Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese

Tatsuki Kuribayashi[1,3], **Takumi Ito**[1,3], **Jun Suzuki**[1,2], **Kentaro Inui**[1,2]
[1]Tohoku University [2]RIKEN [3]Langsmith Inc.
{kuribayashi, t-ito, jun.suzuki, inui} @ecei.tohoku.ac.jp

Abstract

We examine a methodology using neural language models (LMs) for analyzing the word order of language. This LM-based method has the potential to overcome the difficulties existing methods face, such as the propagation of preprocessor errors in count-based methods. In this study, we explore whether the LM-based method is valid for analyzing the word order. As a case study, this study focuses on Japanese due to its complex and flexible word order. To validate the LM-based method, we test (i) parallels between LMs and human word order preference, and (ii) consistency of the results obtained using the LM-based method with previous linguistic studies. Through our experiments, we tentatively conclude that LMs display sufficient word order knowledge for usage as an analysis tool. Finally, using the LM-based method, we demonstrate the relationship between the canonical word order and topicalization, which had yet to be analyzed by large-scale experiments.

1 Introduction

Speakers sometimes have a range of options for word order in conveying a similar meaning. A typical case in English is dative alternation:

(1) a. *A teacher gave a student a book.*
b. *A teacher gave a book to a student.*

Even for such a particular alternation, several studies (Bresnan et al., 2007; Hovav and Levin, 2008; Colleman, 2009) investigated the factors determining this word order and found that the choice is not random. For analyzing such linguistic phenomena, linguists repeat the cycle of constructing hypotheses and testing their validity, usually through psychological experiments or count-based methods. However, these approaches sometimes face difficulties, such as scalability issues in psychological experiments and the propagation of preprocessor errors in count-based methods.

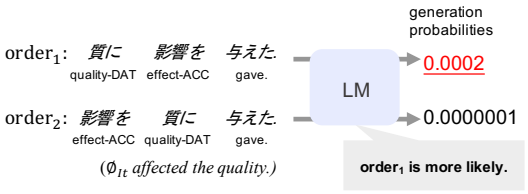

Figure 1: LM-based method for evaluating the canonicality of each word order considering their generation probabilities.

Compared to the typical approaches for evaluating linguistic hypotheses, approaches using LMs have potential advantages (Section 3.2). In this study, we examine the methodology of using LMs for analyzing word order (Figure 1). To validate the LM-based method, we first examine if there is a parallel between canonical word order and generation probability of LMs for each word order. Futrell and Levy (2019) reported that English LMs have human-like word order preferences, which can be one piece of evidence for validating the LM-based method. However, it is not clear whether the above assumption is valid even in languages with more flexible word order.

In this study, we specifically focus on the Japanese language due to its complex and flexible word order. There are many claims on the canonical word order of Japanese, and it has attracted considerable attention from linguists and natural language processing (NLP) researchers for decades (Hoji, 1985; Saeki, 1998; Miyamoto, 2002; Matsuoka, 2003; Koizumi and Tamaoka, 2004; Nakamoto et al., 2006; Shigenaga, 2014; Sasano and Okumura, 2016; Orita, 2017; Asahara et al., 2018).

We investigated the validity of using Japanese LMs for canonical word order analysis by conducting two sets of experiments: (i) comparing word order preference in LMs to that in Japanese speakers (Section 4), and (ii) checking the consistency

	Topic	Time	Location	Subject	(Adverb)	Indirect object	Direct object	Verb
Notation	TOP	TIM	LOC	NOM	-	DAT	ACC	-
Typical particle	“ ” (wa)	“ ” (ni)	“ ” (de)	“ ” (ga)	-	“ ” (ni)	“ ” (o)	-
Related section	6	5.2	5.2	5.2	5.3	5.1	5.1	5.1

Table 1: Overview of the typical cases in Japanese, their typical particles, and the sections where the corresponding case is analyzed. The well-known canonical word order of Japanese is listed from left to right.

between the preference of LMs with previous linguistic studies (Section 5). From our experiments, we tentatively conclude that LMs display sufficient word order knowledge for usage as an analysis tool, and further explore potential applications. Finally, we analyzed the relationship between topicalization and word order of Japanese by taking advantage of the LM-based method (Section 6).

In summary, we:

- Discuss and validate the use of LMs as a tool for word order analysis as well as investigate the sensitivity of LMs against different word orders in non-European language (Section 3);
- Find encouraging parallels between the results obtained with the LM-based method and those with the previously established method on various hypotheses of canonical word order of Japanese (Sections 4 and 5); and
- Showcase the advantages of an LM-based method through analyzing linguistic phenomena that is difficult to explore with the previous data-driven methods (Section 6).

2 Linguistic background

This section provides a brief overview of the linguistic background of canonical word order, some basics of Japanese grammar, and common methods of linguistic analysis.

2.1 On canonical word order

Every language is assumed to have a canonical word order, even those with flexible word order (Comrie, 1989). There has been a significant linguistic effort to reveal the factors determining the canonical word order (Bresnan et al., 2007; Hoji, 1985). The motivations for revealing the canonical word order range from linguistic interests to those involved in various other fields—it relates to language acquisition and production in psycholinguistics (Slobin and Bever, 1982; Akhtar, 1999), second language education (Alonso Belmonte et al., 2000), and natural language generation (Visweswariah et al., 2011) or error correction (Cheng et al., 2014) in NLP. In Japanese, there are also many studies on its canonical word order (Hoji, 1985; Saeki, 1998; Koizumi and Tamaoka, 2004; Sasano and Okumura, 2016).

Japanese canonical word order The word order of Japanese is basically subject-object-verb (SOV) order, but there is no strict rule except placing the verb at the end of the sentence (Tsujimura, 2013). For example, the following three sentences have the same denotational meaning (“A teacher gave a student a book.”):

(2) a. H L ’ k , ’ BR_ .
teacher-NOM student-DAT book-ACC gave.
b. H L , ’ ’ k BR_ .
teacher-NOM book-ACC student-DAT gave.
c. , ’ ’ k H L BR_ .
book-ACC student-DAT teacher-NOM gave.

This order-free nature suggests that the position of each constituent does not represent its semantic role (case). Instead, postpositional case particles indicate the roles. Table 1 shows typical constituents in a Japanese sentence, their postpositional particles, their canonical order, and the sections of this paper where each of them is analyzed. Note that postpositional case particles are sometimes omitted or replaced with other particles such as adverbial particles (Section 6). These characteristics complicate the factors determining word order, which renders the automatic analysis of Japanese word order difficult.

2.2 On typical methods for evaluating word order hypotheses and their difficulties

There are two main methods in linguistic research: human-based methods, which observe human reactions, and data-driven methods, which analyze text corpora.

Human-based methods A typical approach of testing word order hypotheses is observing the reaction (e.g., reading time) of humans to each word order (Shigenaga, 2014; Bahlmann et al., 2007). These approaches are based on the direct observation of humans, but this method has scalability

issues. There are also concerns that the participants may be biased, and that the experiments may not be replicable.

Data-driven methods Another typical approach is counting the occurrence frequencies of the targeted phenomena in a large corpus. This count-based method is based on the assumption that there are parallels between the canonical word order and the frequency of each word order in a large corpus. The parallel has been widely discussed (Arnon and Snider, 2010; Bresnan et al., 2007), and many studies rely on this assumption (Sasano and Okumura, 2016; Kempen and Harbusch, 2004). One of the advantages of this approach is suitability for large scale experiments. This enables considering a large number of examples.

