[*AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, October 28, 1997]

# The Use of the Pegs Computational Discourse Framework as an Interlingua Representation

**Susann Luperfoy**
**luperfoy@mitre.org**

**Keith Miller**
**keith@mitre.org**

**The MITRE Corporation**
**1820 Dolley Madison Blvd.**
**McLean, VA 22102**
**USA**

## 1. Introduction

We treat the spoken dialogue interaction between a command and control system* and the user commanding that system through a conversational interface as an instance of real-time voice-to-voice machine translation (MT) similar to that seen in the Interpreting Telephone (IT) application domain. Voice-to-voice MT and spoken dialogue user interface systems share these important similarities from the perspective of software engineering:

> Translation is bi-directional with SL and TL alternating with each turn in the dialogue.
>
> The dialogue is interactive so the translation must occur at each step in the dialogue without lookahead.
>
> Input (on the user's side) is spoken natural language.
>
> Tracking the discourse is important for achieving correct translations.
>
> Like Verbmobil: there is a shared visual space that often affects interpretation.
> Like the Interpreting Telephone: user's facial expression and other extralinguistic cues are not available to translator.

Our approach to both applications relies on an interlingua representation that encodes properties of the dialogue history, properties of the world of reference and properties of the current state of the discourse only as essential for getting from source language input to target language output. The theoretical discourse framework (based on Heim 1982, Kamp 1981) that underlies the dialogue manager for our spoken dialogue user-interface systems supports the job of extracting and collecting information from the context, and facilitating human-machine language interaction in a multi-user environment. We will present our use of the mode-independent and language-independent 'discourse interlingua' representation (LuperFoy 1991) and the adapta-

---

* In this paper, we use the term "command and control system" in its broadest sense to mean systems in which a user issues commands to a backend system through the medium of human language.

tion of updating procedures needed for applying the framework to the task of translation. Empirical support for the dialogue theory and the implementation we describe comes from an observational study of one human interpreter carrying out the dialogue mediation functions of an interpreted telephone conversation (Miller et al, 1996).

Relative to MT systems that process text documents in batch mode, real-time, interactive, voice-to-voice MT systems give rise to a new set of requirements for the discourse processing component of computer-mediated human-human dialogue systems and automated dialogue systems in general. This paper describes how a prototype system for voice-to-voice interpretation has been used as a model for constructing interlingua-based human-interface 'translation' systems.

## 2. Dialogue Manager Tasks in Bilingual Dialogue Translation

Real-time, interactive spoken dialogue interpretation involves both dialogue management and dialogue tracking tasks. The former include management of turn-taking and ensuring that the correct channels of communication are open at the appropriate times, whereas the latter comprises the tracking of speech acts, surface form utterances, and mentions derived from the logical forms of those utterances. The dialogue manager requires multiple discourse processing strategies for the different tasks with which it is faced. To sufficiently interpret Source Language (SL) input in context, and to generate an appropriate Target Language (TL) result by assisting the generation component in discourse planning and realization, the discourse component must contend with the following three discourse processing tasks.

**1. Manage and Track the User-to-User Interaction**: First, the dialogue manager for real-time voice-to-voice MT systems mediates a potentially complex exchange between two clients. The IT dialogue manager enforces a simple turn-taking protocol so that it can impose a total order on utterances and partition input speech into non-overlapping turns, to ensure its own ability to record an accurate history of the dialogue utterances as they occur. The mediator must consider the unfolding of this collaborative dialogue in order to discern the discourse segment purposes of Grosz and Sidner (1979), speech acts, and rhetorical moves exchanged by the two human users. In complex, distributed modeling and simulation applications it is common to have multiple users who talk "through" the system with each other. Although our design for a dialogue manager does take this type of communication into account, it has been factored out in the current application. We restrict ourselves to focusing on discourse issues arising from a single user's mediated interaction with a backend system.

