TR-I-0267

Discourse Management Mechanism on Spoken Dialogue Processing (in an MT System)

Susann LuperFoy

1992. 6

Abstract

Discourse research can contribute to the Interpreting Telephony (IT) enterprises in two ways: (1) externally - it provides the terminology for appropriate description of the IT dialogue task, a way of defining the properties of the IT system as a tool for managing a highly specialized type of interpersonal interaction, and (2) internally - it provides the computational techniques for integrating a discourse processing module with other components of the IT software system, such as Speech Recognition (SR) and Natural Language (NL) Analysis. I will discuss how to define the essential discourse properties of the IT system and how to approach solutions at the discourse level. Particular attention is given to the possible modes for conveying information to the SR module.

ATR自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

1. Introduction *

Discourse research can contribute to the Interpreting Telephony (IT) enterprises in two ways: (1) externally - it provides the terminology for appropriate description of the IT dialogue task, a way of defining the properties of the IT system as a tool for managing a highly specialized type of interpersonal interaction, and (2) internally - it provides the computational techniques for integrating a discourse processing module with other components of the IT software system, such as Speech Recognition (SR) and Natural Language (NL) Analysis. I will discuss how to define the essential discourse properties of the IT system and how to approach solutions at the discourse level. Particular attention is given to the possible modes for conveying information to the SR module.

---

*) This report summarizes a three-month investigation into the dialogue aspects of the interpreting telephony task. The form of this document is a set of annotated slides taken from my summary presentation to IT staff of ATR, on 2 June 1992.

Overview

Dynamic, inexpensive discourse processing

Dialogue aspects of IT
(external properties)

Dialogue management with other IT modules
(internal properties)

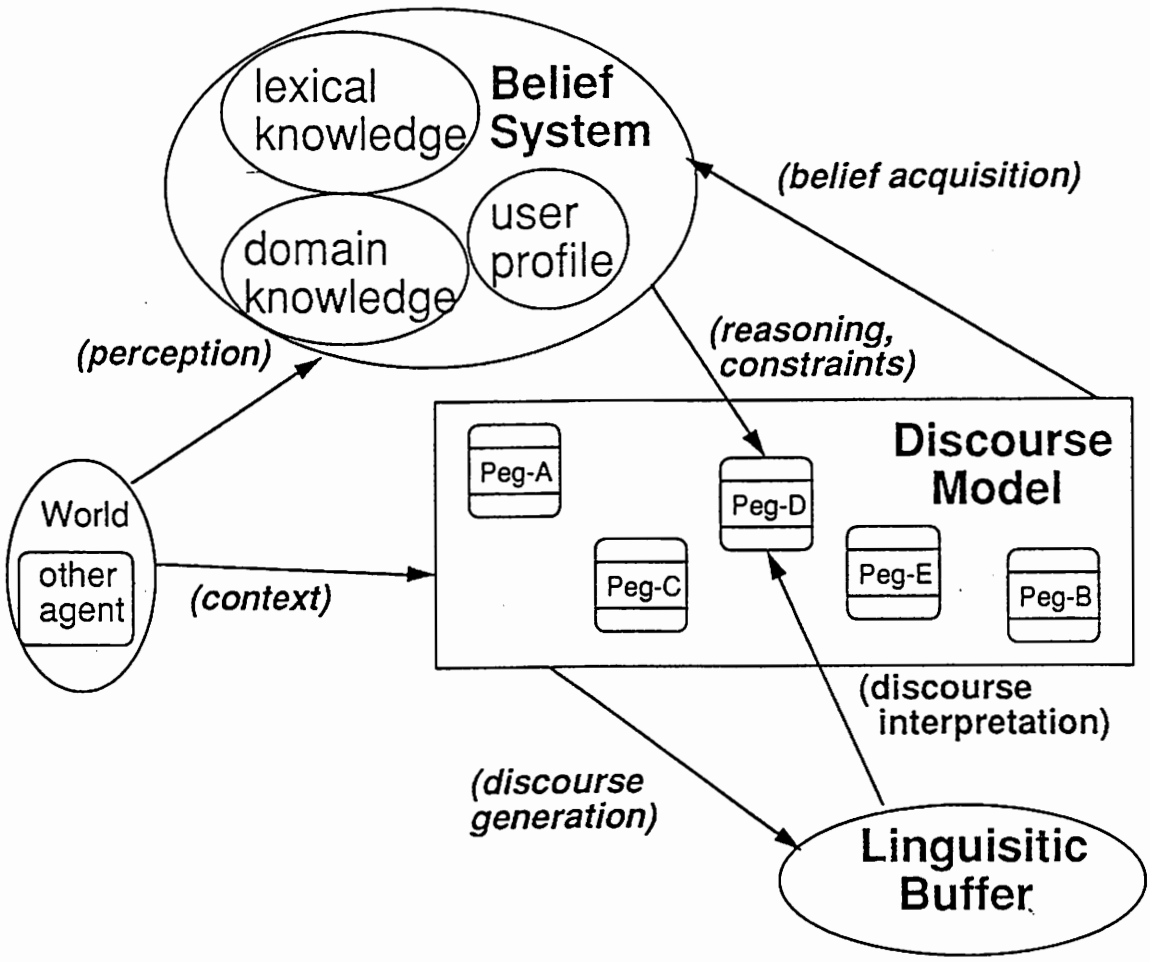
Detailed Symbolic Information Dialogue Segments (DSIDS)

2. This presentation is organized around four issues. First, the guiding purpose behind my research on discourse within the IT project and elsewhere is a search for inexpensive techniques for dynamically incorporating discourse-level information into NL processing. MT researchers acknowledge the importance of discourse information, but there exists an understandable reluctance to incur the computational expense of constructing representations and defining processes for detecting, extracting from, and later consulting, abstract discourse relations reflected in spoken language. I offer an approach to discourse representation and processing that defines the simplest set of information required by the IT task and at the minimum computational cost. I will not claim adequacy for this approach but rely on intuitive arguments that the representation of referring expressions is an aspect of any discourse analysis system.

The second issue raised is that of defining the "external" discourse properties of the IT system. How do we expect the interpreting telephone of the future to behave as a discourse tool? what is its role in the interpersonal task to which it is applied? what is realistic to expect of its users as the two human participants of a three-way conversation?

In contrast, investigation of the "internal" discourse properties of the same system is essential for constraining the design of the discourse component as one of several software modules cooperating on the IT project. We should identify the unique properties of this MT task that will both allow for useful simplifying assumptions and present necessary challenges peculiar to this system. I will give several examples of ways for the discourse component to interact with SR, NL Analysis, NL Generation, and so on.

Finally, as an illustration of the proposed model for conveyance of discourse information to SR I present an analysis of one type of discourse segment, the Detailed Symbolic Information Dialogue Segment (DSIDS). This phenomenon occurs frequently in telephone dialogue where interactive information solicitation is the primary objective, as opposed to friendly chats, sales calls, etc. DSIDS pose unique problems for speech recognizers in general and the incorporation of discourse knowledge is essential to handling these.



3. This diagram is taken from my introductory talk to ATR on 17 March outlining the approach to discourse representation advocated in my prior research. I have argued for separating the information available to the computational discourse system based on (1) the sort of information it is, (2) the operations that create, access, and update the information, (3) the purpose served by the information in the larger language-processing task, and (4) the rate of decay and the operations that cause information to decay.

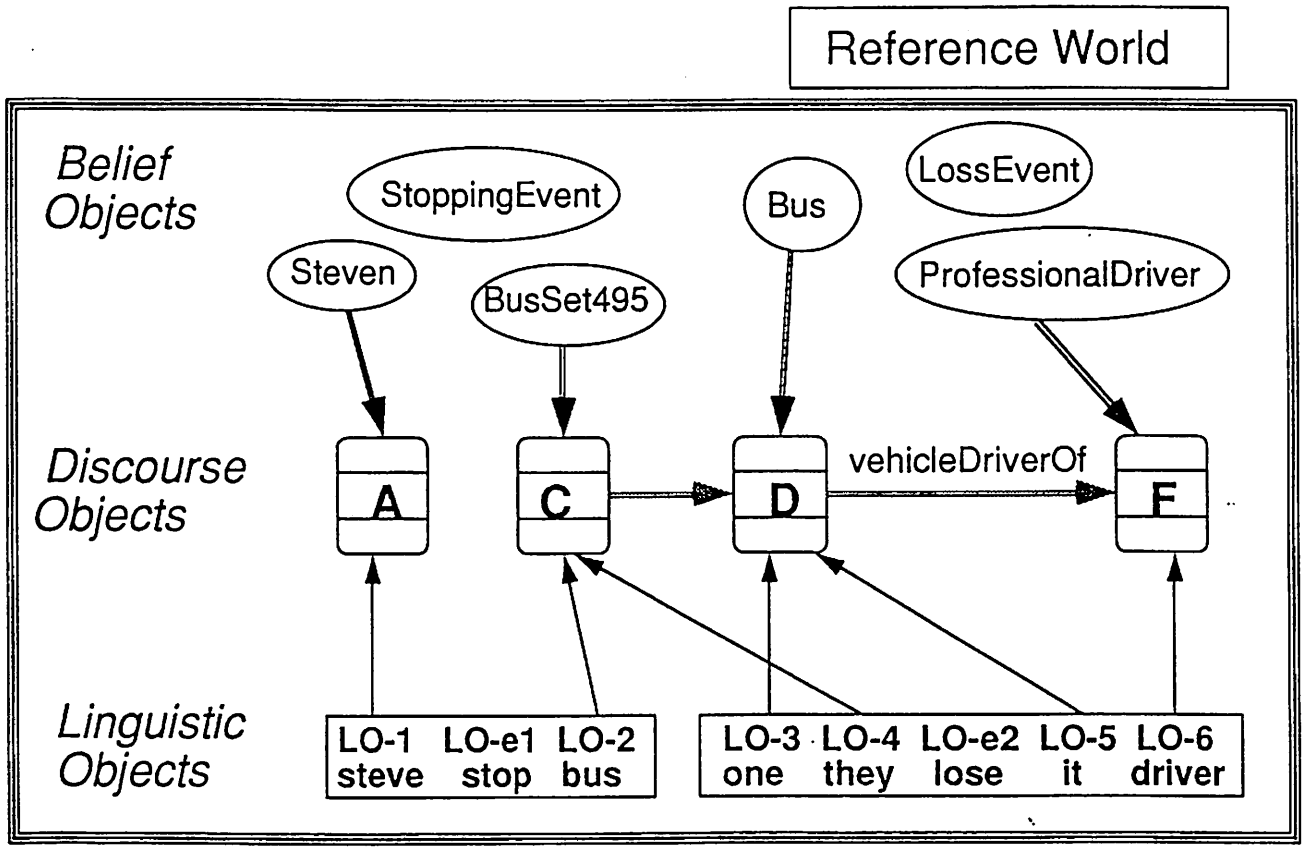
Information in the "Belief System" region (or KB) is generally expensive to create, store, update and access, and should therefore be assumed by the system design only as absolutely necessary to the NL task being undertaken. Discourse information is held in an intermediate tier as the information content of the ongoing discourse, regardless of

(a) its truth with respect to the world of reference
(b) agreement on the part of the other agent (in the case of dialogue discourse)
(c) the match between this information and the beliefs of the interpreter

This distinction between the belief system and the reprsentation of discourse is important for the representation of real discourse because people are able to produce meaningful utterances that aren't true, that their conversational partners don't agree with, or that they themselves don't believe to be true.

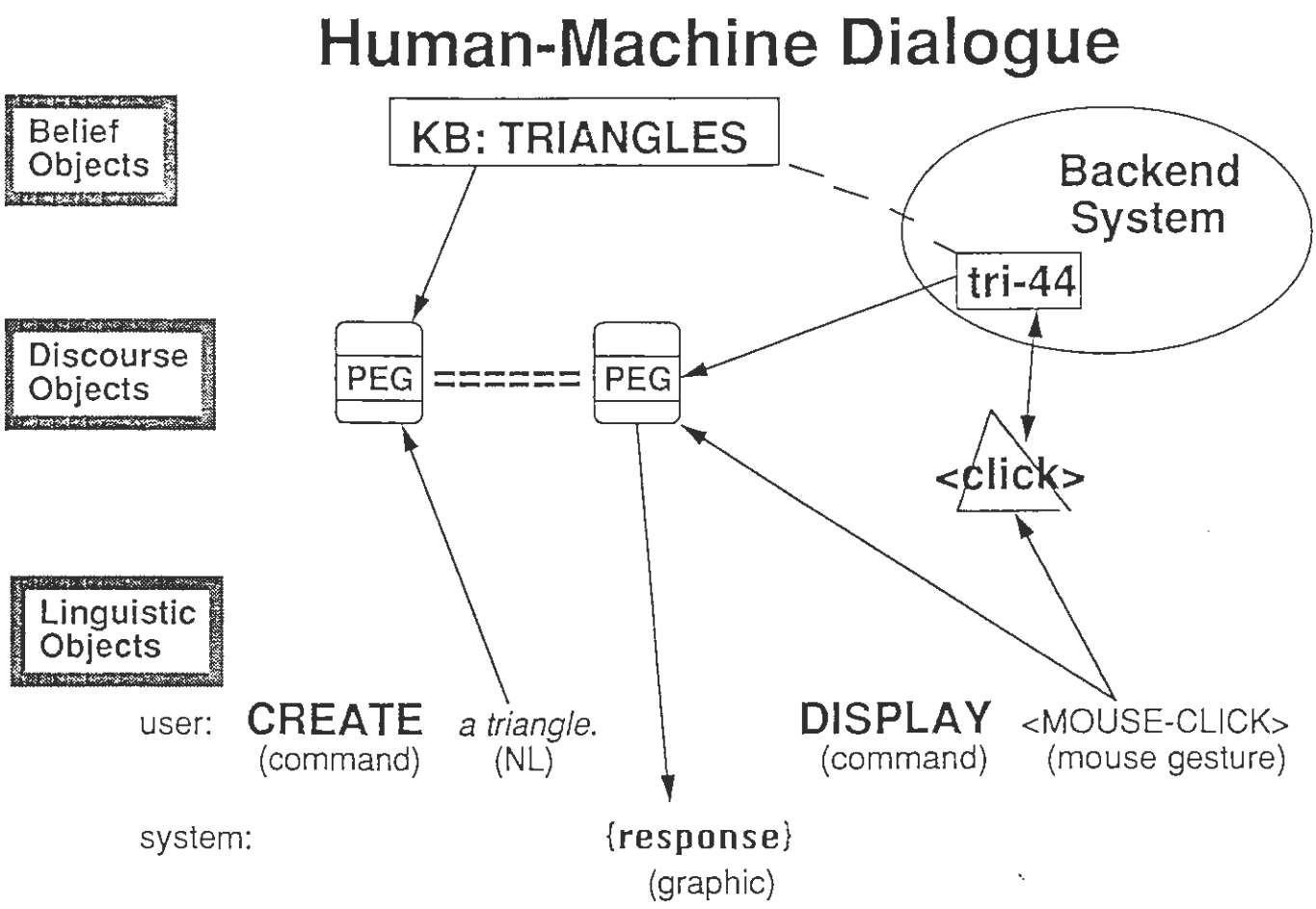
Links between regions on this diagram denote the various operations that relate the three distinct information sources. This partitioning allows us to distinguish understanding a discourse (discourse interpretation) from incorporating into a belief system the information content of a discourse (belief acquisition); it allows for reasoning about discourse information without first incorporating the propositional content of the discourse into the belief system, and comparison of that information to the more stable belief system information.

The 3-Tiered Discourse Structure



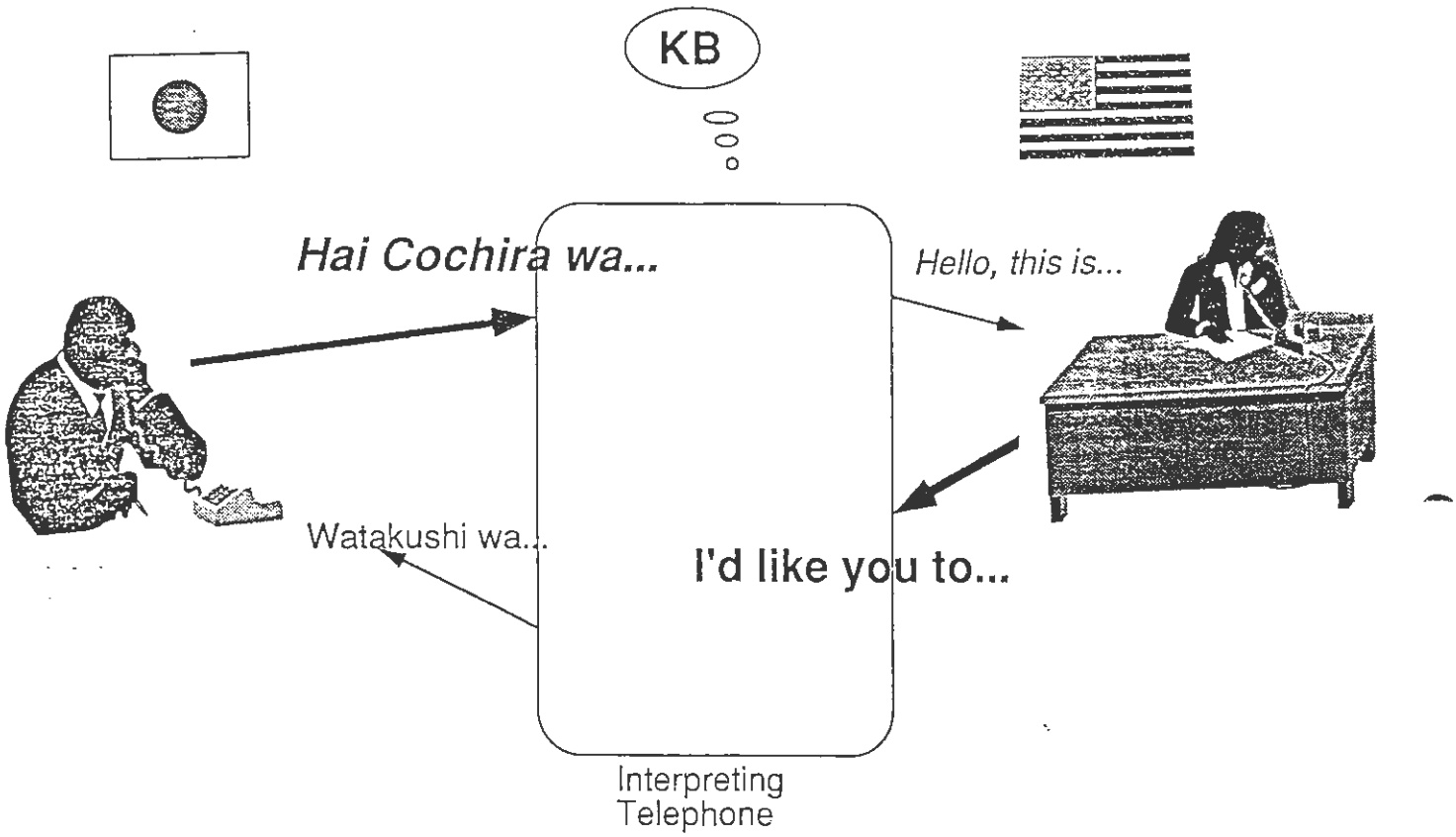
Input Speech Steve-1 stopped the buses-2 because one-3 of them-4 lost its-5 driver-6

4. The 3-tiered representation applies to discourse in general, dialogue as well as monologue, spoken discourse as well as text, and ndependent of the input language. Because the discourse and KB are maintained separately the discourse interpreter need not have knowledge (a stored prior mental representation) of the entities under discussion in order to process the discussion. The discourse tier is also separate from the linguistic tier. My prior research (and my introductory presentation to ATR) gives some justifications for this bifurcation on the basis of anaphoric phenomena in English. The IT project makes the need for this separation especially clear in part due to the need to maintain dual discourse histories for the two alternating input languages. An essential claim of the framework is that the way a peg gets introduced into a discourse is independent of its subsequent behavior in the discourse model. The example on this diagram shows a segment of an extended English monologue, with anaphoric relations represented as links between LO's and Discourse Pegs. The next slide shows an application of the three-tiered framework to multimodal user interface dialogue.



