



Universidad Nacional Autónoma de México
Facultad de Filosofía y Letras
Licenciatura en Lengua y Literaturas Hispánicas

Evaluación del sistema conversacional con español hablado “Golem – Adivina la Carta” en el museo Universum.



Tesis que para obtener el título de
Licenciada en Lengua y Literaturas Hispánicas
presenta

Esther Venegas Briones

Asesor: Dr. Luis Alberto Pineda Cortés

SWAYED

México, Ciudad Universitaria, 2011.



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

A mi padres, Humberto y Esther.

A mis hermanos Humberto, Marcela y Juan.

A mis sobrinos Eduardo, Humberto, Emiliano, Ilana y Elías.

Agradezco al SUAyED la posibilidad de seguir estudiando y ocho semestres maravillosos.

Agradezco al Dr. Luis A. Pineda todo el apoyo, la oportunidad de participar en el grupo y en todo lo que se ha construido en este tiempo.

Agradezco al Dr. Ivan Meza la generosidad de su tiempo, la paciencia, el ánimo y la energía del trabajo de todos los días.

Agradezco el apoyo de los proyectos CONACYT 81965 y PAPPIT-UNAM IN-115710 por proporcionar los recursos necesarios para la realización de esta tesis.

Í N D I C E

Introducción	5
Antecedentes	5
Objetivo	8
Estructura de la tesis	10
Capítulo 1. Construcción de sistemas de diálogo multimodales	12
1.1. Sistemas de diálogo y la arquitectura cognitiva	13
1.2. Modelos de diálogo	16
1.3. Estructura de la conversación	21
Capítulo 2. Golem: Adivina la Carta	27
2.1. Funcionamiento de la aplicación	28
2.2. Descripción de la tarea y objetivo del juego	32
2.3. Estructura de la conversación	37
Capítulo 3. Metodología de evaluación de sistemas de diálogo.	42
3.1. Técnicas y herramientas de evaluación de sistemas de diálogo	44
3.2. El coeficiente <i>kappa</i>	48
3.3. Paradise como metodología integral de evaluación	54
3.3.1 Matriz de atributos	56
3.3.2. Éxito en la tarea	57
3.3.3. Medición	59
3.3.4. La función de desempeño	61
Capítulo 4. Evaluación del juego “Adivina la carta”	64
4.1. Diseño del ejercicio de evaluación.	65
4.2. Resultados de la evaluación.	67
4.2.1. Descripción de la tarea	67
4.2.2. Intercambio de información	68
4.2.3. Estimación de la función de desempeño	73
4.2.4. Costos del diálogo: medidas de eficiencia y efectividad	76
4.2.5. Estimación de la función de desempeño	78
Conclusiones	84
Bibliografía	91
Apéndice	95

INTRODUCCIÓN

Antecedentes

En los últimos años hemos sido testigos de un incremento importante en el desarrollo de programas informáticos que permiten el procesamiento del lenguaje. La integración de conocimientos lingüísticos para el desarrollo de aplicaciones computacionales para el uso cotidiano, es cada vez mayor. Actualmente, a través del teléfono es posible acceder a una serie de servicios, desde un directorio telefónico hasta la reservación de boletos y la obtención de información sobre horarios de transporte.

Estos avances son resultado de décadas de investigación. De acuerdo con John R. Searle (Searle 2006, 1), uno de los avances más importantes en la filosofía en los últimos cien o ciento veinticinco años ha ocurrido en la filosofía del lenguaje; estos avances han ido de la mano del desarrollo de la lógica matemática iniciados por Friedrich Ludwig Gottlob Frege, inventor tanto de la filosofía del lenguaje como de la lógica moderna, y seguido entre otros por Russell, Ludwig Wittgenstein, Willard van Oman Quine, John Austin y sus sucesores.

El estudio del lenguaje ha sido objeto de análisis no sólo de lingüistas y filósofos; la psicología, las neurociencias, las matemáticas y la computación (Meza Ruiz 2002, 7) también han dedicado sus esfuerzos a conocer el procesamiento del lenguaje y la generación del habla. La contribución de todas éstas, y de las ramas científicas que han surgido en el camino, ha sido indispensable para avanzar en el objetivo de reconocer y reproducir el lenguaje de una forma semejante a la que utiliza el ser humano de forma automática.

Particularmente, la relación entre la informática y el lenguaje se estrechó tanto que dio origen, a mediados del siglo XX, al surgimiento de la lingüística computacional, como el área de la computación dedicada a estudiar el lenguaje desde el punto de vista de la información necesaria para su fiel representación e interpretación. Actualmente la lingüística computacional tiene entre sus objetivos reproducir el lenguaje de una manera semejante a como es utilizada por el ser humano, entender el significado y el procesamiento de textos, producir sistemas de síntesis y reconocimiento del habla, procesar y representar el significado del texto y facilitar la interacción oral de las personas con los sistemas informáticos.

Debe reconocerse que el lenguaje humano es un sistema de gran complejidad. Por una parte, la facultad lingüística humana se relaciona con la percepción, el pensamiento, la memoria, la conducta motora y la conciencia y por otra, las conversaciones se desarrollan en un contexto espacial y temporal en medio de deseos, creencias, conocimientos e intenciones, que envuelven los actos lingüísticos que se observan en la conversación.

En este sentido, si bien en el corto plazo no existe la posibilidad de emular la totalidad de las competencias humanas en la conversación, sí ha sido posible construir sistemas que en un contexto determinado provean respuestas necesarias para el cumplimiento de tareas específicas y, aún con esta limitación, los investigadores han logrado avanzar y llevar el objetivo de la interacción de un intercambio de comandos a un intercambio de información cada vez más complejo.

De la mano con los avances en el desarrollo de estos sistemas van los avances en las herramientas de comunicación. La navegación entre páginas web se realiza hoy en día principalmente de forma visual, pero el teléfono se ha convertido en un dispositivo indispensable para muchos usuarios (Callejas and López-Cózar 2005) y la interacción oral es la que más comúnmente usamos para comunicarnos.

Por ello, si la conversación es la forma de comunicación más eficiente y flexible entre los seres humanos, la construcción de sistemas de diálogo constituye la alternativa más prometedora en la comunicación entre humanos y las computadoras y un reto importante en la comunidad científica. La maduración de esta tecnología cambiará radicalmente la manera como se interactúa con las máquinas y tendrá un gran impacto en varios ámbitos de la vida social y económica y aplicación en diversos ámbitos.

Conforme la tecnología de la lingüística computacional se ha ido incorporando al mercado, los mecanismos de evaluación han recibido un estímulo adicional y son ahora esenciales para el desarrollo y optimización; de igual manera es importante para los investigadores contar con evidencia que permita a la comunidad científica y al público entender y hacer juicios confiables con respecto al trabajo realizado.

Durante varios años, los investigadores han buscado permanentemente herramientas que les permitan sustentar de manera confiable los juicios y resultados de sus investigaciones y, aunque el tema de una metodología general de evaluación sigue siendo un reto importante, la valoraciones que se logran en cada ejercicio han sido importantes en los avances logrados, y siguen siendo importantes para obtener información suficiente y relevante sobre el funcionamiento general del servicio evaluado y conocer el desempeño de componentes específicos

Objetivo

Durante los últimos diez años, el Departamento de Ciencias de la Computación del Instituto de Investigación en Matemáticas Aplicadas y en Sistemas (IIMAS) de la Universidad Nacional Autónoma de México (UNAM) se ha dedicado al desarrollo de sistemas de conversacionales en español de México (Aguilar and Pineda 2009; Avilés, Meza, et al. 2010; Rascón, Avilés and Pineda 2010). Los trabajos de investigación han requerido de un trabajo interdisciplinario entre la ingeniería computacional y la lingüística para la construcción de sistemas sencillos pero robustos para el reconocimiento de voz, que han dado como resultado, entre otras, una aplicación que permite al sistema jugar: “Adivina la carta” con usuarios reales, principalmente niños. Este juego puede utilizarse en un módulo permanente ubicado en Universum, el museo de las ciencias de la UNAM (Meza, Salinas and Venegas, et al. 2010).

En este juego, el usuario tiene el objetivo de adivinar una carta, elegida de manera aleatoria por el sistema, entre diez cartas con motivos astronómicos que se localizan en el módulo. El usuario tiene hasta cuatro oportunidades para identificarla planteando preguntas sobre las características de los objetos representados. Una vez agotadas las oportunidades, el usuario debe mostrar la carta que cree que fue elegida por el sistema, el cual interpreta visualmente la carta, confirma a los ganadores y agradece a todos los usuarios su participación.

Este sistema se ha evaluado de forma periódica, ya que existe un interés específico en el DCC para obtener, a través de estos ejercicios, referentes concretos que permitan conocer y analizar de manera objetiva el funcionamiento de las aplicaciones desarrolladas, para planear y organizar mejoras, así como para analizar comparativamente los cambios realizados respecto a versiones anteriores o respecto a otras aplicaciones desarrolladas en el DCC. En general, las evaluaciones han consistido en la generación de métricas que permiten la valoración de los diferentes elementos que conforman y hacen posible el funcionamiento de estos sistemas.

Aunque a la fecha no existe una metodología para la evaluación de sistemas de diálogo que tenga una aceptación generalizada a nivel internacional, elegimos PARADISE, una metodología que se cita en múltiples artículos de investigación y que recoge o integra métodos y herramientas desarrollados como resultado de la experiencia de evaluación de varias áreas del procesamiento de lenguaje natural (Walker, et al. 1997).

El objetivo principal de este trabajo consiste en presentar los resultados de la evaluación del sistema conversacional con español hablado “Golem – Adivina la Carta” desarrollado en el DCC – IIMAS y que se encuentra funcionando en el museo Universum. Para ello presentamos las herramientas teóricas y empíricas que han sido utilizadas tanto para el desarrollo de la aplicación, así como para evaluar el adecuado funcionamiento del sistema. Se trata de la valoración de una aplicación que está en permanente exposición ante el público, con usuarios reales, fuera del laboratorio y en consecuencia, un ejercicio que no se simula y que agradece a los usuarios que acudieron y siguen acudiendo como visitantes al museo de las ciencias y que con su participación contribuyen a la mejora del sistema.

Existen algunas aportaciones que a nivel personal y en el grupo se realizaron a la misma metodología en la instrumentación del ejercicio, que pretenden contribuir al levantamiento de información y, en última instancia, facilitar en el futuro la evaluación de los sistemas que se desarrollan en el DCC.

Por lo pronto, los resultados de la evaluación han sido útiles para conocer el desempeño de una aplicación que se ubica en exhibición permanente y para dirigir y priorizar los esfuerzos de mejora.

Estructura de la tesis

Para el desarrollo de este trabajo y cumplir el objetivo señalado, se presentan cuatro capítulos además de un apartado de conclusiones finales para abordar los temas que se indican a continuación:

El capítulo 1 describe los elementos que conforman un sistema de diálogo. Incluye algunas definiciones que consideramos indispensables para entender el desarrollo de las aplicaciones dentro de un marco teórico y una visión general que guía o motiva los trabajos que se llevan a cabo en el DCC. Partimos describiendo la arquitectura general que da orden y sentido a los elementos que intervienen en la construcción de las aplicaciones y de los diálogos, continuamos con la descripción de los modelos de diálogo y por último abordamos la forma en que se estructura una conversación.

El segundo capítulo aborda de lleno la aplicación, motivo de evaluación en este estudio. Con base en los elementos descritos en el capítulo anterior, se describe el juego “Adivina la Carta, su objetivo y funcionamiento, así como el alcance de la tarea y la construcción del diálogo. Se ejemplifica una conversación a nivel general y se ejemplifican las situaciones de error o recuperación, que finalmente serán motivo de la evaluación.

El capítulo 3 está dedicado a la metodología de evaluación. El ejercicio que se lleva a cabo en este trabajo de tesis se basa en una metodología general, la cual se describe de forma detallada. Para dar un enfoque integral sobre el desempeño de una aplicación, esta metodología toma como base métodos y herramientas que se toman de la experiencia de evaluación de varias áreas del procesamiento de lenguaje natural, así como de las ciencias cognitivas. Se incluye una explicación de éstos a modo de antecedente, ya que se son elementos indispensables de la misma. Durante la exposición de la metodología mencionamos algunos ajustes que hicimos, si no específicamente en la metodología, sí en su aplicación, atendiendo las necesidades particulares de evaluación de “Adivina la Carta”.

En el capítulo 4 se describen las condiciones en las que se llevó a cabo el ejercicio de evaluación, el escenario y las condiciones de interacción. Se presenta un resumen de los resultados de la evaluación tal y como se recuperaron, como resultado de la aplicación de la metodología de evaluación. Aun cuando la metodología hace énfasis en una función única de evaluación, lo cierto es que los resultados han sido útiles, tal vez más en su individualidad que integrados en un resultado numérico único. También es cierto que resulta complicado intentar presentar todos los resultados en un espacio limitado, por ello, los cuadros completos de resultados se integran en un apéndice.

En el capítulo de conclusiones exponemos nuestros hallazgos con respecto a la interpretación de los resultados obtenidos, los beneficios que esperamos obtener referentes a la mejora del sistema Golem – Adivina la Carta, así como las posibilidades de aplicación en otras aplicaciones desarrolladas en el ámbito de los trabajos que se realizan en el DCC.

CAPÍTULO 1.

CONSTRUCCIÓN Y DESARROLLO DE SISTEMAS DE DIÁLOGO MULTIMODALES

Los sistemas de diálogo son sistemas computacionales desarrollados para interactuar a través del lenguaje con los seres humanos. El presente capítulo tiene como objetivo presentar los elementos que conforman los sistemas de diálogo, así como el marco conceptual de su desarrollo tal y como se lleva a cabo en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), con el fin de conocer y delimitar el objeto de estudio de este trabajo y de la evaluación. Para ello tomo como base el trabajo que se ha desarrollado en el contexto del proyecto Diálogos Inteligentes Multimodales en Español (DIME), desarrollados en el DCC del IIMAS, del cual surgió el sistema “Adivina la Carta” (Meza, Salinas and Venegas, et al. 2010).

Para una consulta referente a la construcción y desarrollo de sistemas de diálogo multimodal en el DCC del IIMAS pueden consultarse, entre otros los artículos siguientes: An Interaction Oriented Cognitive Architecture for Service Robots (Pineda 2010), De los sistemas conversacionales a los robots parlantes (Pineda 2011). The obligations and common ground structure of practical dialogues (Pineda, Estrada, et al. 2007)

La descripción de los sistemas de diálogo va unida a la descripción de la arquitectura cognitiva al amparo éstos se conciben para estructurar los procesos computacionales. Los modelos de diálogo constituyen abstracciones que permiten representar de manera genérica protocolos conversacionales. Por último se explica el enfoque que guía la forma en la que se conciben y construyen las conversaciones. Este enfoque concibe las conversaciones como transacciones construidas con aportaciones que realizan en turnos intercalados el agente y el usuario durante la conversación, las cuales deben estar “balanceadas” para que se dé la comunicación. Una vez hechas estas definiciones podremos proceder con un mayor sustento con la explicación del funcionamiento de “Adivina la Carta” en el capítulo siguiente.

1.1. Sistemas de diálogo y la arquitectura cognitiva

Los sistemas computacionales desarrollados para interactuar con seres humanos se denominan sistemas de diálogo. De acuerdo con James Allen (**Allen, et al. 2000**), la visión de lo que debe ser el alcance o la complejidad de la interacción ha ubicado a los investigadores en dos grupos; por un lado están aquellos para quienes la interacción implica una acción limitada, entendiendo como limitada la característica de aquellos sistemas en los que el lenguaje y su uso se conciben como escasos. Pone como ejemplo de este tipo de aplicaciones, aquellos sistemas utilizados en los conmutadores en donde el usuario debe presionar la tecla “1” o decir “uno”. En el extremo opuesto, ubica a aquellos investigadores cuya motivación en el desarrollo de sistemas de diálogo es imitar, en la medida de la posible, una conversación tal y como la mantienen los seres humanos, por lo que buscan expandir las opciones de los usuarios durante la interacción.

Las aplicaciones que se desarrollan en el DCC IIMAS, comparten los motivos del segundo grupo y, aunque la comunidad investigadora reconoce que en el corto plazo no hay posibilidades de emular todas las competencias humanas que intervienen para el desarrollo de una conversación, existe la convicción y la experiencia de que es posible

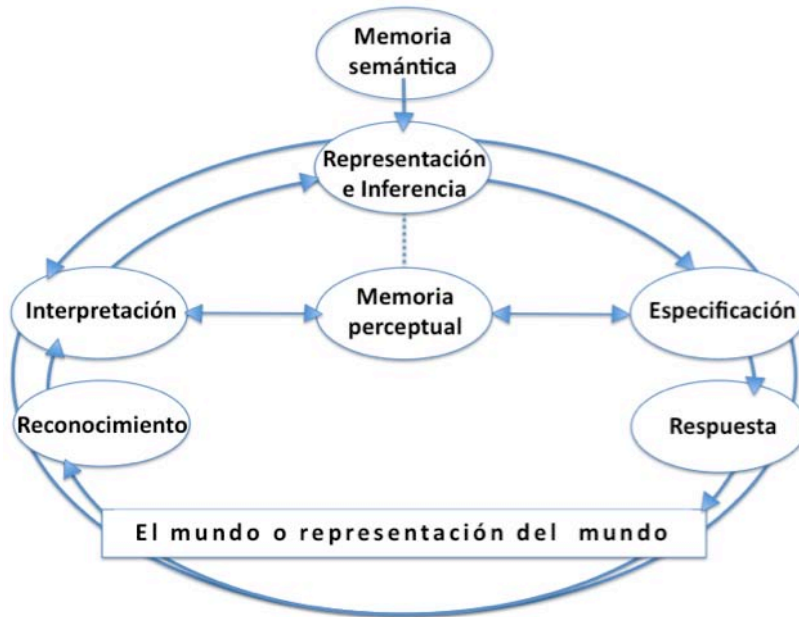
desarrollar sistemas que en un contexto determinado, además de solucionar tareas específicas, utilicen el lenguaje con una variedad de fines para reconocer las intenciones del usuario y que, por ejemplo, permitan a éste tomar la iniciativa de la conversación para acelerar la solución de una tarea sin dejar de asegurar la comunicación.

Cuando hablamos de conversaciones y aplicaciones cuyo fin es solucionar una tarea específica, como obtener información del usuario o solucionar un problema en un contexto específico, hablamos de diálogos prácticos. En los diálogos prácticos las competencias conversacionales requeridas son significativamente más simples de alcanzar, comparadas con la competencia humana general para la conversación. Las aplicaciones que se desarrollan en el DCC constituyen diálogos prácticos, es decir, se trata de aplicaciones desarrolladas para ejecutar una tarea determinada en un contexto previamente especificado utilizando un lenguaje con una variedad de fines. En los diálogos prácticos, la complejidad de la interpretación y el manejo del lenguaje son independientes de la tarea que se ejecuta.

Otra característica que distingue las aplicaciones que desarrolla el DCC del IIMAS, es que éstas están diseñadas con el objetivo de interactuar de manera activa con el mundo y con los seres humanos, lo que ha dado lugar a una arquitectura cognitiva que se presenta a continuación (Pineda 2008; Pineda, Meza and Salinas 2010).

Esta arquitectura comparte algunos aspectos con la arquitectura cognitiva llamada Beliefs, Desires and Intentions o BDI, creencias, deseos e intenciones, (Pineda 2011) y ha sido el resultado de una intensa actividad de investigación para combinar la percepción de los estímulos de otras personas y del mundo, el pensamiento y la acción o interacción con otras personas y con el mundo. Para fines de este trabajo esta arquitectura nos permite acotar las capacidades de los agentes y dar orden y sentido a los elementos que intervienen en la construcción y funcionamiento de las aplicaciones y de los diálogos, tal y como puede observarse en la gráfica 1.1. siguiente.

Gráfica 1.1 Arquitectura cognitiva



Fuente: (Pineda 2010)

En este diagrama están representados por una parte el mundo real, y por otra y con más detalle, los diversos módulos que en un sistema están asociados con la percepción, el pensamiento y la acción. En este sistema cada módulo tiene una función específica. Si se sigue el curso marcado por las flechas puede observarse que el diagrama conforma un círculo, ciclo que se recorre cada vez que hay una interacción entre el sistema y el mundo.

El reconocimiento y la interpretación corresponden a la percepción. El reconocimiento es el proceso mediante el cual el sistema capta, a través de algún dispositivo, la información del mundo y la traduce en una imagen o un código específicos, dependiendo del dispositivo utilizado. El producto del reconocimiento es un código o una imagen no interpretada; es decir, en el caso del lenguaje, el sistema de reconocimiento de voz produce texto en un formato particular, sin otorgar interpretación alguna. Existen distintos dispositivos de interacción, como pueden ser los dispositivos de navegación, visuales o sonoros; cada uno capta y genera códigos a procesarse a través de algoritmos específicos correspondientes a cada modalidad.

En el siguiente nivel de percepción encontramos el intérprete, encargado de traducir las imágenes o códigos captados por el reconocedor y de traducirlos en el marco de las expectativas y del contexto conocidos y delimitados en función de la tarea específica a realizar. El intérprete entrega imágenes interpretadas.

Los módulos que corresponden a la memoria perceptual y semántica toman estos nombres porque cumplen funciones semejantes a las que cada una de éstas tiene con la memoria humana. La memoria perceptual está formada por una serie de asociaciones entre la interpretación (significado) y las imágenes o códigos no interpretados que el intérprete utiliza al momento de realizar su función. La memoria semántica se refiere al conocimiento de las palabras, de su significado, de los conceptos y de sus relaciones, cuando tal conocimiento no está ligado al contexto espaciotemporal de su adquisición. Estos conjuntos de imágenes y significados se crean basados en las necesidades del contexto y la tarea específica a solucionar, y se cargan en el sistema con anticipación, utilizando los mismos dispositivos de reconocimiento con los que cuenta la aplicación. Aquí, la información se guarda en la memoria, tomando el orden de la serie de expectativas que espera se cumplan a lo largo de la conversación, de tal forma que la interpretación adecuada pueda ser recuperada en la situación particular en la que esta información deba ser interpretada. La memoria episódica almacena información de hechos, como nombres de personas o conceptos, no asociados a ninguna otra memoria en particular.

Los módulos de especificación y respuesta corresponden a la acción o interacción. La respuesta, al igual que el reconocimiento, está acotada por la serie de dispositivos que tenga para su uso la aplicación.

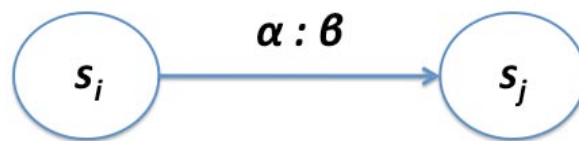
1.2. Modelos de diálogo

Una vez definidos los sistemas de diálogo y la relación que guardan con la arquitectura cognitiva que guía el desarrollo de aplicaciones en el DCC-IIMAS, procederemos a explicar la noción de los modelos de diálogo. Los modelos de diálogo constituyen

abstracciones que permiten esquematizar los protocolos conversacionales que se utilizan de manera recurrente. Estos esquemas permiten representar las situaciones, las intenciones, las expectativas y las acciones que ocurren en el transcurso de una conversación en aplicaciones desarrolladas para ejecutar una tarea determinada en un contexto previamente especificado.