In this method, researchers often have to identify the phenomena of interest with preprocessors (e.g., the predicate-argument structure parser used by Sasano and Okumura (2016)) in order to count them. However, sometimes, identi cation of the targeted phenomena is dif cult for the preprocessors, which limits the possibilities of analysis. For example, Sasano and Okumura (2016) focused only on simple examples where case markers appear explicitly, and only extract the head noun of the argument to avoid preprocessor errors. Thus, they could not analyze the phenomena in which the above conditions were not met. The above issue becomes more serious in low-resource languages, where the necessary preprocessors are often unavailable.

In this count-based direction, Bloem (2016) used n-gram LMs to test the claims on the German two-verb clusters. This method is closest to our proposed approach, but the general validity of using LMs is out of focus. This LM-based method also relies on the assumption of the parallels between the canonical word order and the frequency.

Another common data-driven approach is to train an interpretable model (e.g., Bayesian linear mixed models) to predict the targeted linguistic phenomena and analyze the inner workings of the model (e.g., slope parameters) (Bresnan et al., 2007; Asahara et al., 2018). Through this approach, researchers can obtain richer statistics, such as the strength of each factor's effect on the targeted phenomena, but creating labeled data and designing features for supervised learning can be costly.

3 LM-based method

3.1 Overview of the LM-based method

In the NLP eld, LMs are widely used to estimate the acceptability of text (Olteanu et al., 2006; Kann et al., 2018). An overview of the LM-based method is shown in Figure 1. After preparing several word orders considering the targeted linguistic hypothesis, we compare their generation probabilities in LMs. We assume that the word order with the highest generation probability follows their canonical word order.

3.2 Advantages of the LM-based method

In the count-based methods mentioned in Section 2.2, researchers often require preprocessors to identify the occurrence of the phenomena of interest in a large corpus. On the other hand, researchers need to prepare data to be scored by LMs to evaluate hypothesis in the LM-based method. Whether it is easier to prepare the preprocessor or the evaluation data depends on the situation. For example, the data preparation is easier in the situation where one wants to analyze the word order trends when a speci c postpositional particle is omitted. The question is whether Japanese speakers prefer the word order like in Example (3)-a or (3)-b.[1]

(3) a. ’ k , ’ BR_ .
student-DAT book(-ACC) gave.
b. , ’ ’ k BR_ .
book(-ACC) student-DAT gave.

While identifying the cases (ACC in Example (3)) without their postpositional particle is dif cult, creating the data without a speci c postpositional particle by modifying the existing data is easier such as creating Example (4)-b from Example (4)-a.

(4) a. ’ k , ’ BR_ .
student-DAT book-ACC gave.
b. ’ k , ’ BR_ .
student-DAT book(-ACC) gave.

Thus, in such situation, the LM-based method can be suitable.

The human-based method is more reliable given an example. However, it can be prohibitively costly. While the human-based method requires an evaluation data and human subjects, the LM-based method only requires the evaluation data. Thus, the LM-based method can be more suitable for estimating the validity of hypotheses and considering

[1] Omitted characters are crossed out. (e.g.,

many examples as exhaustively as possible. In addition, the LM-based method can be replicable. The suitable approach can be different in a situation, and broadening the choice of alternative methodologies may be beneficial to linguistic research.

Nowadays, various useful frameworks, language resources, and machine resources required to train LMs are available[2], which support the ease of implementing the LM-based method. Moreover, we make the LMs used in this study available[3].

3.3 Strategies to validate the use of LMs to analyze the word order

The goal of this study is to validate the use of LMs for analyzing the canonical word order. The canonical word order itself is still a subject of research, and the community does not know all about it. Thus, it is ultimately impossible to enumerate the requirements on what LMs should know about the canonical word order and probe the knowledge of LMs. Instead, we demonstrate the validity of the LM-based method by showcasing two types of parallels: (i) word order preference of LMs showing parallels with that of humans, and (ii) the results obtained with the LM-based method and those with previous methods being consistent on various claims on canonical word order. If the results of LMs are consistent with those of existing methods, the possibility that LMs and existing methods have the same ability to evaluate the hypotheses is supported. If the LM-based method is assumed to be valid, the method has the potential to streamline the research on unevaluated claims on word order. In the experiment sections, we examine the properties of Japanese LMs on (i) and (ii).

3.4 CAUTION – when using LMs for evaluating linguistic hypotheses

Even if LMs satisfy the criteria described in 3.3, there is no exact guarantee that LM scores will reflect the effectiveness of human processing of specific constructions in general. Thus, there seems to be a danger of confusing LM artifacts with language facts. Based on this, we hope that researchers use LMs as a tool just to limit the hypothesis space. LM supported hypotheses should then be re-verified with a human-based approach. Furthermore, since there is a lot of hypotheses and corresponding research, we cannot check all the properties of LMs in this study. This study focuses on intra-sentential factors of Japanese case order, and it is still unclear whether the LM-based method works properly in linguistic phenomena which are far from being the focus of this study. This is the first study where evidence is collected on the validity of using LMs for word order analysis and encourages further research on collecting such evidence and examining under what conditions this validity is guaranteed.

3.5 LMs settings

We used auto-regressive, unidirectional LMs with Transformer (Vaswani et al., 2017). We used two variants of LMs, a character-based LM (CLM) and a subword-based LM (SLM). In training SLM, the input sentences are once divided into morphemes by MeCab (Kudo, 2006) with a UniDic dictionary[4], and then these morphemes are split into subword units by byte-pair-encoding. (Sennrich et al., 2016)[5]. 160M sentences[6] randomly selected from 3B web pages were used to train the LMs. Hyperparameters are shown in Appendix A.

Given a sentence s, we calculate its generation probability $p(s) = p_l(s) \cdot p_r(s)$, where $p_l(\cdot)$ and $p_r(\cdot)$ are generation probabilities calculated by a left-to-right LM and a right-to-left LM, respectively. Depending on the hypothesis, we compare the generation probabilities of various variants of s with different word orders. We assume that the word order with the highest generation probability follows their canonical word order.

4 Experiment1: comparing human and LMs word order preference

To examine the validity of using LMs for canonical word order analysis, we examined the parallels between the LMs and humans on the task determining the canonicality of the word order (Figure 2). First, we created data for this task (Section 4.1). We then compared the word order preference of LMs and that of humans (Section 4.2).

[2] For example, one can train LMs with fairseq (Ott et al., 2019) and Wikipedia data on cloud computing platforms.

[3] https://github.com/kuribayashi4/LM_as_Word_Order_Evaluator .

[4] https://unidic.ninjal.ac.jp/

[5] Implemented in sentencepiece (Kudo and Richardson, 2018) We set character coverage to 0.9995 and vocab size to 100,000.

[6] 14GB in UTF-8 encoding. For reference, Japanese Wikipedia has around 2.5 GB of text. Because the focus of this study has context-independent nature, the sentences order is shuffled to prevent learning the inter-sentential characteristics of the language.

Figure 2: Overview of the experiment of comparing human and LMs word order preference. First, we created data for the task of comparing the appropriateness of the word order (left part), then we compare the preference of LMs and humans through this task (right part).

