

# コーパスへの意味・談話情報のアノテーション

福原裕一 乾健太郎  
東北大学 大学院情報科学研究科

## 1 はじめに

情報抽出や質問応答、複数文書要約や情報分析など、大量の文書集合から特定の種類の情報を抽出し、抽出した情報を照合、集約するといった言語情報編集を計算機で実現するためには、形態素、統語解析に加えて、意味解析、談話解析と呼ばれる、言葉の意味に踏み込んだ処理が必要である。語義曖昧性解消、固有表現抽出、照応・共参照解析、述語項構造解析、モダリティ解析、談話関係、時間解析などがそれに当たるが、こうした技術の研究を推し進めるためには、実際のテキストに個々のレイヤの統語情報・意味情報を注釈付けしたコーパスを構築し、技術の開発・評価用のデータとして研究者間で共有することが不可欠である。

コーパスに何らかの注釈付けを行うは、まず注釈情報の仕様を設計する必要がある。統語情報では、例えばどのような品詞体系を採用するか、統語構造の表現方法はどうかといった取り決めが必要である。意味情報についても、どのような範囲のものを固有表現と認めるのか、どのような場合に照応詞と先行詞の関係が成り立つのか、述語の項を何種類に分類するかといった取り決めを合理的に行う必要がある。これによって言語の解析という漠然とした目標が具体的な部分タスクに切り分けられることになる。すなわち、注釈付けの仕様を論じることは言語処理がどんな問題を解くべきか論じることであり、極めて重要な意味を持っている。

意味・談話情報の注釈付けは、形態素・統語情報に比べて大きく遅れていたが、近年急速な発展を見せている。冒頭で挙げた語義曖昧性解消、固有表現抽出、照応・共参照解析といった個々のレイヤにおいてタグ付きコーパスが構築され、それらを利用した解析器の研究も進みつつある。とくに、研究リソースが集中する英語についてはその傾向が顕著であり、言語学サイドからも例えば格文法のFillmore や生成語彙論のPustejovsky といった著名な研究者らが意欲的に仕様設計に携わり、工学系研究者だけでは困難な理論的基盤の構築が進んでいる[3, 11, 28, 29, 31, 30]。

また、複数のコーパスの注釈情報を統合する動きも活発になってきた。とくに、同一の文書集合に対して、統語情報の上に共参照、述語項構造、談話関係など様々なレイヤの意味情報を重層的に付与する試みがすでにいくつか報告されていることは注目すべきである[28, 11, 26]。

OntoNote[26]などはその代表的な一例で、TreeBank (統語情報) やPropBank (述語項構造) , WordNet (動詞や名詞のポリセミー) やcoreference (共参照)などのレイヤが重層的に付与されており、それぞれのレイヤ間における注釈の整合性は90パーセント以上であると言われている。意味情報が重層的に付与されたコーパスは、より総合的な意味情報付きコーパスと見なすことができ、意味談話解析全体の設計をより広い視点から考察するためにも有益である。またレイヤ間の注釈の整合性を分析することによって、仕様の洗練にも繋がると期待されている[28]。

本稿では、こうしたコーパスへの統語・意味情報の注釈付けについて、これまで作られてきた言語資源を軸に、述語項構造、照応・共参照、談話関係についての先行研究を概観し、各情報の注釈付けで検討すべき課題をまとめる。

## 2 照応・共参照

文章内には多くの照応が存在している。では、どのようなものを照応というのか。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。一方、二つ（もしくはそれ以上）の表現が現実世界あるいは仮想世界において同一の実体を指す場合、それらの表現は共参照（あるいは同一指示）の関係にあるという。例えば、

(1) の“横尾 *i*”と“彼 *i*”は照応関係であり、かつ共参照でもある。

- (1) 横尾 *i* は画家でもないし、デザイナーでもない。そんなことは彼 *i* にとってはどうでもよいことなのだ。

一方、(2) の“iPod *i*”は“それ *i*”と照応関係にあると解釈できるが、共参照ではない。

- (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、照応関係と共参照関係は似てはいるが同じではないので、注意が必要である。

文献[23]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。

また照応詞の現れる位置によっても照応は分類される。例(3)のように、照応詞が現れる前に先行詞が既に表れている場合、これを前方照応 (**anaphora**) と言い、逆に(4)のように、先行詞が照応詞よりも後に現れるものを後方照応 (**cataphora**) と言う。前方照応、後方照応ともに文章中において起こる現象であることから、これらはまとめて文脈照応 (**endanaphora**) と呼ばれる。

- (3) 太郎は ipod を買って、それを次郎に貸した。  
(4) これってもしかして？ 今流行ってる ipod？

ところで、照応は必ずしも同一文章内に先行詞を持つ必要はない。先行詞が同一文章内にもない場合もある。このような照応は外界照応 (**exophora**) と呼ばれる。

- (5) それ、ちょっと貸してくれない。

このような場合、照応詞の照応先となる先行詞は、文章中には認められず、話し手と聞き手が共に認識可能な対象は外界（言語外）にある。

照応詞とその先行詞の関係による照応現象の分類をまとめると、以下のようになる。

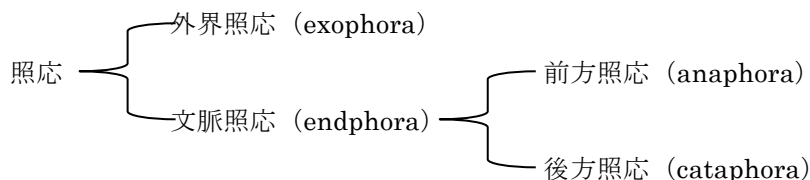


図:1 照応現象の分類（文献[1]）

さらに照応現象は先行詞が明示的かどうかによっても分類される。(6) a のように先行詞と照応詞が同義表現や上位下位関係であるなど、直接的な指示の関係にある場合を直接照応という。一方、(6) b のように、先行詞と照応詞が部分全体関係や属性関係であるなど、間接的な指示の関係にある場合を間接照応 (bridging reference) という。

(6) a 英語の単語帳を買った。この本は初級学習者にピッタリだ。

(6) b 英語の単語帳を買った。その表紙は鮮やかな緑色だ。

(6) bの「表紙」は先行詞「単語帳」の部分であることを指している。このような間接照応は、文章中で表層的には省略されているため、言語表現の読み手（受け手）は常識などの知識に基づいた推論を行って復元する必要がある。

