

日本語書き言葉コーパスへの重層的意味情報付与 ～ 照応・共参照, 述語項構造, モダリティ, 談話関係 ～

乾健太郎 (ツール班分担者: 奈良先端科学技術大学院大学情報科学研究科)

飯田龍 (ツール班分担者: 東京工業大学大学院情報理工学研究科)

Multi-Layered Semantic Annotation to Japanese Text Corpora: Anaphora/Coreference, Predicate-Argument Structure, Modality, and Discourse Relation

Kentaro Inui (Nara Institute of Science and Technology)

Ryu Iida (Tokyo Institute of Technology)

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, 統語解析技術だけでは不十分で, 意味解析, 談話解析と呼ばれるような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意味情報を付与するには, まず付与する意味情報の仕様を設計する必要がある. 意味情報の仕様とは, 例えば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であり, 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 [2, 13, 39, 40, 45, 44].

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 [39, 13, 37]. 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 [39].

一方, 日本語コーパスへの重層的意味情報付与についても, Global Document Annotation (GDA) [10] タグ付与コーパス (以下, GDA コーパス) や京都テキストコーパス第 4.0 版 [19] (京都コーパス 4.0) など, 先駆的な取り組みがあり, また我々のグループでも NAIST テキストコーパス [14] (NAIST コーパス) の開発を通して仕様の洗練をはかってきた. しかしながら, 英語を取り巻く最近の動きに比べると, リソース間の互換性のや理論的基盤の整備などの面で立ち後れつつあることは否めない. こうした背景を鑑みると, 本特定研究の成果として代表性のある日本語書き言葉コーパスが広く研究者間で共有できるようになった今, これを, 言語学の専門家との連携のもとに意味処理課題の重層的な設計を進めるための恰好の研究材料とすべきであろう.

そこで, 本稿では, 意味情報のレイヤとして我々のグループがこれまで携わってきた照応・共参照, 述語項構造, モダリティ, 談話関係の 4 つを取り上げ, それぞれについて注釈付けの動向を概観するとともに, 仕様設計上の主な課題を整理し, 最後に重層的意味情報付与の動向と今後の展望を論じる. 紙面の制約で詳細には立ち入れないが, 重要な関連文献はなるべく引くようにしたので, 関心のある読者をご参照いただきたい. 意味情報のレイヤとしては, この他にも語義 [35] や固有表現 [9] などがあり, それぞれ本特定研究の研究グループが取り組んでおられる. それらとの連携も今後の重要な課題であると考えている.

2 意味情報の注釈付け

2.1 照応・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。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。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。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。

文献 [31] 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。

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 Conference) 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]¹ 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 [47, 32, etc.], 限量子 (every, most など) を伴う名詞句や同格表現 (Julius Caesar_i, a well-known emperor_i, ...) 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 [50]。これに対し、MUC の共参照解析タスクの後継に相当する Automatic Content Extraction (ACE) [5] の Entity Detection and Tracking (EDT) では、コーパス中の個別の言語表現 (mention, 言及) とそれが指す現実/仮想世界中の対象 (entity, 実体) を陽に記述するタグ付け仕様を導入し、共参照関係の厳密化をはかった²。ただし、EDT では、マークアップの対象を人名や組織名など特定の種類の固有名に限定するなど、共参照関係の認定の網羅性に問題が残る。

日本語については、京都コーパス 4.0, GDA コーパス, NAIST テキストコーパスのいずれも照応あるいは共参照関係の注釈付けを行っている。京都コーパス 4.0 では、京都コーパス 3.0 の一部 (新聞 555 記事, 5,127 文) に共参照関係を注釈付けしている。ただし、このコーパスでは ACE で導入されている実体-実体間の共参照関係に加え、次の (3) の実体 (“村山富市”) と属性 (“首相”) のような関係や ISA の関係も区別せずに注釈付けしている。

(3) 村山富市_i 首相_i の年頭記者会見の要旨は次の通り。

これについては GDA コーパスも同様で、(4) の 2 つの “フロン” に対して共参照タグが付与されるなど、IRA と ISA の区別は明確でない。

(4) フロン_i 対策急げ... フロン_i による環境破壊対策は...

