Predicting Japanese Word Order in Double Object Constructions

Masayuki Asahara
National Institute for Japanese
Language and Linguistics, Japan
Center for Corpus Development
masayu-a@ninjal.ac.jp

Satoshi Nambu
Monash University, Australia
School of Languages, Literatures,
Cultures and Linguistics
satoshi.nambu@monash.edu

Shin-Ichiro Sano
Keio University, Japan
Faculty of Business and Commerce
s-sano@keio.jp

Abstract

This paper presents a statistical model to predict Japanese word order in the double object constructions. We employed a Bayesian linear mixed model with manually annotated predicate-argument structure data. The findings from the refined corpus analysis confirmed the effects of information status of an NP as 'given-new ordering' in addition to the effects of 'long-before-short' as a tendency of the general Japanese word order.

1 Introduction

Because Japanese exhibits a flexible word order, potential factors that predict word orders of a given construction in Japanese have been recently delved into, particularly in the field of computational linguistics (Yamashita and Kondo, 2011; Orita, 2017). One of the major findings relevant to the current study is 'long-before-short', whereby a long noun phrase (NP) tends to be scrambled ahead of a short NP (Yamashita and Chang, 2001).

This paper sheds light on those factors in double object constructions (DOC), where either (1) an indirect object (IOBJ) or (2) a direct object (DOBJ) can precede the other object:

(1) Taro-ga Hanako-ni hon-o ageta.
Taro-SBJ Hanako-IOBJ book-DOBJ gave
'Taro gave Hanako a book.'

(2) Taro-ga hon-o Hanako-ni ageta.
Taro-SBJ book-DOBJ Hanako-IOBJ gave
'Taro gave Hanako a book.'

Since both of the word orders are available, studies in theoretical syntax have been disputing about what is the canonical word order under the hypothesis of deriving one word order (i.e., either IOBJ-DOBJ or DOBJ-IOBJ) from another in the context of derivational syntax (Hoji, 1985; Miyagawa, 1997; Matsuoka, 2003). In this paper, we do not attempt to adjudicate upon the dispute solely based on the frequency of the two word orders in a corpus, but aim to detect principal factors that predict the word order in the DOC, which may eventually lead to resolving the issue in theoretical syntax. To that end, we employed a Bayesian linear mixed model with potential factors affecting the word orders as a preliminary study.

Other than the factor 'long-before-short' proposed in previous studies, the key factor in the current study is an information status of an NP in a given context under the theoretical framework of information structure (Lambrecht, 1994; Vallduví and Engdahl, 1996). The framework provides us key categories, such as (informationally) given/old, new, topic, and focus, to classify an NP as how it functions in a particular context. We assume the information status as one of the principle predictors based on the following two reasons; (i) a discourse-given element tends to precede a discourse-new one in a sentence in Japanese (Kuno, 1978, 2004; Nakagawa, 2016), (ii) focused or new elements in Japanese tend to appear in a position immediately preceding the predicate (Kuno, 1978; Kim, 1988; Ishihara, 2001; Vermeulen, 2012). These two claims regarding the general word order of Japanese are combined into the following hypothesis regarding the word orders in the DOC.

(3) Our hypothesis:
In the DOC, a discourse-given object tends to appear on the left of the other object, and a discourse-new object tends to be on the right side.

Incorporating the information status of an NP with another factor 'long-before-short' proposed in the previous studies, we built a statistical model

Proceedings of the Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing, pages 36–40
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

Table 1: Comparison with Preceding Work

	(Sasano and Okumura, 2016)	(Orita, 2017)	The current work
corpus	Web Corpus	NAIST Text Corpus	BCCWJ-PAS and BCCWJ-DepPara
genres	Web	Newspaper	Newspaper, Books, Magazines, Yahoo! Answes, Blog, Whitepaper
target	SUBJ-IOBJ-DOBJ-PRED	SUBJ-DOBJ-PRED	SUBJ-IOBJ-DOBJ-PRED
documents	n/a	2,929	1,980
sentences	around 10 billion	38,384	57,225
tuples	648 types × 350,000 samples	3,103 tokens	584 tokens
features	verb types	syntactic priming, NP length, given-new, and animacy	NP length, and given-new
analysis	linear regression and NPMI	logistic regression (glm)	Bayesian linear mixed model (rstan)

to predict the word orders in the DOC. One important advantage of our study is that, with the latest version of the corpus we used (See Section 3), the information status of an NP can be analyzed not simply by bipartite groups as either pronoun (given) or others (new) but by the number of co-indexed items in a preceding text.