Los modelos de diálogo se definen a partir de por lo menos dos situaciones; una situación inicial (s_i) y otra final (s_j). Gráficamente, cada situación se representa por un nodo y los nodos se relacionan a través de vínculos. Cada vínculo está formado por una expectativa a satisfacer (α) y por las acciones a realizar una vez satisfecha la expectativa (β). Una vez cumplida la expectativa y realizada la acción, el sistema pasa de la situación inicial s_i a la situación s_j , como se ilustra en el siguiente ejemplo en la gráfica 1.2.:

Gráfica 1.2. Representación gráfica de los modelos de diálogo



Fuente: (Pineda, Meza and Salinas 2010, 23)

Además de la situación inicial y la situación final, que no indican más que el inicio y fin del protocolo, los modelos cuentan con un número variado de situaciones intermedias que representan, cada una, parte de las acciones que se deben realizar para solucionar la tarea global.

Las expectativas y acciones pueden estar vacías o ser concretas. Por ejemplo, si el sistema despliega una imagen con dos botones para dos diferentes acciones cada vez que se inicia, no esperará ninguna acción del usuario; haga o no haga nada, el sistema desplegará la imagen con el saludo cada vez que se inicie y llegará a la siguiente situación, donde esperará que se cumpla la expectativa siguiente.

Las expectativas y las acciones, cuando son concretas, pueden adoptar cualquiera de las modalidades que hemos mencionado, es decir, dependerán de los dispositivos disponibles y podrán adoptar una acción perceptible al usuario en forma de texto, lenguaje, una imagen o movimiento o una acción no perceptible como resolver un problema mediante un cálculo interno.

Esta representación de las acciones, toma como premisa el concepto de *acto del habla* introducido por Austin y posteriormente desarrollado por Searle. De acuerdo con esta visión, las acciones lingüísticas y las motoras son igualmente acciones con efectos en el ámbito del conocimiento o en el físico, respectivamente. En este sentido, las intenciones podrán expresarse con oraciones gramaticalmente correctas, con frases sueltas, con interjecciones o simplemente con algún gesto, y la interpretación consistirá en comprender la intención del hablante, independientemente de la manera que haya elegido para expresarse. (Pineda 2011)

Una vez que el sistema atraviese de una situación inicial a una subsecuente, mediante el cumplimiento de expectativas y la realización de acciones, se irán sucediendo una serie de situaciones que juntas conforman la conversación o la interacción. La historia de la travesía del sistema por las distintas situaciones queda guardada en el sistema, constituye el contexto específico de la conversación y puede recuperarse cada vez que sea necesario.

La representación de los modelos de diálogo puede adquirir diversos grados de complejidad; por ejemplo, cuando la acción se especifica con predicados que incluyan variables o funciones. En el primer caso, el sistema ha requerido al usuario su nombre, la expectativa es que escuche el nombre y la acción, saludarlo utilizando el nombre escuchado:

Gráfica 1.3. Representación gráfica de los modelos de diálogo

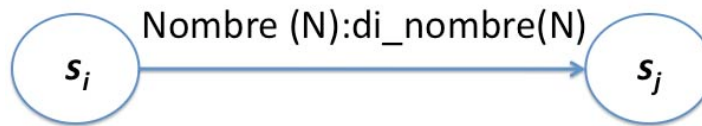


Tabla 1.1. Golem – Adivina la Carta: Ejemplo nombre

Golem	¿Cuál es tu nombre?
Usuario	Rubén
Golem	¡Hola Rubén, me da mucho gusto jugar contigo!

Cuando la acción cambia con base en la historia de la conversación, es posible utilizar funciones para especificar la expectativa y la acción correspondiente, la cual dependerá del resultado del cálculo de la función. Supongamos la misma situación anterior, en la que el sistema le pide al usuario diga su nombre. Si el usuario dice el nombre y el sistema no percibe ningún sonido, le pedirá que repita nuevamente la información. Si después de, por ejemplo, cuatro repeticiones, el sistema sigue sin percibir sonido, informará al usuario que no puede seguir interactuando, debido a que hay un problema mayor.

Por ello, antes de solicitar al usuario que repita su nombre, deberá recuperar la historia y contar las veces que ha solicitado al usuario repita la información para volver a solicitarla o dar el aviso correspondiente y proceder a la salida. En las gráficas 1.3 y 1.4, el punto representa la función que hace el cálculo mencionado.

Gráfica 1.4. Representación gráfica de los modelos de diálogo

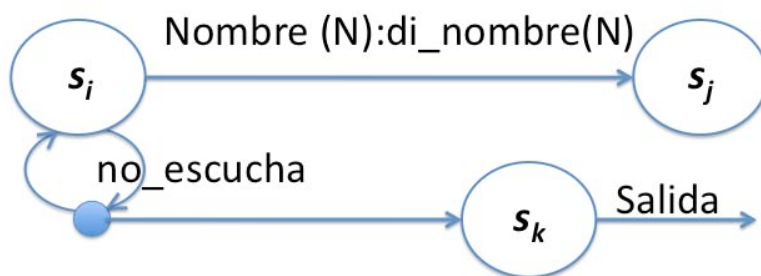
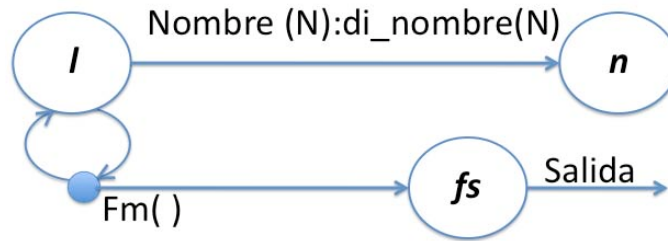


Tabla 1.2. Golem Adivina la carta: Ejemplo reparación

Golem	¿Cuál es tu nombre?
Usuario	
Golem	¿Podrías repetirlo?
Usuario	
Golem	¿Podrías repetirlo?
Usuario	
Golem	¿Podrías repetirlo?
Usuario	
Golem	Salida

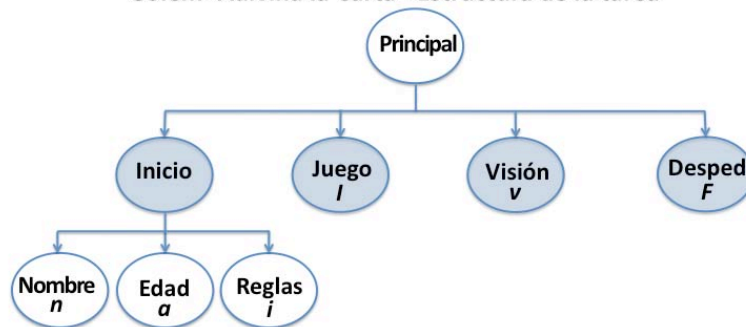
Conforme al concepto de los actos del habla, las expectativas y las acciones pueden adoptar distintas modalidades. No obstante, existe la posibilidad de indicar el tipo de modalidad perceptual que está involucrada en cada una de las situaciones, dependiendo del tipo de dispositivos con que cuenta la aplicación. Las etiquetas *vision (v)* o *listening (l)*, indicarán la modalidad correspondiente, según sea el caso. Otras etiquetas usadas frecuentemente son *neutral (n)*, utilizada cuando la expectativa está vacía o *final (fs)* para señalar el fin del protocolo.

Gráfica 1.5. Representación gráfica de los modelos de diálogo



La representación de los modelos de diálogo es una abstracción que trata de ser simple, aunque las tareas representadas no lo sean. Existen situaciones llamadas recursivas, las cuales contienen modelos de diálogo completos embebidos y subordinados, de tal forma que cuando el sistema llega a una situación recursiva, recorre el modelo completo y una vez terminado regresa al modelo original y continúa hasta llegar a la situación final.

Gráfica 1.6.
Golem- Adivina la Carta - Estructura de la tarea



Esta facilidad de representación, refleja la facilidad para construir las aplicaciones en forma modular y para utilizar y reutilizar algún módulo en la misma aplicación o en otra, conforme se vaya necesitando.

1.3. Estructura de la conversación

La idea principal de presentar los elementos más representativos de una teoría de la estructuración de diálogos, es que constituye el tercero de los sustentos teóricos más

importantes en el desarrollo de aplicaciones y en la especificación de sistemas multimodales inteligentes que se lleva a cabo en el DCC - IIMAS.

La teoría toma como base los conceptos de *estructura conversacional*, *obligaciones* y *acuerdo común*, los cuales cuentan con una larga tradición de estudio en la filosofía, la lingüística, la psicología y la inteligencia artificial, así como en los trabajos de investigación que se llevaron a cabo para el análisis de la estructura de diálogos orientados a resolver tareas y el esquema de transcripción y clasificación DIME-DAMSL que se presenta de manera integral en (Pineda, Estrada, et al. 2007).

Esta teoría sostiene que una conversación está formada por una serie de transacciones que tienen como fin la resolución de una tarea. Las transacciones se forman con actos del habla de los participantes. Los actos del habla pueden tener distintas funciones, por ejemplo, solicitar una información o pedir al interlocutor que realice una acción motora o mental. Estos actos reciben, para fines computacionales y en el marco de la modelación de diálogos prácticos, el nombre de “tipos de actos del habla” y se manejan como un conjunto finito en un dominio delimitado (Pineda 2004).

La teoría toma prestadas las reglas y la terminología contable, en la conducta conversacional, cuando uno de los participantes en una conversación hace una solicitud, un requerimiento, un compromiso o una oferta, crea un cargo que deberá ser satisfecho con un acto del habla para recobrar el equilibrio.

Estas pequeñas estructuras compuestas por actos del habla que se relacionan a partir de la especificación de una intención por parte de un agente y la satisfacción de esa intención por parte del oyente se denominan *obligaciones* y representan el propósito global de una conversación (Tabla 1.3.)

Tabla 1.3. Relaciones de balance de las obligaciones

Cargo	Abono	Participante
Solicitud de información	Respuesta	Otro
Requerimiento de acción	Acción	Otro
Compromiso	Acción	Mismo
Oferta	Acción	Mismo

(Pineda, Estrada, et al. 2007, 3)

En una conversación ideal cada acto del habla se entiende en cuanto se ejecuta y existe un acuerdo o un “piso común” entre los participantes durante toda la conversación. Sin embargo, en la vida real, en una conversación espontánea, el flujo de comunicación se interrumpe frecuentemente por lo que una buena cantidad de elocuciones son necesarias para reparar y restablecer el flujo y proseguir la conversación.

El acuerdo entre los dos participantes en una conversación, también llamado piso común o acuerdo común, puede romperse por dos razones: por una falta de acuerdo entre los participantes, porque uno de los participantes no está de acuerdo con todo o parte del contenido expresado en la conversación; o por problemas de entendimiento, ya sea porque el mensaje no es claro debido a alguna interferencia, como el ruido. En este caso el agente se ve en la necesidad de reafirmar la base de acuerdo en la conversación.

En el plano del *acuerdo común* se ubican los actos del habla a través de los cuales los participantes en una conversación detienen el flujo de la conversación para asegurar que comparten creencias, intenciones y que han entendido la comunicación de la contraparte (Tabla 1.4.).

Tabla 1.4. Relaciones de balance en el plano de los acuerdos

Cargo	Abono	Participante
Solicitud de información	Acuerdo- Acción- Afirmación	Otro
Requerimiento de acción	Acuerdo-Acción	Otro
Oferta	Acuerdo-Acción	Otro
Opción abierta	Acuerdo-Acción	Otro
Afirmación	Acuerdo-Acción	Otro
Reafirmación	Acuerdo-Acción	Otro
Referencia vaga	Reparación	Mismo / Otro
Referencia ambigua	Solución	Mismo / Otro
Acto de entendimiento	Elocución siguiente o actual	Otro
Señal no comprensible	Elocución siguiente misma señal	Otro

(Pineda, Estrada, et al. 2007, 4)

Otro elemento importante a considerar cuando hablamos de la construcción de diálogos prácticos en un dominio específico es el contexto. Por el contexto de una conversación vamos a entender por una parte, un espacio, un tiempo, un tema, una serie de conceptos relevantes, creencias, deseos e intenciones que los interlocutores comparten en cierta medida; por otra parte, también consideraremos como parte del contexto los actos del habla que van surgiendo durante la conversación y que no sólo forman la historia de la misma y que son importantes para dar significado a los actos lingüísticos y mantener la conversación, ya que la condicionan.

Una vez que contamos con una arquitectura cognitiva, con un objetivo global que la tarea específica debe cumplir, con los objetivos parciales claros, un dominio del lenguaje delimitado por el contexto, la técnica de los modelos de diálogo y el conocimiento de la estructura conversacional, quedará todavía pendiente definir la estrategia de la conversación.

La estrategia de la conversación y el modo elegido para transmitir un mensaje son independientes de la estructura; es decir, si la estructura de la conversación que utilizo establece en un momento dado la acción “Saludo”, puedo utilizar no sólo dos modalidades distintas, como vimos en un ejemplo anterior, sino dos mensajes también diferentes:

Participante	Mensaje
Golem	¡Hola!

Participante	Mensaje
Golem	Hola a todos, me llamo Golem y este es el juego de Adivina la Carta. ¡Tenemos jugador!

Ilustremos este hecho con otro ejemplo, retomando el diálogo que utilizamos para ejemplificar el uso de las funciones como parte de las expectativas y las acciones. De acuerdo con la estructura y estrategia definidas, el sistema requerirá hasta cuatro veces a un usuario que repita la información cuando no reciba señal alguna. En el ejemplo anterior, el sistema preguntó el nombre al usuario y utilizamos cuatro veces la misma respuesta ¿Podrías repetirlo? Utilizando cuatro mensajes distintos, el usuario repetirá igualmente la información requerida, pero el efecto del mensaje será distinto.

Tabla 1.5. Golem Adivina la carta: Ejemplo reparación con estrategia

Golem	¿Cuál es tu nombre?
Usuario	
Golem	No te escucho, ¿qué me dijiste?
Usuario	
Golem	Estoy sordo y no escucho, ¿qué me dijiste?
Usuario	
Golem	Me tengo que lavar las orejas, ¿qué me dijiste?
Usuario	
Golem	

Este último ejemplo fue tomado de versiones distintas del juego “Adivina la Carta”. Aunque a simple vista parecerían cambios de etiqueta, están respaldados con cambios que se dan desde la estructura del modelo de diálogo y fueron resultado de una evaluación previa de la aplicación.

Podemos concluir este capítulo diciendo que si bien la experiencia en el análisis de los actos del habla del proyecto DAMSL y en la investigación para el desarrollo del esquema DIME-DAMSL mostró que aun los diálogos prácticos orientados a la solución de tareas específicas y sencillas son de una complejidad considerable, los investigadores han continuado construyendo sistemas conversacionales sencillos con lo que tienen a su disposición: una arquitectura cognitiva, tareas específicas y sencillas con objetivos claros, dominios delimitados, los modelos de diálogo y el conocimiento de la estructura conversacional y la tecnología disponible al momento.

En el capítulo siguiente se explicará con amplitud la aplicación “Adivina la Carta”, tomando como base las definiciones de este capítulo para posteriormente, pasar a los resultados de la evaluación, que es objeto de este trabajo.

CAPÍTULO 2.

GOLEM: ADIVINA LA CARTA

Una vez expuestos los conceptos más importantes para la construcción y desarrollo de los sistemas de diálogo y los elementos que constituyen el marco de referencia que guía los trabajos que se llevan a cabo en el DCC-IIMAS, procederemos a explicar en este capítulo su aplicación práctica a través de la del sistema “Adivina la Carta” describiendo en qué consiste y cómo funciona.

A partir del esquema de la arquitectura cognitiva ejemplificaremos por qué desde su origen está dirigido a su interacción con el mundo o su representación y describiremos los dispositivos y herramientas con los que cuenta para ello. A partir de los modelos de diálogo, explicaremos la estructura general de la tarea y las opciones que por una parte, el sistema tiene para reparar los diálogos que pudieran interrumpir la comunicación y el entendimiento y, en consecuencia, el cumplimiento de la tarea. Por último, a partir de la teoría de la estructura de la conversación se describe una sesión típica del juego, así como los mecanismos de recuperación que existen para equilibrar la conversación y para corregir algún error en el reconocimiento.

Para finalizar la descripción del sistema, haremos un recuento sobre el trabajo que se hace relacionado con el conocimiento sintáctico, semántico y pragmático que dan lugar al modelo de lenguaje. Se hace referencia a la construcción del corpus en un inicio, así

como al trabajo actualización del modelo, que incluye la etiquetación, necesaria para el aprendizaje y su evaluación.

2.1. Funcionamiento de la aplicación

“Golem - Adivina la Carta” es un sistema de diálogo en español que maneja e interpreta señales en diferentes modalidades; mediante el lenguaje y la visión computacional. (Meza, Salinas and Avilés, et al. 2010) El sistema se ubica en un módulo permanente en Universum, el Museo de las Ciencias de la Universidad Nacional Autónoma de México (UNAM).



La idea de que la aplicación esté exhibida de manera permanente es una situación privilegiada, ya que esto permite a los investigadores llevar la aplicación fuera del laboratorio para ensayar y trabajar de manera continua con los usuarios en un ambiente real, en un esquema de colaboración, con beneficios para el museo y para el DCC - IIMAS. El módulo fue inaugurado el 12 de diciembre de 2009 para conmemorar en México el año internacional de la Astronomía.

“Golem – Adivina la Carta” fue hecho de manera especial para los niños, para que jugaran e interactuaran con él, principalmente a través del lenguaje hablado. El módulo lo compone una cabina en la que está todo el equipo necesario para el juego: en una mesa se localizan las 10 cartas con motivos astronómicos, una de las cuales será elegida durante el juego; un monitor conectado a una computadora personal oculta que permite al sistema desplegar imágenes y otro tipo de información que facilite la

comunicación con el usuario; un micrófono hace las veces de oídos y permite captar la comunicación del usuario; una cámara que hace las veces de ojos de la aplicación; dos sillas y un par de bocinas para escuchar al robot. La cabina tiene como fin disminuir un poco el ruido exterior; no obstante, el público que permanece en el exterior de la cabina puede seguir el juego mirando por el cristal que la cabina tiene enfrente, escuchando lo que el robot y el jugador dicen a través de otro par de bocinas ubicadas fuera del módulo y a través de otro monitor que permite ver lo mismo que el jugador.

La cámara y el micrófono son los dispositivos disponibles que permiten al sistema captar las señales sonoras y visuales emitidas por el usuario, que son reconocidas e interpretadas por el sistema y, en el marco de las expectativas y del contexto del juego, el intérprete entrega la respuesta correspondiente a través de las bocinas o el monitor, respectivamente para el caso de la acción sonora o visual.



En “Adivina la Carta” el objetivo del juego para el niño consiste en obtener, a través de la conversación, información suficiente que le permita deducir una carta elegida de manera aleatoria por el sistema.

Para la interacción con el sistema se tomaron inicialmente participantes con edades entre los 10 y los 14, años en virtud de que los niños utilizan un lenguaje más sencillo, aun cuando hablen o se refieran a conceptos complicados. Posteriormente, cuando a través del uso de la aplicación se consideró que la tarea podía ser complicada se extendió la participación a niños de hasta 16 años; más adelante, cuando se presenten

los datos de la evaluación, se verá los efectos de la dificultad del juego en la resolución de la tarea, como en el uso del lenguaje.

Para explicar la creación de diálogos entre la persona y la computadora, es necesario hablar de los procesos de reconocimiento y síntesis, dos tecnologías necesarias para crear un diálogo, que le permiten a la aplicación manejar e interpretar el lenguaje, es decir, escuchar el lenguaje hablado, interpretar, generar una respuesta y comunicarla al usuario. Para el desarrollo de estos sistemas se requiere la participación de ingenieros, computólogos y lingüistas.

En el reconocimiento de voz, el sistema capta una señal sonora a través del micrófono, extrae ciertas características que compara con datos y patrones obtenidos anteriormente, elige el modelo que más se asemeje a la señal de entrada y la convierte a un formato de representación que podrá ser correlacionado con información lingüística.

El reconocimiento requiere de un reconocedor de voz, del entrenamiento de modelos acústicos y la generación de modelos del lenguaje, para que al enfrentarse con una señal sonora, el programa tenga datos y patrones contra los cuales comparar esa señal. Los datos y patrones provienen de un corpus del habla generado previamente. Lo ideal es que el corpus utilizado para el entrenamiento sea extenso, para crear la mayor cantidad de modelos acústicos que reflejen la variación individual de las voces, los acentos y las velocidades de elocución.

Un corpus acústico o para el reconocimiento de voz, es un diccionario que contiene: las palabras que serán utilizadas por el sistema, el conjunto de sonidos que forma cada palabra, así como la transcripción fonética y ortográfica alineada temporalmente con la señal sonora . El fonetista es el responsable de crear la serie de transcripciones o etiquetas que sirven para crear los modelos acústicos y los diccionarios de pronunciación para la creación de sistemas computacionales para el reconocimiento.

En el caso de “Adivina la Carta” se utilizó el Corpus Dimex100 (Pineda, Villaseñor, et al. 2004) y el Corpus Dimex 100 niños. El primero consiste en una base de datos de oraciones habladas en español de México que incluyen su respectivo análisis fonético. Este corpus fue diseñado, recopilado y analizado en el DCC-IIMAS en el contexto del proyecto DIME, en una alianza con fonetistas de la Facultad de Filosofía y Letras de la UNAM y ha sido utilizado para la creación de los modelos acústicos y los diccionarios de pronunciación de los sistemas de reconocimiento. El corpus Dimex100 niños consta de una base de datos de oraciones generadas con las voces de cien niños, con la idea de adaptar la tecnología a la fonética infantil y facilitar al robot el reconocimiento.

Una vez realizado el reconocimiento, el código obtenido se pasa al intérprete. Para realizar su función, el intérprete utiliza los conceptos generales y particulares guardados en la memoria semántica, así como la información de los hechos que el agente guarda en la memoria perceptual o imaginística. La información que requiere el intérprete depende de la situación de diálogo por la que atraviesa en ese momento, es decir, de las expectativas que espera se cumplan y la tarea específica a solucionar. De tal forma que de la memoria semántica recuperará los conceptos necesarios y de la memoria perceptual la asociación entre la interpretación (significado) y las imágenes no interpretadas. La interpretación de la intención corresponde al acto del habla y será un insumo importante para determinar la acción a realizar.

Para la realización de los mensajes orales, el sistema utiliza un sistema de síntesis del habla. La síntesis consiste en transformar un texto escrito en una realización sonora; para ello es necesario que el texto tenga un formato electrónico apropiado para que pueda ser oralizado apropiadamente (López Escobedo 2004, 13). Si la función del sistema fuera únicamente articular mensajes orales, lo más sencillo sería reproducir mensajes grabados previamente. Sin embargo, en estas aplicaciones, la tarea de traducción de un texto a su forma acústica constituye la culminación de un proceso más complejo que relaciona los actos del habla, los conceptos de la memoria semántica y los hechos de la memoria perceptual para generar una conversación espontánea.