4.1 Human annotation

Data We randomly collected 10k sentences from 3B web pages, which are not overlapped with the LM training data. To remove overly complex sentences, we extracted sentences that must: (i) have less than or equal to five clauses and one verb, (ii) have clauses with a sibling relationship in its dependency tree, and they accompany a particle or adverb, (iii) not have special symbols such as parentheses, and (iv) not have a backward dependency path. For each sentence, we created its scrambled version.[7] The scrambling process is as follows:

1. Identify the dependency structure by using JUMAN[8] and KNP[9].
2. Randomly select a clause with several children.
3. Shuffle the position of its children along with their descendants.

Annotation We used the crowdsourcing platform Yahoo Japan!.[10] For our task, we showed crowdworkers a pair of sentences ($order_1$, $order_2$), where one sentence has the original word order and the other sentence has a scrambled word order.[11] Each annotator was instructed to label the pair with one of the following choices: (1) $order_1$ is better, (2) $order_2$ is better, or (3) the pair contains a semantically broken sentence. Only the sentences ($order_1$, $order_2$) were shown to the annotators, and they were instructed not to imagine a specific context for the sentences. We filtered unmotivated workers by using check questions.[12] For each pair instance, we employed 10 crowdworkers. In total, 756 unique, motivated crowdworkers participated in our task.

From the annotated data, we collected only the pairs satisfying the following conditions for our experiments: (i) none of 10 annotators determined that the pair contains a semantically broken sentence, and (ii) nine or more annotators preferred the same order. The majority decision is labeled in each pair; the task is binary classification. We assume that if many workers prefer a certain word order, then it follows its canonical word order, and the other one deviates from it. We collected 2.6k pair instances of sentences.

4.2 Results

We compared the word order preference of LMs and that of the workers by using the 2.6K pairs created in Section 4.1. We calculated the correlation of the decisions between the LMs and the workers; which word order is more appropriate $order_1$ or $order_2$. The word orders supported by CLM and SLM are highly correlated with workers, with the Pearson correlation coefficient of 0.89 and 0.90, respectively. This supports the assumption that the generation probability of LMs can determine the canonical word order as accurately as humans do. Note that such a direct comparison of word order is difficult with the count-based methods because of the sparsity of the corpus.

5 Experiment2: consistency with previous studies

This section examines whether LMs show word order preference consistent with previous linguistic studies. The results are entirely consistent, which support the validity of the LM-based methods in Japanese. Each subsection focuses on a specific component of Japanese sentences.

5.1 Double objects

The order of double objects is one of the most controversial topics in Japanese word order. Examples of the possible order are as follows:

(5) DAT-ACC: 先生 が 学生 に 本 を 渡した。
teacher NOM student DAT book ACC gave.

ACC-DAT: 先生 が 本 を 学生 に 渡した。
teacher NOM book ACC student DAT gave.

Henceforth, DAT-ACC/ACC-DAT denotes the word order in which the DAT/ACC argument precedes the ACC/DAT argument. We evaluate the

[7] When several scrambled versions were possible for a given sentence, we randomly selected one of them.

[8] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[9] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

[10] https://crowdsourcing.yahoo.co.jp/

[11] Crowdworkers did not know which sentence was the original sentence.

[12] We manually created check questions considering the Japanese speakers' preference in trial experiments in advance.

(a) Each verb's ACC-DAT rate.

(b) Relationship between each verb's $R^v_{\text{DAT-only}}$ and the ACC-DAT rate.

(c) Relationship between the degree of co-occurrence of verb and arguments, and the ACC-DAT rate in each example. For the results of LMs, the ACC-DAT rate of each example is regarded as 1 if LMs prefer ACC-DAT order, otherwise we regard the example as 0.

Figure 3: Overlap of the results of Sasano and Okumura (2016) and that of LMs. In figures (a) and (b), each plot corresponds to each verb. In figure (c), each plot corresponds to each example. The legend of figure (a) and (b) is the same as in figure (c). "S&O 2016" refers to Sasano and Okumura (2016).

claims Sasano and Okumura (2016) focused on with the data they collected.[13]

Word order for each verb First, we analyzed the trend of the double object order for each verb. We analyzed 620 verbs following Sasano and Okumura (2016).[14] For each set of examples S^v corresponding to a verb v, we: (i) created an instance with the swapped order of ACC and DAT for each example, and (ii) compared the generation probabilities of the original and swapped instances. A^v is the set of examples preferred by LMs. $R^v_{\text{ACC-DAT}}$ is calculated as follows:

$$R^v_{\text{ACC-DAT}} = \frac{N^v_{\text{ACC-DAT}}}{N^v_{\text{ACC-DAT}} + N^v_{\text{DAT-ACC}}} ;$$

where $N^v_{\text{ACC-DAT}} = N^v_{\text{DAT-ACC}}$ is the number of examples with the ACC-DAT / DAT-ACC order in A^v.

Figure 3-(a) shows the relationship between $R^v_{\text{ACC-DAT}}$ determined by LMs and one reported in a previous count-based study (Sasano and Okumura, 2016). These results strongly correlate with the Pearson correlation coefficient of 0.91 and 0.88, in CLM and SLM, respectively. In addition, "canonical word order is DAT-ACC" (Hoji, 1985) is unlikely to be valid because there are verbs where $R^v_{\text{ACC-DAT}}$ is very high (details in Appendix B.1). This conclusion is consistent with Sasano and Okumura (2016).

Word order and verb types In Japanese, there are show-type and pass-type verbs (details in Appendix B.2). Matsuoka (2003) claimed that the order of double objects differs depending on these verb types. Following Sasano and Okumura (2016), we analyzed this trends.

We applied the Wilcoxon rank-sum test between the distributions of $R^v_{\text{ACC-DAT}}$ determined by LMs in the two groups (show-type and pass-type verbs). The results show no significant difference between the two groups (p-value is 0.17 and 0.12 in the experiments using CLM and SLM, respectively). These results are consistent with the count-based (Sasano and Okumura, 2016) and the human-based (Miyamoto, 2002; Koizumi and Tamaoka, 2004) methods.

Word order and argument omission Sasano and Okumura (2016) claimed that the frequently omitted case is placed near the verb. First, we calculated $R^v_{\text{DAT-only}}$ for each verb v as follows:

$$R^v_{\text{DAT-only}} = \frac{N^v_{\text{DAT-only}}}{N^v_{\text{DAT-only}} + N^v_{\text{ACC-only}}} ;$$

where $N^v_{\text{DAT-only}} = N^v_{\text{ACC-only}}$ denotes the number of examples in which the DAT / ACC case appears, and the other case does not in S^v. A large $R^v_{\text{DAT-only}}$ score indicates that the DAT argument is less frequently omitted than the ACC argument in S^v. We analyzed the relationship between $R^v_{\text{DAT-only}}$ and $R^v_{\text{ACC-DAT}}$ for each verb.

Figure 3-(b) shows that the regression lines from the LM-based method and Sasano and Okumura (2016) corroborate similar trends. The Pearson correlation coefficient between $R^v_{\text{DAT-only}}$ and $R^v_{\text{ACC-DAT}}$ is 0.404 for CLM and 0.374 for SLM. The results are consistent with Sasano and Okumura (2016), where they reported that the correlation coefficient was 0.391.

