社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－177 (10)
2007／1／26

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

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 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
{ryu-i,mamoru-k,inui,matsu}@is.naist.jp

本稿では，日本語書き言葉を対象とした述語項構造と共参照のタグ付与について議論する．述語項構造や共参照解析は形態素・構文解析などの基盤技術と自然言語処理の応用分野とを繋ぐ重要な技術であり，これらの問題の主要な解析手法はタグ付与コーパスに基づく学習ベースの手法である．この手法で利用するための大規模な訓練データが必要となるが，これまでに日本語を対象にした大規模なタグ付きコーパスは存在しなかった．また，既存のコーパス作成に関する研究で採用されているタグ付与の基準は，言語の違いや我々が対象としたい解析と異なるために，そのまま採用することができない．そこで，既存のいくつかのタグ付与の仕様を比較し，我々のタグ付与作業で採用する基準について吟味する．また，実際に京都コーパス第3.0版の文章を対象にタグ付与の仕様について検討した結果とタグ付与の際に問題となった点や今後検討すべき点について報告する．

NAIST Text Corpus: Annotating Predicate-Argument and Coreference Relations

Ryu Iida　Mamoru Komachi　Kentaro Inui　Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma Nara 630-0192 Japan
{ryu-i,mamoru-k,inui,matsu}@is.naist.jp

In this paper, we discuss how to annotate predicate-argument and coreference relations in Japanese written text. Predicate argument analysis and coreference resolution are particularly important as they often provide a crucial bridge between basic NLP techniques such as morpho-syntactic analysis and end-level applications, and they have been mainly developed with corpus-based empirical approaches. In order to train a classification model in such approaches, a large scale corpus annotated with predicate-argument and coreference information is needed. To our best knowledge, however, there is no corpus including plenty of such tags in Japanese. In addition, we have difficulty adopting the traditional specifications for annotating tags due to the problem setting of each task and the difference between Japanese and English. So, we develop a new criteria for our annotating processes by examining the previous work on annotating tasks. This paper explains our annotating specification cultivated through actual annotating processes for the texts in Kyoto Text Corpus version 3.0, and discusses the future directions.

1 はじめに

情報抽出や機械翻訳などのNLPの応用処理への需要が高まる中で，その技術を実現するための中核的な要素技術となる共参照と述語項構造の解析に関して多くの研究者が解析技術を向上させてきた．それらの技術の多くは各情報が付与されたコーパス（以後，タグ付与コーパス）を訓練用データとして教師あり手法を用いるやり方が一般的であり，解析の対象となるコーパス作成の方法論についても議論がなされてきた [4, 6, 2]．

共参照解析については，主に英語を対象にいくつかのタグ付与のスキーマが提案されており，実際にそのスキーマに従ったコーパスが作成されている [4, 16, 3, 12, 2]．例えば，Message Understanding Conference (MUC) の Coreference (CO) タスク [4] や，その後継にあたる Automatic Content Ex-

traction (ACE) program の Entity Detection and Tracking (EDT) タスクでは，数年に渡って主に英語を対象に詳細な仕様が設計されてきた．また，述語項構造解析に関しては，CoNLL の shared task [1] で評価データとして利用されている PropBank[11] を対象に仕様が模索されてきた．

日本語を対象に述語項構造と共参照の研究をするにあたり，分析，学習，評価のための大規模なタグ付きコーパスが必要となるが，現状で利用可能な Global Document Annotation (GDA) [3] タグ付与コーパス（以後，GDA コーパス）や京都テキストコーパス第 4.0 版（以後，京都コーパス 4.0）は，述語項構造や共参照の解析のための十分な規模の評価データとはいえない．

また，タグ付与の仕様についてもおおきく以下の 2 点を考える必要がある．

- MUC や ACE の仕様は情報抽出に特化したものになっているが，この仕様をそのまま導入することの問題点．
- 英語と日本語の言語の差異によって生じる問題のずれ．

そこで，本稿では，タグ付与に関する既存の研究を吟味し，我々が取り組んでいる述語項構造と共参照情報の書き言葉コーパスへのタグ付与に関してどのような仕様を採用するかについて述べる．2 節で照応と共参照の関係について確認し，3 節では述語項構造と共参照のタグ付与に関する先行研究を紹介する．次に，4 節で先行研究を踏まえた上の我々のタグ付与の指針を示す．5 節で現状のタグ付与の問題点を述べ，その問題についての議論を行い，最後に 6 節でまとめる．

また，今回の作業の結果作成された述語項構造と共参照タグ付与コーパスを NAIST テキストコーパスとして公開している[2]．

2 照応と共参照

照応とはある表現が同一文章内の他の表現を指す機能をいい，指す側の表現を**照応詞**，指される側の表現を**先行詞**という．これに対し，二つ（もしくはそれ以上）の表現が現実世界（もしくは仮想世界）において同一の実体を指している場合には**共参照**（もしくは**同一指示**）の関係にあるという．先行詞となる表現が固有表現になる場合など，多くの場合は照応関係かつ共参照の関係が成り立つ．例えば，文章 (1) では，代名詞 “彼$_i$” が “村山首相$_i$” を指しており，かつ同一の人物を指しているため，照応関係かつ共参照関係であるとみなすことができる．

(1) 村山首相$_i$ は...
　　彼$_i$ は ...

