述語の概念フレームと PropBank 形式の意味役割を付与した NPCMJ-PT の構築

竹内 孔一 [1]　バトラー アラステア [2]　長崎 郁 [3]　パルデシ プラシャント [4]
[1] 岡山大学 [2] 弘前大学 [3] 名古屋大学 [4] 国立国語研究所
takeuc-k@okayama-u.ac.jp

概要

日本語のコーパスに対して構文木を付与して Web 上で公開している NPCMJ (NINJAL Parsed Corpus of Modern Japanese) に対して，述語と項の意味的な関係を整理して公開している述語項構造シソーラス (PT) の概念フレームと意味役割を付与した NPCMJ-PT を構築している．現段階で約 7.4 万述語に対して概念フレームを付与し，約 14.9 万件の項に対して意味役割関係を付与している．本稿では現段階で付与したデータについての特徴と今後の展望について記述する．

1　はじめに

文の述語を中心とした係り関係の構造に対して，意味的な関係を付与する意味役割付与データの構築が英語を中心に行われている (例えば PropBank[1] や FrameNet[2])．さらに，意味役割と概念フレームを元にした文の抽象的な意味構造をグラフで表す AMR(Abstract Meaning Representation) が提案されており[1)]，AMR parser の開発 [3][2)] や AMR によるバイオ文献における情報抽出が研究されている [4].

近年では，意味役割ではなく，係り関係のタイプを付与する UD (Universal Dependency) が提案されており，主語と目的語ならば nsubj や obj とした関係で取り出すことが出来る．一方で，意味役割と異なるため下記のような open に対して開けた動作主なのか開けられた対象なのかの違いは処理されない[3)]．

(1)　[$_{nsubj}$ He] open the [$_{obj}$ door] .

(2)　The [$_{nsubj}$ door] opened .

一方，PropBank 形式の意味役割と概念フレームでは下記のように動作主と対象を分ける．

(3)　[$_{Arg0}$ He] [$_{open.01}$ open] [the $_{Arg1}$ door] .

(4)　[The $_{Arg1}$ door] [$_{open.01}$ opened] .

よって，単に意味的な関係だけでなく，辞書 (概念フレーム (この場合 open.01)) と同時に開発するのが意味役割付与の特徴である．

本研究では PropBank 形式の意味役割と名前の意味役割を日本語の述語 (動詞，サ変名詞，形容詞，形容動詞の約 1.2 万語) に対して 2.4 万例文を付与した述語項構造シソーラス PT (Predicate-argument Thesaurus) を辞書として NPCMJ に意味役割を付与して公開している[4)]．概念フレームは現段階で 1097 件定義しており，NPCMJ の付与とともに拡張している．

一方，NPCMJ は言語学研究者および日本語学習者を意識した Web 上で検索可能な大規模ツリーバンクである (約 6.7 万ツリーを公開)．本研究では意味役割と概念フレームを付与することで，事例検索に役立てるとともに自然言語処理における日本語意味役割付与の学習データとしての利用を目標としている．

2　現在付与している NPCMJ-PT の全体像

NPCMJ-PT は NPCMJ の構文木に従って述語に対して述語項構造シソーラスの概念フレームを付与して，項に対して意味役割を付与したデータである．図 1 に事例を示す．述語とそれに対応する項についてはあらかじめ NPCMJ の構文木から計算によって導出されている [5]．構文木データを基にしているためゼロ代名詞[5)]

1）　https://catalog.ldc.upenn.edu/LDC2020T02
2）　その他ツール https://github.com/IBM/transition-amr-parser
3）　Stanford Parser の出力例を参考に記述した．
4）　http://pth.cl.cs.okayama-u.ac.jp/
5）　http://www.ls-japan.org/modules/documents/index.php?content_id=309

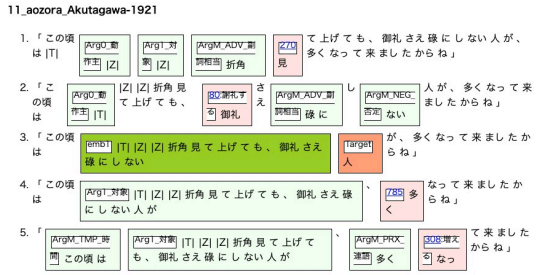

図 1　NPCMJ-PT の付与例 (aozora_Akutagawa-1921 コーパス)

