TR-IT-0232

Echoing in Japanese conversations

Marc Swerts[1], Hanae Koiso[2],
Atsushi Shimojima[2] and Yasuhiro Katagiri[2]
[1]*ATR ITL and* [2]*ATR MIC*

September 5, 1997

ABSTRACT
The study reported upon in this paper focusses on different functions of echoing in Japanese dialogues. Echoing is defined as a speaker's lexical repeat of (parts of) an utterance spoken by a conversation partner in a previous turn. The phenomenon was investigated in three task-oriented, informal dialogues. Taking Traum's model for grounding as a basis, repeats in this corpus were labeled in terms of whether or not the speaker signals that he has integrated the other person's utterance into his own body of knowledge. The investigation brought to light that the level of integration is reflected in a number of lexical and prosodic correlates. These features are discussed regarding their information potential, i.e., their signal accuracy and comprehensiveness. In the future, the research needs to be extended to a larger database and it should be checked experimentally whether the discourse labeling can be reproduced reliably.

1 Introduction

1.1 General

Except for particular settings, the exchange of information through spoken language is usually not an exact data transfer process between a sender and a receiver. A person who talks to another cannot simply take it for granted that all his messages are completely understood by the other party. Communication failures may arise for a variety of reasons: e.g. there may be different types of noise on the channel, a speaker may overestimate the other person's knowledge about a given state of affairs or the listener may simply not have been paying enough attention. Despite this uncertainty in verbal interaction, it is often not possible for a speaker to tell by direct inspection whether a message came across succesfully.

Therefore, dialogue partners constantly negotiate on the information being exchanged in the course of their conversation. Being aware of the fact that spoken communication is a rather risky business, they normally try to reach mutual understanding in a collaborative way. This fits in a view that language is more than a *static* symbolic system, but rather that it is a *dynamic* medium for communication (Clark and Wilkes-Gibbs 1986, Traum 1994). In such a perspective, the process of understanding an utterance is more than one individual's attempt to associate the sound form of incoming speech with a particular meaning. Instead, communication is teamwork which involves the active participation of all conversants who constantly have to seek and provide evidence to coordinate their mental beliefs (Brennan 1990).

This is clearly illustrated by the fact that the most prototypical example of language usage, daily-life conversation, is characterized by different signals that -strictly speaking- do not contribute to the content of the topic at hand, but serve to manage the dialogue. Some of these cues may be non-verbal, e.g. like the use of pointing and other hand gestures, head nodding, gaze, etc. Conversants also use particular utterances in order to acknowledge receipt of a message, to repair it or to ask for clarifications. The ultimate goal of such dialogue management utterances is obviously to support an optimal exchange of information between dialogue partners.

The process of how conversants reach a state of mutual understanding of what was intended by the speaker of an utterance has been defined by different scholars as the so-called grounding phenomenon (Clark and Schaefer 1989; Traum 1994). The current research focusses on the relevance for grounding of repetitive or echoing utterances (see below). This type of utterances has received some attention in recent conversational analyses, but is yet not completely understood. In particular, it needs to be explored in more detail to what extent speakers use prosody to differentiate various functions of echoing.

1.2 Specific

Dialogues between two or more persons sometimes show instances where a speaker in one way or another repeats what his or her conversation partner just said in a previous turn.

Examples of such repetitive utterances are given in fragments (1) and (2) below of dialogues between speakers A and B:

(1) A and then you transfer to the keage line ...
B keage line
A which will bring you to kyoto station

(2) A and that is the keage line ...
B keage line ?
A that's right, keage line

Repetitions can present a problem to linguistic theories that too simply posit that every sentence should contain new information. Indeed, in a literal sense, the turns produced by speaker B are logically redundant, since they are lexically and propositionally identical to (parts of) a previous utterance. A discourse model, however, is descriptively inadequate if it is not able to account for such repeats, in part because they have intuitively clear pragmatic functions in everyday conversations and may even serve multiple goals. In general, they can be explained in terms of dialogue management behaviour.

For instance, the purpose for repeating in (1) is probably to acknowledge that information has been received, in that sense being equivalent to the usage of a simple "*uhuh*", whereas the repeat in (2) signals a communication error, i.e., B appears to be unsure about the information provided by A and wants to have confirmation that he understood A correctly. From the point of view of information flow, the repeats in the examples above are very distinct. Speaker A's incidental miss of an acknowledgment such as in (1) does not necessarily lead to communication problems afterwards. However, to guarantee successful interaction, it seems more crucial that A really detects the request for repair in (2).

An important question is what differentiates various usages of repeats in actual conversations, in order to enable conversation partners to interpret them correctly. This study looks into their prosodic features to see whether these correlate with communicatively different repeats. The section below deals with previous work on repeats and their prosodic correlates. Then, definitions are given of grounding, mainly based on work by Traum (1994), and of different types of repeats. The next three sections elaborate on the specific hypotheses of this study, the methods used and the specific results. The paper ends with a general discussion and conclusion.

