

A Multimodal Dialogue System for Playing the Game “Guess the card”*

*Un sistema de diálogo multimodal para jugar el juego de
“Adivina la Carta”*

Ivan Meza, Elia Pérez, Lisset Salinas, Hector Aviles, Luis A. Pineda

IIMAS, UNAM

Ciudad Universitaria

{ivanvladimir,liz,haviles}@turing.iimas.unam.mx, eliappp@gmail.com, luis@leibniz.iimas.unam.mx

Resumen: En este artículo se presenta un sistema conversacional en español hablado y con visión computacional que juega el juego de “adivina la carta” con el público en una cabina de la exhibición permanente del museo de las ciencias Universum de la Universidad Nacional Autónoma de México. Se presenta el modelo conceptual así como la arquitectura del sistema. Se incluye también la transcripción de un fragmento de un diálogo real colectado en el museo, así como una evaluación preliminar del sistema. Se concluye con una reflexión acerca de los alcances de la presente metodología.

Palabras clave: Sistema de diálogo, Administración del diálogo, Sistemas multimodales con habla y visión

Abstract: In this paper a dialogue system with spoken Spanish and computer vision that plays the game “Guess the card” with members of the general public in a permanent stand of the science museum Universum at the National Autonomous University of Mexico (UNAM) is presented. The conceptual and architectural guidelines for the construction of the system are presented. An excerpt of an actual dialogue collected at the museum is also included, along with a preliminary evaluation of the system. The paper is concluded with a reflection about the scope of the present methodology.

Keywords: Dialogue system, dialogue manager, Multimodal Speech and Vision Systems

1. Introduction

Over the last ten years we have been developing a technological infrastructure for the construction of spoken dialogue systems in Spanish supported by multimodal input and output, including the interpretation of images through computer vision and the display of pictures and animations to support the speech output. We are interested in applica-

tions in fixed stands, like the one presented in this paper, but also with mobile capabilities. In this latter case, we developed the robot Golem, which was able to act as the guide of a poster session through a spoken Spanish conversation. We have now developed a new application to demonstrate this kind of technology in a permanent stand at the Universum science museum of the National Autonomous University of Mexico (UNAM). In this stand the system is able to play the game “Guess the card” with the public, mostly children, through a fluent conversation in spoken Spanish, where the linguistic behavior is coordinated with computer vision and the display of pictures to support the system’s output. The stand has a table with ten cards with astronomical motives on it, and the system chooses one of them; then the human user asks up to four questions to identify the card in ques-

* We thank the support and effort at IIMAS by Fabian Garcia Nocetti, Wendy Aguilar, Hayde Castellanos, and the visiting students Carlos Hernández, Edith Moya, Aldo Fabian, Karen Soriano, Nashielly Vasquez, Miriam Reyes, Ramón Laguna and Tania Pérez. We also thank the support and help provided at the museum Universum by René Drucker, Lourdes Guevara, Gabriela Guzzy, Luis Morales, Emmanuel Toscano, Jimena Reyes, Brenda Flores, Germán Albizuri, Pablo Flores, Esteban Estrada, Esteban Monroy, Ana Lara, María Agonizantes, Diego Álvarez, Addina Cuerva, Claudia Hernández and León Soriano.

tion; when this interrogatory is finished the system asks the user to show the card that he or she thinks was chosen by the system. Finally, the system interprets such a card through computer vision and acknowledges whether the user guessed the card, or tells him or her which one was the right one.

In this project we have focused in the definition and implementation of generic architecture to support multimodal dialogues, and in the quick development of specific applications. The present architecture is centered around the notion of dialogue model specification and interpretation. Dialogue models are representations of small conversational protocols which are defined in advance through analysis. These units are then assembled dynamically during the interaction producing rich and natural conversations. The central component of the system is the dialogue manager (DM); this program interprets the dialogue models continuously, and each interpretation act corresponds to a conversational transaction. The main tenant of our approach is that dialogue acts are expressed and interpreted in relation to a conversational context that is shared between the speaker and hearer. We pose that the interpretation context has two main parts: a global context that holds for whole of the conversational domain, and needs to be identified in advance through analysis, and a specific context that is built dynamically along each particular conversation. Dialogue models can be thought of as representations of the global context, and the history of the interpretations and actions that are performed in every particular conversation constitute the specific context. Each particular application is defined as a set of dialogue models, but the DM is a generic application independent program.