を含めて項が設定されているところが特徴的である．こうした点は Penn Treebank に付与を行っている PropBank と同様の形式である．

現段階[6]の付与データの量を表 1 に示す．現

表 1　NPCMJ-PT 全体の現段階の付与件数

項目	件数
付与した述語の数 (概念フレーム数)	73,849
付与した文	38,774
付与した項の数	143,671

段階で NPCMJ-PT は内部的に 9.5 万件の構文木 (=文) を有している[7]．図 1 で示したとおり，1 文に複数の述語が登録されているため，1 文で複数の概念フレームを付与する場合が多い．また，NPCMJ-PT には PT の例文も構文木とともに記録されている．PT の事例は 2.2 万件程度であるため，PT の付与例を抜いた場合，概念フレーム数は 51,480 件，項の数は 103,033 件である．現在も付与を継続している．

以下の節では具体的にどのような述語に対して，どのような概念フレームや意味役割が付与されているのか具体例を挙げて記述する．

3　述語と概念フレームの付与

表 2 に付与した述語の種類数，概念フレームの種類数を示す．NPCMJ-PT では述語は活用した語のみの記録であるため，基本形は形態素解析 MeCab を利用して推定した基本形を利用する．表 2 に示すように述語の概念フレームには PT だけでなく一部日本語 WordNet の概念フレームを付与している．これは図 1 にもあるように述語として名詞も含まれており，一部の名

6)　2022 年 1 月 12 日現在.
7)　2022 年 1 月 12 日時点では 6.7 万木 (文) が公開されている (https://npcmj.ninjal.ac.jp/).

表 2　付与した述語の上位 10 件の頻度 (基本形に整形)

項目	件数
述語の付与数 (述べ数)	73,849
述語の種類数	12,672
上記の内 WordNet の概念付与数	1181
概念フレームの種類 (PT 場合)	1020
概念フレームの種類 (WordNet)	505

詞に対して実験的に日本語の WordNet を利用して概念を付与している[8]．

概念フレームの付与の仕方であるが図 1 にもあるように，PT の概念フレームに述語の登録がない場合，近い概念フレームを指定して，その概念フレームの中で最も似ている述語を指定して付与している．例えば図 1 の 2 例目の「御礼」は PT には登録がないが，Frame ID 80 番 (提供) の概念フレームにおける「謝礼する」に近いことを示している．このようにアノテーション作業で概念フレームを利用しつつ辞書の拡張を行って概念を付与している．

PT の概念フレームは現段階で 1097 種類定義している．PT のデータが入っているにもかかわらず全概念フレームの種類が入っていない．これは PT の概念フレームが更新されていることや，付与の形式ミスなどで今回のデータには入れなかった事例が存在することが考えられる．

また表 2 には現れないが，図 1 に示すように複合動詞についても付与している．NPCMJ-PT では最後の述語に概念フレームを付与することから，最初の要素には ArgM-PRX を付与し「多くなる」の意味に近い概念フレーム Frame ID 308(程度の変化/多い) を付与する．

次に，表 3 に付与した述語のうち頻度の高いものを順に上位から並べたものを示す．述語は上記のように形態素解析 MeCab を利用して推定した基本形を利用してまとめている．

表 3　上位 10 件の述語 (MeCab による推定)

述語	だ	の	する	ある	なる
頻度	2263	1832	1758	1686	1625
述語	です	に	*	と	思う
頻度	1055	918	754	556	479

8)　サ変名詞は PT に登録されているため，ほとんど PT の概念フレームが付与されている．

表 3 で最も多い述語の「だ」と「の」の例を下記 (5)(6) に示す．どちらもコピュラであることがわかる．

(5) [$_{Arg1}$ 私は] [$_{Arg2}$ 遠藤という書生] [$_{895:です}$だ] (aozora_Akutagawa-1921)
(6) [$_{Arg2}$ インド人] [$_{895:です}$ の] [$_{Arg1}$ 婆さん] (aozora_Akutagawa-1921)