以上のことから、単に照応といっても複雑な様相を呈していることがよくわかる。本稿ではすべての照応について詳しく見ることはせず、大きく分けて4つのタイプの照応について、コーパスごとにどのような注釈付けが行われているかを見ていく。また、照応関係の注釈付けが、文章内、文章間それぞれどのレベルにおいて行われているかもあわせて見ていくことにする。

(7) 本稿で扱う照応現象

identity-of-reference anaphora (IRA, 共参照)

identity-of-sense anaphora (ISA, sense anaphora)

exophora (外界照応)

bridging reference (間接照応)

照応・共参照の注釈付けをした代表的なコーパスには、ACE, GNOME Corpus, OntoNote がある。以下それぞれについての特徴を概観する。

## 2.1 ACE

照応・共参照関係の注釈付けは、情報抽出に関する評価型会議MUC (Message Understanding Confernce) が第6 回会議 (1995 年) および第7 回会議 (1997 年) で提供した共参照解析評価用データ[11]まで遡る。MUCの共参照タグ付きデータは、その後共参照解析手法のベンチマークデータとして長く利用されたが[32, 24, etc.], 限量子 (every, most など) を伴う名詞句や同格表現 (Julius Caesari, a well-known emperori,...) にまで共参照関係を認めるなど、仕様上の問題も指摘されている[33].

Automatic Content Extraction (以下 : ACE) [6]の特徴は、 Entity Detection and Tracking (以下 : EDT) において、それまで混同して扱われてきたコーパス中の個別の言語表現 (mention, 言及) とそれが指す現実あるいは仮想世界の対象 (entity, 実態) を陽に記述するタグ付け仕様を導入し、さらに、The Relation Detection and Characterization (RDC) では、entityのどうしの関係についてRole, Part, At, Near, Socialの5つと、サブタイプを含めた計24種類にマークアップし、共参照関係の厳密化をはかったことである。ただし、EDT では、名詞句へのマークアップの対象が意味的なもの (人名や組織名など特定の種類の固有名) に限定しており、共参照関係の認定の網羅性に問題が残る。また、RDCにおけるentityどうしの関係についても、人と組織の関係、州と国の関係など、意味的なものに限定していることから、bridging referenceなどは対象に含まれない。

照応関係の注釈付けは、文章内および文章間において行われているが、IRAのみを対象としており、ISAは対象に含まれない。これに対しACEの後継に相当するGNOMEでは、名詞句のタイプごとに、より包括的な意味の注釈付けを行うことで、また照応詞と先行詞の関係も包括的にとらえたことにより、bridging referenceもマークアップの対象にしているなど、ACE EDITに比べ共参照関係の認定の網羅性は広い。

## 2.2 GNOME

ここではGNOME corpusについて概観する。GNOME[4,5]ではNE (名詞要素) への意味属性を包括的な形でマークアップしているため、ACEに比べより多くの名詞句について言語表現 (mention, 言及) とそれが指す現実あるいは仮想世界の対象 (entity, 実態) の記述をもとに、ACEではマークアップされていないbridging reference, exophoraについてのマークアップも可能にしている。このような理由で本稿ではGNOMEについてより詳しく概観することにしたい。

GNOME の特徴は、対象とするテキストをはじめに文 (sentence) 単位で、次に独自にディスコース・ユニット (DU) と呼ぶ節に近い単位を使用してマークアップしたことである。このDU (UTYPE 属性) には、main, relative, appositive, parenthetical (主節, 関係節, 同格節, 挿入句) などがある。

(8) DU (UTYPE属性) タグ

main, relative, such-as, appositive,  
parenthetical, paren-rel, paren-app, parenmain,  
subject, complement, adjunct,  
coord-vp, preposed-pp, listitem,  
cleft, title, disc-marker.

文や UD (節) レベルのマークアップは、同一文章内または文章間といった照応・共参照のマークアップには不可欠である。

GNOME では、すべての NPs を NE として扱い、それぞれに ID (番号) を割り振る他に、CAT (定冠詞を伴う, 不定冠詞を伴う, 代名詞である, 固有名詞である) や NUM (数, 人数や数量), PER (人であるかどうか), GEN (性別を有するかどうか), GENERIC (generic かどうか), GF (subject, object, predicate などのどれに相当するか), LFTYPE (数量詞が付くか否か), DEIX (対象指示的かどうか), ONTO (具体的か抽象的か), REFERENCE (直示的か, 暗示的か), STRUCTURE (individual かどうか) など, 全部で 16 種類の属性を表すタグが用意されている。そして、それぞれのタグにはより細かい情報を付与するためのタグも用意されている。(9) は NE の属性タグ (10) は CAT タグのより細かな情報を示したものである。

(9) NE の属性タグ

ID ANI CAT COUNT  
DEIX DEN GEN GENERIC  
GF LFTYPE LOEB NUM  
ONTO PER REFERENCE STRUCTURE

(10) CTA タグの下位分類

PERS-PRO POSS-PRO REFL-PRO Q-PRO WH-PRO  
THIS-PRO THAT-PRO ONE-ANA NULL-ANA PN  
POSS-NP THE-NP THIS-NP THAT-NP A-NP  
ANOTHER-NP BARE-NP Q-NP WH-NP NUM-NP  
MEAS-NP GERUND COORD-NP

これらの NE タグは、マークアップの対象やマールアップされる情報で大きく 3 つの段階に分けることができる (表 1 参照)。これは照応・共参照関係の認定に重要な役割を果たす。

---

タグ	マークアップの対象	マークアップされる情報
----	-----------	-------------

---

CAT,GEN,GF, NUM	NE	固有名詞, 代名詞, subject, object
COUNT,GENERIC,ONTO	antecedent relation	NE どうしの関係, 類似要素, 部分全体
DEIX,REFERENCE,STRUCTRE	anaphoric information	照応の種類, coreference, bridgingreference

表1 使用されるタグとマークアップの対象、情報

### 2.2.1 照応・共参照のマークアップ

では、GNOMEはこの照応・共参照関係をどのようにマークアップしているのだろうか。ここではGNOMEにおける照応・共参照関係のマークアップについて詳しく見ていく。

