名詞句と述語の共起関係から見たコーパス研究

三好伸芳（実践女子大学）

1. はじめに

本研究では、統語情報付きコーパスである NPCMJ を用いて、文中における名詞句と述語の結びつきがどのように分布しているのかを明らかにする。名詞句は、通常、普通名詞、固有名詞、代名詞などから構成され、その語彙的性質の違いから、テキスト内において異なった述語と結びつくことが予想される。しかしながら、文中の名詞句が実際にどのような述語と結びついているのかは、技術的な問題から従来のコーパスでは明らかにすることができなかった。そこで本研究では、以下の 2 点の課題に取り組む。

i. 名詞句とそれを項とする述語はどのように分布しているのか。
ii. 得られた調査結果から、日本語のどのような性質が明らかにされるのか。

2. 先行研究

・従来のコーパス研究

従来のコーパス研究においては、「単独で取り出し可能な語彙／文法項目を設定し、当該項目の特定環境下における分布を調査する」という傾向が強い。例えば、計量国語学会（編）（2017）に所収されている山崎（2017）、丸山（2017）、伊藤（2017）では、そのような手法の事例が紹介されている。

(1) a. 語彙の量的な分布と多様性 〈山崎論文〉
b. 述語の形態変化に現れる文法形式の量的傾向 〈丸山論文〉
c. 品詞の比率に現れる文体差 〈伊藤論文〉

→「動詞／形容詞／名詞」といった品詞情報や「～てもらう／～てくれる」といった形態情報を頼りに、単独で抽出可能な項目のみを扱うのが従来のコーパス研究の基本的なアプローチである。これを仮に**〈単項型〉**のアプローチと呼んでおく。

→これに対し、本研究は「複数の語彙／文法項目の係り受け関係を設定し、当該項目同士の特定環境下における分布を調査する」という、いわば**〈複項型〉**のアプローチをとる。[1]

[1] 従来のコーパスにおいても、〈複項型〉のアプローチが不可能なわけではない（例えば、従属節形式に接続する述語の分布を調査した中俣（2017）などはその好例である）。しかし、形態論情報のみに頼る従来のコーパスで〈複項型〉のアプローチをとる場合、やや煩雑な手法をとるか、隣接した項目間のみの大幅に制約された調査しか行うことができない。前述の中俣（2017）は、従属節に接続する品詞を調査する際に後方共起条件を工夫することでこの問題を部分的に解決している。

・従来のコーパスの検索範囲

前述のようなコーパス研究の方向性は、従来のコーパスが抱える技術的な問題とも関連が深い。例えば、下記のような例における項名詞句と述語を指定して検索することは、現代日本語書き言葉均衡コーパス（BCCWJ）のような形態論情報のみが付加されたコーパスではほぼ不可能であると言ってよい。

(2)　アイヌは、アニミズムといって、山でも川でも沼でも人間同様に考えて、大小二つの沼が並んでいれば、それを親子連れと考えて「親の沼」「子の沼」と呼んだのであります。　34_aozora_Chiri-1956-2;JP

→BCCWJ において、(2)のような例における「アイヌ」と「呼んだ」を何らかの検索条件を指定して取り出すことは、項目間の距離が離れすぎているために事実上不可能である。
→従来のコーパスでは、線状的な形態情報を指定することしかできないため、技術的に〈単項型〉のアプローチに特化せざるをえない。

しかし、本研究が調査に使用する NPCMJ は、上記の「アイヌ」と「呼んだ」のような線状的には捉えにくい階層的関係を取り出すことが可能であり、むしろ〈複項型〉のアプローチを得意とする。以下、実際に本調査の内容を紹介していく。

3. 調査方法

・調査対象

〈単項型〉の研究において、「名詞や動詞（述語）がどのように分布しているのか」を明らかにすることが基礎的なデータとなるように、〈複項型〉の研究においては「名詞や動詞（述語）がどのような関係のもとに分布しているのか」を明らかにすることが重要な基礎的データになると考えられる。そこで本研究では、下記のような述語の項となる名詞（句）の分布を明らかにする。

(3)　{普通名詞／固有名詞／代名詞}　を項とする　{動詞述語／形容詞述語／名詞述語}

→つまり、「普通名詞／固有名詞／代名詞」の 3 項目と「動詞述語／形容詞述語／名詞述語」の 3 項目のかけ合わせになるため、少なくとも 3×3＝9 パターンの検索が必要となる。