しかしながら、ISA を認めると、例えば次の例 (5) の “食べ物” (兵庫県内で不足している食べ物) と “食料” (被災地と離れた場所にある食料) を同じ概念と解釈すべきか否かといった困難な判断に迫られる場合が多数出てくる。

(5) 兵庫県内の暗やみの中で、人々が水と食べ物の不足に苦しんでいる同じ夜、隣接した大阪の繁華街ではネオンが光り、飲食店はにぎわっている。水も食料も、被災地を離れるとふんだんにある。

このため NAIST コーパスでは、厳密に同一実体を参照している場合に限定して共参照関係を認め、京都コーパス 3.0 の全記事約 (2,929 記事, 38,384 文) に対して注釈付けを行っている [14]³。しかし、厳密性を重視するあまり、固有表現間の共参照性にタグが集中し、代名詞や連体指示詞による照応を取りこぼす問題も見られた。さらにデータに基づく検討が必要である。

なお、NAIST コーパスでは、いわゆる bridging reference [4] や間接照応 [54] と呼ばれる照応関係についても注釈付けを試みているが [16]、これについてもさらに仕様の洗練をはかる必要がある。

2.2 述語項構造

述語とその項 (本稿では complement と adjunct の両方を指す用語として「項 (argument)」を用いる) の注釈付けに関しては、表層格レベルから深層格レベルまで様々な提案がある。代表的なコーパスは英語を対象とした PropBank [36] である。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 (6) で

¹http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

²<http://projects ldc.upenn.edu/ace/annotation/>

³<http://cl.naist.jp/nldata/corpus/>

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.

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

こうした意味役割の注釈付けは, FrameNet プロジェクト [2, 34] でも精力的に行われており, 述語項構造解析の重要な研究資源になっている.

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る. こうした背景から, 京都コーパス 4.0 では文間のゼロ照応や外界照応となる項に対しても注釈付けを行っており, 現在のところ, 共参照タグ付与対象と同じ 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述語-項関係が付与されている.

(7) 体の状態_i について健康と 答えたニツイテ_i 人は 87.8% で, 体力に自信を持っているとの回答は 81.1% だった.

一方, GDA コーパスでは agent や theme などの意味役割のラベルを用意してマークアップする点が異なるが, ゼロ照応を含めた項を対象とする点では京都コーパス 4.0 と同様である. ただし, 我々が確認した限り, 現在のコーパスではゼロ照応の先行詞が同一文内にある場合は注釈付けの対象とはなっておらず, 述語項構造解析の訓練事例として利用するには網羅性の点で問題がある.

これらに対し, NAIST コーパスでは, 京都コーパスと同様に文内・文間のゼロ照応, 外界照応を含めて網羅的に述語項関係を同定する仕様になっており, 京都コーパスの数倍の規模の新聞記事コーパスへの注釈付けがすでに完了している [14]. 関係のラベルは, (i) 研究者間で広く合意できる意味役割のセットを定義するのが容易でないこと, (ii) 日本語では表層格が概ね意味役割と対応する程度に細分化されていることから, 京都コーパスと同様, ガ/ヲ/ニ等の表層格で表現することにした. ただし, 京都コーパスとは異なり, 述語の原形, すなわち受動態や使役態, 可能態を無標の形式に戻した状態の述語に対する表層格を同定するという仕様を採用している. 例えば, 京都コーパス 4.0 では文 (8) の述語 “食べさせる” に対して, (8a) のように “私_i”, “彼_j”, “リンゴ_k” をそのガ格, ヲ格, ニ格と認める. 一方 NAIST コーパスでは, (8b) のように述語の原形 “食べる” に対し, “彼_j” をガ格, “リンゴ_k” をヲ格と同定する. また, 使役態の super causer のように述語の原形に対して格要素が増える場合には, (8b) の例のように “追加ガ格” のような注釈付けを行う. これによって, 作業者にとって容易な表層格レベルでの注釈付けを通して意味役割に近い情報を付与することが可能になる.

(8) 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}

ただし, 今後, 語彙概念構造 [17] のような述語の意味分析・記述が進めば, いずれそうした意味的な語彙資源と表層格レベルのマークアップの連携を検討すべきであろう.