[1] http://www.lsi.upc.edu/~srlconll/

[2] コーパスの情報は http://cl.naist.jp/nldata/corpus/を，今回紙面の都合上述べることできなかった仕様の詳細は http://cl.naist.jp/~ryu-i/coreference_tag.html を参照されたい．

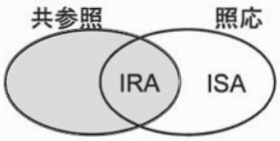

図 1: IRA と ISA の関係

これに対し，文章 (2) では，2 文目の “それ$_i$” は 1 文目の “iPod$_i$” を指しているため照応関係となるが，同じ実体を指していないため共参照関係とはならない．

(2) 太郎は iPod$_i$ を買った。
　　次郎も それ$_i$ を買った。

このように照応関係にある場合でも，同一の実体を指している場合とそれ以外の場合が存在する．文献 [9] では，前者のような共参照かつ照応関係となる関係を identity-of-reference anaphora（IRA），後者を identity-of-sense anaphora（ISA）と呼び区別している（概念間の関係は図 1 を参照）．

照応と共参照は異なる概念であるにもかかわらず，IRA が両方の性質を兼ねるため，既存研究ではそれぞれの概念が混同して扱われてきた．3 節で述べるタグ付与コーパス作成の先行研究でも同様にいくつかの異なる解釈で仕様が設計されている．

3 先行研究

この節では，共参照と述語項構造のタグ付与に関する主な先行研究を説明する．

3.1 共参照のタグ付与

情報抽出の主要な会議である Message Understanding Confernce (MUC) では，第 6 回と第 7 回の会議（以後，MUC-6 と MUC-7）において，情報抽出の部分問題として共参照解析の問題を扱っている[3]．MUC-6，MUC-7 の共参照関係タグ付与コーパスでは名詞句間の共参照関係がタグ付与され（ただし，動名詞（gerund）は除く），Soon ら [13] や Ng ら [10] などさまざまな機械学習に基づく共参照解析手法の gold standard データとして利用されてきた．しかし，このコーパスの仕様では，一般に共参照関係とはみなされないような量化表現（every, most など）を伴う場合や同格表現（Julius Caesar$_i$, the/a well-known emperor$_i$,...）も共参照関係とみなしてタグ付与されているという問題を含んでいる[4]．

MUC の共参照解析タスクの後継に相当する Automatic Content Extraction (ACE) [2] の Entity Detection and Tracking (EDT) では，この過剰な共参照関係の認定を回避するために，mention（**言及**）と entity（**実体**）という 2 つの概念を導入しタスクを設定している．言及とは文章中に出現する

[3] http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

[4] 詳細は van Deemter ら [14] を参照されたい．

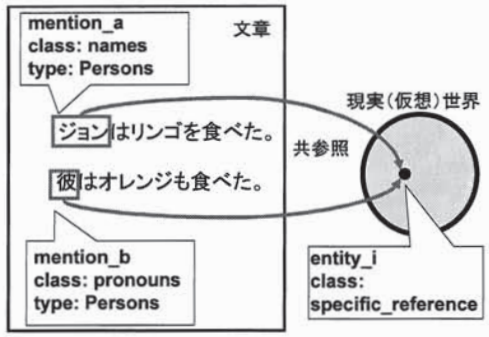

図 2: 言及（mention）と実体（entity）

表現のことで，情報抽出で解析の対象となるいくつかの固有表現を含む．これに対し，実体とは 2 節で述べた意味での実体，つまり現実世界（もしくは仮想世界）で指せるモノを意味する．一例をあげると，図 2 の文章において，“ジョン” と “彼” はそれぞれクラスが names と pronouns の言及であり，その 2 つの表現が実体のレベルではクラスが specific_reference である同一の実体を指しているというタグ付与を行う．

EDT のタグ付与[5]では，言及の型が人名や組織名などいくつかの固有表現に該当する場合で，かつ総称的（generic）でない場合にのみ共参照関係のタグ付与を行う．このため，ACE のデータセットでは文章内に出現する共参照関係に網羅的にタグが付与されず，文章内のすべての実体を対象として解析を行うには不十分なデータとなっている．

日本語に関しても，京都コーパス 4.0[16] や GDA コーパス [3] などのタグ付きコーパスに共参照相当のタグが付与されている．京都コーパス 4.0 には，係り受けの情報に加え，毎日新聞 95 年度版の一部（555 記事，5,127 文）に 114,729 もの共参照タグが付与されている．ただし，このコーパスでは ACE で導入されている実体と実体の間の共参照関係に加え，実体と属性の間にも共参照関係のタグを付与している．例えば，文 (3) において，実体 “村山 $_i$” とその属性 “首相 $_i$” の間に共参照の関係が付与されるという特徴を持つ．

(3) 村山$_i$首相$_i$は...

