A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions

Ryohei Sasano **Manabu Okumura**
Tokyo Institute of Technology
{sasano,oku}@pi.titech.ac.jp

Abstract

The canonical word order of Japanese double object constructions has attracted considerable attention among linguists and has been a topic of many studies. However, most of these studies require either manual analyses or measurements of human characteristics such as brain activities or reading times for each example. Thus, while these analyses are reliable for the examples they focus on, they cannot be generalized to other examples. On the other hand, the trend of actual usage can be collected automatically from a large corpus. Thus, in this paper, we assume that there is a relationship between the canonical word order and the proportion of each word order in a large corpus and present a corpus-based analysis of canonical word order of Japanese double object constructions.

1 Introduction

Japanese has a much freer word order than English. For example, a Japanese double object construction has six possible word orders as follows:

(1) a: *Ken-ga Aya-ni camera-wo miseta.*
Ken-NOM Aya-DAT camera-ACC showed

b: *Ken-ga camera-wo Aya-ni miseta.*
Ken-NOM camera-ACC Aya-DAT showed

c: *Aya-ni Ken-ga camera-wo miseta.*
Aya-DAT Ken-NOM camera-ACC showed

d: *Aya-ni camera-wo Ken-ga miseta.*
Aya-DAT camera-ACC Ken-NOM showed

e: *Camera-wo Ken-ga Aya-ni miseta.*
camera-ACC Ken-NOM Aya-DAT showed

f: *Camera-wo Aya-ni Ken-ga miseta.*
camera-ACC Aya-DAT Ken-NOM showed

In these examples, the position of the verb *miseta* (showed) is fixed but the positions of its nominative (NOM), dative (DAT), and accusative (ACC) arguments are scrambled. Note that, although the word orders are different, they have essentially the same meaning "Ken showed a camera to Aya."

In the field of linguistics, each language is assumed to have a basic word order that is fundamental to its sentence structure and in most cases there is a generally accepted theory on the word order for each structure. That is, even if there are several possible word orders for essentially same sentences consisting of the same elements, only one of them is regarded as the canonical word order and the others are considered to be generated by scrambling it. However, in the case of Japanese double object constructions, there are several claims on the canonical argument order.

There have been a number of studies on the canonical word order of Japanese double object constructions ranging from theoretical studies (Hoji, 1985; Miyagawa and Tsujioka, 2004) to empirical ones based on psychological experiments (Koizumi and Tamaoka, 2004; Nakamoto et al., 2006; Shigenaga, 2014) and brain science (Koso et al., 2004; Inubushi et al., 2009). However, most of them required either manual analyses or measurements of human characteristics such as brain activities or reading times for each example. Thus, while these analyses are reliable for the example they focus on, they cannot be easily generalized to other examples[1]. That is, another manual analysis or measurement is required to consider the canonical word order of another example.

On the other hand, the trend of actual usage can be collected from a large corpus. While it is difficult to say whether a word order is canonical or

[1]Note that in this work, we assume that there could be different canonical word orders for different double-object sentences as will be explained in Section 2.2.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2236–2244,
Berlin, Germany, August 7-12, 2016.

not from one specific example, we can consider that a word order would be canonical if it is overwhelmingly dominant in a large corpus. For example, since the DAT−ACC order[2] is overwhelmingly dominant in the case that the verb is *kanjiru* (feel), its dative argument is *kotoba* (word), and its accusative argument is *aijô* (affection) as shown in Example (2), we can consider that the DAT−ACC order would be canonical in this case. Note that, the numbers in parentheses represent the proportion of each word order in Examples (2) and (3); ϕ_X denotes the omitted noun or pronoun X in this paper.

(2) **DAT−ACC:** *Kotoba-ni aijô-wo kanjiru.*
(97.5%) word−DAT affection−ACC feel

ACC−DAT: *Aijô-wo kotoba-ni kanjiru.*
(2.5%) affection−ACC word−DAT feel

(ϕ_I feel the affection in ϕ_{your} words.)

On the contrary, since the ACC−DAT order is overwhelmingly dominant in the case that the verb is *sasou* (ask), its dative argument is *dêto* (date), and its accusative argument is *josei* (woman) as shown in Example (3), the ACC−DAT order is considered to be canonical in this case.

(3) **DAT−ACC:** *Dêto-ni josei-wo sasou.*
(0.4%) date−DAT woman−ACC ask

ACC−DAT: *Josei-wo dêto-ni sasou.*
(99.6%) woman−ACC date−DAT ask

(ϕ_I ask a woman out on a date.)

