Interaction between Scrambling and Topic Marker *WA* in Centering Theory

Satoshi Imamura and Einar Andreas Helgason

Keywords: Centering Theory; Japanese; Scrambling; Topic Marker

Abstract

The aim of this paper is to investigate the discourse function of scrambling and the topic marker *wa*, using a quantitative approach. We employ Centering Theory to analyze written modern Japanese corpus data. We show that objects in $O_{ACC}S_{NOM}V$ sentences tend to be backward-looking centers (Cb) and that Cf-ranking is based on grammatical function even in scrambled word order. In $O_{TOP}S_{NOM}V$, objects are Cb in general we cannot predict which referent is Cp in the following utterance. Therefore, we hypothesize that scrambling predicts topic shift from object to subject but this phenomenon disappears when scrambled object is marked with topic marker *wa*. This is because the function of scrambling conflicts with that of topicalized object. To put it more concretely, OSV is used to show that Cp is the subject whereas topicalizing object conveys the intention that the object is Cp. Third, in order to examine our hypothesis, we analyzed $O_{ACC}S_{TOP}V$. The results support our hypothesis because topic shift occurs in $O_{ACC}CS_{TOP}V$. Moreover, I propose that animacy also has an influence on Cf-ranking on the basis of the difference between our data and Kameyama (1985)'s example in $O_{TOP}S_{TOP}V$.

1. Introduction

In Japanese, a language with a relatively flexible word order, both SOV and OSV orders have the same basic meaning. Japanese speakers, however, must choose one of the word orders in order to express their intention. This is a general property of all flexible word order languages. What factor, then, determines the ordering choice between SOV and OSV ? One possibility is that OSV is selected when the direct object is given information (Kuno 1978; Imamura 2014, 2015; Imamura, Sato, and Koizumi 2014, 2016). However, given information is a general concept and thus involves various grammatical operations, including topic marker *wa*. Both scrambled objects and topic marker *wa* are generally given information. Thus, it is necessary to examine the discourse function of them more closely. In order to meet this goal, we will investigate the discourse function of scrambling

and topic marker *wa* in the following sections from a viewpoint of Centering Theory. Consequently, we conclude that scrambling predicts 'topic shift' and topic marker *wa* has a strong effect on Cf-ranking.

This paper is organized in the following way. Section 2 presents an overview of previous studies regarding Centering Theory and scrambling operations. Sections 3 and 4 present a Centering analysis of scrambling and topic marker *wa* based on written corpus data. Section 5 is devoted to the conclusion.

2. Previous Studies

2－1. Centering Theory

Centering Theory (CT) is a model of local discourse coherence which was first formulated by computational linguists for tracking participants' center of attention in discourse (Brennan et al. 1987; Grosz et al. 1983, 1995; Joshi and Weinstein 1981; Kameyama 1985). According to CT, discourse coherence is measured by three concepts: backward-looking center, forward-looking center, and preferred center.

The basic assumption of CT is that speakers will make a link between a referent in their current utterance and a referent in their previous utterance so that they can maintain local coherence in their discourse. This referent is called backward- looking center (Cb) which is a link with the previous sentence; it is the most significant discourse referent under discussion in both the current and previous utterances. To put it more informally, Cb is what the sentence is about. Note that Cb must exist both in the immediately preceding utterance and in the current utterance.

CT simultaneously presupposes that speakers draw on referents from their current utterance to frame their next one. These referents are called forward-looking centers (Cfs), which are defined as members of an ordered set of referents corresponding to referents mentioned in the current utterance. They are a list of all discourse entities in a sentence that may be linked to a succeeding utterance. In other words, Cfs are candidates for becoming Cb in the following utterance. The set of Cfs is ordered by salience, ranked most often in terms of a hierarchy of grammatical relations: SUBJECT is higher than OBJECT which is higher than OTHER constituents. Informally speaking, the ranking of Cfs corresponds to the likelihood for them to become the Cb in the following utterance. However, Walker et al. (1994) note that Cf ranking is not so much universal as language-specific, depending on the means which a language provides for indicating discourse functions. Therefore, we will employ Walker et al. (1994)' s Cf rankings for analyzing Japanese sentences: (GRAMMATICAL OR ZERO) TOPIC > EMPATHY > SUBJECT > OBJECT2 > OBJECT > OTHERS (Walker et al., 1994:199). Note that GRAMMATICAL TOPIC in this hierarchy are noun

phrases with the grammatical topic marker *wa*.

