

日本語ゼロ照応関係に対する特徴分類とそのアノテーション

飯田龍

東京工業大学 大学院情報理工学研究科
ryu-i@cl.cs.titech.ac.jp

笹野遼平

東京工業大学 精密工学研究所
sasano@pi.titech.ac.jp

1 はじめに

ある表現が同一文章内の他の表現を指す場合、二つの表現は照応関係にあるという。照応関係にある表現のうち、指す側の表現を照応詞、指される側の表現を先行詞と呼ぶ。また、ある表現の集合が同一実体を指す場合にそれらは共参照関係にあるという。この照応・共参照関係を自動的に解析する処理は言語理解のための重要な問題の一つとして古くから研究が進められている [5]。照応・共参照解析の研究では、性・数の一致や談話要素の顕現性などの言語的な特徴を利用した人手により規則作成に基づく解析手法に加え、照応関係がアノテーションされたコーパス（タグ付きコーパス）を利用した学習ベースの手法が提案されており、他の言語処理の問題と同様に、学習・評価用のタグ付きコーパスを開発し、そのタグ付きコーパスに基づいて研究が進められている [6]。

照応関係もしくは共参照関係は上述の定義に従うが、厳密には照応・共参照関係として認定すべき談話要素やその関係の種類に関してコンセンサスが得られていないため、形態素・係り受け解析などの自然言語処理の基盤処理と比較して、多様な問題設計が行われている。例えば、Message Understanding Conference (MUC)¹で提供された英語の共参照関係タグ付きコーパスでは、共参照解析の問題を情報抽出の部分問題と捉え、一般には共参照関係とは判断されないような同格の関係を共参照関係として扱っている。また、MUCの後継に当たる Automatic Content Extraction (ACE)²では、情報抽出の結果として出力したい表現のクラス（人や組織など）に限定して共参照関係をアノテーションしている。これらの問題設定は共参照解析の出力結果を以降の処理に利用することを前提としているため、定義された共参照関係は必ずしも人間の直感を反映しているとは限らず、また他の応用処理で必要となる関係が含まれているとは限らないため、言語の理解の問

題を解くという観点で考えると必ずしも適切な問題設定になっているとは限らない。

これに対し、日本語を対象とした照応関係のアノテーションでは、主に省略された照応詞（ゼロ代名詞）を検出し、そのゼロ代名詞に関して先行詞をアノテーションするという試みがなされている。日本語では読み手（聞き手）の推論可能な照応詞は頻繁にゼロ代名詞として省略されるため、ゼロ代名詞を検出し、そのゼロ代名詞の先行詞を同定するゼロ照応解析がさまざまな応用処理に必須な処理である。例えば、情報抽出や含意関係認識などでは、文を越えて出現する談話要素（名詞句など）を述語項構造の形式でまとめあげることで、述語項構造を単位とした非常に見通しのよい処理を想定できるが、この述語項構造解析のためにゼロ照応解析は必須な要素技術となる。

ゼロ照応解析の問題設定では、動詞などの述語の項が省略される場合にその項をゼロ代名詞として検出するため、ゼロ代名詞の先行詞の種類を問わず照応関係をアノテーションすることになる。つまり、このアノテーションの手続きを採用することで、照応詞が省略されている場合に関しては漏れなく照応関係をアノテーションできることになる。これにより、ACEで共参照関係とみなす談話要素の意味クラスを限定しているのに対して、日本語のゼロ照応関係のアノテーションでは照応関係の現象を包括的に捉える問題設定が可能となる。

ただし、ゼロ照応関係の場合は照応詞が省略されるために、性・数の一致情報などの手がかりが利用できないため、英語などの明示的に照応詞が出現し、かつ解析に役立つ情報を含んでいる場合と比較して自動解析が困難な例が多数出現する。このため、アノテーションされたすべての問題を対象に自動解析を試みた場合、解析精度が極端に低くなることが報告されている。例えば、ゼロ照応関係が網羅的にアノテーションされた NAIST テキストコーパス中の述語のガ格の省略を対象にゼロ照応解析を行った場合、F 値で約 4 割の解析精度しか得られていない [3]。このように、既

¹http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

²<http://www.itl.nist.gov/iad/mig/tests/ace/>

存のゼロ照応関係のアノテーションでは、英語の照応関係と比較して多様な問題を含むようなアノテーションの方法を採用しているがために、英語の照応・共参照解析の結果と比較して不当に低い解析精度を得ているように見える。英語の場合と同様に、解析対象を限定する方向性も考えられるが、それでは言語理解の問題として本来解かなければならないと考えられていた現象の一部しかコーパスにアノテーションされないことになる。

