Vol.2012-NL-206 No.7
Vol.2012-SLP-91 No.7
2012/5/10

多様な文書の書き始めに対する意味関係タグ付きコーパスの構築

萩行 正嗣[1,a)]　河原 大輔[1,b)]　黒橋 禎夫[1,c)]

概要：現在，自然言語処理では意味・談話解析の本格的な取り組みが始まりつつある．意味・談話解析の研究には意味・談話関係を付与したコーパスが必要であるが，従来の意味・談話関係のタグ付きコーパスは新聞記事を中心に整備されてきた．一方，文書には多様なジャンル，文体のものが存在し，その中には新聞記事では出現しないような言語現象が出現する場合がある．本研究では，Web を利用することで多様な文書の書き始めからなる意味・談話関係タグ付きコーパスを構築し，その分析を行った．

キーワード：タグ付きコーパス，意味・談話関係，照応関係，述語項構造

Building Diverse Document Leads Corpus Annotated with Semantic Relations

Masatsugu Hangyo[1,a)]　Daisuke Kawahara[1,b)]　Sadao Kurohashi[1,c)]

Abstract: Recently, semantic analysis has been actively studied in Natural Language Processing. A corpus which is annotated with semantic relations is necessary for the study of semantic analysis. Although there is such a corpus annotated on newspaper articles, there are texts of various genres and styles which contain linguistic expressions that are not found in newspaper articles. In this paper, we built a diverse document leads corpus annotated with semantic relations and report the statistics of this corpus.

Keywords: annotated corpus, discourse, anaphora, predicate-argument structure

1. はじめに

現在，自然言語処理では意味・談話解析の本格的な取り組みが始まりつつある．これまでの日本語の意味・談話解析の研究は意味・談話関係を付与した新聞記事コーパスを用いて行われてきた．しかし，テキストには新聞記事以外にも百科事典や日記，小説など多様なジャンルがあり，ジャンル内においても多様な文体がある．これらの多様なテキストの中には依頼表現，敬語表現など新聞記事ではあまり出現しない言語現象も出現する．このような言語現象を含むテキストの意味・談話解析を行うためには，多様なテキストからなるタグ付きコーパスの構築とその分析が重要となる．Web ページにはニュース記事，百科事典記事，blog，商用ページなど多様なジャンル，文体のテキストが存在する．そこで本研究は Web から収集したテキストを利用することで，多様なジャンルの文書からなるコーパスの作成を行った．

本研究では意味・談話関係のタグ付けとして述語項構造，照応関係のタグ付けを行う．これらの関係およびそのタグ付けを以下の例 (1) で説明する．なお A←B は A に B というタグを付与することを表す．また以降の例では議論に関係ないタグについては省略する場合がある．

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
[a)] hangyo@nlp.ist.i.kyoto-u.ac.jp
[b)] dk@i.kyoto-u.ac.jp
[c)] kuro@i.kyoto-u.ac.jp

(1)　太郎は時計を買った。
　　(買った ← ガ:太郎，ヲ:時計)
　　弟にそれをあげた。
$$\begin{pmatrix} \text{弟} \leftarrow \text{ノ:太郎} \\ \text{それ} \leftarrow =\text{:時計} \\ \text{あげた} \leftarrow \text{ガ:太郎，ヲ:それ，ニ:弟} \end{pmatrix}$$

述語項構造は述語とその項の関係を記述したもので，例 (1) の「買った」のガ格が「太郎」，ヲ格が「時計」という関係である．この場合「太郎」の格は明示されていないが述語項構造としてはガ格となる．照応関係とは談話中のある表現 (照応詞) が別の表現 (先行詞) を指す現象である[*1]．例 (1)2 文目では「それ」が 1 文目の「時計」を指している．日本語では述語の項が省略されるゼロ照応と呼ばれる現象が頻出する．ゼロ照応と呼ぶのは，そこに「彼」「それ」など何らかの照応詞があると考えられるからであり，その省略された照応詞をゼロ代名詞と呼ぶ．「あげた」の述語項構造のガ格が「太郎」と記述することにより，ガ格にゼロ代名詞が存在し，そのゼロ代名詞の先行詞が「太郎」であることを表現できる．

また，照応関係の中には橋渡し照応と呼ばれる現象がある．これは，照応詞が先行詞を直接指すのではなく，照応詞の何らかの属性が先行詞を指す現象である．例 (1) では，「弟」という語にある「誰かの弟」という属性の「誰か」が「兄」を指していると考えられる．橋渡し照応の指す属性は上位下位関係，部分全体関係，例示，対比関係など多様なものが存在する．

