DS-UCAT: SISTEMA DE DIÁLOGO MULTIMODAL Y MULTILINGÜE PARA UN ENTORNO EDUCATIVO

Ramón López-Cózar[1], Zoraida Callejas[1], Germán Montoro[2], Pablo Haya[2]

[1]Dpto. Lenguajes y Sistemas Informáticos, ETS Ingenierías Informática y Telecomunicaciones, Universidad de Granada, {rlopezc, zoraida}@ugr.es

[2]Dpto. de Ingeniería Informática, Universidad Autónoma de Madrid {German.Montoro, Pablo.Haya}@uam.es

RESUMEN

En este artículo presentamos un sistema de diálogo multimodal y multilingüe que estamos desarrollando para proporcionar asistencia a estudiantes y profesores en algunas de sus actividades habituales en un entorno educativo, p. e. en una Facultad de una Universidad. Tenemos previsto que además de interactuar con el usuario, el sistema pueda interactuar con el entorno en que éste se encuentra en un momento dado, el cual puede cambiar a lo largo de una interacción conforme el usuario se mueve dentro del centro educativo. El artículo describe la arquitectura del sistema, muestra cómo se realiza la interacción con la versión actual del mismo, y comenta cómo tenemos previsto utilizar técnicas de inteligencia ambiental para mejorar su funcionamiento.

1. INTRODUCCIÓN

Recientemente se han desarrollado diversas tecnologías para facilitar el uso de los ordenadores a personas no familiarizadas con la Informática. Por ejemplo, los sistemas de diálogo oral (*Spoken Dialogue Systems*) son programas informáticos diseñados para emular el comportamiento y la capacidad de comunicación de un ser humano ante una tarea dada [1]. Su finalidad es proporcionar información a los usuarios (usualmente a través del teléfono) usando habla como medio de interacción. Algunos sistemas de este tipo que pueden encontrarse en la bibliografía son los siguientes: Pegasus [2], Voyager [3], Dialogos [4], Saplen [5], Jupiter [6], Mercury [7], Dihana [8], Ritel [9] and UAH [10]. Los sistemas de diálogo multimodal (*Multimodal Dialogue Systems*) son el resultado de una sofisticación de los sistemas de diálogo oral, pues permiten que el usuario pueda utilizar diversos canales de comunicación durante la interacción (p. e. habla, texto, sonidos, etc.). Para hacer posible la interacción con usuarios que hablan distintos idiomas, los sistemas de diálogo multilingüe (*Multilingual Dialogue Systems*) permiten que los usuarios puedan comunicarse con el ordenador usando un idioma de entre varios disponibles (p. e. Castellano o Inglés), siendo la funcionalidad del sistema la misma independientemente del idioma utilizado [11].

Por otra parte, los denominados entornos *"activos"* (también llamados entornos de *"inteligencia ambiental"*) tienen como finalidad proporcionar un modo de interacción natural entre el entorno y sus *"habitantes"* [12]. Su objetivo es que el entorno ayude a las personas en su vida diaria, ofreciendo modos no intrusivos de comunicación. De este modo, se pretende que los hogares, aulas, despachos o laboratorios, por ejemplo, sean capaces de asistir a los usuarios en la realización de sus tareas cotidianas [13], [14].

Tras esta breve introducción, el resto del artículo está organizado como sigue. La sección 2 presenta el sistema de diálogo DS-UCAT, comentando su organización mediante documentos X+V, sus interfaces oral y visual, la conexión entre ambas interfaces, y las bases de datos usadas actualmente por el sistema, así como otras que están pendientes de ser creadas. La sección 3 describe cómo se realiza la interacción con la versión actual del sistema, y muestra, a modo de ejemplo, cómo se puede realizar la consulta multimodal de fuentes bibliográficas. La sección 4 comenta cómo tenemos previsto implementar la interacción entre el sistema de diálogo y el entorno educativo. Finalmente, la sección 5 presenta las conclusiones y comenta líneas de trabajo futuro.

