「NP_1 の NP_2」の意味情報のアノテーション

鈴木 莉子[1]　高山 沙也加[2]　北川 舞[3]　田中 リベカ[4]　峯島 宏次[5]　戸次 大介[6]
お茶の水女子大学

{g1520544[1], g1520524[2], g1520510[3], tanaka.ribeka[4], bekki[6]}@is.ocha.ac.jp
mineshima.koji@ocha.ac.jp[5]

1　はじめに

2 つの名詞句をつなぐ助詞「の」を使った表現には様々な意味がある。例えば、(1) の文の内容を「の」を含まない (2) の文でも伝えることができる。

(1) 作家の夏目先生の最初の本は、明治時代の東京の描写の最高峰と評価される。

(2) 作家である夏目先生が最初に書いた本は、明治時代における東京を描写した最高峰と評価される。

ここで、「作家の夏目先生」は「夏目先生は作家である」ということを意味し、「最初の本」は「最初に書いた本」という意味に取れる。このように助詞「の」を使った表現には多義性がある。以下では、この形の表現を「NP_1 の NP_2」と表す。

本研究は、NPCMJ コーパス (NINJAL Parsed Corpus of Modern Japanese) [4] に「NP_1 の NP_2」についての意味情報をアノテートすることを目的とする。NPCMJ は国立国語研究所で開発中の日本語ツリーバンクで、現在、統語構造・意味情報を付与された約 1 万文が検索インターフェイスとともに公開されている[1]。文の階層構造に基づいて、空範疇の情報を含む詳細な統語情報が付与されている点に特徴があり、言語処理だけでなく、言語学分野での言語理論の構築・評価・検証にも用いられることが期待されている。NPCMJ の統語構造は、形式意味論に基づく意味解析とも整合的であり、論理式を出力する意味解析システムへの入力として用いることができる [1]。現在、NPCMJ から範疇文法ツリーバンクへの変換プロジェクトも進行中である [3]。

NPCMJ では、「NP_1 の NP_2」の用法の一部については区別されているものの、その意味の多様性についてはまだ十分に扱われていない。そこで、最近の言語学の研究をふまえて、「NP_1 の NP_2」についての意味情報をアノテートすることを試みる。

本稿の構成は以下の通りである。2 節では、「NP_1 の NP_2」の分類に関する先行研究について述べ、3 節では、本研究で提案する「NP_1 の NP_2」の意味分類について説明する。4 節では、アノテーションの結果を報告し、アノテーションを通して得られた問題点を明らかにする。

[1] http://npcmj.ninjal.ac.jp/

2　先行研究

助詞「の」を伴う名詞句の意味構造については言語学・言語処理の両分野ですでに多くの研究がある。言語学では西山 [9] が、飽和名詞・非飽和名詞という区別を導入した上で、NP_1 と NP_2 の意味関係に基づいて「NP_1 の NP_2」の 5 つのタイプの分類を示した。これを出発点として、様々な言語学的な分類の拡張が試みられている [8, 10]。

飽和名詞・非飽和名詞の区別は、言語処理の文脈でも広く認識されており、イベントや状態を表す名詞の述語項構造 [11] に加えて、この区別を反映した名詞句のアノテーション [2]、辞書データ構築 [7]、格フレームの自動構築と間接照応の解析 [12] などが行われている。しかし、飽和名詞・非飽和名詞の区別を超えて、西山 [9, 10] の詳細な分類が実テキストの解析にどれくらい有効であるのかは、まだ明らかではなく、アノテーションから得られる言語学的フィードバックについても十分な検討がなされていない。本研究では、西山 [9] の分類を土台として、アノテーション・ガイドラインを作成した。分類の詳細は 3 節で説明する。

3　「NP_1 の NP_2」の分類

筆者らが作成した「NP_1 の NP_2」の意味分類を表 1 に示す。各分類の詳細を述べる前に、この分類で重要な概念となる名詞のタイプ分けについて説明する。3.2 節以降で、「NP_1 の NP_2」のラベル体系について述べる。

3.1　飽和名詞と非飽和名詞

西山 [9, 10] による「NP_1 の NP_2」の意味分類では、飽和名詞・非飽和名詞・譲渡不可能名詞の区別が重要な役割を果たす。飽和名詞とは、「俳優」「作家」「会社員」など、論理的には 1 項述語に対応する名詞を指