NPCMJ では「の」が「である」に言い換えられる場合は述語としており，連体修飾の場合などによく出現する．(6) ではこれを「婆さんは/Arg1 インド人/Arg2 です/895」という平叙文を基に付与している．また「*」は MeCab に登録がなかった述語であり，例えば下記の (7) のような動詞があげられる．

(7) [$_{Arg0}$ 亜米利加人は] [$_{Arg2}$ 煙草を] [$_{47}$ 喞え] たなり，(aozora_Akutagawa-1921)

「と」という述語は下記の (8) に示すような場合であり，「である」という述語があるとして概念フレームを付与している．

(8) [$_{Arg1}$ この案件は] [$_{Arg2}$ 調査のみ] [$_{895:です}$と] し，(book_excerpt-30)

NPCMJ の構文木に基づくと，コピュラが「です」「だ」「の」「と」といった異なる形をとって多く出現していることがわかる．

ここで，付録の表 7 に概念フレームの多いものから順に 10 件の例を取り上げる．上述のようにコピュラが最も多く全体の 1 割ほど出現している．続いて，移動 (17)，生成 (124)，伝達 (95)，変更 (407) といった状態変化動詞が続いている[9]．また 9 番目には，「思う」「認識する」など人の判断にかかわる認識 (101) の動詞が出現している．

4 意味役割の付与

NPCMJ-PT における意味役割について記述する．まず付与している意味役割の枠組を記述した後，付与した結果について記述する．

4.1 意味役割の基本的な枠組

基本的に PT の枠組に従った形式である．PT では意味役割として PropBank 形式の意味役割を採用し，必須項が Arg0,1,2 といった数字を利用する．ただし，PropBank では各述語の語義についてそれぞれ別々に Arg0,1,2 の意味的な割り当てを決める (例えば Open.01 など) が，PT では FrameNet[2] と同様に概念フレーム [6] を想定して，概念フレームにおける必須の意味的な関係を Arg0,1,2 で表現する．つまり FrameNet との対応で記述すると Element を数字で表現していることに対応する．概念フレームにおける必須項はイメージとしてその概念で出てくる必須の登場要素を決めて分類したものである．これにより態の異なりなどで表現が変わっても述語に対する項の意味的な関係を指定することができる．

一方で PT では名前の意味役割も同時に付与している．名前の意味役割は文献 [7] の事例にもあるように「〜で」で場所なのか手段なのかといった表現をベースにした解釈で付与する．これにより，例えば，下記の壁塗り構文の場合，PropBank 形式の意味役割ではどちらも違いがないが，名前の意味役割によって，「ペンキで」が手段として捉えている表現である場合で，違いを検索することが可能になる．

(9) [$_{Arg1:対象}$ ペンキを] [$_{Arg2:着点}$ 壁に] [$_{25}$ 塗る]
(10) [$_{Arg1:手段}$ ペンキで] [$_{Arg2:対象}$ 壁を] [$_{25}$ 塗る]

Frame ID 25(覆う) の必須項は「Arg0:動作主」，「Arg1:覆うもの」，「Arg2:覆われるもの」の 3 つである．よって PropBank 形式の意味役割と名前の意味役割は文の表現に応じて様々な対応関係をとる．

4.2 意味役割の付与内容

表 4 に NPCMJ-PT で付与した意味役割 (PropBank 形式) の上位 10 件の頻度分布を示す[10]．表 4 から最も多いのは対象を示す Arg1 であり次いで動作主を表す Arg0，さらに必須項の Arg2 が続く．付加詞 (つまり各概念特有の項ではなく，どんな述語にも付与する係り元) では ArgM-ADV(副詞) や ArgM-TMP(時間)，ArgM-LOC(場所) が多いことがわかる．また，NPCMJ では ArgM-PRX(連語:述語の一部の要素) が多く出現している．下記に ArgM-PRX の事例を示す．

9） 丸括弧内は Frame ID を示す．

10） PropBank 形式の意味役割の一覧は付録の表 6 に示す．

表 4 上位 10 件の意味役割 (PropBank 形式)

