同一指示関係を記述するためのアノテーションモデルの検討

---MUC-7 と MATE を比較して---

吉田悦子　　谷村緑
三重大学人文学部　　京都外国語大学
tantan@human.mie-u.ac.jp　　tani@hat.hi-ho.ne.jp

1. はじめに

本研究の目的は, 英語の談話における同一指示関係を記述するための言語学的情報を整理し, 既存のアノテーションモデルである MUC-7 と MATE のタグ体系を比較しながら, 先行詞の概念を重視する記述よりも, 指示対象の概念を重視する記述のほうが有用性が高いことを明らかにすることである.こうした視点は, テキスト内の首尾一貫性をより正確に分析するためには欠かせないだけでなく, 人間が推論やメト二ミー的な認知能力を介して指示表現を理解していることを明示的に説明する基盤となる.さらに分析が困難とされる話しことばの分析に応用可能な指示体系について考察し, 話しことばと書きことば, 語りと対話のようにジャンルの異なる談話においても利用出来る記述方法を検討する.

2. 概要：同一指示関係のアノテーションスキーマ

本稿は, 談話内において導入された談話要素（discourse entity）を指示する表現形式である指示表現（referring expression）を同一指示関係（co-reference）に基づいて談話内から抽出する際に必要なアノテーションの記述方法について検討する.同一指示関係を記述するために, 従来多くのアノテーションスキーマが提案されている.その主なものに, MUC-7 (Hirschman and Chinchor 1997), MEDSTRACT anaphoric annotation (Castaño et al. 2002), MATE (Poesio 2000, 2004, Poesio et al. 1999), UCREL anaphoric annotation (Figelstone 1992) 等があげられる. 対話コーパスに付与されたアノテーションには談話構造に考慮した HCRC Map Task Corpus が参考になる（Isard 2001）. このコーパスを基にした指示表現に関する研究もあるが(Bard et al 2000), 同一指示関係のタグ付与は実験的なものを除くと利用できるものは残念ながら存在しない.こうして見ると, 言語学的な目的, とりわけ, 意味と指示との関係をその談話コンテクストにおいてとらえることを目指している研究者にとって, 現実的に応用可能なものはほとんどない.その理由は, これらのスキーマはいずれも同一指示関係の記述方法において「先行詞」の情報を重視している点にある.同一指示関係とは基本的に指示される対象とそれを個々独立的に指示している表現の集合との関係のことである.有用なアノテーションにはこうした一貫した指示体系の見方を反映しているものが望ましい.この視点から提案しているアノテーションスキーマに関するモデルは, 対称的なアノテーションスキーマと呼ばれ, 有用であると評価されている（川添・コリアー）. 現行のアノテーションスキーマを検討すると, MUC-7 と MATE においては, 「共参照を記述するところの IDENT 関係は対称的(Symmetrical）かつ推移的(Transitive) な関係であって, 非対称的(Asymmetrical) ではないと明言されている」と, 評価されている.

実際の応用に際して, どんなテキストにどのアノテーションを適用するかも問題になる.谷村ほか（2007）は MEDSTRACT のタグ付与の方法を検討し, 教育応用の視点から英語学習者コーパスにタグ付けを試行的におこなっている.しかしながら, MEDSTRACT は分子生物学および医学分野のテキストからの情報抽出のために開発されたスキーマであり, 話しことば, とりわけ対話ベースのテキストへの応用にはなじみにくい面があると考えられる. こうしたジャンルへの配慮をふまえ, 対話コーパスにおける同一指示関係の記述に関しては, MUC-7 と MATE に限定して, 両者を比較検討したい.

3. MUC-7 と MATE の概要

MUC-7 は用語抽出等さまざまな用途に使用されているアノテーションで書きことばのテキストベースで開発された.一方, MATE は対話コーパスのアノテーションのために開発されたスキーマである.対象とするテキストが異なっていることは, アノテーションスキーマの体系に影響を与えている.つまり, 同一指示表現としてアノテートするものをどこまで

認めるか, さらにテキスト外の情報や推論的なつながりをどこまで体系に組み込んでいくかに相違点がみられるのではないかと考えられる.