5. In a multimodal Human-Computer Interface (HCI) dialogue a peg in the discourse model is independent of the modality used to introduce it into the discourse and independent of which speaker mentioned it, the human user or the backend software system. The various modalities may be intermediated by separate devices and software systems, and may get processed by different semantic interpreters, but the discourse model provides a level for uniform representation of the diverse forms of surface information. Both input and output contribute to the discourse model, and the two are distinguished at the linguistic tier level to keep straight what each 'speaker' said about each discourse peg.

The IT Dialogue Setting



6. This is a sketch of the IT dialogue scenario. There are two speakers, each with a belief system (bubbles above the diagrams of the figures) that is inaccessible to the other speaker and to the IT system. The IT system may or may not have its own stored set of beliefs (a KB) about the domain world or about the two speakers but it would be computationally expensive for it to construct and maintain models of the two users for the purposes of a brief telephone conversation. I will return to this diagram as we discuss the external properties of the IT system in operation.

The Interpreting Telephony Dialogue

speech synthesis (SS) and speech recognition (SR)

NL analysis

NL generation

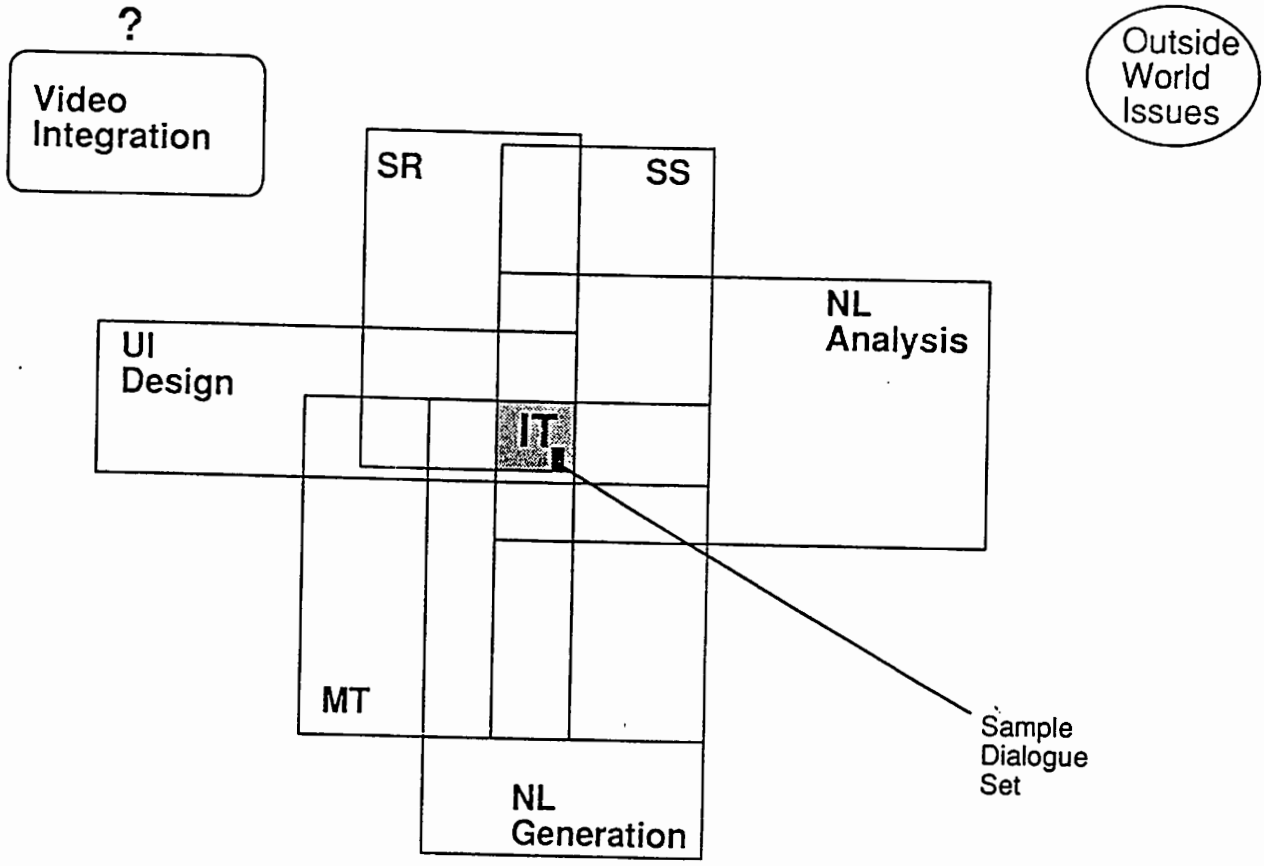
MT

human-human dialogue management

human-machine dialogue management (UI design)

7. This research project involves a unique and very rich combination of interesting computational linguistics research endeavors as the component technologies of the eventual IT system. Speech synthesis (SS) and speech recognition (SR), MT in general, NL analysis and NL generation (both essential components even assuming the EBMT approach to translation) are all being actively pursued at ATR. We must also explore the properties of the human-human dialogues that constitute our chosen language data to be analyzed, translated, and generated, making our data rare among MT projects. And I propose that serious attention be addressed to the human-machine dialogue or (User Interface (UI) design) aspects of this project, a topic to be returned to later in this summary. Obviously, this collection of research projects is too much to be pursued full-scale even for a group of our size. Therefore, what we, like all other research teams, have done is selected a small portion of the problem defined by the intersection of all of these projects via the sample dialogue sets selected as our benchmark test set. The sample set of utterances and the application itself-- registering for an international conference by telephone through free-formed dialogue-- may seem quite contrived, however a system that can process them, if judiciously conceived, will scale up to a general question-answer dialogue management system.

Dialogue Research Efforts



8. I take a brief moment here to argue for very conscious attention to the preliminary task of merely defining our long-term and short-term research enterprise. The first motivation is to be able to determine which problems are of immediate concern and which ones can be delayed. Secondly, we must be able to clearly communicate our immediate and long-term objectives and our accomplishments to the rest of the research community.

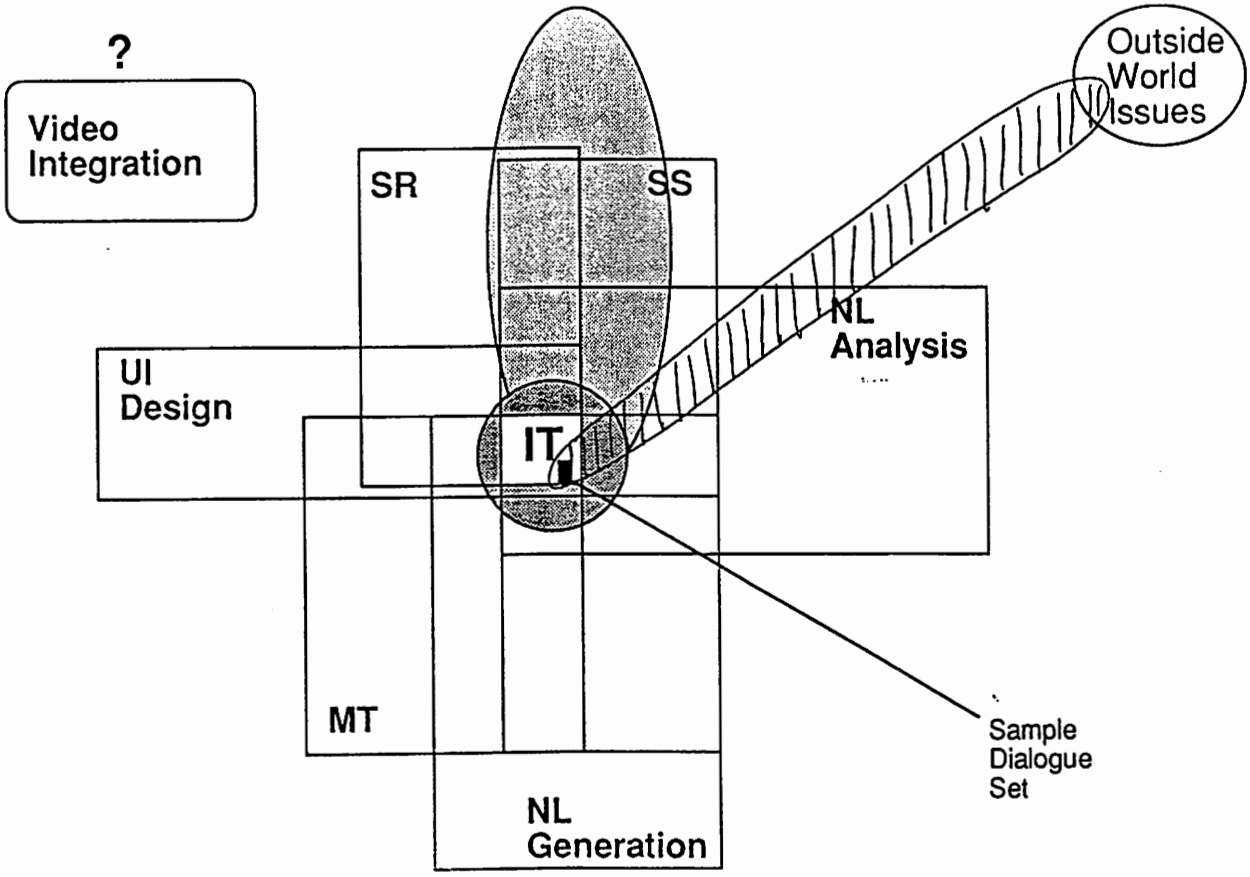
Under the first motivation of analyzing the system under design, a clearly defined goal yields three main benefits:

(1) we are able to define clearly the ways in which the component technologies interact and allow these definitions to constrain our efforts in each of the sub-projects. We want to be sure that our short-term efforts contribute to future work on a course toward the ultimate goal. We also want to be sure that our work represents a balanced sample from the important technologies, i.e., so that we are not spending a disproportionate amount of our energy on NL analysis while ignoring important MT, SR, NL generation, and UI design issues.

(2) We can be sure that we are targeting essential problems of the IT system and not external world tasks such as the conversion of time zone information (see Slide #28 for illustration).

(3) We have an idea of what it would mean to alter some aspect of the system, for example if we were to add video to the IT system what changes would be entailed for speech recognition, for translation, or for representation of the discourse interaction.

Dialogue Research Efforts



9. To illustrate what an unbalanced effort would be consider the case where we apply a disproportionate amount of our resources to speech synthesis research, including problems that don't contribute to the IT task nor even interact with other component technologies. While these are worthy research efforts they may act as distractions from the most efficient path to the IT system. The worst case would be to find ourselves expending valuable research effort on a problem that is not within the realm of the IT task even in the distant future, such as the time zone conversion problem. Our target should be a configuration of research projects represented as a circle around the shaded IT box on this venn diagram such that some of our effort contributes to the component fields in general, MT, NLP, SS, and so on but that most results are directly applicable to the IT task.

Effects of Video

New Data
- Visually introduced pegs (indexicals)
- Hand and facial gestures

New Analysis of Old Data
- Face-to-face "hello" doesn't translate to "moshi moshi"
- Prosody and non-verbal speech used differently

Audio and visual data coordination
- laughter, coughs, facial expressions
- delay for MT processing of speech data

New task may call for entirely different solution
- video with subtitles
- still image with speech

10. It is also important to examine closely any suggested change to the stated goals of the project. For example, before we seriously consider the introduction of a video channel to the IT system, we must decide how that modifies the overall IT project definition and our task as researchers on one or more of the component technologies.

• One change with the arrival of video to the IT system is that information can then be introduced into the discourse context through the visual channel. This includes hand and facial gestures, and deictic pointing to objects or parts of objects in the shared visual field, none of which our system at present is designed to address.

• A second area of change is a modified analysis of data we do currently handle. For example, the English greeting "hello" translates as "moshi-moshi" in telephone dialogues only. This is because of its function in the telephone dialogue as an opening or as a meta-request when the speaker has reason to believe the connection has been lost. In those cases "hello" and "moshi-moshi" are the corresponding telephone dialogue devices of the respective language communities. However, the video channel modifies the dialogue to something more like face-to-face dialogue in which case "hello" translates as something like "konnichi-wa, " or "ohiogozaimasu." What is retained in the video-added system is the uniqueness of the bilingual dialogue setting, a situation with a very specialized set of interaction rules, and one that people in general have little exposure to prior to this technology.

• A third consideration is the way the technology might alter human behavior. People behave differently when they know that their entire message must be compressed into the spoken channel over a telephone line. When we add face-to-face visual information we can expect the linguistic behavior itself to change. For example, we might expect more interruptions and a greater tolerance for interruption creating overlapping speech that must then be handled by the SR component. One precaution against this would be an overt indication from the IT system of which speaker is being attended to at each moment, with the understanding that the other speaker is being ignored by the SR system. However, presumably that listener will be transmitting visual information to the speaker perpetually so many decisions must be

made concerning whether to interpret visual information or just pass it through the system. Speakers may also show changes in the way they use prosodic information to convey intentions.

Visual information such as facial expressions and pointing gestures must also be coordinated in some fashion with sounds, laughter, coughs, silent pauses, and with production of the translated linguistic output. For example, a speaker's intentions may be uninterpretable to the other human if laughter is 'passed through' the system immediately while simultaneous speech data get delayed for processing through the various layers of the IT system.

- Finally, we can expect the introduction of video to require a whole new line of research into how the various modalities are to be coordinated in order to accomplish the IT task. Such a thorough system analysis may reveal substantial changes to the research project. For example, depending on the stated purpose of the video link, it may turn out that video with subtitles yields the most natural system, or perhaps a still image with speech might accomplish the goal most effectively, in which case the need for NL research change dramatically.

Desiderata

External Properties:

Repair dialogues between humans
Robust and Fast
Variety of dialogue segment types
Two speakers, changing roles
Language-specific dialogues must be coherent

Internal Properties:

Dialogue cooperates with SR/SG
Dialogue cooperates with MT
Don't need to resolve references (cf. Voyager)
Must supply pegs for deixis, ellipsis, and anaphora
Detailed Information Dialogue Segments (DIDS)
Interactive:
(1) must take action
(2) cannot unwind

11. This slide presents a set of desiderata that will be assumed for the remainder of the discussion. I mentioned earlier the external discourse properties of the IT system having to do with how the system functions as the mediator of interpersonal interaction between two monolingual speakers. Some desirable properties that I observe (and you are probably aware of many others) are itemized here:

• The two human speakers will not become passive but will actively engage each other in various sorts of dialogue exchanges such as requesting confirmation or repair of a misunderstanding. That is, the two speakers will apply interpersonal communication techniques to help manage their IT dialogue.

• The system must be robust and fast in order to serve the needs of a user who has chosen the IT system over other slower, less interactive options such as written documents or electronic mail. The human user must be able to rely on the IT system to properly translate the utterances on both sides.

• A single IT conversation will be composed of dialogue segments of different sorts; there will be question-reply segments, meta-dialogues for repair and clarification, conversational openings and closings, interruptions, and so on. The dialogue manager must be able to recognize these segment shifts and adapt its behavior.

• The IT system will have to manage the interaction and the alternation of speaker-hearer roles between the two users, making it clear to both which of the two is being attended to at each moment.

• Each of the two human users experiences a monolingual dialogue in their own language. Each will behave according to the linguistic and dialogue interaction rules for that language. This means that the two dialogues must be internally coherent and obey the conventions of the respective languages. So the IT system must be competent in the discourse rules and conventions of both languages and must be able to use cues from one to assist in translation to or from the other.

The desired internal properties of the dialogue manager as a module in the larger IT system include

• The dialogue manager must cooperate with speech recognition and speech generation since (a) the dialogue manager will receive its input from the SR via NL analysis, and (b) there will be some distinctions in SR that can only be determined at the discourse level such as the relative felicity in the current discourse of the various hypotheses returned by SR, and SR can benefit from information available only at the discourse level.

• The discourse processor should also provide information to the MT process, for example many cases of lexical mismatch can only be sorted out with the use of discourse contextual information

• Because the task is translation, full reference resolution is not essential; In contrast to spoken interfaces to robots (DeMori, et al.) and other speech UI systems (Zue, et al.) there may be no strong need for a knowledge base to serve as the foundation of reference resolution.

• Nonetheless, the discourse model must contain representational objects (discourse pegs) for handling context-dependent reference, with zero-pronouns in Japanese being an example of this challenge. However, there needn't be a link to a knowledge base from the objects in the discourse model in order for MT to succeed.

• Speech recognition in DSIDS (Detailed Symbolic Information Dialogue Segments) and other specialized dialogue segments must be supported by the discourse system.