2.3 事象性名詞の項構造

動詞や形容詞などの述語への項構造の付与に加え, 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 [29, 10, 19], 実際にそれらの問題への取り組みも報告されている [18, 22, 23]. 例えば, Meyers らが作成した NomBank [29] では, Penn Treebank [25] を対象に名詞とその項構造のタグ付与を行っている. 例 (9) に示すように, 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 [36] で用いられている意味役割相当の項のラベルを句の中に項が出現している場合に限ってマークアップしている.

(9) There have been [ARGM-NEG *no*] [ARG0 *customer*] [rel *complaints*] [ARG1 *about that issue*].

京都コーパス 4.0 や NAIST コーパスでは, 述語の場合と同様, 事象性名詞とその項に対して表層格レベルの注釈付けを行っている. 例えば, 文 (10) の場合, 事象性名詞 “影響” に対し, “影響する” という事象を想定し, “離党_i” をガ格の要素と解釈する.

(10) 村山富市首相は年頭にあたり首相官邸で内閣記者会と二十八日会見し, 社会党の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. 離党者がいても, その範囲にとどまると思う」と述べ, 大量離党には至らないとの見通しを示した.

名詞の項構造については、名詞と項の関係が「候補_i擁立_{ラ:i}」や「兵士_jの脱走_{ガ:j}」のように、複合名詞句の中や“A / B”などに縮退される場合もあり、こうした場合にどこまでを注釈付けの対象とするかなど、検討の余地がある。

2.4 モダリティ

テキストの中で述べられている個々の事象あるいは命題が述語項構造と照応・共参照で概ね捉えられるとすれば、その次に必要となるのは、個々の事象あるいは命題に対する書き手あるいは他者の態度、すなわち広義のモダリティ情報を認識することである。テキスト中の事象はそれぞれ、実際に成立した事実として語られているかもしれないし、それとも成立しなかったこととして語られているかもしれない。あるいはその成立を書き手が推測したり、希望している場合もある。

(11) a. この夏、ぜひとも京都に旅行に行きたい。

(「この夏、私が京都に旅行に行く」という事象が将来において実現することを 望んでいる)

b. もう遅いから、きっと彼は先に帰ったんだろう。

(「彼が先に帰る」という事象がすでに成立しているであろうと 推量している)

c. 廊下を走らないでください。

(「あなたが廊下を走る」という事象を 否定的に評価し、受け手にそれを実行しないように 働きかけている)

こうしたモダリティ情報を自動的に識別することは、情報抽出や質問応答、文書要約等の広い範囲の言語情報編集に必要な処理であり、その課題設計やコーパス作成にもにわかに関心が高まっている [41, 38, 44, 45, 20, 33]。また、生物・医療情報の分野でもいくつかの動きが見られる [27, 48]。ただし、モダリティと一口に言っても、どのような種類の情報をカバーし、どのような粒度で識別するかは研究グループによって大きく異なり、課題設定間相互の関係も不明確で、やや混沌としているのが現状である。研究者間で一定の合意が得られるまでには、さらに理論的な分析とデータの蓄積を重ねる必要がある。

広義のモダリティ情報には、**伝達の態度**（叙述、意志、働きかけ、問いかけなど）、当該事象の**真偽**やその**確信度**、**態度表明者**（真偽判断の保持者）、態度や真偽のスコープなどが含まれるが、現在これらを最も包括的にカバーしているコーパスは FactBank [43, 45] であろう。FactBank では、テキスト中の各事象に対し、それが事実（と判断されている）かどうかの情報を極性（polarity）、認識的モダリティ（epistemic modality）、態度表明者（source）の組み合わせで注釈付けする。これによって、次のように真偽判断が入れ子になる場合も表現できる。

(12) *Mary regrets that John does not know he is sick.*

書き手は *John is sick* が真であると思っている

John は *John is aware of it* が偽であると思っている

Mary は *Mary knows John is sick* が真であると思っている

ただし、FactBank の仕様だけで言語情報編集に必要なモダリティ情報を十分に記述できるかという点、必ずしもそうではない。例えば、FactBank は意志・働きかけ・許可などの伝達の態度の注釈付けまではカバーしていない。これについては同じ Pustejovsky のグループが進めている TimeBank (TimeML) の MODALITY タグによって部分的にカバーされるが、TimeBank の MODALITY タグは事象動詞に接続する助動詞（must, may, should, would, could）をマークアップするに過ぎず、日本語など他の言語に適用するには何らかの拡張が必要である。