本研究では、どのような言語的な特徴がゼロ照応解析を難しくしているかを調査するために、これまで人手で問題分析してきた経験を生かし、ゼロ照応関係の特徴分類を行う。さらに、その分析結果に基づいて、すでにアノテーションされたゼロ照応関係についても特徴を分類する予定である。このアノテーション作業を通じて、そもそも解析対象とするコーパスにどのような問題がどのくらい出現し、今後何に注力して解析するモデルを考えなければならないかを明らかにする。以降では、まず2節でゼロ照応に関する既存のアノテーションの仕様を紹介し、次に3節で提案するゼロ照応の分類を説明する。さらに、この作業を行う上での作業体制について説明し、最後に5節でまとめる。

2 関連研究

日本語を対象にゼロ照応関係をアノテーションする試みは、文章中の述語に網羅的に述語とその項の関係をアノテーションする作業の中ですでに実現されている。これらのアノテーションの結果と文の係り受け構造を照合することで、項としてアノテーションされているが、直接の係り受け関係にない述語と項をゼロ照応の関係とみなす。

アノテーションの際の述語と項の関係としてはおおきく意味役割レベルのアノテーションと表層格レベルのアノテーションの2つが採用されている。例えば、GDA タグ付きコーパス [2] では動作主、主題など意味役割レベルのアノテーションされており、一方、NAIST テキストコーパス [9] や京都大学テキストコーパス [8] では、表層格レベルでアノテーションされている。意味役割に基づくアノテーションでは、動作主などの理解に関する意味のある粒度で項の関係を定義できる自由度があるが、逆にそれらをどう設計すれば一貫性のあるものになるかを考えるのが難しい。一方で、表層格に基づいてアノテーションする場合は表層格として述語が何を取るかを考えればよいため、実例もしくは人間の内省に基づいて項構造を考えることが

できるが、動詞ごとにどの格要素が必須となるかが異なるため、必須と任意の判断が難しく、また自他交替現象に関して格パターンが複数存在する場合、どのパターンを選択するかで揺れが生じるという問題も起こる。

また、京都大学テキストコーパスでは述語の出現形に対し項をアノテーションするのに対し、NAIST テキストコーパスでは述語の原形を対象に項をアノテーションするという違いも存在する。このため、例えば、(1)で、京都大学テキストコーパスでは「陥落させる」の項をアノテーションするのに対し、NAIST テキストコーパスでは原形の「陥落する」に対してアノテーションを行うことになる。

(1) a. 京都大学テキストコーパス:

薩摩軍₁ は 熊本城₂ を 陥落させる {ガ:1, ヲ:2} ことができなかった。

b. NAIST テキストコーパス:

薩摩軍₁ は 熊本城₂ を 陥落さ {ガ:2} せることができなかった。

このため、前者の場合は主題の連鎖を陽に考慮しやすく、一方後者のアノテーションの方法では動詞の選択選好の情報を利用しやすいという利点がある。このように、アノテーションの基準次第で解析に利用する情報の導入のしやすさが異なるため、この点をどのように問題設計するかは直接的に見た目の解析精度に影響する。例えば、NAIST コーパスではその仕様上、受動態で出現している場合は動詞の原形に戻して項を付与するが、この場合は前方の談話から続く主題連鎖の手がかりと矛盾する出力を行う必要があり、その点を反映した学習は難しいため、受動態を伴って省略が出現する場合は頻繁に解析を誤る傾向にある。

また、どの格が省略されるかも問題の性質を左右する。NAIST コーパスに出現する頻度のうち、ゼロ照応関係を表層格ごとに分けて数えた結果を表1に示す。表1を見ると、アノテーションされたゼロ照応のうち、ほとんどがガ格の省略であることがわかる。つまり、ゼロ照応解析の問題として解かなければならない問題の多くは、省略されたガ格を文脈からいかに補完するかを考える問題ということになる。ただし、ヲ格やニ格の省略が選択選好に基づいて補完すべき先行詞の候補を限定できるのと比較して、ガ格の省略はこの選択選好に関する候補の絞り込みが難しく、また省略が起きる場合にはさまざまな言語現象が関与しているため、選択選好などの単一の知識に基づいて照応関係を説明できない場合が多い。このため、照応解析には複雑な知識の組み合わせ、また推論などの機構を利用する必要があるが、どのような知識や機構が解析に