The highest-ranked member of the current utterance's Cfs is designated as the preferred center (Cp) and it is the most probable candidate for Cb in a succeeding utterance. One of the members of the current utterance's Cfs is a Cb. This Cb is the highest-ranked entity from the previous utterance's Cfs that is realized in the current utterance. In addition to the Cf ranking, zero pronouns and pronouns have an effect on the choice of Cb. If some referent of the preceding utterance is realized as a (zero) pronoun in the current utterance, it must be the Cb in the current utterance (Okumura and Tamura 1996; Walker et al. 1994). Underlying this rule, there is the idea that pronouns are more coherent than other referents because they require an antecedent in the preceding context. In addition, the rule states that zero pronouns and pronouns can override the Cf rankings.

The centering framework described here above is formally summarized in (1), (2), and (3).

(1) a. Each utterance Un has at most one backward-looking center, Cb (Un), and, if Un has Cb(Un), the referent of Cb(Un) must be realized in both Un-1 and Un.

b. Every referent of the forward-looking centers list, Cfs(Un), must be realized in Un.
The highest-ranked member of Cfs(Un) is the preferred center, Cp(Un). Cp (Un) is the most probable candidate for Cb(Un+1).

c. Cb (Un) is determined by Cfs (Un-1), and, to put it more concretely Cb (Un) is the highest-ranked referent of Cf(Un-1) that is realized in Un.

(Grosz et al. 1995:208-210)

(2) Ranking of Cfs for Japanese
(GRAMMATICAL OR ZERO) TOPIC > EMPATHY > SUBJECT > OBJECT2 > OBJECT > OTHERS (Walker et al. 1994:199)

(3) Pronoun Rule
If some referent of Un is also realized as a (zero) pronoun in Un+1, this referent is Cb(Un+1)

(Walker et al. 1994:199)

2－2. Scrambling

Syntactically, it has been proposed that O in OSV is moved from the VP-internal position toward the sentence initial position (Miyagawa 2010; Saito 1985, 2009). However, the conditions under which scrambled word orders are appropriate have yet to be described adequately. From the standpoint of CT, Rambow (1993) proposed that scrambling is permitted when non-scrambled constituents either do not exist in the previous utterance, Un-1, or are lower than scrambled

constituents in the Cf ranking for Un-1. Whether the Cf ranking in a free-word-order language like Japanese is influenced by word orders or not is an issue that has yet to be resolved. Strube and Udo (1999) claim that word order itself may affect Cf ranking. Their proposal is that the degree of givenness and left-to-right order determines Cf ranking. In addition, Gordon et al. (1993) empirically revealed that surface position influences the Cf ranking. However, Miltsakaki (2002) and Hoffman (1998) showed that grammatical relations are the most reliable determinant of Cf ranking. In fact, Hoffman revealed that grammatical relations have the greatest effect on Cfs even for scrambled word order in Turkish. Therefore, further studies are needed to disentangle the issue of whether word order has an influence on Cf ranking.

3. Centering Analysis of Scrambling

We will explore the discourse function of scrambling by using CT. First, we will investigate whether or not scrambled objects are Cbs. If scrambling correlates with Cb (thereby making Rambow's (1993) analysis applicable to Japanese), scrambled constituents in Un should be higher in the Cf ranking of Un-1 than non-scrambled constituents. In other words, a scrambled constituent should be Cb. If not, non-scrambled constituents may be Cbs. Second, we will check the Cb of the succeeding sentence to see if the Cf ranking is based on grammatical relations even in non-canonical word order. If word order has an influence on Cf ranking (Gordon et al. 1993; Strube and Udo 1999), there are two possibilities, depending on the strength of that influence compared to the strength of Cf ranking. If the effect of scrambling partially offsets that of grammatical relations, the following sentence may refer to the referent of a scrambled constituent as frequently as to a non-scrambled one. The effect of scrambling might also completely override that of Cf ranking and the Cb of the following sentence is the referent of a scrambled object. On the other hand, if word order does not have an influence on Cf ranking (Hoffman 1998; Miltsakaki 2002), the antecedent of the Cb in the succeeding sentence should be associated with the highest constituent in Cf ranking largely based on grammatical relations, irrespective of word order. Lastly, we will examine the relationship between word order and grammatical topic marker *wa* because this marker also has a strong connection with Cb and Cp. The basic predictions regarding scrambling are summarized below.

