

El proyecto DIME y el robot conversacional *Golem*: Una experiencia multidisciplinaria entre la computación y la lingüística

Dr. Luis A. Pineda

Departamento de Ciencias de la Computación
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México

1. Introducción

Uno de los objetivos principales de la inteligencia artificial y la lingüística computacional es desarrollar programas de cómputo con la capacidad de sostener una conversación en lenguaje hablado, por ejemplo en español, con los seres humanos. 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. Este tipo de sistemas permitirán sostener conversaciones orientadas a resolver problemas sencillos pero cotidianos en diversos entornos problemáticos; serán de utilidad, por ejemplo, en la enseñanza, especialmente en ambientes virtuales; en servicios telefónicos a través de los llamados “call centers”, o en las industrias del turismo y el entretenimiento. La reducción del tamaño de los dispositivos computacionales así como el aumento de las capacidades de cómputo, tanto en velocidad procesamiento como en el tamaño de la memoria, y el aumento de la velocidad y disponibilidad de los equipos de comunicaciones, que continuará de manera acelerada por varios años todavía, brindarán el entorno tecnológico que hará factible la interacción con dispositivos tecnológicos de índole diversa a través del lenguaje hablado.

El escenario será diferente, muy probablemente, al que nos ofrecen las imágenes de la ciencia ficción, donde las máquinas superan la llamada “prueba de Turing” (Turing, 1950), es decir, donde hay robots humanoides que exhiben capacidades lingüísticas indiscernibles de las de los seres humanos. Sin embargo, en el mundo real no es de esperarse que existan sistemas computacionales capaces de establecer una conversación espontánea acerca de un tema arbitrario, o de otras formas complejas de expresión lingüística, como puede ser el lenguaje empleado en la transmisión de estados afectivos o la poesía. El lenguaje humano tiene diferentes niveles de expresión o de representación, como son el fonético y fonológico, el prosódico, el morfológico, léxico y sintáctico, el semántico y el pragmático, y aunque se conoce mucho de la estructura y funcionalidad de cada uno de estos niveles, hay todavía muchas preguntas abiertas acerca de su naturaleza y de cómo se relacionan entre sí; asimismo, la facultad lingüística humana se relaciona de manera muy compleja con la percepción, el pensamiento, la memoria, conducta motora y la conciencia; adicionalmente, la conversación siempre se da en un contexto espacial y temporal, en el que los interlocutores tienen deseos, creencia, conocimientos e intenciones, formando todo esto un contexto que da significado a los actos lingüísticos que se observan en la conversación. En resumen, el lenguaje humano es un objeto de gran complejidad, y no se conoce lo suficiente de su naturaleza para pensar que va a ser posible la construcción de máquinas parlantes que se desempeñen de manera comparable a los seres humanos, cuando menos en el horizonte de corto y mediano plazo, y probablemente tampoco en el de largo plazo.

Sin embargo, la experiencia en la construcción de máquinas conversacionales, y de modo más genérico, de las tecnologías del lenguaje, muestra que hay aspectos del lenguaje humano que sí se pueden modelar computacionalmente; mientras que desde el punto de vista de la ingeniería computacional la cuestión es qué tan generales y útiles podrán ser estas máquinas, desde el punto de vista lingüístico la pregunta es hasta qué punto la metáfora computacional permite explicar la naturaleza del lenguaje humano. Estas dos preguntas están entrelazadas y el desarrollo de sistemas conversacionales requiere de una investigación interdisciplinaria entre la ingeniería computacional y la lingüística.

Con el fin de abordar estas interrogantes, y tomando en cuenta el interés muy considerable que existe en la creación de sistemas conversacionales en el entorno internacional, tanto desde el punto de vista práctico como teórico, aunado a los esfuerzos muy limitados para el desarrollo de esta tecnología en la lengua española, en particular en el dialecto mexicano, a finales de 1998 se inició el proyecto Diálogos Inteligentes Multimodales en Español (DIME)¹ en el Departamento de Ciencia de la Computación del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM (DCC-IIMAS). Este proyecto tuvo como una de sus premisas originales que la información lingüística utilizada en las labores de análisis y la creación de modelos computacionales debe tener una base empírica, en la medida de lo posible, por lo que uno de sus objetivos iniciales fue la recolección de un corpus multimodal en un dominio problemático específico; como producto de esta tarea se obtuvo el corpus DIME², el cual se describe en la Sección 2 de este artículo. También siguiendo la tradición de las gramáticas generativas, que sugiere enfáticamente que la sintaxis es el nivel principal –y autónomo– de representación lingüística, y que el ingrediente básico para la interpretación del lenguaje es la inducción de la estructura sintáctica de las oraciones, otro objetivo inicial del proyecto fue la definición de una gramática computacional del español así como su sistema de *parseo* (Pineda y Meza, 2004). Sin embargo, el análisis del corpus DIME, así como la dificultad de inferir la intención de un agente a partir de la estructura sintáctica de la oración, en un contexto conversacional dado, nos llevó a cambiar la orientación del proyecto y adoptar al nivel pragmático e intencional como el nivel de representación central para analizar la comunicación lingüística. Con este nuevo enfoque y dada la necesidad de analizar los diálogos del corpus DIME desde el punto de vista de las intenciones expresadas por los agentes conversacionales, y tomando en cuenta la estructura de la conversación, se desarrolló el esquema DIME-DAMSL, el cual se describe brevemente en la Sección 3 (Pineda et al., 2007). Independientemente de que los diálogos en el corpus DIME son orientados a la solución de tareas específicas sencillas (i.e. distribuir el mobiliario de una cocina), éstos presentan una complejidad significativa, por lo que el modelo es todavía teórico y de carácter esencialmente lingüístico. Por otra parte, dado el interés práctico de construir sistemas conversacionales se desarrolló un esquema computacional para representar contextos conversacionales en dominios específicos, cuya interpretación permite llevar a cabo un diálogo sencillo entre un interlocutor humano y una computadora; estos esquemas, llamados “modelos de diálogo”, así como su sistema de interpretación se describen en la Sección 4.

El desarrollo de sistemas conversacionales requiere también contar con tecnología robusta pero flexible para el reconocimiento de voz. Los elementos básicos para la

¹ <http://leibniz.iimas.unam.mx/~luis/DIME/>

² <http://leibniz.iimas.unam.mx/~luis/DIME/CORPUS-DIME.html>

construcción de estos sistemas son los modelos acústico, los diccionarios de pronunciación y los modelos del lenguaje. Al inicio del proyecto se pensó que sería posible adoptar un sistema de reconocimiento de alguna fuente académica o de investigación; sin embargo, esto no fue posible y fue necesario crear una infraestructura propia; desde el punto de vista algorítmico se adoptó el paradigma de los modelos ocultos de Markov para la creación de modelos acústicos, así como los paquetes o *tool kits* para creación de reconocedores de voz genéricos, primero el sistema HTK³ y luego el sistema Sphinx⁴; sin embargo, para utilizar estas herramientas se requiere contar con una base de datos de información fonética con una base lingüística sólida así como de extensión suficiente, pero tampoco se encontró un recurso lingüístico de calidad disponible en nuestra lengua, por lo que fue también necesario crearlo; como producto de este esfuerzo a finales del 2003 y principios del 2004 se diseñó y recolectó el Corpus DIMEx100⁵; este recurso, además de proveer una fuente muy rica de información fonética del español de México, nos ha permitido crear una serie de reconocedores de voz con una calidad suficiente para ser utilizados en sistemas prototipos reales. La plataforma para la construcción de sistemas de reconocimiento, en la que el Corpus DIMEx100 tienen un papel central, se describe brevemente en la Sección 5.

Los diversos recursos tecnológicos y lingüísticos generados a lo largo de la vida del proyecto han madurado lo suficiente para permitirnos crear sistemas conversacionales sencillos pero robustos, y actualmente el proyecto cuenta con una infraestructura básica para la creación este tipo de sistemas. Con el propósito de demostrar esta tecnología en un entorno atractivo se adquirió un robot para fines de investigación, y se programó para que sirviera como guía de una sesión de carteles de los proyectos de investigación del DCC-IIMAS. En esta aplicación el robot, llamado *Golem*, explica dichos proyectos en español hablado en el contexto de una conversación con un usuario humano; el robot se ha demostrado exitosamente en varias ocasiones, tanto en nuestro laboratorio como en varios foros nacionales. En la Sección 6 se describen brevemente la arquitectura y funcionalidad del robot, y también cómo la infraestructura desarrollada en el proyecto DIME ha sido adaptada a *Golem* para que éste pueda sostener una conversación sencilla.

El artículo concluye con una breve reflexión acerca de lo que se ha logrado en los proyectos DIME y *Golem* con énfasis en el trabajo interdisciplinario entre ingenieros computacionales y lingüistas, que ha sido fundamental para lograr nuestros resultados.

2. El Corpus DIME

El Corpus DIME es un recurso lingüístico consistente en un conjunto de diálogos o conversaciones espontáneas, en español hablado, orientadas hacia la solución de una tarea simple de diseño, en la que un cliente (“sujeto” o “usuario”) distribuye el mobiliario de una cocina con la asistencia de un vendedor (“sistema” o “mago”)⁶, quien es un experto en esta clase de diseños. Como en muchas tareas de diseño las conversaciones se apoyaron o

³ <http://htk.eng.cam.ac.uk/>

⁴ <http://cmusphinx.sourceforge.net/html/cmusphinx.php>

⁵ <http://leibniz.iimas.unam.mx/~luis/DIME/CORPUS-DIMEx100.html>

⁶ La terminología se adopta de los experimentos llamados del “Mago de Oz” de interacción hombre-máquina en los que un sujeto interactúa con una computadora a través de una interfaz sin saber que las acciones del sistema son producidas realmente por un ser humano; sin embargo, en nuestro experimento los sujetos sabían que estaban interactuando con un experto humano y no con un programa de cómputo.

contextualizaron con la ayuda de diagramas. Para crear y modificar estos diagramas se adaptó un programa de diseño comercial⁷, por lo que no fue necesario programar la interfaz ni las herramientas de computación gráfica interactiva. La interfaz gráfica utilizada, así como un ejemplo de una situación particular de diseño, se muestran en la Figura 1.

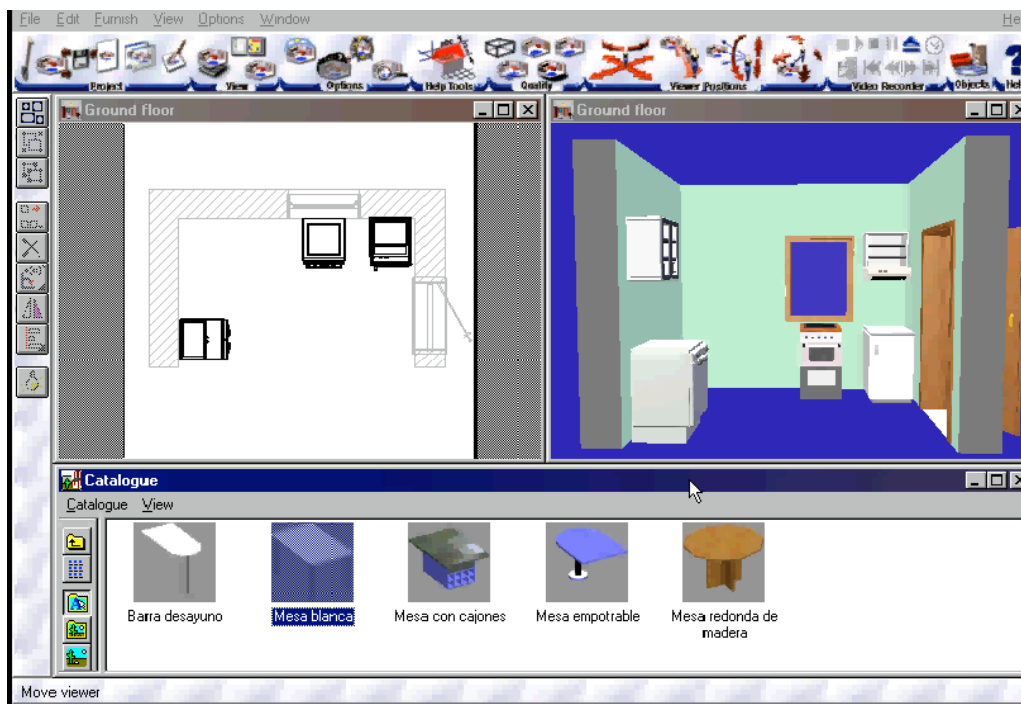


Figura 1. Interfaz para la recolección del Corpus

La interfaz consta de tres partes principales: una vista de planta de la cocina (arriba a la izquierda), una vista tridimensional (arriba a la derecha) y un catálogo de muebles (ventana horizontal interior) con los tipos e instancias de muebles disponibles para incluirse en un diseño particular. La tarea consistía en seleccionar los muebles, así como su disposición espacial, hasta llegar a un diseño que fuera satisfactorio para el cliente y que a su vez cumpliera con un conjunto de restricciones de diseño que eran del conocimiento del mago.

En el experimento se pidió a los interlocutores hablar de la manera más natural posible, y no se especificaron limitaciones para el lenguaje que se podría emplear a lo largo de la sesión de diseño; por lo mismo, los fenómenos típicos del habla espontánea, como interjecciones, pausas, repeticiones, reparaciones, habla simultánea o traslapes entre interlocutores, etc., ocurrieron de manera natural y con mucha frecuencia; asimismo, sólo un conjunto muy reducido de elocuciones fueron “gramaticales” (i.e. oraciones “bien formadas” que expresan una proposición completa y significativa); de hecho, frases nominales o complementos circunstanciales sueltos, así como interjecciones con diversas funciones, además de muchas ocurrencias de frases e incluso palabras incompletas, ocurren frecuentemente. Un ejemplo de un segmento de diálogo (muy simple) se muestra a continuación en la Tabla 1.

⁷ Home Designer, Alpha Software Corp. and Data Becker GmbH & Co KG.