Therefore, in this paper, we assume that there is a relationship between the canonical word order and the proportion of each word order in a large corpus and attempt to evaluate several claims on the canonical word order of Japanese double object constructions on the basis of a large corpus. Since we extract examples of double object constructions only from reliable parts of parses of a very large corpus, consisting of more than 10 billion unique sentences, we can reliably leverage a large amount of examples. To the best of our knowledge, this is the first attempt to analyze the canonical word order of Japanese double object constructions on the basis of such a large corpus.

[2]Since Japanese word order is basically subject-object-verb (SOV) and thus the canonical position of nominative argument is considered to be the first position, we simply call the nominative, dative, accusative order as the DAT−ACC order, and the nominative, accusative, dative order as the ACC−DAT order in this paper.

2 Japanese double object constructions

2.1 Relevant Japanese grammar

We briefly describe the relevant Japanese grammar. Japanese word order is basically subject-object-verb (SOV) order, but the word order is often scrambled and does not mark syntactic relations. Instead, postpositional case particles function as case markers. For example, nominative, dative, and accusative cases are represented by case particles *ga*, *ni*, and *wo*, respectively.

In a double object construction, the subject, indirect object, and direct object are typically marked with the case particles *ga* (nominative), *ni* (dative), and *wo* (accusative), respectively, as shown in Example (4)-a.

(4) a: *Watashi-ga kare-ni camera-wo miseta.*
I−NOM him−DAT camera−ACC showed

b: *Watashi-wa kare-ni camera-wo miseta.*
I−TOP him−DAT camera−ACC showed

c: ϕ_I *kare-ni camera-wo miseta.*
ϕ_I-NOM him−DAT camera−ACC showed

(I showed him a camera.)

However, when an argument represents the topic of the sentence (TOP), the topic marker *wa* is used as a postpositional particle, and case particles *ga* and *wo* do not appear explicitly. For example, since *watashi* (I) in Example (4)-b represents the topic of the sentence, the nominative case particle *ga* is replaced by the topic marker *wa*.

Similarly, an argument modified by its predicate does not accompany a postpositional case particle that represents the syntactic relation between the predicate and argument. For example, since *camera* in Example (5) is modified by the predicate *miseta* (showed), the accusative case particle *wo* does not appear explicitly.

(5) *Watashi-ga kare-ni miseta camera.*
I−NOM him−DAT showed camera

(A camera that I showed him.)

In addition, arguments are often omitted in Japanese when we can easily guess what the omitted argument is or we do not suppose a specific object. For example, the nominative argument is omitted in Example (4)-c, since we can easily guess the subject is the first person.

These characteristics make it difficult to automatically extract examples of word orders in double object construction from a corpus.

2.2 Canonical argument order

There are three major claims as to the canonical argument order of Japanese double object constructions (Koizumi and Tamaoka, 2004).

One is the traditional analysis by Hoji (1985), which argues that only the nominative, dative, accusative (DAT−ACC) order like in Example (1)-a is canonical for all cases. The second claim, made by Matsuoka (2003), argues that Japanese double object constructions have two canonical word orders, the DAT−ACC order and the ACC−DAT order, depending on the verb types. The third claim, by Miyagawa (1997), asserts that both the DAT−ACC order and ACC−DAT order are canonical for all cases.

Note that, the definition of the term *canonical word order* varies from study to study. Some studies presume that there is only one canonical word order for one construction (Hoji, 1985), while others presume that a canonical word order can be different for each verb or each tuple of a verb and its arguments (Matsuoka, 2003). In addition, some studies presume that there can be multiple canonical word orders for one sentence (Miyagawa, 1997). In this paper, we basically adopt the position that there is only one canonical word order for one tuple of a verb and its arguments but the canonical word orders can be different for different tuples of a verb and its arguments.

2.3 Other features related to word order

There are a number of known features that affect word order. For example, it is often said that long arguments tend to be placed far from the verb, whereas short arguments tend to be placed near the verb. The From-Old-to-New Principle (Kuno, 2006) is also well known; it argues that the unmarked word order of constituents is old, predictable information first; and new, unpredictable information last. Note that these types of features are not specific to argument orders in Japanese double object constituents. For example, Bresnan et al. (2007) reported the similar features were also observed in the English dative alternation and useful for predicting the dative alternation.

However, since we are interested in the canonical word order, we do not want to take these features into account. In this work, we assume that these features can be ignored by using a very large corpus and analyzing the word order on the basis of statistical information acquired from the corpus.

3 Claims on the canonical word order of Japanese double object constructions

In this paper, we will address the following five claims on the canonical word order of Japanese double object constructions.