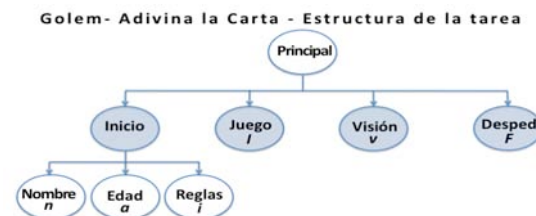
2.2. Descripción de la tarea y objetivo del juego

Los modelos de diálogo descritos en el capítulo anterior, constituyen abstracciones que permiten representar la estructura y el contexto global de una conversación. En el juego “Adivina la Carta” los niños entablan una pequeña conversación con el sistema, con el fin de obtener información suficiente que les permita deducir la carta que el sistema eligió. Durante el juego, el sistema responde utilizando un lenguaje sintetizado, efectos de sonido y despliega en el monitor imágenes acordes con la respuesta.

La conversación está organizada en seis subdiálogos que integran la estructura de la tarea; la gráfica y tablas siguientes ilustran la estructura general de “Adivina la Carta”. El administrador del diálogo interpreta cada uno de los modelos de manera continua de tal manera que se ensamblan de manera dinámica durante una interacción produciendo conversaciones ricas y naturales.

Tabla 3.1. Golem – Adivina la Carta
- Modelo de diálogo general –

- (n) Saludo y el reconocimiento del nombre
- (a) Pregunta y confirmación de la edad
- (i) Explicación y comprensión de las reglas del juego.
- (I) Interrogatorio, el usuario puede formular hasta cuatro preguntas
- (V) Selección y verificación de la carta seleccionada por el usuario
- (F) Despedida



El juego inicia **(n)** cuando el sistema se presenta, saluda y pregunta al jugador su nombre para saludarlo en forma personalizada. Posteriormente, **(a)** el sistema pregunta al jugador su edad y confirma la respuesta. Esta información es útil porque el juego puede no resultar tan divertido para niños menores de 10 años. Una vez que el participante

está identificado por su nombre y edad, puede jugar. El sistema ofrece posteriormente una explicación de las instrucciones del juego **(i)**; el usuario puede elegir escuchar las instrucciones o no. En caso de que el usuario haya elegido escuchar las instrucciones, el sistema pregunta si las ha comprendido o si quiere escucharlas nuevamente. Al finalizar, este módulo inicia el interrogatorio.

Antes de iniciar las preguntas, en este módulo **(I)** el sistema le pide al usuario que se fije bien en las cartas mientras él elige una, después le pide que haga la primera pregunta. El niño tiene hasta cuatro oportunidades para formular preguntas distintas sobre las características, la forma y los colores de los objetos que se representan en las cartas, que le permitan deducir y adivinar de cuál se trata. Las preguntas generalmente son sencillas y cortas, tales como ¿Es redondo? ¿Es de color azul? ¿Es un planeta?. A cada pregunta formulada, el sistema responde si la característica está o no presente en la carta elegida y lo invita a realizar la siguiente pregunta, hasta agotar las oportunidades.

Si alguno de los jugadores cree que puede adivinar la carta antes de formular las cuatro preguntas, será suficiente con que así lo exprese. El sistema entenderá que el jugador está listo para terminar el juego.

Ya sea porque el jugador agote las cuatro oportunidades para preguntar o porque decida anticipar su opción a terminar, deberá mostrar frente a la cámara la carta que cree que el sistema eligió **(v)**; si el jugador intentó anticipar la respuesta, para confirmar la carta que mencionó previamente. Después de mostrar la carta el sistema pide una confirmación verbal de la carta mostrada para comparar y notifica al jugador si ganó el juego porque adivinó la carta o le dice cuál fue la carta seleccionada.

En el módulo final **(F)**, el robot agradece al niño su participación, le pide que sonría, le toma una fotografía y lo despide con su nombre. El juego ha terminado.

De acuerdo con lo definido desde la estructura de la interacción, por lo general, el sistema lleva la iniciativa de la conversación. El usuario responde y todas sus respuestas son importantes, aunque no todas son críticas para el cumplimiento del objetivo; por

ejemplo, el participante se siente gratamente impresionado cuando el robot lo saluda por su nombre al inicio del juego y más aun cuando sigue recordando el nombre al momento de la despedida, ya que se siente aludido directamente. Por otra parte, todas las preguntas del interrogatorio son críticas para cumplir con el objetivo del juego, ya que contar con un número mayor o menor de pistas son importantes para terminar el juego y, en su caso, adivinar la carta.

Existen preguntas que no son rectificaciones, pero que se plantean para confirmar puntos críticos de información que pudieran surgir; por ejemplo, el usuario tiene la opción de terminar de manera anticipada el juego, si cree que puede adivinar la carta que el sistema eligió. Para ello, en lugar de preguntar una característica del objeto, el usuario dirá el nombre de la carta elegida. El sistema debe saber interpretar este acto y dirigirá al usuario hacia el final del juego, no sin antes rectificar el deseo del usuario de adivinar de manera anticipada la carta. Si el usuario no confirma su deseo de terminar, el juego continuará.

Independientemente de que las preguntas sean importantes para la conclusión de la tarea o no, el modelo incluye la posibilidad de reparar cualquier error que pueda surgir en el reconocimiento a lo largo de toda la conversación. En el caso del reconocimiento de voz a través de dos estrategias; una es la que se utiliza cuando el reconocedor no produce ningún resultado, en virtud de que no percibe ninguna señal (por ejemplo porque no hay nada que escuchar o porque el usuario no habló). La segunda estrategia se utiliza cuando el sistema no entiende lo que dice el usuario, es decir cuando lo que percibe no corresponde a ninguna de los modelos aprendidos. El número de veces que puede invocarse el modelo de recuperación se controla mediante una función con el fin de hacer el diálogo más fluido y coherente.

Por último, todas las respuestas son importantes en tanto que constituyen el contexto específico que va construyendo de manera dinámica la historia de cada interacción. La historia es importante para saber cuántas preguntas ha realizado el participante o

cuántas veces debe rectificar la información que no se ha captado o no se captado con claridad.

Como hemos explicado anteriormente, los modelos de diálogo permiten esquematizar los protocolos conversacionales que se utilizan en la construcción de sistemas de diálogo. En este caso hemos revisado el modelo de diálogo general que se utilizó para construir el juego de “Adivina la Carta”. El modelo utilizado nos permite ver que existe una tarea global definida, así como distintas sub-tareas determinadas. Las tareas específicas están esquematizadas en modelos semejantes, cada uno de los cuales incluye expectativas y acciones específicas. El administrador del diálogo se encarga de ir ensamblando cada uno de los modelos, pero el logro de conversaciones ricas y naturales dependerá no sólo de una interpretación subsecuente y continua de los modelos, sino del contenido de los diálogos y la lógica de la conversación.

La tabla siguiente presenta un ejemplo de una interacción entre Sofía, una joven de 13 años y el sistema en el juego “Adivina la Carta” y fue tomado de las evaluaciones que se llevaron a cabo para la evaluación de la aplicación. Esta tabla nos es útil para mostrar la estructura de la conversación; las tres primeras columnas muestran que la conversación está organizada en turnos en los que participan de manera alternada el usuario y el sistema, y toma como base la estructura definida en los modelos de diálogo. El diálogo puede seguirse tal y como ocurrió en la columna Elocución y la última columna nos da idea de las imágenes que el sistema despliega en el monitor durante el juego.

Este diálogo constituye un buen ejemplo de un juego típico, porque Sofía termina la tarea y además adivina la carta que había pensado el sistema. La conclusión de la tarea es un indicador importante para los desarrolladores y para la evaluación del sistema, ya que éste debe ser capaz de recuperar en cada subdiálogo la información esperada, con el fin de pasar subsecuentemente por todos los módulos hasta el final.

Tabla 3.2. Golem - Adivina la carta

Interacción O18 ^{1/}

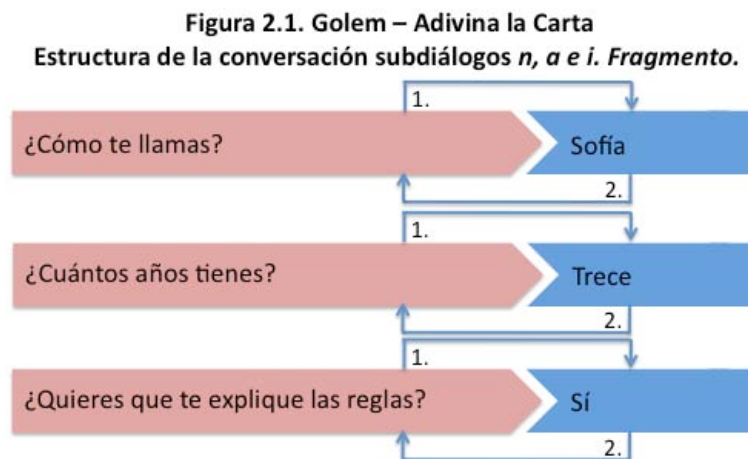
Turno	Partic.	Subdiál.	Elocución	Acción visual
1	Golem	In-(n)	Hola a todos, me llamo Golem y este es el juego de Adivina la Carta. Tenemos jugador. ¿Cuál es tu nombre?	Bienvenido
2	Usuario	In-(n)	Sofía	
3	Golem	In-(n)	Hola Sofía, me da mucho gusto jugar contigo.	
		In-(a)	¿Cuál es tu edad?	
4	Usuario	In-(a)	Trece años	
5	Golem	In-(a)	¿Seguro que tienes trece?	
6	Usuario	In-(a)	Sí	
7	Golem	In-(a)	Muy bien, podemos jugar.	
		In-(i)	¿Quiéres que te explique las reglas del juego?	
8	Usuario	In-(i)	Sí	
9	Golem	In-(i)	Presta atención. Enfrente de ti tenemos distintas cartas.	Imagen con las 10 cartas
		In-(i)	De ellas voy a escoger una al azar y tú adivinarás de cuál se trata. Para tratar de adivinar podrás hacerme cuatro preguntas sobre el color...	Imagen con colores
		In-(i)	... o la forma de las cartas, o si se parece a otro objeto.	Imagen con formas
		In-(i)	Me puedes hacer preguntas como las que aparecen en pantalla. Después de la ronda de preguntas compararemos mi carta y la que tú crees que elegí. ¿Entendiste las instrucciones?	Ejemplo preguntas.
10	Usuario	In-(i)	Sí	
11	Golem	In-(i)	OK. Fíjate muy bien en las cartas mientras escojo una.	
		I	Empecemos el juego, por favor hazme tu primera pregunta.	
12	Usuario	I	¿Es rojo?	
13	Golem	I	No, en la carta que veo, no hay nada rojo. Vamos, hazme la siguiente pregunta.	Imagen con color rojo
14	Usuario	I	¿Es azul?	
15	Golem	I	Excelente, en la carta puedo ver una gran porción de color azul. Vamos, hazme la siguiente pregunta.	Imagen con olas de color azul.
16	Usuario	I	¿Hay vida?	
17	Golem	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 piénsala muy bien.	Imagen con plantas y luz.
18	Usuario	I	¿Tiene verde?	
19	Golem	I	Muy bien, en mi carta hay algunas porciones que son de color verde.	Imagen con plantas de color verde.
		V	Ha finalizado tu ronda de preguntas. Por favor muéstrame la carta.	Imagen con una interrogación.
20	Usuario	V	Muestra la carta con la imagen del planeta Tierra	
21	Golem	V	No tengo muy buena vista. Es ésta la carta que me mostraste	Imagen con el planeta Tierra
22	Usuario	V	Sí	Imagen con el planeta Tierra
23	Golem	V	Muy bien, has adivinado. La carta que escogí fue precisamente la de la tierra. Bien hecho.	Imagen ¡Felicidades!
		F	¡Oh! Pero antes de irte por favor mira a la cámara. ¡Sonríe! Gracias por jugar conmigo Sofía, regresa pronto.	La cámara toma una foto y la despliega

^{1/} Evaluación realizada del 30 de junio al 1 de julio de 2010.

A partir de este mismo ejemplo pasaremos a ilustrar con éste y otros ejemplos prácticos la estructuración de los diálogos, de acuerdo con la teoría de la estructuración de diálogos que presentamos en el capítulo anterior.

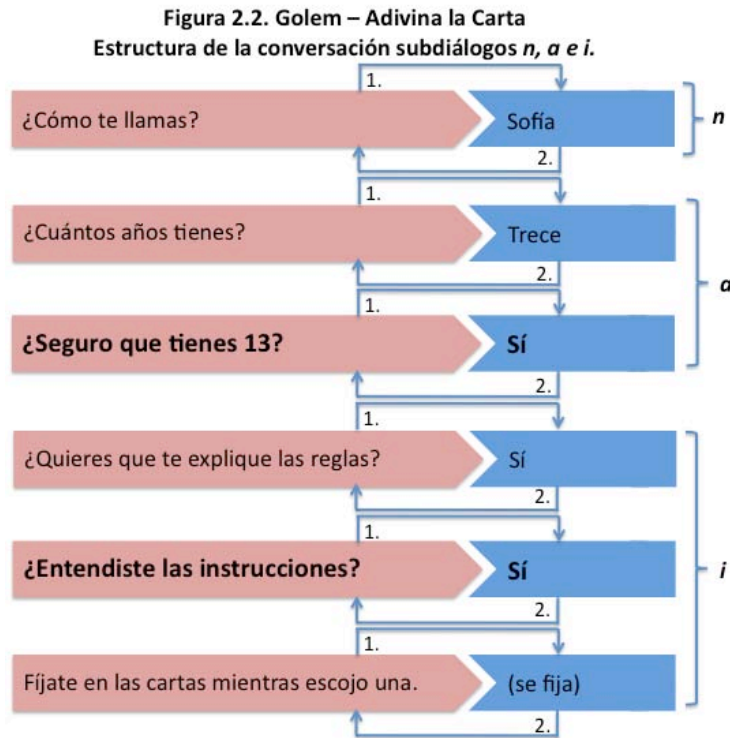
2.3 Estructura de la conversación

La conclusión de la tarea de manera exitosa en el juego “Adivina la Carta” requiere que el participante y el sistema resuelvan una serie de “cargos” que ambos van presentando durante el juego. El corazón de la conversación está en la serie de preguntas que el niño hace para conocer las características de la carta elegida por el sistema, pero antes de este interrogatorio deben cumplirse algunas formalidades de cortesía.



La figura 2.1. muestra la estructura de la conversación que inicia después de un saludo general. El diálogo inicia con dos solicitudes de información, nombre y edad, y una oferta que quedan resueltas de manera inmediata. En virtud de que para el sistema de “Adivina la Carta” la edad del participante y la comprensión de las instrucciones del juego son dos asuntos importantes, aun y cuando haya entendido la respuesta del usuario, interrumpe la conversación con dos elocuciones que le permiten asegurar que

ha entendido la comunicación de la contraparte antes de proceder propiamente al juego (Figura 2.2.).



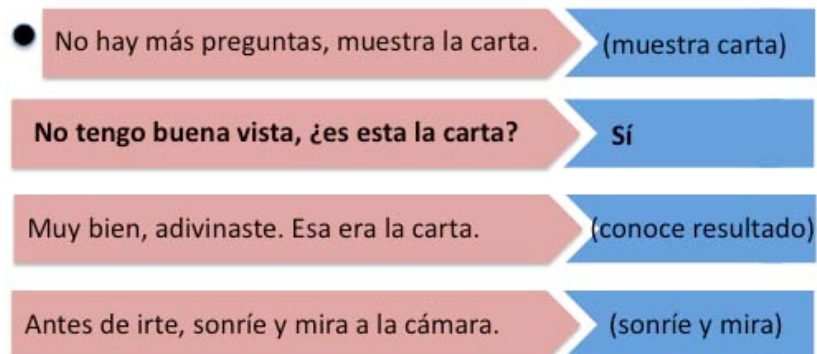
Durante el interrogatorio la estructura de la conversación es más compleja. El sistema requiere al usuario haga una primera pregunta, cargo que queda satisfecho con el planteamiento de una primera pregunta, pero que a su vez crea un cargo nuevo, esta vez de parte del usuario al sistema que quedará satisfecho y cerrará por completo la transacción una vez que el sistema dé la respuesta al usuario, tal y como lo observamos en la figura 2.3. en este fragmento de la conversación:

Figura 2.3. Golem – Adivina la Carta
Estructura de la conversación Interrogatorio. Interacción O18.



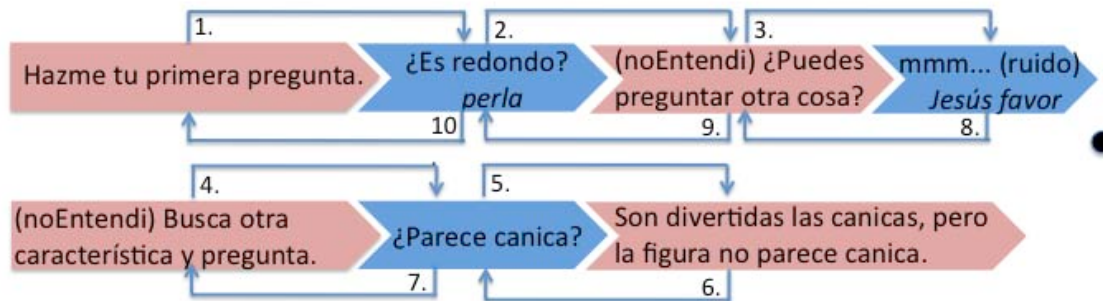
Por lo que respecta a la estructura, la parte final de la conversación, que corresponde a los subdiálogos (v) de visión y (F) de despedida, vuelve a la forma inicial de cargo / satisfacción. El subdiálogo de visión tiene la peculiaridad de que el usuario actúa a través una imagen, en lugar del lenguaje, para mostrar la carta que cree que eligió el robot y que éste confirma de manera verbal antes de dar, finalmente, el resultado del juego al niño.

Figura 2.4. Golem – Adivina la Carta
Estructura de la conversación Subdiálogos v y F. Interacción O18.



En una conversación ideal cada acto del habla se entiende en cuanto se ejecuta y cada solicitud o requerimiento de información tiene una satisfacción inmediata; sin embargo tal y como sucede en la vida real, durante las interacciones con la aplicación el flujo de comunicación se interrumpe frecuentemente por lo que un número importante de elocuciones son necesarias para reparar y restablecer el flujo de la comunicación.

Figura 2.5. Golem – Adivina la Carta
Estructura de una acción de reparación. Fragmento interacción R5.



Por último, este mismo ejemplo sirve para ejemplificar cómo el contexto es importante para dar significado a los actos lingüísticos y mantener la conversación. Si revisamos el registro de la conversación, podemos observar que en lugar de “redondo” el sistema reconoció “perla” y en lugar de “mmm” acompañado de ruido el sistema reconoció “Jesús favor”, palabras que no forman parte del contexto del juego ni del momento de la conversación, por lo que lanza en ambos casos las acciones correspondientes de reparación.

Las acciones de reparación son indispensables para reestablecer la comunicación entre el usuario y el agente y son importantes porque permiten a los usuarios terminar el juego, es decir terminar la tarea; sin embargo, cada acción de reparación alarga la conversación, el número de elocuciones y el tiempo total de la transacción, lo que refleja el nivel de desempeño del sistema y afecta en forma directa la percepción del usuario acerca del mismo.

Para conocer los elementos que son importantes en la evaluación de los sistemas de diálogo, procederemos a explicar en el siguiente capítulo los elementos que se utilizan para evaluar y calificar los sistemas de diálogo para pasar posteriormente de lleno a la evaluación y sus resultados.

CAPÍTULO 3.

METODOLOGÍAS PARA LA EVALUACIÓN DE SISTEMAS DE DIÁLOGO.

La evaluación en general es un proceso técnico a través del cual se valora en forma sistemática y continua el rendimiento y el comportamiento de un servicio. La evaluación puede ser integral o parcial, pero en todo caso, permite obtener información sobre el funcionamiento del servicio evaluado, determinar en qué medida se están cumpliendo las metas de desempeño y calidad, valorar el impacto de cambios realizados en algún proceso y predecir, en alguna medida, el comportamiento futuro.

Para cualquier investigador es importante contar con evidencia que permita a la comunidad científica y al público entender y hacer juicios confiables con respecto al trabajo que realiza, por lo que la confiabilidad de los procedimientos de medición empleados es clave. Si los investigadores no pueden generar algún grado de acuerdo con sus resultados y si no hay oportunidad de replicarlos, éstos tendrán una validez muy relativa.

En el caso de la evaluación de los sistemas de diálogo, como en otros campos, el avance en las metodologías de evaluación ha ido de la mano del avance en el desarrollo de las

aplicaciones, que es indudable. Un impulso importante se ha dado a partir de la introducción en el mercado de aplicaciones que utilizan este tipo de tecnologías, que hacen más urgente la necesidad de conocer las capacidades de estos agentes. Aunque a la fecha no existe consenso sobre una metodología general de evaluación para los sistemas de diálogo, para el ejercicio que sustenta este trabajo elegimos Paradise (Walker, et al. 1997), una metodología de evaluación que ha sido citada en diversos artículos y trabajos de investigación (Skantze 2003; Hajdinjak and Mihelic 2006; Cuayáhuatl 2009; Cuayáhuatl, Renals, et al. 2010), aunque también ha sido objeto de varios señalamientos debido a sus limitaciones (Paek 2001; Hjalmarsson 2002) . Las creadoras de la metodología la presentan como una metodología general de evaluación que integra elementos diversos, utilizados hasta entonces para la medición de desempeño de sistemas conversacionales, así como en otras áreas vinculadas al procesamiento del lenguaje natural. En virtud de que se trata de la metodología que utilizaremos en la evaluación, se hace una descripción de cada una de sus etapas, las cuales se presentarán en el capítulo siguiente con los resultados de la evaluación de “Adivina la Carta”.

Previamente presentamos algunos antecedentes importantes que fueron útiles en la evaluación de los sistemas de diálogo y en la propuesta de metodología que utilizaremos en la evaluación. Asimismo, dedicamos una sección a la descripción del coeficiente *kappa* como una aportación en el camino a encontrar herramientas confiables para la evaluación de sistemas de diálogo. (Garfield 1986; Cerda and Villarroel del P. 2008; López de Ullibarri Galparsoro and Pita Fernández 2004; Molinero 2001).

Durante el ejercicio de evaluación y como parte de las evaluaciones preliminares, fuimos detectando algunas necesidades particulares tanto para el levantamiento de información como para su valoración, que corresponden a las particularidades de la tarea y del sistema, por lo que, en conjunto con los desarrolladores, consideramos necesario tomar algunas decisiones para hacer algunos ajustes a la metodología y su aplicación. Estos ajustes se irán anotando durante la descripción de la aplicación del

coeficiente *kappa* a la evaluación de los sistemas de diálogo y durante la descripción de la metodología de evaluación.

3.1. Técnicas y herramientas de evaluación de sistemas de diálogo.

En el campo de la lingüística computacional y las ciencias cognitivas, los investigadores han buscado herramientas que les permitan sustentar de manera confiable los juicios y resultados de sus investigaciones, haciendo referencia a medidas estadísticas fáciles de interpretar y de comparar. En su artículo “Assessing Agreement on Classification Tasks: The *Kappa* Statistic, Jean Carletta (Carletta 1996, 249) opinaba que a pesar del uso de diversas métricas de evaluación, objetivas y subjetivas, los resultados de las investigaciones realizadas en estos campos estaban basados en juicios subjetivos ya que ninguno tenía una explicación fácil de entender, lo que finalmente hacía que los resultados se basaran en la interpretación del autor y estuvieran sujetos a que la explicación le pareciera aceptable a los lectores.