Word order and semantic role of the dative argument Matsuoka (2003) claimed that the canonical word order differs depending on the semantic role of the dative argument. Sasano and Okumura

[13] We filtered the examples overlapping with the training data of LMs in advance. As a result, we collected 4.5M examples.

[14] We removed verbs for which all examples overlap with the data for training the LMs.

	TIM < LOC	TIM < NOM	LOC < NOM
CLM	.757	.642	.604
SLM	.708	.632	.615
Count	.686	.666	.681

Table 2: The column $a < b$ shows the score $o(a < b)$, which indicates the rate of case a being more likely to be placed before b. The row "Count" shows the count-based results in the dataset we used.

(2016) evaluated this claim by analyzing the trend in the following two types of examples:

(6) Type-A: , ' f ! k Ô W_
book-ACC school-DAT returned.

Type-B: H k , ' Ô W_
teacher-DAT book-ACC returned.

Type-A has an inanimate goal noun (school) as the DAT argument, while Type-B has an animate processor (teacher). It was reported that Type-A is likely to be the ACC-DAT order, while Type-B is likely to be the DAT-ACC order. Following Sasano and Okumura (2016), we analyzed 113 verbs.[15] For each verb, we compared the ACC-DAT rate in its type-A examples and the rate in its type-B examples.

The number of verbs where the ACC-DAT order is preferred in Type-A examples to Type-B examples is significantly larger (a two-sided sign test $p < 0.05$). This result is consistent with that of Sasano and Okumura (2016); Matsuoka (2003) and implies that the LMs capture the animacy of the nouns. Details are in Appendix B.3.

Word order and co-occurrence of verb and arguments Sasano and Okumura (2016) claimed that an argument that frequently co-occurs with the verb tends to be placed near the verb. For each example, the LMs determine which word order (DAT-ACC or ACC-DAT) is appropriate. Each example also has a score NPMI (definition in Appendix B.4). Higher NPMI means that the DAT noun in the example more strongly co-occurs with the verb in the example than the ACC noun.

Figure 3-(c) shows the relationship between NPMI and the ACC-DAT rate in each example. NPMI and the ACC-DAT rate are correlated with the Pearson correlation coefficient of 0.517 and 0.521 in CLM and SLM, respectively. These results are consistent with Sasano and Okumura (2016).

[15] Among the 126 verbs used in Sasano and Okumura (2016), 113 verbs with data that do not overlap with the LM training data were selected.

Model	MODAL	TIME	MANNER	RESULTIVE
CLM	1.	1	0.5	1.
SLM	1.	0.5	1.	0.5

Table 3: The scores denote the rank correlation between the preference of each adverb position in LMs and that reported in (Koizumi and Tamaoka, 2006).

5.2 Order of constituents representing time, location, and subject information

Our focus moves to the cases closer to the beginning of the sentences. The following claim is a well-known property of Japanese word order: "The case representing time information (TIM) is placed before the case representing location information (LOC), and the TIM and LOC cases are placed before the NOM case" (Saeki, 1960, 1998). We examined a parallel between the result obtained with the LM-based and count-based methods on this claim.

We randomly collected 81k examples from 3B web pages.[16] To create the examples, we identified the case components by KNP, and the TIM and LOC cases were categorized with JUMAN (details in Appendix C). For each example, we created all possible word orders and obtained the word order with the highest generation probability $P(\cdot)$. Given S a set of s, we calculated a score $o(a < b)$ for cases a and b as follows:

$$o(a < b) = \frac{N_{a<b}}{N_{a<b} + N_{b<a}};$$

where $N_{k<l}$ is the number of examples where the case k precedes the case l in S. Higher $o(a < b)$ indicates that the case a is more likely to be placed before the case b. The results with the LM-based methods and the count-based method are consistent (Table 2). Both results show that $o(\text{TIM} < \text{LOC})$ is significantly larger than $o(\text{TIM} > \text{LOC})$ ($p < 0.05$ with a two-sided signed test), which indicates that the TIM case usually precedes the LOC case. Similarly, the results indicate that the TIM case and the LOC case precedes the NOM case.

5.3 Adverb position

We checked the preference of the adverb position in LMs. The position of the adverb has no restriction except that it must be before the verb, which is similar to the trend of the case position. However, Koizumi and Tamaoka (2006) claimed that "There is a canonical position of an adverb depend-

[16] Without overlap with the training data of LMs.

Model	long precedes short	short precedes long
CLM	5,640	3,754
SLM	5,757	3,914

Table 4: Changes in the position of a constituent with the largest number of chunks.

ing on its type." They focus on four types of adverbs: MODAL, TIME, MANNER, and RESULTIVE.

We used the same examples as Koizumi and Tamaoka (2006). For each example, we created its three variants with a different adverb position as follows ("A friend handled the tools roughly"):

(10) ASOV: q ´ k Ë T L S w ' q c_
roughly friend-NOM tools-ACC handled.

SAOV: Ë T L q ´ k S w ' q c_
friend-NOM roughly tools-ACC handled.

SOAV: Ë T L S w ' q ´ k q c_
friend-NOM tools-ACC roughly handled.

where the sequence of the alphabet such as "ASOV" denote the word order of its corresponding sentences. For example, "ASOV" indicates the order adverb< subject< object< verb. "A," "S," "O," and "V" denote "adverb," "subject," "object," and "verb," respectively.

Then, we obtained the preferred adverb position by comparing their generation probabilities. Finally, for each adverb type and its examples, we ranked the preference of the possible adverb positions: "ASOV," "SAOV," and "SOAV." Table 3 shows the rank correlation of the preference of the position of each adverb type. The results show similar trends of LMs with that of the human-based method (Koizumi and Tamaoka, 2006).

5.4 Long-before-short effect

The effects of "long-before-short," the trend that a long constituent precedes a short one, has been reported in several studies (Asahara et al., 2018; Orita, 2017) We checked whether this effect can be captured with the LM-based method. Among the examples used in Section 5.2, we analyzed about 9.5k examples in which the position of the constituent with the largest number of chunks[17] differed between its canonical case order[18] and the order supported by LMs.

Table 4 shows that there are signi cantly ($p < 0.05$ with a two-sided signed test) large numbers of examples where the longest constituent moves closer to the beginning of the sentence. This result is consistent with existing studies and supports the tendency for longer constituents to appear before shorter ones.

5.5 Summary of the results

We found parallels between the results with the LM-based method and that with the previously established method on various properties of canonical word order. These results support the use of LMs for analyzing Japanese canonical word order.

6 Analysis: word order and topicalization

In the previous section, we tentatively concluded that LMs can be used for analyzing the intra-sentential properties on the canonical word order. Based on this nding, in this section, we demonstrate the analysis of additional claims on the properties of the canonical word order with the LM-based method, which has been less explored by large-scale experiments. This section shows the analysis of the relationship between topicalization and the canonical word order. Additional analyses on the effect of various adverbial particles for the word order are shown in Appendix F.

6.1 Topicalization in Japanese

The adverbial particle "o " (TOP) is usually used as a postpositional particle when a speci c constituent represents the topic of the sentence (Heycock, 1993; Noda, 1996; Fry, 2003). When a case component is topicalized, the constituent moves to the beginning of the sentence, and the particle "o " (TOP) is added (Noda, 1996). Additionally, the original case particle is sometimes omitted,[19] which makes the case of the constituent dif cult to identify. For example, to topicalize " ' " (book-ACC) in Example (8)-a, the constituent moves to the beginning of the sentence, and the original accusative case particle " " (ACC) is omitted. Similarly, "H L " (teacher-NOM) is topicalized in Example (8)-b. The original sentence is enclosed in the square brackets in Example (8).