(4) Prediction 1
  a. Scrambled constituents are not Cbs.
  b. Scrambled constituents are Cbs.
(5) Prediction 2
  a. If the effect of scrambling is as strong as that of grammatical relations,

Cf ranking should be ambiguous and it should be difficult to predict which entity becomes Cb in following sentences.

b. If scrambling strongly affects Cf ranking, scrambled constituents should become Cbs in following sentences.

c. If scrambling does not have an influence on Cf ranking, Cf ranking should be largely based on grammatical relations.

3－1. Method

3－1－1. Corpus Data

In order to collect OSV sentences, we employed Aozora Bunko, a database of a collection of novels written in modern Japanese.

3－1－2. Materials

Two types of OSV sentences were collected from the Aozora Bunko data. One includes sentences containing a direct object marked with the topic marker *wa* and the other includes sentences containing a scrambled element marked with the accusative case marker o. They were all transitive sentences.

3－1－3. Analyzing Complex Clauses

CT is a theory of local coherence and its basic unit of analysis is a simple sentence with only one predicate. In order to analyze complex clauses which have more than one predicate, it is necessary to divide them into separates utterances. Roughly speaking, a simple utterance coincides with a finite clause, whether subordinate or not, with the exception of relative clauses. (Hadic and Taboada 2006; Kameyama 1998).

3－1－4. Realization

Discourse entities in CT can be realized in two ways. The first is direct realization, which ignores potential bridging between the preceding utterance and the current utterance (Grosz et al. 1995). However, limiting realization only to entities that have been explicitly mentioned in the preceding utterance will cause a large number of ZERO transitions, which have no Cb. Therefore, indirect realizations of discourse entities that handle bridging relations are also required. They occur when there is an indirect relationship between an entity in the preceding utterance and an entity in the current utterance. According to Hadic Zabala and Taboada (2006), indirect realizations can be based on various types of relation: noun relation, synonym and near-synonym relation, superordinate relation, inclusive relation, and part-whole relation.

3-1-5. Procedure

Samples were collected from the whole corpus. In order to avoid garbage data such as center-embedded clauses and double-object constructions, a strict search was conducted for the strings X-*o*-X-*ga*-X-*ta* and X-*wa*-X-*ga*-X-*ta*.

3-1-6. Data Analysis

After the data was assembled, we implemented two analyses. In the first step, using the χ^2-test, we looked at whether the scrambled constituent was Cb or not in order to see if scrambling correlates with Cb. In the second step, we investigated whether scrambling has an effect on Cf ranking by checking the referents of each sentence following a scrambled sentence.

3-2. Results

Table 1. The observed frequencies of the referent of Cb for the current sentences

	$O_{ACC}SV$	$O_{TOPIC}SV$
Cb=O	24	39
Cb=S	1	8

Summary of the observed frequencies of the referent of Cb in the current sentences is shown in Table 1. First, the χ^2 test was conducted for O and S conditions in $O_{ACC}S_{NOM}V$. This revealed that there was a significant difference between them (χ^2 (1)=21. 160, p<. 001). In other words, in $O_{ACC}S_{NOM}V$, the frequency of O as Cb was significantly higher than that of S as Cb. Second, for $O_{TOP}S_{NOM}V$, there was also a significant difference in frequency between O and S as Cb (χ^2 (1)=20. 447, p<. 001), with the frequency of O as Cb higher.