# Eloc	Participante	Elocución
25	Sujeto	<i>Después <sil> ¿me puedes poner <sil> e[l] el extractor de aire encima de la <sil> de la estufa?</i>
26	Mago	<i>okay</i>
27	Mago	<i>¿así está bien?</i>
28	Sujeto	<i>si <sil> así está bien</i>

Tabla 1. Ejemplo de transacción básica

Las cuatro elocuciones de la Tabla 1 constituyen una transacción de diseño en la que se incluye un extractor de aire en la cocina. Este tipo de lenguaje va acompañado normalmente de acciones motoras, como la inclusión del extractor en el diagrama entre las elocuciones 26 y 27 proferidas por el mago. Hay también un componente indexical significativo en el que, por ejemplo, los demostrativos como *así* en 27 hacen referencia a acciones motoras que se extienden de manera espacial y temporal. De manera más general, el corpus DIME exhibe un lenguaje rico en expresiones espaciales y de conductas lingüísticas dirigidas a la resolución de referencias vagas o ambiguas, por lo que es un recurso muy útil para el estudio del lenguaje espacial.

Para recopilar el corpus se utilizó un escenario con dos cuartos separados, uno para cada uno de los interlocutores, por lo que estos no tenían contacto directo visual o auditivo, como se muestra en la Figura 2.

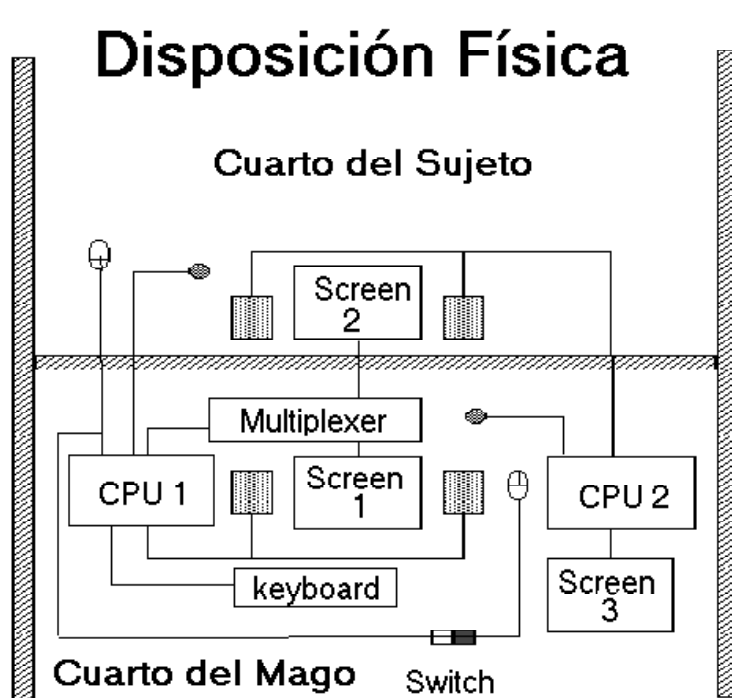


Figura 2. Escenario para recolección del Corpus DIME

Cada cuarto contaba con un monitor, así como de aditamentos para grabar la voz de ambos interlocutores; ambos monitores estaban conectados por un multiplexor al mismo CPU, por lo que tanto el sujeto como el mago veían la misma imagen a lo largo de todo el diálogo. Los interlocutores podían también apuntar a objetos o zonas de la pantalla

mediante un ratón y, al igual que los monitores, ambos ratones estaban conectados al mismo CPU; sin embargo, se contaba con un interruptor (bajo el control del mago) por medio del cual se daba el control del ratón a quien tenía el turno conversacional. En este escenario fue posible grabar el video y el audio de estos diálogos, así como los actos de apuntar; estos archivos digitales constituyen la materia prima del corpus DIME. Como producto de este experimento se obtuvieron 26 diálogos útiles grabados por 15 sujetos; cada diálogo tiene una duración promedio de 14 minutos, 185 elocuciones en 115 turnos y 880 instancias de palabras producidas entre los dos hablantes. Asimismo, el léxico utilizado en dominios conversacionales específicos es normalmente limitado y aunque el corpus consta de aproximadamente 30,000 instancias de palabras, tan sólo se utilizaron 1,120 tipos entre todos los hablantes a lo largo de todos los diálogos.

Una vez que fue recopilado el recurso básico se procedió a transcribirlo o etiquetarlo en varios niveles de representación lingüística; la primera tarea consistió en hacer la transcripción ortográfica y segmentar el corpus en elocuciones; dado que en el lenguaje hablado no existe un concepto análogo a la oración gramatical del lenguaje escrito, se consideró que una elocución es una unidad lingüística a nivel intencional y entonativa o prosódica. La tarea de transcripción fue realizada por varios estudiantes tanto de ingeniería computacional como de lingüística.

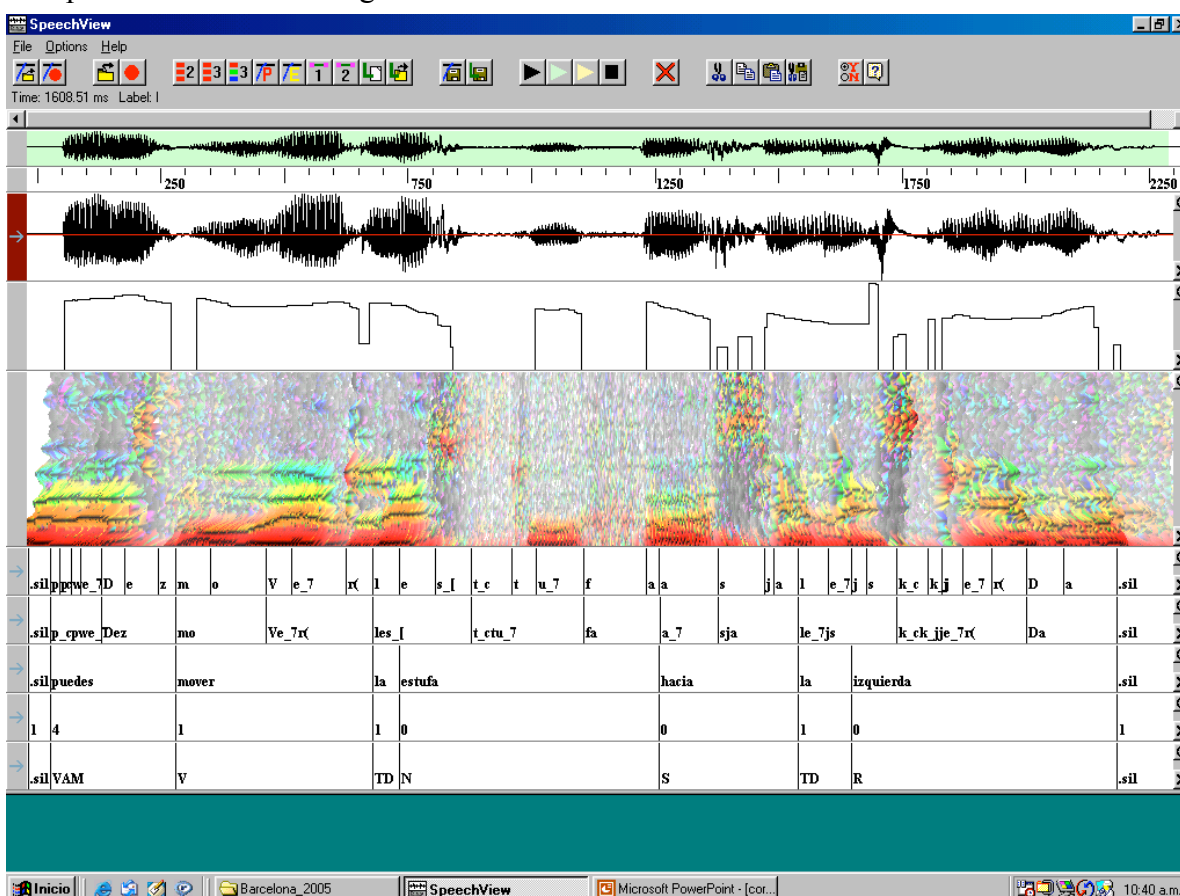


Figura 3. Niveles de transcripción del Corpus

Posteriormente se procedió a transcribir el corpus a nivel fonético y fonológico, prosódico y entonativo, léxico y sintáctico (partes del habla) y finalmente a nivel

pragmático e intencional. En la Figura 3 se muestra la transcripción alineada temporalmente de la elocución *puedes mover la estufa hacia la izquierda* en los niveles fonético, de sílabas fonéticas (que difieren de las sílabas fonológicas canónicas del habla escrita), léxico, prosódico a nivel de la separación de las unidades léxicas (donde “0” indica sinalefas, “1” separación normal entre palabras y “4” separación de unidades entonativas completas o elocuciones); finalmente se muestra la categoría léxica de las palabras en la elocución. Adicionalmente, en la parte superior, se muestra el oscilograma de la elocución, la curva de la frecuencia fundamental, así como el espectrograma, en el cual se pueden distinguir claramente las zonas los formantes de los sonidos vocálicos, así como los silencios y ruidos de los consonánticos. La herramienta principal para realizar estas transcripciones fue *SpeechView* del *CSLU Tool Kit*⁸.

Uno de los temas de mayor interés en el contexto del proyecto es la relación entre la entonación de un enunciado y la intención que éste expresa; por lo mismo, se realizó también la transcripción de una parte muy significativa del corpus bajo el esquema INSINT (Hirst et al., 2000); en la Figura 5 se muestra un ejemplo de la curva de la frecuencia fundamental de un enunciado con los tonos INSINT asociados a los puntos de inflexión de dicha curva.

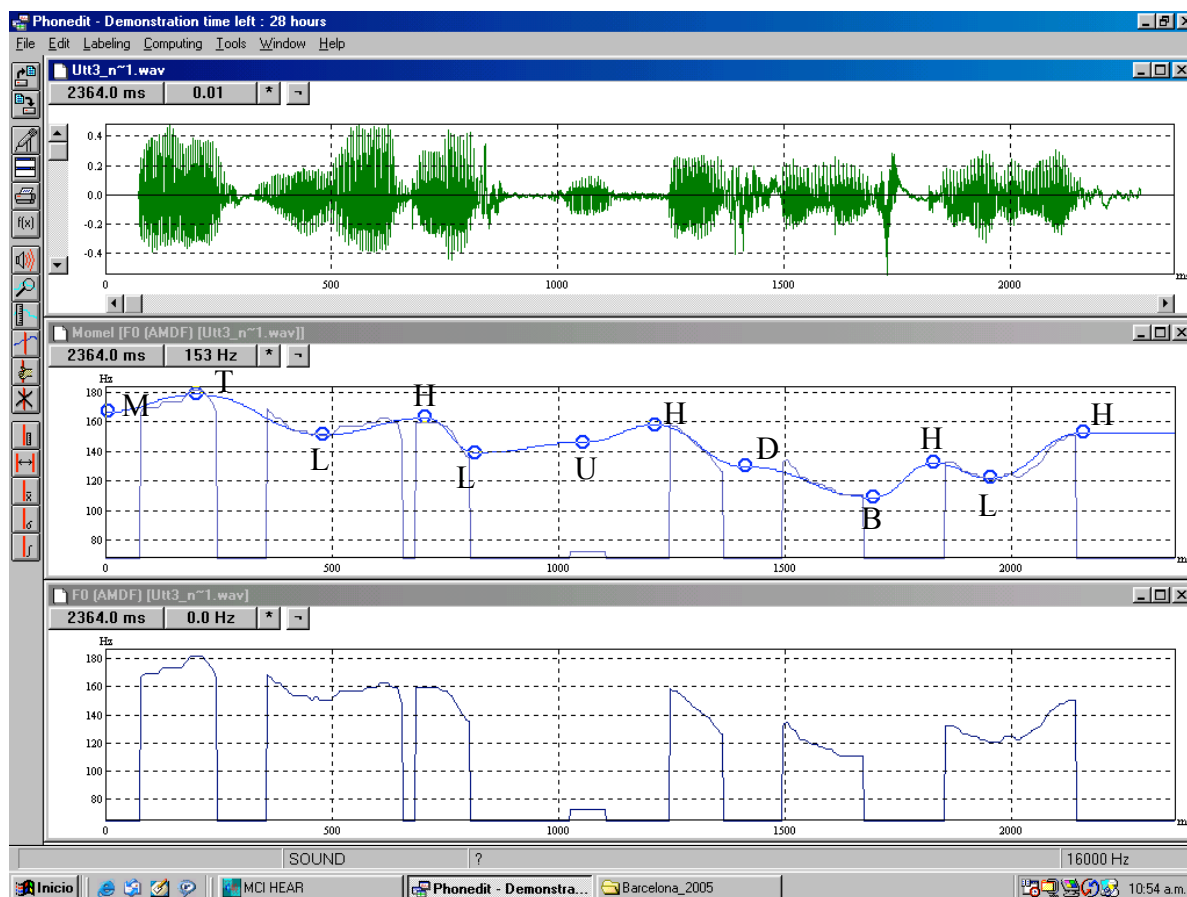


Figura 4. Transcripción entonativa de acuerdo al esquema INTSINT

⁸ <http://www.cslu.ogi.edu/toolkit/>

En esta representación la curva entonativa está referida a los y tonos inferior (*Bottom*), medio (*Medium*) y superior (*Top*), a los puntos de inflexión inferiores (*Low*) y superior (*High*) y a las microinflexiones ascendentes (*Up*) y descendentes (*Down*); asimismo, la secuencia temporal de estos tonos (e.g. la cadena *MTLHLUHDBHLH*) es una representación discreta del perfil entonativo de la elocución. Existen herramientas para hacer esta transcripción de manera automática (Hirst et al., 2000), y se cuenta con la cadena tonal correspondiente a una parte muy significativa del corpus. La estructura de estas cadenas puede asociarse de manera muy directa con la modalidad del enunciado, y también contribuye, aunque en menor medida, a predecir el tipo de acto de habla (e.g. directiva de acción, solicitud de información, etc.) expresado por el mismo (Coria y Pineda, 2008).