2. **Manage and Track the User-to-Backend System Interaction**: The bilingual dialogue is defined by the sequence of user input utterances. For example:

| The source language sequence | is translated into the target language utterance |
|---|---|
| "First Platoon increase speed by two zero miles per hour." | INCREASE(VELOCITY-SETTING, PLATOON-449, 20) |
| "Now halt" | SET(VELOCITY-SETTING, PLATOON-449, 0) |
| "Send an AWACS mission to coordinates 30-00-07N 042-56-32E." | CREATE_MISSION(MISSION1, AWACS, 9C3B, 30-00-07N 042-56-32E, 30000, 000300) |
| "Decrease altitude of that aircraft to 12,000 feet." | SET(AIRCRAFT10, ALTITUDE, 12000) |
| "Remain on that orbit for 6 hours." | SET(MISSION1, DURATION, 000600) |

The mediator thus alternates its attention to participate in and track two monolingual dialogues, shifting back and forth from one to the other, switching dialogue rule systems each time.

3. **Manage and Track the User-to-System Interaction** If the user or system detects a problem the dialogue manager has a monolingual (meta-) exchange with the client indicating the problem. Similarly, if the Speech Recognizer fails, or if the input utterance cannot be translated into a well-formed command in the language of the backend system, the dialogue manager must engage the user in a troubleshooting meta-dialogue. We assert that in its role as the "voice of the system", it is within the domain of the dialogue manager to completely control all dialogue between the user and the system. According to this model, the dialogue manager presents a unified interface to the user, rather than allowing individual components to interact with the user separately. These exchanges are also tracked by the discourse manager, represented internally using the same pegs interlingua (illuminated in section 3.2) used to represent the main bilingual dialogue.*

