Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 113-120 recibido 17-01-11 revisado 11-02-11 aceptado 24-02-11

Evaluation of a Dialogue System for Children based on an Interaction-Oriented Cognitive Architecture*

Evaluación de un sistema de diálogo para niños basado en una Arquitectura Cognitiva Orientada a la Interacción

Esther Venegas-Briones, Ivan Meza-Ruiz, Luis A. Pineda
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
esthervenegas@hotmail.com, ivanvladimir@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx

Resumen: En este artículo presentamos la evaluación del sistema de diálogo "Adivina la Carta" con usuarios finales. Este es un sistema multimodal que se exhibe de manera permanente en el Museo de Ciencias Universum donde sus principales usuarios son niños. Presentamos un resumen de la metodología empleada para la evaluación con resultados objetivos y subjetivos.
Palabras clave: Sistema de diálogo, Sistemas multimodales, Evaluación

Abstract: This paper presents a full evaluation of the "Guess the card" system with final users. This is a multimodal dialogue system in a permanent stand in the Universum Science Museum whose main users are children. We summarise the applied evaluation methodology with the corresponding objective and subjective results.
Keywords: Dialogue system, dialogue manager, Multimodal Speech and Vision Systems

1 Introduction

In this paper we present a full evaluation of the "Guess the card" system (Meza et al., 2010a; Meza et al., 2010b). This is a Spanish spoken dialogue system with multimodal input and output. There are two aspects to highlight about the nature that make the evaluation challenging. First, the main users of the system are children between ages 10 and 16. Second, the dialogue system is based on an Interaction-Oriented Cognitive Architecture (Pineda, Meza, and Salinas, 2010). This architecture allows the development of spoken dialogue systems but it also handles multimodal input and output which has to be considered during the evaluation.

The "Guess the card" system handles spoken language interaction but it also includes the interpretation of images through computer vision and the display of pictures and animations to support linguistic behavior. The system stands in a permanent module at the Universum-UNAM Science Museum of the National Autonomous University of Mexico (UNAM). The system plays a game with the user where the user has the goal of guessing a card chosen that has been chosen randomly by the system; for this he or she is allowed to ask a number of questions about the features of the images on the cards in order to gather enough information about the chosen card. A more detailed explanation of the task and an example is presented in section 2.

The cognitive architecture includes a dialogue model with its interpreter program to represent the interaction protocols. For instance, there is a dialogue model for asking the name of the user, and a another one to handle the questions. Dialogue models capture the structure of the task and specify these protocols as conversational situations that relate potential expectations with the actions that the system needs to perform once one expectation is met. The dialogue manager coordinates the elements of the architecture by traversing dialogue models as required by the interaction. In this model intentions can be expressed by the user through spoken language and other modalities, even by events in the world. For instance, in the present implementation the user discloses the card by showing it to the system's eye

* We thank the enthusiastic support of the members of the Golem group at IIMAS and at the Science Museum Universum-UNAM. We also gratefully thank the support of grants CONACYT 81965 and PAPPIT-UNAM IN-115710.

ISSN 1135-5948

(a camera attached to it). The architecture has been used to build a number of different applications (Avilés et al., 2010b; Avilés et al., 2010a; Rascón, Avilés, and Pineda, 2010). The elements of the architecture and the modeling of the task are presented in section 3.

We based our evaluation on the PARADISE framework (Walker et al., 1997). This is a widely used methodology for evaluating dialogue systems that considers objective and subjective aspects of dialogues. This methodology was used to evaluate quantitative and qualitative elements from real user data. Some objective metrics were directly measured from the logs of the system which were manually labelled. A more exhaustive analysis was done in order to calculate the task success. Finally, a survey was taken to measure the user satisfaction for the subjective aspects of the evaluation. The general methodology and some adaptations to our system and further considerations during the evaluation are presented in sections 4 and 5. The results and analysis of the evaluation are presented on section 6.

2 Task

The motivation of the user in an interaction is to find the card chosen by the system from a set of ten cards. For this, the user can ask up to four questions in spoken Spanish. The cards have astronomical motives (e.g., the sun, a telescope) and are located on a table in front of the user. Examples of questions are: *is the object red?*, *is it a planet?* and *does it have energy?* In a typical session the system introduces itself, asks for the name and age of the child. Next, if required, the system provides the instructions of the game. Then, there is the questions session properly. At the end of the interrogatory, the child is asked to place the card that he or she thinks is the right one in front of the camera; then, the system confirms whether the child has won the game, or reveals which card was the right one. During the game the system responses are rendered using synthesized language, sound effects and images displayed on the screen. Table 1 presents an actual conversation between the system and the user.

