Predicting Japanese Word Order in Double Object Constructions

日本語の二重目的語構文の基本語順について

浅原 正幸 (国立国語研究所) *
南部 智史 (モナシュ大学)
佐野 真一郎 (慶應義塾大学)

Predicting Japanese Word Order in Double Object Constructions

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Satoshi Nambu (Monash University)
Shin-Ichiro Sano (Keio University)

要旨

本稿では日本語の二重目的語構文の基本語順について予測する統計モデルについて議論する。『現代日本語書き言葉均衡コーパス』コアデータに係り受け構造・述語項構造・共参照情報を悉皆付与したデータから、二重目的語構文を抽出し、格要素と述語要素に分類語彙表番号を付与したうえで、ベイジアン線形混合モデルにより分析を行った。結果、名詞句の情報構造の効果として知られている旧情報が新情報よりも先行する現象と、モーラ数が多いものが少ないものに先行する現象が確認された。分類語彙表番号による効果は、今回の分析では確認されなかった。

1. はじめに

日本語は語順が自由な言語である。日本語の語順に影響を与える影響について、主に計算言語学の分野で調査されてきた (Yamashita and Kondo 2011, Orita 2017)。一つの良く知られている知見 'long-before-short' (Yamashita and Chang 2001) として、かき混ぜにより長い名詞句が短い名詞句よりも前に傾向がある。本稿では、そのなかで日本語の二重目的語構文に注目する。二重目的語構文でニ格名詞句 (1) とヲ格名詞句 (2) のどちらが先行するのかを検討する：

(1) 太郎が 花子に 本を あげた

(2) 太郎が 本を 花子に あげた

日本語においてはどちらの語順も可能であるため、理論言語学においては何が正規語順かについて、ある語順は他の語順から派生されているという仮説に基づいて議論されてきた (Hoji 1985, Miyagawa 1997, Matsuoka 2003)。本稿では、単にコーパス中の二重目的語構文の頻度

* masayu-a@ninjal.ac.jp

表 1　先行研究との比較

	(Sasano and Okumura 2016)	(Orita 2017)	本研究
コーパス	ウェブコーパス	NAIST テキストコーパス	BCCWJ-PAS BCCWJ-DepPara
ジャンル	ウェブ	新聞記事	新聞記事, 書籍, 雑誌, 白書 ブログ, Q/A サイト
対象	ガ-ニ-ヲ-述語	ガ-ヲ-述語	ガ-ニ-ヲ-述語
文書数	n/a	2,929	1,980
文数	100 億語	38,384	57,225
文型数	648 types × 350,000 samples	3,103 tokens	584 tokens
分析対象	verb types	syntactic priming, NP length, given-new, and animacy	NP length, and given-new
分析方法	線形回帰・ NPMI	ロジスティック回帰 (glm)	ベイジアン線形混合モデル (rstan)

を数えることにより正規語順について議論するのではなく、理論言語学などの先行研究で言及されている様々な要因を考慮した統計分析を行う。この目的のために、語順に影響を与える要因を考慮したベイジアン線形混合モデルを用いて分析を行う。

'long-before-short' 以外の要因として、文脈中の名詞句の情報状態が語順に影響を与えるという理論的な枠組がある (Lambrecht 1994, Vallduví and Engdahl 1996)。この枠組では、ある名詞句が文脈中で情報の状態としてどのような機能をもつか、情報の新旧・トピック（話題）・フォーカス（焦点）などの観点を導入する。この情報の状態が語順を決める基本的な要因の一つであるということを、次の二つの理由に基づいて仮定する：(1) 日本語の文中、談話に既出の要素は、談話の未出の要素に先行する (Kuno 1978, 2004, Nakagawa 2016)，(2) 日本語の文中、フォーカス（焦点）もしくは新情報は述語の直前に出現する傾向にある (Kuno 1978, Kim 1988, Ishihara 2001, Vermeulen 2012)。一般的な日本語の語順に関するこれらの 2 つの主張に基づいて、二重目的語構文について以下のような仮説を検討する。

(3)　仮説:
二重目的語構文において、談話に既出の要素は他の要素より左に位置する傾向にあり、談話に未出の要素は他の要素より右に位置する傾向にある。

先行研究で提案されている 'long-before-short' 関連の要因と名詞句の情報状態を考慮したうえで、日本語二重目的語構文の語順を推測する統計モデルを構築する。本研究の新規性として、述語項構造と共参照が人手でアノテーションされたものを使うことで、代名詞（既出）か否か（未出）のような二分的な分析でではなく、先行文脈に名詞句の指示対象が出現している回数を含めて考慮する点があげられる。