## 3. Our Approach

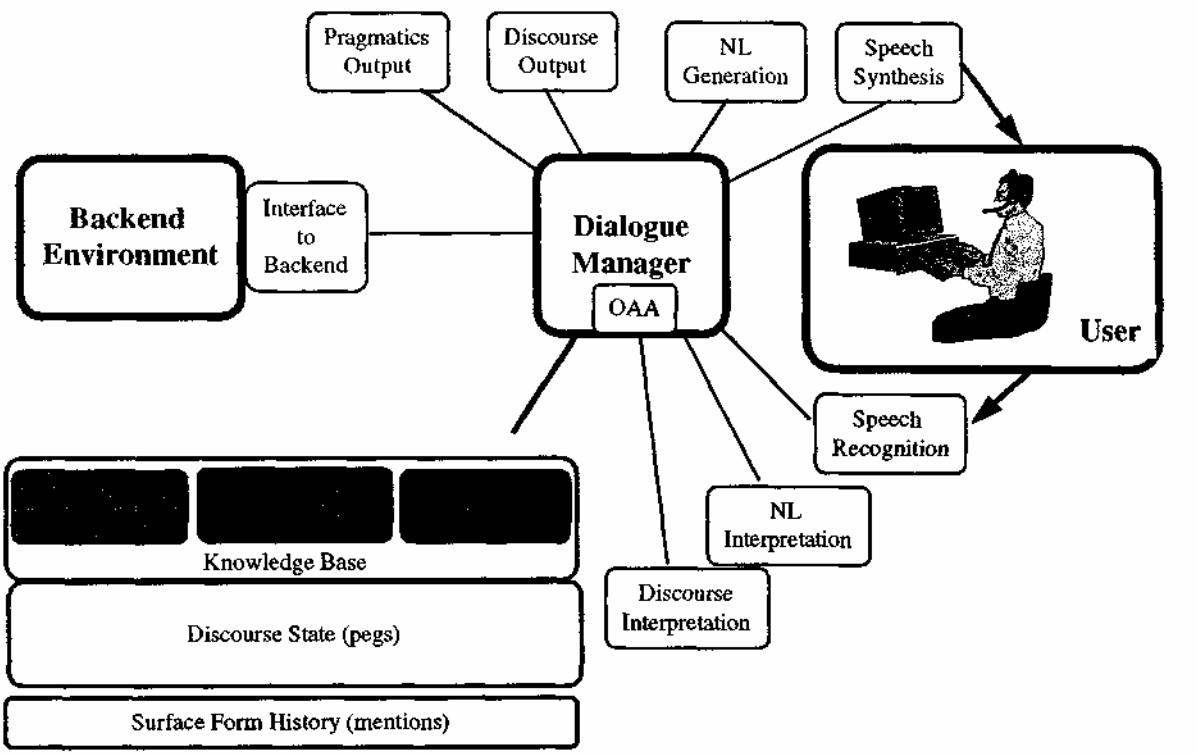


**Figure 1: The bilingual spoken dialogue metaphor applied to user-interface interaction.**

The Interpreting Telephony (Morimoto, et al., 1989) and the Verbmobil project (Kay et al., 1993), (Wahlster et al., 1993), (Quantz, et al., 1994), exhibit a dialogue management task similar to that of a user interface system that translates between spoken input English and the command language of the backend system. The IT application calls for spoken language understanding and generation plus bi-directional translation in a real-time operation environment. The tasks of the

* In the case of command and control systems, the main bilingual dialogue is the User-to-Agent interaction; in the case of the IT, it is the Japanese-English human-human dialogue.

discourse module of an IT system are to maintain an updated representation of the bare essentials from the ongoing discourse, and to make use of that representation to produce in-context interpretation of input utterances and to generate context-appropriate output utterances. Moreover, the current state of the art in speech recognition and Natural Language Processing (NLP), force a restriction of the subject matter of the computer-mediated human-human dialogue similar to the restrictions required in spoken dialogue user interface systems: e.g., the narrow domain of conference registration, or negotiation of date, time and venue for a future business meeting.

This situation is similar to that of command and control systems in two important ways: First, there are two languages involved. In the case of the IT, these two languages are Japanese and English; in the command and control case, the languages are a natural language and the language of the backend system. Secondly, the restricted domain allows for use of a reduced language model. Thus, the task for our software dialogue mediator is similar to that of the voice-to-voice machine interpretation system.

Two specific applications to which we have applied this model (shown graphically in Figure 1) include a Modeling and Simulation (M & S) system and a Web-based employee time reporting system. Whereas the language model for the interpreting telephone is limited to the domain of conference registration, in Modeling and Simulation (M & S) systems, for example, the world of discourse is limited to imperatives regarding objects such as platoons, ships, and aircraft, and the domain model is limited to knowledge about these entities. For the voice interface to the time reporting system, the world of discourse is limited to commands and queries regarding task charge numbers, hours charged, task leaders and their contact information, and the domain model reflects this constrained problem space.

### 3.1 Theoretical Framework

The discourse-processing framework proposed in this paper builds on theoretical and computational models of discourse (Kamp, 1981), (Heim, 1982), (Landman, 1986), (Carlson, 1977), (Sidner, 1979), (Webber, 1978), (Grosz et al., 1985), (Brennan, 1993), with modification to account for problems encountered in real-time processing of bilingual spoken dialogue. The framework was originally developed as a computational adaptation of Heim's File Change Semantics (FCS) which, for our purposes, is equivalent to Kamp's Discourse Representation Theory (DRT). The theoretical framework, hereafter labeled "PCS/DRT," was extended to handle the discourse phenomena encountered when people talk to computers about the contents of a knowledge base, as discussed in (LuperFoy, 1991; LuperFoy and Rich, 1992; LuperFoy, 1992). Factors that imposed new constraints on the discourse framework and require extensions to the semantic theory include uncertain and incomplete world knowledge underlying the semantic system, incomplete logical interpretations of utterances that form the input to the discourse module, and inevitable disfluencies encountered in human-machine dialogue interaction, all of which are faced by the Interpreting Telephone dialogue manager given a particular utterance.

### 3.2 The Model

A key innovation to the semantic theory was the structuring of each File Card (Discourse Marker) into three tiers:

1. Surface Form Tier: Objects on this tier are called mentions. They arise from the logical form representation of the input utterance, as produced by the sentence processor. Each time a construct is mentioned by either speaker in the discourse, a new mention is added to the discourse representation.

2. Discourse Model Tier: Objects on this tier are called Discourse Pegs or simply Pegs. There is one peg maintained in the discourse tier for each construct under discussion though

any peg may have several mentions associated with it, one for each occasion in which it is involved in an input utterance. This is the core of our interlingua representation system. It is here that the dynamic model of the current discourse state is maintained.

3. Belief System Tier: Objects on this tier are the structures of the static knowledge representation system that provides the semantics for the language understanding system. These objects are not normally deleted, created, or modified by the dialogue interpretation process.

