Resolving and Generating Zero Anaphora in Japanese: A Language Learning Aid Perspective

Mitsuko Yamura-Takei†, Miho Fujiwara‡ and Teruaki Aizawa§

†Graduate School of Information Sciences
Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, JAPAN
yamuram@nlp.its.hiroshima-cu.ac.jp

‡Department of Japanese and Chinese
Willamette University
900 State Street, Salem, OR, 97301, USA
Phone: 503-370-6015 / Fax: 503-375-5398
mfujiwar@willamette.edu

§Faculty of Information Sciences
Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, JAPAN
Phone: 82-830-1768 / Fax: 82-830-1768
aizawa@its.hiroshima-cu.ac.jp

Abstract

This paper presents our preliminary work in developing a Natural Language Processing (NLP)-enhanced Computer-Assisted Language Learning (CALL) program for Japanese learners that supports reading comprehension and writing revision process with focus on zero anaphora. We first discuss the pedagogical status of zero anaphora in a Japanese as a Second Language (JSL) learning environment, and argue the potential benefit of a zero anaphora understanding aid. We then propose the anaphora resolution/generation algorithms within the centering framework. We finally test the feasibility of centering and its algorithms on some empirical data from a language classroom. As a result of our data analysis, centering theory has proven to function as a pedagogically plausible language learning aid model in the tasks of resolving and generating zero anaphora.

Keywords: zero anaphora, natural language processing, computer-assisted language learning, centering theory

1 Introduction

Natural Language Processing (NLP) is a fascinating area of research and an emerging technology with a variety of real-world applications. Computer-Assisted Language Learning (CALL) is one such area that NLP techniques can be applied to. The language teaching and NLP communities, however, have tended to pursue their goals quite separately from one another, even though they both deal with natural human languages. Language teachers and NLP researchers could benefit from working more closely with each other in developing an NLP-inspired language learning aid.

Among a range of techniques that NLP can offer, anaphora resolution/generation is of particular importance to many NLP applications and has attracted much attention and extensive

research efforts, mainly in machine translation (MT) and information extraction. It is rather surprising that CALL has not yet seen much interest in anaphora. Zero anaphora is a hard nut to crack for learners but it has not drawn much attention in a Japanese as a Second Language (JSL) instruction, as discussed in Section 2.

Centering theory, originating as a computational model of discourse, has been served as one of the major anaphora resolution algorithms, due to its simplicity and computational tractability. The centering model is immensely language-independent; the model is modified for individual languages by setting its language-specific parameters. The parameters for Japanese were proposed by Kameyama (1985) and Walker *et al.* (1990, 1994). Since then, this Japanese model has been experimented with by many researchers, and has proven to be reasonably feasible.

In light of all this background, we attempt to build an NLP-enhanced CALL program for JSL learners that targets the acquisition of zero anaphora in the centering framework. The aims of this paper, as a preliminary work of our attempt, are: (1) to argue the potential benefit of an NLP-based zero anaphora understanding aid in a JSL learning environment; and (2) to attempt to prove the pedagogical feasibility of centering theory in the tasks of supporting reading comprehension and writing revision process, by providing empirical data directly from a language classroom.

In Section 2 we discuss a challenge that zero anaphora poses for JSL learners by presenting some language classroom data, and in Section 3 we overview the centering formalism and its modification for this study. In Section 4, we present the results of JSL data analysis and discuss the feasibility of the theory as a language learning aid model. Section 5 concludes with a discussion of possible future work.

2 Zero Anaphora for JSL Learners

In English, every argument of a verb must be overtly expressed in a sentence; otherwise the sentence becomes ungrammatical. Japanese, in contrast, allows such arguments to be freely omitted when they are recoverable from a given context or with relevant world knowledge. These unexpressed elements are called zero anaphora or zero pronouns (henceforth *zeros*).