• In this, and all interactive systems (versus non-interactive document processing systems) when faced with inadequate information to process the current utterance, the dialogue manager in an interactive system can neither

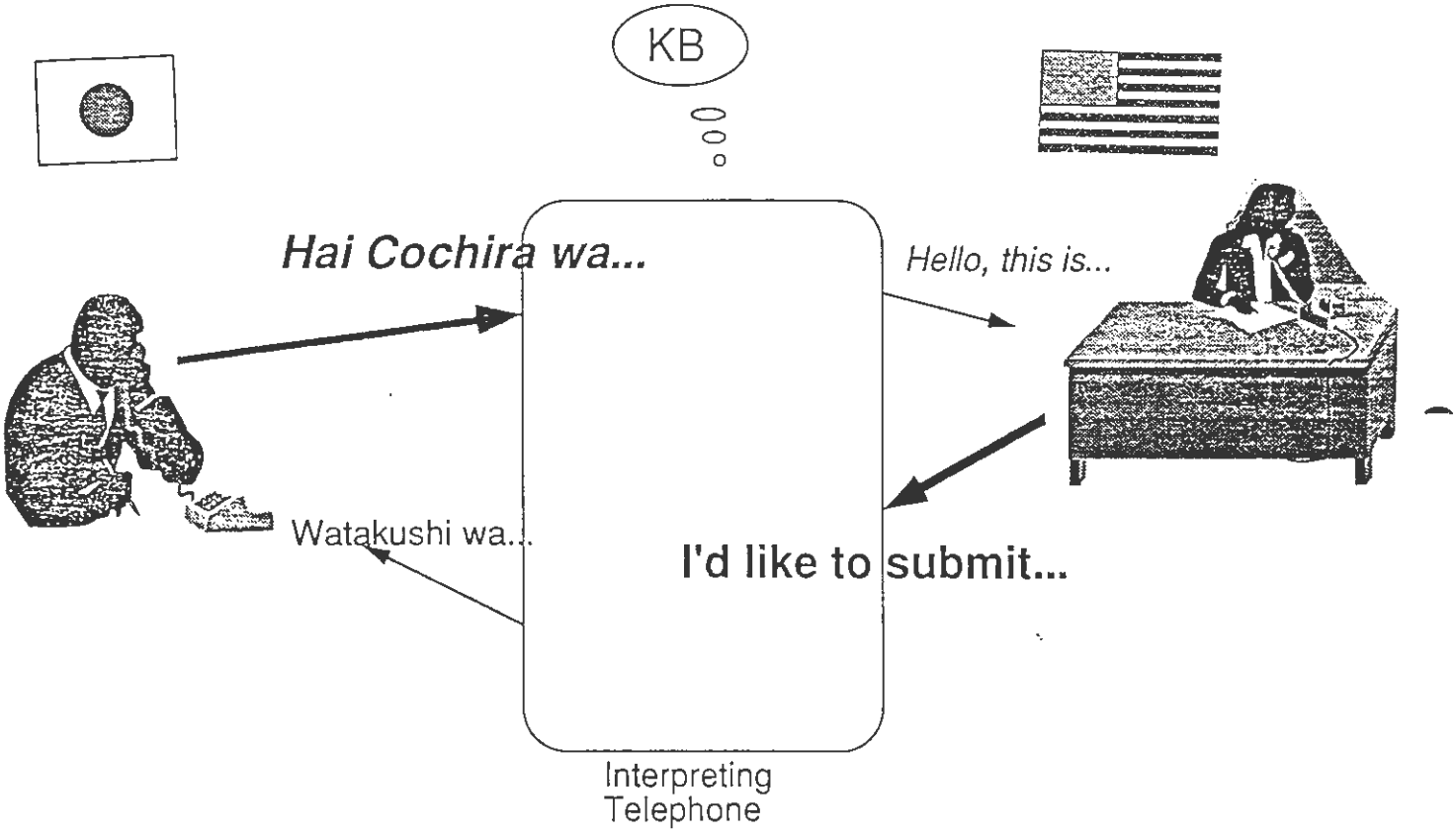
(1) delay until complete information emerges, because action must be taken immediately, i.e., even if there is not certainty in the system's output, some result must be produced so that the dialogue can continue, whereas in a text interpretation system the discourse processor could be defined to handle uncertainty by delaying or looking ahead to a subsequent location of the text for disambiguating information,

nor

(2) count on 'unwinding' later to return to an earlier state of discourse processing. Once an utterance has been produced in a dialogue with human users, it cannot be retracted by the system.

The conclusion is that in this interactive system all uncertainty must be assessed and dealt with at the point where it occurs with the awareness that repairs may be required at some later point in the dialogue.

Bilingual Dialogue 'Eaves-Dropper'



12. External and internal properties for the IT system suggest a metaphor for the IT-mediated bilingual dialogue. Returning to the IT diagram we can examine three possible metaphors for the interaction, including the proposed "Bilingual Dialogue Eaves-Dropper" ("nesumigiki") model.

Three Metaphors

Dialogue Interpreter
full reference resolution
too costly
not necessary
can lead to errors

Dialogue Translator
quick and inexpensive
not adequate (zero pronouns, ellipsis, repairs)

Bilingual Dialogue "Eaves-Dropper'
moderate cost
handles context-dependence
similar to operating system

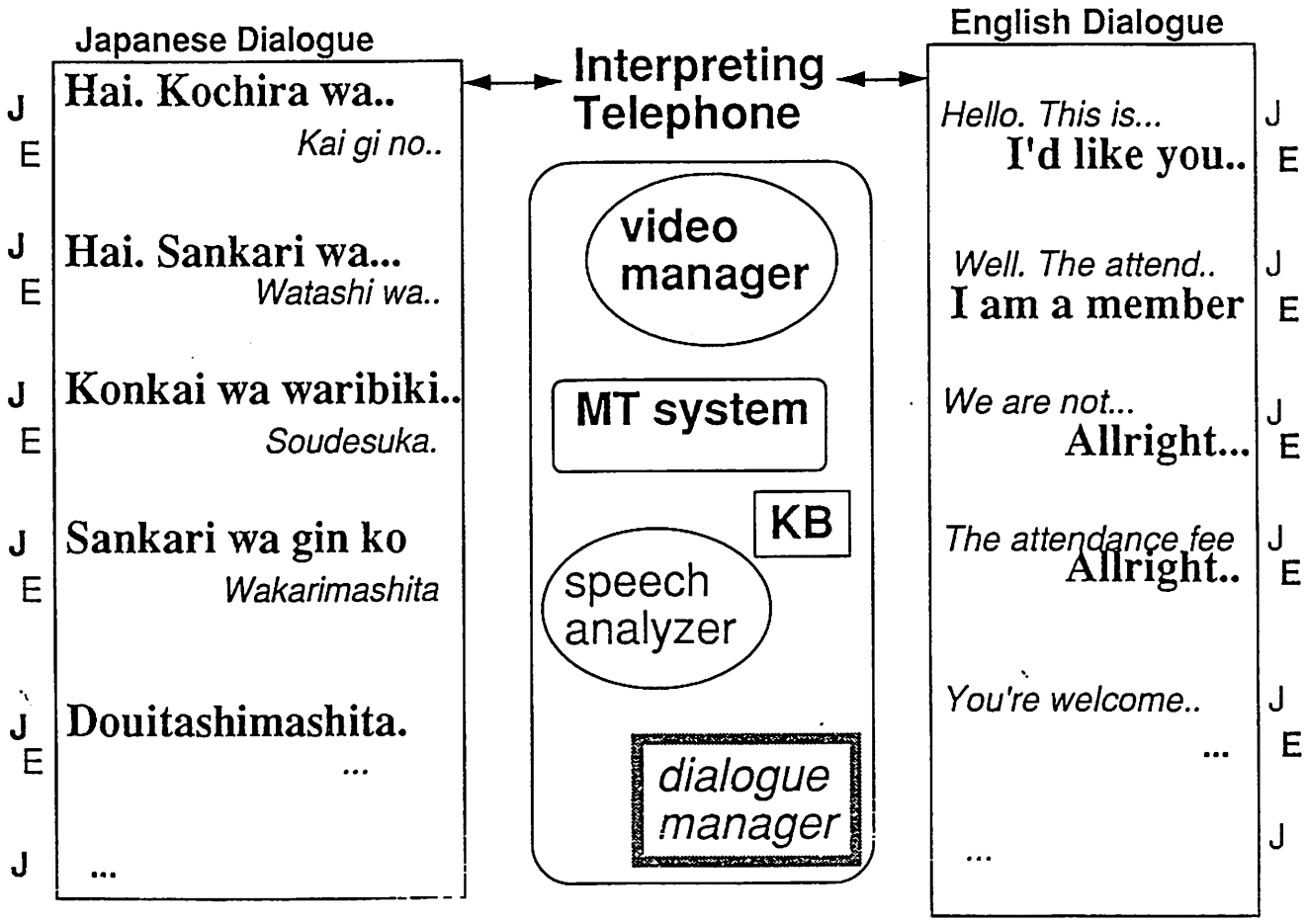
13. The Oviatt et al. study of human interpreters revealed a natural tendency for the interpreter to assume a very active role, effectively transforming the dialogue into a three-way conversation. The interpreter initiated requests for clarification, asked content questions based on prior experience with the domain task, and used third-person pronouns to refer to one speaker when addressing the other. The human interpreter often asked questions in anticipation of the other speaker's upcoming information needs. This human example provides one possible hypothesis for how the IT system might be structured but I suggest that this metaphor is too computationally costly for a real-time system. It requires abundant world knowledge and specific domain knowledge that may need to be acquired rapidly during dialogue processing. The intelligent-agent approach to IT cannot be achieved in the foreseeable future, and it is prone to errors that would be difficult to recover from. Moreover, it is not necessary in order to perform the IT task.

At the other extreme is the simple conduit model of translation in which the IT system would simply translate each sentence in isolation using lexical, syntactic, and sentential semantic processing, without considering contextual information. The assumption would be that the two human participants could perform repairs outside the system to compensate for the limited ability of the translation system. The advantage of this model is that it is relatively quick and inexpensive. The problem is that it is not a sufficient model for translation between Japanese and English as the most challenging phenomena require discourse processing.

The model I advocate is labeled the "Bilingual Dialogue Eaves-Dropper" in which the IT system acts as an observer of the dialogue, recording and making use of discourse-level representations for selected features of the ongoing interaction without participating in the conversation. However, if a communication problem arises the IT dialogue manager takes over control and interrupts the main dialogue to carry on recovery dialogue with one or both users. This model resembles a dialogue mediation system that many users are familiar with, namely the operating system that mediates 'dialogue' between users and software systems. The OS is only heard from when a difficulty arises, such as a printer runnning out of paper or a save operation being blocked by detection of an overfull disk. The OS keeps the simplest representation of the dialogue adequate for the tasks it must manage in order to provide required services to the user.

One alternative that has been suggested, is the assignment of a bilingual human operator to each IT dialogue, standing by to handle any problems that arise during the running of the IT system. And a variation would be to have a bank of human operators to which each dialogue in a state of difficulty would get queued. Both are unacceptable from a user needs point of view since the users would find it annoying to be placed on hold each time a difficulty occurred, and to avoid user waiting by having a dedicated human for each conversation casts doubt on the value of having automated the MT system to begin with.

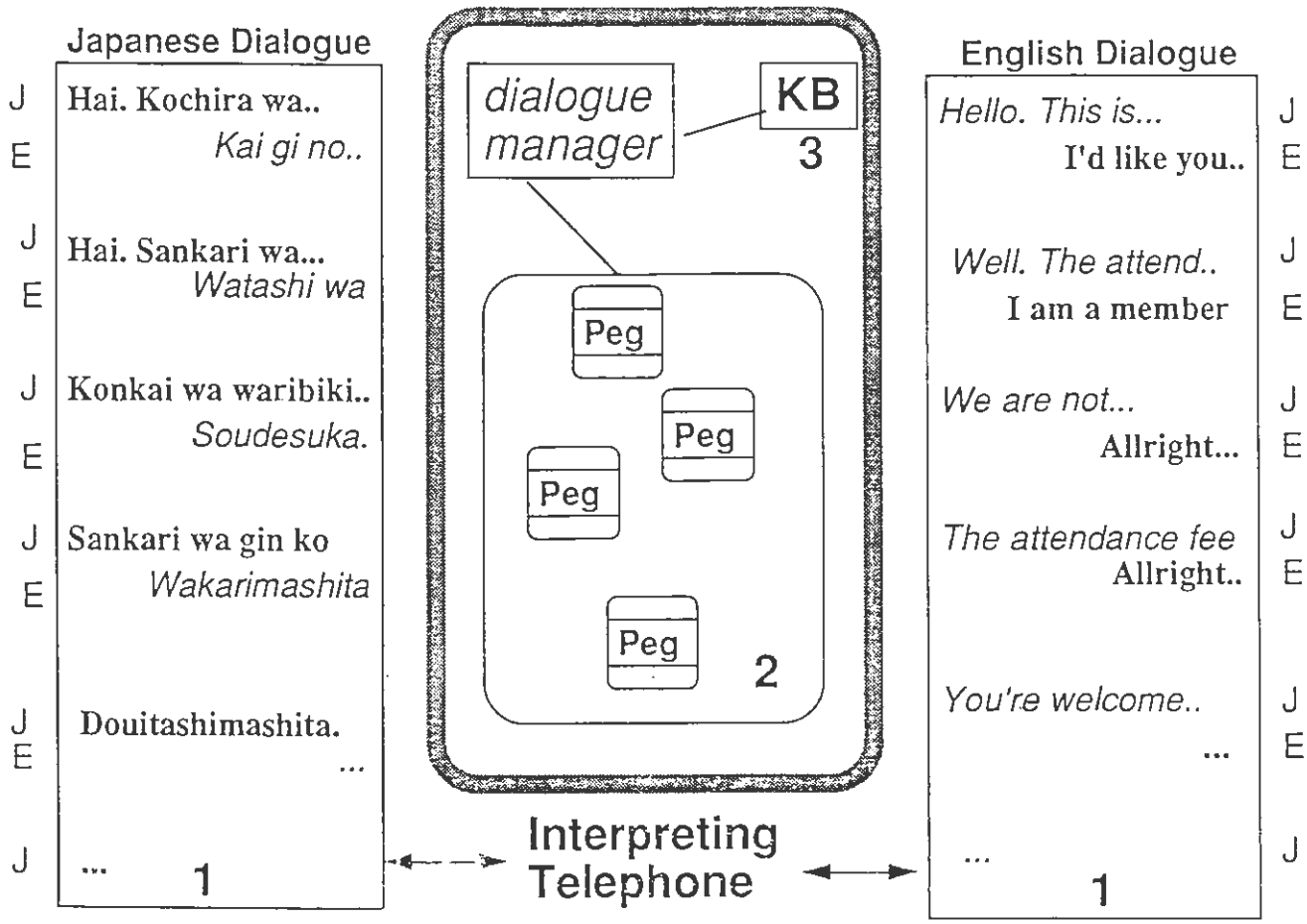
Ultimate IT System



14. In this illustration the IT system is positioned as the mediator of a bilingual dialogue. "J" stands for the speaker of Japanese, "E" for the speaker of English. J experiences a monolingual Japanese dialogue with IT just as E experiences a monolingual English dialogue. The two dialogues are isomorphic in structure but each adheres to the linguistic rules and dialogue interaction rules of the language community it represents. There are thus three dialogues to be studied in the IT project (1) E-J: the bilingual dialogue between the two humans, (2) J-IT: the monolingual Japanese dialogue between the Japanese-speaking user and the system, and (3) E-IT: the monolingual English dialogue between the English-speaking user and the system.

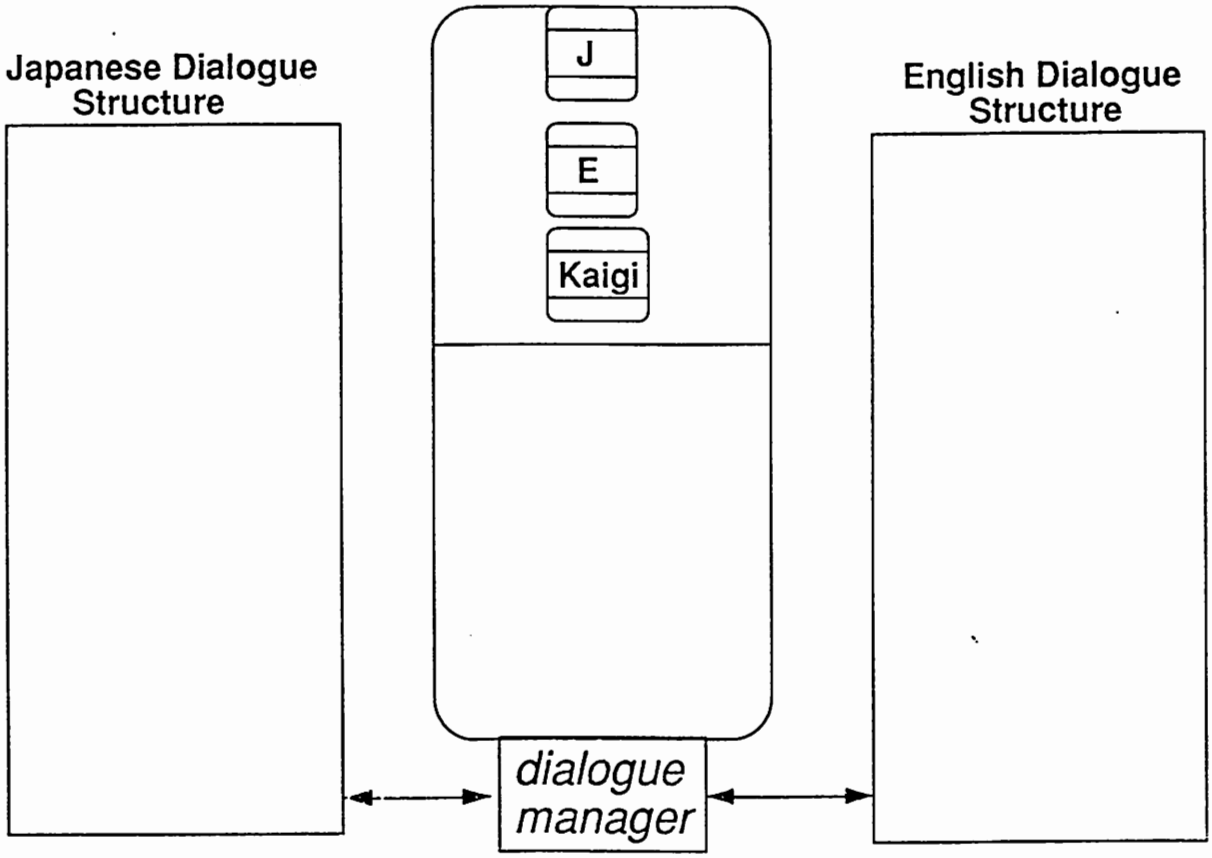
Among the many components making up this IT system there is a dialogue manager, the behavior of which is the topic of this report.

Three Tiered Discourse



15. An expansion of that dialogue manager box reveals the configuration of the three tiers that contribute to discourse processing in the IT system. (1) At the linguistic tier there are two representations of the dialogue, one in Japanese, one in English; (2) at the discourse model tier are the discourse pegs for the language-neutral representation of the ongoing dialogue, and (3) at the belief system tier is the knowledge base, or whatever domain model is determined necessary for the MT task.

