『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの構築

宮内 拓也♠　浅原 正幸♣　中川 奈津子♡　加藤 祥♢

♠♣♢ 人間文化研究機構 国立国語研究所 コーパス開発センター

♠ 東京外国語大学大学院 総合国際学研究科

♡ 日本学術振興会 特別研究員 (PD) / 千葉大学大学院 融合科学研究科

{♠ t-miyauchi, ♣ masayu-a, ♢ yasuda-s}@ninjal.ac.jp

♡ nakagawanatuko@gmail.com

1 はじめに

冠詞がない言語を母語とする者にとって，冠詞がある言語を習得する際の冠詞選択は難しいものである [1]．冠詞選択は，一般に定性や特定性などの情報構造が大きな影響を与えると考えられる．

言語処理の分野では英語母語話者が産出した大量のテキストから，英語学習者の冠詞の誤りを検出する手法が提案されている [2]．しかし，日本語母語話者が産出する他言語の冠詞選択を検討する場合，日本語における名詞句の情報構造を考慮する必要がある．さらに機械翻訳において日本語文を冠詞のある言語に翻訳する際にも，日本語の情報構造が問題となってくる．

本稿では，機械翻訳での冠詞選択の問題に関する基礎研究として,『現代日本語書き言葉均衡コーパス』[3] (以下，BCCWJ) のテキスト内の名詞句に対して情報構造に関わる文法情報のアノテーションを行った結果を報告する．

日本語の情報構造に関する過去のアノテーションは，主としてテキスト中に出現する情報が談話中に既出であること（情報状態）を共参照情報として付与するものであった．本研究は，より汎用性を求めるため，情報状態，定性，特定性のみならず，名詞句に情報構造関係の様々な項目でタグを付与した．

2 関連研究

情報構造のアノテーションには，当該の言語形式の情報構造をどのように決定するかという点で，二つのタイプがある．

まず初めに，当該の言語形式に基づいて情報構造を決める研究がある．例えば，Calhoun ら [4] は，Vallduví and Vilkuna や Steedman の研究 [5, 6] に言及し，韻律を採用した．L+H*LH%の韻律を持つ形式はテーマ (Theme; 概してトピック (Topic) に対応) となり，H*LH%の韻律を持つものはレーマ (Rheme; Focus) となる．そして，彼らは当該の NP が以前に言及されたか否か，またそれが以前述べられた個体から言及可能か否かという点をもとに情報構造をアノテーションした．Hajičová ら [7] は語順を用いた情報構造のアノテーションを提案した．この研究は情報構造についてプラハ学派の伝統に触発されたものであり，それゆえに動詞より左にある言語形式をトピックとする．これらのアノテーション基準は言語依存であり，日本語に適用可能なものではない．

二つ目のタイプの研究では，言語学的なテストを採用している．Götze ら [8] は言語に依存せず，かつ特定の言語理論にもよらず，情報状態 (given/accessible/new) とトピック (aboutness topic/frame setting topic)，フォーカス (new-information focus/contrastive focus) をアノテーションするための基準を策定した．例えば，アバウトネストピックは以下の手続きで決定される [8, p.165]:

(1) An NP X is the aboutness topic of a sentence S containing X if
   a. S would be a natural continuation to the announcement *Let me tell you something about X*
   b. S would be a good answer to the question *What about X?*

c. S could be naturally transformed into the sentence *Concerning X, S′*, where S′ differs from S only insofar as X has been replaced by a suitable pronoun.