2 Previous work

Tannen (1989) remarked that, for particular linguistic schools, repetition is basic to language in general in the sense that speakers constantly utter sentences they heard before or use standard phrases and fixed expressions. The current study, however, will be limited to the functional use of repetition in interactive talk. Though there are cultural and individual

differences (Tannen 1989), repetitive utterances are omni-present in most conversations in which they are used for "accomplishing social goals, or simply managing the business of conversation" (p. 51). This section will give an overview of recent work on repetitions and other informationally redundant utterances in different types of conversations, with a main focus on investigations of prosody.

Couper-Kuhlen (1996) investigated repetitive utterances and specifically explored to what extent these were accompanied by prosodic matching. Her data consist of two hours of recordings of a radio phone-in programme in which listeners call in to the studio and try to guess the answer to a riddle. The study reveals that speakers often match their pitch register, both in a relative and an absolute way. According to Couper-Kuhlen, the former is used to signal "quotation", the latter "mimicry". Useful from this study is the insight that repetition can be defined at different linguistic levels (e.g. lexical and/or prosodic), and that one can specify different types of intonational matching, i.e., in terms of register or tone.

The richest source of research on repetitions is children's discourse (Keenan and Schieffelin, 1976; Ochs 1979; Tannen, 1989). Not only do children often repeat utterances addressed to them, repetition also appears to be a useful didactic technique of adults in teaching language to children. One example in this tradition of research, which explicitly took into account prosody, is Tarplee (1996). She investigated a collection of adults' repetitions of children's utterances, produced in a particular interactional setting – where the children are engaged in labelling from picture books. It appears that repetitions can be used as affirmatory or reparative actions. The latter are prosodically different from the former in that they are marked with a contrastive pitch contour and there is a comparatively long temporal delay in their placement.

The timing result reminds one of other studies that have dealt with overlap and delay of speaking turns. For example, Fais (1994) gives evidence for the argument that natural dialogue is essentially collaborative by showing that conversants often utter (parts of) sentences simultaneously. Such overlapping co-productions suggest that a conversation is the result of the combined efforts of the dialogue partners. Similar research on timing of turns and its relevance for grounding is discussed in Brennan (1990).

Walker (1992) argues that one discourse function of repetitions and other types of informationally redundant utterances such as paraphrases is to provide evidence to support the assumptions underlying the inference of mutual beliefs. According to Walker, a speaker can get evidence about the result of his 'linguistic action' if the conversant repeats or paraphrases his utterance. Interestingly, Walker also discusses the possibility that repeats when realized with question intonation may represent some form of negative evidence, i.e., they point to conflicts in beliefs. Stenström (1994) calls these cases echo-questions that function as checks or express strong surprise. The default, however, is that repeats function to accept information provided by the other party.

Summarizing this short overview of previous work, one can state that several researchers mention the existence of basically two types of repeats, i.e., confirmatory and reparative

ones. Though most studies lack a sufficient degree of explicit phonetic detail, it is suggested that the latter category is characterized by marked prosodic features, in terms of contrastive melodic contours, question intonation or large temporal delay.

3 Definitions

3.1 Grounding

The repeats in this study were analyzed using the framework of grounding, in particular the computational theory of it as proposed by Traum (1994), as a source of inspiration. Therefore, this section will first present the basic ideas of grounding, and the related problems with it as noticed by Traum. Then, it will deal with the core of Traum's computational protocol. The final subsection introduces the extensions to this model, which were useful for the current research.

3.1.1 Basic ideas

Traum's (1994) computational model for grounding is based on, but partly competitive with, Clark and Wilkes-Gibbs (1986), Clark and Schaefer (1989) and Brennan (1990). Starting assumption in these studies is that conversants need to bring a certain amount of common ground to the dialogue in order to understand each other. Clark and Schaefer (1989) propose that this is achieved by means of *contributions* that consist of two parts:

Presentation phase: **A** presents utterance **u** for **B** to consider. He does so on the assumption that, if **B** gives evidence **e** or stronger he can believe that **B** understands what **A** means by **u**.

Acceptance phase: **B** accepts utterance **u** by giving evidence **e'** that he believes he understands what **A** means by **u**. He does so on the assumption that, once **A** registers evidence **e'**, he will also believe that **B** understands.

The evidence given by B may vary in strength, with "continued attention" being the weakest and "a verbatim display" being the strongest form.

In Traum's view, there are basically three sorts of problems with the theory given above. First, though the presentation and the acceptance phases may be intuitively appealing, it is sometimes hard in practice to exactly decide to which of the two an utterance actually belongs (e.g. in the case of an other-initiated self-repair). Second, since every presentation needs to be followed by an acceptance, it is not clear when a grounding act is completed, since acceptance in itself also needs to be accepted, which again needs to be accepted, ad infinitum. Finally, the model is insufficient to use as a guide for an agent in a conversation deciding what to do next based on what has happened before.