表 1: NAIST テキストコーパス中のゼロ照応関係

出現箇所	ガ格	ヲ格	ニ格
ゼロ照応 (文内)	32,270 (0.526)	5,625 (0.800)	1,417 (0.707)
ゼロ照応 (文間)	13,181 (0.215)	1,307 (0.186)	542 (0.270)
ゼロ照応 (文章外)	15,885 (0.259)	96 (0.014)	45 (0.022)
全体	61,336 (1.000)	7,028 (1.000)	2,004 (1.000)

どの程度影響するかについては明らかにされておらず、これを詳しく分析する必要がある。そこで、多様な現象が混在するゼロ照応解析において今後何に注力して研究すべきかを明かにするために、本研究では3節で導入するゼロ照応関係の特徴に基づきアノテーションを行い、各特徴の出現頻度やその解析時の難易度について細かく分析することを考える。

3 ゼロ照応関係の特徴分類

これまでに行ったアノテーション作業やその仕様策定の経験に基づいて、ゼロ照応関係、特にガ格の省略の分類を行う。分類のために、以下に示す6つの特徴を各ゼロ照応関係にアノテーションすることを考える。このアノテーションを行うことで、各特徴がどのくらい出現するか、また、ある特徴が他のどの特徴と共起するかを調査することが可能になる。以降でそれぞれの特徴について、例を示しながら説明する。

1) 談話要素の顕現性を考慮した遷移 日本語に関するセンタリング理論 [1] に基づくゼロ照応関係の分析 [7] では、談話要素の顕現性の強さはその表現が伴う助詞と関連付けられており、典型的には「は (主題) > が (主語) > に (間接目的語) > を (直接目的語) > その他」の順で顕現性が高いと説明されている。つまり、顕現性の高い談話要素ほど以降の文脈で省略される傾向がある。

例えば、下記の例³では、1文目に出現する「小沢一郎幹事長」が係助詞「は」で主題化されているため、次の文で同一の表現を主節の主語として生成する際にゼロ代名詞「 ϕ_1 」⁴が用いられている。同様に、3文目でも同一の表現が主節の主語となるため、省略されたままである。

³以降で説明の際に利用する例は毎日新聞 95 年の記事の抜粋であり、紙面の都合上、必ずしも文章の最初からは掲載していない。また、説明のために一部編集している場合もあることに注意されたい。また、もとの記事で改行されていない場合でも文の区切りがわかりやすくなるよう改行を追加している。

⁴以降で、ゼロ代名詞は ϕ で表し、また対応する下付き文字を物名詞句が先行詞であることを表す。

新進党の**小沢一郎幹事長**₁は十五日午前、当地で開かれた「アジア・太平洋指導者円卓会議」で基調報告。

(ϕ_1 ガ) 昨年の新進党結成の背景や目的などを説明した。

(ϕ_1 ガ) 新進党について、その性格を「生活者に重きを置き、都市に住む人の気持ちをしっかりとつかんだ政党だ」と解説。

このように、日本語のゼロ照応ではゼロ代名詞の直前の主題化された表現が先行詞となる場合が多く、この場合は容易に解析可能であると考えられる。また、上述の顕現性の強さの序列に基づき、主題化された表現が前方文脈に出現しない場合は主語「が」が先行詞となることが知られており、その場合もこの特徴として分類する。また、4) の埋め込み構造の場合と区別するため、ゼロ代名詞とその先行詞は地の文に出現している場合に限定する。

2) A の B の関係を理解 次に、名詞句、特に「A の B」のボタンで出現する場合に、B が主題化もしくは主語として出現しており、後続文脈で出現するゼロ代名詞の先行詞が A である場合について説明する。1) の例では主題化もしくは主語として出現した談話要素そのものが先行詞となっているが、次の例では2文目の「使う」のゼロ代名詞の先行詞は1文目の「民主化政策」ではなく、その名詞句を修飾する「ロシア」となっている。

ロシア₁の民営化政策は**チュバイス**氏の指導で九二年十月から始まった。

(ϕ_1 ガ) 民営化証券を**使って**株式を取得するという第一段階は九四年六月に終了し、現在は現金で株を購入できる第二段階に入っている。

この現象を解釈するためには「A の B」の A と B がどのような関係で記述されているか、またゼロ代名詞と関連する述語の選択選好などの知識を利用することが考えられる。