・調査に際しての前提

(3)の条件に合致するのは、以下のような検索式である。

(4) [N] / [NPR] / [PRO] > ([NP] > ([^PP] > ([^IP] < [VB])))
[ADJI])))
[NP-PRD])))

→例えば動詞述語の項となっている普通名詞を抽出するためには「[N] > ([NP] > ([^PP] > ([^IP] < [VB])))」（「動詞を直接支配する節の直接支配下にある普通名詞」）という検索式を用いることになる。

実際に(4)の組み合わせで得られたデータを示す。

表 1　名詞句と全述語の共起関係

	動詞述語	形容詞述語	名詞述語	総計
普通名詞	46183	2334	2234	50751
	91.0%	4.6%	4.4%	100.0%
固有名詞	5162	162	376	5700
	90.6%	2.8%	6.6%	100.0%
代名詞	4158	123	444	4725
	88.0%	2.6%	9.4%	100.0%

→検索対象が極めて一般的な構造なので、NPCMJ のような規模のコーパスであってもある程度のデータを集めることが可能。

しかし、(4)の検索式には何点か問題がある。以下、その点を確認する。

問題点①：述語ごとにとりうる格の差異が考慮されていない。

(5) a. こうした交流を大切にしてほしい。
109_news_KAHOKU_52;K201403260A0T10XX00001;JP

b. 二人はそんな恰好でグレゴールの部屋へ入っていった。
1145_aozora_Harada-1960;JP

→(5)は「普通名詞-動詞」の組み合わせを検索したものであるが、ヲ格／ニ格／デ格／ヘ格といった、形容詞述語や名詞述語の項になりにくい格が出現している。このような例は動詞述語の検索数を著しく上昇させ、対等な条件下における分布を見えにくくする。
→本研究では助詞「ハ／ガ」が付加された名詞句のみを調査対象とする。

問題点②：述語の現れる環境が考慮されていない。

(6) a. これらは、インストールなどの日常操作の多くを実行するための事前作成された指示プログラムです。 46_nonfiction_IBM-1401;JP

b. 「楽都仙台」は、市民一般に既に浸透していた「学都仙台」と同音であるため、市の音楽事業以外にもこのような仙台の音楽イベント全般を形容する際にも用いられるようになり、さらに仙台市民のライフスタイルを表す言葉としても使用されるようになった。 37_wikipedia_Sendai_Music;JP

→(6)の波線部全てが節の述語と見なされ、下線部の名詞全てが検索対象と見なされる。従属節にはハの出現に制約があるなど、こちらも項となる名詞句の分布が見えづらくなるおそれがある。

→本研究では主節述語のみを調査対象とする。

つまり、名詞句と述語の双方に、さらに特定的な条件を加える必要があるということである。具体的には、以下のような検索式を用いることで、上記の問題がおおむね解消される。

(7) [N] / [NPR] / [PRO] } > ([NP] > (([^PP] < ([^P] << /は/)) > ([IP-MAT] < / > ([NP] > (([^PP] < ([^P] << /が/)) > ([IP-MAT] < { [VB]))) [ADJI]))) [NP-PRD])))

→名詞句に関する条件（「([^PP]<([^P]<</は/))」の部分）と述語に関する条件（「[IP-MAT]」の部分）が補われている。「は／が」の違いを扱うので、最終的には3×3×2（「は／が」）＝18パターンの検索を行うことになる。

4. 調査結果と分析

・調査結果

まず、比較対象として「主節における述語の比率」を示しておく。

表２　主節述語全体の内訳

	動詞述語	形容詞述語	名詞述語	総計
主節述語	20124	1137	3038	24299
	82.8%	4.7%	12.5%	100.0%

→圧倒的に動詞述語が多く、名詞述語、形容詞述語と続く。

続いて、本調査全体の結果を示す。

表 3　名詞句と述語の共起関係[2]

	動詞述語			形容詞述語			名詞述語			総計
	-ハ	-ガ	合計	-ハ	-ガ	合計	-ハ	-ガ	合計	
普通名詞[3]	2768	1579	4347	214	117	331	600	55	655	5333
	51.9%	29.6%	81.5%	4.0%	2.2%	6.2%	11.3%	1.0%	12.3%	100.0%
固有名詞	997	257	1254	20	1	21	98	8	106	1381
	72.2%	18.6%	90.8%	1.4%	0.1%	1.5%	7.1%	0.6%	7.7%	100.0%
代名詞	492	21	513	23	1	24	151	22	173	710
	69.3%	3.0%	72.3%	3.2%	0.1%	3.4%	21.3%	3.1%	24.4%	100.0%