The structuring of discourse objects into three tiers was to address a discovered need to represent the form and behavior of three distinct information types available to a dialogue processor during interpretation. There was a corresponding need for three sets of processes to manipulate these information classes. For example, the collection of pegs in the discourse model is structured by attentional focus, computed for each utterance according to (Grosz et al., 1985) while mentions are maintained and decay in linear order corresponding to their time of introduction into the discourse.

Each input utterance is stored in a Tier 1 structure for the monolingual dialogue in the SL of the utterance. The result of its translation appears on the opposing Tier 1 structure. This surface form data structure is important for constraining within-language discourse phenomena such as Japanese morphological marking of honorifics, English interpretation and generation of one-anaphoric NP's, and the marking of gender, number, case, and definiteness on English noun phrases.

The discourse pegs level (Tier 2) represents language-neutral properties of the bilingual dialogue including attentional focus structuring of constructs under discussion, and thus can serve as a working pivot language. The language of discourse pegs does not qualify as an interlingua in the strong sense defined by Nirenburg et al. (1992) because it is not a complete knowledge representation language capable of encoding full meanings of natural language utterances. It is merely a formalization of the language-neutral computational structures required to perform discourse level transfer from SL to TL in a limited domain. It is an incomplete semantic representation, and its content is determined in part by the task-driven demands of the translation application and by a contrastive analysis of the two languages involved. For example, in the conference registration domain there will be one peg for the Japanese speaker, J, one peg for the English speaker, E, one peg for the conference, and one for the registration form that J intends to send to E. That registration form peg is linked to several mentions, some in Japanese input, some Japanese output, some English input, and others English output.

The third tier of the original framework is the belief system or knowledge base (KB). Opinions vary as to how much knowledge is essential for adequate translation. Without taking a stand on this issue, we chose a minimal KB in this application to explore the adequacy of the dual surface tiers plus discourse model tier for the job of performing discourse-level transfer between English and Japanese. The KB contains a typological model of the of conference registration domain. Full understanding (knowledge-based NLU) is only required during user-system clarification/repair dialogues when the IT is acting as a sort of expert system with regard to its own functionality and the IT situation. In this situation, it must be able to understand and respond to queries about previous utterances, the identity of the participants, etc. However, it does not require full knowledge of conference registration and cannot answer questions about the conference; that is the expertise of the user in the registrar's office.

## 4. Empirical Study

In a “wizard of oz” study, (Miller et al., 1996; Duff et al., 1995), our interpreting telephone wizard used an electronic voice modulator to make his speech resemble synthesized speech, and he “translated” English utterances to English utterances by placing one caller on hold and simply

repeating or paraphrasing the input for the other subject. The wizard's task therefore required no bilingual effort, just the need to retain the content of one user's turn long enough to convey it to the other user. In this way, we were able to isolate the dialogue management requirements of the system from issues of speech processing and translation. We were in effect simulating the highly idealized state of affairs in which all technical problems except discourse processing have been completely overcome; given our experimental design all component processes combined (apart from dialogue management) impose a near zero run-time computational cost while being essentially error free. In this way, the varying cost of dialogue management is exposed.

Conference registration was the topic of conversation, following the original project at Advanced Telecommunications Research (ATR) Laboratories in Kyoto, Japan, (Morimoto, 1992). The scenario was that of a caller phoning the office of the registrar for an international conference. The conversation was limited to topics relevant to the caller's registering for a conference: venue, deadlines, exchange of address information, etc.. The exercise was carried out over the telephone with sequestered subjects so that there was no shared visual context and no exchange of paralinguistic information through intonation, facial expressions, or deictic behavior of any kind. This disallowed exchange of non-linguistic data directly between subjects instead of through the wizard mediator when acting as dialogue agent. Our wizard adhered to a restricted grammar in order to discourage three-agent behavior during the main dialogue. Experimenters video taped wizard and subjects during the main session and during post-session interviews. This experimental model, which is obviously based on a human-human dialogue can also serve to provide insight into the design of the discourse manager in a command and control system.