El Corpus DIME incluye también la transcripción de las disfluencias y reparaciones incurridas por los hablantes en el habla espontánea. Una reparación típica se muestra en la elocución 25 en la tabla 2; en *Después* <sil> ¿me puedes poner <sil> e[l] el extractor de aire encima de la <sil> de la estufa? se observan dos reparaciones; en la primera se repite el artículo *el* y la reparación consiste en descartar la primera instancia, y lo mismo sucede con la secuencia *de la* en el complemento locativo final. Las reparaciones del habla se pueden analizar con cierto grado de generalidad en tres partes: el *reparandum*, es decir la parte que se enuncia pero se descarta, el punto de interrupción, donde se inicia el proceso de reparación, y la alteración, que substituye o modifica al *reparandum*; frecuentemente se observa también un pausa o una interjección después del punto de interrupción, que ocurre, probablemente, para que el locutor formule o “piense” la alteración. Por ejemplo en *puedo eh puedes borrar el fregadero* el punto de interrupción ocurre inmediatamente después del *reparandum* *puedo*, mientras que la alteración *puedes*, sigue la interjección *eh*, que en este ejemplo es un término de edición. En el contexto del proyecto se realizó un análisis de estos fenómenos y se propuso un algoritmo para hacer las reparaciones más frecuentes (Moreno y Pineda, 2006).

La noción de elocución adoptada permitió llevar a cabo el proceso de transcripción; sin embargo, a la luz de los análisis posteriores, resultó limitada y repercutió negativamente en el análisis de las intenciones o actos del habla. En particular, la unidad intención y prosodia adoptada para la segmentación no es suficientemente precisa debido a los diferentes tipos de cláusulas que pueden constituir una elocución, incluyendo las interjecciones, ya que todas éstas pueden ser consideradas unidades intencionales completas si se enuncian en un contexto apropiado. Por esta razón es posible confundir los límites de la elocución y en el análisis de corpus frecuentemente se segmentó en una elocución en dos o más partes, y también se dio el caso de que dos o más elocuciones genuinamente diferentes se segmentaran como una sola. Por lo mismo, una solución más general al problema de cómo segmentar en elocuciones el habla espontánea debiera considerar también la estructura de las reparaciones del habla, ya que tomando en cuenta este fenómeno, la elocución se puede conceptualizar como una unidad prosódica e intencional que contiene una o más reparaciones del habla; bajo este criterio se puede considerar que una elocución “normal” tiene también la estructura de una reparación, pero tal que el *reparandum* y el término de edición son vacíos, el punto de interrupción ocurre al inicio del enunciado y la alteración corresponde al enunciado mismo.

3. El Esquema DIME-DAMSL

El contexto básico para la adquisición y uso del lenguaje es la conversación espontánea cara a cara entre dos personas; esta conducta tiene por objetivo que los interlocutores se comuniquen entre sí. Para que la comunicación sea exitosa se requiere que el escucha interprete la elocución de manera consistente con las intenciones del hablante; esto es, que el hablante emita una elocución con cierta intención y que el escucha infiera dicha intención a partir de la elocución misma y del contexto en que ésta se realiza. El ciclo de la comunicación requiere también que el escucha haga ver al hablante que entendió la intención expresada y qué va a hacer al respecto. Desde este punto de vista una elocución es el vehículo de una acción que se realiza por medio del lenguaje o, de acuerdo a la terminología tradicional, un acto del habla (e.g., Levinson, 1983, 2000). De manera intuitiva los actos del habla pueden clasificarse de acuerdo al tipo de acción que realizan; por ejemplo, su función puede ser presentar o solicitar información, requerir que el interlocutor realice cierta acción motora o mental (e.g. que tome nota de cierta información y la guarde en su memoria), etc.

Los actos del habla pueden también ser pensados como un par consistente en el tipo de acción que el hablante requiere que el escucha haga con la información o proposición expresada por una elocución, y la proposición misma. Por ejemplo, la elocución 25 en la Tabla 1 se interpreta, en el contexto del diálogo, como que el hablante *H* requiere que el escucha *E* haga *P*, donde *P* es *poner el extractor de aire encima de la estufa*. Para que esta interpretación sea posible las identidades de *H* y *E* deben ser accesibles o cuando menos inferibles a partir del contexto conversacional, y la situación debe ser tal que *H* debe estar en condiciones de requerir a *E* que haga la acción nombrada y *E* tiene la obligación de atender a dicho requerimiento. Para fines computacionales y para la modelación de los llamados diálogos prácticos (Allen, et al., 2000, 2001) se asume que los predicados de este tipo (i.e. *requerir*) son un conjunto finito y probablemente pequeño; en el presente contexto a los actos o acciones genéricas nombradas por estos predicados se les denomina “tipos de actos del habla”; asimismo, un dominio conversacional se especifica por un conjunto de estos tipos de actos, las condiciones bajo las cuales se pueden llevar a cabo las realizaciones concretas de estos tipos de actos, y por un conjunto de restricciones sobre sus argumentos; de manera más coloquial el tipo de acto del habla especifica las instrucciones que debe seguir *E* en el proceso de la proposición *P*. Asimismo, “entender” se puede conceptualizar como la inferencia realizada para obtener dicha instrucción con su predicación asociadas a partir de la elocución misma y del contexto conversacional.

Cada tipo de acto del habla se pronuncia con una forma prosódica o entonación preferida; por ejemplo, las afirmaciones se pronuncian de manera declarativa, las solicitudes de información de manera interrogativa y las directivas de acción de forma imperativa. Sin embargo, frecuentemente se observa que un tipo de acto del habla se expresa con una entonación diferente a la esperada; por ejemplo, una directiva de acción puede expresarse en forma interrogativa, como la elocución 25 en la Tabla 1; a los actos del habla que utilizan una “fuerza” expresiva diferente a la básica se les denomina como “actos del habla indirectos”. Los actos del habla indirectos también pueden involucrar una reinterpretación del significado léxico de los términos utilizados; por ejemplo, si *H* dice a *E*, quien va manejando y tiene su ventanilla abierta, *¿podrías cerrarle a tu frijolito?* con la intención de que *E* cierre la ventana, no sólo cambia la modalidad (interrogativa en vez de imperativa) sino también el significado de léxico del complemento directo (i.e., *frijol* por

frío como metáfora de *ventana*, por asociación fonética, y además en diminutivo!). Aunque este ejemplo es un poco extremo (aunque oído recientemente en una conversación espontánea), ilustra de manera muy clara que la interpretación de los actos del habla indirectos es una inferencia sumamente compleja, que depende del contenido léxico y sintáctico de la elocución misma, pero también y de manera crucial, del contexto en que se expresa.

Un factor adicional que complica el proceso de interpretación de los actos del habla es que una elocución puede expresar o realizar más de un acción lingüística al mismo tiempo. Por ejemplo, la elocución 26 en la Tabla 1, *okay*, puede ser interpretada como que el mago hace el compromiso de realizar la directiva de acción expresada por el usuario, pero también simplemente que entendió y acepta dicha expresión. Estas dos acciones son diferentes en varios niveles o planos de expresión: con el acto de comprometerse se adquiere la obligación de realizar la acción, y ésta, en caso de que se lleve a cabo, se realiza de manera posterior al compromiso mismo; por otro lado, aceptar un acto del habla se dirige al discurso previo e implica entender el mensaje y decidir si éste es compatible con lo que ya se ha dicho en la conversación y también con el contexto. En oposición al plano de las obligaciones, en el que se expresan las acciones motoras o lingüísticas que son el propósito global de la comunicación, podemos pensar, cuando menos para fines de análisis, que hay un plano de expresión que tiene por objetivo establecer un “acuerdo mutuo” entre los interlocutores y permitir que estos “estén enganchados” durante este proceso. A su vez, este plano se puede dividir en un sub-plano de acuerdo propiamente y un sub-plano de la comunicación, en el que ocurren actos del habla cuyo único propósito es asegurarse de que los mensajes lingüísticos fluyen normalmente, como las interjecciones que pronuncia el escucha para comunicarle al hablante que le está entendiendo correctamente (e.g. *aja*), o los mensajes para reforzar la comunicación (e.g. *¿me entendiste? ¿me lo repites por favor?*).

Para efectos de nuestra investigación, la noción de “contexto” se entiende en un sentido muy amplio, es decir que la conversaciones se da entre un hablante y un escucha, o se hace referencia a una tercera persona, ya sea singular o plural, en una situación espacial y temporal, y en relación a un conjunto de conceptos relevantes al tema del que se habla, que son compartidos por los interlocutores en cierta medida, y también en relación a un conjunto de creencias, deseos e intenciones que tienen los interlocutores, que también pueden ser compartidos parcialmente, pero que también pueden estar en conflicto. Adicionalmente, los actos del habla no se dan de manera aislada y normalmente están relacionados con los actos del habla proferidos previamente en la conversación y condicionan también los actos que se pueden seguir en la misma. Por ejemplo, una solicitud de información va seguida normalmente de una respuesta y un compromiso se salda cuando se realiza el acto comprometido. Todos estos factores contribuyen al contexto conversacional.

Asimismo, la conducta lingüística es frecuentemente esquemática y sigue protocolos compartidos por comunidades específicas; estos protocolos son apropiados para resolver problemas comunicativos cotidianos y se estructuran en términos de ciertos tipos de actos del habla y del orden preferencial en que estos ocurren. Por ejemplo, en la Tabla 1 se muestra un protocolo o transacción que tiene por objetivo realizar una acción motora (poner el extractor encima de la estufa en el diagrama); las transacciones en el Corpus DIME se componen normalmente de la especificación de una intención, seguida de su satisfacción; en el ejemplo la especificación consiste de la directiva de acción en 25 y la aceptación/compromiso en 26. La satisfacción por su parte se da en 27 y 28; en 27 el mago

realiza la acción motora solicitada a la vez que realiza una solicitud de información con el fin de confirmar que la acción realizada satisface la intención expresada, y en 28 el usuario realiza una respuesta/afirmación con la confirmación requerida. La transacción de ejemplo es muy básica y se realiza de manera directa, sin embargo, frecuentemente se requieren varias elocuciones y turnos conversacionales para realizar cada una de las dos fases, dando lugar a conductas lingüísticas sumamente complejas.

Se puede observar también que los tipos de actos de habla empleados son apropiados al punto de la transacción en el que ocurren y existen pocas alternativas para expresar las cosas de manera diferente. Por ejemplo, la transacción puede iniciarse con una oferta por parte del mago en vez de la directiva de acción expresada por el usuario, y pueden existir algunos patrones adicionales para iniciar la transacción, pero no son muchos; también, como segundo ejemplo, los compromisos ocurren comúnmente al final de la fase de especificación, una vez que se ha interpretado correctamente la intención del usuario y la acción deseada queda totalmente determinada, pero es muy poco probable que un compromiso ocurra al inicio de la transacción, cuando dicha información está todavía por especificarse (aunque se dan casos, por supuesto!). Adicionalmente, la expresión de ciertos tipos de actos del habla abre contextos internos en la transacción que se satisfacen o se cierran con la expresión otros actos de carácter específico; por ejemplo, una solicitud de información se tiene que satisfacer por una respuesta, y una directiva de acción por la acción solicitada. De modo más general, los protocolos conversacionales están conformados no sólo un conjunto de actos del habla que ocurren en cierto orden, sino también por un conjunto de relaciones entre los ellos. Un protocolo completo o “balanceado” es aquel en el todos los actos del habla que lo constituyen se han realizado, y las relaciones entre dichos actos se satisfacen. Por todo esto, el conocimiento del protocolo conversacional, o de la estructura de la transacción, es sumamente importante para determinar el tipo de acto de habla expresado por una elocución; sin este conocimiento es muy difícil de explicar una buena parte de la conducta lingüística.

Tomando en cuenta estas consideraciones y elaborando sobre el esquema de análisis de actos del habla llamado DAMSL (Allen et al., 1997), en la presente investigación se desarrolló el esquema DIME-DAMSL (Pineda et al., 2006, 2007). El análisis de una transacción típica bajo este esquema se ilustra en la Tabla 2. Las columnas en la tabla representan respectivamente el número de elocución dentro de la transacción, el participante que emite la elocución (*Mago* o *Usuario*), la contribución de la elocución en los planos obligaciones y acuerdo mutuo, que se divide en acuerdo y entendimiento, y los tipos de actos de habla expresados por la elocución en ambos planos. Adoptando la terminología contable las contribuciones se dividen en cargos (*crg*) y créditos o abonos (*cdt*); los actos del habla que “hacen cargos” crean expectativas conversacionales que deben ser satisfechas por los actos del habla que hacen los “créditos” correspondientes. Los dígitos en las celdas de las columnas de cargos y abonos indican la elocución que expresa el acto del habla que hace el cargo y las llaves verticales indican la relación con la elocución que expresa el acto del habla que lo acredita, para cada uno de los tres planos de expresión.

La noción de aportación está motivada por la intuición de que la conducta conversacional obedece a un principio de conservación, donde un acto del habla que hace un cargo expresa una necesidad del agente conversacional que debe ser satisfecha por el acto del habla que lo abona, para que el agente recobre su estado de equilibrio. Como muchas otras unidades de representación lingüística las acciones que realizan los actos

tienen una fase tensiva que se resuelve con otros actos que disuelven dicha tensión, retornando al agente a su línea base o de “reposo” conversacional.