また，タグ付与対象の規模が小さいため，新聞記事ではあまり出現しない代名詞に相当するタグは 555 記事中に 276 個しか出現しない[6]．そのため，このコーパスを代名詞の照応解析の訓練事例とみなした場合に，規模が十分に大きいとはいえない．

一方，GDA コーパスでは，実体か総称的な表記かを区別せずに共参照タグを付与している．下記の (4) は GDA コーパスから抜粋したものだが，この文章では総称名詞である 2 つの “フロン $_i$” に対して共参照タグが付与されている．このような例が多数見られたため，GDA の共参照タグは IRA と ISA の両方の関係で付与されていると考えられる．

(4) フロン$_i$対策急げ...フロン$_i$による環境対策は...

3.2 述語項構造のタグ付与

述語とその項[7]のタグ付与に関しては，表層格レベルから深層格レベルまでさまざまなレベルでのタグ付与についての議論がある．例えば，英語を対象とした PropBank [11] では，述語の項のラベルとして，基本的には agent や theme などの意味役割に相当する ARG0, ARG1, ..., ARG5, AA, AM, AM-ADV など，35 種類のタグを用いて文章にタグ付与を行っている．一例をあげると，文 (5) に出現している動詞 “earned” に対し，“the refiner” を agent 相当である ARG0，“$66 million, or $1.19 a share” を theme に相当する ARG1 としてタグが付与されている．

(5) [ARGM−TMP *A year earlier*], [ARG0 *the refiner*] [rel *earned*] [ARG1 *$66 million, or $1.19 a share*].

ただし，対象言語が英語である場合，タグ付与の対象となる述語の項は述語と同一文内に出現するため，PropBank ではタグ付与の範囲を同一文内に限定している．

一方，日本語を対象にする場合は必須格が省略されるゼロ照応の現象が頻繁に起きるため，文を越えて，もしくは文章外の要素にもタグを付与する必要がある．京都コーパス 4.0 では文間，外界照応となる項に関してもタグが付与されている．共参照タグ付与の対象となった 555 記事を対象にガ/ヲ/ニ/カラ/ヘ/ト/ヨリ/マデなどの格助詞相当の表層格に加え，ニツイテのような連語も一つの表層格としてタグが付与されている．例えば，文章 (6) の 2 文目に出現している述語 “帰っ（て）” では 1 文目に出現している “トム” をガ格としてタグ付与し，またニ格は文章中に先行詞が無いため “外界照応” のラベルを付与する．

(6) トム$_i$は今日学校へ行った。
帰つ$_{ガ:i,\ ニ:外界照応}$てすぐに遊びにでかけた。

また，GDA コーパスではゼロ照応に関して agent, theme などの粒度で意味役割のタグが付与されているが，我々が確認した限りでは，述語と係り関係にある場合や，ゼロ照応の場合であっても同一文内に先行詞が出現している場合にはタグが付与されておらず，学習手法の訓練事例として利用するには網羅性の点で問題が残る．

3.3 事態性名詞のタグ付与

動詞や形容詞などの述語への項構造の付与に加え，事態としての意味を伴うような名詞（以後，**事**

[5] http://projects.ldc.upenn.edu/ace/annotation/

[6] `rel="="` が付与された箇所のうち，品詞が「指示詞」である場合と IPADIC[18] の品詞体系の「名詞-代名詞」に該当する文字列を含む場合の個数を求めた．前者が 181 個，後者が 95 個であるが，そのうち節を指す場合が 63 個（全体の 23%）であるため，名詞句間の関係に付与された個数は 213 となる．

[7] 本稿で用いる用語 “項” は complement と adjunct の両方を指す．

態性名詞）に対して項構造情報を付与する試みも報告されている．

例えば，Meyers らが作成した NomBank[8] では，Penn Treebank[7] を対象に事態性名詞とその項のタグ付与を行っている．このコーパスでは英語における動詞の名詞化に着目して，PropBank[11] の仕様に従って項構造を付与している．例えば，句 (7) において，名詞 “growth” はある事態を表しており，その項として名詞句内の “in dividents” と “next year” がそれぞれ theme 相当の項と任意格相当の項として付与されている．

(7) *12% growth in dividends next year* [REL=*growth*, ARG1=*in dividends*, ARGM-TMP=*next year*]

ただし，PropBank の仕様に準拠しているため，タグ付与対象となる項は文内（多くの場合は句の中）に制限される．

日本語に関しても，京都コーパス 4.0 では事態性名詞とその項に対して表層格でタグが付与されている．例えば，文 (8) に示すように “及ぼす” の格要素となっている事態 “影響” に対して，“離党$_i$” が影響することのガ格として付与されている．

(8) 新民主連合所属議員の**離党**$_i$ 問題について「政権に 影響$_{ガ:i}$ を及ぼすことにはならない。...