This striking contrast poses a major challenge not only for Japanese-English MT developers but also for JSL learners who have English or another explicit argument language as their first language. Very few JSL textbooks, however, have a section addressing formal instruction and/or include intensive exercises on this ellipsis mechanism. Yet, *zeros* do exist even in very beginning level materials, not to mention in real-world authentic texts, as shown later in Section 4.1. As a result, many JSL teachers rely heavily on their intuitions about naturalness, rather than depending upon some systematic knowledge, when they explain *zeros*. Intuition is a conventional tool in teaching one's native language, but from a students' perspective, a well-developed systematic theory can be more convincing and more helpful.

This pedagogical contradiction is a trigger of our study for building an autonomous learning aid specially designed for understanding *zeros* based on a theoretically sound foundation, from which both teachers and students can benefit. Also, although there are many programs available that support reading by providing morphological and/or syntactic analyses with automatic dictionary look-up on demand,[1] discourse-level learning aids are not as prevalent. It is probably true that vocabulary plays a central role in both reading and writing. But in order to achieve more fluency and linguistic sophistication, it is also essential to explore discourse and coreference structure.

In what follows, we will present some empirical data from a JSL classroom to verify the claim that zero anaphora is one of the critical issues students face when learning Japanese.

2.1 Interpreting Zero Anaphora

There are voices from both teachers and learners that claim interpreting omissions is not an easy task. In order to verify this claim, ten upper-intermediate JSL students[2] were tested on a twelve-utterance narrative discourse that contains eight *zeros*. They were requested to translate the text into English specifying what each pronoun indicates.

[1]These include *Japanese Language Reading Tutorial System: Reading Tutor* at http://language.tiu.ac.jp and *Asunaro* at http://hinoki.ryu.titech.ac.jp.

[2] They are all native speakers of English.

Out of total 80 *zero* interpretations, only 46% of them turned out to be correct.[3] Some *zeros* were easier to resolve than others; success rates ranged from 90% to 10%. Also, some students performed better than others; their scores varied from 80% to 0%.[4] The result, despite these variants, was defective enough to denote the validity of the claim that *zeros* are hard to process for human learners.

2.2 Producing Zero Anaphora

For those whose first language does not permit sentence parts to be omitted, it would be a perplexing task to identify what contexts allow omission and which elements can be safely omitted.

We might naturally assume that learners tend to underuse *zeros* rather than overuse them, by using the strategy of avoidance, which often results in unnaturalness caused by the redundant use of full Noun Phrases (NPs). In order to examine this assumption, let us present an intermediate student's writing sample in (1).

(1)

1. nomin-ga itsumo tozoku-ni kome-ya okane-o nusumarete

   farmer-SUB often thief-OBJ2 rice-and money-OBJ be-robbed
   "The farmers were often robbed of rice and money by thieves, and"

2. **nomin-ga** tozoku-o taosu tame-ni

   farmer-SUB thief-OBJ beat in-order-that
   "in order that the farmers beat the thieves, "

3. 7-nin-no samurai-o yatoimashita.

   Ø-SUB 7 samurai-OBJ hired
   "Ø hired 7 samurais"

The subject in the second utterance (*nomin* "the farmers") would most plausibly be recommended by teachers to elide in order to perform more natural Japanese discourse.

As instantiated by this example (as well as other examples found in our data), it is often the case in a JSL classroom that such omissions are advised in order to avoid unnaturalness caused by redundancy.

3 Centering Model

In order to computer-assist in solving the problems presented in the previous section, we decided to use the centering model as our theoretical and computational framework.

This section presents the basic definitions and assumptions of centering discussed in the literature (e.g., Grosz *et al*., 1995; Walker *et al.*, 1998). We discuss some modifications we made for this study in Japanese. We later examine how this model fits into the language learning aid tasks, in both interpreting and producing *zeros*.

3.1 Overview of Centering

Centering theory is a computational model of discourse that models the interrelationships between focus, the choice of referring expressions, and perceived coherence of utterances.