The data revealed unpredicted discourse behavior that lead us to two conclusions regarding proper design of the IT dialogue manager. First, even at its unrealistic fastest, the half-duplex transmission (where only one utterance can be processed at a time) was too slow. The impatience of the subjects was not the only problem with the half-duplex transmission, however. Subjects also had frequent difficulty tracking the dialogue as long pauses left them unprepared for resumption of the dialogue. Our voice-only wizard design provided help dialogues, but repair dialogues themselves were problematic. Subjects easily lost track of the main context during sidebar discussions with the wizard and were unable to gracefully terminate a repair dialogue sequence and indicate to the wizard that the next utterance was intended for the other user as part of the main dialogue. Furthermore, since the IT wizard's voice was the same during translation output and sidebar dialogue, there were no overt cues to tell the user whether the source of the system's output utterance was the other user or the translation device itself

Our conclusion was that voice-only interaction would be inadequate for many user-system dialogue tasks. A mixed-modality interface that provided a visualization of the dialogue setting would help users track the main dialogue and distinguish human-human from human-system dialogues, which could be carried out in parallel. Our model calls for a mixed modality system incorporating reactions to these findings, so that various sorts of contextual information are distinguished visually, and the mode of interaction varied with information stream: the history of the dialogue presented in text, immediate status of the system presented as a highlighted icon, visual image of the other user as a still photo or video feed, and user interface dialogue that could be carried on locally on one user's machine in parallel to the main, translated dialogue. Keyboard input would be available for entering symbolic data, such as postal addresses, that do not require translation and are not be easily conveyed through speech even in human-human dialogues. Although these multiple interaction modes are possible, it is crucial to note that all communicative acts (mouse clicks, keying in information, voice interactions) are mediated by the dialogue manager. These observations carry over to command and control systems, in which visual support would help the user to offload the cognitive task of maintaining context for the various modes of communication, and multiple modes of interaction would increase the naturalness of the user's exchanges with the system. In addition to these recommendations, a complete description of this study design and conclusions appears in Duff et al., 1995.

## Bibliography


Brennan, S. E. and E.A. Hulteen. (1993). *Interaction and Feedback in a Spoken Language System.* AAAI Fall Symposium Series Symposium on Human-Computer Collaboration: Reconciling Theory, Synthesizing Practice.

Clark, H. and E. Schaefer. (1987). Collaborating on Contributions to Conversations. *Language and Cognitive Processes,* pp. 19-41.

Cohen, P., M. Dalrymple, D.B. Moran, F.C.N. Pereira, J.W. Sullivan, R.A. Gargan, J.L. Schlossberg and S.W. Tyler. (1989) *Synergistic Use of Direct Manipulation and Natural Language.* In Proceedings of CHI, pp. 227-233.

Dahl, D. and C. N. Ball. (19,90). *Reference Resolution in PUNDIT* (Tech. Report). UNISYS.

Dorr, B., Lin, D., Lee, J.H. and Suh, S. (1994) *A Parameter-Based Message-Passing Parser for MT of Korean and English.* In Lin and Dorr.

Duff, D., E. Kim, S. LuperFoy, K. Miller (1995) *Some Effects of Electronic Mediation: An Observational Study of Dialogue Management for the Interpreting Telephone.* Georgetown University Roundtable presession on Computer-Mediated Communication.

Furuse, O. and H. Iida (1992). *An Example-Based Method for Transfer-Driven Machine Translation.* In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.

Grosz, B. (1977). *The Representation and Use of Focus in a System for Understanding Dialogs.* In Proceedings of IJCAI5.

Grosz, B. and C. Sidner (1985) *The Structures of Discourse Structure.* Technical Report CSLI-85-39, Center for the Study of Language and Information. SRI International

Guha, R. V. and D. Lenat. (1990). Cyc: A Mid-Term Report. *AI Magazine*

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases.* Ph.D. thesis, Department of Linguistics, The University of Massachusetts.

Hobbs, J. and M. Kameyama.(1990) Abduction for Machine Translation. Proceedings of COLING.