Claim A: The DAT−ACC order is canonical.

Claim B: There are two canonical word orders, the DAT−ACC and the ACC−DAT order, depending on the verb types.

Claim C: An argument whose grammatical case is infrequently omitted with a given verb tends to be placed near the verb.

Claim D: The canonical word order varies depending on the semantic role and animacy of the dative argument.

Claim E: An argument that frequently co-occurs with the verb tends to be placed near the verb.

Claim A (Hoji, 1985) presumes that there is only one canonical word order for Japanese double object constructions regardless of the verb type. On the other hand, Claims B and C argue that the canonical word order varies depending on verb, but they still do not take into account the lexical information of the arguments. Thus, these claims can be verified by investigating the distribution of word orders for each verb.

With regard to Claim B, Matsuoka (2003) classified causative-inchoative alternating verbs into two types: show-type and pass-type, and claimed the DAT−ACC order is the canonical order for show-type verbs, whereas the ACC−DAT order is the canonical order for pass-type verbs. The definitions of each verb type are as follows. In the case of show-type verbs, the dative argument of a causative sentence is the subject of its corresponding inchoative sentence as shown in Example (6). On the other hand, in the case of pass-type verbs, the accusative argument is the subject of its corresponding inchoative sentence as shown in Example (7).

(6) **Causative:** *Kare-ni camera-wo miseta.*
him−DAT camera−ACC showed
(ϕ_I showed him a camera.)

Inchoative: *Kare-ga mita.*
he−NOM saw
(He saw $\phi_{something}$.)

(7) **Causative:** *Camera-wo kare-ni watashita.*
camera−ACC him−DAT passed
(ϕ_I passed him a camera.)

Inchoative: *Camera-ga watatta.*
camera−NOM passed
(A camera passed to $\phi_{someone}$.)

Claim C is based on our observation. It is based on the assumption that if an argument of a verb is important for interpreting the meaning of the verb, it tends to be placed near the verb and does not tend to be omitted.

Claims D and E take into account the lexical information of arguments and assume that the canonical word order of Japanese double object constructions is affected by the characteristics of the dative and/or accusative arguments. With regard to Claim D, Matsuoka (2003) asserted that the canonical order varies depending on the semantic role of the dative argument. Specifically, the DAT−ACC order is more preferred when the semantic role of dative argument is animate *Possessor* than when the semantic role is inanimate *Goal*.

Claim E is based on our observation again, which argues that if the dative or accusative argument frequently co-occurs with the verb, it has a strong relationship with the verb, and thus is placed nearby. A typical example that satisfies this claim is idiomatic expressions as will be discussed in Section 5.4.

4 Example collection

A corpus-based analysis of canonical word order can leverage a much larger number of examples than approaches based on theoretical analysis, psychological experiments, or brain science can. However, automatically collected examples sometimes include inappropriate ones. For example, if we extract all sequences of a verb and its preceding argument candidates, the sequence "*Kagi-wo kare-ni iwareta*" (ϕ_I am told the key by him) is mistakenly extracted from Example (8), although *kagi-wo* is not actually an argument of *iwareta* but an argument of *oita*.

(8) *Kagi-wo kare-ni iwareta basho-ni oita.*
key−ACC him−DAT told place−DAT put
(ϕ_I put the key on the place where he told ϕ_{me}.)

As predicted, we can alleviate this problem by using a dependency parser. However, the accuracy of the state-of-the-art Japanese dependency parser is not very high, specifically about 92% for news paper articles (Yoshinaga and Kitsuregawa, 2014), and thus, inappropriate examples would be extracted even if we used one.

Therefore, in this work, we decided to extract examples only from reliable parts of dependency parses. Specifically, we used a corpus consisting of more than 10 billion unique sentences extracted from the Web, selected parse trees that have no syntactic ambiguity, and then extracted examples only from the selected parse trees. This strategy basically follows Kawahara and Kurohashi (2002)'s strategy for automatic case frame construction. The detailed procedure of example collection is as follows:

1. Extract Japanese Web pages using linguistic information, split the Web pages into sentences using periods and HTML tags, and merge sentences that are the exactly same into one sentence to avoid collecting the same example several times, which might be extracted from a mirror site.

2. Employ the Japanese morphological analyzer JUMAN[3] and the syntactic analyzer KNP[4], and extract examples of verbs and their arguments from parse trees that have no syntactic ambiguity[5].

3. Collect the examples if the verb satisfies all the following conditions:

   (a) The verb has an entry in the JUMAN dictionary and appears in the active voice.

   (b) The verb has more than 500 different examples of dative and accusative argument pairs.

