名詞化された事態表現への意味的注釈付け*

飯田 龍　小町 守　乾 健太郎　松本 裕治
奈良先端科学技術大学院大学
{ryu-i,mamoru-k,inui,matsu}@is.naist.jp

1 はじめに

近年，述語項構造のような意味に一歩踏み込んだ構造の自動解析技術が盛んに研究されている [2, etc.]．この課題では，動詞のような述語とその項に対し，PropBank[10] などで定義された役割（ARG1, ARG2 など）を自動的に付与することを目的とする．この述語項構造を高精度で解析することで，情報抽出などの応用分野に広く貢献できると考えられる．

PropBank に代表される述語項構造解析タスクでは，動詞などの用言を対象に問題が設計されているが，動詞派生名詞やサ変名詞などの名詞（以後，**事態性名詞**）についても述語と同様に，項同定の問題が設計され [9, 3, 12, 15]，実際にそれらの問題への取り組みも報告されている [5, 6, 7]．例えば，我々が新聞記事約 3 千記事（約 4 万文）へ事態とその項を人手タグ付与した結果（NAIST テキストコーパス）[15] には，延べ 106,628 の述語がタグ付与されたのに対し，事態性名詞にタグ付与された個数は 28,569 と，文書全体に記述される事態表現の約 2 割が名詞によって表現されていることがわかった．この数値からも述語だけではなく事態性名詞も対象に項構造解析を行うことで，文章中に記述された事態を網羅的に扱うことができ，情報抽出などの事態を扱う応用処理に直接的に貢献できると考えられる．

事態性名詞を対象に項を付与する場合には，用言の場合とは異なり，名詞が事態（**コト**）を指すか実体（**モノ**）を指すかの判断が必要となる．例えば，例 (1) の “電話$_i$” は “彼ガ私ニ電話スル” という事態を表すのに対し，“電話$_j$” はモノとしての電話を指すという違いを区別する必要がある．

(1) **彼**$_a$ からの 電話$_{i(ガ:a,\ ニ:b)}$ によると、**私**$_b$ は彼の家に 電話$_j$ を忘れたらしい。

これまでの我々のタグ付与作業では，コーパス中の個々の事態性名詞の出現について，(i) モノを指しているかとコトを指しているかを判別し，(ii) コトの場合のみ項を付与する，という前提のもとで作業を進めてきたが，実際はこの 2 つの間にはずれがあり，このために述語の場合と比較して人手作業の品質が低下していることがわかった [15]．そこで，本稿では事態性名詞の項付与に関して問題となる事例を紹介し，その対処方法を論じる．以下，2 節で関連研究と我々の取り組みを比較し，3 節でこれまでに策定してきた事態性名詞へのタグ付与の仕様 [15] を概観，またこの仕様に基づいて作業を行った場合に起こる典型的な作業の揺れを示す．4 節でそれらの揺れに対しどう対処するかを説明し，5 節でその変更によってタグ付与作業の品質に影響が出るかを調査した結果をまとめる．最後に 6 節でまとめる．

2 関連研究

我々の取り組み [15] とは独立に，いくつかの言語で事態性名詞の項構造解析のためのデータ作成の試みが報告されている [12, 3, 9, 1]．例えば，Meyers らが作成している NomBank[9] では，英語を対象に事態性名詞や関係名詞に対して項を PropBank と同様の形式（ARG1, ARG2 など）で付与している．タグ付与作業では，あらかじめタグ付与対象となる名詞がどのような語義を持ち，各語義がどのような項を持つかを列挙した辞書を作成し，作業の際にはその情報を参照しながらアノテーションを行うが，名詞句にタグを付与する際には以下の 3 つの条件のうちどれか 1 つを満たす必要がある [8]．

1. 対象となる名詞句が少なくとも 1 つの項をその名詞句内に含む．
2. 名詞句の主辞が事態や状態などを表わす名詞であり，かつその名詞を修飾する語を少なくとも 1 つ含む．
3. 機能動詞結合のような構成で項が出現している．

