様式 C-19

科学研究費補助金研究成果報告書

平成２１年６月２９日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19720144
研究課題名（和文）　否定的なフィードバックメカニズムとアウトプット仮説の実証的研究

研究課題名（英文）　Noticing the gap, hypothesis testing, and the uptake of subsequent feedback

研究代表者
Ｓｈｅｐｐａｒｄ　Ｃｈｒｉｓ（SHEPPARD Chris）
早稲田大学・理工学術院・准教授
研究者番号：60350386

研究成果の概要：
1) Second language learners notice gaps in their linguistic knowledge and test hypotheses about language during oral production.
2) These gaps in knowledge orient language learners to information related to both their noticed gaps and their tested hypotheses about the language.
3) Information related to their noticed gaps and their tested hypotheses about the language is more likely to be incorporated into their subsequent output, perhaps resulting in learning.
4) Feedback is useful to learners when they have experienced a problem with oral production processes and are oriented towards feedback related to their problems.

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	600,000	0	600,000
２００８年度	300,000	90,000	390,000
年度			
年度			
年度			
総　計	900,000	90,000	990,000

研究分野：人文学
科研費の分科・細目：言語学・外国語教育
キーワード：第二言語習得、フィードバック、アウトプット仮説、再生刺激法、言語教育

１．研究開始当初の背景

Feedback, noticing the gap and hypothesis testing

The utility of negative evidence in second language acquisition (and first language acquisition) has been debated in the literature for the last thirty years. Proponents of Chomsky's universal grammar stated that all one needed to learn a first language was positive evidence. Brown and Hanlon (1970) supported this view by demonstrating that explicit negative feedback was not present in child-adult interaction.

Krashen, (1982, 1985) proposed that second language and first language acquisition both use the same mechanisms. He claimed that because positive evidence was sufficient, feedback would not help in the acquisition of implicit knowledge. He even went so far as to say that feedback could damage the learning process as it could 'raise the affective filter' by de-motivating learners. His position was supported by Carroll (2001), Truscott (1996), and White (1993).

However, feedback has been shown to assist in the learning of a second language and there have been similar findings in first language acquisition research. In a meta-analysis of 15 feedback studies, Russell and Spada (2006) found that, on average, the studies had an effect size of d = 1.16, and a range from d = 0.15 to d = 2.18. According to Cohen (1962), an effect size of greater than 0.8 is thought to be large.

Long (1996, 2007) has suggested that recasts are perhaps the best form of feedback from a communicative point of view, as they provide a brief opportunity to focus on form without disturbing the overall communicative purpose of interaction. He defines recasts as "a reformulation of all or part of a learner's immediately preceding utterance in which one or more non-target-like items is/are replaced by the correct target language form(s)" (2007, p. 77).

Research has demonstrated, however, that recasts are not always effective. Lyster and Ranta (1997) found that, while recasts were the most common form of feedback in communicative language classrooms, the learners 'uptook' them the least. In a review article Ellis and Sheen (2006) questioned the 'acquisitional value of recasts in comparison to other forms of feedback' (p. 575). Lyster (2004) found that recasts were not as effective as prompts, which indicate errors to learners but do not correct them. Sheen (2010) found that her oral recast group did not perform better than the control in the acquisition of articles.

Other studies have shown a positive effect for recasts (Doughty and Varela, 1998, and Philp, (2003). Doughty and Varela provided regular recasts and other feedback over the period of several weeks and demonstrated improvement. Philp demonstrated that her participants noticed over 60-70% of the recasts supplied to them.

In order to explain the difference in these results, Long (2003) has suggested that for recasts to be effective, they need to be perceptually salient. The success of recasts used in Doughty and Varela's (1998) study indicates that the learners were able to identify the corrections. Leeman (2003) compared recasts, which she claimed contained both positive and negative evidence, enhanced salience of positive evidence and unenhanced positive evidence. Only the recast enhanced salience groups improved, which suggests that the utility of recasts is, at least in part, determined by their salience. Likewise, Sheen (2010) suggested that it was the lack of salience of the recasts in her study which was responsible for the recast groups lack of improvement.

Thus, to date the research has demonstrated that for second language learners to be able to use recasts to improve their interlanguage, they need to know that they are being corrected and the feedback needs to be salient. However, there has been little focus on how the learner interacts with recasts in the spoken input.

Swain's (1995) output hypothesis suggests four ways in which oral output could have a role in the acquisition of a language. Two of those ways were 'noticing the gap' and hypothesis testing. Both of these functions can be understood through Levelt's (1989) speech model which de Bot (1996) applied to second language acquisition.

Noticing the gap occurs when there is a

problem in the formulator (where sentences are formed in the brain) and a concept cannot be represented. At this point the problem will become conscious and a strategy to deal with it will become necessary. Once the message is formulated, Level's model allows for monitoring it. This monitoring could take place before the message is articulated, or after. During monitoring, the speaker could again 'notice the gap'.

Once speakers are aware of the gap, they have several options, which Oxford (1995) terms 'compensation strategies'. She suggests that the learner could avoid communication altogether or abandon the topic. Another option is to reformulate the message. This reformulation could take place by using explicit knowledge to attempt to fill the gap. Another option is to ignore the problem completely and produce the output as is. Swain's 'hypothesis testing' happens when the speaker produces the utterance when they are aware of a gap in their knowledge.

de Bot (1996) suggested that when learners notice a gap, they could become aware of a problem, which in turn would lead to more attention to information which could fill the gap in subsequent input.

２．研究の目的

This report presents research which examined the roles of noticing gaps and hypothesis testing in the uptake of recasts in subsequent input. It asked the question "do learners incorporate information in input related to their awareness of linguistic problems in their output?" Swain's (1995) Output Hypothesis predicts that learners will incorporate information related to gap awareness at a rate greater than the rate of 'overall incorporation.'

３．研究の方法

To answer this question, two studies were undertaken. First, the design of the First study is described, and then the participants. This is followed by a description of the instruments and procedures used to elicit the data. Finally, there is a discussion of how the data was analyzed.

This study used a factorial repeated-measures experimental design. It was selected as it was deemed to be the most effective way to answer the research questions. The experiment involved four groups; a control and three treatment groups. Pretests were conducted to provide a base-line from which to compare improvement.

All four groups completed a pre-test, and three post tests All the tests involved producing oral narratives based on picture compositions. The Control Group completed the tests only. The Stimulated Recall Group completed all the tests and in addition were given the opportunity to review their pre-test performance to identify problems in their linguistic knowledge. The Input Group completed all the tests and also a 'stimulated recall' procedure. In addition, its members listened to a native-speaker version of the narrative they had produced in the pre-test. The 'Repair' Group also completed all the tests as well as the stimulated recall procedure. However, unlike the other experimental groups, when problems were identified by the learners, repair was provided by the researcher. The overall design is summarized in Figure 1

Participants

The eighty-one participants in the experiment were drawn from optional English language courses held in an institute attached to a large private university in Japan. The participants volunteered their services, after receiving an explanation of the experiment and what was

required of them, then signed a permission form. They were offered 2000 yen in return for their services.

Japanese university students are false beginners, having had six years of compulsory English language education, at both junior and senior high school. The majority of the courses offered at high school are based on a grammar translation method, known in Japan as *yakudoku*. In addition, most of the university's departments include English in the first and sometimes the second year as part of their general curriculum requirements. For the most part, these classes are a continuation of high school. However, as teaching methodology is usually left to the teachers, there are some who focus on English for communication.

The participants were randomly assigned to one of four groups. This was done by lottery. Upon arrival at the research site, they were required to select a folded slip of paper with one of the group names written on it. They were then placed in that group. There were 20 students in each of the Control, Stimulated Recall and Input Groups and 21 students in the Repair Group.

Instruments

In this section the instruments used in the study are described in detail. The instruments consisted of a narrative task that doubled as a test (in the pretest, and posttests one and two) and as material for the experimental treatments. A second narrative task was used in post-test three. In addition, there was a native speaker version of the first narrative.