Hollan, J., E. Rich, W. Hill, D. Wroblewski, W. Wilner, K. Wittenberg, J. Grudin, and members of the Human Interface Laboratory. (1988). *An Introduction to HITS: Human Interface Tool Suite* (Tech Report). MCC

Kamp, H. (1981). A Theory of Truth and Semantic Representation. In J. Gooenendijk, T. Janssen, and M. Stokhof (editors), *Formal Methods in the Study of Language, Part 1.* Mathematisch Centrum, Amsterdam, The Netherlands.

Kay, M., J.M. Gawron, P. Norvig, (1993). *Verbmobil: A Translation System for Face-to-Face Dialog.*

Landman, F. (1986). Pegs and Alecs. *Linguistics and Philosophy,* pp. 97-155.

LuperFoy, S. (1989). The Semantics of Plural Indefinite Anaphors in English. *Texas Linguistic Forum,* pp. 91-136.

LuperFoy, S. (1991). *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions.* Doctoral dissertation, Department of Linguistics, The University of Texas.

LuperFoy, S. (1992). The Representation of Multimodal User-Interface Dialogues Using Discourse Pegs. Proceedings annual meeting of the Association for Computational Linguistics.

Miller, K., S. LuperFoy, E. Kim, D. Duff (1996). Dialogue Management for Computer-Mediated Spoken Bilingual Dialogue. The Electronic Journal of Communication / La Revue Electronique de Communication. Volume 6 (3): Computer-mediated Discourse Analysis, ed. S. Herring.

Miller, S., M. Bates, R. Bobrow, R. Ingria, J. Makhoul, R. Schwartz (1995). Recent Progress in Hidden Understanding Models. Proceedings of the ARPA Spoken Language Systems Technology Workshop.

Morimoto, T., K. Ogura, K. Kita, K. Kogure and K. Kakigahara 1989. Spoken Language Processing in SL-TRANS. ATR Symposium on Basic Research for Telephone Interpretation, Tokyo.

Nirenburg, S., J. Carbonell, M. Tomita, K. Goodman (1992). Machine Translation: A Knowledge-Based Approach. Morgan Kaufman Publishers.

Oviatt, S.L., P.R. Cohen and A. Podlozny (1990) *Spoken Language in Interpreted Telephone Dialogues.* SRI International Technical Note 496.

Polanyi, L. (1985) *A Theory of Discourse Structure and Discourse Coherence.* In Proceedings of the 21st Annual Meeting of the Chicago Linguistics Society.

Quantz, J.J., M. Gehrke, U. Kuessner, B. Schmitz (1994) *The Verbmobil Domain Model.* Technical Report Number 122 of Projektgruppe KIT at Technische Universitaet Berlin.

Rich, E. A. and S. LuperFoy (1988) An Architecture for Anaphora Resolution, *Proceedings of Applied ACL*

Seneff, S., V. Zue, J. Polifroni, C. Pao, L. Hetherington, D. Goddeau, and J. Glass. (1995). Preliminary Development of a Displayless PEGASUS System. Proceedings of the ARPA Spoken Language Systems Technology Workshop.

Sidner, C. L. (1979) *Towards a Computational Theory of Definite Anaphora Comprehension in Discourse.* Doctoral dissertation, Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

des Tombe, L. (1995). *Compensation in Machine Translation.* Presentation to the Hildesheim University workshop on Machine Translation and Translation Theory.

Wahlster, W. (1993) Verbmobil-Translation of Face-to-Face Dialogues. Technical Report, German Research Centre for Artificial Intelligence (DFKI).

Waibel, A., H. Sawai and K. Shikano. (1988). Modularity and Scaling in Large Phonemic Neural Networks. Technical Report (TR-1-0034), ATR, Tokyo.

Walker, M. (1991) *Redundancy in Collaborative Dialogue.* AAAI Fall Symposium Series.

Ward, G., G. McKoon and R. Ratcliff. (1991) *How Morphosyntactic and Pragmatic Factors Affect the Accessibility of Discourse Entities.* AAAI Fall Symposium Series.

Webber, B. (1978) *A Formal Approach to Discourse Anaphora.* Doctoral dissertation, Division of Applied Mathematics, Harvard University.

Y. Yamazaki and Morimoto, T. 1994. ATR Research Activities on Speech Translation. IEEE workshop on Interactive Voice Technology for Telecommunications Applications, Kyoto. pp. 61-66.