(8) a. , 'o [H L , ' BR_ .]
book-TOP teacher-NOM book-ACC gave.

b. H Lo [H L , ' BR_ .]
teacher-TOP teacher-NOM book-ACC gave.

[17] chunks were identi ed by KNP.

[18] In this section, canonical case order is assumed to be TOM< LOC< NOM< DAT< ACC

[19] The particles " " (ACC) and "L " (NOM) are omitted.

With the above process, we can easily create a sentence with a topicalized constituent. On the other hand, identifying the original case of the topicalized case components is error-prone. Thus, the LM-based method can be suitable for empirically evaluating the claims related to the topicalization.

6.2 Experiments and results

By using the LM-based method, we evaluate the following two claims:

(i) The more anterior the case is in the canonical word order, the more likely its component is topicalized (Noda, 1996).

(ii) The more the verb prefers the ACC-DAT order, the more likely the ACC case is topicalized than the DAT case.

The claim (i) suggests that, for example, the NOM case is more likely to be topicalized than the ACC case because the NOM case is before the ACC case in the canonical word order of Japanese. The claim (ii) is based on our observation. It can be regarded as an extension of the claim (i) considering the effect of the verb on its argument order. We assume that the canonical word order of Japanese is TIM < LOC < NOM < DAT < ACC in this section.

Claim (i) We examine which case is more likely to be topicalized. We collected 81k examples from Japanese Wikipedia (Details are in Appendix C). For each example, a set of candidates was created by topicalizing each case, as shown in Example (8). Then, we selected the sentences with the highest score by LMs in each candidate set. We denote the obtained sentences as S^{topic}. We calculated a score $t_{a>b}$ for pairs of cases a and b.

$$t_{a>b} = \frac{N_{a>b}}{N_{a>b} + N_{b>a}}$$

where $N_{a>b}$ is the examples where the case a and b appear, and case a is a topic of the sentence in S^{topic}. The higher the score is, the more the case a is likely to be topicalized than the case b is.

We compared $t_{a>b}$ and $t_{b>a}$ among the pairs of cases a and b, where the case a precedes the case b in the canonical word order. Through our experiments, $t_{a>b}$ was significantly larger than $t_{b>a}$ ($p < 0.05$ with a paired t-test) in CLM and SLM results, which supports the claim (i) (Noda, 1996). Detailed results are shown in Appendix E.

Claim (ii) The canonical word order of double objects is different for each verb (Section 5.1). Based on this assumption and the claim (i), we hypothesized that the more the verb prefers the ACC-DAT order, the more likely the ACC case of the verb is topicalized than the DAT case.

We used the same data as in Section 5.1. For each example, we created two sentences by topicalizing the ACC or DAT argument. Then we compared their generation probabilities. In each set of examples corresponding to a verb v, we calculated the rate that the sentence with the topicalized ACC argument is preferred rather than that with the topicalized DAT argument. This rate and $R^v_{ACC\text{-}DAT}$ is significantly correlated with the Pearson correlation coefficient of 0.89 and 0.84 in CLM and SLM, respectively. This results support the claim (ii). Detailed results are shown in Appendix E.

7 Conclusion and Future work

We have proposed to use LMs as a tool for analyzing word order in Japanese. Our experimental results support the validity of using Japanese LMs for canonical word order analysis, which has the potential to broaden the possibilities of linguistic research. From an engineering view, this study supports the use of LMs for scoring Japanese word order automatically. From the viewpoint of the linguistic field, we provide additional empirical evidence to various word order hypotheses as well as demonstrate the validity of the LM-based method.

We plan to further explore the capability of LMs on other linguistic phenomena related to word order, such as “given new ordering” (Nakagawa, 2016; Asahara et al., 2018). Since LMs are language-agnostic, analyzing word order in another language with the LM-based method would also be an interesting direction to investigate. Furthermore, we would like to extend a comparison between machine and human language processing beyond the perspective of word order.

8 Acknowledgments

We would like to offer our gratitude to Kaori Uchiyama for taking the time to discuss our paper and Ana Brassard for her sharp feedback on English. We also would like to show our appreciation to the Tohoku NLP lab members for their valuable advice. We are particularly grateful to Ryohei Sasano for sharing the data for double objects order analyses. This work was supported by JST CREST Grant Number JPMJCR1513, JSPS KAKENHI Grant Number JP19H04162, and Grant-in-Aid for JSPS Fellows Grant Number JP20J22697.

References

Nameera Akhtar. 1999. Acquiring basic word order: Evidence for data-driven learning of syntactic structure. Journal of child language, 26(2):339–356.

Isabel Alonso Belmonte et al. 2000. Teaching English Word Order to ESL Spanish Students. A Functional Perspective.

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1):67–82.

Masayuki Asahara, Satoshi Nambu, and Shin-Ichiro Sano. 2018. Predicting Japanese Word Order in Double Object Constructions. In Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing, pages 36–40, Melbourne. Association for Computational Linguistics.

Jörg Bahlmann, Antoni Rodriguez-Fornells, Michael Rotte, and Thomas F Münte. 2007. An fMRI study of canonical and noncanonical word order in German. Human brain mapping, 28(10):940–949.

Jelke Bloem. 2016. Testing the Processing Hypothesis of word order variation using a probabilistic language model. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), pages 174–185, Osaka, Japan. The COLING 2016 Organizing Committee.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R Harald Baayen. 2007. Predicting the dative alternation. In Cognitive foundations of interpretation, pages 69–94. KNAW.

Shuk-Man Cheng, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 279–289, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Timothy Colleman. 2009. Verb disposition in argument structure alternations: a corpus study of the dative alternation in Dutch. Language Sciences, 31(5):593–611.

Bernard Comrie. 1989. Language universals and linguistic typology: Syntax and morphology. University of Chicago press.

John Fry. 2003. Ellipsis and wa-marking in Japanese conversation. Taylor & Francis.

Richard Futrell and Roger P Levy. 2019. Do RNNs learn human-like abstract word order preferences? In Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 50–59.

Édouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. 2017. Efficient softmax approximation for GPUs. In Proceedings of the 34th International Conference on Machine Learning volume 70 of Proceedings of Machine Learning Research, pages 1302–1310, International Convention Centre, Sydney, Australia. PMLR.

Caroline Heycock. 1993. Syntactic predication in Japanese. Journal of East Asian Linguistics, 2(2):167–211.

Hajime Hoji. 1985. Logical form constraints and configurational structures in Japanese. PHD Thesis. University of Washington.

Malka Rappaport Hovav and Beth Levin. 2008. The English dative alternation: The case for verb sensitivity. Journal of linguistics, 44(1):129–167.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-Level Fluency Evaluation: References Help, But Can Be Spared! In Proceedings of the 22nd Conference on Computational Natural Language Learning, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Gerard Kempen and Karin Harbusch. 2004. A corpus study into word order variation in German subordinate clauses: Animacy affects. Multidisciplinary approaches to language production, pages 173–181.