また，事態と項の関係が「**候補**$_i$擁立$_{ヲ:i}$」や「**兵士**$_j$ の 脱走$_{ガ:j}$」のように，複合名詞句の中や “A ノ B” などに縮退される場合もあり，このような場合どこまでをタグ付与の対象とするかを明示的に決める必要がある．

4 NAIST テキストコーパスの仕様

3 節で述べた先行研究を踏まえ，今回の作業ではおおきく (1) 述語の基本形とその表層格，(2) 事態性名詞とその表層格，(3)IRA の関係のみを対象とした共参照関係の 3 つの関係を対象にタグを付与した．

4.1 述語と項のタグ付与

述語そのものの認定に関しては，品詞体系として IPADIC[18] を採用し，動詞，形容詞，名詞+“だ（助動詞）” の 3 種類をタグ付けの対象となる述語とみなし，作業を行う．

述語の格要素については，京都コーパス 4.0 が採用しているような表層格，GDA のような深層格，また PropBank で付与されているような独自の基準など，さまざまなタグ付与のレベルが考えられる．この中で我々は「誰が何を何に対してどうする」という情報抽出的な観点でタグを付与することが自然だと考え，述語の原形に対して項のタグを付与する．ただし，表層レベルからなんらかの情報を捨象して意味のレベルでタグを付与することが応用処理に寄与するのかが現状では自明でないため，格交替の情報のみを捨象して表層格でタグ付与を行った．例えば，京都コーパス 4.0 では文 (9) の述語 “食べさせる” に対して “私$_i$”，“彼$_j$”，“リンゴ$_k$” をそれぞれガ，ヲ，ニ格でタグ付与するのに対し，我々の仕様では述語の原形 “食べる” に対して “彼$_j$ ガリンゴ$_k$ ヲ食べる” というタグを付与する．ただし，述語の原形に対してタグを付与する場合には使役者に相当する “私$_i$” と述語 “食べる” の間の関係にタグが付与されないことになる．これを回避するため，格要素を増やす助動詞に対してタグ “追加ガ（ニ）格” を付与した．例えば，文 (9) では，助動詞 “させる” に対し “私$_i$” を追加ガ格でタグ付与し，文 (10) では助動詞 “やる” に対し “彼$_j$” を追加ニ格でタグ付与する．

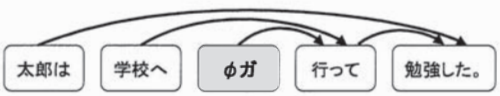

図 3: 文内ゼロ照応の認定

表 1: 述語と項のタグ付与の比較

コーパス	付与の対象	付与の範囲
PropBank	意味役割相当	intra
GDA	意味役割	inter, exo
京都コーパス 4.0	表層格（出現形）	intra, inter, exo
NAIST コーパス	表層格（基本形）	intra, inter, exo

intra:文内照応，inter:文間照応，exo:外界照応

(9) a. **私**$_i$ は**彼**$_j$ に**リンゴ**$_k$ を 食べさせる$_{ガ:i,\ ヲ:k,\ ニ:j}$。
b. **私**$_i$ は**彼**$_j$ に**リンゴ**$_k$ を 食べ$_{ガ:j,\ ヲ:k}$ させる$_{追加ガ格:i}$。

(10) **私**$_i$ は**彼**$_j$ に**本**$_k$ を 読ん$_{ガ:i,\ ヲ:k}$ で やる$_{追加ニ格:j}$。

また，京都コーパス 4.0 では表層格を網羅する形で作業を進められたが，今回の作業では，頻出するガ/ヲ/ニ格のみを対象に作業を進め，どの程度の品質で作業ができるかを調査した．

表層格は項が係り受け関係にある場合に加え，省略がある場合にも区別せずに作業を行う．例えば，図 3 では述語 “行っ（て）” は係り関係にある文節にガ格に相当するものがないため，省略があるとみなすが[8]，この場合にも区別せずに “太郎” にガ格のタグを付与する[9]．当面このような基準で作業を進めることで，今後深層格の情報をタグとして付与する必要がでてきた場合にも，ゼロ照応の関係を含む形で述語の基本形の必須格に付与した情報は，agent，theme のような意味役割を付与する場合や語彙概念構造（LCS）[5] の意味述語の情報を付与する際にも役立つと考えられる．

述語に関する我々の仕様と他のコーパスの仕様の比較をまとめると表 1 のようになる．

[8] この例は並列表現はゼロ照応と区別すべきという議論もあるが，項が係り受け関係にない場合は統一的にゼロ照応とみなすほうが機械処理を行う際には見通しがよいと考えている．

[9] 最終的に付与される情報には係り受け関係は含まないが，京都コーパスの係り受け情報と統合することによりゼロ照応か否かの判別が可能である．表 3 でまとめる統計量も京都コーパスと統合した結果を用いて求めた．

4.2 事態性名詞と項のタグ付与

動詞や形容詞などの述語に加え，事態性名詞に対して述語と同様に必須格となる表層ガ/ヲ/ニ格を付与する．作業者は与えられた名詞（主にサ変名詞）が事態を表しているか否かを判定し，事態性名詞と判断した名詞（句）に対して必須格を付与する．例えば，文 (11) で出現している二つの "電話" という名詞のうち，"電話 $_i$" が「電話する」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 $_i$" のみを事態性名詞と認定し，これに対して "彼 $_a$" をガ格，"私 $_b$" をニ格として付与しなければならない．