A discourse segment consists of a sequence of utterances[5] U_1, …U_n. Each utterance U_i evokes a set of discourse entities, the FORWARD-LOOKING CENTERS, $Cf(U_i)$. The highest ranked entity in $Cf(U_{i-1})$ realized in U_i is the BACKWORD-LOOKING CENTER, $Cb(U_i)$. The highest ranked member of the Cf set is the PREFERRED CENTER, $Cp(U_i)$. The Cp is predicted to be Cb in the following utterance. The members of the Cf list are ranked as in (2), in case of Japanese (Walker *et al.*, 1994):

(2) Cf ranking for Japanese

(GRAMMATICAL OR ZERO) TOPIC > EMPATHY > SUBJECT > OBJECT2 > OBJECT > OTHERS

[3] Incorrect interpretations include the cases that *zeros* are not clearly specified by using passives or generic pronouns.

[4] Interestingly, their scores agree with their overall proficiency of Japanese.

[5] Following Kameyama (1998), Walker (1998) and Iida (1998), we defined utterance as a finite clause with one predicate (e.g. verb, adjective, *na*-adjective and copula), serving as a center updating unit.

Four types of transitions, reflecting the degree of local coherence, are defined, as shown in Table 1.

	$Cb(U_i) = Cb(U_{i-1})$ Or $Cb(U_{i-1}) = [?]$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	CONTINUE	SMOOTH-SHIFT
$Cb(U_i)\neq Cp(U_i)$	RETAIN	ROUGH-SHIFT

Table 1: Centering Transitions.

On top of these definitions, centering also includes the two rules in (3).

(3) Centering rules

Rule 1: If some element of $Cf(U_{i-1})$ is realized as a pronoun in U_i then so is $Cb(U_i)$.
Rule 2: Transition states are ordered. CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT.

Pronominalization (or zero pronominalization) is predicted to comply with Rule 1 that is called Pronoun Rule, on the assumption that discourse is coherent by following the preference ordering given in Rule 2. The Pronoun Rule and the choice of referring expressions over transition types play a significant role in the development of (zero) anaphora resolution and generation algorithms.

3.2 Global Center Model

Centering, as described in the previous section, is a theory of local focusing. The antecedents of *zeros* are assumed to be in the immediately preceding utterances. A study of naturally occurring texts, however, reveals abundant cases that contradict this assumption in various languages.

For example, the use of *zeros* in RETAIN and ROUGH-SHIFT transition states is not rare in Japanese, as data presented by Iida (1998) and Yamura-Takei *et al.* (2000) and our analysis of some JSL reading texts indicate. This is summarized in Table 2.

	CONTINUE	RETAIN	SMOOTH-SHIFT	ROUGH-SHIFT
Iida	76	3	34	23
Y-T	43	4	28	16
JSL	48	3	32	18

Table2: Distribution of centering transitions of utterances with *zeros* in three different texts.

These cases are further classified into two kinds: (1) the antecedent of *zero* (U_i) is an element other than $Cp(U_{i-1})$ in $Cf(U_{i-1})$ list; or (2) no appropriate antecedent can be found in $Cf(U_{i-1})$.

Out of 20 *zeros* in these two transition states observed in Yamura-Takei *et al.* (2000)'s data, for example, only two can be resolved locally, and the others calls for the global center structure in order to search for potential antecedents beyond the local boundary.

Iida (1998) introduced a global Cb list as an expanded search area. This list is comprised of former Cbs ordered by recency. The search in this list is activated when local search fails. We call this the GCB model in this paper.

Yamura-Takei *et al.* (2000) proposed Global Topic Hypothesis (GTH) in order to compensate for the GCB model in treating utterances with multiple *zeros*.[6] In GTH, each utterance is assigned one and only one Global Topic (GT) with a proximate grammatical topic (i.e., *wa*-marked NP). Global Transition: $GT(U_i) = Cp(U_i)$ is applied when one or more *zeros*, in an utterance with multiple pronouns, are retrieved globally. We call this model the GCB/GTH model.

Yamura-Takei *et al.* (2000) also introduced an alternative global structure: a Global Topic (GT) list that is a list of former topics (NPs marked by topic particle –*wa*) ordered by recency. This global structure was derived from Huang (1984)'s Topic NP Deletion that allows a topic of a sentence to be "deleted under identity with a topic of a preceding sentence" (p.549). We call this model the GT model.