分類ラベル	概要	下位分類	例
[飽和]	NP_1 と NP_2 が何らかの関係 R を有する	[一般][空間][時間][順序]	山田先生の本、今の仕事
[コピュラ]	「NP_1 である NP_2」という意味関係にある		病気の父、待望の娘
[非飽和]	NP_1 が NP_2 の意味的な項である	[一般]	太郎の妹、大会の優勝者
		[サ変名詞-ガ格・ヲ格・ニ格]	先生の指示、薬の調合、大学の合格
		[形容詞]	研究の面白さ、世界の不思議
[部分全体]	NP_1 が全体を表し、NP_2 がその部分を表す		キリンの首、服の襟
[時間断片]	時間領域 NP_1 での NP_2 の断片を固定する		大正末期の東京
[同格]	NP_1 の種類を NP_2 で補足する		菊の花、日本の国
[例示]	NP_1 が NP_2 の例を示す		風水害などの自然災害
[形式名詞]	NP_2 が形式名詞である		過去のこと、彼女のおかげ
[数量・限定]	NP_1 もしくは NP_2 が数量・限定を表す		3 人の学生、何らかの原因
[慣用表現]	固有名詞化していたり一語化している		阿吽の呼吸、みどりの窓口

表 1: 「NP_1 の NP_2」の意味分類

す。非飽和名詞は、「主役」「作者」「社員」など、論理的には多項述語（典型的には 2 項述語）に対応する名詞である。

飽和名詞と非飽和名詞を区別するために、いくつかの言語的テストが提案されている [10]。例えば、非飽和名詞である「主役」の場合、「あの人物は主役か」という問いに対しては、「X の主役」のパラメータ X を埋めなければ答えることが難しい。一方、「主役」とよく似た「俳優」の場合、「あの人物は俳優か」という問いは文脈からの補足なしに答えることが可能であり、飽和名詞に分類される。このテストを含めて、いわゆるカキ料理構文テスト [9] など、言語学の先行研究で提案されているテストを飽和名詞と非飽和名詞を区別するために使用する。

譲渡不可能名詞は、「手」「玄関」「ハンドル」のように、それが部分として機能する基体（人・家・車）を必要とする表現である。基体への参照を伴うという点で非飽和名詞と似ているが、「この手は誰の手か」という問いが自然であるのに対して、「この作者は何の作者か」という問いが不自然であることなど、非飽和名詞と譲渡不可能名詞の意味的な違いが指摘されている [6]。本稿では、この議論をふまえて、非飽和名詞と譲渡不可能名詞を区別して「NP_1 の NP_2」のアノテーションを試みる。

3.2 飽和ラベル

NP_1 と NP_2 の間に文脈から決定される何らかの関係 R を補うことで「NP_1 の NP_2」が解釈されるとき、[飽和] ラベルを付与する。R としてとりうる関係の種類はきわめて多様である。

(3) [[$_{NP_1}$ 山田先生] の [$_{NP_2}$ 本]] [一般]

(3) では、NP_1 と NP_2 の間に、文脈に応じて、例えば「山田先生が書いた本」「山田先生が所有している本」など複数の関係が考えられる。

補う関係が比較的に明確なものとして、[空間][時間][順序] の 3 つの下位分類を設けた。具体例を以下に示す。

(4) a. [[$_{NP_1}$ 内側] の [$_{NP_2}$ 壁]] [空間]
b. [[$_{NP_1}$ 今] の [$_{NP_2}$ 仕事]] [時間]
c. [[$_{NP_1}$ 最初] の [$_{NP_2}$ 結婚]] [順序]

アノテーション時には、関係がこのうちのいずれかに特定される場合にのみ [空間][時間][順序] に分類し、それ以外の場合は、[一般] に分類する。

典型的には NP_2 は飽和名詞であるが、NP_2 が非飽和名詞であっても NP_1 が直接の意味的な項になっていない場合はこの分類に該当する。例えば、「北海道の妹」は、「北海道に住む（太郎の）妹」という解釈の場合、[飽和-一般] に分類される。