The first narrative task, which learners performed as the pre-test, also served as the basis for the experimental treatments.

The dimensions of a task define the difficulty students will have, and thus predict performance. Therefore, it is important to describe the task accurately. Skehan (1998) suggests three criteria for defining a task; code complexity, cognitive complexity (including cognitive familiarity and cognitive processing) and communicative stress. Robinson (2001b) suggests that the characteristics of the task, the learner's relationship with the task (task difficulty), and the conditions under which the task is performed constitute a more useful task classification. Ellis (2003) combines several classification frameworks (pedagogic, rhetorical, cognitive and psycholinguistic) to produce a general framework for the classification of tasks. This framework covers Robinson's task characteristics and conditions. In his framework the design features are the task input, conditions, processes, and outcomes. These features were used to describe and evaluate the tasks used in this experiment.

The medium of input for the narrative was pictorial. Participants were required to produce a story from cartoon pictures. The ten-picture 'story' contained four characters, three of them belonging to a family. The overall structure of the story was defined by the pictures. However, the participants needed to structure the information in each picture and determine the links between them. Because it was visual this information was static, and concrete, so it was a 'here and now' context. The topic of the narrative was a boyfriend visiting his girlfriend's home unannounced and being invited for dinner. The family dinner would be a relatively familiar situation for these Japanese participants. However, given that a boyfriend does not usually visit the family's home, especially unannounced, this situation may have been an unusual one.

One of the conditions of this task was that the interlocutor relationship was one-way, with the participant doing all the speaking. The task was a single one, only requiring the creation of a story by

the participant during which the pictures were put in order by the researcher. Unlimited time was given to complete the task. The level of control of the interaction was limited.

The second and third aspects to consider in describing this task were the cognitive processes involved and the outcomes from the completion of the task. As participants create the narrative they determine the contextual connections and the relationships between the picture frames that is, they cognitively process the information.. There are two outcomes. The first is the narrative itself. The second is a set of ordered pictures based on the narrative.

Most of the task characteristics and conditions described above conform to those Ellis (2003) predicted would make a task easy to perform. However, he also hypothesized that one-way flow of information, monologic discourse mode, oral output and the open nature of the output would increase difficulty. In spite of the presence of these negative conditions in the research tasks g it was hoped that they would be easy enough to allow the participants to focus their attention on formulation, not just on the content.

The second task was designed to be as close to the first task as possible in all respects. This task, used in the third post-test, was a ten-frame picture story administered under the same conditions as the first task. Only the content of the pictures was different. Although the general topic, having dinner, was the same, the situation was at a restaurant and among friends rather than at home with the family. There were also four characters. One difference between the tasks is that in the first task, it was relatively simple to label the characters according to family roles. In the second task, labeling was not as simple as the characters were given no names. Thus, to distinguish the four characters it was necessary to identify them linguistically through modification and relative clauses, although a few participants did attempt to invent names for the characters.

Task equivalency was confirmed for fluency and accuracy . However, contrary to predictions, the first narrative task elicited performances which were significantly more complex on average than those of the second narrative, (Cohen's D = 0.68).

A version of the narrative for the first task was scripted and then recorded by the researcher at a medium pace (appendix x). There were 264 pruned words and the task took 3 minutes 11 seconds to complete. The pace of speech was measured at 101.2 syllables per minute. There were no errors (100% Error free t-units per t-unit). The complexity of the passage was recorded at 2.36 phrases per t-unit. As this was a scripted passage, there were no fillers used, or reformulations. The most frequent 2000 words accounted for a little over 85% of the text as shown in Table 2. Also the counting program Vocabulary LC, used to determine the appropriateness of the vocabulary level of the text for Japanese high school students, showed that the text was most appropriate for second year high school students, demonstrating that the vocabulary in the narrative would not challenge the learners, hopefully allowing them to focus their attention on noticing linguistic forms rather than comprehension.

Procedures

This section describes the procedures used in the study. The experimental procedures for each of the four experimental groups: the Control Group, the Stimulated Recall Group, the Input Group, and the Repair Group are explained first, and then this is followed by a description of the testing procedures. These include the pretests and the three posttests. Finally, the procedures used in the working memory tests are explained. The

procedures took place over two sessions, with the pretest, the first posttest and the treatment taking place in the first session, which took between 60 and 80 minutes, depending on the participant. The second session, which included the second and third posttest and the working memory tests, took just under an hour.

The Control Group in this experiment took part in an English conversation equivalent in length to the stimulated recall procedure described below. It should be noted, as this procedure was conducted in English, learners in the Control Group received more oral English input than the other experimental groups. This conversation was on average 28.54 minutes (s = 1.96 minutes.) long.

In this section, a rationale for the choice of the stimulated recall procedure is provided and then an outline of the procedure followed with the Simulated Recall Group is given..

Stimulated recall is a technique which attempts to determine conscious cognitive processes through retrospection. Jourdenais (2001) reports that the use of introspective techniques in language acquisition research is fairly common. It is found in many areas including writing, reading, language use, interpretation, discourse, and, most recently, research on attention and awareness. Research using think-aloud protocols has been particularly useful to demonstrate a role for noticing in language acquisition (Alanen 1995, Journdais et al. 1995, Leow 1998a, 1998b, 2000, Rosa and O'Neill 1999, Rosa and Leow 2004a, 2004b, also see Færch and Kasper 1987). In these studies, even when in many cases the treatments were not particularly effective in developing the learners' interlanguage, this methodology demonstrated a connection between noticing and subsequent acquisition. So using introspection as a technique, would appear to be a useful tool to investigate if and how learners become aware of their gaps in knowledge through monitoring their language production. The following section evaluates the validity of using this technique in this experiment.

Introspective techniques have been criticized for several reasons. It has been suggested that only conscious processes are accessible to verbal report by participants, that the process of giving verbal report may interact with the learning system of the learners, that processing constraints are limiting, and finally that modality introduces limitations. In a re-examination of a series of studies using retrospective reports, Nisbett and DeCamp-Wilson (1977) showed that attempts to elicit participants' explanations for their actions directly after they produced those actions were invalid. Based on these results and others, Ericsson and Simon (1980) suggest that for verbal reports to be valid, they should only be focused on the description of conscious thought processes and not allow the reporter to make inferences in any way about their activity.

A second problem with verbal reports is that they could be what Leow and Morgan-Short (2004) describe as reactive. This is where the act of observation is interfering with the system which is being measured. Although this paradox (the act of system observation in effect alters the system in some way, for example Schrodinger's (1935) cat) is impossible to avoid whatever the measurement technique, introspection is particularly problematic as it requires the participant to observe their own behavior. However, Adams (2003) found that her use of stimulated recall did have an effect. In her examination of the role of feedback, she asked her participants to produce a written narrative from a series of eight pictures. The Control Group completed this task only. Next she had a noticing session where learners were provided with a reformulated version of their

narrative and asked to verbalize their ideas about the differences between the two versions. This constituted the noticing condition. Finally, directly after the noticing session, she conducted stimulated recall sessions where learners described what they were thinking at the time. Findings showed that the noticing group and the noticing + stimulated recall group produced more target like instances of grammatical features in their post-test narrative. After removing the control, she found that the stimulated recall group also produced more target like output. She concluded this provided evidence of a reactive role for stimulated recall in language learning.

This conclusion is different to that of Leow and Morgan-Short who, by comparing two groups completing the same task either with or without think-aloud protocols, found that 'reactivity does not play a significant role in learners' subsequent performances'. The difference between the stimulated recall and the think-aloud is possibly due to processing limitations. One of the criticisms of think aloud is that it may require participants to use attentional capacity which is thus detracted from performing the task (Jourdenais 2001). With no spare capacity, the participants have less opportunity to use the verbalization processes to assist with their noticing. However when this verbalization takes place retrospectively, as in stimulated recall, the participant has additional attentional capacity to focus on noticing items (Ericsson and Simon 1980). Thus, it appears that think-aloud may be superior superior technique when compared to stimulated recall which does not allow 'additional' noticing, and possibly learning. On the other hand, stimulated recall may be more useful as it is less intrusive on the processes it is attempting to measure.