3 The system

The Interaction-Oriented Cognitive Architecture on which the "Guess the card" is based has three conceptual levels: *recognition-rendering*, *interpretation-specification* and *representation-inference*, in addition to the semantic and perceptual memory (Pineda, Meza, and Salinas, 2010). The architecture is shown in Figure 1. The recognition and interpretation modules correspond to the perceptual modalities. For the system we use speech and visual perceptual processes. The goal of these modules is to assign an interpretation to input spoken and visual messages. On the other hand, the specification and rendering modules correspond to a behavioral response. In this application we use speech and visual actions.

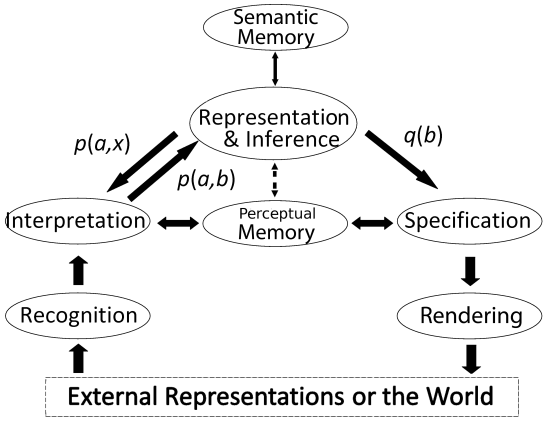

Figure 1: HCI dialogue system architecture

Recognition is a bottom-up process in the recognition-rendering layer that translates the input information into an image or pattern in a predefined coding format. No meaning or interpretation is assigned to the product at this recognition level. We refer to this kind of patterns as "uninterpreted images". For instance the "Guess the card" system uses an Automatic Speech Recognizer (ASR) module to translate speech into text and the computing vision algorithm SURF to codify the cards in terms of a set of salient points invariant to translation and rotation. The interpretation is part of the interpretation-specification layer and it is a heterarquic process that given the *expectations* of the situation (top-down) and an uninterpreted image (bottom-up) assigns an interpretation to the input message. For instance, in the present application, each expected intention has an

Turn	Spk	DM	Spanish Utterance/`Recognition hypothesis`/ *English Translation*	Inter.	Visual actions
1	SYS	M	Hola a todos me llamo Golem y éste es el juego de adivina la carta, tenemos jugador, ¿Cúal es tu nombre?/*Hello everyone, my name is Golem and this is the "guess the card" game, we have a player, what is your name?*		Welcome image
2	USR	n	Sofía/`sofia`/*Sofia*	$notU$	
3	SYS	n	Hola Sofía, me da mucho gusto jugar contigo/*Hello Sofia, I'm very glad to play with you*		
4	SYS	a	¿Cúal es tu edad?/*What is your age?*		
5	USR	a	Trece/`trece`/*Thirteen*	$age(13)$	
6	SYS	a	¿Seguro que tienes trece?/*Are you sure you are thirteen?*		
7	USR	a	Sí/`si`/Yes	ok	
8	SYS	a	Muy bien, podemos jugar/*Very good, we can play*		
9	SYS	i	¿Quieres que te explique las reglas del juego?/*Do you want me to explain the rules of the game?*		
10	USR	i	Sí/`si`/*Yes*	ok	
11	SYS	i	Presta atención ... Después de la ronda de preguntas compararemos mi carta y la que tú crees que elegí. Veremos si adivinaste. ¿Entendiste las instrucciones?/*Pay attention ... After that we will compare my card with the one you think I have chosen. We will see if you have guessed. Did you understand the instructions?*		A slideshow with examples
12	USR	i	Sí/`si`/*Yes*	ok	
13	SYS	i	Okay, fíjate muy bien en las cartas mientras escojo una/*Okay, look at the cards while I choose one.*		
14	SYS	I	Empecemos el juego, por favor hazme tu primera pregunta/*Let's start the game, please make your first question*		
15	USR	I	¿Es rojo?/`es rojo`/Is it red?	$feat(red)$	
16	SYS	I	/*No, en la carta que veo no hay nada rojo. Vamos, hazme la siguiente pregunta*/No, in the card I see there is nothing red. Lets go, make your next questions		
17	USR	I	¿Es azul?/`es azul`/Is it blue?	$feat(blue)$	
18	SYS	I	*Excelente, en la carta puedo ver una gran porción de color azul. Vamos, hazme la siguiente pregunta* /Exelent, I can see a big section of blue in the card. Lets go, make your next questions		
19	USR	I	¿Hay vida?/`hay vida`/Is there life?	$feat(life)$	
20	SYS	I	/ *En el espacio hay muy poca vida, pero este es uno de los lugares donde encontrarás varias formas de vida. Sólo te queda esta última pregunta, así es que piensala muy bien.*/In the space there is little life, but this is one place where you would find several forms of life. This is your last question, so think it very well.		
21	USR	I	¿Tiene verde?/`tiene verde`/Does it has green?	$feat(green)$	
22	SYS	I	*Muy bien, en mi carta hay algunas porcines que son de color verde. Ha finalizado tu ronda de preguntas.*/Very good, in my card there are some portions which are green. Your chances of asking are over.		
23	SYS	V	Por favor muéstrame la carta/*Please show me the card*		
24	USR	V	**User shows card with the earth on it to the camera**	$card(earth)$	
25	SYS	V	No tengo muy buena vista. Es ésta la carta que me mostraste/*I don't have a good sight, is this the card that you showed to me?*		Image of the recognized card
26	USR	V	Sí/`si`/Yes	ok	
27	SYS	V	Muy bien, has adivinado que la carta que escogí fue precisamente la de la tierra, '!bien hecho!/*Very good. You have guessed, the card that I have chosen was precisely the one of the Earth. Well Done!*		Image of the Earth
28	SYS	F	¡Oh! Pero antes de irte por favor mira a la cámara. Gracias por jugar conmigo Sofía, regresa pronto/*Oh! But before you leave look at the camera. Thanks for playing with me Sofía come back soon*		Camera takes a picture and shows it