2. EL SISTEMA DS-UCAT

Una línea de investigación del proyecto UCAT (*Ubiquitous Collaborative Training*)[1] está dedicada a la implementación de un sistema de diálogo multimodal y multilingüe denominado DS-UCAT (*Dialogue System for Ubiquitous Collaborative Training*). La finalidad de este sistema es proporcionar asistencia a profesores y

[1] Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología, mediante el proyecto TIN2004-03140 Ubiquitous Collaborative Training.

estudiantes en algunas de sus actividades cotidianas dentro un centro educativo (p. e. en una Facultad de una Universidad). En el diseño actual consideramos que el sistema podrá prestar servicio en tres tipos de ubicaciones dentro del entorno educativo: biblioteca, despachos de profesores y aulas. En nuestra implementación actual el sistema puede interactuar con el usuario usando sonidos, voz, gráficos y texto. La entrada multimodal permite al usuario proporcionar información al sistema a través de habla, teclado y ratón. Por ejemplo, el usuario puede consultar si existe disponibilidad de libros sobre una determinada materia en la biblioteca pronunciando el nombre de la materia (p. e. Matemáticas), seleccionándola con el ratón en una lista de materias, o escribiendo el nombre de la materia en un campo de un formulario. Dado que la salida del sistema también es multimodal, para esta consulta el sistema genera un mensaje oral indicando que se muestra en pantalla un listado con libros encontrados en la base de datos relacionados con la materia indicada.

La Figura 1 muestra la arquitectura modular de la versión actual del sistema, compuesta por un servidor de documentos XHTML+Voice (también llamados documentos X+V) al que acceden los dispositivos móviles de los usuarios (Tablet PCs, ordenadores portátiles o PDAs) mediante conexiones inalámbricas.

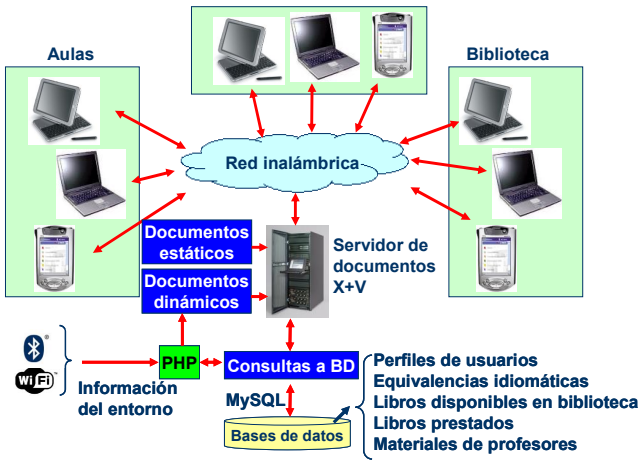

Figura 1. *Arquitectura del sistema DS-UCAT.*

2.1. Documentos XHTML+Voice (X+V)

El sistema está implementado mediante un conjunto de documentos X+V. Algunos de ellos se encuentran almacenados en el servidor de documentos X+V mostrado en la Figura 1, mientras que otros se generan dinámicamente usando programas PHP que tienen en cuenta las características y preferencias de los usuarios, p. e. sexo (masculino/femenino) y lenguaje de interacción preferido (p. e. Inglés), así como información extraída de diversas bases de datos. Los documentos X+V constan de formularios cuyos campos se pueden rellenar mediante habla, texto o selecciones de opciones efectuadas con el ratón.

Para visualizar los documentos, los usuarios deben disponer en sus dispositivos móviles de un navegador que soporte la interacción oral y soporte la especificación X+V. En nuestra implementación usamos el navegador Opera[2], el cual permite introducir comandos mediante voz (p. e. para recargar una página, detener el acceso a Internet, volver hacia atrás, etc.). Además, usando las gramáticas adecuadas para el reconocimiento del habla (RAH), el navegador puede reconocer las frases pronunciadas relacionadas con nuestro dominio de aplicación (entorno educativo). Para implementar la interacción multimodal hemos desarrollado dos interfaces (una para la interacción oral y la otra para la interacción visual) que describimos a continuación.