意味役割	頻度
Arg1	58579
Arg0	38800
Arg2	23611
ArgM-ADV	5696
ArgM-TMP	3683
ArgM-LOC	2004
ArgM-MNR	1613
Arg3	1485
ArgM-PRX	1369
ArgM-CAU	1095

(11) [$_{ArgM-ADV}$ もう] [$_{ArgM-PRX}$ どうでも] [$_{927:無関心だ}$ いい] (aozora_Dazai-2-1940)
(12) [$_{ArgM-PRX}$ 気ガ] [$_{210:気絶する}$ 遠クナッ] テシマウ (aozora_Akutagawa-1921)

事例 (12) に示すように複合動詞だけではなく，共起語や慣用表現の要素も ArgM-PRX タグを付与する．

表 5 に名前の意味役割の上位 10 件の頻度分布を示す．PropBank 形式の意味役割と同様に，最も頻度が多い要素が「対象」で次に「動作主」となっている．以降の意味的な関係として，経験者や補語相当（は）といった必須項に関するものが出現している．表 4 と比較すると，Arg1

表 5 上位 10 件の意味役割 (名前の意味役割)

意味役割	頻度
対象	48,928
動作主	28,551
経験者	10,749
補語相当（は）	7,615
副詞相当	5,696
着点	3,684
時間	3,684
場所	3,211
対象（動作）	2,723
対象（生成物）	2,566

よりも「対象」の頻度が少ない．つまり Arg1 は名前の意味役割では他のラベルが付与されていることがわかる．名前の意味役割は全部で 73 種類設定している．よって PropBank 形式では Arg0 であっても，「動作主」や「原因」，「経験者」など細分化されている事例が付与されている．また Arg1 も単なる対象だけではなく，「経験者」や「補語相当」，「着点」などに付与されている[11]．

下記に「対象」および「補語相当（は）」の例を示す．

(13) [$_{Arg1:対象}$ おれの声は] [$_{Arg2:補語相当（は）}$ 天上に燃える炎の声] [$_{895:です}$ だ] (aozora_Akutagawa-1921)
(14) [$_{Arg2:補語相当（は）}$ 一人] [$_{895:です}$ の] [$_{Arg1:対象}$ 男] (aozora_Doyle-1905)

基本的には「A は B だ」のコピュラの場合がよく見受けられる．例 (14) のように連体修飾の形で「の」を述語と考えた場合にも付与されていることが多い．

5 今後の予定

現在も概念フレームと意味役割付与作業を続けているため，本稿で記述した統計量は変わる予定である．また意味役割を付与した結果をどのようなデータ構造にするか現在議論中である．先の文献 [9] で記述したように FrameNet 準拠の XML 形式の出力は可能であるが一方で，構文木の情報は同時に盛り込まれない．また，本原稿を書くにあたってプログラムを作成して NPCMJ-PT データを xlsx 形式にまとめたものを利用した．表計算ソフトの形式を利用すると集約した情報を得る場合に有利である．どのような形式で NPCMJ-PT を出力するか引き続き検討する．

謝辞

本研究は国立国語研究所機関拠点型基幹研究プロジェクト「統語・意味解析コーパスの開発と言語研究」および JSPS 科研費（課題番号 15H03210）と (課題番号 19K00552) の助成を受けたものである．

11） 「経験者」は Arg0 の場合も Arg1 の場合の付与もある．基本的には意図せずになにか行為を行う主体である．PropBank1.7 の辞書では here.01 や see.01 の Arg0 に Experiencer という注釈が入っている．動作主体でありながら，「経験者」の意味で付与しているように見受けられる．「経験者」は Arg1 にも対応する．例えば「[$_{Arg1:経験者}$ 赤ちゃんが] [$_{Arg0:原因}$ 物音に] [$_{398}$ 驚く]」のように心理的な影響を受ける主体に「経験者」を付与している [8].

参考文献

[1] Olga Babko-Malaya. *PropBank Annotation Guidelines*, 2005.

[2] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 86–90, 1998.

[3] Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. Amr parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[4] Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed S. Elsayed, Skatje Myers, and Martha Palmer. Fine-grained information extraction from biomedical literature based on knowledge-enriched abstract meaning representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 6261–6270, 2019.