Table 1: Example of interaction between the "guess the card" system and a user (Spk, Speaker; DM, Dialogue model; Inter., Interpretation).

associated regular expression, which codifies a large number of forms that can be used to express such intention. So, such regular expressions are applied to the text recovered by the ASR system, and whenever such a match holds, the concept associated to the regular expression is assigned. Visual concepts are recovered by a similar process in which the SURF representation of the external image is matched with the SURF representations associated to the visual concepts that are available to the system. These perceptual associations are stored beforehand in the system's perceptual memory. On the other hand, the propositional knowledge about the properties of the cards is stored in the system's semantic memory.

Output actions associated to expectations are specified through Multimodal Rhetorical Structures (MRS). This specification is part of the interpretation-specification layer in the architecture. Once the specification is done, the specified actions are dispatched to the render mechanisms which belongs to the recognition-render layer of the architecture. The lists of basic actions are defined loosely the lines of the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988).

Finally, the representation-inference layer has the purpose of identifying the adequate action during interpretation. For this purpose, this layer uses protocols of the interaction for representing the task. The dialogue manager is is the module in charge of this task and it uses dialogue models to specify the protocols. These specify a sequence of 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*. Dialogue models are represented through recursive transition networks (RTN) augmented with functions standing for expectations, actions and next situations (F-RTN) (Pineda, 2008). For further detail on the dialogue models specification consult (Pineda, Meza, and Salinas, 2010).

Although one could build a dialogue model for a task, the current framework allows to divide such dialogue into sub-dialogues which will correspond to the necessary sub-taks for a successful interaction. For the task of the present application, we specified six sub-dialogues which capture the structure of the task: name recognition (n), age verification (a), instructions (i), interrogation (I), visualization (V) and farewell (F). In order to capture this structure we defined eight dialogue models. One for each of these sub-dialogues plus the greeting (G) and the main (M) dialogue models. The main dialogue model coordinates the greeting (G), the interrogatory (I), the visualization (V) and the farewell (F) sub-dialogues. The greeting consists of a greeting by the system plus the name recognition, age verification and instructions sub-dialogues, and the choosing of card by the system. Table 1 presents a full dialogue between the system and a child using this structure. The table illustrates the turns and the dialogue model to which they belong, the utterances produced by the user and the corresponding recognition hypotheses produced by the ASR system, the interpretation hypotheses, the visual interpretations and the display actions performed by the system.

4 The evaluation

The evaluation is based on the PARADISE framework which is widely used for dialogue system evaluation (Walker et al., 1997). The adequacy of PARADISE for multimodal dialogue systems has been challenged (Beringer et al., 2002). Mainly, it has been pointed out that the framework has a problem when multiple modules provide interpretations for the same interaction *Which one to select for the evaluation?* However, this situation does not arise in the cognitive architecture since dialogue models explicitly specify the expectations that should match the interpretation, otherwise it means there was an error in the interaction and a recuperation dialogue has to be triggered.