この条件を見てわかるように，NomBank では統語的に制約された条件に基づいて項を付与できるか判断しており，項が文を越えて出現している場合は付与対象としていない．また，裸名詞（bare noun）は項を取るか否かの判断が困難なため付与の対象外となっている．

また，FrameNet[1] でもフレームを喚起（evoke）する表現の一種として名詞を考えており，NomBank と同様に機能動詞結合や there 構文などの文や句の構成を手がかりとして作業を行っている．また，複合語を考える際には，主辞が事態性名詞の場合には項（FrameNet ではフレーム要素に相当）を付与するが，逆に事態性名詞が主辞以外の位置に出現している場合はタグ付与の対象とはしない [11]．このため，例えば，複合名詞 “旅行/客” で “旅行” は主辞の位置に出現していないため，“客ガ旅行スル” という述語-項関係は付与しないことになる．

上述の 2 つの関連研究では主に統語的な手がかりを伴っており，かつ局所的に項が出現する場合を対象にアノテーションを行っている．しかし，情報抽出などの応用処理を考えた場合，項が述語や事態性名詞の近傍に出現するとは限らず，事態性名詞から離れた位置に出現する項も頑健に同定できることが重要な要素技術となる．

*Semantic Annotation of Nominalized Event Mentions
Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto
Nara Institute of Science and Technology

例えば，NAIST テキストコーパス [15] 内の事態性名詞と項の出現分布を見ると，事態性名詞と項（ガ格のみ）が異なる文に出現している割合は全体の 32%，また同一文内に出現しているが近傍に出現していない割合は全体の 18%であり，事態性名詞のガ格の 5 割は近傍を見るだけでは項を同定することができないことがわかる．情報検索のような応用処理を考えた場合，これらの離れて出現する項についてもできるだけ整合性の取れる形で仕様を決める必要がある．

また，石井 [13] は日本語の学術用語集の用語を対象に，和語動詞派生の名詞（例えば，“送出し 効率”，“コンクリート 打ち” など）とその他の構成素がどのような関係にあるかを，“主体” や “結果対象”，“目的” という意味役割の粒度で分析している．こちらの試みでも複合語の中の構成素のみを対象にしており，それ以外の項の出現には言及されていない．

3 事態性名詞への項付与

事態性名詞への項付与について，NAIST テキストコーパス (ver. 1.4β) の作業基準を以下に列挙し，この基準で揺れが生じる典型例を示す．

3.1 項付与のための作業基準

文献 [15] にまとめた作業基準のうち，事態性名詞の項付与に関する項目を以下にまとめる．

項を取る事態性名詞のみを対象にする 事態を表す名詞には “探索” のようなサ変名詞や “笑い” のような和語動詞派生の名詞があり，これらの名詞は “サ変名詞+スル” や派生前の動詞（例えば，“笑い” なら “笑う”）を考えることでどのような項を取るかを作業者が想起できる．一方，“運動会” や “雨” などの名詞も事態を指すと考えられるが，項をどのように定めるかが明確ではないため，今回はサ変名詞や和語動詞派生の名詞のみを対象に作業を進める．

モノとコトを明示的に分けて作業する 事態性名詞は文章中でコトを指しているかそれ以外かを分ける必要がある．例えば，1 節の例 (1) に出現している 2 つの事態性名詞 “電話” のうち，“電話 $_i$” が “電話スル” というコトを表しているのに対し，“電話 $_j$” は “（携帯）電話” というモノを表している．この状況で作業者は “電話 $_i$” のみを項を取ると判定し，これに対して “彼 $_a$” をガ格，“私 $_b$” をニ格として付与する．この分類作業によって作成された事例集合が機械がコトかモノかを自動判別するための有用な学習事例となる．

項付与の粒度 事態性名詞と項となる名詞（句）との関係には Agent，Theme のような抽象度の高い関係や，PropBank[10] で用いられている ARG1，ARG2 のような関係などさまざまな粒度が考えられるが，我々は必須格となる表層ガ/ヲ/ニ格の付与を行っている[1]．項が文章内に出現していない場合は，外界照応の関係とみなし “一人称”，“二人称”，“その他” の粒度で関係を付与する．つまり，必須ガ/ヲ/ニ格には漏れなく項が付与されることになる．