形態素，構文関係のタグ付けは文単位で独立であり，文書が長くなっても作業量は線形にしか増加しない．一方，意味・談話関係のタグ付けでは文をまたぐ関係を扱うため，文書が長くなると作業者が考慮すべき要素が組み合わせ的に増加する．このため 1 文書あたりの作業時間が長くなり，文書全体にタグ付けを行うと，タグ付けできる文書数が限られてしまう．本研究では多様な文書からなるタグ付きコーパスを目的としているため，先頭の数文に限定してタグ付けを行うことで 1 文書あたりの作業量を抑える．意味・談話解析では既に解析した前方の文の解析結果を利用する場合があり，先頭の解析誤りが後続文の解析に悪影響を与える．先頭数文に限定したコーパスを作ることで，文書の先頭の解析精度を上げることが期待でき，全体での精度向上にも寄与できると考えられる．

本論文ではまず 2 章で関連研究について述べる．3 章でコーパスを構成する文書について述べ，4 章でタグ付けについて述べる．5 章でタグ付けされたコーパスの性質について議論し，6 章でまとめとする．

[*1] 照応に類似した概念として共参照が存在するが，共参照は照応で表現できるものがほとんどなので，本論文では特に断りがない限り照応として扱う．

2. 関連研究

日本語の述語項構造および照応関係タグ付きコーパスとしては，京都大学テキストコーパス [4] と NAIST テキストコーパス [5] がある．これらのコーパスは 1995 年の毎日新聞に述語項構造および共参照関係を付与したコーパスである．新聞記事は内容が報道と社説に限られており，文体も統一されているため，新聞記事以外の意味・談話解析への適応には不向きである．

様々なジャンルからなる日本語コーパスとしては現代日本語書き言葉均衡コーパス (BCCWJ)[*2]がある．このコーパスは書籍，雑誌などの出版物やインターネット上のテキストなどからなるコーパスである．このコーパスでは，書籍などについては幅広いジャンルのテキストから構築されているが，インターネット上のテキストは掲示板やブログなどに限定されている．このためインターネット上に多数存在する企業ページなどはコーパスには含まれない．

BCCWJ に意味・談話関係を付与する研究として，日本語 FrameNet を付与するものがある [3]．この研究では BCCWJ のコアデータに含まれる用言に対して FrameNet で定義された述語項構造の記述を行っている．しかし FrameNet ではゼロ代名詞の有無は述語項構造に含まれるものの，先行詞が同一文内にない場合にはその照応先の情報を付与していない．また，照応関係の情報も付与されておらず，文をまたぐ意味・談話関係の情報は付与されていない．

日本語以外で複数のジャンルに渡ってゼロ照応を扱ったコーパスとしては，Z-corpus[1] や LMC(Live Memories Corpus)[2] などがある．Z-corpus はスペイン語の法律書，教科書，百科事典記事に対しゼロ照応の情報を付与したコーパスである．ゼロ照応のみを扱っており，前方照応や述語項構造の情報は付与されていない．これはスペイン語ではゼロ照応は主語のみに発生するため述語項構造の情報とは独立にゼロ照応の情報を記述できるためである．

LMC はイタリア語の Wikipedia と blog に照応関係のタグ付けをしたコーパスである．照応関係としてゼロ照応も扱っているが，述語項構造は扱っていない．イタリア語もゼロ照応は主語のみに発生するので，このコーパスではゼロ照応の起こった用言を照応詞としてタグ付けしている．

3. タグ付与対象の文書

従来，意味・談話関係タグ付きコーパスの構築は新聞記事を中心に行われてきた．しかし，新聞記事にはほとんど出現しないような言語現象も存在し，そのような言語現象を研究するためには多様な文書を対象とする必要がある．本研究ではドメインなどを限定せずに Web を利用することで多様な文書を収集する．多様な文書からなるコーパス

[*2] http://www.tokuteicorpus.jp/

見出し:２００８．０７．１０　Ｔｈｕｒｓｄａｙ
気がつけば梅雨も明けてました。
毎日暑い日が続きますね。
父の手術も無事に終わり、少しだけほっとしてます。
(後略)