The evaluation of the system consisted of three aspects:

- *User satisfaction*: Measures the perception of the performance of the system by the users.
- *Task success*: Measures how well the agent and the user achieve the interchange of information at the end of a dialogue or sub-dialogue. This is different from just considering a binary value about the completion of the task. We followed PARADISE and we used the *kappa* (κ) metric for this aspect.

- *Efficiency and quantitative metrics*: These metrics measure the efficacy and performance of the system in two levels. At the interaction level, for instance, *user's turns*; and at the level of single modules, for instance, the *Word Error Rate* for the Speech Recogniser.

The PARADISE framework proposes to apply a survey for measuring the user satisfaction. This survey measures seven topics of the system: TTS Performance, ASR Performance, Task ease, Interaction pace, User expertise, System response, Expected behavior and Future use. The survey is applied as a questionnaire with a question for each one of the topics. For the answers in the questionnaire we used a Likert scale with four options, two positive and two negative. The scale weights are: 100%, 66%, 33% and 0%, these have the effect of a high reward for the first positive answer (i.e., 100%), but also a high punishment for a last negative answer (i.e., 0%). For some questions, there was a clarification question about the reason why the user answered as he or she did. We added a question about the state of mind of the children since the kids commented in previous preliminary evaluations that they had been nervous during the interaction with the system. The questionnaire applied during the survey was explicitly designed for children. Appendix A presents the questionnaire and its translation.

In the context of dialogue models, the *kappa* metric is used to measure the agreement of the interchange of information between system and user. In particular for the cognitive architecture, the *kappa* metric will measure the agreement of the interpretations invoked at each situation. The intuition is that every time the system reaches the same situation and the user has the same intention, the interpretation processes have to interpret the same intention-meaning. The *kappa* metric will capture that agreement between the different interpretations processes, so that when all of them agree $\kappa = 1$, but if they agree by chance $\kappa = 0$. The elements of the confusion matrix used to calculate the *kappa* metric are directly defined by the expectations of the dialogue models. Thus there was no need to identify the modalities of such interpretations.

Finally, the qualitative metrics have the purpose of measuring the performance of the system. At the level of the interaction, it's possible to collect a wide range of statistics, and with the passing of time and as new dialogue systems get evaluated the repertoire of metrics have grown. For the evaluation we measure statistics of the user utterances, system utterances, the visual perception and the duration of the interactions. On the other hand, the qualitative metrics focus on the performance of specific modules as well. For this we measure the performance of the different modules, mainly the ones in charge of the recognition and interpretation. We compared with what a human would have recognised and/or interpreted.

5 The setting

The evaluation was carried out with real final users. We asked thirty children aged between 10 and 16 year-old which were visiting the museum to take part in the evaluation by playing a game with the system. The history of the interactions were recorded by the system. A video recording was also taken for further analysis, specially for the *kappa* metric. None of the children had previous experience with the system. The group of children were balanced by gender. There were 14 children aged from 10 to 13, the rest were 14 or older. All were native speakers of Spanish. After one interaction with the system, the children were interviewed to apply the questionnaire by one member of the project.

The implementation of the system used during the evaluation is as follows. The system uses Open Agent Architecture framework (OAA) for communication between the modules of cognitive architecture (Cheyer and Martin, 2001). The dialogue manager is implemented in *Prolog*. The ASR system is Sphinx3 (Huerta, Chen, and Stern, 1999). We also developed a speech recognizer for children. For this, we collected the Corpus DIMEx100 Children. This is a corpus based on the Corpus DIMEx100 Adults (Pineda et al., 2009). The speech interpreter uses a word spotting technique based on regular expressions. For visual perception we used feature extraction and matching based on the Speeded-Up Robust Features (*SURF*) algorithm (Bay et al., 2008). The SURF implementation is based on OpenCV (Bradski and Kaehler, 2008) with a naive nearest neighbor search.

Factor	Value
Participants	30
Finished the game	30
Won the game	2
Asked question one	30
Asked question two	30
Asked question three	27
Asked question four	27
Average questions per dialogue	3.77

Table 2: Statistics about the games played by the children.

Factor	Percentage
TTS Performance	93%
ASR Performance	47%
User expertise	60%
Task ease	66.7%
Interaction pace	92%
System response	59%
Expected behavior	65%
Future use	83%

Table 3: Percentage of the positive answers to the user-satisfaction questionnaire.

