詞可能| ^|| ^|^ |^ |^ 調子が<<<良くない>>><<<>>><<<日>>><<<>>>は頭がくらくらし、足がこの通り重くなるんです」と断り、これまでのことを語り始めた。

以下、表４に、本分析に用いた主なコマンドを示す。

表４　本プログラムの主なコマンド一覧

コマンド名	定義
count	指定したカラムで集計、再集計を行う。
select	カラムの文字列としての一致、前方一致、後方一致、部分一致、数値としての大小を判定し、一致、不一致の行を出力する。
columns	カラムの取捨、複製、並べ替えを行う。
sort	文字順、数値の昇順降順で行を並べ替える。
compute	ダイス係数の計算のような簡単な数値計算を行う。
concat	ファイルを連結する。
import	二つの tsv ファイルでカラムが対応したときに指定カラムを取り込む。
findWords	語彙素 and/or 品詞の列で検索する。予備的調査に用いる。
findPattern	語彙素 and/or 品詞の正規表現で検索する。調査に用いる。

以下、4.で本プログラムから得られる結果を用いて日本語学習者に示すことができることの例として、類義語「興味」と「関心」の修飾語を分析した結果について述べる。

４．本プログラムを用いた検索例－「興味」と「関心」－

４．１　辞書による意味の記述

「興味」「関心」については、国語辞典や類語辞典では、表５のように意味が記述されている。

表５　国語辞典・類語辞典による「興味」「関心」の意味の記述（波線筆者）

辞書名	興味	関心
『大辞泉』	①　その物事が感じさせるおもむき。おもしろみ。興。②ある対象に対する特別の関心。「－がわく」③心理学で、ある対象を価値あるものとして、主観的に選択しようとする心的傾向。教育学では、学習の動機付けの一つ。	①ある物事に特に心を引かれ、注意を向けること。「政治に－がある」「幼児教育に－が高まる」「周囲の－の的」
『大辞林』	①物事に心がひかれおもしろいと感じること。おもしろみ。おもむき。「－を覚える」「－をもつ」②〔教・心〕ある対象に対して特別の関心・注意を向ける心的傾向。	①物事に興味をもったり、注意を払ったりすること。気にかけること。「政治に－をもつ」
『類語例解辞典』	(1)「興味」はおもしろいと感じる気持ちや、知りたいと食指を動かされるような気持ちを言う。(2)「興味」が対象のある１点に感情的に向けられることがあるのに対し、「関心」は対象全体に理性的に向けられることが多い。	

このように「興味」と「関心」は、語を定義するのに互いの語が用いられていたり、説明が抽象的であったりなど、辞書の意味の記述だけでは、日本語学習者にとっては必ずしもわかりやすいとは言えない。そこで本プログラムを用いて、より具体的にどのように使用されているかを学習者に示し、両者の違いをさらに明確にさせたいと考える。

４．２　「興味」「関心」と共起する修飾語

本プログラムの結果から、「興味」と「関心」を中心語とする部分を分析する。BCCWJには、「興味」は 7,864 例、「関心」は 6,243 例あった。以下、表６に、検索の対象となる被修飾語の「興味」「関心」が各サブコーパスで実際に出現した頻度（粗頻度）、および各サブコーパスの１億語当たりの頻度（相対頻度）を示す。

表６　「興味」「関心」のサブコーパス別頻度[4]

	頻度	書籍 LB	ベス OB	知恵 OC	法律 OL	国会 OM	広報 OP	教科 OT	白書 OW	ブロ OY	書籍 PB	雑誌 PM	新聞 PN
興味	粗	2597	267	828	0	51	218	59	31	1149	2237	358	69
	相対	7267	6026	6889	0	911	4718	5267	549	8848	6661	6714	4295
関心	粗	2220	179	104	4	301	290	63	413	302	2094	165	108
	相対	6212	4040	865	331	5378	6276	5624	7322	2325	6235	3094	6723