図 1　見出しが本文中に出現しない例

見出し:『ミニスカ宇宙海賊』アニメ化決定！
笹本祐一さんの「ミニスカ宇宙海賊」のアニメ化が決定しました。
監督・シリーズ構成は佐藤竜雄、アニメーション制作はサテライトに決まりました。
放映は２０１１年を予定しています。
ご期待ください！

図 2　見出しの要素が先頭 3 文中に出現する例

見出し:売布神社
どもども、森田です。
さてさて、前回中山寺に行きましたが、その続きです。
中山寺から西にぶらぶらと住宅街を歩いていきます。
たぶん、7，8 分ぐらいです。
すると、でかい池が目の前に出てきます。
この池の左上あたりに歩いていくと、売布神社 に着きます。
(後略)

図 3　見出しを除くと意味・談話関係の理解が困難になる例

の作成のためには，1 文書あたりの作業負荷を低くする必要があるので，各文書の先頭 3 文にタグ付けを限定する．本研究で構築するコーパスの規模は 1000 文書とする．

Web に存在する文書にはコーパスとして利用するには不適切なものも多数存在している．これらを全てを人手で確認し，選別することは非常にコストがかかる．Web に存在する文書の数は本研究で目標とするコーパスの規模に比べて遥かに多い．そのため，人手で不適切な文書を確認する前に簡単なルールで自動フィルタリングを行う．さらにフィルタリングの結果残った文書を人手で確認し，コーパスに含めるのに適切な文書についてのみタグ付けの作業を行う．

3.1　意味・談話関係の理解が困難な文書の判定

発話や文書などの言語使用はある場・状況において行われ，場・状況は基本的に話者・著者と聴者・読者の間で共有されている．また，発話や文書の内容は場・状況となんらかの連続性を持っている．

形態素・構文レベルのタグ付きコーパスでは，各文を独立に扱うので，このような場・状況との連続性を考慮する必要はない．しかし，意味・談話関係コーパスにおいては，この問題を考慮する必要がある．本研究ではコーパスとしては基本的にテキストだけを扱うため，例えば，どの Web サイトかという情報がなければ理解しにくい文書はコーパスとしては不適当である．

文書には先頭に見出しを持つものが存在し，場・状況との連続性において重要な役割を持つ場合がある．しかし，見出しは名詞句の連続など文として成立していないものを多く含むため本研究ではタグ付け対象から除く．新聞記事では文書冒頭において全体の要約にあたる文が存在し，見出しを除いても意味・談話関係を理解できるものがほとんどである．Web においては要約の役目を果たす文が存在しない場合があり，見出しを除くと意味・談話関係を理解できないものも存在する．一方でブログにおける日付けなどが見出しになっている場合，見出しを除いても意味・談話関係の理解に影響がないものも存在する．本研究では見出しを除くと意味・談話関係が理解できないような文書はコーパスから除くこととする．

本研究では，文書が見出しをもつかどうかを自動的に判定する．Web には HTML タグなどの構造情報があるが，見出しを指定する <h> タグ以外で見出しが記述される場合があり，一方で <h> タグでマークアップされていても見出しではない場合もある．そこでテキストの内容から見出しの判定を行う．1 文目が句点で終わっていない場合または体言止めの場合に 1 文目を見出しと判定し，それ以外の場合には見出しなしとする．1 文目が見出しの文書の場合には，見出しを除いた後続の 3 文を抽出し，見出しなしの場合には先頭 3 文を抽出する．ただし，見出しを除くと意味・談話関係の理解が困難になると考えられる文書を自動で除去する．

見出し中の語彙が以降の文書中に出現しない場合には，見出しを除いても意味・談話関係の理解に影響を与えないと考えられる．図 1 の例では見出しが日付であり，このような場合には見出しを除いても以降の意味・談話関係の理解には影響を与えない．また, 見出し中の語彙が文書中に出現する場合でも，先頭 3 文中に出現する場合には，意味・談話関係は理解できると考えられる．図 2 の例では 1 文目が要約の役割を果たしており，見出し中の語彙が先頭 3 文に出現している．このような場合には見出しを除いても先頭 3 文の理解は可能であると考えられる．一方で見出し中の語彙が先頭 3 文以外に出現した場合には，コーパスとして利用する先頭 3 文だけで見出しの情報が復元できず，意味・談話関係の理解が困難となると考えられる．図 3 の例では見出しに含まれる「売布神社」が 6 文目に出現している．しかし先頭 3 文には「売布神社」は出現せず，先頭 3 文だけでは「売布神社」に向かうという意味・談話関係の理解が困難である．そこで見出し中の語彙が先頭 3 文以外に出現する場合には見出しを除くと先頭 3 文の意味・談話