2 Preceding Work

Table 1 shows a comparison with the latest corpus studies on Japanese word ordering.

Sasano and Okumura (2016) explored the canonical word order of Japanese double object constructions (either SUBJ-IOBJ-DOBJ-PRED or SUBJ-DOBJ-IOBJ-PRED) by a large-scale web corpus. The web corpus contains 10 billion sentences parsed by the Japanese morphological analyzer JUMAN and the syntactic analyzer KNP. In their analysis, the parse trees without syntactic ambiguity were extracted from the web corpus, and the word order was estimated by verb types with a linear regression and normalized pointwise mutual information. Their model did not include any inter-sentential factors such as coreference.

Orita (2017) made a statistical model to predict a scrambled word order as (direct) object-subject. She used the NAIST Text corpus which has a manual annotation of predicate-argument structure and coreference information. She explored the effect of syntactic priming, NP length, animacy, and given-new bipartite information status (given was defined as having a lexically identical item in a previous text). Her frequentism statistical analysis (simple logistic regression) did not detect a significant effect of the given-new factor on the order of a subject and an object.

As a preliminary study which features coreferential information as a potential factor, we used manual annotation of syntactic dependencies, predicate-argument structures and coreference information, employing a Bayesian statistical analysis on the small-sized well-maintained data.

3 Experiments

3.1 Corpora: BCCWJ-PAS

We used the 'Balanced Corpus of Contemporary Written Japanese' (BCCWJ) (Maekawa et al., 2014), which includes morphological information and sentence boundaries, as the target corpus. The corpus was extended with annotations of predicate-argument structures as BCCWJ-PAS (BCCWJ Predicate Argument Structures), based on the NAIST Text Corpus (Iida et al., 2007) compatible standard. We revised all annotations of the BCCWJ-PAS data, including subjects (with case marker *-ga*), direct objects (with case marker *-o*), and indirect objects (with case marker *-ni*), as well as coreferential information of NPs. After the revision process, syntactic dependencies of BCCWJ-DepPara (Asahara and Matsumoto, 2016) were overlaid on the predicate-argument structures.

We extracted 4-tuples of subject (subj), direct object (dobj), indirect object (iobj) and predicate (pred) from the overlaid data. Excluding 4-tuples with zero-pronoun, case alternation, or inter-clause dependencies from the target data, we obtained 584 samples of the 4-tuples.

Figure 1 shows an example sentence from BCCWJ Yahoo! Answer sample (OC09_04653). The surface is segmented into base phrases, which is the unit to evaluate the distance between two constituents as in the following pairs of the 4-tuples: subj-pred (dist$^{subj}_{pred}$), dobj-pred (dist$^{dobj}_{pred}$), iobj-pred (dist$^{iobj}_{pred}$), subj-iobj (dist$^{subj}_{iobj}$), subj-dobj (dist$^{subj}_{dobj}$), and iobj-dobj (dist$^{iobj}_{dobj}$). The distance was calculated from the rightmost word in each pair. For example, in Figure 1, dist$^{subj}_{pred}$ is identified as the distance between “” and “” as 4.

Verifying effects of 'long-before-short' as a

Table 2: Basic Statistics

	min	1Q	med	mean	3Q	max
$\text{dist}^{subj}_{pred}$	1.0	4.0	5.0	5.8	7.0	23.0
$\text{dist}^{dobj}_{pred}$	1.0	1.0	1.0	1.7	2.0	13.0
$\text{dist}^{iobj}_{pred}$	1.0	1.0	2.0	2.3	3.0	17.0
$\text{dist}^{subj}_{iobj}$	-14.0	1.0	3.0	3.5	5.0	21.0
$\text{dist}^{subj}_{dobj}$	-10.0	2.0	3.0	4.1	5.0	22.0
$\text{dist}^{iobj}_{dobj}$	-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

general Japanese word-order tendency, lengths of constituents were modeled as fixed effects in the statistical analysis. The lengths of subject, direct object and indirect object were calculated based on a mora count (in pronunciation) available in BCCWJ as N^{subj}_{mora}, N^{dobj}_{mora}, and N^{iobj}_{mora}, respectively. For example, in Figure 1, N^{subj}_{mora} is the number of morae of “ ” (*sono kanojoga*), which is 6. Note that an NP may contain more than one base phrase including an embedded clause. We evaluated the maximum span of the dependency subtree in BCCWJ-DepPara as a length of the NP.