The research reported in this report deals with the oral modality and it is nearly impossible to use think-aloud protocols while producing spoken narratives. It was decided, therefore, to use stimulated recall to determine what problems learners became aware of during production. The task which was the object of the procedure was different to that of Adams (2003). Whereas she asked learners to recall their thought processes during a noticing session, this study examined learners' awareness of gaps in linguistic knowledge resulting from their oral narrative production.

The third criticism that can be directed at stimulated recall in particular is that, given the limited capacity of the working memory (a maximum of 20 seconds, Doughty, 2001; Cowan, 1999), stimulated recall would usually take place at a time long after the memory trace had decayed. This perhaps would lead learners to create inferences rather than report what they were thinking at the time (Nisbett and DeCamp-Wilson, 1977).

Gass and Mackey (2000) suggest that this problem can be avoided by using some kind of stimulus to reactivate the memory trace which they predicted would decay exponentially. So, although learners have verbal access to the content of their working memories for only 20 seconds, Gass and Mackey suggest that with a stimulus such as a video replay of their performance to reactivate the decaying trace, it is still possible for the learners to recover their thought processes up to 24 hours after the activity.

Mackey (2002) examined the validity of this kind of stimulated recall. She coded the participants' interaction data for comprehensible input, feedback, pushed output and hypothesis testing in three different contexts (classroom interaction, NS (Native speaker) -NNS Non –native –speaker) interaction, and NNS-NNS interaction), and then compared her observations

with the participants' stimulated recall reports. She found that, on the whole, agreement between the two measurement methods was between 70% and 80% in two of the interaction contexts (classroom interaction and NS-NNS interaction). However, the third context (NNS-NNS interaction) produced agreements of about 60%, and below 50% for feedback. This was due to a lower orientation to correction by an NNS interlocutor. These results provide evidence that stimulated recall is a valid technique to measure the cognitive processes taking place. Clearly more research needs to be conducted to determine the validity of the use of this instrument.

To sum up, there has been limited research on the use of stimulated recall as an introspective technique. It does enable the researcher to investigate both speaking and listening, which think-aloud does not due to its physical limitations. Mackey (2002) provided some evidence that it is valid. Adams (2003) demonstrated the process is reactive, arguing that it assisted learners to notice more and therefore produce a more accurate subsequent performance.

In order to determine the extent of the threat to validity in this research, Group Two completed the stimulated recall session and took part in no other treatments.

The procedure was as follows. Once the pretest was completed the participants in the experimental groups were instructed to turn to the next page in the booklet and read the instructions for the stimulated recall protocols. Verbal confirmation of understanding was made, and then participants were asked if they would like the recall procedure to take place in their native language (Japanese), or in English. Of the 81 participants, two indicated that they would prefer English. This procedure was carried out by the researcher, who is a second language speaker of Japanese, but a native-speaker of English.

The video tape of the participants' pre-test production was then replayed. The participants were also asked to turn back to the picture story. It was hoped that these two aids would provide the contextualization required to recall what conscious thought processes had taken place during production. During the viewing of the video, the researcher paused the tape at a hiatus in speech, a reformulation or a repetition and told the participant 'You said '…', and then asked, "what were you thinking at this time?" In addition to this, the participants were also told that if anything came to mind during the procedure they were to speak out. If the learner said anything the video was paused. The tape was also replayed whenever it was necessary. Examples of the stimulated recall session are given below.

Example 1: Stimulated recall 1

Trigger:	They are, they, ah, next picture, they look so happy
R：	they are と言った後、あ、next pictureと言った。なにを考えていましたか？
M2(2)：	私の中では次の絵に移ったけれども、いきなりthey areと言ったら、聞いている人が同じ絵の中で動いていると思ってしまうかなあと思って、一応区切った
[R:	First you said 'they are', and then after that you said 'next picture'. What were you thinking?]
[M2(2):	I was already thinking about the next picture, but I thought that if I suddenly said 'they are', then the person listening would think that I was talking about

the same picture, so I changed what I was saying.]

(Translated by the researcher)

Example 2: Stimulated recall 2

Trigger: Ah, Second picture, ah, the man, ah, asked the father, ah, to, ah, where Kate is

R：a skの後にポーズがあった。その時、なにを考えていましたか？

T1(2)： Edが「Kateがどこにいるのか」と聞いているというのをまず日本語で考えて、それを英語に置き換えようとしてポーズができた。

[R: There was a pause after 'ask'. What were you thinking at that time?

[T1(2): I first made the sentence 'Ed asked where Kate was' in Japanese. Then during the pause I was changing the sentence into English.]

(Translated by the researcher)

On average the stimulated recall took 25.71 minutes with a large standard deviation of 7.11 minutes. Correlations showed that this variance could be accounted for, in part, by the length of the pretest passage ($r = .473$, $p < .01$). That is to say the longer the narrative produced by the participant, the longer the stimulated recall session.

The Input Group also completed the stimulated recall, following exactly the same procedure as the Stimulated Recall Group. In addition, the Input Group was given the opportunity to listen to the native speaker version of the pretest narrative task they had performed. It was hoped to determine if noticing gaps in knowledge under these conditions would result in an increased orientation to those items present in the input. For this group, the native English speaker's version was played on a tape recorder to ensure the conditions were equivalent for all participants in this group. They were asked to take notes of any items which they thought were important and these were collected in as a record of their noticing.

Given that the on-line task of listening and note-taking is demanding, a five-minute retrospective interview was also conducted to determine any other items the learner might have noticed. This interview was also used to confirm that the opportunity to take notes had allowed a fair representation of what the learners had noticed.

One difficulty of measuring noticing is that we are not able to determine if we have measured everything that the learners have noticed. In Reber's (et al.) experiments, from which he concluded that learning is unconscious, the measurement of noticing was incomplete. This made it difficult for him to make strong claims that noticing was the necessary and sufficient condition for language learning. Our operationalization is subject to the same weakness. Therefore, rather than attempting to determine whether or not 'noticing' is necessary and sufficient for language acquisition, this research only attempted to determine if there was a link between noticing and the awareness of linguistic problems.

Another weakness of noticing measures is that they can also over-determine the degree of noticing. A possible example of this is underlining, used by Izumi et al. (1999). Here it is possible learners will underline more than they actually notice, as the measure does not represent

information which has been actually processed. The requirement to take notes constitutes a measure of the participants processing of information, and thus this particular problem was avoided in this study. (Also, underlining is less appropriate in an auditory context, even in associated note-taking.)

Like the Input Group, the Repair Group also completed the stimulated recall. What distinguished this group from the Input Group was that, instead of listening to the native speaker version of the narrative, they received feedback on any linguistic problem they drew attention to as they performed the stimulated recall. 3 and 4 provide examples of 'repair' after noticing a gap and hypothesis testing.

Example 3: Noticing the gap repair

Trigger: And ka, ah, he, she, ah, go to, ah, the entrance and hmm, meet Ed

Stimulated Recall Session

R： ここは

A3(4)： 次の場面が頭の中にすごくあった。ここでの喜びを表現するためにもっと適当な動詞があるとは思ったのだが、meetしか思いつかなくて、歯がゆい感じだった

[R: What about here?]

[A3(4): I was thinking really hard about this next part. I really wanted a verb which would express how happy she was, but all that I could think of was 'meet'. It was very frustrating.]

Repair:

R: She runs to Ed happily

(Translated by the researcher)

Example 4.4: Hypothesis testing repair

Trigger: So Kate, Kate down stairs and meet Edy

Stimulated Recall Session:

R： 長いポーズKate downstairsとゆっくり言っているが、そのとき、何を考えていましたか？

S4(4)： 「階段を下りる」がdownstairsでいいのか考えていた

[R: After this long pause, you said 'Kate downstairs' very slowly. What were you thinking at that time?]