(11) **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。

また，タグ付与の対象が複合語の場合はその構成素を構成的に分解した上でそれぞれの構成素に対して事態性判別の作業を行う．例えば，「紛争仲裁」は構成素 "紛争" と "仲裁" のそれぞれの意味を構成的に組み合せてできた複合語だとみなし，"仲裁" を事態性名詞と判断する．一方，「フランス革命」のような分解すると複合語の持つ意味が欠落する場合にはこれ以上分解せず，事態性名詞のタグを付与しない．

4.3 名詞句間の共参照関係のタグ付与

共参照のタグ付与では，2 節で述べた IRA に加え ISA の関係も含めてタグを付与するか否かの選択肢があるが，ISA の関係まで含めてしまうと，総称名詞間の包含関係のような複雑な関係を考慮して作業を行う必要がある．例えば，文章 (12) では，総称名詞 "図書館 $_a$" と総称名詞 "図書館 $_b$" が同一の概念を指しているために共参照関係として設定すべきかもしれない．しかし，"本 $_i$" と "本 $_j$" の場合は "本 $_i$" が「本を意味する類に属するすべての要素」を指すのに対し，"本 $_j$" は「図書館の本（図書館に置いてある本)」を指し，二つの総称名詞の間には "本 $_i \supset$ 本 $_j$" という包含関係が成り立つため，"本 $_i$" と "本 $_j$" の間に共参照関係を認めるか否かに揺れが生じることになる．

(12) 図書館$_a$ には 本$_i$ が置いてある．
図書館$_b$ の 本$_j$ は借りることができる．

そこで，述語や事態性名詞が ISA も含めた関係にタグ付与しているのに対し，共参照に関しては IRA の関係にのみタグを付与する．ただし，EDT の仕様のように，実体が組織名や場所名など数種の固有表現に限定して共参照関係のタグを付与することは，さまざまな応用分野で必要となる共参照の表現を網羅できないため望ましくない．そこで，今回の作業では，作業者にはいくつか作業の具体例とともに以下の 3 つの基準を提示するだけで，表現のクラスを限定せずに共参照関係のタグ付与を行い，どのような問題が生じるのかを調査した．

表 2: 共参照タグ付与の差異

コーパス	特徴
GDA	IRA と ISA の関係両方に付与.
ACE EDT	IRA の関係にのみ付与. ただし，実体がいくつかの固有表現のクラスに限定されている.
京都コーパス 4.0	IRA と ISA の関係両方に付与.
NAIST コーパス	IRA の関係にのみ付与.

1. **照応詞は文節の主辞（最右の名詞自立語）のみに限定する.**
2. **談話内に出現した名詞句のみを先行詞とする.**
3. **総称名詞は照応詞，先行詞とみなさない.**

既存の共参照関係のタグ付与の研究と比較すると表 2 のようになる．

4.4 統計

4.1，4.2，4.3 の仕様に従い，京都コーパス 3.0 の全記事（2,929 記事，38,384 文）を対象に，2 人の作業者が述語項構造と共参照の関係についてタグ付与作業を行った．述語/事態性名詞とその項に付与されたタグの個数を表 3 にまとめる．ただし，項の出現位置によって，同一文節内[10]，係り関係にある場合[11]，文内のゼロ照応関係，文間のゼロ照応関係，文章内に項が出現しない文章外ゼロ照応の 5 つに分類して頻度を求めた．表 3 より，述語の項目ではヲ格，ニ格の項のほとんどは係り関係にあるのに対し，ガ格の約 6 割はゼロ照応の関係にあることがわかる．これに対して，事態性名詞のヲ格，ニ格は同一文節内，つまり複合語の構成素として項が出現している割合が高く，ガ格に関しては約 8 割がゼロ照応の関係にあり，述語の場合と比較して項の出現箇所がおおきく異なっていることがわかる．

共参照関係のタグについては，タグ付与された実体の総数が 10,531，最初に出現した表現を先行詞，その他を照応詞とみなしたときの照応詞の個数が 25,357 であった．京都コーパス 4.0 より圧倒的に個数が少ないが，これは実体間の関係にのみ限定して作業を行ったためだと考えられる．また，共参照に関与する代名詞の個数は 622 と，京都コーパスと比較してタグ付与した記事に対する割合が小さい．これは作業者が (1) 節照応の付与を行わずに，(2) 共参照タグ付与に関して厳密な実体の一致を強いたために完全に一致するとみなせない場合にはタグが付与されなかったためだと考えられる．代名詞に関しては現状の仕様のように IRA の関係でタグを付与する立場と，実体の一致を問題としない ISA の関係で付与するという立場の二つを考えることができ，それぞれ必要となる仕様は応用分野によって異なる．そのため，代名詞については追加的に ISA 関係のタグを分けて付与することも今後検討したい．