   (c) The proportion of examples that include both the dative and accusative arguments out of all examples that include the target verb is larger than 5%.

We employ the syntactic analyzer KNP with options "-dpnd-fast -tab -check." KNP with these

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

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

[5]Murawaki and Kurohashi (2012) reported that 20.7% of the dependency relations were extracted from a newspaper corpus and the accuracy was 98.3% when they adopted Kawahara and Kurohashi (2002)'s strategy.

options outputs all head candidates for each *bunsetsu*[6] on the basis of heuristic rules. We then extract the example of a verb and its argument if the argument candidate have only one head candidate.

For example, since Japanese is a head-final language and only the verb *bunsetsu* can be the head of the most noun *bunsetsu* in Japanese, *basho-ni* in Example (8) has only one head candidate *oita* (put), whereas *kagi-wo* and *kare-ni* have two head candidates *iwareta* (told) and *oita* (put). Thus, we extract only the example "*basho-ni oita*" from Example (8). In addition, when an argument consists of a compound noun, we only extract the head noun and its postpositional particle as the argument to avoid data sparsity.

Condition 3-(c) is set in order to extract only ditransitive verbs, which take both dative and accusative arguments. Although the threshold of 5% seems small at first glance, most verbs that satisfy it are actually ditransitive. This is because arguments are often omitted in Japanese, and thus, only some of the examples explicitly include both dative and accusative arguments even in the case of ditransitive verb.

Out of a corpus consisting of more than 10 billion unique sentences, 648 verbs satisfied these conditions. Hereafter, we will focus on these 648 verbs. The average number of occurrences of each verb was about 350 thousand and the average number of extracted examples that include both dative and accusative arguments was about 59 thousand.

5 Corpus-based analysis of canonical word order

Here, we present a corpus-based analysis of the canonical word order of Japanese double object constructions. We will address Claims A and C in Section 5.1, Claim B in Section 5.2, Claim D in Section 5.3, and Claim E in Section 5.4.

5.1 Word order for each verb

Let us examine the relation between the proportion of the DAT only example $R_{\text{DAT-only}}$ and the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$ for each of the 648 verbs to inspect Claims A and C.

[6]In Japanese, *bunsetsu* is a basic unit of dependency, consisting of one or more content words and the following zero or more function words. In this paper, we segment each example sentence into a sequence of *bunsetsu*.

$R_{\text{DAT-only}}$ is calculated as follows:

$$R_{\text{DAT-only}} = \frac{N_{\text{DAT-only}}}{N_{\text{DAT-only}} + N_{\text{ACC-only}}},$$

where $N_{\text{DAT/ACC-only}}$ is the number of example types that only include the corresponding argument out of the dative and accusative arguments. For example, we count the number of example types like Example (9) that include an accusative argument but do not include a dative argument to get the value of $N_{\text{ACC-only}}$. Accordingly, the large $R_{\text{DAT-only}}$ value indicates that the dative argument is less frequently omitted than the accusative argument.

(9) *Gakuchô-ga gakui-wo juyo-shita.*
president−NOM degree−ACC conferred
(The president conferred a degree on $\phi_{someone}$.)

However, if we use all extracted examples that include only one of the dative and accusative arguments for calculating $R_{\text{DAT-only}}$, the value is likely to suffer from a bias that the larger $R_{\text{ACC-DAT}}$ is, the larger $R_{\text{DAT-only}}$ becomes. This is because the arguments that tend to be placed near the verb have relatively few syntactic ambiguity. Since we extract the examples from the reliable parts of parses that have no syntactic ambiguity, these arguments tend to be included in the extracted examples more frequently than the other arguments.

To avoid this bias, we use only these examples in which the nominative case is also extracted for calculating $R_{\text{DAT-only}}$. This is based on the assumption that since Japanese word order is basically subject-object-verb order, if the nominative argument is collected but one of the dative and accusative arguments is not collected, the argument is actually omitted. Through a preliminary investigation on Kyoto University Text Corpus[7], we confirmed the effect of this constraint to avoid the bias.

On the other hand, $R_{\text{ACC-DAT}}$ is calculated as follows:

$$R_{\text{ACC-DAT}} = \frac{N_{\text{ACC-DAT}}}{N_{\text{DAT-ACC}} + N_{\text{ACC-DAT}}},$$

where $N_{\text{DAT-ACC/ACC-DAT}}$ is the number of example types that include both the dative and accusative arguments in the corresponding order.