3.1.2 Computational model

As a solution, Traum presents a protocol for grounding in the form of a finite-state grammar, which is based on previous work, but also modifies it in order to minimize its deficiencies. It basically is a theory on how a **discourse unit** (DU) is constructed from a sequence of grounding acts between two agents, i.e. an Initiator (I) and a Responder (R). The grounding acts specified below all operate at the level of utterance units (UU), which Traum defines as a continuus stretch of speech by the same speaker, punctuated by prosodic boundaries (including pauses of significant length and boundary tones).

initiate act The opening utterance of a DU, which usually corresponds to the first utterance in the presentation phase, as defined by Clark and Schaefer (1989).

continue act Subsequent utterances which add new material in a presentation.

acknowledgment An utterance which claims or demonstrates understanding of a previous utterance.

repair Acts that change the content of the current DU, either in correcting previously uttered materials, or in adding omitted material which will change the interpretation of the speaker's intention.

request-repair A request for a repair by the other party.

request-acknowledgment Attempt to get the other agent to acknowledge the previous utterance.

cancel Closes off the current DU as ungrounded.

A finite-state grammar which covers all these grounding acts, and which was implemented in the TRAINS conversation system, is visualized in Figure 1. This network, which basically represents the core of Traum's theory, can be extended by additional depths of nesting to model subdialogues. For instance, to allow the possibility of a repair request by the initiator of the DU after some sort of a response by the receiver, one needs a request-repair network from state F in Figure 1.

3.1.3 Extensions

The model sketched above served as a useful starting point for the current investigation, because it is able to elegantly assign different roles to utterances in a grounding sequence. However, after some experimentation with it, it soon became clear that the model needed to be extended in a couple of ways for the purpose of the current study of repeats.

First, Traum's theory seems to imply that only those utterances can act as genuine grounding acts that are other-directed. That is, the Responder's underlying communicative

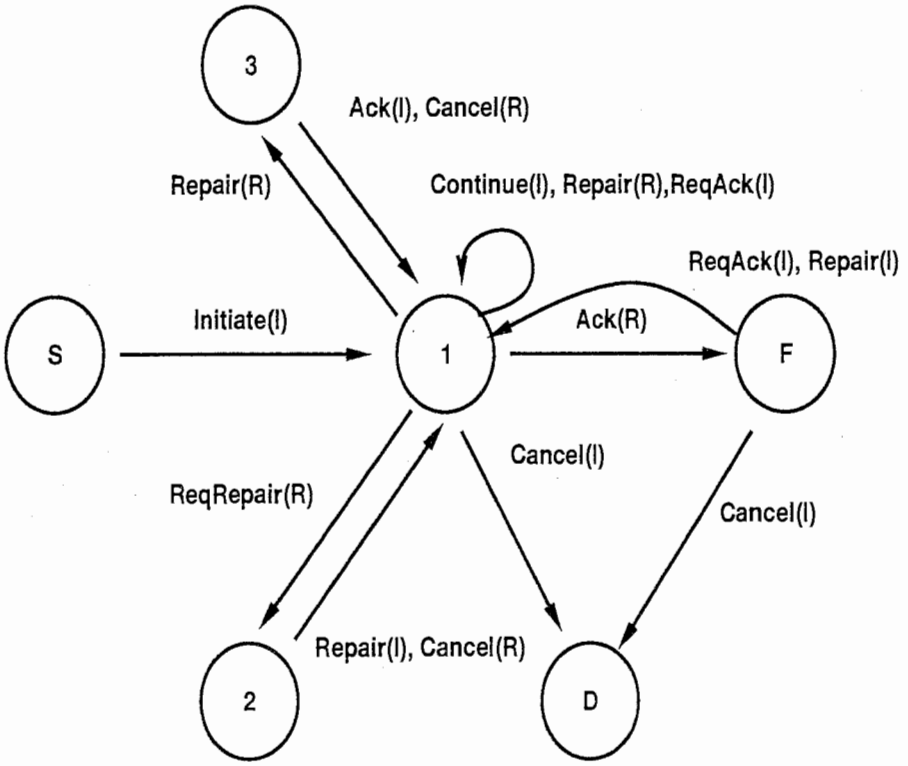

Figure 1: Finite-State Network of a DU (equivalent to Figure 3.8 in Traum (1994))

intention should be to explictly provide evidence to the communication partner. However, one often has the impression that speakers also produce utterances that strictly speaking are not "communicative" in this sense. For instance, a speaker sometimes may just mumble because he is surprised to receive particular information. Yet, although such utterances may not intentionally be used to give evidence about information received, they can still have signal value, since the other party can interpret them as cues to the Responder's mental beliefs.

Second, within Traum's framework, there is room for the specification of more grounding acts. First, one might consider to define a category ("Request-continue") which covers a form of acknowledgment occuring before the total DU is completed. Basically, the Responder uses it to confirm the acceptance of particular information, at the same time signaling that he would like to receive more. Second, there are cases that seem to be (intentionally) ambiguous between an acknowledgment or a request-continue on the one hand and a request-repair on the other. In this sense, they represent a separate category ("Display") by means of which a Responder indicates that he leaves it up to the Initiator to interpret them as either of two options: if the information is correct, the Initiator can continue; if it is not, he needs to repair.

Because of these drawbacks, a labeling protocol will be proposed below (see section 5.2),

which uses "level of integration" as a central concept. The scheme is claimed to be more general than Traum's proposal because (i) it is able to capture more grounding acts, i.e., those that are ambiguous or those acknowledge information before the DU is completed, and (ii) it can account for utterances that mainly seem to be self-directed.