[10] 「～(ここを埋める) になる」のような表現がコーパス中では一文節であるのに対し，作業者が「～に」と「なる」を分けて付与した場合などを含む．

[11] 「サンマを焼く男」の「男」が「焼く」のガ格となるような，連体修飾の関係も含む．

表 3: 述語項構造に関するタグの統計

	出現箇所	ガ格		ヲ格		ニ格	
述語	同一文節内	177	(0.002)	60	(0.001)	591	(0.027)
106,628	係り関係	44,402	(0.419)	35,882	(0.835)	18,912	(0.879)
	ゼロ照応 (文内)	32,270	(0.305)	5,625	(0.131)	1,417	(0.066)
	ゼロ照応 (文間)	13,181	(0.124)	1,307	(0.030)	542	(0.025)
	ゼロ照応 (文章外)	15,885	(0.150)	96	(0.002)	45	(0.002)
	全体	105,915	(1.000)	42,970	(1.000)	21,507	(1.000)
事態性名詞	同一文節内	2,195	(0.077)	5,574	(0.506)	846	(0.436)
28,569	係り関係	4,332	(0.152)	2,890	(0.263)	298	(0.154)
	ゼロ照応 (文内)	9,222	(0.324)	1,645	(0.149)	586	(0.302)
	ゼロ照応 (文間)	5,190	(0.183)	854	(0.078)	201	(0.104)
	ゼロ照応 (文章外)	7,525	(0.264)	42	(0.004)	10	(0.005)
	全体	28,464	(1.000)	11,005	(1.000)	1,941	(1.000)

表 4: タグの一致率

	再現率		精度	
述語	0.921	(806/875)	0.944	(806/854)
ガ格	0.823	(683/830)	0.829	(683/824)
ヲ格	0.899	(329/366)	0.954	(329/345)
ニ格	0.724	(105/145)	0.890	(105/118)
事態性名詞	0.965	(247/256)	0.792	(247/312)
ガ格	0.735	(191/260)	0.743	(191/257)
ヲ格	0.827	(86/104)	0.869	(86/99)
ニ格	0.389	(7/18)	0.583	(7/12)
共参照	0.813	(126/155)	0.813	(126/155)

次に，実際に作業を行っている 2 人の作業者間のタグ付与の一致率を調査するため，ランダムに選択した報道 30 記事を対象に作業を行った．評価は一方の作業者のタグ付与の結果を正解，他方の作業結果をシステムの出力とみなし再現率と精度で評価する．ただし，それぞれのタグの一致率は各タグの終了位置の一致で評価した．また，述語と事態性名詞の項の一致率については，2 人の作業者の述語（事態性名詞）が一致した箇所のみを対象に評価した．また，共参照の一致率については MUC score[15] を用いて再現率と精度を求めた．これらの基準で評価した結果を表 4 に示す．表 4 よりわかるように，それぞれのタグ付与は多くの場合 8 割を越える品質で作業ができているが，改善の余地は大きい．5 節では，各タグ付与において，問題となった主要な点を説明し，その問題を解決するための今後の方向性について議論する．

5 タグ付与の問題点と今後の展望

この節では述語，事態性名詞，共参照のそれぞれのタグ付与作業中に生じた問題を説明し，それに対する今後の対応などをまとめる．

5.1 述語のタグ付与の問題点

まず，述語そのもののタグ付与に関してだが，タグ付与対象となる述語が「〜とし て」のような機能語相当表現と表現上では同一の場合に述語認定に揺れが生じることがわかった．土屋ら [19] は，機能語相当表現（複合辞）単体を対象に作業者間の一致率を評価しており，ある程度揺れなく作業できていることを示しているが，今回のように対象となる述語が項をとるか否かを判断しながら述語の認定を行う場合は，例えば「会社 A が会社 B を子会社とし て」では「として」が “ある一つの側面からの価値付け・意味付け” という意味の機能語相当表現なのか，それとも「会社 A が会社 B を子会社と する」と解釈すべきなのかを判断することが難しい．この問題はついては，各表現ごとにどちらに解釈すべきかをあらかじめ決めておき，できるだけ多くのタグ付与の例を作業者に提示することで対応したい．

5.2 事態性名詞タグ付与の問題点

事態性名詞の認定に関しては，述語の場合とは異なり，対象となる名詞（句）が出現文脈でモノとコトのどちらを表しているのを認定する作業が必要となるが，これに加え，「投資 率」のような複合語をどの程度構成的に分解できるかを判断しなければならず，この分解についての基準が作業者間で異なったために一致率が低下した．また，“契約”，“規制”，“投資” のような表現は事態として解釈可能であるが，文脈によってコトを表しているのか，事態が起こった結果できた結果物としてのモノなのかの判定が困難な場合が存在した．例えば，文 (13) では “インセンティブ規制” を結果物とみなすか，“規制” を事態とみなすかで揺れが生じた．

(13) 料金規制当局と公益事業者が、一種の社会契約を結んだという考えに立つもので、経営効率化促進のための社会契約インセンティブ 規制 とも言われる。

また，現状では主にサ変名詞を対象に，ガ/ヲ/ニ格の表層格で項との関係を付与しているが，事態性名詞は述語と異なり，助詞 “を” で述語に係っている格要素が基本的にはヲ格となるといった表層格による統語的な制約を受けない．そのため，表層格でタグを付与するほうが必ずしもよいというわけではなく，今後 “運動会” のような広い意味での事態をサ変名詞の事態といっしょに扱う場合には，各事態に関して agent，theme のような意味役割でタグ付けすることが望ましいかもしれない．