Later, we will examine the performance of each of the following models; (a) Local model, (b) GCB model, (c) GCB/GTH model[7], and (d) GT model.

3.3 Centering Algorithms

In this section, we present anaphora resolution/generation algorithms based on the implications from the centering model.

[6] Multiple pronouns in an utterance disclose deficiency of centering that tracks down one center, as pointed out in Grosz and Sidner (1998).

[7] The model (c) is a GCB model enhanced by Global Topic Hypothesis. GTH is included in (d) GT model.

3.3.1 Resolution Algorithm

The core of the resolution algorithm is the standard centering algorithm which we call the Local model. The global search mechanism, either in a GCB or GT list, is activated when the local search fails. The algorithm is given in (4).

(4) Resolution Algorithm

1. Accept utterance U_i as input.
2. **If** U_i contains a *zero*, generate possible Cb-Cf combinations, i.e., $Cb(U_i)$ and a list of $Cf(U_i)$.
   Else, update center data structure by creating $Cf(U_i)$ and computing its transition state and continue on to U_{i+1}.
3. Filter the list by linguistic constraints:
   (1) contra-indexing, and
   (2) selectional restriction of the predicate.
4. Rank by transition state preference ordering.
5. Select $Cp(U_i)$ in a highest ranked transition as the antecedent.
6. **If** no appropriate antecedent can be found, go to global check:
   (1) Go through a global list.
   (2) Select the most recent Cb/GT that does not contradict the constraints.[8]
   (3) **If** no appropriate antecedent can be found, return NAF.[9]
   Else, continue on to U_{i+1}.

   Else, update data and continue on to U_{i+1}.

When NAF is returned, it would most probably be the cases that require relevant world knowledge or backward anaphora.[10]

3.3.2 Generation Algorithm

Centering theory has mainly developed as an algorithm for anaphora resolution. We assume that the principles underlying the constraints and rules of the theory can be inverted and facilitated as an anaphora generation operator. In an attempt to fit the principles into the generation task, we define the hypothesis in (5), simply assuming that *zeros* are used when they are locally recoverable, and then incorporate the hypothesis into the algorithm given in (6).

(5) Zero Generation Hypothesis

Cbs in the CONTINUE/SMOOTH-SHIFT transition states are zero-pronominalized; Cbs in the RETAIN/ROUGH-SHIFT transition states are not.

(6) Zero Generation Algorithm

1. Accept utterance U_i as input.
2. Create a set of $Cb(U_i)$ and $Cf(U_i)$ and compute its transition state.

   (If U_i contains a *zero*, go to the resolution algorithm and return with the retrieved entity.)

3. **If** $Cp(U_i)$ is not realized as *zero* in either the CONTINUE or SMOOTH-SHIFT transition state, where $Cb(U_i)=Cp(U_i)$, return "message"[11] and continue on to U_{i+1}.
   Else, continue.

4 Data Analysis and Discussion

In this section, we present an empirical study of naturally occurring texts in a JSL setting, and discuss the results from a language learning perspective.

4.1 Zero Anaphora in JSL Reading

We first examined occurrence of *zeros* in texts taken from JSL textbooks.[12] The texts are 16 multi-paragraphed narrative discourses, eight of which are randomly selected from Book 1 (beginning level) and another eight from Book 2 (lower-intermediate level).

The result is shown in Table 3. This reveals that very beginning level Japanese does exhibit *zeros*. Also, proportional occurrence of *zeros* increases as instructional level rises, from 17% to 34%.

[8] GTH is applied here, when applicable.

[9] No Antecedent Found

[10] Backward anaphora is not rare in Japanese, but it is outside the scope of this paper.

[11] See section 4.2 for details.

[12] "Hiroko-san no Tanoshii Nihongo 1 & 2" (1999, 1998), Bonjinsha, Tokyo.

	Utterances	Utterances with *zeros*
Book 1	103	18 (17%)
Book 2	231	83 (34%)
Total	334	101(30%)

Table 3: Occurrence of *zeros* in JSL texts.