#	T	Elocución	Obligaciones		Acuerdo Mutuo				Tipos de actos del habla	
					Acuerdo		Entendimiento		Obligaciones	Acuerdo mutuo
			Crg	Cdt	Crg	Cdt	Crg	Cdt		
1	M	Quieres que traiga algún mueble a la cocina?			1					oferta
2	U	Si	1			1			oferta	aceptar
3		Necesito una estufa	3		3				directiva de acción	directiva de acción
4	M	Un segundo				3				aceptar
5		Estos son los modelos de estufas que tenemos, estufas sencillas y estufas con alacenas laterales			5					Opción abierta
6	U	Mmmm <sil> voy a seleccionar esa estufa			6					aceptar afirmar
7	M	Okay				6				aceptar
8	U	eh, por favor la necesito en <sil> en la pared del fondo			8		8			afirmar ambigua
9	M	¿cuál es la pared del fondo?	9						sol. de inf	espera
10	U	A ver, aquí		9	10				respuesta	afirma apunta- zona
11	M	¿ahí?	11						sol. de inf	espera
12	U	si		11				8	respuesta	resuelve
13	M	Un segundo	13			10 8			compromiso	aceptar
14		<acción gráfica: agrega objeto>		13 3 1	14				action-graf	afirma
15		¿ahí está bien?	15						sol. de inf	
16	U	Si, de momento, si		15		14			respuesta	aceptar

Tabla 2. Análisis de transacción bajo el esquema DIME-DAMSL

De manera más general, el esquema DIME-DAMSL postula que un dominio conversacional se define en función de un conjunto de tipos de actos del habla, de la relación de equilibrio entre las instancias o realizaciones de los mismos, y de un conjunto de restricciones tanto sobre los argumentos de los tipos de actos como sobre la relación de equilibrio. El esquema postula también que los diálogos prácticos se estructuran en

términos de transacciones, y que una transacción es completa y balanceada cuando se han llevado a cabo las fases de especificación y satisfacción de la intención, y se satisface la relación de equilibrio para todos los actos del habla que ocurren en la misma.

Pasamos ahora a ilustrar el funcionamiento del esquema DIME-DAMSL mediante la descripción de la estructura de la transacción en la Tabla 2. La especificación de la intención se realiza de la elocución 1 a la 13 y la satisfacción de la 14 a la 16. En la fase de especificación *U* expresa una directiva de acción así como sus argumentos a partir de la oferta inicial expresada por *M*, mientras que la satisfacción consiste en la realización de la acción gráfica por parte de *M* y la confirmación de que la acción realizada es la deseada por parte de *U*. La aportación de cada acto del habla en la transacción es como sigue: la elocución 1 expresada por el mago realiza una “oferta” que inicialmente hace un cargo en el plano del acuerdo mutuo, ya que *M* espera que *U* tome nota de la misma. En 2 el usuario *U* acepta la oferta y abona el cargo de acuerdo realizado en 1, satisfaciendo de este modo la relación de equilibrio en el plano del acuerdo mutuo, como lo indica la llave correspondiente; por otra parte, al aceptarse la oferta en 2, *M* adquiere la obligación de realizar la acción ofertada, por lo que se hace el cargo del acto expresado en 1 en el plano de las obligaciones. Mientras que algunos actos del habla hacen un cargo en las obligaciones de manera inmediata, como las directivas de acción, otros tipos de actos sólo hacen un cargo si se aceptan por el interlocutor, como en este caso. En 3 *U* expresa una “directiva de acción” (i.e., un comando) que crea un cargo sobre *M* en ambos planos; por un lado *M* debe realizar la acción solicitada, pero también debe tomar nota de que entendió la directiva y que está dispuesto a actuar en consecuencia, es decir que la acepta. Asimismo, mientras algunos actos crean un cargo en el plano de las obligaciones sobre quien los realiza, como la oferta, otros hacen el cargo sobre el interlocutor, como en el caso de las directivas de acción; la información de quien debe abonar un cargo para cada tipo de acto del habla se especifica en la relación de equilibrio. En 4 *M* expresa que entendió lo que se le está solicitando y “acepta” realizar dicha acción, abonando de este modo el cargo sobre el plano de acuerdo expresado en 3.

En este punto, a pesar de que *M* entendió y aceptó lo que se le solicita, es decir, el tipo de acto del habla expresado (i.e. directiva de acción) así como la proposición asociada (i.e. poner una estufa), no le es posible realizar dicha acción ya que es todavía necesario especificar qué estufa en particular se requiere y dónde debe colocarse; es decir, es necesario especificar los argumentos de la acción “poner”. Con el fin de obtener esta información *M* toma la iniciativa conversacional y expresa la opción abierta en 5. Este acto del habla no genera ninguna obligación en *U*, ya que la opción abierta sólo presenta un conjunto de alternativas o un “menú de opciones” sin que se requiera seleccionar ninguna de éstas necesariamente (i.e. *U* puede simplemente ignorar las opciones implícitamente pasando a otra cosa, o rechazarlas explícitamente), pero sí genera un cargo sobre *U* en el plano de acuerdo ya que éste debe dar acuse de recibo de las opciones que se le están ofreciendo. En 6 *U* “acepta” la opción abierta y especifica mediante una expresión deíctica, apoyada por un acto de apuntar, cuál es la estufa que desea (seleccionándola del catálogo); es decir, realiza una “afirmación”, la cual se acepta por *M* en 7.

En 6 y 7 se completa la especificación del argumento acusativo de la acción de poner, pero es todavía necesario especificar el argumento locativo, y esto se lleva a cabo entre 8 y 13. En 8 *U* “afirma” dónde se requiere colocar la estufa, a la cual se hace referencia de forma anafórica mediante el clítico *la*; sin embargo, la referencia espacial es ambigua ya que en la representación gráfica hay varios referentes potenciales para *la pared del fondo*.

Este tipo de expresiones ocasionan una falla en la comunicación: si el intérprete no es capaz de resolver una ambigüedad referencial se rompe momentáneamente el acuerdo mutuo y éste se tiene que reestablecer para que la conversación pueda proseguir; de otro modo los interlocutores estarían hablando de cosas diferentes, probablemente sin notarlo. Aunque la ambigüedad no es un acto del habla propiamente, esta propiedad de la afirmación se marca también como una etiqueta. Por estas razones 8 hace dos cargos de acuerdo mutuo, uno en el sub-plano del acuerdo propiamente y otro, causado por la ambigüedad, en el sub-plano del entendimiento. En 9 *M* hace una “solicitud de información” para resolver cuál es el referente de *la pared del fondo*; este acto del habla requiere una respuesta y hace un cargo en el plano de las obligaciones; por otro lado, la pregunta en sí expresa que la afirmación ambigua en 8 no se acepta y que el acto de aceptarla se pospone o se pone en “espera”. Al igual que la ambigüedad no es un acto del habla sino una propiedad del acto que la expresa, la espera tampoco es un acto del habla propiamente, sino una propiedad de este tipo de solicitud de información (i.e. pregunta de clarificación), y recibe también una etiqueta (i.e. “espera”). Por lo mismo las llaves del acuerdo que se inician en 8 en la Tabla 2 no se cierran en 9, sino que se extienden hasta que se resuelve la ambigüedad, se acepta la afirmación y se reestablece el acuerdo mutuo⁹.

La “respuesta” de *U* en 10 hace el abono en el plano de las obligaciones a la pregunta en 9, como se indica con la llave correspondiente. Esta respuesta es también una afirmación, que se expresa con una expresión deíctica, y hace un cargo en el plano de acuerdo, ya que *M* tiene que dar acuse de recibo (i.e. identificar la pared del fondo mediante el punto señalado asociado al demostrativo *aquí*). En este caso la acción de apuntar con el cursor se realiza sobre una zona de la pantalla y se etiqueta con “apunta-zona”. En 11 *M* hace una nueva solicitud de información (i.e. la pregunta si/no ¿*ahí*?) con el cargo correspondiente en las obligaciones; esta pregunta expresa a su vez un nuevo “espera” que pospone la aceptación de 10. En 12 *U* responde afirmativamente con lo que se abona el cargo en las obligaciones; también, con esta afirmación se resuelve la ambigüedad espacial introducida en 8, abonando de este modo el cargo pendiente en el plano del entendimiento, como se indica con la llave que va de 8 a 12 en este plano; resolver la ambigüedad es una propiedad de la afirmación que, como en los casos anteriores, se marca con la etiqueta “resuelve” en el acuerdo mutuo. Finalmente, *M* está en condiciones de aceptar 8 y 10, lo cual realiza en 13 con la elocución *un segundo*. Aquí se completa también la determinación de los argumentos de la intención “*poner x en y*” y *M* está en condiciones de comprometerse a realizar la acción solicitada por *U*, por lo que 13 expresa también un “compromiso” y hace un cargo sobre *M* mismo en el plano de las obligaciones. Con esta elocución se completa la fase de especificación de la intención; en este punto se tiene que satisfacer la relación de equilibrio para todos los actos del habla en el plano de acuerdo, y tiene que haber uno o más cargos pendientes en el de las obligaciones.

La satisfacción de las obligaciones pendientes se realiza con la acción gráfica realizada por *M* en 14; esta acción satisface simultáneamente la oferta hecha por *M* mismo en 1, la directiva de acción hecha por *U* en 3 y el compromiso adquirido por *M* en 13, como se indica por las tres llaves que se cierran en 13 en el plano de las obligaciones. Por otra parte 13 expresa también nueva información (i.e. el diagrama resultante) por lo que esta acción

⁹ Las referencias vagas (no presentes en el este ejemplo) en las que los referentes espaciales no están suficientemente determinados para actuar sobre los mismos (e.g. moverlos) la vaguedad se marca y se determina por un método análogo al utilizado para resolver la ambigüedad.

gráfica es también un acto del habla (expresado con una acción motora) que realiza una afirmación, que a su vez hace el cargo correspondiente en el plano del acuerdo mutuo. Asimismo, *M* requiere confirmar que la situación gráfica resultante de la acción gráfica es consistente con la intención expresada por *U*, por lo que realiza la solicitud de información en 15, haciendo el cargo correspondiente en el plano de las obligaciones; finalmente, en 16 *U* responde a la pregunta de confirmación de manera afirmativa; esta respuesta abona el cargo pendiente en las obligaciones, y también el cargo pendiente en el acuerdo mutuo al aceptar la información introducida por la acción gráfica en 14. Con esta elocución concluye la fase de satisfacción de la intención así como la transacción misma. Como se puede observar en la Tabla 2 todos los actos del habla satisfacen la relación de equilibrio, por lo que la transacción está balanceada. La tabla ilustra también que las relaciones de cargo y abono en cada plano obedecen a una disciplina computacional de pila (o *stack*) y que la relación de equilibrio se puede modelar mediante tres pilas que proceden de forma paralela. En este modelo una transacción es completa y balanceada cuando se han realizado las fases de especificación y satisfacción, y las tres pilas están vacías.

En este tipo de diálogos la iniciativa conversacional se comparte entre ambos interlocutores; aunque el usuario tiene la iniciativa principal, ya que el fin de la transacción es satisfacer la directiva de acción que éste expresa, independientemente de que *M* realice el primer turno conversacional, la iniciativa pasa a *M* cuando se requiere clarificar algunas informaciones. En la transacción se observan dos clases de situaciones en las que *M* toma la iniciativa: para determinar los argumentos de la predicación expresada por un acto del habla, como sucede en 5, y para reestablecer el acuerdo mutuo cuando este se rompe por una referencia ambigua (o vaga), o por una falla en el canal de la comunicación, como sucede en 8. Como se puede observar en la Tabla 2, las llaves que representan a la promesa y a la directiva de acción (la intención principal de la transacción) dominan o envuelven a las llaves en el plano del acuerdo mutuo; sin embargo, cuando éste se interrumpe, las llaves del acuerdo mutuo dominan a las llaves de las obligaciones que corresponden a las preguntas y respuestas de clarificación. Esta observación sugiere que en este tipo de diálogos la iniciativa principal es del usuario, quien comanda al sistema, pero cuando se requiere determinar los argumentos de los actos del habla, o se interrumpe el acuerdo mutuo, con el subsecuente “desenganche” entre los interlocutores, el agente que requiere información adicional (i.e. determinar los argumentos) o detecta la interrupción del acuerdo, toma la iniciativa conversacional para recabar la información faltante o iniciar la reparación correspondiente. Una vez que la reparación se realiza *M* regresa la iniciativa a *U*; por otro lado, si *U* es quien detecta la falla, se queda con la iniciativa una vez que ésta se repara, ya que el usuario es quien comanda la interacción de forma global.

Con esto se concluye la descripción del esquema DIME-DAMSL; este esquema nos ha permitido modelar 20 de los 26 diálogos útiles del corpus DIME de manera exitosa (más de doscientas transacciones) y se dispone tanto el análisis genérico de todas las transacciones, como se ilustra en la Tabla 2, así como el análisis detallado de cada transacción, el cual incluye una tabla que relaciona el conjunto completo de etiquetas del esquema con cada una de las elocuciones en la transacción¹⁰. Estos resultados nos permiten postular de manera preliminar que el esquema DIME-DAMSL es suficientemente expresivo para modelar el conjunto de los llamados prácticos.

¹⁰ http://leibniz.iimas.unam.mx/~luis/DIME/dimedamsl/CORPUS_DIME.htm

4. Modelos de Diálogos

Los diálogos prácticos con iniciativa compartida ilustrados en la sección anterior son sumamente complejos y la construcción de un sistema computacional conversacional que siga los lineamientos del esquema DIME-DAMSL es por el momento un proyecto que contemplamos realizar a mediano y largo plazo. Sin embargo, el proyecto DIME tiene también el objetivo de construir sistemas conversacionales en el corto plazo, por lo que se desarrolló un esquema para la modelación computacional de diálogos más sencillos. En estos diálogos, como en los diálogos prácticos en general, el sistema computacional funge como un asistente que ofrece un servicio al usuario humano y la iniciativa conversacional no es realmente compartida, ya que el agente computacional se limita a ofrecer un servicio mientras que el usuario se limita a informar al sistema sus requerimientos y restricciones. También, y en oposición a los diálogos ilustrados en la sección anterior, en los que las predicciones se expresan a lo largo de varios turnos y elocuciones, y el acuerdo se negocia entre los interlocutores, los actos de obligaciones que se presentan en esta sección se expresan en un solo turno y los actos de acuerdo se reducen a aceptar o rechazar a estas intenciones. Asimismo, los modelos de diálogo que se describen a continuación se estructuran en términos de protocolos conversacionales que satisfacen la relación de equilibrio, aunque de manera implícita.