3.2 Repeats

Repeats can be defined formally according to several criteria. For instance, Tannen (1989) makes a first distinction between self-repetition and allo-repetition. A second dimension is a scale of fixity in form, ranging from exact repetition to paraphrase. Finally, one may distinguish between differently timed repetitions, from immediate to delayed repetition.

For the present study, only repetitions of the other conversant's utterances are considered. Taking utterance units (UUs) as a unit of analysis, "echoing" was operationalized in the following way:

> Let X be a sequence of utterance units made in a single speaking turn, and Y be another sequence of utterance units made in the directly following turn. Then, X and Y are echoic pairs if and only if a sequence of morae that occupies 50 percent or more of Y already appears in X or is a semantic paraphrase of a part of X.

The definition explicitly did not capture utterances repeating previous materials when one or more intervening turns occurred. Although such echoing sequences occurring with a considerable delay are potentially interesting, it is virtually impossible to differentiate real echo's from incidental repeats, if one does not constrain repeats to those that appear in previous turns. The concept of semantic variant in the definition may sound vague at first sight, but only clear-cut cases were included being only slightly different from the repeated utterance unit (see Appendix A for some examples).

There were two more constraints. First, only repeats coming from the Responder were considered, thus excluding cases that served as "initiates", i.e., starting a new grounding act or being part of an answer to a question. Finally, the research did not take into account typical standardized adjacency pairs such as in greetings, like the repeating of "moshimoshi" (hello). Given these constraints, the definition given above resulted in a total of 71 repeats in the corpus (see below)).

4 Hypotheses

One could argue that one set of grounding acts represent a form of confirmation (acknowledgment, request-continue), whereas another set consist of cases that hint towards a (possible) conflict (display, request-repair); the former reflect integration of information, the latter non-integration. Because they serve as a flag to an understanding problem, one expects that repeats expressing non-integration are more crucial for information flow, because if the

Table 1: List of prosodic features and their expected settings for integrating and non-integrating repeats

Features	Integration	Non-integration
Pitch register	low	high
Intonation contour	declarative	interrogative
Loudness	soft	loud
Delay	short	long
Tempo	fast	slow

problem is not solved by the other party further communication may break down. Based on the literature presented above, it is logical to assume that to some extent these utterances share the same prosodic features as the corrections discussed in Swerts and Ostendorf (1997). They found that subjects talking to a speech understanding system tend to provide their utterances with marked intonational and durational characteristics if they have to correct a previous query due to an understanding problem of the system.

Therefore, one would hypothesize that non-integrative instances of echoing utterances are provided with features that are marked, whereas the integrating ones are expected to exhibit unmarked features. Though this does not pretend to be an exhaustive list, this could lead to the prosodic predictions for the two categories as summarized in Table 1.

Next to these prosodic features, there may be differences in the lexical context of repeats. There are basically two different possibilities. On the one hand, one could expect that repeats signaling non-integration are embedded in a longer utterance unit. Since it is essential that these utterances are picked up by the information giver, one could argue that they are more likely to occur in a turn in which the understanding problem is flagged with additional lexical materials. On the other hand, it is also possible that the Responder tends to limit the repeat to just that part of the other speaker's turn which was problematic, in other words explicitly focusing on the information for which a repair is requested, which would lead to relatively short repeats. Therefore, to find out which of these opposite expectations is valid, the current study will also deal with surrounding lexical context of repeats.

5 Method

5.1 Data

Analyses were based on three elicited, Japanese dialogues recorded at ATR-MIC, between each time two male undergraduate students who were familiar to each other. While they were seated in a sound-isolated studio, one participant was given the task to orally instruct

the other on how to build a particular construction, like a "duck", using differently coloured blocks. The result had to be similar to a construction shown on a picture, which only the instruction-giver could see. Both participants were allowed to gesture during communication, but the instructor could not physically touch any of the blocks. Using a head-set with microphone for both participants, the speech materials were recorded on separate channels, so that even when their speech overlapped in time, their voices were separate on tape. Additionally, during the dialogues, the participants' faces and hands were recorded on video, though these images would not be used in the current study.

In the original set-up, there were different conditions. Subjects could communicate face-to-face (1a), while seated at both sides of a window (1b) or through video display (1c). In one set of dialogues ("open" condition) the instructor could see the result of his instructions, whereas in another set ("closed" condition) he could not. After a short introduction in which the subjects were informed about the task and their respective roles, followed by a practice session, the actual experiment consisted of three dialogues between the same pair of subjects, the first two in the "open" condition, the third in "closed" condition. Sessions 1 and 2 were always different with respect to 1a, 1b or 1c, whereas sessions 2 and 3 were always the same regarding 1a, 1b, and 1c. The subjects had to end their conversation after 15 minutes. They did not have to complete all the constructions, but were instructed to perform them in a particular order, going from easy (3 pieces) to more difficult (12 pieces).