2.　先行研究

表 1 は、コーパスに基づく日本語の語順の分析に関する直近の先行研究について示す。

Sasano and Okumura (2016) は日本語の二重目的語構文の単文レベルの正規語順について、大規模なウェブコーパスを用いて「ガ-ニ-ヲ-述語」と「ガ-ヲ-ニ-述語」のどちらが優勢かにつ

いて調査した。形態素解析器 JUMAN と係り受け・格解析器 KNP により自動解析した 100 億文を準備し、統語的な曖昧性がない部分木を取り出して分析資料を準備した。この資料に対し、動詞タイプ (SHOW 型 or PASS 型) を要因とした語順の分析を線形回帰と相互情報量 (normalized pointwise mutual information) に基づいて分析した。彼らのモデルでは共参照のような文間の関係を調査していない。

Orita (2017) は、述語項構造と共参照情報が人手で付与された NAIST テキストコーパスのアノテーションを用いて、直接目的語と主語の語順のかき混ぜを予測する統計モデルを構築した。彼女の調査では、プライミング・名詞句の長さ・有生性・情報状態（既出／未出）の効果を調査した。頻度主義的な統計分析（単純なロジスティック回帰）では、主語と目的語の語順について、情報状態に関する効果は観察されなかった。

本研究では、共参照情報が二重目的語構文の語順に影響を与える要因であると想定して、人手による係り受け・述語項構造・共参照を重ね合わせたデータを用いた、ベイジアン線形混合モデルに基づく調査を実施する。

3. 実験

3.1 データ：BCCWJ-PAS

本研究では、『現代日本語書き言葉均衡コーパス』(以下、BCCWJ) (Maekawa et al. 2014) に対する述語項構造と共参照のアノテーション BCCWJ-PAS を研究資料として用いる。このアノテーション基準は NAIST テキストコーパス (Iida et al. 2007) のものに準じる。このデータに、文節係り受けアノテーション BCCWJ-DepPara (Asahara and Matsumoto 2016) を重ね合わせることで、直接係り受け関係がある 主語（ガ）・直接目的語（ヲ）・間接目的語（ニ）・述語の 4 つ組を抽出する。ゼロ照応や格交替の事例を排除した結果、57,225 文から 584 例の 4 つ組を抽出した。

図 1 に Yahoo! 知恵袋サンプル (OC09_04653) の例を示す。表層文字列は、我々の語順を評価する際の相対距離の基本単位である文節単位に区切られている。相対距離は 4 つ組のなかから以下の 6 対を評価する：ガ-述語 ($\mathrm{dist}^{subj}_{pred}$), ヲ-述語 ($\mathrm{dist}^{dobj}_{pred}$), ニ-述語 ($\mathrm{dist}^{iobj}_{pred}$), ガ-ニ ($\mathrm{dist}^{subj}_{iobj}$), ガ-ヲ ($\mathrm{dist}^{subj}_{dobj}$), ニ-ヲ ($\mathrm{dist}^{iobj}_{dobj}$)。距離は対の左要素と右要素の文節に基づく距離を評価する。例えば、図 1 において、$\mathrm{dist}^{subj}_{pred}$ は「彼女が」と「使います」の距離を 4 とする。

日本語の語順の傾向として ‘long-before-short’ の効果を調査するために、項名詞句の長さを統計モデルの固定効果として導入する。項名詞句の長さは、BCCWJ-PAS でラベルづけされている項名詞句の右端を最右要素とし、係り受け木上で項名詞句を根とした場合の部分木の左端を最左要素とし、この 2 要素間の発音形のモーラ数に基づきガ格モーラ数 (N^{subj}_{mora}), ヲ格モーラ数 (N^{dobj}_{mora}), ニ格モーラ数 (N^{iobj}_{mora}) を定義する。例えば、図 1 において N^{subj}_{mora} は「その彼女が（ソノカノジョガ）」のモーラ数 6 と定義する。なお、部分係り受け木の最大スパンに基づくために、項名詞句の長さは 2 文節以上に対して定義する場合もある。

さらに、given-new ordering を確認するために、先行文脈における共参照要素数を固定効果に入れる。BCCWJ-PAS のアノテーションから得られたガ格名詞句・ヲ格名詞句・ニ格名詞