2.1.1. Interfaz oral

La interfaz oral permite realizar el RAH y la síntesis del habla. Ambos procesos se llevan a cabo mediante el reconocedor y sintetizador de habla incluidos en el navegador Opera. En nuestra implementación actual el RAH se realiza mediante la técnica tap-&-talk, es decir, el usuario debe hacer clic en un determinado icono del navegador y mantenerlo pulsado mientras está hablando (no obstante, esta es una características que puede ser cambiada fácilmente en la sección de "preferencias" del navegador). Para realizar el RAH y la comprensión del habla utilizamos gramáticas JSGF (Java Speech Grammar Format). Algunas de estas gramáticas las usamos a nivel de formulario (para que se puedan rellenar varios campos del mismo en una única interacción oral) mientras que otras las usamos a nivel de campo (para que sólo se pueda rellenar un campo del formulario en una interacción). Algunas gramáticas son estáticas, mientras que otras se generan dinámicamente usando programas PHP que realizan consultas MySQL a las bases de datos e incluyen los datos obtenidos en el vocabulario de dichas gramáticas (p. e. títulos de libros). Por ejemplo, utilizando una de estas gramáticas para reconocer consultas de libros efectuadas en el idioma Inglés, si el usuario pronuncia la frase "I need books about math please", el sistema rellena el campo "materia" con la palabra "math".

Las gramáticas usadas para el reconocimiento de consultas de libros deben ser actualizadas conforme se producen cambios en el catálogo de libros disponibles en la biblioteca, de ahí que se generen de forma dinámica a partir del contenido de la base de datos Libros Disponible en biblioteca (ver Figura 1). Para actualizar estas gramáticas hemos implementado un programa PHP que realiza dos tareas. En primer lugar, realiza consultas MySQL a las base de datos para obtener la información relativa a los libros disponibles,

[2] http://www.opera.com

como autores, títulos y materias. En segundo lugar, crea gramáticas para reconocer tanto frases completas como datos de forma aislada (p. e. títulos, autores o materias) usando la información obtenida en la etapa anterior.

En la salida del sistema, la síntesis de habla se realiza incluyendo frases en formato de texto en sentencias VoiceXML[3] del tipo <prompt> … </prompt>. Estas frases se transforman en voz mediante el sintetizador de voz integrado en el navegador. Algunas de estas frases tienen un formato fijo, mientras que otras se crean en tiempo de ejecución teniendo en cuenta el tipo de usuario (profesor o estudiante), su sexo (masculino o femenino) y la información extraída de algunas bases de datos.

2.1.2. Interfaz visual

En la entrada al sistema, la interacción visual se usa para obtener datos del usuario mediante campos y botones para la selección de opciones mediante el ratón, típicamente usados en XHTML (ver Figuras 2 y 3). En la salida del sistema, la interacción visual se utiliza para proporcionar al usuario los datos obtenidos de las bases de datos (p. e. lista de libros disponibles relacionados con una determinada materia), así como información acerca del usuario actual (ver Figura 2).