また、否定や疑問のスコープをどのように扱うかも大きな問題である。FactBank の現仕様では、次の例のように否定の焦点が修辞関係（因果関係）だけに当たるような場合の扱いについて論じられていない。

(13) *John thinks that Mary will get cured **not because** she took the medication (but because she has started practising yoga).*

この例のような否定や疑問のスコープについては、修辞関係の否定や疑問に限れば、Penn Discourse TreeBank の attribution タグ [38] がカバーしている。しかし、否定や疑問の焦点になるのは修辞関係だけとは限らず、次のような部分否定も含め、さらに検討が必要である。

(14) **全員**がこの案に賛成しているというわけではない。

もちろん、こうした議論の先には常に時相論理のような精密な意味表現の体系が視野に入ってくるが、我々の目的はそうした過度に複雑な意味表現を導入することなく、広い範囲の言語情報編集に有益で現実的な注釈付けの枠組みを設計し、それを実現する解析モデルを開発することである。

こうした背景から、我々のグループでも広義のモダリティ情報を注釈付けする枠組みを設計し、実際にコーパスへのタグ付けも進めている [42, 6, 26]. 我々の枠組みでは、FactBank がカバーする情報に伝達の態度や否定・疑問のスコープを加えたものになっており、次のような7つ組みで広義のモダリティ情報を表現する。

(15) a. これからはお酒を飲むことを控えようと思います。

判断主体	時制	仮想	態度	真偽判断	価値判断	焦点
wr	未来	-	意志	高確率から低確率	ネガティブ	-

b. 全員がこの案に賛成しているというわけではない。

判断主体	時制	仮想	態度	真偽判断	価値判断	焦点
wr	非未来	-	叙述	成立	-	焦点 (否定, 全員)

2.5 談話関係

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の研究だけでなく、談話単位間の意味的關係を把握する必要がある。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の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とするかでいくつかの異なる理論的な枠組みが提案され [24, 12, 52], 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その組み合わせについてどのような条件が成り立つべきかを定義している。

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、こうした定義に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 [49].

一方、接続表現に着目して談話単位間の関係の情報を付与する試みも報告されている [21, 30]. 例えば、Penn Discourse TreeBank (PDTB) [30] では、主に“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がかりに談話単位間の関係を付与する。関係を付与する際には、どのセグメントがどのセグメントと関係しているかを接続表現の項としてタグ付与する。

(16) *After* [*arg2* adjusting for inflation] the Commerce Department said [*arg1* spending didn't change in September]

例えば、例 (16) では接続詞 *After* に関して、“spending...September”を ARG1, “adjusting...inflation”を ARG2 として関係付ける。PDTB では、このように接続表現を手がかりに関係を認定するが、接続表現が出現していない場合や特定の範囲内での関係の認定しか行わないため、網羅性の点で課題が残る。

また、日本語でも横山ら [55], 新森ら [46] 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。また我々も、原因、理由、根拠等の因果関係に限定して注釈付けを試みている [15]. しかしながら、こうした注釈付けの成果の中には公開されていないものも多く、得られた資源や知見の共有・蓄積が進んでいないのが現状である。

3 意味情報の重層的付与に向けて

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 [8, 7] が挙げられる。Prague Dependency Treebank では、形態素と依存構造の注釈付けに加え、Tectogrammatcs と呼ばれるレイヤを用意し、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を付与する。

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 [40]. 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 [39]. Pustejovsky らはその後、XBank ブラウザ⁴と呼ばれるツールを開発し、PropBank, NomBank,

⁴<http://timeml.org/ula/xbank-browser/>

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 LDC (Linguistic Data Consortium) から配布されている⁵。

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、OntoNotes プロジェクト [37] では、Penn Treebank スタイルの統語構造、PropBank/NomBank スタイルの意味役割割他、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 [13] 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。