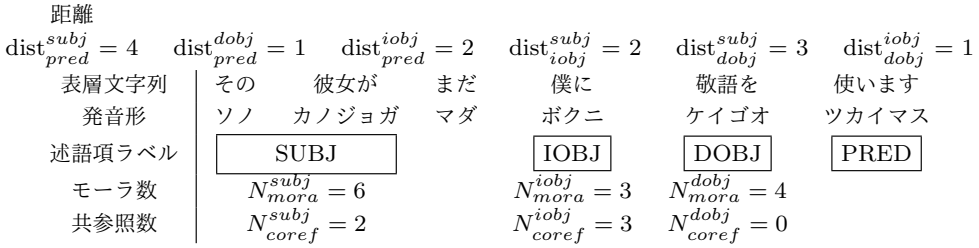

図 1　BCCWJ Yahoo! 知恵袋サンプルの例: (OC09_04653)

表 2　基礎統計

	min	1Q	med	mean	3Q	max
$\mathrm{dist}^{subj}_{pred}$	1.0	4.0	5.0	5.8	7.0	23.0
$\mathrm{dist}^{dobj}_{pred}$	1.0	1.0	1.0	1.7	2.0	13.0
$\mathrm{dist}^{iobj}_{pred}$	1.0	1.0	2.0	2.3	3.0	17.0
$\mathrm{dist}^{subj}_{iobj}$	-14.0	1.0	3.0	3.5	5.0	21.0
$\mathrm{dist}^{subj}_{dobj}$	-10.0	2.0	3.0	4.1	5.0	22.0
$\mathrm{dist}^{iobj}_{dobj}$	-12.0	-1.0	1.0	0.6	2.0	16.0
N^{subj}_{mora}	2.0	4.0	5.0	6.5	8.0	32.0
N^{dobj}_{mora}	2.0	3.0	4.0	5.3	6.0	37.0
N^{iobj}_{mora}	2.0	4.0	5.0	6.1	7.0	52.0
N^{subj}_{coref}	0.0	0.0	1.0	6.9	6.0	105.0
N^{dobj}_{coref}	0.0	0.0	0.0	0.5	0.0	44.0
N^{iobj}_{coref}	0.0	0.0	0.0	3.1	1.0	99.0

句の共参照数を $N^{subj}_{coref}, N^{dobj}_{coref}, N^{iobj}_{coref}$ として定義する。表 2 に、距離・モーラ数・共参照数の基礎統計量を示す。

3.2　統計処理

ベイジアン線形混合モデル (Sorensen et al. 2016) (BLMM) に基づき、2 つの項の間の距離もしくは項と述語の間の距離を評価する。具体的には、次の式に基づき統計モデルを作成する：

$$\begin{aligned}\mathrm{dist}^{left}_{right} &\sim \mathrm{Normal}(\mu,\sigma)\\ \mu &\leftarrow \alpha+\beta^{subj}_{mora}\cdot N^{subj}_{mora}+\beta^{subj}_{coref}\cdot N^{subj}_{coref}\\ &\quad +\beta^{dobj}_{mora}\cdot N^{dobj}_{mora}+\beta^{dobj}_{coref}\cdot N^{dobj}_{coref}\\ &\quad +\beta^{iobj}_{mora}\cdot N^{iobj}_{mora}+\beta^{iobj}_{coref}\cdot N^{iobj}_{coref}.\end{aligned}$$

ここで $\mathrm{dist}^{left}_{right}$ は left 要素と right 要素の文節を単位とした距離を意味する。例えば

$\text{dist}_{iobj}^{subj}$ は、主語 subj (left) と間接目的語 iobj(right) の間の文節に基づく距離 (隣接は 1) を表す。左右が反対の場合には負の値を持つ。これを、平均値 μ と標準偏差 σ の正規分布によりモデル化する。平均値 μ は切片 α と、2 タイプの変数の線形式により定義する。1 つ目のタイプの変数 $N_{mora}^{subj}, N_{mora}^{dobj}, N_{mora}^{iobj}$ は、主語・直接目的語・間接目的語のモーラ数に対するものである。それぞれ 2 文節以上の場合には、文節境界を越えてモーラ数を数える。2 つ目のタイプの変数 $N_{coref}^{subj}, N_{coref}^{dobj}, N_{coref}^{iobj}$ は、主語・直接目的語・間接目的語の共参照先行詞数に対するものである。β_b^a は、変数 N_b^a に対する傾きを表す。

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 2000 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。