本研究は Götze ら [8] に沿うものであるが，いくつかの点で彼らの研究とは大きく異なっている．まず，本研究ではトピックとフォーカスを直接アノテーションしない．これはトピックやフォーカスがそれぞれに多次元であるためである [9]．実際に Götze ら [8, p.163] は様々な種類のアバウトネストピックを区別している (指示的 NP，特定解釈や総称解釈を持つ不定の NP など)．定性や特定性のような要因は独立であると考え，そのようにアノテーションする方がより単純である．第 2 にトピックやフォーカスと相関すると知られている要因はより多くある (例えば，有生性や動作主性 [10, 11])．そのため，本研究では情報構造アノテーションの一環としてこれらの要因をアノテーションする．

3 タグとアノテーション基準

BCCWJ では，長単位と短単位という二つの単位が採用されているが，本研究では，短単位の名詞をアノテーション対象とする．

以下の (2) で示す項目についてラベルを設定した．

(2) a. 情報状態 (information status)
b. 定性 (definiteness)
c. 特定性 (specificity)
d. 有生性 (animacy)
e. 有情性 (sentience)
f. 動作主性 (agentivity)
g. 共有性 (commonness)

(2a) の情報状態とは，いわゆる旧情報と新情報の区別である．ある談話において，新たな情報は「新情報 (discourse-new)」となり，聞き手が知っている情報は「旧情報 (discourse-old)」となる．

(3) a. 担任だった池田弘子先生は違った
b. スクールカウンセラーでもあった先生の授業は
(読売新聞 [BCCWJ: PN1c_00001])

(3a) の下線部の名詞は新情報であり，(3b) の下線部の名詞は旧情報である．これらの名詞は共参照関係にある．

(2b) の定性とは，指示対象を聞き手が同定できるか否かを示すカテゴリーである．指示対象を聞き手が同定できると話し手が想定していれば「定 (definite)」であり，同定できないと想定している場合は「不定 (indefinite)」である．本研究では，スコープとして前後 3 文を見ることとする．

(4) そんな薄い$_{a.}$ かばんじゃ$_{b.}$ 遊び道具も入らないよ
(読売新聞 [BCCWJ: PN1c_00001])

(4) の下線部 *a.* の名詞は定であり，(4) の下線部 *b.* の名詞は不定である．

(2c) の特定性は，話し手が特定の事物を想定しているか否かを示す意味論的カテゴリーである．話し手が特定の事物を想定しているならば「特定 (specific)」となり，想定していなければ「不特定 (unspecific)」となる．定性と同様，スコープとして前後 3 文を見ることとする．

(5) 行き場を失った$_{a.}$ 廃タイヤがあぜ道や$_{b.}$ 納屋の横に放置されてきた
(北海道新聞 [BCCWJ: PN2e_00001])

(5) の下線部 *a.* の名詞は特定であり，(5) の下線部 *b.* の名詞は不特定である．

(2d) の有生性とは，生きているか否かを示すカテゴリーである．生物（人間，動物など）は「有生 (animate)」であり，無生物（植物を含む）は「無生 (inanimate)」である．有生性は名詞句レベルに付与されるものとする．有生性と似た概念として (2e) の有情性がある．これは，情意があるか否かを示すパラメターである．自由意志による移動が可能な場合は「有情 (sentient)」となり，自由意志による移動はないなら「非情 (insentient)」となる．日本語については，有生/無性の区別よりも有情/無情の区別が重要であるとする研究もありまた，有生性と有情性の値が異なる場合もあることから，このパラメターの設定が必要となる．情意の有無は名詞句単体では判定できない場合があるため，有情性は述語-項レベルに付与されるものとする．

(6) オオクチバスなどの$_{a.}$ ブラックバス類が、少なくとも四十三都道府県の七百六十一のため池や$_{b.}$ 湖沼に侵入し、
(読売新聞 [BCCWJ: PN4c_00001])

(6) の下線部 *a.* の名詞は有生・有情であり，(6) の下線部 *b.* の名詞は無生・無情である．

(2f) の動作主性は，事態に関わる人がその事態で果たしている役割を示す．行為を意図的に実現するもの

は「動作主 (agent)」とし，行為によって変化を被るものを「被動作主 (patient / theme)」とする．このパラメターは節レベルに付与し，主節と従属節の両方を考慮することとする．また,「どちらでもよい」「どちらでもない」を許す．