日本語でも、GDA は形態素、統語情報から意味役割、省略・共参照、修辞構造までをカバーする重層的なマークアップ言語になっており、上述のような海外での動きに比べても極めて先駆的な提案であったと言える。京都コーパス 4.0 も形態素、依存構造、述語項構造 (格構造)、共参照をカバーする重層的注釈付きコーパスであり、早い時期からデータが公開され、これらのレイヤの解析技術の研究に重要な役割を果たしている。一方、NAIST コーパスでは、述語項構造および共参照の注釈の規模を京都コーパス 4.0 の数倍に拡大した他、仕様面でも京都コーパス 4.0 や GDA コーパスの問題点を分析し、改善をはかった。

これらの重層的な注釈付けによって期待される効果の一例を示そう。述語項構造解析において項がゼロ照応のとき、次の例 (17) のように、先行詞と認め得る名詞句が前方文脈中に複数個出現する場合がある (この例では 1 文目の“村山富市首相”と 2 文目の“首相”)。

(17) 就任後初めて地元の大分県へ里帰りしていた**村山富市首相**_i は三十一日夕、三泊四日の日程を終えて日航機で羽田空港に到着した。**首相**_i は記者団に対し、「突然大分に帰ったが、温かい歓迎に接し_{ガ:i} 『地元はいいなあ』という感謝の気持ちでいっぱい、期待に応じてしっかり頑張らないといかんという気持ちを一層強く持った」と感想を述べた。

このような場合、それら複数の先行詞候補が共参照関係にあることが分かっているならば、解析器がいずれを選んで正解と判定できるし、むしろこうした例が示唆するように、述語項構造解析と共参照解析は同時並行的に解くべき問題である可能性もある。こうしたレイヤ間の相互作用を考えることは、仕様の設計・洗練のみならず、解析手法の改善に繋がる可能性も秘めている。

もちろん課題も多い。それぞれのレイヤで解決すべき課題が残っていることは本稿でこれまで見てきた通りである。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、2.1 節で触れた IRA と ISA の問題をやはり再検討する必要が出てくる。モダリティや談話関係、語義等の他のレイヤへ実際の注釈付け作業を拡張することも急がねばなるまい。さらに、これらのコーパスはいずれも新聞記事を対象にしており、例えば代名詞による共参照が少ないなど、ジャンルによる偏りは免れない。本特定研究のコアデータへの注釈付けを進めることによって、新聞記事とは異なる言語現象に現在の仕様が適合するかどうかを分析し、仕様の洗練をはかる必要がある。

異なる注釈仕様間の互換性 (compatibility) や相互連携性 (interoperability) も重要な課題である。国際的には、異なるリソース間の互換性に関する調査 [28] が行われたり、相互連携性を確保するための枠組み UIMA (Unstructured Information Management Architecture) [51] によるリソースの共有も進んでいる。2008 年には言語資源の相互連携性に関する初めての国際会議 [1] が開催されるなど、研究者の関心も高まっている。こうした動向にも注意を払いながら、日本語のリソース同士はもとより、他言語との相互連携性を築く活動にも注力していく必要がある。

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。また、一部は、(独) 情報通信研究機構の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」、科研費若手研究 (スタートアップ) 「類義述語句同定のための語彙的知識の体系化と集積」(課題番号: 20800029, 代表: 松吉俊) の支援を受けている。

参考文献

- [1] *The First International Conference on Global Interoperability for Language Resources*, 1998.

⁵Unified Linguistic Annotation Text Collection, LDC2009T07. <http://www ldc.upenn.edu/Catalog/>