3) **選択選好に基づく理解** 2) が「A の B は/が」のように、主題化された、もしくは主語で出現する名詞句の中に出現している A が先行詞であるのに対し、この特徴では動詞の選択選好に依存して解釈が可能な場合を考える。次の例では、「結成される」のガ格には典型的に政党などの組織が出現するという知識に基づいて解釈可能である⁵。このような知識に基づいて、この文章では「新進党」が「結成され(る)」の先行詞となることがわかる。

現職の党首は脱落し、羽田孜副党首と小沢一郎幹事長による事実上の一騎打ちに――。
新進党₁の党首公選から目が離せなくなってきた。今回の党首選挙は、(φ₁ ガ) 昨年十二月十日に結成されて以来、初めての公選である。

4) **埋め込み構造・引用内に先行詞が出現** 1)~3) で示した例では主に主節に談話要素が出現しているが、これに対して先行詞が連体修飾節の中や引用節の中に出現する場合がある。談話要素が連体修飾節や引用節に出現する場合は、格助詞に基づく談話要素の顕現性の序列は必ずしも先行詞同定に反映されない。例えば、下記の例で 4 文目の引用の中に出現する動詞「支援する」の先行詞は直前で主題化された「大木田守・政調会長」ではなく、1 文目に出現している「東京都」である。

最も苦しい立場に立たされるのは、**東京都**₁。
監督官庁としての責任があり、(φ₁ ガ) 不良債権処理を目的とした債権回収機関に、約二百億円の収益支援をしなければならぬ。
都議会は二月七日に開会予定だが、支援策を盛り込んだ議案に対し、与党の公明党や社会党が厳しく追及する構えだ。
都議会公明の大木田守・政調会長は「(φ₁ ガ) この二組合だけに支援する必然性がない。支援にはきちんとした基準、ルールが必要だ」と批判している。

また、次の例では 2 文目の「明言する」のガ格が省略されており、先行詞は 1 文目のターノフ米国務長官の発話内の「米国」が先行詞となっている。この例では発話内で主題化されている談話要素が先行詞となっているが、他の例ではその発話の発話者（この例では、ターノフ米国務長官）が後続文脈に出現するゼロ代

名詞の先行詞となる場合もあり、その判断は容易ではない。

中国側によれば、最近訪中したターノフ米国務次官は「最近の出来事によって**米国**₁は台湾問題の重要性と敏感さへの理解を深めた。台湾問題についての中国の立場を重視し尊重する」と表明した、とされる。
(φ₁ ガ) 中国が要求しているように「李総統の私的訪米を二度と認めない」と公式には明言できないが、今後は慎重に対応するという意思表示だ。

このように文章中で誰かが発話した内容が引用される場合、その文脈で発話される意図や後続文脈とのつながりが理解できてはじめて、埋め込まれた構造に出現するゼロ代名詞（もしくは先行詞）との関係が理解できると考えられる。このため、この種の現象は他のゼロ照応関係とは区別して扱う必要がある。

5) **大域的な主題として捕捉** これまでに述べてきた内容は談話の局所的な把握に関してゼロ照応関係をどう考えるかという特徴に関してだが、一方で談話全体でどのような内容が書かれているかという大域的な主題を把握し、それを文脈に応じて補完する必要がある場合がある。

例えば、下記の例では 5 文目と 6 文目では「宗教活動に熱心な人々」という総称的な対象に関する言及が挿入されており、次の 7 文目では「ラルフ・リード氏」に関する話題に立ち返り、以降では「ラルフ・リード氏」に関する記述が続く⁶。つまり、この文章に対するゼロ照応解析では、「ラルフ・リード氏」の政治活動に関する文章であることをあらかじめ捉え、「ラルフ・リード氏」が離れた位置にしか出現していない場合でも文脈に応じて先行詞として補完する機構が必要となる。

⁵ここでは「結成する」という動詞の原形と「結成される」という受動態に関してそれぞれ異なる格パターンで項が出現し、それぞれにどのような表現が入り得るかを把握できていると仮定する。