(7) a. 編み笠をかぶった人なつっこい笑顔を見るだけで、
b. 独特な雰囲気の写真になりました
(産経新聞 [BCCWJ: PN1d_00001])

(7a) の下線部の名詞は，主節では被動作主であり従属節では動作主である．このような場合に「どちらでもよい」というタグを付与する．(7b) の下線部の名詞は動作主でも被動作主でもないため,「どちらでもない」となる．

(2g) の共有性は，情報を聞き手が既に知っていると話し手が想定しているか否かを示すパラメターである．聞き手が既に知っていると話し手が想定している情報は「共有 (hearer-old)」であり，知らないと想定している情報は「非共有 (hearer-new)」である．なお，この判断の際はアノテータの世界知識を使ってもよいとし,「想定可能」というラベルも許す．このラベルは，ブリッジング (bridging) を起こしている際に付与される．

(8) a. キャンティ街道を抜け、b. オリーブ畑に囲まれた田園地帯のc. レストランで、
(読売新聞 [BCCWJ: PN4c_00001])

(8) の下線部 *a.* の名詞は共有であり，下線部 *b.* の名詞は非共有である．下線部 *c.* の名詞はブリッジングを起こしているため,「想定可能」となる．

固有名詞については，アノテーションの際，有名の度合いを考慮してよいこととし，アノテータの持つ世界知識を参照してもよいとする．(9a) の形式名詞や (9b) のような慣用表現は対象から外し，それぞれ「形式名詞」,「慣用表現」タグを付与する．なお，慣用表現であるか否かについてはアノテータによる揺れを許すこととする．

(9) a. 様々な人がいるということが
b. 聞く耳を持たせてくれるんです
(読売新聞 [BCCWJ: PN1c_00001])

4 基礎統計

対象は BCCWJ の新聞 (PN) コアデータ 16 サンプルに出現する名詞 2023 件とした．サンプルの選択は `BCCWJ-ANNOTATION-ORDER` に基づく．作業者は BCCWJ-DepParaPAS[12, 13] に付与された共参照情報を確認しながら作業を行う．定性，特定性，有生性，有情性，動作主性については，与えられた文脈から判断できない場合に「どちらでもよい」というタグを認めた．特定性，動作主性，共有性については，その概念が認めがたい場合に「どちらでもない」というタグを認めた．

表 1 にタグの基礎統計を示す．情報状態のラベルは以前アノテーションされた共参照情報に基づいている[1]．情報状態と他の分布は異なっている．ゆえに，この差異は日本語からの翻訳の際の冠詞選択に影響を与えると考えられる．

表 2 に情報状態と定性の分割表を示す．不定のラベルは新情報のラベルと共に現れることが多いが，興味深いことに定のラベルは新情報，旧情報のどちらのラベルとも現れうるという傾向がある．これは共参照情報の冠詞選択への貢献が限界的であることを示している．

表 3 は情報状態と特定性の分割表である．これについても情報状態と定性のものと同様の分布を示している．

5 おわりに: まとめと今後の課題

本稿では，BCCWJ に対する情報構造のアノテーションデータを紹介した．本研究では日本語の名詞句に対し，七つの情報構造に関する概念を導入した．これらのアノテーションラベルの分布は日本語母語話者の冠詞選択や冠詞誤りの修正に対し共参照情報だけでは十分でないことを示しており，新たに導入されたラベルは翻訳における冠詞選択に役立つと思われる．

今後の課題は以下に示すとおりである．

まず，情報構造をアノテーションする被験者実験を行う．本稿で示したラベルは言語学者によってアノテーションされたものである．非言語学者により情報構造のラベルを判定する質問を作成し，非言語学者にもわかるような言語学的なテストを設計する．それにより，アノテーションの数だけでなく，標的サンプルをも増やすことができる．