検討するアノテーションスキーマの概要については以下を参照している.
MUC-7:
http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

MATE:
http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_8. html

以下の章で, 基本的な条件としてどんな言語表現を同一指示とみなすかについて, 両者のアノテーションの記述方法を比較検討する.

4. 基本的な条件

4.1 名詞句

談話に最初に導入される談話要素は, 基本的に名詞句(NP)である. 私たちが使用する指示表現とは, 'any expression in an utterance to refer to something or someone , i.e. a particular referent in mind in uttering any word' (Hurford et al. 2007) と定義されるように, いかなる語を発話していても頭の中には特定の指示対象（referent）があり, それを指し示すために発話において用いられる言語表現なのである.特定の明示的な言語形式によりその対象にアクセス出来る場合には, 不定名詞句, 定名詞句, 固有名詞, 代名詞などが指示表現となるうるが, 照応表現の解釈は一つに決まらない場合もある.

(1) John likes Bill. He is crazy.
(2) John is a policeman. He is crazy.

(1)の場合, 二つの談話要素が導入され, 後続の同一指示表現の解釈は文法情報と語用論的情報との相互作用によりあいまいになる可能性がある.また, 定義より, (2)のようないわゆるコピュラ文において, a policeman は, 談話要素にはなりえない.

すでに定義したように, 同一指示関係は, 指示対象とそれを指示している個々の表現の集合との関係, であるので, 先行詞がどの表現と結びつくかは問題ではない.このことは MATE において明確である. MATE は最初に導入された名詞句に<coref:de>(de: discourse entity) というタグを付与し, 後続談話でその名詞句が指し示す対象に対して同一指示表現としてのタグを<coref:link> や <coref:anchor> として付与している.さらに現場照応における指示対象へのタグである<coref:ue> や 談話指示（discourse deixis）についての<coref:seg>を付与する. MATE, MUC-7 共に裸名詞も同様に扱い, oranges, orange juice, bananas（MATE）, drug, drugs, its contract drilling business, the contract drilling business（MUC-7）などもマークされる.また固有名詞についても同様にマークされる: the independent Parades Commission (MATE), Reuters Holding PLC – Reuters(MUC-7)

4.2 拡張する名詞句

MATE と MUC-7 ともに最大の名詞句をマークすることが強調されている.つまり, 名詞句中の主要部をふくめた全体を記述する：the coreference task, the last contract you will ever get, a large quantity of sugar , about 200,000 tons of sugar, Frederick F. Fernwhistle Jr. Ford Motor Co., George Rath など, 固有名も含む, 名詞群をマークする(MUC-7). このように名詞句は修飾部を含めた最大の名詞句をふくめる. たとえば, 同格 'Fred Frosty, the ice cream king of Tyson's Corner', 前置詞句つき名詞句 'a joint venture with Sony' などである.MATE, MUC-7 共に 'of construction' は全体の名詞句としてマークされることが可能である：some of the symptoms (MATE), the senior of the executives who will assume Holland's duties (MUC-7)

接続詞つき名詞句(conjoined NPs)は, 二つ以上の名詞句の結合であるが, MATE は初出の名詞をそのままマークするが, MUC-7 では後続の表現形式によって名詞のマークを考慮することを提案している. すでに例(2)で触れたが, 述語名詞は談話要素を導入しないのでコピュラ文において, 定・不定にかかわらず, マークしない MATE に対して, MUC-7 は定・不定共に認めており, コピュラとみなすことの可能な動詞で結ばれた述語名詞もマークしている.

同格句の扱いについては, MUC-7 が部分的な重なりを示すものから（'The criminals, often legal immigrants'）最大の名詞句として扱う可能性も示唆している. 一方, 否定に導かれる名詞句 'Ms Ima Head, never a great MUC fan,'はマークしないとしている. MATE では, 同格句が隣接している場合は後者を埋め込むとするが, 同格句が隣接していない場合は, 別々にタグをふると定めている.

MUC-7, MATE ともに, 代名詞や指示詞, 定名詞句の先行詞が節の場合, 指示関係は対象外として扱っていない.