Subdialogue	κ
Name	0.61
Age	0.91
1st question	0.72
2nd question	0.63
3rd question	0.64
4th question	0.33
Whole dialogue	0.74

Table 4: κ metrics for sub-dialogue and the whole dialogue

6 Results

Table 2 presents some statistics of the interactions played by the children. These results show that the game was a hard task since 9 out of 30 children guessed the chosen card. However, all of the children were able to reach the end of the game, and 10 of them tried to guess the card before their fourth chance.

The results of the survey are presented in Table 3. It was agreeable to see than the children seemed to have no problem assessing the performance of the system. A recurrent problem in the system is the perception of the speech recognition performance. Children perceive that the system does not understand them (*ASR Performance*). These results reinforce that the game is perceived as a difficult task (*task ease*). We believe the main cause is because children do not know what question to pose to the system (*user expertise*). Children also think the system response was not fast enough (*System response*), actually 85% of children think the system is slow. On the other hand, most of the children will play again with the system (*future use*), and they even think the system behaved better than they expected, 27 children gave a positive answer (*Expected behavior*). We also found that 33% of the children felt nervous while playing with the system.

The results for the task success can be measured for the whole interaction or for the sub-dialogues. We focused on the sub-dialogue for asking the name (n), the age (a), and the sub-dialogue for asking a question (I). For the latter, for each chance the children had we consider it to be a different realisation since he or she could have guessed before his or her four chances. The results for the task success are shown in 4. They measure the agreement of the interpreting process at each of the subtasks. From the results we can see that understanding a name is harder than an age. This makes sense since the number of possibilities for the names are much lager than the ages. However, something that we did not expect was that as the posed questions gets harder for the system. We believe this has its origin on the phenomena that the questions get more complex after some tries. In particular, during the evaluation we observed that at first children repeated the examples provided by the system given in the instructions stage, but later they explored more options. The last *kappa* result of the tables corresponds to metric for the whole dialogue. For this case, the interaction on which the user shows the card to the system was included among the other linguistic interactions. For the cognitive architecture the interactions once they get interpreted are not different, in this case a visual act can be compared with a speech act. These results are quite motivating. Despite the problems on the speech recogniser, there is a good interchange of information between the user and the system.

For the qualitative metrics we recorded the interaction logs and audio said by the children. With this information we were able to measure durations of different events and

Average duration dialogue	$4m\ 54s$
Maximum duration dialogue	$6m\ 53s$
Minimum duration dialogue	$3m\ 32s$
System's utterances	1257
User's utterances	501
Average user's utterances	16.7
Average duration of user's utterance	$7s$
Maximum user's utterances	27
Minimum user's utterances	10

Table 5: Utterance and duration metrics for the whole interaction

Speech recogniser	
Word error rate	59%
Total words recognized	1,001
Total types recognized	264
Most common token	"SI"
Average words per utterance	2
Errors no sound	0.06%
Language interpreter	
Accuracy	57%
Accuracy for names	45%
Accuracy for ages	90%
Accuracy for questions	51%
No interpretation found	27%
Visual interpreter	
Accuracy	48%
Average of seen tries per dialogue	1.7

Table 6: Metrics for recogniser and interpreters

the utterances by the system and users. Table 5 shows the main metrics from these recordings. Additionally, these records were extended with manual transcription to evaluate specific modules of the system. In particular, the speech recogniser, the language interpreter and the visual interpreter. These results are presented in 6. We confirm that the speech recogniser performance is quite poor, however we found that the language interpreter had a similar performance. On the other hand, when the interpreter does not find an interpretation for the recognised utterance, it triggers a recovery strategy and this happened 27% of the times. So in case of confusion, the system will ask for some clarification. This allowed the system to reach the end of the interaction for all users, even though it does not understand what has been said 40% of the times.

7 Conclusions

In this paper we have presented a full evaluation of a dialogue system based on a Cognitive Architecture oriented to interaction. The "Guess the card" system is a multimodal system which plays a game with children aged between 10 and 16 on which the user tries to guess a card by asking questions to the system. Once the child has guessed the card or his/her chances are finished he/she will show the guessed card to the system's eye. The evaluation was based on the PARADISE framework. Previously, the adequacy for multimodal dialogue system has been put into doubt. However, the evaluation for our system was possible given that the cognitive architecture abstracts the task from the interpretation processes.