Next, we hand-simulated the resolution algorithm described in 3.3.1 on the text and examine the distribution of centering transitions in a total of 101 utterances with *zeros*. The result is presented in Table 4.

	CONTINUE	RETAIN	SMOOTH-SHIFT	ROUGH-SHIFT
Book1	5	1	11	1
Book2	43	2	21	17
Total	48	3	32	18

Table 4: Distribution of centering transitions.

This distribution roughly matches that of data from newspaper articles given in Iida (1998) and data from short novels in Yamura-Takei *et al.* (2000) as presented in 3.2. Again, the use of *zeros* in RETAIN/ROUGH-SHIFT transition states is not rare. One *zero* in RETAIN and ten *zeros* in ROUGH-SHIFT, both from Book 2, call for the search in a global center list, but none from Book 1 does. In other words, all the *zeros* in Book 1 can be resolved locally. This provides us with a pedagogical suggestion that a local center model could be first introduced at the rudimentary stage of learning, followed by a global center model.

The result of using the models presented in Section 3.3 on this data is shown in Table 5.[13]

	Local	GCB	GCB /GTH	GT
Book 1	100%	100%	100%	100%
Book 2	82%	92%	93%	92%

Table 5: Success rates of the models.

All the *zeros* at a very beginning level (Book 1) were successfully retrieved by the local model alone. At a more advanced level (Book 2), however, the global models apparently outperformed the local model, although there were still some cases unsolved. The GCB/GTH model was slightly better than the other two global models since GTH saved utterances with multiple *zeros*.

4.2 Zero Anaphora in JSL Writing

In order to test the feasibility of our centering zero generation algorithm, two upper intermediate students' essays were manually analyzed, and their choice of anaphors were examined. The results are summarized in Table 6.

CONTINUE/SMOOTH-SHIFT		RETAIN/ROUGH-SHIFT	
ZERO	FULL NP	ZERO	FULL NP
27	6	3	23

Table 6: Choice of anaphors in transitions.

The six cases of full NPs in CONTINUE or SMOOTH-SHIFT transition state indicate the underuse of *zeros* according to our hypothesis in (5), and the algorithm returns "message" to suggest revision. The message would be that those NPs "may"[14] be omitted. The revised essays according to this suggestion successfully satisfied native speakers' intuitive judgment about naturalness. This implies an implemental validity of the algorithm output.

4.3 Discussion

This paper has highlighted some implications of centering theory for supporting language learning in respect to understanding *zeros* in Japanese. Studying our empirical data, centering has proven to function in a reasonably adequate way as a language learning aid model in the tasks of resolving and generating *zeros*.

As for the global center models, there was no single outstanding one among the three, as shown in Table 5.[15] There were two *zeros* that could be resolved by GT, but not by GCB. On the other hand, there were three *zeros* that could be retrieved by GCB, but not by GT. It is difficult to rank these two models. It should be noted, however, that out of four utterances with multiple *zeros*, three were dealt with locally, and the other one was successfully saved by GTH.

[13] This result is slightly better than that in Yamura-Takei *et al.* (2000), probably because of less complexity of our language data in this paper.

[14] The use/unuse of *zeros* is not a grammaticality issue, as in English, but rather an issue of discourse-level naturalness; therefore, omission is not a requirement but an option.

[15] GT model provides a most intuitively plausible center structure for a topic-prominent language like Japanese, though.

Still, there were two (non-cataphoric) *zeros* that could be resolved neither locally nor globally. The antecedent of one such *zero* was neither in a local Cf list nor in any global center list, but an entity in the previous discourse. This one negative example prescribes a new search area: a list of entities that are not in any center lists, as a last resort. The other case involves the process of filtering the candidates with linguistic constraints. Selectional restriction does not always succeed in eliminating inappropriate antecedents.

In the task of generation, the unuse of *zeros* in the CONTINUE/SMOOTH-SHIFT transition states could be served as a redundancy indicator in our data.