Masatoshi Koizumi and Katsuo Tamaoka. 2004. Cognitive processing of Japanese sentences with ditransitive verbs. Gengo Kenkyu (Journal of the Linguistic Society of Japan), 2004(125):173–190.

Masatoshi Koizumi and Katsuo Tamaoka. 2006. The Canonical Positions of Adjuncts in the Processing of Japanese Sentence. Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 13(3):392–403.

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.jp.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Mikinari Matsuoka. 2003. Two Types of Ditransitive Consturctions in Japanese. Journal of East Asian Linguistics, 12(2):171–203.

Edson T Miyamoto. 2002. Sources of difficulty in the processing of scrambling in Japanese. Sentence processing in East Asian languages, pages 167–188.

Natsuko Nakagawa. 2016. Information structure in spoken japanese: Particles, word order, and intonation.

Keiko Nakamoto, Jae-ho Lee, and Kow Kuroda. 2006. Preferred Word Orders Correlate with Sentential Meanings That Cannot Be Reduced to Verb Meanings: A New Perspective on Construction Effects in Japanese. Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 13(3):334–352.

Hisashi Noda. 1996. Wa to ga [Wa and ga]. Kurosio Publishers.

Marian Olteanu, Pasin Suriyentrakorn, and Dan Moldovan. 2006. Language models and reranking for machine translation. In Proceedings of the Workshop on Statistical Machine Translation, pages 150–153, New York City. Association for Computational Linguistics.

Naho Orita. 2017. Predicting japanese scrambling in the wild. In Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics, pages 41–45.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Tetsuo Saeki. 1960. Gendaigo ni okeru gojun no kenkyū – iwayuru hogo no baai [The trend of word order in modern writing– in so-called complements]. Gengo seikatsu [Language life], (111):56–63.

Tetsuo Saeki. 1998. Yōsetsu Nihongo no Gojun [Essentials of Japanese word order]. Kurosio Publishers.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yasumasa Shigenaga. 2014. Canonical Word Order of Japanese Ditransitive Sentences: A Preliminary Investigation through a Grammaticality Judgment Survey. Advances in Language and Literary Studies, 5(2):35–45.

Dan I Slobin and Thomas G Bever. 1982. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. Cognition, 12(3):229–265.

Natsuko Tsujimura. 2013. An introduction to Japanese linguistics. John Wiley & Sons.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. 2011. A Word Reordering Model for Improved Machine Translation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 486–496, Edinburgh, Scotland, UK. Association for Computational Linguistics.

A Hyperparameters and implementation of the LMs

We used the Transformer (Vaswani et al., 2017) LMs implemented in fairseq (Ott et al., 2019). Table 5 shows the hyperparameters of the LMs. The adaptive softmax cutoff (Grave et al., 2017) is only applied to SLM. We split 10K sentences for dev set. The left-to-right and right-to-left CLMs achieved a perplexity of 11.05 and 11.08, respectively. The left-to-right and right-to-left SLMs achieved a perplexity of 28.51 and 28.25, respectively. Note that the difference in the perplexities between CLM and SLM is due to the difference in the vocabulary size.

B Details on Section 5.1 (double objects)

B.1 Word order for each verb

It is considered that different verbs have different preferences in the order of their object. For example, while the verb "compare" prefers the ACC-DAT order (Example (9)-a), the verb "express" prefers the DAT-ACC order (Example (9)-b).

(9) a. person-ACC color-DAT compared.
(compared a person to color.)

b. shopkeeper-DAT respect-ACC expressed.
(expressed a respect to a shopkeeper.)

Table 6 shows the verbs with the top five and the five worst $R^{v}_{\text{ACC-DAT}}$.

B.2 Word order and verb types

There are two types of causative-inchoative alternating verbs in Japanese: show-type verbs and pass-type verbs. The verb types are determined by the subject of the sentence where the corresponding inchoative verb is used. For the show-type verbs, the DAT argument of a causative sentence becomes the subject in its corresponding inchoative sentence (Example (10)). On the other hand, the ACC argument of a causative sentence becomes the subject in its corresponding inchoative sentence for the pass-type verbs (Example (11)).

(10) Causative: student-DAT book-ACC showed.
(showed a student a book.)

Inchoative: student-NOM saw.
(A student saw $_{\text{something}}$)

(11) Causative: student-DAT book-ACC showed.
(passed a student a book.)

Inchoative: book-NOM passed.
(A book passed to $_{\text{something}}$)

Matsuoka (2003) claims that the show-type verb prefers the DAT-ACC order, while the pass-type verb prefers the ACC-DAT order.

Table 7 shows $R^{v}_{\text{ACC-DAT}}$ of the show-type and pass-type verbs. The results show no significant difference in word order trends between show-type and pass-type verbs, which are consistent with that of Sasano and Okumura (2016).

B.3 Word order and semantic role of the dative argument

As described in Section 5.1, Sasano and Okumura (2016) reported that type-A examples prefer the ACC-DAT order and type-B examples prefer the DAT-ACC order. We used the same examples as Sasano and Okumura (2016) used. We analyzed the difference in the trend of argument order between type-A and type-B examples in each verb. Table 8 shows the verbs, which show a significant change in the argument order between type-A and type-B examples ($p < 0.05$ in a two-proportion z-test). In the experiment using CLM, 31 verbs show the trend that type-A examples more prefer the ACC-DAT order to type-B, and 17 verbs show contrary trends. In the experiment using SLM, 38 verbs show the trend that type-A examples more prefer the ACC-DAT order to type-B, and 11 verbs show contrary trends. These results show that the number of verbs, where the ACC-DAT order is preferred by type-A examples rather than type-B, is significantly larger ($p < 0.05$ with a two-sided sign test). This experimental design follows Sasano and Okumura (2016).

B.4 Word order and co-occurrence of verb and arguments

We evaluate the claim that an argument frequently co-occurring with the verb tends to be placed near the verb. We examine the relationship between each example's word order trend and NPMI.

NPMI is calculated as follows:

Fairseq model	architecture	transformer_lm
	adaptive softmax cut off	50,000, 140,000
Optimizer	algorithm	Nesterov accelerated gradient (nag)
	learning rates	1e-5
	momentum	0.99
	weight decay	0
	clip norm	0.1
Learning rate scheduler	type	cosine
	warmup updates	16,000
	warmup init lrarning rate	1e-7
	max learning rate	0.1
	min learning rate	1e-9
	t mult (factor to grow the length of each period)	2
	learning rate period updates	270,000
	learning rate shrink	0.75
Training	batch size	4608 tokens
	epochs	3

Table 5: Hyperparameters of the LMs.