We found that the main problem of the system is the speech recognition performance. The qualitative metrics of its performance and the perception of the children about it are low. Currently, we are performing some work to improve the acoustic and language models of such module. The task by itself is considered a difficult task for the children. The main problem for the children seems to be to come up with a question. We see this as an opportunity window and we plan to study the behaviour of the children using different strategies to help them with the questions. Despite all these problems the system was able to finish the interactions and most of the children were impressed by it. We believe this is in part possible because of the cognitive architecture which allows us to capture the task structure directly into dialogue models, and separates interactions from the interpretations process by means of expectations.

We are also exploring the relations between the objective measurements and the user satisfaction as proposed by (Callejas and López-Cózar, 2008). On preliminary results, we have found the tasks success has the highest correlation with the user satisfaction.

A Appendix 1: Questionnaire

TTS Performance *¿Entendiste lo que el sistema decía?*/Did you understand what the system said?
ASR Performance *¿El sistema entendió lo que tú decias?*/Did the system understood what you said?
User expertise *¿Sabías qué preguntale al*

sistema?/Did you know what to ask to the system?
Task ease *¿Tuviste tiempo suficiente para pensar las preguntas?*/Did you have enough time to think the questions?
Interaction pace*¿Te gustó el ritmo del juego?*/Did you like the pace of the game?
System response *¿Te gustó la velocidad de respuesta del sistema?*/Did you like the speed of the system's answers?
Expected behavior *¿El sistema funcionó como te lo imaginabas?*/Did the system work as you imagined?
Future use *¿Volverías a jugar con el sistema?*/Would you play again with the system?
State of mind *¿Cómo te sentiste en la interacción con el sistema?*/How did you feel in the interacion with the system?

References

Avilés, H., M. Alvarado, E. Venegas, C. Rascón, I. Meza, and L. Pineda. 2010a. Development of a tour-guide robot using dialogue models and a cognitive architecture. In *Proceedings IBERAMIA 2010. LNCS (LNAI)*, pages 512–521.

Avilés, H., I. Meza, W. Aguilar, and L. Pineda. 2010b. Integrating Pointing Gestures into a Spanish-spoken Dialog System for Conversational Service Robots. In *Second International Conference on Agents and Artificial Intelligence (ICAART)*.

Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool. 2008. SURF: Speeded-Up Robust Features. *Computer Vision and Image Understanding*, 110(3):346–359.

Beringer, N., L. Katerina, V. Penide-López, and U. Turk. 2002. End-to-end evaluation of multimodal dialogue systems – can we transfer established methods? In *Proc. 3rd Language Resources and Evaluation Conference*, pages 558–563.

Bradski, G. and A. Kaehler. 2008. *Learning OpenCV: Computer Vision with the OpenCV Library*. ORilley.

Callejas, Z. and R. López-Cózar. 2008. Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication*, pages 646–665.

Cheyer, A. and D. Martin. 2001. The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1):143–148, March. OAA.

Huerta, J. M., S. J. Chen, and R. M. Stern. 1999. The 1998 carnegie mellon university sphinx-3 spanish broadcast news transcription system. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*.

Mann, W. C. and S. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, pages 243–281.

Meza, I., L. Salinas, H. Avilés, and L. Pineda. 2010a. A multimodal dialogue system for playing the game "guess the card". *Procesamiento de Lenguaje Natural*, pages 131–138.

Meza, I., L. Salinas, E. Venegas, H. Castellanos, A. Chavarria, and L. Pineda. 2010b. Specification and Evaluation of a Spanish Conversational System Using Dialogue Models. In *Proceedings IBERAMIA 2010. LNCS (LNAI).*, pages 346–355.

Pineda, L., I. Meza, and L. Salinas. 2010. Dialogue Model Specification and Interpretation for Intelligent Multimodal HCI. In *Proceedings IBERAMIA 2010. LNCS (LNAI)*, pages 20–29.

Pineda, L. A. 2008. Specification and interpretation of multimodal dialogue models for human-robot interaction. In *Artificial Intelligence for Humans: Service Robots and Social Modeling*, pages 33–50.

Pineda, Luis, H. Castellanos, J. Cuétara, L. Galescu an J. Juárez, J. Llisterri, P. Pérez, and L. Villase nor. 2009. The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation*.

Rascón, C., H. Avilés, and L. Pineda. 2010. Robotic orientation towards speaker for human-robot interaction. In *Proceedings of Iberamia 2010. Ed. Kuri, A. LNAI (2010)*, pages 10–19.

Walker, M. A., D. J. Litman, C. A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. pages 271–280.