[1]粒度選択の理由は文献 [15] を参照．

格パタンの曖昧性 項はガ/ヲ/ニ格などの表層格で付与するため，例えば，自他交替が可能な事態性名詞（例えば，“実現”）が単体で出現した場合，“X ガ Y ヲ実現スル” と “Y ガ実現スル” のどちらの格パタンで付与するかの曖昧性が生じる．そこで，作業時にはできるだけ多く項を付与できる格パタンを選択し，その格パタンで必須となる格要素を付与する．

3.2 作業の揺れの典型例

3.1 でまとめた基準にしたがって作業を行った結果，ある程度の作業者間の一致率を得たが，

1. 事態か否かを明示的に分類しなければならない
2. 述語の場合とは異なり，格要素が格標識（格助詞）を持たない

という 2 つの問題のため，述語の場合に比べ作業の一致率は良いとは言えない．特に 1 つ目の問題に関して，モノとコトの 2 値に分類することが困難な事例が多数出現し，文献 [15] に示した揺れの主な原因となった．例えば，例 (2) の名詞 “報告” は “文化庁ガ報告スル” というコトを表していると同時に報告された結果（内容物）というモノを表していることになり，この例からもわかるように，いくつかの事態性名詞は解釈によってモノとコトのどちらとも判断できる場合がある．

(2) **文化庁**の 2005 年の 報告 によると、各宗教団体の報告による信者数は合計 2 億 1100 万人である。

つまり，項となり得る表現（例 (2) では “文化庁”）が近傍に出現しているか否かがコトを指すか否かの判定におおきく影響し，上のような場合でも項となる表現が近くに出現していない場合はモノと判断されるなど，一貫した作業結果が得られていない．

4 モノとコトの重複解釈を許す注釈付与仕様

3.2 の例 (2) に示した作業の揺れが起こる例では，事態性名詞 “報告” が文脈から項を付与できるという判断とモノとしての解釈もあり得るという判断が独立に行われるため，「あらかじめ明示的にコトかモノかを区別し，コトへのみ項を付与する」という作業の前提とは対応しない問題が起こる．この点に着目して議論した結果，以下に示すの 2 つの提案をタグ付与の仕様として採用した．

提案 1: モノを指す表現へも項を付与する モノとコトの境界を項付与できるか否かで弁別することは困難であり，3.2 の例 (2) のようにモノとして解釈できる場合にも項を付与可能である．そこで，今回の仕様ではモノである場合でも項を持つと判断できた場合には，モノ/コトの判別とは独立に項を付与する．

提案 2: モノとコトを指す表現を区別するためモノと判断した根拠もタグ付与する 提案 1 で述べた仕様を採用すると，モノの場合も項を付与するため項を付与したことがコトを指すという情報と等価ではなくなる．しかし，文章中の事態のみを抽出したい応用分野も存在するため，作業結果には事態性名詞がコトを指すという情報もできる限り残しておくことが望ましい．そこで，まず我々はあらかじめ揺れが起こった事態性名詞を人手分析し，モノとコトの解釈で曖昧性が生じる名詞を ⟨ 結果物/

内容〉，〈モノ〉，〈役割〉，〈ずれ〉の 4 種のクラスに分類した．事態性名詞をモノとして解釈できる場合には，これら 4 つのうちいずれかをタグ付与することでモノとしての証拠を残す．逆にこれらのタグが付与されないことでコトを指す事態性名詞を表現する．また，名詞クラスのタグを用意することで，項付与が困難な場合に作業者が無理に項を付与しようとする事態を回避することができる．

上述の 2 つの提案を採用することにより，これまでモノに分類するか項を取るかどちらか一方の情報しか付与できなかった例 (2) の “報告” についても，“モノ（具体物）” としての解釈と “文化庁ガ報告スル” という項構造の両方を情報を付与できるようになる．

以下で，今回付与する 4 種の名詞クラスについて説明する．

結果物/内容 典型的には例 (3) のような内容節をとる（トノ，トイウを伴って出現する）場合，“意見” のような名詞は内容節が表す内容と同格であり，“意見スル” というコトを指すとは考えにくい[2]．

(3) 党内には「社会党会派の離脱者は従来通り除名すべきだ」との 意見 が根強く...