Figure 1 shows the results. The left figure shows the relation between the proportion of the DAT

[7]Kyoto University Text Corpus 4.0: http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Kyoto University Text Corpus

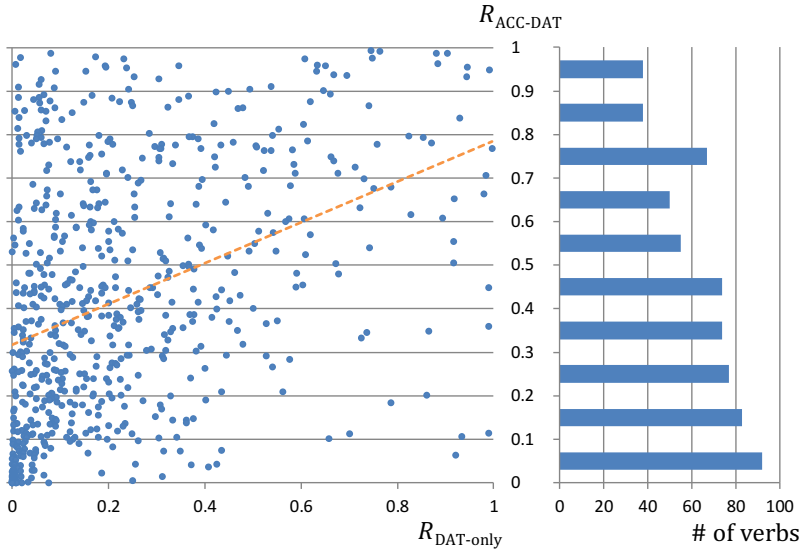

Figure 1: The left figure shows the relation between the proportion of the DAT only example $R_{\text{DAT-only}}$ (x-axis) and the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$ (y-axis). The right figure shows the number of verbs in the corresponding range of $R_{\text{ACC-DAT}}$.

only example $R_{\text{DAT-only}}$ and the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$ for each of the 648 verbs. The x-axis denotes $R_{\text{DAT-only}}$, the y-axis denotes $R_{\text{ACC-DAT}}$, and each point in the figure represents one of the 648 verbs. The dashed line is a linear regression line. The right figure shows the number of verbs in the corresponding range of $R_{\text{ACC-DAT}}$.

Pearson's correlation coefficient between $R_{\text{DAT-only}}$ and $R_{\text{ACC-DAT}}$ is 0.391, which weakly supports Claim C: an argument whose grammatical case is infrequently omitted with a given verb tends to be placed near the verb. The proportion of the ACC−DAT order for all 648 verbs is 0.328. Thus, if we presume that there is only one canonical word order for Japanese double object constructions, this result suggests that the DAT−ACC order is the canonical one, as claimed by Hoji (Claim A). However, the right figure shows that the proportions of the ACC−DAT order differ from verb to verb. Moreover, the values of $R_{\text{ACC-DAT}}$ for 435 out of 648 verbs are between 0.2 and 0.8. From these observations, we can say the preferred word order cannot be determined even if the verb is given in most cases.

5.2 Word order and verb type

To inspect Matsuoka (2003)'s claim that the DAT−ACC order is canonical for show-type verbs, whereas the ACC−DAT order is canonical for pass-type verbs, we investigated the proportions of the ACC−DAT order for several pass-type and show-type verbs. In this investigation, we used 11 pass-type verbs and 22 show-type verbs that were used by Koizumi and Tamaoka (2004) in their psychological experiments[8].

Table 1 shows the results. Although we can see that the macro average of $R_{\text{ACC-DAT}}$ of pass-type verbs is larger than that of show-type verbs, the difference is not significant[9]. Moreover, even in the case of pass-type verbs, the DAT−ACC order is dominant, which suggests Matsuoka (2003)'s claim is not true. Note that this conclusion is consistent with the experimental results reported by both Miyamoto and Takahashi (2002) and Koizumi and Tamaoka (2004).

5.3 Relation between word order and semantic role of the dative argument

Next, let us examine the relation between the category of the dative argument and the word order to verify the effect of the semantic role of the dative argument. We selected eight categories in the JUMAN dictionary[10] that appear more than 1 million times as dative arguments. Table 2 shows the results. We can see that there are differences in the

[8] We excluded a show-type verb *hakaseru* (dress), since it is divided into two morphemes by JUMAN. Instead, we added two show-type verbs *shiraseru* (notify) and *kotodukeru* (leave a message).

[9] The two-tailed p-value of permutation test is about 0.177.

[10] In JUMAN dictionary, 22 categories are defined and tagged to common nouns.