4. 結果と考察

4.1 結果

表 3 距離の推定結果

距離	α	β_{mora}^{subj}	β_{mora}^{dobj}	β_{mora}^{iobj}	β_{coref}^{subj}	β_{coref}^{dobj}	β_{coref}^{iobj}	σ
$\text{dist}_{pred}^{subj}$	4.814***	0.146***	-0.031	0.040	0.002	-0.056	-0.009	3.323
	(0.375)	(0.040)	(0.042)	(0.032)	(0.011)	(0.043)	(0.016)	(0.100)
$\text{dist}_{pred}^{dobj}$	1.593***	-0.009	0.061***	-0.032**	-0.001	0.037**	-0.005	1.072
	(0.128)	(0.013)	(0.014)	(0.011)	(0.004)	(0.014)	(0.005)	(0.032)
$\text{dist}_{pred}^{iobj}$	2.100**	-0.022	-0.056**	0.112***	-0.018***	-0.045	0.037***	1.861
	(0.217)	(0.022)	(0.023)	(0.018)	(0.006)	(0.024)	(0.009)	(0.055)
$\text{dist}_{iobj}^{subj}$	2.668***	0.171***	0.026	0.071**	0.020	-0.011	-0.046**	3.577
	(0.420)	(0.043)	(0.045)	(0.035)	(0.012)	(0.047)	(0.017)	(0.108)
$\text{dist}_{dobj}^{subj}$	3.205***	0.155***	-0.092**	0.072**	0.003	-0.094**	-0.004	3.452
	(0.404)	(0.041)	(0.043)	(0.034)	(0.012)	(0.046)	(0.017)	(0.103)
$\text{dist}_{dobj}^{iobj}$	0.502	-0.013	-0.117***	0.143***	-0.017**	-0.081**	0.041***	2.436
	(0.287)	(0.029)	(0.030)	(0.024)	(0.008)	(0.033)	(0.011)	(0.071)

** > ±2SD, *** > ±3SD

表 3 に BLMM により推定されたパラメータ値を示す。値は平均値と標準偏差（カッコ内）による。

まず、主語と述語間の距離 ($\text{dist}_{pred}^{subj}$) は、主語のモーラ数にのみ影響を受ける。主語のモーラ数が多いほどその述語との距離が長くなる傾向が見られた。

直接目的語と述語間の距離 ($\text{dist}_{pred}^{dobj}$) は、直接目的語のモーラ数・共参照先行詞数と、間接目的語のモーラ数の影響を受ける。i) 直接目的語のモーラ数が多いほど、述語からの距離が遠くなる、ii) 直接目的語の共参照先行詞数が多いほど、述語からの距離が遠くなる、iii) 間接目的語のモーラ数が多いほど、直接目的語と述語との距離が近くなる傾向が見られた。

間接目的語と述語間の距離 ($\text{dist}_{pred}^{iobj}$) は、間接目的語のモーラ数・共参照先行詞数と直接目的語のモーラ数・共参照先行詞数の影響を受ける。i) 間接目的語のモーラ数が多いほど、述語からの距離が遠くなる、ii) 間接目的語の共参照先行詞数が多いほど、述語からの距離が遠くなる、iii) 直接目的語のモーラ数が多いほど、間接目的語と述語との距離が近くなる、iv) 直接目的語の共参照先行詞数が多いほど、間接目的語と述語との距離が近くなる傾向が見られた。

二つの項名詞句間の距離 ($\text{dist}_{iobj}^{subj}$, $\text{dist}_{dobj}^{subj}$, and $\text{dist}_{dobj}^{iobj}$) も項名詞句ー述語間の距離と同じ傾向がある。しかしながら、項名詞句のモーラ数は項の長さ（構成する文節数）と相関があるため、最左項名詞句と最右項名詞句距離（例えば、主語と直接目的語間）が、間にある項名詞句（間接目的語）のモーラ数の影響を受ける。

4.2 考察

結果より、二重目的語構文において、主語は直接目的語や間接目的語より先行する傾向がわかった。間接目的語は、直接目的語より先行するが、有意ではなかった (p=0.09)。

共参照先行詞数に対する推定された係数 ($\text{dist}_{pred}^{dobj}$ に対する N_{coref}^{dobj} や、$\text{dist}_{pred}^{iobj}$ に対する N_{coref}^{iobj}) をみると、直接目的語と間接目的語間の語順に対して 'given-new ordering' の仮説を支持していることがわかる。共参照先行詞の数が多い目的語は、述語からの距離が遠くなる傾向がみられる。