5 Future Work

This study can be seen as a preliminary stage of research into the development of NLP-enhanced CALL. In analyzing data, our study has brought our attention to some areas for several further investigations as the next stage of research.

First, a full study will require a larger empirical corpus from a variety of proficiency levels. For example, more JSL writing samples should be analyzed to assess the performance of the generation algorithm, by comparing the algorithm output with human JSL teachers' judgments.

Secondly, we should seek for a more explicit empirical evidence to determine which global model is the most applicable in the resolution task, from theoretical, computational and pedagogical perspectives. Also, we should further examine how we can incorporate the global model into the generation task. For example, the use of *zeros* in the RETAIN and ROUGH-SHIFT transitions (i.e., globally recoverable *zeros* or possible overuse of *zeros* that might cause ambiguity) is a critical issue in generation task as well as in interpretation task. Our data included three of such cases, as shown in Table 6. There needs to be criteria about which *zeros* in these transition states can be used without causing excessive inference load on the hearer/reader.

Last, in interpretation, it should be further investigated how to treat *zeros* that cannot be (correctly) resolved either locally or globally.[16] Some cases might be solved by expanding the search area or introducing finer semantic constraints within the framework, as mentioned in 4.3, while others should be treated in some pedagogically responsible way, probably outside the scope of centering. Insights from Nerbonne (in press) might lead to a solution. He states that the users (language learners) are intelligent in using the information provided even when the program cannot account for the full complexity of natural human languages. It might be a more practical and efficient way from a pedagogical perspective to list all the potential referential candidates from both local/global center lists and to let learners choose one utilizing their knowledge available and other relevant information that the program can offer. Effective collaboration between computers and humans, supplementing each other's abilities, would be a promising area for future research in building CALL systems.

Acknowledgments

We are grateful to all the JSL teachers and students who provided us with insights for our study. We would like to thank the NLP lab members at Hiroshima City University for helpful discussions. We would also like to thank the anonymous reviewers for their valuable comments.

References

Grosz, B.J., Joshi, A.K. and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21/2, 203-225.

Grosz, B.J. and Sidner, C.L. (1998). Lost intuitions and forgotten intentions. In M. Walker, A. Joshi & E. Prince (eds.), *Centering Theory in Discourse*. Oxford University Press.

Huang, C.-T.J. (1984). On the distribution and reference of empty pronouns. *Linguistics Inquiry* 15, 531-574.

Iida, M. (1998). Discourse coherence and shifting centers in Japanese texts. In M. Walker, A. Joshi & E. Prince (eds.), *Centering Theory in Discourse*. Oxford University Press.

Kameyama, M. (1985). *Zero Anaphora: The Case of Japanese.* Ph.D dissertation, Stanford University.

[16] This is when NAF (no antecedent found) is returned in the resolution algorithm in 3.3.1 or when centering makes incorrect predictions.

Kameyama, M. (1998). Intrasentential centering: A case study. In M. Walker, A. Joshi & E. Prince (eds.), *Centering Theory in Discourse*. Oxford University Press.

Nerbonne, J. (in press). Computer-assisted language learning and natural language processing. To appear in R. Mitkov (ed.), *Handbook of Computational Linguistics.* Oxford University Press, 2001. Available at http://odur.let.rug.nl/~nerbonne/paper.html.

Walker, M. (1998). Centering, anaphora resolution, and discourse structure. In M. Walker, A. Joshi & E. Prince (eds.), *Centering Theory in Discourse*. Oxford University Press.

Walker, M., Iida, M. and Cote, S. (1990). Centering in Japanese discourse. In *Proceedings of the Thirteenth International Conference on Computational Linguistics (COLING-90)*, 1-6.

Walker, M., Iida, M. and Cote, S. (1994). Japanese discourse and the process of centering. *Computational Linguistics* 20/2, 193-232.

Yamura-Takei, M., Takata, M. and Aizawa, T. (2000). The role of global topic in Japanese zero anaphora resolution (in Japanese). In *IPSJ SIG Notes, Natural Language* No.135, 71-78.