⁶紙面の都合上、8 文目以降の文は省略する。

「まだ大学生のようにフレッシュ」と言われる童顔と、他人に警戒心を抱かせないような柔らかな語り口が、この若々しい男の最大の武器だ。

ラルフ・リード氏₁。

(ϕ_1 ガ) 共和党保守派最強の集票マシンといわれる「クリスチャン・コアリション」のリーダーとして、昨年の中間選挙を踏み台に反リベラル運動の最前線に躍り出てきた。

タイム誌の年末特集では、(ϕ_1 ガ) 将来の米国の指導者五十人の一人に選ばれている。

キリスト教の国である米でも、宗教活動に熱心な人々₂はこれまで政治の世界では(ϕ_2 ガ) 異端視されることが多かった。

(ϕ_2 ガ) (ϕ_2 ガ) 神学的価値観を持ち込むのでは、と警戒されたためだ。

しかし、(ϕ_1 ガ) 「日曜日に欠かさず教会に行き、毎日聖書を読んで暮らしている人間が、なぜ二級市民に甘んじなければならないのか」と声を上げる。

なると考えられる。この例はそれが顕著に見える例だが、実際にはより高度な推論を行うことで関係が理解できる場合も存在し、これの事例は知識獲得とその応用のよい例題となると考えられる。

いつもは動きすぎて墓穴を掘ることの多い琴錦₁が、この日は一気の出足だけを心に決めてかかっていた。逆に迷ったのは若乃花₂の方だった。

...

「(ϕ_2 ガ) 考え過ぎてしまった。(ϕ_2 ガ) 相手の足首のことは忘れよう、忘れよう 思ったんだけどね。それが中途半端な立ち合いにつながった。

いつも相手の機先を制する若乃花₂なのに、(ϕ_2 ガ) フワッと 立って、(ϕ_2 ガ) 胸で 受けてしまった。そこへ (ϕ_1 ガ) 最高の出足で 突っ込み、(ϕ_1 ガ) 左右を 差す。

あとは (ϕ_1 ガ) 右からおっつけるような形で一直線に前へ。

6) 常識的な知識に基づいた推論が必要 ゼロ照応関係の理解で機械での解析が最も困難な事例は、常識的な知識やその知識を利用した推論が必要となる場合である。例えば、次の例のように「香港」が中国に返還されるという歴史的な背景を知らなければ、適切に「返還される」の省略を補完することはできない。これは、ある用語に関する知識を収集し、それを適切に運用できるかという点で評価されるべきである。このため、少なくとも 1) の主題連鎖に基づく現象とは区別されるべきである。

香港₁の輸出入の拠点、ビクトリア港を隔てた九龍半島と香港島の両側でいま、埋め立て工事が急ピッチで進められている。

(ϕ_1 ガ) 中国に 返還される 一九九七年に開港予定の新空港関連の鉄道や道路、商業用地を確保する工事で、これらを含めたプロジェクトの総額は二兆円にのぼる。

上述の例は知識との照合に関する例だが、照合だけでは理解が困難な事例が存在する。例えば、下記の例は相撲の取組の説明だが、「いつも相手の～」の文で「若乃花」が「フワッと立って、胸で受けてしまった」後で、「そこへ～」の文から「最高の出足で突っ込み、左右を差す」行為を行ったのが「琴錦」であることを理解するためには、この状況が明確に理解でき、ある行為を行った場合にそれと対となる他の行為は他者が行うといった複雑な知識とそれに基づく推論が必要に

これら 6 種の特徴の説明からわかるように、この 6 種の特徴は排他的ではない点に注意されたい。例えば、4) の埋め込み構造の特徴は 2) の「A の B は/が」のパターンや 5) の大域的な主題を考慮しなければならない場合と同時に出現可能であり、また、3) の選択選好を用いて解析できる場合は、2), 4), 5) と同時に起こり得る。このため、これらの特徴をアノテーションする場合は複数の可能性を考慮しながら、作業を行う必要がある。

4 作業体制

3 節で示したゼロ照応関係の特徴をアノテーションするために、現代日本語書き言葉均衡コーパス⁷の一部を対象に作業を行う予定である。このコーパスには書籍、新聞、白書、blog などさまざまな種類の記事が含まれているため、これらを対象にゼロ照応関係をアノテーションすることで、新聞記事にのみアノテーションする場合と比較して多様な種類の照応の現象が収集できると考えられる。

このうち、ゼロ照応関係を含む述語項構造のアノテーションに関しては奈良先端大学院大学で作業が進行中であり、この結果を利用したアノテーションを行うことを考えている。ただし、このまま利用すると述語と項が係り受け関係にある場合と区別できないため、