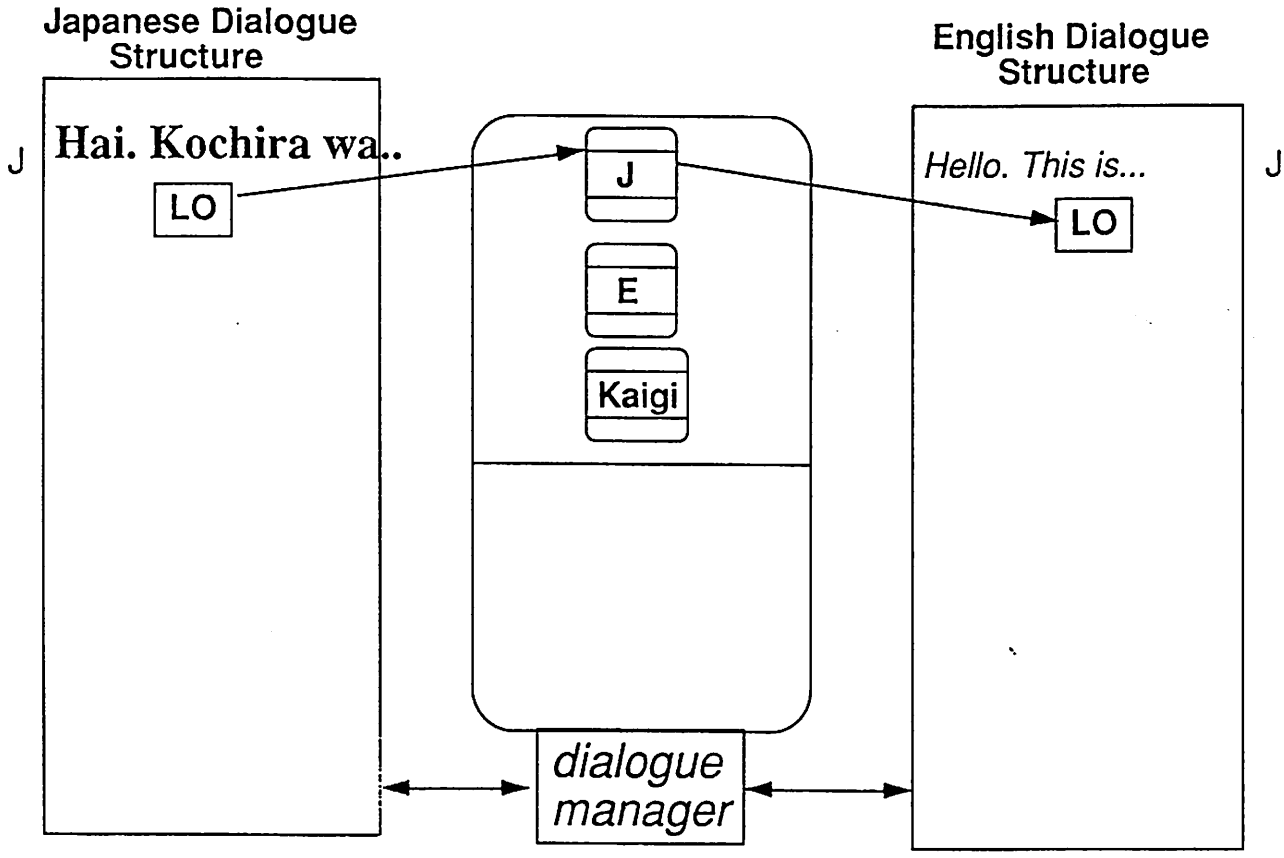
Initialized Dialogue



16. The next several slides present a sequence of snapshots from one of the sample dialogues as it might be represented using a three-tiered model of discourse with pegs. First of all the discourse representation is initialized to contain pegs for the two participants, in this case J and E, and for any other entities that are known to be highly salient and likely targets for reference in the dialogue. One example is the peg, Kaigi, standing for the conference that this IT system has been customized to manage. These objects may or may not be associated with KB objects that imbue them with additional information. Such knowledge might bear on discourse language such as the personal names, titles, current geographical locations, relative social status, of E and J, etc.

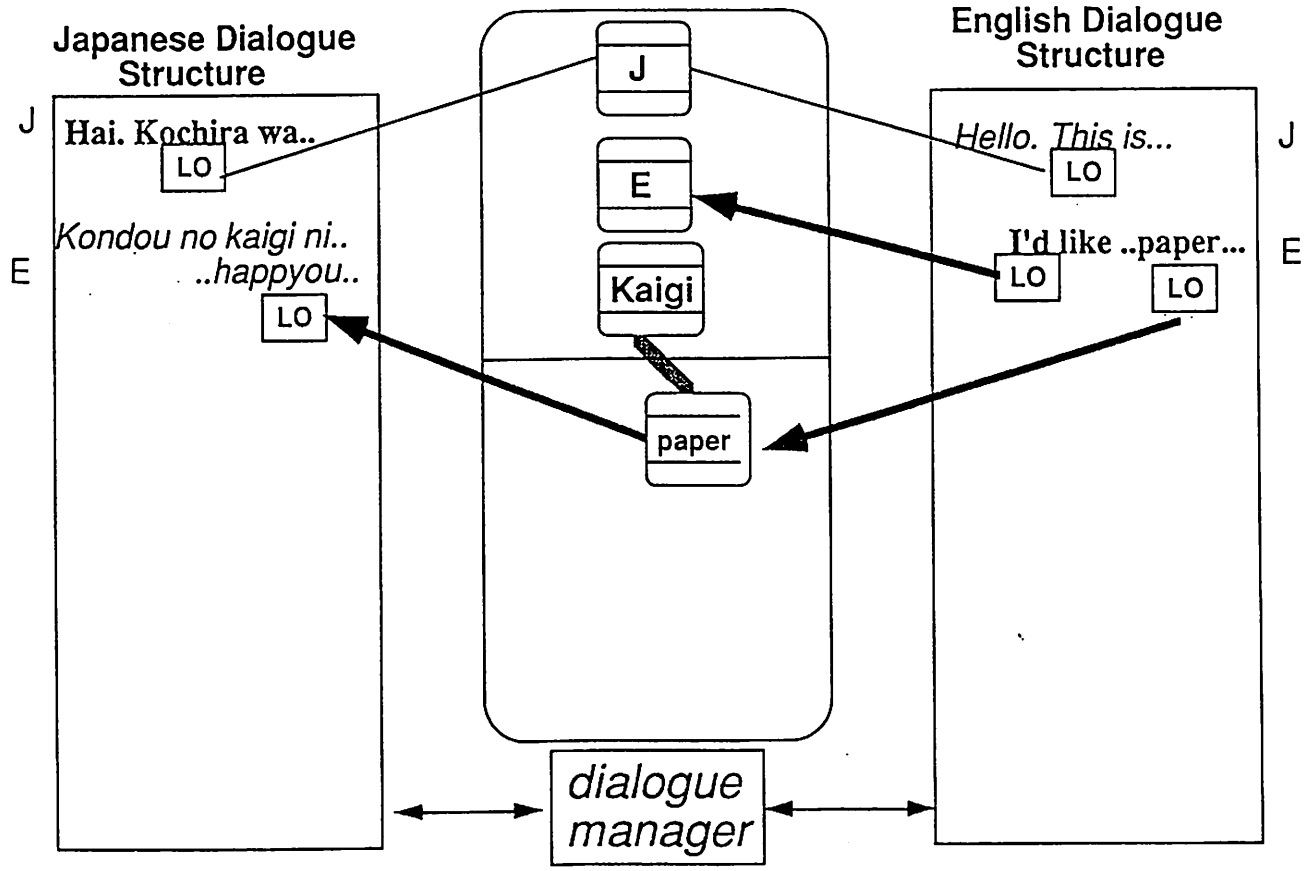
Discourse pegs can be introduced through non-verbal speech, and through other channels, if for example there is a video link. Pegs are partial, potentially incorrect, potentially in conflict with the KB information available to the IT system, and receptive to non-monotonic updating during subsequent discourse processing.

After First Utterance



17. After the opening utterance of the sample dialogue has been processed we see that no additional pegs have been introduced, but links between linguistic objects (LO's) and an existing peg have been created. Since this is an indexical reference by speaker J to themself so there is no need for execution of a centering algorithm for anaphora resolution, rather, the indexical term "kochira" indicates the peg named J directly. For proper generation of the English translation an LO for "This" is established in the English dialogue structurea and linked to the peg labeled J.

Second Utterance



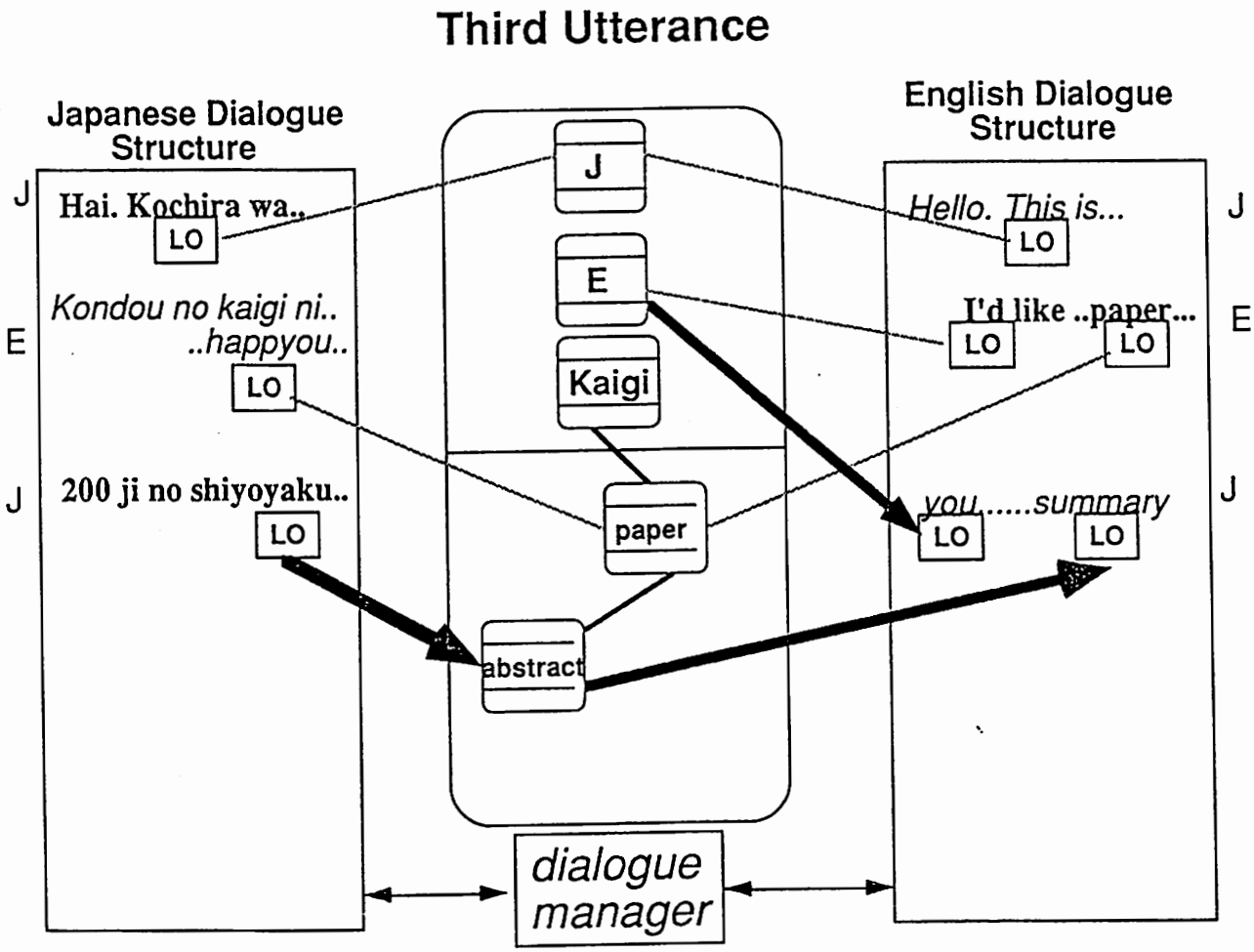
18. In the second utterance the term, "paper" introduces a peg that is related to the existing peg, Kaigi. This peg gets mentioned in the Japanese linguistic tier with the translation result "happyou." The first-person pronoun, "I" in this utterance mentions the peg E but due to zero-pronoun effects in the Japanese no LO is created and there is no link from E to the monolingual Japanese translation for this noun phrase.

Notice that in this framework there are three sources of information available to the MT system:

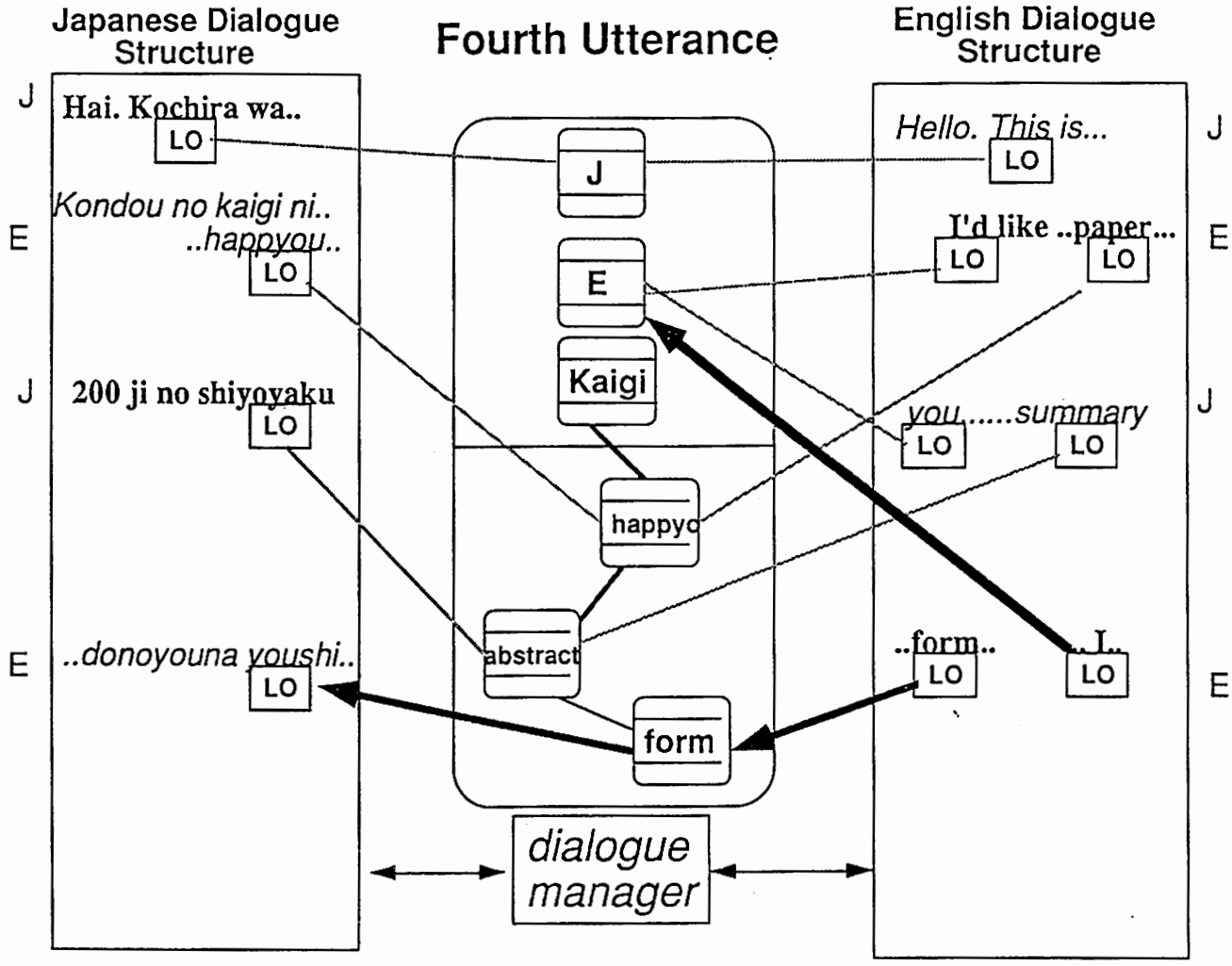
(1) the immediate source-language input utterance that is to be translated and the monolingual representation of the dialogue context in which it occurred.

(2) the pegs in the language-neutral discourse model and any additional world-knowledge information associated with them through a peg-KB link

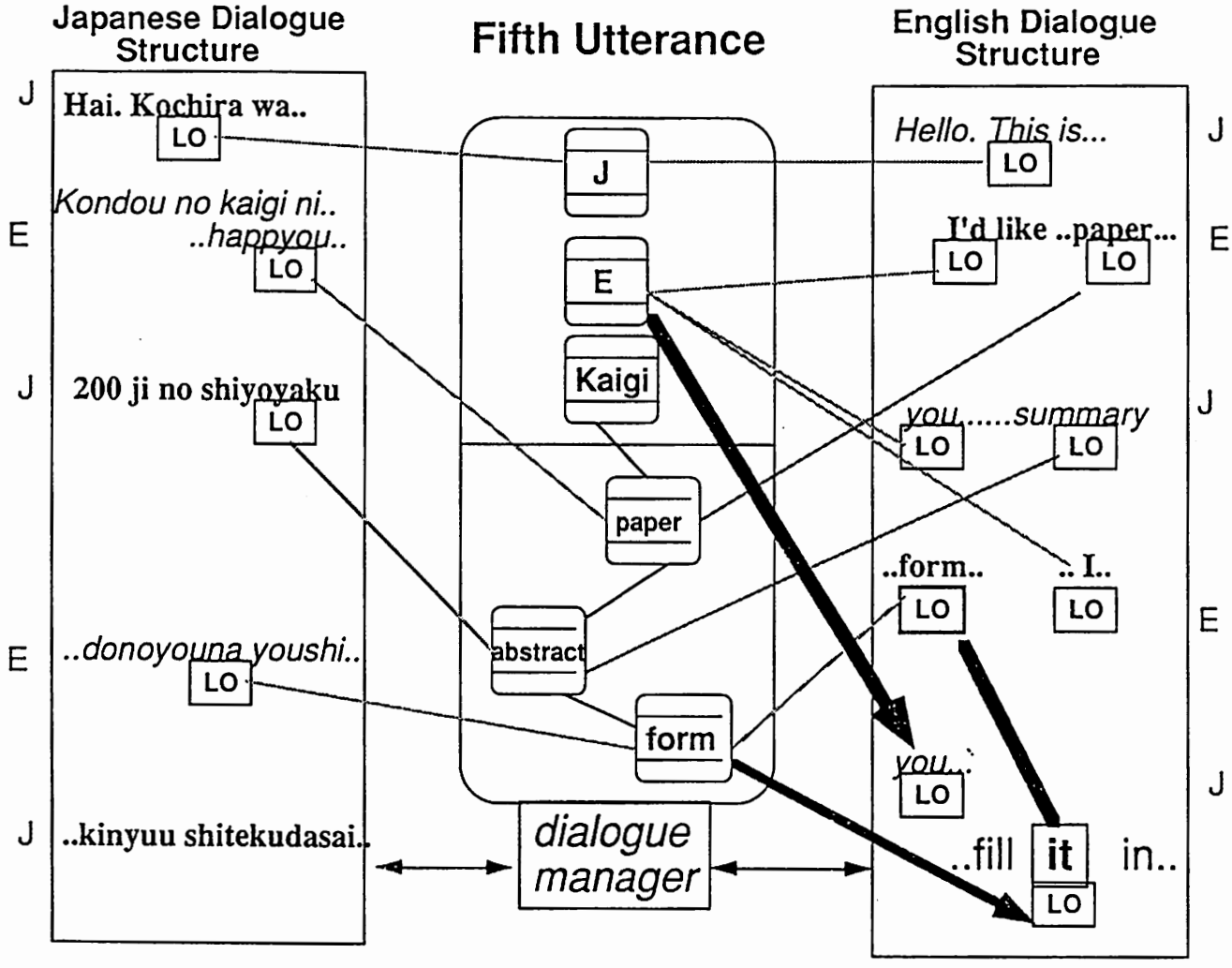
(3) the monolingual target-language dialogue context into which the translated reference must fit



19. The dialogue continues with the current-speaker shifted from E to J, current-listener from J to E, and source-language from English to Japanese. J uses the term "shiyoyaku" to introduce the concept of a written abstract related to the manuscript introduced by E's term "paper" in the previous utterance. At the linguistic representation level these forms are maintained separately and the assertions about each peg remain associated with the speaker who originated them, but at the discourse model tier the concepts have a single representation. Utterance 3 contains a zero-pronoun reference to E which must be translated into English as "you."