For the current investigation, three sessions of block dialogues by different pairs of speakers were analysed in terms of repeats. All the dialogues were of the "closed" type, but the first two were recorded under condition 1b, whereas the last one was under condition 1a. The data were first fed into the computer with a 20 kHz sampling frequency and converted into Xwaves format. Using the power measurements, the speech materials were automatically divided into "utterance units," defined as consecutive stretches of speech bounded by silence. Start and end time of each unit were extracted automatically.

5.2 Labeling

The dialogue act specification of the different repeats was achieved by means of a consensus labeling between the three authors affiliated with MIC. To this end, they could listen to the speech and read the transcribed texts of the repeats as often as they liked and take into account whatever dialogue context, until consensus was reached.

The following procedure was used. First, those echoic UUs considered to be "initiates" and "repairs" were identified and excluded from further analyses. The other instances of repeats were rated in terms of the degree to which the Responder had integrated the given information into his body of knowledge[1], using the following 5-point scale:

[1]Apart from acceptance rate, the repeats were also labeled in terms of Reception Rate (RR) ("To what extent did the speaker receive the sound of the repeated part?") and in terms of Communicative Intention (CI) ("Is the repeat mainly self-directed or other-directed?"). Since the repeats were mostly given the same RR and CI labels, these will not be discussed further in the present paper.

1 the speaker expresses that he is far from integrating the given information into his body of knowledge

2– the speaker expresses that he has some difficulty in integrating the given information

2 the speaker expresses that he has not yet integrated the given information, but is ready to do so

2+ the speaker expresses that he has almost integrated the given information, but not completely yet

3 the speaker expresses that he has fully integrated the given information into his body of knowledge

Due to the sparsity of the data, the current study did not explore to what extent all categories on the 5-points scale presented above are reflected in distinct prosodic and lexical features. The analysis was reduced to a two-fold distinction - thought to be very basic - between repeats that expressed integration and those that did not. The former comprised the categories 2+ and 3 of the original scale, the latter categories 1, 2– and 2. This yielded a distribution of 23 integrating and 48 non-integrating repeats.

5.3 Selected features

Both categorical and continuous variables were taken into account. The former were obtained by manual labeling, and comprised specifications of length category and boundary tone.

Length category Repeats were classified as to their lexical make-up, considering (i) whether they were completely identical to the repeated UU, (ii) whether they contained some additional lexical materials, (iii) whether they consisted of fewer words or (iv) whether they were paraphrases. Next, cases were specified that (v) only paraphrased part of a previous UU or (vi) more information than in the repeated UU. A final category (vi) comprised instances of multiple repeats. For examples, see Appendix A. Categories (iii) and (v) were later collapsed into one set ("short repeats") and (i), (ii), (iv) and (vi) into another ("long repeats"), whereas (vi) did not occur in the selected data.

Boundary tone Intonation of the repeats was labeled in terms of a slight variant of J-ToBI (Venditti, 1994) by an independent researcher who was not aware of the purpose of the current research. Focusing on the final boundary tones, there appeared to be one set of high-ending contours: the simple rise (H%) and the fall-rise (L%H%), and another set consisting of low boundary tones: the simple fall (L%) and the rise-fall (L%HL%).

Next to these categorical classifications, continuous features were obtained through automatic procedures, consisting of measures for pitch register, tempo, loudness and delay.

Pitch register Pitch register, which refers to the phenomenon that a speaker can utter his sentence in a rather low or high voice (Grosz and Hirschberg, 1992), was measured as the F_0 mean per utterance unit using Xwaves.

Tempo Normalized average mora duration per utterance unit was chosen as a measure of articulation rate. Using the transcriptions of the speech data, phone labels were first automatically time-aligned. After the phones were further grouped into a smaller set of morae, the normalized mora durations were calculated.

Loudness Loudness was defined as the measured energy, more specifically the mean RMS amplitude per utterance unit, as obtained from Xwaves.

Delay Delay was measured on the basis of the automatically obtained start and end times of the utterance units. In particular, the time distance was calculated between the offset of the repeated fragment and the onset of the repeating fragment. In this way, a large negative number reflects overlap, whereas a positive number a considerable delay.

To ease comparisons between prosodic variables and between speakers, all the prosodic features were normalized per speaker in terms of the distance of a give value from the mean in units of standard deviations.

6 Results

6.1 Descriptive analysis

Tables 2 and 3 give the distribution of integrating and non-integrating repeats as a function of type of boundary tones and length category, respectively. There appears to be some dependency of these categorical variables on the type of repeat, both distributions being statistically significant (Boundary tone: $\chi 2 = 4.094$ (df $= 1$, $p<0.05$); Length type: $\chi 2 = 4.802$ (df $= 1$, $p<0.05$)). There is a comparatively stronger preference for integrating repeats to be provided with a low boundary tone. The majority of non-integrating repeats also exhibit a low tone, but the relative frequency of high boundary tones is higher for this category than for the other. This finding is in agreement with the predictions. Regarding length category, the data allow one to choose between the two alternatives discussed in the hypotheses section regarding the lexical context of repeats: non-integrating repeats tend to focus on the problematic part of a previous utterance only, rather than being embedded in a longer turn which flags a communication problem with additional means.