以下、表７に「興味」「関心」とそれぞれの修飾語の共起についてダイス係数上位 20 語を示す。なお便宜上、ダイス係数は 100 万倍して表示している。

表７　「興味」「関心」と修飾語のダイス係数上位 20 語

興味				関心			
順	修飾語	ダイス係数	共起頻度	順	修飾語	ダイス係数	共起頻度
1	強い	3,533	29	1	無	59,736	622
2	デートレードに	3,315	13	2	強い	20,347	169
3	株に	3,306	13	3	重大な	12,047	45
4	子供達の	2,570	13	4	国民の	11,218	60
5	車関係に	2,297	9	5	深い	10,348	61
6	販売に	2,293	9	6	人々の	9,187	44
7	車に	2,186	10	7	大きな	5,768	91
8	物に	2,123	21	8	政治に	3,693	11
9	歴史に	1,895	8	9	高い	3,618	37
10	子供の	1,860	17	10	問題に	3,373	21
11	生徒の	1,817	8	11	問題への	3,194	9
12	話に	1,807	9	12	消費者の	3,032	10
13	科学に	1,761	7	13	性的	2,951	9
14	深い	1,731	12	14	研究者の	2,685	8

[4]表中では省略した BCCWJ のサブコーパスの名称は、以下の通りである。出版サブコーパス（書籍（PB）、雑誌（PM）、新聞（PN））、特定目的サブコーパス（白書（OW）、教科書（OT）、広報紙（OP）、ベストセラー（OB）、Yahoo!知恵袋（OC）、Yahoo!ブログ（OY）、法律（OL）、国会会議録（OM））

15	女性に	1,730	8	15	世間の	2,552	8
16	方に	1,704	10	16	ボランティア活動に	2,506	7
17	食に	1,519	6	17	環境問題への	2,489	7
18	自分に	1,509	10	18	市民の	2472	9
19	純粋な	1,414	6	19	最大の	2,450	10
20	事に	1,385	63	20	主要な	2,397	8

(1)「興味」「関心」を持つ主体を表す修飾語

まず「名詞＋の」で「興味」や「関心」を持つ主体を表す修飾語を比較する。「興味」は、特に「子供（達）の」「生徒の」など年少者に多く使用されている一方、「関心」は「国民の」「人々の」「消費者の」「研究者の」「市民の」など社会的立場や役割を示す集合名詞が多く共起していた。以下、これらの語の一部のサブコーパス別の出現状況を表８に示す。

表８　サブコーパス別「興味」「関心」を持つ主体を表す修飾語（相対共起頻度）

サブコーパス		子供達の(興味)	国民の(関心)	市民の(関心)	人々の(関心)
書籍	LB	2	11	5	89
ベストセラー	OB		22		
yahoo!知恵袋	OC		8		
国会議事録	OM	17	160		
広報紙	OP	21	21	43	
教科書	OT		89		
白書	OW		531		
yahoo!ブログ	OY	7	7		
書籍	PB	23	17	2	32
雑誌	PM	18	18		18
新聞	PN		311	249	
全サブコーパス相対共起語頻度		88	1195	299	139

「興味」と共起する「子供達の」は「書籍(PB)」などで多く見られ、この傾向は「子供の」「生徒の」でも同様であった。また、「関心」と共起する「国民の」は「白書（OW)」で、「市民の」は「新聞（PN)」や「広報紙（OP)」で、「人々の」は「書籍（LB)」で特徴的に出現していた。

(2)「興味」「関心」の強さ・重要性を表す修飾語

次に「興味」「関心」の強さ・重要性を表す修飾語については、「興味」には「強い」「深い」が共起していたのに対し、「関心」には「強い」「重大な」「深い」「大きな」「高い」などが共起しており、語の種類も頻度も「関心」のほうが多いことが観察できた。

以下、強さや重要性を表す修飾語のサブコーパス別の出現状況を表９に示す。「強い」は特に「関心」において頻度が高く、多くのサブコーパスで出現している。一方で「大きな」「重大な」は「国会議事録（OM)」で、「高い」は「白書（OW)」で特徴的に出現しており、「深い」は「ベストセラー（OB)」「書籍（LB)」などで特徴的に出現している。

表９　サブコーパス別「興味」「関心」の強さ・重要性を表す修飾語（相対共起頻度）

サブコーパス		強い(興味)	強い(関心)	深い(興味)	深い(関心)
書籍	LB	36	204	19	75
ベストセラー	OB	90	90	22	135
yahoo!知恵袋	OC		8		
国会議事録	OM		178		53
広報紙	OP	21	21		
教科書	OT				
白書	OW		195		35
yahoo!ブログ	OY	15	15	7	
書籍	PB	26	181	8	59
雑誌	PM		75		37
新聞	PN		124		62
全サブコーパス相対共起語頻度		188	1091	56	456
サブコーパス		大きな(関心)	重大な(関心)	高い(関心)	最大の(関心)
書籍	LB	78	13	22	16
ベストセラー	OB		22		
yahoo!知恵袋	OC				
国会議事録	OM	321	553	53	35
広報紙	OP				
教科書	OT				
白書	OW	141	88	283	
yahoo!ブログ	OY	23		15	
書籍	PB	95	8	11	5
雑誌	PM			56	
新聞	PN	124		62	
全サブコーパス相対共起語頻度		782	684	502	56