In addition, the numbers of coreferent items in a preceding text were modeled as fixed effects. The numbers of coreferent items for subject, direct object and indirect object were obtained from the BCCWJ-PAS annotations as N^{subj}_{coref}, N^{dobj}_{coref}, and N^{iobj}_{coref}, respectively. Table 2 shows the basic statistics of the distance, mora, and number of coreferent items.

3.2 Statistical Analysis

We used Bayesian linear mixed models (Sorensen et al., 2016) (BLMM) for the statistical analysis on the distance between arguments as well as an argument and its predicate. We modeled the following formula:

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

$\text{dist}^{left}_{right}$ (e.g. $\text{dist}^{subj}_{iobj}$: distance between subject (left) and indirect object (right)) stands for the distance between left and right elements, which is modeled by a normal distribution with average μ and stdev σ. μ is defined by a linear formula with an intercept α and two types of interest coefficients. N^{subj}_{mora}, N^{dobj}_{mora}, and N^{iobj}_{mora} are the number of morae of a subject, a direct object, and an indirect object, respectively. The subject and objects can be composed of more than one phrase, and when they contain a clause, the number of morae was defined with the clause length.

N^{subj}_{coref}, N^{dobj}_{coref}, and N^{iobj}_{coref} stand for the number of preceding coreferent NPs of a subject, a direct object, and an indirect, respectively. β^{a}_{b} are the slope parameters for the coefficients N^{a}_{b}. Note that the distance was measured by the number of base phrase units, and a minus value indicates a distance in an opposite direction.

We ran 4 chains × 2000 post-warmup iteration, and all models were converged.

4 Results and Discussions

4.1 Results

Table 3 shows the estimated parameters by the BLMM; the values are means with standard deviations (in brackets). The findings are summarized as follows.

First, the distance between a subject and its predicate ($\text{dist}^{subj}_{pred}$) is affected only by the number of morae of a subject, which indicates that a longer subject NP has a longer distance from its predicate.

Second, the distance between a direct object and its predicate ($\text{dist}^{dobj}_{pred}$) is affected by the number of morae of the direct object, the number of its preceding coreferent items, and the number of morae of the indirect object. It indicates that i) a longer direct object has a longer distance from its predicate, ii) a direct object with more coreferent items in a preceding text has a longer distance from its predicate, and iii) a longer indirect object makes shorter the distance between the direct object and its predicate.

Third, the distance between an indirect object and its predicate ($\text{dist}^{iobj}_{pred}$) is affected by the number of morae of the indirect object, the number of its preceding coreferent items, the number of morae of a direct object, and the number of preceding coreferent items of a subject. It indicates that i) a longer indirect object has a longer distance from its predicate, ii) an indirect object with more coreferent items in a preceding text has a longer distance from its predicate, iii) a longer direct ob-

$\text{dist}^{subj}_{pred}$ = 4 $\text{dist}^{dobj}_{pred}$ = 1 $\text{dist}^{iobj}_{pred}$ = 2 $\text{dist}^{subj}_{iobj}$ = 2 $\text{dist}^{subj}_{dobj}$ = 3 $\text{dist}^{iobj}_{dobj}$ = 1

surface					ł	
pronunciation	sono	kanojoga	mada	bokuni	keigoo	tsukaimasu
translation	that	she	yet	me	honorific-OBJ	use
predicate-argument	SUBJ			IOBJ	DOBJ	PRED
morae	$N^{subj}_{mora} = 6$			$N^{iobj}_{mora} = 3$	$N^{dobj}_{mora} = 4$	
coreference	$N^{subj}_{coref} = 2$			$N^{iobj}_{coref} = 3$	$N^{dobj}_{coref} = 0$	

Figure 1: Example sentence (BCCWJ Yahoo! Answers:OC09_04653)

Table 3: Evaluation of distances

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

** > ±2SD, *** > ±3SD

ject makes shorter the distance between the indirect object and its predicate, and iv) a subject with more coreferent items makes shorter the distance between the indirect object and its predicate.