表 1　ストップフレーズの例

ボタンを押してください
自動的に移動します
検索できます
ログイン
相互リンク

関係の理解が困難になるとし，自動で除去する．

3.2　タグ付けに不適切な文書の判定

Web から収集された文書には様々なものが含まれる．本研究では以下のようなものはタグ付けが困難であるとして，コーパスに含めない．

理解に専門知識を必要とする　理解に専門的な知識を必要とする文書は作業者が理解できない場合があり，正しいタグ付けが困難である

文章に意味的連続性がない　収集された文書には本来は離れた位置に配置されたテキストを連続したテキストとして抽出してしまったものが含まれる．このような文書は文をまたぐ意味・談話関係のタグ付けができない

過度にくだけた文体で記述されている　過度にくだけた表現は形態素のタグ付けですら困難である

これらを除くために，先頭 3 文の中に以下の要素を含む文書を自動で除去する．

- 体言止めの文：修辞的な文や箇条書きの一部であることが多い
- 句点で終わっていない文：テキストの抜き出し誤りなど非文であることが多い
- 10 文節以上ある：形態素解析の誤りであることが多い
- ローマ字を含む：略語や伏せ字であることが多い
- 表 1 のストップフレーズを含む：自動生成ページや Web 独特の表現を除くため

また，ミラーページや引用ページを除去するために，編集距離が 50 以下の文書があった場合には一方を除去する．

4.　タグ付け内容と基準

4.1　タグ付け内容と手法

本コーパスでは形態素，構文構文，述語項構造，照応関係，固有表現のタグ付けを行う．このうち述語項構造，照応関係が意味・談話関係のタグ付けにあたる．これらの意味・談話関係のタグを付与するためには，タグ付け単位の設定などのために形態素，構文のタグ付けが必要となる．固有表現は意味・談話関係のタグ付けには必要ないが，意味・談話解析の際には重要な手掛かりとなるのでタグ付けを行う．

形態素，構文は京都大学テキストコーパスと同様の基準によりタグ付けを行う．

述語項構造と照応関係のタグ付けの単位として，京都大学テキストコーパスと同様に，基本句という単位を設定する．基本句とは自立語 1 語を核として，前後の付属語を付加したものである．例 (2) に基本句単位での分割の例を示す．述語項構造と照応関係の情報は基本句ごとに付与し，照応関係の照応先も基本句とする．照応先が複合語の場合には，その主辞の基本句を照応先とする．例 (2) では，下線部の「党」の照応先は「国民新党」なので，その主辞である「新党」を照応先としてタグ付けする．

(2)　７月/１７日、/国民/新党/災害/対策/事務/局長と/して、/党 を/代表して/現地へ/向かいました。
(党 ←=:新党)

述語項構造は基本的に京都大学テキストコーパスと同様の基準で付与する．述語項構造の取る項は直接係り受け関係にある項，ゼロ前方照応の項，ゼロ外界照応の項の 3 つに分類される．この内，ゼロ前方照応の項，ゼロ外界照応の項においては，ゼロ代名詞の有無に加え，その照応先も合わせて項の情報として付与する．ゼロ外界照応の照応先を表 2 に示す．不特定同士を区別したい場合には，後ろに整数値を付与し，不特定:人 1，不特定:人 2，のようにする．例 (3) では手術をする人も受ける人も，不特定の人であるが，明確に別の人物である．このような場合に不特定:人 1，不特定:人 2 を照応先としてタグ付けする．

(3)　豊胸/手術 を/ためらう/理由に/痛みへの/不安が/多いようです。
(手術 ← ガ:不特定:人 1，ヲ:不特定:人 2)

京都大学テキストコーパスでは，いわゆる二重主語構文に対するタグ付けとしてガ 2 格を設定し，以下の例のようにタグ付けを行っている．

(4)　彼は/ビールが/飲みたい。
(飲みたい ← ガ 2:彼，ガ:ビール)

京都大学テキストコーパスの基準では，例 (5) では「象が長い」とは言えないので，「象」は「長い」のガ 2 格と扱わないこととなっている．一方，本コーパスでは主題を表す表現の場合にはガ 2 格とすることにし，例 (5) では，「長い」に対して「ガ 2:象，ガ:鼻」というタグを付与した．

(5)　象は/鼻が/長い。
(長い ← ガ 2:象，ガ:鼻)