(3)「興味」「関心」の対象を表す修飾語[5]

「興味」「関心」の対象を表す語については、「興味」は「科学」「歴史」といった分野を表す語のほかに、「デートレード」「株」といった具体的な事象や「車」「話」「食」など、日常的な語が多く共起しているのに対し、「関心」は「政治」「問題」や、さらに具体的な「環境問題」「ボランティア活動」など、社会的な問題や取り組むべき課題などを表す語と多く共起している。

上記のことから、「興味」と「関心」は、ともに「物事に心ひかれる」「おもしろさを感じる」という意味を表す言葉であることとともに、BCCWJ の検索結果を基に、以下のような特徴を日本語学習者に示すことができると考える。

① 「興味」はより日常的、個人的、「関心」はより社会的、集団的な文脈で用いられることが多いこと
② したがって「関心」では、「関心」を持つ主体を表す言葉には、社会的立場や集団などの意味を表す「国民の」「人々の」などが多く用いられる傾向があること
③ 「興味」「関心」の強さについては、ともに「強い」が最も多く、かつ比較的広い文脈で使用されること
④ さらに「関心」の強さの表現としては、「高い」「大きな」「重大な」「深い」などが共起

[5] 対象を表す修飾語（ニ格）は、「興味」「関心」を修飾する連体修飾語ではないが、本稿では含めて分析を行う。

していたが、これらは特定のサブコーパスで多く使用されており、「強い」に比べると、使用する文脈に注意が必要なこと

⑤ 「興味」「関心」の内容は、「興味」のほうがより具体的、日常的な内容を、「関心」はより社会的な問題や取り組むべき課題に使用される

などが挙げられる。ただし、上記の結果は、書き言葉のコーパスの特徴を反映したものと考えられ、「興味」「関心」の特徴として示すには、話し言葉のコーパスによる調査も必要と考えられる。

５．おわりに

以上、日本語学習者のための名詞と修飾語の検索プログラムの開発の概要と、その検索例と分析例について具体的に記述した。今後は、プログラムの検証、および改善を継続するとともに、日本語学習者が自身で検索できるためのデータの整理とインターフェイスの開発が課題である。

謝　辞

本研究は JSPS 科研費（基盤研究(c) 23520638、および 25370591）の助成を受けています。

文　献

石川慎一郎（2008）『英語コーパスと言語教育』大修館書店
神田靖子、佐尾ちとせ、佐藤由紀子、山田あき子（2011）『連語を使おう』古今書院
坂井美恵子、中溝朋子、金森由美（2011）「類義語『決心』『決意』『決断』の使い分け ―コーパスから見たコロケーションの特徴―」『跨文化交際中的日語教育研究１異文化コミュニケーションのための日本語教育』pp.835-837
中溝朋子、坂井美恵子、金森由美（2012）「共起表現から見た『決定』『決心』『決意』『決断』『判断』の異同について」『日本語教育国際研究大会予稿集（第１分冊）』 p.70
山口久代、竹沢美樹、崔美貴（2012）『コロケーションが身につく日本語表現練習帳』研究社
李在鎬、石川慎一郎、砂川有里子（2012）『日本語教育のためのコーパス調査入門』くろしお出版

資　料

『現代日本語書き言葉均衡コーパス（BCCWJ）DVD 版公開データ』(2011) 国立国語研究所
『大辞泉』　goo 辞書（『デジタル大辞泉』小学館）
『大辞林』　BIGLOBE サーチ（『大辞林　第二版』三省堂）
『使い方のわかる類語例解辞典』（小学館）

関連 URL

現代日本語書き言葉均衡コーパス　http://www.ninjal.ac.jp/corpus_center/bccwj/
日本語学習者のためのコロケーション学習サイト「コロケーション彗星」
http://nagareboshi.susi.oita-u.ac.jp/index.html