	ACC-DATis preferred			DAT-ACCis preferred		
Model	Verb	$R^v_{\text{ACC-DAT}}$	S&O	Verb	$R^v_{\text{ACC-DAT}}$	S&O
CLM	“‹ H‹ ” (compare)	0.993	0.945	ħ Y‹ ” (to table)	0.001	0.013
	“Û —Y‹ ” (converted)	0.992	0.935	„ ~Y ” (put on air)	0.000	0.017
	“¼Wú Y ” (extruded)	0.979	0.923	ħ „Y ” (cook inside)	0.000	0.019
	“‹ Ë f‹ ” (mitateru)	0.994	0.919	“ ‹ ” (close the eyes)	0.001	0.021
	“ Û ” (conversion)	0.975	0.898	æ•‹ ” (shrug)	0.002	0.022
SLM	“‹ H‹ ” (compare)	0.993	0.926	‰ Y‹ ” (kissuru)	0.003	0.018
	“¼Wú Y ” (extruded)	0.979	0.914	ħ Y‹ ” (to table)	0.001	0.018
	“ã • ” (con nement)	0.885	0.912	„ ~Y ” (put on air)	0.000	0.021
	“y Ë f‹ ” (help)	0.933	0.904	œKY ” (leave out)	0.002	0.022
	“0 Y ” (attributable)	0.838	0.903	“ • e Œ‹ ” (step into)	0.002	0.025

Table 6: The verbs with the top ve and the worst ve $R^v_{\text{ACC-DAT}}$ in each LM. The “S&O” columns show the ACC-DATrate reported in Sasano and Okumura (2016).

$$\text{NPMI} = \text{NPMI}(n_{\text{DAT}}; v) \quad \text{NPMI}(n_{\text{ACC}}; v) \;;$$

$$\text{where NPMI}(n_c; v) = \frac{\text{PMI}(n_c; v)}{\log(p(n_c; v))} \;;$$

$$\text{PMI}(n_c; v) = \log\frac{p(n_c; v)}{p(n_c)p(v)} \;;$$

where,v is a verb andn$_c$ (c 2 DAT, ACC) is its argument.

C Data used in Section 5.2, Section 6, and Appendix F

First, we randomly collected 50M sentences from 3B web pages. Note that there is no overlap between the collected sentences and the training data of LMs. Next, we obtained the sentences that satisfy the following criteria:

There is a verb (placed at the end of the sentence) with more than two arguments (accompanying the case particlega, o, ni, or de), where dependency distance between the verb and arguments is one.
Each argument (with its descendant) has fewer than 11 morphemes in the argument.

In each example, the verb (satisfying the above condition), its arguments, and the descendants of the arguments are extracted. Example sentences are created by concatenating the verb, its argument, and the descendants of the arguments with preserving their order in the original sentences.

In the experiments in Section 5.2, we analyzed the word order trend of theTIM and LOC constituents. We regard the constituent (argument and its descendants) satisfying the following condition as theTIM constituent:

Accompanying the postpositional case particle “k ” (DAT).

Show-type				Pass-type							
Verb	CLM	SLM	S&O	Verb	CLM	SLM	S&O	Verb	CLM	SLM	S&O
“å ‰[‹ ” (notify)	.718	.754	.522	‚“ Y ” (put back)	.366	.395	.771	“ ‰Y ” (leak)	.152	.207	.332
“ Q‹ ” (deposit)	.426	.391	.399	b •‹ ” (lodge)	.638	.704	.748	ñ Ky‹ ” (oat)	.387	.406	.255
“‹ [‹ ” (show)	.353	.429	.301	“ € ” (wrap)	.316	.356	.603	“ Q‹ ” (direct)	.291	.319	.251
“« [‹ ” (cover)	.240	.224	.256	“ H‹ ” (inform)	.419	.460	.522	‹“ Y ” (leave)	.323	.318	.238
“Y H‹ ” (teach)	.297	.293	.235	₩[‹ ” (place or)	.556	.498	.496	Ë •‹ ” (bury)	.405	.430	.223
“^ Q‹ ” (give)	.101	.084	.186	J Q‹ ” (deliver)	.364	.419	.491	÷ \‹ ” (blend)	.336	.276	.200
“t s[‹ ” (shower)	.113	.121	.177	& y‹ ” (range)	.423	.485	.481	S f‹ ” (hit)	.287	.320	.185
“¸ Y ” (lend)	.253	.213	.118	vdQ‹ ” (knock)	.333	.344	.436	,“ Q‹ ” (hang)	.285	.288	.108
“¸@[‹ ” (dress)	.115	.109	.113	Ø Q‹ ” (attach)	.326	.329	.368	f m‹ ” (pile)	.226	.263	.084
-	-	-	-	“! Y ” (pass)	.349	.336	.362	û f‹ ” (build)	.117	.099	.069
-	-	-	-	“= hY ” (drop)	.379	.397	.351	-	-	-	-
Macro Avg.	.291	.291	.305	Macro Avg.					.347	.364	.361

Table 7: Overlap of the results of LMs and that of Sasano and Okumura (2016) on the relationship between the ACC-DAT rate and verb types. Each score corresponding to a verb denotes its ACC-DAT rate. The “S&O” columns show the ACC-DAT rate reported in Sasano and Okumura (2016). There is no significant difference between the distributions of the ACC-DAT rate in two verb types.

Containing time category morphemes[20]

We regard the constituent (argument and its descendants) satisfying the following condition as the LOC constituent:

Accompanying the postpositional case particle “g ”.
Containing location category morphemes[20]

81k examples were created. The averaged number of characters in a sentence was 45.1 characters. The number of occurrences of each case is shown in Table 9. The scrambling process conducted in the experiments (Sections 5.2 and 6) is the same as described in Section 4.

D Details on Section 5.3 (adverb)

Table 10 shows the correlation between the result of LMs and that of Koizumi and Tamaoka (2006). The column “Canonical” shows the position, which is significantly preferred over the other positions. “A,” “S,” “O,” and “V” denote “adverb,” “subject,” “object,” and “verb,” respectively. The sequence of the alphabets corresponds to their order; for example, “ASOV” indicates the order: adverb < subject < object < verb. Following Koizumi and Tamaoka (2006), we examined the three candidate positions of the adverb: “ASOV,” “SAOV,” and “SOAV.” The score r denotes the Pearson correlation coefficient of the preferred ranks of each adverb position to that reported in Koizumi and Tamaoka (2006).

E Details on Section 6.2 (topicalization)

We topicalized a specific constituent by moving the constituent to the beginning of the sentence and adding the adverbial particle “” (TOP). Strictly speaking, conjunctions are preferentially placed at the beginning of the sentence rather than topicalized constituents. The examples we used do not include the conjunctions at the beginning of the sentence. The adverbial particle was added according to the rules shown in Table 12.

Figure 4: Correlation between the ACC-DAT rate and the rate that the ACC argument is more likely to be topicalized than DAT for each verb. Each plot corresponds to the result of each verb.

Claim (i): Table 11 shows the t_{ajb} for each pair of the case a (row) and b (column). The results show that the more anterior the case a is and the more posterior the case b is in the canonical word order, the larger the t_{ajb} is.

Claim (ii): Figure 4 shows that the more a verb prefers the ACC-DAT order, the more ACC case tends to be topicalized. The X-axis denotes the ACC-DAT rate of the verb, and the Y-axis denotes the trend that ACC is more likely to be topicalized than DAT.