[S4(4): I was thinking that I wanted to say 'run downstairs' and I wasn't sure if downstairs was right.]

Repair:

R: Kate ran downstairs

(Translated by the researcher)

It was hypothesized that learners who noticed gaps in their output during task performance, would be orientated toward subsequent input containing information which they could use to fill those gaps. Feedback, one possible form of subsequent input, was given to the Repair Group when, during the stimulated recall procedure, participants indicated that they were aware that they could not formulate a meaning as intended.

Initially, participants were asked to complete the pre-test which was on the first page of the booklet they were given. This was the first narrative task. During the pre-test an outline was attached to the mini digital video camera, fed into a video cassette recorder and the test was recorded on video as well as on audiotape and camera. The video-recording was used in the

stimulated recall.

The learners were asked to read the instructions on the first page of the booklet. They were also informed that the researcher had access to the same ten pictures, but in a different order. The researcher would listen to the produced narrative and attempt to place the ten pictures in the correct order as the participants narrated their story. This activity was included in an attempt to increase the communicative-value of the task. Comprehension of the instructions was checked and an opportunity to ask questions was given. Once a participant understood the narrative task, s/he was instructed to turn to the picture-story and given one minute to prepare. Because this experiment examines 'learners' noticing gaps in their output' and the role this plays in their learning, it was thought that one minute to prepare the narrative would provide enough time to plan the content, and thus allow more processing capacity to be focused on language during task performance. No time limit was placed on the completion of the task. This was also to ensure that the participants had the opportunity to plan on-line (Yuan and Ellis, 2003) and thus focus more on grammatical accuracy and complexity.

The researcher did not take an active role during the task. However, he did provide non-verbal and verbal back-channeling when he felt it was necessary. This was not in response to the accuracy of statements, but there was a possibility that this back-channeling was interpreted as positive feedback. However, the provision was the same for all four groups. In total this task took an average of 4.40 minutes (s.d. = 1.89 min.) to complete.

The first post-test was held immediately after the stimulated recall, during the first session. The test used the same task as the pre-test and followed the same procedures. One minute's preparation time was given to view the picture prompts, before an unlimited time was provided for students to orally produce their narratives. Although the participants were again informed that the researcher would order the pictures according to their narratives, it was clear that, as this was a task repetition, the communicative value of this activity was much reduced. This task took an average of 3.89 minutes with a standard deviation of 1.67 minutes to complete.

Nine of the eighty-one participants could not attend the second session at the arranged time, two weeks after the first session, so other times were arranged for them. The shortest time between sessions for any student was one week, while the longest was 22 days. All participants who came to the first session attended the second. (They were not paid until both sessions were completed.).

The procedures and tasks used in the second post-test were a repeat of those used for the pre and first post tests. Again there was one minute preparation, but unlimited time for completion of the task. The mean completion time was 3.61 minutes (s = 1.66 minutes).

The third-post test used the second task described in the instruments section. This was to determine if any gains in language performance could be realized in the performance of a second task. As with the other tasks, there was one minute preparation time provided before the task began, and unlimited time provided for completion of the task. The average time to complete the task was 5.56 minutes. The standard deviation was 2.97 minutes. Preparation time is not included in any timing.

Analysis

This final section describes how the information derived from the tests described

above was processed and then analyzed. Once the data was collected, this information was then uploaded from the digital recorders to a computer. The files were transcribed broadly, using conventional spelling, including all fillers, repetitions and reformulations. Time taken was also marked. There were five transcripts for each of the participants, one for the pretest, one for each of the posttests, and one for the stimulated recall. The pre- and posttest transcription was completed by the researcher, but the stimulated recall transcription was done by a native Japanese speaker. All analysis for the first and the second research question was carried out on these transcripts. The third research question investigates this data and the results from the working memory tests.

The second research question asked if awareness of problems influences the subsequent incorporation of input. This was examined by first determining if the participants noticed their gaps and then created language hypotheses, and if there was input available to fill these gaps. Second, the analysis of the connection between noticing problems and incorporation was described.

The first sub-question was answered by simply counting the number of times the participants attended to their output problems, based on their comments in the stimulated recall sessions. The problems were tallied into two categories for each participant: noticing the gap and hypothesis creation. Following this the rate of noticing the gaps and hypothesis creation per t-unit was calculated to allow for cross-group comparisons.

Noticing the gap was defined as any time in the pretest production of the narrative that the participant indicated that he or she had difficulty in formulating a message and the intended message was then abandoned. Examples 5 and 6 demonstrate noticed gaps. Intrarater reliability was measured at .967.

Example 5: Noticing the gap I

R： in the houseポーズがあって two peopleと自信がなさそうに言っていますが

H4(3)： 「夫婦」と言いたかったが、分からなかった

[R: After you said ‘in the house’ you paused before you said ‘two people’]

[H4(3): I wanted to say husband and wife, but I didn’t know how to.]

Stimulating sentence:

(Eto, two, in the house, two people sitting on the sofa)

(Translated by the researcher)

Example 4.6: Noticing the gap II

R： 長いポーズの後the man

E1(4)： この人の特徴について説明できないか考えていた。ひげもあるし、髪も立っているし、そういうことを説明したかったが、単語が分からなかった

[R: After a long pause, the man…]

[E1(4): I thought that I would not be able to explain the man’s characteristics. I couldn’t explain that the man had a beard, and that his hair was standing, I didn’t know the words.]

Stimulating sentence:

(ah, next, hmm, the man asked, asked, hmm, ah, ah, in the home, in the home, hmm, maybe father and mother, eh, the man asked, ask, asked to father, hmm, where is Kate?)

R: Researcher H4(3), E1(4): Participant Code (group Number)

(Translated by the researcher)

The second measure of attending to problems counted the participant's constructions of target language hypotheses. This was when they indicated an awareness that they were not sure whether what they had produced was correct, or was a good formulation of what they had actually intended to say. For this to be counted as a language hypothesis the ensuing output had to be either grammatically incorrect, or not an accurate representation of what the participants indicated they had wanted to say (see Example 7 and 8 below). Intrarater reliability was measured at .862.

Example 7: Hypothesis testing I

R： parents don't know ポーズwho is私を見てthe man

J1(4)： 文法を考えていた。だれだかを知らなかった。

R: Parents don't know, pause, who is, then you looked at me, the man

J1(4): I was thinking about the grammar. 'They didn't know who he is'.

Sentence: And, uh, but parents don't know who ah, who is the, the man (laugh)

(Translated by the researcher)

Example 8: Hypothesis testing II

R： Kate says ポーズshall weの繰り返し

M3(4)： 「食事を一緒にしようよ」と表現しようとして、dinnerをそのまま動詞として使ってもいいのかどうか考えていた

R: Kate says, pause, then you repeat shall we, shall we…

M3(4): I wanted to say 'lets have dinner together, but I wasn't sure if I was able to use dinner as a verb.

Sentence: The next pictures, the Kate says shall we, shall we dinner

R: Researcher J1(4), M3(4): Participant Code (group Number)

(Translated by the researcher)

Next the number of gaps noticed and hypotheses tested for which there was linguistic information available in the text were counted. Then the total number of the available words related to each of the gaps and hypotheses were also counted. These numbers represent the total possible incorporation for both noticed gaps and tested hypotheses.

Cross-group comparisons, based on the individual group treatments, were made using one-way ANOVAs. These were made to check that all groups were noticing to the same extent and were not being influenced by their treatments. This analysis was conducted only on the three groups which completed the stimulated recall procedure.

Analysis for the incorporation of input will be described next. First the degree of incorporation was measured for both groups. This was then calculated as a percentage of the total possible incorporation and the necessary comparisons were made. The first comparison determined if the rate of incorporation of information related to noticed problems was greater than could normally be expected. The second compared the rates of incorporation between the two experimental groups.