3.3 非飽和ラベル

NP_1 が NP_2 の意味的な項となっている場合、[非飽和] に分類する。NP_2 の種類に対応して [一般][サ変名詞][形容詞] の下位分類を設けた。NP_2 がサ変名詞あるいは形容詞由来の場合、それぞれ [サ変名詞][形容詞] とする。NP_2 がそれ以外の非飽和名詞である場合、例えば「太郎の妹」は、[一般] とする。

[サ変名詞] に該当する NP_2 と項となる NP_1 との関係に応じて、さらに、[ガ格][ヲ格][ニ格] に分類する。それぞれの例を以下に示す。

(5) a. [[$_{NP_1}$ 先生] の [$_{NP_2}$ 指示]] [ガ格]
b. [[$_{NP_1}$ 物理学] の [$_{NP_2}$ 研究]] [ヲ格]
c. [[$_{NP_1}$ 組織] の [$_{NP_2}$ 協力者]] [ニ格]

(5a) は「先生が指示する」より [ガ格] であり、(5b) は「物理学を研究する」であるから [ヲ格] となる。一方、(5c) では、NP_2 が複合名詞となっているが、「組織に協力する者」と解釈できるので、[ニ格] となる。

NP_2 が形容詞由来の名詞である場合は、[形容詞] とする。例として「発表の面白さ」「市民の安全」「世界の平和」などが挙げられる。

3.4 部分全体ラベル

NP_2 が譲渡不可能名詞で NP_1 がその基体を表す表現であるとき、[部分全体] に分類する。何を部分と全体の関係とみなすかには解釈の幅があるため、ここでは、「太郎の手」「車のエンジン」のように、NP_1 も NP_2 もそれぞれ単独で具体物を指示するものに限定する。「大学の文学部」のように NP_2 が組織のような抽象物を指示する場合や、「この年の 11 月」のように NP_1 も NP_2 も時間表現である場合は、[部分全体] には含めない[2]。

3.5 その他のラベル

[コピュラ][時間断片] は、西山 [9] の分類に従う。[同格] は NP_1 がどのような種類の対象であるのかを NP_2 で補足的に明示するケースであり、「菊の花」「君が代の歌」など先行研究 [8] で挙げられている例のほか、「松子と梅子の二人の少女」のようなケースを含める[3]。

[数量・限定] は NP_1 もしくは NP_2 が数量を表す表現であるものや、「例のウワサ」「問題の建物」のように直示的・照応的な限定表現であるものが該当する。この他、「りんごなどの果物」のように NP_1 が NP_2 の例となっているものを [例示]、NP_2 が形式名詞であるものを [形式名詞] に分類する。該当する形式名詞の一覧はガイドラインに明示した。なお、「NP_1 の NP_2」の形をしているが、固有名詞化していたり一語化している表現、例えば「氷山の一角」「鰻の寝床」「針の筵」などは [慣用表現] とした。

4 アノテーション結果と考察

4.1 アノテーション

NPCMJ の 1482 文（886 事例）について、5 名の作業者でアノテーションを行った。使用したデータは、小説、新聞記事、Wikipedia 記事など複数ジャンルにわたる。このうち 195 事例については、非エキスパートによるアノテーションの品質を調査するため、言語学の専門知識を持たない 3 名の作業者によるアノテーションを実施した。195 事例を 3 分割し、各パートを 1 名の作業者が担当した。その結果を予測ラベルとする。

正解ラベルは、言語学の専門知識をもつ 2 名の作業者が作成した。まず、195 事例を 2 分割し、各パートについて一方の作業者がアノテーションを行い、他方がチェックを行った。このうち判定が一致しなかったものが 19 件あり、議論の上で正解ラベルを決定した。

[2]「大学の文学部」は [非飽和-一般]、「この年の 11 月」は [飽和-時間] に分類した。詳細は公開予定のガイドラインを参照のこと。

[3][同格] には制限があり、松子が少女であったとしても、「松子の少女」は不自然である。NP_2 が「二人の」のような数量表現を伴うとき [同格] が自然となるのは興味深い。