En el proyecto se explora también la hipótesis de que los diálogos prácticos se estructuran en términos de un conjunto de situaciones conversacionales, cada una de las cuales corresponde a un estado informacional en el que el agente computacional tiene un conjunto de expectativas acerca de las intenciones que puede expresar su interlocutor humano y que son relevantes para el proceso de solución de tareas en la situación; la situación contiene asimismo la especificación de la acción o conjunto acciones que el agente computacional debe realizar como respuesta a cada una de dichas intenciones, y la situación conversacional a la que se llega después de realizar de las acciones correspondientes. Estas acciones pueden ser lingüísticas, como contestar a una pregunta en lenguaje hablado; gráficas, cuando la respuesta se da a través del despliegue de una imagen o un video; motoras, en el caso de que el agente sea un robot situado que se mueve como respuesta a una solicitud expresada por el usuario, o multimodales, cuando la acción que realiza el agente conversacional combina más de una modalidad.

Desde de le punto de vista del modelo computacional, en estos diálogos los interlocutores “visitan” un conjunto de situaciones, cada una de las cuales representa que cierta parte de la tarea ya se ha realizado; hay una situación inicial en que todavía no se ha llevado a cabo ninguna acción, una o más situaciones finales, y un conjunto de situaciones intermedias que se visitan durante la conversación y de manera simultánea, durante el proceso de solución de tareas. En esta investigación, al conjunto de situaciones así como sus relaciones de intención/acción para un dominio conversacional se le denomina “modelo de diálogo”. Los modelos de diálogo pueden también ser pensados como la especificación del contexto de interpretación de los actos del habla en el dominio conversacional. La especificación de este contexto, aunque de manera idealizada, simplifica significativamente la interpretación de los actos del habla en el dominio conversacional.

Los modelos de diálogo se representan como redes de transición recursiva, donde los nodos representan situaciones y los arcos las relaciones intencionales entre las mismas; cada arco tiene una etiqueta de la forma $\alpha:\beta$, donde α representa el acto del habla que tiene

que expresar el usuario para recorrer dicho acto, y β representa la acción, probablemente multimodal, que realiza el agente conversacional como respuesta a α . En el esquema actual cada acto del habla α tiene asociada una expresión regular que define un conjunto de alternativas de palabras o expresiones que pueden ser utilizadas de manera potencial para expresar la intención α en la situación correspondiente. En el proceso de interpretación la entrada textual recuperada mediante el reconocedor de voz (probablemente con ruido y algunos errores) se procesa por las expresiones regulares asociadas a los actos del habla de cada uno de los arcos que emanan de la situación, y en caso de la cadena de entrada satisfaga a una de estas expresiones, se identifica al acto del habla asociado al arco correspondiente, el cuál se recorre, ejecutando la acción β correspondiente.

Asimismo, las acciones realizadas por el agente computacional se conceptualizan como actos retóricos que tienen por objetivo realizar una función informativa, de acuerdo con los lineamientos generales de la teoría de la estructura retórica (Mann y Thompson, 1983); estos actos se piensan de manera análoga a estructuras discursivas o párrafos textuales que tienen una intención global, y cada una de sus partes, como las elocuciones u oraciones, contribuyen a realizar dicha función. Una estructura retórica cuyo fin sea explicar un concepto, por ejemplo, puede constituirse por una presentación del concepto, una elaboración, una ejemplificación y una generalización. En nuestro esquema hay un conjunto de este tipo de actos básicos definido de antemano, los cuales se definen de manera paramétrica y están asociados a una modalidad de expresión específica. En nuestro modelo cada β se define como una secuencia arbitraria de estos actos básicos y en cada modelo de diálogo se definen de antemano todos los actos retóricos utilizados por el propio modelo. Incluso, la conducta motora asociada a un acto comunicativo o intencional se conceptualiza como un acto básico que se realiza en el contexto de un acto retórico.

Tanto los actos del habla α que expresa el usuario como los actos retóricos β que expresa el sistema, así como el estado hacia el estado hacia el que se realiza la transición, se especifican normalmente por predicados que representan intenciones y acciones concretas, pero también se pueden especificar mediante funciones que tienen como argumentos a la situación actual y a la historia de la conversación (i.e. el conjunto de actos del habla expresados, de las acciones realizadas y de las situaciones recorridas desde el inicio de la conversación); estas funciones se evalúan dinámicamente durante el proceso de interpretación y tienen como valor los actos del habla o las acciones específicas que se tienen que expresar por el usuario y realizar por el sistema para navegar a través del arco correspondiente; debido a la inclusión de estas funciones en el formalismo, nos referimos a estos grafos como redes de transición recursivas funcionales. Este recurso permite incrementar significativamente la fuerza expresiva de estos protocolos o esquemas conversacionales.

Asimismo, las situaciones están tipificadas de acuerdo a modalidad de información de entrada o salida que se procesa en las mismas; los tipos de situaciones definidos en la versión actual del sistema son *de escucha*, *de habla*, *recursivas* y *finales*; las situaciones recursivas corresponden a modelos de diálogos subordinados que se evocan cuando se llega a una situación de este tipo durante el recorrido de un grafo. Cada modelo tiene una situación inicial, que puede ser de habla o de escucha, y una o más situaciones finales. Con este formalismo, un dominio conversacional se define de manera modular a partir de un conjunto de modelos de diálogo, donde estos modelos pueden incluir cero, una o más situaciones recursivas. En la Figura 5 se presenta un ejemplo de un modelo diálogo.

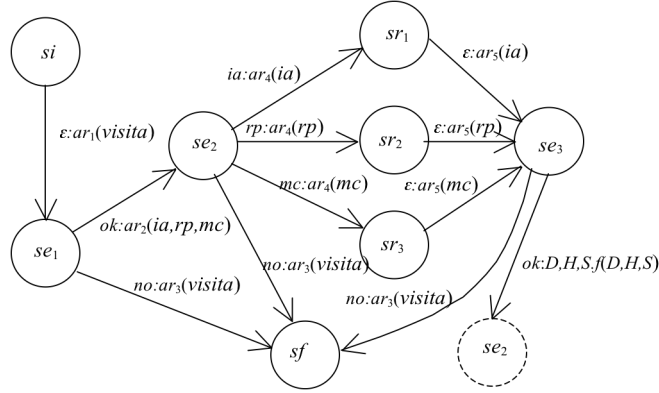


Figura 5. Modelo de diálogo

Este modelo representa el protocolo conversacional principal en el que un agente conversacional ofrece una visita guiada en la que se explican las áreas y proyectos de investigación desarrollados en el DCC-IIMAS. El diálogo se inicia en la situación de habla *si* de donde se pasa a la situación de escucha *se1*; en la etiqueta $\alpha:\beta$ entre estos dos nodos α es ϵ o vacía y la acción β (i.e. el predicado $ar_1(visita)$) se realiza incondicionalmente ya que sólo hay un arco entre estos dos nodos; esta predicación codifica al acto retórico con el que se inicia la conversación (i.e. ar_1) y como resultado de transitar por dicho arco, el sistema realiza los actos básico que constituyen a dicho acto retórico, en este caso una presentación y una oferta. Estos actos se realizan en español hablado y el sistema pronuncia *este es el departamento de ciencias de la computación* (la presentación) seguido de *quieres hacer una visita* (la oferta). En el predicado $ar_1(visita)$ “visita” es un argumento del acto retórico e indica lo que el agente conversacional está ofreciendo.

Con esta acción el agente pasa a la situación de escucha *se1*; en este punto de la conversación el agente sólo tiene dos expectativas conversacionales: que el usuario acepte o rechace la visita, y hay un arco de salida por cada una de estas opciones. En las situaciones de escucha se activa el sistema de reconocimiento de voz y el sistema espera una respuesta por parte del usuario. En el ejemplo el acto del habla $\alpha = ok$ del arco que va de *se1* a *se2* está asociado a la cadena “* γ ” donde γ se define como “*si*”, “*okay*”, “*por favor*”, “*quiero hacer la visita por favor*”, “*si, gracias*”, e incluso “*aja*”, y “*” se puede reemplazar por cualquier texto; si cualquiera de estas formas es parte de la cadena recuperada por el reconocedor de voz, el sistema asume que el usuario ha expresado la intención *ok* y selecciona la trayectoria correspondiente. A continuación se realiza el acto retórico β , especificado como $ar_2(ia, rp, mc)$, el cual se define en nuestro ejemplo como una oferta, que se realiza como *el departamento tiene las áreas de inteligencia artificial, reconocimiento de patrones y modelos combinatorios, ¿qué área quieres visitar?* Los parámetros de este acto corresponden a las opciones que se ofrecen (i.e. *ia* a inteligencia artificial, *rp* a reconocimiento de patrones y *mc* a modelos combinatorios). Después de hacer la oferta el sistema llega a la situación recursiva *sr1*, la cual corresponde a un modelo de diálogo subordinado, posiblemente con varios modelos anidados, que define el protocolo para visitar el área de inteligencia artificial. El proceso de navegación continua de manera similar hasta que se llega a la situación final del grafo. La conversación concluye cuando se alcanza la situación final del modelo de diálogo principal; en caso de que se alcance la

situación final de un diálogo anidado la conversación se resume en el modelo subordinante, en la situación que se alcanza al salir de la situación recursiva que contiene al modelo diálogo anidado.

La Figura 5 ilustra también la transición entre se_2 y se_3 en la que el acto retórico β no está especificado directamente sino mediante la función f . Esta función tiene como parámetros el identificador del modelo de diálogo actual D , la historia de toda la conversación H y la situación actual S . Por ejemplo, en caso de que este arco se atravesara una vez que ya se hubiera visitado inteligencia artificial (i.e. la situación recursiva sr_1), lo cual quedaría registrado en el valor actual de H , la evaluación de f produciría el predicado $ar_2(rp, mc)$, que a su vez se realizaría como la oferta *departamento tiene las áreas de reconocimiento de patrones y modelos combinatorios, ¿qué área quieres visitar?* Como se indica arriba, los actos retóricos son paramétricos, y los parámetros de una oferta son los identificadores de las cosas ofertadas. Asimismo, el intérprete de las situaciones conversacionales registra en la historia H los actos del habla y las acciones concretas de cada transición, así como las situaciones visitadas, a lo largo de toda la conversación. También, al igual que los actos retóricos de cada modelo, las funciones intencionales (i.e. f) se definen de antemano cuando se especifica cada modelo de diálogo.

En este esquema, los actos del habla realizados por el usuario humano se interpretan por un programa intérprete o manejador de diálogo en relación a los modelos de diálogo, como se ilustra en la Figura 6.

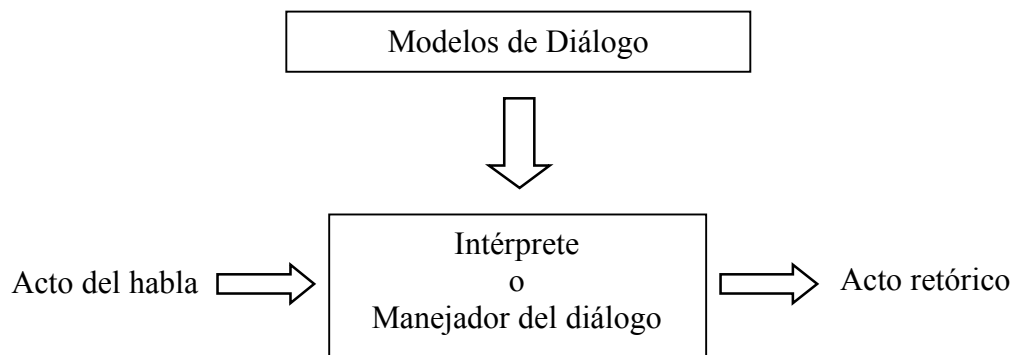


Figura 6. Sistema de administración de diálogos

El sistema intérprete sigue los lineamientos de diseño de los intérpretes de los lenguajes de programación (e.g. Basic, Prolog, Lisp) y las expresiones de entrada, en este caso los actos del habla realizados por el usuario, se interpretan a un ambiente de evaluación definido por los propios modelos de diálogo. Para la creación de este ambiente se diseñó un lenguaje para la especificación de modelos de diálogos y se programó su sistema intérprete (en el lenguaje de programación Prolog); las expresiones en este lenguaje tienen una estructura del tipo atributo valor, y permiten la definición de las redes de transición recursivas funcionales de manera declarativa. El sistema intérprete junto con el lenguaje de especificación de modelos de diálogos constituyen la herramienta central desarrollada en el proyecto DIME para la construcción de sistemas conversacionales orientados a la solución de tareas simples en español hablado, y es componente principal del robot conversacional *Golem*, desarrollado en el proyecto, como se explica más adelante.

En este esquema la conversación en español hablado procede a la par del proceso de interpretación de los modelos de diálogo; en cada ciclo de interpretación el sistema recibe

una acto del habla del sistema de reconocimiento de voz, lo interpreta en relación a la situación actual y realiza un acto retórico en una o más modalidades, y este ciclo de interpretación procede hasta que se alcanza la situación final del modelo de diálogo principal.

5. Reconocimiento de voz en español y el Corpus DIMEx100

El desarrollo de sistemas conversacionales en lenguaje hablado requiere contar una infraestructura para la creación de sistemas de reconocimiento de voz. Estos sistemas tienen por función convertir una elocución a su representación textual. Actualmente existe una tecnología sumamente desarrollada para desarrollar esta función, basada en la aplicación del modelo del canal ruidoso y la inferencia bayesiana (e.g. ver Jurafsky y Martin, 2000). Los elementos principales de este tipo de sistemas son (1) un conjunto de modelos acústicos, cada uno de los cuales modela a una unidad fonética, que a su vez se representa con un símbolo del alfabeto fonético de la lengua o léxica dialecto en cuestión, (2) un diccionario de pronunciaciones, que asocia palabra del lenguaje a una o varias pronunciaciones (representadas como secuencias de símbolos del alfabeto fonético) y (3) un modelo de lenguaje, que asigna una probabilidad a cada palabra o secuencias de palabras respecto a un corpus, normalmente recopilado en el dominio conversacional.