En la reseña que hicieron sobre la evaluación de los sistemas para el procesamiento del lenguaje natural J. R. Galliers y K. Sparck Jones (Sparck - Jones and Galliers 1996), atribuyeron a la limitación en las capacidades de los procesadores de lenguaje y de recuperación de información que la evaluación en estos campos avanzara menos que en otros campos de investigación. Sin embargo, estas autoras identifican un cambio importante generado a partir de principios de los años 90 en los avances logrados en todas las áreas relacionadas con el procesamiento del lenguaje y, por lo tanto, también en la evaluación. Explican este avance, por una parte, como el resultado de la introducción en el mercado de aplicaciones que utilizan estas tecnologías, y por otra al interés de la DARPA (Defense Advanced Research Projects Agency), la agencia del Departamento de Defensa de los Estados Unidos dedicada a promover la investigación, que colocó un foco de atención sobre el procesamiento y reconocimiento del lenguaje.

Aunque los objetivos de la evaluación en áreas como la traducción automática, la recuperación y extracción de información, la comprensión del lenguaje, el reconocimiento del habla, la síntesis de voz y la comprensión del lenguaje son específicos, los esfuerzos de evaluación realizados en las distintas áreas han generado experiencias y aprendizajes que se comparten. Los avances más importantes se atribuyen a la traducción automática, ya que es el área que cuenta desde hace más tiempo con sistemas que trabajan y se comercializan, por lo que cuentan con una actividad permanente dirigida a mejorar los sistemas existentes con respecto a ellos mismos y a los de la competencia. En este caso los objetivos de las mejoras y de la evaluación tienen un fin muy claro, las aplicaciones deben ser mejores, más rápidas y más económicas.

En comparación con los sistemas de traducción automática, la evaluación de sistemas que utilizan el reconocimiento del habla enfrentan la dificultad y el reto adicional de manejar la señal acústica, y como hemos visto, existe un gran progreso en el desarrollo de métricas y metodologías de medición para muchos componentes individuales, pero el progreso ha sido menor cuando se trata de evaluar todos estos elementos de manera simultánea. Los sistemas de diálogo pueden ser evaluados midiendo la habilidad del sistema para ayudar a los usuarios a alcanzar sus metas, o de la posibilidad para recuperarse de errores en caso de un error en el reconocimiento, pero entre más completa es la interacción entre el usuario y el sistema, se vuelve más difícil de abordar la evaluación del sistema como un todo.

Durante el trabajo de investigación realizado para esta tesis revisamos principalmente los ejercicios de evaluación para sistemas de diálogo, identificamos las mejores prácticas y distintas métricas que tienen como fin generar los indicadores principales de desempeño de los sistemas de diálogo: de éstas seleccionamos aquéllas que tuvieron sentido para nuestra evaluación, las cuales se resumen a continuación:

- El registro del comportamiento y uso del sistema que se genera cada vez que se utiliza una aplicación es un buen ejemplo de una práctica común que genera

información objetiva útil para la medición, para el diagnóstico y la evaluación directa. A través del registro se recupera información que no requiere el juicio o participación del humano y permite determinar indicadores tales como:

- **Éxito en el logro de la tarea:** Este indicador es uno de los más relevantes para la evaluación, por una parte, porque es importante que los sistemas respondan razonablemente, de tal manera que el usuario alcance la meta que se propone, independientemente de que exista una consideración económica u operacional. Pero por otra, porque la definición de la tarea determina la aplicabilidad de otros indicadores. **(Walker, Fromer and Shrikanth 1998; Kamm, Litman and Walker 1998; M. A. Walker 2000; Hajdinjak and Mihelic 2006)**
- **Duración de los diálogos:** Este indicador está muy relacionado con la descripción de la tarea. Este indicador tiene sentido cuando una tarea es específica e igual para todos los usuarios. En nuestro caso, la duración de un diálogo sólo puede extenderse por acciones de reparación. **(Walker, Fromer and Shrikanth 1998; Kamm, Litman and Walker 1998; Litmann and Pan 1999; M. A. Walker 2000; Walker, Kamm and Boland 2000; Skantze 2003; Hajdinjak and Mihelic 2006; Cuayáhuitl 2009)**
- **Número de elocuciones (usuario):** Número de actos en lengua hablada realizados por el usuario durante la interacción. **(Walker, Fromer and Shrikanth 1998; Kamm, Litman and Walker 1998; Litmann and Pan 1999; M. A. Walker 2000; Walker, Kamm and Boland 2000; Hjalmarsson 2002; Hajdinjak and Mihelic 2006; Cuayáhuitl 2009)**
- **Número de elocuciones (sistema):** Número de actos en lengua hablada realizados por el sistema durante la interacción. En el caso de Adivina la Carta, considerando la estructura del diálogo, estos dos indicadores están muy relacionados. **(Walker, Fromer and Shrikanth 1998; Litmann and Pan 1999; M. A. Walker 2000; Walker, Kamm and Boland 2000; Hjalmarsson 2002; Hajdinjak and Mihelic 2006; Cuayáhuitl 2009)**
- **Número de acciones de visión:** Número de actos realizados por el usuario durante la interacción utilizando el módulo de visión del sistema. Este es un indicador nuevo que no aparece en ninguno de los artículos revisados. Esta situación es posible debido a que la arquitectura del sistema trabaja con expectativas que se definen con anticipación y que son independientes de la modalidad.

- El registro digital de las elocuciones de los participantes y su posterior transcripción permite comparar lo dicho por el usuario con lo que el sistema reconoce y elaborar otros cálculos sobre los niveles de reconocimiento. En esta labor es indispensable la participación de evaluadores humanos, y permite determinar indicadores como:
 - % error en reconocimiento por palabra:
 - % acierto en reconocimiento por palabra:
 - % Reconocimiento por diálogo:
 - Representaciones con predicado:
 - Representaciones con argumentos:
 - Promedio error en predicados:
 - Promedio error en argumentos:

- La grabación externa de audio y video de las interacciones constituye una evidencia útil para validar el curso de la interacción y corregir cualquier error que haya surgido en la interacción, así como para conocer la actitud de los usuarios al momento de interactuar con el sistema. Las apreciaciones de este tipo de material requieren y están sujetas indispensablemente a la valoración humana.

- La evaluación de la percepción de los usuarios con respecto al uso del sistema, aunque subjetiva, es una práctica con varios beneficios, ya que representa el punto de vista de los usuarios. Para ello se utiliza generalmente un cuestionario para recoger, entre otros aspectos, la opinión de los usuarios respecto a la usabilidad de la aplicación y las intenciones de continuar usando el sistema. Más adelante, cuando se aborde la descripción de la metodología de evaluación retomaremos el tema de esta evaluación.

- En virtud de que no todos los errores generados en las interacciones son observables a través del registro automático, diseñamos una bitácora para el levantamiento de algunos comportamientos de los usuarios que son motivo de

fallas en la comunicación; por ejemplo, palabras incompletas, titubeo en la respuesta, ruido de fondo, dejar de hablar, entre otros. Para este tipo de situaciones, la bitácora manual permite registrar acontecimientos relevantes que en opinión del evaluador deban quedar registrados en el momento de la interacción. La bitácora, que también fue objeto de mejoras durante las primeras evaluaciones, permite documentar la historia de la interacción con pocas anotaciones, permite conocer con relativa rapidez y precisión los resultados de la evaluación y hacer estimaciones puntuales, antes del procesamiento total de la información.

Durante la investigación comprobamos que existen diversos criterios de evaluación que pueden aplicar a un sistema, y existen diversos sistemas con distintos objetivos que pueden alcanzarse a través de la consecución de diferentes tareas; sin embargo, la evaluación para ser efectiva debe ser precisa y rigurosa, representada en medidas y métodos que puedan replicarse con un costo beneficio adecuado, ya que los costos de la evaluación afectan no sólo los experimentos, sino también la investigación.

La metodología de evaluación que utilizamos para este trabajo combina varios elementos importantes que hemos expuesto; en primer lugar, una serie de mediciones objetivas y subjetivas que constituyen los costos involucrados en la ejecución del diálogo; por otra, una medición para calcular el éxito en la realización de la tarea específica, basada en el cálculo del coeficiente *kappa* que explicaremos a continuación, así como la percepción del usuario con respecto a la satisfacción en el uso del sistema. La combinación de estos tres elementos da sustento a la metodología de evaluación que se explicará hacia el final del capítulo.

3.2. El coeficiente *kappa*

La medición constituye una actividad presente en nuestra vida diaria, ya sea de manera muy práctica o más compleja, como en la investigación; como ejemplos de medición tenemos actividades relativamente simples como el registro del peso, la temperatura o

de la presión arterial, hasta actividades más sofisticadas como la evaluación de la calidad de vida de un enfermo. En el entorno de un laboratorio, la medición se ocupa generalmente de fenómenos objetivos. En este contexto, la evaluación del error de medición se realiza a partir de pautas perfectamente estandarizadas, que incluyen la comprobación de dos propiedades de la medición: su reproducibilidad y su exactitud.

En algunos ámbitos y sobre todo fuera del laboratorio, las condiciones en las que se realiza la medición no suelen ser perfectamente controlables; el proceso de medición se puede ver influenciado por la subjetividad del observador y con frecuencia no existe un patrón de referencia aceptable para comprobar la exactitud de la medición. La calidad de las mediciones condiciona no sólo la calidad de la investigación, sino también la calidad de las decisiones que se apoyan en dichas mediciones. El concepto de fiabilidad y los diversos índices usados para estimarla se comprenden mejor si se hace explícito el modelo de medida utilizado.

El coeficiente *kappa* fue propuesto originalmente por Jacob Cohen en 1960 para medir el grado de acuerdo entre dos métodos o dos evaluadores que clasifican el resultado de una observación. Este método fue generalizado para el caso de más de dos evaluadores, por Fleiss, por lo que a veces también se habla del índice *kappa* de Fleiss.

El caso más sencillo se presenta cuando la variable cualitativa es dicotómica, y se comparan dos métodos de clasificación. Esta situación se puede representar en una tabla de frecuencias:

Tabla 3.1. Ejemplo de concordancias

		Método B		
		Positivo	Negativo	
Método A	Positivo	a	cd	f1
	Negativo	b		f2
		c1	c2	n

Fuente: (Molinero 2001)

Las respuestas que indican concordancia son las que se sitúan sobre la diagonal principal; si un dato se sitúa sobre dicha diagonal, ello significa que ambos observadores han clasificado el ítem en la misma categoría del sistema de clasificación. De esta observación surge naturalmente la más simple de las medidas de concordancia, la suma de las probabilidades a lo largo de la diagonal principal frente al total de sujetos: $(a + d) / n$.

En un inicio, este índice fue propuesto como una medida de concordancia por su sencillez; sin embargo, su interpretación no estuvo exenta de problemas al momento de interpretar las concordancias. Resulta que aunque no existiera ninguna relación entre los dos métodos de clasificación, se podría encontrar algún grado de concordancia entre ellos por puro azar. Así, si el método A consiste en clasificar al paciente con resultado positivo, si al lanzar una moneda al aire cae un águila, y negativo en el caso contrario, y hacemos lo mismo en el método B, es previsible encontrar en promedio un 50 % de coincidencias. Aun si supusiéramos que el sistema A es un método científico de diagnóstico y el método B es la opinión de un vidente, también es previsible encontrar un cierto grado de concordancia debido al azar. Con el fin de determinar hasta qué punto la concordancia observada es superior a la que es esperable obtener por puro azar, se definió el índice de concordancia *kappa*.

El coeficiente *kappa* mide el acuerdo existente entre una serie de observadores, corrigiendo aquellos acuerdos que puedan darse por azar. En virtud de que tradicionalmente la variabilidad entre observadores se ha reconocido como una fuente importante de error, es importante estimar el grado de dicha variabilidad y el coeficiente *kappa* refleja y permite calcular la concordancia, en este caso de dos observadores. (Aun cuando hemos hablado que existe la evaluación de concordancia para tres o más observadores, para lo cual se utiliza el coeficiente *kappa* de Fleiss, la metodología de evaluación que utilizamos en esta tesis utiliza la definición que toma como base dos observadores). El coeficiente *kappa* representa una aportación en esta

dirección, básicamente mediante la incorporación en su fórmula de una corrección que excluye la concordancia debida exclusivamente al azar.

Este indicador se construye con base en un cociente, el cual incluye en su numerador la diferencia entre la sumatoria de las concordancias observadas y la sumatoria de las concordancias atribuibles al azar, mientras que su denominador incluye la diferencia entre el total de observaciones y la sumatoria de las concordancias atribuibles al azar. En términos conceptuales, la fórmula del coeficiente *kappa* puede expresarse de la siguiente forma:

$$K = \frac{\text{Suma de \% de concordancias observadas} - \text{Suma de \% de concordancias por azar}}{\text{Total de observaciones} - \text{Suma de \% de concordancias por azar}}$$

Fuente: **(Cerde and Villarroel del P. 2008)**

Como fórmula el coeficiente *kappa* se expresa de la siguiente manera:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

En donde P(A) es la proporción de concordancia observada y P(E) es la proporción de concordancia esperada por puro azar.

El coeficiente *kappa* puede tomar valores entre -1 y +1. En caso de acuerdo perfecto la proporción de concordancia será 1, por lo que 1-P(E) representa el margen de acuerdo posible no atribuible al azar. Mientras más cercano a +1, mayor es el grado de concordancia entre los observadores, por el contrario, mientras más cercano a -1, mayor es el grado de discordancia entre ellos. Un valor de $\kappa = 0$ refleja que la concordancia observada es precisamente la que se espera a causa exclusivamente del azar.

Sobre los niveles aceptables de acuerdo existen también diferencias; sin embargo, a la hora de interpretar los valores de k es útil disponer de una escala como la siguiente, la cual, a pesar de su arbitrariedad, se toma como aceptable.

Tabla 3.2. Valoración del Índice Kappa

Valor de k	Fuerza de la concordancia
< 0.20	Pobre
0.21 – 0.40	Débil
0.41 – 0.60	Moderada
0.61 – 0.80	Buena
0.81 – 1.00	Muy buena

Fuente: (López de Ullibarri Galparsoro and Pita Fernández 2004)

En el caso de la lingüística computacional y las ciencias cognitivas, Carletta propuso en 1996 utilizar el coeficiente *kappa*, tomando como base la experiencia y los resultados que los investigadores en el campo del análisis de contenidos obtuvieron mediante su utilización y por lo que extendió su uso a la evaluación de sistemas de diálogo. La metodología de evaluación Paradise, propone el uso del coeficiente *kappa* a partir de la construcción de una matriz que resume qué tan bien logran el usuario y el agente intercambiar la información necesaria para concretar una tarea en particular. Utilizaremos el ejemplo de la tabla siguiente, tomada del mismo ejemplo que citan Walker et al (Walker, et al. 1997, 273) en la presentación de la metodología. En este ejemplo observamos el uso de la matriz de frecuencias para resumir 100 interacciones de usuarios hipotéticos que utilizaron dos versiones distintas de una aplicación que brinda a los usuarios de una estación de tren en Italia horarios de salida para los trenes que viajan desde y hacia las ciudades de Milán, Turín, Roma y Trento.

Los valores de la matriz se basan en una comparación entre el conjunto de expectativas y el desarrollo efectivo del diálogo en el escenario evaluado. Las columnas representan los valores que se espera intercambien usuario y agente: la información de las ciudades

de salida y destino, el rango de salida, así como la hora efectiva de salida del tren. Las filas integran la información reunida efectivamente durante las interacciones.

Siempre que un valor obtenido de un diálogo coincide con alguna de las expectativas, el número en la diagonal correspondiente en el centro de la matriz se incrementa en 1. Las celdas fuera de la diagonal representan los “malos entendidos” que no hayan sido corregidos durante el diálogo.

La utilización del coeficiente *kappa* para determinar el acuerdo entre los observadores, normaliza el monto de acuerdo esperado por azar y permite unificar la medida entre varios observadores. Esta medida ha mostrado ventajas sobre otras herramientas de evaluación, por lo que ha sido ampliamente usada en el campo del análisis de contenidos. Además de ser fácilmente interpretable permite comparar resultados diferentes y sugiere resultados que han resultado útiles para los propósitos requeridos en la evaluación. A continuación veremos el uso que se da a la matriz y al coeficiente *kappa* en el marco de la evaluación general.

Tabla : Matriz de confusión. Agente A

Datos	Datos llave													
	Ciudad salida				Ciudad destino				Rango salida		Hora de salida			
	Milán	Roma	Turín	Trento	Milán	Roma	Turín	Trento	Mañana	Tarde	6am	8am	6pm	8pm
Milán	22		1		3									
Roma		29												
Turín	4		16	4			1							
Trento	1	1	5	11			1							
Milán	3				20									
Roma						22								
Turín			2		1	1	20	5						
Trento			1		1	2	8	15						
Mañana									45	10				
Tarde									5	40				
6am											20		2	
8am											1	19	2	4
6pm											2		18	
8pm											2	6	3	21
Suma	30	30	25	15	25	25	30	20	50	50	25	25	25	25

Tabla : Matriz de confusión. Agente B

Datos	Datos llave													
	Ciudad salida				Ciudad destino				Rango salida		Hora de salida			
	Milán	Roma	Turín	Trento	Milán	Roma	Turín	Trento	Mañana	Tarde	6am	8am	6pm	8pm
Milán	16		1		4				3	2				
Roma	1	20	1			3								
Turín	5	1	9	4	2		4	2						
Trento	1	2	6	6			2	3						
Milán	4				15				2	3				
Roma	1	6				19								
Turín			5	2	1	1	15	4						
Trento		1	3	3	1	2	9	11						
Mañana	2				2				39	10				
Tarde									6	35				
6am											20	5	5	4
8am												10	5	5
6pm											5	5	10	5
8pm												5	5	11
Suma	30	30	25	15	25	25	30	20	50	50	25	25	25	25

3.3. Paradise como metodología integral para la evaluación.

En 1997 Marilyn A. Walker, Diane J. Litman, Candace A. Kamm y Alicia Abella presentaron una metodología de evaluación para sistemas de diálogo (Walker, et al. 1997). Llamaron a esta metodología Paradise (Paradigm for Dialogue System Evaluation), como una propuesta integral de evaluación necesaria en un mundo en el que los avances en el reconocimiento del habla y el procesamiento del lenguaje natural han permitido el desarrollo de agentes conversacionales para una gran variedad de fines, pero en el que, aseveraban, existe una carencia importante de metodologías de evaluación que permitan comparar el desempeño de diferentes agentes de diálogo, así como la posibilidad de generalizar los resultados. Paradise se presenta como un marco general de evaluación que permite no sólo la comparación entre diferentes versiones del mismo sistema, sino entre diferentes sistemas en distintos o en el mismo campo. Aseguran que permite desarrollar modelos predictivos de usabilidad a través de una función que toma en cuenta las propiedades principales del sistema y permite generalizar los resultados determinando los factores que tienen un mayor impacto en la usabilidad.

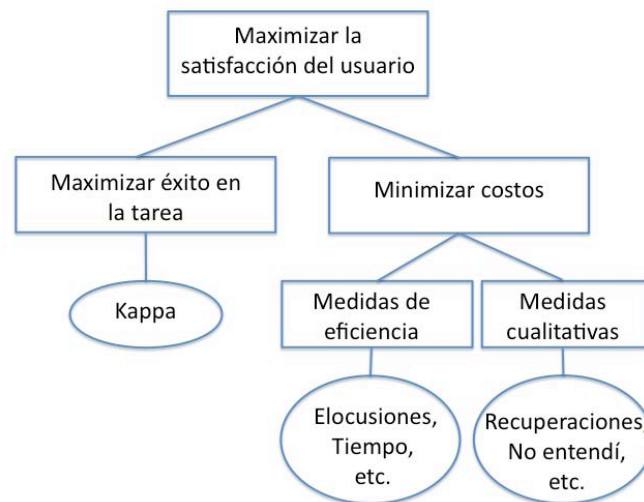
La metodología de evaluación que Paradise propone, parte de la idea de que el desempeño de una aplicación es resultado de una función de tres elementos:

- el éxito en la realización de la tarea específica
- los costos involucrados en la ejecución del diálogo
- la percepción del usuario con respecto a la satisfacción en el uso del sistema

Para llevar a cabo la combinación de estos elementos, Paradise utiliza métodos de la teoría de las decisiones la cual requiere determinar: a) los objetivos del problema de decisión y b) una serie de atributos medibles para hacer operativos dichos objetivos.

En este sentido, Paradise ubica la maximización de la satisfacción del usuario como el objetivo general a alcanzar por los sistemas de diálogo (hablado), equiparándola a un concepto general de usabilidad. Dos objetivos secundarios contribuyen al logro del objetivo general; maximizar el éxito en el cumplimiento de la tarea y minimizar los costos involucrados, como puede observarse en la imagen siguiente.

Gráfica 3.1. Paradise: Estructura de objetivos para la evaluación de sistema de diálogo



Además de los objetivos primario y secundarios, la gráfica 3.1. muestra en los óvalos las mediciones que se utilizarán para hacer operativa la cuantificación del cumplimiento de

estos objetivos, mediciones que explicaremos a continuación. Una vez determinados una serie de cálculos parciales, la autora sugiere la utilización de una función de desempeño, la cual resume en un solo número la medición de desempeño.

3.3.1. Matriz de atributos

La matriz de atributos tiene como fin resumir en qué consiste la tarea y cómo debe llevarse a cabo. La matriz se construye, por una parte, con el tipo de información que deben intercambiar el usuario y el sistema durante el dialogo, así como los valores posibles. Ejemplifica la construcción de la matriz con los dos diálogos hipotéticos que hemos mencionado que fueron utilizados por Morena Danieli y Elisabetta Gerbino en un sistema que brinda información sobre la salida de trenes en hora pico entre cuatro ciudades (Los diálogos completos pueden consultarse en el apéndice). La tarea consiste en que durante el diálogo el agente obtenga del usuario tres valores: la ciudad de salida (DC), la ciudad de destino (AC) y el rango para determinar la hora de salida (DR) para dar al usuario la información que solicita, que es la hora de salida del tren de interés (DT). La matriz integra los atributos, sus valores, junto con la dirección del flujo de información que resumen la tarea a realizar.

Tabla 3.5 Matriz de atributos simplificada de una aplicación de horarios de trenes

Atributo	Valores posibles	Flujo de información
Ciudad de salida DC	Milán, Roma, Turín, Trento	Al agente
Ciudad de llegada AC	Milán, Roma, Turín, Trento	Al agente
Rango de salida DR	Mañana, Tarde	Al agente
Hora de salida DT	6am, 8am, 6pm, 8pm	Al usuario

Esta matriz es la misma para los diálogos que se ejemplifican, lo que confirma que la representación de la tarea es independiente de la estrategia de diálogo utilizada y es

indispensable para la medición del siguiente indicador, que es el del éxito en el cumplimiento de la tarea.