分類ラベル	適合率	再現率	F1	件数
飽和-空間	0.75	0.80	0.77	15 (29)
飽和-時間	0.89	0.80	0.84	10 (66)
飽和-順序	1.00	1.00	1.00	2 (5)
飽和-一般	0.60	0.58	0.59	36 (131)
コピュラ	0.82	0.64	0.72	14 (133)
非飽和-サ変名詞-ガ格	0.67	0.29	0.40	7 (52)
非飽和-サ変名詞-ヲ格	0.72	0.93	0.81	14 (73)
非飽和-サ変名詞-二格	0.00	0.00	0.00	2 (3)
非飽和-形容詞	0.00	0.00	0.00	0 (3)
非飽和-一般	0.81	0.69	0.75	55 (171)
部分全体	0.33	1.00	0.50	4 (24)
時間断片	0.33	1.00	0.50	1 (6)
同格	0.67	0.50	0.57	4 (8)
例示	0.33	1.00	0.50	1 (10)
形式名詞	0.95	0.90	0.93	21 (103)
数量・限定	0.50	0.62	0.56	8 (66)
慣用表現	0.00	0.00	0.00	1 (3)
飽和	0.74	0.73	0.74	63 (231)
非飽和	0.85	0.77	0.81	78 (302)
全体	0.73	0.70	0.71	195 (886)

表 2: 各ラベルの適合率、再現率、F 値。括弧内の数字は 1482 文全体での出現件数を表す。ただし、NPCMJ ではコピュラ、形式名詞、数量・限定の一部について統語的な区別があり、総件数はそれらの事例も含む。

正解ラベルに対する予測ラベルの適合率、再現率、F 値を算出した。分類ラベルごとの結果を表 2 に示す。16 ラベル全体のマイクロ平均を求めると全体の F 値は 0.71 であった。上位ラベルのうち、[飽和] の F 値は 0.74、[非飽和] の F 値は 0.81 であり、一定の品質が得られたと言える。

4.2 考察

[飽和] と [非飽和] の区別の問題点：抽象名詞

エラー分析の結果、最も多かったのが [飽和] と [非飽和] の対立である。正解ラベルが [非飽和] だが [飽和] が振られていた例が 55 件中 8 件、[飽和] だが [非飽和] が振られていたのが 36 件中 6 件あった。

3.1 節で述べたように、言語的テストによって、飽和名詞と非飽和名詞は典型的なケースでは判別可能である。しかし、これらのテストは、NP_2 が人や物のような具体物を指示する場合には容易に適用できるが、抽象物を指示する場合、適用が困難となる。例えば、「その地方の文化」における「文化」の場合、「あなた（これ）は文化か」という形の自然な問いを作ることがそもそも難しい。また、NP_2 を非飽和名詞と判定したとしても、NP_1 がその意味的な項であるのか否かは自明ではない。他にも、判定の不一致が見られた例として、「鈴木さんの言葉」「肉親の情」などがある。

[飽和] の下位分類の問題点

その他に多かったのが、[飽和-一般] と [飽和] のそれ

以外の下位分類の対立である。[飽和-一般] が正解ラベルでその他の下位分類が対立ラベルとなっている例が 36 件中 1 件、その逆の例が 27 件中 2 件あった。

3.1 節にあるように、関係 R が文脈に応じて様々な関係となり得るものは、[飽和-一般] に分類するが、R が複数の関係を取り得るか否か、意見が分かれる例があった。例えば、「部屋の灯」の場合、「ある部屋の中の灯」という解釈では [飽和-空間] に分類できる。しかし、「部屋」を全体として、「灯」をその一部と見なすこともできるという考えもあり、複数の関係を持ち得るか意見が分かれた。

[部分全体] の問題点

本質的に判定が困難なものとして、[部分全体] と [飽和] が挙げられる。元は全体を構成する一要素だが、構成することを放棄するのが当然となる「死体」「抜け毛」のような変化系は [飽和]、単体で見ると飽和名詞のようである「ビルのエレベーター」「部屋のカーテン」などの名詞句は [部分全体] のように、難しいケースはあるが判別可能である。

一方で、判別が困難な例もある。まず、NP_2 が「入れ歯」「眼鏡」のような、基体に付加され着脱可能なものがある。また、体を構成する部分だが体に対する機能を放棄するのが当然となる「太郎の汗」や、NP_2 が「よだれ」「涙」のようなケースでも、[部分全体] か否かの判別が難しい。その他にも、NP_2 が「灰」「燃えかす」「残り」「汚れ」のような場合も、[部分全体] か [飽和] かの区別は明らかではない。