The first comparison was made using the Input Group's performance. First the total incorporation of information was measured based on the first post-test performance. The number of words incorporated from the input passage which had not been present in the pre-test oral narrative performance was counted to produce a total word incorporation count. These words were then compared with each linguistic problem noticed, and when a single word or more was matched with a noticed problem this was tallied to give a score for the total number of noticed problems 'filled'. The overall number of words was also tallied to give the total word incorporation related to noticed problems. Following this, the total word incorporation count was produced as a percentage of the total possible incorporation. Likewise, the total words actually incorporated related to noticed problems was given as a percentage of the total number of words available to fill these perceived problems in the input. As before, these ratios were produced both for noticed gaps and for created hypotheses and compared in two paired t-tests. The same comparison was not possible for the Repair Group as the rates of total incorporation of information and the total incorporation of information related to noticed problems were the same. This was because all information available to this group from repair was related to noticed problems.

Next the comparison between the two groups was made. The ratio of total noticed gaps incorporated and the total problem-related words incorporated were calculated for the Repair Group Finally, these ratios were compared using a simple t-test.

４．研究成果

Results

In order to respond to the question, does awareness of problems influences the subsequent incorporation of input?. several steps were thought necessary (see Fig. 2). First it was determined that the participants notice problems in their linguistic knowledge as they were producing a narrative task. Secondly, the presence of information in the input was confirmed for both the Input Group and the Repair Group. Subsequently, the participants' incorporation of these items was investigated.

Figure 2: A role for output in subsequent language production

Noticing the gap → (Available Input) → Incorporation

The results showed that the average number of gaps noticed and hypotheses formed during production for each participant was almost the same in each of the groups. There was an average of 5.5 gaps noticed and an average of 5.6 incorrect hypotheses produced. There was also a large range, with between zero and nineteen gaps (SD = 2.7) and zero and fourteen hypotheses (SD = 3.0). Although the minimum count was 0 for both gap awareness and hypothesis creation, the minimum combined count was three, indicating that all participants had attended to some kind of problem with their interlanguage knowledge. It was also found that the Stimulated Recall Group produced significantly more hypotheses that the other groups.

The analysis of the native-speaker narrative provided to the Input Group and the repair afforded to the Repair Group demonstrated clearly that sufficient appropriate information was available for the learners. The results are summarized in Table 1.

Based on these results, the data from three

participants was subsequently removed from further analysis. The first, a participant in the Input Group, had no information provided in the native speaker text relevant to filling the two gaps that she noticed. In this case, then, there was no information for her to notice. She did have information relevant to her language hypotheses however and her data was not removed for these analyses. Likewise another two participants, one each in the Input Group and the Repair Group, had not been provided with enough information to confirm their language hypotheses, of which they produced 2 and 3 respectively. Again, both were included in the 'gaps' analysis but removed from the analysis relating to language hypotheses.

The results of this section demonstrated that, for the majority of participants, information which could potentially fill gaps and hypotheses was available.

This next section presents results which demonstrate that information pertaining to the noticed gaps and hypotheses was incorporated into subsequent output, and thus represents 'a filling of the gap'. As before, first, the analytical procedures will be described. This will be followed by a presentation of the results.

The analytical procedures used to determine if the participants were actually incorporating information related to gaps in their linguistic knowledge are described here. Ten measures were used. The procedures used to calculate these are described below:

Overall Incorporation. The first measure was 'overall incorporation'. This was produced by adding the total number of words incorporated. The incorporation was defined as the words which appeared in the first posttest which were in the input narrative, but were not present in the pretest. This measure was produced for the Input Group only, as the Repair Group was only supplied input related to their noticed problems.

Rate of Overall Incorporation. The second measure was the 'rate of overall incorporation'. This was calculated by dividing the overall incorporated words with the total words available for input, resulting in an overall incorporation percentage. Again, this measure was only calculated for the Input Group.

Gaps and Hypotheses Filled. The number of gaps filled was calculated by matching the incorporated words with each noticed gap. Each time there was a match, this was counted (regardless of the number of words in each match). The total number of hypotheses filled was calculated in the same way.

Rate of Gaps and Hypotheses Filled. The rate of gaps filled was calculated by dividing the gaps filled by the total number of gaps noticed which had information available to incorporate. Likewise the percentage of hypotheses filled was calculated by dividing the hypotheses filled by the total number of hypothesis noticed which had information available to be incorporated. These rates were both expressed as percentages.

Words Related to Gaps and Hypotheses incorporated. For the words related to noticed gaps incorporated, the total number of words incorporated which matched the participants' noticed gaps was counted. The measure for the number of words incorporated which were related to hypotheses was calculated by matching the words incorporated to the words available to fill tested hypotheses.

Rate of Words Related to Gaps and Hypotheses incorporated. The rate of words related to gaps incorporated was calculated by dividing the number of words related to gaps incorporated by the total number of words available to fill the gaps. The rate of words related to hypotheses incorporated was calculated by

dividing the number of words related to hypotheses incorporated by the total number of words available to fill hypotheses.

These measures were calculated for the three experimental groups: the Stimulated Recall Group, the Input Group and the Repair Group. However, for the Stimulated Recall Group, the term 'change' rather than 'incorporation' should be used, as there was no 'treatment' input to incorporate. The original trigger causing the output problem in the pretest for each noticed gap and hypothesis was compared with the output in the first posttest. If there was no change in the utterance in any way, or it was not present (avoided) in the posttest, it was not counted. However, if there was some alteration in the utterance, then it was tallied as a change. All the words which were replaced or altered were also counted.

The analyses, first, confirmed that the rate of incorporation of information related to noticed gaps, was greater than the rate of overall incorporation. This was also true of hypothses. Analysis was done by using paired t-tests to compare the rate of overall incorporation from the input narrative with the rate of incorporation of information related both to noticed gaps, and to the hypotheses. It was, as explained previously, not possible to measure 'overall incorporation for the Repair Group.

The second analysis compared the three groups using two one-way ANOVAs. This method not only allowed the Input Group and the Repair Group to be compared but also determined if the levels of change were more than could be normally expected if no information was available to participants from either input or repair.

As shown in Table 2 below, in their second output attempt, the Stimulated Recall Group made changes to an average of 1.55 noticed gaps or 29.9% of the total. These changes involved an average of 4.05 words (17.2%). Hypotheses altered were on average only 0.6, or 15% of those created. This corresponded to 0.6 of a word per participant, or a mere 5.5% of the words available.

The overall incorporation came to 27.4% of the available input for the Input Group. These participants incorporated a mean of 3.6 of their noticed gaps into their first posttest. This represented 66.5% incorporation. 12.0 (37.3%) words were incorporated from those available in the input directly corresponding to the learners' perceived gaps. The number of hypotheses incorporated was lower, at an average of 1.3 (41.3%) per participant, involving a mean 3.2 (23.5%) incorporated words.

The Repair Group filled 3.8 (77.1%) noticed gaps and incorporated 12.1 (62.4%) words per individual. The hypotheses filled came to 4.1 on average (81.3%) and involved 10.1 (66.4%) words per person.

The next stage was to determine whether the rate of incorporation was greater than could be expected if problem awareness was not playing a role (Table 3). The incorporation of words that were related to noticed gaps reported during the stimulated recall session was 42.4% of the total items it would have been possible to incorporate. This was found to be significantly greater than the 27.4% overall incorporation of the input in the first posttest ($t(17) = 2.569$, $p = .02$). However, the percentage of items incorporated from hypotheses was 23.5%, which was lower than the general rate of noticing of 27.4%. This difference was not significant ($t(17) = -.678$, $p = .507$).

The one-way ANOVAs comparing the incorporation (or change in the case of the Stimulated Recall Group) of the three groups demonstrated significant differences for all of the

four measures: gaps ($F(2, 57) = 18.7$, $p < .001$) and gap related words incorporated ($F(2, 57) = 12.2$, $p < .001$), and hypotheses ($F(2, 56) = 23.6$, $p < .001$) and hypothesis related words incorporated ($F(2, 56) = 31.4$, $p < .001$).