この類例としては “提案”，“決定”，“報告” などがある．

また，“連合シタ” 結果，“連合” という実体が存在するという解釈に基づき，例 (4) のような実体を指すのみで事態を表すとは考えにくい “連合” についても〈結果物/内容〉のタグを付与し，項は付与しない．この類例としては “組織” などの表現がある．

(4) 十日夜には、自由 連合 の新年会に自民党から森喜朗幹事長、島村宜伸国対委員長らが出席した。

同様に，例 (5) の “規制” も “規制スル” 事態よりも規制の内容そのものを指すと判断した場合は，項は付与せず〈結果物/内容〉タグのみを付与する．

(5) また、経済問題については日本経済の構造変革のため 規制 緩和に積極的に取り組むと訴える。

モノ（具体物） 事態性名詞が文脈中でモノ（具体物）を指しているかを判定する．前述の例 (1) のモノとしての “電話 $_j$” や場所としての “施設”，道具としての “装備” などの表現がこれに相当する．例えば，ある文脈で “携帯” という表現が “携帯電話” というモノを指す場合は〈モノ〉タグを付与する．

役割 “課長 補佐”，“松本 教授”，“オシム 監督” などの表現は，名詞句全体で個人を指示しており，名詞句内の事態性名詞がコトを指すとは考えにくく，この場合には〈役割〉ラベルを付与することで項の付与を回避する．このような表現は典型的に主辞の位置に出現している場合が多い．

述語と事態性名詞との語義のずれ 事態性名詞が派生前の動詞の意味と異なる場合には，項を付与することができない．例えば，例 (6) のサ変名詞 “一定” は動詞 “一定スル” と異なった意味で用いられており，このような場合には〈ずれ〉タグを付与し，項は付与しない．

(6) 一定 の得票で議席を占めた後に今回と同様「除名」などの騒動が起きれば、...

表 1: タグ付与の作業結果

	項を取る	モノ	結果物/内容	ずれ	役割
作業者 1	558	0	111	35	16
作業者 2	582	7	196	27	9
一致	531	0	92	10	5
一致 / 作業者 1	0.95	0	0.83	0.29	0.31
一致 / 作業者 2	0.91	0	0.47	0.37	0.56

表 2: 項タグ付与の一致率

	ガ格	ヲ格	ニ格
作業者 1	528	317	82
作業者 2	567	219	54
一致	415	182	44
一致/作業者 1	0.786	0.574	0.537
一致/作業者 2	0.732	0.831	0.815

5 人手タグ付与の評価

4 節にまとめた作業方法を採用することで人手でのタグ付与品質にどのような影響が出るかの調査を行った．具体的には，作業者 2 人が新聞報道 50 記事中のサ変名詞 665 箇所に対し，その名詞が項を持つか否かの判定と項を持つ場合は項の付与を行った．この作業とは独立に 4 節に示した名詞クラスの付与を行った．今回の作業では頻出するサ変名詞を対象に作業し，和語動詞派生の名詞は対象外とした．作業者 2 名の作業結果とその一致率を表 1 に示す．表 1 より，665 件のサ変名詞のうちどちらの作業者も 550 を越えるサ変名詞に対して項を持つと判定しており，文章中のほとんどのサ変名詞は項付与対象となっていることがわかる．また，項を持つか否かの作業者間の一致率はそれぞれの作業者について見た場合 0.95 と 0.91 と以前の作業品質の調査 [15]（一致率は 0.905 と 0.810）と比較して一致率が向上しており，今回の作業方針が品質向上に有効であったことがわかる．また，項を持つか否か Kappa 値で評価したところ 0.522 という結果を得た．名詞クラスの一致率については良いとは言い難いが，これは作業者間で〈結果物/内容〉と〈ずれ〉にそれぞれ付与するなど，クラス間の揺れが生じたためであり，またそもそもスル接続で表現する頻度が低い “確証” などの事態性名詞に関する解釈の異なりも作業の揺れの原因となった．また，項を取るか否かのタグ付与が不一致だった 78 事例を調べたところ，44 事例は項を付与するか否かに作業者間で解釈が異なる事例であり，残りは付与の誤りとみなせる事例であった．