Los modelos acústicos se construyen comúnmente mediante modelos ocultos de Markov o HMM¹¹, y permiten la conversión de la señal de audio (representada como un conjunto de vectores de características, que a su vez se produce a partir de la señal original mediante el uso de técnicas de procesamiento de señales) en una secuencia de símbolos fonéticos, es decir, en su representación fonética; el reconocimiento se logra al alinear dicha representación con una secuencia de símbolos fonéticos formada a partir de la concatenación de una o más pronunciaciones previamente almacenadas en el diccionario de pronunciación. Cuando se logra realizar esta alineación se resuelven de manera simultánea el problema de segmentar la señal en una secuencia de palabras y el reconocimiento de la voz propiamente. Sin embargo, dada la naturaleza estocástica de los fenómenos involucrados, cada elocución proferida por el hablante da lugar no a una sino a varias hipótesis fonéticas, cada una de las cuales tienen una probabilidad en relación a todas las elocuciones posibles. El modelo del canal ruidoso pondera estas hipótesis a través del producto de la probabilidad de cada hipótesis fonética con la probabilidad *a priori* de la oración correspondiente, de acuerdo al modelo del lenguaje. Este producto representa el mejor compromiso entre la evidencia acústica que se obtiene del evento comunicativo (tomando en cuenta las características de elocución proferida, ruido en la línea, distorsiones de emisión y recepción, etc.), y la expectativa que dicha oración sea expresada en el dominio conversacional. De acuerdo con esto, por ejemplo, una elocución con mucho ruido puede ser interpretada correctamente si la oración correspondiente tiene una probabilidad alta de ocurrir en el lenguaje, y de la misma forma, una oración con baja probabilidad puede interpretarse correctamente si hay buena evidencia acústica. En general, el teorema de Bayes se utiliza para maximizar esta relación y en cada evento de reconocimiento todas las secuencias de pronunciaciones posibles (formadas con palabras en el diccionario de pronunciación) se ponderan mediante el producto de la probabilidad condicional de cada una de éstas dada la señal de audio --que se calcula por algoritmos de optimización, como

¹¹ *idem*

la programación dinámica o Viterbi¹², en base a los modelos acústicos y la información contenida en el diccionario de pronunciaciones-- con la probabilidad a priori de la secuencia léxica correspondiente, dada por el modelo de lenguaje.

En la práctica existen diversos paquetes o *toolkits* de acceso libre, como HTK o Sphinx, ambos basados en los HMM, que incluyen tanto los algoritmos para crear dichos modelos como para realizar el reconocimiento. Sin embargo, para usar estas herramientas es necesario contar con una base de datos de información fonética del lenguaje que se desea reconocer, así como del diccionario de pronunciaciones. Estos recursos son muy vastos y su recopilación es un proceso sumamente laborioso; por un lado, para crear cada modelo acústico se requiere un número muy significativo de muestras de la unidad fonética correspondiente; por el otro, un diccionario de pronunciaciones puede contener varios miles de palabras y cada una de estas puede tener varias pronunciaciones. Asimismo, para que estos recursos sean de utilidad deben recopilarse de manera empírica y analizarse a partir de una base fonética sólida para el lenguaje. Dado el estado de la tecnología, para crear reconocedores de calidad se requiere, además de conocer las herramientas y algoritmos de cómputo mencionados, contar con un recurso lingüístico con una buena base empírica.

Cuando se inició el proyecto DIME se pensó que sería posible hacerse de un reconocedor de voz para el español de México que pudiera ser incorporado de manera flexible a los sistemas conversacionales; sin embargo, esto no fue posible y se pensó en crear un reconocedor propio, mediante el uso de tecnología comercial así como de algún corpus fonético público. Sin embargo, esto tampoco fue posible debido no sólo a la escasa cantidad de recursos fonéticos disponibles, sino también a la ausencia de una base fonética sólida para crearlos. En este contexto el M. en C. Javier Cuétara realizó un análisis de los alfabetos fonéticos computacionales internacionales, así como un estudio empírico del Corpus DIME para identificar las unidades fonéticas que ocurren de manera frecuente y sistemática (Cuétara, 2004); de este estudio surgió tanto el alfabeto fonético computacional Mexbet como un censo y caracterización de las unidades fonéticas que es necesario considerar para la construcción de reconocedores de voz para el español de México. Este conjunto consiste de 37 unidades fonéticas, como se ilustra en la Tablas 3 y 4. Con esta base teórica se decidió crear un nuevo corpus fonético para fortalecer y consolida nuestra infraestructura para la construcción de sistemas de reconocimiento de voz. De esta iniciativa surgió el Corpus DIMEx100 (Pineda et al., 2004, 2008). Para la construcción del recurso se partió del Corpus-230 (Villaseñor *et al.*, 2004); este es un recurso textual extraído de Internet del cual se utilizaron las 15,000 frases de entre 5 y 15 palabras con el menor valor de perplejidad¹³; estas frases se depuraron eliminando abreviaturas, así como

¹² *idem*

¹³ Intuitivamente, la perplejidad es una medida del número de unidades en una secuencia o cadena de símbolos (e.g. palabras) que pueden seguir a una unidad dada en relación a un corpus, por lo que mientras más bajo es este valor aumenta la capacidad de predicción de la siguiente unidad. Aunque la perplejidad de las palabras es una medida indirecta de la riqueza fonética del corpus, su uso nos permitió recopilar un corpus fonético en el que aparecen todas las unidades fonéticas del español de México con un número significativo de muestras de cada unidad y con una representación proporcional a la ocurrencia de estas unidades en el habla; por lo mismo, postulamos que el corpus es un recurso empírico adecuado para la construcción de sistemas de reconocimiento de voz en dicho dialecto, lo cual fue verificado con la construcción de sistemas de reconocimiento de voz, como se explica más adelante.

palabras de otros idiomas, etc., y se editaron para su mejor lectura. De este ejercicio quedaron 5,010 frases, las cuales fueron leídas por 100 hablantes y grabadas en un estudio de alta calidad. Cada hablante grabó 50 frases asignadas individualmente y las restantes 10 frases fueron grabadas por los 100 hablantes. Los archivos de audio recopilados de esta forma con sus textos asociados constituyen la base empírica del Corpus DIMEx100.

	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar
Oclusivas sordas	[p] <i>papá</i>		[t] <i>tio</i>		[k_ j] <i>queso, kilo</i>	[k] <i>cama</i>
Oclusivas sonoras	[b] <i>van, bien</i>		[d] <i>diente, un día</i>			[g] <i>gato, un gato</i>
Africadas sordas					[tS] <i>hacha</i>	
Africadas sonoras					[dZ] <i>lluvia, yunque un yunque</i>	
Fricativas sordas		[f] <i>foco</i>	[s_ []] <i>asta</i>	[s] <i>sol, cielo</i>		[x] <i>paja, geranio</i>
Fricativas sonoras				[z] <i>mismo</i>	[Z] <i>ayer, el yunque</i>	
Aproximantes	[V] <i>haba</i>		[D] <i>hada</i>			[G] <i>el gato</i>
Nasales	[m] <i>más</i>		[n_ []] <i>antes</i>	[n] <i>nene</i>	[n~] <i>año</i>	[N] <i>angel</i>
Laterales				[l] <i>loco</i>		
“Tap”				[r(] <i>pero</i>		
“Trill”				[r] <i>perro</i>		

Tabla 3. Sonidos consonánticos

	Palatal				Cent.	Velar			
Semi-vocales / semi-consonantes	[j] <i>viene, hay</i>								[w] <i>suave, aura</i>
Cerradas		[i] <i>ahí</i>						[u] <i>su</i>	
Medias			[e] <i>meta</i>				[o] <i>lo</i>		
				[E] <i>erre</i>			[O] <i>sol</i>		
Abiertas				[a_ j] <i>aire</i>	[a] <i>la</i>	[a_ 2] <i>aunque, alma</i>			

Tabla 4. Sonidos vocálicos

Una vez creado el recurso se procedió a su análisis fonético para la identificación de las unidades acústicas, así como para la creación de los diccionarios de pronunciación. El análisis se llevó a cabo en tres niveles de granularidad fonética, los tres derivados del conjunto básico de 37 unidades. El nivel básico de análisis: T-54, T-44 y T-22. El nivel T-54 incluye 54 unidades fonéticas y distingue, además de las unidades en las Tablas 3 y 4, las vocales acentuadas de las no acentuadas tanto en sus formas abiertas como cerradas, así como los silencios de las consonantes oclusivas. El nivel medio (T-44) incluye 44 unidades e incluye también los silencios de las oclusivas y las formas básicas de los sonidos vocálicos tanto acentuados como no acentuados, además de una unidad por cada par *p/b*,

t/d, k/g, m/n y *r/rr* que, cuando aparecen en coda silábica, no tienen valor contrastante y sus realizaciones no se pueden distinguir auditivamente (e.g. opción/obción para el par *p/b*). Finalmente el nivel T-22 incluye tan sólo las unidades fonéticas que corresponden a los 22 fonemas del español, los 17 sonidos consonánticos y los 5 vocálicos. La especificación de los tres niveles está disponible en (Pineda et al., 2008). En la Figura 7 se muestra la herramienta *SpeechView* del *CSLU Toolkit* utilizada para el proceso de transcripción. En la parte superior se muestran dos espectrogramas, el primero corresponde a la frase completa (i.e. *un día llegó a casa tarde*) y el segundo es una vista aumentada a un segmento de la misma. En la figura se muestran también las transcripciones temporalmente alineadas de las unidades fonéticas en los tres niveles de transcripción; finalmente, en el último nivel se muestra la transcripción ortográfica de la palabra alineada temporalmente con sus unidades fonéticas inicial y final.

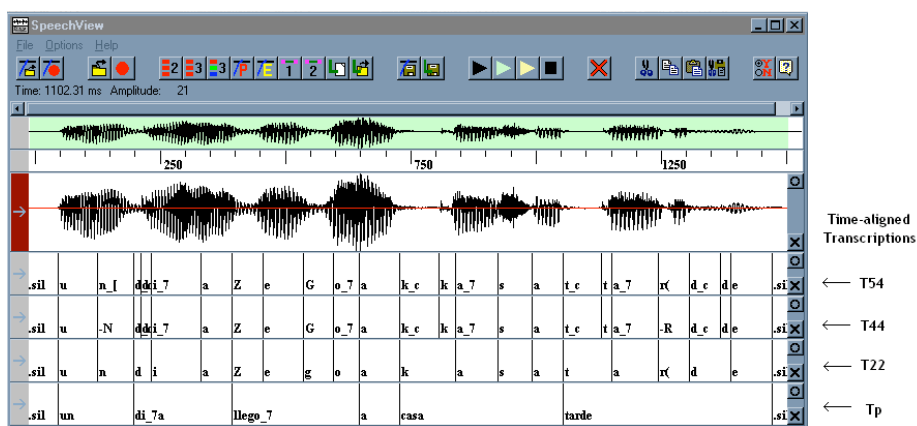


Figura 7. Análisis fonético de una frase en los tres niveles de transcripción

El análisis fonético se realizó de manera manual por fonetistas expertos – principalmente estudiantes de licenciatura y servicio social-- capacitados en el contexto del proyecto, con la ayuda de varias herramientas computacionales para editar y verificar la transcripción. El proceso se inició a principios de 2004 cuando se recolectó el Corpus y concluyó a principios de 2008, después de varias rondas de revisión. El proceso fue supervisado de manera continua por una fonetista experta y el corpus se validó mediante la construcción de varios reconocedores de voz, con diferente volumen de datos para el proceso de entrenamiento en los tres niveles de transcripción.

La granularidad de la transcripción es un parámetro que incide en la dimensión concreto versus abstracto tanto del análisis fonético como computacional para el reconocimiento de voz: mientras más fino es el análisis (i.e. el nivel T-54) más detallada es la información fonética y aumenta el número de pronunciaciones de cada palabra en el diccionario de pronunciación; por el contrario en el nivel más T-22 el número de distinciones fonéticas es menor, y disminuye el número de pronunciaciones en cada palabra. Intuitivamente, una unidad fonética en el nivel T-22 (i.e. un fonema) representa una abstracción sobre algunas unidades del nivel T-44 y algunas más del nivel T-54. De la misma manera, las pronunciaciones del diccionario en el nivel T-22 pueden considerarse como genéricas, y en caso de que sólo exista una, ésta sería la pronunciación canónica de la palabra; estas pronunciaciones pueden también pensarse como abstracciones de las formas que aparecen en los diccionarios de los niveles T-44 y T-54. La diferencia en los niveles de granularidad de la transcripción puede pensarse también en términos del costo de procesamiento:

mientras menos unidades mayor la abstracción, y las realizaciones de palabras y frases tienen que asociarse con la clase más cercana con un costo computacional menor; por otro lado, un mayor número de unidades fonéticas y pronunciaciones de palabras permite reconocer la forma exacta realizada por hablantes concretos, pero el espacio de pronunciaciones se puede incrementar significativamente. La pregunta que nos interesa en éste espectro es qué es mejor, abstraer y minimizar los requerimientos de memoria y velocidad de cómputo, por un lado, o aprovechar la información empírica detallada recabada en el corpus para hacer un reconocimiento más concreto, aunque con un costo adicional de memoria y pasos de procesamiento. Para resolver esta pregunta, a lo largo del proyecto se han programado varios reconocedores de voz que tienen como parámetros el volumen de datos (el porcentaje del Corpus utilizado para crear los modelos acústicos y los diccionarios de pronunciación) y el nivel de granularidad en la transcripción. Estos reconocedores se denominan *DIMExVol-TX*, donde *Vol* representa el porcentaje del corpus utilizado para crear el reconocedor y *X* representa el nivel de granularidad. Por ejemplo, el reconocedor *DIMEx50-T22* está construido con el 50% del corpus en el nivel T-22. Para responder a la pregunta planteada anteriormente se han construido los reconocedores *DIMEx30*, *DIMEx50*, *DIMEx75* y *DIMEx100* en los tres niveles de transcripción. Para este efecto se han utilizados modelos de lenguajes neutros (construidos a partir del propio Corpus *DIMEx100*) para medir la calidad fonética de los sistemas independientemente del dominio conversacional. En este momento tenemos resultados preliminares y todo apunta a que en nuestra plataforma el uso de la información concreta en los niveles de transcripción con mayor granularidad gana, si el volumen de datos para el entrenamiento de los modelos acústicos es suficientemente grande. Este resultado es muy interesante ya que la mayoría de los sistemas de reconocimiento de voz utilizan un número mínimo de unidades fonéticas para maximizar la abstracción, pero normalmente no se cuenta con información empírica detallada y los diccionarios de pronunciación contienen sólo las formas canónicas. Sin embargo, como lo muestran nuestros resultados, la información empírica detallada del Corpus *DIMEx100* se puede capitalizar para mejorar el desempeño fonético de los sistemas de reconocimiento con un costo moderado en memoria y poder de cómputo.