Turning to the discussion of the continuous variables, the results for pitch, delay, average mora duration and energy are visualized in figures 2, 3, 4 and 5, respectively. It can be seen that low pitch, short delay and low energy are more often associated with integrating repeats. Conversely, higher pitch, long delay and high energy are more likely to reflect non-integrating repeats. T-tests revealed significant effects of utterance category on each of these features:

Table 2: Number of integrating and non-integrating repeats as a function of type of boundary tone

Boundary tone	Integrating	Non-integrating	Total
L%	19	28	47
H%	4	20	24
Total	23	48	71

Table 3: Number of integrating and non-integrating repeats as a function of length type

Length type	Integrating	Non-integrating	Total
Long	15	18	33
Short	8	30	38
Total	23	48	71

F_0 (T=-2.1969, df=69, p<0.05), delay (T=-2.5219, df=69, p<0.05) and energy (T=-2.3328, df=69, p<0.05). They are all in the expected direction, more prominent/marked features being more typical for the repeats that flag a (potential) communication problem. The difference in mora duration was not significant (T=-1.5248, df=69, *n.s.*). This may partly be due to the fact that temporal variation is already reserved for other functional purposes. Koiso *et al* (1997) report that acceleration and deceleration patterns are exploited by speakers of Japanese to signal opening versus non-opening of new information units.

6.2 Information potential

To analyse the signaling value of the different features in relation to integration level, the repeats were further explored in terms of two concepts borrowed from information-theoretical science, i.e., comprehensiveness and accuracy. The former is a measure of the coverage of the information signaled, the latter refers to the correctness of the signaling (see also Koiso *et al.*, 1997). In a way, accuracy and comprehensiveness reflect the perspectives of the traditional participants in the communication chain, the speaker and the listener, in the sense that the first indicates to what extent prosodic and lexical features can be predicted from type of integration, whereas the second gives the probability of integration level given a set of prosodic and lexical features.

Specifically, accuracy and comprehensiveness were studied in relation to"markedness", defined as the different feature settings which are not default and therefore (highly) promi-