5.3 項のタグ付与の問題点

項のタグ付与に関しては述語が取り得る格パタンが複数存在するために作業者間で揺れが生じることがわかった．この問題の典型的な例が自動詞と他動詞の交替である．例えば，述語 “実現する” は同じ語義に対して表層格レベルで “agent ガ theme ヲ実現する” と “theme ガ実現する” の 2 つの格パタンが存在するため，文章中ですべての格要素が省略されている場合は，作業者はどちらの解釈でもタグ付与が可能になってしまう．自他交替の問題と類似して，制度などの表現に動作主性（agentivity）を認めるか否かで解釈が異なるために揺れが生じる場合もある．例えば，文 (14) において，述語 “しばる” は，直前に出現している “規制” の動作主性を認め，“規制ガ theme ヲしばる” と “agent ガ規制デ theme ヲしばる” の 2 つの解釈が存在する．

(14) 我々の生活が知らず知らずにどれだけ規制で<u>しばら</u>れているか、規制緩和によって豊かさが変わっていくのかを考えてみた。

このような交替を伴う場合の揺れに関しては，どちらかのパタンを優先するという規則をあらかじめ決めておき作業することで対応できると考えられる．

動作主性の問題に関連して，組織のような実体にどのくらい動作主性を認めるかが作業者間で異なるために揺れが生じる場合も頻繁に起こった．例えば，文 (15) では，組織 “与野党” もしくはその組織の “党首” が事態 “協力（する）” のガ格として解釈可能である．

(15) ... 自民、さきがけ、新進各党の**与野党**$_a$ の**党首**$_b$ 会談を呼び掛けて <u>協力</u> を求めるべきだ。

このような組織とその組織の関係者のような対立や，また文 (16) の “北朝鮮” と “同指導部” のような組織とその部署の関係など，ある名詞が他の名詞と関連しているために，複数の解釈が可能な場合は “（与野党ノ）党首” のように詳細化されているほうにタグを付与することによって作業者間の揺れを少なくすることができると考えられる．このように扱うことで，もし〈与野党，所属，党首〉のような名詞間の関係解析が実現できれば，他方の名詞（(15) では “与野党”）と対象となる述語を関連付けて扱うことができる．

(16) **北朝鮮**$_a$ における新年の辞は、**同指導部**$_b$ の施政方針 <u>発表</u> に当たる重要行事である。

また図 4(a) のように，項としてタグ付与されるべき名詞句が IRA の関係で他の名詞句と関連付けられている場合は，共参照関係にある名詞句のいずれかを項として同定する問題とみなすことができるが，一方図 4(b) の “子供” と “児童” のような ISA の関係で出現している名詞句については，ラベルが付与されていない “子供” は述語の項としてタグが付与されないという問題が起こる．

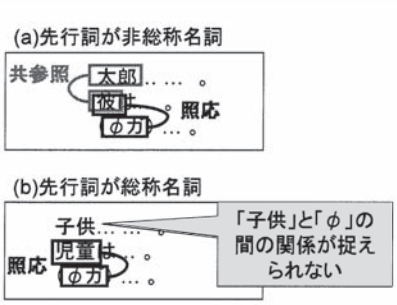

図 4: 先行詞が総称名詞の場合のタグ付与の漏れ

5.4 共参照タグ付与の問題点

IRA のみを対象に共参照のタグを付与する作業に関してもいくつかの問題が残る．まず 1 つ目の問題を文章 (17) を例に説明しよう．

(17) グロズヌイからの報道によると三日、大統領官邸の北西一・五キロの鉄道駅付近で<u>ロシア軍部隊$_i$</u>とチェチェン側部隊が衝突したが、<u>ロシア側$_i$</u>は中心部への進撃を阻まれて苦戦。... <u>ロシア政府$_j$</u>は三日、戦況に関する声明を発表し、大統領官邸を含む首都中心部は依然としてロシア側が支配していると強調した。しかし現地からのテレビ映像では、官邸はじめ中心部は依然としてドゥダエフ政権部隊の兵士が警戒に当たっており、<u>ロシア側$_j$</u>の発表と食い違いを見せている。

この例で，最初に出現する “ロシア側 $_i$” が “ロシア軍部隊 $_i$” の換喩に相当するのに対し，次に出現する “ロシア側 $_j$” は “ロシア政府 $_j$” の換喩として解釈できる．現状の仕様では “ロシア軍部隊 $_i$” と “ロシア側 $_i$”，“ロシア政府 $_j$” と “ロシア側 $_j$” それぞれに共参照関係のタグを付与することになるが，換喩の解釈しながらタグ付与作業を行なうことが困難であることに加え，実際の自動解析する際にも非常に困難な問題設定となる．この問題を回避するために，換喩の解釈で共参照のタグを付与するのではなく，文章に出現している 4 つの “ロシア” を同一の実体としてタグ付与するなどの方法が考えられるが，どのように仕様を決めた方がよいかは明らかでないため今後検討していく必要がある．

また，IRA の認定に関しても，実体が具体名詞である場合は 2 つの言及が同一の実体を指すか否かの認定が容易であるが，抽象名詞の場合は同じものを指しているかの判定が困難である．4.3 で共参照関係のタグ付与にはあらかじめ名詞のクラスを指定して作業を行うことは望ましくないと述べたが，抽象名詞に関してはいくつかの意味クラスに限定して作業を行い，どのくらい揺れなく作業できるかを調査したい．