3.3.2. Éxito en la tarea

El éxito en el cumplimiento de la tarea mide la culminación efectiva en el intercambio de información que el agente y el usuario logran al finalizar el diálogo. Para su medición Paradise propone, como una aportación novedosa, la adaptación del coeficiente *kappa*, utilizado originalmente para cuantificar el grado de acuerdo entre dos observadores y corregir este resultado determinando un factor de acuerdo que puede atribuirse al azar. El resultado del cálculo genera un indicador del éxito en el cumplimiento de una tarea.

El indicador se calcula a partir de una matriz de frecuencias, también llamada matriz de confusión, que se construye a partir de la matriz de atributos. La matriz de frecuencias resume el intercambio de información necesaria que logran el usuario y el agente para concretar una tarea en particular.

Los valores de la matriz se basan en una comparación entre el conjunto de expectativas y el desarrollo efectivo del diálogo en el escenario evaluado, en donde los valores esperados se atribuyen a un observador experto, y las observaciones generadas en la interacción, a un observador común. Es decir, los valores que se espera intercambien usuario y agente en un escenario particular se ubican en las columnas, y en las filas, la información reunida efectivamente durante las interacciones. Las coincidencias se ubican en la diagonal principal, y los errores, fuera de ésta. Es importante recordar que aquellos intercambios con errores que logren corregirse, se ubicarán en el centro de la diagonal, ya que representan un intercambio exitoso de información.

A este esquema de anotación le hicimos dos adecuaciones que representan de mejor manera lo que ocurre en las interacciones de “Adivina la Carta”. Para ello vale la pena recordar que el desarrollo de aplicaciones en el DCC del IIMAS, está motivado por la idea de imitar, en la medida de lo posible, la conversación tal y como la mantienen los seres

humanos, por ello, si bien en un contexto determinado, motivan la utilización de un lenguaje variado para dar al usuario la facilidad de expresarse libremente durante el diálogo. Sin embargo, existen situaciones que no considera la metodología; por una parte, cuando el usuario utiliza palabras que no están en la base de conocimiento de Golem; y por otra, cuando el usuario decide no hablar.

De acuerdo con lo previsto en el desarrollo, si el sistema no detecta ningún sonido o si les es imposible asignar una representación semántica al sonido, debe iniciar el procedimiento de recuperación y requerir al usuario repita su respuesta hasta en cuatro ocasiones, respetando el número de preguntas realizadas. Cualquier comportamiento distinto se contabilizó como un error, si el sistema respondió como estaba previsto, anotamos una respuesta correcta. De esta manera, cuando así ocurrió en el intercambio, agregamos a las matrices las columnas "N.B." y "-" para los casos correspondientes a las expectativas de las palabras fuera de la base de datos y a los silencios de los participantes.

Por otra parte, aunque Paradise está previsto para sistemas de diálogo que utilizan únicamente el lenguaje, integramos los actos de visión a la matriz, para registrar los actos de visión y las incidencias de errores, lo cual fue posible debido a que para la arquitectura cognitiva utilizada, los actos son independientes de la modalidad.

Una vez hecha esta anotación, podemos seguir con el ejemplo de Morena y Gerbino citado anteriormente para medir la proporción de veces que la información del diálogo coincide con las expectativas $P(A)$ y la proporción de veces que se espera que la información del diálogo y las expectativas coincidan por azar $P(E)$ para el cálculo del coeficiente *kappa* con la fórmula mencionada:

Una vez realizados los cálculos con la información de estas dos tablas, $P(E)$ es la misma para los dos agentes; 0.079. Para el agente A $P(A) = 0.795$ y $k = 0.777$; mientras que para el agente B $P(A) = 0.59$ y $k = 0.555$, por lo que el agente A permite cumplir con mayor éxito el objetivo que el diálogo del agente B.

3.3.3. Medición de los costos del diálogo y estimación de la función de desempeño

De acuerdo con la figura 3.1. el desempeño es resultado de una función que depende de los costos del diálogo. El cálculo de los costos involucrados en cada diálogo considera distintas medidas de eficiencia, así como medidas relacionadas con la calidad. Para determinar la contribución que cada una de estas medidas tiene con el desempeño, es necesario combinar el resultado obtenido del cálculo del coeficiente k con la serie de resultados, para estimar una función de desempeño. Las medidas que tradicionalmente se han utilizado, incluyen medidas tales como el número de turnos, el tiempo transcurrido para cumplir la tarea o el número de reparaciones durante el diálogo.

Específicamente para el cálculo de costos involucrados, en nuestra evaluación elegimos aquellas mencionadas en la sección 3.1. *Técnicas y herramientas de evaluación de los sistemas de diálogo*. Se trata de medidas obtenidas en parte, directamente del registro del comportamiento y uso del sistema que se genera cada vez que se utiliza una aplicación, y por otra, a partir de la transcripción de las elocuciones de los usuarios.

La satisfacción del usuario ocupa el lugar más importante en la evaluación por lo que el resultado de esta medición es indispensable. Comúnmente el índice de satisfacción se obtiene de la aplicación de cuestionarios a los usuarios y, de acuerdo con la metodología, puede tomarse el resultado de una de las preguntas más relevantes o el promedio general. De acuerdo con la metodología que utilizamos, el cuestionario debe contener preguntas relativas a ocho rubros importantes:

- a) El desempeño del sintetizador de voz: ¿Es fácil comprender lo que dice el sistema?
- b) El desempeño del reconocimiento de voz: ¿El sistema entendió lo que tú decías?
- c) La facilidad de la tarea: ¿Fue fácil terminar la tarea?
- d) El ritmo de la interacción: ¿Fue apropiado el ritmo de la interacción?
- e) La experiencia del usuario: ¿Sabías qué hacer durante el diálogo?
- f) La respuesta del sistema: ¿Fue adecuada la velocidad de respuesta del sistema?
- g) El comportamiento esperado ¿El sistema funcionó como lo esperabas?
- h) El uso futuro: ¿Volverías a interactuar con el sistema en el futuro?

Para nuestro ejercicio de evaluación hicimos también algunos ajustes al cuestionario. En primer lugar adaptamos el cuestionario de evaluación, para hacer más claras las preguntas para los niños y para nosotros, con el fin de obtener la respuesta buscada. Es decir ajustamos preguntas como “¿Es fácil terminar la tarea?” por “¿Qué tan fácil o difícil crees que es / sea adivinar la carta?”.

Precisamente este rubro nos pareció tan importante, que en lugar de una utilizamos tres preguntas para conocer dónde podía estar la dificultad de los niños para concluir la tarea, así que además de “¿Qué tan fácil o difícil crees que es / sea adivinar la carta?” incluimos “¿Tuviste tiempo suficiente para pensar las preguntas?” y “¿En general, qué tan fácil o difícil fue terminar el juego?”.

A los rubros que requiere la metodología de Paradise agregamos una pregunta que surgió de manera espontánea en las evaluaciones preliminares, ya que a una pregunta abierta sobre cualquier comentario que quisieran hacer los niños, comentaron que se habían sentido desesperados, cuando nuestro sistema presentó más fallas en el reconocimiento; o nerviosos, sobre todo cuando los amigos o los familiares o desconocidos los observaban desde afuera.

Por último, hicimos algunos cambios en la escala de evaluación. Sobre la escala encontramos diversas modalidades en las evaluaciones revisadas. Algunas utilizan escalas numéricas del uno al ocho o del uno al cinco; otros utilizan escalas de Likert a las que después asignan un valor numérico (casi nunca = 1, pocas veces = 2, algunas veces = 3, seguido = 4, casi siempre = 5). Nosotros tomamos una escala de Likert de cuatro valores: dos positivos y dos negativos, para eliminar los valores neutrales que muchas veces deben eliminarse al momento de realizar los conteos. A la escala asignamos valores que premian el buen desempeño y castigan un mal desempeño, quedando la escala de la siguiente manera:

Todas las veces	Muchas veces	Pocas veces	Nunca
100	66	33	0

Junto con esta escala de medición utilizamos en dos preguntas una respuesta que requería una respuesta terminante: sí o no, y en su caso, el usuario tenía oportunidad de aclarar las causas de su respuesta. El cuestionario aplicado quedó finalmente como muestra la Tabla 3.6.

Una vez aplicado el cuestionario puede calcularse un índice general de satisfacción a través del promedio de las respuestas obtenidas o el resultado de una de las preguntas en particular. Nosotros utilizamos el promedio de las respuestas obtenidas, excluyendo la pregunta relativa a su sentir durante el juego, ya que esa pregunta no se refiere al desempeño de la aplicación.

3.3.4. La función de desempeño

De acuerdo con la metodología, cuando el evaluador conoce tanto las distintas medidas de desempeño objetivas y el resultado obtenido de la aplicación del cuestionario de satisfacción, debe calcular la contribución de cada uno de estos factores a la satisfacción del usuario, para lo cual se utiliza la regresión lineal múltiple con el fin de determinar el peso que describe la contribución relativa de cada factor en la satisfacción del usuario.

Sin embargo, en virtud de que cada una de las medidas son distintas y tienen escalas diferentes, es necesario normalizar tanto los resultados obtenidos para k como para cada uno de los resultados. La función para normalizar cada uno de los factores X es la siguiente:

$$N_{(x)} = \frac{x - \bar{x}}{\sigma_x}$$

Tabla 3.6. Adivina la carta: Cuestionario de evaluación

Propósito: Evaluar la adecuada estructura de interacción del sistema; en particular, el modelo de diálogo, la efectividad del reconocedor de voz, reconocedor visual y la estructura de la conversación, incluyendo acciones de confirmación y reparación.

Fecha de evaluación: 30-06-2010 al 01-07-2010.

Participantes: 30 niños y niñas entre 10 y 16 años (15/15).

Tipo de evaluación: Cualitativa.

Pregunta	Mide	Pregunta	Mide																				
<p>¿Entendiste lo que Golem decía?</p> <p>Todas las veces Muchas veces Pocas veces Nunca N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Facilidad para entender al robot	<p>¿Por qué? Preguntar específicamente razones de satisfacción o insatisfacción.</p> <p>Muy lento Poco lento Poco rápido Muy rápido N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											
<p>¿Golem entendió lo que tú decías?</p> <p>Todas las veces Muchas veces Pocas veces Nunca N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Entendimiento del sistema	<p>8 ¿Te gustó la velocidad de respuesta de Golem?</p> <p>Sí No N.C.</p> <table border="1"> <tr><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td></tr> </table>							Respuesta del sistema				
<p>¿Sabías qué preguntarle a Golem?</p> <p>Siempre Muchas veces Pocas veces Nunca</p> <table border="1"> <tr><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td></tr> </table>									Experiencia del usuario	<p>¿Por qué? ¿Específicamente, qué fue lo que te gustó o no te gustó?</p> <p>Muy lento Un poco lento Un poco rápido Muy rápido N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>													
<p>¿Tuviste tiempo suficiente para pensar las preguntas?</p> <p>Todas las veces Muchas veces Pocas veces Nunca N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Facilidad para terminar la tarea	<p>9 ¿El sistema funcionó como te lo imaginabas?</p> <p>Mucho mejor Mejor Peor Mucho peor N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Comportamiento esperado
<p>¿En general, qué tan difícil o fácil fue terminar el juego?</p> <p>Muy fácil Algo fácil Algo difícil Muy difícil N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Facilidad para terminar la tarea	<p>10 ¿Volverías a jugar con Golem?</p> <p>Seguramente sí Probablemente sí Probablemente no Seguramente no N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Uso futuro
<p>¿Y qué tan difícil o fácil crees que es/ sea adivinar la carta?</p> <p>Muy fácil Algo fácil Algo difícil Muy difícil N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											Facilidad para terminar la tarea	<p>11 ¿Cómo te sentiste en tu la interacción con Golem?</p> <p>Tranquilo Desesperado Nervioso N.C.</p> <table border="1"> <tr><td></td><td></td><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td><td></td><td></td></tr> </table>											
<p>¿Te gustó el ritmo del juego?</p> <p>Sí No N.C.</p> <table border="1"> <tr><td></td><td></td><td></td></tr> <tr><td></td><td></td><td></td></tr> </table>							Ritmo del juego																

Antes de entrar de lleno al cálculo de la función de desempeño se recomienda incluir preferentemente los factores de más peso y eliminar aquéllos que puedan ser redundantes, por lo que se recomienda llevar a cabo una segunda regresión para obtener tanto el ponderador de k como el factor más significativo en la satisfacción del usuario.

En el cálculo de la regresión lineal múltiple, la satisfacción del usuario es tratada como la variable dependiente y los demás factores como las variables independientes. Del resultado será posible determinar los factores que contribuyen en mayor medida en la de la satisfacción del usuario.

$$P = (\alpha * N(k)) - \sum_{i=1}^n w_i * N(c_i)$$

En el capítulo siguiente presentaremos un ejemplo de la aplicación de la metodología tal y como se llevó a cabo durante la evaluación de “Golem – Adivina la Carta”.

CAPÍTULO 4.

EVALUACIÓN DEL JUEGO “ADIVINA LA CARTA”

La evaluación del juego “Adivina la Carta” consiste en hacer que diversos usuarios interactúen con el sistema, con el fin de integrar información que permita hacer un juicio sobre el desempeño de la aplicación. El sistema ha sido objeto de evaluaciones periódicas y de supervisión permanente las cuales han sido útiles no sólo en su objetivo central de probar la aplicación, sino también para probar las herramientas que facilitan la recuperación de datos, así como los dispositivos que conforman el sistema, los cuales se describieron en el capítulo anterior. La evaluación se llevó a cabo del 30 de junio hasta el 2 de julio de 2010.

Este capítulo hace una breve descripción del estudio de campo y del proceso de levantamiento de información tal y como se describe en la metodología aplicada, que se describe en el capítulo anterior, desde la definición de la tarea hasta el resultado final de la función general de evaluación: Iniciamos este capítulo con la definición de la tarea y la descripción del escenario de la evaluación, seguidamente presentamos la matriz de atributos que resume la representación de la tarea, así como el cálculo del éxito en su

cumplimiento. Posteriormente damos a conocer los resultados de la evaluación de la satisfacción por parte de los usuarios, así como los resultados obtenidos a partir del registro automático de datos. Antes de pasar a la función general de desempeño es menester realizar un paso previo y de gran importancia que consiste en la determinación de los factores que inciden en mayor medida en el resultado de satisfacción del usuario, identificado como el más importante para la evaluación, para cerrar finalmente con el cálculo final de la función de desempeño.

4.1. Diseño del ejercicio de evaluación

El módulo de “Golem – Adivina la carta” está ubicado en el Museo de las Ciencias de la Universidad Nacional Autónoma de México. El módulo utilizado para la realización del estudio es el mismo que se ha utilizado desde su inauguración en diciembre de 2009 y desde entonces forma parte de la exhibición permanente al público.

El módulo está formado por una cabina con un relativo aislamiento de ruido externo. En ella está todo el equipo necesario para el juego: una mesa, un monitor conectado a una computadora personal, un micrófono y diez cartas con motivos astronómicos, así como una cámara y un par de bocinas para escuchar al robot. En el módulo hay lugar para el jugador y para un becario que asiste a los jugadores durante la interacción en el módulo. Es importante indicar que la evaluación que se describe en este capítulo, así como las evaluaciones anteriores que se mencionan, se llevaron a cabo fuera del laboratorio con usuarios reales.

Como parte de la experiencia obtenida de las evaluaciones y del monitoreo del sistema, se instrumentó el registro automático de uso y funcionamiento del sistema para reconstruir la historia a partir de los indicadores objetivos de la evaluación, asimismo se decidió guardar un registro digital de las elocuciones de cada participante, útil sobre todo para valorar el nivel de reconocimiento. Esta información ha sido utilizada además

para la actualización del modelo del lenguaje. Los registros se generaron desde abril y ahora son una herramienta permanente de conocimiento del sistema.

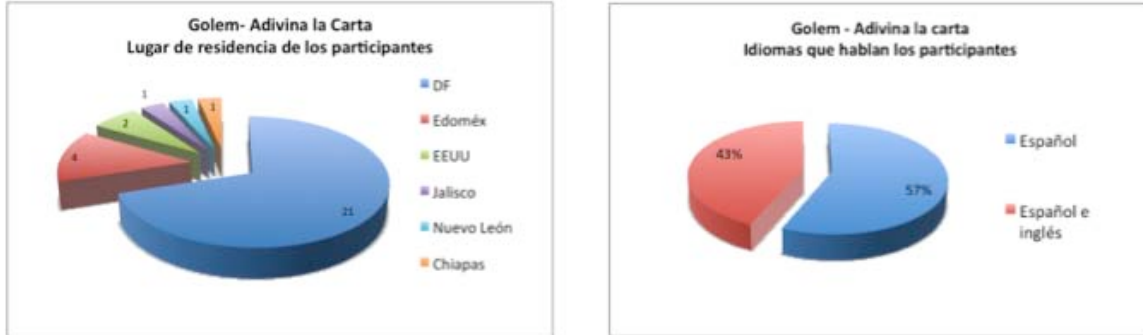
Asimismo, se instaló una videocámara con el fin de grabar las interacciones de manera general. Por ahora, este material está pensado únicamente como evidencia de las evaluaciones realizadas y como respaldo en caso de que haya dificultad en entender la elocución por parte del usuario; sin embargo, en un futuro puede ser útil para conocer la actitud del usuario frente al sistema y durante el diálogo. Por otra parte, utilizamos la bitácora manual para el registro de otros eventos relevantes que surgieran durante la evaluación, así como un cuestionario de evaluación en papel que el usuario respondió al personal del equipo evaluador.

El equipo evaluador estuvo conformado por tres personas responsables de: a) acompañar al usuario en la interacción dentro del módulo, llevar la bitácora manual y videograbar cada una de las interacciones; b) requisitar el registro y permiso de participación correspondiente del padre o tutor que acompañara al menor y hablar, en su caso, con los padres de los menores cuando así fue necesario; c) invitar a los usuarios a participar y levantar la encuesta de satisfacción una vez concluida la interacción.

Ninguno de los jóvenes evaluados contó con experiencia previa en la interacción con sistemas de diálogo. Cada uno fue informado en el momento de la invitación sobre la evaluación que estábamos llevando a cabo y recibió las instrucciones únicamente antes de iniciar la interacción. Fuera de lograr el número previsto de evaluados y la distribución por sexo no hubo otro tipo de selección en el público objetivo.

Para la evaluación del juego “Adivina la Carta”, pedimos a 30 visitantes del museo que cumplieran con los requerimientos de edad y género, que interactuaran con el sistema. Para la evaluación se integraron interacciones de 30 usuarios, 15 varones y 15 mujeres, entre 10 y 16 años.

Golem – Adivina la Carta: Residencia e idiomas que hablan los participantes



Como parte del registro de los usuarios evaluados se obtuvo el nombre completo, dirección, sexo, edad, nivel de escolaridad, lugar de nacimiento, lugar de residencia y los idiomas que hablan o que están aprendiendo en la escuela. A todos los participantes que acudieron al museo en compañía del padre o tutor, se les solicitó la autorización de éstos y a quienes que contaban con una identificación se les pidió una copia de la misma.

Los usuarios evaluados fueron los 30 primeros que cubrieron los criterios de elegibilidad, se repitieron únicamente las evaluaciones cuando falló la grabación del video.

4.2. Resultados de la evaluación

4.2.1. Descripción de la tarea

Cada usuario tuvo la oportunidad de jugar una sola vez. En una interacción típica para el niño, el objetivo es adivinar la carta y ganar. Para ello enfrenta varias dificultades; la dificultad principal fue elaborar una serie de preguntas que le permitieran identificar las características más representativas de los objetos representados para deducir la carta que eligió el sistema. Durante la explicación del funcionamiento del juego, el sistema despliega en la pantalla ejemplos de las preguntas que el niño puede hacer; sin embargo, es común que cuando el usuario no sabe qué preguntar o cuando necesita

más tiempo para pensar, simplemente no hable, lo cual queda marcado en el registro de uso del sistema como un error atribuible al reconocimiento.

Cada vez que el niño enfrenta la situación de que el sistema no entienda su respuesta se inicia un proceso de recuperación de información y bastará con que la repita para subsanar el error, siempre y cuando se trate de un error de reconocimiento. Sin embargo, puede ocurrir que el nombre o la característica empleada no estén en el diccionario de la aplicación y, en este caso, el usuario no tendrá forma de saberlo y podrá insistir o cambiar la pregunta para seguir jugando, cuando perciba que el sistema no lo entiende. Las preguntas relativas al nombre y la edad no son críticas para el cumplimiento de la tarea, por ello, si el sistema no entiende el nombre o la edad del participante, lo dejarán continuar el juego.

Por último, el niño tiene la dificultad de enfrentar al público que lo mira desde afuera, ya sea la familia, los amigos de la escuela o la gente que pasa, muchos niños se sienten intimidados o nerviosos de que los vean y saberse escuchados a través de las bocinas ubicadas en el exterior.

4.2.1. Intercambio de información y medición del éxito en el cumplimiento del objetivo

Una vez descrita y acotada la tarea, se procede a identificar la información que el usuario y el agente deben intercambiar durante cada subdiálogo para llegar al final de la conversación. Como hemos visto en los capítulos anteriores, la tarea está organizada en subdiálogos, que constituyen la estructura de la tarea.

La tabla 4.1., muestra el tipo de información que el usuario y el sistema deben intercambiar en cada subdiálogo como condición para pasar al subsecuente y concluir la tarea. En el capítulo 3 observamos que el intercambio de información en estas conversaciones es mutuo; la información que el usuario brinda al robot es útil para concluir la tarea, pero la que el robot proporciona es indispensable no sólo para concluir con éxito la tarea, sino para cumplir con el objetivo del niño de ganar el juego. Sin

embargo, conforme lo establece la metodología, utilizaremos únicamente la información que la aplicación requiere para la conclusión del éxito en la tarea.

La primera columna indica el subdiálogo correspondiente (**n**: Saludo y reconocimiento del nombre; **a**: pregunta y confirmación de la edad; **i**: Explicación de las reglas del juego; **I**: Interrogatorio; **V**: Selección y verificación de la carta; **F**: Despedida), la segunda muestra la dirección de la información, es decir la que el usuario hace de conocimiento de Golem. La tercera columna señala el tipo de información que deben intercambiar y la última el número de valores posibles, considerando el conjunto de imágenes y significados creados con base en el contexto y la tarea específica a solucionar, que se cargan con anticipación en el sistema.

En el subdiálogo **n** Golem debe reconocer el nombre del usuario, para ello cuenta en su base de conocimiento con 622 nombres almacenados. Esta es de las pruebas más difíciles para el sistema. En el subdiálogo **a**, el sistema pide la edad y la confirmación de la edad. Aunque para esta evaluación sólo consideramos niños entre 10 y 16 años, es posible reconocer edades desde 1 a 30 años, aunque los niños menores de 10 años son descartados de manera automática. Para la confirmación de la edad tenemos 15 formas registradas para la afirmación y la negación (Ej. Ok, claro, ajá, nel). El subdiálogo **i** se refiere a las instrucciones del juego; el niño debe decir si requiere la explicación del juego y confirmar que ha comprendido la explicación. La base de conocimiento en estas respuestas es la misma. El subdiálogo **I** se refiere al juego, propiamente dicho. En el cuadro están representadas las cuatro oportunidades con las que cuenta cada participante para deducir la carta que eligió el sistema o, en su caso, para adivinar de manera anticipada la carta que cree que eligió. En este caso, la base de conocimiento incluye 83 características que describen los motivos que muestran las cartas, así como las 10 cartas.