⁷<http://www.tokuteicorpus.jp/>

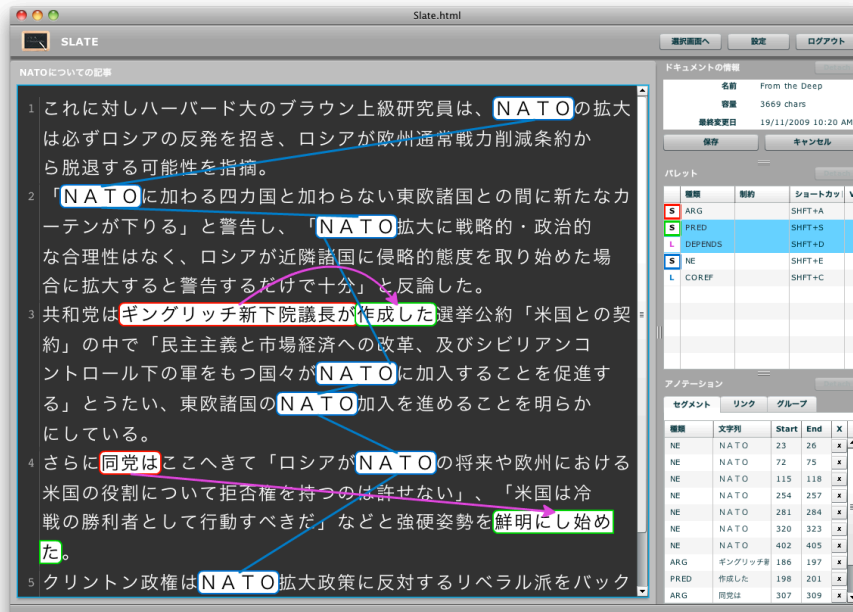


図 1: Slate のアノテーション作業画面の例

Cabocho⁸や KNP⁹などを利用した係り受け解析の結果と照合し、どの場合がゼロ照応関係にあるかという情報を付加する。

アノテーション作業では、アノテーションツールとして Slate[4] を利用する。Slate では名詞句や述語に関するセグメンテーション、またそれらの間の述語と項の関係を直感的な操作でアノテーション可能である。Slate を用いた作業の例を図 1 に示す。この例では、文章中で「NATO」は青色の線分で関連付けられており、これらが共参照関係にあることを表す。また、「作成する」と「ギングリッチ新下院議長」、「鮮明にし始める」と「同党」はそれぞれ述語とガ格であることを矢印で関連付けられている。このようにゼロ照応関係を Slate に表示することで、ゼロ照応関係に関連した述語項構造関係を概観することが可能になる。この状況で作業者は表示されたゼロ照応関係のそれぞれに対し、3 節で定義されたゼロ照応関係の特徴のどれに該当するかのアノテーションを行う。

5 おわりに

本稿では、ゼロ照応関係の特徴を分類し、それらを実データにアノテーションする研究の方向性について提案した。3 節で示したように、ゼロ照応の問題は顕現

性に基づく主題の遷移から常識的な知識に基づく推論まで、多様な現象が関連しており、もし揺れなくそれらを分類することができれば、それぞれの特徴ごとにどう解けばよいかを区別して考えることができる。本稿では、ゼロ照応の関係を 6 種類の特徴に分類したが、4 節に示した広範な種類の文章にアノテーションをした場合、ここで示した分類で網羅できているとは限らない。このため、今後はアノテーション作業を通じて分類そのものについても洗練する予定である。

参考文献

- [1] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–226, 1995.
- [2] K. Hasida. 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>.
- [3] R. Iida and M. Poesio. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of ACL-HLT 2011*, pp. 804–813, 2011.
- [4] D. Kaplan, R. Iida, K. Nishina, and T. Tokunaga. Slate - a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, Vol. 26, No. 2, pp. 89–101, 2012.
- [5] R. Mitkov. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education, 2002.

⁸<http://code.google.com/p/cabochoa/>

⁹<http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP>

- [6] M. Poesio, S. Ponzetto, and Y. Versley. Computational models of anaphora resolution: A survey. Unpublished.
- [7] M. Walker, A. K. Joshi, and E. Prince (eds.). *Centering Theory in Discourse*. Oxford Univ. Press, 1997.
- [8] 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.
- [9] 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, 2010.