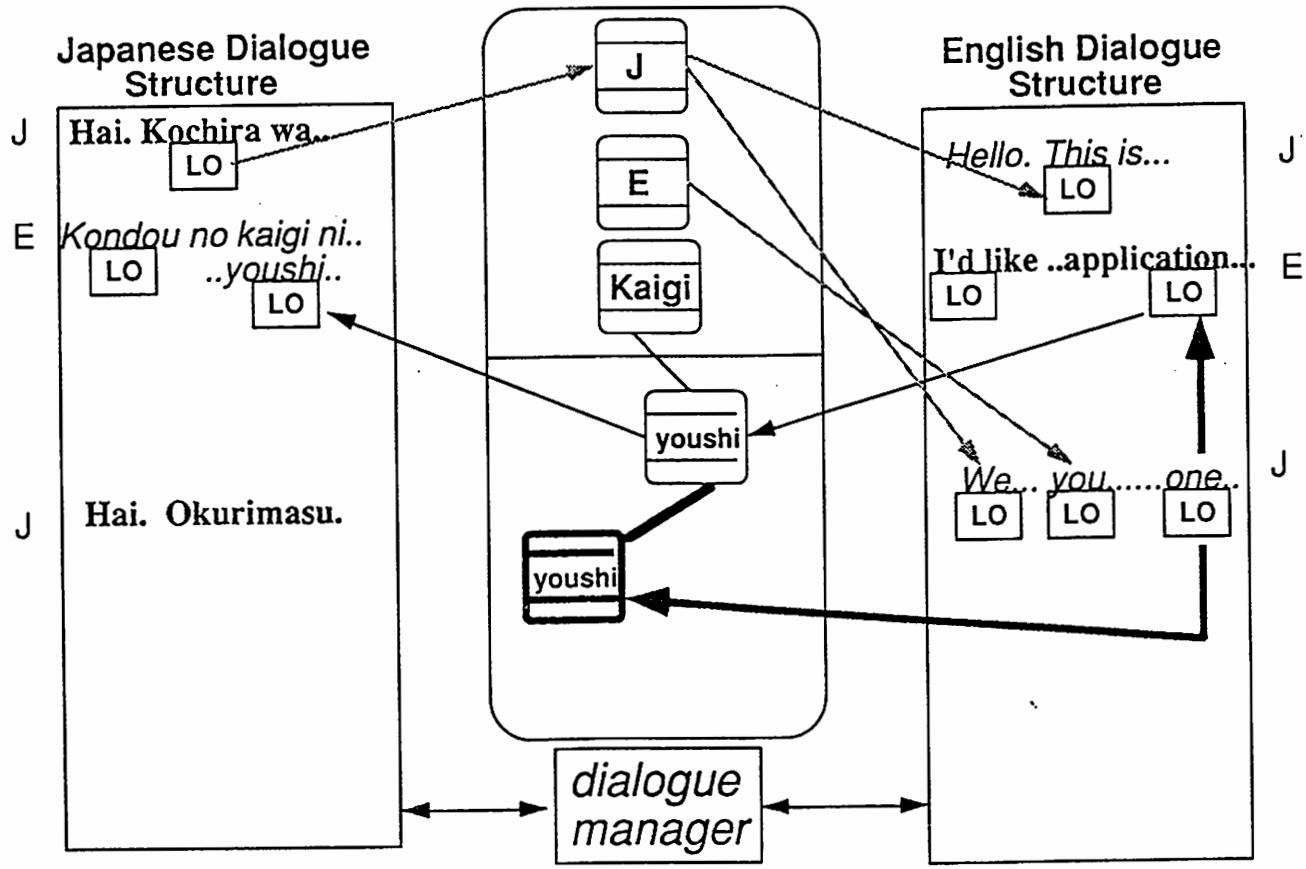


20. We notice that after the fourth utterance has been issued by the English speaker the peg E has been mentioned explicitly three times by that speaker and not at all by J. By separating the linguistic tier into two isomorphic monolingual representations we are able to encode this asymmetry in our discourse representation and make use of it in resolving zero-pronoun references.



21. The fifth utterance illustrates the need for both language-specific and language-neutral representations of the dialogue in order to translate discourse-dependent forms such as zero-pronouns from Japanese into English. Even though there is no LO in the source language utterance, rules of dialogue in the target language demand an explicit form. The specific features of the IT situation and the representation we have chosen for it provide the MT system with two sources of help in generating the correct translation, the language-neutral discourse representation structured by attentional focus and the discourse history in the target language that helps to constrain the form of the resulting output.

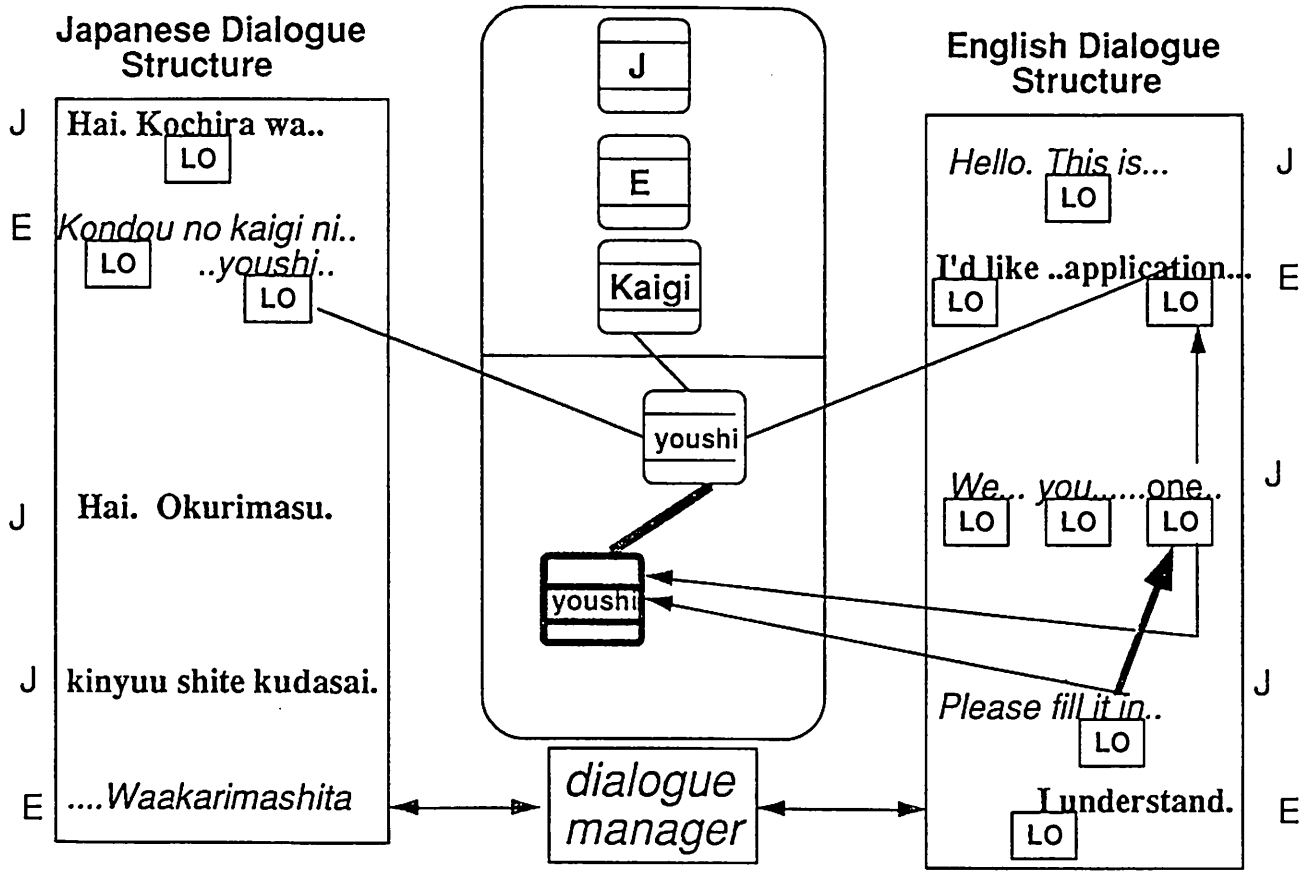
Language-Specific Phenomena: One-Anaphora



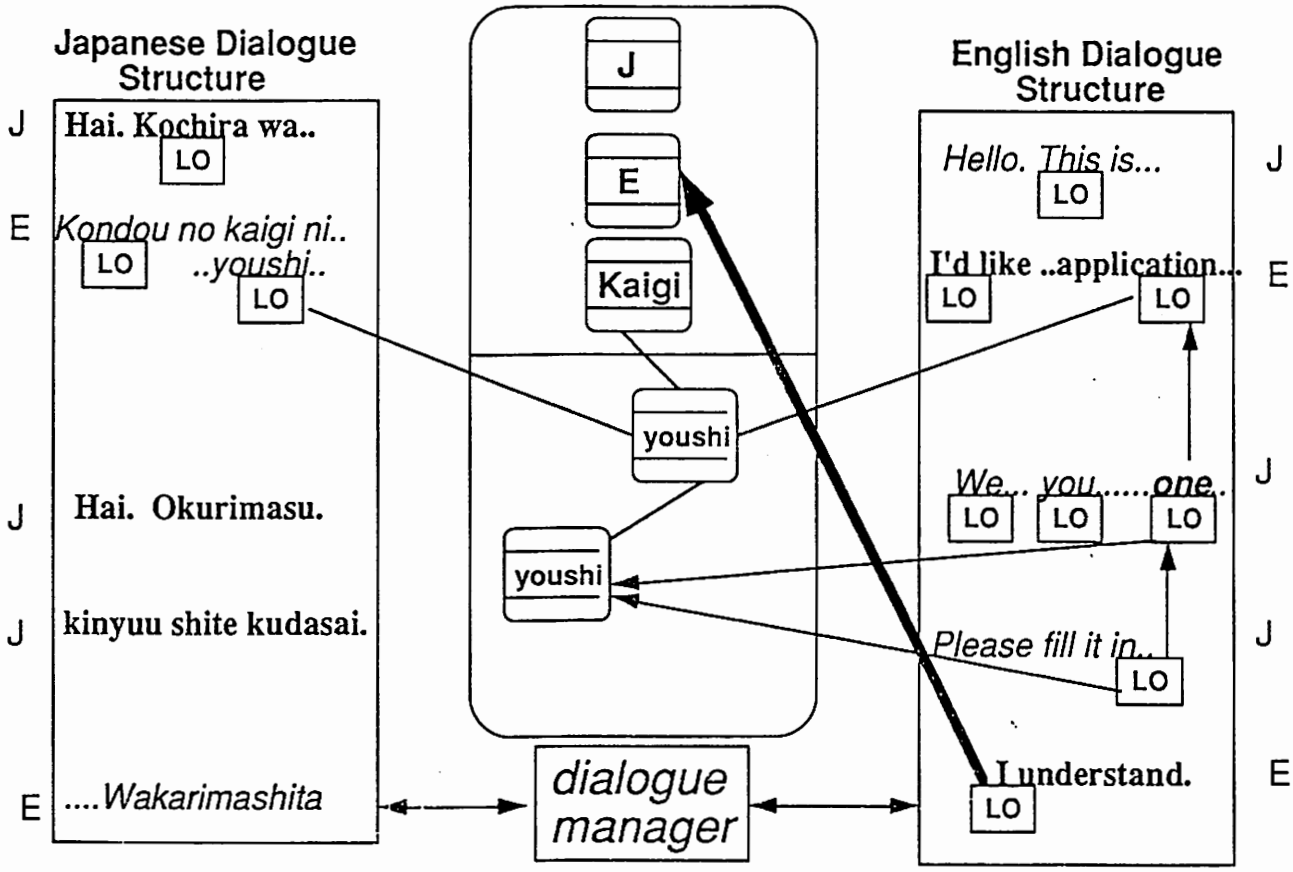
22. This diagram illustrates the value of the encoded asymmetry between the two monolingual dialogues. In order to process and translate such language-dependent discourse phenomena as English one-anaphora there must be a representation of the English dialogue apart from the language neutral representation. In this example, in order to translate "okurimasu" to something like "We will send you one" rules of English are applied to the English representation of the bilingual dialogue (the right hand side of this diagram). For example, the sponsor (to use terminology from my earlier talk) of the generated one-anaphor must be a count noun in English and it must be located in the very recent discourse history.

The result of one-anaphor processing is a new peg introduced into the discourse model, in this case the second (darkly printed) peg labeled "youshi." It is now available to sponsor a subsequent definite anaphor "it."

Language-Specific Phenomena: Pronouns



Language-Specific Phenomena:Zero Pronouns

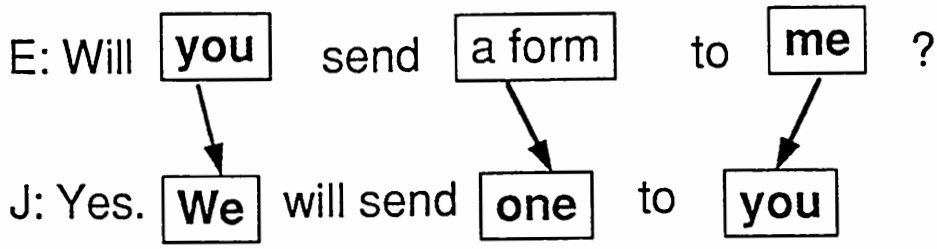


23. Zero-pronouns are a well studied language-specific Japanese phenomenon that make use of the focus structure represented in discourse model.

Take Advantage of Bilingual Dialogue

Ellipsis in dialogue- using English-specific rules to help translate Japanese input.

"Hai. Okurimasu."



24. Given the representation argued for thus far we can examine the MT task faced by the IT system and suggest specific techniques available to it though not to MT systems in general. Some factors to consider in looking for shortcuts to take advantage of in this unique project include, the fact that the data are from a dialogue, i.e., the speaker and hearer roles alternate, the dialogue is bilingual, the two languages are Japanese and English, the source and target languages alternate, for the immediate term this is a telephone conversation so there is no visual information, and the language use will be specialized for the domain of information requests to a conference secretary's office.

In this example, the Japanese reply "Hai. Okurimasu" must be translated into English in such a way that the result is compatible with the preceding English discourse. The problem is that the Japanese source utterance is lacking arguments that are required to be explicit in the English result. Contextual information and rules of English anaphora can help determine the missing arguments and their proper expression in the target language. This is the standard approach to zero-anaphora.

Odd/Even Information Content

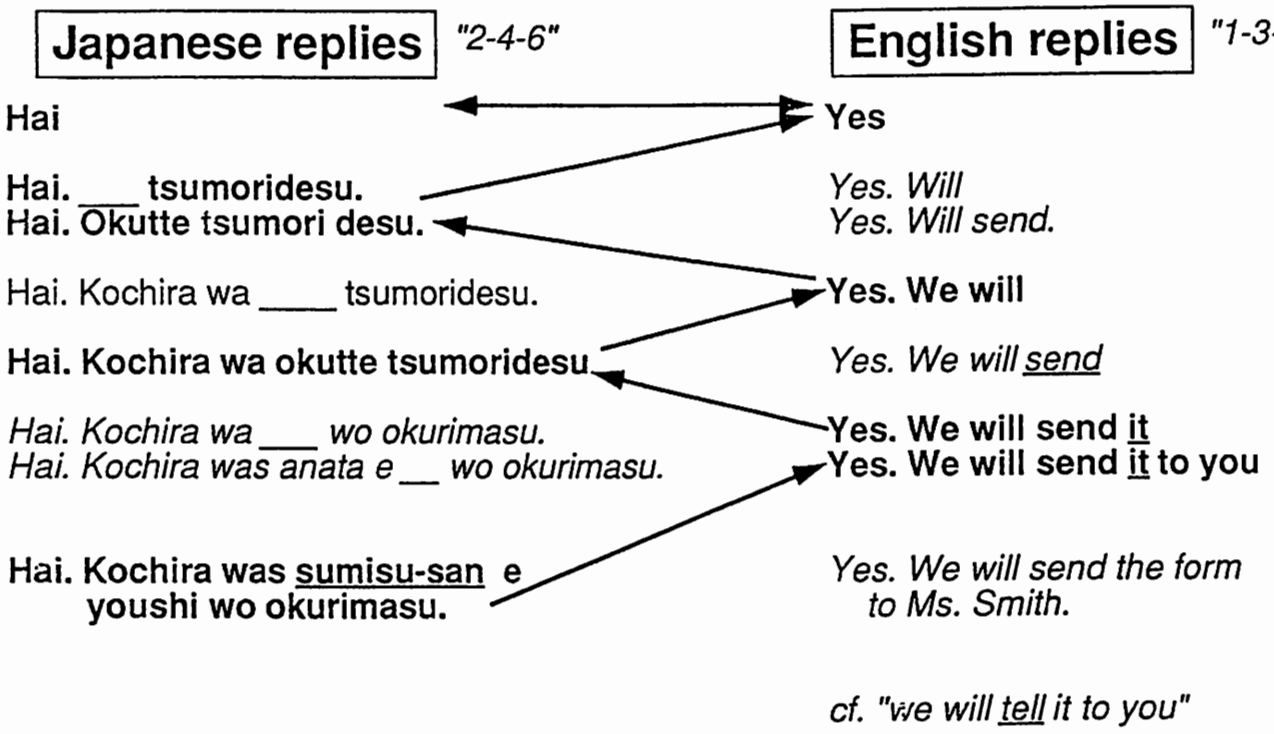
E: Will you send a form to me?