6 おわりに

本稿では，日本語を対象とした述語項構造・共参照タグ付与コーパスに関して，我々が今回採用したタグ付与の基準について報告した．4 節の議論に基

づき，述語項構造のタグに関してはISAとIRAの関係両方で，共参照関係はIRAの関係でタグ付与作業を行い，京都コーパス3.0を対象にこれまでにない大規模な述語項構造・共参照タグ付きコーパスを作成した．また，作業の過程で起こった問題について考察し，作業の詳細化のための項目を述べた．

今回作業では述語と事態性名詞の表層ガ/ヲ/ニ格と共参照関係のタグ付与を行ったが，情報抽出などの応用分野を想定した場合，今回作業したラベルに加え，以下に示す内容に取り組む必要があると考えている．

まず，述語項構造の関係に関しては，必須ガ/ヲ/ニ格に加え，カラ格やデ格などその他の表層格の付与が必要だと考えており，これについては現在作業が進行中で，作業結果を反映して再公開する予定である．

また，今回の作業では名詞間の関係として共参照関係のみを作業対象としたが，上位下位や部分全体，所属関係など，さまざまな関係の解析も述語項構造・共参照解析と同様に応用処理のための重要な構成素となる．この名詞間の関係について，京都コーパス4.0で採用されている“A ノ B”の粒度でタグを付与した場合，この“ノ”で付与した結果には上位下位関係や部分全体関係などさまざまな関係を含んでしまうため，関係抽出の粒度としては不十分である．また，この名詞句間の関係解析は，bridging reference[1]や間接照応[17]などの用語で表現される場合もあるが，bridging referenceは一般に英語の定情報（definite）の存在が仮定された上で述べられることが多い．つまり，“the”を伴った名詞句があるにもかかわらず，参照する先行詞が文章中に出現していない場合にどう解釈すればよいかという点が議論の中心となっている．一方，日本語などの冠詞のが利用できない言語の場合，“the”のような手がかりがないために，どの名詞句の対に対して間接照応の関係を付与するかという課題設計そのものが困難になると考えられる．ACEのRelation Detection and Characterization (RDC)タスクでは，3.1で述べた実体の間の関係にのみ抽出対象となる関係を定義しているが，実体のクラスをオープンにした場合に揺れなく作業できるかについても今後調査したい．

さらに，今回の作業では新聞記事を対象に作業を行ったが，例えば代名詞の出現が少ないなど，このコーパス内の用例だけを学習手法の訓練事例として利用すると，blogなどの照応解析，述語項構造解析を適用したい記事との異なりのために適切に解析できない恐れがあり，今後はタグ付与作業をいくつかの領域に拡張して進める必要がある．

また，タグ付与に関する仕様書に関して，それぞれ個別の仕様について，外延的に例を示すだけで仕様をまとめるのではなく，それぞれのタグがどのような性質を持っているために付与されているかという内包的な仕様も明示的に記述することで，実際に解析に利用した研究者が問題の性質を分析するのに役立つ仕様書を作成することが重要だと考えており，今後の作業内容については順次Webページ[12]にまとめていく予定である．

参考文献

[1] Clark, H. H.: Bridging, *Thinking: Readings in Cognitive Science* (Johnson-Laird, P. N. and Wason, P.(eds.)), Cambridge University Press (1977).

[2] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. and Weischedel, R.: Automatic Content Extraction (ACE) program - task definitions and performance measures, *Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 837–840 (2004).

[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] Hirschman, L.: *MUC-7 coreference task definition*. Version 3.0 (1997).

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

[6] Kingsbury, P. and Palmer, M.: From TreeBank to PropBank, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 1989–1993 (2002).

[7] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A.: Building a Large Annotated Corpus of English:The Penn Treebank, *Computational Linguistics*, pp. 313–330 (1993).

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

[9] Mitkov, R.(ed.): *Anaphora Resolution*, Studies in Language and Linguistics, Pearson Education (2002).

[10] Ng, V. and Cardie, C.: Improving Machine Learning Approaches to Coreference Resolution, *Proceedings of the 40th ACL*, pp. 104–111 (2002a).

[11] 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).

[12] Poesio, M.: Discourse Annotation and Semantic Annotation in the GNOME Corpus, *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, pp. 72–79 (2004).

[13] Soon, W. M., Ng, H. T. and Lim, D. C. Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases, *Computational Linguistics*, Vol. 27, No. 4, pp. 521–544 (2001).

[14] van Deemter, K. and Kibble, R.: What is coreference, and what should coreference annotation be?, *Proceedings of the ACL '99 Workshop on Coreference and its applications*, pp. 90–96 (1999).

[15] Vilain, M., Burger, J., Aberdeen, J., Connolly, D. and Hirschman, L.: A Model-Theoretic Coreference Scoring Scheme, *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52 (1995).

[16] 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).

[17] 山梨正明: 推論と照応, くろしお出版 (1992).

[18] 浅原正幸, 松本裕治: ipadic version 2.6.3 ユーザーズマニュアル (2003).

[19] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一: 日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).

[12] http://cl.naist.jp/~ryu-i/coreference_tag.html