2.1.3. Conexión entre ambas interfaces

Para crear la conexión entre ambas interfaces usamos manejadores de eventos que incluimos en la sección body de los documentos X+V. De esta forma, cuando el navegador carga el documento para realizar consultas bibliográficas, por ejemplo, se produce el evento on load y, como consecuencia, se ejecuta un formulario VoiceXML al que hemos llamado initial_vform para atender dicho evento. Para que una frase del usuario pueda rellenar varios campos de un formulario en una misma interacción, usamos una etiqueta del tipo: <vxml:initial name="initial_vform" … </initial> (típicamente empleada en VoiceXML) y empleamos una gramática a nivel de formulario. Así, mediante el documento para la consulta bibliográfica, el sistema puede generar el mensaje oral "Por favor, realice una consulta bibliográfica", y el usuario puede responder a este mensaje pronunciando un número variable de ítems de información, como por ejemplo: autores, autores y año de publicación, autores, año de publicación y materia, etc. También usamos en nuestra implementación eventos del tipo: ev:event="onclick", cuya finalidad es generar un evento conforme el usuario hace clic en un campo de un formulario. El manejador para este tipo de evento es un código escrito en VoiceXML para obtener el valor proporcionado por el usuario para dicho campo.

[3] http://www.w3.org/TR/voicexml2.0

2.2. Bases de datos

Para proporcionar información a los usuarios e interactuar con ellos adecuadamente, el sistema realiza consultas MySQL a varias bases de datos. La base de datos Perfiles de usuarios (ver Figura 1) contiene información acerca de los usuarios del sistema, incluyendo su nombre, sexo, dirección y número de teléfono. Esta base de datos también almacena tres tipos de preferencias: i) lenguaje para la interacción (Castellano o Inglés, de momento), ii) interacción oral (habilitada/deshabilitada) y iii) tipo de voz del sistema (masculina/femenina).

La base de datos Equivalencias idiomáticas almacena expresiones en diversos idiomas correspondientes a tipos concretos de frases que el sistema puede usar a lo largo de la interacción. La selección de un tipo de expresiones u otro se realiza en base al idioma seleccionado por el usuario en su perfil. Un ejemplo de tales expresiones es el mensaje de bienvenida que el sistema proporciona al usuario confirme éste accede al sistema: "Welcome to the DS-UCAT system" para la interacción en Inglés, o "Bienvenido al Sistema DS-UCAT" para la interacción en Castellano.

Adicionalmente, en nuestra implementación actual el sistema usa dos bases de datos de forma provisional, que en una implementación posterior deberán ser substituidas por bases de datos reales. Por una parte, la base de datos Libros disponibles en biblioteca almacena información relativa a fuentes bibliográficas supuestamente disponibles en la biblioteca del centro educativo. Por otra, la base de datos Libros prestados almacena datos de libros supuestamente prestados por la biblioteca a usuarios del sistema (profesores o estudiantes). Usando ambas bases de datos el sistema puede responder a preguntas como la mostrada en la Figura 3 para una interacción en Inglés. Empleando el formulario de la figura, cuando el usuario haciendo clic en un campo, el sistema genera un mensaje oral para solicitar al usuario que introduzca un dato para ese campo (p. e. "Book title?"). El usuario puede proporcionar este dato de forma oral o mediante texto. Tras consultar la base de datos Libros disponibles en biblioteca, el sistema puede generar oralmente un mensaje del tipo "The following books are available ….", o bien, el mensaje "No books were found" si no encontró ningún registro en la base de datos que satisfaga los criterios de búsqueda.

La base de datos Materiales de profesores es una base de datos pendiente de ser creada cuya finalidad será almacenar documentos proporcionados a los estudiantes por los profesores. Nuestro objetivo es que los documentos X+V para los entornos de trabajo Clase y Despacho de Profesor (también pendientes de ser creados) muestren una vista de esta base de datos, de forma similar a como lo hace el documento X+V mostrado en la Figura 3 usado para consultar la base de datos de Libros disponibles en biblioteca.

3. USO DEL SISTEMA DS-UCAT

Para poder interactuar con la versión actual del sistema el usuario debe comenzar por identificarse mediante un nombre de usuario y una clave. A partir de esta identificación el sistema determina el tipo de usuario de que se trata, consultando la base de datos Perfiles de usuarios. En función de las características del usuario, el sistema determina el idioma que debe usar durante la interacción, así como el uso (o no) de interacción oral, y el tipo de voz que debe emplear (masculina o femenina). Por ejemplo, la Figura 2 presenta la ventana mostrada inicialmente por el sistema para un usuario que seleccionó Inglés como idioma de interacción. Además, dado que este usuario indicó en su perfil que deseaba utilizar la modalidad oral, a la vez que el sistema muestra la pantalla de bienvenida, genera el mensaje oral "Hello Ramón, welcome to the DS-UCAT system".