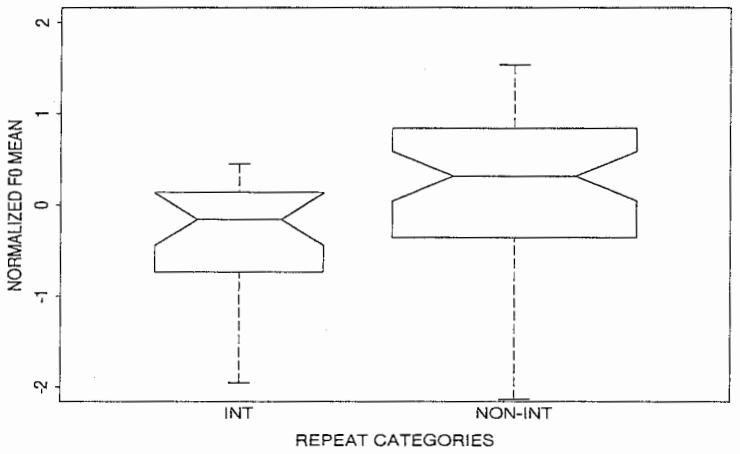

Figure 2: Normalized F_0 mean per utterance category

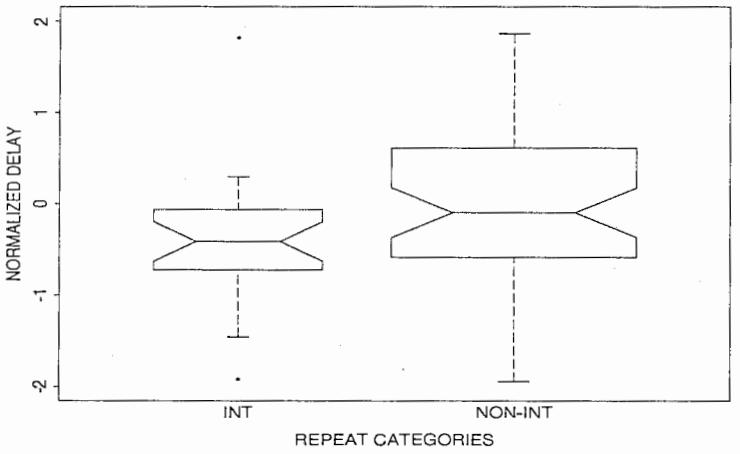

Figure 3: Normalized delay per utterance category

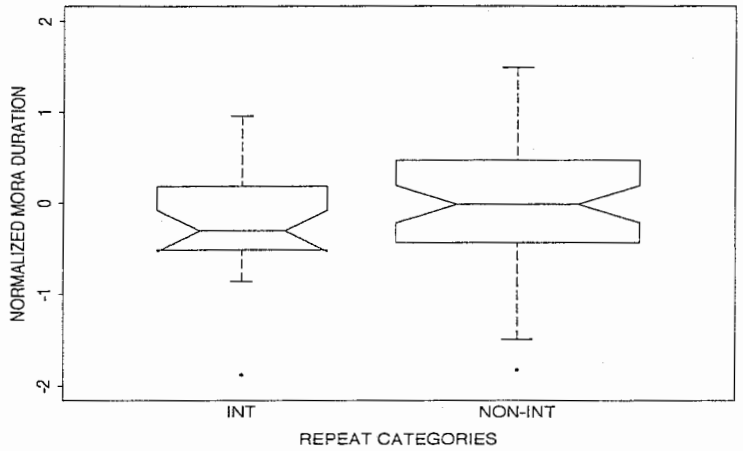

Figure 4: Normalized mora duration per utterance category

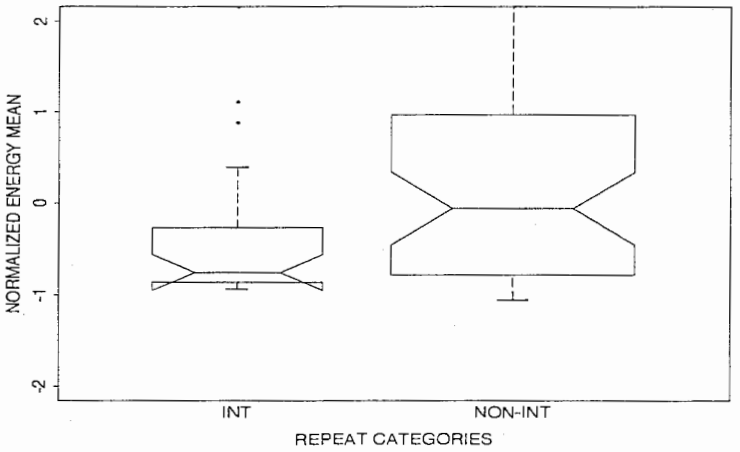

Figure 5: Normalized energy mean per utterance category

Table 4: Signal comprehensiveness of non-integrating repeats as a function of the presence of one or more marked prosodic features

Features			Classification	
Boundary Tone	Length	Continuous variable	Comprehensiveness	Chance
H%	Short	> 0 sd	100%	91.5%
H%	Short	> 0.5 sd	80.3%	89.6%
H%	Short	> 1 sd	79.1%	70.4%

nent (see hypothesis section). For the categorical variables, that relates to the H% type of boundary tones and to the short length category. For the current analysis, three degrees of markedness were specified for the continuous variables, the weakest being the cases where the values were simply higher than average, the middle where values were at least 0.5 standard deviation above average, and the strongest where values were at least 1 standard deviation above average. As already remarked above, normalization of the different prosodic features facilitated the comparative analysis of the effect of these variables on level of integration.

A quantitative formula for comprehensiveness is given in (1)

$$\frac{NonInt^{+}}{NonInt} \tag{1}$$

where $^{+}$ refers to the presence of at least one marked feature. A signal would be completely comprehensible if there would be no integrating repeats having marked features. Chance level for comprehensiveness is computed as the number of all repeats having one or more marked prosodic features divided by all repeats.

The equation for accuracy is given in (2)

$$\frac{NonInt^{+}}{NonInt^{+} + Int^{+}} \tag{2}$$

where $^{+}$ is defined in the same way as before. A signal would be completely accurate if all the non-integrating repeats would have at least one marked feature. Chance level for accuracy is operationalized as the number of non-integrating repeats divided by all repeats.

Given these definitions, the results for comprehensiveness and accuracy for level of integration are given in Tables 4 and 5, respectively. As can be seen, the features analysed in this study appear to have information potential, since both measures are always above chance level, irrespective of the degree of markedness of the continuous features. There is also a logical trade-off between the two measures, as accuracy increases when comprehensiveness decreases, and vice versa, as a function of the degree of markedness.

Table 5: Signal accuracy of non-integrating repeats as a function of the presence of one or more marked prosodic features

Features			Classification	
Boundary Tone	Length	Continuous variable	Accuracy	Chance
H%	Short	> 0 sd	73.8%	67.6%
H%	Short	> 0.5 sd	75.4%	67.6%
H%	Short	> 1 sd	76.0%	67.6%

7 Discussion and conclusion

Summarizing the results of this investigation, it appears that echoing utterances in Japanese, informal dialogues may serve at least two distinct communicative goals: to signal that information has been integrated succesfully by the Responder or to express the fact that he has some difficulty in integrating it in his body of knowledge. Phonetic measurements reveal that these repeat categories are reflected in different prosodic and lexical features: in agreement with what could be expected from the literature, the non-integrating cases are more likely to have one or more marked prosodic variables. Explorations of their information potential brought to light that these features have significant signal capacity in terms of accuracy and comprehensiveness. From these results, it can be concluded that repeats are potentially useful in spoken communication because they represent different dialogue management acts: they function as different types of "evidence" from the Responder about the information being presented by the Initiator.

This leads one to reflect on the differences between human-human and human-machine interaction. In the former, all the conversants are very well aware of each other's limited resources, which is evidenced by the fact that they constantly seek and provide evidence about mutual mental beliefs (Brennan, 1990; Walker, 1992). Repeating is a clear example of such communicative behaviour. In this perspective, it seems unrealistic to expect from spoken dialogue systems that they would be able to act as "perfect communication partners" that can achieve errorless understanding, since there will always be types of noise that are too severe to be solved by machines. Alternatively, in order to make the interaction with spoken dialogue systems more efficient, it might be worthwhile to model particular strategies that are typical of human-human interaction, such as repeating, which have been proven to be useful to handle the intrinsic uncertainty of spoken communication.