[20] identified by JUMAN

Model	Verbs whose type-A examples prefer the DAT-ACC order	Verbs whose type-B examples prefer the DAT-ACC order
CLM	“ Q‹ ” (deposit), “n O ” (place), “ d ” (have), “e Œ‹ ” (put in), “ •‹ ” (pay), “õ ” (mail), “› f ” (supply), “ú Y ” (put out), “K v ” (transport), “A Y ” (shed), “› Q‹ ” (hang), “þ ‹ ” (decorate), “ƒ R‹ ” (spread), “û Y ” (transfer), “‹ Y ” (leave), “M ” (deliver), “ ‹ ” (send), “• R‹ ” (throw), “ Ø ” (send), “Ô t ” (return), “J Q‹ ” (send), “; Y ” (return), “@Q‹ ” (wear), “ R‹ ” (increase), “= h Y ” (drop), “ [‹ ” (publish), “ ô ” (change), “ e ” (deliver), “x Y ” (unload), “² ” (publish), “ Y ” (get X through)	“M ” (distribute), “! Y ” (pass), “×ì¼óÈ ” (present), “ •[‹ ” (match), “‹ [‹ ” (show), “Ð › ” (offer), “ H‹ ” (give), “S f‹ ” (hit), “Þ Y ” (turn), “ý ” (add), “¸ Y ” (lend), “U : ” (exhibit), “n H‹ ” (lay), “• < ” (request), “? e ” (insert), “• •‹ ” (collect), “Ë B ” (claim)
SLM	“ Q‹ ” (deposit), “n O ” (place), “< € ” (ask), “e Œ‹ ” (put in), “ •‹ ” (pay), “õ ” (mail), “ú Y ” (put out), “K v ” (transport), “A Y ” (shed), “› Q ‹ ” (hang), “ƒ R‹ ” (spread), “û Y ” (transfer), “‹ Y ” (leave), “ê¯¹È ” (request), “M ” (deliver), “ ‹ ” (send), “• R‹ ” (throw), “ Ø ” (send), “B •‹ ” (ask), “Ð ú ” (submit), “J Q‹ ” (deliver), “• B ” (request), “; Y ” (return), “Ä Ø ” (donate), “Ä ” (donation), “@Q‹ ” (wear), “W[‹ ” (place), “ R ‹ ” (increase), “= hY ” (drop), “¼‹ ” (stick), “ Q ‹ ” (divide), “p‰~O ” (scatter), “o•‹ ” (t), “/ U F ” (pay), “M T ” (deliver), “x Y ” (unload), “• • ‹ ” (collect), “ Y ” (get X through)	“×ì¼óÈ ” (present), “ d ” (have), “ •[‹ ” (match), “‹ [‹ ” (show), “ Q‹ ” (point), “Ð › ” (offer), “Å ™” (equip), “ý ” (add), “U : ” (exhibit), “n H‹ ” (lay), “¡ (” (adopt)

Table 8: The verbs which show a signi cant change in the argument order trend depending on the semantic role of its dative argument. The scores denote DAT-ACC rate. Type-A corresponds to the examples with an inanimate goal dative argument. Type-B corresponds to the examples with an animate processor dative argument. The number of type-A verbs is signi cantly larger than that of type-B verbs.

Case	#occurrence
TIM	11,780
LOC	15,544
NOM	55,230
DAT	56,243
ACC	57,823

Table 9: The number of occurrence for each case in the data used in Section 5.2, Section 6, and Appendix F

F Additional analysis: adverbial particles and their effect for word order

The adverbial particles We can add supplementary information with adverbial particles. The adverbial particle “ó ” (TOP) is the typical one. In Example (12), the adverbial particle “ ” (also), instead of “ ” (ACC), implies that there is another thing the teacher gave to the student (that teacher gave not only but also a book to a student).

(12) ’ k , BR_ .
student-DAT also book-ACC gave.

Experiments A constituent accompanying the adverbial particle “ó ” (TOP) is moved to the beginning of the sentence (Noda, 1996). However, it is not clear whether other adverbial particles also have the above property. In this section, we evaluate the following claim: a different adverbial particle shows different degrees of the effects for the word order.

For each example S 2 S collected from Japanese Wikipedia, we replaced the postpositional particle with a speci c adverbial particle, following the rules in Table 12. We used four typical adverbial particles: “ó ” (TOP), “S] ” (emphasis), “ ” (also), and “ Q ” (only). Two variants of word order, Non-moved and Moved were created for each example. Example (13) is an example focusing on the ACC case with the particle, “ ” (also).

(13) Original: ’ k , BR_ .
student-DAT book-ACC gave.

Non-moved: ’ k , BR_ .
student-DAT also book-ACC gave.

Moved: , ’ k BR_ .
also book-ACC student-DAT book-ACC gave.

We compared the generation probabilities between the Non-moved and Moved orders. We calculated the rate that the Moved order is preferred in each combination of the case types and the adverbial particles.

Model	MODAL		TIME		MANNER		RESULTIVE	
	Canonical	r	Canonical	r	Canonical	r	Canonical	r
CLM	ASOV	1.	ASOV, SAOV	1.	SAOV, SOAV	0.5	SAOV, SOAV	1.
SLM	ASOV	1.	SAOV	0.5	SAOV, SOAV	1.	SOAV	0.5
Koizumi(2016)	ASOV	-	ASOV, SAOV	-	SAOV, SOAV	-	SAOV, SOAV	-

Table 10: Overlap of the preference of the adverb position of LMs and that of Koizumi and Tamaoka (2006). The column "Canonical" shows the adverb position, which is signi cantly preferred over the other positions. The score r denotes the Pearson correlation coef cient of the preferred rank of three possible adverb positions obtained from LMs to that of Koizumi and Tamaoka (2006).

	TIM	PLC	NOM	DAT	NOM
TIM	-	.490	.329	.720	.698
PLC	.510	-	.484	.748	.742
NOM	.671	.516	-	.804	.852
DAT	.280	.252	.196	-	.536
NOM	.302	.258	.148	.464	-

(a) CLM

	TIM	PLC	NOM	DAT	NOM
TIM	-	.538	.402	.676	.711
PLC	.462	-	.553	.757	.749
NOM	.598	.447	-	.774	.834
DAT	.324	.243	.226	-	.552
NOM	.289	.251	.166	.448	-

(b) SLM

Table 11: The scores denote $t_{a|b}$. The row corresponds to the case a, the column corresponds to b. Higher $t_{a|b}$ suggests the trend that the case a is more likely to be topicalized than the case b.

Original case particle	After the adverbial particle "o" (TOP) is added
L (TOP)	~~Lo~~
k (TIM, DAT)	ko
' (ACC)	~~'o~~
g (LOC)	go

Table 12: Rules of deleting the original case particle when the adverbial particle "o" (TOP) is added. This rule is also applied when adding the other adverbial particles (Appendix F).

Results The results are shown in Table 13. When using "o" (TOP) as a postpositional particle, the Movedorder is preferred to Non-moved, which is consistent with the well-known characteristics of topicalization described in Section 6. In addition, the degree of preference between Moved and Non-moved differs depending on the adverbial particles. Furthermore, the results indicate that the anterior case in the canonical word order is likely to move to the beginning of the sentence by the effect of the adverbial particle.

Additional experiments and results We analyzed the trend of double object order when a speci c case accompanies an adverbial particle. Figure 5 shows the result when the ACC argument accompanies an adverbial particle, and Figure 6 shows the result when the DAT argument accompanies an adverbial particle. The left parts of these gures show the result of CLM, and the right part of these gures shows the result of SLM. The X-axis denotes the ACC-DAT/ DAT-ACC rate of the verb when both of the arguments do not accompany an adverbial particle. The Y-axis denotes the ACC-DAT/ DAT-ACC rate when a speci c case accompanies an adverbial particle. The results show that the case accompanying an adverbial particle is likely to be placed near the beginning of the sentence. In addition, the degree of the above trend depends on the adverbial particles. These results suggest that some adverbial particles have a effect for word order.