Table 2. The observed frequencies of the referent of Cb for the succeeding sentences

	$O_{ACC}SV$	O_{TOPIC} SV
Cb(Un+1)=O(Un)	4	17
Cb(Un+1)=S(Un)	14	15

Next, a χ^2 test was conducted for the referent of Cb in the following sentences. Firstly, for $O_{ACC}S_{NOM}V$ condition, the differences between O and S resulted in a significant chi-square (χ^2 (1)=5. 556, p<. 05) with the frequency of S becoming Cb higher than that of O. Secondly, for $O_{TOP}S_{NOM}V$, there was no significant difference between the frequency of O becoming Cb and that of S becoming Cb (χ^2 (1)=0. 125, p=. 724).

3－3. Discussion

The first point is that we have empirically demonstrated using corpus data that a scrambled direct object is more often a Cb than not. Namely, as for Prediction 1, it is borne out that scrambled constituents correlate with Cb. This coincides with what is theoretically claimed by Rambow (1993) that a direct object is more often scrambled when it is Cb. In other words, it is revealed that $O_{ACC}S_{NOM}V$ is preferred when direct objects are Cb. This tendency also applies to $O_{TOP}S_{NOM}V$ because topicalized objects are generally Cb.

The second point of this analysis is that Cf ranking is mainly based on grammatical relations when $O_{ACC}S_{NOM}V$ occurs because most of the Cbs of the succeeding sentences have a link with subject (i. e. the highest ranked element) in the current sentences. That is, for Prediction 2, the hypothesis that Cf ranking is based largely on grammatical relations in $O_{ACC}S_{NOM}V$ is found to be true. However, this tendency was not observed for $O_{TOP}S_{NOM}V$ because the Cbs of the succeeding sentences are realized as both subject and object in the current sentences with equal frequency.

In short, from the viewpoint of CT, we found two facts about OSV. First, OSV prefers objects as Cb whether it is $O_{ACC}S_{NOM}V$ or $O_{TOP}S_{NOM}V$. Second, the behavior of Cb (Un+1) depends on whether O in OSV (Un) has a topic marker *wa* or not. To put it more concretely, subjects of $O_{ACC}S_{NOM}V$ generally become Cb(Un+1) but the behavior of Cb (Un+1) is unpredictable in $O_{TOP}S_{NOM}V$. Why do topical objects compete with nominative subjects for Cb (Un+1) ? One possibility is that the function of OSV is to make the subject the Cb (Un+1), i. e. it predicts topic shift from object to subject; while the function of the topic marker *wa* is to make the *wa*-marked NP the Cb (Un+1), i. e. to continue to talk about the same referent. Thus, the function of OSV seems to be inconsistent with the one of topical objects. If our analysis is on the right track, $O_{ACC}S_{TOP}V$ should be employed when speakers continue to talk about the referent of a subject. This is because the function of both OSV and topical subjects is to convey the intention that subjects will be Cb (Un+1). We will explore this prediction in the next section.

4. Centering Analysis of Scrambling with a Topical Subject

In section 3, we hypothesized that the purpose of scrambling is to show the intention of topic shift from object to subject, and that the topic marker *wa* is an option for speakers to highlight referents. If our hypotheses are right, sentences immediately after OACCSTOPV will have the referent of the topical subject as Cb. In contrast, if our hypotheses are wrong, we should not observe such results. Our prediction is shown in (6).

(6) Prediction 3
Subjects of $O_{ACC}S_{TOP}V$ should become Cb in the following sentence.

4－1. Method

4－1－1. Corpus Data

Aozora Bunko was employed again to assemble $O_{ACC}S_{TOP}V$ sentences.

4－1－2. Materials

Only one type of sentence was assembled: $O_{ACC}S_{TOP}V$.

4－1－3. Procedure

The procedure was the same as the one used in section 3 except that the search string used was X-*o*-X-*wa*-X-*ta*.

4－1－4. Data Analysis

We conducted a χ^2-test for the frequency of Cb in the current sentence and the following sentence in order to confirm Prediction 3.