モーラ数に対する推定された係数 ($\text{dist}_{pred}^{subj}$ に対する N_{mora}^{subj} や、$\text{dist}_{pred}^{dobj}$ に対する N_{mora}^{dobj} や、$\text{dist}_{pred}^{iobj}$ に対する N_{mora}^{iobj}) をみると、二重目的語構文の全ての項名詞句が 'long-before-short' の仮説を支持していることがわかる。ある目的語と述語間の距離に対して、もう一つの目的語のモーラ数に対する推定された係数が負の値であること ($\text{dist}_{pred}^{iobj}$ に対する N_{mora}^{dobj} や $\text{dist}_{pred}^{dobj}$ に対する N_{mora}^{iobj}) から、長い目的語が他の目的語に先行する傾向が見られた。

5. おわりに

本稿ではベイジアン線形混合モデルを用いて述語項構造・共参照アノテーションデータを分析することにより日本語の二重目的語構文の語順について検討した。結果、直接目的語と間接目的語の間に 'given-new ordering' の傾向が見られることがわかった。さらに全ての項名詞句について、'long-before-short' の傾向が確認された。

今後、名詞句の有生性や述語動詞の分類が二重目的語構文の語順に与える影響について分析する。このために項名詞句と述語動詞に対する分類語彙表番号付与の作業を進めている Kokuritsukokugokenkyusho (1964)。

謝　辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP15K12888, JP17H00917, JP18H05521 によるものです。

文　献

Hiroko Yamashita, and Tadahisa Kondo (2011). "Linguistic Constraints and Long-before-short Tendency." *IEICE Technocal report (TL)*:TL2011-19, pp. 61–65.

Naho Orita (2017). "Predicting Japanese scrambling in the wild." *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, pp. 41–45. Valencia, Spain: Association for Computational Linguistics.

Hiroko Yamashita, and Franklin Chang (2001). ""Long Before Short" Preference in the Production of a Head-final Language." *Cognition*, 81:2, pp. B45–B55.

Hajime Hoji (1985). “Logical Form Constraints and Configurational Structures in Japanese.” Unpublished doctoral dissertation, University of Washington.

Shigeru Miyagawa (1997). “Against Optional Scrambling.” *Linguistic Inquiry*, 28, pp. 1–26.

Mikinari Matsuoka (2003). “Two Types of ditransitive constructions in Japanese.” *Journal of East Asian Linguistics*, 12, pp. 171–203.

Knud Lambrecht (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Vol. 71. *Cambridge Studies in Linguistics.*: Cambridge University Press.

Enric Vallduví, and Elisabet Engdahl (1996). “The linguistic realization of information packaging.” *Linguistics*, 34:3, pp. 459–520.

Susumu Kuno (1978). *Danwa no bunpoo [Grammar of discourse]*. Tokyo: Taishukan Shoten.

Susumu Kuno (2004). “Empathy and direct discourse perspectives.” Lawrence Horn, and Gregory Ward (Eds.), *The handbook of pragmatics.*: Oxford: Blackwell. pp. 315–343.

Natsuko Nakagawa (2016). “Information Structure in Spoken Japanese: Particles, word order, and intonation.” Unpublished doctoral dissertation, Kyoto University.

Alan Hyun-Oak Kim (1988). “Preverbal focusing and type XXIII languages.” Jessica Wirth Michael Hammond, Edith A. Moravcsik (Ed.), *Studies in syntactic typology.*: Amsterdam: John Benjamins. pp. 147–169.

Shin-ichiro Ishihara“Stress, focus, and scrambling in Japanese.” Ora Matushansky Elena Guerzoni (Ed.), *MITWPL 39.*: Cambridge, MA: MITWPL. pp. 142–175.

Reiko Vermeulen (2012). “The information structure of Japanese.” Renate Musan Manfred Krifka (Ed.), *The expression of information structure.*: Berlin: De Gruyter Mouton. pp. 187–216.

Ryohei Sasano, and Manabu Okumura (2016). “A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2244. Berlin, Germany: Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto (2007). “Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations.” *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139. Prague, Czech Republic: Association for Computational Linguistics.

Masayuki Asahara, and Yuji Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annota-

tion Treebank on the 'Balanced Corpus of Contemporary Written Japanese'." *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 49–58. Osaka, Japan: The COLING 2016 Organizing Committee.

Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth (2016). "Bayesian Linear Mixed Models using Stan: A tutorial for psychologists, linguists, and cognitive scientists." *Quantitative Methods for Psychology*, 12:3, pp. 175–200.

国立国語研究所 (編) (1964).『分類語彙表』　秀英出版.