Show-type		Pass-type			
verb	$R_{\texttt{ACC-DAT}}$	verb	$R_{\texttt{ACC-DAT}}$	verb	$R_{\texttt{ACC-DAT}}$
shiraseru (notify)	0.522	*modosu* (put back)	0.771	*otosu* (drop)	0.351
azukeru (deposit)	0.399	*tomeru* (lodge)	0.748	*morasu* (leak)	0.332
kotodukeru (leave a message)	0.386	*tsutsumu* (wrap)	0.603	*ukaberu* (float)	0.255
satosu (admonish)	0.325	*tsutaeru* (inform)	0.522	*mukeru* (direct)	0.251
miseru (show)	0.301	*noseru* (place on)	0.496	*nokosu* (leave)	0.238
kabuseru (cover)	0.256	*todokeru* (deliver)	0.491	*umeru* (bury)	0.223
osieru (teach)	0.235	*naraberu* (range)	0.481	*mazeru* (blend)	0.200
sazukeru (give)	0.186	*kaesu* (give back)	0.448	*ateru* (hit)	0.185
abiseru (shower)	0.177	*butsukeru* (knock)	0.436	*kakeru* (hang)	0.108
kasu (lend)	0.118	*tsukeru* (attach)	0.368	*kasaneru* (pile)	0.084
kiseru (dress)	0.113	*watasu* (pass)	0.362	*tateru* (build)	0.069
Macro average	0.274			Macro average	0.365

Table 1: Proportions of the ACC−DAT order for each pass-type verb and show-type verb.

Category	# of examples	$R_{\texttt{ACC-DAT}}$	Typical examples
PLACE-FUNCTION	1376990	0.499	*shita* (bottom), *yoko* (side), *soto* (outside), *hôkô* (direction), ...
ANIMAL-PART	1483885	0.441	*te* (hand), *mi* (body), *atama* (head), *hada* (skin), *mune* (chest), ...
PERSON	5511281	0.387	*tomodachi* (friend), *hito* (human), *shichô* (mayor), *watashi* (I), ...
ARTIFACT-OTHER	2751008	0.372	*pasokon* (PC), *fairu* (file), *furo* (bath), *hon* (book), ...
PLACE-INSTITUTION	1618690	0.342	*heya* (room), *mise* (shop), *tokoro* (location), *gakkô* (school), ...
PLACE-OTHER	2439188	0.341	*basho* (place), *sekai* (world), *ichi* (position), *zenmen* (front), ...
QUANTITY	1100222	0.308	*zu* (figure), *hyô* (table), *hanbun* (half), *atai* (value), ...
ABSTRACT	10219318	0.307	*blog* (blog), *kokoro* (mind), *list* (list), *shiya* (sight), ...
Total	26500582	0.353	

Table 2: Proportions of the ACC−DAT order for each category of dative argument.

proportions of the ACC−DAT order. In particular, when the dative argument's category is PLACE-FUNCTION such as *shita* (bottom) and *yoko* (side) or ANIMAL-PART such as *te* (hand) and *mi* (body), the ACC−DAT order is more preferred than otherwise.

As mentioned in Section 3, Matsuoka (2003) claimed the DAT−ACC order is more preferred when the semantic role of the dative argument is animate *Possessor* than when the semantic role is inanimate *Goal*. Thus, we thought the DAT−ACC order would be preferred when the dative argument's category is PERSON, but we could not find such a trend. We think, however, this is due to that dative arguments of the PERSON category do not always have the semantic role of an animate *Possessor*. Thus, we conducted a further investigation in an attempt to verify Matsuoka (2003)'s claim.

First, we collected examples that satisfied the following two conditions: the accusative argument belongs to ARTIFACT-OTHER category, and the dative argument belongs to either PLACE-INSTITUTION or PERSON category. We call the former Type-A[11], and the latter Type-B hereafter, and consider that the semantic role of the dative argument is inanimate *Goal* in most cases of Type-A, whereas it is animate *Possessor* in most cases of Type-B. Example (10) shows typical examples of Type-A and Type-B. Here, the categories of *hon* (book), *gakkô* (school), and *sensei* (teacher) are ARTIFACT-OTHER, PLACE-INSTITUTION, and PERSON, respectively, and the semantic roles of dative arguments are considered to be *Goal* in (10)-a and *Possessor* in (10)-b.

(10) a: *Hon-wo* *gakkô-ni* *kaeshita.*
book−ACC school−DAT returned
($\phi_{someone}$ returned the book to school.)

b: *Sensei-ni* *hon-wo* *kaeshita.*
teacher−DAT book−ACC returned
($\phi_{someone}$ returned the book to the teacher.)