4－2. Results

Table 3. The observed frequencies of the referent of Cb (Un) and Cb (Un+1) in $O_{ACC}S_{TOP}V$

	Cb(Un)	Cb(Un+1)
O	19	5
S	9	24

In Table 3, we demonstrate the summary of the observed frequencies of the referent of Cb in target sentences and Cb in following sentences. Firstly, the χ^2 test revealed that there was a marginal difference between the frequency of S and O as Cb (Un) (χ^2 (1)=3.571, p<.10). In other words, O seems more likely to be Cb in current sentences. Secondly, there was a significant difference between the frequency of S and O as Cb (Un+1) (χ^2 (1)=12.448, p<.001). This means that a topical subject is preferable to O for becoming Cb in following sentences.

4－3. Discussion

Our prediction in (6) is borne out because topical subjects in $O_{ACC}S_{TOP}V$ are usually Cb (Un+1). This supports our claim that scrambling predicts topic shift, because, as we saw above, scrambled objects are more likely to be Cb (Un) than topical subjects. It has been said that *wa* is a topic marker, but, when it co-occurs with scrambling, scrambled constituents are preferable as discourse-anaphoric referents to *wa*-marked constituents.

5. General Discussion

In section 3 and 4, we have revealed that scrambled objects tend to be Cb but subjects are generally Cp. In other words, scrambled sentences with accusative objects are accompanied by a shift of topic from the accusative object to the subject. Yet, this function is not observed in scrambled sentences with a topicalized subject. Why did we observe such a result ? Walker et al. (1994) observed that a grammatical topic (*wa*-marked entity) is salient and weakens the discourse function of zero objects. In other words, after a sentence which contains both a zero object and a topic-marked subject, you cannot predict which referent will be Cb in the following sentence. In our opinion, this analysis can be extended to OTOPSNOMV. This is because, after OTOPSNOMV sentences, we cannot predict which referent will be carried over to succeeding sentences. In fact, in 17 examples it was the sentence initial object, while in 15 cases, it was the nominative subject. However, this tendency cannot be explained by a Cf ranking where *wa*-marked objects are ranked higher than ga-marked subjects. Why are *wa*-marked objects as high as ga-marked subjects in the Cf ranking ? In order to solve this problem, we compared actual examples we assembled with Kameyama (1985:114), which is the important basis for Cf ranking in Japanese. Kameyama's example is shown (7a). In this example, both the subject and object are human. In contrast, almost all of our actual examples (50 out of 53 examples) have a human subject and an inanimate object as in (7b). This fact suggests that animacy may have an influence on Cf ranking.

(7) a. Jirō-*wa* Tarō-ga kinō sanzan hihanshi-ta.
Jiro-TOP Taro-NOM yesterday severely criticize-PAST
'Jiro, Taro severely criticized [him] yesterday.'

b. Murasaki-no Joō-no tegami-*wa* Minamoto-no Chūshō-ga motteki-ta.
purple-GEN queen-GEN letter-TOP Minamoto-GEN chusho-NOM bring-PAST
'Lieutenant general Minamoto brought the Purple Queen's letter.'

Therefore, the reason why our conclusion does not match Kameyama (1985) is because the data she uses to construct the original Cf ranking for Japanese was based on constructed examples with animate objects while the naturally-occurring examples in our corpus mostly had inanimate objects.

Returning to our main point, we would like to contend that topicalized objects and nominative subjects compete with each other. Namely, word order change (OSV) makes subjects more salient while topic-marking direct objects strengthens their salience. As a result, the salience of the subject competes with that of the direct object in $O_{TOP}S_{NOM}V$ sentences, and we cannot predict which referent will be referred to in the following sentence. On the basis of the interactions between

topic markers, OSV word order, and zero objects, we hypothesize that topic markers compete with not only zero objects but also the word order.

6. Conclusion

In this paper, we have compared the discourse function of scrambling and topic marker *wa*. Based on written Japanese data, we have demonstrated that scrambled accusative objects are Cb(Un) but do not become Cb(Un+1). To put it more concretely, in OACCSNOM/TOPV, speakers tend to shift the Cb from the referent of a scrambled object after that sentence. However, this account cannot be applied to OTOPSNOMV because the discourse function of topic marker *wa* competes with OSV word order. In fact, if OSV word order arises with a topic-marked subject, we cannot predict which referent (subject or object) will be referred to again in the following sentence.