照応関係のタグ付けは京都大学テキストコーパスに準拠する．京都大学テキストコーパスでは，照応関係を 3 つに分けてタグ付けを行っている．1 つ目は共参照関係にある照応関係である．例 (6) では，下線部の「自分」は前方の「ティーンエージャー」と共参照関係にあるので，「自分」

表 2 ゼロ外界照応の照応先

著者
読者
不特定:人
不特定:物
不特定:状況

表 3 固有表現の種類

ORGNIZATION
PERSON
LOCATION
ARTIFACT
DATE
TIME
MONEY
PERCENT

に対して「=:ティーンエージャー」というタグを付ける．

(6) ティーンエージャーが、/懸命に/ライトセーバーを/振り回している/自分 の/姿を/密かに/ビデオに/収めた。
(自分 ←=:ティーンエージャー)

2 つ目は橋渡し照応のうち名詞の項として「A の B」として表現できるものであり，名詞 B の項として「ノ:A」というタグ付けがされる．例 (7) では，下線部の「相手」では,「ラズナーの相手」と表現できるので,「相手」に対して「ノ:ラズナー」というタグを付与する．

(7) アタマの/先発は/ラズナー、/相手 は/陽と/なっています。
(相手 ← ノ:ラズナー)

3 つ目は共参照関係にない照応関係,「A の B」と表現できないような橋渡し照応であり「　」というタグ付けがされる．例 (8) では，下線部の「語学」の下位概念が 1 文目の「英語」を照応している橋渡し照応であり,「英語の語学」とは表現できないので,「語学」に対して「　:英語」と付与する．

(8) 英語/力を/つけたい/読者の/ために/毎月/さまざまな/学習法を/特集します。
語学 は/モチベーションも/大事。
(語学 ←　:英語)

京都大学テキストコーパスでは照応先は文章内の表現に限定されていたが，本コーパスでは新たに著者，読者への外界照応を付与した．この詳細は 4.2 節で述べる．

固有表現は IREX[*3]の基準に準拠して付与する．固有表現タグは固有表現を表す範囲と固有表現の種類によって表現される．固有表現の種類は表 3 に示す 8 種類である．例 (9) では「ラズナー」に人を表す固有表現である「PERSON」,「ホークス」に組織を表す固有表現である「ORGNIZATION」というタグが付与される．

(9) そこで、ラズナー と ホークス の今季対戦成績を掲載します。
$\begin{pmatrix}\text{ラズナー} \leftarrow\text{PERSON}\\ \text{ホークス} \leftarrow\text{ORGNIZATION}\end{pmatrix}$

タグ付け作業の際にはまず JUMAN[*4]，KNP[*5]で自動でタグ付けを行い，その後 GUI のツールを利用してタグの修正を行った．

4.2 著者・読者表現

談話において文書の著者・読者は特別な要素である．モダリティ，敬語など著者・読者に強く影響される言語現象の存在や著者・読者は省略されやすいなど，著者・読者は他の談話要素と異なった振舞いをする．そのため文書中のどの要素が著者・読者にあたるかは，意味・談話解析において重要な手がかりとなる．

従来の新聞記事を扱ったコーパスでは著者・読者が談話に出現することはほとんどなく，文書中の著者・読者については扱われてこなかった．しかし新聞記事以外の文書では談話に著者・読者が出現することがある．談話に著者・読者が出現する場合でも，著者・読者を示す表現が文書中に出現しない場合がある．例えば前記の図 1 では，談話に著者が出現しているが，著者を示す表現は出現していない．一方，文書中で著者・読者が出現する場合には人称代名詞に限らず様々な表現で出現する．例えば例 (10) の「こま」のように固有名である場合や「主婦」や「母」などのように役職などである場合が存在する．

(10) 東京都に/住む/「お気楽/主婦」/こま です。
$\begin{pmatrix}\text{主婦} \leftarrow\text{=:著者}\\ \text{こま} \leftarrow\text{=:主婦}\end{pmatrix}$
0 歳と/ 6 歳の/男の子の/母 を/してます。
(母 ←=:主婦)