In this paper we illustrate the present conceptual architecture with the application “guess the card” at Universum meseum. Section 2 presents the main characteristics of the dialogue manager. The architecture of the system is presented and discussed in section 3. Section 4 specifies the task and shows an excerpt of an actual dialogue collected at the stand. Section 5 illustrates the dialogue models for this application. A preliminary evaluation of the system is presented in Section 6. The implementation details are presented in Section 7. Finally, in section 8 we present our conclusions about the implications of the

present theory and methodology for the construction of this and similar systems.

2. Dialogue Manager

The dialogue manager interprets the conversational protocols codified in the dialogue models, and coordinates the system’s perceptions, both linguistic and visual, with the system’s actions. It also keeps track of the dynamic conversational context, which is required to make interpretations and perform actions that depend on the previous communicative events in the current conversation. We are interested in modeling practical dialogues in which the conversational partners “visit” conversational *situations* with highly structured expectations about what can be expressed by the interlocutor or about the visual events that can occur in the world, which we call expected intentions or *expectations*. This information forms a part of the global context and is used in all interpretation acts, and also to produce the corresponding relevant *actions*.

Situations, expectations and actions of an application domain are encoded through *dialogue models*. A dialogue model is represented as a directed graph (cycles are permitted). Situations are represented as nodes and edges are labeled with expectation and action pairs. If the expectation of an edge is satisfied by the current interpretation, then the corresponding action is performed. Situations can have one or more input and output expectation-action pairs. Situations are typed according the modality of the expected input; the main types are *listening* or linguistic and *seeing* or visual. There is also a special type of situation in which a full dialogue model is embedded. Situations of this type are called *recursive*. When a recursive situation is reached, the current dialogue model is pushed down into a stack, and the embedded model is interpreted, so the conversation as a whole has a stack structure. All dialogue models have one or more final situations, and when these are reached, the model’s interpretation process is terminated. If there is a dialogue at the top of the stack it is pop up and its interpretation is resumed; otherwise the dialogue as a whole is terminated. In this sense, dialogue models correspond to recursive transition networks (RTN), which have the same expressive power of context free grammars.

All dialogue models have an *error* situa-

tion. When the input message or event is not expected, or cannot be assigned an interpretation, the system reaches an error situation, and starts a recovery conversational protocol. In the default case, it produces a recovery speech act (e.g., *I didn't understand you, could you repeat it please?*); at this point the dialogue reaches again the situation in which the communication failure occurred and resumes the conversation with the same context. However, the error situation can also embed a full recovery dialogue to handle specific recovery patterns to achieve grounding at the communication and agreement conversational layers (Clark y Schaefer, 1989; Pineda et al., 2007).

Expected intentions and actions are expressed through abstractions that are independent of the expression used by the interlocutor and of the actual patterns that appear on the visual field of the system. These abstractions allow to capture a wide range of possible concrete communication behavior. Accordingly, the analysis of a task domain corresponds to the identification of possible speech act protocols that are observed empirically in the domain, and this analysis is codified in the dialogue models.

In our implementation, expectations are expressed through a declarative notation representing speech acts and actions. Actions are also specified declaratively through Multimodal Rhetorical Structures (MRS); these are lists of basic rhetorical acts, defined along the lines of the Rhetorical Structure Theory (RST) (Mann y Thompson, 1988). Although the specification of MRS is also modality independent, the basic rhetorical acts have an associated output modality. Accordingly, a MRS is thought of as "paragraph" in which some of its sentences are rendered through speech, but others may be rendered visually, as texts, pictures, animations and video. The specification of speech acts and rhetorical acts can be expressed through concrete expressions (e.g., constants and grounded predicates), but also these can be expressed through propositional functions.

The notation for transitions is illustrated in Figure 1, where the situation s_f is reached from s_i if the corresponding expectation is satisfied; during this transition the *MRS* is performed by the system. The specification of a transition with a concrete expectation and a concrete action, that can be named through

constants, is illustrated in Figure 2. The specification of a transition involving propositional functions is illustrated in Figure 3. In this case the expectation has some concrete information defined in advance in the dialogue model, but its satisfaction requires that some content information, represented by the variable x , is collected from the actual message or event in the world. Expectations and actions in dialogue models are parametric objects, as illustrated in Figure 3. Situations can also have parameters, and this mechanism permits the specification and flow of information along the conversation.