Japanese replies	English replies
Hai	Yes
Hai. ___ tsumoridesu. Hai. Okutte tsumori desu.	*Yes. ___ Will* *Yes. ___ Will send.*
Hai. Kochira wa ____ tsumoridesu.	Yes. We will
Hai. Kochira wa okurimasu	*Yes. We will send*
Hai. Kochira wa ___ wo okurimasu. Hai. Kochira was anata e __ wo okurimasu.	Yes. We will send it Yes. We will send it to you
Hai. Kochira wa sumisu-san e youshi wo okurimasu.	*Yes. We will send it to Ms. Smith.*
	cf. "we will tell it to you"

25. An alternative approach is to consider the grammaticality of expanded or contracted forms of the source utterance. In this diagram boldface print indicates grammatical forms in either language. Illegal forms are indicated with reduced, italic type. Comparing left to right sides we see a mismatch in the number of 'pieces of information' required to be expressed in the two languages. So while Japanese allows "Hai. Okutte tsumori desu" the transliteration, "Yes. Intend to send." is ungrammatical. So it is as if the two languages were always off by one in their required number of arguments required to be made explicit in the translation using context. However, the MT system has two options, (a) supply an additional bit of information (n+1) using context and anaphora resolution techniques such as centering, to yield in this case "Yes. I intend to send one," or (b) delete one piece of information (n-1) to yield simply "Yes".

MT Strategy Using Odd/Even Contrast

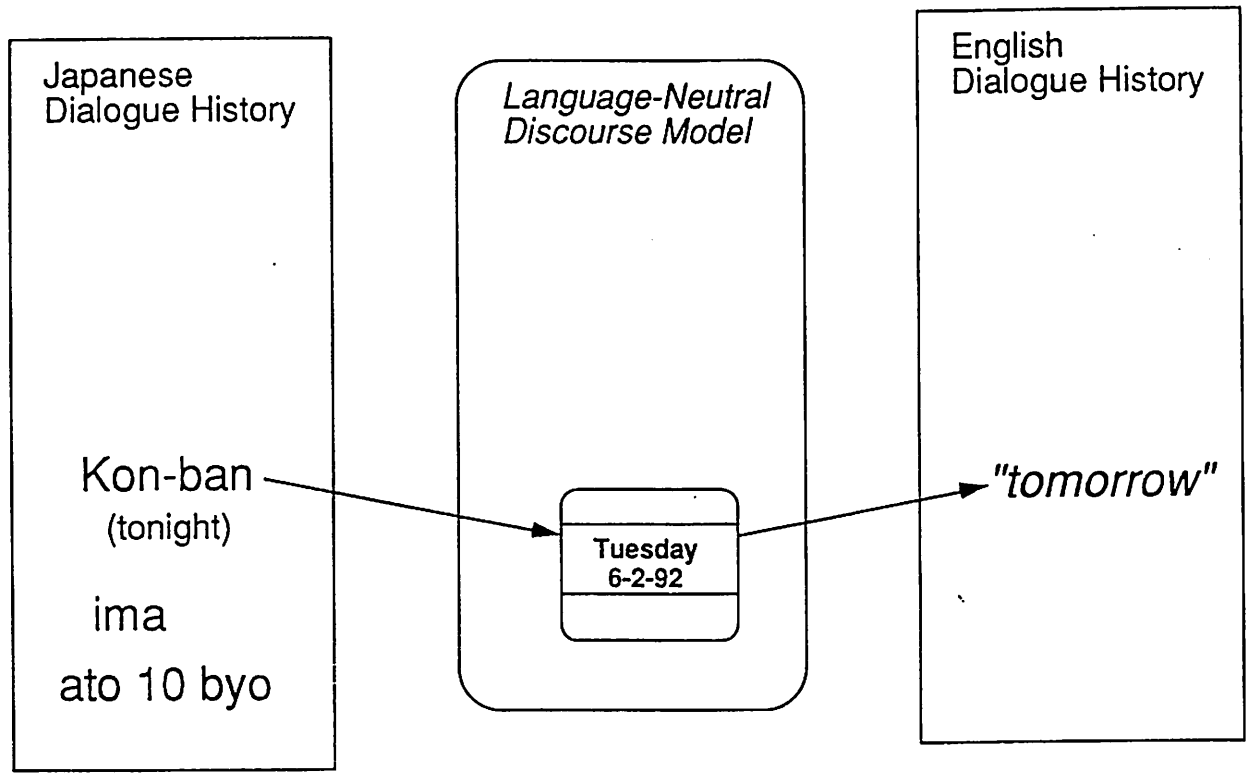
E: Will you send a form to me?



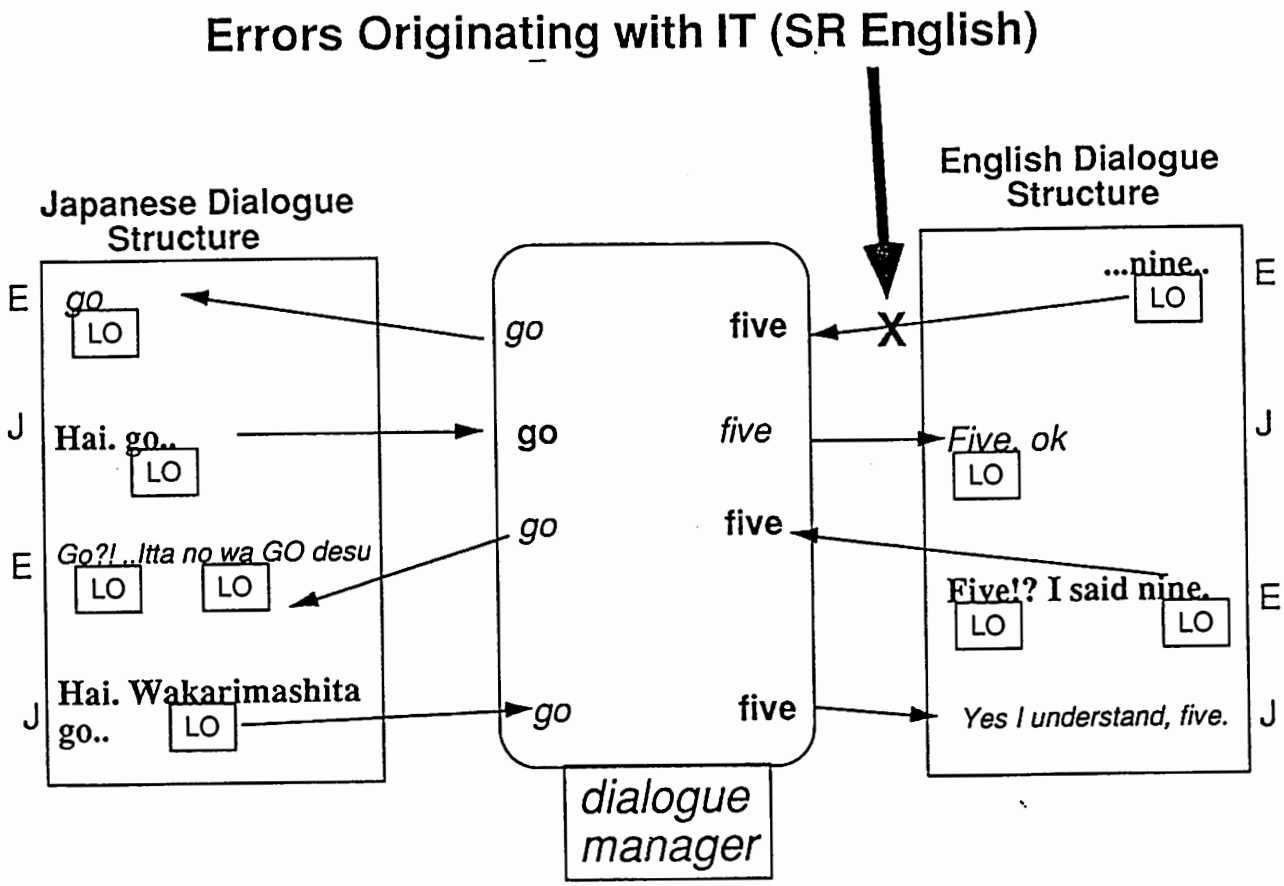
26. The "delete one" approach indicated by arrows in this diagram produces grammatical utterances in the target language. The assumption is that the human recipient of the result can supply missing information from context. Though this may be controversial, the "add one" approach has the inherent risk of being prone to error because it is supplying information from context just as the human must, often with a high degree of uncertainty.

Outside World Problems

(e.g., temporal indexicals across time-zones)

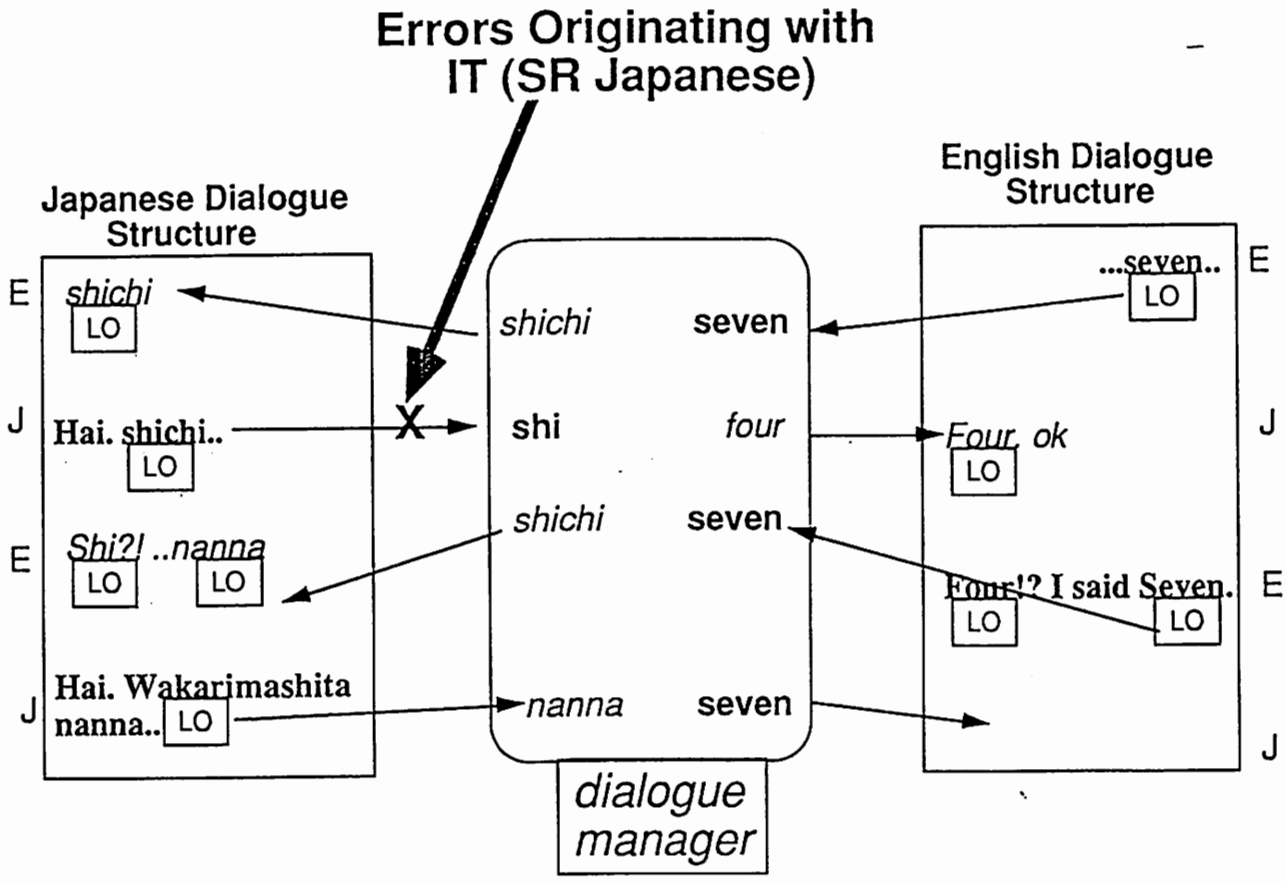


27. This slide illustrates one dialogue interaction problem that should be ignored by the IT research effort because it is better handled by the system-external intelligence of the human users. If on Monday morning in Kyoto (Sunday night in Ohio) the Japanese speaker promises to fax a document to the English speaker in Ohio "konban" the IT output ought to be "tonight"/"this evening" and the system should not attempt to compute the time-zone conversion between Ohio and Japan to get "tomorrow evening" in the thought that the referent time, Monday evening, is would be described from the Eastern Time Zone perspective of the English speaker in Ohio as, "tomorrow evening." This is a non-linguistic (real world) problem for any two people speaking across distances that involve time-zone discrepancies and humans are accustomed to handling the conversion cooperatively. "Do you mean you'll send it Monday night your time, which is Monday morning my time?" More importantly, such intrusive interpretation is prone to error and counter-intuitive. Users would have a difficult time determining what was intended by the other speaker if the system deviated from language translation and strayed into these areas. In other words, this is not a problem for the IT system.



28. In contrast, the next situation does represent a problem for the IT system, and one that I will use to argue for directing attention to user interface (UI) issues of the IT system. Here the English SR system has failed on English speaker's input "nine." If the error goes undetected it can pass through the system past several stages until the problem becomes very difficult to repair. "Nine," having been recognized as "five," gets translated (correctly) to "go" which gets synthesized (correctly) and presented to the Japanese user who hears (correctly) "go" and offers a confirmation statement, "Hai. go" ("Yes/ok. Five"). That utterance gets recognized correctly by the Japanese SR system, translated (correctly) from "go" to "five" and synthesize intelligibly by the English SG system. When the English speaker hears the confirmation utterance "Ok. five" they recognize the error and initiate a repair subdialogue by generating a correction statement. Notice that the English speaker has been conditioned to a level of meta-linguistic awareness that "five" and "nine" sound alike and are often confused among English speakers. (This example was suggested to me by N. Campbell). The Japanese speaker, on the other hand, is not prepared for the confusion among "go" and "kyu" ("nine") and cannot help with detecting or diagnosing such as recognition error.

The solution I suggest is to design the UI dialogue on each side to include an 'echo' mechanism that reproduces (possibly as text on a screen) each input utterance as it is recognized, and allow the speaker an opportunity to clarify before the utterance is passed on through the system.

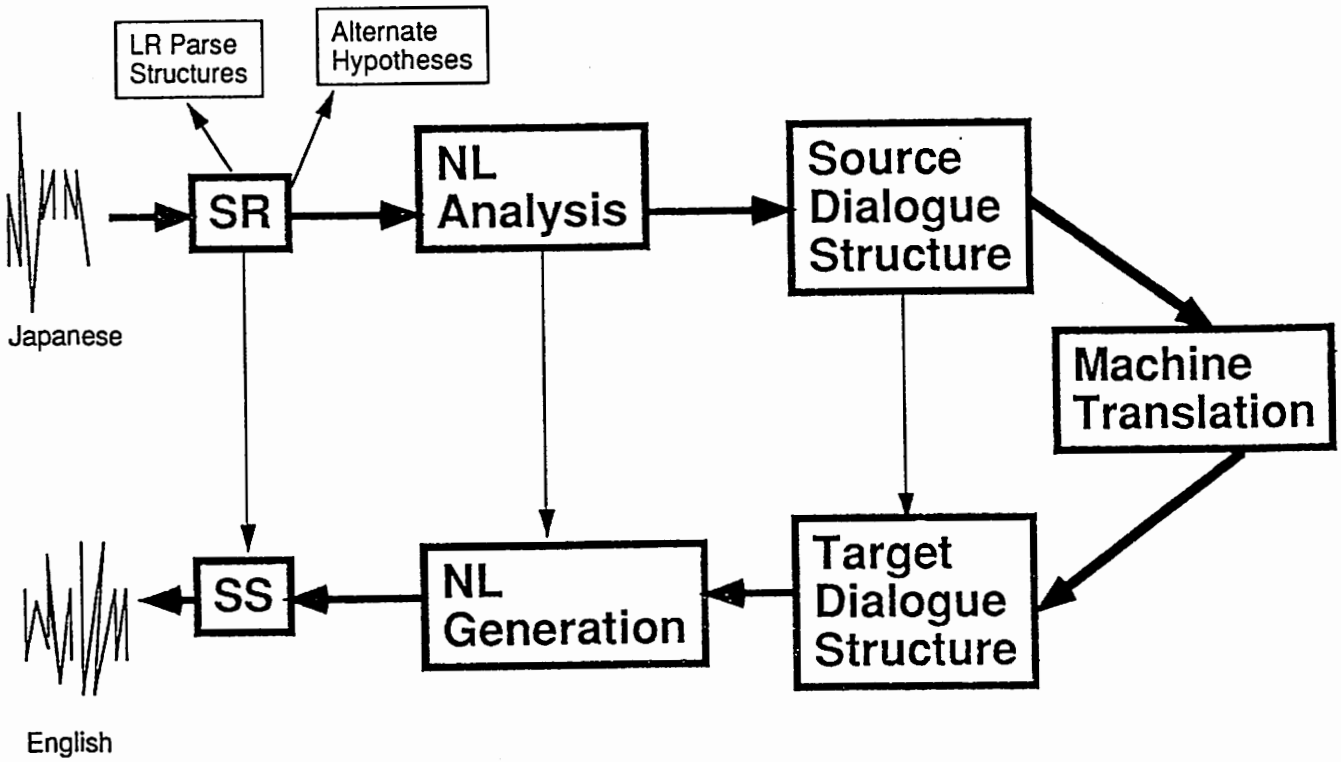


29. An even more dramatic example, one quite likely to occur, involves an error in recognition of a confirmation utterance. Here "seven" is recognized and translated and the Japanese speaker offers confirmation. That input is misrecognized as "shi" and translated (correctly) to "four." Here the English speaker is not prepared for confusion between "seven" and "four" but the problem is difficult for either user to diagnose since the error and evidence of an error occur in remote locations during processing. Both errors illustrated here are IT system errors not human errors and in both cases the IT system must recognize, diagnose, design a repair, and execute a repair dialogue.

Here again, an input echo system would have alerted the source language speaker and allowed for immediate repair. Regardless of the preventative measures designed into the IT system, there will always be errors of various sorts. The UI system must perform four operations:

- detection: recognize that an error has occurred
- diagnosis: classify the error as resulting from speaker mistake, SR, NL analysis, MT, NL generation, SS.
- repair plan: devise a dialogue solution by engaging one or both users in a repair subdialogue to clarify the analysis.
- repair execution: carry out the repair plan with the cooperation of the human(s) responding to their attempts to clarify.

A View of the Processes



30. The remainder of this report has to do with the question of how discourse processing can assist the process of SR in particular. This slide is a version of a diagram seen often around ATR which illustrates the internal structure of the IT process. I have included a dialogue component and indicated the two results of SR in the ATR IT system.