次に，2 人の作業者が項を持つと判断した 531 事例について，項（ガ/ヲ/ニ格）がどのくらい一致するかを評価した結果を表 2 に示す．項を取ると判断された事態性名詞のうち付与された項が一致しなかった 265 事例を人手で分析，揺れの原因を調査した結果を表 3 にまとめる[3]．作業の揺れはおおきく 2 つの問題に起因している．一つは，各事態性名詞を述語化して考えた際に作業者間で異なった格パタンを想起したためである．特に，ある事態性名詞に対して，一方の作業者は必須格としてヲ格

[2] もちろん “意見” という表現だからといって必ず〈結果物/内容〉タグを付与するわけではなく，文脈から “X ガ Y ト意見スル” という事態と判断できる場合は項を付与し，〈結果物/内容〉タグはしない．

[3] 1 事例を複数の誤りの原因に割り割り振ったため，合計は 265 事例より多くなる．

表 3: タグ付与不一致の原因分析の結果

揺れの原因		頻度
格パタンの認識誤り	(ヲ格が不足)	88
	(ニ格が不足)	25
外界 or 文章内談話要素	(ガ格)	59
	(ヲ格)	6
	(ニ格)	1
タグ付与誤り		52
部分-全体の関係		17
正解		10
その他		13

を取ると判断したが，他方はそれを取らないと判断した場合が揺れの大部分を占めていることがわかった．この問題に関しては，語彙概念構造 [4] を考慮して作成されている動詞辞書 [14] のような情報を作業の際に提示することで，作業者が想起できない格パタンを網羅的に把握することができ，揺れが少なくなると考えられる．このような作業者支援については開発中のアノテーションツール [16] でどのように情報を提示するかという問題と同時に考えていきたい．

また，もう一つの主要な揺れの原因は，同定すべき項の粒度に関するものである．例えば，例 (7) で項を持つと判断された “整備” には，前方文脈に出現している “日本鉄道建設公団” という組織が “整備スル” という解釈と，文章中に出現しない “特定の誰か（もしくは集団）” が “整備スル” という 2 つの解釈が存在する．

(7) **日本鉄道建設公団**は十一日、整備新幹線の北海道新幹線について、ルート公表に向けた函館市と小樽市付近の調査に一月下旬から着手すると発表した。調査は、整備 新幹線建設費とは別枠の、建設推進準備事業費三十億円の中で行われる。

この問題は「できるだけ文章内から項を選択する」という基準を用いた場合でも，作業結果は作業者の解釈に左右されるため，解決はできず，述語の場合も同様に問題となる．この問題と関連して，これまでのタグ付きコーパス構築の方法論はできるだけ揺れを無くすことが前提であり，解析はその厳密な設定のもと問題を解くという立場で研究が進められてきたが，今後はその代替案として作業者一人もしくは複数人の揺れを許容するような学習・分類の枠組みを検討すべきかもしれない．

6 おわりに

本稿では，日本語の名詞化された事態表現に対する意味情報付与，特に述語の項構造関係に相当する関係の付与について議論を行った．事態表現となり得る候補としてサ変名詞を取り上げ，それらに対し 4 つの名詞クラスと，項構造を独立に付与することで，名詞の項付与について起こる曖昧性に柔軟に対処する仕様を設計した．仕様の信頼性を評価するために 2 人の作業者で作業の一致率を調査し，以前の一致率の評価 [15] と比較してタグ付与の一致率が向上したことを報告した．今回策定した事態性名詞に関する仕様に基づき，次版の NAIST テキストコーパスでは項情報と名詞クラスの情報を付与して公開する予定である．