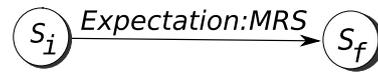


Figure 1: Specification of a situation's transition.

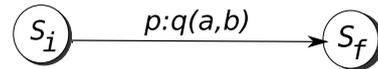


Figure 2: Specification of a concrete intention-action pair.

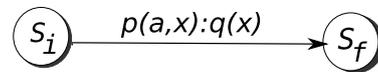


Figure 3: Specification through propositional functions.

Concrete expectations and actions interpreted and performed by the system (i.e., grounded interpretations and action specifications) are collected in the conversation history, where the stack structure of the dialogue is also preserved.

In order to access the dialogue history, expectations and actions can also be specified through domain specific functions, as illustrated in Figure 4. The arguments of these functions are domain specific information, the current dialogue model, and the conversation history. When the labels of an edge are specified through functions, these are evaluated first, and the values of the functions determine the system behavior (i.e., the actual expectation that needs to be satisfied to follow the corresponding edge, the action that is produced by the system or the conversational situation that is reached if the expectation is satisfied). This functional machinery per-

mits also the resolution of terms an expressions on the basis of the discourse information (i.e., anaphoric inferences). The definition of these functions extends the expressive power of the formalism, but preserves an implicit graph directed process, with the corresponding computational advantages. For this reason we call this formalism Functional Recursive Transition Networks (F-RTN).



Figura 4: Example of functional arc.

3. Architecture

Linguistic and visual events need to be recognized and interpreted before they can be matched with the expectations defined in the dialogue models. This is, the input information needs to be interpreted in order to be used in conversation. To model the relation between linguistic and visual interpretations and the information codified in the dialogue models we have developed a three-layers conceptual architecture; the top-level corresponds to the interpretation of dialogue models, as mentioned. The bottom-level corresponds to the recognition level in which the external information (e.g., speech or images) is translated into an internal data structure. However, the product of a recognition process at this bottom level is thought of as an “uninterpreted image” or “pattern”; this is, as an instance of a data-structure which has not been assigned meaning (e.g., If a person does not know Greek, but is familiar with the Greek alphabet, he or she can recognize that a text written in Greek is in fact a text, but is unable to tell what does it mean. So, for this person, a string of Greek symbols is an uninterpreted image). The architecture contains also an intermediate level which is constituted by modality specific interpreters; the role of these interpreters is to assign meanings to the uninterpreted images in terms of the expectations of the current conversational situation. In our scheme, expectations are also used as indices to memory objects, and an interpretation act consists of matching a memory object indexed by a current expectation with the uninterpreted image produced by the recognition device. The output of this process is the interpretation of the speech act performed by the interlocutor, or the inter-

pretation of the event perceived in the world. Interpretations (i.e., the output of interpretation process) and expectations are codified in the same format in the dialogue models. In our scheme, the bottom and intermediate levels correspond to low and high level perception, and we think of perception as a process that takes an external stimuli and produces its interpretation in relation to the context. The interpretation process is illustrated in Figure 5.

For the present application we define a speech and a visual perception process. The goal of speech perception is to assign an interpretation to the speech act performed by the human user. When the DM reaches a listening situation it feeds the language interpreter with a set of expectations in a top-down fashion. In the present implementation each expected intention has an associated regular expression stored in memory, which codifies a large number of ways to state such intention. The language’s interpreter recovers such a regular expression and applies it to the text produced by the ASR system; if this match is successful, the expected intention, with the values recovered from the input speech, becomes the interpretation. Figure 5 illustrates this flow of information.

Visual interpretation proceeds in the same way. In this case, when the DM reaches the seeing situation, it expects to see one among the ten cards, and this information is feed top-down from the DM to the visual interpreter. This in turn activates the vision recognition module, which provides a set of features codifying the image of the card in a bottom-up fashion. Visual expectations are also indices the image of the card, and this association is also codified in memory. For visual interpretation, the features of the external image are matched with the cards codified in memory, and the interpretation corresponds to the expectation with the largest number of matches.