Obviously, the research needs to be extended in a number of ways. First, one might consider in a larger corpus of dialogues whether the results generalize to other speech materials. Specifically, repeats may be interesting from a cross-linguistic point of view, given that other languages may exhibit different conversational rules, e.g. in terms of politeness

principles or in the degree to which repeating is socially accepted (Tannen, 1989). Similarly, one could expect that the structure of repeating varies as a function of the discourse setting, e.g. mother-child interaction being very different from a classroom situation. Second, the dialogue act classification presented here needs to be evaluated experimentally to see to what extent the proposed labels are reproducible.

Acknowledgments

Sponsored as a postdoc with the Flemish Fund for Scientic Research (FWO - Flanders), Marc Swerts is also affiliated with the Center for Research on User–System Interaction (IPO) and with Antwerp University (UIA). The research reported upon in this paper was carried out while he was a visiting researcher at department 2 of ATR-ITL in summer 1997. Marc Swerts would like to thank Nick Campbell and Norio Higuchi for giving him the opportunity to visit their lab, and Nick Campbell for his constant help. Further thanks to Harald Singer, Kyohko Shimoda, Satoshi Kitagawa, Miwako Kurihara and Yoshinori Michijiri for technical assistance and help with labeling of prosody and dialogue structure.

References

S. Brennan (1990): Seeking and providing evidence for mutual understanding. PhD thesis. Stanford University.

H.H. Clark and E.F. Schaefer (1989): Contributing to discourse. Cognitive science 13, pp. 259-294.

H.H. Clark and D. Wilkes-Gibbs (1986): Referring as a collaborative process. Cognition 22, pp. 1-39.

E. Couper-Kuhlen (1996): The prosody of repetition: on quoting and mimicry. in E. Couper-Kuhlen and M. Selting (Eds.): Prosody in conversation, pp. 366-405, Cambridge: Cambridge University Press.

L. Fais (1994): Conversation as collaboration: Some syntactic evidence. Speech communication 15, pp. 231-242.

B. Grosz and J. Hirschberg (1992): Some intonational characteristics of discourse structure. Proc. ICSLP, Banff 1992, pp. 429-432.

E.O. Keenan and B.B. Schieffelin (1976): Topic as a discourse notion: a study of topic in the conversation of children and adults. in C.N. Li (Ed.): Subject and topic, pp. 335-384, New York: Academic Press.

H. Koiso, A. Shimojima and Y. Katagiri (1997): Informational potentials of dynamic speech rate in dialogue. In Proc. of the Nineteenth Annual Conference of the Cognitive Science Society, August, pp. 394-399.

E. Ochs (1979): Planned and unplanned discourse. in T. Givon (Ed.): Syntax and semantics 12: Discourse and syntax, pp. 51-80, New York: Academic Press.

A.-B. Stenström (1994): An introduction to spoken interaction. London, New York: Longman.

M. Swerts and M. Ostendorf (1997): Prosodic and lexical indications of discourse structure in human-machine interactions. Speech Communication 22 (1), pp. 25-41.

D. Tannen (1989): Talking voices: repetition, dialogue, and imagery in conversational discourse. Cambridge: Cambridge University Press.

C. Tarplee (1996): Working on young children's utterances: prosodic aspects of repetition during picture labelling. in E. Couper-Kuhlen and M. Selting (Eds.): Prosody in conversation, pp. 406-435, Cambridge: Cambridge University Press.

D. Traum (1994): A computational theory of grounding in natural language conversation. Unpublished Ph.D. thesis, University of Rochester

J.J. Venditti (1995): Japanese ToBI labelling guides. Technical report. Ohio State University.

M.A. Walker (1992): Redundancy in collaborative dialogue. Fourteenth International Conference on Computational Linguistics, pp. 345-351.

Appendix A:
Examples of repeats in different dialogue contexts
(<...> refers to repeated fragment, [...] to repeating fragment)

Type	Time	Speaker	Utterance
equal	00:35:296-00:35:984	F:	<megamieru>
	00:35:792-00:36:432	G:	[megamieru]
subset	00:27:104-00:27:872	F:	orenjino<hananoyatsuwo>
	00:27:648-00:28:464	G:	[hananoyatsuwo]
additional	00:42:672-00:43:280	G:	a<nosankakkei>noyatsuwo
	00:45:519-00:46:048	F:	naniiro[nosankakkei]
paraphrase	01:39:216-01:39:920	F:	<akaarimasu>
	01:40:160-01:40:640	G:	[akaaru]
paraphrase-part	04:48:832-04:50:112	G:	<tokkigadeteruhouwoshita>nishite
	04:51:060-04:52:448	F:	[tokkigadeteruhoushita]
paraphrase-add	00:53:264-00:53:568	G:	<chigau>
	00:53:536-00:53:728	F:	[chau]no