The post-hoc multiple Tukey (HSD) comparisons (Table 4) showed that the Stimulated Recall group made changes to their output related to noticed gaps significantly less often than the other two groups, i.e. gaps (Input, D = 1.72; Repair, D = 2.39) and words (Input, D = 1.23; Repair, D = 2.33). The Input Group and the Repair Group incorporated similar percentages of information related to noticed gaps (D = 0.24), but the Repair Group incorporated more words than the Input Group (D = 0.55).

The proportion of hypotheses repaired was significantly lower for the Stimulated Recall Group than for the Input Group (D = 0.851), and both of these groups had much lower proportions than the Repair Group (D = 3.26). This was replicated for the proportion of words (Input, D = .781; Repair, D = 3.51). Finally the Repair group incorporated significantly more information related to hypotheses (D = .792) and words associated with it (D = .838) than the Input Group.

To conclude, the analysis showed that 1) the information available in the input to fill gaps was incorporated into the first posttest production at a significantly higher rate than 'overall incorporation' (the rate of all the words incorporated from the input narrative into the first posttest). This translated into a rate of 67% for the Input Group, and 77.1% for the Repair Group. 2) Both the Input Group and the Repair groups changed their second narrative production (the first post-test) with respect to the pre-test at much greater rates than the Stimulated Recall Group. This result shows that providing input is effective in inducing change. 3) Comparisons between the two groups showed no significant differences between them, indicating that they were essentially processing information related to noticed gaps in the same manner. The results are summarized in Table 5 above.

The results of hypothesis testing were different from those for noticed gaps (see Table 6). 1) The provision of repair during the treatment phase of the experiment resulted in the Repair Group reporting more hypotheses. This meant that there were more hypotheses which could be filled than for the other groups. 2) The rate of incorporation for both the Input Group and the Repair Group was much greater than the rate of change for the Stimulated Recall Group. 3) However, the Input Group's rate of incorporation was not greater than could be expected for the rate of 'overall incorporation' (the rate of all the words incorporated from the input narrative into the first posttest), which suggests that this group was not incorporating information related to the hypotheses that they had created earlier. Finally, 4) the Repair Group incorporated items more than the Input Group.

Experiment 2

Although Experiment 1 demonstrated that repair is uptaken and then incorporated in subsequent output, it still cannot be concluded that this is due to noticed gaps. The repair was provided to the participants every time that they indicated they noticed a gap, or were testing a hypothesis. There is no evidence that they were incorporating input because they were testing a hypothesis. The possibility remains that they would have uptaken and incorporated repair regardless of whether it was aimed at tested hypothesis or not.

Experiment 2 compares the uptake and

incorporation of repair aimed at tested hypotheses with repair aimed at random errors. Based on Swain's output hypothesis, it is predicted that repair aimed at tested hypotheses will be uptaken and incorporated at a greater rate than random repair.

Method

Experiment two followed the same procedure as experiment one, with new participants placed in three groups. These are explained below.

Participants

In order to determine the optimal samples size a power analysis of the results in the previous experiment was conducted. For a significance level of $p < .05$ and with statistical power of .8, a sample size of 8.52 participants per group was necessary.

The participants were drawn from the science and engineering department of the same large private university and were all enrolled in the first or second year of compulsory English language courses. The participants volunteered their services after receiving an explanation of the experiment and what was require of them. They were offered 2000 yen in return for their services. Their background was the same as for the participants in experiment 1.

Procedure

Three groups were formed for the experiment. The first group acted as a control. The first was the stimulated recall group and the second group was the repair group. The procedures for these groups was the same as with the groups of the same name from experiment one. The third group, random repair, was different however.

While the repair in the repair group was provided every time the participants indicated that they noticed a gap or were testing a hypothesis, the repair in the random repair group was provided every time they produced an error. Care was taken to ensure the amount of repair between the two groups was comparable. Based on the data from experiment 1, not more than 15 instances of repair was provided. The experimental procedure, the instruments used and the analysis were all the same as the previous experiment.

Results

The results for the repair group for experiment 2 were similar to that of experiment 1, as can be expected. On average, participants in this group noticed 4.8 (1.9) gaps and tested 5.2 (2.5) hypotheses. Of that the repair contained 78.0% of the information related to hypothesis and 87.5% related to their gaps. In contrast, although the random repair group noticed a similar level of gaps and tested hypotheses, only 24.4% and 25.6% of the information was available in the input. This information is summarized on table 7

The results replicate that for experiment 1, participants are both noticing gaps in their linguistic knowledge, and testing hypotheses. The differences between the two groups are with the linguistic information available for incorporation. The random repair group has limited information available for filling gaps, or confirming hypotheses. However, participants in this group received addition repair randomly 8.1 times on average containing 31.2 words.

The next section examines the extent to which each of the groups incorporated information available in the input, or in the case of the stimulated recall only group, the extent to

which their output related to their noticed gaps and tested hypotheses was altered.

Similar analytical procedures were employed to experiment 1. In addition, incorporation of other repair was added for the random repair group. Statistical analyses were performed on total incorporation which is a sum of all information incorporated including that related to noticed gaps, tested hypotheses, and, in the case of the random repair group, other repair.

The results, shown on table 8 demonstrated that there was a significant difference in the levels of uptake between groups. The repair group incorporated significantly more gap related information than the Stimulated recall group ($F(2, 23) = 4.615$, $p = .021$). Incorporation related to hypotheses tested was significantly greater for the repair group when compared to the other two groups. ($F(2, 24) = 10.489$, $p = .001$). The random repair group also received "other repair". This was incorporated at a much lower rate of 30%. This information is summarized on Table 9.

In order to determine if the repair group was incorporating more information than the random repair group, as predicted, the three types of incorporation (repair related to noticing the gap, repair related to hypothesis testing, and other repair) were combined and compared using a one-way ANOVA for the percentage of incorporated repair and a t-test for the percentage of incorporated words. The descriptive statistics are summarized on Table 10.

The results of the ANOVA demonstrated a significant difference between the three groups ($F(2, 24) = 30.338$, $p < .001$). Post hoc tests (Tukey HSD) indicated that the Repair group was incorporating significantly more of the available information than the Stimulated Recall Group and the Random Repair Group. The T-test demonstrated a similar results, with the percentage of incorporated words being significantly greater for the repair group than the random repair group ($t(16) = 5.326$, $p < .001$).

These results demonstrate very clearly that participants were more likely to incorporate information related to their noticed gaps and tested hypotheses at a much greater rate than when repair was provided at random. Interestingly, the random repair group did not outperform the stimulated recall group which did not receive any input whatsoever.

Discussion

The discussion has been conducted in two parts. The first examines the mechanism behind the noticing of gaps and the creation of language hypotheses during language production. This is followed by a discussion of whether an awareness of problems in their output orients learners' towards related information in the input, leading to its incorporation.

Noticing gaps and creating hypotheses

The results showed that learners both noticed linguistic problems and produced language learning hypotheses. This supports the conclusions of Swain and Lapkin (1995) who found in a dictogloss task that their 'learners may encounter a linguistic problem leading them to notice what they do not know, or know only partially' (Swain, 1995, p. 129). This present study extends Swain and Lapkin's research (whilch examined dialogic discourse) by showing that noticing occurs in oral on-line monologic production.

The process involved in the learner's recognition of output difficulties can be explained by Levelt's (1989) speech production model. First, many of the stimulated recall comments were related to the planning of the content, at the level

Levelt labeled 'the conceptualizer'. Although one minute was provided in the present study to determine content, this was often not enough to plan finer details. The participants also reported that they sometimes changed their intentions as they realized their initial plan was inadequate to communicate what they intended.

Once the preverbal message had been determined, the speakers often reported that they were unable to create the message. In other words, they noticed a gap in their interlanguage knowledge. This reflected Levelt's 'formulator' which he claims is responsible for the grammatical and phonological encoding of the message.