Como se puede observar en la Figura, tras haberse identificado en el sistema, el usuario debe seleccionar el entorno de trabajo que desea utilizar (biblioteca, aula o despacho de profesor). Esta selección inicial le permite trabajar en un entorno que no es aquél en que se encuentra en un momento dado, lo que le permite, por ejemplo, realizar consultas bibliográficas desde un aula.

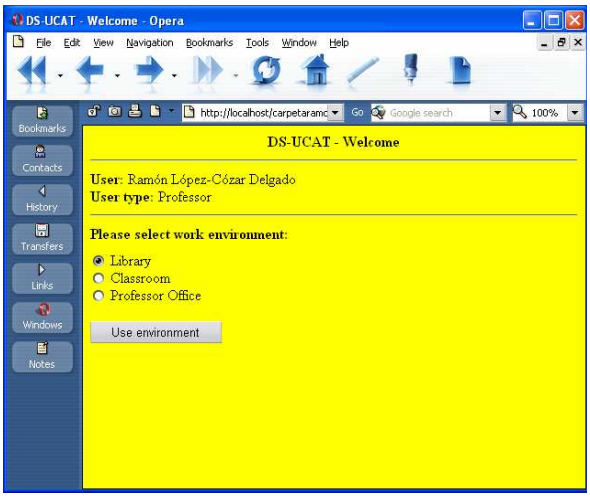

Figura 2. *Mensaje de bienvenida para interacción en Inglés.*

La Figura 3 muestra el documento X+V que aparece en pantalla conforme el usuario comienza a usar el entorno de trabajo Biblioteca. Dado que el usuario habilitó el uso de interacción oral en su perfil, el sistema general el mensaje oral "Please enter a book query". Para realizar la entrada mediante voz, el usuario debe mantener pulsada una determinada tecla o hacer clic en un determinado botón del navegador conforme pronuncia cada frase (técnica tap-&-talk). Además, puede efectuar la consulta proporcionado los datos de la consulta mediante el teclado y el ratón.

Hemos implementado en el sistema las dos estrategias de interacción típicamente empleadas en sistemas de diálogo: mixta y dirigida por el sistema. El uso de una u otra depende del tipo de documento X+V con el que interactúa el usuario en un momento dado. Por ejemplo, si la interacción se realiza con el documento diseñado para realizar consultas bibliográficas, el sistema utiliza interacción mixta para permitir el relleno de varios campos del formulario en una misma interacción (p. e. mediante la pronunciación de la frase "I need information about books written by Prieto in 2006"). Si la frase no puede ser reconocida correctamente, el sistema solicita datos al usuario con objeto de rellenar los campos del formulario uno a uno (iniciativa dirigida por el sistema).

El sistema tiene en cuenta tres tipos de eventos que suelen ocurrir en una interacción oral con un sistema de diálogo: i) que el usuario solicite ayuda, ii) que el usuario no pronuncie ninguna palabra cuando el sistema lo espera, y iii) que la entrada obtenida del usuario no pueda ser reconocida por ninguna gramática activa. Para procesar estos eventos, usamos contralores de eventos de tipo help, noinput y nomatch, respectivamente, de forma similar a como se usan en VoiceXML.