まずGNOMEでは、それぞれのNEがどのような属性を持つかがマークアップされる。ACE EDT では、名詞句へのマークアップの対象が意味的なもの（人名や組織名など特定の種類の固有名）に限定しており、これが広く共参照関係を認定するうえで問題となっていたが、GNOMEでは意味的なものに加え、NUM、PERなどの統語的な属性の注釈付けを行うなど、より包括的な属性が注釈付けされている。

NEについてのマークアップが終わると、各NEの属性を基にNEどうしの関係（antecedent relation）がマークアップされる。これはNEに付与された意味レベルのタグによって行われる。例えば、先行詞と照応詞が「保険」と「特約」の関係ならば、後者は前者の一部を成す関係にあることがわかる。このような「全体」と「一部」を表す関係や、「家」と「屋根」のように「所有」を表す場合、GNOMEのanaphoric informationは、（bridging reference, 間接照応）としてマークアップされる。一方で、同義表現や上位下位関係などの直接的な指示の関係にある場合には、coreferenceとしてマークアップされる。

最後に各照応詞と先行詞の関係がマークアップされるのであるが、これは<rel=>によって示されている。文中の先行詞と照応詞が一つあるいは二つ以上の同一実体を指す共参照になる場合は、<rel=“ident”>としてマークアップされる。照応関係をマークアップする際、照応詞の照応先となる先行詞が明示的に現れている場合は、共参照<rel=“ident”>としてマークアップされるが、先行詞が明示的に認められず、表層的に省略されている（bridging reference, 間接照応）の場合には、<rel=“poss”>（所有の関係）や<rel=“element”>（全体と一部の関係）として示される。

(11)

John has bought a new car.

The indicators use the latest laser technology.

(12)

The Italian team didn't play well yesterday until the centre-forward was replaced in the 30th minute.

(11) は、照応詞“The indicators”の先行詞に相当する要素が先行文脈において明示的に出現していない。照応詞の照応先として“a new car”が認定されるのは、文献[7]によると、The indicators と a new car は一般的な知識に基づいて、両者は所有の関係（一方が他方を所有する）にある、という推論ができるためである。このような場合は、<rel= “poss” >としてマークアップされる。(12) は、先行詞と照応詞が、全体と一部の関係にあるため、これは<rel=“element” >としてマークアップされる。

さらにGNOMEでは、(exophora 外界照応)のマークアップも行っている(例13)。

(13) *pass me the salt, please.*

(13) の“the salt”の照応先である先行詞は、文章中には出現しておらず、話し手と聞き手が共に認識可能な対象として言語外に存在している。GNOMEでは、exophoraの場合、先行詞と照応詞の関係は<rel= >を使用してマークアップせずに、NEのDEIXタグ (DEIX=yes 対象指示的である) でマークアップしている。

GNOMEでは、間接照応の同定に関して、(i) 照応詞が先行詞の一部を成している、(ii) 照応詞が先行詞の性質、部分、状態を表している、(iii) 照応詞が先行詞と所有の関係にある、といった規則に基づき間接照応のマークアップをしているところ、また、exophoraについてもNEのDEIXタグによってマークアップしている点で評価できる。このようなexophoraのマークアップを可能にしているNEの包括的な機能タグもACEにはないGNOMEの特徴といえる。ただし、先行詞と照応詞が同義表現や上位下位関係であるなど、直接的な指示の関係にある場合にはISAを考慮せず、共参照関係を陽に認めてしまっている点は腑に落ちない。

### 2.3 OntoNotes

OntoNotes プロジェクト[26]は、Eduard Hovyらを中心としたグループによって行われている。コーパスは、主に新聞記事を対象にしたものであり、言語は英語だけではなく中国語によっても同様の注釈付けが行われている。OntoNoteの特徴は、同一のコーパスに様々なレベルの意味情報を重層的に付与しているところである。Treebankスタイルの統語構造や、PropBankスタイルの述語項構造および意味役割などの注釈付けの上に、これらの情報を基にして、NomBank (動詞や形容詞の名詞化) やWordSense (語義)、固有名 (Nameと呼ばれ、18種の固有名詞の実態を定めたもの)、そしてcoreference (共参照の注釈付け) などの情報を重層的に付与している。また、語義の曖昧性の解消に至っては、多くの言語研究者らによるWordSenseの注釈付けが行われ、これは各言語研究者間の注釈付けの一致率が90パーセントになるまで徹底して行われた。

照応・共参照関係の注釈付けであるが、GNOMEでは以下の例にある“one”については、John と tree TVsのどちらにも捉えられるとし、共参照関係 (identな関係) の注釈付けの対象に含めなかったが、OntoNoteでは、照応詞に対して複数の先行詞候補を共参照関係として認定するな

どし、このような問題を解決している。

(14) *(John) has (three TVs). (He) keeps (one) in the kitchen.*

また、述語項構造解析と共参照解析を並列的に行うことによって、生成文法において音形を持たない（または、目に見えない）文法要素（名詞句）である空範疇（PRO）においても共参照関係を認定することが可能になっている。

もちろん課題も多い。OntoNotes では、共参照関係認定の厳密化を図るという理由で、文章内の共参照関係のみマークアップの対象となっている。さらに、それぞれのレイヤには解決すべき課題も残っている。このことについては3章で取り上げる、PropBank, NomBank で詳しく触れるとして、上で述べた述語項構造と共参照の組み合わせにしても、述語項構造のゼロ照応は一般にISAを許すので、2章で触れたIRAとISAの問題をやはり再検討する必要がある。

## 2.4 The Copenhagen Dependency Treebanks (CDT)

CDT[41, 42]はKromannらによって、デンマーク語を中心に、英語、ドイツ語、イタリア語、そしてスペイン語を並行して扱い、依存構造の上に談話情報、照応、形態素、シンタックスなどの注釈付けを行っている。この章ではCDTの全ての注釈付けについて詳細に見ることはせず、照応・共参照の注釈付けについてのみ紹介する。

CDTの特徴は他のコーパスとは違い、すべての照応・共参照関係の注釈付けを行っていることである。それでは、以下でCDTの照応・共参照の扱いについて見てみようと思う。