Three Ways to Use Discourse Information

1. Top-down discourse information *run-time*
   (to predict next discourse event)
   plans, intentions, communicative acts

2. Left-right and Right-to-Left *run-time*
   (to incorporate new speech input
   into model of ongoing discourse)
   zero pronouns
   attentional focus

3. Right-Left loops from discourse to SR *training-time*

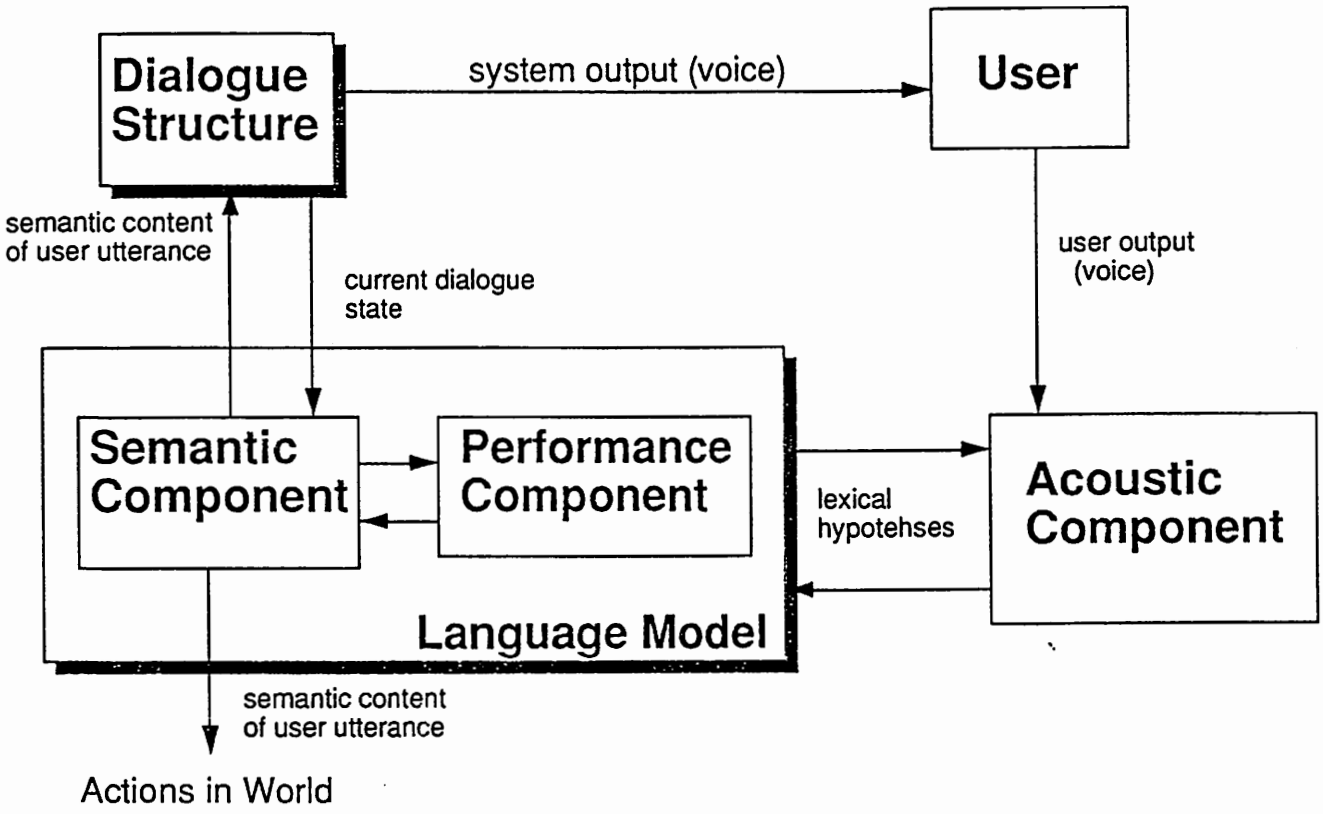
31. We can characterize the contribution of discourse to the SR process as being of three kinds. Top-down discourse information is predictive. The IT system can be used to anticipate language behavior at the level of intentions, speech acts, utterances, and even NP forms. The discourse information used can be dynamically collected information and may require access to stored libraries of plans and realizations of plans. (see Yamaoka, 1992 presented at COLING)

Left-to-right information transfer refers to accumlation of information from SR, NL analysis, etc., into the discourse representation. Right to left information is then the use of discourse information to filter multiple returned values of a module prior to (to the 'left' of) itself.

"Right-to-Left Loops" is a term I use to describe training of a system using stochastic information from prior output. This is the basis of stochastic approaches to SR in general and the specific approach being explored by Demori et al., and by Weibel et al. in attempting to extend the statistical methods of SR beyond phoneme and word recognition and into discourse-level interpretation of spoken utterances. The looped link is to indicate that the information about discourses in general is fed back to the SR module before run time thereby modifying the behavior of the SR process by training it on what is determined to be representative data.

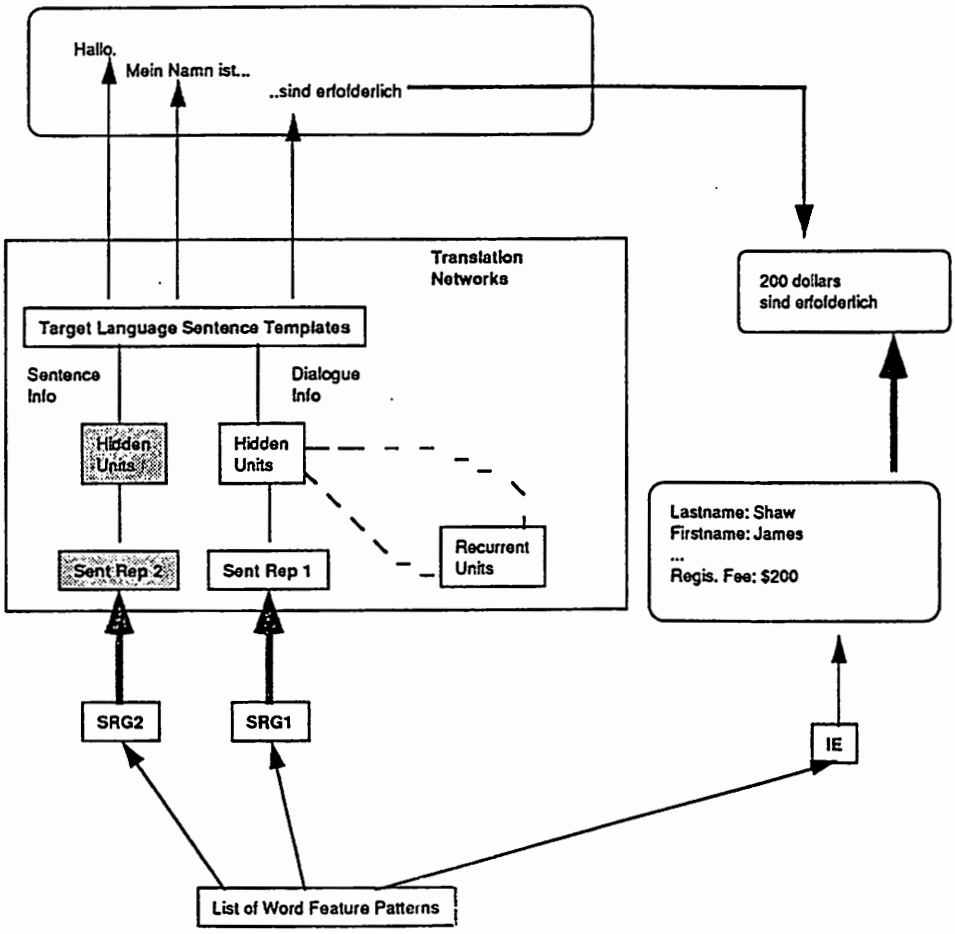
2.1 DeMori, Kuhn, Lazzari

A Probabilistic Approach to Person-Robot Dialogue



32. This diagram was taken from DeMori et al. and describes their speech UI to a robotic system. The task presented to their SR system differs significantly from the IT SR task in that the set of utterances that their robot UI SR system needs to recognize is highly constrained to be that set of utterances that are meaningful to the backend system. Any input that does not match a pattern from a meaningful utterance can be ignored by the SR system. This is the significance of their performance module. In contrast, the IT system's SR component (1) may be faced with any of a large number of input utterances, and (2) unexpected input not included in that set cannot be ignored by the future IT system; unexpected or not, it must be conveyed to the target speaker.

Wang and Waibel



33. This diagram was taken from Janus: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. It uses a dialogue corpus for data but its analysis is not affected by the dialogue phenomena therein, i.e., the data get processed as a sequence of spoken utterances. It translates individual utterances from English to Japanese and to German using an interlingua KB-based technique.

2.3 Zue et al.

Integration of Speech Recognition and Natural Language Processing in the MIT Voyager System

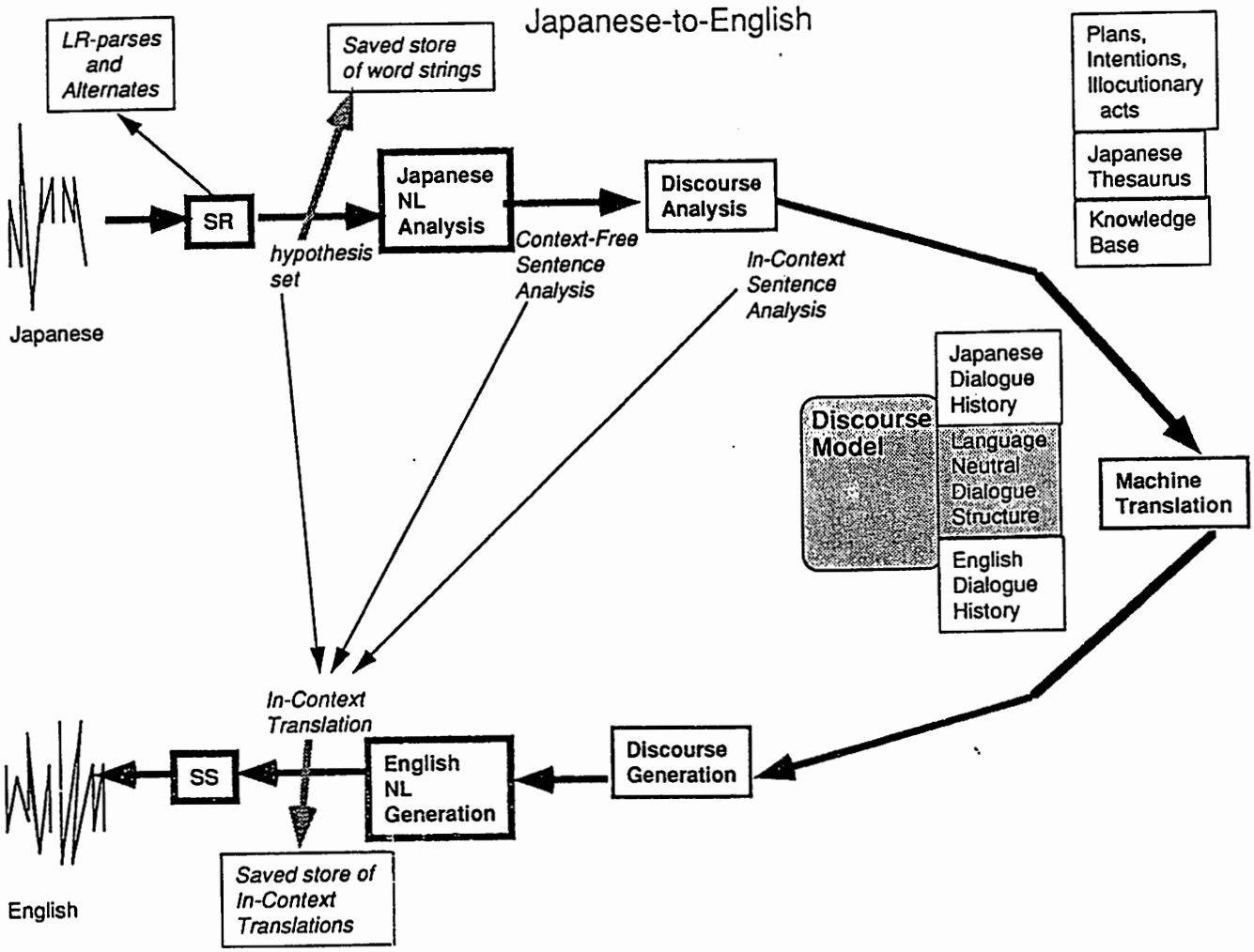
used generator to construct word-pair language model

generates top N word strings to NL component (integrate SR and NL)

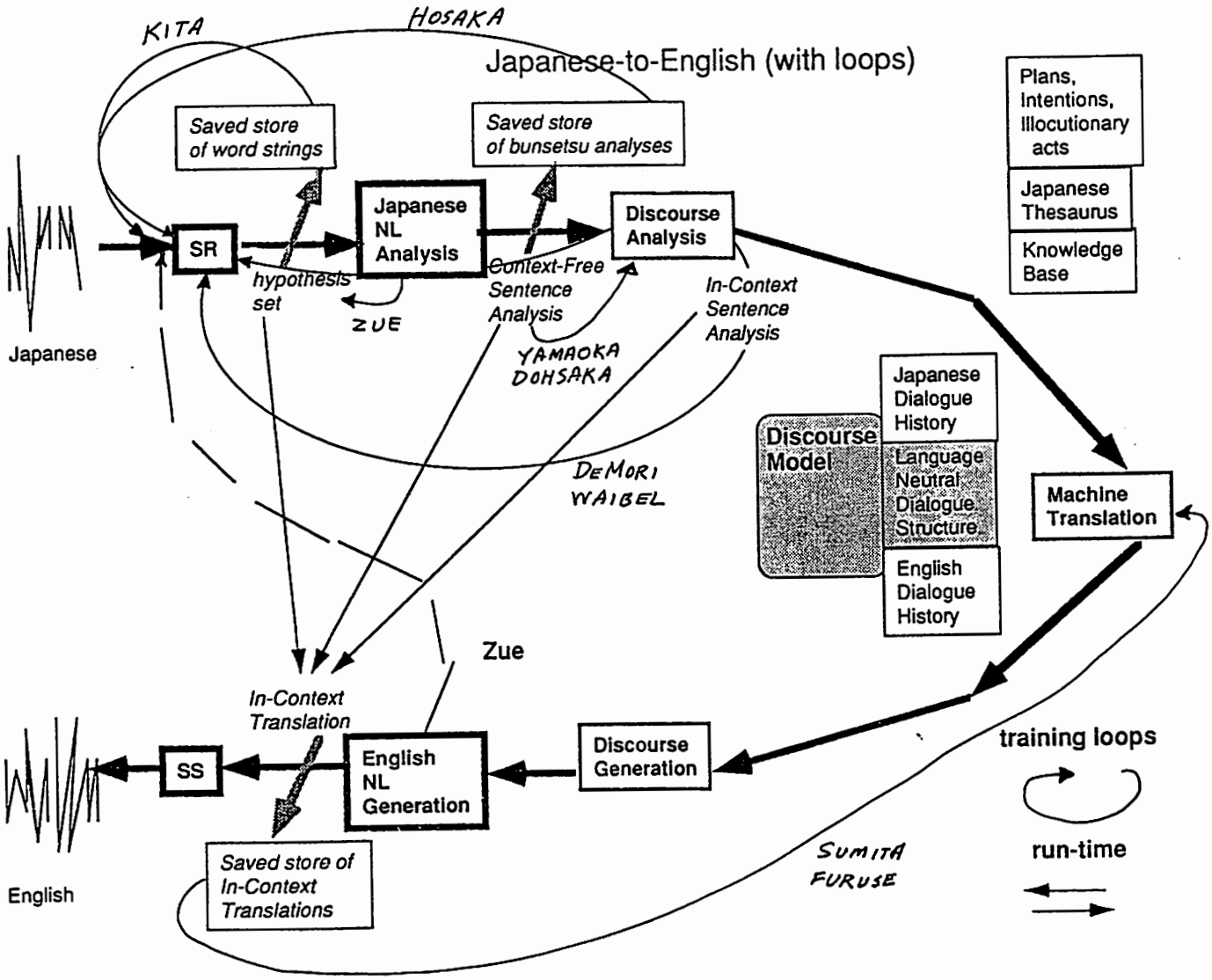
modified Viterbi to N-best search (A* algorithm)

future: dynamic adaptation of NL constraints, e.g., increase KB; control perplexity of SR by limiting vocabulary-based on the discourse history

34. The Voyager system represents another effort to integrate speech recognition and NLP including discourse processing. It is a natural language user interface to a map system that supports user queries regarding navigation directions between locations in Cambridge, Massachusetts. Developers used the NL generation component as the training set for the language model to be used during recognition. (In my terminology, this would be represented as a right-to-left loop from the NL component to the SR component.) Voyager uses an N-best search and returns the n most favored hypotheses to the NL component. Zue et al. suggest as future methods for feeding discourse information to the SR component at run-time (which would be a right-to-left arrow from discourse to SR in my diagrams) by using word occurrences in the discourse history to affect expectations of the SR process.

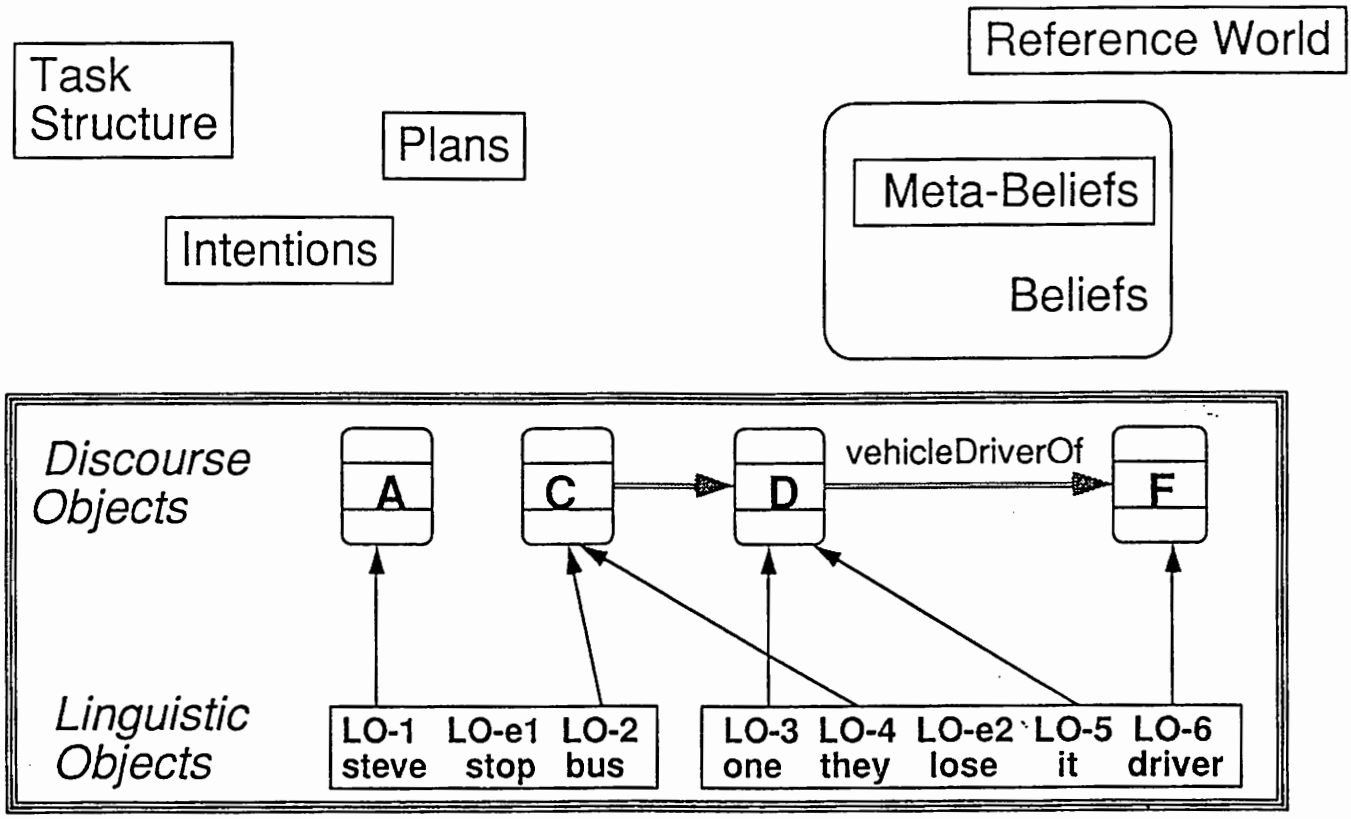


35. In this elaboration of the standard diagram in Slide #31, I have added a representation for the discourse histories and discourse model and have included labels on outputs of each process. So, for example, the SR component's LR parser produces a syntactic parse for each utterance but discards it after completion of the SR task. Arcs from intermediate points during analysis of the Japanese utterance are meant to indicate translation of partially analyzed source language whenever resources are sufficient. For example, idiomatic phrases such as "I would appreciate it" may get translated to "onegaishimasu" without being given a full, in-context semantic analysis. Higher level discourse information such as plans, intentions, and communicative acts, are indicated in boxes in the upper right region of the diagram. While I want to acknowledge their value for thorough discourse analysis and MT, my goal is to accomplish the particular MT task faced by the IT system, with less expensive techniques and structures.