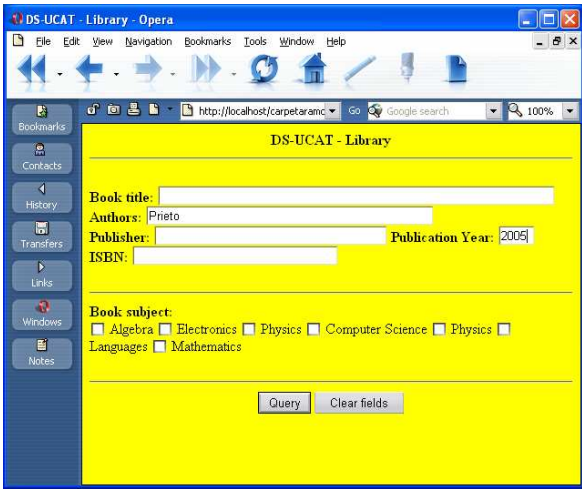

Figura 3. *Consulta bibliográfica para interacción en Inglés.*

4. INTERACCIÓN ENTRE EL SISTEMA DS-UCAT Y EL ENTORNO EDUCATIVO

Otra línea de trabajo dentro del proyecto UCAT (*Ubiquitous Collaborative Training*) está relacionada con la implementación de un sistema de identificación y localización automática de usuarios en el entorno educativo. Su finalidad es personalizar diversos servicios ofrecidos dependiendo del usuario y de la situación de éste dentro del entorno. Para implementar este sistema estamos utilizando una combinación de diversas tecnologías cuya integración permita obtener un conocimiento más preciso del contexto del usuario. Concretamente, por una parte hemos desplegado un

sistema de identificación y localización mediante radio frecuencia que permite conocer los lugares a los que acceden las personas. La infraestructura del sistema consta de un conjunto de antenas RFID (radio frecuencia) situadas en los marcos de entrada de las distintas estancias. Para ser localizado dentro del entorno, cada usuario dispone de una tarjeta RFID pasiva que le identifica unívocamente. Cuando una persona accede a una estancia, la antena correspondiente lee la información de la tarjeta RFID y comunica a un servidor central la nueva localización del usuario. Este mecanismo ha sido desarrollado y utilizado dentro de un entorno aplicado al hogar (ver Figura 4), lo que nos ha permitido comprobar su funcionalidad en condiciones reales para controlar oralmente lámparas, equipos de música y otros dispositivos [15]. Por otra parte, dado que puede no ser posible situar antenas en todas las estancias, también hemos desarrollado un sistema que permite detectar la actividad de las personas en su equipo de trabajo. De este modo se puede determinar la localización de una persona si se encuentra trabajando con su ordenador de la oficina.

Tenemos previsto hacer uso de estos mecanismos de localización e identificación en el sistema DS-UCAT con objeto de mejorar su funcionamiento. De esta forma, se evitará la necesidad de realizar una identificación inicial mediante nombre de usuario y contraseña para acceder al sistema, pues éste podrá determinar qué usuario está interactuando a partir de la información obtenida del servidor central de localización.

Figura 4. *Entorno de interacción para el hogar desarrollado en el proyecto Odisea.*

Además, se podrá adaptar automáticamente diversos aspectos del funcionamiento del sistema. Por ejemplo, el sistema de RAH quizás se pueda adaptar al usuario y al entorno en que éste se encuentra, empleando modelos acústicos y de lenguaje específicos para cada tipo de ubicación dentro del entorno educativo. También se podría adaptar la forma en que el sistema interacciona con el usuario, usando para ello diferentes modelos de diálogo, específicos para el lugar concreto dentro del centro educativo en que se encuentre el usuario en cada momento.

5. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo se ha centrado, fundamentalmente, en nuestra tarea dentro del proyecto UCAT relacionada con la implementación de un sistema de diálogo multimodal y multiligüe que proporcione asistencia a profesores y estudiantes en algunas de sus actividades cotidianas dentro de un centro educativo (p. e. en una Facultad de una Universidad). El artículo ha abordado la implementación actual del mismo, basada en documentos X+V. Algunos de estos documentos están almacenados en un servidor central, mientras que otros se generan dinámicamente en tiempo de ejecución mediante programas PHP. Este servidor interactúa con los dispositivos móviles de los usuarios mediante conexiones inalámbricas. A modo de ejemplo, hemos mostrado cómo se puede realizar la consulta multimodal de fuentes bibliográficas.

