日本語における名詞句の情報構造と語順の相関について

宮内 拓也
東京外国語大学大学院 /
日本学術振興会 特別研究員
miyauchi.takuya.k0 @ tufs.ac.jp

浅原 正幸
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
masayu-a @ ninjal.ac.jp

1 はじめに

日本語は格要素ではなく情報構造により語順が決まる傾向にある．Asahara ら [1] は,『現代日本語書き言葉均衡コーパス』[2] (以下，BCCWJ) のコアデータに対する述語項構造・共参照情報アノテーション [3] を用いて，ガ・ヲ・ニすべての格要素を含む二重目的語構文を抽出し，その共参照情報を用いて情報の新旧を付与し，直接目的格 (ヲ) か間接目的格 (ニ) かよりも，情報の新旧が語順に影響を与える傾向を明らかにした．しかしながら，名詞句の特徴は，共参照に基づく情報の新旧以外にも多様な情報構造があるほか，二重目的語構文以外の場合についてどのような語順になるのかを調査する必要がある．

そこで，本稿では，BCCWJ 内の名詞句に対して情報構造に関わる文法情報のタグをアノテーションした研究 [4] を利用して，情報構造に影響する文法情報がどのように語順に影響を及ぼすのかについて調査する．具体的には，名詞句と名詞句の係り先との距離を，情報状態・共有性・定性・有生性の観点を特徴量として，ベイズ線形混合モデル (Bayesian Linear Mixed Model; [5]) により回帰する．

結果，名詞句は文中で,「旧情報 ＞ 新情報」,「共有 ＞ 想定可能 ＞ 非共有」,「不定 ＞ 定」,「有生 ＞ 無生」の順で並ぶという推定結果が得られた．

2 方法

2.1 使用したデータ

Miyauchi ら [4] は BCCWJ のテキスト (新聞 (PN) コアデータ 16 サンプル，全 16,657 語 (短単位)，5,195 文節，739 文) 内の名詞句 2,023 に対し，情報構造に関係する文法情報のラベル (情報状態，共有性，定性，特定性，有生性，有情性，動作主性) をアノテーションしている．各名詞句に付与されたラベルの値は以下 (1) の通りであり，アノテーション作業は図 1 に例示するような形式で人手でなされた[1]．

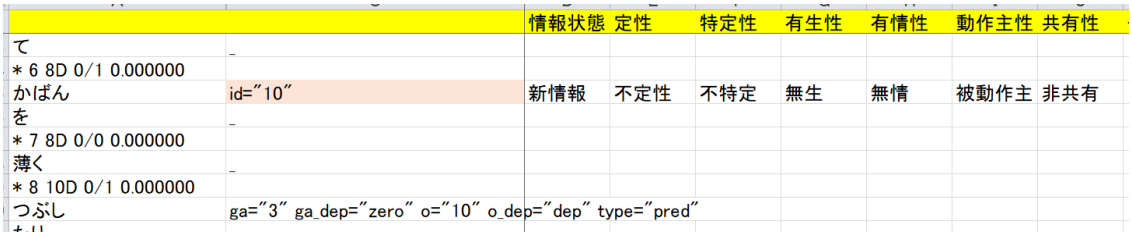
(1) a. 情報状態: 「新情報」/「旧情報」
b. 定性: 「共有」/「非共有」/「想定可能」
c. 特定性: 「定」/「不定」
d. 有生性: 「特定」/「不特定」
e. 有情性: 「有生」/「無生」
f. 動作主性: 「有情」/「非情」
g. 共有性: 「動作主」/「被動作主」

本研究では (1) の内，情報状態 (新情報/旧情報), 共有性 (共有/想定可能/非共有), 定性 (定/不定)，有生性 (有生/無生) の値を用いた．情報状態は書き手の観点に基づく情報の新旧で，共有性は読み手の観点に基づく情報の新旧である．定性は日英翻訳の際に問題となる情報であり，語順との関連が分かることにより冠詞推定に寄与する可能性がある．有生性は，存在表現の差異 (いる／ある) のほか，格交代を表出する使役・受動態との関連があり，さらには語順に影響を及ぼす可能性がある．