6. El robot conversacional Golem

Los resultados del proyecto DIME se conjuntaron para la creación del robot conversacional *Golem*. El robot¹⁴ cuenta con un conjunto de sensores y actuadores, así como de una interfaz de programación a nivel de sistema mediante la cual es posible leer el estado de cada sensor en un momento dado y activar el mecanismo de control de movimiento del robot; cuenta también con un sistema de navegación básico por medio del cual es posible controlar su movimiento en un espacio cartesiano; en particular, se asume que el robot está en el origen de dicho sistema al activarse, y es posible especificar la posición y orientación a la que debe moverse, así como la velocidad con la que debe desplazarse.

En el proyecto *Golem* se integró la tecnología de sistemas conversacionales y reconocimiento de voz desarrollada en el contexto del proyecto DIME para dirigir la acción del robot en el contexto de una conversación en español hablado. El elemento central de esta integración es el intérprete de los modelos de diálogos (i.e. el manejador de diálogo),

¹⁴ Un *RWI Magellan Pro*, se adquirió con el apoyo de PAPIIT, UNAM, en el año 2001.

que dirige la conducta del sistema a nivel intencional. El intérprete interpreta un conjunto de modelos hechos *ex profeso* para la aplicación de demostración, en la que *Golem* funge como guía de una sesión de carteles de los proyectos de investigación que se desarrollan en el departamento de Ciencias de la Computación del IIMAS, UNAM. Para desarrollar esta aplicación se creó un conjunto de modelos de diálogo en los que se representa tanto el esquema conversacional para llevar a cabo las visitas como el contenido conceptual al que es posible referirse durante la conversación. Como parte del ejercicio se construyeron un conjunto de carteles de los proyectos a explicar; esta labor la llevaron a cabo técnicos académicos, estudiantes de licenciatura y servicio social, tanto de ingeniería como de lingüística, familiarizados con la estructura de los modelos de diálogos. La labor de análisis se llevó a cabo en base a entrevistas a los investigadores del departamento, quienes facilitaron los textos, imágenes y videos para la construcción de los carteles; estos contenidos se hicieron accesibles al sistema mediante su inclusión en las estructuras retóricas de los modelos de diálogos y se utilizan en las respuestas y explicaciones dadas por el robot durante la visita, como se explica arriba en la Sección 4. La tecnología está suficientemente desarrollada para explorar su aplicación a otros dominios, como dar visitas a museos, construir módulos de información turística, diseñar tutores multimodales interactivos, etc.

Para efectos de la implementación computacional se adoptó el sistema de agentes distribuidos *Open Agent Architecture*, disponible como software libre Internet¹⁵. En esta arquitectura es posible asociar procesos computacionales arbitrarios a “agentes” dirigidos a realizar funciones específicas en el contexto del sistema como un todo. Estos procesos pueden definirse en diferentes lenguajes de programación como Prolog, C, C++, Perl, Java, etc., y ejecutarse en diversos sistemas operativos, como *Linux* y *Windows*, como es el caso en nuestra aplicación. El agente principal ejecuta el intérprete de los modelos de diálogos y subordina a todos los demás agentes. Se define también un agente para ejecutar cada uno de los procesos asociados con cada una de las modalidades de entrada y de salida: el agente reconocedor de voz, el agente de interpretación del texto en función de las expectativas intencionales del robot en cada situación conversacional, el agente sintetizador de voz, los agentes para el despliegue de las imágenes y videos, y el agente que controla el movimiento del robot propiamente. La arquitectura de agentes del robot *Golem* se muestra en la Figura 8.

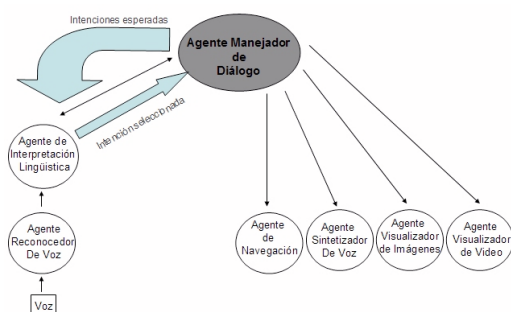
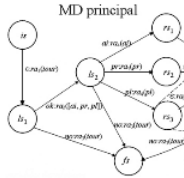


Figura 8. Arquitectura de agentes del robot *Golem*

¹⁵ <http://www.ai.sri.com/~oaa>

El robot *Golem* se presentó en el museo *Universum* en Junio de 2007¹⁶ y a partir de entonces ha causado un interés muy considerable por parte de los medios (televisión, radio, prensa escrita e Internet) como se puede apreciar en la página del proyecto¹⁷. En esta página está también disponible un video realizado por la Dirección General de Divulgación de la Ciencia de la UNAM, en el que se da una visión de las ideas generales del proyecto, se muestra el robot en funcionamiento y se da una visión detallada de los módulos que lo constituyen. Un segmento de conversación real con *Golem* se ilustra en la Tabla 5.

# TUR- NO	LENGUAJE HABLADO	DESPLIEGUE MULTIMODAL
1	GOLEM: Inteligencia artificial tiene áreas, personal y proyectos. ¿Qué quieres que te explique?	
2	USUARIO: Primero explícame los proyectos, por favor.	
3	GOLEM: Ok. Inteligencia artificial tiene los proyectos Dime, Dimex 100, Pitágoras, Golem y Reconocimiento Visual de Objetos. ¿Qué proyecto quieres que te explique?	
4	USUARIO: El tuyo Golem.	
5	GOLEM: Perdón, no te entendí.	
6	USUARIO: Que me expliques el tuyo Golem!	
7	GOLEM: Ok. Yo soy Golem. Tengo poco más de 5 años y me programaron para darte esta visita guiada. Estoy programado bajo una estructura de agentes abiertos. Tengo un agente reconocedor de voz, un ... ¿Quieres que te explique de mi manejador del diálogo?	
8	USUARIO: Si, si quiero.	
9	GOLEM: El manejador de diálogo es un intérprete de modelos de diálogo. Un modelo de diálogo es una representación de una conversación esquemática y un dominio conversacional. Se representa por un conjunto de modelos de diálogo ...	
10	GOLEM: Terminamos con el proyecto Golem. ¿Quieres que te siga explicando los proyectos de inteligencia artificial?	
12	USUARIO: Claro que si.	
	GOLEM: Inteligencia artificial tiene los proyectos Dime, Dimex 100, Pitágoras y Reconocimiento Visual de Objetos. ¿Qué proyecto quieres ver?	

¹⁶ Gaceta UNAM, No. 3995, 25 de junio de 2007, pp. 8-9.

¹⁷ <http://leibniz.iimas.unam.mx/~luis/golem/>

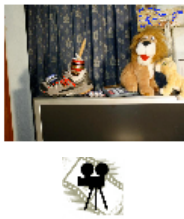
			
13	USUARIO: Muéstrame reconocimiento visual de objetos, por favor.		
14	GOLEM: Los seres humanos son capaces de reconocer objetos y además pareciera que lo hacen con gran facilidad ... Este proyecto tiene participantes, objetivo, descripción y resultados. ¿Qué sección quieres que te explique?		
	USUARIO: La de resultados.		
	GOLEM: Este video es un ejemplo de cómo una computadora ve el mundo y de cómo puede ir reconociendo los objetos que previamente ya conocía ...		
	GOLEM: Terminamos con resultados. ¿Quieres que te siga explicando reconocimiento visual de objetos?		
	USUARIO: No, ya no.		
15	GOLEM: Ok. ¿Quieres que te siga explicando los proyectos de inteligencia artificial?		
16	USUARIO: No, gracias. Ya me voy.		
17	GOLEM: ... Gracias por visitar el departamento de ciencias de la computación. Hasta luego.		

Tabla 5. Segmento de conversación con *Golem*

7. Conclusiones

En este artículo se describen los antecedentes, desarrollos y resultados principales de los proyectos DIME y *Golem*. El proyecto DIME es un proyecto a largo plazo que tiene por objetivo la creación de sistemas conversacionales en español hablado, con apoyos multimodales, para la solución cooperativa de tareas simples en dominios de interacción específicos entre seres humanos y sistemas computacionales; el proyecto *Golem*, por su parte, fue concebido para demostrar esta tecnología en un entorno útil y atractivo. De manera más teórica, estos proyectos tienen también la finalidad de explorar en que medida es posible comprender el lenguaje a través de la metáfora computacional.

Una de las premisas fundamentales de esta investigación es que el desarrollo de sistemas conversacionales requiere, además del desarrollo y uso de tecnología computacional de punta, contar con recursos lingüísticos muy amplios en los diversos niveles de representación lingüística, tanto para llevar a cabo las labores de análisis como para la creación de los sistemas propiamente; por lo mismo, una parte muy significativa del esfuerzo se ha centrado en el diseño, recolección y análisis del Corpus DIME y el Corpus DIMEx100. El primero es un recurso para el análisis de la conversación en varios niveles de representación lingüística, como el fonético y fonológico, el entonativo y prosódico, el léxico, morfológico, sintáctico y semántico, y finalmente, el nivel pragmático-intencional.

El Corpus DIMEx100, por su parte, es un recurso fonético y léxico de carácter productivo para la creación de sistemas de reconocimiento de voz, especialmente para la creación de modelos acústicos y diccionarios de pronunciación. La creación de estos recursos es compleja ya que además de su diseño, así como el diseño y construcción de las metodologías y herramientas para recolectarlos, es necesario contar con esquemas de transcripción apropiados, los cuales no siempre están disponibles de antemano. Estos esquemas presuponen una teoría, no siempre muy elaborada, acerca del nivel de representación lingüístico correspondiente, así como su relación con otros niveles, y en nuestra experiencia, aunque normalmente se parte de una concepción teórica inicial, ésta se modifica y enriquece con el avance de la tarea de transcripción; por lo mismo, la labor de análisis se desarrolla a la par de la teoría, y este proceso coordinado madura de manera incremental. Asimismo, el grado de madurez de estas teorías depende de qué tanto las intuiciones lingüísticas, reflejadas en el esquema de transcripción, tengan una base empírica, es decir, en los fenómenos que se exhiben en los recursos lingüístico. Por estas razones, la recolección y análisis de los Corpus DIME y DIMEx100 ha sido un proceso incremental que aún está abierto.

Es importante señalar también que la recolección y análisis de dichos recursos, así como su uso en la construcción de sistemas computacionales, involucra la participación tanto de científicos e ingenieros computacionales como de lingüistas en sus diferentes especialidades, por lo que el proyecto tiene una orientación multidisciplinaria. Asimismo, el proyecto involucra a varias especialidades computacionales, y también de varios niveles de representación lingüística, así como las relaciones entre los mismos, lo que da al proyecto un carácter interdisciplinario, tanto en el la computación como en la lingüística.

En el transcurso del proyecto se han recorrido diversas concepciones acerca del lenguaje y de la forma de abordar la construcción de sistemas conversacionales. Inicialmente se exploró una concepción formalista del lenguaje, que sostiene que la sintaxis es el nivel principal de representación lingüística y que este nivel de representación es, en gran medida, independiente de otros niveles de representación. Esta visión sugiere una arquitectura de procesamiento “en línea” (*pipeline*) y de “afuera hacia adentro” o *bottom-up*, en el que la señal acústica se convierte a una representación fonética/fonológica, que a su vez se interpreta como un objeto léxico, y a partir de este se obtiene la estructura sintáctica de la oración, ya que ésta expresa la proposición o significado conceptual de la oración de manera independiente del contexto; esta estructura es posteriormente enriquecida con procesos locales que resuelven las referencias anafóricos e indexicales, para obtener la forma lógica de la oración. De acuerdo con esta arquitectura, una inferencia posterior asocia dicho significado a un contexto de interpretación, y finalmente a las intenciones que tienen los hablantes al proferir dichas elocuciones u oraciones. Esta visión da también un peso muy importante a las intuiciones subjetivas acerca de la forma de “la oración gramatical” que, dicho sea de paso, está influenciada por una tradición muy profunda basada en el estudio del lenguaje textual. Como consecuencia de esta visión, en la primera etapa del proyecto se hizo un esfuerzo muy considerable en definir una gramática del español así como su sistema de parseo. Sin embargo, y a pesar del éxito razonable alcanzado en estos objetivos, quedaba todavía pendiente el problema de cómo asociar el significado “canónico” de la oración a las intenciones y expectativas intencionales que tienen un agente en un estado dado de interpretación; nuestra experiencia sugiere que en esta arquitectura “pipe-line” la complejidad de esta tarea es muy grande, y probablemente insuperable.