El subdiálogo **v** presenta aparentemente una dificultad adicional ya que combina una acción de visión con una de habla. La experiencia más amplia en metodologías de evaluación existe en los sistemas de diálogo hablado y se señala que existen problemas

de interpretación cuando los sistemas utilizan múltiples medios de interacción. Sin embargo, esta situación no se presenta en “Adivina la Carta”, ya que el modelo de diálogo espera que la expectativa se cumpla a través de la visión. Una vez cumplida la expectativa es indistinta la forma en que el sistema obtuvo la información. Por ello, durante toda la evaluación, este acto recibirá el mismo tratamiento que los demás que utilizan el lenguaje hablado. En virtud de que el juego es de 10 cartas, son 10 las opciones con las que el sistema cuenta para el reconocimiento de esta información. En el caso de la confirmación aplica lo comentado con anterioridad. Durante la despedida no hay intercambio.

Tabla 4.1. GOLEM: Juego Adivina la carta

Intercambio de información: Diálogo estándar

Sub diálogo	Intercambio de información		Valores posibles
n	Usuario a Golem	Nombre del usuario	Nombre = 622
a	Usuario a Golem	Edad	Edad = 30
	Usuario a Golem	Confirmación	Afirmación + Negación =12
i	Usuario a Golem	sí/no reglas del juego	Afirmación + Negación =12
	Usuario a Golem	Confirmación	Afirmación + Negación =12
l	Usuario a Golem	Características o carta	Características o carta = 83 +10 = 93
	Usuario a Golem	Características o carta	Características o carta = 83 +10 = 93
	Usuario a Golem	Características o carta	Características o carta = 83 +10 = 93
	Usuario a Golem	Características o carta	Características o carta = 83 +10 = 93
v	Usuario a Golem	Muestra carta	Cartas = 10
	Usuario a Golem	Confirmación	Afirmación + Negación =12
F	Usuario a Golem	Despedida	

La identificación de esta información que intercambian el usuario y el sistema, es indispensable para la construcción de la matriz de atributos y posteriormente para el cálculo del indicador que mide el éxito en el cumplimiento de la tarea.

La matriz de atributos que se muestra en la tabla 4.2 resume el intercambio de información de los treinta participantes de la evaluación y su comparación con el escenario de expectativas. La matriz refleja la estructura de la conversación, constituida por subdiálogos e incluye, de acuerdo con la metodología, las acciones de reparación. Es decir que, si después de varios intentos el usuario y el sistema logran intercambiar la información necesaria para concretar la tarea parcial, el intercambio quedará registrado como exitoso en la matriz.

La matriz resume de una manera muy adecuada lo sucedido con los participantes durante la evaluación. Con respecto a la estructura descrita en los capítulos anteriores podemos observar que se trata de la misma que hemos venido revisando en los capítulos anteriores, con excepción de las dos columnas señaladas como v19 y v 21, que corresponden a los intentos de los participantes de terminar de manera anticipada el juego o, eventualmente, a errores en el reconocimiento por algunas frases que no hayan podido ser ubicadas de manera correcta en el contexto; por ejemplo a las preguntas “¿Está cerca del Sol?” o “¿Gira alrededor de la Tierra?” el sistema puede entender que el participante desea adivinar la carta “Sol” o “Tierra”.

Si bien no todos los participantes intentaron adivinar la carta de manera anticipada, no quisimos dejar de reflejar en esta matriz general cómo transcurrieron las conversaciones durante la prueba.

Por otra parte, cabe mencionar que a pesar de que cada subdiálogo cuenta con una base de conocimientos, formada por una serie de imágenes y significados para cada situación, en la evaluación se utilizan únicamente los datos generados durante el intercambio de información.

Observamos también, sin llegar a un análisis más profundo o cuantificable, que conforme avanzaron las preguntas del interrogatorio el nivel de reconocimiento fue cada vez menor. Como comentamos hace algunos párrafos, la base de conocimiento se ajusta en cada subdiálogo, por lo que no es posible que las confusiones se den

únicamente en el contexto de la misma pregunta, cualquier otro malentendido lanza la acción de reparación, solicitando al participante repita lo dicho.

Tabla 4.2. Golem: Adivina la carta - Matriz de confusión y cálculo del coeficiente Kappa
Resumen de resultados

Datos registrados	Datos llave												
	n		a			i		l				v	
	v1	v3	v5	v7	v9	v11	v13	v15	v17	v19	v21	v23	v24
	Nombre del usuario	Edad	Confirma edad	¿Instrucciones ?	Confirma instrucciones	Pregunta1	Pregunta2	Pregunta3	Pregunta4	Confirma terminar	Confirma carta	Muestra carta	¿Confirma?
OK + No base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	
	622	30	15	15	15	110	110	110	110	15	15	10	15
v1	22												
v2 Otro	8												
v3		28											
v4 Otro		2											
v5			24										
v6 Otro			4										
v7				29									
v8 Otro				1									
v9					26								
v10 Otro					2								
v11						25							
v12 Otro						5							
v13							21						
v14 Otro							9						
v15								19					
v16 Otro								8					
v17									13				
v18 Otro									14				
v19										10			
v20 Otro										0			
v21											10		
v22 Otro											0		
v23												25	
v24 Otro												5	
v25													26
v26 Otro													3
Suma	30	30	28	30	28	30	30	27	27	10	10	30	29

Utilizando la fórmula para el cálculo del porcentaje de concordancias observadas y el porcentaje de concordancias atribuibles al azar obtenemos concordancias en un 76.11, un 8.2% de aciertos atribuibles al azar y por lo tanto una k de 0.74.

Este ejercicio y el cálculo correspondiente también se llevó a cabo a nivel de cada uno de los subdiálogos, considerando los intercambios por nombre, edad y cada una de las preguntas del interrogatorio por separado, obteniéndose los resultados que se resumen en la tabla 4.3. siguiente.

Es fácilmente observable que la prueba más sencilla tanto para el usuario como para el sistema es el reconocimiento de la edad. Constituye una respuesta fácil de generar por parte del usuario y con universo de respuestas básicamente limitado. Prácticamente el

Tabla 4.3. Golem - Adivina la Carta

Cálculo del coeficiente Kappa - Subdiálogos

Subdiálogo	Kappa	Aciertos	Acuerdo por azar
ln(n)	0.611	0.63	0.06
ln-(a)	0.917	0.93	0.20
I primera	0.724	0.83	0.11
I segunda	0.663	0.70	0.11
I tercera	0.643	0.67	0.07
I cuarta	0.339	0.41	0.10

sistema reconoció la edad de 28 de los 30 participantes en la evaluación. El nombre que aparentemente era de las respuestas más difíciles, tuvo un mejor desempeño que la cuarta pregunta del interrogatorio. En nuestra opinión aquí hay que sumar los problemas de reconocimiento a los problemas del usuario para plantear las preguntas al sistema.

En el apéndice incluimos la información detallada con los resultados del coeficiente *kappa* para cada uno de los subdiálogos que se resumen en la tabla 4.3, así como el cálculo llevado a cabo para el cálculo del coeficiente para cada uno de los participantes.

4.2.3. Resultados de la satisfacción del usuario

Una vez terminada la interacción, un miembro del equipo evaluador solicitó a cada uno de los participantes que contestaran un cuestionario para recoger la percepción que tuvieron después de interactuar con el sistema. Ningún niño fue obligado a contestar ninguna de las preguntas y los resultados de la evaluación pueden observarse en la tabla 4.4.

Tabla 4.4. Adivina la carta: Resultados de la evaluación

Propósito: Evaluar la adecuada estructura de interacción del sistema; en particular, el modelo de diálogo, la efectividad del reconocedor de voz, reconocedor visual y la estructura de la conversación, incluyendo acciones de confirmación y reparación.

Fecha de evaluación: 30-06-2010 al 01-07-2010.

Participantes: 30 niños y niñas entre 10 y 16 años (15/15).

Tipo de evaluación: Cualitativa.

Pregunta	Mide	Pregunta	Mide																														
<p>¿Entendiste lo que Golem decía?</p> <table border="1"> <thead> <tr> <th>Todas las veces</th> <th>Muchas veces</th> <th>Pocas veces</th> <th>Nunca</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>26</td> <td>2</td> <td>2</td> <td>0</td> <td>0</td> </tr> <tr> <td>86.7</td> <td>6.7</td> <td>6.7</td> <td>0.0</td> <td>0.0</td> </tr> </tbody> </table>	Todas las veces	Muchas veces	Pocas veces	Nunca	N.C.	26	2	2	0	0	86.7	6.7	6.7	0.0	0.0	<p>Facilidad para entender al robot</p>	<p>¿Por qué? Preguntar específicamente razones de satisfacción o insatisfacción.</p> <table border="1"> <thead> <tr> <th>Muy lento</th> <th>Poco lento</th> <th>Poco rápido</th> <th>Muy rápido</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>12</td> <td>6</td> <td>1</td> <td>11</td> </tr> <tr> <td>0.0</td> <td>63.2</td> <td>31.6</td> <td>5.3</td> <td>57.9</td> </tr> </tbody> </table>	Muy lento	Poco lento	Poco rápido	Muy rápido	N.C.	0	12	6	1	11	0.0	63.2	31.6	5.3	57.9	<p>30</p> <p>100</p> <p>19</p> <p>100</p>
Todas las veces	Muchas veces	Pocas veces	Nunca	N.C.																													
26	2	2	0	0																													
86.7	6.7	6.7	0.0	0.0																													
Muy lento	Poco lento	Poco rápido	Muy rápido	N.C.																													
0	12	6	1	11																													
0.0	63.2	31.6	5.3	57.9																													
<p>¿Golem entendió lo que tú decías?</p> <table border="1"> <thead> <tr> <th>Todas las veces</th> <th>Muchas veces</th> <th>Pocas veces</th> <th>Nunca</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>3</td> <td>7</td> <td>20</td> <td>0</td> <td>0</td> </tr> <tr> <td>10.0</td> <td></td> <td>66.7</td> <td>0.0</td> <td>0.0</td> </tr> </tbody> </table>	Todas las veces	Muchas veces	Pocas veces	Nunca	N.C.	3	7	20	0	0	10.0		66.7	0.0	0.0	<p>Entendimiento del sistema</p>	<p>8 ¿Te gustó la velocidad de respuesta de Golem?</p> <table border="1"> <thead> <tr> <th>Sí</th> <th>No</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>16</td> <td>10</td> <td>3</td> </tr> <tr> <td>61.5</td> <td>38.5</td> <td>11.5</td> </tr> </tbody> </table>	Sí	No	N.C.	16	10	3	61.5	38.5	11.5	<p>Respuesta del sistema</p> <p>30</p> <p>77</p> <p>26</p> <p>100</p>						
Todas las veces	Muchas veces	Pocas veces	Nunca	N.C.																													
3	7	20	0	0																													
10.0		66.7	0.0	0.0																													
Sí	No	N.C.																															
16	10	3																															
61.5	38.5	11.5																															
<p>¿Sabías qué preguntar a Golem?</p> <table border="1"> <thead> <tr> <th>Siempre</th> <th>Muchas veces</th> <th>Pocas veces</th> <th>Nunca</th> </tr> </thead> <tbody> <tr> <td>7</td> <td>10</td> <td>13</td> <td>0</td> </tr> <tr> <td>23.3</td> <td>33.3</td> <td>43.3</td> <td>0.0</td> </tr> </tbody> </table>	Siempre	Muchas veces	Pocas veces	Nunca	7	10	13	0	23.3	33.3	43.3	0.0	<p>Experiencia del usuario</p>	<p>¿Por qué? ¿Específicamente, qué fue lo que te gustó o no te gustó?</p> <table border="1"> <thead> <tr> <th>Muy lento</th> <th>Un poco lento</th> <th>Un poco rápido</th> <th>Muy rápido</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>4</td> <td>13</td> <td>3</td> <td>0</td> <td>10</td> </tr> <tr> <td>20.0</td> <td>65.0</td> <td>15.0</td> <td>0.0</td> <td></td> </tr> </tbody> </table>	Muy lento	Un poco lento	Un poco rápido	Muy rápido	N.C.	4	13	3	0	10	20.0	65.0	15.0	0.0		<p>30</p> <p>100</p> <p>20</p> <p>100</p>			
Siempre	Muchas veces	Pocas veces	Nunca																														
7	10	13	0																														
23.3	33.3	43.3	0.0																														
Muy lento	Un poco lento	Un poco rápido	Muy rápido	N.C.																													
4	13	3	0	10																													
20.0	65.0	15.0	0.0																														
<p>¿Tuviste tiempo suficiente para pensar las preguntas?</p> <table border="1"> <thead> <tr> <th>Todas las veces</th> <th>Muchas veces</th> <th>Pocas veces</th> <th>Nunca</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>12</td> <td>8</td> <td>10</td> <td>0</td> <td>0</td> </tr> <tr> <td>40.0</td> <td>26.7</td> <td>33.3</td> <td>0.0</td> <td></td> </tr> </tbody> </table>	Todas las veces	Muchas veces	Pocas veces	Nunca	N.C.	12	8	10	0	0	40.0	26.7	33.3	0.0		<p>Facilidad para terminar la tarea</p>	<p>9 ¿El sistema funcionó como te lo imaginabas?</p> <table border="1"> <thead> <tr> <th>Mucho mejor</th> <th>Mejor</th> <th>Peor</th> <th>Mucho peor</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>26</td> <td>2</td> <td>0</td> <td>1</td> </tr> <tr> <td>3.4</td> <td>89.7</td> <td>6.9</td> <td>0.0</td> <td>3.4</td> </tr> </tbody> </table>	Mucho mejor	Mejor	Peor	Mucho peor	N.C.	1	26	2	0	1	3.4	89.7	6.9	0.0	3.4	<p>Comportamiento esperado</p> <p>30</p> <p>100</p> <p>29</p> <p>100</p>
Todas las veces	Muchas veces	Pocas veces	Nunca	N.C.																													
12	8	10	0	0																													
40.0	26.7	33.3	0.0																														
Mucho mejor	Mejor	Peor	Mucho peor	N.C.																													
1	26	2	0	1																													
3.4	89.7	6.9	0.0	3.4																													
<p>¿En general, qué tan difícil o fácil fue terminar el juego?</p> <table border="1"> <thead> <tr> <th>Muy fácil</th> <th>Algo fácil</th> <th>Algo difícil</th> <th>Muy difícil</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>11</td> <td>8</td> <td>9</td> <td>2</td> <td>0</td> </tr> <tr> <td>36.7</td> <td>26.7</td> <td>30.0</td> <td>6.7</td> <td></td> </tr> </tbody> </table>	Muy fácil	Algo fácil	Algo difícil	Muy difícil	N.C.	11	8	9	2	0	36.7	26.7	30.0	6.7		<p>Facilidad para terminar la tarea</p>	<p>10 ¿Volverías a jugar con Golem?</p> <table border="1"> <thead> <tr> <th>Seguramente sí</th> <th>Probablemente sí</th> <th>Probablemente no</th> <th>Seguramente no</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>19</td> <td>8</td> <td>2</td> <td>1</td> <td>0</td> </tr> <tr> <td>63.3</td> <td>26.7</td> <td>6.7</td> <td>3.3</td> <td>0.0</td> </tr> </tbody> </table>	Seguramente sí	Probablemente sí	Probablemente no	Seguramente no	N.C.	19	8	2	1	0	63.3	26.7	6.7	3.3	0.0	<p>Uso futuro</p> <p>30</p> <p>100</p> <p>30</p> <p>100</p>
Muy fácil	Algo fácil	Algo difícil	Muy difícil	N.C.																													
11	8	9	2	0																													
36.7	26.7	30.0	6.7																														
Seguramente sí	Probablemente sí	Probablemente no	Seguramente no	N.C.																													
19	8	2	1	0																													
63.3	26.7	6.7	3.3	0.0																													
<p>¿Y qué tan difícil o fácil crees que es/ sea adivinar la carta?</p> <table border="1"> <thead> <tr> <th>Muy fácil</th> <th>Algo fácil</th> <th>Algo difícil</th> <th>Muy difícil</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>4</td> <td>9</td> <td>11</td> <td>6</td> <td>0</td> </tr> <tr> <td>13.3</td> <td>30.0</td> <td>36.7</td> <td>20.0</td> <td></td> </tr> </tbody> </table>	Muy fácil	Algo fácil	Algo difícil	Muy difícil	N.C.	4	9	11	6	0	13.3	30.0	36.7	20.0		<p>Facilidad para terminar la tarea</p>	<p>11 ¿Cómo te sentiste en la interacción con Golem?</p> <table border="1"> <thead> <tr> <th>Tranquilo</th> <th>Desesperado</th> <th>Nervioso</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>17</td> <td>3</td> <td>10</td> <td>0</td> </tr> <tr> <td>56.7</td> <td>10.0</td> <td>33.3</td> <td>0.0</td> </tr> </tbody> </table>	Tranquilo	Desesperado	Nervioso	N.C.	17	3	10	0	56.7	10.0	33.3	0.0	<p>30</p> <p>100</p> <p>30</p> <p>100</p>			
Muy fácil	Algo fácil	Algo difícil	Muy difícil	N.C.																													
4	9	11	6	0																													
13.3	30.0	36.7	20.0																														
Tranquilo	Desesperado	Nervioso	N.C.																														
17	3	10	0																														
56.7	10.0	33.3	0.0																														
<p>¿Te gustó el ritmo del juego?</p> <table border="1"> <thead> <tr> <th>Sí</th> <th>No</th> <th>N.C.</th> </tr> </thead> <tbody> <tr> <td>24</td> <td>2</td> <td>4</td> </tr> <tr> <td>92.3</td> <td>7.7</td> <td></td> </tr> </tbody> </table>	Sí	No	N.C.	24	2	4	92.3	7.7		<p>Ritmo del juego</p> <p>26</p> <p>100</p>																							
Sí	No	N.C.																															
24	2	4																															
92.3	7.7																																

Los resultados muestran un desempeño menor en el reconocimiento y, por lo tanto, la necesidad de atender este problema lo antes posible, pero muestran también que para los niños la tarea de adivinar la carta no es sencilla y que una de las principales dificultades consiste en plantear las preguntas al sistema, aunque tengan tiempo suficiente para preguntar.

Entre los resultados alentadores fue conocer que más del 90% de los usuarios estuvo satisfecho con el ritmo de juego de la interacción, que 90% probablemente volvería a jugar con Golem, que 86% entendió todas las veces lo que el robot decía, que más de la mitad se sintió tranquilo durante la interacción y que 93.1 de los evaluados consideraron que el sistema funcionó mejor o mucho mejor de lo que esperaban.

Una vez obtenidos estos resultados procedimos a ponderarlos según el valor aplicable; Todas las veces / Mucho mejor = 100 Muchas veces / Mejor = 66 Pocas veces / Peor 33 y Nunca / Mucho peor 0. A nivel general el resultado de la evaluación da una calificación de 67.78.

La escala de evaluación es estricta, pero lo importante es que intenta ser equitativa bajo el principio de castigar en mayor medida los malos resultados y premiar los buenos. Sea cual fuere la escala numérica aplicada se trata de un primer referente, por lo que será necesario esperar los resultados de futuras evaluaciones para poder comparar los resultados y verificar alguna mejora.

La tabla 4.5. muestra un resumen de las calificaciones ordenadas de acuerdo con el criterio de evaluación. Los resultados de la evaluación se generaron también a nivel de cada uno de los participantes y fueron un insumo importante para la determinación de los factores que más inciden en la satisfacción del usuario.

Tabla 4.5. Golem - Adivina la Carta Resultados de satisfacción por pregunta	
	Calificación
Facilidad para entender al robot	93.27
Entendimiento del sistema	47.40
Experiencia del usuario	59.63
Facilidad para terminar la tarea (¿Sabías qué preguntar?)	68.60
Facilidad para terminar la tarea (¿Tuviste tiempo suficiente para pensar las preguntas?)	64.17
Facilidad para terminar la tarea (¿En general, qué tan difícil o fácil fue terminar el juego?)	45.23
Ritmo del juego	92.31
Respuesta del sistema	59.19
Comportamiento esperado	64.90
Uso futuro	83.13

4.2.4. Costos del diálogo: medidas de eficiencia y efectividad

Los resultados de la tabla 4.6. resumen los resultados de un reporte generado a partir de los datos del registro automático de interacciones, que facilita la integración de información de cada uno de los usuarios del sistema y, en este caso, específicamente de los participantes en la evaluación. La información completa del reporte puede consultarse también en el apéndice.

Como hemos mencionado con anterioridad, parte de la información se genera directamente de los registros de uso y comportamiento del sistema y otra parte combina esta información con la transcripción de las elocuciones generadas por los usuarios. La idea de generar reportes automáticos o semi automáticos tiene la finalidad de facilitar el monitoreo y evaluación una aplicación que se encuentra en exhibición permanente ante el público, generando información cuantificable para calificar el desempeño o el resultado de la aplicación de mejoras en el sistema.

Como puede observarse, de los 30 usuarios que participaron en la evaluación, todos concluyeron con la tarea encomendada, terminar el juego. De los treinta, sólo 9 lograron adivinar la carta, lo cual confirma que ésta no resulta siempre fácil para todos los participantes.

Tabla 4.6. Gólem - Adivina la Carta

Resumen de información del registro automático de interacciones

Total diálogos:	30
Total terminaron el juego:	30
Total despedida con nombre:	29
Total juegos ganados:	9
Total juegos perdidos:	21
% error en reconocimiento por palabra:	0.59
% acierto en reconocimiento por palabra:	0.41
Máxima duración de un diálogo:	00:06:53
Mínima duración de un diálogo:	00:03:32
Máximo número de elocuciones (usuario):	27
Mínimo número de elocuciones (usuario):	10
Máximo número de elocuciones (sistema):	53
Mínimo número de elocuciones (sistema):	33
Promedio número de acciones de visión:	1.73
Máximo número de acciones de visión:	5
Mínimo número de acciones de visión:	1
% Reconocimiento por diálogo:	0.57
% Máximo de reconocimiento por diálogo:	1
% Mínimo de reconocimiento por diálogo:	0.35
Máximo número sin representación semántica:	10
Mínimo número sin representación semántica:	0
Máximo número sin detección de sonido:	5
Mínimo número sin detección de sonido:	0

Los tiempos de duración de las interacciones son variables y están relacionadas directamente con el número de elocuciones. En un juego estándar el usuario utiliza sus cuatro oportunidades antes de mostrar la carta que cree que el sistema eligió, pero el juego puede acortarse si el participante decide arriesgarse a adivinar antes. Por el contrario, el juego sólo puede alargarse por causa de las acciones de reparación. El juego más corto, de un poco más de los tres minutos y medio, es de un jugador que quiso adivinar la carta después de la segunda pregunta y el juego más largo, de casi siete

minutos, es de un participante que utilizó todas sus oportunidades, pero además tuvo el máximo número de problemas en el módulo de visión.

En esta serie de interacciones tuvimos un diálogo en el que el reconocimiento fue de 100%, aunque por otra parte el diálogo con el porcentaje mínimo de reconocimiento fue de 35%.