→普通名詞はおおむね平均的な振る舞いを見せるのに対し、固有名詞と代名詞はやや予測に反する傾向が見られた。

まず、比較的要因がはっきりしている点について確認する。

観察①：形容詞述語と名詞述語と共起するガ格名詞句は少ない傾向にある。

(8)　a. ところが、ドアはもう二度と開かれず、グレゴールが待っていたこともむなしかった。　486_aozora_Harada-1960;JP
　　b. 狩野山楽（1559-1635）がその中心人物である。　46_nonfiction_IBM-1401;JP
(9)　わあがない。　118_aozora_Miyazawa-1934;JP

→恒常的性質を表す述語なので、「ガ」による標示が少なくなるのは当然と言える[4]。

観察②：「ガ格普通名詞-形容詞述語」の組み合わせは相対的に生起しやすい。

(10)　a. あの子は身体の工合がよくないんです。　188_aozora_Harada-1960;JP
　　b. 7 月から 8 月にかけては、熱帯夜が多く、暑さが厳しい。 18_wikipedia_Shanghai;JP

[2] 集計に際しては、「辞書」、「その他」、「教科書」、「法文」の 4 ジャンルは除いている。これらのコーパスにおいては、特定の構文が連続して使用されるなど、サンプルとして適さない性質がある。
[3] 普通名詞の集計に際しては、「こと／ころ／とき／ところ／の／もの／はず／つもり」（漢字表記を含む）などの形式化しているものは除いた。形式化の程度は連続的であり、全ての形式的な名詞を排除したわけではないが、上記以外のものはそれほど数も多くなく、結果への影響は軽微であると考えられる。
[4] ただし、「代名詞-名詞述語」の組み合わせは、助詞の別を問わず多く観察される。収集した例を見ると、次のようなプレゼンテーション（テッドトーク）での用例が目立つことから、なんらかの提示的な機能により、代名詞が用いられていると考えられるが、詳細は改めて論じたい。

(i)　a. これがその様子です　52_ted_talk_9;AnthonyAtala_2009P;JP
　　b. これはサンショウウオの写真です　28_ted_talk_9;AnthonyAtala_2009P;JP

→原理的に部分主語を取れるのが普通名詞に限られるため、ガ格と形容詞述語が生起しやすい。

→以上の点は、いわば直感的に自明のことであるが、このような名詞句と述語の関係に関する直感を表 3 のような量的なデータとして示せることは、〈複項型〉のアプローチを得意とする NPCMJ の極めて重要な特徴である。

・分析

表 3 の中でも特筆すべき点として、次のような特徴が挙げられる。

観察③:「ガ格代名詞-動詞述語」の組み合わせは、相対的に極めて生起しにくい。

→「ガ格固有名詞-動詞述語」が普通に生起していることからすると、このような事実は、直感的に自明とは言えない。以下、この振る舞いの要因について考察する。

「ガ格代名詞-動詞述語」に該当するのは次のような例である。

(11) a. 私がレンズの上に指を置いたに違いない。 444_fiction_DICK-1952;JP
b. そして、商品見本はまだ包装してないし、彼自身がそれほど気分がすぐれないし、活溌な感じもしないのだ。 61_aozora_Harada-1960;JP
c. 落ちるときに立てるにちがいない大きな物音のことを考えると、それがいちばん気にかかった。 139_aozora_Harada-1960;JP

→「ガ格代名詞-動詞述語」は、ほとんどの例が総記の解釈の傾向を持つ。

このような傾向は固有名詞には見られない。

(12) a. 「何か用かね？」と、ザムザ氏がたずねた。 1201_aozora_Harada-1960;JP
b. 窓の下の耕助が言いました。 39_aozora_Miyazawa-1934;JP
c. 1926 年(大正 15 年)に仙台市電が開業した。 157_wikipedia_Sendai_City;JP