### 2.4.1 CDTの照応情報への注釈付け

CDTは依存文法をもとにした統語構造をもとに、照応や共参照情報の注釈付けを行っている。さらにQualia structureと呼ばれる独自のタグセットの使用により、照応情報の注釈付けも行っている。Qualiaとは、我々が、意識にある状態で感じる様々な質感のことである。日本語では、しばしば感覚質などと訳されている。Qualia structureにおけるFOMAL qualiaとは、先行詞となる名詞に備わっている、大きさ、形や色などの特徴を指し、例えば、ASSOC-FORMALであれば、先行詞となる名詞に備わっている質感、形や色、味、香り、大きさなどの特徴が照応詞側に確認できる場合、これを照応関係にあると認定する。

例えば「日本酒を買った。香り、味ともに文句なしの旨さだ。」のような場合、クオリアをもとにしてその関係（香りや味から日本酒の質感を表現できることから、日本酒と香りや味は部分全体の関係にある）を判断する。

(15) associative-anaphor のタグセット

- |                     |                   |                  |
|---------------------|-------------------|------------------|
| 1. Qualia structure | 2. Semantic roles | 3. Other types   |
| ASSOC-FORMAL        | ASSOC-AGENT       | ASSOC-LOC(ation) |



ASSOC-CONST(titutive)	ASSOC-PATIENT	ASSOC-TIME
ASSOC-AGENTIVE	ASSOC-EXPER(iencer)	ASSOC-EVENT
ASSOC-TELIC	ASSOC-REC(ipient)	
	ASSOC-INST(rument)	

(16) The ham to be used in the dish must not be too salty. You cannot use *the thin slices*, which are packaged in the refrigerated counter. *They* are too salty and too wet and ***the flavor*** [ASSOC-FOMAL] is not good enough. (斜体が先行詞, 太字が照応詞)

先行詞 “the thin slices” の備える特徴である味覚という要素が “the flavor” という照応詞であることが理解できる。

次に, Semantic roles についてであるが, これは, 先行詞に対して照応詞が agent となるか, または patient となるか, 先行詞と照応詞の関係を記述している。

(17a) *The operation* itself requires general anesthesia ... the patient is asleep for the entire course of the operation. ***The surgeon*** [ASSOC-AGENT] opens the chest by dividing the breast bone or sternum. (斜体が先行詞, 太字が照応詞)

(17b) *The operation* itself requires general anesthesia ... ***the patient*** [ASSOC-PATIENT] is asleep for the entire course of the operation. The surgeon opens the chest by dividing the breast bone or sternum. (斜体が先行詞, 太字が照応詞)

例えば (17a) であれば, The operation に対して The surgeon は agent であることから, ASSOC-AGENT anaphor となり, (17b) なら, The operation に対して The patient は patient であるから, ASSOC-PATIENT anaphor となる。

#### 2.4.2 CDT における共参照情報のマークアップ

次に, 共参照情報の注釈付けであるが, GNOME では単に ident としてマークアップされていた (18) のような固有名詞と代名詞の関係や, 上位下位関係「例, 車→その乗り物」, 「ある黄色い車→この車」について, CDT では, 共参照関係をさらに詳しく分類し情報を付与している。また, 先行する文や節がそのまま先行詞となる共参照 (表 2 D 参照) 「例, 「大阪のゲーム販売会社の元社長らが不正に資産を水増ししたとして逮捕された事件で, 元社長が価格をつり上げるため虚偽の事業計画を公表していたことがわかりました。 *i* この事件 *i* は, …」」や, 発

話行為の共参照 (表 2 E 参照) 「例, 「今度あったらただじゃおかないからな *i*, その *i* 言葉が



intra: 文内照応, inter: 文間照応, exo: 外界照応

表3 各コーパスの照応関係認定の範囲と注釈付けの対象

IRA (共参照) の注釈付けに関しては, ACE, GNOME, OntoNotes, CDTのいずれのコーパスでも行っているが, ISA (照応) については, CDTのみが注釈付けしている。また, bridging reference (間接照応) については, GNOME, CDTが注釈付けの対象にしているが, exophora (外界照応) については, GNOMEのみであった。

照応関係の認定範囲については, ACE, GNOME, CDTは, 文章内, 文章間の照応関係の注釈付けを対象にしているが, OntoNotesのみ厳密な共参照関係の認定を行うという理由から, 文章内の共参照関係のみを対象にしている。

コーパスによって, その目的や設計は大きく異なっており, 一概にどのコーパスが優れているということとはできない。何をどの程度明らかにするか, その目的に応じたコーパスの選択が重要である。

### 3 意味情報の注釈付け

コーパスへのもっとも基本的な注釈付けは, 形態素, 統語情報である。形態素, 統語情報が注釈付けされた代表的なコーパスに, PennTreeBank[20]がある。PennTree Bankでは, 文中の各単語に品詞情報が付与されている。NNP, JJ, VBPなどの記号はそれぞれ普通名詞複数形, 形容詞, 動詞過去のような品詞情報を表している。また, 各文に対してNP, VPなどの句構造情報が付与されている。

文の統語的な構造を階層的にとらえることができるPennTreeBankは, 統語解析研究の標準データとして長年利用されてきた。しかし, 階層性だけではとらえられない情報も文の中には存在している。たとえば, 動作主や受益者などの, それぞれの動詞の項がとる意味的な性質の違いは, このような構造を見るだけではわからない。そこで, こうした意味的な性質の違いを明らかにするために, 動詞の項がとる意味役割情報を付与する必要性がでてくる。

#### 3.1 FrameNet

言語学の一分野に, 語の意味を記述する方法を論じる研究分野がある。その一つが, Fillmoreが提唱したフレーム意味論[43]である。フレーム意味論では, 言語の理解や産出には言語知識のみならず百科事典的知識が必要であると考え, それらの知識を「フレーム」と呼ばれる知識構造の体系として記述した。FrameNet[3]は, Fillmore率いる研究チームが, このフレーム意味論をもとに開発を続けているコーパスである。

FrameNetの知識記述を構成する単位はフレームである。個々のフレームは, 動詞格フレームのある種の抽象化になっており, 意味役割に相当する「フレーム要素 (frame element)」の集合とそのフレームを喚起する語の集合によって規定される。例えば, “購買”を表す

Commerce\_buyというフレームは次のように記述される。

### Commerce buy フレーム

フレーム要素（必須）：

- ・ Buyer: The Buyer wants the Goods and offers Money to a Seller in exchange for them.
- ・ Goods: The Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.