2.2 モデル化の方法

本研究では，情報状態，共有性，定性，有生性の各ラベルをもとに，係り元名詞句とその係り先文節との距離 dist を間に入る文節数によりベイズ線形混合モデルで評価した．アノテーションされた名詞句 2,023 のうち，係り先がある[2]名詞句 1,939 を分析対象とした．係り受けの情報として，BCCWJ-DepPara [6] の情報を用いた．

[1] 各ラベルの付与の基準などは [4] 等を参照のこと．
[2] 係り先がないものの多くは体言止めである．



		情報状態	定性	特定性	有生性	有情性	動作主性	共有性
て	_							
* 6 8D 0/1 0.000000								
かばん	id="10"	新情報	不定性	不特定	無生	無情	被動作主	非共有
を	_							
* 7 8D 0/0 0.000000								
薄く	_							
* 8 10D 0/1 0.000000								
つぶし	ga="3" ga_dep="zero" o="10" o_dep="dep" type="pred"							

図 1: 情報構造のアノテーション

具体的には以下のような線形式でモデル化を行った：

$$\text{dist} \sim \text{Normal}(\mu, \sigma),$$
$$(\text{但し}, \mu \leftarrow \alpha + \beta^{*}_{\text{情報状態}} + \beta^{*}_{\text{共有性}} + \beta^{*}_{\text{定性}} + \beta^{*}_{\text{有生性}}).$$

ここで Normal は平均 μ 標準偏差 σ の正規分布とし，切片 α と各カテゴリパラメータ $\beta^{\text{旧情報}}_{\text{情報状態}}, \beta^{\text{新情報}}_{\text{情報状態}}, \beta^{\text{共有}}_{\text{共有性}}, \beta^{\text{想定可能}}_{\text{共有性}}, \beta^{\text{非共有}}_{\text{共有性}}, \beta^{\text{定}}_{\text{定性}}, \beta^{\text{不定}}_{\text{定性}}, \beta^{\text{有生}}_{\text{有生性}}, \beta^{\text{無生}}_{\text{有生性}}$ の取りうるパラメータ割り当ての線形結合で平均 μ を定式化した．

これを rstan パッケージを用いて推定する．warmup 後のイテレーションを 15,000 回に設定し，4 回シミュレーションを実施した．全てのモデルは収束した．

3 結果

統計処理の結果，表 1 の結果が得られた．

表 1: 結果

Parameter	Rhat	mean
α	1.172	0.243
$\beta^{\text{新情報}}_{\text{情報状態}}$	1.072	-0.632
$\beta^{\text{旧情報}}_{\text{情報状態}}$	1.022	1.147
$\beta^{\text{定}}_{\text{定性}}$	1.002	1.147
$\beta^{\text{不定}}_{\text{定性}}$	1.021	0.488
$\beta^{\text{有生}}_{\text{有生性}}$	1.068	0.442
$\beta^{\text{無生}}_{\text{有生性}}$	1.155	-0.697
$\beta^{\text{共有}}_{\text{共有性}}$	1.123	1.332
$\beta^{\text{想定可能}}_{\text{共有性}}$	1.202	0.764
$\beta^{\text{非共有}}_{\text{共有性}}$	1.002	-1.975
σ	1.148	0.063
log-posterior	1.019	28198.286

Rhat は収束判定パラメータで 1.2 以下を収束とみなす．mean は事後平均であり，各ラベルにより係り先との距離が長くなるか（+ 方向）短くなるか（− 方向）を数値で示す．日本語は主辞が後置されるために係り先との距離が長くなる要素が，語順においては先行する要素となる．これをもとに各名詞句の語順を推定すると，表 2 の通りとなり，名詞句は文中で，「旧情報 ＞ 新情報」，「共有 ＞ 想定可能 ＞ 非共有」，「不定 ＞ 定」，「有生 ＞ 無生」の順で並ぶという推定結果が得られた．