⁶<http://www.anc.org/>

- [2] 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.
- [3] 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.
- [4] H. H. Clark. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press, 1977.
- [5] 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.
- [6] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. 日本語文章の事象に対する判断情報アノテーション. 情報処理学会研究報告書, 2009-NL-193, 2009. (to appear).
- [7] J. Hajič, E. Hajičová, J. Hlaváčová, V. Klimeš, J. Mirovský, P. Pajas, J. Štěpánek, B. Vidová, V. Hladká, and Z. Žabokrtský. PDT 2.0 — Guide, 2006. <http://ufal.ms.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/pdf/pdt-guide.pdf>.
- [8] E. Hajičová. Prague Dependency Treebank: From analytic to tectogrammatical annotation. *Proceedings the First International Conference on Text, Speech, Dialogue*, pp. 45–50, 1998.
- [9] 橋本泰一. 拡張固有表現タグ付きコーパスの構築に向けて—白書、書籍、yahoo! 知恵袋コアデータ—. 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ予稿集, pp. 71–76, 2009.
- [10] H. 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>.
- [11] L. Hirschman. *MUC-7 coreference task definition*. version 3.0, 1997.
- [12] J. R. Hobbs. On the coherence and structure of discourse. Technical Report 85-37, CSLI, 1985.
- [13] 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.
- [14] 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.
- [15] 飯田龍, 乾健太郎, 松本裕治. 根拠情報抽出の課題設計と予備実験. 言語処理学会第 15 回年次大会発表論文集, pp. 817–820, 2009.
- [16] Naoya Inoue, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Resolving direct and indirect anaphora for japanese definite noun phrases. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, 2009.
- [17] R. Jackendoff. *Semantic Structures*. Current Studies in Linguistics 18. The MIT Press, 1990.
- [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] 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.
- [20] 川添愛, 齊藤学, 片岡喜代子, 戸次大介. 確実性判断に関わる意味的文脈アノテーション. 情報処理学会研究報告書, 2009-NL-189, pp. 77–84, 2009.
- [21] A. Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, University of Edinburgh, 1995.
- [22] M. Komachi, R. Iida, K. Inui, and Y. Matsumoto. Learning based argument structure analysis of event-nouns in japanese. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp. 120–128, 2007.
- [23] 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.
- [24] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.
- [25] 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.
- [26] 松吉俊, 佐尾ちとせ, 江口萌, 乾健太郎, 松本裕治. 判断情報タグ付与コーパス作成の作業基準 第 0.7 版, 2009. <http://cl.naist.jp/nltools/modality/manual.pdf>.
- [27] Ben Medlock and Ted Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 992–999, 2007.
- [28] A. Meyers, A. C. Fang, L. Ferro, S. Kübler, T. Jia-lin, M. Palmer, M. Poesio, A. Dolbey, K. K. Schuler, and E. Loper. Annotation Compatibility Working Group Report. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, pp. 38–53, 2006.

- [29] 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.
- [30] 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.
- [31] R. Mitkov, editor. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education, 2002.
- [32] V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th ACL*, pp. 104–111, 2002a.
- [33] 信本浩司, 木下恭子, 黒橋禎夫. モダリティおよび用言のガ格情報を付与したコーパスの作成. 言語処理学会第 6 回年次大会発表論文集, pp. 20–23, 2000.
- [34] 小原京子, 斎藤博昭. フレーム意味論と『日本語コーパス』に基づく日本語彙情報資源『日本語フレームネット』の構築. 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ, デモ・ポスターセッション, 2009.
- [35] 奥村学, 白井清昭. Bccwj を用いた新しい語義曖昧性解消タスク. 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ予稿集, pp. 143–146, 2009.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] V. Rubin, E. Liddy, and N. Kando. Certainty identification in texts: Categorization model and manual tagging result. In *Computing Attitude and Affect in Text: Theories and Applications*, chapter 7, pp. 61–74. Springer-Verlag, 2005.
- [42] 佐尾ちとせ, 江口萌, 松吉俊, 乾健太郎. 日本語文のモダリティ・極性情報を捉えるために. 言語処理学会第 15 回年次大会発表論文集, pp. 793–796, 2009.
- [43] R. Saurí. *FactBank 1.0 Annotation Guidelines*. http://www.cs.brandeis.edu/roser/pubs/fb_annotGuidelines.pdf, 2008.
- [44] 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/anguide.1.2.1.pdf>, 2006.
- [45] R. Saurí and J. Pustejovsky. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 2009.
- [46] 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.
- [47] 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.
- [48] György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 38–45, 2008.
- [49] 竹内和広. テキスト修辭構造タグ付けの半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.
- [50] 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.
- [51] Graham Wilcock. *Introduction to Linguistic Annotation and Text Analytics: Synthesis Lectures on Human Language Technologies*. Morgan & Claypool, 2009.
- [52] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, Vol. 31, No. 2, pp. 249–287, 2005.
- [53] F. Wolf, E. Gibson, A. Fisher, and M. Knight. The discourse graphbank: A database of texts annotated with coherence relations, 2005.
- [54] 山梨正明. 推論と照応. くろしお出版, 1992.
- [55] 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.