Levelt's model also includes monitoring. Once the message is formulated, it can be monitored through the speech comprehension system. In this current study, the participants often reported that they were unsure of the accuracy of the message they had formulated. This was defined as 'hypotheses testing' in this dissertation. Levelt refers to monitoring occurring either after formulation of the message in 'internal speech' or after articulation in 'overt speech'. The occurrence of both of these monitoring events was found in this study, with speakers identifying output problems both in long pauses before production occurred, and through repetitions and reformulations.

It is also possible to explain monitoring in Krashen's (1985) sense where explicit knowledge of the language is involved. Given the existence of two separate types of linguistic knowledge, explicit and implicit (i.e. Paradis, 1994; Ellis, 2005), it is also possible that monitoring of formulated language takes place utilizing explicit knowledge (instead of or in addition to implicit knowledge).

One of the assumptions of this research was that the three groups would become aware of production problems at similar rates. This assumption was borne out for the rates of gaps noticed but not for hypothesis testing. A one-way ANOVA demonstrated there was a significant difference in the rate of hypotheses created across the groups. Post-hoc tests showed the difference lay with the Repair Group which produced significantly more hypotheses than the Input Group. The Repair Group also produced more hypotheses than the Stimulated Recall Group but this was not significant.

It is possible that the difference between the groups was due to the provision of repair. It also seems that there were individual differences in the learners' ability to exploit repair as Levene's test of equality of variance showed a significantly larger variance ($F = 5.241$, $p = .028$) among members in the Repair Group.

It appears, then, that the provision of feedback in the form of repair led the participants to produce more hypotheses. What is not clear is the reason for this. It is possible that the Repair Group participants were induced to recall more of their original problems. However, it is also possible that the provision of repair assisted them to create even more hypotheses during the post-monitoring of their performance as a result of the stimulated recall procedure itself. Whichever is the case, it is clear that the Repair Group noticed more linguistic problems.

Incorporation of Subsequent Input

The hypothesis tested was based on Swain's Output Hypothesis. It predicted that noticing gaps and creating hypotheses will orient participants to related information in the input, resulting in incorporation. This section discusses the degree of incorporation for each of the treatment groups and concludes by confirming the

hypothesis for both the Input and Repair Groups.

The two groups behaved quantitatively and qualitatively differently based on the treatments they received. Although, in the first task, the Repair Group did not incorporate input related to noticed gaps at a rate greater than that of the Input Group, the rate of incorporation of information related to hypotheses was significantly greater. Also, in incorporation of information related to language hypotheses, the Repair Group outperformed the Input Group. This next section attempts to explain these disparities.

Positive and Negative Evidence

The first difference is in the form of the input. The Input Group received only positive evidence, whereas the Repair Group received both positive and negative evidence. It appears that the negative evidence contained in the repair is important for the participants to be able to confirm or disconfirm their language hypotheses (see Leeman, 2003).

The provision of feedback to the participants in the Repair Group provided them with explicit information as to the correctness of their language hypothesis. Their hypotheses were only repaired when they were incorrect which enabled them to either confirm or disconfirm them. Once they were able to disconfirm a hypothesis, they were then in a position to incorporate the information encoded into the repair into subsequent performance.

The content of the native speaker's narrative on the other hand was not fine-tuned to the participants' hypotheses. As a result, the information was not necessarily present to check all their hypotheses. In addition, the information was present for both the correct and incorrect hypotheses, in contrast to the Repair Group, where the input was only available for the inaccurate hypotheses. This combination most likely made it to difficult to first disconfirm a hypothesis, and then incorporate the information related to it.

Processing Constraints

N. Ellis (2005) states that one of the benefits of recasts is that they occur immediately after a problematic utterance. Doughty (2001), for example, calculates that given that working memory traces decay exponentially a recast should be within about 20 seconds of activation. The Input Group obtained their input between twenty and thirty minutes after their first narrative task, which may have made retention of noted gaps in conscious memory difficult. Repairs were also provided well after the original noting of the problem. However, the difference here was that awareness of the problems had been reactivated (and possibly created) and verbalized during the stimulated recall procedure, Thus the repair would have come at close to the optimum time for processing in working memory.

In summary, two possible explanations have been put forward for the differences found between the Input Group and the Repair Group. The first was the distinction between repair which contained both negative and positive evidence and the native speaker narrative which was positive evidence alone. Processing restrictions were, then, put forward as a possible reason for group differences. The capacity and temporal limitations of working memory meant that repairs which came directly after noticing speech production problems were more likely to be incorporated than subsequent input which was made available some time later.

References

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research, 7*(3), 347-376.

Alanen, R. (1995) Awareness in Language Learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning*.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11-53). New York: Wiley.

Carroll, S. E. (1991). *Input and Evidence: The raw material of second language acquisition.* Amsterdam: John Benjamins.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). Cambridge: Cambridge University Press.

de Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning, 46*(3), 529-555.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition in second language instruction* (pp. 206-257). Cambridge: Cambridge.

Doughty, C., & Williams, J. (1998a). Issues and terminology. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 1-12). Cambridge: Cambridge University Press.

Ellis, N. C. (2005a). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*(2), 305-352.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141-172.

Ellis, R. and Sheen, Y. (2006). *Reexamining the role of recasts in second language acquisition*. Studies in Second Language Acquisition, 28, 575-600.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Faerch, C., & Kasper, G. (1987). From product to process - introspective methods in second language research. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 5-21). Philadelphia: Multilingual Matters.Jourdenais, R. (2001). Cognition, instruction and protocol analysis. In P. Robinson (Ed.), *Cognition in second language instruction* (pp. 354-376). Cambridge: Cambridge University Press.

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*(3), 421-542.

Jourdenais, R., Ota, M., Stauffer, S., Boyson, B., & Doughty, C. (1995). Does textual enhancement promote noticing? A think-aloud protocol analysis. In R. Schmidt

(Ed.), *Attention and awareness in foreign language learning* (pp. 183-216). Honolulu: University of Hawai'i Press.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. New York: Pergamon Press.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Lincolnwood, Il.: Larendo.

Leeman, J. (2003). Recasts and second language development: Beyond negative evidence. *Studies in Second Language Acquisition, 25*, 37-63.

Leow, R. P. (1998a). The effects of amount and type of exposure on adult learners' L2 development in SLA. *The Modern Language Journal, 82*(1), 49-68.

Leow, R. P. (1998b). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) fine-grained analysis of attention. *Applied Psycholinguistics, 19*, 133-159.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition, 22*(4), 557-584.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26*, 35-57.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Ma: MIT Press.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego, Ca: Academic Press.

Long, M. H. (2007). *Problems in SLA*. New Jersey: Lawrence Erlbaum.

Lyster, R. (2004). Differential Effects of Prompts and Recasts in Form-Focused Instruction. Studies in Second Language Acquisition, 26, 399-432.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learning uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19*(1), 37-66.

Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research, 37*, 379-394.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Oxford, R. (1995). *Language Learning Strategies: What every teacher should know*. Boston: Heinle & Heinle.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: Implications for bilingualism and SLA. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 393-420). San Diego, Ca.: Academic Press.

Philp, J. (2003). Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition, 25*, 99-126.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set.

Journal of Experimental Psychology: Human Learning and Memory, 2(1), 88-94.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition in second language acquisition* (pp. 287-318). Cambridge: Cambridge University Press.

Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Rosa, E., & O'Neill, M. D. (1999). Explicitness, intake, and the issue of awareness: Another piece in the puzzle. *Studies in Second Language Acquisition, 21*(4), 511-556.

Rosa, E., & Leow, R. P. (2004a). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics, 25*, 269-292.

Rosa, E., & Leow, R. P. (2004b). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *The Modern Language Journal, 88*(2), 192-216.