フレーム要素（選択的）：

- ・ Seller: The Seller has possession of the Goods and exchanges them for Money from a Buyer.
- ・ Recipient: The individual intended by the Buyer to receive the Goods.
- ・ Money: Money is the thing given in exchange for Goods in a transaction.

フレーム喚起語: buy.v, purchase.v, purchase ((act)).n

関連フレーム：

- ・ Inherits From: Getting
- ・ Is Inherited By: Renting
- ・ Is Used By: Importing, Shopping
- ・ Perspective on: Commerce goods-transfer

図2 Commerce\_buyフレーム

フレームとフレームの間には上位下位関係（Inferit）や部分全体関係（Subframe），手段目的関係（Use），因果関係，時間的順序関係などが規定され，フレームの集合はFrameNetと呼ばれるネットワークを構成している．例えば，Commerce buyフレームでは，いわゆる対象に相当するフレーム要素はGoods と呼ばれるが，その上位フレームであるGettingフレームではTheme と呼ばれる．このような上位下位関係や部分全体関係の関連するフレーム間には，どのフレーム要素がどのフレーム要素に対応するかも記述されており，それによってフレーム間の関係を利用した多様な推論が可能になっている．

### 3.2 述語項構造（PropBank）

PropBank[25]は，PennTreeBankの統語情報上に述語（主に動詞）とその項の意味役割を付与したものである．PropBankもFrameNetもFillmoreのフレーム意味論をもとにしているが，両者の意味役割の記述方法は異なっている．

PropBankでは，agent やtheme などの意味役割に相当するARG0, ARG1, . . . , ARG5, AA, AM-ADV などの35 種類のラベルを用いて文内の述語と項の関係をマークアップする．例えば次の例(19) では，動詞“earned” に対し，“the refiner” をagent 相当のARG0，“\$66 million, or \$1.19 a share” をtheme 相当のARG1 としてマークアップする．

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

Propbankのフレームは、動詞ごとに個別のフレームが用意され、その動詞の項構造の数に応じて一動詞あたり平均1.4個のフレームが作られている。たとえば、leaveについては以下のようなになる（20参照）。

(20) leave

Frameset leave.01 "move away from":

Arg0: entity leaving

Arg1: place left

Frameset leave.02 "give":

Arg0: giver

Arg1: thing given

Arg2: beneficiary

PropBankでは、一つの動詞が取る意味ごとにフレームを用意しており、FrameNetに見られるようなフレーム間の関係は定義されておらず、各フレームは定義上独立な関係にある（表4）。そのため、FrameNetに比べ、多様な意味を持つ動詞の意味役割を詳細に記述できる点で優れている。ただし、動詞の全ての意味について、このフレームが用意されているわけではないため、未知のものについては扱えないという問題もある。その点、FrameNetでは、未知の意味役割についても、各フレームの上位下位関係や部分全体関係から推論によって導き出すことができるため、動詞の意味役割に対する網羅性は高い。

FrameNet, PropBankはともにFillmoreのフレーム意味論をもとにしているが、フレームの記述の方法は異なっている。このような差異が、述語（動詞）の意味記述の網羅性や精度の違いとして表れている。

PropBank		FrameNet	
<i>buy .01</i>	<i>sell .01</i>	<i>COMMERCE_buy</i>	<i>COMMERCE_sell</i>
Arg0: buyer	Arg0: seller	Buyer	Seller
Arg1: thing bought	Arg1: thing sold	Goods	Goods
Arg2: seller	Arg2: buyer	Seller	Buyer
Arg3: price paid	Arg3: price paid	Money	Money
Arg4: benefactive	Arg4: benefactive	Recipient	Recipient

表4 PropBank と FrameNet における動詞 “buy” と “sell” の定義

また, PropBank やFrameNet コーパスは文内の述語と項の関係だけを対象としているが, 日本語のように必須格も含めて頻繁に省略が起こる言語では, 述語の項が文境界を越えて出現する場合(すなわちゼロ照応)や, 一人称や二人称など同一文章内に明示的に出現しない場合(外界照応)への対応も必要であろう.

### 3.3 NomBank (事象性名詞の項構造)

Meyers らが作成したNomBank[21]では, Penn Treebank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, PropBank の仕様に従って項構造を付与している. 例えば, 句(21)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 dividends” 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.

(21) 12% growth in dividends next year.

[REL=growth, ARG1=in dividends, ARGM-TMP=next year]

また, 名詞化した動詞や形容詞が支援動詞を伴う場合についても, PropBankの意味役割相当のラベルが付与されるわけであるが, 支援動詞は “SUPPORT” と記述される(22参照).

(22) The judge made demands on his staff.

[REL=demands, SUPPORT=made, ARG0=The judge, ARG2=on his staff]

PropBank を用いた動詞に対する意味役割付与は, 述語項構造解析において, その有効性が確認されている. NomBankは, PropBankの仕様を基に作成され, 名詞の語幹やクラスといった事態性名詞についての意味素性や, 支援動詞構文を認識するための述語との位置関係といった統語素性, そして項構造を正しく認識することを可能にした. しかしながら, NomBankは, 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.

### 3.4 PropBankにおける照応・共参照の注釈付け

本稿の目的は, 述語項構造, 照応・共参照, 談話関係についての先行研究を概観し, 各情報の注釈付けで検討すべき課題をまとめるとともに, 各レイヤ間の関係についてもまとめることにある. 近年, 複数のコーパスの注釈情報を統合する動きも活発になっており, 同一の文書集合に対して, 統語情報の上に共参照, 述語項構造, 談話関係など様々なレイヤの意味情報を重層的に付

与する試みがすでいくつか報告されていることは冒頭で触れた通りである。そこで、ここでは PropBankにおいて照応・共参照がどのようにマークアップされているのかを紹介する。その前に、照応・共参照の定義を軽くおさらいしておく。

照応と共参照の定義は、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。一方、二つ（もしくはそれ以上）の表現が現実世界あるいは仮想世界において同一の実体を指す場合、それらの表現は共参照（あるいは同一指示）の関係にある、というものであった。

PropBankでは、（照応 以下: ISA）,（共参照 以下: IRA）について、先行詞と照応詞が厳密に同一実体を示す場合に限り、IRAとしてマークアップしている（23参照）。

(23) John Smith of Company X arrived yesterday. Mr.Smith said that..."