4.3 代名詞

代名詞の扱いはほぼ共通しているが, 人称代名詞に加えて, MATE は指示代名詞, 所有代名詞, 不定代名詞をマークしていることを明示している. 再帰代名詞については MATE は動詞の項として目的語に置かれる場合のみ認めているが（Julie washed *herself*. Bill was talking to *himself*）, MUC-7 は emphatics 用法についても認めている(He is, *himself*, unsure of the outcome.). 所有代名詞である 'its chairperson' や属

格形名詞 ‘the man’s arm’はそのままマークされるか, 別々にするかは決まっていない. MATE, MUC-7 共にゼロ代名詞についてはマークしないと記述されている. また, コンマで同格句が区切られていない場合は分けないでマークする.

Bound anaphor と呼ばれるタイプの照応関係について, MATE, MUC-7 双方ともマークしている：
Nobody likes to lose *his* job.; *Every* man who knows *his* own mind.

MATE は, 照応関係の拡張として所有関係 (‘Louise and *Her* graduation; ‘Aime Jacquet praised *his* team’s’) のほか, 部分をあらわすもの (‘the seat of the chair’) もマークしている. MUC-7 では対象外である.

4.4 関係詞節を含む名詞句

関係詞は制限用法 ‘the rumor that the war had ended’は問題なくマークするが, 非制限用法は, 全体でマークすると定める MUC-7（‘the Penn Central Co., which used to run a railroad’）に対して, MATE は扱いについては関係詞節を含めるか含めないかの選択肢を残している.

4.5 動名詞

動名詞について, MATE はその機能から名詞句同様に扱う方針である(‘They had been accused of ignoring the environment’). MUC-7 では現在分詞の動詞形にあたるものはマークしないが (‘Slowing the economy is supported by some Fed officials; it is repudiated by others.’), 名詞として機能しているもの ‘Slowing of the economy’ はマークする点では MATE と同様である.

4.6 否定文あるいは疑問文

MATE は談話要素の存在が否定される文脈の場合, 後続する文脈で照応関係が成り立つときのみ（’I don’t want to buy *a car*. *It* would cost me too much money.）, タグをふる. MUC-7 のほうは, 次のどの場合でも名詞句をマークする.

(3)
I have a machete.
I don’t have a machete.
Do you have a machete?

疑問文については, MATE ではマークされている一方, MEU-7 では対象外とされている.以下のような疑問文が対象となる：

(4)
A: *Which route* do you want to take?
B: The Corning to Elmira route.

4.7 メトニミー

M ATE では対象外とされているが, MUC-7 では以下のような場合, メトニミーによる照応関係を考慮した記述が提案されている：

(5)
The White House sent its health care proposal to Congress yesterday. Senator Dole said the *administration*’s bill had little chance of passing.
(6)
Ford announced a new product line yesterday. *Ford* spokes man John Smith said *they* will start manufacturing widgets.

こうしたメトニミーは, 推論によって引き起こされる潜在的な情報に基づいて理解される照応現象であり（駒田・吉田 2008）, とりわけトポニミーと解釈することができる. この照応関係は重要な談話指示であり, とりわけ対話内での首尾一貫性にかかわっており, MATE で対象外となっているのは疑問である.

4.8 事象関係（Event relations）と間接照応（bridging）

MUC-7 では対象外とされているが, MATE では以下のような場合, 間接照応（bridging）を考慮した記述が提案されている：

(7)
There was an explosion. *The noise* was tremendous.

この指示には, 語用論的推論とよばれる「文脈や言語外の知識を背景として誘引される推論」（山梨 1992: 14）が働いている, ここでの the noise は, 一般的な知識のフレームによる同一指示関係を維持しているだけでなく, 談話の展開において焦点移動の連鎖を形成する役目もある（山梨 2004）. 以下の例も同様に扱われる：

(8)
Muslims from all over the world *were taught gun-making and guerrilla warfare* in Afghanistan.
The instructors were members of some of the most radical Islamic militant groups in the region.

4.9 流暢さに欠ける発話 (Disfluencies)

MED において, 繰り返されているほうの名詞, もしくは完全な名詞句を採用する. MUC-7 では書きことばのテキストを扱うためこのような現象は現れてこない.