Russell, J. and Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In Norris, J. M. and Ortega, L. (Eds.) Synthesizing Research on Language learning and Teaching. (pp. 133-164). Amsterdam: John Benjamins.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Sheen, Y. (2010). *Differential Effects of Oral and Written Corrective Feedback in the ESL Classroom.* Studies in Second Language Acquisition, 32, 203-234.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125-144). Oxford: Oxford University Press.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16*(3), 371-391.

Truscott, J. (1996). *The case against grammar correction in L2 writing classes*. Language Learning, 46, 327-369.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning in fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1-27.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１件）

① Sheppard, Chris. (2008). Noticing the gap, hypothesis testing, and the uptake of subsequent feedback. Poster presented at the Second Language Research Forum 2008, October, 19.

６．研究組織
(1)研究代表者
シェパード・クリス（Sheppard, Chris）
早稲田大学・理工学術院・準教授
研究者番号：**60350386**

Figure 1: Experimental design

Group / **Test**	**Control**	**Stimulated Recall**	**Input**	**Repair**
Pre-test	Narrative 1 ↓	Narrative 1 ↓	Narrative 1 ↓	Narrative 1 ↓
Treatment	No Treatment	Stimulated Recall	Stimulated Recall ↓ Input Note-taking	Stimulated Recall and Embedded Feedback
Posttest one	↓ Narrative 1	↓ Narrative 1	↓ Narrative 1	↓ Narrative 1
Two-week interval				
Posttest two	Narrative 1 ↓	Narrative 1 ↓	Narrative 1 ↓	Narrative 1 ↓
Posttest three	Narrative 2	Narrative 2	Narrative 2	Narrative 2

Table 2: Word frequencies for the 'noticing' passage

Word List	Tokens	%	Types	%
One Thousand	208	76.19%	100	77.52%
Two Thousand	26	9.52%	17	13.18%
Three Thousand	1	0.37%	1	0.78%
Other	9	3.30%	9	6.98%
Names	29	10.62%	2	1.55%
Total	273	100.00%	129	100.00%

Table 1: Means and standard deviations of the available information

Group		Total Input	Gap Count	Gap %	Gap Words	Hypotheses Count	Hypotheses %	Hypotheses Words
Input	Mean	264	5.32	85.0	29.4	3.21	62.9	15.8
	S.D.	-	2.71	24.0	16.3	2.18	30.4	11.0
Repair	Mean	-	4.90	90.3	20.1	5.10	75.7	16.1
	S.D.	-	2.19	25.3	10.7	2.81	23.7	11.9
Total	Mean	-	5.10	87.8	24.5	4.20	69.6	16.0
	S.D.	-	2.43	24.6	14.24	2.67	27.5	11.3

Table 2: Means and standard deviations of incorporated information related to noticed gaps and hypotheses.

Group		Overall Wo.	Overall %	Gaps No.	Gaps %	Gaps W.	Gaps %	Hypotheses No.	Hypotheses %	Hypotheses Wo.	Hypotheses %
Stim. Recall	M	-	-	1.55	29.9	4.05	17.2	.600	15.0	.600	5.5
	N	-	-	20	20	20	20	20	20	20	20
	S.D.	-	-	1.15	22.0	4.61	19.1	.821	22.1	.940	11.7
Input	M	72.3	27.4	3.56	66.5	12.0	37.3	1.28	41.3	3.22	23.5
	N	18	18	18	18	18	18	18	18	18	18
	S.D.	20.5	7.45	1.29	20.5	7.32	13.1	1.18	37.7	3.56	30.4
Repair	M	-	-	3.76	77.1	12.1	62.4	4.15	81.3	10.2	66.4
	N	-	-	21	21	21	21	20	20	20	20
	S.D.	-	-	1.79	17.3	6.62	19.7	1.90	18.4	6.16	21.6
Total	M	-	-	3.67	57.6	12.1	39.0	2.79	46.6	6.87	32.6
	N	-	-	59	59	59	59	58	58	58	58
	S.D.	-	-	1.56	28.6	6.86	25.8	2.15	38.5	6.13	34.2

Table 3: Paired samples test for the % total noticing and % of noticing related to perceived gaps and hypotheses.

	Mean	Std. Deviation	T	Df	Sig. (2-tailed)
Noticing	.150	.248	2.57	17	.020
Hypotheses	-.0497	.311	-.678	17	.507

Table 4: Tukey's (HSD) multiple comparisons for the proportions of noticed gaps and hypotheses incorporated

		(I) Group	(J) Group	Difference	Sig.
Gaps Noticed	% Gaps	Stim	Input	-0.366	0.000
			Repair	-0.472	0.000
		Input	Repair	-0.106	0.241
	% words	Stim	Input	-0.202	0.003
			Repair	-0.453	0.000
		Input	Repair	-0.251	0.000
Hypotheses Tested	% hypothesis	Stim	Input	-0.263	0.012
			Repair	-0.663	0.000
		Input	Repair	-0.400	0.000
	% words	Stim	Input	-0.180	0.046
			Repair	-0.609	0.000
		Input	Repair	-0.429	0.000

Table 5: The process of incorporation (noticed gaps)

Group	Gaps Noticed	Fill-able Gaps	Incorporation No. %
Stim Recall (SR)	5.2	-	0.30
Input (I)	6.2	5.3	0.67
Repair (R)	5.2	4.9	0.77
Group Comparisons	N.S.	-	SR < I, R

Table 6: The process of incorporation (hypothesis formation)

Group	Hypotheses	Fill-able Hypotheses	Incorporation No %
Stim. Recall (SR)	5.0	-	15.0
Input (I)	5.0	3.2	41.3
Repair (R)	6.7	5.1	81.3
Group Comparisons	SR, I < R	I < R	SR < I, R; I < R

Table7: Means and standard deviations of the available information

Group		Gap			Hypotheses		
		Count	%	words	Count	%	Words
Repair	Mean	4.8	87.5	27.7	5.2	78	16
	S.D.	1.9	12.9	8.3	2.5	10.1	5.7
Random Repair	Mean	5.1	24.4	8.8	5.1	25.6	6
	S.D.	1.6	39.6	4.4	2	13.6	3.5

Table 8: Means and standard deviations of incorporated information related to noticed gaps and hypotheses.

		Gaps				Hypotheses			
Groups		No.	%	Words	%	No.	%	Words	%
Stim. Recall	M	1.4	0.4	4.9		0.8	0.2	1.2	
	N	9	9	9		9	9	9	
	S.D.	0.9	0.3	3.1		0.7	0.2	1.1	
Repair	M	3.4	0.9	15.7	0.6	3.2	0.8	12.4	0.8
	N	9	9	9	9	9	9	9	9
	S.D.	1.5	0.2	4.9	0.1	1.4	0.2	4.7	0.1
Random Repair	M	0.8	0.7	3.8	0.4	0.8	0.5	3.9	0.6
	N	9	8	9	8	9	9	9	9
	S.D.	0.7	0.5	3.0	0.3	0.7	0.4	2.9	0.3
Total	M	1.9	0.6	8.1	0.5	1.6	0.5	5.9	0.4
	N	27	26	27	17	27	27	27	27
	S.D.	1.6	0.4	6.5	0.2	1.5	0.4	5.8	0.4

Table 9: Other repair incorporated by the random repair group

		Other Repair			
Group		No.	%	Words	%
Random Repair	M	2.8	0.3	7.6	0.2
	N	9	9	9	9
	S.D.	1.6	0.2	5.3	0.2

Table 10: Total Incorporation by group

		Total Incorporation			
Group		No.	%	Words	%
Stim. Recall	M	2.2	0.3	6.1	
	N	9	9	9	
	S.D.	0.7	0.1	3.1	
Repair	M	6.7	0.8	28.1	0.6
	N	9	9	9	9
	S.D.	2.1	0.2	7.4	0.1
Random Repair	M	4.3	0.4	15.2	0.3
	N	9	9	9	9
	S.D.	1.8	0.2	6.4	0.1
Total	M	4.4	0.5	16.5	0.3
	N	27	27	27	27
	S.D.	2.4	0.3	10.8	0.3