今後の課題として，今回設計した仕様に基づき大規模な訓練事例を作成し自動解析の評価に用いることで，どのように解析品質に影響するかの調査を計画している．具体的には，人手付与された ⟨ 結果物/内容 ⟩ や ⟨ 役割 ⟩ などのラベル付き事例と大規模な生コーパスから抽出したモノ/コトの教師無し判別モデルを組み合わせることで，事態性名詞のモノ/コトの分類精度が向上するかを調査したい．

また，これまでの作業では主に新聞報道記事を対象に具体的な作業指針を検討してきたが，blog のような異なる記述スタイルの場合にも頑健に作業できるかを調査する必要があり，これについては特定領域研究「日本語コーパス」で公開予定の Web 文書に対してもタグ付与作業することで，タグ付与の仕様そのものの品質を吟味したい．

謝辞

本研究は科研費特定領域研究「代表制を有する大規模日本語書き言葉コーパスの構築」，ツール班「書き言葉コーパスの自動アノテーションの研究」（研究代表者: 松本裕治）の支援を受けた．記して謝意を表する．

参考文献

[1] Baker, C. F., Fillmore, C. J. and Lowe, J. B.: The Berkeley FrameNet project, *Proceedings of the ACL-COLING*, pp. 86–90 (1998).

[2] Gildea, D. and Jurafsky, D.: Automatic Labeling of Semantic Roles, Vol. 28, No. 3, pp. 245–288 (2002).

[3] Hasida, K.: GDA 日本語アノテーションマニュアル 草稿 第 0.74 版 (2005). http://i-content.org/gda/tagman.html.

[4] Jackendoff, R.: *Semantic Structures*, Current Studies in Linguistics 18, The MIT Press (1990).

[5] Jiang, Z. P. and Ng, H. T.: Semantic Role Labeling of NomBank: A Maximum Entropy Approach, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp. 138–145 (2006).

[6] Komachi, M., Iida, R., Inui, K. and Matsumoto, Y.: Learning Based Argument Structure Analysis of Event-nouns in Japanese, *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp. 120–128 (2007).

[7] Liu, C. and Ng, H. T.: Learning Predictive Structures for Semantic Role Labeling of NomBank, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 208–215 (2007).

[8] Meyers, A.: Annotation Guidelines for NomBank - Noun Argument Structure for PropBank 2007 (2007). http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf.

[9] Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. and Grishman, R.: The NomBank Project: An Interim Report, *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation* (2004).

[10] Palmer, M., Gildea, D. and Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106 (2005).

[11] Rupppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. and Scheffczyk, J.: FrameNet II: Extended Theory and Practice (2006). http://framenet.icsi.berkeley.edu/book/book.pdf.

[12] 河原大輔, 黒橋禎夫, 橋田浩一:「関係」タグ付きコーパスの作成, 言語処理学会第 8 回年次大会発表論文集, pp. 495–498 (2002).

[13] 石井正彦: 現代日本語の複合語形成論, ひつじ書房 (2007).

[14] 竹内孔一, 乾健太郎, 藤田篤: 語彙概念構造に基づく日本語動詞の統語・意味特性の記述, レキシコンフォーラム (影山太郎 (編)), No. 2, ひつじ書房, pp. 85–120 (2006).

[15] 飯田龍, 小町守, 乾健太郎, 松本裕治: NAIST テキストコーパス: 述語項構造と共参照関係のアノテーション, 情報処理学会研究報告 (自然言語処理研究会) NL-177-10, pp. 71–78 (2007).

[16] 野口正樹, 三好健太, 徳永健伸, 飯田龍, 小町守, 乾健太郎: 汎用アノテーションツール SLAT, 言語処理学会第 14 回年次大会発表論文集 (2008).