Uno de los resultados más importantes que refleja el desempeño por tratarse de un sistema de diálogo básicamente hablado, es el indicador de reconocimiento por palabra (Word Accuracy) (wa). Éste es un indicador para reconocedores utilizado ampliamente, se calcula de manera automática y requiere la existencia de referencias para cada una de las elocuciones generadas por los usuarios. Este indicador tuvo una efectividad de 41%; es decir que del total de palabras reconoció correctamente este porcentaje.

Adicionalmente, podemos observar algunos indicadores de errores comunes en la interpretación. Los errores por falta de representación semántica se refieren a la imposibilidad del sistema de asignar un significado a la señal acústica recibida; otro tipo de errores ocurre cuando no es posible detectar ninguna señal acústica relacionada con la voz, para la asignación de significado. En ambos casos tuvimos interacciones sin errores, el número máximo de errores fue de 5 y 10, respectivamente.

4.2.5. Estimación de la función de desempeño.

De acuerdo con Paradise, la función de desempeño combina una serie de costos para determinar su contribución relativa al desempeño de una aplicación.

$$P = (\alpha * N(k)) - \sum_{i=1}^n w_i * N(c_i)$$

,

El desempeño se calcula tomando en cuenta, por una parte, el éxito en el cumplimiento de la tarea y por otro, una serie de costos que influyen en la satisfacción que los usuarios manifiestan después de usar la aplicación.

Por ello, antes de entrar de lleno en el cálculo de la función, la metodología sugiere conocer la contribución que cada uno de los factores evaluados tiene sobre la satisfacción del usuario, identificada como el fin máximo de la aplicación. De la información recuperada durante las interacciones se elaboró un reporte que resume los resultados para cada uno de los participantes. El reporte incluye, como puede observarse en la tabla 4.7, la identificación del participante por número, el sexo, la edad, el resultado obtenido de la encuesta de satisfacción, el cálculo del coeficiente k , el éxito en adivinar la carta, el número de elocuciones del sistema, el número de elocuciones del usuario, el tiempo de interacción, así como los errores de reconocimiento, de entendimiento y de visión, entre otros.

Aunque es posible especular y seguramente resultaría interesante investigar la relación de la edad, el género o el nivel socioeconómico en el desempeño de los jugadores y en la determinación de los resultados de satisfacción, en este ejercicio dejamos de lado este tipo de consideraciones, ya que el trabajo de esta tesis se refiere únicamente a la evaluación del desempeño de la aplicación, de tal forma que mantuvimos para la evaluación únicamente los criterios siguientes:

- a) Éxito en el cumplimiento de la tarea, coeficiente *kappa*
- b) Número de elocuciones del usuario
- c) Número de elocuciones del sistema
- d) Tiempo de interacción
- e) Número de errores de reconocimiento
- f) Total de errores (reconocimiento, entendimiento y visión)

Los resultados se normalizaron con el fin de solventar el problema de que cada uno se presenta en una escala diferente, incluyendo también los resultados de satisfacción de

los usuarios, tal y como lo proponen Melita Hajdjinak y France Mihelic (Hajdjinak and Mihelic 2006). La regresión lineal permite determinar la contribución relativa de cada factor. Lo ideal es incluir en la función factores significativos y eliminar las redundancias, por lo que se recomienda llevar a cabo una segunda regresión.

Tabla 4.7. Juego: Adivina la carta - Registro de participantes y resultados

	No. Part.	Sexo	Edad	Satisfacción	k	Adivinó	USRINPUT ^{/1}	SYSUTT	Tiempo	Error Recono ^{/2}	Error visión	Error total
1	O1	Masculino	10	79.8	1.00	0	12	34	00:03:32	2	0	2
2	O2	Femenino	10	89.8	0.92	1	19	42	00:04:59	6	0	6
3	O3	Femenino	10	89.0	0.70	0	14	40	00:04:26	1	0	1
4	O4	Masculino	16	79.9	0.92	1	23	44	00:05:36	10	0	10
5	O5	Masculino	16	53.1	0.75	0	19	43	00:05:03	6	0	6
6	O6	Masculino	15	73.1	0.65	0	28	47	00:06:17	13	0	13
7	O7	Femenino	16	69.7	0.80	0	18	42	00:04:47	7	0	7
8	O8	Femenino	12	29.4	0.70	0	23	49	00:05:18	7	0	7
9	O11	Masculino	10	63.2	1.00	0	18	40	00:04:36	5	0	5
10	O12	Femenino	10	73.0	0.80	0	18	42	00:04:49	3	2	5
11	O13	Femenino	16	59.8	0.80	0	22	46	00:05:37	10	1	11
12	O14	Femenino	16	69.7	0.40	0	13	37	00:03:57	4	0	4
13	O15	Femenino	16	56.5	0.80	0	25	48	00:06:02	7	4	11
14	O16	Femenino	12	54.0	0.78	0	15	39	00:04:36	5	1	6
15	O17	Femenino	13	63.1	0.70	0	14	46	00:04:38	2	0	2
16	O18	Femenino	13	89.8	1.00	1	11	34	00:03:38	0	0	0
17	O19	Femenino	12	93.2	0.80	1	13	36	00:03:54	0	1	1
18	O20	Femenino	13	62.8	0.90	0	17	41	00:04:42	7	0	7
19	O21	Masculino	15	66.4	0.70	0	14	38	00:04:05	4	0	4
20	O22	Masculino	12	69.7	0.90	0	15	39	00:04:14	2	1	3
21	O23	Masculino	16	66.2	0.70	0	14	37	00:04:01	4	0	4
22	O24	Masculino	10	79.8	0.89	1	13	42	00:03:36	4	0	4
23	O25	Masculino	14	55.3	0.70	0	26	50	00:06:12	11	2	13
24	O27	Masculino	10	48.0	0.50	0	30	53	00:06:54	11	4	15
25	O30	Masculino	15	70.1	0.89	1	17	33	00:04:11	5	0	5
26	R1	Femenino	16	56.4	0.70	0	19	44	00:05:09	5	1	6
27	R2	Femenino	15	73.0	0.91	1	21	42	00:05:15	11	0	11
28	R3	Masculino	15	63.0	0.91	1	22	46	00:05:37	5	1	6
29	R4	Masculino	14	66.4	0.91	0	15	39	00:04:22	0	1	1
30	R5	Masculino	15	56.3	0.80	1	21	44	00:05:35	11	0	11

/1 Incluye lenguaje y visión. /2 Incluye No entendí y Nada

Para los cálculos utilicé el programa de análisis estadístico AnalystSoft., StatPlus:mac, versión 2009. La tabla 4.8 presenta en la columna los resultados de una primera regresión. De acuerdo con este ejercicio, la columna “Coeficientes” muestra los atributos que influyen en mayor medida en el satisfacción del usuario. De mayor a menor, el resultado fue el siguiente:

- Número de elocuciones del sistema.
- El tiempo de interacción.
- Número de elocuciones del usuario.
- El número total de errores.
- El coeficiente *kappa*.
- El número de errores de reconocimiento.

Para una segunda regresión, atendiendo directamente al resultado numérico, hubiera tomado directamente las tres primeras variables; sin embargo, por la estructura de la conversación el número de elocuciones del sistema y del usuario están relacionadas directamente. Por otra parte, la cifra de los errores totales (reconocimiento de voz y visual) contiene el valor de los errores de reconocimiento de voz. De esta manera, para la segunda regresión consideramos únicamente las variables relativas a:

- Las elocuciones del sistema
- El tiempo de interacción
- El número total de errores

El resultado relativo al coeficiente *kappa* está incluido de cualquier forma en el cálculo de la función, así que quedó incluido en el segundo ejercicio.

Tabla 4.8. Golem - Adivina la Carta - Relación de variables tomadas para la evaluación con el resultado de satisfacción

Estadísticos de Regresión							
R	0.6209						
R Cuadrado	0.38552						
R Cuadrado Ajustado	0.22522						
S	0.89527						
Número Total de Casos	30						
$\text{tsifacci?n} = -0.0000 + 0.1806 * k + 0.0953 * \text{Error Recono} * -0.2423 * \text{Error total} + 0.4824 * \text{Tiempo} - 0.2777 * \text{USRINPUT} * -0.5949 * \text{SYS}$							
ANOVA							
	d.f.	SS	MS	F	nivel p		
Regresión	6.	11.56551	1.92758	2.40497	0.05958		
Residuo	23.	18.43449	0.8015				
Total	29.	30.					
	Coefficientes	Error Estándar	LCL	UCL	Estadístico t	nivel p	HO (2%) rechazado?
Intercepto	-1.47551E-16	0.16345	-0.40861	0.40861	-9.02716E-16	1.	No
k	0.18056	0.18197	-0.27433	0.63545	0.99229	0.33138	No
Error Recono *	0.09533	0.66478	-1.56654	1.7572	0.1434	0.88722	No
Error total	-0.24226	0.88862	-2.46369	1.97917	-0.27263	0.78757	No
Tiempo	0.48239	0.81092	-1.54479	2.50957	0.59488	0.55773	No
USRINPUT *	-0.27772	0.78875	-2.24949	1.69404	-0.35211	0.72797	No
SYSUTT	-0.59493	0.37233	-1.52569	0.33583	-1.59788	0.12372	No
T (2%)	2.49987						
LCL - Valor inferior de un intervalo de confianza (LCL)							
UCL - Valor superior de un intervalo de confianza (UCL)							

De acuerdo con los resultados que observamos en la tabla 4.9., salvo que algunos valores quedan excluidos, las conclusiones son las mismas:

Tabla 4.9. Golem - Adivina la Carta - Relación de variables principales tomadas para la evaluación con el resultado de satisfacción

Estadísticos de Regresión							
R	0.61775						
R Cuadrado	0.38161						
R Cuadrado Ajustado	0.28267						
S	0.86143						
Número Total de Casos	30						
$\text{Satisfacci?n} = -0.0000 + 0.1700 * k - 0.2004 * \text{Error total} + 0.2519 * \text{Tiempo} - 0.5975 * \text{SYSUTT}$							
ANOVA							
	d.f.	SS	MS	F	nivel p		
Regresión	4.	11.44842	2.86211	3.85696	0.01419		
Residuo	25.	18.55158	0.74206				
Total	29.	30.					
	Coefficientes	Error Estándar	LCL	UCL	Estadístico t	nivel p	HO (2%) rechazado?
Intercepto	-3.12547E-16	0.15728	-0.39085	0.39085	-1.98726E-15	1.	No
k	0.17001	0.1703	-0.25319	0.59322	0.99834	0.32768	No
Error total	-0.20037	0.35849	-1.09125	0.6905	-0.55894	0.58118	No
Tiempo	0.25192	0.51245	-1.02157	1.52541	0.4916	0.62729	No
SYSUTT	-0.59754	0.35811	-1.4875	0.29241	-1.66858	0.10768	No
T (2%)	2.48511						
LCL - Valor inferior de un intervalo de confianza (LCL)							
UCL - Valor superior de un intervalo de confianza (UCL)							

Los valores de los coeficientes, que miden el impacto marginal de cada regresor sobre la variable dependiente manteniendo todo lo demás constante, muestran que los atributos que influyen en mayor medida en el satisfacción del usuario son:

- Las elocuciones del sistema
- El tiempo de interacción
- El número total de errores
- El coeficiente *kappa*

En la columna “nivel p”, referente al nivel de fiabilidad de los resultados, podemos observar que ninguna variable es por sí sola un indicador determinante y confiable en relación con el resultado de la evaluación, todos los resultados son mayores al valor de referencia recomendado (0.05). De igual manera, el indicador R^2 (varía entre 0 y 1) aumenta cuando se incrementa el número de variables explicativas, lo que confirma que las variables elegidas en su conjunto contribuyen con poca información a predecir el valor de la satisfacción del usuario.

Por último, para concluir los pasos previstos en la metodología, tomamos los valores generados en este ejercicio para sustituir en la ecuación del índice de desempeño, lo cual genera como resultado **15.08**.

CONCLUSIONES

En este trabajo hemos presentado los resultados de la evaluación del sistema conversacional con español hablado “Golem- Adivina la Carta” desarrollado en el DCC del IIMAS de la UNAM. El sistema fue desarrollado con base en una arquitectura cognitiva orientada a la interacción “Adivina la Carta”. Este robot es capaz de sostener una conversación, para lo cual utiliza el lenguaje y la visión para la comunicación con los usuarios, principalmente niños entre 10 y 16 años.

El objetivo de conocer y valorar el desempeño de una aplicación que tiene una ubicación permanente frente al público, sobre todo un público infantil, ha sido importante tanto para los miembros del equipo responsable del desarrollo de “Adivina la Carta”, así como para las autoridades del museo.

La evaluación toma como marco de referencia la metodología Paradise, a la cual hicimos algunas adecuaciones, atendiendo las características particulares de nuestra aplicación. La metodología ha sido utilizada ampliamente en evaluaciones de sistemas de diálogo y, aunque no tiene una aceptación a nivel general, constituye un procedimiento establecido que permite hacer de la valoración general un ejercicio claro y ordenado.

El recorrido para la preparación y desarrollo de la evaluación fue un ejercicio muy revelador que mostró beneficios desde la planeación y en cada una de las etapas, no sólo en la generación de los resultados finales que presentamos como resultado del cuestionario de evaluación o del cálculo de la función de desempeño.

Para Paradise, el objetivo principal de las aplicaciones debe ser maximizar la satisfacción del usuario. A este objetivo principal contribuyen dos objetivos secundarios: maximizar el éxito en el cumplimiento de la tarea y minimizar los costos involucrados. A través de una serie de métricas se cuantifican tanto el éxito en la tarea como los costos de transacción, para determinar la contribución de cada uno en la satisfacción del usuario y calcular un indicador final de desempeño.

En general, podemos afirmar que el sistema funciona y agrada a los niños, ya que los 30 participantes evaluados terminaron el juego y la mayoría manifestó estar dispuesta a jugar nuevamente con el sistema, lo cual indica un grado de satisfacción razonable.

Los resultados objetivos mostraron la necesidad de mejorar el reconocimiento de voz para contar con un sistema más robusto. Estos resultados son coherentes con la percepción que los usuarios tuvieron sobre el mismo tema, por lo que se han iniciado acciones para mejorar los modelos acústicos y del lenguaje. En aspectos como el tiempo de interacción y el número de elocuciones, los datos recuperados reflejan mejoras con respecto a las evaluaciones preliminares y serán útiles para entender mejor tanto la tarea como lo que los niños valoran después de interactuar con el sistema.

En este primer ejercicio, ninguna de las variables utilizadas mostró una relación determinante con respecto al resultado general de satisfacción, por lo que quedamos obligados e invitados a explorar y a ampliar la investigación en los temas que mencionaremos a continuación, que aunque no forman parte de la metodología, seguramente permitirán complementar el diagnóstico y conocimiento de la aplicación y de los usuarios.

En primer lugar, la tarea que plantea Adivina la Carta tiene muchas diferencias respecto a otros sistemas evaluados y no necesariamente porque la aplicación esté dirigida a niños. En primer lugar, la interacción tiene únicamente un fin lúdico y no hay un menú de tareas para elegir, ni búsquedas de información en las bases de datos y no hay transacciones por concretar. La dificultad de la tarea es inherente a la misma y tan importante es la eficiencia de la aplicación como la motivación de los usuarios por adivinar la carta elegida por el sistema.

La metodología equipara el éxito en el cumplimiento de la tarea, con el intercambio de información que el agente recibe por parte del usuario. En “Adivina la Carta” el intercambio de información es importante en los dos sentidos, el sistema requiere la información del usuario para atravesar exitosamente los subdiálogos de la aplicación, pero la información del sistema determina el comportamiento del usuario, en tanto que obtiene información útil para decidir las acciones siguientes que le faciliten deducir y, en su caso, adivinar la carta elegida. Esta situación constituye una oportunidad para considerar conceptos como la idoneidad de las respuestas de un agente, la estrategia de diálogo y la calidad de la conversación para la delimitación del concepto de tarea.

Continuando con el tema de la tarea, en opinión de los niños, la tarea es compleja y es un elemento difícil de valorar. Éste fue el único tópico al cual dedicamos tres preguntas en el cuestionario de satisfacción. La tabla 4.10 muestra un resumen de los resultados obtenidos; sin embargo no contamos con los elementos necesarios que nos permitan determinar con claridad qué elementos conforman o distinguen los distintos grados de dificultad en un diálogo que debe ser al mismo divertido y retador.

Tabla 4.10 Golem Adivina la Carta - Facilidad para terminar la tarea
- Promedio participantes -

¿Sabías qué preguntar	¿Tuviste tiempo suficiente para pensar la pregunta?	¿Qué tan difícil o fácil fue terminar el juego?	Promedio
69	64	47	60

Detectamos que uno de los principales problemas para los niños es la formulación de las preguntas. De acuerdo con la información que integramos en la tabla 4.11., la mitad de los niños utilizó para la primera pregunta los mismos tres atributos que son los que se despliegan durante las instrucciones del juego; conforme avanzó la interacción, los atributos utilizados se fueron diversificando, pero también aumentó el número de casos en los que los niños no formularon ninguna pregunta.

Tabla 4.11. Golem - Adivina la Carta - Interrogatorio

Subdiálogo Interrogatorio	Kappa	Atributos utilizados	Participantes	50% utilizó	No preguntaron
I pregunta	0.724	15	30	3	0
II pregunta	0.663	19	30	4	4
III pregunta	0.643	22	27	7	3
IV pregunta	0.339	23	27	10	7

Finalmente y casi de manera espontánea surgió lo que podríamos llamar el “componente emocional” (ver tabla 4.12.) que fue integrado como una pregunta más al cuestionario de satisfacción, sin más pretensión que recoger el sentir de un poco más del 40% de los participantes, pero que en el futuro puede ser útil para conocer el peso que este tipo de circunstancias tienen en la dificultad de la tarea.

**Tabla 4.12. Golem - Adivina la Carta -
¿Cómo te sentiste en la interacción con Golem?**

Tranquilo	Desesperado / Nervioso
17	13
56.7%	43.3%

La matriz de atributos resultó una herramienta útil que pudimos adaptar para representar actos en distintas modalidades, habla y visión, para registrar los errores por falta de representación semántica o por falta de detección, así como para reflejar la historia de la interacción, de acuerdo con la forma en que haya decidido jugar cada participante. Especialmente el resultado de cálculo del coeficiente *kappa* para cada subdiálogo fue valioso para darnos cuenta no sólo que formular las preguntas del juego es difícil, sino que esta dificultad es creciente y tiene un efecto medible en el desempeño del jugador. No obstante, fuimos menos exitosos en encontrar la relación del cálculo en el éxito en la tarea con los resultados generales de satisfacción.

Con respecto al cuestionario de satisfacción, encontramos que existen opiniones que dudan respecto a la científicidad de los tópicos que conforman el cuestionario de satisfacción. Manifiestan reparos no sólo respecto a los tópicos incluidos, sino también respecto al peso específico que debe asignarse a cada uno. La misma metodología resta importancia al tema cuando sugiere, sin mayor complicación, tomar el promedio de los resultados obtenidos las preguntas o elegir el resultado de alguna pregunta particular.

En mi opinión, el cuestionario de satisfacción es una prueba costosa, pero muy necesaria; por ello, es conveniente aprovechar la oportunidad de tener contacto con usuarios reales, dispuestos a contestar para obtener la información que nos permita conocer las razones de satisfacción o insatisfacción con respecto al desempeño y a la interacción con la aplicación.

Seguramente en algunos casos será posible lograr alguna coherencia entre los resultados del cuestionario de satisfacción, subjetivo por definición, con los datos del registro del sistema, objetivos por circunstancia, como muestra la tabla 4.13. relativa a los resultados sobre el desempeño en el reconocimiento. En otros casos, será necesario buscar las explicaciones obteniendo información adicional del usuario, o descubriendo nuevas relaciones estadísticas entre los datos objetivos del sistema y las causas de satisfacción entre los usuarios tal y como lo proponen Zoraida Callejas y Ramón López Cózar (Callejas and López - Cózar 2008).

Tabla 4.13. Golem - Adivina la Carta Resultados sobre desempeño del sistema

Cuestionario de satisfacción		Registro automático	
¿Entendió el Golem lo que tú decías? (promedio)	47.40	% reconocimiento por diálogo (promedio)	57.00

Aunque no es parte del trabajo de esta tesis, también queda pendiente comprobar la facilidad de comparar los resultados obtenidos de versiones distintas de un mismo sistema o de sistemas distintos, así como sobre la facilidad de generar modelos predictivos de la usabilidad.

Por lo pronto contamos con información de desempeño recuperada a partir de las evaluaciones preliminares de “Adivina la Carta”, que ha sido útil como un referente claro para priorizar y apreciar las acciones de mejora de versiones distintas del sistema y esperamos eventualmente estar en posibilidades de predecir cambios en la satisfacción del usuario, pero esto requerirá un trabajo importante en la investigación.

En cualquier caso, como vimos en los capítulos iniciales de este trabajo, la construcción de sistemas de diálogo orientados a la solución de tareas específicas, es una tarea de una complejidad considerable y en ese sentido la evaluación también lo es.

No hay necesidad de esperar al futuro; a la fecha, los agentes conversacionales que se están desarrollando no sólo abordan situaciones cada vez más complejas, sino que están utilizando más modalidades de interacción. También en el IIMAS, la arquitectura cognitiva utilizada para el desarrollo de “Adivina la Carta” ha sido utilizada para la construcción de diversas aplicaciones que igualmente integran distintas modalidades de interacción, por ejemplo el caso de un robot conversacional capaz de guiar una sesión para la presentación de carteles relacionados con los proyectos que se desarrollan en el DCC del IIMAS, el cual utiliza en la interacción el gesto de apuntar. Para estas evaluaciones tomamos como base algunos de los principios que sugiere la evaluación de PARADISE. De estas evaluaciones han surgido tres artículos de investigación que fueron

publicados en medios especializados en temas de inteligencia artificial y procesamiento del lenguaje natural. (Avilés, Alvarado, et al. 2010; Meza, Salinas and Venegas, et al. 2010; Venegas, Meza and Pineda 2011).

Por lo anterior será necesario encontrar pronto nuevas formas para evaluar estos agentes o construir evaluaciones a partir de ejercicios parciales orientados a la valoración de agentes o aspectos particulares que después puedan integrarse a un sistema general.

BIBLIOGRAFÍA

Allen, et al. "An Architecture for a Generic Dialogue Shell." *Natural Language Engineering*, 2000: 213-228.

Aguilar, Wendy, and Luis A. Pineda. "Integrating Graph-Based Vision Perception to Spoken Conversation in Human Robot Interaction." Edited by Joan Cabestany and Juan M. Corchado. *Bio-Inspired Systems: Computational and Ambient Intelligence, 10th International Work - Conference on Artificial Neural Networks, ISANN 2009*. Salamanca, Spain: Springer Verlag, 2009. 789-796.

Avilés, Héctor, Iván V. Meza, Wendy Aguilar, and Luis A. Pineda. "Integrating Pointing Gestures into a Spanish-spoken Dialog Systems for Conversational Service Robots." Edited by Joaquim Felipe and Ana L. N. Fred and Bernardette Sharp. *ICAART 2010 Proceedings of the International Conference on agents and Artificial Intelligence*. Valencia: INSTICC Press, 2010. 585-588.