(23) の“John Smith”と“Mr.Smith”は共に固有名詞であり、先行詞と照応詞が同一実体を指すため、IRAとしてマークアップされる。一方、(24)のように、照応詞が代名詞である場合には、IRAではなくISAとしてマークアップされる。このように、PropBankでは先行詞と照応詞が厳密に同一実体を指す場合以外はISAとしてマークアップされる。

(24) John Smith of Company X arrived yesterday. He said that..."

照応詞がpossessive,あるいは、adjectiveの関係にある場合もISAとしてマークアップされる（25参照）。

(25) Many companies raised their payouts by more than 10%.

(25) は、先行詞“Many companies”の照応詞である“their”は、adjectiveの関係にある。このような場合もISAとしてマークアップされる。

(26) I want to buy a car. I need it to go to work.

(26) の“a car”は話し手が実際に所有していない、いわば仮想世界における車である。このような場合、共参照の定義では、先行詞と照応詞が同一の実体を指すのであればIRAを認めるが、PropBankでは先行詞“car”に対する照応詞“it”が先行詞と同一の実体以外を指す可能性も示唆できるとし、ISAとしてマークアップしている。

(27) Larry is a university lecturer.

(28) Larry, the chair of his department, became president.

ただし、先行詞が句 (NP) ではなく節 (CLAUSE) になっているものや、predicate nominal (27) や appositive (28) などは、IRA, ISAというよりは先行詞がどのようなものなのか、といった性質を記述しているにすぎないという理由で、マークアップの対象から外れている。

### 3.4.1 空範疇と共参照

空範疇とは、生成文法において音形を持たない（または、目に見えない）文法要素（名詞句）のことを指す。空範疇には、その先行詞との間に移動関係が見られないPRO (29) や、ある要素が移動という操作を受けた結果として、もとの位置に残される痕跡などがある。後者の例としては、NP痕跡 (30) , wh痕跡などがあげられる。空範疇および統語構造の注釈を付与した代表的なものに、TreeBankコーパスがある。

(29) I made a decision [\*] to leave.

(30) Active: Mary hit John

Passive: John was hit [\*T\*] by Mary.

PropBankにおいても空範疇 (empty category) はマークアップされており、すべての空範疇は同一文内において存在する特定のNPと共参照関係を持っている。例えば (31) は、空範疇[\*] が “I” と同一の実体を指すことを意味し、IRAとしてマークアップされる。

(31) I made a decision [\*] to leave.

Rel: leave

Arg0: [\*] -> I

### 3.4.2 各コーパスにおける照応・共参照情報の注釈付け

PropBankの特徴は、GNOMEやOntoNotesはIRAのみをマークアップの対象にしていたが、ISAや空範疇にもマークアップをしている点であろう。また、IRAやISAの関係は、文章内、文章間両方においてマークアップされており、GNOMEやOntoNotesに比べ、IRA, ISAの網羅性は広いといえる。ただし、PropBankのIRAとISAの基準は、GNOMEやOntoNotesに比べ曖昧なため、先行詞と照応詞が厳密に同一実体を指している場合のみIRAを認め、そうでないものは全てISAとしている。そのため実際にはIRAであっても、ISAとしてマークアップされるなどの問題もある。

GNOME, OntoNotes, PropBankそれぞれのマークアップの範囲と対象にしている照応関係をまとめると、以下のようになる。

---

コーパス	照応関係の認定範囲	マークアップの対象
------	-----------	-----------

---



ACE	intra inter	IRA
GNOME	intra inter exo	IRA bridging reference exophora
OntoNotes	intra	IRA
PropBank	intra inter	IRA ISA PRO

intra: 文内照応, inter: 文間照応, exo: 外界照応

表5 各コーパスの照応関係認定の範囲とマークアップの対象

## 4 談話関係

### 4.1 RST corpus

文章中の事態の情報を言語処理の応用処理に利用するためには、談話単位内に出現している述語項構造や照応関係の研究だけではなく、談話単位間の意味的关系を把握する必要が出てくる。談話単位間の関係については“背景”や“原因”といった事態間の関係ラベルをどのように定義するか、また文章の構造を木構造とするか、より一般的なグラフ構造にするかでいくつかの異なる論理的な枠組みが提案され、それぞれの枠組みに基づいたタグ付きコーパスが整備されている[53. 3]。例えば、RSTでは関係の種類として“詳細化”，“原因”，“結果”，“対比”など約23種類が提案されている。RSTでは、次のように、個々の関係について2つの談話単位それぞれとの組み合わせにおいて、どのような条件が成り立つかを定義している。

また、日本語でも横山ら[38]，新森ら[39]，杉浦[40]などが談話単位間の関係ラベルを定義してそれぞれ独自にタグ付きコーパスを構築している。

#### (32) RSTにおける根拠関係の定義

- nuclear (帰結) 側の制約: 書き手が満足できるほど読み手が帰結について信じていないかもしれない。
- satellite (根拠) 側の制約: 読み手は根拠を信じられる、もしくは根拠に信憑性がある。
- 帰結と根拠の組み合わせについての制約: 読み手の根拠への理解が帰結についての信念を増す。
- 書き手の意図: 読み手の帰結への信念を増す。

しかし、こうした定義は曖昧な記述も多く、人手による関係付与も揺れる傾向が強いことが報告されている[36]。

### 4.2 Penn Discourse Treebank (PTSD)

Penn Discourse Tree Bank (PDTB) [22]では、主“because”や“but”など明示的に文章中に出現している接続表現を手掛かりとし、RSTにおける木構造といったような、談話構造の形を仮定せず、談話片間の2項関係で表現している。図3はPDTBの仮定している関係セットであり、特徴は3段階の階層的な関係分類となっている点である。これにより、コーパス利用者の求める粒度に柔軟に対応できる。