4.10 不連続な談話要素（Discontinuous elements）

対話コーパスを扱う MATE では, コーパスデータが chunk レベルで切ってある場合, ひとつの談話要素が複数の発話にまたがる現象としての不連続性を

無視できないので, 照応表現に ‘next’ および ‘prev’ という attributes のタグを付与して関連づけておく. これは, 対話の特徴としてあげられる現象で, 談話要素を導入する際に起こるもので, 「両属連鎖」と呼ばれる. 関与する発話者が一人か, 複数かによって二つのパターンに分けられる. 同一の発話者が一度発話を中断したあと, 再びもとの発話を継続する発話を産出する場合の多くは, 相手の対話者からのあいづち (back-channels) をはさんで起こる（吉田 2007）. 以下がその例である：

(9)
Giver: Curving, just curving round *the diamond*
Follower: uh-huh
Giver: *mine*
…uh-huh

5. 結び

以上, 本稿では同一指示関係の「対称的モデル」に基づくアノテーションスキーマの中から, MATE と MUC-7 を取り上げ, 現在なされているタグ付与について概観し, 条件項目ごとに両者の比較を試みた.

既に述べたが, 同一指示関係を出来るだけ忠実に記述するためには, 先行詞を規定するのではなく, 指示対象を談話コンテクストの変化に応じて復元できるような一貫した記述体系を確立することが非常に重要であることがわかった. 今後は, 両者についてより詳細な検討を加え, 双方からの有用性をとりこみながらより包括的で, かつ対話コーパスのジャンルに応用可能なタグ付けの記述方法を決定する必要がある. そしてさらに, 実際のデータに基づいたタグ付与による分析をおこない, 修正を加えた上で具体的な提案に結びつけていきたい.

参考文献

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., and Newlands, A. (2000). ‘Controlling the intelligibility of referring expressions’. *Journal of Memory and Language*, 42 (1): 1-22.

Castaño, J., Zhang, J., and Pustejovsky, J. (2002) Anaphora Resolution in Biomedical Literature. In International Symposium on Reference Resolution. Alicante, Spain. http://medstract.org/papers/coreference.pdf

Fligelstone, S. (1992) ‘Developing a Scheme for Annotationg Text to Show Anaphoric Relations’. In G. Leitner (ed.), *New Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter. 153-170.

Hirschman, L and N.Chinchor (1997) MUC-7 Coreference Task Definition, Version 3.0 in *Proceeding of MUC-7*. http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

Hurford, J.H., B.Heasley, and M.Smith. (2007) *Semantics: a coursebook* Cambridge University Press

Isard, Amy (2001) ‘An XML architecture for the HCRC Map Task Corpus’ (BI-DIALOG 2001), In *The HCRC Map Task Corpus Annotations Version 1.0*, Human Communication Research Centre/University of Edinburgh & University of Glasgow http://www.hcrc.ed.ac.uk/maptask/originalxml.html

川添愛, ナイジェル・コリアー(2003) 「対称モデルに基づく共参照関係アノテーションスキーマ」『情報処理』vol.2003. No.23, 39-45.

駒田ゆき子・吉田悦子(2008)「日本人英語学習者にみられる照応表現の特徴と問題点--センタリングモデルによる分析の可能性 --」『言語処理学会第 14 回年次大発表論文集』（投稿中）.

Poesio, Massimo (2000) *MATE Dialogue Annotation Guidelines (2.4Coreference)* http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html

Poesio, Massimo, Florence Bruneseaux, and Laurent Romary (1999) "The MATE meta-scheme for coreference in dialogues in multiple language", *Proceeding of the ACL Workshop on Standards for Discourse Tagging*. Maryland, 65-74.

Poesio, Massimo (2004) "The MATE/GNOME Proposals for anaphoric annotation, revisited", *Proceeding of SIGDIAL*, Boston, April.

谷村緑, 和泉絵美, 竹内和広, 井佐原均(2007) 「日本人英語学習者の談話における共参照関係の記述方法の検討」『言語処理学会第 13 回年次大会発表論文集』（478-481）

山梨正明（1992）『推論と照応』東京：くろしお出版

山梨正明（2004）『ことばの認知空間』東京：開拓社

吉田悦子(2007)「対話における逸脱文のパターンと発話解釈について」『言語処理学会第 13 回年次大会発表論文集』(262-265)