Las principales líneas de trabajo futuro están relacionadas con las siguientes tareas:

- Creación de los documentos X+V y las bases de datos necesarias para permitir la interacción con el sistema en los entornos de trabajo Aula y Despacho de Profesor.

- Desarrollo de la funcionalidad del sistema que permita interactuar con los dispositivos del entorno. Para ello tenemos previsto utilizar el gestor de diálogos Odisea (usado en el entorno del hogar citado anteriormente [15]), en el cual estamos trabajando actualmente para adaptarlo a los nuevos entornos de trabajo dentro del centro educativo [16]. De esta forma lograremos, por ejemplo, que los profesores puedan interactuar de forma multimodal y multilingüe con el proyector de transparencias, las luces del aula, o el equipo de música de su despacho.

6. BIBLIOGRAFÍA

[1] McTear, M. F. *Spoken Dialogue Technology: Toward the Conversational User Interface*, Springer, 2004.

[2] Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goodine, D., Goddeau, D., Glass, J. "PEGASUS: A spoken dialogue interface for on-line air travel planning". *Speech Communication*, 15, pp. 331-340, 1994.

[3] Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V. "Multilingual spoken-language understanding in the MIT Voyager system". *Speech Communication*, 17 (1-2), pp. 1-18, 1995.

[4] Danielli, M. “On the use of expectations for detecting and repairing human-machine miscommunication”. Working notes of the AAAI Workshop on Detecting, Repairing, and Preventing Human-Machine Miscommunication, pp. 87-93, 1996.

[5] López-Cózar, R., García, P., Díaz, J., Rubio, A. J. “A voice activated dialogue system for fast-food restaurant applications”. Proc. Eurospeech, pp. 1783-1786, 1997.

[6] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L. “Jupiter: A telephone-based conversational interface for weather information”. *IEEE Trans. on Speech and Audio Proc.*, 8(1), pp. 85-96, 2000.

[7] Seneff S., Polifroni J. “Dialogue management in the Mercury flight reservation system”. Proc. ANLP-NAACL Satellite Workshop, pp. 1-6, p. 2000.

[8] Torres, F., Sanchis, E., Segarra, E. “Learning of stochastic dialog models through a dialogue simulation technique”. Proc. Interspeech, pp. 817-820, 2005.

[9] Galibert, O., Illouz, G., Rosset, S. “Ritel : An open-domain, human-computer dialog system”. Proc. Interspeech, pp. 909-912, 2005.

[10] Callejas, Z., López-Cózar, R. “Implementing modular dialogue systems – A case of study”. Proc. ISCA Research Workshop on Applied Spoken Language Interaction in Distributed Environments, 2005.

[11] López-Cózar, R., Araki, M. *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. John Wiley & Sons Publishers, 2005.

[12] Coen, M. “Building brains for intelligent environments”. Proc. National Conference on Artificial Intelligence, 1998.

[13] Adler, A., Davis, R. “Speech and Sketching for Multimodal Design”. Proc. 9th International Conference on Intelligent User Interfaces, pp.214-216, 2004.

[14] Milward, D., Beveridge, M. A. “Ontologies and the structure of dialogue”. Proc. 8th Workshop on the Semantics and Pragmatics of Dialogue, 2004.

[15] Haya, P. A., Montoro, G., Alamán, X. “A prototype of a context-based architecture for intelligent home environments”, Proc. International Conference on Cooperative Information Systems (CoopIS 2004). pp. 477-491, 2004.

[16] Montoro, G., Haya, P. A., Alamán, X., López-Cózar, R., Callejas, Z. “A proposal for an XML definition of a dynamic spoken interface for ambient intelligence”, International Conference on Intelligent Computing (ICIC 06), pp. 711-716, 2006.