また，日本語の場合には人称代名詞の使用が少なく，照応関係の情報からどの要素が著者・読者にあたるかの同定も困難である[*6]．

談話中に出現する著者・読者の表現をタグ付けするために，文書中の著者・読者の表現に対して外界照応として,「=:著者」,「=:読者」のタグを付与する．著者・読者は文書中で 1 人と仮定し，文書中で「=:著者」,「=:読者」それぞれ最大でも 1 表現にしか付与しないこととする．共参照関係にあり，著者・読者が複数回言及されている場合にはいずれか 1 つに付与することとする．例 (10) では下線部の 3 つの表現が著者を表す表現だが,「主婦」に対して「=:著者」とタグ付けしている．

*3 http://nlp.cs.nyu.edu/irex/NE/df990214.txt
*4 http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
*5 http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
*6 英語であれば “I” などと照応関係にある表現が著者であると推測できる．

企業など組織のホームページではサイト管理者などが組織を代表して文書を記述している場合がある．そのような場合には，その組織が著者であるとしてタグを付与することとする．例 (11) ではサイト管理者が「神戸徳洲会病院」を代表して記述していると考えられるので，その主辞である「病院」に対し「=:著者」を付与する．

(11) 神戸/徳洲会/病院 では/地域の/医療/機関との/連携を/大切にしています。
(病院 ←=:著者)

店舗のページなどでは店舗を表す表現と店員を表す表現が共に出現する場合がある．このような場合には厳密には店員が著者と考えられるが，組織を代表している場合には組織を著者とするという規則を優先し，お店を著者としてタグ付けする．例 (12) では「スタッフ」ではなく「館」に「=:著者」を付与する．

(12) タウン/ロフト/館 の/店舗/情報を/お伝えします。
(館 ←=:著者)
ご来店/予定の/際に/アクセス等で/お困りでしたら、/当店/スタッフ まで/お気軽に/ご連絡下さい。

4.3 曖昧性のあるタグ付けに対する基準

意味・談話関係のタグ付けでは，付与するタグを一意に決められない場合が存在する．本研究では以下のようにタグ付け基準を定め，タグ付けの曖昧性を解消している．

述語項構造では複数の候補がある場合がある．例 (13) では，「買えない」に対して「ガ 2:監督，ガ:サンマ」と「ガ:監督，ヲ:サンマ」という 2 通りのタグ付けが考えられる．このような場合には格助詞の格を優先してタグ付けする．例 (13) の場合，「サンマ」の格助詞のガ格を優先し，「ガ 2:監督，ガ:サンマ」を付与する．

(13) サンマが/買えない/監督の/畑中です。
(買えない ← ガ 2:監督，ガ:サンマ)

係助詞「は」でかかる場合など格が明示されていない場合には，ガ 2 格ではない格を優先する．例えば例 (14) の「ある」では「ガ 2:ココア，ガ:効果」と「ガ:効果，ニ:ココア」という 2 種類のタグ付けが考えられるが，「ガ:効果，ニ:ココア」をタグ付けする．

(14) ココアは/さまざまな/効果が/ある。
(ある ← ガ:効果，ニ:ココア)

複合動詞に述語項構造をタグ付けをする場合には，本動詞に対して付与するか，付属動詞を含む複合動詞に対して付与するかによってタグが異なる．例 (15) の場合，付与するタグは，本動詞に対して付与すると「ガ:あなた」となり，複合動詞に対して付与すると「ガ:私，ニ:あなた」を付与することとなる．本コーパスでは例 (16) の格助詞との一貫したタグ付けの観点から複合動詞全体に対して自然な格を付与する．

(15) 来て頂ければ、/私は/あなたに/会います。
(来て頂ければ ← ガ:私， ニ:あなた)

(16) 私は/あなたに/来て頂ければ/助かります。
(来て頂ければ ← ガ:私，ニ:あなた)

表 4 コーパスの統計

文書数	1000
文数	3000
形態素数	59644
文節数	18905
基本句数	23938
タグが付与された基本句数	14865

表 5 文書ごとの著者・読者の出現

	表現あり	表現なし	出現なし
著者	258	364	378
読者	105	290	605

5. コーパスの統計

現在，3 人の作業者により 1000 文書のタグ付け作業が終了している．タグ付けされたコーパスの統計を表 4 に示す．これより全基本句のうちおよそ半数になんらかのタグが付与されたことが分かる．実際のタグ付けの例を図 4 と図 5 に示す[*7]．コーパスには個人 Web サイト，blog，ニュース記事，自治体の広報ページ，企業の広報ページ，レシピサイトなど多様な文書が含まれる．この中には企業ページ内の広報用 blog のような，一意にジャンル分けができないようなものも存在する．