For the system’s output, when an expectation is matched with the interpretation of the current input stimuli the systems performs the corresponding action (i.e., as defined by its associated MRS). MRSs are also abstract specifications that need to be instantiated in terms of the specific input interpretation and the dynamic contexts. This specification is performed by modality specific programs, and rendered by modality spe-

cific devices (e.g., the speech synthesizer and the display drivers to render texts, pictures or videos).

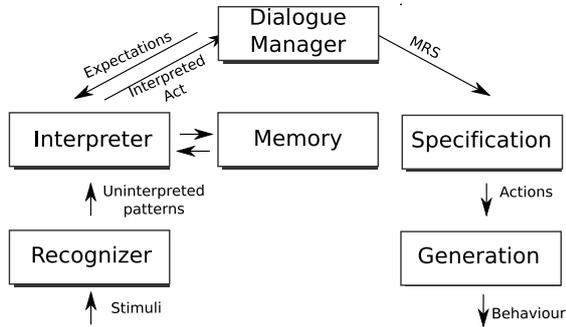


Figure 5: Dialogue system architecture

4. Task

The present application plays the game “guess the card”. The goal of the human user is to guess a card that is chosen by the system from a set of ten cards. The cards have astronomical motives (e.g., the sun, a telescope, etc). Since the system is placed in a science museum oriented to children most users are children aged between 10 and 14 years old. In the stand there is a table with the cards on it; in a typical session the system introduces itself, asks for the name and age of the child, and explains the rules of the game. At this point, the system tells the user that it has chosen one of the cards. The child is then allowed to ask up to four questions about the features of the card in question. At the end of the interrogatory, the child is asked to take the card that he or she thinks is the right one and place it in front of the system’s camera; then, the system confirms whether the child has won the game, or tells the child which card was the right one.

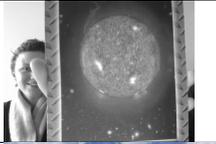
Table 1 shows an excerpt of a dialogue between a user and the system. Each row in the table contains the turn number, the speaker, the elocution, and the visual input and output produced by the user and system. In turns 1 to 3 the system greets the user and asks for his or her name. Turns 40 to 47 correspond to two questions with the corresponding system’s answer. This fragment includes two failures with the corresponding recovery strategies. Finally, from turns 48 to 52 the user shows the card with the corresponding system answer. In particular, in this dialogue, the system identified the card suc-

cessfully at the second time, and it was in fact the card chosen by the system.

5. Dialogue Models for the Application

This application has two dialogue models. The main defines the greeting part, the choosing of the card, the actual cycle of questions and answers (as a recursive situation) and the final salutation. The embedded model handles the interrogatory, and the verification of the child’s choice.

Figure 6 illustrates the embedded dialogue model ($r_1(C)$), which has the card chosen by the system as its parameter. At this point the system asks the first question; no input is required since the input edge is labeled with an empty expectation (the ϵ symbol). The system reaches the listening situation l_4 and waits for an answer. This situation has two output edges, so there are two active expectations at this point of the dialogue. One is that the user asks for a feature of the card (e.g., “is it red?”) and the other is that the user states what is the card in question (e.g., “is it the sun?”). This latter expectation is satisfied when the child feels that he or she has enough information, regardless whether the four questions have been made. Notice that this is an indirect speech act, as an assertion has been made through a question; however, the system performs the right interpretation in terms of the reference to the entity (i.e., the sun). Questions about features are interpreted through a similar referential heuristics. If the user asks for a feature, the expectation $feature(x)$ is satisfied, the MRS $validate(C, X)$ is performed and the situation n_3 is reached. This MRS checks whether the card chosen by the system has such a feature, and renders a relevant feedback text (through speech) and picture (displayed on the screen). However, if the child states the card, the system ask him or her to confirm whether he or she wants to the end the game, and situation l_5 is reached. Situation n_3 has only one output edge with an empty expectation; however, both the MRS and next situation of this edge are specified through functions that depend on the dialogue history. In particular, how many questions have been made. So, the actual rhetorical act produced at each transition would vary depending on how advanced is the dialogue; in case there are still questions to be made the situation l_4 is reached