4 考察

4.1 情報状態，共有性

情報状態に関しては，旧情報の名詞より新情報の名詞のほうが後方に来やすく，共有性に関しては共有の名詞が前方に，非共有の名詞が後方に来やすいという結果であった．つまり，「談話中に指示される否かにかかわらず，聞き手が知っている (と話し手が想定している) 情報は聞き手が知らない (と話し手が想定している) 情報より文中で前方に位置する」と一般化できることになる．

これらの結果は，機能的文眺望 (Functional Sentence Perspective) の研究で議論される，担う伝達情報の量に応じて少ない情報量のものから多いものへ順に並べられるという Firbas [7] による「伝達のダイナミズム」(Communicative Dynamism) や，文中で旧情報を前に置き新情報を後ろに置くという久野 [8] による「旧から新へのインフォーメイションの流れ」(2) を支持することとなった．

(2) 「旧から新へのインフォーメイションの流れ」文中の語順は，古いインフォーメイションを表す要素から，新しいインフォーメイションを表す要素へ進むのを原則とする． [8, p.59]

表 2: 名詞句とその係り先との距離

	推定結果	推定された語順
情報状態	$\beta^{\text{新情報}}_{\text{情報状態}} < \beta^{\text{旧情報}}_{\text{情報状態}}$	旧情報 ＞ 新情報 の順
共有性	$\beta^{\text{非共有}}_{\text{共有性}} < \beta^{\text{想定可能}}_{\text{共有性}} < \beta^{\text{共有}}_{\text{共有性}}$	共有 ＞ 想定可能 ＞ 非共有 の順
定性	$\beta^{\text{不定}}_{\text{定性}} < \beta^{\text{定}}_{\text{定性}}$	定 ＞ 不定 の順
有生性	$\beta^{\text{無生}}_{\text{有生性}} < \beta^{\text{有生}}_{\text{有生性}}$	有生 ＞ 無生 の順

4.2 定性

定性については，定の名詞句より不定の名詞句のほうが後に来やすいという結果であった．つまり,「定の名詞句は不定の名詞句より文中で前方に位置する」と一般化できることになる．

前述の「伝達のダイナミズム」[7] によれば，担う伝達情報の量に応じて少ない情報量のものから多いものへ順に並べられるため，文脈的依存度が相対的に高い定の名詞句が前方に現れやすく，逆に文脈的独立度が高い不定の名詞句が後方に現れやすいことが予測される．実際に，日本語同様に冠詞による定性の表示を行わないスラヴ系の言語ではかつてより定性と語順の関係が指摘されており [9, 10, 11]，それは伝達のダイナミズムに沿うものである．例えば，(3) はロシア語の例文だが，文の後方に位置する (3a) の *lampa*「ランプ」は不定で解釈されやすく，文の前方に位置する (3b) の *lampa* は定で解釈されやすいとされる．

(3) a. Na stale stojala lampa.
on table stood lamp

b. Lampa stojala na stole.
lamp stood on table
「テーブルの上にランプがあった.」
[10, p.266]

本研究の結果からは，日本語においても，同様の傾向が見られると言える．

4.3 有生性

有生性に関しては，有生の名詞より無生の名詞のほうが後方に来やすいという結果であった．この結果から見れば,「有生の名詞句は無生の名詞句より文中で前方に位置する」と一般化できることになる．

これは一般的な有生性階層 (有生 ＞ 無生) に従う形となっている．例えば，Silverstein [12] は，文の成立において名詞句の階層でより高い位置にあるものが主語となる文構造が優先されるとする．Silverstein の階層をもとに，Dixon の研究 [13] などを受けて角田 [14] が修正した名詞句階層を図 2 に示す．図中の左方が階層の上位を，右方が下位を表す．

日本語ではスクランブリング等による語順の変更がない限り主語は他の要素より前方に現れるため，この結果は Silverstein の主張を支持することになると考えられる．