タグ付けされたコーパスにおける著者，著者の文書ごとの出現数を表 5 に示す．ここで「表現あり」とは文書中に著者・読者にあたる表現があり，外界照応によりタグ付けされている文書の数を表す．「表現なし」とは著者・読者にあたる表現はないが外界ゼロ照応の照応先として出現している文書の数を表す．著者の場合は約 6 割，読者の場合は約 4 割の文書において談話に出現することが分かる．

文書中で著者表現が出現した回数は 358 回，読者表現が出現した回数は 134 回であった．その例と出現回数を表 6 と表 7 に示す．著者表現では，「私」が 63 回と多いが，「管理人」「主婦」「監督」などの立場を表す表現や「協会」「病院」などの組織を表す表現，「こま」「カーブス」など固有名など多様な表現で出現することが分かる．またコーパス

[*7] 表層の表現では曖昧性がある場合には説明のため添字を付与した．

表 6 著者表現の例

著者表現	出現回数
私	63
弊社	12
店	10
会	10
当社	9
自分	8
当店	6
管理人	5
協会	3
病院	3
主婦	2
監督	1
カーブス	1
こま	1

表 7 読者表現の例

読者表現	出現回数
皆様	28
客	24
あなた	23
方	9
自分	8
人	7
皆さん	6
会員	5
自身	3
患者	2
読者	1
生徒	1
贈り主	1
市民	1

表 8 ゼロ照応の個数

	文章内ゼロ照応	外界ゼロ照応	合計
ガ格	1703	2488	4191
ヲ格	594	100	694
ニ格	409	388	797
ガ２格	72	116	188
合計	2778	3092	5870

全体で 1 度しか出現しなかった表現が 96 表現，2 度しか出現しなかった表現が 24 表現と，文脈により著者表現となるものが多い．読者表現では二人称代名詞の「皆様」に次いで「客」が多い．これは Web ページで読者を想定するのは企業ページなどであることが多いためと考えられる．また，「生徒」「贈り主」「市民」など文書特有の読者を想定する表現も見られる．著者，読者両方の表現で用いられるものとしては「自分」が見られた．

タグ付けされたゼロ照応の個数を表 8 に示す．表 8 から特にガ格においてゼロ照応が多く，その約 6 割が外界ゼロ照応であることが分かる．またヲ格とニ格ではゼロ照応の数には大きな差はないがニ格ではヲ格に比べゼロ照応の割合が高いことが分かる．また，文章内ゼロ照応の照応先の内訳を表 9 に，外界ゼロ照応の照応先の内訳を表 10 に示す．表 9 で著者，読者とは，ゼロ代名詞の照応先が著者，読者と共参照であることを表す．表 9 と表 10 から，ガ格のゼロ照応の照応先のうち著者が約 1/3，読者が約 1/6 を占めていることが分かる．一方，ニ格においては外界ゼロ照応において照応先として著者よりも読者が多いことが分かる．ヲ格の照応先としては著者，読者ともに少なく，外界ゼロ照応においては不特定:人と不特定:物が約 8 割を占めた．

タグ付けされた照応関係を表 11 に示す．また，照応先の内訳を表 12 と表 13 に示す．表 11 から照応関係においては種類を問わず文章内照応が多くを占めることが分かる．また「　」よりも「ノ」が多いことから，橋渡し照応の多くは「A の B」の形に言い換えるられることが分かる．表 13 から不特定:状況が照応先とならないことが分かる．

表 9 文章内ゼロ照応の内訳

	著者	読者	その他	合計
ガ格	602	176	925	1703
ヲ格	8	4	582	594
ニ格	78	44	287	409
ガ２格	23	8	41	72
合計	711	232	1835	2778

表 11 照応関係の数

照応の種類	文章内照応	外界照応	合計
＝	2201	363	2564
ノ	3185	201	3386
	757	43	800
合計	6143	607	6750

表 12 文章内照応の内訳

	著者	読者	その他	合計
＝	100	29	2072	2201
ノ	256	96	2833	3185
	31	24	702	757
合計	387	149	5607	6143

6. まとめ

本研究では Web を利用することで多様な文書からなる意味・談話関係タグ付きコーパスを構築した．本研究では意味・談話関係のタグとして，述語項構造と照応関係の付与を行った．また，文書の著者・読者に着目し，その表現に対してタグ付けを行った．タグ付けを先頭 3 文に限定することで 1 文書あたりの作業量を減らし，1000 文書へのタグ付けを行った．構築されたコーパスを分析した結果，多くの文書において談話に著者・読者が出現し，多様な表現で記述されること，また特にゼロ照応において重要な役割を持つことを確かめた．

謝辞 本コーパスのタグ付け作業に協力していただいた，石川真奈見氏，二階堂奈月氏，堀内マリ香氏に心から感謝致します．

参考文献

[1] L. Rello and I. Ilisei. A comparative study of spanish zero pronoun distribution. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, pp. 209–214, 7 2009.