<b>Temporal</b>	<b>Contingency</b>	<b>Expansion</b>
Asynchronous	Cause	Instantiation
precedence	reason	Restatement
succession	result	specification
Synchronous	Pragmatic Cause	equivalence
<b>Comparison</b>	justification	generalization
Contrast	Condition	Alternative
juxtaposition	hypothetical	conjunctive
opposition	general	disjunctive
Pragmatic Contrast	unreal present	chosen alternative
Concession	unreal past	Exception
expectation	factual present	Conjunction
contra-expectation	factual past	List
Pragmatic Concession	Pragmatic Condition	
	relevance	
	implicit assertion	

図3 PDTBの関係セット

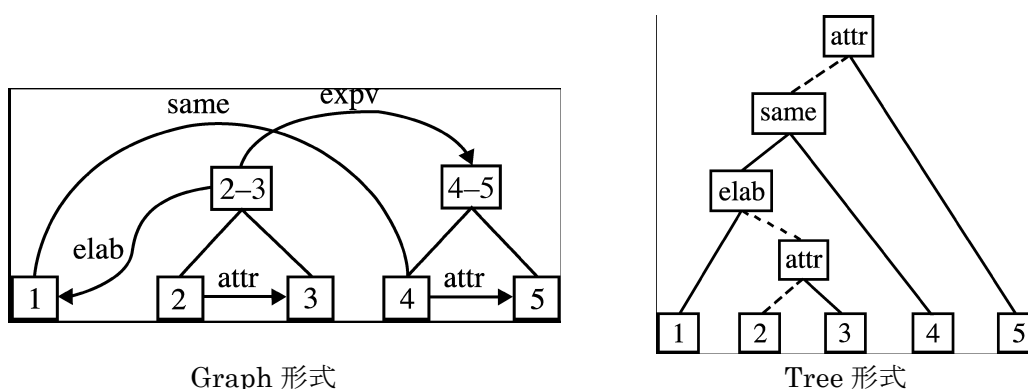
PDTB では語彙的な手がかりを持つ談話関係及びローカルな談話関係に注目して談話関係を付与するというアプローチの方法をとっている。ある関係をもった談話片間に、その関係を示す接続表現(connective) が明示的に存在するときの関係を**explicit** な関係、接続表現がない時の関係を**implicit**な関係と呼ぶが、関係付与者は、まず明示的な手がかりである接続表現 (e.g., *because*, *for example*など) をもとに**explicit** な関係を付与する。加えて同一段落内の隣り合った文章間といったローカルな談話片間に対して**implicit** な関係を付与する。このように、**explicit** である場合と**implicit** である場合とで関係付与の方法を分けている点はPDTB の大きな特徴である。

### 4.3 Discourse Graphbank

RST では文章の構造を木構造としたが、より一般的なグラフ構造を導入したのが Wolf & Gibson による Discourse Graphbank[35]である。

- (34) 1. 1a[ Mr. Baker's assistant for inter-American affairs, ] 1b[ Bernard Aronson, ]  
2. while maintaining  
3. that the Sandinistas had also broken the cease-fire,  
4. acknowledged:

## 5. “It’s never very clear who starts what.”



expv = violated expectation; elab = elaboration; attr = attribution

図4 Graph形式とTree形式での談話構造

図4は、上段の文章に対する談話構造をグラフ形式、Tree形式でそれぞれ表したものである。Tree形式と比べるとGraph形式のほうがsameの関係やexpvの関係をより詳細に付与することができるのがわかる。

## 5 マルチレイヤ

同一のコーパスに様々なレベルの意味情報を重層的に注釈付けする試みもすでいくつか報告されている。その先駆けの一つに、Prague Dependency Treebank [9, 8] が挙げられる。Prague Dependency Treebank (PDT) は、Penn Tree Bankの有用性に感銘を受けたチェコのJan Hajicらによって提唱された。PDTの特徴は、PennTreeBankの句構造文法とは異なる、基本的な記述の単位に句を用いない依存文法を基にした記述を行っていることである。句構造に依存しない依存文法は、特定の語順による定義がされないため、チェコ語のような比較的語順の自由な言語の記述に適している。では、PDTはどのようなレイヤを使用しているのだろうか。

PDTのレイヤであるが、形態素と依存構造の注釈付けに加え、Tectogramaticsと呼ばれるレイヤが用意され、深層の依存構造(PropBankスタイルの意味役割付与に概ね相当)から省略・共参照、新旧情報など、幅広い意味談話情報が付与されている。

一方、異なる研究グループによって個別に開発された異なるレイヤの注釈付けを垂直に統合する試みも報告されている。代表的な例は、PropBank, NomBank, TimeBank, Penn Discourse Treebank, FactBank等の注釈情報を統合するPustejovskyらの試みであろう[29]。Pustejovskyらはこうした統合によって異なるレイヤ間で注釈情報を付き合わせ、調整することが可能になり、コーパス全体の仕様の整合性、注釈情報の品質の改善が期待できると論じている[28]。Pustejovskyらはその後、XBankブラウザ5と呼ばれるツールを開発し、PropBank,

NomBank, TimeBank, Penn Discourse Treebank, MPQA の注釈情報をレイヤ横断的に調べることができる環境を開発するに至っている。これら異なるレイヤの言語情報を統一的に記述するための Unified Linguistic Annotation と呼ばれる枠組みも開発され、小規模ながらこの枠組みで注釈付けされたテキストデータも LDCLDC (Linguistic Data Consortium) から配布されている。

この他にも複数の研究サイトが協調的に意味的注釈付けの垂直統合をはかる試みがいくつかある。例えば, Eduard Hovy らを中心として行われた OntoNotes プロジェクト[26] では, Penn Treebank スタイルの統語構造, PropBank/NomBank スタイルの意味役割の他, 語義, 固有名, 共参照を重層的に付与する。また, 多様なジャンルのテキストへの注釈付けを目的として, American National Corpus (ANC) の一部に重層的に注釈付けするプロジェクト[11] も進行中であり, その一部は Open ANC としてすでに公開されている。