5 おわりに

本稿では，情報構造に関係する文法情報がどのように語順に影響を及ぼすのかについて調査するため，BCCWJ 内の名詞句に対して付与された情報構造に関わる文法情報のラベル [4] の各値を利用し，値が付与された名詞句の文末からの距離をベイズ線形混合モデル [5] によりモデル化した結果を報告した．

本研究では，名詞句は文中で,「旧情報 ＞ 新情報」,「共有 ＞ 想定可能 ＞ 非共有」,「不定 ＞ 定」,「有生 ＞ 無生」の順で並ぶという推定結果が得られた．これらの結果は，先行研究の「伝達のダイナミズム」[7] や「旧から新へのインフォーメイションの流れ」[8],「名詞句階層」[12] を支持するものである．

謝辞

本研究は JSPS 科研費 (課題番号: 17J07534) の助成を受けている．本研究の一部は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」によるものである．

参考文献

[1] Masayuki Asahara, Satoshi Nambu, and Shin-Ichiro Sano. Predicting japanese word order

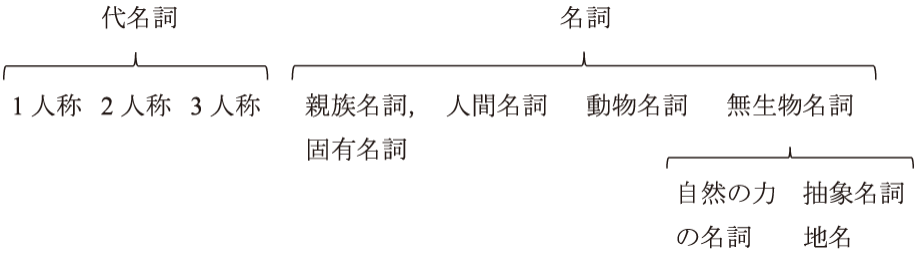


図 2: 名詞句階層

in double object constructions. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pp. 36–40, Melbourne, July 2018. Association for Computational Linguistics.

[2] 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.

[3] 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.

[4] Takuya Miyauchi, Masayuki Asahara, Natsuko Nakagawa, and Sachi Kato. Information-structure annotation of the “Balanced Corpus of Contemporary Written Japanese”. In Kôiti Hasida and Win Pa Pa, editors, *Computational Linguistics*, Vol. 781 of *Communications in Computer and Information Science*, pp. 155–165, Singapore, 2018. Springer.

[5] Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, Vol. 12, No. 3, pp. 175–200, 2016.

[6] 浅原正幸, 松本裕治. 『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション. 自然言語処理, Vol. 25, No. 4, pp. 331–356, 2018.

[7] Jan. Firbas. On the concept of communicative dynamism in the theory of functional sentence perspective. In *Sborník prací Filozofické fakulty Brněnské univerzity A19*, pp. 135–144. 1971.

[8] 久野すすむ. 談話の文法. 大修館書店, 東京, 1978.

[9] Jiří Krámský. *The article and the concept of definiteness in language*. Mouton de Gruyter, Hague, 1972.

[10] Catherine V. Chvany. Notes on ‘root’ and ‘structure-preserving’ in Russian. In C. Corum, T.C. Smith-Stark, and A. Weiser, editors, *You take the high node and I will take the low node*, pp. 252–290. Chicago Linguistic Society, Chicago, IL, 1973.

[11] Aleksander Szwedek. A note on the relation between the article in English and word order in Polish. In Fisiak Jacek, editor, *Papers and Studies in Contrastive Linguistics*, Vol. 2, pp. 213–225. Adam Mickiewicz University Press, Poznań, 1974.

[12] Michael Silverstein. Hierarchy of features and ergativity. In Robert M. W. Dixon, editor, *Grammatical categories in Australian languages*, pp. 112–171. Australian National University, Canberra, 1976.

[13] Robert M. W. Dixon. Ergatuvuty. *Language*, pp. 59–138, 1979.

[14] 角田太作. 世界の言語と日本語. くろしお出版, 東京, 1991.