Por otro lado, el análisis del corpus DIME y el tener que dar cuenta de los fenómenos típicos del habla espontánea, mostraron de manera muy clara que muchas intuiciones sintácticas tradicionales son idealizaciones que carecen de un sustento empírico sólido en el habla espontánea; por ejemplo, aunque la oración textual bien formada corresponde a la elocución normalmente hay muchas excepciones, y no se cuenta con una noción formal de elocución que tome en cuenta a la entonación, a las disfluencias y reparaciones, y a la intención, que sea ampliamente aceptada. Asimismo, los estudios fonéticos y prosódicos sobre el Corpus DIME y el Corpus DIMEx100 sugieren que la información contenida en estos niveles de representación puede tener un impacto significativo en el proceso de la información léxica, sintáctica y pragmática. Por ejemplo, en el español el acento tienen una función fonológica (e.g. *ánimo*, *animo*, *animó*) y la distinción entre vocales acentuadas y no acentuadas permite resolver a nivel fonético la ambigüedad léxica en la que se incurre si no se toma en cuenta esta distinción, como en el caso de los reconocedores de voz que sólo consideran los 22 fonemas básicos del español, así como en los sistemas de parseo asociados que tienen que resolver esta ambigüedad en el nivel sintáctico. Adicionalmente, los resultados del proyecto muestran que existe una gran variabilidad fonética en el habla, incluso en el caso de la lectura controlada, y cada palabra tiene varias formas alofónicas empleadas comúnmente, y esto es más notorio si el nivel de transcripción es más fino; por ejemplo, las palabras *universidad* e *información* tienen más de 30 pronunciaciones diferentes en el nivel T-54 del corpus DIMEx100; por lo mismo, un diccionario de pronunciación rico en pronunciaciones puede aumentar la tasa de reconocimiento de manera significativa con un costo computacional razonable. La entonación es también una fuente sumamente rica de información y nuestros resultados actuales muestran que este factor contribuye de una manera muy significativa (aunque no es suficiente en sí mismo) para identificar el tipo de acto del habla o la intención expresada por una elocución.

Por otra parte, el análisis del Corpus DIME muestra que una cantidad muy significativa de elocuciones son formas incompletas, que van desde las interjecciones, palabras y frases incompletas, cláusulas que funcionan como complementos directos e indirectos, hasta las predicaciones completas, aunque en todos estos niveles se observan los fenómenos de disfluencias y reparaciones, y cada unas de estas unidades puede considerarse una unidad entonativa e intencional completa.

Finalmente, la arquitectura de “pipe-line” presupone también, cuando menos en la práctica, un role relativamente limitado para la noción de “contexto de interpretación”, el cual, en algunas vertientes, se reduce simplemente al conjunto de oraciones expresadas de manera previa a la oración que se requiere interpretar. Esta perspectiva se enfoca también a la solución de problemas muy específicos de la interfaz semántico-pragmática, y en particular a la resolución de las referencias anafóricas; sin embargo, sin devaluar la importancia de este fenómeno, este es apenas uno de los factores que conforman el contexto de interpretación.

Por todas estas razones durante el 2002 y 2003 se revisó y replanteó el enfoque del proyecto y se decidió traer al primer plano el problema de la comunicación; es decir, se decidió centrar el análisis en el nivel de representación pragmático, en el que se representan las intenciones de los agentes embarcados en la interacción lingüística. Independientemente que el conjunto de intenciones posibles que puede tener un ser humano está abierto, el número de tipos de intenciones posibles es mucho más limitado; más aún, los diálogos prácticos, cuya modelación computacional parece ser viable, se enfocan a la solución cooperativa de una tarea relativamente sencilla entre dos agentes en dominios

problemáticos específicos, y se propone que los tipos de intenciones que se expresan en estos diálogos constituyen un conjunto finito, y que el número y restricciones sobre sus argumentos se pueden también identificar de antemano a través del análisis.

Esta reflexión motivó aumento el interés en el análisis de la conversación, especialmente a nivel intencional, tanto desde el punto de vista teórico como práctico; del primero surgió el esquema DIME-DAMSL y del segundo el concepto de modelos de diálogos y su programa intérprete. En ambos frentes, la experiencia del proyecto sugiere que el nivel de interpretación pragmático actúa de “arriba hacia abajo” o de manera *top-down* y las expectativas intencionales codificadas en los modelos de diálogo intervienen en la determinación de la interpretación que se da a los texto recuperados por el sistema de reconocimiento de voz de manera *bottom-up*, en función de las expectativas intencionales del agente computacional en la situación conversacional.

Finalmente, la experiencia acumulada en el proyecto sugiere una noción amplia de contexto de interpretación que incluye, además del discurso previo y la situación espacial y temporal, esenciales para resolver las referencias anafóricas e indexicales, el conjunto de protocolos conversacionales en los que se lleva a cabo la conducta lingüística así como las expectativas intencionales de los agentes. Por esta razón, en esta investigación se concibe a los modelos de diálogo como representaciones del contexto conversacional, y estas representaciones hacen posible la interpretación del lenguaje de manera efectiva, a pesar de que las representaciones en los niveles fonético/fonológico y léxico/sintáctico/semántico sean incompletas; además, esta representación permite definir procesos para atender a las fallas de la interpretación en los diferentes niveles, con lo que es posible la construcción de sistemas razonablemente robustos. Todos estos razonamientos se llevaron a la práctica en el proyecto *Golem* con resultados muy prometedores, como ya se ha mencionado.

Por otra parte, nuestro enfoque, basado en la concepción de que el sistema de administración de diálogo como un intérprete, en el sentido formal de los lenguajes de programación, y en la noción de modelos de diálogo definidos a nivel intencional, en los que se relacionan las expectativas intencionales con las acciones realizadas por el agente en cada acción de interpretación, y donde las acciones complejas dependen de acciones simples, codificadas como estructuras retóricas multimodales, sugiere también que, en oposición a los procesos de pensamiento, como las inferencias deductivas o procesos de planeación, por ejemplo, los modelos de diálogo contienen información de carácter lingüístico, y que las inferencias que permiten modelar son también de este tipo. De la misma manera que generar una explicación multimodal requiere la realización de actos retóricos básicos, orientados a modalidades específicas, los procesos de pensamiento, como resolver un problema, realizar un tarea de planeación, o “pensar” una movida en un juego, son acciones que realiza el agente como parte de la respuesta a la interpretación de una intención, y pueden conceptualizar como conductas embebidas dentro de los actos retóricos multimodales; de hecho, en la arquitectura de agentes del robot *Golem* podrían agregarse de manera muy sencilla agentes para cada una de estas actividades (e.g., probadores de teoremas, verificadores de modelos, sistemas de planeación, etc.), enriqueciendo de la variedad de conductas desplegadas por el *robot*.

Agradecimientos

El autor agradece la participación de varios colaboradores y estudiantes en los proyectos DIME y *Golem* a lo largo de los años. Se agradece en particular la participación del Dr.

Luis Villaseñor y el M. en C. Antonio Massé, quienes recolectaron y coordinaron la transcripción y segmentación del Corpus DIME, y a todos los técnicos académicos y estudiantes que participaron en estas tareas; del M. en C. Javier Cuétara quien identificó el conjunto de alófonos para el español de México y sus contextos fonéticos en base al análisis empírico del Corpus DIME y propuso la forma actual del alfabeto fonético Mexbet; Se agradece también de manera muy especial la participación de Varinia Estrada en la definición del esquema DIME-DAMSL y la transcripción detallada de 20 diálogos del Corpus DIME en este esquema, así como a los demás estudiantes y colegas que participaron en este proceso; al Dr. Sergio Coria por su contribución al estudio de la entonación y su relación con los actos del habla, por su contribución al desarrollo del esquema DIME-DAMSL y por la transcripción entonativa del Corpus DIME en el esquema INTSINT; del Ing. Ivan Moreno por el análisis de las disfluencias y reparaciones de habla en el Corpus DIME y a los estudiantes que participaron en la transcripción de estos fenómenos; de la M. en C. Fernanda López, la Lic. Ana Helena Ceballos y la Lic. Alejandra Espinoza por sus contribuciones a la fonética del español de México en base al análisis empírico del Corpus DIME. Se agradece muy especialmente la participación del M. en C. Ivan Meza quien desarrolló las gramáticas y parser del español en HPSG y LKG, por su contribución al desarrollo de la teoría de clíticos y la perífrasis del español, por la programación del primer manejador de diálogo de *Golem*, escrito como un autómata de estados finitos, y por la integración del primer sistema de demostración del robot *Golem*. Se agradece también la participación del Dr. Arturo Espinoza, quien habilitó originalmente el sistema de navegación del robot *Golem* y del M. en C. Miguel Salas y el Ing. Paulino Ochoa, quienes definieron el ambiente OOA y el puente de programación entre OOA y el sistema de navegación de *Golem* escrito en Corba. Se agradece también la participación del Dr. Luis Villaseñor por su participación en el diseño del Corpus DIMEx100, de la Lic. Hayde Castellanos quien coordinó la recolección de dicho corpus y supervisó su transcripción, así como a los 100 hablantes que donaron su voz para la creación de este recurso. Asimismo, a los numerosos estudiantes que participaron en el proceso de transcripción fonética, especialmente a Varinia Estrada, Fernanda López, Ana Helena Ceballos y Alejandra Espinoza. Se agradece muy especialmente la participación de la Ing. Elia Patricia Pérez Pavón por la definición del ambiente para la creación de sistemas de reconocimiento de voz con Sphinx y el Corpus DIMEx100, por la creación del reconocedor DIMEx30-T22, incorporado en la versión actual del robot *Golem*, y por la programación de los agentes de reconocimiento y síntesis de voz; se agradece también de manera muy especial la participación de la M. en C. Wendy Aguilar por la integración de la plataforma actual de *Golem*, la programación de los agentes para el despliegue de imágenes y videos, y la creación de varios de los modelos de diálogo; asimismo, por su apoyo en las presentaciones *en vivo* del robot *Golem* en diversos foros nacionales, así como a todos los técnicos y estudiantes que han participado en estas demostraciones. Se agradece también la participación de Varinia Estrada así como de varios otros estudiantes por la creación de los contenidos y carteles, y por la creación de varios modelos de diálogo del sistema de demostración del robot; a Janet Juárez por la creación de los reconocedores DIMEx50, DIMEx75 y DIMEx100 y a Irving Alvarez, por su contribución en la validación computacional del Corpus DIMEx100. Se agradece también la colaboración del Dr. Joaquim Llisterri de la Universidad Autónoma de Barcelona en los aspectos de fonética, fonología y entonación del proyecto a lo largo de varios años; del Dr. Lucian Galescu del Institute for Human and Machine Cognition (IHMC) en Florida por su apoyo constante

para el desarrollo de la plataforma de reconocimiento de voz con *Sphinx*, y al Dr. James Allen de la Universidad de Rochester y del IHMC, por su colaboración constante desde el inicio del proyecto. Se agradece también la participación de muchos otros colaboradores y estudiantes en diversos aspectos a lo largo de la vida del proyecto. Asimismo, se agradecen los comentarios y sugerencias por parte de varios colegas tanto nacionales como en el extranjero que han visitado al DCC-IIMAS y participado en las reuniones de trabajo de los proyectos DIME y *Golem*. También al Depto. de Video de la Dirección General de Divulgación de la Ciencia y el museo *Universum* de la UNAM por la creación del video del robot *Golem*. Finalmente, se agradece el apoyo de los proyectos NSF/CONACyT C092 y 39380-A, del proyecto CONACyT 27948-A y de los proyectos PAPIIT-UNAM IN-111700 e IN-121206.

Referencias

Allen, J.F., & Core, M. 1997. Draft of DAMSL: Dialogue Act Markup in Several Layers, Technical Report, The Multiparty Discourse Group. University of Rochester, Rochester, USA.

Allen, J.F.; Byron, D. K.; Dzikovska, M.; Ferguson, G.; Galescu, L. & Stent, A. 2000. An Architecture for a generic dialogue shell. *Natural Language Engineering*, 6(34):213–228.

Allen, J.F.; Byron, D. K.; Dzikovska, M.; Ferguson, G.; Galescu, L. & Stent, A. 2001. Toward Conversational Human-Computer Interaction. *AI Magazine*, 22(4):27–38, Winter.

Coria, S. R y Pineda, L. A. 2008. An analysis of prosodia information for the recognition of dialogue acts in a multimodal corpus in Mexican Spanish, Enviado a *Computer Speech and Language*, Elsevier.

Cuétara, J. 2004. Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla. Tesis de Maestría, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México, México.

Hirst, D. DiCristo, A. Espesser, R. 2000. Levels of representation and levels of analysis for the description of intonation systems, in M. Horne (ed) *Prosody: Theory and Experiment*, Kluwer-Dordrecht.

Jurafsky, D., Martin, J. H. 2000. *Speech and Language Processing*, Prentice Hall, New Jersey.

Levinson, S. C. 1983. *Pragmatics*, Cambridge: Cambridge University Press.

Levinson, S. C. 2000. *Presumptive Meanings*, Cambridge: A Bradford Book, The MIT Press.

Mann, W. C., Thompson, S. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization, *Text* 8(3), pp. 243—281.

Moreno, I., Pineda, L. A. 2006. Speech repairs in the DIME Corpus, *Research in Computing Science* Vol. 20: 63 – 76.

Pineda, L. A., Villaseñor, L., Cuétara, J., Castellanos, López, I. 2004. DIMEx100: A new phonetic and speech corpus for Mexican Spanish, en *Advances in Artificial Intelligence*,

Iberamia-2004, C. Lemaitre, C. A Reyes & J. A. Gonzalez (Eds.), *Lectures Notes in Artificial Intelligence* 3315, pp. 974–983. Springer-Verlag, Berlin Heidelberg.

Pineda, L. A., Castellanos, H., Coria, S., Estrada, V., López, F., López, I., Meza, I., Moreno, I., Pérez, P., Rodríguez, C. 2006. Balancing transactions in practical dialogues, *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 331–342.

Pineda, L. A., Castellanos, H., Cuétara J., Galescu, L., Llisterri, J., Pérez P., y Villaseñor, L. 2008. The Corpus DIMEx100: Transcription and Evaluation. Enviado a *Language Resources and Evaluation*, Springer-Verlag.

Pineda, L. A., Estrada, V., Coria, S. y Allen, J. 2007. The obligations and common ground structure of practical dialogues, *Revista Iberoamericana de Inteligencia Artificial*, Vol. 11 (36): 9–17.

Pineda, L. A., Meza, I. 2005. The Spanish Pronominal Clitic System, *Procesamiento del Lenguaje Natural* 34: 67–103.

Turing, A. 1950. Computing Machinery and Intelligence, *Mind*, 59: 433-460.

Villaseñor, L., Montes & Gómez, M., Vaufreydaz, D. & Serignat, J. F. 2004. Experiments on the Construction of a Phonetically Balanced Corpus from the WEB, Proceedings of CICLING2004, LNCS, Springer-Verlag, Vol. 2945, 416-419.