Turn	Speaker	Interaction	Screen
1	S	Hola a todos me llamo Golem y éste es el juego de adivina la carta. Tenemos jugador, ¿cómo te llamas?/ <i>Hello everyone, my name is Golem and this is the game of guess the card. We have a player, what is your name?</i>	
2	U	Olimpia	
3	S	Hola Olimpia, gusto en conocerte, ¿cuántos años tienes?/ <i>Hello Olimpia, nice to meet you, how old are you?</i>	
...			
40	U	¿redondo?/ <i>round?</i>	
41	S	Sí, vas bien. Lo que hay en la carta es redondo. Sólo te queda esta última pregunta así es que piénsala muy bien/ <i>Yes, you are right. In the card there is something round. This is your last chance think it well</i>	
42	U	–	
43	S	¿Podrías repetirlo?/ <i>Could you repeat it?</i>	
44	U	–	
45	S	¿Podrías repetirlo?/ <i>Could you repeat it?</i>	
46	U	¿Es un planeta?/ <i>Is it a planet?</i>	
47	S	Lo siento, en mi carta no aparece un planeta/ <i>I am sorry, there is no planet in the card</i>	
48	S	Ha finalizado tu ronda de preguntas, por favor muéstrame la carta que crees que escogí/ <i>You have no chances left, please show me the card you think I chosed.</i>	
49	U	[<i>The user shows the card</i>]	
50	S	No tengo muy buena vista, es ésta la carta que me mostraste/ <i>I do not have good eye sight, is this the card you showed me?.</i>	
51	U	No no/ <i>No no</i>	
52	S	Ok, muéstrame la carta/ <i>Okay, show me the card</i>	

Cuadro 1: Excerpt of a dialogue between the user (U) and the system (S).

again; however, when the child has exhausted the four chances, the dialogue reaches the situation n_4 ; this situation has also only one output edge with an empty expectation, and the situation s_1 is reached deterministically. This is a seeing situation where the system asks the child to show it the card, entering

the final part of the dialogue. The remaining nodes and edges of the graph are understood in a similar way.

6. Evaluation

We have performed a preliminary evaluation of our system using the user-satisfaction

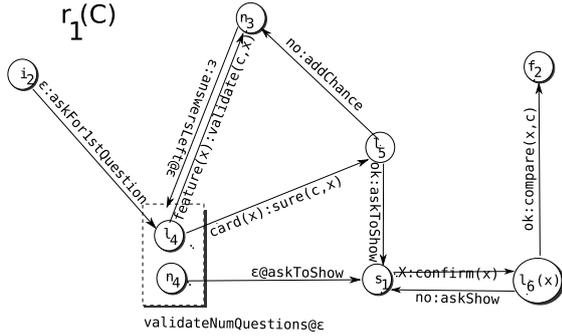


Figure 6: Example of a dialogue model for guessing a card

Factor	Percentage
TTS Performance	90 %
ASR Performance	50 %
Task ease	50 %
Interaction pace	80 %
User expertise	50 %
System response	70 %
Expected behavior	60 %
Future use	90 %

Cuadro 2: Percentage of the “yes” answers to the user-satisfaction questionnaire.

questionnaire from the PARADISE framework (Walker et al., 1997). For this, ten children played the game. All the children finished the game, and four of them guessed the right card. In average, there were 33,64 user turns. The visual system was able to identify the card at a rate of 1,18 tries per card. Table 2 summarizes the results obtained for the user satisfaction questionnaire. The game by itself is hard for the children, as can be seen in the *task ease* and *user expertise* factors. The children usually hesitate about what to ask, even though they are presented with suggestions by the system. The quality of ASR system needs to be improved considerably; in particular, the interpretation of the names and children’s ages has proven difficult. However, despite these shortcomings, the majority of children would like to play with the system again.