Next, we extracted verbs that had at least 100 examples of both types, calculated the proportion of the ACC−DAT order $R_{\texttt{ACC-DAT}}$ for each verb and type, and counted the number of verbs for which the values of $R_{\texttt{ACC-DAT}}$ were significantly different between Type-A and Type-B[12]. Out of 126 verbs that have at least 100 examples for both types, 64 verbs show the trend that Type-A prefers the ACC−DAT order more than Type-B does, and only 30 verbs have the opposite trend. This fact supports Matsuoka (2003)'s claim.

[11] That is, the categories of the accusative and dative arguments of a Type-A example are ARTIFACT-OTHER and PLACE-INSTITUTION, respectively.

[12] We conducted a two-proportion z-test with a significance level of 0.05.

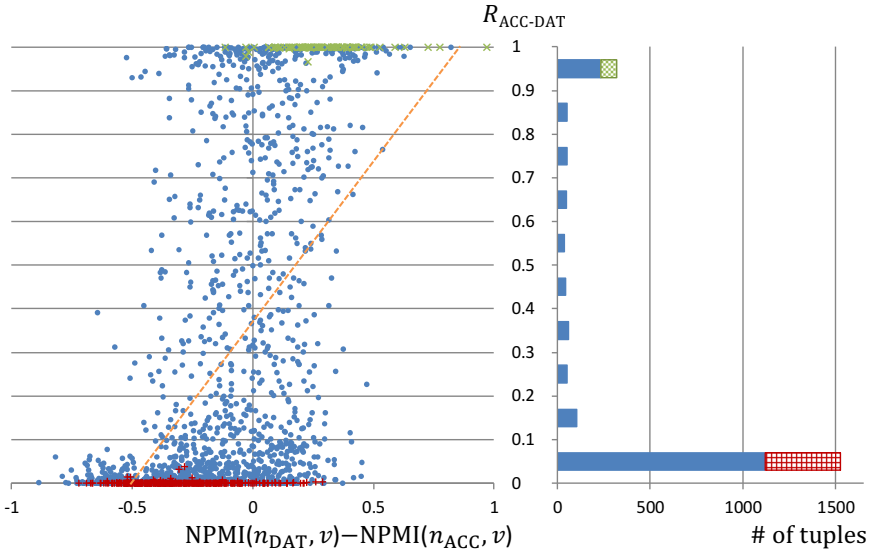

Figure 2: The left figure shows the relation between the difference of NPMI(n_{DAT}, v) from NPMI(n_{ACC}, v) (x-axis) and the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$ (y-axis). The tuples whose verb and accusative/dative argument are used as an idiom are represented by +/×. The right figure shows the number of tuples of a verb and its dative and accusative arguments in the corresponding range of $R_{\text{ACC-DAT}}$.

5.4 Relation between word order and degree of co-occurrence of verb and arguments

Now let us turn to the relation between the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$ and the degree of co-occurrence of a verb and its argument to verify Claim E. Here, we leverage the normalized pointwise mutual information (NPMI) for measuring the degree of co-occurrence between a verb and its argument. NPMI is a normalized version of PMI. The value ranges between [-1,+1] and takes -1 for never occurring together, 0 for independence, +1 for complete co-occurrence. The NPMI of a verb v and its argument n_c ($c \in \{\text{DAT}, \text{ACC}\}$) is calculated as

$$\text{NPMI}(n_c, v) = \frac{\text{PMI}(n_c, v)}{-\log(p(n_c, v))},$$

$$\text{where PMI}(n_c, v) = \log \frac{p(n_c, v)}{p(n_c)p(v)}.$$

We investigate the relation between the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$ and the difference of NPMI(n_{DAT}, v) from NPMI(n_{ACC}, v), i.e., NPMI(n_{DAT}, v) − NPMI(n_{ACC}, v). If Claim E is true, when the dative argument co-occurs with the verb frequently, the dative argument tends to be placed near the verb and thus the proportion of the ACC−DAT order would take a large value.

We investigated 2302 tuples of a verb and its dative and accusative arguments that appear more than 500 times in the corpus. The average number of occurrences of each tuple was 1532. Figure 2 shows the results. The left figure shows the relation between the difference of NPMI(n_{DAT}, v) from NPMI(n_{ACC}, v) and the proportion of the ACC−DAT order $R_{\text{ACC-DAT}}$. Each point in the figure represents one of the 2302 tuples. The dashed line is a linear regression line. The right figure shows the number of tuples in the corresponding range of $R_{\text{ACC-DAT}}$.