一方、以下の文の「ジェームズ・ヘプバーンの血」では、NP_2 の「血」は、抽象的な意味を含む。

(6) ジョゼフが [[$_{NP_1}$ ジェームズ・ヘプバーン] の [$_{NP_2}$ 血]] をひいていたという事実はない。

このような例では、アノテーションの判定は分かれず [非飽和-一般] で一致した。

5 おわりに

本研究では言語学での先行研究を基に「NP_1 の NP_2」の意味分類を再考した。従来よりも細かく区別することで分類の網羅性を確保した。アノテーションの結果、4.2 節で述べたような興味深い境界事例が見つかった。アノテーション結果はガイドラインとともに公開予定である。今後は本研究で提案した分類を NPCMJ に統合することを目指す。

また、「NP_1 の NP_2」の意味分類はこの表現の名詞句を含む文間の含意関係 [5] と密接な関係がある。例えば、[飽和-一般] に分類される「あれは [[$_{NP_1}$ 山田先生] の [$_{NP_2}$ 車]] である」の場合、この文は「あれは車である」を含意するが、「あれは山田先生である」を含意しない。一方、[コピュラ] の場合、NP_1 を除去する推論も NP_2 を除去する推論も成り立ち、「彼女は [[$_{NP_1}$ 看護婦] の [$_{NP_2}$ 花子]] だ」は「彼女は看護婦だ」も「彼女は花子だ」も含意する。ただし、含意関係の成立には様々な要因が関与しているため、「NP_1 の NP_2」の各事例について含意関係が成り立つかどうかをアノテートすることは、「NP_1 の NP_2」の意味分類の精緻化につながるだけでなく、理論的にも興味深い課題である。それにより、NPCMJ から実テキストの複雑さをもつ含意関係認識データセットを構築することにもつながる。今後、「NP_1 の NP_2」を含む文の含意関係のアノテーションも同時に進める予定である。

謝辞 Alastair Butler、Stephen Wright Horn、長崎郁、窪田悠介、林部祐太の各氏に感謝します。

本研究は JSPS 科研費 15J11772 の助成、及び JST CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域「知識に基づく構造的言語処理の確立と知識インフラの構築」プロジェクトの支援を受けたものである。

参考文献

[1] Alastair Butler. *Linguistic Expressions and Semantic Processing*. Springer, 2015.

[2] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In *Proceedings of LREC*, pp. 2008–2013, 2002.

[3] Yusuke Kubota and Koji Mineshima. From Keyaki to ABC: A treebank conversion project, 2017. 「統語・意味解析コーパスの開発と言語研究」第 2 回研究発表会.

[4] アラステア・バトラー, 吉本啓, 岸本秀樹, プラシャント・パルデシ. 統語・意味解析情報付き日本語コーパスのアノテーション. 言語処理学会第 22 回年次大会発表論文集, pp. 589–592, 2016.

[5] 川添愛, 田中リベカ, 峯島宏次, 戸次大介. 日本語意味論テストセットの構築. 言語処理学会第 21 回年次大会発表論文集, pp. 704–707, 2015.

[6] 西川賢哉. 「NP_1 の NP_2」タイプ F—譲渡不可能名詞 NP_2 とその基体表現 NP_1. 西山佑司（編）, 名詞句の世界, pp. 65–82. ひつじ書房, 2013.

[7] 竹内孔一, 宮田周, 河村一希. 述語項構造を意識した名詞データの構築. 第 7 回コーパス日本語学ワークショップ予稿集, pp. 143–146, 2015.

[8] 三宅知宏. 「主要部」の概念と “X の Y” 型名詞句. 日本語研究のインターフェイス, pp. 79–87. くろしお出版, 2011.

[9] 西山佑司. 日本語名詞句の意味論と語用論—指示的名詞句と非指示的名詞句—. ひつじ書房, 2003.

[10] 西山佑司（編）. 名詞句の世界. ひつじ書房, 2013.

[11] 飯田龍, 小町守, 乾健太郎, 松本裕治. 名詞化された事態表現への意味的注釈付け. 言語処理学会第 14 回年次大会発表論文集, pp. 227–280, 2008.

[12] 笹野遼平, 河原大輔, 黒橋禎夫. 名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析. 自然言語処理, Vol. 12, No. 3, pp. 129–144, 2005.