7. Implementation

The dialogue manager is implemented in *Prolog*. The modality specific interpreter and recognition modules are defined as independent processes, implemented with different programming languages and environments. For the control and communication between pro-

cesses we use the Open Agent Architecture framework (Cheyer y Martin, 2001). For the ASR system we use the Sphinx3 system (Huerta, Chen, y Stern, 1999). In particular for the system presented here, we developed a speech recognizer for children which are our main users. For this we collected the Corpus DIMEx100 Children. This is a corpus based on our previous work with the Corpus DIMEx100 Adults (Pineda et al., 2009). For this corpus, 100 children were-recorded the same 5,000 sentences of the adults version. With this setting we were able to get a 47,5 % word error rate (WER) with a basic language model based on the sentences of the corpus. This performance is comparable with the 48,3 % WER we obtain with the adult version of the corpus which has been further validated (Pineda et al., 2009).

For visual perception we use feature extraction and matching based on the Speeded-Up Robust Features (*SURF*) algorithm (Bay et al., 2008). This algorithm consists of three main steps: i) detection of interest points, ii) description of interest points, and, iii) object matching. Detection of interest points is based on the determinant of Hessian matrix (approximated by simple weighted box filters) to detect extrema pixels (i.e., pixels with darkest or lightest values) across a scale-space representation of the image. This representation is useful to achieve size invariance. Description of each interest point is composed by sums of 2D Haar wavelets responses to reflect intensity changes of square patches around the interest point. Integral images (Viola y Jones, 2004) are used to speed-up convolution. Object matching is carried out by nearest neighbor search and the trace of the Hessian matrix to distinguish between bright interest points on dark backgrounds and the inverse setting. Although in this system we are considering card identification only, this ideas can be extended to include different tasks of visual analysis. For example, in (Aviles et al., 2010) we have used this architecture to identify posters and also posters’ regions chosen by users through pointing gestures within the context of a tour-guide robot. SURF implementation is based on OpenCV (Bradski y Kaehler, 2008) with a naive nearest neighbor search.

8. Conclusions

In this paper we have presented a multimodal application with spoken language and vision developed with the framework of dialogue models specification and interpretation that we have developed over the years. The present system shows that this methodology and programming environment is mature enough to build real applications for the general public in a museum and similar kind of environments in a relatively short amount of time. Our methodology is focused on practical dialogues for task oriented applications that can be characterized through generic conversational protocols that can be found through analysis. The notion of global and specific context permits to interpret dialogue or speech acts in a simple way, without extensive syntactic and semantic analysis, as the context constraints very heavily the possible interpretations. The expressive power of F-RTN and the specification of abstract expectation and actions permit to model the conversational domain through simple protocols, that nevertheless generate rich and diverse conversational behavior. The present application has been evaluated in a preliminary way, and current results suggest that the quality of ASR system needs to improve considerably for the construction of robust applications. Although the present system is operational, and most children are able to complete the game, and are willing to play it again, showing a reasonable degree of user satisfaction, communication failures are still quite high, and there is a considerable conversational effort expended on recovery dialogues. Nevertheless we are confident that the present methodology has good potential for future applications.

Bibliografía

- Aviles, Hector, Ivan Meza, Wendy Aguilar, y Luis Pineda. 2010. Integrating Pointing Gestures into a Spanish-spoken Dialog System for Conversational Service Robots. To appear.
- Bay, Herbert, Andreas Ess, Tinne Tuytelaars, y Luc Van Gool. 2008. SURF: Speeded-Up Robust Features. *Computer Vision and Image Understanding*, 110(3):346–359.
- Bradski, G. y A. Kaehler. 2008. *Learning OpenCV: Computer Vision with the OpenCV Library*. ORilley.
- Cheyner, Adam y David Martin. 2001. The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1):143–148, March. OAA.
- Clark, H.H. y E.F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.
- Huerta, J. M., S. J. Chen, y R. M. Stern. 1999. The 1998 carnegie mellon university sphinx-3 spanish broadcast news transcription system. En *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*.
- Mann, W. C. y S. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281.
- Pineda, L., V. Estrada, S. Coria, y J. Allen. 2007. The obligations and common ground structure of practical dialogues. *Revista Iberoamericana de Inteligencia Artificial*, 11(36):9–17.
- Pineda, Luis, H. Castellanos, J. Cuétara, L. Galescu an J. Juárez, J. Llisterri, P. Pérez, y L. Villase nor. 2009. The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation*.
- Viola, Paul A. y Michael J. Jones. 2004. Robust Real-time Object Detection. *International Journal of Computer Vision*, 57(2):137–154.
- Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, Ace A. Kamm, y Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. páginas 271–280.