Pearson's correlation coefficient between the difference of NPMI and $R_{\text{ACC-DAT}}$ is 0.567, which supports Claim E: an argument that frequently co-occurs with the verb tends to be placed near the verb. Moreover, the values of $R_{\text{ACC-DAT}}$ are larger than 0.9 or smaller than 0.1 for 1631 out of 2302 tuples. This result indicates that if a tuple of a verb and its dative and accusative arguments is given, the preferred word order is determined. This is contrastive to the conclusion that the preferred word order cannot be determined even if the verb is given as discussed in Section 5.1.

One of the typical examples that satisfy Claim E is an idiomatic expression. Indeed, a verb and its argument that are used as an idiom co-occur frequently and are usually placed adjacent to each other. In addition, it is well known that if the argument order is scrambled, the idiomatic meaning disappears (Miyagawa and Tsujioka, 2004). Thus, we investigated to what extent idiomatic expres-

sions affected the findings discussed above. For all 2302 tuples, we manually judged whether the verb and the adjacent argument are used as an idiom in most cases. As a result, the verbs and their accusative arguments are judged as idiomatic for 404; the verbs and their dative arguments are judged as idiomatic for 84 out of 2302 tuples. We show these tuples by $+$ and $\times$ in Figure 2, respectively. As predicted, the values of $R_{\texttt{ACC-DAT}}$ are smaller than 0.1 for all of the former examples, and larger than 0.9 for all of the latter examples. However, even if we ignore these idiomatic examples, Pearson's correlation coefficient between the difference of NPMI and $R_{\texttt{ACC-DAT}}$ is 0.513, which is usually considered as moderate correlation.

6 Conclusion

This paper presented a corpus-based analysis of canonical word order of Japanese double object constructions. Our analysis suggests 1) the canonical word order of such constructions varies from verb to verb, 2) there is only a weak relation between the canonical word order and the verb type: show-type and pass-type, 3) an argument whose grammatical case is infrequently omitted with a given verb tends to be placed near the verb, 4) the canonical word order varies depending on the semantic role of the dative argument, and 5) an argument that frequently co-occurs with the verb tends to be placed near the verb.

Acknowledgments

We would like to thank Sadao Kurohashi and Daisuke Kawahara for helping us to collect examples from the Web. This work was supported by JSPS KAKENHI Grant Number 25730131 and 16K16110.

References

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer, and Joost Zwarts, editors, *Cognitive foundations of interpretation*, pages 69–94. Amsterdam: Royal Netherlands Academy of Science.

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

Tomoo Inubushi, Kazuki Iijima, Masatoshi Koizumi, and Kuniyoshi L. Sakai. 2009. The effect of canonical word orders on the neural processing of double object sentences: An MEG study. In *Proceedings of the 32nd Annual Meeting of the Japan Neuroscience Society*.

Daisuke Kawahara and Sadao Kurohashi. 2002. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 425–431.

Masatoshi Koizumi and Katsuo Tamaoka. 2004. Cognitive processing of Japanese sentences with ditransitive verbs. *Gengo Kenkyu*, 125:173–190.

Ayumi Koso, Hiroko Hagiwara, and Takahiro Soshi. 2004. What a multi-channel EEG system reveals about the processing of Japanese double object constructions (in Japanese). In *Technical report of IEICE. Thought and language 104(170)*, pages 31–36.

Susumu Kuno. 2006. Empathy and direct discourse perspectives. In Laurence Horn and Gergory Ward, editors, *The Handbook of Pragmatics*, Blackwell Handbooks in Linguistics, pages 315–343. Wiley.

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

Shigeru Miyagawa and Takae Tsujioka. 2004. Argument structure and ditransitive verbs in Japanese. *Journal of East Asian linguistics*, 13:1–38.

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

Edson T. Miyamoto and Shoichi Takahashi. 2002. Sources of difficulty in processing scrambling in Japanese. In Mineharu Nakayama, editor, *Sentence processing in East Asian languages. Stanford, Calif*, pages 167–188. CSLI Publications.

Yugo Murawaki and Sadao Kurohashi. 2012. Semi-supervised noun compound analysis with edge and span features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1915–1932.

Keiko Nakamoto, Jae-Ho Lee, and Kow Kuroda. 2006. Cognitive mechanisms for sentence comprehension preferred word orders correlate with "sentential" meanings that cannot be reduced to verb meanings: A new perspective on "construction effects" in Japanese (in Japanese). *Cognitive Studies*, 13(3):334–352.

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.

Naoki Yoshinaga and Masaru Kitsuregawa. 2014. A self-adaptive classifier for efficient text-stream processing. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2014)*, pages 1091–1102.