## 6 まとめ

本稿では, 自然言語処理研究に欠かせない注釈付きコーパスの現状を概観し, それぞれのコーパスの特徴や, そしてそれらの抱える課題について紹介した。

何を固有表現と認めるのか, 述語の項を何種類に分類するか, 照応関係の認定条件をどうするかなど, 問題の具体的な取り決めは主として注釈付きコーパスの設計者が行ってきた。統計的自然言語処理と呼ばれる経験主義的アプローチが拡大したこの20年は, ベンチマークの開発による問題設計の歴史としてもたどることができる。これらの問題設計の多くが具体的な応用の課題分析から始まっている点も重要である。たとえば, 固有表現の重要性が広く認識されるようになったのは, 米国の評価型研究プログラムMUC の情報抽出タスクがきっかけだった。情報抽出の研究プログラムACE からは, entity と mention を区別した共参照解析の課題仕様が提案され, こうして問題は発見され, 創造され, 修正されてきた。具体的な応用から出発して, 実際の言語データと格闘しながら問題を練っていく。問題の切り方が良ければ, やがては応用横断的な基本問題に一般化され, 技術のモジュール性, そして統合のしやすさにも繋がる。問題を設計することの貢献は大きい。とは言え, 本稿で見たように, 注釈付きコーパスの設計にはまだ多くの課題が残っている。とくに, 意味レベルの注釈はまだ仕様設計の模索段階にあると言っても過言でない。そこには, 言語処理研究者だけでなく, 言語学研究者にとっても極めて興味深い研究課題がいくつも埋まっているはずである。言語処理と言語学の垣根を越えた連携による研究のより深化が望まれる。

### 参考文献

- [1] 山梨正明. 推論と照応. くろしお出版, 1992.
- [2] Charles J. Fillmore. Frame semantics and the nature of language. In *In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language*

- and Speech*, pp. 20–32, 1976.
- [3] C.F. Baker, C.J. Fillmore, and J.B. Lowe. The berkeley framenet project. In *Proceedings of the International Conference on Computational Linguistics* (COLING/ACL-98), pp.86–90, 1998.
- [4] M.Poesio. *The GNOME Annotation Scheme Manual*. University of Edinburgh, HCRC and Infomatics, Scotland, fourth version edition, July. Available from <http://www.hcrc.ed.ac.uk/~gnome>. 2000.
- [5] M.Poesio. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston, May, 2004.
- [6] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. Automatic content extraction (ace) program - task definitions and performance measures. In *Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 837–840, 2004.
- [7] H. H. Clark. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press, 1977.
- [8] J. Hajič, E. Hajičcová, J. Hlaváčová, V. Klimeš, J. Mírovský, P. Pajas, J. Štěpánek, B. Vidová, V. Hladká, and Z. Zabokrtský. PDT 2.0 — Guide, 2006. <http://ufal.ms.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/pdf/pdtguide.pdf>.
- [9] E. Hajičová. Prague Dependency Treebank: From analytic to tectogrammatical annotation. *Proceedings the First International Conference on Text, Speech, Dialogue*, pp. 45–50, 1998.
- [10] L. Hirschman. *MUC-7 coreference task definition*. version 3.0, 1997.
- [11] N. Ide, C. Baker, C. Fellbaum, C. Fillmore, and R. Passonneau. MASC: The manually annotated sub-corpus of american english. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pp. 2455–2460, 2008.
- [18] Z. P. Jiang and H. T. Ng. Semantic role labeling of nombank: A maximum entropy approach. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp. 138–145, 2006.
- [19] C. Liu and H. T. Ng. Learning predictive structures for semantic role labeling of nombank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 208–215, 2007.
- [20] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. In *Computational Linguistics*, pp. 313–330, 1993.
- [21] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The nombank project: An interimreport. In *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, 2004.

- [22] E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. The penn discourse treebank. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 2237–2240, 2004.
- [23] R. Mitkov, editor. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education, 2002.
- [24] V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th ACL*, pp. 104–111, 2002a.
- [25] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106, 2005.
- [26] S. Pradhan, E. Hovy, MS Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: A unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing*, pp. 517–526, 2007.
- [27] R. Prasad, N. Dinesh, A. Lee, A. Joshi, and B. Webber. Annotating attribution in the Penn Discourse TreeBank. In *Proceedings of the COLING/ACL-2006 Workshop on Sentiment and Subjectivity in Text*, pp. 31–38, 2006.
- [28] J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 5–12, 2005.
- [29] J. Pustejovsky, Martha P., and A. Meyers. Introduction to Frontiers in Corpus Annotation II Pie in the Sky. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 1–4, 2005.
- [30] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. *TimeML Annotation Guidelines Version 1.2.1*. [http://www.timeml.org/site/publications/timeMLdocs/annguide 1.2.1.pdf](http://www.timeml.org/site/publications/timeMLdocs/annguide%201.2.1.pdf), 2006.
- [31] R. Saurí and J. Pustejovsky. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 2009.
- [32] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, Vol. 27, No. 4, pp. 521–544, 2001.
- [33] K. van Deemter and R. Kibble. What is coreference, and what should coreference annotation be? In *Proceedings of the ACL '99 Workshop on Coreference and its applications*, pp. 90–96, 1999.
- [34] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, Vol. 31, No. 2, pp. 249–287, 2005.
- [35] F. Wolf, E. Gibson, A. Fisher, and M. Knight. The discourse graphbank: A database of texts annotated with coherence relations, 2005.
- [36] 竹内和広. テキスト修辞構造タグ付けの自動化に関する研究. PhD thesis, 奈良先端科学

- 技術大学院大学, 1999.
- [37] L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pp. 1–10, 2001.
- [38] 横山憲司, 難波英嗣, 奥村学. Support vector machine を用いた談話構造解析. 情報処理学会研究報告書, 2003-NL-153, pp. 193–200, 2003.
- [39] 新森昭宏, 奥村学, 丸山雄三, 岩山真. 手がかり句を用いた特許請求項の構造解析. 情報処理学会論文誌, Vol. 45, No. 3, pp. 891–905, 2004.
- [40] 杉浦 純. 日本語談話関係コーパスの構築に向けた関連研究の調査と試験的アノテーション. 卒業論文. 東北大学, 2010.
- [41] Iørn Korzen and Matthias Buch-Kromann. *Anaphoric relations in the Copenhagen Dependency Treebanks*. To be presented at Beyond Semantics 2011, DGfS Workshop, Göttingen, February 23-25, 2011.
- [42] Matthias Buch-Kromann and Iørn Korzen. *The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks*. In Proc. of Linguistic Annotation Workshop, ACL 2010,
- [43] Fillmore, C. J. “Frame semantics and the nature of language.” *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280, pp. 20–32. 1976.