The distance between arguments ($\text{dist}^{subj}_{iobj}$, $\text{dist}^{subj}_{dobj}$, and $\text{dist}^{iobj}_{dobj}$) represents nearly the same tendency as the combination of the predicate-argument distance. However, the number of morae of an argument is correlated with the length of the argument (i.e., the number of base phrases), and thus, the distance between the leftmost and rightmost arguments (e.g. subject, direct object) is affected by the number of morae of the middle argument (e.g. N^{iobj}_{mora}).

4.2 Discussions

The results revealed that the subject tends to precede the direct and indirect objects in the double object constructions. Although the indirect object tends to precede the direct object, it is not significant (p=0.09).

The estimated coefficients for the number of coreferent items (N^{dobj}_{coref} for $\text{dist}^{dobj}_{pred}$ and N^{iobj}_{coref} for $\text{dist}^{iobj}_{pred}$) support our hypothesis in (3) as 'given-new ordering' for the direct and indirect objects. An object with many preceding coreferent items tends to be farther from a corresponding predicate.

The estimated coefficients for the number of morae (N^{subj}_{mora} for $\text{dist}^{subj}_{pred}$, N^{dobj}_{mora} for $\text{dist}^{dobj}_{pred}$ and N^{iobj}_{mora} for $\text{dist}^{iobj}_{pred}$) indicate that the orders of all arguments in the DOC follow 'long-before-short'. It is also confirmed by the minus values as the estimated coefficients for the number of morae of one object in relation to the order of the other object and its predicate (N^{dobj}_{mora} for $\text{dist}^{iobj}_{pred}$ and N^{iobj}_{mora} for $\text{dist}^{dobj}_{pred}$), suggesting that a longer object tends to precede the other object in the DOC.

5 Conclusions

This article presents a Bayesian statistical analysis on Japanese word ordering in the double object constructions. It revealed the 'given-new ordering' for the indirect and direct objects and also confirmed the 'long-before-short' tendency for all of the arguments in the constructions.

Setting off from the current preliminary study, our future work is to investigate effects of verb type and animacy of an NP. We are currently annotating the labels of a Japanese thesaurus 'Word List by Semantic Principles' (WLSP) (Kokuritsukokugokenkyusho, 1964), which enables us to explore those effects.

Acknowledgments

This work was supported by JSPS KAKENHI Grants Numbers JP15K12888 and JP17H00917 and is a project of the Center for Corpus Development, NINJAL.

References

Masayuki Asahara and Yuji Matsumoto. 2016. BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 49–58, Osaka, Japan. The COLING 2016 Organizing Committee.

Hajime Hoji. 1985. *Logical Form Constraints and Configurational Structures in Japanese*. Ph.D. thesis, University of Washington.

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

Shin-ichiro Ishihara. 2001. Stress, focus, and scrambling in japanese. In Ora Matushansky Elena Guerzoni, editor, *MITWPL 39*, pages 142–175. Cambridge, MA: MITWPL.

Alan Hyun-Oak Kim. 1988. Preverbal focusing and type xxiii languages. In Jessica Wirth Michael Hammond, Edith A. Moravcsik, editor, *Studies in syntactic typology*, pages 147–169. Amsterdam: John Benjamins.

Kokuritsukokugokenkyusho, editor. 1964. *Bunruigoihyo [Word List by Semantic Principles]*. Shuei Shuppan.

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

Susumu Kuno. 2004. Empathy and direct discourse perspectives. In Lawrence Horn and Gregory Ward, editors, *The handbook of pragmatics*, pages 315–343. Oxford: Blackwell.

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

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*, 48:345–371.

Mikinari Matsuoka. 2003. Two types of ditransitive constructions in japanese. *Journal of East Asian Linguistics*, 12:171–203.

Shigeru Miyagawa. 1997. Against optional scrambling. *Linguistic Inquiry*, 28:1–26.

Natsuko Nakagawa. 2016. *Information Structure in Spoken Japanese: Particles, word order, and intonation*. Ph.D. thesis, Kyoto University.

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

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

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):175–200.

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

Reiko Vermeulen. 2012. The information structure of japanese. In Renate Musan Manfred Krifka, editor, *The expression of information structure*, pages 187–216. Berlin: De Gruyter Mouton.

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

Hiroko Yamashita and Tadahisa Kondo. 2011. Linguistic constraints and long-before-short tendency. In *IEICE Technocal report (TL)*, TL2011-19, pages 61–65.