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/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 Theory 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 plausible language learning aid model in the tasks of resolving and generating zero anaphora.

Keywords: zero anaphora, NLP, CALL, 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 so far seen much interest in anaphora, considering that zero
anaphora is a hard nut to crack for learners that has not drawn much attention in a Japanese as a Second Language (JSL) instruction, as discussed in Section 2.

Centering Theory, originating as a discourse structure model, 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; an individual language is modified by setting its language-specific parameters within the model. 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 NLP-based zero anaphora understanding aid in a JSL learning environment; and (2) to attempt to prove the feasibility of Centering Theory in the task 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 posits 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, and Section 5 concludes with a discussion of possible future work.

2 Zero Anaphora for JSL Learners

In English, all the arguments that the verb subcategorizes for are required to be expressed in a sentence; otherwise the sentence becomes ungrammatical. Japanese, in comparison, does allow such arguments to be freely omitted when they are recoverable from a given context or relevant knowledge. These unexpressed elements are called zero anaphora or zero pronouns (henceforth zeros).

This striking contrast posits a major challenge not only for Japanese-English MT developers but also for JSL learners who, in particular, have English or other explicit languages as their first languages. However, very few JSL textbooks spare their chapters or sections for the formal and organized instruction and/or 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 intuition about naturalness, rather than some systematic knowledge, when they explain zeros. Intuition is a conventional tool in teaching one’s native language, but from a students’ perspective, a systematically well-developed theory can sometimes be more convincing.

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 quite a few programs available that support reading by providing morphological and/or syntactic analysis with automatic dictionary look-up on demand, discourse-level learning aids are not as prevalent. It is probably true that vocabulary plays a central role in reading and writing, but in order to achieve more fluency and linguistic sophistication, exploring discourse and coreference structure would be essential.

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 in learning Japanese.

2.1 Interpreting Zero Anaphora

There are voices both from teachers and learners that claim interpreting omissions is not an easy task. In order to verify this claim, 10 upper-intermediate JSL students were tested on a 12-utterance narrative discourse that contains 8 zeros, by letting them translate the text into English specifying what each pronoun indicates.

Out of total 80 zero interpretations, only 46% of them turned out to be correct. Some zeros

---

1These 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.
3 Incorrect interpretations include the cases that zeros are not clearly specified by using passives or generic pronouns.
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 also be a perplexing task to identify which elements can be safely omitted and what contexts allow omission.

It would be easily assumed 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 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 yatoimasita.
   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 single example among many others found in our data, it is often the case in a JSL classroom that such omission is advised 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.

In this section, we present the basic definitions and assumptions in Centering discussed in the literature (e.g., Walker et al., 1998), including some modifications we made for this study in Japanese, and later examine how this model could fit into the language learning aid task, both in interpreting and producing zeros.

3.1 Overview of Centering

Centering Theory is a theory of discourse structure 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\(^5\) \(U_1, \ldots, U_n\). Each utterance \(U_i\) evokes a set of discourse entities, the FORWARD-LOOKING CENTERS, \(Cf(U_i)\). The highest ranked entity in \(Cf(U_i)\) 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

\[
\text{(GRAMMATICAL OR ZERO) TOPIC} > \text{EMPATHY} > \text{SUBJECT} > \text{OBJECT2} > \text{OBJECT} > \text{OTHERS}
\]

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

| \(Cb(U_i) = Cb(U_{i-1})\) or \(Cb(U_{i-1}) = \text{Ø}\) | CONTINUE |
| \(Cb(U_i) ≠ Cb(U_{i-1})\) | SMOOTH-SHIFT |
| \(Cb(U_i) ≠ Cb(U_{i-1})\) | RETAIN |
| \(Cb(U_i) ≠ Cb(U_{i-1})\) | ROUGH-SHIFT |

Table 1: Centering Transitions

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

\(^4\) 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), serving as a center updating unit.
(3) Centering rules

**Rule 1**: If some element of Cf(U\textsubscript{i-1}) is realized as a pronoun in U\textsubscript{i}, then so is Cb(U\textsubscript{i}).

**Rule 2**: Transition states are ordered. CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT.

Pronominalisation (or zero pronominalisation) 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, which assumes that the antecedents of zeros are in the immediately preceding utterances. A study of naturally-occurring texts, however, reveals abundant cases that contradict this assumption.

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, as summarized in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>CONTINUE</th>
<th>RETAIN</th>
<th>SMOOTH-SHIFT</th>
<th>ROUGH-SHIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iida</td>
<td>76</td>
<td>3</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>Y-T</td>
<td>43</td>
<td>4</td>
<td>28</td>
<td>16</td>
</tr>
<tr>
<td>JSL</td>
<td>48</td>
<td>3</td>
<td>32</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 2: 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\textsubscript{i}) is other element than Cp(U\textsubscript{i-1}) in Cf(U\textsubscript{i-1}) list, or (2) no appropriate antecedent can be found in Cf(U\textsubscript{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 model in order to search for potential antecedents beyond the local boundary.