[5] Alastair Butler. Treebank Semantics. Technical report, Hirosaki University, 2021. (`https://entrees.github.io/index.html` accessed 2022/1/12).

[6] Adele E. Goldberg. *Constructions*. The University of Chicago Press, 1995.

[7] 日本語記述文法研究会. 現代日本語文法 2 第 3 部格と構文 第 4 部ヴォイス. くろしお出版, 2009.

[8] 竹内孔一, バトラーアラステア, 長崎郁, パルデシプラシャント. PropBank 形式を考慮した NPCMJ に対する意味役割付与〜態の違いと経験者の付与〜. 言語処理学会 第 26 回年次大会発表論文集, pp. 633–636,, 2020.

[9] 竹内孔一, アラステアバトラー, 長崎郁, プラシャントパルデシ. Npcmj への propbank 形式の意味役割と概念フレームの付与の進捗報告. 言語処理学会第 27 回年次大会発表論文集,E8-4, 2021.

A　意味役割ラベルの一覧

NPCMJ-PT を構築するにあたり利用している意味役割を示す．表 6 に示すように PropBank 形式の意味役割は 28 種類である．

表 6　PropBank 形式の意味役割の一覧

タグ	説明
Arg0	必須項 (動作主)
Arg1	必須項 (対象)
Arg2	必須項
Arg3	必須項
Arg4	必須項
Arg5	必須項
ArgA	使役態で追加される使役者
ArgE	受動態で追加される経験者
ArgM-ADV	副詞
ArgM-AND	順接
ArgM-BUT	逆接
ArgM-CAU	原因
ArgM-CMP	補語相当
ArgM-CND	条件
ArgM-CRT	基準
ArgM-DIR	方向
ArgM-EXT	程度
ArgM-LOC	場所
ArgM-MDF	修飾
ArgM-MNR	様態
ArgM-MNS	手段
ArgM-NEG	否定
ArgM-PRP	目的
ArgM-PRX	連語
ArgM-REC	相互
ArgM-SCP	領域
ArgM-SPK	話者
ArgM-TMP	時間

B　上位 10 件の概念フレーム

付与した概念フレームについて上位 10 件の頻度分布を表 7 に示す

概念フレームは最大 5 階層のシソーラス構造になっている．記号の'_' が階層の境界を表している．概念フレーム内の括弧内の数字は Frame ID を表している．詳細な事例は Web 上で確認できる (http://pth.cl.cs.okayama-u.ac.jp/).

表 7　上位 10 件の概念フレーム

順位	概念フレーム	頻度
1	状態変化なし（状態）_コピュラ _コピュラ (895)	7748
2	状態変化あり_位置変化_位置変化（物理）_着点への移動 (17)	2663
3	状態変化あり_生成・消滅_生成（物理）_生成 (124)	2634
4	状態変化あり_位置変化_位置変化（情報）（人物間）_他者への伝達 _伝達 (95)	2332
5	状態変化あり_変更_変更_変更 (407)	1938
6	状態変化あり_位置変化_位置関係の変化（物理）_絡まる・ほどく _絡まる (29)	1589
7	状態変化なし（状態）_位置_存在 (529)	1289
8	状態変化あり_主体の変化（判断・認識の変化）_判断（認識）_認識・判断 (101)	921
9	状態変化なし（活動）_実行_実行 _実行 _行う (251)	850
10	状態変化あり_対象の変化（主体の判断に伴う変化）_判断（認定）_決定 (333)	761

表 7 内の頻度を見ると特徴的であることがわかる．本文でも述べたようにコピュラが他の述語の概念と比較して突出して大きいことがわかる．一方で,「着点への移動」から「伝達」まで大きく頻度に差が見られない．これは資料をみると人やものの移動 (「行く」「来る」「帰る」) という表現が青空文庫だけでなくニュースにも見られており，記述する内容として取り上げられることがわかる．また「生成」もどうようになにか発生したり，作るといった表現がどのジャンルにも表現している．こうした表現が単なる存在 (529) よりも文書で表現されることが多いことがわかる．