36. This diagram emphasizes the exchange of processing information between components of the IT system. Left-to-right arcs coincide with the predominant flow of control from acoustic signal processing through source utterance analysis, translation, and generation of target utterance output. Loops indicate the use of prior output language as data to train IT components. So for example, the saved store of bunsetsu analyses are used to train the SR module so that its future a priori predictions will be in line with those previous training data, and Zue et al. use the utterance strings produced by Voyager's own grammar as data for training the SR module. The primary concern here is with the use of discourse information to enhance the performance of lower level components, especially the SR module.

Inexpensive Discourse Structures



37. The three-tiered discourse representation does not require that the discourse model be supported by a knowledge base, i.e., the third tier. What is essential is that the pegs level be viewed as partial, possibly incorrect and therefore correctable, and as a representation of the discourse itself apart from any grounding in the real (reference) world or in beliefs about the world. Intentions, plans, beliefs about constructs in the domain world, and task structure may affect the processes that update the discourse model but the model itself is not integrated with those knowledge sources. Therefore, we can conceive of the IT dialogue manager as functioning in the three-tiered framework without assuming that the NL and discourse processing routines have access to a knowledge base and without integrating information on intentions with the representation of the discourse.

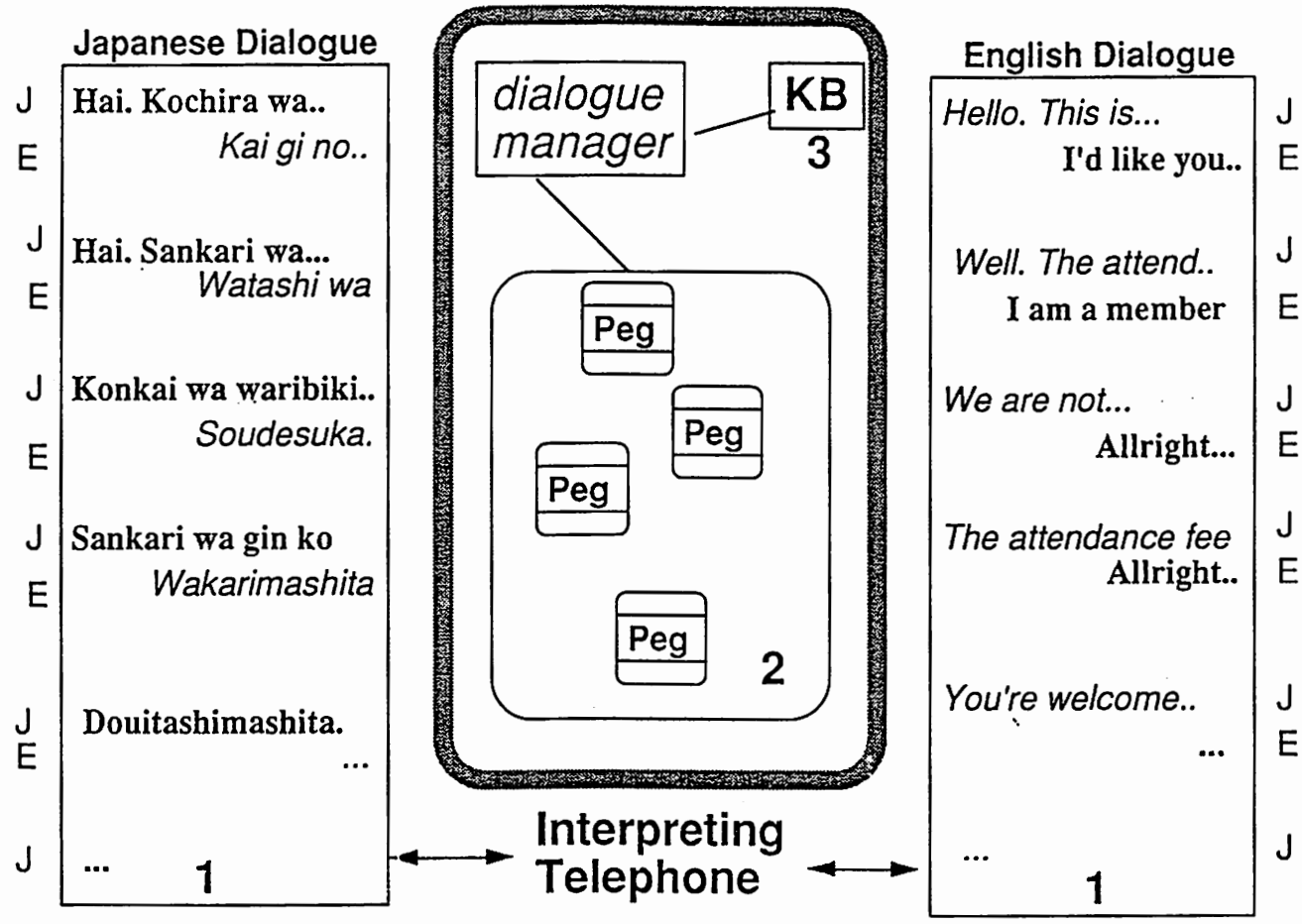
Information Acquisition Dialogue

E from J:
Is this the conference office
What if someone doesn't have a credit card
Registration fee
US dollars?
How much?
Hotel accommodation
arrange ourselves
Travel from Tokyo to Kyoto (fly or trian) how long does it take
Taxi from Kyoto station to the conference center
Sightseeing tour
Deadline date for registraion form application
Group rates for 50 attendees
Purpose or main theme of conference
When to submit paper or summary, how long should it be, J or E,
Presentation, slides or overhead
Outline of the conference
What to do t o participate, what about students?
Date, place,
Overseas speakers? translation to Japanese? to English?
Address of Secretariat
Registration form

J from E:
Name? Address?
Do you have a registration form?

38. Turning to the data in the sample dialogue we see that these are aptly characterized as question-answer dialogues where the English-speaking caller elicits information from the Japanese receptionist at the conference secretary's office. The exception occurs when the Japanese-speaker requests name and mailing address information from the prospective conference attendee. These dialogue segments are the focus of the remainder of this presentation.

Three Tiered Discourse



39. I will examine the representation of these address-elicitation dialogue segments in the three-tiered discourse framework using the dual monolingual discourse histories for Japanese and English and a single language-neutral representation using discourse pegs.

Address Ellicitation in Samples

[A] and [1]
We'll send you a registration form.
Your name and address, please?
My address is 23 Chayamachi, Kita-ku, Osaka.
My name is Mayumi Suzuki.
All right.
We'll send you a registration form.

[B]
Your name and address, please?
My address is 2-2 Tokui-machi, Higashi-ku, Osaka
My name is Taro Shimizu.
All right.

[4]
We'll send you the announcement of the conference, so please refer to it.
Your name and address, please?
Adam Smith.
My address is 2-27-7 Tamatsukuri, Higashi-ku, Osaka
All right.
Wed like to ask your phone number also.
Yes.
372-8018.
372-8018, right?.
Yes.
That's right.
Thank-you very much.
Good-bye.

40. Looking through the 10 sample dialogues we find in each sample, one such address-elicitation segment and they take the form on this slide.

Address Ellicitation in the Future

We'll send you the announcement of the conference, so please refer to it.
Your name and address, please?

Oveissi Mohammed

Oveissi Mohammed And is "Mohammed" your last name?

No. Sorry. Oveissi is my last name.

Would you please spell that?

Yes. That's O-V-E-I

O-V-E-I. umhmm.

S-S-I. Oveissi.

S-S-I. Oveissi
And the first name was Mohammed?.

Yes. Mohammed.

All right. And your mailing address.

My address is Grona gatan 7

B-R-

No. G-R-O-N-A. And gatan is G-A-T-A-N. And it's number 7

ok. Grona gatan 7.

And the city name is "Virserum" spelled V as in "Victor"-I-R

V-I-R. ok.

S-E-R-U-M as in "Mary"

Allright. Virserum.

The country is Sweden.

And that's in Sweden is it? Ok. And is there a postal code?

Yes. It's 57 (pause) 027/

57 027. All right. And wed like to ask your phone number also.

Yes. It's 372-8018.

372-8018, right?.

Yes.That's right. Thank-you very much. Good-bye.

41. The sample dialogues were not selected in order to exercise a dialogue management component and thus they are represented as quite simple exchanges. However, this slide is a constructed illustration of the sorts of exchanges that naturally occur when two people are negotiating detailed or symbolic information such as names and addresses. For example we see

- corrections "No. Sorry. Oveissi is my last name."
- requests for clarification: "Would you spell that please?"
- spelling (letters as lexical items): "O-V-E-I"
- clarification and repetition: "..Mohammed?" "Yes. Mohammed."
- lexical items associated with locations in various places in the world (note that the English-speaking caller may be a non-native speaker phoning from any country in the world, using English because it is one of the languages the IT has been programmed to handle, so for example, "gatan" means "street" in Swedish)
- numerals: "7"
- packaging of symbolic information, bounded by silent pauses:

"Virserum. V-I-R <pause> S-E-R-U-M"

- confirmation from listener during pause: "V-I-R.." <begin pause> "V-I-R ok" <end pause>
- language/culture-dependent clarification devices: "V as in Victor" and "niner" for 9 in English, or "Tokyo no Kyo" and "nanna" for 7 in Japanese.
- opening and closing of the address-elicitation segment signalled at the discourse level: "Your name and address please" ... "Thank-you very much."

DSIDS

Degree of accuracy increases dramatically

Lexical expectations
- numerals
- letters
- help phrases "s as in Sam," 9--> "niner"
- domain-specific "Higashi," "Shi," "Road," "Department"

Packaging of information (Clark)
- pauses *("five-seven <pause> o-five-seven")*
- clarification subdialogues *("Was that M as in Mary?")*
- temporary suspension of MT *("Department of Japan Studies")*

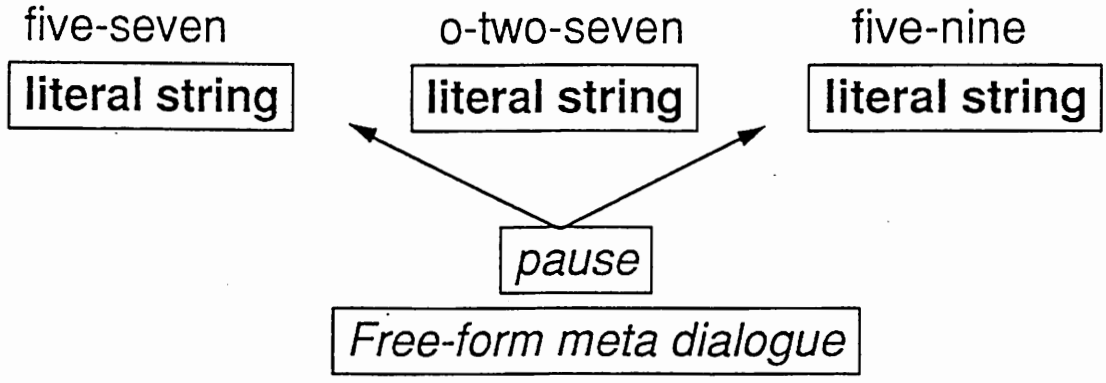
Expect Proper Nouns and Novel terms
("Department of Japan Studies")

Opening and closing indicated at discourse level

42. I have labelled these dialogue segments "DSIDS" or Detailed Symbolic Information Dialogue Segments. We can expect them to occur and take very similar forms in all sorts of dialogue where the purpose is information elicitation, e.g., making dinner or airline reservations, requesting detailed street directions, and so on. This slide reviews the properties of DSIDS language with emphasis on the special demands they place on the SR component.

The degree of accuracy required is much greater here than in dialogue in general and the penalty for error is also greater. For example, in the address-elicitation domain the misunderstanding of a single digit can cause an important letter to fail to be delivered. The SR system should expect to see more letters, more numerals, and more clarification devices in conjunction with them both. The weights or probabilities on grammar rules or lexical items should increase for domain-specific words so that, for example "avenue" would increase in likelihood relative to "haven't you" during DSIDS segments in this domain. Proper nouns and novel terms will occur frequently and may need to be passed through the MT system without translation. In order to recognize them the SR component needs to operate in a "phonetic typewriter" mode (H. Singer, personal communication) assuming that typed input is not available. Openings and closings as well as the internal structure of DSIDS can be recognized at the discourse level only, and are critical to proper behavior of the SR system. This fact establishes the necessity for a reliable protocol for dynamic communication between the dialogue processor and the SR component.

Symbolic Information Packaging



43. This slide illustrates the use of pauses by the information provider so that the recipient can take note and possibly request clarification during pauses. Thus, in spite of the symbolic nature of the dialogue inside the specialized DSIDS temporary environment, not all language is of a symbolic nature; since free-form dialogue can occur during pauses. Pauses as communication devices are useful between humans and the empathetic speaker can be expected to employ them cooperatively. This is one example of facts that emerge when we study the human-human dialogue as opposed to either human-machine dialogue surrounding the IT project. Pauses can also be exploited by the SR system for recognition of symbolic information in speech.

Discourse and SR

Opening and closing of DIDS

Cache recent words in discourse lexicon

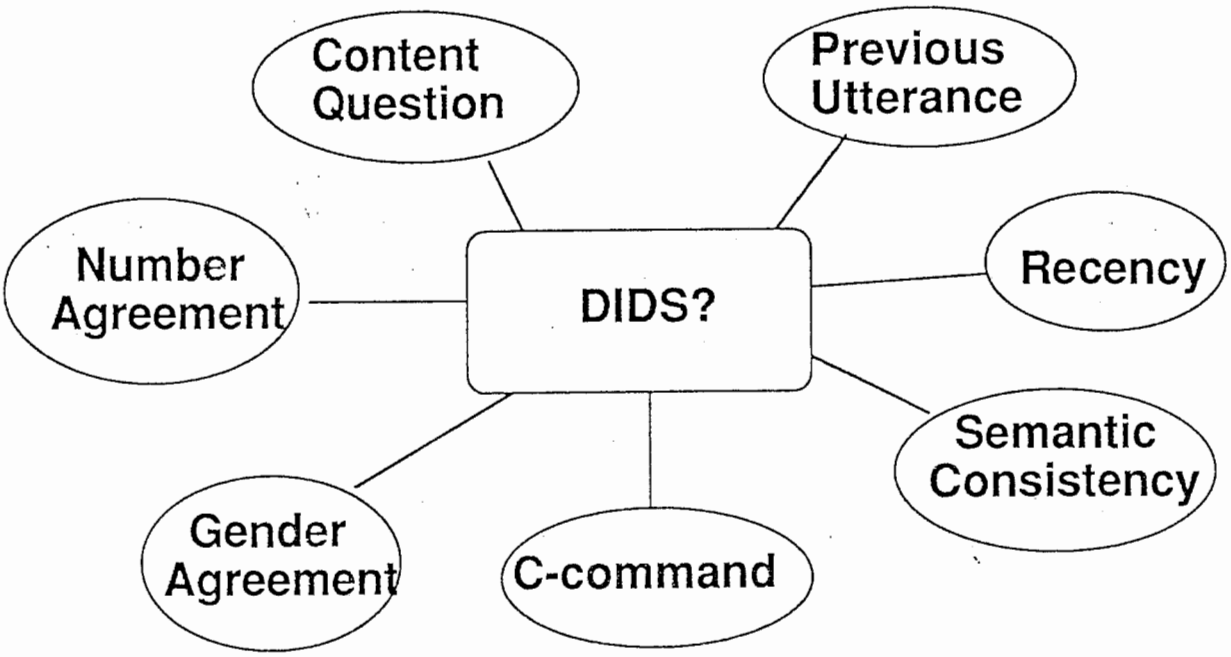
Pegs in focus increase probabilities for
- alternate descriptive phrases
- synonyms
- pronouns (she versus he)

As posterior process prioritize SR hypotheses based on discourse history

Pass through proper nouns

44. The important observation about DSIDS is that SR needs to know that the dialogue is temporarily in the environment of the specialized language (speech) behavior, but only discourse level observer is capable of detecting the entrance to or exit from such a segment. This is the motivation behind the search for an efficient technique for dynamic communication between the dialogue manager and SR components. How the SR component actually adjusts to the news that the IT dialogue has moved into a DSIDS environment is one area where work must be done. I plan to continue work with DSP researchers at MITRE, exploring the possibilities in this area.

Blackboard Algorithm for DIDS Opening/Closing



45. To close this discussion of DSIDS, we notice that the transition of the dialogue into such an environment is itself a probabilistic matter. So even though it is a determination for discourse-level processing to make, it is not always certain and can be handled through a combination of heuristics compiled by a simple voting algorithm and reported to SR as a probabilistic value.

Goals

Cooperate with and improve performance of SR

Cooperate with and improve performance of MT

Improve the UI dialogue attributes of IT system

46. This closing slide restates what I take to be three high-level goals for the discourse research component of the IT research effort.