Iida (1998) introduced a global Cb list that is a list of former Cbs ordered by recency, as an expanded search area. We call this the GCB model in this paper.

Yamura-Takei *et al.* (2000) introduced Global Topic Hypothesis (GTH) in order to compensate the GCB model in treating utterances with multiple zeros. In GTH, each utterance is assigned one and only Global Topic (GT) with a proximate grammatical topic. Global Transition: GT(U\textsubscript{i}) = Cp (U\textsubscript{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 (marked by topic marker –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, and (d) GT model.

### 3.3 Centering Algorithms

In this section, we present anaphora resolution/generation algorithms based on the implications from Centering 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\textsubscript{i} as input.
2. If U\textsubscript{i} contains zero pronoun, generate possible Cb-Cf combinations, i.e., Cb(U\textsubscript{i}) and a list of Cf(U\textsubscript{i}).
   
   **Else**, update center data structure by creating Cf(U\textsubscript{i}) and computing its transition state and continue on U\textsubscript{i+1}.

---

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. 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.  
   (3) If no appropriate antecedent can be found, return NAF.  
       Else, continue on U_{i+1}.
   Else, update data and continue on U_{i+1}.

When NAF is returned, it would most probably be the case that requires relevant world knowledge or backward anaphora.

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) 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, return “message”  
   and continue on 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 a JSL textbook. The texts are 16 multi-paragraphed narrative discourses, 8 of which are randomly selected from Book 1 (beginning level) and another 8 from Book 2 (lower-intermediate level).

The result is shown in Table 3, which reveals very beginning level Japanese does exhibit zeros. Also, proportional occurrence of zeros increases as instructional level rises.

<table>
<thead>
<tr>
<th></th>
<th>Utterances</th>
<th>Utterances with zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book 1</td>
<td>103</td>
<td>18 (17%)</td>
</tr>
<tr>
<td>Book 2</td>
<td>231</td>
<td>83 (34%)</td>
</tr>
<tr>
<td>Total</td>
<td>334</td>
<td>101 (30%)</td>
</tr>
</tbody>
</table>

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, as presented in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>CONTINUE</th>
<th>RETAIN</th>
<th>SMOOTH-SHIFT</th>
<th>ROUGH-SHIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book 1</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Book 2</td>
<td>43</td>
<td>2</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
<td>3</td>
<td>32</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4: Distribution of centering transitions

This distribution roughly matches that of data from newspaper articles given in Iida (1998) and

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.
data from short novels in Yamura-Takei et al. (2000) as we mentioned earlier. Again, the use of zeros in RETAIN/ROUGH-SHIFT transition states is not rare. One zero in RETAIN and 10 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 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

<table>
<thead>
<tr>
<th></th>
<th>Local</th>
<th>GCB</th>
<th>GCB/GTH</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book 1</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Book 2</td>
<td>82%</td>
<td>92%</td>
<td>93%</td>
<td>92%</td>
</tr>
</tbody>
</table>

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 Centering zero anaphora generation algorithm, two upper-intermediate students’ essays were manually analyzed, and their choice of anaphors were examined, as summarized in Table 6.

<table>
<thead>
<tr>
<th>CONTINUE/SMOOTH-SHIFT</th>
<th>RETAIN/ROUGH-SHIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZERO</td>
<td>FULL NP</td>
</tr>
<tr>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>ZERO</td>
<td>FULL NP</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 6: Choice of anaphors in transitions

The use of full noun phrases (NPs) in 6 CONTINUE or SMOOTH-SHIFT transition states indicates 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” be omitted.

4.3 Discussion

This paper has highlighted some implications of Centering Theory for supporting language learning in respect to understanding anaphora. Studying our empirical data, Centering Theory has proven to function in a reasonably adequate way as a language learning aid model in the tasks of resolving and generating zero anaphora.

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 could 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 4 utterances with multiple zeros, 3 were dealt locally, and the other one was successfully saved by GTH.

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 list. The other case involves the process of filtering the candidates with linguistic constraints. Selectional restriction does not always succeed in excluding 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 our data, our study prescribes a path for several further investigations as the next stage of research.

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, 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.
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 algorithm output with human JSL teachers’ judgment.

Secondly, we should seek for a more explicit empirical evidence to determine which global model is the most applicable in the resolution task, from a theoretical, computational and pedagogical perspective. 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 3 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.

Last, in interpretation, it should be further investigated how to treat zeros that cannot be resolved either locally or globally. Some cases might be solved by expanding the search area 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) who notes 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, might lead to a solution. Listing all the potential referential candidates from both local/global center lists and letting learners to choose one utilizing their knowledge available, might be a more practical and efficient way from a pedagogical perspective. 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 students who provided us with their writing samples full of 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

---

16 This is when NAF (no antecedent found) is returned in the resolution algorithm in 3.3.1.