Avilés, Héctor, Montserrat Alvarado, Esther Venegas, Caleb Rascón, Iván Meza Ruiz, and Luis A. Pineda. "Developmento of a Tour-Guide Robort Using Dialogue Models and a Cognitive Architecture." Edited by Ángel Fernando Kuri Morales and Guillermo Ricardo Simari. *Advances in Artificial Intlligence - IBERAMIA 2010, 12th Ibero-American Conference on AI*. Bahía Blanca, Argentina: Springer, 2010. 512-521.

Callejas, Zoraida, and Ramón López-Cózar. "Nueva Técnica de Generación Automática de Gramáticas para Sistemas de Diálogo." Edited by Sociedad Española para el Procesamiento del Lenguaje Natural. *Pvocesamiento del Lenguaje Natural*, no. 35 (Septiembre 2005): 205-212.

Callejas, Zoraida, and Ramón López - Cózar. "Relations between de facto critaria in the evaluation of a spoken dialogue system." Edited by Elsevier B. V. *Speech comunication* 50, no. 8-9 (2008): 646-665.

Carletta, Jean. "Assessing agreement on classification tasks: the kappa statistics." *Computational Linguistics*, January 1, 1996: 249-254.

Cerda, Jaime, and Luis Villarroel del P. "Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa." *Revista chilena de pediatría* 79, no. 1 (Febrero 2008): 54-58.

Cuayáhuitl, Heriberto. *Hierarchical Reinforcement Learning for Spoken Dialogue Systems*. Edinburgh: University of Edinburgh, 2009.

Cuayáhuitl, Heriberto, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. "Evaluation of a hierarchical reinforcement learning spoken dialogue system." *Computational Speech Language* (Academic Press Ltd.) 24, no. 2 (April 2010): 395-429.

Garfield, Eugene. "Citation Classic - A coefficient of agreement for nominal scales." *This Week Citation Classic*. January 18, 1986. <http://www.garfield.library.upenn.edu/classics1986/A1986AXF2600001.pdf> (accessed March 10, 2011).

Hajdinjak, Melita, and France Mihelic. "The PARADISE Evaluation Framework: Issues and Findings." *Computational Linguistics* (MIT Press) 32, no. 2 (June 2006): 263-272.

Hjalmarsson, Anna. *Evaluating AdApt, a multi-modal conversational, dialogue system using PARADISE*. Estocolmo: Royal Institute of Technology, 2002.

Kamm, Candace A., Diane J. Litman, and Marilyn A. Walker. "From Novice To Expert: The Effect Of Tutorials On User Expertise With Spoken Dialogue Systems." *Proceedings of the International Conference on Spoken Language Processing, ICSLP98*. 1998. 1211-1214.

Litmann, Diane J., and Shimei Pan. "Empirically Evaluating an Adaptable Spoken Dialogue System." *Proceedings of the 7th International Conference on User Modeling*. Banff, Canada: Springer-Verlag New York, Inc., 1999. 55-64.

López de Ullibarri Galparsoro, I., and S. Pita Fernández. "Medidas de concordancia: el índice de Kappa." *Fisterra.com*. Septiembre 24, 2004. <http://www.fisterra.com/mbe/investiga/kappa/kappa.asp> (accessed March 11, 2011).

López Escobedo, Fernanda. *El estudio de los diptongos del español de México para su aplicación en un reconocedor de habla*. México: Universidad Nacional Autónoma de México, 2004.

Meza Ruiz, Ivan Vladimir. *Sistema de verbos auxiliares del español en HPSG*. México: Facultad de Ingeniería UNAM, 2002.

Meza, Ivan V., Lisset Salinas, Esther Venegas, Hayde Castellanos, Alejandra Chavarría, and Luis A. Pineda. "Specification and Evaluation of a Spanish Conversational System

Using Dialogue Models." Edited by Ángel Fernando Kuri Morales and Guillermo Ricardo Simari. *Advances in Artificial Intelligence - IBERAMIA 2010 12th Ibero - American Conference on AI*. Bahía Blanca: Springer, 2010. 346-355.

Meza, Ivan V., Lisset Salinas, Héctor Avilés, and Luis A. Pineda. "a multimodal Dialogue System for Playing the Game "Guess the Card"." *Porcesamiento de lenguaje Natural*. 2010. 131-138.

Moliner, Luis M. "Medidas de concordancia para variables cualitativas." *Sociedad Española de Hipertensión - Liga Española para la Lucha contra la Hipertensión Arterial*. Septiembre 1, 2001. <http://www.seh-lelha.org/concor2.htm> (accessed Marzo 11, 2011).

Paek, Tim. "Empirical methods for evaluating Dialog Systems." *Proceedings of the workshop on Evaluation for Language and Dialogue Systemsorkshop on* . Toulouse, France: Association for Computational Linguistics, 2001. 1-8.

Pineda, Luis A. "Diálogos Inteligentes Multimodales en Español." --, 1 1, 2007: 1.

—. "An Interaction Oriented Cognitive Architecture for Service Robots." *Proceedings of Iberamia*. Bahía Blanca: ACL, 2010. 1-10.

—. "Specification and Interpretation of Multimodal Dialogue Models for Human Hobot interaction." *Artificial Intelligence for Humans, Service Robots and Social Modeling*, 2008: 20-29.

Pineda, Luis A. "De los sistemas conversacionales a los robots parlantes." *Ciencia*, Julio 2011: (por aparecer).

—. *El proyecto DIME y el robot conversacional Golem: una experiencia multidisciplinaria entre la computación y la lingüística*. 2004. <http://leibniz.iimas.unam.mx/~luis/golem/papers/DIME-Golem.pdf> (accessed March 6, 2011).

Pineda, Luis A., Luis Villaseñor, Javier Cuétara, Hayde Castellanos, and López Ivonne. "A New Phonetic and Speech Corpus for Mexican Spanish." Edited by Christian Lemaître, Carlos A. Reyes and Jesús A. González. *Advances in Artificial Intelligence - IBERAMIA 2004, 9th Ibero-American Convergence on AI*. Puebla, México: Springer, 2004. 974-984.

Pineda, Luis A., et al. "The Corpus DIMEx100: trnascryption and Evaluation." *Language Resources and Evaluation*, January 1, 2010: 347-370.

Pineda, Luis A., Ivan V. Meza, and Lisset Salinas. "Dialogue Model Specification and Interpretation for Intelligent Multimodal HCI." Edited by Angel Fernando Kuri Morales and Guillermo Ricardo Simari. *Advances in Artificial Intelligence - IBERAMIA 2010, 12th Ibero-American Conference on AI*. Bahía Blanca, Argentina: Springer, 2010. 20-29.

Pineda, Luis A., Varinia Estrada, Sergio Coria, and James Allen. "The obligations and common ground structure of practical dialogues." *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 2007: 9-17.

Rascón, Caleb, Héctor Avilés, and Luis A. Pineda. "Robotic Orientation towards Speaker for Human - Robot Interaction." Edited by Ángel Fernando Kuri Morales and Guillermo Ricardo Simari. *Advances in Artificial Intelligence - IBERAMIA 2010, 12th Ibero-American Conference on AI*. Bahía Blanca: Springer, 2010. 10-19.

Searle, John R. "What is Language? Some Preliminary Remarks." *Proffessor John Searle. Home page at the University of Carlifornia, Berkeley*. Novembre 6, 2006. <http://socrates.berkeley.edu/~jsearle/whatislanguage.pdf> (accessed Mazo 6, 2011).

Skantze, Gabriel. "Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems." *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*. Chateau-d'Oex-Vaud, Switzerland, 2003. 71-76.

Sparck - Jones, Karen, and Julia R. Galliers. *Evaluating Natural Language processing Systems*. Cambridge: Springer, 1996.

Venegas, Esther, Ivan V. Meza, and Luis A. Pineda. "Evaluation of a Dialogue System for Children based on a Cognitivo Architecture oriented to Interaction." *Procesamiento del Lenguaje Natural*, 2011: Por aparecer.

Walker, Marilyn A, Candace Kamm, and Julie Boland. "Developing and Testing General Models of Spoken Dialogue System Performance." *Proceedings Language Resources and Evaluation Conference, LREC-2000*. Malta: MIT Press, 2000.

Walker, Marilyn A. "An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email." *Journal of Artificial Intelligence Research (AI Access Foundation)* 12, no. 1 (June 2000): 387-416.

Walker, Marilyn A., Diane J. Litmann, Candace , A. Kamm, and Alicia Abella. "Paradise: A framework for evaluating spoken dialogue agents." *International Conference on Computational Linguistics*. Madrid: ACL, 1997. 271-280.

Walker, Marilyn, A., Jeane C. Fromer, and Narayanan Shrikanth. "Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email." *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics, COLING/ACL 98*. 1998. 1345-1352.

APÉNDICE

1. Diálogos hipotéticos utilizados por Morena Danieli y Elisabetta Gerbino

Interacción del Agente A

- A1 Hola. Este es el Servicio de Información Trenes
 Por favor, hable después del tono
 ¿Qué información es la que necesita?
- U1 Quiero ir de Turín a Milán.
- A2 ¿Quiére usted ir de Trento a Milán?
 ¿Sí o no?
- U2 No.
- A3 ¿Quiére usted salir de Trento?
 ¿Sí o no?
- U3 No.
- A4 ¿De dónde quiere usted salir?
- U4 Yo quiero salir de Turín.
- U5 ¿Quiére usted salir de Turín?
 ¿Sí o no?
- A6 ¿Quiere usted ir a Milán?
- U6 Sí.
- A7 ¿A qué hora quiere usted salir?
- U7 Quisiera salir por la tarde.
- A8 ¿Quiére usted salir entre 6 y 9 PM?
 ¿Sí o no?
- U8 Sí.
- A9 Hay un tren que sale a las 8:00 PM.

Interacción del Agente B

- B1 Hola. Este es el Servicio de Información Trenes
 Por favor, hable después del tono
 ¿Qué información es la que necesita?
- U1 Quiero viajar de Turín a Milán.
- B2 ¿A qué hora quiere usted salir de Merano a Milán?
- U2 No, quiero salir de Torino en la tarde.
- B3 ¿Quiére usted salir de Torino entre 6 y 11 PM?
 Por favor conteste sí o no.
- U3 Sí.
- A4 Hay un tren que sale a las 8:00 PM.

2. Bitácora de interacción para el levantamiento de información

Juego: Adivina la carta - Bitácora de interacción

SD	Información requerida	Real	Entendido	No entendi / No escuché	Golem	Observaciones
S1	Nombre del usuario					
	Edad					
S2	Confirma edad sí /no					
S3	Instrucciones del juego sí /no					
	Confirma instrucciones sí /no					
	a) Característica o carta					
	Confirma terminar sí /no					
	Confirma carta sí /no					
	b) Característica o carta					
	Confirma terminar sí /no					
	Confirma carta sí /no					
S4	c) Característica o carta					
	Confirma terminar sí /no					
	Confirma carta sí /no					
	d) Característica o carta					
	Confirma terminar sí /no					
	Confirma carta sí /no					
S5	Muestra carta					
	Confirma carta sí /no					
S6	Información requerida	Recuerda nombre	Ganó			Observaciones
	Despedida					

3. Golem: Adivina la carta - Matriz de confusión y cálculo de Kappa - Nombre del usuario

Resumen de resultados

	aaron	abigail	ada	ainara	amos	ana	angel	antonio	carla	emilio	eric	frida	herminio	ines	isabel	hugo	javier	jesus	karen	kati	luis	manuel	maria	mariana	moises	olga	ramiro	ricardo	roman	ruben	sergio	sofia	sol	NB				
aaron	1																																					
abigail		1																																				
ada																									1													
ainara																																						
amos																																					1	
ana																																						
angel							1																															
antonio								2																														
carla									1																													
emilio																																						
eric											1																											
frida												1																										
herminio																																						
ines																																					1	
isabel																1																						
hugo																	1																					
javier																																						
jesus																																					1	
karen																					1																	
kati																						1																
luis																							1															
manuel																								1														
maria																									1													
mariana																										1												
moises																											1											
olga																												1										
ramiro																													1									
ricardo																														1								
roman																																						
ruben																																						
sergio																																						
sofia																																						
sol																																						
NB																																						1
ti	1	1	0	0	0	1	1	2	1	2	1	1	0	0	1	1	0	0	1	1	1	1	0	2	1	0	0	0	1	0	2	1	2	0	4			
T	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30		
ti/T	0.03	0.03	0	0	0	0	0	0.067	0	0.067	0	0.03	0	0	0.03	0.03	0	0	0	0	0	0	0.03	0	0.067	0.03	0	0	0.033	0	0.07	0.03	0.1	0	0.13			
(ti/T)^2	0.001	0.001	####	0.000	0.000	####	0.001	0.004	####	0.004	####	0.001	0.000	####	0.001	0.001	0.000	0.000	0.001	####	####	0.001	0.000	0.004	0.001	0.000	0.000	0.001	0.000	0.004	0.001	####	####	0.018				
Intercambio	30																																				30	
Aciertos																																						19
PA																																						0.633
PE																																						0.058
k																																						0.611

4. Golem: Adivina la carta - Matriz de confusión y cálculo de Kappa - Subdiálogo Edad

Resumen de resultados

	diez,10	once,11	doce,12	trece,13	catorce,14	quince,15	dieciseis,16
diez,10	6						
once,11							
doce,12			3				
trece,13				3			
catorce,14					3		
quince,15						5	
dieciseis,16							8
Error			1				1

ti	6	0	4	3	3	5	9	Intercambio	Aciertos	PA
T	30	30	30	30	30	30	30	30	28	0.93
ti/T	0.2	0	0.133	0.1	0.1	0.1667	0.3			
(ti/T)^2	0.040	0.000	0.018	0.010	0.010	0.028	0.090	0.196	PE	
								0.92	k	

5. Golem: Adivina la carta - Matriz de confusión y calculo de Kappa - Pregunta a

Resumen de resultados

		aparato	aro	azul	blanc[[^]]]+, blanco	canica	colores	espiral	estrellas	gas	naranj a	onda	planet[[^]]]+, planeta	redond[[^]]]+, redondo	roj[[^]]+, rojo	NB			
aparato		1																	
aro																			
azul				2															
blanc[[^]]+,blanco					1														
canica						4													
colores							1												
espiral								1											
Estrellas									1										
gas													1						
naranja											1								
onda																			
planet[[^]]+,planeta										1			1						
redond[[^]]+,redondo			2									1		6					
roj[[^]]+,rojo															5				
NB																1	Intercambio:	Aciertos	PA
	ti	1	2	2	1	4	1	1	1	1	1	1	2	6	5	1	30	25	0.83
	T	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30			
	ti/T	0.03	0.07	0.07	0.03	0.13	0.03	0.03	0.03	0.03	0.03	0.03	0.07	0.20	0.17	0.03			
	(ti/T)^2	0.001	0.004	0.004	0.001	0.018	0.001	0.001	0.001	0.001	0.001	0.001	0.004	0.040	0.028	####	0.109	PE	
																	0.724	k	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			

6. Golem: Adivina la carta - Matriz de confusión y cálculo de Kappa - Pregunta b

Resumen de resultados

	--	arco	azul	blanc[^]+, blanco	calor	canica	colores	espiral	grande	luz	negr[^]+, negro	nube	ojo	onda	planeta	puntos	redond[^]+, redondo	roj[^]+, rojo	sol	tierra			
--																							
arco	1															1							
azul			6																				
blanc[^]+, blanco				4																			
calor					1																		
canica						1																	
colores							1																
espiral								1															
grande									1														
luz	1			1																			
negr[^]+, negro											1												
nube												1											
ojo	1																						
onda																							
planeta																							
puntos									1														
redond[^]+, redondo																							
roj[^]+, rojo	1																						
sol																							
tierra																							
ti	4	0	6	5	1	1	2	1	1	0	1	1	0	0	0	1	1	2	1	2	30	21	0.7
T	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30			
ti/T	0.13	0.00	0.20	0.17	0.03	0.03	0.07	0.03	0.03	0.00	0.03	0.03	0.00	0.00	0.00	0.03	0.03	0.07	0.03	0.07			
(ti/T)^2	0.018	0.000	0.040	0.028	#####	0.001	0.004	0.001	0.001	#####	0.001	0.001	#####	0.000	0.000	0.001	0.001	0.004	0.001	0.004	0.109	PE	
																					0.663	k	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19			

7. Golem: Adivina la carta - Matriz de confusión y cálculo de Kappa - Pregunta c

Resumen de resultados

	--	azul	blanc[^]+, blanco	cafe	canica	circul[^]+, circular	gas	gris	irregular	Júpiter	Luz	Luna	naranja	nube	ojo	pelota	planet[^]+, planeta	queso	redond[^]+, redondo	roj[^]]+,rojo	rosa[^]]+,rosa	sol	vida					
--																												
azul		2		1																								
blanc[^]+, blanco			3																									
cafe																												
canica					1																							
circul[^]+, circular						1																						
gas								1																				
gris									1																			
irregular										1																		
Júpiter											1																	
Luz												1																
Luna	1																											
naranja	1																											
nube													1															
ojo																												
pelota																												
planet[^]+, planeta																												
queso																												
redond[^]+, redondo																												
roj[^]+, rojo																												
rosa[^]+, rosa																												
sol																												
vida																												
ti	3	2	3	1	1	2	0	1	1	1	1	0	1	1	0	1	1	0	1	3	1	1	1		Intercambio	Acertos	PA	
T	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27		27	18	0.667	
ti/T	0.11	0.07	0.11	0.04	0.04	0.07	0.00	0.04	0.04	0.04	0.04	0.00	0.04	0.04	0.00	0.04	0.04	0.00	0.04	0.11	0.04	0.04	0.04					
(ti/T)^2	0.012	0.005	0.012	0.001	0.001	0.005	0.000	0.001	0.001	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.012	0.001	0.001	0.001		0.067	PE		
																										0.643	k	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22					

8. Golem: Adivina la carta - Matriz de confusión y cálculo de Kappa - Pregunta d

Resumen de resultados

	agua	blanc[^]+, blanco	bola	canica	caracol	flor	gigante	irregular	Júpiter	larga	luz	negr[^]+ ,negro	nube	ocho	ojo	planet[^]+ planeta	puntos	queso	redond[^]+ redondo	roj[^]+ rojo	sol	verde	NB					
-																												
agua					1														1									
blanc[^]+,blanco		1																										
bola	1							1																				
canica				1																								
caracol																												
flor										1					1													
gigante																												
irregular																												
Júpiter									1																			
larga																												
luz	2						1																					
negr[^]+ ,negro												1																
nube													1															
ocho																												
ojo	1																											
planet[^]+ planeta																	2											
puntos																												
queso																												
redond[^]+ redondo																				1								
roj[^]+ rojo																					2							
sol	3																					1						
verde																							1					
NB																								1				
ti	7	0	1	0	1	1	0	1	1	1	0	1	1	1	0	2	1	0	2	2	1	1	1		Intercambio	Aciertos	PA	
T	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27		27	11	0.407	
ti/T	0.26	0.00	0.04	0.00	0.04	0.04	0.00	0.04	0.04	0.04	0.00	0.04	0.04	0.04	0.00	0.07	0.04	0.00	0.07	0.07	0.04	0.04	0.04					
(ti/T)^2	0.067	0.000	0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.001	0.000	0.001	0.001	0.001	0.000	0.005	0.001	0.000	0.005	0.005	0.001	0.001	0.001		0.103	PE		
																									0.339	k		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23					

10. Golem - Adivina la Carta

Resultado del cálculo del coeficiente Kappa por subdiálogo

Usuario	Kappa	Aciertos	Acuerdo por azar
1	1.000	1.000	0.100
2	0.917	0.923	0.077
3	0.700	0.727	0.091
4	0.917	0.923	0.077
5	0.750	0.769	0.077
6	0.654	0.692	0.110
7	0.796	0.818	0.110
8	0.700	0.727	0.091
11	1.000	1.000	0.091
12	0.800	0.818	0.091
13	0.800	0.818	0.091
14	0.400	0.455	0.091
15	0.800	0.818	0.091
16	0.782	0.800	0.083
17	0.700	0.727	0.091
18	1.000	1.000	0.091
19	0.800	0.818	0.091
20	0.900	0.909	0.091
21	0.700	0.727	0.091
22	0.900	0.909	0.091
23	0.700	0.727	0.091
24	0.891	0.900	0.083
25	0.700	0.727	0.091
27	0.500	0.545	0.091
30	0.891	0.900	0.083
r1	0.700	0.727	0.091
r2	0.914	0.923	0.107
r3	0.914	0.923	0.107
r4	0.914	0.923	0.107
r5	0.800	0.818	0.091
PROMEDIO	0.80	0.82	0.09

11. Resumen de información obtenida del registro automático de interacciones

=====Información básica:	
Total diálogos:	30
Promedio duración de los diálogos:	04:53.6
Promedio número de elocuciones (usuario):	16.7
Promedio número de elocuciones (sistema):	41.9
=====Máximos y mínimos - información por diálogo:	
Máxima duración de un diálogo:	00:06:53
Mínima duración de un diálogo:	00:03:32
Máximo número de elocuciones (usuario):	27
Mínimo número de elocuciones (usuario):	10
Máximo número de elocuciones (sistema):	53
Mínimo número de elocuciones (sistema):	33
=====Información sobre visión:	
Total número de acciones de visión:	52
Promedio número de acciones de visión:	1.73
Máximo número de acciones de visión:	5
Mínimo número de acciones de visión:	1
=====Información sobre la tarea:	
Total terminaron el juego:	30
Total despedida con nombre:	29
Total juegos ganados:	9
Total juegos perdidos:	21
Promedio número de preguntas:	3.77
Máximo número de preguntas:	4
Mínimo número de preguntas:	2
=====Información sobre el reconocimiento de voz:	
% error en reconocimiento por palabra:	0.59
% acierto en reconocimiento por palabra:	0.41
=====Interpretación de información:	
% Reconocimiento por diálogo:	0.57
% Máximo de reconocimiento por diálogo:	1
% Mínimo de reconocimiento por diálogo:	0.35
Representaciones con predicado:	501
Representaciones con argumentos:	180
Promedio error en predicados:	0.58
Promedio error en argumentos:	0.55
Total predicados categoría CARACTERISTICA:	103
Total predicados categoría NOMBRE:	29
Total predicado categoría EDAD:	31
% Aciertos predicados categoría CARACTERISTICA:	0.51
% Aciertos predicado categoría NOMBRE:	0.45
% Aciertos predicado categoría EDAD:	0.9
=====Interpretación de errores:	
Sin representación semántica:	135
Sin detección de sonido:	30
Máximo número sin representación semántica:	10
Mínimo número sin representación semántica:	0
Máximo número sin detección de sonido:	5
Mínimo número sin detección de sonido:	0

12. Golem - Adivina la Carta

Resultados de satisfacción por usuario

No.	Participantes	Calificación
1	O1	79.80
2	O2	89.80
3	O3	89.00
4	O4	79.90
5	O5	53.10
6	O6	73.10
7	O7	69.70
8	O8	29.44
9	O11	56.40
10	O12	73.00
11	O13	63.20
12	O14	73.00
13	O15	59.80
14	O16	69.70
15	O17	56.50
16	O18	54.00
17	O19	63.10
18	O20	89.80
19	O21	93.20
20	O22	62.78
21	O23	66.40
22	O24	69.70
23	O25	66.20
24	O27	79.80
25	O30	55.33
26	R1	63.00
27	R2	48.00
28	R3	66.40
29	R4	56.30
30	R5	70.11