[2] Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. Anaphoric annotation of wikipedia and blogs in the live memories cor-

表 10 外界ゼロ照応の内訳

	著者	読者	不特定:人	不特定:物	不特定:状況	合計
ガ格	930	637	734	95	92	2488
ヲ格	3	9	32	52	4	100
ニ格	66	153	140	27	2	388
ガ２格	43	44	25	4	0	116
合計	1042	843	931	178	98	3092

表 13 外界照応の内訳

	著者	読者	不特定:人	不特定:物	不特定:状況	合計
=	258	105	0	0	0	363
ノ	95	52	28	26	0	201
	16	18	4	5	0	43
合計	369	175	32	31	0	607

ワイヤー$_1$・/フォックス$_1$・/テリア$_1$は/数多い/テリア$_2$/犬$_1$/種$_1$の/中で/世界的に/最も/知られている/犬$_2$/種$_2$である。

テリア$_2$← :テリア$_1$
犬$_1$← 修飾:テリア$_2$
種$_1$← ノ:犬中 ← ノ:種$_1$
知られている ← ガ:種$_2$，デ:中
種$_2$ である ← ガ:テリア，ノ:犬$_2$，=:テリア$_1$

典型的な/テリア$_3$の/気質を/現代に/伝える/「イギリス/犬$_3$/種$_3$」と/して/人気が/高い。

典型 ← ガ:テリア$_3$
テリア$_3$←=:テリア$_2$
気質 ← ノ:テリア$_3$
伝える ← ガ:種$_3$，ヲ:気質，ニ:現代
種$_3$←=:テリア$_1$
高い ← ガ２:テリア$_1$，ガ:人気

ワイヤー$_2$・/フォックス$_2$・/テリア$_4$は/穴に/潜む/小型/害/獣の/駆除/犬$_4$であったが、/特に/狐/狩りで/活躍した。

テリア$_4$←=:テリア$_1$
潜む ← ガ:獣，ニ:穴
小型 ← ガ:獣
獣 ← 修飾:害
駆除 ← ガ:犬$_4$，ヲ:獣
犬$_4$← ガ:テリア$_4$，=:テリア$_4$
狐 ←=:フォックス$_2$， :獣
狩り ← ガ:不特定:人，ガ:テリア$_4$，ヲ:狐
活躍した ← ガ:テリア$_4$，デ:狩り

図 4 タグ付け例 1

ハピ/猫$_1$には/現在/１６匹の/猫$_2$/スタッフ$_1$が/みなさまの/ご来店を/お待ちしています。

猫$_1$←=:著者
１６匹 ← 時間:現在
猫$_2$← :猫$_1$
スタッフ$_1$← 修飾:16 匹，修飾:猫$_2$，=:16 匹
みなさま ←=:読者
ご来店 ← ガ:みなさま，ニ:猫$_1$
お待ちしています ← ガ:スタッフ$_1$，ニ:猫$_1$，ヲ:ご来店

猫$_3$/スタッフ$_2$は/体調/管理の/ため、/ローテーションで/お店に/出ています。

スタッフ$_2$←=:スタッフ$_1$
体調 ← ノ:スタッフ$_2$
管理 ← ガ:著者，ヲ:体調
ため ← ノ:管理
お店 ←=:猫$_1$
出ています ← ガ:スタッフ$_2$，ニ:お店，デ:ローテーション

日に/よって/お休みの/子も/いますので/ご了承くださいませ。

お休み ← ガ:子，ヲ:お店
子 ←=:スタッフ$_2$
います ← ガ:子
ご了承くださいませ ← ガ:読者，ヲ:います

図 5 タグ付け例 2

pus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 5 2010.

[3] 小原京子. 日本語フレームネットの全文テキストアノテーション: Bccwj への意味フレーム付与の試み. 言語処理学会 第 17 回年次大会, pp. 703–704, 3 2011.

[4] 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, pp. 495–498, 3 2002.

[5] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関 係のアノテーション: Naist テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, 4 2010.