→固有名詞の場合には、同様の環境でも問題なく中立叙述として解釈できることが多い。

以上のような事実は、内省によっても確認される。この点について観察するため、日本語の存在文について取り上げる。西山（1994: 118-124）は、日本語の存在文には一般的に不定名詞句が現れる環境で特定の対象や固有名詞が現れると指摘している。

(13) a. 机の上に洋子のバイオリンがあった。（西山 1994: 118）
b. （人混みのなかで）おや、あんなところに妹の洋子がいる、いったい何をしているんだろう。（西山 1994: 123）

しかし、主語名詞句を代名詞に置き換えると、同じ文脈でも不自然になるか、総記の解釈でなければ容認が難しくなる。

(14) a. ?公園に彼女がいた。
（cf. え、公園に彼女がいるの？）
b. ?おや、あんなところに彼がいる。
（cf. え、あんなところに彼がいるの？）

→同じ定指示とされる固有名詞と代名詞であっても、総記の解釈の現れ方には差異があり、動詞述語と共起した場合、照応的な代名詞は中立叙述の解釈と馴染まない[5]。

以上の観察から、次のような一般化が示せる。

(15) ガ格を伴って動詞述語と共起した際、固有名詞は中立叙述の解釈を容認するが、代名詞は原則として容認しない。

→日本語の場合、固有名詞は定性制約に抵触せず、普通名詞と同じような振る舞いをする場合があるが、代名詞の場合には定性制約に抵触しやすい。

ガ格の解釈については、久野（1973: 32）による、次のような述語に着目した一般化が受け容れられている。

(16) 述語が恒常的性質を表す場合、ガ格名詞句は総記解釈となる。

→本研究では、NPCMJ で得られたデータの分布から、項となった名詞句と総記解釈との対応関係という、これまでにない見方を提示した。

5. おわりに

以上、本研究では「名詞句と述語の分布」、「得られたデータに基づく質的分析」を行った。結果として、名詞句とそれを項とする述語の大まかな全体像が明らかにされ、その結果から名詞句と総記解釈の対応関係について従来とは違った角度からの一般化を提示した。

[5] 特にデータは挙げられていないが、西山（1994: 119）にも同趣旨の指摘が見られる。

一方、本研究でやり残したことは多い。今後の課題として挙げられる点を以下に挙げる。

① 〈複項型〉のアプローチによる文体論的研究
今回の研究はジャンルとの関係について部分的な指摘に留まっている。コーパスの規模の問題もあるが、文体的特徴と並行して、名詞句と述語の分布にも差異が現れる可能性がある。

② 検索の精緻化・体系化
名詞句の側では、形式的な名詞や特殊な語彙的意味を持つ名詞（身体部位名詞など）に対する配慮、述語の側では付属する文法形式に対する配慮などと、さまざまな要因を検討する必要がある。

③ 直感的に説明ができない分布に対する分析
本研究においては、代名詞と名詞述語が相対的に結びつきやすいという点については明確な説明が与えられなかった。これについては改めて分析の機会を設けたい。

→従来のコーパスではもっぱら〈単項型〉のアプローチをとることしかできなかったが、NPCMJ により、〈複項型〉のアプローチに対する可能性が大きく拓かれたと言える。

［付記］
本研究は国立国語研究所の共同研究プロジェクト「統語・意味解析コーパスの開発と言語研究」の研究成果を報告したものである。
本研究は JSPS 科研費による若手研究「現代日本語における述語と補部の相互作用に関する研究」（課題番号：19K13155）の成果の一部である。

【参考文献】

伊藤雅光（2017）「第 5 章 文章・文体」，計量国語学会（編）（2017）『データで学ぶ日本語学入門』，pp. 45-55.
久野暲（1973）『日本文法研究』大修館書店.
計量国語学会（編）（2017）『データで学ぶ日本語学入門』朝倉書店.
中俣尚己（2017）「接続助詞の前接語に見られる品詞の偏り―コーパスから見える南モデル―」『日本語の研究』13-4，pp. 1-17.
西山佑司（1994）「日本語の存在文と変項名詞句」『慶應義塾大学言語文化研究所紀要』第 26 号，pp. 115-148，慶應義塾大学言語文化研究所.
丸山直子（2017）「第 4 章 文法・意味」，計量国語学会（編）（2017）『データで学ぶ日本語学入門』，pp. 33-44.
山崎誠（2017）「第 3 章 語彙」，計量国語学会（編）（2017）『データで学ぶ日本語学入門』，pp. 22-32.