Acknowledgements

We wish to acknowledge the contributions of everyone who has assisted us in our research. In particular, we wish to express our gratitude to the researchers at the Japanese Studies Department at the University of Oxford. Bjarke Frellesvig pointed out a drawback of our first analysis and his comments contributed to the general discussion. We also thank Stephen Wright Horn for a number of helpful suggestions.

References

Brennan, S. M., Friedman, M. W. & Pollard, C. J. (1987). A centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, 155-62.

Gordon, P. C., Grosz, B. J. & Gilliom, L. A. (1993). Pronouns, names, and centering of attention in discourse. Cognitive Science 17: 311-47.

Grosz, B., Joshi, A. K. & Weinstein, S. (1983). Providing a Unified Account of Definite Noun Phrases in Discourse. In Proceedings of the 21st Meeting of the Association of Computational Linguistics, 44-50.

Grosz, B., Joshi, A. K. & Weinstein, S. (1995). Centering: A Framework for Modelling the Local Coherence of Discourse. Computational Linguistics 21: 203-26.

Hadic, Z. L. & Taboada, M. (2006). Centering Theory in Spanish: Coding Manual. Unpublished manuscript, Simon Fraser University. Available from: http://www.sfu.ca/~mtaboada.

Hoffman, B. (1998). Word order, information structure, and centering in Turkish. In M. A. Walker, A. K. Joshi & E. E Prince (eds.), Centering Theory in Discourse, 251-71. Oxford, England: Oxford University Press.

Imamura, S. (2014). The Influence of givenness and heaviness on OSV in Japanese. In W. Aroonmanakun, P. Booonkwan & T. Supnithi, (eds.), Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, 224-233. Bangkok: Chulalongkon University.

Imamura, S. (2015). The effects of givenness and heaviness on VP-internal scrambling and VP-external scrambling in Japanese. Studies in Pragmatics 17: 1-16.

Imamura, S., Sato, Y. & Koizumi, M. (2014). Influence of Information Structure on Word Order Change and Topic Marker *WA* in Japanese. In W. Aroonmanakun, P. Booonkwan & T. Supnithi (eds.), Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, 432-441.

Bangkok: Chulalongkon University.

Imamura, S., Sato, Y. & Koizumi, M. (2016). The processing cost of scrambling and topicalization in Japanese. Frontiers in Psychology 7: 531. doi: 10. 3389/fpsyg. 2016. 00531

Joshi, A. K., Weinstein, S., 1981. Control of inference: Role of some aspects of discourse structure centering. In Proceedings of International Joint Conference on Artificial Intelligence, 385-7.

Kameyama, M., 1985. Zero anaphora: the case of Japanese. Doctorial dissertation, Stanford University.

Kameyama, M., 1998. Intrasentential Centering: a case study. In: M. A. Walker, A. K. Joshi, E. F. Prince. (eds.), Centering Theory in Discourse, 89–112. Oxford: Clarendon Press.

Kuno, S., 1978. Danwa-no Bunpō [Grammar of Discourse]. Tokyo: Taishūkan.

Miltsakaki, E., 2002. Toward an aposynthesis of topic continuity and intrasentential anaphora. Computational Linguistics 28, 319-55.

Okumura, M., Tamura, K., 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In Proceedings of the 16th Annual Meeting of the Association for Computational Linguistics, 871-6.

Rambow, O., 1993. Pragmatic aspects of scrambling and topicalization in German. In Workshop on Centering Theory in Naturally-Occurring Discourse. Philadelphia, PA, May 1993.

Saito, M., 1985 Some Asymmetries in Japanese and their Theoretical Implications. Doctoral dissertation, MIT.

Saito, M., 2009. Optional A-scrambling. Japanese/Korean Linguistics 16, 44-63.

Strube, M., Hahn, U., 1999. Functional centering: Grounding referential coherence in information structure. Computational Linguistics 25, 309-44.

Walker, M., Iida, M., Cote, S., 1994. Japanese discourse and the process of centering. Computational Linguistics 20, 193–232.