第 2 に，機械学習に基づきシソーラスを用いて情報構造の評価モデルを開発する．それにより，日本語テクストに基づき，機械翻訳による冠詞選択の評価を行う．

[1] 共参照情報アノテーションについての論文は未発行である．

表 1: タグの基礎統計

情報状態	新情報	旧情報	-	-
	1345	678	-	-
定性	定	不定	どちらでもよい	-
	1122	899	2	-
特定性	特定	不特定	どちらでもよい	どちらでもない
	1157	749	116	1
有生性	有生	無生	どちらでもよい	-
	342	1680	1	-
有情性	有情	無情	どちらでもよい	-
	337	1678	8	-
動作主性	動作主	被動作主	どちらでもよい	どちらでもない
	192	338	2	1491
共有性	共有	非共有	想定可能	どちらでもない
	1036	494	489	4

表 2: 情報状態と定性

	新情報	旧情報
定	497	625
不定	846	53
どちらでもよい	2	0

表 3: 情報状態と特定性

	新情報	旧情報
特定	531	626
不特定	705	44
どちらでもよい	108	8
どちらでもない	1	0

最後に，本稿で示した情報構造アノテーションを視線計測のデータ [14] と重ね合わせる．これにより，情報構造と視線の関係がより深く解明されることが期待される．

謝辞

本研究は JSPS 科研費 (課題番号：25284083，研究代表者：浅原正幸) の助成を受けたものである．

参考文献

[1] Tania Ionin, Heejeong Ko, and Kenneth Wexler. Article semantics in L2 acquisition: The role of specificity. *Language Acquisition*, Vol. 12, No. 1, pp. 3–69, 2004.

[2] Ryo Nagata, Tatsuya Iguchi, Fumito Masui, Atsuo Kawai, and Isu Naoki. A statistical model based on the three head words for detecting article errors. *IEICE TRANSACTINOS on Information ans Systems*, Vol. E88-D, No. 7, pp. 1700–1706, 2005.

[3] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[4] Sasha Calhoun, Malvina Nissim, Mark Steedman, and Jason Brenier. A framework for annotating information structure in discourse. In Adam Meyers, editor, *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pp. 45–52, Ann Arbor, 2005. The Association for Computational Linguistics.

[5] Enric Vallduví and M Vilkuna. On rheme and kontrast. In P. W. Culicover and L. McNally, editors, *The Limits of Syntax*, pp. 79–108. Academic Press, San Diego, 1998.

[6] Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, Vol. 34, pp. 649–689, 2000.

[7] Eva Hajičová, Jarmila Panevová, and Petr Sgall. A manual for tectogrammatical tagging of the Prague Dependency Treebank. Technical report, ÚFAL/CKL, 2000. (TR-2000-09).

[8] Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. Information structure. In Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors, *Information structure in cross-linguistic corpora: annotation guidelines for phonology, morphology, syntax, semantics and information structure*, Vol. 7, pp. 147–187. Universitätsverlag Potsdam, 2007.

[9] Natsuko Nakagawa. *Information structure in spoken Japanese: Particles, word order, and intonation.* PhD thesis, Kyoto University, Kyoto, 2016.

[10] Talmy Givón. Topic, pronoun, and grammatical agreement. In Charles N. Li, editor, *Subject and Topic*, pp. 149–187. Academic Press, New York, 1976.

[11] Edward L. Keenan. Towards a universal definition of “subject”. In Charles N. Li, editor, *Subject and Topic*, pp. 303–334. Academic Press, New York, 1976.

[12] 植田禎子, 飯田龍, 浅原正幸, 松本裕治, 徳永健伸. 『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション. 第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214, 2015.

[13] 浅原正幸, 大村舞. BCCWJ-DepParaPAS:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化. 言語処理学会第 22 回年次大会発表論文集, pp. 489–492, 2016.

[14] Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. Reading-time annotations for balanced corpus of contemporary written japanese. In *Proceedings of COLING-2016*, 2016.